VIRTUAL APPLIED DATA SCIENCE TRAINING INSTITUTE

VADSTI 2021 Theme: “Data Science Approaches to Better Understand Clinical and Genomic Informatics”

Module 2: Data Exploration & Visualization

02/18/2020

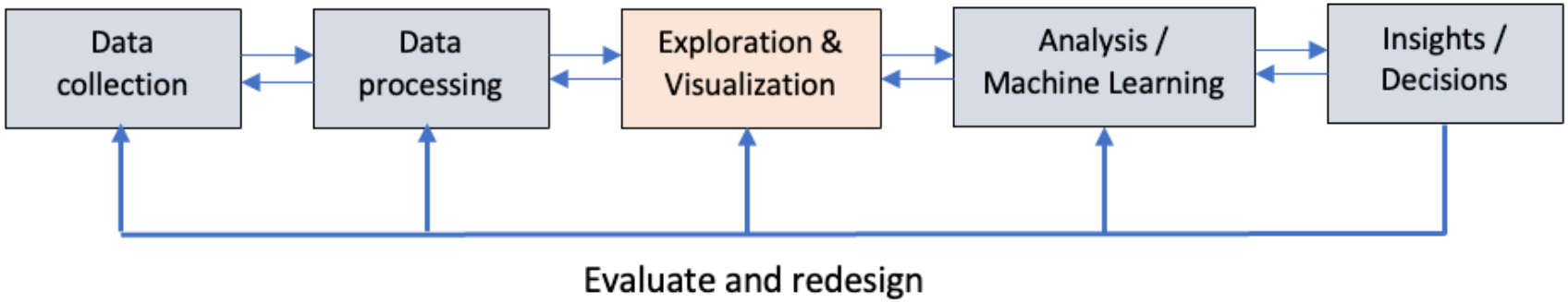
Anas Belouali

“If you don't reveal some insights soon, I’m going to be forced to slice, dice, and drill!”

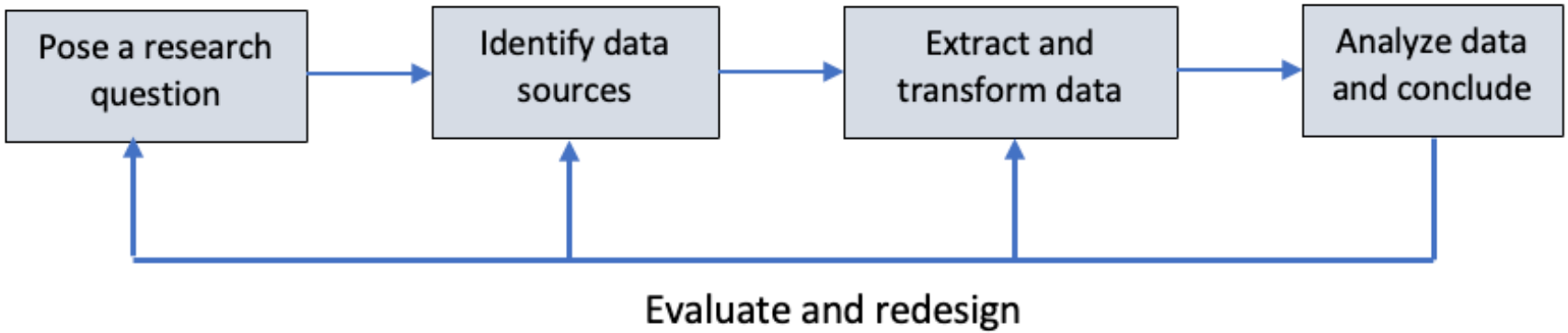
Outline

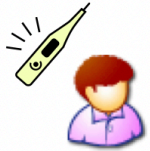
- Intro to clinical data and challenges of clinical data mining
- Basics of data exploration and visualization
- Data types and visualization types
- Hands-on

Data science project lifecycle

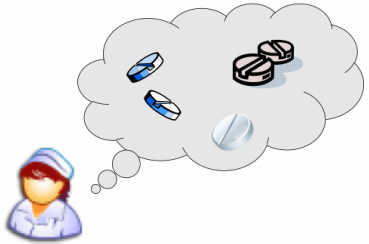


Clinical data mining workflow

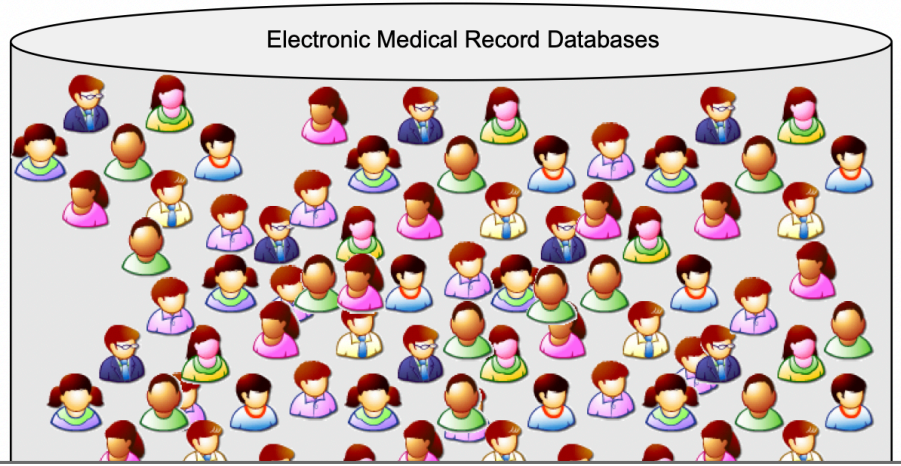




Patient



Clinician

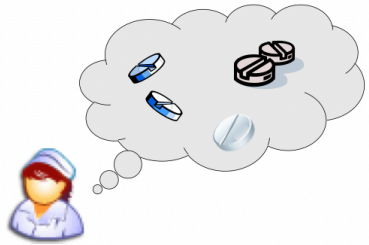


Thousands or Millions of Patients

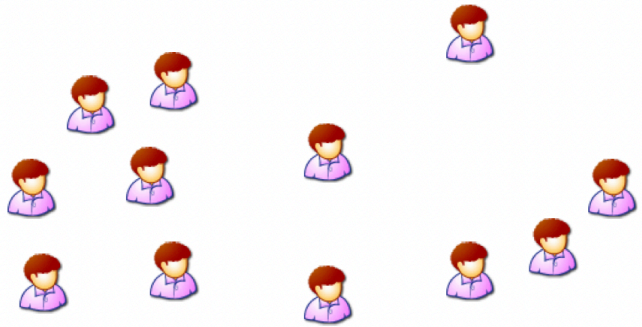
- 10+ Years of Data Per Patient
- Tens of Thousands of Features
  - Demographics
  - Diagnoses
  - Labs
  - Procedures
  - Claims
- Unstructured Physician Notes

