AIBridge

Lecture 2

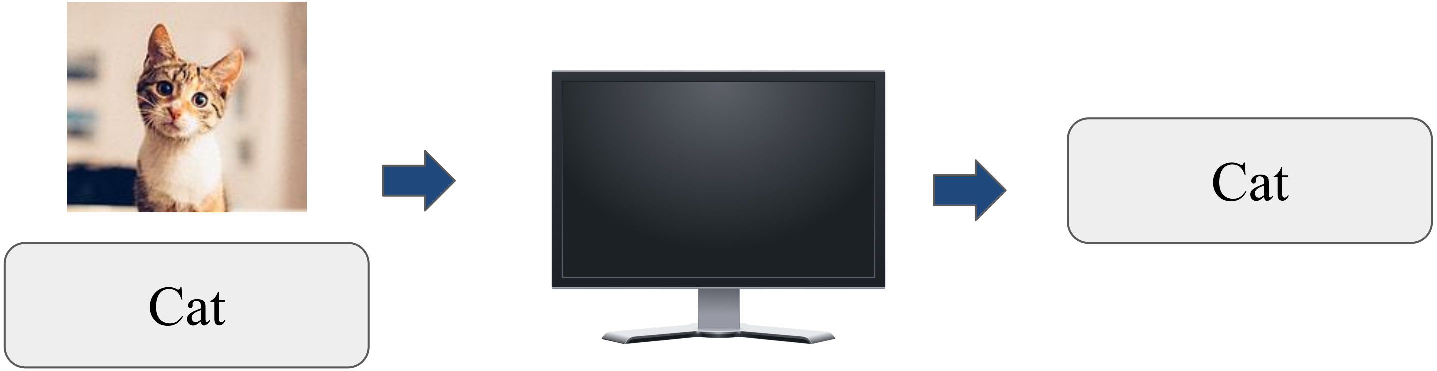
SUPERVISED LEARNING

Supervised Learning

- Classification:
  - Predicting a label/class/category
  - Ex: spam or not, cancer or not, cat or dog, red wine vs. white wine
- Regression:
  - Predicting a (continuous) quantity
  - Ex: Survival rate, wine quality, yield prediction

Supervised learning

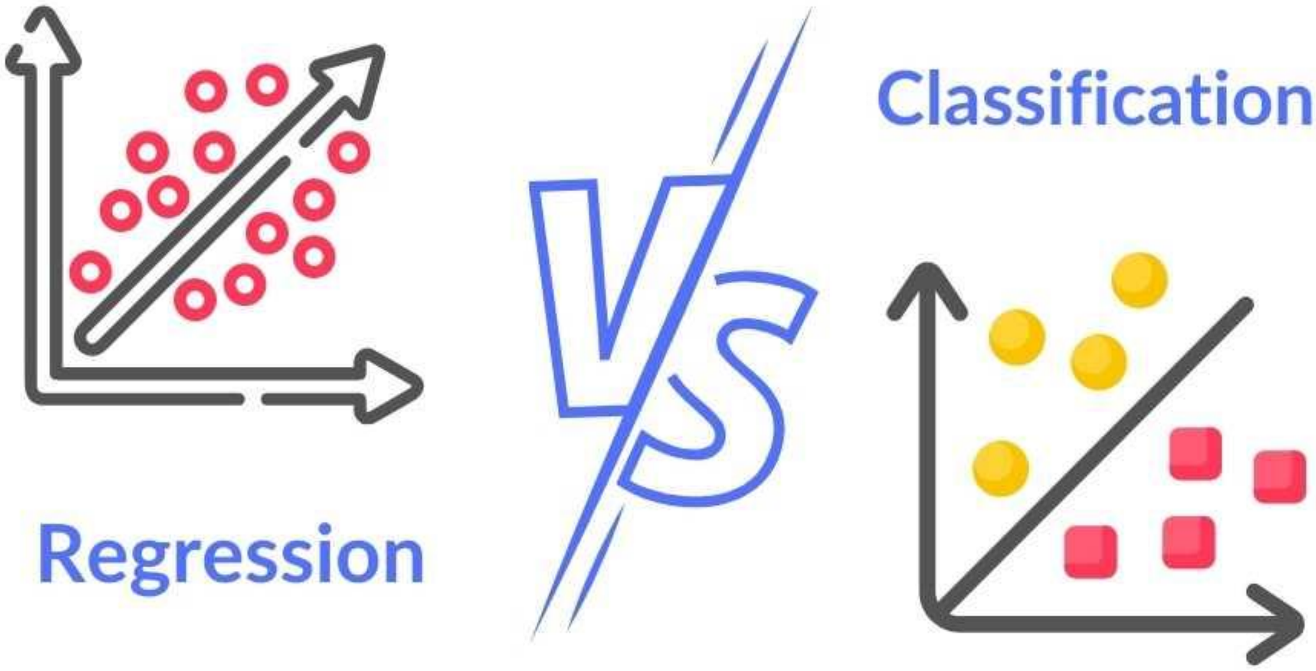
- training the model with labeled data
- Most widely used ML techniques in real world applications.



Cat

Cat

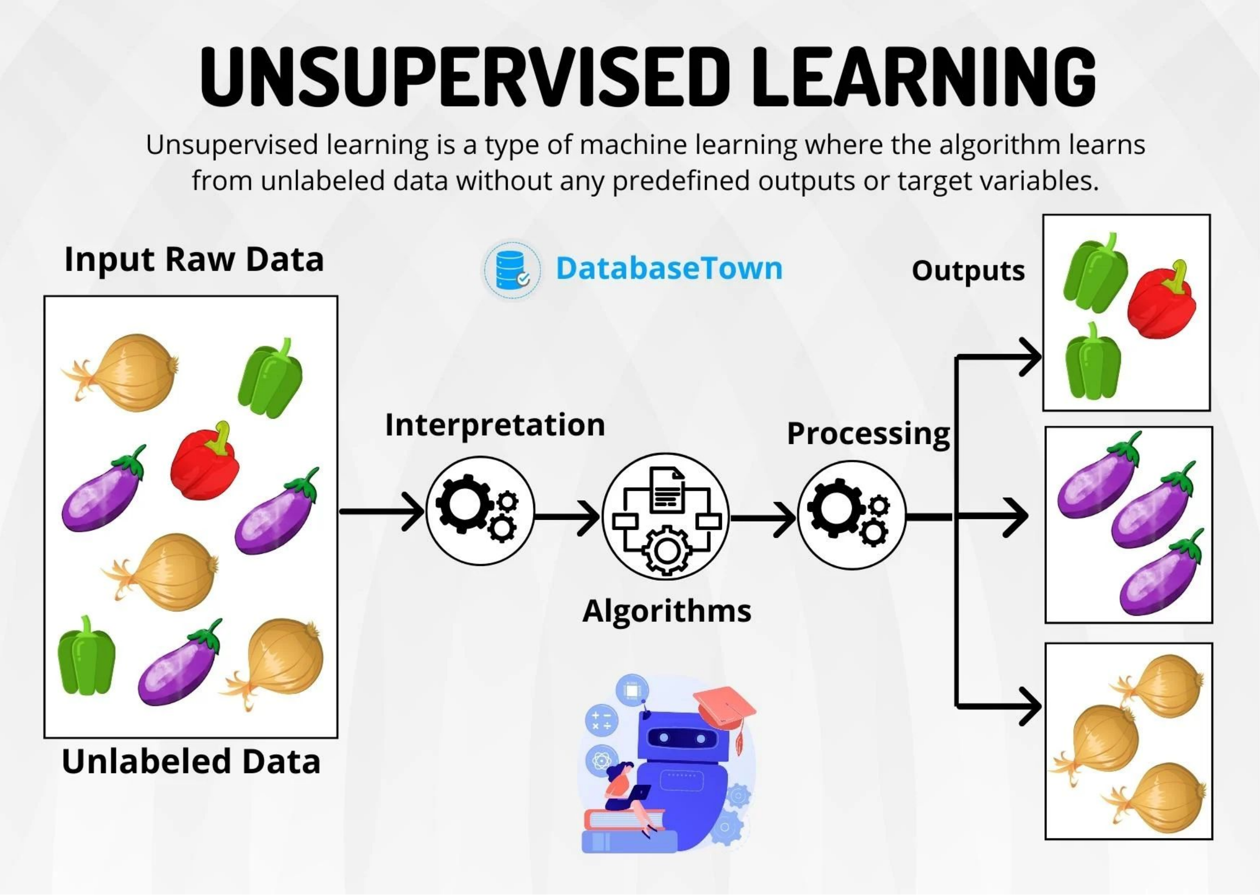
Supervised Learning



https://spotintelligence.com/2023/05/02/regression-vs-classification/

UNSUPERVISED LEARNING

Unsupervised Learning

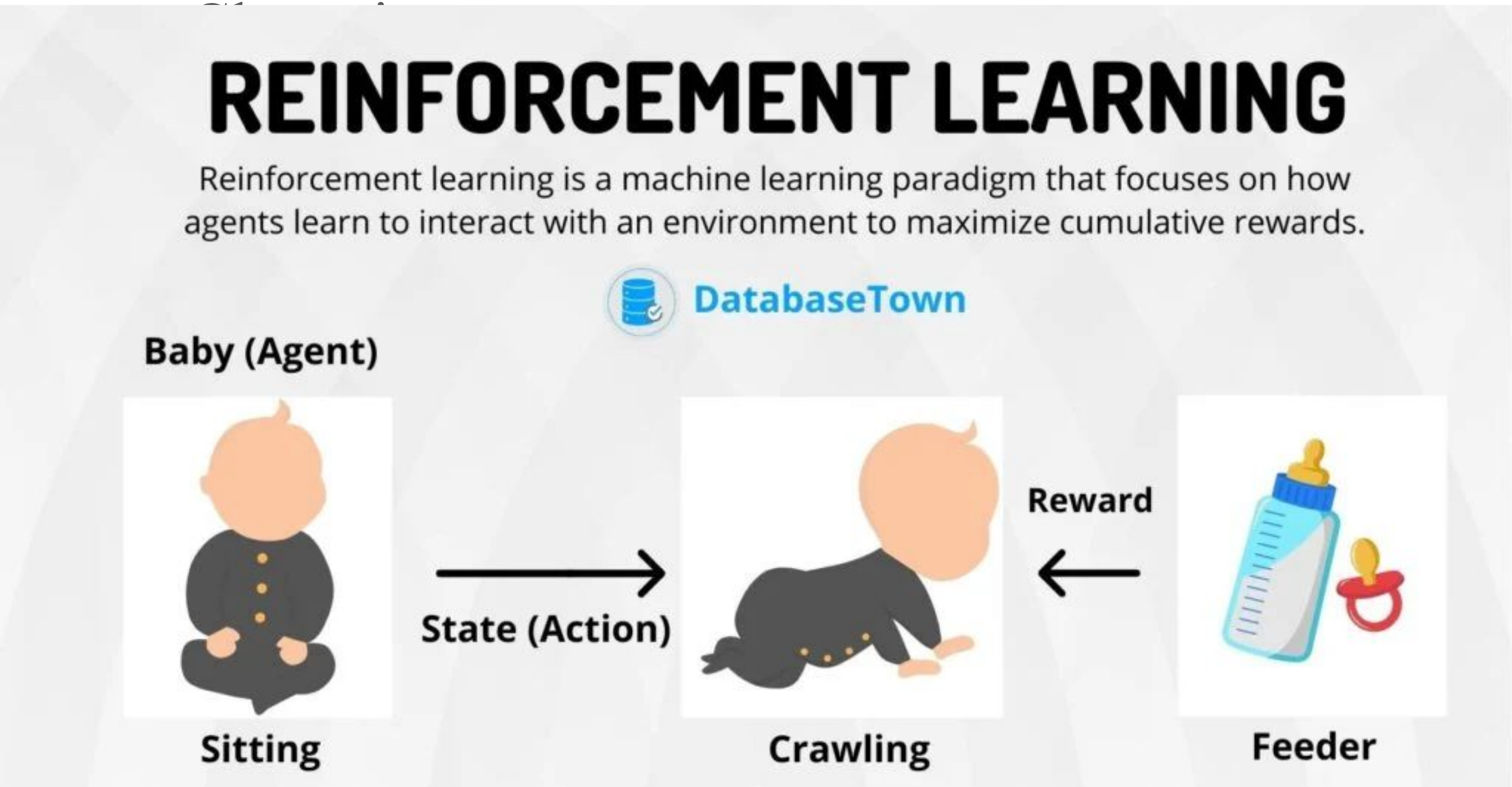


UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.

