Data Science Training – Introduction and Data Visualization

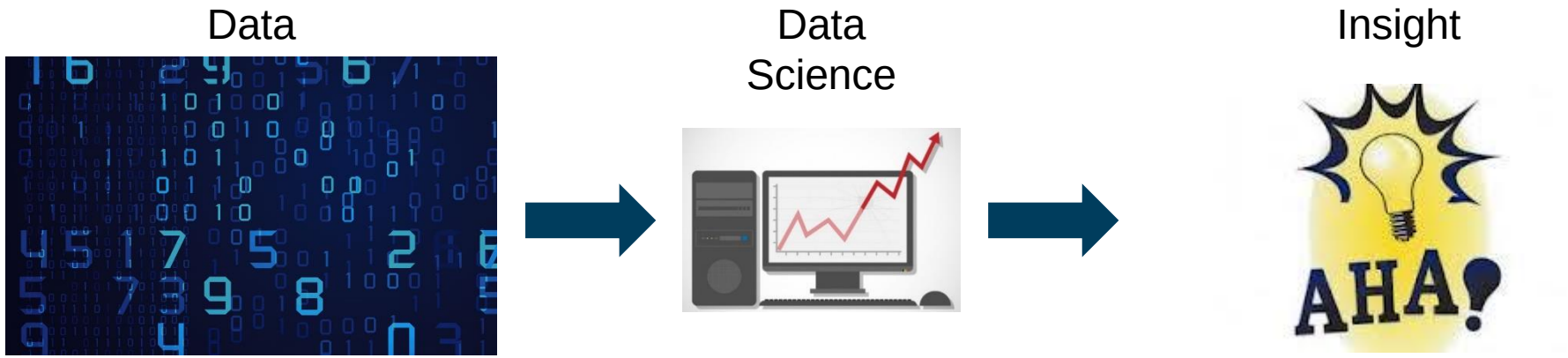
Presenter: Max Cohen

Adviser: Dionisios G. Vlachos

6/23/2020

What is Data Science?

- Field of study focusing on extracting insights from data
- Common related phrases:
  - Machine learning, statistics, data analytics
- Extremely interdisciplinary field:
  - Mathematics, statistics, computer science, information science
  - **Domain knowledge**
- Set of tools and techniques to learn from data



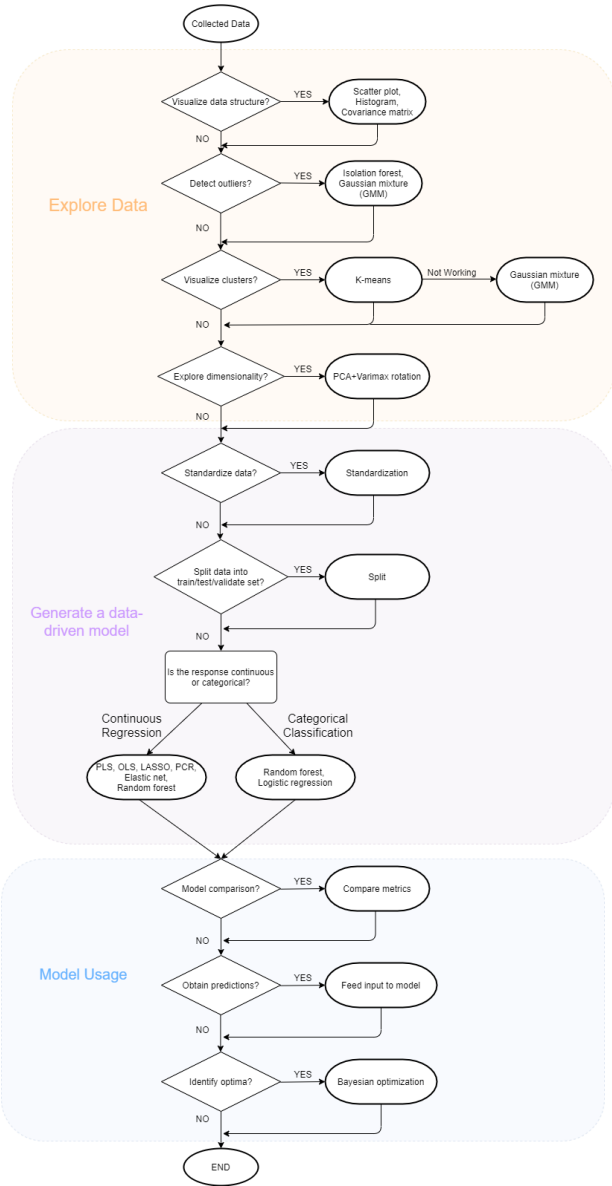
How to Apply Data Science?

- All tools can be misused
- Created a flowchart to provide structure for applying techniques
- **Goal of these training sessions: teach you how to use the flowchart and listed techniques**
  - Provide foundation

