

Beyond the Actuarial Tables: A Primer on Data Science and Statistics for Medical Directors

Dr. Prajakt Mahajan, PartnerRe

Data Scientist

(Noun)

Definition:

Someone who does precision
guesswork based on unreliable
data provided by those of
questionable knowledge

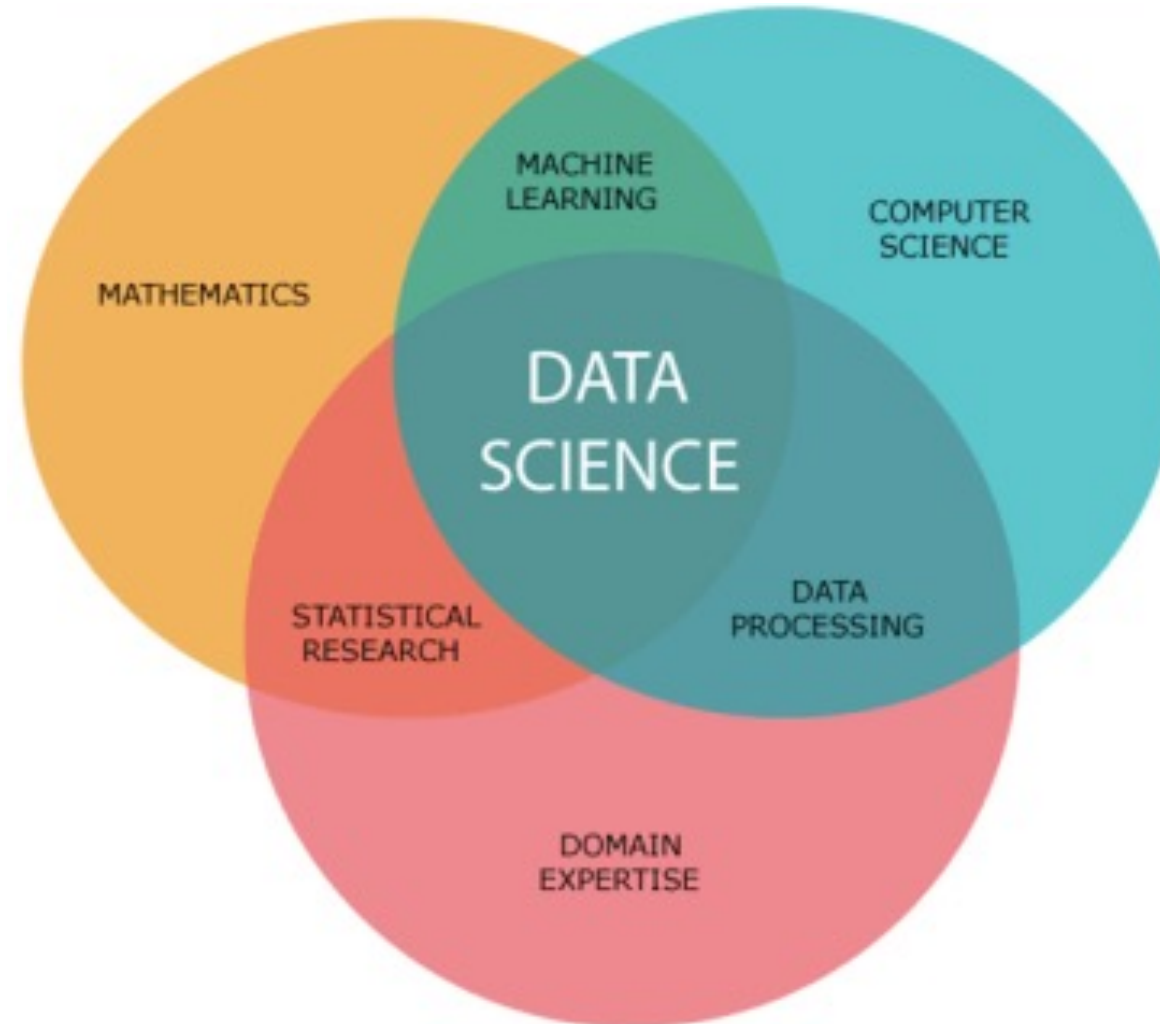
Also see: Wizard, Magician

Agenda

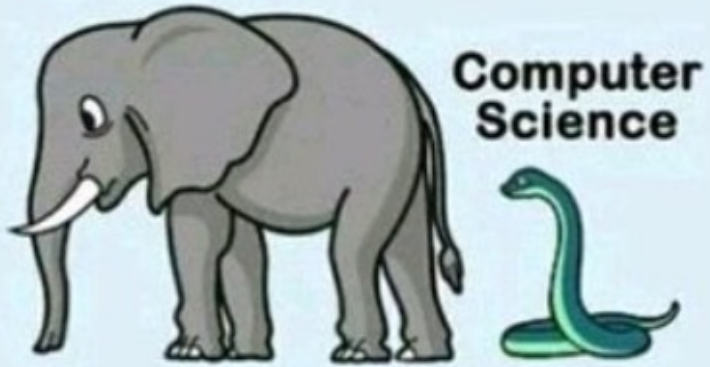
- ▶ Basics of Data Science
- ▶ Data Science Life Cycle
- ▶ Tools and Structure
- ▶ Key Statistical Concepts
- ▶ Types of Predictive Models
- ▶ Evaluating a Model
- ▶ Trends and Future Directions
- ▶ Role of Medical Director

