Course Announcements

Due Sunday (11:59 PM):

- D7
- Q8
- Weekly Project Survey (*optional*)

Notes:

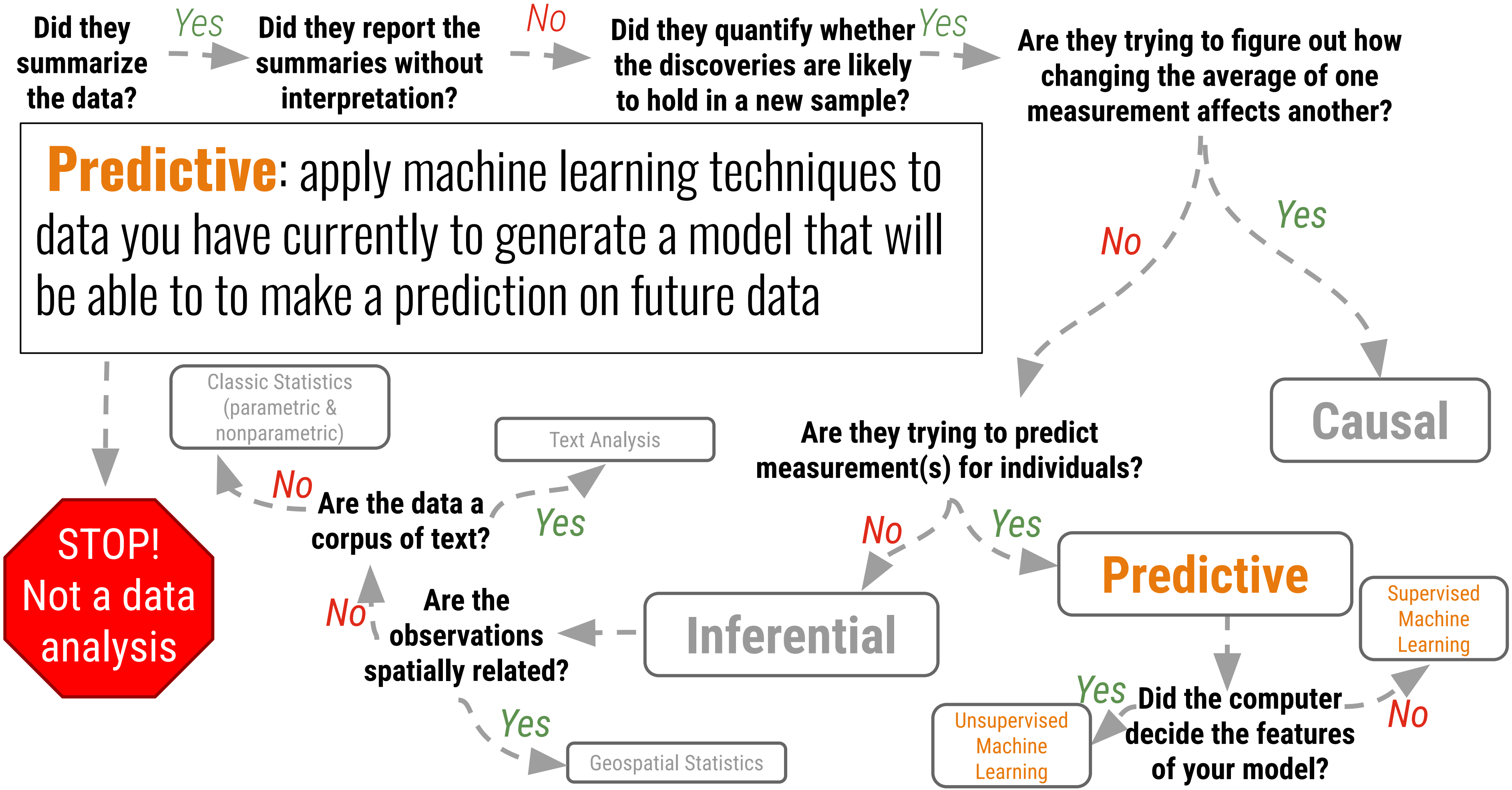
- A3 and D6 grades posted
- Data Checkpoint grading underway (discuss)
- update: A4 now available; due *Wed of week 10* (3/13)
    - Note: “Validate” will fail; use Kernel > Restart & Run All instead

Machine Learning I

Sid Joshi
UC San Diego

Department of Cognitive Science
s1joshi@ucsd.edu

Predictive: apply machine learning techniques to data you have currently to generate a model that will be able to to make a prediction on future data



Did they summarize the data?

Did they report the summaries without interpretation?

Did they quantify whether the discoveries are likely to hold in a new sample?

Are they trying to figure out how changing the average of one measurement affects another?

Are they trying to predict measurement(s) for individuals?

Causal

Predictive

Inferential

Are the observations spatially related?

Are the data a corpus of text?

Did the computer decide the features of your model?

Supervised Machine Learning

Unsupervised Machine Learning

Classic Statistics (parametric & nonparametric)

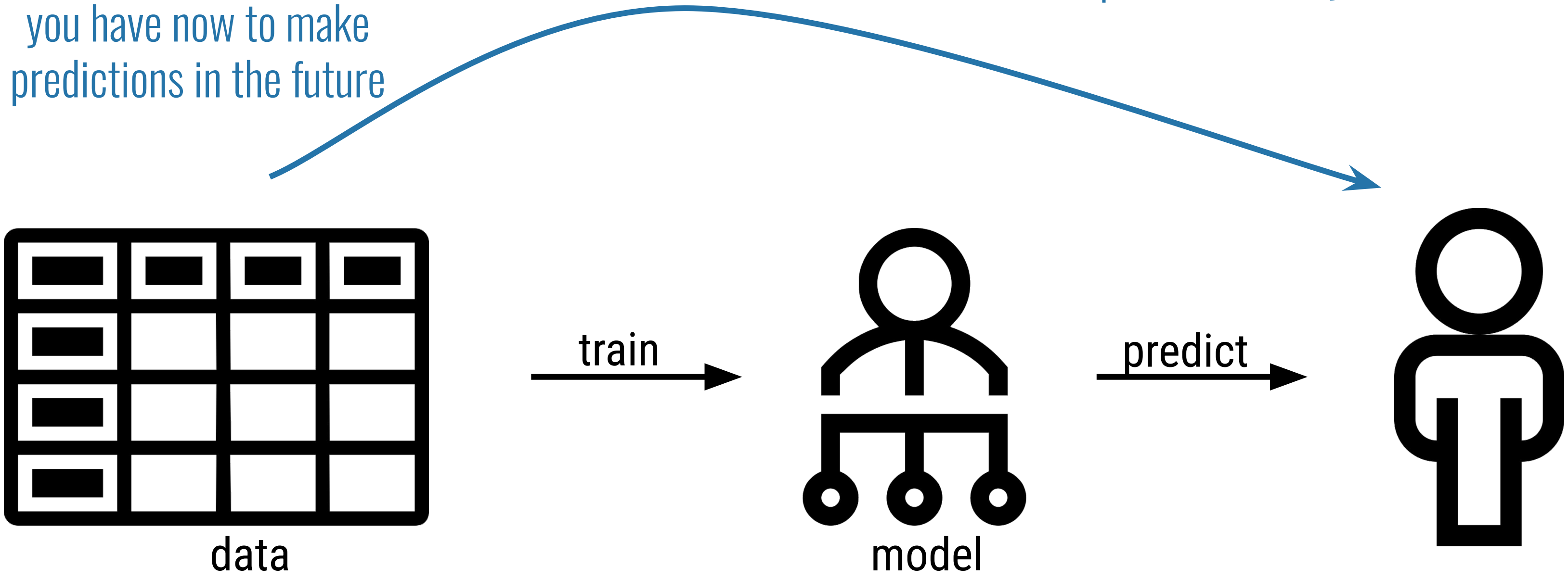
Text Analysis

Geospatial Statistics

STOP! Not a data analysis

predictive analysis uses data you have now to make predictions in the future

machine learning approaches are used for predictive analysis!



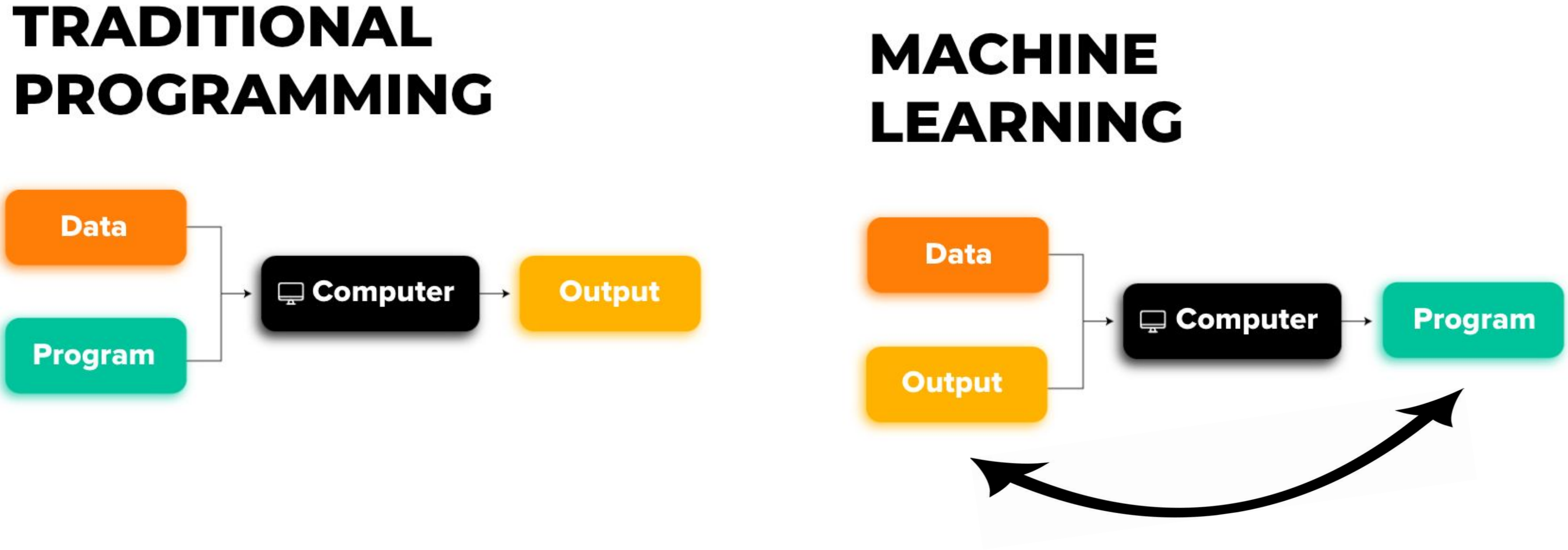
What is machine learning?

“Machine learning is the science of getting computers to act without being explicitly programmed”

- Andrew Ng, Stanford, ex-Google, chief scientist at Baidu, Coursera founder, Stanford Adjunct Faculty

adapted from Brad Voytek

A Shift in Programming Paradigm



TRADITIONAL PROGRAMMING

Data

Program

Computer

Output

MACHINE LEARNING

Data

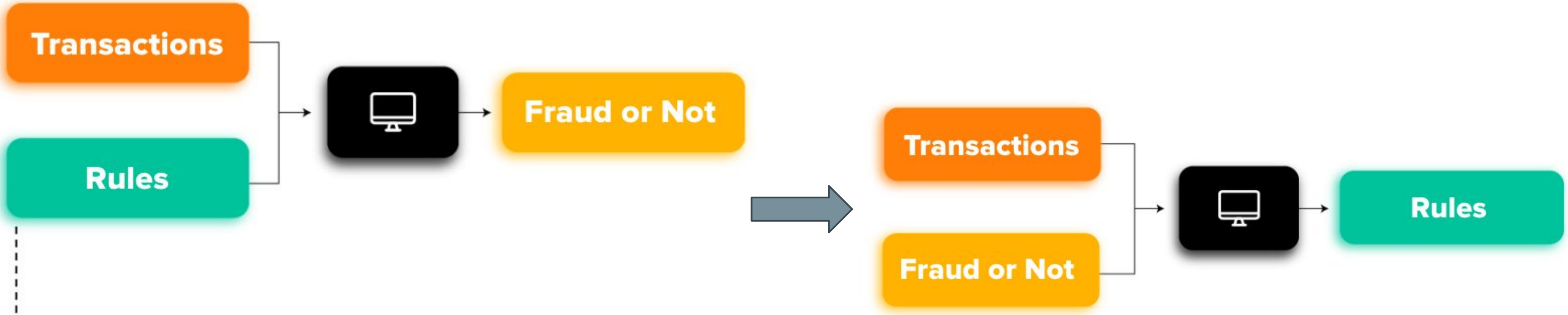
Output

Computer

Program

https://www.avenga.com/magazine/machine-learning-programming/

- **Problem:** Detecting whether credit card charges are fraudulent.
- **Data science question:** Can we use the time of the charge, the location of the charge, and the price of the charge to predict whether that charge is fraudulent or not?
- **Type of analysis:** Predictive analysis



Transactions

Rules

Fraud or Not

Transactions

Fraud or Not

Rules

Rule1. Claim time - Submit time < 1 h
Rule2. Agreement review time > 5 m
Rule3. ...

Prediction Questions

Which of these questions is most appropriate for machine learning?

A How common is watching Sesame Street in the US?

B What is the effect of watching Sesame Street on children's brains?

C What is the relationship between early childhood educational programming and success in elementary school?

D Can we use information about one's early childhood to predict their success in elementary school?

E How does Sesame Street cause an increase in educational attainment?

