ELEG 6913: Machine Learning for Big Data

Fall 2016

Lecture 7: Machine Learning: From Theories to Applications

Dr. Xishuang Dong

Outline

- **Review of Machine Learning**
- **Machine Learning Based Applications**
- **Natural Language Processing via Machine Learning**

(Acknowledgment: some parts of the slides are from Internet and various other sources. The copyright of those parts belongs to their original owners.)

Outline

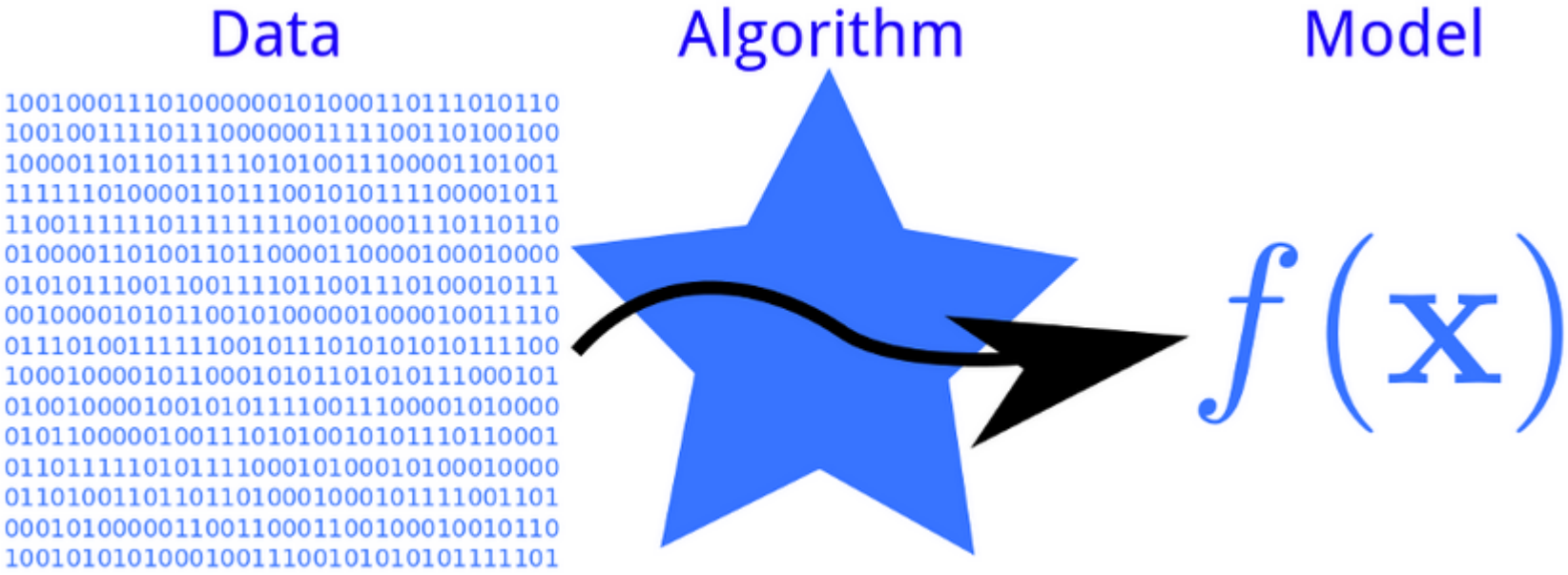
- **Review of Machine Learning**
- **Machine Learning Based Applications**
- **Natural Language Processing via Machine Learning**

Machine Learning

- **It is a branch of artificial intelligence (AI).**
- **It is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.**
- **Behaviors such as recognizing faces, translating, and searching.**

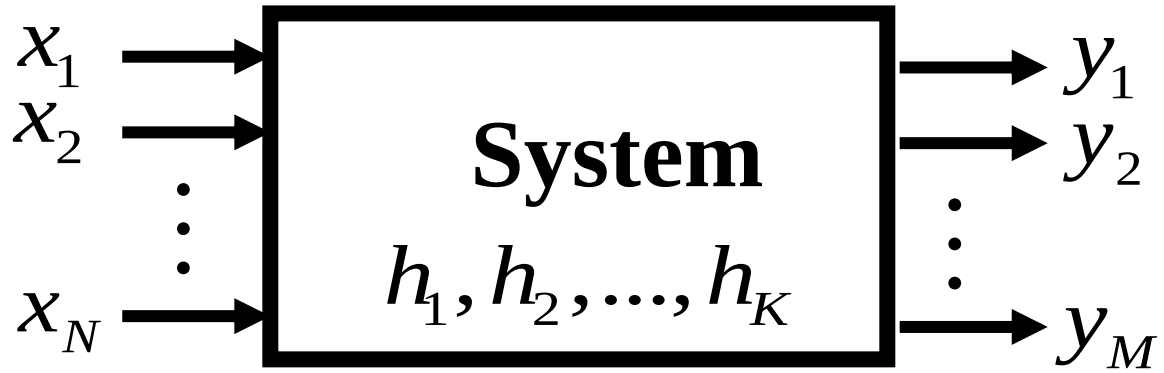
Machine Learning

- **Machine learning systems automatically learn programs from data to generate a model.**



