CSC345/M45:
Big Data & Machine Learning
(intro & some fundamentals)

Dr. Xianghua Xie

x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
224 Computational Foundry, Bay Campus

Course Information

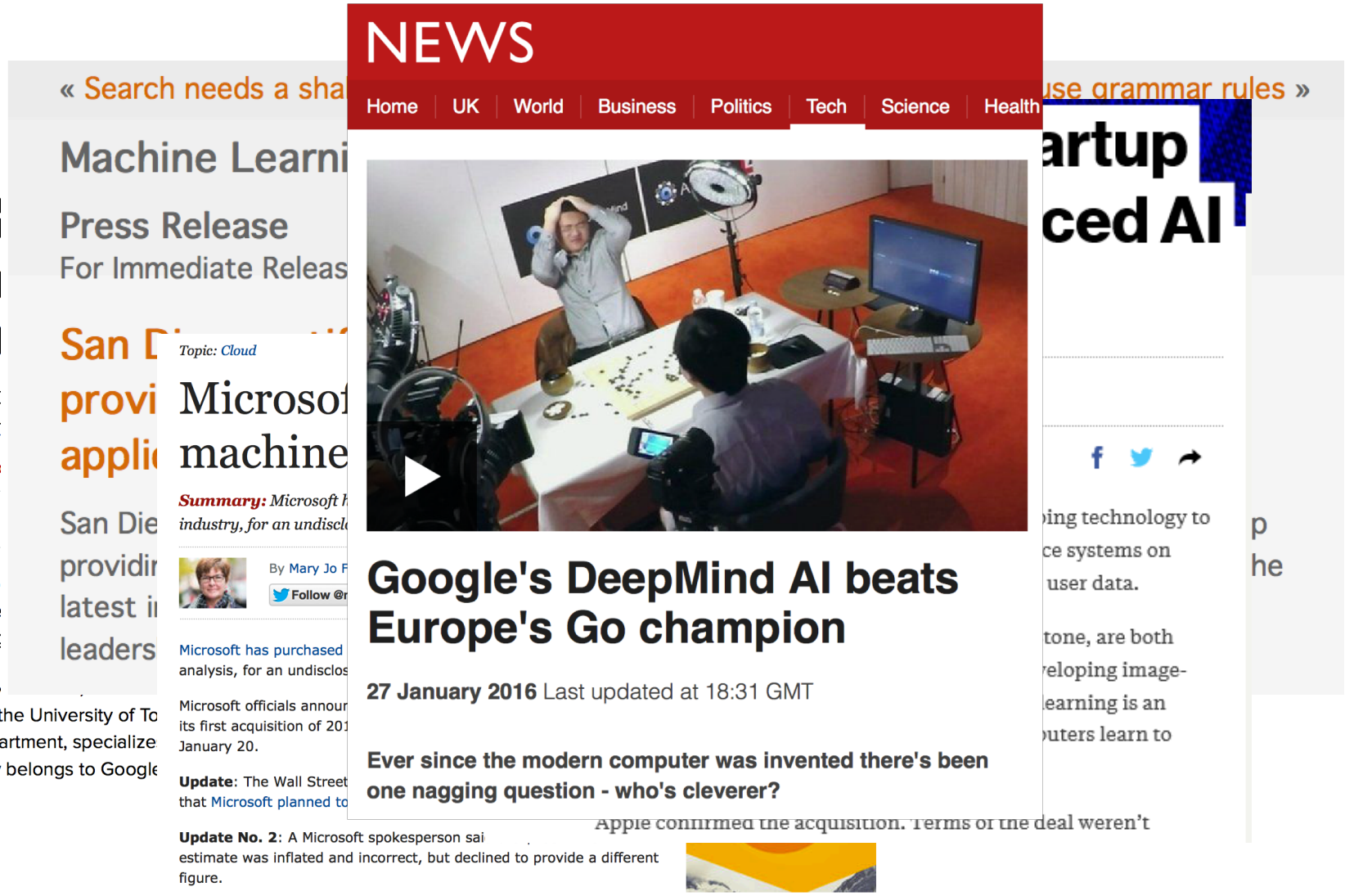
- Lectures
  - 3 lectures a week for 7 or 8 weeks; hence we will finish before the term ends
- Labs
  - 1 hour each week, supervised session
- Assessment
  - Lab work: 20% (**Upload** lab work to Blackboard & **Sign off** in lab classes)
  - Coursework: 20% (**Late submission: 0%)**
  - Exam: 60%
- Textbooks (recommended reading)
  - C. Bishop, Pattern Recognition and Machine Learning, Springer 2006
- Tutor: questions related to lab classes
  - Point of contact: Dr. Mike Edwards (Michael.Edwards@Swansea.ac.uk)
  - Gavin Tsang (658679@swansea.ac.uk) Michael Kenning (788486@Swansea.ac.uk)
  - Dave George (654214@swansea.ac.uk) Ali Alqahtani (884714@Swansea.ac.uk)

- Lecture notes handout:
  - Download your copy from Blackboard and print if required

- A few words about the lab sessions…

Research in the “wild”

Terminologies that have been widely used in the news, e.g. ”Deep Learning”.



Research in the “wild”

- Goal: Predict how a viewer will rate a movie

- 10% improvement = 1 million dollars

- Overarching theme:

Data + Learning

Big Data: definitions

- “Big Data” is not a scientific term

- extremely large data sets that may be analysed computationally to
  - reveal patterns, trends, and associations,
  - especially relating to human behaviour and interactions.

- a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include
  - analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy.

Big Data: definitions cont.

- an evolving term that describes any voluminous amount of
  - structured,
  - semi-structured
  - and unstructured data
- that has the potential to be mined for information.

- Characteristics:
  - the extreme volume of data (petabytes and exabytes),
  - the wide variety of types of data (can not be integrated easily)
  - and the velocity at which the data needs to processed.

- Data
- Information
- Knowledge

Does data speak for itself?

This module is concerned with Machine Learning techniques that are essential for analysing complex, heterogeneous datasets.

Is this module right for me?

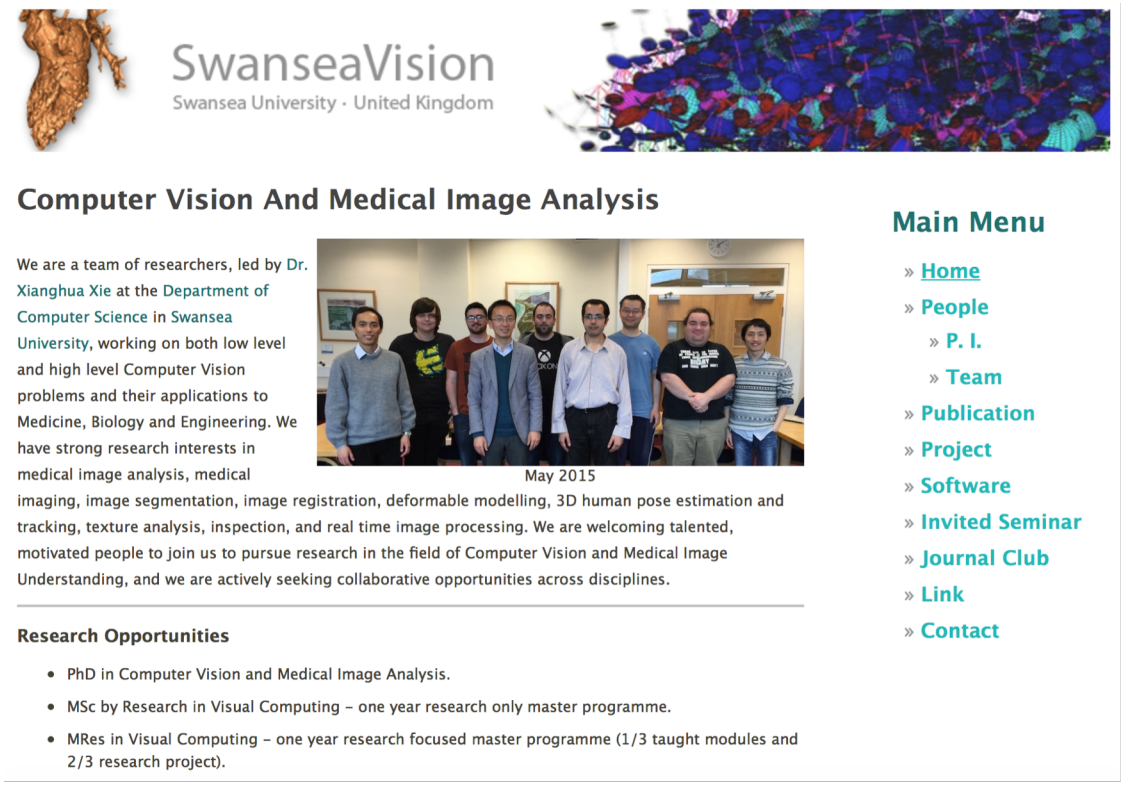
- Mathematics: lots of them!
  - willingness to learn;
  - topics include linear systems, calculus, probability
- Programming
  - You will learn Python (some introduction in lab but mostly self-taught)
    - Similar syntax to Java
    - Powerful and versatile; used extensively in both academia and industry
  - Lab classes: step by step examples and demos

- **All lecture notes are available NOW on Blackboard** (subject to minor changes): Scan to see if it is too mathemaical for you...

