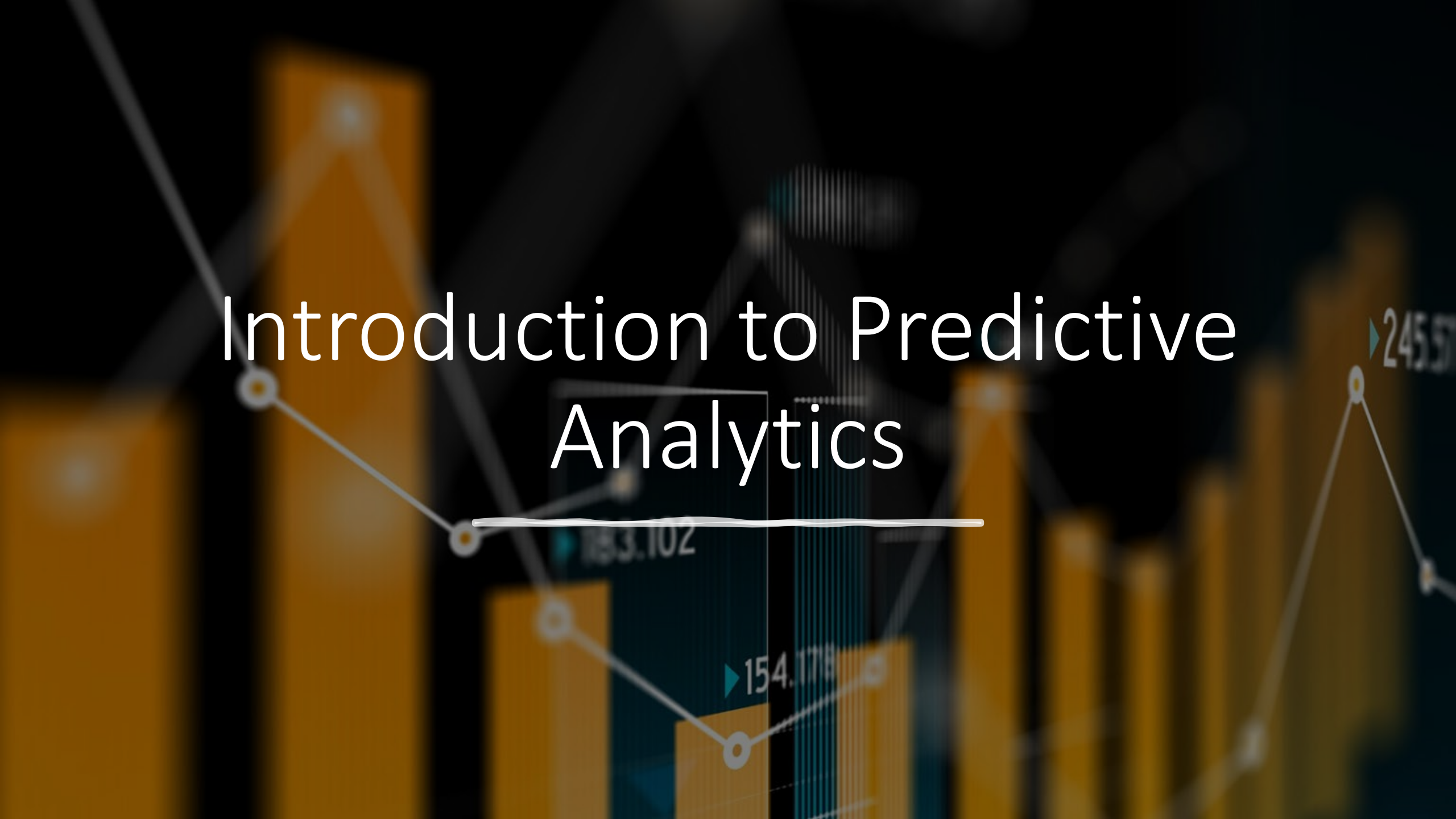


Introduction to Predictive Analytics

ABOUT ME:

Dr. Brittany Green

Asst. Professor of Information Systems, Analytics, and Operations at University of Louisville

Background: PhD in Business Analytics from University of Cincinnati

MS in Operations Research from University of Pittsburgh

MS in Industrial Engineering from Auburn University

BS in Math with minor in CS from Birmingham Southern College

ABOUT ME:

Dr. Brittany Green

Asst. Professor of Information Systems, Analytics, and Operations

Analytics Consulting:

Samsung.com, VA healthcare system, Trilogy Health, UC Sports Medicine

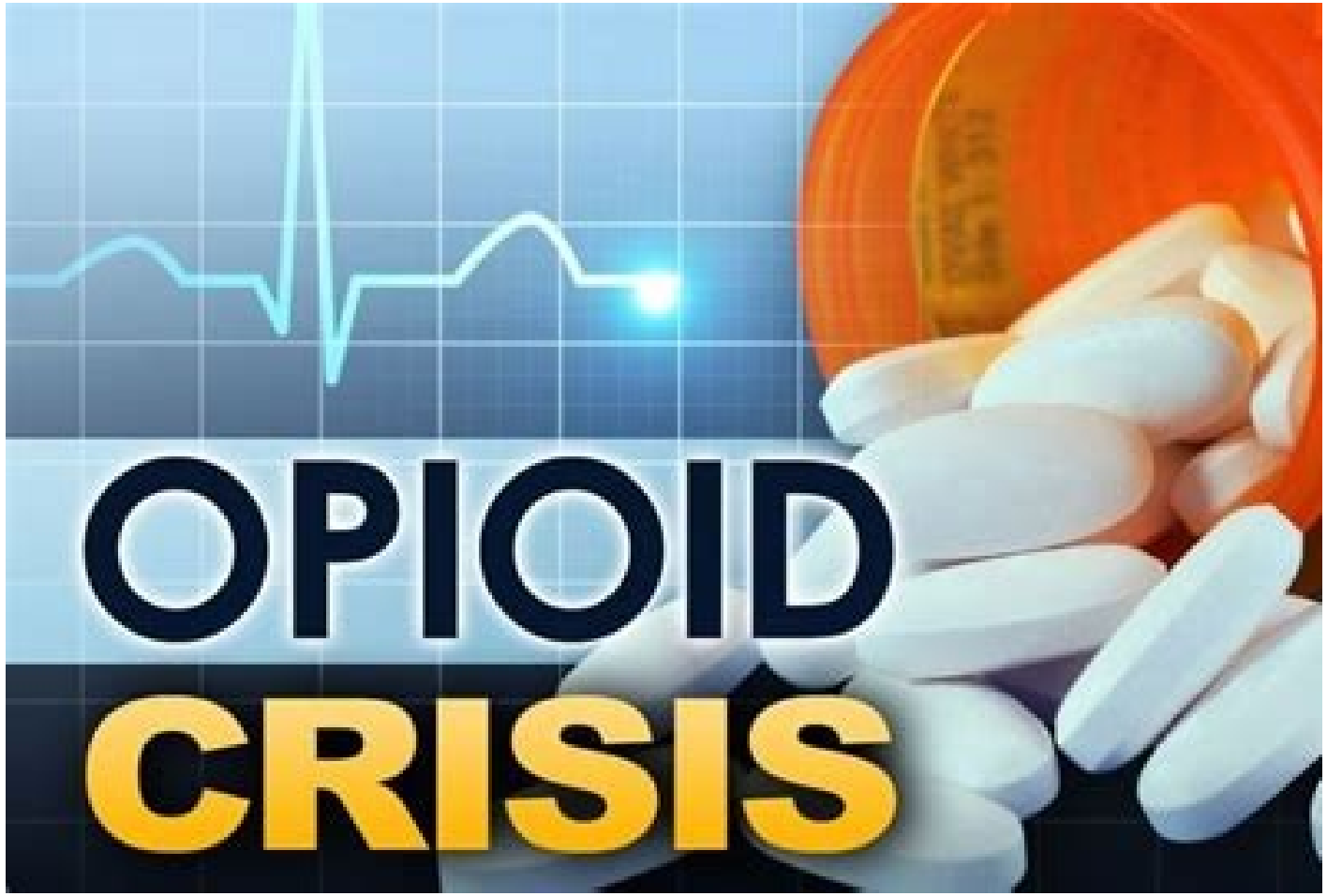
Example research:

Ultra-high Dimensional Quantile Regression for Longitudinal Data: an Application to Blood Pressure Analysis *Journal of the American Statistical Association*

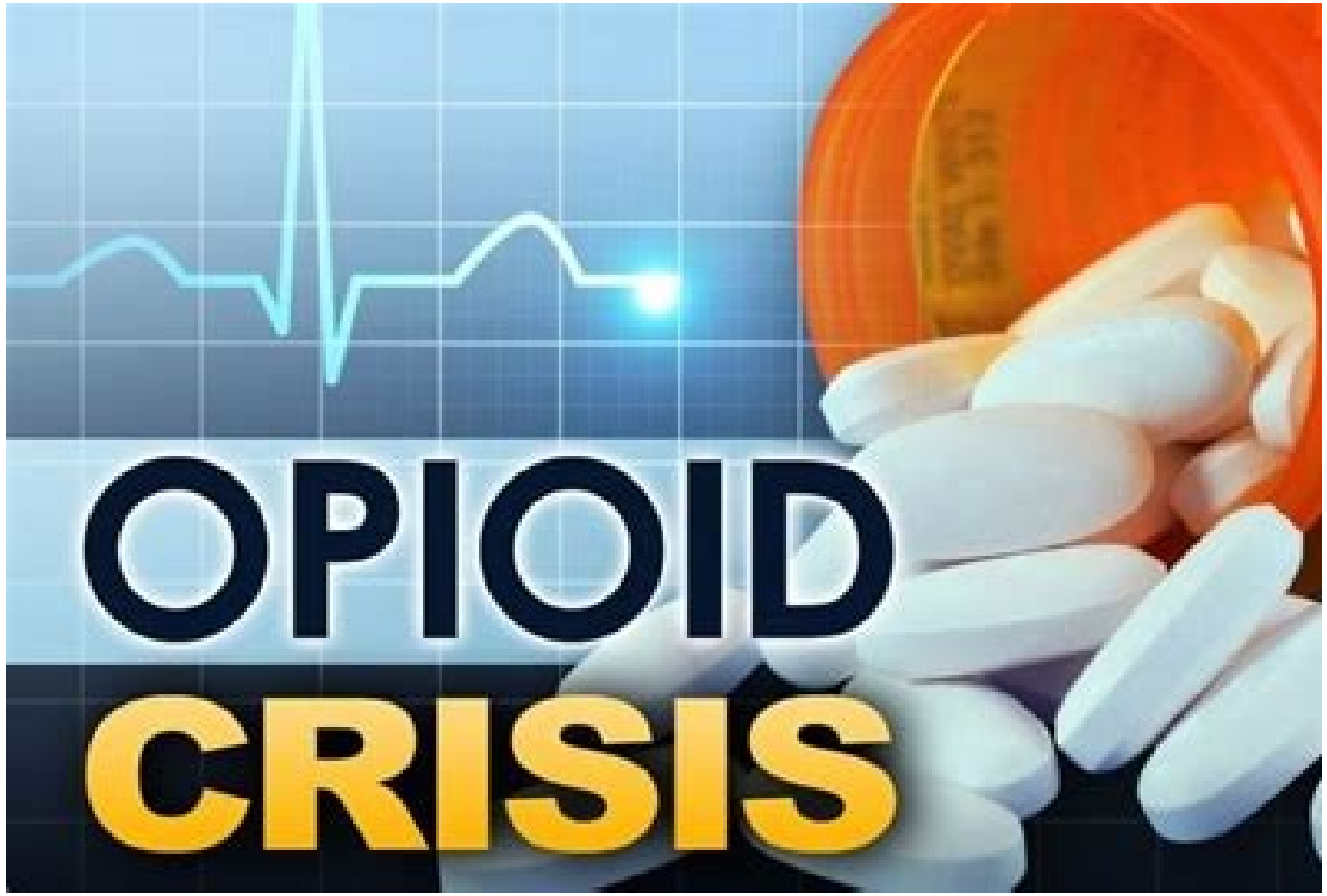
Machine learning classification of verified head impact exposure strengthens associations with brain changes. *Annals of biomedical engineering*

Syllabus Overview

- Any questions?



