What is Machine Learning, Anyway

Jim Box, SAS Institute

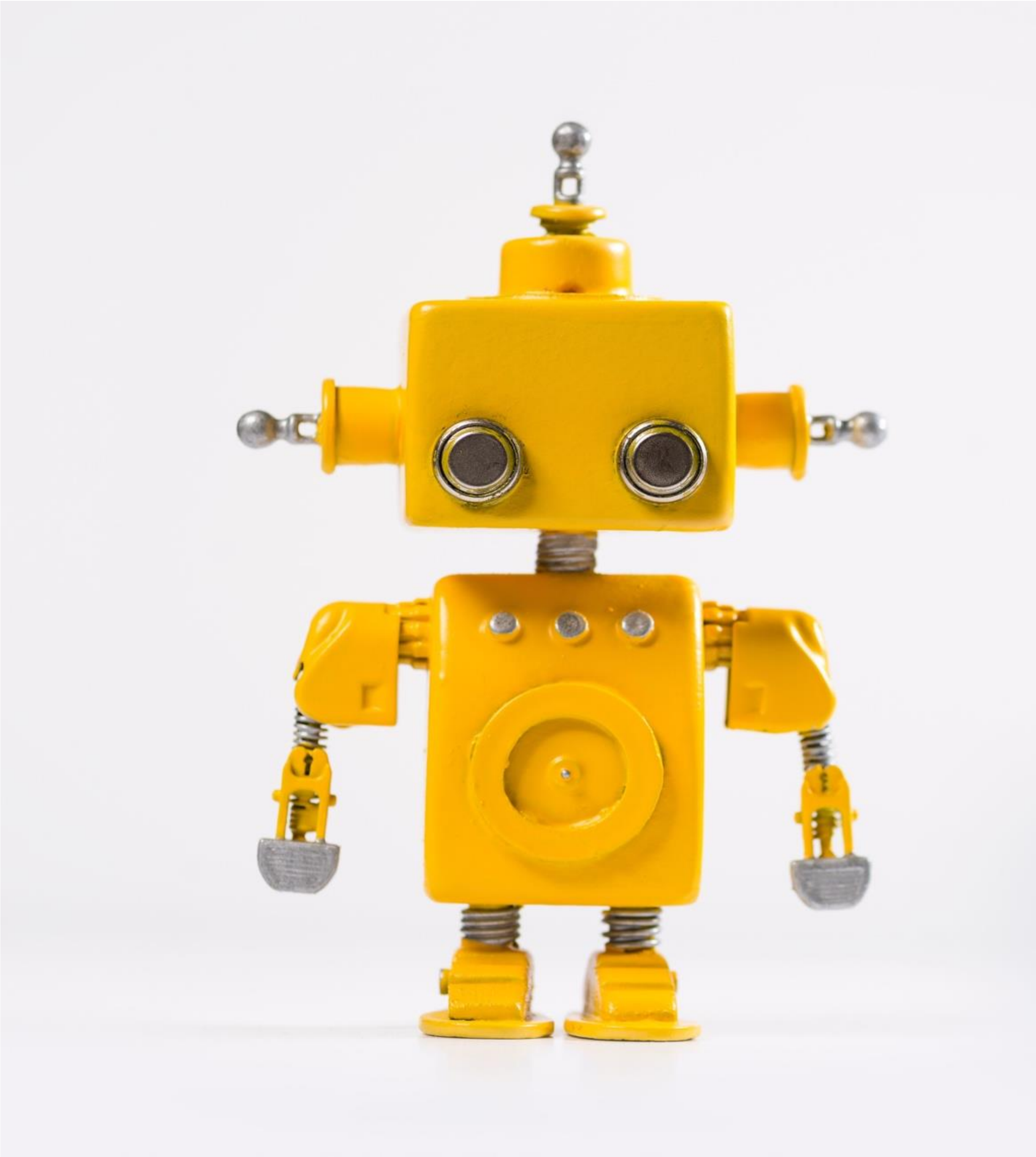
Let’s Start with a Definition

Machine Learning

- Algorithms that train computer systems to improve their performance on specific tasks without being given explicit rules on how to do that
- The “computer systems” up there is an Artificial Intelligence
- ML is a branch of AI, but can also be used outside an AI application

Let's Start with a Definition

Machine Learning



Learning from machines,
We know what it means,
To use algorithms that can reason,
And make decisions with precision.

It is the power of data,
That can be used to inform,
To detect patterns and trends,
And keep us all in the firm.

The ability to recognize,
The patterns we choose,
Enables us to take action,
And respond to the news.

Machine learning is the future,
It will bring us so far,
We can build smarter systems,
That are better by far.

Machine Learning vs Statistical Inference

What’s the Difference?

	Statistical Inference	Machine Learning
Data	Data has some underlying probability distribution	Wide range of data types
Goals	Test hypotheses / explore relationships	Make predictions / decisions
Approach	Specify a model and estimate parameters	Learn patterns by itself
Scope	Smaller data with a specific question	Large, complex data with many variables
Validation	Check the underlying assumptions about the data	See how the model does on data from a hold-out sample

Machine Learning Methods

Learning Types

- Supervised Learning – Algorithms are trained on a labeled dataset and work to predict outcomes
- Unsupervised Learning – Techniques where data is not labeled or classified, so there is no target to predict
- Deep Learning – ML techniques that use multiple layers of nonlinear processing to analyze data and create predictions

Supervised Learning

Popular Models and Use Cases

Supervised Learning

Use Cases

- Generally used for predictive models, mostly binary (Yes/No) predictions
- Can be used for continuous variable predictions
- Data is partitioned into a Training Set (~70%) and a Validation set (~30%) to see how the model does on data it was not exposed to in training -- this is to minimize overfitting the model.

