Massachusetts Institute of Technology

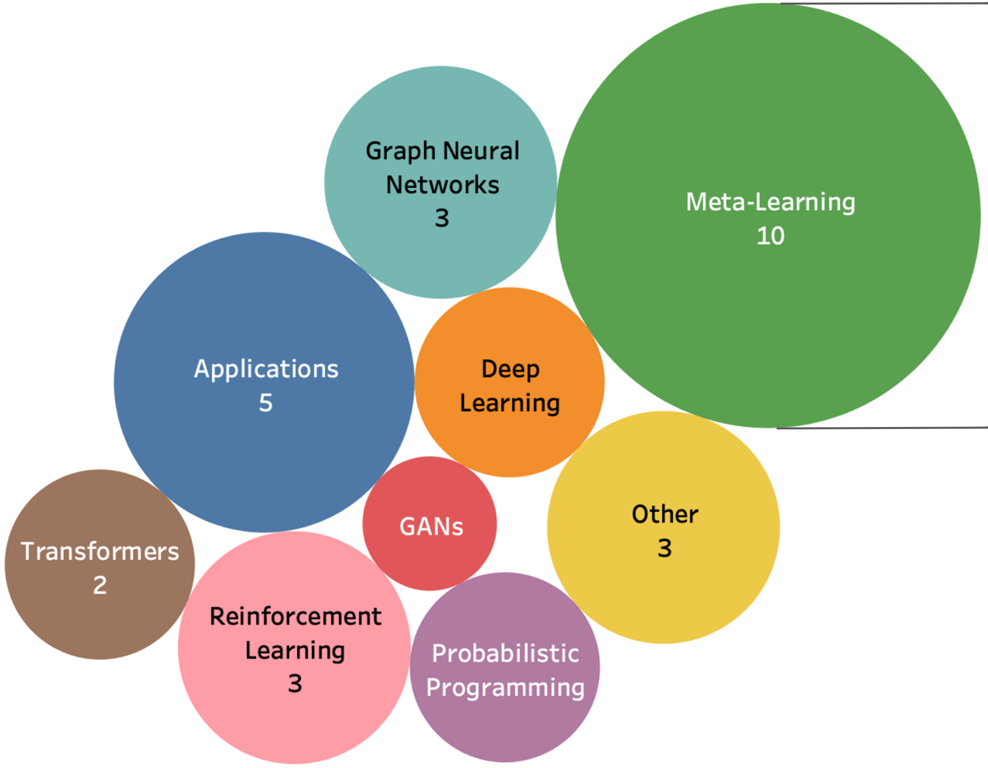
Meta Learning

MIT

Iddo Drori, Fall 2020

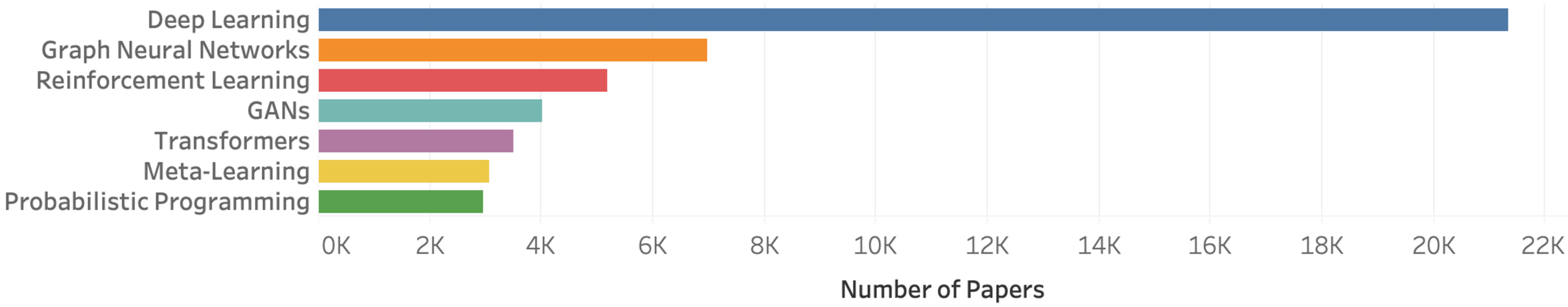
Massachusetts Institute of Technology

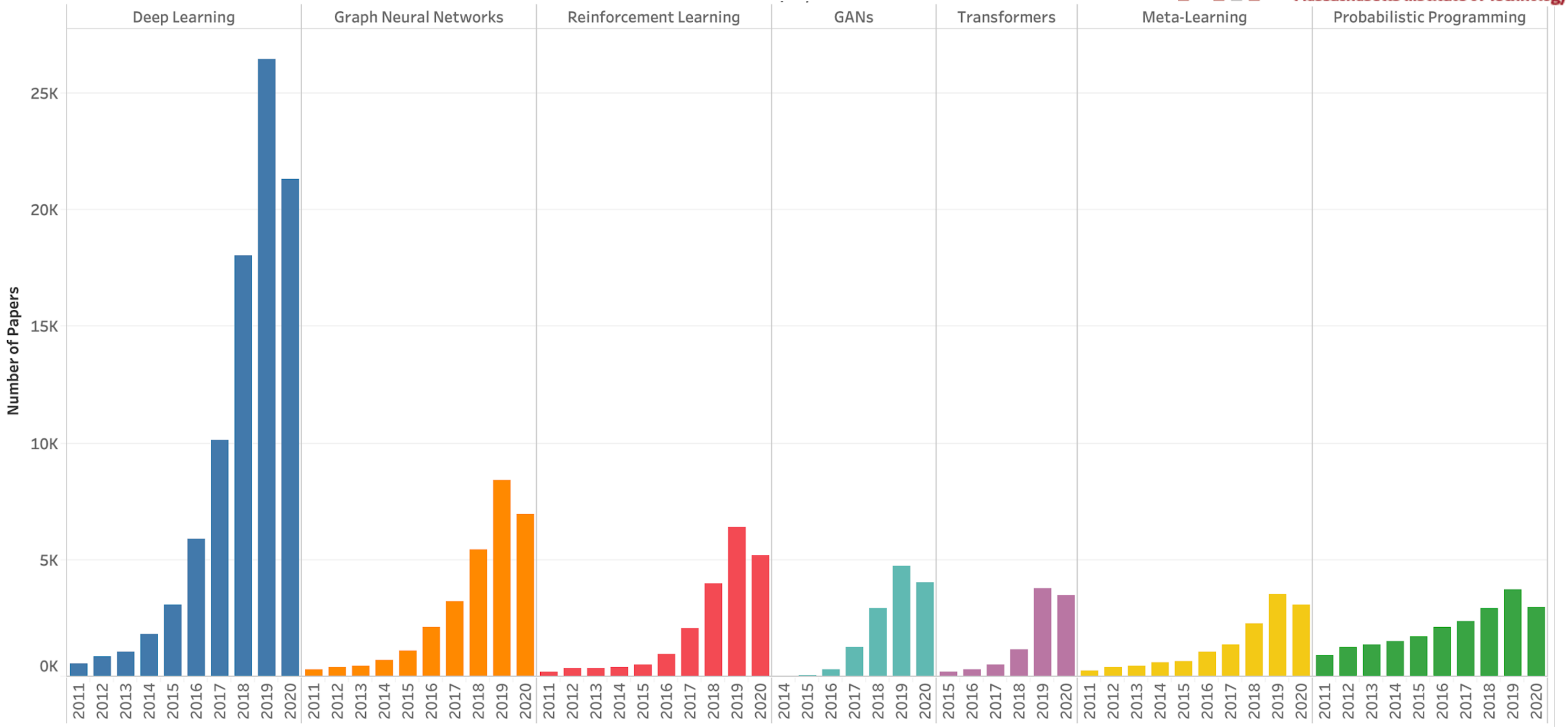
Course Overview: Lecture Topics



Transfer learning
Adaptation
Multi-task learning
Meta learning
Few-shot learning
Online learning
Automated Machine Learning
...

of Papers in 2020 (so far)





Data Source: IBM Science Summarizer

Motivation

Human Brain Connectome

- 100 Billion neurons (1 Billion neurons in cat brain)
- 100 Trillion connections: each neuron connects to 5k-200k
- 10k different types of neurons
- 1k new neurons per day our entire life

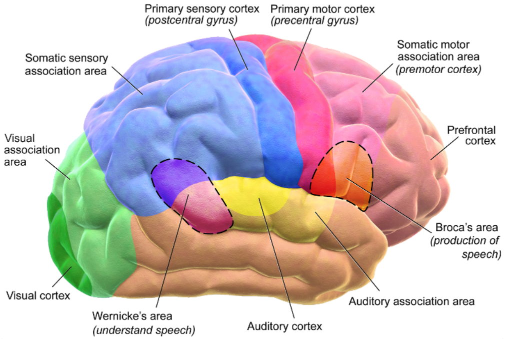
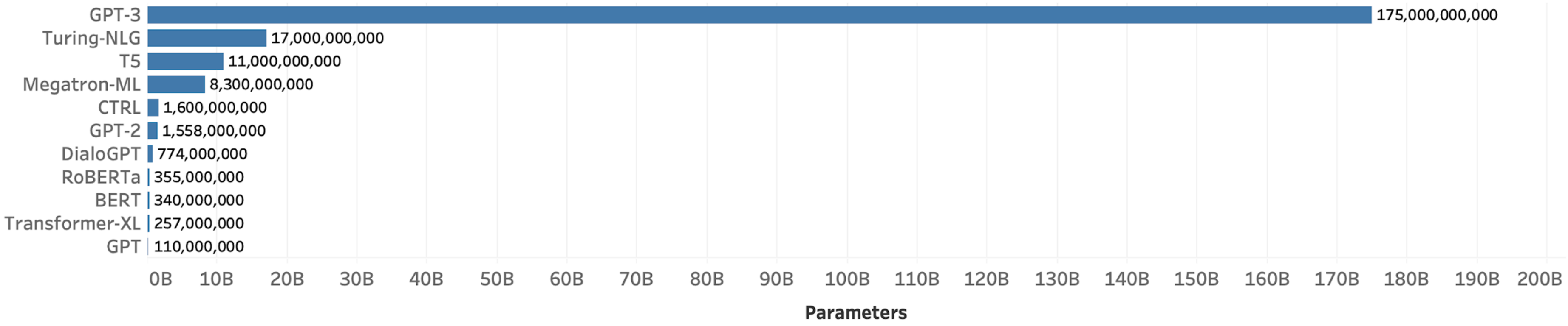


