Course Announcements

Due Sunday (11:59 PM):

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

Notes:

- A4 will be posted tomorrow
- I will catch up on Piazza/email today and tomorrow

Feb Fridays: Valerie Thomas (1943-)

Data Scientist (Mathematician & Data Analyst)

- Developed the digital media formats for image processing systems used by NASA
- Led teams of up to 50 people at NASA
- Invented the illusion transmitter (holds patent)
- Has mentored students from elementary through university; judged science fairs for decades
- Recipient of the Goddard Space Flight Center Award of Merit and the NASA Equal Opportunity Medal

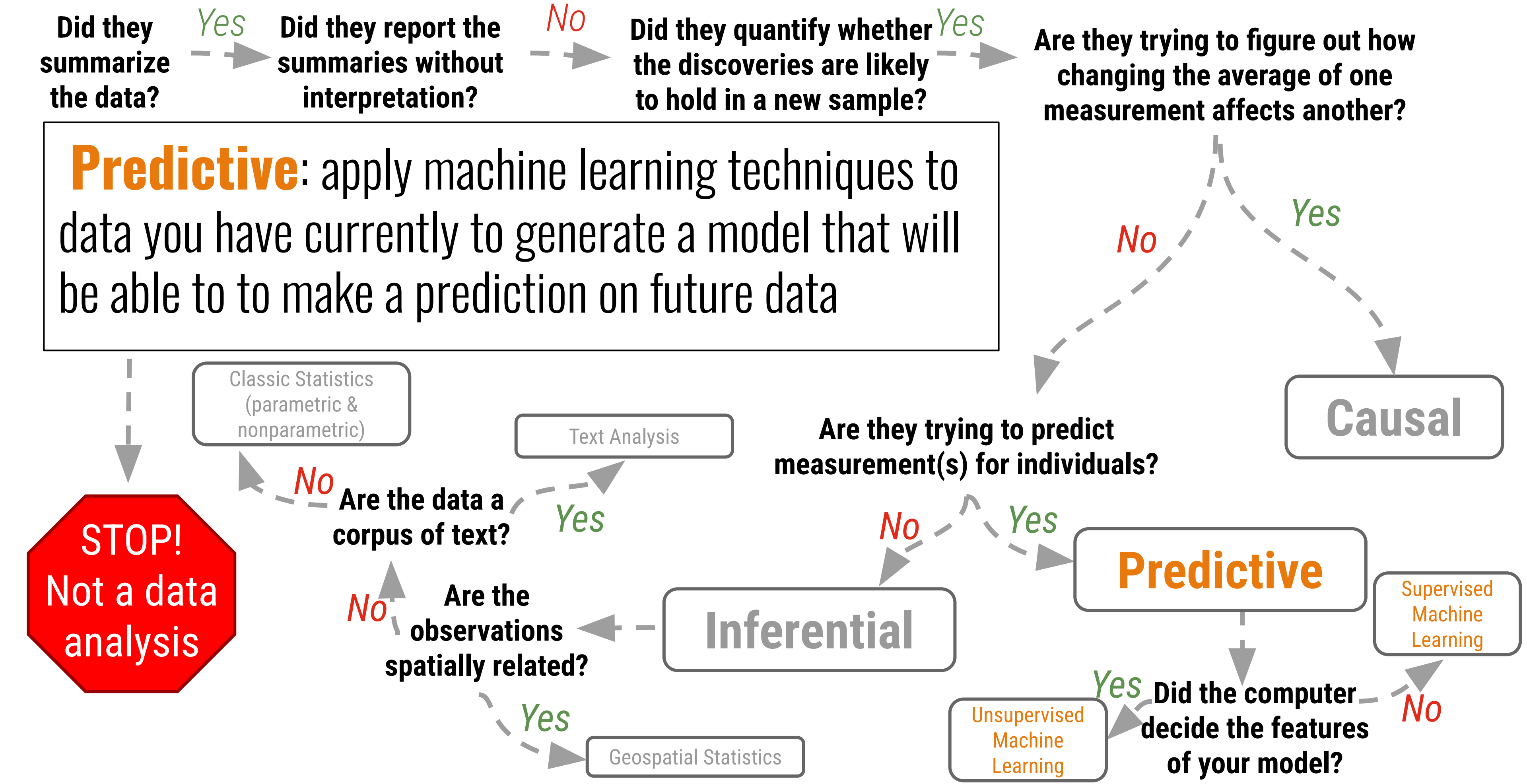
Also see: Mary Jackson, Katherine Johnson and Dorothy Vaughan (computers behind NASA's early space missions; depicted in Hidden Figures)

Machine Learning I

Sid Joshi
UC San Diego

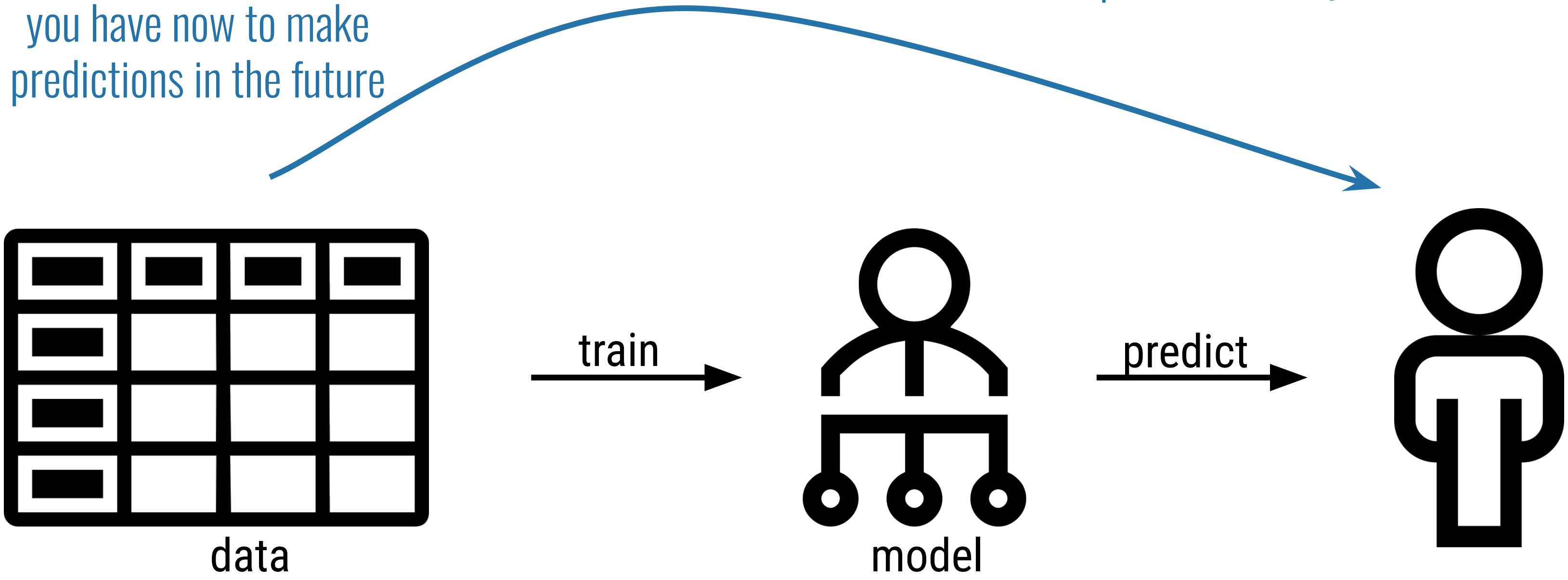
Department of Cognitive Science
s1joshi@ucsd.edu

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



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

machine learning approaches are used for predictive analysis!

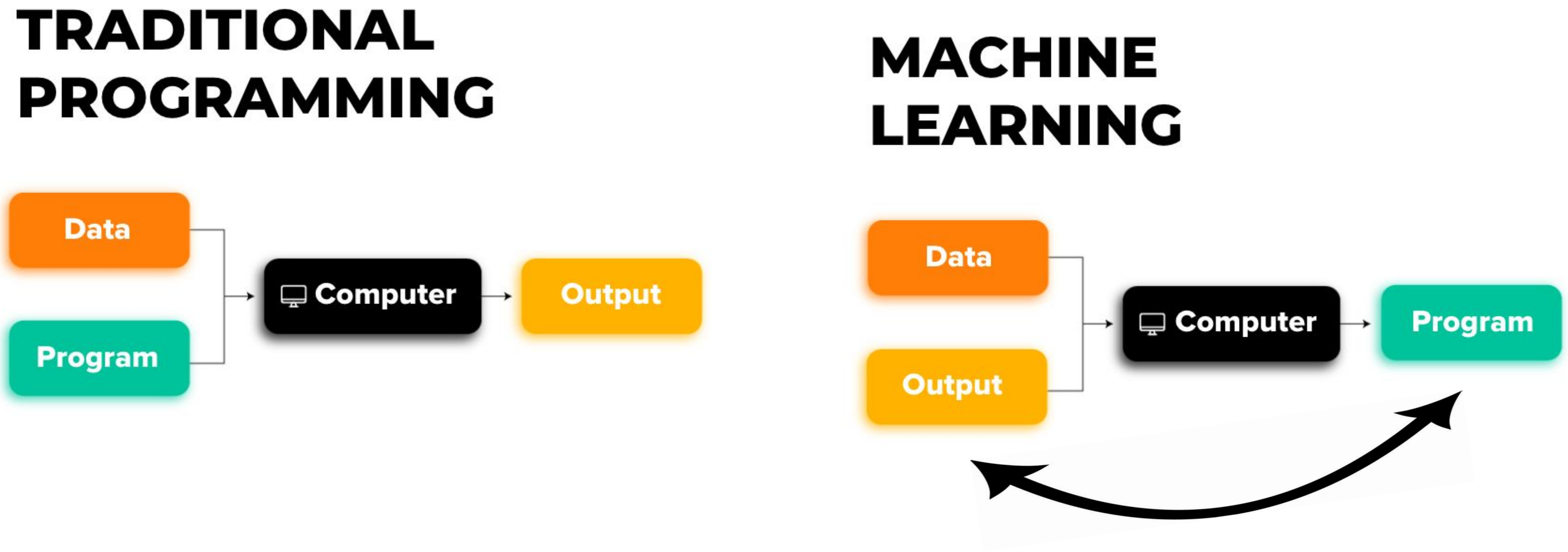


What is machine learning?

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

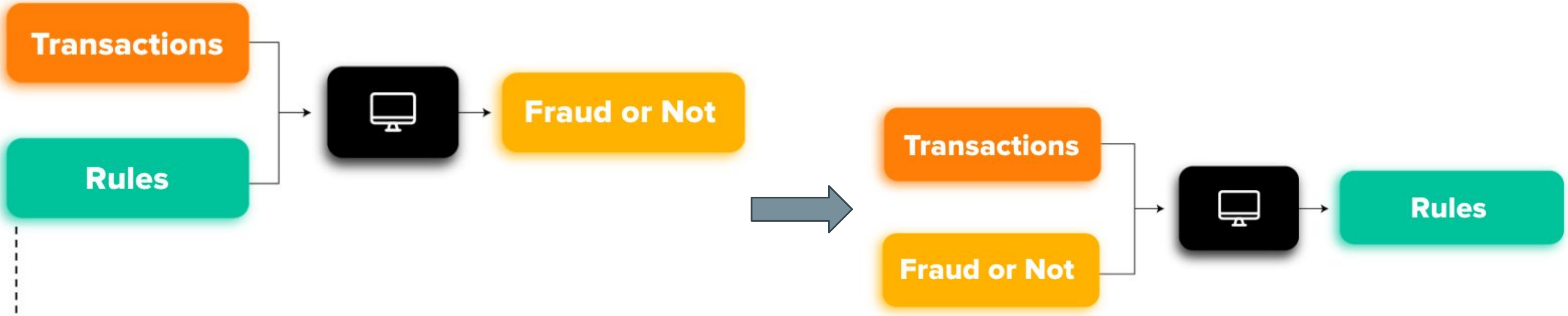
- Andrew Ng, Stanford, ex-Google, chief scientist at Baidu, Coursera founder, Stanford Adjunct Faculty

A Shift in Programming Paradigm



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



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?

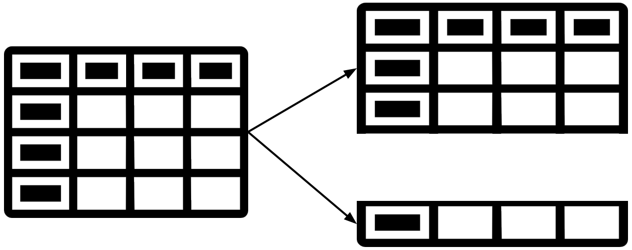
D Can we use information about one's early childhood to predict their success in elementary school?

