COMS 6998-7 Spring 2025: "Theoretical Foundations of Large Language Models"

Instructor: Prof. Daniel Hsu

TAs: Jingwen Liu, Arman Ozcan

"Classical" Machine Learning (ML)

- Spam filtering (from email text)
- Ad click prediction (from user profile and context)
- Gene expression level prediction (from upstream DNA)
- …
- ML strategy:
  - Think about useful data that would be useful for prediction
    - "Feature engineering"
  - Think about how data is combined to make a good prediction
    - "Model selection" / picking a "hypothesis class"
  - Collect this data for many examples
  - Fit model to the examples
    - "Parameter estimation" / "learning"

Deep Learning (DL) circa 2012+

- Computer vision, speech recognition, AlphaGo, …
- DL strategy:
  - Use "raw" data (pixels, waveform, board state, …) as much as possible, with a multi-layer neural network that "learns" the useful features
  - Model now has many parameters
    - No real interpretation of any single parameter
- Compared to "classical" ML: may need a lot more data, but **performance ceiling is much higher**

ML/DL today

- Age of "pre-trained" models (a.k.a. "foundation" or "base" model)
- Strategies today:
  - "Fine-tuning":
    - Procure a base model
    - Represent your "raw" data in language "understood" by base model
    - Minimally adjust parameters of base model for your prediction task
  - "Prompting":
    - Procure a base model
    - Describe your task and provide examples --- in language "understood" by base model --- all as input to the base model
    - Execute base model on this input to get desired predictions
- Somehow the **base model already encodes many useful ways of combining data to make predictions**!

Language models

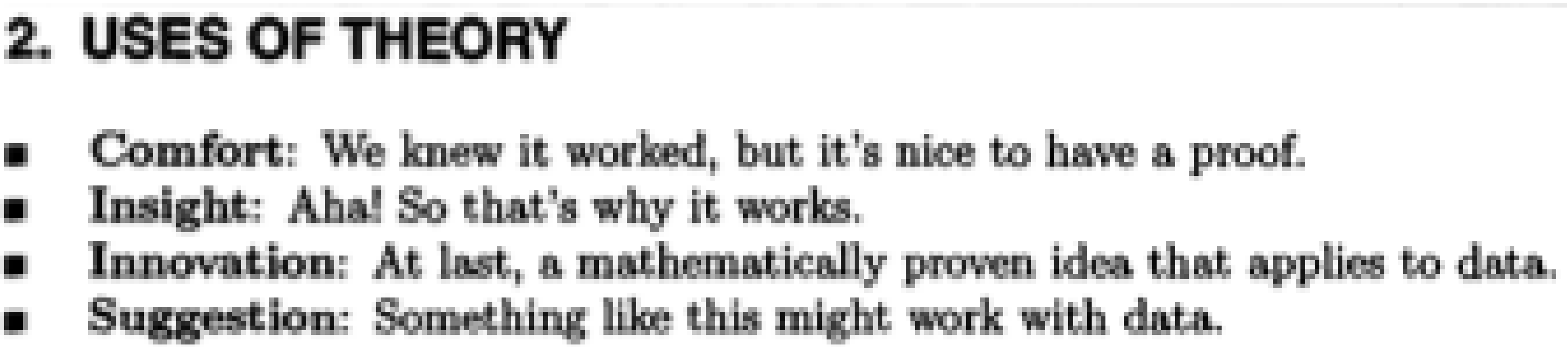
- Large Language Models are perhaps the most well-known base models (e.g., BERT, GPT-*, Llama)
    - Can be regarded as a type of multi-layer neural network
    - Pre-trained on huge amounts of language data – text from many sources (general web scrape, arXiv, GitHub, Wikipedia, StackExchange, …)
    - Fitting strategy is extremely simplistic
        - E.g., treat data as sequence $x_1, x_2, x_3, \ldots$; train model to predict x_{t+1} from $x_1, x_2, \ldots, x_t$
- Unclear why complex ways of combining information for other tasks (e.g., solve logic puzzles) should emerge from this kind of training

This class: Theoretical Foundations of LLMs

- Sorry, this class will not give any definitive answers
- "Theoretical foundations":
  - Important (mathematical) ideas that helped get us here
  - Ideas that **might be** useful for understanding of how LLMs work
- Theory of LLMs: future work ☺

