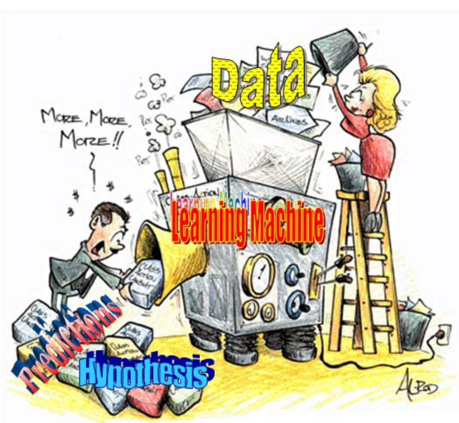
Machine Learning

10-701, Fall 2016

Introduction to ML and Density Estimation



Eric Xing

Lecture 1, September 7, 2016

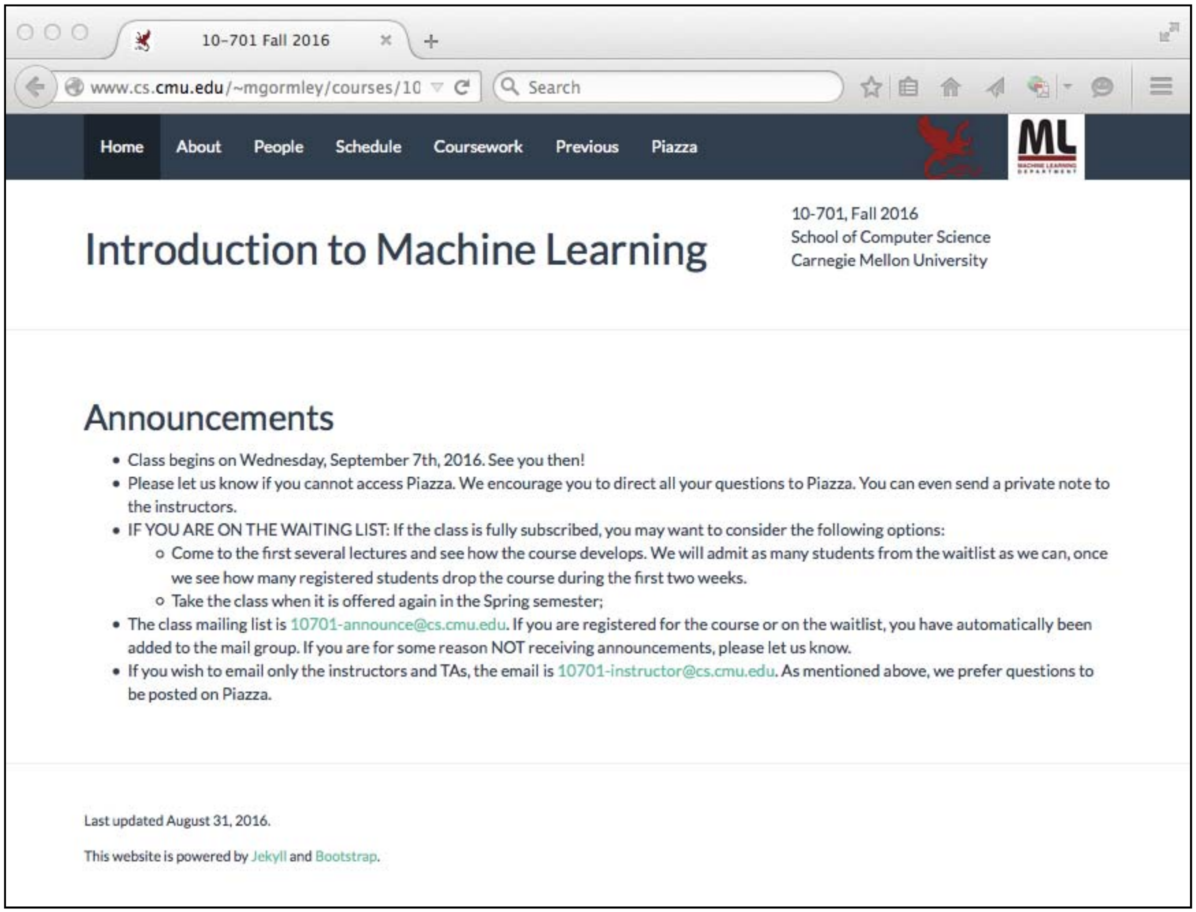
Reading: Mitchell: Chap 1,3

Class Registration

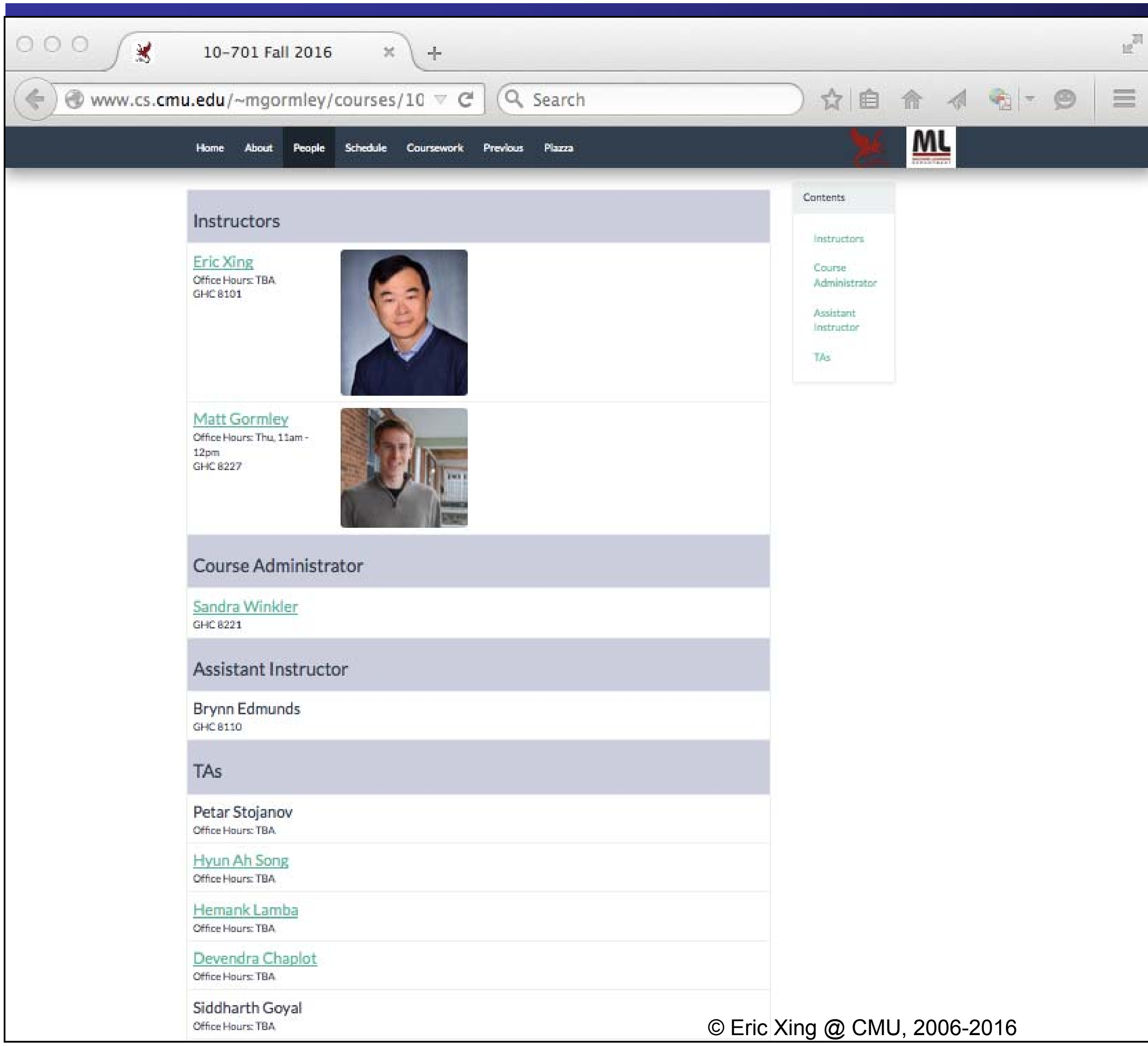
- IF YOU ARE ON THE WAITING LIST: This class is now fully subscribed. You may want to consider the following options:
  - Take the class when it is offered again in the Spring semester;
  - Come to the first several lectures and see how the course develops. We will admit as many students from the waitlist as we can, once we see how many registered students drop the course during the first two weeks.