Supervised Learning Models

Logistic Regression

- Similar to Linear Regression, except you are predicting the **probability** of an outcome
- Produces an equation to predict that probability of success
- Hypothesis tests for each variable's significance
- You must specify a **cutoff point** that determines if the prediction is a success or failure
- Cutoff point selection is critical, but often overlooked
- Case complete model

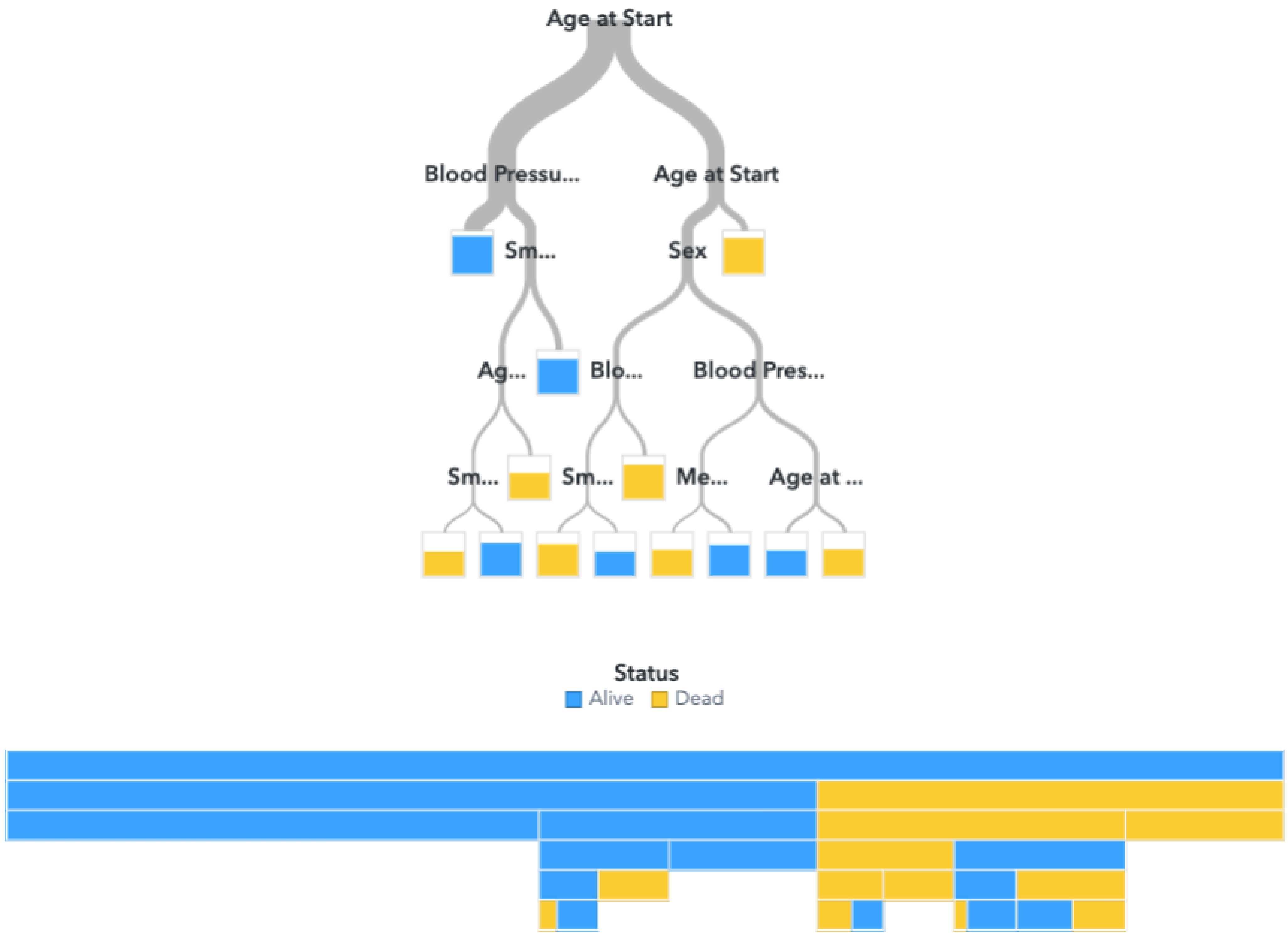
Supervised Learning Models

Decision Trees

- Find the values of variables that can split the data into branches that lead to leaves where there are a high number of success or failures
- Algorithm searches through all values of all variables
- Handles missing values like any other value
- Very easy to interpret and use
- Can limit the numbers of splits in a branch, the number of levels possible and the smallest leaf size

Supervised Learning Models

Decision Trees



Supervised Learning Models

Random Forests

- Data is sampled (with replacement) to create multiple subsets of observations (bootstrapped samples)
- Make a decision tree for each bootstrapped sample, using a random selection of variables
- Once all trees are trained, predictions are made by running new data through each tree. Final prediction is based on majority vote
- Essentially crowdsourcing predictions
- Less prone to overfitting than a single tree