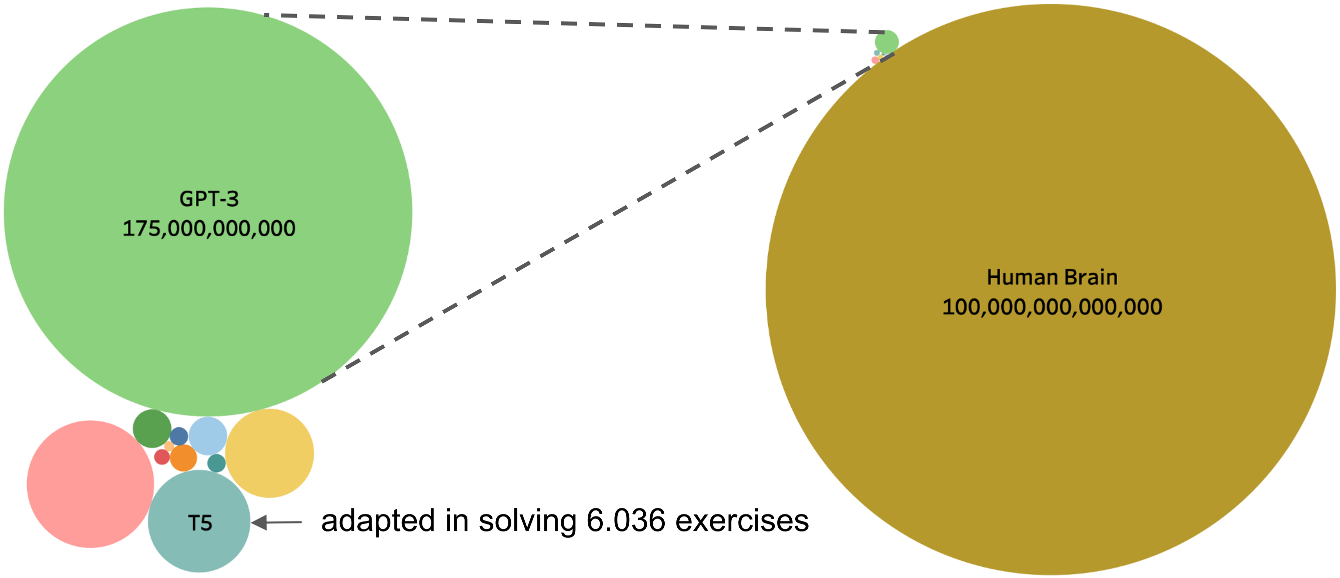
Image Source: Wikipedia

Transformer Parameters



- 3 orders of magnitude less parameters than number of connections in human brain

Number of Connections or Parameters



- Transformers have 3 orders of magnitude less parameters than number of connections in human brain

Super-Human ML Systems: AlphaX

- AlphaZero: board games
- AlphaStar: multiplayer online games
- AlphaFold: protein structure prediction
- AlphaD3M: automated machine learning
- AlphaStock: stock trading
- ..
- AlphaDogfight: fighter pilot

DARPA Programs

- Self driving grand challenge 2 decades ago: competitive.

Recent collaborative efforts

- Data-Driven Discovery of Models (D3M): AutoML
- Learning with Less Labels (LwLL): few shot learning
- Lifelong Learning Machines (L2M): online learning
- Machine Common Sense (MCS)

Automated machine learning, few shot learning, online learning, learn to read, natural language understanding

Meta Learning Definitions

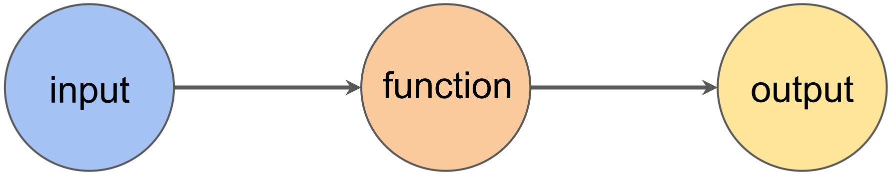
Definitions

- Supervised learning
- Transfer learning
- Meta learning
- Automated machine learning

- Adaptation
- Multi-task learning
- Few-shot learning
- Online learning

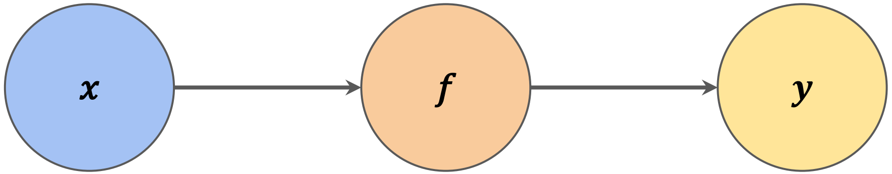
Observation

- Input $\mathbf{x}_{d\text{x}1}$
- Function $\mathbf{f}$
- Output $y_{1\text{x}1}$



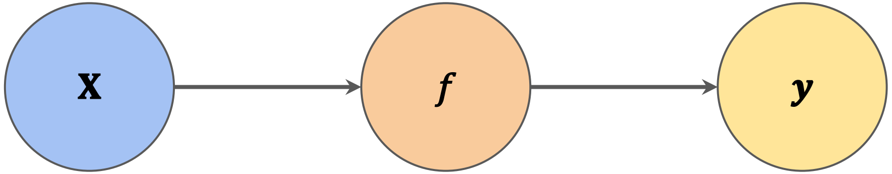
Observation

- Input $\mathbf{x}_{dx1}$
- Function **f**
- Output y_{1x1}



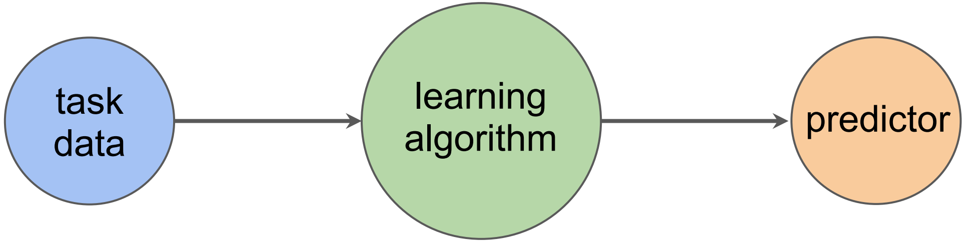
Observations

- Input $\mathbf{X}_{dxm}$
- Function **f**
- Output $\mathbf{y}_{mx1}$

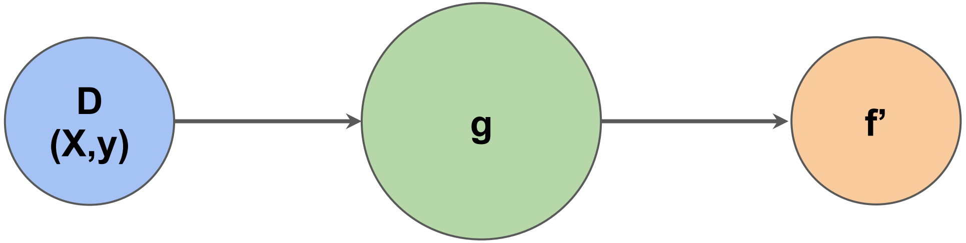


$$\boldsymbol{y} = f(\boldsymbol{X})$$

Supervised Learning

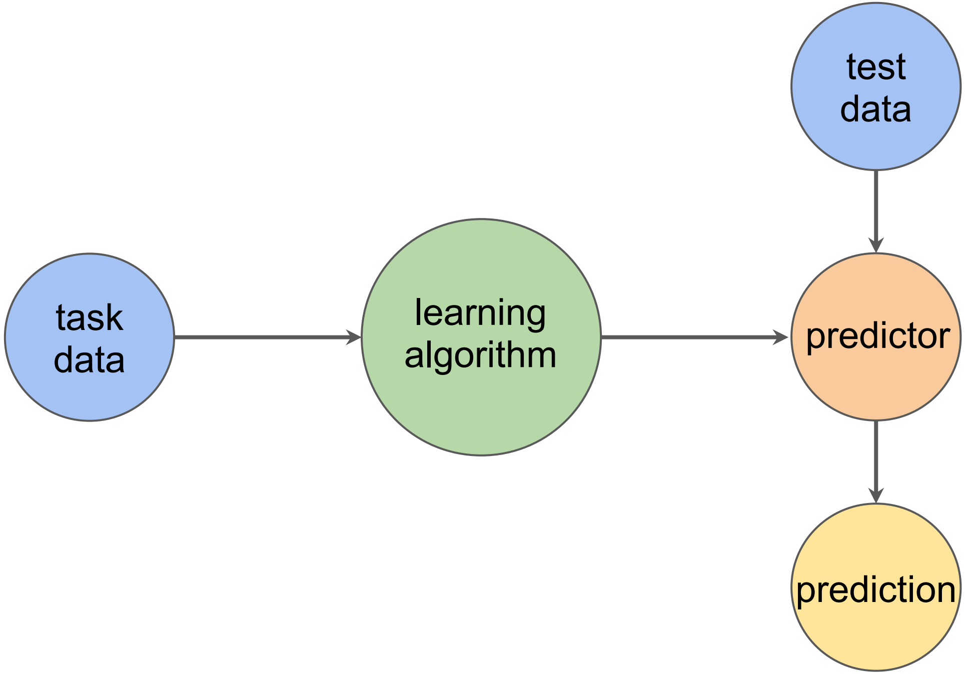


Supervised Learning

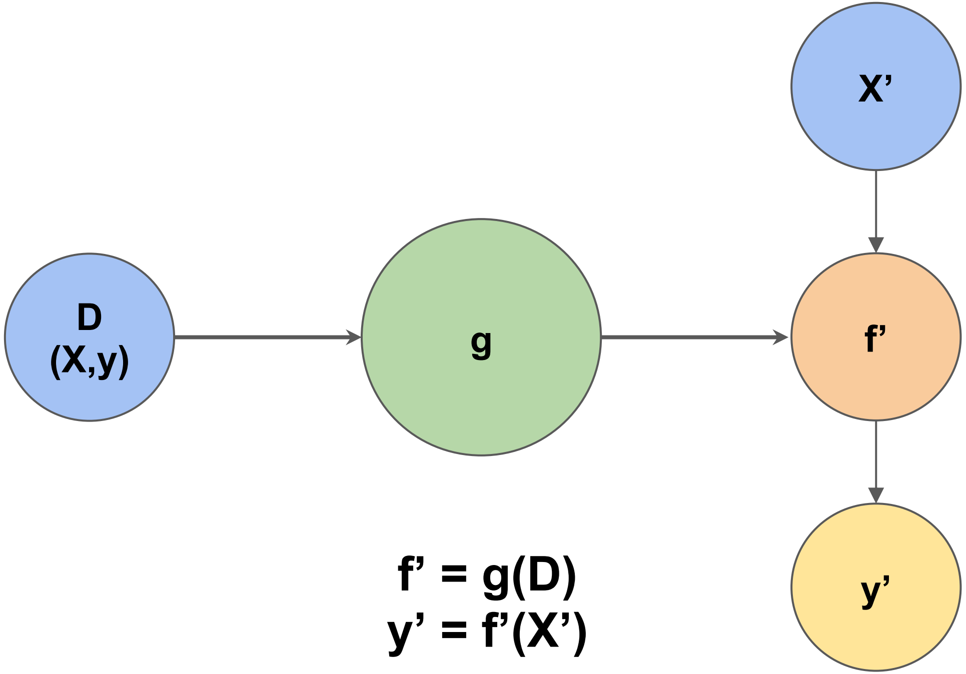


f’ = g(D)

