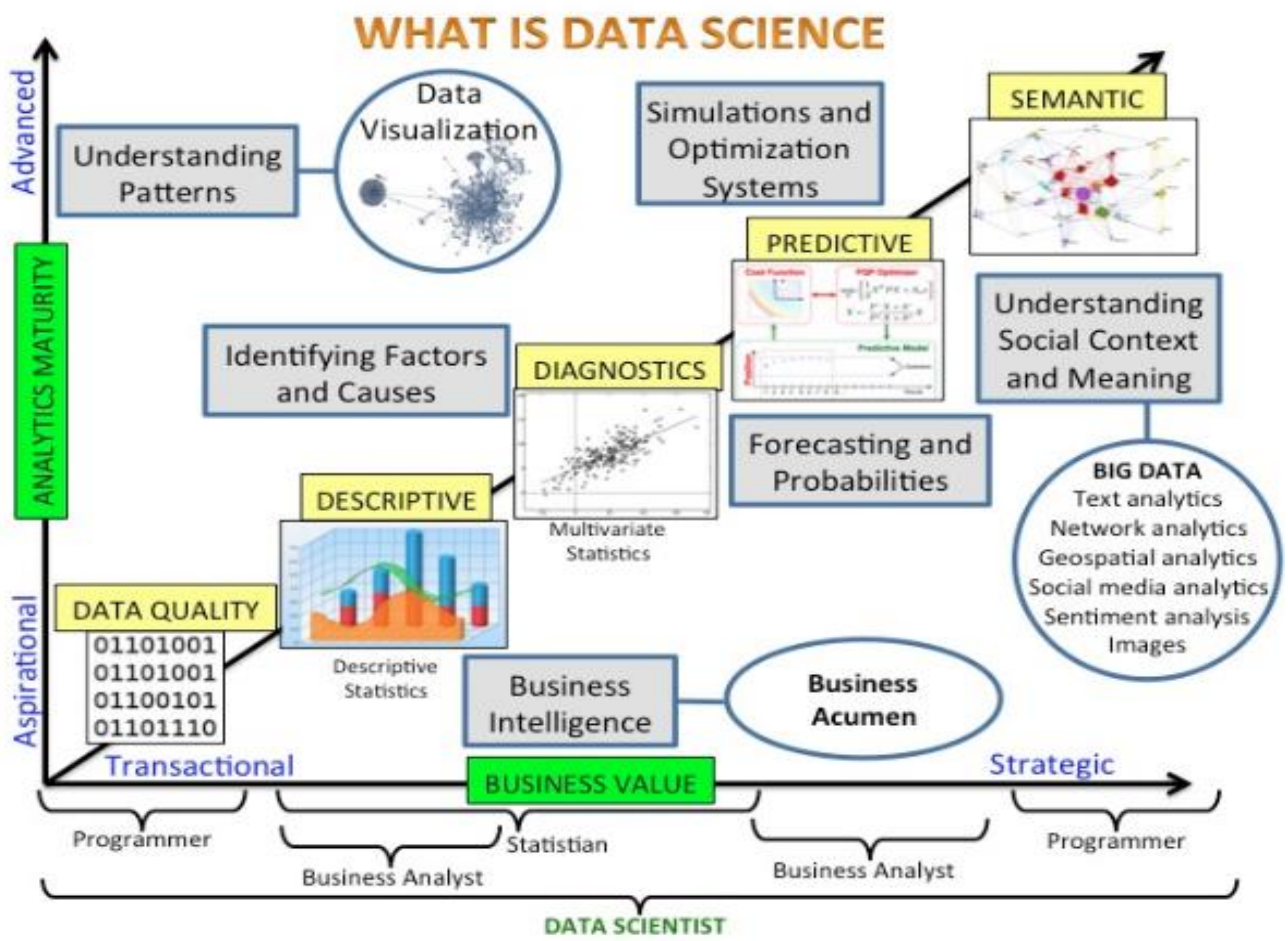
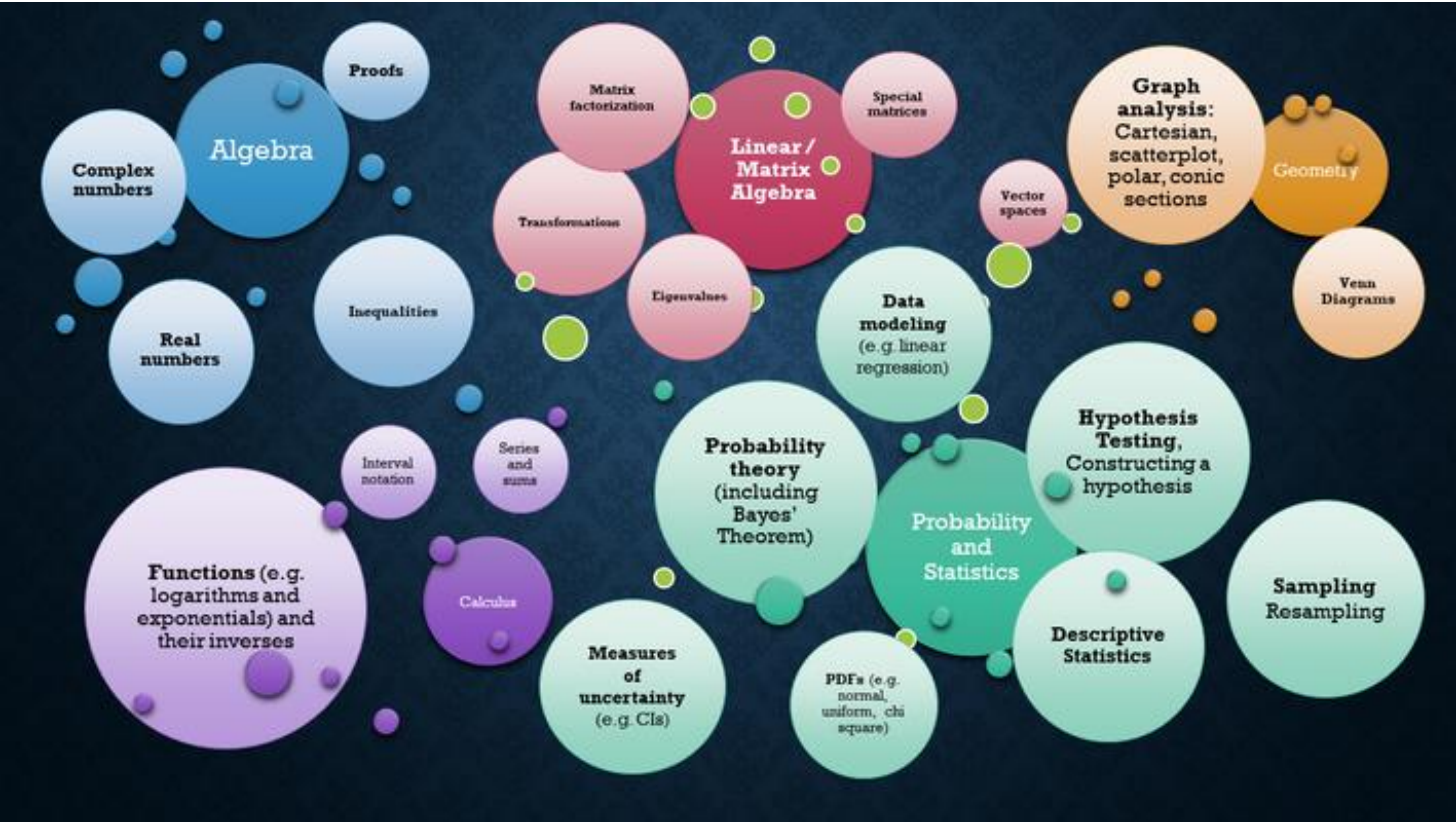
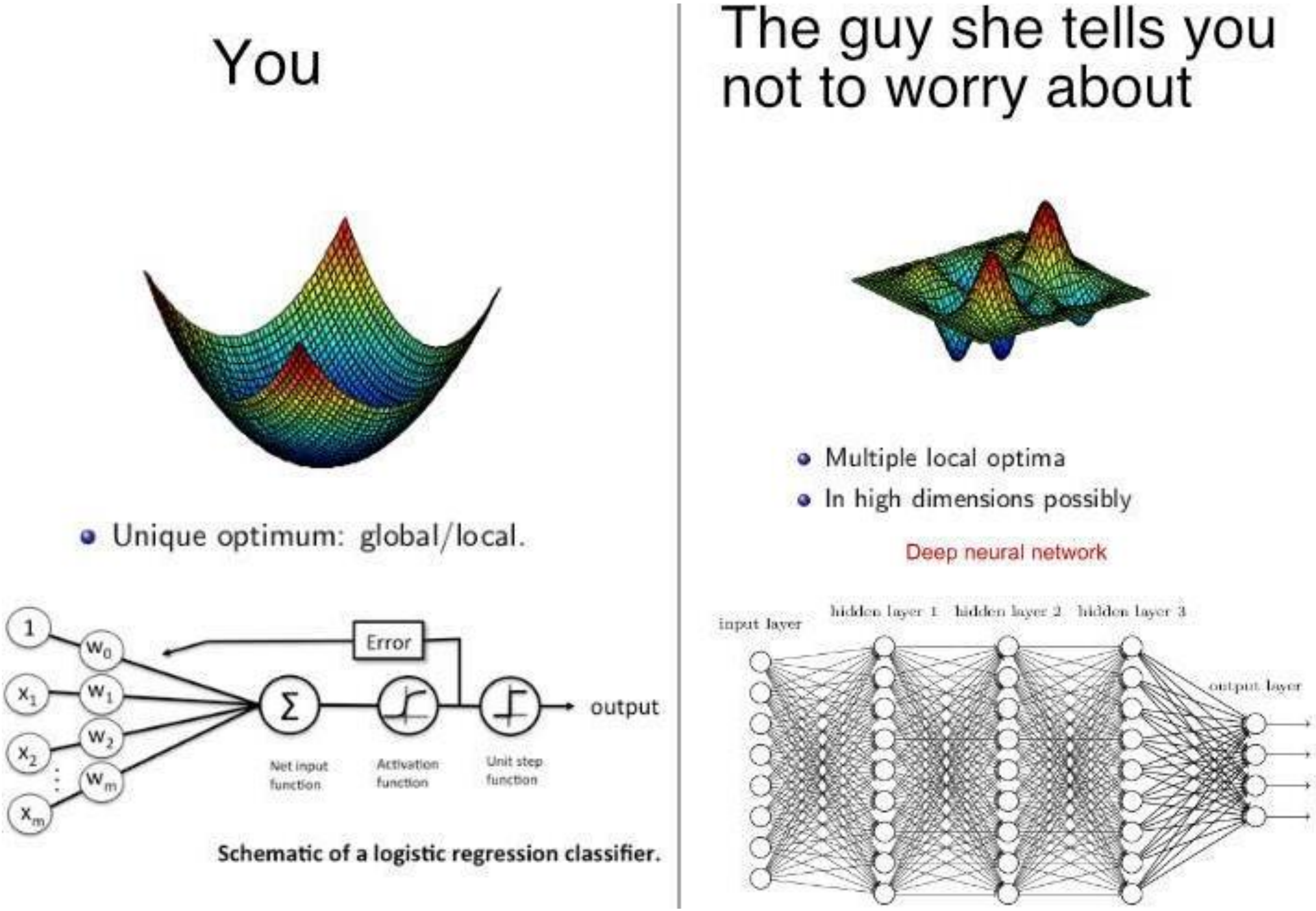

WHAT IS DATA SCIENCE
Advanced
Aspirational
ANALYTICS MATURITY
Understanding Patterns
Data Visualization
Simulations and Optimization Systems
SEMANTIC
PREDICTIVE
Identifying Factors and Causes
DIAGNOSTICS
Understanding Social Context and Meaning
Forecasting and Probabilities
DESCRIPTIVE
Multivariate Statistics
BIG DATA
Text analytics
Network analytics
Geospatial analytics
Social media analytics
Sentiment analysis
Images
DATA QUALITY
01101001
01101001
01100101
01101110
Descriptive Statistics
Business Intelligence
Business Acumen
Transactional
BUSINESS VALUE
Strategic
Programmer
Statistian
Business Analyst
Business Analyst
Programmer
DATA SCIENTIST

Math for Data Science

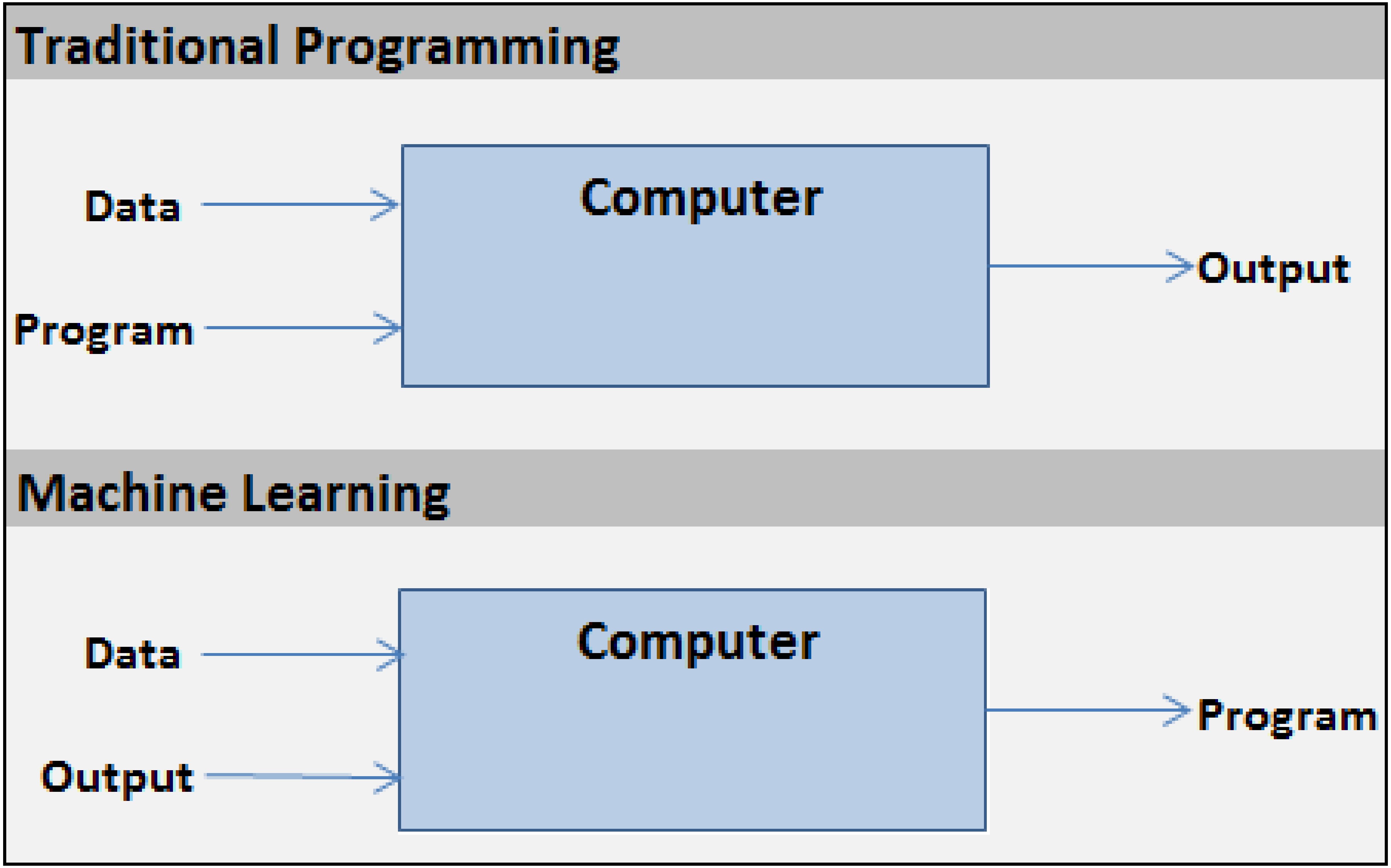


Deep learning vs Machine learning in One Picture

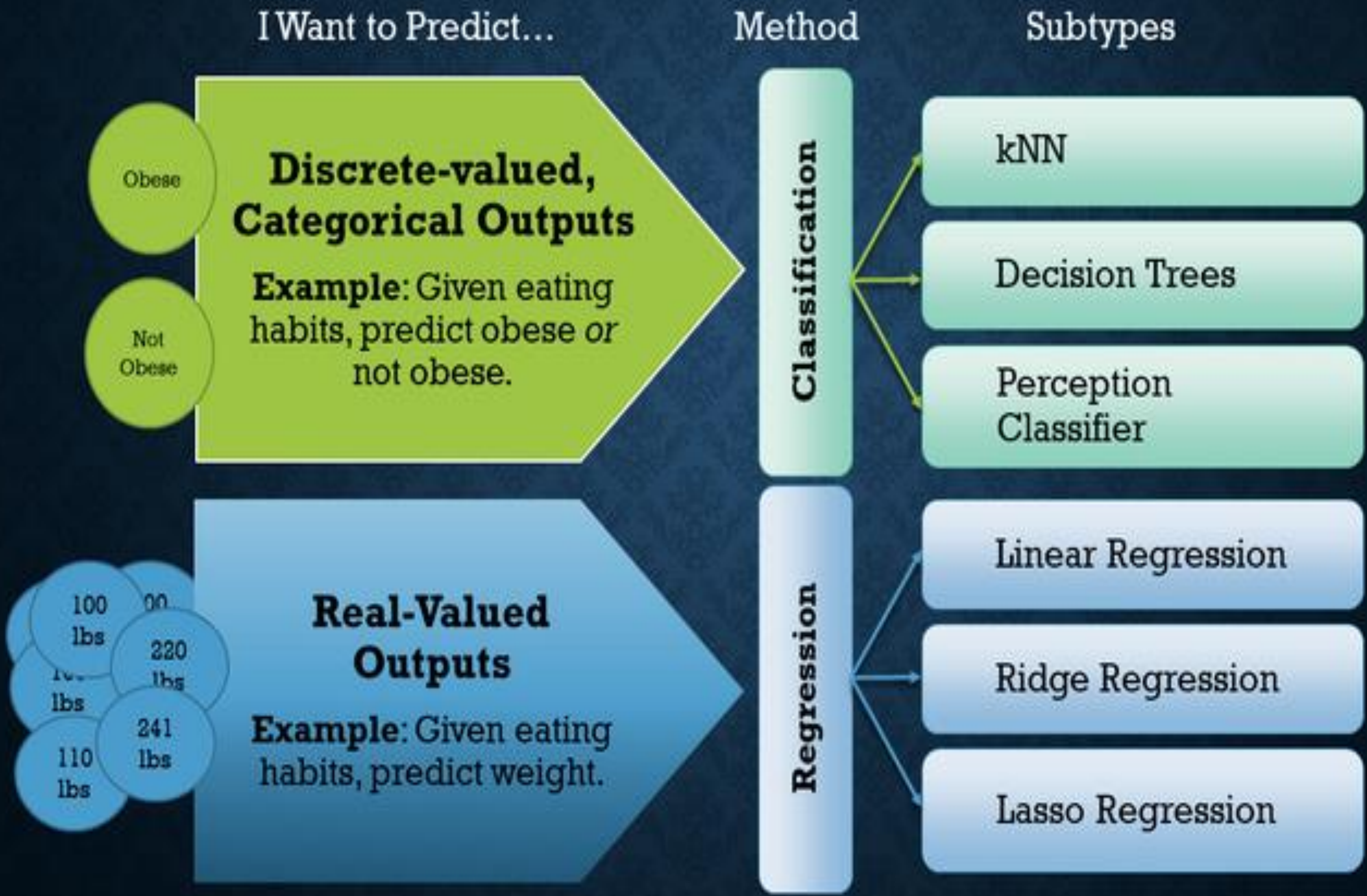


Which Software Should I Choose?	Python	R	SAS	SQL
Best for:	General programming; Data analysis; Deep learning; Repeated tasks	Statistical analysis; Data analysis; Single passes of data	Statistical analysis; Data analysis	Database manipulating, updating, querying; Extracting, wrangling data
Availability	Free, open source	Free, open source	Paid (free for university edition); Closed source	Open and closed source versions available (free and paid)
Easy to learn?	Yes, especially for software engineers	Steep learning curve; Relatively easier if no prior coding experience	Yes, especially if you already know SQL	Relatively easy for basic level; Learning curve for more complex tasks
Advantages	Easy to deploy; General purpose language; Widely used by corporations	Minimal coding required for statistical models	Highly reliable, secure and stable	Very readable
Disadvantages	Requires rigorous testing	Very statistics oriented; Not a general-purpose program	Relatively expensive	Not general purpose; very specific, limited capability

Traditional Programming vs ML in One Picture



	Prediction: *Predicting Y from X*	**Inference:** *Understand relationship between X and Y*
Goal	Develop a "best" model (considering all predictors) to predict Y with high accuracy, low error.	Estimate an association between and outcome variable and a predictor variable (while adjusting for confounders).
Answers the question	How can I accurately predict new data points?	What do the relationships between the variables mean?
Example	What mortality levels does the model predict given a certain income and education level?	Which has the biggest impact on mortality: income or education?


