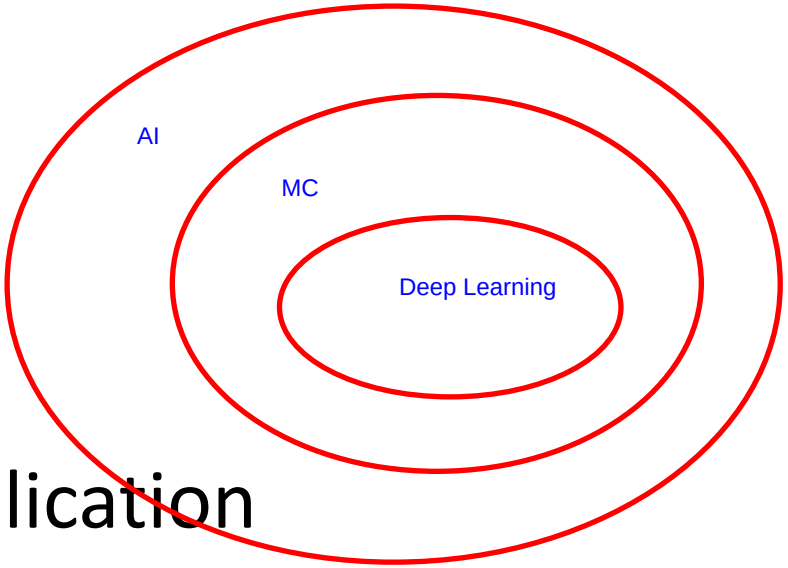
Fpt Software | CAMPUS

Introduction to
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

May 21,
2019

Agenda



- What is Machine Learning
- Learning model
- Architecture of an ML application
- Algorithms in ML
- Problems of Machine Learning
- Application of AI & ML

What is machine learning?

- A branch of **artificial intelligence**, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.
- As intelligence requires knowledge, it is necessary for the computers to acquire knowledge.

  - **Machine Learning (Mitchell 1997)**
    - Learn from past experiences
    - Improve the performances of intelligent programs
  - **Definitions (Mitchell 1997)**
    - 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 the tasks improves with the experiences

Fpt Software | CAMPUS

Classification

FROM DATA TO DISCRETE CLASSES

Spam filtering

data

Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anç show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

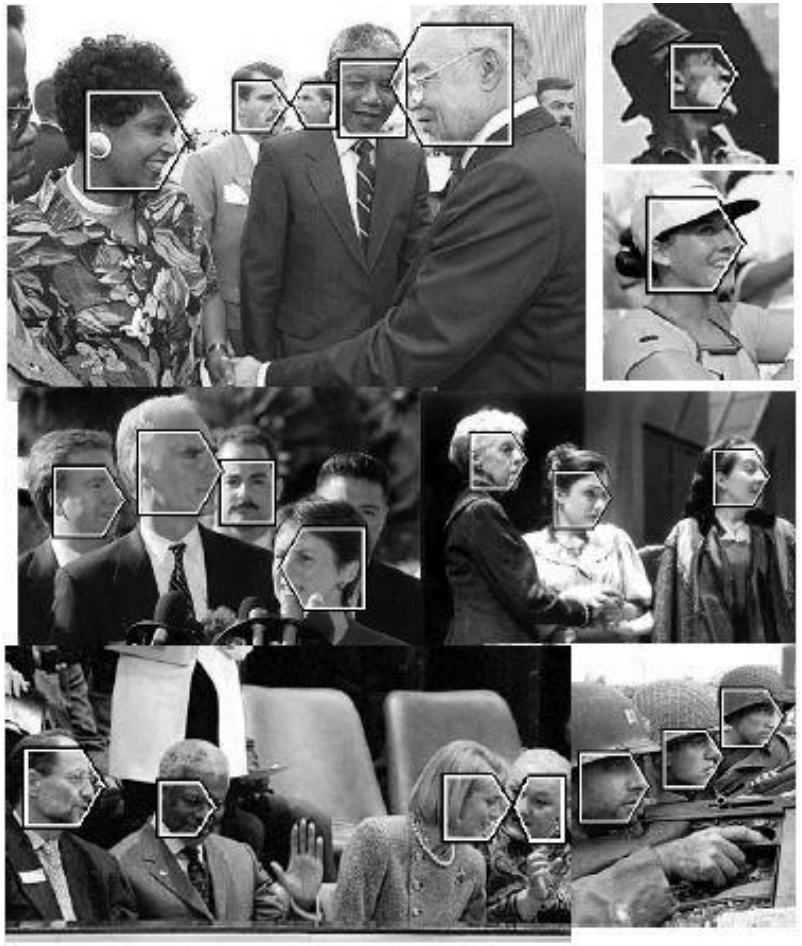
Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