Module supplementary materials

- Everything will be on Blackboard

Module supplementary materials (tutorials and source code): http://csvision.swan.ac.uk/bdml.html (also on Blackboard)

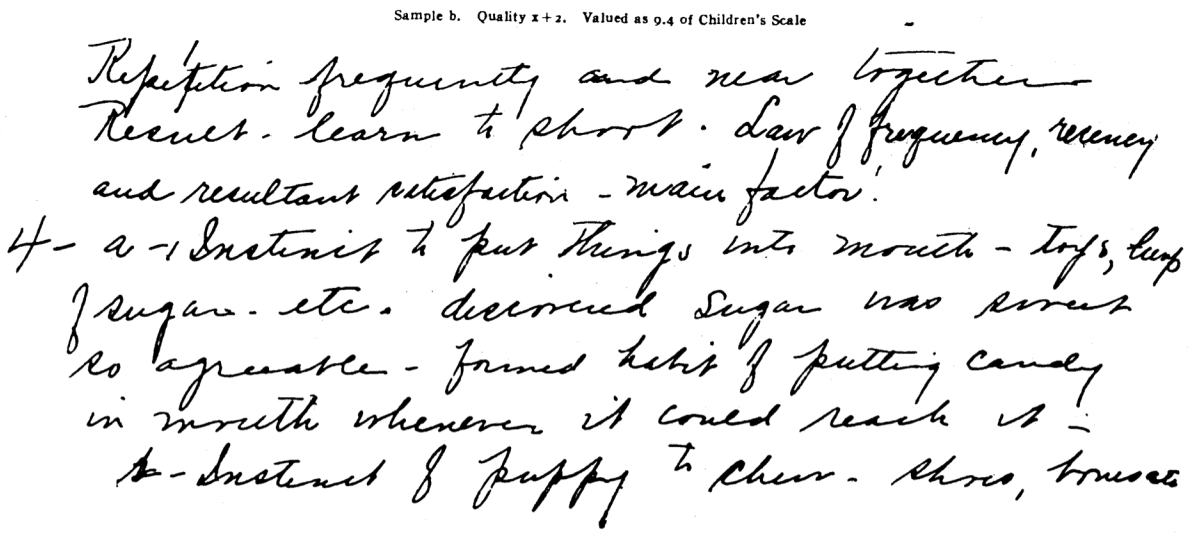


Some “Learning” Fundamentals

- **Attend lectures and lab sessions!**
- Be curious and Do Ask Questions
- Invest time in course work and lab work
- Please do NOT use the following in lectures (unless special needs)
  - Mobile
  - Tablet
  - Laptop

What is Machine Learning

- Example task: character recognition
  - A much harder problem: handwriting recognition (probably more useful)
  - A robust solution requires a large amount of data

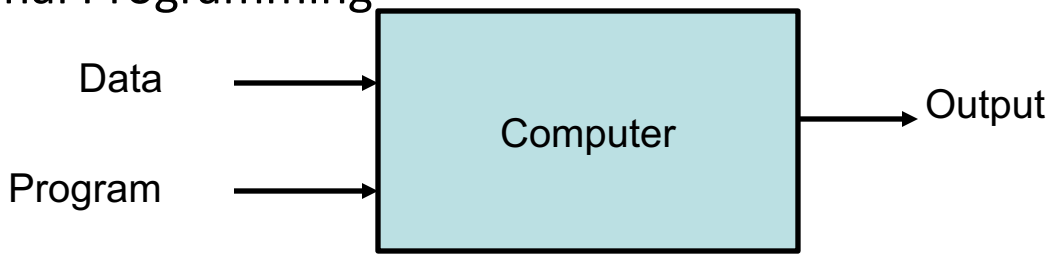


What is Machine Learning

- Why quantity of data is important, besides quality?

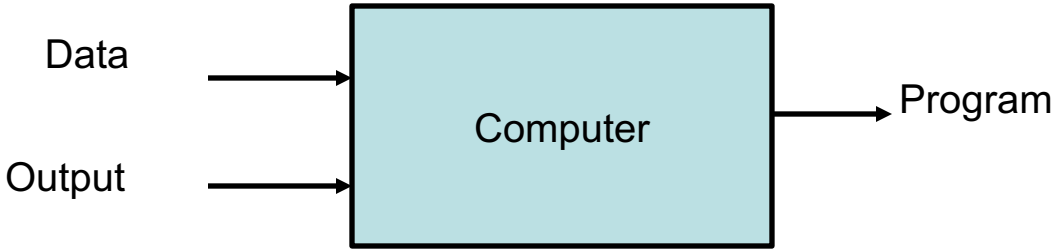
What is Machine Learning

- Traditional Programming



Example: Sorting numbers

- Machine Learning

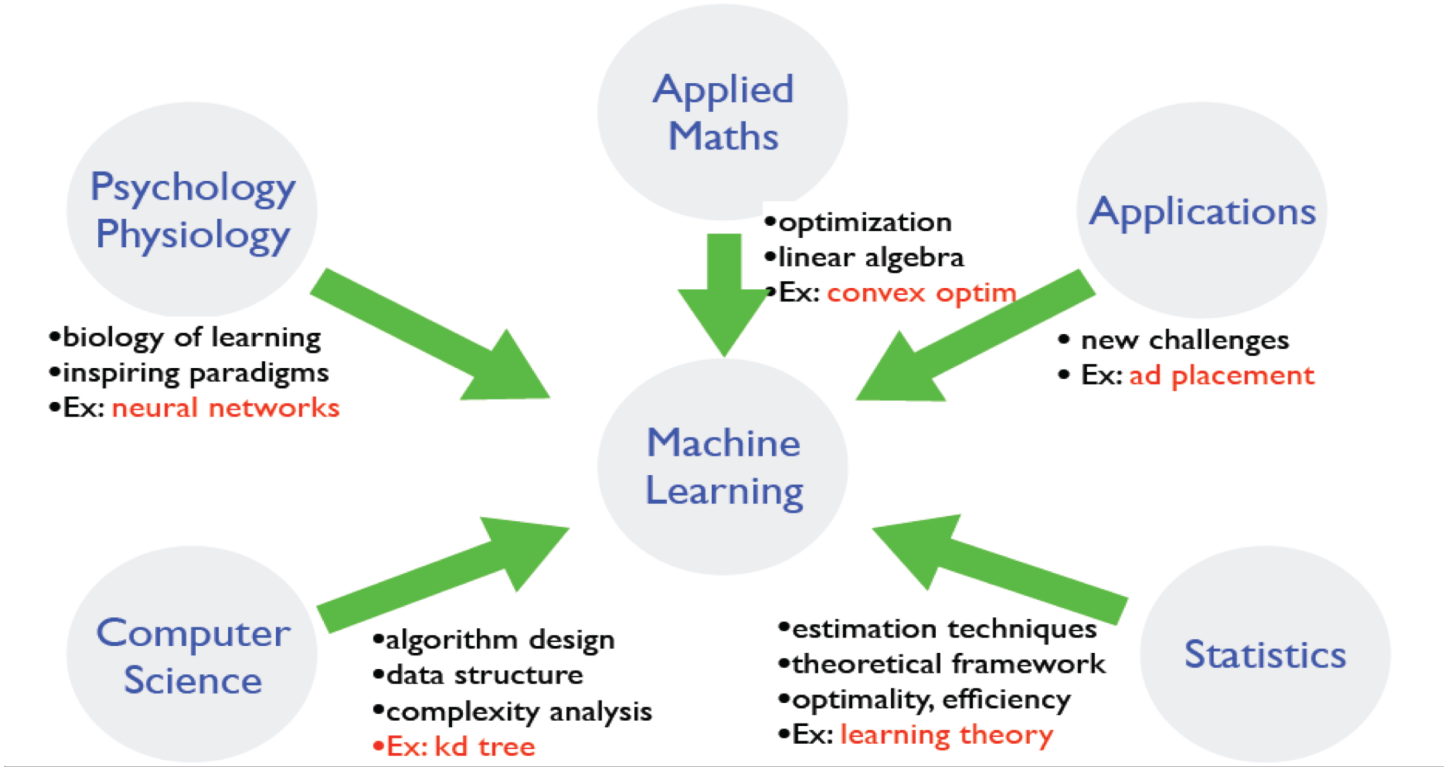


Machine Learning: definitions

- Learning
  - "the acquisition of knowledge or skills through experience, study, or by being taught."
- Machine Learning
  - [Arthur Samuel, 1959]
    - Field of study that gives computers the ability to learn without being explicitly programmed
  - [Kevin Murphy] algorithms that
    - automatically detect patterns in data
    - use the uncovered patterns to predict future data or other outcomes of interest
  - [Tom Mitchell] algorithms that
    - improve their performance (P)
    - at some task (T)
    - with experience (E)

