CS F441
Data Visualization

SUMANTA PATTANAIK

WEEK 2

Preparing and Familarizing with Data

- Acquisition:
  - Download, or
  - Manually gather, or
  - Extract from a Database, from a website or some document and Consolidate
- Examination:
  - Metadata
  - Completeness
  - Quality (errors, unusual data, …)
- Transforming for quality: Cleaning the data
- Transforming for analysis
- Dimension reduction

Data Examination: Metadata

- Information regarding a data set of interest.
- Provides information that can help in its interpretation
  - the format of individual fields within the data records.
  - the base reference point from which some of the data fields are measured,
  - the units used in the measurements,
  - the symbol or number used to indicate a missing value,
  - and the resolution at which measurements were acquired.
- Important in selecting the appropriate preprocessing operations, and in setting their parameters.

A Sample Metadata

- IMAGE DATA: Image files within this directory contain 2 dimensional views of a male cadaver, as collected for the National Library of Medicine's Visible Human Program.
- Anatomical Area:
  - TOTAL BODY Three sets (t1,t2,pd) of MRI male images.
  - Type image: GE MRI, Signa v5.2 Frame size: 256, 256
    - Specifies the image size (width, height) in pixels.
    - Pixel size: SEE FILE HEADER,, Specifies the pixel size (width, height, separation) in millimeters.
- Image format: GE 16 BITS, Compressed Unix compressed, use "uncompress [filename]" to restore.
  - Header size: 7900. The header block size in bytes.
  - Coordinate offset: NONE,NONE
  - If images files are cropped to remove empty pixels, these offsets are provided, in pixels, relative to a fixed coordinate plane.
  - …

Data Quality

- High-quality data needs to pass a set of quality criteria.
  - **Validity**
  - **Accuracy**
  - **Completeness**
  - **Consistency**
  - **Uniformity**

See:
https://en.wikipedia.org/wiki/Data_cleansing#Data_quality

Data Quality

- **Validity:** The degree to which the data conform to defined domain rules or constraints.
  - ***Data-Type Constraints:*** values in a particular column must be of a particular datatype, e.g., boolean, numeric, date, etc.
  - ***Range Constraints:*** typically, numbers or dates should fall within a certain range.
    - Ex: India lies to the north of the equator between 6° 44′ and 35° 30′ north latitude and 68° 7′ and 97° 25′ east longitude.
  - ***Mandatory Constraints***: certain columns cannot be empty.
  - ***Unique Constraints:*** a field, or a combination of fields, must be unique across a dataset.
  - ***Set-Membership constraints*:** values of a column come from a set of discrete values, e.g. enum values. For example, a person's gender may be male or female.
  - **Regular expression patterns:** text fields that have to be in a certain pattern. For example, phone numbers may be required to have the pattern "X-999-9999999".
  - **Cross-field validation:** certain conditions that span across multiple fields must hold. For example, a patient's date of discharge from the hospital cannot be earlier than the date of admission.

Data Quality

- **Accuracy:** The degree of conformity of a measure to a standard or a true value
  - **ex: 6** digit Pin Code in an address
  - Accuracy is very hard to achieve in the general case, because it requires accessing an external source of data that contains the true value: such "gold standard" data is often unavailable.

Data Quality

- **Completeness**: The degree to which all required measures are known.
  - Incompleteness is almost impossible to fix : one cannot infer facts that were not captured when the data in question was initially recorded.
    - Ex: User response

Data Quality

- **Consistency**: The degree to which the data is consistent, within the same data set or across multiple data sets.
  - Inconsistency occurs when two values in the data set contradict each other.
    - **Ex:** Age and Marital status

Data Quality

- **Uniformity**: The degree to which a set data measures are specified using the same units of measure in all systems.
  - ex: Weight or height data related to an international event (say Olympics)

Data Cleaning

- **Data Cleaning**: (also called **Data wrangling**)
  - the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database
    - identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and
    - replacing, modifying, or deleting the dirty data

Data Cleaning

- Main Steps of Data Cleaning:
  - **Inspection:** Detect unexpected, incorrect, and inconsistent data.
  - **Cleaning:** Fix or remove the anomalies discovered.
  - **Reporting:** A report about the changes made and the quality of the currently stored data is recorded.
- Verify and Repeat

Source: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Data Cleaning

- **Inspection:** Detect unexpected, incorrect, and inconsistent data.
  - **Data profiling: Generate** A summary statistics about the data
    - Is the data column recorded as a string or number?.
    - How many values are missing?
    - How many unique values in a column, and their distribution?
  - **Statistical Analysis and Visualization of Distribution**
    - mean, standard deviation, mean, range, or quantiles.

Source: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Data Cleaning

- **Cleaning:** Fix or remove the anomalies discovered.
  - **Missing Values**:
    - Drop Row: missing values in a column rarely happen and occur at random
    - Drop Column: most of the column's values are missing, and occur at random
    - Assign Value (impute): Mean/Median value or prediction using Linear regression.
    - Do nothing but Flag

Source: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Data Cleaning

- **Cleaning:** Fix or remove the anomalies discovered. (Continued)
  - Remove Irrelevant/Duplicate data
  - Convert data type: ex: "Make sure numbers are stored as numerical data types"
  - Massage string data: Fix typos, remove extra white space, Capitalize etc..
  - Standardize data: Same unit of measurement (ex: M or cm or mm), European or USA version (date format)

Data Cleaning

- **Reporting:** A report about the changes made and the quality of the currently stored data is recorded.