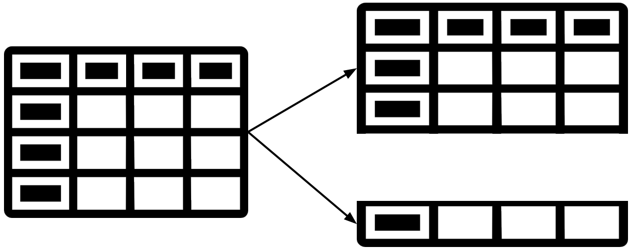
Machine Learning Generalizations

All models are wrong
but some are useful

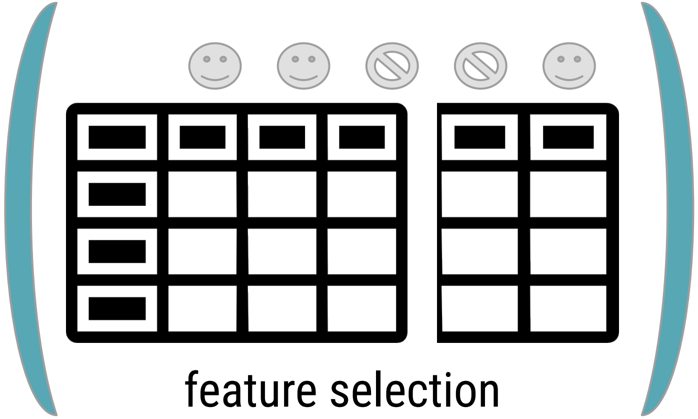
George E.P. Box

The goal of modeling is to simplify, not replicate

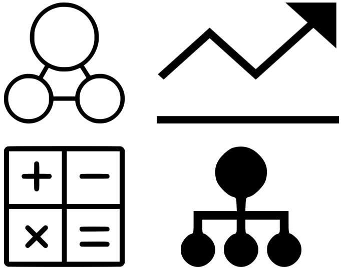
Basic Steps to Prediction



data partitioning

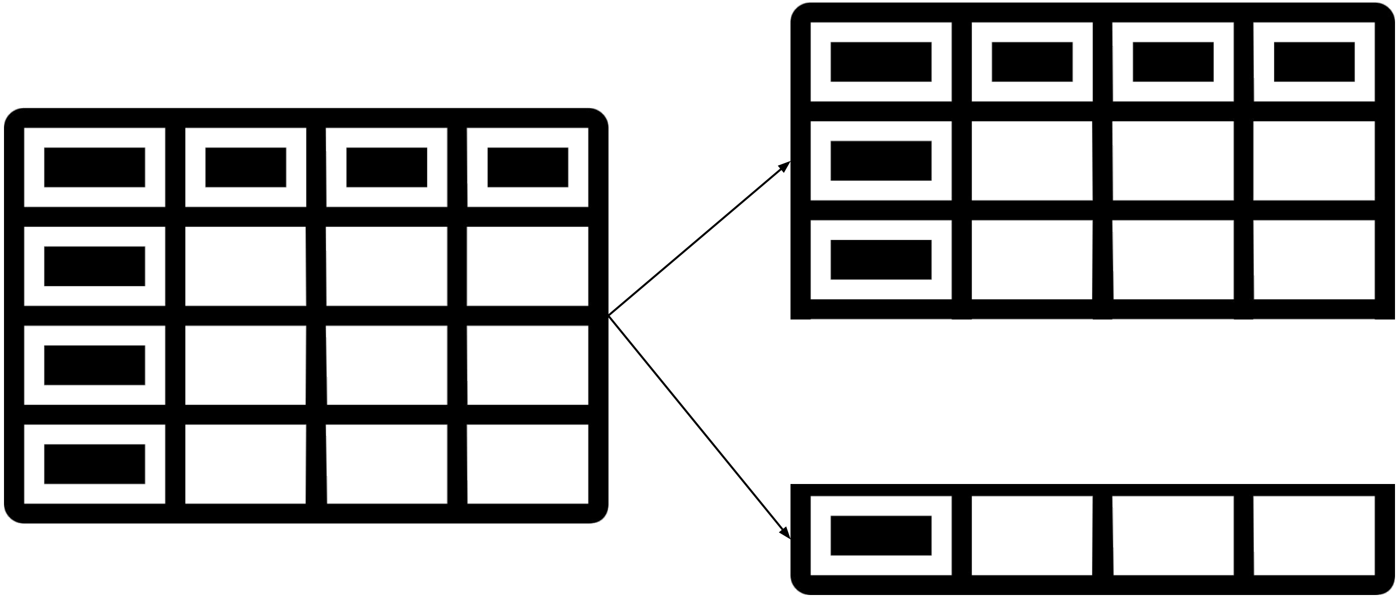


feature selection



model selection

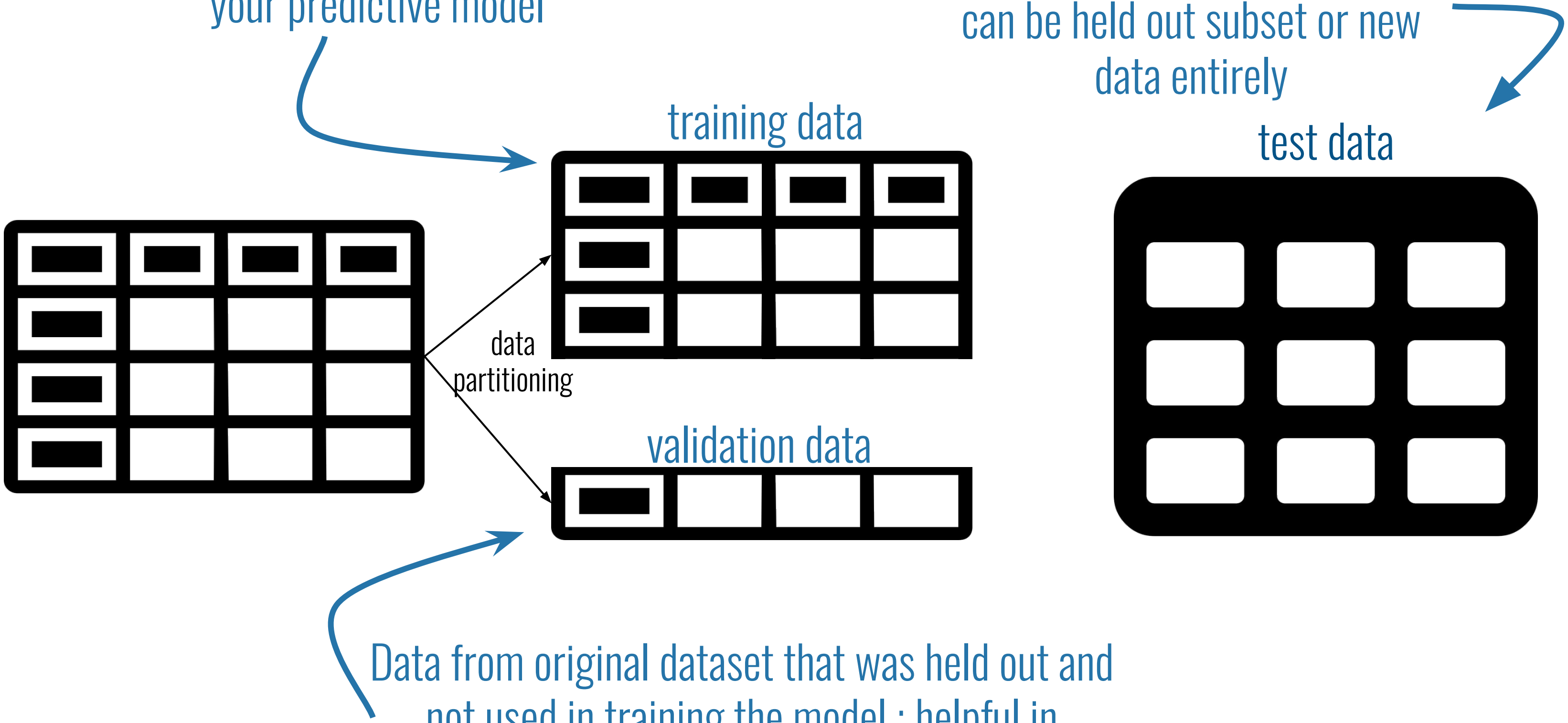
model assessment



data partitioning

the data used to build your predictive model

Data set used to assess if prediction model is generalizable; can be held out subset or new data entirely



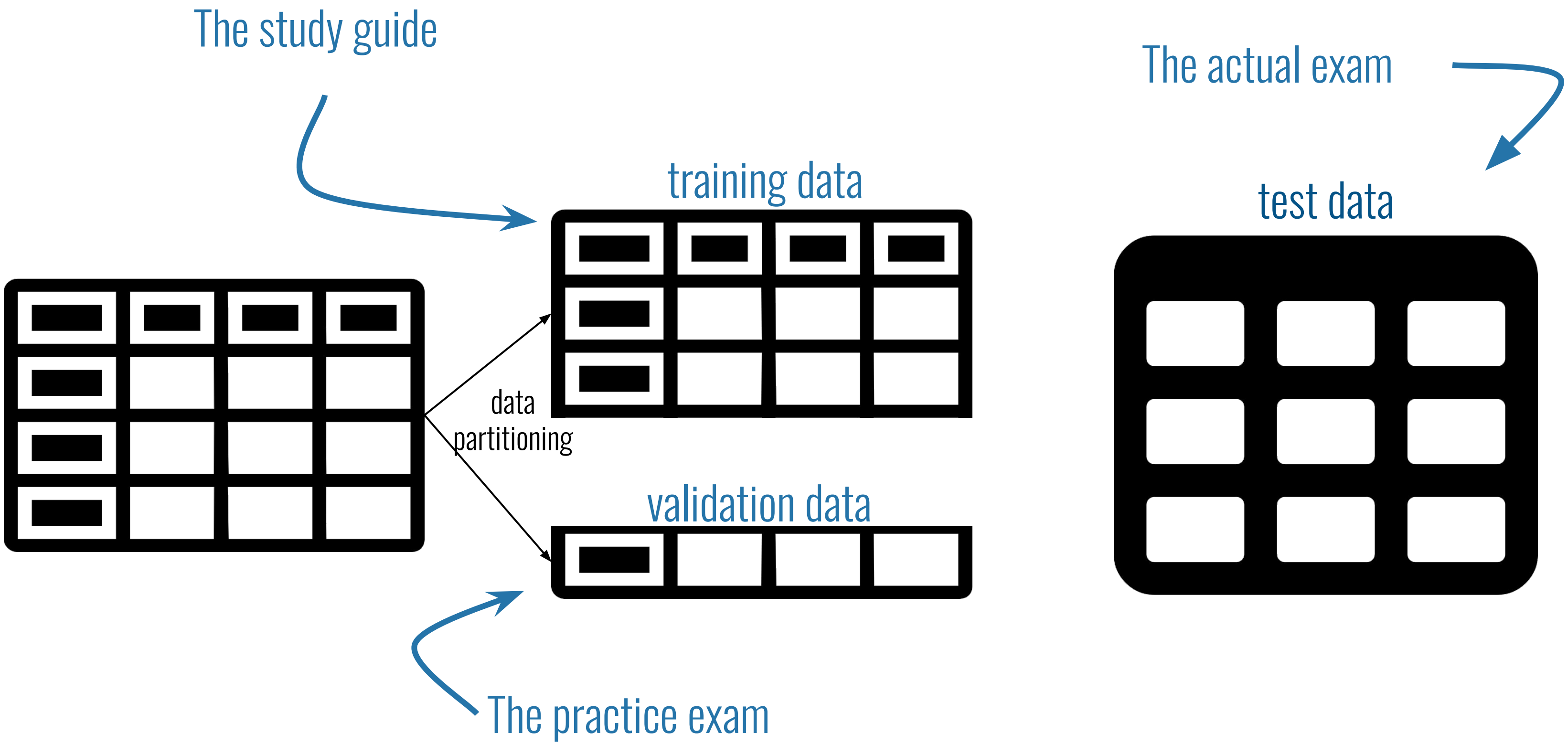
training data

test data

data partitioning

validation data

Data from original dataset that was held out and not used in training the model ; helpful in fine-tuning prediction accuracy



The study guide

training data

data partitioning

validation data

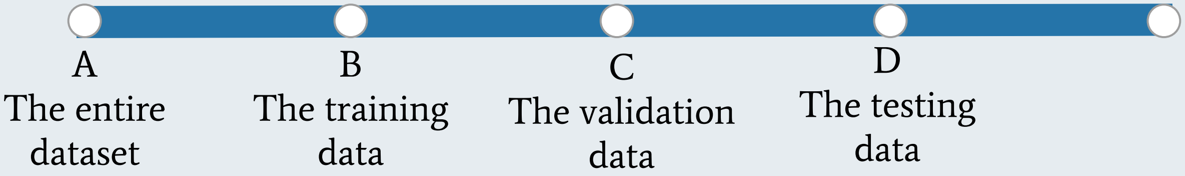
The practice exam

The actual exam

test data

Data Partitioning

What portion of the data are typically used for generating the model?

