

# **Lecture 1**

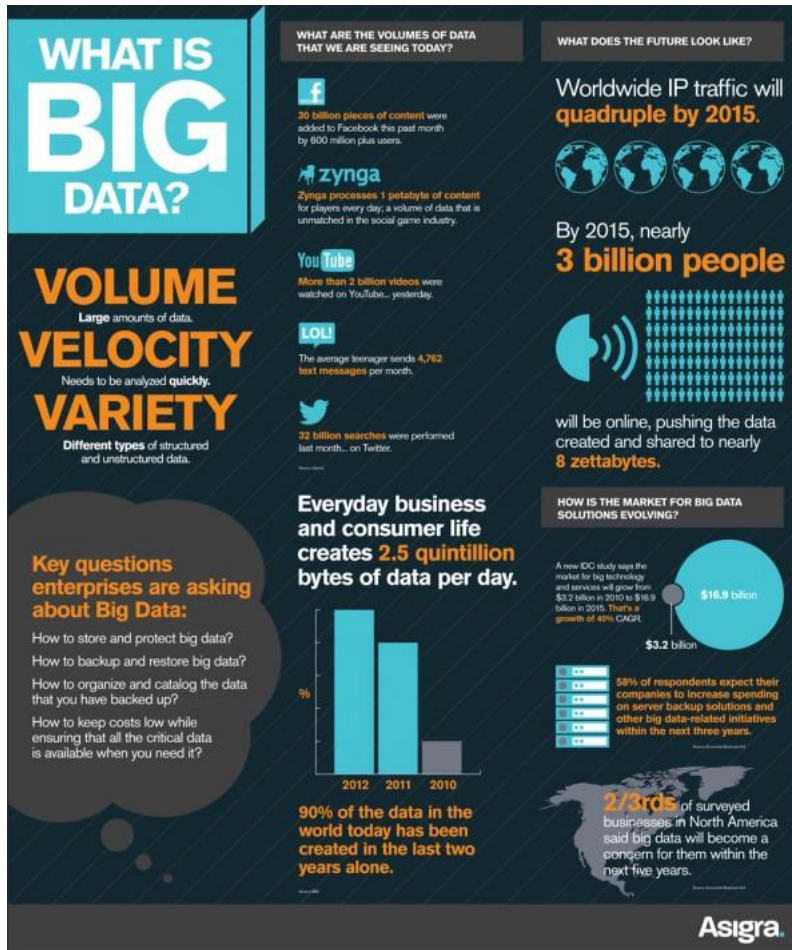
# **An Introduction to Machine Learning**

Rui Xia

School of Computer Science & Engineering  
Nanjing University of Science & Technology

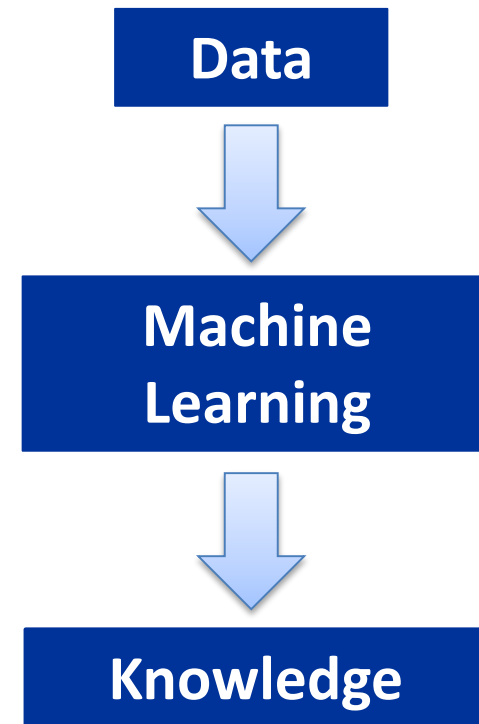
<http://www.nustm.cn/~rxia>

# Big-Data World



- 500 million tweets are sent per day. That's around 6,000 tweets per second.
- Facebook has more than 2 billion active users generating social interaction data.
- More than 5 billion people are calling, texting, tweeting and browsing websites on mobile phones.
- Walmart handles more than 1 million customer transactions every hour.
- VISA processes more than 172,800,000 card transactions each day.
- United Parcel Service receives on average 39.5 million tracking requests from customers per day.
- RFID (radio frequency ID) systems generate up to 1,000 times the data of conventional bar code systems.

# What is Machine Learning?



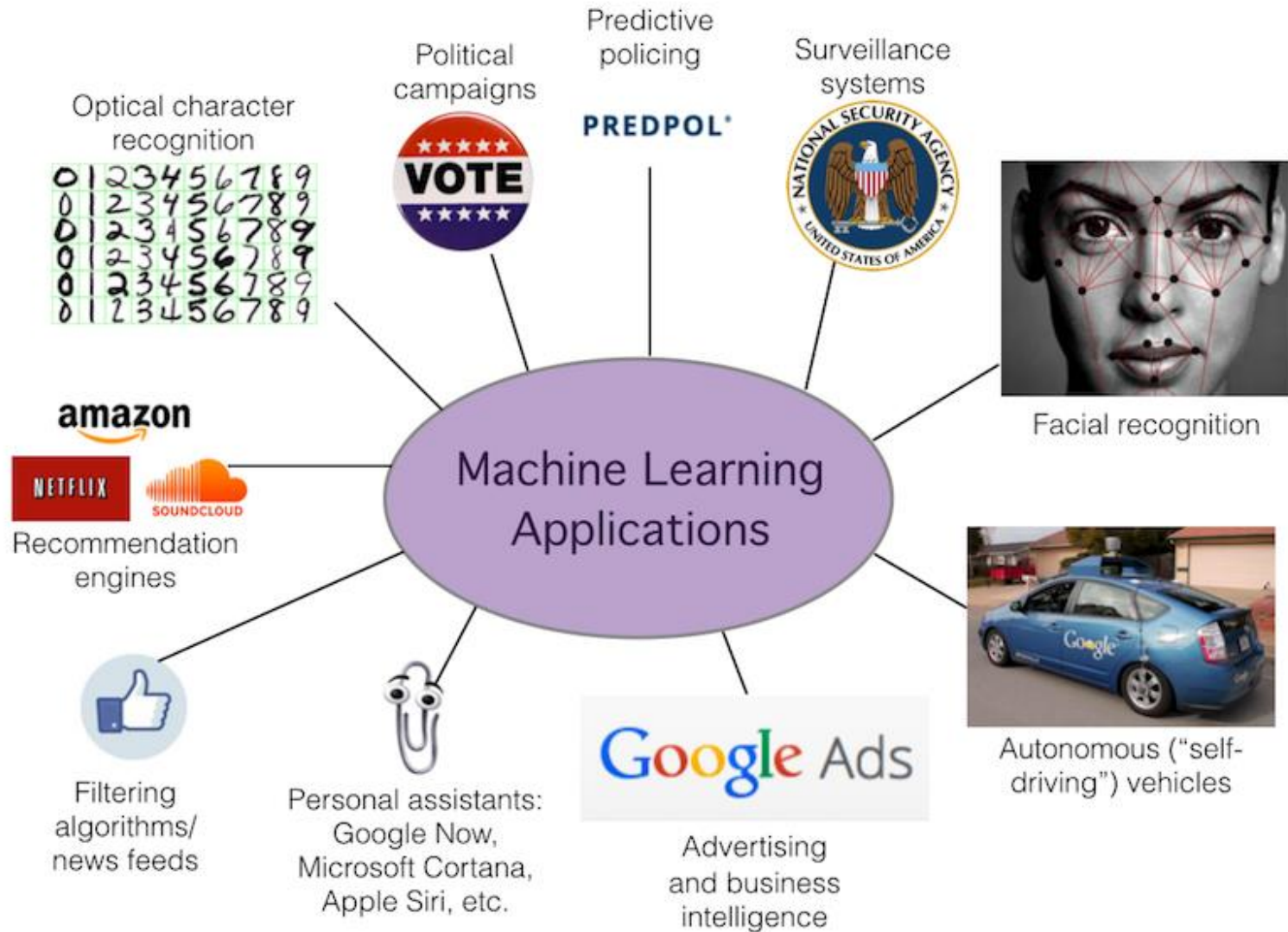
# Definition of Machine Learning

- Arthur Samuel (1959) defined machine learning as a  
*“Field of study that gives computers the ability to learn without being explicitly programmed”*
- Tom M. Mitchell (1997) provided a widely quoted, more formal definition  
*“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  $E$ ”*



**Can machines (i.e., computers) learn what we (human beings) can learn (from data)?**

# Machine Learning Applications





# Text Categorization



U.S. [dropdown]

**Top Stories**

**★ Obama's budget proposal draws rapid fire from legislators**

USA Today - [Richard Wolf](#), [Steve Hebert](#) - 2 hours ago

WASHINGTON - President Obama's proposed \$3.8 trillion budget ran into immediate trouble in Congress on Monday among lawmakers who said it tries to do too much while cutting the deficit too little.