Supervised Learning

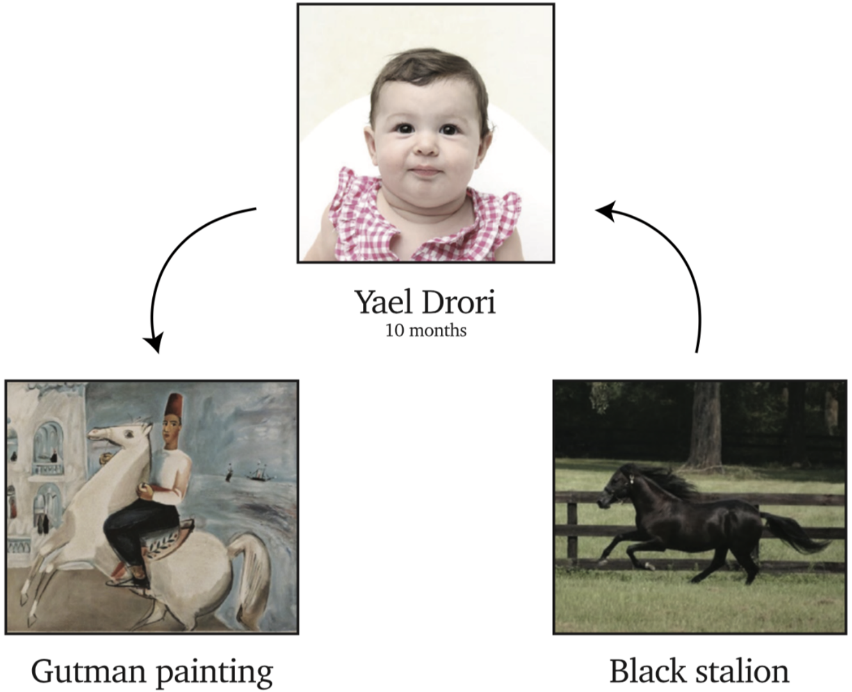


Supervised Learning



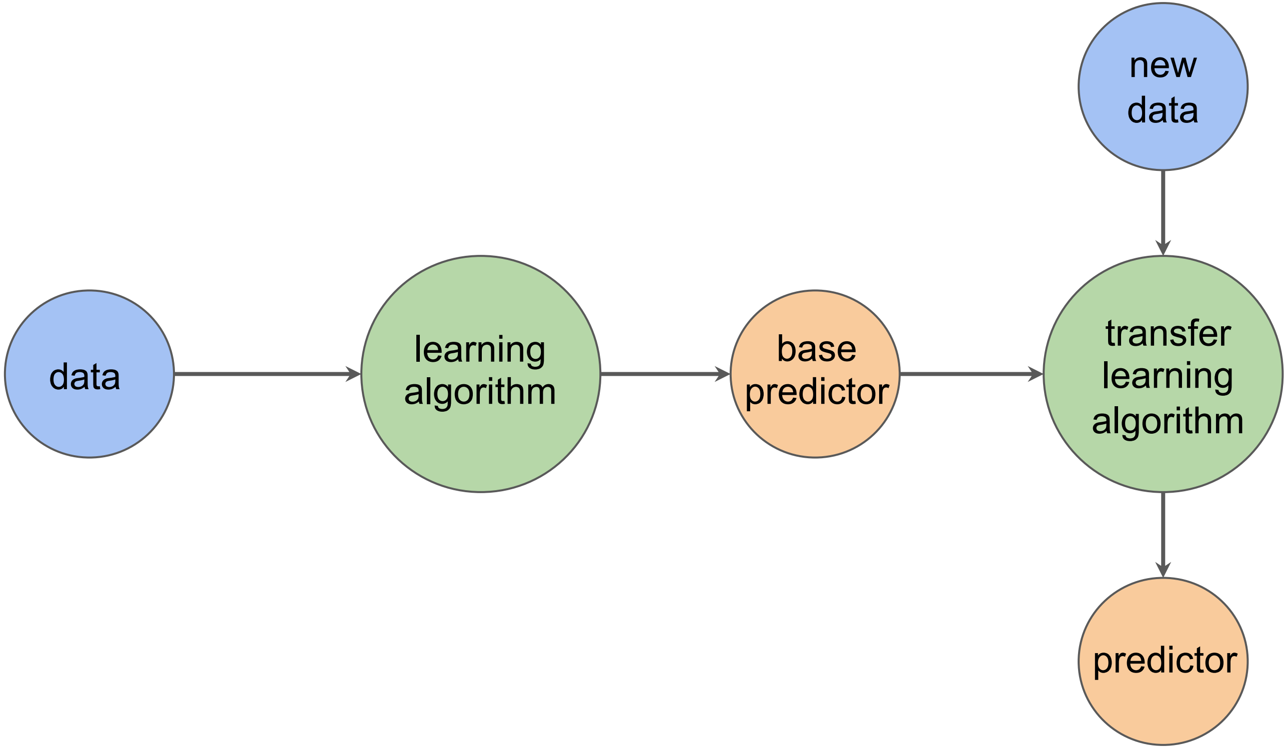
f’ = g(D)
y’ = f’(X’)

