# Preparing Data for Machine Learning

#### UNDERSTANDING THE NEED FOR DATA PREPARATION



Janani Ravi CO-FOUNDER, LOONYCORN www.loonycorn.com

#### Overview

Need for data preparation in machine learning

Insufficient data

Excessive or overly complex data

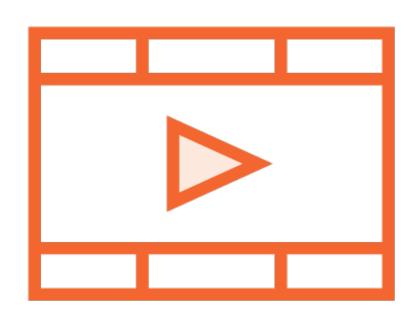
Non-representative data, missing data, outliers

Oversampling and undersampling

Overfitting and underfitting models

# Prerequisites and Course Outline

#### Prerequisites

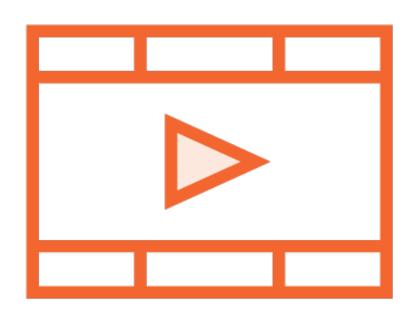


**Basic Python programming** 

Basic understanding of the machine learning workflow

Built and trained simple machine learning models

#### Prerequisites



Python Fundamentals

Understanding Machine Learning

Building Your First scikit-learn Solution

#### Course Outline



Need for data preparation

Data cleaning and transformation

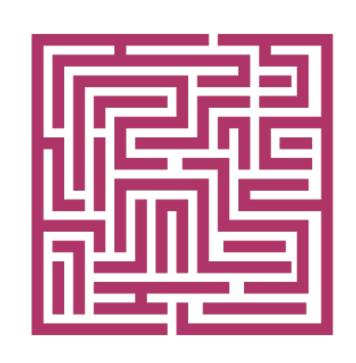
Continuous and categorical Data

Understanding feature selection