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

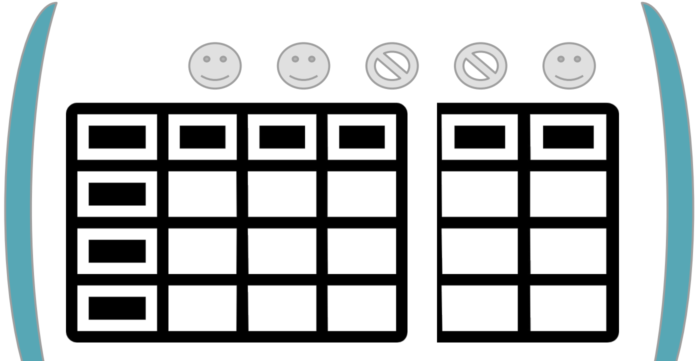
Machine Learning Generalizations

The goal of all models is to simplify, not replicate

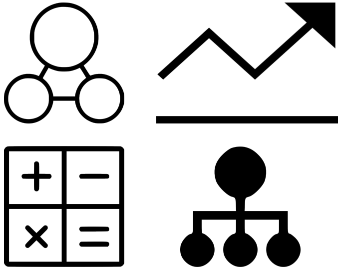
Basic Steps to Prediction



data
partitioning

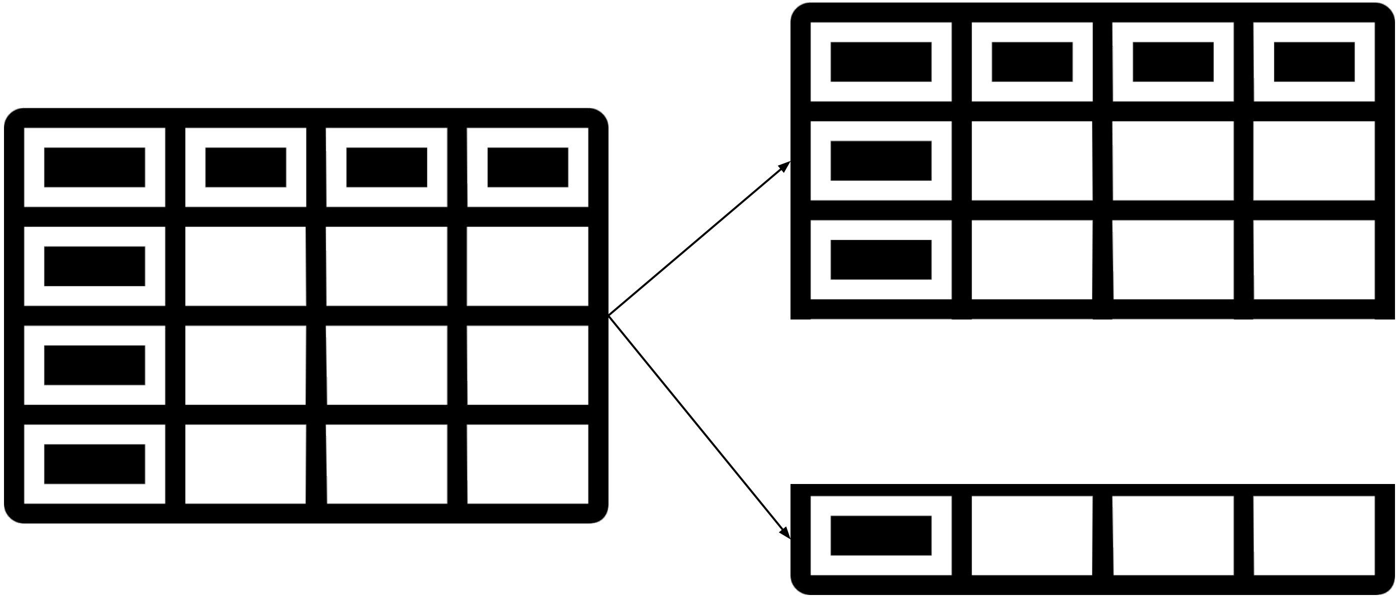


feature selection

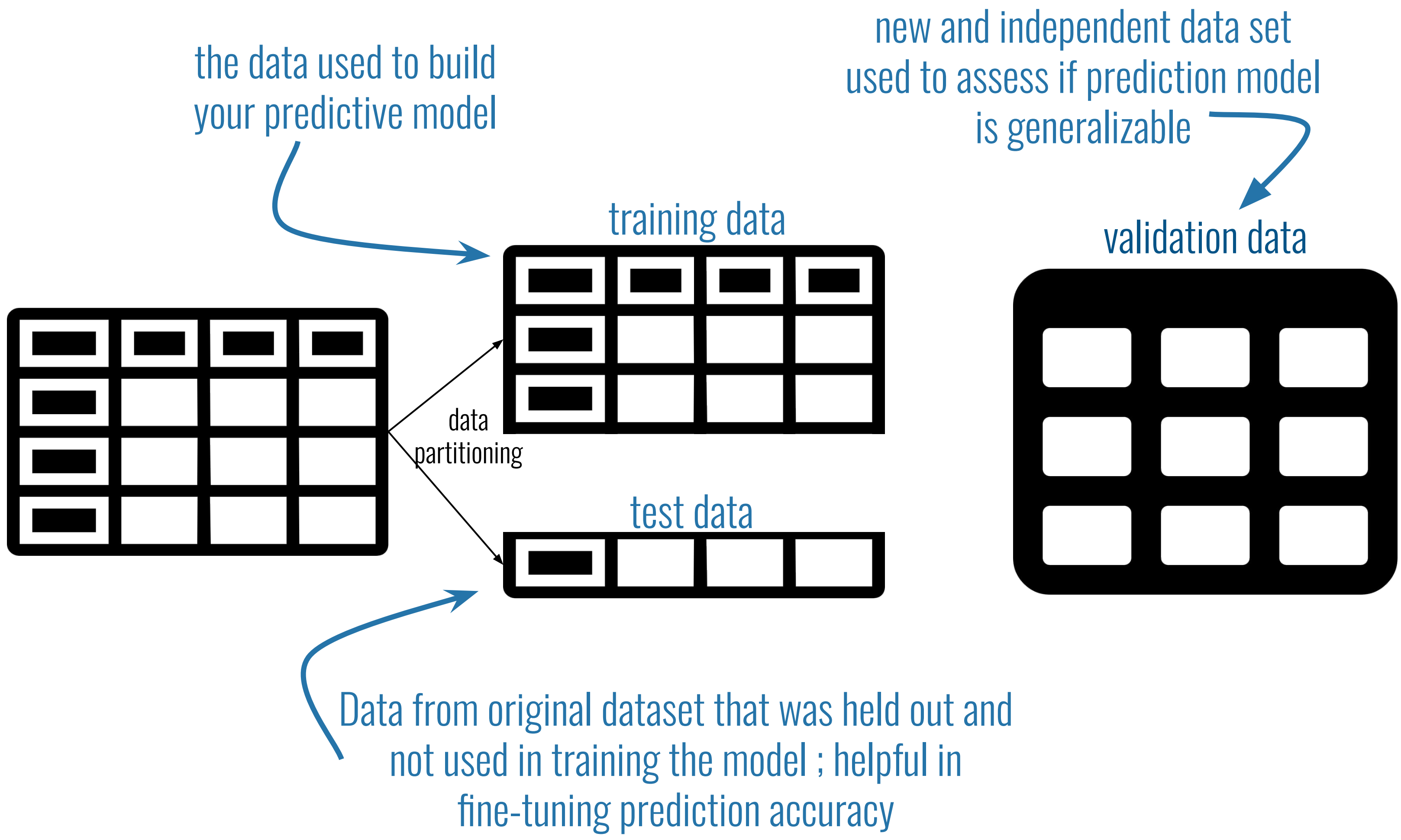


model selection

model assessment



data partitioning


the data used to build your predictive model
new and independent data set used to assess if prediction model is generalizable
training data
validation data
data partitioning
test data
Data from original dataset that was held out and not used in training the model ; helpful in fine-tuning prediction accuracy

Data Partitioning

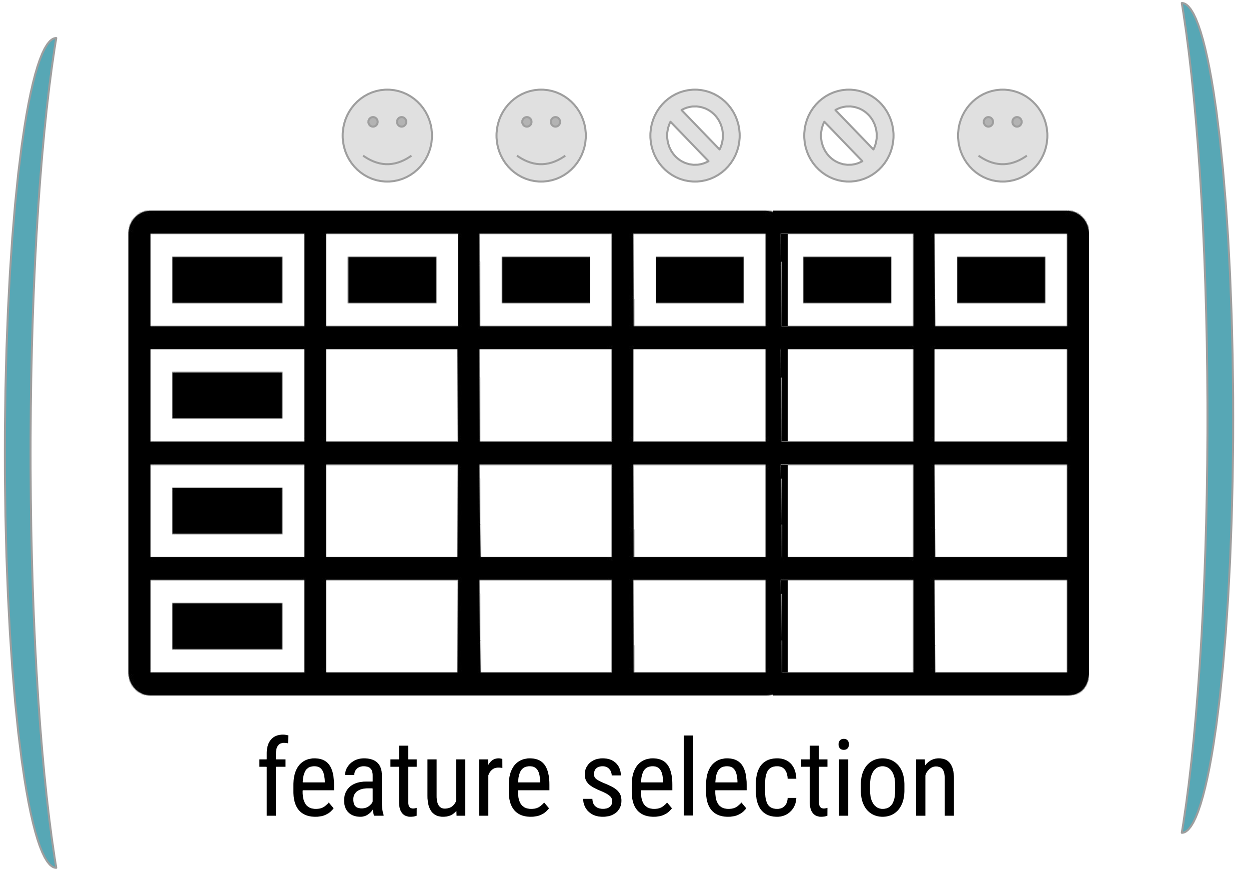
What portion of the data are typically used for generating the model?

