Lecture 1 : Introduction

§1 Data science

1. What's data science

1° deal with vast volumes of data
2° using modern tools and techniques
3° to find unseen patterns / derive meaningful information / make business decisions.

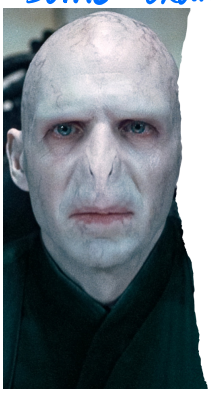
2. Data science is not one single subject!

incorporates
- probility
- statistics
- operations research
- machine learning
- business applications
- .....

3. What do you need

1° Analytical skills: solid mathematical / technical foundation
2° Bussiness understanding: industrial expertise

4. Some examples



Blind Box Example

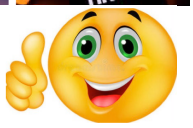
- Ten of you went to a blind box seller for the following blind box. You were told the chance to get a Harry Potter versus a Lord Voldemort is fifty-fifty.
- You all wish to get a Harry Potter.

Blind Box Example

- If the seller told the truth, the probability that all ten of you got a Voldemort is < 0.001. → Probability
- Given the above observation, you are pretty confident that the seller was lying. What's the true chance of getting a Harry Potter versus Voldemort? → Statistics
- Suppose you are a honest seller. How should you design the blind box and its price to maximize your own profit? → Optimization
- Design a Cartoon for Harry Potter. Which one is more likely to be popular?  → Machine learning

- Jack opened a shoe store and he didn't want to miss this opportunity.
- He plans to do some publicity, in order to increase the exposure and attract more customers to buy his products.
- There are many advertising channels.

- How much shall Jack invest in each channel?
- Do you have any advices?



- Collect other people's revenue performance with different investments.
- Let investment be X, and the revenue be R.
- Data: (X_i, R_i), ...

- From data, find the relation between X and R.
- Given the pattern as R(X), Jack's decision shall be

$\boxed{\text{argmax}_X R(X)}$

The value of X such that R(X) is maximized

§2 Key steps in data science

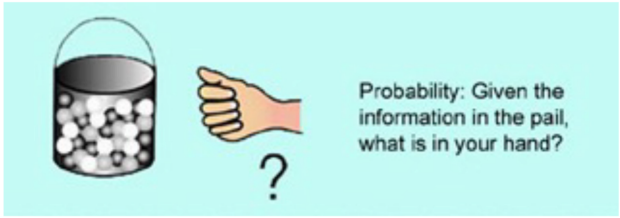
1. Programing (编程) (How to collect the past data?)

1° creating instructions that tell a computer how to perform a task
2° fundamental part

2. Probability (概率) theory (How to describe the data pattern?)

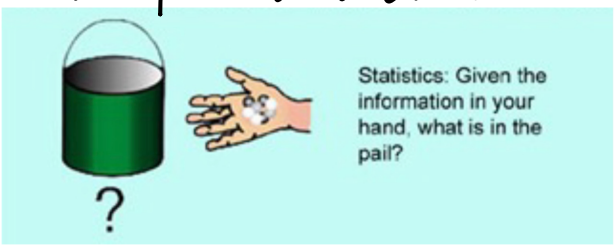
1° A formality to make sense of the world in terms of uncertainty

2° Numerical descriptions of how likely an event is to occur

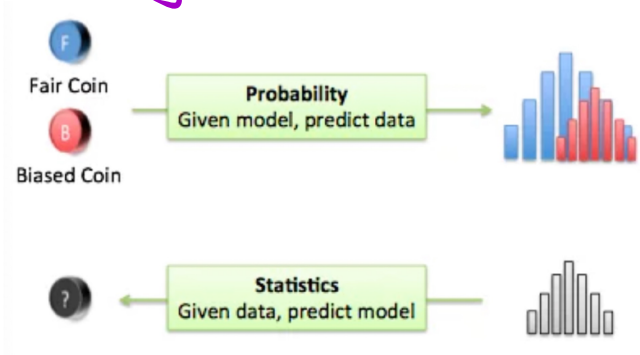


Statistics (统计学) (Is the data pattern we found reliable?)