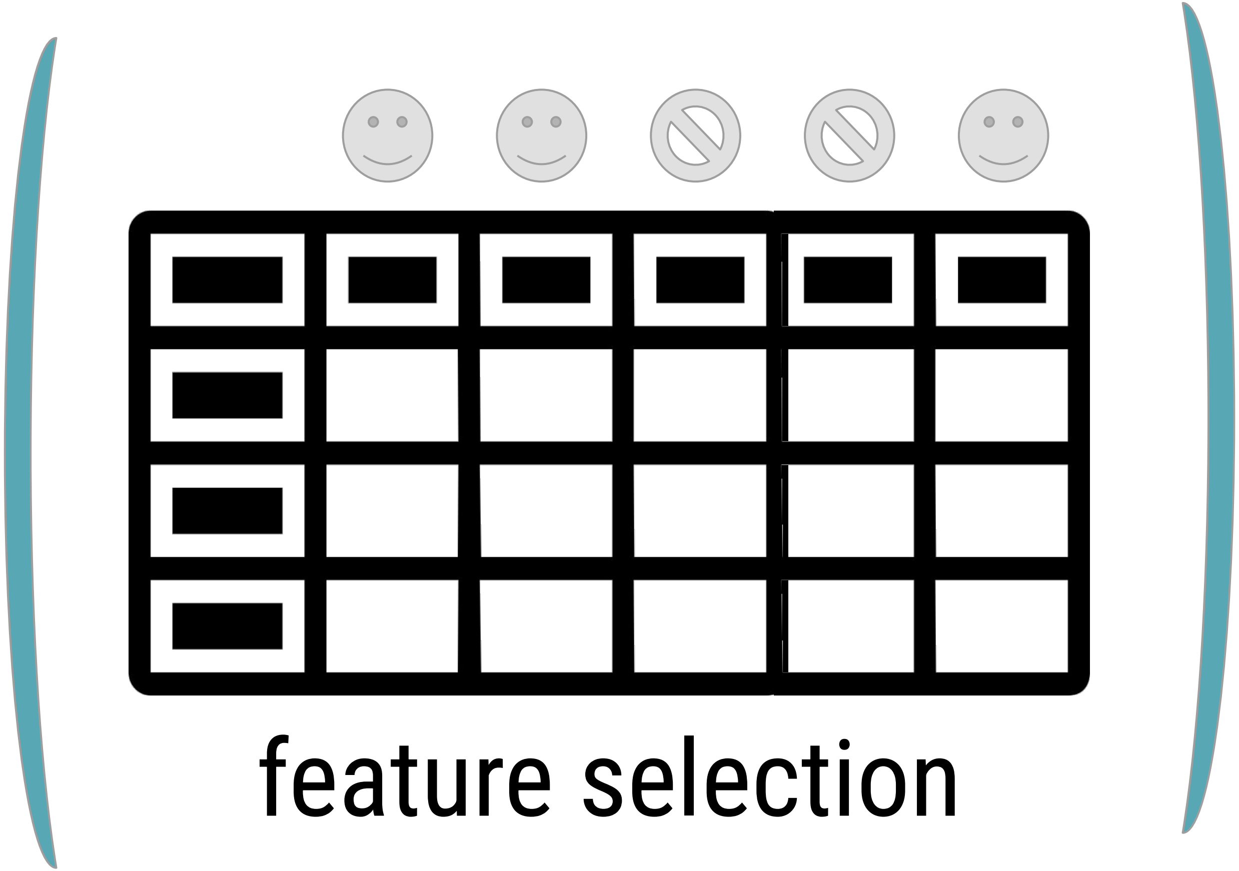


A
The entire dataset

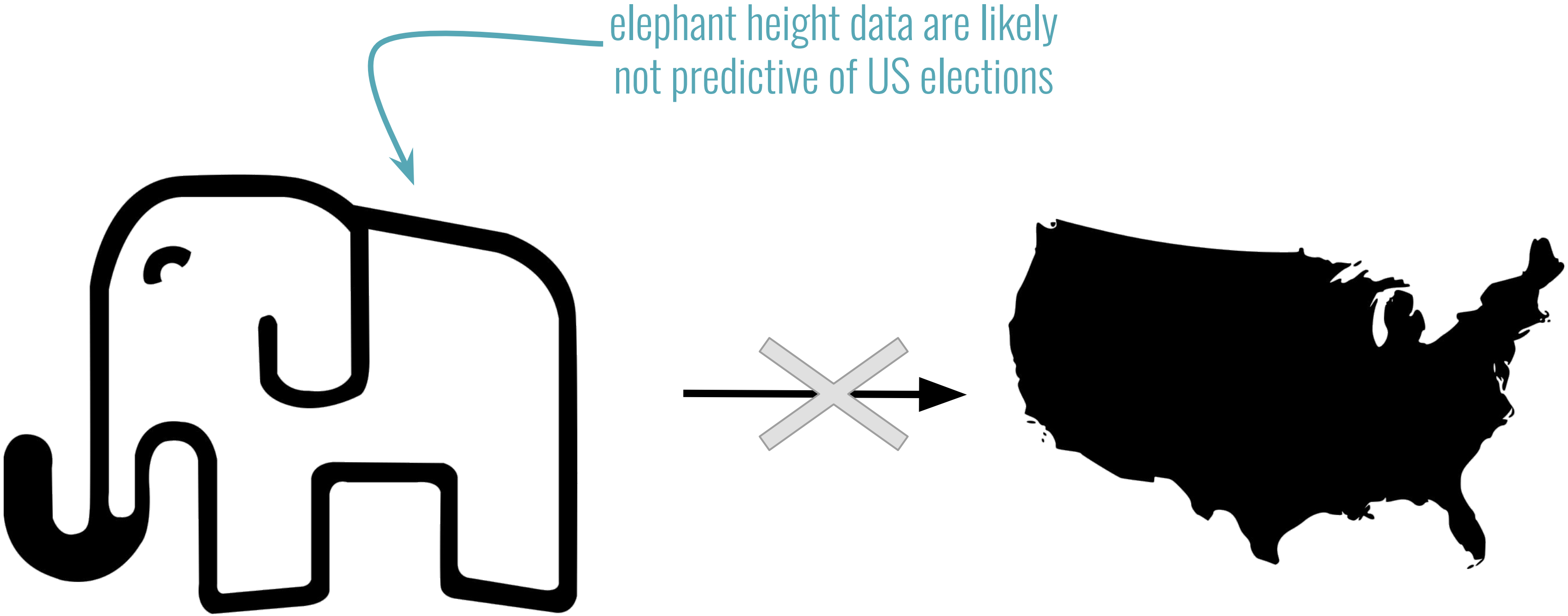
B
The training data

C
The validation data

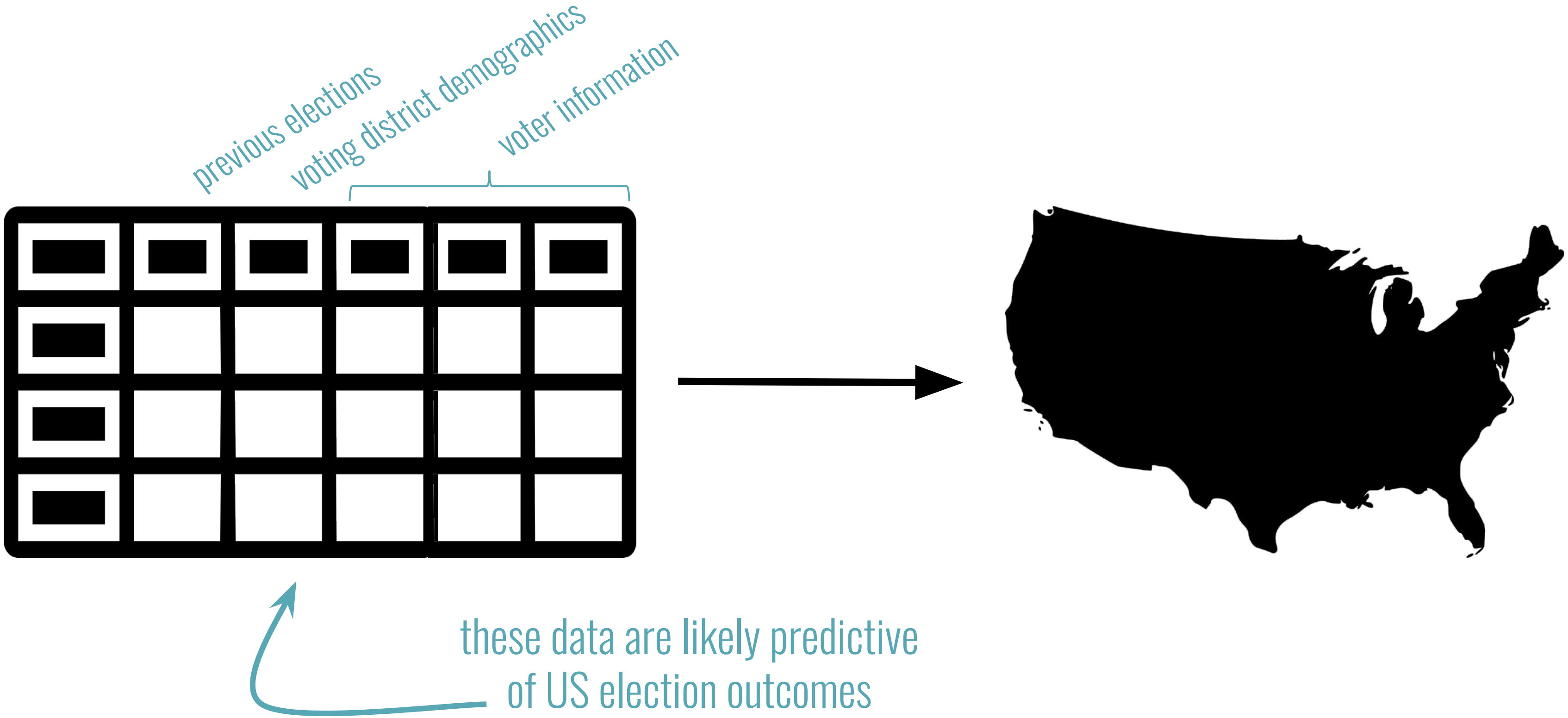
D
The testing data



feature selection



elephant height data are likely not predictive of US elections

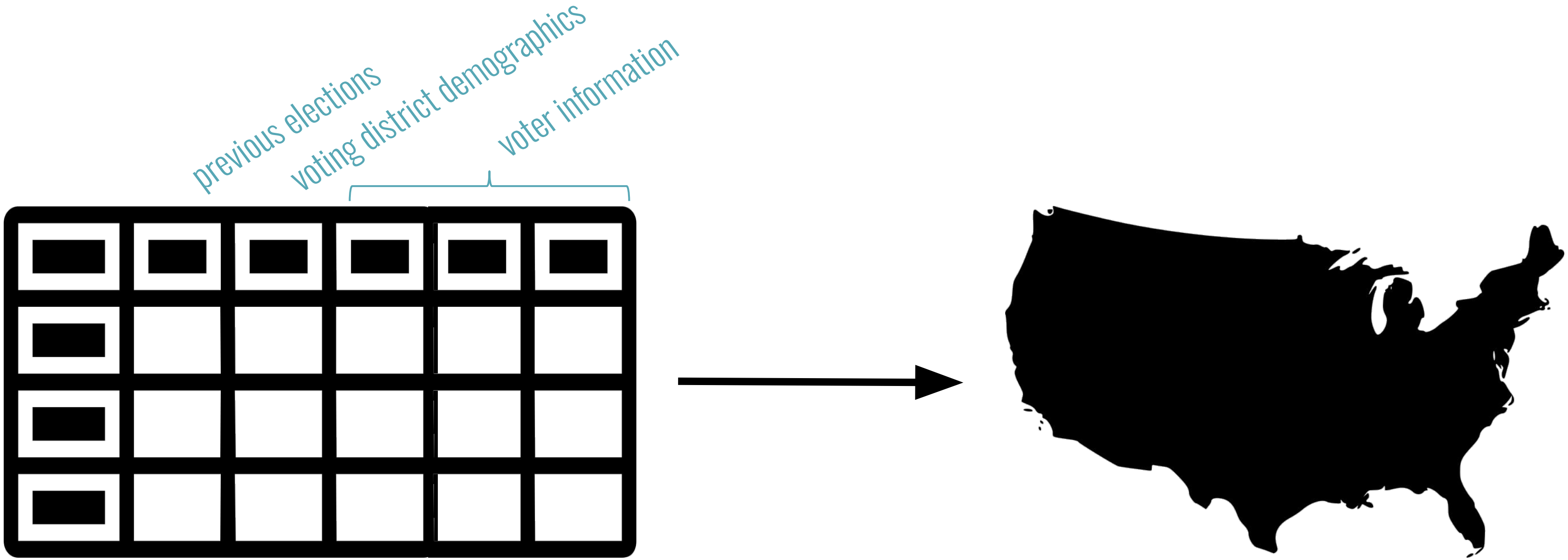


previous elections

voting district demographics

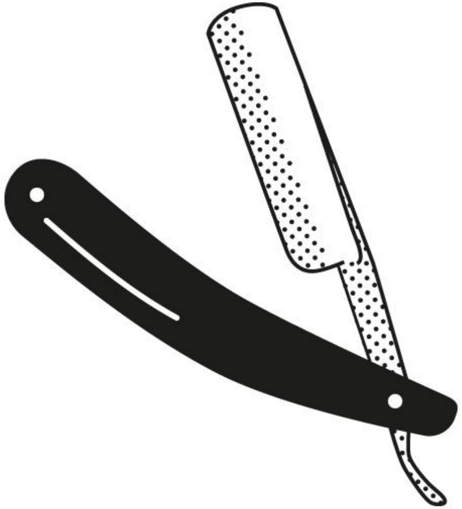
voter information

these data are likely predictive
of US election outcomes



feature selection determines which variables are most predictive and includes them in the model

Note: Occam’s Razor



- When faced with two explanations for the *same evidence*, we prefer the explanation that makes the fewest assumptions
- Given *equal performance*, choose the simpler model

model selection

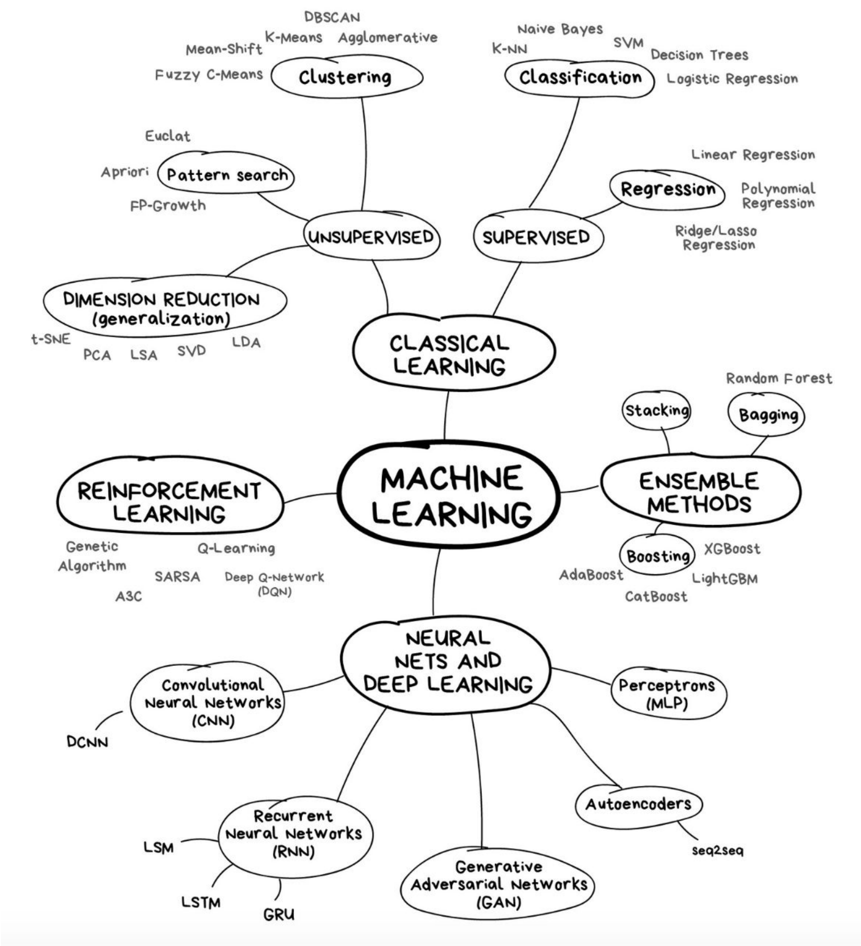


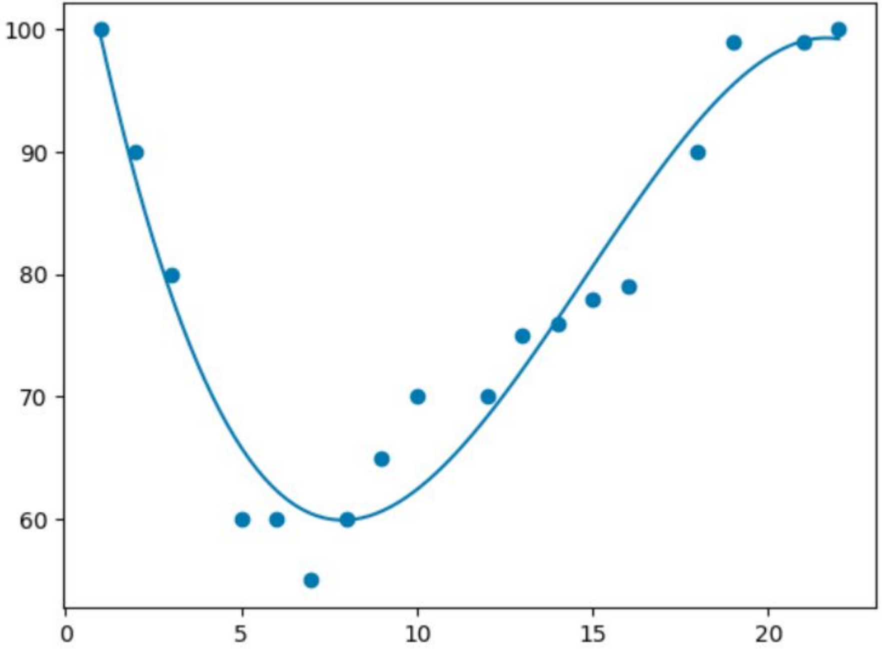
Image source: https://vas3k.com/blog/machine_learning/

Selecting the "Right" Model

○ **A** Thought about it and have a thought

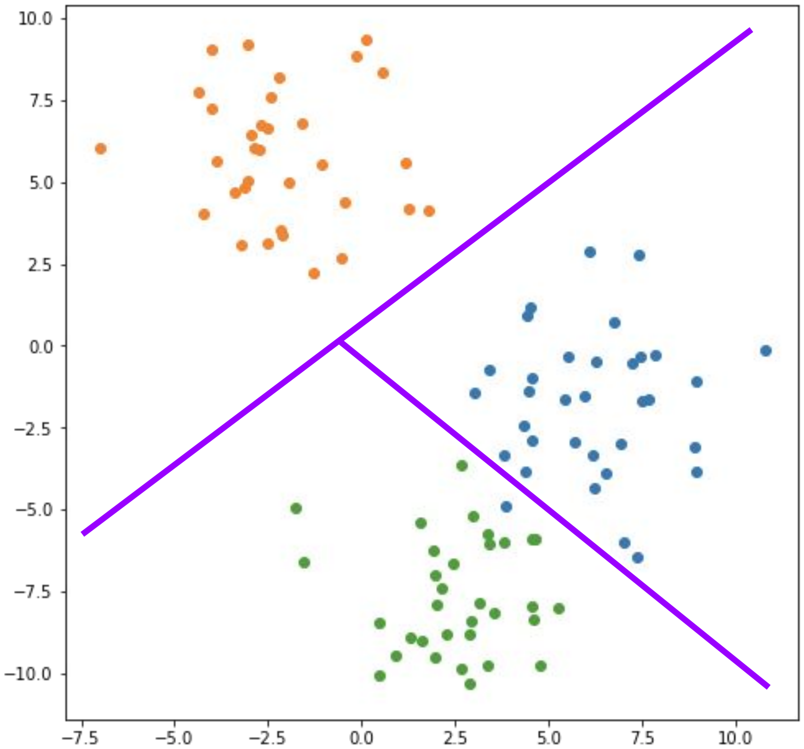
○ **B** Thought about it and have no thought

○ **C** I'm confused



Selecting the “Right” Model

- ○ **A** Thought about it and have a thought
- ○ **B** Thought about it and have no thought
- ○ **C** I’m confused