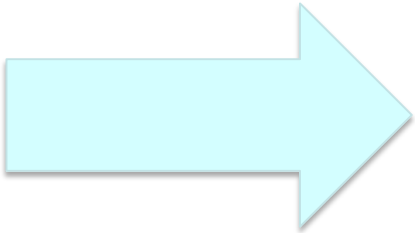
prediction

Face recognition

Example training images
for each orientation



Weather prediction



REGRESSION

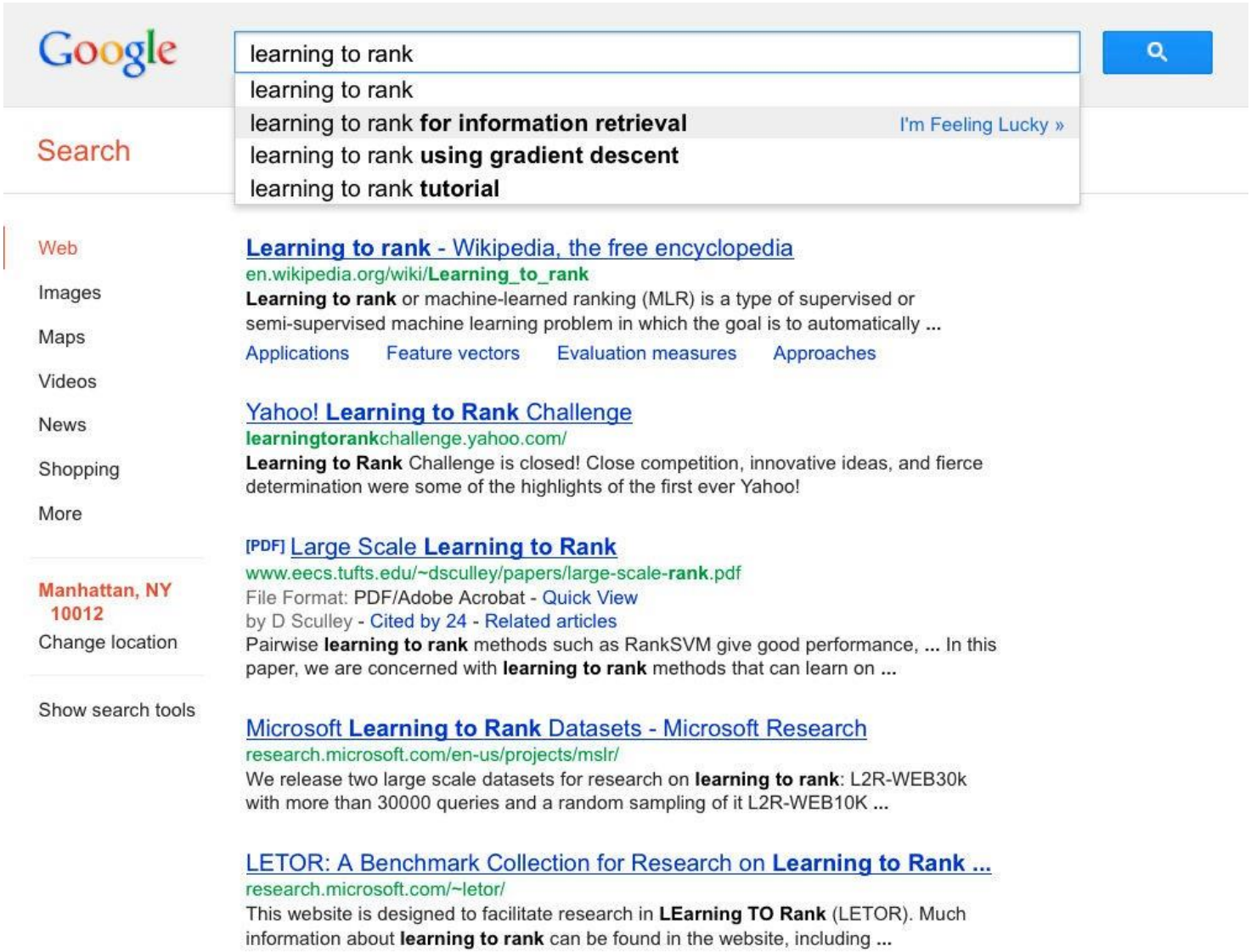
predicting a numeric value

Stock market

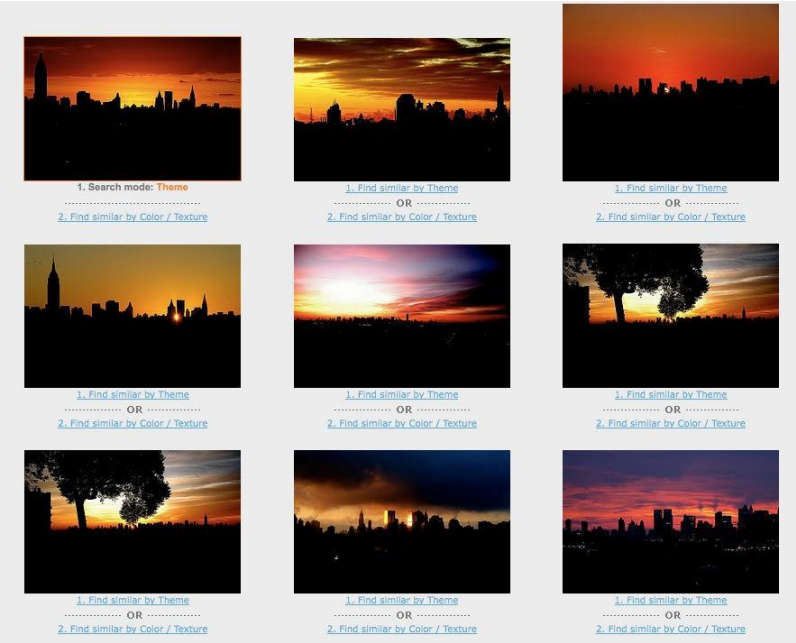
RANKING

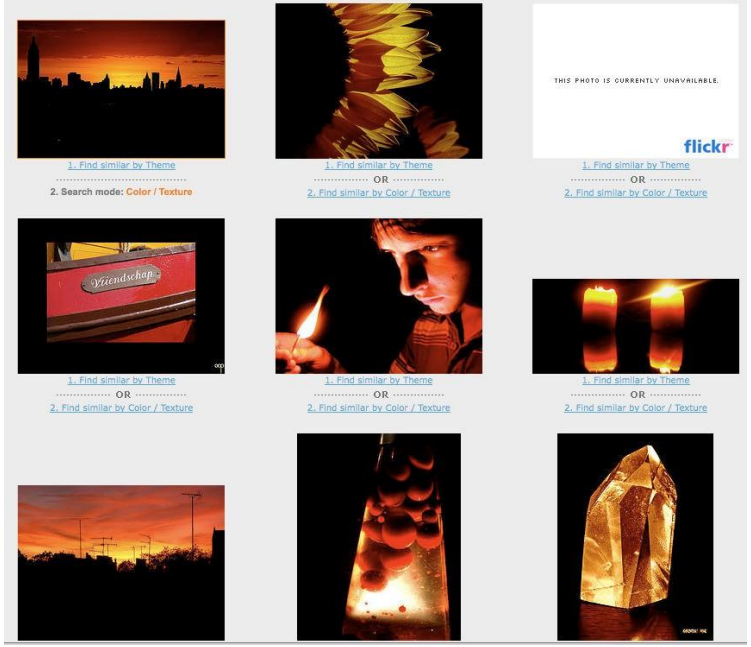
comparing items

Web search



Given image, find similar images

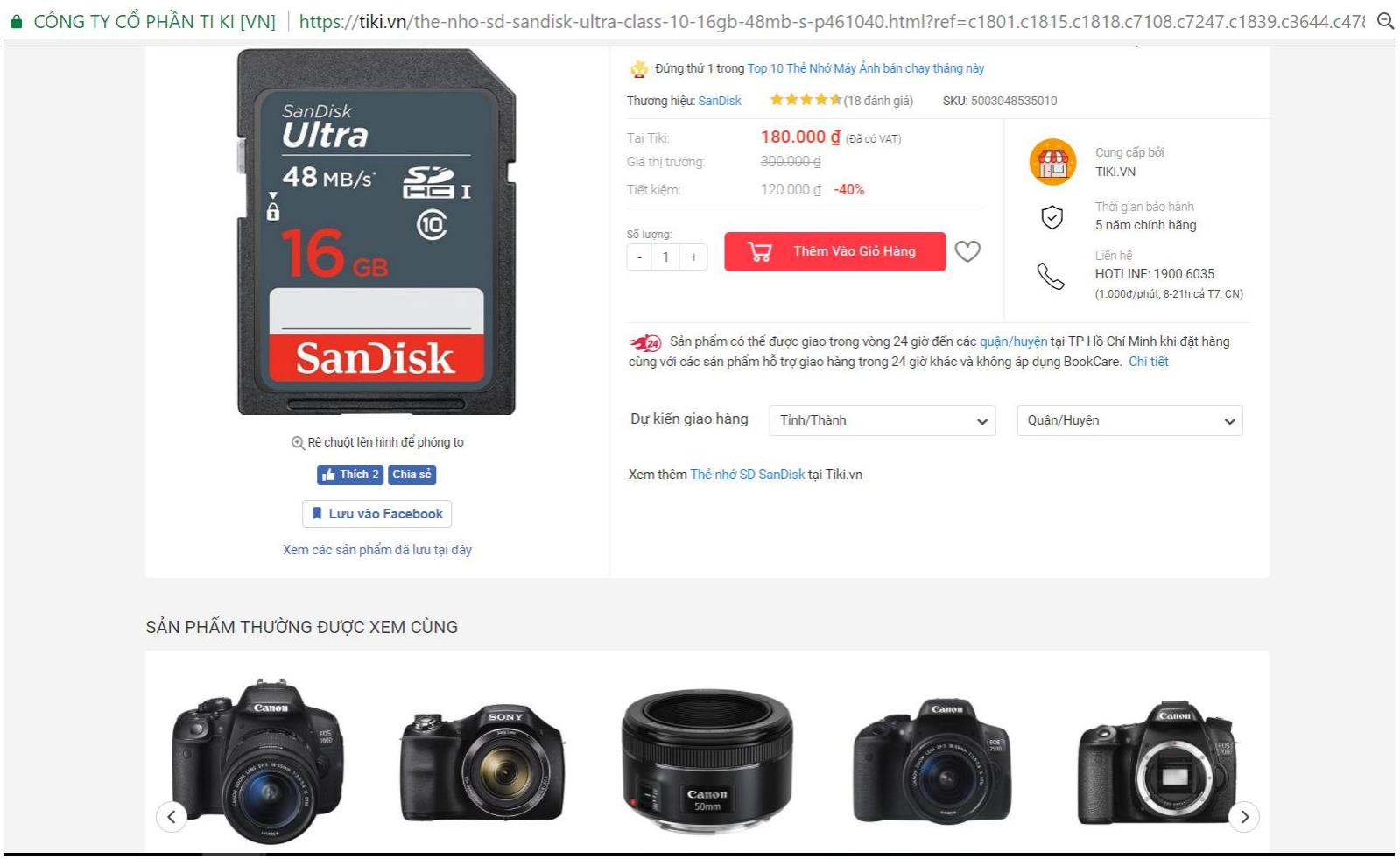




http://www.tiltomo.com/

COLLABORATIVE FILTERING

Recommendation systems



Recommendation systems

Machine learning competition with a $1 million prize

Leaderboard

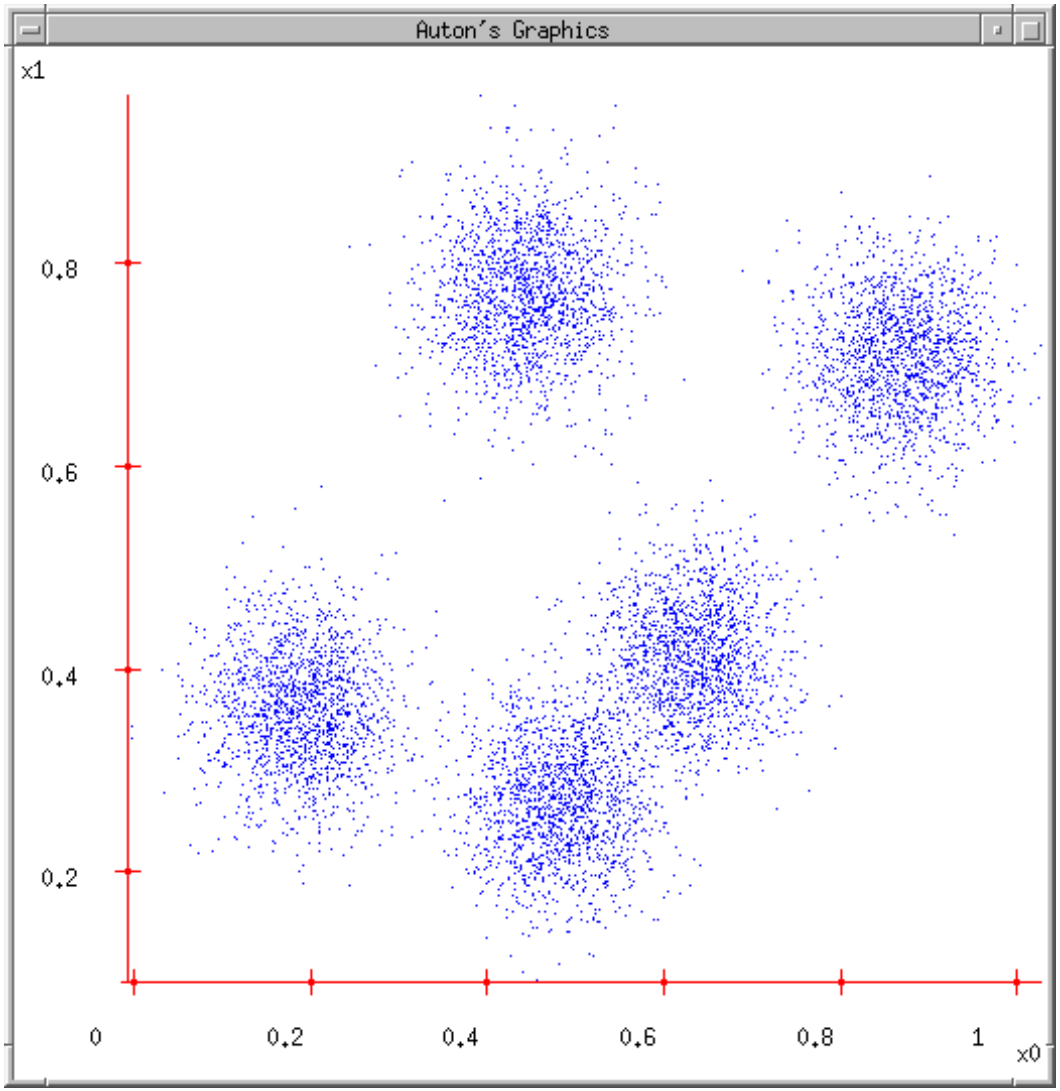
Display top 20 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
	Grand Prize - RMSE <= 0.8563			
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25 20:05:52
5	Vandelay Industries !	0.8579	9.83	2009-07-26 02:49:53
6	PragmaticTheory	0.8582	9.80	2009-07-12 15:09:53
7	BellKor in BigChaos	0.8590	9.71	2009-07-26 12:57:25
8	Dace	0.8603	9.58	2009-07-24 17:18:43
9	Opera Solutions	0.8611	9.49	2009-07-26 18:02:08
10	BellKor	0.8612	9.48	2009-07-26 17:19:11
11	BigChaos	0.8613	9.47	2009-06-23 23:06:52
12	Feeds2	0.8613	9.47	2009-07-24 20:06:46
	Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos			
13	xiangliang	0.8633	9.26	2009-07-21 02:04:40
14	Gravity	0.8634	9.25	2009-07-26 15:58:34
15	Ces	0.8642	9.17	2009-07-25 17:42:38
16	Invisible Ideas	0.8644	9.14	2009-07-20 03:26:12
17	Just a guy in a garage	0.8650	9.08	2009-07-22 14:10:42
18	Craig Carmichael	0.8656	9.02	2009-07-25 16:00:54
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	acmehill	0.8659	8.99	2009-04-16 06:29:35
	Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
	Cinematch score on quiz subset - RMSE = 0.9514			

