Machine Learning Overview

“We are entering the era of big data. For example, there are about **1 trillion web pages** ; **one hour of video is uploaded to YouTube every second**, amounting to 10 years of content every day; the genomes of 1000s of people, each of which has a length of 3.8 × 109 base pairs, have been sequenced by various labs; **Walmart handles more than 1M transactions per hour and has databases containing more than 2.5 petabytes** (2.5 × 1015) of information (Cukier 2010); and so on. This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, **we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)**.”

-- Kevin Murphy, Google

Supervised Learning

You are given labels associated with each data point. Your goal is to predict those labels.

Example Problems:

- Regression
- Classification
- Structured Prediction
- Ranking

Unsupervised Learning

You are not given any labels. Goal is to recover the underlying structure of the data.

Example Problems:

- Clustering
- Topic Modeling
- Generative Models

Reinforcement Learning

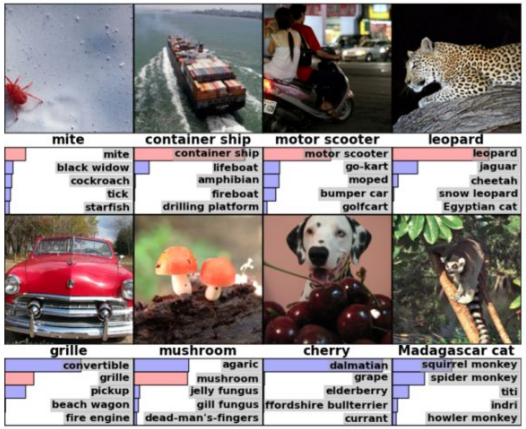
Your goal is to maximize long-term rewards by taking an action at each time step.

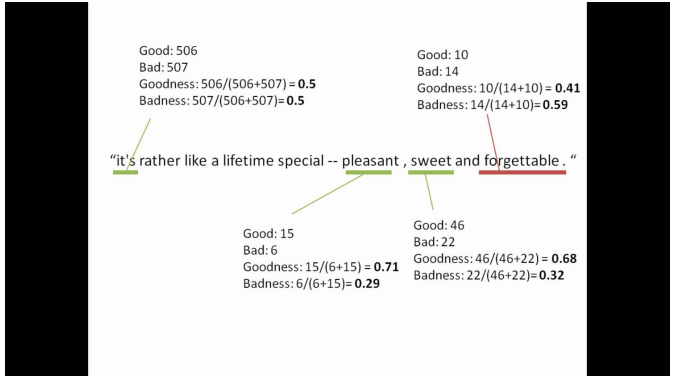
Example Problems:

- Sequential Decision Making
- LIFE!

