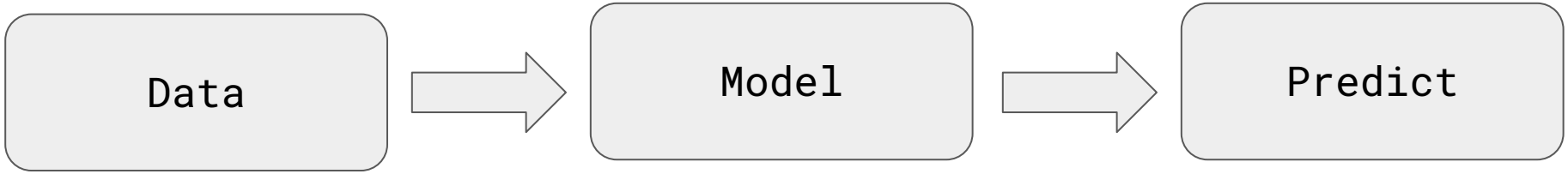
Fundamentals of Machine Learning

Outline

- Definition
- Terminologies
- Workflow

Acknowledgement : Most of slide credits go to CSE455, University of Washington

What is machine learning?



Using Data To Train the model For better prediction !

Machine Learning basic terminologies

- Learning
- Examples / Samples / Data
- Features
- Targets / Labels
- Model
- Training
- Testing / Prediction
- Inference
- Hyperparameters

What is learning in machine learning?

- Algorithms to approximate functions
  - Usually use lots of statistics
  - Usually minimize some form of *loss function*

What is learning in machine learning?

- Algorithms to approximate functions
  - Usually use lots of statistics
  - Usually minimize some form of *loss function*
- Supervised learning
  - Given inputs to a function, try to predict the output
  - Have lots of labelled examples

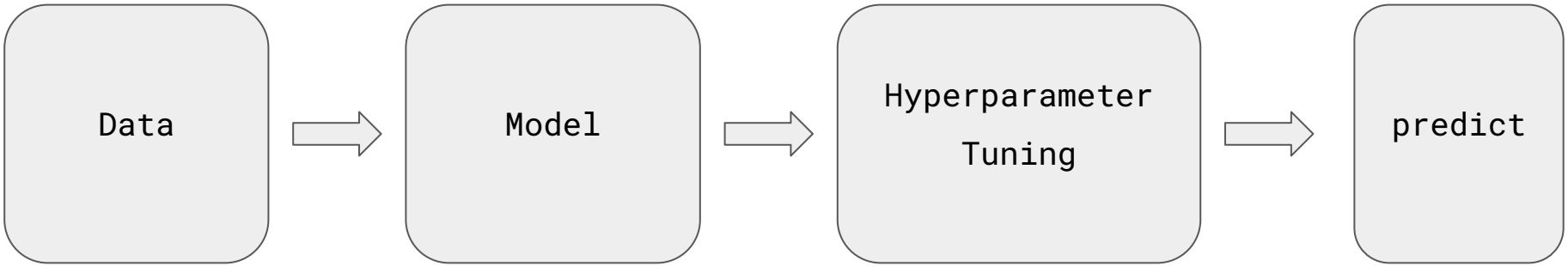
What is learning in machine learning?

- Algorithms to approximate functions
  - Usually use lots of statistics
  - Usually minimize some form of *loss function*
- Supervised learning
  - Given inputs to a function, try to predict the output
  - Have lots of labelled examples
- Semi-supervised learning
  - Same but number of labelled examples < number of examples

What is learning in machine learning?

- Algorithms to approximate functions
  - Usually use lots of statistics
  - Usually minimize some form of *loss function*
- Supervised learning
  - Given inputs to a function, try to predict the output
  - Have lots of labelled examples
- Semi-supervised learning
  - Same but number of labelled examples < number of examples
- Unsupervised learning
  - Want to model unlabelled data
  - Find similarities and differences between subgroups of data
  - Learn functions to generate new data

