Introduction

Instructor: Steven C.H. Hoi

School of Information Systems

Singapore Management University

Email: chhoi@smu.edu.sg

Outline

- What is Machine Learning?
- Why study Machine Learning?
- Machine Learning Applications
- Machine Learning Overview
- History of Machine Learning

Machine Learning is hot!!

A breakthrough in machine learning would be worth ten Microsofts

(Bill Gates, Chairman of Microsoft)

Machine learning is the next Internet

(Tony Tether, Director, DARPA)

Machine learning is today's discontinuity

(Jerry Yang, Founder of Yahoo!)

Machine learning is the hot new thing

(John Hennessy, President, Stanford)

Machine learning is going to result in a real revolution

(Greg Papadopoulos, Former CTO, Sun)

Web rankings today are mostly a matter of machine learning

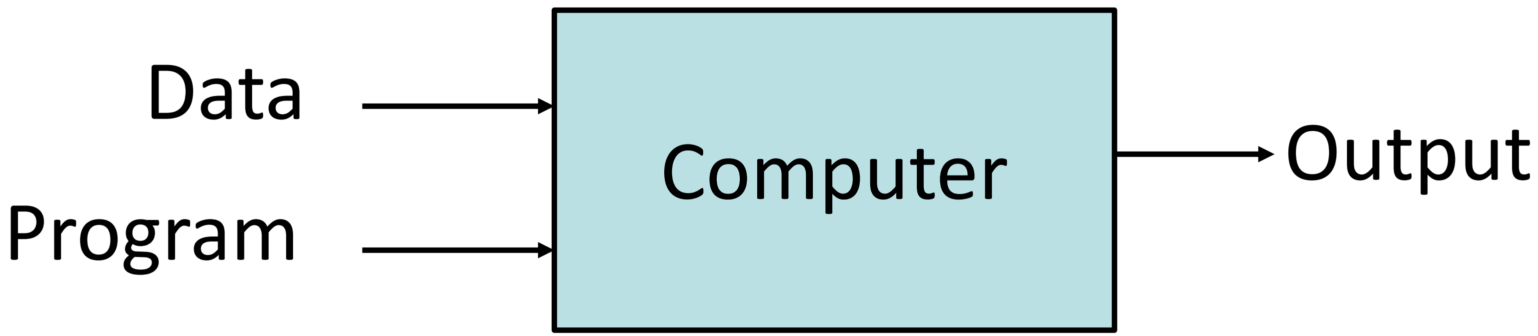
(Prabhakar Raghavan, Director, Yahoo!)

The next era of computer science is going to focus on machine learning

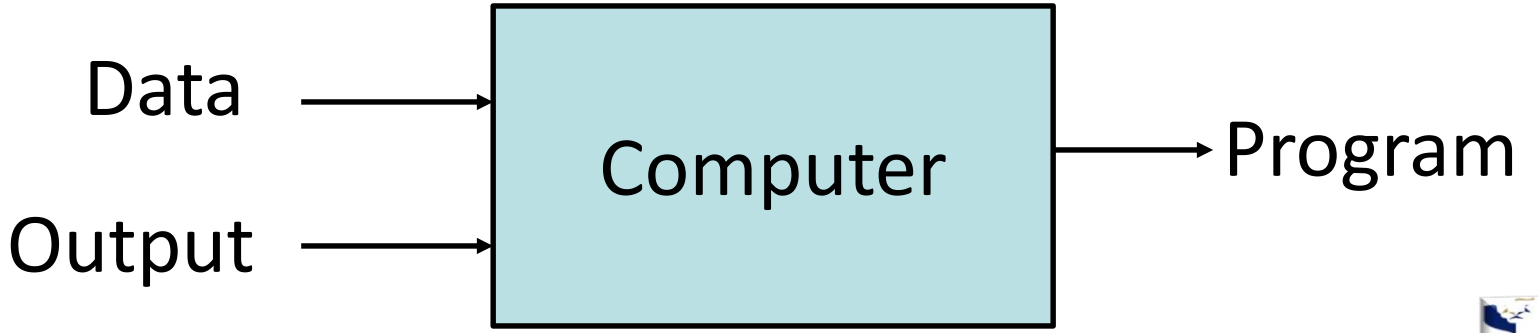
(Steve Ballmer, former CEO, Microsoft)

What is Machine Learning

- **Traditional Programming**



- **Machine Learning**

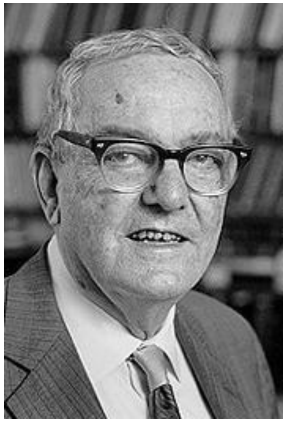


Magic?

- **No, more like gardening**

- **Gardener** = You
- **Seeds** = Algorithms
- **Nutrients** = Data
- **Plants** = Programs

Machine Learning: Definitions



- **Herbert Simon**: "Learning is any process by which a system improves performance from experience."

- **Tom M. Mitchell**: "A computer program is said to learn from experience **E** with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

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

Example: Learning to Play Checker Game

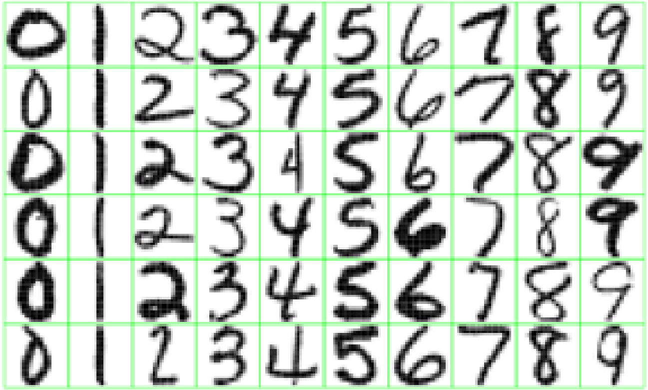
Arthur Samuel (1952)

- T: Playing checker games
- P: Percentage of games won against an arbitrary opponent
- E: Playing practice games against many people or itself

Example: Automated Email Spam Filtering

- T: Categorize email messages as spam or legitimate
- P:
- E:

Example: Handwritten Recognition



- T: Recognizing hand-written words
- P:
- E:

Example: Movie Recommendation

- T: Predict how a user favorites any given movie / Recommend a user top favorite movies
- P:
- E:

Why study Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Recent Trends

- Recent progress in algorithm and theory
- Growing flood of online data and big data
- Computational power is increasing rapidly
- Growing industry in data science and AI

Three Niches for Machine Learning

- Data mining
  - Use historical data to improve decisions, e.g., Medical records → medical knowledge
- Software applications that are hard to program by hand
  - Speech recognition
  - Image classification
  - Autonomous driving
  - etc
- User modeling
  - Automatic recommender systems

Typical Data Mining Task

Patient103 time=1 →	*Patient103* time=2 ··· →	*Patient103* time=n
Age: 23	Age: 23	Age: 23
FirstPregnancy: no	FirstPregnancy: no	FirstPregnancy: no
Anemia: no	Anemia: no	Anemia: no
Diabetes: no	Diabetes: YES	Diabetes: no
PreviousPrematureBirth: no	PreviousPrematureBirth: no	PreviousPrematureBirth: no
Ultrasound: ?	Ultrasound: abnormal	Ultrasound: ?
Elective C–Section: ?	Elective C–Section: no	Elective C–Section: no
Emergency C–Section: ?	Emergency C–Section: ?	**Emergency C–Section: Yes**
	...	...

Given:

- 9147 patient records, each describing pregnancy and birth
- Each patient contains 215 features

Task:

- Classes of future patients at high risk for Emergency Cesarean Section

Data Mining Results

Patient103 time=1	*Patient103* time=2	*Patient103* time=n
Age: 23	Age: 23	Age: 23
FirstPregnancy: no	FirstPregnancy: no	FirstPregnancy: no
Anemia: no	Anemia: no	Anemia: no
Diabetes: no	Diabetes: YES	Diabetes: no
PreviousPrematureBirth: no	PreviousPrematureBirth: no	PreviousPrematureBirth: no
Ultrasound: ?	Ultrasound: abnormal	Ultrasound: ?
Elective C–Section: ?	Elective C–Section: no	Elective C–Section: no
Emergency C–Section: ?	Emergency C–Section: ?	**Emergency C–Section: Yes**
	...	...

One of 18 learned rules:

If no previous vaginal delivery
abnormal 2nd Trimester Ultrasound
Malpresentation at admission

Then probability of Emergency C-Section is 0.6

Credit Risk Analysis

Customer103: (time=t0)	*Customer103:* (time=t1)	...	*Customer103:* (time=tn)
Years of credit: 9	Years of credit: 9		Years of credit: 9
Loan balance: $2,400	Loan balance: $3,250		Loan balance: $4,500
Income: $52k	Income: ?		Income: ?
Own House: Yes	Own House: Yes		Own House: Yes
Other delinquent accts: 2	Other delinquent accts: 2		Other delinquent accts: 3
Max billing cycles late: 3	Max billing cycles late: 4		Max billing cycles late: 6
Profitable customer?: ?	Profitable customer?: ?		**Profitable customer?: No**
...	...		...

