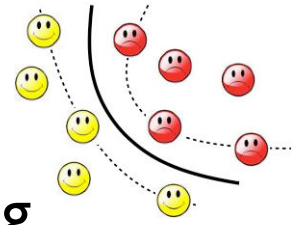
THE UNIVERSITY OF SYDNEY



Machine Learning and Data Mining

(COMP 5318)

School of Computer Science

Nguyen Hoang Tran

Team

THE UNIVERSITY OF SYDNEY

- Unit Coordinator and Lecturer:
  - Dr Nguyen Hoang Tran, room 428, J12
- Teaching Assistants: Canh Dinh, Zhiyi Wang
- Tutors

Format of the lectures

THE UNIVERSITY OF SYDNEY

- 10-15 min review from previous week
- 1h-1h30min of new content
- 5-10 min of examples

Assumed knowledge

THE UNIVERSITY OF SYDNEY

- Linear algebra, calculus
- Basics of probability theory
- Programming skills (Python)

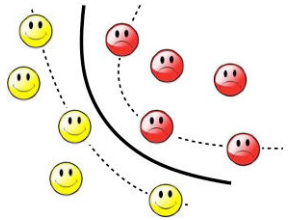
Labs: Python

- Python is a high-level programming language designed to enforce good coding practices.
- Interactive and very natural to use.
- Extremely versatile and excellent for prototyping.
- Great libraries for machine learning eg. scikit-learn, TensorFlow, Keras, Pytorch

www.python.org

Tutorial 1

- Check on canvas
- Introduction to Python
- Bring questions to your tutor next week



Let's talk about Machine Learning

What is Machine Learning?

Informally: Making predictions from data

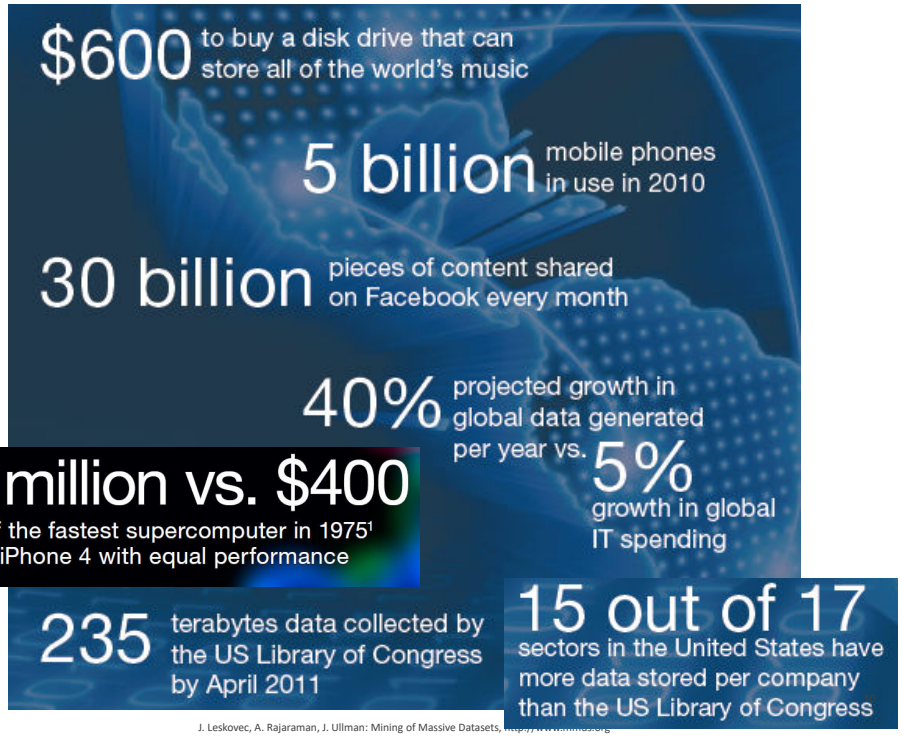
Formally: The construction of a statistical model that is an underlying distribution from which the data is drawn from, or using which we can classify the data into different categories..