- As you watch video, listen for two things
  - What's the outcome we care about?
  - And what decisions could be improved if we could predict that outcome earlier?



- 1 in 4 patients receiving long term opioid therapy will struggle with opioid addiction and dependence.
- Health insurance costs have skyrocketed
- If you were a healthcare company,
  - what would you predict,
  - what action would you take because of that prediction,
  - what data would you use, and how would you model?

Goal: “identify individuals who are at risk of getting addicted to opioids.”

Source: Humana-Mays Healthcare Analytics 2019 Case Competition

Humana

*we need to go beyond just thinking in terms of which model to use to thinking about how our tools can benefit the business and society.

Goal: “Proactively identify individuals who are at a higher risk of getting addicted to opioids for tracking and appropriate servicing. This will not only help improve the well being of the members but also reduce the cost burden on the individual as well as the Humana. Humana will have a better idea on a member’s likelihood of long term use of opiods and will thus be able to make more informed decisions on prescription of opioids as well as timely interventions.”

Source: Humana-Mays Healthcare Analytics 2019 Case Competition

This course is all about…

Turning data into useful business insights.

- Business questions:
  - How can our client improve sales?
  - How can a company reduce customer churn?
  - How can banks reduce fraud?
  - How can an airline company reduce the length of flight delays?

This course is all about…

How can we best estimate the relationship between data and an outcome of interest (e.g., sales)? Can we find a model and variables that are helpful in predicting an outcome of interest?

This course is all about…

Random Forest

Logistic Regression

How can we best estimate the relationship between data and an outcome of interest?

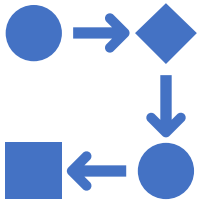
Can we find factors that are helpful in predicting an outcome of interest?

Gradient Boosting

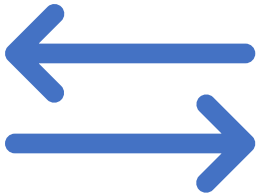
LASSO

But… it’s not simply applying techniques

A powerful, flexible model will always find some pattern in data, but does the model predict new observations well?



Does your overall approach for solving the business problem have errors? Can you identify errors from other proposals?



Can you clearly report and present the results and relate back to the business problem?

Goal of this course: Further prepare you for Data Science in Business

Consulting assignments

Data science interview questions

Evaluate data science proposals and approaches

Case competition

Syllabus Overview

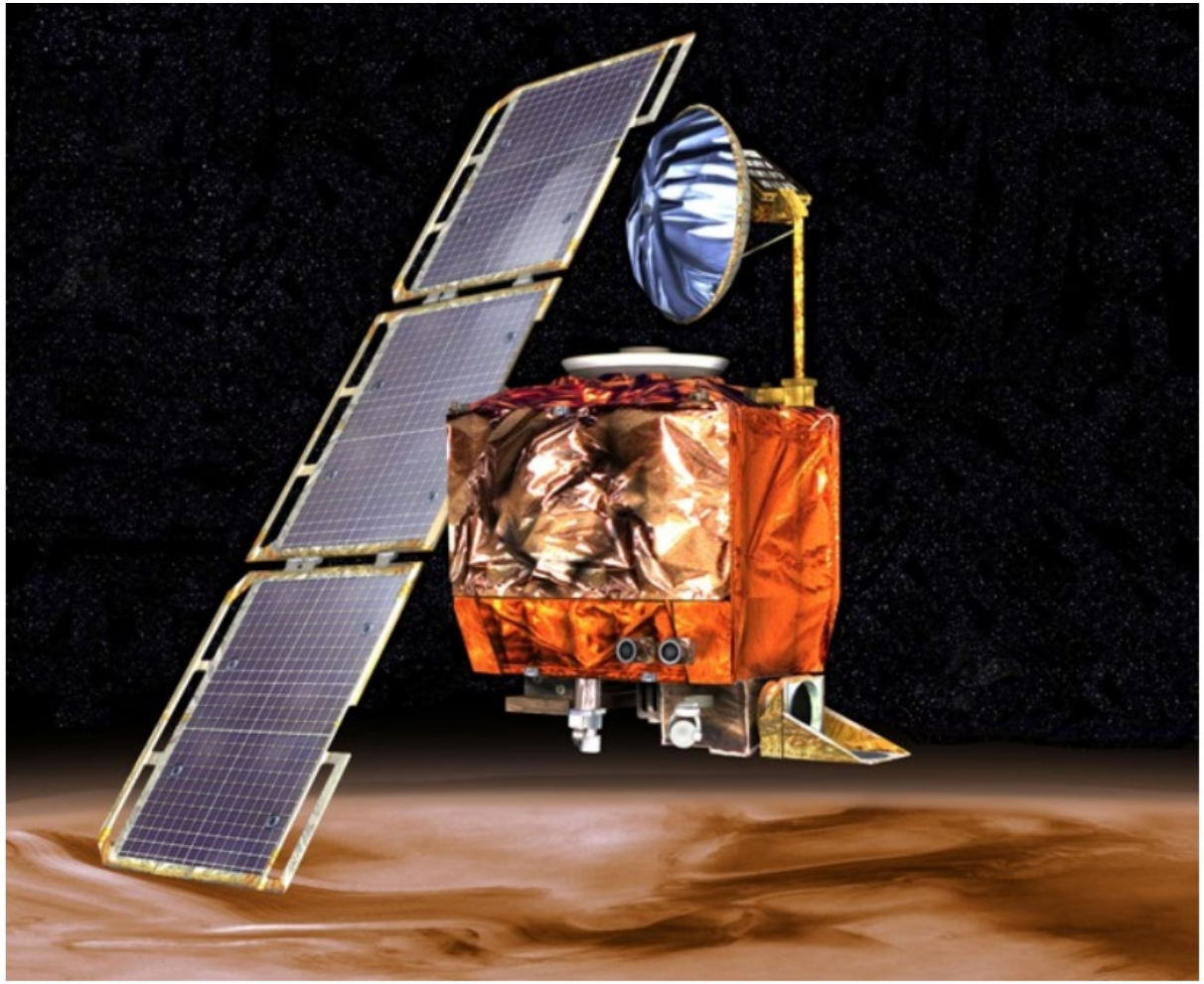
- Any questions?

Today's Lecture: Key Concepts

1. Exploratory Data Analysis
2. Supervised vs. Unsupervised Methods
3. Data mining vs using the results of data mining
4. How to choose one model over another?
   - Overfitting and Generalization
   - Partitioning data
5. How to evaluate performance for regression problems?
6. Bias Variance Tradeoff

Exploratory Data Analysis: Get to know your data and provide insights

$125 million gone…



1999: A disaster investigation board reports that NASA's Mars Climate Orbiter burned up in the Martian atmosphere because engineers failed to convert units from English to metric.

https://www.wired.com/2010/11/1110mars-climate-observer-report/

Exploratory Data Analysis: Get to know your data

- Check number of variables, number of observations, number of missing data points
  - If more than 30-40% missing in one column could consider deleting
- Check data quality
  - Incorrect units, inconsistencies (e.g., age > 150), misspelled categories, duplicate rows, etc.
- To do this consider using:
  - For each continuous/numeric variable: histogram, boxplot, mean, standard deviation, other summary statistics
  - For each categorical variable: table to get the number in each category, bar plots

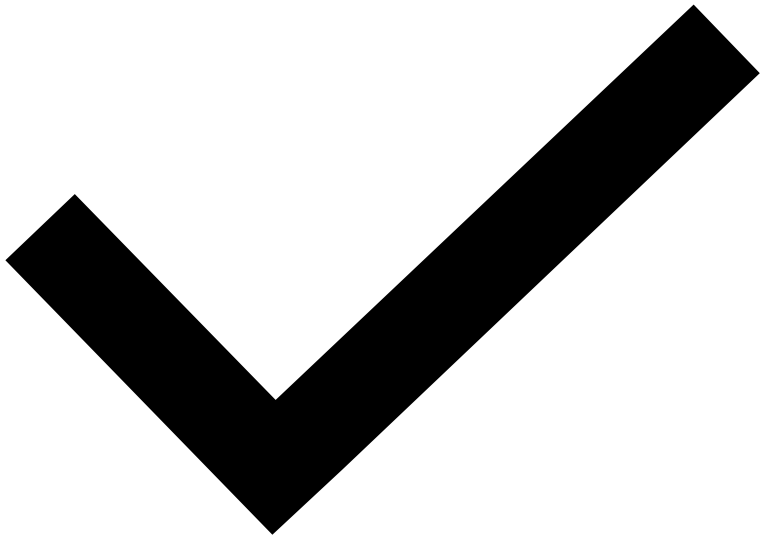
Exploratory Data Analysis: Provide insights

- Find interesting relationships or patterns between response and predictor variables
- If the response is continuous
  - **Numeric predictor variable**: scatterplots between numeric predictor variable and numeric response
  - **Categorical predictor variable**: create histogram, boxplot of response variable separated by categories of predictor variable
- If the response is binary
  - **Numeric predictor variable**: create histogram, boxplot of predictor variable separated by categories of response variable
  - **Categorical predictor variable**: Create two-way tables, bar plots

Exploratory Data Analysis: Provide insights

- Before or even after predictive modeling, can do exploratory data analysis to visualize the most important relationships
- Critical when presenting to management

Do this every time!



EDA Consulting Exercise

(Over the next 45 minutes)

EDA Consulting Exercise

- You are the analytics team for an online chocolate retailer. Use EDA to figure out what's associated with higher spending and propose one actionable next step.

- Download the HeavenlyChocolatesInClass.csv dataset and the R-and-Python.html for the lab.

