Machine Learning Overview

Brian Wright, PhD

University of Virginia | School of Data Science

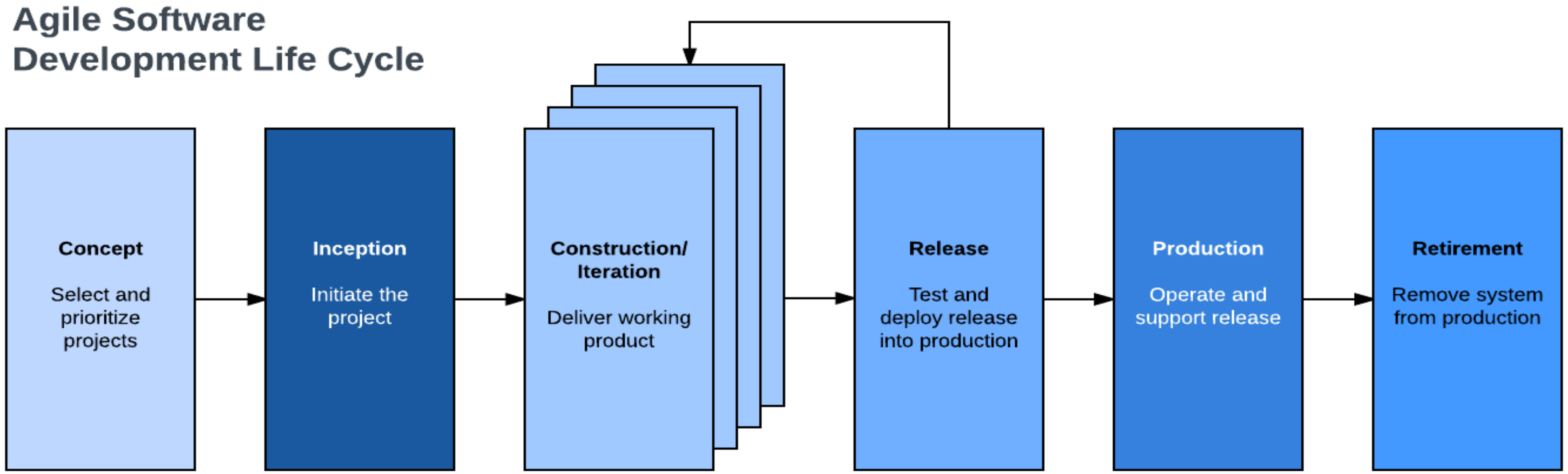
Themes

Machine Learning Lifecycle

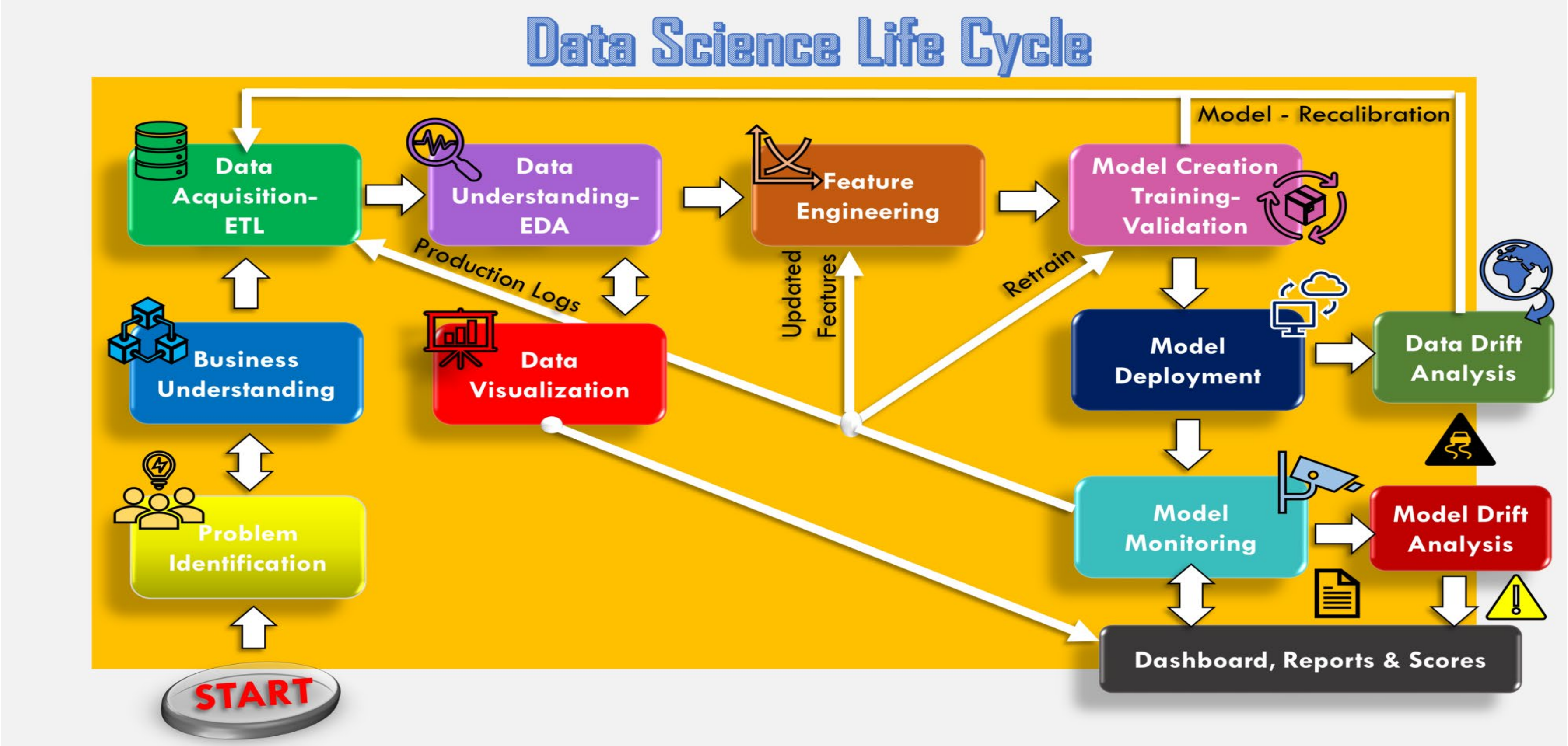
Are you ready for Machine Learning?

