Introduction to R for Data Science

Modules	Objective
1. Introduction to Data Science	To understand the process of data science
2. Introduction to R	To familiar with R environment
3. Data Manipulation in R	To know various manipulation function in R
4. Graphics in R	To know how to graphically present data in R

Module 1
Introduction to Data Science

The Power of Data

"In the 21st century, **the candidate with [the] best data**, merged with the best messages dictated by that data, **wins**."

Andrew Rasiej, Personal Democracy Forum

"…the **biggest win came from good old SQL** on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

Source: from Potter, D., Binnig, C., Upfal, E.

A nice TED idea:
https://www.ted.com/talks/gary_flake_is_pivot_a_turning_point_for_web_exploration

What is Data?

- Datum
- Raw facts
- Always unstructured
- Numeric or non-numeric
- Transformation of data

My purchase list → Make payment to the cashier → Stock Deduction – Alert Stock Manager to make new order

→ End of month - sales report – Sales manager will decide whether to have promotion

What is Data? cont…

- Data is valuable but it needs people to work on it
- Good communication: data transform successfully from source to recipient
- Recipient receives information
- In computer, smallest unit of information is bit. 8 bits = byte
- How to know whether the data is right or wrong? --- through science

What is Science?

- Systematic and logical approach to discover how things work in the universe
- It allows us to explore (identify) data, make prediction, infer (quantify what you know)