Credit @Adam Perer, https://www.fda.gov/

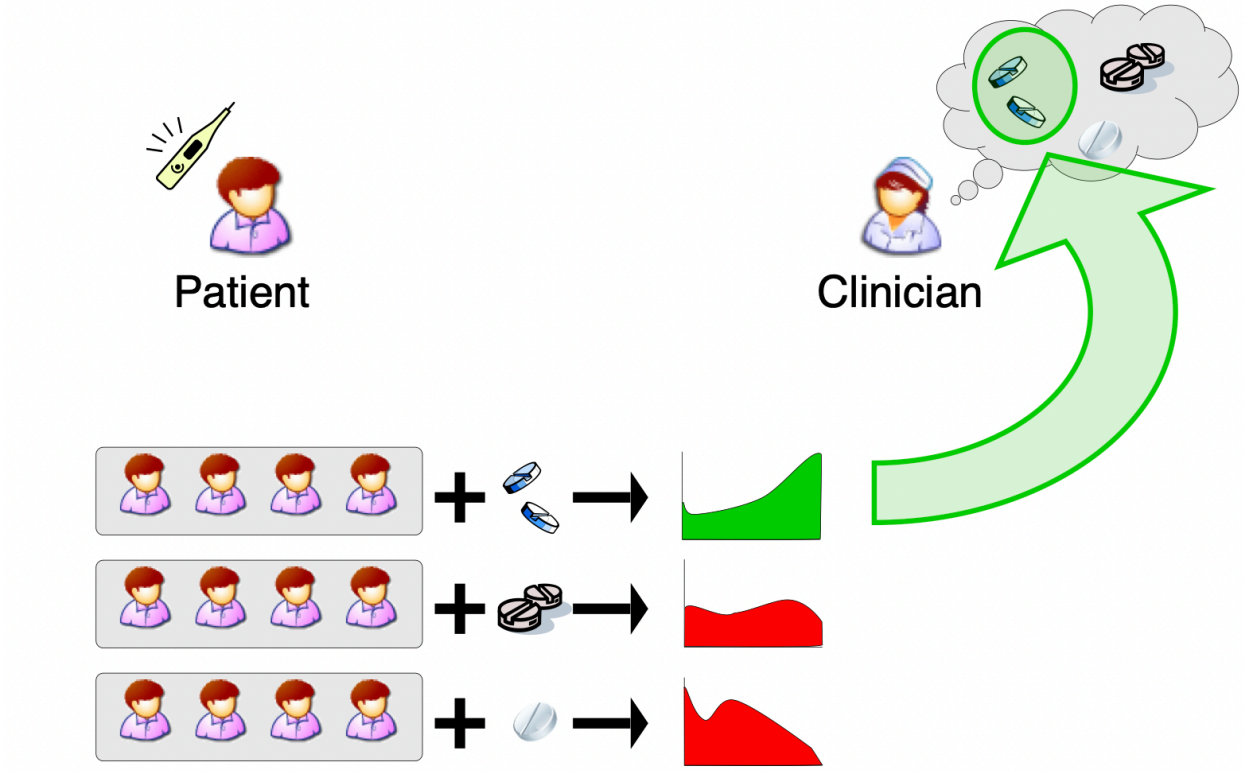
Patient



Clinician

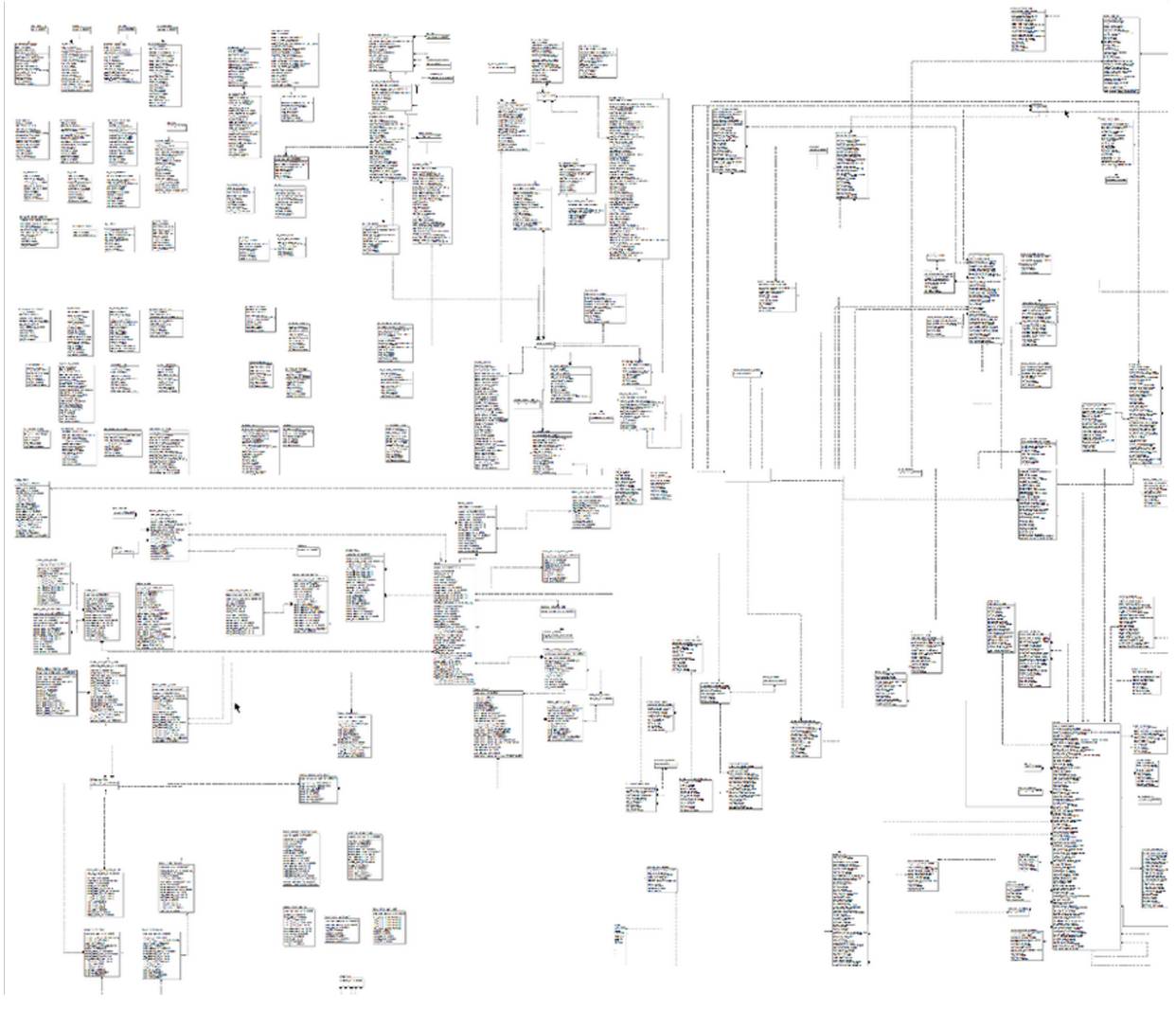


Credit @Adam Perer, https://www.fda.gov/



Credit @Adam Perer, https://www.fda.gov/

EHR Database Schema (Cerner)

