Introduction to Machine Learning and Data Mining

DCS310

Sun Yat-sen University

Chang-Dong Wang

wangchd3@mail.sysu.edu.cn

Slides:
Eric Eaton@University of Pennsylvania
Jure Leskovec@Stanford University
Chang-Dong Wang@Sun Yat-sen University

What is Machine Learning?

“Learning is any process by which a system improves performance from experience.”

- Herbert Simon (Nobel Prize,Turing Award)

Definition by Tom Mitchell (1998):

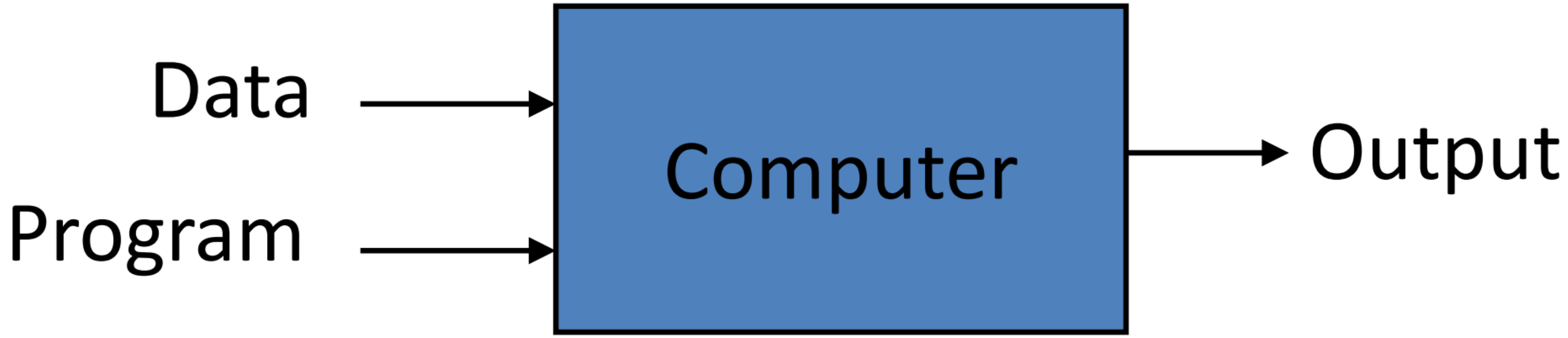
Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E.

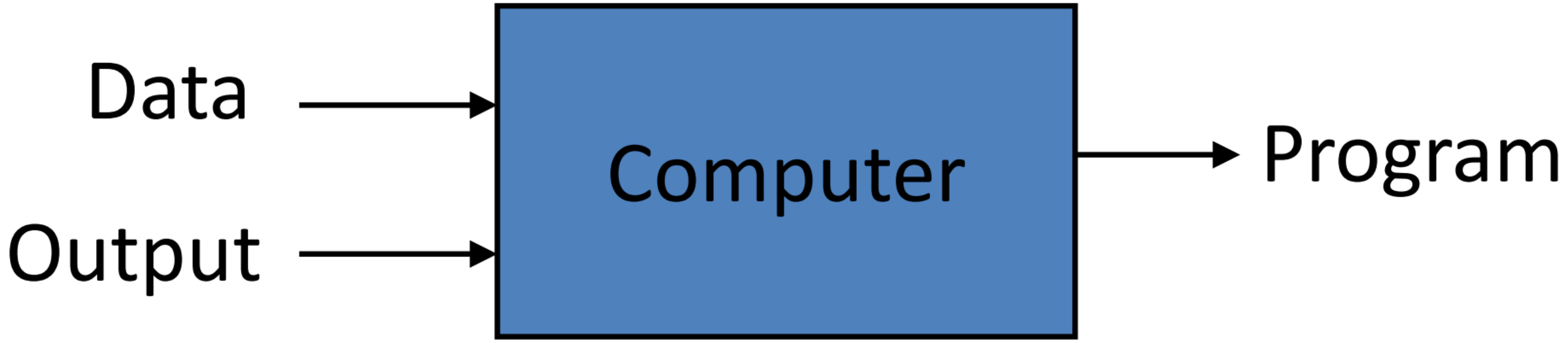
A well-defined learning task is given by <P, T, E>.

Machine Learning vs. Traditional Programming

Traditional Programming



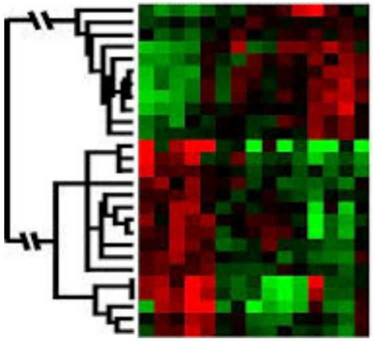
Machine Learning



When Do We Use Machine Learning?

ML is used when:

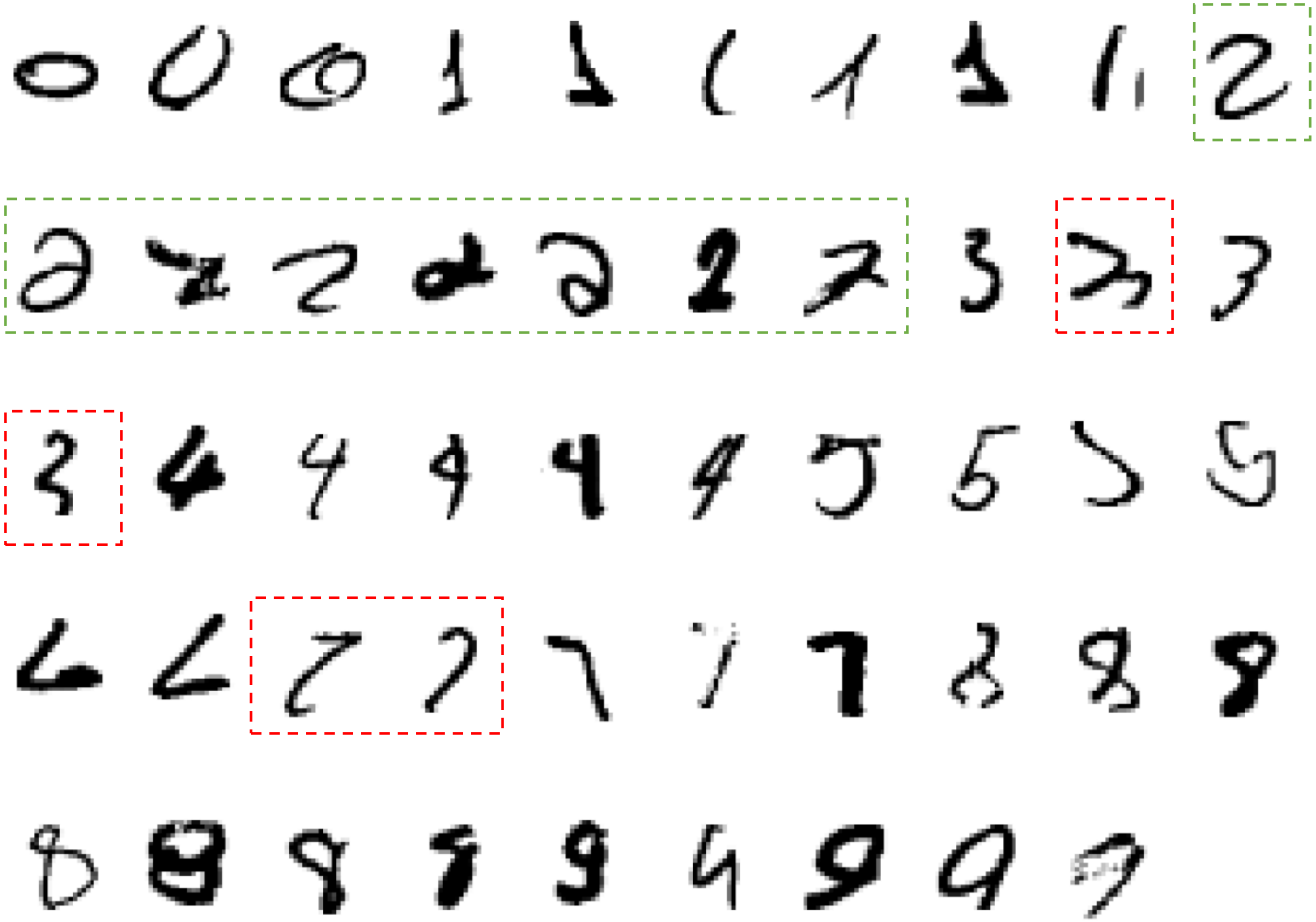
- Human expertise does not exist (navigating on Mars)
- Humans can’t explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)



Learning isn’t always useful:

- There is no need to “learn” to calculate payroll

A classic example of a task that requires machine learning: It is very hard to say what makes a 2



Some more examples of tasks that are best solved by using a learning algorithm

- Recognizing patterns:

- Facial identities or facial expressions

- Handwritten or spoken words

- Medical images

- Generating patterns:

- Generating images or motion sequences

- Recognizing anomalies:

- Unusual credit card transactions

- Unusual patterns of sensor readings in a nuclear power plant

- Prediction:

- Future stock prices or currency exchange rates

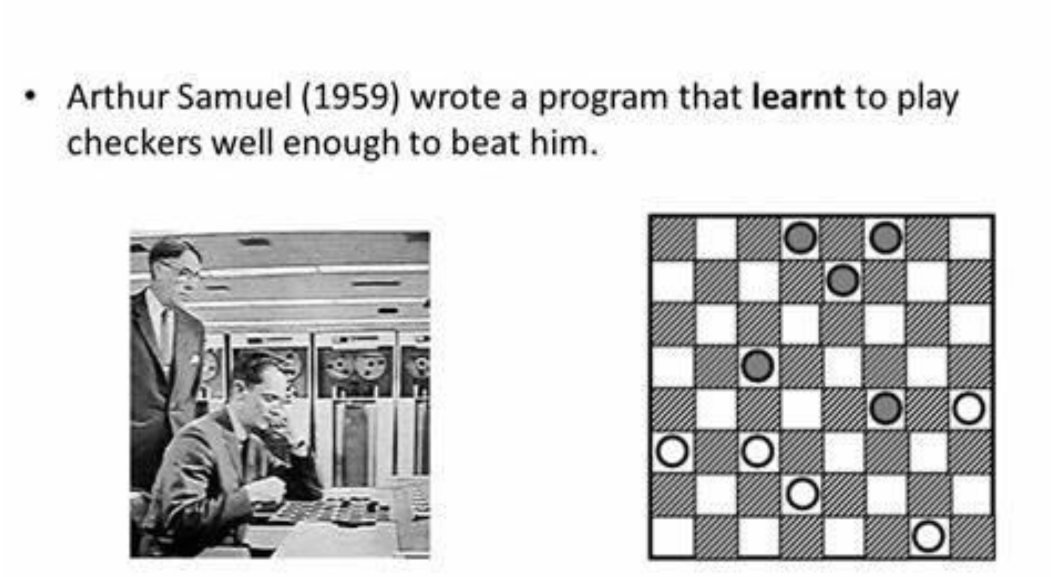
Sample Applications

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

Samuel’s Checkers-Player

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.”