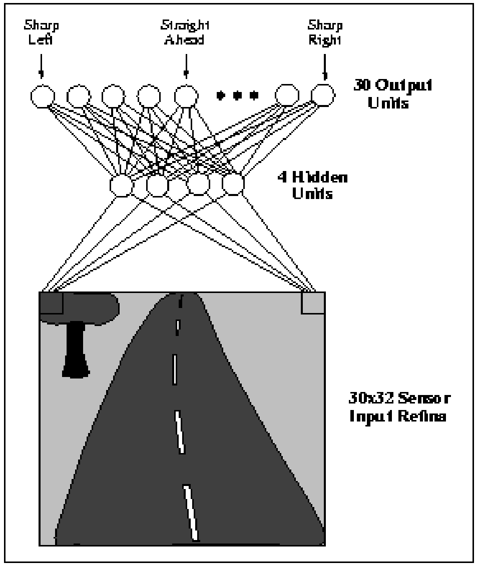
Learned Rules:

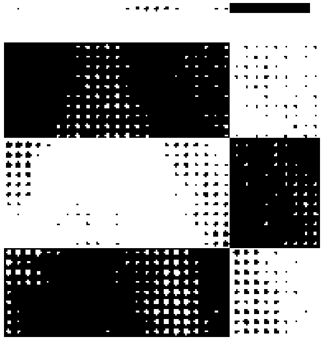
If Other-Delinquent-Account > 2
Number-Delinquent-Billing-Cycles > 1
Then Profitable-Costumer ? = no

If Other-Delinquent-Account = 0
(Income > $30K or Years-of-Credit > 3)
Then Profitable-Costumer ? = yes

Programs too Difficult to Program By Hand

- ALVINN drives 70mph on highways

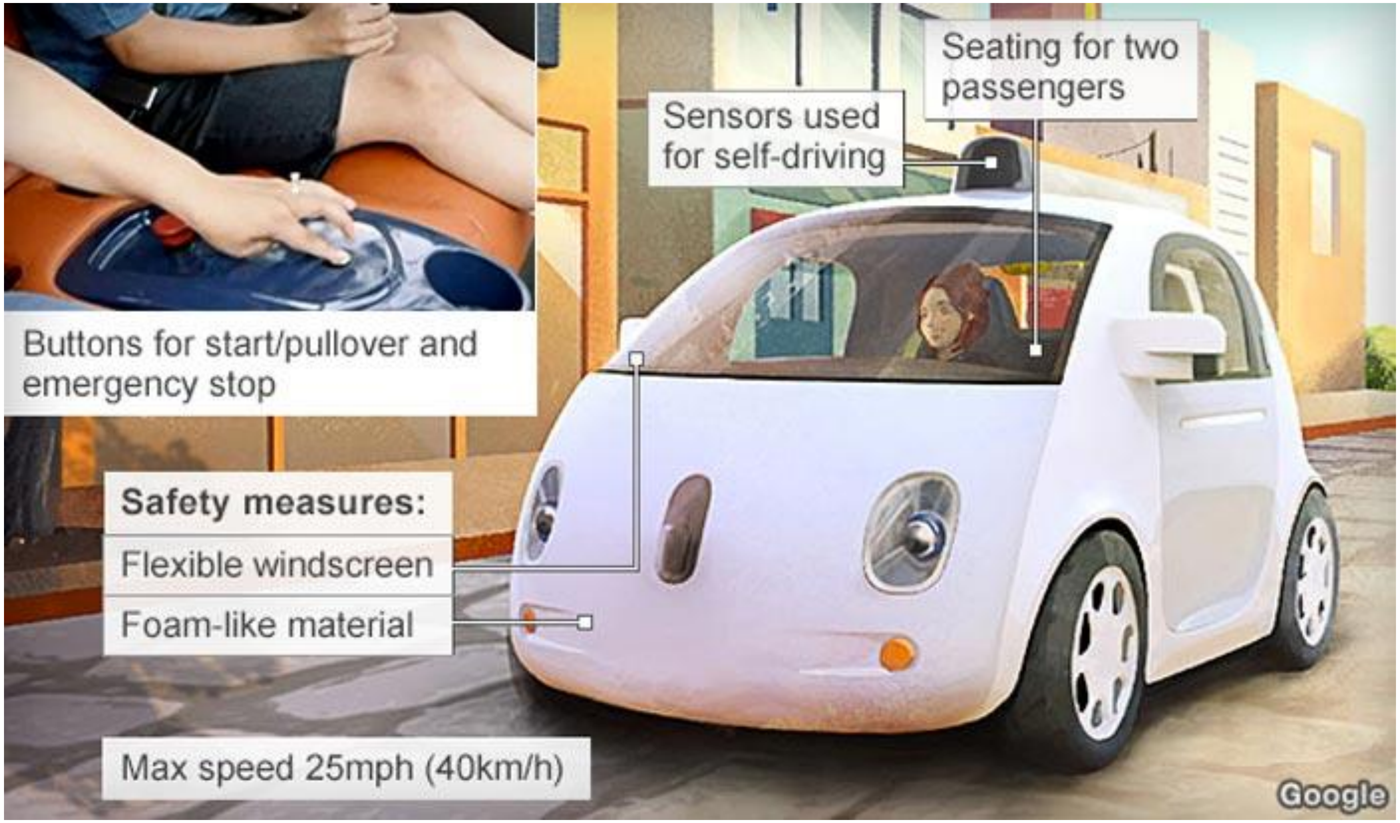




ALVINN: an autonomous land vehicle in a neural network (1989)

Programs too Difficult to Program By Hand

- Google's self-driving car

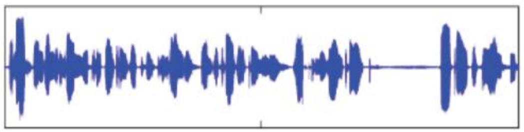


Google's new, completely autonomous vehicle.

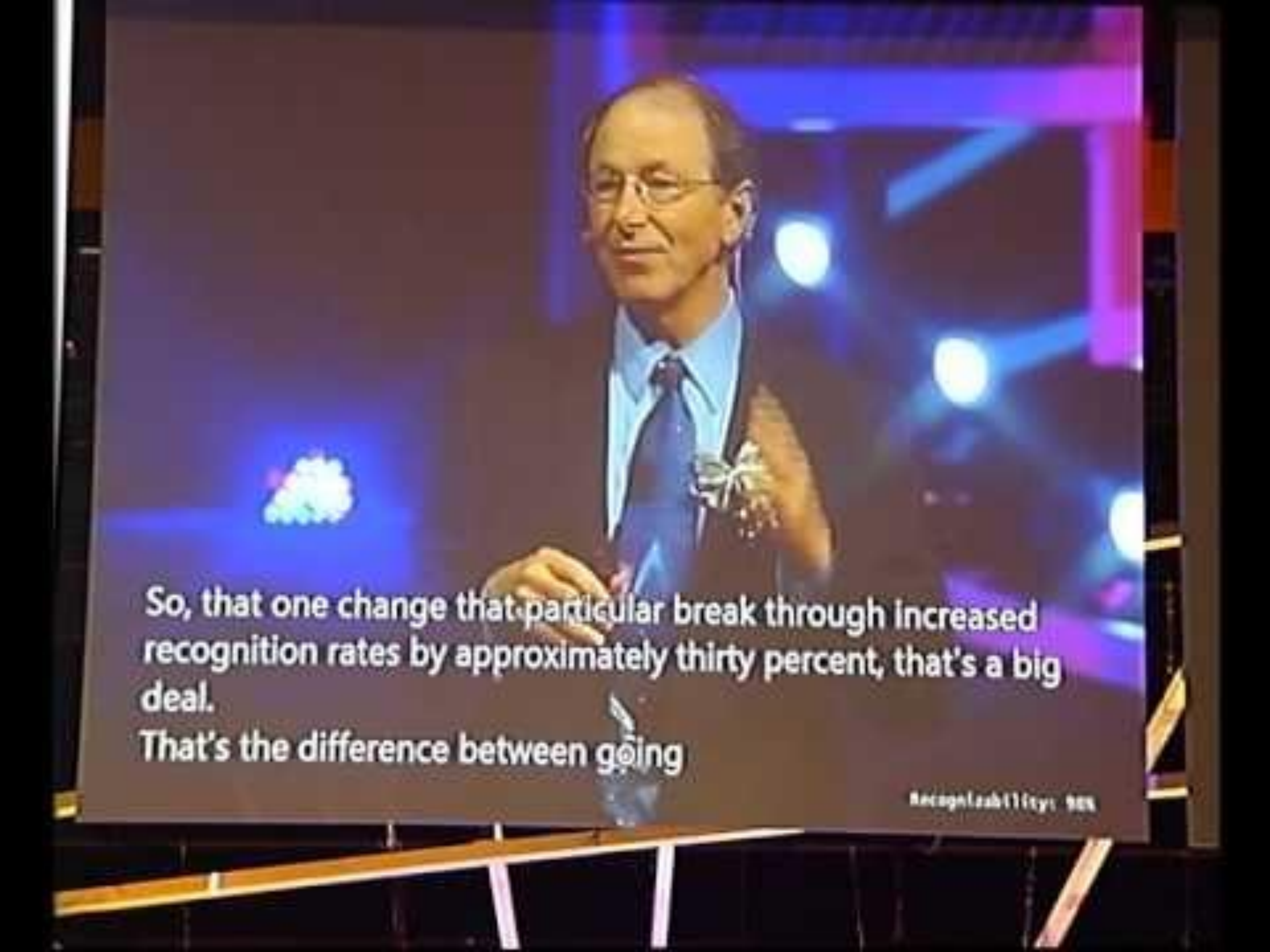

THE VERGE

Programs too Difficult to Program By Hand

- Speech Recognition

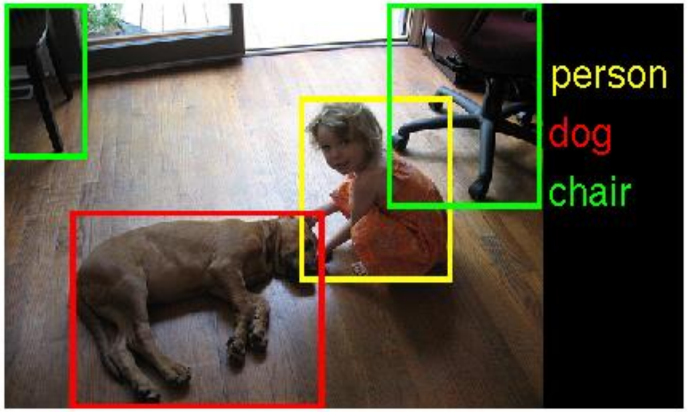


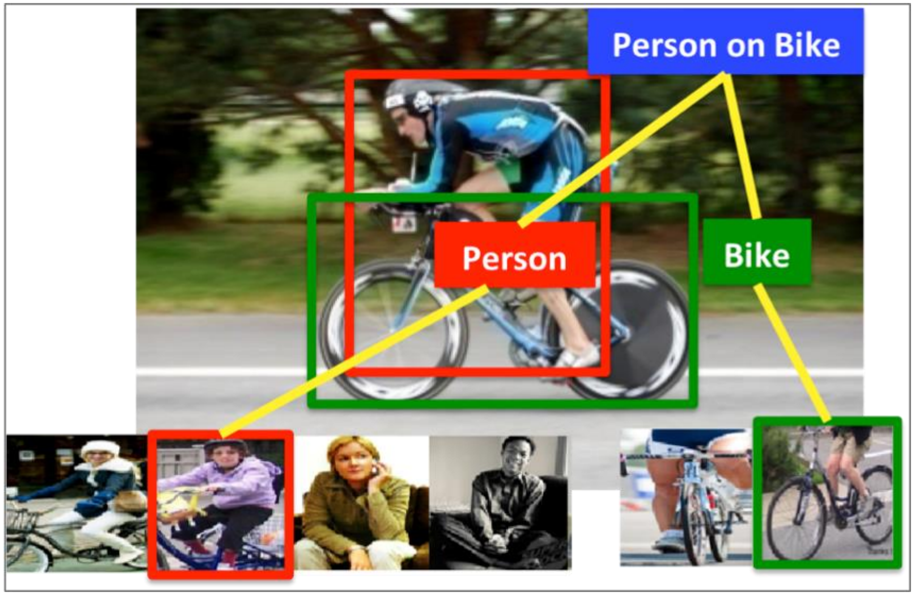
- Speech Translation


So, that one change that particular break through increased recognition rates by approximately thirty percent, that's a big deal.
That's the difference between going