I Want to Predict…
Method
Subtypes
Obese
Not Obese
Discrete-valued, Categorical Outputs
Example: Given eating habits, predict obese or not obese.
Classification
kNN
Decision Trees
Perception Classifier
100 lbs
220 lbs
241 lbs
110 lbs
lbs
Real-Valued Outputs
Example: Given eating habits, predict weight.
Regression
Linear Regression
Ridge Regression
Lasso Regression

1. State the Null Hypothesis

"Null" means that it's a **commonly accepted fact** that you are working to *nullify*.

Example: the null might be "Data scientists earn an average of $113,309"

$H_0: \mu = 133,309$

2. State the Alternate Hypothesis

The **alternative to the null**.

For example, you think that data scientists earn a lot less than 100k, and you're going to set out to prove your theory.

$H_1: \mu < 133,309$

3. Choose Your Alpha Level

The alpha level α is **the probability of making the *wrong* decision** when the null hypothesis is true.

Not sure what alpha level to set? 5% is very common.

$\alpha = 0.05$

4. Collect Your Data

You may already have the data at hand.

If not, you **have four main choices**:

- *Census*
- *Sample survey*
- *Experiment*
- *Observational survey*

5. Calculate the Test Statistic

Which test statistic you use depends on which test you're running. **Choosing the right test** is probably the biggest challenge in hypothesis testing.

Four of the most common:

- *Z-Statistic*
- *T-Statistic*
- *F-Statistic (ANOVA)*
- *Chi-Square Statistic*

6. Make Accept / Reject Regions

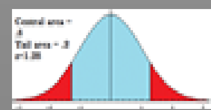
The **rejection region** tells you at what values you should reject the null hypothesis.

The **critical value** (the line that separates accept from reject) is calculated with statistical tables (or software, based on those tables).

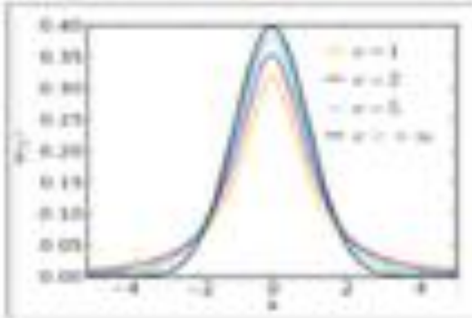
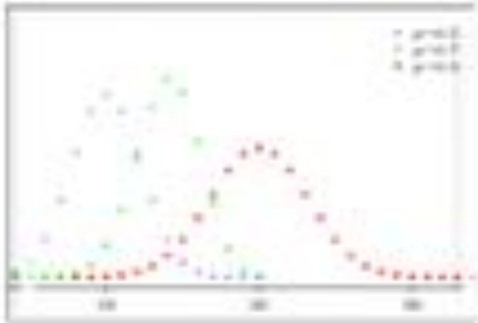
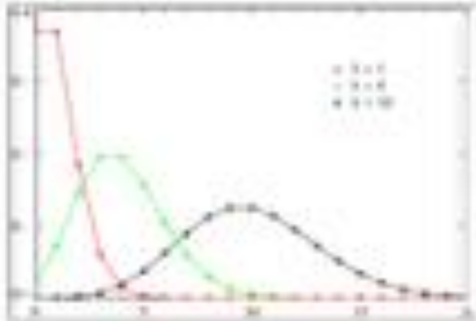
7. Accept or Reject the Null Hypothesis

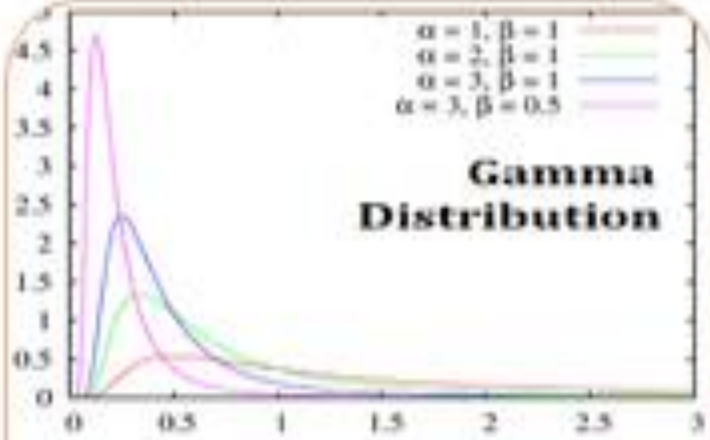
The basic idea is that your test statistic is going to fall into one of two areas: *accept* or *reject*.

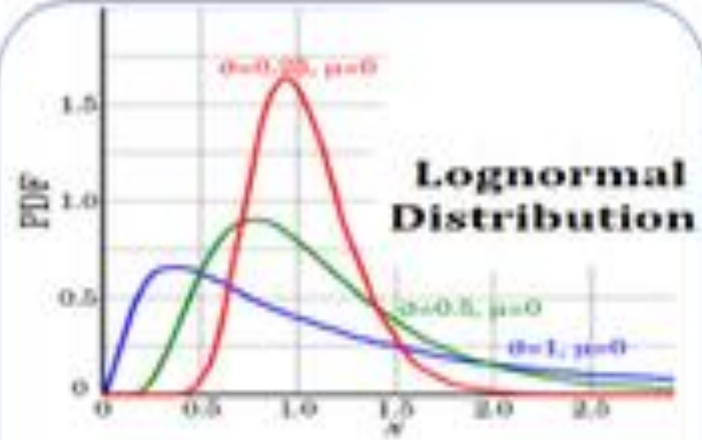
In this graph, you would **reject the null hypothesis** if your test statistic falls into the red area.

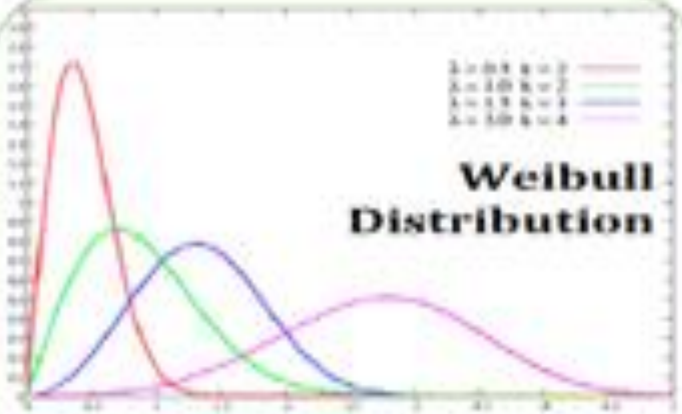


Type of Distribution	Binomial *How many Successes in n trials?*	Poisson *Good for rare events (large n, small p)*	Hypergeometric *Use for small populations, without replacement*
Number of trials, n	Fixed n	Unknown n (it is a random variable) and potentially infinite.	Fixed n
Probability, p	Known p. Probability of Success is constant from trial to trial	Unknown p for each trial (but known average p).	Probability for each trial changes (because of no replacement).
With or without replacement?	With replacement	With replacement	Without replacement
When to use it?	You know the *exact* probability of an event happening; you want to find the probability of that event happening k times out of n.	You know the *mean* probability of an event and you want to find the probability of n events happening.	Samples are small compared to the population ($n \geq 5\%$) . Binomial is easier and provides a good approximation if you have a large population.
Example	Number of defects in a box of 1,000 factory produced widgets	Number of innocent people convicted of a crime.	30 people (14W, 16M) apply for two jobs. What is the probability both positions are filled by women?

	Normal Distribution	Student's T-Distribution	Binomial Distribution	Poisson Distribution	Exponential Distribution
What does it look like?					
Defining Characteristics	Distinctive Bell Shape	Shorter, fatter than the normal distribution.	Two outcomes: Success/Failure	Various shapes, but valid only for integers on the x-axis.	Models Time Between Events
Example of When to Use It	Modeling natural phenomena (height, weight, IQ, test scores etc.)	When you have small samples or don't know the population variance (σ^2).	Coin Toss Probability (Heads, Tails)	Gives probability of number of events in a fixed interval.	"How much time will go by before a major hurricane hits the Atlantic Seaboard?"
Example of DS Application	Least squares fitting or propagation of uncertainty.	Unknown σ^2 is common in real life data, you you'll have to use the T instead of the normal in that case.	Anywhere where binary (yes/no, black/white, vote/don't vote) data is used.	Anywhere there is a waiting time between events.	Building continuous-time Markov chains.







	Gamma Distribution	Lognormal Distribution	Weibull Distribution
Parameters	Rate(β); Shape (α)	Location (μ); Shape (σ)	Location (μ); Shape (γ); Scale (α)
Overview	A family that includes the exponential distribution	Exponential (anti-log) of the normal distribution	A natural extension of the exponential distribution
Common Uses	Event waiting times when the event process isn't completely random (Mun, 2008)	Model lifetimes in survival analysis /reliability (Bromideh, 2012)	Describe the distribution of lifetime data (Bromideh, 2012) and the time to wait for the next event.
When to Use	Most often used as time to a *r*th hit in a Poisson process	Use when data has small bumps/perturbations that additively affect the result	Use when the exponential doesn't quite fit or when you have steep drop-offs.

THE GAUSSIAN CORRELATION INEQUALITY

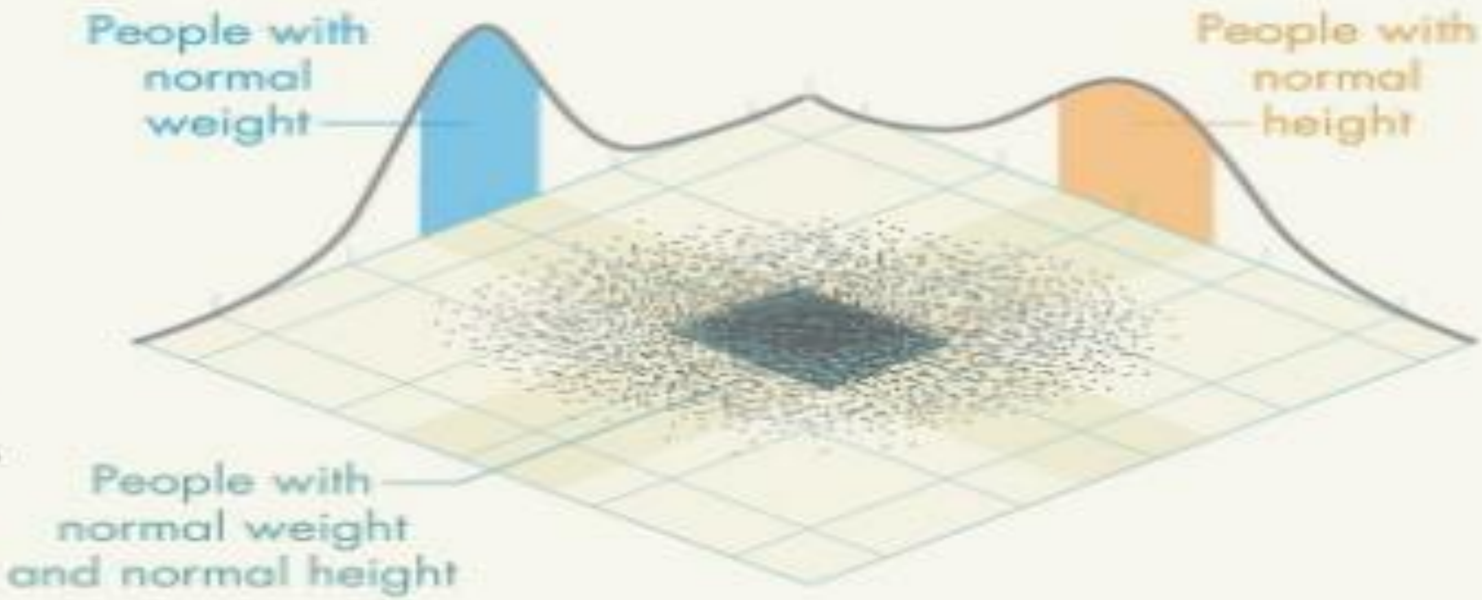
... in convex geometry:

Consider any two convex symmetrical shapes in any number of dimensions that are centered on the same point, which forms a target. Darts thrown at the target will land in a bell curve or "Gaussian distribution" of positions. The overlap of the two shapes increases your probability (P) of striking both.

P(Dart lands in both circle + rectangle) ≥ P(Dart lands in circle) x P(Dart lands in rectangle)

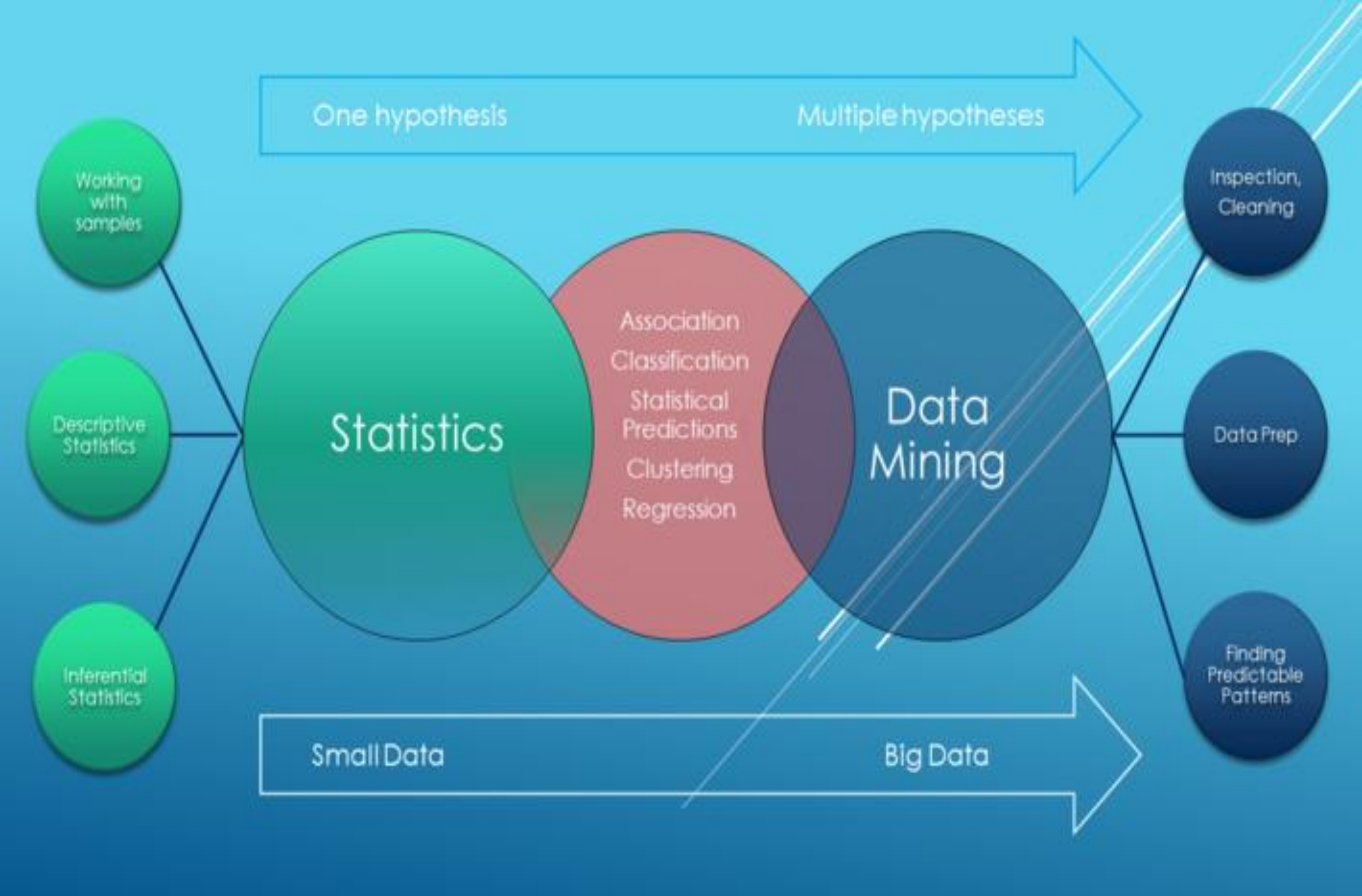
... in multivariate statistics:

Consider a sample of people's weights and heights plotted on *x-y* axes. Because these variables are correlated, the odds that someone's weight and height will both fall within a combined range is greater than or equal to the product of the independent odds of falling in each range. (The general inequality holds for any number of variables.)

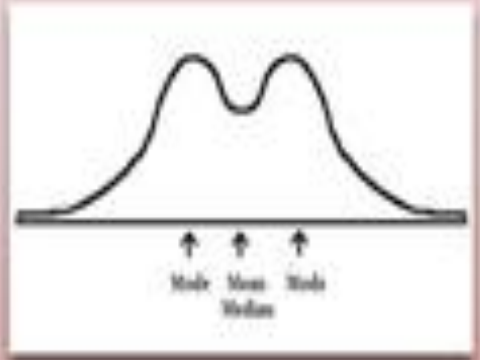


P(Person has both normal weight + height) ≥ P(Person has normal weight) x P(Person has normal height)

	Overview	Process	Subtypes	Examples
Supervised Learning	Majority of algorithms. Machine is trained using **well-labeled data**; inputs and outputs are matched.	Mapping function takes inputs and matches to outputs, creating a target function.	Classification, Regression	Linear regression, Random forest, SVM.
Unsupervised Learning	**Unlabeled data** (inputs only) is analyzed. Learning happens without supervision.	Inputs are used to create a model of the data.	Clustering, Association.	PCA, k-Means, Hierarchical clustering.
Semi supervised	**Some data is labeled**, some not. Goal: better results than labeled data alone. Good for real world data.	Combination of above processes.	All the above.	Self training, Mixture models, Semi-supervised SVM



One hypothesis
Multiple hypotheses
Working with samples
Descriptive Statistics
Inferential Statistics
Statistics
Association
Classification
Statistical Predictions
Clustering
Regression
Data Mining
Inspection, Cleaning
Data Prep
Finding Predictable Patterns
Small Data
Big Data

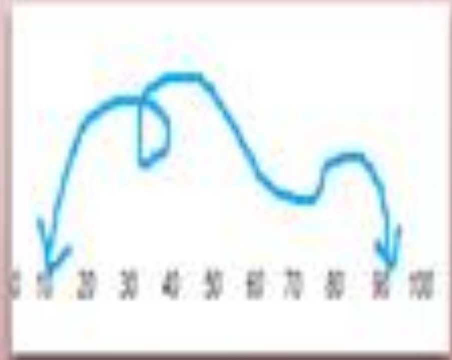


Central Tendency

Mean, Median, Mode, Outliers

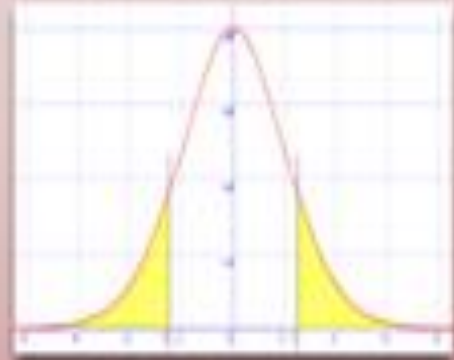
Measures of Spread

Range, Standard deviation, Variance, Quartiles



Percentiles

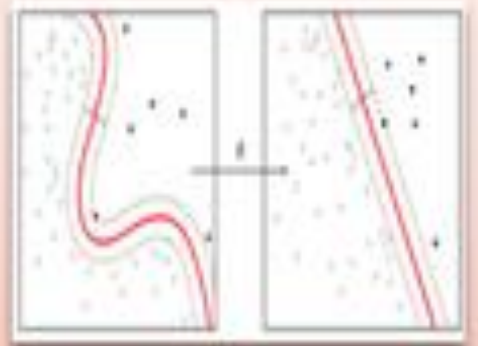
Position of data, percentile rank, percentile range



Probability Distributions:

Uniform, normal (Gaussian), Poisson

Basic Probability and Statistics



Dimensionality reduction

Pruning, PCA

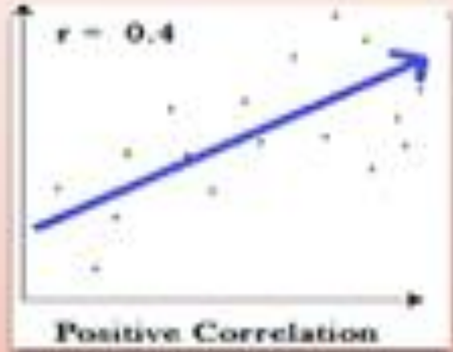
Sampling

SRS, Reservoir, Undersampling, Oversampling.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesian statistics