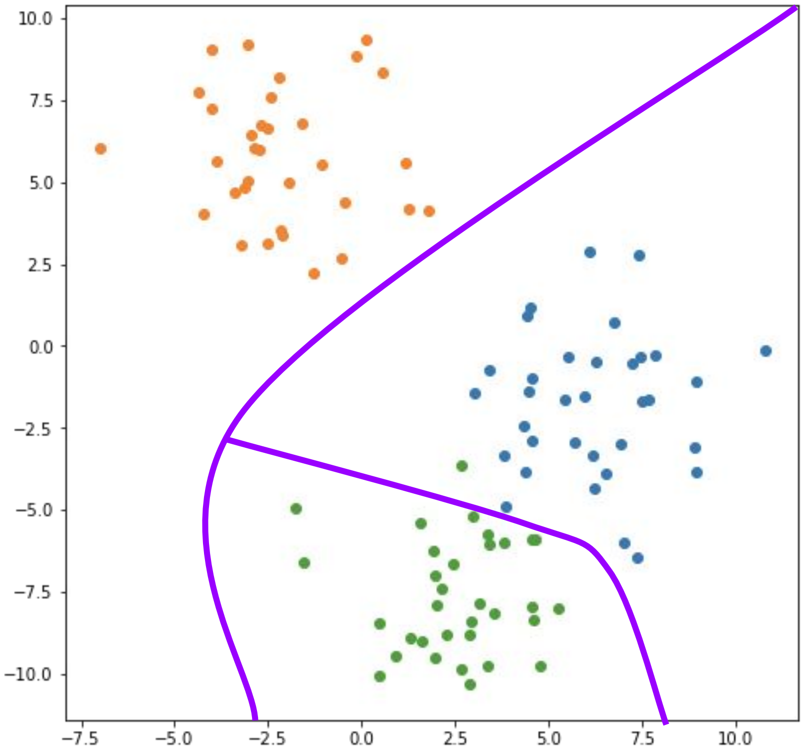


Selecting the "Right" Model

○ **A** Thought about it and have a thought

○ **B** Thought about it and have no thought

○ **C** I'm confused

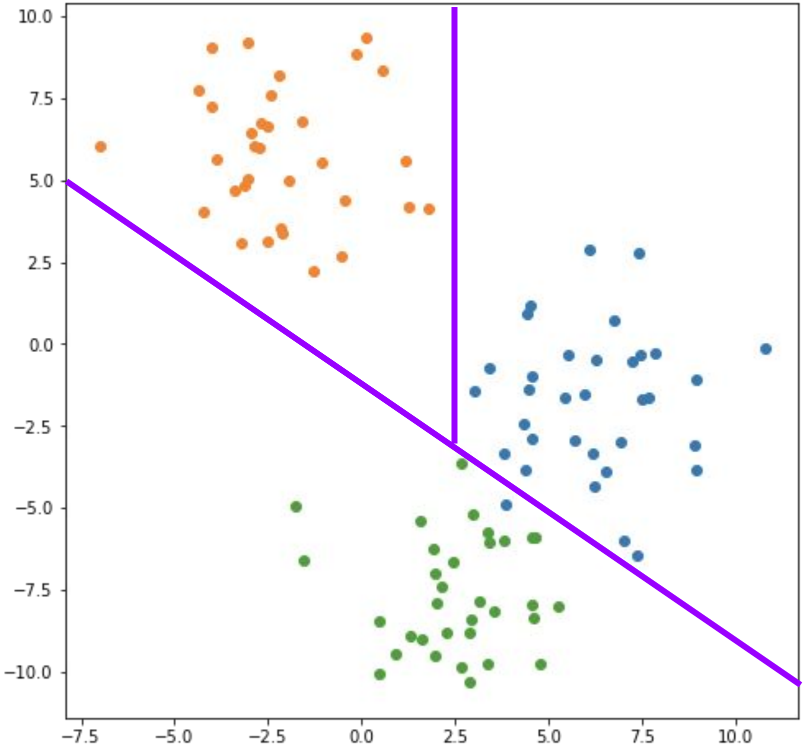


Selecting the “Right” Model

○ **A** Thought about it and have a thought

○ **B** Thought about it and have no thought

○ **C** I’m confused

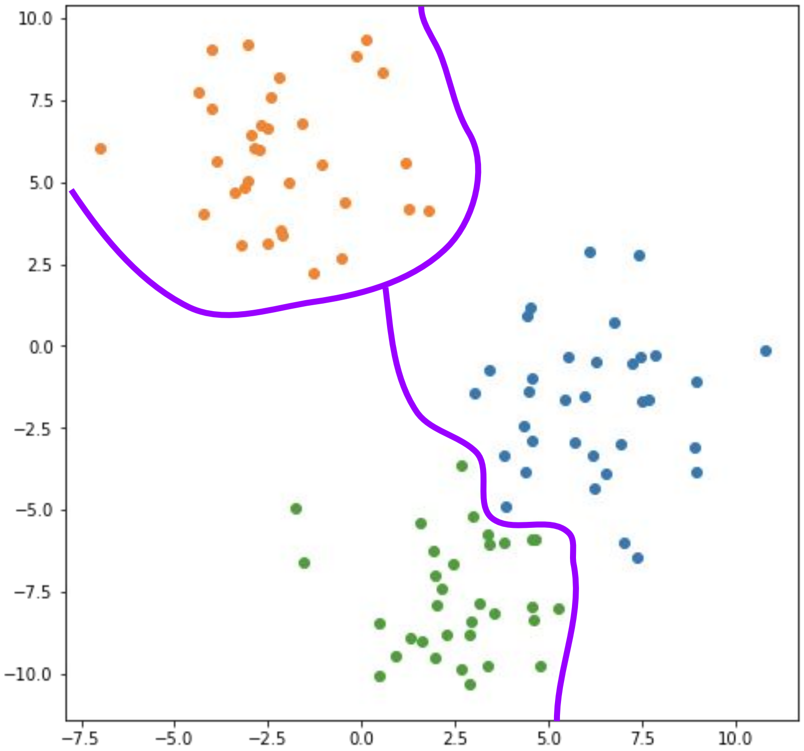


Selecting the “Right” Model

○ **A** Thought about it and have a thought

○ **B** Thought about it and have no thought

○ **C** I’m confused



Cross-Validation

In reality, our eyeball meter won’t (and sometimes can’t) cut it

Real data are messy and live in dimensions we cannot even comprehend (let alone visualize)

Cross-validation offers a systematic way of assessing various models and determine which ones meet our requirements the closest

- Validation Set
- Leave-one-out
- K-Fold

model assessment

Many Metrics!

Assessing machine learning models involves gauging its performance on novel data

- Novel data come from the testing set we left out earlier!
- Performance assessments depend on the specific problem/question
- Different metrics measure different aspects of models
    - e.g. MSE, MAE, accuracy, specificity/sensitivity, etc.

Classical Machine Learning

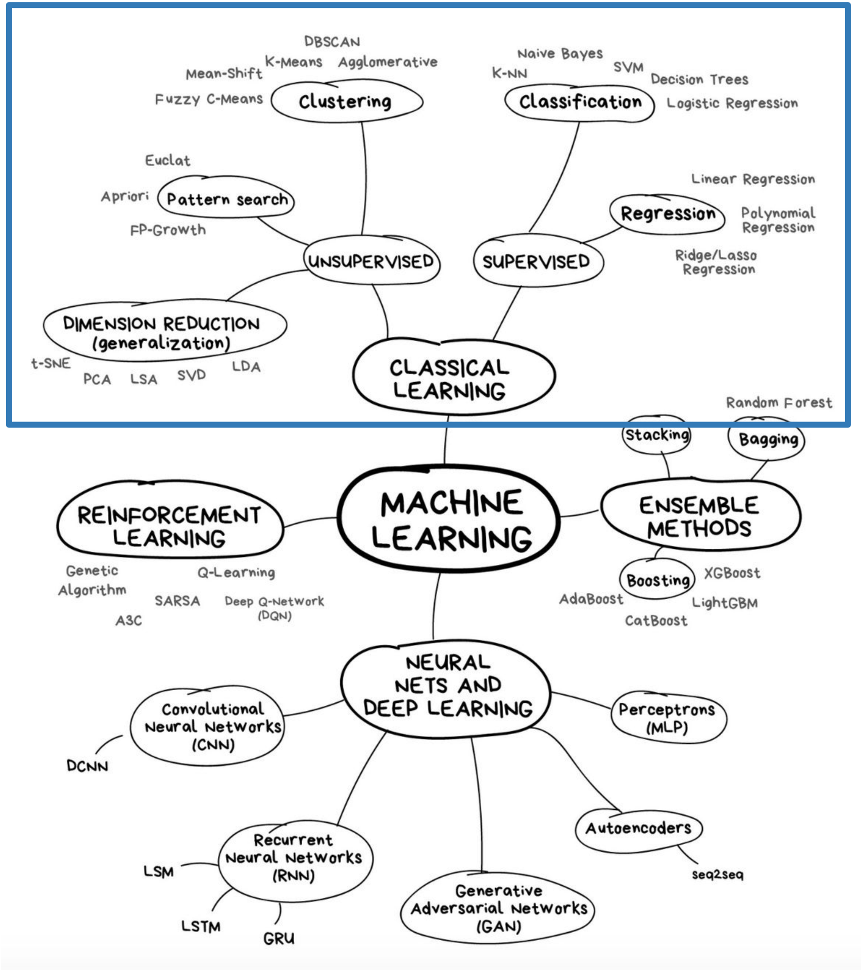


Image source: https://vas3k.com/blog/machine_learning/

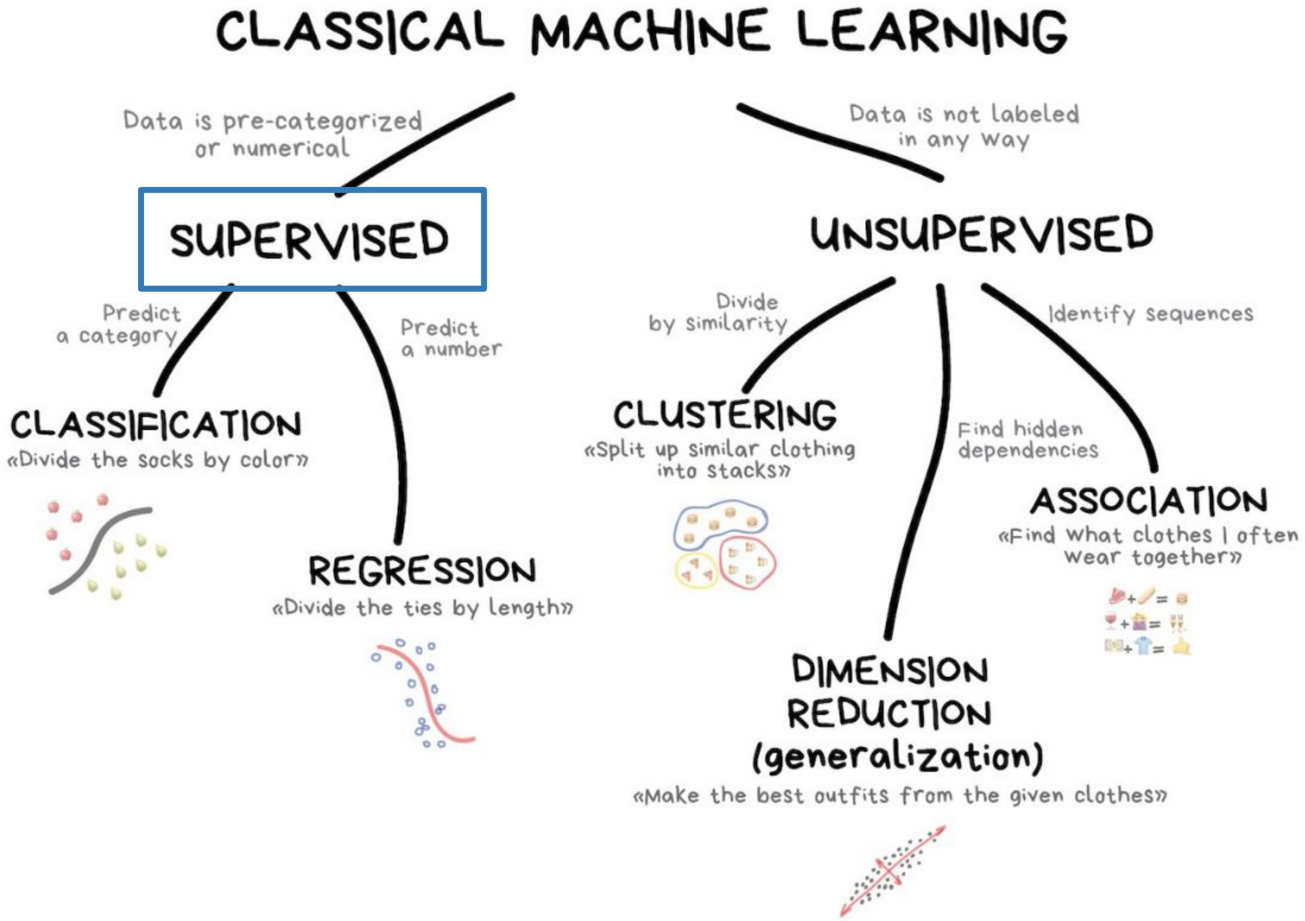

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar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

Image source: https://vas3k.com/blog/machine_learning/

Regression

- Predicting a **continuous** outcome by finding a relationship between inputs and outputs
- Makes predictions on new data based on the learned patterns from the training set
- Used for:
  - Stock price forecasts
  - Demand and sales volume analysis
  - Number-time correlations
- Popular regression models include:
  - Linear regression
  - Polynomial regression

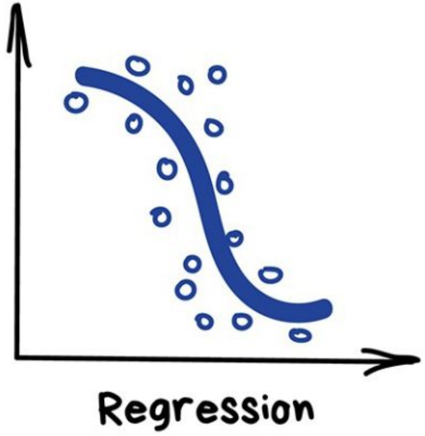


Image source: https://vas3k.com/blog/machine_learning/

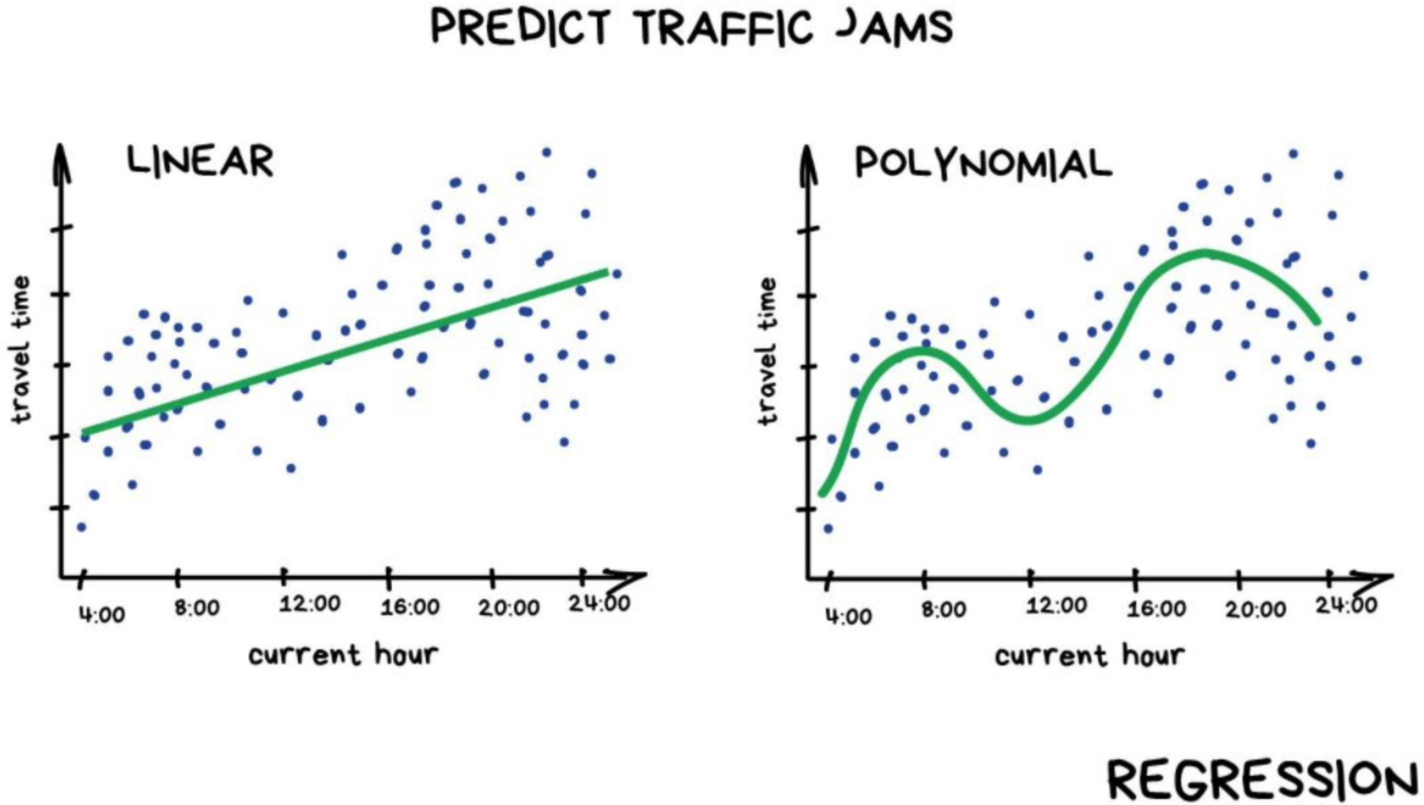


Image source: https://vas3k.com/blog/machine_learning/

Classification

- Predicting a **categorical** outcome by finding boundaries between classes of data points
- Makes predictions on new data based on where within the decision boundary it falls
- Used for:
  - Spam filtering
  - Fraud detection
  - Language detection
- Popular classification models include:
  - Logistic regression
  - Support Vector Machine
  - Decision Tree