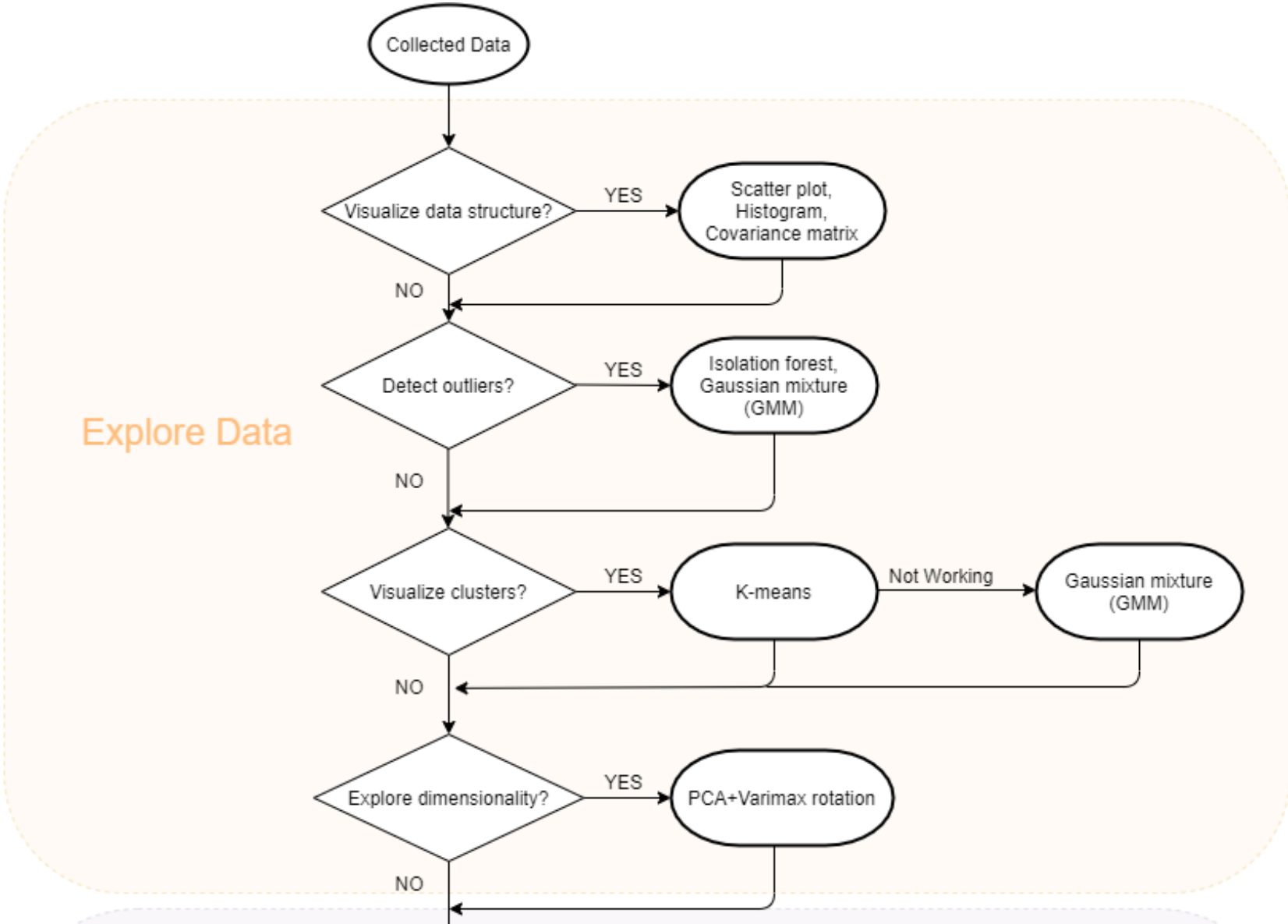
Explore Data

Generate a data-driven model

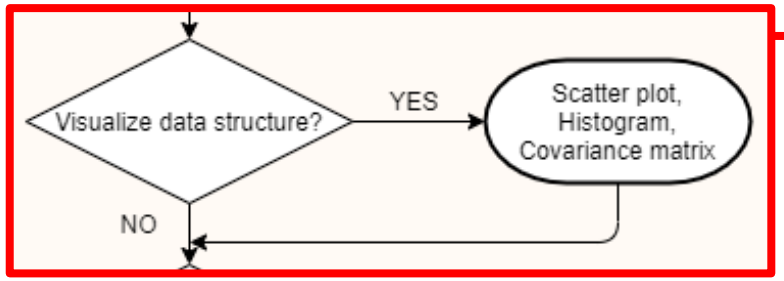
Model Usage

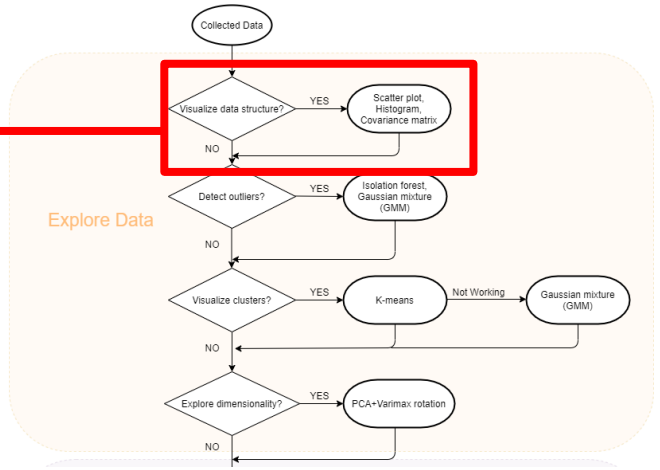


Explore Data

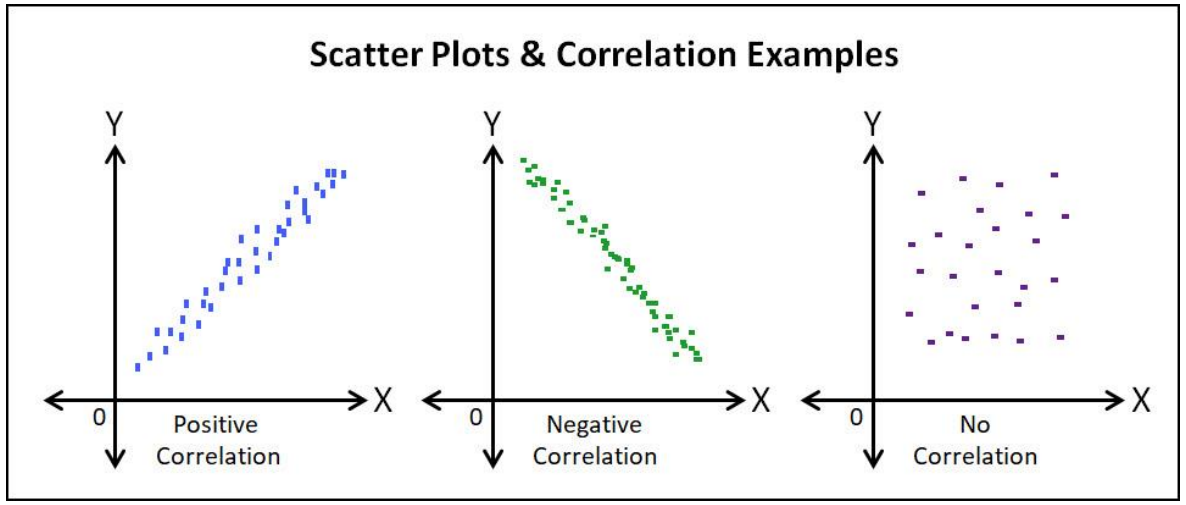


Explore Data – Visualize

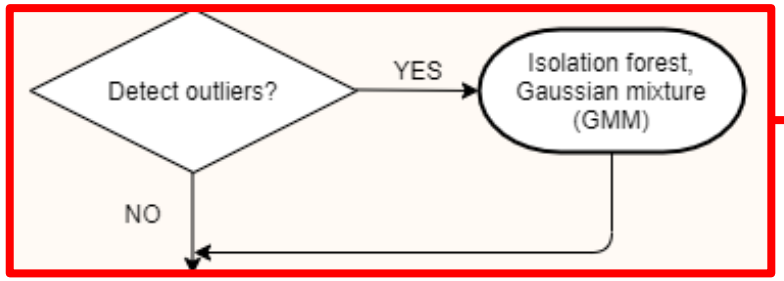


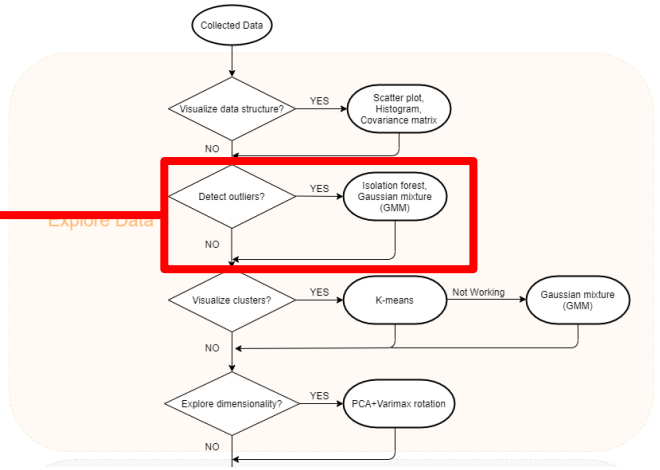


- Obtain visual understanding of data trends
- Identify possible concerns before modeling