Data Science

- ▶ Definition
- ▶ Importance

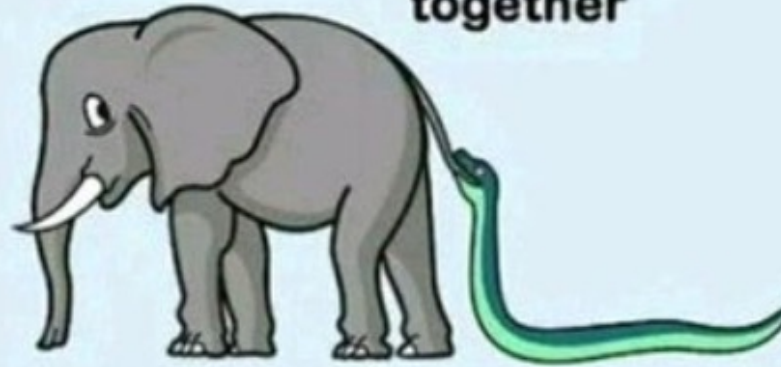


Statistics

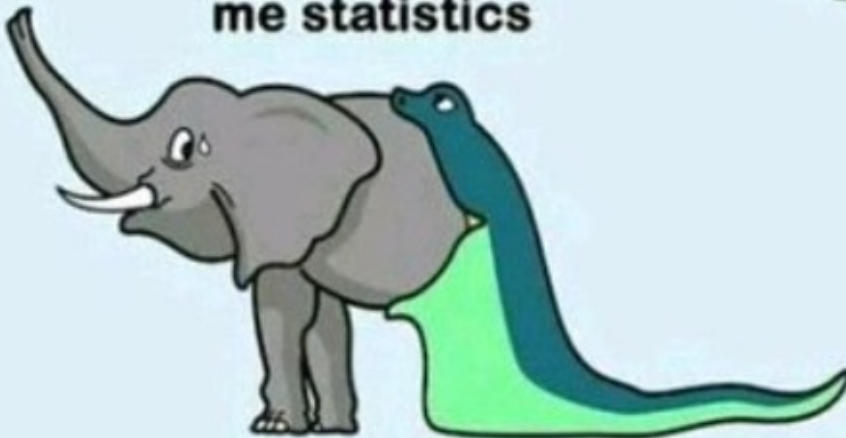


**Computer
Science**

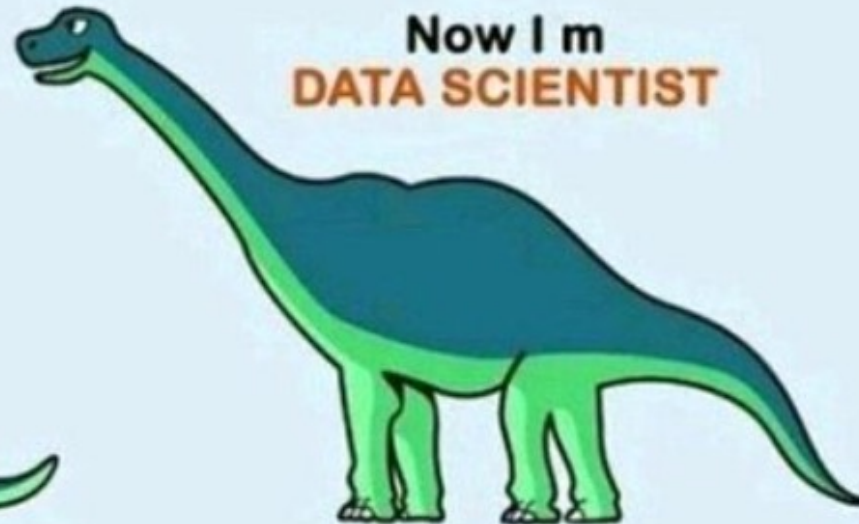
**We will work
together**



**Please teach
me statistics**



**Now I m
DATA SCIENTIST**



Lifecycle of Data Science

Problem Definition

Data Collection

Data Preprocessing

**Exploratory Data
Analysis**



Lifecycle of Data Science

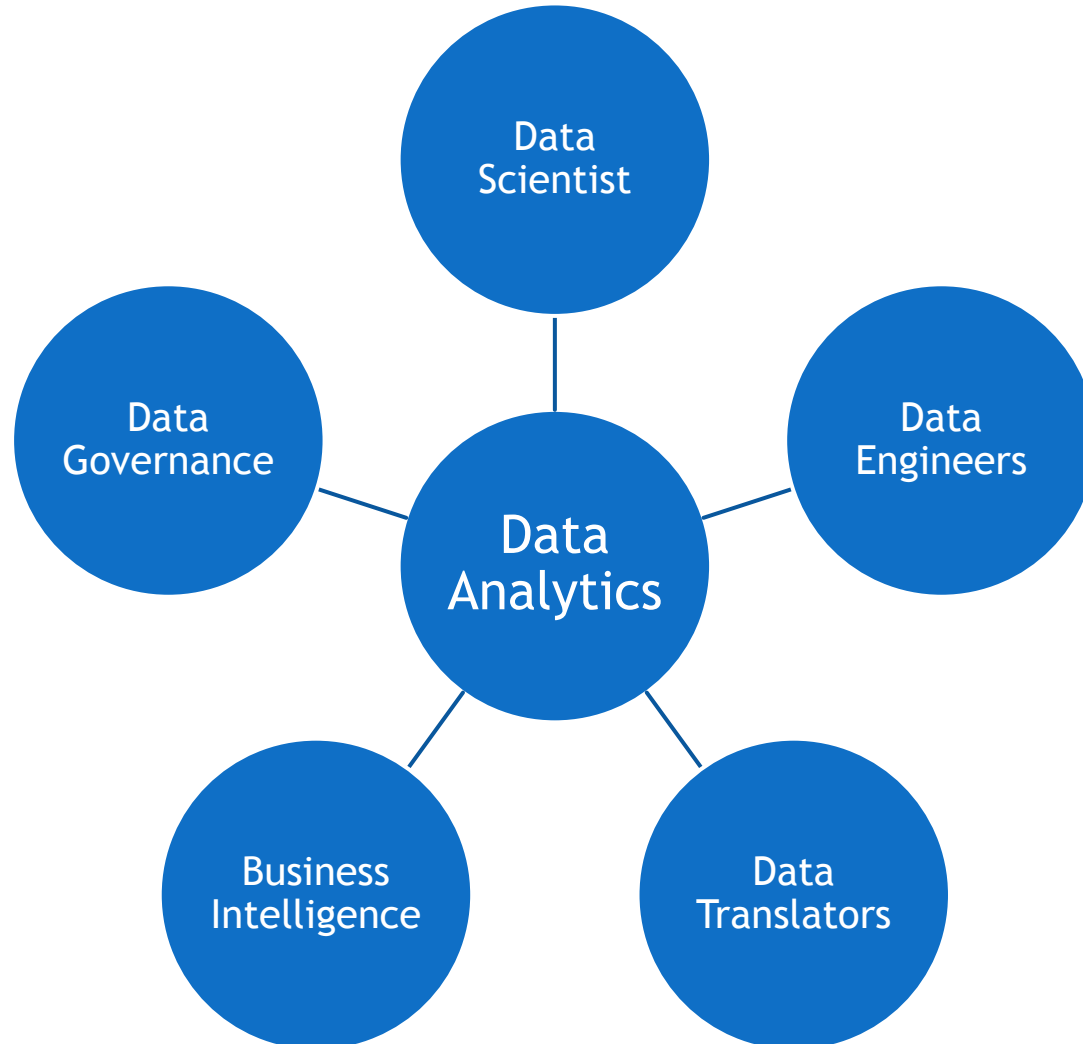


Tools

- ▶ Data Storage/Collection
 - ▶ Data Warehouse
 - ▶ Cloud Storage
- ▶ Data Extraction
 - ▶ SQL
- ▶ Data transformation and analysis
 - ▶ Microsoft Excel
 - ▶ R programming
 - ▶ Python
- ▶ Visualization
 - ▶ Tableau
 - ▶ Power BI
- ▶ Modeling
 - ▶ Python
 - ▶ R programming