[Video: Obama Budget Would Create Highest-ever Deficit](#) YouTube The Associated Press

[Wealthy Face Tax Increase](#) Wall Street Journal

[Milwaukee Journal Sentinel](#) - [Bradenton Herald](#) - [ABC News](#) - [Daily Caller](#)

[all 4,573 news articles >](#) [Email this story](#)

**☆ Md. stands to gain despite Obama budget cuts**

Baltimore Sun - [Paul West](#) - 1 hour ago

WASHINGTON - President Barack Obama wants to end the nation's troubled program to return a but NASA officials indicated Monday

**Left sidebar:**

- Top Stories
- Starred ★
- World
- U.S.
- Business
- Sci/Tech
- Entertainment
- Sports
- Health
- Spotlight
- Most Popular

**Bottom left:**

- All news
- Headlines



Google in:spam [search]

Gmail - [compose] [refresh] [more]

**COMPOSE**

**Inbox (7)**

- Starred
- Important
- Sent Mail
- Drafts (15)
- All Mail
- Spam (46)**
- Trash

**Circles**

**Spam messages:**

- me
- no1.gr
- PayPal
- EdFed
- LoopGalaxy
- LinkShare
- WESTERN UNION MONEY TR
- Miss Beauty Musa
- American Musical Supply

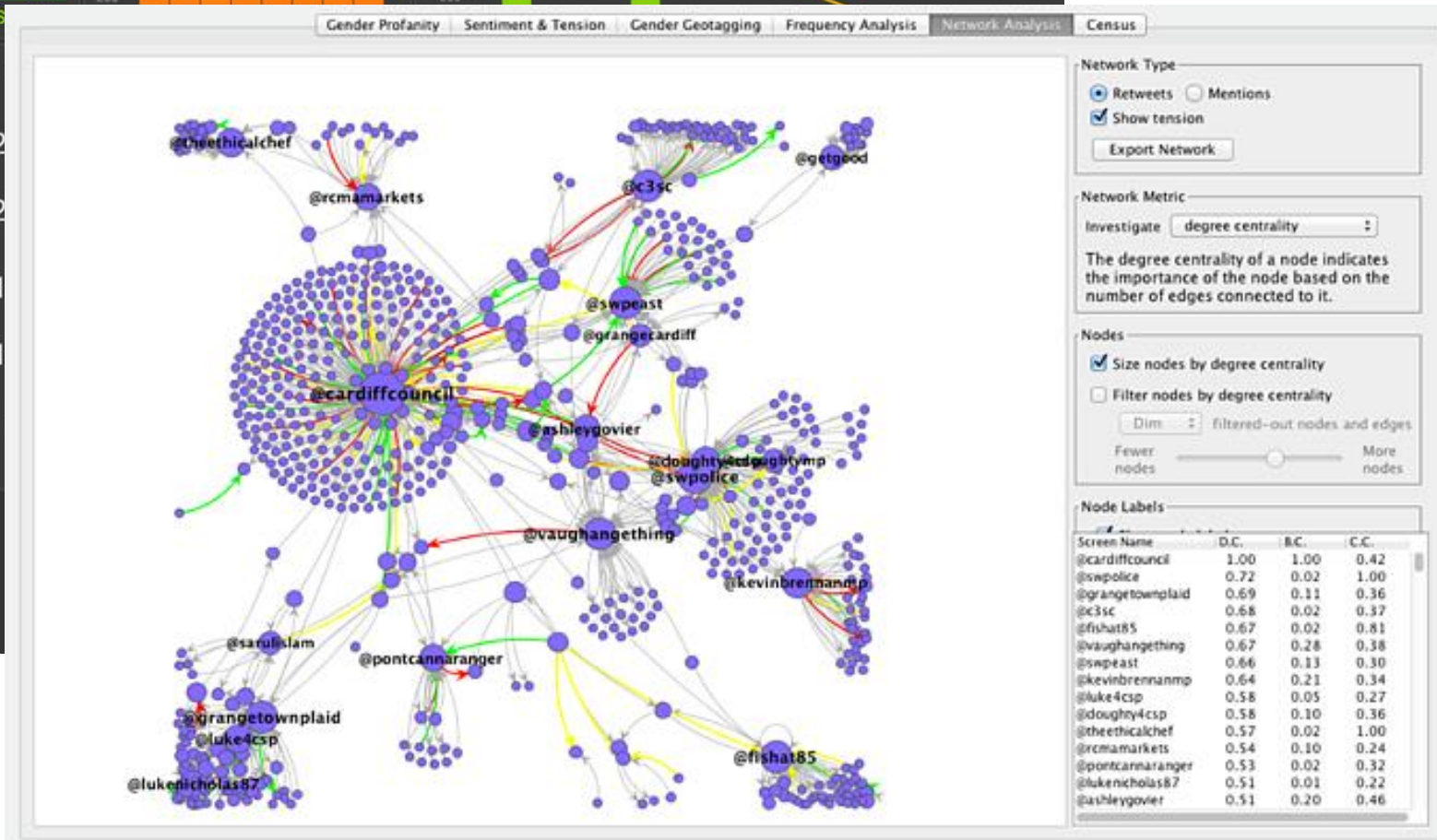
**Spam message content:**

- Delete all spam messages now (messages that have been in Spa
- New submission from Quick Poll: Facebook Pre-Fill - I would u
- Προσπαθήστε το κινητό σας.... - Ean den mporeite na delte to ne
- Your PayPal account has been limited! - Warning Notification De
- "What NOT TO DO During Your Interview" - To ensure prompt d
- March Madness Sale! 50% Off All Sample Packs - Share Embe
- Register Now: Social & Mobile Technologies Webinar - Social I
- WESTERN UNION - Attn, We are grateful to contact you and anno
- Dearest - Dearest I know this mail will come to you as a surprise s
- Live Loud on Stage with Pro Gear up to 66% off - Speaker Syst