http://phdp.github.io/posts/2013-07-05-dtl.html

A Generic ML System



Input Variables: $\mathbf{x} = (x_1, x_2, ..., x_N)$

Hidden Variables: $\mathbf{h} = (h_1, h_2, ..., h_K)$

Output Variables: $\mathbf{y} = (y_1, y_2, ..., y_M)$

Other Definitions of Machine Learning

- **Machine Learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system**
- **These algorithms originate form many fields:**
  - ✓ Statistics, Mathematics, Physics, Neuroscience, etc.

ML Terminology

- **Samples**: Items or instances used for learning (or training) or evaluation (or testing).
- **Features**: Set of attributes represented as a vector associated with an sample.
- **Labels**: Values or categories assigned to examples. For classification the labels are categories; For regression the labels are real numbers.
- **Output**: Prediction labels by using a model of the machine learning algorithm.
- **Model**: Information that the machine learning algorithm stores after training. The model is used when predicting labels of new, unseen examples.

ML Terminology

- **Training sample**: Examples used to train a machine learning algorithm.
- **Testing sample**: Examples used to evaluate the performance of a learning algorithm. The test sample is separated from the training samples and not available in the learning stage.

The Sub-Fields of ML

- **Supervised Learning**
- **Unsupervised Learning**
- **Reinforcement Learning**

The Sub-Fields of ML

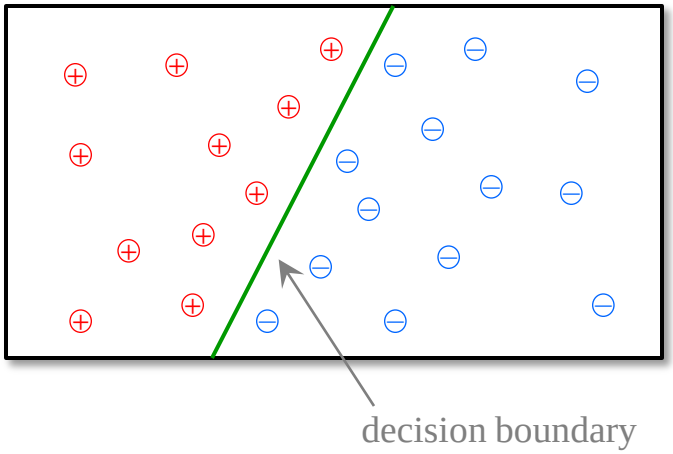
- **Supervised Learning**
- **Unsupervised Learning**
- **Reinforcement Learning**

Supervised Learning

- **We have training samples with correct answers (labels)**
- **Use the training samples and labels to learn the algorithm (model)**
- **Then apply it to data without a correct answer (labels)**

