Introduction to Data Science

@cambridgespark

CAMBRIDGE SPARK

How our trainings work...

- Hands on approach (teach-code…)
- Work in pairs
- Ask us a lot of questions

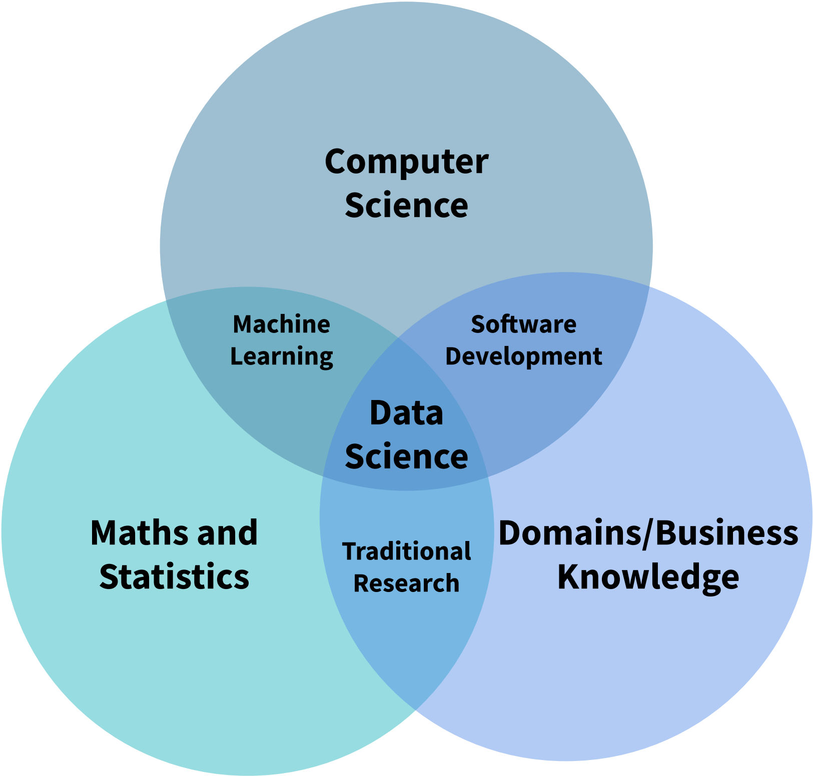
Data Scientist

Harvard
Business
Review

“The sexiest job of the 21st century.”

(T. Davenport, D.J. Patil, HBR)

Data Scientist: a jack of all trade



CAMBRIDGE SPARK

What is Machine Learning?

CAMBRIDGE SPARK

Paul, the psychic octopus

84.6% correct predictions during the 2010 FIFA World Cup

Understanding the components of learning

How do you know it’s a cat?

Learning by experience

- Algorithms are training to recognise **patterns** in data
- Once trained, they can be used on **new** data
  - To categorise data
  - To automatise decision taking
  - ...

Learning by experience

- Algorithms are training to recognise **patterns** in data
- Once trained, they can be used on **new** data
  - To categorise data
  - To automatise decision taking
  - …

Key element: **DATA**

Big Data, a big buzzword

Intuition: Big Data = anything that will break Excel

Further:

- **Velocity**: data pulled from rapid stream
- **Volume**: data stored on multiple machines
- **Variety**: different format, structured, unstructured, ..

Mostly engineering issues

Big Data, a word of warning

Big Data does not mean big insight and can mean big mistakes:

- Is it really **representative**? (e.g.: Polls, Trends on Google, …)
- The uncertainty about your decisions will decrease with the size of the data, but it can still be **biased**
- Remain cautious

Two broad categories of learning

Supervised Learning

- Data: (input point, response)
- Aim:
  - Assign response to new points

Example: Recommender system

Unsupervised Learning

- Data: (input point)
- Aim:
  - Group similar input points

Example: Market segmentation

Quiz

Supervised vs Unsupervised Learning

CAMBRIDGE SPARK

Supervised or Unsupervised?

- Face recognition

Supervised or Unsupervised?

- Face recognition
- Spam/Fraud detection

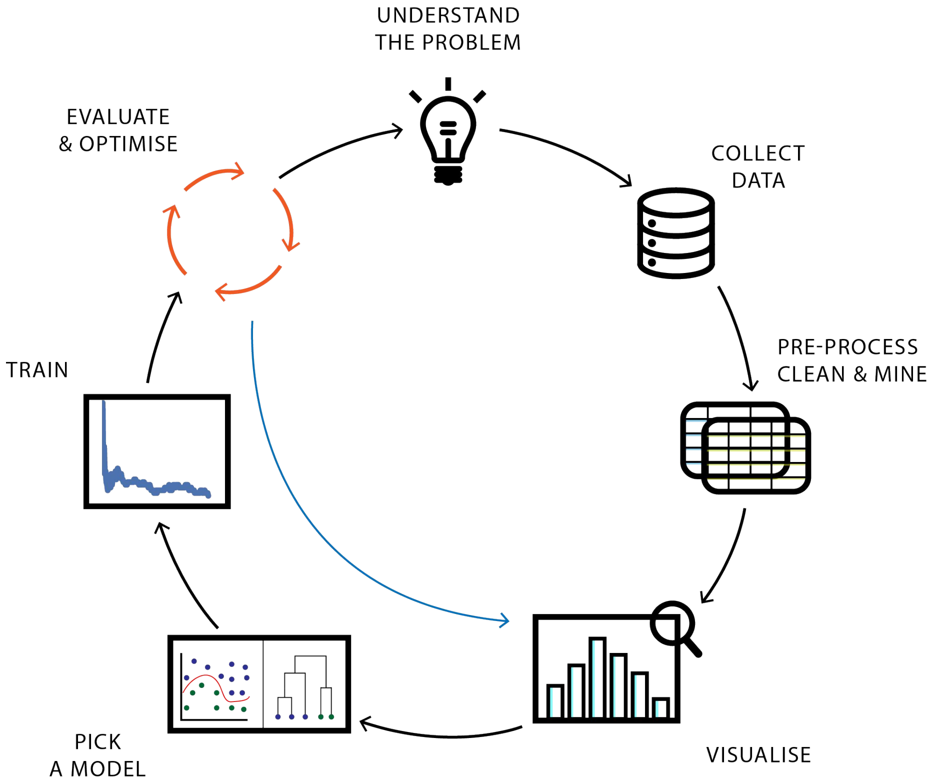
Supervised or Unsupervised?

- Face recognition
- Spam/Fraud detection
- Automatic tabs in GMail

Supervised or Unsupervised?

- Face recognition
- Spam/Fraud detection
- Automatic tabs in GMail
- Online advertisement

Data Science Pipeline



CAMBRIDGE SPARK