Machine Learning workflow



- Collection
- Preparation

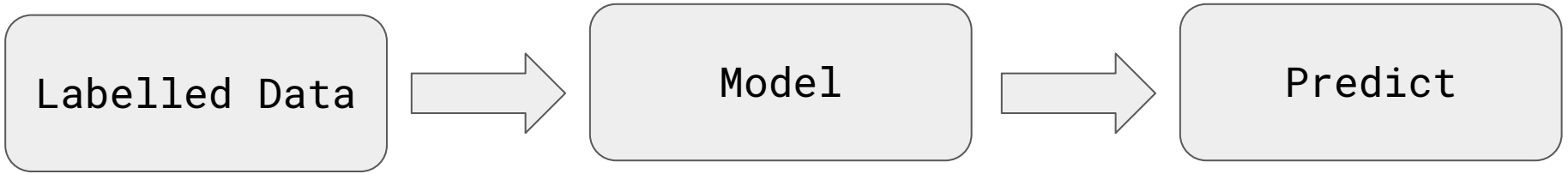
- Model Choosing
- Model Training

- Tune model hyperparameters for better performance

- Approximation of model performance in real world problem

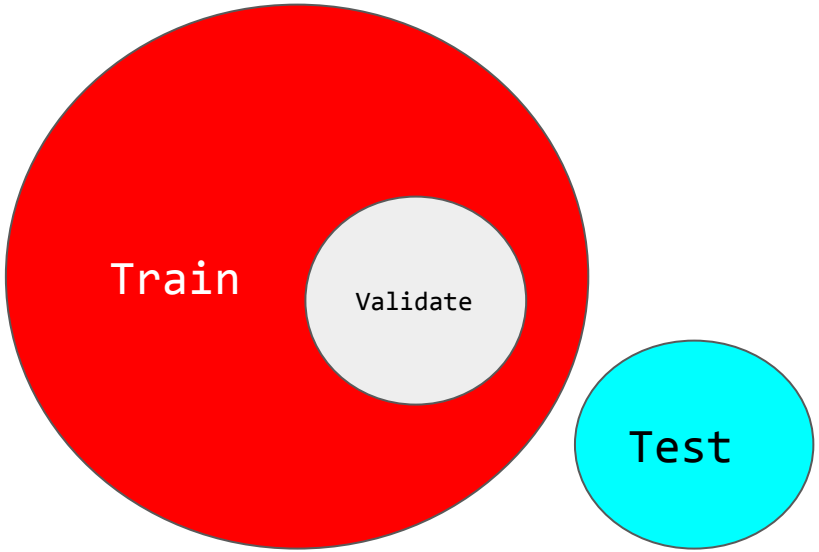
Supervised learning

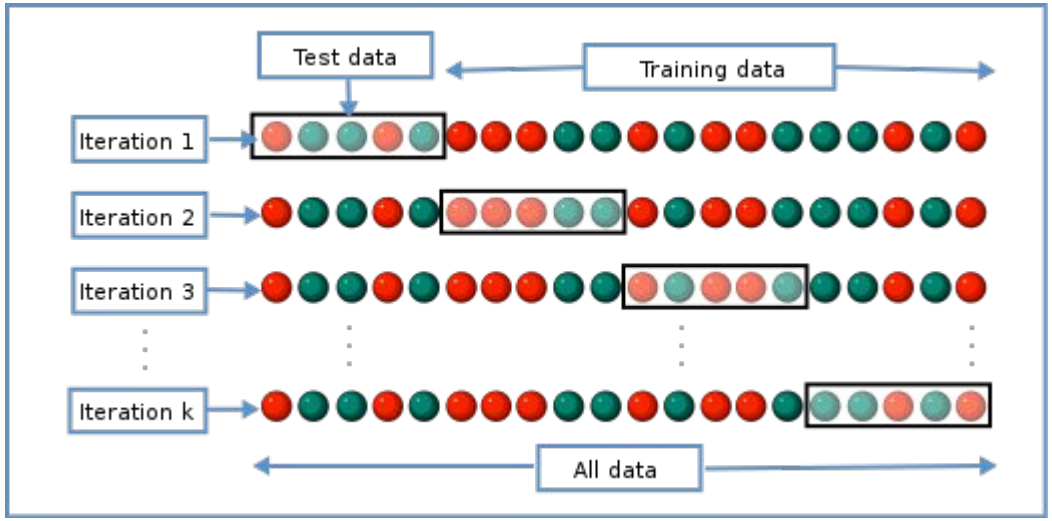
- Given inputs to a function, try to predict the output
- Have lots of labelled examples