# Sentiment Analysis and Opinion Mining

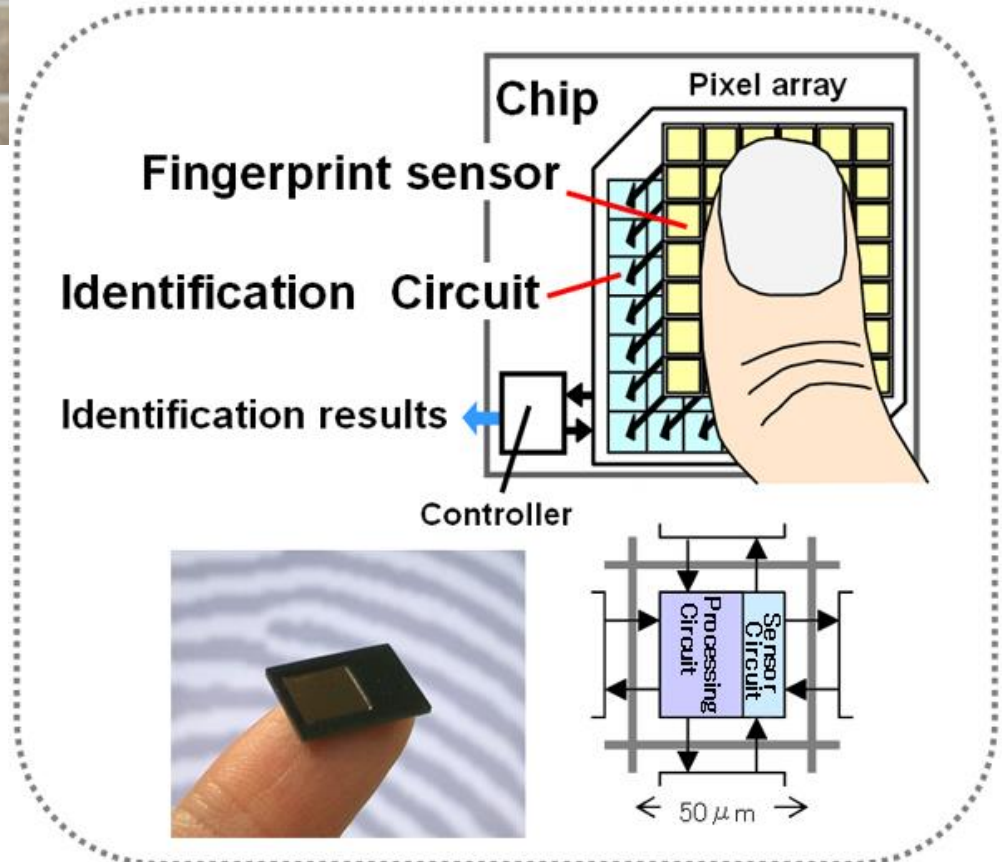
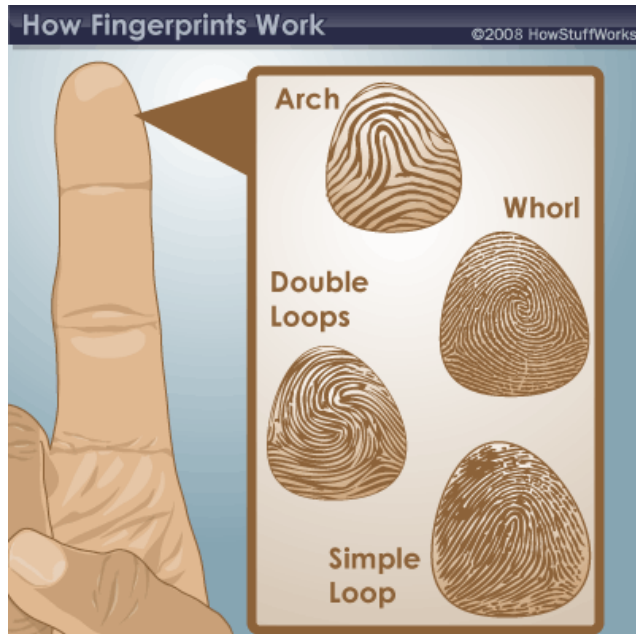


# Social Media Analysis





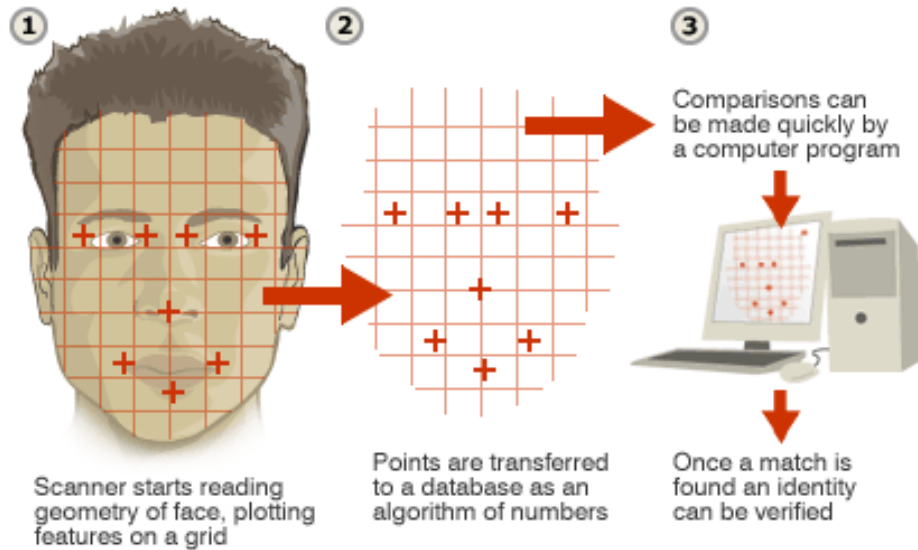
# Fingerprint Identification



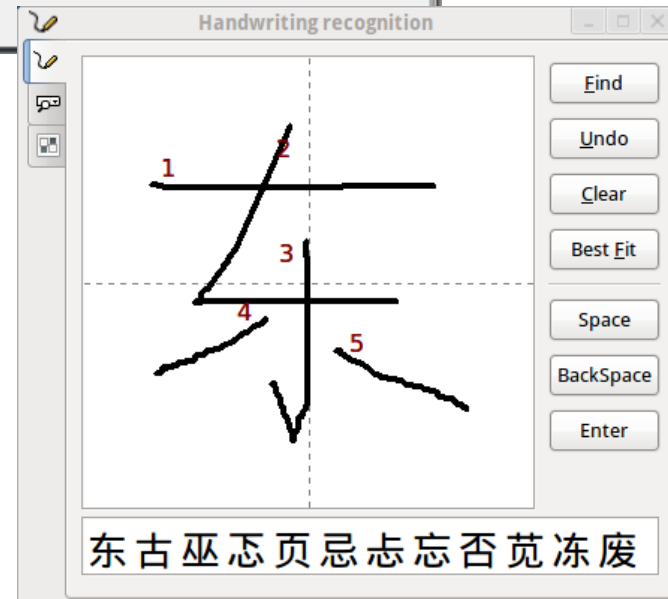
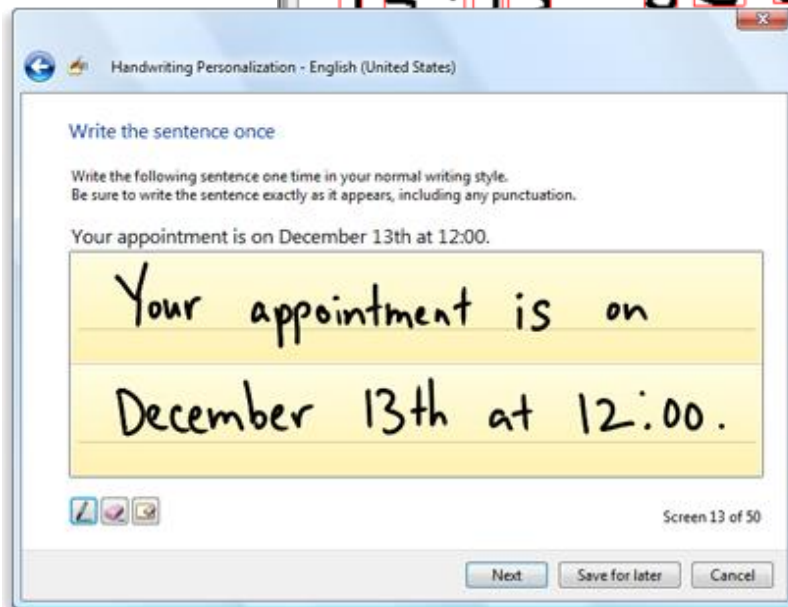
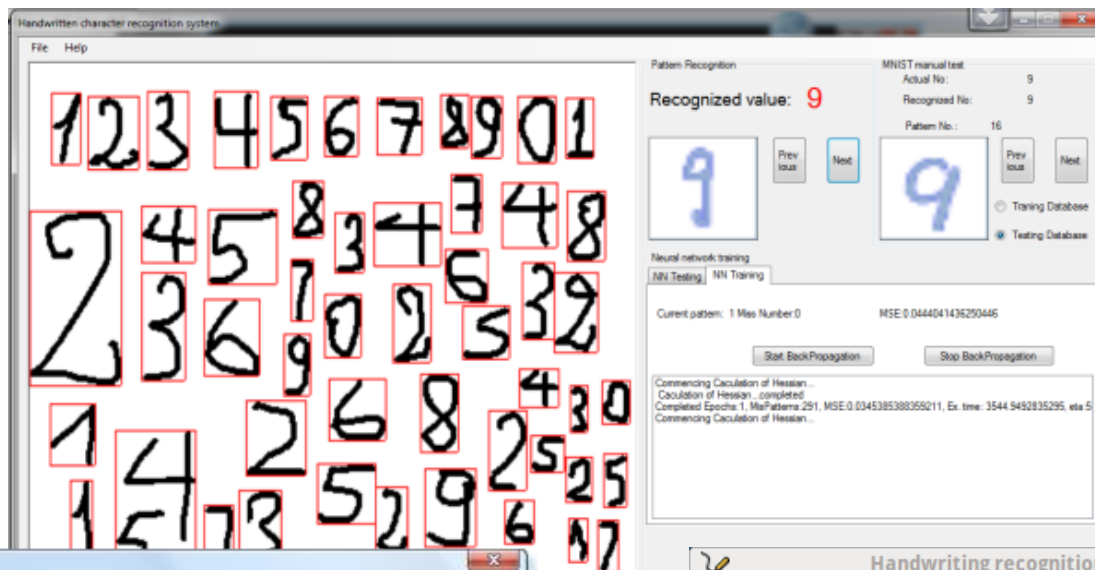
# Face Recognition