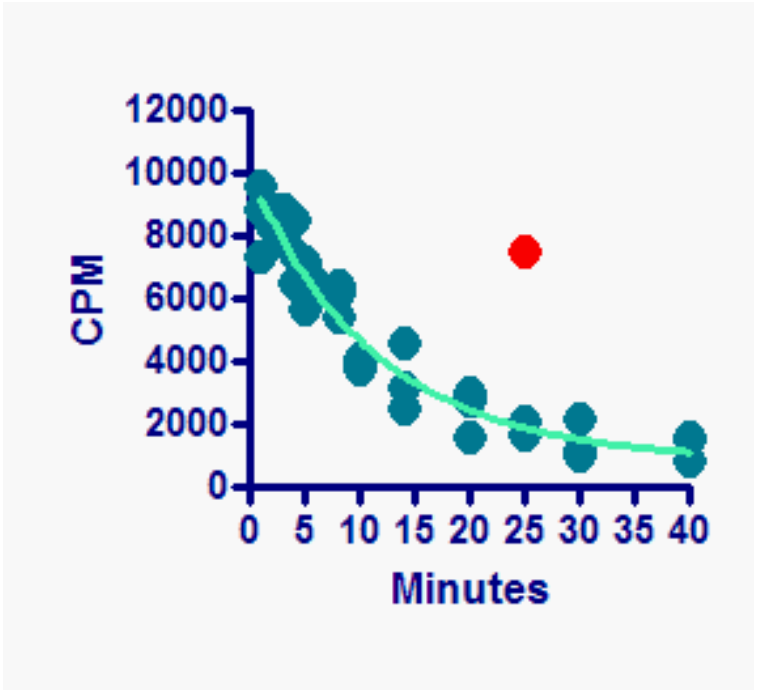


Explore Data – Outliers

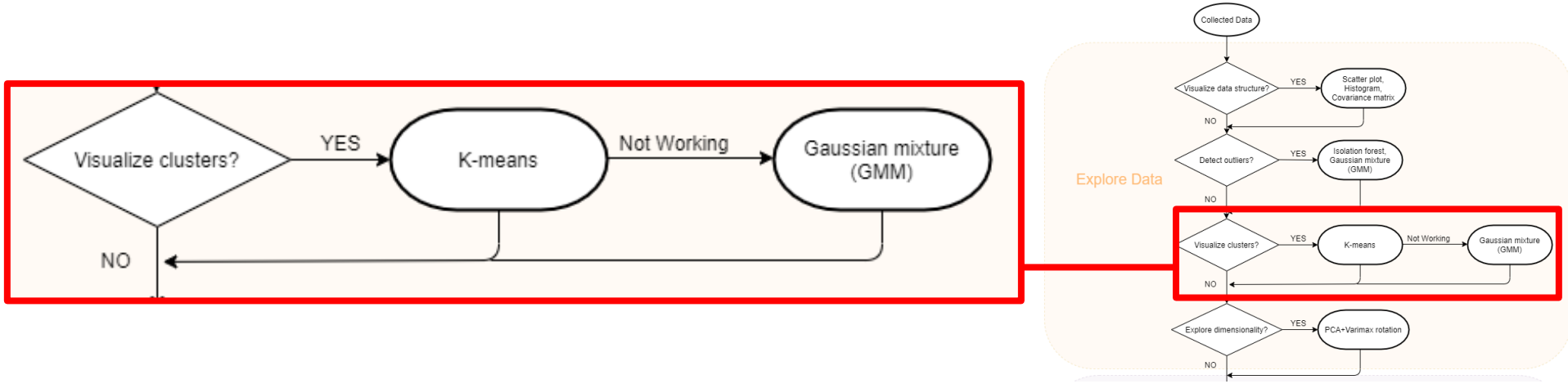




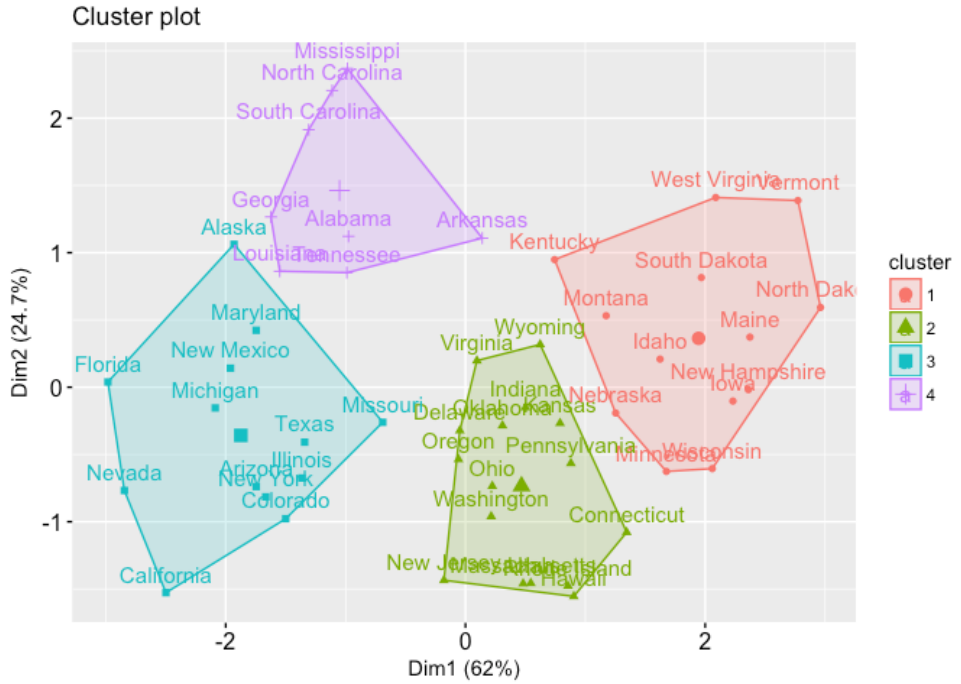
- Outlier removal
  - Improve training for models
- Outlier identification and focus
  - Determine what causes outliers, if desirable



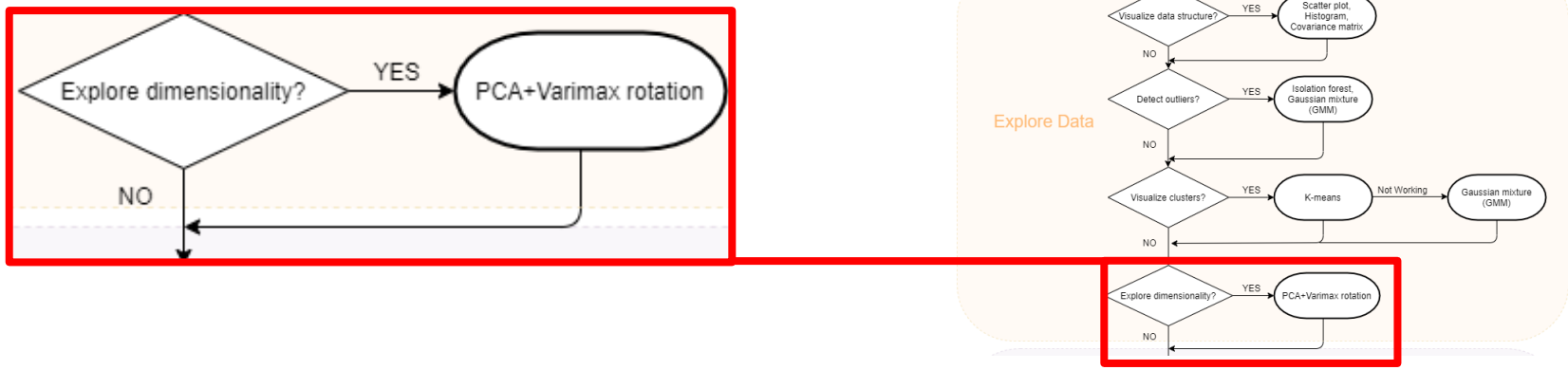
Explore Data – Clusters



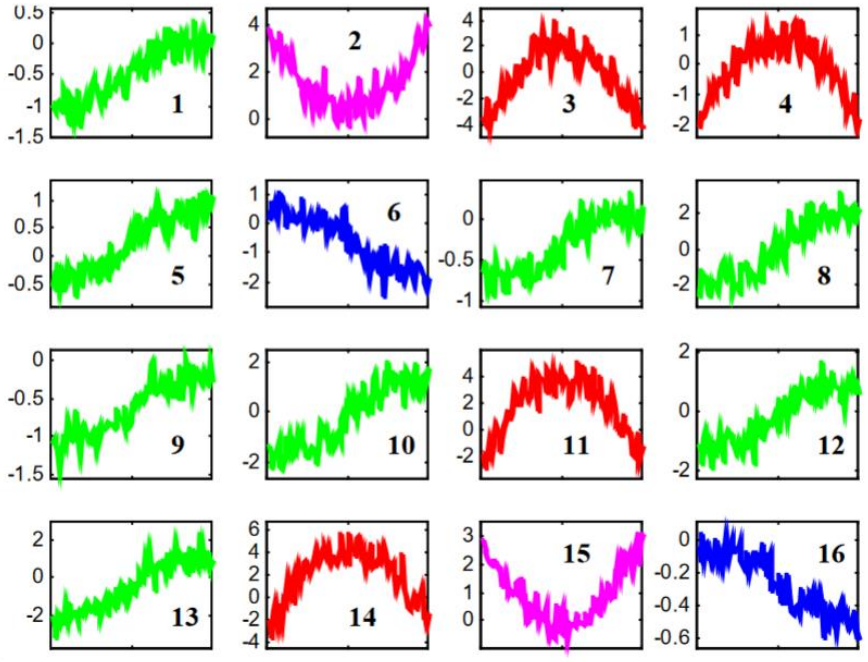
- Identify clusters in data
- Clusters reveal trends and subsets in data
  - Can help discriminate subsets to model



