Introduction

Jun Zhu

dcszj@mail.tsinghua.edu.cn

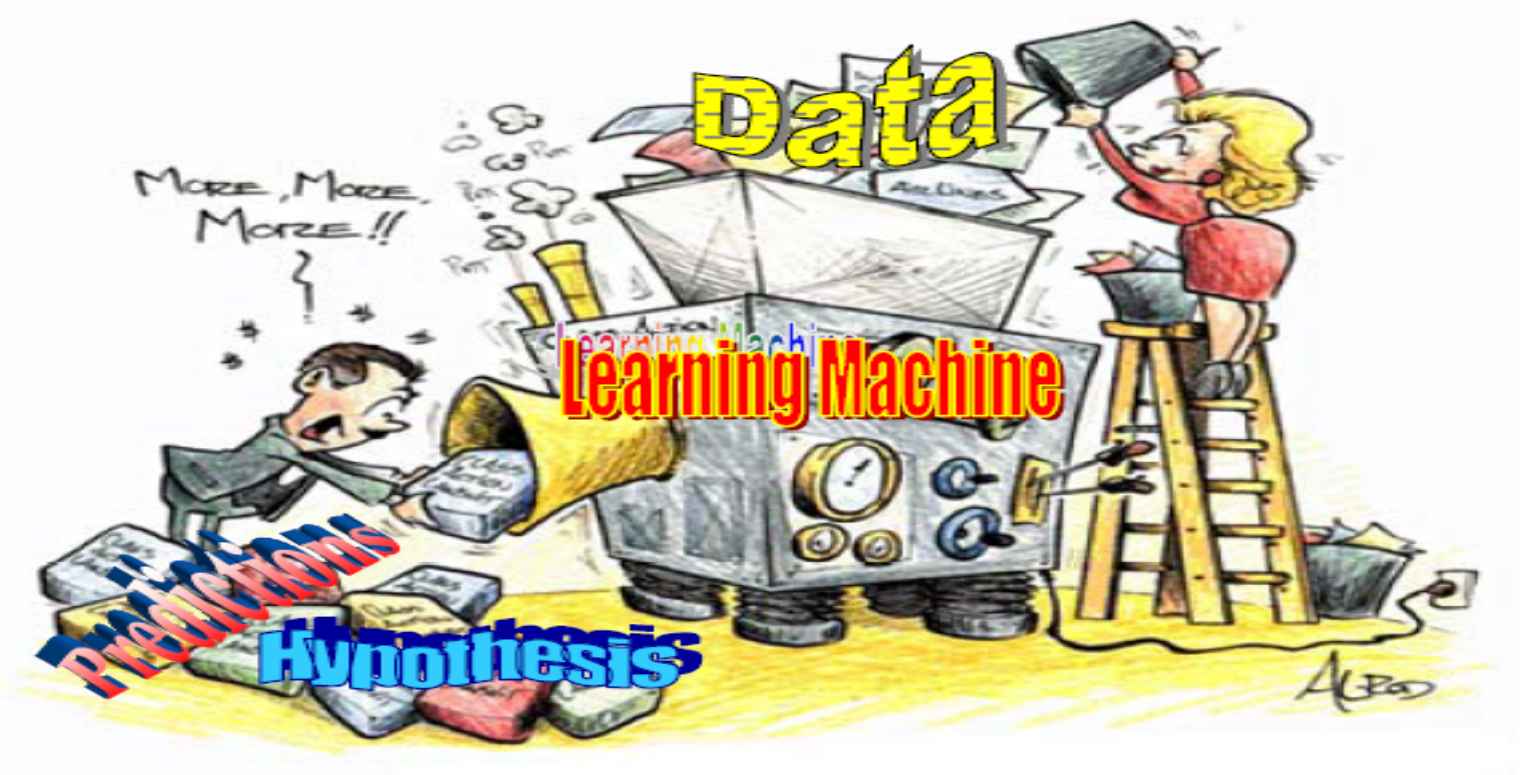
Sate Key Lab of Intelligent Tech. & Systems,

Tsinghua University

What is Machine Learning?

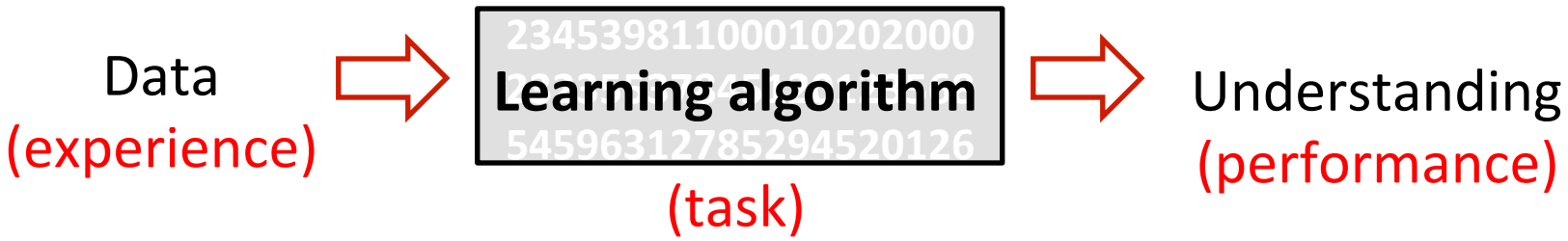


Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that take as input empirical data, and yield patterns or predictions thought to be features of the underlying mechanism that generated the data

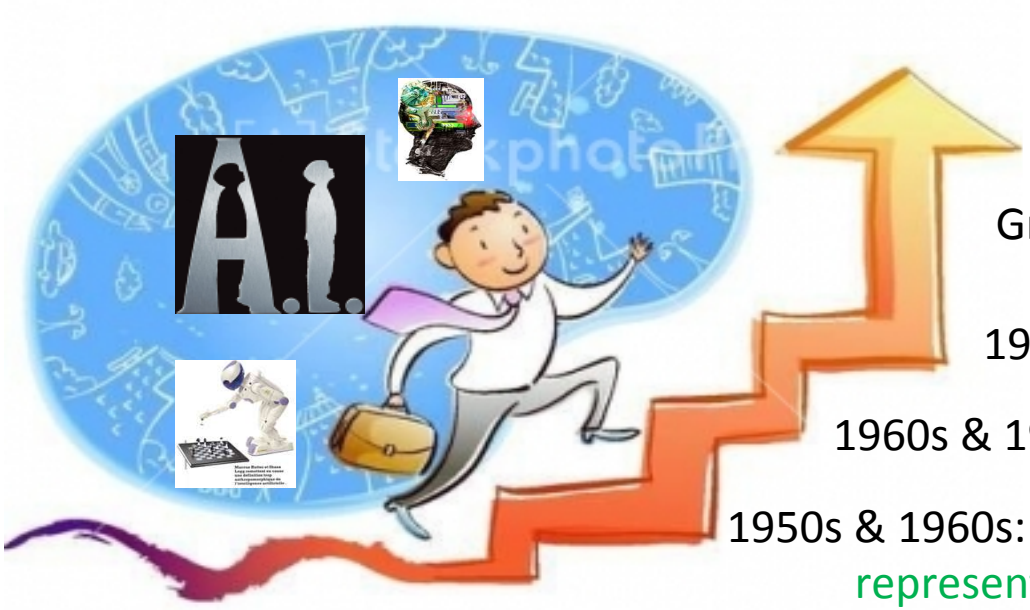


What is machine learning?

- Study of algorithms that
  - (automatically) improve their performance
  - at some task
  - with experience



(Statistical) Machine Learning in AI



since 1990s
Great growth in machine learning

1970s & 1980s: expert systems

1960s & 1970s: success on NLP systems

1950s & 1960s: symbolic reasoning, representation, search

pre-computer age
thoughts on symbolic reasoning



[Judea Pearl, Turing Award 2011]

- For "innovations that enabled remarkable advances in the partnership between humans and machines that is the foundation of Artificial Intelligence (AI)"
- "His work serves as the standard method for handling uncertainty in computer systems, with applications from medical diagnosis, homeland security and genetic counseling to natural language understanding and mapping gene expression data."
- "Modern applications of AI, such as robotics, self-driving cars, speech recognition, and machine translation deal with uncertainty. Pearl has been instrumental in supplying the rationale and much valuable technology that allow these applications to flourish."