Measuring belief or confidence

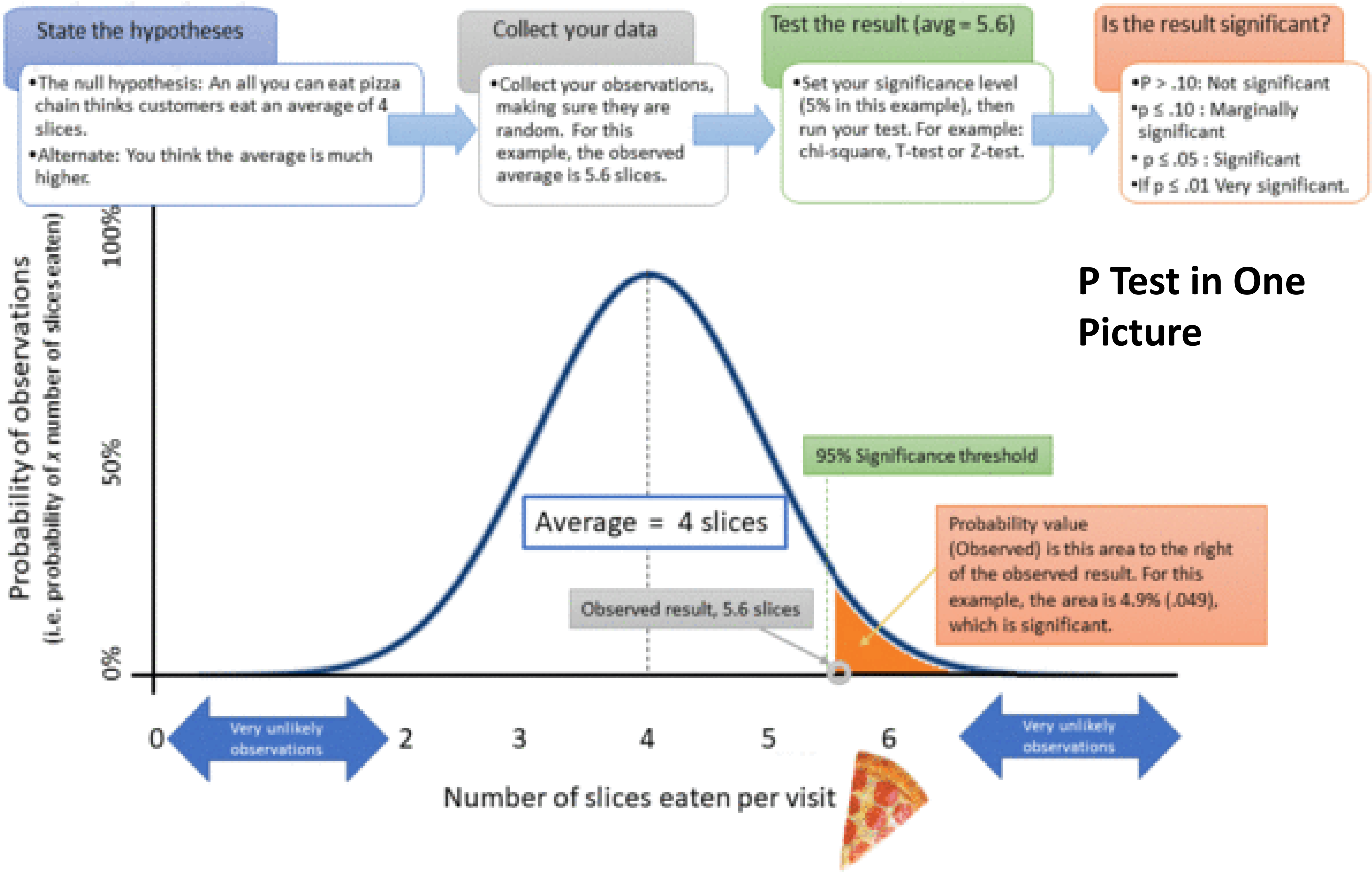


Covariance & correlation

How data is related

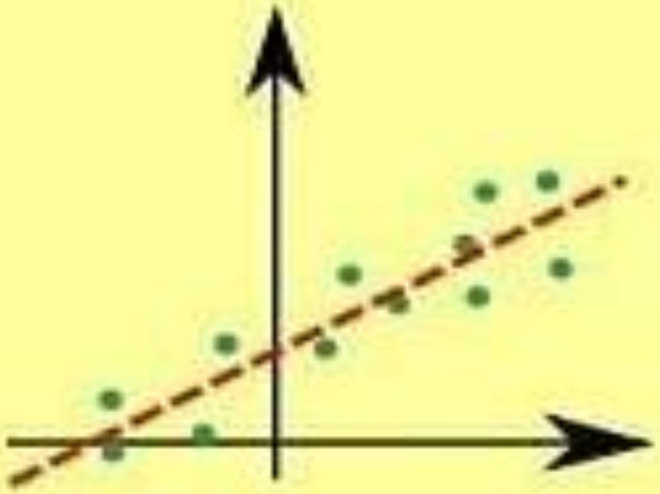
More Advanced Probability and Statistics

P Test in One Picture



LINEAR REGRESSION

1. Econometric modelling
2. Marketing Mix Model
3. Customer Lifetime Value



Continuous ⇒ Continuous

$$y = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$$

lm(y ~ x1 + x2, data)

1 unit increase in x increases y by α

LOGISTIC REGRESSION

1. Customer Choice Model
2. Click-through Rate
3. Conversion Rate
4. Credit Scoring

Continuous ⇒ True/False

$$y = \frac{1}{1 + e^{-z}}$$

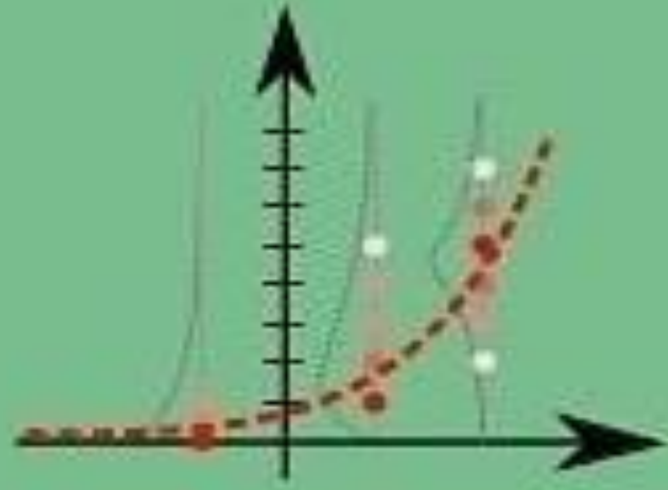
$$z = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$$

glm(y ~ x1 + x2, data, family=binomial( ))

1 unit increase in x increases log odds by α

POISSON REGRESSION

1. Number of orders in lifetime
2. Number of visits per user



Continuous ⇒ 0,1,2,...

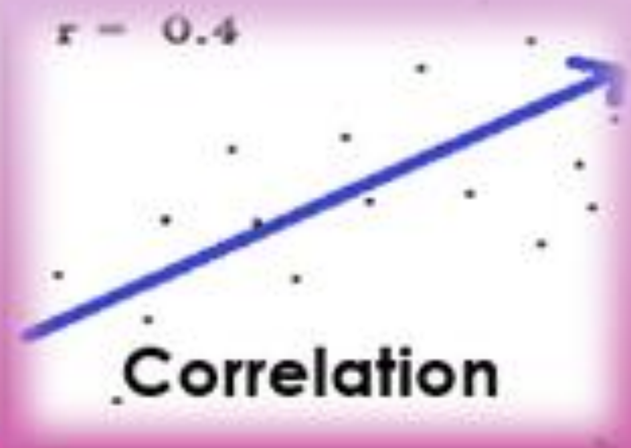
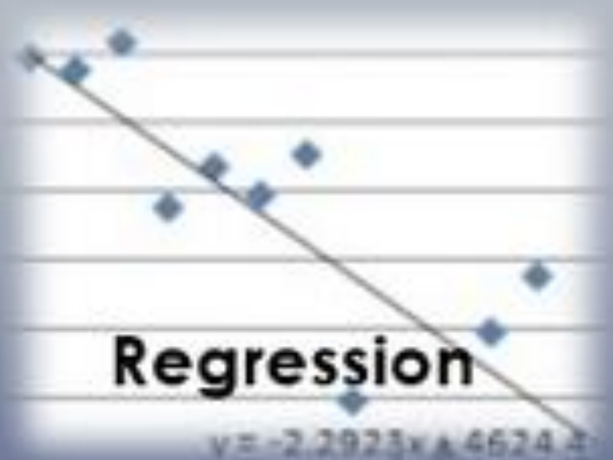
$$y \sim Poisson(\lambda)$$

$$ln\lambda = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$$

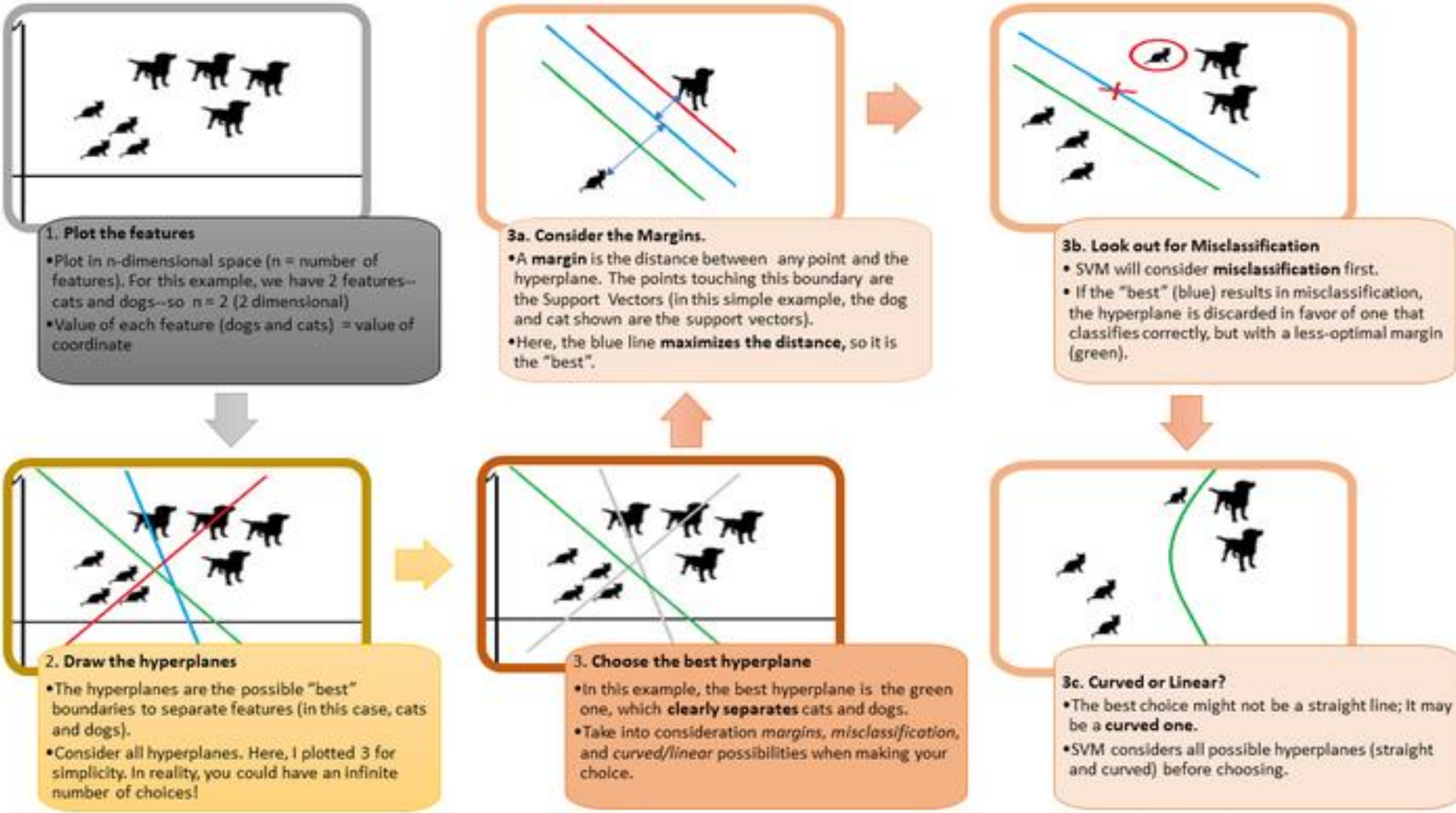
glm(y ~ x1 + x2, data, family=poisson( ))

1 unit increase in x multiplies y by e^{α}

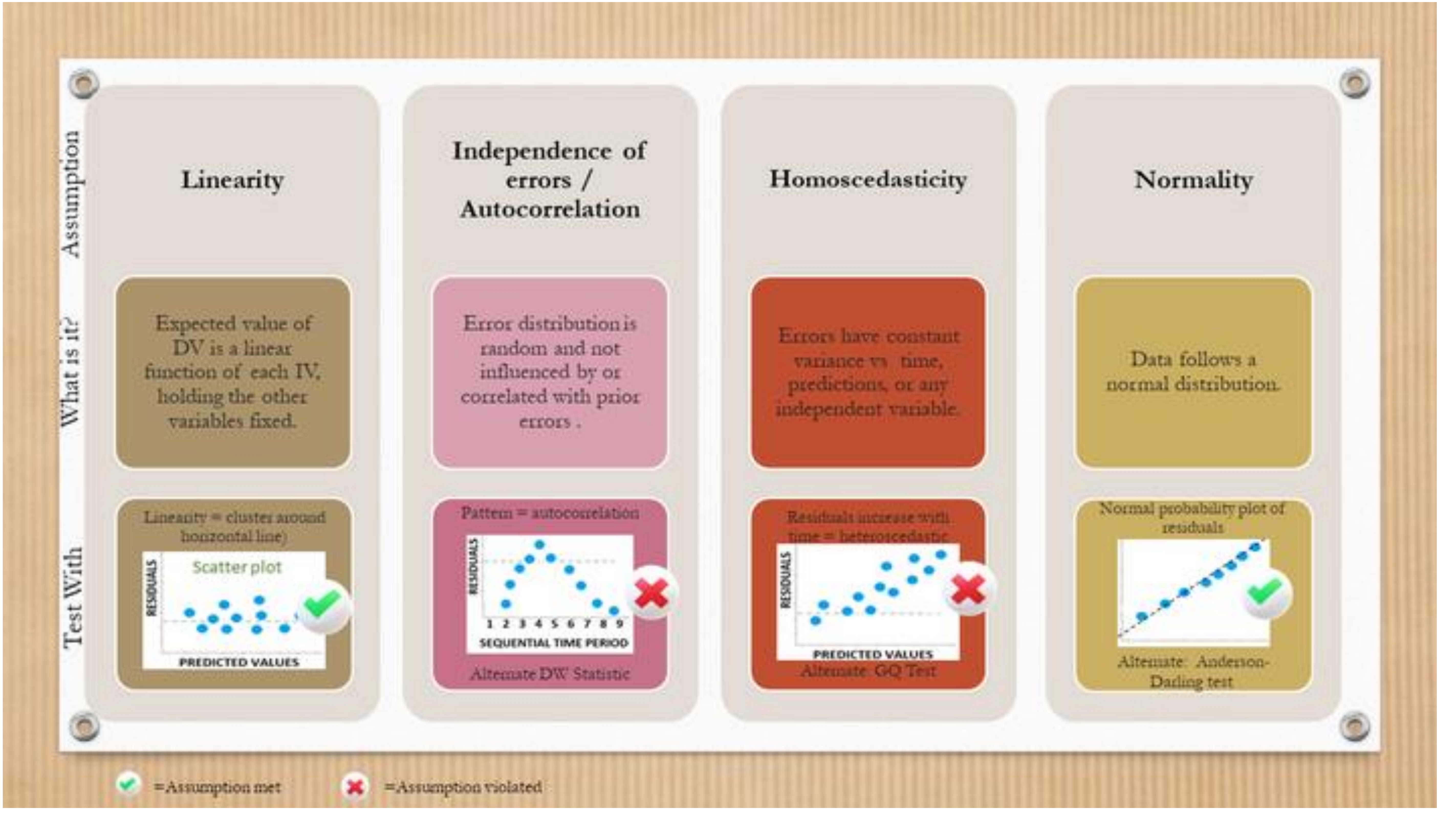
Differences Between Pearson Correlation and Linear Regression

	Use when...	Results	Assumptions (Requirements)
	You want to know if there's **an association between two variables.**	**Strength & direction of relationship (r):** -1 = perfect negative, +1. = perfect positive, 0 = no correlation.	**Validity:** Valid measurements, a good sample, unconfounded comparisons. **Distribution:** Linear relationship between the two.
	You want to predict **how one variable will change with another.**	**Estimates of parameters for a regression equation:** The B_0 and B_1 in the linear regression equation $Y = B_0 + B_1 * X$	**All the above, plus:** 1. Quantitative Data 2. Outlier Condition 3. Independence of Errors 4. Homoscedasticity 5. Normality of Error Distribution

SVM in One Picture



Assumption of Logistic Regression in One Picture



Logistic Regression in One Picture

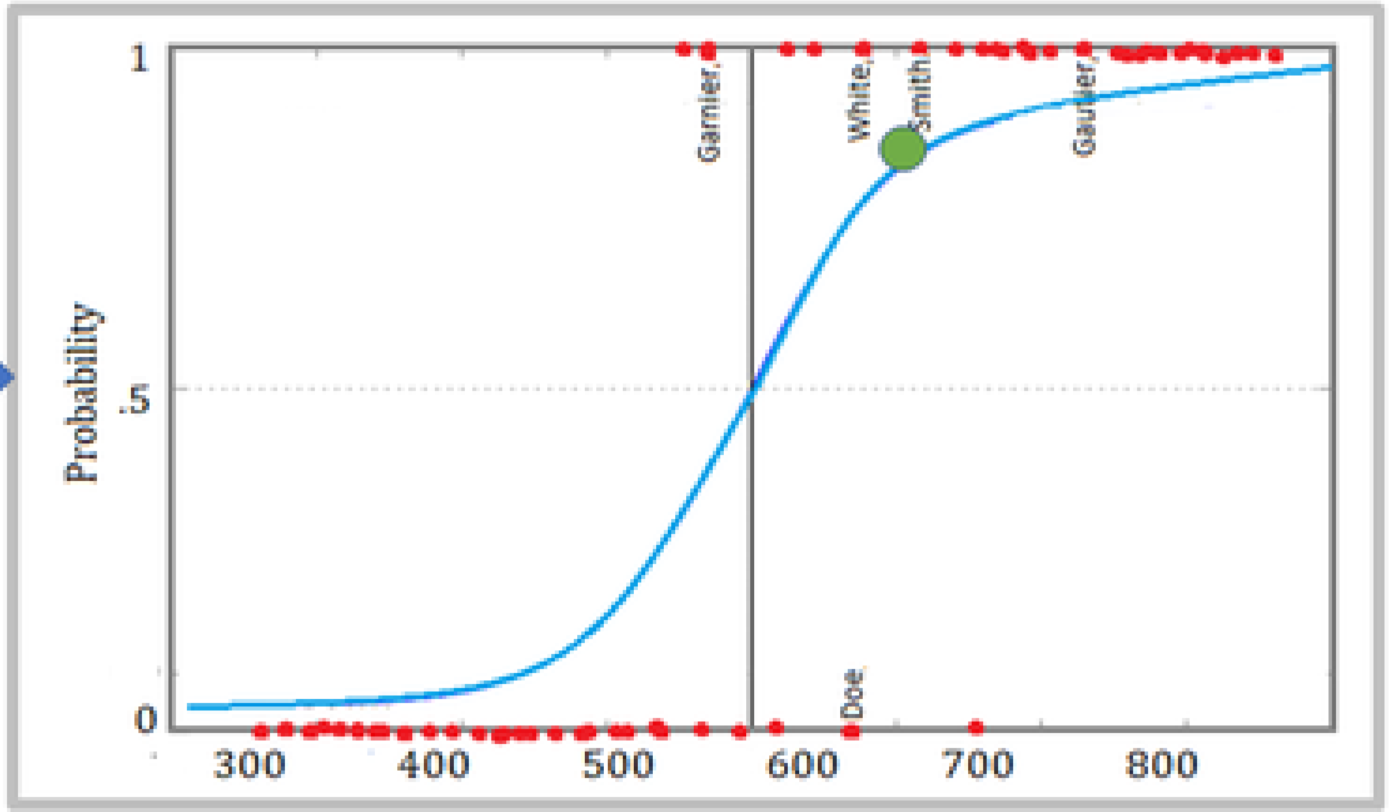
1. Collect Observations

Observations should be classified by success (1) or failure (0). For example, data on 1,000 people, their credit scores, and whether they got approved (1) or denied (0) for a car loan.

