EXPLORATORY DATA ANALYSIS

A BRIEF PRIMER

Curated by

HARPREET SAHOTA

THE ARTISTS OF DATA SCIENCE

The Artists of Data Science
with Harpreet Sahota

HOW CAN WE GENERATE MEANINGFUL AND USEFUL INFORMATION FROM DATA?

The answer is **Exploratory Data Analysis (EDA)**. EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures.

THE SIGNIFICANCE OF EDA

EDA actually reveals ground truth about the content without making any underlying assumptions. This is the fact that data scientists use this process to actually understand what type of modeling and hypotheses can be created. Key components of exploratory data analysis include summarizing data, statistical analysis, and visualization of data.

STEPS IN EDA

PROBLEM DEFINITION

The main tasks involved in problem definition are **defining the main objective of the analysis**, defining the main deliverables, outlining the main roles and responsibilities, obtaining the current status of the data, defining the timetable, and performing cost/benefit analysis

DATA ANALYSIS

The main tasks involve summarizing the data, **finding the hidden correlation and relationships among the data**, developing predictive models, evaluating the models, and calculating the accuracies. Some of the **techniques used** for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

COMMUNICATION OF RESULTS

This step involves **presenting the dataset to the target audience** in the form of graphs, summary tables, maps, and diagrams. The result analyzed from the dataset should be i**nterpretable by the business stakeholders**, which is one of the major goals of EDA. Most of the graphical analysis techniques include scattering plots, character plots, histograms, box plots, residual plots, mean plots, and others.

The Artists of Data Science
with Harpreet Sahota

TYPES OF DATA

It is crucial to identify the type of data under analysis. Here we're going to examine numerical data, categorical data, and their subtypes

NUMERICAL DATA

This data has a **sense of measurement involved in it**; for example, a person's age, height, weight, blood pressure, heart rate, temperature, number of teeth, number of bones, and the number of family members. The numerical dataset can be either *discrete* or *continuous* types.

DISCRETE DATA

This is data that is **countable and its values can be listed out**. The discrete variable takes a *fixed number of distinct values*. For example, the Country variable can have values such as Canada, England, Malta, and Croatia. It is fixed. The Rank variable of a student in a classroom can take values from 1, 2, 3, 4, 5, and so on.

CONTINUOUS DATA

A variable that can have an **infinite number of numerical values within a specific range** is classified as continuous variable. For example, what is the temperature of your city today? *Can we be finite?*

CATEGORICAL DATA

This type of data represents the **characteristics of an object**; for example, gender, marital status, type of address, or categories of the movies. Here are some types of categorical data you might find in data:

- Gender (Male, Female, Other, or Unknown)
- Marital Status (Annulled, Divorced, Interlocutory, Legally Separated, Married,
- Polygamous, Never Married, Domestic Partner, Unmarried, Widowed, or
- Unknown)
- Blood type (A, B, AB, or O)

These types of variables can have one of a limited number of values. A **binary categorical variable** can take exactly two values and is also referred to as a *dichotomous variable*. **Polytomous variables** are categorical variables that can take more than two possible values

MEASUREMENT SCALES

You might be thinking why should you care about the scale of your data? Can't I load the data and just get on with my analyzing ? Well, you *could*. But think about this: you have a dataset, and you want to analyze it. **How will you decide whether you can make a pie chart, bar chart, or histogram?**

Understanding the type of data is relevant in understanding what type of computation you can perform, what type of model you should fit on the dataset, and what type of visualization you can generate. There are four measurement scales that data can take on: nominal, ordinal, interval, and ratio

NOMINAL DATA

This data has **no natural ordering** among the categories such as gender, eye color, political party, or ethnicity.

- **Frequency** is the rate at which a label occurs over a period of time within the dataset.
- **Proportion** can be calculated by dividing the frequency by the total number of events.
- Then, you could compute the **percentage** of each proportion.
- And to **visualize** the nominal dataset, you can use either a pie chart or a bar chart.

ORDINAL DATA

Ordinal data has a natural order among the categories but the distance between values is undeterminable. Examples might include:

- Likert scales (strongly agree, neutral, strongly disagree)
- Preferences ranking (low, medium, high)
- Placement in a contest (first, second, third)

The main difference in the ordinal and nominal scale is the order. In ordinal scales, the order of the values is a significant factor. The *mode* or the *median* is acceptable as the measure of central tendency; but the average wouldn't make much sense.

INTERVAL DATA

For interval data the **values have order and the distance between units is the same**, but there is not a zero value. Examples include SAT scores or miles per hour (MPH)—the difference between 100 MPH and 90 MPH is the same difference as between 70 MPH and 60 MPH.

Measure of central tendencies that are acceptable are the classical summary statistics: Mean, median, mode, and standard deviations.