Machine learning in Action

- Document classification

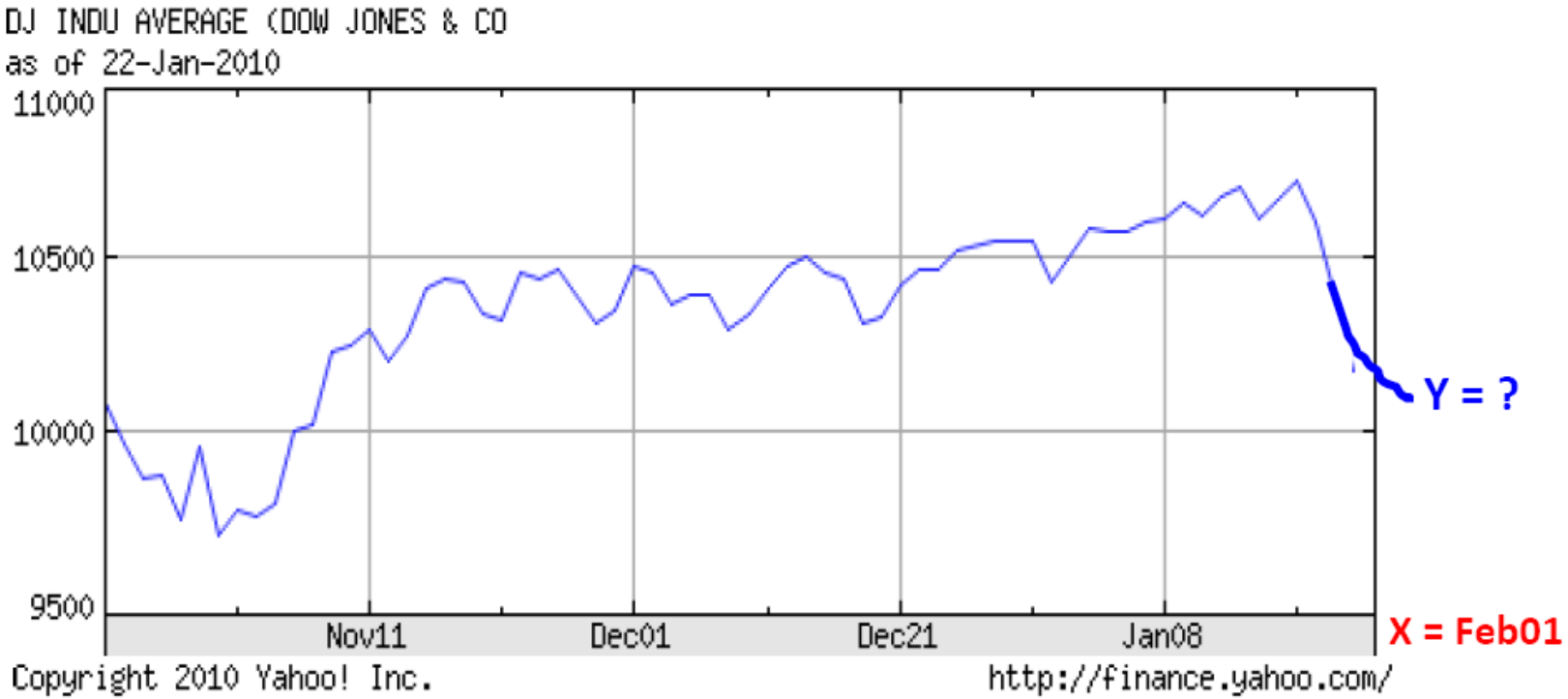


Sports
News
Politics
…

classification

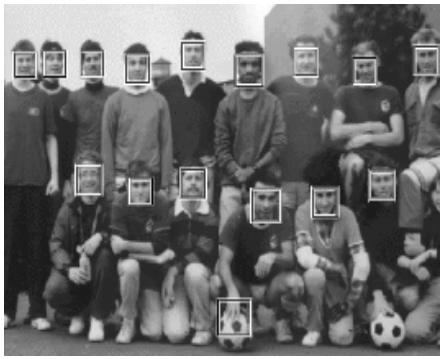
Regression

- Stock market prediction



Computer Vision

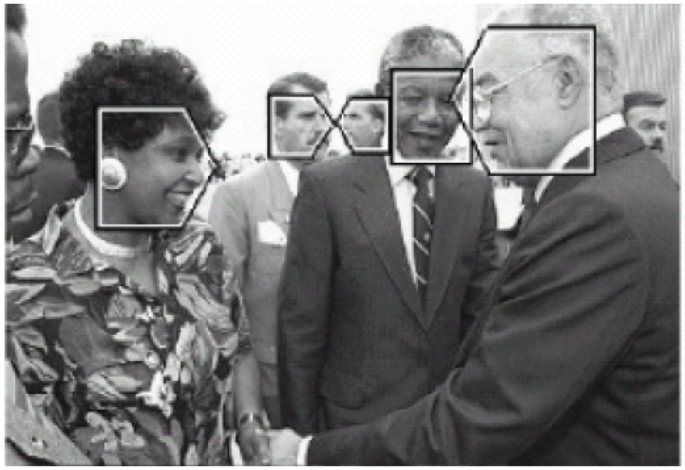
- Face recognition
- Scene understanding
- Action/behavior recognition
- Image tagging and search
- Optical character recognition (OCR)

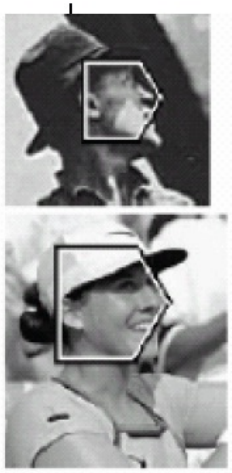






ImageNet Challenge: 1000 categories, 1.2 million images for training





Speech Recognition

- A classic problem in AI, very difficult!
  - "Let's talk about how to wreck a nice beach"
  - small vocabulary is – easy
  - challenges: large vocabulary, noise, accent, semantics







Natural Language Processing

- Machine translation
- Information Extraction
- Information Retrieval, question answering
- Text classification, spam filtering, etc….





Name: Honray For Barney
Price: $12.95
Picture: http://xxxx.jpg
Description: Three Barney board books will

Control

- Cars navigating on their own





  - DAPA urban challenge

  - Tsinghua Mobile Robot V (THMR-V):



Control (cont'd)

- The best helicopter pilot is now a computer!
  - it runs a program that learns how to fly and make acrobatic maneuvers by itself!
  - no taped instructions, joysticks, or things like that …





http://heli.stanford.edu/

Control (cont'd)

- Robot assistant?

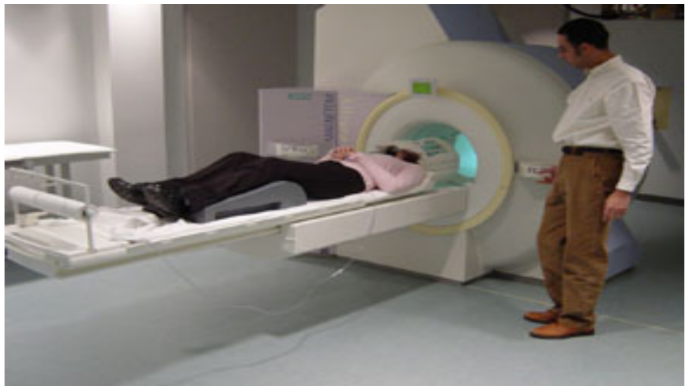


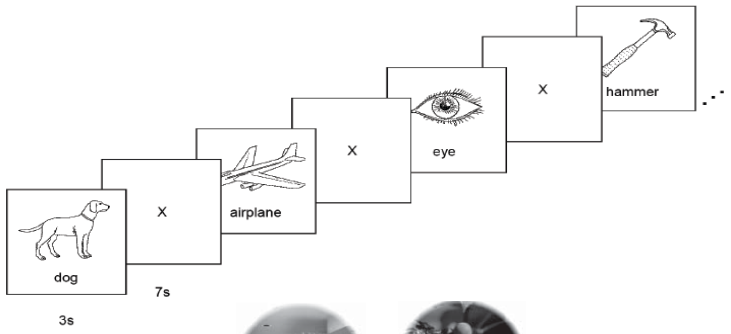


http://stair.stanford.edu/

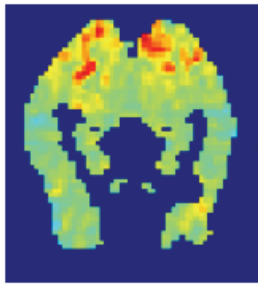
Science

- Decoding thoughts from brain activity









Tool
Animal
…

[Mitchell et al, Science 2008]
[Kay et al., Nature, 2008]

Science (cont'd)

- Bayesian models of inductive learning and reasoning [Tenenbaum et al., Science 2011]
  - Challenge:
    - *How can people generalize well from sparse, noisy, and ambiguous data?*
  - Hypothesis:
    - *If the mind goes beyond the data given, some more abstract background knowledge must generate and delimit the possible hypotheses*
  - Bayesian models make structured abstract knowledge and statistical inference cooperate
  - Examples
    - Word learning [Xu & Tenenbaum, Psychol. Rev. 2007]
    - Causal relation learning [Griffiths & Tenenbaum, 2005]
    - Human feature learning [Austerweil & Griffiths, NIPS 2009]
    - …

J. Tenenbaum et al., "How to grow a mind: Statistics, Structure, and Abstraction". Science 331, 1279 (2011)

More others …

- Many more
  - Natural language processing
  - Speech recognition
  - Computer vision
  - Computational biology
  - Social network analysis
  - Sensor networks
  - Health care
  - Protest ??
  - …

Machine learning in Action

- Machine learning for protest?

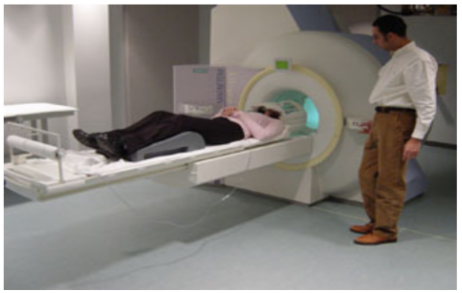


CMU ML students and post-docs at G-20 Pittsburgh Summit 2009

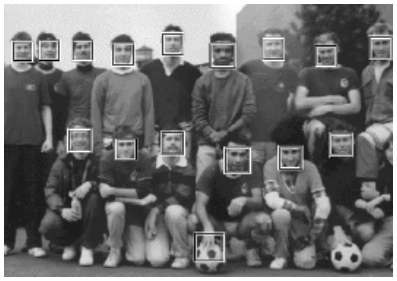
Machine Learning – practice



document classification



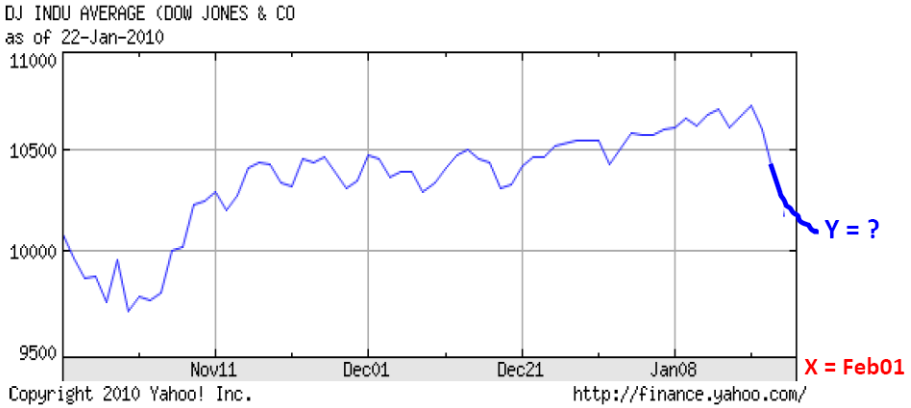
decoding brain signal



face recognition



robot control



stock market prediction

Machine Learning – theory

PAC Learning Theory
(supervised concept learning)

examples (*m*)

representational complexity (*H*)

error rate (ε)

failure probability (δ)

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

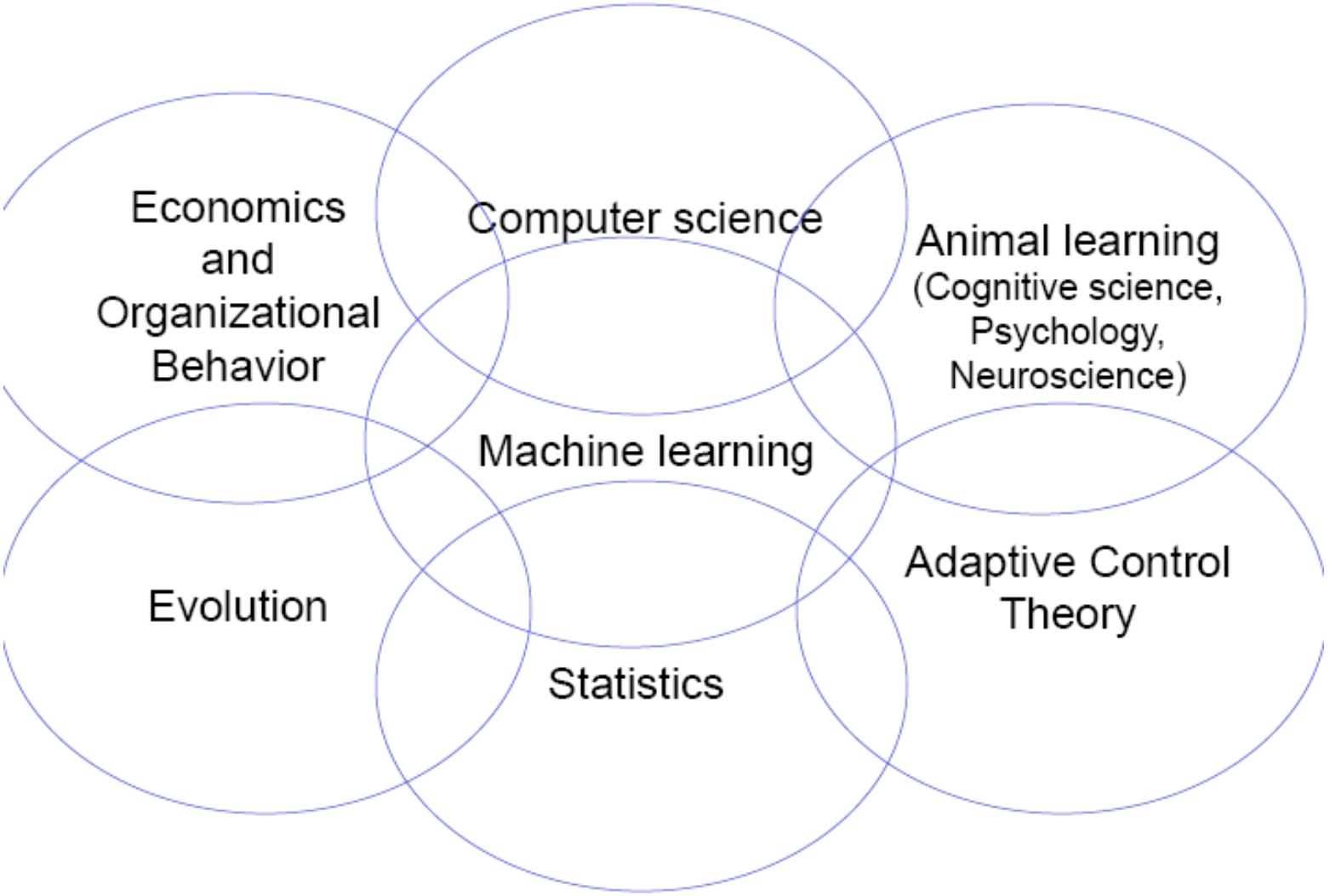
Other theories for
- semi-supervised learning
- reinforcement skill learning
- active learning
- …

