Data Exploration

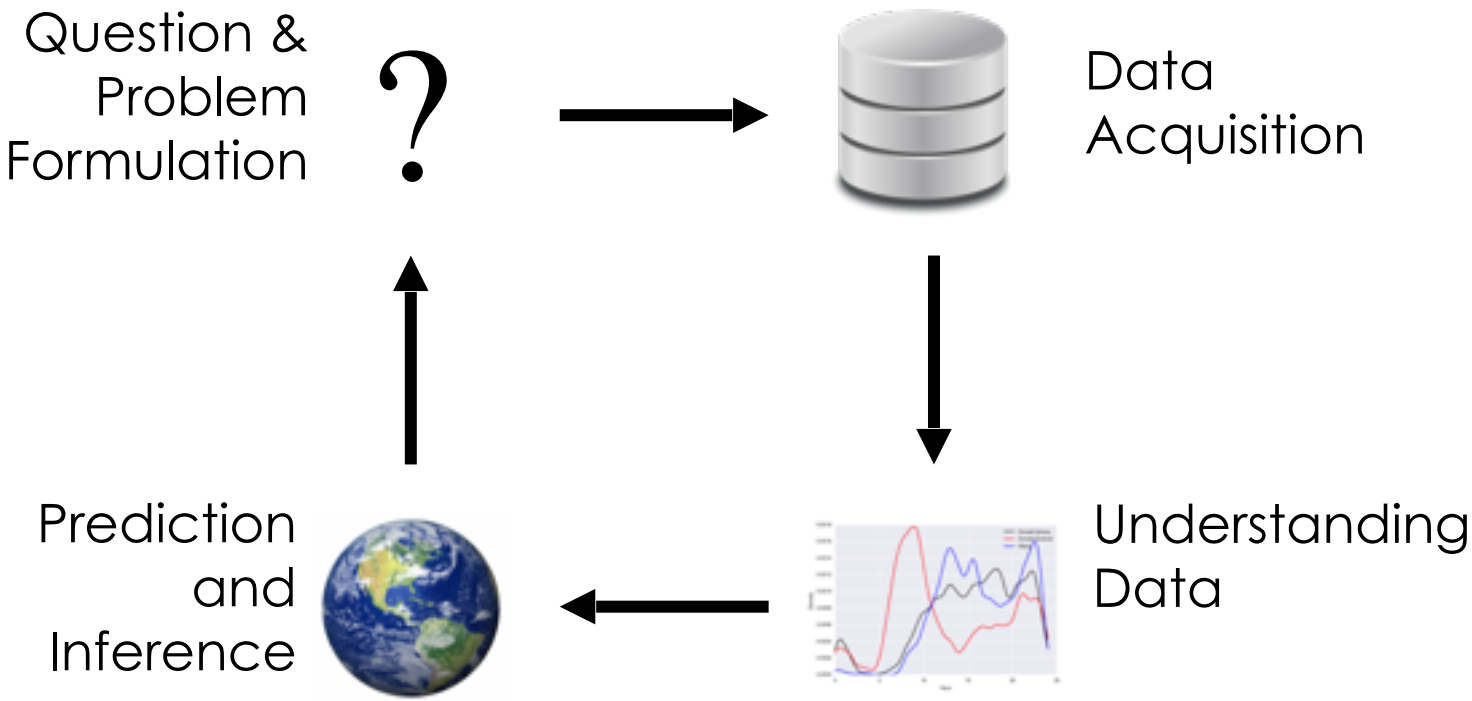
Seyed Abbas Hosseini
Sharif University of Technology