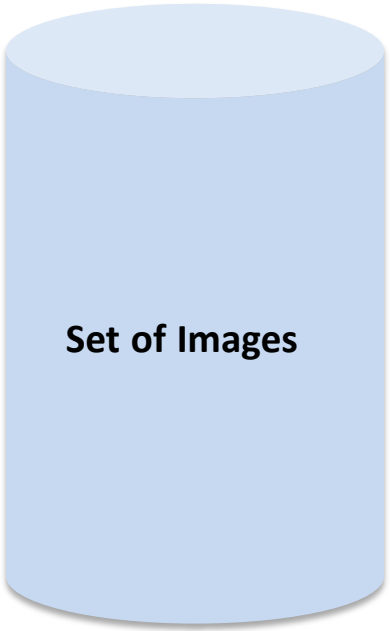
CLUSTERING

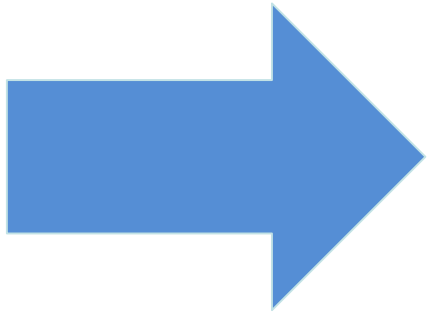
discovering structure in data

Clustering Data | Group similar things

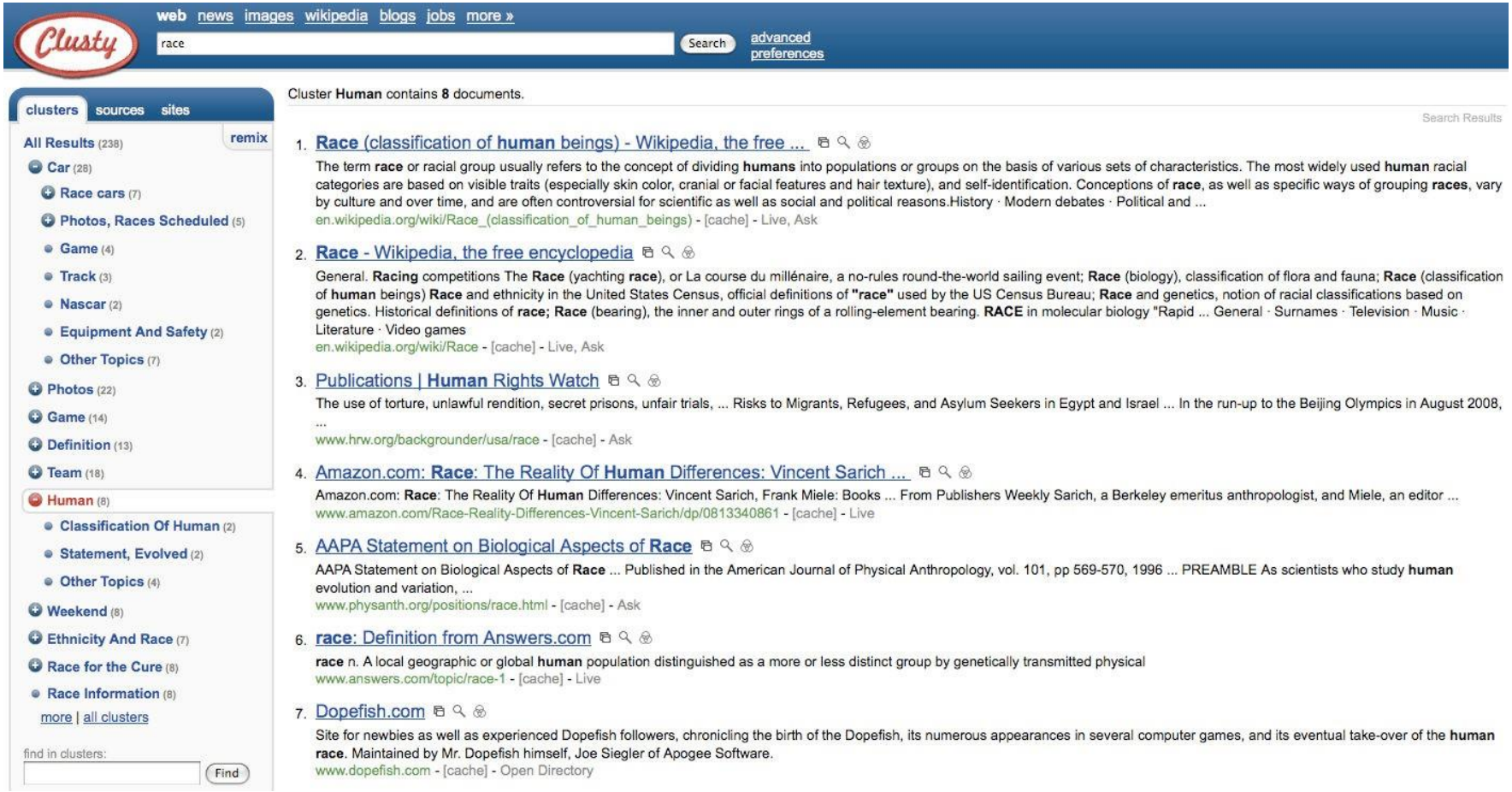


Clustering images





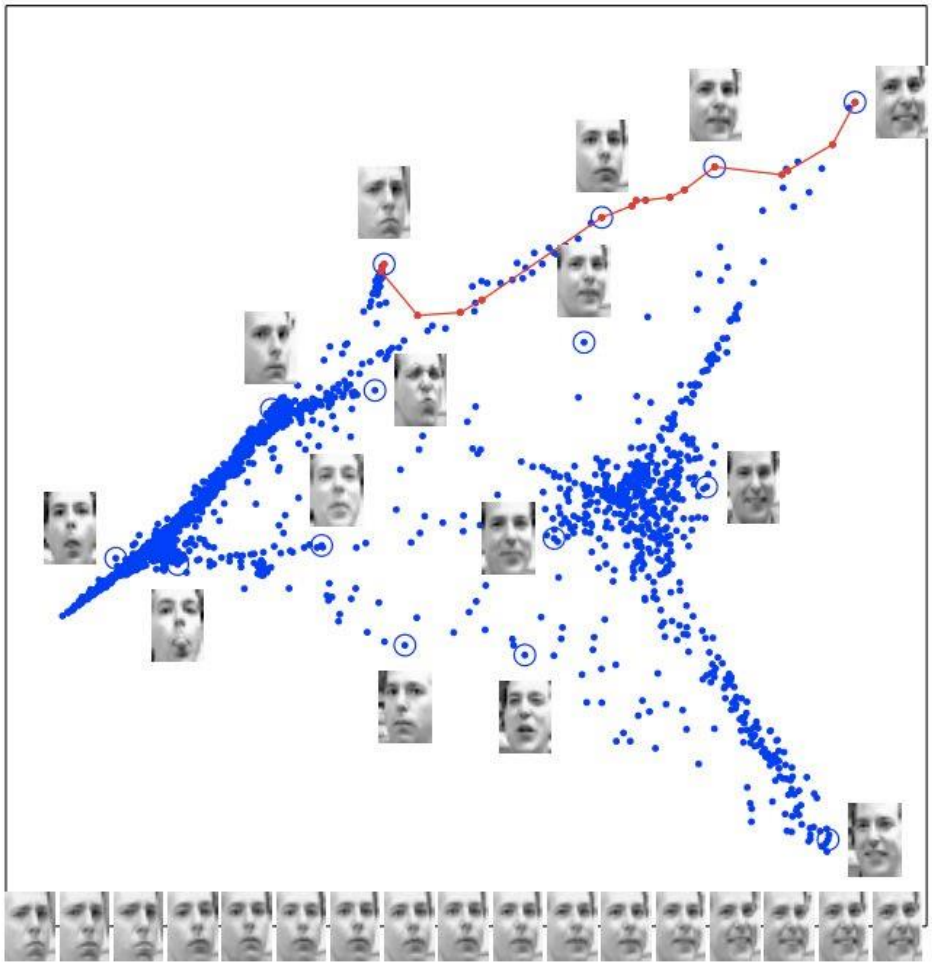
Clustering web search results



EMBEDDING
visualizing data

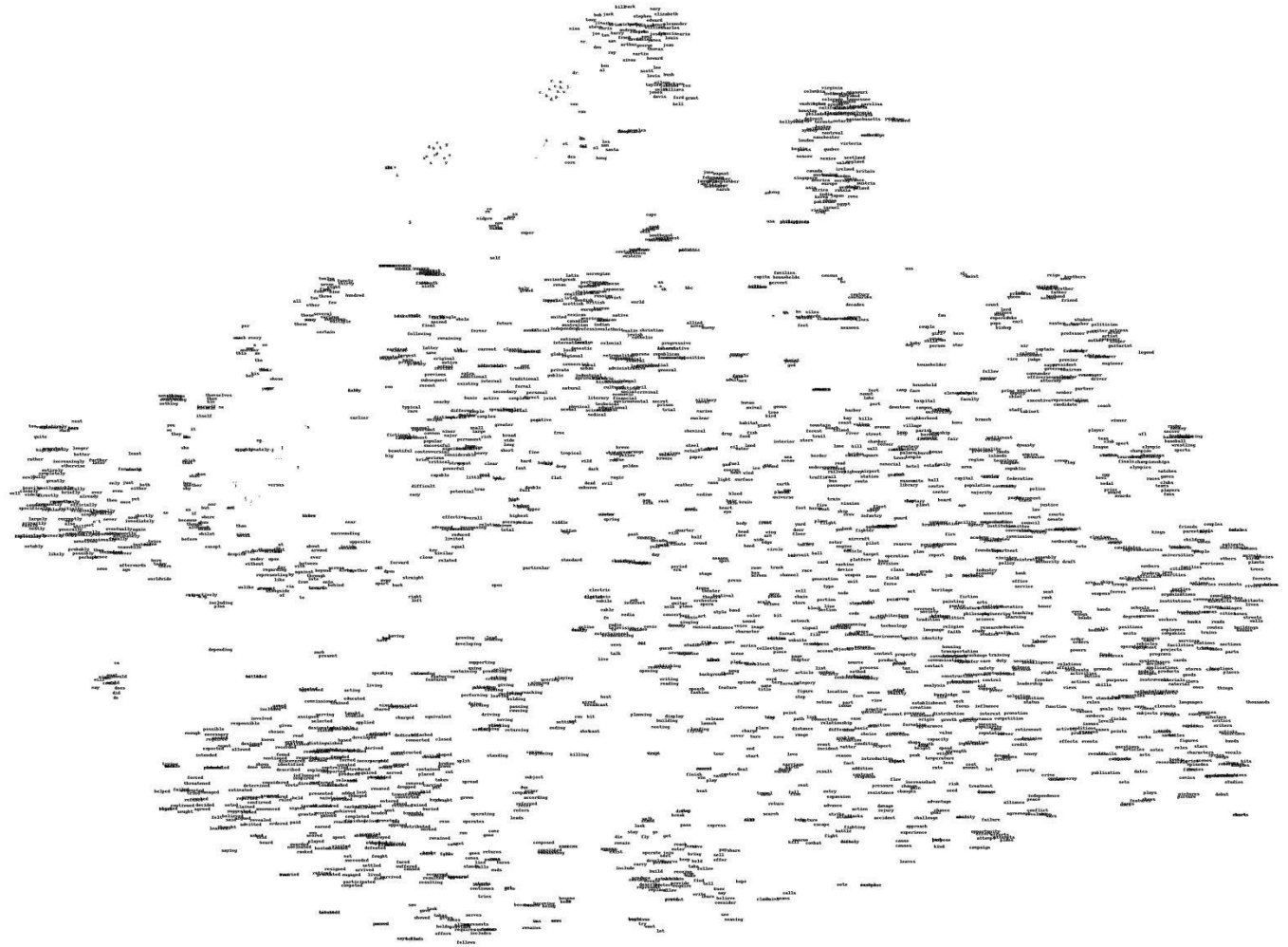
Embedding images

- Images have thousands or millions of pixels.
- Can we give each image a coordinate, such that similar images are near each other?