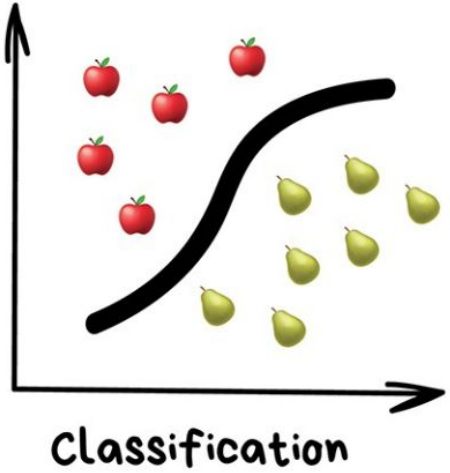


Image source: https://vas3k.com/blog/machine_learning/

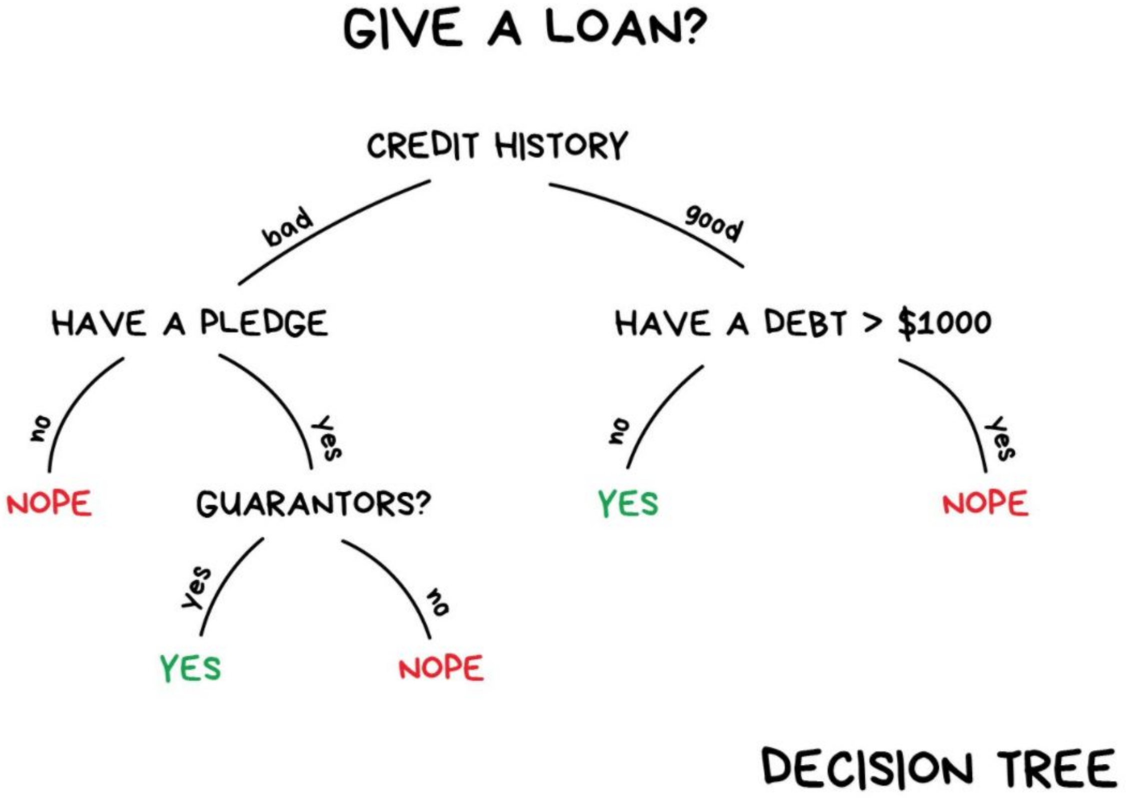


Image source: https://vas3k.com/blog/machine_learning/

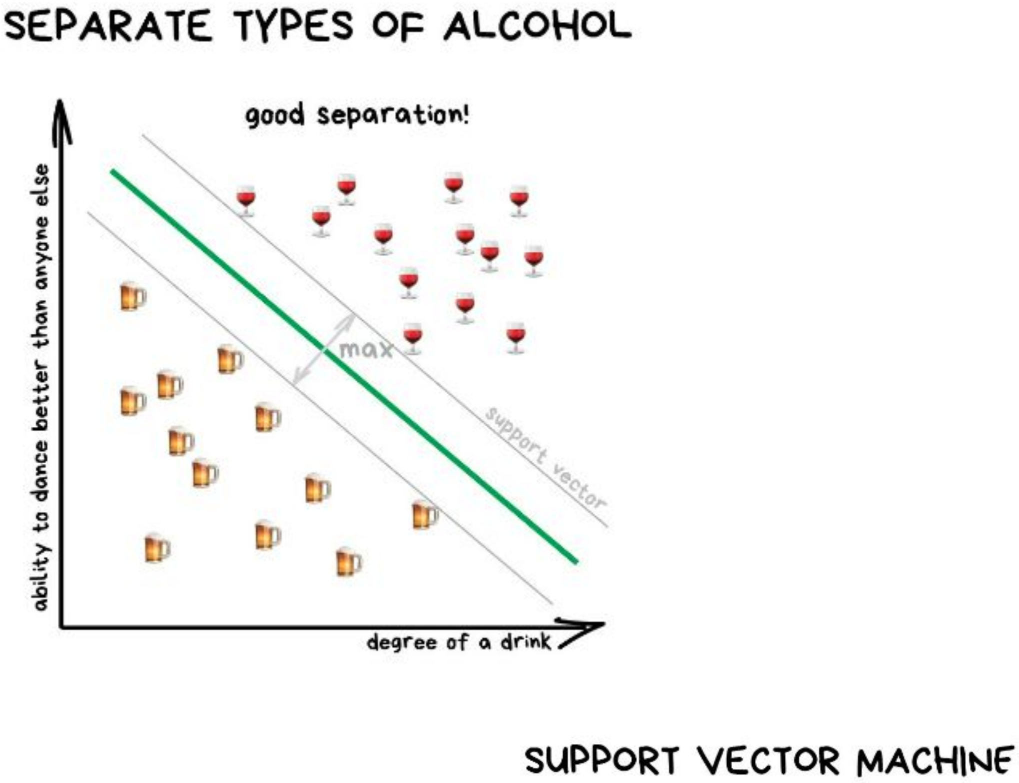


Image source: https://vas3k.com/blog/machine_learning/

Supervised vs. Unsupervised

Supervised algorithms use labeled data → Checks answers and improves over time

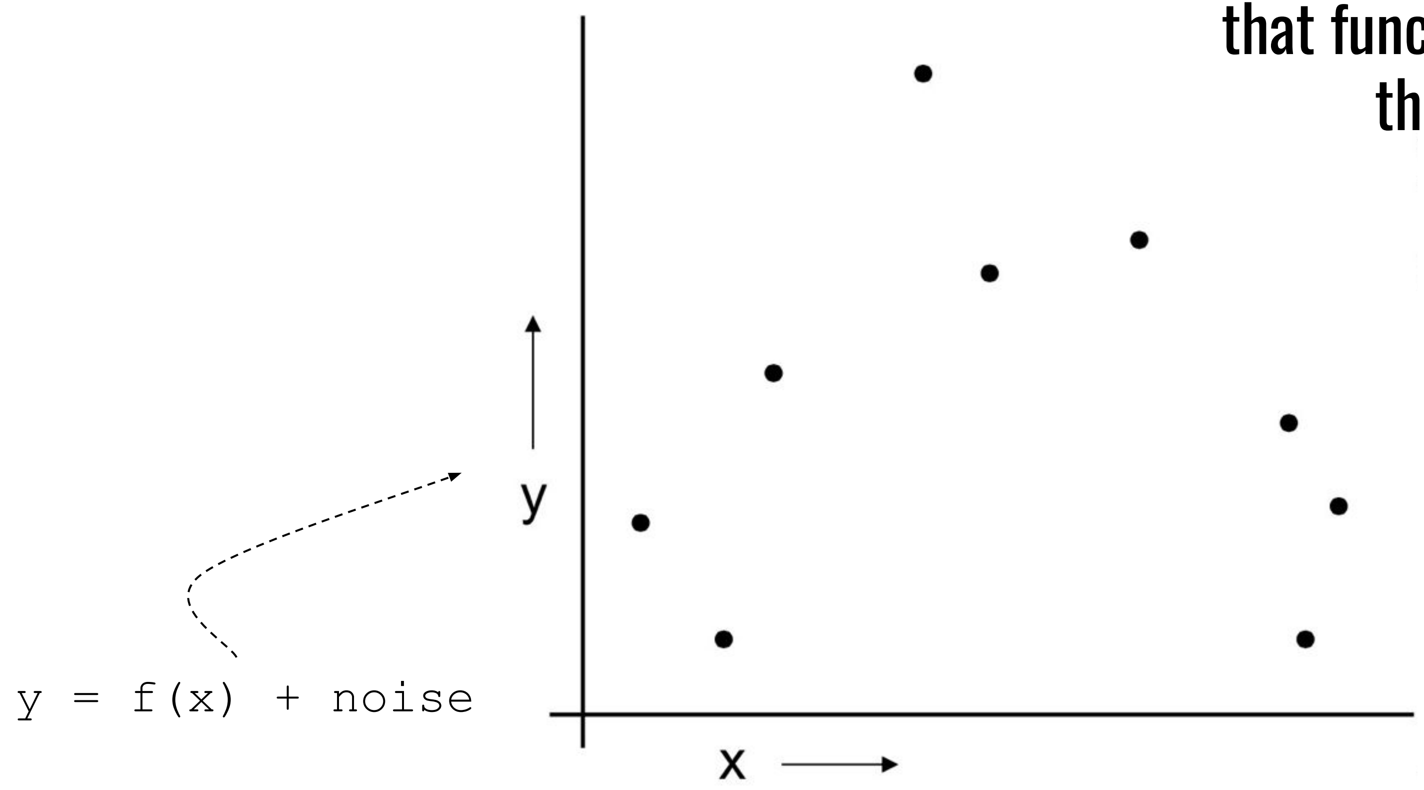
- Learning relationships between inputs and outputs to make predictions
- Goal is to minimize error or maximize accuracy in predictions

Unsupervised algorithms use unlabeled data → No "correct" answers

- Commonly used to discover new patterns where they are unknown → research!!
- Verifies itself with similarity/stability scores

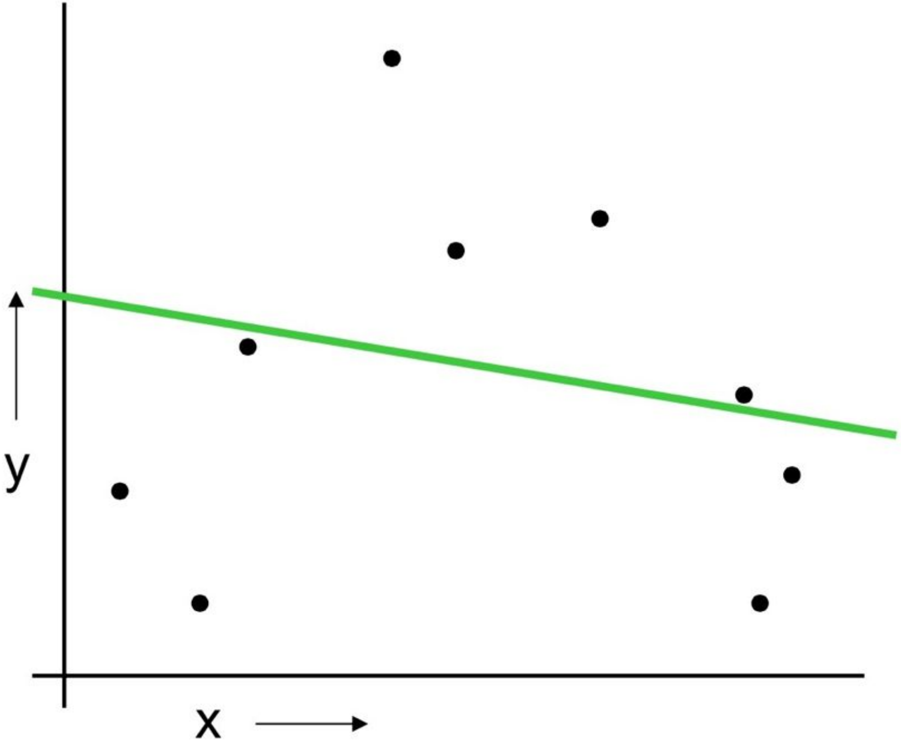
Regression Walkthrough

Can we determine what that function (`f`) *is* using these data?