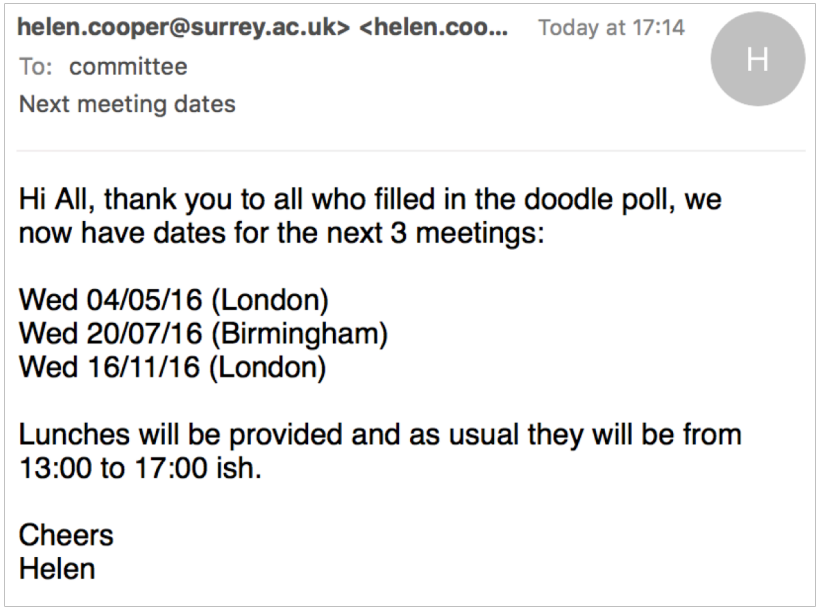
Machine Learning

- Where does Machine Learning fit in relevant subjects?



Machine Learning in a nutshell

- An oversimplified example: spam detection (classification)
  - Learn a mapping from input to output f: X -> Y
  - X: emails, Y: {Spam, NotSpam}



helen.cooper@surrey.ac.uk> <helen.coo... Today at 17:14

To: committee

Next meeting dates

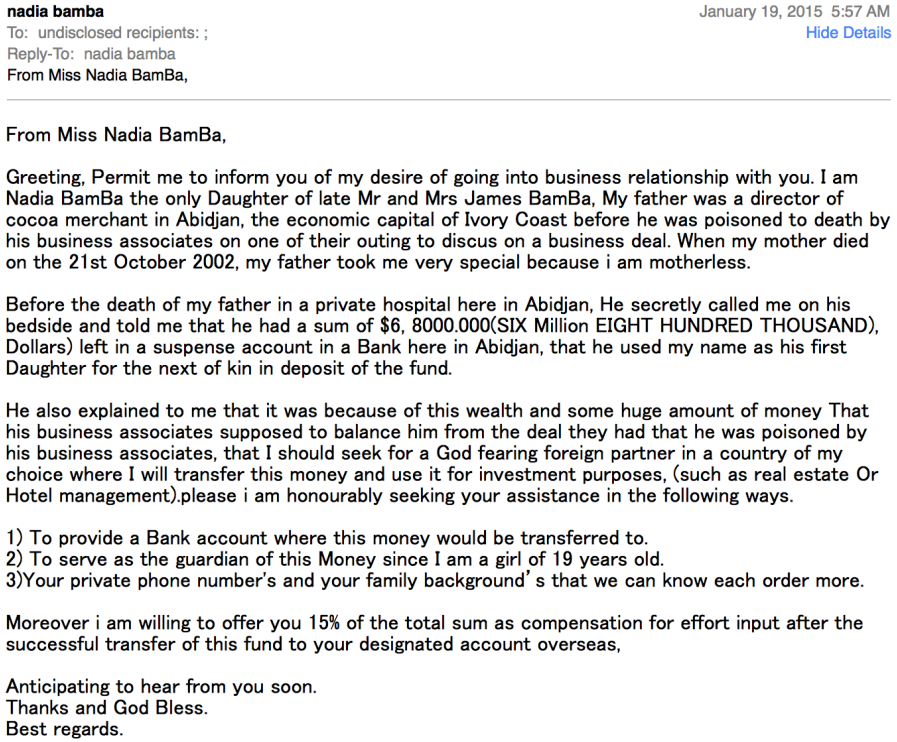
H

Hi All, thank you to all who filled in the doodle poll, we now have dates for the next 3 meetings:

Wed 04/05/16 (London)
Wed 20/07/16 (Birmingham)
Wed 16/11/16 (London)

Lunches will be provided and as usual they will be from 13:00 to 17:00 ish.

Cheers
Helen



nadia bamba January 19, 2015 5:57 AM

To: undisclosed recipients: ; Hide Details

Reply-To: nadia bamba

From Miss Nadia BamBa,

From Miss Nadia BamBa,

Greeting, Permit me to inform you of my desire of going into business relationship with you. I am Nadia BamBa the only Daughter of late Mr and Mrs James BamBa, My father was a director of cocoa merchant in Abidjan, the economic capital of Ivory Coast before he was poisoned to death by his business associates on one of their outing to discus on a business deal. When my mother died on the 21st October 2002, my father took me very special because i am motherless.

Before the death of my father in a private hospital here in Abidjan, He secretly called me on his bedside and told me that he had a sum of $6, 8000.000(SIX Million EIGHT HUNDRED THOUSAND), Dollars) left in a suspense account in a Bank here in Abidjan, that he used my name as his first Daughter for the next of kin in deposit of the fund.

He also explained to me that it was because of this wealth and some huge amount of money That his business associates supposed to balance him from the deal they had that he was poisoned by his business associates, that I should seek for a God fearing foreign partner in a country of my choice where I will transfer this money and use it for investment purposes, (such as real estate Or Hotel management).please i am honourably seeking your assistance in the following ways.

1) To provide a Bank account where this money would be transferred to.
2) To serve as the guardian of this Money since I am a girl of 19 years old.
3)Your private phone number's and your family background' s that we can know each order more.

Moreover i am willing to offer you 15% of the total sum as compensation for effort input after the successful transfer of this fund to your designated account overseas,

Anticipating to hear from you soon.
Thanks and God Bless.
Best regards.

Machine Learning in a nutshell

- An oversimplified example: spam detection (classification)
  - Intuition
    - Spam Emails
      - a lot of words like
        - » “money”
        - » “free”
        - » “bank account”
    - Regular Emails
      - word usage pattern is more spread out

Machine Learning in a nutshell

- An oversimplified example: spam detection (classification)
  - Let us simply count the keywords

Greeting, Permit me to inform you of my desire of going into business relations
Nadia BamBa the only Daughter of late Mr and Mrs James BamBa, My father w
cocoa merchant in Abidjan, the economic capital of Ivory Coast before he was
his business associates on one of their outing to discus on a business deal. Wh
on the 21st October 2002, my father took me very special because i am mothe

Before the death of my father in a private hospital here in Abidjan, He secretly
bedside and told me that he had a sum of $6, 8000.000(SIX Million EIGHT HUNI
Dollars) left in a suspense account in a Bank here in Abidjan, that he used my r
Daughter for the next of kin in deposit of the fund.

He also explained to me that it was because of this wealth and some huge amo
his business associates supposed to balance him from the deal they had that h
his business associates, that I should seek for a God fearing foreign partner in
choice where I will transfer this money and use it for investment purposes, (suc
Hotel management).please i am honourably seeking your assistance in the follo

1) To provide a Bank account where this money would be transferred to.
2) To serve as the guardian of this Money since I am a girl of 19 years old.
3)Your private phone number's and your family background' s that we can know

now have dates for the next 3 meetings:

Wed 04/05/16 (London)
Wed 20/07/16 (Birmingham)
Wed 16/11/16 (London)

Lunches will be provided and as usual they will be
13:00 to 17:00 ish.