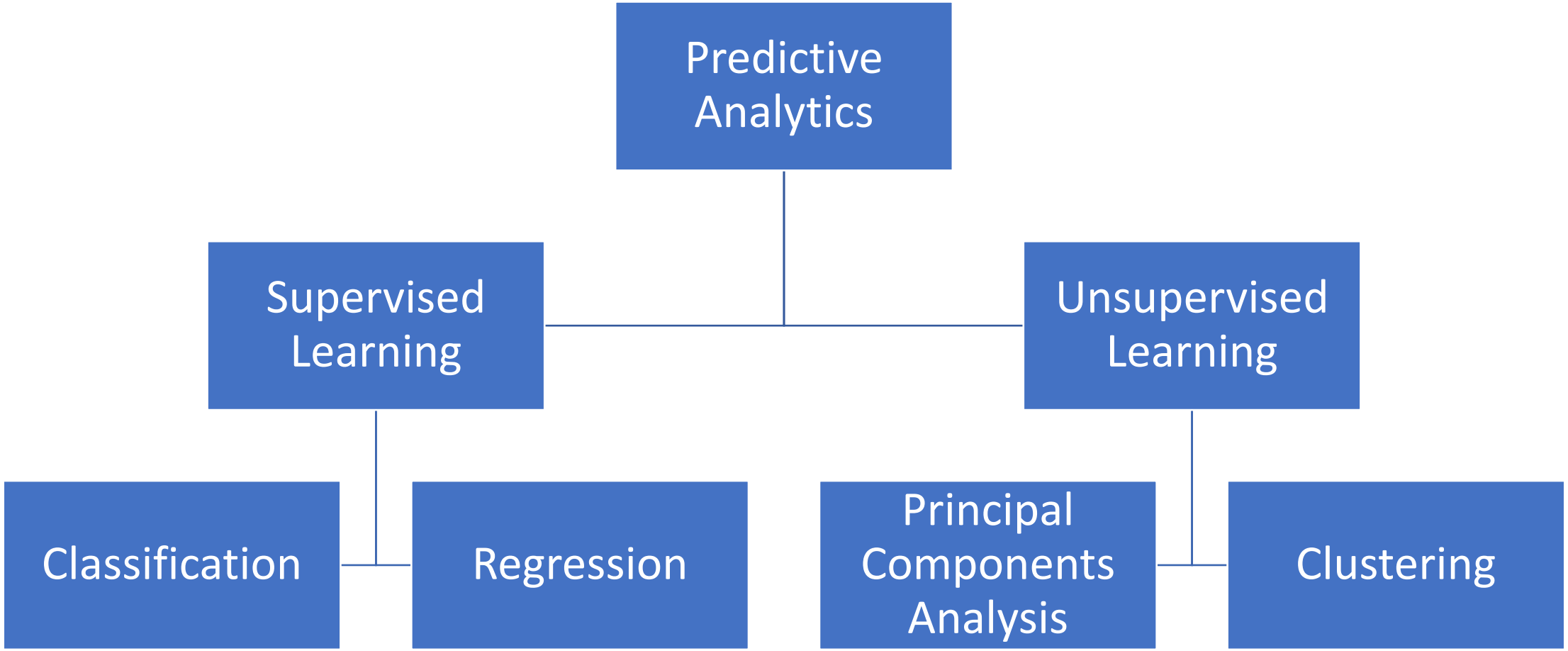
EDA Consulting Exercise

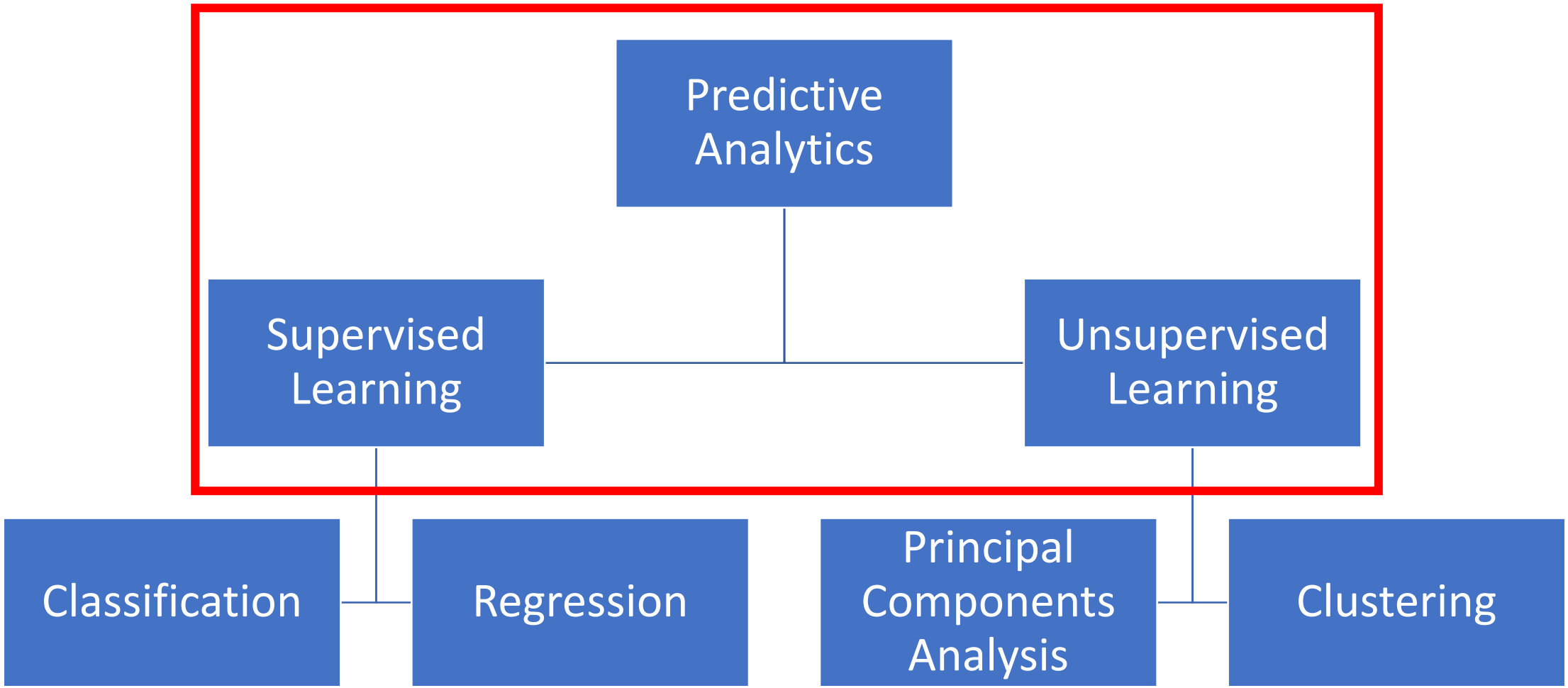
- Step 1: Every member in the team should create one univariate plots:
  - Distribution of a numeric variable using a histogram/boxplot
  - Distribution of a categorical variable using a frequency table
- Step 2: Every member in the team should create one relationship plot:
  - Numeric response by numeric variable (use scatter plot)
  - Numeric response by categorical variable (use boxplot/histogram faceted by categorical variable)
  - Numeric response by numeric by categorical or numeric reponse by categorical by categorical (see last section of the EDA lab for examples)
- Step 3: Decide on 1 final plot/business insight as a team. Be prepared to share:
  - What do you see in this plot? (describe the pattern)
  - So what? (why it matters + how it connects to sales in one statement)
  - Now what should the business do to improve sales given your findings? (one specific action)

Before we go further, let's talk data terminology.

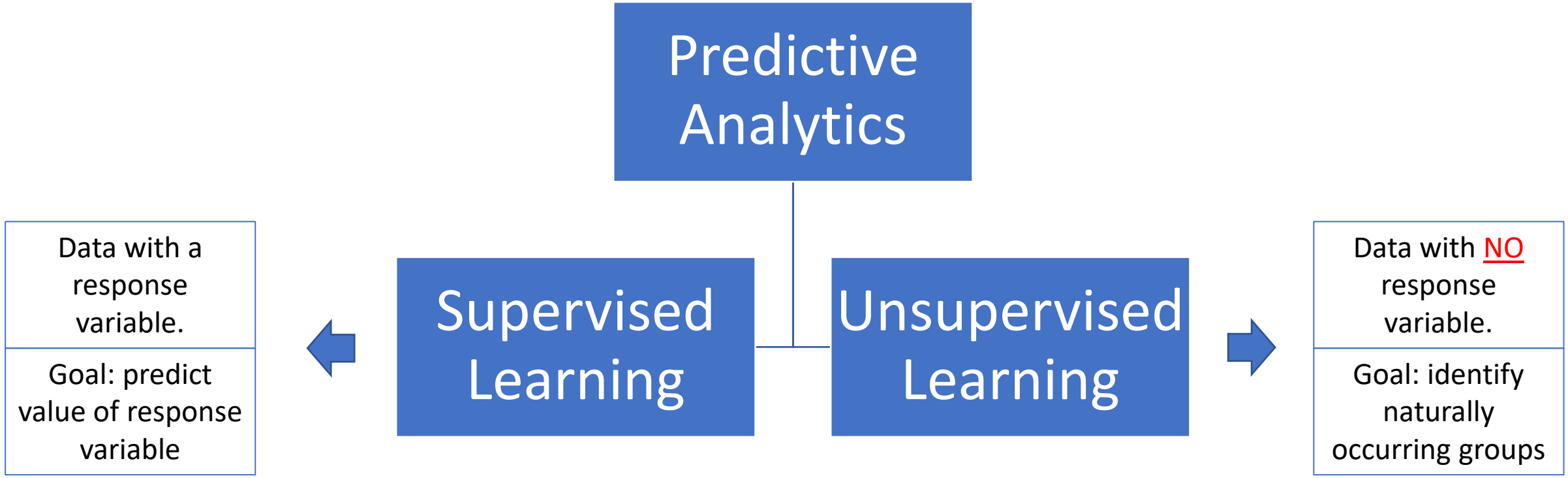
Response variable or dependent variable	Predictor variables or independent variables	Observations
• The variable you are trying to predict	• Those variables that can be related to changes in response variable	• Number of rows of a dataset

Major Concepts




Predictive Analytics
Supervised Learning
Unsupervised Learning
Classification
Regression
Principal Components Analysis
Clustering

Key Concept: Supervised vs. Unsupervised Learning



Supervised or Unsupervised?

Predict if a customer will churn or not

Estimate housing price given data on the characteristics of a house

Segment similar clients for a marketing campaign

Supervised or Unsupervised?