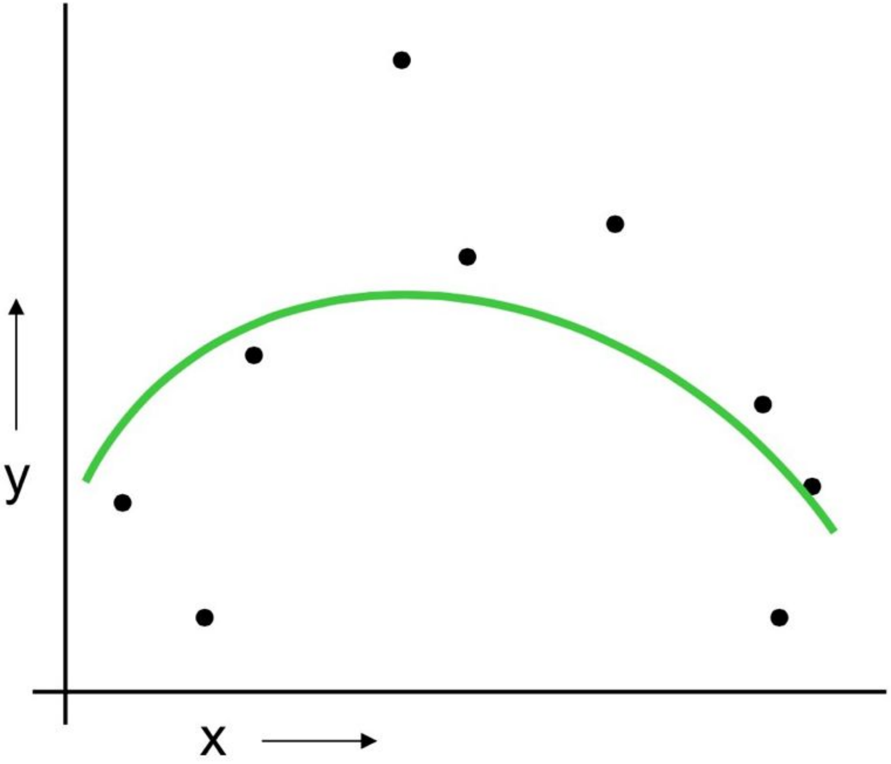
adapted from Brad Voytek

Linear regression



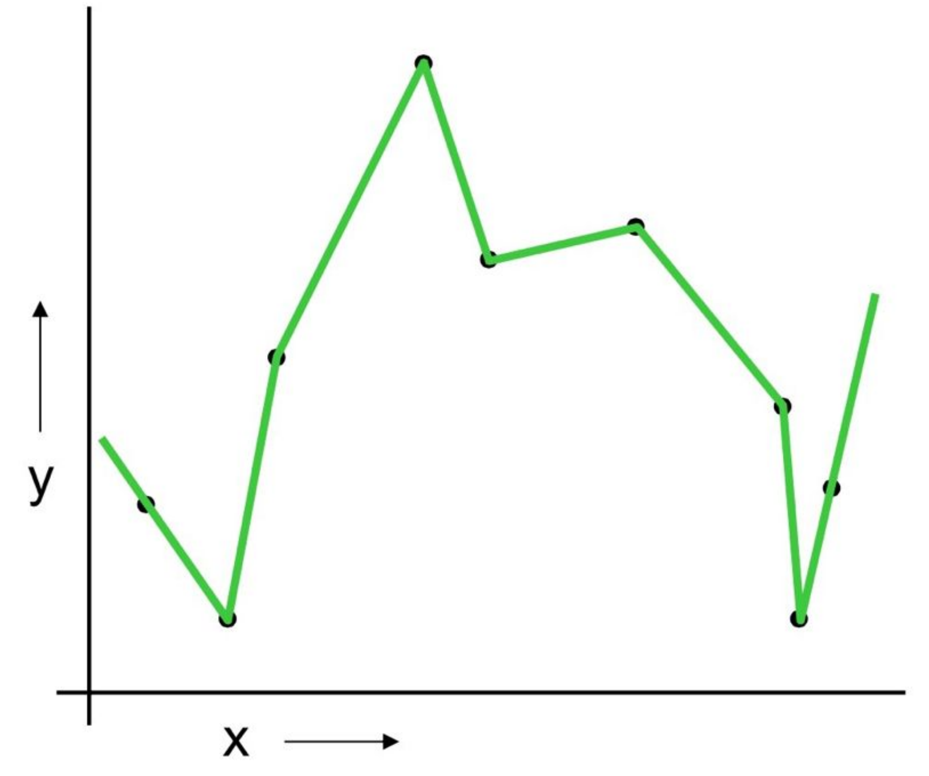
adapted from Brad Voytek

Quadratic regression



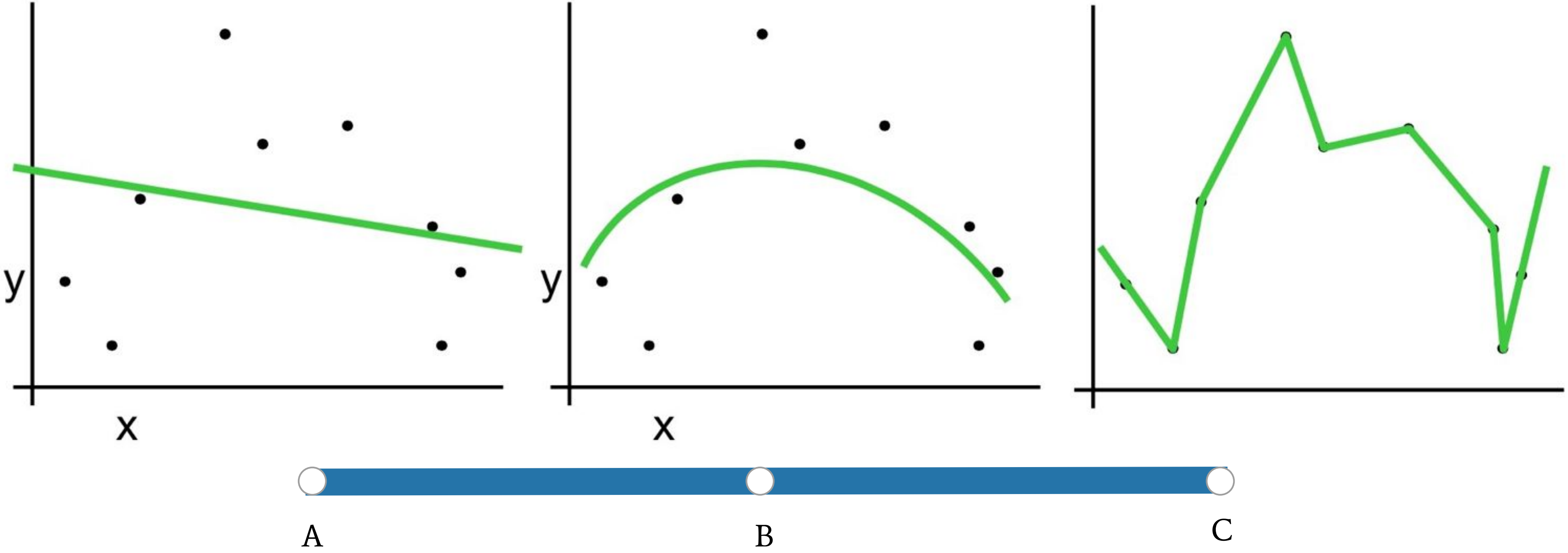
adapted from Brad Voytek

Piecewise linear nonparametric regression



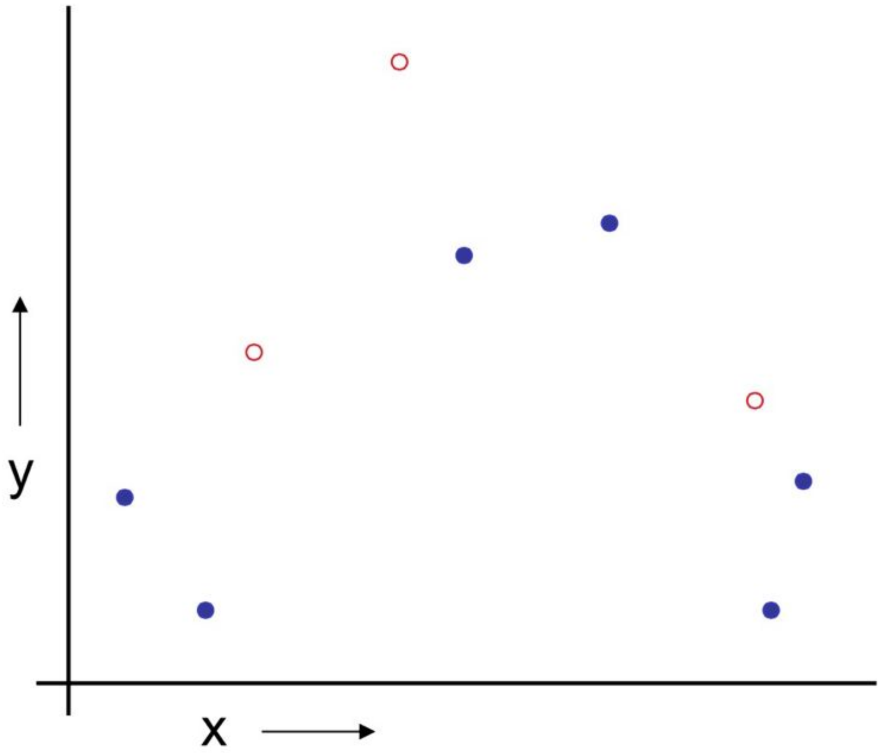
adapted from Brad Voytek

Which to choose?



adapted from Brad Voytek

The data partition method

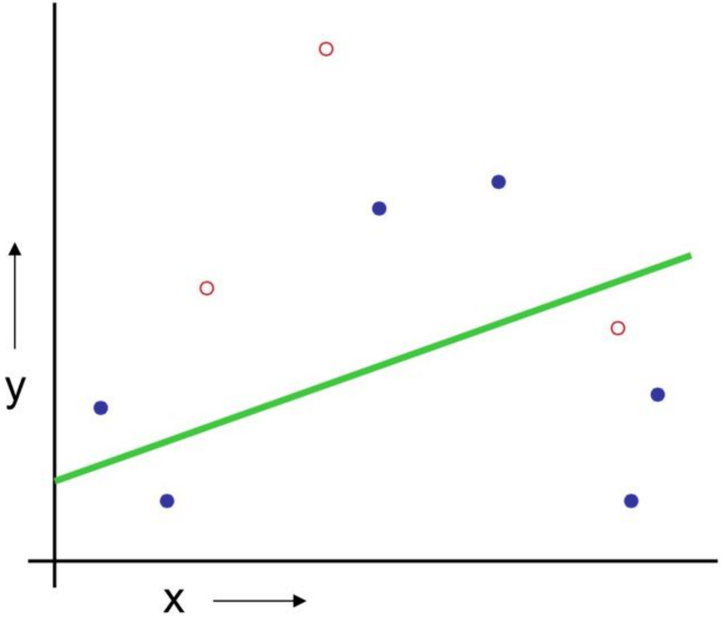


1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

adapted from Brad Voytek

Train the model on your training set



(Linear regression example)

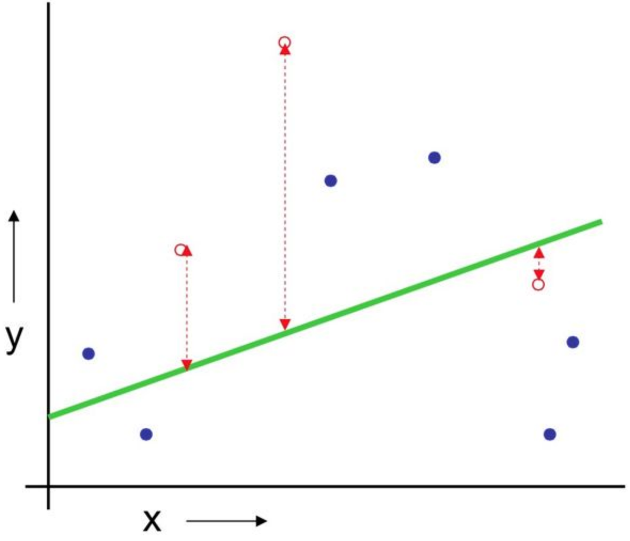
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

adapted from Brad Voytek

Assess future performance using the test set



(Linear regression example)

Mean Squared Error = 2.4

1. Randomly choose 30% of the data to be in a test set

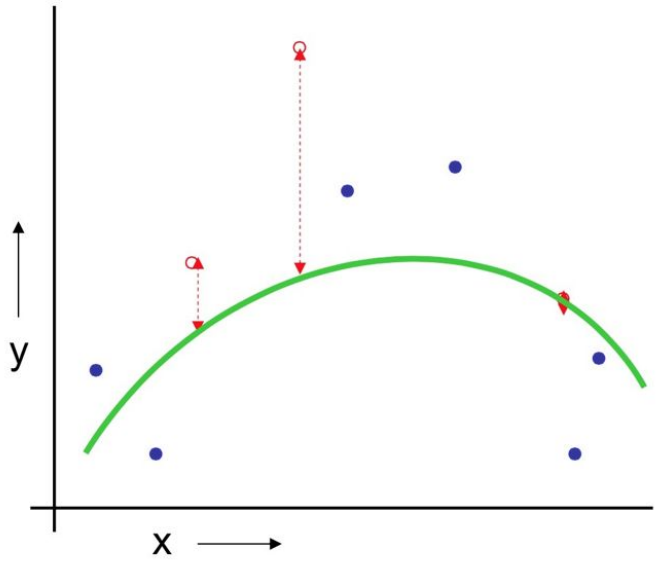
2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

adapted from Brad Voytek

Go through this process for each possible model



(Quadratic regression example)

Mean Squared Error = 0.9

1. Randomly choose 30% of the data to be in a test set

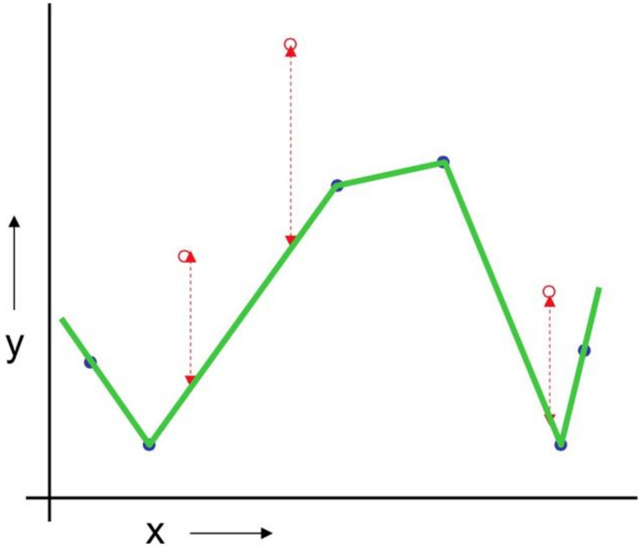
2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

adapted from Brad Voytek

Go through this process for each possible model



(Join the dots example)

Mean Squared Error = 2.2

1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

adapted from Brad Voytek

Pros and cons of data partitioning

Pros:

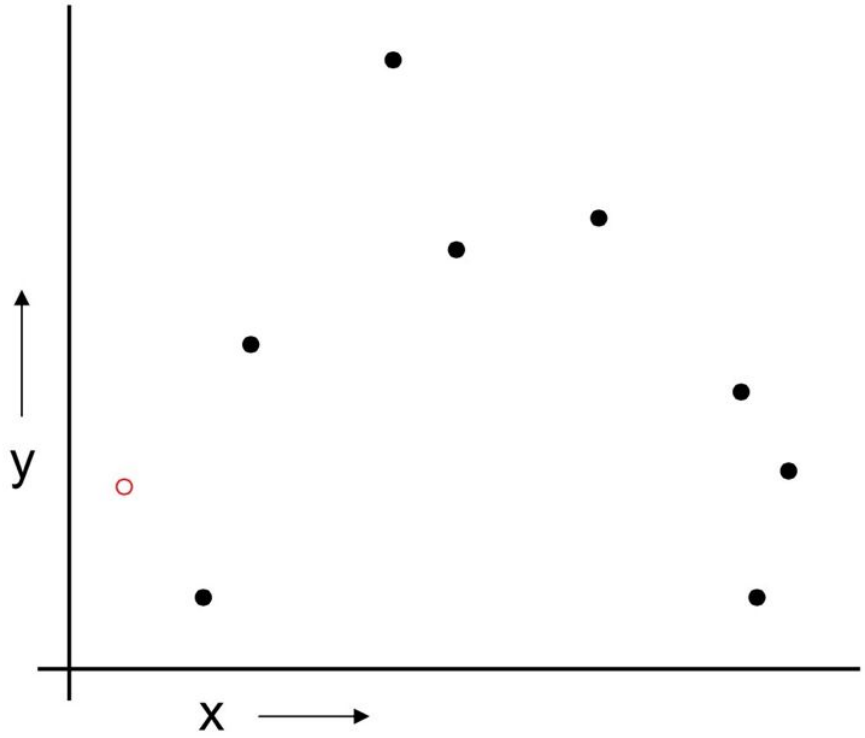
- Simple approach
- Can choose model with best test-set score

Cons:

- Model fit on 30% less data than you have
- Without a large data set, removing 30% of the data could bias prediction

adapted from Brad Voytek

Leave one out cross validation (LOOCV)

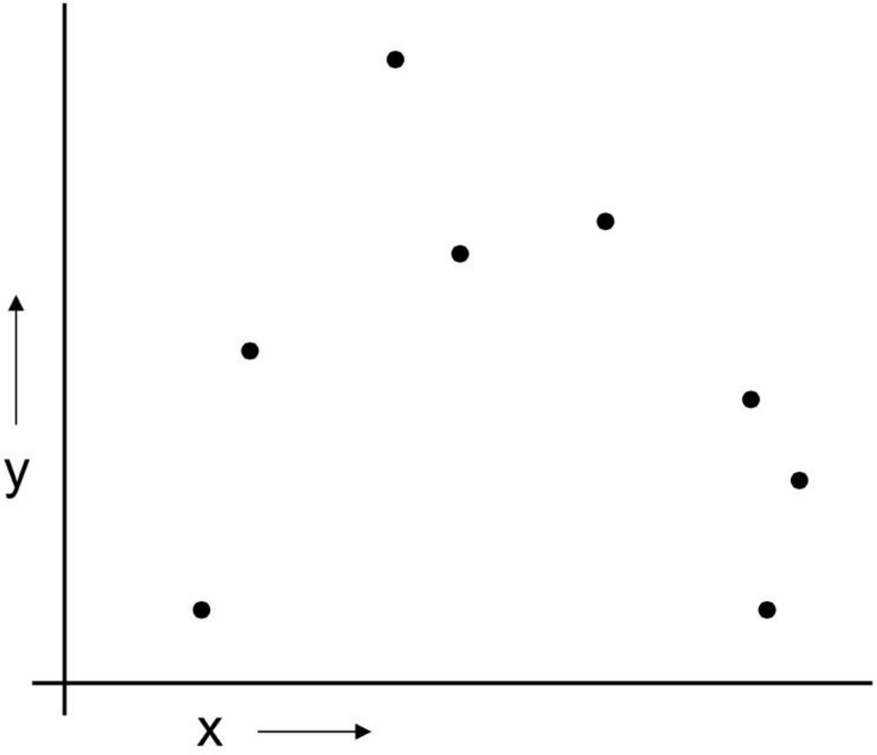


For k=1 to R

1. Let (x_k, y_k) be the kth record

adapted from Brad Voytek

Leave one out cross validation (LOOCV)

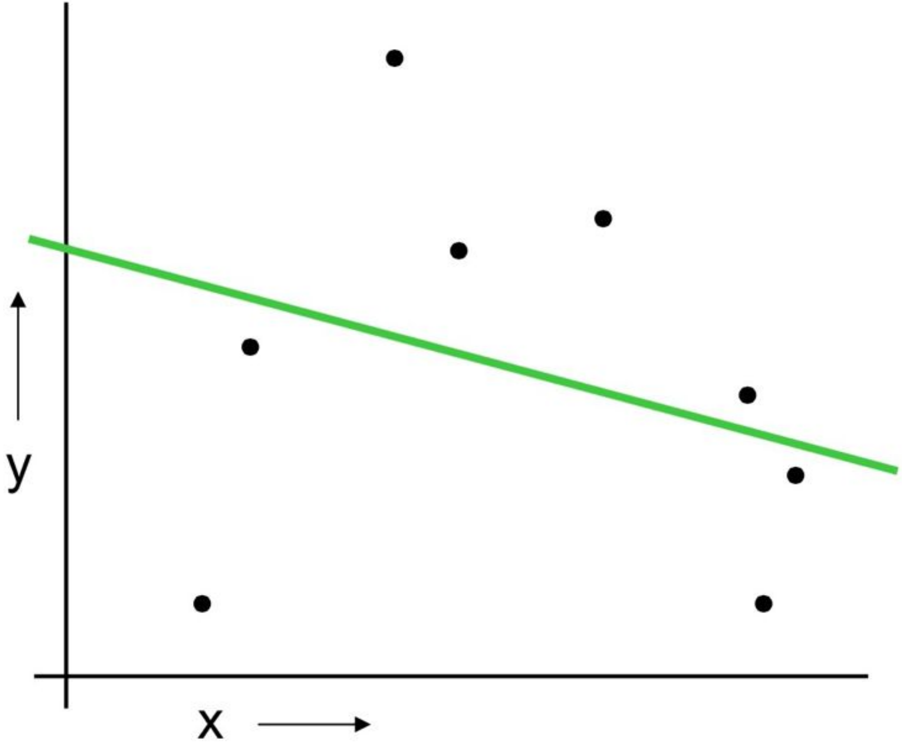


For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset

adapted from Brad Voytek

Leave one out cross validation (LOOCV)