Data we feed

- Train Test Split
- Cross Validation



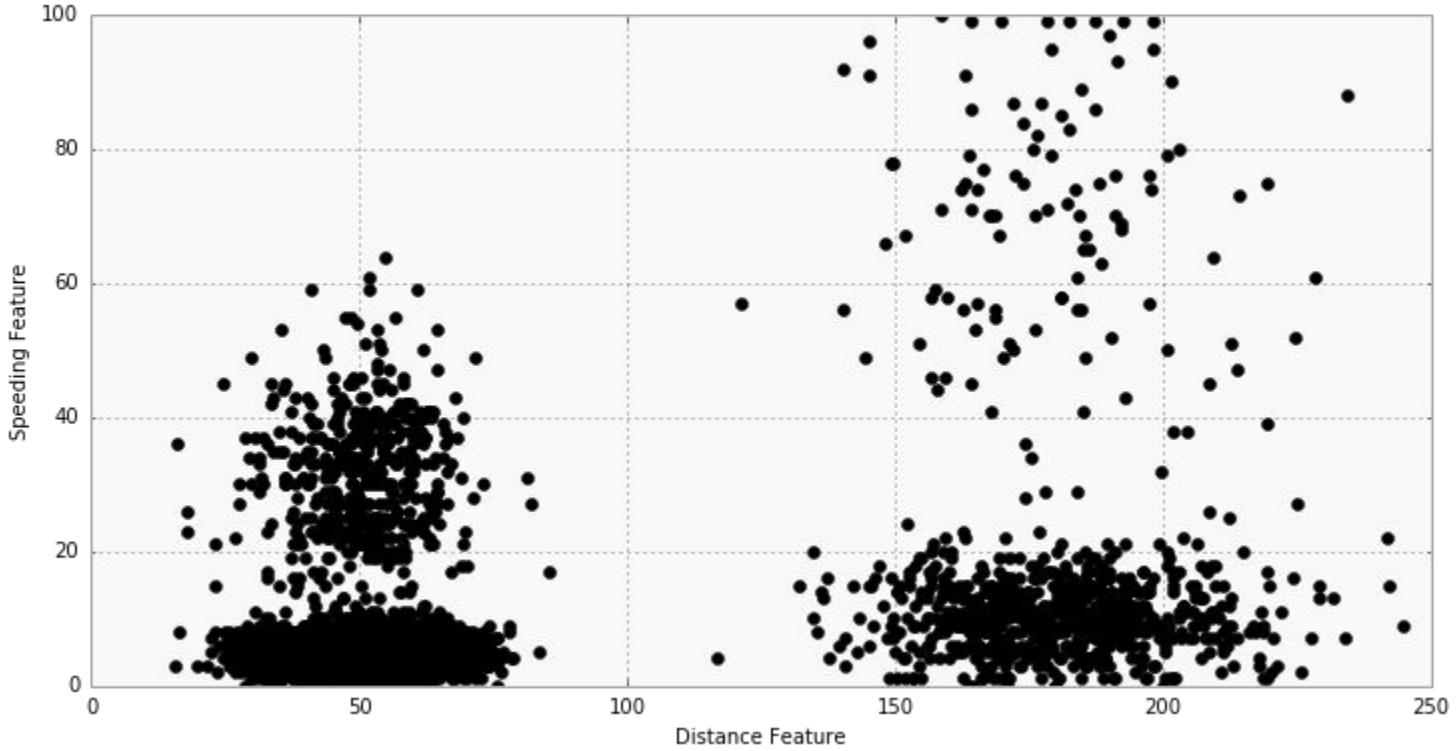


Unsupervised learning

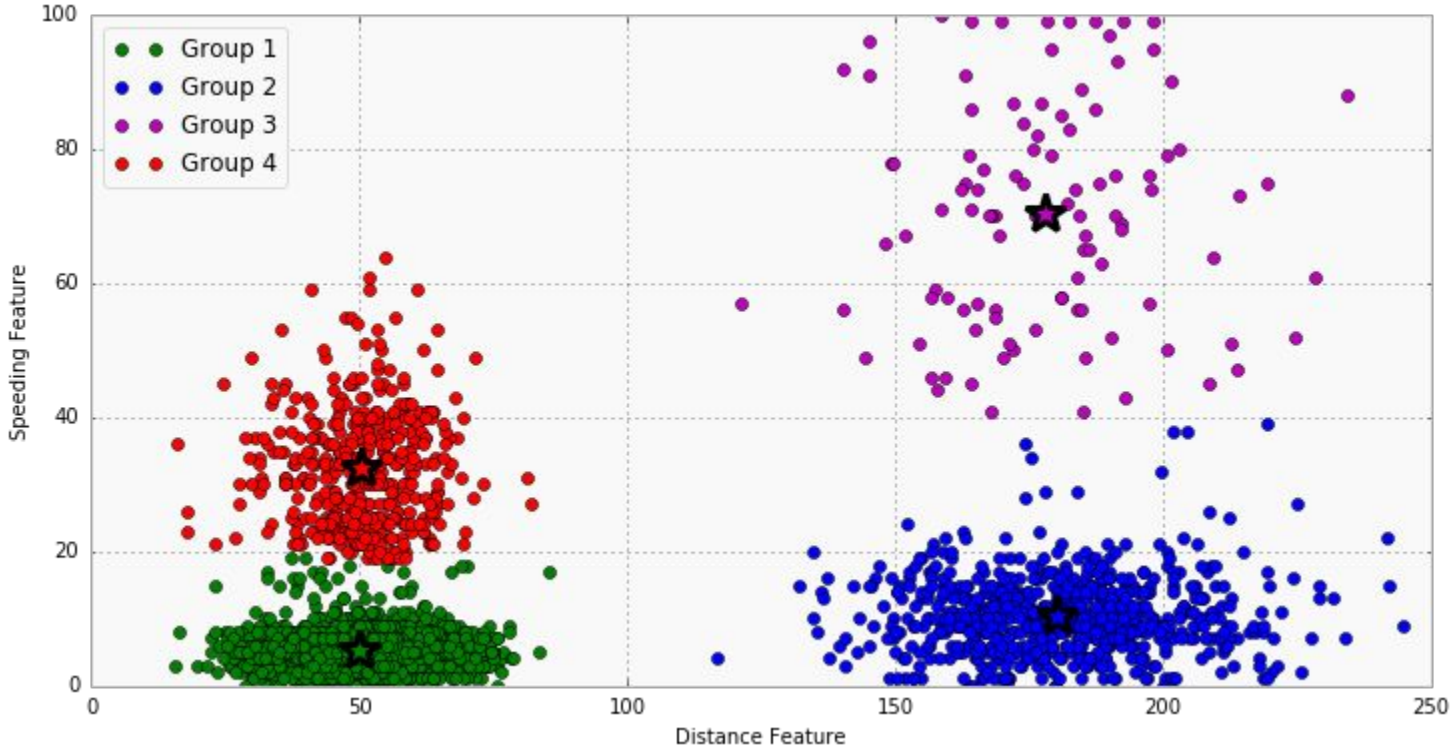
No labels, just looking for patterns in data

E.g. clustering, in data with multiple clusters, what are they, how big, etc.

Clustering: finding groups in data



Clustering: finding groups in data



K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers

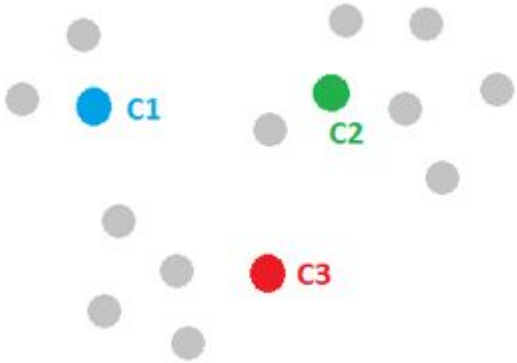
https://healthcare.ai/step-step-k-means-clustering/

K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers

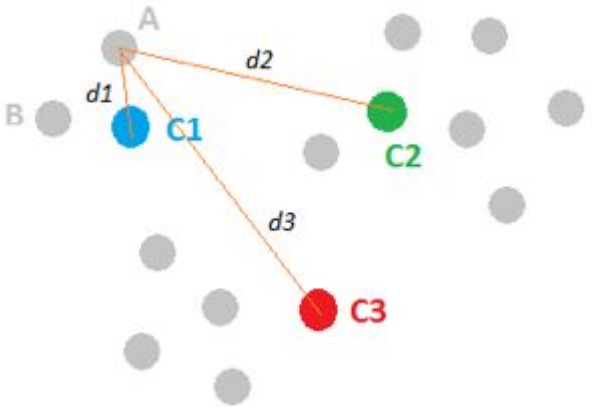


K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers