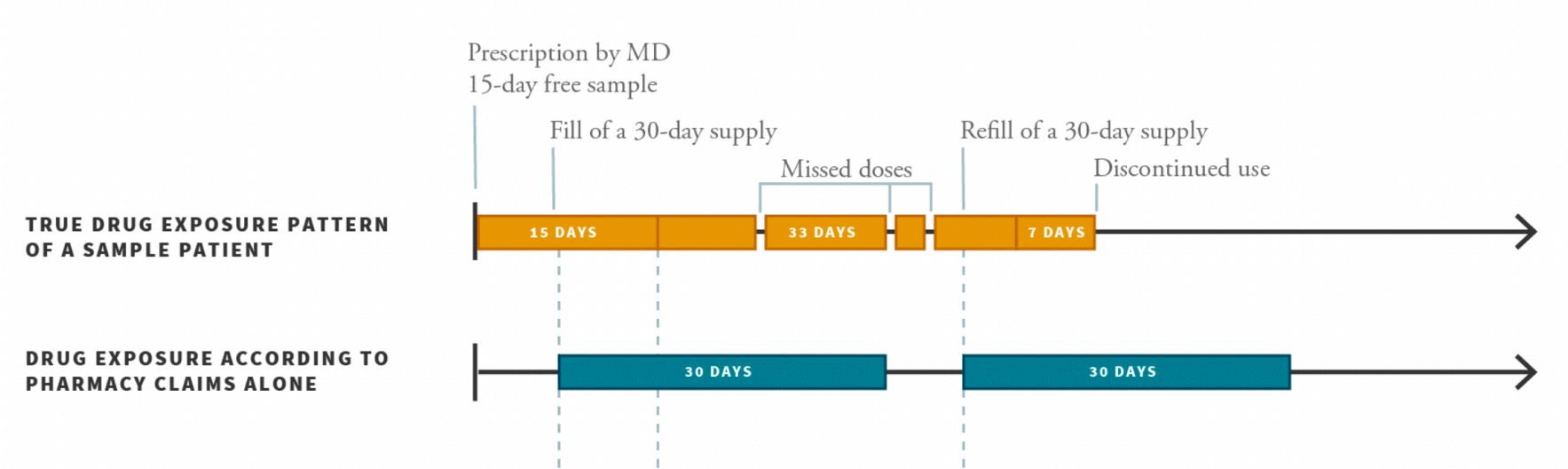


Challenges of data in healthcare

- Messy
- Incomplete
- Fragmented
- Ever-changing Data
- Biased
- Privacy and Security Regulations

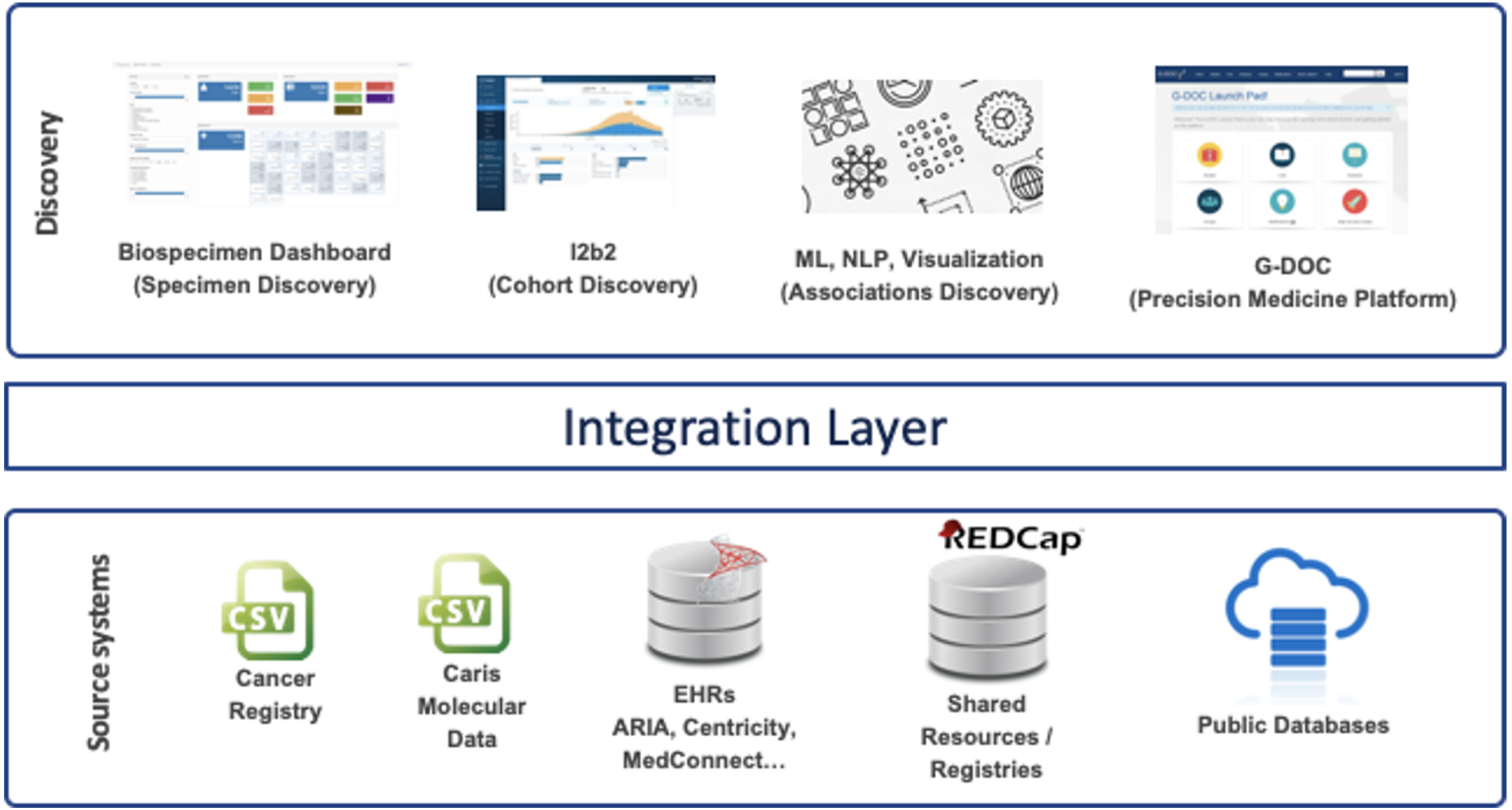
“How am I supposed to analyze this?”