N = 1,000
Smith, J., 660, approved (1)
Doe, J. , 630, denied (0)
Garnier, B., 550, approved, (1)
Gautier, P., 750, approved, (1)
White, Z., 640, approved (1) . . .

2. Build Model

Data points (shown in red in this example) are plotted on a traditional scatter plot. They are binary, so will be clustered on either the "0" line at the bottom, or the "1" line at the top. A regression equation is calculated, which typically follows an S-shape (shown in blue).

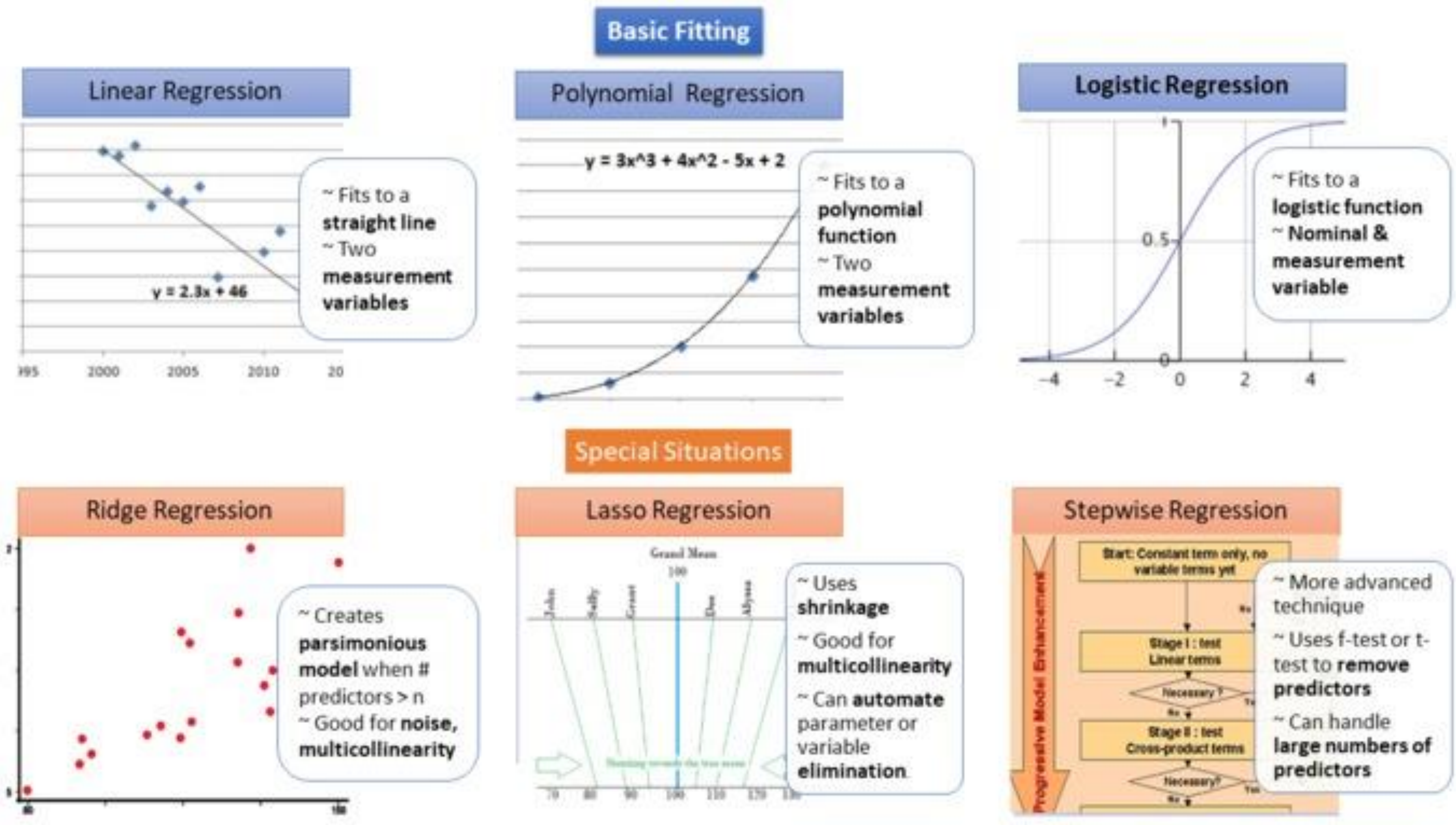


3. Read model's output

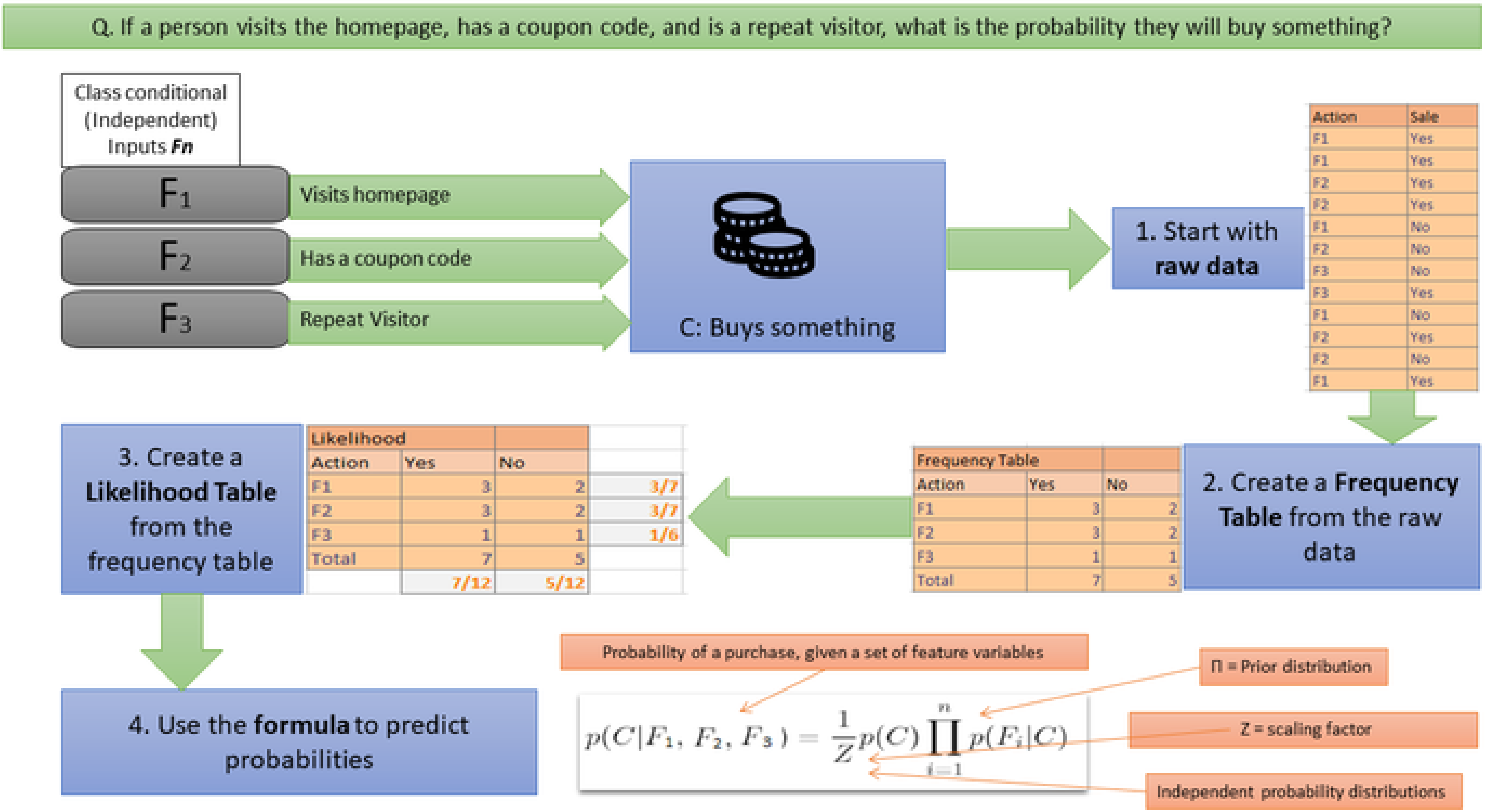
For example, "If I feed a credit score of 660 into the model, what is the probability the person will get approved for a car loan?"

Score of 660 (green dot on graph) = .8

Regression Analysis in One Picture



Naïve Bayes in One Picture



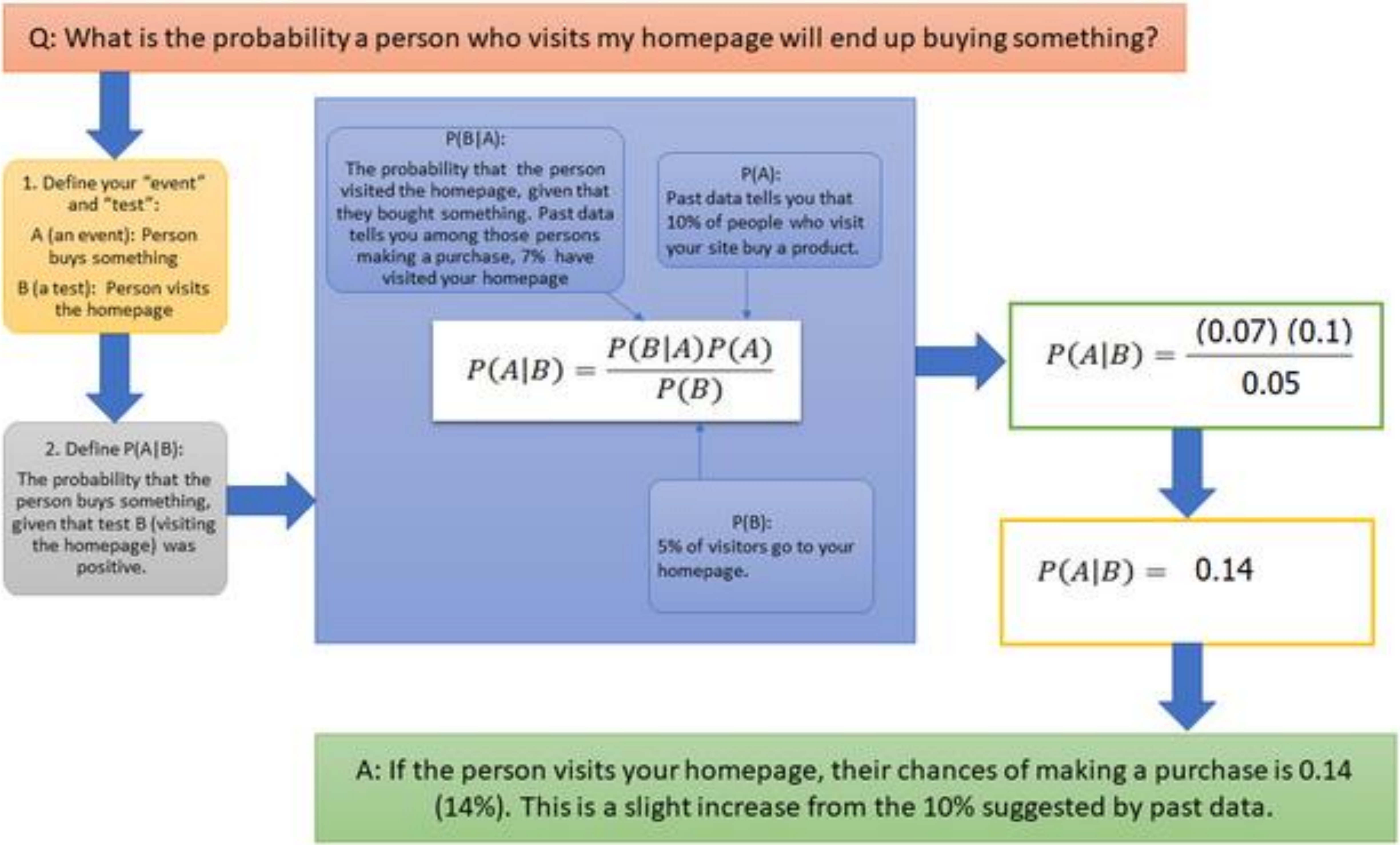
Action	Sale
F1	Yes
F1	Yes
F2	Yes
F2	Yes
F1	No
F2	No
F3	No
F3	Yes
F1	No
F2	Yes
F2	No
F1	Yes

Frequency Table		
Action	Yes	No
F1	3	2
F2	3	2
F3	1	1
Total	7	5

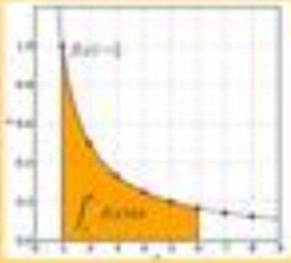
Likelihood			
Action	Yes	No	
F1	3	2	3/7
F2	3	2	3/7
F3	1	1	1/6
Total	7	5	
	7/12	5/12	

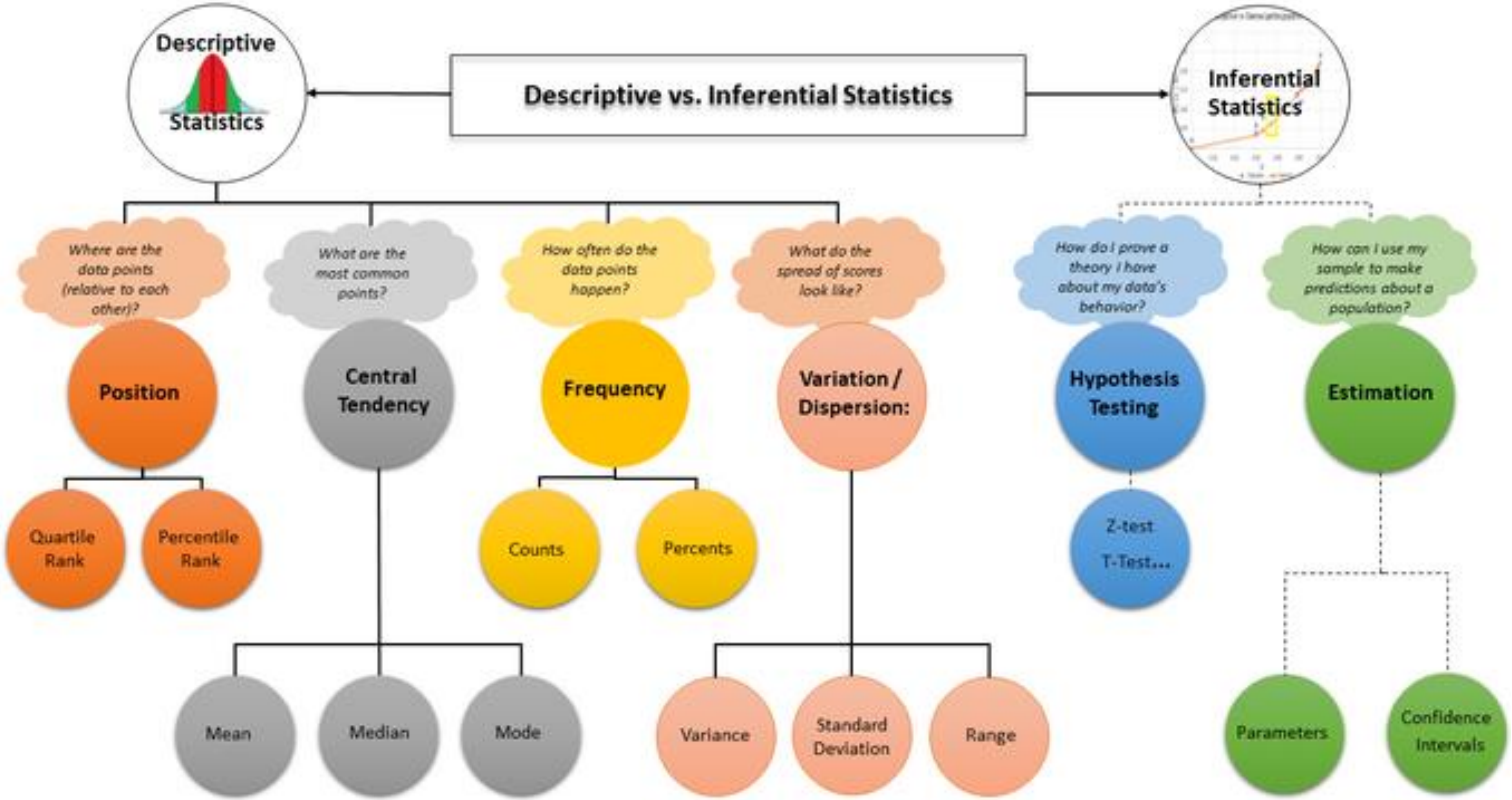
$$p(C|F_1, F_2, F_3) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

Bayes Theorem in One Picture

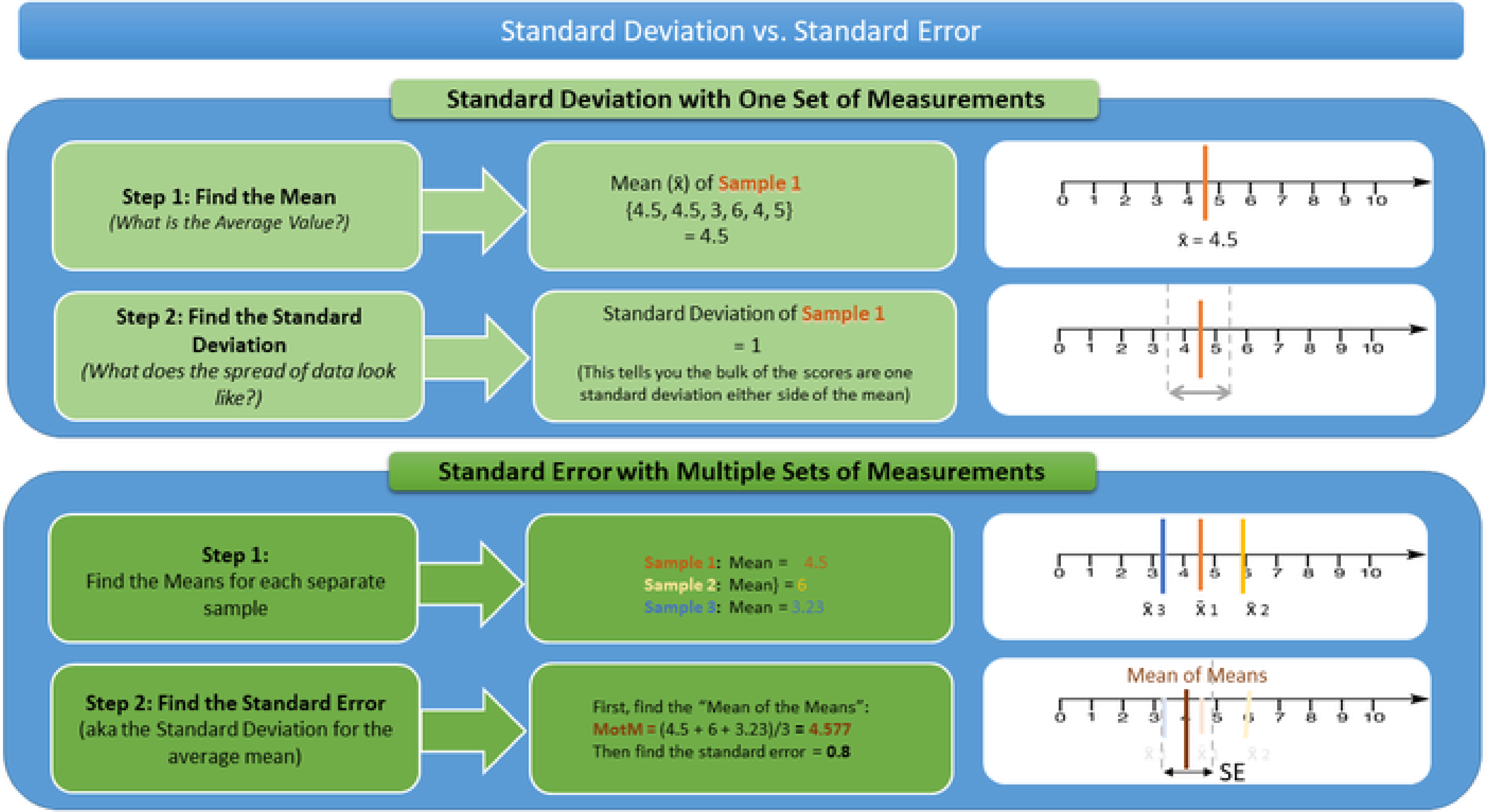


Statistics and Machine learning
In One Picture

	 Statistics	 Machine Learning	$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$ Statistical Learning
Subfield of...	Mathematics	Computer Science (AI)	Statistics & Machine Learning
Focus on...	Building models with explicitly programmed instructions	Creating systems that learn from data	Sets of tools for modeling and understanding complex data
Purpose	Inferences; Relationships between variables	Optimization; Prediction accuracy	Building statistical models for prediction; understanding data
Prior assumptions about data	Some knowledge about population usually required	None	Some knowledge about population may be required
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high dimensional data; ML learns from data	Usually applied to high dimensional data
Knowledge overlap	No ML knowledge required	Some stats knowledge usually needed: stats is basis for algorithms	Knowledge of statistics and ML required