Supervised Learning Algorithms

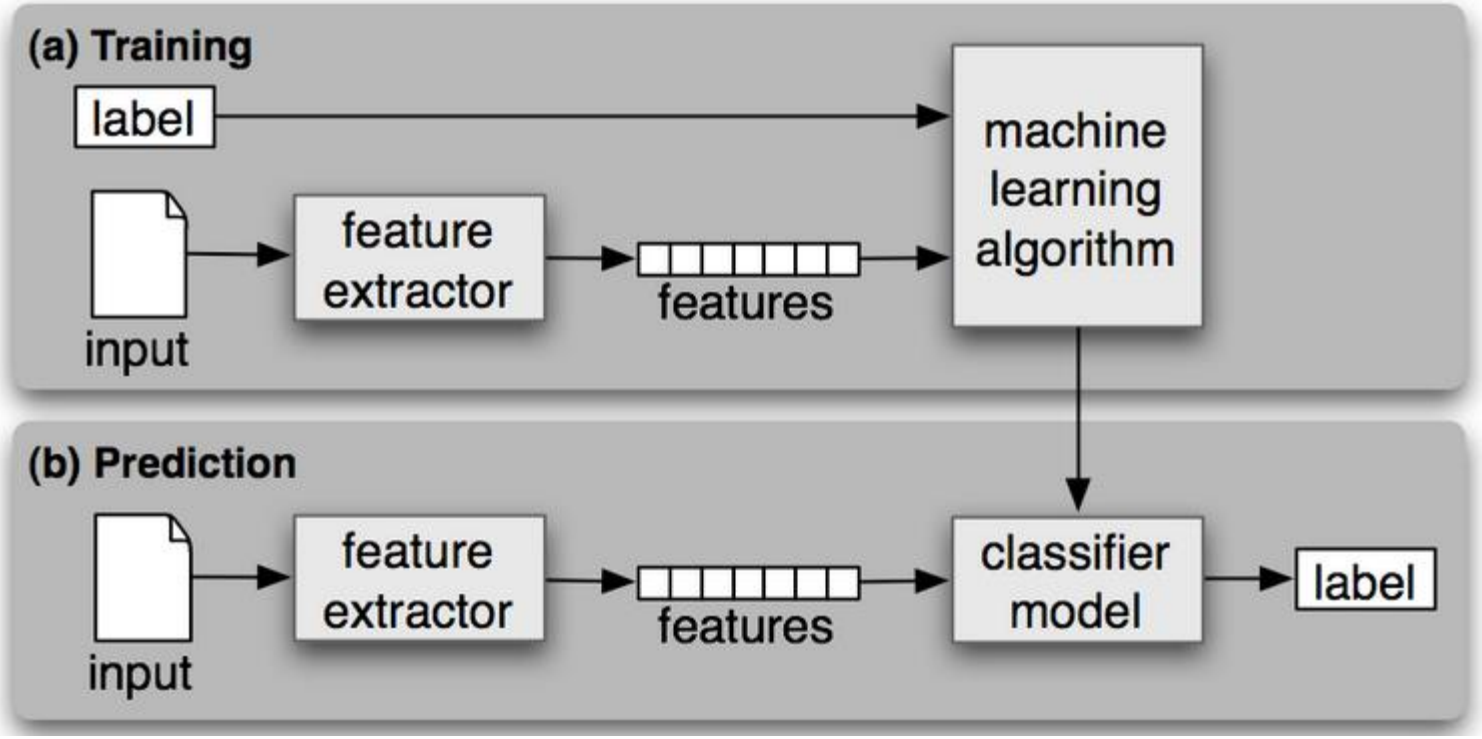
Classification



Target: $y \in \{\ldots, -1, 1, \ldots\}$

Supervised Learning Algorithms

Classification



http://www.nltk.org/book/ch06.html

Supervised Learning Algorithms

- **Classifier Models**
  - ✓ **Logistic Regression**
  - ✓ **Na ïve Bayes**
  - ✓ **Neural Networks**
  - ✓ **Maximum Entropy**
  - ✓ **Support Vector Machine**
  - ✓ **… …**

Supervised Learning Algorithms

- **Classifier Models**
  - ✓ **Logistic Regression**
  - ✓ **Na ïve Bayes**
  - ✓ **Neural Networks**
  - ✓ **Maximum Entropy**
  - ✓ **Support Vector Machine**
  - ✓ **… …**

Supervised Learning Algorithms

- **Performance Rank**
  - Support Vector Machine
  - Maximum Entropy
  - Neural Networks ???
  - Na ïve Bayes
  - Logistic Regression

- **Speed Rank**
  - Logistic Regression
  - Na ïve Bayes
  - Maximum Entropy
  - Support Vector Machine
  - Neural Networks

Evaluating Models

- **Infinite data is best, but…**
- **N (N=10) Fold cross validation**
  - ✓ Create N folds or subsets from the training data (approximately equally distributed with approximately the same number of samples).
  - ✓ Build N models, each with a different set of N-1 folds, and evaluate each model on the remaining fold
  - ✓ Error estimate is average error over all N models

Unsupervised Learning

- **No labels are involved in the learning procedure (unlike supervised learning)**
- **Clustering**
  - ✓ Task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
  - ✓ K-means