Example of exposure misclassification



https://www.coursera.org/learn/introduction-clinical-data

Data Landscape for research at GUMC



Common Data Models in Healthcare

- A Common Data Model allows for systematic analysis of disparate observational databases.
- Transform data into a common format
- Some CDMs use a common representation (terminologies, vocabularies, coding schemes)
- Perform systematic analyses using a library of standard analytic tools
- Examples distributed health data networks: *i2b2 (ACT), OHDSI, Trinetx, PCORnet, FDA Adverse Event Reporting System (FAERS)...*

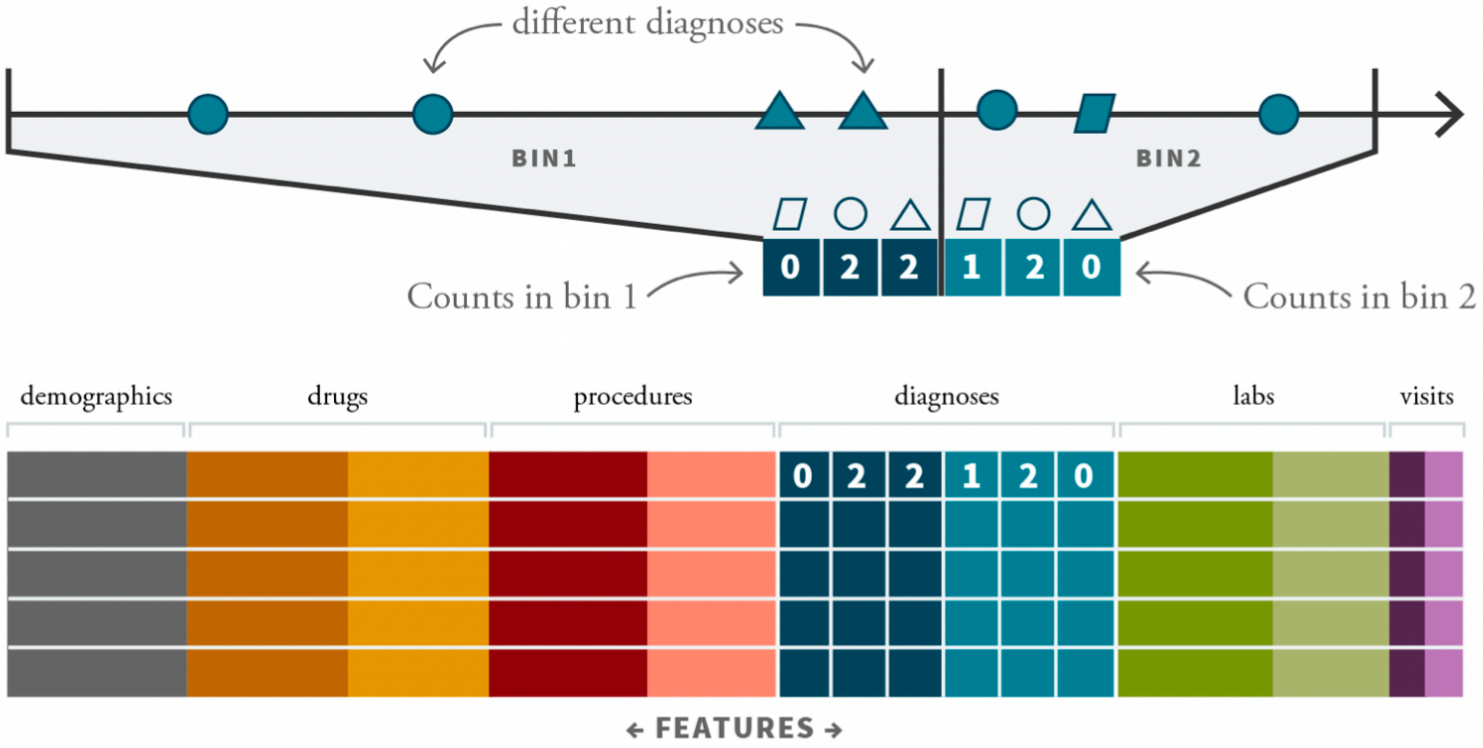
Data in healthcare is heterogenous

- **Structured:** consistent organization, tabular format
- **Unstructured data:**
  - *Clinical text:* different from natural language; contains several acronyms
  - *Images:* MRI, X-rays, CT..
  - *Signals:* measurements coming from a sensor, usually at regular intervals (e.g. EKG)

Information as a patient timeline

https://www.coursera.org/learn/introduction-clinical-data

Patient feature matrix

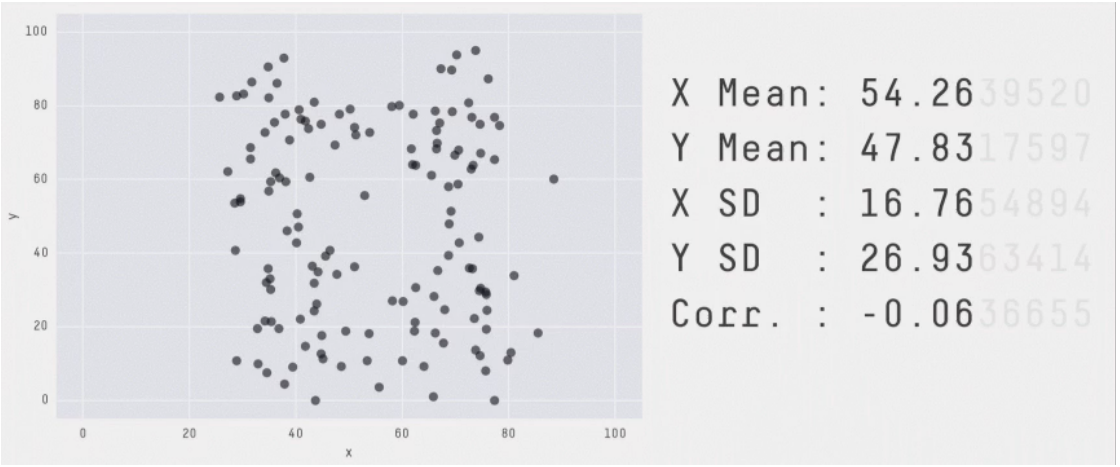


https://www.coursera.org/learn/introduction-clinical-data

Data Exploration Visualization

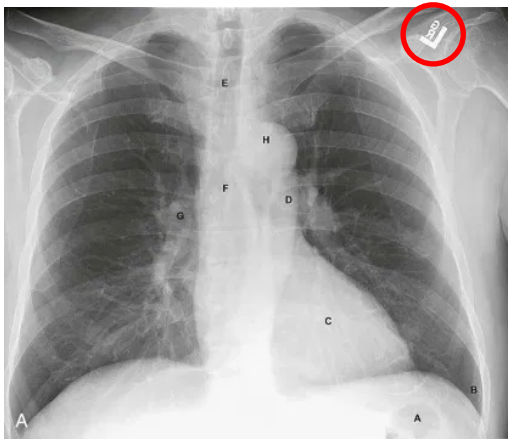
- **Two type of visualization:**
  - **Data exploration visualization:** Figuring out what is true
  - **Data presentation visualization:** storytelling with the data to convince other people something is true
- **Data exploration** is much broader than just visualization.
- **Key motivations:** select the right tools for analysis; recognize patterns using human abilities; recognize biases and errors…
- **Related to Exploratory data analysis (EDA):** an approach to analyzing datasets to summarize their main characteristics, often with visual methods. Created by statistician John Tukey.