REINFORCEMENT LEARNING

Reinforcement Learning

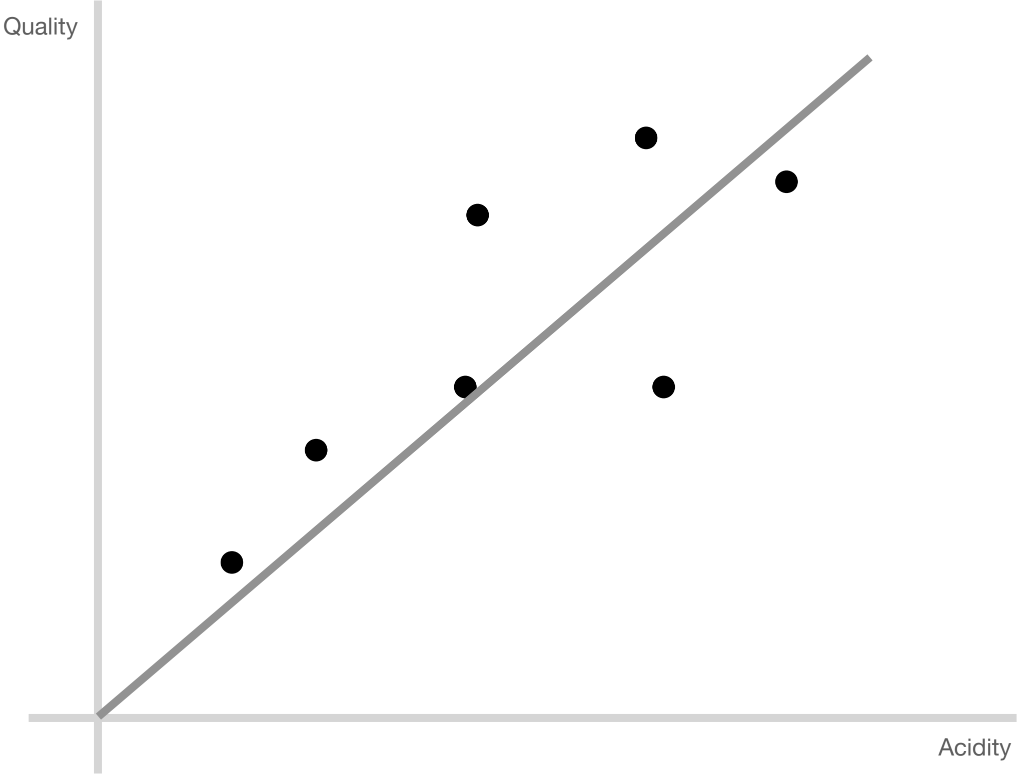


REINFORCEMENT LEARNING

Reinforcement learning is a machine learning paradigm that focuses on how agents learn to interact with an environment to maximize cumulative rewards.

Linear Regression

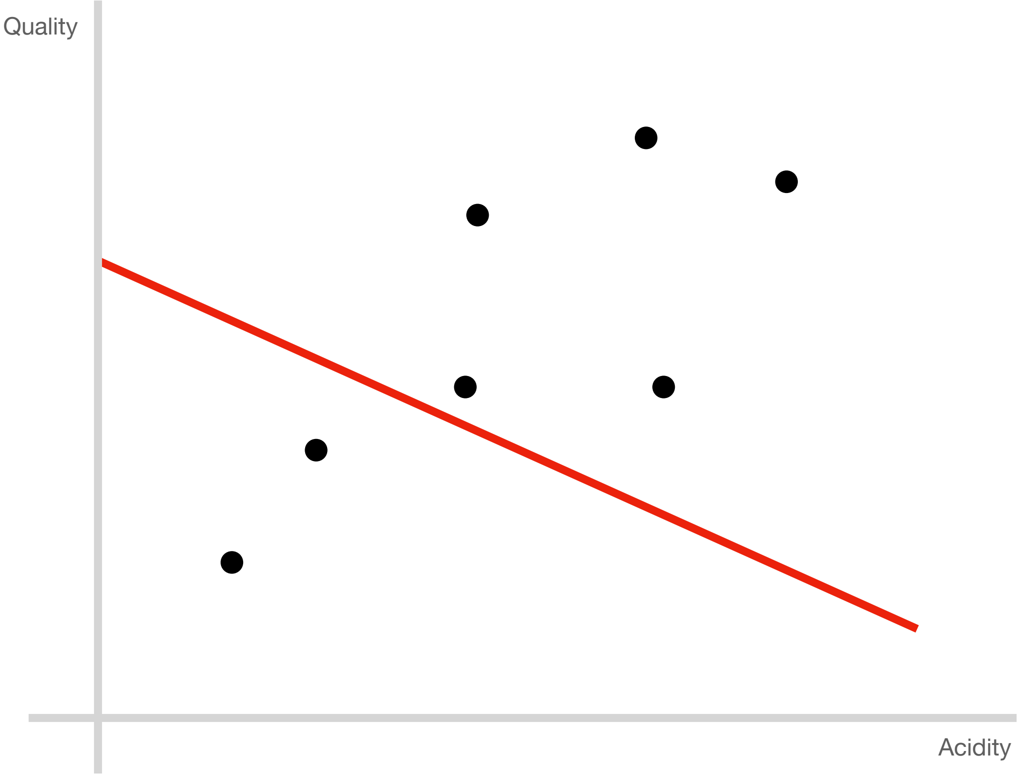
STATISTICS!



this model could be a function

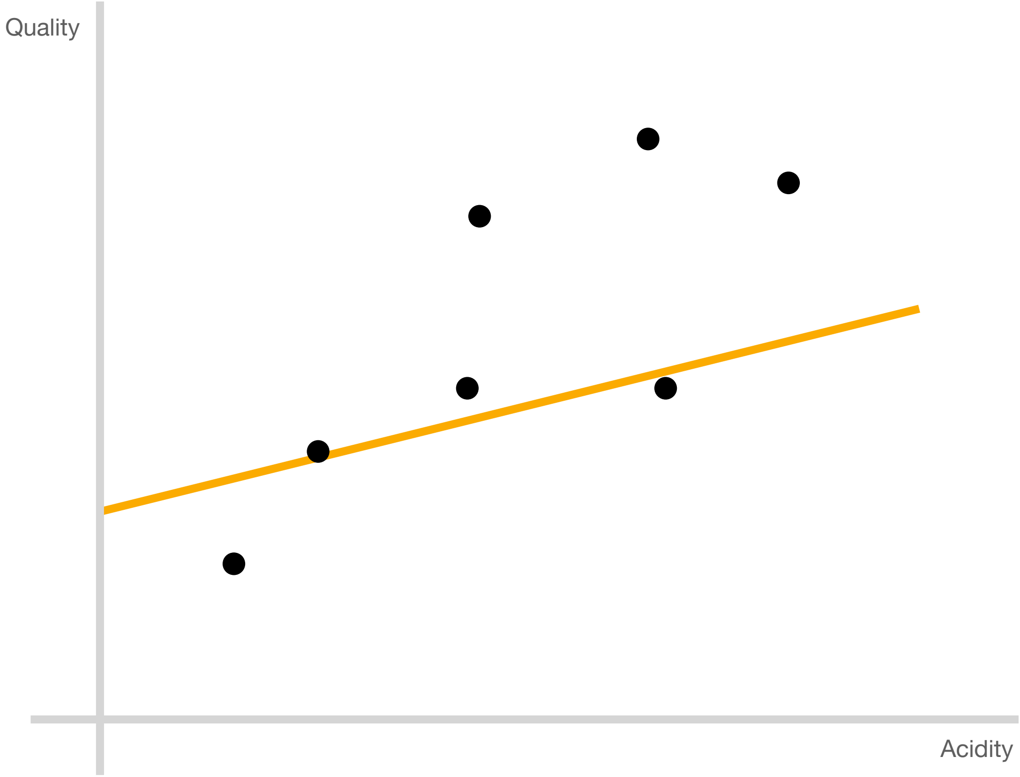
Linear Regression

STATISTICS!



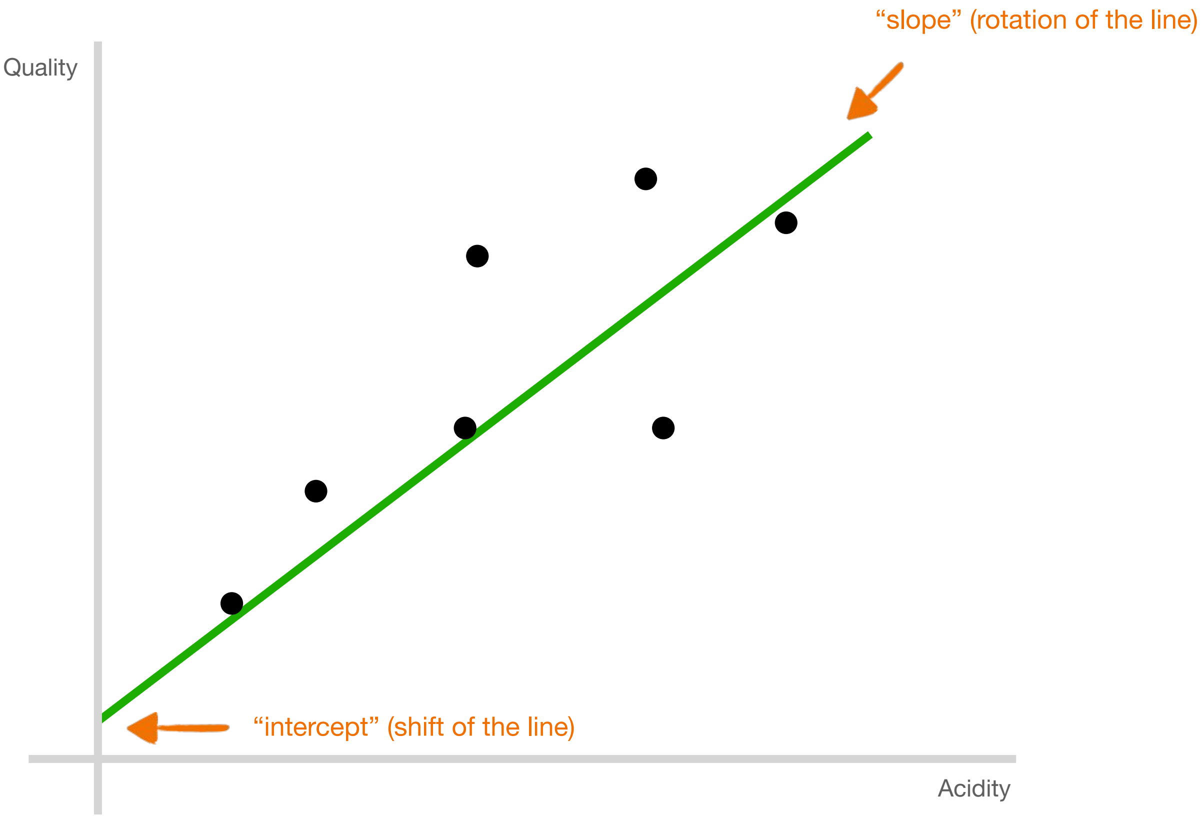
Linear Regression

STATISTICS!



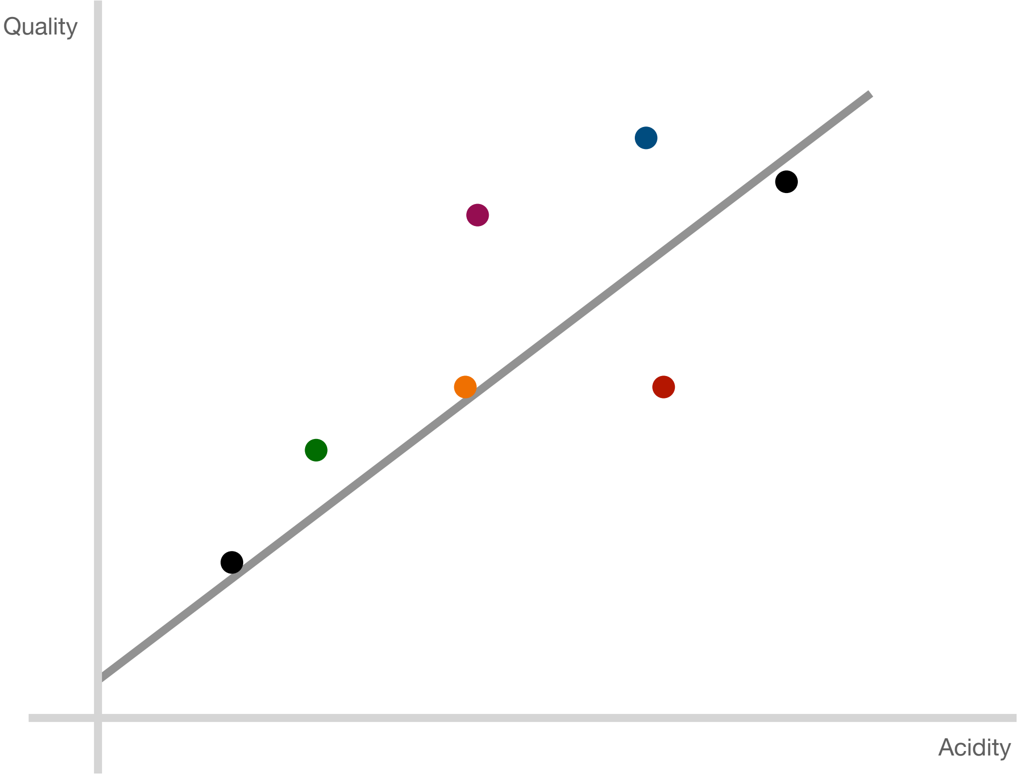
Linear Regression

STATISTICS!