A
The entire dataset

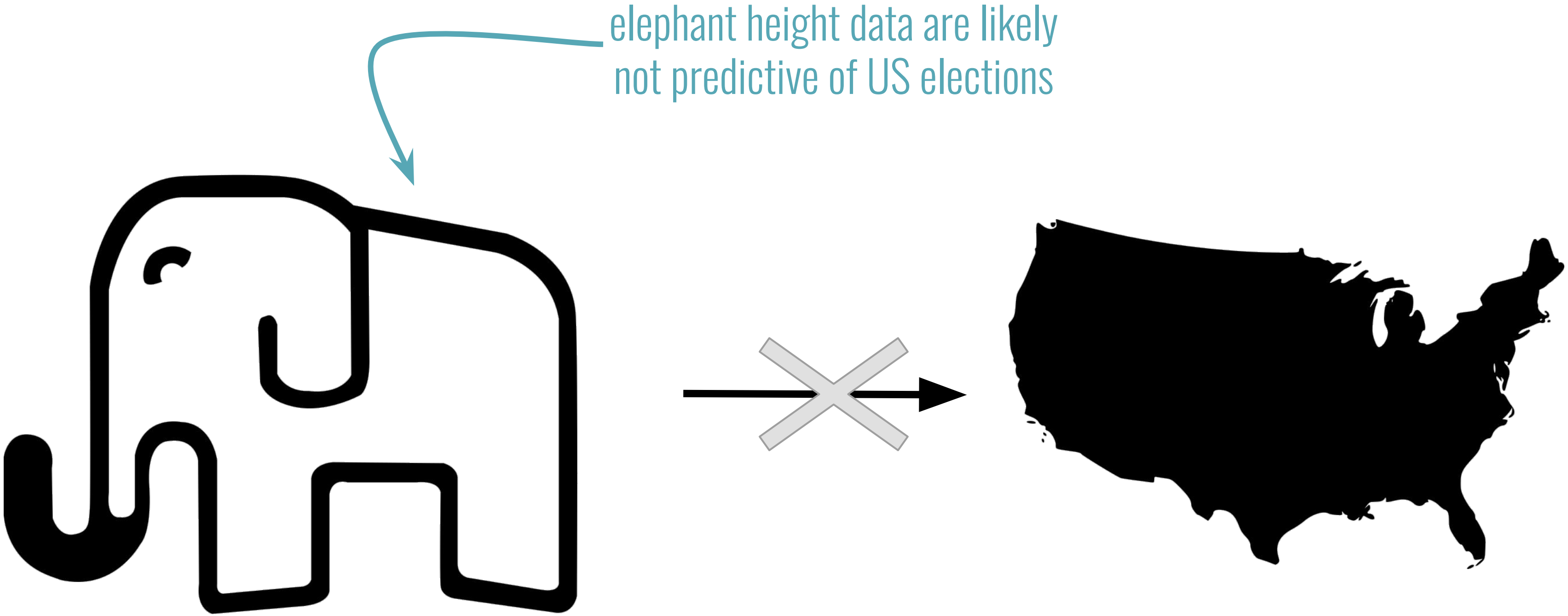
B
The training data

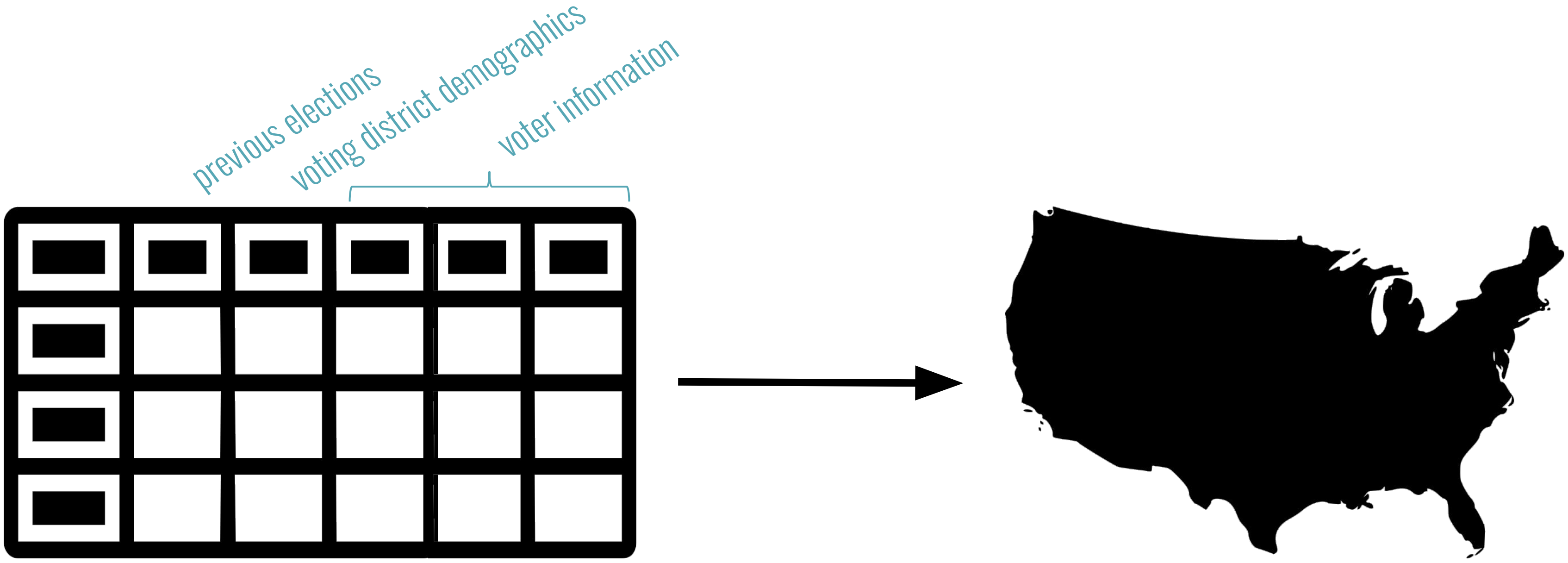
C
The testing data

D
The validation data

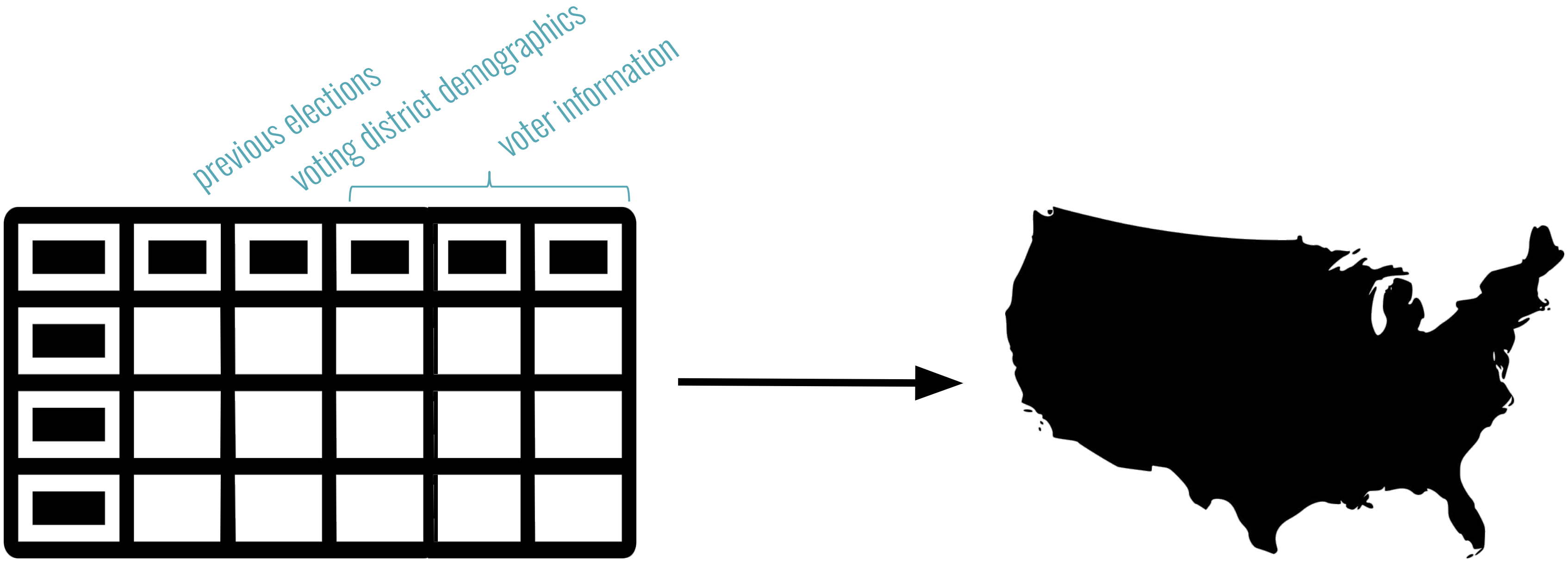


feature selection

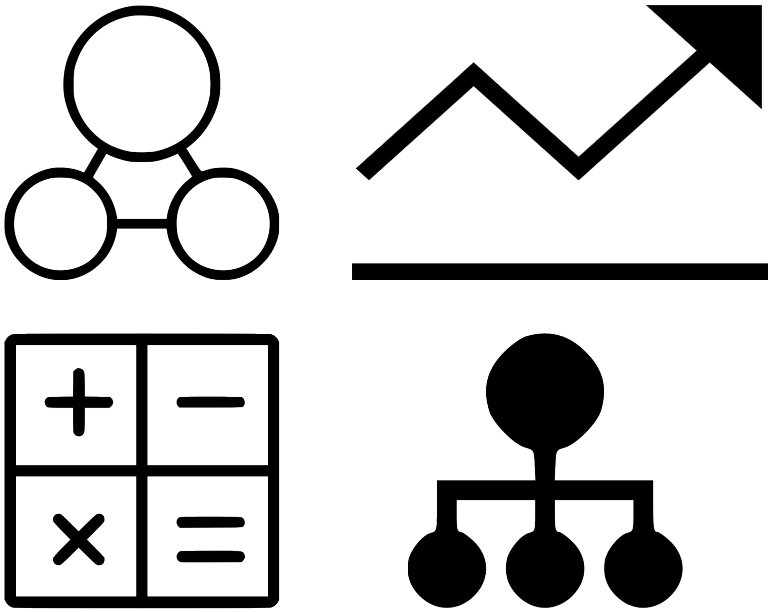

elephant height data are likely
not predictive of US elections



these data are likely predictive
of US election outcomes



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



model selection

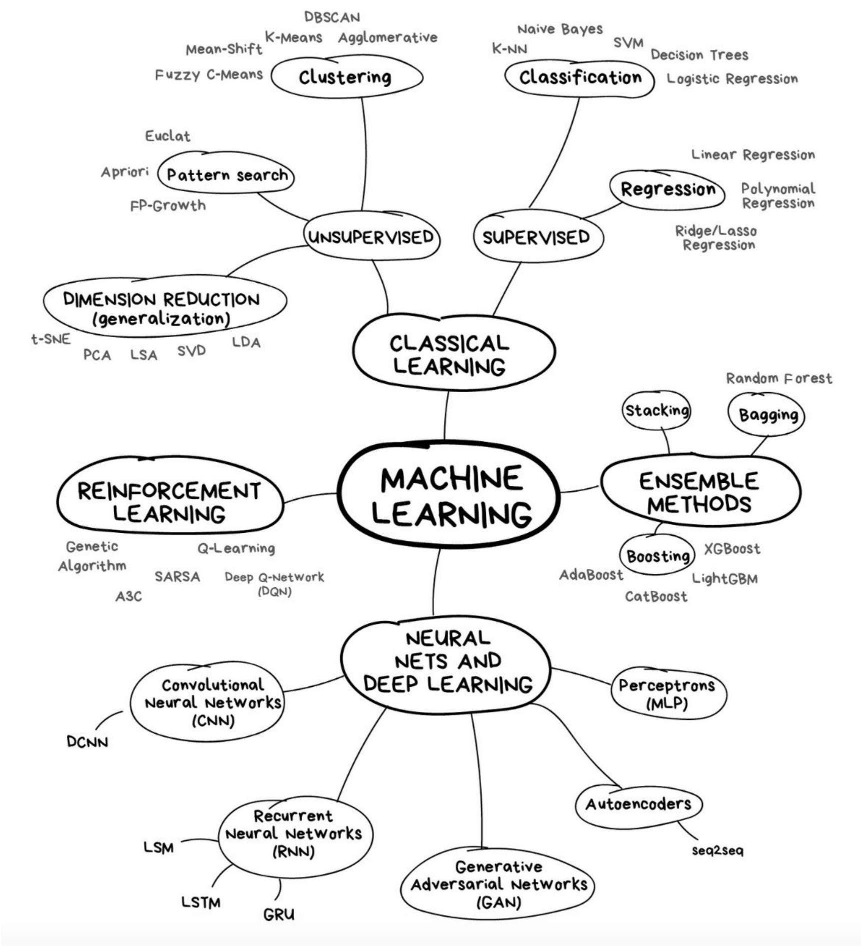


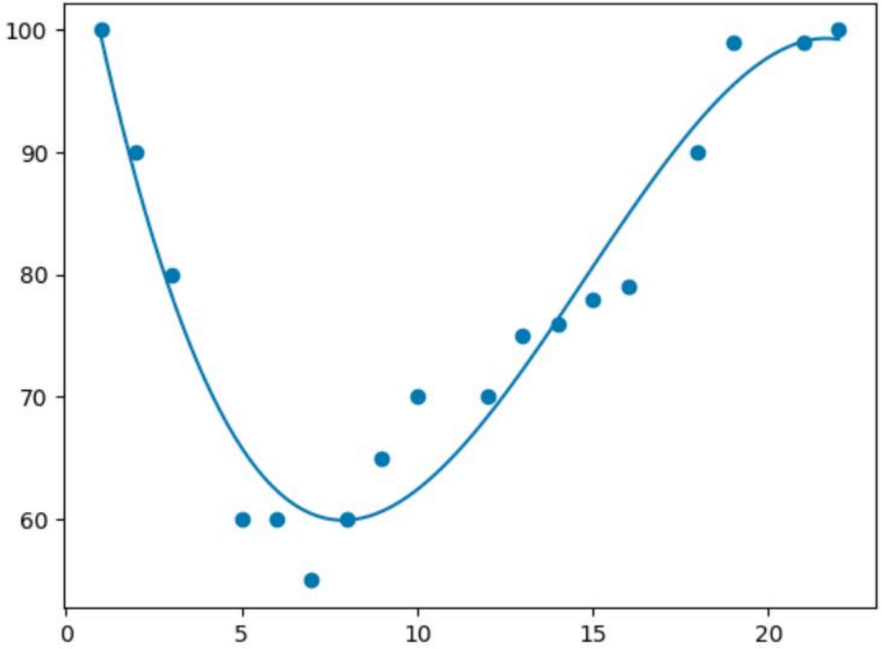
Image source: https://vas3k.com/blog/machine_learning/