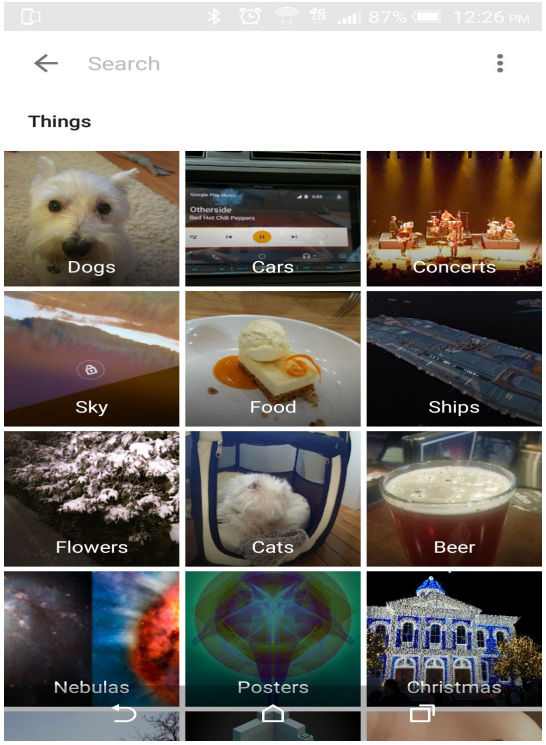
Supervised Learning: Classification

Imagenet Challenge





Sentiment Analysis



Google Photos

Supervised Learning: Regression

OFF MARKET

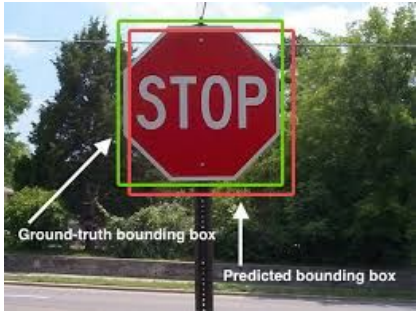
Zestimate®:

$263,528

Price this home

Rent Zestimate®: $1,750 /mo

Home Valuation



Bounding Boxes

Word = ‘apple tv’

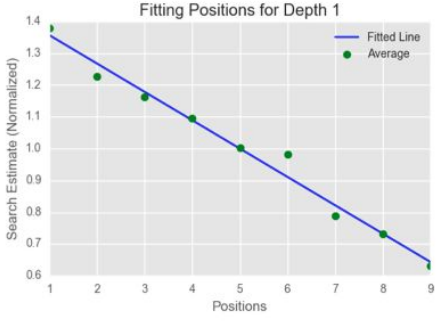
Depth = 1

Semantic Lookup= ‘electronics:apple’

Scaling Factor = 1.28

Base Estimate = 52,784,300

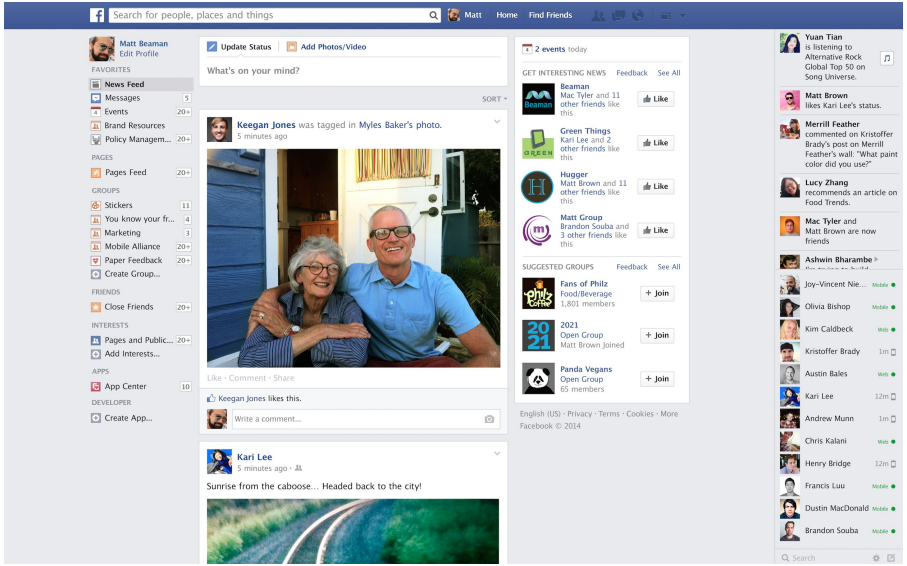
Search Estimates for Keywords on e-commerce sites



