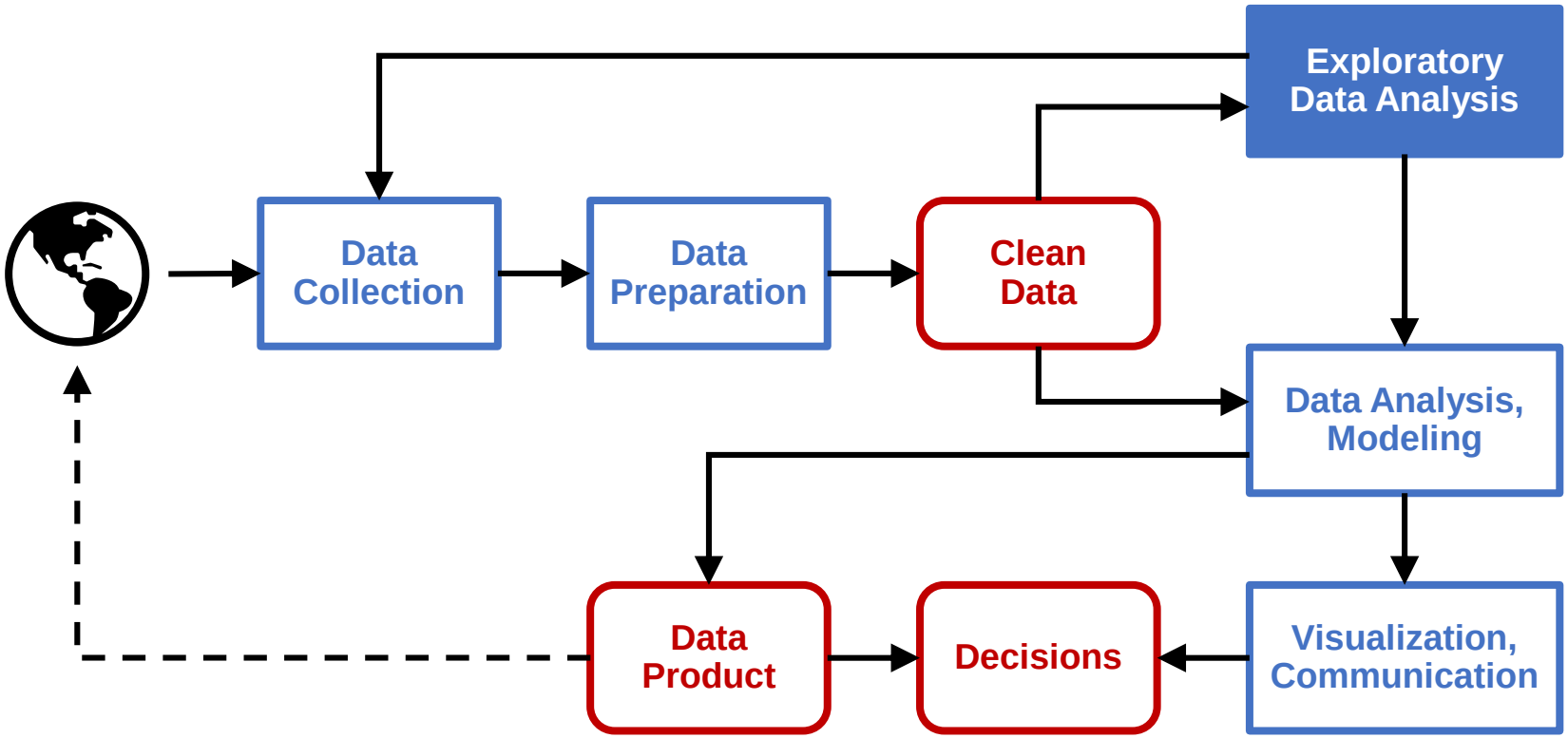
Exploratory Data Analysis

CS 418. Introduction to Data Science

Exploratory Data Analysis
The Data Science Process

- The goal of the **data science process** is to **extract knowledge or insights from data**.



Adapted from: Cathy O'Neil and Rachel Schutt, *Doing Data Science* (2013)

Exploratory Data Analysis
Introduction (I)

- **Exploratory data analysis** (**EDA**) is the **first step** toward building a model and it is a critical part of the **data science process**.
- In **exploratory data analysis**, there is no hypothesis and there is no model. The goal is to **understand and summarize the data and its underlying structure**.
- The basic tools of **exploratory data analysis** are **summary statistics** and **graphs**.
- **Exploratory data analysis** was developed by John Tukey as a contrast to **confirmatory data analysis**, which is concerned with modeling and hypotheses testing.

"[M]y central interest is in **data analysis**, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate..."

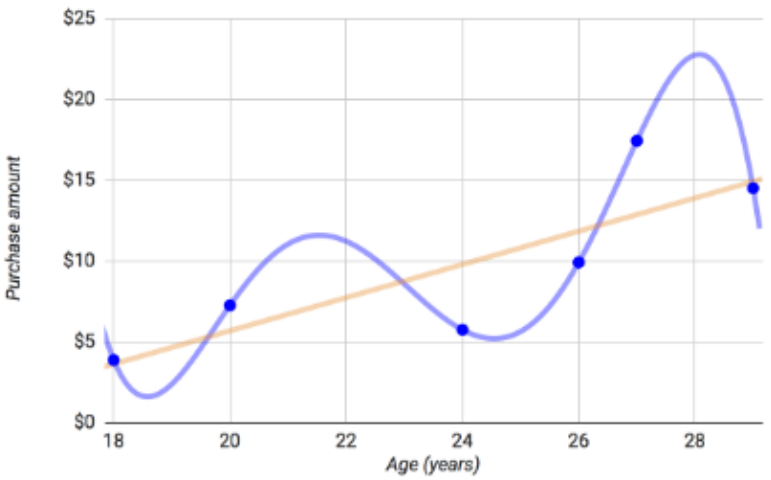
John W. Tukey
The future of data analysis (1962)

Exploratory Data Analysis
Introduction (II)

- **Exploring one-dimensional data:**
  - Compute **summary statistics** (mean, standard deviation, max, min).
  - Create a **histogram**.
- **Exploring two-dimensional data:**
  - Create a **scatter plot**.
  - Compute **correlation**.
    - **Correlation** is a value **between -1 and 1** that measures the **linear relationship** between two variables
    - A **positive correlation** indicates that one variable increases as the other increases. A **negative correlation** indicates that one variable decreases as the other one increases.
- **Exploring high dimensional data:**
  - Create a **scatterplot matrix** showing all pairwise scatter plots.
  - Create a **correlation matrix** showing all pairwise correlations.

Exploratory Data Analysis
What Next?

- Depending on the conclusions of the **exploratory data analysis**, we may return to **data collection** or proceed to **data modeling**.
- In the **data modeling** step, we will construct a **statistical model** that represents the data.
- We will construct (or **fit**) this model by **estimating its parameters** using the observed data.
- **Overfitting** means the model follows the observed data too closely and is not good at representing unseen data.
  - **Overfitting** has several causes, such as selecting a model with **too many parameters** or **too many variables**.



Exploratory Data Analysis
References

- Daniel Chen. *Pandas for Everyone* (2018).
- Joel Grus. *Data Science from Scratch* (2015).
- Cathy O'Neil and Rachel Schutt. *Doing Data Science* (2013).