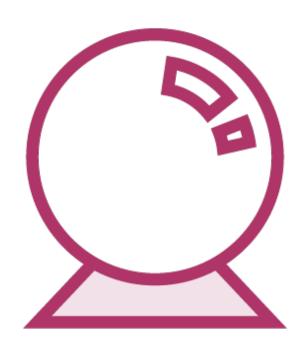
Implementing feature selection

# The Need for Data Preparation

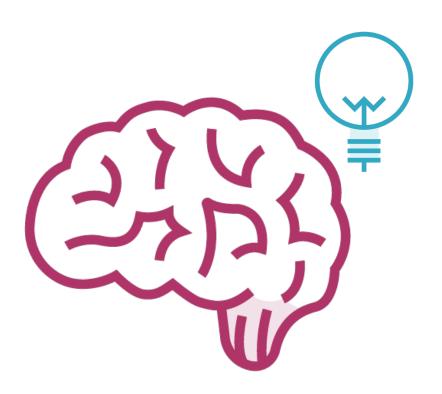
# Machine Learning







Find patterns

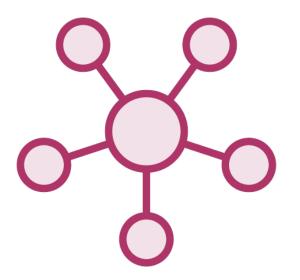


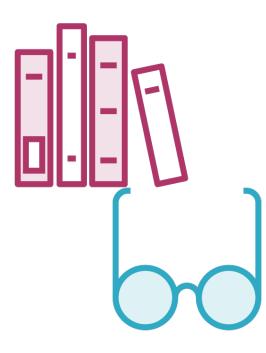
Make intelligent decisions

# Types of Machine Learning Problems









Classification

Regression

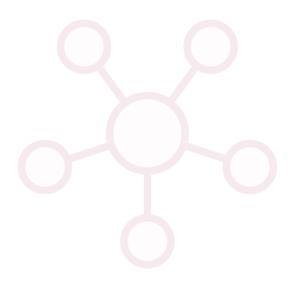
Clustering

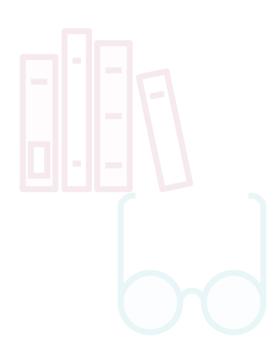
Dimensionality Reduction

## Types of Machine Learning Problems









Classification

Regression

Clustering

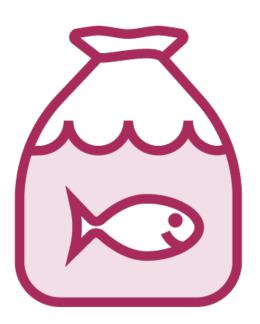
**Dimensionality Reduction** 

#### Whales: Fish or Mammals?



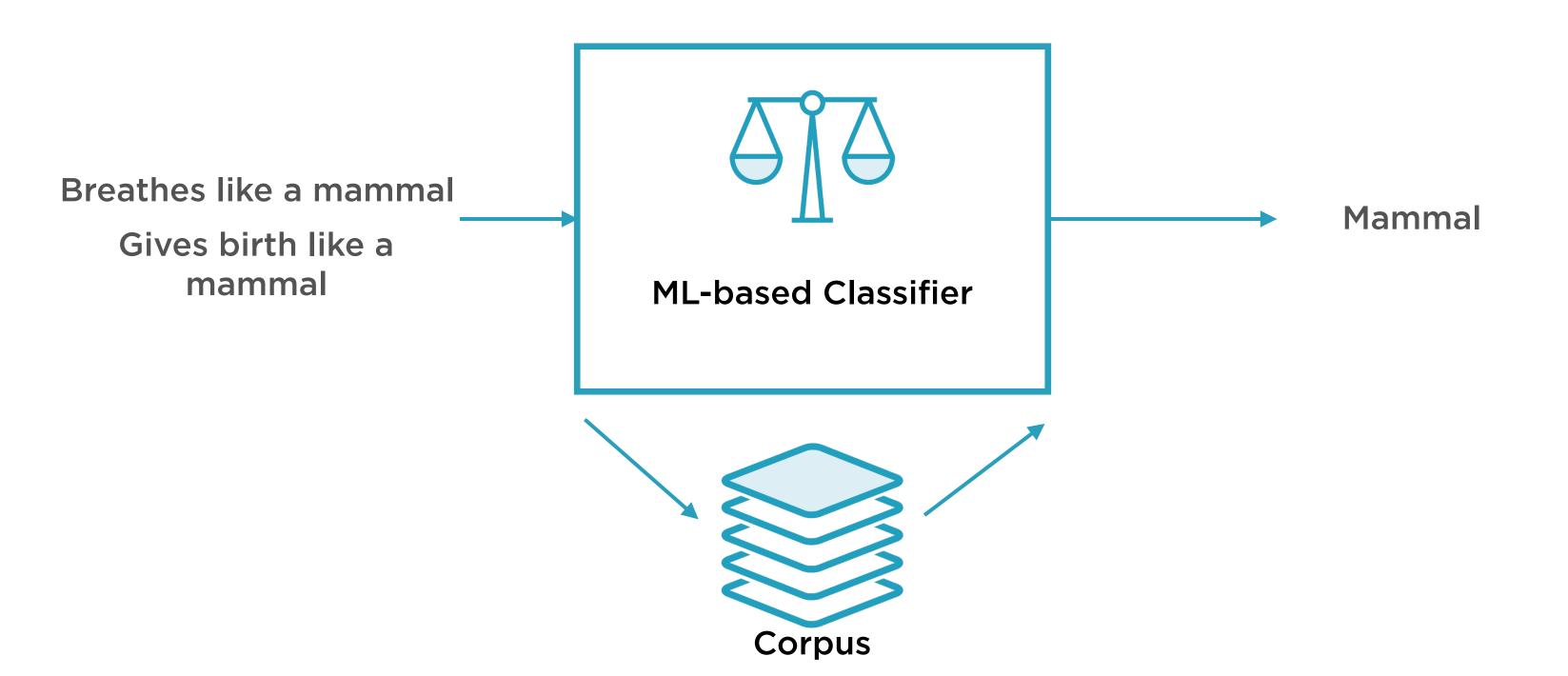
**Mammals** 

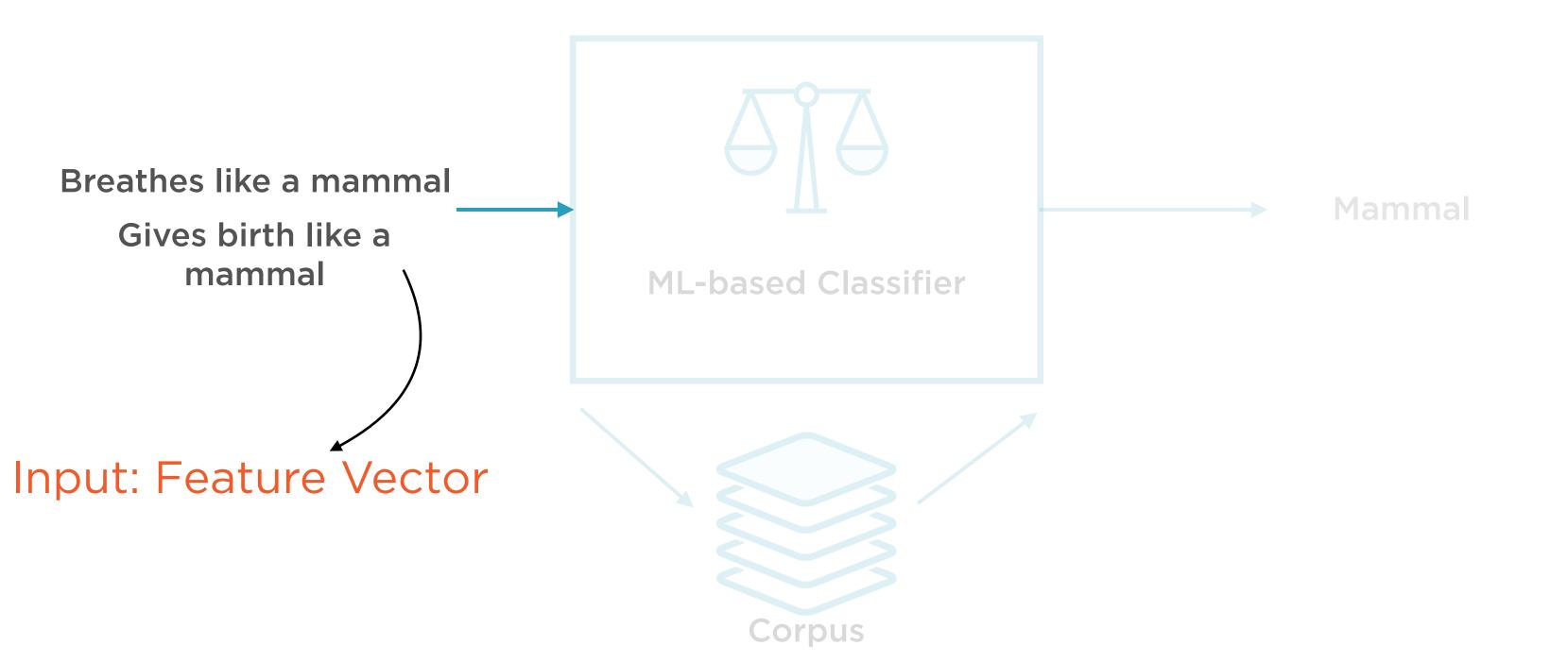
Members of the infraorder Cetacea

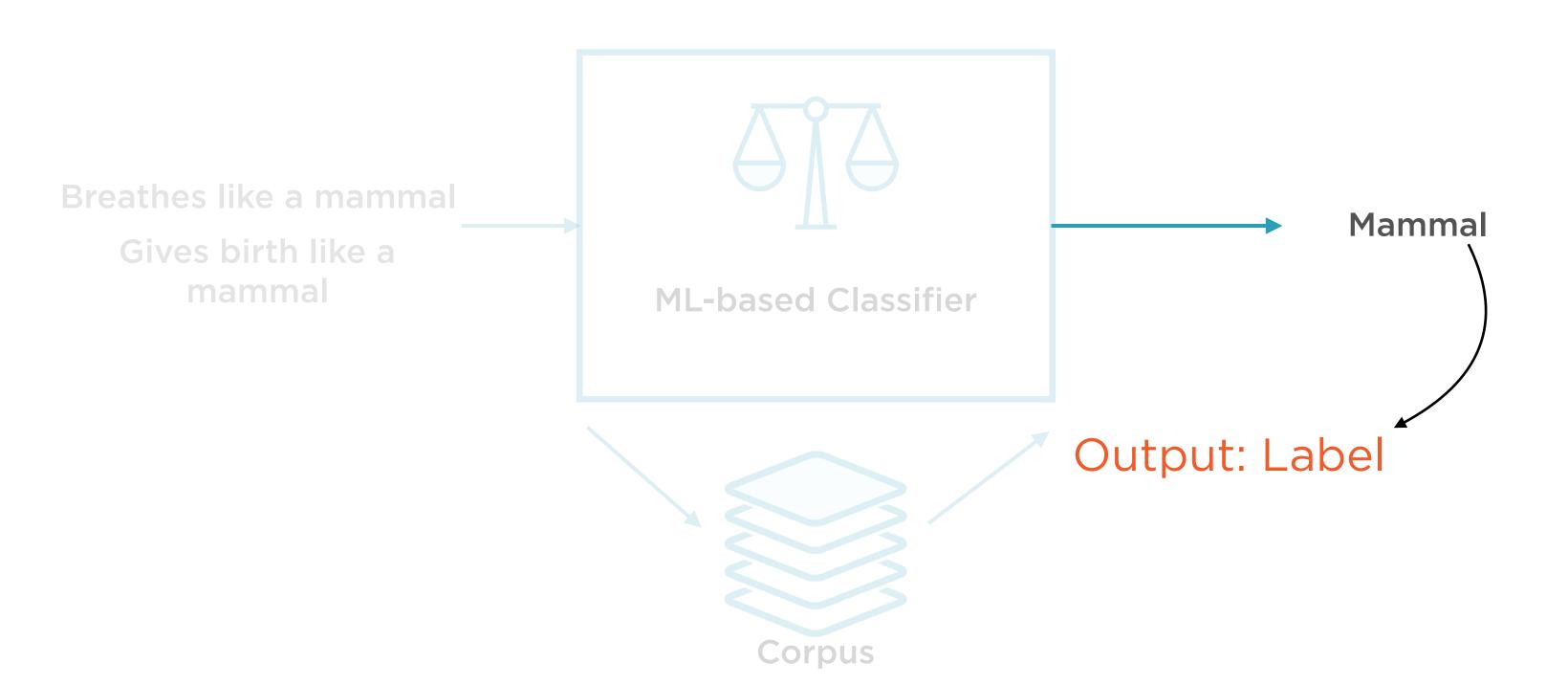


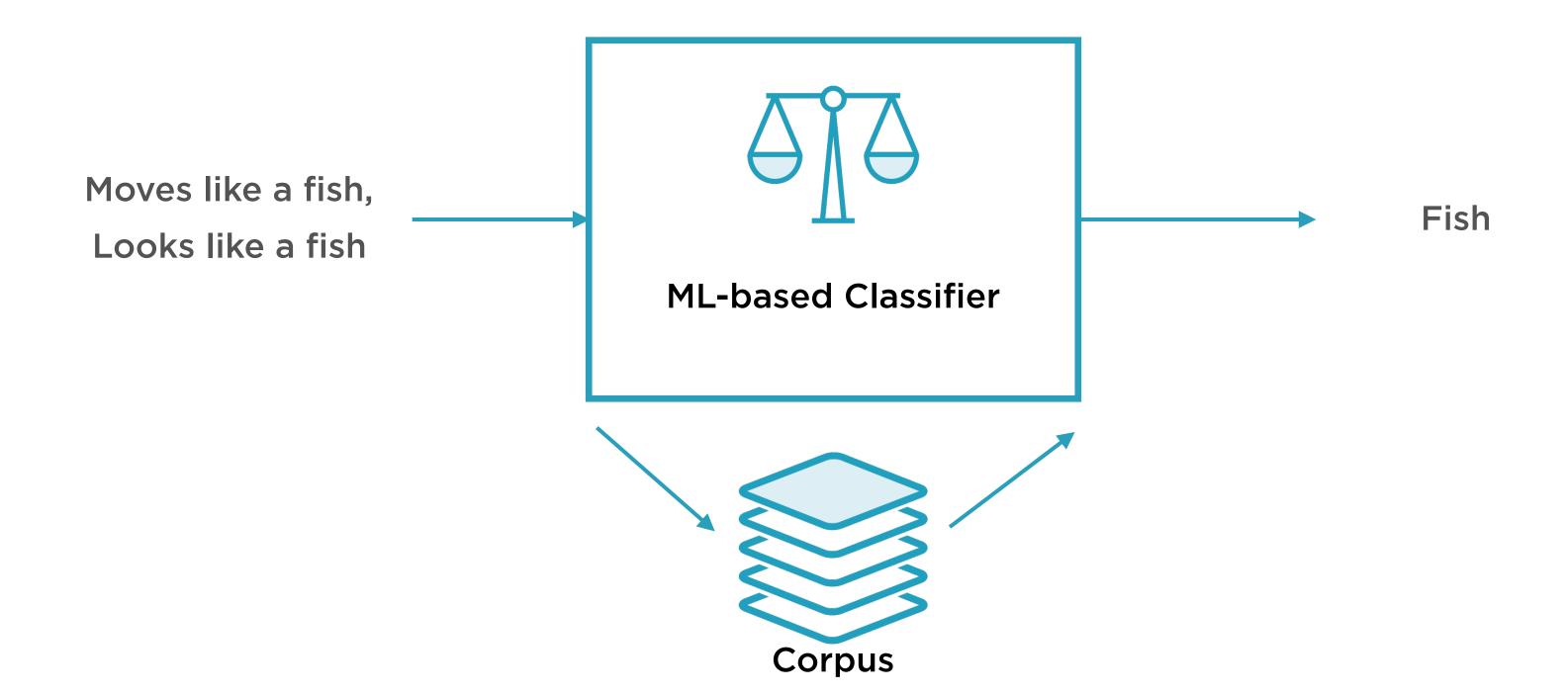
Fish

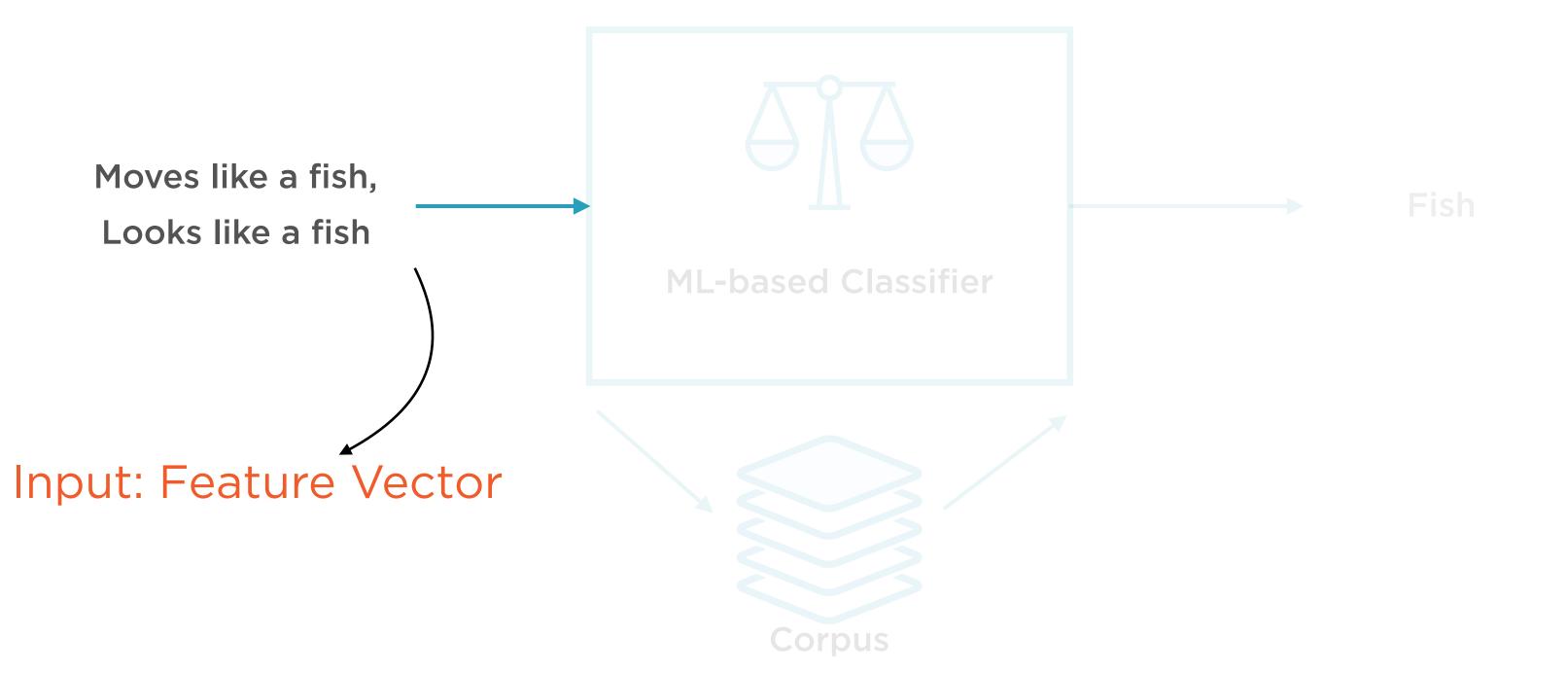
Look like fish, swim like fish, move with fish