Terms and Phases

Engineering of Machine Learning Algos versus Software Development

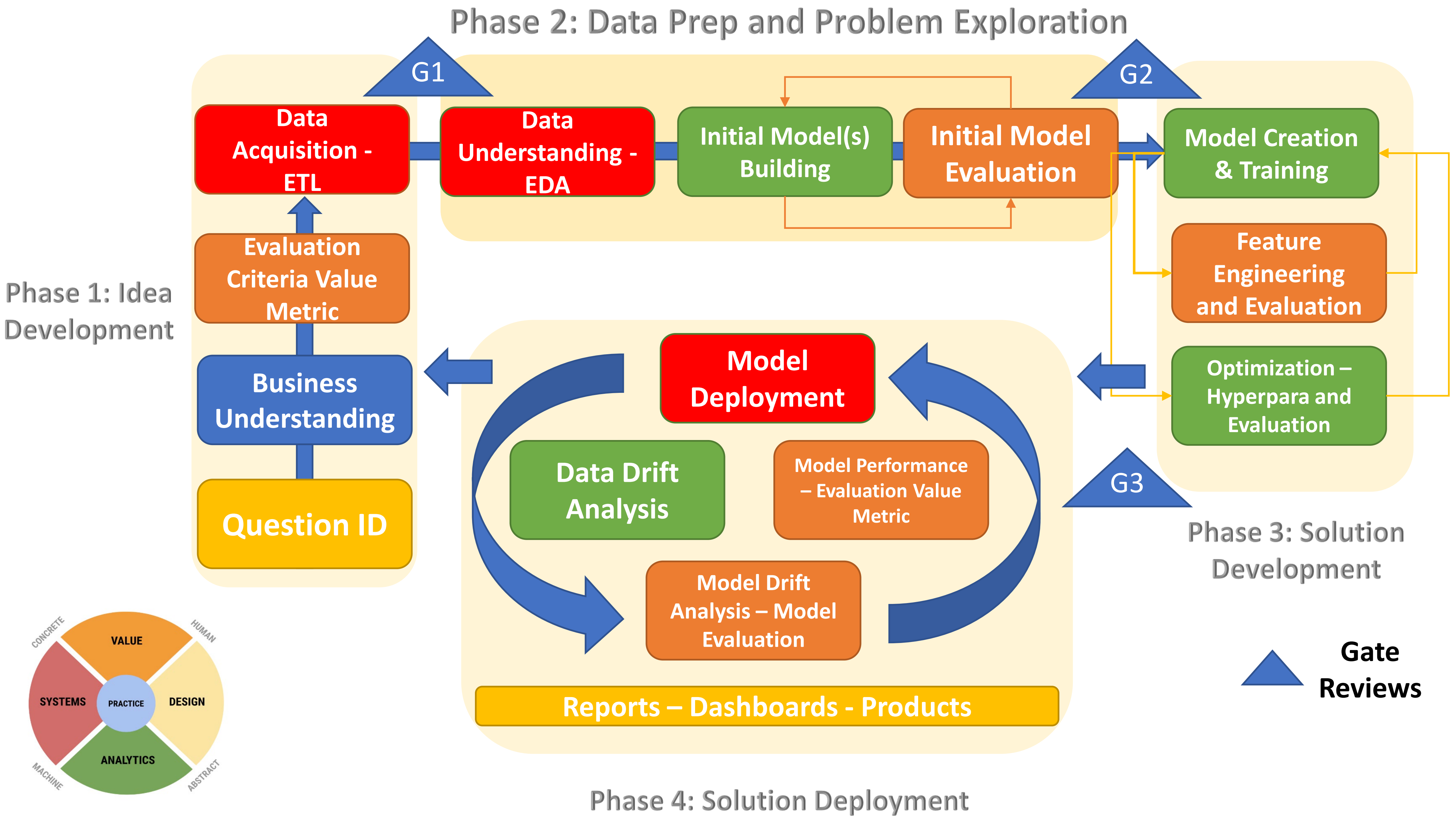


Made in Lucidchart



Source: https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc

Brian's Version of Data Science Lifecycle



Machine Learning Time

“A field of Computer Science that gives computers the ability to learn without being explicitly programmed.”

- Arthur Samuel (Coined the term in 1959 at IBM)

“The ability [for systems] to acquire their own knowledge, by extracting patterns from raw data.”