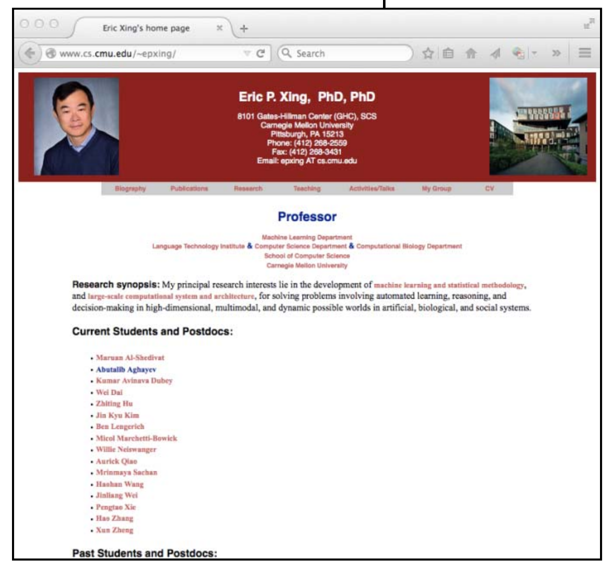
Machine Learning 10-701

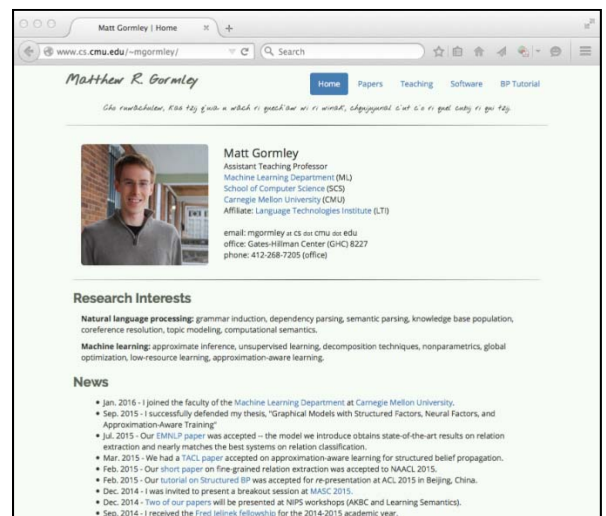
- Class webpage:
  - http://www.cs.cmu.edu/~mgormley/courses/10701-f16/



The instructors







Brynn Edmunds

- Previous Research
  - Medical Physics with specific interest in Radiotherapy and Radiation Oncology
    - Examination of DVH parameters for prostate treatments
    - Comparing clinicians with different training to look for treatment variability
- Currently: ML Assistant Instructor

- Devendra Chaplot
- Office Hour: Friday 11:00am -12:00pm
- Location: GHC 5412
- Interests: Concept Graph Learning, Computational models of human learning, Reinforcement Learning

- **Siddharth Goyal**
- **Office Hour: Tue$_{th}$ 4:00pm -5:00pm**
- **Location: GHC 5 floor common area**
- **Interests: Bayesian optimization, Reinforcement learning**

Hemank Lamba

Office Hours: Tuesday, 11 to Noon

Location: TBD

Research

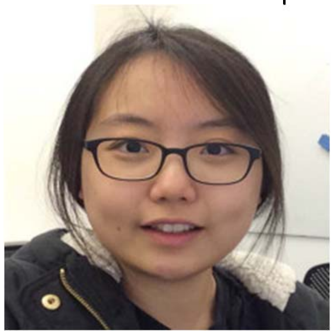
- **Graph Mining**
- **Data Mining**
- **Anomaly Detection**
- **Social Good Applications**

Hyun Ah Song

Office hour: Friday 1pm-2pm

Office: GHC 8003

Interests: time series analysis



Petar Stojanov

Office Hours: Wednesday, 4:30 to 5:30pm
(starting next week)

Location: TBD

Research
- **Transfer Learning**
- **Domain Adaptation**
- **Multitask Learning**

Logistics

- Text book
  - Chris Bishop, **Pattern Recognition and Machine Learning (required)**
  - Kevin Murphy, **Machine Learning, a probabilistic approach**
  - Tom Mitchell, **Machine Learning**
  - David Mackay, **Information Theory, Inference, and Learning Algorithms**
- Mailing Lists:
  - To contact the instructors: 10701-instructors@cs.cmu.edu
  - Class announcements list: 10701-announce@cs.cmu.edu.
- **Piazza …**

Logistics

- 5 homework assignments: 35% of grade
  - Theory exercises
  - Implementation exercises
- **Final project: 35% of grade**
  - Applying machine learning to your research area
    - NLP, IR,, vision, robotics, computational biology …
  - Outcomes that offer real utility and value
    - Search all the wine bottle labels,
    - An iPhone app for landmark recognition
  - Theoretical and/or algorithmic work
    - a more efficient approximate inference algorithm
    - a new sampling scheme for a non-trivial model …
  - 3-member team to be formed in the first two weeks, proposal, mid-way report, poster & demo, final report.
- One Midterm: 30%
  - Theory exercises and/or analysis. Dates already set (no "ticket already booked", "I am in a conference", etc. excuse …)
- Policies …

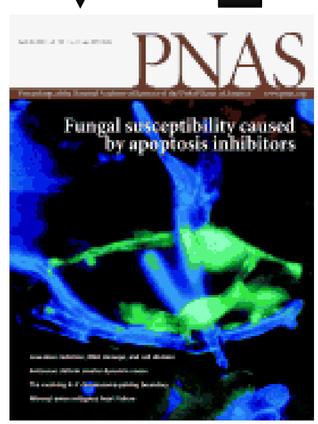
What is Learning

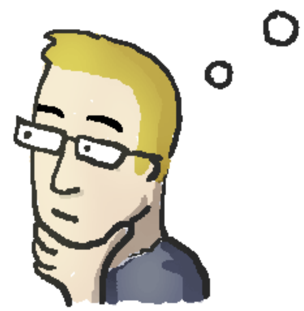
Learning is about seeking a **predictive** and/or **executable** understanding of natural/artificial subjects, phenomena, or activities from …

Apoptosis + Medicine

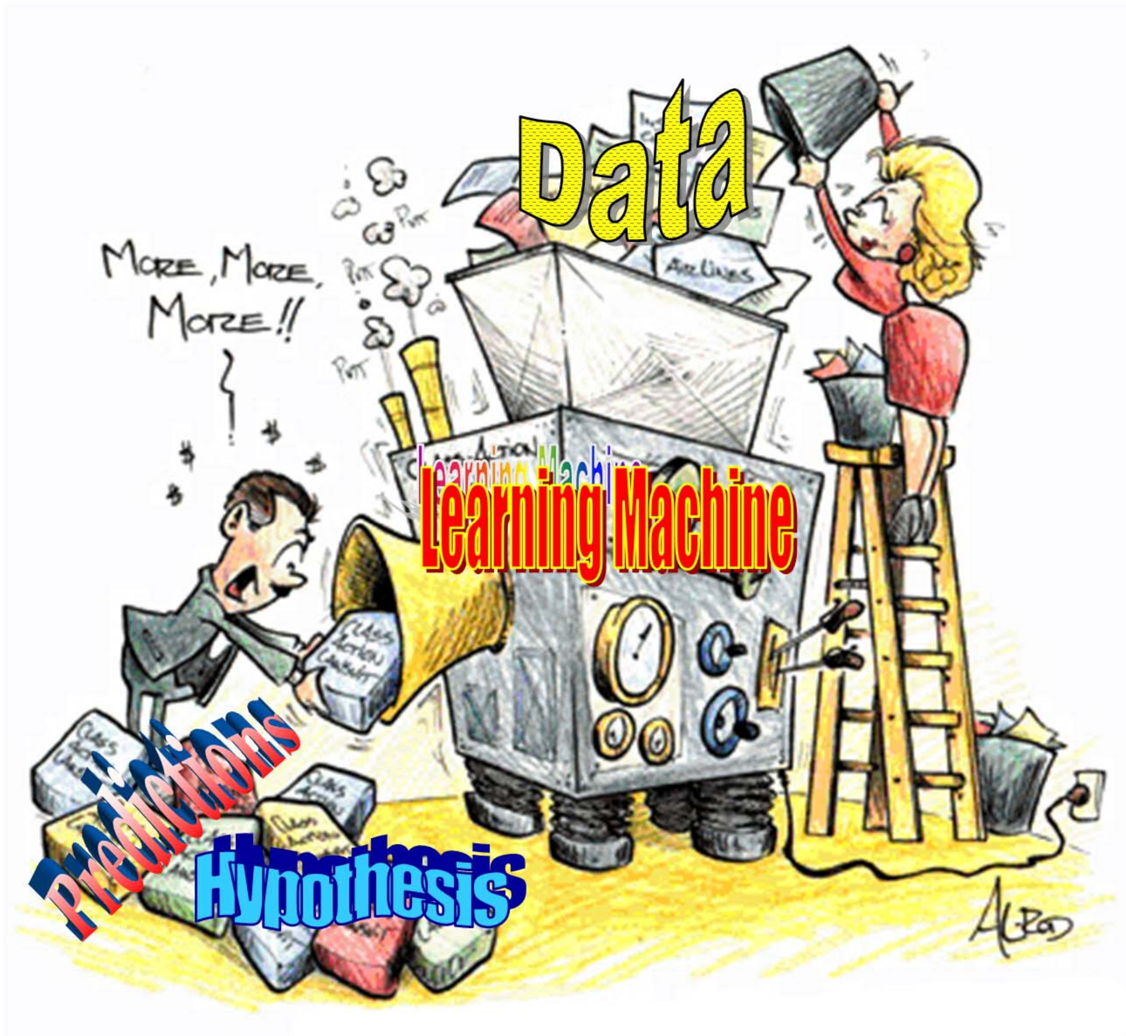
Grammatical rules
Manufacturing procedures
Natural laws
…

Inference:
what does this mean?
Any similar article?
…





Machine Learning (ML)