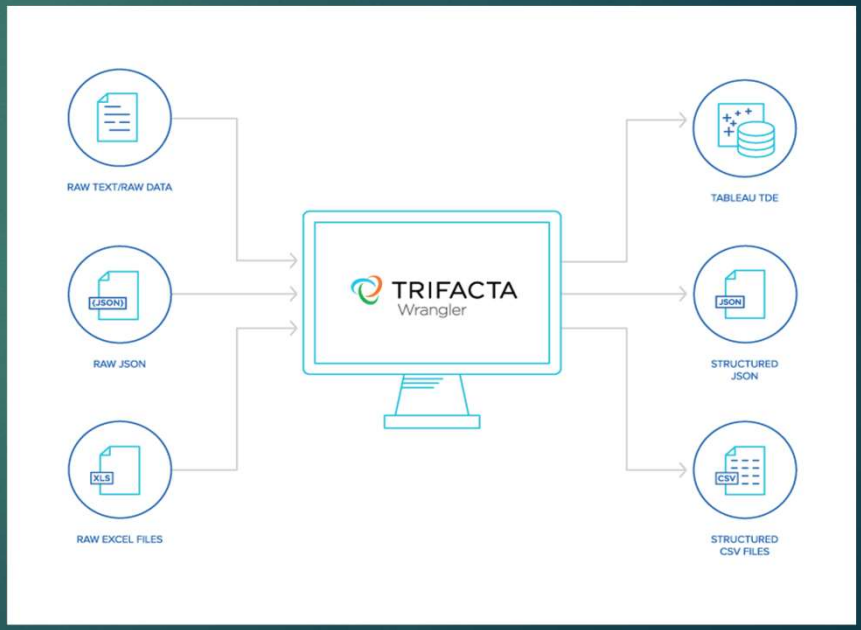
Source: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Data Cleaning

- Tools:
  - Drag and Drop Tools
  - Script Based

Drag-and-Drop app for Data Cleaning

- Trifacta Wrangler: Messy Data Accepted
  https://www.trifacta.com/products/wrangler/
- Originally Stanford/Berkeley Data Wrangler project



Drag-and-Drop App for Data Cleaning

OpenRefine: http://openrefine.org/

Originally Google Refine.

Links to other free/commercial Data cleaning tools can be found from https://en.wikipedia.org/wiki/Data_cleansing

Based on Scripting Languages

- **Functions in Python Pandas**
- Javascript Array functions and additional Tools in D3
- Functions in R Tidyverse

Visualization Tools

Two Categories of Tools

- **Drag-and-Drop Tools:**
- **Based on Scripting Languages**

Two Categories of Tools

- **Drag-and-Drop Tools:**
  - Rely on a Graphical User Interface
  - Make assumptions about what you may like to do
    - Ex: You may draw SALES to Y-axis and DATE to the X-axis. The tools assumes that you are interested in graphing total sales per month.
  - Examples: Tableau, PowerBI
- **Based on Scripting Languages**
  - You decide what and how you want to create visualization.

Next Monday's Quiz

- Topics
  - Data Quality
  - Python Pandas related

Tableau

#1 most-used Business Intelligence software tool

- **Pro**: Simple and easy to be a beginner user
  - connect to your data source, whether an Excel file, a database connection, or any of the dozens of other connection options
  - drag the variable names you want onto a graph object (a "sheet") and customize as you see fit.
  - combine sheets into a *dashboard* in whatever configuration you like, and get creative with parameters, filters, or other customization options.
  - Allows public and private hosting
  - Free license for students
- **Cons**: "A minute to learn, a lifetime to master!"
  - Less flexible

Microsoft Power BI

Picking up to be the most successful analytic and business intelligence platform

- **Pro**: Like Tableau Simple and easy to be a beginner user
  - connect to your data source, whether an Excel file, a database connection, or any of the dozens of other connection options
  - drag the variable names you want onto a graph object (a "sheet") and customize as you see fit.
  - Easy integration with analysis
- **Cons**:
  - Development time can be long
  - Expensive

Two Categories of Tools

- **Drag-and-Drop Tools:**
- **Based on Scripting Languages**
  - Ex: matplotlib and Plotly in Python; D3, Observable Plot, Plotly in Javascript, ggplot, Plotly in R, …
  - Better control on the result, but you have to be explicit about what you want.

This class will use Scripting Languages

- **Python**: Scripting language
- **Plotly** and **matplotlib** : For data visualization

Matplotlib

- Matplotlib is a popular Python library for creating visualizations.
  - matplotlib.pyplot: This is the primary module used for creating visualizations. It provides a simple interface for creating plots and charts
- Pros:
  - Versatile and Customizable
  - Wide Adoption: Being one of the oldest and most popular plotting libraries in Python, Matplotlib is widely adopted and often considered the starting point for many data visualization tasks.
- Cons:
  - Steep Learning Curve.
  - Limited interactivity
  - Its default styles might not always produce the most aesthetically pleasing plots compared to some other libraries.
  - As in any scripting-based visualization tool, a lot of code to write.
- We will mostly use Pandas.plot and Seaborn: Libraries developed on the top of MatplotLib. Reduce coding load.

https://matplotlib.org/

Plotly

- Open Source graphics library for creating interactive, publication-quality graphs. It has a concise and (hopefully) memorable functions to foster fluency
- Pros:
  - Interface is available to Python, R, Javascript, Matlab, Julia
  - Great support for interaction
  - Beautiful visualizations
- Cons:
  - As in any scripting-based visualization tool, a lot of code to write.
- We will mostly use Plotly.express: A library developed on the top of Plotly. Reduces coding load.

https://plotly.com/graphing-libraries/