Theory

Breiman (1995) “Reflections After Refereeing Papers for NIPS”



2. USES OF THEORY

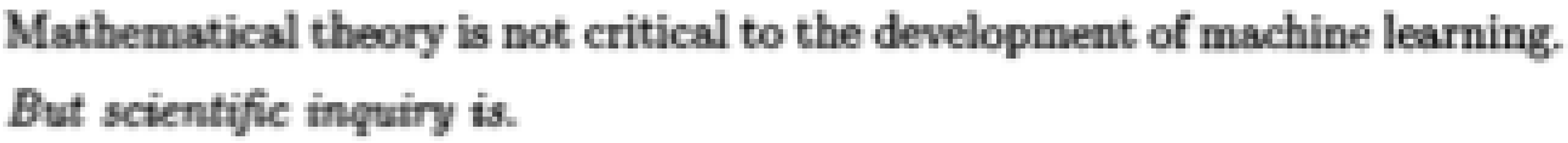
- **Comfort:** We knew it worked, but it's nice to have a proof.
- **Insight:** Aha! So that's why it works.
- **Innovation:** At last, a mathematically proven idea that applies to data.
- **Suggestion:** Something like this might work with data.

Breiman’s “Post World War II” Examples (in 1995):

- Asymptotic analyses of decision trees, nearest neighbor, universal approximation
- Nonparametric regression, sparsity in inverse problems
- Spectral analysis in time series, information theory, bootstrap
- Theory-inspired heuristics for function fitting

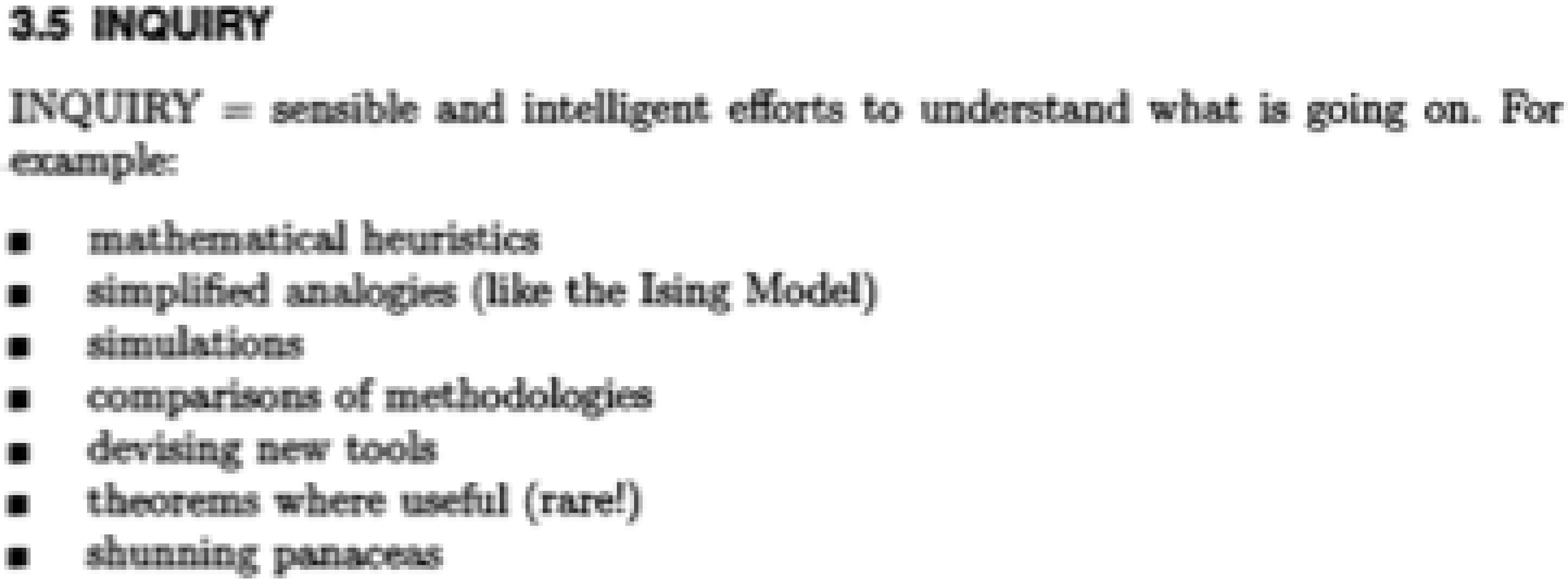
Inquiry

Breiman (1995) “Reflections After Refereeing Papers for NIPS”



Mathematical theory is not critical to the development of machine learning.