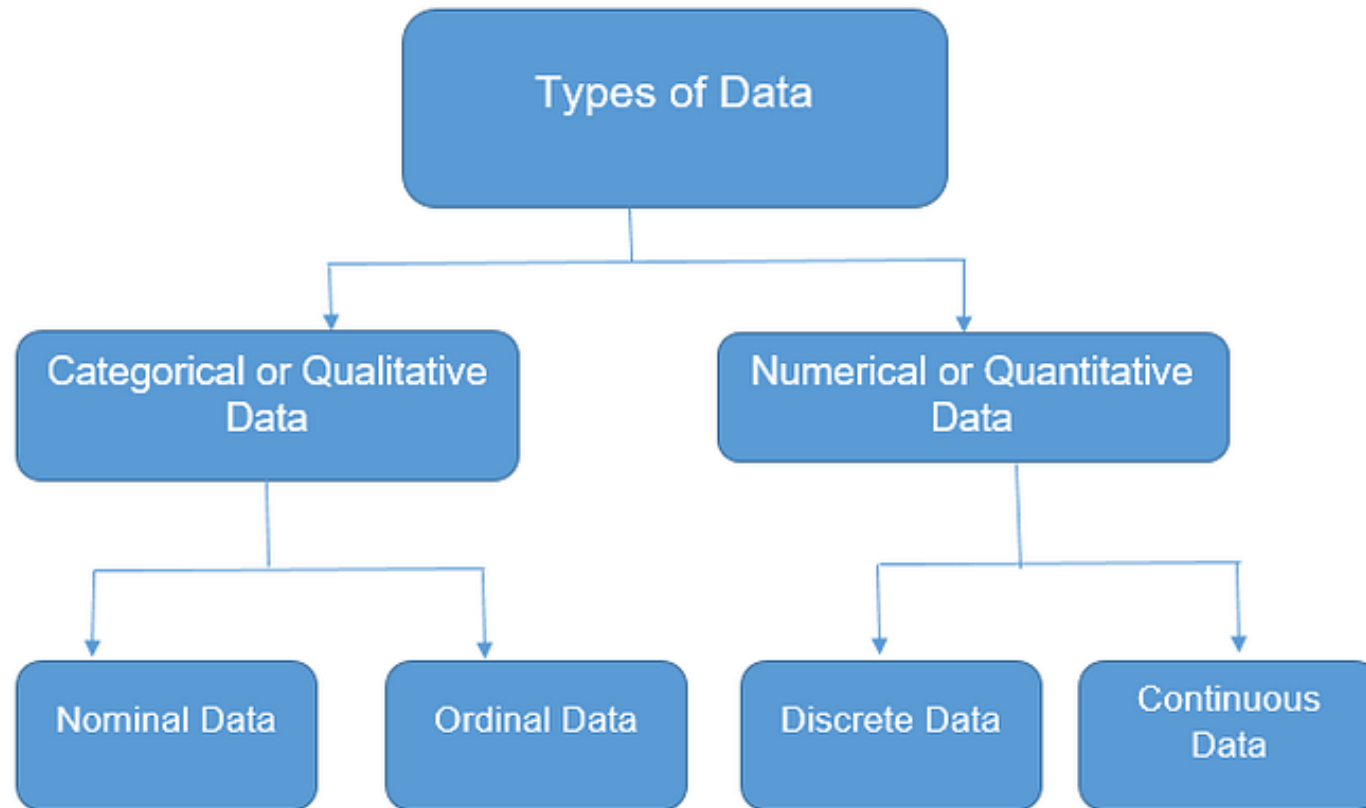
Data Analytics - Team Structure



Key Statistical Concepts

- ▶ Types of Data
- ▶ Descriptive and Inferential Statistics

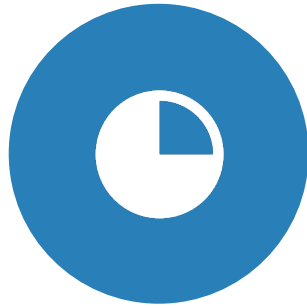
Types of Data



Role of Statistics in Data Science



Statistics helps collect, analyze and draw actionable insights from data



Descriptive statistics helps summarize and describe data



Inferential statistics helps make predictions and forecasts

Statistics forms the foundation of data science by providing the tools to work with data

Descriptive Statistics



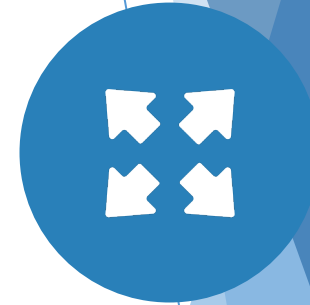
What is descriptive statistics?

Descriptive statistics summarize and organize characteristics of a data set.



Measures of central tendency

Mean, median, and mode describe the central position of a data set.



Measures of variability

Range, variance, and standard deviation describe how spread out a data set is.

Descriptive statistics help visualize and summarize data in a meaningful way.

Descriptive Statistics

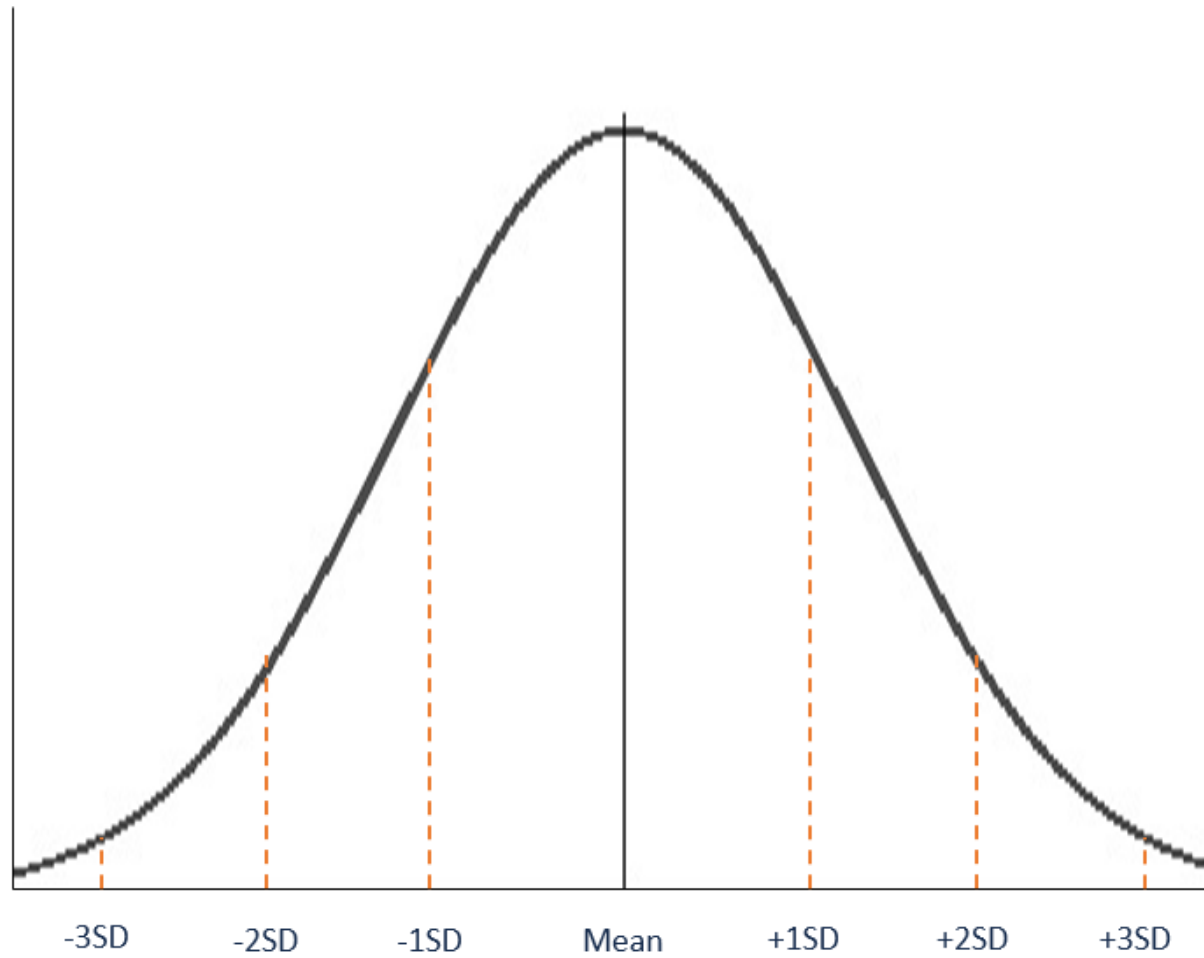
Measures of central tendency

- **Mean (Average):** The sum of all values divided by the number of values
- **Median:** The middle value in a sorted dataset
- **Mode:** The most frequently occurring value in a dataset