## HOW 2D FACIAL SCANNERS RECORD IDENTITIES

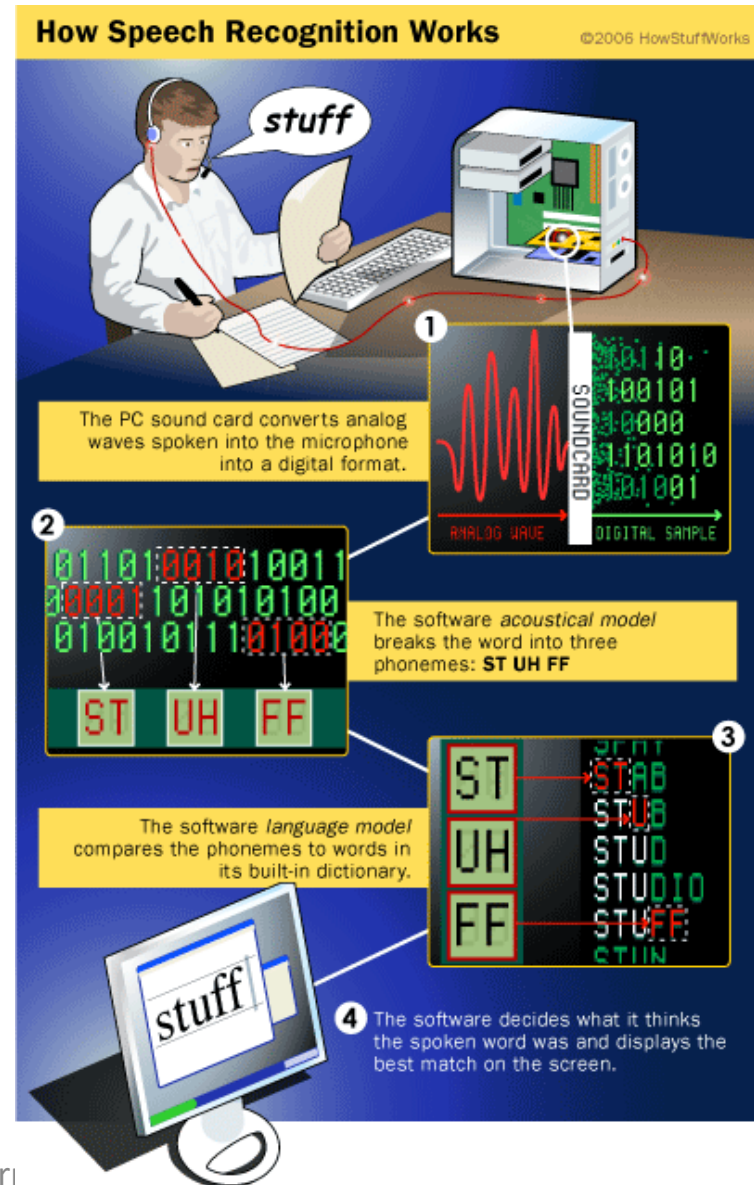
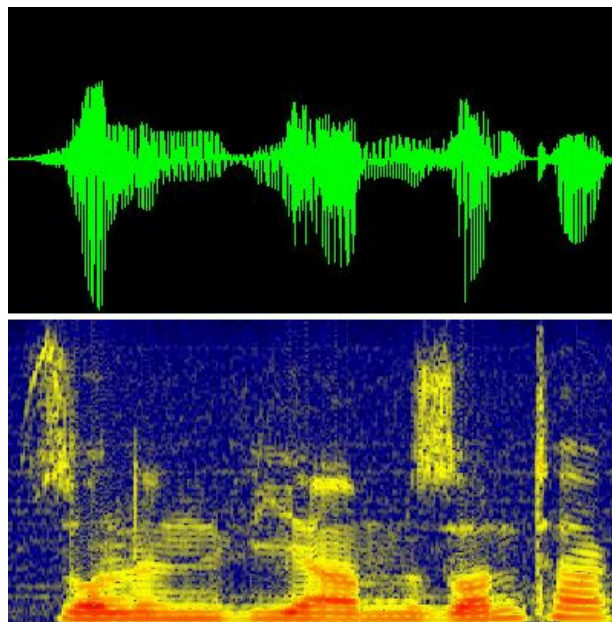
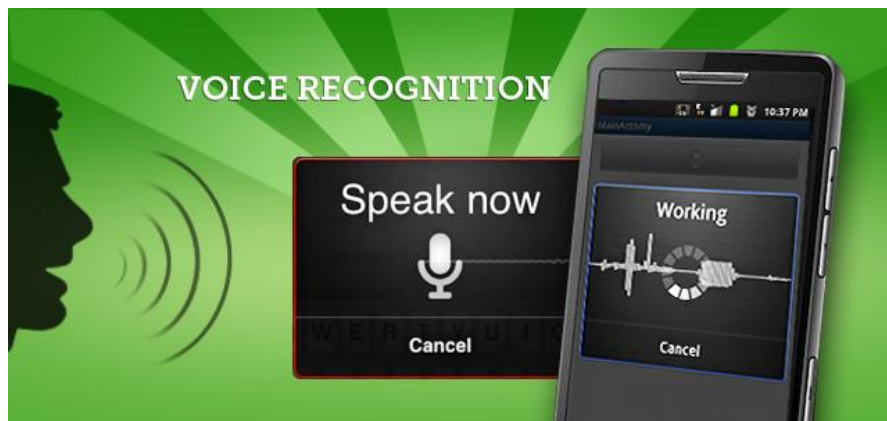