But scientific inquiry is.



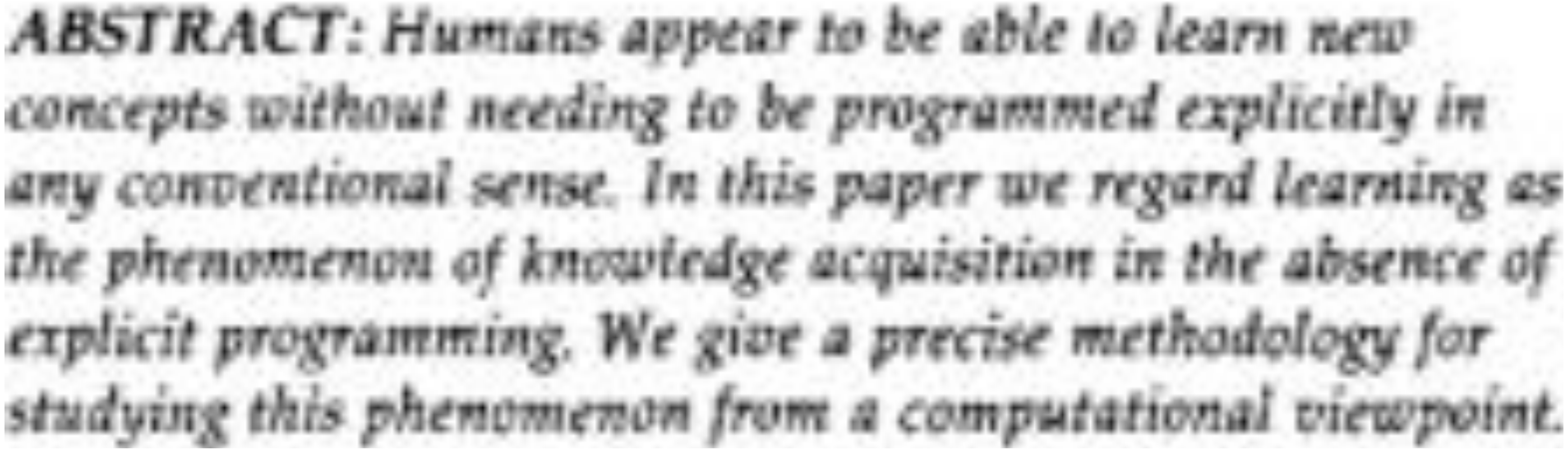
3.5 INQUIRY

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Explaining phenomena

Valiant (1984) “A Theory of the Learnable”



ABSTRACT: *Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming. We give a precise methodology for studying this phenomenon from a computational viewpoint.*