Selecting the Right Model

A - Thought about it and have a thought

B - Thought about it and have no thought

C - I'm confused

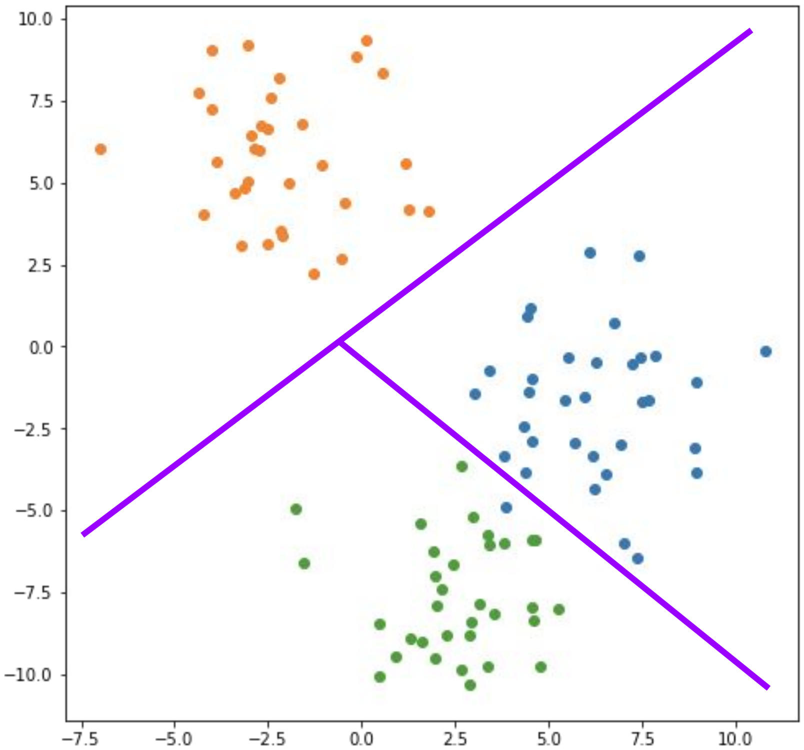


Selecting the Right Model

A - Thought about it and have a thought

B - Thought about it and have no thought

C - I'm confused

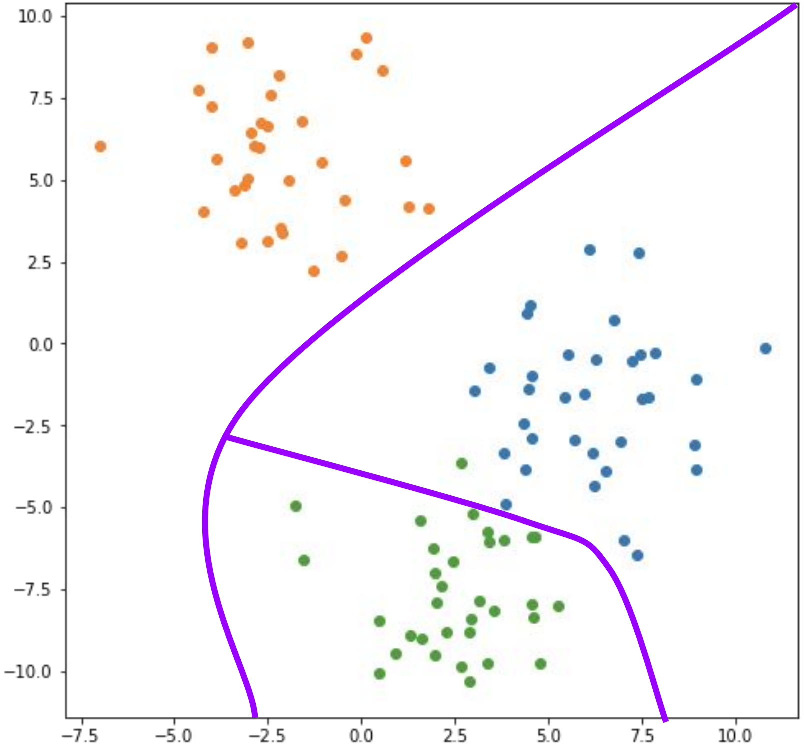


Selecting the Right Model

A - Thought about it and have a thought

B - Thought about it and have no thought

C - I'm confused

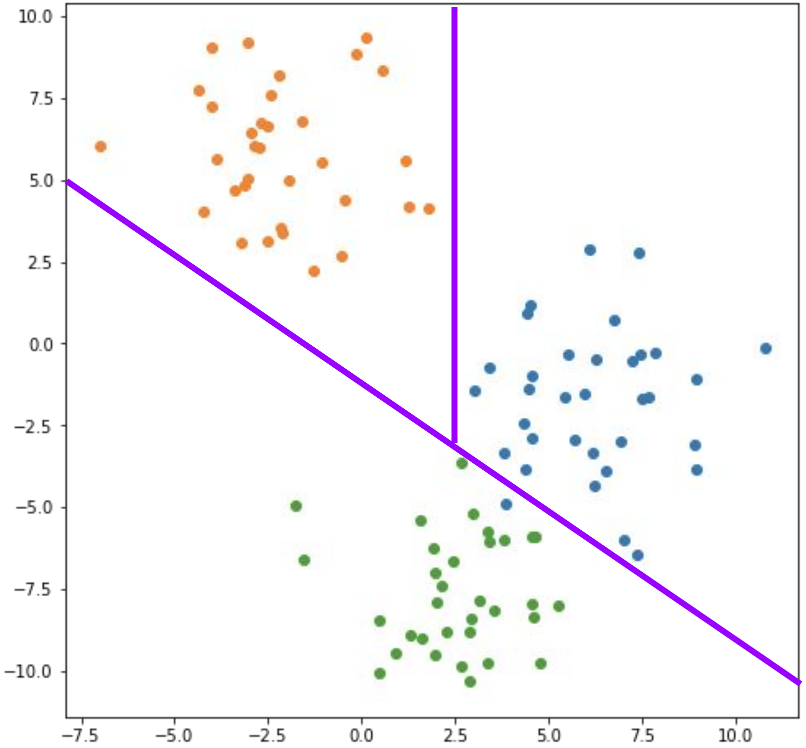


Selecting the Right Model

A - Thought about it and have a thought

B - Thought about it and have no thought

C - I'm confused

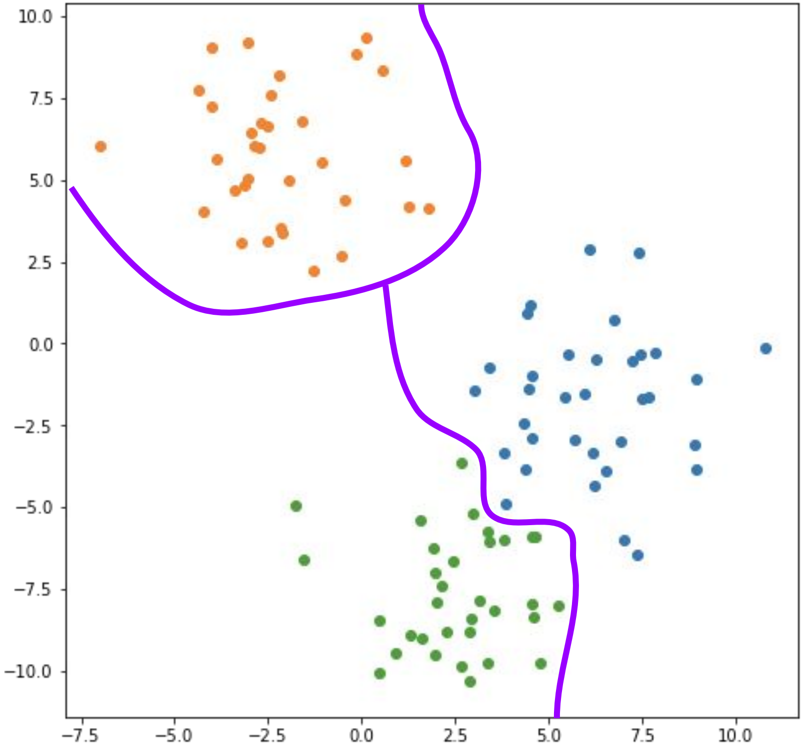


Selecting the Right Model

A - Thought about it and have a thought

B - Thought about it and have no thought

C - I'm confused



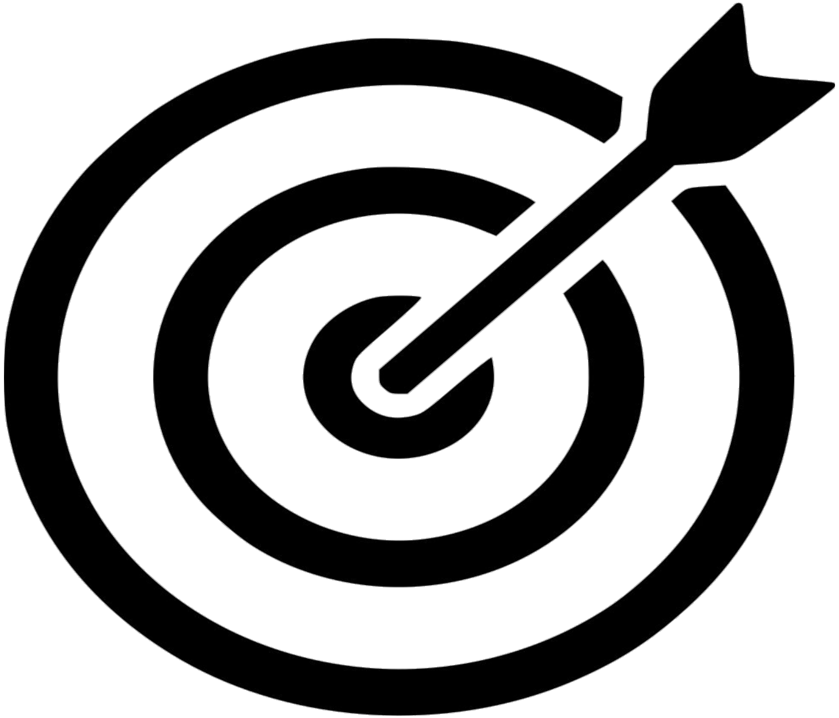
Cross-Validation

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 partition we left out earlier!
  - Training data was the study guide
  - Validation data was the practice exam
  - Testing data are the actual exam
- Performance assessments depend on the specific problem/question
- Different metrics measure different aspects of models
  - e.g. RMSE, MAE, accuracy, specificity/sensitivity, etc.

Classical Machine Learning

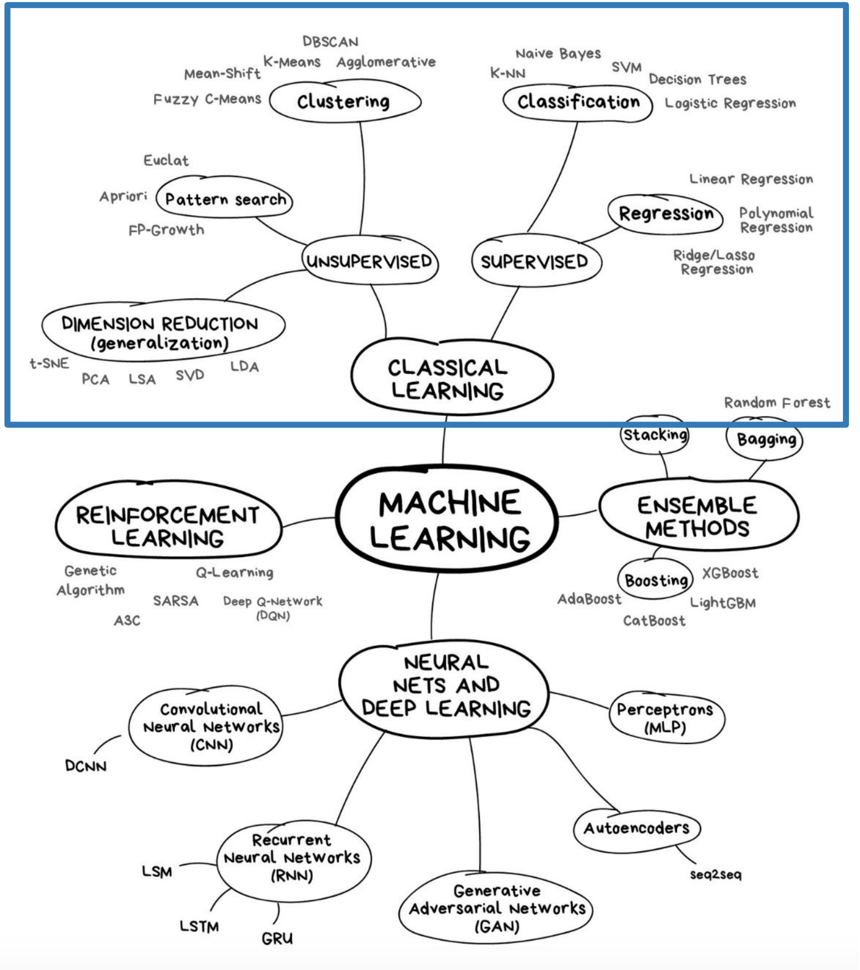


Image source: https://vas3k.com/blog/machine_learning/

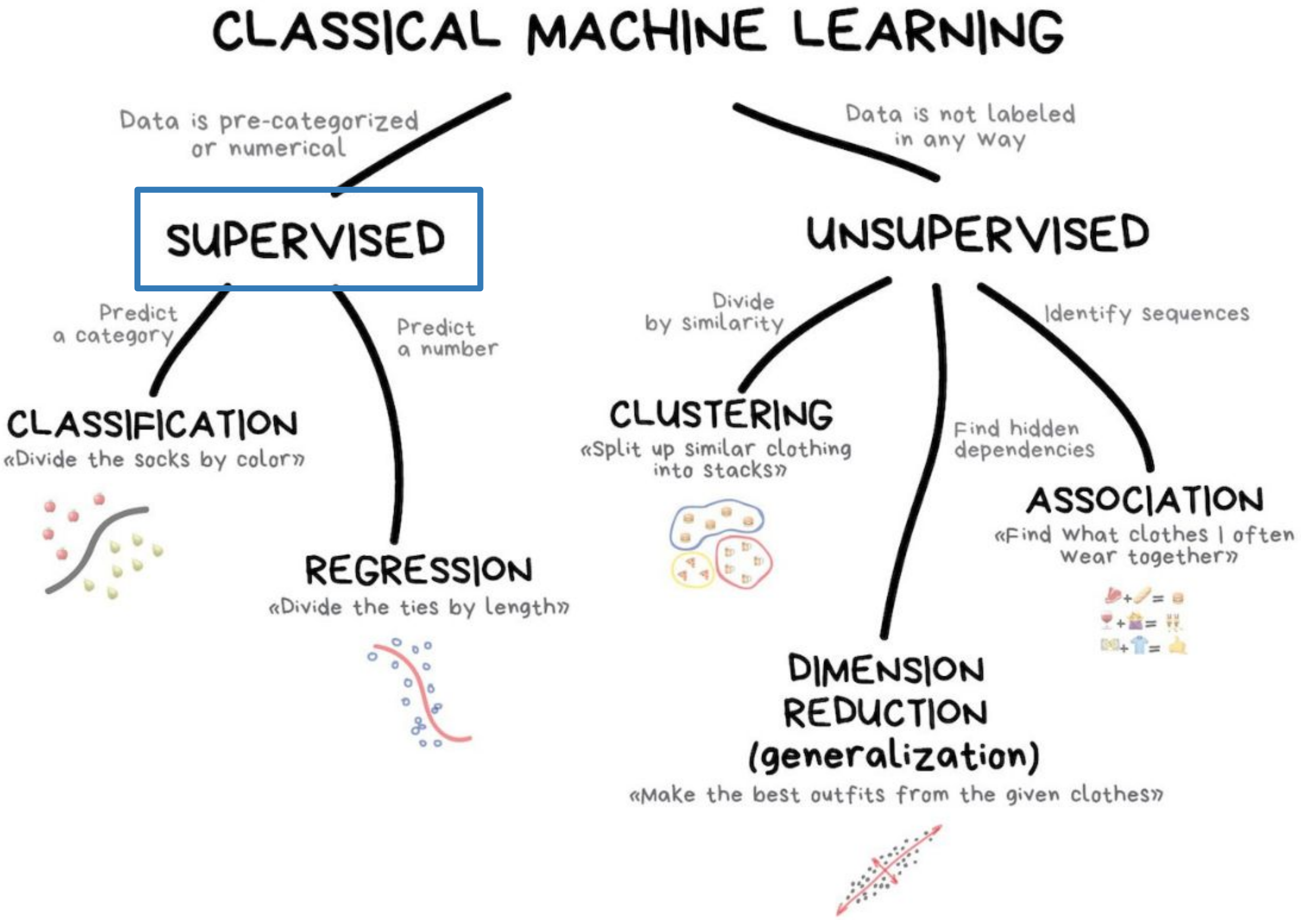
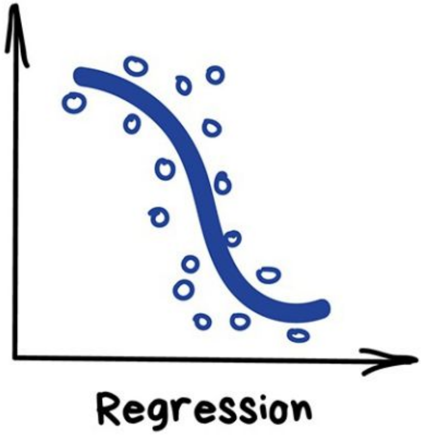