… also relating to
- # mistakes during training
- asymptotic performance
- convergence rate
- bias, variance tradeoff
- …

[Leslie G. Valiant, 1984; Turing Award, 2010]

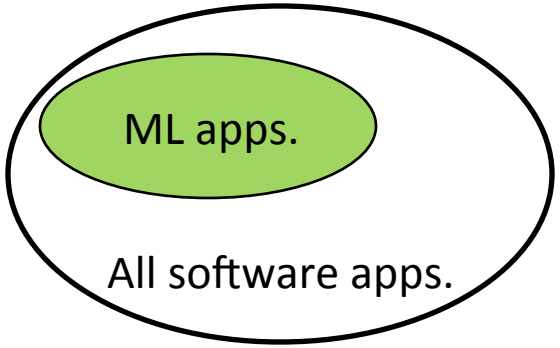


"For transformative contributions to the theory of computation, including the theory of probably approximately correct (PAC) learning, the complexity of enumeration and of algebraic computation, and the theory of parallel and distributed computing."


Economics
and
Organizational
Behavior
Computer science
Animal learning
(Cognitive science,
Psychology,
Neuroscience)
Machine learning
Evolution
Statistics
Adaptive Control
Theory

Growth of Machine Learning in CS

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



- This ML niche is growing (why?)

Growth of Machine Learning in CS

- Machine learning already the preferred approach to
  - Speech recognition, natu
  - Computer vision
  - Medical outcomes analy
  - Robot control
  - …

Huge amount of data …

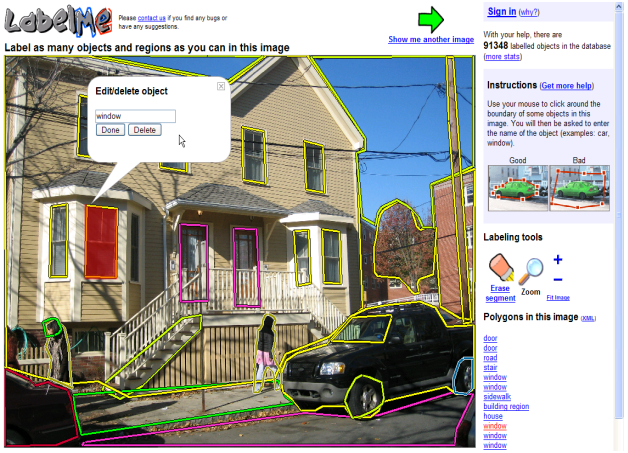
- Web: estimated Google index 45 billion pages
- Transaction data: 5-50 TB/day
- Satellite image feeds: ~1TB/day/satellite
- Biological data: 1-10TB/day/sequencer
- TV: 2TB/day/channel;
- YouTube 4TB/day uploaded
- Photos: 1.5 billion photos/week uploaded

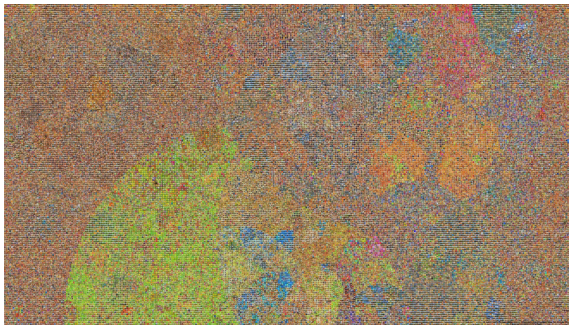
- This ML niche is growing
  - Improved machine learning algorithms
  - Increased data capture, networking, new sensors
  - Software too complex to write by hand
  - Demand for self-customization to user, environment

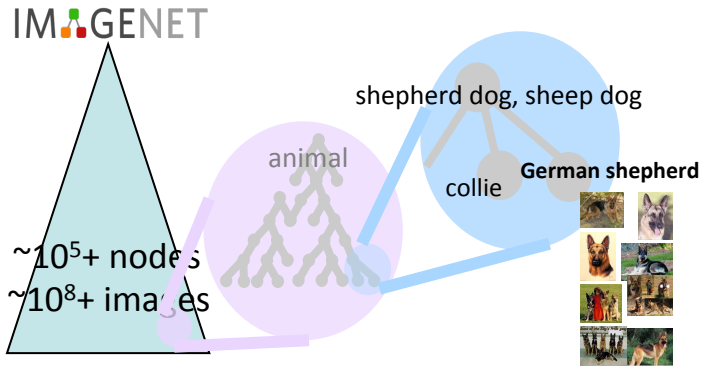
ML has a long way to go …

- Very large-scale learning in rich media









10^5 images 10^{1-2} categories

10^5 images 10^{2-3} categories

10^{6-7} images 10^{3-4} categories

Machine Learning Tasks

- Broad categories
  - Supervised learning
    - Classification, Regression
  - Unsupervised learning
    - Density estimation, Clustering, Dimensionality reduction
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning
  - Transfer learning
  - Many more …

Supervised Learning

- Task: learn a predictive function $h: \mathcal{X} \rightarrow \mathcal{Y}$

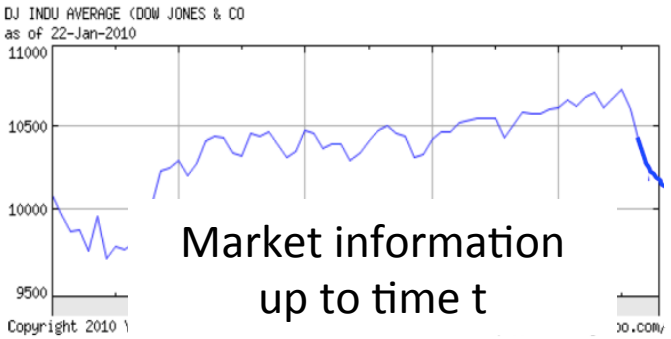
Feature space $\mathcal{X}$





Label space $\mathcal{Y}$

"Sports"
"News"
"Politics"
…





Share price
"$ 20.50"

- "Experience" or training data:

$$\{< x_d, y_d >\}_{d=1}^{D},\ x_d \in \mathcal{X}, y_d \in \mathcal{Y}$$

Supervised Learning – classification

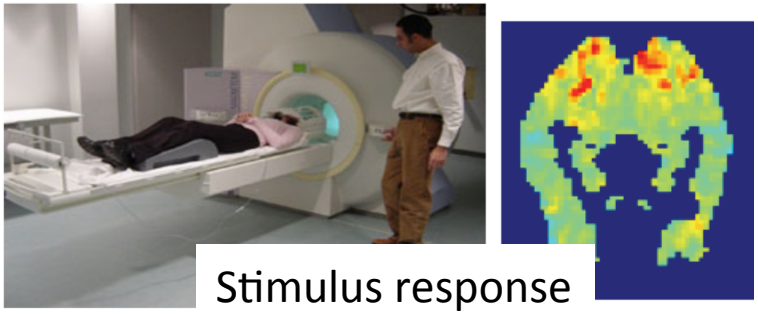
Feature space $\mathcal{X}$

Label space $\mathcal{Y}$





"Sports"
"News"
"Politics"
…





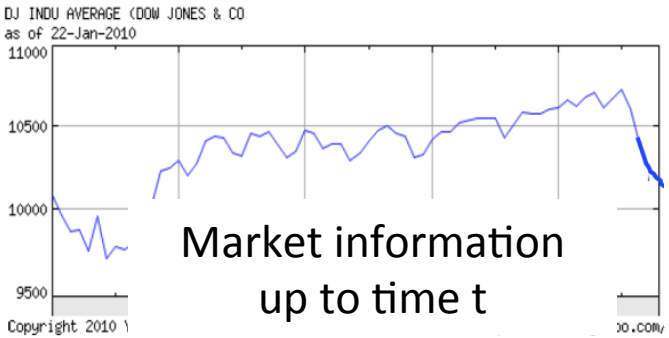
"Tool"
"Animal"
…

Discrete Labels

Supervised Learning – regression

Feature space $\mathcal{X}$

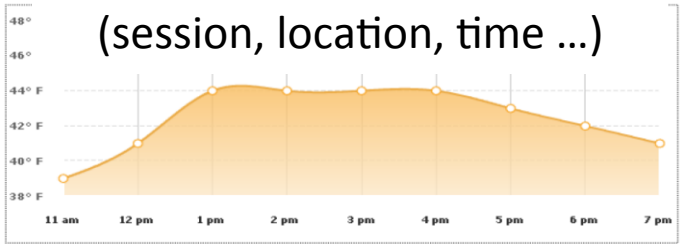
Label space $\mathcal{Y}$



Market information up to time t



(session, location, time …)

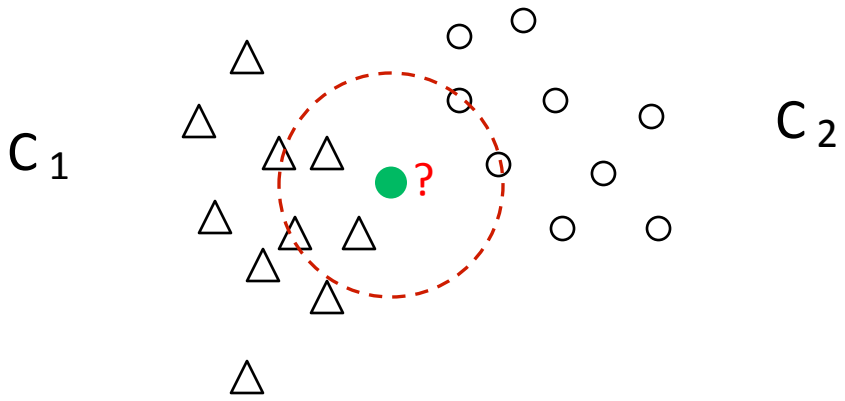


Share price
"$ 20.50"

Temperature
"42° F"

Continuous Labels

How to learn a classifier?

