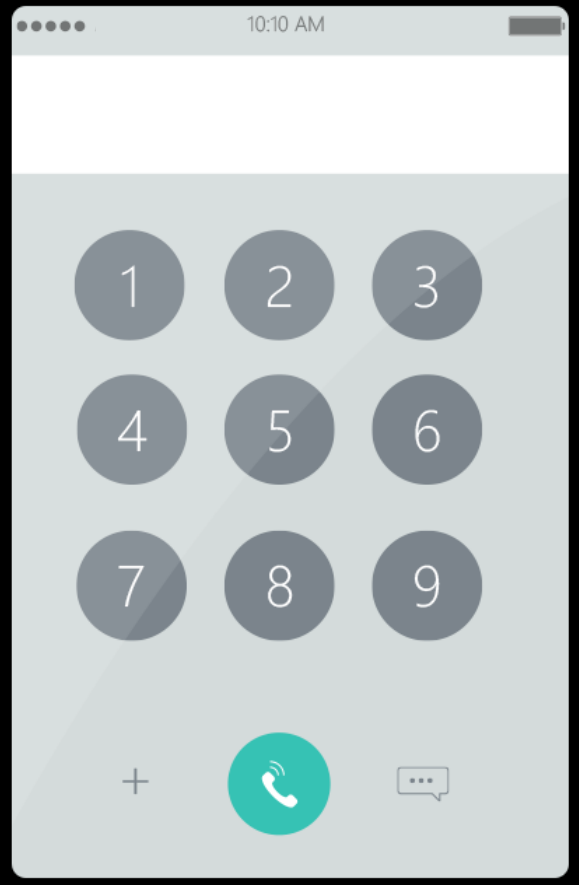
An Introduction to Machine Learning

Blaine Sundrud

Agenda

- Define key terms and concepts
- Explain the machine learning (ML) pipeline
- Discuss using the ML pipeline to solve a real-world business problem



Agent 01
Customer care for returns

Agent 02
Customer care for Kindle

Agent 03
Customer care for ….

ML Problem Framing

Is ML the right solution for the business problem?

Business Problem: How to route customers to agent with right skill?

?

Machine Learning

Call center example

- What do these skills represent?
- How much overlap between skills?
- Can skills be combined?
- What happens for incorrect routes?
- Could the question still be answered?

ML Pipeline

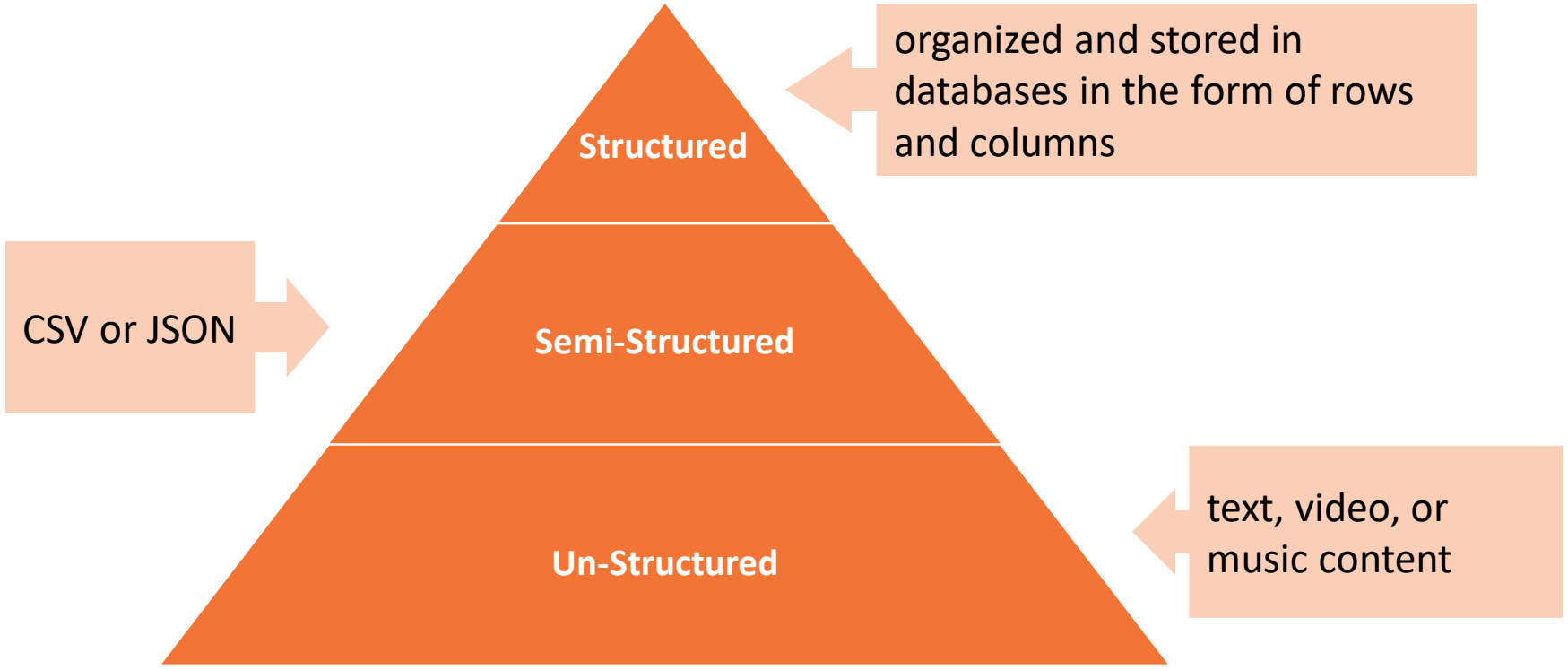

Data Collection & Integration
Data Preparation
Data Visualization & Analysis
Feature Selection & Engineering
Model Training
Model Evaluation
Prediction