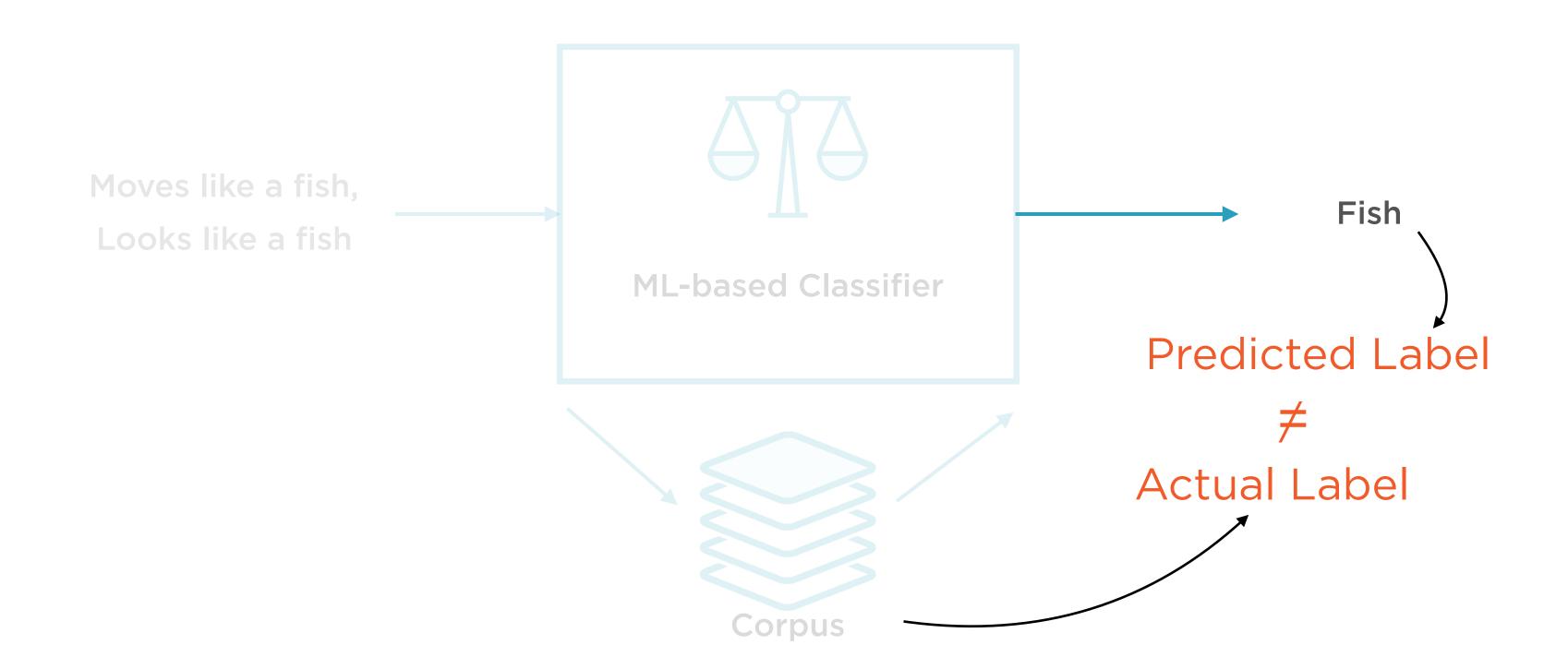












Garbage In, Garbage Out
If data fed into an ML model is of
poor quality, the model will be of
poor quality

#### Problems with Data

Insufficient data

Too much data

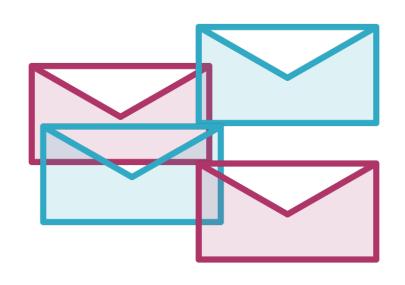
Non-representative data

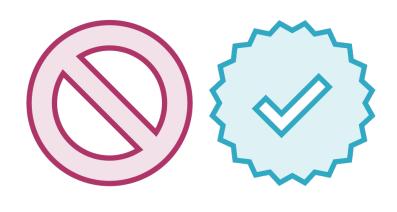
Missing data

Duplicate data

Outliers

# Machine Learning





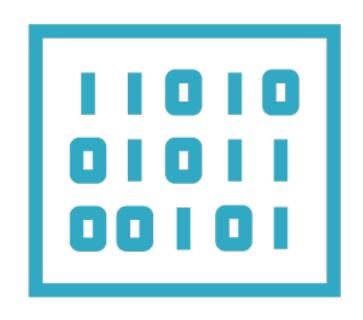


Emails on a server

Spam or Ham?

Trash or Inbox

# Machine Learning







Images represented as pixels

Identify edges, colors, shapes

A photo of a little girl

#### Problems with Data

Insufficient data

Too much data

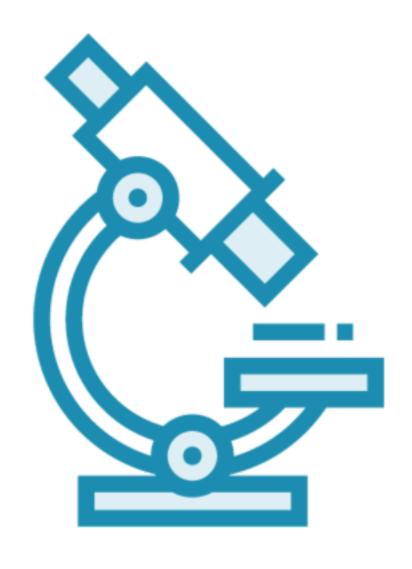
Non-representative data

Missing data

Duplicate data

Outliers

#### Insufficient Data

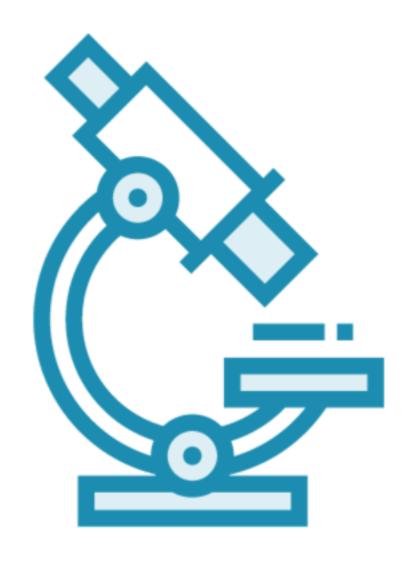


# Models trained with insufficient data perform poorly in prediction

#### Paradoxically leads to either

- Overfitting: Read too much for too little data
- Underfitting: Build overly simplistic model from available data

#### Insufficient Data

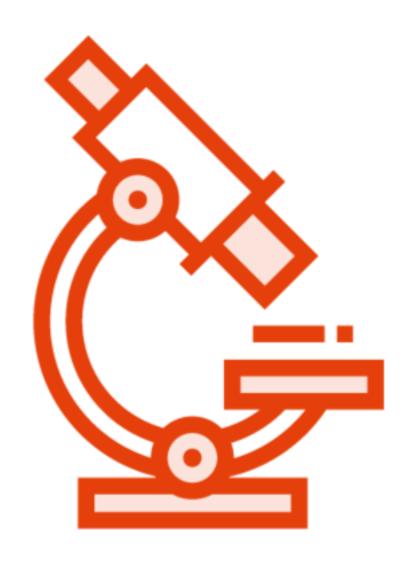


Common struggle for projects in the real world

Relevant data may not be available

Collection process difficult and timeconsuming

#### Insufficient Data



No great solution for insufficient data Simply need to find more data sources

#### Dealing with Small Datasets

Model complexity

Transfer learning

Data augmentation

Synthetic data

#### Dealing with Small Datasets

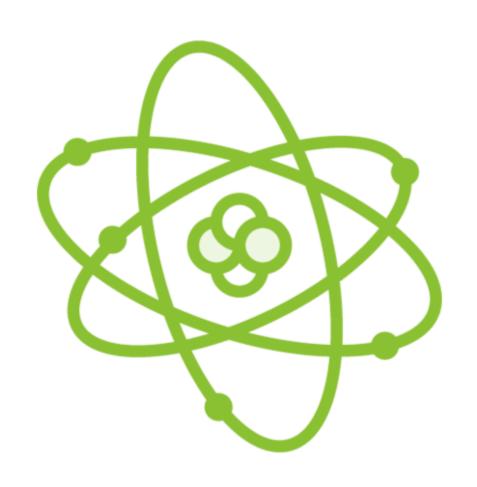
Model complexity

Transfer learning

Data augmentation

Synthetic data

#### Model Complexity



Simpler model with fewer model parameters

Less susceptible to overfitting

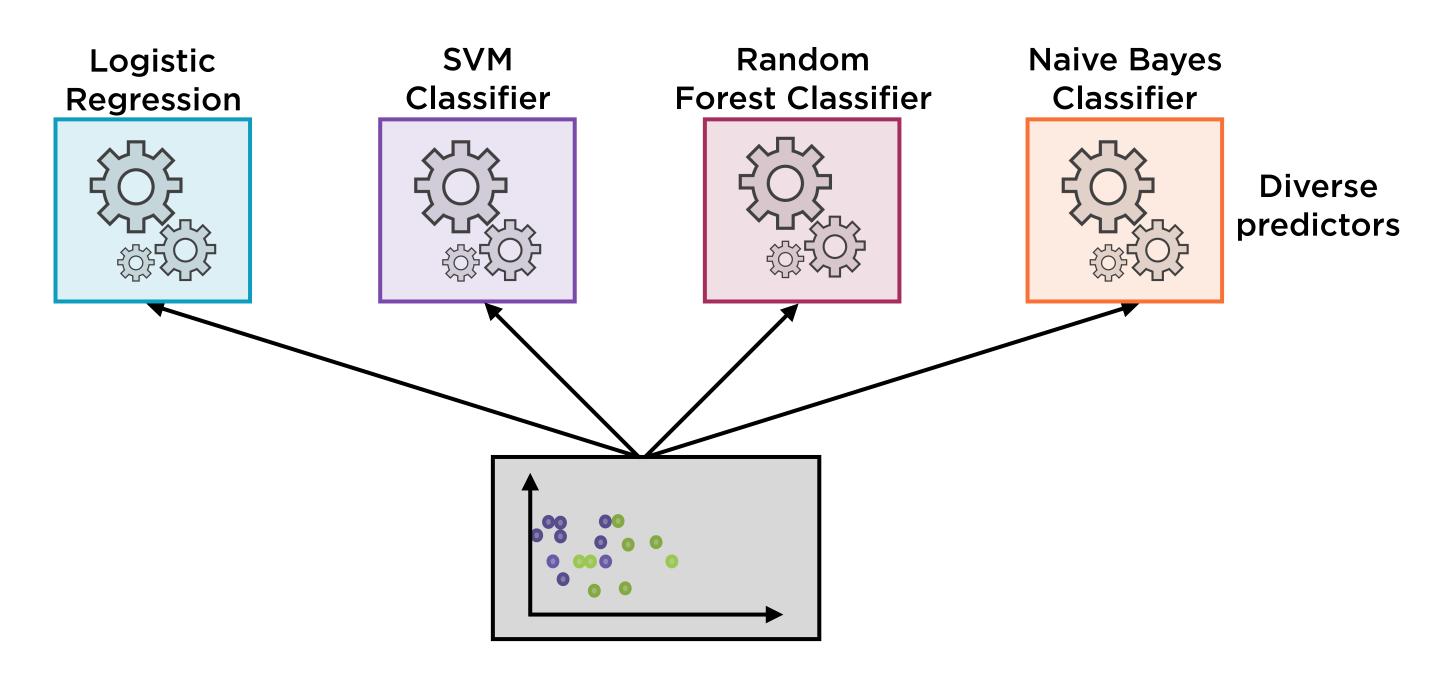
e.g. Naive Bayes classifier, logistic regression

Use ensemble techniques

# Ensemble Learning

Machine learning technique in which several learners are combined to obtain a better performance than any of the learners individually.