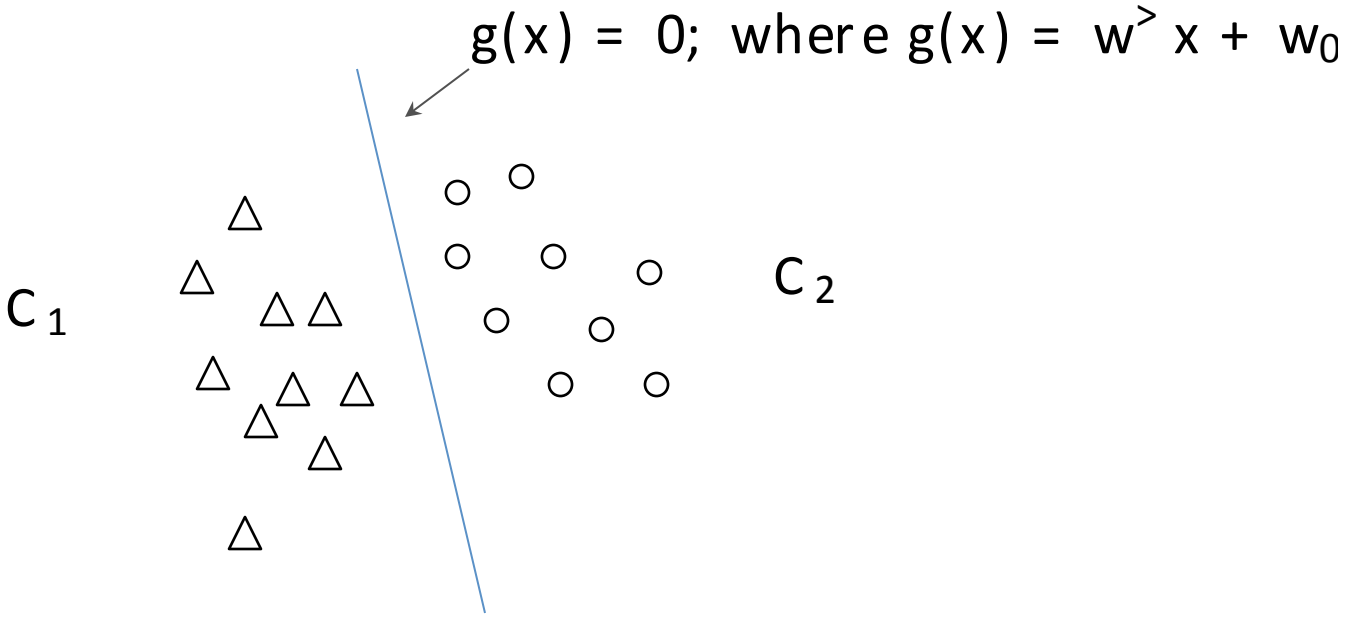


K-NN: a Non-parametric approach

Distance metric matters!

How to learn a classifier?

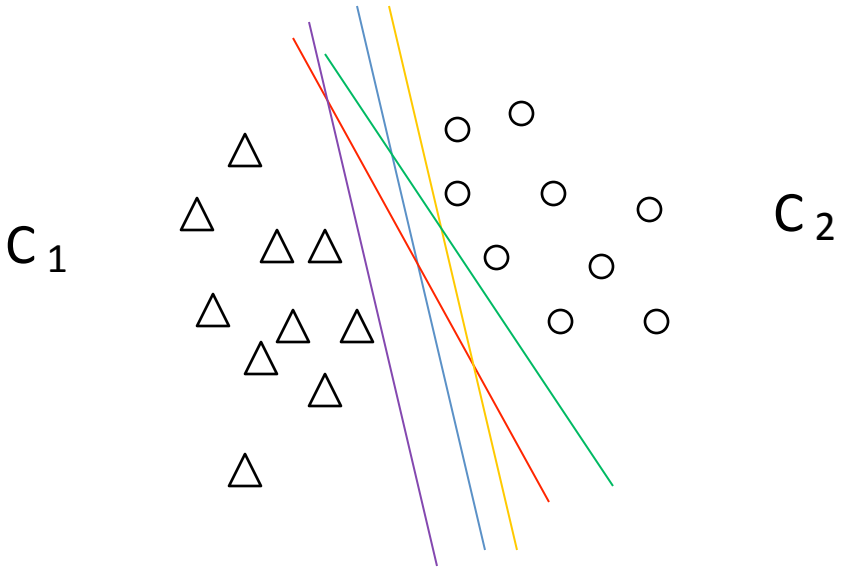
Parametric (model-based) approaches:



a good decision boundary

$$y^{\S} = \begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{if } g(x) < 0 \end{cases}$$

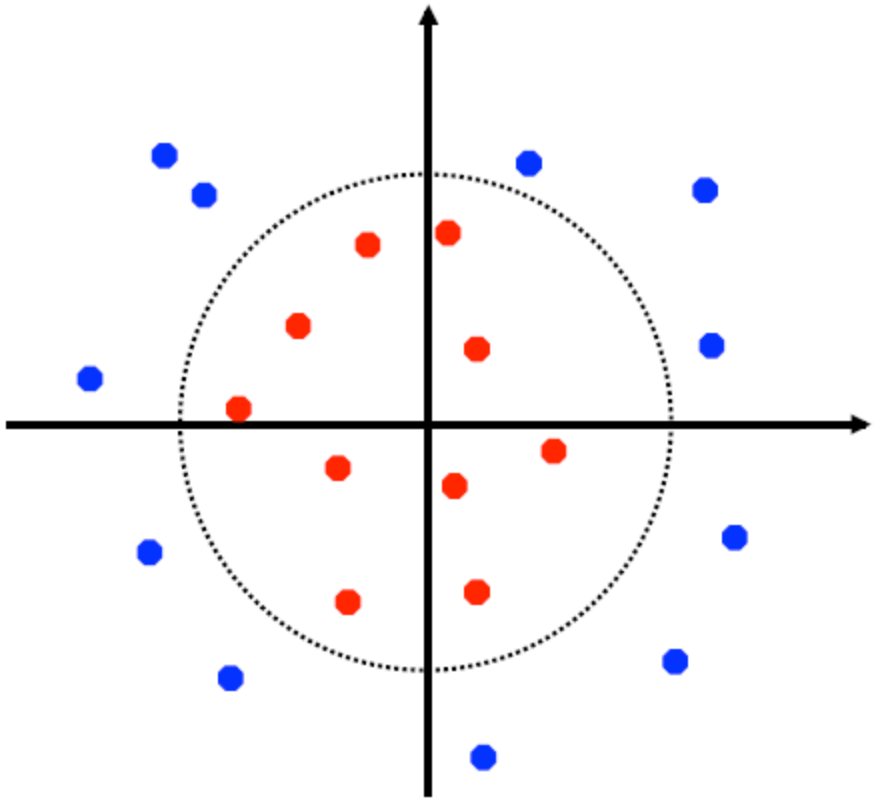
How to learn a classifier?



Many good decision boundaries

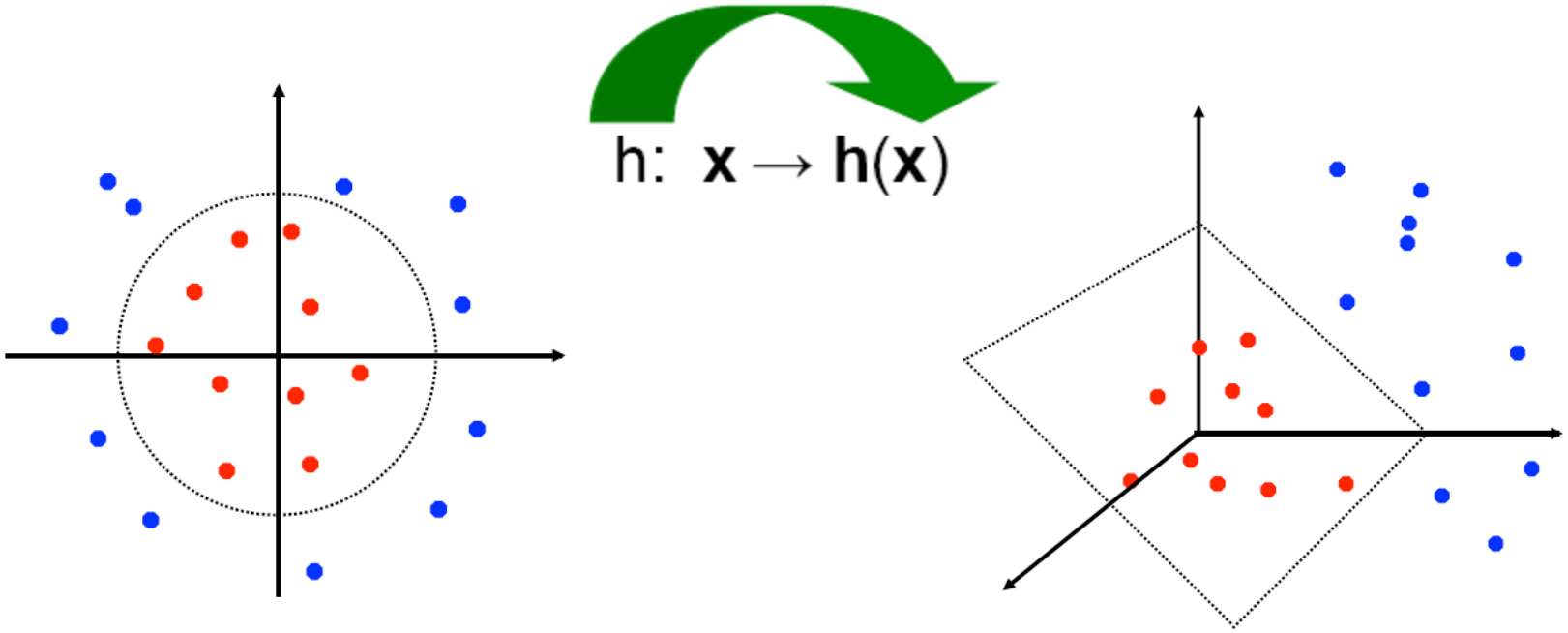
which one should we choose?

How to learn a classifier?



How about non-linearity?

How to learn a classifier?