Programs too Difficult to Program By Hand

- Visual object recognition

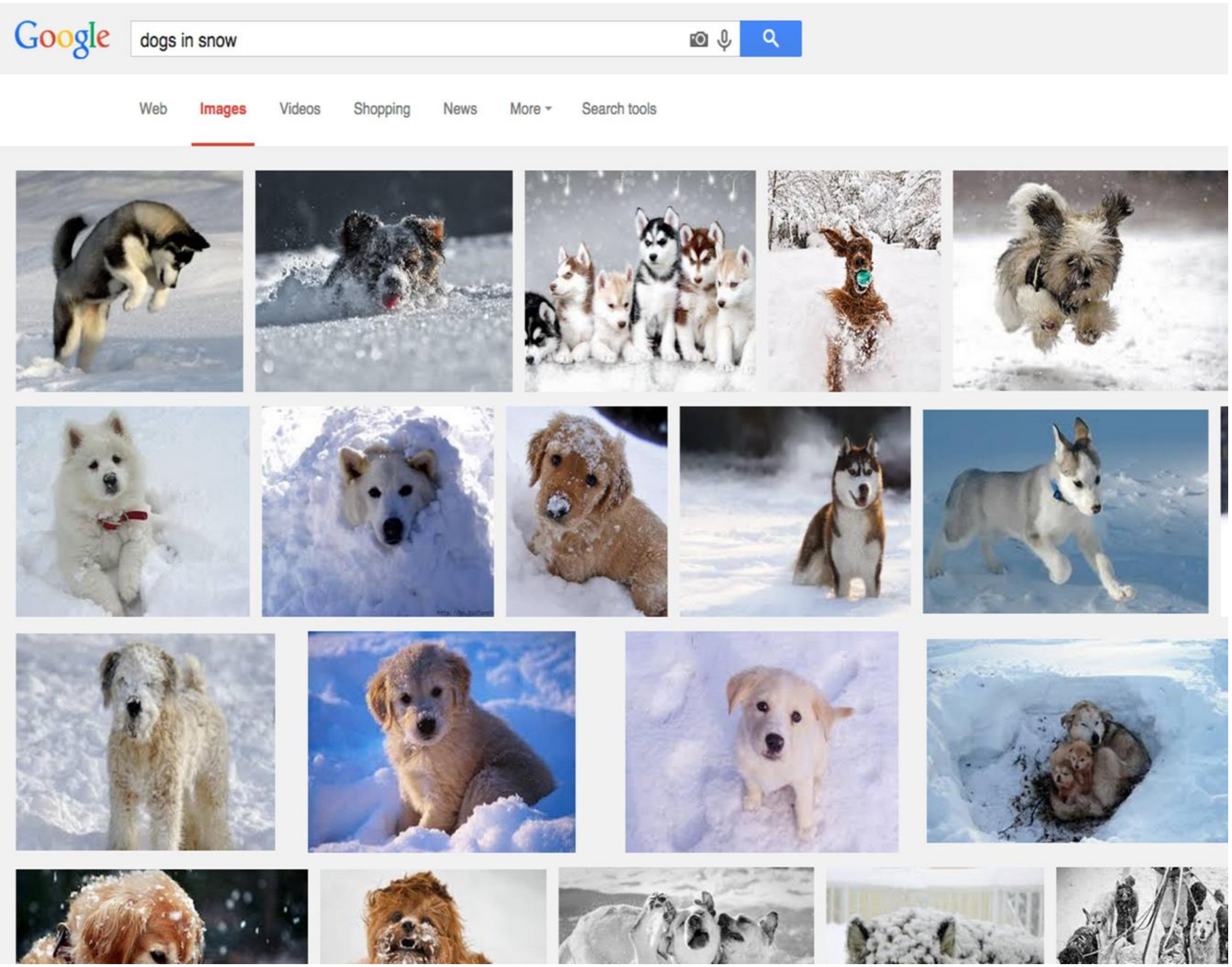




- Image Tagging/Captioning

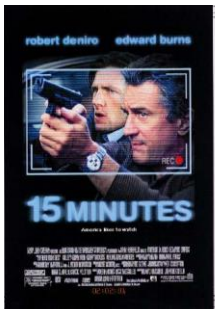
Automatically captioned: *“Two pizzas sitting on top of a stove top oven”*

Image Retrieval using Texts



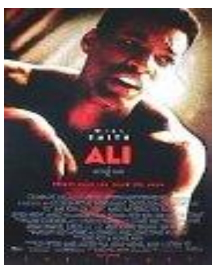
Software that Models Users

History



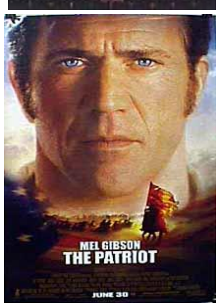
Description: A homicide detective and a fire marshall must stop a pair of murderers who commit videotaped crimes to become media darlings

Rating: ★★★★



Description: A biography of sports legend, Muhammad Ali, from his early days to his days in the ring

Rating: ★★★



Description: Benjamin Martin is drawn into the American revolutionary war against his will when a brutal British commander kills his son.

Rating: ★★

What to Recommend?

Description: A high-school boy is given the chance to write a story about an up-and-coming rock band as he accompanies it on their concert tour.

Recommend: No

Description: A young adventurer named Milo Thatch joins an intrepid group of explorers to find the mysterious lost continent of Atlantis.

Recommend: Yes

Netflix Recommendation Contest

- Machine learning competition with $1million prize
  - Movie recommendations

And if You Liked the Movie, a Netflix Contest May Reward You Handsomely

By KATIE HAFNER
Published: October 2, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS
SAVE

ARTICLE TOOLS SPONSORED BY
FOR YOUR CONSIDERATION LITTLE MISS SUNSHINE

Netflix, the popular online movie rental service, is planning to award $1 million to the first person who can improve the accuracy of movie recommendations based on personal preferences.

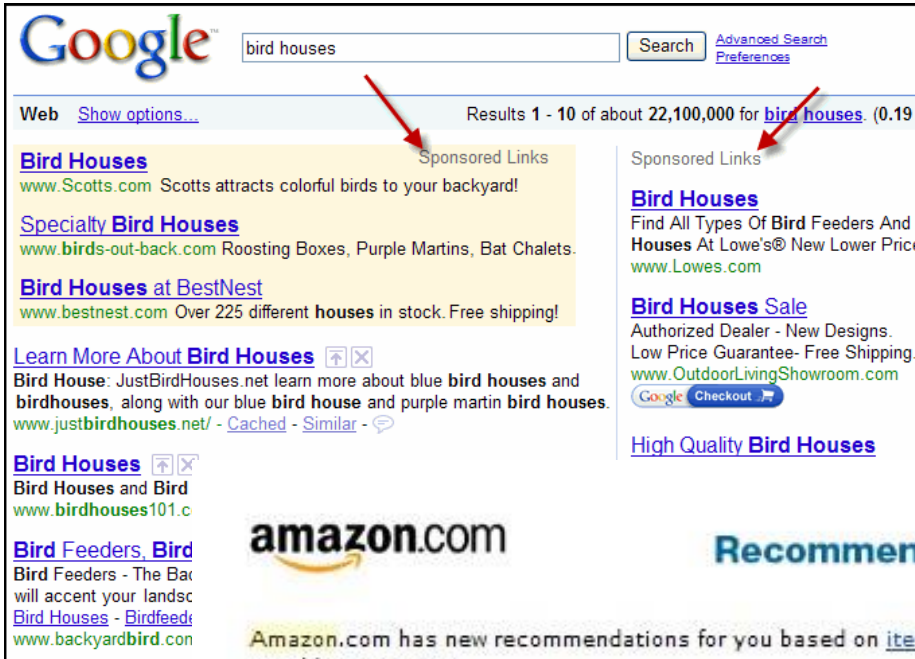
Enlarge This Image

To win the prize, which is to be announced today, a contestant will have to devise a system that is more accurate than the company's current recommendation system by at least 10 percent. And to improve the quality of research, Netflix is making available to the public 100 million of its customers' movie ratings, a database the company says is the largest of its kind ever released.

users →, movies ↓

1		?	3	5	?
?	1				2
	4		4	5	?