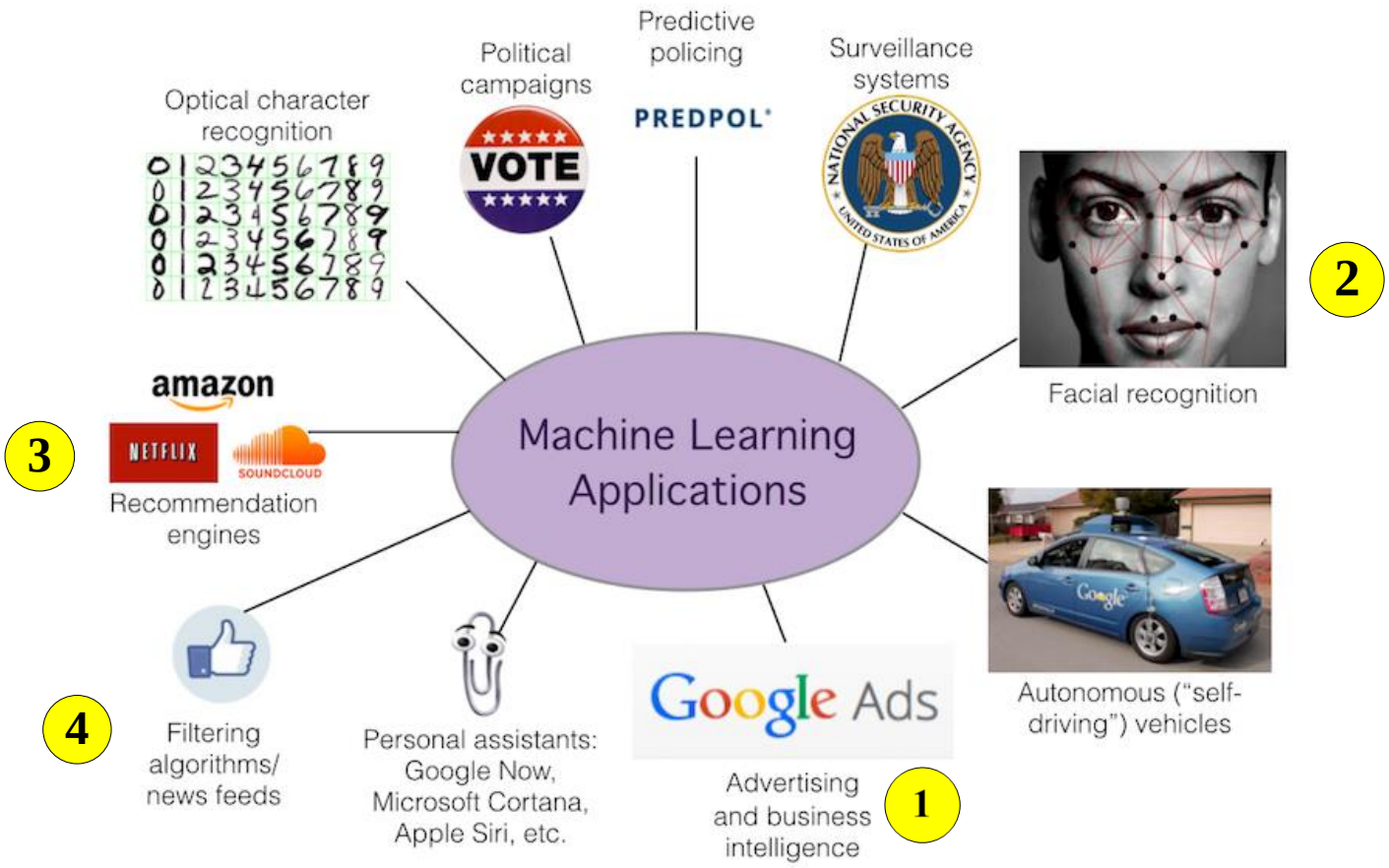
Reinforcement Learning (RL)

- **Autonomous agent learns to act "optimally" without human intervention**
- **Agent learns by stochastically interacting with its environment and getting infrequent rewards.**
- **Goal: maximize rewards**

Outline

- Review of Machine Learning
- **Machine Learning Based Applications**
- Natural Language Processing via Machine Learning