Transfer Learning

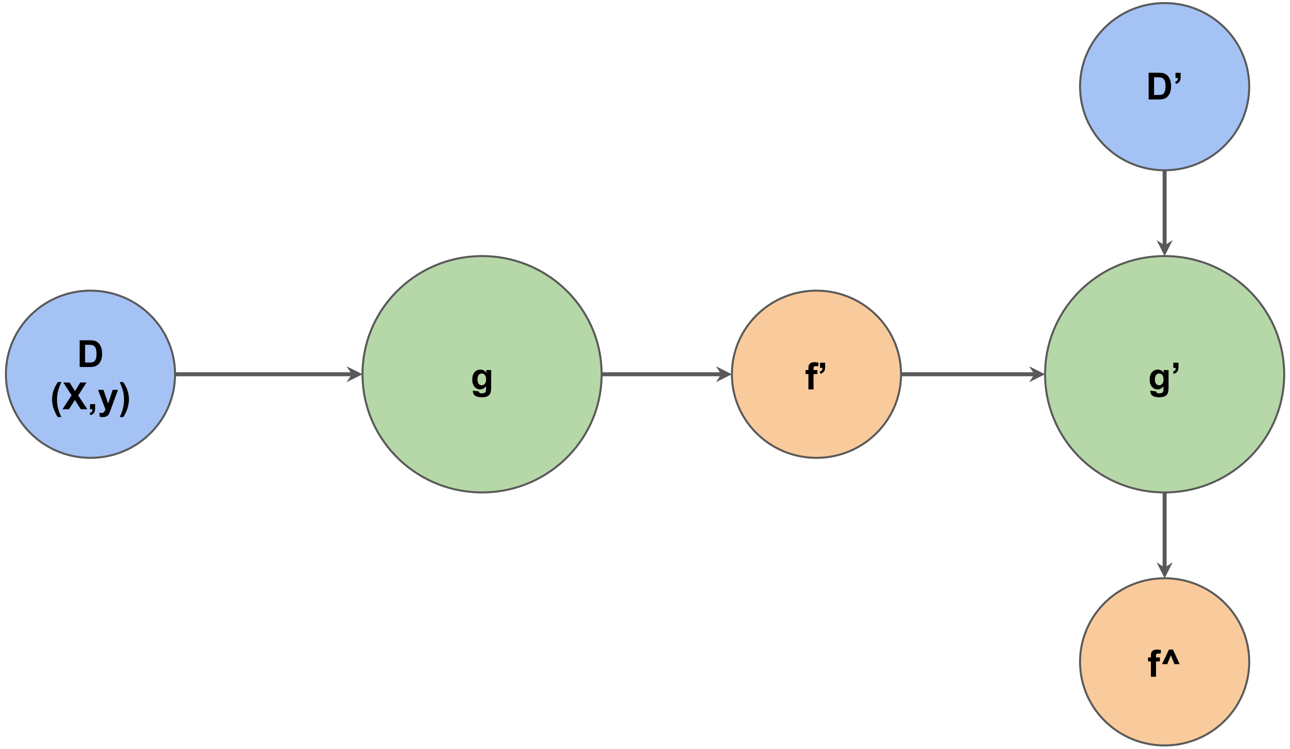


Source: Deep Learning course 2017, Iddo Drori

Transfer Learning

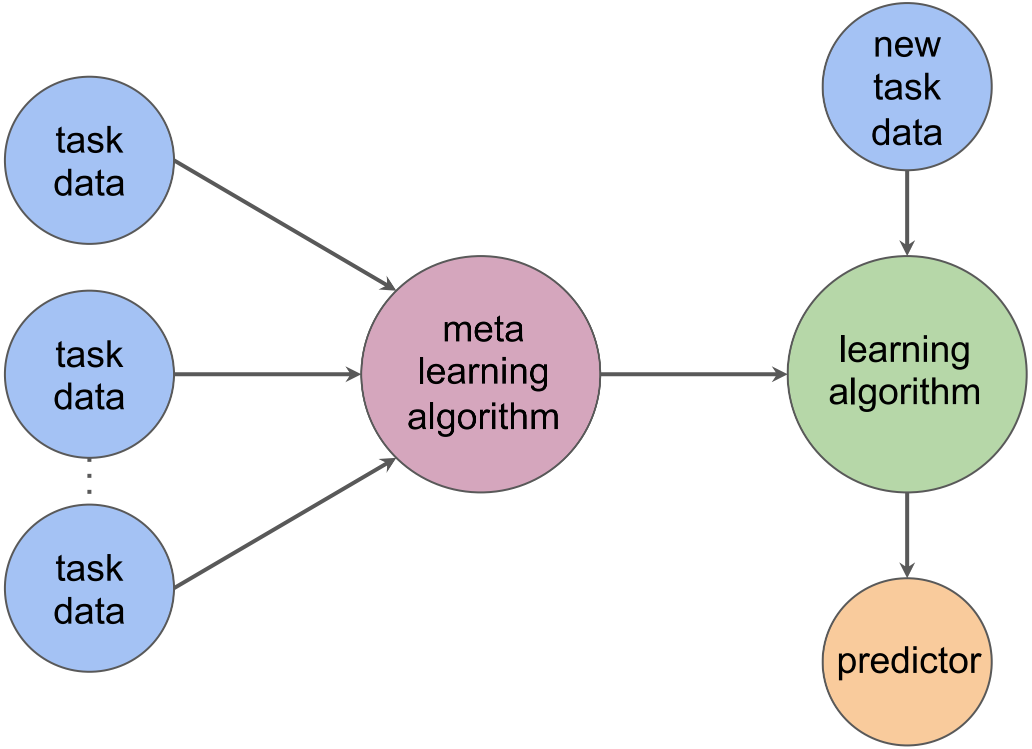


Transfer Learning

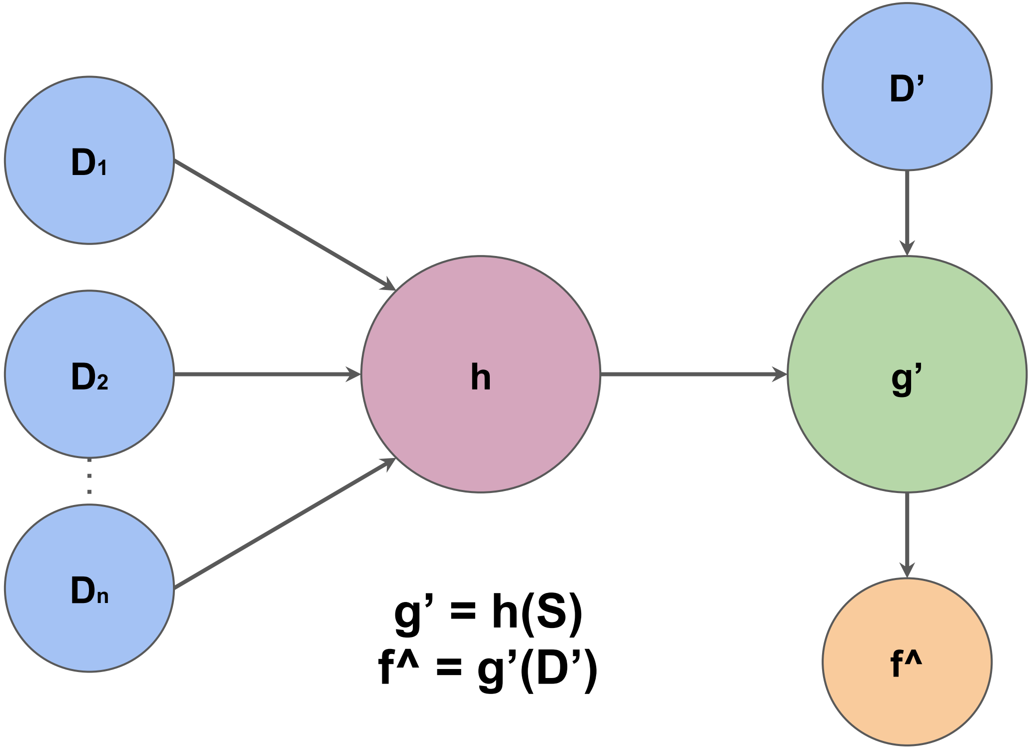


Meta Learning

task = data splits, priors

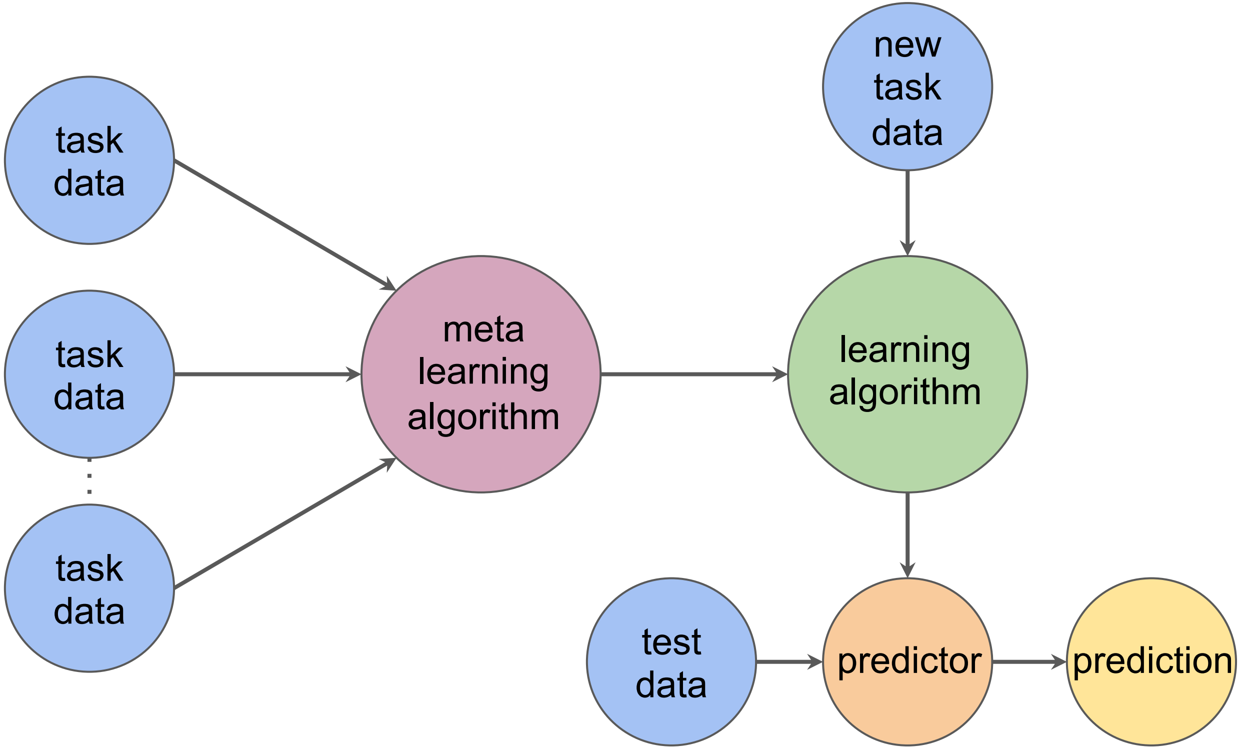


Meta Learning

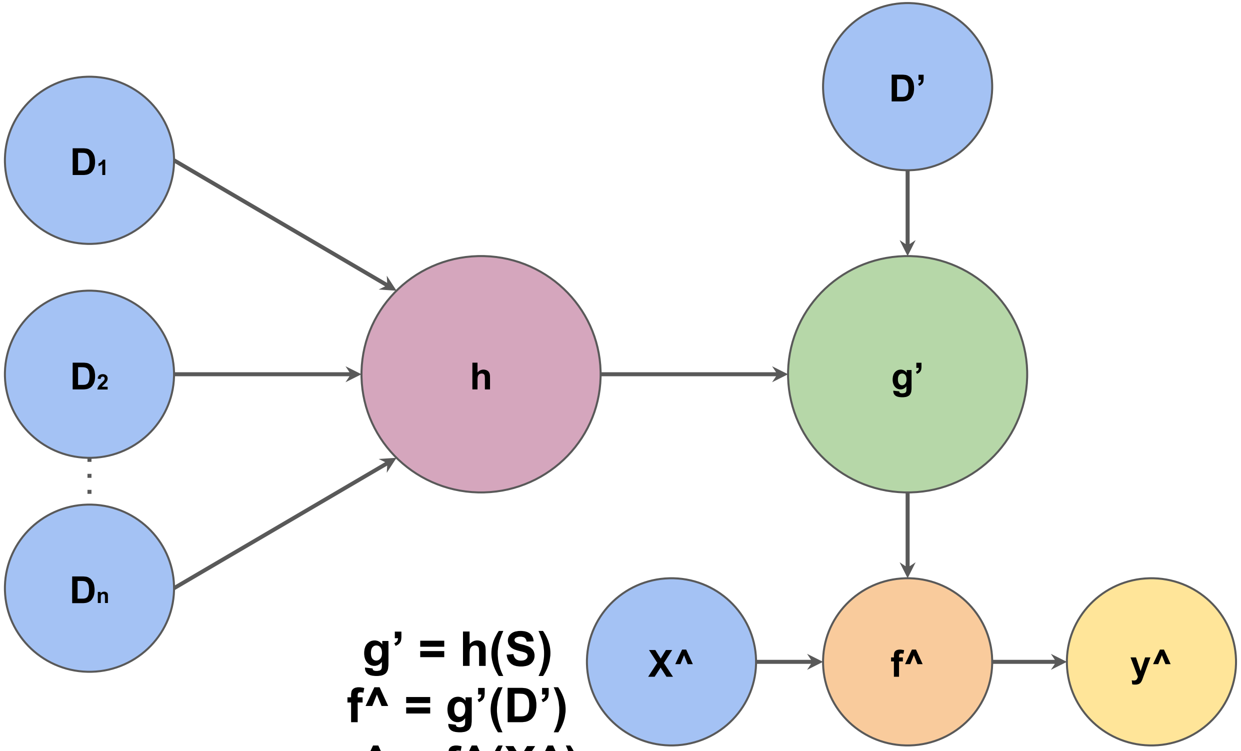


Meta Learning

task = data splits, priors



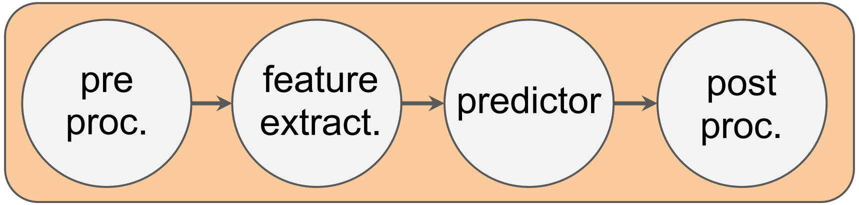
Meta Learning



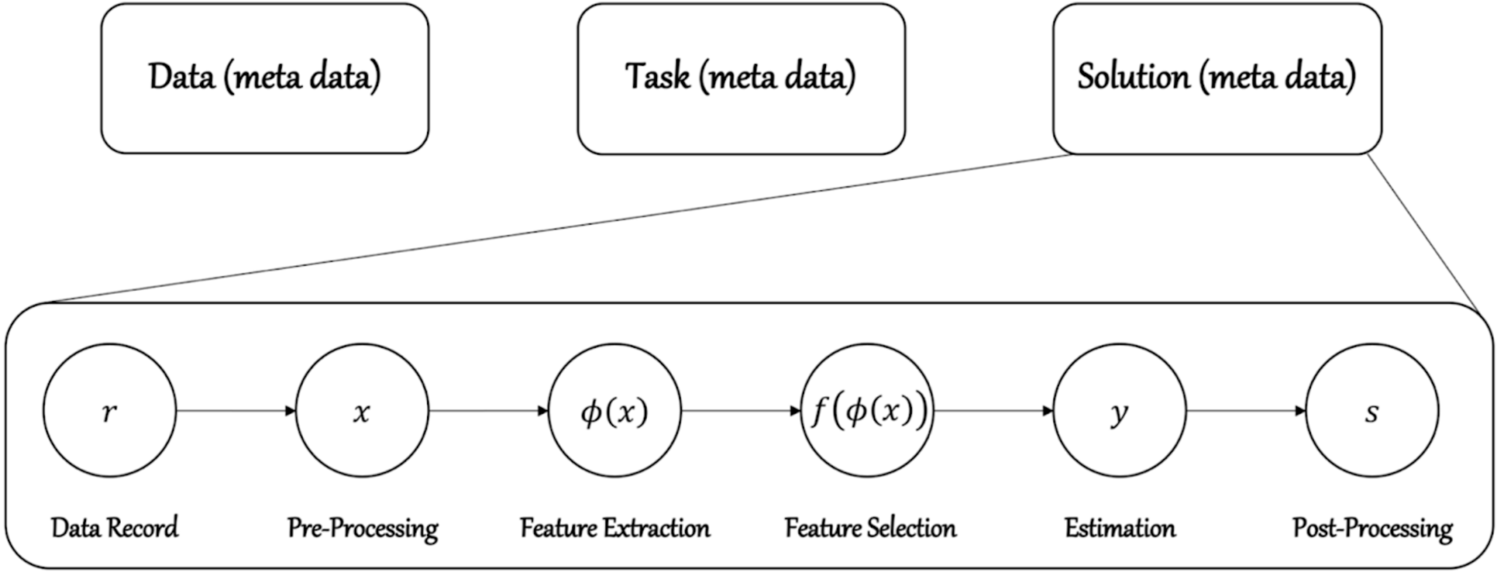
$$g' = h(S)$$