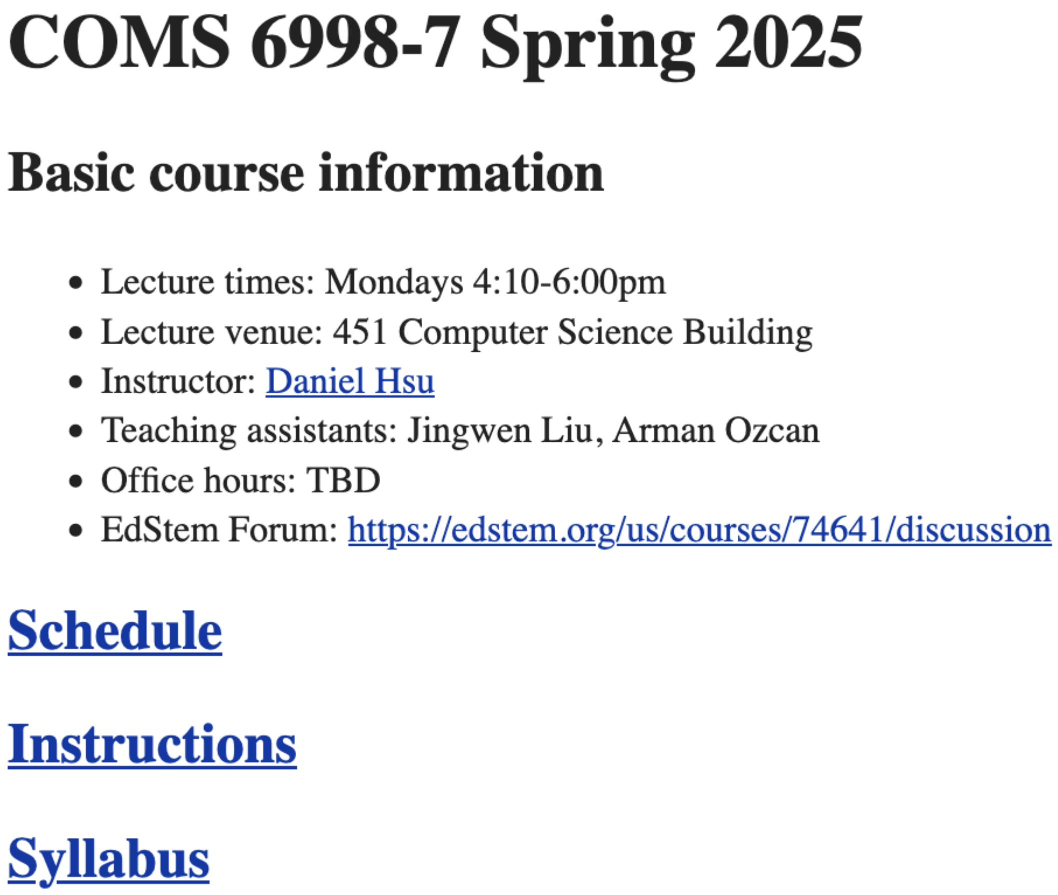
Advice

- Many ways to develop a theory of learning, LLMs, etc.
  - Mathematics is best tool for being precise
    - Theorem + proof = demonstration of understanding
  - May need to work at different levels of abstraction
- I think the ideas around LLMs are beautiful in their own right!

Class logistics

Course information

- Webpage: https://www.cs.columbia.edu/~djhsu/coms6998-s25/
  - Used to be on notion.io, but that was too hard to maintain!!!
- EdStem message board
- TAs: Jingwen and Arman
  - Will be helping you with your class presentations (more later)



COMS 6998-7 Spring 2025

Basic course information

- Lecture times: Mondays 4:10-6:00pm
- Lecture venue: 451 Computer Science Building
- Instructor: Daniel Hsu
- Teaching assistants: Jingwen Liu, Arman Ozcan
- Office hours: TBD
- EdStem Forum: https://edstem.org/us/courses/74641/discussion

Schedule

Instructions

Syllabus

Topics

- Some "classics" on language modeling and word embeddings
- "Neural" language and computation models
- Learning theory for neural computation
- Other topics "around" LLMs where there is some theory
  - Post-training
  - "Chain-of-thought" learning
  - Calibration
  - Compositional learning