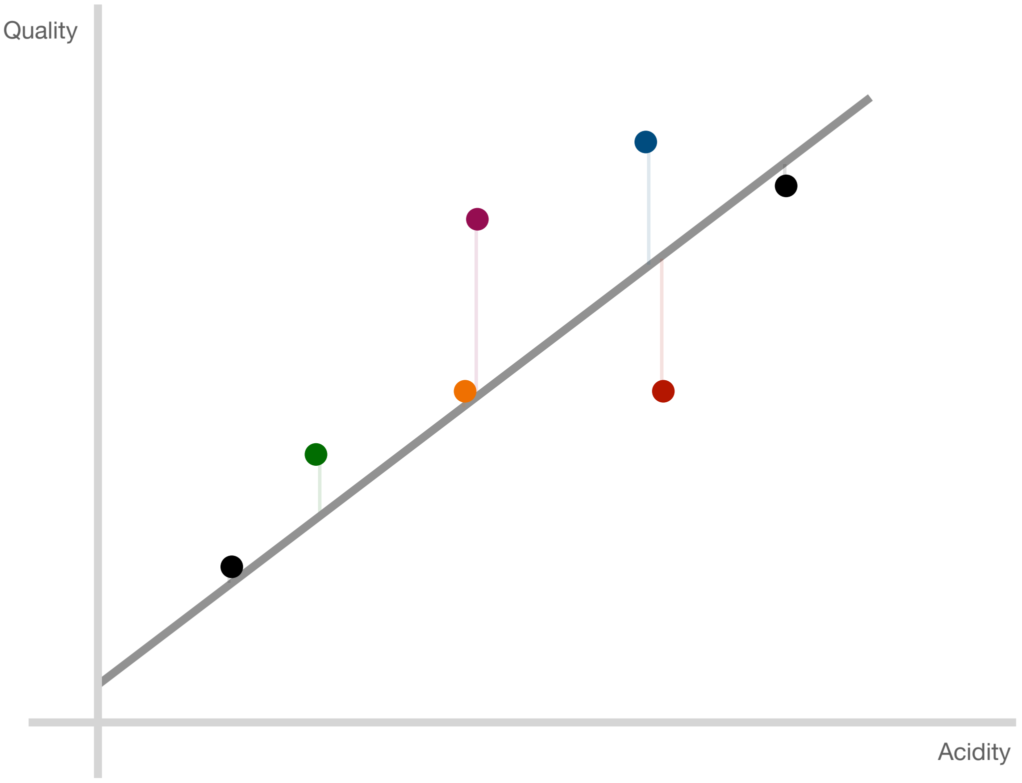
Linear Regression

STATISTICS!



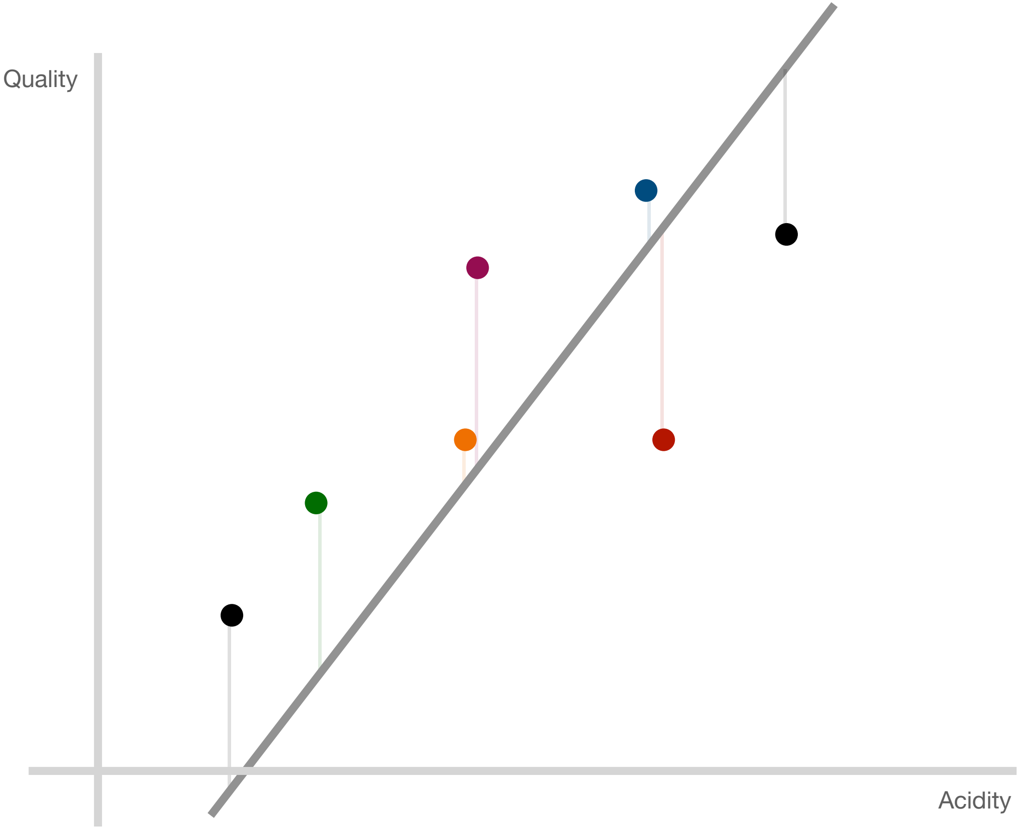
Linear Regression

STATISTICS!



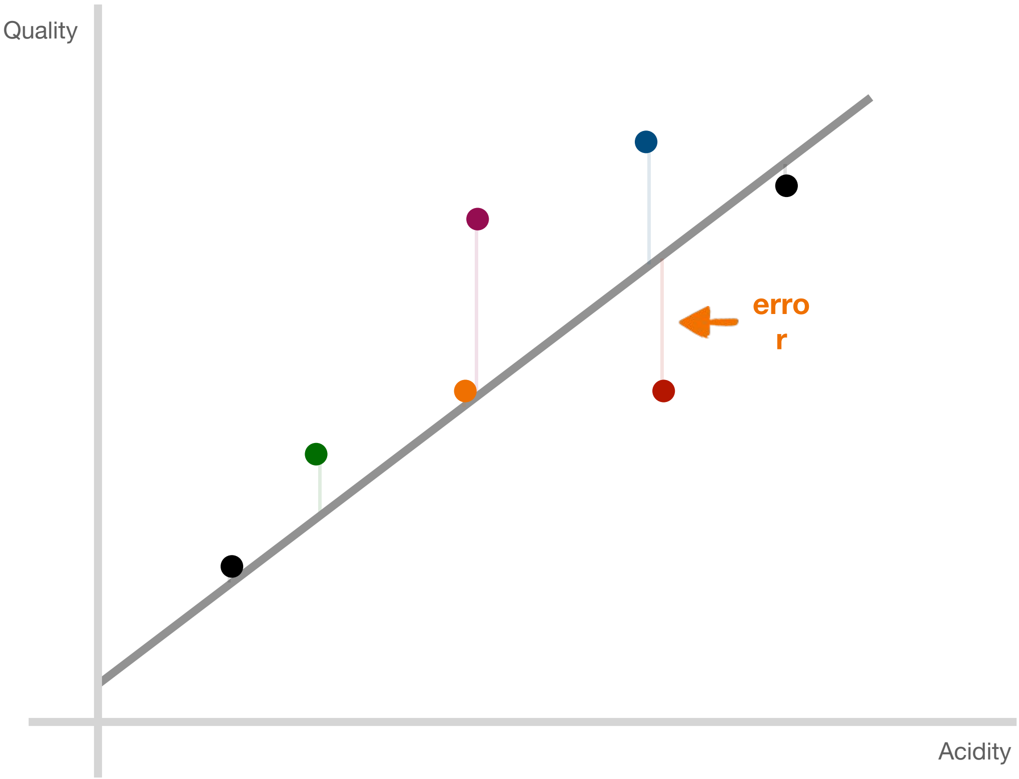
Linear Regression

STATISTICS!



Linear Regression

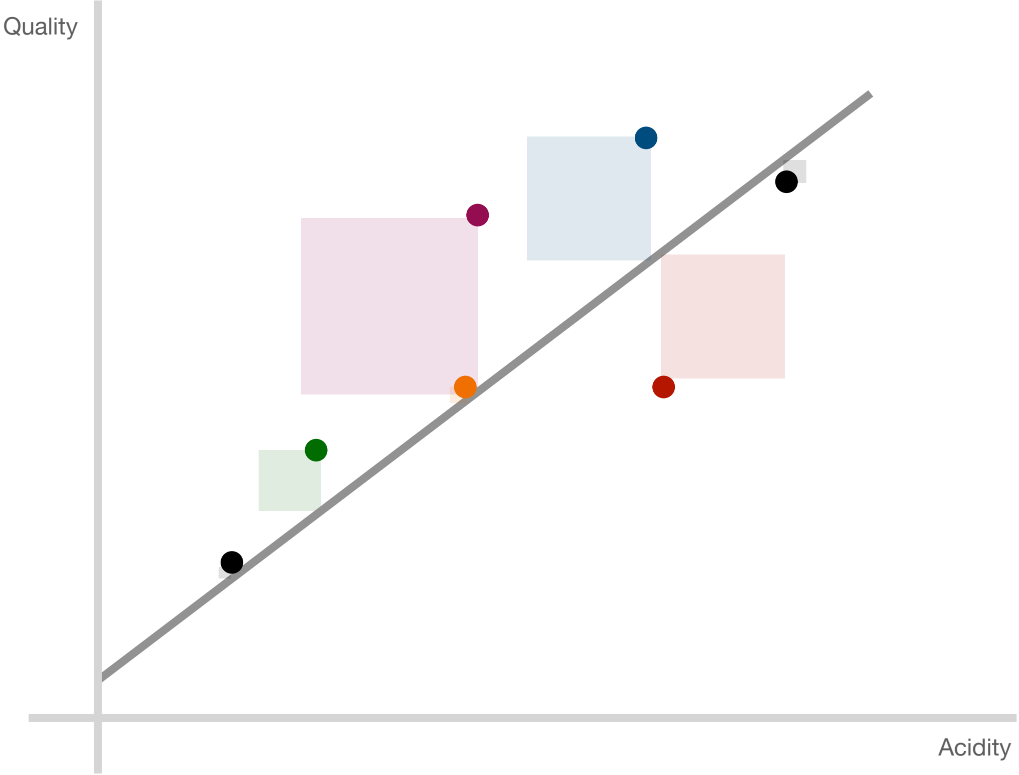
STATISTICS!



- **error** is a measure of the “incorrectness” of a line

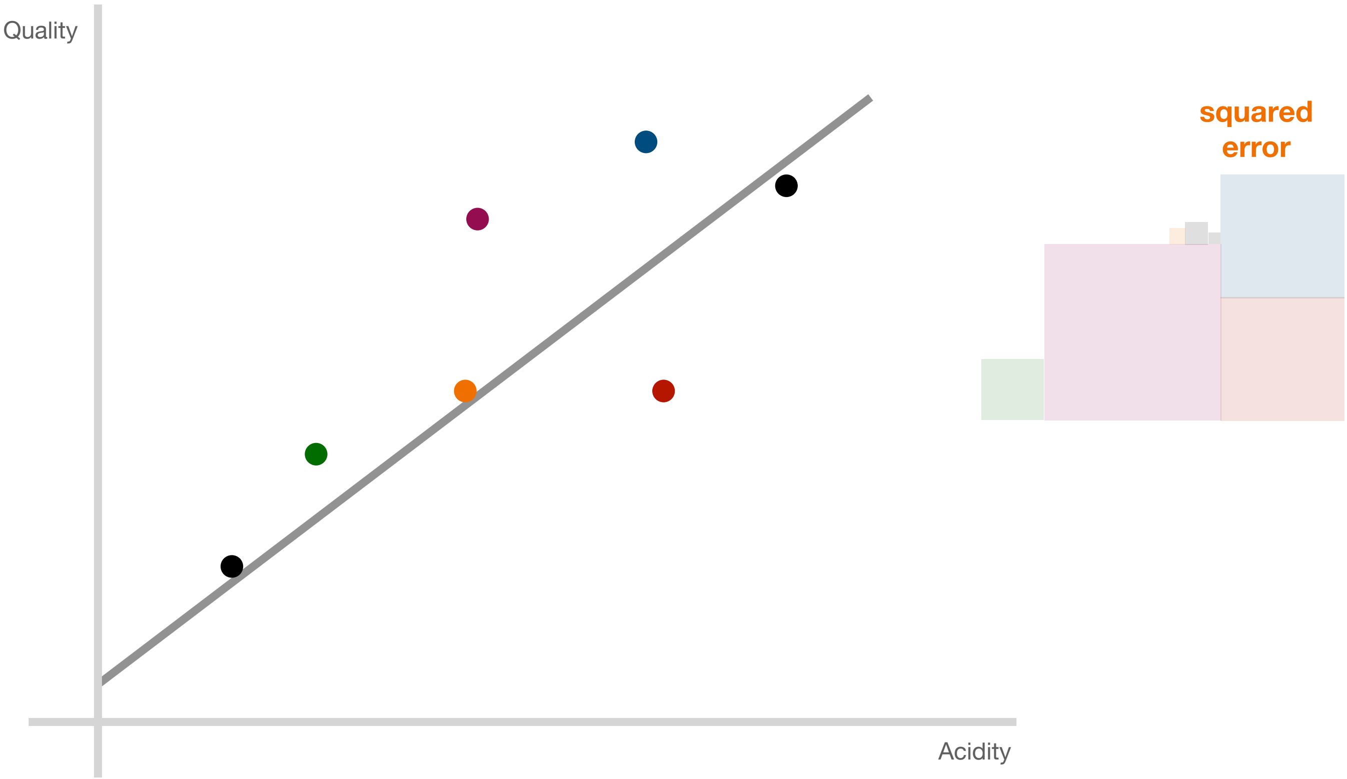
Linear Regression

STATISTICS!