Three ML problems

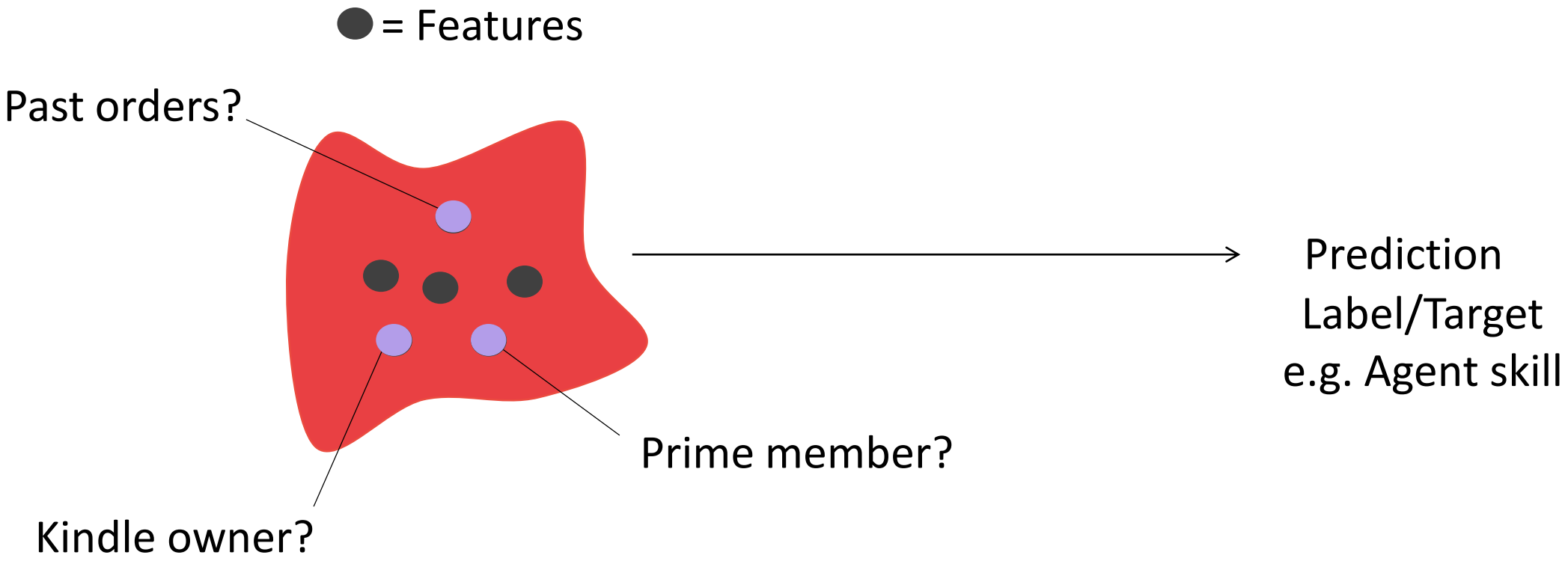
1. Binary: Two groups
2. Multi-class: More than two groups
3. Regression: Continuous vales

Types of data

There are three types of data.