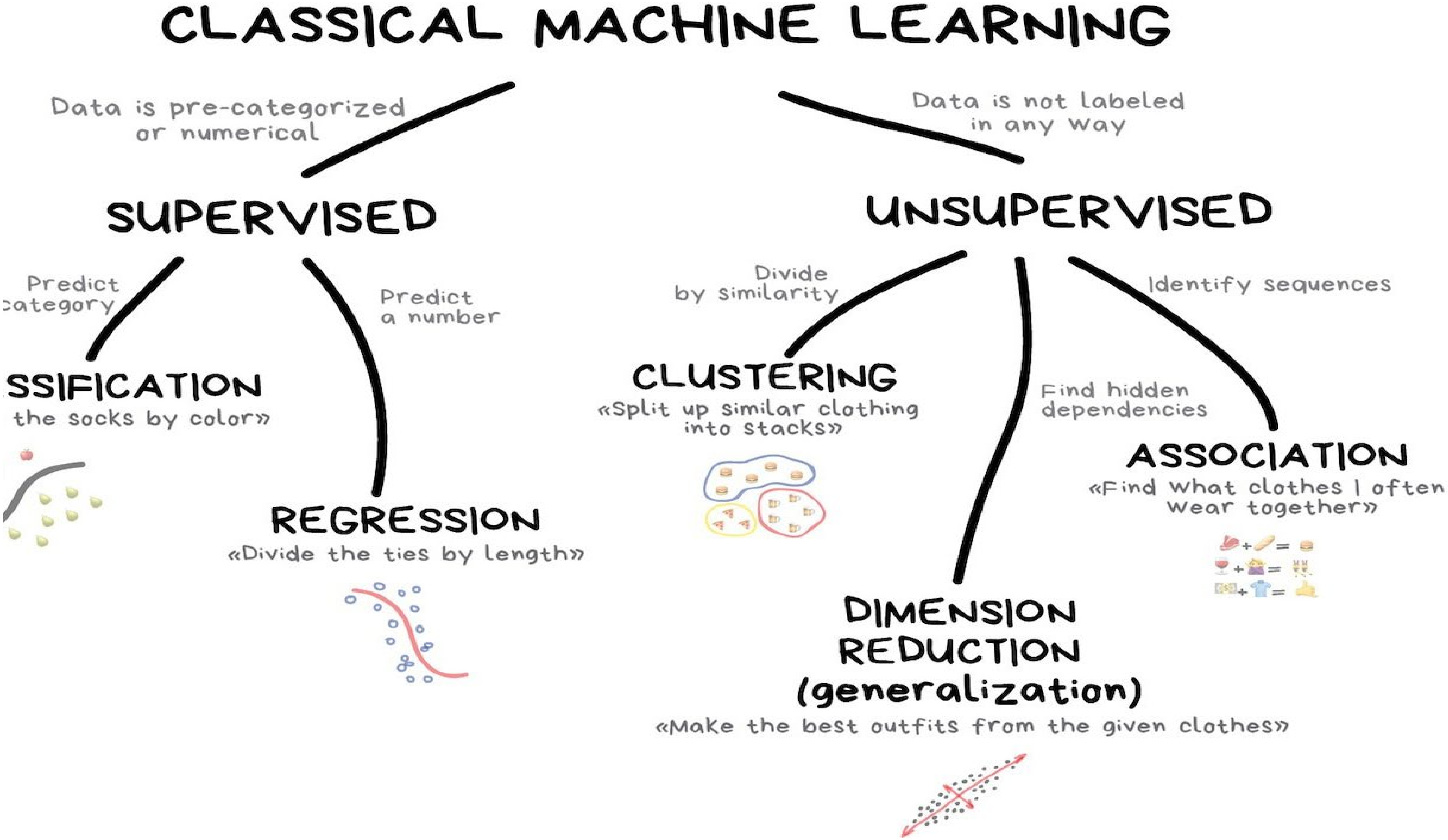
- *Deep Learning*, Goodfellow et al

“A computer program is said to learn from experience E with respect to some set of tasks T and performance measure P if its performance tasks in T, as measured by P, improves with experience E.”

- Tom Mitchell (Computer Scientist & Professor at Carnegie Mellon)

Machine vs. human

	Machine	Human
Understanding context		✓
Thinking through the problem		✓
Asking the right questions		✓
Selecting the right tools		✓
Performing calculations quickly	✓	
Performing repetitive tasks	✓	
Following pre-defined rules	✓	
Interpreting results		✓

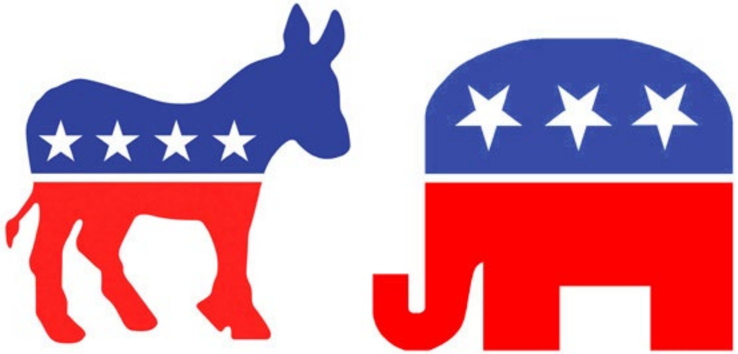


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict category
Predict a number
Divide by similarity
Identify sequences
Find hidden dependencies
SSIFICATION
the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