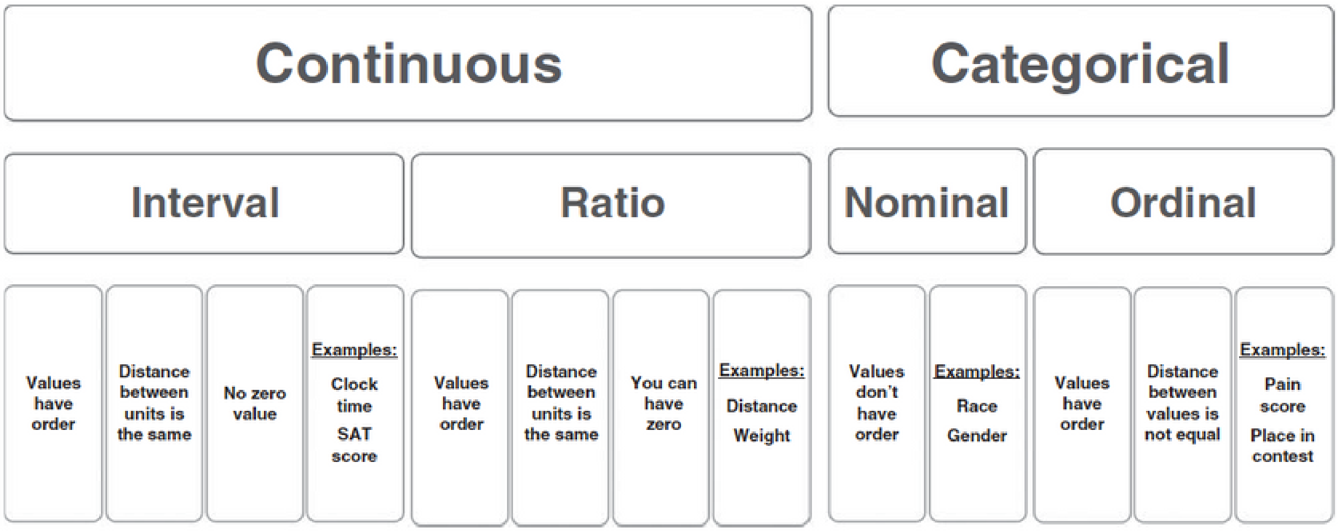
RATIO DATA

Ratio scales contain **order, exact values, and absolute zero**, which makes it possible to be used in descriptive and inferential statistics.

These scales provide numerous possibilities for statistical analysis. Mathematical operations, the measure of central tendencies, and the measure of dispersion and coefficient of variation can also be computed from such scales.

SUMMARY OF DATA TYPES

Provides:	Nominal	Ordinal	Interval	Ratio
The “order”of values is known		✓	✓	✓
“Counts,” aka “Frequency of Distribution”	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has “true zero”				✓



Continuous

Interval: Values have order; Distance between units is the same; No zero value; Examples: Clock time, SAT score

Ratio: Values have order; Distance between units is the same; You can have zero; Examples: Distance, Weight

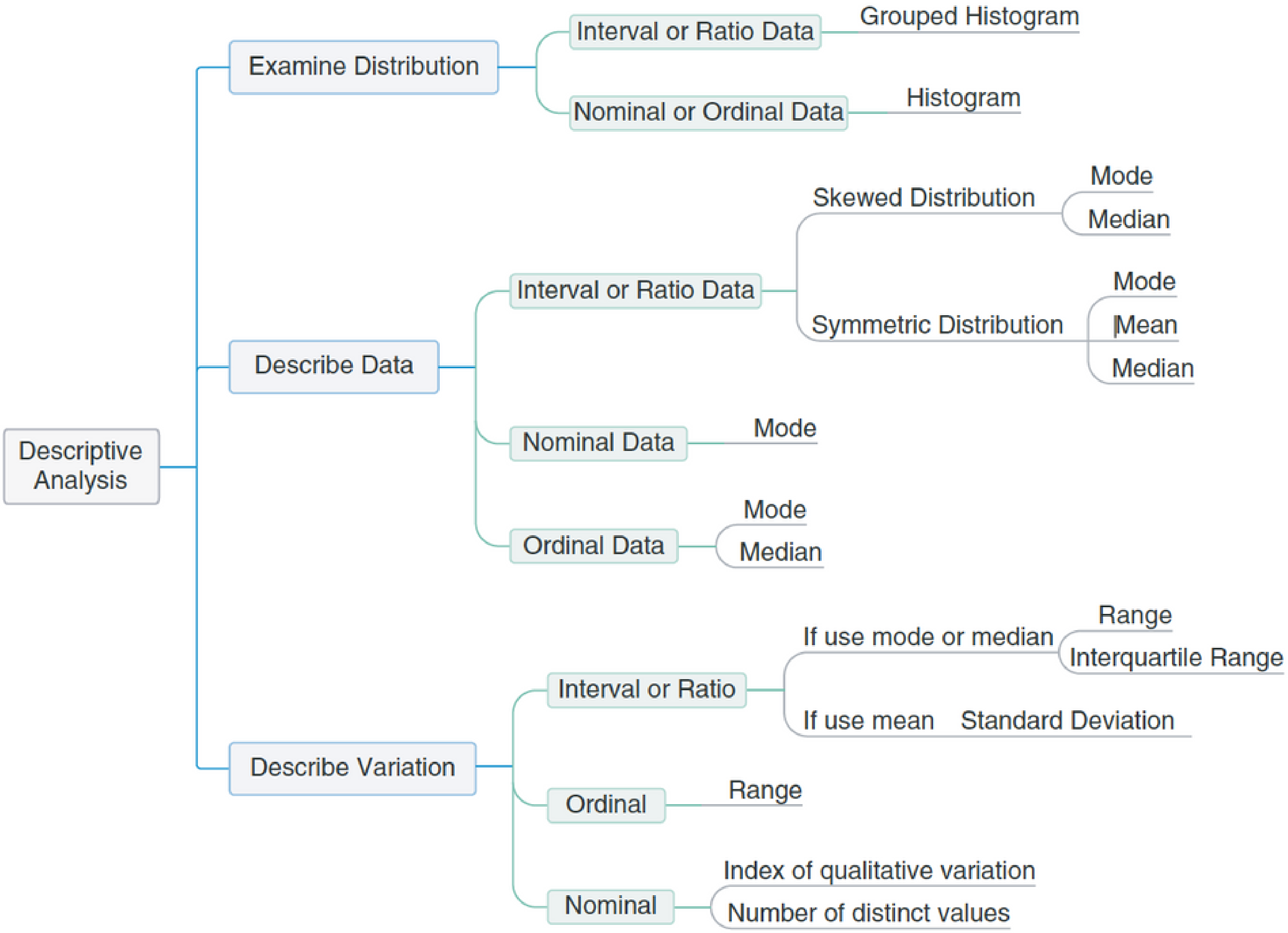
Categorical

Nominal: Values don’t have order; Examples: Race, Gender

Ordinal: Values have order; Distance between values is not equal; Examples: Pain score, Place in contest