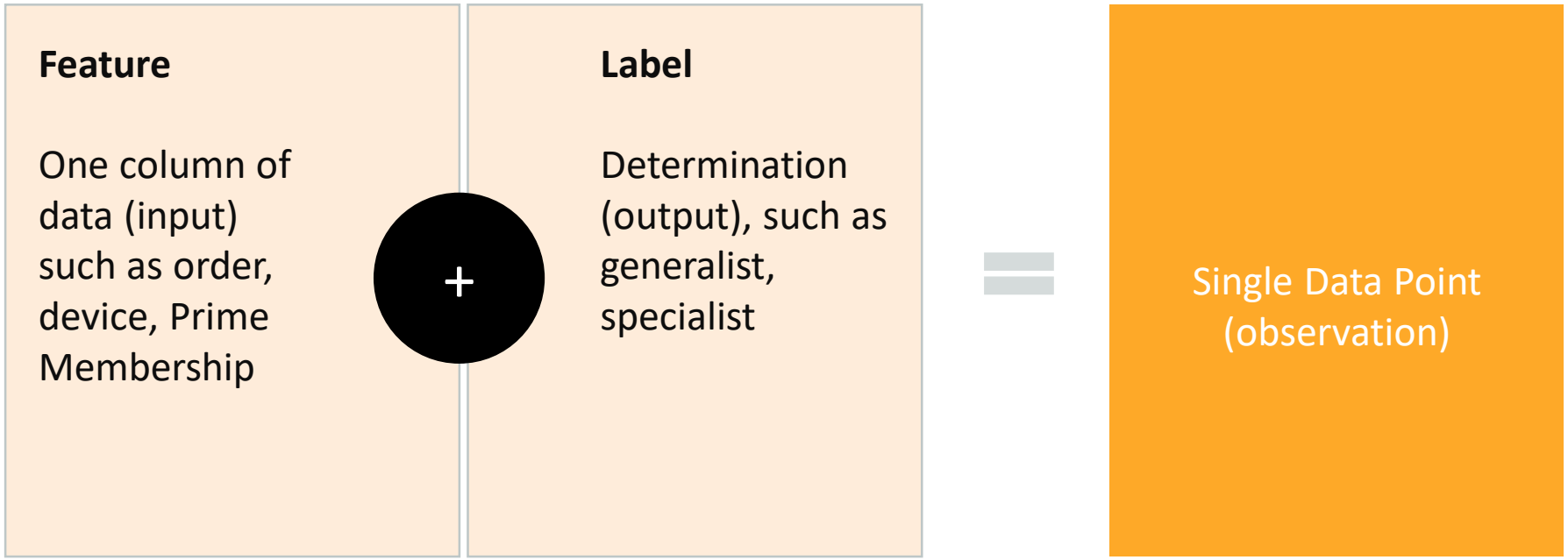


From features to prediction

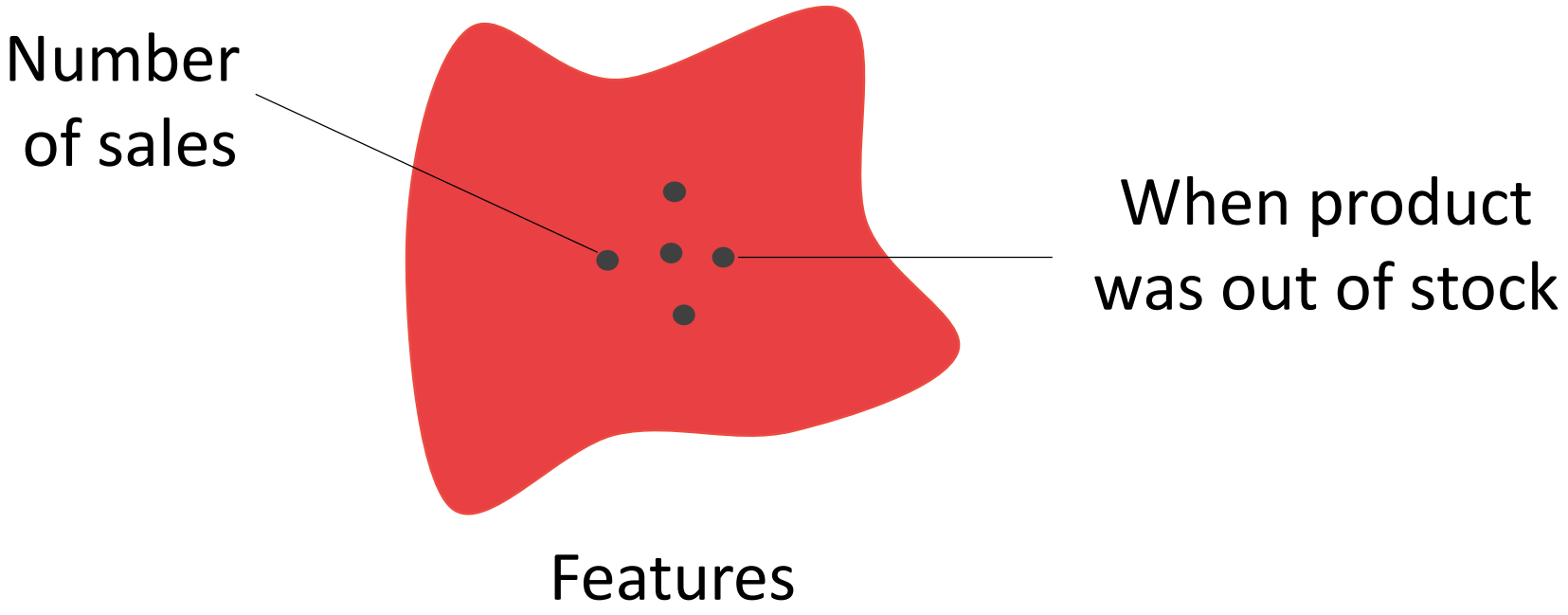


Single data point

Together, the features and the labels make up a single data point.

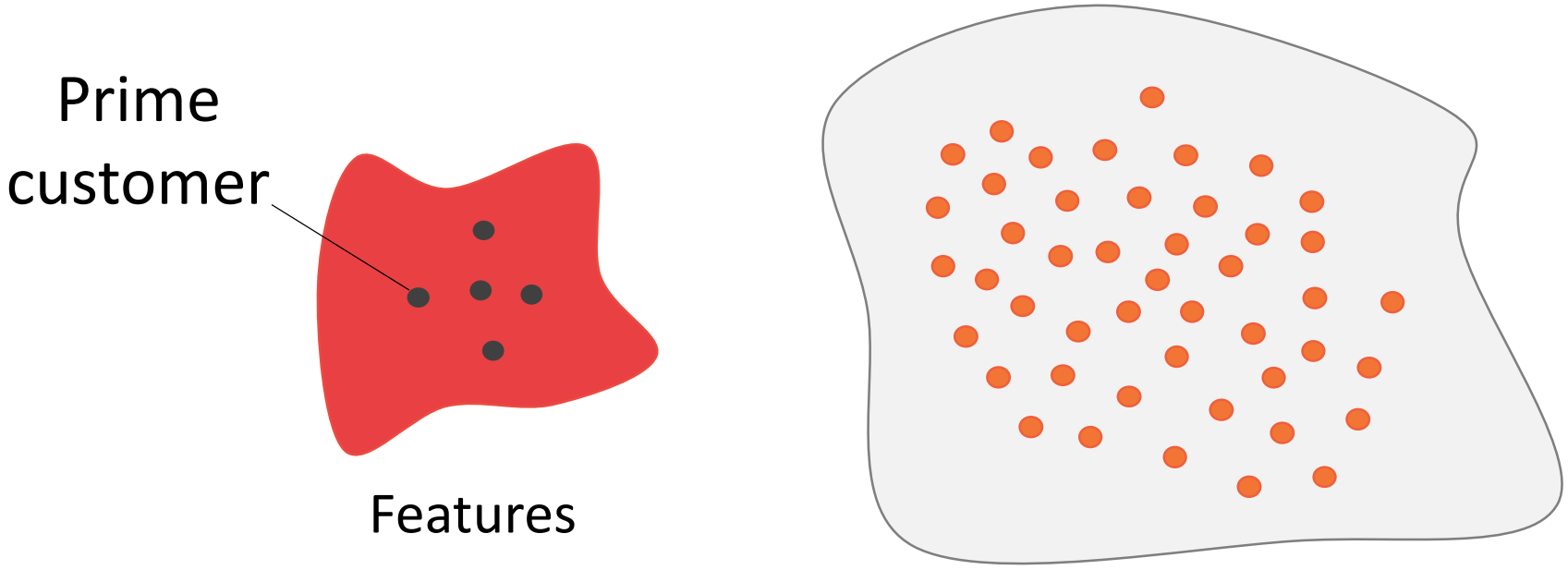


Representative features



Ratio of features to data points

A general rule of thumb here is you should have 10 times the number of data points as features.



Data Preparation

Role in the data prep phase

Your job in the data prep phase is to *manually* and critically explore your data.

- What features are there?
- Does it match expectations?
- Is there enough information to make accurate prediction?

Role in the data prep phase

Your job in the data prep phase is to *manually* and critically explore your data.

Confirm all labels are relevant to the ML problem.

- What features are there?
- Does it match expectations?
- Is there enough information to make accurate prediction?