Always visualize your data



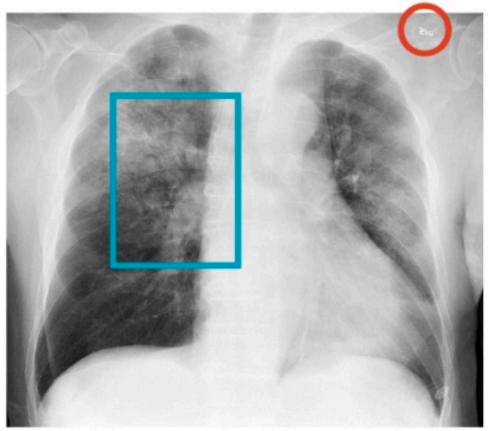
[Source: https://twitter.com/JustinMatejka/status/770682771656368128 Credit: @JustinMatejka, @albertocairo]

Always visualize your data



Hospital A

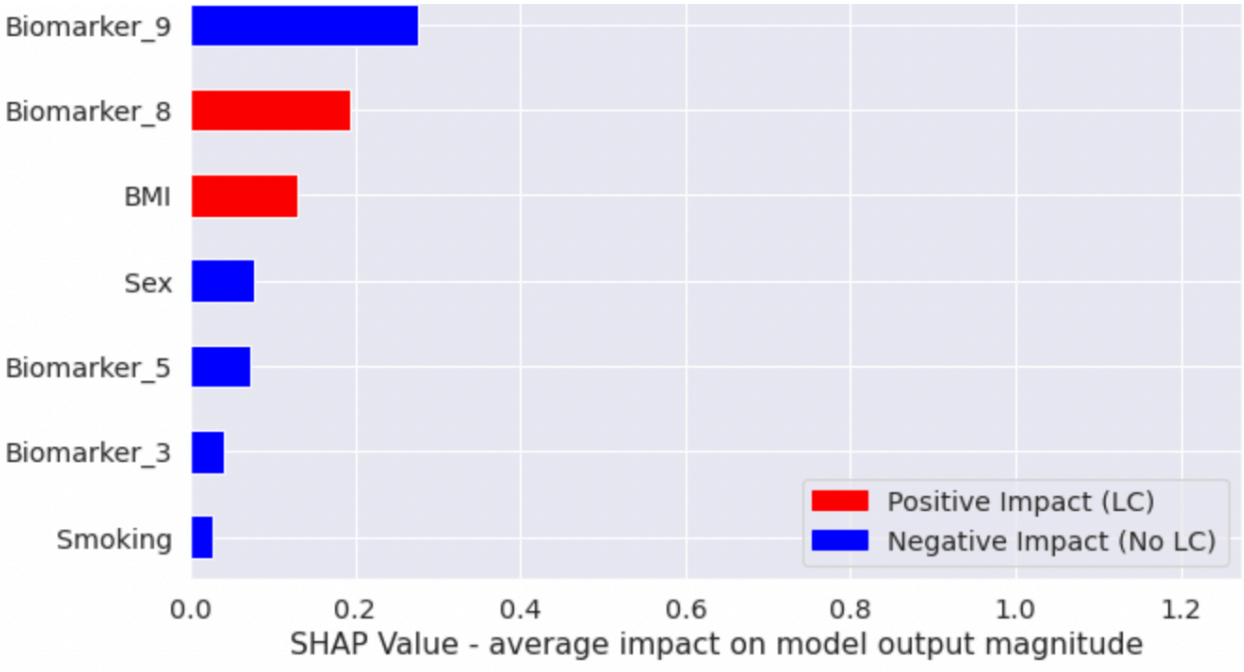
1% prevalence of pneumonia



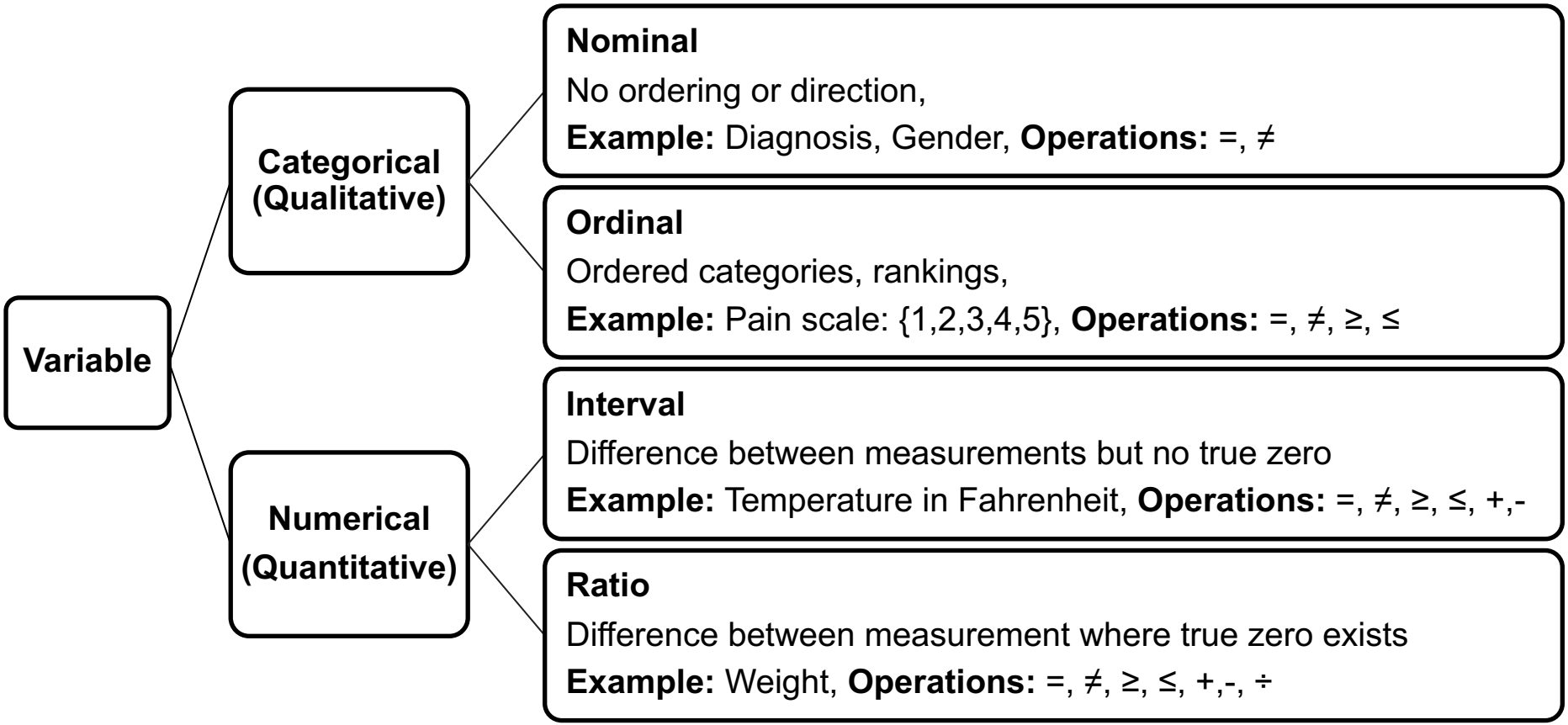
Hospital B

38% prevalence of pneumonia

Always visualize your results



Understanding your data types



Visualization Types

We are going to map the visualization types to the type and dimensionality of the underlying data.