The Artists of Data Science
with Harpreet Sahota

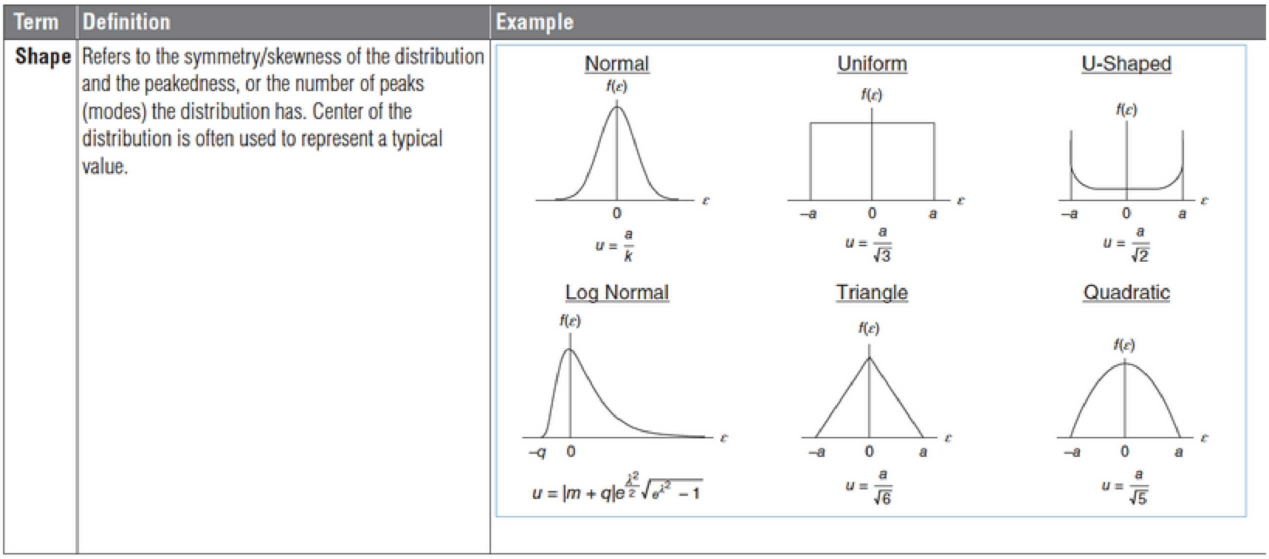
FLOW CHART OF ANALYSIS OPTIONS



- Descriptive Analysis
  - Examine Distribution
    - Interval or Ratio Data
      - Grouped Histogram
    - Nominal or Ordinal Data
      - Histogram
  - Describe Data
    - Interval or Ratio Data
      - Skewed Distribution
        - Mode
        - Median
      - Symmetric Distribution
        - Mode
        - |Mean
        - Median
    - Nominal Data
      - Mode
    - Ordinal Data
      - Mode
      - Median
  - Describe Variation
    - Interval or Ratio
      - If use mode or median
        - Range
        - Interquartile Range
      - If use mean
        - Standard Deviation
    - Ordinal
      - Range
    - Nominal
      - Index of qualitative variation
      - Number of distinct values

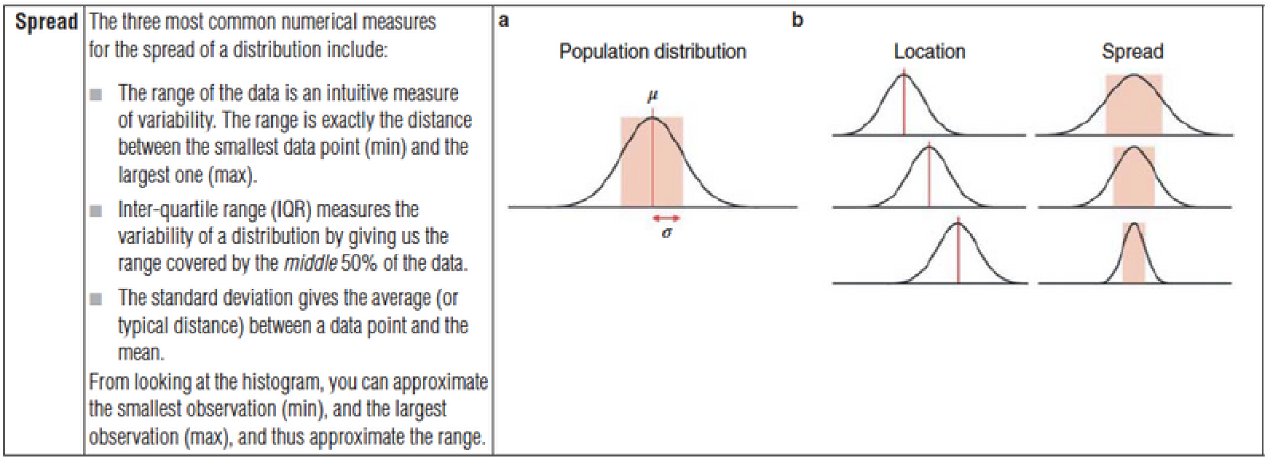
The Artists of Data Science
with Harpreet Sahota