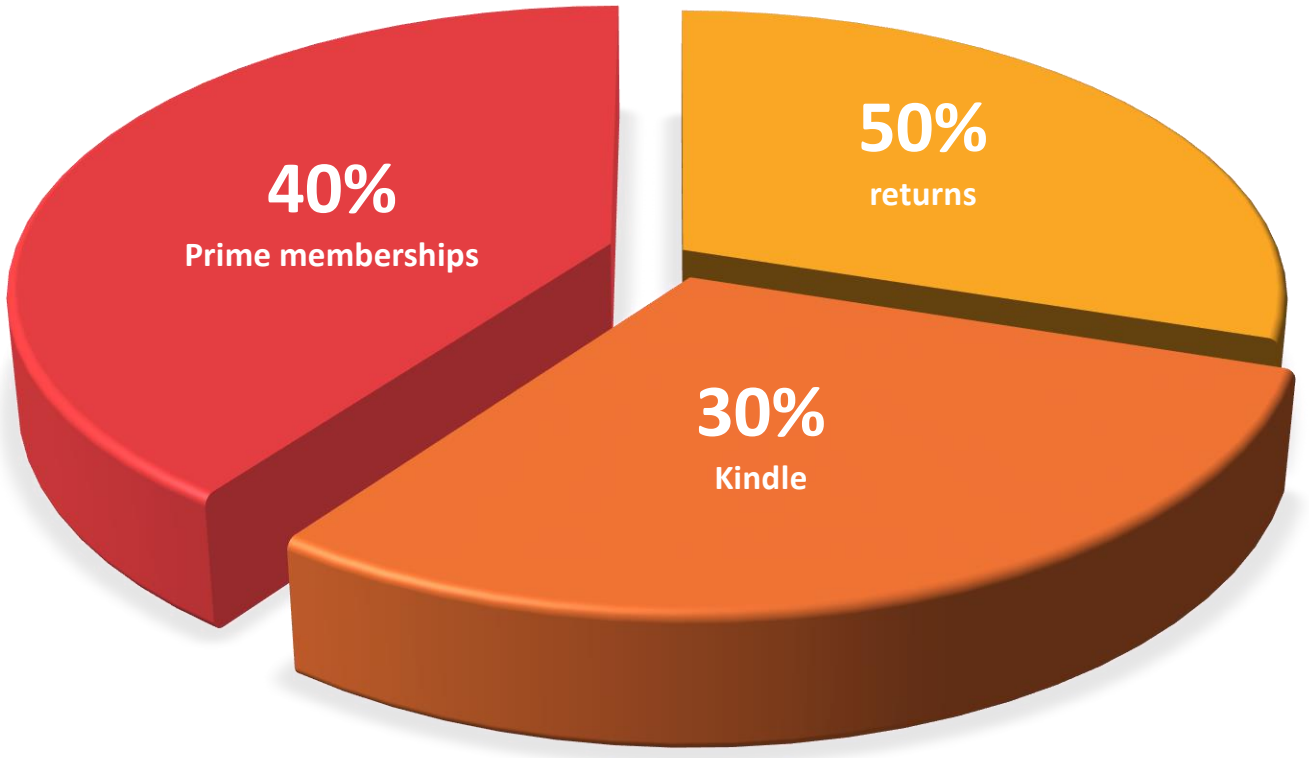
- Should any labels be excluded?
- Are any labels not entirely accurate?
- What skills?
- Are there similar skills?
- Can skills be combined?
- What happens for incorrect routes?
- Could agent answer questions from incorrect routes?

Data Visualization & Analysis

Choosing the right algorithm

1. Supervised
2. Unsupervised
3. Reinforcement
4. Deep Learning

A programmatic analysis

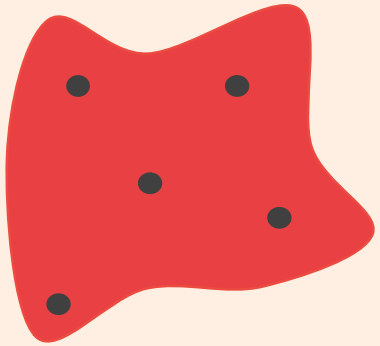


*Percentages are greater than 100% because some callers called about more than one issue.