Linear Regression

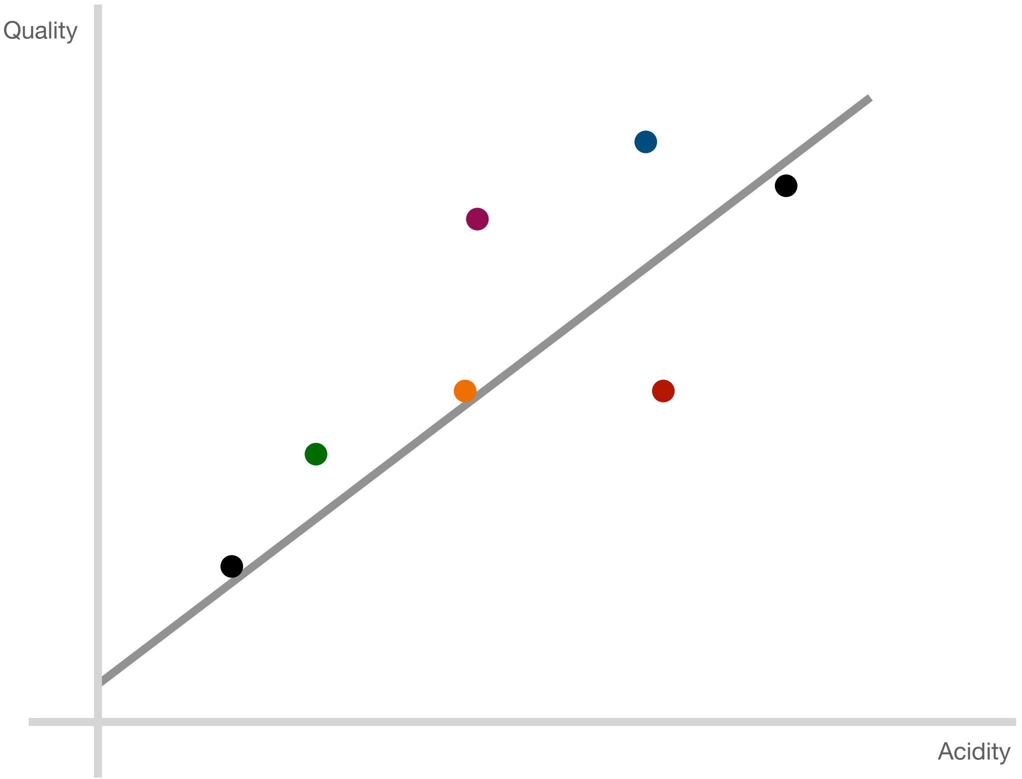
STATISTICS!



- **sum-of-squared** error is a common error metric for linear regression
- **sum-of-squared** error is also known as “L2 Penalty”

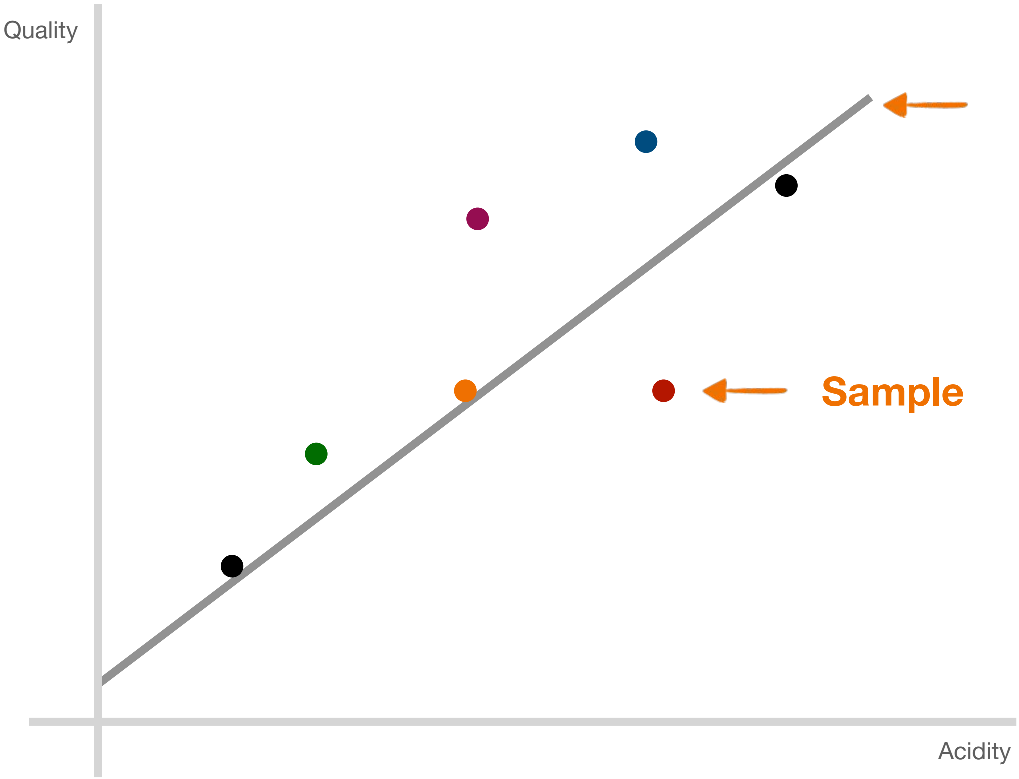
Linear Regression

STATISTICS!



Linear Regression

Output



Sample

Input

Logistic Regression

a possible solution…

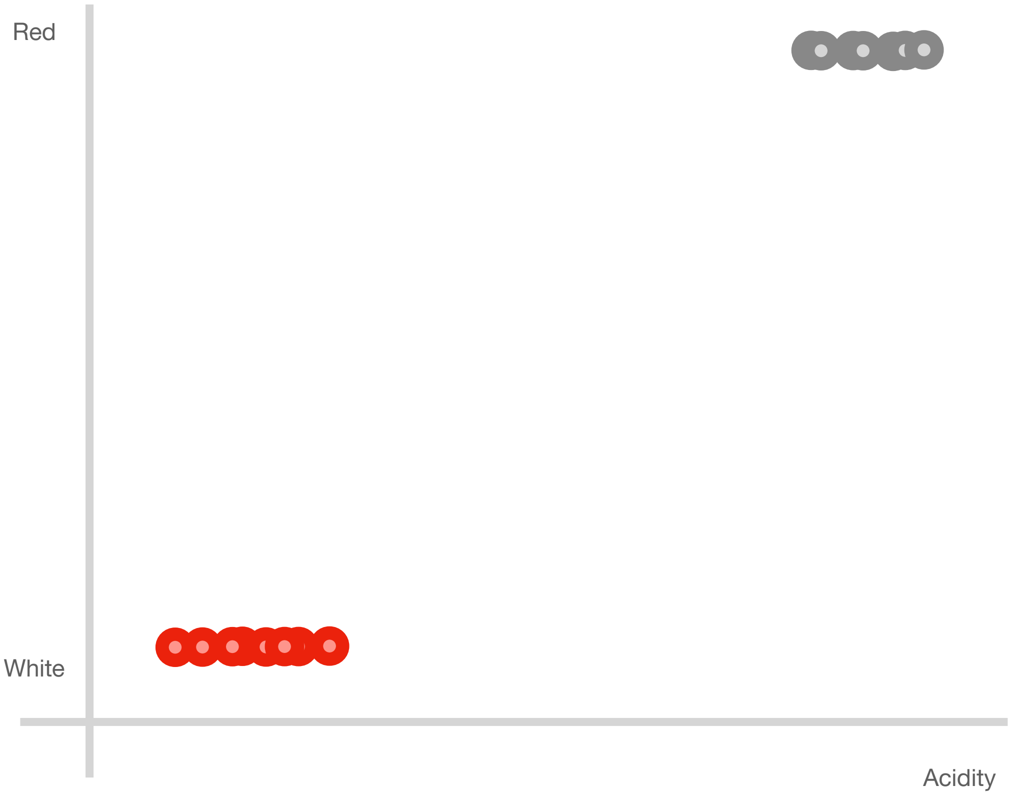
Red = 0

White = 1

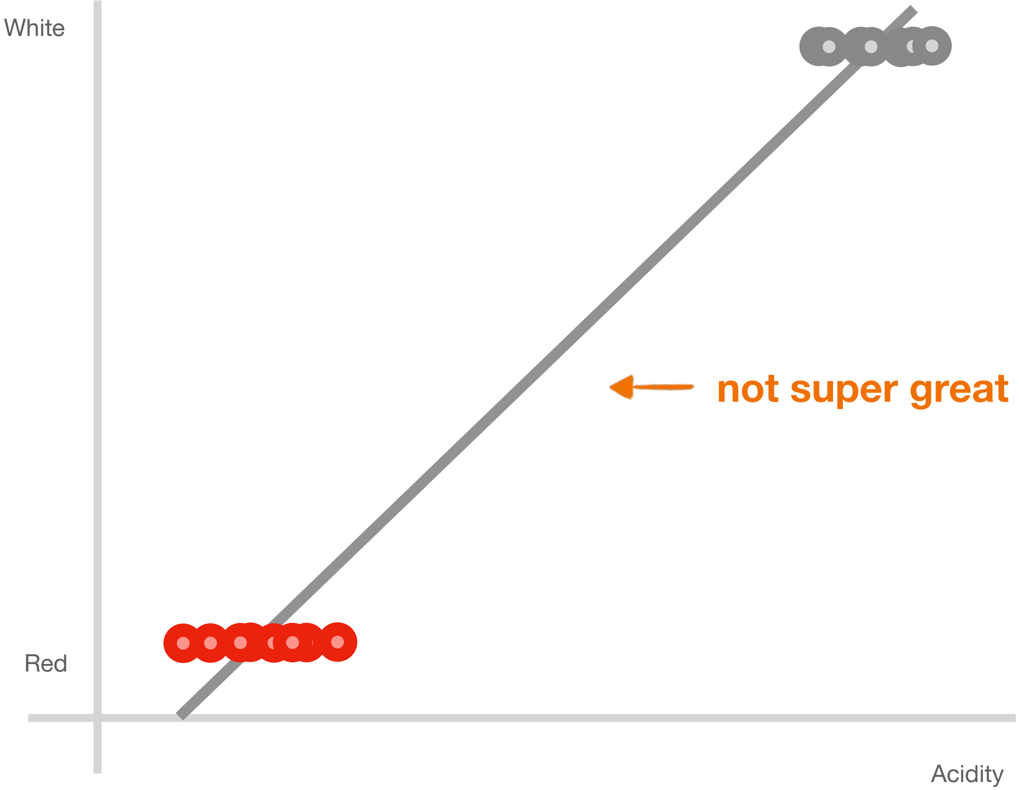
- categorical label outputs are named "**classes**"

Logistic Regression

Logistic Regression

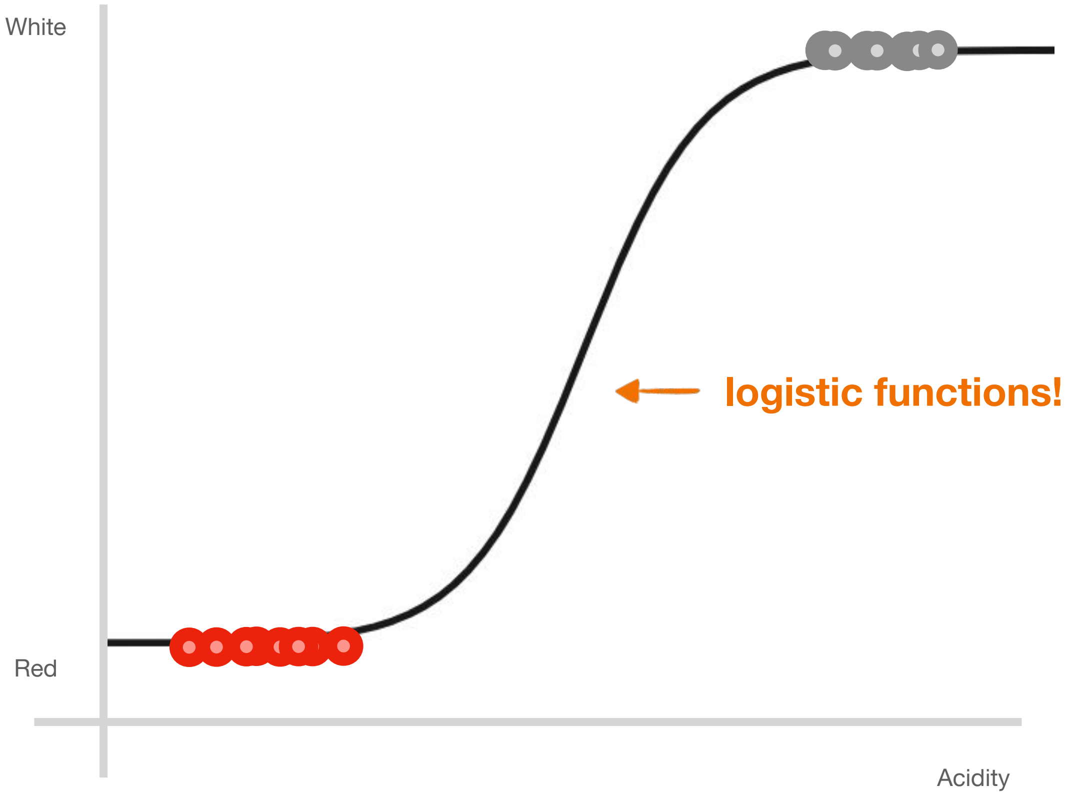


Logistic Regression



not super great

Logistic Regression



logistic functions!