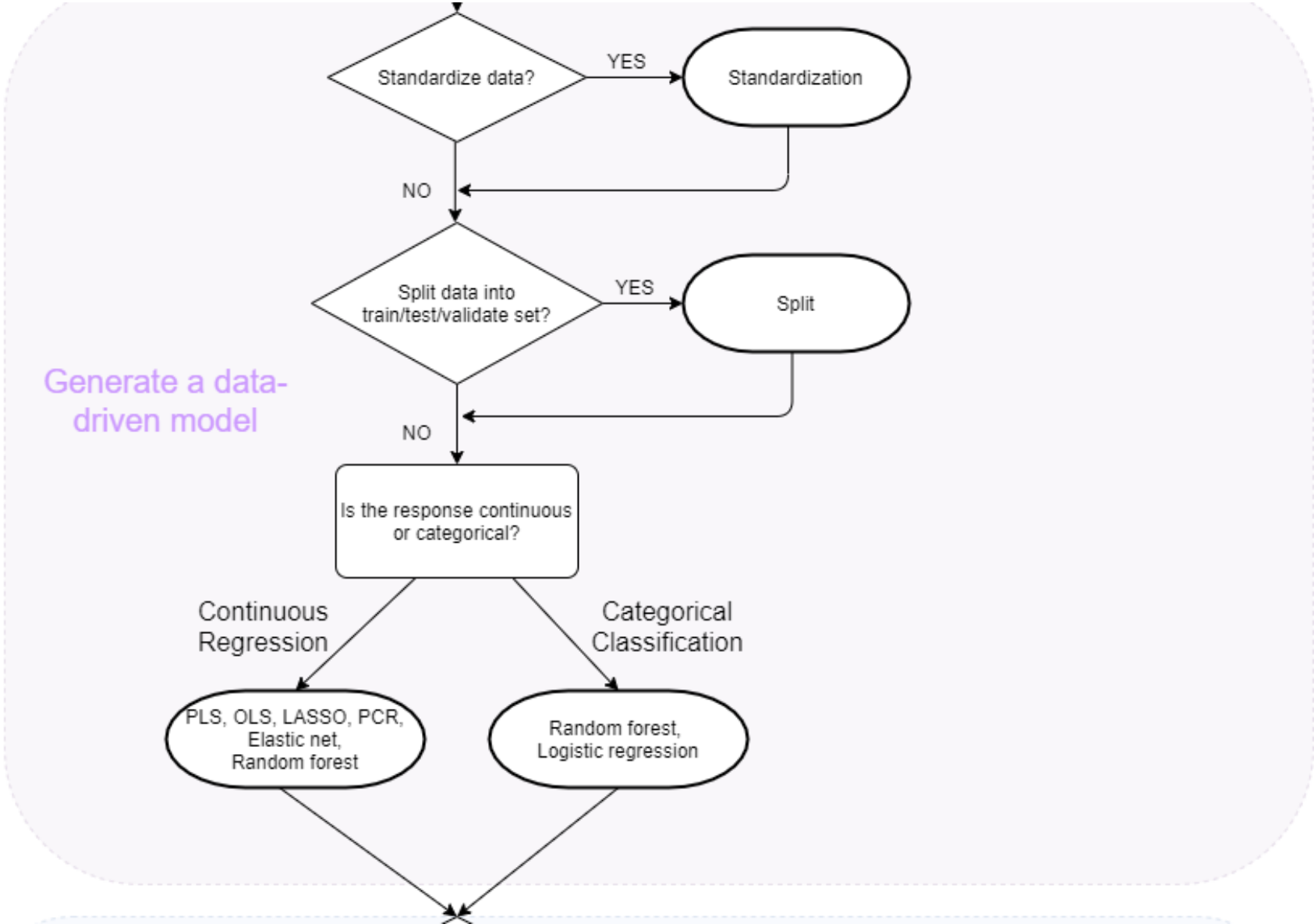
Explore Data – Dimensionality



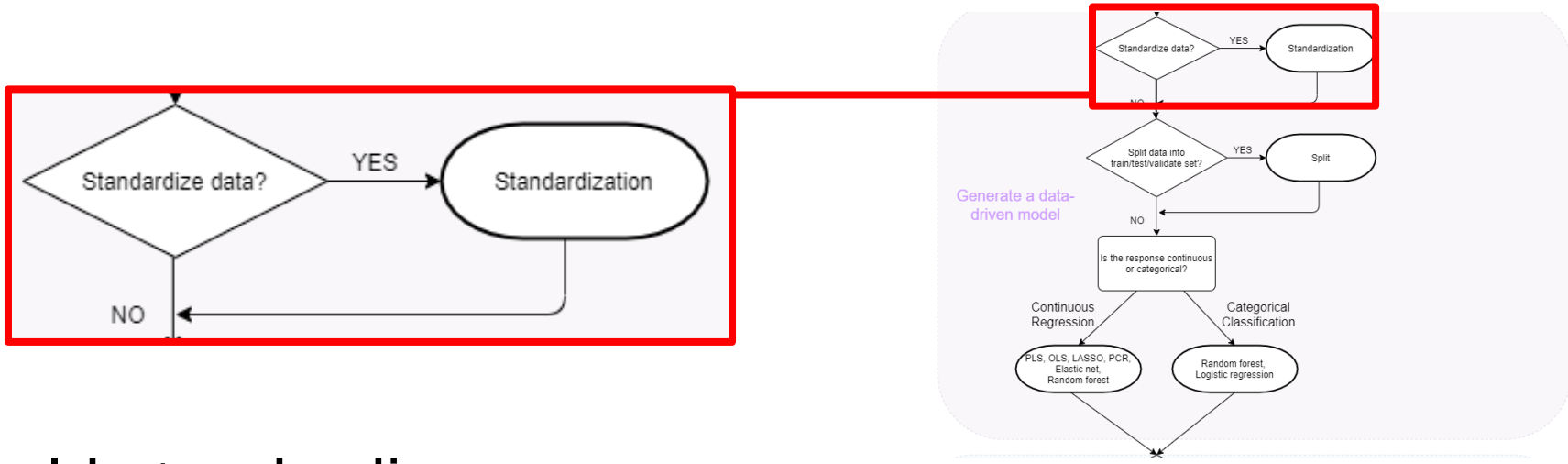
- Determine how many intrinsic dimensions exist in data
  - Identify trends in data
  - Reduce inputs into models



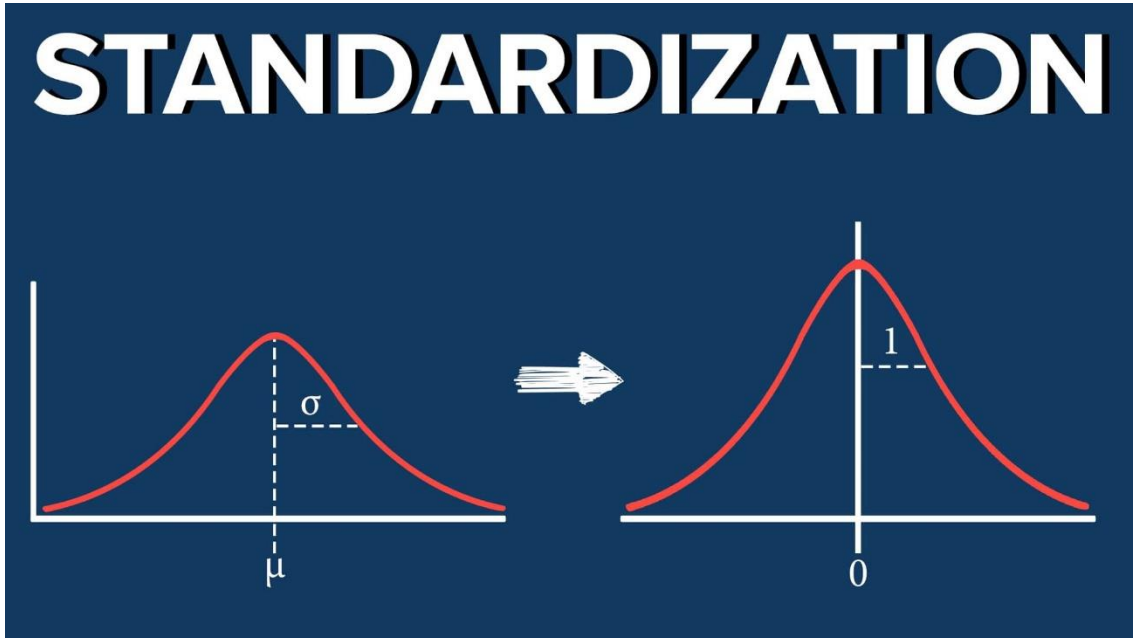
Data Driven Models



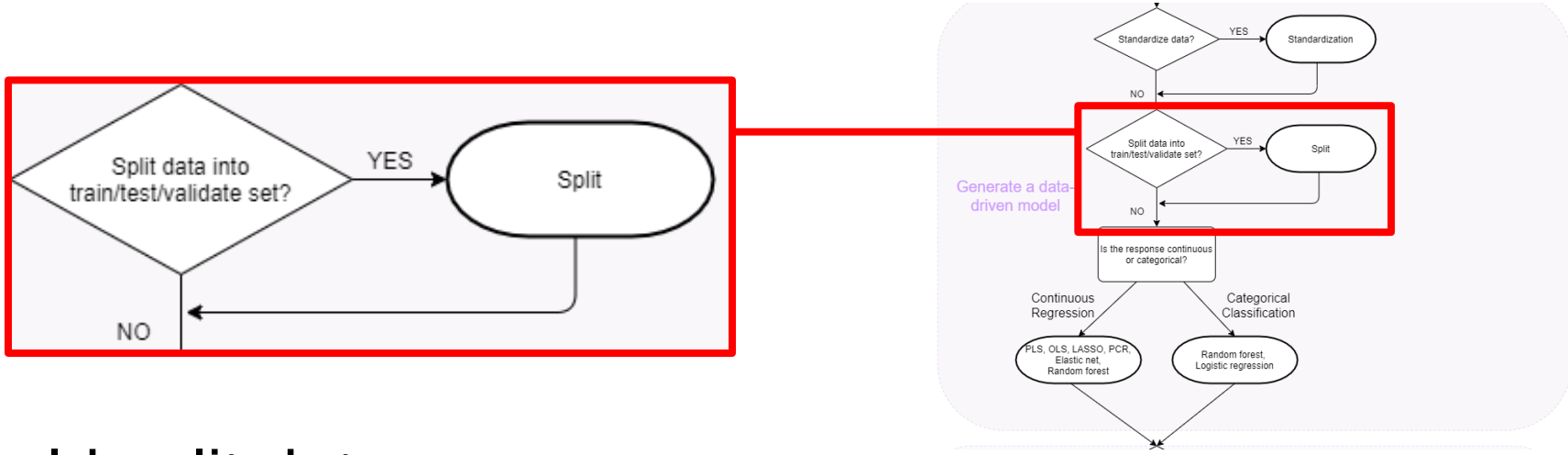
Data Driven Models – Standardization



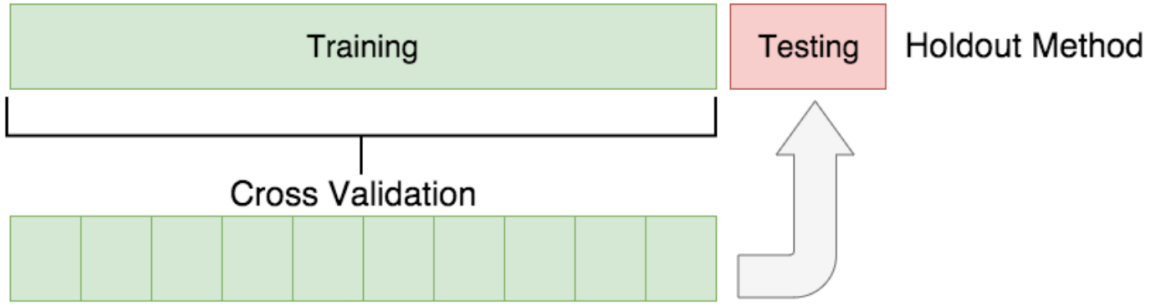
- Should standardize or normalize before building a model
- Think of it as “pre-processing” data
  - Scales data into comparable range