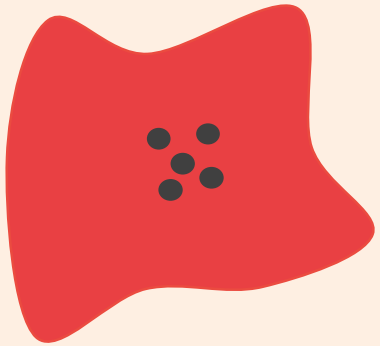
Data Visualization & Analysis

Supervised algorithms

Input/output relationship is **known.**



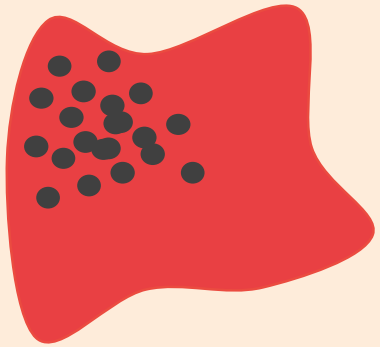
➡ Credit card fraud



➡ Voter fraud

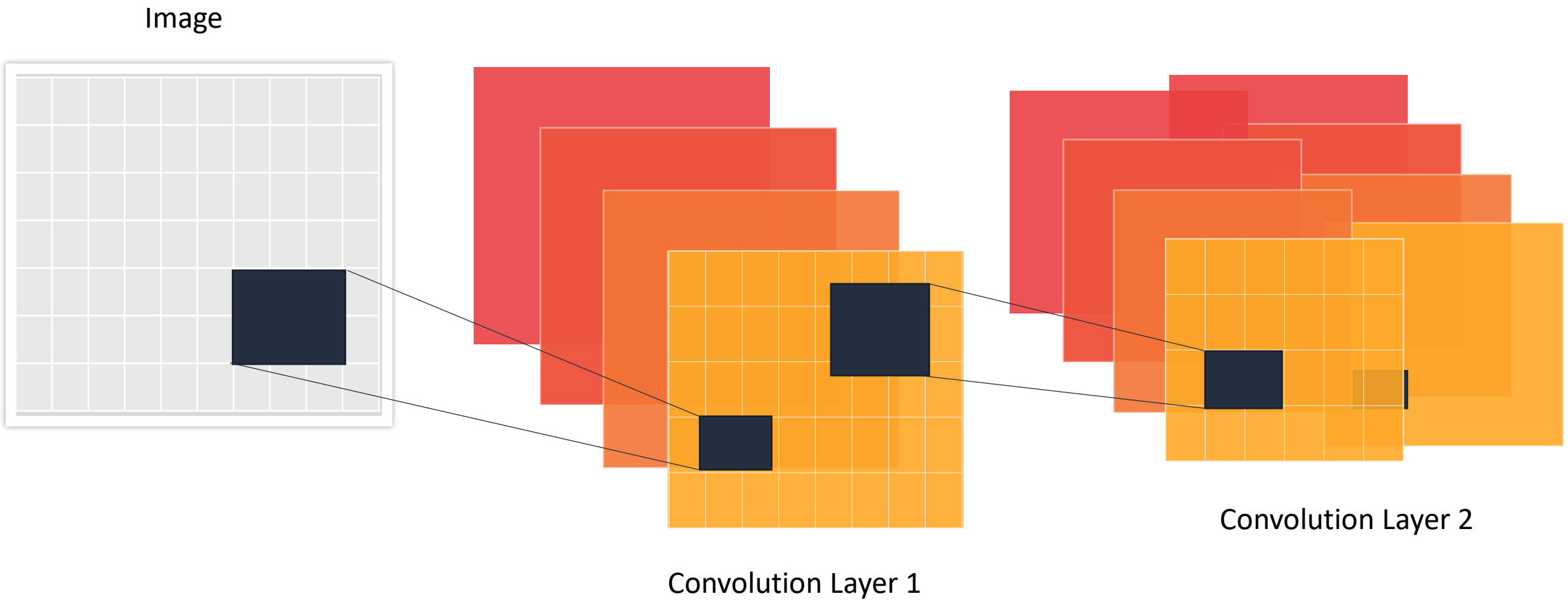
Unsupervised algorithms

Input/output relationship is **unknown.**



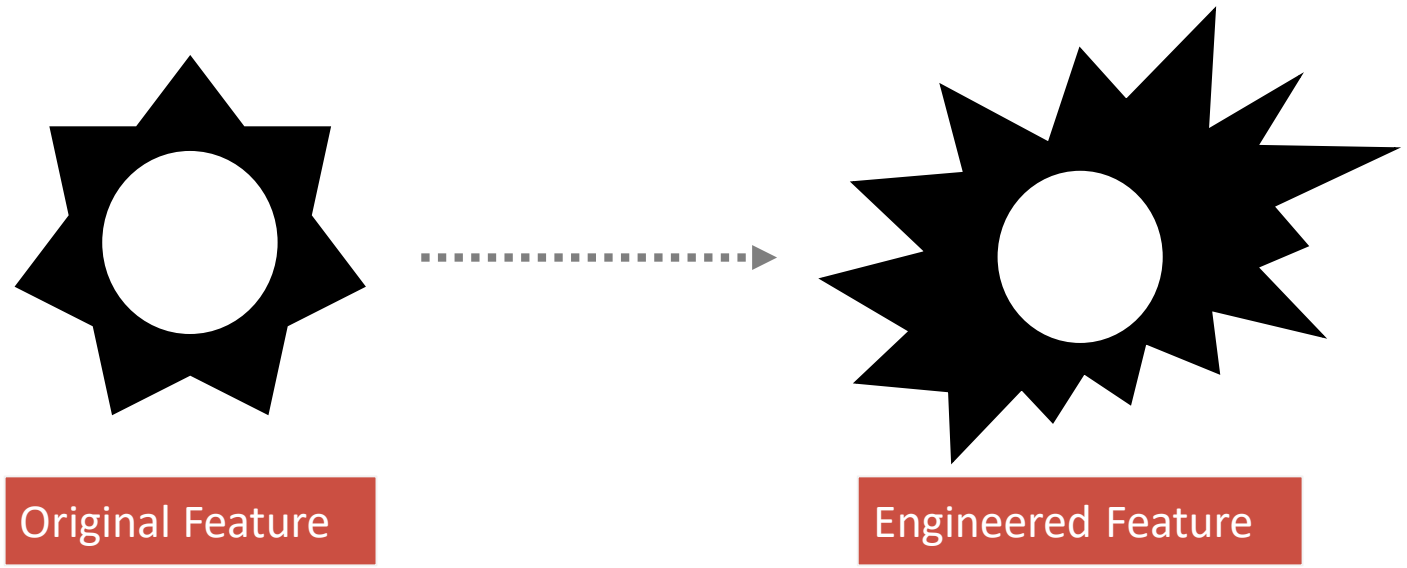
➡ Large order from suspicious address

Invention of convolutional neural networks



Feature Selection & Engineering

Feature engineering



Original Feature

Engineered Feature

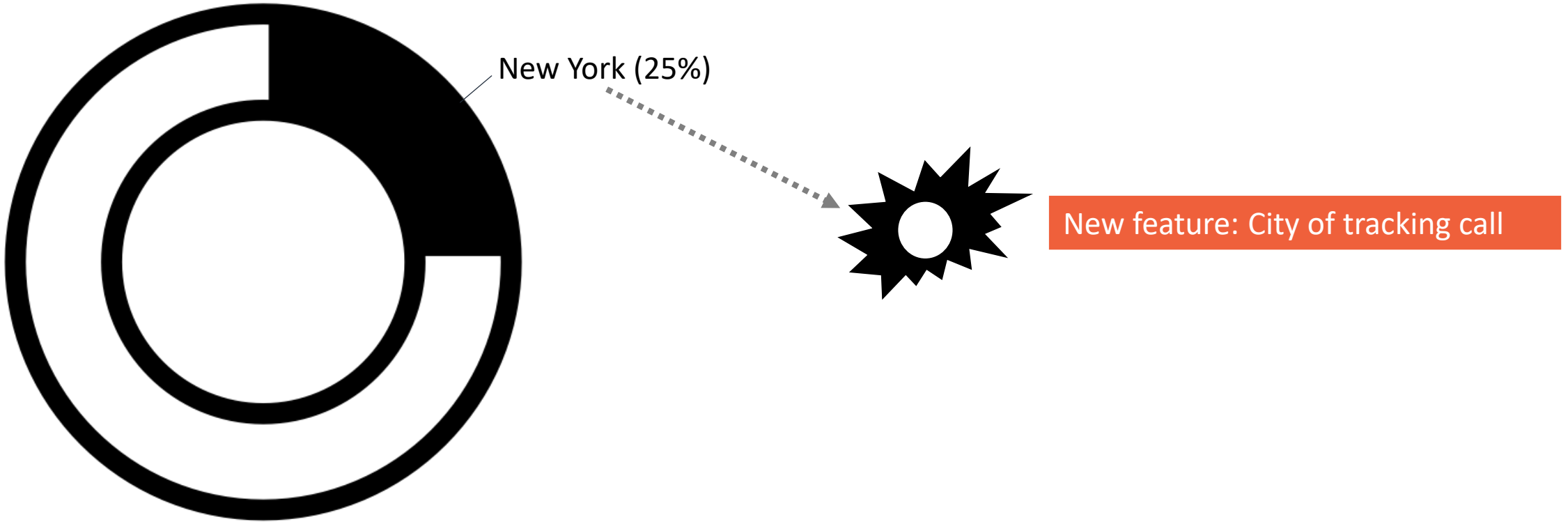
Helps to answer questions like:

Do these features make sense for my prediction?

How can I engineer features based on visualizations?