-Arthur Samuel (1959)



- Arthur Samuel (1959) wrote a program that **learnt** to play checkers well enough to beat him.

Defining the Learning Task

Improve on task T, with respect to performance metric P, based on experience E

T: Playing checkers

P: Percentage of games won against an arbitrary opponent

E: Playing practice games against itself

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors

P: Average distance traveled before a human-judged error

E: A sequence of images and steering commands recorded while observing a human driver.

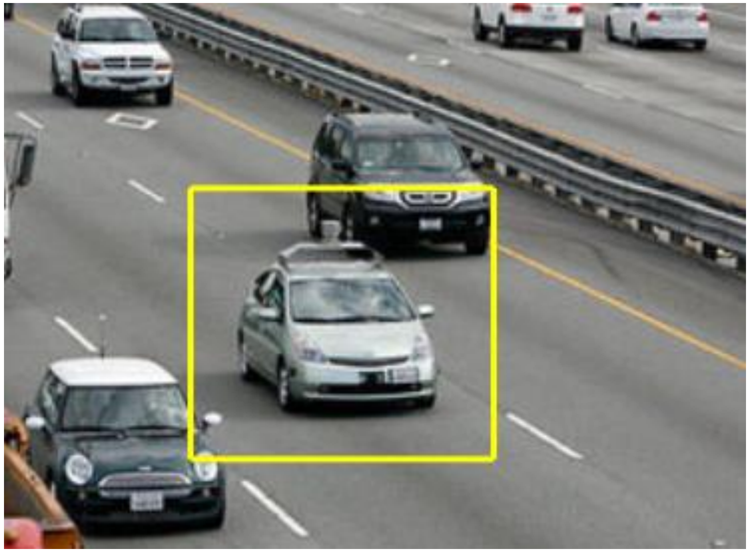
T: Categorize email messages as spam or legitimate

P: Percentage of email messages correctly classified

E: Database of emails, some with human-given labels

State of the Art Applications of Machine Learning

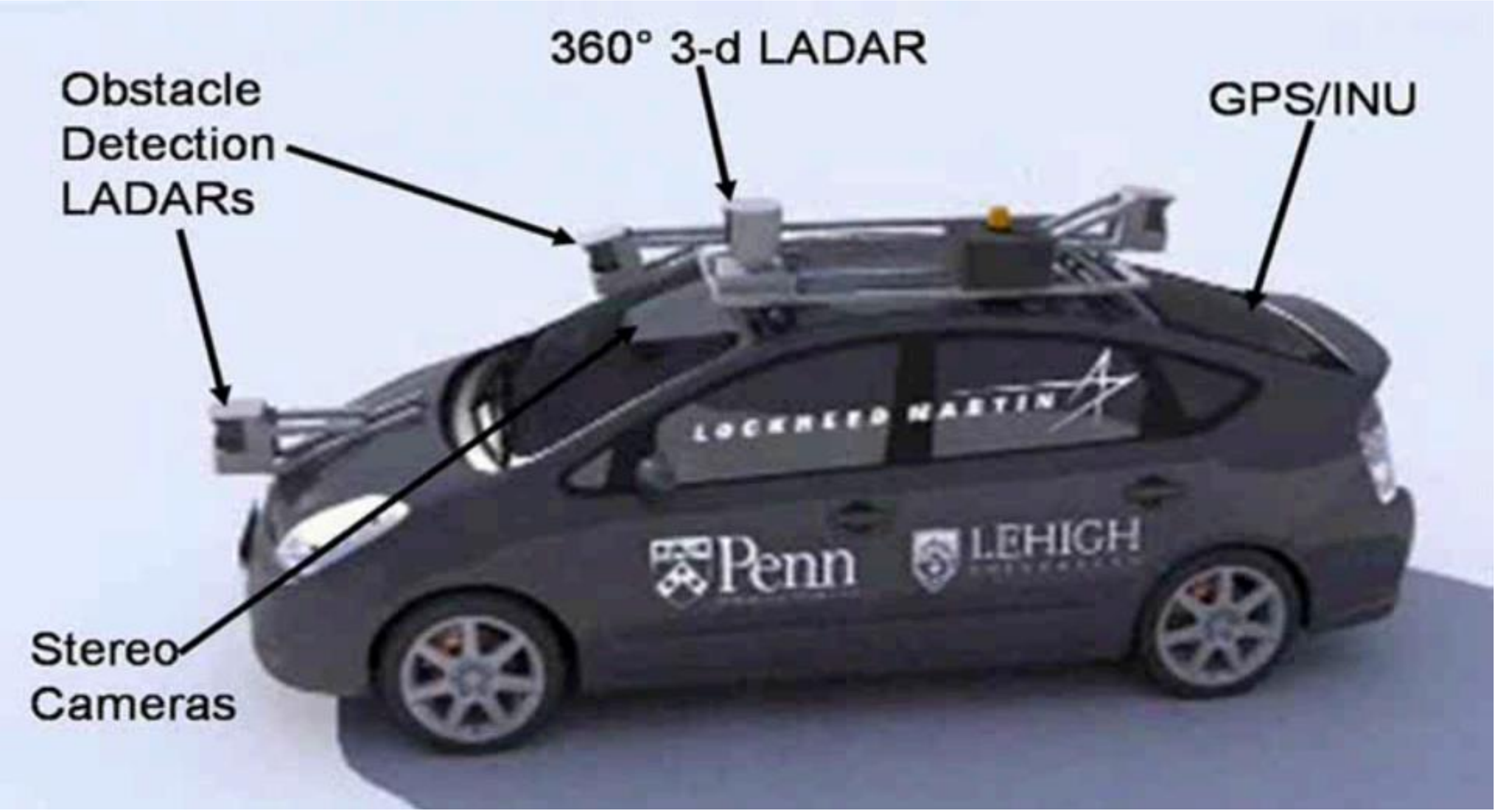
Autonomous Cars



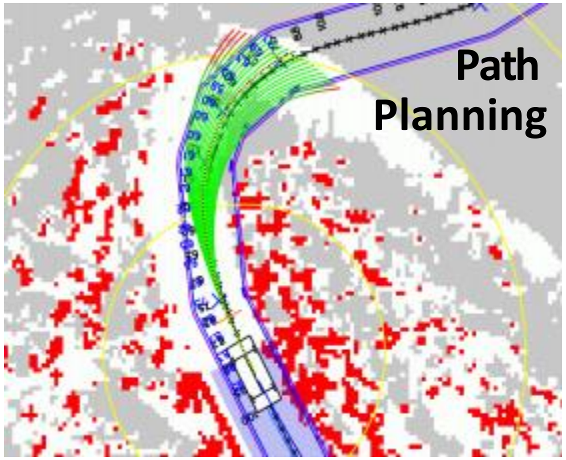
- Nevada made it legal for autonomous cars to drive on roads in June 2011
- As of 2013, four states (Nevada, Florida, California, and Michigan) have legalized autonomous cars

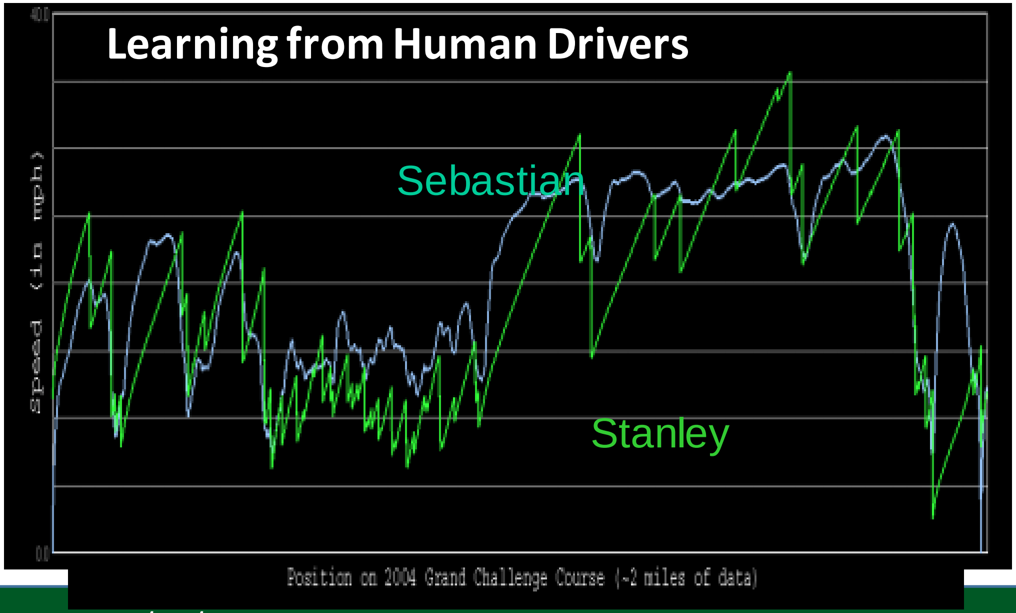
Penn's Autonomous Car →
(Ben Franklin Racing Team)