Supervised Learning Models

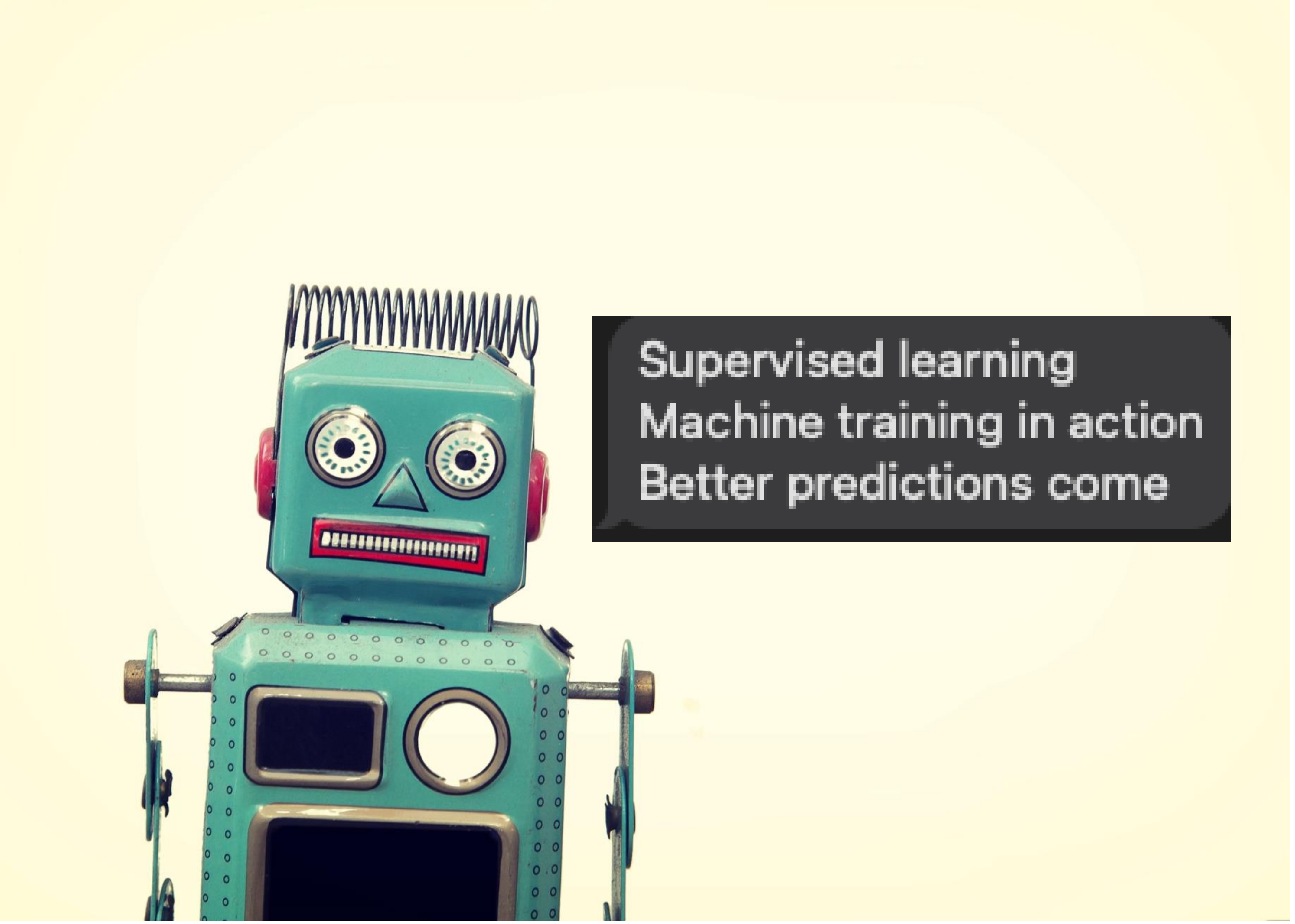
Gradient Boosting

1. Train a tree to make predictions
2. Calculate errors between predicted and actual value
3. Build a new model to predict errors in step 2
4. Combine the base model and error model to make new predictions and calculate new error rates
5. Repeat 2-4 until error rate stops improving
6. Final version gives prediction

