Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Outline

- 4 courses of 3 hours each, with practical examples
- 3 Practical works of 4 hours each, with R/RStudio and python/anaconda programs / libraries
- 3 short quizzes
- Slides from several MOOCs, advanced materials

Outline

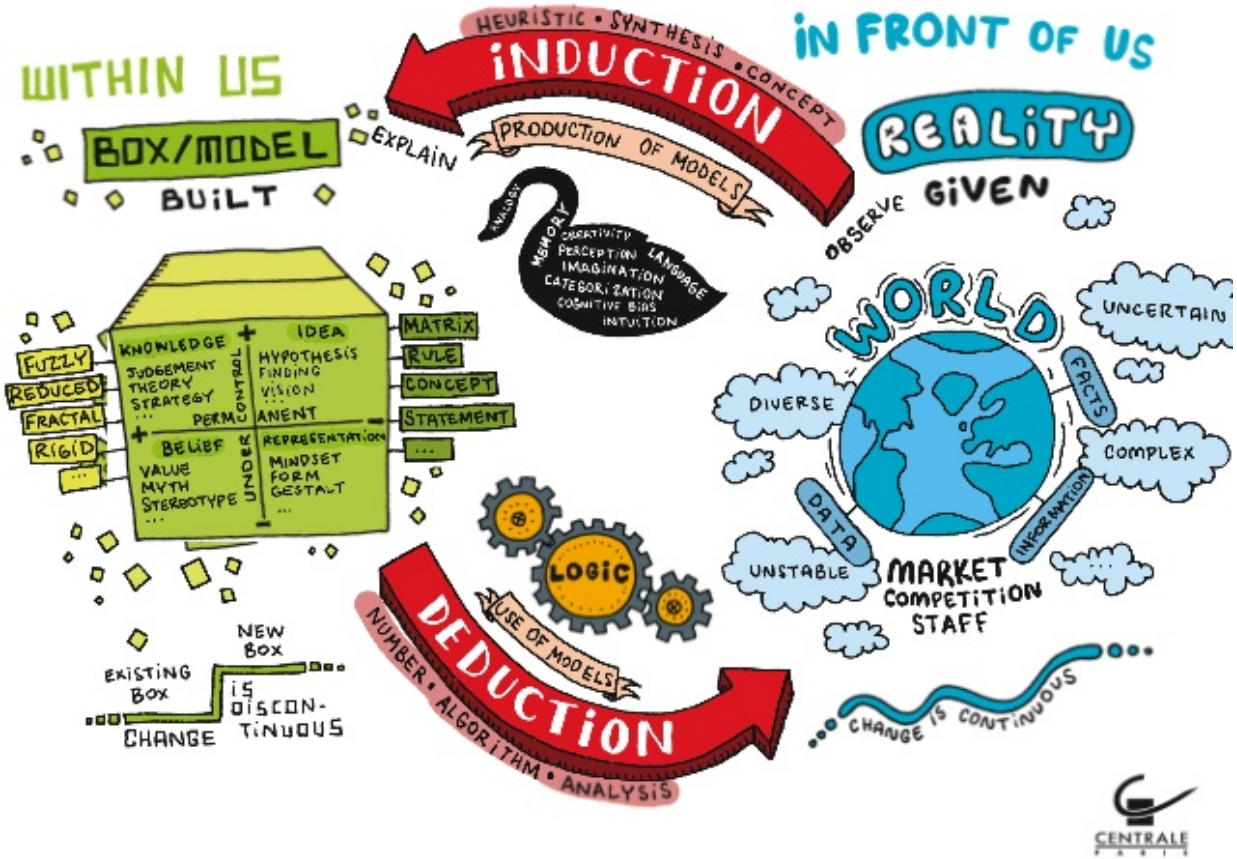
- How to understand data sets, correlation but not causation
- How to handle variety of data sets (see 3V description later)
- See materials for Supervised (*) and Unsupervised (*) use-cases