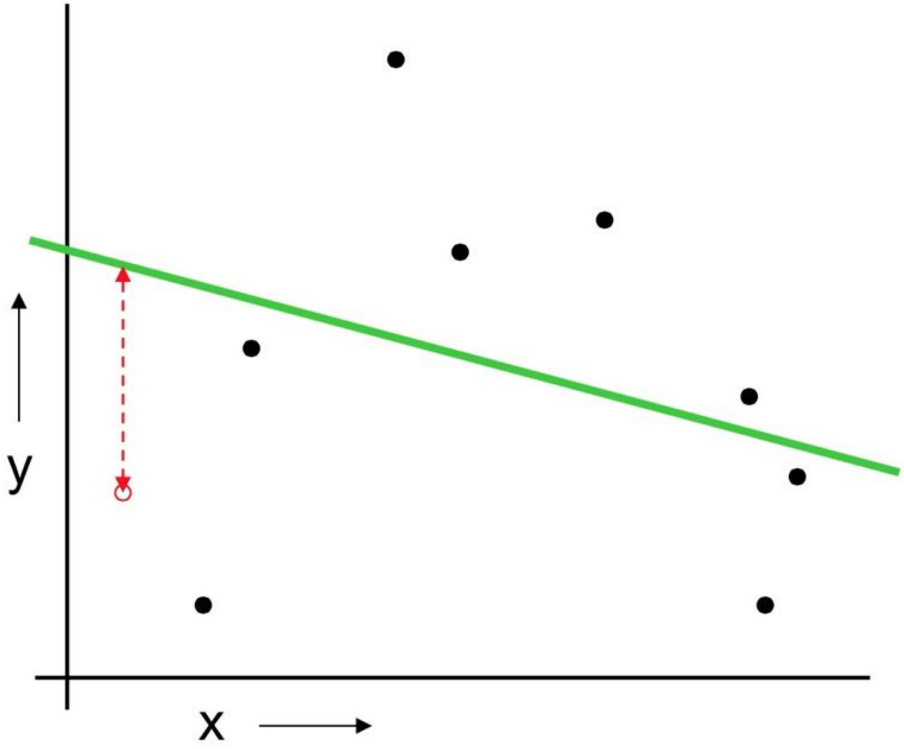


For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints

adapted from Brad Voytek

Leave one out cross validation (LOOCV)

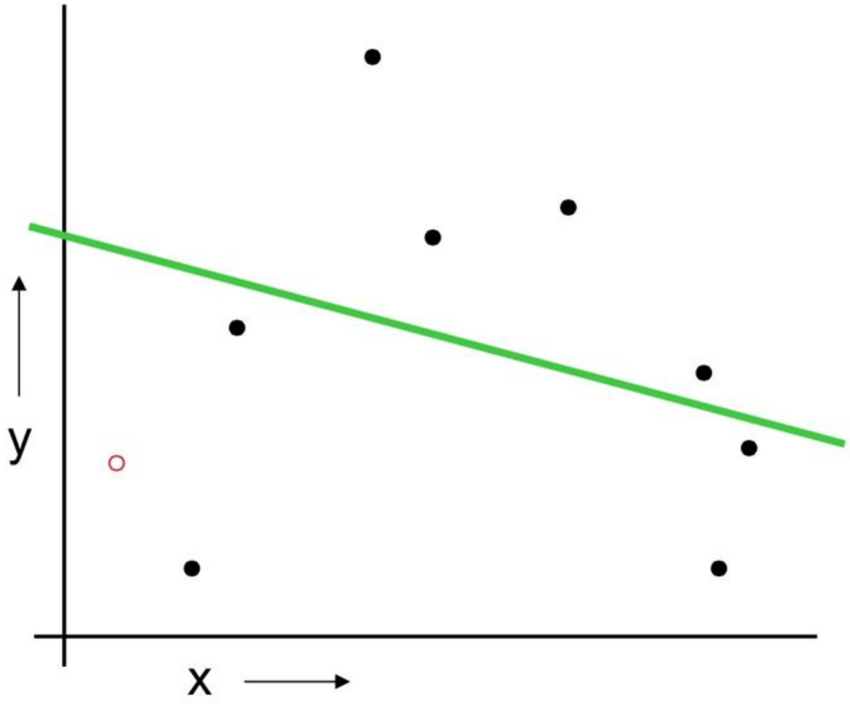


For k=1 to R

1. Let (x_k,y_k) be the k^{th} record
2. Temporarily remove (x_k,y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k,y_k)

adapted from Brad Voytek

Leave one out cross validation (LOOCV)



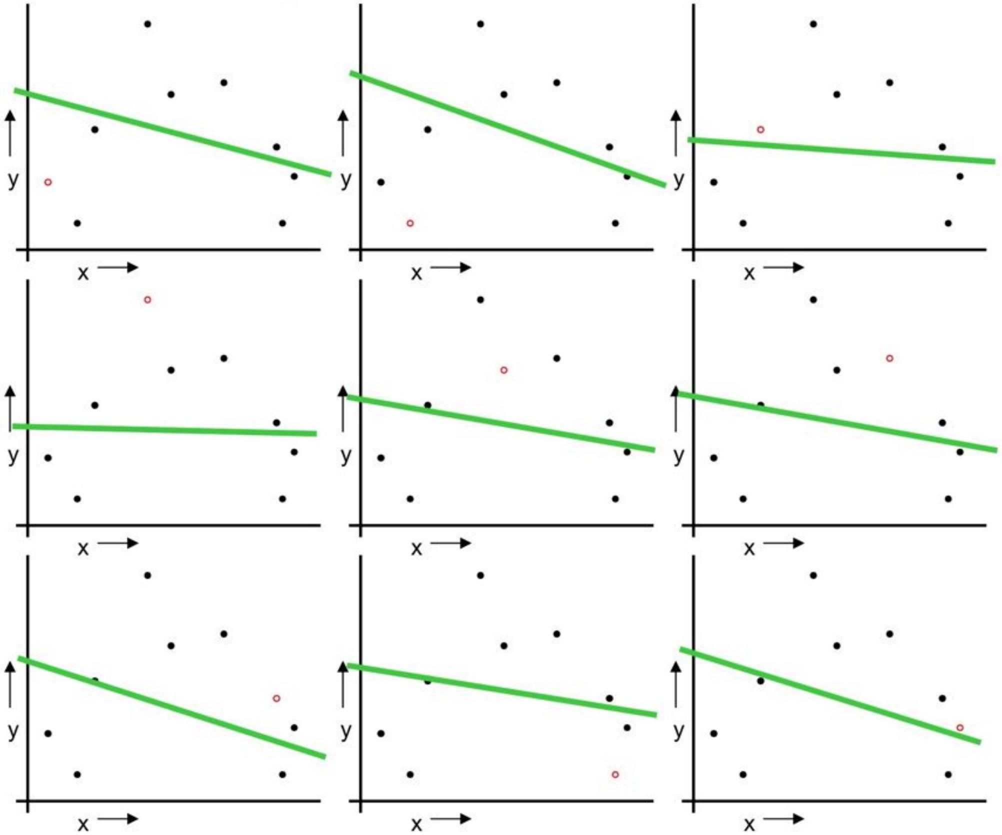
For k=1 to R

1. Let (x_k,y_k) be the k^{th} record
2. Temporarily remove (x_k,y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k,y_k)

When you've done all points, report the mean error.

adapted from Brad Voytek

Leave one out cross validation (LOOCV)



For k=1 to R

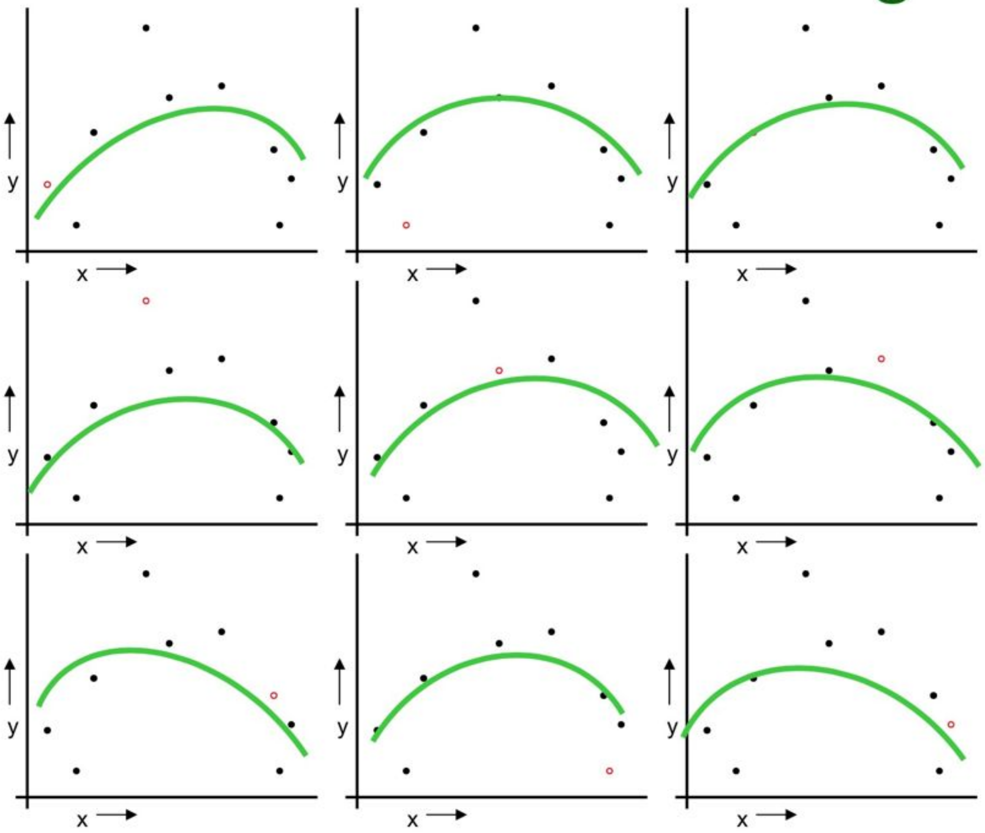
1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV} = 2.12$

adapted from Brad Voytek

Leave one out cross validation (LOOCV)



For k=1 to R

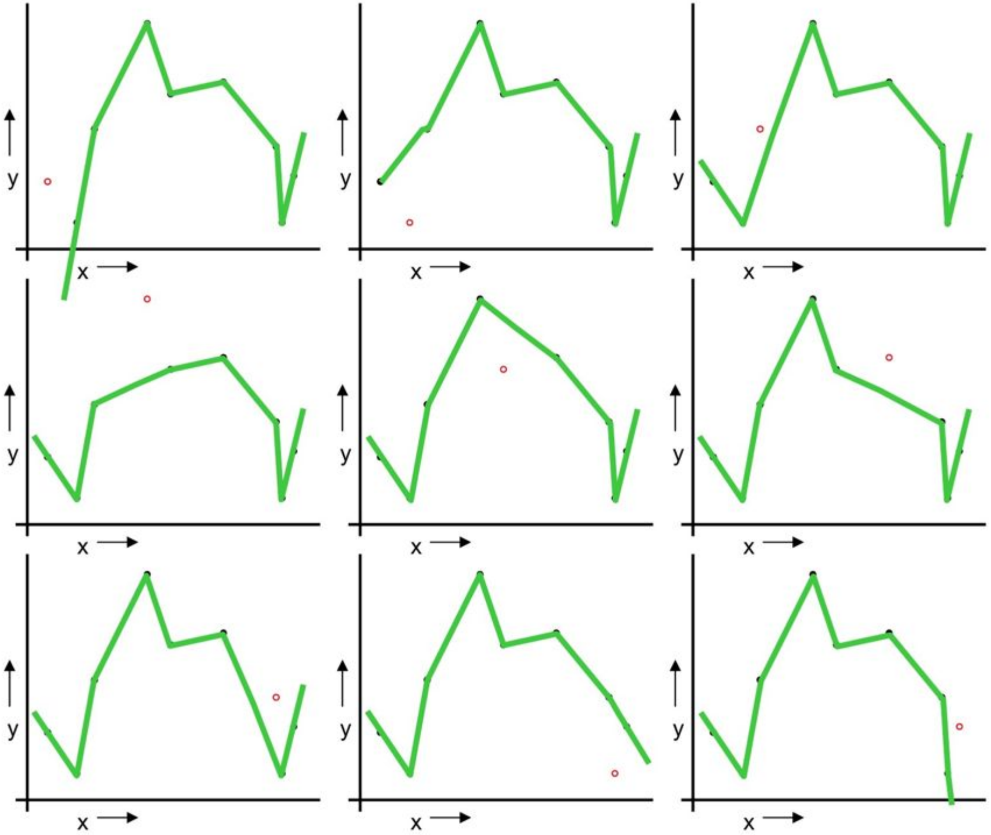
1. Let (x_k,y_k) be the k^{th} record
2. Temporarily remove (x_k,y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k,y_k)

When you've done all points, report the mean error.

MSE_{LOOCV} $=0.962$

adapted from Brad Voytek

Leave one out cross validation (LOOCV)



For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV} = 3.33$

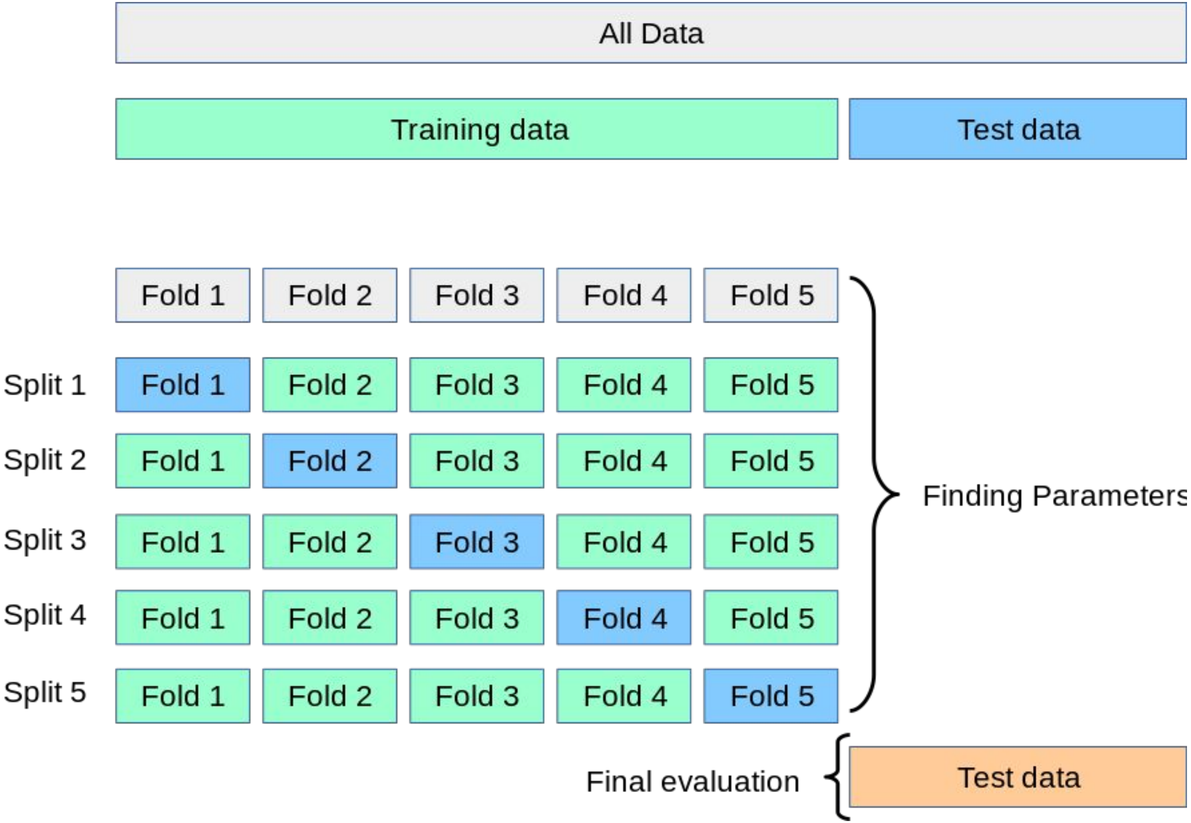
adapted from Brad Voytek

Method Comparison

	Cons	Pros
Data partitioning	Variance: unreliable estimate of future performance	Cheap
LOOCV	Computationally expensive; has weird behavior	Uses all your data