Image source: https://vas3k.com/blog/machine_learning/

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



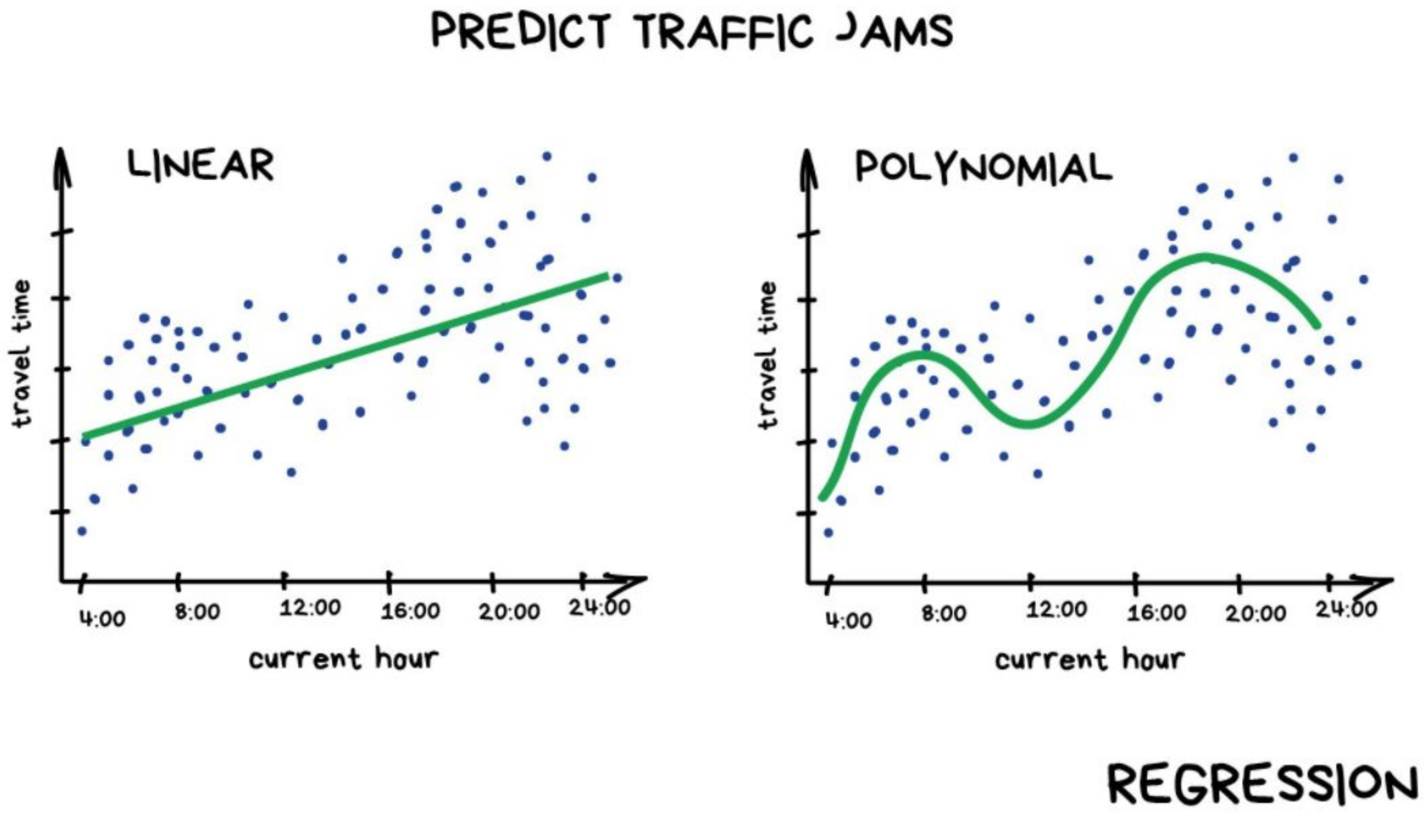


Image source: https://vas3k.com/blog/machine_learning/

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

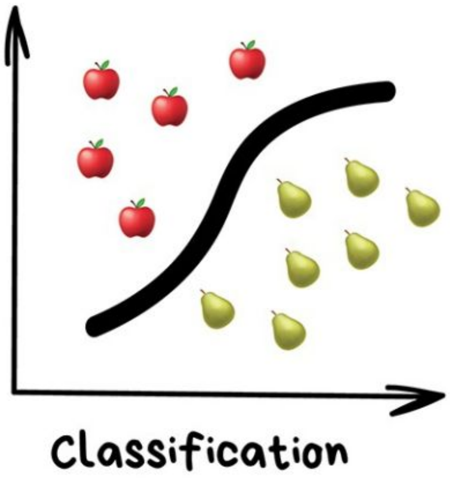


Image source: https://vas3k.com/blog/machine_learning/

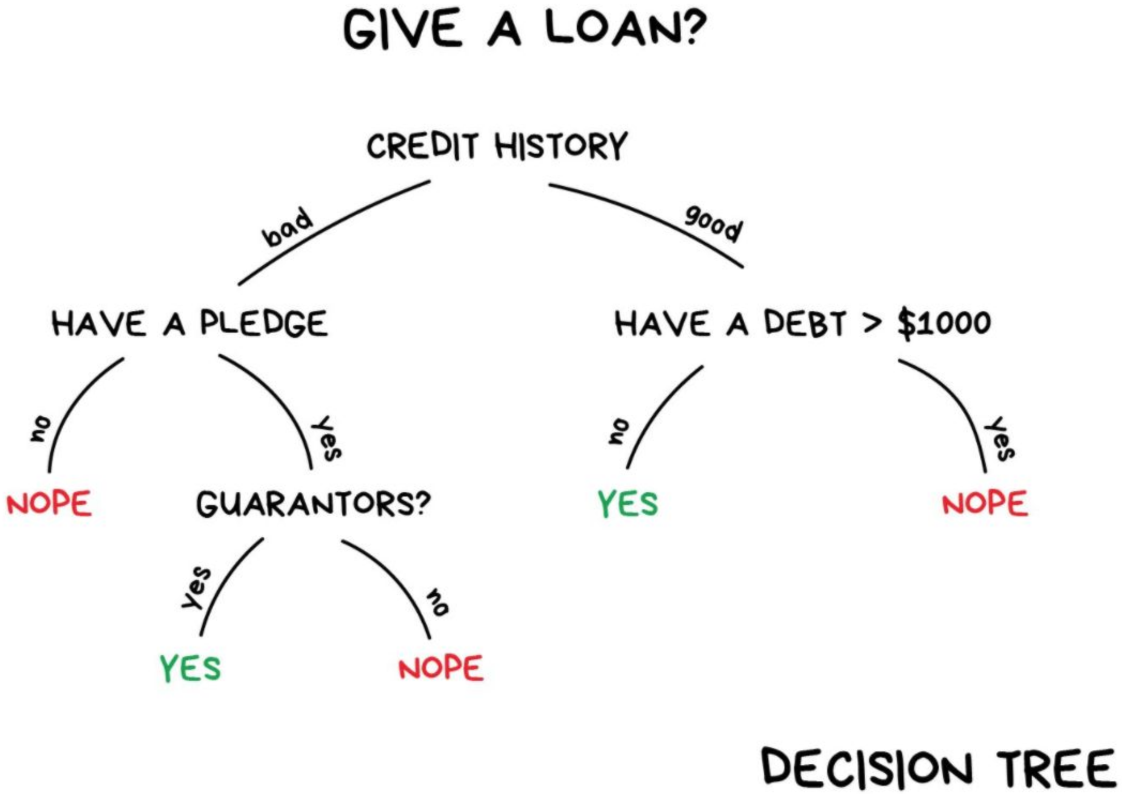


Image source: https://vas3k.com/blog/machine_learning/

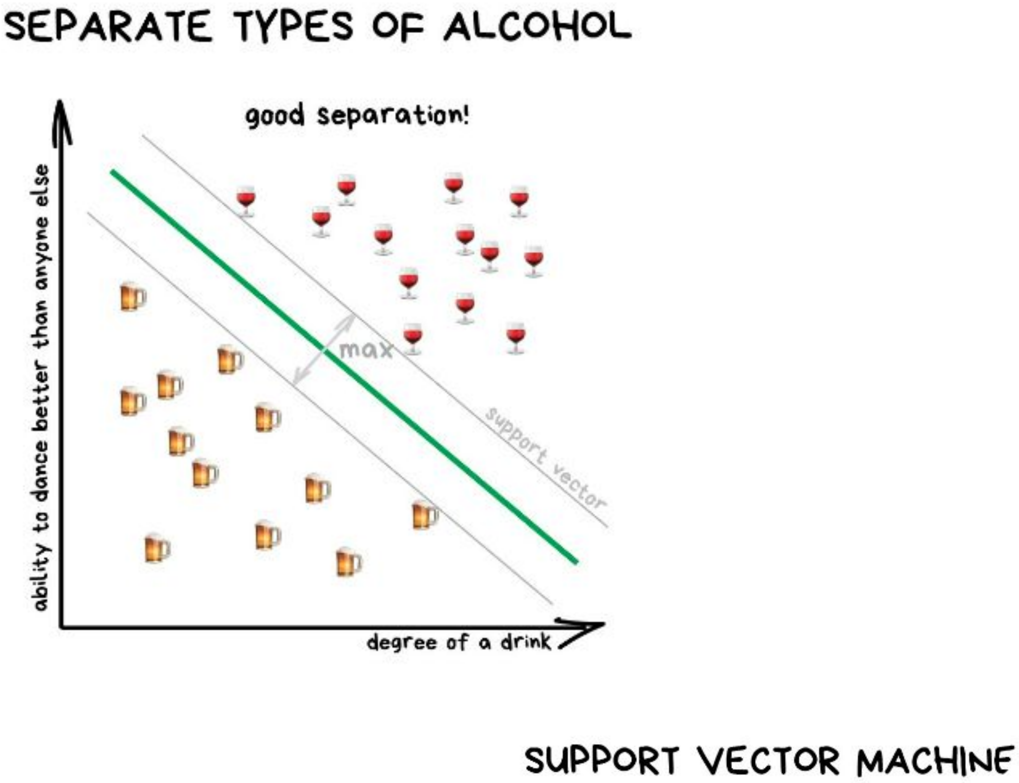


Image source: https://vas3k.com/blog/machine_learning/

Supervised vs. Unsupervised

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

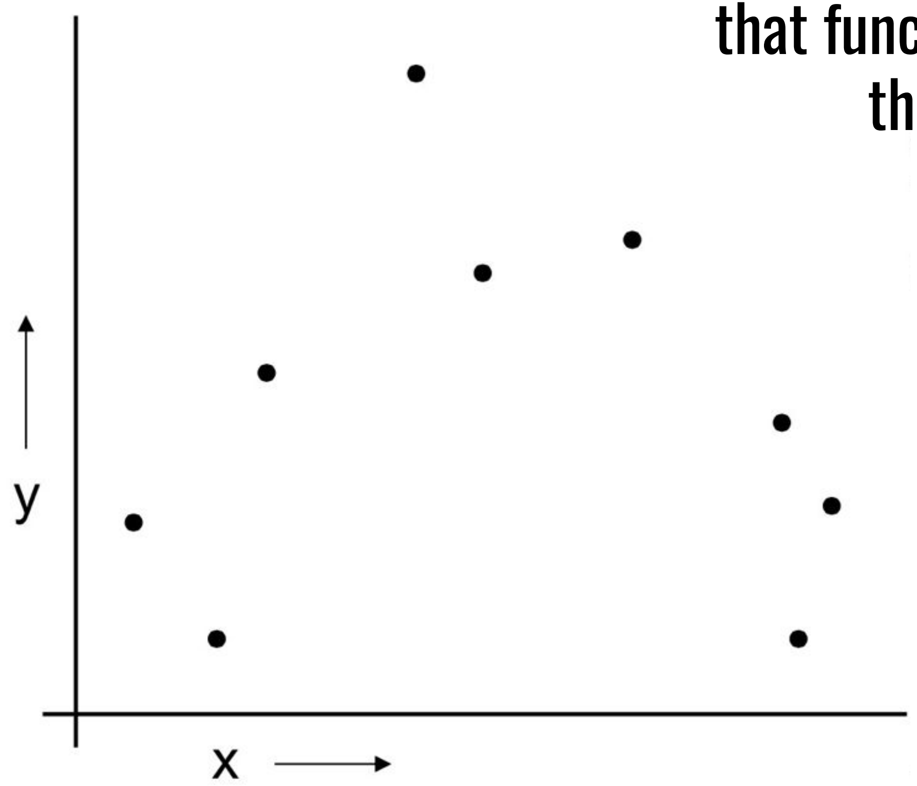
- Learning relationships between inputs and outputs to make predictions
- Goal is to minimize error or maximize accuracy in predictions

