Data Analysis Process

CIS-2266

Good to Have a Process

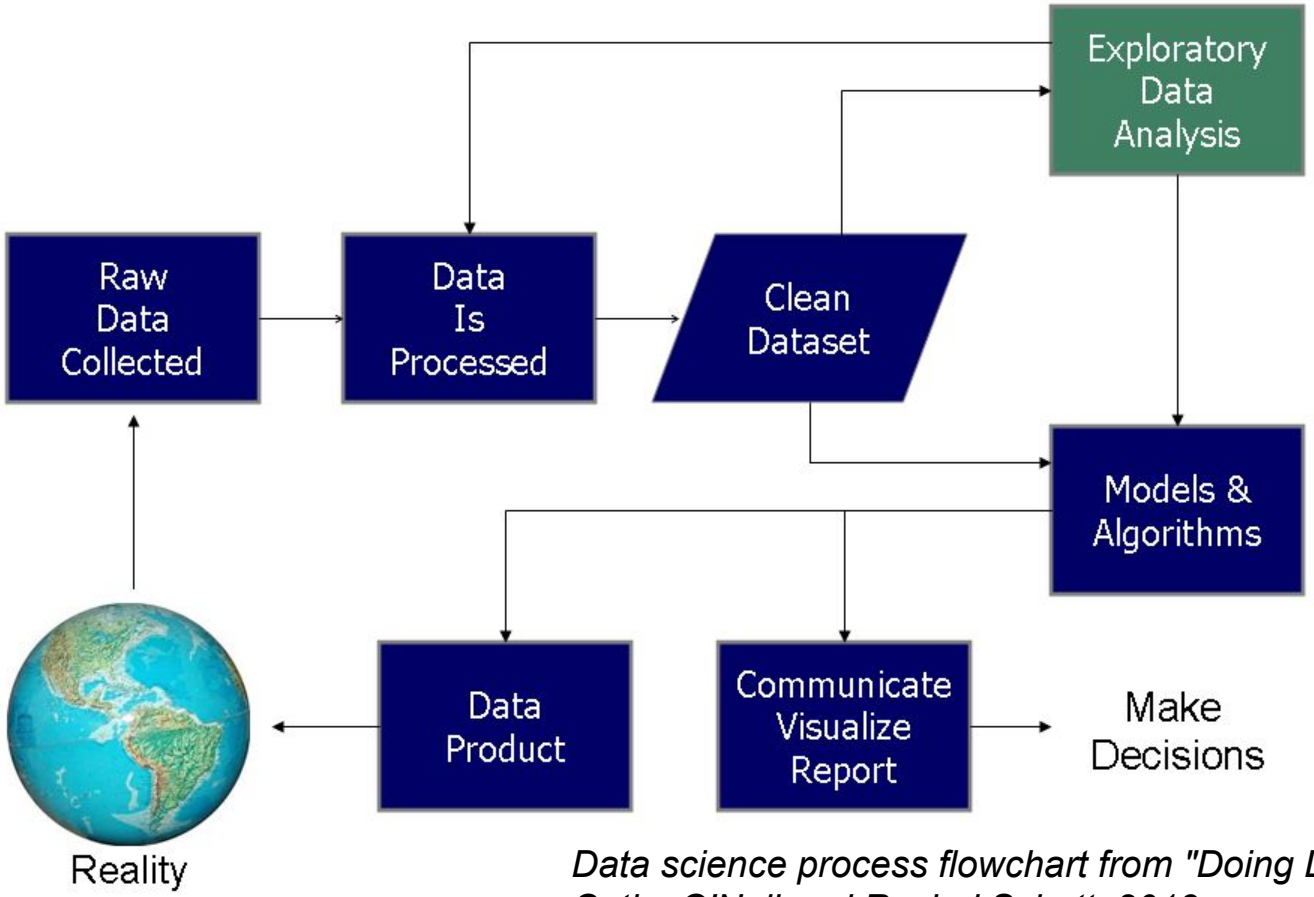
- Want to:
  - Show integrity
  - Non-bias
  - Quality of Data
  - Integrity of the Process

Data Analysis

"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, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

- John Turkey, Statistician, 1961

Data Science Process



Data science process flowchart from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013

Process

1. Requirements
2. Collection
3. Processing
4. Cleaning
5. Exploration
6. Modeling
7. Product
8. Communication

Data Requirements

1. What data is needed to answer any pending questions?
2. What is the nature of the data? i.e. numerical, categorical, boolean…
3. Nature of data:
    a. Qualitative data - Information about qualities; information that can't actually be measured.
    b. Quantitative data - Information that can be measured.
4. How much is needed to make a significant analysis?

Data Collection

1. The act of gathering the raw data for analysis.
2. May be directly measured and collected or..
3. Previously done and pulled from a data source
    a. Example:
        i. Using U.S. census data
        ii. Actually canvassing a neighborhood with questionnaires.
4. Be sure to document the data source for any independent analysis.

Data Processing

1. May need to stage collected data into format for further steps.
2. Example: Get XML data into a table format.
3. Keep original data set and process into new datasets
4. Document the process done

Data Cleaning

1. Check Processed Data for:
    a. Duplicates
    b. Errors in data types
2. Correcting or handling these errors
3. Document the findings and how they were handled

Data Exploration

1. Understand the data
2. Max, min, median and mean
3. Distribution:
    a. Normality
    b. Skewness
    c. Kurtosis
4. Visualization may be good
5. Document Results

Data Models

1. Using a documented method such as:
    a. Mathematical Algorithm
    b. Statistical Analysis
        i. Correlation
        ii. Causation
        iii. Regression Analysis
2. Generate Results
3. Infer Answers to Questions from first step

Data Products

1. Results may be used for further analysis
    a. Examples:
        i. After performing correlation analysis “*showing that beer sales were higher when home team was winning more home games*”
            1. Sort values based on an ordinal metric - What was the price of the beers purchased (top 5)
        ii. Computer a derived value - The average spent on beer seen in the data.

Communication

1. Reports
2. Presentations
3. May need more than one kind:
    a. Executive Summary
    b. Detail Report on Analysis (showing the gritty details and math)