Supervised machine learning

Pattern discovery when inputs (x) and outputs (y) are known

Input x:
Voter



Output y:
Political affiliation

Examples: Classification and regression are supervised machine learning

Unsupervised machine learning

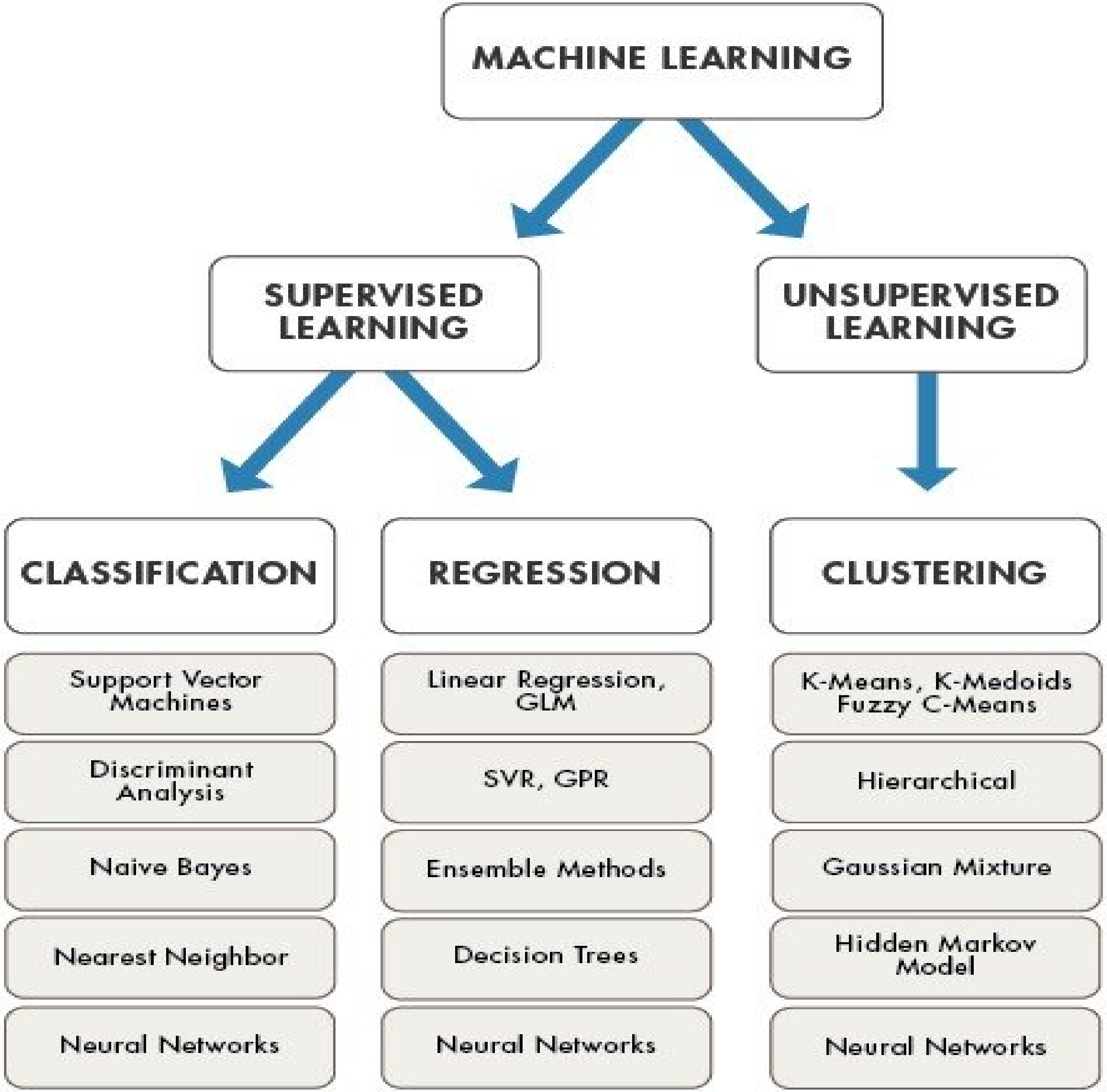
The data inputs *(x)* have no target outputs *(y)*

Input x:
Voter

→ ?

Output y:
Not given
(to be discovered)

We want to impose structure on the inputs *(x)* to say something meaningful about the data


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Neural Networks
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks

Machine Learning is a general use technology what does that mean?

Machine Learning Overview

- A **general-purpose technology** or **GPT** is a term coined to describe a new method of producing and inventing that is important enough to have a protracted aggregate impact.

- Similar to electricity or the internet, in that it can be applied across domains and work to improve market outcomes.

Machine Learning Overview

- Twitter Data Usage

- Error rates on ImageNet (10,000 labelled images) have been driven down from 30% in 2010 to less than 3% today.
  - Below 5% is important why?

- Chess: Deep Blue (IBM AI) searched some 200 million positions per second, Kasparov was searching not more than 5–10 positions probably, per second. Yet he played almost at the same level….why?

Machine Learning Overview

- However, before we all turn into robots consider two important facts:
  1. We remain remarkably far away from what would be consider a similar general intelligence that can be compared to humans
  2. Machines cannot do the full range of tasks that humans can do

We can then refer to jobs or activities that might be good cases for Machine Learning as SML or Suitable for Machine Learning

What are examples of tasks that might be SML and how do we know if our organizations are ready?

Machine Learning Overview

- Successful implementation of ML requires very **detailed specifications** on what is to be learned and data to support that learning activity.

- Including the development of engineering features through a series of **trial-and-error** and..