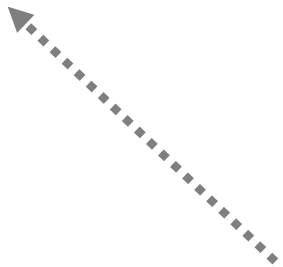
Feature engineering by visualizing data

Location of customers calling about tracking



Feature engineering from our use case

Most recent order	Date/Time of most recent order	Owns a Kindle
hat	01/13/2018, 1PM	yes

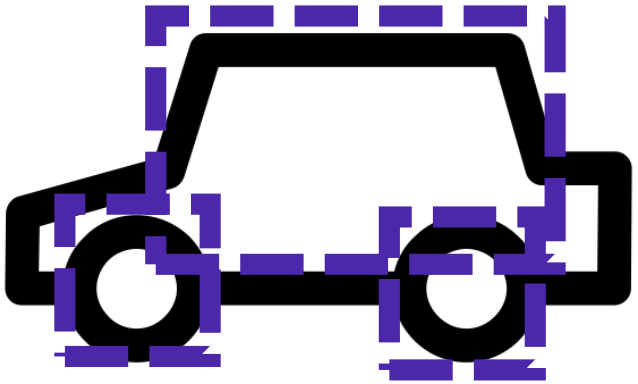


Days since last order
72 days

Feature engineering in image classification

Raw Image

High level features



= edges

Model Training

A helpful data analysis tool

https://scikit-learn.org/stable/

Types of hyperparameter tuning

- Loss function
- Regularization
- Learning parameters

https://scikit-learn.org/stable/

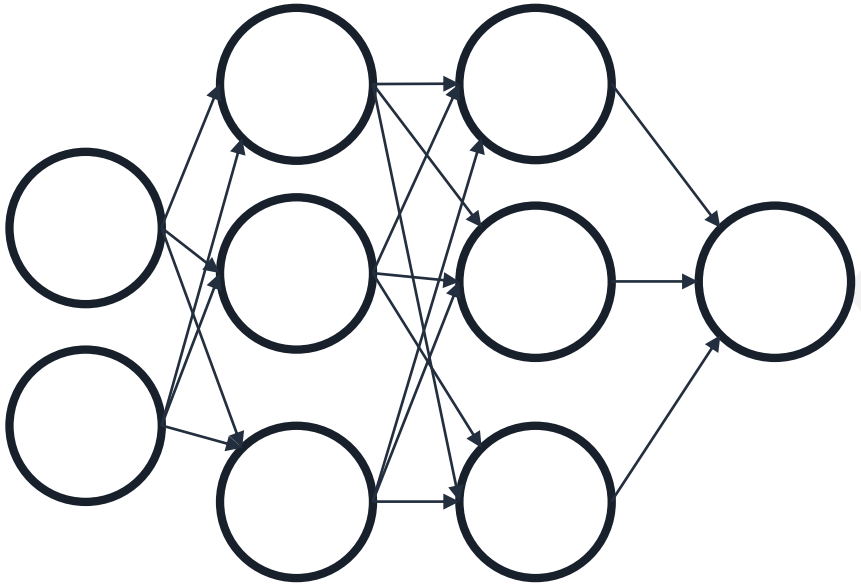
Bias

The gap between predicted value and actual value

Variance

How dispersed your predicted values are

Hyperparameter



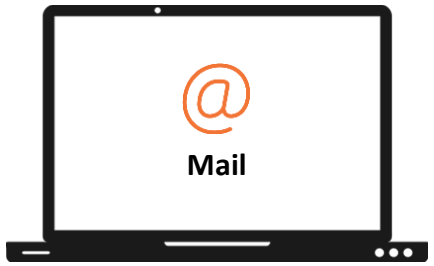
Parameter

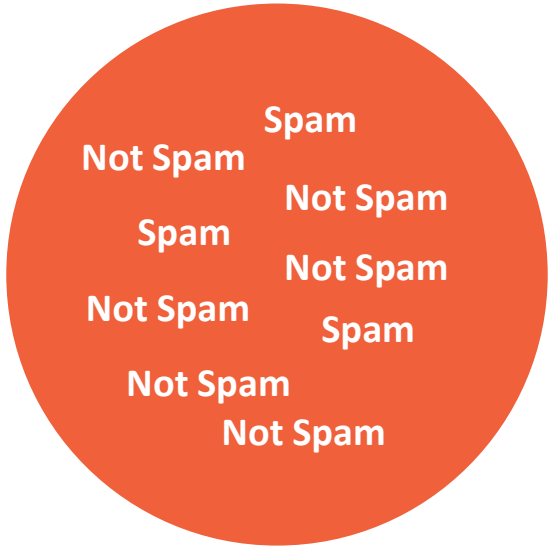
$$\bigcirc = \sum_{i=1}^{3} W_i X_i$$

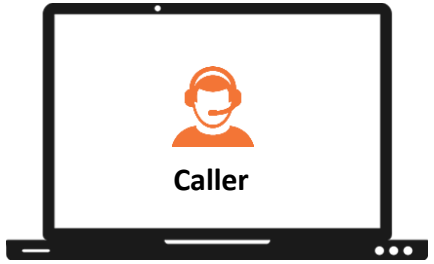
Hyperparameter

How quickly the model learns the weights

Label unknown?







Model Evaluation

Accuracy and precision

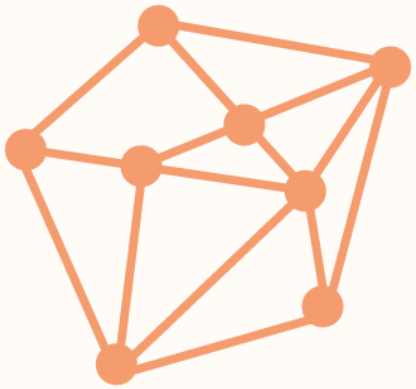
Accuracy $= \dfrac{\text{Correct Predictions}}{\text{Total \# Predictions}}$

Precision $= \dfrac{\text{True Positives}}{\text{True Positives + False Negatives}}$