Descriptive vs. Inferential Statistics
Descriptive Statistics
Inferential Statistics
Where are the data points (relative to each other)?
What are the most common points?
How often do the data points happen?
What do the spread of scores look like?
How do I prove a theory I have about my data's behavior?
How can I use my sample to make predictions about a population?
Position
Central Tendency
Frequency
Variation / Dispersion:
Hypothesis Testing
Estimation
Quartile Rank
Percentile Rank
Counts
Percents
Z-test
T-Test...
Mean
Median
Mode
Variance
Standard Deviation
Range
Parameters
Confidence Intervals

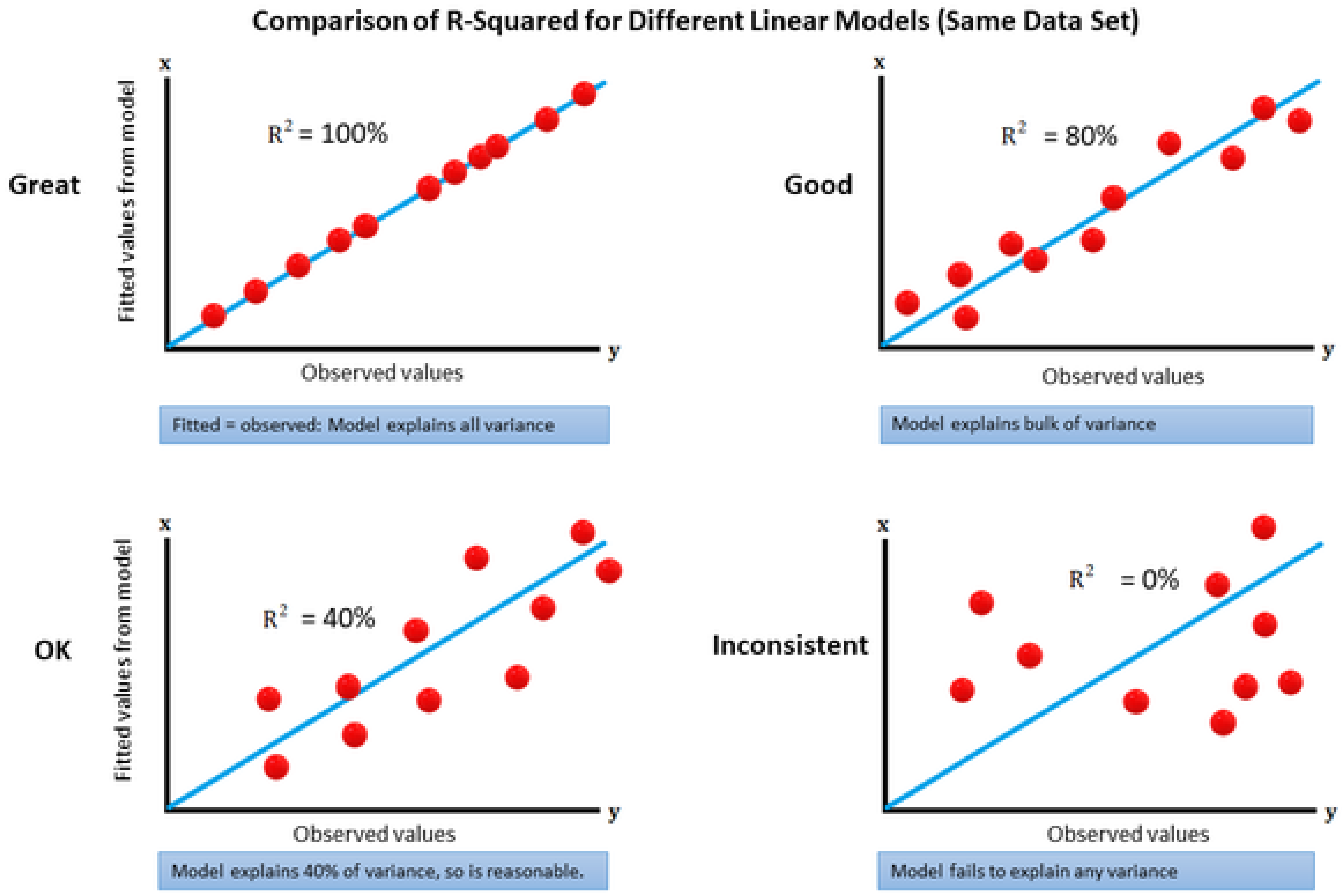


Standard Deviation vs. Standard Error
Standard Deviation with One Set of Measurements
Step 1: Find the Mean
(What is the Average Value?)
Mean (x̄) of Sample 1
{4.5, 4.5, 3, 6, 4, 5}
= 4.5
0 1 2 3 4 5 6 7 8 9 10
x̄ = 4.5
Step 2: Find the Standard Deviation
(What does the spread of data look like?)
Standard Deviation of Sample 1
= 1
(This tells you the bulk of the scores are one standard deviation either side of the mean)
0 1 2 3 4 5 6 7 8 9 10
Standard Error with Multiple Sets of Measurements
Step 1:
Find the Means for each separate sample
Sample 1: Mean = 4.5
Sample 2: Mean} = 6
Sample 3: Mean = 3.23
0 1 2 3 4 5 6 7 8 9 10
x̄ 3
x̄ 1
x̄ 2
Step 2: Find the Standard Error
(aka the Standard Deviation for the average mean)
First, find the "Mean of the Means":
MotM = (4.5 + 6 + 3.23)/3 = 4.577
Then find the standard error = 0.8
Mean of Means
0 1 2 3 4 5 6 7 8 9 10
SE

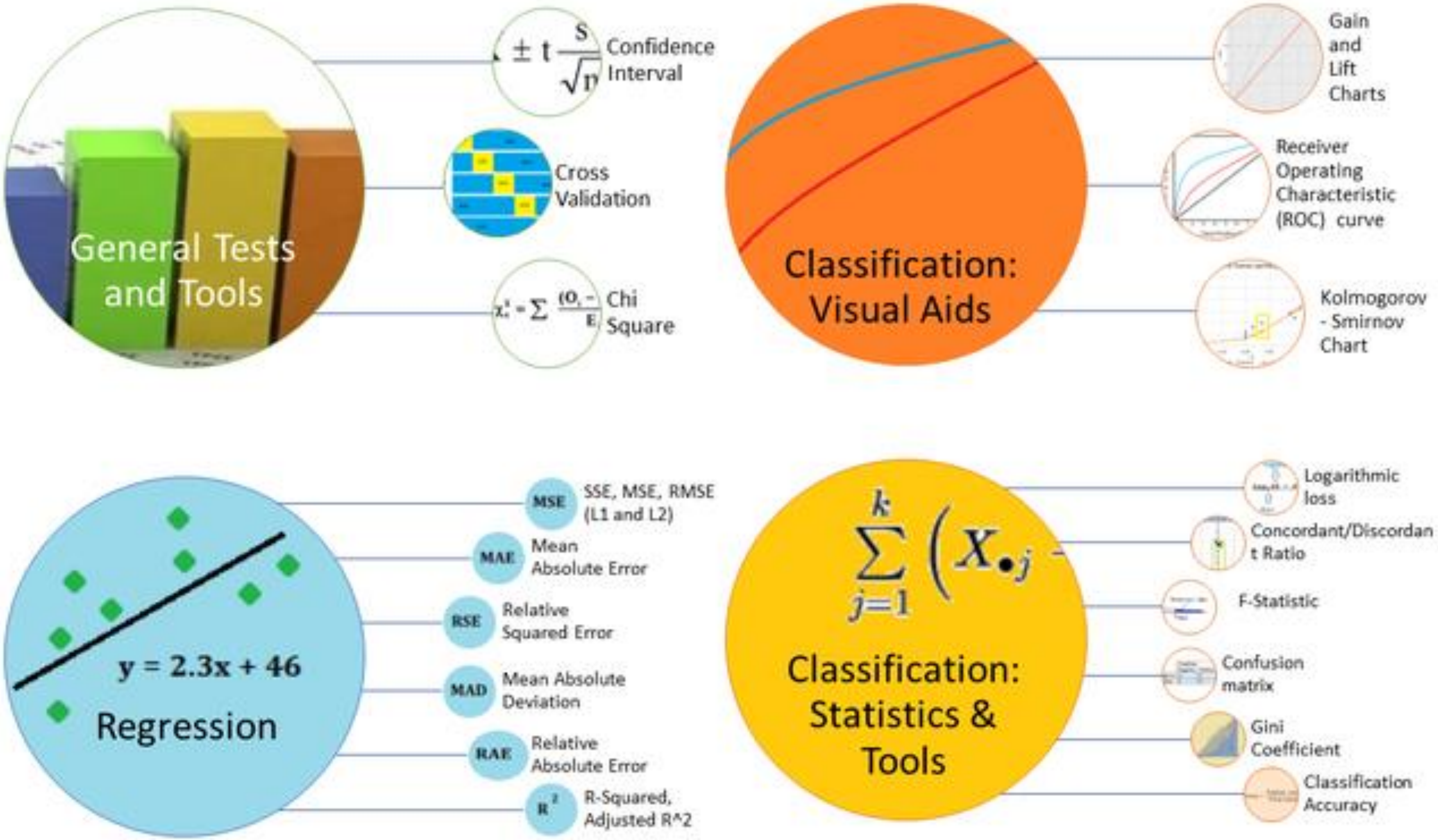
Correlation Coefficients in One Picture

	Pearson's	Spearman's Rank	Kendall's Tau	Gamma	Yule's Q
Relationships evaluated	Linear	Monotonic	Monotonic	Monotonic	Monotonic
Parametric or Non Parametric	Parametric	Non Parametric	Non Parametric	Non Parametric	Non Parametric
Input variables	Continuous	Continuous, ordinal, or ratio	Ordinal	Ordinal	Ordinal (2x2)
Advantages	Easy to use and understand Widely available	Normal distribution not required Widely available	Insensitive to errors More accurate p-value for smaller samples than Spearman's	Good for tied ranks Not affected much by outliers	Simplified version of Gamma for 2x2 table