- Then most importantly embedding these products into **normal business operations** in such a way that efficiencies can be realized.

Machine Learning Overview

- What tasks are most suitable for ML to take over:
  - Most recent successes are predicated on **supervised learning**
  - Competency is narrow as compared to the complexity of **human** decision making

1. Learning a function that maps well-defined **inputs** to well-defined **outputs**
   - If can predict Y given any value of X – still might not produce the actual causal effect
2. **Large Data** is present or can be created containing input-output pairs
   - The more training data available the more arcuate the model
3. Task provides **clear feedback** with well definable goals and metrics
   - If we know what to achieve – (optimize flight patterns not a single flight)
4. Where **reasoning** and diverse background knowledge is not necessary
   - Good at empirical associations but terrible at decision making that requires common sense of historical knowledge
5. No need for **why** the decision was made to be clear
   - NN could use millions numerical weights

Machine Learning Overview

6. A tolerance for error or sub-optimal solutions
   - ML use probabilistic outputs which means some error is always assumed
7. Function of item being learned should not **change rapidly** over time
   - Work best when the distribution of future test examples is the same roughly as the training set over time
   - If not the case systems need to be in place to refresh algorithms

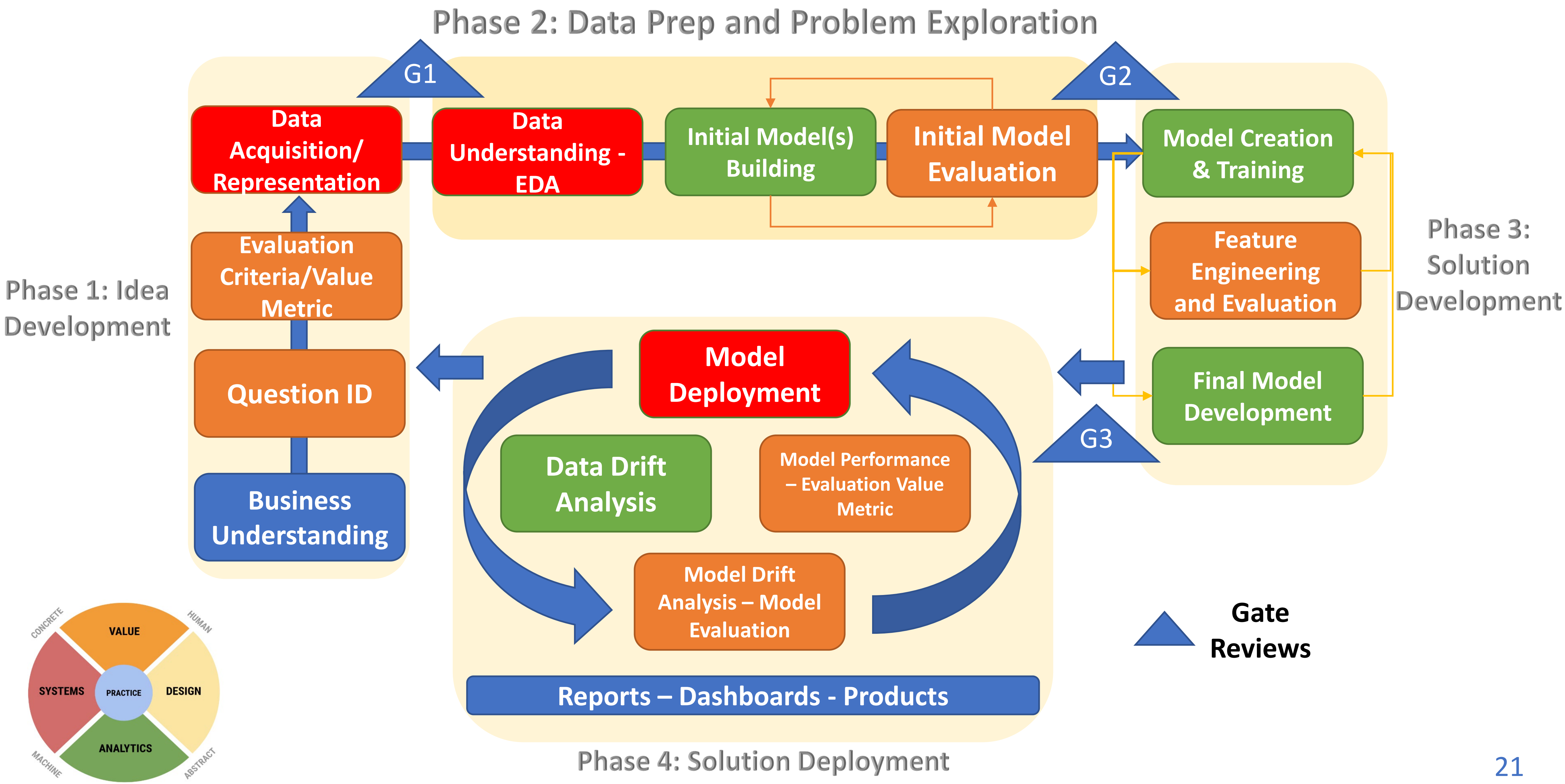
How do machines learn?