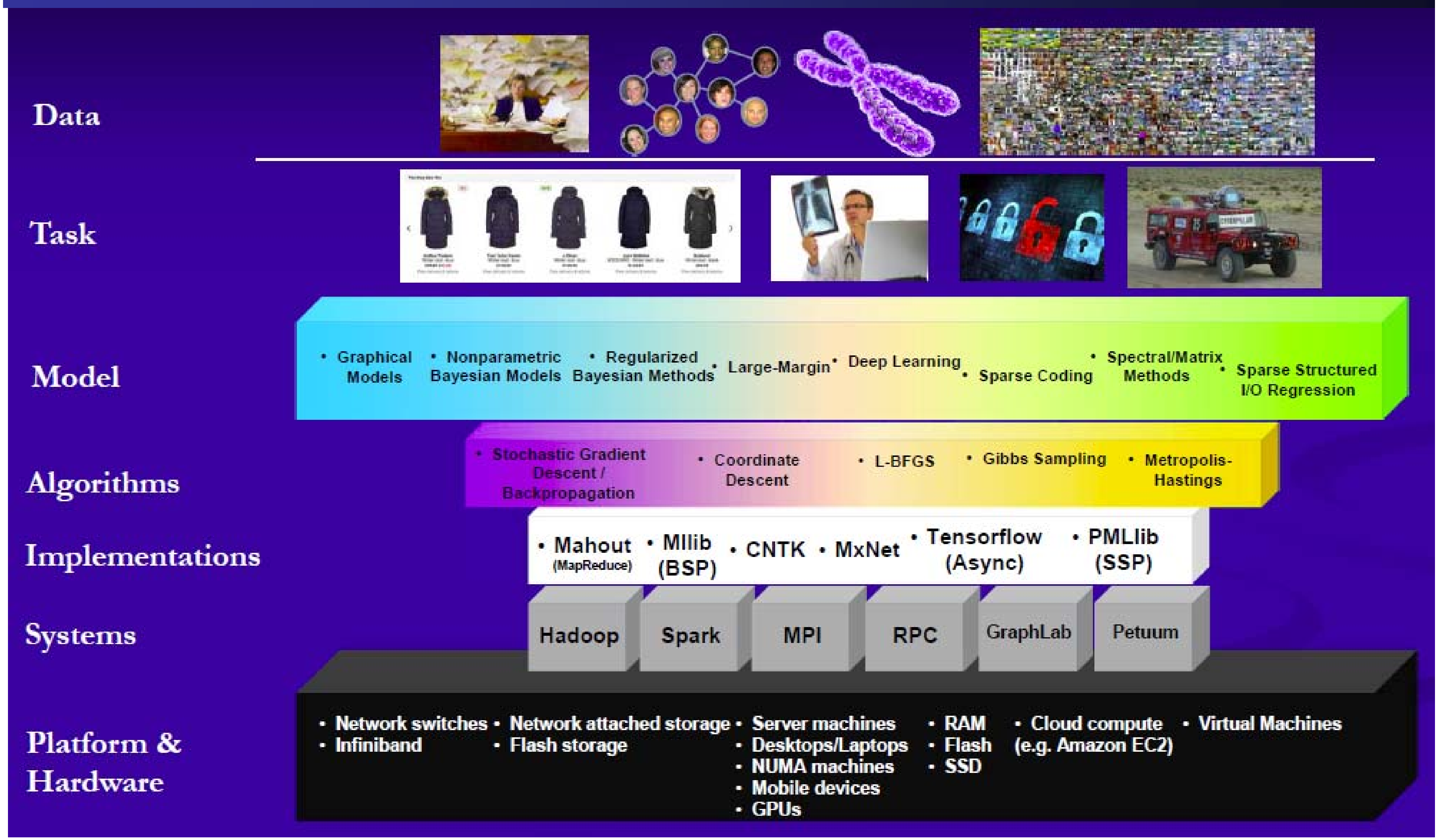


A short definition

- Study of **algorithms** and **systems** that
  - improve their performance P
  - at some task T
  - with experience E

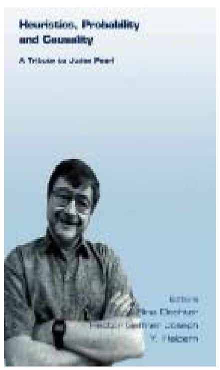
well-defined learning task: <P,T,E>

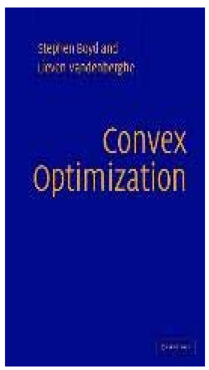
Elements of Modern ML

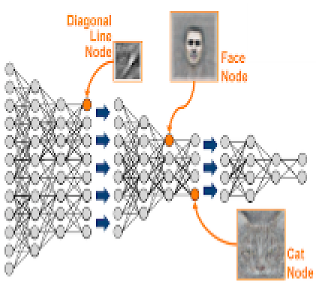


ML methodologies, system paradigms, & hardware infrastructure

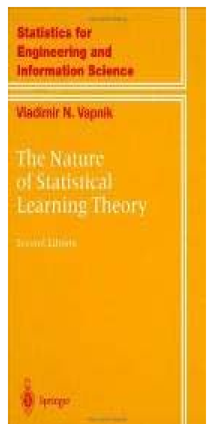
- New mathematical tools

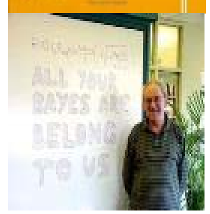




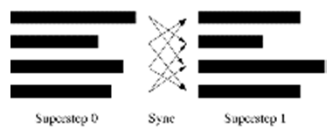
- New theory and algorithms



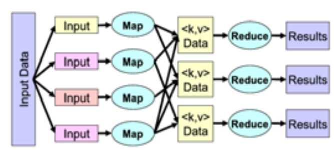



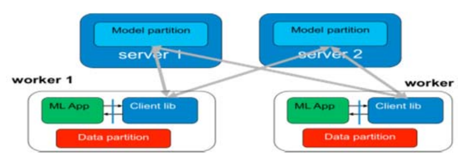
- New system architecture

BSP



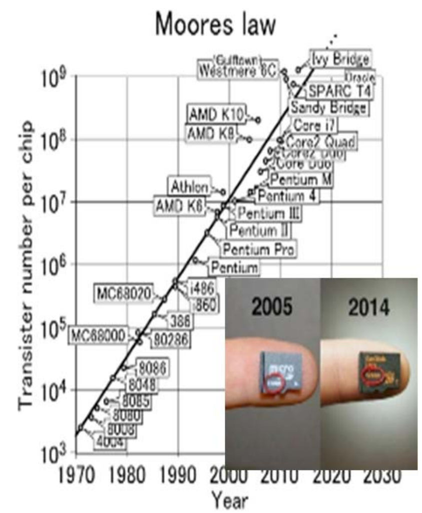
MapReduce



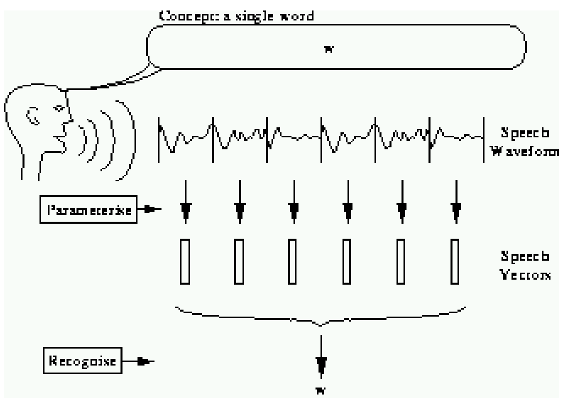



Parameter Server and SSP

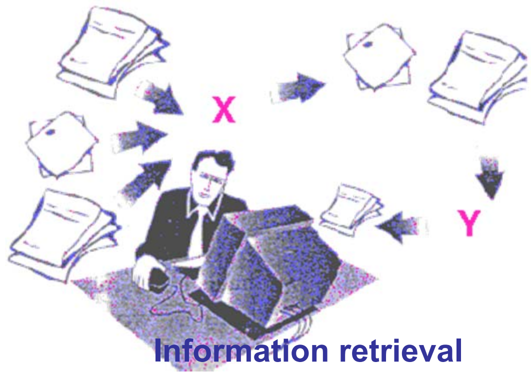
- Moore's Law



Where Machine Learning is being used or can be useful?