Calculate distance points <-> centers

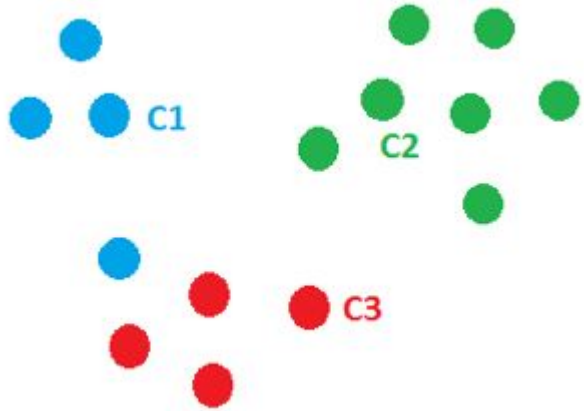


K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers
Calculate distance points <-> centers
Assign each point to closest cluster center

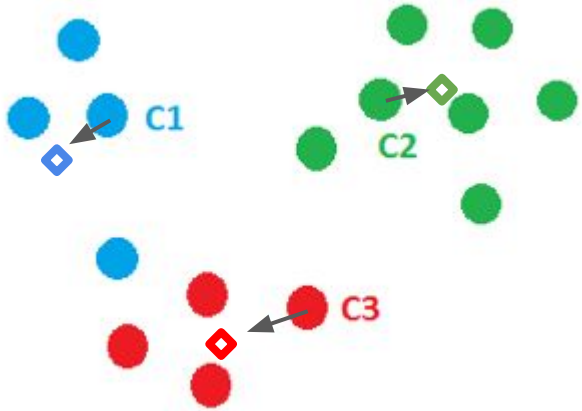


K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers
Calculate distance points <-> centers
Assign each point to closest cluster center
Update cluster centers: avg of points

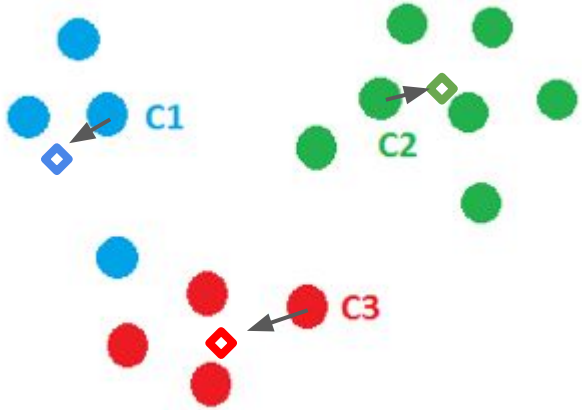


K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers
Calculate distance points <-> centers
Assign each point to closest cluster center
Update cluster centers: avg of points
Repeat!