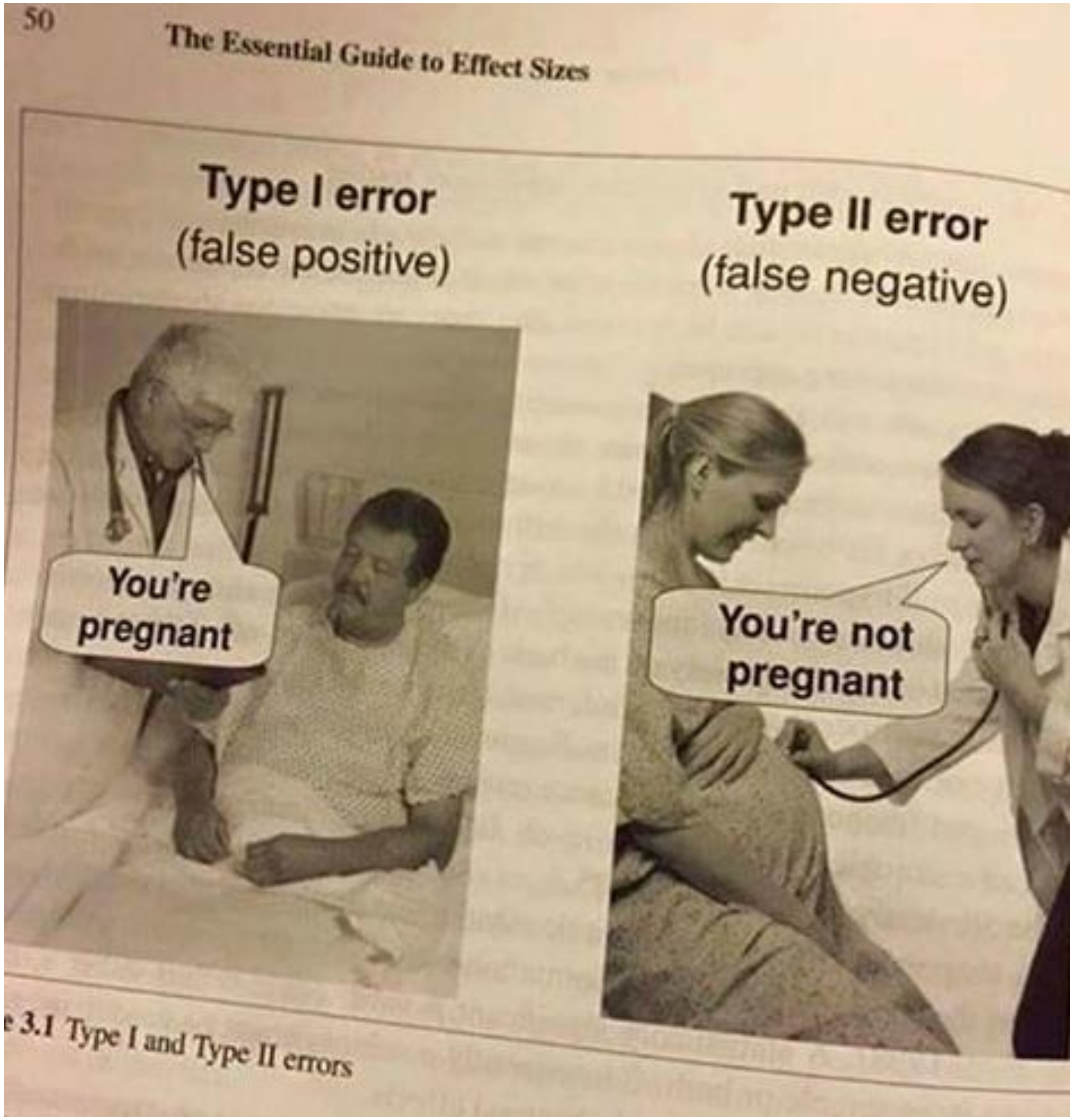
R-Square metrics in One Picture



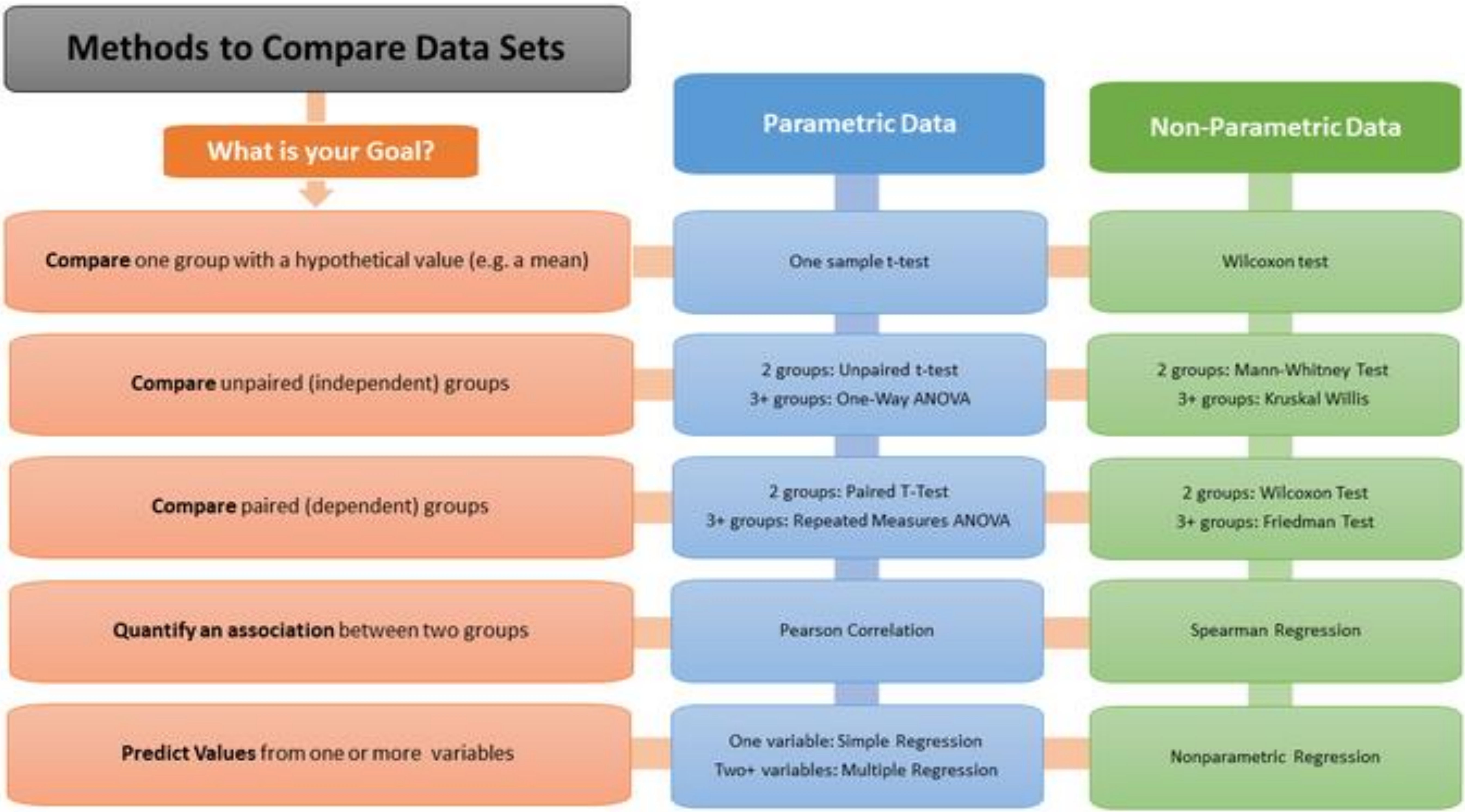
Evaluation metrics in One Picture



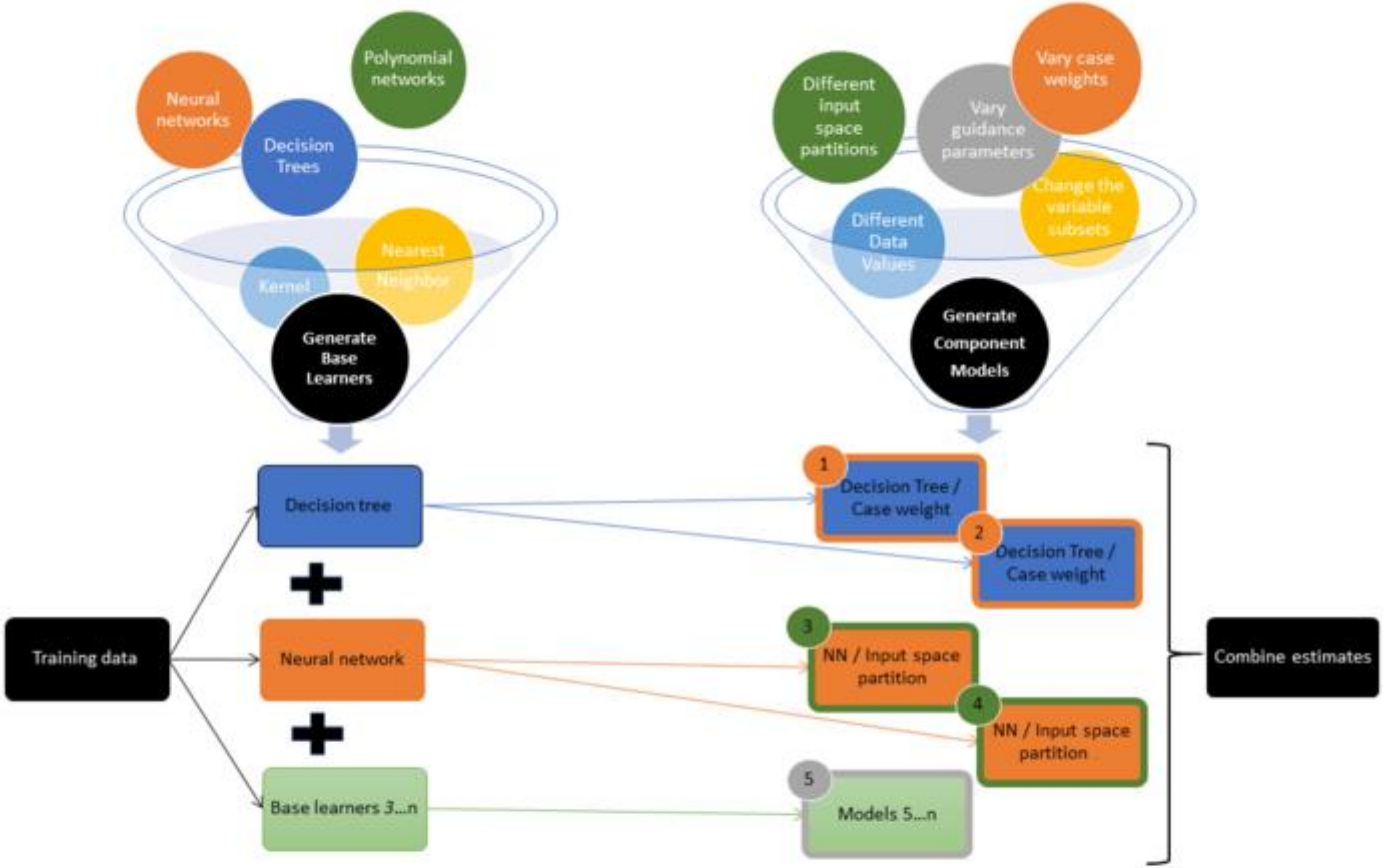
Type-I and Type-II in One Picture



Comparing Dataset in One Picture



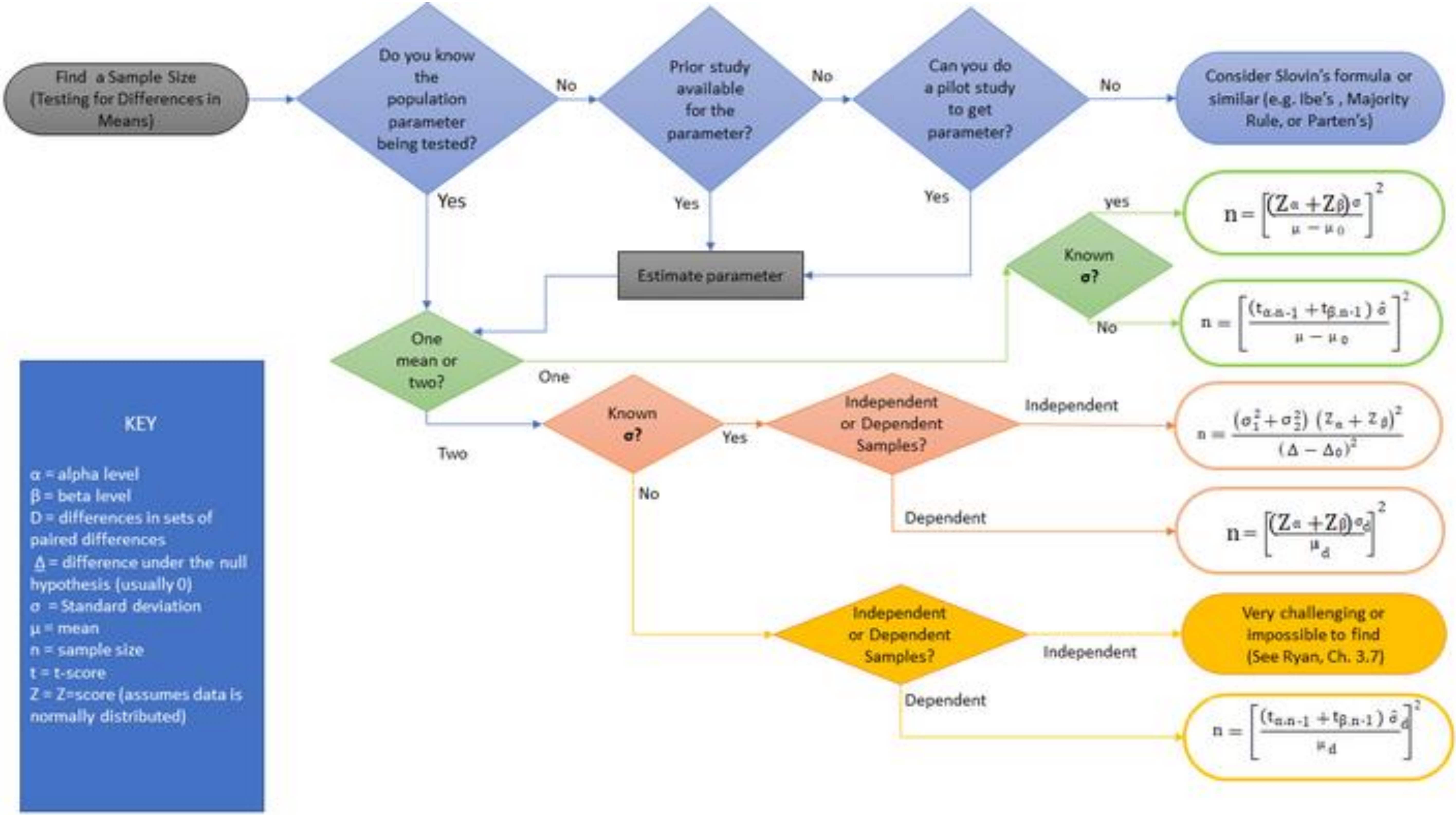
Ensemble Dataset in One Picture



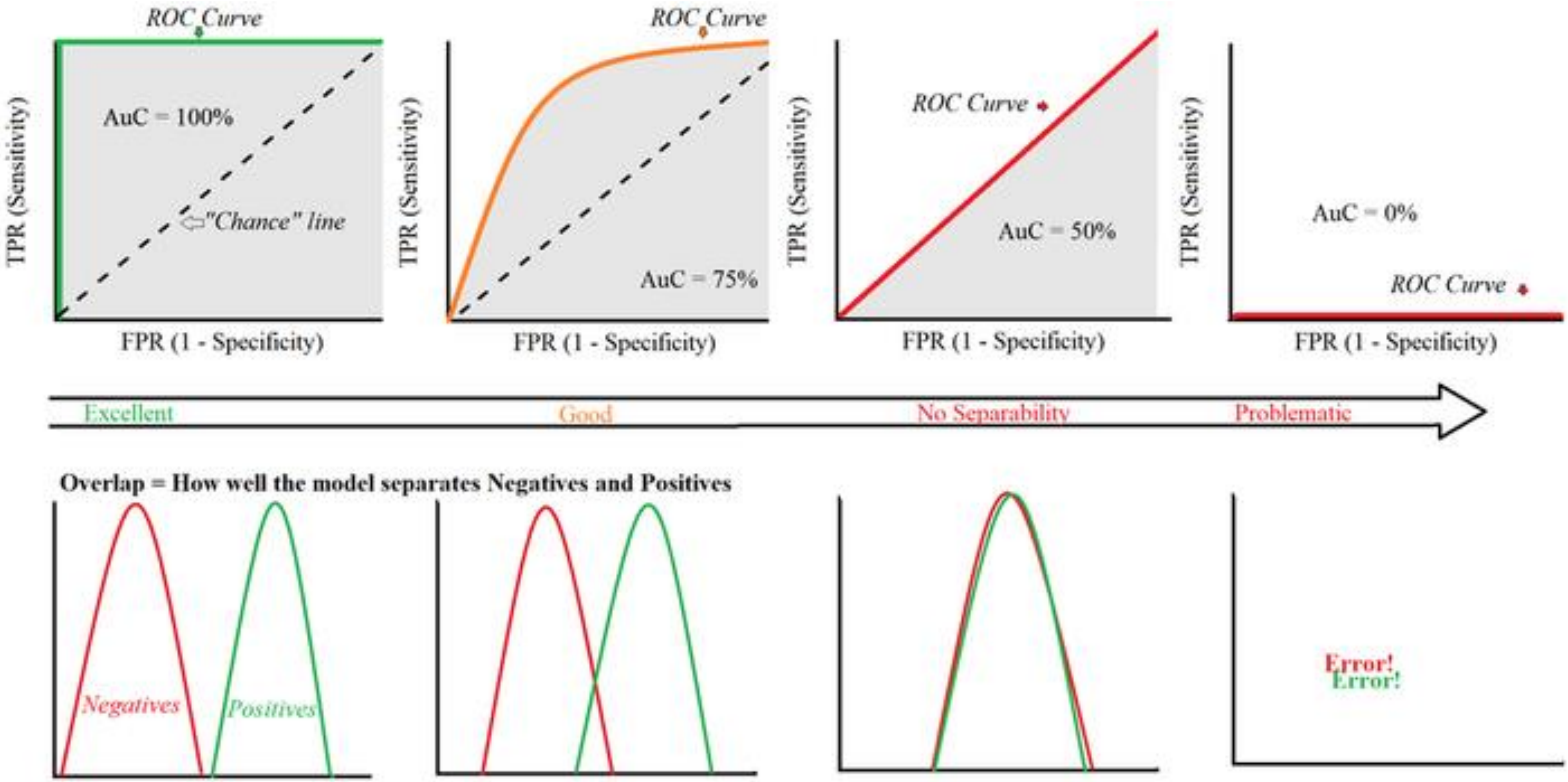
Parametric and Non-Parametric in One Picture

	Nonparametric ("Common") Bootstrap	Parametric Bootstrap	Parametric (Traditional)
Assumptions / Requirements	Large sample sizes	You know something about the distribution (e.g. the shape)	Usually, many assumptions about data (e.g. standard deviation, distribution of population) are required
Use because...	Traditional formulas not available, are too difficult, or may be less accurate.	Traditional formulas not available, are too difficult, or may be less accurate.	You have population data and are able to use traditional formulas.
Samples from...	Empirical distribution (i.e. the observations), with replacement	Parametric / hypothetical estimate of the true distribution	True population
Advantages	Can be used to estimate any sampling distribution	More powerful than nonparametric bootstrap Gives key distribution info (e.g. tail behavior)	Formulas are in wide use, many popular statistical programs / spreadsheets can do calculations
Disadvantages	Not suitable for small sample sizes: underestimates variation	Incorrect distribution choice introduces extra error-bias. Can be difficult to choose distribution.	Can only estimate sampling distributions under tight conditions & assumptions

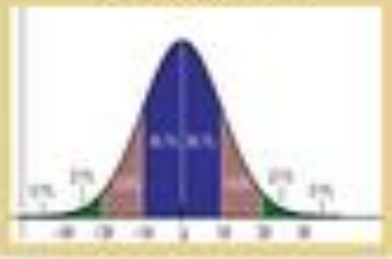
Determining Sample Size in One Picture



ROC CURVE in One Picture



Z-test and T-test in One Picture

	Subtypes	Distribution Used	Population Standard Deviation	Sample Size	Assumptions
Z Test	• One sample • Two independent samples • Two proportions	• Normal Distribution 	• Known population standard deviation	• Large samples only • (n > 30)	• Interval or ratio data • Random selection • Independent samples • Normality
T Test	• One sample • Paired T-test (dependent means) • T-test for independent means	• T-Distribution 	• Unknown population standard deviation	• Small samples to large samples	• Continuous or ordinal data • Random selection • Normality • Homogeneity of Variance
Notes	• More than two groups or levels? Use ANOVA instead	• The t-distribution is slightly shorter and fatter than the normal	• In real life, population standard deviation is usually unknown	• Sample size is a rule of thumb • For n > 30, you can choose either the T or Z	• Assumptions can differ for specific tests (e.g. independent vs. dependent means)

Anova in One Picture

	One Way ANOVA	Factorial ANOVA	Two Way ANOVA
Basic Description	Identifies differences between the means of 3+ independent (unrelated) groups	Compares mean differences between groups split on two or more independent variables	Special Case of Factorial ANOVA with 2 factors
Independent Variables (Factors)	1	2 or more (although 3/4 is usually the max due to complexity of interpreting results and higher probability of Type I errors)	2 IVs: Factor A (2 or more levels) crossed with Factor B (2 or more levels)
What is being compared in the test?	Means of three or more IV's groups on a dependent variable (although 2 groups is possible, 3 or more groups is the norm).	Effects of multiple groups of multiple IVs on a dependent variable, and on each other.	Effects of multiple groups of two IVs on a dependent variable, and on each other.
Assumptions	Continuous dependent variable Normality Sample independence Homogeneity of Variance	Continuous dependent variable Normality Sample independence Homogeneity of Variance Categorical independent variables	Continuous dependent variable Normality Sample independence Homogeneity of Variance Categorical independent variables

Predictive Analytics in One Picture

Technique	Purpose	Advantages / Disadvantages	What questions can be answered?
Decision Trees	Predict future classes of data.	**Advantages**: Easy to implement and understand. **Disadvantage**: Can be overly simplistic for many problems.	"Which one" (i.e. one discrete variable) or answers to Yes/No questions.
Neural Networks	Cluster and classify by recognizing patterns in data.	**Advantage**: Outperforms most ML algorithms. **Disadvantages**: Tough to implement; requires parallel processor.	Pretty much any question (as long as there's sufficient data and some kind of correlation/causation)
Regression (Linear and Logistic)	Estimates relationships between variables, uses those to predict future outcomes.	**Advantage**: Results are easy to understand. **Disadvantage**: Limited to linear/logistic data.	How much or how many of a certain target variable?
Time Series	Forecasts continuous variables over time.	**Advantage**: Easy to understand results. **Disadvantage**: Limited to time-dependent data.	What is the result going to be at a future data?
Clustering (K-means)	Organization of data into similar groups.	**Advantage**: Easy to implement. **Disadvantage**: Can be hard to predict "k", the number of clusters.	What kinds of patterns are in the data?
Factor Analysis	Find explanations for outcomes / correlations.	**Advantage**: Reduction of data to concise "picture". **Disadvantage**: Factors can be hard to interpret; Information can be lost.	What are the explanations for the themes in the data?

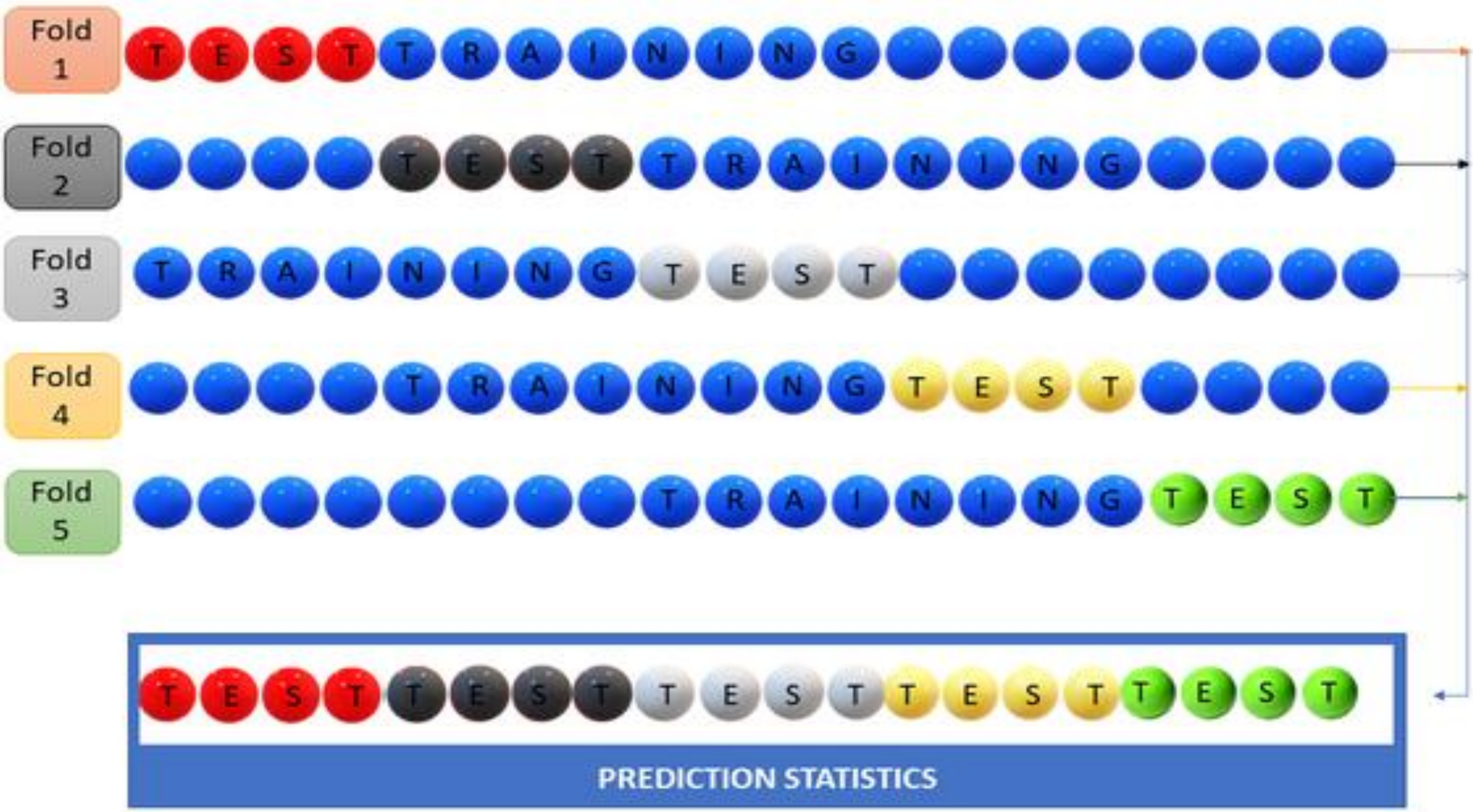
Time Series Method in One Picture

List of Time Series Method

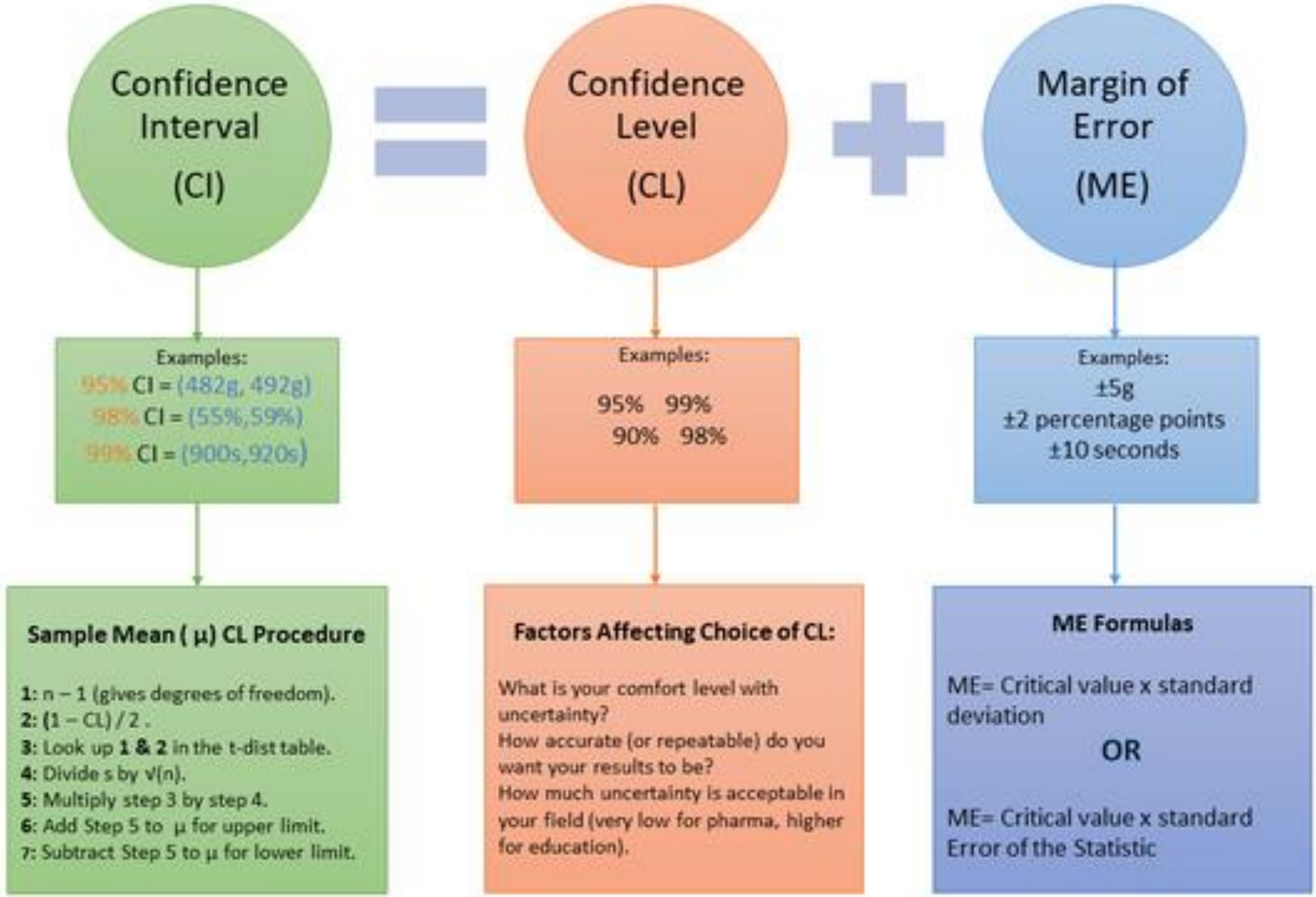
Datanest

1. Autoregression (AR)
2. Moving Average (MA)
3. Autoregressive Moving Average (ARMA)
4. Autoregressive Integrated Moving Average (ARIMA)
5. Seasonal Autoregressive Integrated Moving-Average (SARIMA)
6. Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX)
7. Vector Autoregression (VAR)
8. Vector Autoregression Moving-Average (VARMA)
9. Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX)
10. Simple Exponential Smoothing (SES)
11. Holt Winter's Exponential Smoothing (HWES)
12. Prophet
13. Naive method
14. LSTM (Long Short Term Memory)
15. STAR (Space Time Autoregressive)
16. GSTAR (Generalized Space Time Autoregressive)
17. LSTAR (Logistic Smooth Transition Autoregressive)
18. Transfer Function
19. Intervention Method
20. Recurrent Neural Network
21. Fuzzy Neural Network