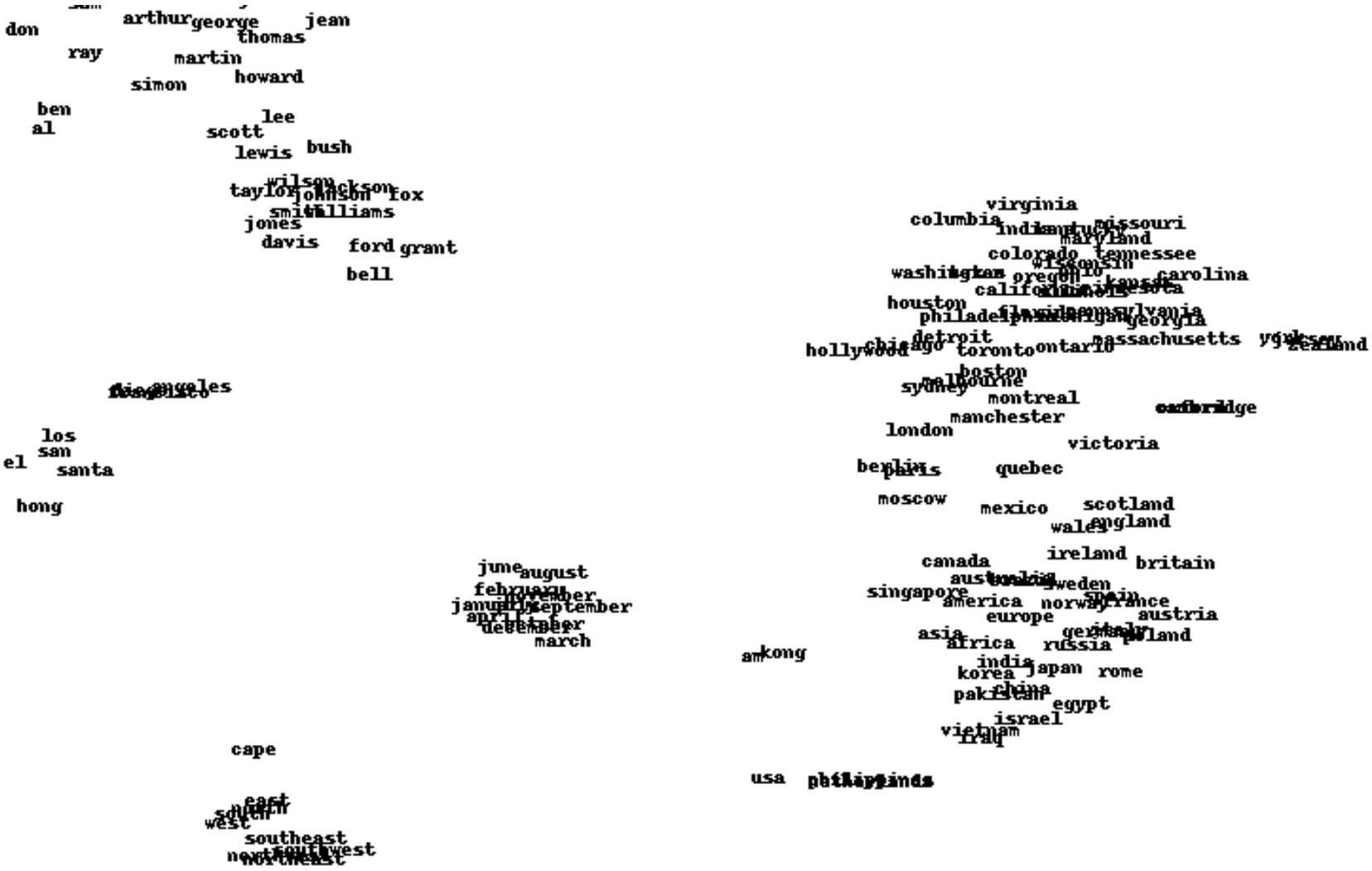
[Saul & Roweis ‘03]

Embedding words



[Joseph Turian]

Embedding words (zoom in)

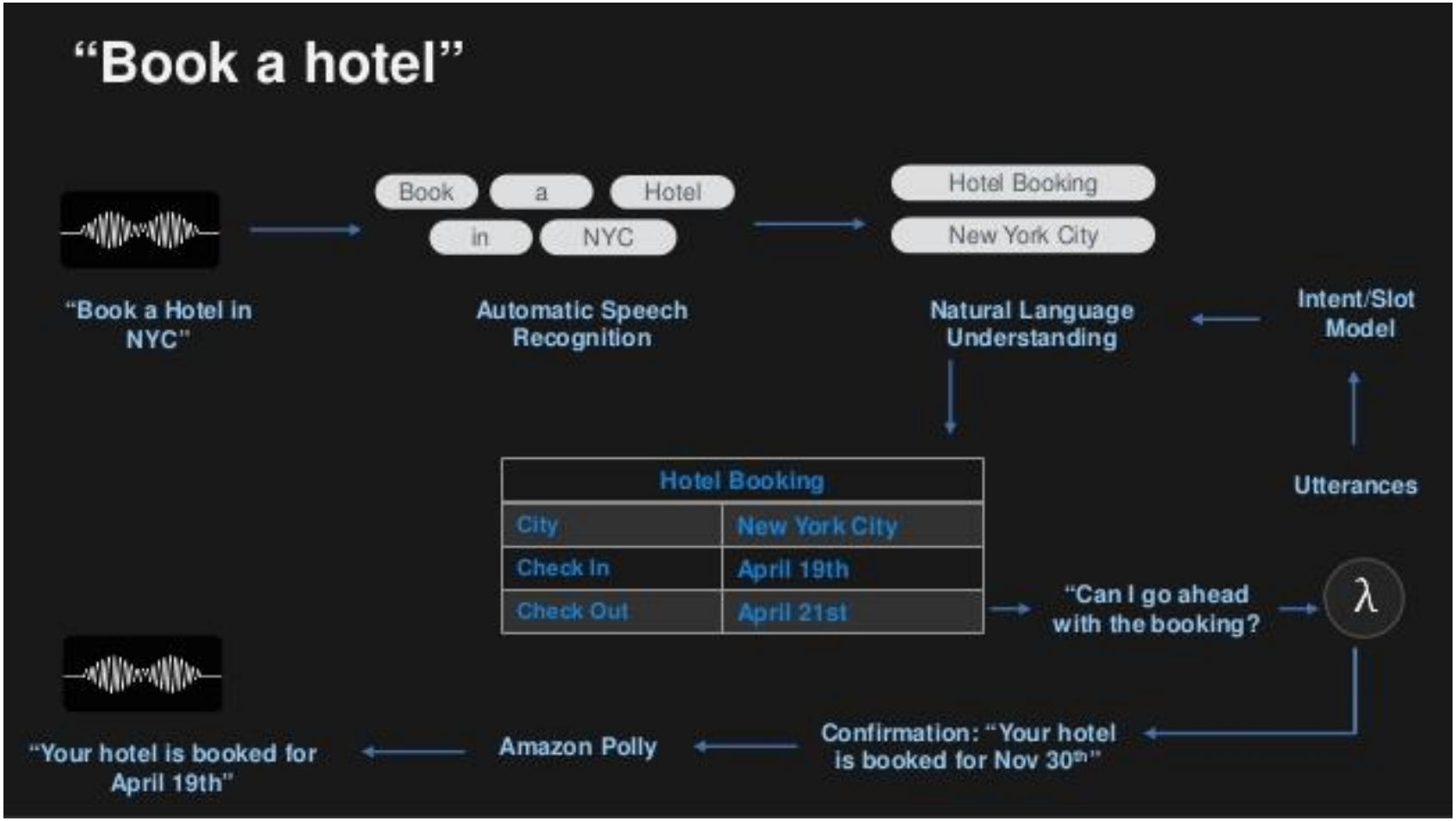


[Joseph Turian]