- **logistic functions** are a good way of making predictions on binary/categorical data
- **sum-of-squared** error is *still a* common error metric for logistic regression

Logistic Regression

White Red Champagne

pair them up!

Logistic Regression

White White Red

Red Champagne Champagne

Logistic Regression

White	White	Red
Red	Champagne	Champagne

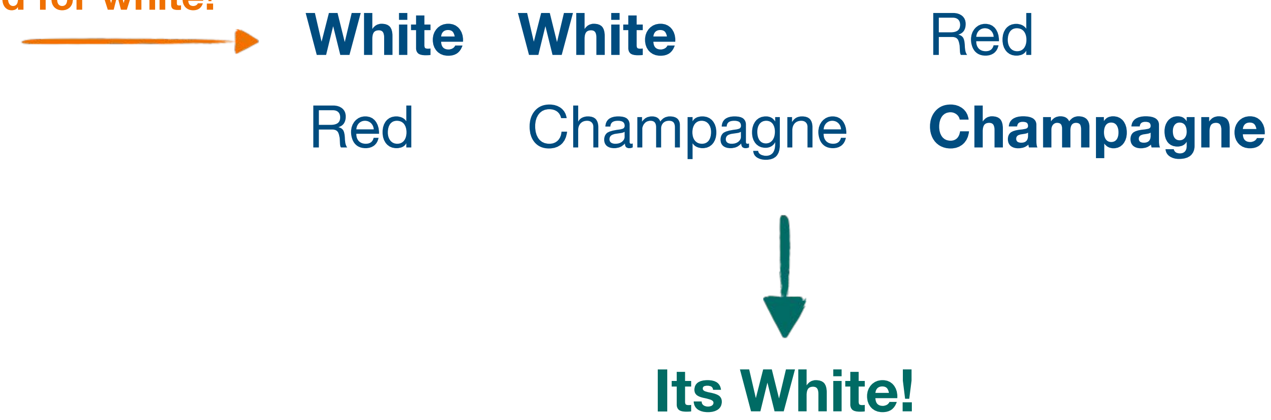
Logistic Regression

White	**White**	Red
Red	Champagne	Champagne

Logistic Regression

White	**White**	Red
Red	Champagne	**Champagne**

Logistic Regression



more pairs voted for white!

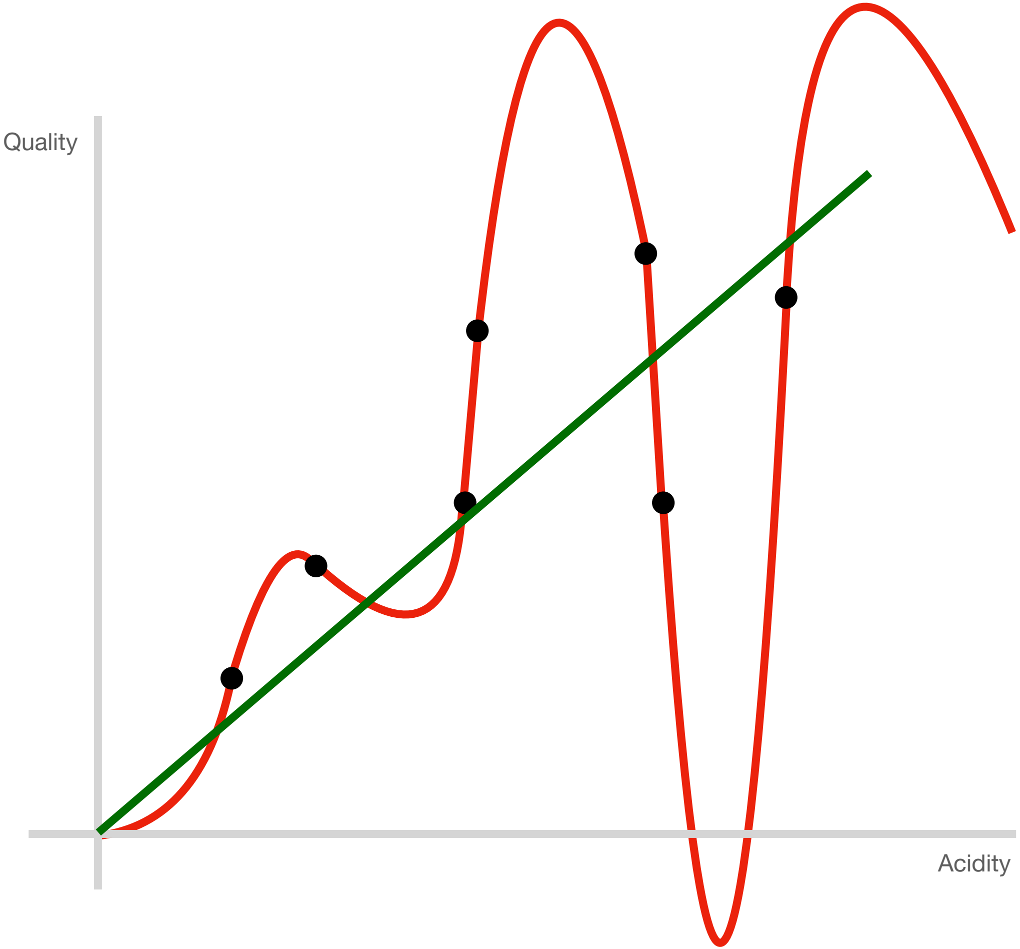
White **White** Red

Red Champagne **Champagne**

Its White!

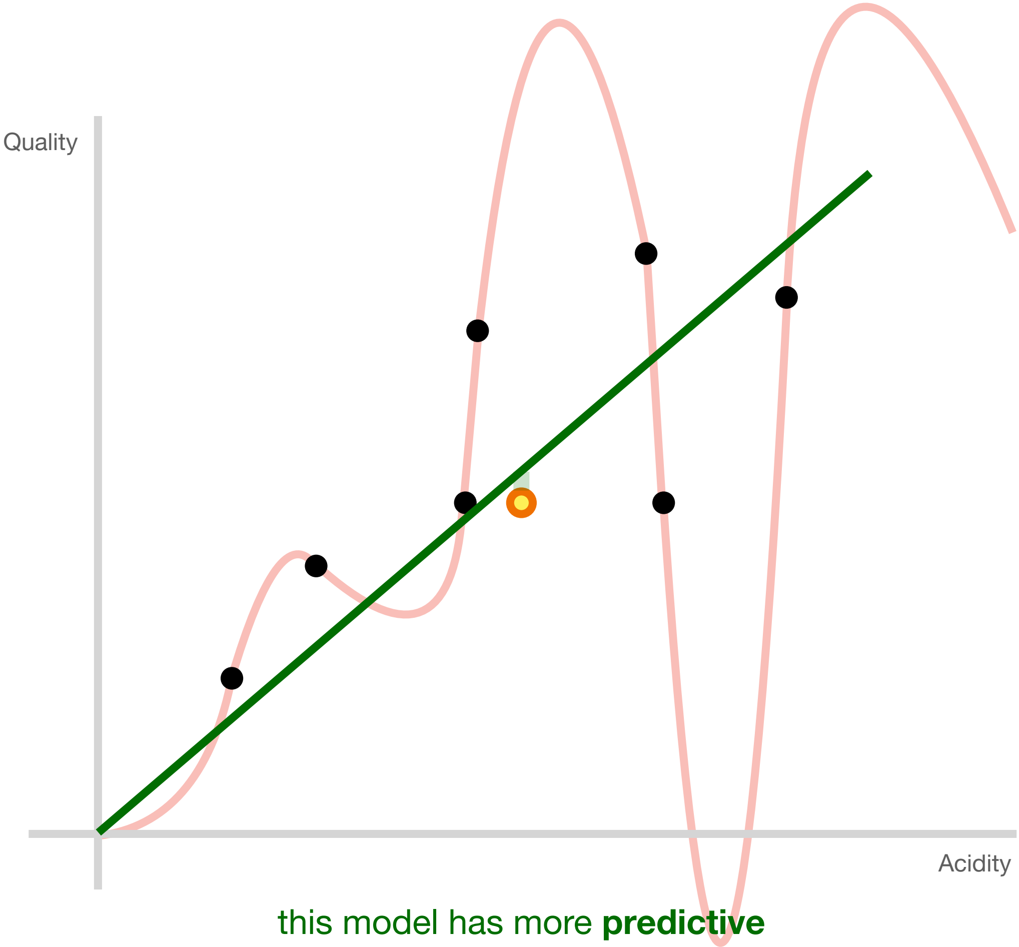
- **One-vs-one** multiclass classification uses the most “voted for” class among paired models

overfitting



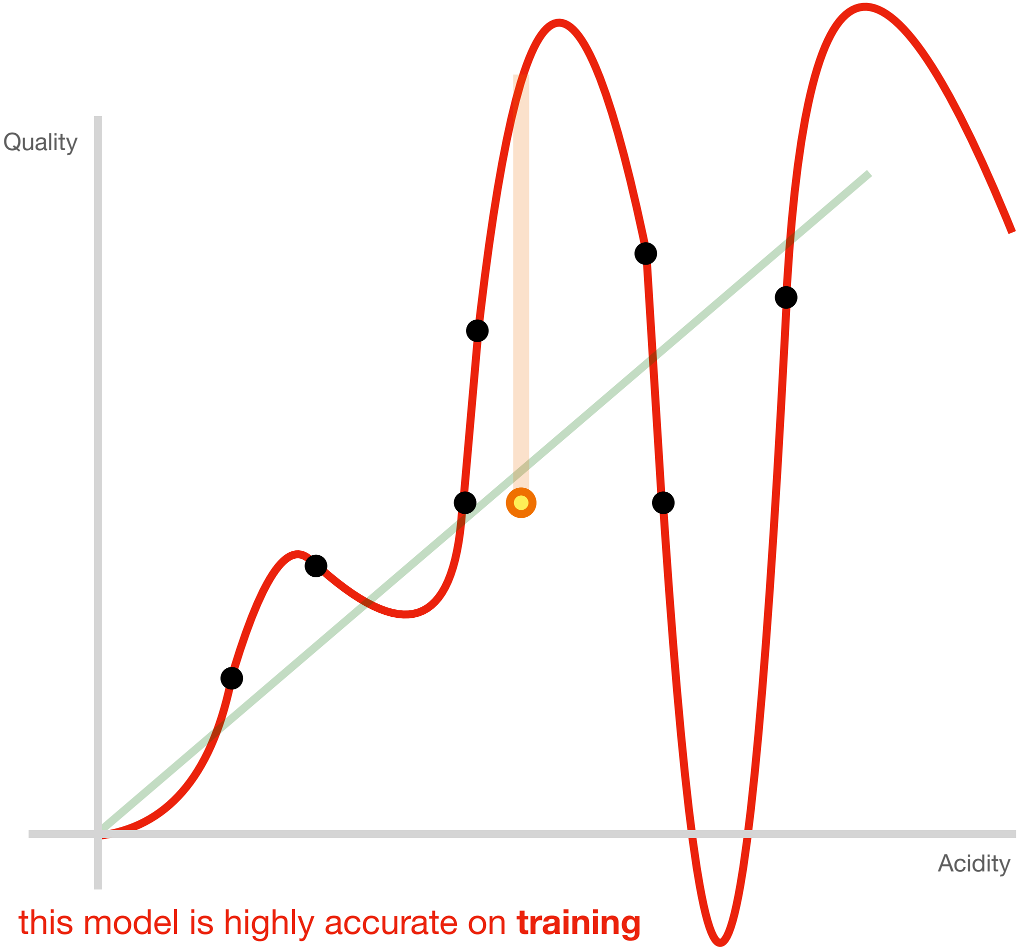
Which one is a better line?