# Ensemble Learning



#### Dealing with Small Datasets

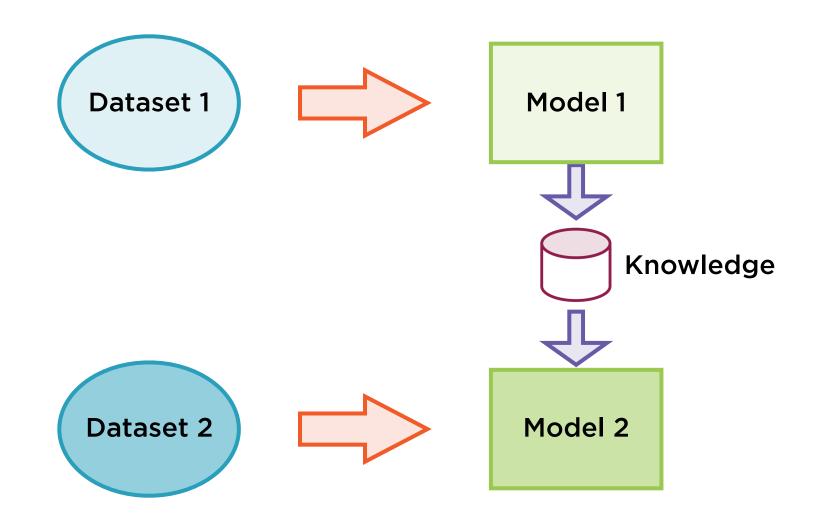
Model complexity

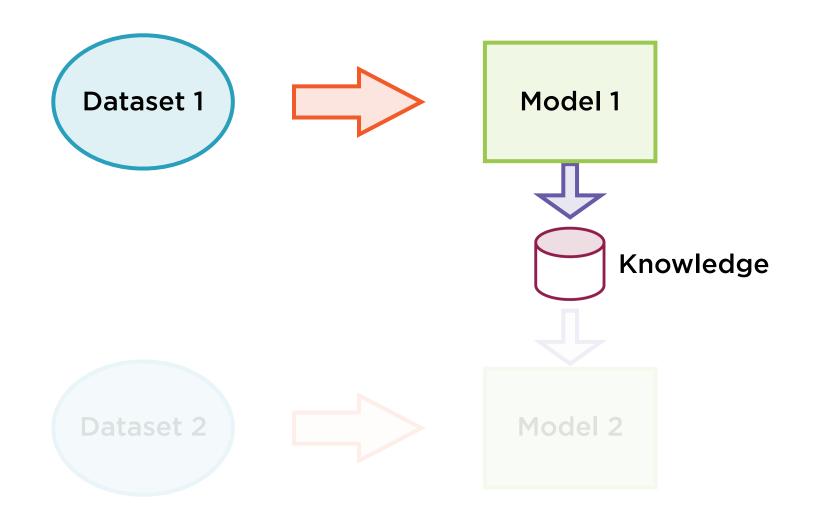
Transfer learning

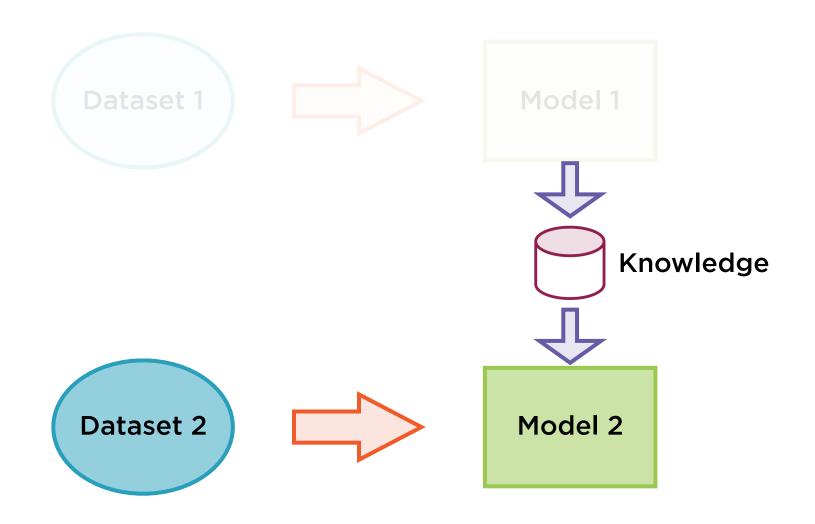
Data augmentation

Synthetic data

The practice of re-using a trained neural network that solves a problem similar to yours, usually leaving the network architecture unchanged and re-using some or all of the model weights.







Transferred knowledge is especially useful when the new dataset is small and not sufficient to train a model from scratch

#### Dealing with Small Datasets

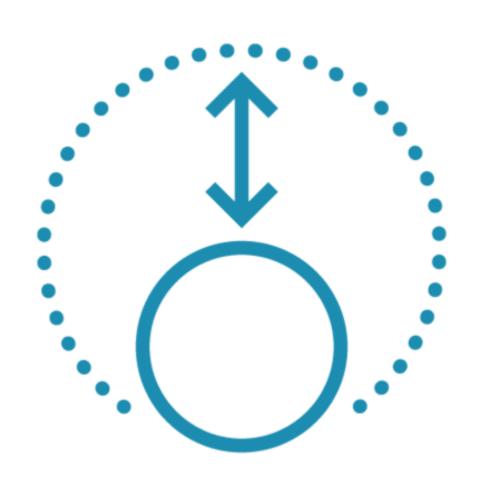
Model complexity

Transfer learning

Data augmentation

Synthetic data

## Data Augmentation



Increase the number of training samples

Perturbed images are a form of data augmentation

Scaling, rotation, affine transforms

Makes CNN training more robust

### Dealing with Small Datasets

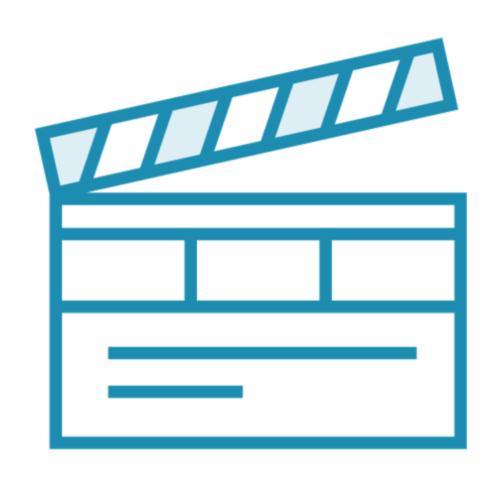
Model complexity

Transfer learning

Data augmentation

Synthetic data

## Synthetic Data



Artificially generate samples which mimic real world data

Oversampling of existing data points

Can introduce bias in existing data

#### Problems with Data

Insufficient data

Too much data

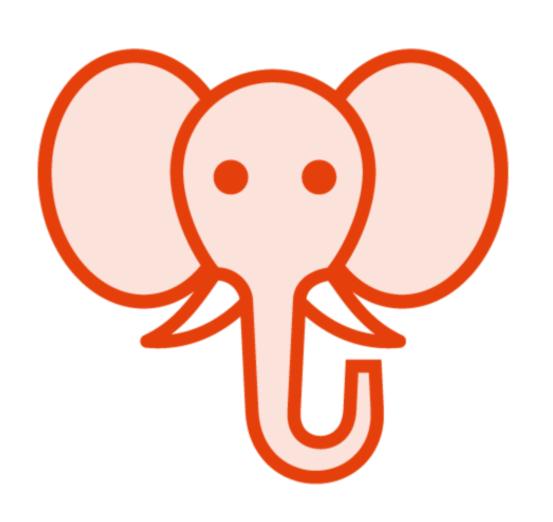
Non-representative data

Missing data

Duplicate data

Outliers

#### Too Much Data



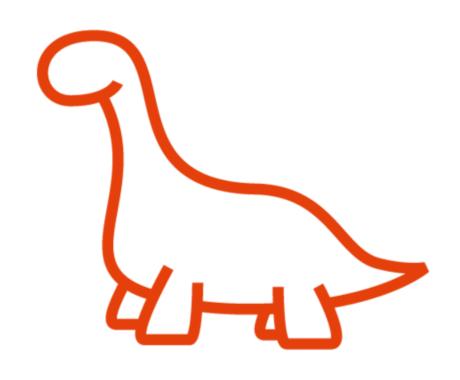
#### Data might be excessive in two ways

- Curse of dimensionality: Too many columns
- Outdated historical data: Too many rows

## Concept Drift

The relationship between features (X-variables) and labels (Y-variables) changes over time; ML models fail to keep up, and consequently their performance suffers

#### Outdated Historical Data



If not eliminated, leads to concept drift

Outdated historical data is a serious issue in specific applications

Financial trading

Usually requires human expert to judge which rows to leave out

## Curse of Dimensionality



## Two specific problems arise when too much data is available

- Deciding which data is actually relevant
- Aggregating very low-level data into useful features

## Curse of Dimensionality



#### Easier problems to solve

- Feature selection: Deciding which data is actually relevant
- Feature engineering: Aggregating very low-level data into useful features
- Dimensionality Reduction: Reduce complexity without losing information

## Concept Hierarchy

A mapping that combines very low-level features (e.g. latitudes and longitudes) into more general, usable features (e.g. zip codes)

#### Problems with Data

Insufficient data

Too much data

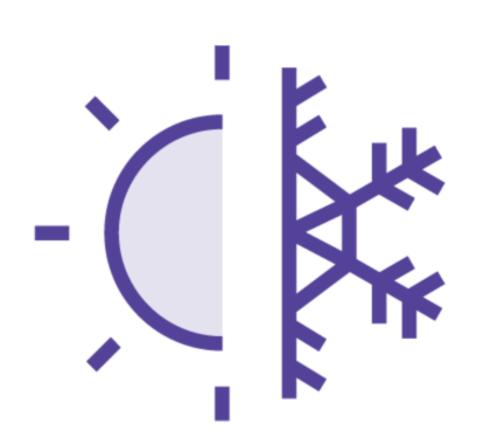
Non-representative data

Missing data

**Duplicate data** 

**Outliers** 

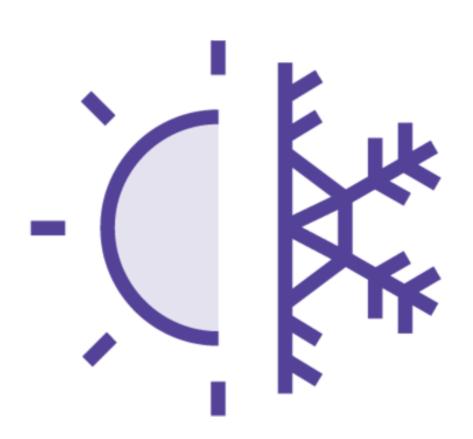
## Non-representative Data



Data is inaccurate, small errors have significant impact

Account for data cleaning and processing time

## Non-representative Data



Data not representative of the real world i.e. biased

Leads to biased models that perform poorly in practice

Mitigate using oversampling and undersampling

#### Problems with Data

Insufficient data

Too much data

Non-representative data

Missing data

Duplicate data

Outliers

## Cleaning Data



# Data cleaning procedures can help significantly mitigate effect of

- Missing data
- Outliers

#### Problems with Data

Insufficient data

Too much data

Non-representative data

Missing data

Duplicate data

Outliers

## Duplicate Data



# If data can flagged as duplicate, problem relatively easy to solve

- Simply de-duplicate

# Can be hard to identify in some applications

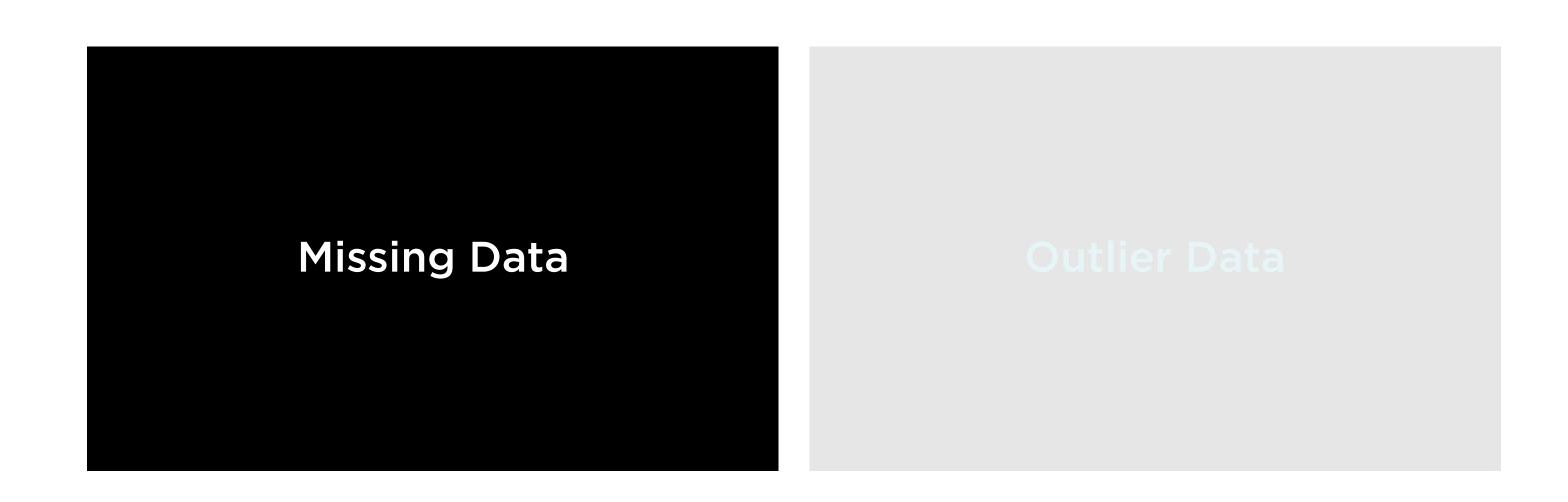
- Real-time streaming

## Missing Values and Outliers

## Data Cleaning and Preparation

Missing Data Outlier Data

## Data Cleaning and Preparation



# Missing Data Deletion Imputation

## Deletion a.k.a. Listwise Deletion

Delete an entire record (row) if a single value (column) is missing. Simple but can lead to bias.

