Designing Exploratory Search Systems that Stimulate Memory and Reduce Cognitive Load

Savvas Dimitrios Petridis

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Designing Exploratory Search Systems that Stimulate Memory and Reduce Cognitive Load

Savvas Dimitrios Petridis

From music fans finding new songs in a genre, graphic designers brainstorming ways to depict a message, and journalists scrutinizing documents for angles, people often conduct exploratory searches to understand complex topics. In contrast to traditional search, which is done to quickly answer simple questions, exploratory search is an iterative learning process that involves understanding an information space in order to find useful pieces of information. Exploratory search is composed of two, closely-related sub-processes: (1) *information foraging*, choosing sources and collecting information, and (2) *sensemaking*, organizing this information into a mental framework. Both of these sub-processes are cognitively taxing and heavily rely on our memory. For information foraging, users need to read long, complex resources and recognize useful pieces of information. For sensemaking, as users encounter more information, it becomes harder to relate new information to their current knowledge. The spreading activation theory of memory purports that the information we encounter materializes in our working memory, which spreads activation into our long-term memory, enabling us to recall related semantic information to make sense of newly found information. From this theory, this thesis introduces three strategies for creating organizations that better stimulate memory: (1) constructing overviews that are *association networks* that mimic our memory's structure, (2) incorporating our *prior knowledge* in these overviews, and (3) providing *concrete* information to help us make sense of abstract ideas. This thesis demonstrates how to employ these strategies through three exploratory search systems across three domains: (A) SymbolFinder helps graphic designers explore visual symbols for abstract concepts, (B) TastePaths helps music fans explore artists within a genre, and (C) AngleKindling supports journalists explore story angles for a press release. Through this body of work, I demonstrate that by designing exploratory search systems to stimulate our memory, we can make acquiring and making sense of knowledge less cognitively demanding.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I'd like to thank Lydia, who on top of being an insightful and supportive advisor, is a genuinely kind person and friend. Thanks to her guidance, I can frame research questions, write a solid CHI introduction, and quote most of Bojack Horseman. Next, my co-PhDer Katy, who made graduate school infinitely more enjoyable through her humor, warmth and impeccable taste in comics and bright green pants. Vivian, who was always thoughtful, helpful, and kind; thank you for the Tupperware of Nous Cold Brew. Nous Cold Brew, for always being there for me. Columbia University and all those who were part of my Bachelors and PhD experience. My parents, Elpida and Dimitri, who through their steadfast support and love, have made me the person I am today. And finally, Auriane, this PhD is as much yours as it is mine; I could not have done it without you.

To my parents and Auriane.

Chapter 1: Introduction

The advent of the Internet has made an extraordinary amount of information easily accessible. To help users quickly sift through this data, early information retrieval (IR) researchers developed look-up based systems where users input questions, such as “Who are the members of the The Beatles?” and the system responds with an answer [1]. While this look-up pattern is effective for question-answering, it does not adequately support search tasks that involve learning about a *complex* topic, such as students learning about dense subject areas, work teams researching solutions for products, and scientists investigating complex phenomena [2]. Unlike answering a simple question, these tasks require an open-ended exploration of a broad topic and require support for understanding the information space surrounding the topic, as well as identifying useful sources of information [3] [4]. Ultimately, exploratory search is a learning process and requires support beyond looking up answers to questions.

Exploratory search consists of two intertwined processes: (1) information foraging and (2) sensemaking. Information foraging is the process of collecting and extracting information from resources [5]. As one explores, they opportunistically “forage” for information, making decisions on which web page or resource to explore based on its “information scent”, which is their “(imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues, such as bibliographic citations, WWW links, or icons representing the sources” [6]. After choosing a resource based on its scent, exploratory searchers then “exploit” it, e.g. carefully read it [7] and begin the sensemaking process. While information foraging involves collecting data, sensemaking involves organizing this data: “sensemaking is the process of searching for a representation and encoding data in that representation to answer task-specific questions” [8]. Finally, these two processes bleed into each other: users forage for information, organize it, realize that there is a gap in this organization, and then forage for more information to fill this gap [7].

Exploratory search is a loop that involves collecting and mentally organizing information.

Information foraging and sensemaking are cognitively taxing processes that heavily rely on our memory. [8]. A fundamental challenge of information foraging is the “cost of having to actually work through the material and eventually exploit it” [7]. Reading takes time and mental energy. And as users read through long, complex resources, they rely on their memory to recognize useful pieces of information, through information scent, to direct their search [6]. And if information scent is misleading or not immediately understandable, users will inevitably expend time and energy exploiting resources that contain irrelevant or uninteresting information [6]. Finally, when users find useful information, they have to add it to their mental organization. One challenge with this is that “human working memory has inherent capacity limits” [7]. As they encounter more and more information, users have a hard time recalling and relating their current knowledge to new information. Because of this, it becomes difficult to organize information and make informed decisions on where to explore next. Overall, exploratory search is cognitively difficult and relies on our ability to recall related semantic information.

To help users learn more easily, exploratory search systems should stimulate our memory so that we can better forage and make sense of new information. However, current exploratory search systems do not do enough to stimulate our memory. There are three broad categories of exploratory search systems: (1) query-expansion-based interfaces, (2) cluster-based interfaces, and (3) network-based interfaces, and each could better help us recall related information. Query expansion provides related keywords and images that are close to the original query. And while these keywords might help users recognize a useful direction to take their search, they often are unorganized and **could be connected together to create a stronger information scent** and stimulus for our memory. Cluster-based interfaces go a step further than query expansion and proactively organize pieces of information into groups to construct an overview. However, within these clusters, **information pieces are still very abstract**, like entire documents, and are often very hard to compare and connect to each other. Finally, network-based interfaces explicitly link information together, often in more understandable ways than cluster-based interfaces, but they still operate with pieces of information that are too abstract. And at the same time, these interfaces do not incorporate the user's prior knowledge to help them **connect new information to what they already know**. As a result, users are (1) left to tediously rediscover knowledge they have already encountered and (2) made to make sense of new information without context and background information. By designing exploratory search systems with our memory in mind, we can make learning less cognitively taxing.

This thesis applies memory theory to better organize information in exploratory search systems so that we can make learning less cognitively taxing. We dive deep into a prominent theory of memory, which posits that human semantic memory is a network, where each node is a piece of information and each edge represents an association [9] [10]. When we encounter a piece of information, its node is activated in our memory, and then this activation spreads to that node's neighbors, triggering related thoughts and ideas. From this spreading activation theory of memory, we derive three memory strategies to incorporate into exploratory search systems:

- **Association network**. Organize the information overview into an association network. By doing so, the overview's structure mimics that of our memory and helps group information together to create a stronger stimulus for our memory.
- **Prior knowledge**. Incorporate the user's prior knowledge into the overview. Illustrating connections between the user's prior knowledge and new information helps users recall additional semantic information to make the new information stick better to what they already know.
- **Concreteness**. Concretize abstract information. Concrete information is more salient in our memory, and by providing concrete examples for abstract information, we better integrate this abstract knowledge with our current knowledge.

These three memory strategies are incorporated into three exploratory search systems: **SymbolFinder**, **TastePaths**, and **AngleKindling** (Figure 1.1). All three systems incorporate *association network* and *concreteness*, while **TastePaths** also incorporates *prior knowledge*:

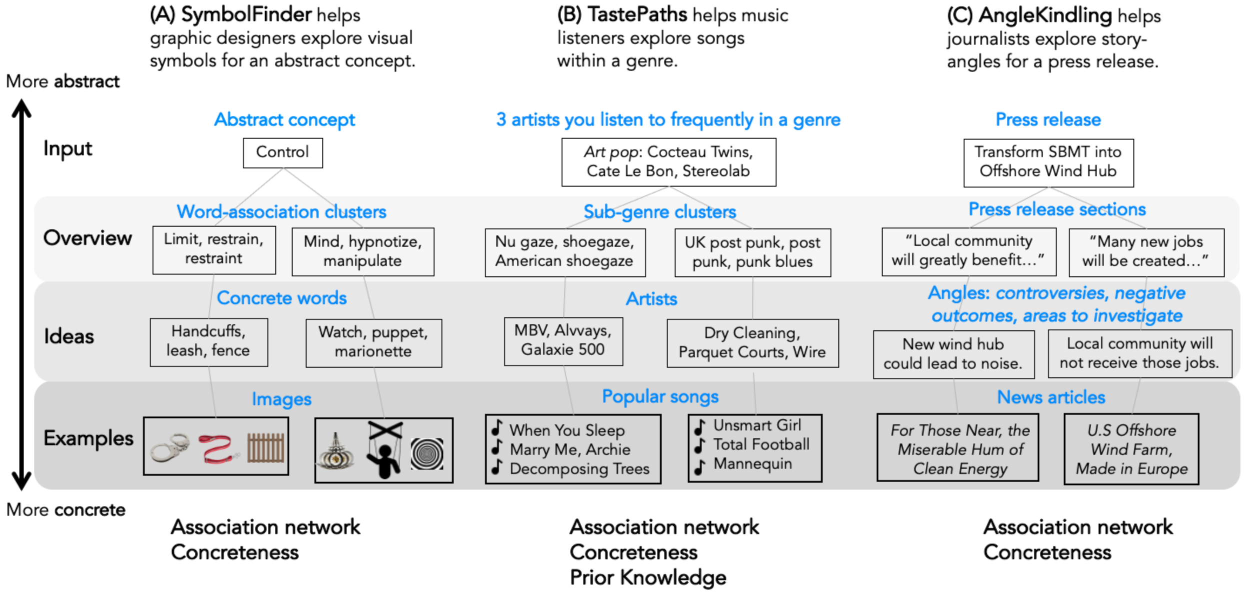

(A) SymbolFinder helps graphic designers explore visual symbols for an abstract concept.
(B) TastePaths helps music listeners explore songs within a genre.
(C) AngleKindling helps journalists explore story-angles for a press release.
More abstract
More concrete
Input
Overview
Ideas
Examples
Abstract concept
Control
Word-association clusters
Limit, restrain, restraint
Mind, hypnotize, manipulate
Concrete words
Handcuffs, leash, fence
Watch, puppet, marionette
Images
3 artists you listen to frequently in a genre
Art pop: Cocteau Twins, Cate Le Bon, Stereolab
Sub-genre clusters
Nu gaze, shoegaze, American shoegaze
UK post punk, post punk, punk blues
Artists
MBV, Alvvays, Galaxie 500
Dry Cleaning, Parquet Courts, Wire
Popular songs
When You Sleep
Marry Me, Archie
Decomposing Trees
Unsmart Girl
Total Football
Mannequin
Press release
Transform SBMT into Offshore Wind Hub
Press release sections
"Local community will greatly benefit…"
"Many new jobs will be created…"
Angles: controversies, negative outcomes, areas to investigate
New wind hub could lead to noise.
Local community will not receive those jobs.
News articles
For Those Near, the Miserable Hum of Clean Energy
U.S Offshore Wind Farm, Made in Europe
Association network
Concreteness
Association network
Concreteness
Prior Knowledge
Association network
Concreteness

Figure 1.1: **Three exploratory search systems that better stimulate our memory: *SymbolFinder*, *TastePaths*, and *AngleKindling*.** Each takes in an abstract input, such as a concept, entire genre, or a convoluted press release, and constructs an association network to serve as an overview. Each system guides users to progressively more concrete information. And finally, TastePaths also incorporates the user's prior knowledge into this overview.

- **SymbolFinder** helps graphic designers create visual metaphors by exploring multiple, diverse symbols for abstract concepts. From an abstract concept, SymbolFinder constructs an association network consisting of word-association clusters that capture the meanings and contexts of the abstract concept. When diving deeper into each cluster, users explore progressively concrete information like concrete words which are potential ideas for symbols, as well as images (Figure 1.1A).
- **TastePaths** helps music fans explore and discover new artists and songs within a genre. From the user's prior knowledge, i.e. three artists the user listens to frequently in a genre, TastePaths constructs an association network consisting of the many sub-genres within a larger genre. Users explore progressively concrete information like artists and songs within each cluster to find new music (Figure 1.1B).
- **AngleKindling** helps journalists explore angles for story ideas, given a press release. From

a long press release, AngleKindling splits the document into an overview of sections. Each section is then associated with a set angles like, potential controversies and negative outcomes, which might inspire a story. And finally, for each angle, a concrete news article is provided for background context (Figure 1.1C).

1.0.1 Contributions

The complete list of this dissertation's contributions include the following:

Concepts and Techniques

- Three design strategies for stimulating memory in exploratory search: (1) association network, (2) prior knowledge, and (3) concreteness.
- Incrementally generating an association network from an item in a knowledge graph and using network centrality and other network-properties to create an organized view of the data.
- Using a few-shot LLM-prompt to construct a search space given a document.
- Sorting items within clusters by concreteness and relevance to help users quickly make sense of abstract information.

Artifacts

- SymbolFinder, a system which helps novice graphic designers *explore* visual symbols for abstract concepts.
- TastePaths, a system which helps music listeners *explore* and find songs to listen to in a genre.
- AngleKindling, a system which helps journalists *explore* story angles for a press release.

Experimental Results

- Three formative studies which illustrate that users have trouble exploring the diverse elements of an information space because of fixation and the limits of their memory.

- A comparative user study with 10 novice designers, which demonstrates that SymbolFinder helps users find 50% more symbols with significantly less mental demand and effort. This result supports that an association network and concreteness helped users remember and explore the diverse meanings associated with an abstract concept.
- A study with two versions of TastePaths (with and with-out prior knowledge) demonstrating that prior knowledge is very useful for exploring and understanding an overview.
- A study illustrating that AngleKindling was perceived to be significantly more helpful for thinking of story ideas with less mental demand than a prior journalistic angle-ideation tool that provides less concrete suggestions.

Chapter 2: Background

In the following section, I introduce exploratory search and distinguish it from the most common interaction pattern for accessing data, lookup. Then I explain two essential processes of exploratory search: (1) information foraging and (2) sensemaking. These two processes depend on our memory, and to explain this, I go over a prominent theory of memory that posits our memories are nodes in a semantic network. Finally, I look at current, interactive exploratory search systems and discuss how they do and do not apply memory theory to their design.

2.1 Exploratory search

There is an extraordinary abundance of information available on the web, in libraries, newspapers, and encyclopedias, too much for an individual to sift through without help. To support users in quickly surfacing the content they need, information retrieval (IR) researchers developed lookup-based retrieval models, where users enter a text query and the system surfaces relevant documents in a sorted list [1]. This lookup interaction pattern is used by most major Web search engines, such as Google [1] and Bing [2], and is the most prominent way we seek information.

While lookup search effectively supports question-answer scenarios, it does not suffice when users' goals are less certain and more open-ended. Instead of answering a specific question, users conduct exploratory searches to "improve their understanding of a topic", and they often tackle problems that are "ill-structured" and require "additional information from external sources" to help clarify their goals and actions [11]. The exploratory search process is essentially a learning process, where users learn more about the topic they are exploring and refine their goals: "During exploratory searches, it is likely that the problem context will become better understood by the

[1] https://www.google.com/

[2] https://www.bing.com/

searcher, allowing them to make more informed decisions about interaction of information use" [4]. Finally, as the user has learned more about the topic, they might transition to "focused searching", where they conduct more look-up oriented searches and spend time extracting information from targeted areas in the information space. All in all, exploratory search involves learning about a topic, and as users learn, their goals and interests become more concrete and targeted.

To support learning, exploratory search systems need to help users understand the diversity of information around the topic and easily understand the information they encounter. By presenting users a diverse set of results for the topic, the system helps them "learn more about an entire subject area topic" [4] and also helps them with "orienteering", the process through which users employ their "recall and recognition skills" to determine relevant areas in the information space [12]. As well as providing an overview to help users situate themselves, exploratory search systems should also provide assistance for deriving insights from the documents they find [3]. Extracting and synthesizing information from documents is time consuming and mentally taxing, and supporting this process would help users further explore the information space. Finally, users have different knowledge and skill-levels, so exploratory search systems should provide information and documents personalized to that user so that they best learn [4]. Toward helping users learn, exploratory search systems can help orient them and extract information according to their needs and skills.

2.2 Information foraging and sensemaking

Exploratory search consists of two, closely-related processes: information foraging and sensemaking, both of which are cognitively taxing. Information foraging is broadly the process of collecting information. There are two main challenges in information foraging: (1) information scent and (2) information exploitation [5]. As one explores, they opportunistically "forage" for information, jumping from resource to resource and extracting information from them. In information foraging theory, these resources are known as "patches" and each patch has a different information value to that user [5]. Users make decisions on which patch to visit based on its "information scent", which is their "(imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues, such as bibliographic citations, WWW links, or icons representing the sources" [6]. After choosing a resource based on its scent, exploratory searchers then "exploit" it, e.g. carefully read or extract information from it [5]. While choosing a patch is relatively quick, exploiting it, or "actually work[ing] through the material" is quite costly and time consuming [7] [3]. This cost compounds when users misinterpret information scent and accidentally exploit patches with low information value, acquiring minimal information for a great deal of effort. Overall, information foraging is the collecting component of exploratory search, and it's main cognitive cost is extracting information from a resource.

While information foraging is the process of collecting information, sensemaking is the process of organizing information into a mental framework. Exploratory search is a learning process, of which a key component is sensemaking, the mental loop through which users fit data to a framework and fit a framework to data [13]. As users come across information, they begin initiating frames that can potentially organize the data, but as they explore further, this framework might be challenged by new data that does not quite fit. They must then either adjust the framework or collect more data to address this "cognitive gap" [14]. Adjusting a framework is cognitively taxing; it involves assessing all the information collected and generating a new organization. This becomes increasingly difficult as the amount of encountered data increases, because "human working memory has inherent capacity limits", limiting "the number of hypotheses, the amount of evidence… that can be simultaneously heeded" [7]. In summary, sensemaking is the iterative process of organizing information into a mental framework and is limited by one's ability to recall and relate information.

2.3 The spreading activation of memory

Memory plays a crucial role in exploratory search. In information foraging, users recall and recognize information in titles and text snippets to evaluate information scent and "orient" themselves [12] [15]. In sensemaking, users rely on their memory to recall and relate the information they have seen to create mental frameworks of their new knowledge [13]. In the following section,

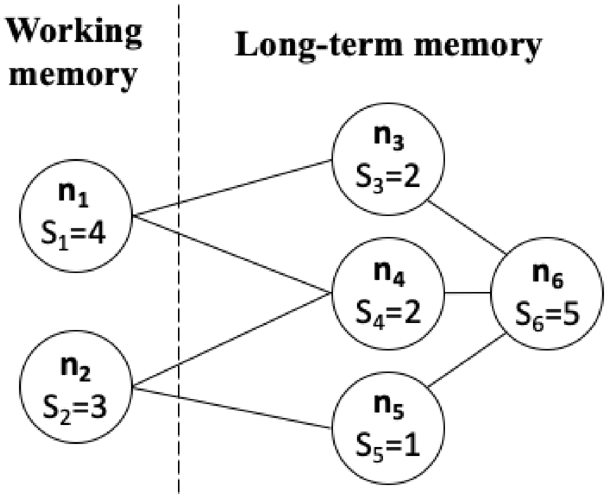


Figure 2.1: **Memory is composed of two parts: (1) working and (2) long-term.** When our minds are stimulated, nodes in working memory are "activated" and spread this activation to nodes in long-term memory, recalling semantically related information.

I discuss a prominent theory of memory used by researchers to study and predict how users forage for information.

Anderson's theory of memory states that human semantic memory can be represented as a network where each node, or "cognitive unit", contains a sentence-worth of information, and edges between nodes represent associations [9] [15]. Many psychological studies have been conducted that provide evidence for this theory [16] [17] [18]. There are two types of memory present in this network: working and long-term. Working memory consists of the information currently being processed in the system and is composed of information pertaining to our current environment and what we are currently thinking about. Working memory nodes are connected to nodes in long-term memory, which contains information that is permanently stored in the system. The relationship between these two types of memory is responsible for how we recall information.

We recall information through a "spreading activation", which starts in our working memory and spreads and activates nodes in our long-term memory. While our minds are stimulated, a node in working memory can serve as a source of activation. This activation then spreads to nodes in long-term memory, diminishing in strength as it spreads through the network until it disappears.

For example, we might be reading an article which activates node n_1 in our working memory (Figure 2.1). From n_1, its activation would spread to n_3 and n_4, which would both in turn spread activation to n_6. Finally, n_6 would spread an activation to n_5. The activation strength received by each node is dependent on (1) the activation strength of the node in working memory, in this case s_1, (2) its base-level activation strength, s_i, and (3) the strengths of each of its neighbors in long-term memory. The stronger the original activation, the greater the recall.

To study how users forage for information [10], Pirolli and Card applied Anderson's theory of memory [9] to predict which clusters users would select in the *Scatter/Gather* [19] document browsing system. *Scatter/Gather* is an interface which enables users to browse a large document collection. The system starts with a set of clusters that summarize the entire collection. Users select or "gather" clusters they are interested in, which are then "scattered" or re-clustered into a new set. Critically, each cluster is labeled with a "cluster summary": a set of topics (frequently occurring keywords) and a set of document titles with their respective topics; this summary is the cluster's information scent.

Pirolli and Card predict which clusters users will select based on their spreading activation relative to the query, which they defined as: $A_i = B_i + \sum_j W_j S_{ij}$. A_i is the activation of query word i. B_i is the base-level activation of i, S_{ij} is the association strength between cluster word j and query word i, and W_j is the base level activation of cluster word j. That is, the activation strength of a query word i for a particular cluster summary depends on its base-level strength and the strength of its associations with that cluster's words. The clusters which had a higher activation for the query were predicted to be selected more often. To evaluate these predictions, they analyzed extensive user logs and found that these predicted clusters correlated highly with those that were actually selected by users, beating two strong baselines, including (1) word overlap and (2) the co-occurrence of the query words and the summary words. Therefore, this study provides initial evidence that the strength of spreading activation a scent provides is predictive of what we recognize and choose to explore, suggesting we should design exploratory search systems so that they stimulate our memory.

2.4 Exploratory search interfaces

A plethora of interactive, exploratory search systems have been developed. In the following section, I'll first go over systems which aim to expand a user's search either by query expansion or by incorporating images, then systems which use clustering to summarize the search space, and finally systems which have users explore and make sense of networks.

One of the first methods implemented to help users explore beyond their original query was query expansion. Common techniques for expanding a user's query include providing keywords that were extracted from the most relevant documents associated with the query [20] [21], words that commonly co-occur with past queries made to the system [22], or highly associated concepts from a knowledge graph [23] [24]. As well as suggesting keywords, other interactive tools like IdeaWall [25] and *Idea expander* [26] provided images related to the current brainstorm to spark new ideas. While related keywords and images can stimulate the user's memory to broaden their search, these suggestions are often very close to the user's original query, assisting them more in query disambiguation rather than helping them view the diversity of information around their initial query.

Instead of providing suggestions close to the user's original query, other interactive systems generate an overview of results by clustering the query's results. *Recipescape* generates clusters meant to capture the different methods for cooking a particular dish [27]. To help users differentiate these clusters, the system provides a multitude of statistics for users to examine. In addition, many exploratory search systems were developed to help users explore large corpora of text. Systems like *Overview* [28], *Exploratory Labeling Assistant* [29], *Scatter/Gather* [19] and *Topic-relevance map* [30], construct bag-of-words vector-representations of documents and cluster them. While these methods are able to organize the space of documents, these systems do not do much to engage a user's memory. Often the clusters are not interpretable, either because the documents within them are not so related or the keywords are broad and provide little information scent. At the same time, these systems do little to reduce the cognitive load of each item in the cluster. Users are often still left to read entire documents, which are not accurately summarized to stimulate concepts in their memory.

Finally, a number of exploratory search systems support users in exploring and making sense of large networks. *Apolo* helps users construct their own organization of a citation network [31]. Users start with a paper of interest and begin manually clustering related works, while the system suggests papers that fit the current organization. Instead of helping users construct an organization, *VIGOR* automatically constructs an organization by creating an embedding for each paper using its network features, projecting these embeddings in a 2d space, then clustering them [32]. Both of these systems rely on the user to either create or make sense of clusters without much guidance. *Apolo* assumes the user already has knowledge of the papers they're organizing and provides no assistance in summarizing them or reducing their cognitive load. Meanwhile, *VIGOR* creates clusters that must be deciphered by users from their feature distributions. In both cases, to better stimulate user's memories, these systems should better summarize their clusters into coherent and interpretable concepts as well as reduce the amount of information per cognitive unit.

Additionally for each following chapter, we include related work that pertains to its particular domain.

Chapter 3: SymbolFinder: Brainstorming Diverse Symbols Using Local Semantic Networks

3.1 Introduction

Visual symbols play a vital role in daily communication. They are used in public signs, user interfaces, logos, and advertisements to convey important information (Figure 3.2). Symbols quickly and effectively convey abstract ideas using representative concrete objects. For example, “lost and found” is represented by an umbrella and a glove and “search” is represented by a magnifying glass. These concrete objects are also the building blocks for more creative symbolic illustrations, such as the “global warming” PSA in Figure 3.2, which combines ice cream, a symbol of *melt*, with a symbol of *Earth*. Through these concrete objects, we can enable communication that is quickly understood and universal.

While there has been a great deal of work in the graphic arts and in icon design on how to create icons from an image of an object [33] [34], the problem of how to find these symbolic objects for an abstract concept has been relatively overlooked. Visual language is constantly evolving. New symbols are constantly being created to represent new experiences, organizations, and interactions on interfaces [35]. Novices with little to no experience in graphic design are also creating symbols for logos, websites, slide decks, mobile apps and games. Novices have difficulty not only designing icons from concrete objects, but also finding concrete objects to represent the concepts they want to symbolize in the first place.

Finding symbols is particularly challenging for novice designers when (1) the concepts they would like to represent are very abstract and (2) they want to combine them to create more complicated meanings. One such visual design challenge that inspired SymbolFinder and embodies both of these problems is visual metaphors: illustrations that combine symbols to convey a complex

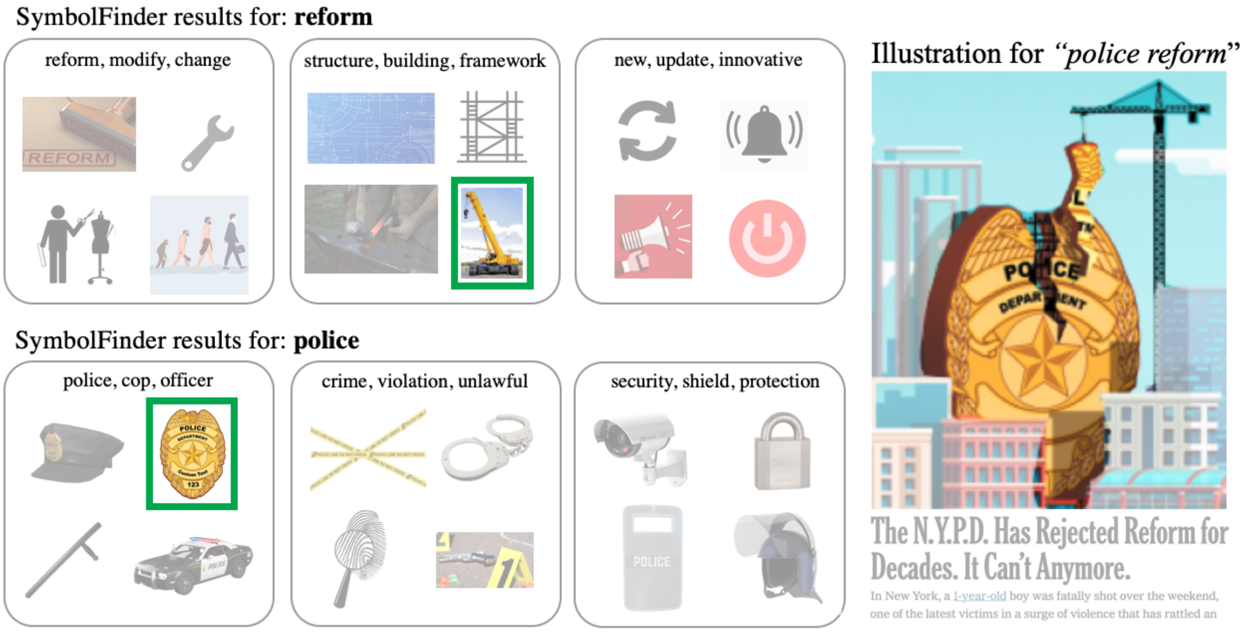

SymbolFinder results for: **reform**
reform, modify, change
REFORM
structure, building, framework
new, update, innovative
Illustration for *"police reform"*
SymbolFinder results for: **police**
police, cop, officer
crime, violation, unlawful
security, shield, protection
POLICE
The N.Y.P.D. Has Rejected Reform for Decades. It Can't Anymore.
In New York, a 1-year-old boy was fatally shot over the weekend, one of the latest victims in a surge of violence that has rattled an

Figure 3.1: SymbolFinder takes as input abstract concepts like *reform* and *police* and helps users brainstorm many diverse objects that symbolically represent those concepts. With this diverse set of symbols, novice designers find it easier to make compelling symbolic illustrations.

meaning, like the "police reform" illustration in Figure 3.1. The two symbols in a visual metaphor must be combined in such a way that their shapes blend naturally and the combined design accurately reflects the emotional tone of the message [36] [37]. To accommodate such constraints and create many design alternatives, it is essential to find a diverse set of symbols for the abstract concepts being depicted. However, converting these abstract concepts into a diverse set of visual symbols is hard for novice designers, preventing them from effectively combining them to convey a message.

In order to understand the challenges and workflow of novice designers, we conducted a formative study, where novice participants used Google Images to find symbols for abstract concepts. We observed that novices relied almost exclusively on recalling their own associations about the concept to search for related images. They often had difficulty brainstorming many different related words, and ended up fixating on a narrow set of associations, which represented a limited aspect of the concept being symbolized. Novices needed help to explore diverse ideas, which is crucial to finding an effective and creative solution [38] [39]. Finally, novices struggled to convert

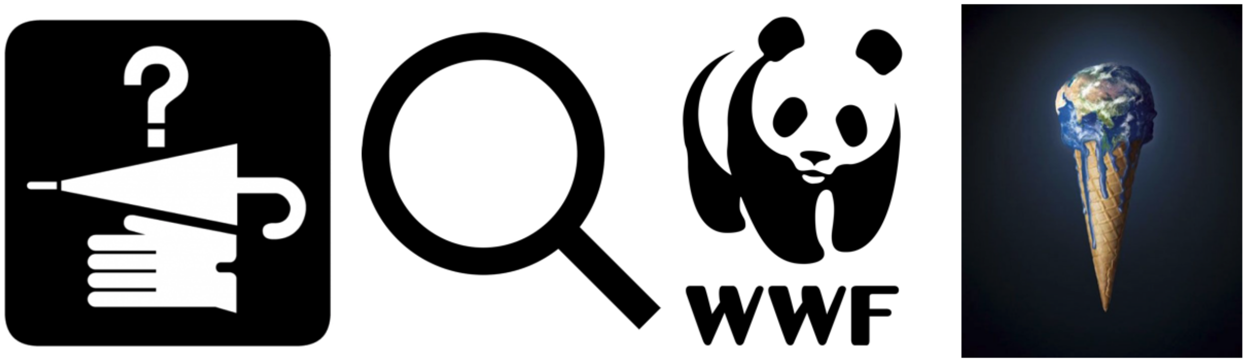


Figure 3.2: Four visual symbols, from four domains: (1) transportation hubs, (2) human-computer interfaces, (3) logos, and (4) public service announcements.

abstract associations into concrete objects and actions that could visually represent the concept.

Inspired by these observations, we created SymbolFinder to help novices find compelling visual symbols for abstract concepts. SymbolFinder helps users brainstorm associations by providing related words from an expansive word association data set. By clustering the related words into groups, each of which represents a related but distinct aspect of the concept, SymbolFinder encourages users to explore a broad range of related contexts, rather than fixating on a narrow set of associations. To create these clusters, SymbolFinder constructs a semantic network of word associations and detects highly connected communities of words. Finally, SymbolFinder helps users find imageable objects and actions by organizing words related to each cluster by word-concreteness.