X

$$\begin{pmatrix} \text{free} & 100 \\ \text{money} & 2 \\ \vdots & \vdots \\ \text{account} & 2 \\ \vdots & \vdots \end{pmatrix}$$

$$\begin{pmatrix} \text{free} & 1 \\ \text{money} & 1 \\ \vdots & \vdots \\ \text{account} & 2 \\ \vdots & \vdots \end{pmatrix}$$

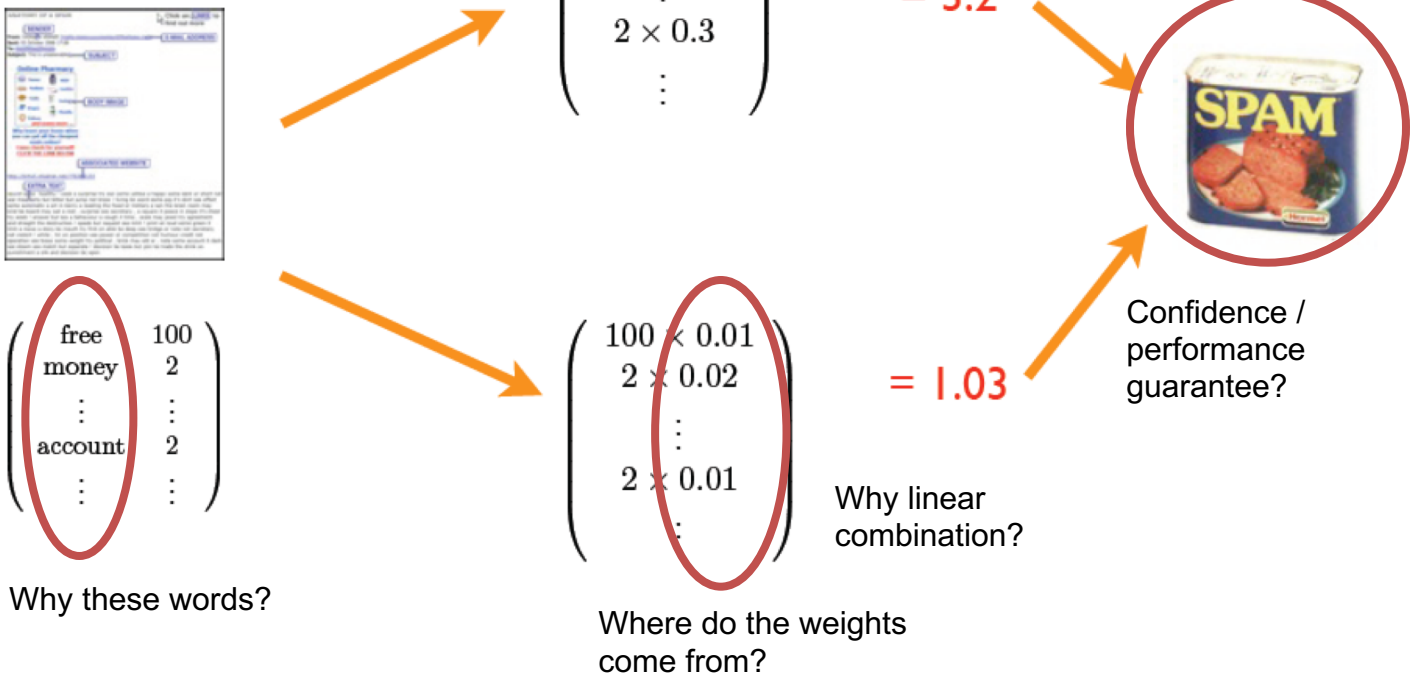
Y

Spam

NotSpam

Machine Learning in a nutshell

- An oversimplified example: spam detection (classification)



Typical Steps in Machine Learning

- Data collection
  - "training data", optionally with "labels"
- Representation
  - how the data are encoded into "features" when presented to a learning algorithm.
- Modelling
  - choose the class of models for the learning algorithm
- Estimation
  - find the model that best explains the data
- Validation
  - evaluate the learned model and compare to solutions found using other model classes.
- Apply learned model to new "test" data

Types of Learning

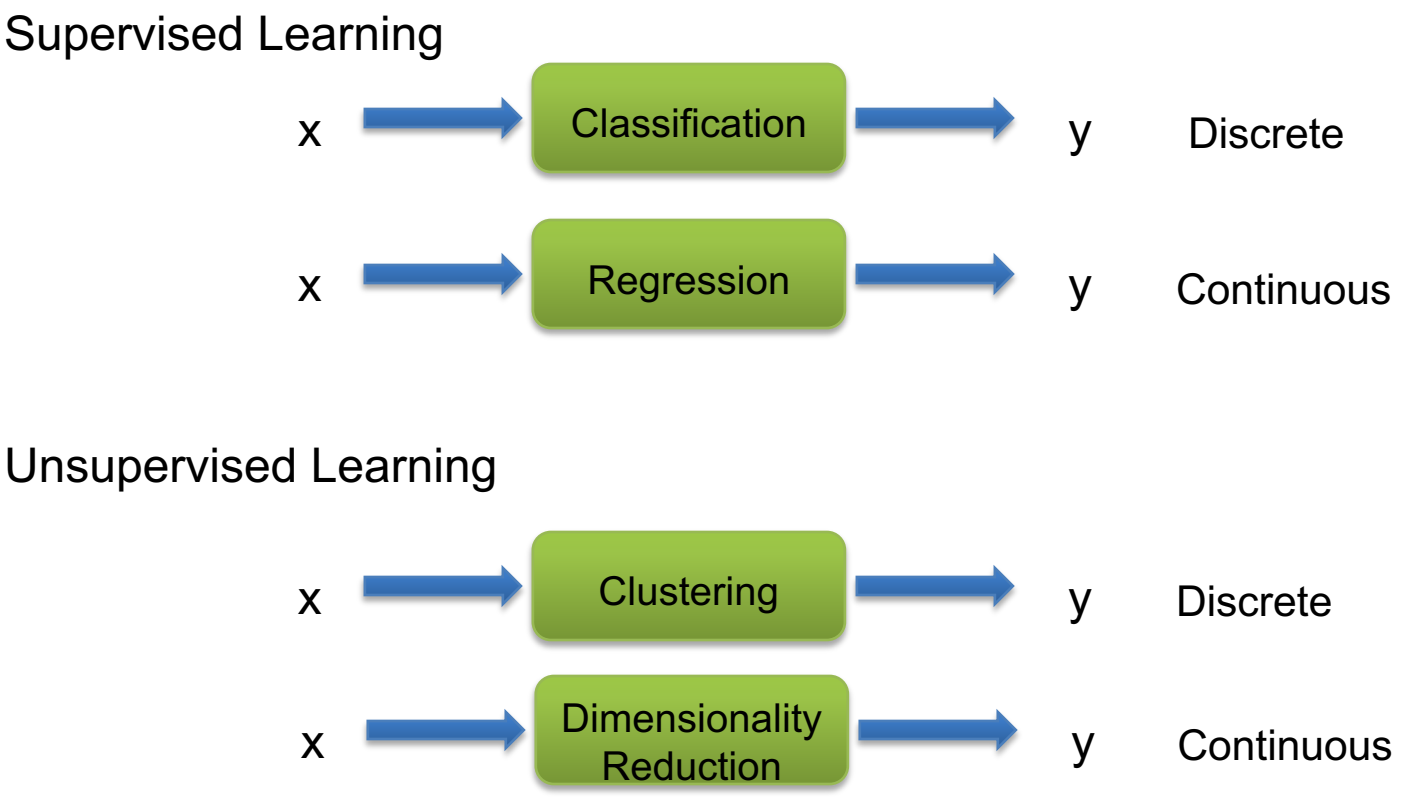
- Supervised learning
  - Training data includes desired outputs
    - Training data provided as pairs (x,y)
    - The goal is to predict an “output” y from an “input” x
    - Output y for each input x is the “supervision” that is given to the learning algorithm.
    - Often obtained by manual “annotation” of the input x
    - Can be costly to do
    - Most common examples
      - Classification
      - Regression

Types of Learning cont.