Measures of variability

- **Range:** The difference between the maximum and minimum values
- **Variance:** A measure of data dispersion
- **Standard Deviation:** A measure of the spread of data points around the mean

Normal Distribution



“A data scientist can have his head in an oven and his feet in ice, and he will say that, on average, he feels fine.”

Inferential Statistics



Inferential statistics are used to analyze a sample of data and make inferences about a larger population



Common examples of inferential statistics include hypothesis testing, confidence intervals, ANOVA, and regression analysis



Inferential statistics require making assumptions about the sample and population distribution

Inferential statistics are powerful techniques to make conclusions about populations from sample data but require thoughtful use and checking of assumptions.

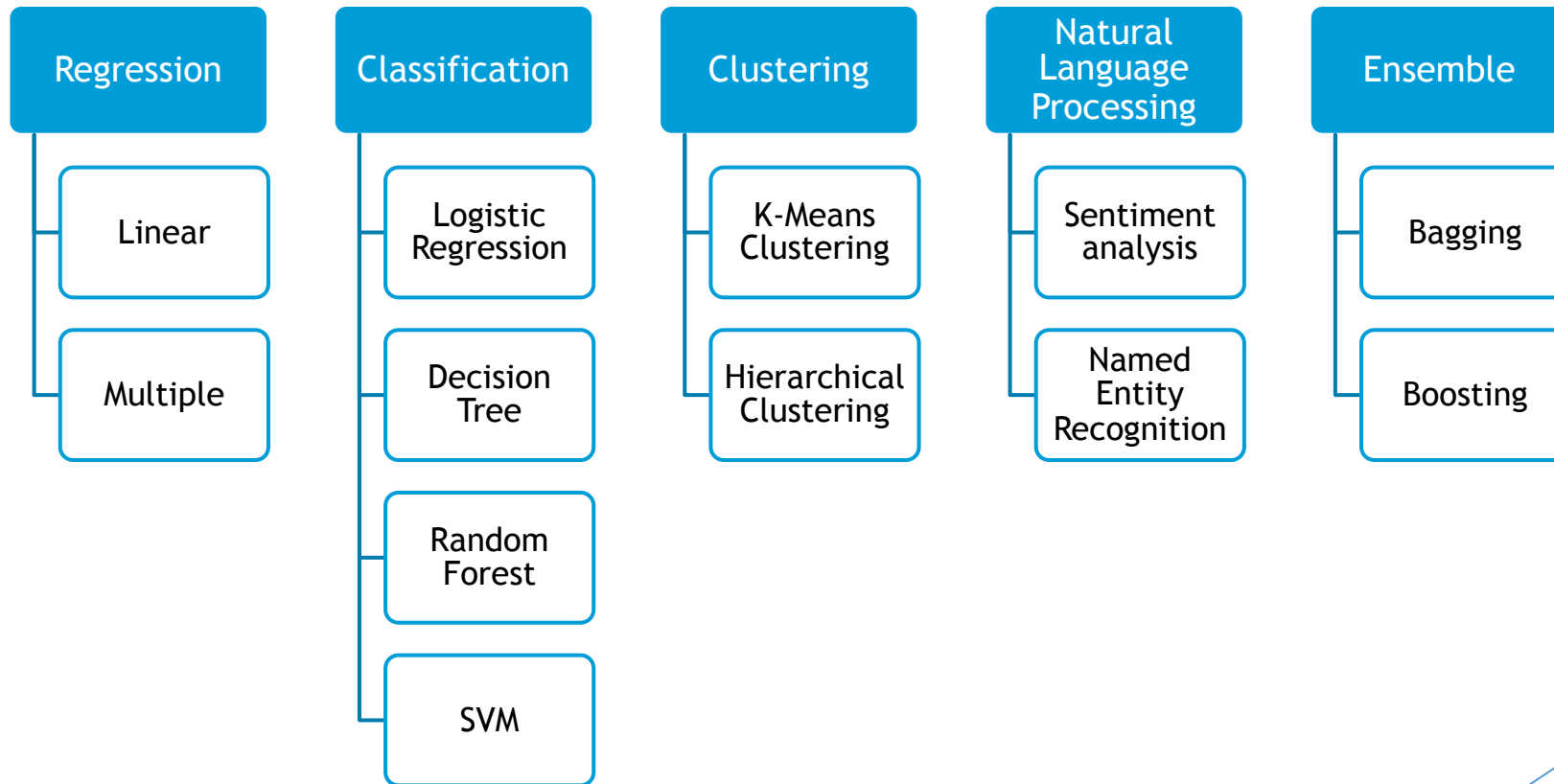
Inferential Statistics

- **Population**
- **Sample**
- **Hypothesis testing:** Checks if an observed sample effect reflects a genuine population effect or random chance.
- **Sample error:** the difference between a population parameter and the sample statistic, which is typically due to random variability
- **Confidence interval:** A range of values within which a population parameter is likely to fall, along with a level of confidence associated with that range
- **Null hypothesis:** A statement that there is no effect, no difference, or no relationship in the population
- **Alternate hypothesis:** A statement that contradicts the null hypothesis and suggests that there is an effect, difference, or relationship in the population
- **P value:** The probability of obtaining a result as extreme as, or more extreme than, the observed result, assuming the null hypothesis is true

Classifying Predictive Models

- ▶ Types of output
- ▶ Complexity
- ▶ Data Type
- ▶ Use Case
- ▶ Method of Ensemble
- ▶ Interpretability
- ▶ Data Availability
- ▶ Data Processing
- ▶ Data Distribution

Types of Predictive Models



Developing and Evaluating a Model

- ▶ Data Splitting
 - ▶ Training and Testing Data
- ▶ Evaluation Metrics
- ▶ Model Testing
- ▶ Check for overfitting
- ▶ Cross Validation
- ▶ Bias and Fairness
- ▶ Tuning and Iterating

