Data and Algorithms

LECTURE 2

Data in Data Science

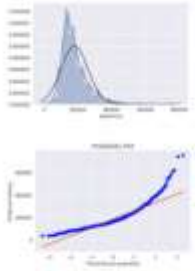
- Data needs: Analyze, Acquire, Organize and Structure
- Data format:
  - Databases – store and integrate data
  - Textual and Natural Language (Qualitative Data)
  - Numeric Data (Quantitative)
  - Interconnected Data (Networks)
  - Visualized (data → information)
  - Sample data for predictions (Learning data sets for Machine Learning)
  - Time Data (Digital Signal Processing)
  - Big Data (Usually unstructured – may include text, audio, video over 1 TB)

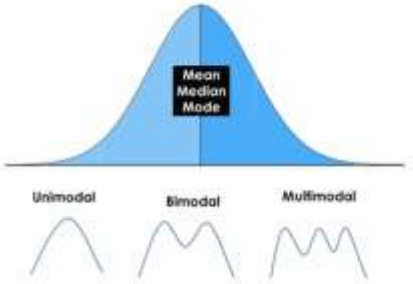
Data Analysis Process

Steps similar to steps followed in scientific discovery

1. Identify question(s) to be answered & **type** of analysis to apply
2. Identify best data set and obtain raw data
3. Clean data / “regularize” the data set
4. Perform analysis
5. Prognosticate
6. Determine if results significant/ Review work
7. Produce Report

Types of Analyses

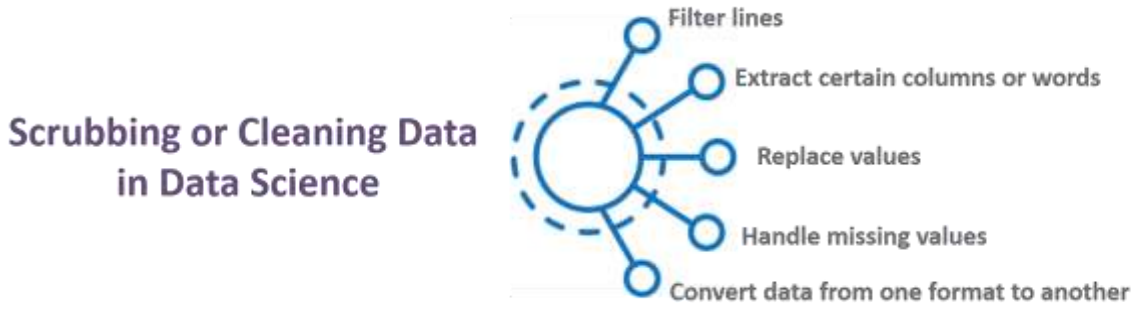




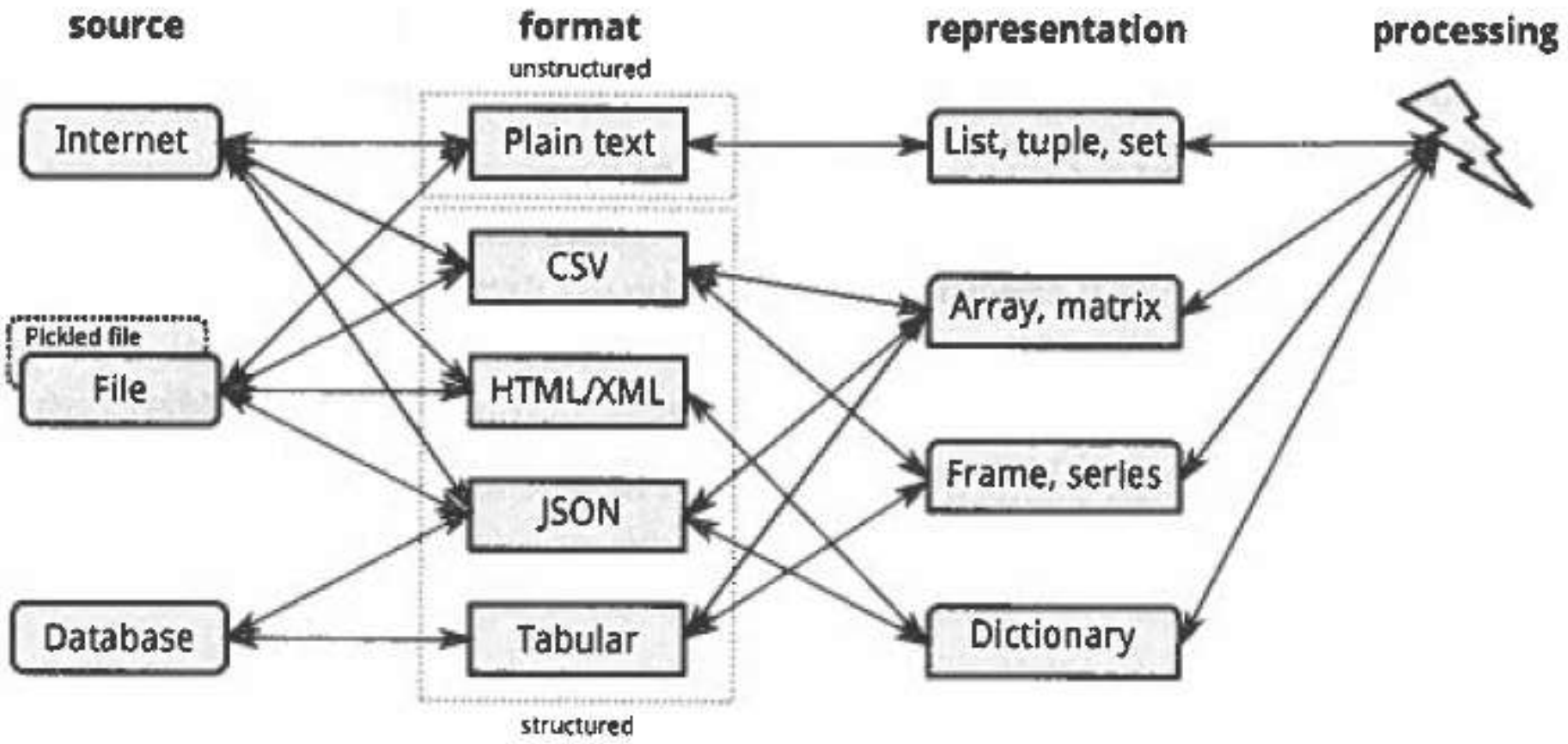
- Descriptive – data set is described by reporting aggregate measures (often visual)
- Exploratory – goal is to find relationships between existing variables
- Inferential – statistical-based, used when you have small data sample and want to describe bigger population
- Predictive – use past data to predict future
- Causal – identifies variables that affect each other
- Mechanistic – explores how ONE variable affects another