- Unsupervised learning
  - Training data does not include desired outputs
- Weakly or Semi-supervised learning
  - Training data includes a few desired outputs
- Reinforcement learning
  - Rewards from sequence of actions
  - Example:
    - There is only one “supervised” signal at the end of the game.
    - But you need to make a move at every step
    - RL deals with “credit assignment”

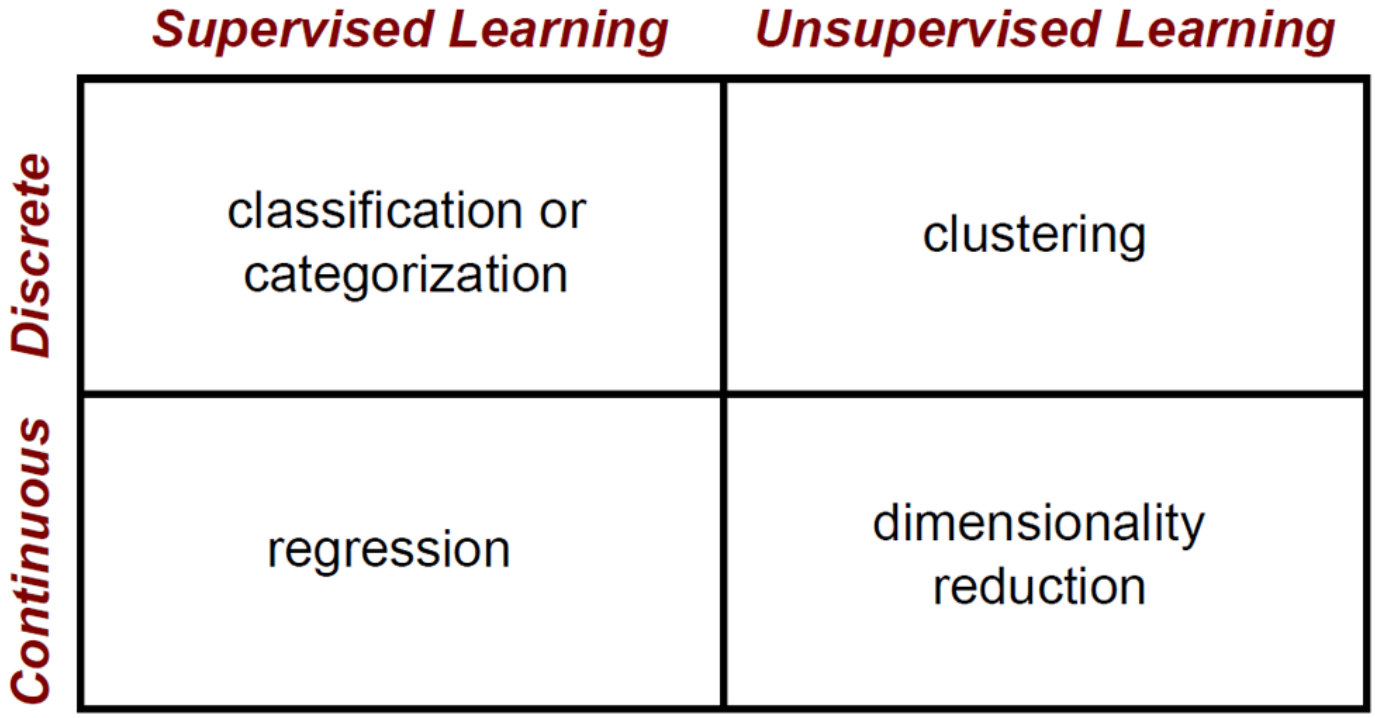
Machine Learning

- Important application areas



Machine Learning

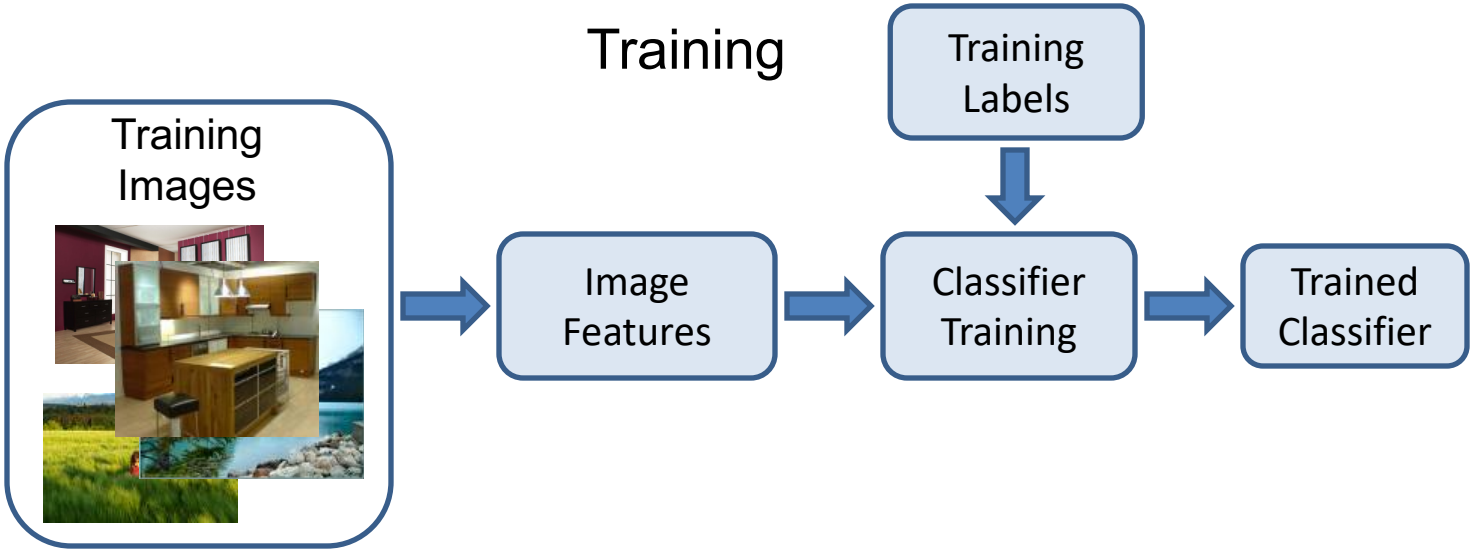
- Important application areas



Note, this is a very crude classification! Lots of techniques cover multiple domains. E.g. regression can be discrete as well.