Autonomous Car Sensors

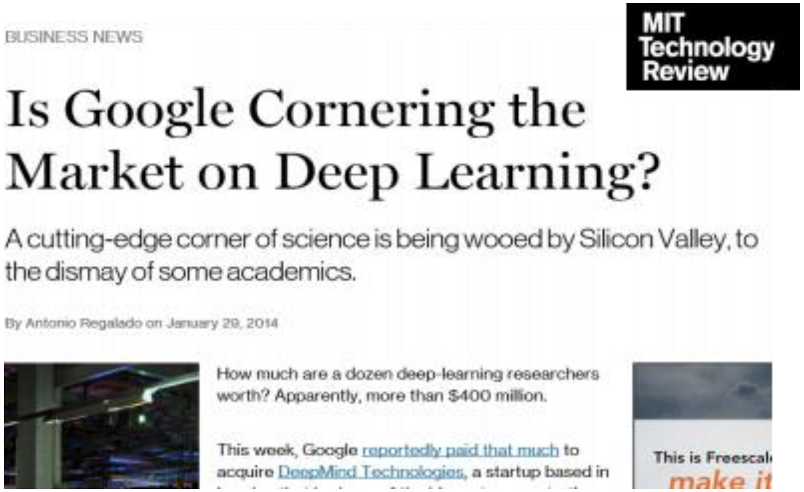


Autonomous Car Technology





Deep Learning in the Headlines



BUSINESS NEWS

MIT Technology Review

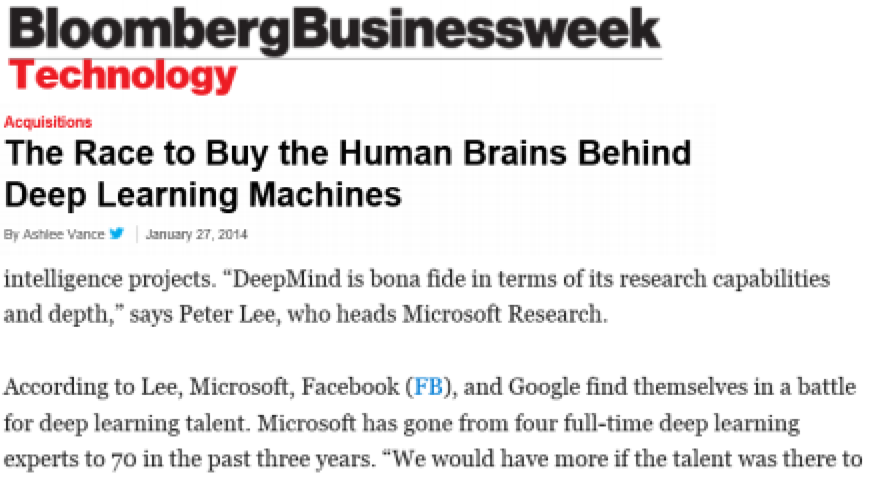
Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

This week, Google reportedly paid that much to acquire DeepMind Technologies, a startup based in



BloombergBusinessweek
Technology

Acquisitions

The Race to Buy the Human Brains Behind Deep Learning Machines

By Ashlee Vance | January 27, 2014

intelligence projects. "DeepMind is bona fide in terms of its research capabilities and depth," says Peter Lee, who heads Microsoft Research.

According to Lee, Microsoft, Facebook (FB), and Google find themselves in a battle for deep learning talent. Microsoft has gone from four full-time deep learning experts to 70 in the past three years. "We would have more if the talent was there to

Deep Learning's Role in the Age of Robots

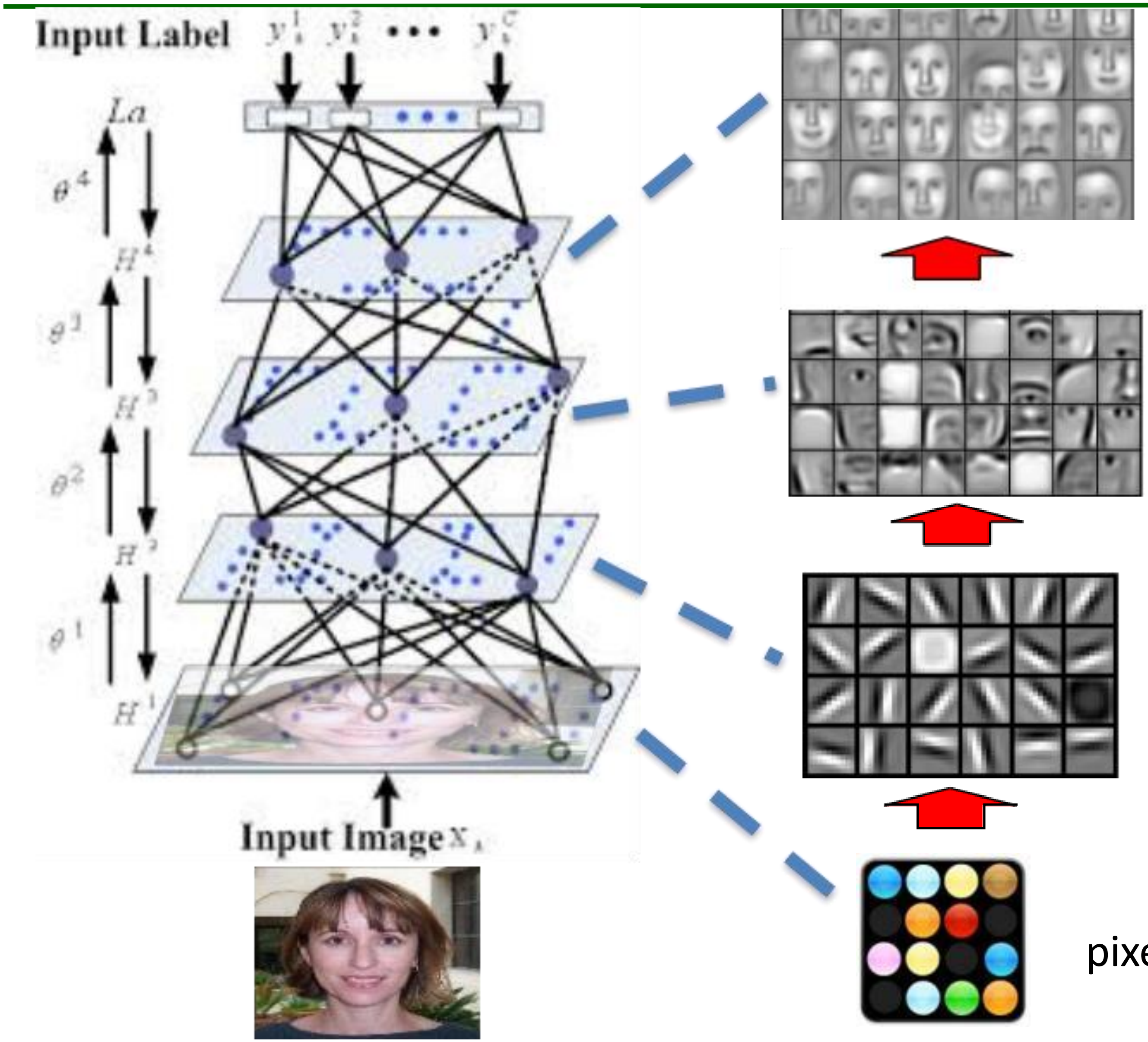
BY JULIAN GREEN, JETPAC 05.02.14 2:56 PM

DEEP LEARNING

- Computers learning and growing on their own
- Able to understand complex, massive amounts of data

DATA ECONOMY
DEEP LEARNING

Deep Belief Net on Face Images



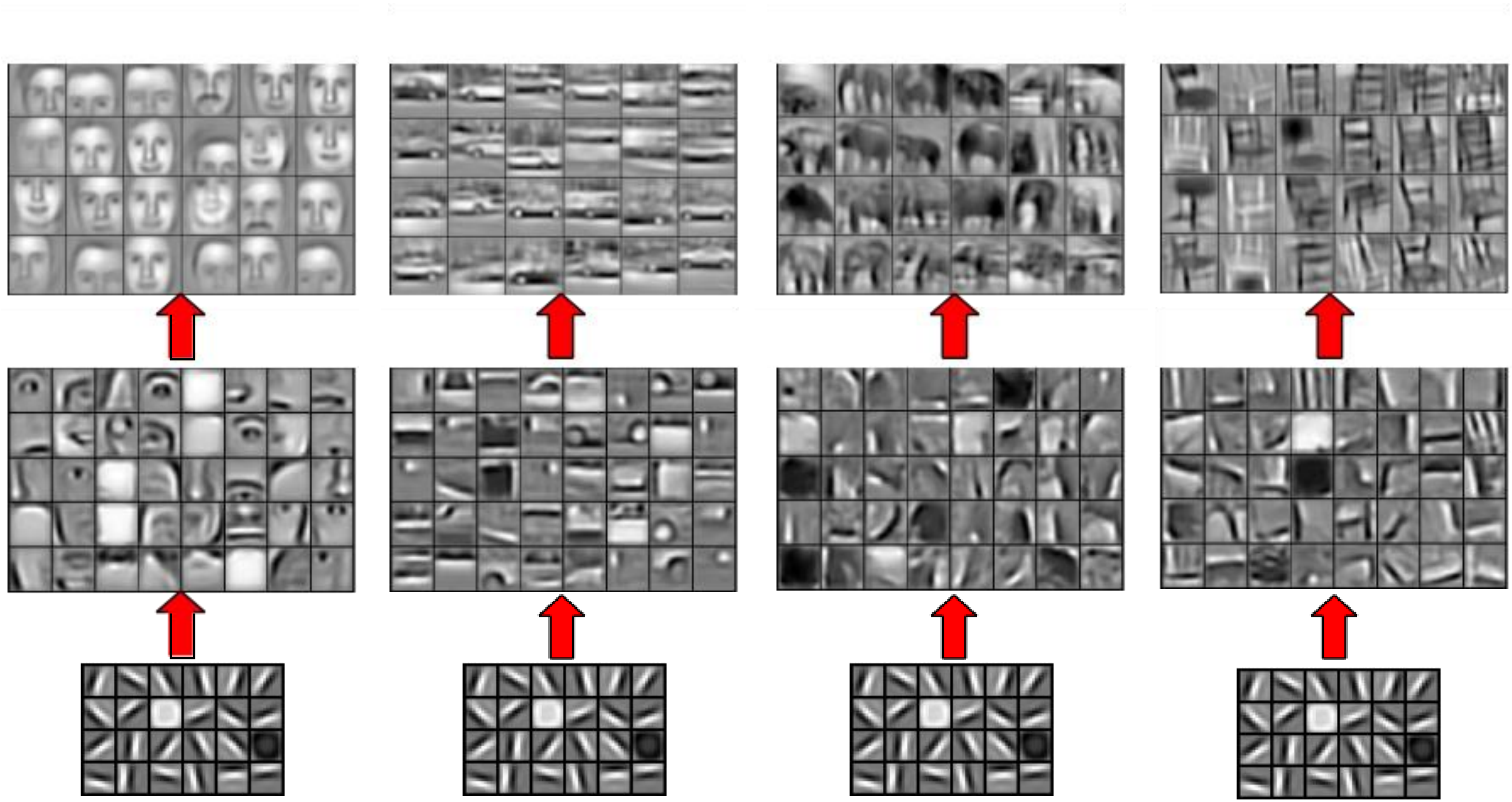
object models

object parts (combination of edges)