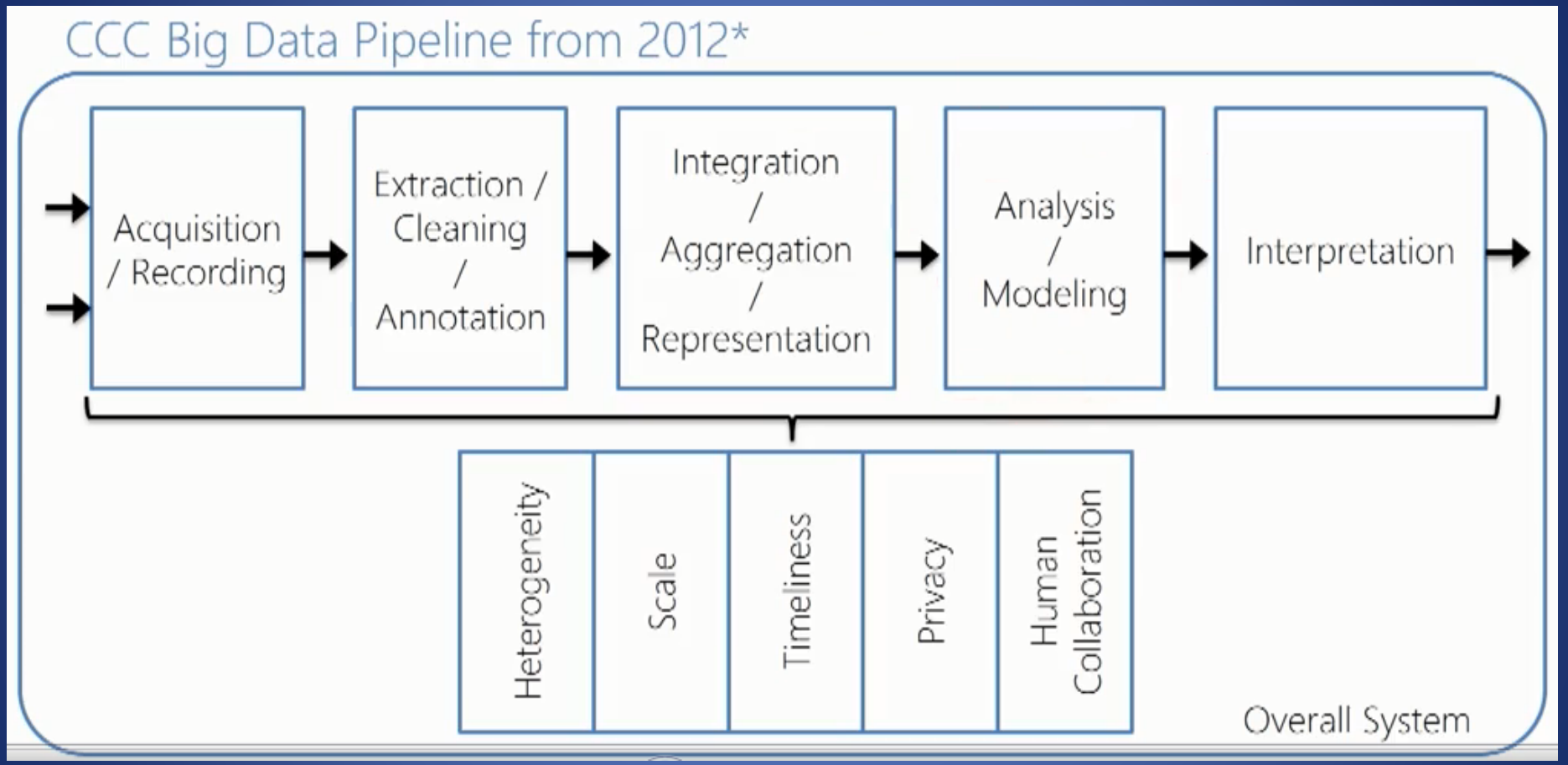
Data Science

- Various definition
  - Using data to make decision that drive actions
  - Application of computational and statistical techniques to address or gain insight into some problem in the real world
  - Emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information. (Stanton, 2013)

Data Science cont…

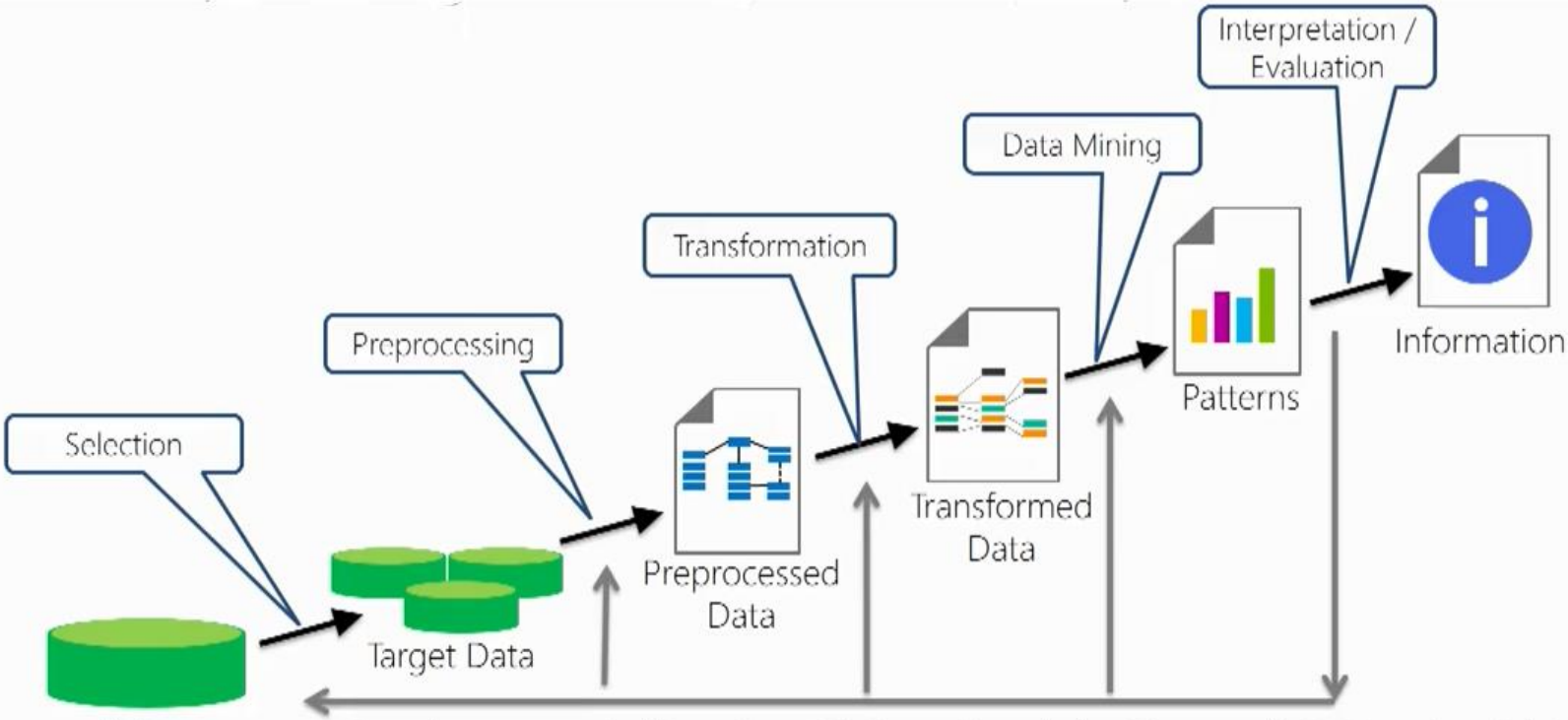
- What is data science?
  - Data-driven (the more the better)
  - Interdisciplinary (mathematic, statistic, computer science….)
  - Extract knowledge from data