This work presents the following contributions:

- SymbolFinder: an interactive interface for finding concrete images to represent abstract concepts.
- A technique for applying local semantic networks to word association data to help users perform a broad and deep brainstorm.
- An evaluation showing that users found on average 49% more unique symbols using SymbolFinder than they did using Google Images. Additionally, SymbolFinder was perceived to require significantly less effort and mental demand.
- A case study of novice designers using SymbolFinder to find the assets they need to create

more than 10 different visual metaphor prototypes for each of 3 news articles.

3.2 Related Work

3.2.1 Visual Symbols

Symbols are fundamental in visual communication and are used in a variety of contexts. They accompany headlines in news articles [35], represent actions in computer interfaces [40] [33], guide people in transportation hubs [41], represent corporations in logos [42], and form associations with products in advertisements [43]. There are many advantages in communicating ideas with symbols. Symbols often require less space to encapsulate an idea than using the word itself, saving space in interfaces, maps, and signs [44]. People can more quickly and easily recognize symbols than words because of our innate visual processes [45] [40]. Symbols are more universally understood than words across cultures, which is why they are used and designed for international transportation hubs [41] [46]. Finally, depicting ideas pictorially aids their memorability and recognition [47] [48]. For these reasons, we built SymbolFinder, to help convey more abstract ideas visually.

3.2.2 Query Expansion and Exploratory Image Search

The queries that users enter when searching for images are often ambiguous and can refer to many different real-world entities. Many researchers have recognized this problem and created tools to help users either clarify their search or find what they're looking for by providing a diverse set of image results. Textual query suggestions are a common technique for helping a user clarify their search. Keywords can be extracted from the most relevant documents associated with the query [49] [21] or taken from commonly occurring pairs of queries from search logs [22] [50]. Zha et. al. improved upon this technique by showing clusters of visually similar images for each keyword to help users preview and compare the images for each keyword [51]. *IGroup* employs a similar technique, by extracting common phrases (n-grams) from the most relevant documents associated with the query and presenting clusters of images for each of these phrases

[52]. While keyword suggestions help users disambiguate their queries, they do not let them explore the broader associated meanings and contexts of their search. By using a word associations dataset, SymbolFinder presents broader contexts that expand the user's idea of the query to help them brainstorm.

Instead of using the documents associated with the images, other tools expand queries with knowledge bases to capture diverse intentions. *PARAgrab* takes synonyms, hyponyms, and hypernyms from WordNet [53] and presents these as related searches to users [24]. Hoque et. al. use both the incoming and outgoing of links of the query's Wikipedia page to provide a list of related queries [54]. A separate knowledge base is used to cluster these associations into categories like person, place, and location. *CIDER* adds to this work by spatially arranging the images from these different queries based on their visual attributes [55]. These tools serve to quickly disambiguate a user's search, like separating Denzel Washington the actor from Washington D.C., the place. However, the organization of these related concepts does not capture different meanings and greater contexts associated with the query. For example, for a query like *reform*, instead of returning a list of specific types of *reform*, SymbolFinder presents a set of diverse clusters, each encapsulating a different sense of *reform* like "fix, amend, redo" and "new, update, innovative", to help the user brainstorm. Lastly, ICONATE [56], a system for automatically generating compound icons, expands an abstract query with a manually created concept map. SymbolFinder instead enables users to conduct their own search over the possible associations related to the concept, which helps them understand the space better and find a symbol better suited for their goal.

There exist also a multitude of exploratory image search tools that help users explore diverse results by clustering images. Cai et. al. use text, link, and visual features to cluster a query's image results [57]. Leuken et. al. create a similar system, involving a dynamic weighting function for the visual features, creating clusters that better align with a human's idea of image diversity [58]. Fan et. al. create a visual summary of image results on Flickr by creating a topic network from user-generated tags. This enabled users to view an overview of the various images connected to their query and explore highly connected clusters [59]. By providing clusters of word-associations

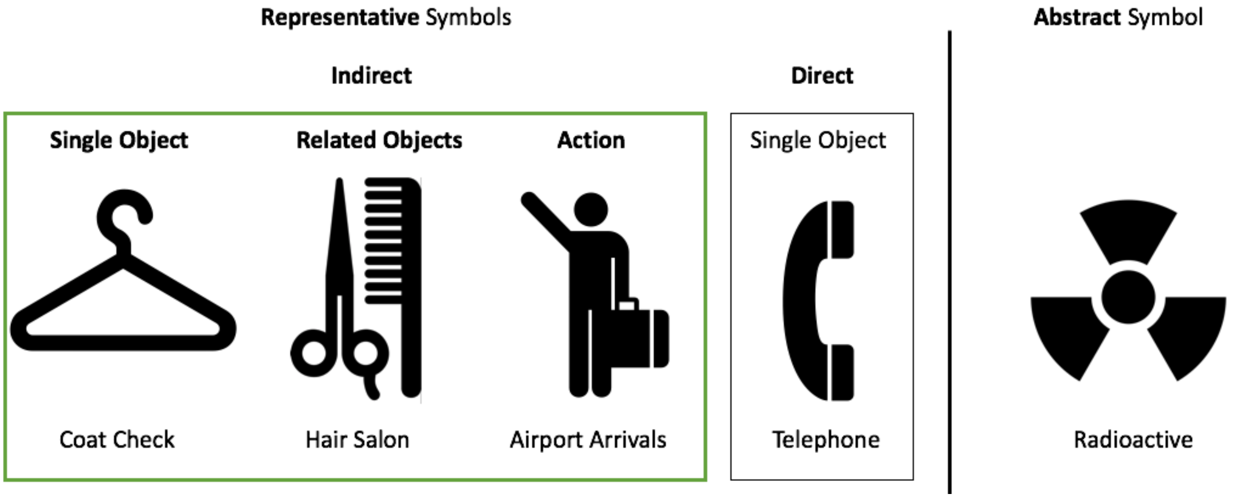


Figure 3.3: The types of symbols include: representative (indirectly and directly), as well as abstract (radioactive symbol).

associated with the query, SymbolFinder also presents an overview of diverse contexts related to the query. However, a crucial component of SymbolFinder is the ability to dive deeper into each cluster and explore concrete words. By incorporating concreteness and in-depth exploration of each cluster, SymbolFinder helps users find objects to symbolize their abstract query.

3.3 Background: What makes a good symbol?

According to the theory of symbols, there are three basic types of symbols: abstract, directly representational, and indirectly representational [60] (Figure 3.3). A symbol is abstract when an abstract pattern represents the idea, like the radioactive symbol. A symbol is directly representational when its content is an exact representation of its idea, like the telephone symbol in Figure 3.3. A symbol is indirectly representational when the image content is associated with but not an exact representation of the idea, like the coat hanger, which represents a *coat check* (Figure 3.3). SymbolFinder was built to help people find indirectly representational symbols for abstract concepts that have a variety of meanings and contexts associated with them. These types of symbols do not require a new design like the *radioactive* symbol and are difficult to find with current image databases, unlike directly representational symbols, as these databases do not enable an exploration

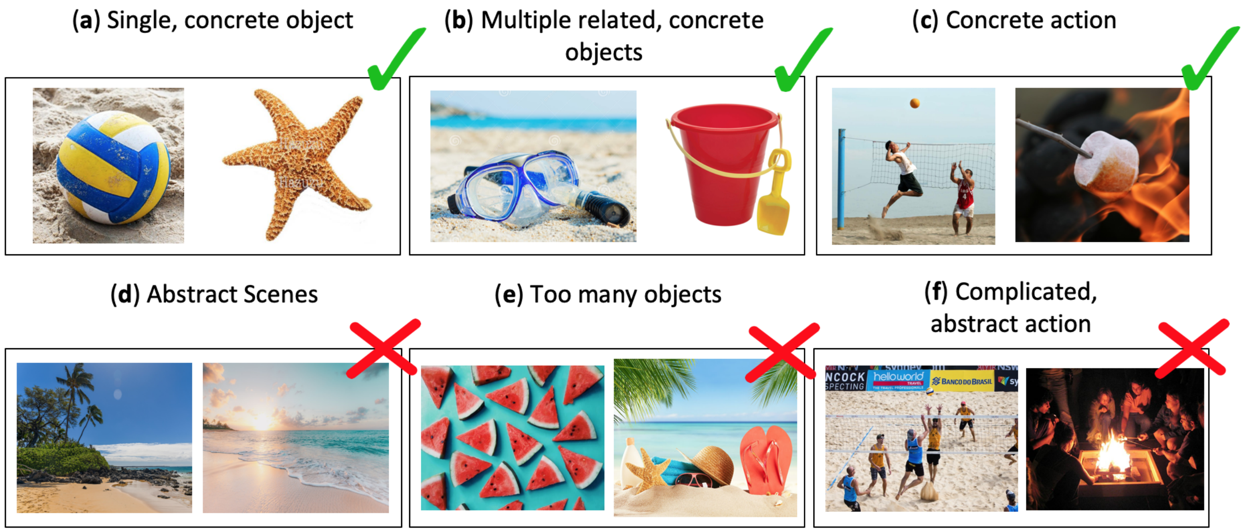

(a) Single, concrete object
(b) Multiple related, concrete objects
(c) Concrete action
✓
✓
✓
(d) Abstract Scenes
(e) Too many objects
(f) Complicated, abstract action
✗
✗
✗

Figure 3.4: The rules for what makes a good symbol, derived from theory on icons and symbols, and explained with the concept: *summer*. These rules were shown to both raters and participants.

of various ideas related to the concept.

A representational symbol can contain three things: a single object, a few related objects, or an action [33]. For example, the coat hanger is the most essential object related to a *coat check*, and thus makes a good single object symbol. Sometimes an extra object makes a symbol more specifically related to the idea it represents. For example, a scissor and a comb together represent a *hair salon* better than either one alone. The two of them together effectively represent the tools a hair stylist uses. Finally, a symbol can also contain an action, like the *airport arrivals* symbol, in which there is a man hailing a taxi. These three categories make up the vast majority of the content displayed in representational symbols.

A good representational symbol is simple and concrete in the content it depicts [34] [33]. The most essential quality of a symbol is that it is recognizable. Its content should contain no more than what is necessary to depict the idea. Visual complexity and extra entities only make them slower to interpret and recognize. From this symbol theory, we establish a set of rules to help users of our system find good symbols (Figure 3.4). A good, indirectly representational symbol can be:

- **A single concrete object**. The object must be able to represent the concept on its own

(Figure 3.4a).

- **Multiple related objects**. The objects should be related to the concept and to each other, like the combination of the scissor and comb in Figure 3.3. However, they should not be the same object, like the watermelon slices (Figure 3.4e), since one is enough to convey the idea. Symbols with more than two representative objects, like the collection of unrelated beach objects in (Figure 3.4e) are too complicated and can be separated into separate symbols.
- **A concrete action**. The action should be concrete and shown clearly, like the volleyball spike in (Figure 3.4c), as opposed to the more complex volleyball scene in Figure 3.4f.
- **No abstract scenes**. Symbols should depict a concrete object or action, instead of abstract landscapes (Figure 3.4d).

3.4 Formative Study

Novice graphic designers are designing symbolic illustrations, like logos and visual metaphors, to convey complex meanings. To create these illustrations, each concept requires a diverse set of symbols. This diversity is important to overcome constraints that occur later on in the design process [36]. To better understand the challenges novices face when searching for symbols and how to help them, we conducted a formative study in which we observed participants search for symbols with a popular image database, Google Images. Google Images is the primary tool used by novice and professional icon designers alike to look up visualizations of concepts [56]. Its interface also has many powerful features for exploring related queries including: query suggestions in the search bar, related queries with representative images above the image results, and filters for color and style. Finally, by appending "symbol" or "icon" to a Google search query, users can easily view a wide array of iconography for a particular concept. Therefore, we study how novices use Google Images when searching for symbols because of its ubiquity and its powerful search features.

3.4.1 Methodology

We recruited 5 participants (3 male, 2 female, average age 24.8) through an email mailing list for recent graduates of a local university. Every participant identified as a novice in graphic design and had used Google Images many times before. Participants were explained that the study was about understanding how novice designers brainstorm different visual representations of abstract concepts. In the task introduction, each participant was shown a slide deck, which introduced them to the concept of good, unique symbols. The slides include a step-by-step introduction to the symbol rules in Figure 3.4 and explained the importance of finding unique symbols that display different concrete objects and actions, not just the same objects in different colors. To ensure that they understood the rules, participants were asked to complete a quiz where they selected good and unique symbols of *summer* from a set of images. Incorrect answers were discussed with the experimenter.

Once participants felt they understood the task, they moved to the experiment phase of the study. Participants were given 10 minutes to find as many good, unique symbols for three abstract concepts. In a brainstorm, more ideas are generally better, even if some ideas are not perfect, as they can inspire other, better ideas and can be iterated over later. To encourage a “more is better” brainstorming mindset, we asked them to find at least 20 symbols, which seemed both challenging but doable. The three abstract concepts were *old*, *exciting*, and *innovation*. These concepts were randomly selected from a visual messaging dataset, which contains the most common concepts symbolized in online messages [43]. Participants were asked to think aloud to convey their thought process. After each concept, they were asked to explain the benefits and drawbacks of Google Images, what search terms helped their brainstorming, and their general strategy. The study took at most 1 hour and participants were paid $20 for their time.

3.4.2 Observations

One author annotated the collected symbols for goodness, according to the symbol rules, and duplicates. Two symbols were considered duplicates if they conveyed the same object or action, regardless of color or image style. On average, participants found 14.6 (standard deviation=2.8) unique symbols for *old*, 10.8 (1.7) for *innovation*, and 11.8 (3.9) for *exciting*. Every participant searched for symbols during the entire allotted 10 minutes.

All five participants were frustrated by the lack of conceptual diversity in the images presented when searching the concept as is on Google Images. P1, P2 and P4 all mentioned that the results for *old* predominantly contained images of old people. Similarly, upon seeing the image results for *excited*, P1 states, “These are all images of the word ‘excited’. Or just people looking excited.” While there was generally a couple representations of the concept in the first set of images produced by Google, users found that they needed to brainstorm on their own to find different symbols.

The most common strategy to find different images was to search terms related to the concept and scan the image results for new visualizations. For example, P1 searched *ancient*, which he recalled on his own, and met many images of the Parthenon, the Colosseum, and pyramids. This turned out to be a fruitful context, from which he was able to collect an additional three symbols for *old*. Similarly, when seeing only images of excited people in the results for *exciting*, P2 subsequently searched *fun* and *adventure*. In doing so, he found other contexts related to *exciting* like extreme sports. Users had to recall these associations on their own. Therefore, our first design goal for SymbolFinder was to **help users brainstorm related words**, in order to enable recognition over recall.

Users however also struggled to find related words that presented different images and concrete contexts related to the concept. For example, when searching for symbols of *exciting*, P2 searched for images of *adventure* and *explore* and was met with similar images of hiking and camping. While he was able to collect a number of symbols from these searches, it was difficult for him to think of another related word that encapsulated a different flavor of *exciting*. Eventually, he searched the word *suspenseful* and found images of horror movies and theatre which inspired more symbols. From this issue we formed our second design goal: when helping users brainstorm associations, we should ensure that we present diverse ideas in order to help them collect **diverse symbols**.

Once users found a fruitful context, their strategy shifted to searching concrete objects and actions that they would select as their symbols. For example, while searching for symbols of *innovation*, P2 started searching for advanced technology like virtual reality goggles and hovercrafts. Similarly for *old*, P1 and P3 searched for objects old people use like canes and wheelchairs. While more abstract searches like *elder* and *technology* served as inspiration, these highly concrete searches contained the images that would end up being their symbols. When exploring related contexts, users should be able to explore concrete words within these contexts to find representative objects and actions. Thus, our third design goal was to help users **concretize abstract concepts**.

Design Goals. In summary, from the formative study we formed three design goals for the SymbolFinder:

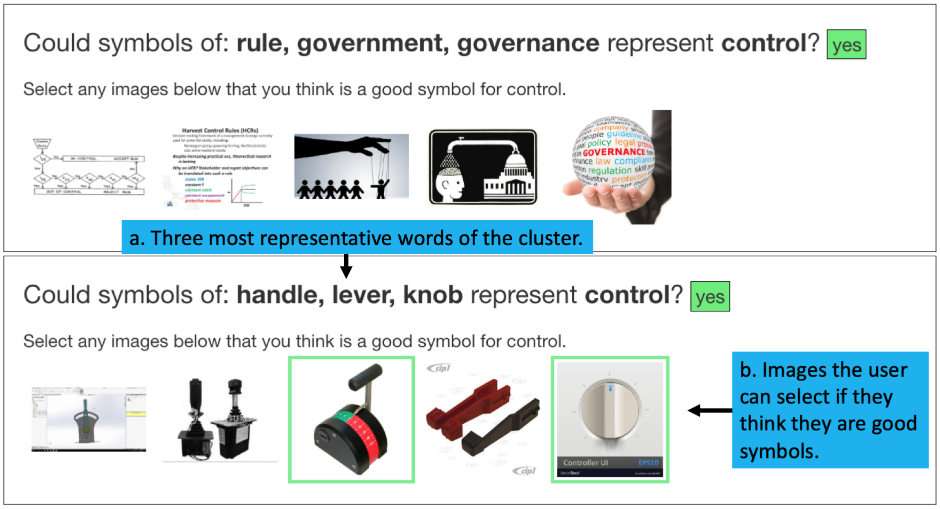
Design Goal 1: Help brainstorm related words to encourage recognition over recall. Users often recalled related terms to see new visualizations of the concept. By relying on their own memory, they miss obvious symbols and contexts associated with the concept.

Design Goal 2: Symbol diversity. When helping users brainstorm related terms we should present them a variety of diverse associations so that they can collect diverse symbols from these associations.

Design Goal 3: Concretize abstract concepts. As well as enabling users to explore diverse associations, they should also be able to explore related concrete terms for each association. This way, users can better find objects and actions to represent the concept.

3.5 SymbolFinder Interface

To address these design goals we present SymbolFinder – an interactive tool that enables novices to find multiple, diverse symbols for abstract concepts. It uses a local semantic network to organize word association data into a hierarchical structure so users can explore diverse contexts associated with a concept. SymbolFinder's interface consists of two phases. Phase 1 is a breadth-first exploration of clusters of associations related to the concept; users select clusters they would like to explore further (Figure 3.5). Phase 2 is an in depth exploration of associated images and


Could symbols of: **rule, government, governance** represent **control**?

Select any images below that you think is a good symbol for control.

Could symbols of: **handle, lever, knob** represent **control**?

Select any images below that you think is a good symbol for control.

a. Three most representative words of the cluster.

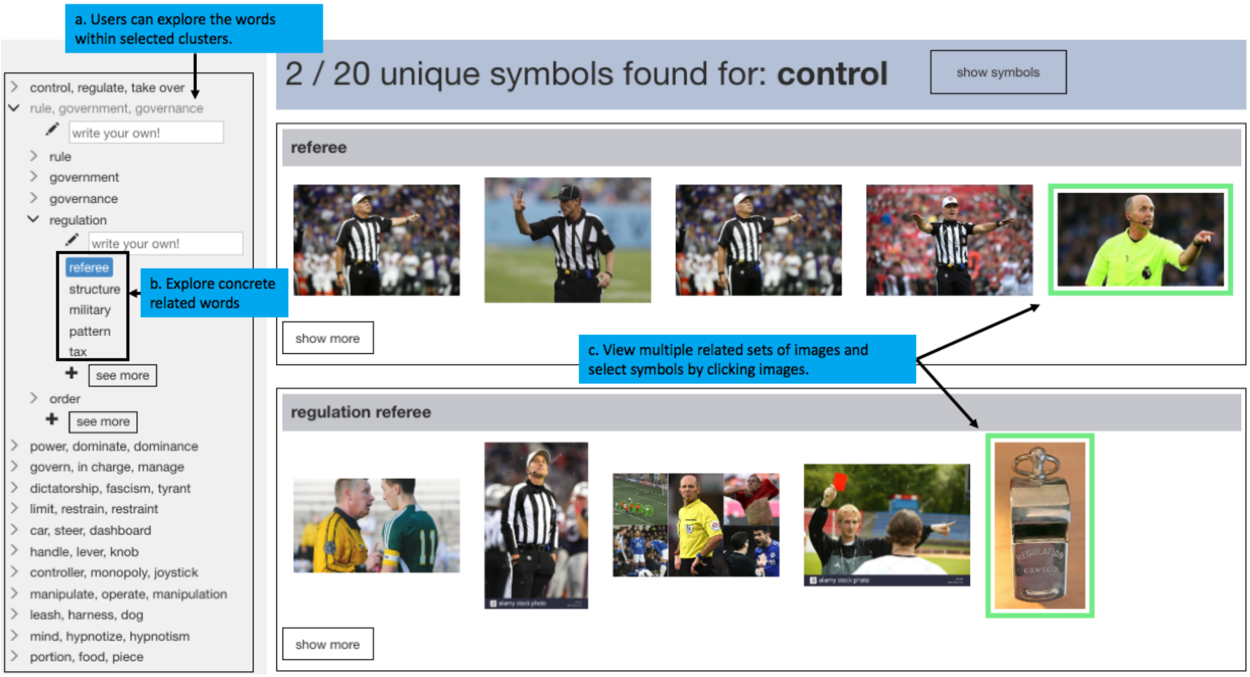
b. Images the user can select if they think they are good symbols.

Figure 3.5: **Phase 1** for the concept: *control*. Users select relevant clusters they would like to explore further in phase 2. Each cluster conveys a different association related to *control*, like the government (top) or physical tools we use to control machines. (bottom).

concrete words for each selected cluster (Figure 3.6).

3.5.1 Phase 1: Breadth-first concept exploration

Phase 1 addresses D1 (help brainstorm related words) and D2 (symbol diversity). To help users brainstorm a broad set of associations, we enable users to explore clusters of words that represent different aspects of the concept's meaning. Figure 3.5 shows a snippet of the Phase 1 interface for the abstract concept *control*. In this example, the first cluster is "rule, government, governance" and the second cluster is ”handle, lever, knob", which are two distinct aspects of *control*. Users scroll through 10 such clusters and select ones to explore further in Phase 2. For each cluster, the user is posed the following question: "Could symbols of [word 1], [word 2], [word 3] represent [concept]?". The user is instructed to press "yes" if they think it might contain symbols for the concept. There are also 5 images related to these words. Users have the option to select an image if they think it is a good symbol. These images come from three Google Image searches, one for each of the words, where each query is formulated as follows: "[concept] [word]". This is done to keep the results relevant to the concept. The queries for top cluster in figure 3.5 were: "control rule", "control government", and "control governance". By having users explore a broad set of


a. Users can explore the words within selected clusters.

b. Explore concrete related words

c. View multiple related sets of images and select symbols by clicking images.

2 / 20 unique symbols found for: control
show symbols
control, regulate, take over
rule, government, governance
write your own!
rule
government
governance
regulation
write your own!
referee
structure
military
pattern
tax
see more
order
see more
power, dominate, dominance
govern, in charge, manage
dictatorship, fascism, tyrant
limit, restrain, restraint
car, steer, dashboard
handle, lever, knob
controller, monopoly, joystick
manipulate, operate, manipulation
leash, harness, dog
mind, hypnotize, hypnotism
portion, food, piece
referee
show more
regulation referee
show more

Figure 3.6: **Phase 2** for the concept *control*. Users dive into the clusters they chose from phase 1. (A) On the left sidebar are the clusters, where users can explore related words. (B) Users can also view concrete associations for each related word in the cluster. While *regulation* is quite abstract, *referee* and *military* are more concrete and easier to visualize. (C) On the right are a few Google Image searches for the selected word: *referee*. The user selected two images, indicated by the green boxes.

clusters briefly, we quickly expose them to a diverse set of associations, preventing them from fixation on a single one.

3.5.2 Phase 2: Image selection within clusters

Phase 2 further supports D1 (help brainstorm related words) as well as D3 (concretize abstract concepts). In phase 2, users further explore the clusters they selected from phase 1 (Figure 3.6a) and select symbols (Figure 3.6c). The key part of this interface is the sidebar on the left which is where users explore the clusters (Figure 3.6a) and recursively explore concrete words related to them (Figure 3.6b). To support D3, when users select the top level cluster words, they are shown related words sorted by concreteness (Figure 3.6b). In Figure 3.6, the user selected the "rule, government, governance" cluster. They then expanded *regulation*, one of the cluster words, and selected *referee*, a related concrete term. They could also view more associations of *regulation* by pressing the "see more" button. As well as exploring the clusters, users can also type associations they think of themselves in the "write your own" text boxes and view images and associations related to their entry. In this way, the sidebar enables users to recognize good symbols as well as use their own thought processes.

The second key part of this interface is the set of Google Image search results that populate the screen when a word is selected (Figure 3.6c). Four queries are made per word, and they include the word on its own [referee], the word and its parent [regulation referee], the concept and the word [control referee], and finally the word and "icon" appended to the search [referee icon]. We include the parent and concept queries as they help keep images on topic. We include the icon query as they often provide simple images of the action or item we are looking for. When users select an image, its link and metadata are saved. When they are done searching, they can download this data. Together, the sidebar of clusters and the multiple image searches effectively help users to find concrete symbols.

3.6 Implementation

SymbolFinder is implemented in the Flask web-framework. In the back-end, SymbolFinder uses the Small World of Words (SWOW) word association dataset [61] and a dataset of concreteness ratings for English words [62]. Calculating the network clusters and eigenvector centrality of nodes is done with the python library NetworkX[1]. Image search is implemented with Google's Custom Search API[2].

3.6.1 D1: Helping users brainstorm related words

To help users brainstorm a broad set of concrete associations, SymbolFinder uses word association data to find words that are related to the concept, have diverse connotations, and are concrete. We explored two different options for creating word associations: (1) Glove word embeddings,

[1] https://networkx.org/

[2] https://developers.google.com/custom-search/v1/introduction

trained on Common Crawl [63] and (2) Small World of Words (SWOW), a crowd-sourced word association database [61]. Word embeddings are commonly used for comparing the similarity between words [64] and have been used in a number of brainstorming tools to compare the similarity of ideas [65] [66]. SWOW is a large English word association dataset. The dataset was created by having thousands of participants complete a word association task, in which each participant records the first three words they think of when seeing a cue word.

In initial testing, we found that SWOW produced words that were more relevant, diverse, and concrete than those by Glove. For example, for the abstract word *help*, the most related words that Glove produces include words like: *helping* and *need*, which are related, but are not diverse or concrete. Meanwhile, SWOW produces terms like: *donation*, *red cross*, and *tutor*. These terms are all related to *help* and they are diverse in that they represent actions, organizations, and people that *help*. Additionally, at least one term (*red cross*) is concrete. Thus, we chose SWOW to be our dataset for providing related words.

3.6.2 D2: Diversity using Local Semantic Networks

The SWOW word association dataset contains hundreds of associations for common abstract concepts, which when represented as an unorganized list, is too many for users to go through. Also, many of these associations are similar, which increases time spent and frustration parsing through them. To help users find diverse symbols from this data, these associations must be organized to identify a small number of diverse, yet highly relevant associations with the concept. To do this, we first create a local semantic network, to find all the words relevant to the concept. Then, a network clustering algorithm is run to identify sets of highly connected words that represent distinct associations of the concept.

For example, Figure 7 shows a 2-level local semantic network for the concept *control*. The first level of the network includes words strongly associated with *control* such as *government*, *rule*, *power* and *dominate*. The second level has words associated with the words in the first level. Including two levels introduces a greater variety of words while keeping words relevant. When this

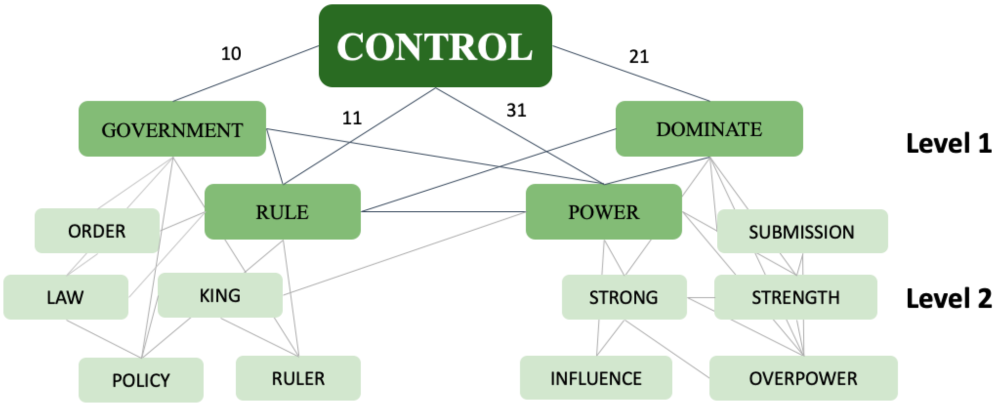


Figure 3.7: The two-level local semantic network for the word *control*. The first level of the network contains words like *government*, *rule*, *power* and *dominate*. The second level contains words like *king*, *law*, and *policy*, stemming from first layer words.

network is clustered, some first-level words such as *government* and *rule* are highly interconnected, and thus are merged into one association to present to users.

Constructing a local semantic network

To convert the word association data into a network, each word in the dataset is treated as a node and each association is an edge. The weight of an edge is the number of users who made that association in the word-association task [61]. From the concept being symbolized, a 2-level network is created (Figure 3.7). To create the first level, the first 60 strongest associations of the root concept, *control*, are added as nodes. We choose the concept's 60 strongest associations to include many associations and to keep the first-level highly related to the root concept. To create the second level, the first 5 strongest associations of each node in the first level is added. We create a second level to include a greater variety of words, but limit it to 5 per node so that few irrelevant words are added. A third level is not constructed, because associations this far from the root tend to introduce a lot of irrelevant words and make the clusters less interpretable.

Clustering the network

Our goal is to create a set of clusters, where each cluster contains highly related words that capture a distinct association of the concept. To cluster the network we considered two algorithms: the Clauset-Newman-Moore [67] and Louvain [68] network clustering algorithms. From initial experimentation we determined that the Louvain algorithm produced more interpretable clusters, as the Clauset-Newman-Moore algorithm tended to produce fewer clusters with a greater number of words, often combining clusters that the Louvain algorithm separated. The Louvain algorithm optimizes modularity, a measure which compares the edge density of the nodes in a cluster to the edge density of the same nodes in a randomly generated network. In our case, the algorithm identifies communities of highly related words by grouping words connected with high edge weights. The algorithm returns a hierarchy of clusters. We use the highest level of clusters (i.e. largest cluster size). For our dataset, the final pass of the algorithm generates about 12 to 20 clusters, which is small enough for a user to explore and large enough to contain a variety of unique ideas related to the concept.

To show users the most relevant clusters first, we sort the clusters by the average importance of their nodes. In this case, importance is defined as the eigenvector centrality of a word. A node has a higher eigenvector centrality if (1) it is connected to many other nodes and (2) its connections are also connected to many other nodes [69]. Sorting clusters in this way prioritizes the most relevant associations for a concept. For example, after sorting clusters for *control*, the most highly associated clusters are at the top of the sidebar in Phase 2, like “rule, government, governance” (Figure 3.6a). Meanwhile, more niche and perhaps less relevant clusters like “leash, harness, dog” are at the bottom of the list.

3.6.3 D3: Concretize abstract concepts

Although the clusters contain relevant and diverse associations, the words within them are not guaranteed to be concrete, like the “rule, government, governance” cluster in Figure 3.6a. Users should be able to explore concrete associations for each related word in the cluster. For example,

Exciting		Future	
Related	Concrete	Related	Concrete
Fun	Motorcycle	Past	Crystal Ball
Thrill	Race car	Tomorrow	Hovercraft
Happy	Roller coaster	Crystal Ball	Robot
Interesting	Package	Someday	Car
New	Firework	Prediction	Spacecraft

Table 3.1: Associations sorted by strength of association (related) and by concreteness (concrete) for two abstract concepts: *Exciting* and *Future*. Sorting by concreteness highlights concrete objects that can serve as symbols.

the word *regulation* is very abstract, but *referee*, *military*, and *tax* are more concrete and thus easier to visualize (Figure 3.6b). By enabling users to view concrete associations for each cluster word, SymbolFinder helps users find imageable objects, actions, and people to symbolize the concept.