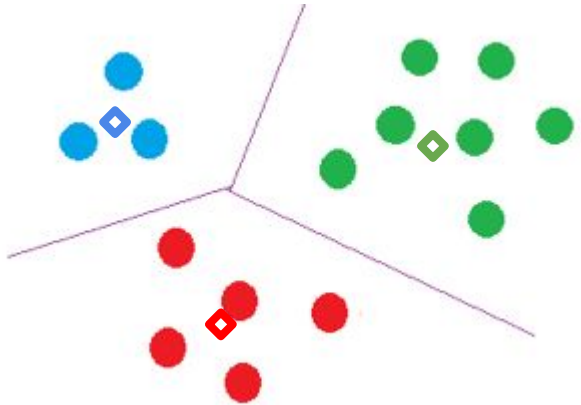


K-means clustering

Assume points are close to other points in group, far from points out of group

Algorithm:

Randomly initialize cluster centers
Loop until converged:
    Calculate distance points <-> centers
    Assign each point to closest center
    Update cluster centers: avg of points



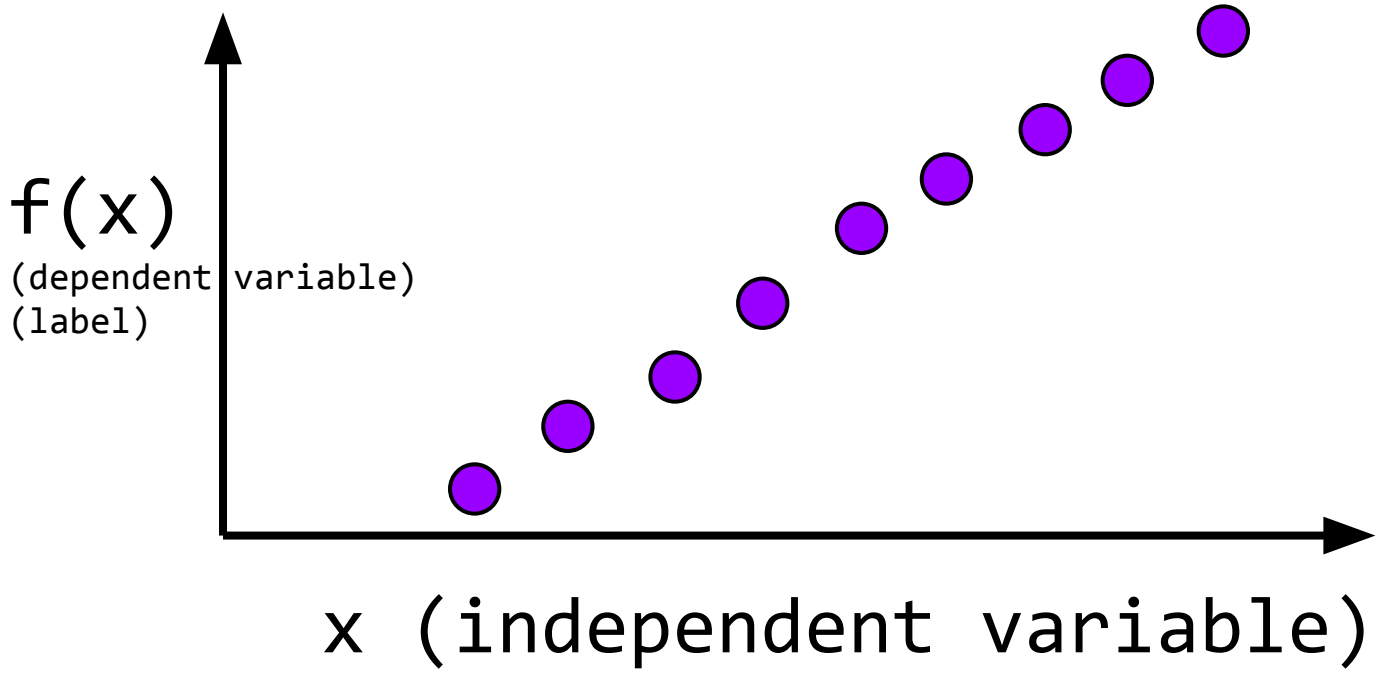
Clustering on images

Group together pixels by color, automatic segmentation: k-means, k = 2

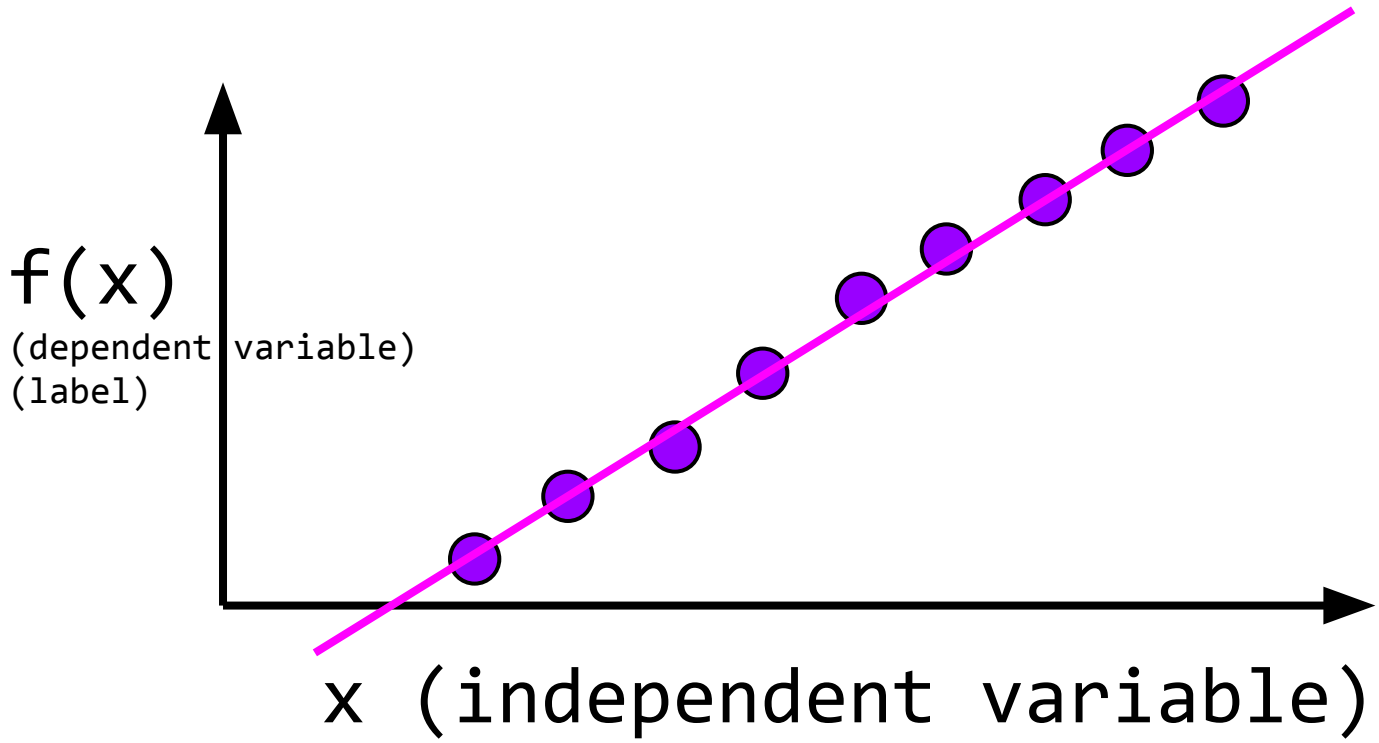
Clustering on images

Group together pixels by color, automatic segmentation: k-means, k = 4

Supervised Learning: Want to estimate f

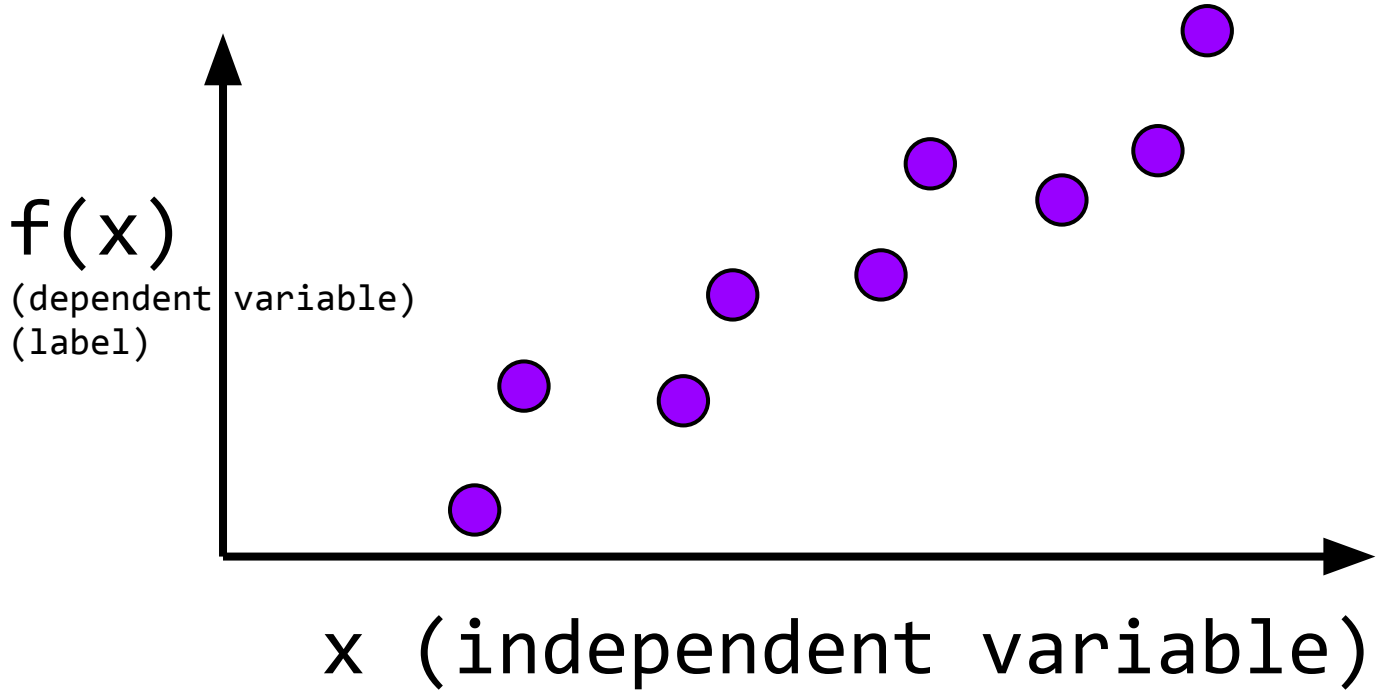


Here’s one possible f*



Data often has *noise*

Why?