Good for complex relationships between features and targets

Supervised Learning Models

Summary



Unsupervised Learning

Popular Models and Use Cases

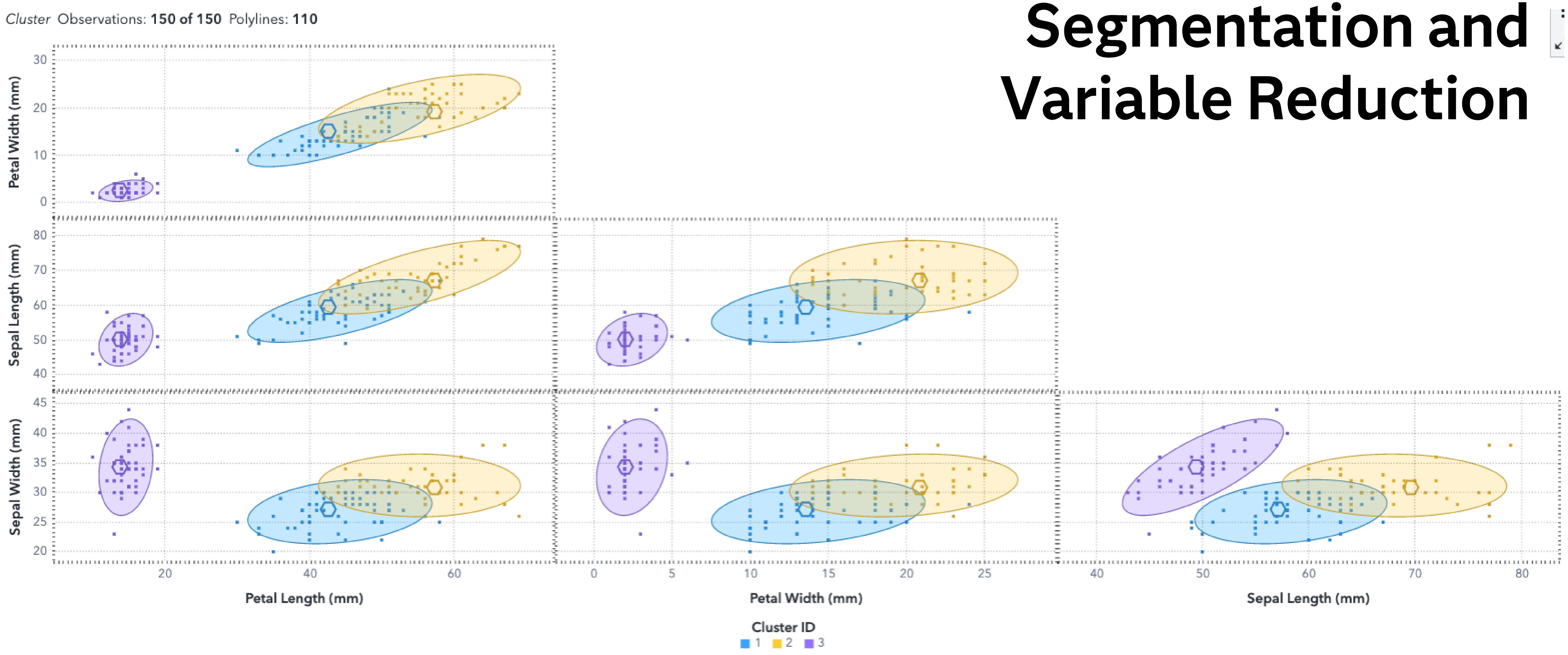
Unsupervised Learning

Use Cases

- Goal: Discover patterns, relationships, and structures in the data
- Most Common use cases:
  - Customer Segmentation
  - Anomaly Reduction
  - Variable Reduction
  - Recommendation Systems
  - Fraud Detection
  - Network Analysis

Unsupervised Learning

Clustering



Segmentation and Variable Reduction

Unsupervised Learning

Clustering

1. Pick a starting number of clusters
2. Randomly assign center points (**centroids**)
3. Determine the distance from each point to each centroid
4. Assign each point to the closest cluster
5. Find the new centroid of points assigned to each cluster
6. Repeat 3-5 until no observations move
7. Repeat the whole process with a different number of clusters

Unsupervised Learning

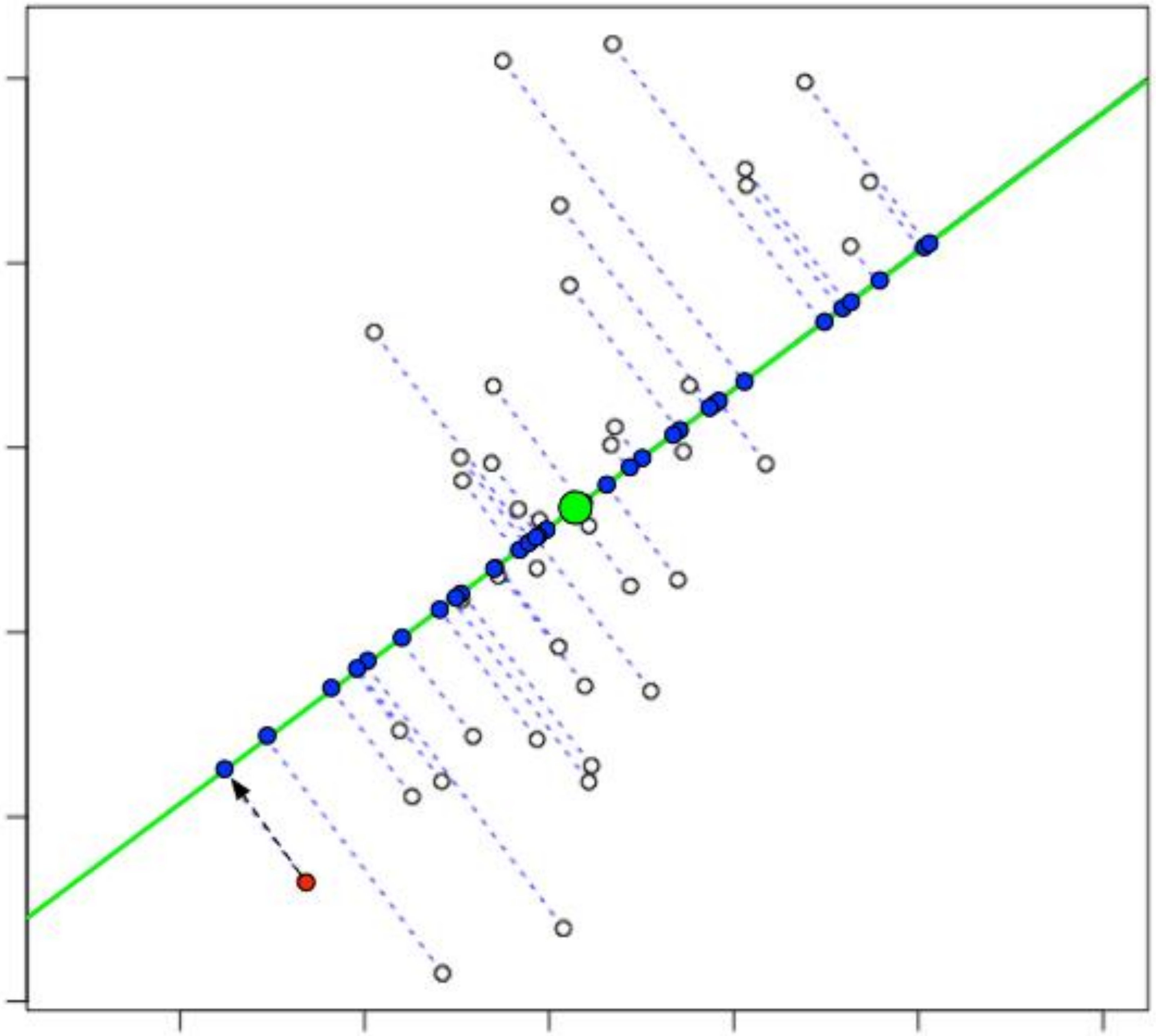
Principal Components Analysis (PCA)