Data Driven Models – Splitting Data



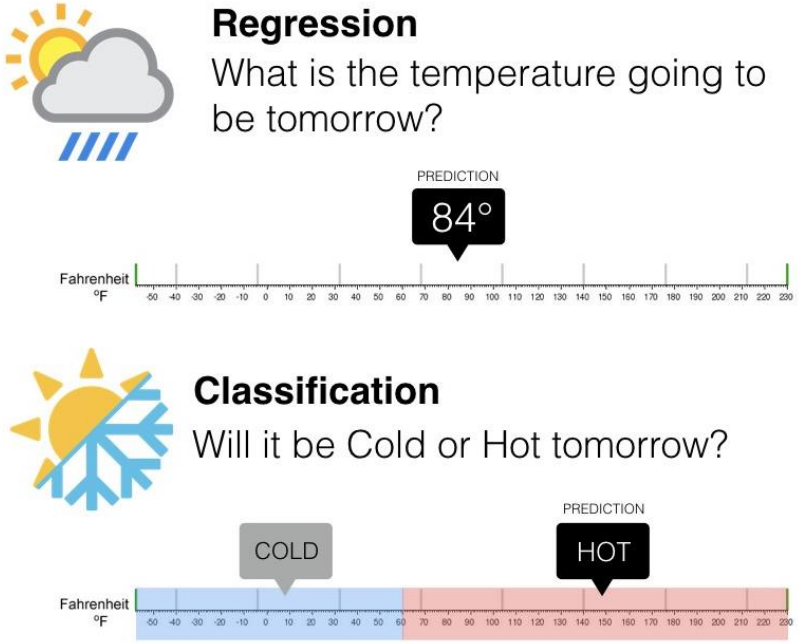
- Should split data to ensure model is appropriate trained and validated
  - Guards against overfitting
  - Gives confidence in model accuracy



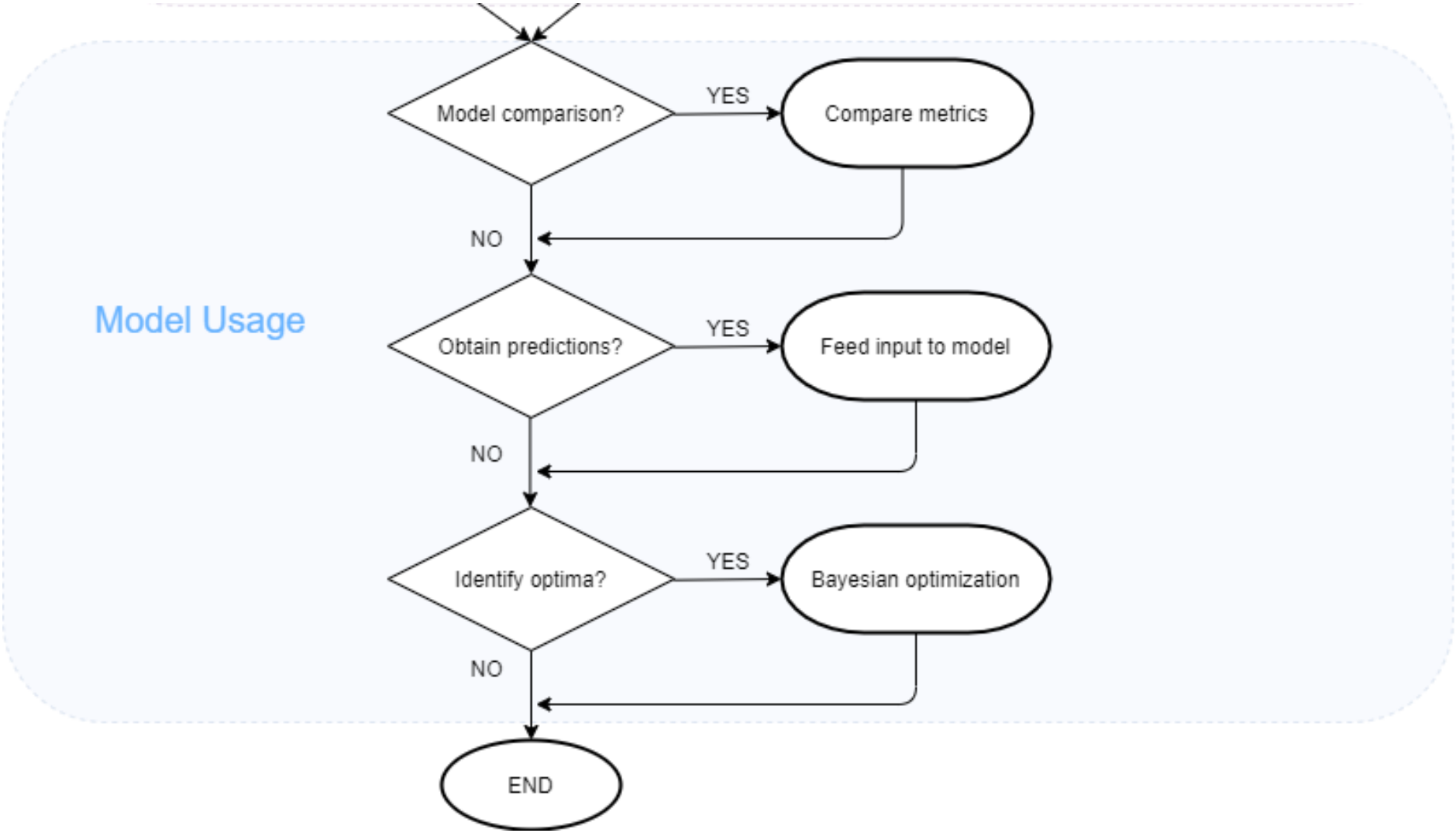
Data Permitting:

Data Driven Models – Regression/Classification

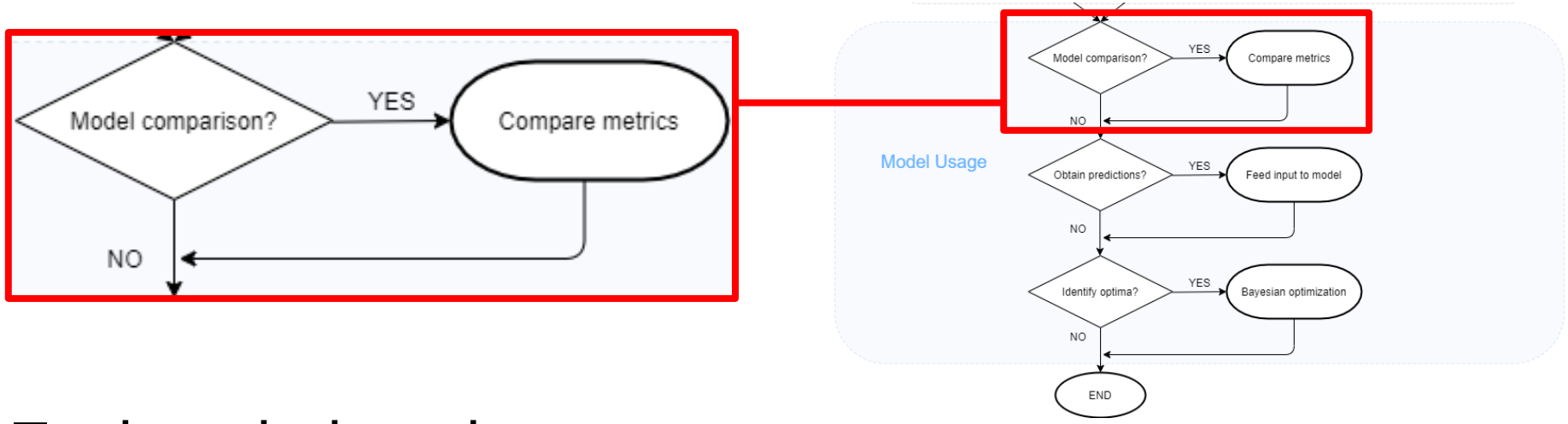
- Constructing model to accurately predict
  - Prediction can be a class or a continuous variable