HOW TO EXAMINE VARIABLES



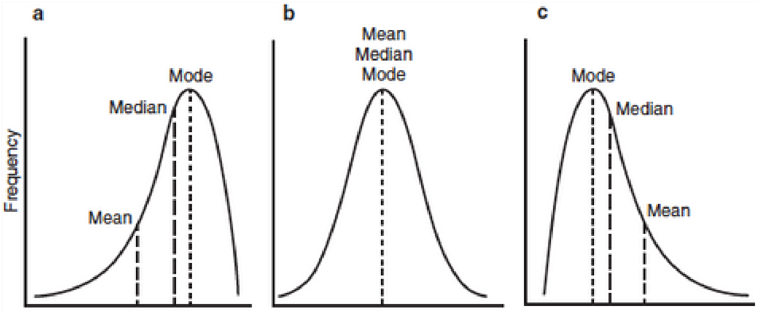
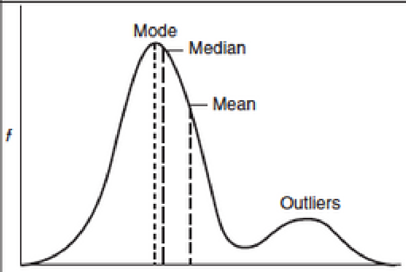
Term	Definition	Example
Shape	Refers to the symmetry/skewness of the distribution and the peakedness, or the number of peaks (modes) the distribution has. Center of the distribution is often used to represent a typical value.	Normal: $u = \frac{a}{k}$ Uniform: $u = \frac{a}{\sqrt{3}}$ U-Shaped: $u = \frac{a}{\sqrt{2}}$ Log Normal: $u = \lvert m + q \rvert e^{\frac{\lambda^2}{2}} \sqrt{e^{\lambda^2} - 1}$ Triangle: $u = \frac{a}{\sqrt{6}}$ Quadratic: $u = \frac{a}{\sqrt{5}}$

Normal
f(ε)
0
ε
Uniform
f(ε)
–a
0
a
ε
U-Shaped
f(ε)
–a
0
a
ε
Log Normal
f(ε)
–q
0
ε
Triangle
f(ε)
–a
0
a
ε
Quadratic
f(ε)
–a
0
a
ε



Spread	The three most common numerical measures for the spread of a distribution include: ■ The range of the data is an intuitive measure of variability. The range is exactly the distance between the smallest data point (min) and the largest one (max). ■ Inter-quartile range (IQR) measures the variability of a distribution by giving us the range covered by the *middle* 50% of the data. ■ The standard deviation gives the average (or typical distance) between a data point and the mean. From looking at the histogram, you can approximate the smallest observation (min), and the largest observation (max), and thus approximate the range.	a Population distribution b Location, Spread