$$\hat{f} = g'(D')$$
$$\hat{y} = \hat{f}(\hat{X})$$

Machine Learning System

- Predictor is part of a machine learning system
- Built from data science / machine learning primitives
- Machine learning primitive = {PCA, SVM, NN,...}
- Example machine learning pipeline:

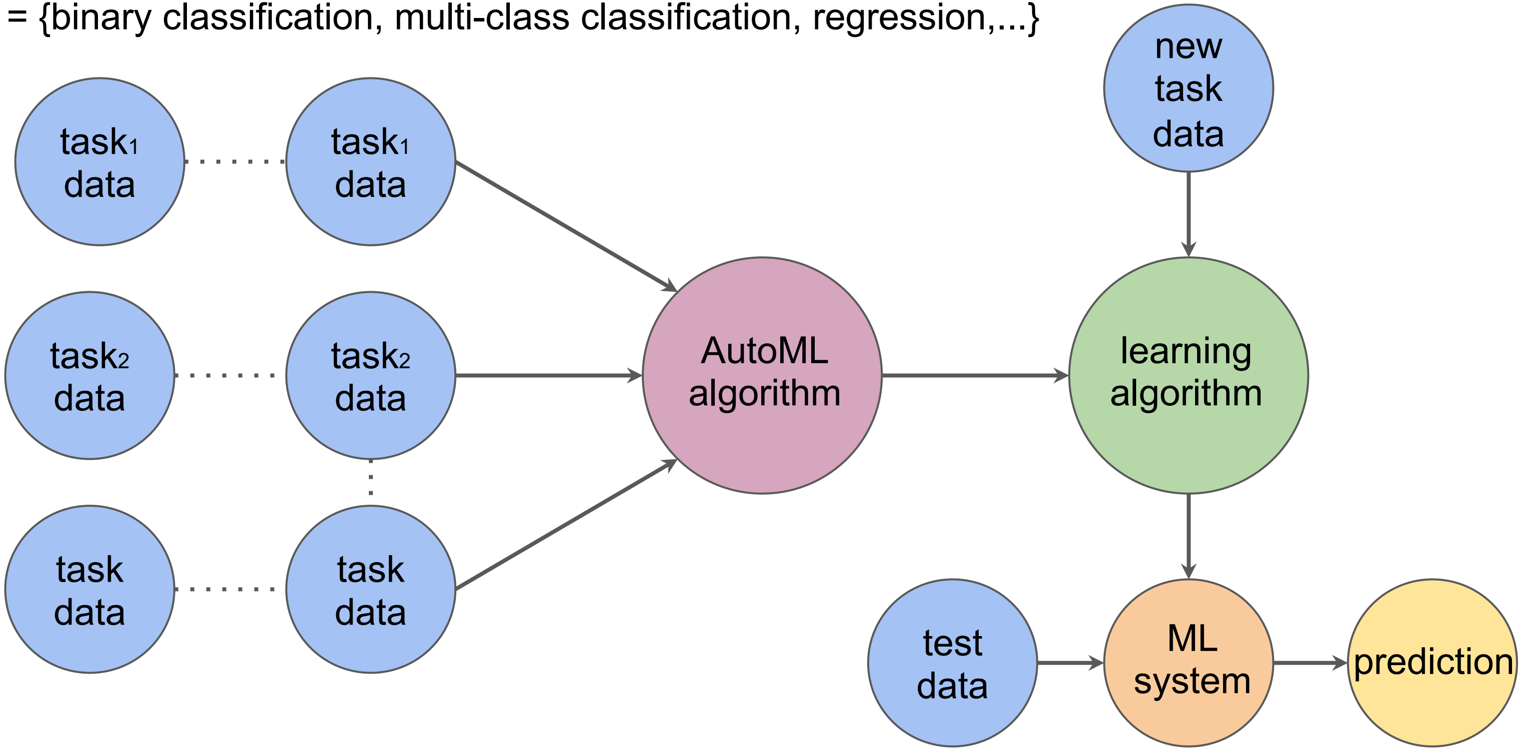


Machine Learning Pipeline

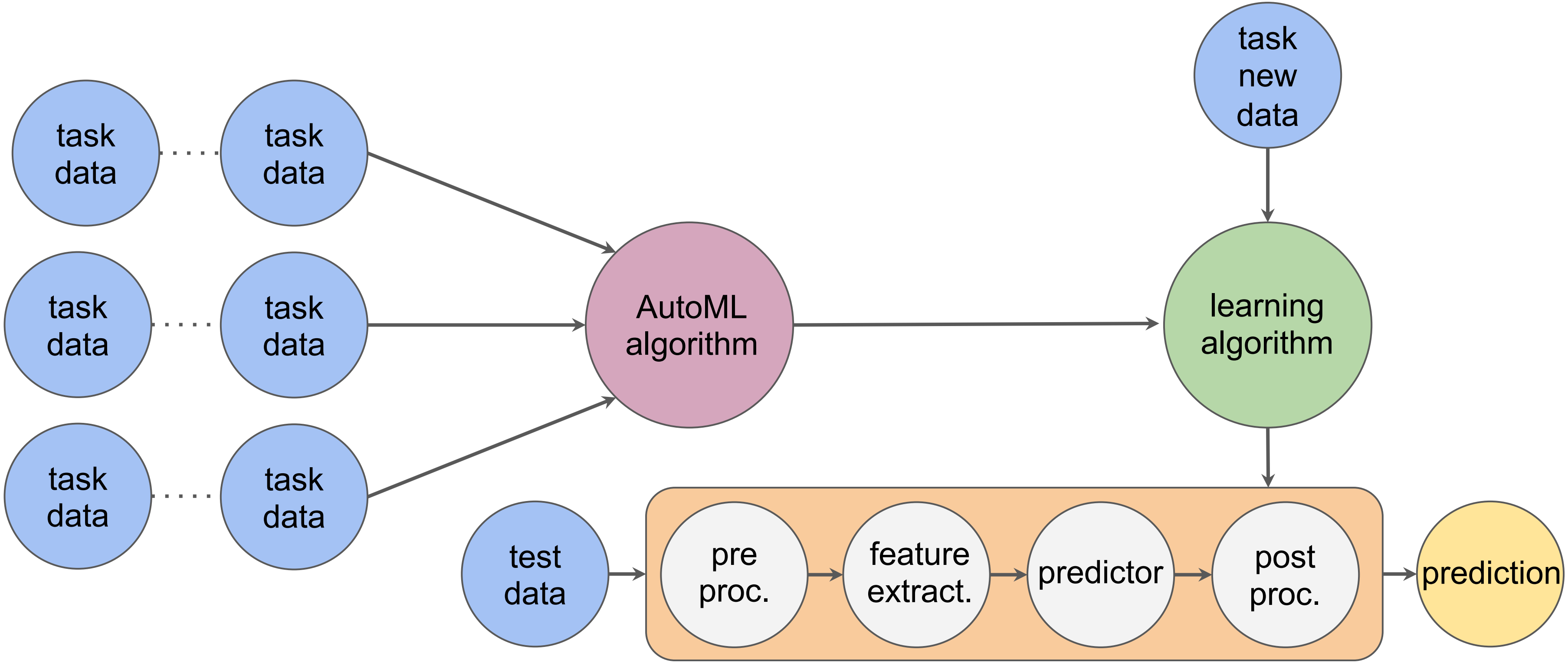


Automated Machine Learning

tasks = {binary classification, multi-class classification, regression,...}



Automated Machine Learning (AutoML)