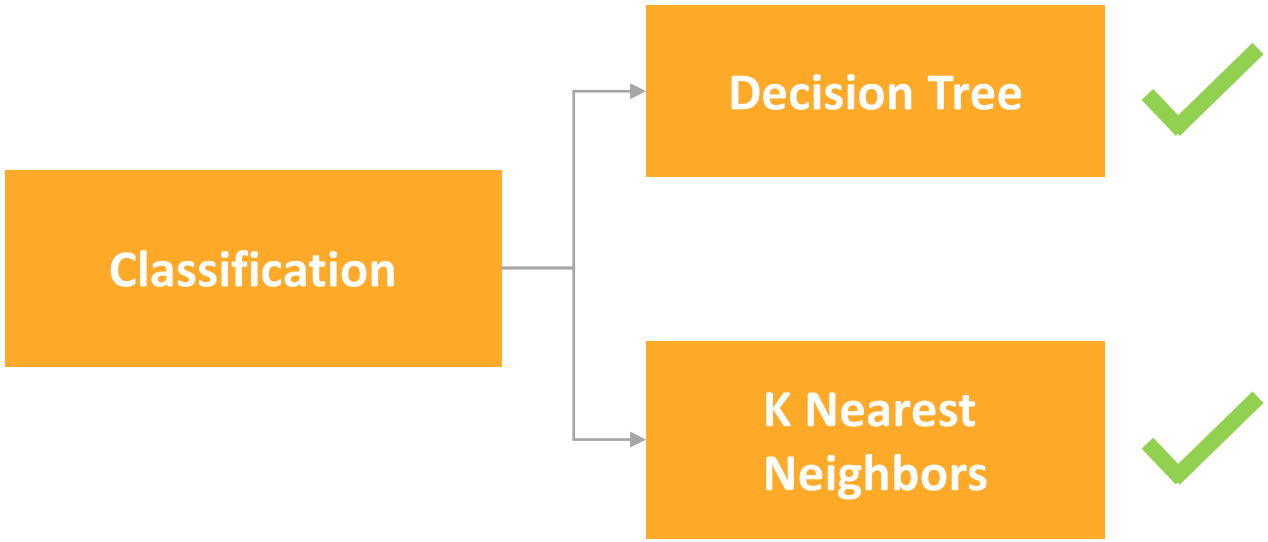
Compare your algorithm to others in its class

Supervised	Unsupervised	Reinforcement
• Regression analysis • Decision trees • K-nearest neighbors • Neural networks	• K-means clustering • Anamoly detection • Neural networks	• Q-learning • SARSA

See how the model does with other algorithms



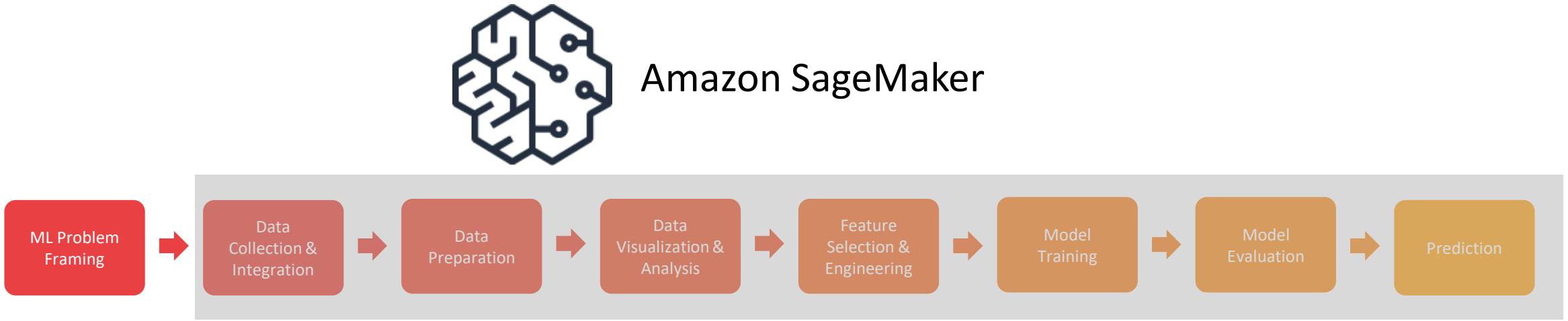
Supervised algorithm



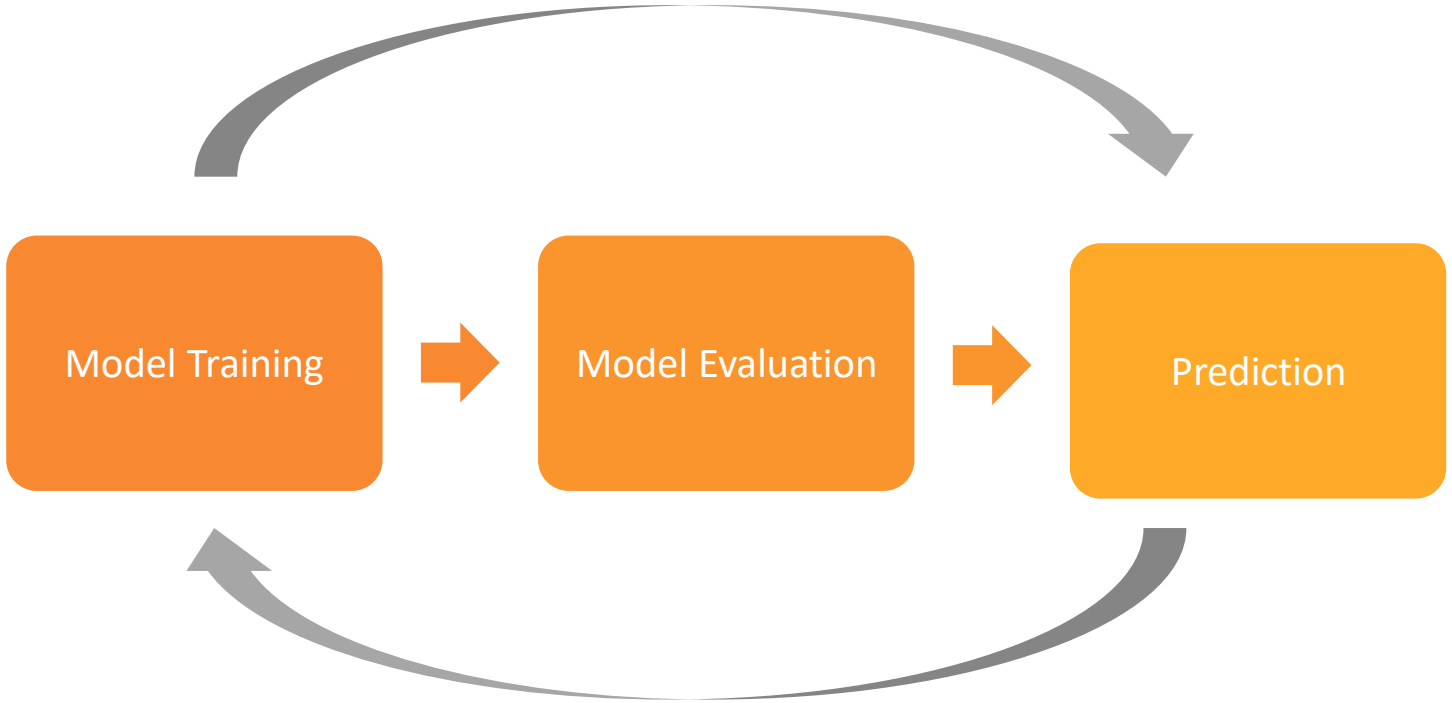
Prediction


ML Problem Framing
Data Collection & Integration
Data Preparation
Data Visualization & Analysis
Feature Selection & Engineering
Model Training
Model Evaluation
Prediction