To incorporate concreteness, we use a crowd-sourced dataset of concreteness ratings for 40,000 English words and phrases [62]. Crowd-workers rated words on a scale from 1 (abstract) to 5 (concrete). In phase 2, for each top-level cluster word like *regulation*, we resort its associations by incorporating both concreteness and strength of association (Figure 3.6b). To include concreteness, we normalize the word's association strength and multiply this value by the word's concreteness score. We sort by this product. Consider the examples shown in Table 3.1. Instead of abstract related terms like *fun*, the concrete lists provide imageable words like *motorcycle* that can symbolize the abstract concept: *exciting*.

3.7 Evaluation

We conduct a within-subjects study to evaluate whether users can find more unique symbols with SymbolFinder than with Google Images for abstract concepts. We also compare the perceived difficulty of finding symbols with SymbolFinder to finding symbols with Google Images. As stated in the formative study, Google Images is a good baseline because of its ubiquity and powerful search features.

3.7.1 Methodology

We recruited 10 students via e-mail mailing list at a local university (2 female, 8 male), with an average age of 26.6. The participants had no formal training in graphic design. The study took at most 2 hours and participants were paid 40 dollars for their time. First they were introduced to the problem of finding unique symbols through a 10-minute warm-up task. Next, they were introduced to both systems. They were asked to use both systems to find as many unique symbols as they could within a time limit and rate the difficulty of the task. Lastly, they answered questions in a semi-structured interview about their experience.

In the task introduction, as in the formative study, each participant was shown a slide deck, which introduced them to the concept of good, unique symbols. To ensure that they understood the rules, participants were asked to complete a quiz where they selected good and unique symbols of *summer* from a set of images. Incorrect answers were discussed with the experimenter.

To set up the experiment, six concepts were selected for participants to symbolize. They were randomly selected from the same visual messaging dataset used in the formative study [43], from three levels of concreteness. The most concrete concepts were *fast* (concreteness=0.66) and *art* (0.83). The medium concrete concepts were *dangerous* (0.46) and *rugged* (0.55). The least concrete concepts were *control* (0.38) and *simple* (0.32). Every participant found symbols for the concepts in the following order: *fast*, *dangerous*, *control*, *art*, *rugged*, *simple*. To counter-balance the study, half of the participants were asked to use SymbolFinder for the first three concepts then Google images for the second set of three concepts. The other half did the opposite. This ensured that each condition had one concept from each level of concreteness.

In the experiment phase, participants were randomly assigned to a condition: SymbolFinder-first or Google-first. In both conditions, participants were given a short introduction to the interface, then given 10 minutes to find at least 20 good, unique symbols for each of the three concepts in that condition. From the formative study, we found 20 to be a challenging but appropriate target for the 10-minute time limit.

While they searched for symbols, participants were able to refer back to the good symbol rules,

	SymbolFinder	Google	p-value
Mental Demand	5.13 (1.41)	6.8 (1.97)	<**0.001**
Physical Demand	1.97 (1.43)	3.97 (2.58)	<**0.001**
Temporal Demand	5.13 (2.55)	6.17 (2.44)	0.11
Performance	6.77 (2.03)	5.9 (1.8)	0.03
Effort	4.87 (1.67)	7.43 (1.52)	<**0.001**
Frustration	3.0 (1.93)	4.33 (1.92)	0.057

Table 3.2: Comparison of SymbolFinder and Google Images for each category in the NASA-TLX questionnaire. 6 paired-sample Wilcoxon tests, with Bonferroni correction, show that mental demand, physical demand, and effort are significantly lower with SymbolFinder than with Google. Standard deviation is in parenthesis.

which were printed on a sheet of paper. After each concept, participants were asked to complete a NASA-TLX survey, where they rated their perceived work-load on a 10-point scale. After the experiment phase, we interviewed participants about their experience. They were asked questions which elicited feedback on which system they preferred and how their preferred system helped them complete the task.

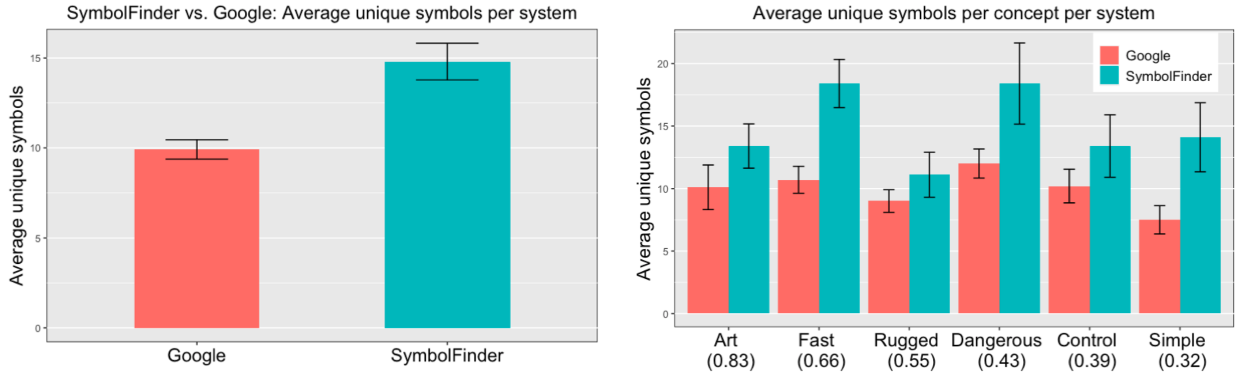
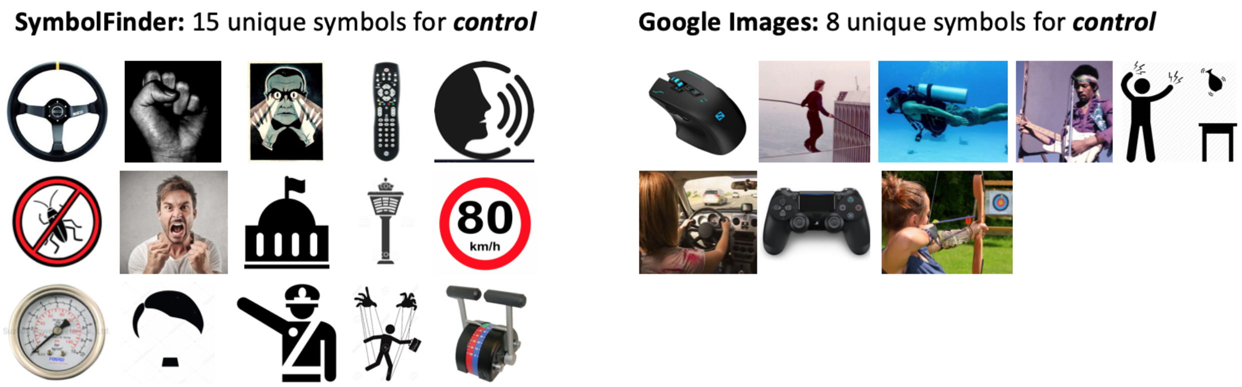


Figure 3.8: Left) Comparison of SymbolFinder and Google across all concepts. On average, users collected significantly more unique symbols with SymbolFinder than with Google. Bars are standard error. Right) Comparison of SymbolFinder and Google for each concept. On average, users collected more unique symbols per concept with SymbolFinder. The bars are standard error. Concreteness is in parenthesis.


SymbolFinder: 15 unique symbols for ***control***

Google Images: 8 unique symbols for ***control***

Figure 3.9: Users collected more unique symbols for each concept with SymbolFinder. Above are results for *control*, where the SymbolFinder user collected 15 unique symbols and the Google Image user collected 8.

3.7.2 Results

To evaluate the performance of the two systems, we needed to count the good, unique symbols found by participants. Although participants were asked to focus on finding only good, unique symbols, some of them did find duplicate symbols or symbols that did not conform to the rules. This is to be expected during a brainstorm, where participants are not supposed to edit their ideas, but focus on generating more ideas in an attempt to get better ideas. To eliminate unrelated and duplicate symbols, we recruited two graduate students in design (who did not participate in the study) to annotate the collected images of each participant. For each image collected, they determined if it was a good symbol or not based on the criteria in Section 3 and the examples in Figure 3.4. Because of the natural subjectivity of this task, we had the annotators label two practice sets of images for good and unique symbols together. Based on the calculated Cohen's Kappa coefficient, the two raters had substantial agreement in their annotations for both goodness and uniqueness. The raters had a 94% agreement on goodness (Cohen's Kappa 0.74) and a 96% agreement on uniqueness (Cohen's Kappa 0.75). To determine the number of good and unique symbols for each participant and concept, we averaged the count annotated by the two raters.

Participants found significantly more unique symbols with SymbolFinder than with Google. To assess whether SymbolFinder helps people find unique symbols compared to using Google, we conducted an analysis of variance on a generalized linear mixed model (GLMM) with Poisson function, where the number of unique symbols is the response variable. This model can account for repeated measures from participants, as well as other factors that could potentially affect the number of unique symbols collected, such as (1) concept concreteness and (2) the order in which the tools were used. Thus, the fixed effects include: *System*, *Concept concreteness*, and *Order*. The random effect is *Participants*. The results indicate a significant effect of *System* ($\chi^2(1) = 57.3, p < 0.001$). There was neither a significant effect of *Concept concreteness* nor *Order*, confirming that the counter-balancing was effective. Following this result, we conducted a paired-sample Wilcoxon test and found that SymbolFinder users collected significantly more symbols than Google users ($V = 59.0, p < 0.001$). With SymbolFinder, participants collected on average 14.8 (stdev=5.5) unique symbols per concept, while Google Image users collected 9.92 (stdev=2.9) (Figure 3.8). Figure 3.9 shows unique symbols found for *control* by one participant using Google Images and another using SymbolFinder; the SymbolFinder participant found almost twice as many unique symbols.

SymbolFinder required significantly less mental demand and effort than Google. An analysis of variance test on a GLMM (with Poisson function) was conducted for each NASA-TLX category, with the same fixed and random effects from before. *System* had a significant effect for mental demand ($\chi^2(1) = 9.4, p < 0.005$), effort ($\chi^2(1) = 15.8, p < 0.001$), frustration level ($\chi^2(1) = 7.2, p < 0.01$), and physical demand ($\chi^2(1) = 19.9, p < 0.001$). Neither *Order* nor *Concreteness* had a significant effect on any category. Following these results, we conducted paired-sample Wilcoxon tests, with Bonferroni correction and found: mental demand ($V = 13.0, p < 0.001$), effort ($V = 21.0, p < 0.001$), and physical demand ($V = 4.0, p < 0.001$) were significantly lower with SymbolFinder (Table 3.2). Users often hit dead-ends of redundant symbols with Google. After exhausting the related searches at the top of the screen, they relied on their own brainstorming to find more symbols, increasing mental demand and effort. Physical demand is less meaningful. The significant result is likely due to users having to copy and paste images from Google into a slide-deck.

SymbolFinder helped participants most when it encouraged multi-faceted interpretation of concepts. While participants found more unique symbols for every concept with SymbolFinder than with Google, the difference was greatest for three concepts in particular: *fast*, *dangerous*, and *simple* (Figure 3.8). SymbolFinder users found 72% more unique symbols for *fast*, 53% more for *dangerous*, and 96% more for *simple*. For the more concrete concept, *fast*, the SymbolFinder clusters mapped broadly to different categories of fast things, such as animals, vehicles, fast-food, natural events, scientific processes, etc. SymbolFinder users dove into these clusters and found many concrete examples listed for each one. Similarly, for the least concrete concept *simple*, the SymbolFinder clusters mapped to adjacent meanings of *simple*, like *primitive* and *pure*. These associations broadened participants' conception of *simple*, and they were able to collect concrete objects like the caveman wheel and water droplet from them. Finally, SymbolFinder was less useful for *rugged*, where clusters mapped to redundant ideas, such as "hard, rock, stone" and "mountain, craggy, rocky". By broadening participant's interpretation of concepts, SymbolFinder helped users collect more symbols, regardless of the concept's concreteness.

3.8 Case Study

To understand how SymbolFinder helps novice designers in practice, we deployed the system to a group of three students who make cover illustrations for a university science publication. Many of their illustrations were visual metaphors that combine symbols of two abstract concepts. We observed their process and interviewed them about their experience using SymbolFinder as a brainstorming tool. The team members were not co-located, but did much of their work together synchronously over teleconference software that allows screen sharing. We joined their conference call and observed them in three 90-minute sessions over a period of three months.

We selected three recent articles for them to make illustrations for. One article was selected from their school science publication and the other two were selected from The New York Times (NYT) for content diversity. For each headline, the team had to select the concept pair to represent it:

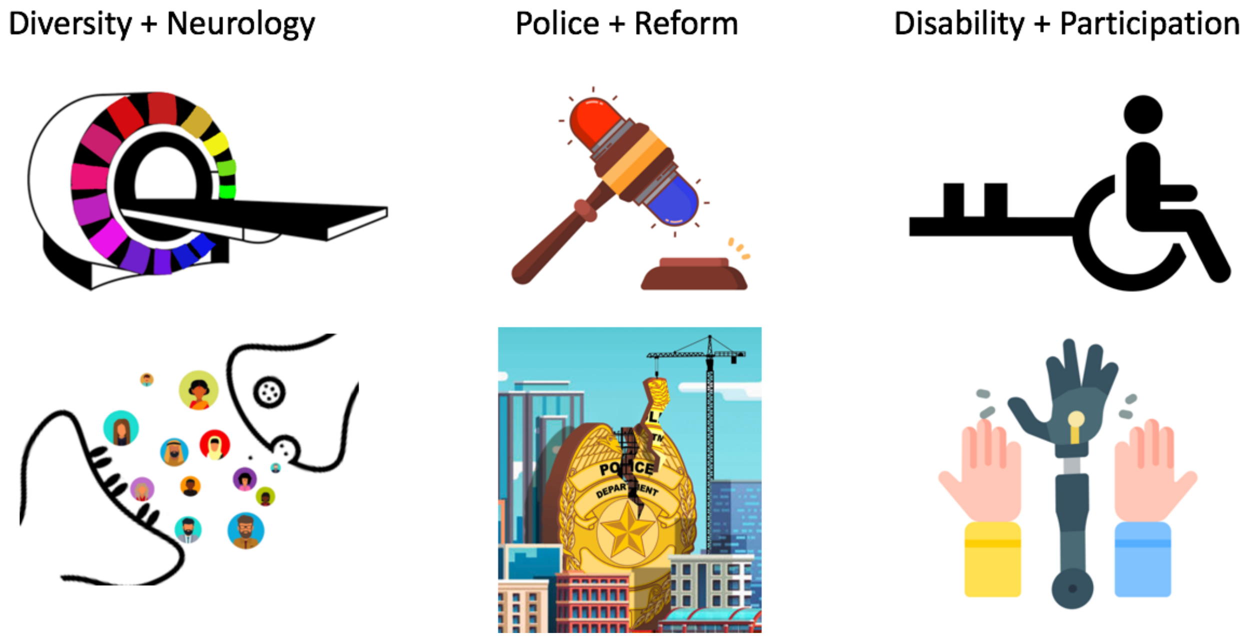


Figure 3.10: Cover illustrations consisting of two combined symbols for the three articles. The team of student designers found each symbol idea from SymbolFinder and used PowerPoint or Photoshop to create these prototypes. For *diversity* + *neurology*, the top illustration consists of an MRI machine and a color wheel, and the bottom illustration consists of a synapse and diverse people. For *police* + *reform*, the top illustration consists of police sirens and a gavel and the bottom illustration consists of a police badge and crane. For *disability* + *participation*, the top illustration consists of a person on a wheelchair and key, and the bottom illustration consists of a prosthetic arm and raised hands.

1. "Public Health Messaging in Minority Communities and COVID-19's Neurological Effects" (science publication)

   Concept pair: *Diversity* (concreteness = 0.45) + *Neurology* (0.5)

2. "The N.Y.P.D. Has Rejected Reform for Decades. It Can't Anymore." (NYT)

   Concept pair: *Police* (0.96) + *Reform* (0.4)

3. "When the World Shut Down, They Saw It Open - The pandemic has made work and social life more accessible for many. People with disabilities are wondering whether virtual accommodations will last." (NYT)

   Concept pair: *Disability* (0.69) + *Participation* (0.52)

From our observations we wanted to answer the following questions:

1. **Picking concepts**. What kind of concepts do they enter into SymbolFinder? Are they abstract or concrete? How do they choose them?

2. **Picking symbols**. What is their process of finding symbols and how does SymbolFinder help?

3. **Combining symbols**. How do they use the symbols to make the illustrations?

3.8.1 Findings

During the three observation sessions, the team made a total of 32 low-fidelity prototypes that were iterated into 6 high-fidelity illustrations. Figure 3.10 shows the two high-fidelity illustrations they made for each headline. Their general process involved first discussing what two concepts they would combine from the headlines, then using SymbolFinder together to find multiple symbols for each concept. Next, they viewed all the symbols to consider which of them might be combined in the illustration. They then created a low-fidelity prototype by copying the images into PowerPoint, and if they liked the idea, they would discuss how to improve it to higher fidelity. One person would then create a high-fidelity illustration in Photoshop. Sometimes they would use Google Images to conduct a secondary search for an image that met some stylistic criteria (color, perspective, etc.)

As expected, SymbolFinder was indeed a part of the early brainstorm part of their design process while they were still exploring multiple possibilities. A somewhat surprising observation is that although the tool was built with a single-user in mind, they used it collaboratively with one person "driving" and screen-sharing while the two others looked at the results and commented on terms and images that interested them. In the future, it might be useful for SymbolFinder to be a multi-user system. However, it's also possible that the tool may work well with one main user driving the system and other users contributing ideas. Brainstorming sessions can be run with or without a leader. Having multiple users selecting concepts independently could lead to redundant symbols that would then need to be deduplicated.

Picking concepts

When selecting two concepts to represent a headline there are many possibilities. They could pick very concrete concepts that are easy to represent visually or abstract concepts that are more difficult to visualize. To select concepts, the team read the article title and text to extract multiple potential concepts that best capture the meaning of the article. For example, while working on A2, "The N.Y.P.D. Has Rejected Reform for Decades. It Can't Anymore", the team quickly picked *police*, a very concrete concept (concreteness = 0.96), as the first concept. For the second concept, they identified a few candidates that were all relatively abstract. This included: *reform* (0.4), *law* (0.51), and *scrutiny* (0.45), which are all very abstract words. Ultimately they made illustrations by combining symbols for *police* and *reform* (Figure 3.10). For the other two articles, they picked two fairly abstract concepts to combine: for A1. they picked *diversity* (0.45) and *neurology* (0.5), for A3 they picked *disability* (0.69) and *participation* (0.52). Given the content of the articles, **at least one of the two concepts they chose to represent it was abstract, thus making SymbolFinder an apt tool for their process.**

Picking symbols

We observed that when they collected symbols, **their focus was to find multiple, diverse representations of the concept**, as opposed to finding the perfect image for one particular representation. SymbolFinder helped them find different representations in two ways: (1) by exposing them to multiple different ideas through the clusters and (2) by presenting them with concrete objects within clusters to find symbols from these different ideas.

In the first phase of SymbolFinder for *reform*, the team found multiple distinct contexts associated with *reform*, many of which led to symbols in phase 2. They selected 7 of 18 clusters shown to them in phase 1, opting for clusters that captured different aspects of *reform*, like "reform, modify, change", "fix, amend, redo" and "new, update, innovative". Meanwhile, they rejected clusters that brought in connotations that did not fit with the overarching message: "police reform", like "political party, progressive, republican", which while related to reforming politics, is not so relevant to police reform. **Phase 1 of SymbolFinder helped the team quickly eliminate these irrelevant clusters**. Of the clusters they chose, the first obvious cluster “reform, modify, change” provided the most symbols at about 40%. However, the team found many useful symbols from the less obvious clusters as well, where 60% of their symbols were spread across 6 other clusters. The second most fruitful cluster was “fix, amend, redo”, containing 22.5% of their total number of symbols. On average, the team used 6.83 (standard deviation = 2.8) clusters presented in SymbolFinder, demonstrating that **the clusters were useful for finding multiple distinct visual representations of the concepts**.

In the second phase of SymbolFinder for *reform*, **the team took advantage of the concrete sub-words to collect many different objects associated with that cluster**. For example, while exploring the “structure, building, framework” cluster, the team collected symbols of a *crane*, *scaffolding*, and *blueprint*, which appeared in the concrete sub-words of *building*. Similarly, from the “fix, amend, redo” cluster, they collected images from many concrete sub-words like *toolbox*, *saw*, and *screwdriver*. Concreteness helped the team quickly convert diverse clusters of concepts into symbols.

While picking symbols, a second constraint on the symbol space became apparent: the connotation and tone of the symbol. In phase 2 for *reform*, the team found symbols like the “update bell icon” and the “cycle refresh button” from the “new, update, innovative” cluster. Although both of these symbols did not have the tone or gravity they wanted for a illustration conveying “police reform”, they collected them anyway, thinking they could be useful for future illustrations with *reform*. In the end, they were most excited by the *scaffolding* symbol, since its tougher tone and “New Yorkness” fit the article tone well. **Thus, while the team is predominantly seeking a variety of representations for each concept, they do keep in mind the tone of the article when looking for appropriate symbols**.

Combining symbols

To combine two symbols, the team employs a **matching strategy**. After finding symbols for both *police* and *reform*, the team placed the images side-by-side to ideate combinations between them. Commenting on their overall process, P2 explained "we start by choosing a symbol we like from one concept. Then we match that symbol with one from the other concept, usually based on shape or function." They ultimately made 10 initial prototypes, 2 of which were iterated over, shown in Figure 3.10. Across these 10 prototypes, they used 8 unique symbols of *police*, which came from 4 different clusters. They combined these *police* symbols with 8 unique symbols of *reform*, which came from 5 different clusters. **By having multiple diverse symbols, the team is (1) able to successfully find a match between two concepts with a higher probability and (2) create a diverse set of prototypes to show their client, using 6-8 unique symbols from each concept.**

3.9 Discussion

In the following section we discuss limitations, future work to improve the system, and generalizable insights for future brainstorming tools.

3.9.1 Emerging vocabulary

Currently, SymbolFinder is constrained to the associations present in the SWOW dataset. There are two limitations to this: (1) symbols are defined culturally and the associations in SWOW are limited to the backgrounds of the users who built it (2) new or esoteric concepts will not appear in the dataset even though it is quite large. For example, in the past, the team worked on an article where one of the concepts was COVID-19, which did not exist in SWOW. In order to find symbols, the team brainstormed on their own, used Google Image Search, and also tried inputting related terms like *virus* into SymbolFinder. To include a new concept, we could extract related keywords from web search results or from a frequently updated knowledge base like Wikipedia. These keywords could then be linked to current entries in the SWOW dataset. We could also support symbol finding for different cultures by including international word association datasets and by helping users extract associations from international corpora.

3.9.2 Finding the perfect image

While SymbolFinder is effective for finding many diverse representations of an abstract concept, it is less useful for finding a specific image, once a particular representation is chosen. In the case study, after the team came up with an idea for an illustration using symbols they found with SymbolFinder, they would sometimes perform a secondary search with Google Images to find a particular version of the symbol. For example, while making the police badge and scaffolding illustration (Figure 3.10), the designer did not use the images of scaffolding they found with SymbolFinder. After imagining the symbol illustration, she had a specific idea for how she wanted the scaffolding to look. She wanted a “consistent background color so it would be easy to remove it”. As well as a removable background, she wanted a 2d image that was neither “super busy”, containing ”overlapping scaffolding”, nor too simple and “unnatural” looking. She ended up scanning many images to find the image she used. As well as finding the perfect image that contains the right visual detail, the team mentioned other constraints they consider in the secondary search, including finding images that are free to use and finding symbols of a particular shape (to increase more blend combinations). Fundamentally, SymbolFinder is a brainstorming tool, but in the future, we can incorporate tools to help users find particular versions of an image that fits their purpose.

3.9.3 Applying SymbolFinder to other visual media and databases

Though built on top of Google Images, we can add many other image databases to SymbolFinder, like the Noun Project [3], Flickr [4], or Shutterstock [5]. These datasets can often provide

[3] https://thenounproject.com/

[4] https://www.flickr.com/

[5] https://www.shutterstock.com/

different types of images, like black and white iconography, to expand the diversity of images users see for a given concept. And beyond images, SymbolFinder can also be expanded to search GIFS and videos. Regardless of the image database or datatype, SymbolFinder's clustered local semantic network will help users explore diverse representations of an abstract concept.

3.9.4 Generalizable insights for future brainstorming tools

For future brainstorming tools, we believe that word-association networks (like SWOW) can be powerful tools to provide people related words that are relevant, concrete, and diverse. Traditionally, brainstorming tools have used word embeddings like GloVe [63] and word2vec [64] to suggest related ideas. However, current word embedding associations do not have all these desirable properties. This is likely because word embeddings are based on the distributional hypothesis of words: two words are similar if they often appear close together in a corpus [70]. However, the closest words related to an abstract concept can often include other abstract words and antonyms [71]. For example, calculating *control*'s closest words with word2vec yields antonyms like *uncontrollable*, different forms of the same word like *controlling*, and similarly abstract terms like *regulate*. These associations are not concrete and do not capture different associations of *control* and thus are not useful for brainstorming. Meanwhile, the word-associations in SWOW are created by people; they are more closely aligned to people's mental perceptions of words and have been shown to capture word relatedness better than word2vec [72]. For *control*, SWOW includes diverse associations like *government* and *dominate*, as well as concrete associations like *traffic light* and *leash*. For future brainstorming tools, word-association networks can be used to better generate related ideas and organize them, as opposed to using word embeddings trained on large corpora.

A fundamental hurtle of any brainstorming and idea-generation system is preventing fixation. Users can be tempted to dive into a sub-problem and ignore the overall solution space. And while brainstorming, it is critical to first rapidly go through many, different ideas prior to iterating and focusing on a subset. This was an issue in earlier iterations of SymbolFinder; in pilot studies users would spend most of their time parsing through the first few clusters of associations and ignore others that might have been more useful. To prevent fixation it was critical to split the workflow into a breadth and depth phase. Once users were made aware of the solution space in the breadth phase, they spread their attention more evenly across clusters in the depth phase and were able to find more solutions.

Concretizing abstract ideas is a cognitive challenge and is useful for many brainstorming and design tasks. We can use concreteness to convert a vague problem like, “How do we help people eat healthier” into a more actionable and specific question like “How can we make vegetables a more convenient snack food?”. The abstract idea of “eating healthy” has many concrete associations, such as consuming more vegetables and cooking more instead of eating take-out. We can apply concreteness to this next set of associations to further refine the question into multiple concrete options like: “How do we make vegetables more snack-like?” and “How can we make cooking as convenient and tasty as take-out?”. With a more concrete framing, it is now easier to brainstorm real solutions. Concreteness can be incorporated in future brainstorming tools to help users first refine and better specify the question they are brainstorming.

3.10 Conclusion

This chapter covered SymbolFinder, our interactive tool that enables users to find diverse symbols for abstract concepts. In our user study we show that users can find significantly more unique symbols for abstract concepts with significantly less effort with SymbolFinder than with Google Images. We also conduct a case study showing how SymbolFinder is useful for creating cover illustrations of news articles. In the future, SymbolFinder can be applied to other media types, like GIFs, and other image databases. Also SymbolFinder could include tools to help users find a perfect image after a representation is chosen and expand its word association dataset automatically with new concepts.

Chapter 4: TastePaths: Enabling Deeper Exploration and Understanding of Personal Preferences in Recommender Systems

4.1 Introduction

Recommender systems play an essential role in determining the information we consume in our daily lives; they provide suggestions on movies to watch, articles to read, music to listen to, and more [73, 74, 75]. The goal of these systems is to help users quickly find content they like in a vast library of information. As such, considerable work has been done to improve the algorithms that predict what unseen content the user is likely to consume [76, 77]. While improving these algorithms generally increases user satisfaction [78], there is a danger that users might get stuck in what is called the "filter bubble", an overly personalized area in the recommendation space that isolates users from other content [79]. These bubbles could in turn reduce users' creativity and individuality by leading many to similar content that is "easy" to consume and fulfills short-term goals, instead of content that helps them further explore and understand their interests [80]. Recommender systems are a vital aspect of our current and future digital world, and so they can be further leveraged to help users grow and develop their personal preferences.

Toward improving recommender systems from a user-centric point of view, a few avenues have been explored. Researchers have developed evaluation frameworks that not only consider the algorithm but also the user experience, like user-choice satisfaction and presentation of recommendations [81, 82]. Besides reconsidering their evaluation, researchers have also pushed the scope of their purpose, calling for recommender systems for "self-actualization" [83, 84]. They argue that instead of focusing on consumption, recommender systems should give agency back to the user: they should enable users to develop their individuality and preferences by helping them understand the item space, explore further, and learn about their interests.

Toward supporting self-actualization in recommender systems, there are still many questions to answer. Existing informational retrieval systems tend to have linear interfaces, presenting a ranked list of items based on some relevance score that is often not clear to the user [85]. While this approach works well for surfacing relevant content for a specific query, it is unclear if it is as effective toward exploring an unfamiliar topic. The role of personalization in exploring new areas in the item-space remains unclear, even in linear interfaces. Recent work has shown that a fully personalized linear interface is not very helpful when introducing users to new content; there are many factors at play, including the user's background and goals [86]. The role of personalization is even less clear when users are given the freedom to explore openly, without the restrictions of a linear interface. Numerous interactive systems have been created for exploring a recommendation space [87, 88, 89]. While they make innovations in interactivity, they do not address how users explore and would like to explore within one of their interests. Finally, to help users learn about themselves and their habits, past work has focused on assisting them to understand what part of the item-space they consume content from at a high level [90, 91, 92]. However, the role of learning in exploration can be further examined.

To shed light on these areas, we study how to support users in exploring and understanding a music genre they commonly listen to. Like other information retrieval systems, the linear nature of these systems in the context of music limits them when they are being used as tools for exploration. We also focus on music because it is a common use case of recommender systems, with hundreds of millions of users [93]. Toward self-actualization, music is highly linked with self-identity and supporting users in understanding their interests will only strengthen that [94]. At the same time, consuming music requires much less time than watching a movie or reading an article, making it a good medium to observe how people prefer to explore their interests in real-time. Finally, music genres are a natural way to convey and communicate an interest. Genres encapsulate information about the sounds, culture, and time-period a particular group of artists belongs to [95]. Because of this, people commonly use genres to communicate their musical interests to others [96]. For these reasons, we study how to help users deeply explore and understand their interests through music genres.

Accordingly, we pose the following research questions:

- **RQ1: Personalization.** What is the role of personalization in helping users explore and understand the music genres they listen to?
- **RQ2: Exploration.** If you remove the linear constraint of a music information retrieval system, what strategies do users employ to explore the genres that interest them? How can we better support users in their exploration processes?
- **RQ3: Learning.** How does learning about their music preferences help users? What would they like to learn?

To investigate how to best explore a music genre and what it means to understand one better, we conducted a formative study in which we observed music experts explore genres. We learned what experts look for when exploring a genre, which helped us formulate three design goals. Using those goals, we built TastePaths: an interactive genre-exploration web tool. TastePaths helps non-music experts explore and understand genres that interest them by presenting an overview of the genre landscape as a clustered graph of related artists. To help users quickly make sense of these clusters, each is labeled with its three most representative sub-genres. Finally, there are two versions of TastePaths: personalized and non-personalized. These versions provide different reference points for exploration. In the personalized version, the graph of artists is grown from three artists the user frequently listens to within that genre. In the non-personalized version, the graph is grown from the three most popular artists in the genre.

To answer our research questions, we conducted a within-subjects user study with 16 participants, where we compared the two versions of TastePaths. For RQ1, we found that participants greatly preferred the personalized version and wanted even more of their listening data reflected in the interface. Regarding RQ2, participants also had a variety of exploration strategies. They also wanted more control, desiring to prune and grow the graph to guide their exploration, and they made meaningful discoveries between or on the edge of clusters. Finally, with respect to RQ3, participants found that exploring with TastePaths improved their mental map of the underlying recommender system's organization; they felt better able to verbalize and search for the music that interested them within a genre.

Our work contributes in three ways: first, we derived three design goals for supporting interactive exploration of a music genre. These were identified from expert interviews with professional music curators who have years of experience exploring new music genres. Second, we present insights addressing our three research questions derived from our prototype-based study of how participants used TastePaths, our interactive genre-exploration tool. These insights cover why personalization was useful to participants, how they envisioned personalization guiding them, users' exploration strategies and where they found their best discoveries. Lastly, we discuss opportunities for incorporating more control and expressive feedback in music recommender systems and for utilizing this feedback to improve their underlying algorithms. We also discuss incorporating finite, goal-based consumption into these systems to encourage meaningful and active exploration.

