Data-Driven Problem Solving in Mechanical Engineering

Exploratory Data Analysis

Masoud Masoumi

ME 364 - Spring 2022

Department of Mechanical Engineering
The Cooper Union for the Advancement of Science and Art

October, 2020

Introduction

In *hypothesis-driven* science, we formulate a theory of how the world works, and then seek to support or reject this hypothesis based on data.

Data-driven science starts by assembling a data set, and then hunts for patterns that ideally will play the role of hypothesis for future analysis.

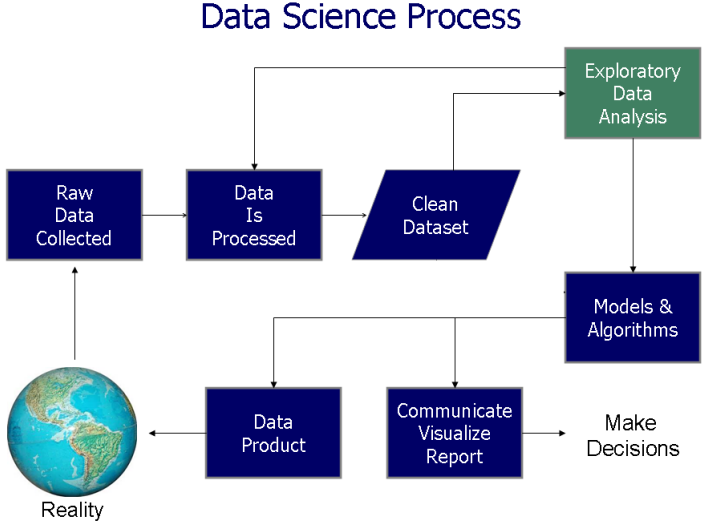
Exploratory Data Analysis (EDA) is the search for patterns and trends in a given data set.

EDA is exploring a data set with the very broad question "what is going on here?"

So, the purpose of EDA is

- To use summary statistics and visualizations to better understand data
- To identify properties of the data and highlight which data values should be treated as noise or outliers

Data-Driven Process



[link:https://en.wikipedia.org/wiki/Exploratory_data_analysis#/media/File:Data_visualization_process_v1.png (modified)]

Confronting a New Data Set

Here are the general steps you need to take when facing a new data set

Answer the Basic Questions

- Who constructed this data set, when, and why?
- How big is this data set?
- What do the fields mean?

Look for familiar and Interpretable Records

- You can get to know the data better
- You can test the soundness of data
- Sometimes you can create them if they don't exist

Data Cleaning

Summary Statistics

Pairwise Correlations

Plots of Distributions