- Reduce the number of variables (features) in a dataset to make modeling smoother (**reducing dimensionality**)

- Basically, instead of needing 12 variables to explain the variance in the data, make a new set of coordinates and only need 3 features

Unsupervised Learning

Principal Components Analysis (PCA)



Unsupervised Learning

Association Analysis

- Identifies sets of items that are frequently obtained together (Beer and pretzels, milk and cereal, similar types of movies streamed, that sort of thing)

- Generates a bunch of association rules

- Ranks the rules based on how often things are consumed together

- Besides market basket analysis, also used in fraud detection and supply chain optimization

Unsupervised Learning

Summary

Deep Learning

Popular Models and Use Cases

Deep Learning

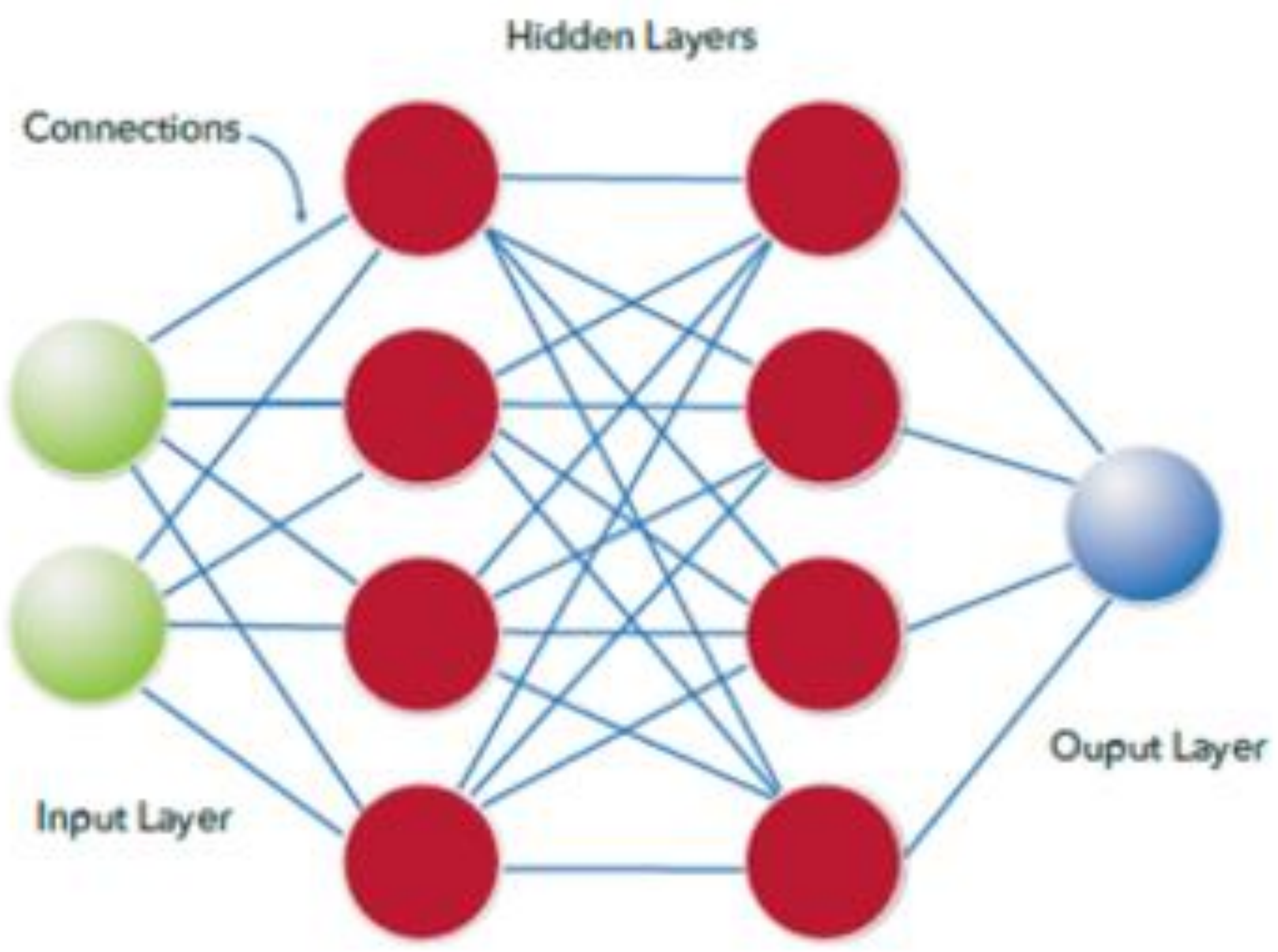
Definition and Use Cases

- Neural Networks – complex networks of neurons that make many steps of models to get from inputs to predictions