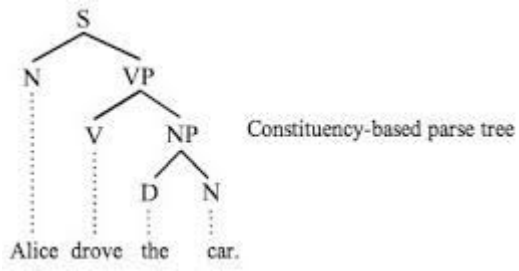
Predicting Equities

Supervised Learning: Ranking, Structured Prediction

Ranking

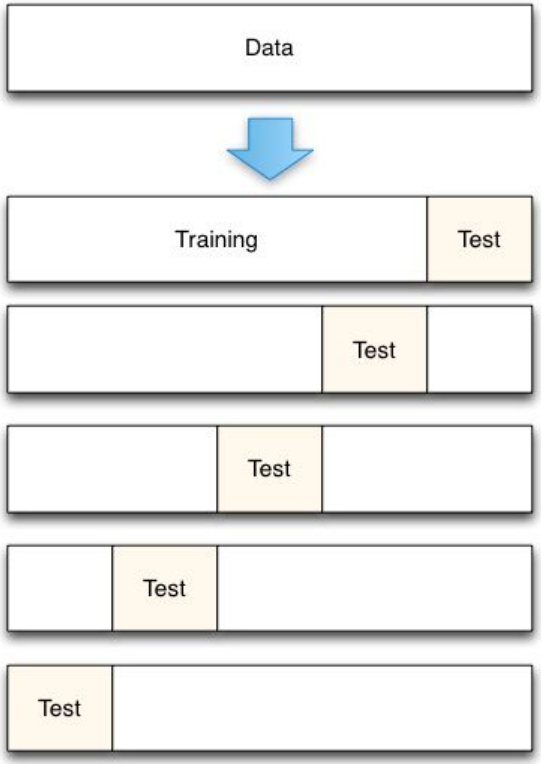


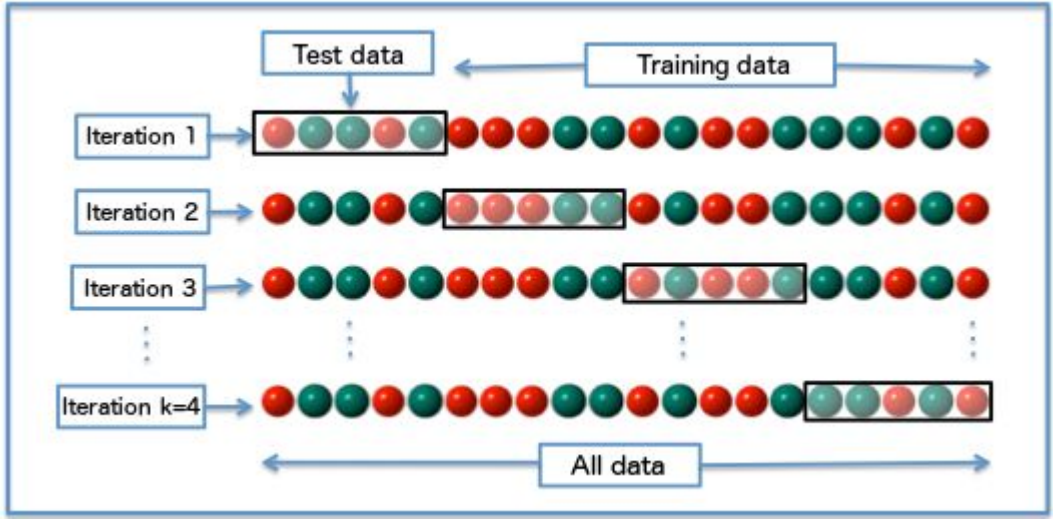
Structured Prediction



Constituency-based parse tree

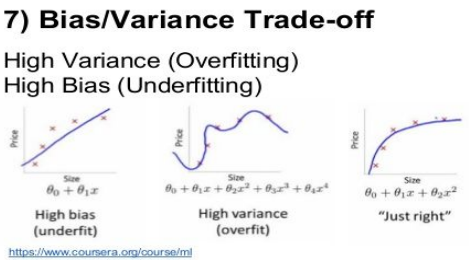
Machine Learning: Value of Different Datasets