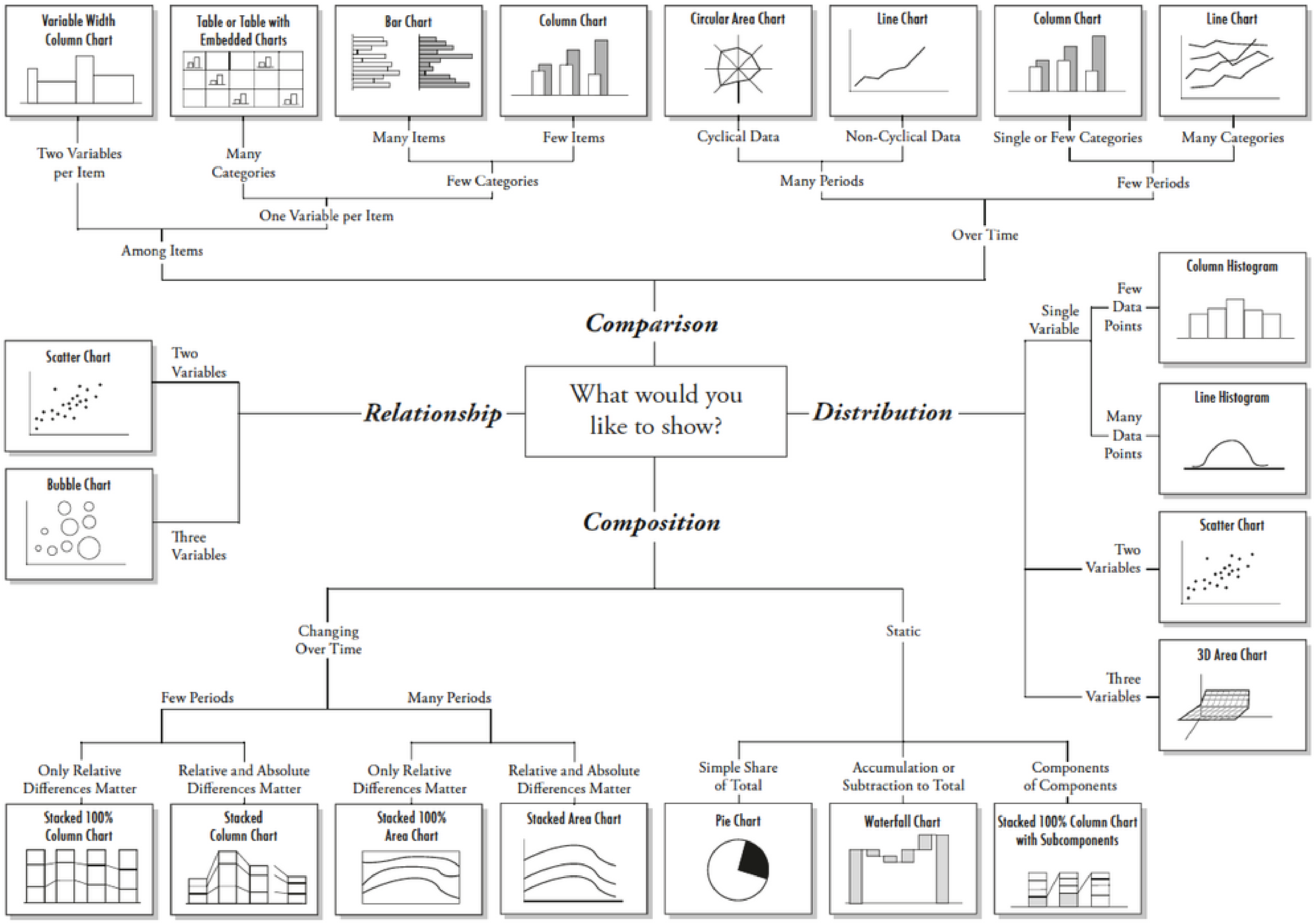
HOW TO EXAMINE VARIABLES

Term	Definition	Example
Position	Measures of position also allow us to compare values from different distributions. - *Percentiles*—The *P*-th percentile can be interpreted as a location in the data for which approximately *P*% of the other values in the distribution fall below the *P*-th percentile and (100 – *P*)% fall above the *P*-th percentile. - *Five-number summary*—The combination of the five numbers (min, Q1, Median, Q3, Max) provides a quick numerical description of both the center and spread of a distribution. - *Standardized scores (Z-scores)*—The standardized scores tell us how many standard deviations the raw score for that individual deviates from the mean and in what direction.	
Outliers	Outliers are observations that fall outside the overall pattern.	

The Artists of Data Science
with Harpreet Sahota

HOW TO VISUALIZE DATA

Chart Suggestions—A Thought-Starter



What would you like to show?

Comparison

Among Items

Two Variables per Item — Variable Width Column Chart

One Variable per Item

Many Categories — Table or Table with Embedded Charts

Few Categories

Many Items — Bar Chart

Few Items — Column Chart

Over Time

Many Periods

Cyclical Data — Circular Area Chart

Non-Cyclical Data — Line Chart

Few Periods

Single or Few Categories — Column Chart

Many Categories — Line Chart

Relationship

Two Variables — Scatter Chart

Three Variables — Bubble Chart

Distribution

Single Variable

Few Data Points — Column Histogram

Many Data Points — Line Histogram

Two Variables — Scatter Chart

Three Variables — 3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter — Stacked 100% Column Chart

Relative and Absolute Differences Matter — Stacked Column Chart

Many Periods

Only Relative Differences Matter — Stacked 100% Area Chart

Relative and Absolute Differences Matter — Stacked Area Chart

Static

Simple Share of Total — Pie Chart

Accumulation or Subtraction to Total — Waterfall Chart

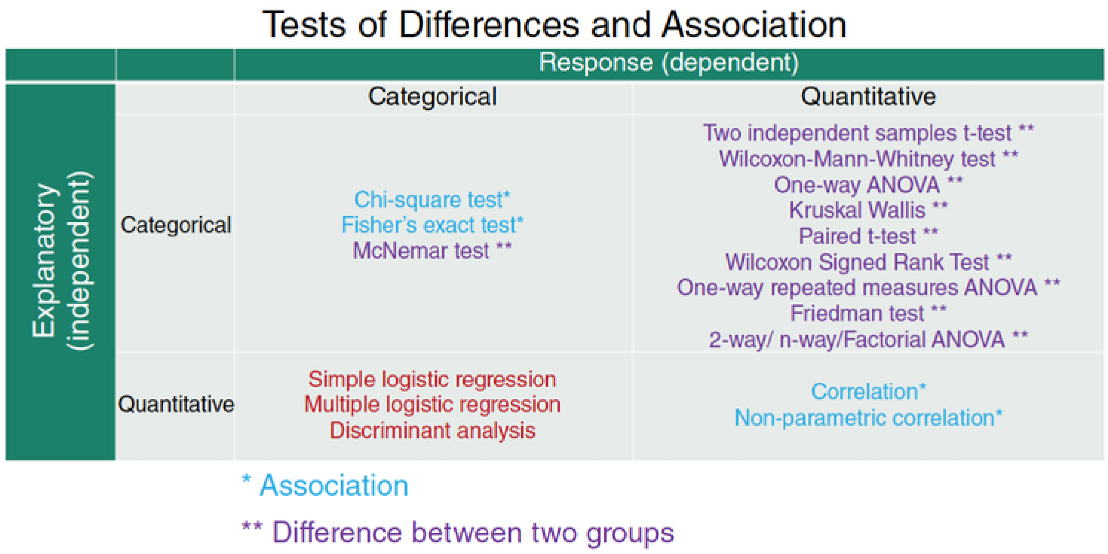
Components of Components — Stacked 100% Column Chart with Subcomponents

The Artists of Data Science
with Harpreet Sahota

FOUR SCENARIOS

The type of variable matters in the methods you can use to explore the relationship. There are four basic types of comparisons that can be made, depending on whether the response variable is categorical or continuous (aka quantitative), and whether the explanatory variable is categorical or continuous. Note: **Response variable**—the outcome or dependent variable. **Explanatory variable**—the variable that claims to explain, predict, or affect the response (independent variable, predictor variable, or covariate). When confronted with a research question that involves exploring the relationship between two variables, the first and most crucial step is to determine which of the four cases represents the data structure of the problem. Each of the four scenarios will determine what methods can be used to understand the relationship.