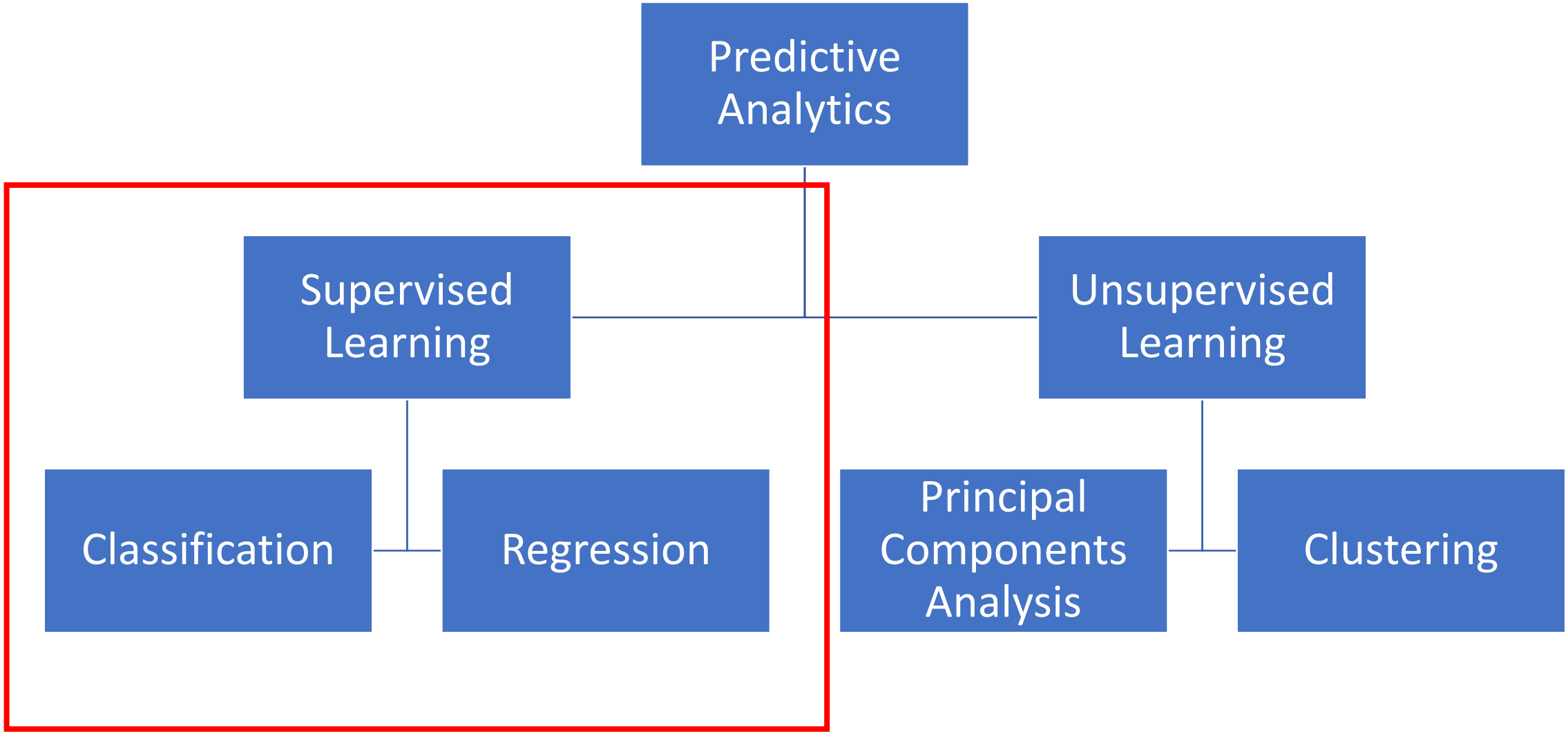
Predict if a customer will churn or not

Estimate housing price given data on the characteristics of a house

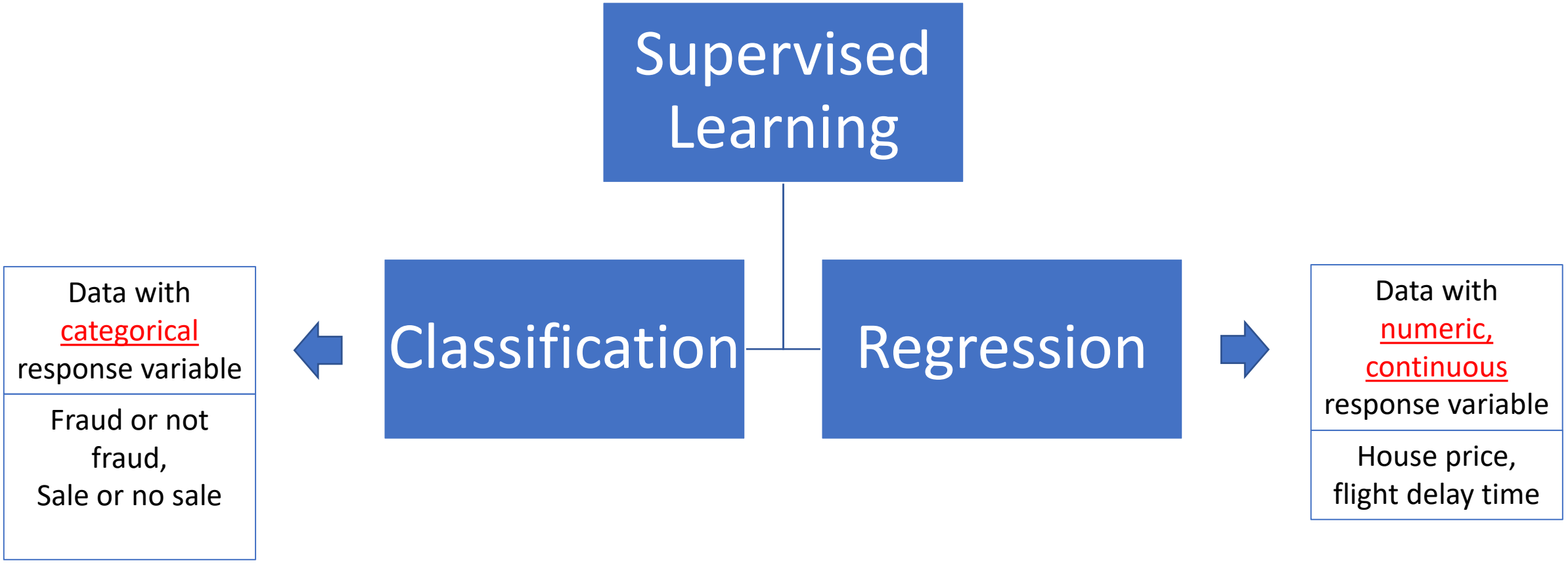
Supervised

Segment similar clients for a marketing campaign

Unsupervised


Predictive Analytics
Supervised Learning
Unsupervised Learning
Classification
Regression
Principal Components Analysis
Clustering

Classification vs. Regression



Classification or Regression?

Predict if a company will enter bankruptcy or not

Estimate housing price given data on the characteristics of a house

Predict sales given advertising dollars spent

Predict if an email is spam or not

Classification or Regression?

Predict if a company will enter bankruptcy or not

classification

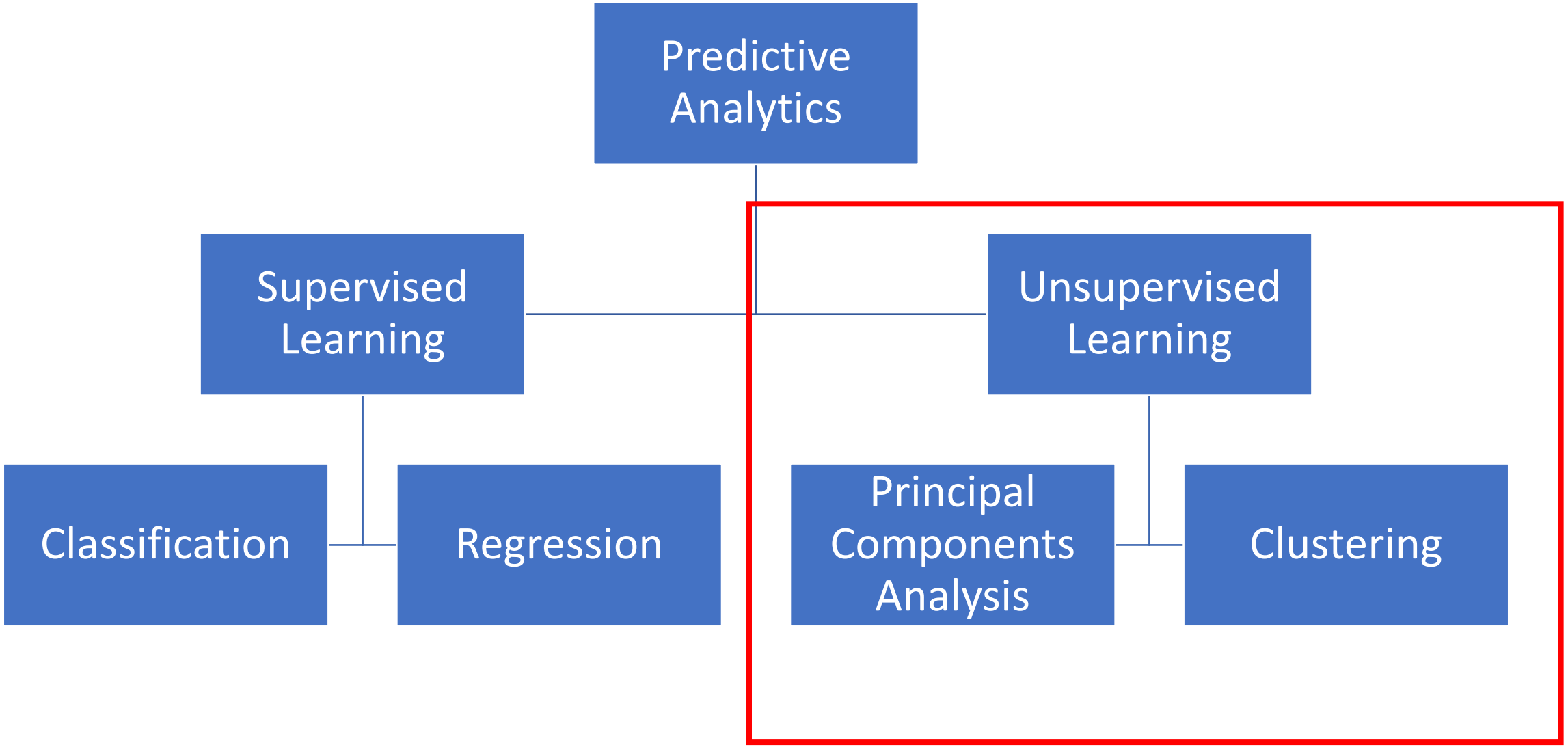
Estimate housing price given data on the characteristics of a house

Predict sales given advertising dollars spent

regression

Predict if an email is spam or not

classification


Predictive Analytics
Supervised Learning
Unsupervised Learning
Classification
Regression
Principal Components Analysis
Clustering

Key Concept: Building the statistical learning model vs using the results from the model

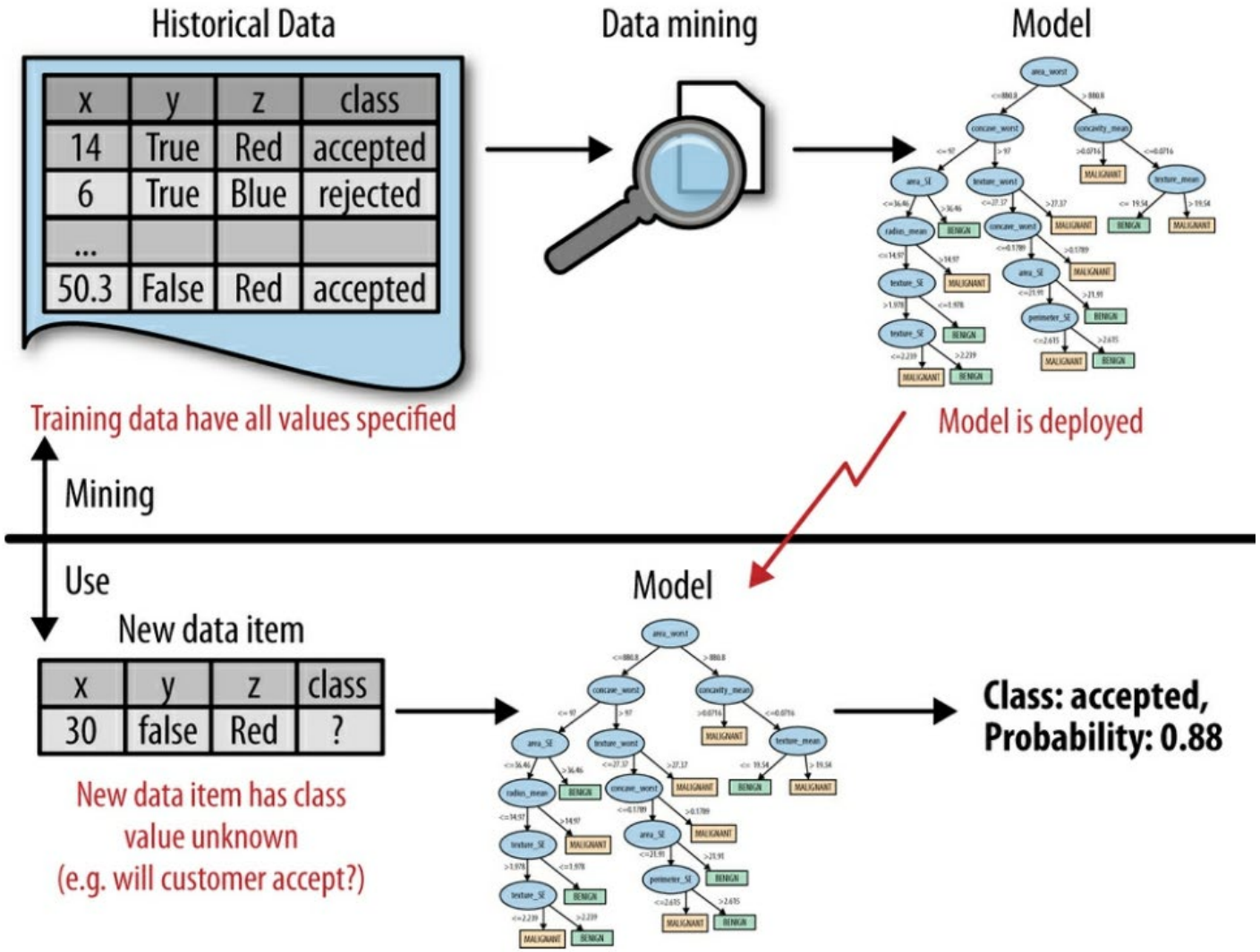
There is an important distinction between:

1. Building the model to find patterns
2. Using previously built model to predict new observations

*Often confused by students

Building the statistical learning model vs Using the Results from the model

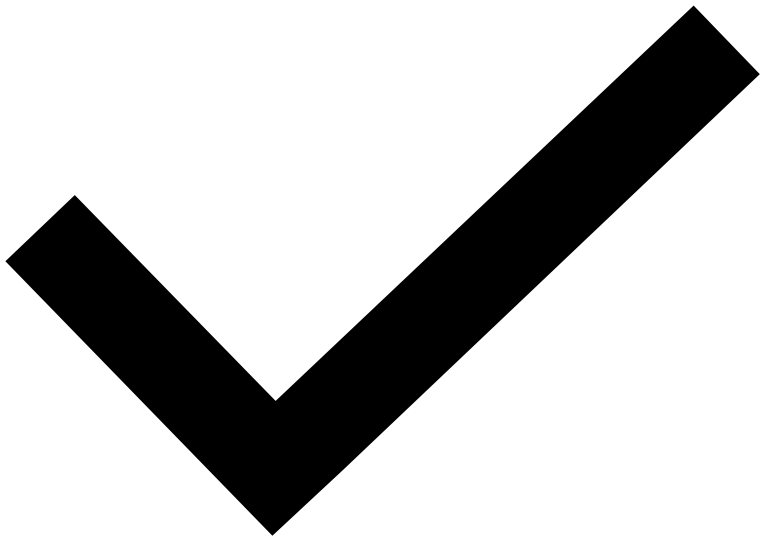
Building the statistical learning model



Using the Results from the model

In **supervised learning**, we have one main job in this class: build models that can predict new data well.

- Given this, how do we choose one model over another when we are interested in predicting new observations?



Evaluate Data Science Proposal: Customer Churn

Customer Churn

- Customer churn is one of the biggest problems telecom companies face
- Customer churn—losing customers to another company—directly impacts:
  - *revenue and growth* and acquiring more customers is *usually more expensive than retaining existing companies*

Customer Churn

To survive in highly competitive markets, almost every telecommunication company tracks consumers' data usage

Then predictive analytics is used to predict customers likely to churn

Then customers likely to churn are targeted with marketing to try and prevent churn

Question:
Supervised or unsupervised?
Classification or Regression?

Customer Churn

To survive in highly competitive markets, almost every telecommunication company tracks consumers' data usage

Then predictive analytics is used to predict customers likely to churn

Then customers likely to churn are targeted with marketing to try and prevent churn

Recall: this is supervised learning and classification

Data Science Proposal

- You’re a manager at TelCoOne, responsible for reducing customer churn
- My consulting team mines your historical data and builds a model to distinguish customers who are likely to churn based on some variables

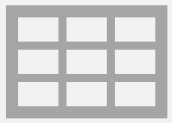
Source: Provost, Foster; Fawcett, Tom. Data Science for Business. O'Reilly Media.

Data Science Proposal

- Given my model, your team checks the performance of the model on the same historical data used to build the model (since experience tells you that churn patterns are relatively stable)
- The model is **100% accurate**
- Not a single mistake, identifying correctly all the churners as well as the nonchurners
- Are you comfortable with this result? Is this a lucky fluke? Why or why not?

Data Science Proposal

It’s not a fluke. My data science team can do this everytime!



How? And is it a useful model?

Here’s how my model works

Given a customer’s data…

Single	54	No	No	Monthly contract	$ 33.75	$ 1,782.00

The model searches for the row with the same data. If it’s found then predict churn if not found predict no churn.

historical data of customers who churn

Marital_Status	Term	Phone_service	Internatio	Agreement_period	Monthly_Charges	Total_Charge	Churn
Married	16	Yes	Yes	Monthly contract	$ 98.05	$ 1,410.25	Yes
Married	36	Yes	Yes	Monthly contract	$ 94.65	$ 3,266.00	Yes
Married	5	Yes	Yes	Monthly contract	$ 97.10	$ 497.55	Yes
Married	42	Yes	Yes	One year contract	$ 95.45	$ 3,944.50	Yes
Married	43	Yes	No	Monthly contract	$ 104.15	$ 5,067.45	Yes
Single	54	No	No	Monthly contract	$ 33.75	$ 1,782.00	Yes
Married	2	Yes	No	Monthly contract	$ 76.00	$ 81.25	Yes

Here’s how my model works

Given a customer’s data…

Single	54	No	No	Monthly contract	$ 33.75	$ 1,782.00

The model searches for the row with the same information from the historical data. If it’s found then predict churn if not found predict no churn.

historical data of customers who churn

Marital_Status	Term	Phone_service	Internatio	Agreement_period	Monthly_Charges	Total_Charge	Churn
Married	16	Yes	Yes	Monthly contract	$ 98.05	$ 1,410.25	Yes
Married	36	Yes	Yes	Monthly contract	$ 94.65	$ 3,266.00	Yes
Married	5	Yes	Yes	Monthly contract	$ 97.10	$ 497.55	Yes
Married	42	Yes	Yes	One year contract	$ 95.45	$ 3,944.50	Yes
Married	43	Yes	No	Monthly contract	$ 104.15	$ 5,067.45	Yes
Single	54	No	No	Monthly contract	$ 33.75	$ 1,782.00	Yes
Married	2	Yes	No	Monthly contract	$ 76.00	$ 81.25	Yes

Predict this customer churns

Here’s how my model works

Given another customer’s data…

Single	68	No	No	Two year contract	$ 40.05	$ 3,260.10

The model searches for the row with the same information from the historical data. If it’s found then predict churn if not found predict no churn.

historical data of customers who churn

Marital_Status	Term	Phone_service	Internatio	Agreement_period	Monthly_Charges	Total_Charge	Churn
Married	16	Yes	Yes	Monthly contract	$ 98.05	$ 1,410.25	Yes
Married	36	Yes	Yes	Monthly contract	$ 94.65	$ 3,266.00	Yes
Married	5	Yes	Yes	Monthly contract	$ 97.10	$ 497.55	Yes
Married	42	Yes	Yes	One year contract	$ 95.45	$ 3,944.50	Yes
Married	43	Yes	No	Monthly contract	$ 104.15	$ 5,067.45	Yes
Single	54	No	No	Monthly contract	$ 33.75	$ 1,782.00	Yes
Married	2	Yes	No	Monthly contract	$ 76.00	$ 81.25	Yes

Not in table, predict this customer doesn’t churn

But is this model useful?

- It memorizes the historical data it was built on…
- What is the problem with this?

But is this model useful?

- It memorizes the data it was built on…
- What is the problem with this?
  - For ALL new customers that are not part of the historical dataset, the model will predict not churn for EVERYONE, since these new customers are not in the table.
  - Model is terrible at predicting new observations!

A model that looked perfect would be completely useless in practice with new data!

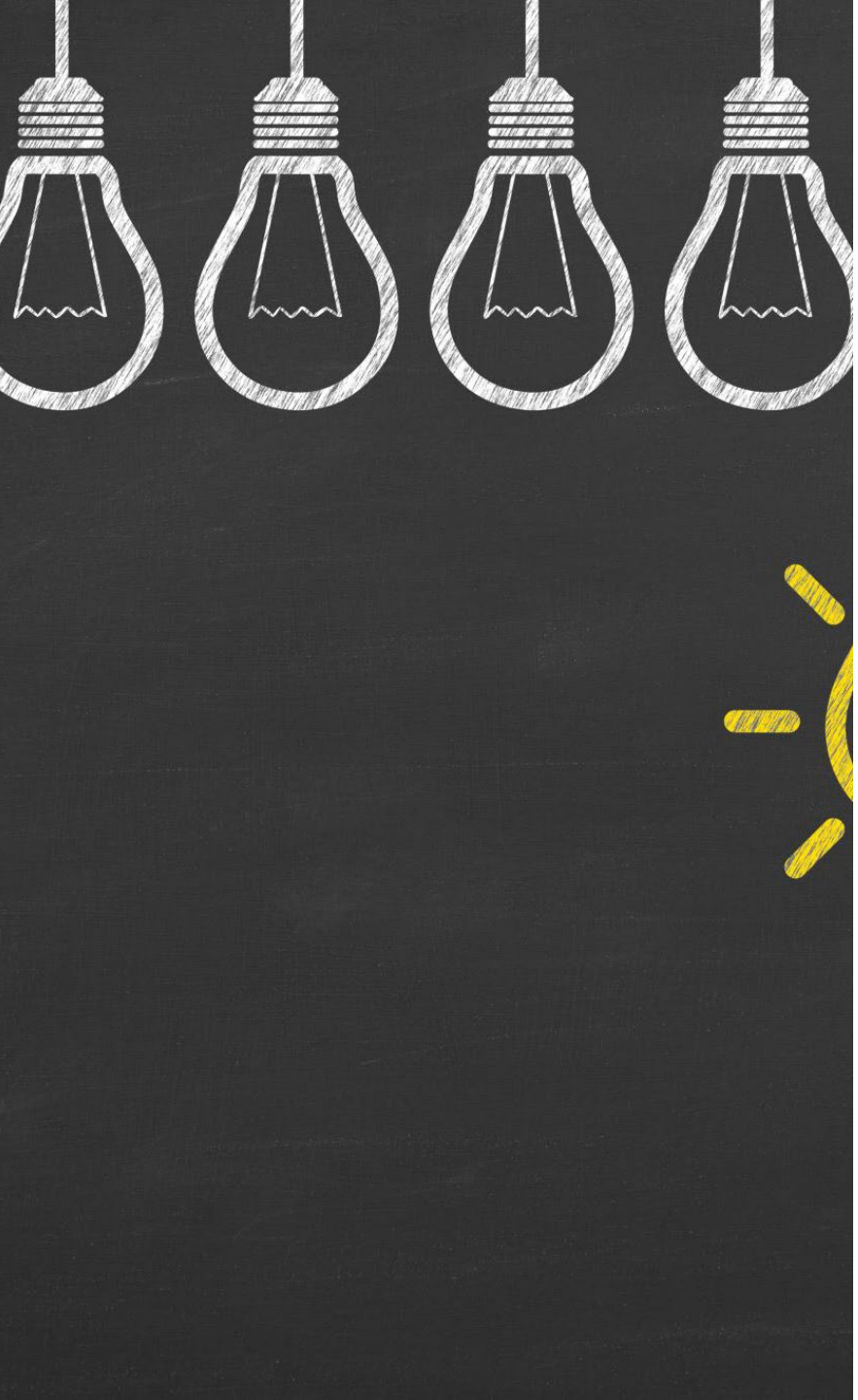
Here’s another model used

- Gradient boosted decision trees (e.g., XGBoost)
- Very flexible and powerful predictive model

- Again we evaluate the performance on the same data we build the model with and get very close to 100% accuracy. Thoughts?