#### Listwise Deletion



Most common method in practice

Can reduce sample size significantly

If values are not missing at random, can introduce significant bias

## Imputation

Fill in missing column values, rather than deleting records with missing values. Missing values are inferred from known data.

## Imputation



Methods range from very simple to very complex

Simplest method: Use column average

Can interpolate from nearby values

Can even build model to predict missing values

## Multivariate Imputation



Univariate imputation: Rely only on known values in same feature

Multivariate imputation: Use all known data to infer missing data

- Construct regression models from other columns to predict this column
- Iterative repeat for all columns

## Hot-deck Imputation



Sort records based on any criteria

For each missing value, use immediately prior available value

"Last Observation Carried Forward"

For time series, equivalent to assuming no change since last measurement

#### Mean Substitution

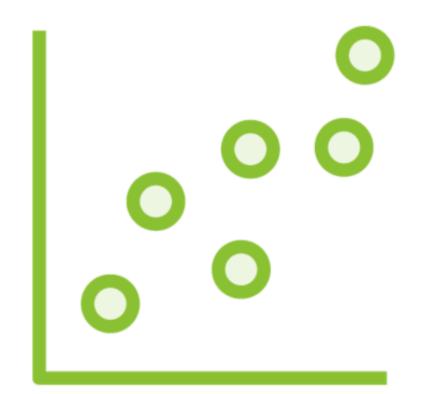


For each missing value, substitute mean of all available values

Has effect of weakening correlations between columns

Can be problematic when bivariate analysis required

## Regression

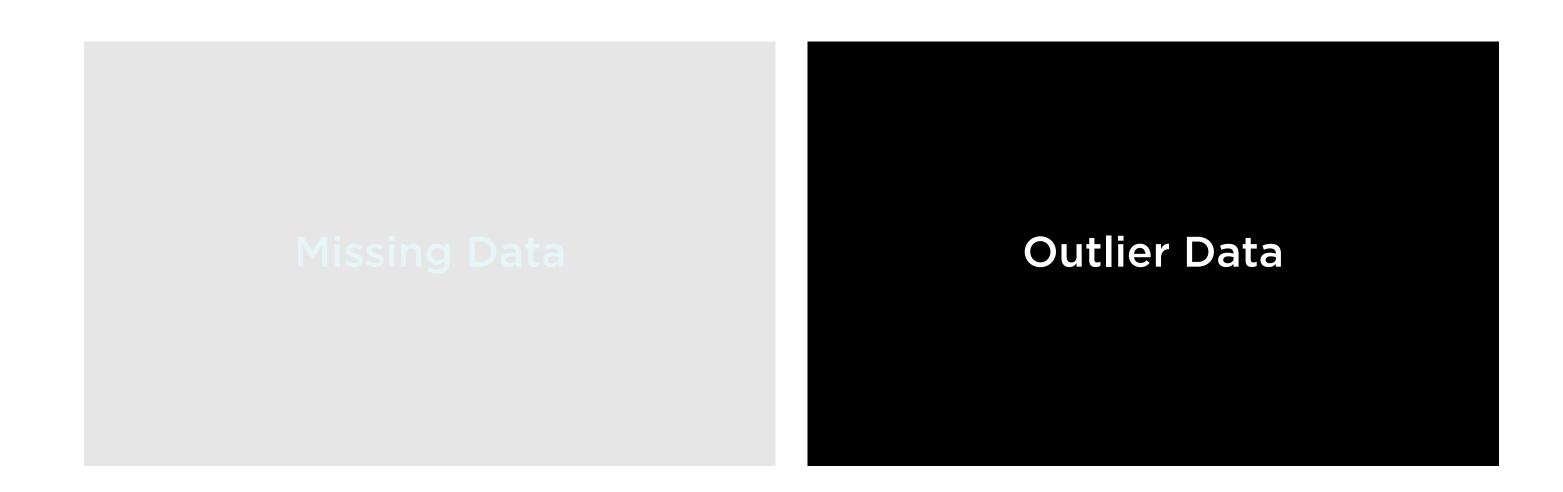


Fit model to predict missing column based on other column values

Tends to strengthen correlations

Regression and mean substitution have complementary strengths

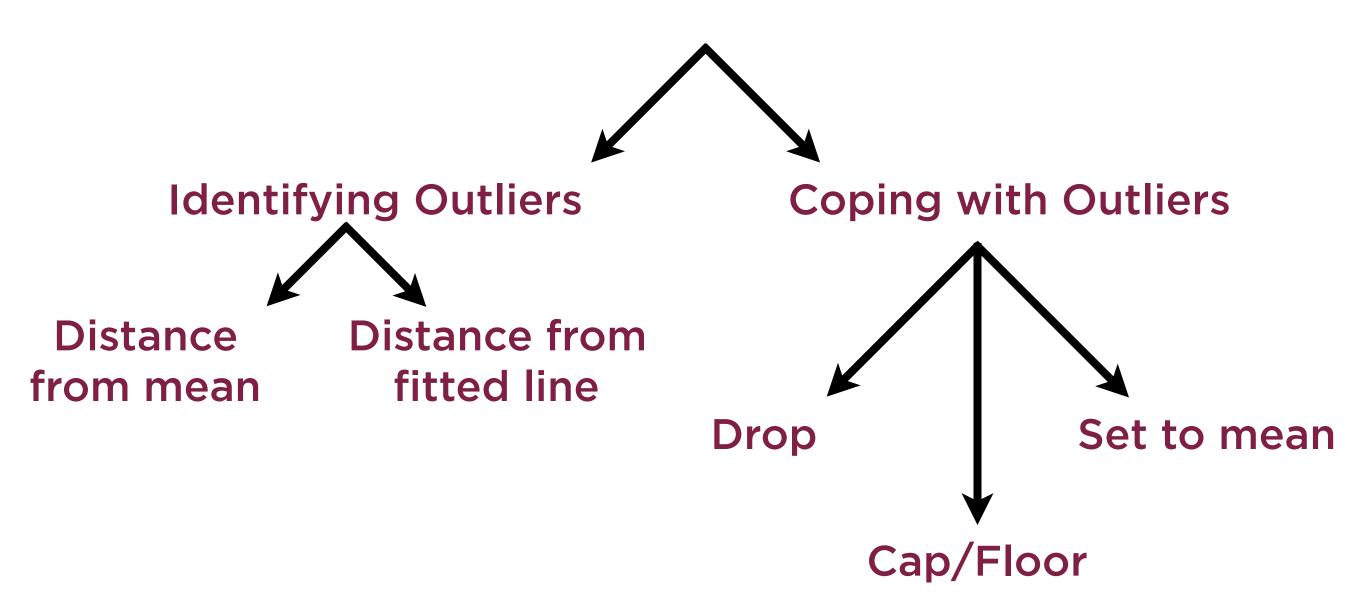
## Data Cleaning and Preparation



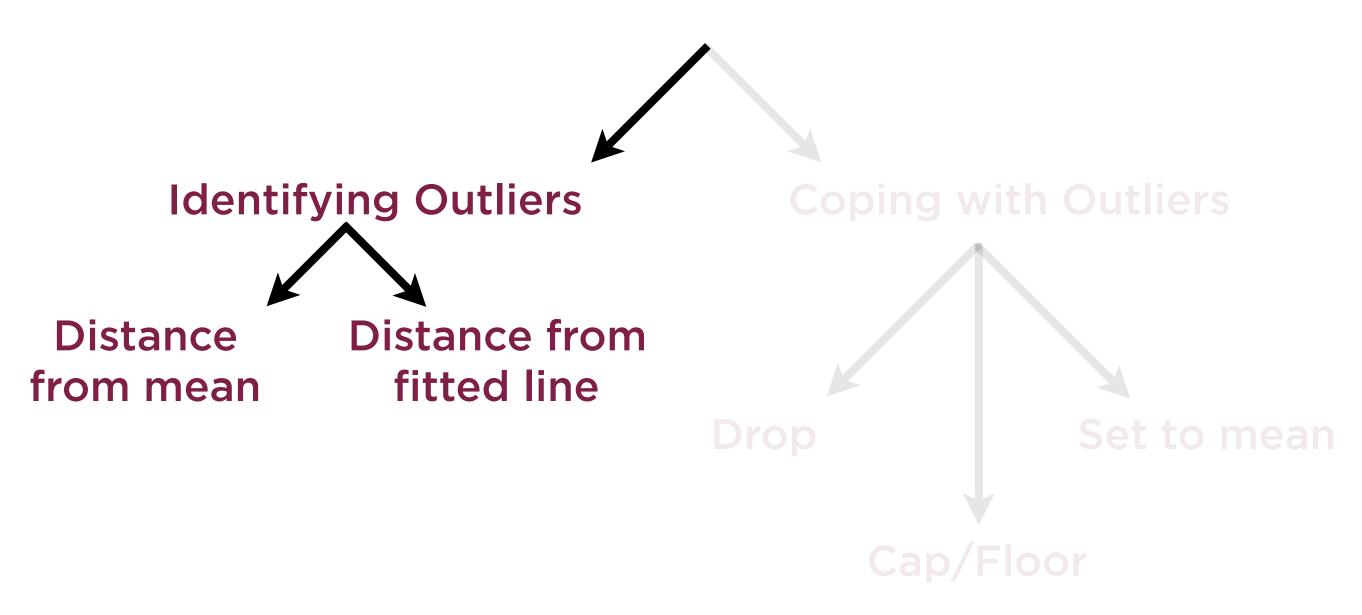
## Outlier

A data point that differs significantly from other data points in the same data set.