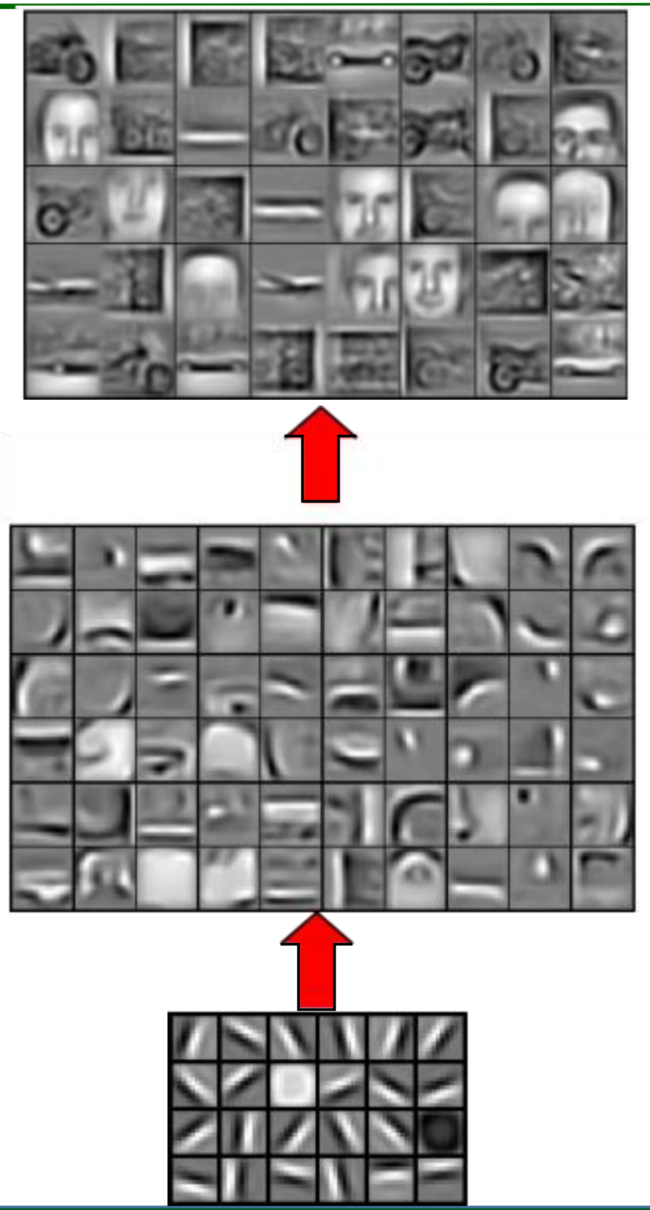
edges

pixels

Learning of Object Parts



Training on Multiple Objects

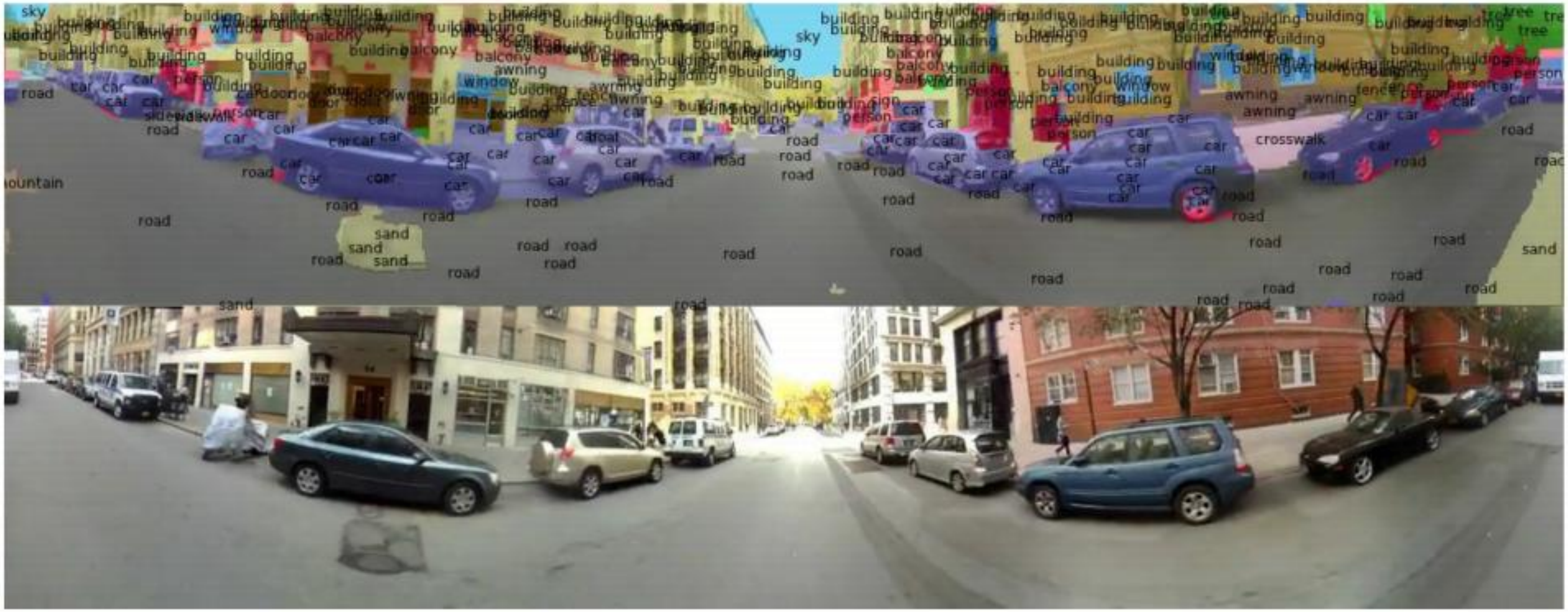


Trained on 4 classes (cars, faces, motorbikes, airplanes).

Second layer: Shared-features and object-specific features.

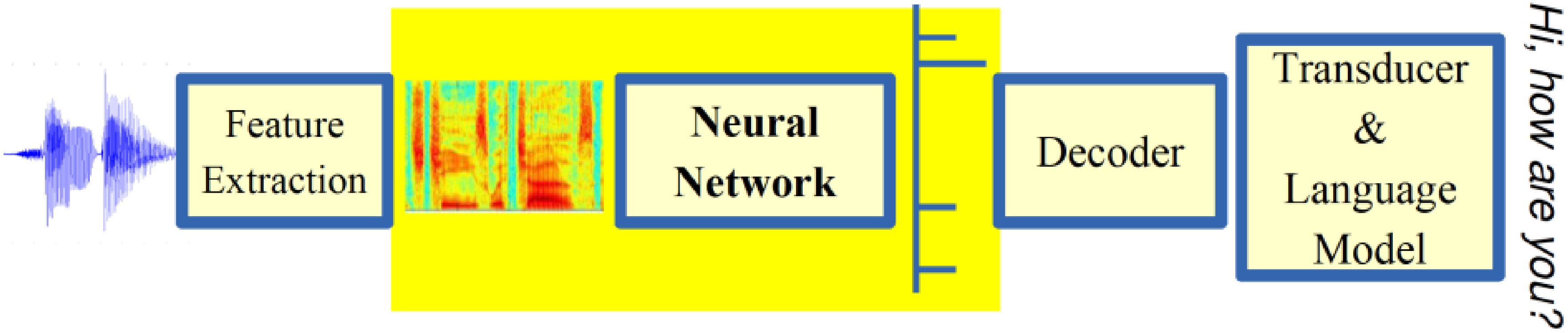
Third layer: More specific features.

Scene Labeling via Deep Learning

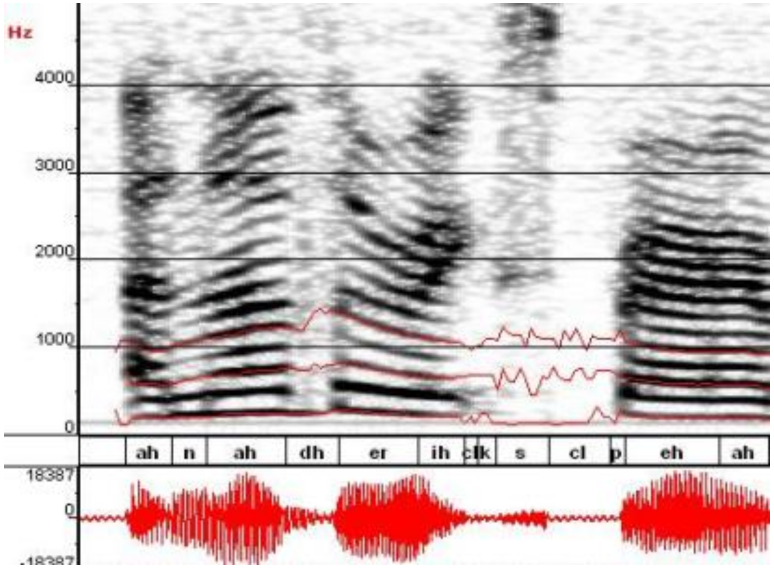


Machine Learning in Automatic Speech Recognition

A Typical Speech Recognition System



ML used to predict phone states from the sound spectrogram



Deep learning has state-of-the-art results

# Hidden Layers	1	2	4	8	10	12
Word Error Rate %	16.0	12.8	11.4	10.9	11.0	11.1

Baseline GMM performance = 15.4%

[Zeiler et al. “On rectified linear units for speech recognition” ICASSP 2013]

Impact of Deep Learning in Speech Technology

您需要做什么?

打开手机中国

正在打开手机中国

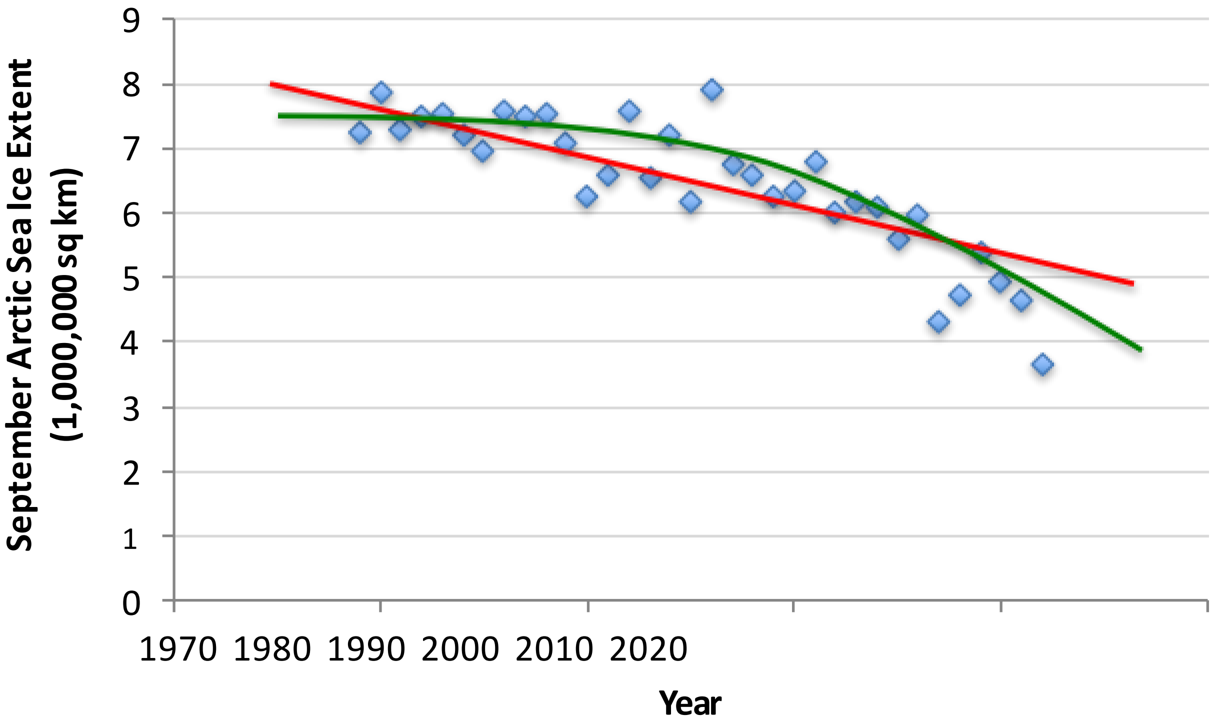
Types of Learning

Types of Learning

- **Supervised (inductive) learning**
  - Given: training data + desired outputs (labels)
- **Unsupervised learning**
  - Given: training data (without desired outputs)
- **Semi-supervised learning**
  - Given: training data + a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning: Regression