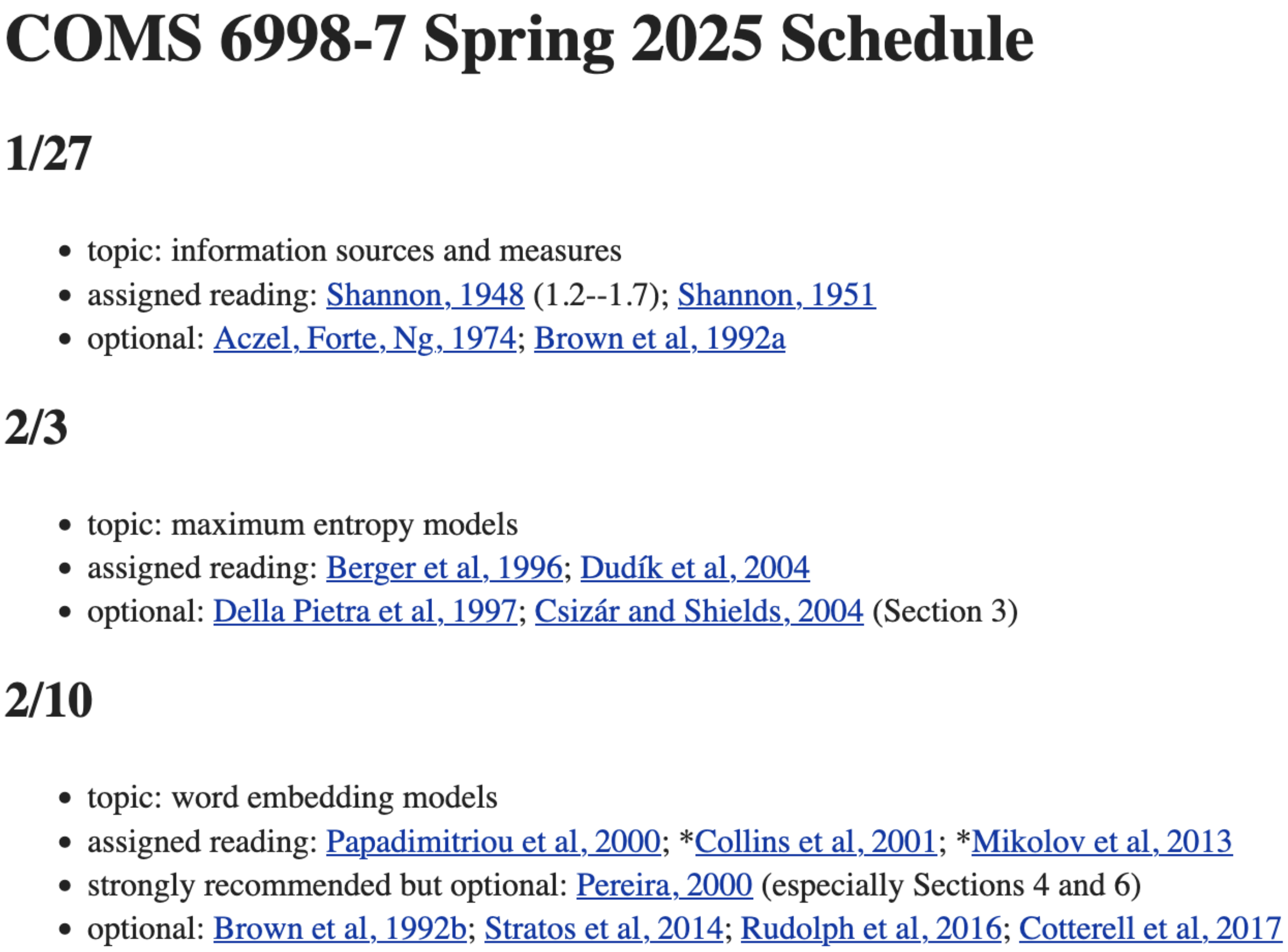
Selection of topics/papers heavily biased by what I know about and find interesting.
If you have suggestions, let me know!

How this class will operate

- Every week: read, reflect, respond to papers
- First couple weeks: some lectures from me
- Rest of semester: short presentations from students + discussion

Reading

https://www.cs.columbia.edu/~djhsu/coms6998-s25/schedule.html



COMS 6998-7 Spring 2025 Schedule

1/27

- topic: information sources and measures
- assigned reading: Shannon, 1948 (1.2--1.7); Shannon, 1951
- optional: Aczel, Forte, Ng, 1974; Brown et al, 1992a

2/3

- topic: maximum entropy models
- assigned reading: Berger et al, 1996; Dudík et al, 2004
- optional: Della Pietra et al, 1997; Csizár and Shields, 2004 (Section 3)

2/10

- topic: word embedding models
- assigned reading: Papadimitriou et al, 2000; *Collins et al, 2001; *Mikolov et al, 2013
- strongly recommended but optional: Pereira, 2000 (especially Sections 4 and 6)
- optional: Brown et al, 1992b; Stratos et al, 2014; Rudolph et al, 2016; Cotterell et al, 2017