#### Outliers



#### Outliers



## Identifying Outliers

Distance from mean

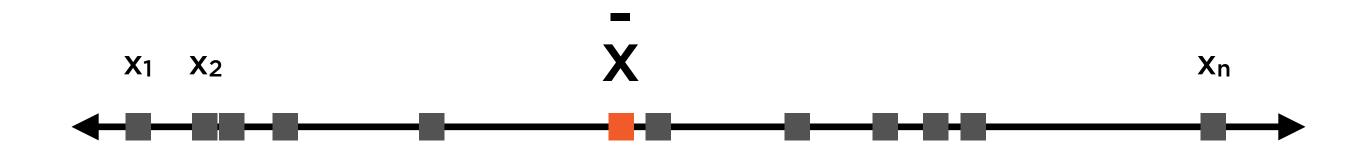
Distance from fitted line

## Identifying Outliers

Distance from mean

Distance from fitted line

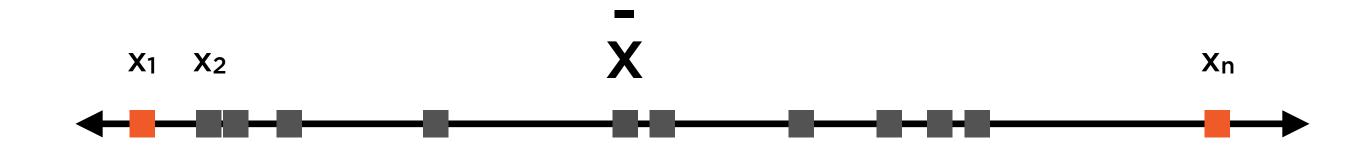
#### Mean as Headline



# The mean, or average, is the one number that best represents all of these data points

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

### Variation Is Important Too

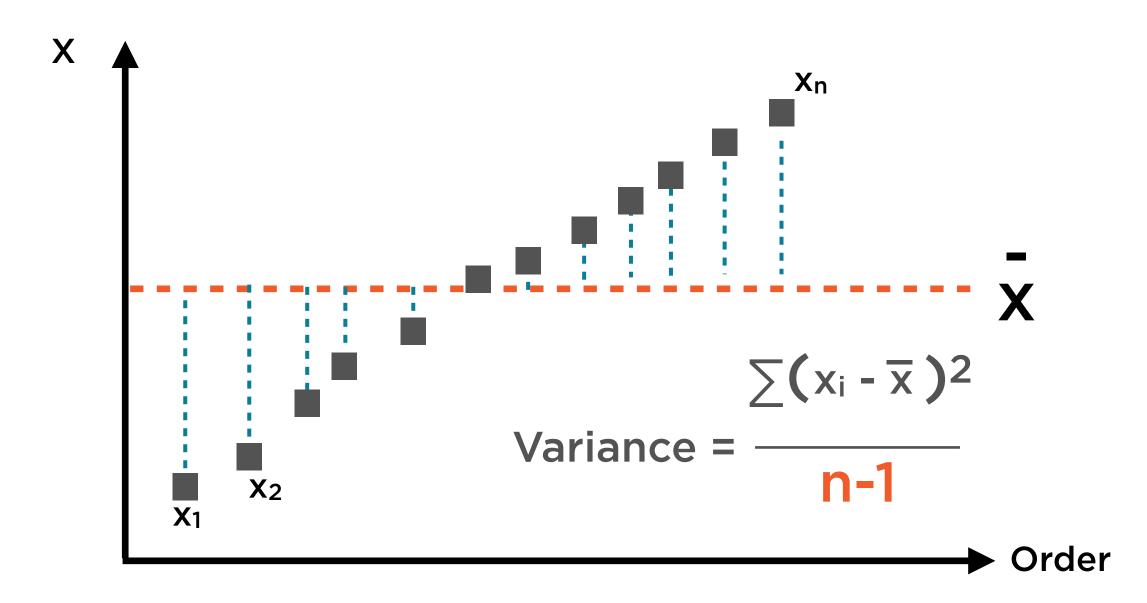


"Do the numbers jump around?"

Range =  $X_{max} - X_{min}$ 

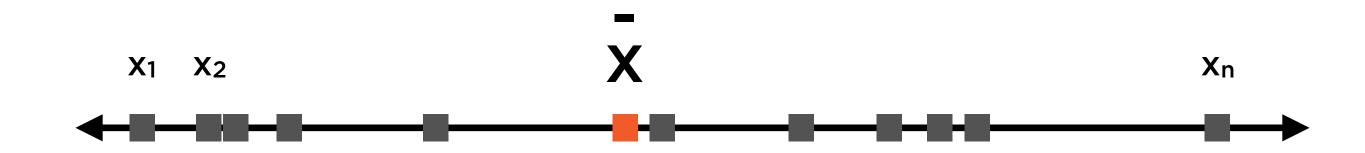
The range ignores the mean, and is swayed by outliers - that's where variance comes in

### Variance as Asterisk



Variance is the second-most important number to summarize this set of data points

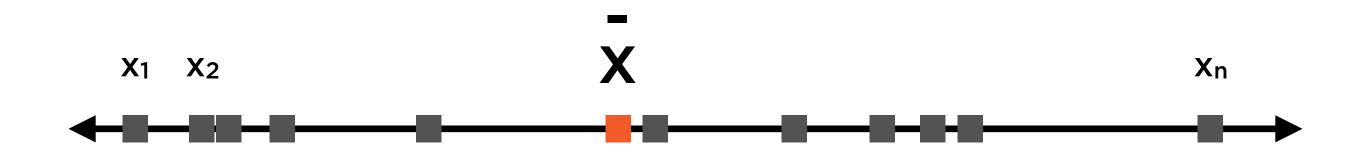
### Mean and Variance



# Mean and variance succinctly summarize a set of numbers

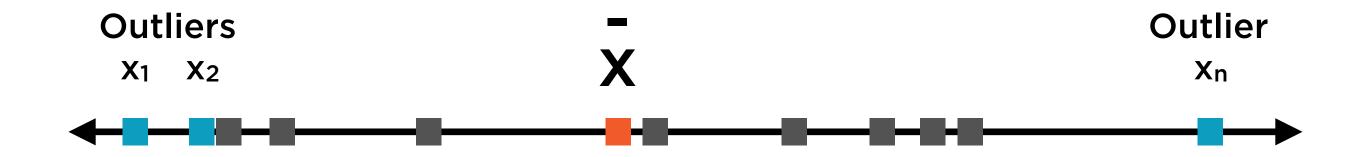
$$\frac{1}{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$
 Variance =  $\frac{\sum (x_i - \overline{x})^2}{n-1}$ 

### Variance and Standard Deviation



Standard deviation is the square root of variance

Variance = 
$$\frac{\sum (x_i - \overline{x})^2}{n-1}$$
 Std Dev = 
$$\sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$



Points that lie more than 3 standard deviations from the mean are often considered outliers

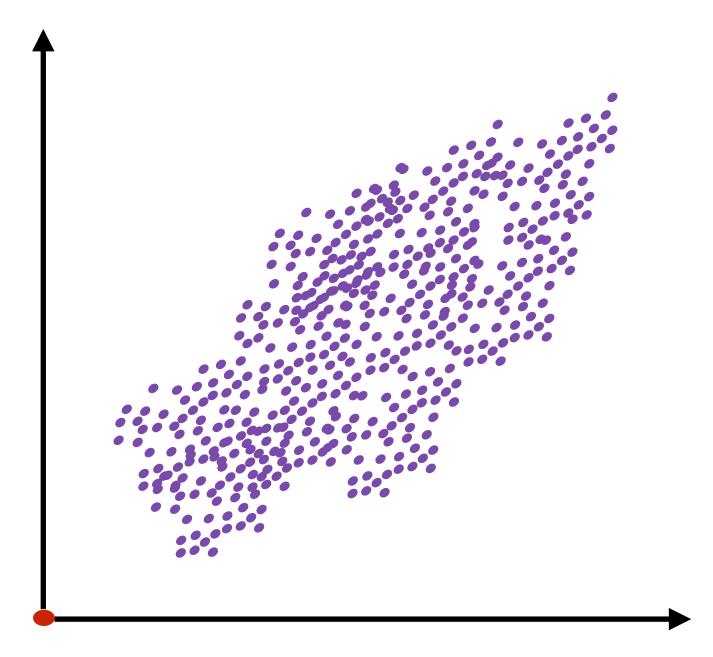


Points that lie more than 3 standard deviations from the mean are often considered outliers

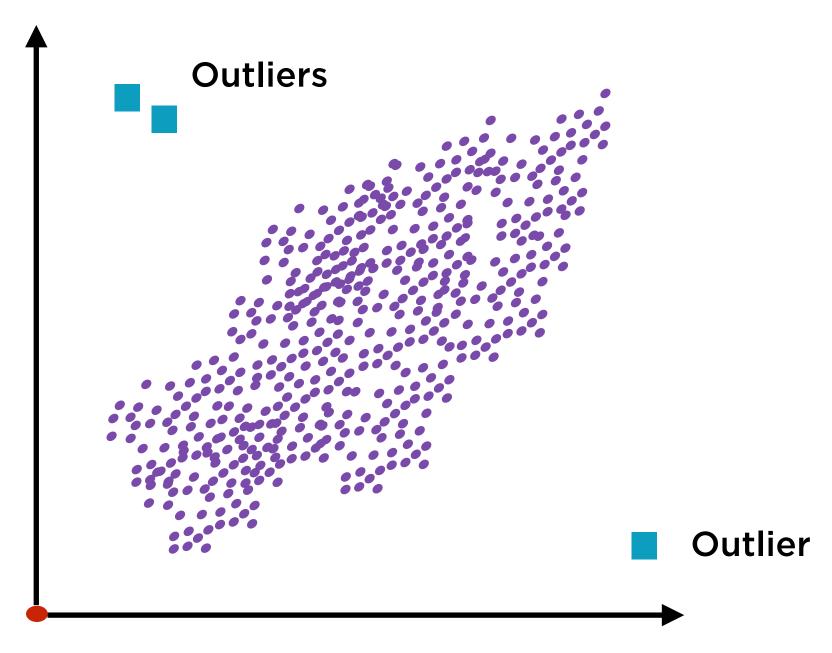
# Identifying Outliers

Distance from mean

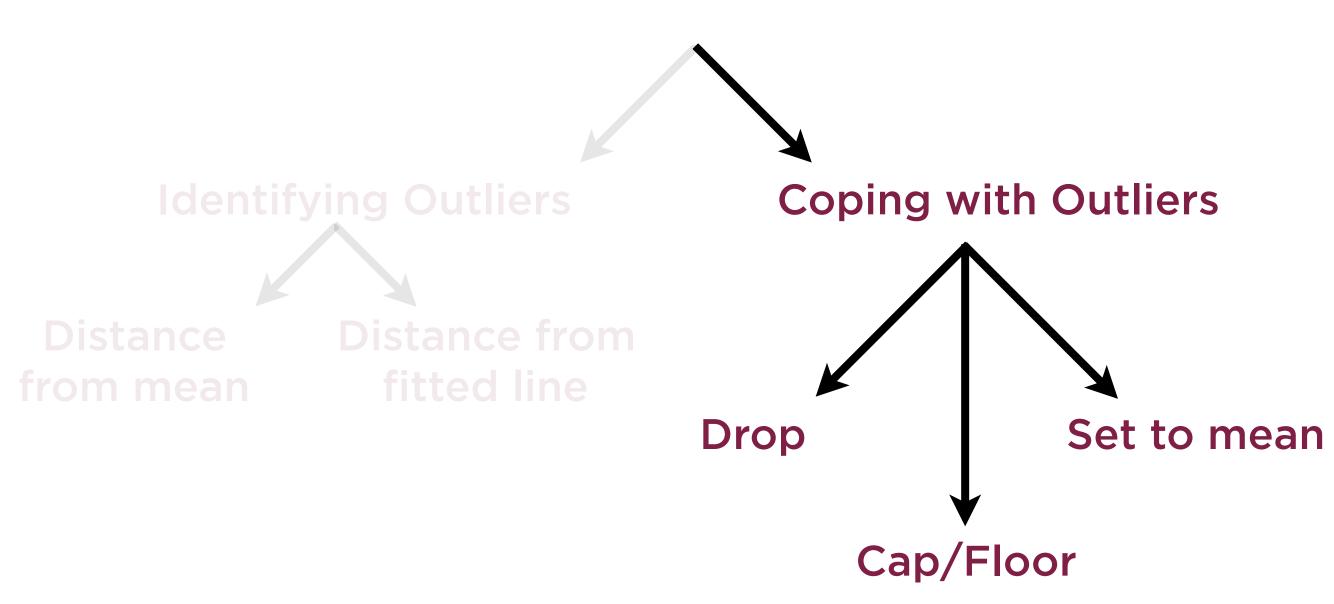
Distance from fitted line



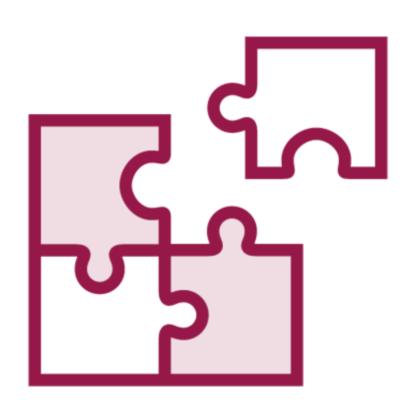
Outliers might also be data points that do not fit into the same relationship as the rest of the data