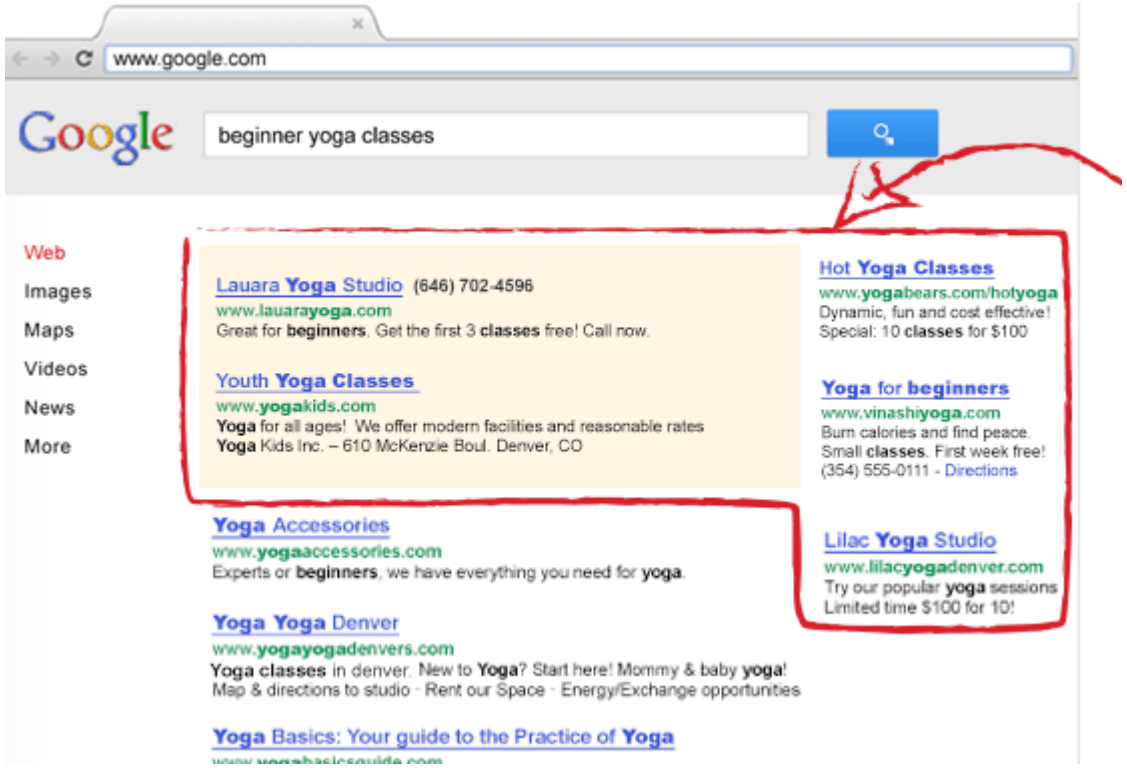
Machine Learning Based Applications



https://redshiftzero.github.io/2015/08/29/Manipulation-and-Machine-Learning/

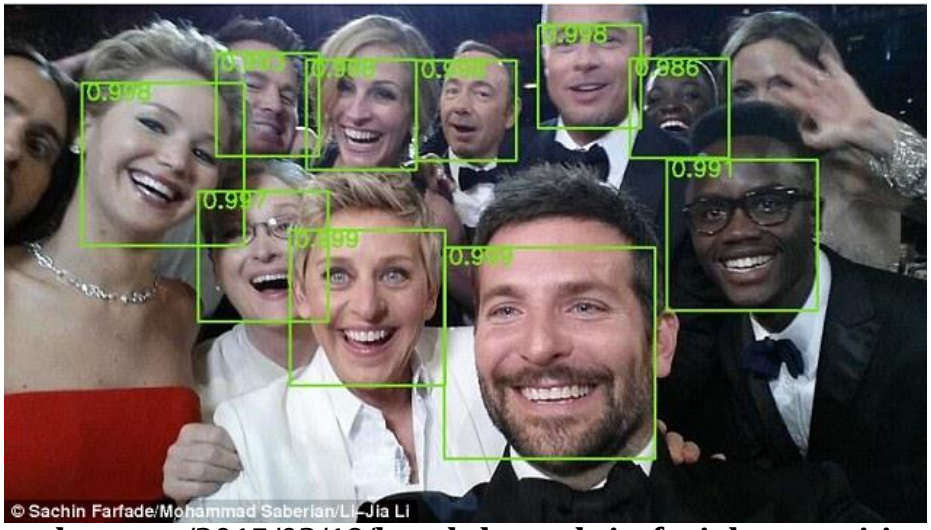
Google Ads

- **User Behavior Analysis**



Facial recognition

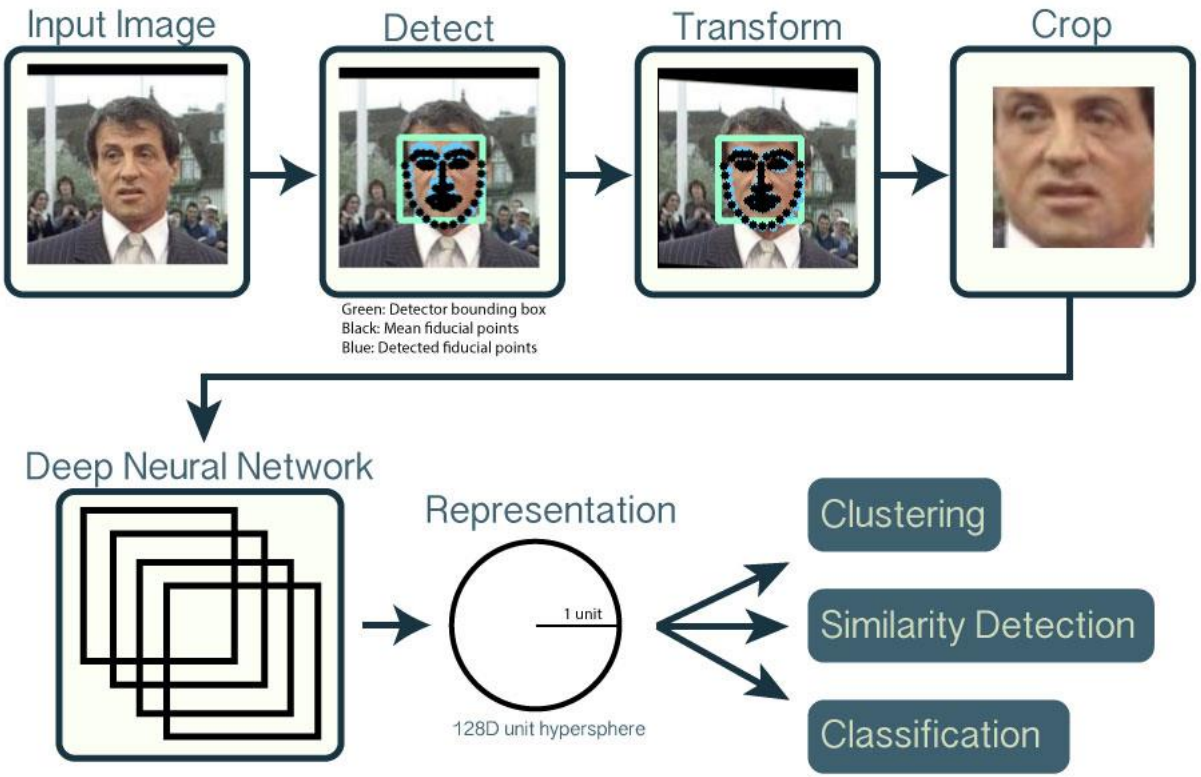
- **A facial recognition system is a computer application capable of identifying or verifying a person from a digital image or a video frame from a video source.**



https://nakedsecurity.sophos.com/2015/02/19/breakthrough-in-facial-recognition-the-deep-dense-face-detector/