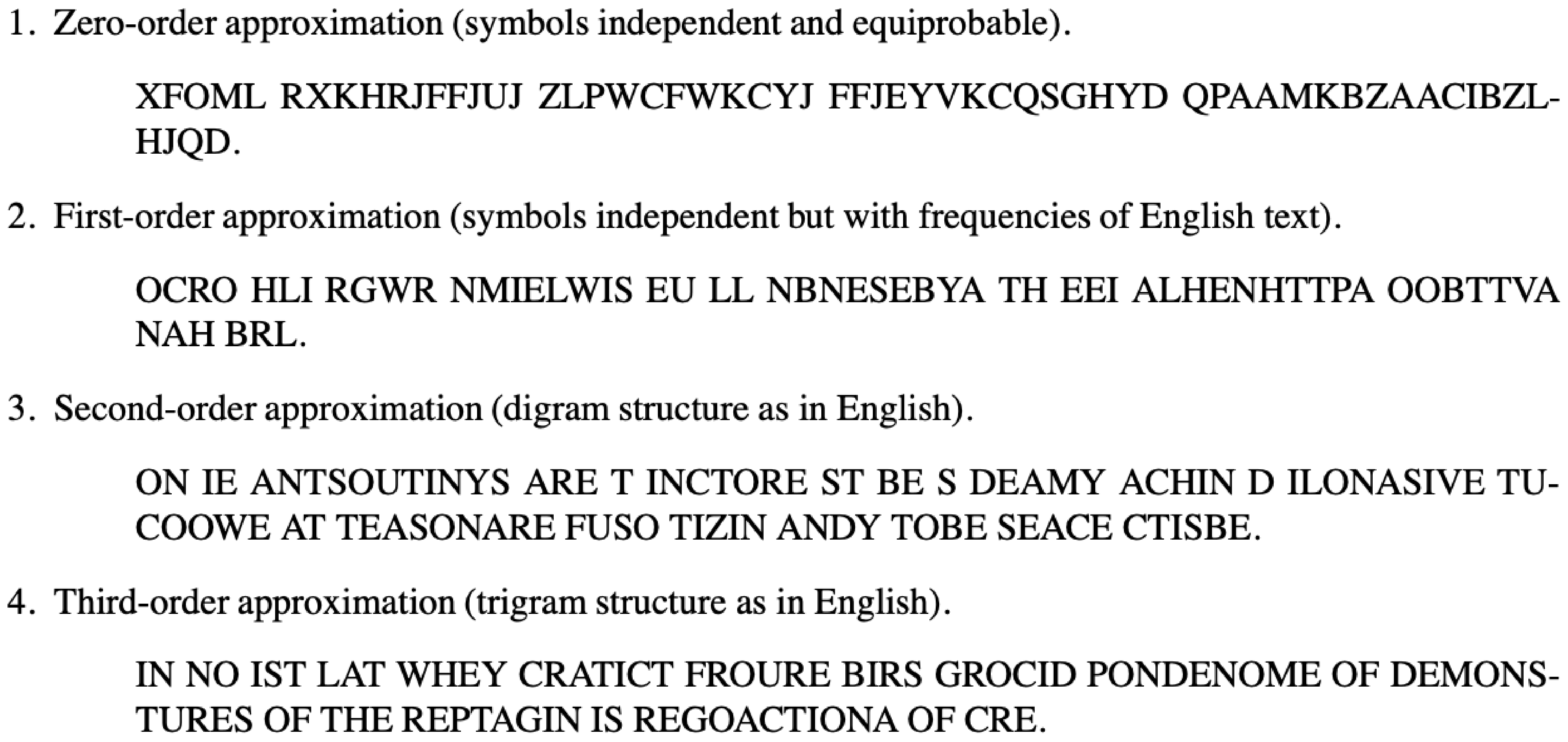
Paper presentations

- **Sign-up for paper presentations with TAs**
    - Shorter papers: 1 student
    - Longer papers: 2 students (but don't need both students to present)
        - Also prepare "scribe notes" / "handout", available ideally before the presentation
- **Goal of (20-25 min) presentation**: we've all read the paper …
    - Recapitulate key ideas of paper (& clear up points others found confusing)
    - Put paper in context of rest of the course so far
    - Present your own insights
    - Stimulate discussion
- **Meet with TAs and instructor to plan presentation**
- Rest of class: **give feedback to presenters every week** (Google Form)

Course project

- Engage with research on theoretical foundations of LLMs
  - Prove something new
  - Formulate a new model or way of thinking about something
  - Use theory to suggest a new experiment
  - Conduct a new experiment inspired by theory
- Project proposal before spring break (short, but think hard about it)
- Progress report (very short)
- Final report (~8 pages + references)

Shannon (1948)



1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.