But is this approach useful?

- Since this is a very **flexible** model, it can still tend to memorize some randomness in the data it was built on!
- If we evaluate this model on the same data it was built on, the performance will be **extremely good**.
- But how does it perform on new data the model hasn't seen before?



Takeaway

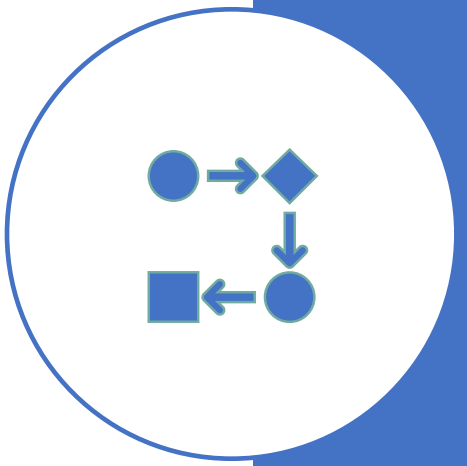
- We need an approach that can evaluate how well a model generalizes to new data.
- **Overfitting** occurs when a flexible model finds chance or irrelevant occurrences in the data it was built on but does not perform well on new datasets.
  - If we allow the model enough flexibility in searching for patterns in a particular dataset, it will find patterns. Unfortunately, these “patterns” may be just chance occurrences in the data causing the approach to perform poorly on unseen data.

How can we assess if a model generalizes well to new data??

Key Concept: Partitioning Data to Avoid Overfitting and Pursue Generalization

1. Split the data into two parts
2. Build each model on the ***training set***
3. Use this previously built model to get predictions for the ***testing set*** and evaluate the performance of these predictions on the testing set

- Why?
  - Model may perform well the dataset it is built on but fails to accurately predict on a new dataset.
  - We minimize this risk using both a train and test set.
  - Simulate real world environment where we don't know anything about the data the final model will be applied to



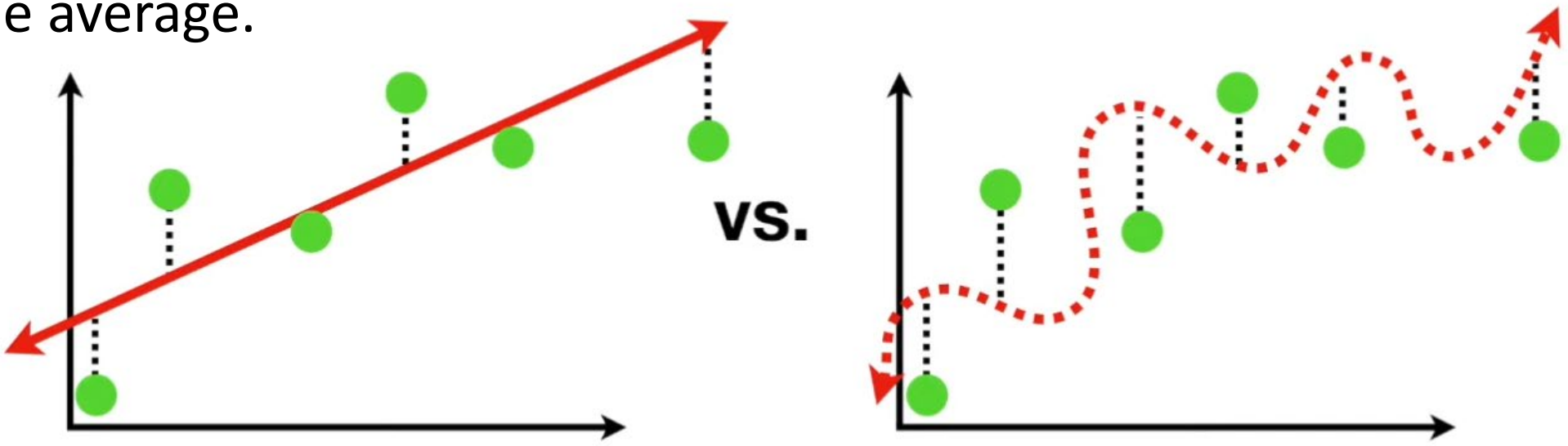
Note: there are other more advanced ways to partition data that we will not cover in this course.

Now, how to evaluate performance for regression problems?

No one method can dominate all others over all possible data sets. Selecting the best approach can be challenging… we next discuss how to evaluate performance on the testing set.

Key Concept: How to Evaluate performance for Regression Problems?

Using the mean squared error (average squared error) on the testing set. That is, we get the sum of squared differences from the model to the points then take the average.

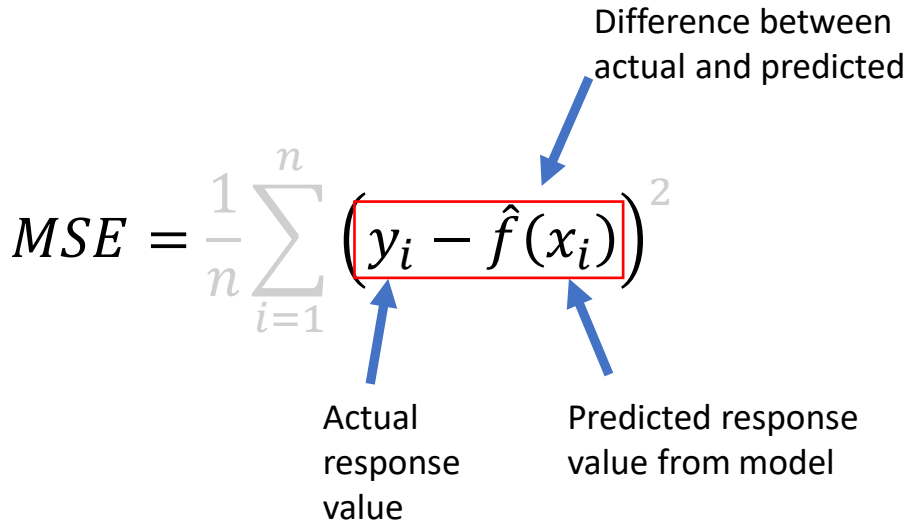


We want the smallest testing MSE as possible.

Key Concept: How to Evaluate Performance for Regression Problems?

We need some way to measure how well any model $\hat{f}(x)$ predicts the response.

- **Mean squared error** (also known as average squared error) on the testing set.



Key Concept: How to Evaluate Performance for Regression Problems?

We need some way to measure how well any model $\hat{f}$ predicts the response.

- **Mean squared error** (also known as average squared error) on the testing set.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

Squared so differences don't cancel each other out

Average over all observations.

Actual response value

Predicted response value from model

*We will cover how to determine the best model for classification problems next week.

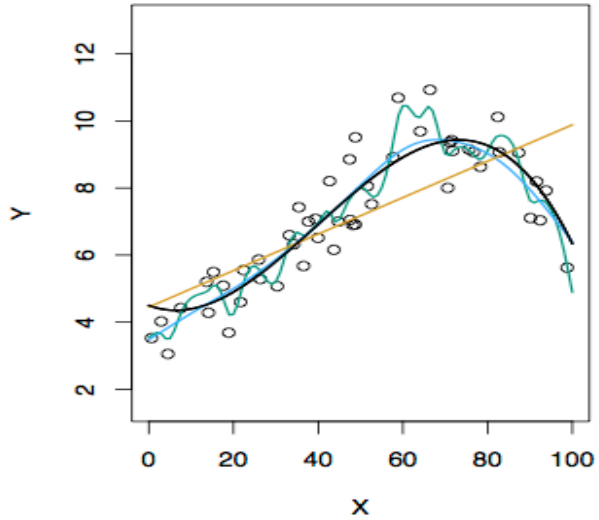
We want the smallest testing MSE as possible.

Training MSE vs. Test MSE

- There is no guarantee that the method with the smallest training MSE will have the smallest test (i.e. new data) MSE.
- In general the more flexible a method is the lower its training MSE will be i.e., it will “fit” or explain the training data very well.
- However, the test MSE may in fact be higher for a more flexible method than for a simple approach like linear regression.

Example Training and Testing MSE Evaluation of different models

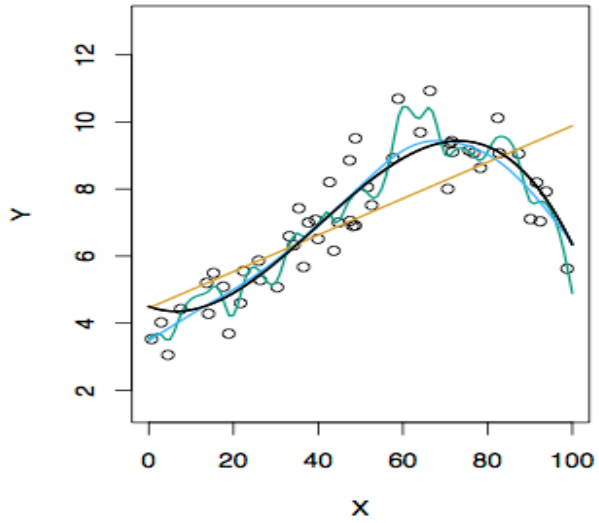
Dataset 1

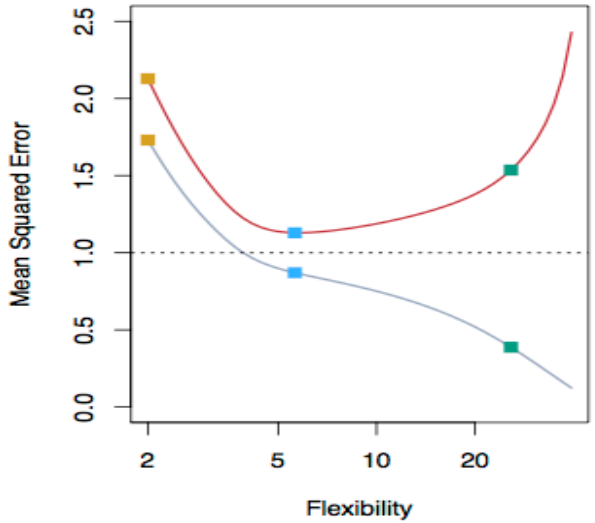


<u>Different models</u>
Black: Truth
Orange:
Blue:
Green:

Which color is most flexible (e.g., most wiggly or allowed the most turns)?