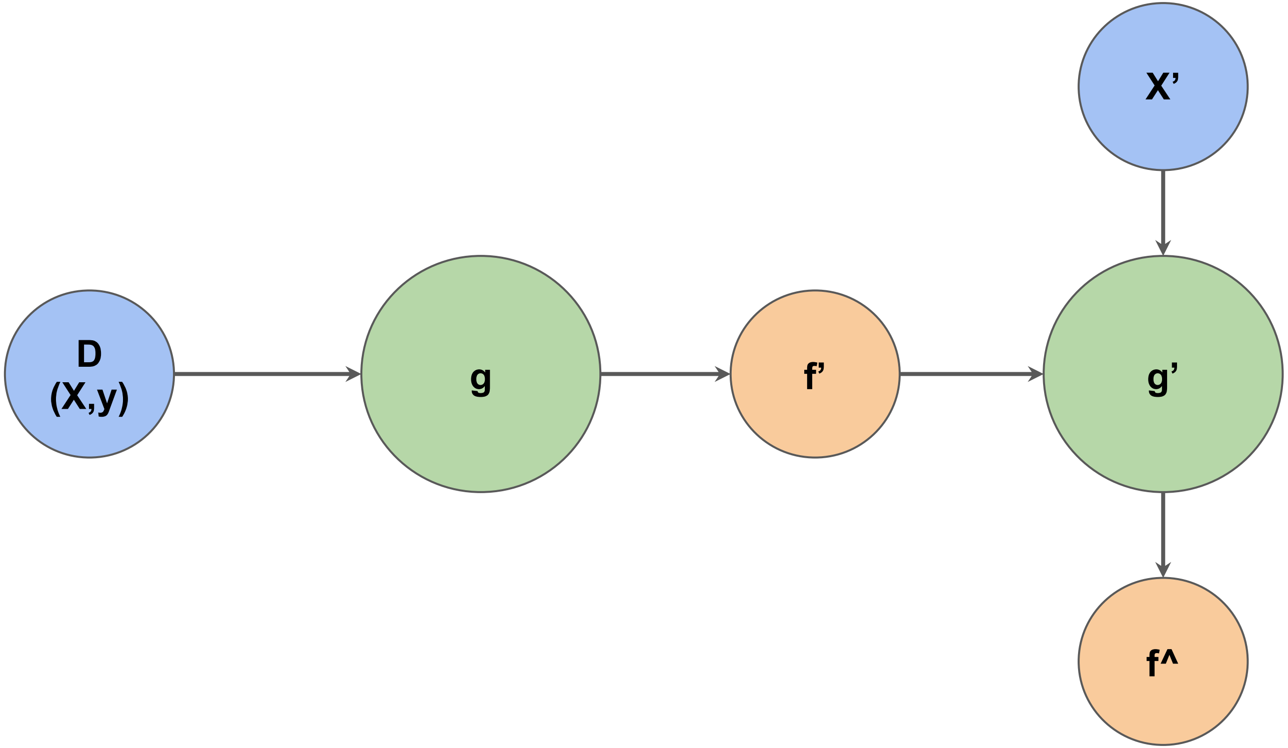
Learning to Learn

- Machine Learning: learn parameters of $\mathcal{M}$

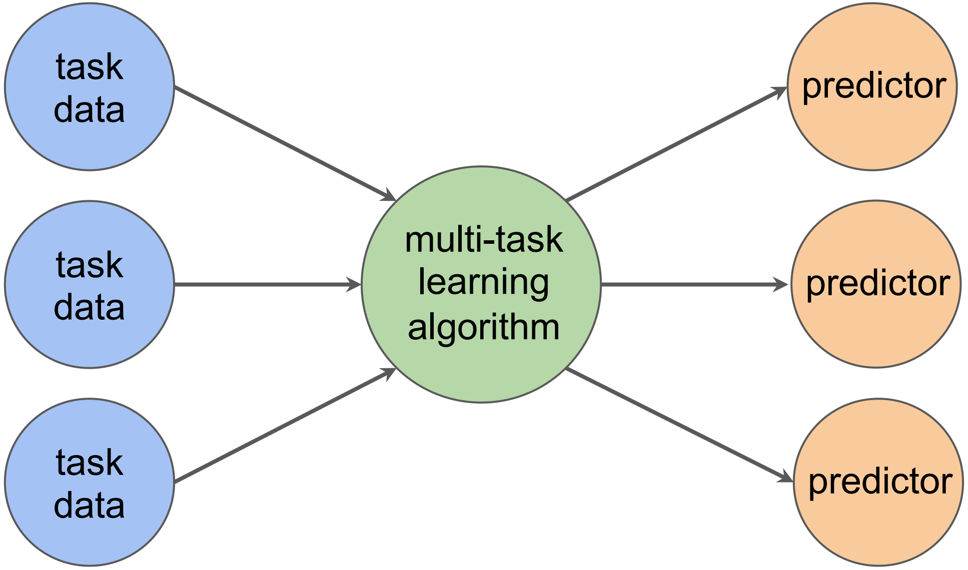
- Learning to learn: learn $\mathcal{M}$ and parameters

  where $\mathcal{M}$ is a classifier or machine learning pipeline or machine learning algorithm or reinforcement learning method, ect.

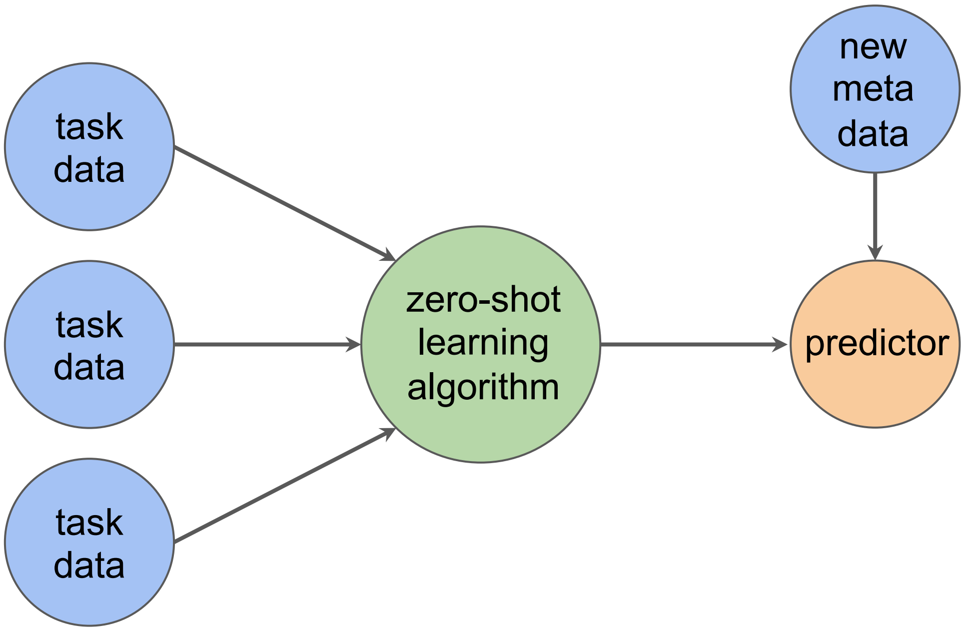
Adaptation (Unsupervised Transfer Learning)



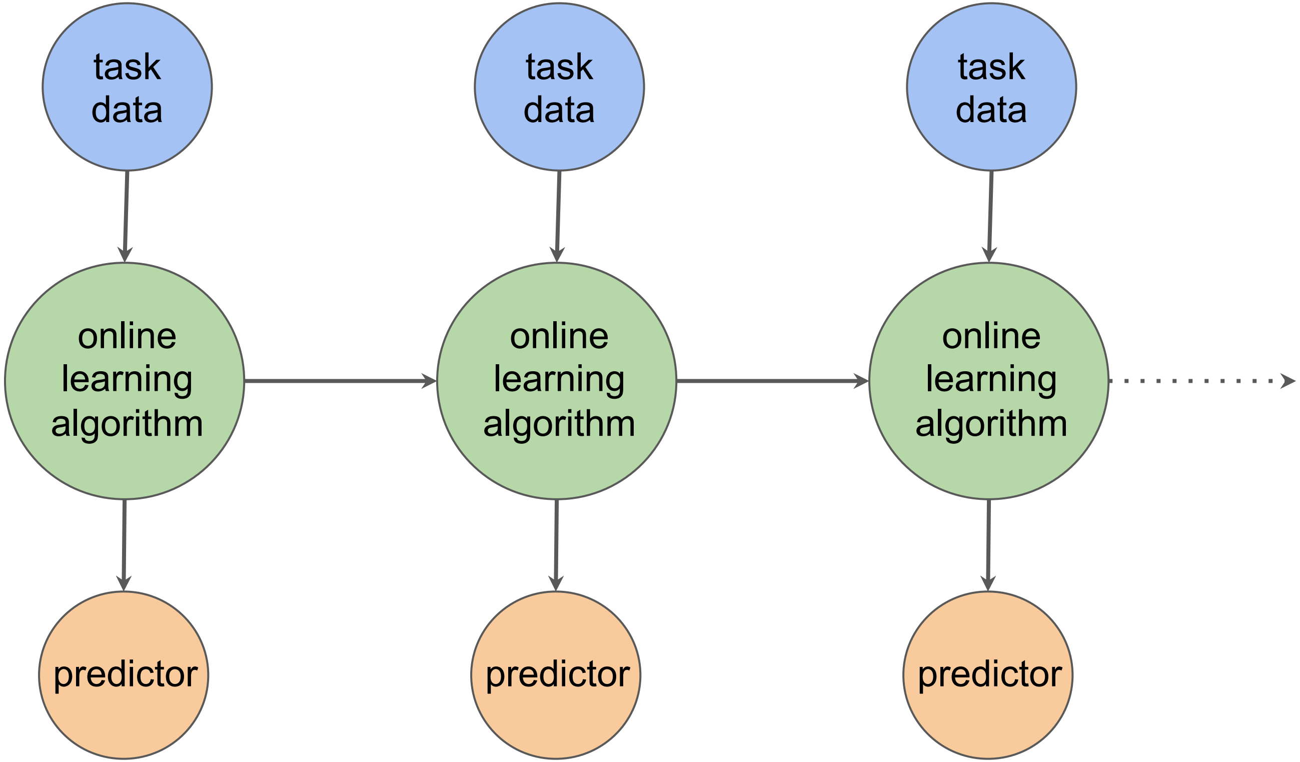
Multi-Task Learning



Zero-Shot Learning



Online Learning (Sequential, Lifelong learning)



Applications

Massachusetts Institute of Technology

Applications

- Automated machine learning
- Computer vision: learning from small data, ..
- Robotics: learning from a few examples, ..
- Information retrieval: adaptation between domains
- Natural language processing
- Cross-lingual generalization
- Machine translation
- Mobile data analysis
- Discovering physics formulas
- Education, answering and generating math questions
- Learning to learn courses
- Learning to code
- Combinatorial optimization
- Autonomous vehicle

DARPA Data Driven Discovery of Models (D3M)

- AutoML goal: solve any task on any dataset specified by a user.

- Broad set of computational primitives as building blocks.

- Automatic systems for machine learning, synthesize pipeline and hyperparameters to solve a previously unknown data and problem.

- Human in the loop: user interface that enables users to interact with and improve the automatically generated results.

- Pipelines: pre-processing, feature extraction, feature selection, estimation, post-processing, evaluation

Human Example

Learning to Code

- Background
- Human performance similar to sports..
- Machines will be in a league of their own.
- Code machines to learn to code.