ML / DM

- **But to extract the knowledge data needs to be**
  - **Stored**
  - **Managed**
  - **And ANALYSED ← this course**

Data Mining ≈ Big Data ≈ Statistics ≈ Machine Learning ≈ Data Science

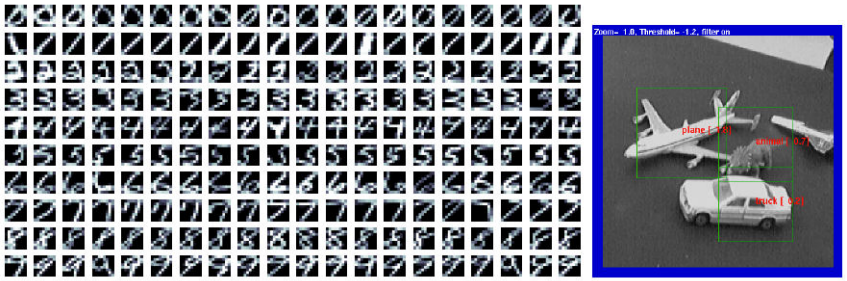
J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Speech recognition

Object and handwriting recognition



(NORB image from Yann LeCun)

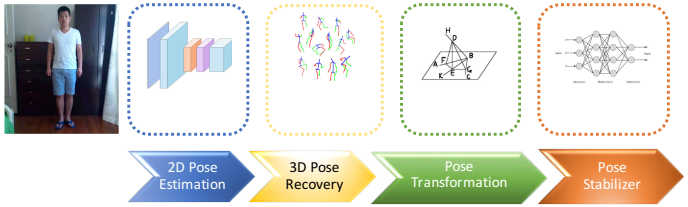
Information retrieval

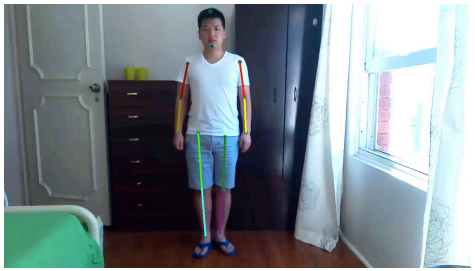
Web Pages

Retrieval
Categorisation
Clustering
Relations between pages

Financial prediction

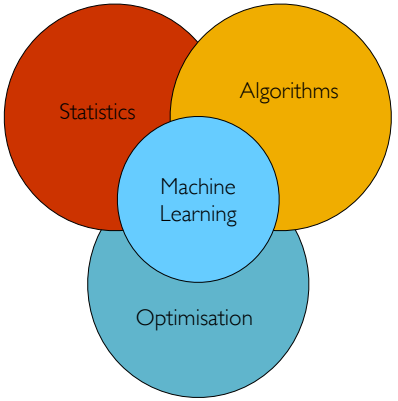
Robotics: pose estimation





This Course: COMP 5318

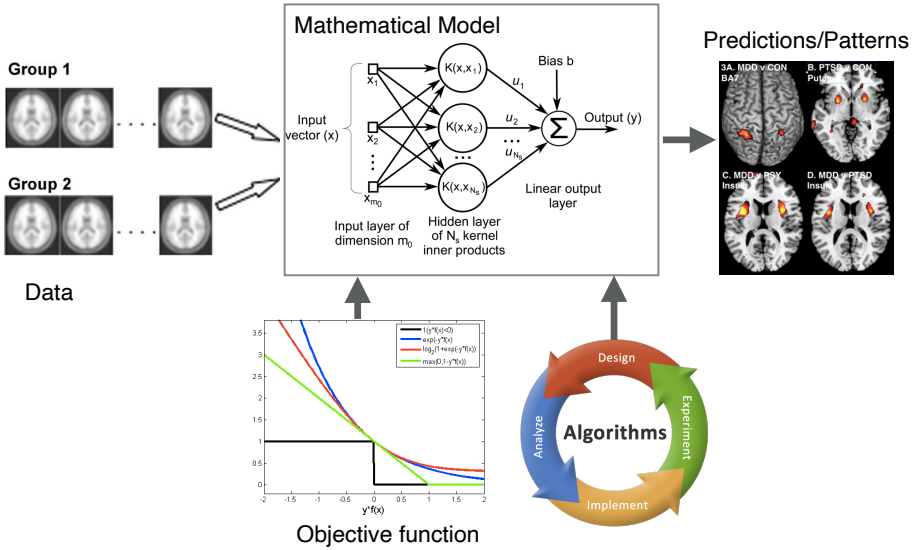
- This course overlaps with statistics, artificial intelligence, databases but more stress on
  - Algorithms
  - Mathematical modelling
  - Automation for handling large data