K-fold Cross Validation

Supervised Learning: Bias Variance Trade Off



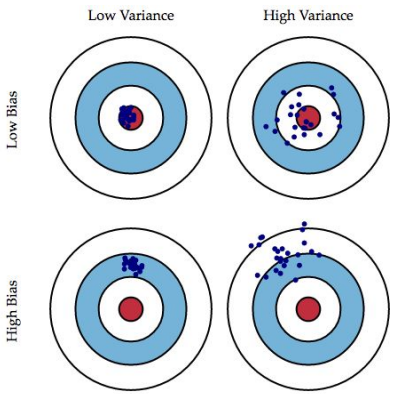
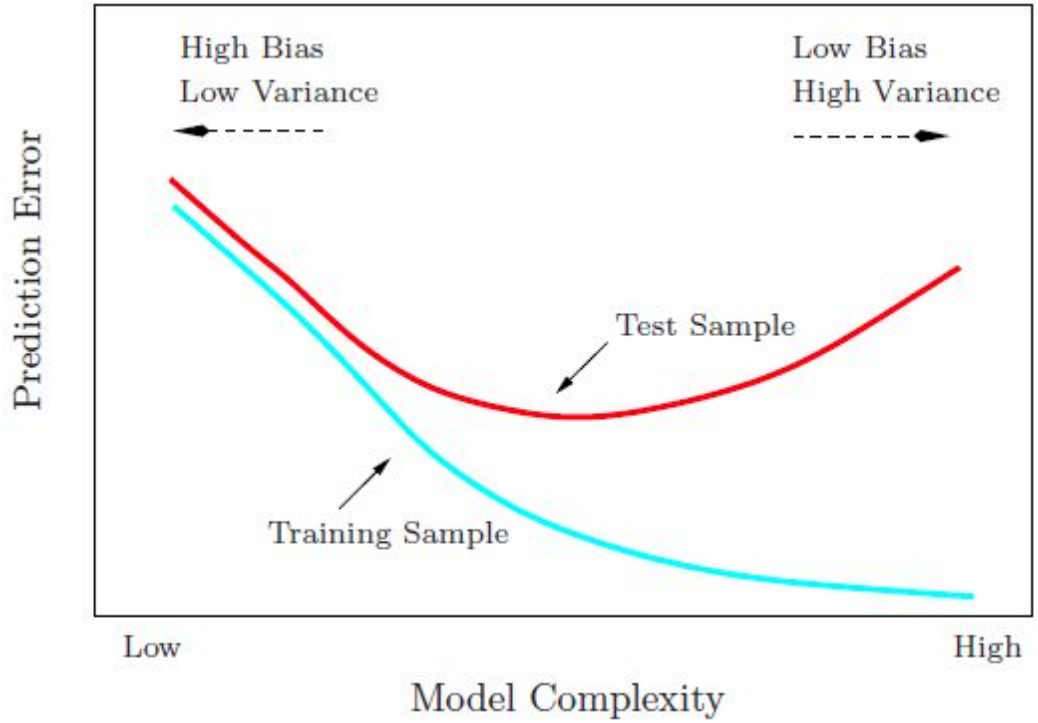


Fig. 1 Graphical illustration of bias and variance.



Supervised Learning: Metrics

Task	Metrics
Classification (Binary *and* Multi-class)	Accuracy, ROC Area Under the Curve, Precision, Recall, Confusion Matrices, log-loss, F-1 Score
Regression	R-squared, Mean Squared Error, Mean Absolute Error, Median Abs Error
Ranking	Spearman's rho, Mean Average Precision