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

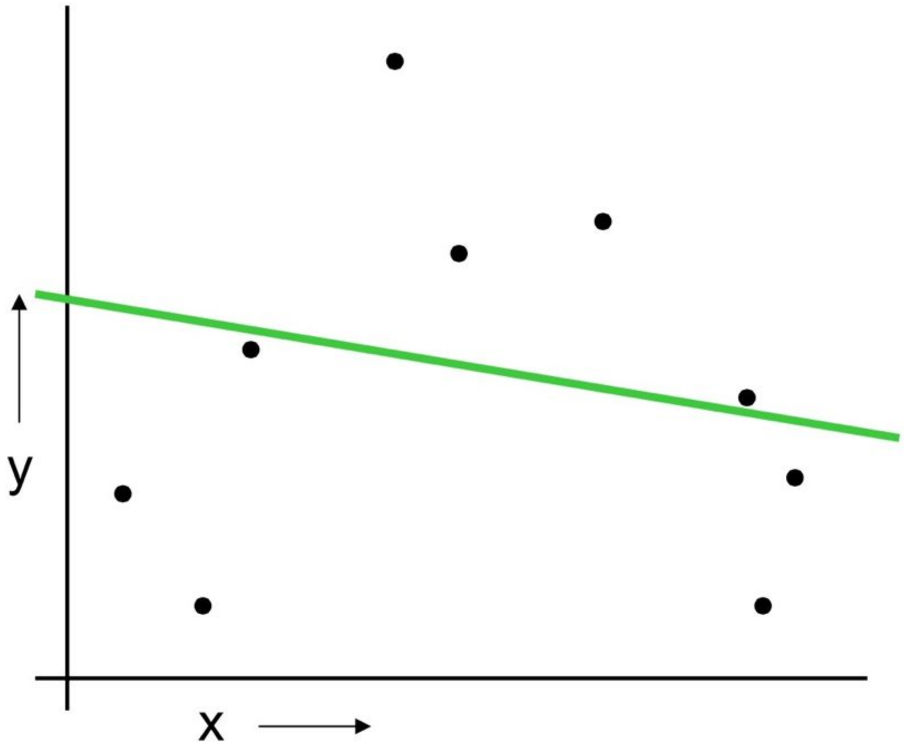
Regression Walkthrough

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

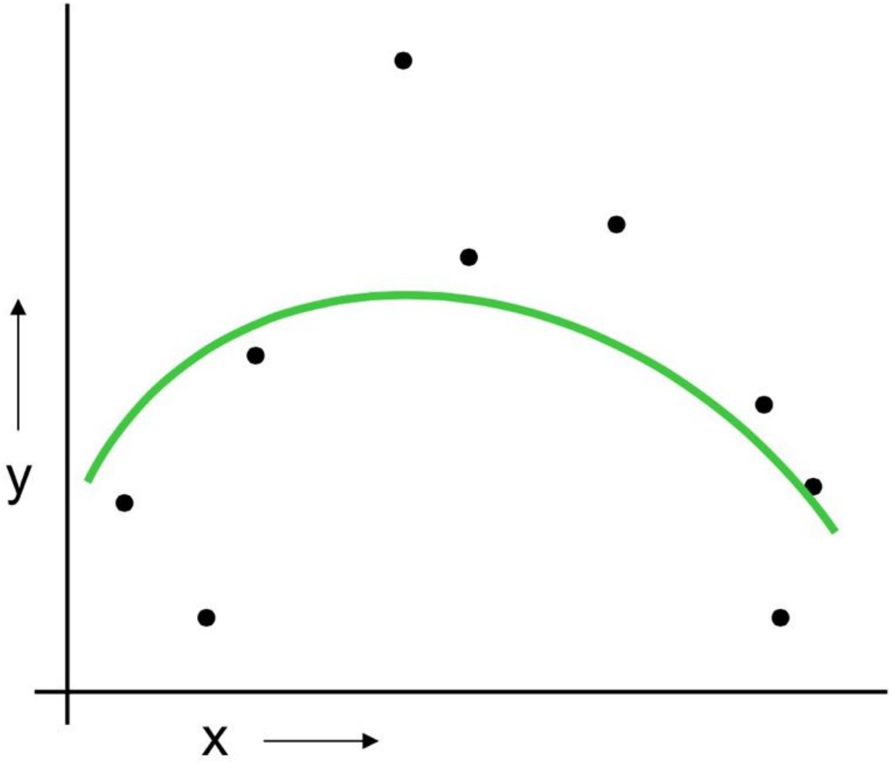


`y = f(x) + noise`

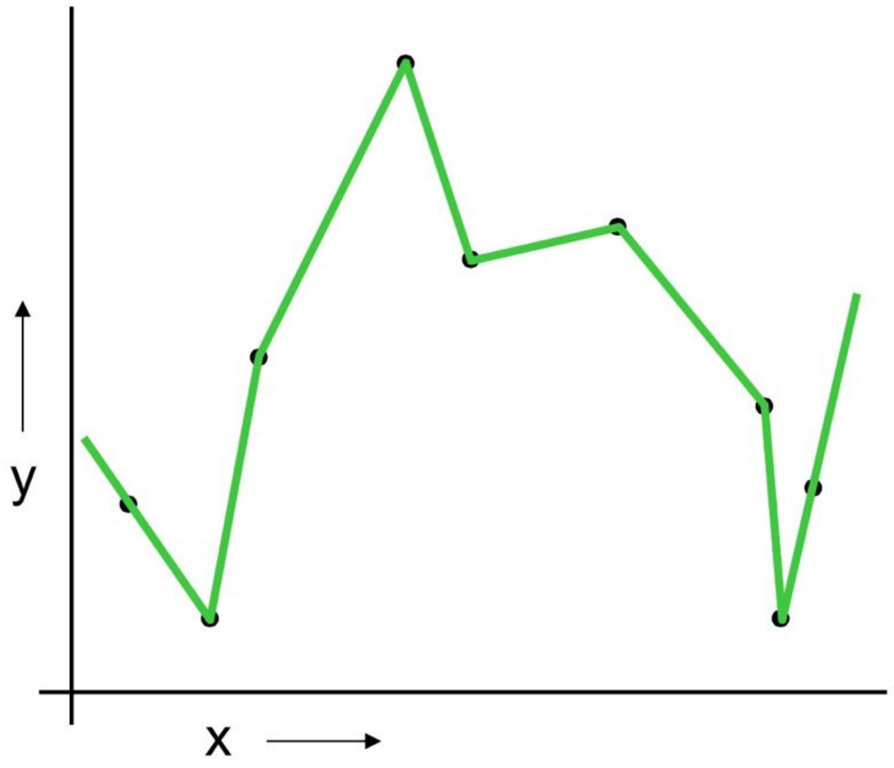
Linear regression



Quadratic regression

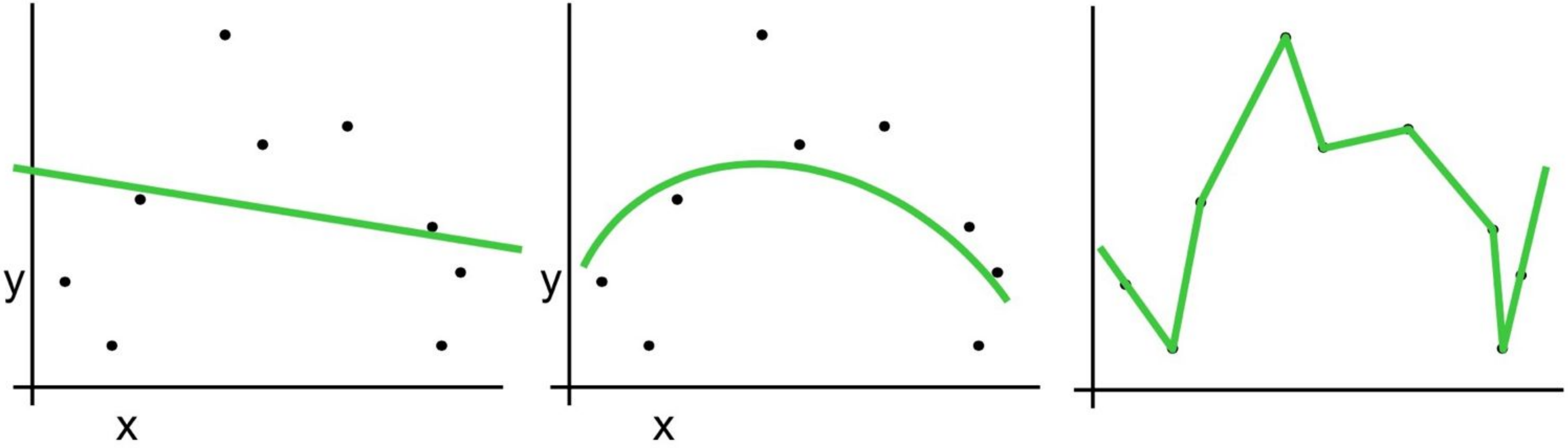


Piecewise linear nonparametric regression



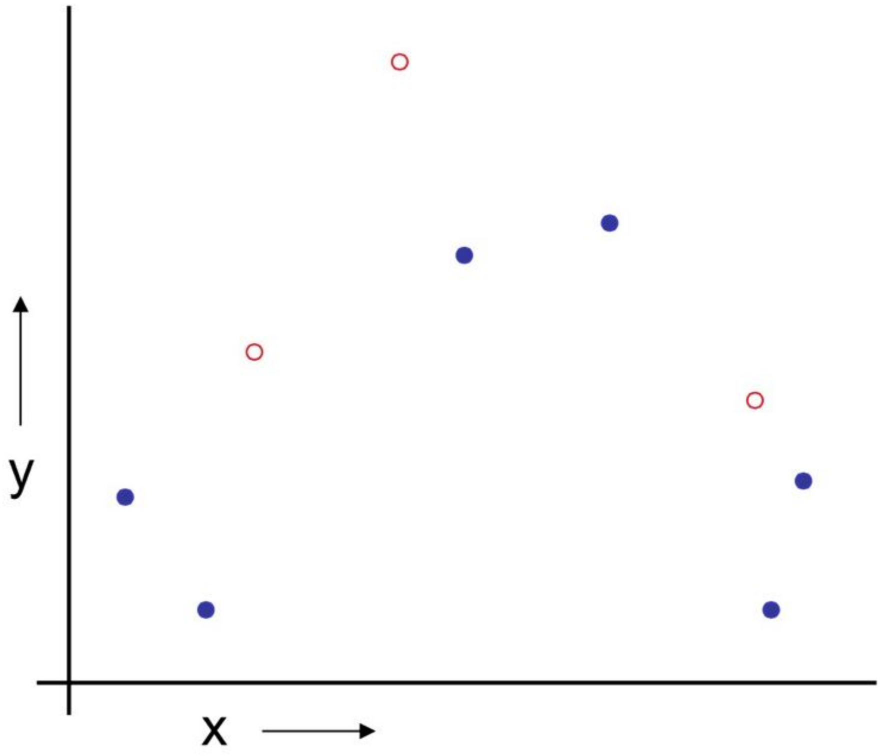
adapted from Brad Voytek

Which to choose?



adapted from Brad Voytek

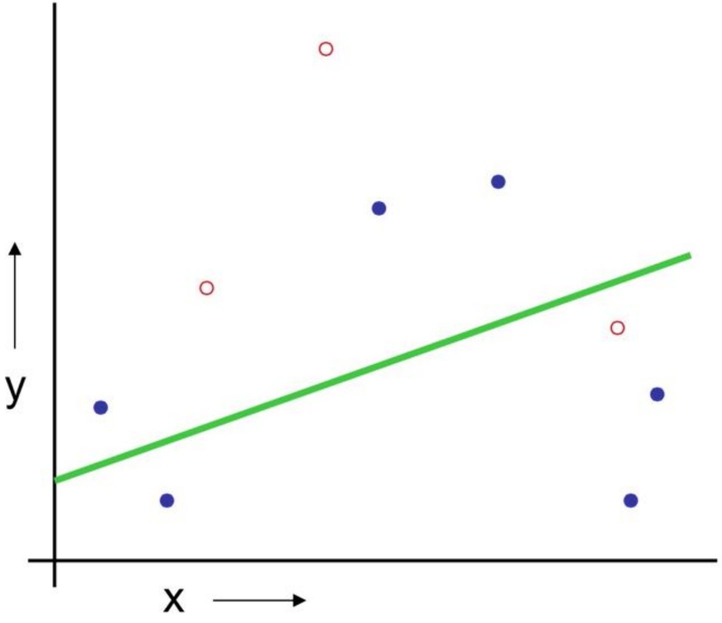
The data partition method



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

2. The remainder is a training set

Train the model on your training set



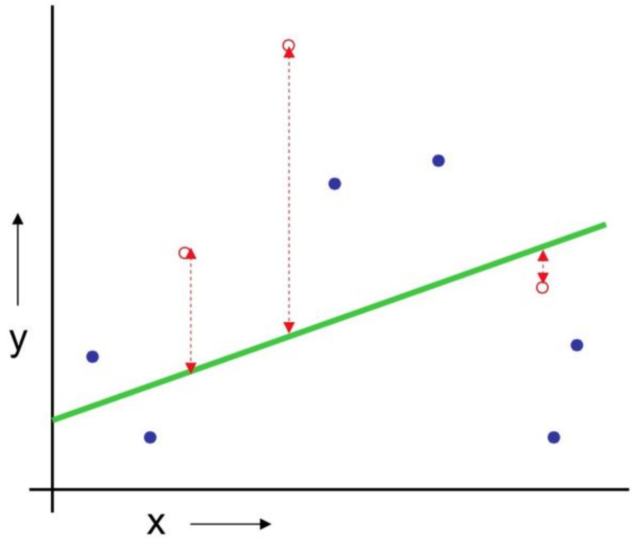
(Linear regression example)

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

Assess future performance using the test set



(Linear regression example)

Mean Squared Error = 2.4

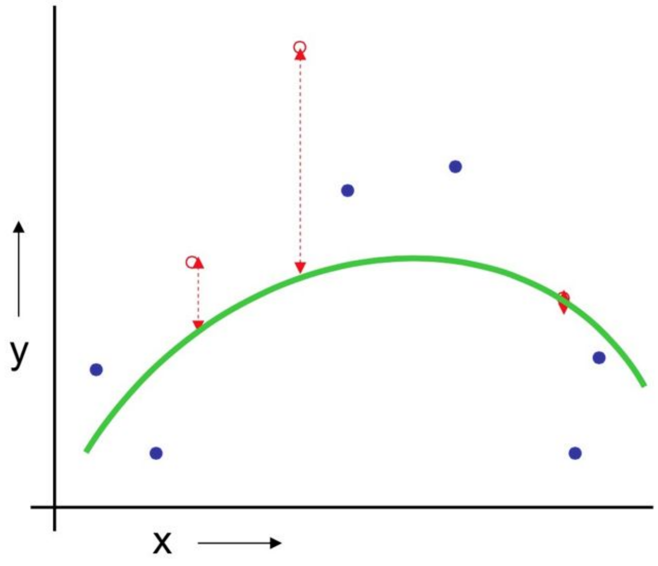
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

Go through this process for each possible model



(Quadratic regression example)

Mean Squared Error = 0.9

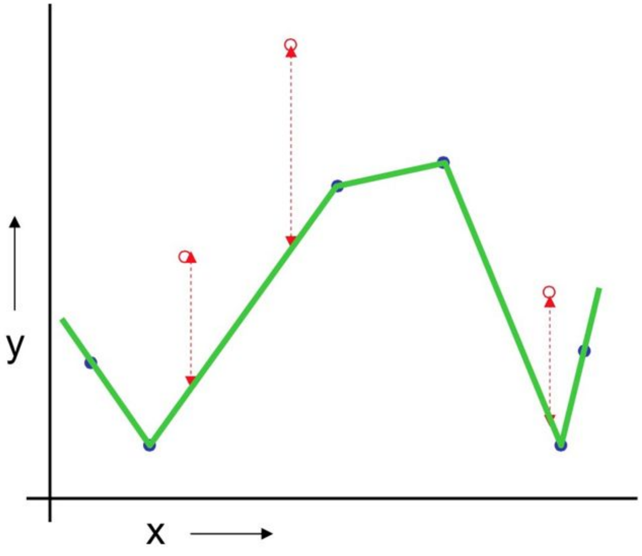
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

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

Pros and cons of data partitioning

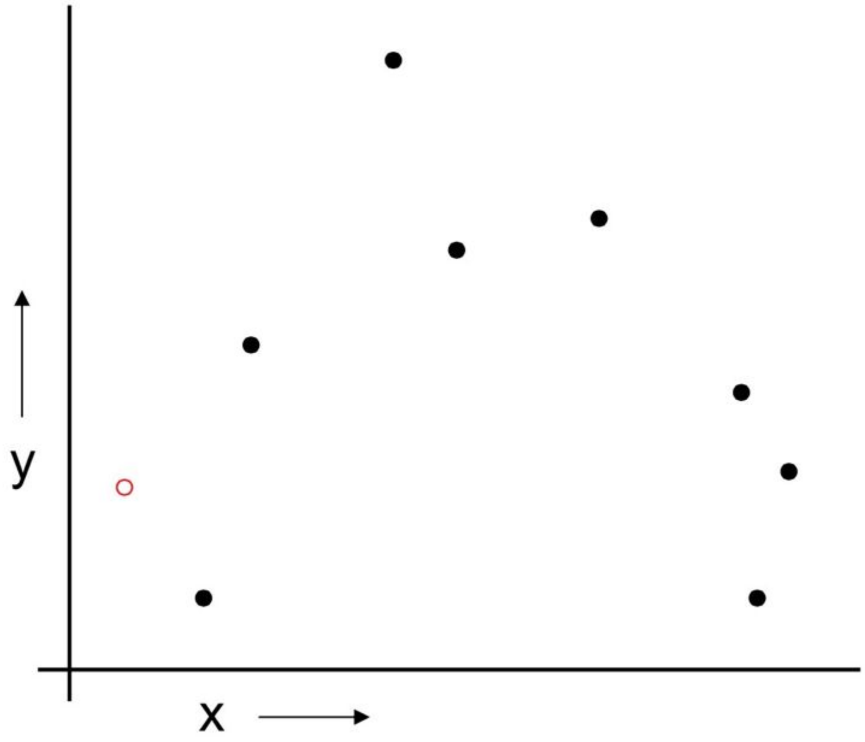
Pros:

- 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

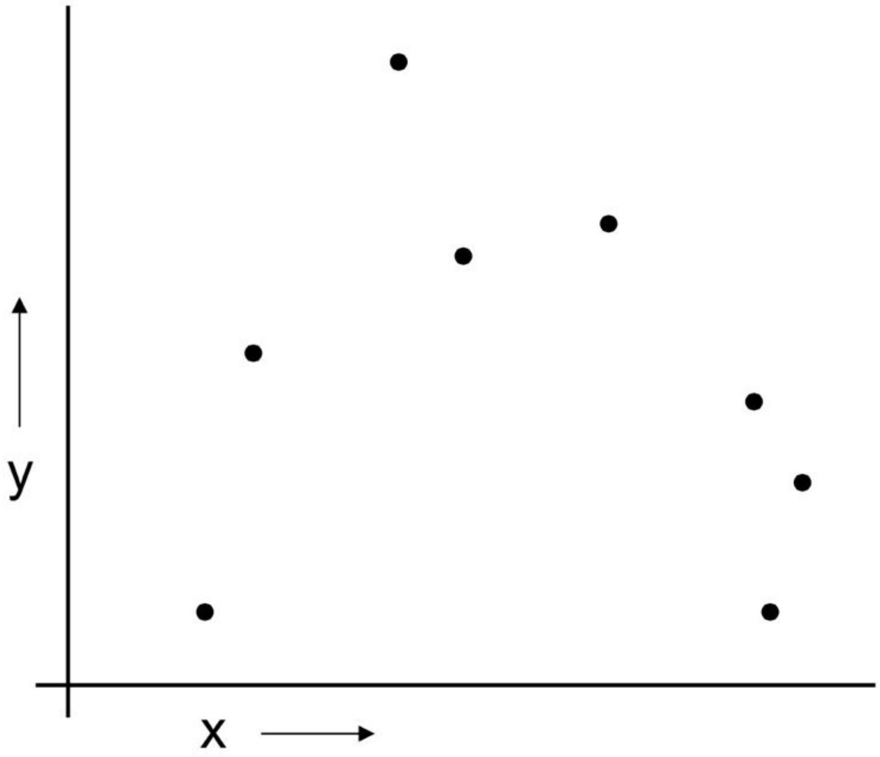
Leave one out cross validation (LOOCV)