# Handwritten Character Recognition





# Speech Recognition





# Image Identification



<https://www.zhihu.com/question/51020471>

# Bioinformatics

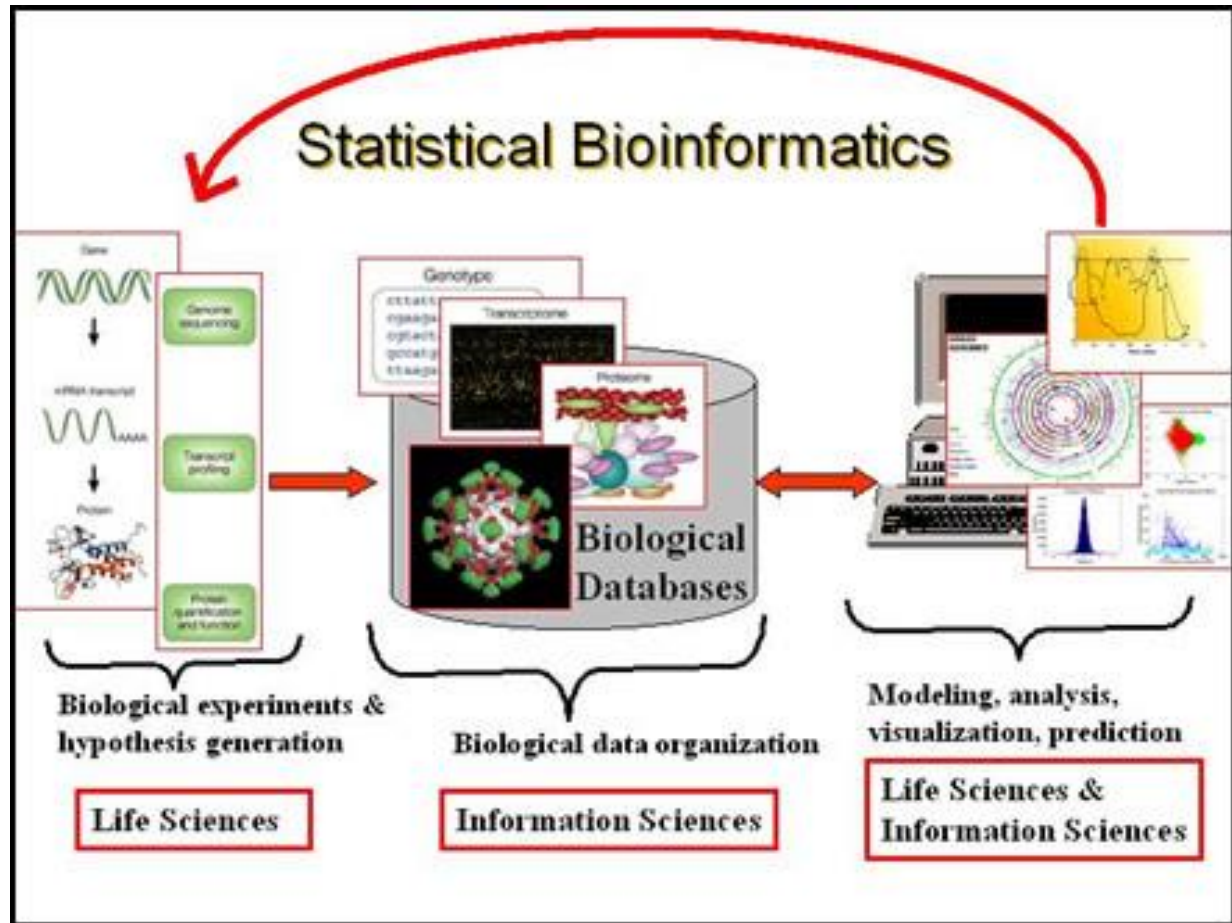


**HUMAN GENOME, IN NUMBERS**

6 billion DNA LETTERS  
46 CHROMOSOMES  
\$9,500 COST TO SEQUENCE

22,000 GENES  
500GB SIZE ON DISK

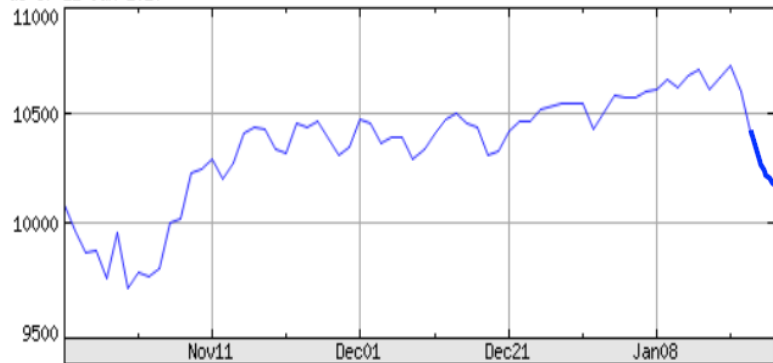
Sources: NIH, Illumina



# Stock Market Prediction



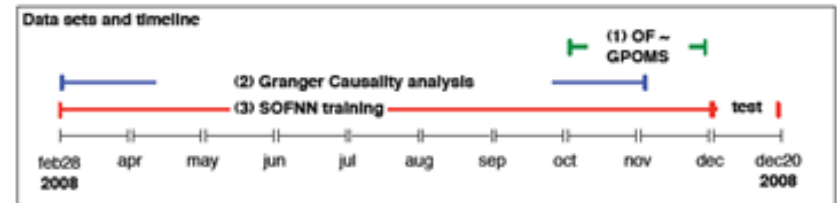
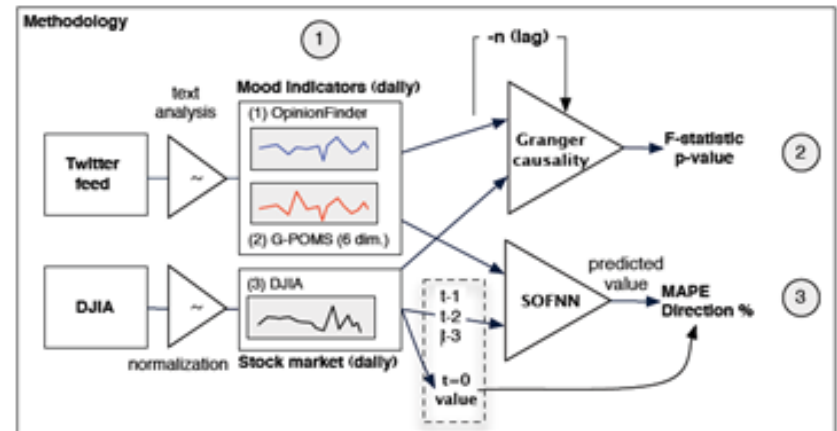
DJ INDU AVERAGE (DOW JONES & CO  
as of 22-Jan-2010



Copyright 2010 Yahoo! Inc.

<http://finance.yahoo.com/>

X = Feb01





# Human-machine Competition



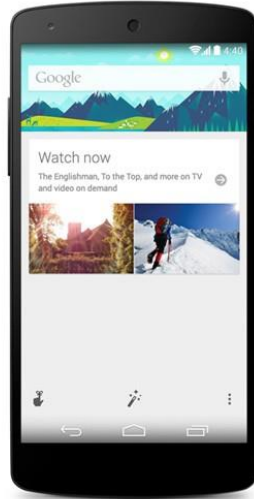


# Dialogue System

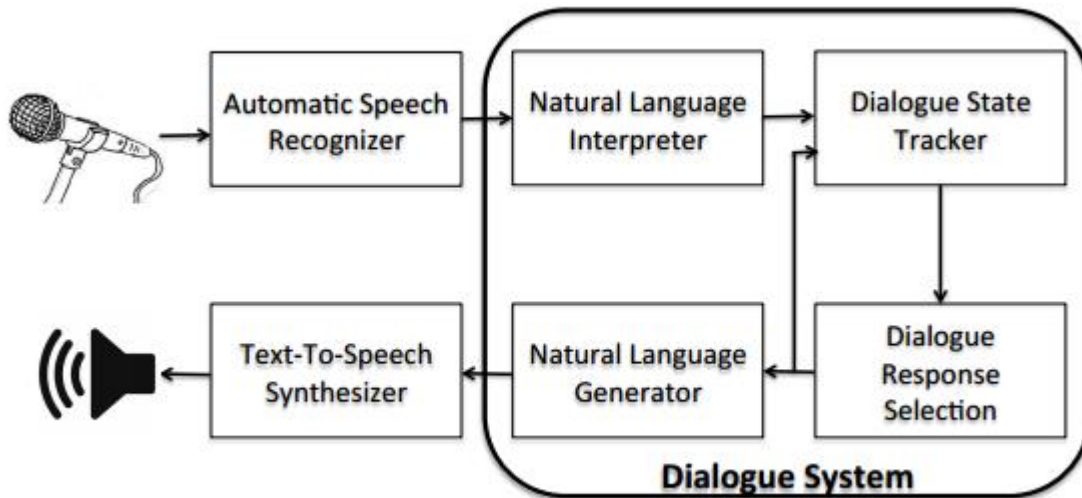
Apple Siri



Google Now



Windows Cortana



# Machine Translation



Translate

Turn off instant translation



Chinese English Spanish Detect language



English Chinese (Simplified) Spanish

Translate

## 【谷歌NMT，见证奇迹的时刻】

微信最近疯传人工智能新进展：谷歌翻译实现重大突破！值得关注和庆贺。mt 几乎无限量的自然带标数据在新技术下，似乎开始发力。报道说：

十年前，我们发布了 Google Translate（谷歌翻译），这项服务背后的核心算法是基于短语的机器翻译（PBMT:Phrase-Based Machine Translation）。

自那时起，机器智能的快速发展已经给我们的语音识别和图像识别能力带来了巨大的提升，但改进机器翻译仍然是一个高难度的目标。

今天，我们宣布发布谷歌神经机器翻译（GNMT Neural Machine Translation）系统，该系统使用先进的训练技术，能够实现到目前为止机器翻译的大提升。我们的全部研究结果详情请参阅我们《Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation》。

[Google NMT, witness the miracle of the moment]

Recent advances in microblogging crazy biography of artificial intelligence: Google translation to achieve a major breakthrough! Worthy of attention and celebration. Mt almost unlimited number of natural standard data in the new technology, it seems to start force. The report says:

Ten years ago, we released Google Translate, the core algorithm behind this service is PBMT: Phrase-Based Machine Translation.