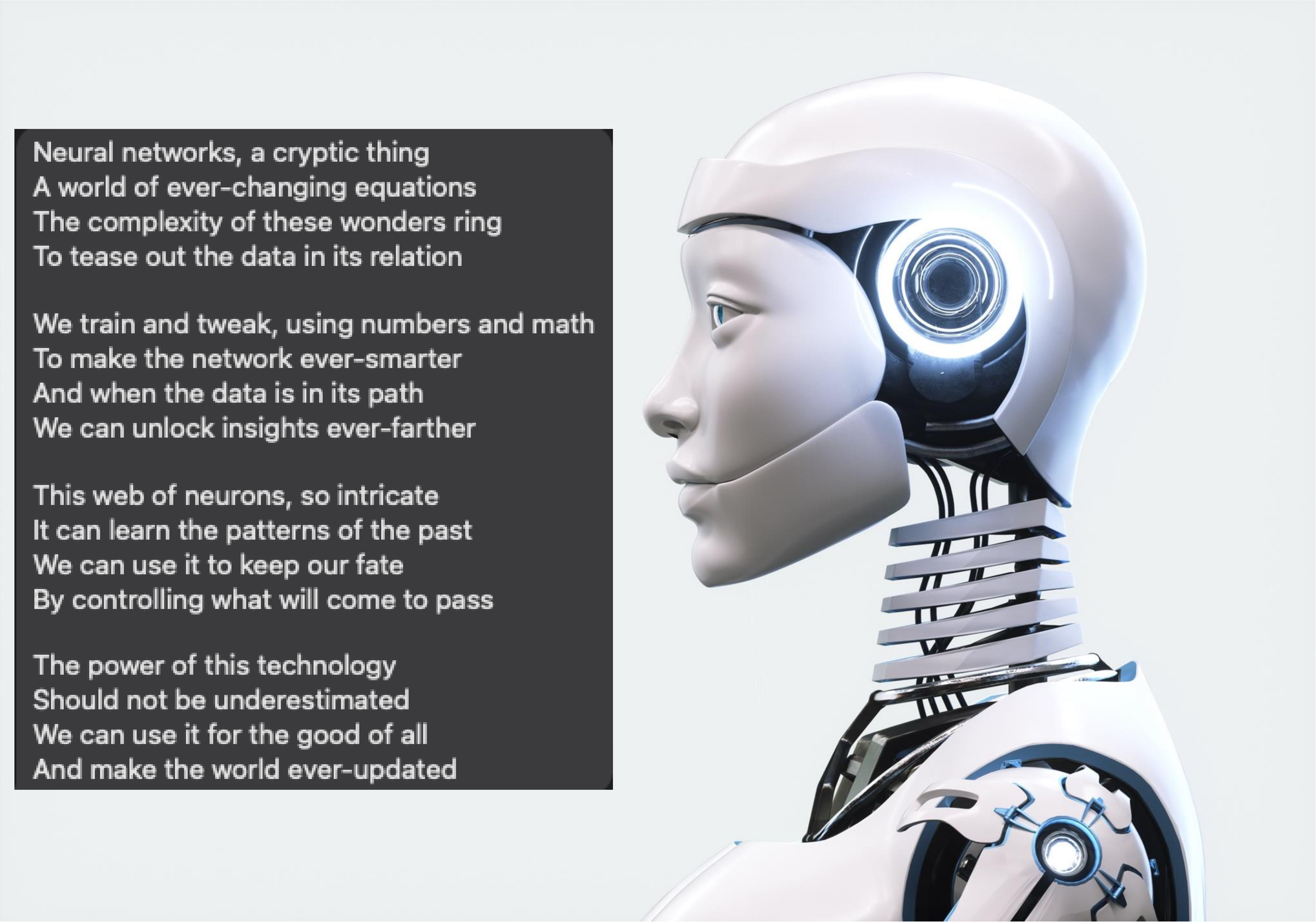
- Computationally intensive and very hard to understand what's happening (black box models)

- Common Use cases:
  - Computer vision & pattern recognition
  - Natural language translation and sentiment analysis
  - Generating new data, such as images or videos