- The basic machine learning process can be divided into three parts.
  - Data Input: Past data or information is utilized as a basis for future decision-making
  - Abstraction: The input data is represented in a broader way through the underlying algorithm
  - Generalization: The abstracted representation is generalized to form a framework for making decisions

(reference Introduction to ML by Subramanian Chandramouli, Saikat Dutt, Amit Kumar Das (https://learning.oreilly.com/library/view/machine-learning/9789389588132/xhtml/chapter001.xhtml#ch1_1)


Input
Data
Abstraction
Generalization

Brian's Version of Data Science Lifecycle

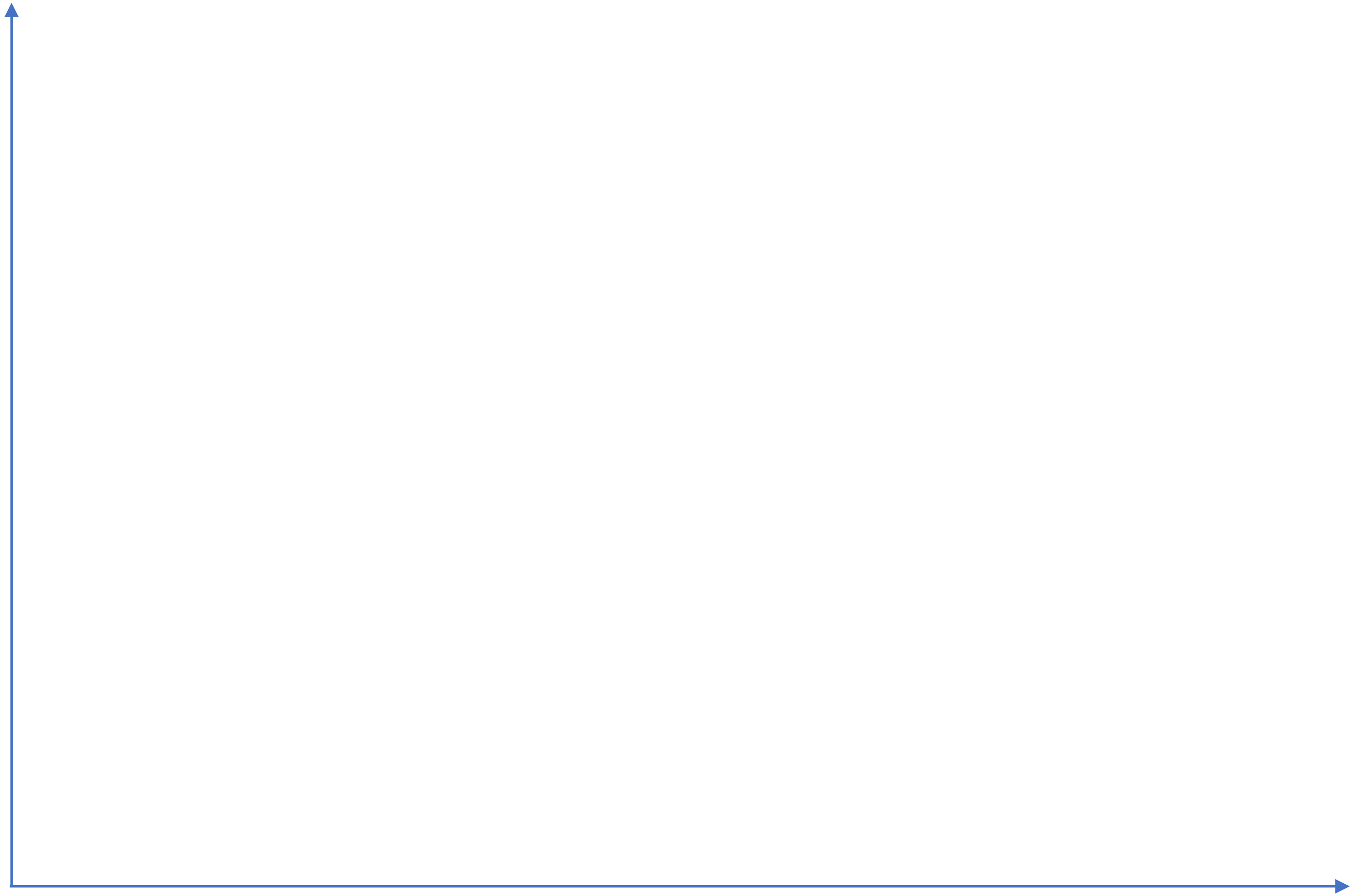


Overview of SOME Key ML Methods/Terms

- Phase – 1 Idea Development
  - Prediction versus Inference
  - Independent Metric for Business Value
  - Target Variable and features
  - Classification versus Regression
  - Probabilistic Interpretation
  - Data acquisition/gathering
- Phase – 2 Data Prep and Problem Exploration
  - Variable classes/types
  - Scaling and/or Normalizing Data/One-Hot Encoding
  - Missing Data
  - Baseline – prevalence
  - Data Partitioning/Sampling
  - EDA and (Summary Stats and Visuals)
- Phase – 3 – Solution Model Development
  - Parameters versus Hyperparameters
  - Thresholding
  - Feature Engineering
  - Bias versus Variance Tradeoff
  - Model Evaluation
  - Non-parametric modelling (random state)

Phase I

Prediction versus Inference



Prediction versus Inference

- Goals of prediction are not centered on how the features are interacting or resulting in an event but are instead focused on the ability of the model to predicted an event.
- Almost all ML methods are focused on predication not causation or inference.
- This is why model performance is based largely on how well a model predicts not necessarily how much individual variables are contributing to error reduction.

Overview of Key ML Methods/Terms

- Independent Metric for Business Value
  - A key part of building a solution using **Machine Learning Techniques** is having a metric that is independent of the model that can be used to determine if the model is providing value.