1° The descipline that extracts correct information from data.



* Probability versus statistics



3. Simulation (模拟) (Given pattern, can we calculate objective for different strategy)

1° a computerized mathematical technique

2° generate random samples

3° aid quantitative analysis and decision making

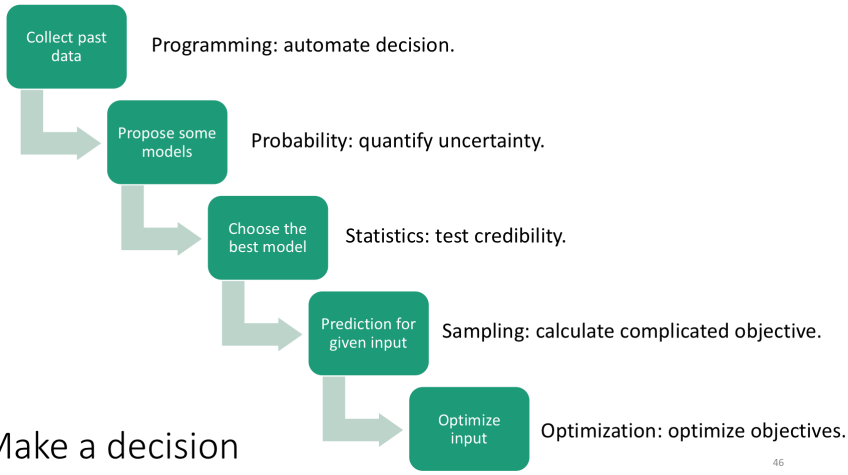
4. Optimization (最优化) (What decision should be made to maximize objective?)

1° The selection of a best element, with regard to some criterion, from some set of available alternatives

$$\min_{x \in \mathbb{R}} (x^2 + 1)$$

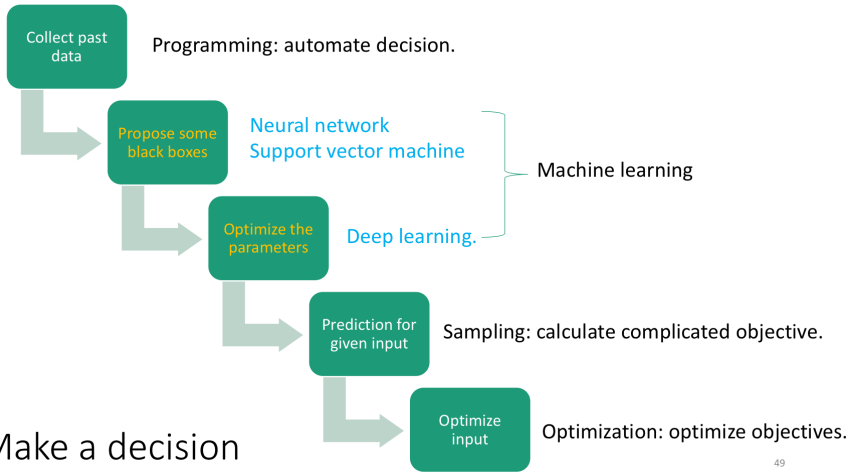
5* Machine learning (机器学习) (If step 2 is too complicated)

Allow software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.



Make a decision

46



Make a decision

49

Questions

1. Given the data, you want to predict how the data is generated. Which theory you may use?
   a) Probability
   b) Statistics

2. Given the profit as a function of strategies, you want to maximize the profit. Which theory you may use?
   a) Simulation
   b) Optimization

§3 Why study data science

1. Data science can be used in...
   1° retail (零售)
   2° healthcare
   3° finance
   4° education
   5° transportation
   6° agriculture
   7° startups (创业)
   8° pharmaceutical industry (医药产业)
   9° art industry