LEADING 3-Day Camp
Data Cleaning and Preprocessing

Rezvaneh (Shadi) Rezapour, Ph.D.
Assistant Professor
Drexel University
sr3563@Drexel.edu
https://www.shadirezapour.com/

About the Instructor

Education:

- PhD in Information Sciences
- MSc in Information Management
- BSc in Electrical Engineering

Research:

- Computational Social Science
- Natural Language Processing
- Human-centered Data Science
- Social Network Analysis
- Machine Learning

Concepts Covered

- What is Data Science?
- Perspectives on Data Science
- Typical Data Science Process
- Real-world Data
- Structured and Unstructured Data
- Data Objects
- Basic Data Types
- Data Cleaning
- Data Cleaning Processes

What is Data Science?

What is Data Science?

*"**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains."*

(https://en.wikipedia.org/wiki/Data_science)

*" A **data Scientist** is someone who creates programming code, and combines it with statistical knowledge to create insights from data"*

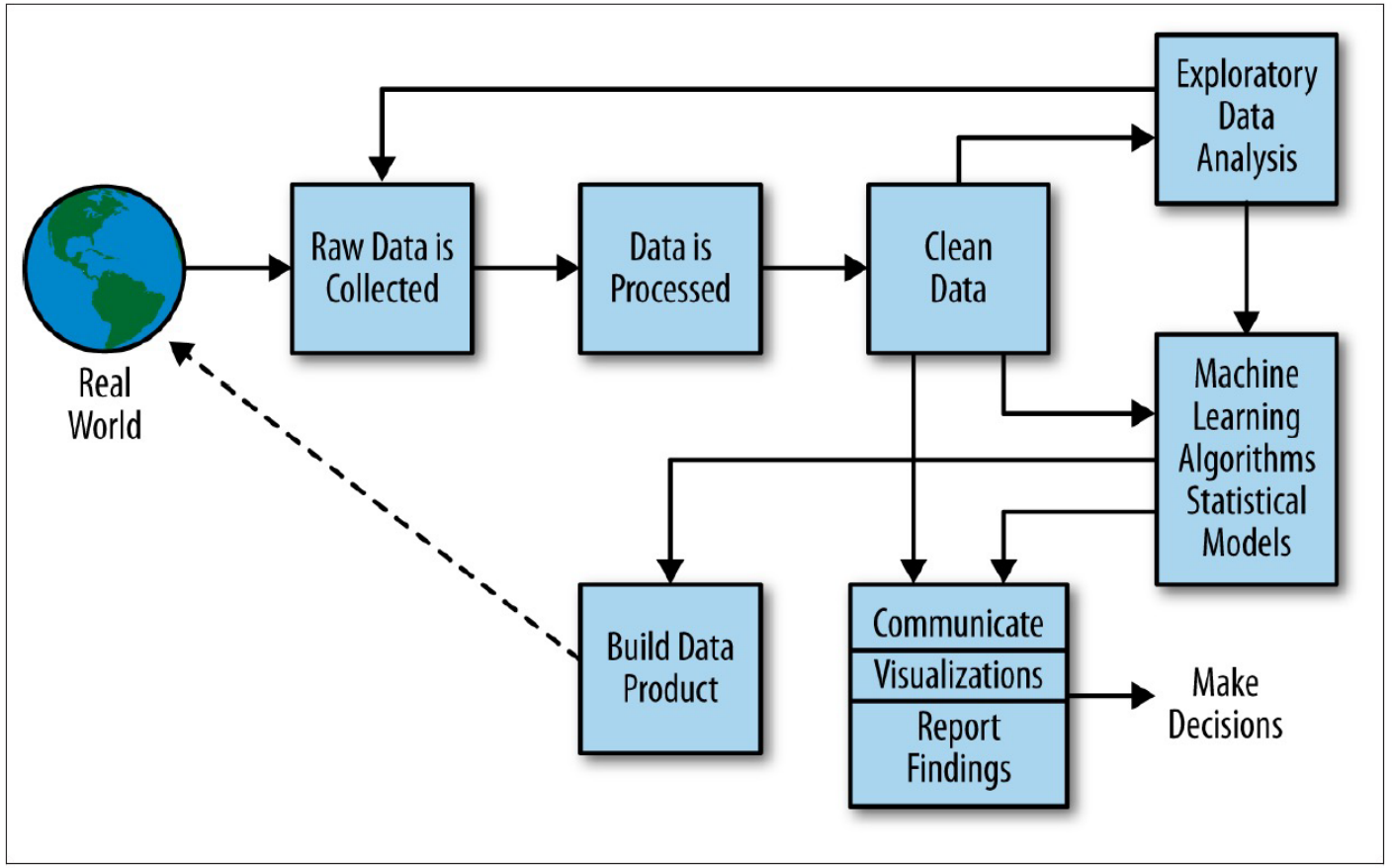
(https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century)