(*) Definition to come later

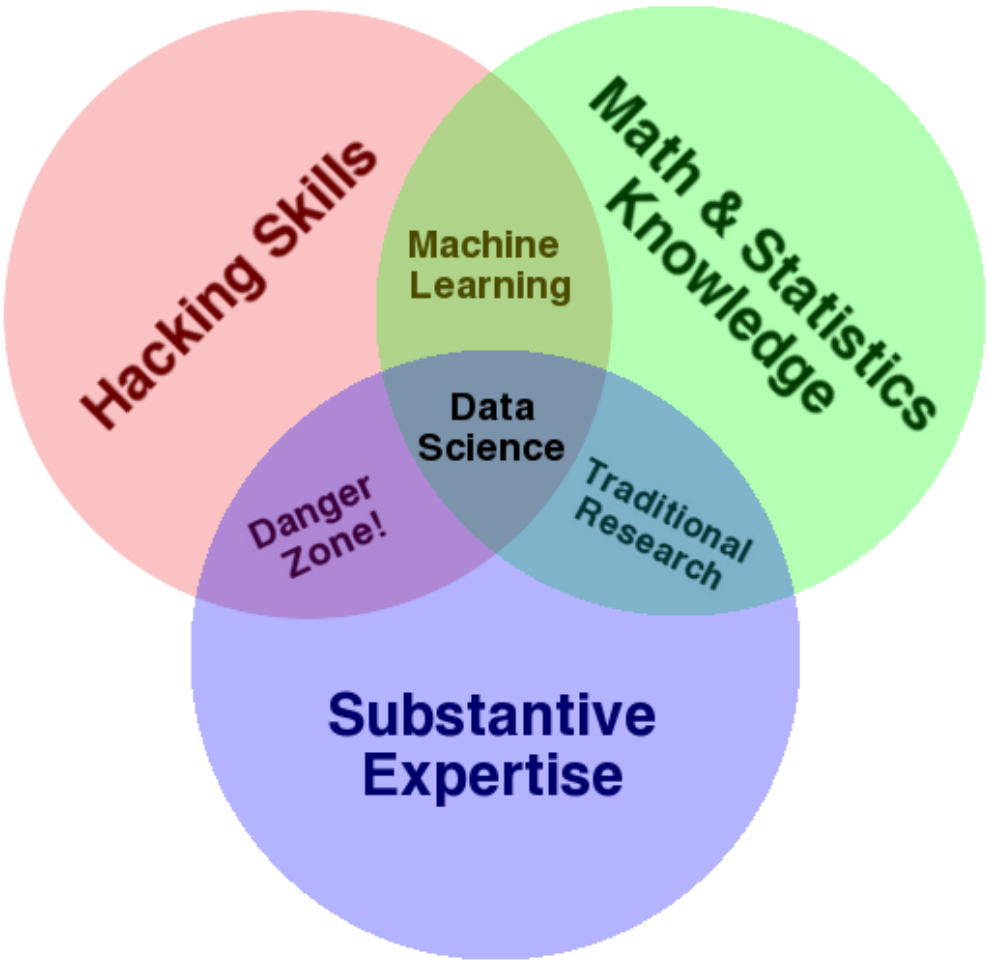
Outline

- 3V definition :
  - Volume, Velocity, Variety
  - + Veracity, ++
- Position of data mining vs ML vs Statistical Learning vs AI

Outline



- From a MOOC entitled What the manager can learn from Philosophy



- Blog post http ://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Outline

Starting point :

- Outcome measurement Y (also called dependent variable, response, target) ;
- Vector of p predictor measurements X_i (also called inputs, regressors, covariates, features, independent variables). X is a matrix of dimension (N,p), where N is the number of measurements ;
- In the **regression problem**, Y is quantitative (e.g price, blood pressure) ;
- In the **classification problem**, Y takes values or levels (categories) in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample) ;
- We have training data $(x_1,y_1),\ldots,(x_N,y_N)$. These are observations (examples, instances) of these measurements.

Outline

- In unsupervised learning, we observe only the features X1, X2, . . . , Xp.
- We are not interested in prediction, because we do not have an associated response variable Y.

Outline

- The goal of Unsupervised Learning is to discover interesting things about the measurements : is there an informative way to visualize the data ? Can we discover subgroups among the variables or among the observations ?
- We discuss two methods :
  - **principal components analysis (PCA)**, a tool used for data visualization or data pre-processing before supervised techniques are applied, and
  - **clustering**, a broad class of methods for discovering unknown subgroups in data.

Outline

- In reinforcement learning, observations are done one by one. (e.g. temporal series)
- After each prediction a reward is given
- Use Case : Chess or Go player
- toward common sense...

Outline

The 3 types of Machine Learning

From the most difficult to the less

Unsupervised Learn.	Supervised Learn.	Reinforcement Learn.
Observations Understand Clustering	+ targets (labels) Predict Classification or Regression	++ rewards Decide Action Policy Strategy

ML Cake

From the Yann Le Cun's lesson at College de France

Reinforcement Learning (cherry)

- The machine predicts a scalar reward given once in a while.
- **A few bits for some samples**

Supervised Learning (icing)

- The machine predicts a category or a few numbers for each input
- **10→10,000 bits per sample**

Unsupervised Learning (cake)

- The machine predicts any part of its input for any observed part.
- Predicts future frames in videos
- **Millions of bits per sample**