Facial recognition

- **Framework**

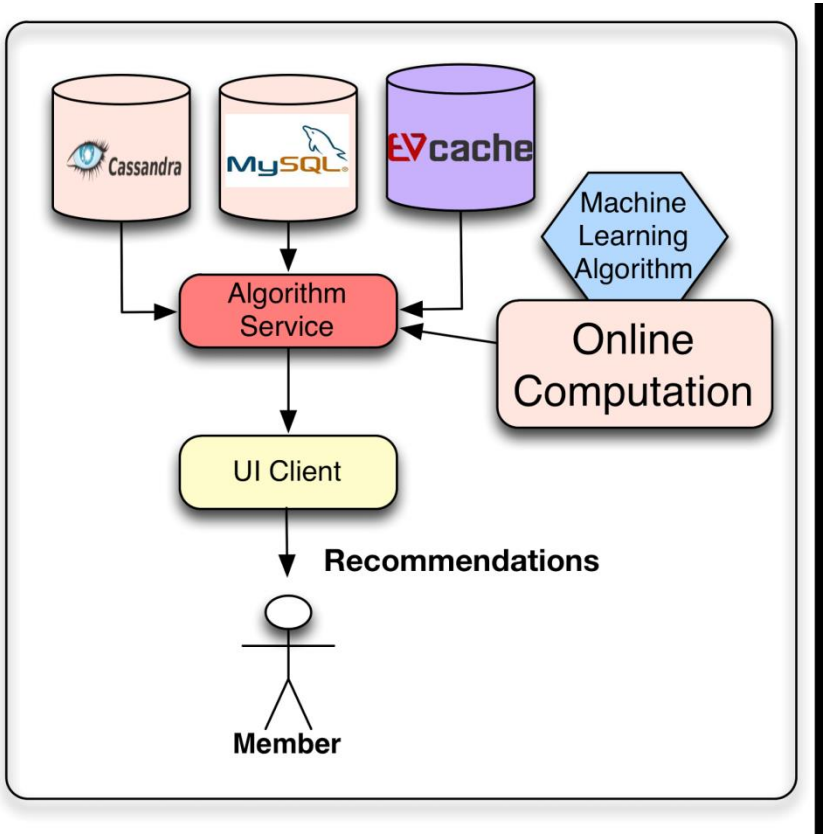


http://www.i-programmer.info/news/105-artificial-intelligence/9375-openface-face-recognition.html

Recommendation System

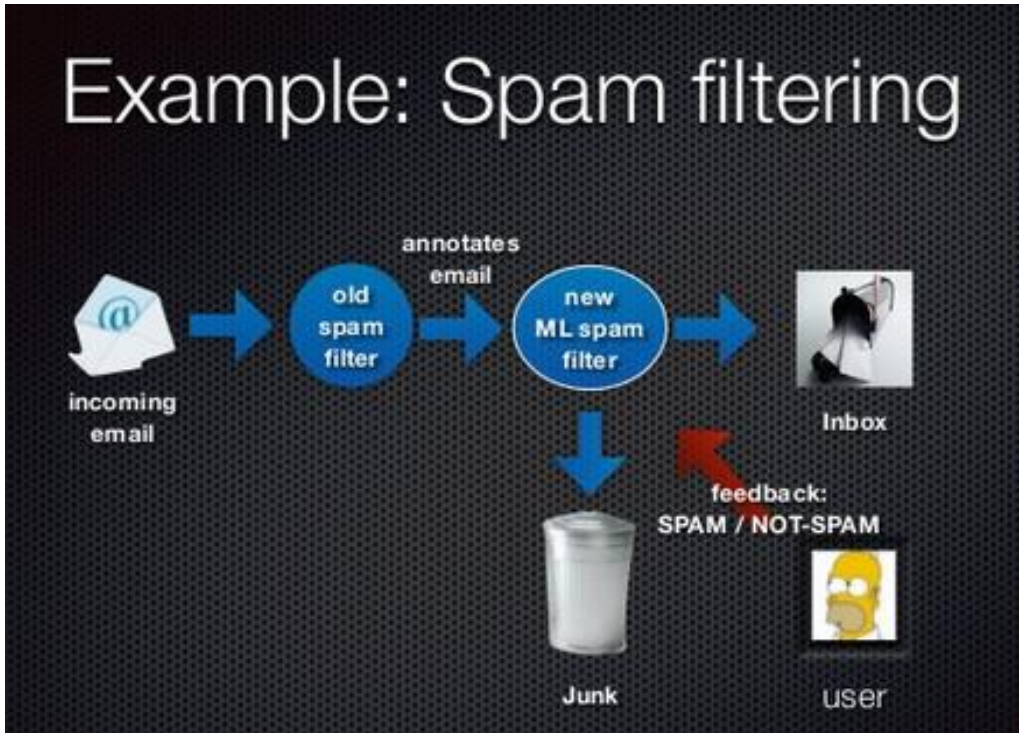
- **Recommender systems or recommendation systems are a subclass of information filtering system that seek to predict the "rating" or "preference" that a user would give to an item.**
  - ✓ Analysis on user behaviors
  - ✓ Analysis on user reviews of products in social networks