Deep Learning
Neural Networks



Deep Learning
Summary



Model Selection

Methods of Evaluation

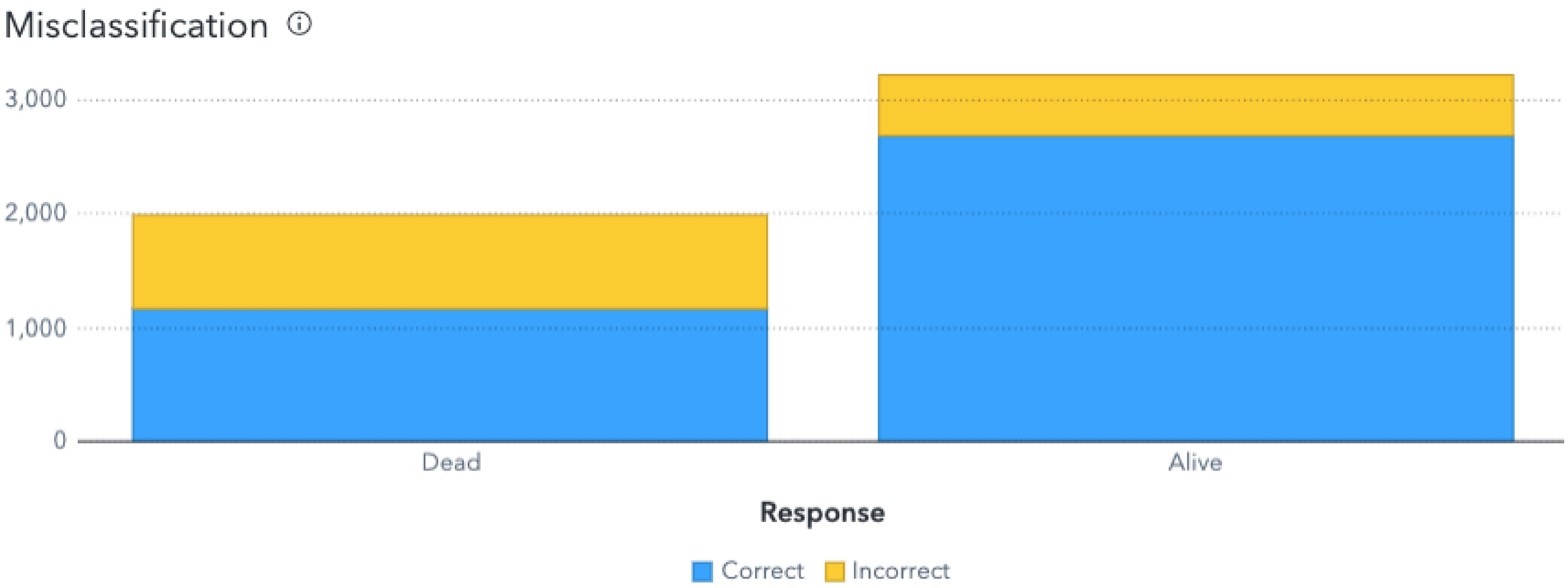
Model Selection

Model Process

- Partition the data into Training and Validation
- Train a bunch of models in the Training data
- Use the model to generate predictions on the Validation data
- Compare the Predicted values to the Actual Values in the Validation partition
- Use different methods to decide which is the “best” model

Model Selection

Misclassification Rate



Dependent on that Cutoff Point!

Model Selection

Confusion Matrix

Confusion Matrix

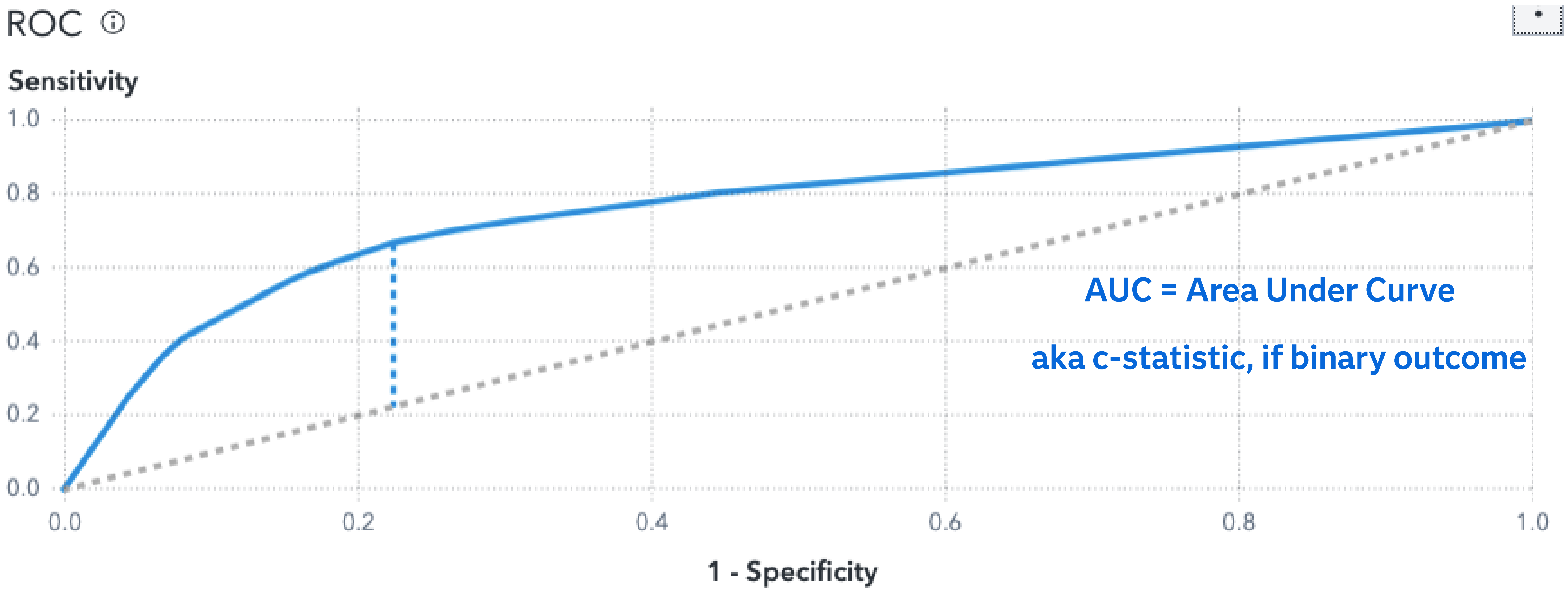
Observed \ Predicted	Alive	Dead
Alive	True Positive 2,686	False Negative 532
Dead	False Positive 821	True Negative 1,170

- **Sensitivity: True Positive / Actually Positive = .8347**
- **Specificity: True Negative / Actually Negative = .5876**

Dependent on that Cutoff Point!