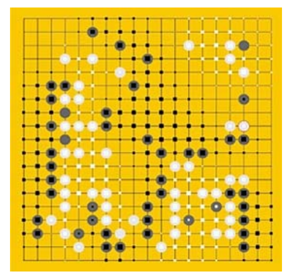


Speech recognition



Information retrieval

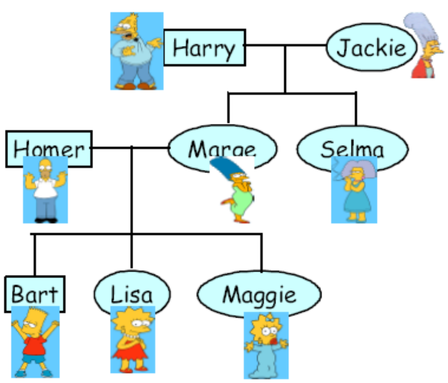
Computer vision



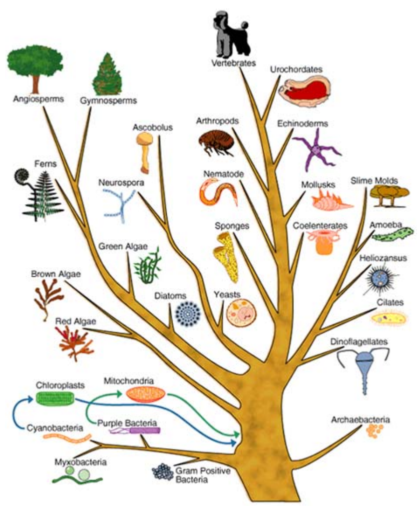
Games

Robotic control

Planning

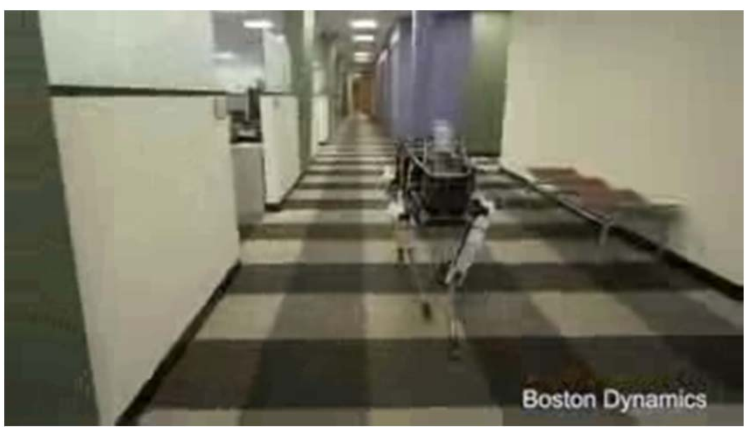


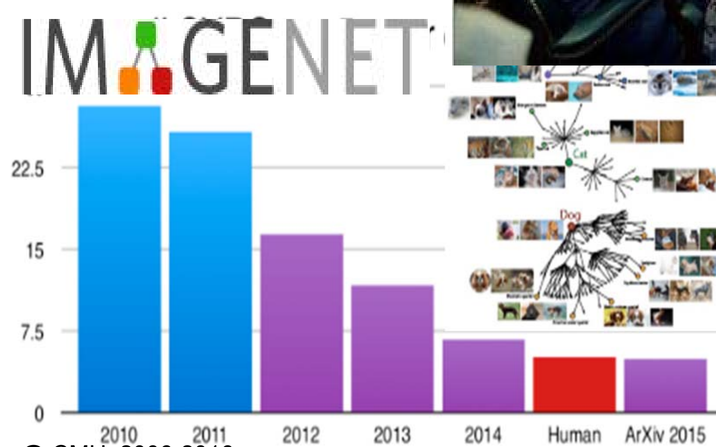
Pedigree



Evolution

Amazing Breakthroughs





Paradigms of Machine Learning

- Supervised Learning
  - Given $D = \{\mathbf{X}_i, \mathbf{Y}_i\}$, learn $f(\cdot): \mathbf{Y}_i = f(\mathbf{X}_i)$, s.t. $D^{\text{new}} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$
- Unsupervised Learning
  - Given $D = \{\mathbf{X}_i\}$, learn $f(\cdot): \mathbf{Y}_i = f(\mathbf{X}_i)$, s.t. $D^{\text{new}} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$
- Semi-supervised Learning
- Reinforcement Learning
  - Given $D = \{\text{env, actions, rewards, simulator/trace/real game}\}$

    learn $\begin{matrix} \text{policy}: e, r \rightarrow a \\ \text{utility}: a, e \rightarrow r \end{matrix}$, s.t. $\{\text{env, new real game}\} \Rightarrow a_1, a_2, a_3 \ldots$
- Active Learning
  - Given $D \sim G(\cdot)$, learn $D^{\text{new}} \sim G'(\cdot)$ and $f(\cdot)$, s.t. $D^{\text{all}} \Rightarrow G'(\cdot), \text{policy}, \{\mathbf{Y}_j\}$
- Transfer learning
- Deep xxx …

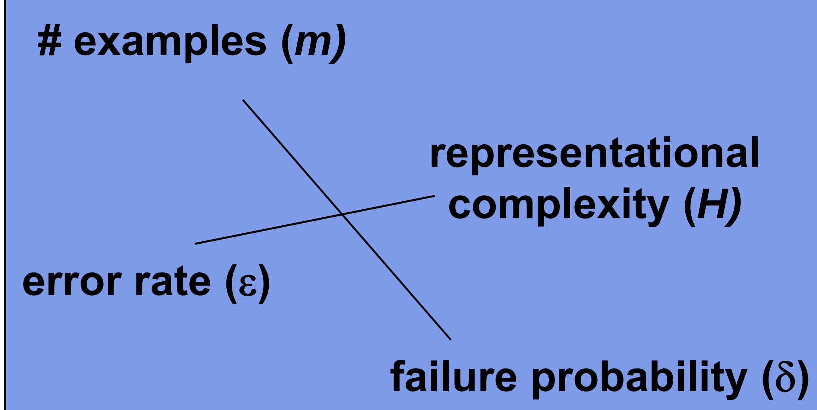
Machine Learning - Theory

For the learned F(; θ)

- Consistency (value, pattern, …)
- Bias versus variance
- Sample complexity
- Learning rate
- Convergence
- Error bound
- Confidence
- Stability
- …

PAC Learning Theory

(supervised concept learning)



$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Why machine learning?

1B+ USERS
30+ PETABYTES

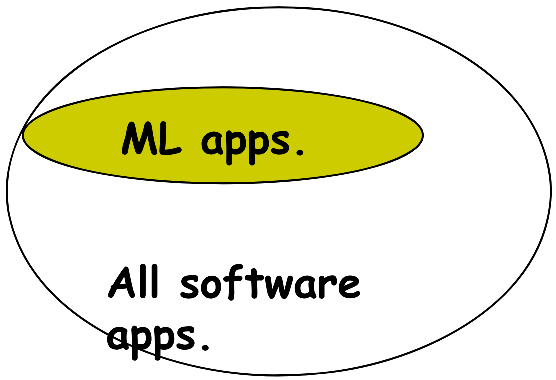
32 million pages

100+ hours video uploaded every minute

645 million users
500 million tweets / day

Growth of Machine Learning

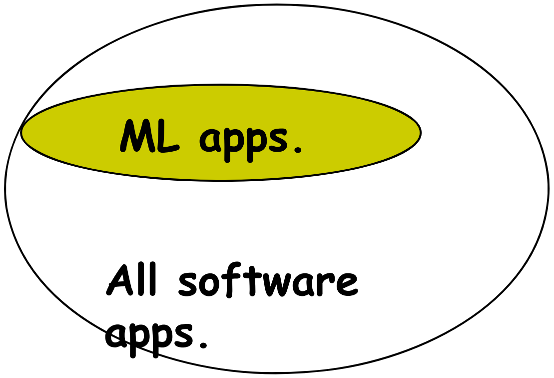
- Machine learning already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing (why?)

Growth of Machine Learning

- Machine learning already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing
  - Improved machine learning algorithms
  - Increased data capture, networking
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for **self-customization to user, environment**

Summary:
What is Machine Learning

Machine Learning seeks to develop theories and computer systems for

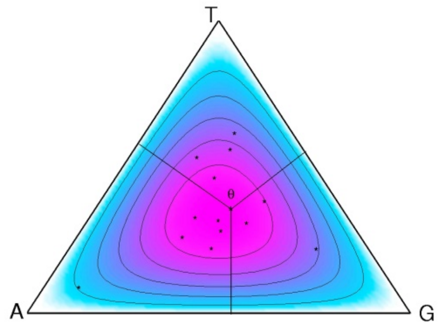
- **representing;**
- **classifying, clustering, recognizing, organizing;**
- **reasoning under uncertainty;**
- **predicting;**
- **and reacting to**
- **…**

complex, real world data, based on the system's own experience with data, and (hopefully) under a unified model or mathematical framework, that

- **can be formally characterized and analyzed**
- **can take into account human prior knowledge**
- **can generalize and adapt across data and domains**
- **can operate automatically and autonomously**
- **and can be interpreted and perceived by human.**

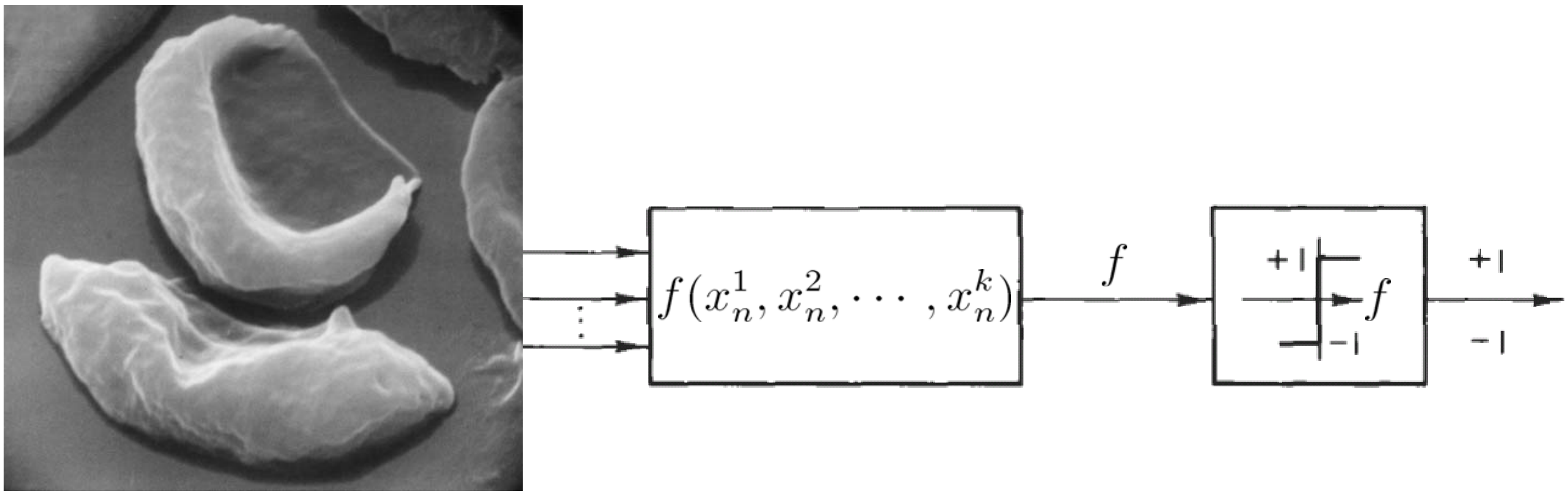
Inference
Prediction
Decision-Making under uncertainty
…