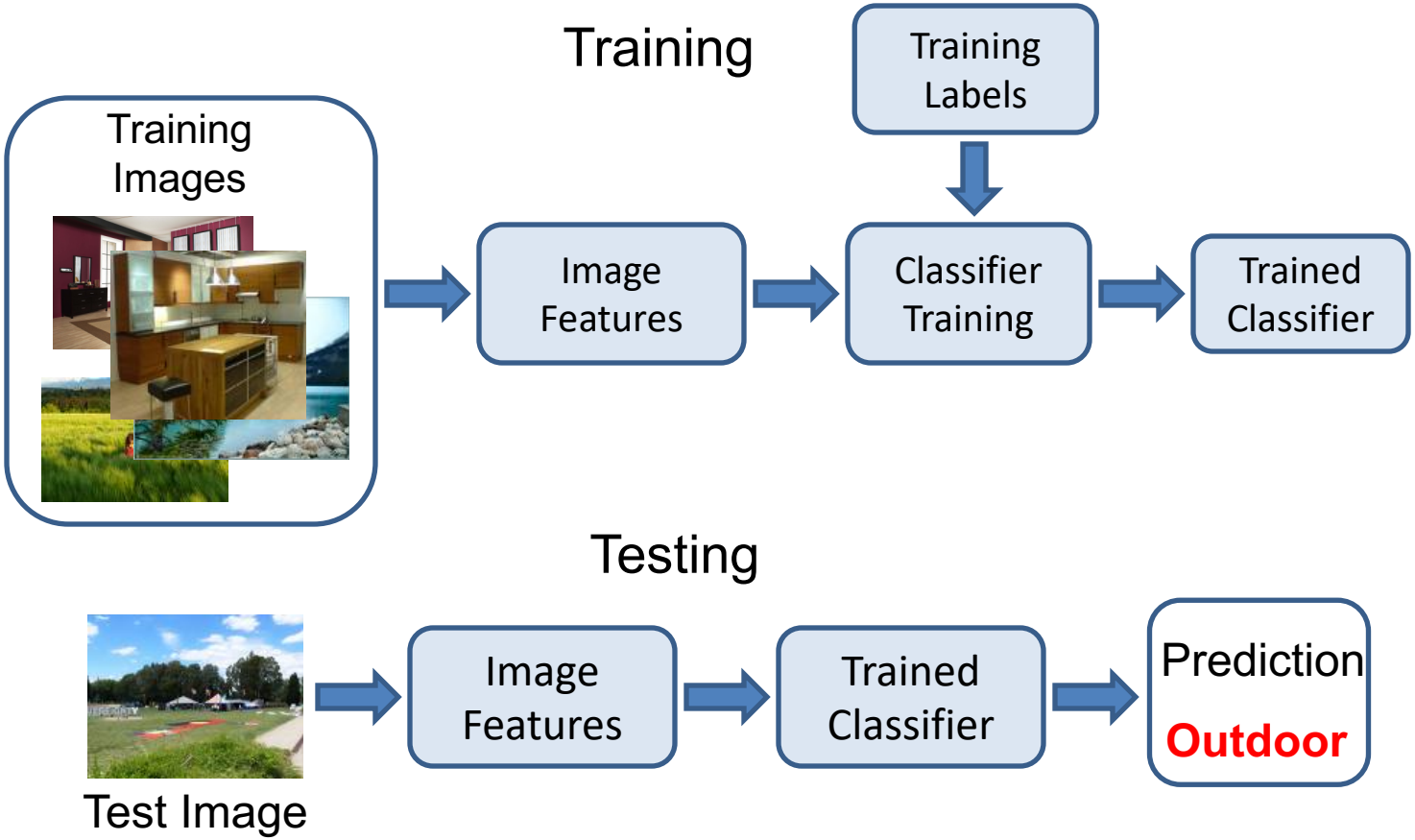
Classification: example

- Image classification



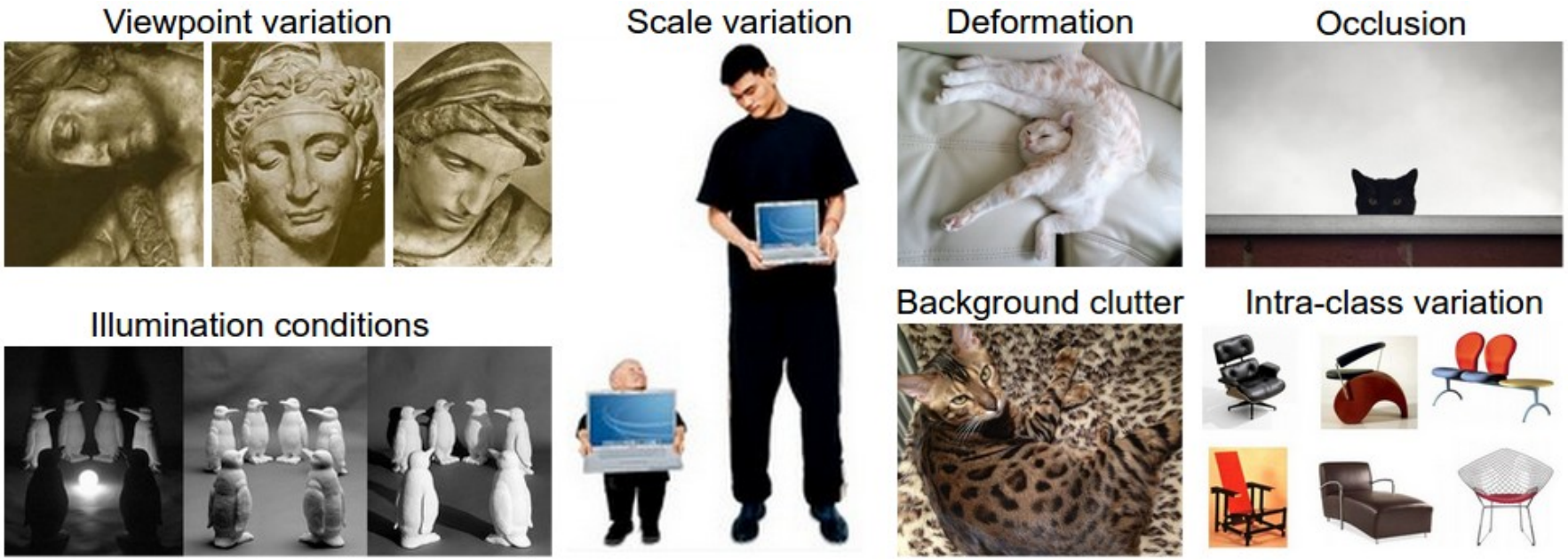
Classification: example

- Image classification



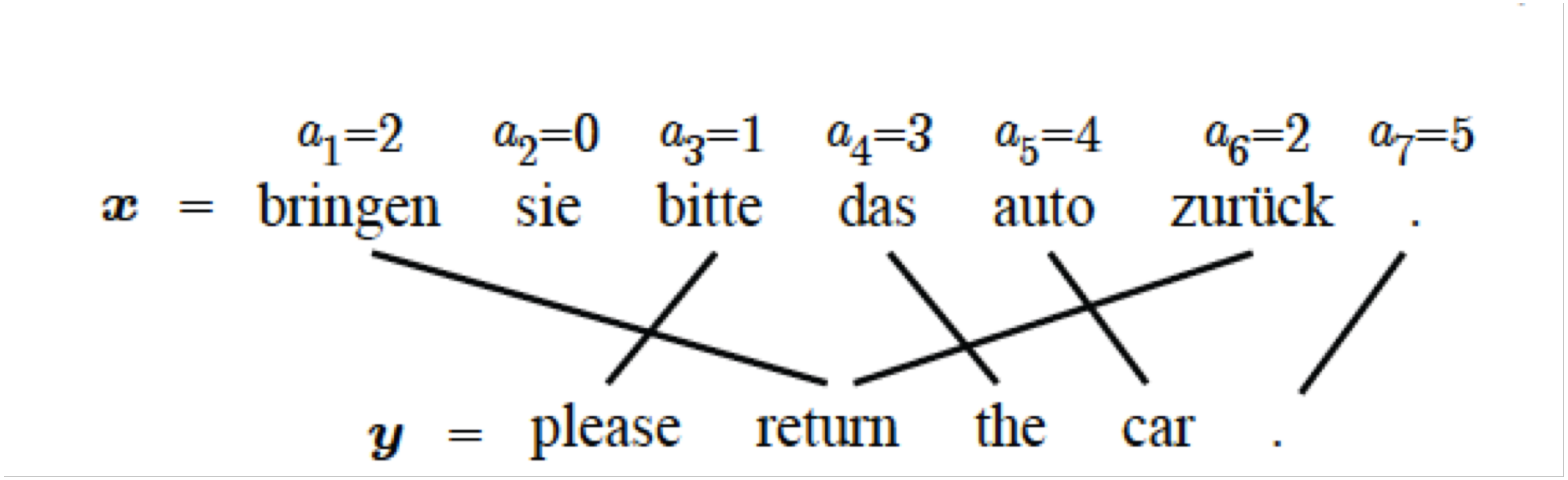
Classification: example

- Image classification
  - Not an easy task!



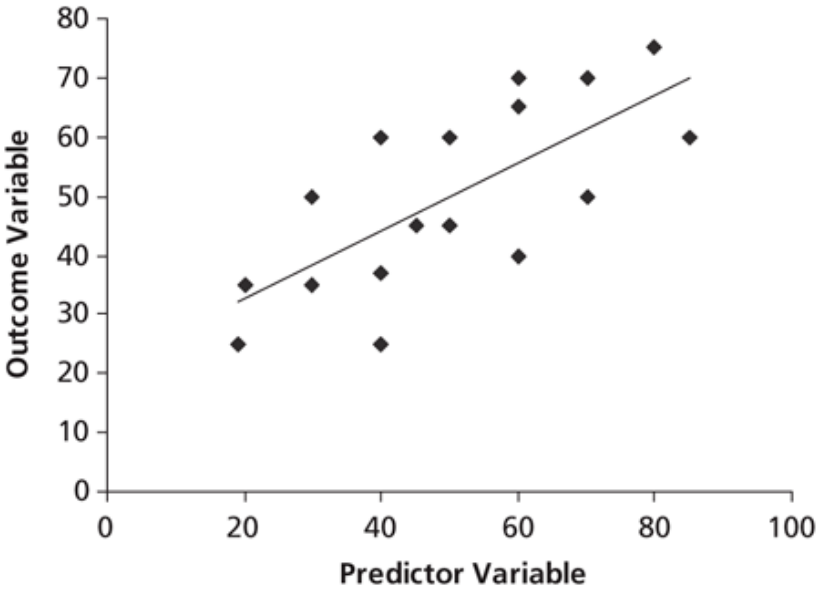
Classification: example

- Machine translation

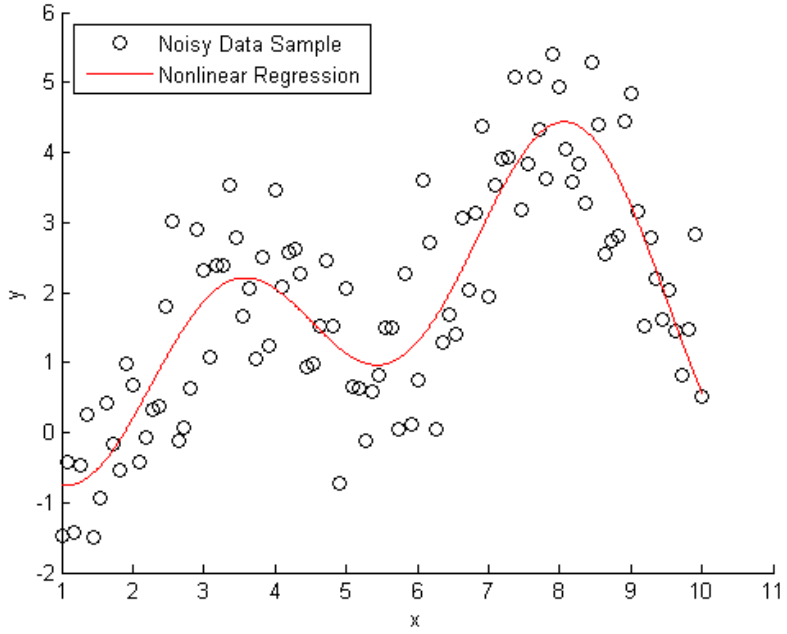


Regression (supervised)