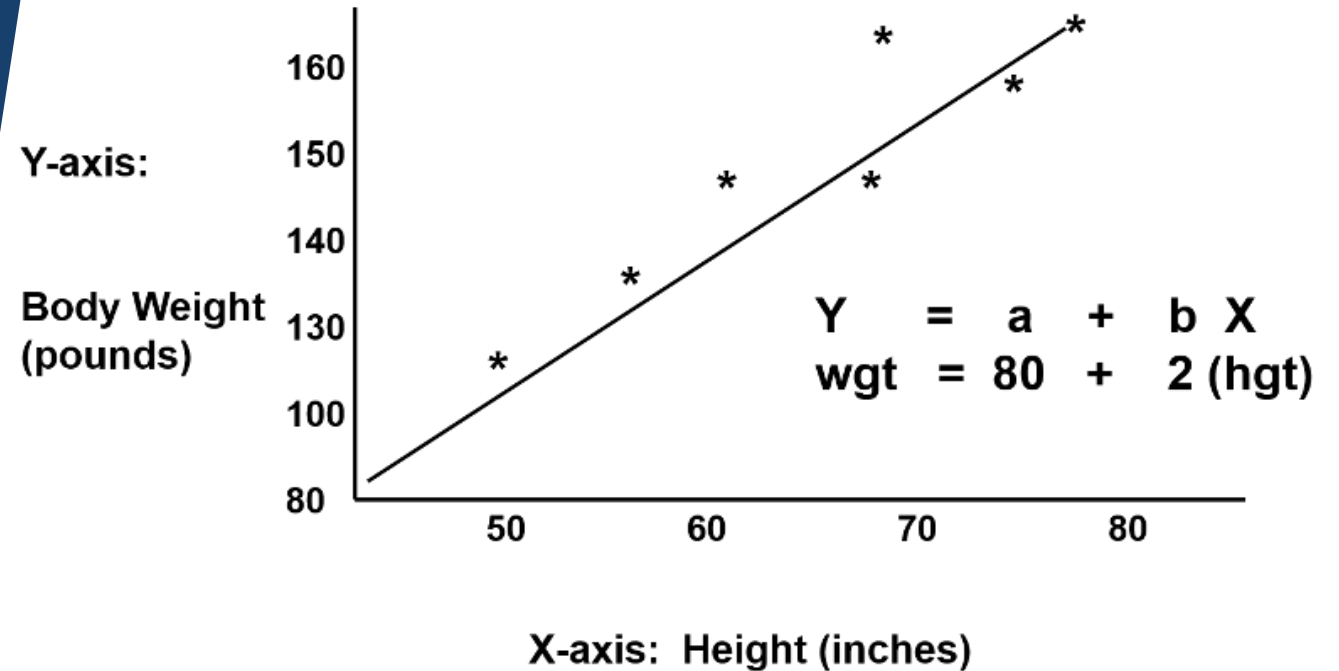
THE #1 MACHINE LEARNING EXCUSE
FOR LEGITIMATELY SLACKING OFF:

"MY MODEL IS TRAINING."



Evaluating a Model

- ▶ Regression Model
 - ▶ Mean Absolute Error (MAE)
 - ▶ Mean Squared Error (MSE)
 - ▶ Root Mean Squared Error (RMSE)
 - ▶ R-squared (R^2)
 - ▶ Coefficient of Determination (Adjusted R-squared)



Evaluating a Model

▶ Regression Model

- ▶ Mean Absolute Error (MAE)
 - ▶ average absolute differences between predicted and actual values
- ▶ Mean Squared Error (MSE)
 - ▶ average of the squared differences between predicted and actual values
- ▶ Root Mean Squared Error (RMSE)
 - ▶ square root of the MSE and provides a measure of error in the same units as the dependent variable
- ▶ R-squared (R^2)
 - ▶ quantifies the proportion of variance explained by the model
- ▶ Coefficient of Determination (Adjusted R-squared)
 - ▶ modification of R-squared that accounts for the number of predictor variables in the model

Evaluating a Model

- ▶ Classification Model
 - ▶ Percent correction classification (PCC)
 - ▶ Confusion Matrix
 - ▶ Area Under the ROC Curve (AUC - ROC)
 - ▶ Lift and Gain Charts

Confusion Matrix

- ▶ A table that summarizes the performance of a classification model on a set of test data. The matrix shows the number of correct and incorrect predictions made by the model, broken down by each class
- ▶ Is used in machine learning and statistics to assess the performance of a classification model

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Confusion Matrix

- ▶ **Accuracy:** The proportion of correct predictions out of the total predictions
- ▶ **Precision:** The ability of the model to correctly identify positive cases
- ▶ **Recall (Sensitivity or True Positive Rate):** The ability of the model to capture all the positive cases
- ▶ **Specificity (True Negative Rate):** The ability of the model to correctly identify negative cases
- ▶ **F1 Score:** The harmonic mean of precision and recall, which balances the trade-off between the two metrics. It is calculated as $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$
- ▶ **False Positive Rate (FPR):** The proportion of actual negative cases that were incorrectly classified as positive, calculated as $FP / (FP + TN)$

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Confusion Matrix

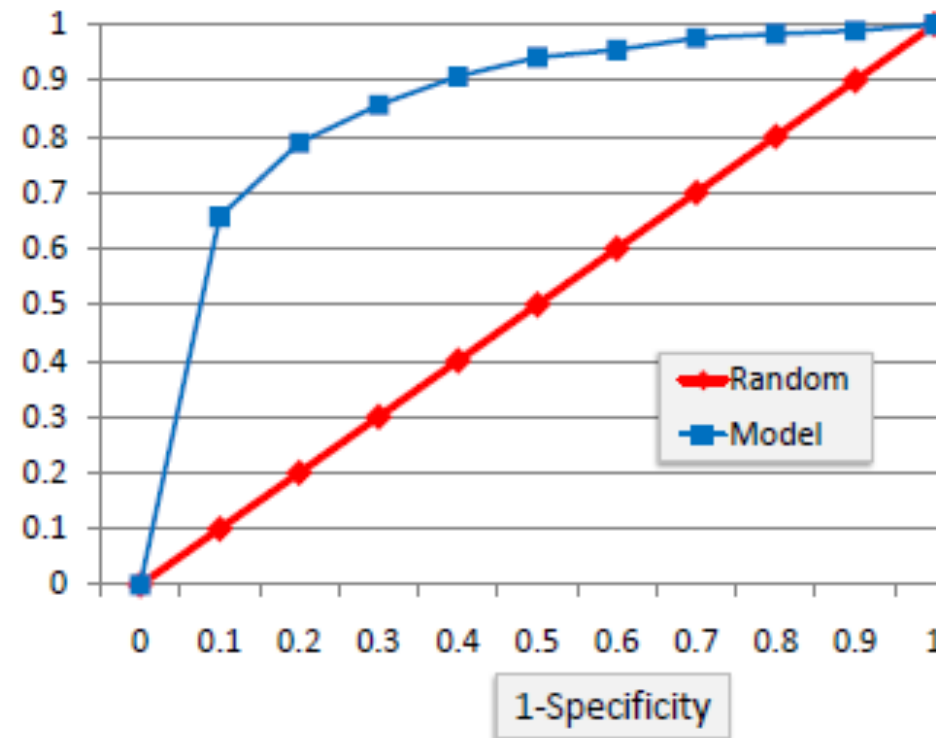
Predicted Classes

Actual Classes

	A	B	C	D	E	F
A	19	4	12	48	25	2
B	50	15	40	25	10	9
C	21	32	12	5	32	15
D	20	29	5	19	45	35
E	2	17	25	36	23	50
F	6	30	25	14	32	23