Web Advertising and Recommendation



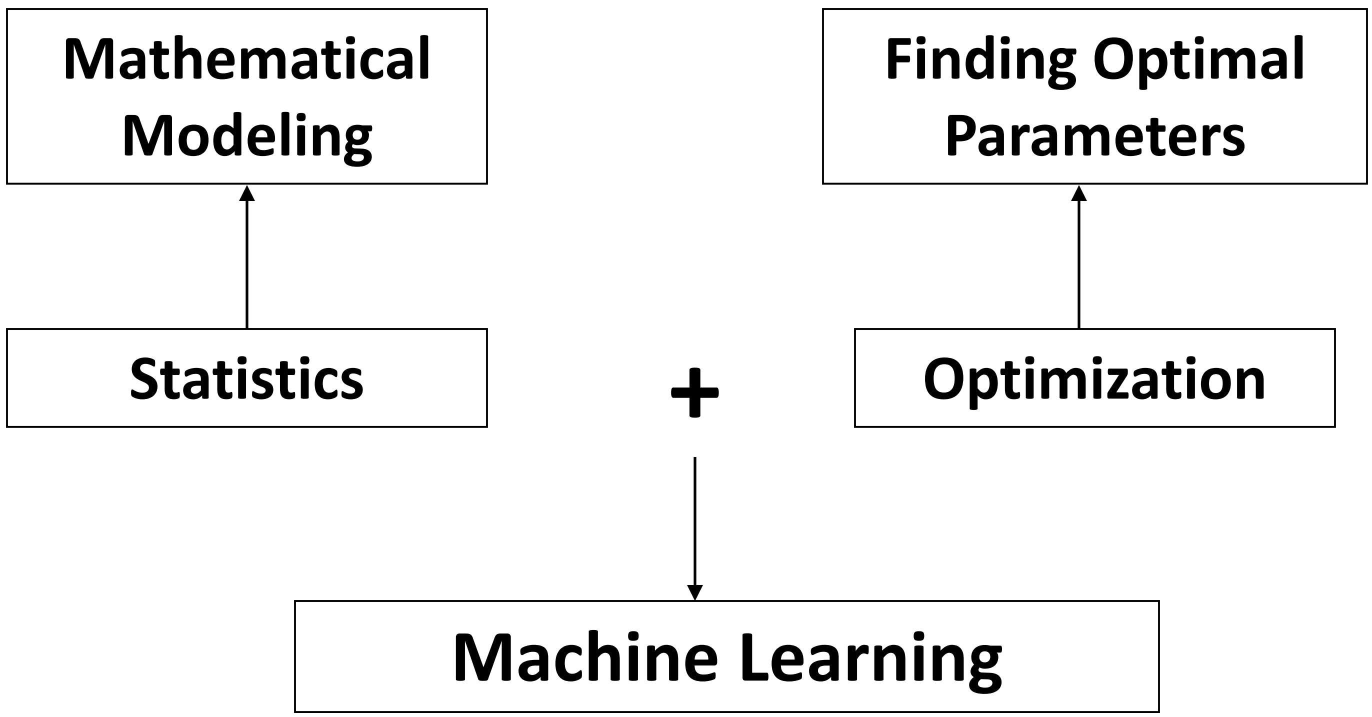
More ML Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Software Debugging
- [Your favorite area]

Relevant Disciplines

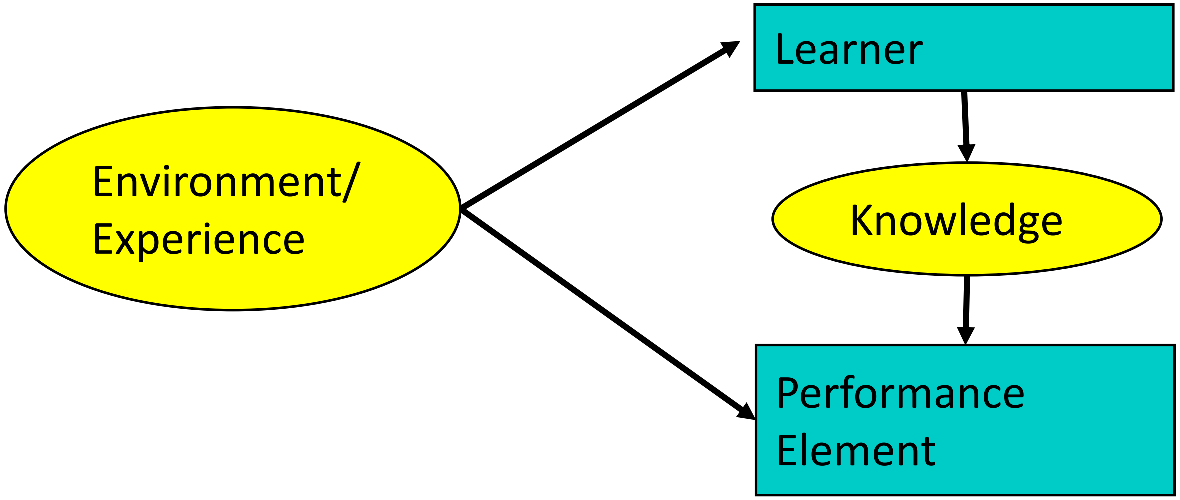
- Artificial Intelligence
- Statistics (particularly Bayesian Stat.)
- Computational complexity theory
- Information theory
- Optimization theory
- Data Mining
- Psychology
- Philosophy
- Neurobiology
- …

A General Framework



Designing a Learning System

- Choose the training experience
- Choose exactly what is too be learned, i.e. the ***target function***.
- Choose how to represent the target function.
- Choose a learning algorithm to infer the target function from the experience.



ML in a Nutshell

- Tens of thousands of machine learning algorithms
- Hundreds new every year
- Every machine learning algorithm has three components:
  - **Representation**
  - **Evaluation**
  - **Optimization**

Representation

- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Optimization

- Combinatorial optimization
  - E.g.: Greedy search
- Convex optimization
  - E.g.: Gradient descent
- Constrained optimization
  - E.g.: Linear programming

Types of Machine Learning

- **Supervised (inductive) learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Semi-supervised learning**
  - Training data includes a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning

- Given (input, correct output), predict (input, ?)

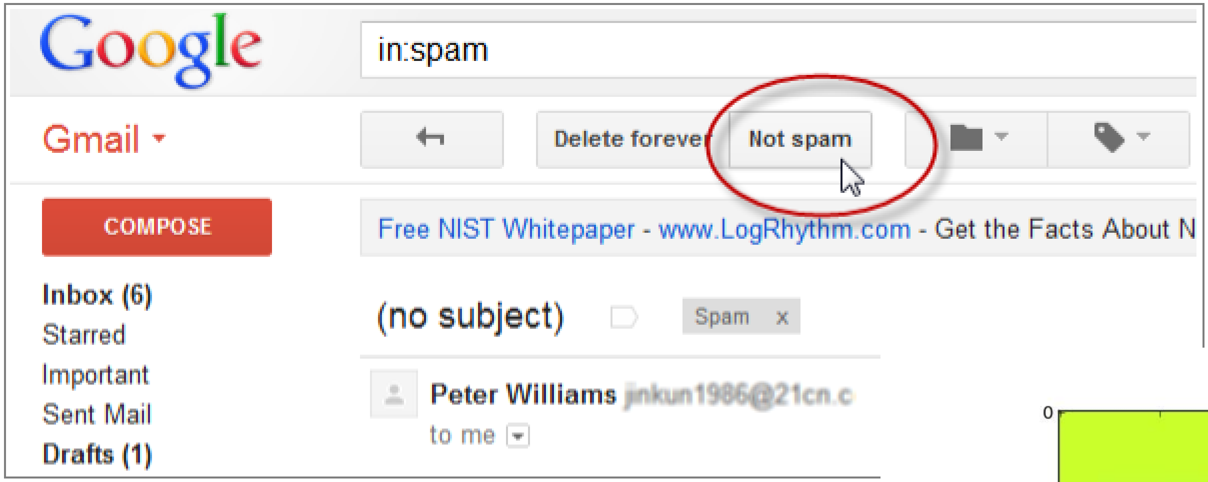
$$h(x) \approx f(x)$$

x $\qquad\qquad$ $y = f(x)$

- **Classification:** discrete output
  - Binary Classification: input x, find y in {-1, +1}
  - Multi-class classification: input x, find y in {1, …, k}
- **Regression:** continuous output
  - Given input x, find y in real-valued space R (R^d)

Binary Classification

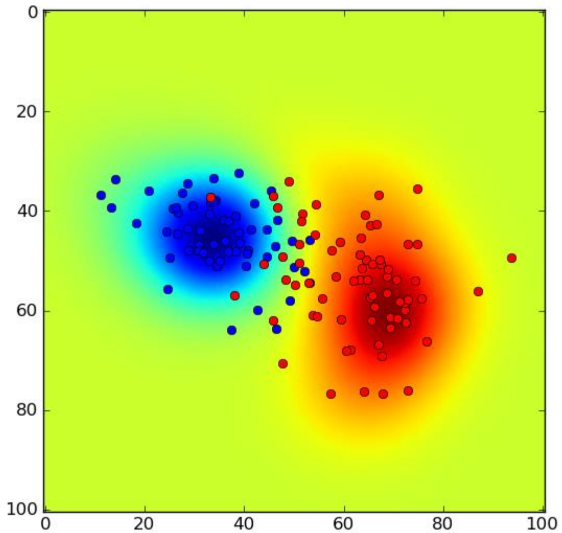
- Spam Email Filtering



- Two classes

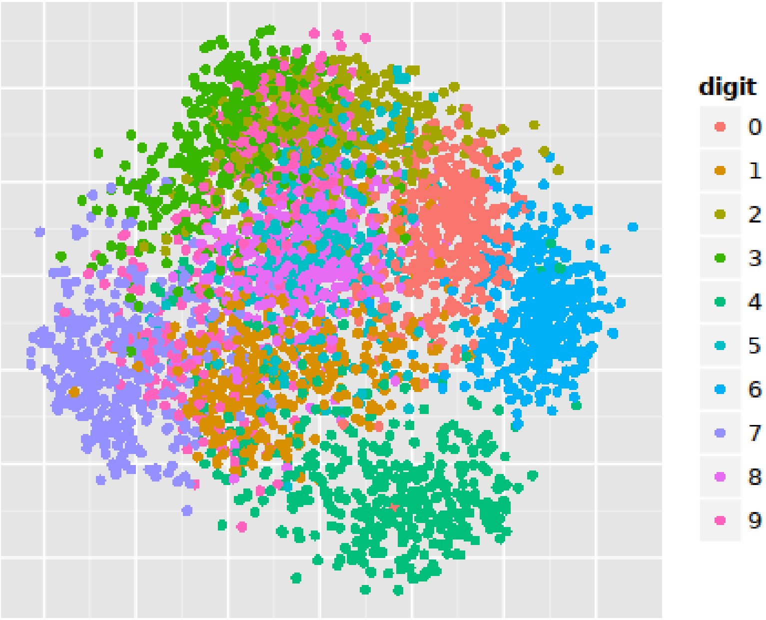
  "+"Spam emails

  "-" Normal/non-spam emails



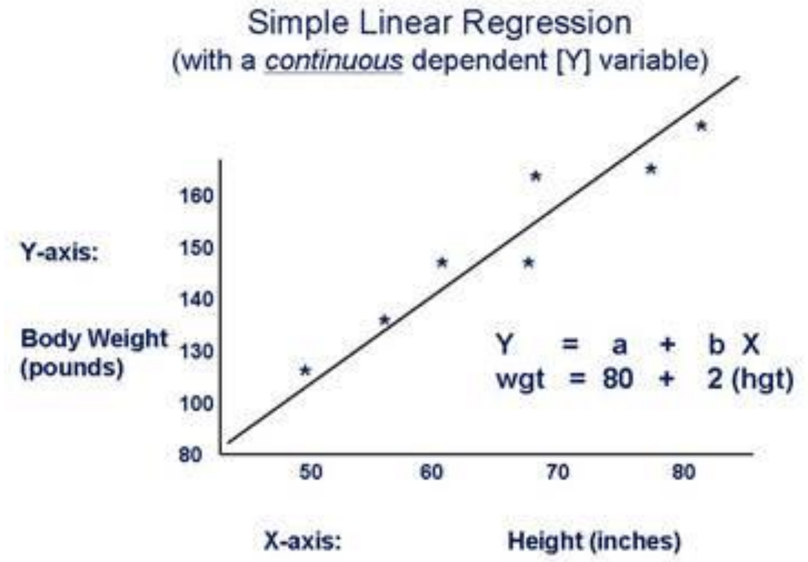
Multi-class Classification

- Digit Recognition
  - Map each image x to one of ten digits [0,…,9]