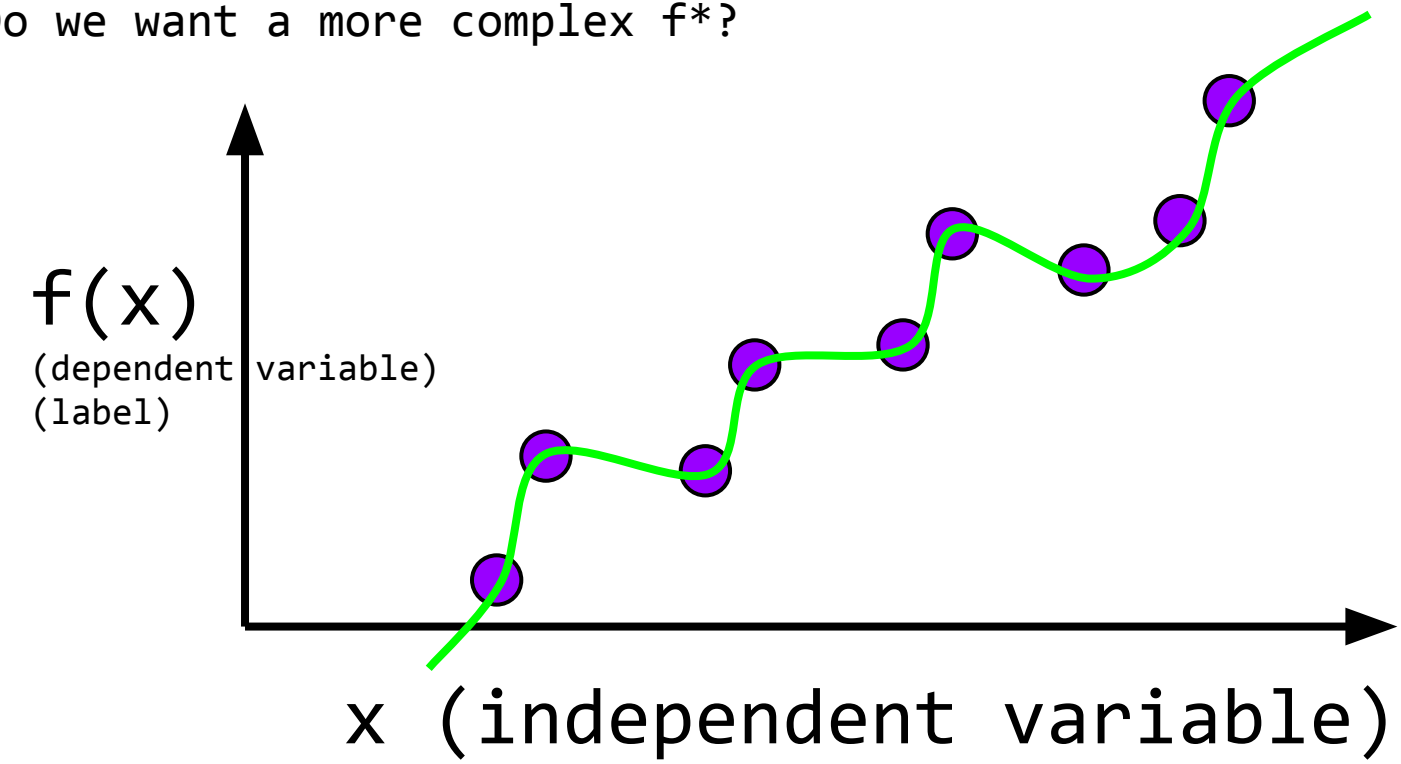
Data often has *noise*

Why?

- Randomness
  - Static in phone lines, random distribution of photons hitting sensor, sensors aren’t precise
- Mislabelled data
  - Common when humans label lots of data
- Variables outside of model
  - Variations might look like noise but are explained by a hidden, unknown variable

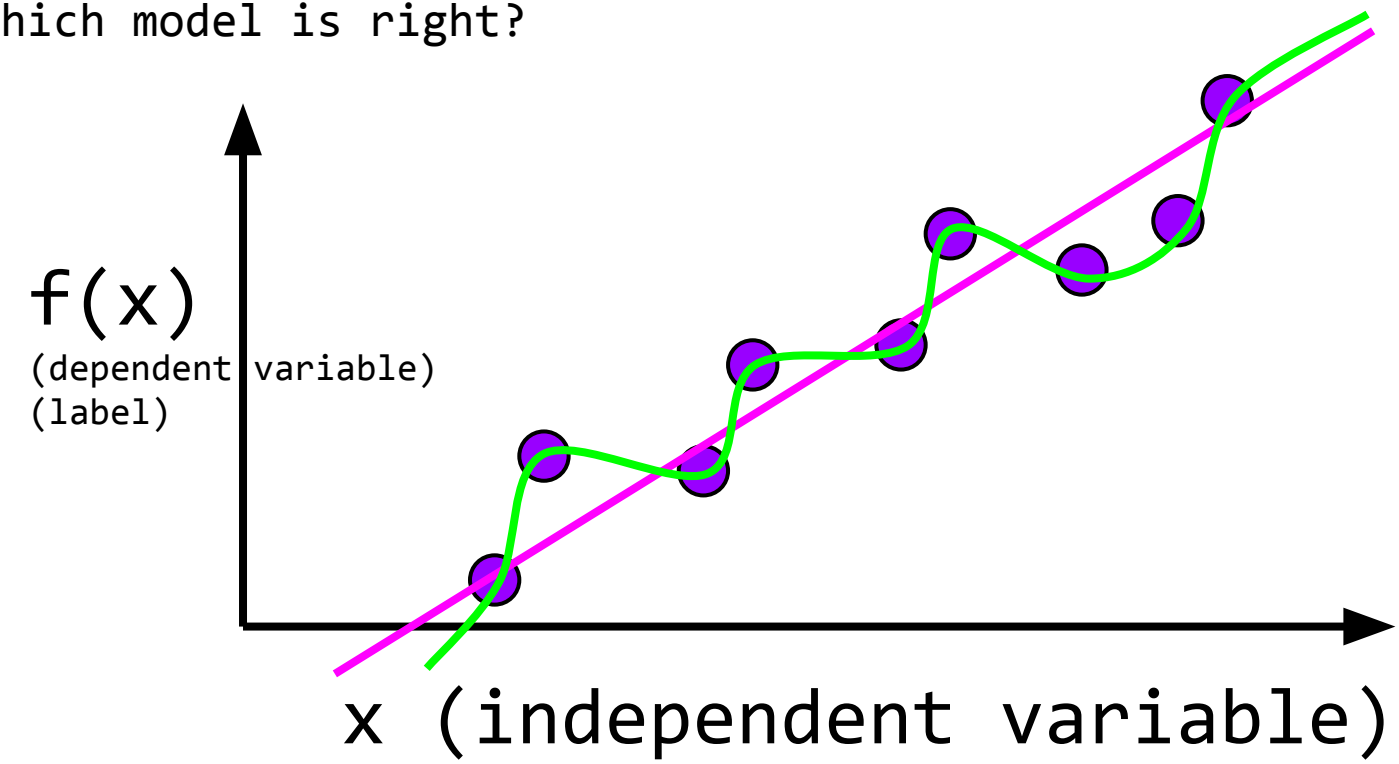
Data often has *noise*

Do we want a more complex f*?