Recommendation System



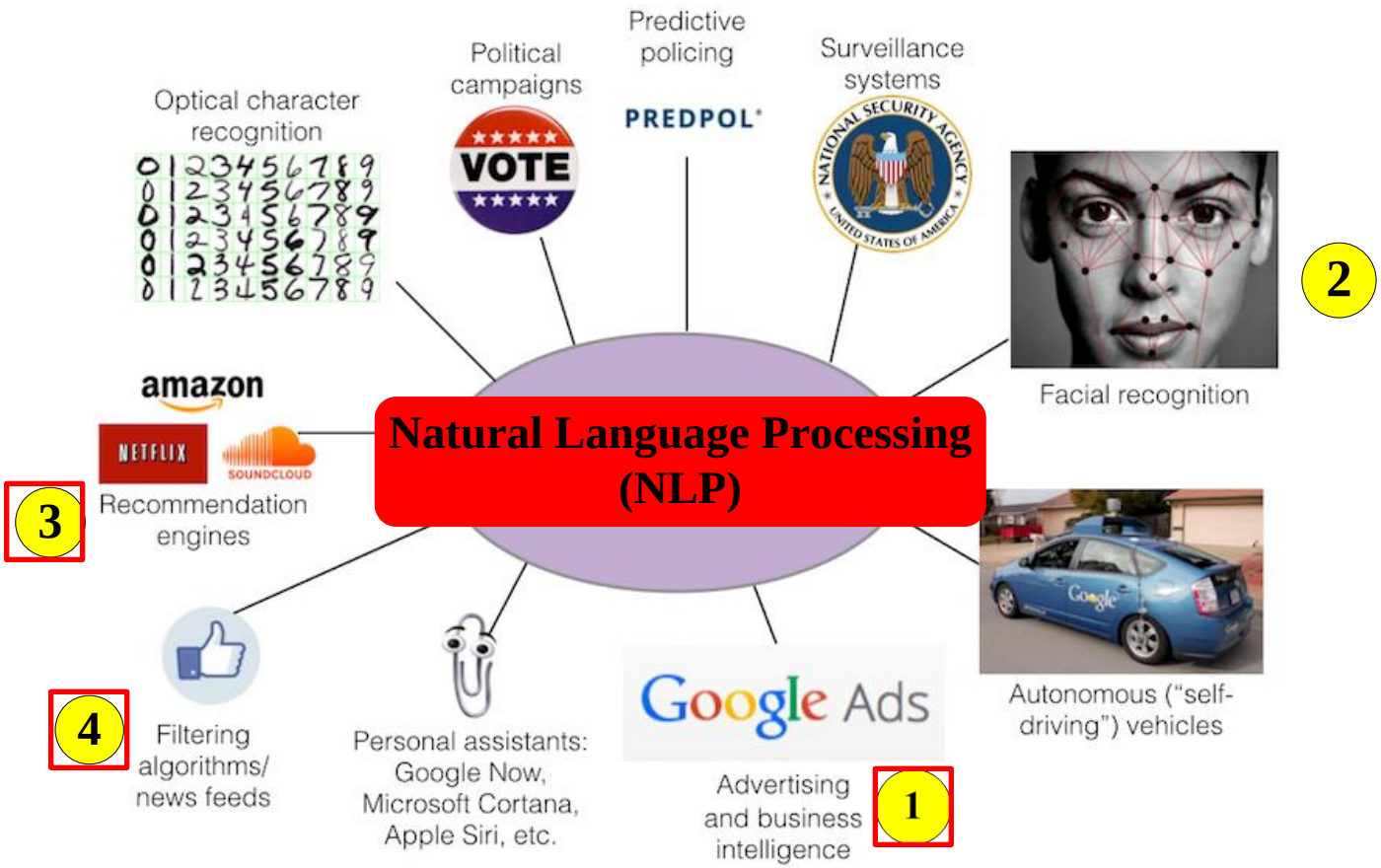
http://techblog.netflix.com/2013/03/system-architectures-for.html

Filtering



http://www.slideshare.net/StampedeCon/making-machine-learning-work-in-practice-stampedecon-2014