For k=1 to R

1. Let (x_k, y_k) be the k^{th} record

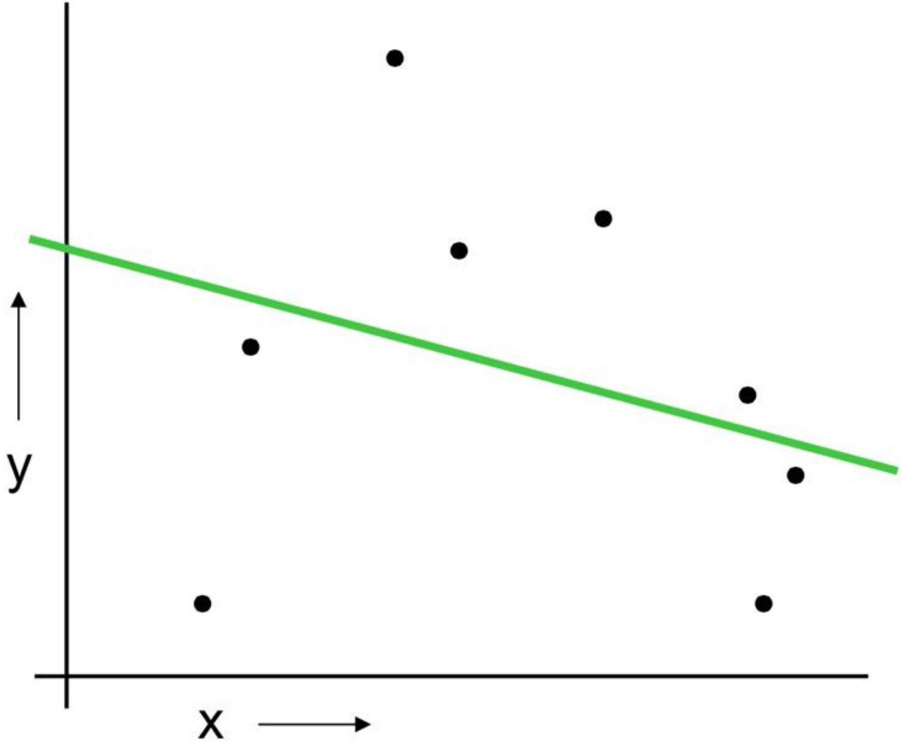
Leave one out cross validation (LOOCV)



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

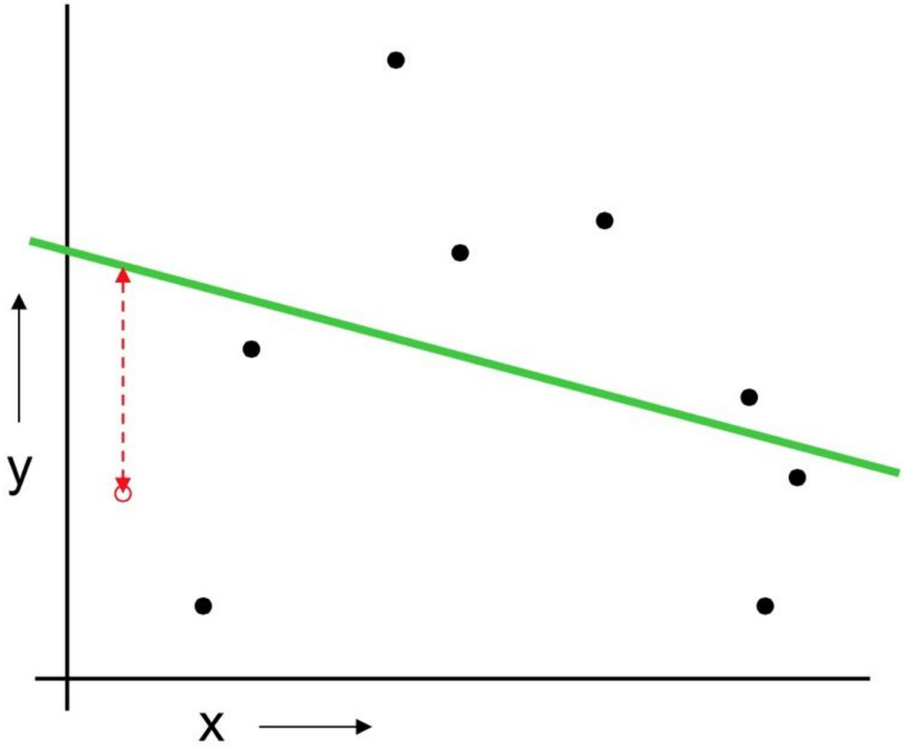
Leave one out cross validation (LOOCV)



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

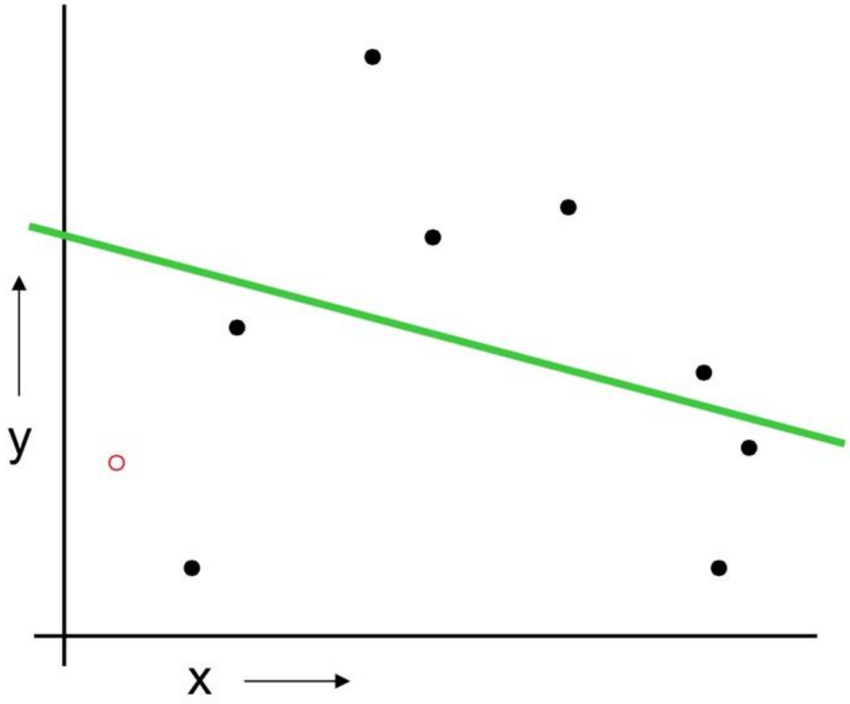
Leave one out cross validation (LOOCV)



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)

Leave one out cross validation (LOOCV)

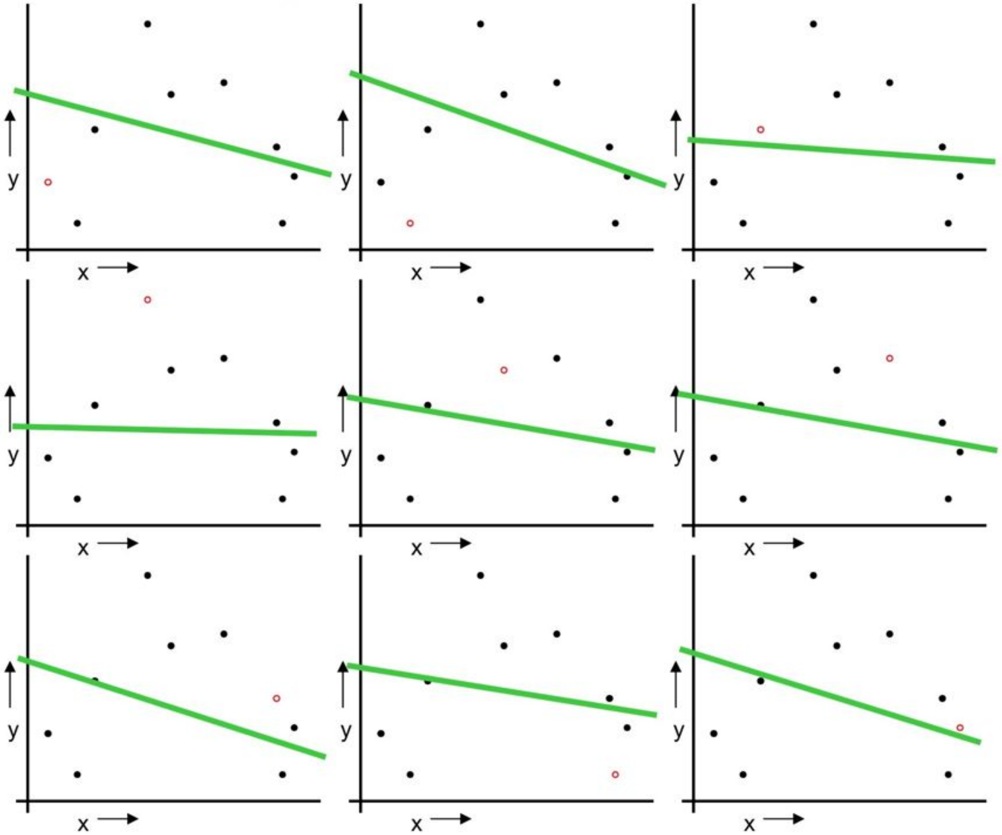


For k=1 to R

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.

Leave one out cross validation (LOOCV)



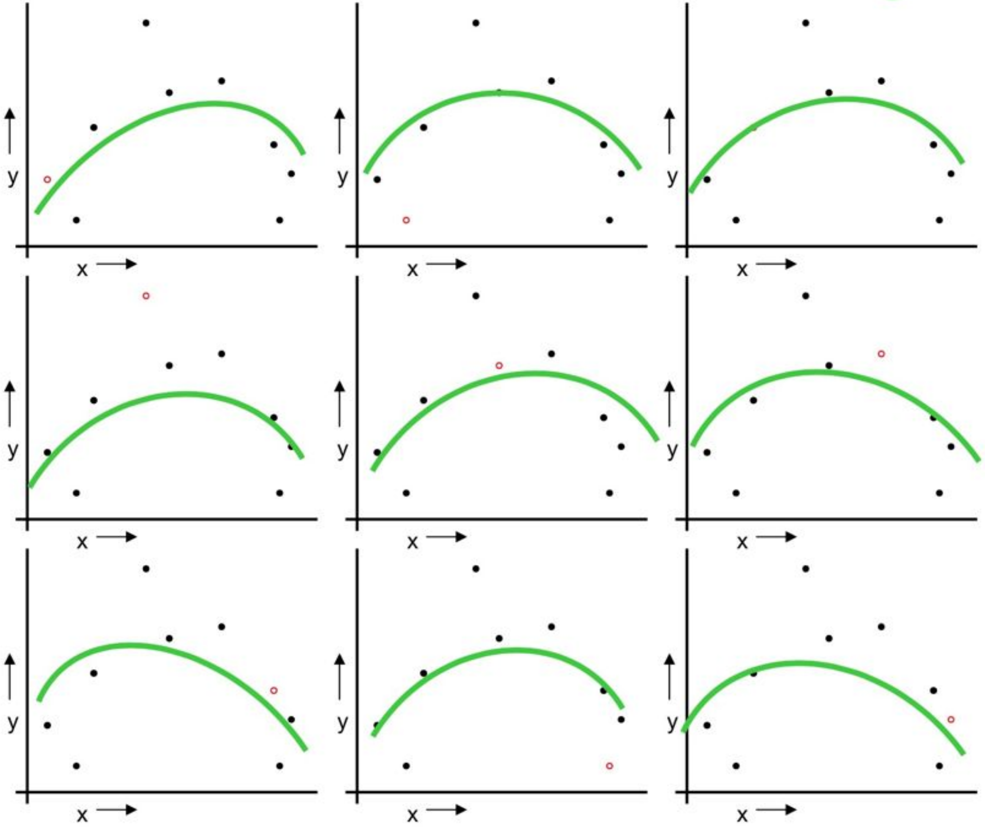
For k=1 to R

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.

$$MSE_{LOOCV} = 2.12$$

Leave one out cross validation (LOOCV)



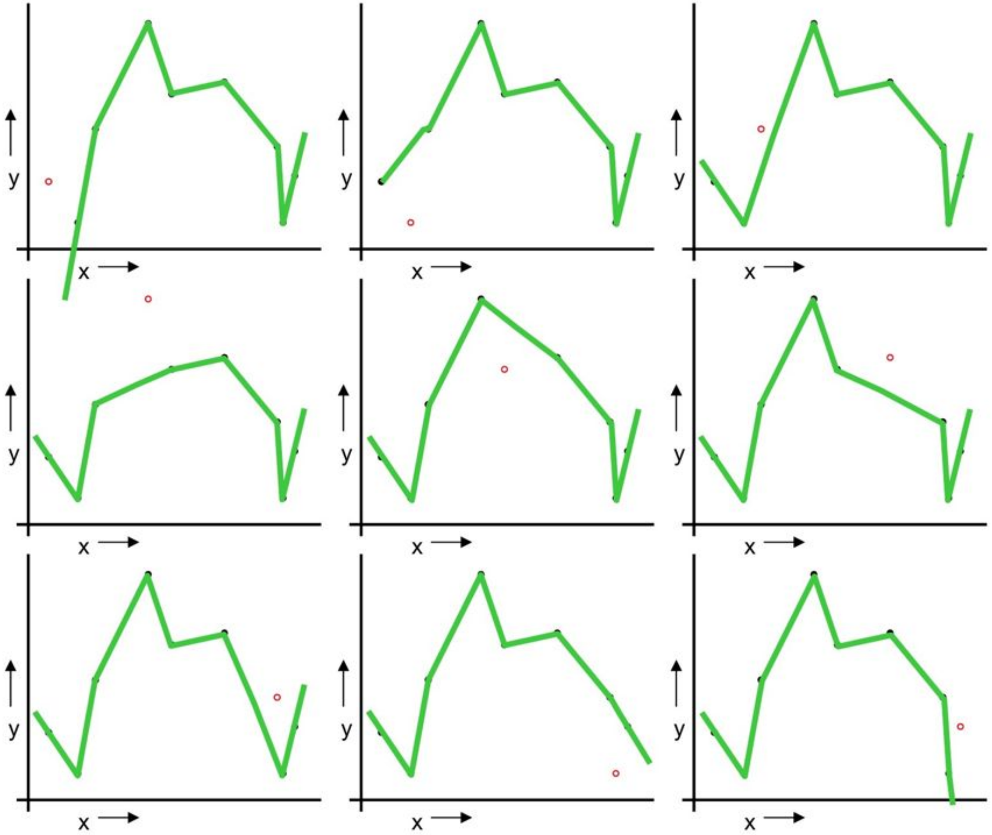
For k=1 to R

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

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)