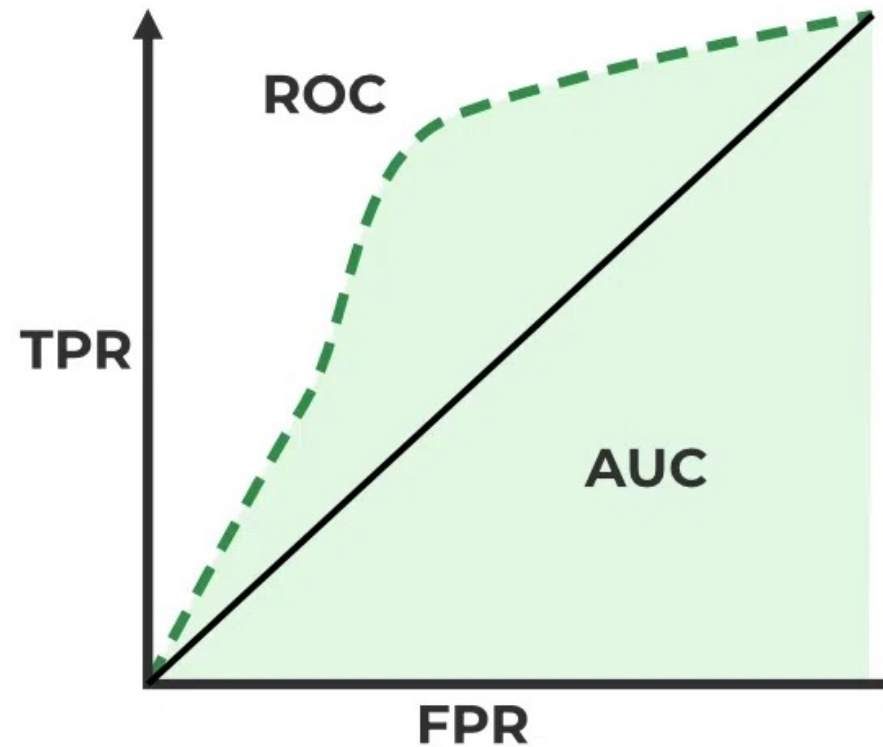
ROC and AUC

- ▶ Receiver Operating Characteristic (ROC) Curve
 - ▶ Charts binary classification model's performance as the discrimination threshold is varied
 - ▶ Plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds
 - ▶ The closer the ROC curve is to the top-left corner, the better the model's performance



ROC and AUC

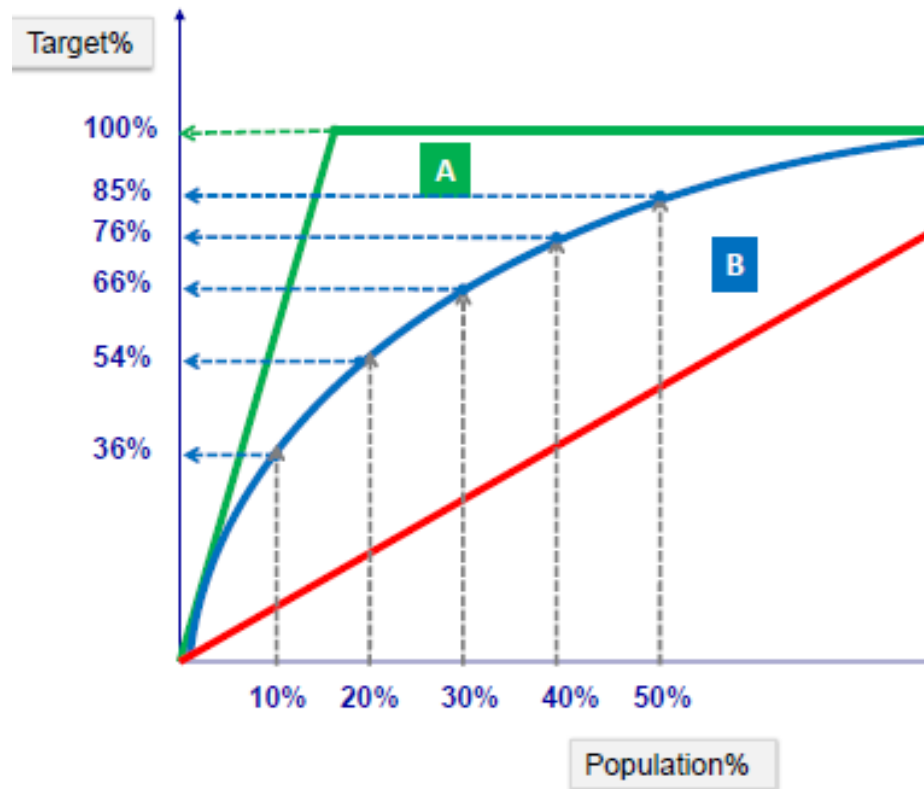
- ▶ Area Under the Curve (AUC)
 - ▶ Single scalar value that quantifies the overall performance of a binary classification model based on its ROC curve
 - ▶ A way to summarize the model's performance across all possible threshold settings, offering a single metric to compare models



Gain and Lift Charts

► Gain Chart

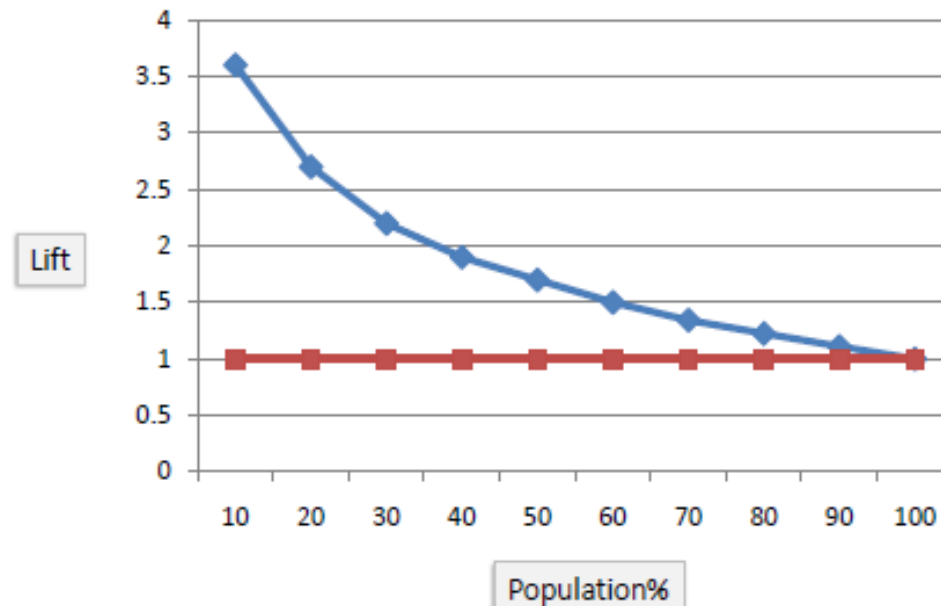
- Helps in evaluating the model's performance by showing the proportion of positive outcomes captured as you target a certain percentage of the population
- Allows you to understand how well your model ranks or prioritizes individuals
- Ideal situation is a steep curve
- Baseline is represented by a diagonal line - the line of random guessing