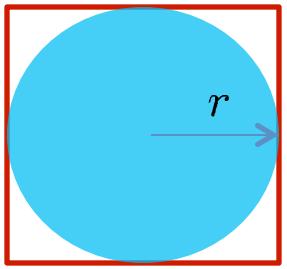


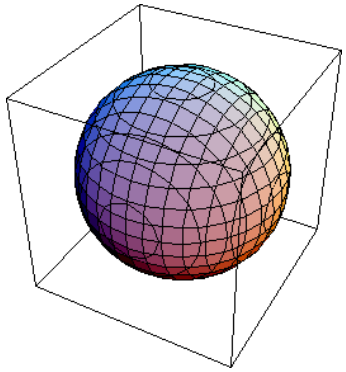
How about non-linearity?

The higher dimension, the better?

How to learn a classifier?

- Curse of dimensionality
  - A high dimensional space is always almost empty





d dimensional space

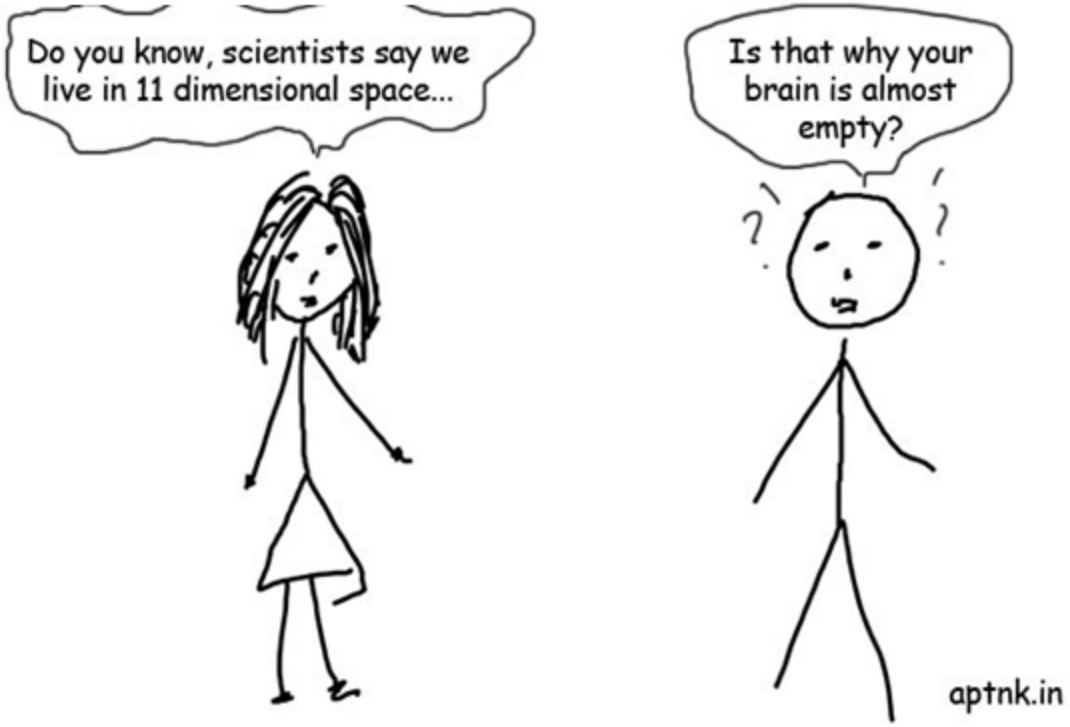
$$\frac{\pi r^2}{(2r)^2} = \frac{\pi}{4} \qquad \frac{\frac{2r^3\pi^{3/2}}{3\Gamma(3/2)}}{(2r)^3} = \frac{\pi^{3/2}}{12\Gamma(3/2)} \qquad \frac{\frac{2r^d\pi^{d/2}}{d\Gamma(d/2)}}{(2r)^d} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)}$$

$$\uparrow d \to \infty \quad 0$$

when one wants to learn pattern from data in high dimensions no matter how much data you have it always seems less!

How to learn a classifier?

- Curse of dimensionality



when one wants to learn pattern from data in high dimensions no matter how much data you have it always seems less! A high dimensional space is always almost empty

How to learn a classifier?

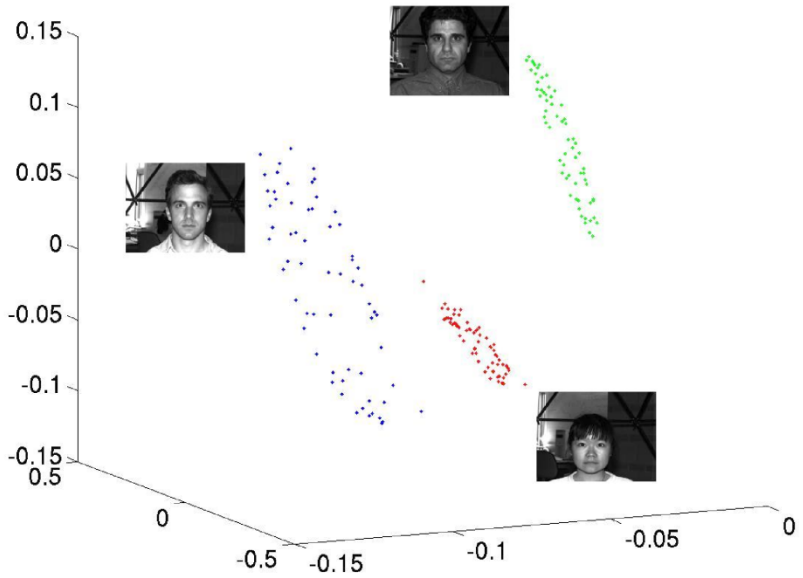
- Curse of dimensionality
  - A high dimensional space is always almost empty
  - … in high dimensions no matter how much data you have it always seems less!

- The blessing of dimensionality
  - *… real data highly concentrate on low-dimensional, sparse, or degenerate structures in the high-dimensional space.*

- But no free lunch: *Gross errors and irrelevant measurements are now ubiquitous in massive cheap data.*

How to learn a classifier?

- The blessing of dimensionality
  - *… real data highly concentrate on low-dimensional, sparse, or degenerate structures in the high-dimensional space.*

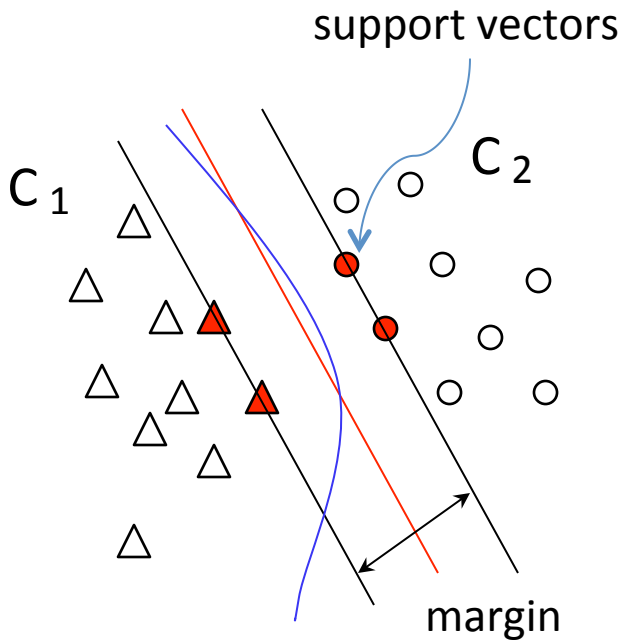


Images of the same face under varying illumination lie approximately on a low (nine)-dimensional subspace, known as the harmonic plane [Basri & Jacobs, PAMI, 2003].

How to learn a classifier?

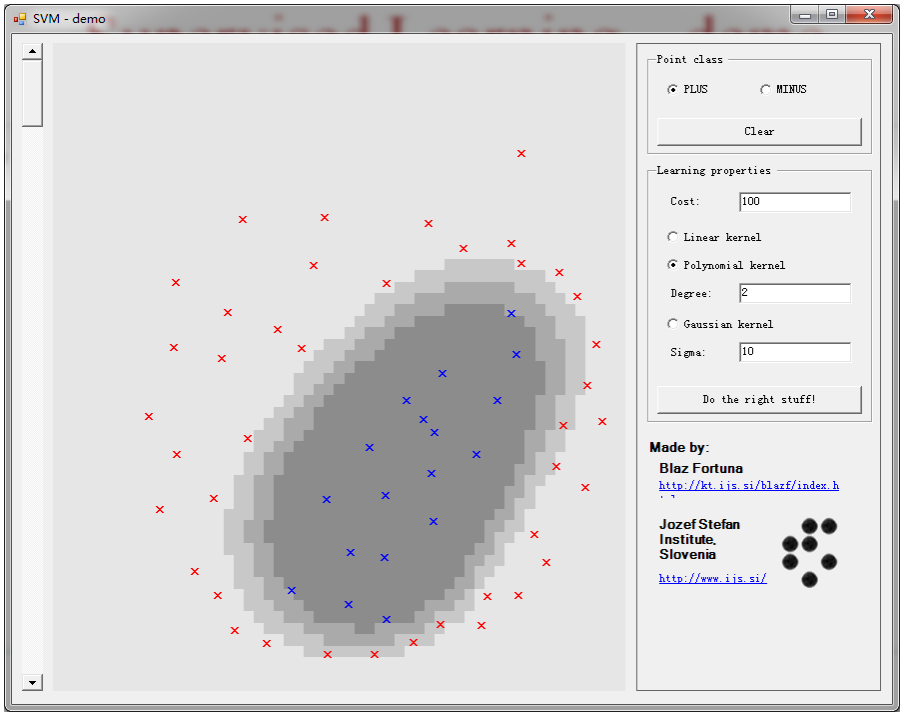
- Support vector machines (SVM) – basics
  - SVM is among the most popular/successful classifiers
  - It provides a *principled way* to learn a *robust* classifier (i.e., a *decision boundary*)

- SVM
  - chooses the one with *maximum margin principle*
  - has sound *theoretical guarantee*
  - extends to *nonlinear decision boundary* by using *kernel* trick
  - learning problem efficiently solved using convex optimization techniques



How to learn a classifier?

- Support vector machines (SVM) – demo



Good ToolKits: [1] SVM-Light: http://svmlight.joachims.org/
[2] LibSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

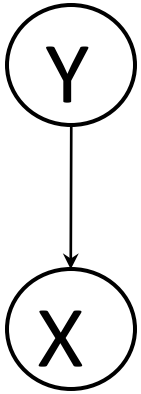
How to learn a classifier?