Bayesian Inference

Probability

- Observed data x, latent variables z
- Inference about hidden variables given by posterior conditional distribution p(z|x)

$$p(z, x) = p(z|x)p(x) = p(x|z)p(z) = p(x, z)$$

- Extending likelihood p(x|z) times prior p(z) to multiple layers

$$p(x|z_1)p(z_1|z_2)\cdots p(z_{l-1}|z_l)p(z_l)$$

- Bayes rule

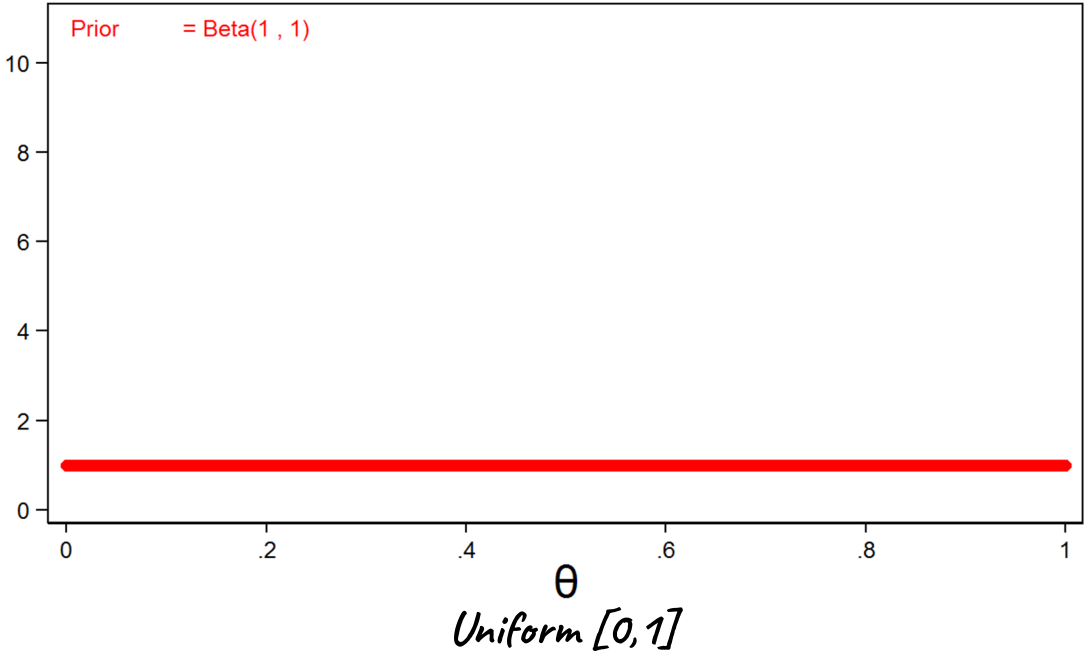
$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

Probability

- High dimensional intractable integral over exponential number of terms for z:

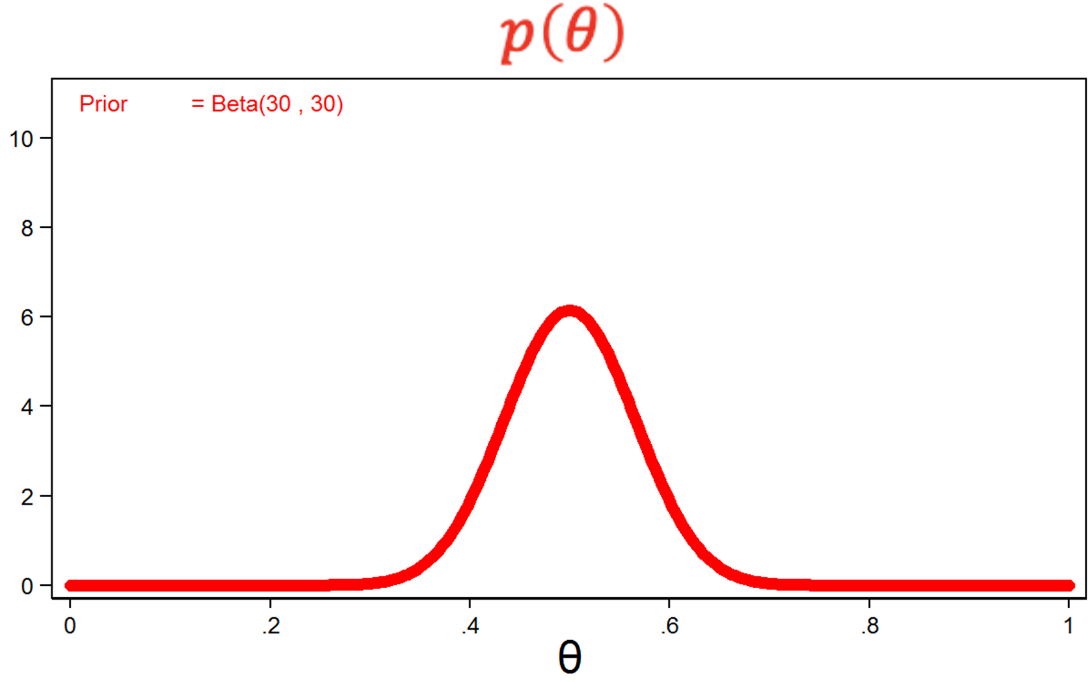
$$p(x) = \int p(x|z)p(z)dz$$

Uninformative Beta prior



Animation Source: Stata

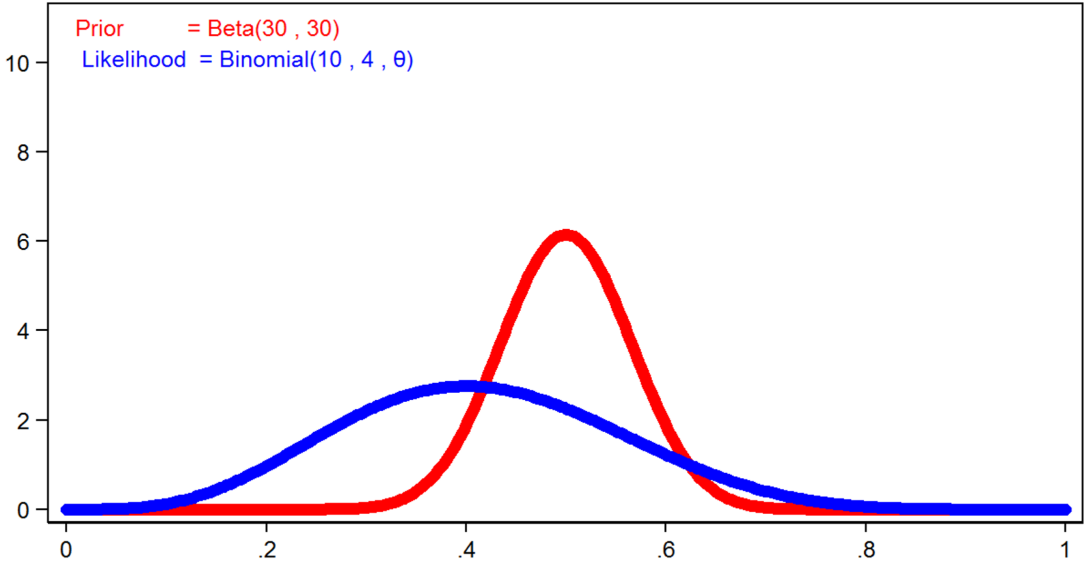
Informative Beta prior distribution



Animation Source: Stata

Binomial likelihood and Beta prior

$p(\theta)$ $p(y|\theta)$



Observe 4 heads out of 10 coin flips

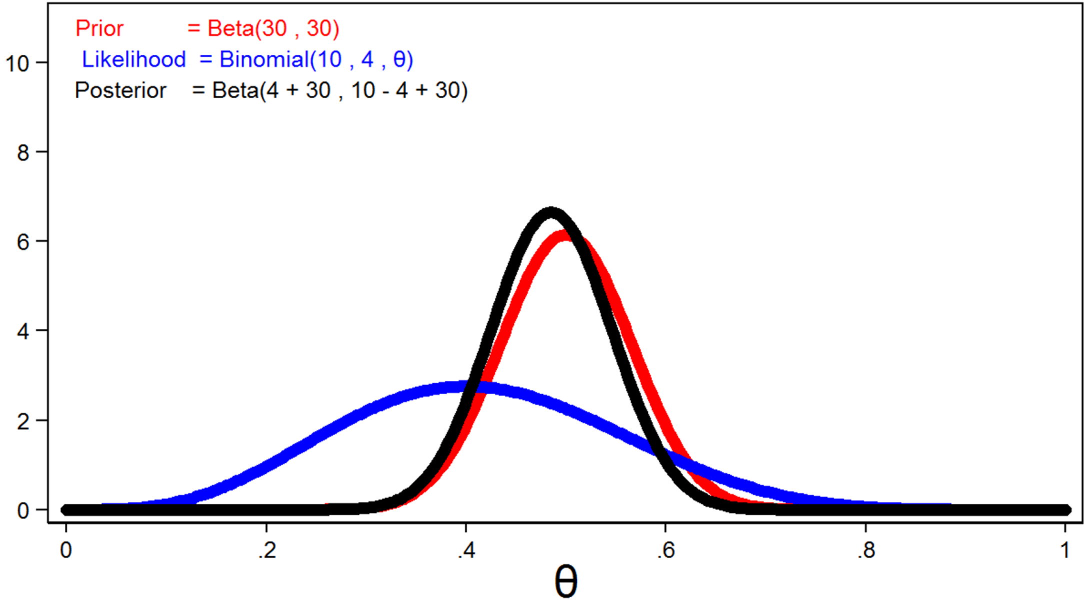
Binomial likelihood function

Animation Source: Stata

Update belief based on result of experiment

$$Posterior \propto Prior \ x \ Likelihood$$

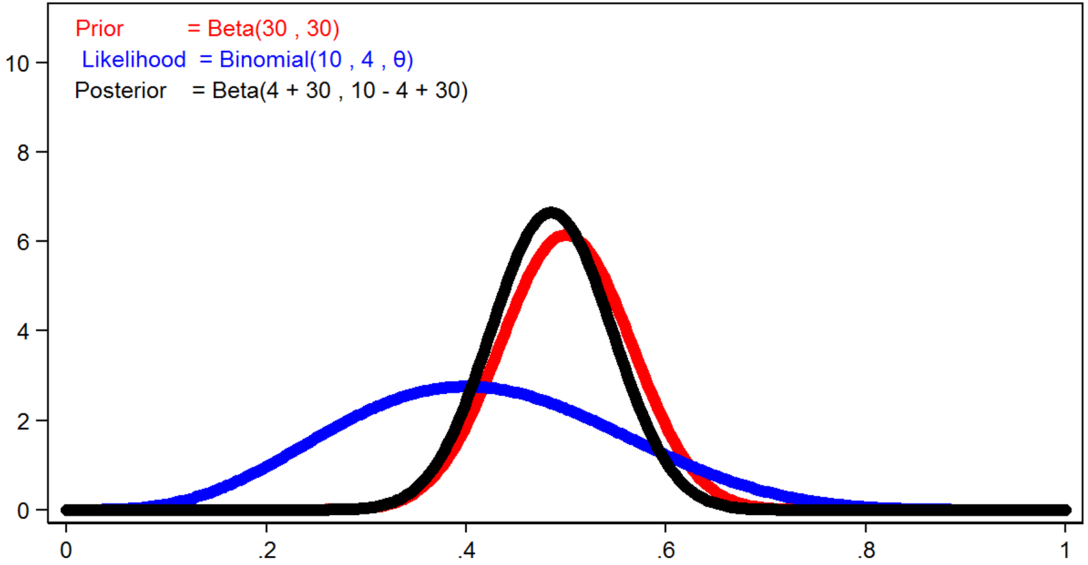
$$p(\theta|y) \propto p(\theta)p(y|\theta)$$