Since then, the rapid development of machine intelligence has given us a great boost in speech recognition and image recognition, but improving machine translation is still a difficult task.

Today, we announced the release of the Google Neural Machine Translation (GNMT) system, which utilizes state-of-the-art training techniques to maximize the quality of machine translation so far. For a full review of our findings, please see our paper "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."



# Autonomous Vehicles



# Many, Many More

- Internet Search and Information Retrieval
- Natural Language Understanding
- Machine Translation
- Locating/tracking/identifying objects in images & video
- Financial prediction and Business Intelligence
- Medical diagnosis, media image analysis
- Recommendation Systems
- ...



# Key words/concepts in Machine Learning

bayesian clustering conditional-distribution cost-function cross-entropy  
decision discriminative distribution em gaussian  
generative graphical-model inference joint-  
distribution least-square likelihood logistic-regression  
map ml model model-selection multinomial naive-bayes  
over-fitting predictive-function regression semi-supervised sequential-model  
supervised unsupervised

**Tag-crowd of Bi-shop's PRML Book**

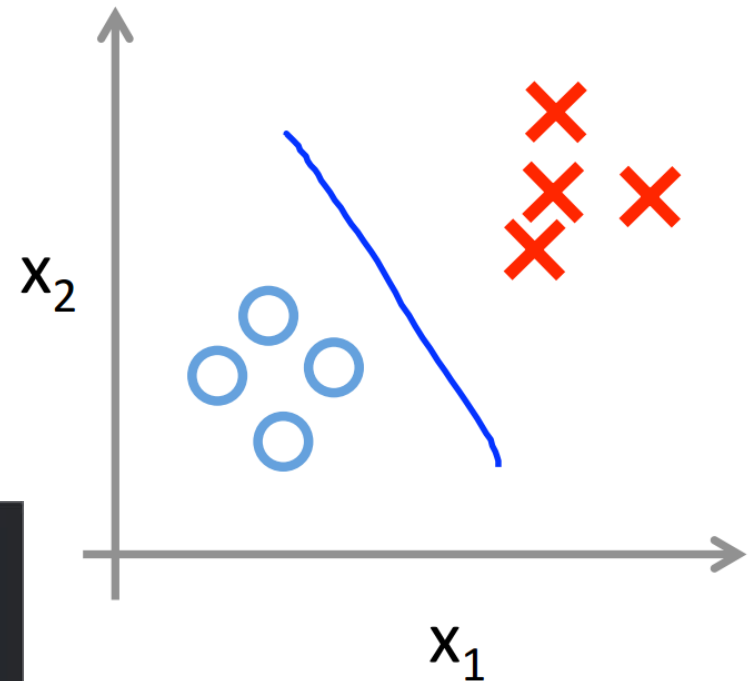
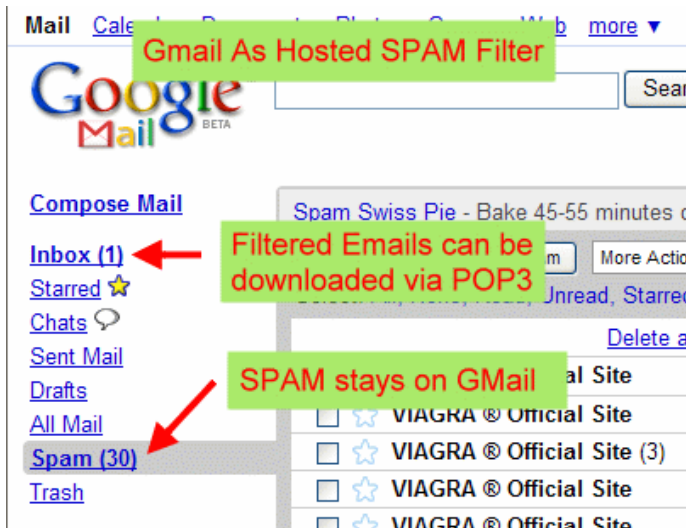
**(Pattern Recognition and Machine Learning)**

<http://research.microsoft.com/en-us/um/people/cmbishop/PRML>

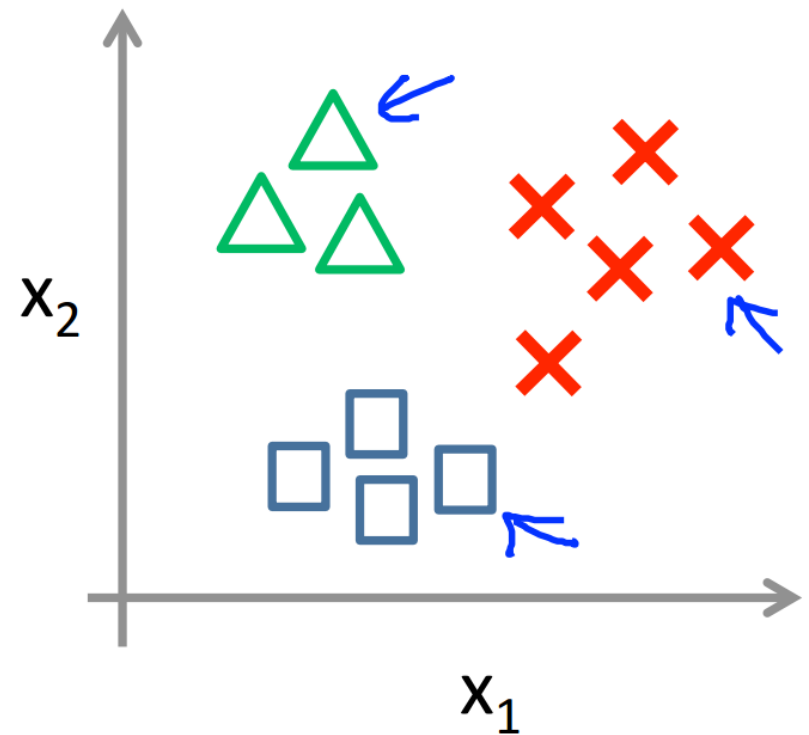
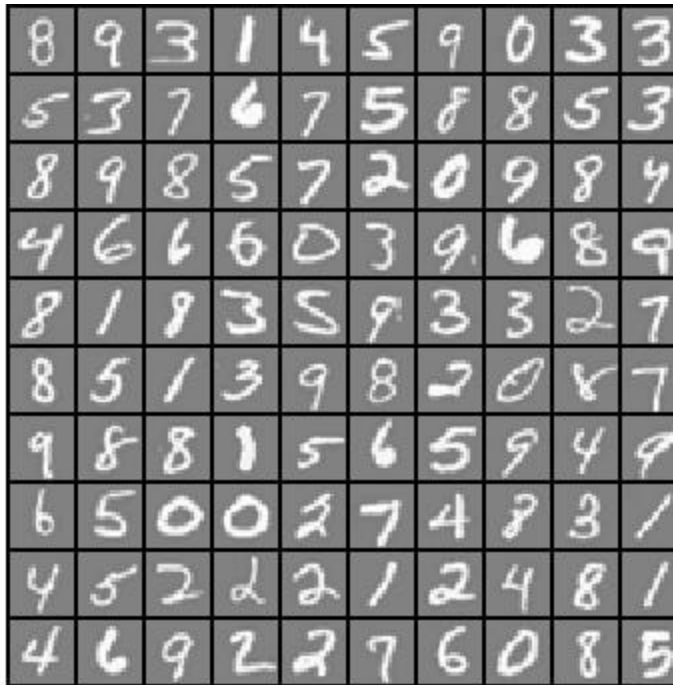
# Machine Learning Categorization

- Supervised Learning: given examples of inputs and corresponding outputs, predict outputs on new inputs
  - Classification, Regression, etc.
- Unsupervised Learning: given only inputs, automatically discover knowledge (labels, features, structure, etc.)
  - Clustering, Density Estimation, etc.
- Semi-supervised Learning
- Ensemble Learning
- Active Learning
- Transfer Learning
- Reinforcement Learning
- Deep learning
- ...