  - Examples
    - Recommender Engine for Netflix: Number of user clicks
    - Spam Block Predictor: Number of viruses in the network
    - Market Clustering: Did sales increase
    - Others?

Overview of Key ML Methods/Terms: Target Variable versus Features

- **Target variable** – Is the variable that includes the patterns the machine learning algorithm is trying to learn. It is the variable of interest and key to evaluating the model output.
  - More simply it is the variable we are trying to predict.

- **Feature variables** – Are the variables the model will use to learn the patterns of the target variable. The process of feature engineering can result in additional features.
  - More simply these are the variables used for predicting the target

Overview of Key ML Methods/Terms: Classification versus Regression

- **Classification** is the process of developing a model to predict whether a target variable is in defined categories. This is driven by having either a binary or multi-level categorical variable as the target variable.
  - Examples:
    - Predicting whether someone is male, or female based on 1,000s of pictures.
    - Predicting whether a team will have a winning season or not based on player performance
    - Predicting whether a person will default on a loan or not
  - Key point: The predications of the model are not binary (1s or 0s) but are given as **percentages** indicating the likelihood that any one row of data belongs to any one category. In the case of target variables with multiple categories each row will get the same number of percent predictions as categories.

Overview of Key ML Methods/Terms: Classification versus Regression

- **Regression** is the process of developing a model to predict a specific number or range of numbers. This is driven by having a continuous variable as the target variable for the model
  - Examples:
    - Predicting the score given the players playing a game.
    - Predicting an amount of rain given weather conditions
    - Predicting a persons weight based on various personal statistics

Overview of Key ML Methods/Terms: Probabilistic Interpretation

- A significant portion of this class will focus on building models for classification. Classification is a much more common machine learning goal versus regression.
- We all know the range of values for probabilities, 0 to 100, the key to understanding these outputs is to think of them as **risk measures**, with 100 being no risk and 0 being all the risk!
- How the outputs are used will depend on your question.
  - Example: How certain do you want to be that a drug is effective as compared to whether a customer will open a marketing email? The results could both yield 75% probabilities but is that high enough?
- Could also think of the outputs as a quantification of uncertainty, the question becomes given your problem how much uncertainty are you willing to accept?

Overview of Key ML Methods/Terms: Data Brainstorming

- Data to Concept – Does the data available support the algo target and goal
  - How difficult is the data to gather?
  - Is the data large enough?
  - What is the rate of change of the data?
  - Do we believe this is the correct source and data content to address the problem?

- Learning Difficulty – How complex or vague is the target variable?
  - Are there imbalances in the classes?
  - Does the data clearly link to the problem?
  - Has this data been used in the past, to what success?
  - Is the target difficult to measure or break into smaller components?
  - What risk level are you willing to accept given the question?

Phase II

Overview of SOME Key ML Methods/Terms

- Phase – 1 Idea Development
  - Prediction versus Inference
  - Independent Metric for Business Value
  - Target Variable and features
  - Classification versus Regression
  - Probabilistic Interpretation
  - Data acquisition/gathering
- **Phase – 2 Data Prep and Problem Exploration**
  - **Variable classes/types**
  - **Scaling and/or Normalizing Data/One-Hot Encoding**
  - **Missing Data**
  - **Baseline – prevalence**
  - **Data Partitioning/Sampling**
  - **EDA (Summary Stats and Visuals)**
- Phase – 3 – Solution Model Development
  - Parameters versus Hyperparameters
  - Thresholding
  - Feature Engineering
  - Bias versus Variance Tradeoff
  - Model Evaluation
  - Non-parametric modelling (random state)

Overview of Key ML Methods/Terms

- # Baseline – prevalence

Phase III

Overview of SOME Key ML Methods/Terms

- Phase – 1 Idea Development
  - Prediction versus Inference
  - Independent Metric for Business Value
  - Target Variable and features
  - Classification versus Regression
  - Probabilistic Interpretation
  - Data acquisition/gathering
- Phase – 2 Data Prep and Problem Exploration
  - Variable classes/types
  - Scaling and/or Normalizing Data/One-Hot Encoding
  - Missing Data
  - Baseline – prevalence
  - Data Partitioning/Sampling
  - EDA (Summary Stats and Visuals)
- **Phase – 3 – Solution Model Development**
  - **Parameters versus Hyperparameters**
  - **Thresholding**
  - **Feature Engineering**
  - **Bias versus Variance Tradeoff**
  - **Model Evaluation**
  - **Non-parametric modelling (random state)**

Overview of Key ML Methods/Terms

- # Feature Engineering – Combining or exploring different levels of variable that best work in your model. Likely going to dedicate a week to just this topic.