4.2 Related Work

In the following section, we discuss the need for the intelligibility of recommendations and the understanding of personal preferences in recommendation systems and current tools that support this. Next, we situate TastePaths in recent work that supports interactive music exploration tools. Finally, we discuss TastePaths in the greater context of interactive systems which aim to support users in exploring and making sense of large datasets.

4.2.1 Understanding Personal Preferences in Recommender Systems

Recommender systems help users navigate a sea of content by showing them items close to their current preferences. While they effectively prevent information overload, there is a concern that recommender systems might be guiding users to an overly personalized space, called the "filter bubble" [80]. This concern has motivated research toward providing evidence that recommendation algorithms actually lead users to filter bubbles, and so far, the results are conflicting at best [79, 97, 98]. Regardless, users are concerned about the content they consume, and in response, there has been research to investigate how to help users understand the items that are recommended to them.

Toward helping users understand the content they consume, past work has mainly focused on making users aware of where their recommendations come from at a global level. In this vein, Tintarev et al. investigated how to help users understand their movie-genre consumption profiles and found that visualizing the distribution of genres helped users understand their broader "blind-spots" in the recommendation space [91]. Nagulendra and Vassileva extended these ideas to the social media domain and created an overview visualization that reveals what categories of content the system is and is not showing in the user's newsfeed, as well as the friends who shared that content [90]. This visualization increased users' awareness that they were viewing a small subset of the recommendation space, helping them feel more in control and more knowledgeable of the content they see. Finally, *NewsViz* also produces an overview of an entire recommendation space as a tree-map, with which users can interactively resize categories to change the distribution of their recommendations [92]. This interactive overview made the system more transparent to users and helped them feel more in control of their recommendations. While these works focus on helping users understand the recommendation space at a global level, there is still more to understand about how to support users in learning about a specific interest of theirs. Our work investigates how to help users learn more about the music genres they listen to frequently.

4.2.2 Supporting Music Discovery and Exploration

By understanding their own consumption profiles, users are better equipped to discover new content; they know where to look for less familiar content and where to go to dive deeper. Toward self-actualization, discovery is an essential component for supporting growth and development [84], and in music recommender systems, discovery has been identified as an important need for music listeners [99, 100, 101, 102]. Currently, however, popular music streaming platforms do not support discovery and exploration that well, but optimize search instead [103]. To fill these gaps, researchers have proposed many different solutions to support exploration of new content in music recommender systems.

One approach, requiring minimal user effort, is to help users find new content through discovery playlists, which introduce users to new areas in the music-recommendation space. These playlists can be generated and sequenced in many different ways. To ease users into new content, Taramigkou et al. generated a playlist that gradually takes a user from their current listening preferences to a new desired genre [104]. Instead of taking users on a gradual path with the playlist, Liang and Willemsen experimented with playlists that immediately introduce the user to a new genre [86]. They found that discovery playlists should be personalized enough so that users can have a smooth entry into the genre but also need to be representative enough so that users can understand the genre's sound. In another study, Liang and Willemsen found that personalization can help nudge listeners towards new and more distant genres [105]. However, while discovery playlists can be an effective means of introducing new content with minimal effort, such playlists and other linear lists of recommendations in general are neither transparent nor controllable [85], which reduces users' acceptance of these recommendations.

To place users at the center of the recommendation process and increase intelligibility, researchers have developed many tools which incorporate interactivity to actively modify recommendations. One such system is *TagFlip* [87], which lets users specify social tags that are associated with the next song. Compared to the mobile Spotify interface, *TagFlip* was perceived to enable more control and transparency over recommendations. Exploring the interplay of interactivity and explanations, Millecamp et al. studied how users with different personalities perceived an interactive playlist generator, where sliders mapped to acoustic attributes [106]. They found that while explanations were beneficial for most users, they were less beneficial for those who felt that the explanation did not help them generate a better playlist. Supporting more fine-grained control, *TasteWeights* is a recommender system that lets users adjust weights from multiple sources to get artist recommendations [88]. Unlike TastePaths, *TasteWeights* does not create an overview of a music genre; one cannot view the different parts of a genre and the artists within them. Meanwhile, *PivotPaths* enables exploration of faceted information related to artists in a particular genre, but unlike TastePaths, does not include the user's past consumption history to orient them in this information space. To study how users explore a genre and to understand the role of personalization in this process, we developed TastePaths, which embeds a user's past data into a genre-overview. With TastePaths, we sought to understand challenges they faced while they explored and how personalization could best guide users.

To help users easily locate and specify their preferences, researchers have used overview visualizations, which depict a larger portion of the music-space for users to explore. These overviews can be created and organized in many different ways. *Moodplay* organizes artists within a two-dimensional mood-space [89]. Users found navigating by mood to be fun and intuitive; they were generally able to find a sub-space that fit their current mood preference. These overview visualizations can also serve to situate and compare a user's preferences to a greater area in the music space, such as a genre. Liang and Willemsen created a mood-based contour-plot visualization that plots songs from a genre and the user's profile in the space [107]. Users found navigating a new genre with the contour plot more helpful than with a bar chart visualization which did not offer the same comparison. Finally, these overview visualizations can also be a more intuitive way to elicit feedback from users on their preferences. Kunkel et al. developed a 3D visualization that presents the entire item space in a map [108]. Users could delineate their preferences by either raising or lowering the elevation in certain regions of the map, and they generally found doing so natural and easy. In this work, we further study overview visualizations in the context of helping users deeply explore an established interest.

4.3 Formative Study

To understand how to best help music listeners explore a novel music space, we wanted to gain insights from how expert music curators explore genres and what information they think is necessary to learn about a genre. Interviewing experts and distilling their process for design goals to help novices is a common practice and has been used to help novices generate compound icons

[56] and explore recipes for dishes [27]. Professional music curators focus on labeling, organizing, and describing large volumes of music. Their job necessitates expertise in multiple genres and involves exploring genres daily. Because of their in-depth expertise, we observed music curators' process as they explored two genres of music they were interested in but less familiar with.

4.3.1 Procedure

We interviewed five professional music curators who all worked at the same large music streaming company and had between 2 to 10 years of experience (five male, average age 37.6). For the interview, we asked each expert to explore two genres of music for 15 minutes each. We encouraged them to use any tool or service they would normally use and to think aloud describing their process as they explored. For the first genre, they were asked to choose anything they wanted to explore. We were interested in seeing what kind of genre they picked and how granular or broad that genre would be. For the second genre, they were asked to pick from a set of the most popular genres from Rate Your Music[1], a large online music database: ambient, blues, classical music, country, electronic, experimental, folk, hip hop, jazz, metal, pop, punk, r&b, rock, and singer/songwriter. We wanted to see the experts' strategies for exploring these well-known genres. Finally, after exploring both genres, the curators were asked a series of questions, focusing on more details around their process, such as what information they were specifically looking for, why they used certain tools, and their overall strategy.

4.3.2 Findings

Overall, we found that the five experts we interviewed had a similar process for exploring a new genre. We describe their process in detail below, and building on it, we extract three design goals that we later applied to TastePaths.

When they first started exploring a genre, the experts generally looked for lists of representative artists and tried to identify ones they were already familiar with. This gave them some context

[1] https://rateyourmusic.com/genres/

around what the genre might sound like. For example, when P5 first searched for *outlaw country* on Google, he recognized Willie Nelson as one of the artists on the returned list. Related to that, P3 already knew a few of the notable artists in *drum and bass*, so his first step in exploring that genre was to search for them on Spotify. He explained that *"It's more you have an artist that interests you, and then you become interested in [the] genre once you find a collective of artists."* This implies that having a familiar artist within the genre provides a helpful starting point for exploration. This finding led us to our first design goal for TastePaths: **to anchor genre exploration in artists the user is already listening to**.

After listening to a few key artists to get a sense of the general sound, the experts focused on the genre at a higher level. They looked for information on its history and variety, such as the different sounds and subgenres that comprised it or its stylistic origins. For example, while reading about *outlaw country*, P5 examined its related genres because *"it helps contextualize this genre. I'm looking at these [related] genres, seeing if I recognize them and thinking about their musical or other types of qualities and trying to relate that back to what I just read about outlaw country."* Thus, he was using his prior knowledge to better understand how the new genre fits in the greater music landscape. In the same vein, P2 explained that knowing a genre better is being able to *"identify things about it that were different from other things."* Overall, this ability to contextualize a genre, be aware of its components, and know what it's related to and different from, is an essential part of understanding a genre. Accordingly, our second design goal for TastePaths was **to help users get an overview of the genre-space in order to give them an idea of what it contains**.

Finally, throughout their process of exploring a genre, the experts enjoyed diving deeper into certain artists. For example, P2 explored a band's related artists in the "fans also like" feature on Spotify. From there, he selected a few artists and spent time looking at their artist photos, reading a few lines from their descriptions, and then would sample a few of their most popular songs. P4 also listened to a few unfamiliar artists in more detail and noted that *"what I'm learning, is I need to redefine my definition of tango, because what I'm hearing is not what I was expecting. I was* *expecting something more danceable. So that's super interesting."* A common strategy among all the experts was to scan through an artist's most popular tracks on Spotify, which helped them evaluate if they wanted to explore this artist further. Being able to dive deeper and listen to an artist's music helped the experts direct their search towards a subsection of the genre they liked more. This inspired our third design goal for TastePaths: **to help users quickly and easily dive deeper into an artist's work**.

Design goals. In summary, we elucidated three design goals for TastePaths from the formative study:

D1: Anchor artists. To give users a meaningful starting point for exploring a genre, we should help anchor them with artists they already know and listen to.

D2: Genre-space overview. To contextualize the genre and help users understand it and its components, we should present an overview depicting the genre-space and its subgenres.

D3: Deep-dive. To allow users to easily assess what parts of the genre they like, we should have a quick and convenient way to deep-dive into an artist's work while being able to seamlessly go back to exploring.

4.4 TastePaths Interface

To address the design goals identified by the expert interviews, we created TastePaths: an interactive web tool, which enables users to better explore and understand a genre they are interested in (Figure 4.1). TastePaths visualizes a force-directed graph of related artists within a genre and assists the user in exploring and making sense of it. To address D1, TastePaths helps a user explore a genre by basing exploration from either three of their most frequently listened to artists in that genre or that genre's three most popular artists. We call these artists a user's "anchor artists"; they appear as black dots in the graph (Figure 4.1A). To address D2, TastePaths generates a graph consisting of 150 related artists stemming from the three anchor artists. To make this graph more of an overview, TastePaths clusters the artists and presents a legend, displaying each cluster's three most representative genres (4.1C). To help users navigate the graph, we relate each anchor artist to

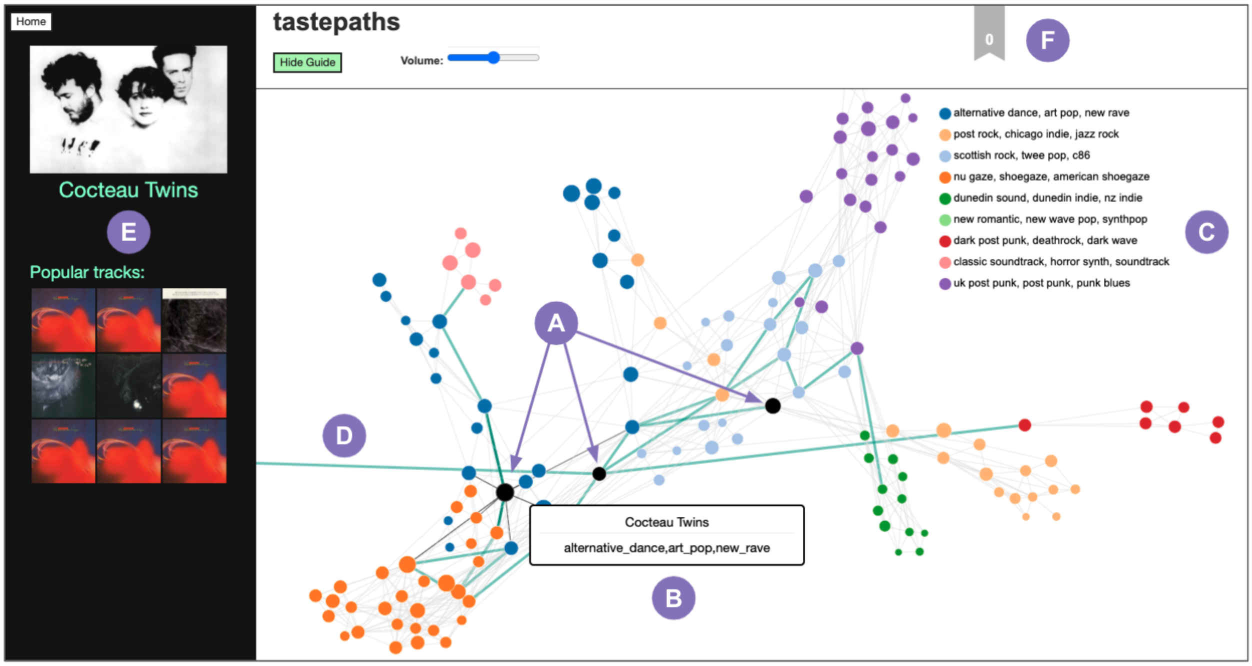

Home
Cocteau Twins
E
Popular tracks:
tastepaths
Hide Guide
Volume:
0
F
● alternative dance, art pop, new rave
● post rock, chicago indie, jazz rock
● scottish rock, twee pop, c86
● nu gaze, shoegaze, american shoegaze
● dunedin sound, dunedin indie, nz indie
● new romantic, new wave pop, synthpop
● dark post punk, deathrock, dark wave
● classic soundtrack, horror synth, soundtrack
● uk post punk, post punk, punk blues
C
A
D
Cocteau Twins
alternative_dance,art_pop,new_rave
B

Figure 4.1: **TastePaths interface for the genre *art pop*.** TastePaths displays a network of artists grown from the user's anchor artists, which are displayed as black nodes (A). When hovering over a node, the artist's name, here "Cocteau Twins", appears and a preview of their most popular track is automatically played (B). This network is clustered to capture different groups of artists within the genre. The three most representative genres for each cluster are shown in the legend (C). To help users navigate the network, there is a green path, called the "guide" (D), which connects the users' anchor artists to important artists within each cluster. By pressing a node, the sidebar displays the artist's cover art and their nine most popular tracks (E). To add a song to the playlist, the user has to click on one of these popular track images. The playlist itself is displayed when the ribbon is clicked (F). The ribbon also displays the number of songs currently in the playlist.

important artists, based on node-centrality, within each of the clusters through a highlighted green path called the "guide" (Figure 4.1D). Finally, to address D3, we enable a deeper dive into each artist by presenting an artist's image and top-9 tracks in the sidebar (Figure 4.1E) when a user clicks on their node in the graph. Users can listen to these tracks by hovering their mouse over the album covers, and they can add them to a playlist by clicking on them. Finally, users can view their current playlist by clicking on the ribbon on the top-right (Figure 4.1F).

4.5 TastePaths Implementation

An essential part of exploring a genre is understanding its components and the different collections of artists that comprise it. To support this, TastePaths presents an overview of the genre-space via a graph of related artists within the genre. This graph is constructed from either the user's personalized anchor artists or the top-3 artists in the genre. To help users distinguish the different groups of artists in the graph, we cluster the graph and assign each cluster a label with its three most representative sub-genres. However, even with these labeled clusters, the graph is large and potentially difficult to explore. Therefore, to help users venture into these different clusters from their anchor artists, we highlight a path from their anchor artists to important artists within each cluster. Below, we describe each element of TastePaths' implementation.

Overall, TastePaths is implemented in the Flask web framework. To cluster the graph, calculate the betweenness centrality of each node, and create the Steiner tree for the guide, we use the python library NetworkX[2]. To get the song previews for each artist, the artist-to-genre data, and each artist's related artists, we use the Spotify Web API[3]. Finally, to visualize the force-directed graph of artists in the genre, we use D3.js[4] [109].

4.5.1 Identifying Genres to Explore and Anchor Artists

The first element of TastePaths is identifying what genre to visualize for a given user since TastePaths was built to help users deeply explore genres they have a demonstrated interest in. To fulfill D1 and anchor their exploration within the genre, we identify genres the user listens to often. Then we anchor their exploration with either (1) artists they listen to frequently within the genre (personalized version) or (2) the three most popular artists in that genre on a music streaming service (non-personalized version). To identify these genres, we access the list of tracks the user has listened to in the past 90 days on a popular music streaming service, and we retrieve the user's top-50 highest-affinity tracks. A user has a higher affinity toward a track if they have listened to

[2] https://networkx.org/

[3] https://developer.spotify.com/documentation/web-api/

[4] https://d3js.org/

it often and intentionally interacted with it (e.g. by adding it to a playlist or playing it). We then get the associated genres for each artist of these top-50 tracks and count the number of different artists and tracks that appear per genre. Any genre with less than three artists is removed from the resulting list; fewer than three artists might rather indicate an interest in those artists instead of the genre. Finally, we sort this list of genres by the number of tracks, since a greater number indicates a stronger interest in the artists and the genre. Any of these genres are suitable for exploration. In the personalized version, we take the three artists within a genre with the greatest number of tracks in the top-50 to serve as anchors. With these anchor artists, we can construct an overview through which the user can explore a genre.

4.5.2 Constructing the Related-artist Graph for a Genre

Next, we use the anchor artists to build a related-artist graph for the genre. To do so, we leverage a music streaming service's artist knowledge graph, where nodes are artists and edges between artists indicate that they share many listeners. We first construct an initial graph, which connects the three anchor artists (either personalized or non-personalized). To do this, we find the shortest path (via the bidirectional version of Dijkstra's algorithm) from the most popular anchor artist to each of the other anchor artists; all the intermediary nodes from these shortest paths are added to the initial graph. For each node in the initial graph, we add its two most related artists and their connections to the graph, creating a second layer of nodes. Next, from this second layer we do the same process and add two nodes for each artists, growing the graph layer by layer until there are 150 nodes in the graph, to ensure a reasonably large overview. By adding only two related artists per node, we grow the graph more deeply, capturing more groups of artists.

4.5.3 Clustering and Labeling the Graph

The graphs generated for a particular genre often revealed groups of densely connected artists that have many sub-genres in common, such as the dark orange *shoegaze* cluster in the bottom-left of Figure 4.1. To identify these densely connected groups, we clustered the graph using the

Louvain graph clustering algorithm [68]. This algorithm clusters the graph hierarchically. We use the top-level clusters returned by the algorithm, which generally returned about 5-10 clusters for the graphs we generated.

To label these clusters in the overview, we select their top-3 most representative genres to include in the legend (Figure 4.1C). To do this, we access the set of genres associated with each artist in the graph, using the Spotify Web API[5]. Our first attempt to label a cluster was to pick the top-3 most common genres among all of the artists in the cluster. However, this top-3 would often include the name of the overarching genre, which would be shared across all the clusters, making them indistinguishable from each other. To find the genres that are common and unique to a particular cluster, we leverage a technique normally used in information retrieval called the term frequency-inverse document frequency (TF-IDF) [110]. This technique is used to determine how relevant a term is to a particular document in a collection of documents. Its output is a set of scores per term per document, where higher scores indicate that a term is specific to a document and lower scores indicate that the term appears often across all the documents. In our case, we treat each cluster as a document and its set of genres as its terms, and we select the top-3 genres with the highest TF-IDF scores to represent that genre. This method creates genre-labels that better distinguish clusters at the sub-genre level.

4.5.4 Guiding Users with a Highlighted Path

Finally, the resulting graph can sometimes be densely connected, and it can be hard for users to see the connections from the anchor artists to each of the clusters. To make this overview easier to navigate, we highlight a simple path connecting the user's anchor artists to important artists in each cluster. This is visualized as a green path, called the "guide", in the graph (Figure 4.1D). We want the minimal number of edges connecting the anchor artists to each cluster to minimize visual complexity. This set of edges is called a Steiner tree [111], and we calculate it using an approximation algorithm in the NetworkX Python library[6].

[5] https://developer.spotify.com/documentation/web-api/

[6] https://networkx.org/

To determine the important nodes in each cluster to include in the tree, we experimented with several node-centrality measures, including basic edge count, eigenvector centrality [69], and betweenness centrality [112]. Edge count and eigenvector centrality classified highly connected nodes at the center of clusters to be important. While accurate, the generated Steiner tree would then include many edges within each cluster to get to this central node. Meanwhile, betweenness centrality emphasized influential nodes at the edge of clusters that acted as "gateway artists" into that cluster. We decided to use this measure of centrality because the resulting Steiner tree was less visually complex.

4.6 Evaluation

To answer our research questions and learn how TastePaths helps users understand and explore their interests, we conducted a within-subjects study, comparing a personalized version of TastePaths to a non-personalized one. We chose a within-subject design because we wanted to understand users' preferences and qualitative reflections between the two conditions. To understand how users perceived the two versions of TastePaths, we analyzed data from multiple sources. This included a questionnaire that measured their perceived helpfulness of each version and user-logs containing their actions as they interacted with the systems. We also asked conducted semi-structured interviews and a thematic analysis on the resulting data. This gave us insights into how they perceived each version, explored the graphs, and what they learned about their interests.

4.6.1 Procedure

The general outline of the study was the following: (1) users were first interviewed on their music preferences and methods for finding new artists, (2) they then used the two versions of TastePaths to find new artists in two genres they commonly listen to, (3) after each version they filled out a questionnaire, rating their perceived engagement and helpfulness of the tool, (4) they were asked a series of questions on their thoughts of the tool in a semi-structured exit interview.

To set up the experiment, we selected two genres for each user to explore, one for each version of TastePaths. To do this, we followed the procedure outlined in Section 4.5.1 to get the user's top genres from their 90-day listening history. From this list of top genres, we selected their top two genres for them to explore. Sometimes these genres were similar and contained a few of the same anchor artists. In that case, we took their top genre-interest and then the next strongest genre-interest that featured no intersection with the top genre's anchor artists. To confirm that the participants were in fact interested in the genres we selected, we asked them to rate on a 7-point Likert scale how knowledgeable they are in that genre and how interested they are in exploring artists in that genre.

In the experiment phase of the study, participants were randomly assigned to a condition which determined which version they will interact with first: either personalized-anchors-first or popular-anchors-first. Condition order was counter-balanced to prevent a learning effect. After being told the two genres they would explore, participants picked the order they explored them in. When interacting with TastePaths for the first time, users were given a short explanation of the interface and its features. For each version, users were told if the anchor artists were either personalized or the most popular artists in that genre. They were then given ten minutes to find five new songs from five new artists. We wanted to see if either version of TastePaths would help users explore a genre more deeply and thus encourage them to find new artists; five artists seemed challenging enough but also doable given the time-limit. We also emphasized that this number was only to encourage them and that they should only add songs to their playlist if they were genuinely interested in that song. Participants were encouraged to talk and explain their process and actions as they explored the visualization.

After exploring each genre, participants were asked if they would like to save the playlist to their personal account associated with a music streaming service. They were also asked to fill out a questionnaire (Table 5.1) to understand their perception of the tool for the task. Finally, after experimenting with both versions of TastePaths, we asked them a series of questions that probed at their preference for each system, exploration strategies, and knowledge they gained from using the tool.

Metric	Statement (7-point Likert scale)
Engagement	Q1. It was entertaining and interesting to explore artists in the {first, second} network.
Interest	Q2. With the {first, second} network, I was able to find artists that matched my interest.
Serendipity	Q3. With the {first, second} network, I found artists that I had not considered in the first place but turned out to be a positive and surprising discovery.
Music Discovery	Q4. The {first, second} network helped me discover new artists.
Guidance	Q5. With the {first, second} network, it was easy to determine which artists I'd be interested in.
Confidence	Q6. I am confident I will like the songs in the playlist I made using the {first, second} network.
Learning	Q7. With the {first, second} network, I feel like I know more about the artists and sounds of the genre better than when I started.
Understanding	Q8. With the {first, second} network, I feel like I understand the artists and sounds of the genre better than before.

Table 4.1: Post-task questionnaire filled out by participants after they used a version of TastePaths to explore a genre. Each statement was rated on a 7-point Likert scale.

4.6.2 Questionnaire

The questionnaire we gave participants borrows ideas from a few common evaluation frameworks for recommender systems (Table 5.1). To understand how the system helped users find music, we measure user-perceived interest in the artists (Q2), music discovery (Q4), and user confidence (Q6), which are adapted from Cai et al. [113]. In the same vein, we also include a question on guidance (Q5), to see if one version made identifying interesting artists easier than the other. To understand if one version was more interesting to use than another, we also include Q1 to measure engagement [114]. Finally, to understand if one version of TastePaths helped users learn about the genre more than the other, we added Q7 and Q8.

	Personal	Non-personal	p-value
Engagement	6.56 (0.89)	6.75 (0.58)	.48
Interest	6.5 (0.89)	5.69 (1.7)	.047
Serendipity	6.25 (1.13)	5.31 (1.66)	.03
Music Discovery	6.75 (0.45)	5.94 (1.95)	.26
Guidance	5.94 (1.48)	5.69 (1.7)	.61
Confidence	5.94 (1.06)	6.25 (0.77)	.21
Learning	5.44 (1.55)	5.63 (1.54)	.46
Understanding	5.31 (1.7)	5.63 (1.75)	.27

Table 4.2: **Comparison of the Personal and Non-personal version of TastePaths for each category in the post-task questionnaire.** 8 paired-sample Wilcoxon tests, with Bonferroni correction, show no significant differences for any metric in the questionnaire. Across the metrics the biggest difference is in *Interest* and *Serendipity*. TastePaths participants using the personalized version found more artists in their interest. In parenthesis is standard deviation.

	Personal	Non-personal	p-value
Listening duration (minutes)	5.18 (1.4)	5.26 (1.5)	.82
Nodes deeply explored	7.75 (2)	7.12 (2.6)	.39
Nodes hovered	68.56 (24)	74.5 (28.2)	.25
Green-path nodes deeply explored	1.63 (1.6)	1.93 (1.9)	.58
Clusters explored	4.31 (1)	3.5 (1.2)	.036
Clusters	7.31 (2.5)	6.13 (1.4)	.12
Playlist songs	7.25 (2.6)	5.68 (2.6)	.015
Playlist artists	5.94 (1.5)	4.94 (2.4)	.079
Playlist saved	.875 (.33)	.625 (.48)	.045

Table 4.3: **Comparison of the Personal and Non-personal version of TastePaths across a number of measurements taken from the recorded user logs in the study**. We conducted nine paired-sample Wilcoxon tests, with Bonferroni correction, and found no significant differences in each of these metrics. These logs indicate that in both conditions, users were highly engaged with the system, hovering over many artists and listening to music for over 50% of their time with each system. Participants using the personalized TastePaths made longer playlists on average, with more unique artists, and were more likely to save it. In parenthesis is standard deviation. All metrics are means of counts, except listening duration.

4.6.3 Participants

We recruited 16 participants (P1-P16), 7 female, 8 male, and 1 non-binary person, from the dscout platform for remote studies[7]. Their ages ranged from 19 to 53, with the average being 31 years old, and they had diverse backgrounds (including diverse occupations, locations within the US, music interests, income levels). To be eligible for the study, they had to be over 18 years old, reside in the US and speak English, have a paid premium account on a music streaming service for at least a year, be interested in exploring new music, and listen to discovery-focused playlists at least once in the last three months on that music streaming service. The interviews were conducted remotely, and participants had to have a computer with a Google Chrome web browser. Consistent with internal guidelines, participants were reimbursed $100 for the 60-minute interview, paid via the dscout app.

4.6.4 Analysis

We performed inductive thematic analysis on the qualitative data from the semi-structured interviews [115, 116]. Through an iterative coding process, two of the authors coded the interview data and discussed any disagreements. Examples of codes included ‘sonic comparison between different clusters’ and ‘helpful aspects of the legend.’

To analyze the questionnaire data, we conducted paired-sample Wilcoxon tests with Bonferroni correction, since we compared two paired groups with ordinal data. We found no significant differences between the two versions of TastePaths for any of the metrics (Table 4.2). To analyze the user-log data, we also conducted paired-sample Wilcoxon tests for the same reasons and once again found no significant differences between the two versions for any of the metrics (Table 4.3).

4.7 Findings

Through our analysis, we identified four main themes: (1) personalization is key, (2) best discoveries are between or on the edge of genres, (3) users want more control: human-in-the-loop

[7] https://dscout.com/

growing and pruning of the graph, (4) improved recommendation explainability through mental map. The first theme helps to address RQ1 by explaining why users preferred the personalized-anchors version of TastePaths. The second and third themes help to address RQ2 by summarizing (1) what exploration strategies were most effective and (2) how users imagined themselves interacting with the graph. Finally, the fourth theme addresses RQ4 by summarizing what users learned and wished they had learned.

4.7.1 Personalization is Key

Twelve out of 16 said they preferred the version of TastePaths with personalized anchor artists. Overall, participants found the personalized version more interesting since it featured more artists they knew and liked. P5 explained that the "*personalized was more useful because it was based on [my] specific taste*." On average, participants collected more songs they liked for their playlists with the personalized version. They collected an average of 7.25 (stdev=2.6) songs per playlist with the personalized version and 5.68 (stdev=2.6) with the non-personalized one (Table 4.3). This aligns with the questionnaire results, where users rated the personalized version higher for both music discovery and interest, suggesting that the personalized version of TastePaths was more helpful to find new and interesting artists.

Participants wanted even more personalization to better guide their exploration in the graph. Beyond the three personal anchor artists, participants imagined more ways their data could be visualized in the graph to help them explore. One idea posed by P9, P5, and P6 was to have the graph indicate which artists the user has listened to before. This way, they could focus on exploring unknown artists nearby the ones they had interacted with before. Other participants imagined even more advanced ways the visualization could be more personalized. P7, for example, wanted TastePaths to prioritize clusters based on her affinity towards them: "*Maybe if there were a couple of artists it knew I liked [and I could] find more direct correlations... [I] want to see a heat map almost - this is the hot spot of what you might like*." Participants wanted TastePaths to incorporate more personal listening data to better guide them to new content they are likely to enjoy in the graph.

4.7.2 Best Discoveries are Between or on the Edge of Genres

Participants found new artists they really liked between genre clusters or at the outskirts of a cluster. Artists between two clusters captured essences of two musical styles, which led to exciting discoveries when these were styles the participant enjoyed. For example, while exploring a personalized network for *dance pop*, P9 found an artist she had never heard of before between two clusters she typically enjoys: "*I've never heard of this person...and I like this because it's a blur between EDM and dance pop. This is definitely on the edge between those genres. I have a lot music that sits in the middle*." Similarly, while exploring a personalized network of *classic rock*, P8 found that he already knew most of the artists in the graph, and so the most interesting finds were at the edges of clusters: "*Outliers are most fun to discover because I haven't heard them in the past*." The richest discoveries for users when exploring a familiar genre were those that helped them dive deeper and united multiple aspects of that genre that they liked.