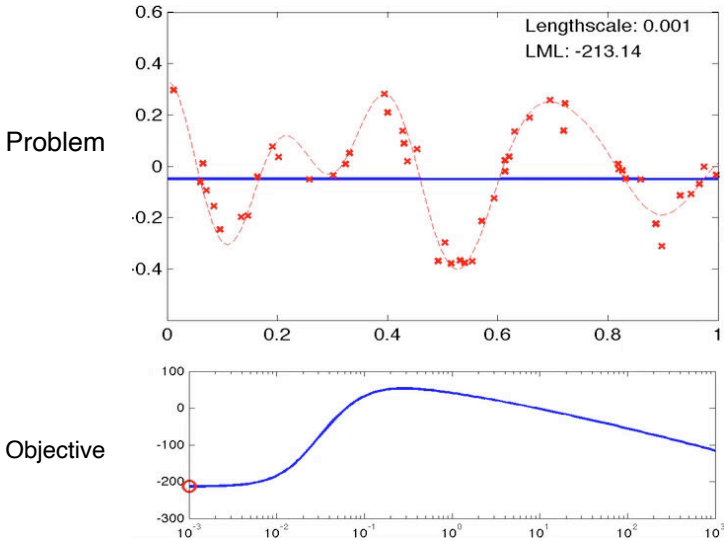
Machine Learning Problems

- Prediction
  - Classification and Regression
- Clustering, segmentation and summarisation
  - Find patterns in the data
- Outlier/anomaly detection
  - Find unusual patterns