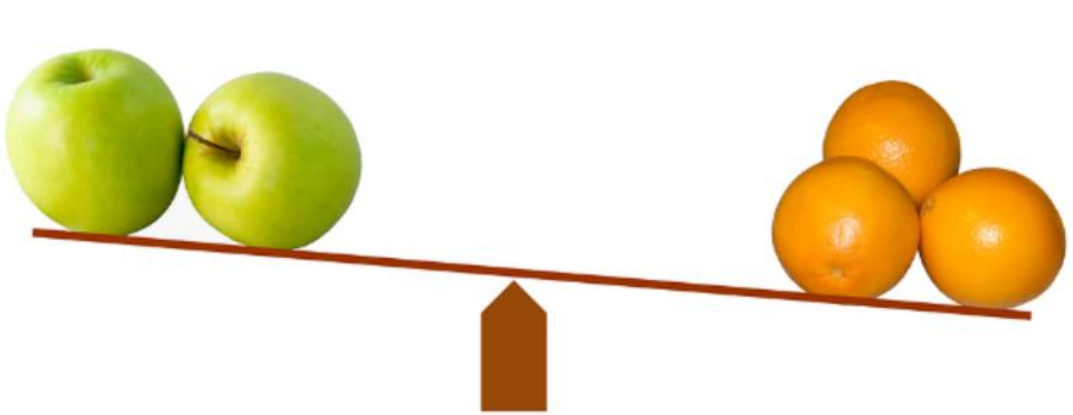
Model Usage


Model Usage
Model comparison?
YES
Compare metrics
NO
Obtain predictions?
YES
Feed input to model
NO
Identify optima?
YES
Bayesian optimization
NO
END

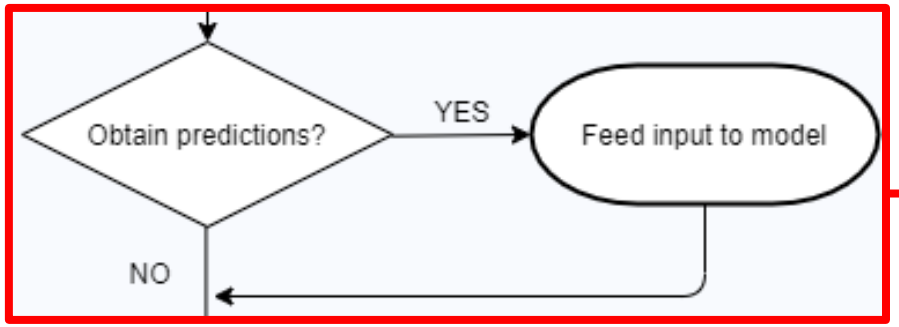
Model Usage – Comparison

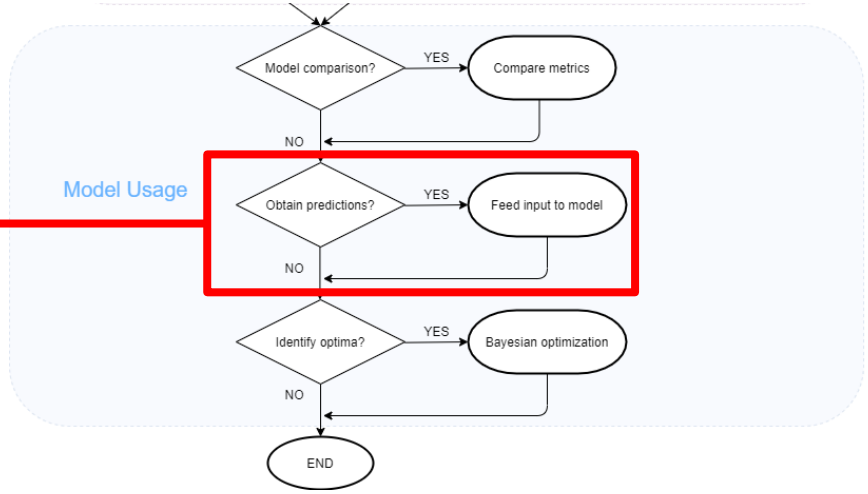


- Each technique has the appropriate metric to be used for evaluation
  - Some can be compared against each other

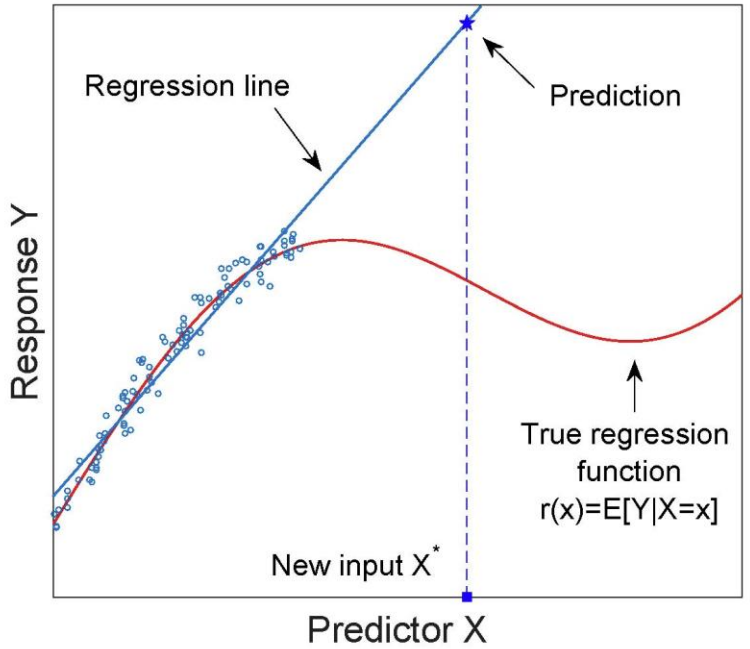


Model Usage – Predictions

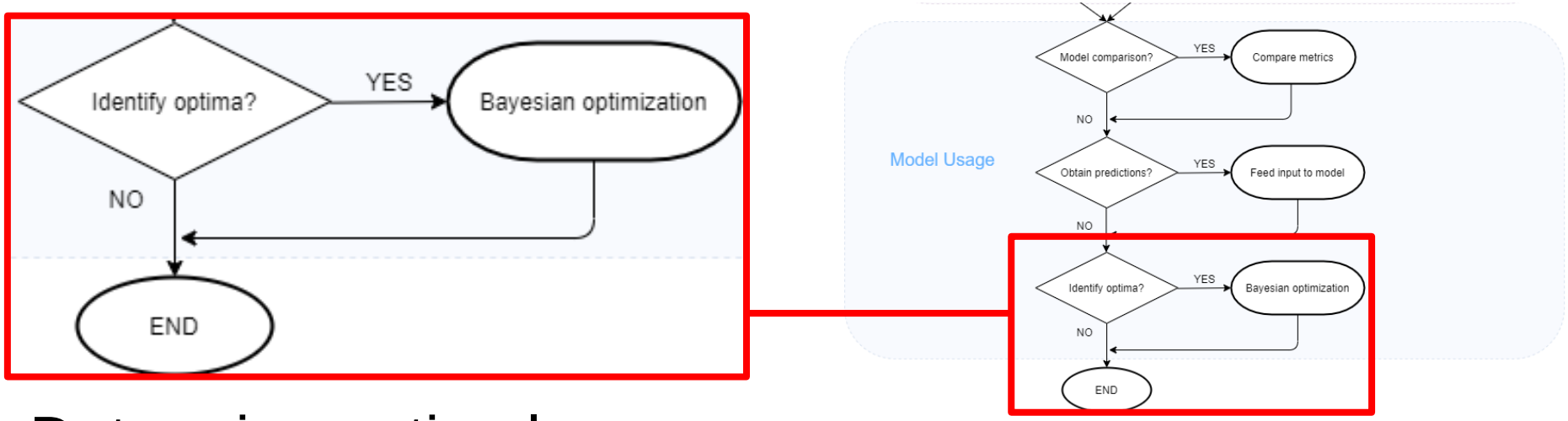




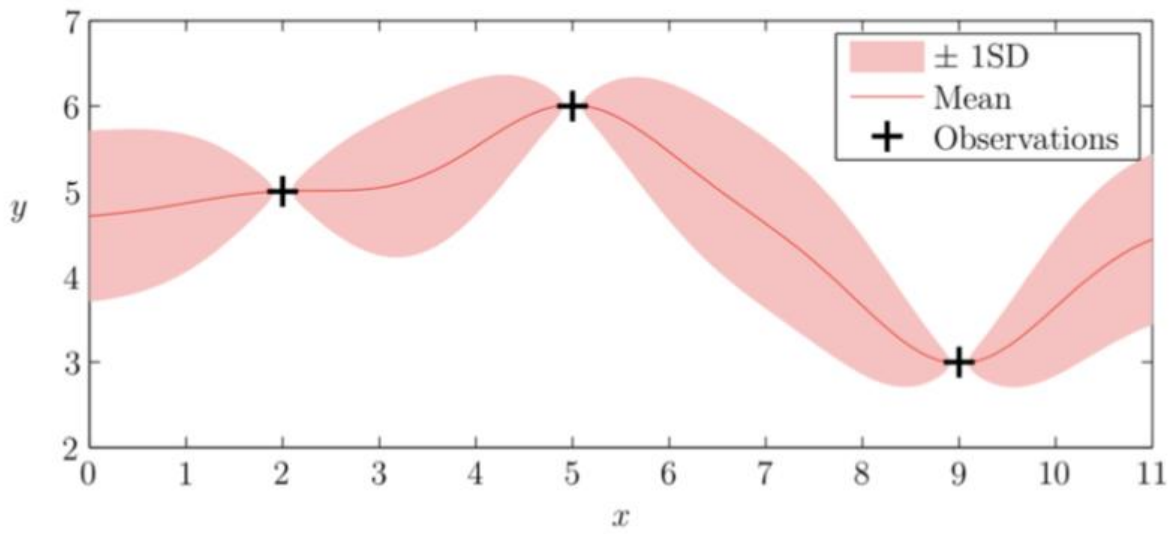
- Process input data to obtain predictions
- Interpolate, don't extrapolate