To make these meaningful discoveries, **participants employed several different strategies to explore the graph visualization**. The most common strategy was to use the legend to identify an interest in the graph (8 out of 16 participants). Since the legend summarized the three sub-genres that best describe each cluster, participants used their preexisting knowledge of sub-genres to pick a cluster that looked interesting. Along with the legend, participants also used the anchor artists to direct their exploration. A common strategy was to start with the anchor artists and explore their direct connections. This was more popular in the non-personalized condition, where participants would first inspect the popularity-based anchor artists, and if they liked them, explored nodes close to them (Figure 4.2B). A few participants also used the green path as a guide (Figure 4.2C). But this was generally an uncommon strategy; on average, participants clicked on less than two artists in the guide in both conditions (Table 4.3). Participants were less inclined to follow the guide to different clusters and more willing to jump around from cluster to cluster. Finally, a couple of the participants did not care for any guidance at all and started from one cluster and systematically

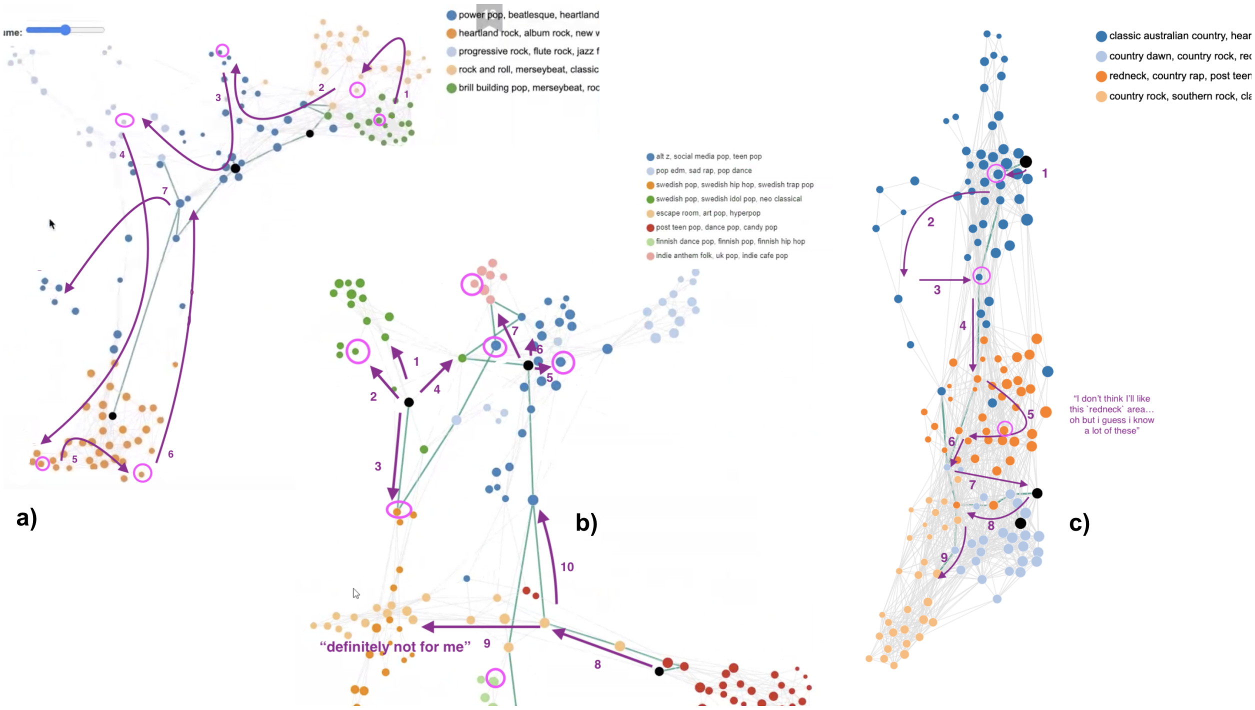

ıme:
power pop, beatlesque, heartland
heartland rock, album rock, new w
progressive rock, flute rock, jazz f
rock and roll, merseybeat, classic
brill building pop, merseybeat, roc
1
2
3
4
5
6
7
a)
alt z, social media pop, teen pop
pop edm, sad rap, pop dance
swedish pop, swedish hip hop, swedish trap pop
swedish pop, swedish idol pop, neo classical
escape room, art pop, hyperpop
post teen pop, dance pop, candy pop
finnish dance pop, finnish pop, finnish hip hop
indie anthem folk, uk pop, indie cafe pop
1
2
3
4
5
6
7
8
9
10
"definitely not for me"
b)
classic australian country, hear
country dawn, country rock, rec
redneck, country rap, post teen
country rock, southern rock, cla
1
2
3
4
5
6
7
8
9
"I don't think I'll like this `redneck` area… oh but i guess i know a lot of these"
c)

Figure 4.2: **Three exploration strategies users employed: a) systematic exploration, b) anchoring, and c) following the guide.** In systematic exploration (a), users picked a cluster to start from and then systematically explored the other clusters one-by-one. In anchoring (b), users explored artists stemming directly from the three anchor artists. In following the guide (c), users followed the nodes along the green path guide. The numbered arrows indicate the order in which users explored the nodes.

explored the clusters one-by-one (Figure 4.2A). To them, it was more important to see all the clusters rather than to focus on any cluster in particular.

4.7.3 Users want more Control: Human-in-the-loop Growing and Pruning of the Graph

As participants explored, they expressed a desire for more control to shape and direct their search. They wanted to remove artists they had heard before and did not like. Also, a few participants mentioned removing or minimizing entire graph clusters that were less interesting. For example, P4 explained that while he would not want to completely remove a cluster, he wanted to "*de-emphasize it*" and interactively control which clusters to view by "*checking which of these genres to even see*." P7 echoed this idea and added that by pruning, she would be able to better explore other parts of the graph: "*It would be helpful [if i could prune this cluster], and get a better* *view of the pathways in Swedish pop*." By pruning, users would be able to create a graph that is (1) easier to explore and (2) more reflective of their interests.

In addition to removing portions of the graph, users also expressed a desire to grow the graph and imagined an adaptive guide that would lead them. For example, P6 discovered a cluster with a sub-genre she had not heard of called *soul flow* with many artists she liked; she wanted to expand the graph from this cluster and continue to explore it. Besides growing the graph, participants also wanted a more intelligent green-path guide that would change according to their input. P2 explained, "*I wish there was an adaptive guide: once you press on an artist, it would create one [and show] other things that might be on the same pace based on that song that you like*." P7 also imagined giving feedback on what she did not like along the path of the guide, which would redirect it to a sub-section she liked more. Overall, participants wanted even more interactivity to better explore the graph.

4.7.4 Improved Recommendation Explainability through Mental Map

With both the non-personalized and personalized versions of TastePaths, users gained a better understanding of how much variance exists within a genre, were able to understand what they liked and disliked within the genre, and grew their vocabulary to describe their interests. For example, after exploring a non-personalized graph for *pop*, P2 learned about the many genres within it: "*I did not really know these genre names... and now I know what it's called. Hopefully after this I will explore them a little more... Genre is usually an afterthought, so it was nice to see what genre they were in*." After exploring a personalized graph for *alternative r&b*, P10 felt like she better understood which part of the genre she actually liked, pinpointing artists with "*hints [of] underground and with hints of jazz*" as her strongest interest. She felt it was important to understand these sub-genres to better reason about where her recommendations were coming from in a music streaming service. Overall, TastePaths helped users better understand the different sounds within a genre as well as their own preferences.

However, beyond becoming acquainted with its sub-genres, participants wanted to learn specific information about the genre itself, including its sonic characteristics, history, and influences. From the questionnaire results, on a scale of one to seven, participants rated "learning" on average 5.44 (stdev=1.55) with the personalized version and 5.63 (stdev=1.54) with the non-personalized version (Table 4.2). While they generally felt they had learned something, they wished they had learned more. For example, P15 wanted a greater understanding of the graph's organizational structure, including more information on why artists were grouped together and descriptors for each cluster's sonic characteristics. Meanwhile, P9 wanted specific information at the artist level: "*I know more artists, but I don't necessarily know more about them... a little bit more about the artists or their background, their process, how they make their music, things generally about the genre. Something about influence - how house music came along, the lineage from Detroit EDM to house music etc*." While not implemented in the current version, adding more information could improve the visualization in future iterations.

Interestingly, a couple of participants felt that the experience with TastePaths made them reevaluate their knowledge of the genre, sometimes even confusing them. After exploring a non-personalized graph of *pop* music, P10 felt overwhelmed at the vastness of that genre: "*I know nothing about pop now. I feel like pop has just become more confusing and now I'm lost in a sea of subgenres... I feel like I was sitting on a step and now they invited me in the house, and I'm like 'what'*." Without being alienated by the entire genre, some participants felt a disconnect between the sub-genre names of the cluster and how they perceived the music. For example, while exploring a *chopped and screwed* cluster in a non-personalized *hip hop* network, P6 noticed a few artists that did not have that quintessential sound of the sub-genre: "*[chopped and screwed] is kind of slowed down and altered in some way... but this one does not sound so slowed down. So maybe within an artist they have a different vibe*." Across their discographies, artists can make music that touches multiple genres, perhaps not making them the perfect fit for a cluster. Overall, a couple of participants felt a disconnect with TastePaths, in some cases at the entire genre level and in other cases with the label of a cluster.

4.8 Discussion

4.8.1 Informing Future Recommender Systems

From the user study, we learned that participants wanted even more control than what TastePaths already provided to provide in-depth feedback to the system. They viewed the graph as the system's representation of their taste, and they imagined more clusters they could explore outside of their local interests at the graph's boundaries. Participants wanted to extend the graph in directions they liked while also editing this representation by de-emphasizing or pruning certain artists and clusters. This willingness to provide richer feedback has been shown in prior work [108], and is in stark contrast to the current methods offered by recommender systems to elicit feedback. Currently, systems either collect implicit feedback, such as play length and skips, which are not transparent to the user, or explicit feedback like ratings, which are cumbersome to collect [117, 118, 119] and perhaps even misleading [120]. Future recommender systems can include support for more expressive and natural feedback from users to tailor how the system represents and understands their interests.

By enabling expressive feedback, we can better inform the algorithms powering popular recommendation systems. One finding from our user study was that particularly rich and interesting discoveries would lie either between two clusters or on the edge of a cluster. This information could be used as implicit feedback to generate discovery playlists to enable further exploration with minimal effort. In the future, larger studies can be conducted to collect this implicit feedback and understand where users are making discoveries to help design better recommender systems and discovery playlists.

Finally, in addition to being useful for understanding how novices explore, TastePaths can also support experts in generating better curated playlists to improve recommendations. User-created playlists and their metadata are often used to calculate the similarity of tracks and artists [121]. From the formative study interviews with experts, we learned that they often used many different resources to explore a less familiar genre and generate a playlist, which requires a lot of time and effort. To provide more and better training data for models using these manually created playlists, we can enable experts to explore faster and more easily with TastePaths. Future work can include understanding how experts use TastePaths, what they discover, and using their exploration results to power recommender systems.

4.8.2 Providing Users Closure to Facilitate more Responsible AI

Participants in our study appreciated that the graph was both expansive and finite. They felt a sense of accomplishment having explored most of the clusters and pride if they realized they knew most of the artists within them. For example, while exploring a personalized graph of *mathcore*, P15 stated: "*I'm very proud of myself, in my metal fandom. Having seen a lot of these bands, I'm happy with the amount I have been able to recognize*." Currently, many recommender systems do not design for an end to the experience but instead aim to maximize their share of the user's time. Because of this, users often consume content to their detriment, neglecting their other plans and goals [122] and losing their sense of agency [123]. Recent work has shown that users prefer versions of recommender systems that promote active interaction and agency when they have a specific intention in mind. One promising way to support agency is through planning [124]. By setting and following goals, users feel more in control of their consumption, as they feel there is an end to the process, unlike in an endless feed of media. Future work can extend these principles to music recommender systems. During a specific task like exploring a genre, music services could help users form goals on how far or how long to explore during the session to encourage growth and agency as opposed to longer listening sessions.

4.8.3 Guidelines for Helping Users Deeply Explore their Interests

While recommender systems are very useful for helping users find content that closely aligns with their current preferences, they can be augmented further to support users in deeply exploring and expanding their interests. By doing so, we could limit the effects of the filter bubble and promote creativity and individuality instead. From the user study, we established two general guidelines for helping users interactively explore and understand their interests: (1) anchor exploration with content the user knows well and help them venture out in many different ways and (2) help users learn about their interests so that they can recognize and consciously interact with their bubble.

In the user study, participants' prior knowledge helped them navigate the space more confidently. From the artists they knew, they were able to identify opportunity spaces; participants were excited to see an unknown artist, or cluster of artists, connecting two other artists they already liked. Future systems can provide multiple ways for users to explore new content from what they currently enjoy. This could include suggesting what lies between two items they know well (either articles, movies, or artists), or suggesting "gateway" items that are connected but less similar to their current interests to introduce them to a new cluster or adjacent genre within the space. These kind of recommendations can help users confidently explore new content outside of their immediate bubble.

As well as anchoring exploration from content the user currently enjoys, recommender systems can also help users learn about this content to help them understand and interact with their filter bubble. Past work has shown that providing a broad overview of the user's consumption increases their awareness of the content they consume and their feeling of control over it [92, 90]. In addition to helping users acknowledge what content they consume at a high level, we show that overviews can also help users better understand what exactly they like about a sub-area in the space, such as a genre. Future systems can provide users with a more fine-grained understanding of their taste by specifying both the broader categories the user is interested in, and the sub-categories that better reflect the user's taste. By incorporating this information, users will be aware of the system's representation of their interests; they can then consciously choose to remain within this bubble or to explore elsewhere.

4.8.4 Limitations and Future Work

While we carefully designed the study, it is not without limitations. One factor that varied across participants and the two conditions in the evaluation was the edge density and number of clusters in the graphs. While each graph was created in the same way and included exactly 150 nodes, the resulting structure of the graph and number of edges was variable. Because of this, some networks were very densely connected with fewer clusters like Figure 4.2C, while others were more spread-out with more clusters like Figure 4.2B. From the interviews, we found that participants generally preferred to explore networks that were more spread-out and had more clusters, and so there might be graph attributes affecting how participants explore. Future work could investigate how different graph shapes and clusters affect how users explore them. TastePaths also required users to hover over nodes in order to see the artist information, and this required extra effort for exploration, especially when the graphs were denser. Future work could investigate how to best highlight important information in the network to reduce the need of node-hovering.

Another limitation in this work is that we recruited participants interested in exploring new music and who have done so in the past three months. Therefore, our results apply more towards those who are open to exploring rather than the general populace of music listeners. We also conducted a relatively small study with 16 participants and focused on qualitative insights. Future work can involve conducting a larger study, to study how to support users who are less willing to explore their interests. In addition, in the formative study, we only interviewed expert music curators, but it would be valuable to also learn more about the tools used by people who are less experienced or less interested in music. For example, it might be less important for them to get a sense of the range of artists in a genre and more important to know what's popular, what the social connections are, etc.

Finally, TastePaths can be altered to study how users transition from one genre to another. In this work we focused on studying how users explore a single, familiar genre, but there is still a lot to learn about how users would like to explore a completely new genre. TastePaths could visualize a familiar genre, as well as a highly related genre the user currently does not listen to; from this setup, we could gain insights toward how to best guide users to new genres from their current interests.

4.9 Conclusion

This chapter covered TastePaths, our interactive web tool that helps users deeply explore and understand the music genres they listen to. We conducted a qualitative study where participants used a personalized and non-personalized version of TastePaths to explore two music genres they listen to often. Our study aimed to understand if TastePaths helps users explore their genres of interest and more broadly, how to better support users in exploring and understanding their preferences. We found that participants greatly preferred the personalized version and wanted even more personalization. They also wanted more control of the graph, including the ability to expand or prune sections of it to better reflect their interests. Finally, they also gained a better mental model of what they liked within their interests and desired to learn even more. Future tools in this space can investigate how to better incorporate learning into exploratory search, how to incorporate more closure and goal-fulfillment in recommendation systems, and how to support users in modifying the system's representation of their taste and interests.

Chapter 5: AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models

5.1 Introduction

Journalists often write stories by carefully analyzing claims made in documents for newsworthiness. These documents can come from freely-shared documents, like press releases, private information leaks, like the Enron email dataset and Panama Papers, as well as public records accessed by Freedom of Information Act (FOIA) requests. Writing stories from these documents is currently mentally taxing and requires a careful consideration of each claim's potential controversy and newsworthiness in order to brainstorm potential angles for stories. An angle is a *framing* of an event or document that "call[s] attention to some aspects of reality while obscuring other elements, which might lead audiences to have different reactions" [125]. Each angle forms a perspective from a few key claims of a document and sets the groundwork for developing a story. Then an angle is substantiated through interviews and information gathering from relevant resources. And by considering multiple angles for a document, journalists can make better decisions on what kind of story to write. But with shortages in newsrooms [126] and the abundance of these documents, journalists do not have the time to comprehensively explore multiple framings for each document.

Current computational tools for journalists predominantly support computational news discovery (CND) - the data-driven identification of potentially newsworthy information [127] [128]. CND tools are often built atop data streams like social media feeds and government websites to direct journalists to anomalous information that could lead to a story, such as a recent high-volume of posts or a recent document detailing a new algorithm the government is using. While these tools help direct journalists' attention to interesting information, they do not help them explore many story angles for that piece of information. From a three-month long co-design with four professional journalists, we learned that in the early stages of a story, an essential part of a journalist's process is to brainstorm multiple angles that can be substantiated into coherent and verifiable stories. For example, a new government policy or initiative might lead to many different controversies and negative outcomes, and currently, journalists rely on their expertise, which may be limited or biased, to consider these effects. In this work, we study how to help journalists explore multiple different angles, given an interesting document.

Large language models (LLMs) have shown great potential in many ideation tasks. Pre-trained on billions of text sources from the Internet, LLMs are a fundamental shift in natural language processing (NLP) [129]. They contain vast world-knowledge and are often able to generate fluent natural language text. With few to no examples, LLMs can reliably execute a number of complicated tasks, including summarization, information extraction, and ideation with remarkably fluent and accurate completions [130]. Within human-computer interaction, LLMs have been used for a variety of tasks, including helping science writers brainstorm ways to communicate their findings [131] and enabling creative writers to explore many ways of writing a story by generating character arcs [132]. In this work, we study if and how LLMs can help professional journalists brainstorm story ideas from a document.

While journalists write stories from many different documents, we focus on press releases. The journalists in the co-design emphasized that press releases, especially those released by government administrations, are a timely and important source of information for writing stories, which is in line with findings from prior work [133]. During a few of the co-design sessions, we observed the journalists as they brainstormed angles for press releases, and we formed four design goals for AngleKindling, our interactive web tool that supports angle exploration for press releases. To help journalists cut through the fluff of the press release, AngleKindling summarizes it into a set of main points. To support angle ideation, AngleKindling employs a LLM to suggest potential controversies and negative outcomes, which call into question the claims and positive bias of the press release. Then, to help journalists verify these angles, AngleKindling links them to the source text, pointing to parts of the press release relevant to each angle. And finally, to provide context for each angle, AngleKindling provides a related news article as historical background.

To evaluate AngleKindling we conducted a within-subjects user study with 12 professional journalists, comparing AngleKindling to INJECT [134], a recent creativity support tool for journalists that also supports angle exploration. INJECT utilizes more traditional natural language processing techniques to supply related articles and extracted entities relevant to the source text; these articles are grouped by their broader, overarching angles. Our findings show that participants found AngleKindling significantly more helpful for brainstorming ideas, while also requiring significantly less mental demand than INJECT.

To summarize, this work contributes the following:

- Four design goals for helping journalists explore angles for press releases, based on findings from our three-month long co-design with four professional journalists.
- AngleKindling, our interactive tool for exploring angles given a press release, which uses a LLM to generate numerous angles like controversies, facilitates trust by linking these angles to the source text, and provides historical background for each angle via a related news article.
- Findings from our evaluation, demonstrating that AngleKindling was perceived to be significantly more helpful for brainstorming story ideas, while requiring significantly less mental demand than the baseline. This was primarily due to AngleKindling (1) helping journalists recognize angles they had not considered, (2) providing angles that were useful for multiple types of stories, (3) helping journalists quickly and deeply engage with the press release, and (4) providing contextualized historical context.
- A discussion that highlights rich areas of future work, including enabling angle customization and prioritizing angles that are more promising than others. We also discuss how the techniques used in this work can be applied to other domains like case law and academia, where LLMs can be used to explore how the decision made in one case might affect the outcomes of similar disputes and the ethical implications of academic papers, respectively.

5.2 Related Work

5.2.1 Computational Journalism

The role of computation in journalism has predominantly been studied in two overarching categories: (1) understanding how technology shapes how consumers and audiences interact with and consume news media and (2) designing tools to improve journalists' capabilities and understanding how these tools affect their workflows [135]. Within the first category, past work has examined how news is shared [136], specifically the role search engines play in biasing the content we consume [137], as well as how users interact with and consume news [138] [139] [140]. Within the second category, numerous tools have been built that support a diverse set of journalistic tasks, including categorizing and understanding large document collections [28], identifying claims to be fact-checked [141] [142], and examining events and content on social media [143] [144] [145] [146] [147] [148]. This work contributes to this growing body of research on supporting journalists' abilities with technology.

CND tools help orient journalists' attention to newsworthy information with algorithms [127] [128]. In more open-ended tasks, like exploring a large set of documents, CND tools often incorporate visualizations so that journalists interactively identify newsworthy information. For instance, *Overview* visualizes a hierarchy of document clusters in a tree-structure, which helped journalists better grasp a birds-eye view of the information within a corpora and identify branches that interested them [28]. Similarly, *Jigsaw* helped users explore document sets at the entity level, by visualizing their relationships across documents in a network [149]. As well as guiding users to newsworthy information via visualizations, other CND tools monitor data-streams of data for trends and anomalies to explicitly highlight newsworthy data. For example, *CityBeat* monitors geotagged social media data to direct journalists to budding local events via abnormal spikes in posts [143]. Another system in this vein, *SRSR* (Seriously Rapid Source Review), directs journalists' attention to user-accounts on Twitter that might be useful sources for breaking news events [150]. Instead of social media data, *Algorithm Tips* monitors government websites for documents describing new algorithmic decision making systems [151]. To present promising leads to journalists, the system employs a crowd to rate each document on a few newsworthy categories. While these CND systems effectively direct journalists' attention to interesting documents and pieces of information, their support often ends there. Multiple narratives and angles can be spun from an interesting document, and in contrast to these CND systems, AngleKindling supports this process of brainstorming angles after an interesting document has been found.

The most comparable system to AngleKindling is INJECT, which also supports the creative process of brainstorming journalistic angles [134]. To help users discover angles for a particular topic, INJECT employs traditional natural language processing (NLP) techniques to provide suggestions of relevant people and articles with different types of angles: causal, quantifiable, and ramifications. The system also uses template-based "creative sparks", which are general suggestions that encourage journalists to consider how a related article's angle might be applied or related to the journalist's story. AngleKindling has similar design goals and also provides related articles and suggestions, but structures them differently. AngleKindling employs a LLM to generate potential *controversies*, *negative outcomes*, and *areas to investigate* from a press release; these are more specific suggestions directly tied to the text, compared to INJECT's sparks. Then to provide further context for these angles, AngleKindling connects each one to a related news article. Therefore, the two systems provide different types of creativity support: AngleKindling is *generative* and *specific*, synthesizing angles tailored to the press release, while INJECT is *associative* and *general*, providing references to previous, related angles and sparks that are more general. In this work, we conduct a user study comparing AngleKindling to INJECT to understand (1) which kind of creativity support journalists prefer and (2) how LLM-based suggestions compare to providing articles with different angles in terms of helpfulness and cognitive load (3) and how to better design angle brainstorming systems.

5.2.2 Creativity Support with Large Language Models

Generative models are being successfully applied to support a number of creative tasks, including music composition [152], designing visual art [153] [154], and writing [131] [132] [155]. Large language models, in particular, are transforming a number of creative tasks. Trained on billions of documents, LLMs contain a vast amount of general world knowledge and can perform numerous NLP tasks without requiring pre-training [130] [156]. LLMs have been used as open-ended collaborative writing tools; with *Wordcraft*, users can view multiple completions from a LLM as well as explore portions of text written in different styles [157]. LLMs have also been shown to be effective in more constrained contexts, including generating suggestions for science communication. Science writers found *Sparks* LLM-suggestions both interesting and useful as a means to understand a reader's perspective [131]. *BunCho*, also uses a LLM to generate titles and synopses from keywords [158]. Finally, LLMs are also currently being used to enable end-users to develop their own AI-infused applications in the form of LLM-chains, where the output of one LLM-step is fed into another [159] [160] [161]. In all of these applications, LLMs have been shown to be effective tools for increasing creativity and useful even when they provide unintended or incorrect outputs. In this work, we apply an LLM to generate angles for journalists, a context in which trust and verification is essential. To help journalists verify angles, we connect them back to the source text. In this work, we investigate if an LLM's common sense reasoning can help journalists think of angles they would not have otherwise.

5.3 Co-Design with Professional Journalists

We conducted a three-month long co-design with four professional journalists (3 male, average age = 38.25), with journalistic experience ranging from 2 to 28 years. The purpose of the co-design was broadly to develop a tool that would help journalists be more productive in writing stories. We met with the journalists once a week, for an hour, to discuss potential problems we could address and datasets to experiment with. In this section, we describe (1) how we identified the need to support angle exploration, given a potentially interesting piece of information, (2) our early experimentation with GPT-3 to support angle-exploration, and (3) a formative study where we derive design goals for a tool that supports angle exploration for press releases.

5.3.1 Supporting Angle Ideation for an Interesting Source

In the earlier sessions of the co-design, we learned that finding an interesting piece of information did not immediately constitute a story, which is in-line with prior work [162]. Initially, the journalists were interested in writing stories about posts made on Gab[1], a conservative social network that operates similarly to Twitter. Our first goal was to help journalists find newsworthy posts on the website, much like the CND tools described in Section 5.2.1. We developed a simple interface that helped journalists search for posts on Gab, based on a few news values: power elite, impact, and timeliness [163]. To include the *power elite* news value, the interface consisted of an updating list of posts made by prominent Gab users, including politicians, people running for office, CEO's, and wealthy individuals. To include the *impact* and *timeliness* news values, users could respectively sort posts by their number of likes and the date they were posted. While this interface helped the journalists explore potentially newsworthy posts, we learned that a post by itself is not a story. For example, the journalists found an interesting post by Marjorie Taylor Greene (MTG), a Georgia Congresswoman. She posted how she had accumulated over "60k in fines" from refusing to wear a mask on the House floor (which was approximately true[2]). While this is an interesting and potentially newsworthy claim, a richer story needs to do more than recount what a powerful individual claims.

There are many ways an interesting piece of information can be spun into a story, and to find a compelling narrative, journalists benefit by exploring many angles. For the MTG post, the journalists brainstormed a few angles. One angle focused on health and included questions like, "How many times did Marjorie Taylor Greene not wear her mask to accrue those fines? Have other members of Congress gotten COVID-19 while she did this?" The answers to these question

[1] https://gab.com/

[2] https://www.huffpost.com/entry/marjorie-taylor-greene-mask-rules-2021_n_6181d57be4b06de3eb6d964f

could lead to a story on MTG endangering the health of prominent politicians. Next, another angle focused on her appeal to her base and involved questions like, “Is 60k a lot of money to Marjorie Greene? Is she taking these fines to make an easy appeal to her base or because she genuinely does not believe in mask wearing?”. Depending on the answers to these questions, a story could be written on MTG’s motivations and behaviors. From a single interesting post, different lines of reasoning could lead to very different stories, each with their own follow-up research and interviews. To help journalists assess more angles more easily, we aimed to support the process of brainstorming angles given an interesting piece of information.

Given their world knowledge and ability to generate fluent text, we experimented with LLMs as an approach to support brainstorming angles. We prompted GPT-3, OpenAI’s large language model, to ideate a connection between a post and a number of angles, including health, technology, and economics. For example, for the economic angle, we prompted GPT-3 to “List the potential implications of the actions described in the post on the American economy”. Even with no training examples, GPT-3 produced compelling connections. For instance, Paul Gosar, a US representative for Arizona, praised the new “Don’t say gay bill” Florida had just passed, which forbids elementary school teachers from discussing gender and sexual orientation with their students. For the economic angle, GPT-3 aptly connected this post to past events where companies had moved offices and ultimately jobs from states where anti-LGBTQ laws had been passed and concluded the same could happen in Florida. The journalists had not considered this connection and ultimately found it helpful as potential story inspiration, giving an early indication that LLMs could be useful for angle exploration.

This initial experimentation with Gab posts showed promise that GPT-3 could help journalists read between the lines of text to reveal its implications, but ultimately, these posts contained too little text to extrapolate on for angles. Instead, the journalists pointed to press releases as a more applicable data source for angle exploration. Press releases contain more text and are often rife with positively biased claims for which LLMs can be used to identify implications. At the same time, writing stories from press releases is a very common task [164] that is not well-supported by technology. Newsrooms are inundated with press releases, where individuals or teams are focused on churning out stories from them. Because of their prevalence in journalism and their claim-filled content, we focus on bettering angle exploration for press releases.

5.3.2 Formative Study: Brainstorming Angles from Press Releases

To understand how to best support angle exploration for press releases, we observed the journalists brainstorm angles for two press releases and interviewed them on their process. They were asked to imagine that their editor had handed them the press release to come up with a few potential story ideas. For each press release, they were given 15 minutes to come up with multiple, different ideas, reflecting the time constraints of a newsroom. As they brainstormed, they were asked to record their ideas in a separate document. We chose press releases distributed by New York City's mayor, Eric Adams, since the journalists lived in or nearby the city and would have the requisite background knowledge to identify the locations and individuals mentioned in the document. One press release (PR1) was about a new safety plan for the city's subway system and the other (PR2) was about plans for a new offshore wind hub to supply electricity for the city.

Findings: Design Goals

Each journalist started by carefully reading the press release. They noted how press releases are typically biased and filled with fluff, or less informative phrases that only praise the government's actions. The journalists tried to quickly skim through this fluff to (1) quickly understand what the press release is addressing and (2) to collect important information. This important information often included specific details pertaining to the plans described in the release. For the wind hub release, the journalists collected information on which companies were contracted to help with construction, the length of these contracts, the number of projected jobs, and other concrete information. These pieces of information were the foundation for the angles they brainstormed, but required a tiresome and potential error-prone process of scanning the document for them. Therefore, our first design goal was to **summarize the press release into a set of main points, to help**

journalists quickly cut through the fluff and identify important details.

From the claims and concrete details they collected from the press release, the journalists ideated story angles, like potential controversies and negative outcomes. For the subway safety plan press release, P1 collected information pertaining to the number of police that would be deployed at each station and their new role to remove homeless people from the subway at the end of each stop. From this information, he wrote down two controversies: (1) the increase in police presence may not reduce violence in the subway and (2) there might only be an increase in police brutality toward the homeless population. In addition to these potential controversies, he also wrote down questions for follow-up investigation, including "Have there been past subway plans and were they effective? Does increasing police presence normally reduce violence?" Similarly, for the wind hub press release, both P2 and P4 recognized that the city had employed a petroleum company to build the wind hub. P2 questioned how the petroleum company landed the contract as well as "how long they had been lobbying the city for this contract". P4 questioned if a petroleum company should be leading the city's green energy movement. Thinking of these controversies and questions was mentally demanding and therefore, **our second design goal was to provide angles that focus on elements of conflict and controversy**. Finally, the journalists brainstormed these angles directly from claims made in the press release, and thus **to facilitate trust in the angles we provide, our third design goal was to ground them in the source material**.

All four journalists emphasized the importance of getting historical background to either (1) think of new angles or (2) to get supporting evidence for the angles they brainstormed. While working on the subway safety plan release, P3 had questions including, "What have other cities done for subway safety plans?" and "How has New York's subway policy changed over the years?" P3 stated that these are questions he would then answer by consulting past news articles written about New York's and other city's subway policies. He explained that by acquiring this historical background, new angles might appear, like "New York is trying the same, ineffective methods to mitigate subway violence" or "New York's new subway plan is radically different from that of other cities." As well as inspiring new angles, historical background can also provide supporting evidence for angles already brainstormed. For the wind hub press release, P2 became interested in the petroleum company's role in constructing the wind farm; he hypothesized that there could be tension from the local community and split opinions on the new hub. While he did not have time during the 15-minute time limit, he explained his next step would be to read past articles on this deal to either validate or refute this hypothesis. Therefore, **our last design goal was to provide relevant historical background to validate and spark new angles**.

Design goals. In summary, we formed four design goals for AngleKindling from the co-design:

D1: Cut through the fluff by summarizing the article into a set of main points.