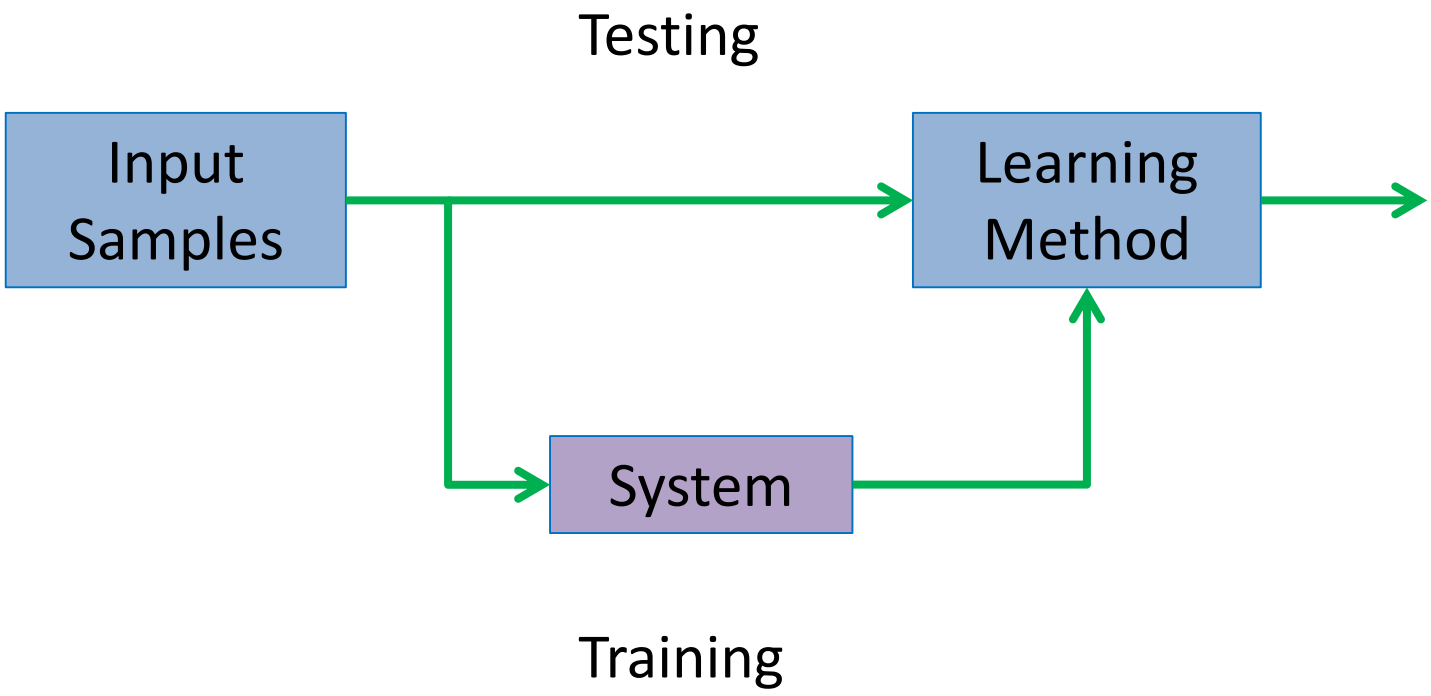
STRUCTURED PREDICTION

from data to discrete classes

Speech recognition

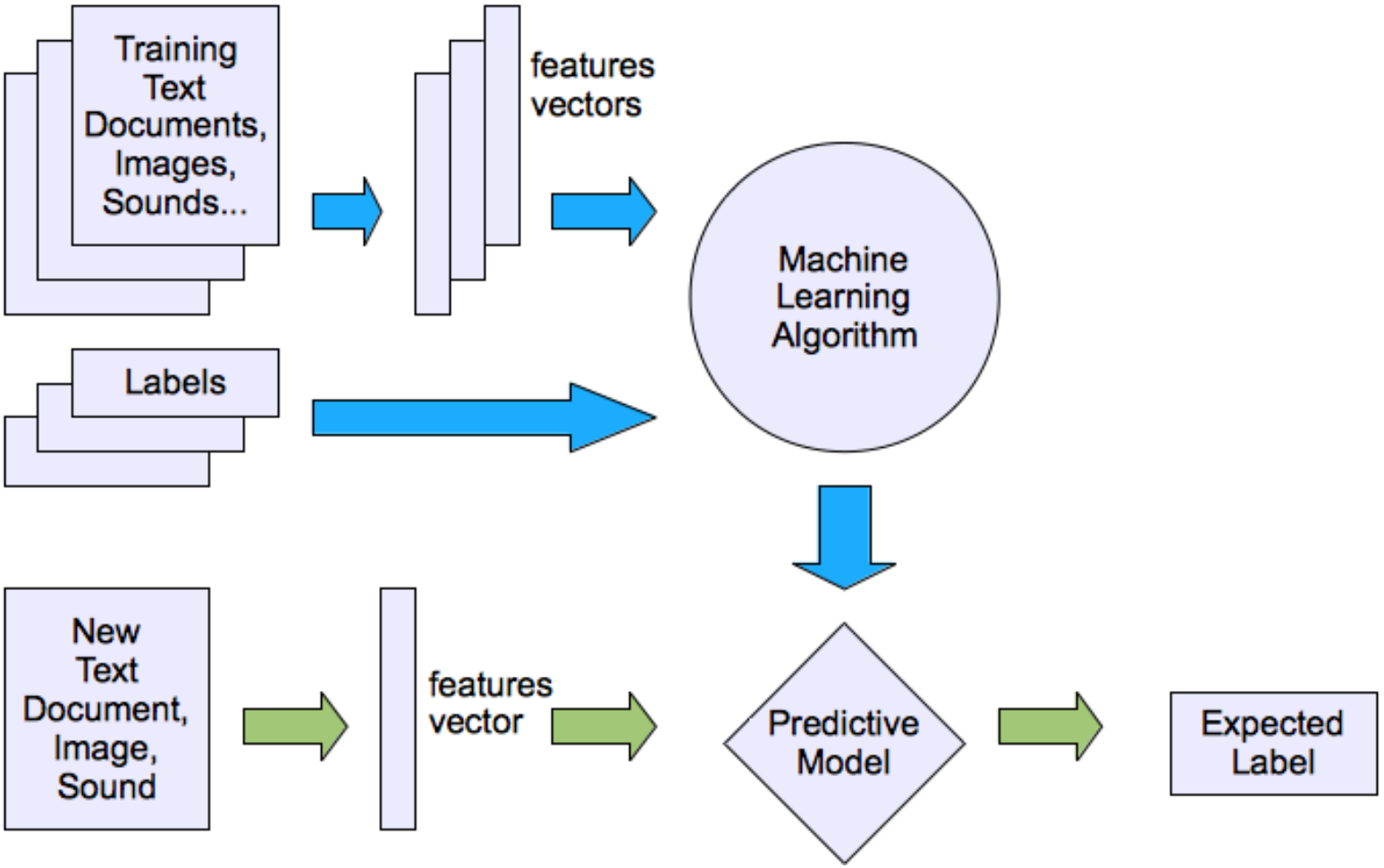


Model of a Machine Learning System



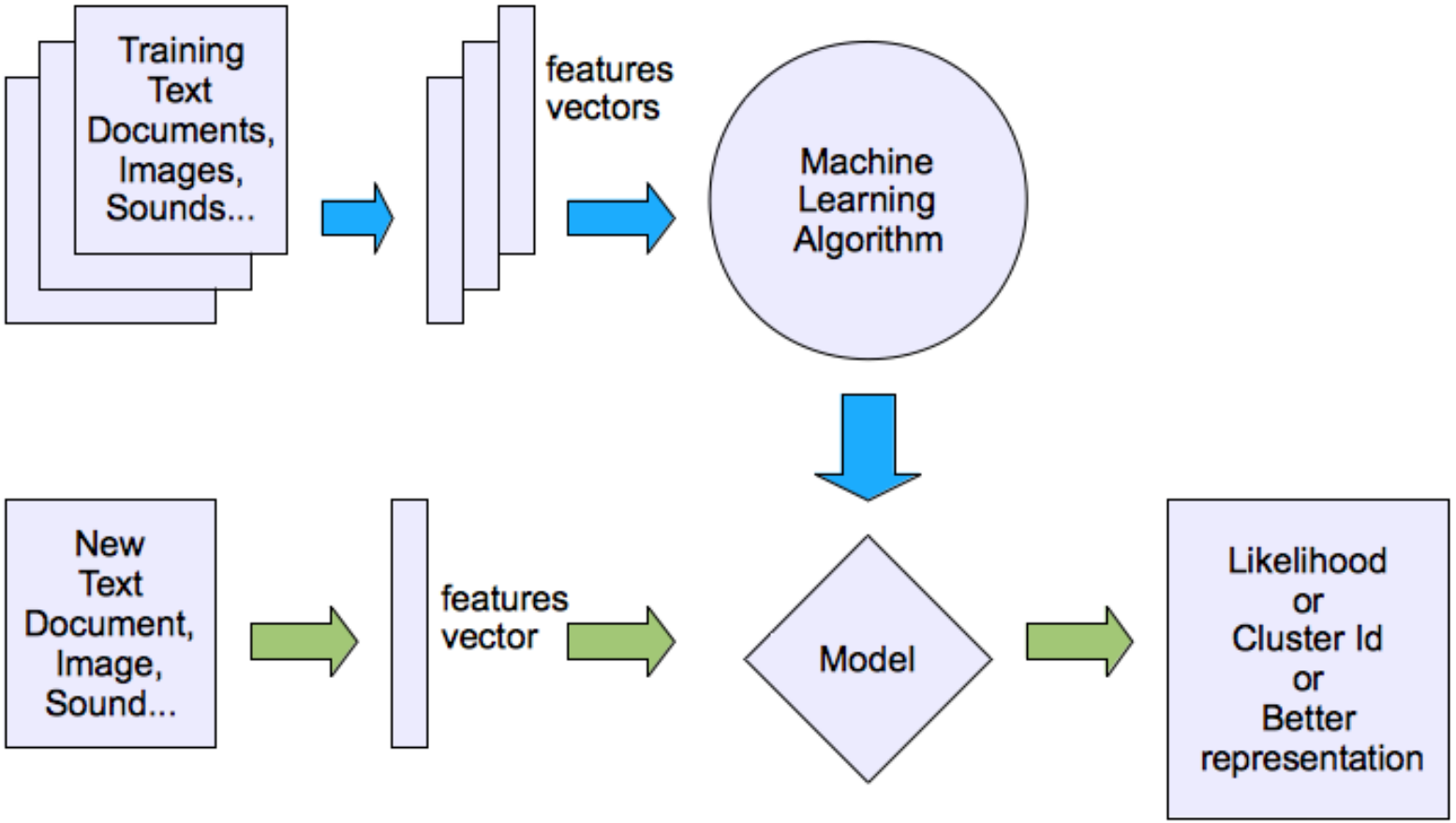
Architecture of an ML application

Supervised learning



Architecture of an ML application

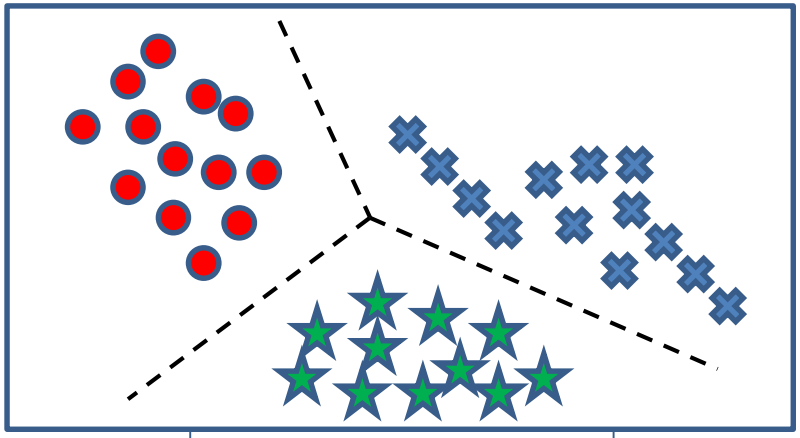
Unsupervised learning



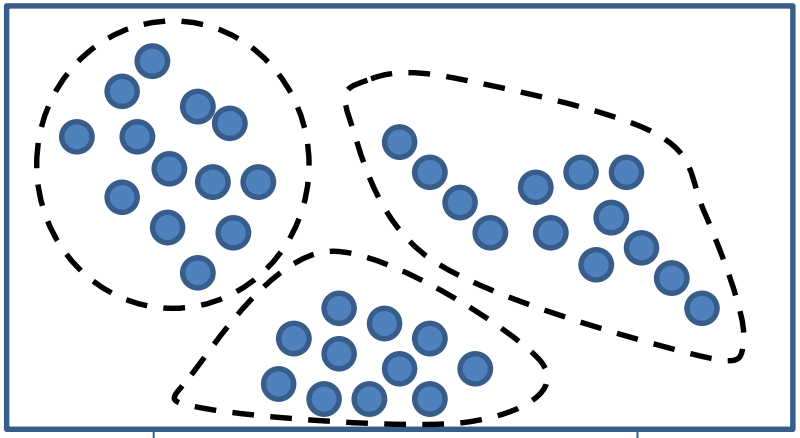
Algorithms of ML

- **Supervised learning**
  - Prediction
  - Classification (discrete labels), Regression (real values)
- **Unsupervised learning**
  - Clustering
  - Probability distribution estimation
  - Finding association (in features)