- Similar to classification, but output y has the form of one or more real numbers.
- Goal is to predict for input x an output f(x) that is close to the true y.
- Learn a continuous function



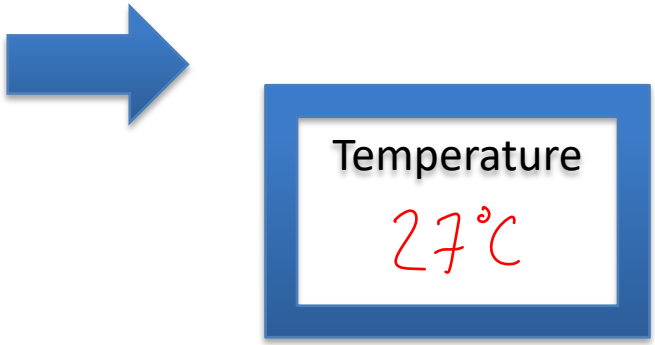
Linear regression



Nonlinear regression

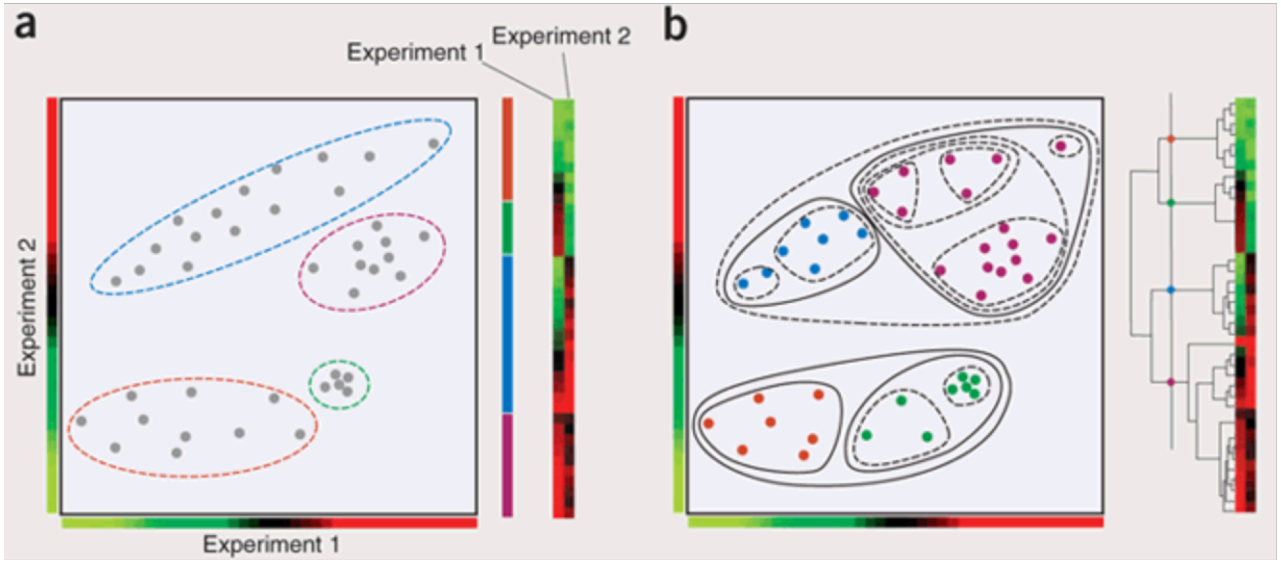
Regression (supervised)

- Example applications



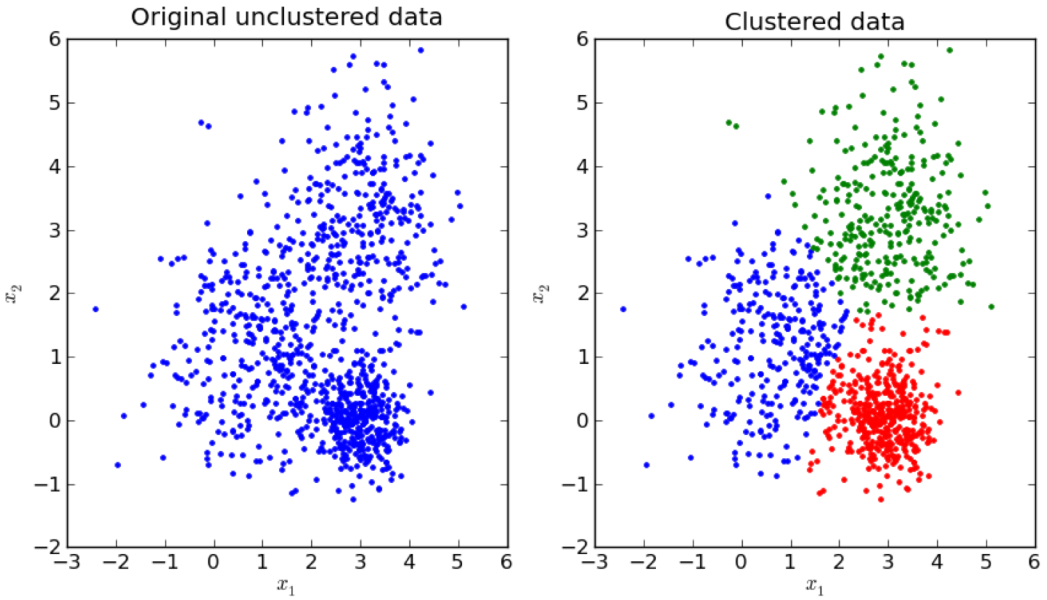
Clustering (unsupervised)

- Clustering: group together *similar* items
  - Finding a group structure in the data
  - Data in one cluster similar to each other
  - Data in different clusters dissimilar
  - To define similarity is not always trivial
- Effect: Map each data point to a discrete cluster index
  - "flat" methods find k groups (k known, or automatically set)
  - "hierarchical" methods define a tree structure over the data

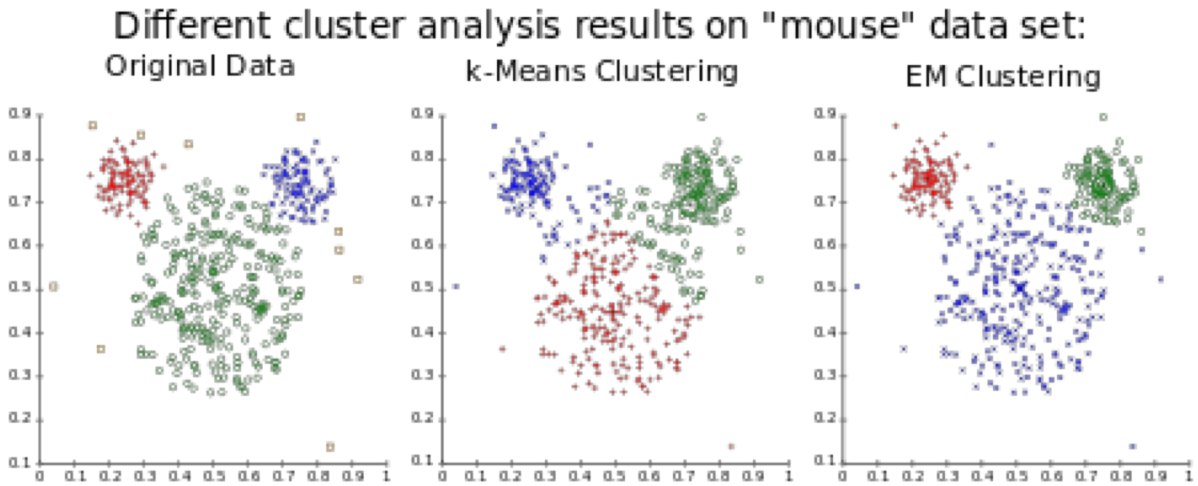


Clustering (unsupervised)