The Process of Data Science



Source: https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

Knowledge Discovery in Database



Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Typical Steps for Data Analysis

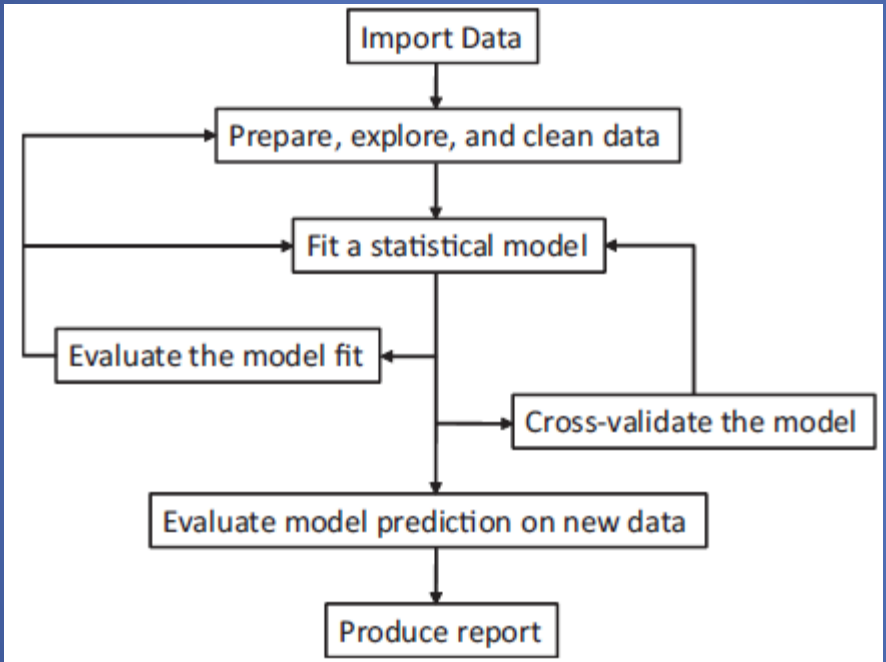


Figure: Data Analysis (Kabacoff, 2011)

Data Scientist

"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

-- Hilary Mason, chief scientist at bit.ly

Data Scientist cont…

- Play vital role in
  - Data acquisition (selection): identifying the problem and collecting data;
  - Extraction/Cleaning/Annotation (preprocessing): pick useful sets of information, clear all the outliers, solve the missing data, standardized the data
  - Integration/Aggregation/Representation (transformation): linking, grouping data
  - Data analysis/modeling (Data mining) : most heavily involved – summarize data; use small set of data to inferences the larger context; visualize the data
  - Interpretation/Evaluation

Data! Data! Data

- Airline industry has great improvement on safety
  - Obtain the accidents happened for several years
  - Obtain the total flights of each region
  - If the result provides information such as 1 death over 45 million flights, so it is highly safe in a flight
- Where to get the cheapest supermarket items?
  - Obtain recent prices of all supermarket items
  - Obtain distant to all the shopping malls within list
  - Make comparison
  - Conclusion: the cheapest supermarket items
- Success story: automatic image captioning; product recommendation (Netflix); spam detection
- The one who ignore data: Nokia

Exercise 1.1 – 5 mintues

- Before data could be analysed, the data must be (please select all that apply):
  - A. Recorded
  - B. Cleansing
  - C. Modeling
  - D. Transforming
  - E. Predicting