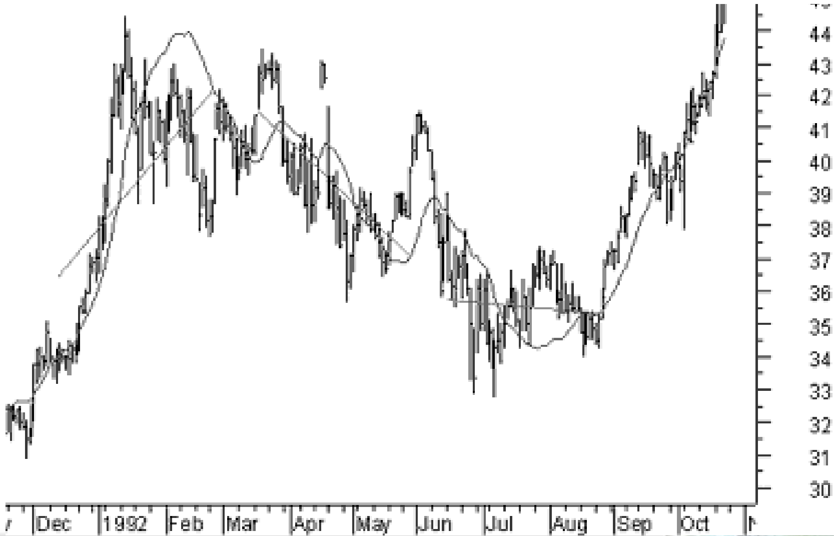


Regression

- Linear Regression
  - Assume linear dependence

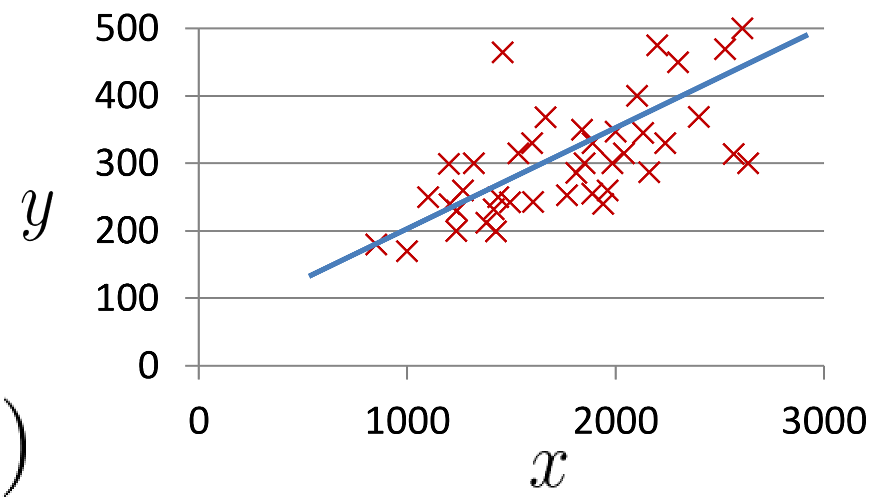


- Nonlinear Regression
  - Time series forecasting



Linear Regression

Hypothesis: $h(x) \approx f(x)$



- **Linear models**

$$h(x) = ax + b \approx f(x)$$

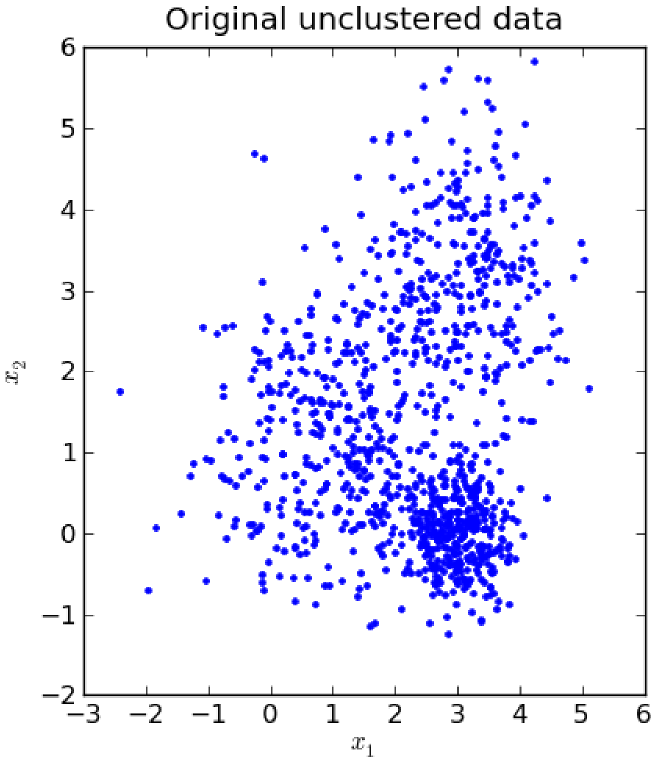
Infinite possible hypotheses!
Any choice of coefficient a and b forms a possible hypothesis

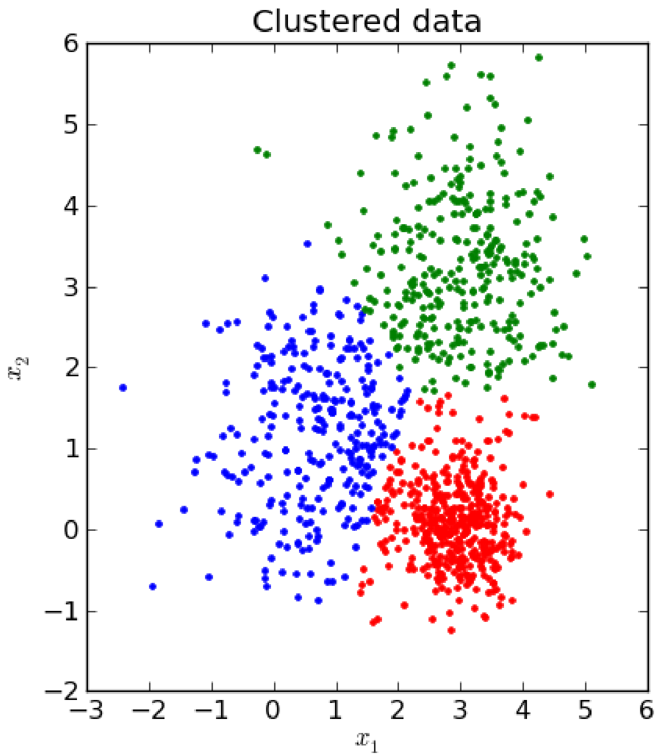
Unsupervised Learning

- Given (input, ~~correct output~~), (input, ?)
- Clustering
  - Find a set of prototypes representing the data
- Dimension Reduction / Principal Components
  - Find a subspace representing the data
- Independent components / dictionary learning
  - Find (small) set of factors for observation
- Novelty/Anomaly detection
  - Find the odd one out

Clustering

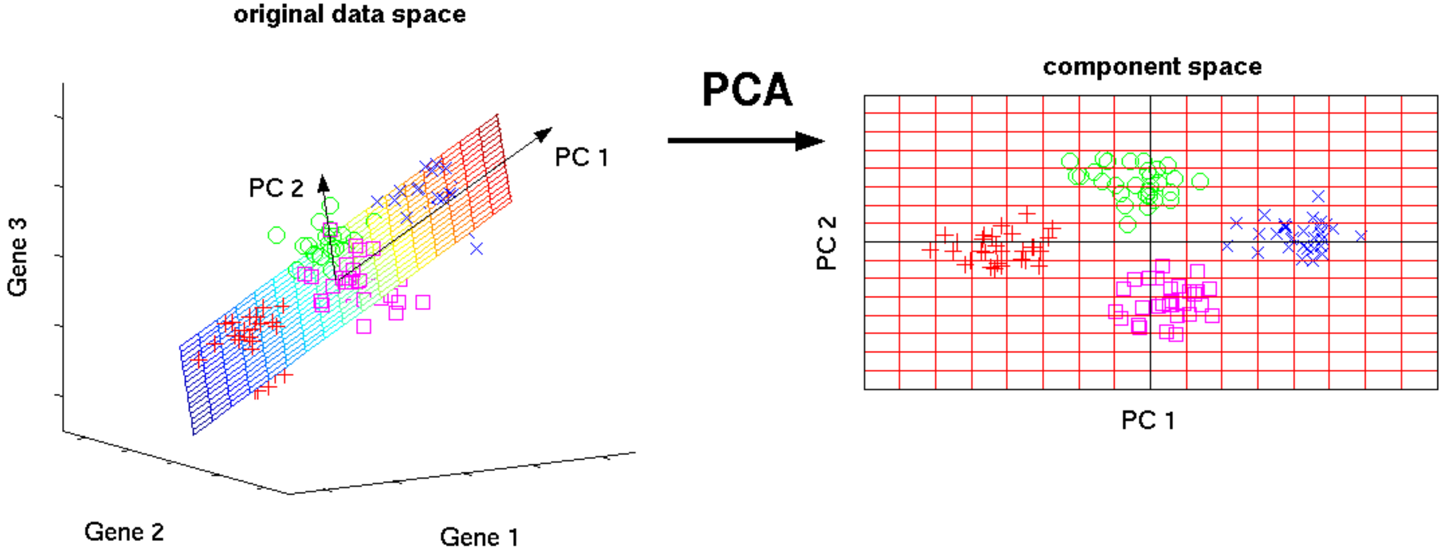
- Clustering Applications
  - Marketing segmentation, group of insurance interests, web news, pictures, city-planning, etc.