- Examples

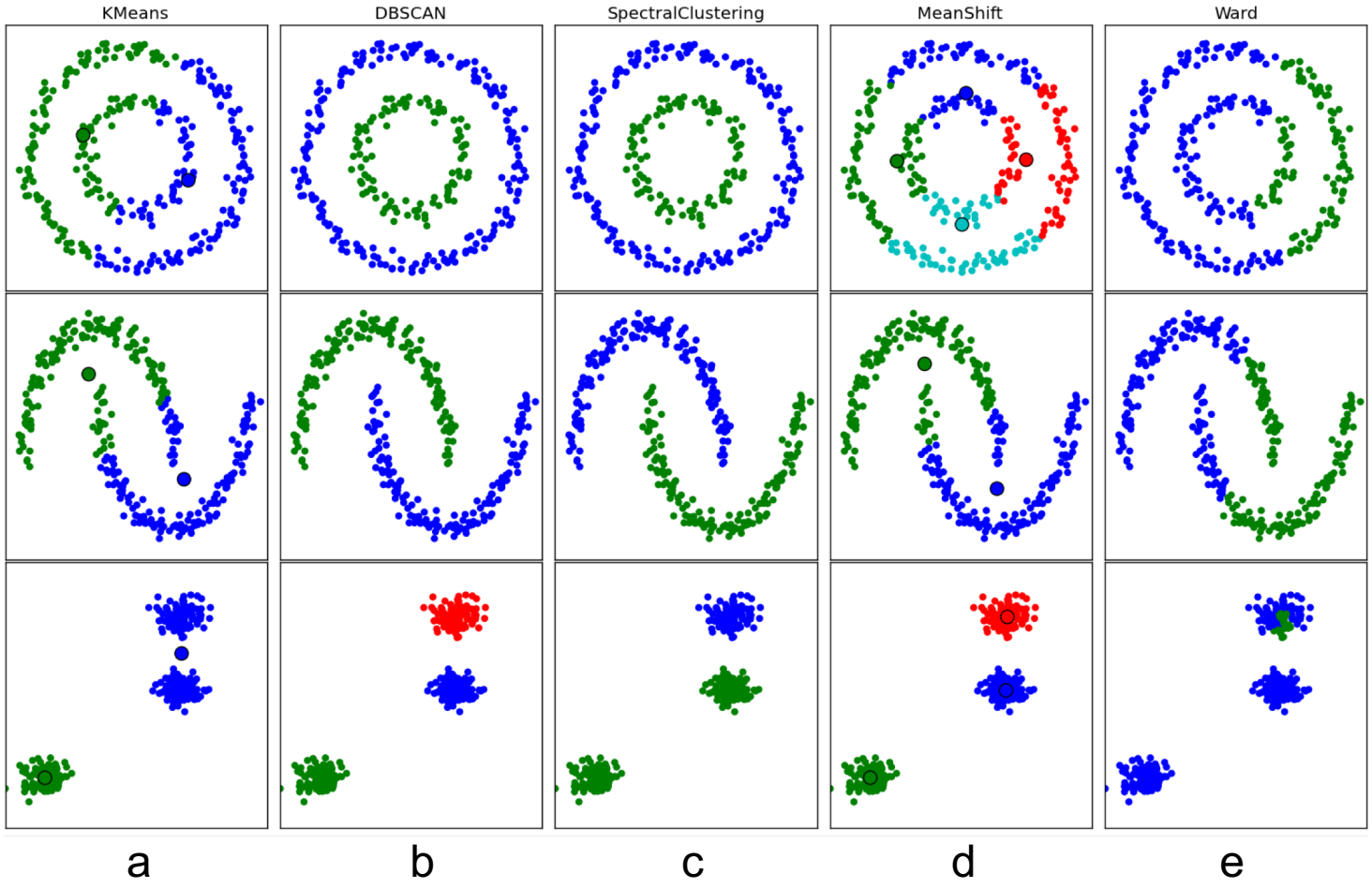


Which one is better?



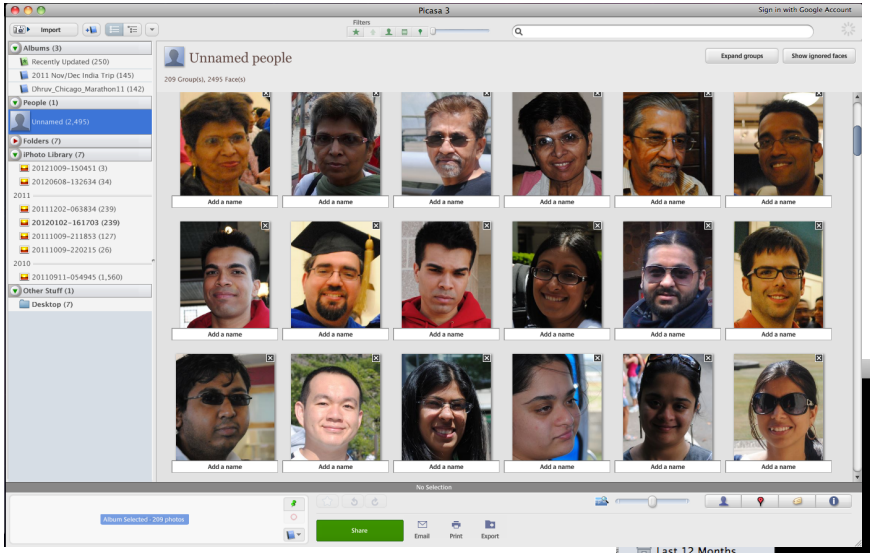
Clustering (unsupervised)

- Examples: evaluate the quality of clustering



Clustering (unsupervised)

- Examples



Picassa

iPhoto

Clustering

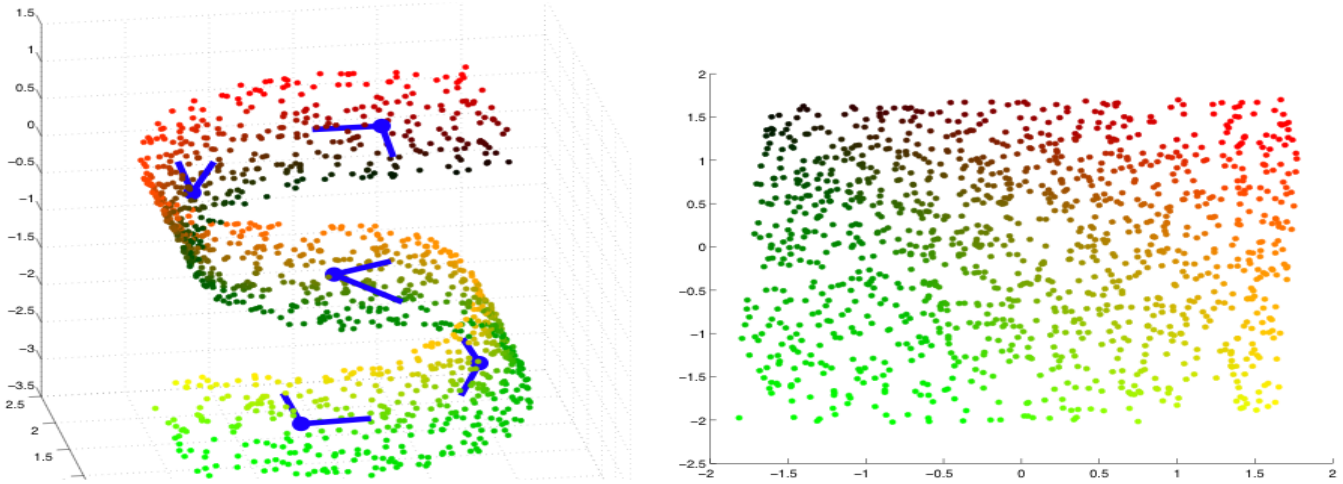
- Learn face similarity from training pairs labelled as same/different
- Cluster faces based on identity

[Guillaumin, Verbeek, Schmid, ICCV 2009]

Dimensionality Reduction (unsupervised)

- Finding a lower dimensional representation of the data
  - Useful for compression, visualization, noise reduction

- Unlike regression: target values not given, i.e. unsupervised



Syllabus

- Clustering
- Regression
- Dimensionality Reduction
- Linear Discriminant Analysis
- Logistic Regression
- Support Vector Machine
- Neural Network