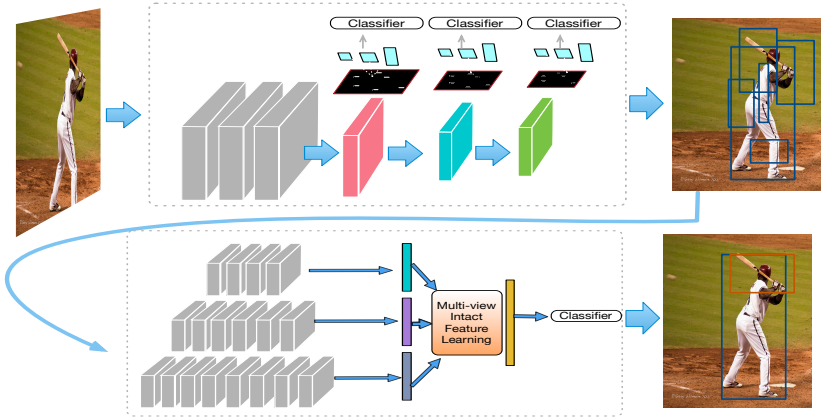
Elements of Machine Learning



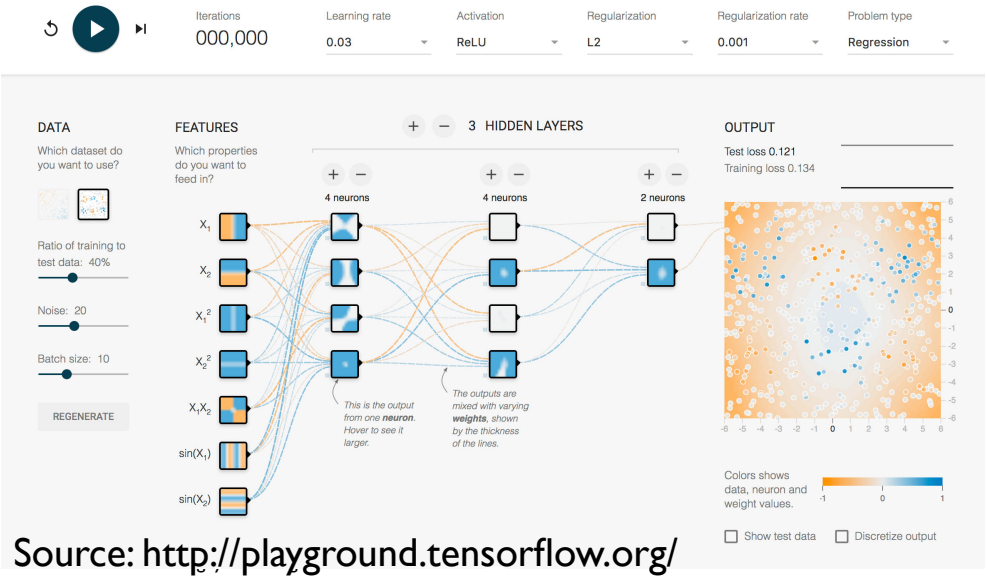
Regression



Classification for object detection

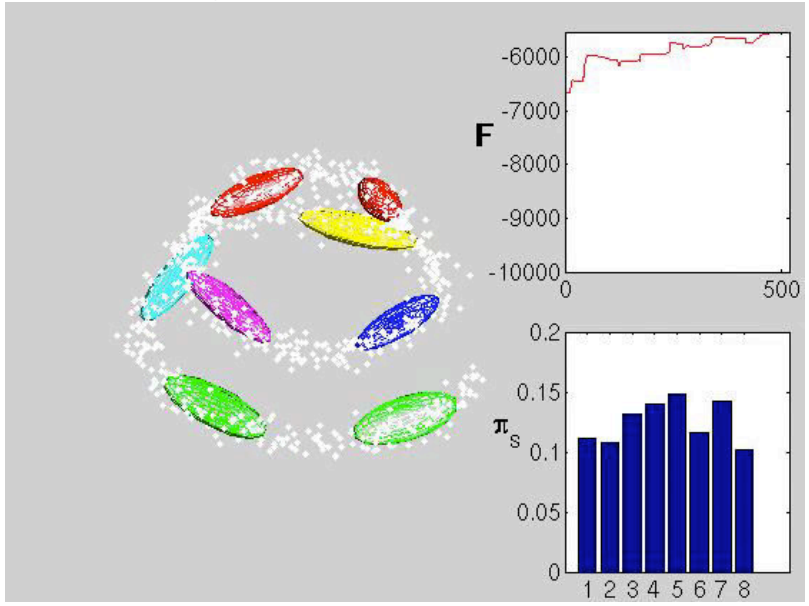


Neural Networks



Source: http://playground.tensorflow.org/

Clustering



Common representation

IMAGE/ VIDEO

TEXT/ COMMENT

TIME SERIES

SYSTEM LOGS

NETWORK

TABULAR/ RATING

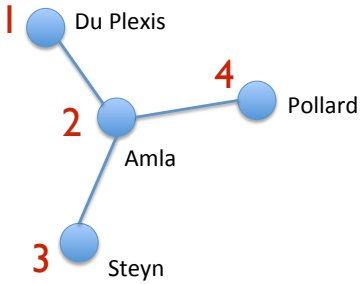
Is there a common way to represent data of different modalities ?

Text to matrix

- Document- Word Matrix
- Document 1: "AACCBBAAA"
- Document 2: "CCAABBDD"

$$\begin{bmatrix} A & B & C & D \\ 5 & 2 & 2 & 0 \\ 2 & 2 & 2 & 2 \end{bmatrix}$$