- Naïve Bayes classifier – basics
  - an representative method from the very important family of *probabilistic graphical models* and *Bayesian methods*

A joint distribution: $p(x, y) = p(y)p(x|y)$

Inference using Bayes rule: prior, likelihood, evidence

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(y)p(x|y)}{p(x)}$$



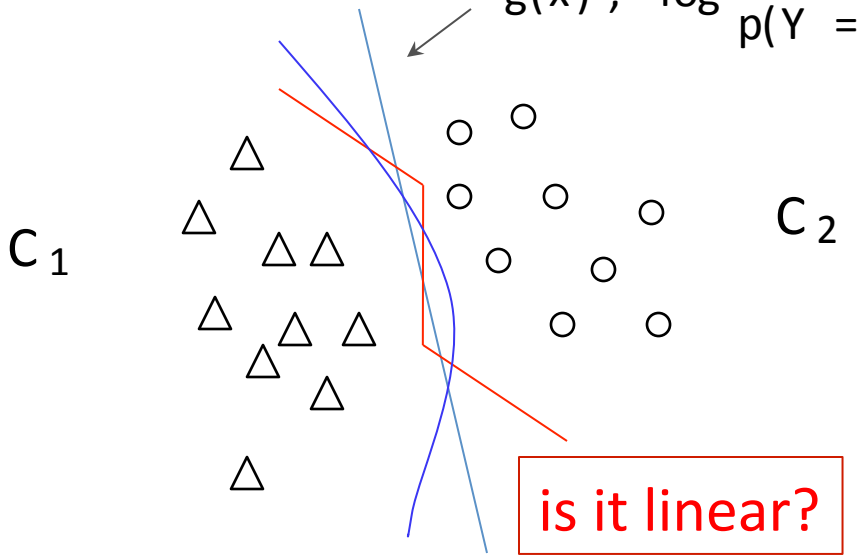
Prediction rule: $y^{\S} = \arg\max_{y \in Y} p(y|x)$

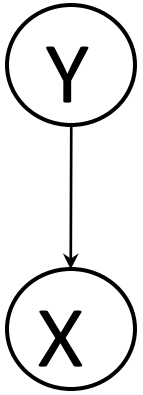
  - fundamental building blocks for *Bayesian networks*
  - nice illustrative example of Bayesian methods

How to learn a classifier?

- Naïve Bayes classifier – basics
  - binary example

$$g(x) , \log \frac{p(Y = C_1 | x)}{p(Y = C_2 | x)} = 0$$





is it linear?

$$y^{\S} = \begin{cases} C_1 & \text{if } p(Y = C_1 | x) > 0{:}5 \\ C_2 & \text{if } p(Y = C_1 | x) < 0{:}5 \end{cases}$$

It is for generalized linear models (GLMs)

How to learn a classifier?

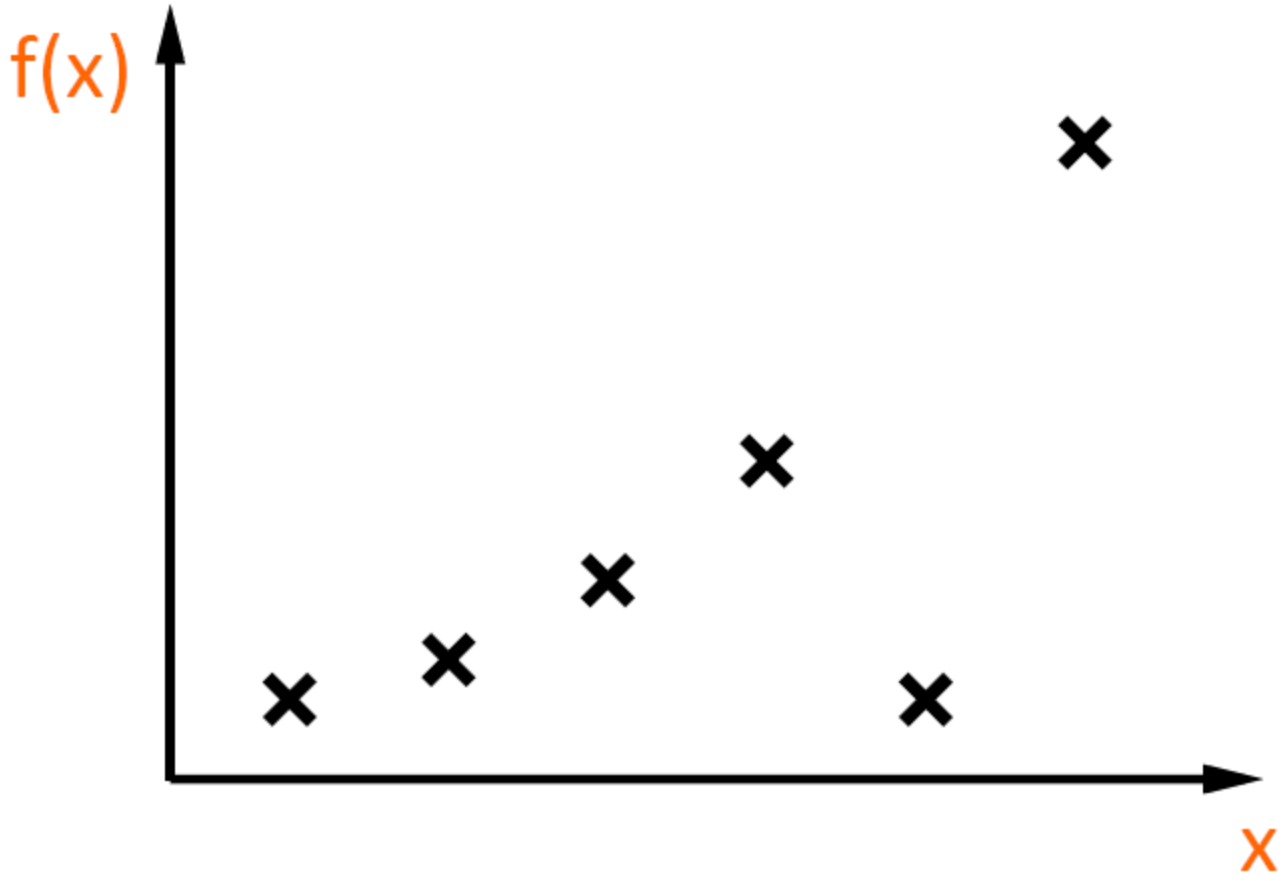
- Many other classifiers
  - K-nearest neighbors
  - Decision trees
  - Logistic regression
  - Boosting
  - Random forests
  - Mixture of experts
  - Maximum entropy discrimination (a nice combination of max-margin learning and Bayesian methods)
  - …

Advice #1:

All models are wrong, but some are useful. – G.E.P. Box

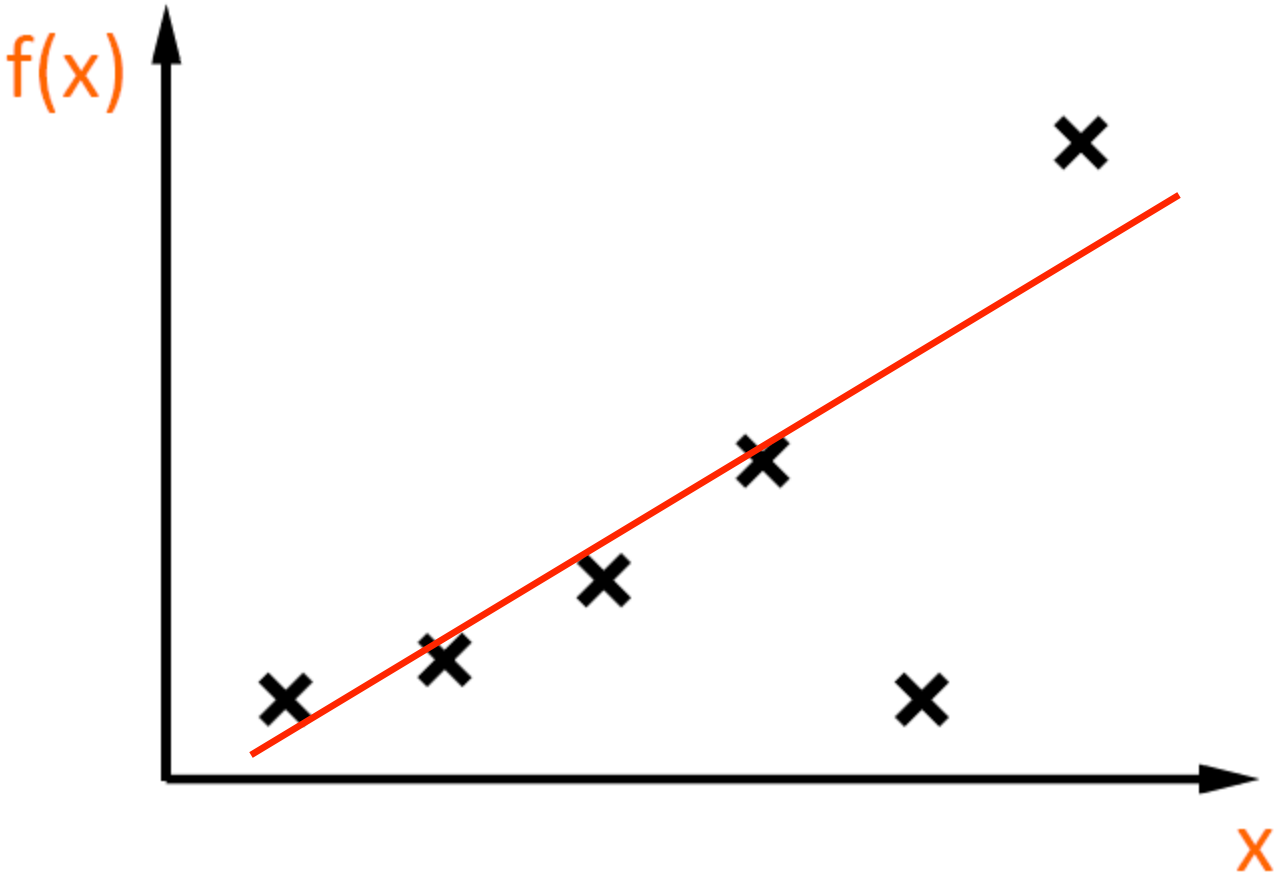
Are complicated models preferred?