Overview of Key ML Methods/Terms

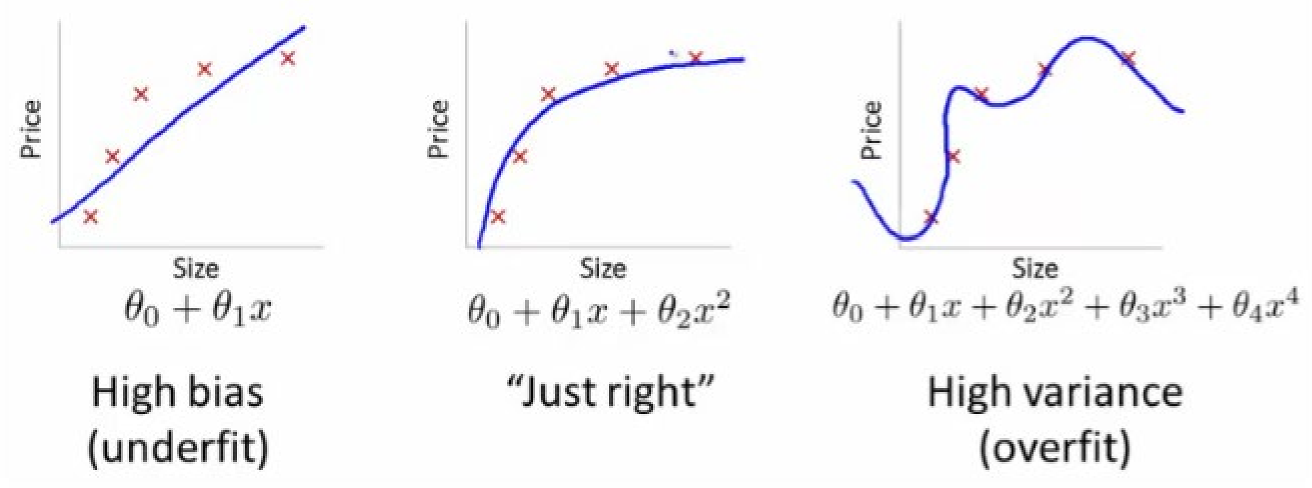
- Thresholding – The percentage point where our models will predict the result to be either a 0 or 1, in the typical binary case.

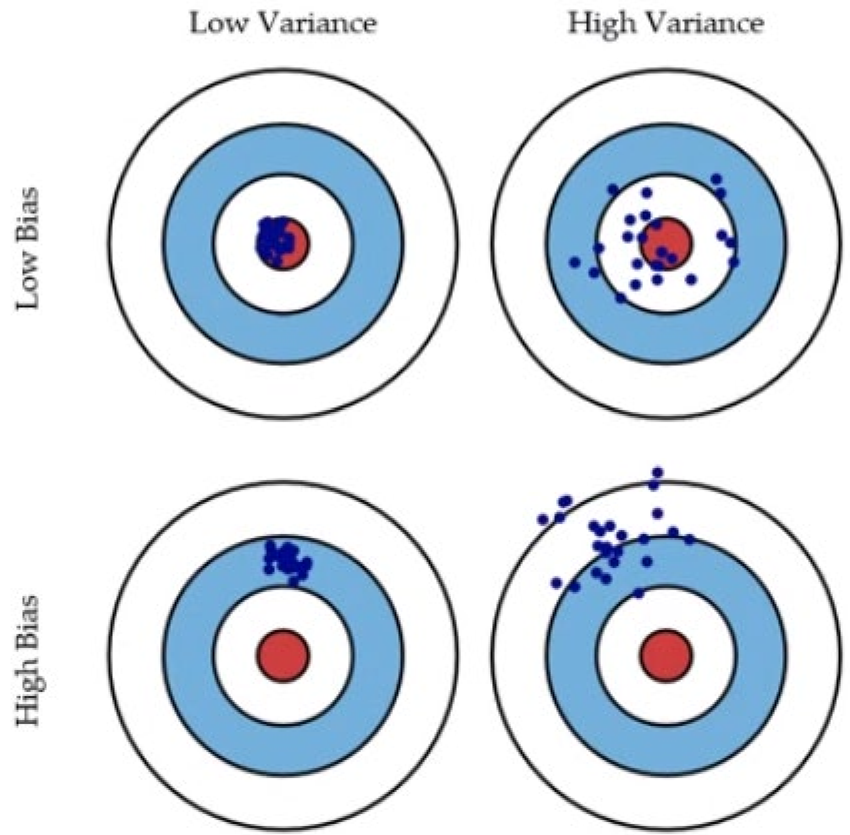
- Adjust the threshold associated with indication of a positive class. The default is 50%, could be that we want to be extra careful and instead adjust that measure up to 75% or 90%.

Overview of Key ML Methods/Terms

- Evaluation – The metrics you use to assess model quality. There are a ton of this measures, and we are dedicating an entire week to the exploring these further. I'll show some examples in the code for this week.

Bias Versus Variance





Extra Material

Bookings.com

Lesson Learned: Booking.com

Bookings.com

- Swiss Army Knife– Their approach to ML is highly **adoptable**, meaning it can be used in a variety of settings – generate specific results or more generalizable depending on the inputs (data)

- Offline Health Check– Use Randomized Control Trails (RCT) to test model outputs aligned with normative business metrics to assess quality (customer conversion)
    - Increase model performance doesn’t necessary translate to better gain in value

Bookings.com

- Make a Target Before you Shoot – Develop a clear understanding of the business case and target variable (what is date flexibility)
  - Learning Difficulty – How complex or vague is the target
  - Data to Concept – Does the data available support the algo target and goal
  - Selection Bias – Does the model perform better for a subset of the target

- Speed Kills – ML algos, even simple ones, take a lot of computing power – to reduce user weight time (latency) measures should be taken
  - See page 1748 (sparsity, model redundancy, caching…etc.)

Machine Learning Overview

- **Keep a watchful eye** – Used specialized monitoring tools to understand how the models are performing in practice (even when the result was unclear)

- Traditional Research Methods (Experimental Design) is a Best Practice Approach to ML –
  - "Experimentation through Randomized Controlled Trials is ingrained into Booking.com culture"