Amazon SageMaker

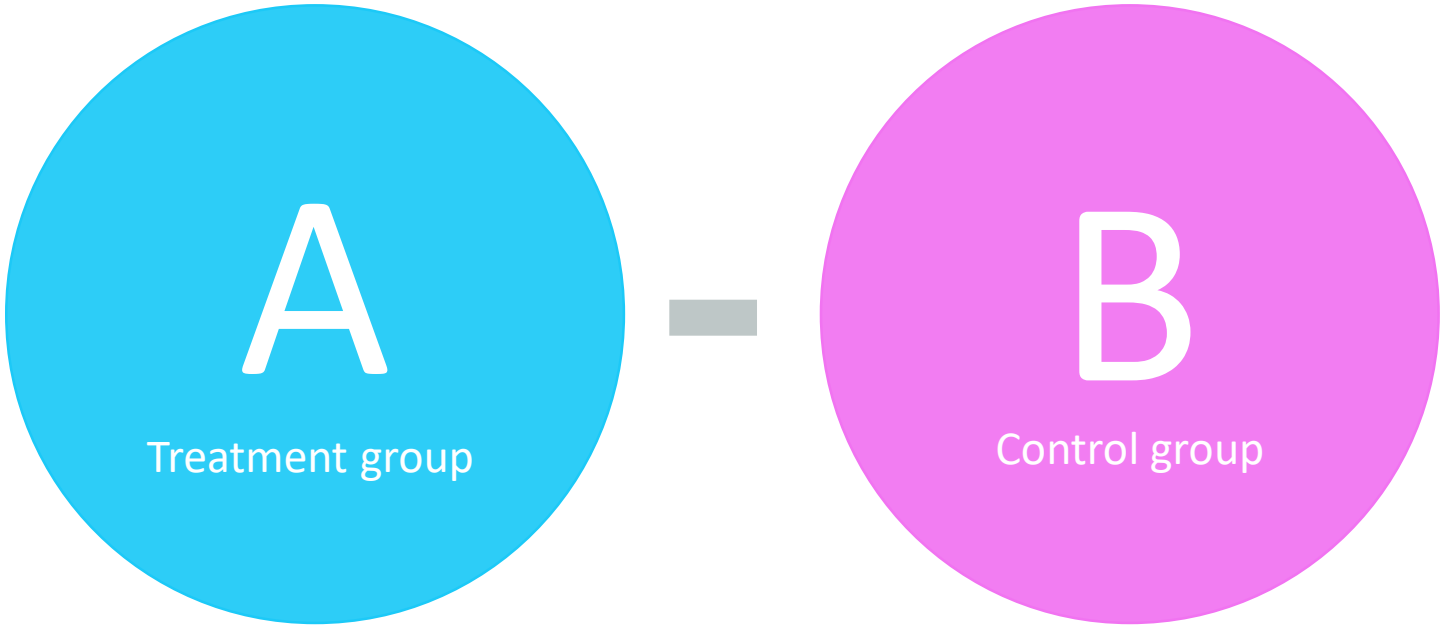


Model production data and re-train



Pre-check

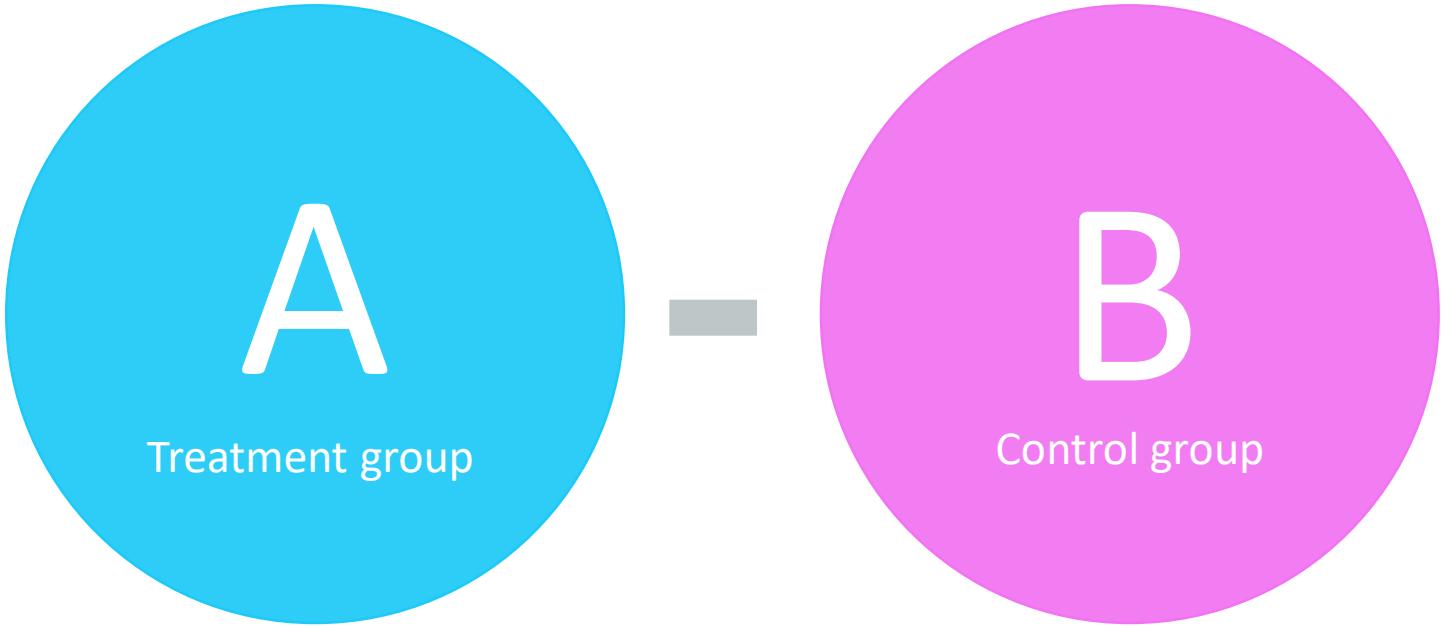
Make sure your new ML solution is compared against your existing baseline in a fair manner



Is the difference significant?

Amazon’s intelligent routing solution

Was based on a simple classification task



Is the difference significant?

20%

reduction