Unsupervised Learning: No Labels

Topic Models

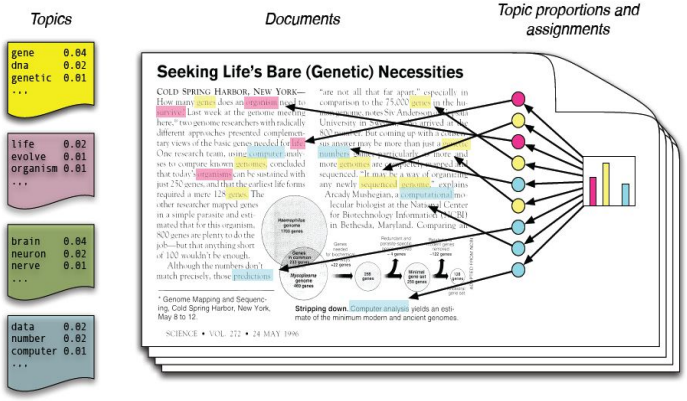
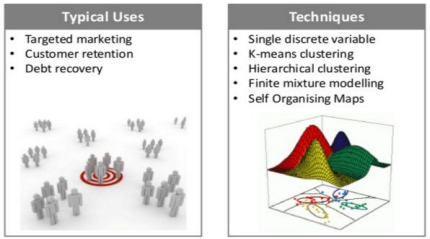


Image Generation

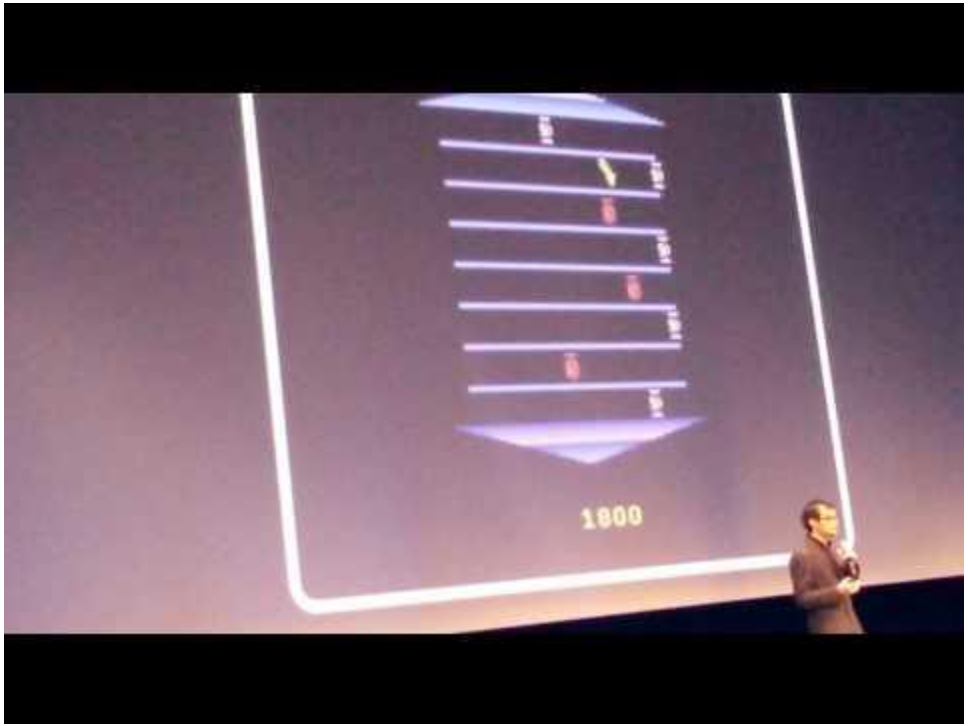
Customer Segmentation



Neural Network Learns to Select Potential Anticancer Drugs

Tue, 02/14/2017 - 11:00am by Moscow Institute of Physics and Technology

Reinforcement Learning



Checks for Understanding: What kind of learning problem is this?

- Find the most interesting patterns from a data dump of Hillary Clinton's emails
- Build a system to detect fractures from X-rays. You are given 1 million X-ray images with no or bad labels and 50,000 images from world class labelers.
- Predict future prices of a particular stock given past history of the stock and similar stocks
- Create an algorithm that predicts treatments for a life-threatening illness with only goal being to cure the disease
- Create a driving simulator that generates images of the road similar to what we see in the real world