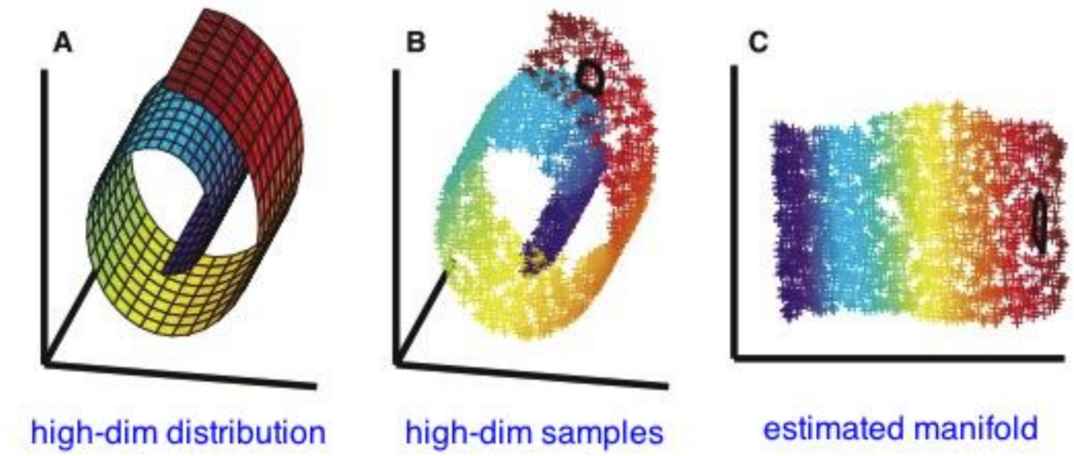


Dimension Reduction / Embedding

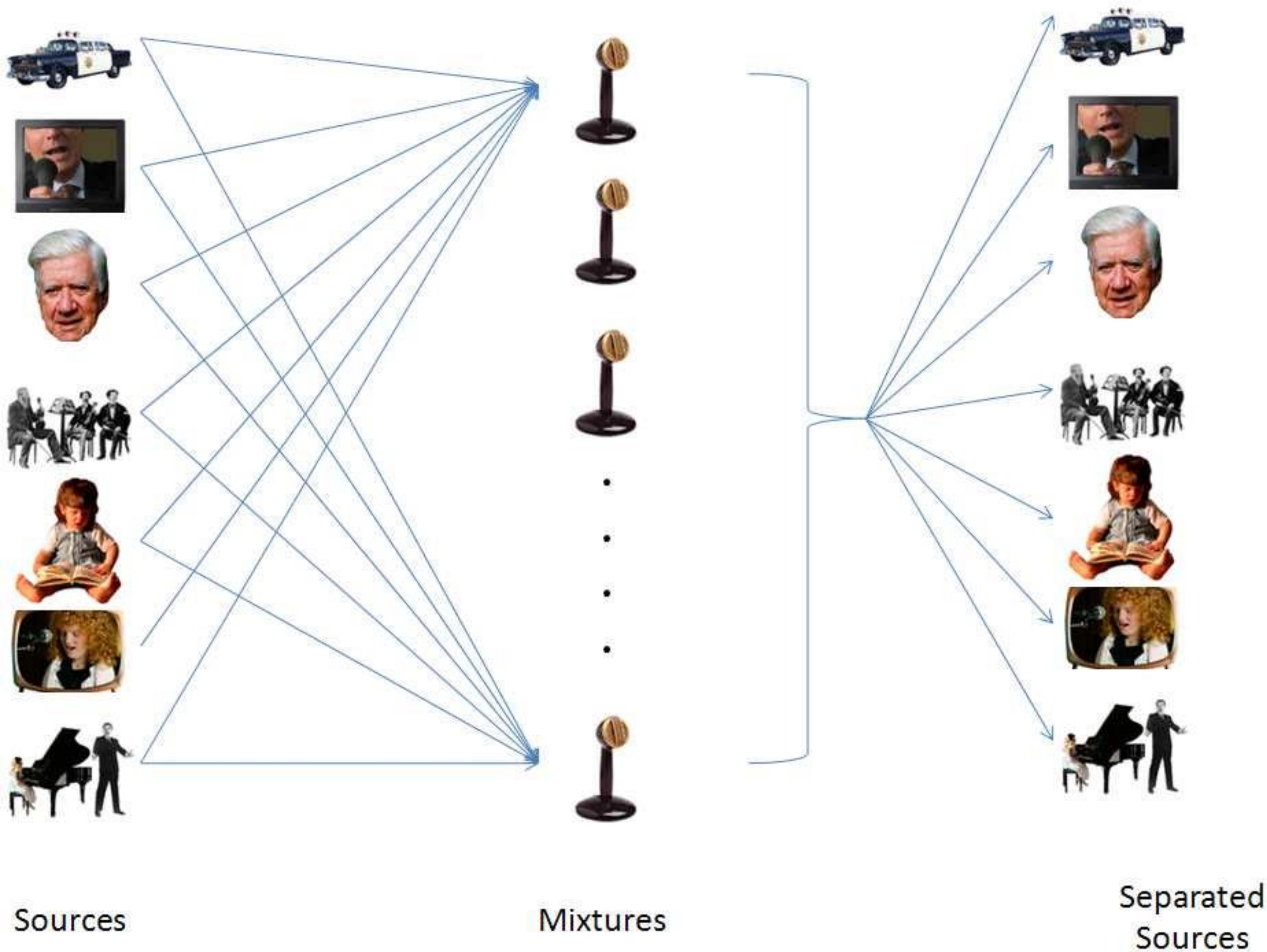
- Principal Component Analysis



- Nonlinear Embedding



Independent Components



Novelty / Anomaly Detection

- Novelty Detection: Identification of new or unknown patterns
  – Parametric approach
  – Non-parametric approach

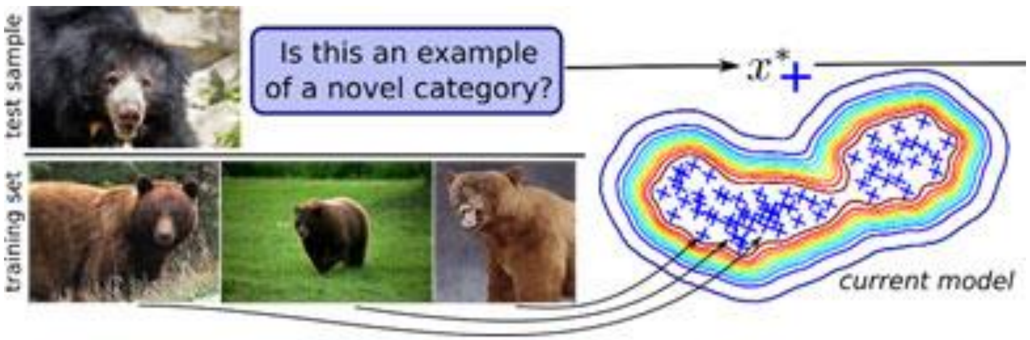
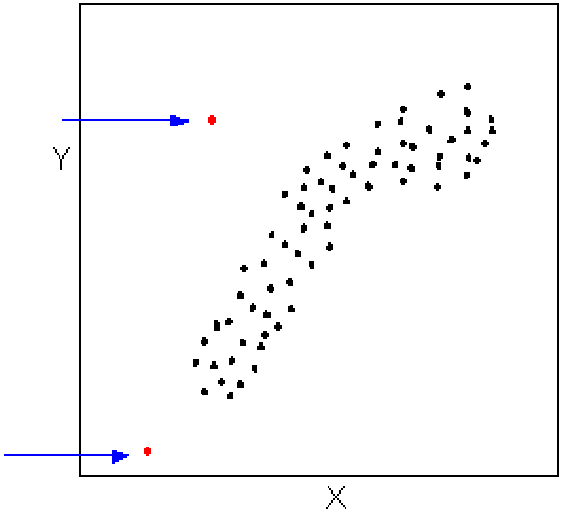


FIGURE 1 EXAMPLE OF NOVELTY DETECTION, BODESHEIM (2012).



Learning by Interacting with Environment

- Batch Learning
  - Observe training data (x1,y1) ... (xl,yl), then deploy
- Online Learning
  - Sequential: Observe x1, predict f(x1), observe x2, ...
  - E.g., Stock market forecasting
- Active learning
  - Query y for x, improve model, pick new x
  - E.g., ask questions in class
- Reinforcement Learning
  - Take action, environment responds, take new action
  - E.g., play chess, drive a car

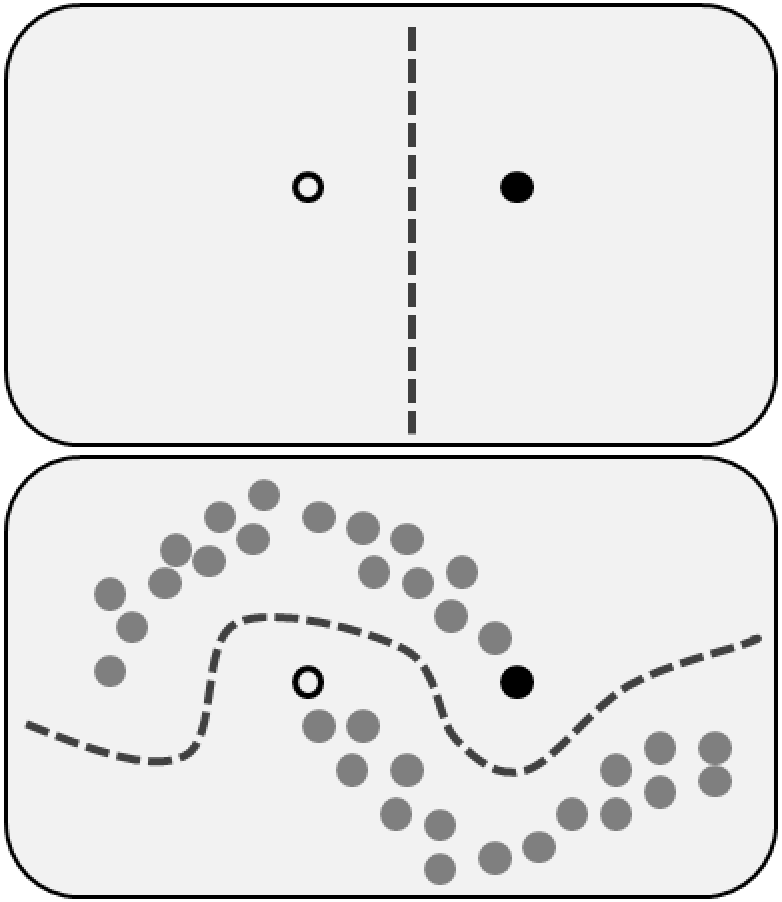
Reinforcement Learning

- Repeat
  - Take action
  - Environment reacts
  - Observe stuff
  - Update model