Machine Learning Applications



https://redshiftzero.github.io/2015/08/29/Manipulation-and-Machine-Learning/

Outline

- Review of Machine Learning
- Machine Learning Based Applications
- **Natural Language Processing via Machine Learning**

Natural Language Processing via Machine Learning

- **Text Analytics**
  - ✓ **Coarse Analytics**
    - ➢ **Text Classification**
    - ➢ **Text Clustering**
    - ➢ **… …**
  - ✓ **Fine Analytics**
    - ➢ **Lexical analysis: Word Segmentation (Chinese), Part-of-speech (POS), Named Entity Recognition**

Natural Language Processing via Machine Learning

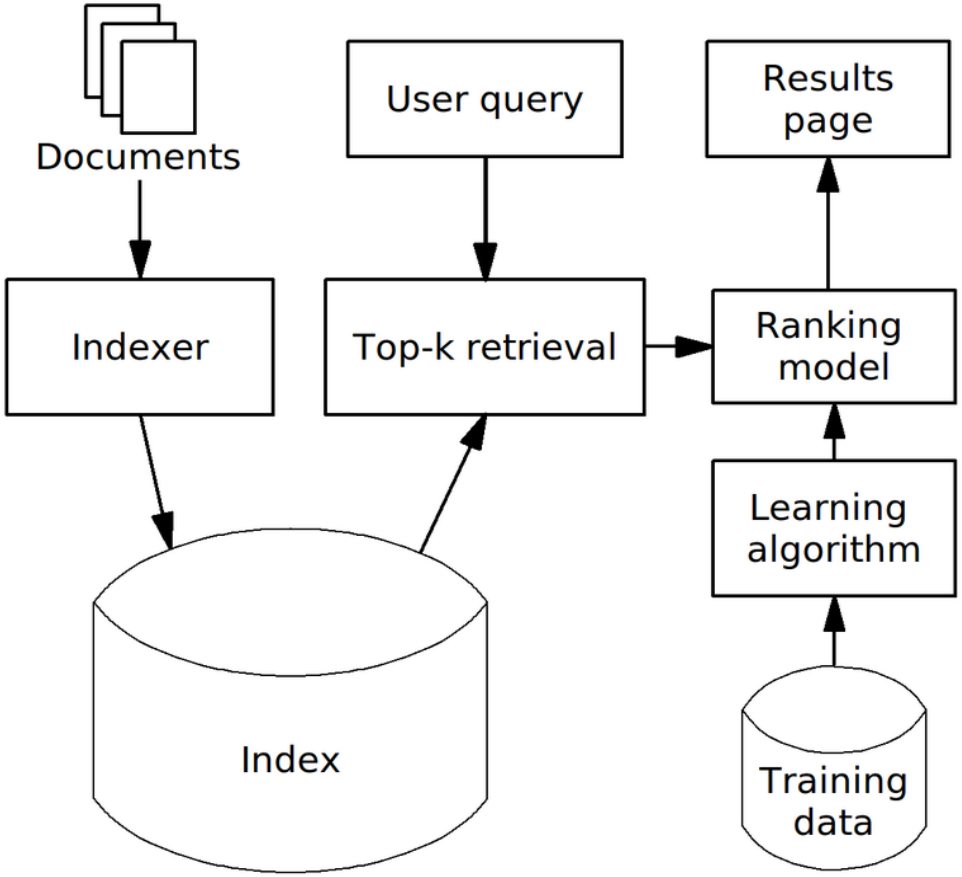
- **Text Analytics**
  - ✓ **Fine Analytics**
    - ➢ **Lexical analysis: Word segmentation (Chinese), Part-of-speech (POS), Named entity recognition, … …**
    - ➢ **Syntactic analysis: Dependency parsing, … …**
    - ➢ **Semantic analytics: Semantic role labeling, Semantic dependency analysis, … …**

Information Retrieval

- **Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources.**

Information Retrieval

- **Framework (Wikipedia)**



Information Retrieval

- **Information Retrieval via Machine Learning**
  - ✓ Learning to rank or machine-learned ranking (MLR) is the application of machine learning, typically supervised, semi-supervised or reinforcement learning, in the construction of ranking models for information retrieval systems.

Machine Translation

- **Machine translation (MT) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.**

Machine Translation

- **Neural machine translation (NMT) is a new approach to machine translation, where we train a single, large neural network to maximize the translation performance.**

Question and Answer

- **Question Answer (Q AND A) is a computer science discipline within the fields of natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language.**

Question and Answer

- **Question Answer System**

IBM Watson

IBM WATSON

- **IBM Watson** is an automated question answering system.
- It competed against Jeopardy!'s two all-time greatest champions.
- This match appeared on television in February of 2011.
- Watson won the match, outscoring both opponents combined.

More recent work on IBM Watson focuses on business applications such as medicine and customer service.

Outline

- **Review of Machine Learning**
- **Machine Learning Based Applications**
- **Natural Language Processing via Machine Learning**

Thank you!

Q&A