Perspectives on Data science

- Data science projects are highly non-linear, commonly requiring course adjustments along the way
- The output of such projects might include papers, reports, prototypes, patents, or company products
- Exploratory Data Analysis (EDA) is the agreed upon term for “making sense” of data

Typical Data Science Process

- **Initiation:** what does the project consist of?
- **Project Framing:** what is the topic of interest?
- **Data Collection:** what data will be the object of study?
- **Exploratory Analysis:** in what condition are the data and what patterns exist?
- **Project Design:** where do the data and project's interest overlap?
- **Pre-processing:** how can the data be collected, modified, or enriched to better satisfy the project interest?
- **Hypothesis Generation:** what patterns in the data can be leveraged?
- **Model Development:** how are patterns in data to be leveraged?
- **Evaluation:** how well did the project perform at satisfying the interest?
- **Output:** how ill the project's results achieve maximum impact?
- **Operations and Optimization:** how can the project's output stay relevant?



"The hardest part of data science is getting good, clean data. Cleaning data is often 80% of the work."
(DJ Patil's talk on Building Great Data Products)

Question

- Have you worked with any datasets before?
- Describe the data.
- Any challenges?

Real-World Data

- **Volume**
- **Velocity**
- **Variety**
- **Veracity**
- **Messy**
- **Incomplete**
- **etc.**

The Four Vs of Big Data

Structured vs. Unstructured Data

- Structured data has consistent, prespecified organization or order
  - Examples include lists, dictionaries, and spreadsheets
- Unstructured data refers to data that does not have a consistent, prespecified organization or order
  - Examples include texts, images, and audio recordings

Data Objects

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- Datasets are made up of data objects.
- A data object represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called samples, examples, instances, data points, objects, tuples.
- Data objects are described by attributes.
- Database rows -> data objects; columns ->attributes.

Basic Data Types

- Boolean: Either True or False
- Numerical: Could be integers (ints) or floating points (floats)
- Nominal: categories, states, or “names of things”
- Ordinal: Letter grades in the exam (A, B, C, D, etc.) (have natural ordering where a number is present in some kind of order by their position on the scale)
- Textual: Represented by strings (i.e., a sequence of characters)

Data Cleaning

- Identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data
- Data cleaning procedure can help with removing major errors and inconsistencies that are inevitable when multiple sources of data are being pulled into one dataset.
- Every data cleaning consists of two phases:
  - Error detection
  - Error repair

Working with Data

- Data Observation
- Data Cleaning and Processing
  - Zero Variance Attributes
  - Columns with Very Few Unique Values
  - Duplicates
  - Outliers
  - Missing Values
  - Scaling features
  - Dealing with Categorical Data

Coding time!

- Jupyter Notebook

or

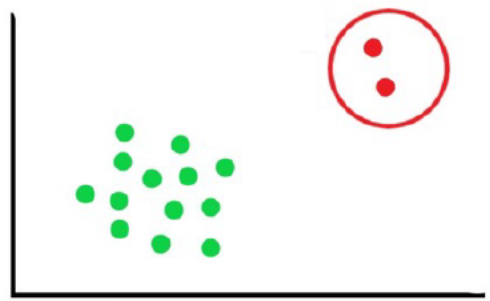
- Google Colab

Acknowledgment:

Coding Materials: Lei Wang (Assistant Professor, Drexel University)