- A simple curve fitting task



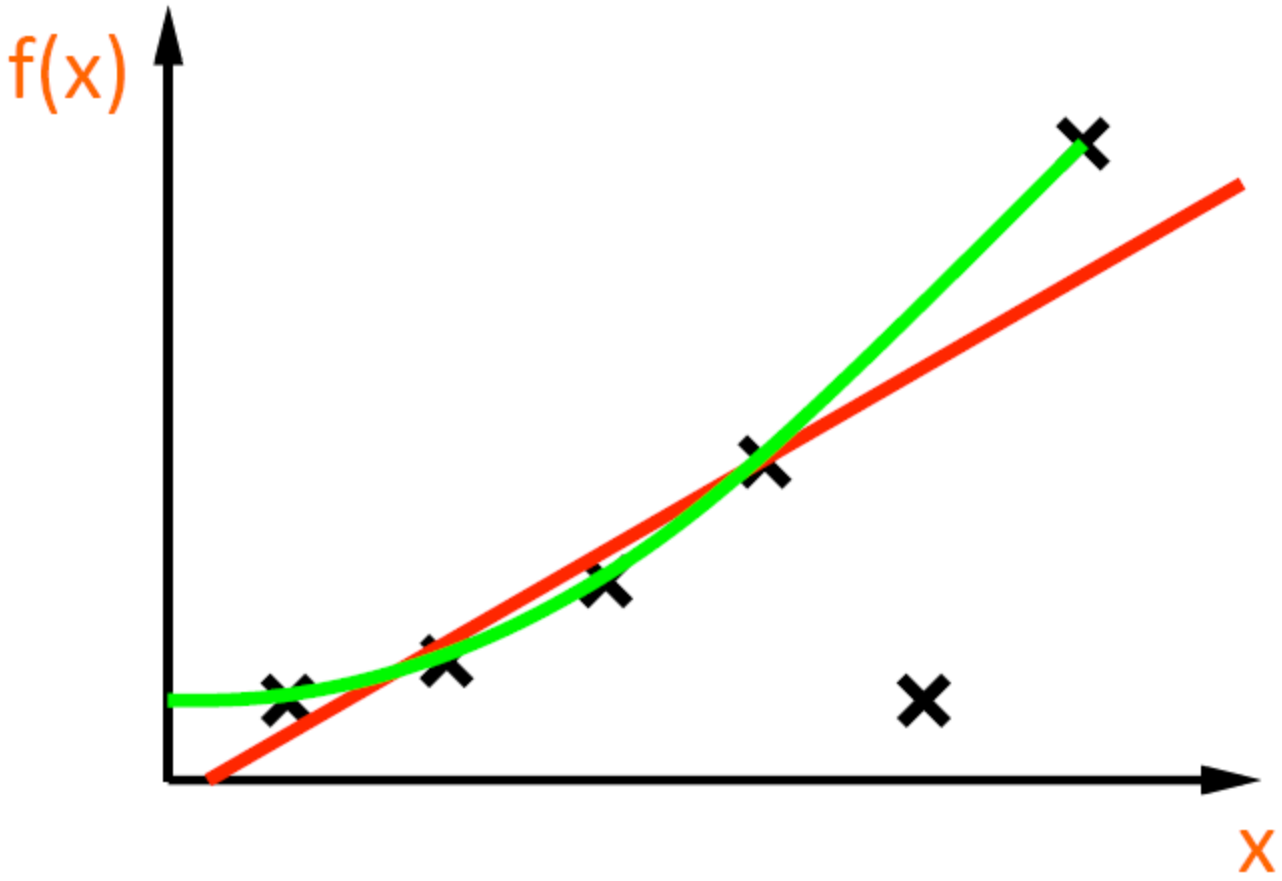
Are complicated models preferred?

- Order = 1



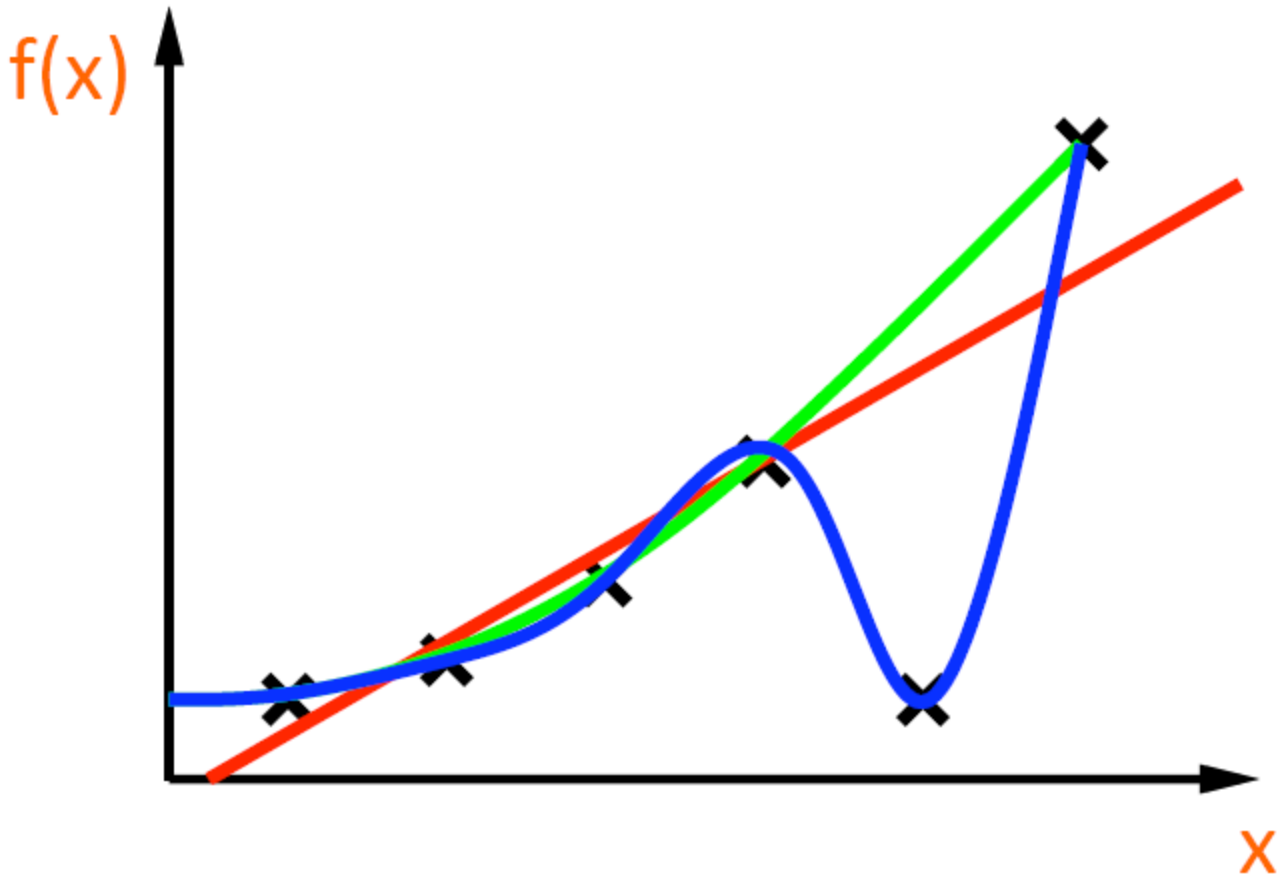
Are complicated models preferred?

- Order = 2



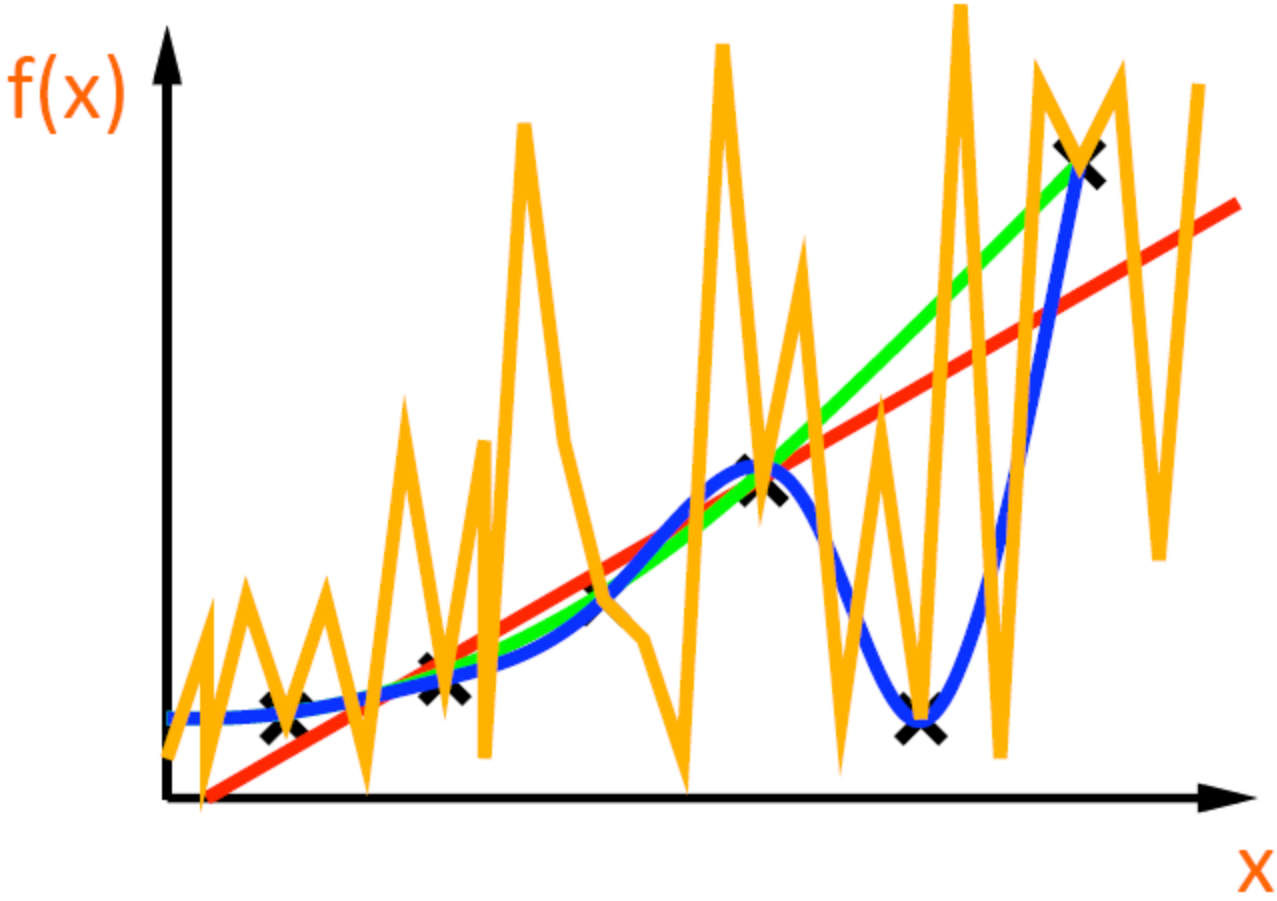
Are complicated models preferred?

- Order = 3



Are complicated models preferred?

- Order = 9?



Are complicated models preferred?

Advice #2: use ML & sophisticated models when necessary



- Issues with model selection!!