→ Statistical Machine Learning
→ Function Approximation: $F(\ |\theta)?$
→ Density Estimation



Classification

- sickle-cell anemia

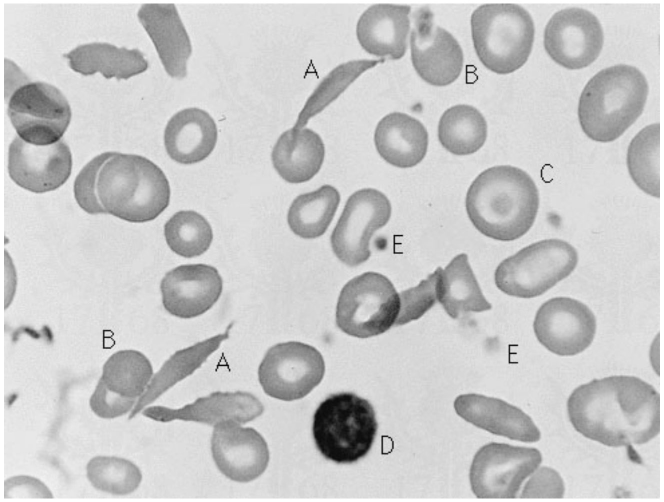


Function Approximation

- **Setting**:
  - Set of possible instances X
  - Unknown target function $f: X \rightarrow Y$
  - Set of function hypotheses $H = \{ h \mid h: X \rightarrow Y \}$

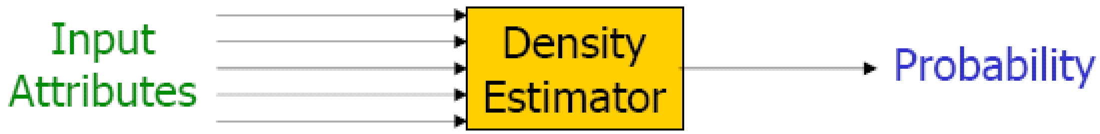
- **Given**:
  - Training examples $\{<x_i, y_i>\}$ of unknown target function f

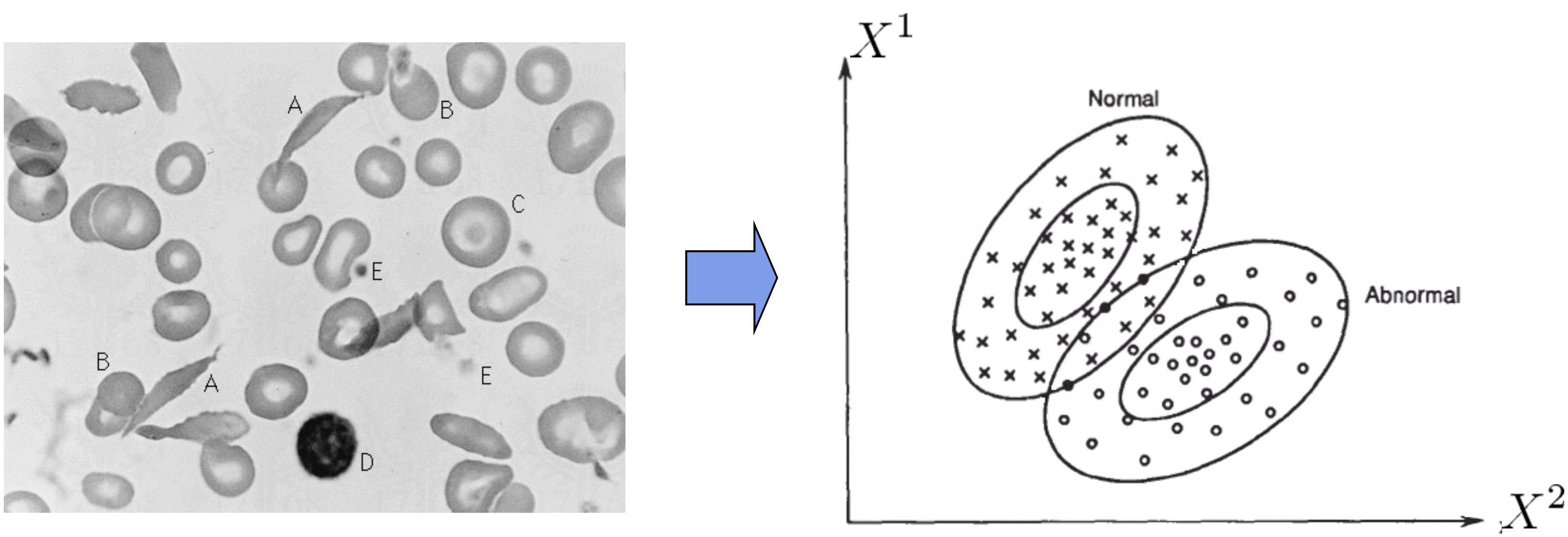
- **Determine**:
  - Hypothesis $h \in H$ that best approximates f



Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a **Probability**





Basic Probability Concepts

- A *sample space* S is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite.)

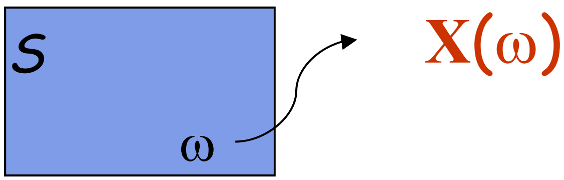
  - E.g., S may be the set of all possible outcomes of a dice roll: $S \equiv \{1,2,3,4,5,6\}$

  - E.g., S may be the set of all possible nucleotides of a DNA site: $S \equiv \{A, T, C, G\}$

  - E.g., S may be the set of all possible positions time-space positions of a aircraft on a radar screen: $S \equiv \{0, R_{max}\} \times \{0, 360^{o}\} \times \{0, +\infty\}$

Random Variable

- A *random variable* is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)



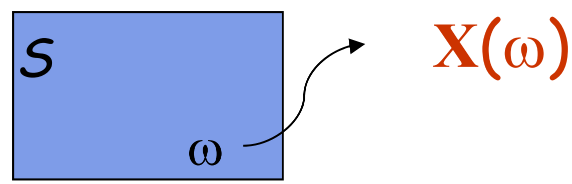
  - Discrete r.v.:
    - The outcome of a dice-roll
    - The outcome of reading a nt at site *i*: X_i
  - Binary event and indicator variable:
    - Seeing an "A" at a site ⇒ X=1, o/w X=0.
    - This describes the true or false outcome a random *event*.
    - Can we describe richer outcomes in the same way? (i.e., X=1, 2, 3, 4, for being A, C, G, T) --- think about what would happen if we take expectation of X.
  - Unit-Base Random vector

    $X_i=[X_i^A, X_i^T, X_i^G, X_i^C]'$, $X_i=[0,0,1,0]'$ ⇒ seeing a "G" at site *i*

  - Continuous r.v.:
    - The outcome of **recording** the **true** location of an aircraft: X_{true}
    - The outcome of **observing** the **measured** location of an aircraft X_{obs}

Random Variable

- Notational convention



  - Univariate

  - Multivariate (random vector)

Discrete Prob. Distribution

- (In the discrete case), a probability distribution P on S (and hence on the domain of X) is an assignment of a non-negative real number $P(s)$ to each $s \in S$ (or each valid value of x) such that $\Sigma_{s \in S} P(s)=1$. $(0 \leq P(s) \leq 1)$
  - intuitively, $P(s)$ corresponds to the *frequency* (or the likelihood) of getting s in the experiments, if repeated many times
  - call $\theta_s = P(s)$ the *parameters* in a discrete probability distribution

- A probability distribution on a sample space is sometimes called a *probability model*, in particular if several different distributions are under consideration
  - write models as M_1, M_2, probabilities as $P(X|M_1)$, $P(X|M_2)$
  - e.g., M_1 may be the appropriate prob. dist. if X is from "fair dice", M_2 is for the "loaded dice".
  - M is usually a two-tuple of {dist. family, dist. parameters}

Discrete Distributions

- Bernoulli distribution: Ber(p)

$$P(x)=\begin{cases}1-\theta & \text{if } x=0\\ \theta & \text{if } x=1\end{cases} \quad\Rightarrow\quad P(x)=p^x(1-p)^{1-x}$$