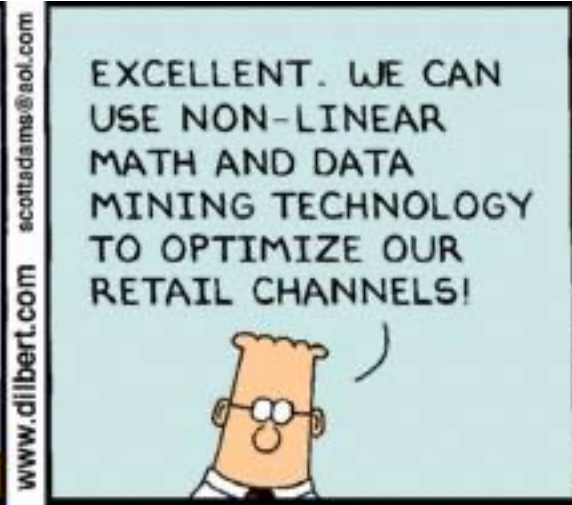


Gain and Lift Charts

▶ Lift Chart

- ▶ Response chart
- ▶ Focuses on how much better the model is compared to random guessing.
- ▶ Compares the cumulative proportion of positive outcomes for a predictive model to the baseline (random) model
- ▶ Help in evaluating the model's performance by showing the proportion of positive outcomes captured as you target a certain percentage of the population





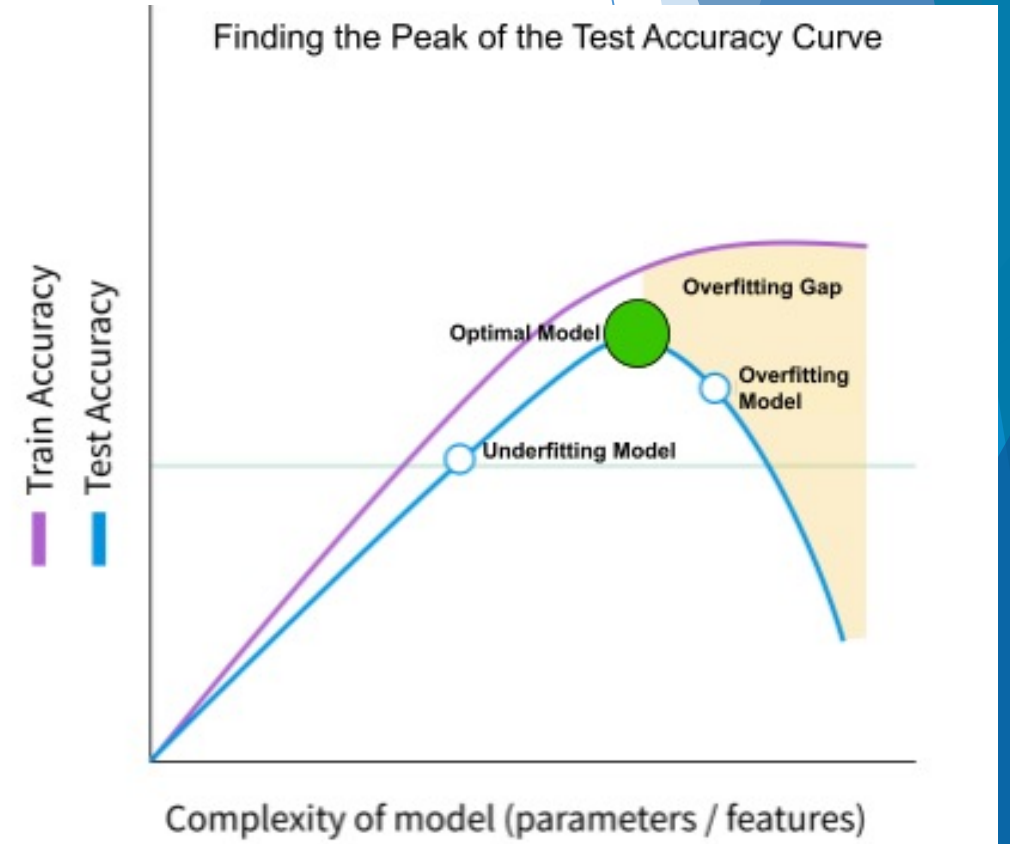
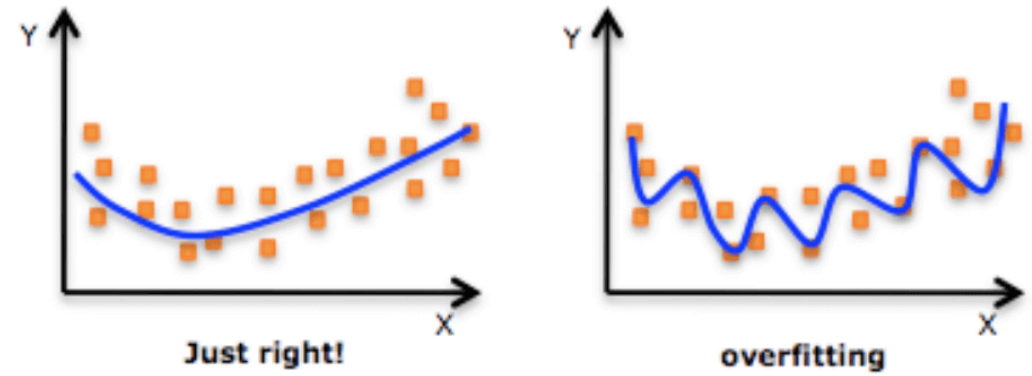
www.dilbert.com scottadams@aol.com

11/13/00 © 2000 United Feature Syndicate, Inc.

Overfitting and Underfitting

▶ Overfitting

- ▶ Model learns the training data too well, capturing noise or random fluctuations in the data rather than the underlying pattern or relationship
- ▶ Loses its ability to generalize and make accurate predictions on new, unseen data
- ▶ Overfitting can lead to poor model performance in practice



Trends and Future Direction



Accelerated Underwriting



Evidence Optimization



Evaluating newer sources of data



Large Language Models and
Underwriting

Accelerated Underwriting

- ▶ Components
- ▶ Revamping efforts
- ▶ Change in the approach
- ▶ Monitoring AU - RHO and PIA
- ▶ Slippage