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function f(x) to predict y given x
  - y is real-valued == regression

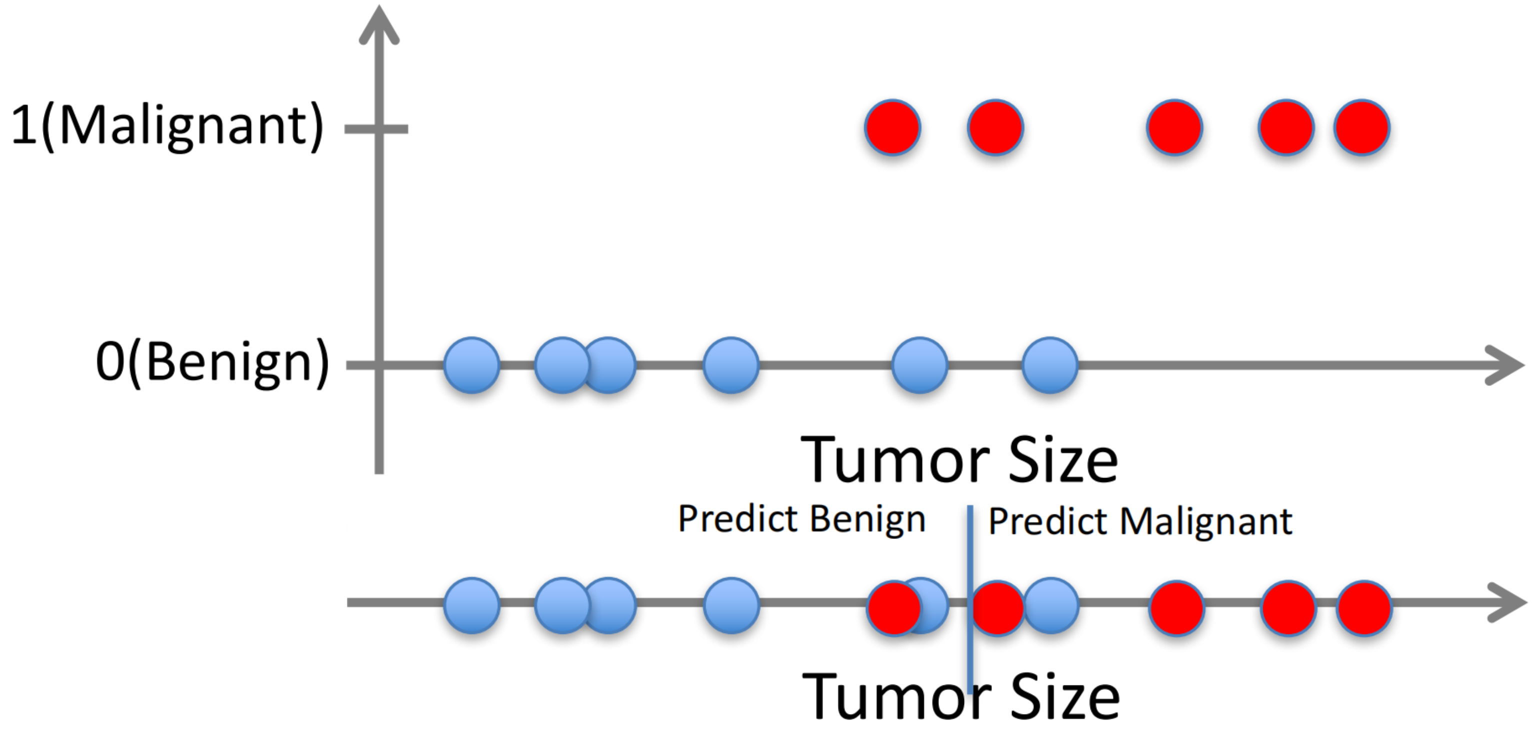


Supervised Learning: Classification

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- Learn a function f(x) to predict y given x
  - y is categorical == classification

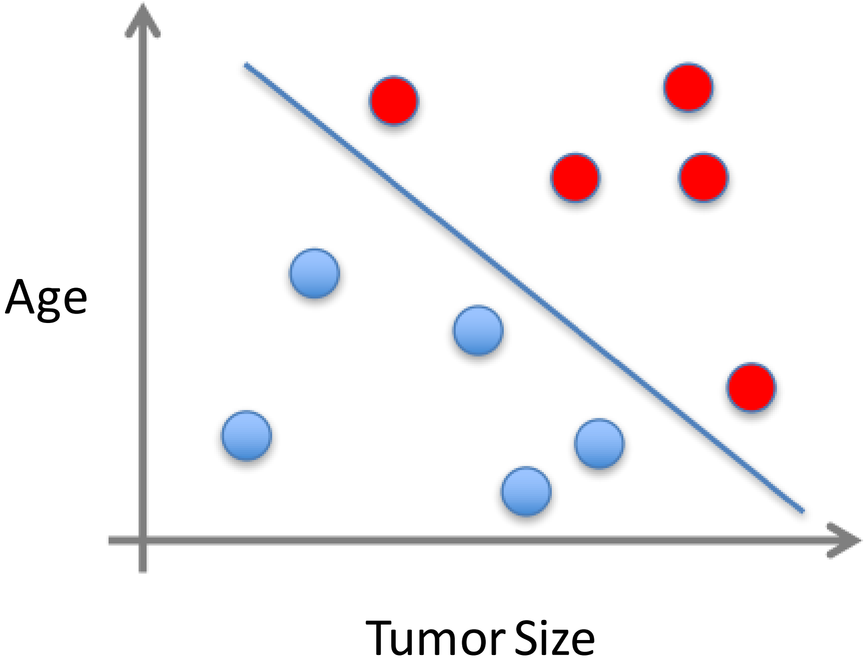
Breast Cancer (Malignant / Benign)