- Multinomial distribution: Mult($1,\theta$)
  - Multinomial (indicator) variable:

$$X=\begin{bmatrix}X^1\\X^2\\X^3\\X^4\\X^5\\X^6\end{bmatrix}, \quad \text{where} \quad \begin{array}{l} X^j=[0,1], \quad \text{and} \quad \sum_{j\in[1,\ldots,6]} X^j=1 \\ X^j=1 \ \text{w.p.}\ \theta_j, \quad \sum_{j\in[1,\ldots,6]}\theta_j=1 \ . \end{array}$$

$$p(x(j))=P\left(\{X^j=1, \text{where } j \text{ index the dice-face}\}\right)$$
$$=\theta_j=\theta_A^{x^A}\times\theta_C^{x^C}\times\theta_G^{x^G}\times\theta_T^{x^T}=\prod_k \theta_k^{x^k}=\theta^x$$

Discrete Distributions

- Multinomial distribution: Mult(n, θ)
  - Count variable:

$$X = \begin{bmatrix} x^1 \\ \vdots \\ x^K \end{bmatrix}, \qquad \text{where} \sum_j x^j = n$$

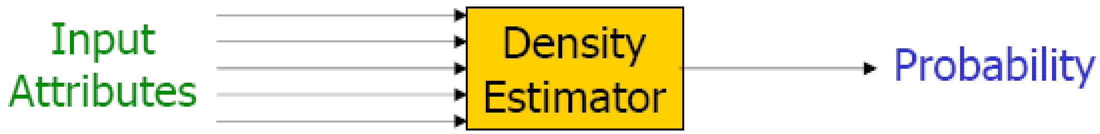
$$p(x) = \frac{n!}{x^1! x^2! \cdots x^K!} \theta_1^{x^1} \theta_2^{x^2} \cdots \theta_K^{x^K} = \frac{n!}{x^1! x^2! \cdots x^K!} \theta^x$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a **Probability**

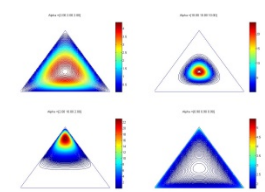


- Often know as *parameter estimation* if the distribution form is specified
  - Binomial, Gaussian …

- Three important issues:
  - Nature of the data (iid, correlated, …)
  - Objective function (MLE, MAP, …)
  - Algorithm (simple algebra, gradient methods, EM, …)
  - Evaluation scheme (likelihood on test data, predictability, consistency, …)

Density Estimation Schemes

Data	Learn parameters	Algorithm	Score param
$(x_1^1, \ldots, x_1^n)$	Maximum likelihood	Analytical	10^{-5}
$(x_2^1, \ldots, x_2^n)$	Bayesian	Gradient	10^{-3}
$\ldots$	Conditional likelihood	EM	10^{-15}
$(x_M^1, \ldots, x_M^n)$	Margin	Sampling	$\ldots$
	…	…	



Parameter Learning from *iid* Data

- Goal: estimate distribution parameters θ from a dataset of N independent, identically distributed (*iid*), fully observed, training cases

$$D = \{x_1, \ldots, x_N\}$$

- Maximum likelihood estimation (MLE)
  1. One of the most common estimators
  2. With iid and full-observability assumption, write $L(\theta)$ as the likelihood of the data:

$$\begin{aligned} L(\theta) &= P(x_1, x_2, \ldots, x_N; \theta) \\ &= P(x_1;\theta)P(x_2;\theta), \ldots, P(x_N;\theta) \\ &= \prod_{i=1}^{N} P(x_i;\theta) \end{aligned}$$

  3. pick the setting of parameters most likely to have generated the data we saw:

$$\theta^* = \arg\max_\theta L(\theta) = \arg\max_\theta \log L(\theta)$$

Example: Bernoulli model

- Data:
  - We observed *N* ***iid*** coin tossing: *D*={1, 0, 1, …, 0}
- Representation:

  Binary r.v: $$x_n = \{0,1\}$$

- Model:

$$P(x) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}$$

- How to write the likelihood of a single observation x_i?

$$P(x_i) = \theta^{x_i} (1-\theta)^{1-x_i}$$

- The likelihood of dataset *D*={$x_1, \ldots, x_N$}:

$$P(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \left(\theta^{x_i} (1-\theta)^{1-x_i} \right) = \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}} (1-\theta)^{\#\text{tails}}$$

Maximum Likelihood Estimation

- Objective function:

$$\ell(\theta; D) = \log P(D \mid \theta) = \log \theta^{n_h}(1-\theta)^{n_t} = n_h \log \theta + (N - n_h)\log(1-\theta)$$

- We need to maximize this w.r.t. θ

- Take derivatives wrt θ

$$\frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \widehat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \widehat{\theta}_{MLE} = \frac{1}{N}\sum_i x_i$$

Frequency as sample mean

- Sufficient statistics
  - The counts, n_h, where $n_h = \sum_i x_i$, are **sufficient statistics** of data *D*

$p(D|\theta)$

Bayesian Parameter Estimation

- Treat the distribution parameters θ also as a *random variable*
- The *a posteriori* distribution of θ after seem the data is:

$$p(\theta | D) = \frac{p(D | \theta) p(\theta)}{p(D)} = \frac{p(D | \theta) p(\theta)}{\int p(D | \theta) p(\theta) d\theta}$$

This is Bayes Rule

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

The prior p(.) encodes our prior knowledge about the domain

Overfitting

- Recall that for Bernoulli Distribution, we have

$$\hat{\theta}_{ML}^{head} = \frac{n^{head}}{n^{head} + n^{tail}}$$

- What if we tossed too few times so that we saw zero head?

  We have $\hat{\theta}_{ML}^{head} = 0,$ and we will predict that the probability of seeing a head next is zero!!!

- The rescue: *"smoothing"*
  - Where n' is know as the pseudo- (imaginary) count

$$\hat{\theta}_{ML}^{head} = \frac{n^{head} + n'}{n^{head} + n^{tail} + n'}$$

  - But can we make this more formal?

Frequentist Parameter Estimation

Two people with different priors $p(\theta)$ will end up with different estimates $p(\theta|D)$.

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

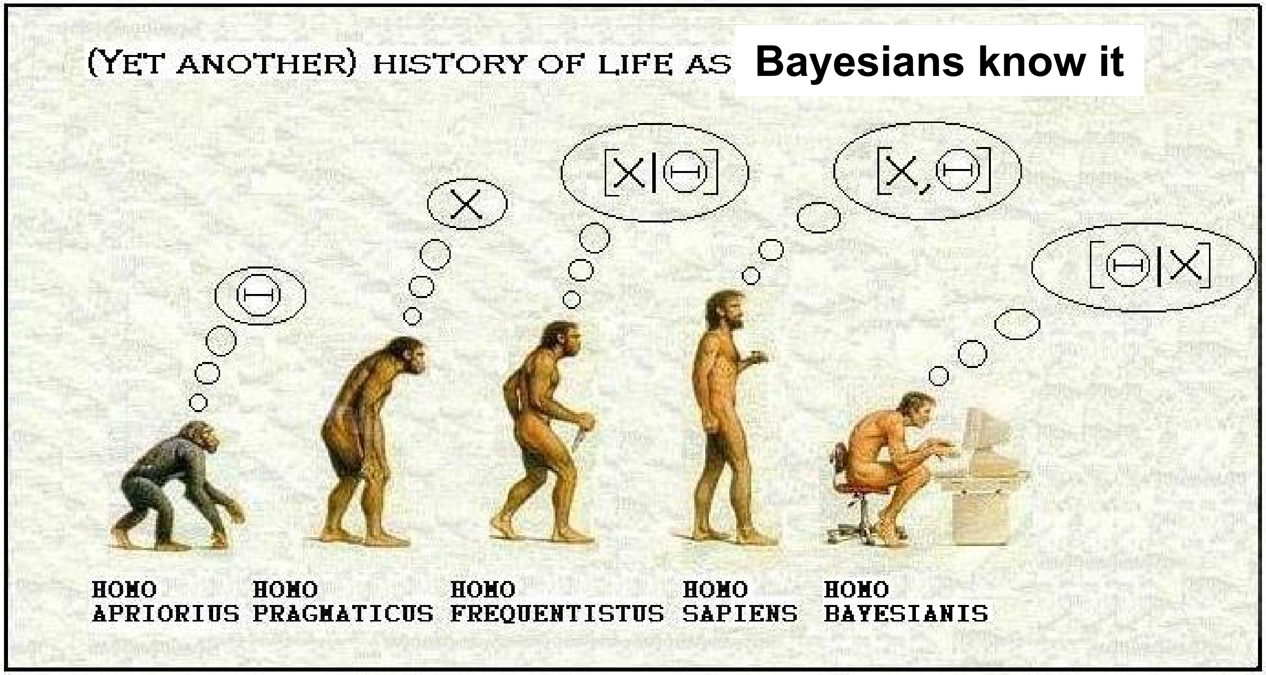
- Frequentists dislike this "subjectivity".
- Frequentists think of the parameter as a fixed, unknown constant, not a random variable.
- Hence they have to come up with different "objective" **estimators** (ways of computing from data), instead of using Bayes' rule.
  - These estimators have different properties, such as being "unbiased", "minimum variance", etc.
  - The maximum likelihood estimator, is one such estimator.