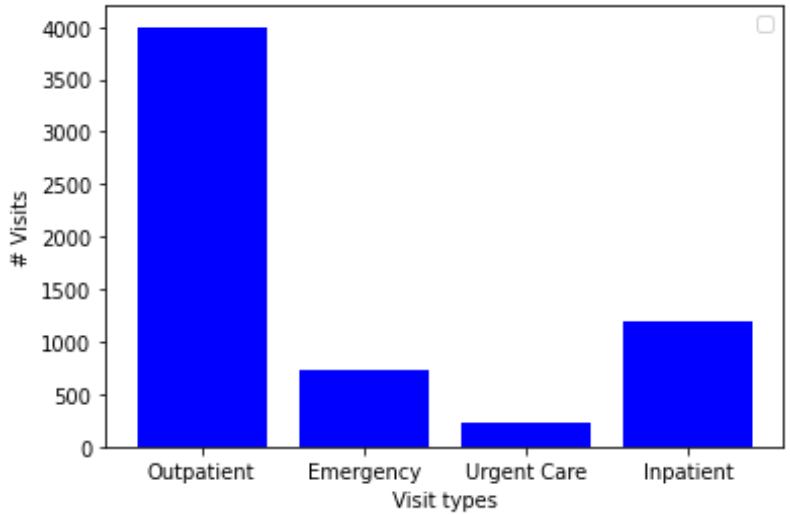
Visualization Types

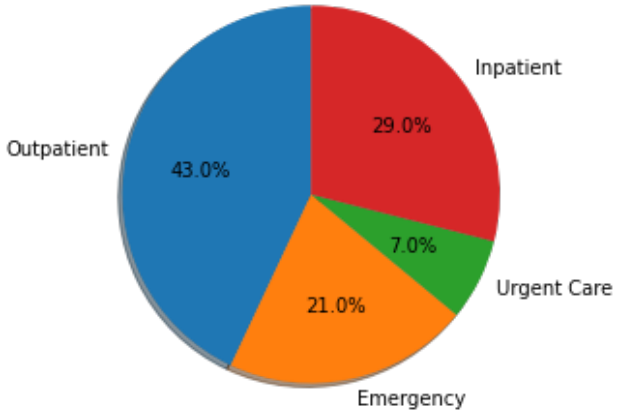
- 1D: bar chart, pie chart, histogram
- 2D: scatter plot, line plot, box plots, heatmaps
- 3D(+): scatter matrix
- **...** (Not an exhaustive list)

1D data – Bar Plots (or pie charts)

Categorical data:

Nominal	Ordinal	Interval	Ratio
Yes	Yes	No	No

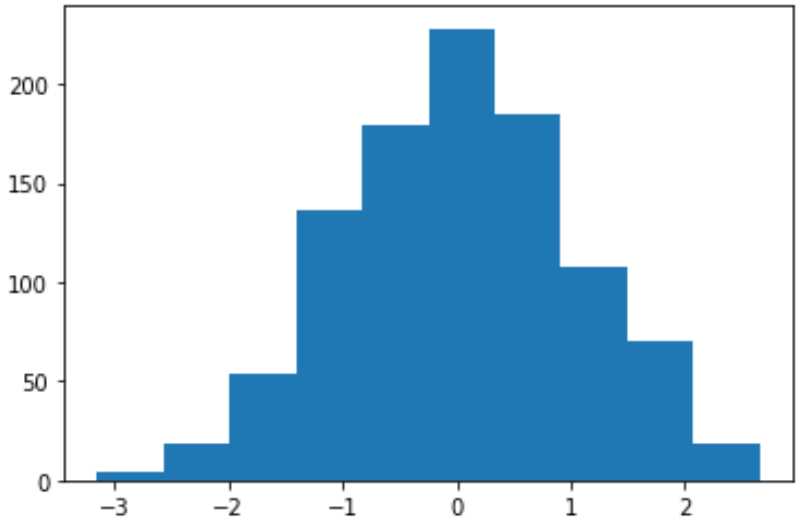




1D data – Histograms

Numerical data:

Nominal	Ordinal	Interval	Ratio
No	No	Yes	Yes

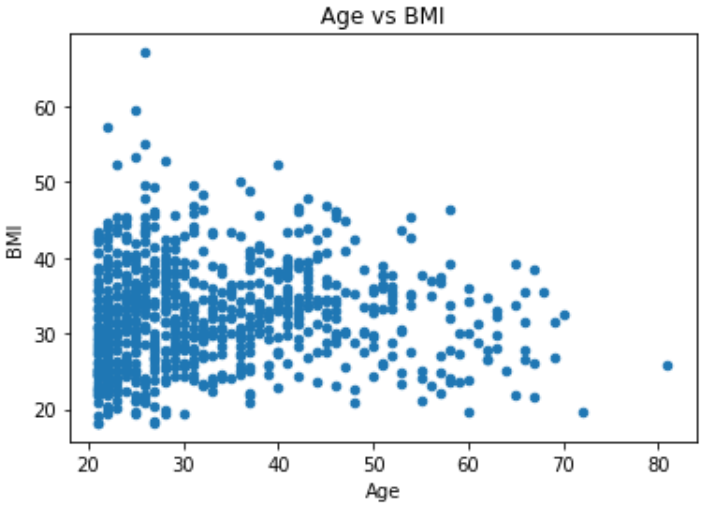


2D data – Scatter Plot

Both

Numerical data:

Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1	No	No	Yes	Yes
Var 2	No	No	Yes	Yes

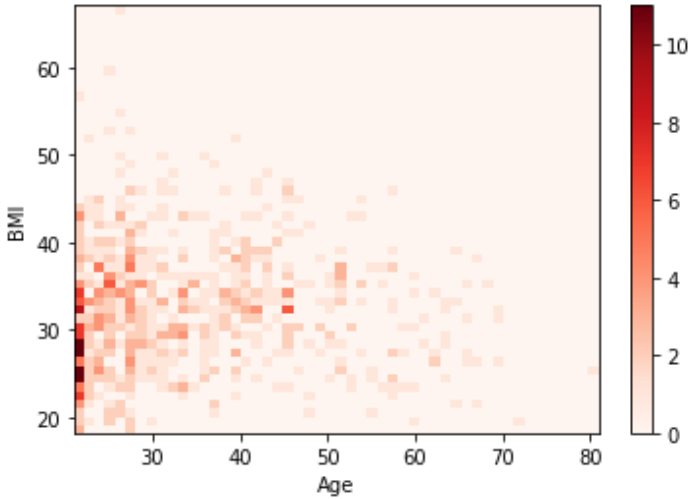


2D data – Density plot (2D histogram)

Both

Numerical data:

Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1	No	No	Yes	Yes
Var 2	No	No	Yes	Yes