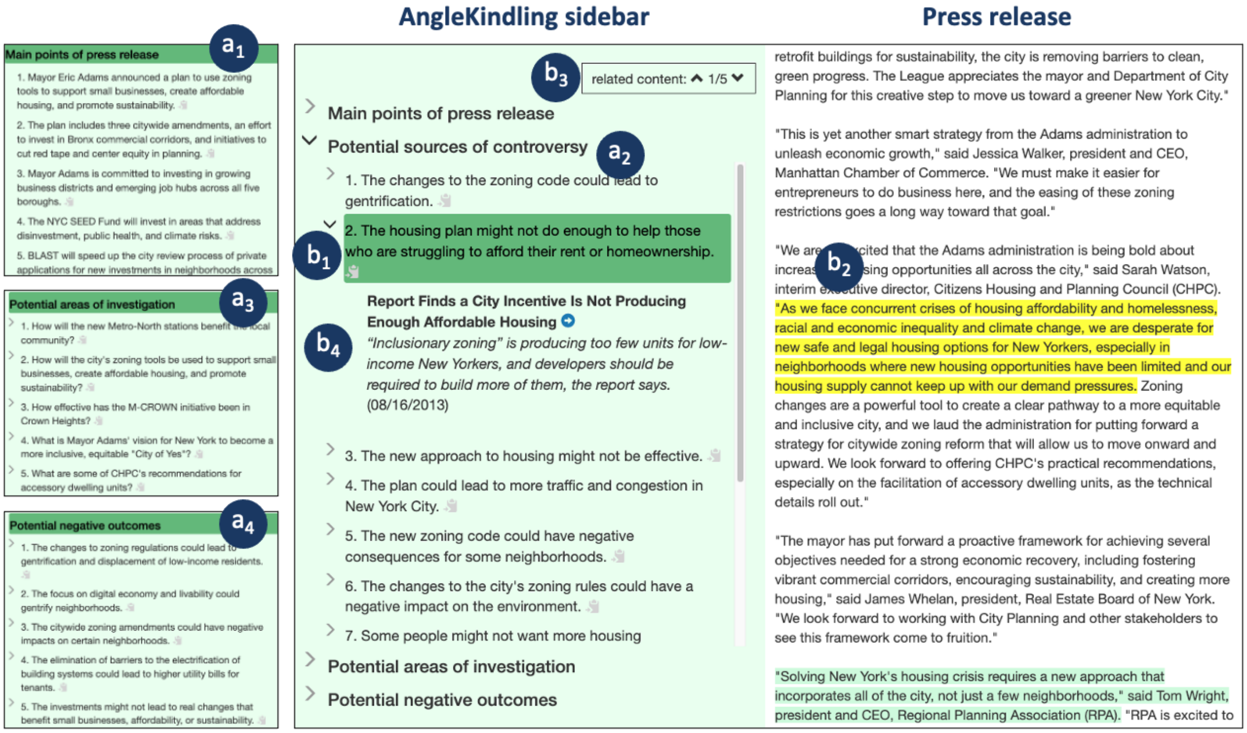
D2: Provide angles focused on conflict and controversy to help journalists call in to question the positive bias of the press release and inspire story ideas.

D3: Facilitate trust by connecting the provided angles directly to the source text (the press release).

D4: Provide relevant historical background to assess angles the journalists brainstorm, show what's been written, and inspire new angles.

5.4 AngleKindling

To address these design goals, we created AngleKindling: an interactive web tool that supports journalists in brainstorming angles, given a press release (Figure 5.1). AngleKindling displays the input press release on the right and the angle suggestions in the green sidebar on the left. The press release in this example is another by Eric Adams, announcing new zoning changes to improve New York's affordable housing and energy efficiency. To address D1, *summarize the press release*, AngleKindling provides a list of the press release's main points (a_1), to help journalists skim the content quickly. To address D2, *provide angles*, AngleKindling provides a list of potential *controversies* (a_2) and *negative outcomes* (a_4) to offer an alternative perspective to the claims made in the press release, as well as *areas of investigation* (a_3) to offer questions the journalist might consider for inspiration. To address D3, *facilitate trust*, AngleKindling connects each angle and main point to five relevant portions of the press release. In this case, the user selected the second


AngleKindling sidebar

Press release

Main points of press release

1. Mayor Eric Adams announced a plan to use zoning tools to support small businesses, create affordable housing, and promote sustainability.
2. The plan includes three citywide amendments, an effort to invest in Bronx commercial corridors, and initiatives to cut red tape and center equity in planning.
3. Mayor Adams is committed to investing in growing business districts and emerging job hubs across all five boroughs.
4. The NYC SEED Fund will invest in areas that address disinvestment, public health, and climate risks.
5. BLAST will speed up the city review process of private applications for new investments in neighborhoods across

Potential areas of investigation

1. How will the new Metro-North stations benefit the local community?
2. How will the city's zoning tools be used to support small businesses, create affordable housing, and promote sustainability?
3. How effective has the M-CROWN initiative been in Crown Heights?
4. What is Mayor Adams' vision for New York to become a more inclusive, equitable "City of Yes"?
5. What are some of CHPC's recommendations for accessory dwelling units?

Potential negative outcomes

1. The changes to zoning regulations could lead to gentrification and displacement of low-income residents.
2. The focus on digital economy and livability could gentrify neighborhoods.
3. The citywide zoning amendments could have negative impacts on certain neighborhoods.
4. The elimination of barriers to the electrification of building systems could lead to higher utility bills for tenants.
5. The investments might not lead to real changes that benefit small businesses, affordability, or sustainability.

related content: 1/5

Main points of press release

Potential sources of controversy

1. The changes to the zoning code could lead to gentrification.

2. The housing plan might not do enough to help those who are struggling to afford their rent or homeownership.

Report Finds a City Incentive Is Not Producing Enough Affordable Housing

"Inclusionary zoning" is producing too few units for low-income New Yorkers, and developers should be required to build more of them, the report says.

(08/16/2013)

3. The new approach to housing might not be effective.

4. The plan could lead to more traffic and congestion in New York City.

5. The new zoning code could have negative consequences for some neighborhoods.

6. The changes to the city's zoning rules could have a negative impact on the environment.

7. Some people might not want more housing

Potential areas of investigation

Potential negative outcomes

retrofit buildings for sustainability, the city is removing barriers to clean, green progress. The League appreciates the mayor and Department of City Planning for this creative step to move us toward a greener New York City."

"This is yet another smart strategy from the Adams administration to unleash economic growth," said Jessica Walker, president and CEO, Manhattan Chamber of Commerce. "We must make it easier for entrepreneurs to do business here, and the easing of these zoning restrictions goes a long way toward that goal."

"We are excited that the Adams administration is being bold about increasing housing opportunities all across the city," said Sarah Watson, interim executive director, Citizens Housing and Planning Council (CHPC). "As we face concurrent crises of housing affordability and homelessness, racial and economic inequality and climate change, we are desperate for new safe and legal housing options for New Yorkers, especially in neighborhoods where new housing opportunities have been limited and our housing supply cannot keep up with our demand pressures. Zoning changes are a powerful tool to create a clear pathway to a more equitable and inclusive city, and we laud the administration for putting forward a strategy for citywide zoning reform that will allow us to move onward and upward. We look forward to offering CHPC's practical recommendations, especially on the facilitation of accessory dwelling units, as the technical details roll out."

"The mayor has put forward a proactive framework for achieving several objectives needed for a strong economic recovery, including fostering vibrant commercial corridors, encouraging sustainability, and creating more housing," said James Whelan, president, Real Estate Board of New York. "We look forward to working with City Planning and other stakeholders to see this framework come to fruition."

"Solving New York's housing crisis requires a new approach that incorporates all of the city, not just a few neighborhoods," said Tom Wright, president and CEO, Regional Planning Association (RPA). "RPA is excited to

Figure 5.1: **AngleKindling's interface** displays the press release on the right and the article's main points (a_1) along with angle suggestions in the green sidebar on the left. The angle suggestions include potential *controversies* (a_2), *areas of investigation* which are questions to consider (a_2), and *negative outcomes* (a_4) that could arise. To help users trust these angles, they can select them (b_1) to view related content from the press release (b_2), and they can skim through up to five pieces of text with the related content button (b_3). Finally, each angle is connected to a New York Times article from the past decade (starting in 2012) to provide historical background (b_4). The title, lead paragraph, and publication date are provided for the article, as well as a link to the article itself, via the blue arrow.

controversy (b_1): "The housing plan might not do enough to help those who are struggling to afford their rent or homeownership". AngleKindling then highlighted a relevant portion of the press release (b_2), which in this case, is a quote that directly opposes the *controversy*, claiming that the new zoning laws will improve the housing opportunities in less fortunate neighborhoods. Users can continue to skim through connected content with the related content button (b_3); the portion in-focus is highlighted yellow, while the rest are green. Finally, to fulfill D4, *provide relevant historical background*, a relevant article from The New York Times is retrieved for each angle (b_4). In this case, the article is from 2013, discussing how a past zoning measure had not

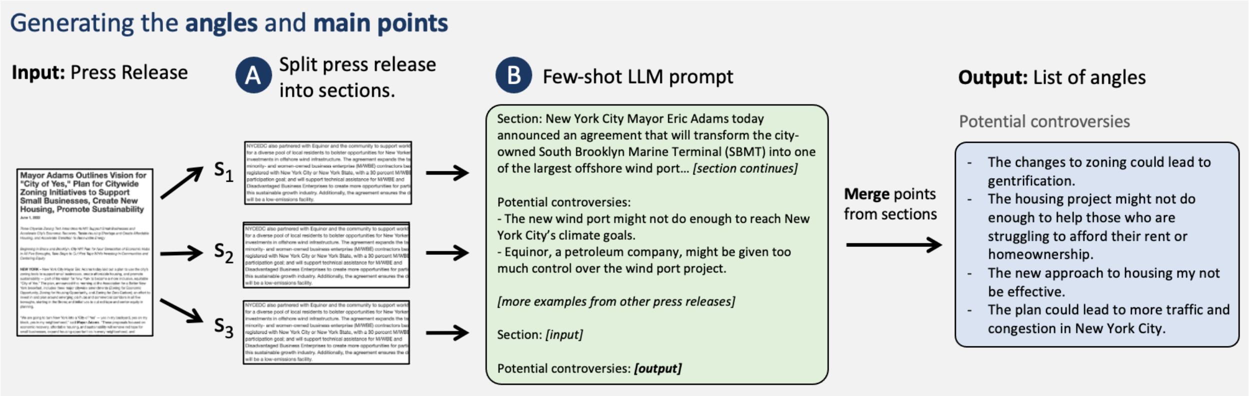

Generating the angles and main points
Input: Press Release
A
Split press release into sections.
B
Few-shot LLM prompt
Output: List of angles
Mayor Adams Outlines Vision for "City of Yes," Plan for Citywide Zoning Initiatives to Support Small Businesses, Create New Housing, Promote Sustainability
S_1
S_2
S_3
Section: New York City Mayor Eric Adams today announced an agreement that will transform the city-owned South Brooklyn Marine Terminal (SBMT) into one of the largest offshore wind port... [section continues]
Potential controversies:
- The new wind port might not do enough to reach New York City's climate goals.
- Equinor, a petroleum company, might be given too much control over the wind port project.
[more examples from other press releases]
Section: [input]
Potential controversies: [output]
Merge points from sections
Potential controversies
- The changes to zoning could lead to gentrification.
- The housing project might not do enough to help those who are struggling to afford their rent or homeownership.
- The new approach to housing my not be effective.
- The plan could lead to more traffic and congestion in New York City.

Figure 5.2: To **generate the angles and main points**, AngleKindling first splits the press release into a set of sections, to fit the input length of the LLM (A). Each section is then inputted to a set of four LLM prompts, to (1) extract the main points of the section (2) ideate potential *controversies*, (3) identify *areas to investigate*, and (4) ideate potential *negative outcomes* (B). Each LLM prompt is few-shot and contains three examples of converting a section into a set of main points or angles. The examples are taken from the angles thought of by the journalists in the formative study. Finally, the angles ideated from each section are then merged together into a single list.

improved conditions for low income New Yorkers, providing evidence for the *controversy* in (b_1) and contradicting the official's quote in (b_2). From here, the journalist can click on the blue arrow in (b_4) to read the article in full, and see if Eric Adam's new zoning plan proposes significant changes to past plans, or continue exploring other angles. Together, these features help journalists take an interesting source of information, like a press release, and explore multiple different story directions.

5.5 Implementation

AngleKindling is implemented in the Flask web-framework. To summarize the press release and generate the angles, AngleKindling employs GPT-3, OpenAI's large language model, via their API [3]. While we use this LLM, our work can be replicated with any other LLM. A central feature of AngleKindling is also connecting each angle to relevant sentences in the press release and a New York Times article. To connect content we embed the angles and press release content using

[3] https://openai.com/api/

Sentence-BERT [165], with the *all-mpnet-base-v2* pre-trained model in particular, via their API [4]. Finally, we use the New York Times API to link a relevant article to each angle [5]. In the following section, we describe how we use these tools to implement AngleKindling's core features.

5.5.1 Providing Angles and Main Points

Given the promise they showed in the co-design, we continued to use a LLM, specifically GPT-3, to fulfill D1 and D2 and generate angles for press releases. The press releases we collected were too long to fit in the input length of GPT-3. To generate angles across the entire document, we split the press release into a set of sections (Figure 5.2A) and generated angles for each section (Figure 5.2B). Each section contained as many complete paragraphs that could fit, along with the prompt, in the input length of GPT-3. Initially, we used zero-shot prompts to ideate controversies. For each section, GPT-3 was prompted to "Create a list of controversies that could potentially arise from the following article section", without any training examples. The completions for the zero-shot prompt would sometimes produce a compelling result, but would mostly output generic, unhelpful controversies like "The plan will fail." At the same time, the completions were often phrased as facts, and instead, we wanted to hedge each controversy to facilitate trust. Therefore, we switched to a few-shot prompt, for which the examples consisted of a press release section, paired with a list of controversies that the journalists thought of from the formative study (Figure 5.2B). The resulting angles, like "The plan could lead to more traffic and congestion in New York City" were more specific and hedged to emphasize that they were *possible* controversies instead of facts.

Extracting the main points of the press release is done similarly to the angles, but involved an extra challenge in removing the press release fluff within each point. Once again, as illustrated in Figure 5.2, the press release is split into sections, where for each one, a few-shot LLM prompt extracts the main points. However, each main point tended to include superfluous information that only served to further the document's positive bias. For instance, from the offshore wind press release in the formative study, one main point was that "New York City Mayor Eric Adams

[4] https://www.sbert.net/

[5] https://developer.nytimes.com/

today announced an agreement that will transform the city-owned South Brooklyn Marine Terminal (SBMT) into one of the largest offshore wind port facilities in the nation" To simplify this point, we use another few-shot LLM prompt to rewrite it with fewer words, generating the simpler, less-biased sentence: "Mayor Adams announced that the South Brooklyn Marine Terminal will be turned into an offshore wind port." With this extra step, we are able to provide a summary that is easier to read and better cuts through the fluff.

5.5.2 Connecting Angles to the Source Text and Historical Background

While these main points and angles might be accurate and inspire ideas, they are difficult for journalists to trust without explicitly tying them back to the source material. To help facilitate this trust, we identify each angle's top five most related sentences in the press release. To do so, we compare the similarity between each angle to each sentence in the press release. For each angle we compute a vector, using Sentence-BERT. Next, we split the press release into sentences using spaCy's[6] built-in sentence segmentation; each sentence is then embedded, also with Sentence-BERT. Finally, we compute the cosine-similarity between each angle to each sentence, and the top five sentences are selected to be highlighted by the related content button (Figure 5.1b_3). By explicitly connecting each angle and main point to the source text, we help journalists quickly verify their relevance.

As well as connecting each angle to the source text, another crucial feature of AngleKindling is bringing historical background by connecting each angle to a past news article. We use The New York Times (NYT) as our source of news articles, as it is (1) a reputable and exemplary news source trusted by journalists and (2) likely to cover the important problems and plans that pertain to New York City. To connect each angle to a news article, we first collect a set of relevant NYT articles for the press release. To do so, we extract the top five most relevant keywords from the press release, once again with a few-shot LLM-prompt. Each keyword is then used to query New York Times articles from the past decade, using their developer API. Through this process we normally

[6] https://spacy.io/

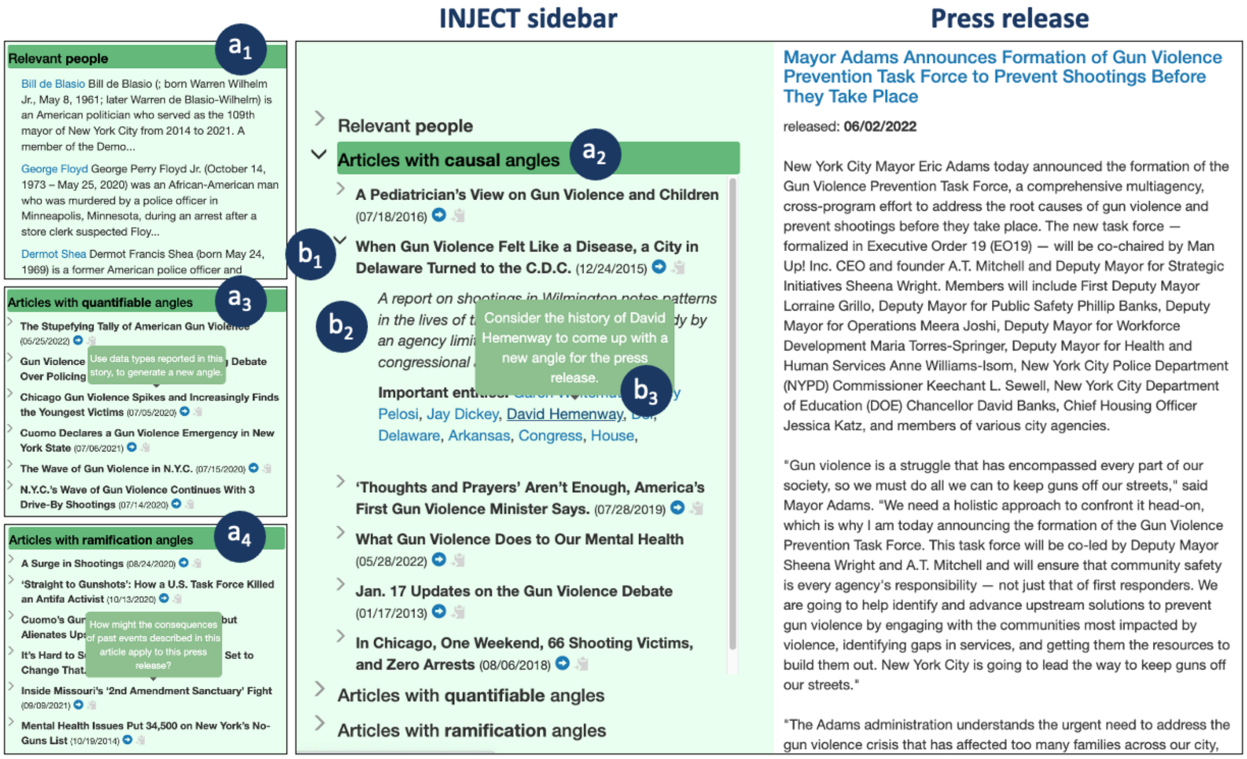
collect approximately 300 relevant articles. For each relevant article, we concatenate its headline and first paragraph to compute an embedding using Sentence-BERT. Often the first paragraph of a news article will convey the most important facts of the story, which along with headline, can be used as the representative material for the article. We then compute the cosine-similarity of these headline embeddings with each angle embedding, and choose the highest scoring article to use as historical background. By doing so, we help journalists gather context for each angle through relevant historical knowledge.

5.6 Evaluation

To understand how AngleKindling may help journalists brainstorm story ideas, we conducted a within-subjects study, comparing AngleKindling to INJECT, a comparable creativity support tool for journalists. To understand what participants liked and disliked about these systems we (1) include a questionnaire to get quantitative measures for each tool's features and helpfulness as well as (2) conduct a semi-structured interview to get qualitative insights on participants' preferences.

5.6.1 INJECT Interface

INJECT is a creativity support tool created to help journalists write stories faster by helping them discover creative angles for stories. In its evaluation, INJECT was deployed in multiple news outlets and used to develop multiple, published stories; INJECT helped journalists come up with new ideas and angles for their stories quickly, "often in less than 3 minutes for each story" [134]. INJECT was originally implemented as a Google Docs Add-on sidebar, so that journalists could seamlessly get creative support as they wrote their story. The original INJECT includes six sources of creativity support, of which we include four (all based on prior news articles): (1) **Quantifiable**: articles that contain quantified information, such as actual numbers, and keywords like *Sterling* and *population*, (2) **People**: information on individuals (from Wikipedia) extracted from related news articles, (3) **Causal**: articles that discuss the background or causes of a story, identified through keywords like *cause*, *impact* and *studies*, and (4) **Ramifications**: articles that



Relevant people

Bill de Blasio Bill de Blasio (; born Warren Wilhelm Jr., May 8, 1961; later Warren de Blasio-Wilhelm) is an American politician who served as the 109th mayor of New York City from 2014 to 2021. A member of the Demo...

George Floyd George Perry Floyd Jr. (October 14, 1973 – May 25, 2020) was an African-American man who was murdered by a police officer in Minneapolis, Minnesota, during an arrest after a store clerk suspected Floy...

Dermot Shea Dermot Francis Shea (born May 24, 1969) is a former American police officer and

Articles with quantifiable angles

- The Stupefying Tally of American Gun Violence (05/25/2022)
- Gun Violence … Debate Over Policing
- Use data types reported in this story, to generate a new angle.
- Chicago Gun Violence Spikes and Increasingly Finds the Youngest Victims (07/05/2020)
- Cuomo Declares a Gun Violence Emergency in New York State (07/06/2021)
- The Wave of Gun Violence in N.Y.C. (07/15/2020)
- N.Y.C.'s Wave of Gun Violence Continues With 3 Drive-By Shootings (07/14/2020)

Articles with ramification angles

- A Surge in Shootings (08/24/2020)
- 'Straight to Gunshots': How a U.S. Task Force Killed an Antifa Activist (10/13/2020)
- Cuomo's Gun … but Alienates Ups…
- It's Hard to S… Set to Change That.
- How might the consequences of past events described in this article apply to this press release?
- Inside Missouri's '2nd Amendment Sanctuary' Fight (09/09/2021)
- Mental Health Issues Put 34,500 on New York's No-Guns List (10/19/2014)

Relevant people

Articles with causal angles

- A Pediatrician's View on Gun Violence and Children (07/18/2016)
- When Gun Violence Felt Like a Disease, a City in Delaware Turned to the C.D.C. (12/24/2015)

  A report on shootings in Wilmington notes patterns in the lives of t… …dy by an agency limi… congressional …

  Consider the history of David Hemenway to come up with a new angle for the press release.

  Important enti… …y Pelosi, Jay Dickey, David Hemenway, …, Delaware, Arkansas, Congress, House,
- 'Thoughts and Prayers' Aren't Enough, America's First Gun Violence Minister Says. (07/28/2019)
- What Gun Violence Does to Our Mental Health (05/28/2022)
- Jan. 17 Updates on the Gun Violence Debate (01/17/2013)
- In Chicago, One Weekend, 66 Shooting Victims, and Zero Arrests (08/06/2018)

Articles with quantifiable angles

Articles with ramification angles

Mayor Adams Announces Formation of Gun Violence Prevention Task Force to Prevent Shootings Before They Take Place

released: **06/02/2022**

New York City Mayor Eric Adams today announced the formation of the Gun Violence Prevention Task Force, a comprehensive multiagency, cross-program effort to address the root causes of gun violence and prevent shootings before they take place. The new task force — formalized in Executive Order 19 (EO19) — will be co-chaired by Man Up! Inc. CEO and founder A.T. Mitchell and Deputy Mayor for Strategic Initiatives Sheena Wright. Members will include First Deputy Mayor Lorraine Grillo, Deputy Mayor for Public Safety Phillip Banks, Deputy Mayor for Operations Meera Joshi, Deputy Mayor for Workforce Development Maria Torres-Springer, Deputy Mayor for Health and Human Services Anne Williams-Isom, New York City Police Department (NYPD) Commissioner Keechant L. Sewell, New York City Department of Education (DOE) Chancellor David Banks, Chief Housing Officer Jessica Katz, and members of various city agencies.

"Gun violence is a struggle that has encompassed every part of our society, so we must do all we can to keep guns off our streets," said Mayor Adams. "We need a holistic approach to confront it head-on, which is why I am today announcing the formation of the Gun Violence Prevention Task Force. This task force will be co-led by Deputy Mayor Sheena Wright and A.T. Mitchell and will ensure that community safety is every agency's responsibility — not just that of first responders. We are going to help identify and advance upstream solutions to prevent gun violence by engaging with the communities most impacted by violence, identifying gaps in services, and getting them the resources to build them out. New York City is going to lead the way to keep guns off our streets."

"The Adams administration understands the urgent need to address the gun violence crisis that has affected too many families across our city,

Figure 5.3: **INJECT's interface** incorporates four types of information sources: relevant people (a_1), articles with *causal* angles (a_2), articles with *quantifiable* angles (a_3), and articles with *ramification* angles (a_4). Each article (b_1) is clickable to reveal its first paragraph (b_2), as well as a list of its extracted entities (people, places, organizations, and events) that are linked to their corresponding Wikipedia pages. Each entity can also be hovered over to reveal a inspirational "spark" (b_3). These sparks also appear when a user hovers over an article title (a_4).

discuss the future consequences of a story, identified through keywords like *outcome*, *consequence*, and *aftermath*. INJECT also has features to include comics and data visualizations, but these were less core features, while the other four were referenced the most as useful features in INJECT's deployment. Overall, INJECT is very relevant and powerful tool for angle ideation, and therefore serves as a good baseline for AngleKindling.

We tailored INJECT's core functionality to provide creativity support for press releases and embedded it in the same interface as AngleKindling (Figure 5.3). To provide the articles and people for each press release, we use the same dataset of articles pulled from The New York Times that we collected for AngleKindling, described in Section 5.5.2. INJECT originally has search functionality, but in our case, we assume the articles have already been searched for, using keywords from the press release. To minimize the visual difference of the two interfaces, we incorporate INJECT's articles as drop-downs for each category: *people* (Figure 5.3a_1), *causal* (Figure 5.3a_2), *quantifiable* (Figure 5.3a_3), and *ramifications* (Figure 5.3a_4). The *people* are extracted from the articles shown in the other categories and sorted by frequency. They are also linked to their Wikipedia pages. Next, the articles are assigned to a category (*causal*, *quantifiable*, and *ramifications*) based on a set of pertinent keywords for each one. When users select a category, they can view its articles (Figure 5.3b_1), along with each article's publication date, a link to its page, its first paragraph, as well as its extracted entities: people, places, organizations, and events. Each entity is linked to its corresponding Wikipedia page and like INJECT, includes a hover-over "spark" related to its category (Figure 5.3b_3). These sparks also exist for the article headlines (Figure 5.3a_3), and are generated using templates provided from the original paper. Overall, while our implementation is not an exact copy of INJECT (i.e. the "Quirky" and ”Data visualization" angles were not implemented), we argue that it features enough of its core functionality to compare the strategies of the two tools.

5.6.2 Procedure

The general outline of the study was the following: (1) participants were first interviewed on their journalism background and experience, (2) they then used AngleKindling and INJECT to brainstorm story angles for two press releases by New York City's mayor, (3) after brainstorming with each tool, they filled out a questionnaire rating each tool's features and their experience coming up with ideas, (4) in a semi-structured interview, they were then asked a series of questions on their preferences and thoughts on each tool.

In the experiment phase of the study, participants were randomly assigned to a condition that determined which tool and press release they would brainstorm story ideas with first. Tool and press release order were counter-balanced to prevent a learning effect. Participants were asked to imagine that their editor had assigned them the press release and asked them to come up with many different story ideas for the press release. Before using each tool, they were shown a video

Metrics (Both Conditions)	**Statement (7-point Likert scale)**
Helpfulness	The system as a whole was helpful for coming up with story ideas.
Pursuable Angles	I would pursue some of the angles from this system.
Mental Demand	Coming up with story ideas was mentally taxing with this system.

AngleKindling Metrics	**Statement (7-point Likert scale)**
Main Points	The main points were helpful for skimming the press release.
Related Content	The "related content" button helped me find relevant information in the press release.
Controversies	The controversies were helpful for coming up with story ideas.
Areas to Investigate	The areas to investigate were helpful for coming up with story ideas.
Negative Outcomes	The negative outcomes were helpful for coming up with story ideas.
Historical Background	The articles were helpful for coming up with story ideas.

INJECT Metrics	**Statement (7-point Likert scale)**
People	The relevant people provided were helpful for coming up with story ideas.
Causal	The articles with causal angles were helpful for coming up with story ideas.
Quantifiable	The articles with quantifiable angles were helpful for coming up with story ideas.
Ramifications	The articles with ramification angles were helpful for coming up with story ideas.
Sparks	The hover-over creative sparks were helpful for coming up with story ideas.

Table 5.1: **Post-task questionnaire** filled out by participants after using either AngleKindling or INJECT. For both systems, participants were asked to rate its *Helpfulness*, *Pursuable Angles* and requisite *Mental Demand*. Each system also had its own specific statements for rating each of its features, to gauge what was most helpful of each tool.

demonstrating its features, using the offshore wind press release as an example. Also, since both tools used New York Times articles, participants were given login information for the publication if they did not have a subscription. After they felt they understood each tool's features, they were then given 15 minutes to brainstorm story ideas for the press release. From the co-design, we found that this time-limit was reasonable and reflective of the time constraints that many journalists face in practice at daily news publications. Participants recorded their story ideas in a document and were encouraged to explain their process and reasoning as they came up with ideas. We define "story ideas" loosely as questions or lines of thought they were genuinely interested in pursuing.

	AngleKindling	INJECT	p-value
Helpfulness	6.17 (0.99)	3.92 (1.38)	**<.05**
Pursuable Angles	6.33 (0.75)	4.5 (2.25)	.058
Mental Demand	1.83 (0.9)	3.42 (1.89)	**<.05**

Table 5.2: Comparison of AngleKindling and INJECT across the three categories from the questionnaire. We conducted three paired-sample Wilcoxon tests with Bonferroni correction, and found that **AngleKindling was perceived to be (1) significantly more *helpful* and (2) significantly less *mentally demanding* to use for brainstorming story ideas**. Average scores are shown with standard deviation in parenthesis. Significant p-values are bolded.

After coming up with story ideas with each tool, participants were asked to fill out a questionnaire (Table 5.1) to understood how each tool and its features helped them brainstorm ideas. And once they brainstormed ideas for both press releases, they were asked a series of questions that probed their preference of each system, how each tool did and did not help them, and how they can be improved.

5.6.3 Participants

We recruited 12 professional journalists (average age = 37, 3 male, experience in the field ranging from 5 to 29 years) via e-mail and social media calls for participation. Eligible participants included journalists that work in any medium, including digital publications, newspapers, magazines, radio or TV. Since the press releases were in English and from New York City, we required participants to be English speakers and based in the United States. The interviews were conducted remotely, and participants had to have a computer with Google Chrome. Participants were compensated $30 for up to 60 minutes of their time.

5.7 Results

From the exit-interviews, all 12 participants preferred AngleKindling to INJECT for brainstorming story ideas. Since the study was within-subjects and the questionnaire involved ordinal data, we conducted three paired sample Wilcoxon tests with Bonferroni correction to compare the two systems' *helpfulness*, how *pursuable* their angles were, and their requisite *mental*

AngleKindling Feature Ratings	
Main Points	6.17 (1.14)
Controversies	5.92 (1.26)
Negative Outcomes	5.75 (1.42)
Related Content	5.67 (1.18)
Areas to Investigate	5.33 (1.75)
Historical Background	4.67 (1.31)

INJECT Feature Ratings	
Quantifiable	4.83 (1.62)
Causal	3.83 (1.07)
Ramifications	3.75 (1.83)
Sparks	3.08 (2.1)
People	2.5 (1.61)

Table 5.3: The questionnaire results for each condition's features. **AngleKindling's** highest rated features were the *Main Points* and potential *Controversies*. The *Main Points* along with the *Related Content* helped users deeply engage with and understand the press release quickly, while the *Controversies* provided many, different ideas for stories. **INJECT's** highest rated feature were the *Quantifiable* articles which many journalists appreciated as a source of data and ideas for incorporating analysis in their stories. Average scores are shown with standard deviation in parenthesis.

demand. We found that **AngleKindling was perceived to be significantly more helpful for coming up with story ideas** (W = 55, Z = 2.96, p < .05), scoring on average 6.17 (std = 0.99) on the questionnaire, while INJECT scored 3.92 (1.38) (Table 5.2). Furthermore, while also more helpful for brainstorming ideas, **AngleKindling also required significantly less mental demand** (W = 0, Z = -2.74, p < .05), scoring on average 1.83 (0.9) compared to INJECT's 3.42 (1.89). Finally, while not significant, participants on average also rated AngleKindling's angles as more pursuable (avg = 6.33, std = 0.75) than those by INJECT (avg = 4.5, std = 2.25). In the following subsections we provide greater context to these results and illustrate that AngleKindling was more helpful because it (1) helped participants recognize angles they originally did not consider, (2) provided angles that were useful for *multiple*, *different* types of stories, (3) helped journalists *quickly* and *deeply* engage with the press release, and (4) incorporated *contextualized* historical background.