COMMON STATISTICAL TESTS FOR DIFFERENCES AND ASSOCIATIONS



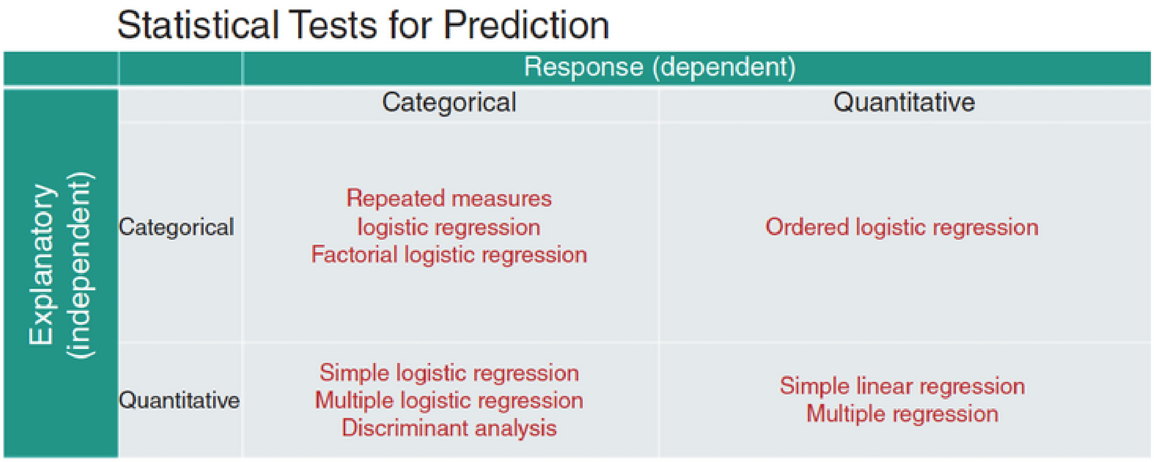
Tests of Differences and Association

		Response (dependent)	
		Categorical	Quantitative
Explanatory (independent)	Categorical	Chi-square test* Fisher's exact test* McNemar test **	Two independent samples t-test ** Wilcoxon-Mann-Whitney test ** One-way ANOVA ** Kruskal Wallis ** Paired t-test ** Wilcoxon Signed Rank Test ** One-way repeated measures ANOVA ** Friedman test ** 2-way/ n-way/Factorial ANOVA **
	Quantitative	Simple logistic regression Multiple logistic regression Discriminant analysis	Correlation* Non-parametric correlation*

* Association

** Difference between two groups

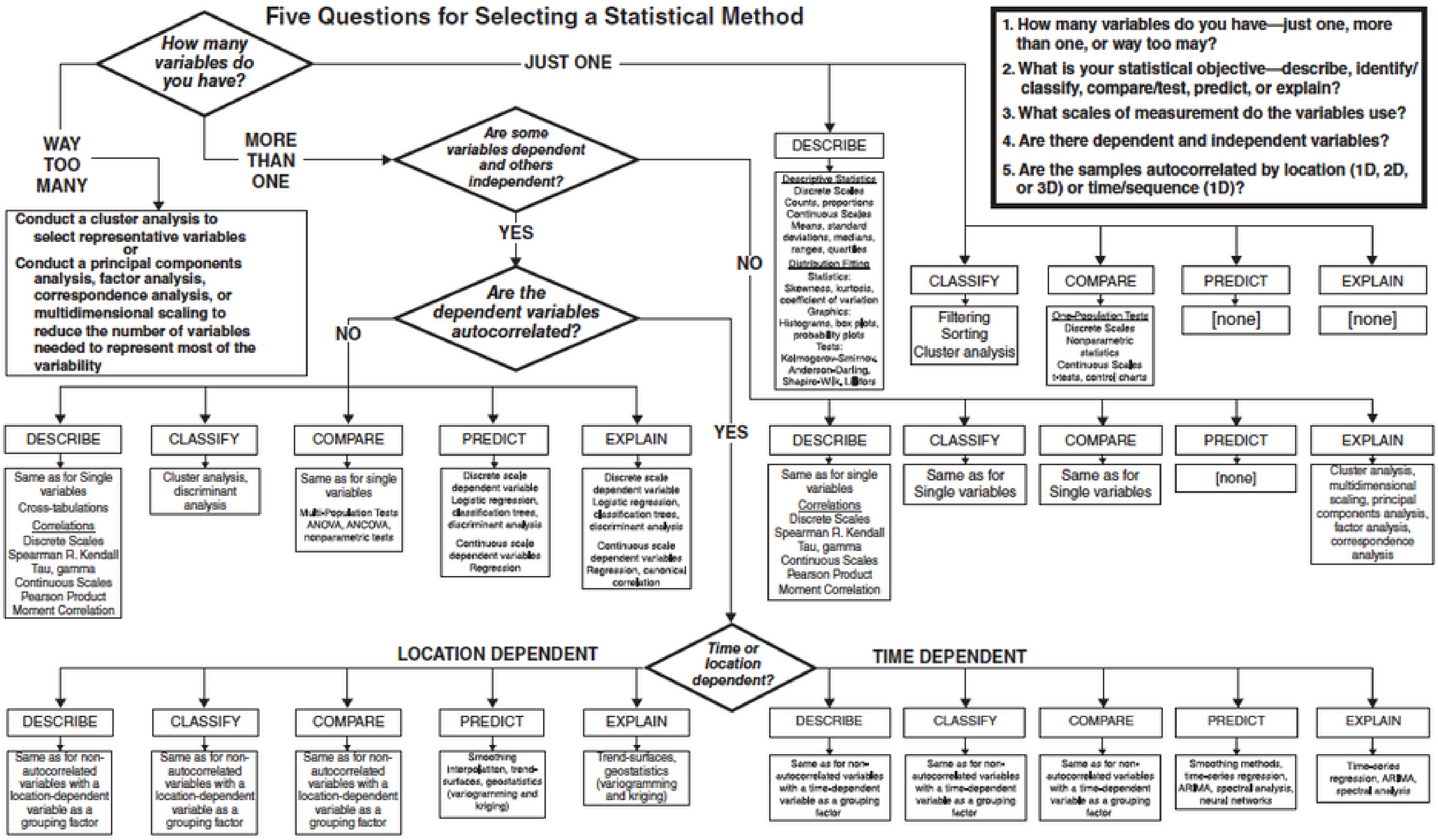
COMMON STATISTICAL TESTS FOR PREDICTION



Statistical Tests for Prediction

		Response (dependent)	
		Categorical	Quantitative
Explanatory (independent)	Categorical	Repeated measures logistic regression Factorial logistic regression	Ordered logistic regression
	Quantitative	Simple logistic regression Multiple logistic regression Discriminant analysis	Simple linear regression Multiple regression

SELECTING A STATISTICAL METHOD



Five Questions for Selecting a Statistical Method

1. How many variables do you have—just one, more than one, or way too may?
2. What is your statistical objective—describe, identify/classify, compare/test, predict, or explain?
3. What scales of measurement do the variables use?
4. Are there dependent and independent variables?
5. Are the samples autocorrelated by location (1D, 2D, or 3D) or time/sequence (1D)?