Model Usage – Optimization



- Determine optimal values from model
- Apply Bayesian optimization to identify optima



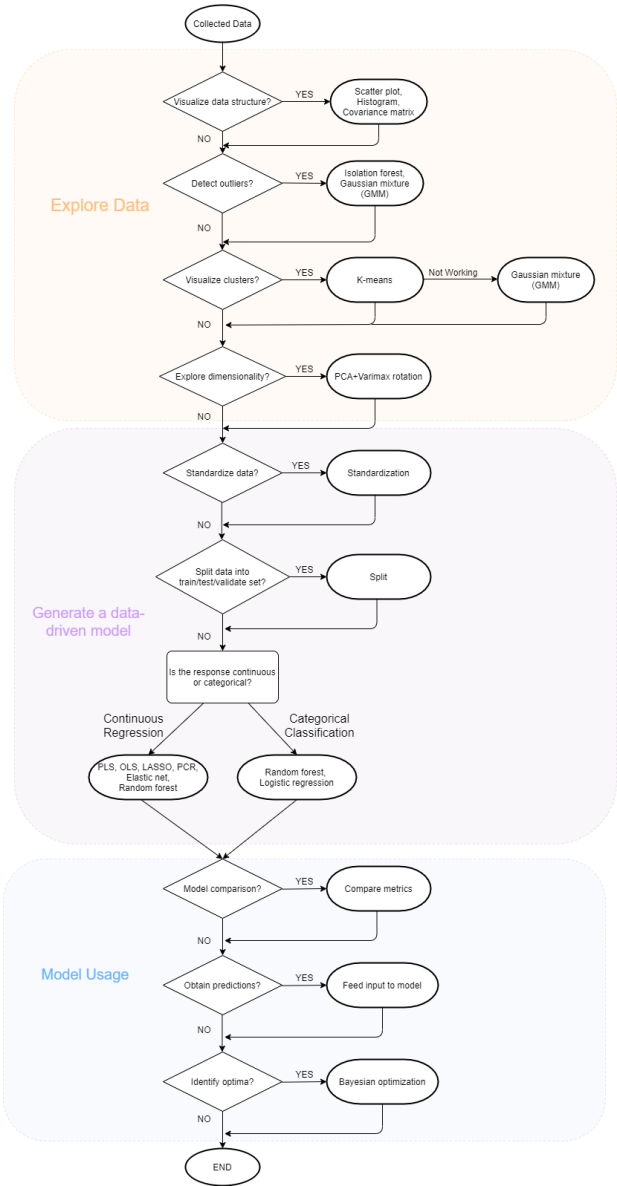
How to Apply Data Science?

- All tools can be misused
- Created a flowchart to provide structure for applying techniques
- **Goal of these training sessions: teach you how to use the flowchart and listed techniques**
  - Provide foundation

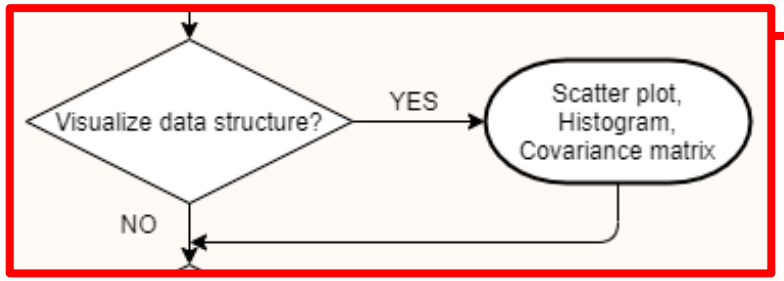
Explore Data

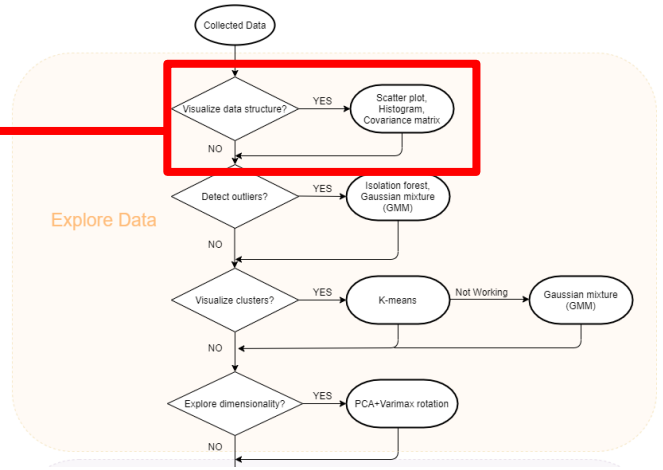
Generate a data-driven model

Model Usage



Explore Data – Visualize

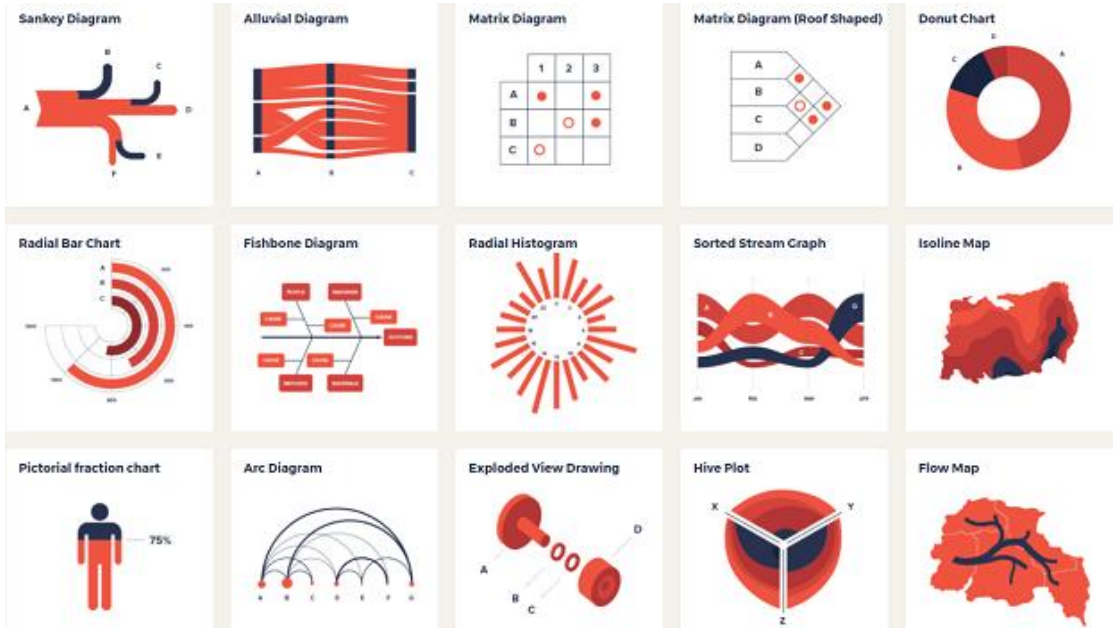




- Why visualize your data?
- Cover fundamentals of:
  - Histograms
  - Scatter plots
  - Correlation matrices
- Workshop examples of each in Python