Data Acquisition & Cleaning

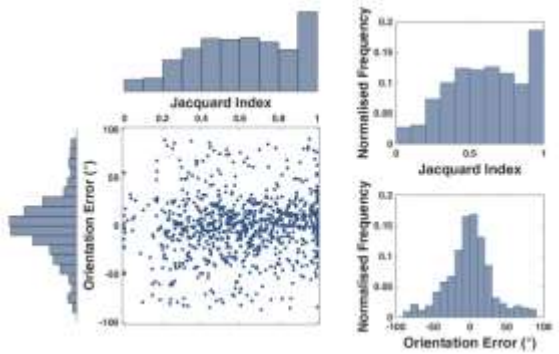
- Study will only be as good as data
- May be no ideal data set
- Raw data from web, database, local file
- No perfect data
  - Missing values, outliers, “non-standard” data
  - e.g. negative ages, not enough digits for ids, bad dates 2/30/2012



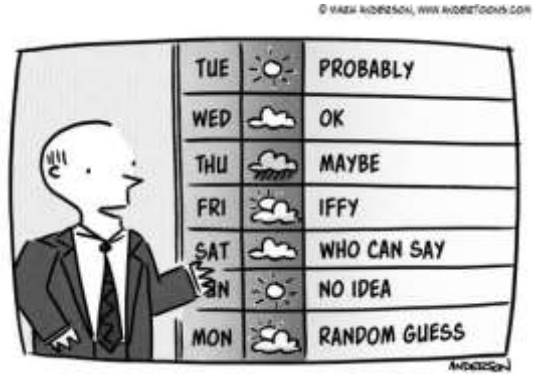
Data Representations



Actual Analysis



- Start with descriptive and exploratory analysis
- Usually will output scatter plots, histograms, statistical summaries
- Will give you a sense of what direction to take
- Next prognostication?
  - Learn from the past, predict the future
- Validate/assess quality of models and their accuracy



Review Work & Report

- Focus not on stats or programming
- Focus on domain
- Are results – significant, will anyone care, what did you do right or wrong, what would you do better or differently?

- Report
  - Explain how and why you processed data, what models you built/used, your conclusions and predictions
  - A report has a structure

Report Structure

- Depends on Data Sponsor/Supervisor/Publisher/Customer
- Typical Structure
  - Abstract
  - Introduction
  - Methods used for data acquisition and processing
  - Results (intermediate and insignificant results go in appendix)
  - Conclusion
  - Appendix

- https://cs.msutexas.edu/~stringfellow/papers/SEKE2006.pdf
- https://cs.msutexas.edu/~stringfellow/papers/SRGM.pdf

Appendix includes

- Non-essential results
- Reproducible code to process data
  - Well-commented scripts, list of tools/commands used
- Raw data/link to raw data (unless it is proprietary)

- Might include README file if report is on a CD/DVD or shareable drive

Your Turn

Install Python (with the Anaconda distribution) and the packages on page xv in Table 1

Write a Hello World Program

Write a Program to read in two numbers, sum them and output them

HOMEWORK DUE: 9/2/21 (Thursday)