# Binary Classification



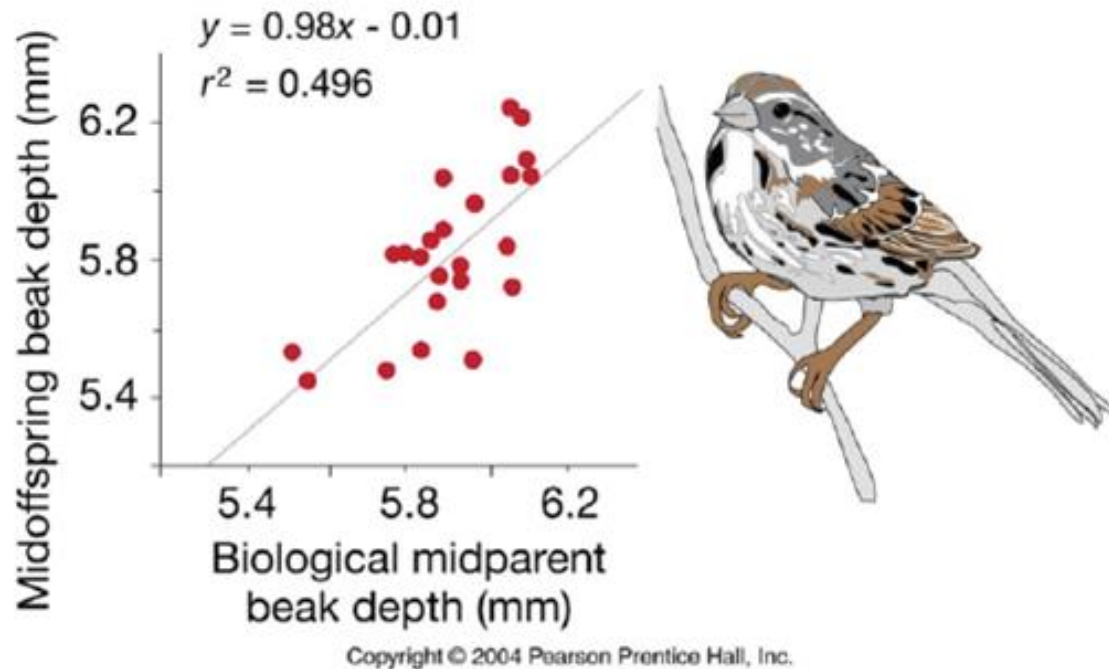
# Multi-class Classification



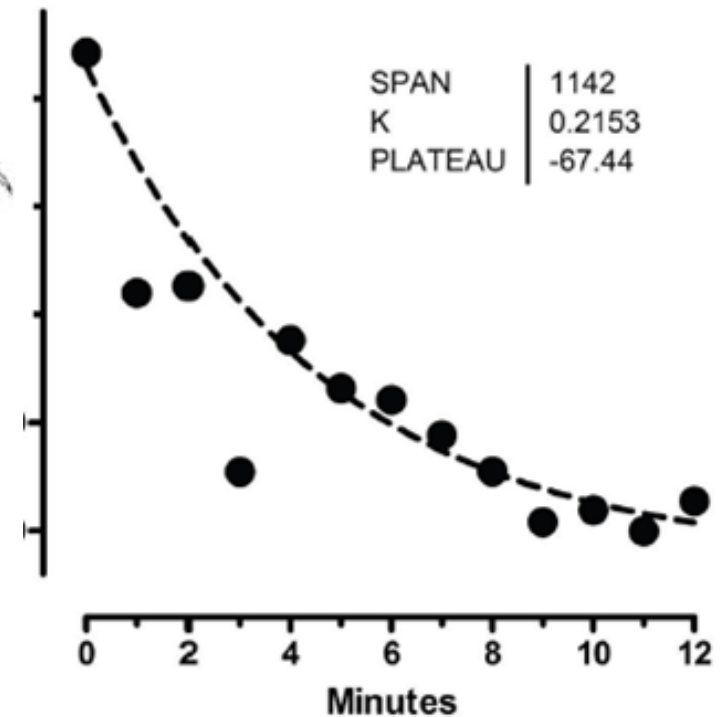


# Regression

## Linear Regression



## Nonlinear Regression

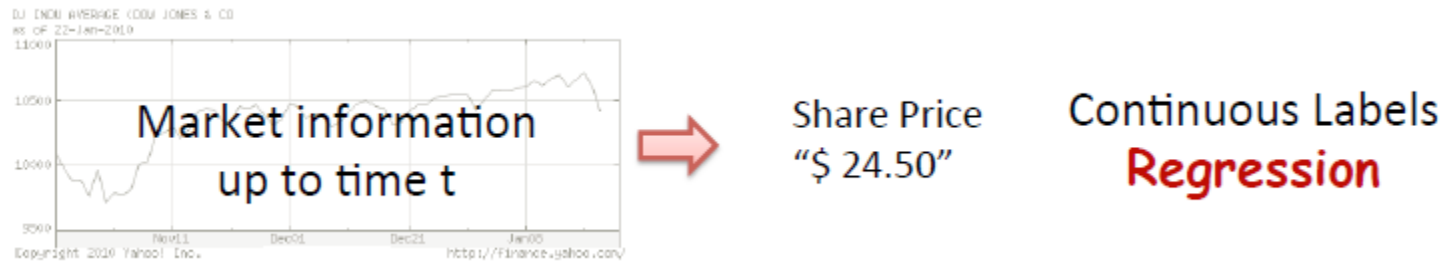


# Classification vs. Regression

- Classification



- Regression



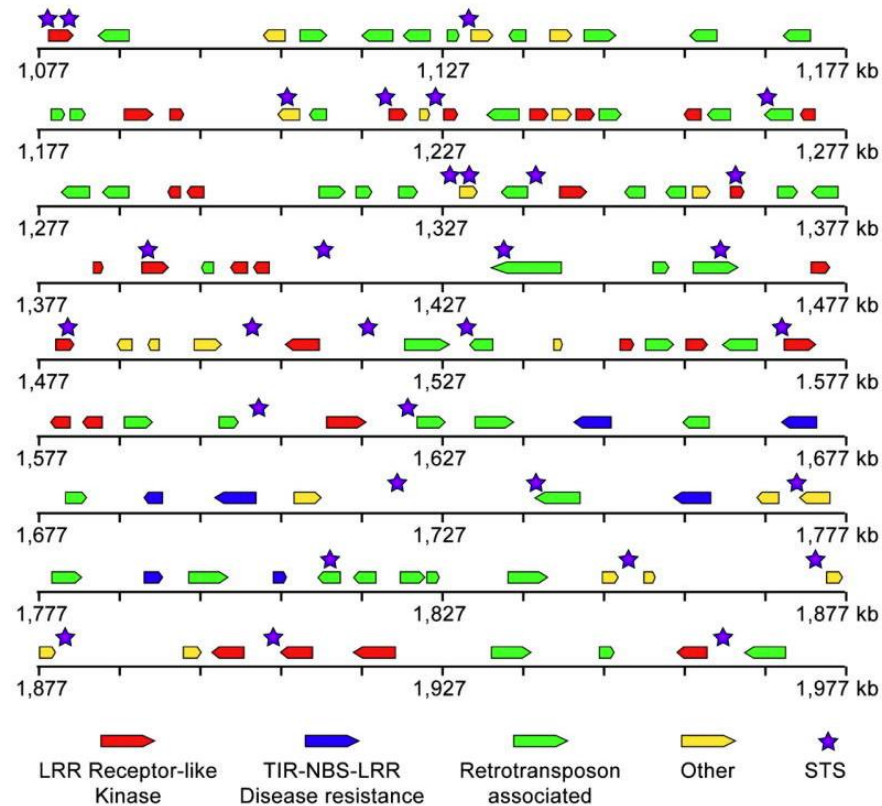
# Clustering

Group similar patterns (words, images, ...)