overfitting



this model has more **predictive power**

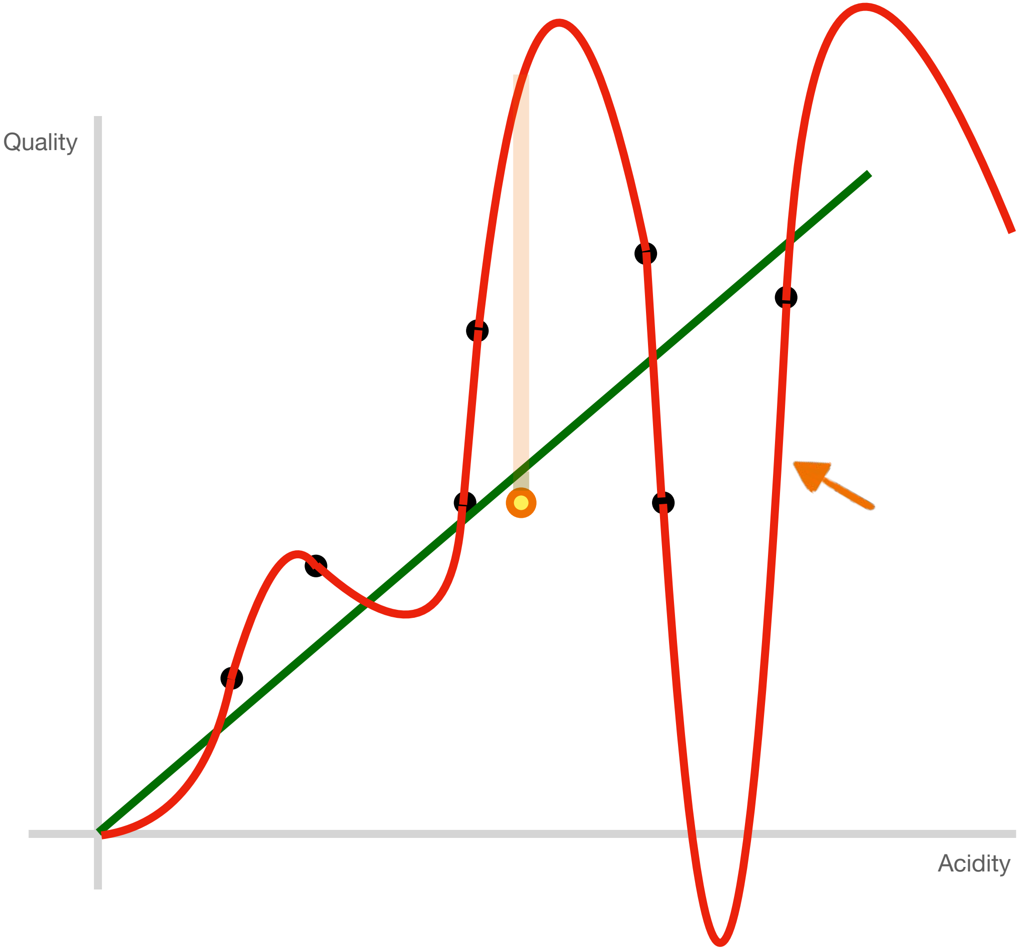
overfitting



this model is highly accurate on **training data**

but bad at predictions anywhere else

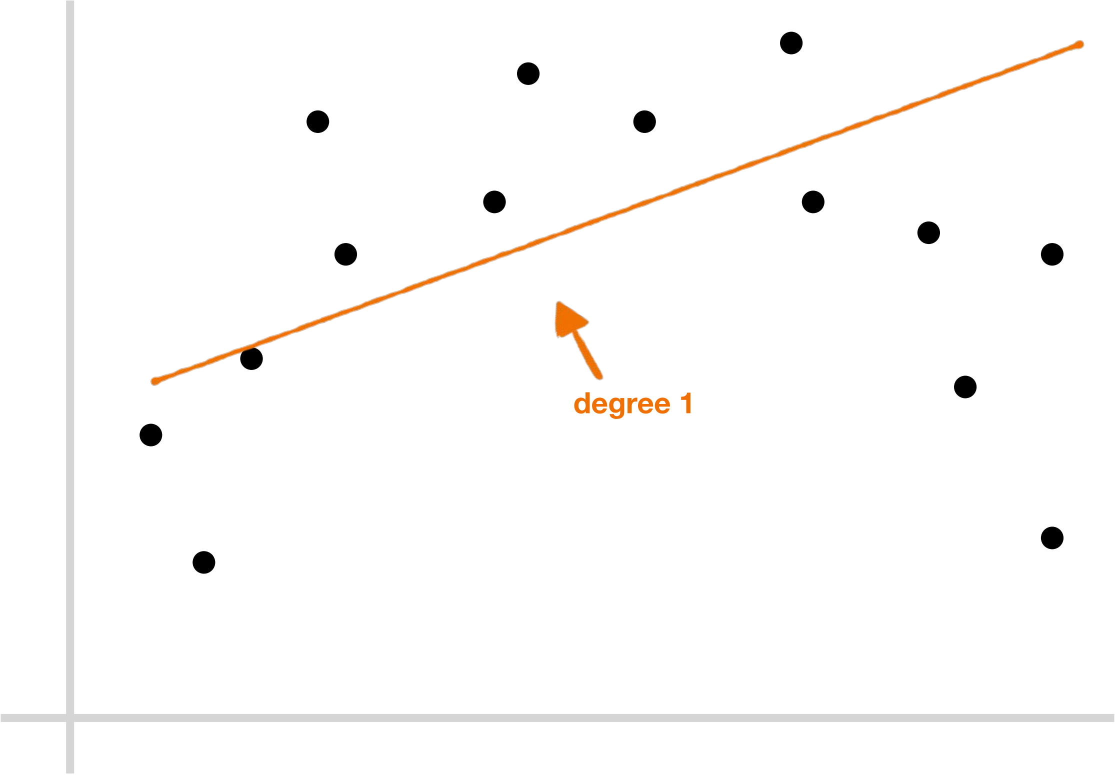
overfitting



- too-precise fits to original data without generalization is called **overfitting**

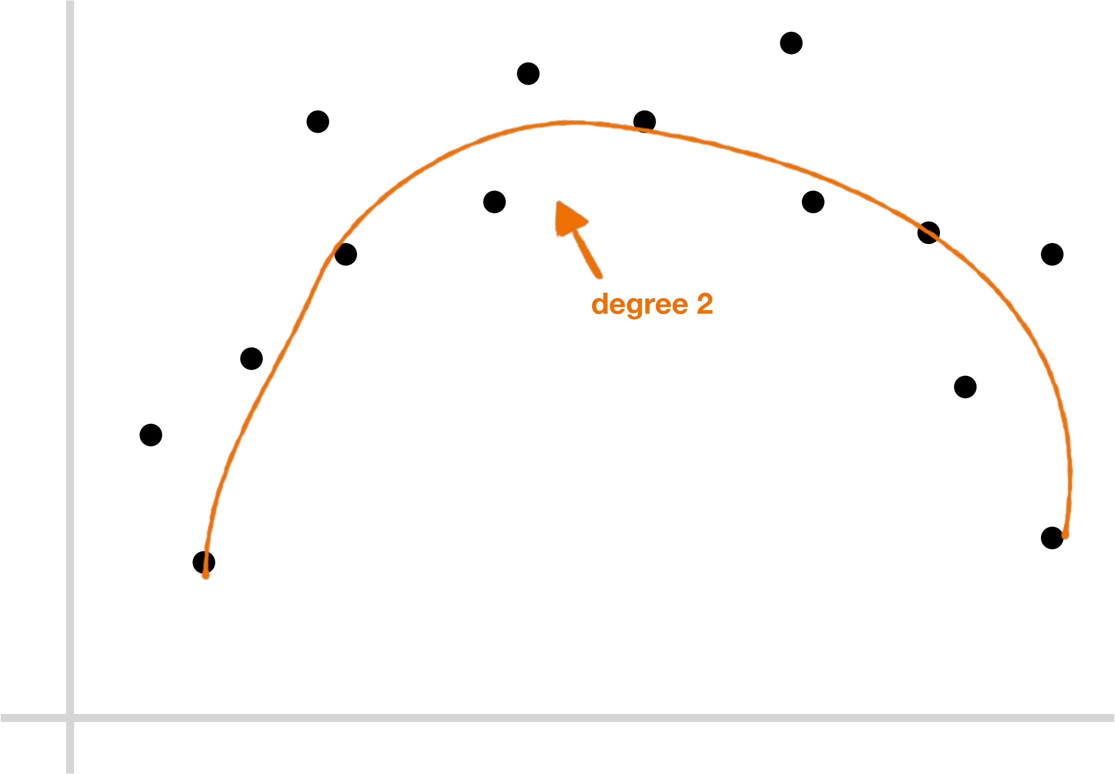
overfitting

underfit!