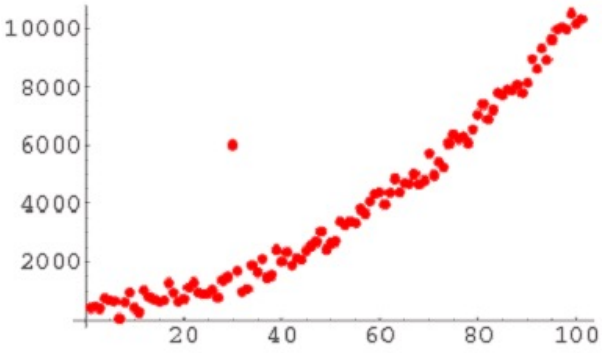
What Are Outliers?

- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
  - Ex.: Unusual credit card purchase
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: It violates the mechanism that generates the normal data
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis



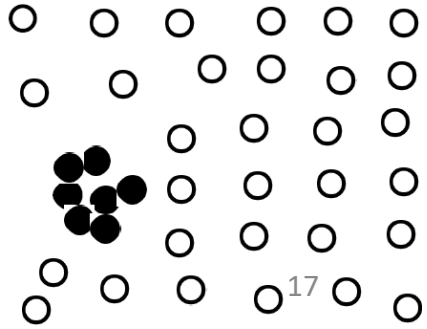
Types of Outliers

- Three kinds: global, contextual and collective outliers
- Global outlier (or point anomaly)
  - Object is Og if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation
- Contextual outlier (or conditional outlier)
  - Object is Oc if it deviates significantly based on a selected context
  - Ex. 80o F in Urbana: outlier? (depending on summer or winter?)
- Collective Outliers
  - A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Applications: E.g., intrusion detection:
    - When a number of computers keep sending denial-of-service packages to each other
- A data set may have multiple types of outlier
- One object may belong to more than one type of outlier