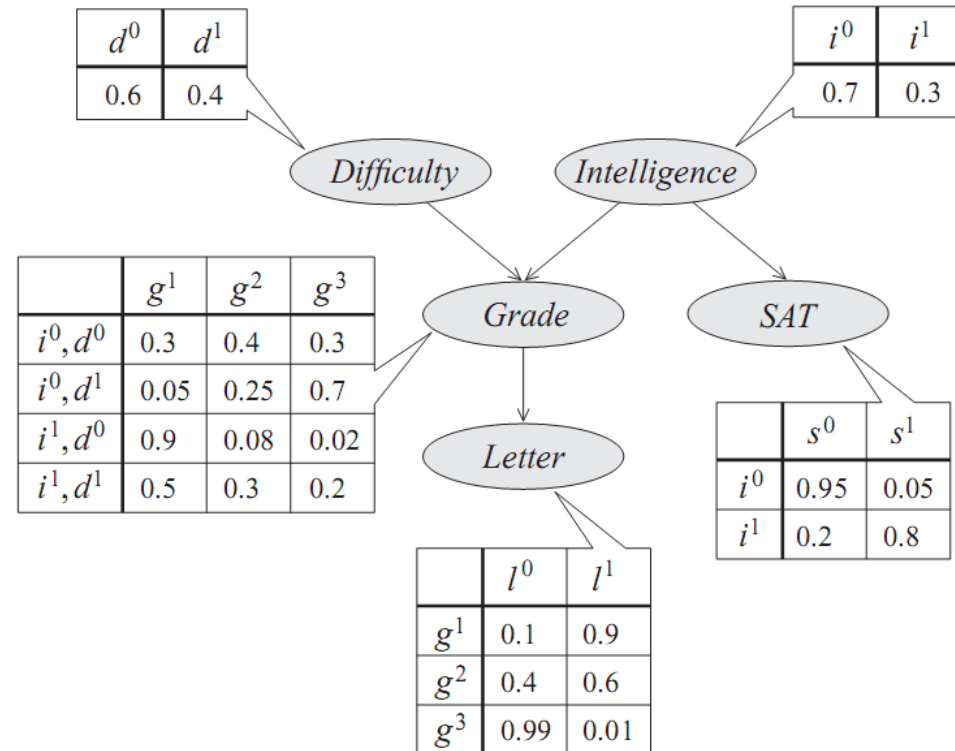


# Sequential, Graphical Learning

- Sequential Model



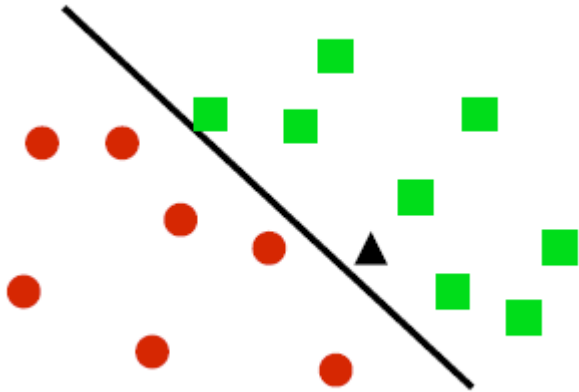
- Graphical Model





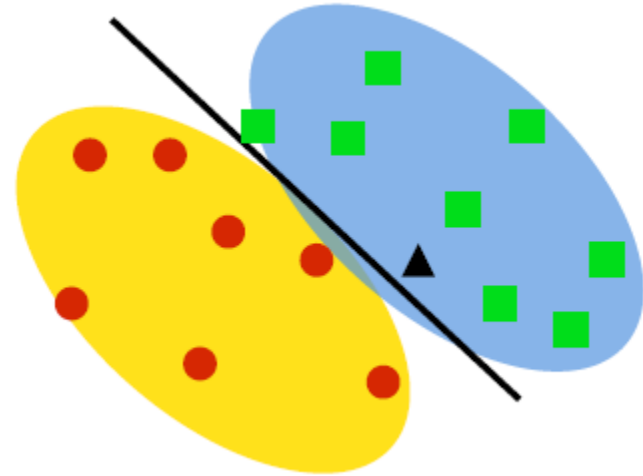
# Generative vs. Discriminative

- Discriminative Model



Model the posterior probability of label given observation  $p(y|x)$

- Generative Model



Model the joint probability of label and observation  $p(x, y)$ , and then use the Bayes rule  $p(y|x) = p(x, y)/p(x)$  for prediction.

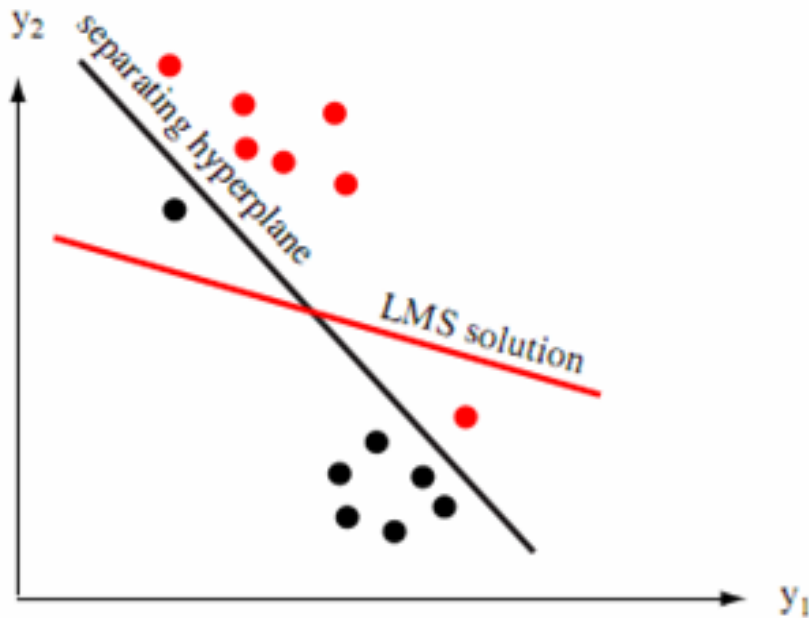
# Machine Learning Phases

- Training phase (using **training data**)  
You present your data from your "gold standard" and train your model, by pairing the input with expected output
- Test phase (using **test data**)  
In order to estimate how good your model has been trained, and to estimate model properties (such as mean error for regression, accuracy for classification)
- Application phase (for **future data**)  
Now you apply your freshly-developed model to the real-world data and get the results

# Hypothesis - Learning - Decision

- Hypothesis
  - Math models with (unknown) **parameters** (or structures)
- Learning (**to estimate the parameters**)
  - Maximum Likelihood Estimation (MLE), MAP, Bayesian Estimation
  - Cost Function Optimization
- Decision
  - Bayes decision rule
  - Direct prediction function

# Learning Criteria



- Perceptron Criterion
- Least Mean Square (LMS)
- Minimum Cross Entropy (CE)
- Maximum Margin Criterion
- Maximum Likelihood
- ...

Which linear hyper-plane  
is better?

Which learning criterion to  
choose?





# Optimization Methods

- Analytic Solution
- Gradient Descent
- Stochastic Gradient Descent
- Newton Method
- Quasi-Newton Method (BFGS)
- Limited Memory BFGS (L-BFGS)
- Conjugate Gradient
- GIS
- IIS
- ...

# Basic Mathematic Knowledge

- Calculus (Differentiation, Integration)
- Linear Algebra
- Probability Theory
- Optimization Methods



**Any Questions?**