Supervised Learning

x can be multi-dimensional

- Each dimension corresponds to an attribute

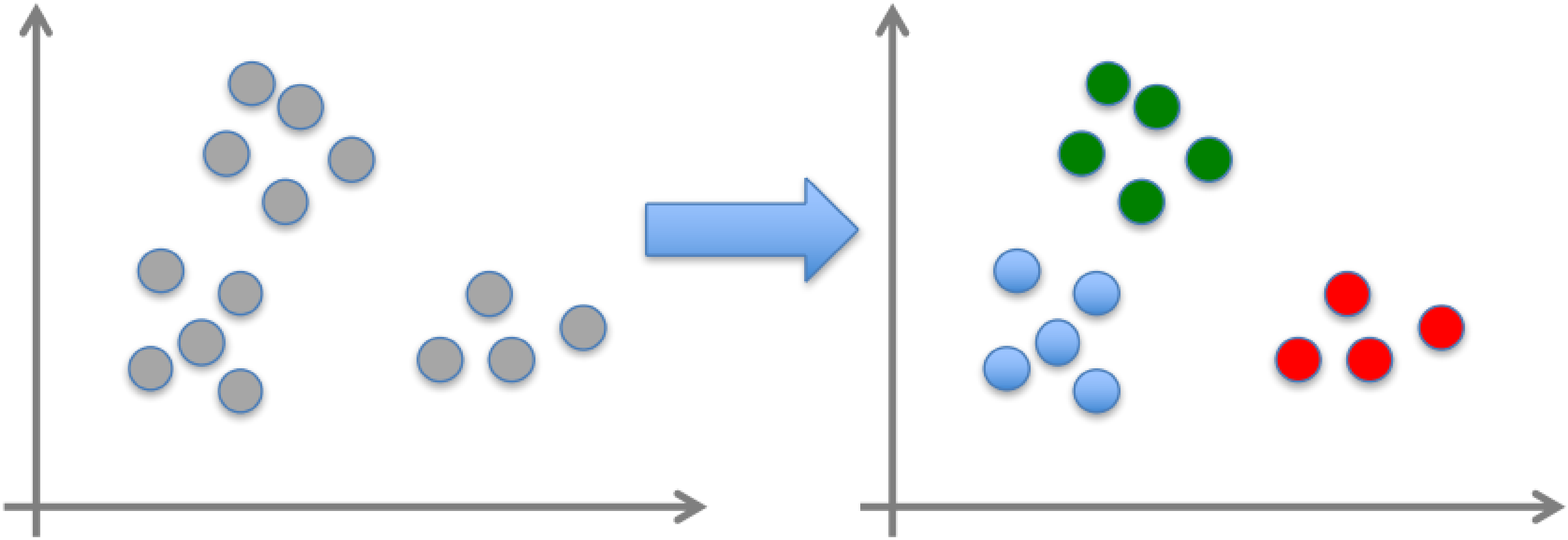


- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

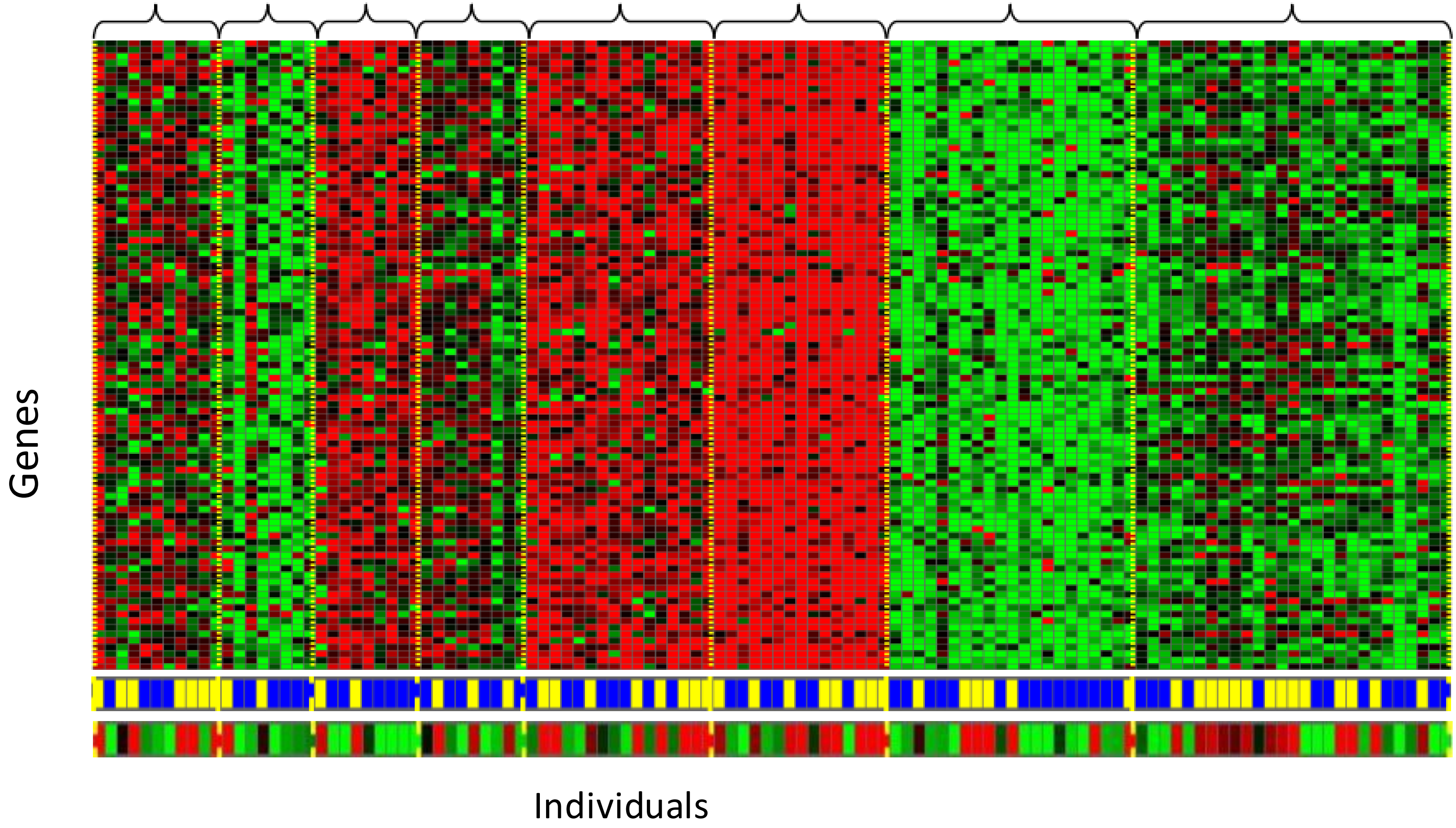
Unsupervised Learning

- Given x_1, x_2, ..., x_n (without labels)
- Output hidden structure behind the x's
  - – E.g., clustering



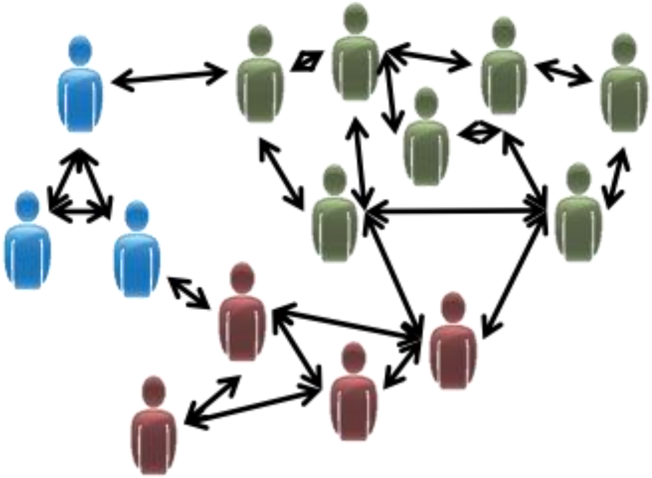
Unsupervised Learning

Genomics application: group individuals by genetic similarity



Unsupervised Learning

Organize computing clusters



Social network analysis

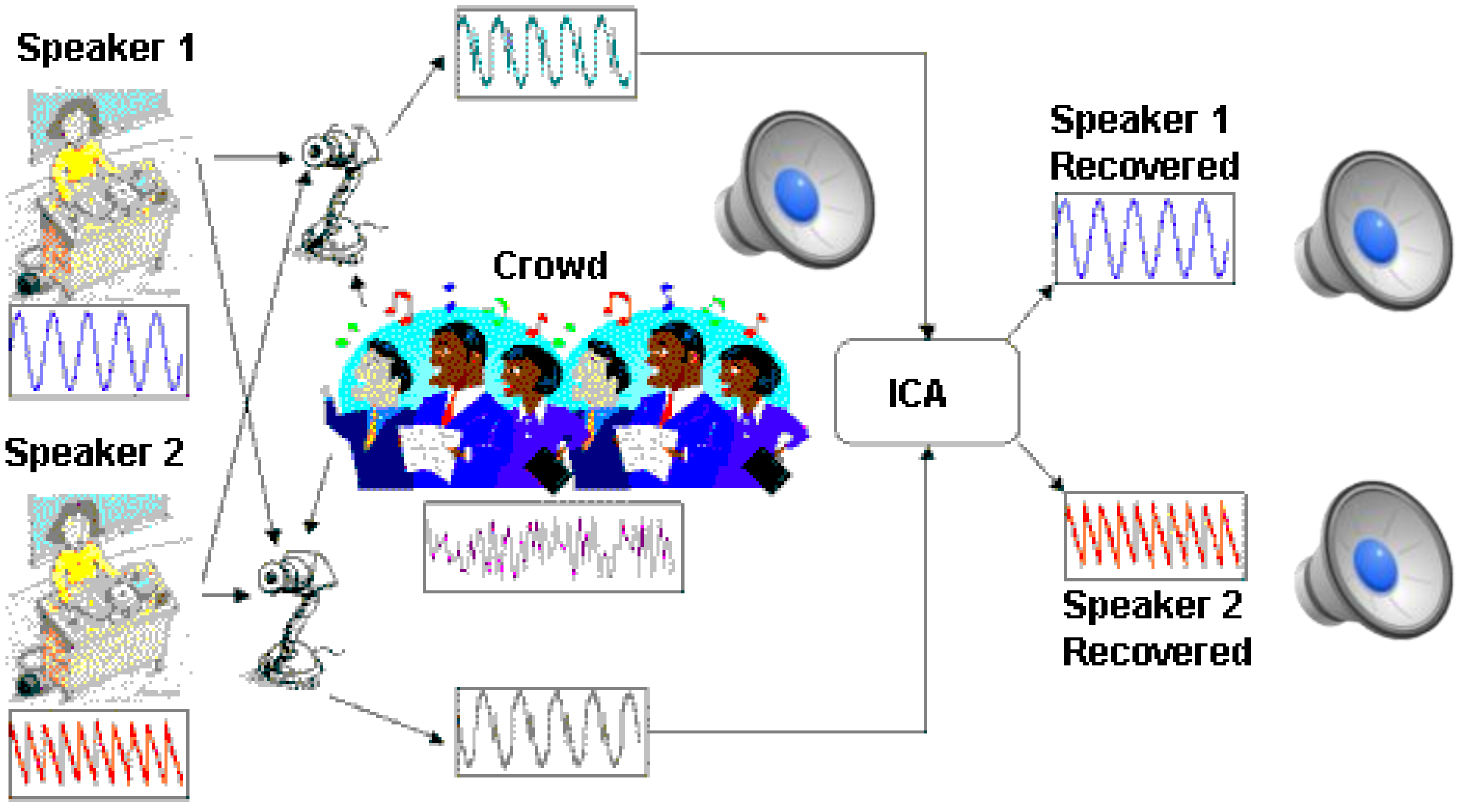
Market segmentation

Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

Astronomical data analysis

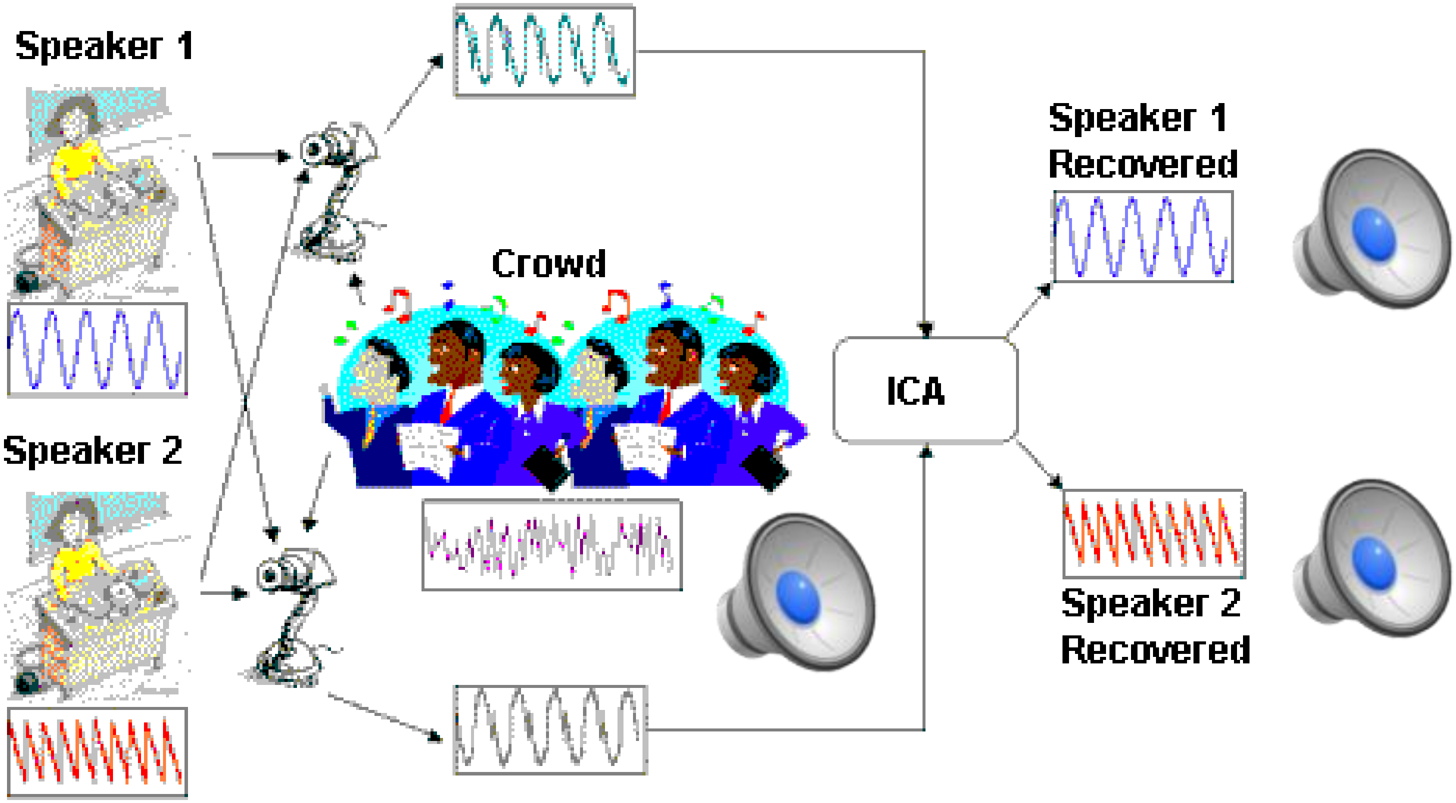
Unsupervised Learning

- Independent component analysis – separate a combined signal into its original sources