Outline

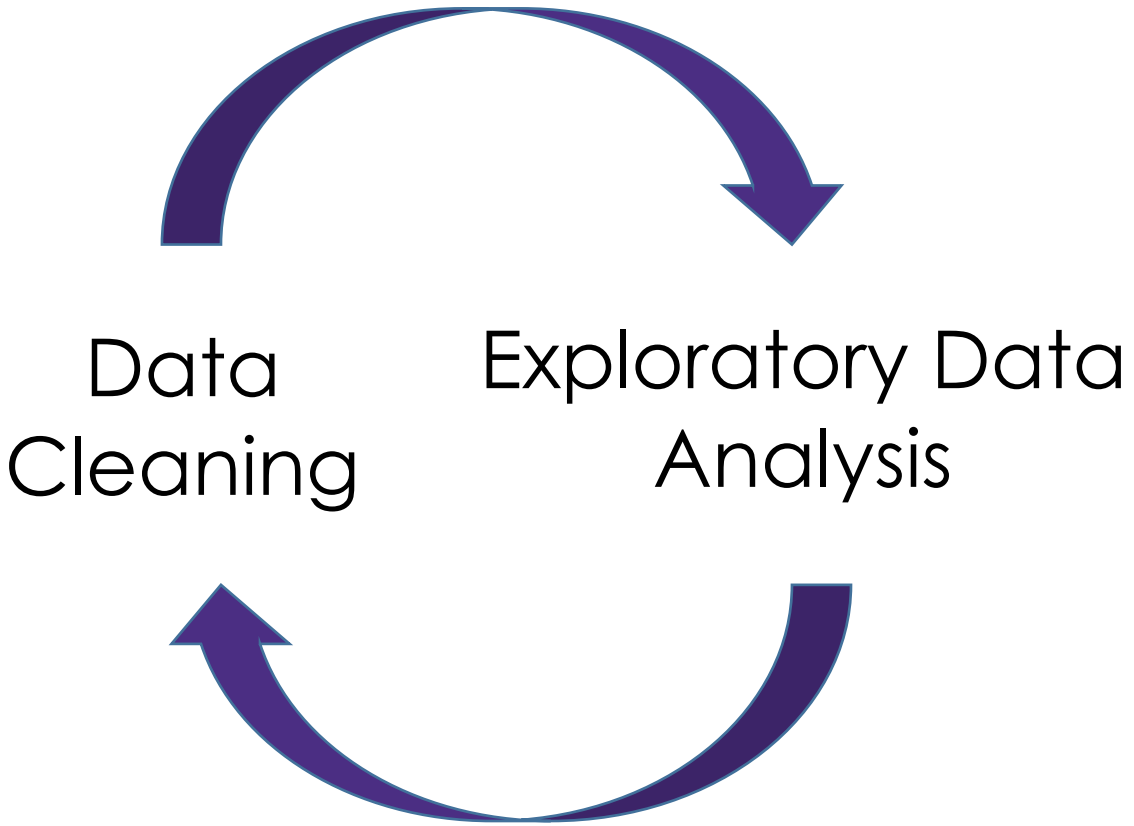
- ❑ Data Science Lifecycle
- ❑ Data Cleaning
- ❑ Pandas
- ❑ Demo
- ❑ Exploratory Data Analysis
- ❑ Demo