Cross Validation in One Picture



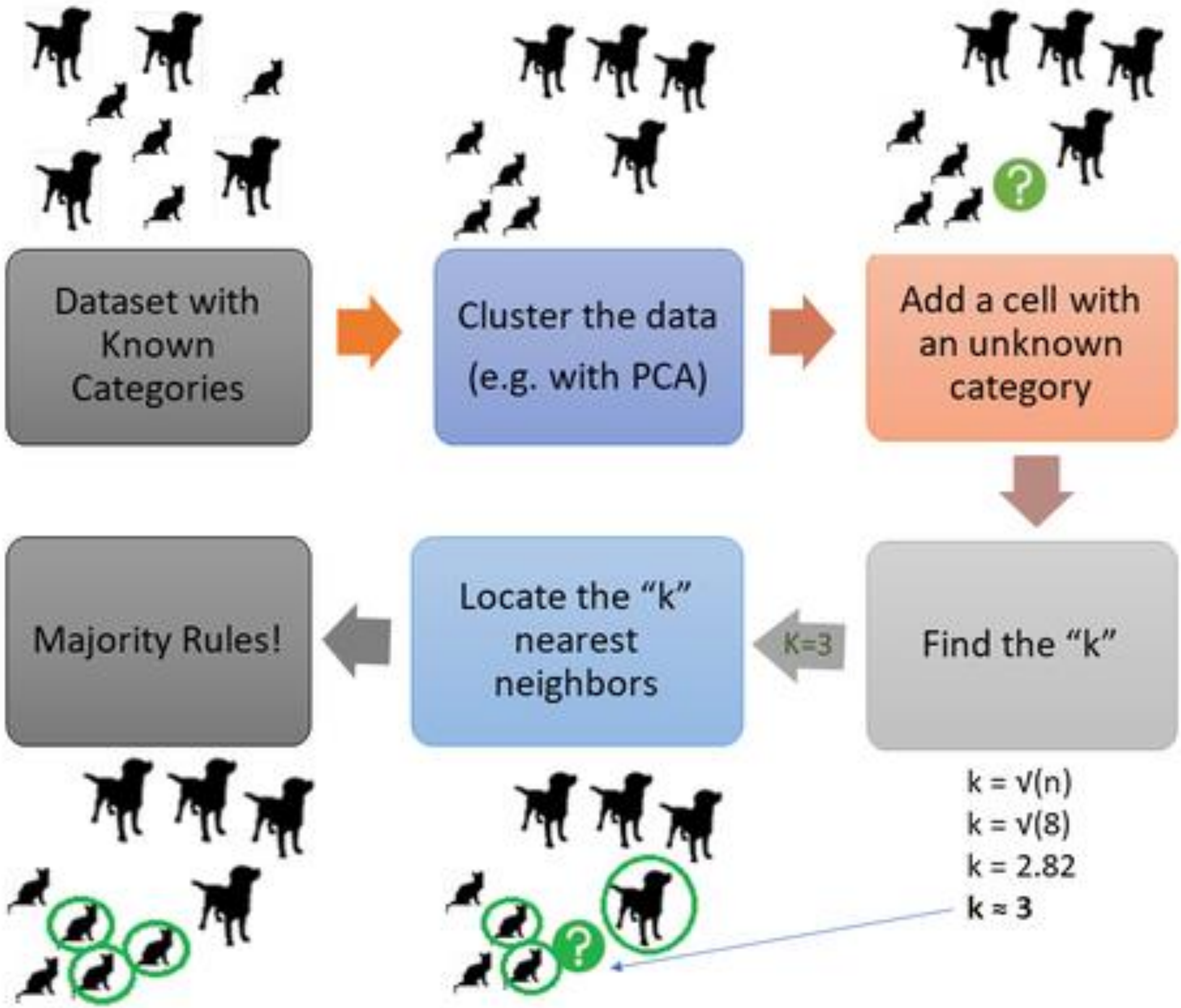
Confidence Interval in One Picture



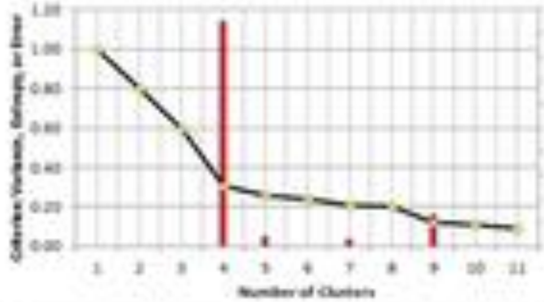
Unsupervised Learning in One Picture

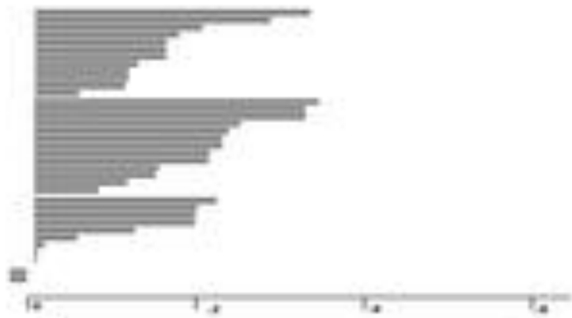
	K Means	Hierarchical	PCA
			
Type	Clustering	Clustering	Dimension Reduction
Overview of process	Nonhierarchical method that finds mutually exclusive spherical clusters based on distance.	Repeatedly links pairs of clusters until every data object is included.	Transforms high dimensional data into low dimensional data using orthogonal transformations.
Disadvantages	• Requires known "k" clusters; Can be difficult to choose. • Initial cluster choices and order of data strongly affect results. • Can be difficult to reproduce results due to random initial "centroid" choice.	• Initial seeds, data order have strong effect. • Merges cannot be undone. • No statistical / theoretical foundation for results. • Sensitive to outliers.	• Principal components (a linear combination of the original features) can be challenging to interpret and read, compared to the original features. • Data must be standardized beforehand.
Advantages	• Easy to implement. • Fast (for small k). • Clusters can be recalibrated.	• Easy to implement. • Dendogram makes for easy visualization of "k". • Results easily reproducible.	• Good visualization tool. • Reduces irrelevant or redundant features. • Reduces overfitting (by reducing features).
Works well for	Big data; Hyper spherical (e.g. 3D sphere).	Small to medium data. Performance and execution time increase dramatically for large data sets.	Extracting important features (components) from big data.

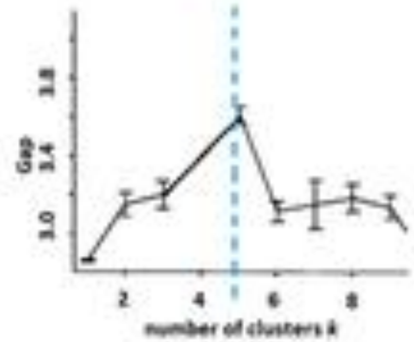
KNN in One Picture



Number of Clusters Selection Method in One Picture

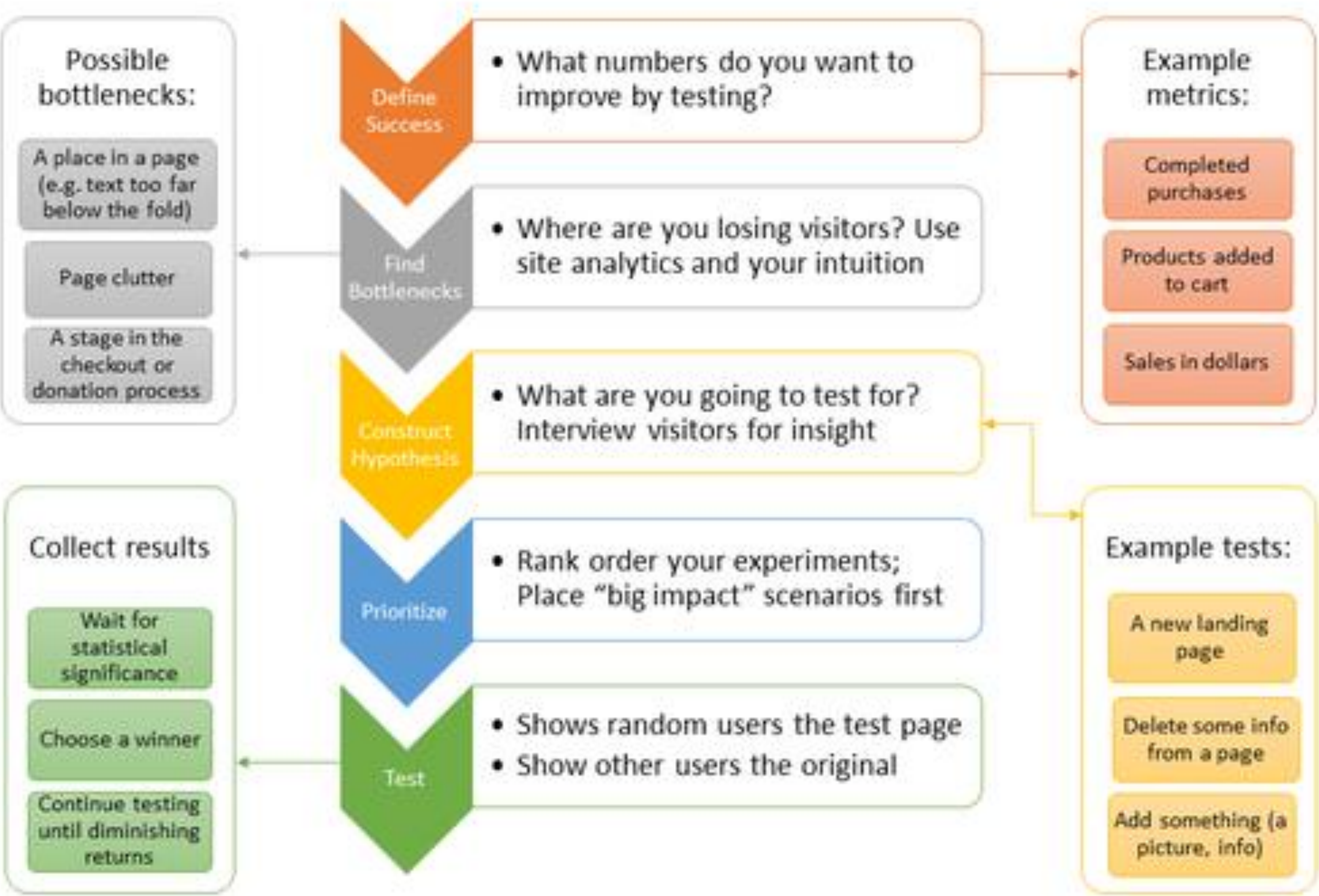




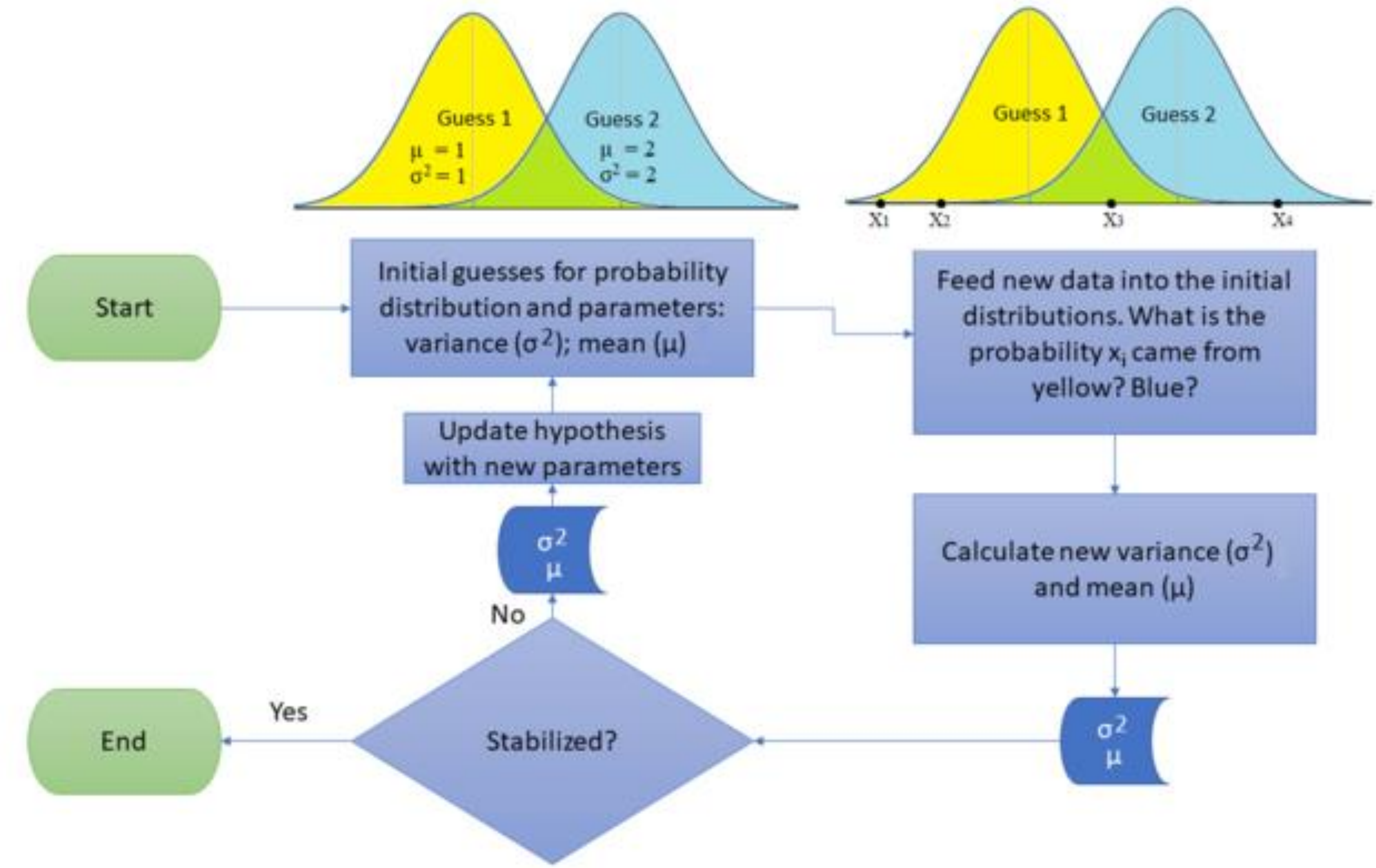


Technique	Elbow Rule	Silhouette Rule	Gap Statistic
How it works	Finds optimal (minimized) within cluster sum of squares (SSE) as a function of number of clusters	Measures how well objects are contained within clusters. Ranges from −1 to +1. High values mean well-matched to own cluster vs. other clusters.	Compares different "k" values with expected values from a dataset with no obvious clustering. The larger the "gap", the better.
Advantages	Ease of use and verification: Clear visual shows an "elbow" (red vertical line in the above image), beyond which more clusters add very little value	Easy to use. Popularity, which means there is plenty of available documentation.	Ease of use: optimal **k** value clear to see (a dashed blue line in the above image) May outperform other methods, according to authors.
Disadvantages	Sometimes ambiguous	More challenging to use and interpret; visual not too intuitive	Functions are poorly documented in R (Steorts, 2017)

AB Testing in One Picture



EM Algorithm in One Picture



ALPHAGO ZERO CHEAT SHEET

The training pipeline for AlphaGo Zero consists of three stages, executed in parallel

SELF PLAY

Create a 'training set'

The best current player plays 25,000 games against itself

See MCTS section to understand how AlphaGo Zero selects each move

At each move, the following information is stored

RETRAIN NETWORK

Optimise the network weights

A TRAINING LOOP

Sample a mini-batch of 2048 positions from the last 500,000 games

Retrain the current neural network on these positions

– The game states are the input (see 'Deep Neural Network Architecture')

Loss function

Compares predictions from the neural network with the search probabilities and actual winner

PREDICTIONS				ACTUAL
	p	Cross-entropy +	π	
	v	Mean squared error + Regularisation	(winner)	

After every 1,000 training loops, evaluate the network

EVALUATE NETWORK

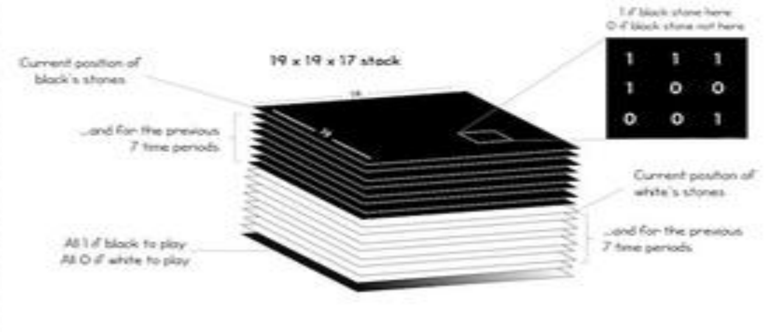
Test to see if the new network is stronger

Play 400 games between the latest neural network and the current best neural network

Both players use MCTS to select their moves, with their respective neural networks to evaluate leaf nodes

Latest player must win 55% of games to be declared the new best player

WHAT IS A 'GAME STATE'



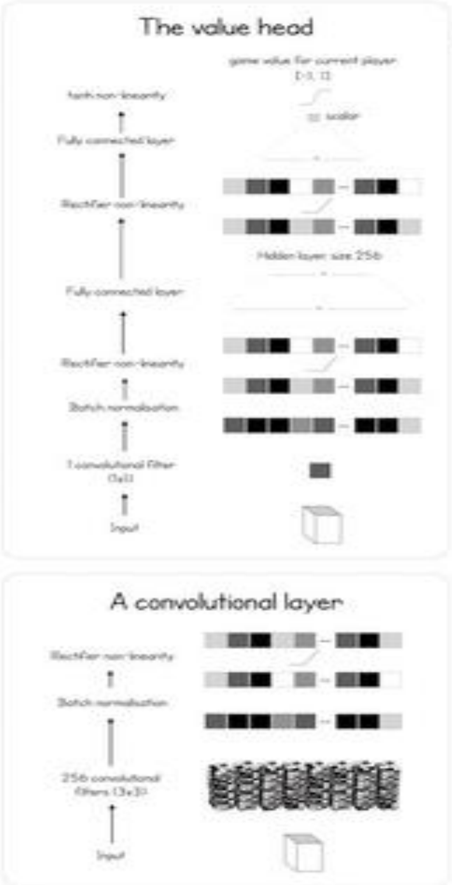
This stack is the input to the deep neural network