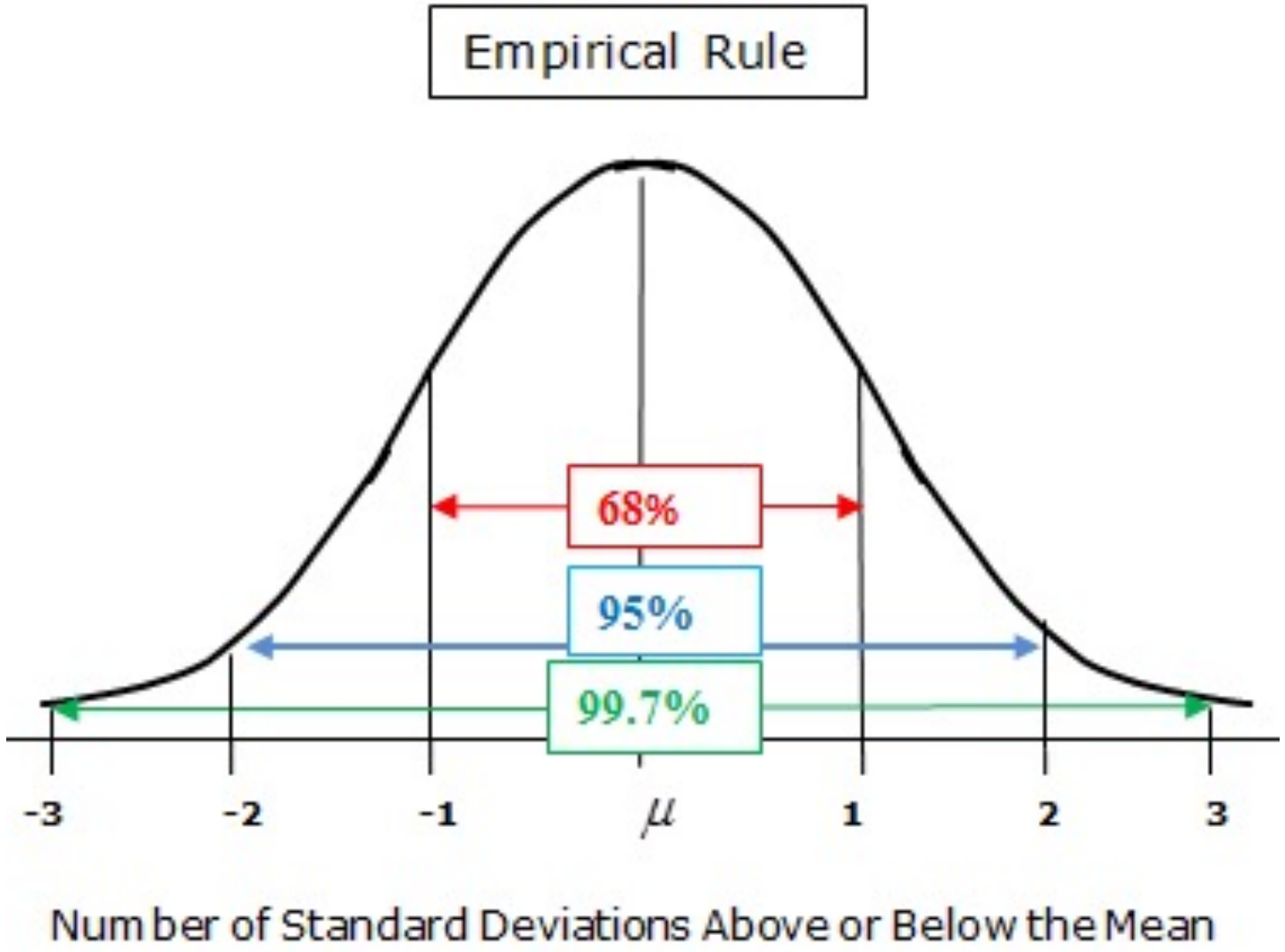
Global Outlier

Collective Outlier

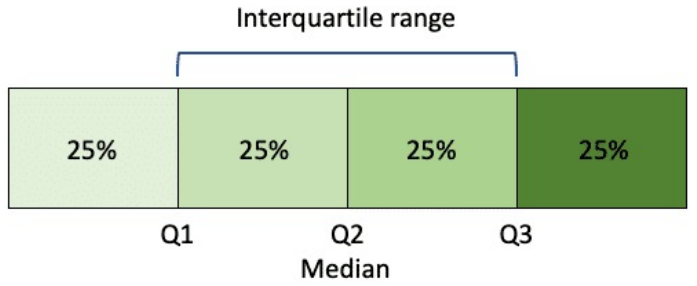


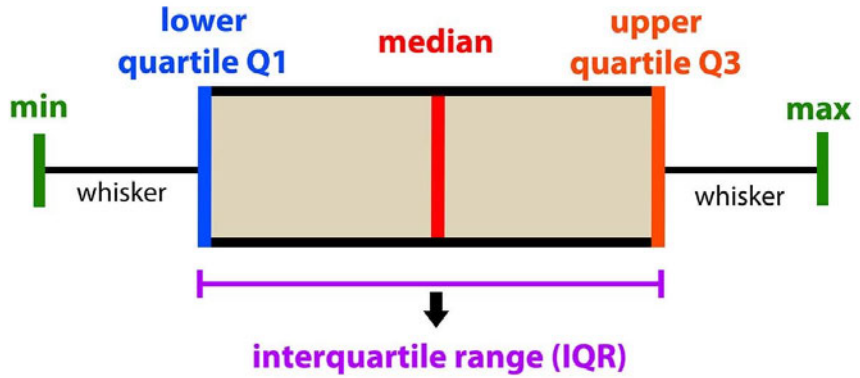
Outlier Detection Methods

- Two ways to categorize outlier detection methods:
  - Based on whether user-*labeled* examples of outliers can be obtained:
    - supervised
    - semi-supervised
    - unsupervised methods
  - Based on *assumptions about normal data and outliers*:
    - statistical
    - proximity-based
    - clustering-based methods


Empirical Rule
68%
95%
99.7%
-3
-2
-1
μ
1
2
3
Number of Standard Deviations Above or Below the Mean

Boxplot Analysis





- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles,
    i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually (1.5 times IQR above the 75th percentile and below the 25th percentile)