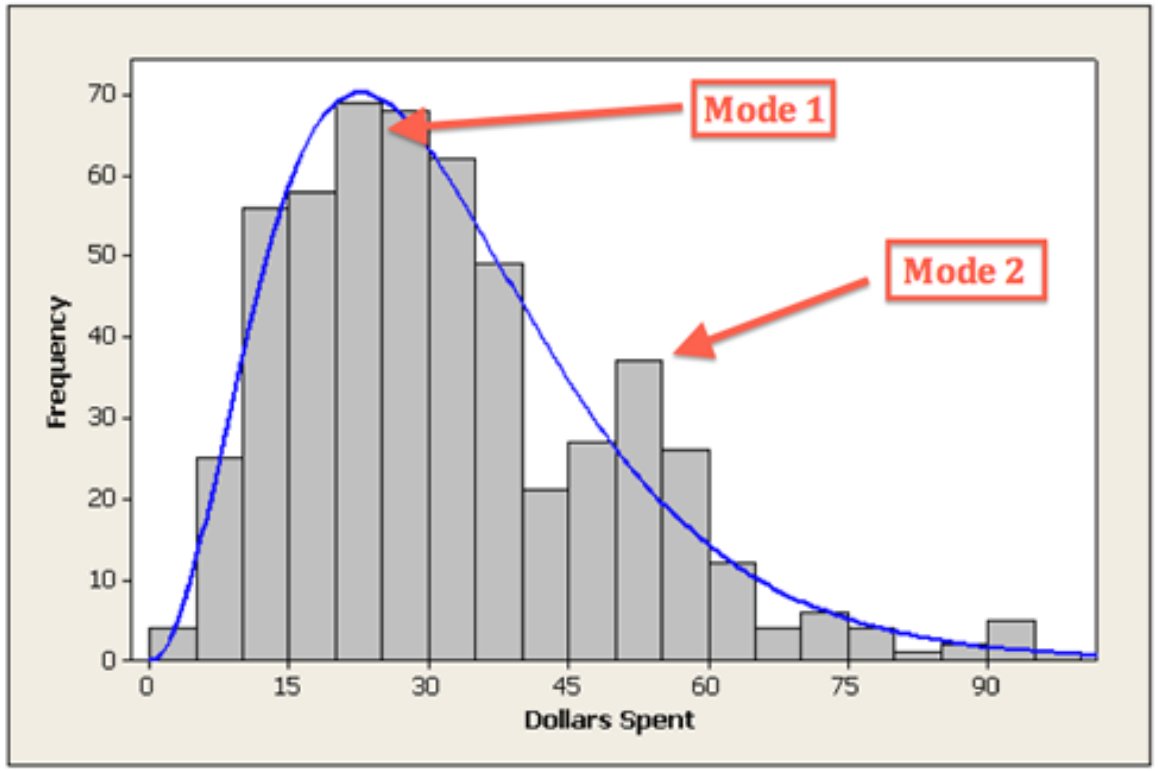
Why Visualize Data?

- Identify important patterns and trends
  - Helps target your model construction
- Identify outliers or errors in data entry
  - Help eliminate troublesome datapoints
- Gives a familiarity with the data structure



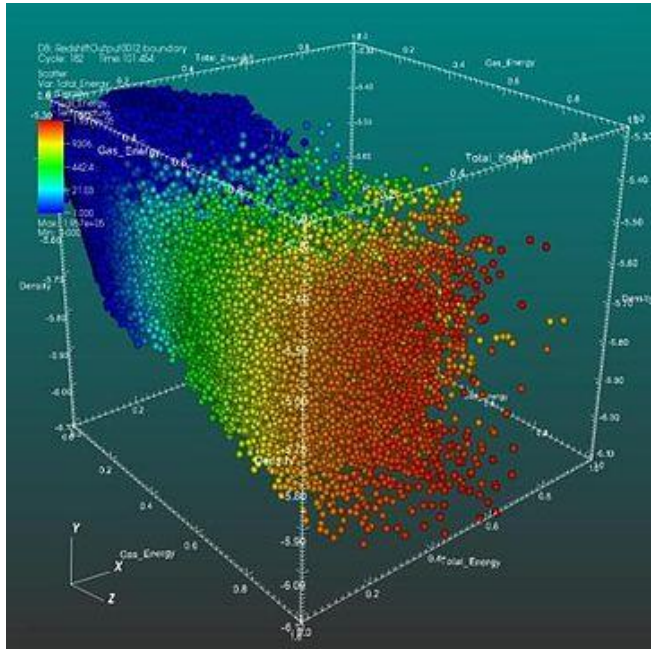
Fundamentals – Histogram

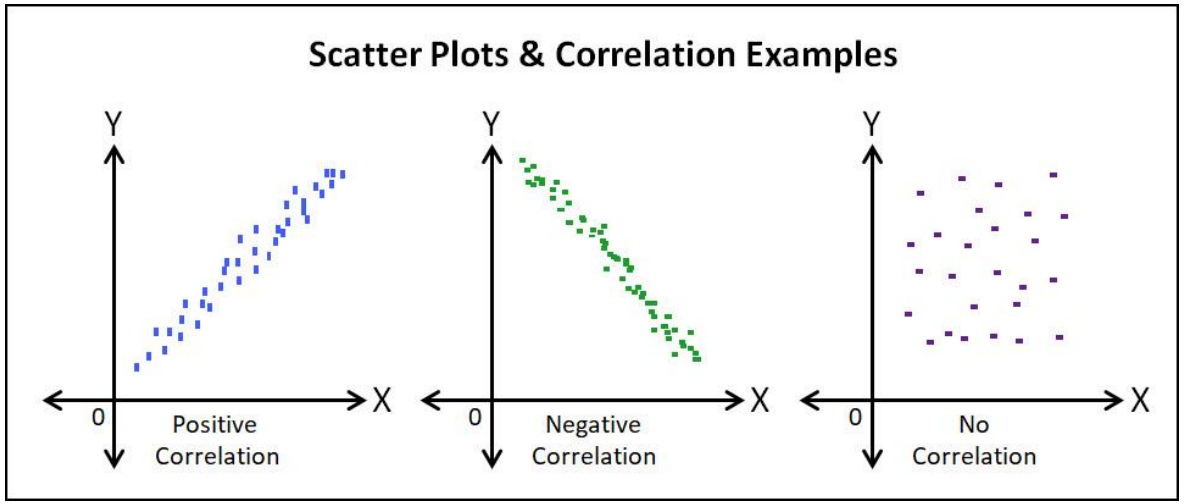
- Approximate representation of the distribution of data
  - Visualizes a single variable at a time
  - Visualization changes with different bin sizes
  - Normalized histograms = probability distributions



Fundamentals – Scatter Plot

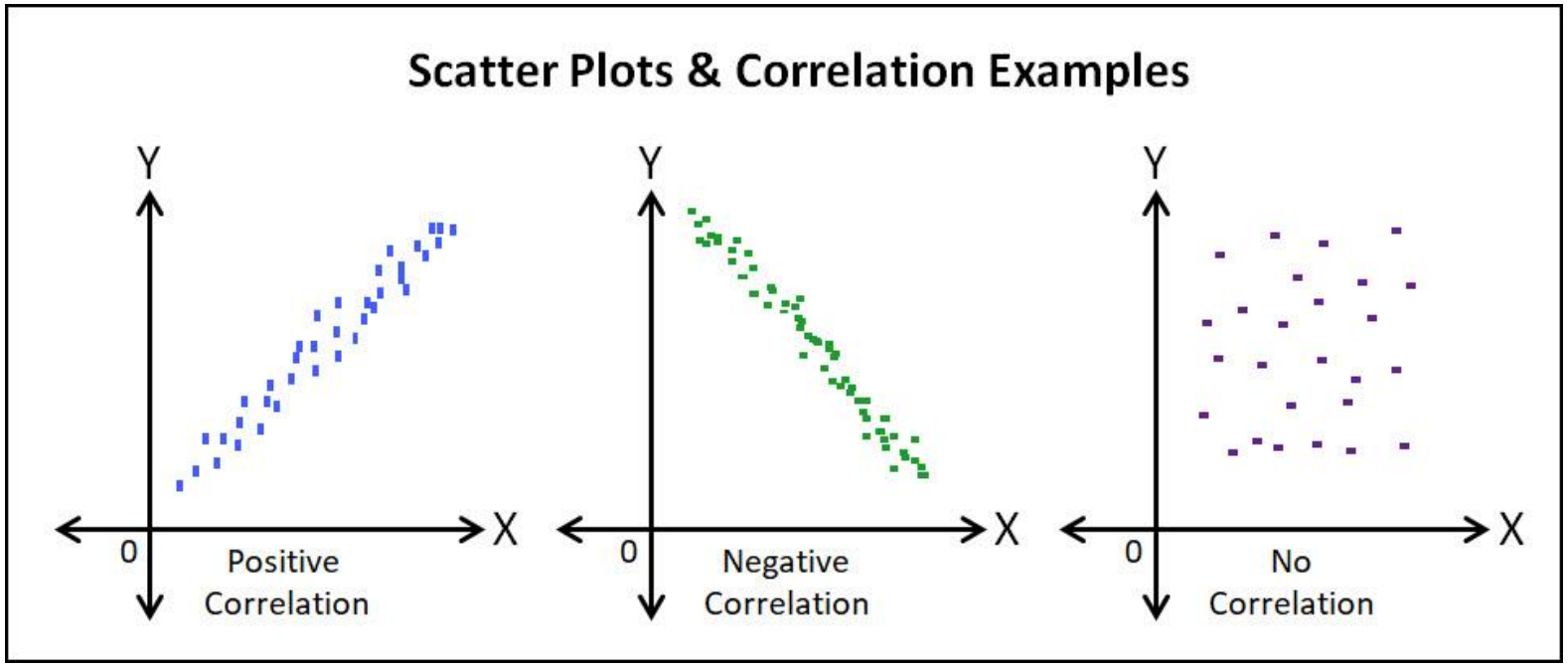
- Mathematical diagram representing multivariate data as discrete points
  - Can be various dimensions
  - We will stick to two dimensions
- Visualize relationships between variables





Fundamentals – Correlation Matrix

- Measurement of how when one variable changes another changes along with it
  - Miles run vs. calories burned will have a positive correlation
  - Miles run vs. shows watched will have a negative correlation
- Ranges from -1 to 1 for ease of comparison



To Jupyter!