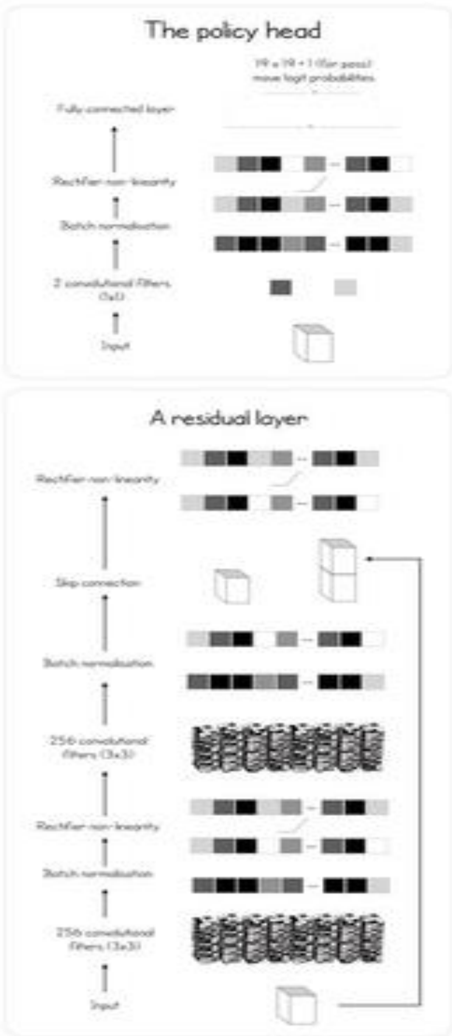
THE DEEP NEURAL NETWORK ARCHITECTURE

How AlphaGo Zero assesses new positions

The network learns 'tabula rasa' (from a blank slate)

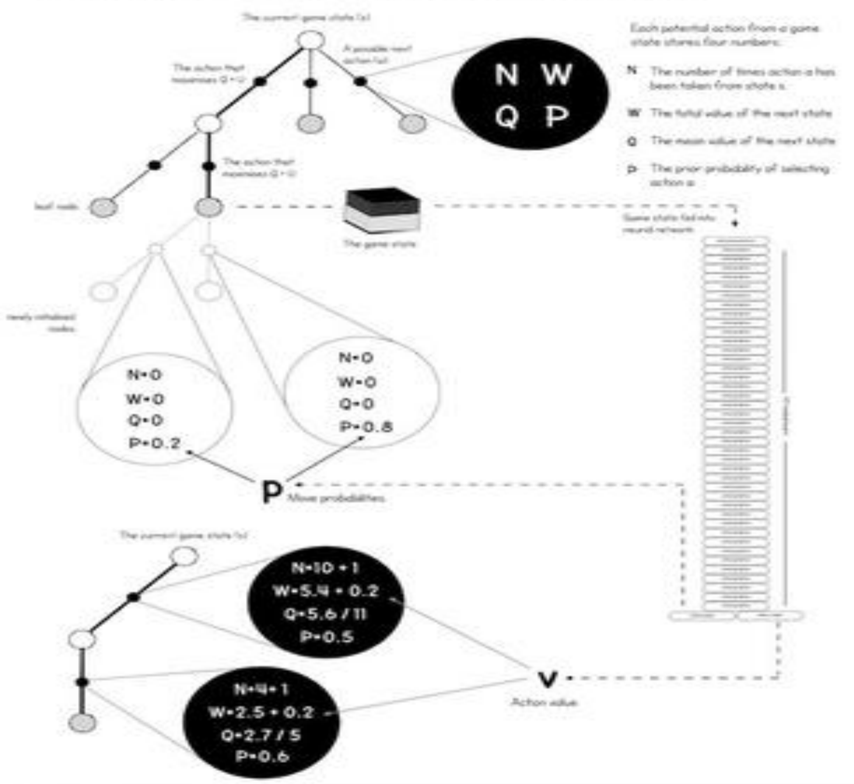
At no point is the network trained using human knowledge or expert moves





MONTE CARLO TREE SEARCH (MCTS)

How AlphaGo Zero chooses its next move



First, run the following simulation 1,600 times...

Start at the root node of the tree (the current game state)

1. Choose the action that maximises...

$$Q + U$$

Q: The mean value of the next state

U: A function of P and N that increases if an action hasn't been explored much, relative to the other actions, or if the prior probability of the action is high

Early on in the simulation, U dominates (more exploration), but later, Q is more important (less exploration)

2. Continue until a leaf node is reached

The game state of the leaf node is passed into the neural network, which outputs predictions about two things:

p Move probabilities

v Value of the state (for the current player)

The move probabilities p are attached to the new feasible actions from the leaf node

3. Backup previous edges

Each edge that was traversed to get to the leaf node is updated as follows:

$$N \rightarrow N + 1$$
$$W \rightarrow W + v$$
$$Q = W / N$$

...then select a move

After 1,600 simulations, the move can either be chosen:

Deterministically (for competitive play)

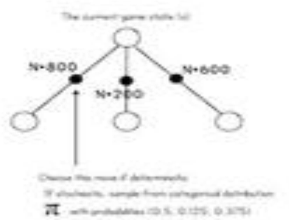
Choose the action from the current state with greatest **N**

Stochastically (for exploratory play)

Choose the action from the current state from the distribution

$$\pi \sim N^{1/\tau}$$

where τ is a temperature parameter controlling exploration



Other points

- The sub-tree from the chosen move is retained for calculating subsequent moves
- The rest of the tree is discarded

DEMYSTIFYING ARTIFICIAL INTELLIGENCE

by Swami Chandrasekaran (@swamichandra)

learns & fits to

Building Blocks

These are the core technologies and essential building blocks that are used in the design • build of AI systems. One or many of these are typically combined together to help realize AI systems.

- Natural Language Processing
  - NL Understanding
  - NL Generation
- Computer Vision
  - Classification
  - Handwriting Detect
  - Text Recognition
  - Instance Segment
  - Facial Detection
  - Object Detection
  - OCR Extraction
  - Motion Estimation
- Machine Translation
- Pattern Recognition
- Knowledge Engineering
- Robotics
- Machine Learning
  - Deep Learning
- Speech Processing
  - Speech Recognition
  - Speaker Verification
  - Speaker Identification
  - Speech Synthesis
- Planning, Scheduling & Optimization

Supporting Cast: Scale Out Infrastructure (Compute, Storage, Platform) · GPU · Ontologies · Knowledge Graph · Domain Experts · Data (doh!) · Ground Truth Management · Data Acquisition & Integration · Scalable Ingestion & Enrichment Pipelines · OCR Conversion · Analytics Engine · Metadata Catalog · Business Rules & Process

AI Systems work on various types of data

Structured Data: Datawarehouse · Credit Card Purchases · Vehicle Data · Billing Records · Order / Purchase History · Machine Generated · CRM Records · Census Data

Medical Reports · Survey Reponses · Trouble Tickets · Many more to come ....

Unstructured Data: Manuals, News & Research Reports · OCR, X-Rays, MRI, CAT Scans · SEC Filings, Loan, Regulatory Documents · Social Media Images · Product Reviews & Images · Call Center Recordings · Multimedia Content · Resume, Emails, Tweets, Blogs

Primary Characteristics of an AI System

AI Systems must exhibit the following characteristics. Call them the DNA, these characteristics are aligned with the key goals of an AI system that include reasoning, knowledge, planning, learning, natural language processing, perception and the ability to move and manipulate objects.

- Identifies new **patterns & relationships** from the data
- Can be **trained** by experts & also **learn** from experience
- Can capture & **represent** the learned **knowledge**
- **Continuously** learns & **reinforces** its behavior based on feedback
- Ingest **large** volume of diverse data in **variety** of formats
- **Converse** in natural language and detect **emotions**
- **Understand** natural forms of **expression** incl language, speech, images etc.
- **Hypothesize** & ranks alternatives and optimize
- **Reason › Infer ›** Generate conclusions from learned knowledge
- Have **Perception** abilities like humans to interpret data
- Responds with a degree of **confidence**
- **Plan** a set of actions to reach a desired **goal** & solve problems

Theoretically, different types of AI

Artificial Narrow Intelligence Specialist in one task. Pretty much all of what we know as AI is ANI.

Artificial General Intelligence Specialist in a wide range of cognitive functions. Continuously reason & improve automatically.

Artificial Super Intelligence Outperform humans. Very serious ethical considerations a must. Well, this may not happen also !!!

most of the current AI is ANI

Eliminate Unconscious Bias · Augment Human Intelligence · Explainable & Evidence Driven

"ARTIFICIAL INTELLIGENCE IS THE SCIENCE AND ENGINEERING OF MAKING INTELLIGENT MACHINES" - JOHN MCCARTHY

Used Responsibly & Ethically · Transparent & Accountable · Respecting Privacy of the individual

Author: Swami Chandrasekaran

depending on the nature of data, choose a learning algorithm · realized using · implemented using

Types of Learning Approaches

3 major approaches through which machines learn without being explicitly programmed

Supervised Learning *Learning from set of labeled data provided by an expert. Also classified into 2 subgroups: regression & classification.*

Raw Data → Features & Labeled Input Data → Train → Learned Model / Algorithm → Deploy, Monitor, Fix and Fine Tune

Unsupervised Learning *Uncovering inherent structures, patterns and relationships hidden in collections of unlabeled data*

Raw Data → Algorithm → Clusters, Anomalies → Review & Use

Reinforcement Learning *Learning what to do in an environment and how to map situations to actions to maximize a reward signal.*

Goals → Initialize Q & Select Actions → Agent → Measure, Reward & Update → Environment → Perform Action

Top Algorithms

- Linear Regression
- Logistic Regression
- Naive Bayes
- Hidden Markov
- XG Boost
- SVM

Deep Learning Algorithms
- Convolutional Neural Nets (CNN)
- Deep Belief Networks (DBN)
- Auto Encoders
- LSTM
- RBM

- K-Means Clustering
- Principal Component Analysis
- Anomaly Detection
- Singular Value Decomp (SVD)

- Markov Decision Process
- Q-Learning
- Temporal Difference Learning

Most common AI Workloads / Tasks

These are some of the most common tasks that AI systems are taught to perform and also learn continuously over time. Each of these tasks MUST provide responses with a degree of confidence or probability associated with them along with underlying evidence.

Named Entity Recognition Ability to identify and extract semantic concepts such as people, clinical drugs, clauses, stock prices, organizations, and many other types of entities from text.

Passage Retrieval Find relevant pieces of information / passages within large & varied corpus based on deep understanding of the query. Intelligent scoring algorithms to rank passages.

Object Class & Detection Given a raw image, identify which class does the image belong to. In object detection, you task is to identify where in the image does the objects lies.

Classification Coming together of machine learning & NLP to assign one or more classes to a span of text, document, including a confidence score associated with the results.

Emotion Detection Much deeper than bipolar sentiment, where in you detect the emotion (anger, joy, irritation etc) of the user based on the text he/she have authored.

NLP Pre-processing, taxonomy, tokenization, POS tagging, syntactic parsing, semantic role labeling, word sense disambiguation, meta concepts, relationships, anaphora resolution.

Language Translation Using neural net techniques like NMT to auto-detect & translate between languages. Also trainable using forced glossary, parallel phrases, corpus-level customization etc.

Speech Transcription Automatically transcribe audio to text including identifying speakers as well as handling multiple languages, understand abbreviations etc with a low error rate. Human error rates are around 5.1%

Knowledge Extraction Being able to learn, represent & extract knowledge from large bodies of unstructured data. Includes knowledge graph construction and learning knowledge base embeddings.

Common Examples of AI Systems in Action

Review Contract Docs	Audio Call Transcription	Conversational Bots	Search	Customer 360	Threat Prevention
Product Recommendations	Sentiment Analysis	Precision Marketing	Equipment Maintenance	Abnormalities from MRI, x-ray etc	Visual Search
Fraud Detection	Customer Churn Prediction	Enforce Compliance	Malware & Spam Detection	Insurance Claims Damage Assessment	Case Law
Social Listening	Employee Attrition Prediction	Lead Generation	Identify Skin Cancer from Lesions	Multi-modal, no touch Commerce	Amazon Alexa
Taxi Dispatch	Home Automation Control	Sort Vegetables	Patient Readmission Prediction	AI enabled RPA	Drug Repurposing

built using

DevOps for AI - How are these systems built?

Here is a view of how to plan, build, test and operationalize AI using an agile approach.

- ✓ CLEARLY DEFINE YOUR PROBLEM STATEMENTS
- ✓ LIST YOUR OVERALL BUSINESS GOALS & OBJECTIVES?
- ✓ KEY QUESTIONS YOU WISH TO SEEK ANSWERS TO
- ✓ EDUCATE & MANAGE STAKEHOLDERS
- ✓ PLAN FOR ORGANIZATIONAL SKILLS
- ✓ DEFINE A CHANGE MANAGEMENT STRATEGY
- ✓ DO NOT UNDERESTIMATE NEED FOR DOMAIN EXPERTS

Assess
- Define Business & Transformation Objectives
- Identify Hills, Backlog & Representative Users
- Perform preliminary data analasis & needs
- Identify & line up SME's & Experts
- Conduct Cost-Benefit Analysis Workshop
- Define Iterative Success Criteria
- Rapidly Prototype

Design
- Use Cases & Actors
- System Context
- Overall Architecture
- Identify Data Sources
- Meta-model / Ontology Design
- Design Comprehensive data & model security & protection schemes

Data Understanding
- Data Source Analysis
- Verify Data Quality
- Design data acquisition strategy
- Design Data Ingestion & Integration Pipeline
- Identify special needs such as OCR
- Plan for out validation, cross-validation, and bootstrapping

Build
- Prepare Ground Truth
- Prepare Blind & Validation Set
- Perform Feature Engineering
- Select appropriate algorithm
- Choose model parameters
- Iteratively Train Models
- Instrument for Usage & Feedback
- Integrate with / build consuming apps

Test & Evaluate
- Model Evaluation & Validation
- Blind Tests. Test with representative users
- Analyze Performance Metrics
- Hyperparameter Tuning
- User Acceptance Test
- Evaluate against business criteria & objectives

Deploy
- Promote & Stage Trained & Tested Models
- Version control & deploy trained models to analytical engines
- Models available as RESTful services
- Roll out communication & training plans for the user personas
- Deploy applications, processes and other components that use the deployed models

Monitor
- Continuously monitor performance
- Setup agile & proactive support mechanisms
- Constantly monitor & measure against business objectives
- Watch out for usage issues
- Setup model Governance approach

Capture &Use Feedback
- Continuously monitor & evaluate feedback
- SME & Experts committee evaluate the feedback
- Collect / define new features
- Governance around what becomes new ground truth

Adjust & Iterate

learned models evaluating using

Measuring AI Systems

The most common measures and other approaches to evaluating the performance of AI systems. Note to self: "Accuracy alone is a very misleading metric"

Precision Proportion of information found that is correct or relevant. A measure of accuracy often on the positive labels of interest. When testing "phone issue" classification model on 1000 examples, 700 predicted as "battery drain". But only 500 of the were correct predictions. So your precision is 500/700 = 0.71. *Low Precision ⇒ Lots of False Positives*

Accuracy Number of correct predictions made divided by the total number of predictions made over all examples you are test the model on. So you tested your model on 1000 examples. The model correctly predicted the label on 880 examples. Your accuracy is 880/1000 = 0.88.

Recall Proportion of how many examples of interest the model has identified. Recall is a measure of how many true positives were found. In the "phone issues" model, say there are 100 SIM card examples. If your model predicted 90 out of those 100, then recall is 90/100 = 0.9. **Low Recall ⇒ Lots of False Negatives**

F1 Score F1 score conveys the balance between the precision and the recall. F1 score reaches its best value at 1 and worst score at 0. **Low F1 ⇒ Low P & R**

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Confusion Matrix Very commonly used to describe the performance of a classification model on a set of test data for which true values are known. Say you are evaluating an "Apple" detection image recog model ...

Actual \ Predicted	Apple	Fish
Apple	True Positive	False Negative (Type II Error)
Fish	False Positive (Type I Error)	True Negative

Word Error Rate Metric for measuring speech recognition & machine translation systems. Derived from the Levenshtein distance, between word sequence H (hypothesis) and word sequence R (reference / ground truth).

Other Evaluation Criteria / Approaches: Bias - Variance Tradeoff · Homogeneity Score · Akaike information criterion(AIC) · Completeness Score · Passage relevancy rating · Jaccard Similarity Score · Average Steps per Conversation · Inter Annotator Agreements · Problem Benchmarks · Conversation abandonment rates · Hamming Loss · Peer Confrontation · Adjusted Rand Index (ARI) · In-Sample Error · Log-Loss · Lift

Receiver operating characteristics (ROC): Technique for visualizing, organizing and selecting classifiers based on their performance. ROC curves show the tradeoff between false positive and true positive rates. We plot **'True Negative Rate'** on X-axis and **'True Positive Rate'** on Y-axis.

Area Under the Curve (AUC): Used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

Turing Test: Suppose that we have a person, a machine, and an interrogator. The interrogator is in a room separated from the other person and the machine. The object of the game is for the interrogator to determine which of the other two is the person, and which is the machine.

Chinese Room: Does the machine *literally* "understand" Chinese? Or is it merely simulating the ability to understand Chinese? Programming a digital computer may make it appear to understand the language but does not produce real understanding.

API's, Libraries & Frameworks

Microsoft CNTK · IBM Watson · Tensorflow · Keras · System ML · PyBrain · NuPic · Chainer · DeepLearning4J · H2O.ai · Oryx 2 · NLTK · MLib · Torch · IBM DSX · Scikit-Learn · PyLearn2 · Mahout

Pragmatically, what's next for AI?

Capsule Networks · Hyper Imaging · Conscious Machines · Hyper Imaging · Contextual Learning · Sparse Data Learning · Ground Truth Methods · Quantum Enabled AI · Custom AI Chips · Resistive Processing Units · Adversarial networks · Generative Models · Deep Q Learning · Auto Feature Elicitation · 3-D Object Detection · Debating Systems

Take time to understand these Key Topics & Concepts

GROUND TRUTH · FEATURE ENGINEERING · CLASSIFICATION & REGRESSION · CLUSTERING · KNOWLEDGE GRAPHS · TRANSFER LEARNING · BIAS & OVERFITTING · GRADIENT DESCENT · AGENTS · INFERENCE · BACK PROPAGATION · ENTITY RESOLUTION · COST FUNCTION · AUGMENTED INTELLIGENCE · REASONING · TENSORS · MODEL EVALUATION CRITERIA · CNN & RNN · NLP BASICS · NEURAL MACHINE TRANSLATION · PRECISION & RECALL · ANNOTATIONS · INGESTION & ENRICHMENT · MODEL GENERALIZATION · SIGMOID FUNCTION

Number Representation Systems in One Picture

	Logistic Map ($p = 1$)	Logistic Map ($p = 0.5$)	Nested Square Root	Base b ($b > 1$, integer)	Base b ($b > 1$, not an integer)	Continued Fraction
Support domain for $x = x(1)$, $x(n)$ and $g(x)$	[0, 1]	[0, 1]	[1, 2]	[0, 1]	[0, 1]	[0, 1]
Support domain for digits $a(n)$	{0, 1}	{0, 1}	{0, 1, 2}	$\{0, 1, \cdots, b-1\}$	$\{0, 1, \cdots, \lfloor b \rfloor\}$	1, 2, 3, ...
$g(x)$ [$x(n+1) = g(x(n))$]	$4x(1-x)$	$\sqrt{4x(1-x)}$	$x^2 - \lfloor x^2 \rfloor + 1$	$bx - \lfloor bx \rfloor$	$bx - \lfloor bx \rfloor$	$1/x - \lfloor 1/x \rfloor$
$h(x)$ [$a(n) = h(x(n))$]	$\lfloor 2x \rfloor$	$\lfloor 2x \rfloor$	$\lfloor x^2 \rfloor - 1$	$\lfloor bx \rfloor$	$\lfloor bx \rfloor$	$\lfloor 1/x \rfloor$
$x = f(\{a(n)\})$	Unknown	Unknown	$\sqrt{a(1) + \sqrt{a(2) + \cdots}}$	$\sum_{k=1}^{\infty} \frac{a(k)}{b^k}$	$\sum_{k=1}^{\infty} \frac{a(k)}{b^k}$	$\frac{1}{a(1) + \frac{1}{a(2) + \cdots}}$
Digits distribution $P(a(n) = k)$	Uniform on {0, 1}	$\sqrt{2}/2$ if $k = 1$	$r(k+2) - r(k+1)$ $r(k) = \sqrt{5\sqrt{k} - 1}$	Uniform on {0, 1, ... , b - 1}	Non uniform	$\log_2\left(\frac{(k+1)^2}{k(k+2)}\right)$
Equilibrium distribution $P(x(n) < y)$	$\frac{1}{\pi}\int_0^y \frac{1}{\sqrt{x(1-x)}}dx$	$1 - \sqrt{1-y}$	$-2 + \sqrt{5y - 1}$	Uniform on [0, 1]	Non uniform	$\frac{\log(1+y)}{\log 2}$
Correlation between $x(n+1)$ and $x(n)$	0	-1/2	Non zero	1/b	Non zero	Non zero
Correlation between $a(n+1)$ and $a(n)$	0	-1/4	Non zero (but close)	0	Non zero	E[$a(n)$] is infinite

The equilibrium distribution is the one satisfying $P(X < y) = P(g(X) < y)$. The statistical properties listed here, are valid for almost all seeds x, except for a set of seeds of measure 0 (depending on the system), known as bad seeds. This document was produced by www.datasciencecentral.com.