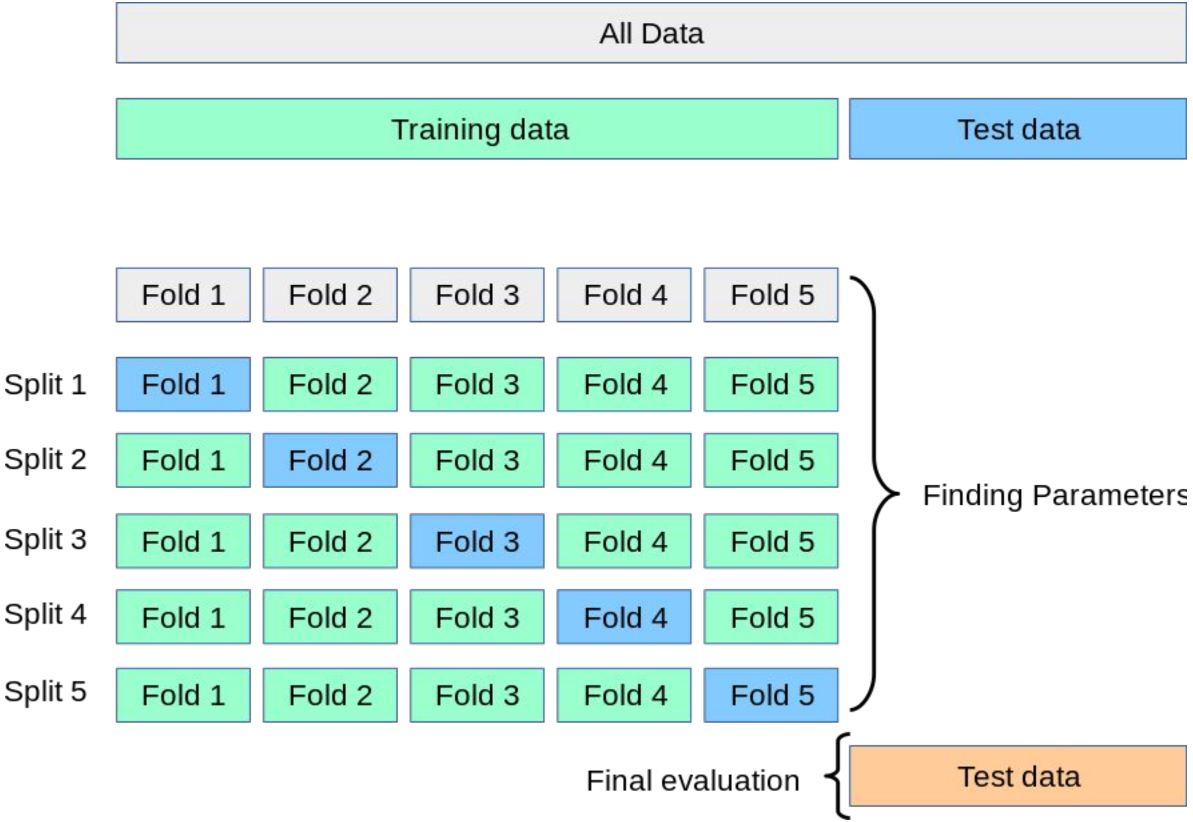
When you've done all points, report the mean error.

MSE_{LOOCV} $=3.33$

Method Comparison

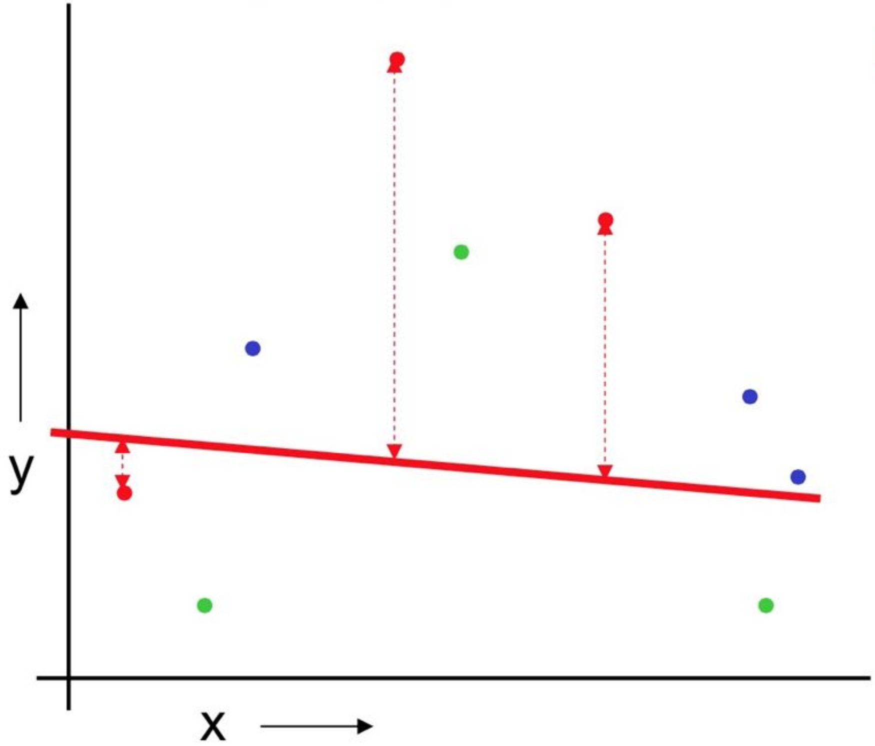
	Cons	Pros
Data partitioning	Variance: unreliable estimate of future performance	Cheap
LOOCV	Computationally expensive; has weird behavior	Uses all your data

k-fold cross validation



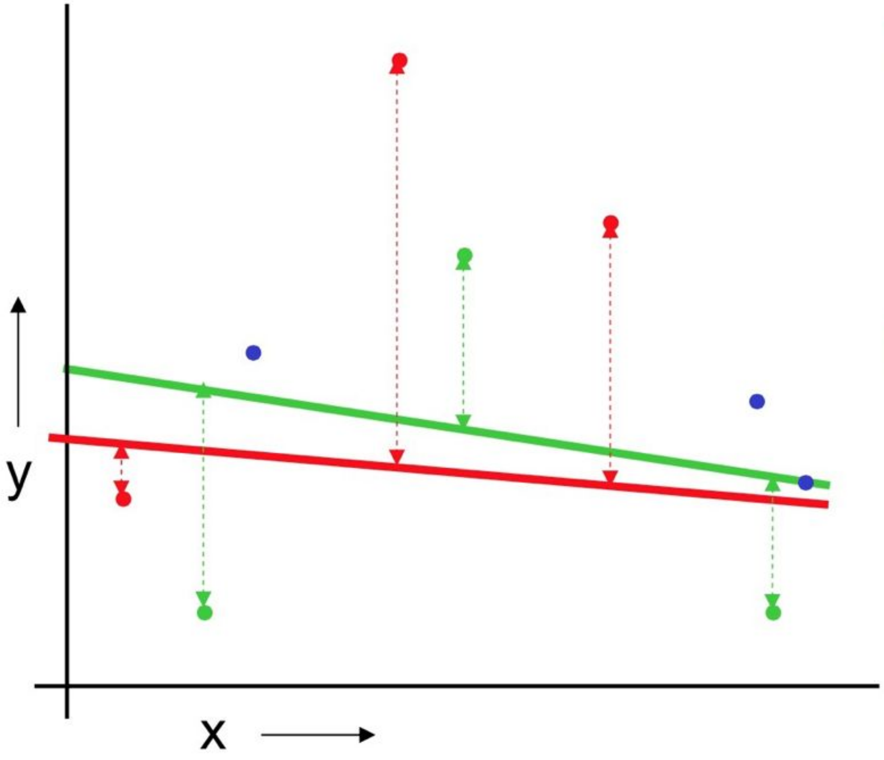
https://scikit-learn.org/stable/modules/cross_validation.html

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.

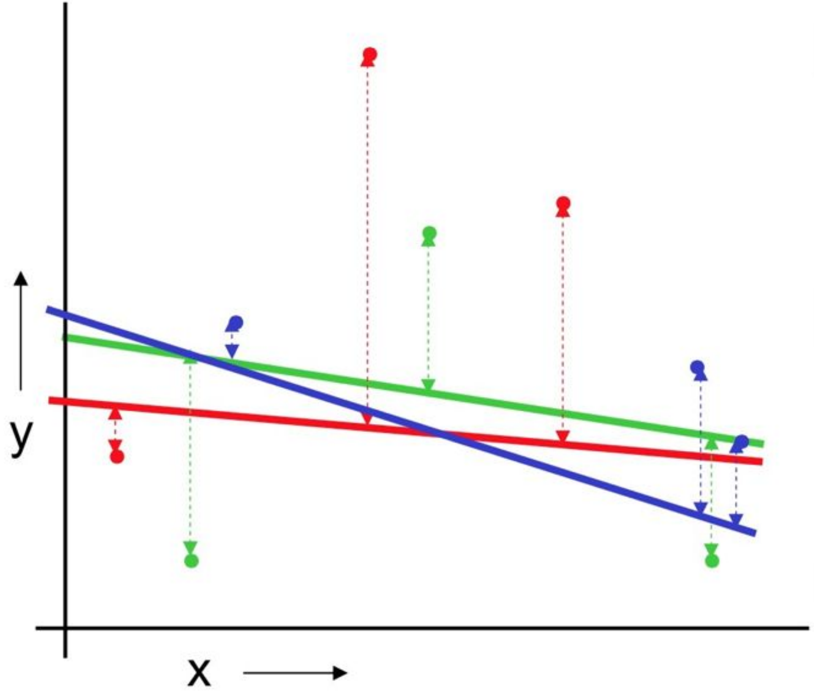
k-fold cross validation



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.

k-fold cross validation

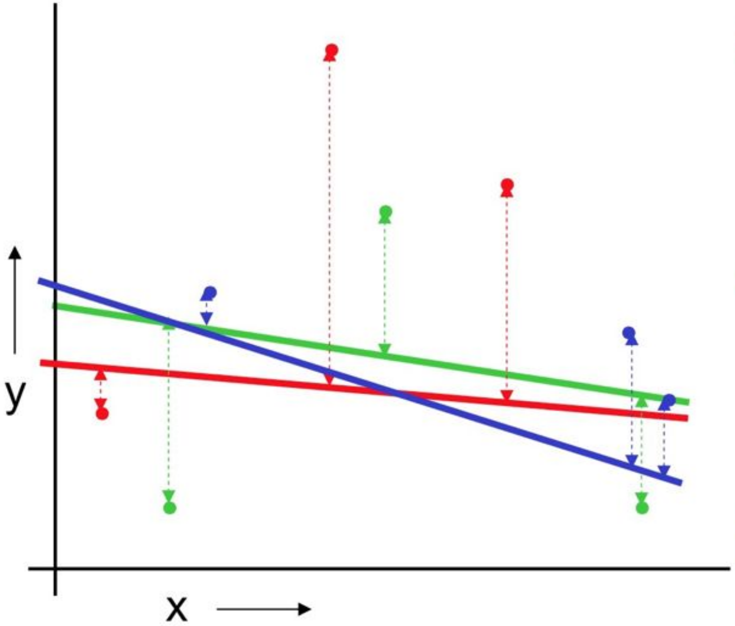


For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

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.

k-fold cross validation



Linear Regression
$MSE_{3FOLD}=2.05$

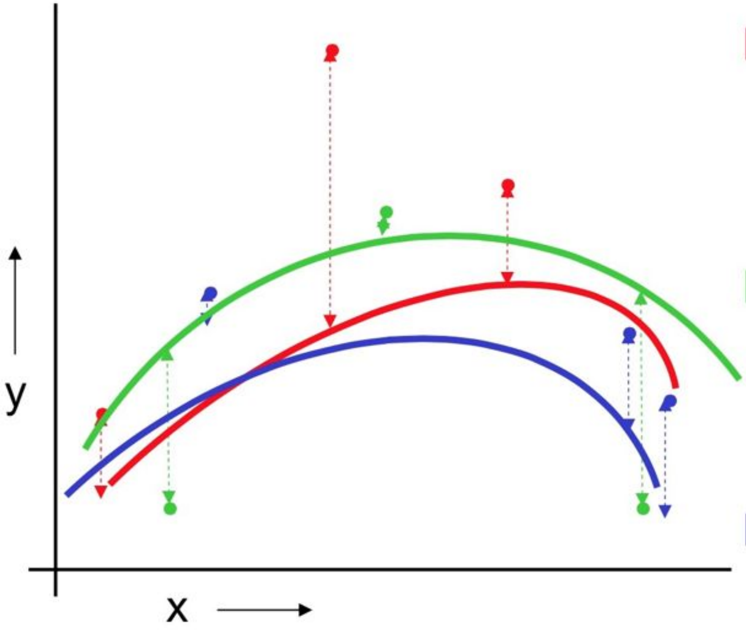
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

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

k-fold cross validation



Quadratic Regression
$MSE_{3FOLD}=1.11$

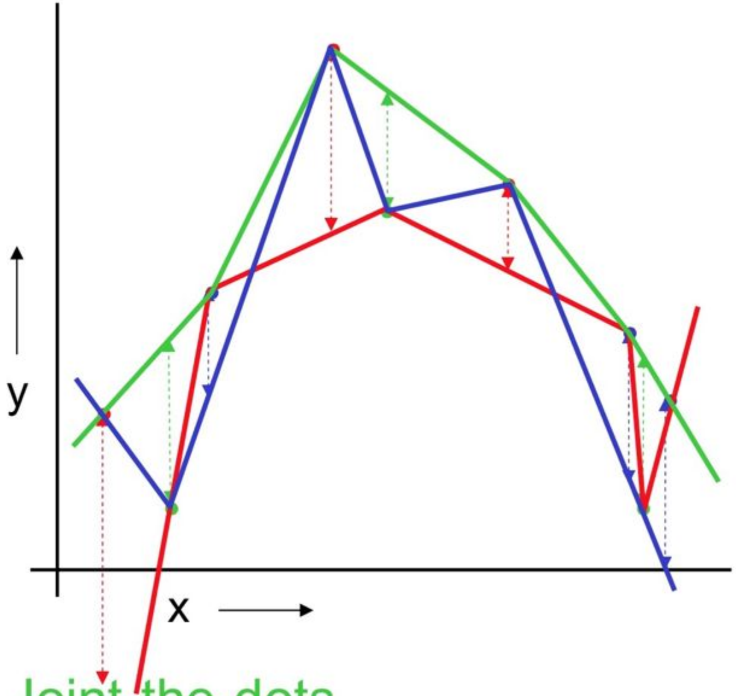
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

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

k-fold cross validation



Joint-the-dots

$MSE_{3FOLD}=2.93$

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.

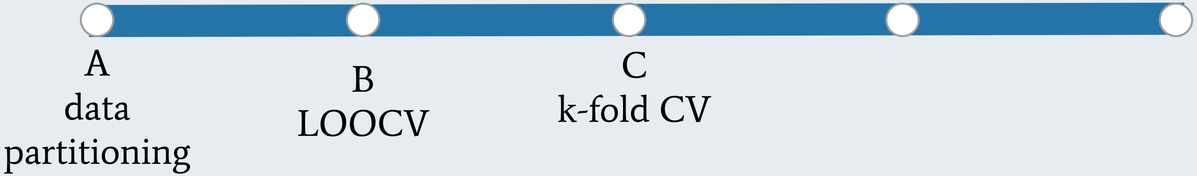
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

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

Validator

Which approach would *you* use to limit overfitting?



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