How many variables do you have?

JUST ONE

DESCRIBE

Descriptive Statistics
Discrete Scales
Counts, proportions
Continuous Scales
Means, standard deviations, medians, ranges, quartiles
Distribution Fitting
Statistics:
Skewness, kurtosis, coefficient of variation
Graphics:
Histograms, box plots, probability plots
Tests:
Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, Lilliefors

CLASSIFY

Filtering
Sorting
Cluster analysis

COMPARE

One-Population Tests
Discrete Scales
Nonparametric statistics
Continuous Scales
t-tests, control charts

PREDICT

[none]

EXPLAIN

[none]

WAY TOO MANY

Conduct a cluster analysis to select representative variables
or
Conduct a principal components analysis, factor analysis, correspondence analysis, or multidimensional scaling to reduce the number of variables needed to represent most of the variability

MORE THAN ONE

Are some variables dependent and others independent?

NO

DESCRIBE

Same as for single variables
Correlations
Discrete Scales
Spearman R. Kendall Tau, gamma
Continuous Scales
Pearson Product Moment Correlation

CLASSIFY

Same as for Single variables

COMPARE

Same as for Single variables

PREDICT

[none]

EXPLAIN

Cluster analysis, multidimensional scaling, principal components analysis, factor analysis, correspondence analysis

YES

Are the dependent variables autocorrelated?

NO

DESCRIBE

Same as for Single variables
Cross-tabulations
Correlations
Discrete Scales
Spearman R. Kendall Tau, gamma
Continuous Scales
Pearson Product Moment Correlation

CLASSIFY

Cluster analysis, discriminant analysis

COMPARE

Same as for single variables
Multi-Population Tests
ANOVA, ANCOVA, nonparametric tests

PREDICT

Discrete scale dependent variable
Logistic regression, classification trees, discriminant analysis
Continuous scale dependent variables
Regression

EXPLAIN

Discrete scale dependent variable
Logistic regression, classification trees, discriminant analysis
Continuous scale dependent variables
Regression, canonical correlation

YES

Time or location dependent?

LOCATION DEPENDENT

DESCRIBE

Same as for non-autocorrelated variables with a location-dependent variable as a grouping factor

CLASSIFY

Same as for non-autocorrelated variables with a location-dependent variable as a grouping factor

COMPARE

Same as for non-autocorrelated variables with a location-dependent variable as a grouping factor

PREDICT

Smoothing interpolation, trend-surfaces, geostatistics (variogramming and kriging)

EXPLAIN

Trend-surfaces, geostatistics (variogramming and kriging)

TIME DEPENDENT

DESCRIBE

Same as for non-autocorrelated variables with a time-dependent variable as a grouping factor

CLASSIFY

Same as for non-autocorrelated variables with a time-dependent variable as a grouping factor

COMPARE

Same as for non-autocorrelated variables with a time-dependent variable as a grouping factor

PREDICT

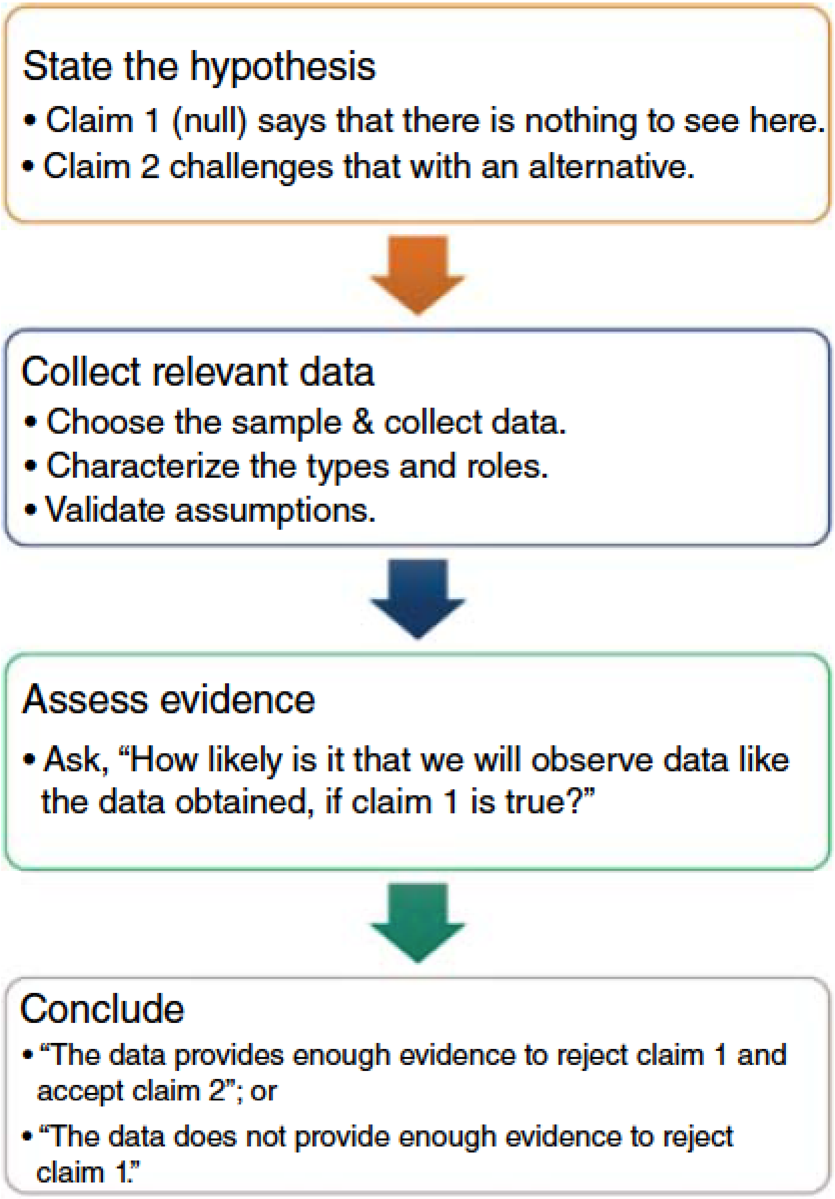
Smoothing methods, time-series regression, ARIMA, spectral analysis, neural networks