Unsupervised Learning

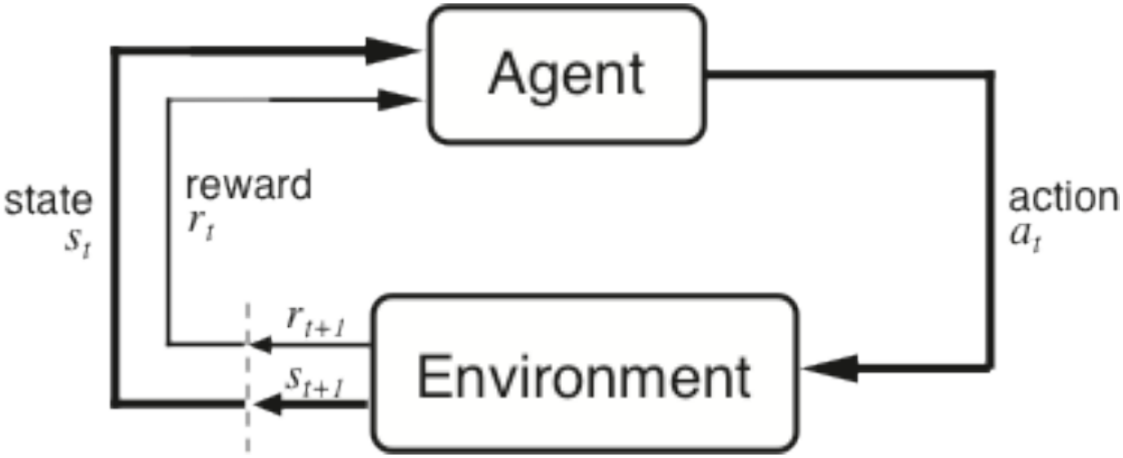
- Independent component analysis – separate a combined signal into its original sources



Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Examples :
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand

The Agent-Environment Interface



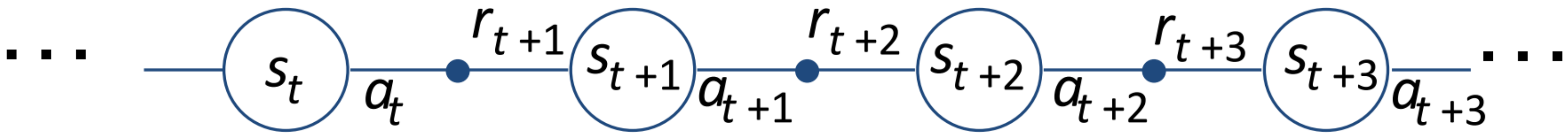
Agent and environment interact at discrete time steps : $t = 0, 1, 2, K$

Agent observes state at step t : $s_t \in S$

produces action at step t : $a_t \in A(s_t)$

gets resulting reward : $r_{t+1} \in \Re$

and resulting next state : s_{t+1}



Reinforcement Learning

AlphaGo vs. 李世石

Inverse Reinforcement Learning

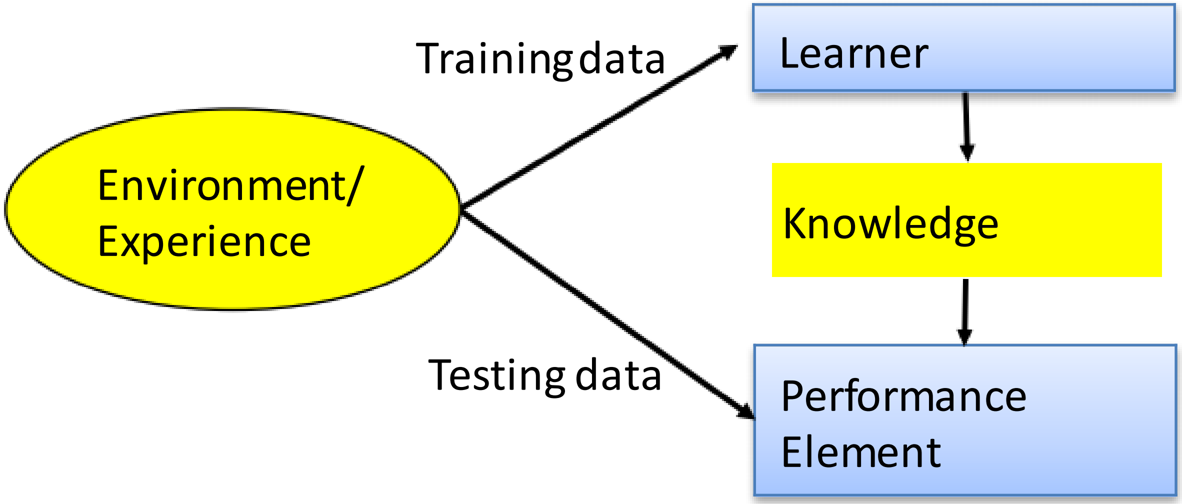
- Extract the reward function from an agent's observed behavior.

Stanford Autonomous Helicopter

Framing a Learning Problem

Designing a Learning System

- Choose the training experience
- Choose exactly what is to 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



Training vs. Test Distribution

- We generally assume that the training and test examples are independently drawn from the same overall distribution of data

  – We call this "i.i.d." (i.e. "independent and identically distributed")

- If examples are not independent (e.g interconnected with each other via linkage structure), requires *collective classification*

- If test distribution is different, requires *transfer learning*

ML in a Nutshell

- Tens of thousands of machine learning algorithms
  - Hundreds new every year

- Every ML algorithm has three components:
  - Representation
  - Optimization
  - Evaluation

Various Function Representations

- Numerical functions
  - Linear regression
  - Neural networks
  - Support vector machines
- Symbolic functions
  - Decision trees
  - Rules in propositional logic
  - Rules in first-order predicate logic
- Instance-based functions
  - Nearest-neighbor
  - Case-based
- Probabilistic Graphical Models
  - Naïve Bayes
  - Bayesian networks
  - Hidden- Markov Models (HMMs)
  - Probabilistic Context Free Grammars (PCFGs)
  - Markov networks

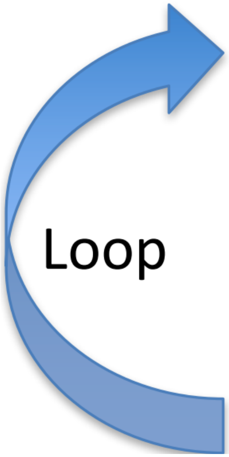
Various Search/Optimization Algorithms

- Gradient descent
  - Perceptron
  - Backpropagation
- Dynamic Programming
  - HMM Learning
  - PCFG Learning
- Divide and Conquer
  - Decision tree induction
  - Rule learning
- Evolutionary Computation
  - Genetic Algorithms (GAs)
  - Genetic Programming (GP)
  - Neuro-evolution

Evaluation

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

ML in Practice



Loop

- Understand domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learn models
- Interpret results
- Consolidate and deploy discovered knowledge

Lessons Learned about Learning

- Learning can be viewed as using direct or indirect experience to approximate a chosen target function.
- Function approximation can be viewed as a search through a space of hypotheses (representations of functions) for one that best fits a set of training data.
- Different learning methods assume different hypothesis spaces (representation languages) and/or employ different search techniques.

A Brief History of Machine Learning

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)
  - E-mail management
  - Personalized assistants that learn
  - Learning in robotics and vision
- **2010s**
  - Deep learning systems
  - Learning for big data
  - Bayesian methods
  - Multi-task & lifelong learning
  - Applications to vision, speech, social networks, learning to read, etc.
  - ???

What is Data Mining?

What is Data Mining?

Knowledge discovery from data

But to extract the knowledge

data needs to be

- Stored
- Managed
- And ANALYZED ← this class

Data Mining ≈ Big Data ≈

Predictive Analytics ≈ Data Science

What is Data Mining?

Given lots of data

Discover patterns and models that are:

- **Valid:** hold on new data with some certainty
- **Useful:** should be possible to act on the item
- **Unexpected:** non-obvious to the system
- **Understandable:** humans should be able to interpret the pattern

What is Data Mining?

The discovery of “models” for data, e.g.

1. Statistical Modeling
2. Machine Learning
3. Computational Approaches to Modeling
4. Summarization
5. Feature Extraction

Statistical Modeling

Statisticians were the first to use the term “data mining.”

Statisticians view data mining as the construction of a statistical model , that is, an underlying distribution from which the visible data is drawn.

Example: Does a data set come from a Gaussian distribution? If yes, what is the parameter for the model?

Machine Learning

Machine learning vs Data mining

Machine-learning practitioners use the data as a training set, to train an algorithm, e.g.

1. Bayes nets
2. Support-vector machines
3. Decision trees
4. Hidden Markov models

Data Mining uses algorithms to discover interesting patterns from data.

Computational Approaches to Modeling

More recently, computer scientists have looked at data mining as an algorithmic problem.

In this case, the model of the data is simply the answer to a complex query about it.

E.g. Computing mean and std. dev.

Summarization

To summarize the data in a simplified form, e.g.

1. PageRank: In this form of Web mining, the entire complex structure of the Web is summarized by a single number for each page.
2. Another important form of summary is clustering

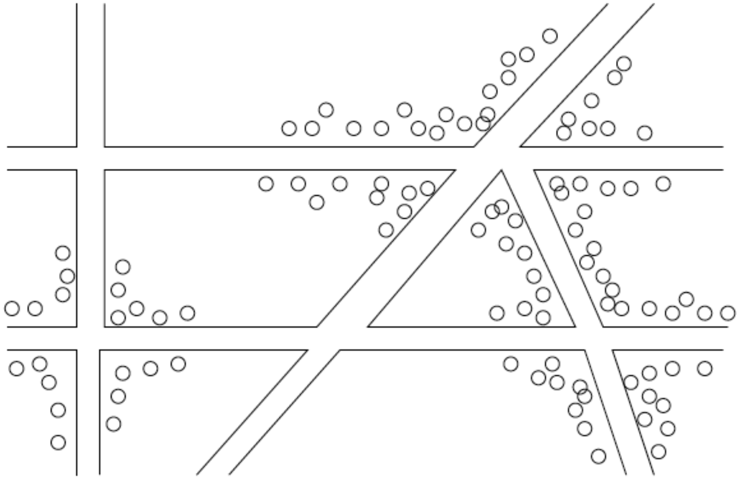


Figure 1.1: Plotting cholera cases on a map of London

Feature Extraction

The typical feature-based model looks for the most extreme examples of a phenomenon and represents the data by these examples, e.g.