Model Selection

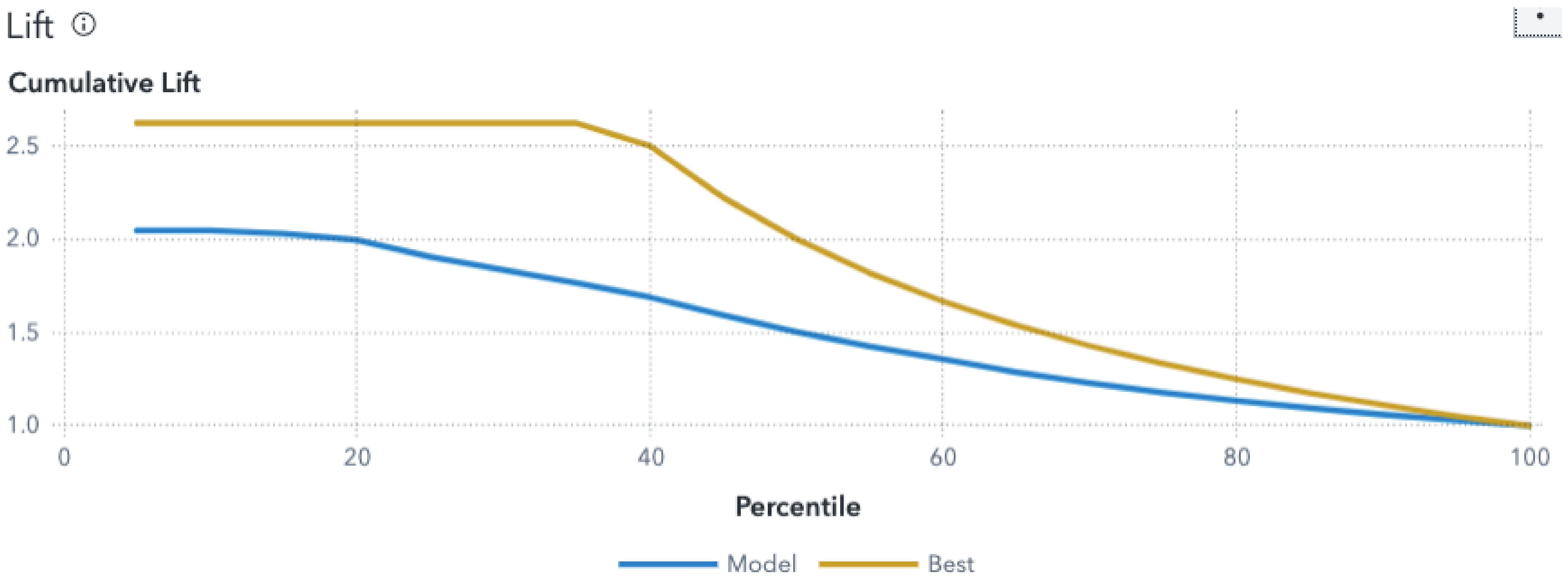
ROC Curves



Looks across all possible cutoff points

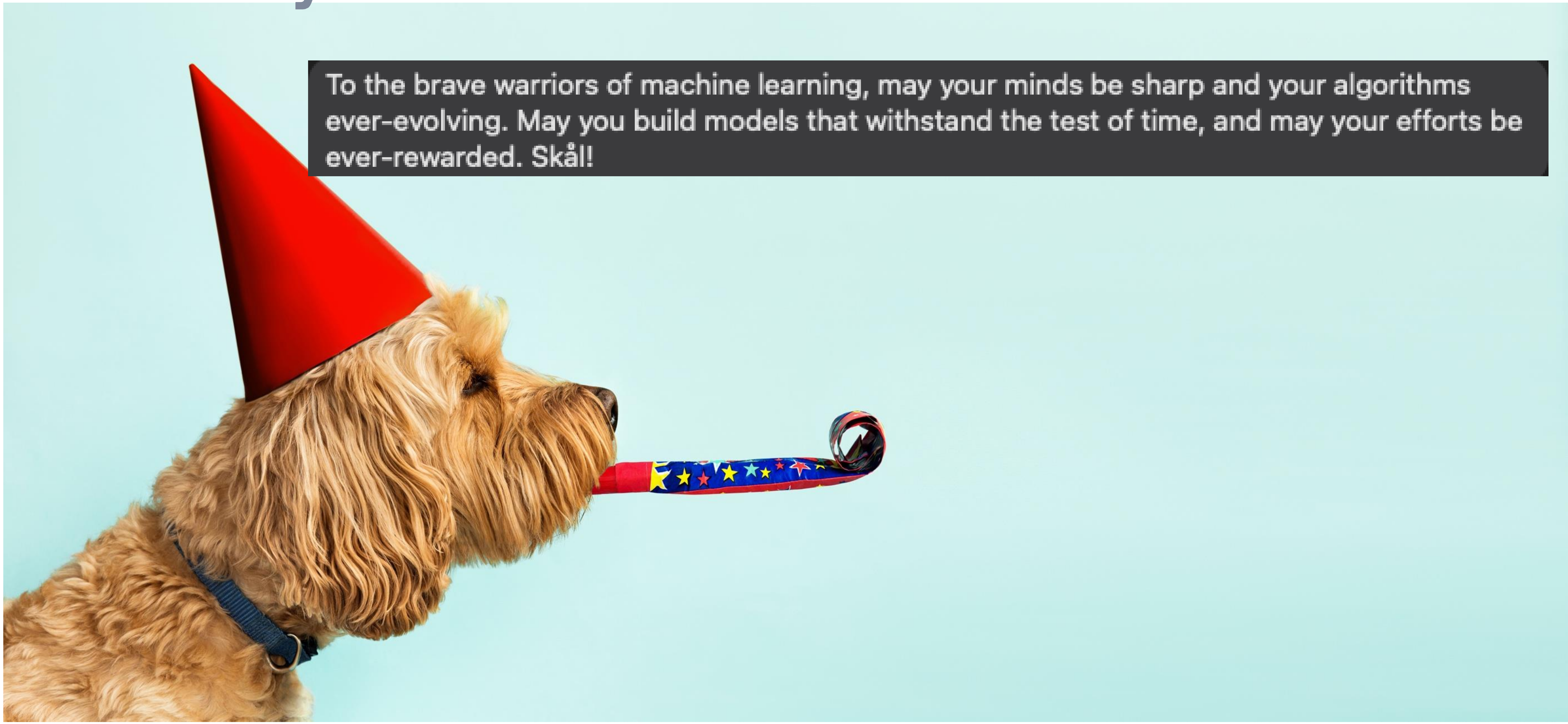
Model Selection

Lift Curves



Independent of cutoff points

Summary


To the brave warriors of machine learning, may your minds be sharp and your algorithms ever-evolving. May you build models that withstand the test of time, and may your efforts be ever-rewarded. Skål!

Thanks!

Jim Box
SAS Institute
Jim.Box@sas.com