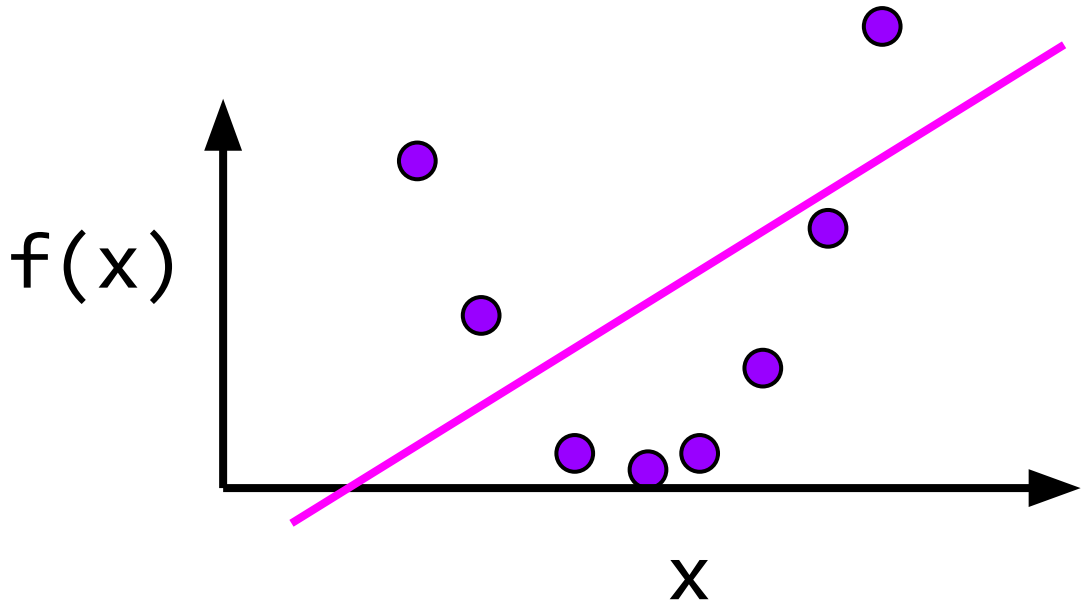
Sometimes the data is more complex

Which model is right?



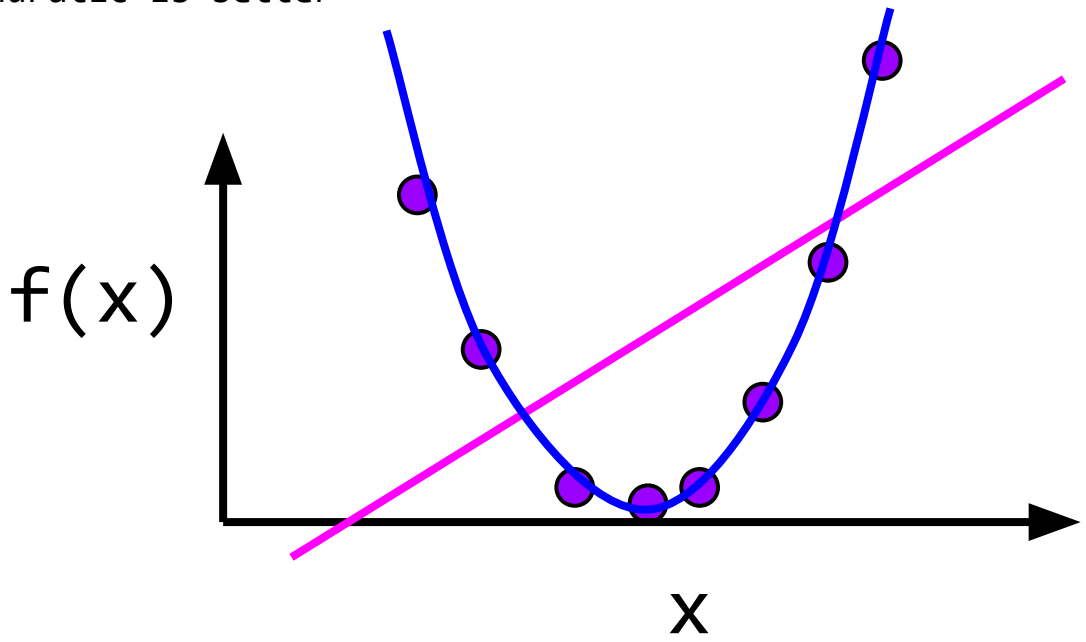
Bias / Variance tradeoff

- Bias
  - Error from assumptions model makes about data
  - Linear model assumes data is linear, bad for data that isn’t



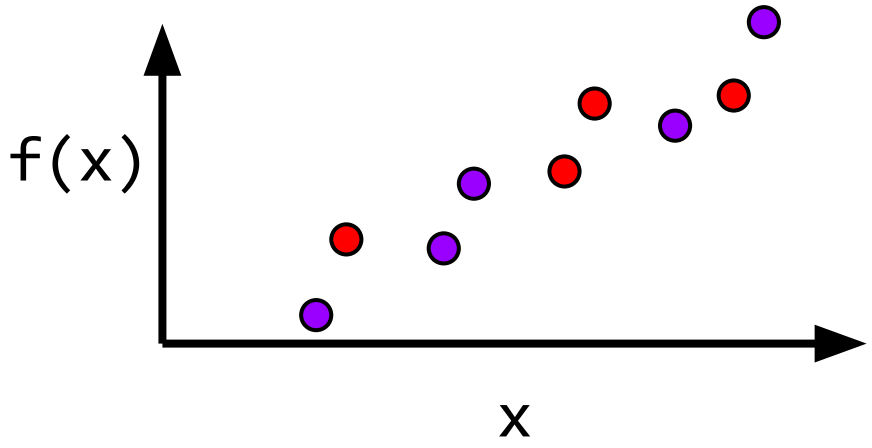
Bias / Variance tradeoff

- Bias
  - Error from assumptions model makes about data
  - Linear model assumes data is linear, bad for data that isn’t
  - Quadratic is better