Outliers might also be data points that do not fit into the same relationship as the rest of the data



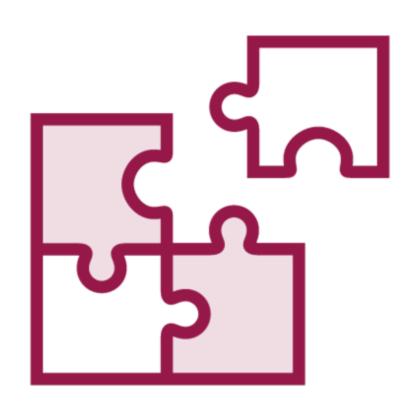
# Coping with Outliers



# Always start by scrutinizing outliers If erroneous observation

- Drop if all attributes of that point are erroneous
- Set to mean if only one attribute is erroneous

# Coping with Outliers



### If genuine, legitimate outlier

- Leave as-is if model not distorted
- Cap/Floor if model is distorted
  - Need to first standardize data
  - Cap positive outliers to +3
  - Floor negative outliers to -3

# Oversampling and Undersampling

# From Sample to Population





All the data out there in the universe



Sample

A subset - hopefully representative - of the population

# From Sample to Population







**Population** 

Representative Sample

**Biased Sample** 

# When Unbiased Samples Make It Hard



A study seeks to measure health effect of a certain chemical

Exposure to chemical is random and extremely rare

For a meaningful test, an unbiased sample would need to be huge

Could we focus on the few exposed instances? (Case studies)

# When Unbiased Samples Make It Hard



Image classifier looking for photos with Hawaiian Crow

One of the rarest birds on earth, looks a lot like the Common Crow

### When Unbiased Samples Make It Hard



Training corpus has millions of images with the Common Crow

Only a dozen images with the Hawaiian Crow

Could we re-use images of the Hawaiian Crow?

# Oversampling and Undersampling are techniques that intentionally add bias to the data in order to make it balanced

### Balancing Datasets

Oversampling of uncommon x or y values

Undersampling of common x or y values

### Forcibly Balanced Datasets



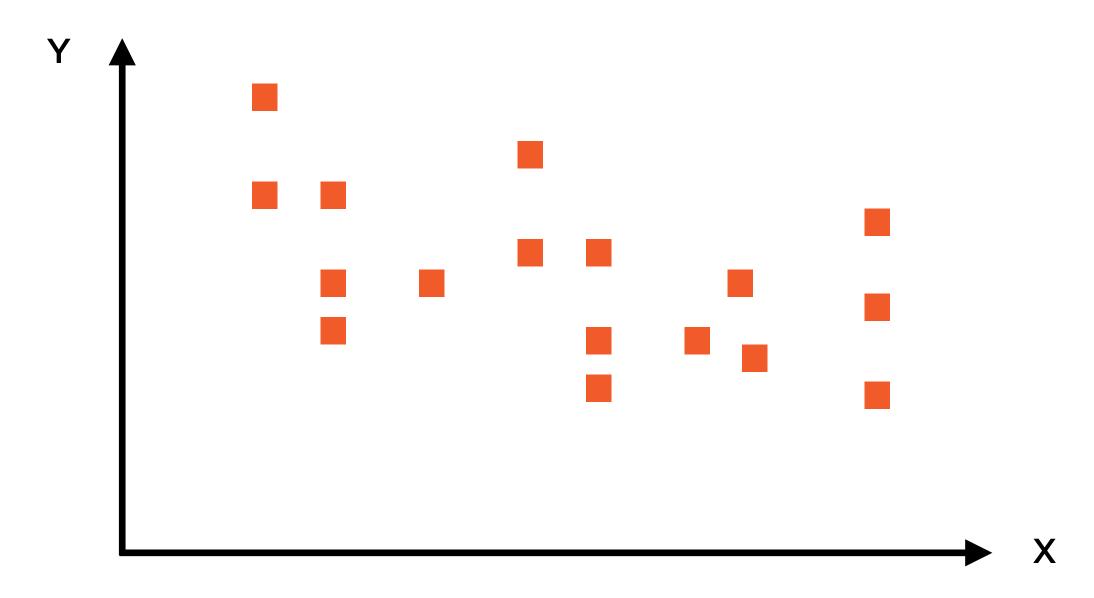
# Oversampling and undersampling tend to

- Reduce accuracy
- Increase precision and recall

#### Related techniques include

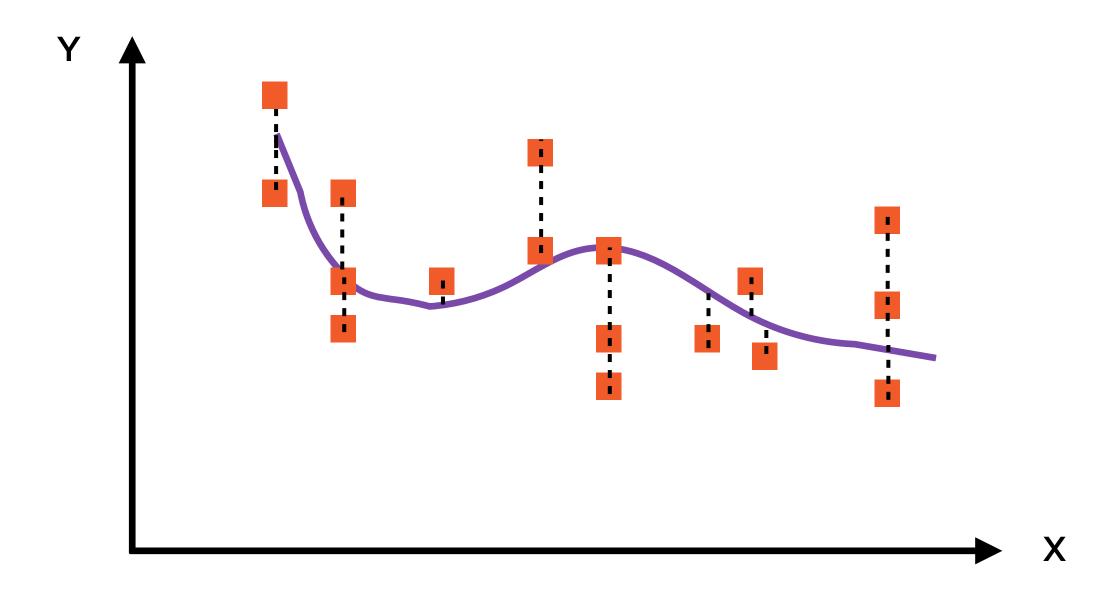
- Case studies
- Stratified sampling

# Overfitting and Underfitting

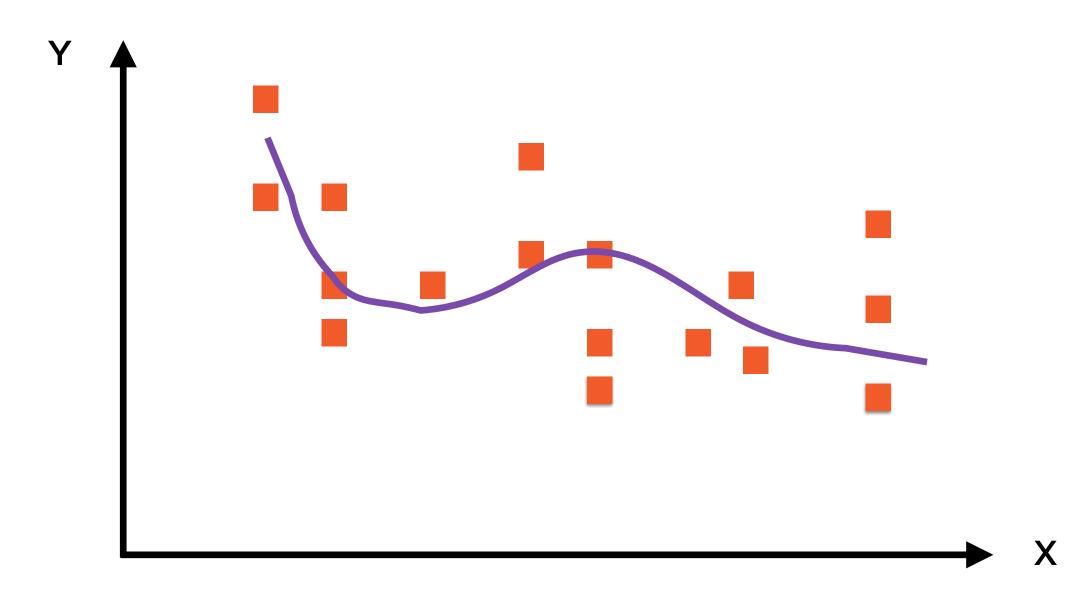


Challenge: Fit the "best" curve through these points

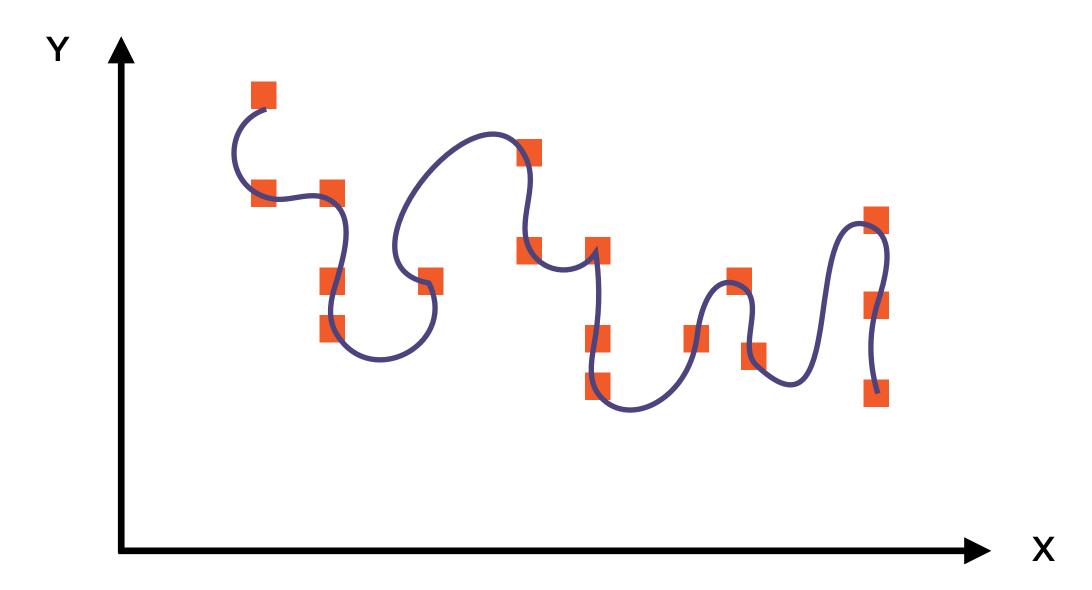
### Good Fit?