Dataset 1





LEFT

Black: Truth

Orange: Least flexible

Blue: More flexible

Green: Most flexible

RIGHT

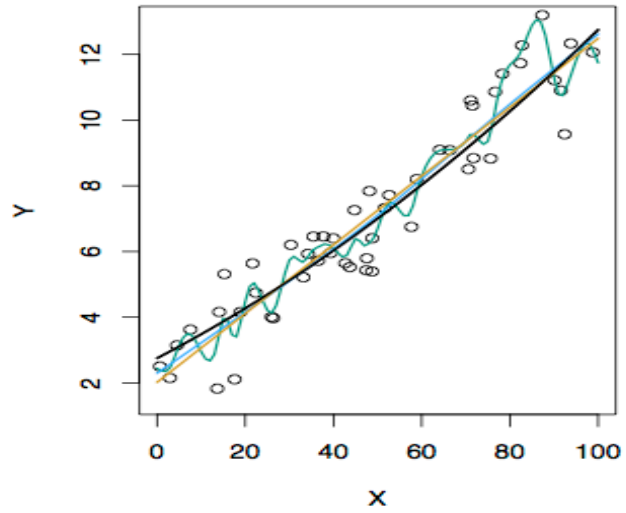
RED: Test MSE

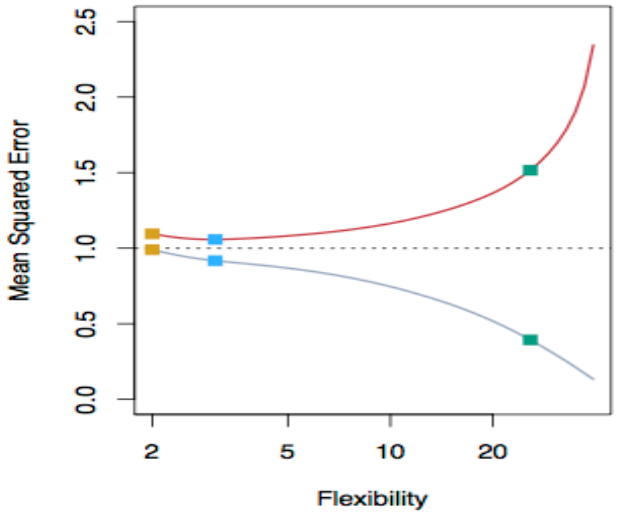
Grey: Training MSE

Dashed: Irreducible error

1. What metric (red,grey or dashed) do we use to determine the best model?
2. Which model performs the best?

Dataset 2





Here the truth is smoother, so the less flexible models do really well.

LEFT

Black: Truth

Orange: Least flexible

Blue: More flexible

Green: Most flexible

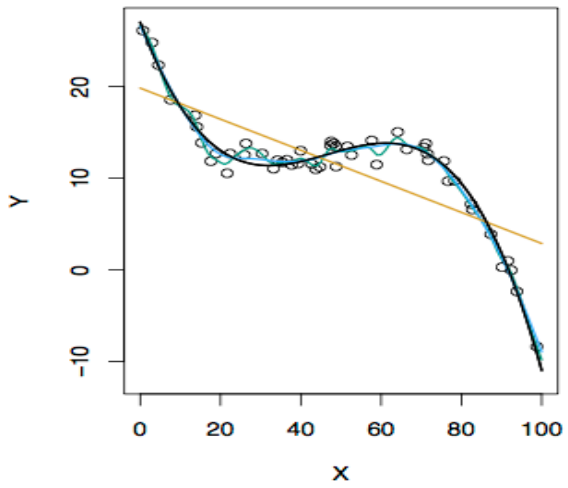
RIGHT

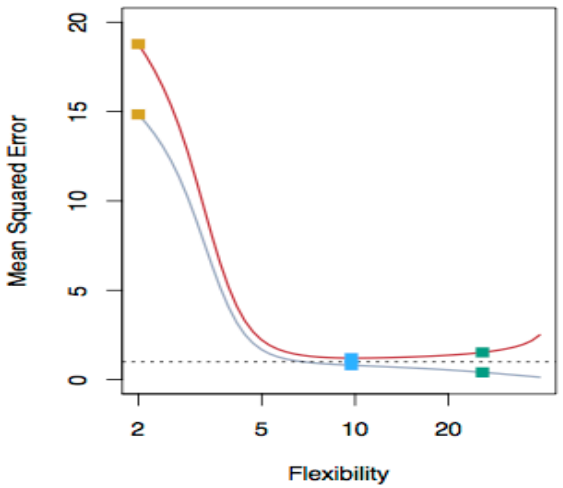
RED: Test MSE

Grey: Training MSE

Dashed: Irreducible error

Dataset 3





Here the truth is wiggly and the noise is low, so the more flexible fits do best.

LEFT

Black: Truth

Orange: Least flexible

Blue: More flexible

Green: Most flexible

RIGHT

RED: Test MSE

Grey: Training MSE

Dashed: Irreducible error

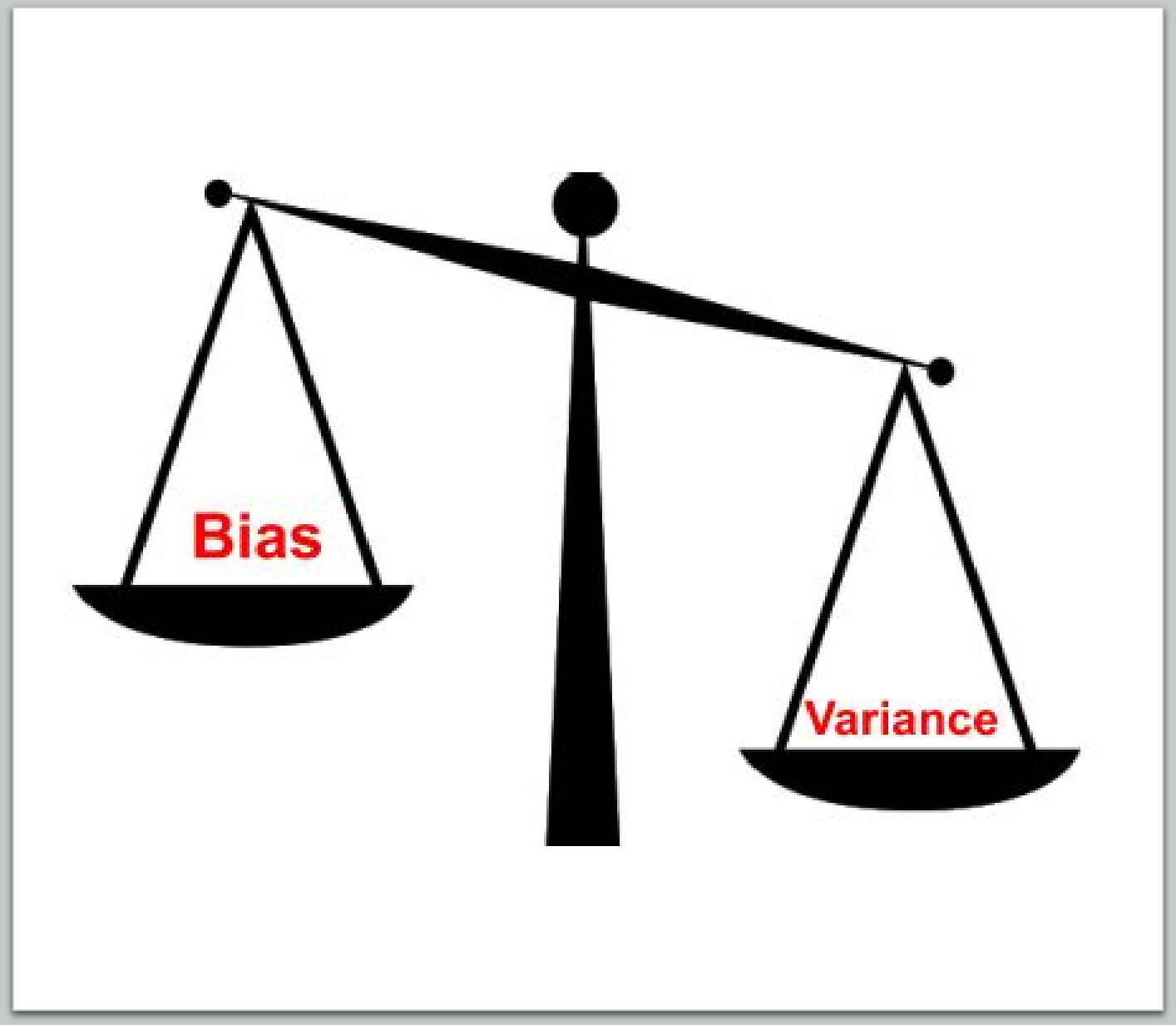
The Test MSE (how we choose the best model) is made up of 2 competing forces…

Key Concept: Bias Variance Tradeoff

- The **test MSE** is made up of two competing forces that govern the choice of learning method:

1. the **variance** of predictions
2. **bias** of predictions

* There will always be irreducible error (model can't reduce this at all)



Bias

- Bias refers to the error that is introduced by modeling a real-life problem (that is usually extremely complicated) by a much simpler problem.
  - In other words, how well your model fits the data.

- For example, linear regression assumes that there is a linear relationship between Y and X. It is unlikely that, in real life, the relationship is exactly linear so some bias will be present.

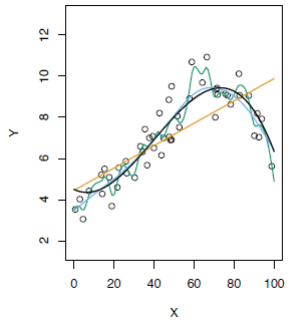
- **The more flexible/complex a method is the less bias it will generally have.**

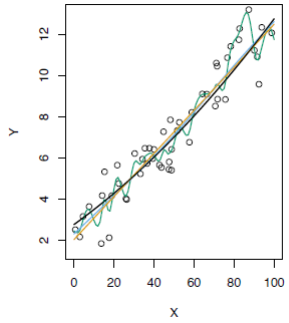
Variance

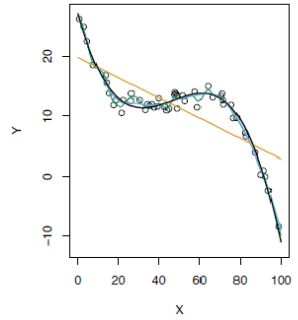
- Variance refers to how much your model and predictions change if you built your model on a different data set.

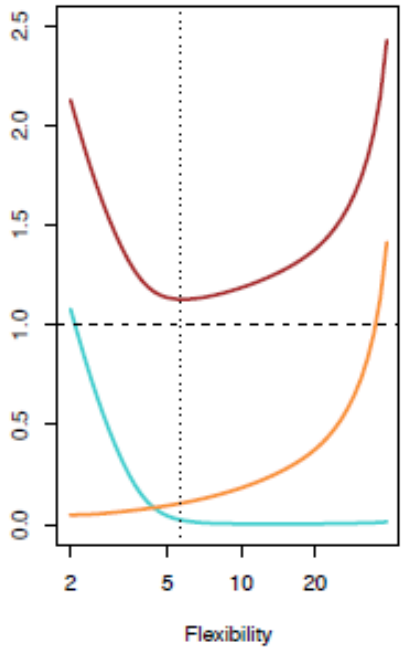
- **Generally, the more flexible a method is the more variance it has.**

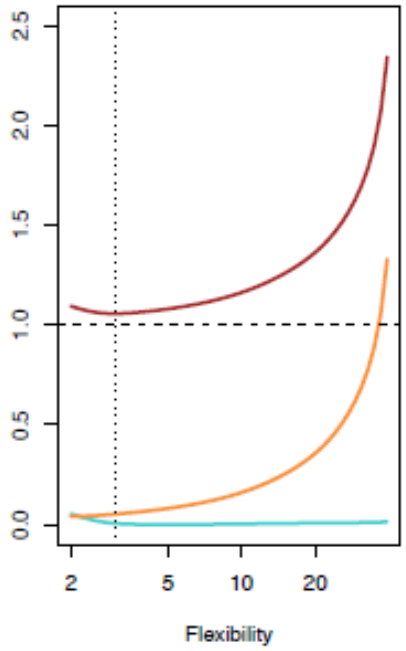
Test MSE, Bias, and Variance

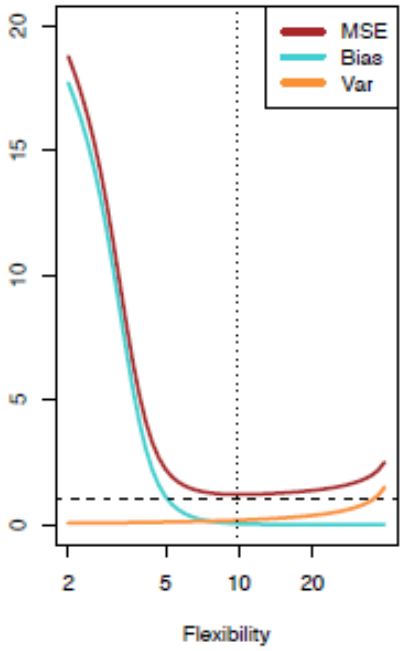












What trends do you notice for bias and variance?