Discussion

θ or $p(\theta)$, this is the problem!

Discussion

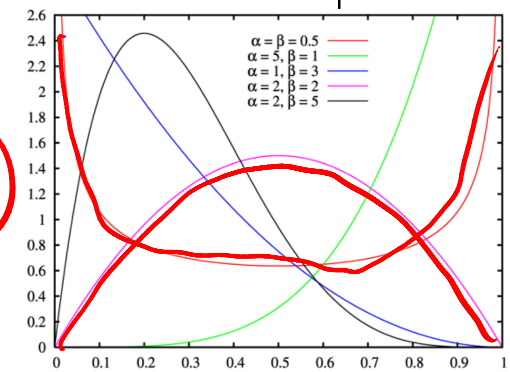
θ **or** $p(\theta)$**, this is the problem!**



Bayesian estimation for Bernoulli

- Beta distribution:

$$P(\theta;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}=B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



  - When x is discrete $\Gamma(x+1)=x\Gamma(x)=x!$

- Posterior distribution of θ:

$$P(\theta\mid x_1,\ldots,x_N)=\frac{p(x_1,\ldots,x_N\mid\theta)p(\theta)}{p(x_1,\ldots,x_N)}\propto\theta^{n_h}(1-\theta)^{n_t}\times\theta^{\alpha-1}(1-\theta)^{\beta-1}=\theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

  - Notice the isomorphism of the posterior to the prior,
  - such a prior is called a **conjugate prior**
  - α *and* β are hyperparameters (parameters of the prior) and correspond to the number of "virtual" heads/tails (pseudo counts)

Bayesian estimation for Bernoulli, con'd

- Posterior distribution of θ:

$$P(\theta \mid x_1,\ldots,x_N) = \frac{p(x_1,\ldots,x_N \mid \theta)\, p(\theta)}{p(x_1,\ldots,x_N)} \propto \theta^{n_h}(1-\theta)^{n_t} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

- Maximum *a posteriori* (MAP) estimation:

$$\theta_{MAP} = \arg\max_{\theta} \log P(\theta \mid x_1,\ldots,x_N)$$

- Posterior mean estimation:

$$\theta_{Bayes} = \int \theta p(\theta \mid D) d\theta = C\int \theta \times \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1} d\theta = \frac{n_h+\alpha}{N+\alpha+\beta}$$

Bata parameters can be understood as pseudo-counts

- Prior strength: A=α+β
  - A can be interoperated as the size of an imaginary data set from which we obtain the **pseudo-counts**

Effect of Prior Strength

- Suppose we have a uniform prior ($\alpha=\beta=1/2$), and we observe $\vec{n} = (n_h = 2, n_t = 8)$
- Weak prior A = 2. Posterior prediction:

$$p(x = h \mid n_h = 2, n_t = 8, \vec{\alpha} = \vec{\alpha}' \times 2) = \frac{1+2}{2+10} = 0.25$$

- Strong prior A = 20. Posterior prediction:

$$p(x = h \mid n_h = 2, n_t = 8, \vec{\alpha} = \vec{\alpha}' \times 20) = \frac{10+2}{20+10} = 0.40$$

- However, if we have enough data, it washes away the prior. e.g., $\vec{n} = (n_h = 200, n_t = 800)$. Then the estimates under weak and strong prior are $\frac{1+200}{2+1000}$ and $\frac{10+200}{20+1000}$, respectively, both of which are close to 0.2

Continuous Prob. Distribution

- A continuous random variable $\mathbf{X}$ can assume any value in an interval on the real line or in a region in a high dimensional space
  - **A random vector** $\mathbf{X}=[x_1, x_2, \ldots, x_n]^T$ usually corresponds to a real-valued measurements of some property, e.g., length, position, …
  - It is not possible to talk about the probability of the random variable assuming a particular value --- $P(\mathbf{x}) = 0$
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval
    - $P(X \in [a,b])$,

      $P(X < x) = P(X \in [-\infty, x])$
    - Arbitrary Boolean combination of basic propositions

Continuous Prob. Distribution

- The probability of the random variable assuming a value within some given interval from *a* to *b* is defined to be the area under the graph of the probability density function between *a* and *b.*

  - Probability mass: $P(X \in [a,b]) = \int_a^b p(x)dx,$

    note that $\int_{-\infty}^{+\infty} p(x)dx = 1.$

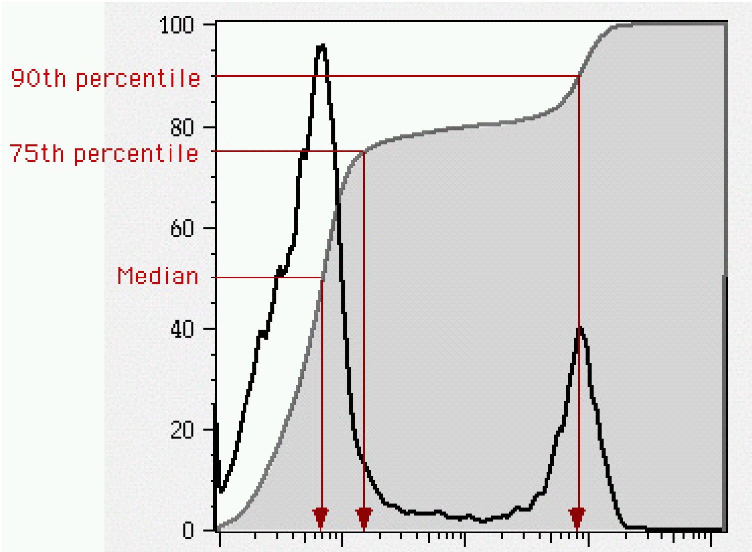
  - Cumulative distribution function (CDF):

$$P(x) = P(X < x) = \int_{-\infty}^{x} p(x')dx'$$

  - Probability density function (PDF):

$$p(x) = \frac{d}{dx}P(x)$$

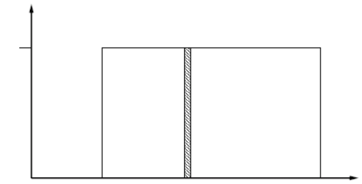
$$\int_{-\infty}^{+\infty} p(x)dx = 1; \quad p(x) > 0, \forall x$$



Car flow on Liberty Bridge (cooked up!)

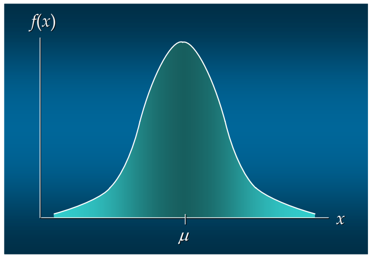
Continuous Distributions

- Uniform Probability Density Function



$$p(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$

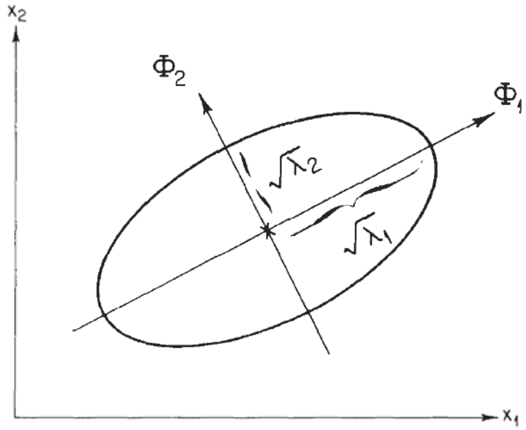
- Normal (Gaussian) Probability Density Function



$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

  - The distribution is symmetric, and is often illustrated as a bell-shaped curve.
  - Two parameters, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.
  - The highest point on the normal curve is at the mean, which is also the median and mode.
  - The mean can be any numerical value: negative, zero, or positive.

- Multivariate Gaussian



$$p(X;\vec{\mu},\Sigma) = \frac{1}{\left(\sqrt{2\pi}\right)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(X-\vec{\mu})^T\Sigma^{-1}(X-\vec{\mu})\right\}$$

Example 2: Gaussian density

- Data:
  - We observed N ***iid*** real samples:
    D={-0.1, 10, 1, -5.2, …, 3}
- Model: $$P(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(x-\mu)^2 / 2\sigma^2\right\}$$
- Log likelihood:

$$\ell(\theta; D) = \log P(D \mid \theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{\sigma^2}$$

- MLE: take derivative and set to zero:

$$\frac{\partial \ell}{\partial \mu} = (1/\sigma^2)\sum_n (x_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n (x_n-\mu)^2$$

$\Rightarrow$

$$\mu_{MLE} = \frac{1}{N}\sum_n (x_n)$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_n (x_n - \mu_{ML})^2$$

MLE for a multivariate-Gaussian

- It can be shown that the MLE for μ *and* Σ is

$$x_n = \begin{pmatrix} x_n^1 \\ x_n^2 \\ \vdots \\ x_n^K \end{pmatrix}$$

$$\mu_{MLE} = \frac{1}{N}\sum_n (x_n)$$

$$\Sigma_{MLE} = \frac{1}{N}\sum_n (x_n - \mu_{ML})(x_n - \mu_{ML})^T = \frac{1}{N} S$$

$$X = \begin{pmatrix} --- & x_1^T & --- \\ --- & x_2^T & --- \\ & \vdots & \\ --- & x_N^T & --- \end{pmatrix}$$

where the scatter matrix is

$$S = \sum_n (x_n - \mu_{ML})(x_n - \mu_{ML})^T = \left(\sum_n x_n x_n^T\right) - N\mu_{ML}\mu_{ML}^T$$