Network data



Nodes

Nodes $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$

Image data

700 x 500

4	45	6
6	12	33
22	17	44

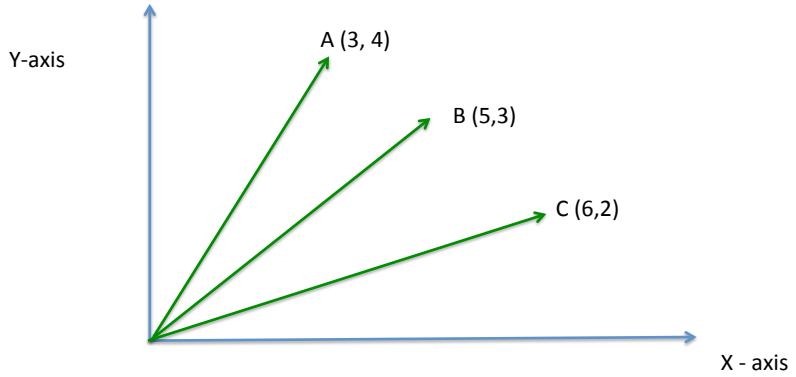
4	45	6	6	12	33	22	17	44

Similarity Computation

- We can now represent most data types as a matrix.
- A special case of a matrix is a vector.
- Now lets compute similarities with these objects.

Similarity Computation

How can we quantify similarity between A, B and C ?



Similarity Computation

- Dot product
  $x = (x_1, x_2, ..., x_n); \quad y = (y_1, y_2, ..., y_n);$
  $x.y = (x_1 y_1 + x_2 y_2 + ... + x_n y_n);$

- Norm (length) of a vector
  $\|x\| = (x.x)^{1/2} = (x_1.x_1 + x_2.x_2 + x_n.x_n)^{1/2}$

Similarity Computation

- The similarity between two vectors x and y is given by

  $sim(x, y) = x.y/(\|x\| \, \|y\|)$

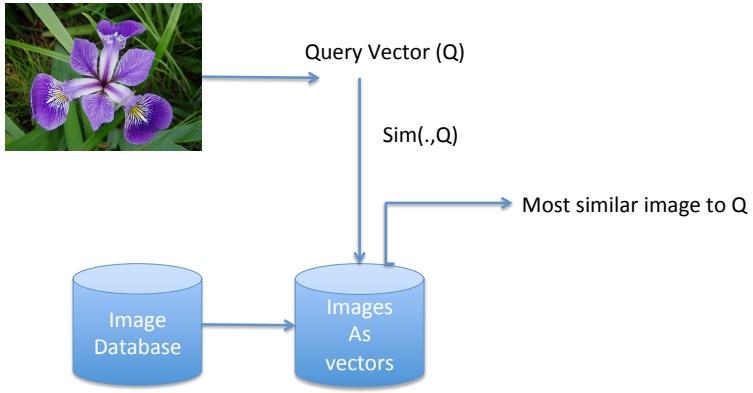
Example

- Let $x =< 3, 1, 2, 4 >$, $y =< 1, 2, 1, 2 >$
- Step 1: Compute the dot-product
  $x.y = 3.1 + 1.2 + 2.1 + 4.2 = 15$
- Step 2: Compute length of x vector
  $\|x\| = (3^2 + 1^2 + 2^2 + 4^2)^{0.5} = 5.477$
  $\|y\| = 3.162$
- $sim(x, y) = x.y/(\|x\| \, \|y\|) = 0.8660$

Properties

- When is $sim(x, y) = 0$?
- When is $sim(x, y) = 1$?
- Can $sim(x, y) < 0$?
- Can $sim(x, y) > 1$?

Image search engine



Object recognition

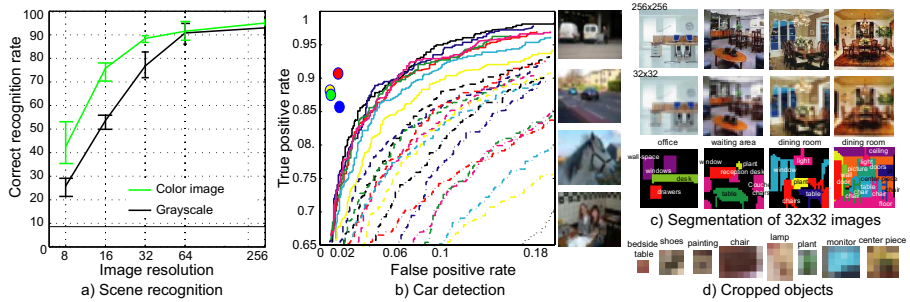
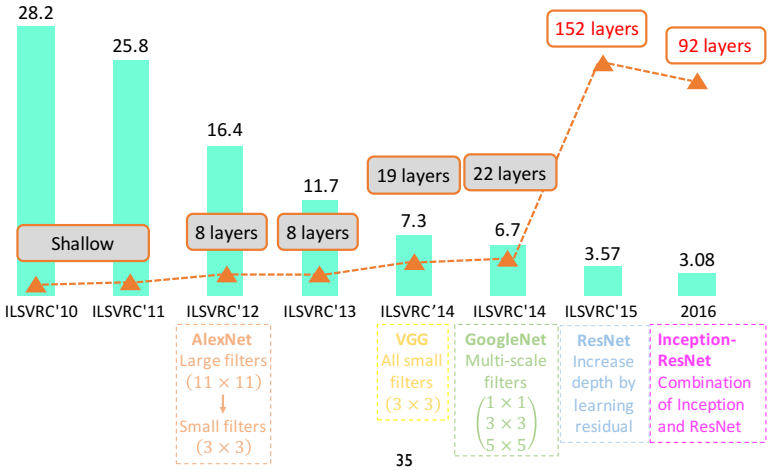