- Applications
  - Game playing
  - Self-driving cars
  - Autonomous plane flight

Inductive vs Transductive Learning

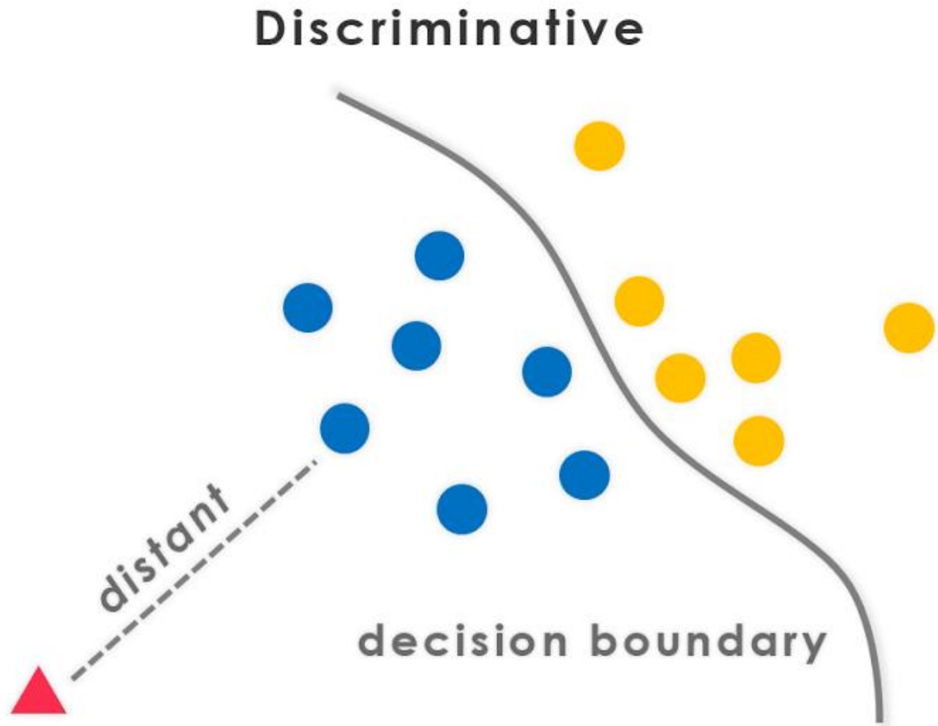
- **Induction**
  - Reasoning from observed training cases to general rules, which are then applied to the test cases.
  - Only have the training data. Do the best with it.
- **Transduction**
  - Reasoning from observed, specific (training) cases to specific (test) cases.
  - We have lots more problems that need to be solved with the same method.

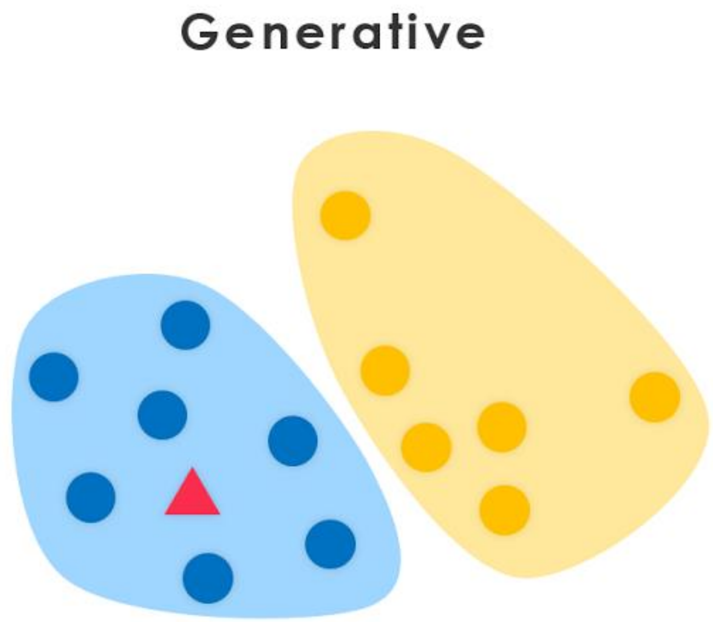


Discriminative vs. Generative (supervised learning)

- Discriminative Models
  - Estimate y|x directly
  - Often better convergence + simpler solutions
- Generative models
  - Estimate joint distribution over (x,y)
  - Use conditional probability to infer y|x
  - Often more intuitive
  - Easier to add prior knowledge

Discriminative vs. Generative

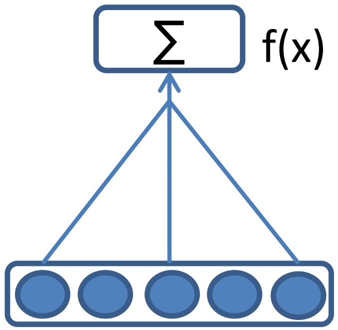




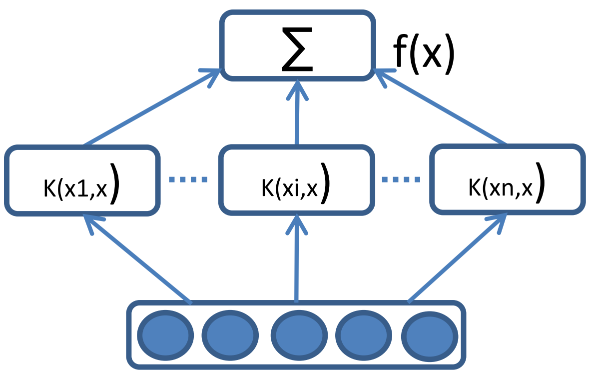
- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

- Model observations (x,y) first, then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

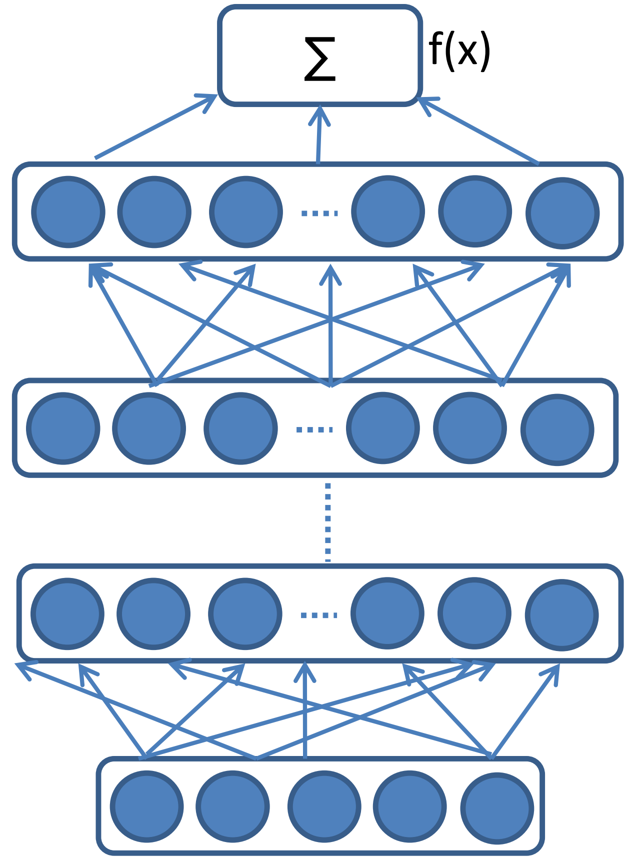
Shallow Learning vs Deep Learning



(a) Linear models



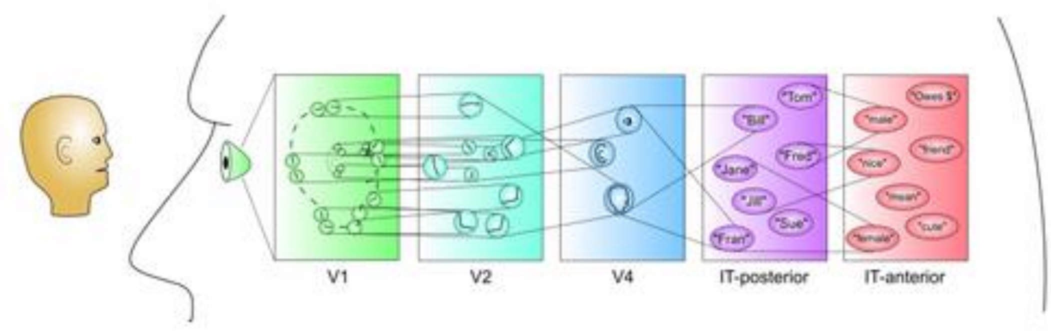
(b) Non-linear models
with shallow architecture