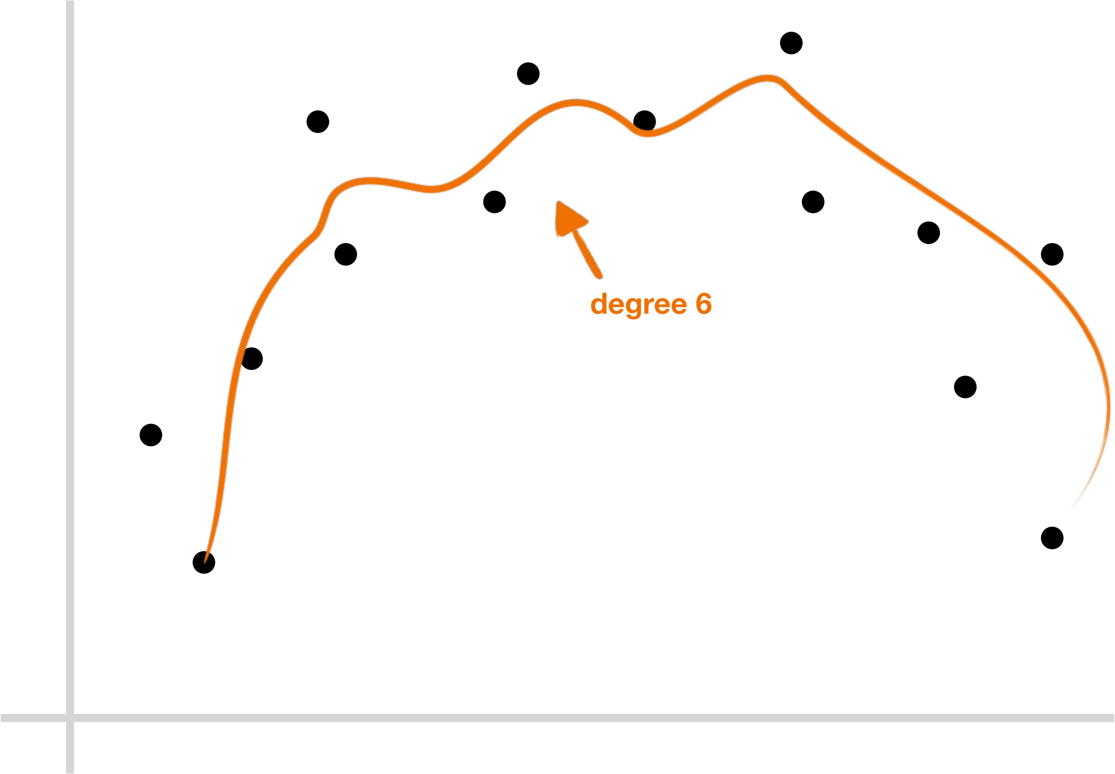


- model is unable to capture relationship between variables

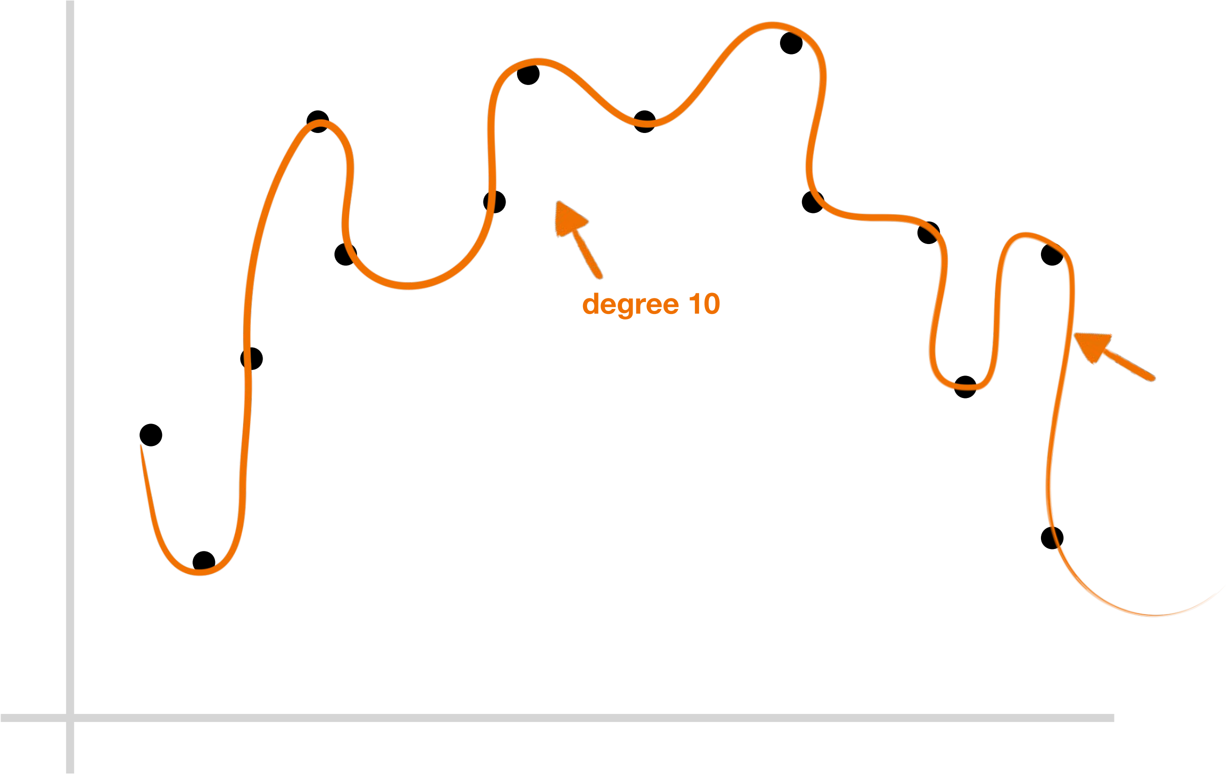
overfitting



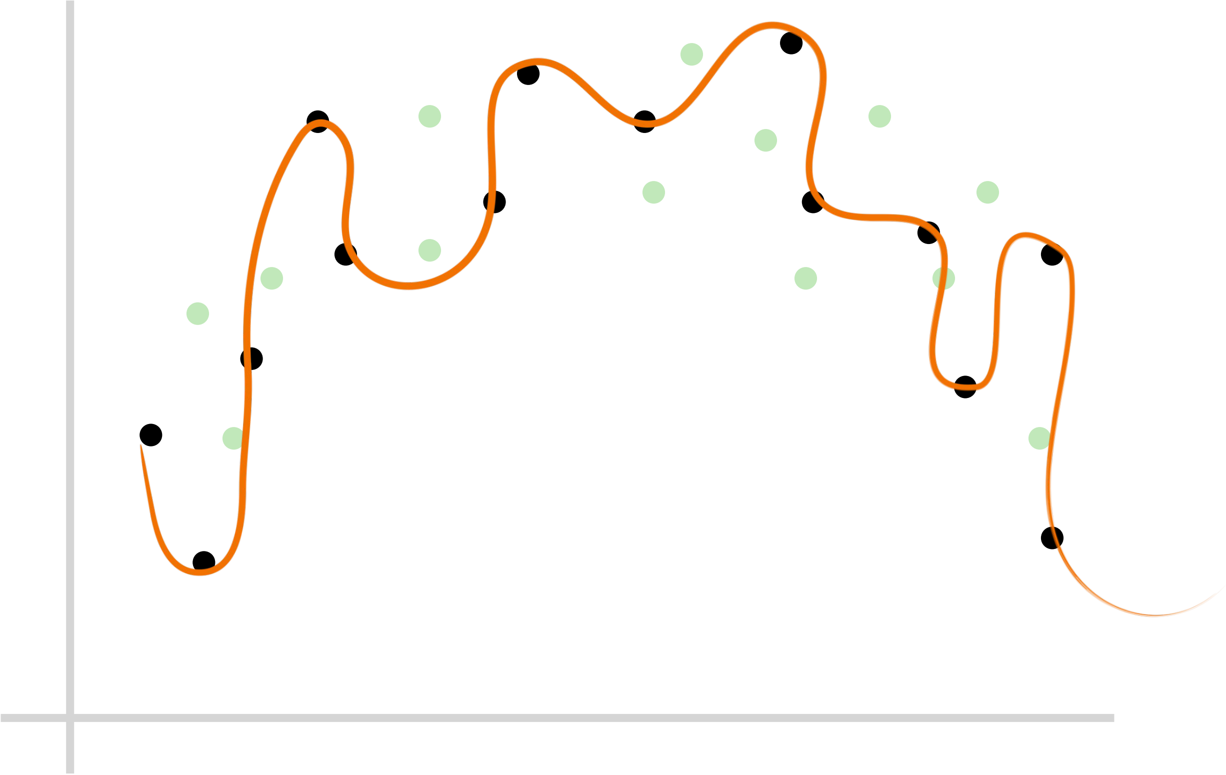
overfitting



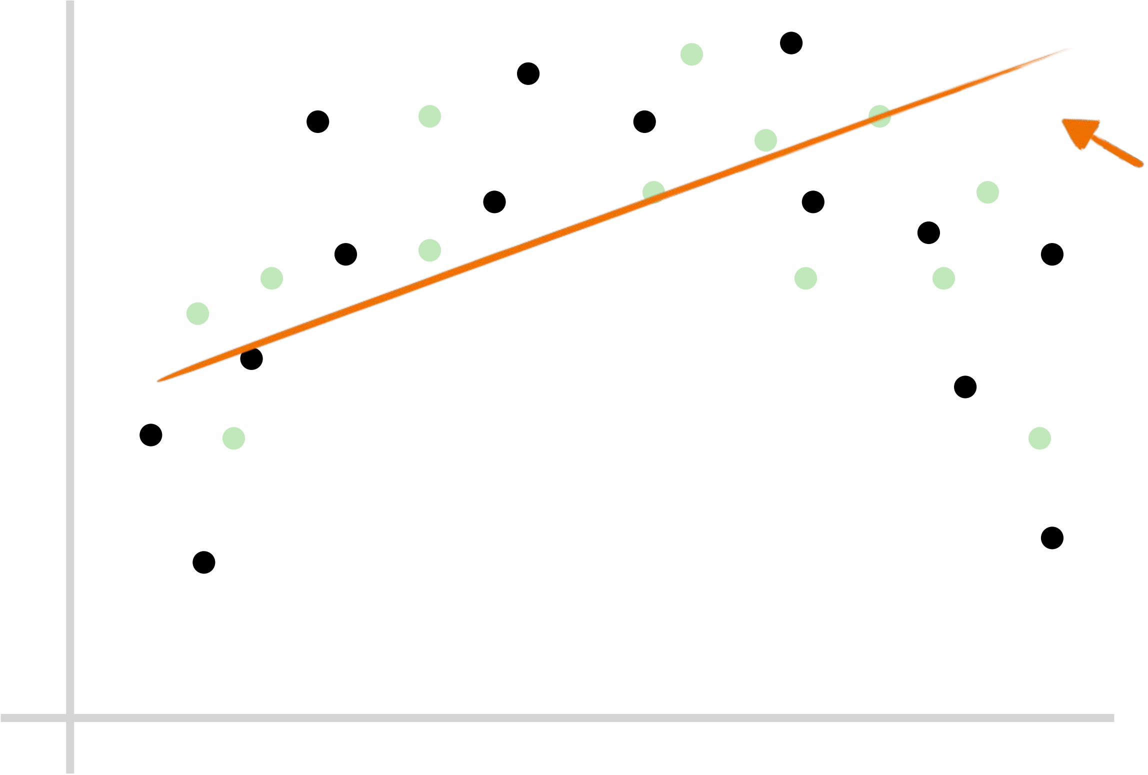
overfitting



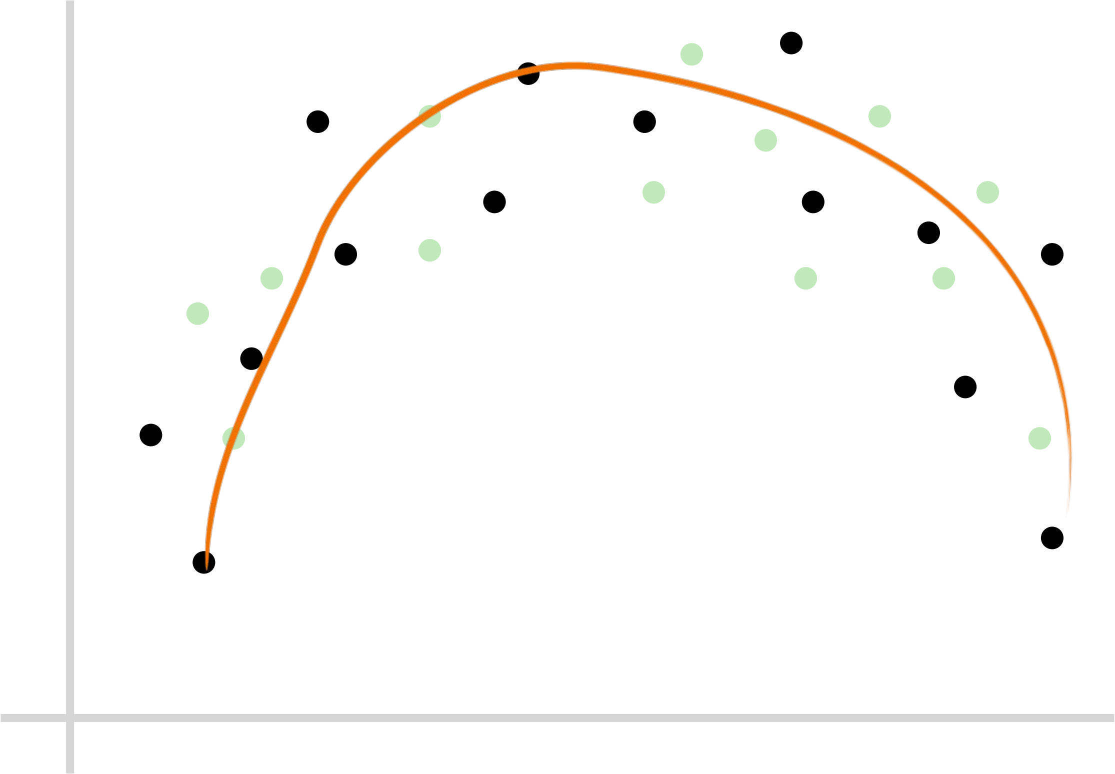
overfitting



overfitting

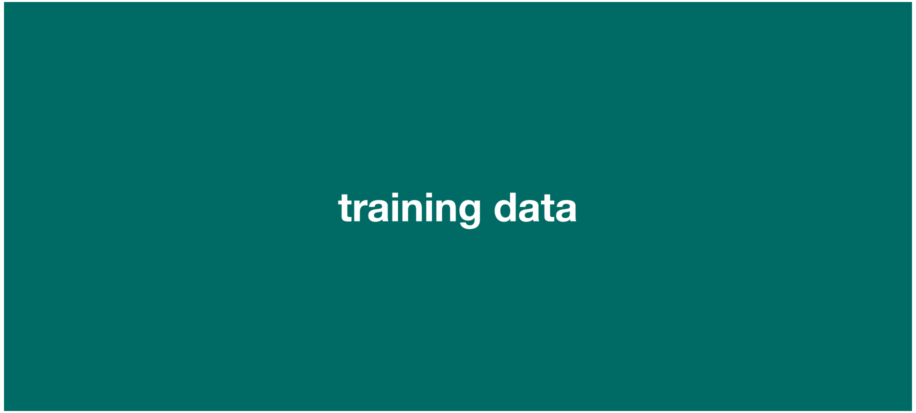


overfitting



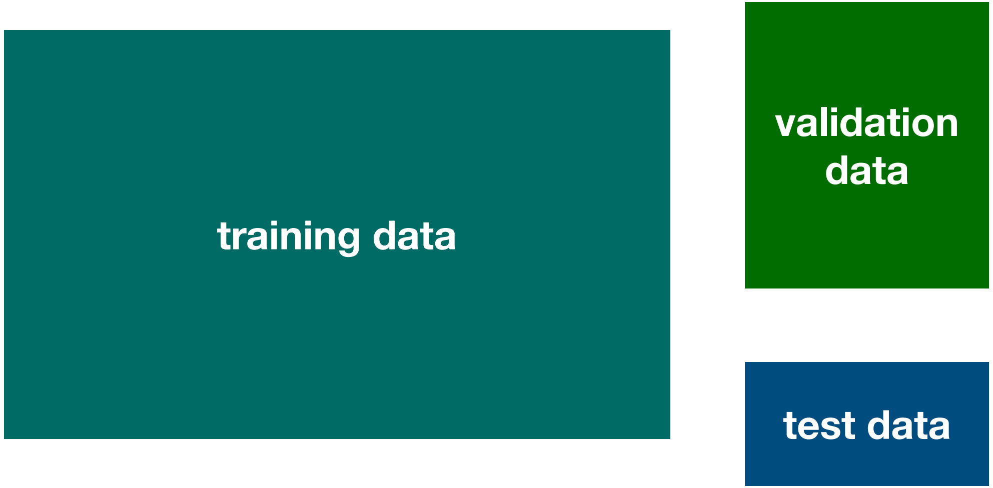
How do we address under/overfitting?