Accelerated Underwriting

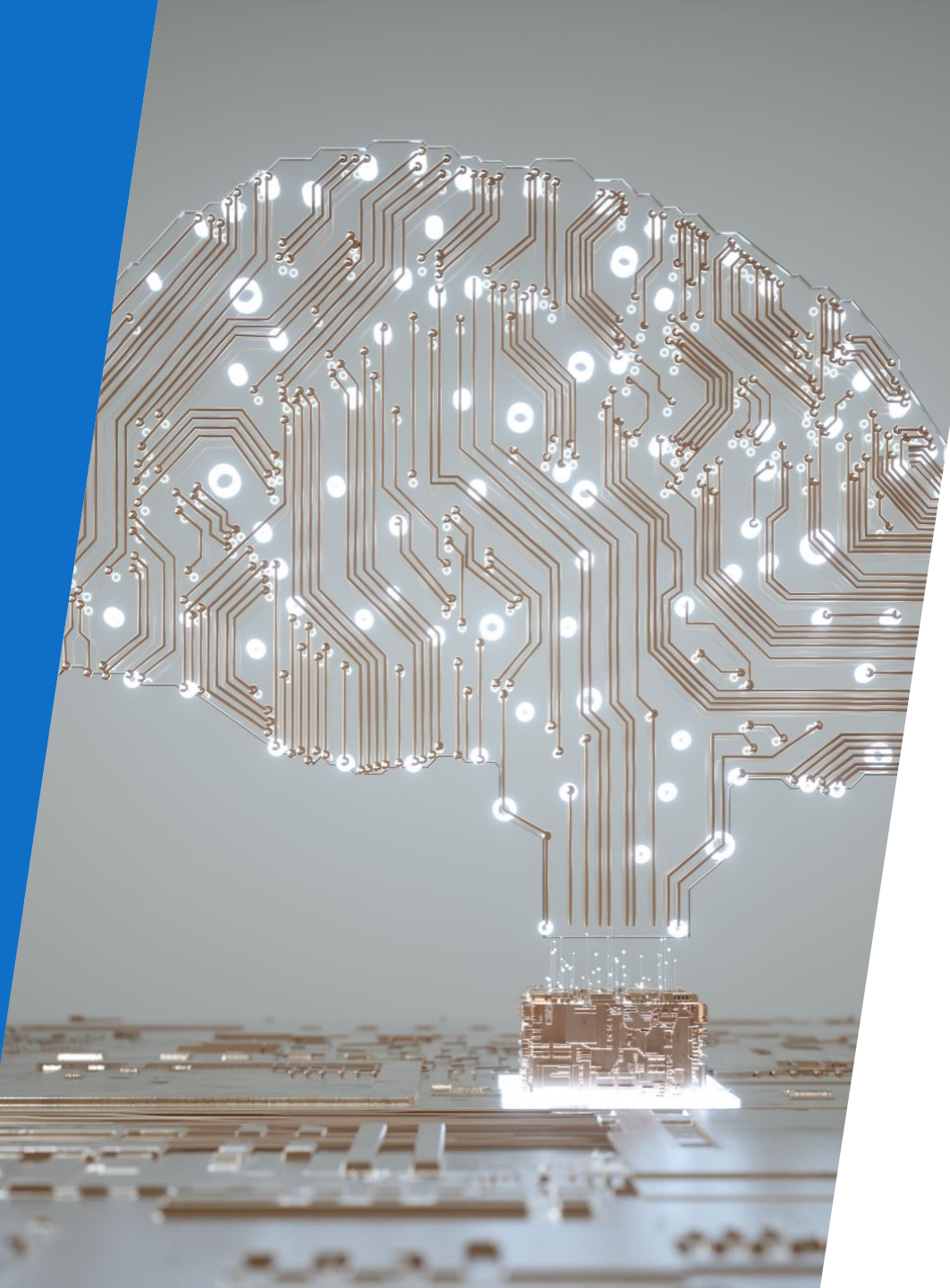
Audits

AUW

AUW/RHO or PIA	NT/SP	NT/P	NT/SS	NT/STD	T/P	T/STD	Sub STD	Decline	
NT/Super Preferred	80%	5%	5%	0	0	3%	3%	4%	100%
NT/Preferred	5%	78%	5%	2%	0	0	5%	5%	100%
NT/Super STD	0	5%	40%	40%	5%	3%	4%	3%	100%
NT/STD	1%	3%	0	90%	3%	0	0	3%	100%

Evidence Optimization

- ▶ Numerous Options
- ▶ Evaluating Data Sources
- ▶ Goal

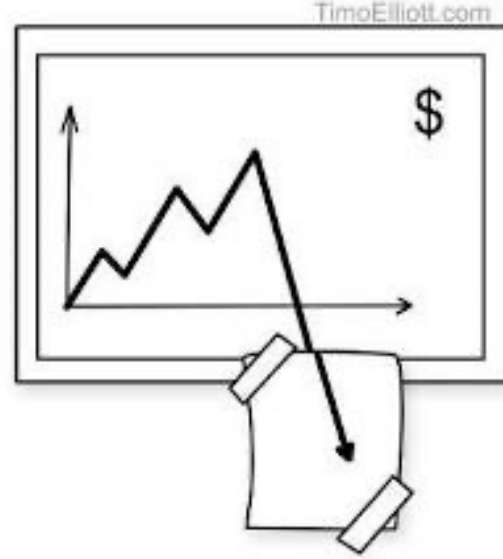
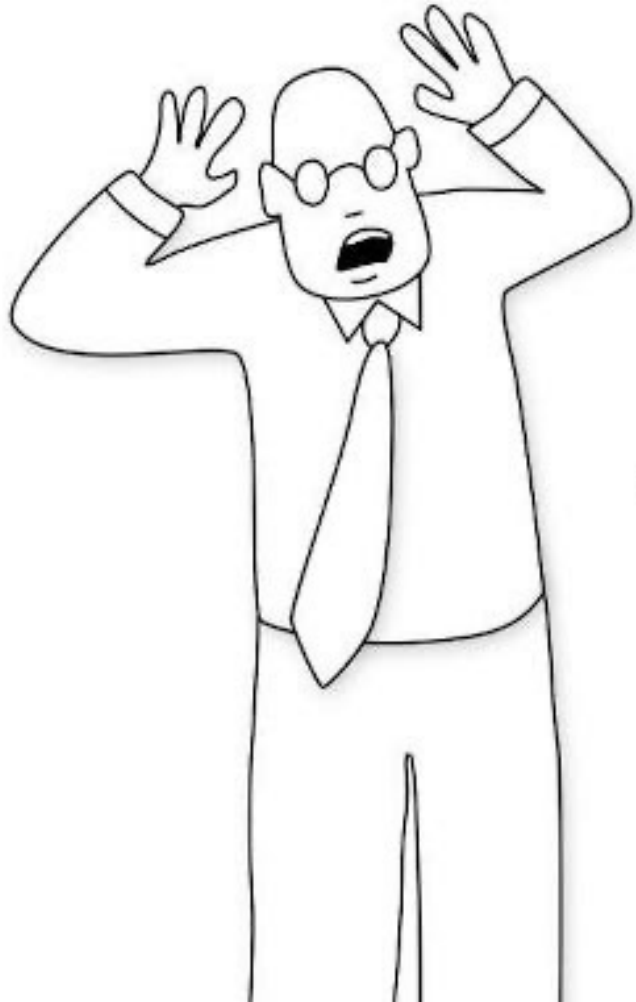


Large Language Modeling and Underwriting

- ▶ Type of artificial intelligence (AI) that uses deep learning algorithms to process and understand natural language
- ▶ Understand, Summarize, Generate and Predict
- ▶ Scope

Role of Medical Director

- ▶ Medical Knowledge
- ▶ Skills
 - ▶ Interdisciplinary communication
 - ▶ Critical thinking
 - ▶ Project management
 - ▶ Leadership
- ▶ Business acumen
- ▶ Data Driven Or Data Informed?



*"Quick! Somebody
find me a
data scientist!"*

Questions?