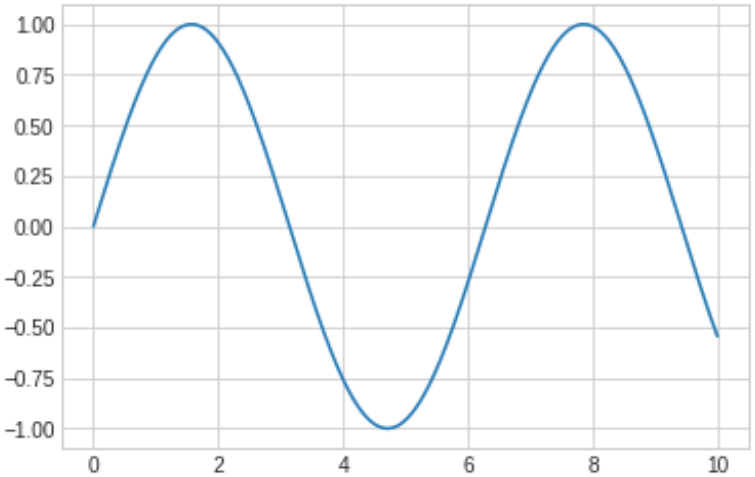


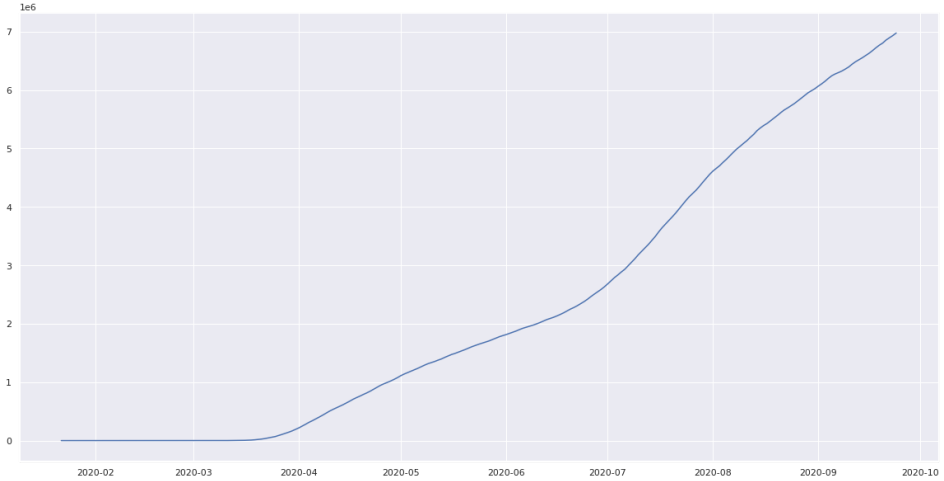
2D data – Line Plot

Both

Numerical data:

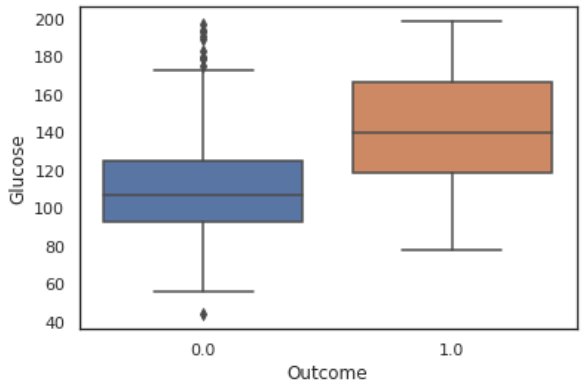
Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1	No	No	Yes	Yes
Var 2	No	No	Yes	Yes

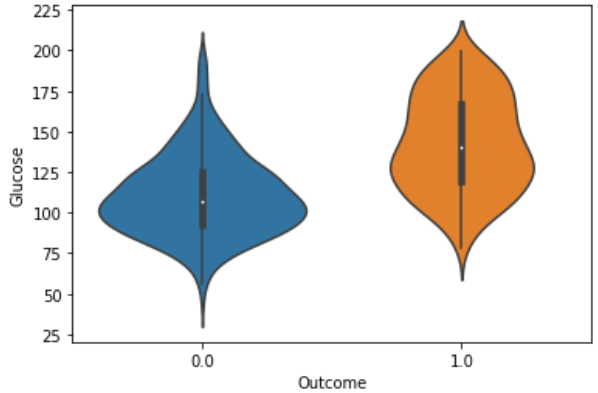




2D data – Box plots / Violin plots

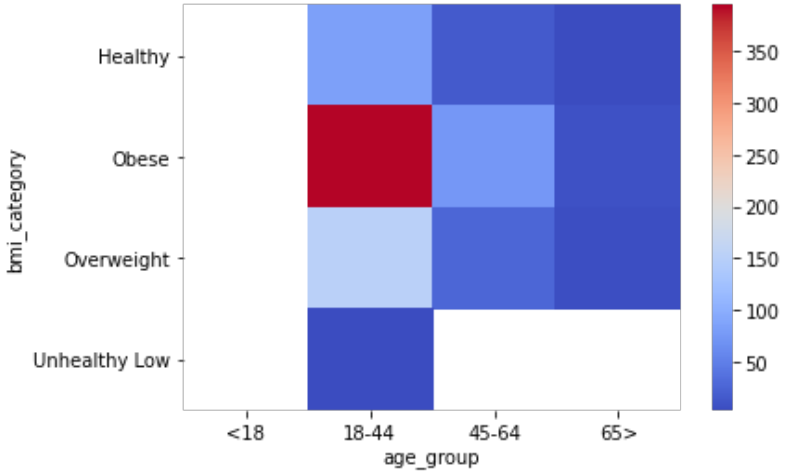
Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1 (categorical)	Yes	Yes	No	No
Var 2 (numerical)	No	No	Yes	Yes

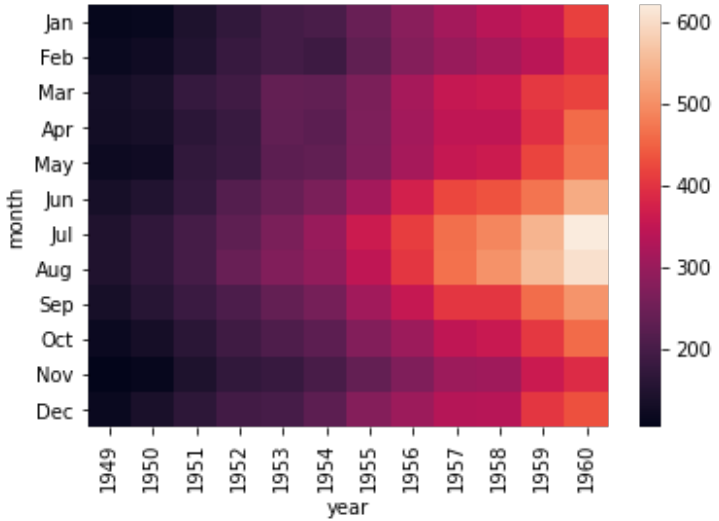




2D data – Heatmap (matrix)