Data Science Life Cycle

Data Science Life Cycle



Data Understanding Loop

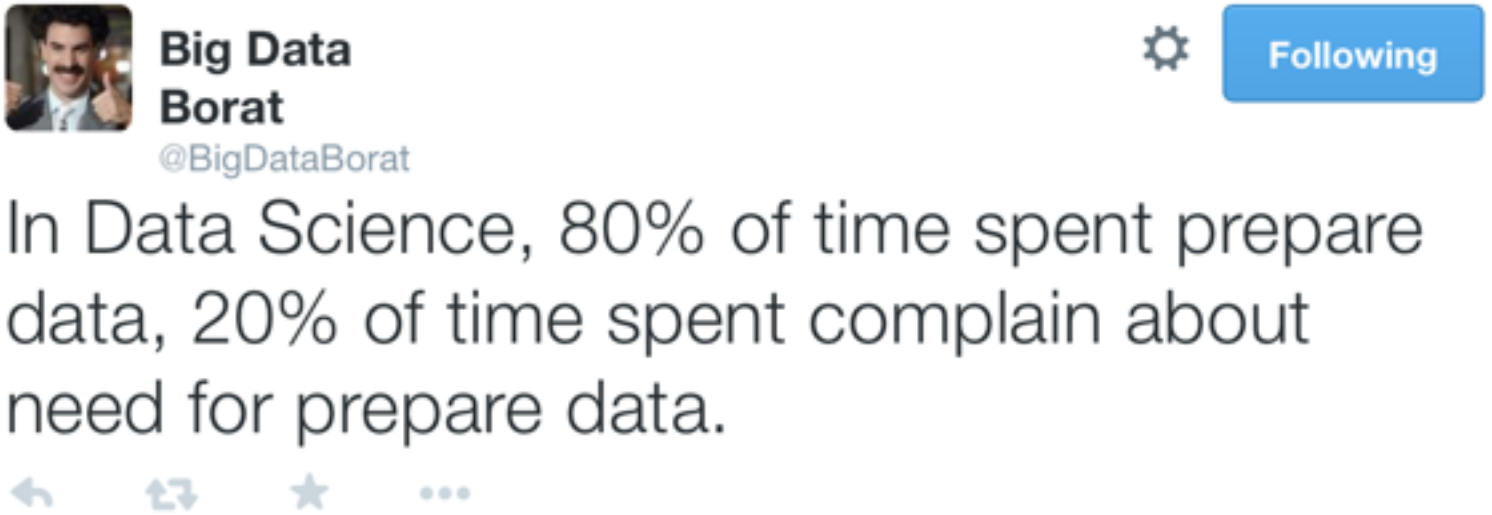


… the infinite loop of data science.

Data Cleaning

- The process of transforming **raw data** often to a tabular form to facilitate subsequent analysis
- Data cleaning often addresses **issues**
  - structure / formatting
  - missing or corrupted values
  - unit conversion
  - encoding text as numbers
  - …
- Sadly, data cleaning is a big part of data science…

Data Cleaning



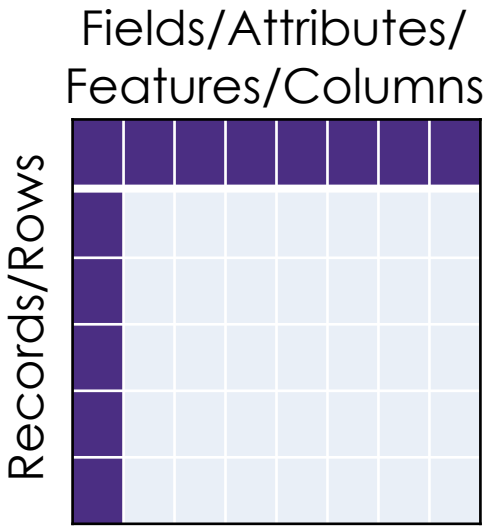
EDA

- The process of **transforming**, **visualizing**, and **summarizing** data to:
  - Build/confirm understanding of the data and its provenance
  - Identify and address potential issues in the data
  - Inform the subsequent analysis
  - discover *potential* hypothesis … (be careful)

- **EDA is an open-ended analysis**
  - Be willing to find something surprising

Data Cleaning

Rectangular Data



We prefer rectangular data for data analysis (why?)

- Regular structures are easy manipulate and analyze
- A big part of data cleaning is about transforming data to be more rectangular

Two kinds of rectangular data: *Tables and Matrices*

(what are the differences?)

- **Tables** (a.k.a. data-frames in R/Python and relations in SQL)
  - Named columns with different types
  - Manipulated using data transformation languages (map, filter, group by, join, …)
- **Matrices**
  - Numeric data of the same type
  - Manipulated using linear algebra (We will go through its details after review on linear algebra)

Pandas Demo

Any Questions?!

Read The docs and

Google it !