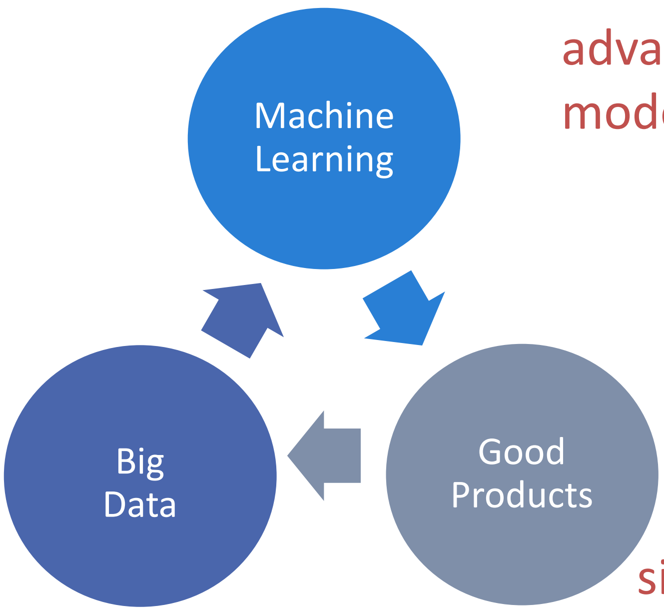


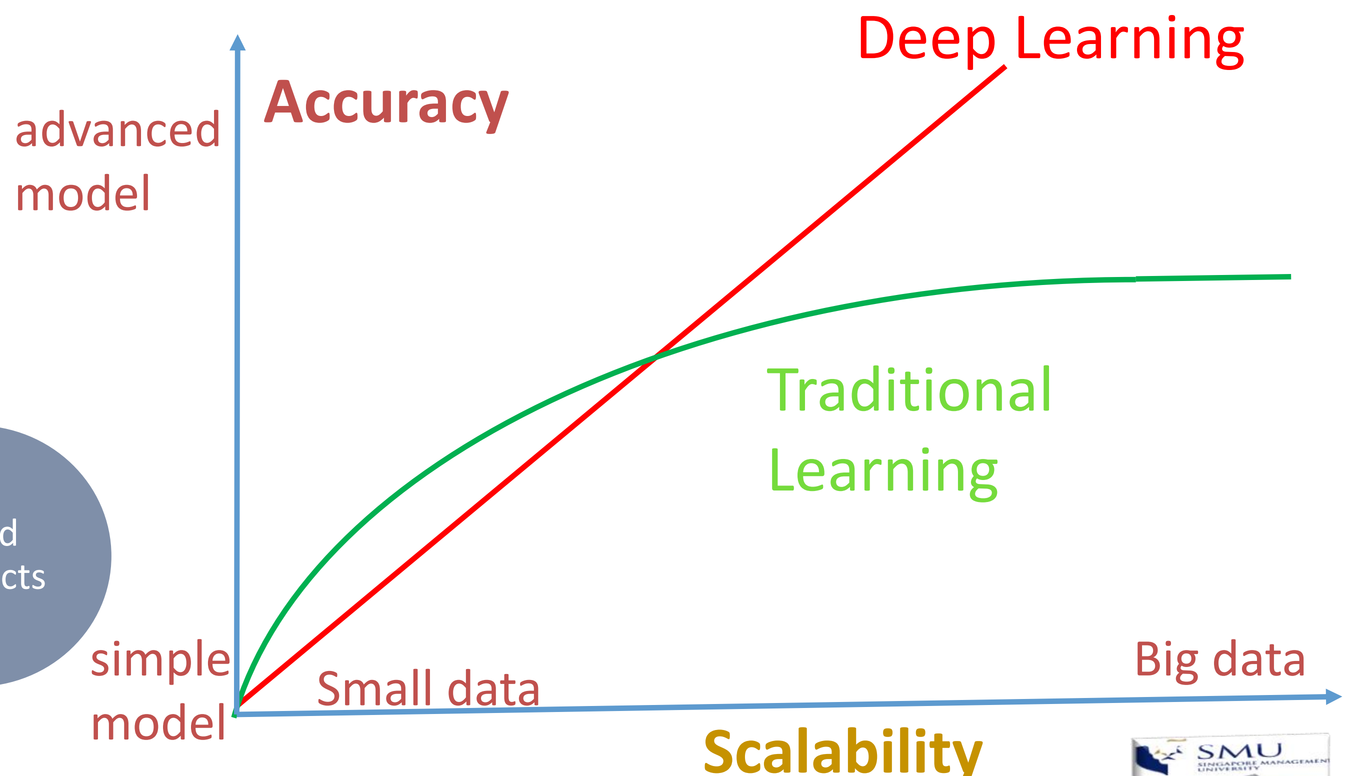
(c) Non-linear Model
with deep architecture

Why Deep Learning?

- A family of machine learning algorithms based on multi-layer networks
- Inspired by the biological architecture of brain in neuroscience







Evaluation of Learning Systems

- **Experimental**
  - Conduct controlled cross-validation experiments to compare various methods on a variety of benchmark datasets.
  - Gather data on their performance, e.g. test accuracy, training-time, testing-time.
  - Analyze differences for statistical significance.
- **Theoretical**
  - Analyze algorithms mathematically and prove theorems about their:
    - Computational complexity
    - Ability to fit training data
    - Sample complexity (number of training examples needed to learn an accurate function)

What We’ll Cover

- **Supervised learning**
  - Linear models for regression
  - Linear models for classification
  - Bayesian learning
  - Support vector machines
  - Neural networks and deep learning
  - Model ensembles
- **Unsupervised learning**
  - Clustering
  - Dimension Reduction/Feature learning

Important Issues in Machine Learning

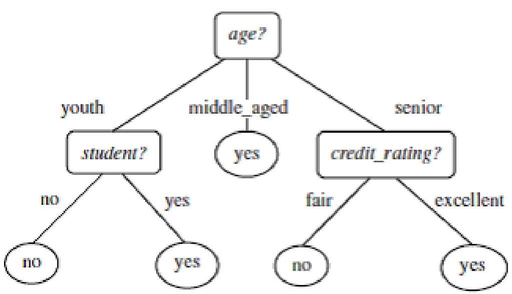
- **Obtaining experience**
  - How to obtain experience?
    - Supervised learning vs. Unsupervised learning
  - How many examples are enough?
    - PAC learning theory
- **Learning algorithms**
  - What algorithm can approximate function well, when?
  - How does the complexity of learning algorithms impact the learning accuracy?
  - Whether the target function is learnable?
- **Representing inputs**
  - How to represent the inputs?
  - How to remove the irrelevant information from the input representation?
  - How to reduce the redundancy of the input representation?

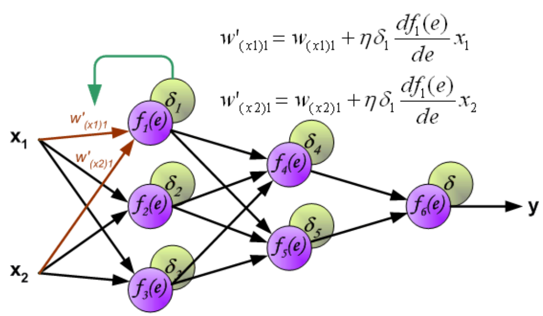
ML in Practice

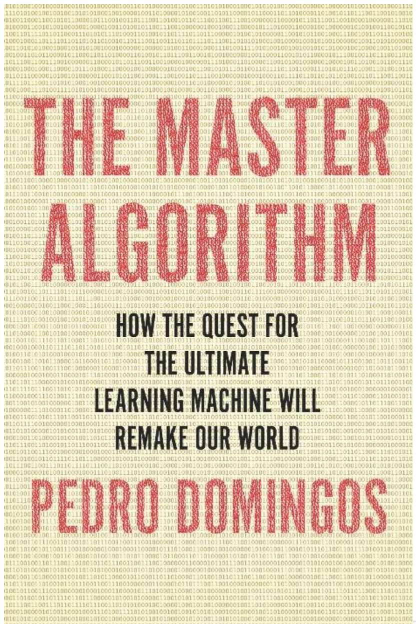
- Understanding domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learning models
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop

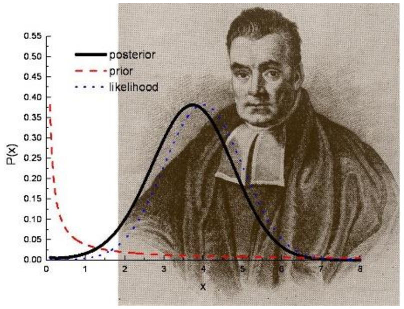
The Five Tribes of Machine Learning

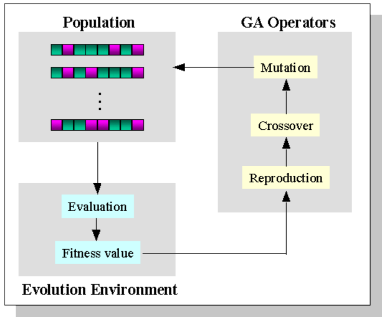
Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

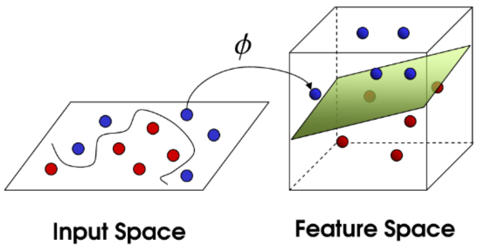












History of Machine Learning

- 1950s
  - Samuel’s checker player
  - Selfridge’s Pandemonium
- 1960s:
  - Neural networks: Perceptron
  - Pattern recognition
  - Learning in the limit theory
  - Minsky and Papert prove limitations of Perceptron
- 1970s:
  - Symbolic concept induction
  - Winston’s arch learner
  - Expert systems and the knowledge acquisition bottleneck
  - Quinlan’s ID3
  - Michalski’s AQ and soybean diagnosis
  - Scientific discovery with BACON
  - Mathematical discovery with AM

History of Machine Learning (cont.)

- 1980s:
  - Advanced decision tree and rule learning
  - Explanation-based Learning (EBL)
  - Learning and planning and problem solving
  - Utility problem
  - Analogy
  - Cognitive architectures
  - Resurgence of neural networks (connectionism, backpropagation)
  - Valiant's PAC Learning Theory
  - Focus on experimental methodology
- 1990s
  - Data mining
  - Adaptive software agents and web applications
  - Text learning
  - Reinforcement learning (RL)
  - Inductive Logic Programming (ILP)
  - Ensembles: Bagging, Boosting, and Stacking
  - Bayes Net learning

History of Machine Learning (cont.)

- 2000s
  - Support vector machines
  - Kernel methods
  - Graphical models
  - Statistical relational learning
  - Transfer learning
  - Sequence labeling
  - Collective classification and structured outputs
  - Computer Systems Applications
    - Compilers, Debugging
    - Graphics
    - Security (intrusion, virus, and worm detection)
  - E mail management
  - Personalized assistants that learn
  - Learning in robotics and vision
  - Online Learning
  - Large-scale Machine Learning
  - Deep Learning / Deep Neural Networks

Publication Venues

- Journals
  - Journal of Machine Learning Research
  - Machine Learning
  - Neurocomputing
  - Neural Computation
  - IEEE Trans on Neural Networks & Learning Systems
  - Artificial Intelligence, Neural Networks, etc
- Conferences
  - ICML
  - NIPS
  - COLT
  - ECML, ACML, AISTATS, ALT, AAAI, IJCAI, KDD, etc

Recommended References

- Machine Learning (by Tom Mitchell, McGraw Hill, 1997).
- Pattern Recognition and Machine Learning (by Chris Bishop 2006).
- Machine Learning: a Probabilistic Perspective (by Kevin Murphy 2012).
- The Elements of Statistical Learning (Trevor Hastie and Robert Tibshirani)
- The Nature of Statistical Learning Theory (Vladimir N. Vapnik)
- Learning from Data (Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin)

QUESTIONS?!