5.7.1 AngleKindling helped participants recognize angles they originally did not consider.

The angles produced by GPT-3 in AngleKindling often contained new connections the journalists had not considered as they read the press release. For example, while working on the gun violence release, P2 found a new *area to investigate* that she found promising: "How effective have similar task forces been in other cities?" From this question, she imagined a story that would compare and assess gun violence task forces in cities comparable to New York. For the zoning press release, P9 was surprised by the *controversy* that "The plan could lead to more traffic and congestion in New York City." She had not considered this effect and became interested in interviewing city-planning experts about the new train lines proposed by the plan. However, not all of AngleKindling's angles were immediately useful, such as the following *controversy*: "The plan does not do enough to address the housing crisis". P11 explained that this angle could be "made by anyone about anything". These generic angles did not inhibit participants however; they were quickly able to scan each set of angles for anything interesting. The participants appreciated being able to recognize interesting angles instead of coming up with their own. P8 described AngleKindling as "proactive", helping her to immediately "see how I could write several stories from this one press release". Overall, AngleKindling was able to produce angles that surprised even professional journalists.

Meanwhile, with INJECT, participants had to work harder to think of story ideas. Their process involved assessing angles other journalists had used in the news articles and determining if they could be applied to the press release. As P5 describes, "This one [INJECT] is more: think about what other people did on a similar story and apply it here." For example, for the article "New Jersey Town says 'No Thanks' to Development", P5 explained she might read this story to find out the reasons why the residents of this town wanted less development and see if they're applicable to New York residents where the zoning changes were being made. Thinking of angles this way was more mentally demanding and likely led to the significant difference in rating (Table 5.2); to come up with ideas, participants were required to skim through the article, collect information, and mentally reason if it was applicable to the press release. The hover-over sparks did little to make this process easier, as shown by their low average rating of 3.08 (Table 5.3). Participants found sparks like "Make your angle more similar to the causal angle in this story" too high-level to be helpful. Overall, coming up with story ideas with INJECT involved a few more mentally taxing steps, while AngleKindling preemptively processed the press release to provide actionable, concrete angles.

5.7.2 AngleKindling's different angles were useful for multiple types of stories.

Participants found that AngleKindling's angles to be useful for multiple different stories, including (1) day-of stories, (2) next-day or week-long investigations, and finally (3) months-long retrospective stories. Multiple participants, including P4, P7, P9, and P12 found the *areas to investigate* particularly useful for **day-of stories**: briefer pieces that aim to summarize the key takeaways of the press release. The *areas to investigate* often included questions that P4 described as "aiming to clarify" the press release and useful for gathering information, such as the following: "What types of services and programs will be offered through this task force?". However, these kinds of stories were less exciting to many of the journalists who instead, preferred investigations that probe what the administration "does not want revealed", as P4 stated. This reasoning likely led to *areas to investigate* have the lowest average usefulness of the GPT-3 completions incorporated in AngleKindling (Table 5.3). Meanwhile, more investigative story ideas stemmed from the potential sources of *controversy* and *negative outcomes*. For example, P12 pointed to the *controversy* that "AT Mitchell may not be qualified to lead the task force" as a potential **next-day or week long story**. She explained that, over the course of a few days, she would research the communities that would be most affected by the new gun violence prevention measures Mayor Adams enacted and then interview organizations or prominent members of those communities to get their take on AT Mitchell. Finally, P8 pointed to a *negative outcome* that could potentially become a **months-long retrospective story**: "The task force could be used to unfairly target communities of color". She explained that after a few months after the press release was distributed, she might gather some data on who was arrested and where police were being stationed to gauge if this *negative outcome* had come into fruition. Overall, AngleKindling's different angles lent themselves to multiple types of stories, from shorter, summarization pieces to longer, deeper investigations.

5.7.3 AngleKindling helped journalists deeply engage with the press release quickly.

Many participants noted that the press releases from Mayor Adams' administration were complex and filled with unnecessary details that diverted their attention from important information. The main points (D1) and the highlighted related content (D3), which participants rated on average 6.17 and 5.67 respectively (Table 5.3), helped them quickly skim and understand the press release, despite this distracting fluff. Participants predominantly used the main points not to replace the press release but to supplement their reading of it. One common strategy they employed was to use the main points as a reading guide: they first scanned the main points to get a (1) high level view of the claims and (2) a quick idea of the information they found interesting, then read the press release in its entirety. As well as guiding their reading, participants also used the main points as a quick reminder of the press release's content as they thought of story ideas. After reading the press release, and throughout his brainstorming process, P10 would reread the main points to regain a "holistic view" of the press release as he assessed potential sources of *controversy* and *negative outcomes*. By doing so, he could better contextualize and make sense of each potential angle. However, the main points were not perfect. P12, P5, and P8 mentioned that the main points missed information that they were very interested in from the press release, particularly the specific implementation details and statistics included in the document. These concrete details are very useful for critically examining the feasibility of the plans mentioned in press releases. Thus, the main points helped participants quickly understand the press release, but at the same time, can be improved to prioritize the statistics mentioned in the document.

Highlighting related content was critical to helping journalists trust both the main points and generated angles. As P2 explains, "The highlighted text is useful. It takes me there right to it... I would not trust these main points without the highlighted text." The related-content button and highlighted text helped the participants quickly verify the veracity of each main point, and without this feature, they would be concerned that the main points might be erroneous or misleading. The related content button also helped journalists acquire evidence to better understand a potential *controversy's* source. While working on the gun violence press release, P12 came upon a *controversy* that was completely unexpected: “There might be infighting among the various agencies involved”. She did not immediately understand why this *controversy* might be related, so she used the related content button to scrub the press release and was taken to a claim in the text that explained the new gun violence task force would work with multiple agencies, including the departments of health, social services, and housing. Being able to verify these *controversies* enabled journalists to both (1) better trust AngleKindling and (2) find interesting information they had not previously considered in the press release.

5.7.4 AngleKindling provided *contextualized* historical background, which helped with brainstorming story ideas.

Connecting a prior news article to an angle helped journalists better understand how the article was related and how it could be applied to inspire new story ideas. For example, for the zoning press release, P4 selected a potential *negative outcome* that stated, “The increased housing opportunities might not be affordable for low and middle-income New Yorkers.” The connected news article was entitled “Some ‘Affordable’ Units Too Costly, Report Says” and detailed how new affordable homes being built in the Bronx required household incomes above the median in New York City. The combination of this angle and news article inspired the potential idea of comparing the new plan with this past attempt to create affordable housing, to answer questions like: Does Eric Adam’s plan avoid the pitfalls of past plans? Are these more empty promises? However, sometimes articles did not provide useful background because they were tenuously connected to the angle. For example, in the gun violence task force press release, the *negative outcome*: “The recommendations of the task force might not be implemented properly” was connected to an article about who is on U.S. Coronavirus Task Force. P2 stated this could be potentially interesting toward a broad story on the general effectiveness of task forces, but ultimately found this article less useful because it was describing a federal task force instead of a city task force, created for a very different problem. Overall, when the news articles were closely related to the angle, participants were able to get relevant background information that sparked new ideas.

Meanwhile, in the INJECT condition, the separation of news articles into causal, quantifiable, and ramification angles was not very conducive to coming up with story ideas. Participants had trouble discerning why a certain article belonged to one of the categories, especially the causal and ramification groups. P4 stated, “I did not really get the causal or ramification angles. This information didn’t come through the article headlines”. He was unsure of how the article “A Pediatrician’s View on Gun Violence and Children” belonged to the causal category; it was not immediately clear what background or causes this article would reference. However, most participants appreciated the quantifiable category, aligning with the findings from INJECT’s own evaluation [134]. Among INJECT’s features, the quantifiable articles were the highest rated, receiving an average score of 4.83, compared to 3.83 for the causal articles and 3.75 for ramifications articles (Table 5.3). The articles within the quantifiable category, were more explainable, often containing a statistic in their headline or lead paragraph, like: “It has been nearly a quarter century since New York City experienced as much gun violence in the month of June as it has seen this year.” The quantifiable articles also provided inspiration on potential datasets to use or analyses to conduct for the press release. P11 stated, “I really like the quantifiable angles, they include numbers and even trends that help give me context for my story.” Finally, the participants also appreciated that the articles appeared together in longer lists, which helped provide great coverage of the topic as a whole. P9 explained, “[INJECT] is a bit broader. It helps me better understand the topic...this a great tool for background information.” INJECT’s list-organization, while not immediately useful for brainstorming ideas, helped participants better learn about the topic as a whole.

Finally, for both systems, participants wanted more contextual information, beyond historical news articles. Many participants mentioned that their goal is to go from the source material to interviewing relevant people and organizations as fast as possible. P4 explained, “The best story ideas will come from people who are smarter than me on the topic.” He wanted both tools to go beyond providing angles and provide *local* organizations, leaders, and experts to interview. From these interviews, journalists can identify the most important questions to answer in a story. While INJECT extracted people and organizations for its related articles, often the extracted individuals were too famous to easily interview or not related or local enough to the press release. As well as individuals to interview, P6 who has a background in law, wanted excerpts of relevant laws brought into each tool. For the zoning press release, P6 wanted a list of each new update to the zoning policy in New York City. She specializes in month-long investigative stories, and incorporating this kind of context would greatly benefit that work. Thus, participants wanted more information pulled into these tools to (1) help them get to interviewing faster and (2) have a deeper understanding of the topic.

5.8 Discussion

In the following section we discuss a few areas of future work, including enabling journalists to write their own LLM-prompts to customize angle exploration, helping journalists prioritize angles given their time constraints and the likelihood of an angle actually yielding a story, and applying LLMs to read between the lines of other source material, like case law and academic papers. Finally, we end by discussing the limitations of this work.

5.8.1 Customizing the LLM angle suggestions

Currently, AngleKindling includes a pre-defined set of angles: *controversies*, *negative outcomes*, and *areas to investigate*, but participants expressed interest in customizing AngleKindling to suggest angles that better aligned with their own and their editor's interests. Our user study provides additional evidence for the need of personalization in computational tools for journalists [151]. P4 stated that he normally likes to write stories about "finance or the economy" and that being able to "push the angles in that direction" would be really useful. One way to help journalists personalize their angles could be to help them write their own LLM prompts. However, there are many challenges novices face when writing LLM prompts, including (1) phrasing the prompt so that it best fulfills the task, (2) providing a diverse set of training examples, and (3) scoping the prompt so that it does not ask for too much in one completion [166] [161]. A first step toward helping journalists customize the LLM-prompts could be having them write their own angles as they read press releases. For example, P4 could record financial-impact angles they thought of as they read, as well as highlight the portion of the press release that inspired each angle. These training samples could then be used as examples for a few-shot prompt, similar to the one shown in Figure 5.2, which could generate financial angles for new press releases. Past work has shown that with support, novices can write their own prompts to create simple AI-applications [159] [160], but has so far been only studied with UX designers and product managers. Future work can examine what specific challenges professional journalists face when writing their own prompts and how to best support them.

Helping journalists write their own prompts could also help them better understand how the system creates its angle suggestions. While using AngleKindling, P7 and P4 both explained that they might trust the suggestions more if they understood how they were generated. P4 was concerned that the angle suggestions might bias him toward certain types of stories and was worried that by spending time examining suggestions, he might be blinded to other angles he might have come up with on his own. Future work can address if letting journalists write their own prompts and familiarizing themselves with LLMs alleviates or exacerbates these anxieties. Perhaps by writing their own prompts, journalists feel they more thoroughly and naturally explore the space of angles, or alternatively, they might realize the LLM's limitations and trust it less as a source for angles.

5.8.2 Prioritizing different angles based on journalistic constraints

In addition to helping journalists personalize angles that better match their own or their editor's interests, AngleKindling can also support journalists in prioritizing which of these angles to pursue, based on time-constraints or evidence. As explained in the user study findings, AngleKindling provides angles that lend themselves to different types of stories, including day-of, next-day, and months-later. Instead of organizing angles by their type, such as *controversy* or *negative outcome*, they could be organized by how much time and work they would take to fulfill. For journalists with just a day to produce a story, AngleKindling could prioritize angles that can be fulfilled quickly, like public reactions and summarization pieces, instead of more investigative stories that require a deep dive into past legislation or interviews with experts. Another potential avenue is to prioritize angles that are more likely to pique reader interest; this was a feature that P6, P10, and P11 explicitly mentioned would be really useful in a system like AngleKindling. Even with a custom LLM-prompt producing angles that are more aligned with their reader's interests, AngleKindling could support highlighting the most interesting ones from this set. If AngleKindling was deployed in a newsroom, one interesting direction to take is to use click-through rates for articles to train a classifier that could identify which angles would lead to stories their readers might most be interested in. Thus, one rich area for future research is helping journalists sort the system's generated angles to satisfy these important constraints.

As well as sorting angles by projected reader interest and required effort, another concern participants had was determining which angles would actually lead to interesting stories. While P7 thought the *controversies* presented interesting ideas, she said it would be difficult to choose which ones to conduct follow-up research for in practice. For the zoning press release, she pointed at the *controversy*: "The plan could lead to gentrification", and asked, "Why are you feeding me that angle? Why would I go down that route? That would take a lot of time to verify that route". While she thought that gentrification was a potential outcome of the new zoning policies, she had no conception of how *likely* this was the case or if there was any recent evidence that could support this angle. Meanwhile, the provided historical background was a 2015 news article, which provided evidence that past zoning plans did not include enough affordable housing. However, this information was too old and did not help her assess the new plan. Thus, another line of future work could involve understanding how to best gather initial evidence for angles, so that journalists can quickly see which are most viable. Past work has shown user generated content, such as comments and posts on social media, can be filtered to help journalists find sources and information for their stories [148]. A similar strategy can be used to help provide evidence, such as recent posts or users to interview from social media platforms, for angles. Overall, beyond helping journalists realize the many stories that can be written from a press release, future work can investigate how to help prioritize angles that already have evidence to support them.

5.8.3 The potential for LLMs to read between the lines

Our evaluation of AngleKindling provides initial evidence that LLMs can identify the hidden implications of a source text. These implications were sometimes completely unexpected and appreciated by professional journalists, like "There might be infighting among the various agencies involved" and "The plan could lead to more traffic and congestion in New York City." While we applied LLMs to read between the lines for press releases, they can be applied to many other domains where reporters may ground a story in a specific document, such as law and academia. Case law, for instance, would be an interesting source of documents to assess the capabilities of an LLM for unveiling implications. Each case consists of a lengthy reasoning portion that incorporates the relevant circumstances and facts as well as relevant prior law to explain the court's decision. An LLM can be applied to dissect the court's argument and generate implications on (1) how this reasoning might affect the outcomes of similar disputes and more broadly, (2) how this decision might affect our lives. In addition to case law, LLMs can also be applied to unearth the implications of findings in academic papers. LLMs are already being used to help those without a scientific background better understand papers [167] by summarizing findings. But this can be taken a step further to help the authors of these papers explore the ethical implications of their findings, implications for other fields of research, and implications for our daily lives. These are ripe domains for future work in understanding if and how LLMs can help us unearth the implicit connections within a source text.

5.8.4 Limitations

While we carefully designed our study, it is not without limitations. First, our implementation of INJECT does not include all of its features, specifically the ability to search over a large corpus of news articles from a variety of different sources. While we did not include this feature, we do believe that our implementation was enough to compare the two broader strategies of both tools, which for INJECT is providing relevant articles organized by their angle type. That being said, there is the possibility that being able to search over a larger corpus might have improved participants' perceptions of INJECT. Though from the qualitative results we don't think this is the case; participants preferred AngleKindling because it was more "proactive" and provided concrete ideas as opposed to only news articles.

The next set of limitations pertains to our participants. We recruited professional journalists across the United States but had them all come up with angles for press releases from the New York City Mayor's office. This means that many participants lacked the additional context about New York, its prominent politicians, and its history when coming up with angles. However, this is not an unrealistic scenario, as many journalists are plunged into a new area's politics and history when they move or have recently started their career. Future work can examine how useful tools like AngleKindling are for journalists reading press releases that are well within their beat and expertise. Finally, we also only included journalists from the United States, whereas journalists in other countries might have different opinions on what kind of angles they value and how they prefer to be supported when they come up with story ideas.

Lastly, this work does not include a formal analysis for how well or how often GPT-3 can create insightful angles. Our evaluation shows initial promise and evidence that LLMs are capable of helping individuals read between the lines of a source text. Future work can conduct a more rigorous assessment of how well LLMs perform at this task. Finally, LLMs are fundamentally limited by their training data. GPT-3 might not function equally well across beats of science, politics and local news. Perhaps LLMs can be fine-tuned or at least have their prompts tuned [168] to support different beats. Next, publicly available LLMs are often not up to date with the latest news; GPT-3 only has world knowledge up until 2021, limiting its ability to generate angles on very recent events. LLMs also reflect the bias in their training data [169] and future work can elucidate if and how this bias bleeds into the story brainstorming process.

5.9 Conclusion

Informed by a three-month long co-design, we created AngleKindling, an interactive web tool which employs a LLM to help journalists come up with angles for a press release. We conducted a within-subjects study with 12 professional journalists, comparing AngleKindling to a very relevant and recent creativity support tool for journalists, INJECT. We found that AngleKindling was perceived to be significantly more helpful for coming up with ideas, with significantly less mental demand. This was primarily due to AngleKindling (1) helping journalists recognize angles they had not considered, (2) providing angles that were useful for multiple types of stories, (3) helping journalists quickly and deeply engage with the press release, and (4) providing contextualized historical context. Future work can explore how creating their own LLM-prompts might help journalists customize angle exploration and affect their trust of the system, how we might best help journalists recognize the most viable angles within their time-limit, and how LLMs can be used to read between the lines of other source material, like case law and academic papers.

Chapter 6: Conclusion and Future Work

This dissertation illustrates how to design exploratory search systems that better stimulate our memory. In this final chapter, we restate the contributions made by the three systems and discuss future work.

6.1 Restatement of Contributions

Toward improving the design of exploratory search systems so that they better help us learn with less cognitive load, this thesis presents three systems that embody three strategies for stimulating our memory. The three strategies include: (1) constructing an **association network** for the overview, so that it mimics our memory's structure and helps users explicitly relate information, (2) incorporating the user's **prior knowledge** into this overview, so that new information sticks better to what they already know, and (3) **concretizing** abstract information so that we can better integrate abstract knowledge with our current knowledge. The contributions of this thesis are as follows:

Concepts and Techniques

- Three design strategies for stimulating memory in exploratory search: (1) association network, (2) prior knowledge, and (3) concreteness.
- Incrementally generating an association network from an item in a knowledge graph and using network centrality and properties to create an organized view of the data.
- Using a few-shot LLM-prompt to construct a search space given a document.
- Sorting items within clusters by concreteness and relevance to help users quickly make sense of abstract information.

Artifacts

- SymbolFinder, a system which helps novice graphic designers *explore* visual symbols for abstract concepts.
- TastePaths, a system which helps music listeners *explore* and find songs to listen to in a genre.
- AngleKindling, a system which helps journalists *explore* story angles for a press release.

Experimental Results

- Three formative studies which illustrate that users have trouble exploring the diverse elements of an information space because of fixation and the limits of their memory.
- A comparative user study with 10 novice designers, which demonstrates that SymbolFinder helps users find 50% more symbols with significantly less mental demand and effort. This result supports that an association network and concreteness helped users remember and explore the diverse meanings associated with an abstract concept.
- A study with two versions of TastePaths (with and with-out prior knowledge) demonstrating that prior knowledge is very useful for exploring and understanding an overview.
- A study illustrating that AngleKindling was perceived to be significantly more helpful for thinking of story ideas with less mental demand than a prior journalistic angle-ideation tool that provides less concrete suggestions.

6.2 Future Work

There are three broad avenues of future work that stem from this dissertation: (1) further exploring the effectiveness of the three memory strategies, (2) building exploratory search systems that use LLMs instead of task-specific datasets, and finally (3) enabling end-users to build their own exploratory search systems.

6.2.1 Further exploring the three memory strategies

Expanding association networks

Each system in this dissertation constructs and organizes an association network to help users learn about and explore an information space. Two of these systems, SymbolFinder and TastePaths, rely on human-curated datasets to construct their association networks: SymbolFinder uses a word association dataset, SWOW [61], and TastePaths uses the Spotify Knowledge Graph. While these datasets are high quality, they are expensive to make and incomplete. For example, SWOW is missing many pop-culture associations, such as movies and TV shows, and at the same time, the dataset is missing concepts that emerged after its creation, such as COVID-19. Similarly, Spotify's Knowledge Graph is (1) missing countless brand new artists that are emerging everyday across many different music platforms, as well as (2) artists that are lesser known to western populaces. Thus, to make these systems more complete and enable exploratory search across all concepts and sub-genres, we need to develop automatic methods to expand these association networks.

One method for expanding association networks is to use the current network as training data and scan other databases to extract new items with similar properties. For example, consider the concept *summer* in SWOW and its association *watermelon*. From some large corpus of text, such as Wikipedia, we could find documents that include both the concept *summer* and each of its associations. We could then take the text surrounding the concept and association in each document, and perhaps some extracted linguistic features, as training data for a classifier. The goal for this classifier is to encode the textual relationship between concepts and their associations in text. Perhaps the text surrounding *summer* and *watermelon* encodes some linguistic relationship that can be used to find other associations. Then, given a new concept like *Star Wars*, this classifier can be used to crawl a large corpus and find associations that have a similar relationship with the new concept, like *lightsaber*. One challenge with creating such a classifier is that even with a lot of training data and a sophisticated model, it might still suggest incorrect associations. SymbolFinder's clusters captured the different meanings of each concept so well because of the quality of the underlying associations. Thus, in addition to supplying new associations automatically, future work can also investigate how to help users quickly inspect and edit them to ensure quality. Such a tool will help expand these networks and enable exploratory search over long-tail items.

Including collective prior knowledge

From TastePath's qualitative study, we learned that incorporating the user's prior knowledge is incredibly valuable for helping them make sense of new information. The participants of this study were Spotify users with extensive listening experience on the application. From their prior listening interactions, we were able to understand their preferences and determine which artists to serve as their anchors as they explored. However, sometimes the user's prior knowledge is incomplete or not good enough to help them navigate the information space. For example, journalists have to consider multiple perspectives when choosing a story angle for a particular story. As well as their own interests, journalists have to consider their editor's preferences for angles as well as their general audience's interests. Therefore, prior knowledge in this case should be *collective* and encapsulate the preferences of these multiple parties to help journalists choose the best angles. The same idea also applies to SymbolFinder. Symbols, as a form of visual communication, should be selected so that they are recognizable to the most people. And thus, we should also incorporate collective prior knowledge in SymbolFinder, to highlight symbols that have commonly been used to represent abstract concepts. Overall, sometimes the prior knowledge incorporated in exploratory search systems needs to go beyond that of its immediate user.

One way to include *collective* prior knowledge is to incorporate the knowledge provided by the interactions of multiple users. For example, for AngleKindling, we could use the publication's website and analyze its click-through rates to determine the kinds of angles its audience prefers. We could take articles with high click-through rates and train a classifier to identify which angles best aligned with reader interest in AngleKindling. Similarly, for SymbolFinder, we could take existing repositories of ads and other visual messaging to identify common visual symbols used to represent abstract concepts, like a skull and crossbones for *dangerous*. By conducting this analysis among numerous abstract concepts, we could highlight words and images in SymbolFinder that are particularly iconic and easily-recognizable for each cluster. This way, users can create visual blends that are easier to understand. In conclusion, we could analyze external datasets, like website interactions and visual messaging datasets, to incorporate collective prior knowledge in exploratory search systems.

Satisfying secondary constraints while exploring concrete information

For all three systems, incorporating concrete examples helped users make sense of more abstract information. However, for each system, secondary constraints emerged as users explored concrete information. For example, as users explored images for concrete words in SymbolFinder, they wanted to find symbols with specific visual characteristics, like a transparent background, a cartoon depiction, or a circular silhouette. Similarly for TastePaths, when users found an artist they were interested in, they wanted to quickly explore the wide range of sounds produced by that artist. By default, TastePaths incorporates the artist's popular tracks as concrete examples, but these tracks often came from one popular album and users were not sure if they received a holistic perspective on the artist's music. Thus, a secondary constraint for these concrete, popular tracks was to ensure that they sampled varied tracks of the artist's discography. And finally, for AngleKindling, the journalists wanted to filter the provided concrete, background articles so that they were about a particular topic or city. Future exploratory search systems should not only incorporate concrete examples but enable end-users to sift through these examples in meaningful ways to fulfill secondary constraints for their particular task.

Anticipating the complete set of secondary constraints for a particular search task is not feasible, but we can help users flexibly fulfill secondary constraints by helping them construct simple classifiers to sort concrete items. For example, in SymbolFinder, we could enable users to construct simple training sets to build classifiers that identify symbols with a particular visual characteristic. A user interested in filtering for black and white images might first select a few they have seen as positive examples, in addition to colorful images as negative examples. Similarly for AngleKindling, journalists could select background articles that discuss a particular city or topic in length to train a text classifier to find similar articles. By incorporating functionality to help users meet their specific secondary constraints within concrete items, we can make exploratory search systems more flexible and useful for a variety of different goals.

6.2.2 LLMs as a general dataset for exploratory search systems

AngleKindling makes the first step toward exploratory search systems that do not rely on datasets pertaining to a particular task. For SymbolFinder and TastePaths, the choice of underlying dataset was both critical to the system's success but also fundamentally limited both systems. For example, while word-associations led to better clusters in SymbolFinder, the SWOW dataset was small and incomplete, lacking associations for pop-culture concepts, like *Star Wars*, and emerging concepts like *COVID-19*. Similarly, TastePaths relied on Spotify's artist knowledge graph, where the only association between artists was if they *shared listeners*, while there can be many more, like *influence* and *shared sonic characteristics*. Unlike these two systems, AngleKindling employed an LLM, which was trained on billions of documents, and therefore contains a general world knowledge that could be useful for many different exploratory search tasks. While we focused on *controversies*, the LLM could have been prompted to generate any kind of angle, from *sports*, *economics*, *immigration*, and more. And beyond journalistic angles, LLMs may be able to construct search spaces for other documents and domains like law and academia. Thus, one long-term line of future work lies in **examining the potential of LLMs as the general, foundational dataset for exploratory search systems**, where the first step is extending AngleKindling's techniques to other domains.

Extending AngleKindling's techniques to documents in law and academia

AngleKindling was able to read between the lines of a press release and construct a search space of angles for journalists. This ability of LLMs can be extended to other knowledge-work domains where people analyze documents to explore their consequences and implications, such as law and academia. Case law, for instance, would be an interesting source of documents to assess the capabilities of an LLM for unveiling implications. Each case consists of a lengthy reasoning portion that incorporates the relevant circumstances and facts as well as relevant prior law to explain the court's decision. An LLM can be applied to dissect the court's argument and generate implications on (1) how this reasoning might affect the outcomes of similar disputes and more broadly, (2) how this decision might affect our lives. In addition to case law, LLMs can also be applied to unearth the implications of findings in academic papers. LLMs are already being used to help those without a scientific background better understand papers by summarizing findings [167]. But this can be taken a step further to help the authors of these papers explore the ethical implications of their findings, implications for other fields of research, and implications for our daily lives. By extending AngleKindling to these related domains, we can gather additional evidence that LLMs are able to generally support many exploratory search tasks.

Rigorously evaluating how well LLMs generate implications from text sources

Journalists found some of AngleKindling's suggestions to be insightful and novel and others to be a bit generic and unhelpful. While building AngleKindling, we experimented with a few types of prompts, including (1) a zero-shot prompt which asked the model to provide a list of controversies, as well as (2) a few-shot prompt which provided three examples of controversies to the model. Through a few informal experiments, we found that the few-shot prompts produced more compelling angles and fewer generic ones. Future work can more thoroughly assess how to design better prompts that produce high-quality and diverse potential controversies (or implications in general) from a source text. Perhaps with 20 examples of controversies and employing a technique like prompt-tuning could lead to better angles, or alternatively, there could be some chain of prompts that performs better. Going further, LLMs might be better at producing implications for some domains more so than others, based on the distribution of content they were fed through their training data. Perhaps because of the proliferation of news and journalism on the web, LLMs are more adept at producing implications that align with journalistic angles, as opposed to implications from case law or machine learning literature. Therefore, future work can also rigorously assess the domains where LLMs are better able to generate implications for than others.

Making LLM generations more explainable

During AngleKindling's evaluation, journalists often wondered why the system was producing certain angles. For example, for the zoning press release, the LLM suggested that the plans for affordable housing could backfire and lead to gentrification. Was this suggestion due to the model having seen this type of effect before, perhaps in prior news articles it consumed in its training data? Or had the model simply learned to contradict the claims made by the press release? To help users understand and better evaluate LLM suggestions, future work can investigate how to identify particularly influential text in the training data that led to the LLM suggestion. Perhaps LLMs can cite their sources to give journalists reason to believe in the suggestion. As mentioned in the prior section, journalists often have little time to write a story and making LLM suggestions more explainable would lead to faster evaluations of the generated angles.

6.2.3 Toward end-users creating their own exploratory search systems with LLMs

In addition to being large repositories of general world knowledge, LLMs also make this knowledge readily accessible to those without programming expertise through natural language prompting. Users can specify the task they want the LLM-prompt to complete in English, like "summarize this paragraph" or "come up with controversies that could stem from this press release". Given this ability, novices could potentially create their own exploratory search systems like AngleKindling on their own. All three exploratory search systems discussed in this thesis required lengthy formative studies and co-design to understand the needs and pain-points of users for each exploratory task. By supporting novices in constructing their own such systems, we can (1) save the time and effort involved with computer scientists understanding the difficulties of end-users and (2) enable end-users to create exploratory search systems that better support their specific needs. Thus, a second long-term line of future work lies in **helping end-users create their own exploratory search**

systems with LLMs.

Customizing the LLM-prompts in AngleKindling

Our implementation of AngleKindling helped journalists uncover potential controversies that could emerge from press releases. But journalists have a variety of backgrounds and interests, and two journalists might pursue wildly different angles from the same press release, such as economics, sports, immigration, etc. One clear next step to help journalists tailor AngleKindling's suggested angles to their own interests is to help them customize the LLM-prompts. A possible way to support this is to have journalists create training samples for the types of angles they would like generated. Similar to the few-shot prompt used in AngleKindling, journalists could highlight sections of the press release and record the angles those sections inspired. These pairs of sections and angles could then serve as training examples.

Exploratory search for prompt-debugging

Beyond journalism, to help users construct their own exploratory search systems with LLMs, they will need to be able to write effective LLM-prompts. For example, if we were to recreate SymbolFinder but with an LLM, one prompt a user might write is "List 10 different broad associations of control". Or similarly for TastePaths, one prompt might be "List the sub-genres of Art-rock, with 10 representative bands for each one". Past work has shown that novices, individuals without machine learning experience, face many challenges when writing their own prompts [161] [166]. For example, a journalist might want to summarize a section of a press release without its positive bias, with the prompt, "Summarize the following section in unbiased language: [section]", but the LLM is neither pulling out the pieces of information the journalist thinks is most important nor removing the bias. An LLM-prompt can be modified in many ways to improve its performance, and novices do not have great intuition on which path to take. In this case, the journalist could do away with the natural language prompt, and provide the sentences he wants extracted and rewritten as training samples. Alternatively, he could keep the natural language prompt and edit its phrasing.

Perhaps “TLDR” is a better phrase to prime the model than “Summarize”. But novices do not have an intuitive understanding of the underlying training data from which the model was trained to optimize the natural language prompt. Finally another option is to decompose this prompt into two, simpler prompts. Perhaps the LLM should first extract important information in one prompt, then pass this result to another prompt which rewrites it in simpler language. Ultimately, the potential changes that can be made to improve an LLM-prompt forms quite a large search space. Future work can investigate how to help novices explore variations to their prompts to help make them more robust.