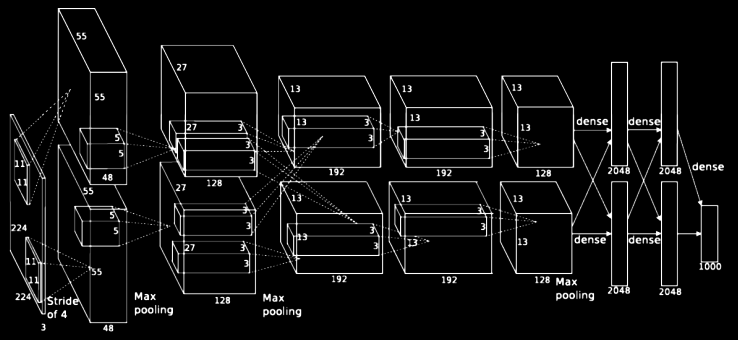
Fig. 1. a) Human performance on scene recognition as a function of resolution. The green and black curves show the performance on color and gray-scale images respectively. For color 32 × 32 images the performance only drops by 7% relative to full resolution, despite having 1/64th of the pixels. b) Car detection task on the PASCAL 2006 test dataset. The colored dots show the performance of four human subjects classifying tiny versions of the test data. The ROC curves of the best vision algorithms (running on full resolution images) are shown for comparison. All lie below the performance of humans on the tiny images, which rely on none of the high-resolution cues exploited by the computer vision algorithms. c) Humans can correctly recognize and segment objects at very low resolutions, even when the objects in isolation can not be recognized (d).

Torralba et al. 80 million tiny images: a large dataset for non-parametric object and scene recognition, PAMI 2008

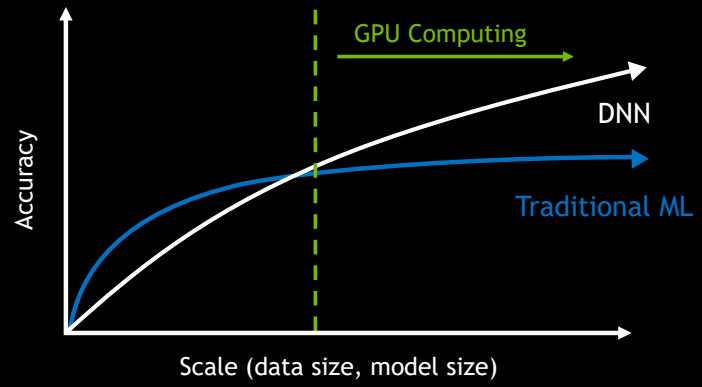
ImageNet: Over **15M** labeled high resolution images; Roughly **22K** categories.



Deep Convolutional Neural Networks

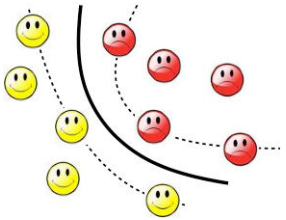


The rise of Deep Learning



https://blog.statsbot.co/deep-learning-achievements-4c563e034257

THE UNIVERSITY OF SYDNEY



Thanks!