adapted from Brad Voytek

k-fold cross validation



All Data

Training data

Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

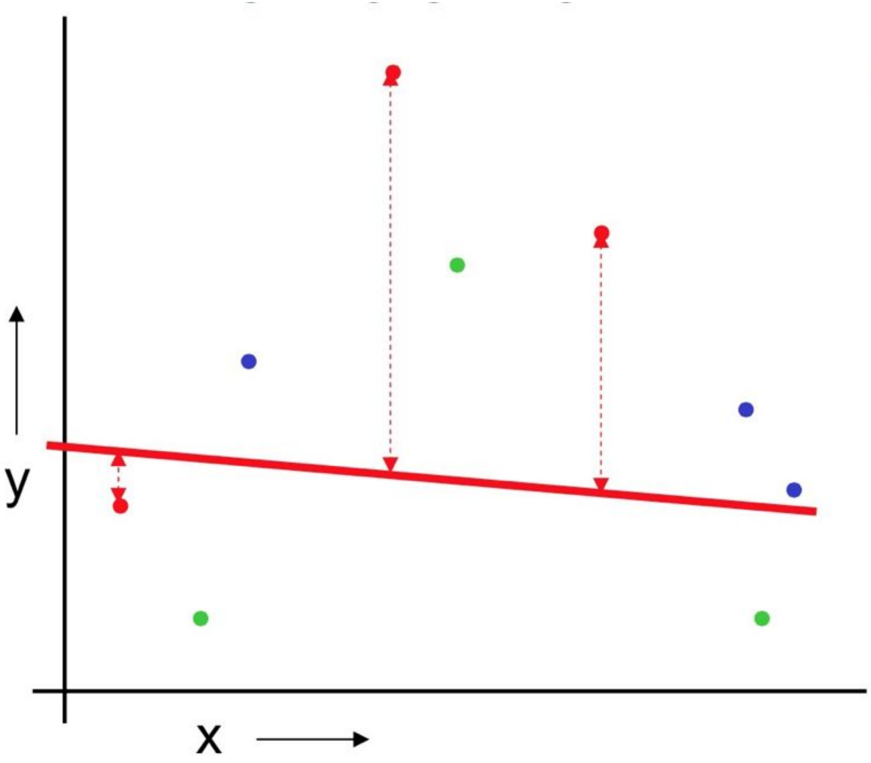
Finding Parameters

Final evaluation

Test data

https://scikit-learn.org/stable/modules/cross_validation.html

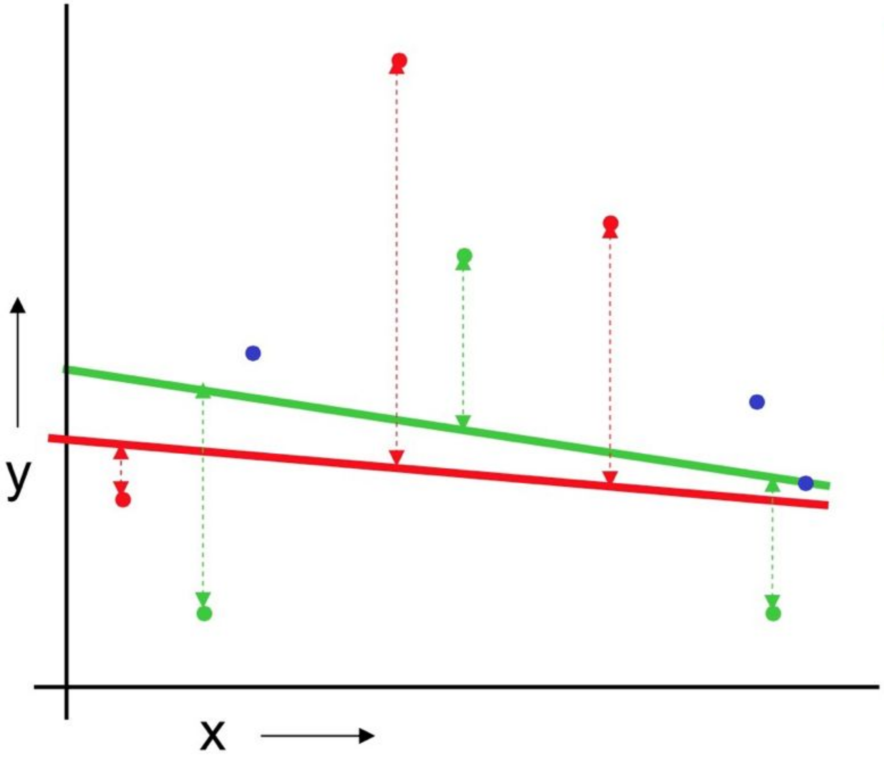
k-fold cross validation



For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

adapted from Brad Voytek

k-fold cross validation

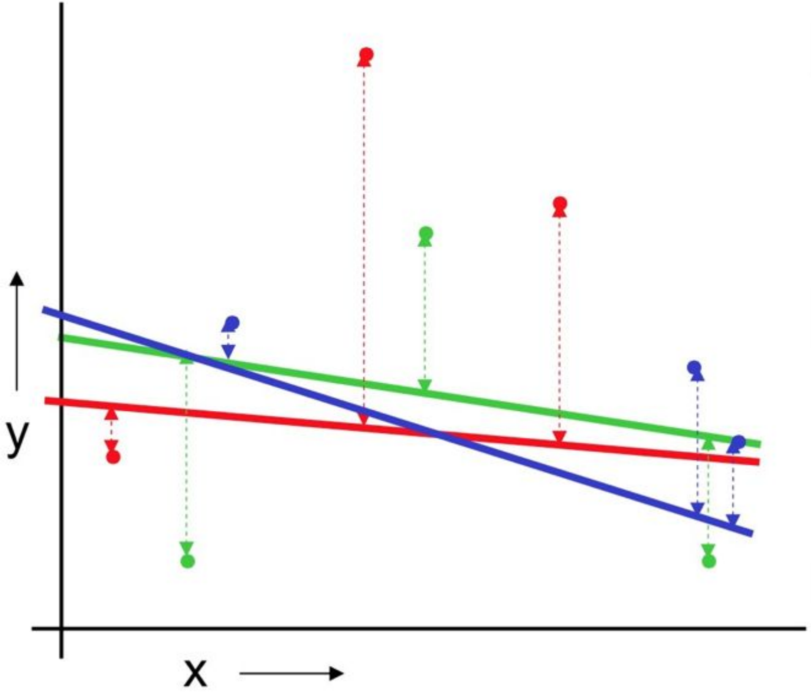


For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

adapted from Brad Voytek

k-fold cross validation



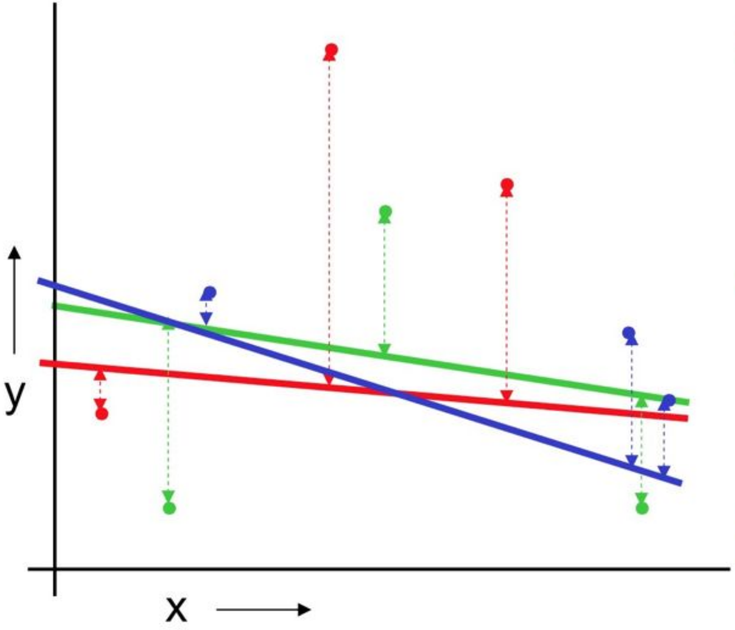
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

adapted from Brad Voytek

k-fold cross validation



Linear Regression

$MSE_{3FOLD}=2.05$

For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

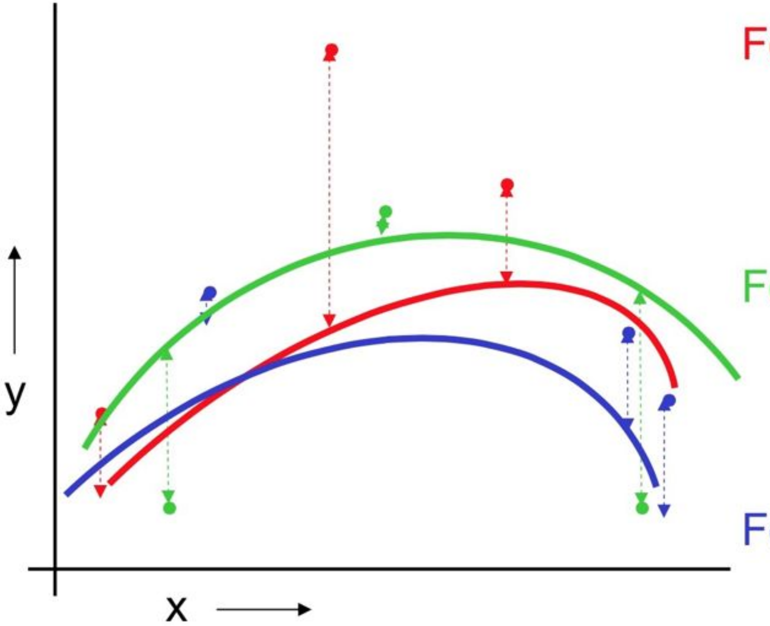
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

adapted from Brad Voytek

k-fold cross validation



Quadratic Regression
MSE_{3FOLD}=1.11

For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

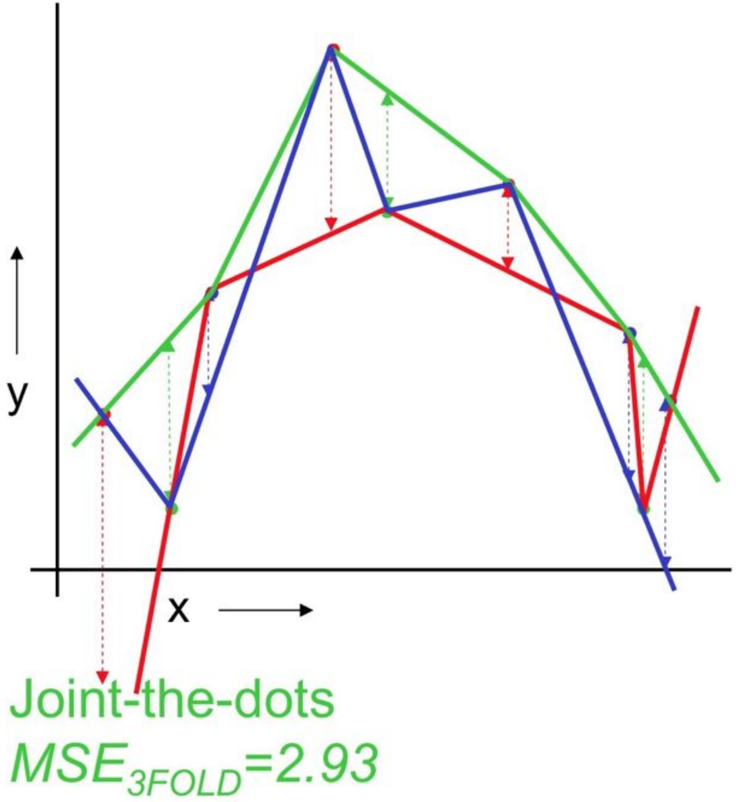
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

adapted from Brad Voytek

k-fold cross validation



For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

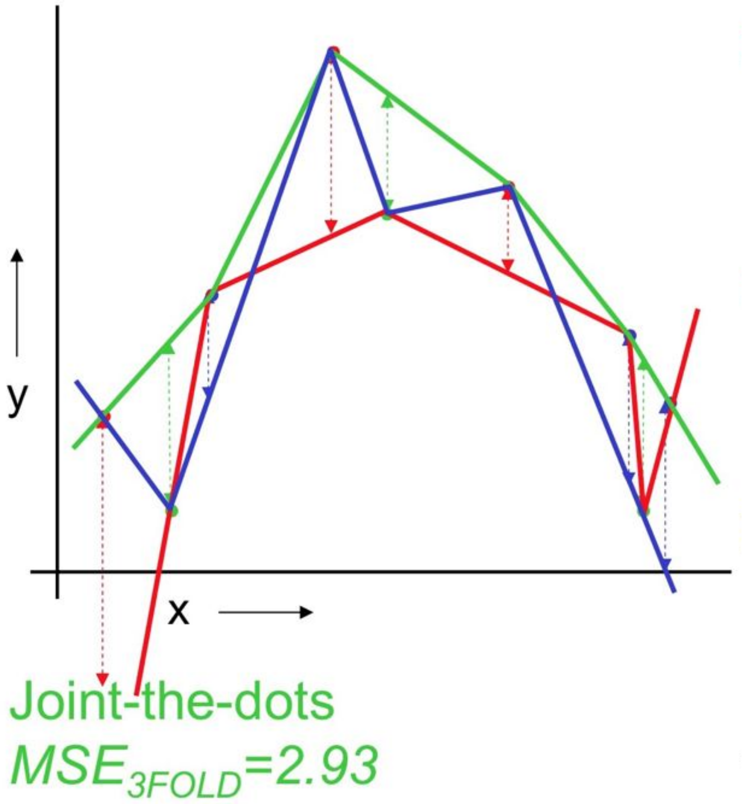
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

adapted from Brad Voytek

k-fold cross validation



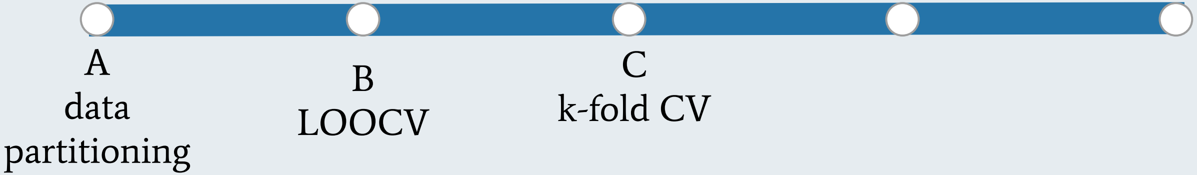
NOTE: Notice each fold generates a dramatically different model than the others?

- This model has *overfit* the training data
- i.e. You only memorized the study guide
- Perfect training performance, but poor predictive power on test set

adapted from Brad Voytek

Validator

Which approach would *you* use to limit overfitting?



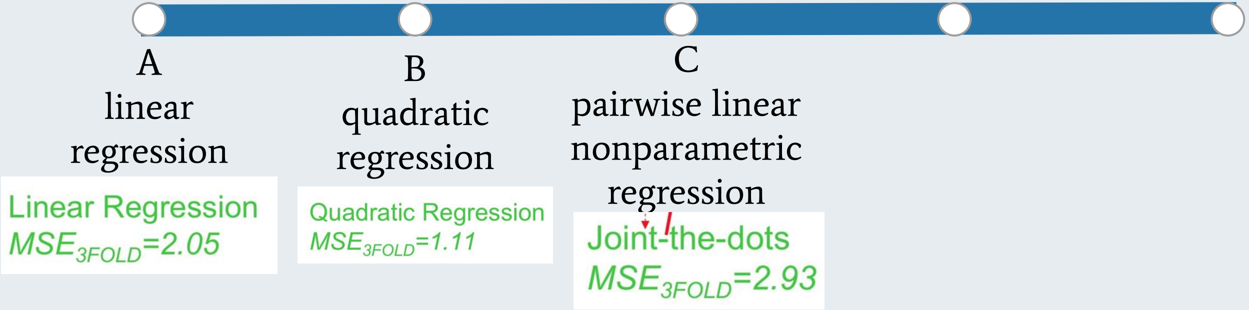
A
data
partitioning

B
LOOCV

C
k-fold CV

Validator

Given the example we just worked, how would *you* model these data?



A
linear
regression

Linear Regression
MSE_{3FOLD}=2.05

B
quadratic
regression

Quadratic Regression
MSE_{3FOLD}=1.11

C
pairwise linear
nonparametric
regression

Joint-the-dots
MSE_{3FOLD}=2.93