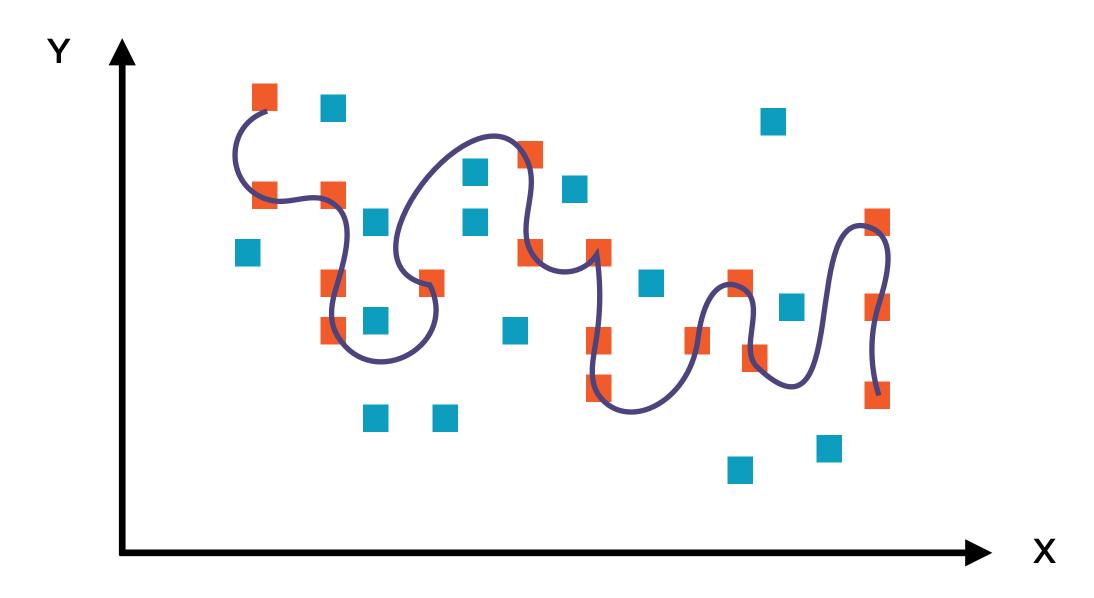
A curve has a "good fit" if the distances of points from the curve are small



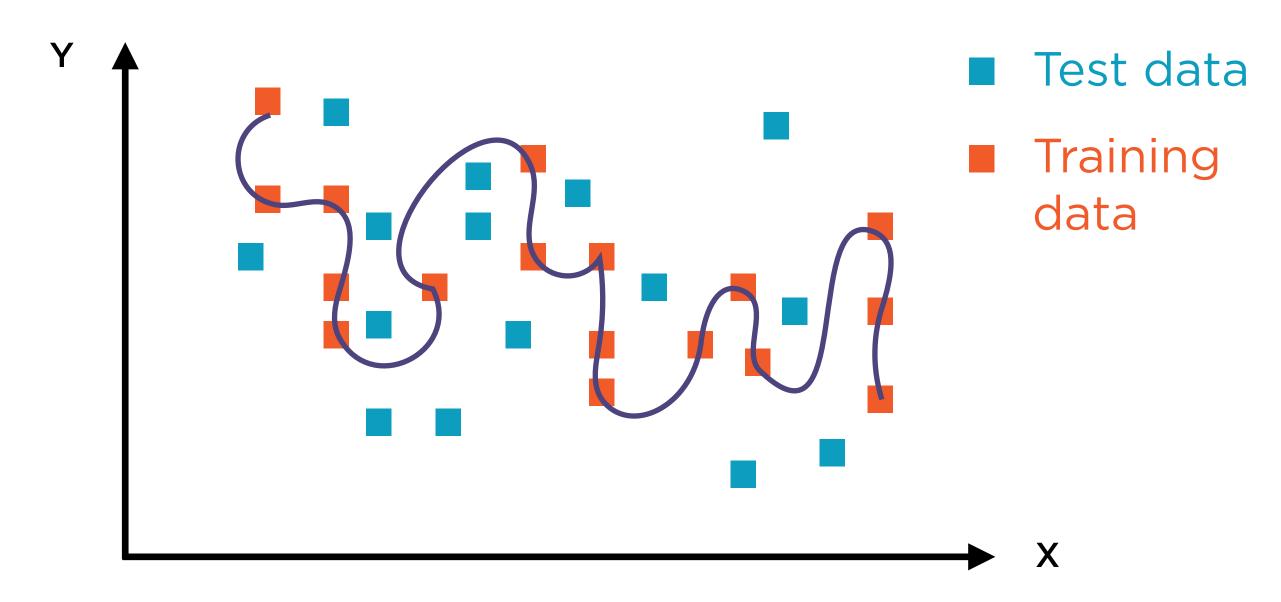
We could draw a pretty complex curve



We can even make it pass through every single point

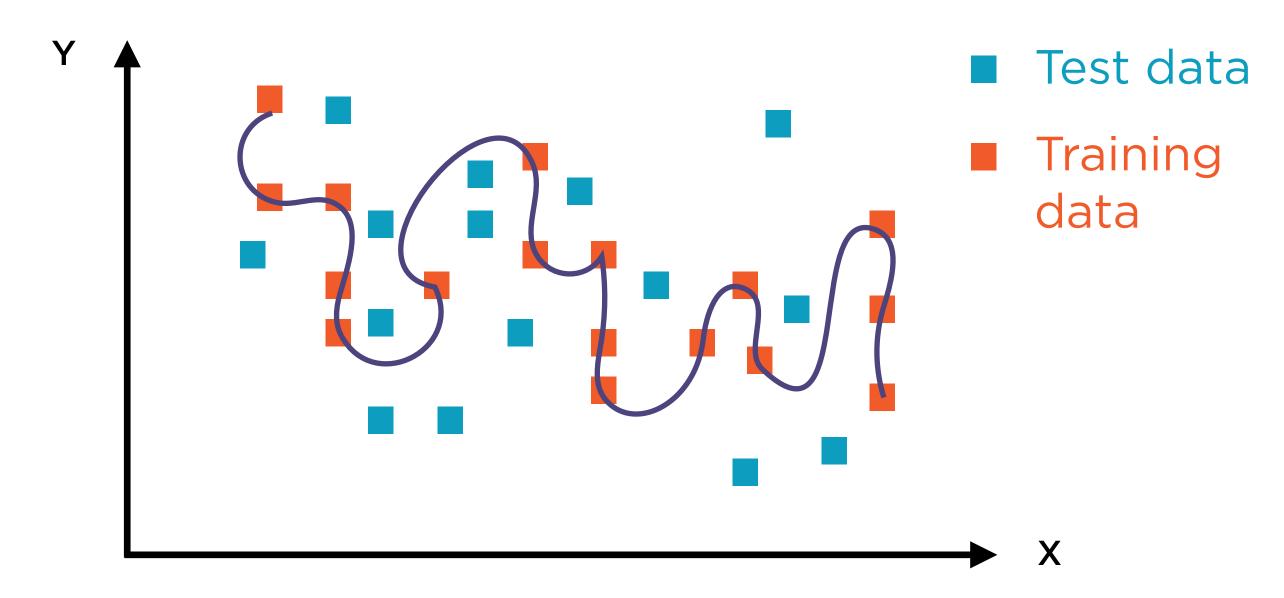


But given a new set of points, this curve might perform quite poorly

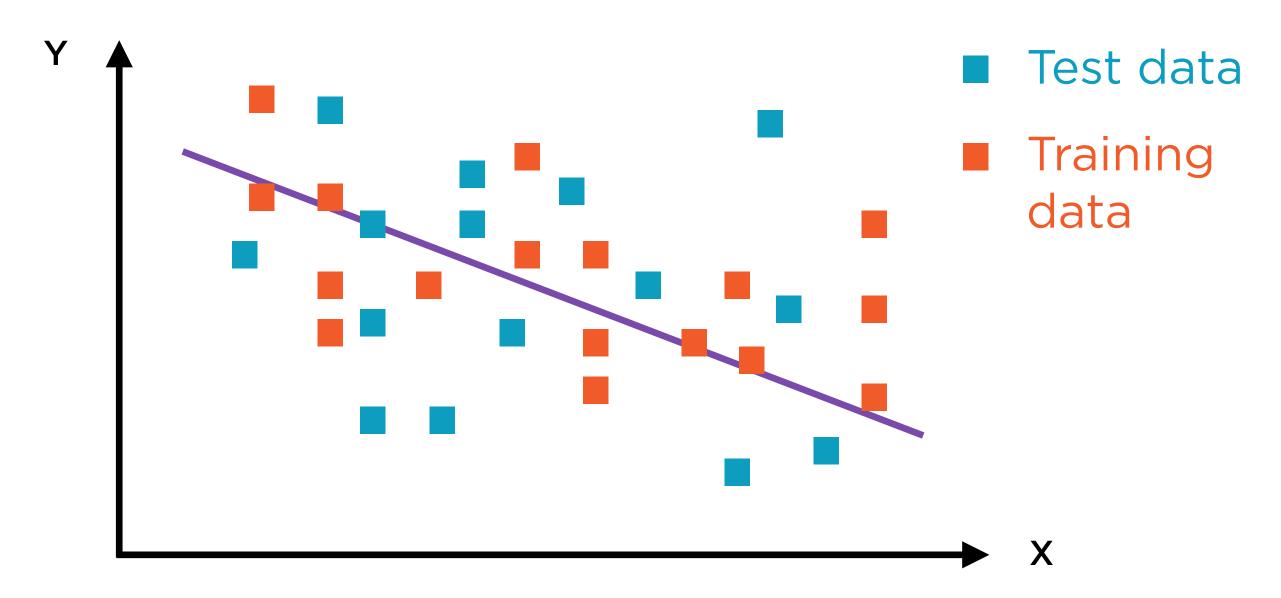


The original points were "training data", the new points are "test data"

### Overfitting

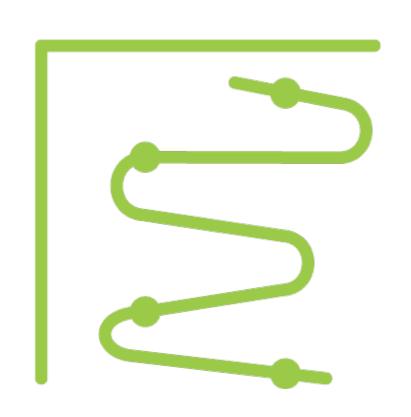


Great performance in training, poor performance in real usage



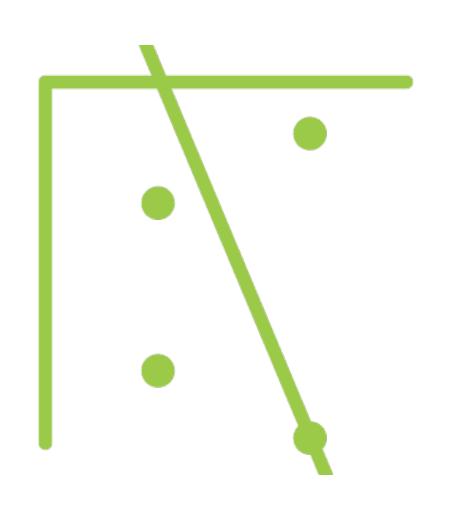
A simple straight line performs worse in training, but better with test data

### Overfitting



Model has memorized the training data
Low training error
Does not work well in the real world
High test error

### Underfitting



Model unable to capture relationships in data

Performs poorly on the training data

Model too "simple" to be useful

### Preventing Overfitting



Regularization - Penalize complex models



Cross-validation - Distinct training and validation phases



Dropout (NNs only) - Intentionally turn off some neurons during training

### Summary

Need for data preparation in machine learning

Insufficient data

Excessive or overly complex data

Non-representative data, missing data, outliers

Oversampling and undersampling

Overfitting and underfitting models