  - Dimension reduction
- **Semi-supervised learning**
- **Reinforcement learning**
  - Decision making (robot, chess machine)

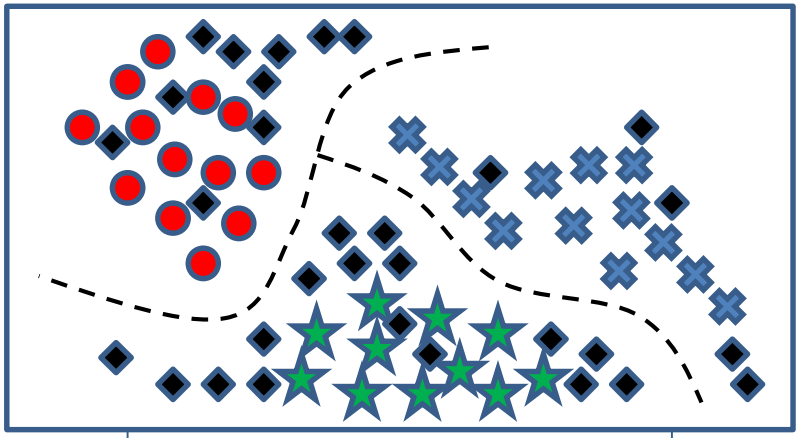
Algorithms of ML



Supervised learning



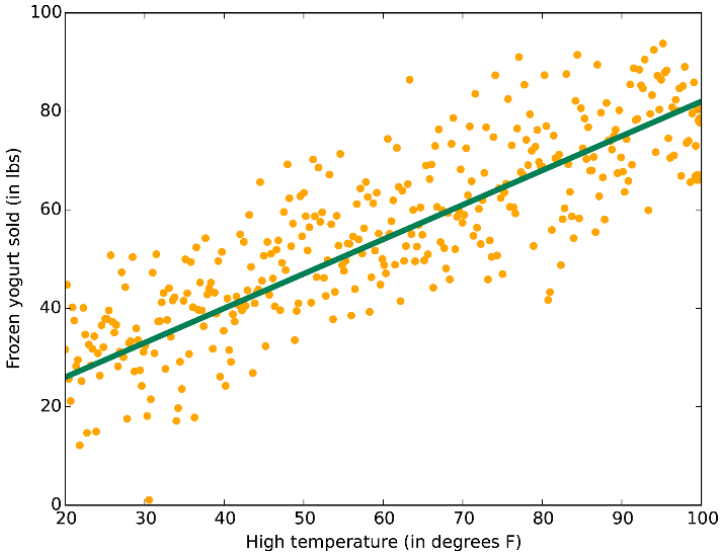
Unsupervised learning

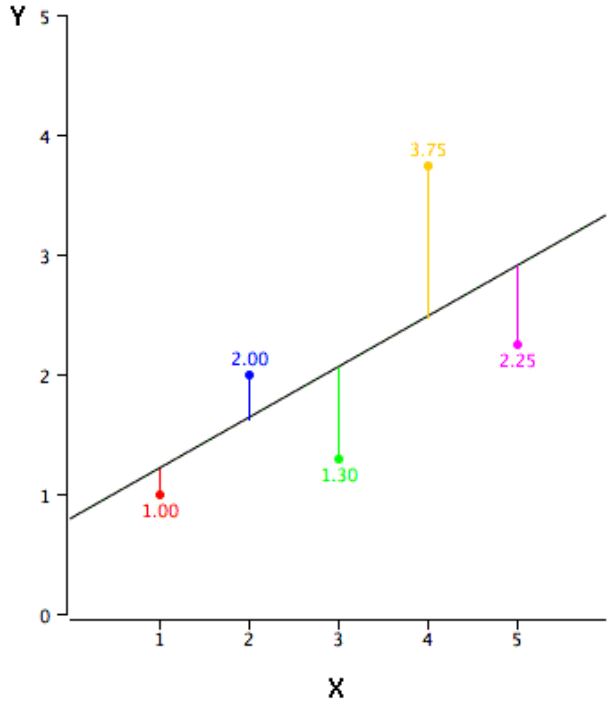


Semi-supervised learning

Linear Regression

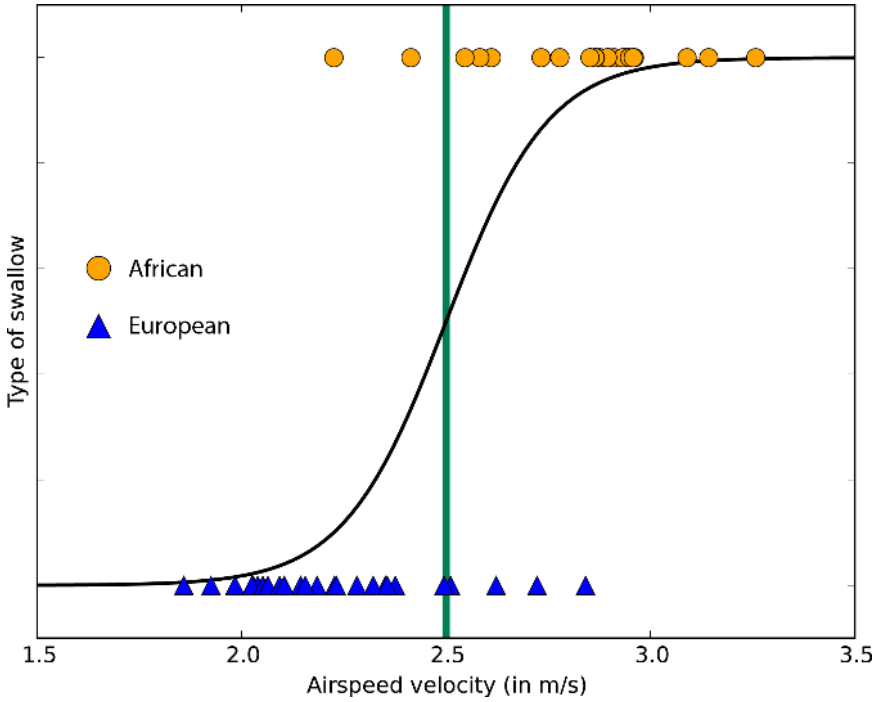
- Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).
- There must be **linear relationship** between independent and dependent variables
- Sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- **Least Square Method** is the most common method used for fitting a regression line.