EXPLAIN

Time-series regression, ARIMA, spectral analysis

STEPS USED IN THE EVALUATION OF A STATISTICAL HYPOTHESIS

Begin with a claim about the population (the null hypothesis), and check whether the data obtained from the sample provide evidence AGAINST this claim.



State the hypothesis

- Claim 1 (null) says that there is nothing to see here.
- Claim 2 challenges that with an alternative.

Collect relevant data

- Choose the sample & collect data.
- Characterize the types and roles.
- Validate assumptions.

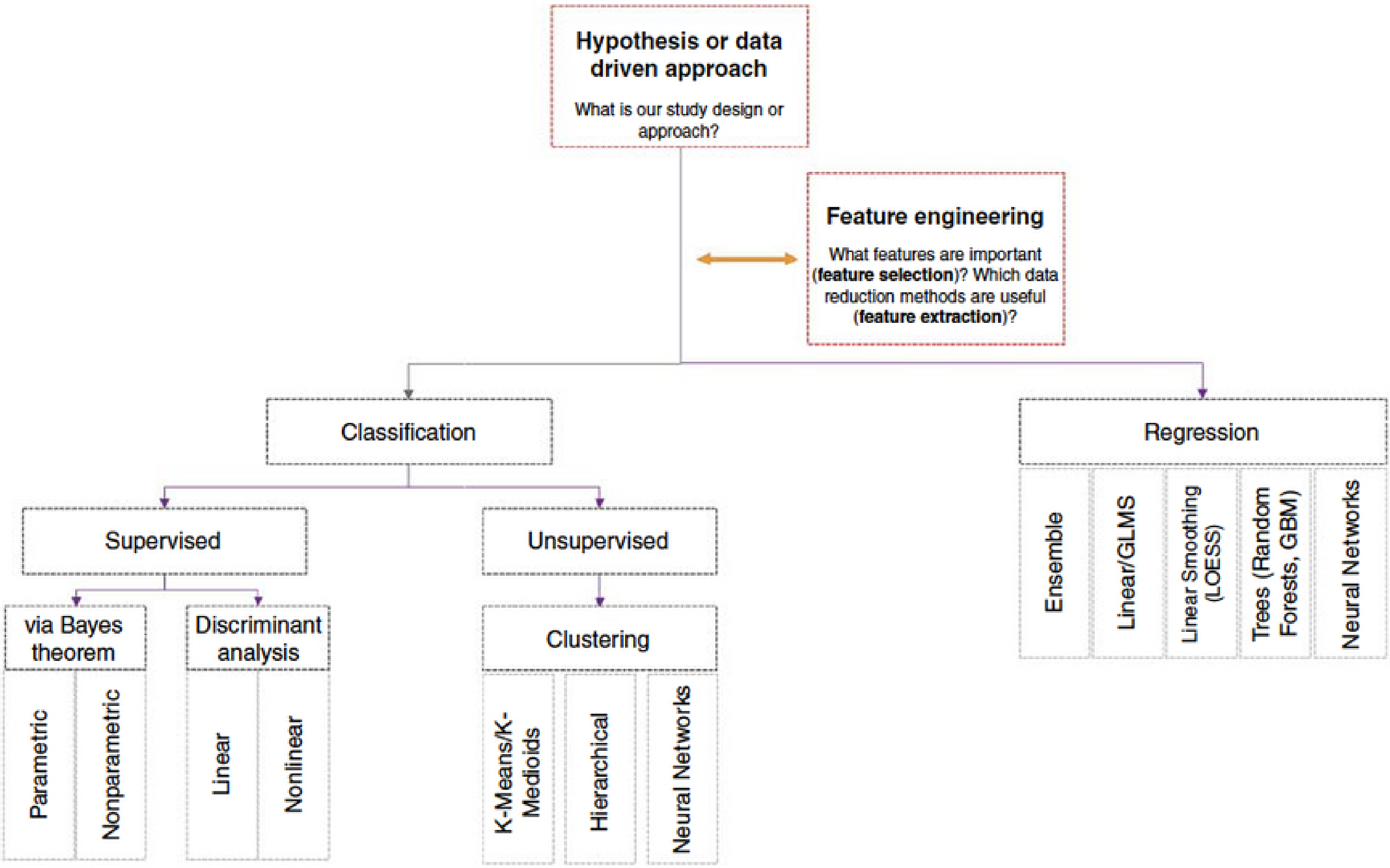
Assess evidence

- Ask, “How likely is it that we will observe data like the data obtained, if claim 1 is true?”

Conclude

- “The data provides enough evidence to reject claim 1 and accept claim 2”; or
- “The data does not provide enough evidence to reject claim 1.”

WHERE DOES FEATURE ENGINEERING FIT IN?



Hypothesis or data driven approach

What is our study design or approach?

Feature engineering

What features are important (**feature selection**)? Which data reduction methods are useful (**feature extraction**)?

Classification

Supervised

via Bayes theorem

Parametric

Nonparametric

Discriminant analysis

Linear

Nonlinear

Unsupervised

Clustering

K-Means/K-Medioids

Hierarchical

Neural Networks

Regression

Ensemble

Linear/GLMS

Linear Smoothing (LOESS)

Trees (Random Forests, GBM)

Neural Networks