*Dotted horizontal line is irreducible error

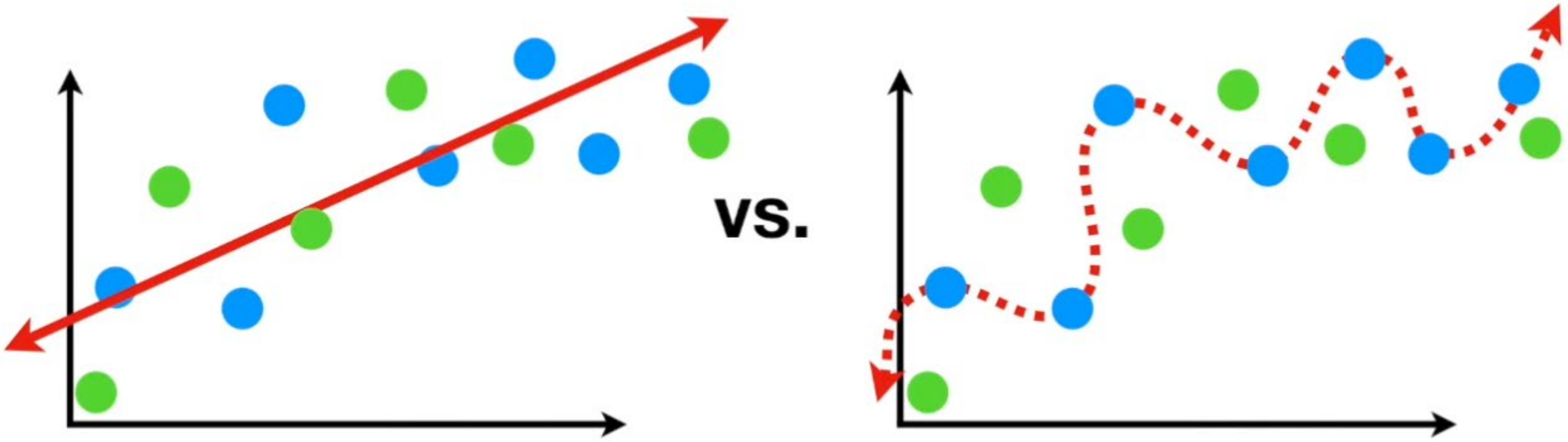
Our Goal

- Best model is the sweet spot of low bias and relatively low variance.
  - As the flexibility of the model increases, its variance increases, and its bias decreases.
- Three commonly used methods for finding the sweet spot between simple and complicated models are: random forest, LASSO, and boosting.
- We will cover each of these methods in the future.

Example: Bias/Variance Tradeoff with Training and Testing sets: Which Model to Choose?

Linear Regression Model

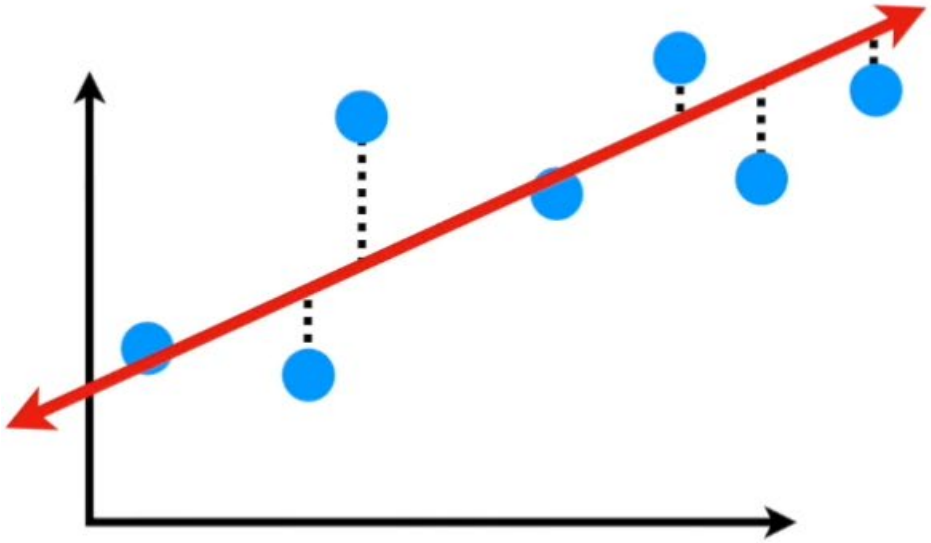
Very Flexible Model: Squiggly Model



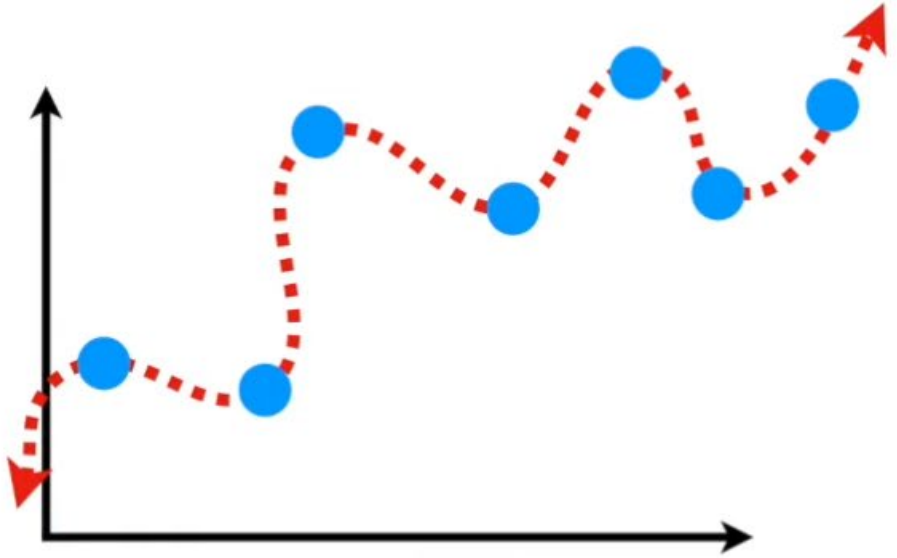
Blue dots are training data and green dots are testing data.

Bias

High bias: model doesn't fit the data well.



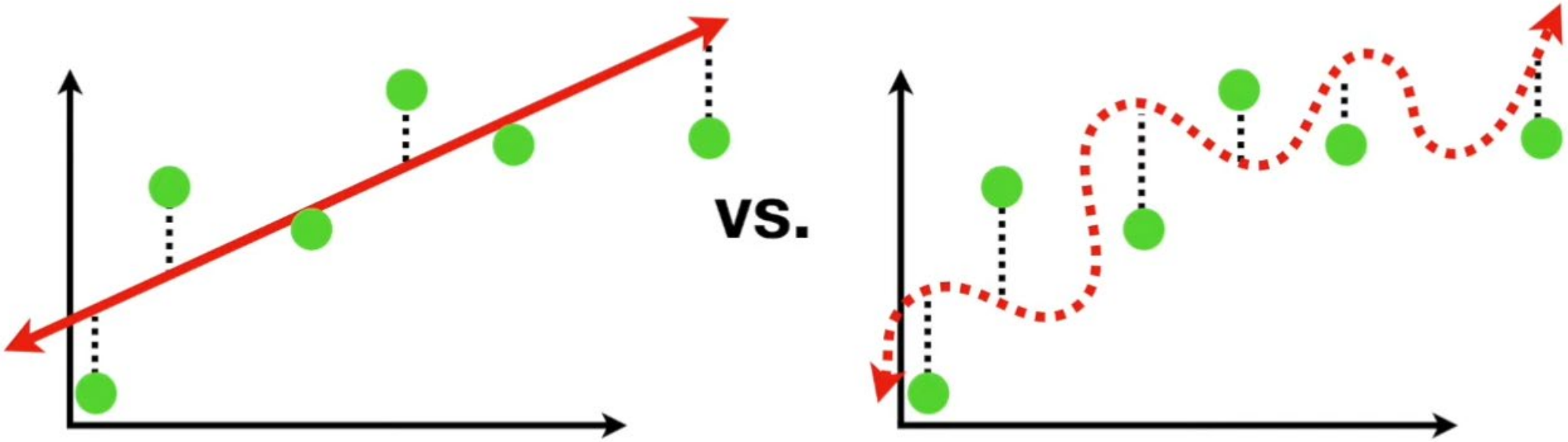
Low bias: model fits the data very well.



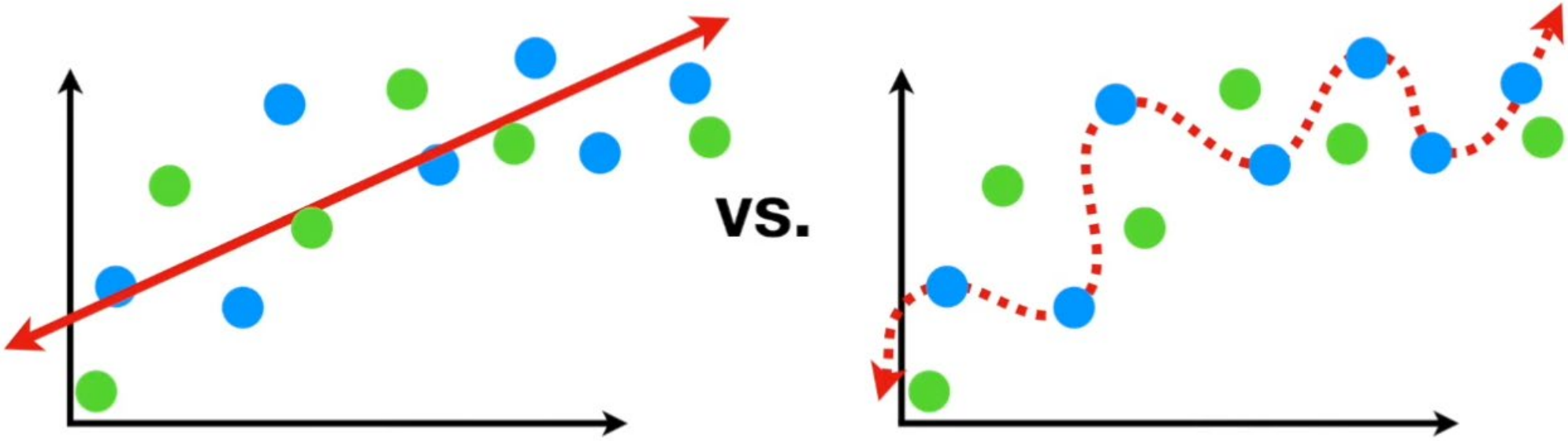
Variance

Lower variance: model fit on one dataset performs about the same on new data.

High variance: model fit on one dataset performs much worse on new data.



Which Model to Choose?



Our Goal

- Best model is the sweet spot of low bias and relatively low variability.