address overfitting



training data

address overfitting

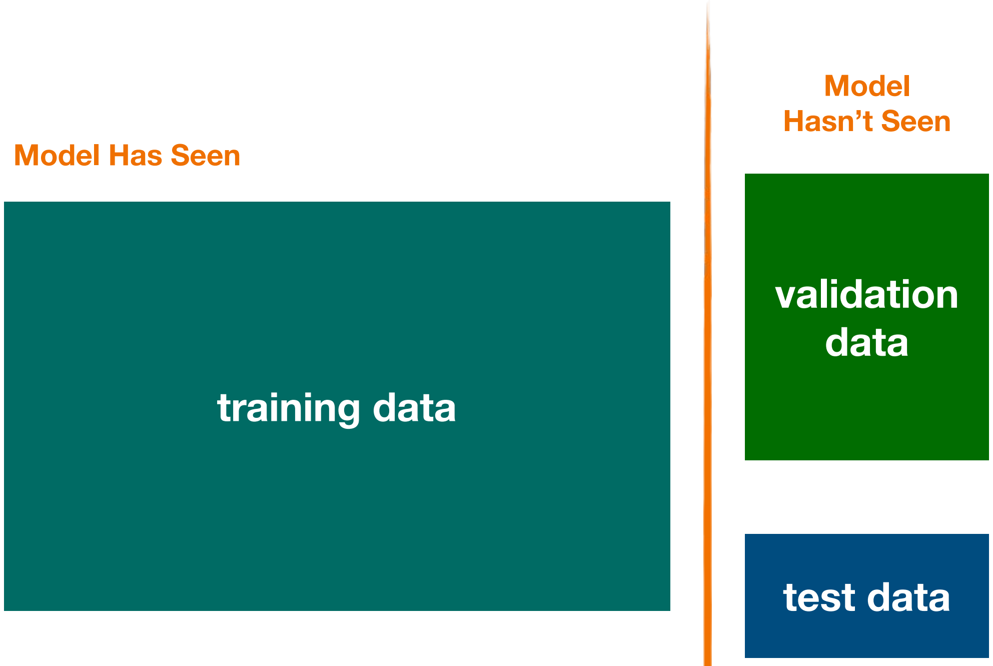


training data

validation data

test data

address overfitting



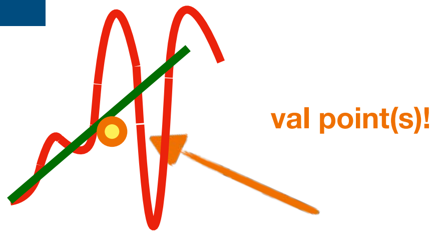
Model Has Seen

training data

Model Hasn't Seen

validation data

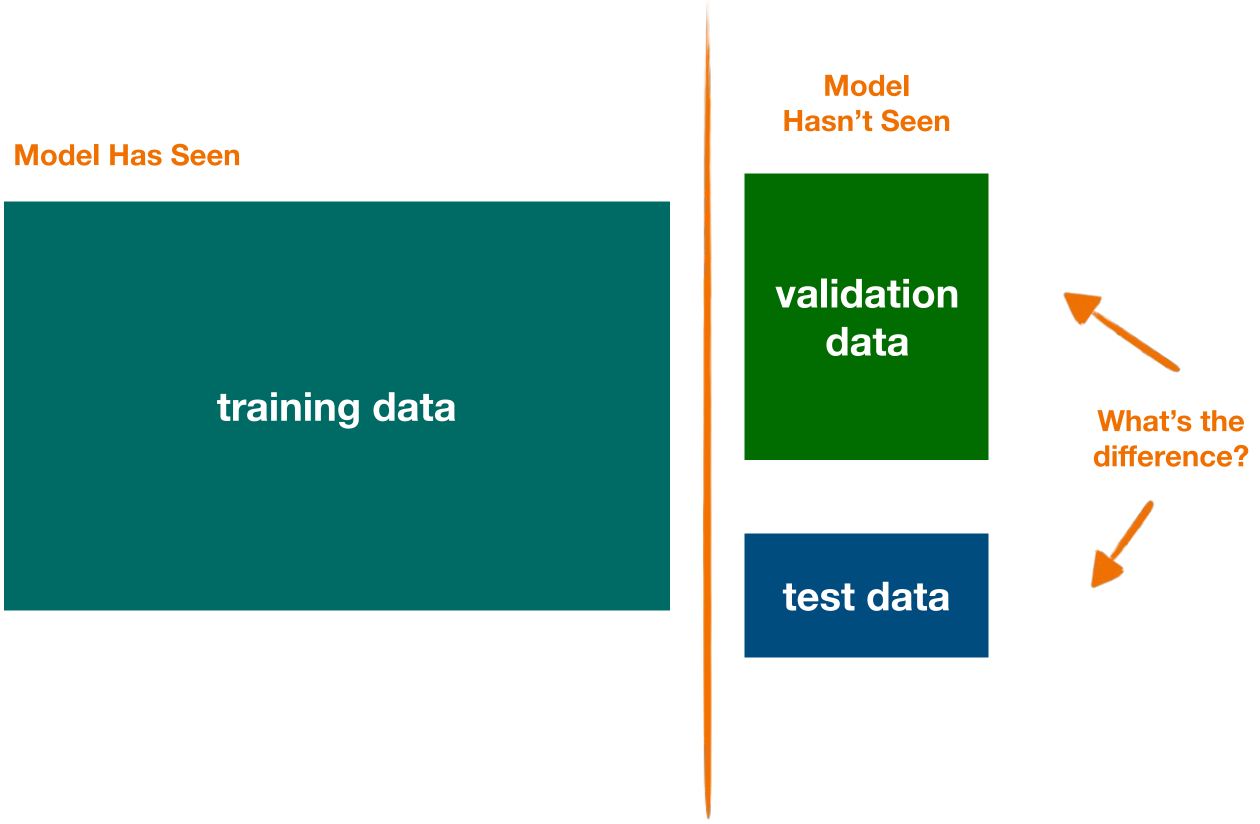
test data



val point(s)!

- we use **validation** and **test** sets, small subsets of data the model hasn't seen before,

address overfitting



Model Has Seen

training data

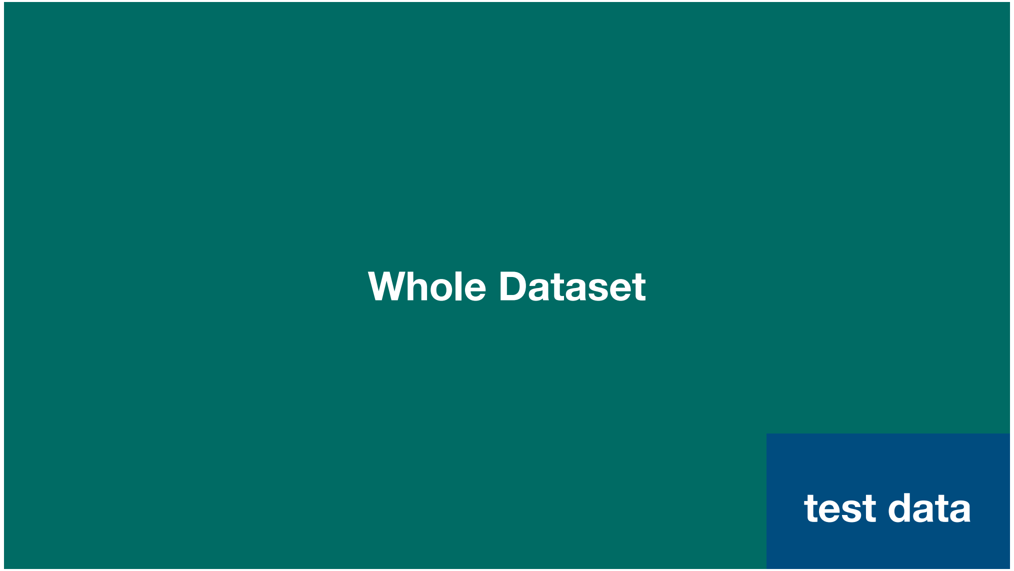
Model Hasn’t Seen

validation data

test data

What’s the difference?

address overfitting



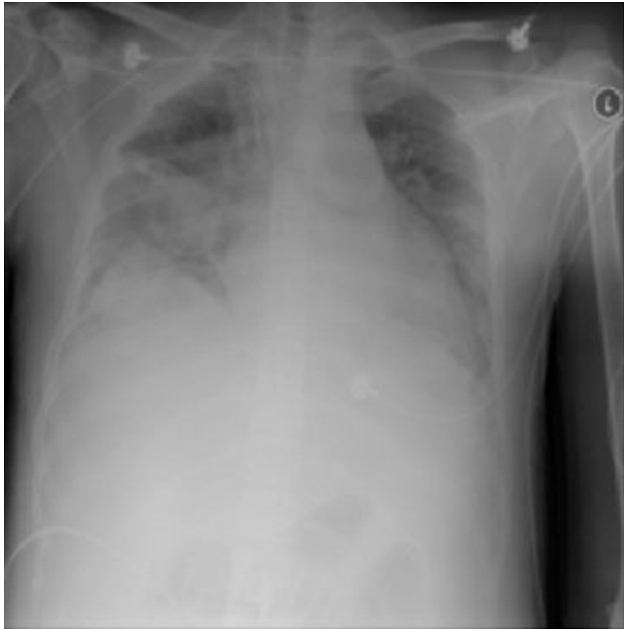
standardized for benchmarking!

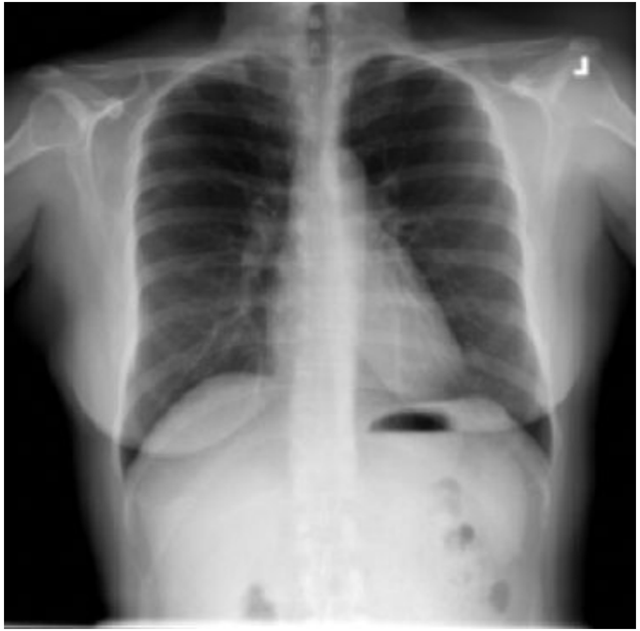
- **test sets** are, unlike validation sets, usually set by the data creator as common, unseen benchmark data.

overfitting can be dangerous

data ethics

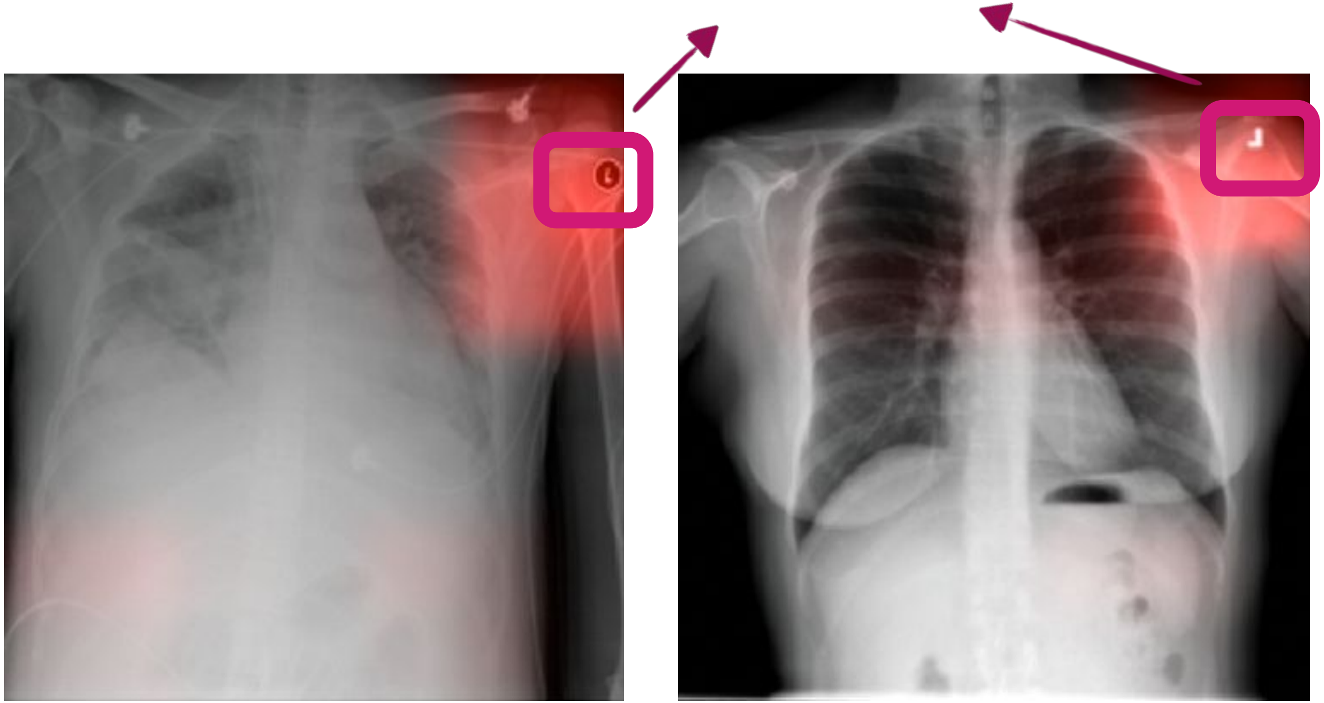
data ethics





which one has pneumonia?

data ethics



different hospitals used different markers!

- models, when not controlled for external factors, often **overfit** on easy targets