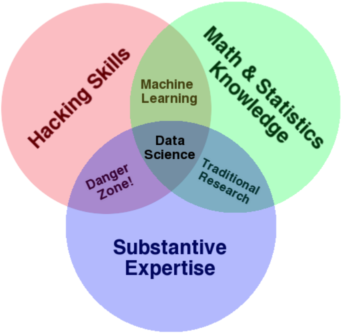
Introduction to Introduction to Data Science

Alipio Jorge

9/29/2020

What is Data Science?

- Hard to define (very broad)
- Good label?
- Bringing disciplines together



(Source: [Drew Conway](http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram).

What is Data Science?

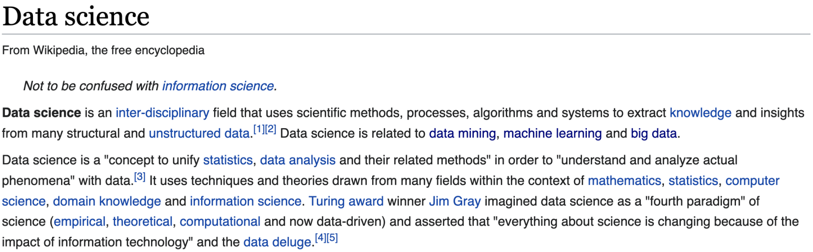
- Statistics / Mathematics
- Computer Science / High Performance Computing
- Machine Learning / Artificial Intelligence
- Data-driven Decision Support
- Data Bases / Big Data
- Business
- Research vs. Applications
- What else?

So, what is Data Science?

- My pop up definition
  - Is is about using methods to get value from data
- Value?
  - Health, Wealth, Pleasure, Happyness, ...

So, what is Data Science?

Given it is an **umbrella term**, disputed by different fields, wikipedia should get most of it



Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.[1][2] Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

Empirical vs Data-driven

- Empirical
  - from **Hypothesis** to Data
- Data driven
  - from **Data** to Hypothsis
  - sailing the sea of data can be **dangerous**
  - beware: don't be fooled by data (and don't use data to fool others)
  - but data can bring you new knowledge (. . . only data)

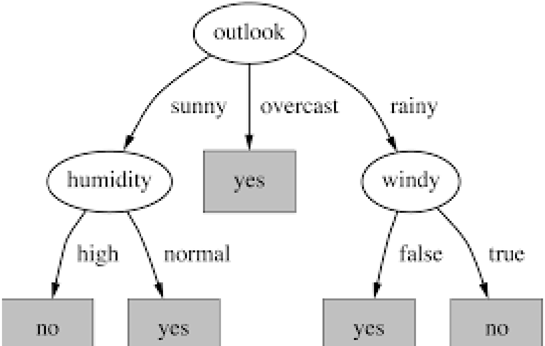
Figure 1: Saildrone

Data

- Real world data
  - messy, complex, voluminous
- Popular data
  - Surveys, Corporate data bases
  - Images, Sounds, Text
  - Clickstreams, Social Networks
- Tabular Data
  - One size fits all?
  - Vectors and Matrices: the universal language of data?

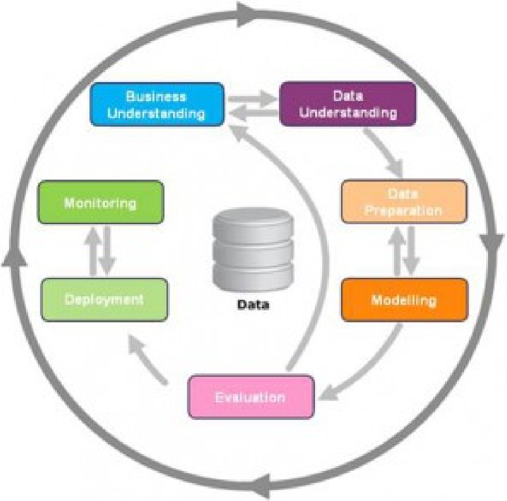
Models

- **Model**: answers questions
  - *is it a good day to play golf?*
  - *which types of clients do I have?*
  - *when is this patient going to develop symptoms?*
- Machine Learning **Algorithm**: builds models
  - given **data**, builds a model
  - also known as **method**
  - a ML **program** implements one algorithm (or more)
  - a ML algorithm learns or trains a model from data



The Process

- Given a problem we can approach with data science, what is the methodology to adopt?
  - CRISP-DM is **one** answer (the most consensual)



The Course

- Content
  - Pre-processing, Classification, Regression, Clustering, Evaluation
- Assessment
  - Test (13 Nov), Assignment (16 Nov to 15 Jan), Exam
- Programming
  - Python, R
  - "I know very little programming!"
- Classes
- Team
  - Alípio Jorge, Inês Dutra