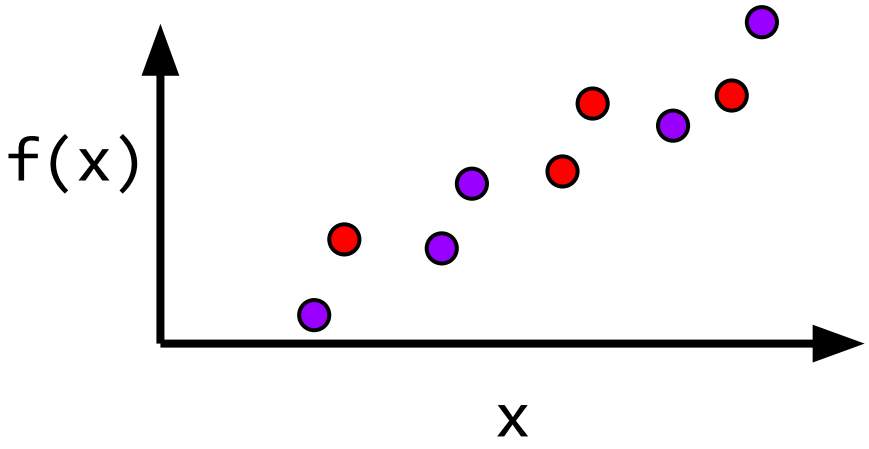
Bias / Variance tradeoff

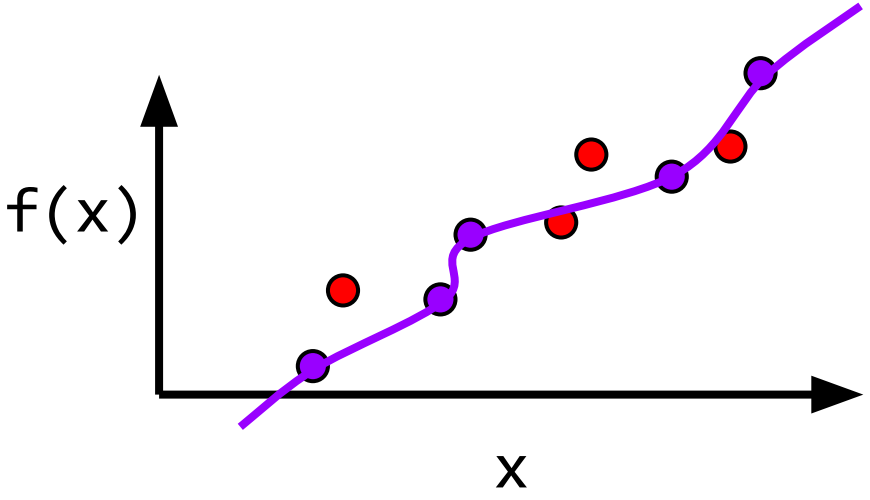
- Variance
  - Algorithm’s sensitivity to noise
  - More complex algorithms are more sensitive!



Bias / Variance tradeoff

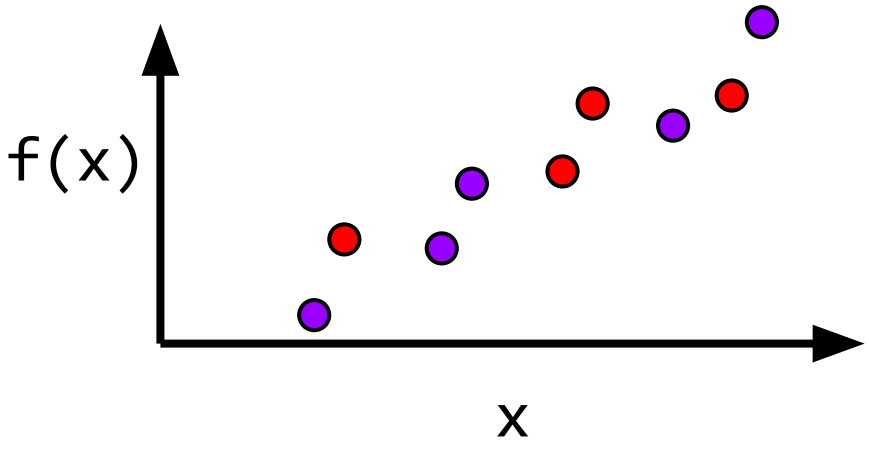
- Variance
  - Algorithm’s sensitivity to noise
  - More complex algorithms are more sensitive!

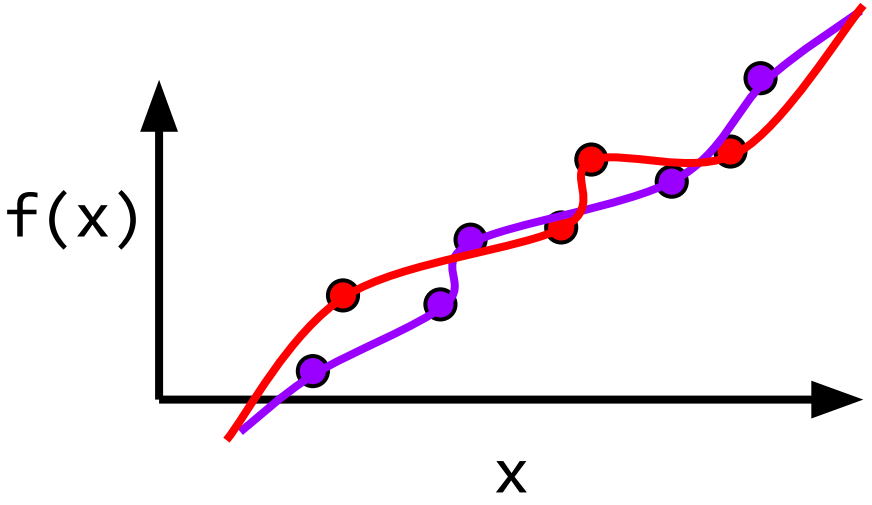




Bias / Variance tradeoff

- Variance
  - Algorithm's sensitivity to noise
  - More complex algorithms are more sensitive!
  - High variance hurts generalization, *overfitting*





Bias / Variance tradeoff

- Noise
  - Random variations in data
- Bias
  - Error from assumptions model makes about data
  - Less complex algorithms -> more assumptions about data
- Variance
  - Algorithm’s sensitivity to noise
  - More complex algorithms are more sensitive!
  - High variance hurts generalization

Q & A

External Reading :

- Introduction to Probability

https://thuraaung-1601.medium.com/introduction-to-probability-7b884750aaa1

- Introduction to Linear Regression

https://thuraaung-1601.medium.com/introduction-to-linear-regression-with-normal-equation-98e6c1f839f8

- Machine Learning and Artificial Intelligence Intro

  (Engr.Thet Naing Tun and Dr.Zaw Min Khaing)

https://youtu.be/oqr6Kh6ywQk

Thank You !