1. Frequent Itemsets
2. Similar Items

Data Mining Tasks

Descriptive methods

- Find human-interpretable patterns that describe the data
  - **Example:** Clustering
  - **More examples?**

Predictive methods

- Use some variables to predict unknown or future values of other variables
  - **Example:** Recommender systems
  - **More examples?**

Meaningfulness of Analytic Answers

A risk with “Data mining” is that an analyst can “discover” patterns that are meaningless

Statisticians call it **Bonferroni’s principle (邦弗朗尼原理)**:

- Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Bush’ Total Information Awareness

DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
scottadams@aol.com
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
1/3/00 © 2000 United Feature Syndicate, Inc.
...SALES TO LEFT-HANDED SQUIRRELS ARE UP... AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

Meaningfulness of Analytic Answers

Example:

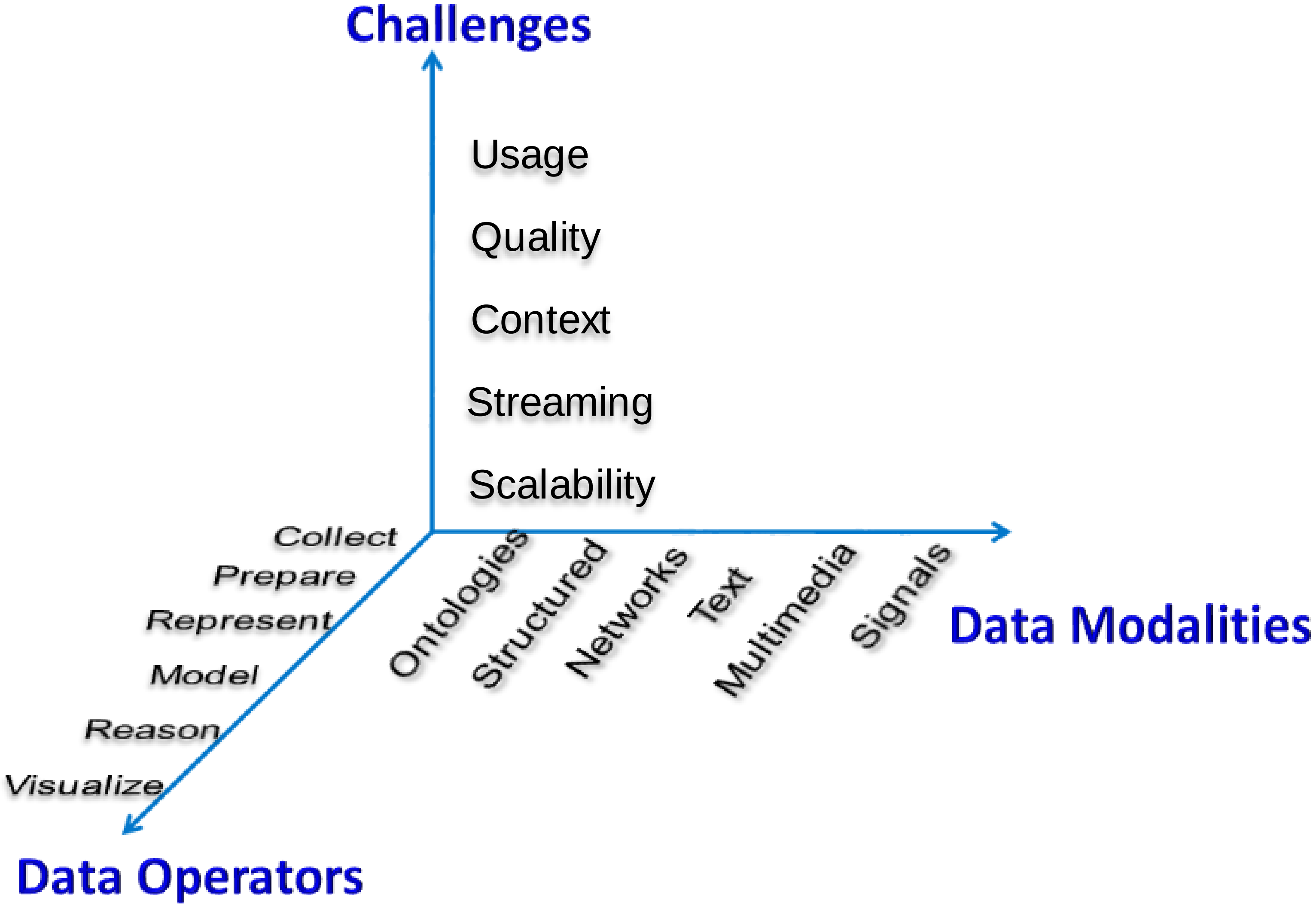
We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**

- 10^9 (1 billion) people being tracked
- 1,000 days
- Each person stays in a hotel 1% of time (1 day out of 100)
- Hotels hold 100 people (so 10^5 hotels)
- **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**

Expected number of "suspicious" pairs of people:

- 250,000
- … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

What Matters When Dealing with Data?



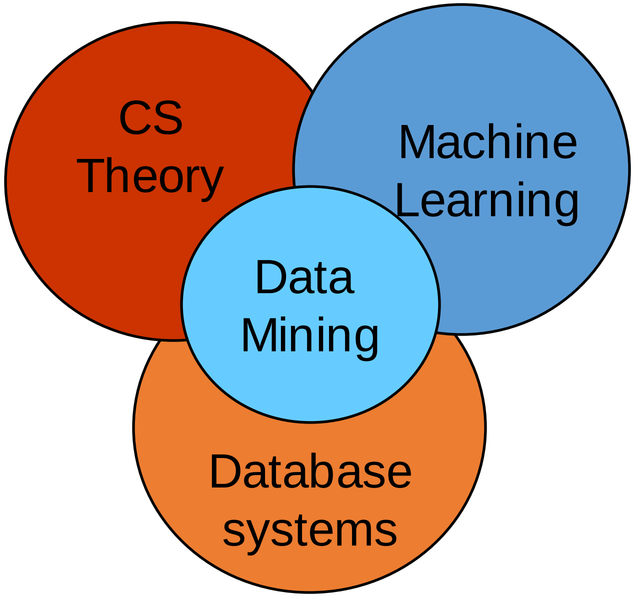
Data Mining: Cultures

Data mining overlaps with:

- **Databases:** Large-scale data, simple queries
- **Machine learning:** Small data, Complex models
- **CS Theory:** (Randomized) Algorithms

Different cultures:

- To a DB person, data mining is an extreme form of **analytic processing** – queries that
  examine large amounts of data
  - Result is the query answer
  - Most of Data miner is from DB person
- To a ML person, data-mining
  is the **inference of models**
  - Result is the parameters of the model
  - ML results can be published in DM Conf./Jou.

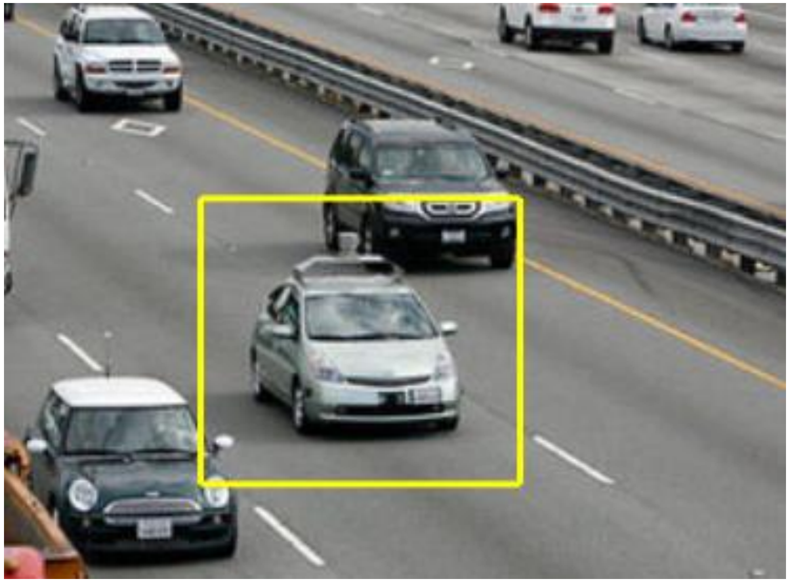


In this class we will do both of ML and DM!

国内机器学习与数据挖掘发展

自动驾驶例子

自动驾驶例子



2011美国内华达州批准自动驾驶汽车上路

2018年广州生物岛自动驾驶出租车试运营

国内机器学习&数据挖掘顶尖科学家

- 南京大学 周志华：机器学习、数据挖掘

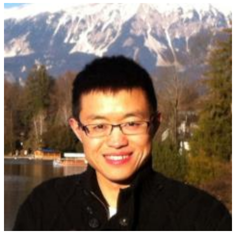
- AAAI/IEEE/IAPR Fellow

- ACML发起人

- AAAI、IJCAI等国际顶级会议PC Chair

- 《机器学习》教材

- 清华大学 唐杰：数据挖掘、机器学习



- IEEE/ACM Fellow

- ACM TKDD执行主编

- KDD、WWW等国际顶级会议Chair/PC Chair

- Many more...

About the Course

- Wechat Group
- TA: 何振宇
- 总分100分，期末考试成绩占40%、平时成绩占60%，最后总成绩 = 期末考试（40%）+ 平时成绩（60%）。
- 平时成绩满分60分，由4次大作业（最后一次为期末课程报告）和平时考勤成绩构成。考勤满分9分，4次大作业满分51分，4次大作业的分值分别为12分、12分、12分、15分。

群二维码名片

19软工保密-ML&DM课程群

该二维码7天内(2月23日前)有效，重新进入将更新

课程教材

- 《机器学习》(作者：周志华, 出版社：清华大学出版社, ISBN:9787302423287, 2016)
- Mining of Massive Datasets， 2th Edition, 2014 (作者：Jure Leskovec, Anand Rajaraman, Jeff Ullman， 出版社：Cambridge University Press)
- Pattern Recognition and Machine Learning (作者：Christopherf M. Bishop, 出版社：Springer Science+Business Media, LLC)