Unsupervised Learning

- Task: learn an explanatory function $f(x),\ x \in \mathcal{X}$
- Aka "Learning without a teacher"

Feature space $\mathcal{X}$

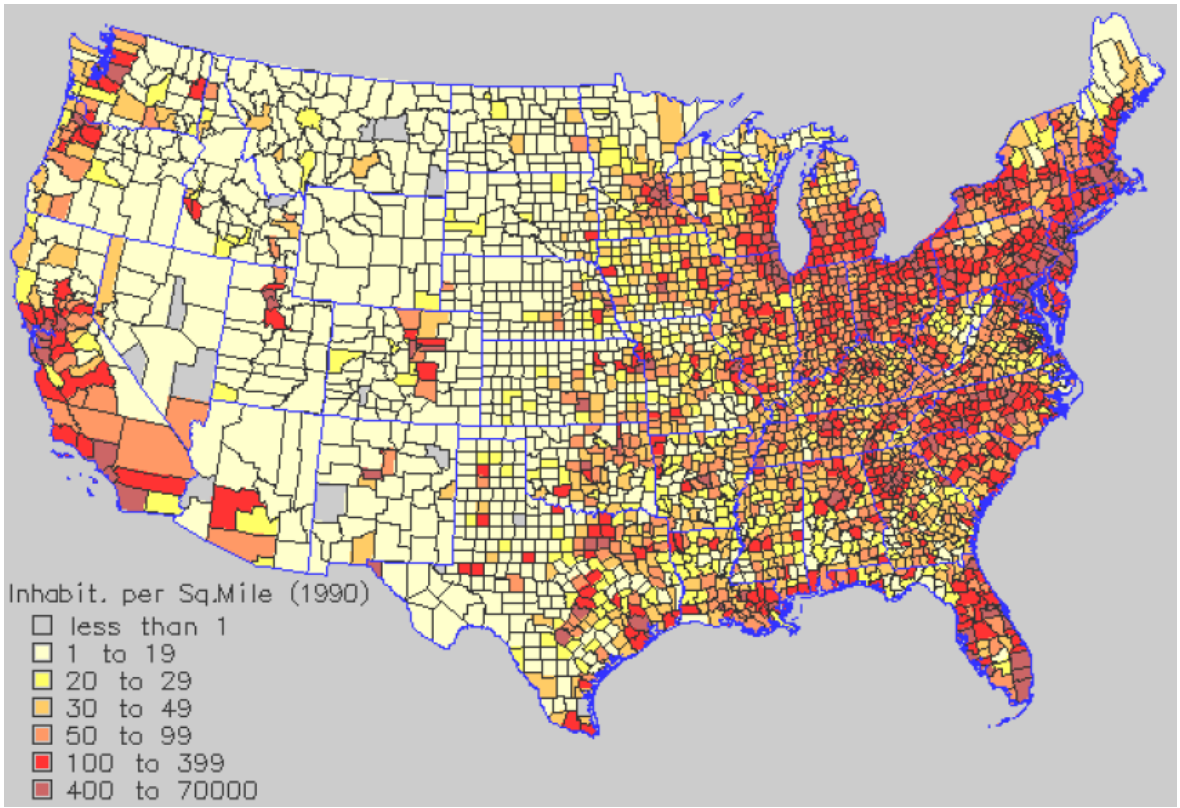




Word distribution
(probability of a word)

- No training/test split

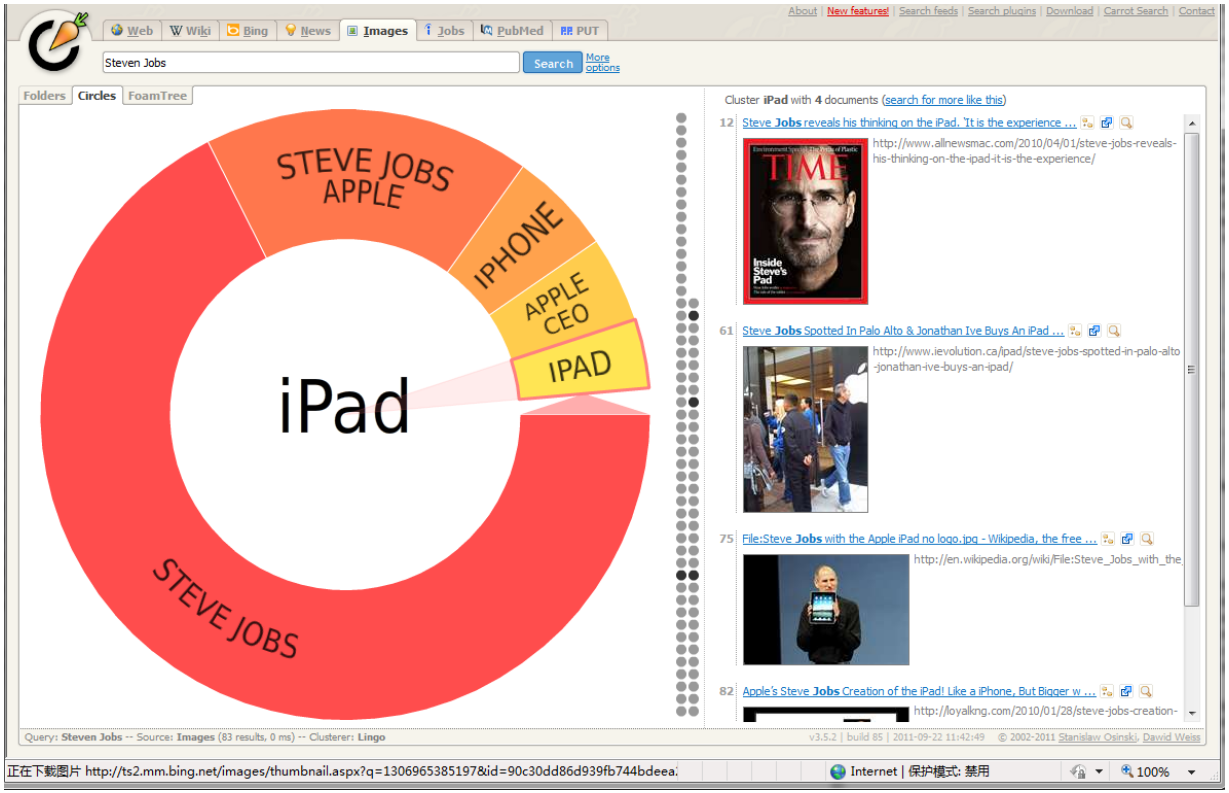
Unsupervised Learning – density estimation



Feature space $\mathcal{X}$
geographical information of a location

Density function
$f(x),\ x \in \mathcal{X}$

Unsupervised Learning – clustering



http://search.carrot2.org/stable/search

Feature space $\mathcal{X}$

Attributes (e.g., pixels & text) of images

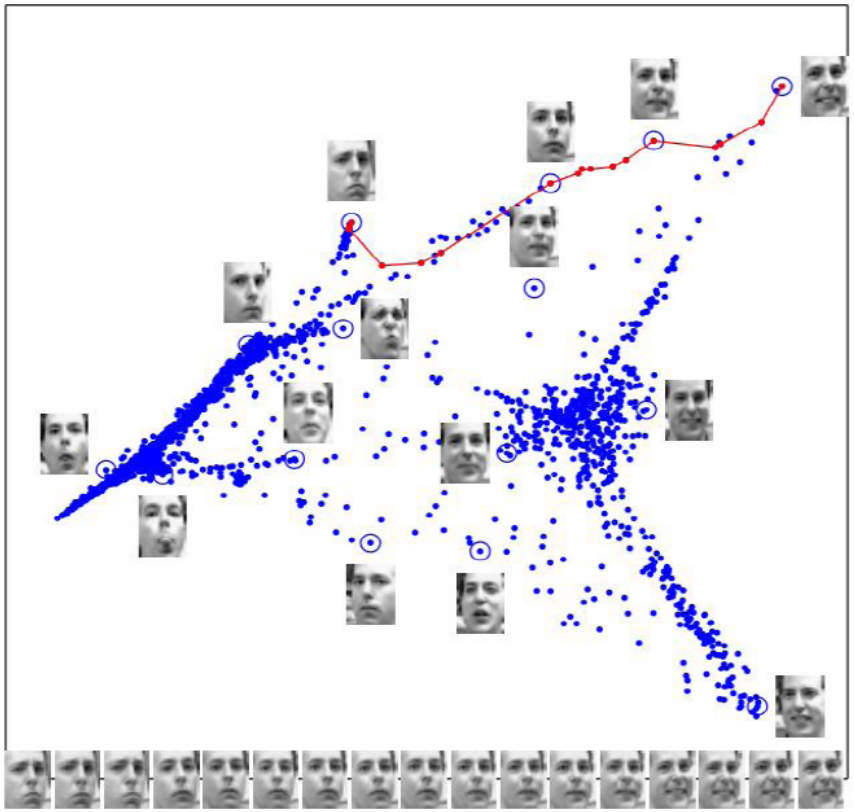
Cluster assignment function

$$f(x),\ x \in \mathcal{X}$$

Unsupervised Learning – dimensionality reduction

Images have thousands or millions of pixels

Can we give each image a coordinate, such that similar images are near each other ?



Feature space $\mathcal{X}$
pixels of images

Coordinate function in 2D space

$$f(x),\ x \in \mathcal{X}$$

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
  - …

dealing with

- 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.

have nice properties.

- ML covers algorithms, theory and very exciting applications
- It's going to be fun and challenging ☺

Interdisciplinary research

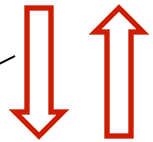
cognitive power



Understanding human brain

- brain activity under various stimulus
- visual & speech perception
- efficient coding, decoding
- …

biological inspiration & support



Tools & theories to deal with complex data



computational power



Statistical machine learning

- various learning paradigms
- sparse learning in high dimension
- learning with deep architectures
- theories & applications
- …

"the only real limitations on making 'machines which think' are our own limitations in not knowing exactly what 'thinking' consists of." – von Neumann

Resources for Further Learning

- Top-tier Conferences:
  - International Conference on Machine Learning (ICML)
  - Advances in Neural Information Processing Systems (NIPS)
  - Uncertainty in Artificial Intelligence (UAI)
  - International Joint Conference on Artificial Intelligence (IJCAI)
  - AAAI Annual Conference (AAAI)
  - Artificial Intelligence and Statistics (AISTATS)

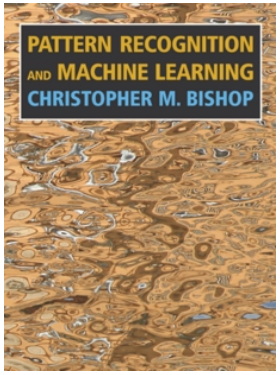
- Top-tier Journals:
  - Journal of Machine Learning Research (JMLR)
  - Machine Learning (MLJ)
  - IEEE Trans. on Pattern Recognition and Machine Intelligence (PAMI)
  - Artificial Intelligence
  - Journal of Artificial Intelligence Research (JAIR)
  - Neural Computation

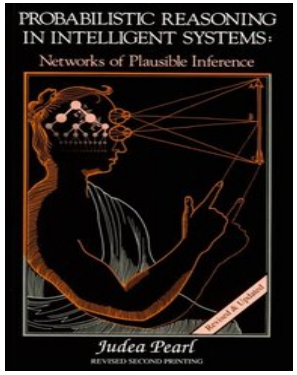
Hot Topics from ICML & NIPS

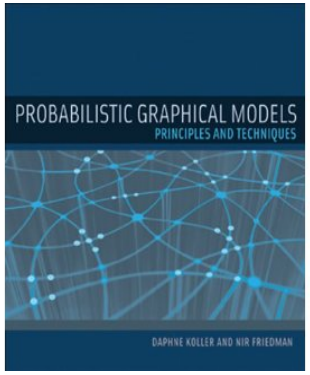
- Hot topics:
  - Probabilistic Latent Variable Models & Bayesian Nonparametrics
  - Deep Learning with Rich Model Architecture
  - Sparse Learning in High Dimensions
  - Large-scale Optimization and Inference
  - Online learning
  - Reinforcement Learning
  - Learning Theory
  - Interdisciplinary Research on Machine Learning, Cognitive Science , etc.

Resources for Further Learning

- Text books:
  - Pattern Recognition and Machine Learning
  - Probabilistic Reasoning in Intelligent Systems
  - Probabilistic Graphical Models (http://pgm.stanford.edu/)







- Public lectures:
  - CMU :
    - http://www.cs.cmu.edu/~guestrin/Class/10708-F08/projects.html
  - Stanford:
    - http://cs228.stanford.edu/
    - http://cs228t.stanford.edu/
  - UPenn:
    - http://www.seas.upenn.edu/~cis620/