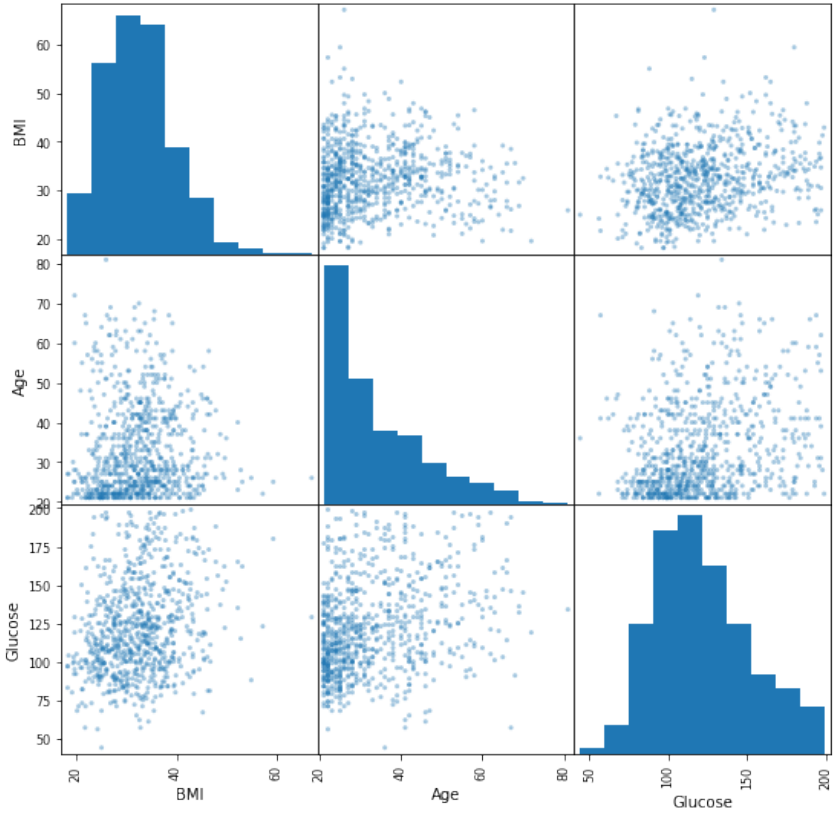
Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1 (categorical)	Yes	Yes	No	No
Var 2 (categorical)	Yes	Yes	No	No





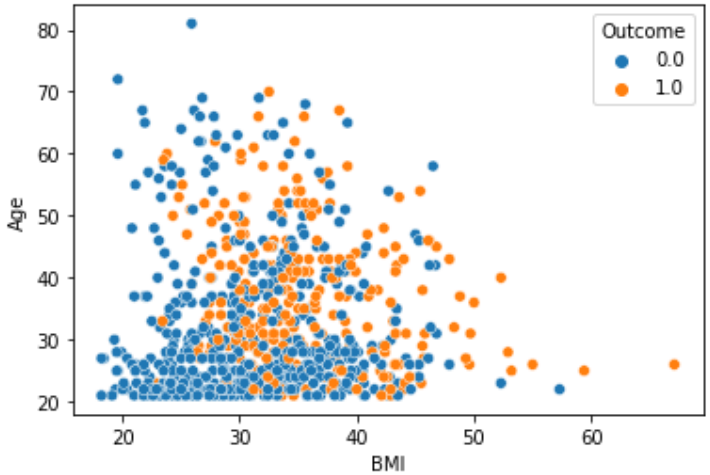
3D data – Scatter plot matrix

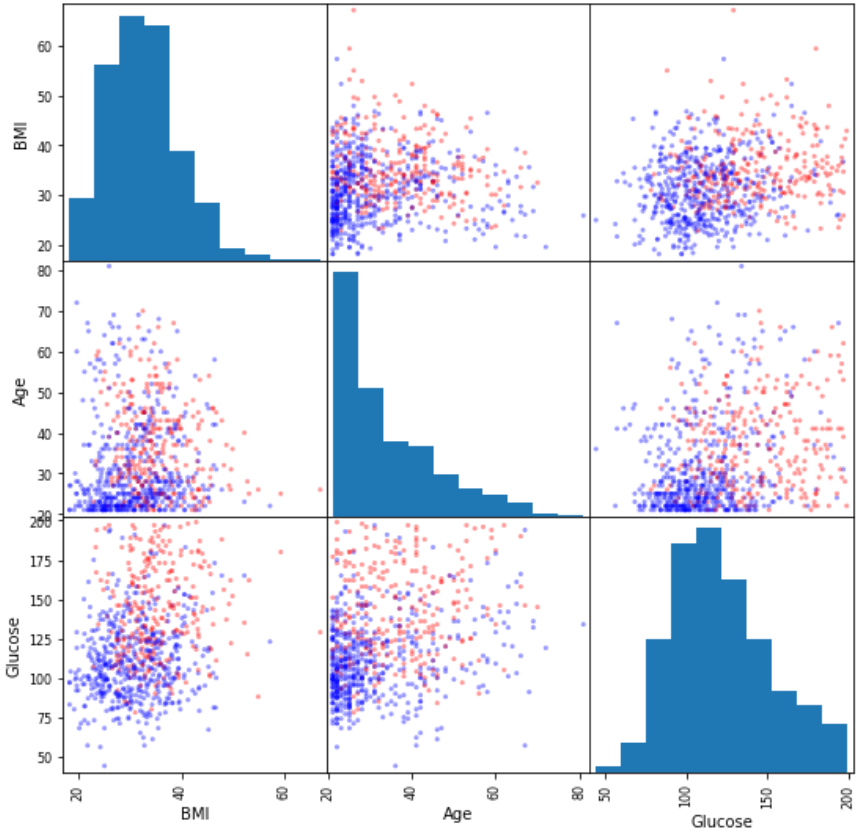
Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1 (numerical)	No	No	Yes	Yes
Var 2 (numerical)	No	No	Yes	Yes
Var 3 (numerical)	No	No	Yes	Yes



3D+ data – Scatter plot matrix

Dimensions	Nominal	Ordinal	Interval	Ratio
Var 1 (numerical)	No	No	Yes	Yes
Var 2 (numerical)	No	No	Yes	Yes
Var 3 (numerical)	No	No	Yes	Yes
Var 4 (numerical)	Yes	Yes	No	No





Python Libraries: Visualization

Few popular plotting libraries:

- **Matplotlib:** low level, provides lots of freedom
- **Pandas Visualization:** easy to use interface, built on Matplotlib
- **Seaborn:** high-level interface, great default styles
- **ggplot:** based on R’s ggplot2, uses Grammar of Graphics
- **Plotly:** can create interactive plots

Python Libraries: Data manipulation, scraping, ML…

- Pandas
- NumPy
- SciPy
- SQLAlchemy
- NLTK

- BeautifulSoup
- Scrappy

- Scikit-learn
- TensorFlow
- Keras
- PyCaret
- PyTorch

Hands-on