- The sufficient statistics are $\Sigma_n x_n$ and $\Sigma_n x_n x_n^T$.
- Note that $X^T X = \Sigma_n x_n x_n^T$ may not be full rank (eg. if $N < D$), in which case Σ_{ML} is not invertible

Bayesian estimation

- Normal Prior:

$$P(\mu) = \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left\{-(\mu-\mu_0)^2 / 2\sigma_0^2\right\}$$

- Joint probability:

$$P(x,\mu) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$
$$\times \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left\{-(\mu-\mu_0)^2 / 2\sigma_0^2\right\}$$

- Posterior:

$$P(\mu \mid x) = \left(2\pi\tilde{\sigma}^2\right)^{-1/2} \exp\left\{-(\mu-\tilde{\mu})^2 / 2\tilde{\sigma}^2\right\}$$

where $\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2}\bar{x} + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2}\mu_0$, and $\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$

Sample mean

Bayesian estimation: unknown μ, known σ

$$\mu_N = \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2}\bar{x} + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2}\mu_0 , \qquad \tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

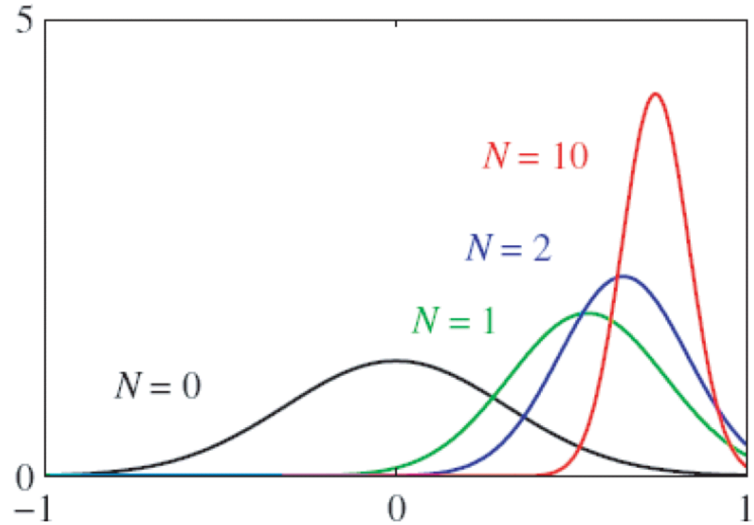
- The posterior mean is a convex combination of the prior and the MLE, with weights proportional to the relative noise levels.
- The precision of the posterior $1/\sigma^2_N$ is the precision of the prior $1/\sigma^2_0$ plus one contribution of data precision $1/\sigma^2$ for each observed data point.

- Sequentially updating the mean
  - $\mu* = 0.8$ (unknown), $(\sigma^2)* = 0.1$ (known)
  - Effect of single data point

$$\mu_1 = \mu_0 + (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} = x - (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}$$

  - Uninformative (vague/ flat) prior, $\sigma^2_0 \to \infty$

$$\mu_N \to \mu_0$$



Summary

- Machine Learning is Cool and Useful!!

- Learning scenarios:
  - Data
  - Objective function
  - Frequentist and Bayesian

- Density estimation
  - Typical discrete distribution
  - Typical continuous distribution (recitation)
  - Conjugate priors

Some suggestions …

How ML facilitates Applications (say, NLP)

NLP

Linguistic Theory: Syntax, Semantics …

- …
- Question Answering
- Machine Translation
- Language Modeling
- POS tagging & Parsing
- Name Entity Recognition
- Sentiment Analysis
- Topic Clustering

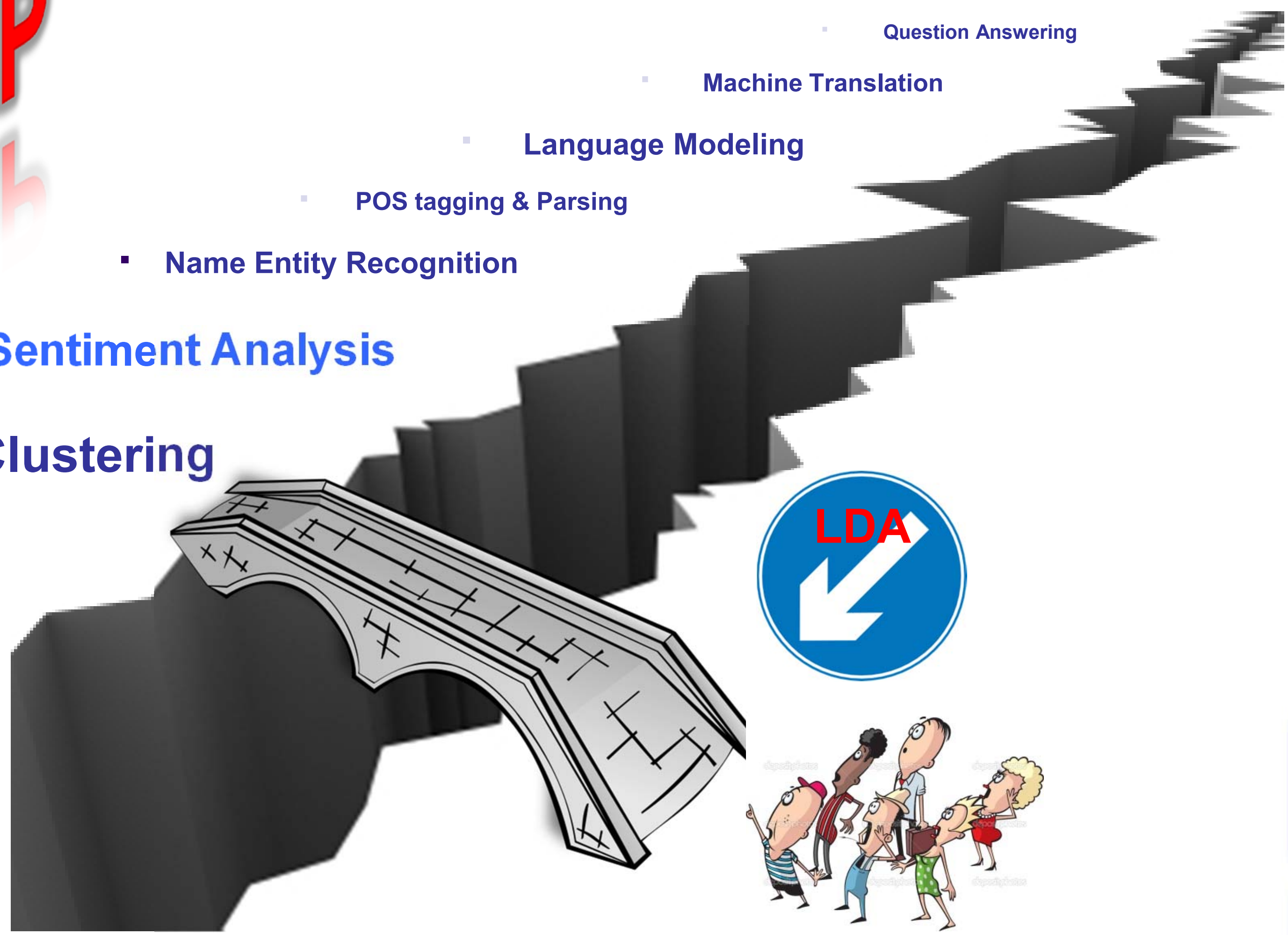
CRF, RNN, SVM, LSA, HMM

ML

One way …

NLP

- …
- Question Answering
- Machine Translation
- Language Modeling
- POS tagging & Parsing
- Name Entity Recognition
- Sentiment Analysis
- Topic Clustering



ML

Maybe highway …?

NLP

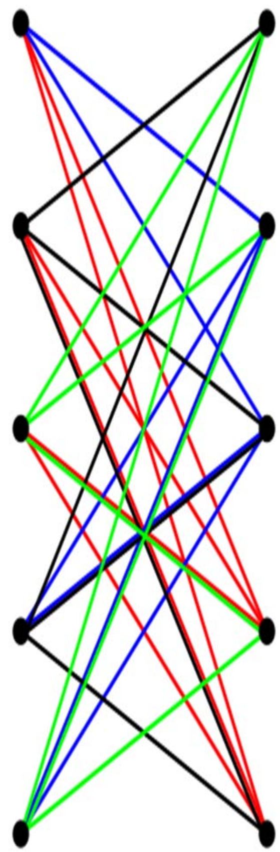
- …
- Question Answering
- Machine Translation
- Language Modeling
- POS tagging & Parsing
- Name Entity Recognition
- Sentiment Analysis
- Topic Clustering

ML


专业深度学习调试
caffe安装 10元
CNN 5元/层
RNN 8元/层
专业

Solution = deep domain knowledge + sounds methodology

- Topic Clustering
- Sentiment Analysis
- POS tagging
- Name Entity Recognition
- Parsing
- Machine Translation
- Question Answering
- …



- Topic models/Latent space models
- Structured input/output predictive models
- Spectrum models
- Deep network models
- Distance metric
- Convex and non-convex optimization algorithms
- Monte Carlo algorithms
- Distributed ML systems
- Consistency/identifiability/convergence theories
- …