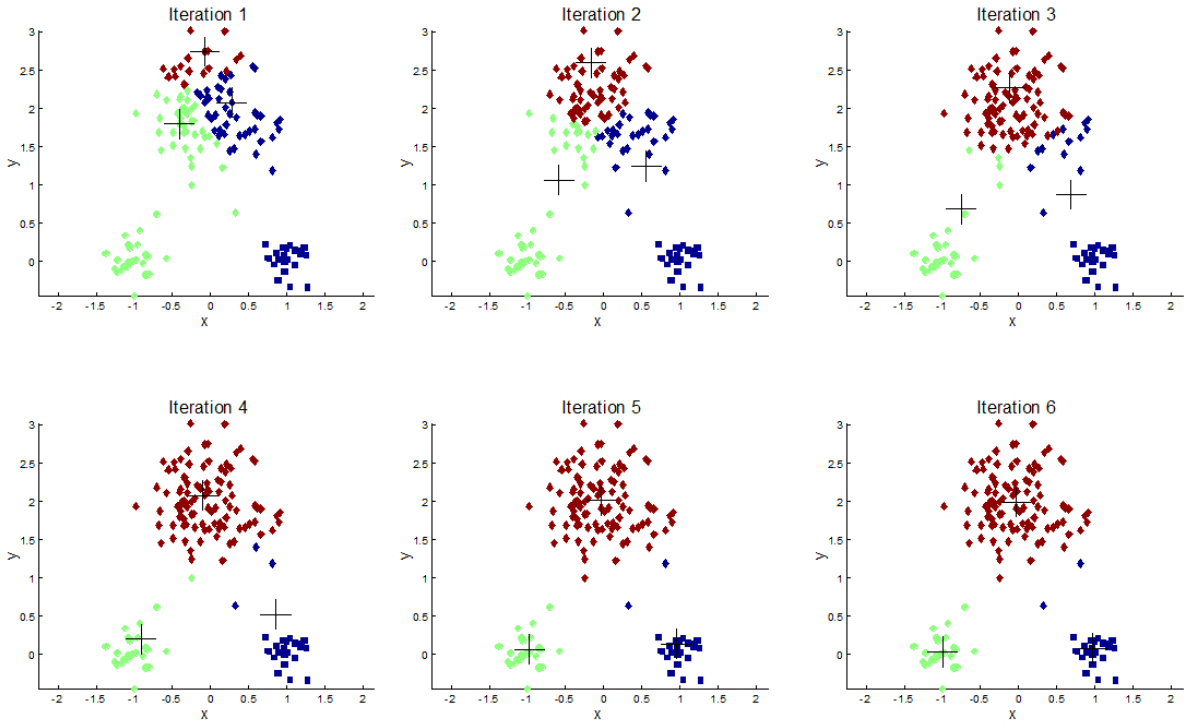
Logistics Regression

- Logistic regression is a binary classification model that predicts a binary.
- Widely used for **classification problems**
- Doesn't require linear relationship between dependent and independent variables.
- Requires **large sample sizes**
- The independent variables should not be correlated with each other i.e. **no multi collinearity**.
- If dependent variable is multi class then it is known as **Multinomial Logistic regression**.



K-means Clustering

- K means is an iterative clustering algorithm that aims to find local maxima in each iteration.
- This algorithm works in these 5 steps :
  1. Specify the desired number of clusters K
  2. Randomly assign each data point to a cluster :
  3. Compute cluster centroids
  4. Re-assign each point to the closest cluster centroid
  5. Re-compute cluster centroids
  6. Repeat steps 4 and 5 until no improvements are possible



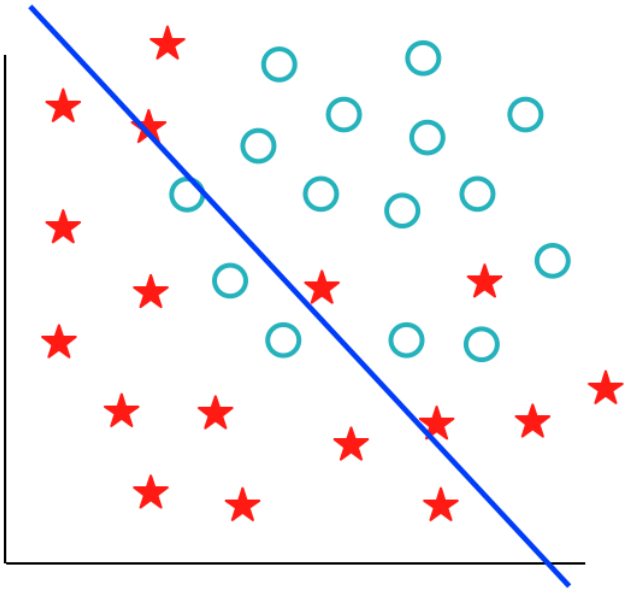
Choosing a model

- How do we choose a hypothesis space?
  - Often we use **prior knowledge** to guide this choice
- How can we gauge the accuracy of a hypothesis on unseen data?
  - **Occam's razor:** use the *simplest* hypothesis consistent with data!

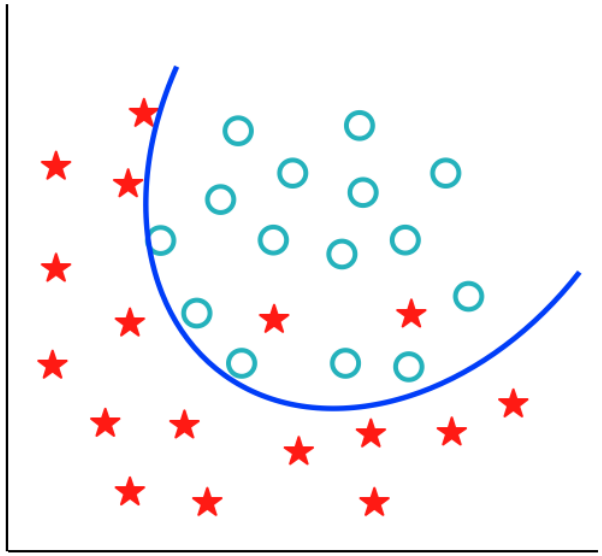
This will help us avoid overfitting.

  - ***Learning theory*** will help us quantify our ability to ***generalize*** as a function of the amount of training data and the hypothesis space
- How do we find the best hypothesis?
  - This is an **algorithmic** question, the main topic of computer science
- How to model applications as machine learning problems? (engineering challenge)

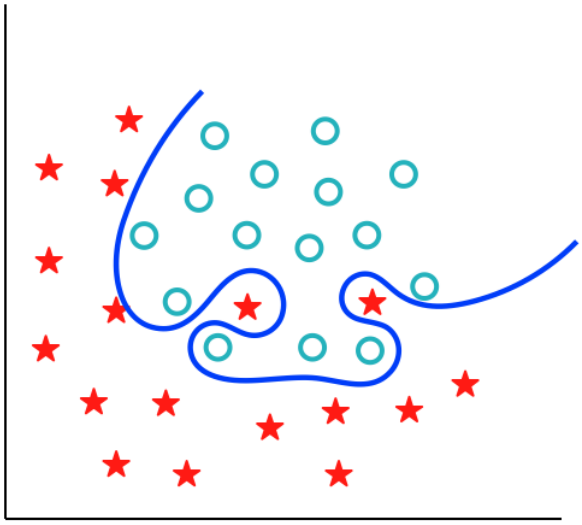
Problems in Machine Learning



Underfitting



Fit



Overfitting

Performance

- There are several factors affecting the performance:
  - **Types of training** provided
  - The form and extent of any initial **background knowledge**
  - The **type of feedback** provided
  - The **learning algorithms** used
- Two important factors:
  - Modeling
  - Optimization

Applications of AI & ML

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision, Face recognition
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - …
- This trend is accelerating
  - Big data
  - Improved machine learning algorithms
  - Faster computers
  - Good open-source software

THANK YOU!