Animation Source: Stata

Update belief based on result of experiment

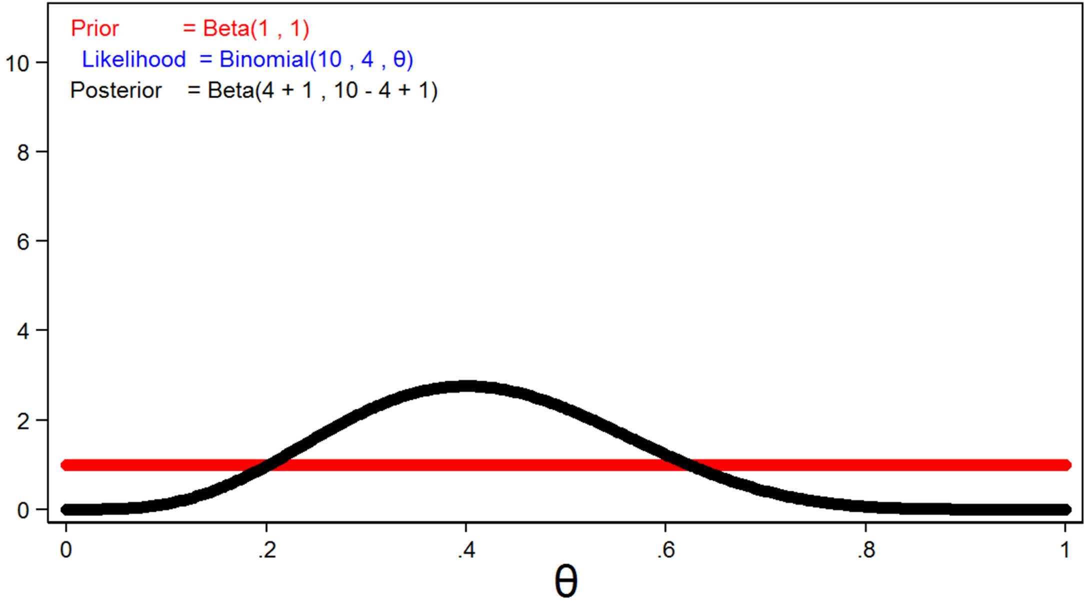
$$p(\theta|y) = Beta(\alpha, \beta) x Binomial(n, \theta) = Beta(y + \alpha, n - y + \beta)$$



Beta distribution is a **conjugate** prior for binomial likelihood function since **posterior distribution belongs to same family as prior distribution**

Animation Source: Stata

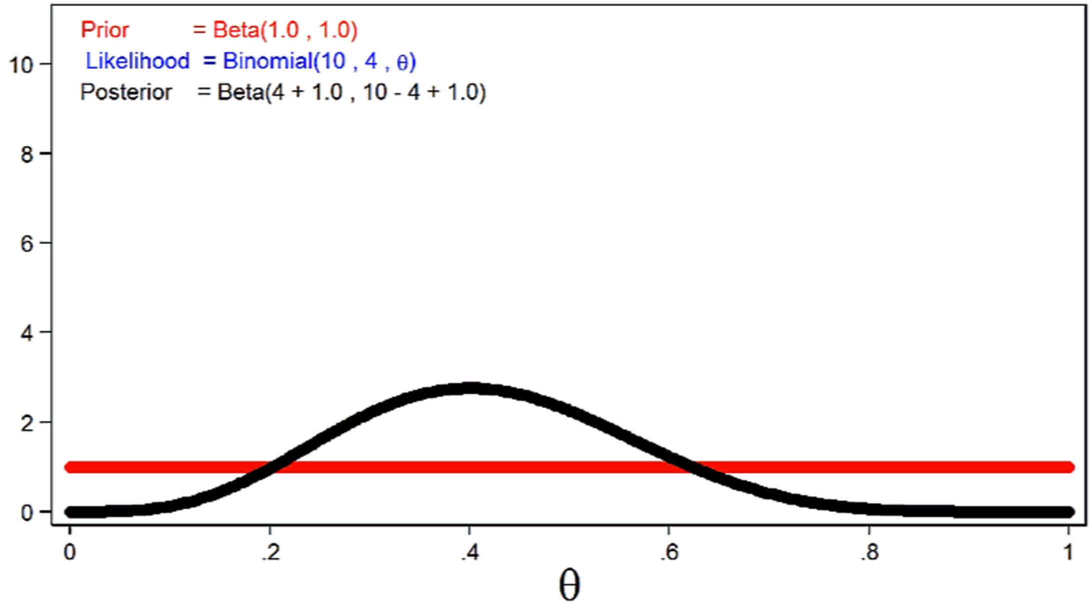
Posterior for Beta(1,1) prior



$$p(\theta|y) = Beta(\alpha,\beta) x Binomial(n,\theta) = Beta(y+\alpha, n-y+\beta)$$

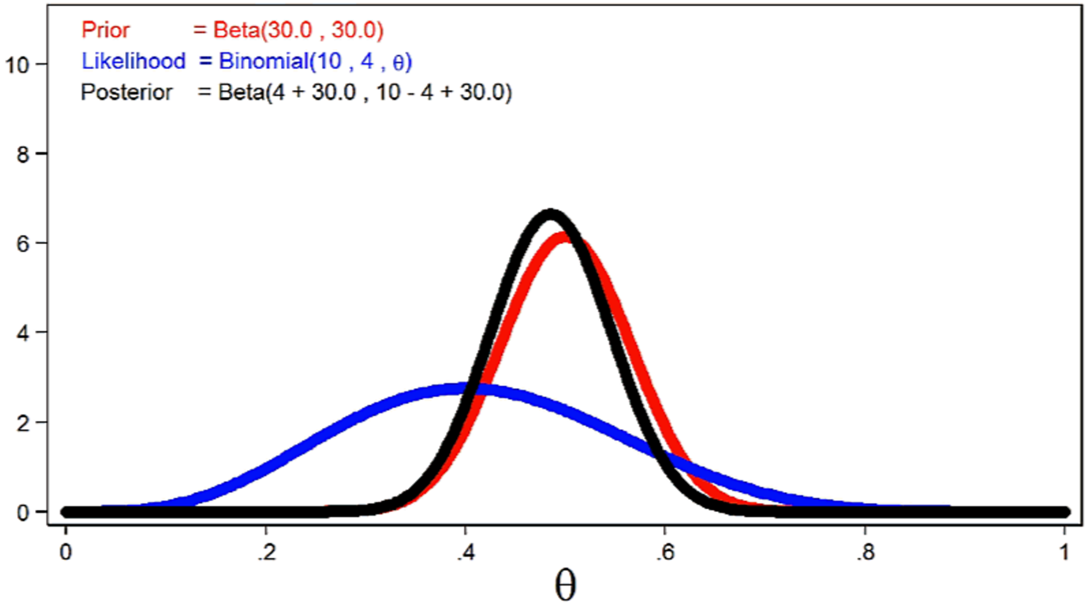
Animation Source: Stata

Effect of more informative prior distribution on posterior distribution



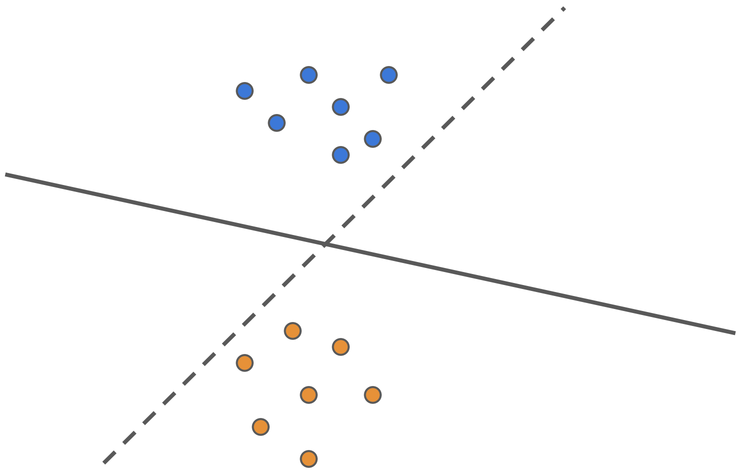
Animation Source: Stata

Effect of larger sample size on posterior distribution



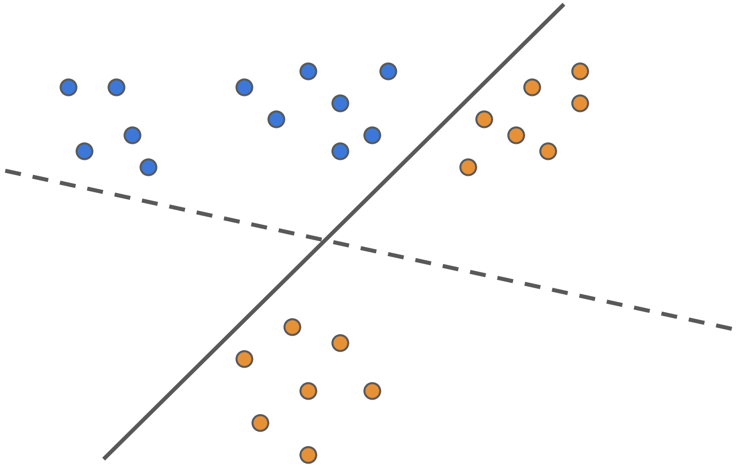
Animation Source: Stata

Example



Example

- Set prior to previous posterior
- Recompute



Meta Learning

MIT
Iddo Drori, Fall 2020