Helping users construct overviews from abstract inputs

A clear pattern from the three exploratory search systems illustrated in this thesis is to take an abstract piece of information (such as a concept, genre, or entire press release), and to construct an overview that enables users to drill down into progressively more concrete information. An exploratory search system generator could guide users in constructing this overview through a chain of LLM-prompts. The system could help users brainstorm the high-level areas of the information space for the given input, which in the case of SymbolFinder, is the abstract concept’s different meanings and contexts. From here, an LLM prompt can assist users in brainstorming more concrete subdivisions for each high-level area. In addition to helping users construct an overview, this exploratory search system generator would need to be able to handle different kinds of textual inputs, such as single concepts, or entire documents. At the same time, it would also need to visualize the overview as it’s being created to help users identify sections of the information space that are less complete and need to be broken down further. By guiding users in decomposing their problem into concrete subdivisions, we can help them form an overview from which they can locate their solution.

Helping users address secondary constraints in exploration

Exploratory search tasks often have emerging, secondary constraints. For SymbolFinder, these secondary constraints included the *tone* of the symbol and its *visual characteristics* (color, shading, background, etc). For AngleKindling, these constraints were the amount of *time and effort* required to substantiate an angle, as well as the *potential audience interest* for that angle. Toward creating a tool to help users make their own exploratory search system, there will need to be features that help users address these secondary constraints. This, in part, could also be addressed with LLMs. For example, during AngleKindling's evaluation, journalists were able to identify angles that aligned better with their editor's or audience's interests. Future systems could help journalists label these angles to train a few-shot LLM prompt that will then attempt to classify other angles that are similar.

6.3 Summary

Exploratory search is a crucial activity we do on a daily basis. ES involves gathering and making sense of information to learn about a complicated topic. Currently, ES is a cognitively taxing process, but by designing ES systems to better stimulate our memory, we can make learning easier and less mentally demanding. This thesis introduced three strategies to better stimulate memory: (1) building an *association network* overview, (2) incorporating the user's *prior knowledge* and (3) *concretizing* abstract information. Embodying these three strategies are three prototypes also introduced and examined by this thesis: *SymbolFinder*, *TastePaths*, and *AngleKindling*. Across these projects, we support exploration for three separate sets of users and domains: *SymbolFinder* helps graphic designers explore diverse symbols for abstract concepts, *TastePaths* helps music fans explore new artists within a genre, and *AngleKindling* helps journalists explore story angles given a press release. Looking toward thee future, we can investigate broadly (1) the potential of LLMs as the general, foundational dataset for exploratory search systems and (2) how to help end-users create their own exploratory search systems with LLMs.

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.,” Stanford InfoLab, Tech. Rep., 1999.

[2] G. Marchionini and R. W. White, “Information-seeking support systems [guest editors’ introduction],” *Computer*, vol. 42, no. 3, pp. 30–32, 2009.

[3] G. Marchionini, “Exploratory search: From finding to understanding,” *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.

[4] R. W. White and R. A. Roth, “Exploratory search: Beyond the query-response paradigm,” *Synthesis lectures on information concepts, retrieval, and services*, vol. 1, no. 1, pp. 1–98, 2009.

[5] P. Pirolli and S. Card, “Information foraging.,” *Psychological review*, vol. 106, no. 4, p. 643, 1999.

[6] S. K. Card *et al.*, “Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’01, Seattle, Washington, USA: Association for Computing Machinery, 2001, 498–505, ISBN: 1581133278.

[7] P. Pirolli and S. Card, “The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis,” in *Proceedings of international conference on intelligence analysis*, McLean, VA, USA, vol. 5, 2005, pp. 2–4.

[8] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, “The cost structure of sensemaking,” in *Proceedings of the INTERACT ’93 and CHI ’93 Conference on Human Factors in Computing Systems*, ser. CHI ’93, Amsterdam, The Netherlands: Association for Computing Machinery, 1993, 269–276, ISBN: 0897915755.

[9] J. R. Anderson, “A spreading activation theory of memory,” *Journal of verbal learning and verbal behavior*, vol. 22, no. 3, pp. 261–295, 1983.

[10] P. Pirolli, “Computational models of information scent-following in a very large browsable text collection,” in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 1997, pp. 3–10.

[11] H. A. Simon, “The structure of ill structured problems,” *Artificial intelligence*, vol. 4, no. 3-4, pp. 181–201, 1973.

[12] R. W. White and S. M. Drucker, “Investigating behavioral variability in web search,” in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW ’07, Banff, Alberta, Canada: Association for Computing Machinery, 2007, 21–30, ISBN: 9781595936547.

[13] G. Klein, B. Moon, and R. R. Hoffman, “Making sense of sensemaking 1: Alternative perspectives,” *IEEE intelligent systems*, vol. 21, no. 4, pp. 70–73, 2006.

[14] B. Dervin and M. Nilan, “Information needs and uses,” *Annual review of information science and technology*, vol. 21, pp. 3–33, 1986.

[15] J. R. Anderson and P. L. Pirolli, “Spread of activation.,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 10, no. 4, p. 791, 1984.

[16] I. Fischler, “Semantic facilitation without association in a lexical decision task,” *Memory & cognition*, vol. 5, no. 3, pp. 335–339, 1977.

[17] D. E. Meyer and R. W. Schvaneveldt, “Meaning, memory structure, and mental processes,” *Science*, vol. 192, no. 4234, pp. 27–33, 1976.

[18] J. H. Neely, “Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention.,” *Journal of experimental psychology: general*, vol. 106, no. 3, p. 226, 1977.

[19] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, “Scatter/gather: A cluster-based approach to browsing large document collections,” *SIGIR Forum*, vol. 51, no. 2, 148–159, 2017.

[20] J. Xu and W. B. Croft, “Quary expansion using local and global document analysis,” in *Acm sigir forum*, ACM New York, NY, USA, vol. 51, 2017, pp. 168–175.

[21] C. Carpineto, R. de Mori, G. Romano, and B. Bigi, “An information-theoretic approach to automatic query expansion,” *ACM Trans. Inf. Syst.*, vol. 19, no. 1, 1–27, Jan. 2001.

[22] C.-K. Huang, L.-F. Chien, and Y.-J. Oyang, “Relevant term suggestion in interactive web search based on contextual information in query session logs,” *Journal of the American Society for Information Science and Technology*, vol. 54, no. 7, pp. 638–649, 2003.

[23] S. Andolina, K. Klouche, D. Cabral, T. Ruotsalo, and G. Jacucci, “Inspirationwall: Supporting idea generation through automatic information exploration,” in *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, ser. C&C ’15, Glasgow, United Kingdom: Association for Computing Machinery, 2015, 103–106, ISBN: 9781450335980.

[24] D. Joshi *et al.*, “Paragrab: A comprehensive architecture for web image management and multimodal querying,” in *VLDB*, vol. 6, 2006, pp. 1163–1166.

[25] Y. Shi, Y. Wang, Y. Qi, J. Chen, X. Xu, and K.-L. Ma, “Ideawall: Improving creative collaboration through combinatorial visual stimuli,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW ’17, Portland, Oregon, USA: Association for Computing Machinery, 2017, 594–603, ISBN: 9781450343350.

[26] H.-C. Wang, D. Cosley, and S. R. Fussell, “Idea expander: Supporting group brainstorming with conversationally triggered visual thinking stimuli,” in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW ’10, Savannah, Georgia, USA: Association for Computing Machinery, 2010, 103–106, ISBN: 9781605587950.

[27] M. Chang, L. V. Guillain, H. Jung, V. M. Hare, J. Kim, and M. Agrawala, “Recipescape: An interactive tool for analyzing cooking instructions at scale,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’18, Montreal QC, Canada: Association for Computing Machinery, 2018, 1–12, ISBN: 9781450356206.

[28] M. Brehmer, S. Ingram, J. Stray, and T. Munzner, “Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2271–2280, 2014.

[29] C. Felix, A. Dasgupta, and E. Bertini, “The exploratory labeling assistant: Mixed-initiative label curation with large document collections,” in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’18, Berlin, Germany: Association for Computing Machinery, 2018, 153–164, ISBN: 9781450359481.

[30] J. Peltonen, K. Belorustceva, and T. Ruotsalo, “Topic-relevance map: Visualization for improving search result comprehension,” in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, ser. IUI ’17, Limassol, Cyprus: Association for Computing Machinery, 2017, 611–622, ISBN: 9781450343480.

[31] D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos, “Apolo: Making sense of large network data by combining rich user interaction and machine learning,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2011, 167–176, ISBN: 9781450302289.

[32] R. Pienta *et al.*, “Vigor: Interactive visual exploration of graph query results,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 215–225, 2018.

[33] W. K. Horton, *The icon book: Visual symbols for computer systems and documentation*. John Wiley & Sons, Inc., 1994.

[34] R. S. Easterby, “The perception of symbols for machine displays,” *Ergonomics*, vol. 13, no. 1, pp. 149–158, 1970.

[35] N. Holmes and R. De Neve, *Designing pictorial symbols*. Rac Books, 1985.

[36] L. B. Chilton, S. Petridis, and M. Agrawala, “Visiblends: A flexible workflow for visual blends,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–14, ISBN: 9781450359702.

[37] C. Forceville, “Pictorial metaphor in advertisements,” *Metaphor and symbol*, vol. 9, no. 1, pp. 1–29, 1994.

[38] M. Nagasundaram and A. R. Dennis, “When a group is not a group: The cognitive foundation of group idea generation,” *Small Group Research*, vol. 24, no. 4, pp. 463–489, 1993.

[39] H.-C. Wang, S. R. Fussell, and D. Cosley, “From diversity to creativity: Stimulating group brainstorming with cultural differences and conversationally-retrieved pictures,” in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ser. CSCW ’11, Hangzhou, China: Association for Computing Machinery, 2011, 265–274, ISBN: 9781450305563.

[40] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, *Designing the user interface: strategies for effective human-computer interaction*. Pearson, 2016.

[41] A. I. of Graphic Arts, *Symbol Signs*. National Technical Information Service, 1974.

[42] A. Marcus, “Corporate identity for iconic interface design: The graphic design perspective,” *Interfaces in Computing*, vol. 2, no. 4, pp. 365–378, 1984.

[43] Z. Hussain *et al.*, “Automatic understanding of image and video advertisements,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1705–1715.

[44] K. Hemenway, “Psychological issues in the use of icons in command menus,” in *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, ser. CHI ’82, Gaithersburg, Maryland, USA: Association for Computing Machinery, 1982, 20–23, ISBN: 9781450373890.

[45] W. Huggins and D. R. Entwisle, *Iconic Communication: An Annoted Bibliography*. Johns Hopkins University Press, 1974.

[46] P. A. Kolers, “Some formal characteristics of pictograms,” *American scientist*, vol. 57, no. 3, pp. 348–363, 1969.

[47] M. A. Borkin *et al.*, “What makes a visualization memorable?” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.

[48] M. A. Borkin *et al.*, “Beyond memorability: Visualization recognition and recall,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 519–528, 2016.

[49] J. Xu and W. B. Croft, “Quary expansion using local and global document analysis,” *SIGIR Forum*, vol. 51, no. 2, 168–175, Aug. 2017.

[50] B. M. Fonseca, P. Golgher, B. Pôssas, B. Ribeiro-Neto, and N. Ziviani, “Concept-based interactive query expansion,” in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’05, Bremen, Germany: Association for Computing Machinery, 2005, 696–703, ISBN: 1595931406.

[51] Z.-J. Zha *et al.*, “Visual query suggestion: Towards capturing user intent in internet image search,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 3, Aug. 2010.

[52] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma, “Igroup: Web image search results clustering,” in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM ’06, Santa Barbara, CA, USA: Association for Computing Machinery, 2006, 377–384, ISBN: 1595934472.

[53] G. A. Miller, “Wordnet: A lexical database for english,” *Commun. ACM*, vol. 38, no. 11, 39–41, Nov. 1995.

[54] E. Hoque, G. Strong, O. Hoeber, and M. Gong, “Conceptual query expansion and visual search results exploration for web image retrieval,” in *Advances in Intelligent Web Mastering–3*, Springer, 2011, pp. 73–82.

[55] E. Hoque, O. Hoeber, and M. Gong, “Cider: Concept-based image diversification, exploration, and retrieval,” *Information Processing Management*, vol. 49, no. 5, pp. 1122 –1138, 2013.

[56] N. Zhao *et al.*, “Iconate: Automatic compound icon generation and ideation,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20, Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–13, ISBN: 9781450367080.

[57] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, “Hierarchical clustering of www image search results using visual, textual and link information,” in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA ’04, New York, NY, USA: Association for Computing Machinery, 2004, 952–959, ISBN: 1581138938.

[58] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, “Visual diversification of image search results,” in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW ’09, Madrid, Spain: Association for Computing Machinery, 2009, 341–350, ISBN: 9781605584874.

[59] J. Fan, Y. Gao, H. Luo, D. A. Keim, and Z. Li, “A novel approach to enable semantic and visual image summarization for exploratory image search,” in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ser. MIR ’08, Vancouver,

British Columbia, Canada: Association for Computing Machinery, 2008, 358–365, ISBN: 9781605583129.

[60] S. J. Mcdougall, M. B. Curry, and O. de Bruijn, "Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols," *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 3, pp. 487–519, 1999.

[61] S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, and G. Storms, "The "small world of words" english word association norms for over 12,000 cue words," *Behavior research methods*, vol. 51, no. 3, pp. 987–1006, 2019.

[62] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior research methods*, vol. 46, no. 3, pp. 904–911, 2014.

[63] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[64] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[65] Y. Shi, Y. Wang, Y. Qi, J. Chen, X. Xu, and K.-L. Ma, "Ideawall: Improving creative collaboration through combinatorial visual stimuli," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '17, Portland, Oregon, USA: Association for Computing Machinery, 2017, 594–603, ISBN: 9781450343350.

[66] J. Chan *et al.*, "Semantically far inspirations considered harmful? accounting for cognitive states in collaborative ideation," in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, ser. C&C '17, Singapore, Singapore: Association for Computing Machinery, 2017, 93–105, ISBN: 9781450344036.

[67] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066 111, 2004.

[68] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008.

[69] M. E. Newman, "The mathematics of networks," *The new palgrave encyclopedia of economics*, vol. 2, no. 2008, pp. 1–12, 2008.

[70] Z. S. Harris, “Distributional structure,” *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[71] M. Ono, M. Miwa, and Y. Sasaki, “Word embedding-based antonym detection using thesauri and distributional information,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 984–989.

[72] S. De Deyne, A. Perfors, and D. J. Navarro, “Predicting human similarity judgments with distributional models: The value of word associations.,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1861–1870.

[73] C. A. Gomez-Uribe and N. Hunt, “The netflix recommender system: Algorithms, business value, and innovation,” *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, Dec. 2016.

[74] F. Amat, A. Chandrashekar, T. Jebara, and J. Basilico, “Artwork personalization at netflix,” in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys ’18, Vancouver, British Columbia, Canada: Association for Computing Machinery, 2018, 487–488, ISBN: 9781450359016.

[75] J. McInerney *et al.*, “Explore, exploit, and explain: Personalizing explainable recommendations with bandits,” in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys ’18, 2018, pp. 31–39.

[76] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” *ACM Comput. Surv.*, vol. 52, no. 1, Feb. 2019.

[77] J. A. Konstan and J. Riedl, “Recommender systems: From algorithms to user experience,” *User modeling and user-adapted interaction*, vol. 22, no. 1, pp. 101–123, 2012.

[78] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, “Explaining the user experience of recommender systems,” *User Modeling and User-Adapted Interaction*, vol. 22, no. 4, pp. 441–504, 2012.

[79] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, “Exploring the filter bubble: The effect of using recommender systems on content diversity,” in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW ’14, Seoul, Korea: Association for Computing Machinery, 2014, 677–686, ISBN: 9781450327442.

[80] E. Pariser, *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[81] B. P. Knijnenburg and M. C. Willemsen, “Evaluating recommender systems with user experiments,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA: Springer US, 2015, pp. 309–352, ISBN: 978-1-4899-7637-6.

[82] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, ser. RecSys '11, Chicago, Illinois, USA: Association for Computing Machinery, 2011, 157–164, ISBN: 9781450306836.

[83] D. Jannach and G. Adomavicius, "Recommendations with a purpose," in *Proceedings of the 10th ACM Conference on Recommender Systems*, ser. RecSys '16, Boston, Massachusetts, USA: Association for Computing Machinery, 2016, 7–10, ISBN: 9781450340359.

[84] B. P. Knijnenburg, S. Sivakumar, and D. Wilkinson, "Recommender systems for self-actualization," in *Proceedings of the 10th ACM Conference on Recommender Systems*, ser. RecSys '16, Boston, Massachusetts, USA: Association for Computing Machinery, 2016, 11–14, ISBN: 9781450340359.

[85] M. Yeomans, A. Shah, S. Mullainathan, and J. Kleinberg, "Making sense of recommendations," *Journal of Behavioral Decision Making*, vol. 32, no. 4, pp. 403–414, 2019.

[86] Y. Liang and M. C. Willemsen, "Personalized recommendations for music genre exploration," in *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '19, Larnaca, Cyprus: Association for Computing Machinery, 2019, 276–284, ISBN: 9781450360210.

[87] M. Kamalzadeh, C. Kralj, T. Möller, and M. Sedlmair, "Tagflip: Active mobile music discovery with social tags," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ser. IUI '16, Sonoma, California, USA: Association for Computing Machinery, 2016, 19–30, ISBN: 9781450341370.

[88] S. Bostandjiev, J. O'Donovan, and T. Höllerer, "Tasteweights: A visual interactive hybrid recommender system," in *Proceedings of the Sixth ACM Conference on Recommender Systems*, ser. RecSys '12, Dublin, Ireland: Association for Computing Machinery, 2012, 35–42, ISBN: 9781450312707.

[89] I. Andjelkovic, D. Parra, and J. O'Donovan, "Moodplay: Interactive music recommendation based on artists' mood similarity," *International Journal of Human-Computer Studies*, vol. 121, pp. 142–159, 2019, Advances in Computer-Human Interaction for Recommender Systems.

[90] S. Nagulendra and J. Vassileva, "Understanding and controlling the filter bubble through interactive visualization: A user study," in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, ser. HT '14, Santiago, Chile: Association for Computing Machinery, 2014, 107–115, ISBN: 9781450329545.

[91] N. Tintarev, S. Rostami, and B. Smyth, "Knowing the unknown: Visualising consumption blind-spots in recommender systems," in *Proceedings of the 33rd Annual ACM Symposium*

on Applied Computing, ser. SAC '18, Pau, France: Association for Computing Machinery, 2018, 1396–1399, ISBN: 9781450351911.

[92] J. Kunkel, C. Schwenger, and J. Ziegler, "Newsviz: Depicting and controlling preference profiles using interactive treemaps in news recommender systems," in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '20, Genoa, Italy: Association for Computing Machinery, 2020, 126–135, ISBN: 9781450368612.

[93] MusicAlly. "Ifpi report reveals 7.4% growth in global recorded music revenues." (2021).

[94] E. Ruud, "Music and identity," *Nordisk Tidsskrift for Musikkterapi*, vol. 6, no. 1, pp. 3–13, 1997. eprint: `https://doi.org/10.1080/08098139709477889`.

[95] J. Frow, *Genre*. Routledge, 2014.

[96] M. J. Delsing, T. F. Ter Bogt, R. C. Engels, and W. H. Meeus, "Adolescents' music preferences and personality characteristics," *European Journal of Personality: Published for the European Association of Personality Psychology*, vol. 22, no. 2, pp. 109–130, 2008.

[97] F. Zuiderveen Borgesius, D. Trilling, J. Möller, B. Bodó, C. H. De Vreese, and N. Helberger, "Should we worry about filter bubbles?" *Internet Policy Review. Journal on Internet Regulation*, vol. 5, no. 1, 2016.

[98] J. Möller, D. Trilling, N. Helberger, and B. van Es, "Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity," *Information, Communication & Society*, vol. 21, no. 7, pp. 959–977, 2018. eprint: `https://doi.org/10.1080/1369118X.2018.1444076`.

[99] A. Laplante and J. S. Downie, "The utilitarian and hedonic outcomes of music information-seeking in everyday life," *Library & Information Science Research*, vol. 33, no. 3, pp. 202–210, 2011.

[100] J. H. Lee, H. Cho, and Y.-S. Kim, "Users' music information needs and behaviors: Design implications for music information retrieval systems," *Journal of the Association for Information Science and Technology*, vol. 67, no. 6, pp. 1301–1330, 2016. eprint: `https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23471`.

[101] J. H. Lee and N. M. Waterman, "Understanding user requirements for music information services.," in *ISMIR*, Citeseer, 2012, pp. 253–258.

[102] J. Garcia-Gathright, B. St. Thomas, C. Hosey, Z. Nazari, and F. Diaz, "Understanding and evaluating user satisfaction with music discovery," in *The 41st International ACM SIGIR Conference on Researc & Development in Information Retrieval*, ser. SIGIR '18, Ann Arbor, MI, USA: Association for Computing Machinery, 2018, 55–64, ISBN: 9781450356572.

[103] C. Hosey, L. Vujović, B. St. Thomas, J. Garcia-Gathright, and J. Thom, "Just give me what i want: How people use and evaluate music search," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, 1–12, ISBN: 9781450359702.

[104] M. Taramigkou, E. Bothos, K. Christidis, D. Apostolou, and G. Mentzas, "Escape the bubble: Guided exploration of music preferences for serendipity and novelty," in *Proceedings of the 7th ACM Conference on Recommender Systems*, ser. RecSys '13, Hong Kong, China: Association for Computing Machinery, 2013, 335–338, ISBN: 9781450324090.

[105] Y. Liang and M. C. Willemsen, "The role of preference consistency, defaults and musical expertise in users' exploration behavior in a genre exploration recommender," in *Fifteenth ACM Conference on Recommender Systems*, 2021, pp. 230–240.

[106] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert, "To explain or not to explain: The effects of personal characteristics when explaining music recommendations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19, Marina del Ray, California: Association for Computing Machinery, 2019, 397–407, ISBN: 9781450362726.

[107] Y. Liang and M. C. Willemsen, "Interactive music genre exploration with visualization and mood control," in *26th International Conference on Intelligent User Interfaces*, ser. IUI '21, College Station, TX, USA: Association for Computing Machinery, 2021, 175–185, ISBN: 9781450380171.

[108] J. Kunkel, B. Loepp, and J. Ziegler, "A 3d item space visualization for presenting and manipulating user preferences in collaborative filtering," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, ser. IUI '17, Limassol, Cyprus: Association for Computing Machinery, 2017, 3–15, ISBN: 9781450343480.

[109] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, 2301–2309, Dec. 2011.

[110] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[111] F. K. Hwang and D. S. Richards, "Steiner tree problems," *Networks*, vol. 22, no. 1, pp. 55–89, 1992.

[112] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001. eprint: `https://doi.org/10.1080/0022250X.2001.9990249`.

[113] W. Cai, Y. Jin, and L. Chen, "Critiquing for music exploration in conversational recommender systems," in *26th International Conference on Intelligent User Interfaces*, ser. IUI

’21, College Station, TX, USA: Association for Computing Machinery, 2021, 480–490, ISBN: 9781450380171.

[114] P. Wärnestål, “User evaluation of a conversational recommender system,” in *Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2005.

[115] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[116] J. Saldaña, *The coding manual for qualitative researchers*. Sage, 2015.

[117] P. Cremonesi, F. Garzottto, and R. Turrin, “User effort vs. accuracy in rating-based elicitation,” in *Proceedings of the Sixth ACM Conference on Recommender Systems*, ser. RecSys ’12, Dublin, Ireland: Association for Computing Machinery, 2012, 27–34, ISBN: 9781450312707.

[118] F. M. Harper, X. Li, Y. Chen, and J. A. Konstan, “An economic model of user rating in an online recommender system,” in *User Modeling 2005*, L. Ardissono, P. Brna, and A. Mitrovic, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 307–316, ISBN: 978-3-540-31878-1.

[119] E. I. Sparling and S. Sen, “Rating: How difficult is it?” In *Proceedings of the Fifth ACM Conference on Recommender Systems*, ser. RecSys ’11, Chicago, Illinois, USA: Association for Computing Machinery, 2011, 149–156, ISBN: 9781450306836.

[120] X. Amatriain, J. M. Pujol, N. Tintarev, and N. Oliver, “Rate it again: Increasing recommendation accuracy by user re-rating,” in *Proceedings of the Third ACM Conference on Recommender Systems*, ser. RecSys ’09, New York, New York, USA: Association for Computing Machinery, 2009, 173–180, ISBN: 9781605584355.

[121] G. Bonnin and D. Jannach, “Automated generation of music playlists: Survey and experiments,” *ACM Comput. Surv.*, vol. 47, no. 2, Nov. 2014.

[122] L. Delaney and L. K. Lades, “Present bias and everyday self-control failures: A day reconstruction study,” *Journal of Behavioral Decision Making*, vol. 30, no. 5, pp. 1157–1167, 2017. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.2031`.

[123] E. P. S. Baumer, R. Sun, and P. Schaedler, “Departing and returning: Sense of agency as an organizing concept for understanding social media non/use transitions,” *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, Nov. 2018.

[124] K. Lukoff *et al.*, “How the design of youtube influences user sense of agency,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2021, ISBN: 9781450380966.

[125] R. M. Entman, "Framing: Towards clarification of a fractured paradigm," *McQuail's reader in mass communication theory*, vol. 390, p. 397, 1993.

[126] D. H. Weaver, L. Willnat, and G. C. Wilhoit, "The american journalist in the digital age: Another look at us news people," *Journalism & Mass Communication Quarterly*, vol. 96, no. 1, pp. 101–130, 2019.

[127] N. Diakopoulos, "Computational news discovery: Towards design considerations for editorial orientation algorithms in journalism," *Digital Journalism*, vol. 8, no. 7, pp. 945–967, 2020.

[128] N. Thurman, "Computational journalism," in *The handbook of journalism studies*, Routledge, 2019, pp. 180–195.

[129] R. Bommasani *et al.*, *On the opportunities and risks of foundation models*, 2021.

[130] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[131] K. I. Gero, V. Liu, and L. Chilton, "Sparks: Inspiration for science writing using language models," in *Designing Interactive Systems Conference*, ser. DIS '22, Virtual Event, Australia: Association for Computing Machinery, 2022, 1002–1019, ISBN: 9781450393584.

[132] J. J. Y. Chung, W. Kim, K. M. Yoo, H. Lee, E. Adar, and M. Chang, "Talebrush: Visual sketching of story generation with pretrained language models," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391566.

[133] H. Örnebring, "Sourcing the news: Key issues in journalism-an innovative study of the israeli press," *Journalism and Mass Communication Quarterly*, vol. 87, no. 3/4, p. 682, 2010.

[134] N. Maiden *et al.*, "Making the news: Digital creativity support for journalists," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, Montreal QC, Canada: Association for Computing Machinery, 2018, 1–11, ISBN: 9781450356206.

[135] S. Cohen, J. T. Hamilton, and F. Turner, "Computational journalism," *Commun. ACM*, vol. 54, no. 10, 66–71, 2011.

[136] C. E. Smith, E. Nevarez, and H. Zhu, "Disseminating research news in hci: Perceived hazards, how-to's, and opportunities for innovation," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, 1–13, ISBN: 9781450367080.

[137] D. Trielli and N. Diakopoulos, “Search as news curator: The role of google in shaping attention to news information,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–15, ISBN: 9781450359702.

[138] F. Bentley, K. Quehl, J. Wirfs-Brock, and M. Bica, “Understanding online news behaviors,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–11, ISBN: 9781450359702.

[139] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran, “Falling for fake news: Investigating the consumption of news via social media,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’18, Montreal QC, Canada: Association for Computing Machinery, 2018, 1–10, ISBN: 9781450356206.

[140] C. Oh *et al.*, “Understanding user perception of automated news generation system,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, 1–13, ISBN: 9781450367080.

[141] N. Hassan, F. Arslan, C. Li, and M. Tremayne, “Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1803–1812.

[142] T. Alhindi, S. Petridis, and S. Muresan, “Where is your evidence: Improving fact-checking by justification modeling,” in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 85–90.

[143] C. Xia *et al.*, “Citybeat: Real-time social media visualization of hyper-local city data,” in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW ’14 Companion, Seoul, Korea: Association for Computing Machinery, 2014, 167–170, ISBN: 9781450327459.

[144] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, “Twitinfo: Aggregating and visualizing microblogs for event exploration,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’11, Vancouver, BC, Canada: Association for Computing Machinery, 2011, 227–236, ISBN: 9781450302289.

[145] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, “Diamonds in the rough: Social media visual analytics for journalistic inquiry,” in *2010 IEEE Symposium on Visual Analytics Science and Technology*, 2010, pp. 115–122.

[146] N. Diakopoulos, S. Goldenberg, and I. Essa, “Videolyzer: Quality analysis of online informational video for bloggers and journalists,” in *Proceedings of the SIGCHI Conference on*

Human Factors in Computing Systems, ser. CHI ’09, Boston, MA, USA: Association for Computing Machinery, 2009, 799–808, ISBN: 9781605582467.

[147] X. Liu, A. Nourbakhsh, Q. Li, S. Shah, R. Martin, and J. Duprey, “Reuters tracer: Toward automated news production using large scale social media data,” in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1483–1493.

[148] Y. Wang and N. Diakopoulos, “Journalistic source discovery: Supporting the identification of news sources in user generated content,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966.

[149] J. Stasko, C. Görg, and Z. Liu, “Jigsaw: Supporting investigative analysis through interactive visualization,” *Information visualization*, vol. 7, no. 2, pp. 118–132, 2008.

[150] N. Diakopoulos, M. De Choudhury, and M. Naaman, “Finding and assessing social media information sources in the context of journalism,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’12, Austin, Texas, USA: Association for Computing Machinery, 2012, 2451–2460, ISBN: 9781450310154.

[151] N. Diakopoulos, D. Trielli, and G. Lee, “Towards understanding and supporting journalistic practices using semi-automated news discovery tools,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, 2021.

[152] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, “Novice-ai music co-creation via ai-steering tools for deep generative models,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, 1–13, ISBN: 9781450367080.

[153] V. Liu, H. Qiao, and L. Chilton, “Opal: Multimodal image generation for news illustration,” 2022.

[154] V. Liu and L. B. Chilton, “Design guidelines for prompt engineering text-to-image generative models,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391573.

[155] H. Shakeri, C. Neustaedter, and S. DiPaola, “Saga: Collaborative storytelling with gpt-3,” in *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW ’21, Virtual Event, USA: Association for Computing Machinery, 2021, 163–166, ISBN: 9781450384797.

[156] H. Li, “Language models: Past, present, and future,” *Commun. ACM*, vol. 65, no. 7, 56–63, 2022.

[157] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, “Wordcraft: Story writing with large language models,” in *27th International Conference on Intelligent User Interfaces*, ser. IUI ’22, Helsinki, Finland: Association for Computing Machinery, 2022, 841–852, ISBN: 9781450391443.

[158] H. Osone, J.-L. Lu, and Y. Ochiai, “Buncho: Ai supported story co-creation via unsupervised multitask learning to increase writers’ creativity in japanese,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380959.

[159] T. Wu, M. Terry, and C. J. Cai, “Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391573.

[160] T. Wu *et al.*, “Promptchainer: Chaining large language model prompts through visual programming,” in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391566.

[161] E. Jiang *et al.*, “Promptmaker: Prompt-based prototyping with largenbsp;languagenbsp;models,” in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391566.

[162] Z. Reich, “The process model of news initiative,” *Journalism Studies*, vol. 7, no. 4, pp. 497–514, 2006.

[163] T. Harcup and D. O’Neill, “What is news?” *Journalism Studies*, vol. 18, no. 12, pp. 1470–1488, 2017. eprint: `https://doi.org/10.1080/1461670X.2016.1150193`.

[164] J. Boumans, “Subsidizing the news?” *Journalism Studies*, vol. 19, no. 15, pp. 2264–2282, 2018.

[165] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[166] E. Jiang *et al.*, “Discovering the syntax and strategies of natural language programming with generative language models,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391573.

[167] T. August, L. L. Wang, J. Bragg, M. A. Hearst, A. Head, and K. Lo, *Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing*, 2022.

[168] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059.

[169] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, “Towards understanding and mitigating social biases in language models,” in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 6565–6576.