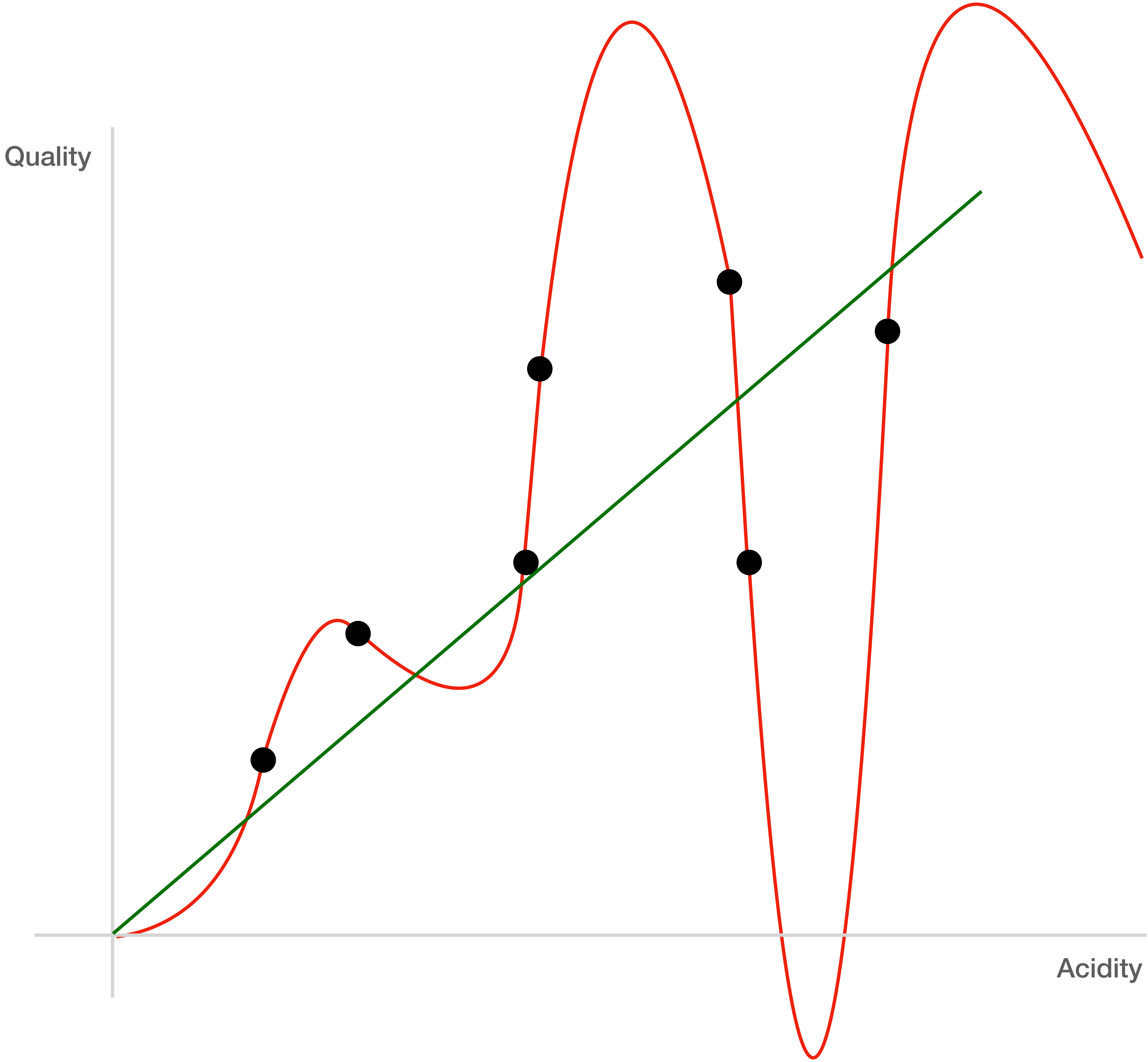
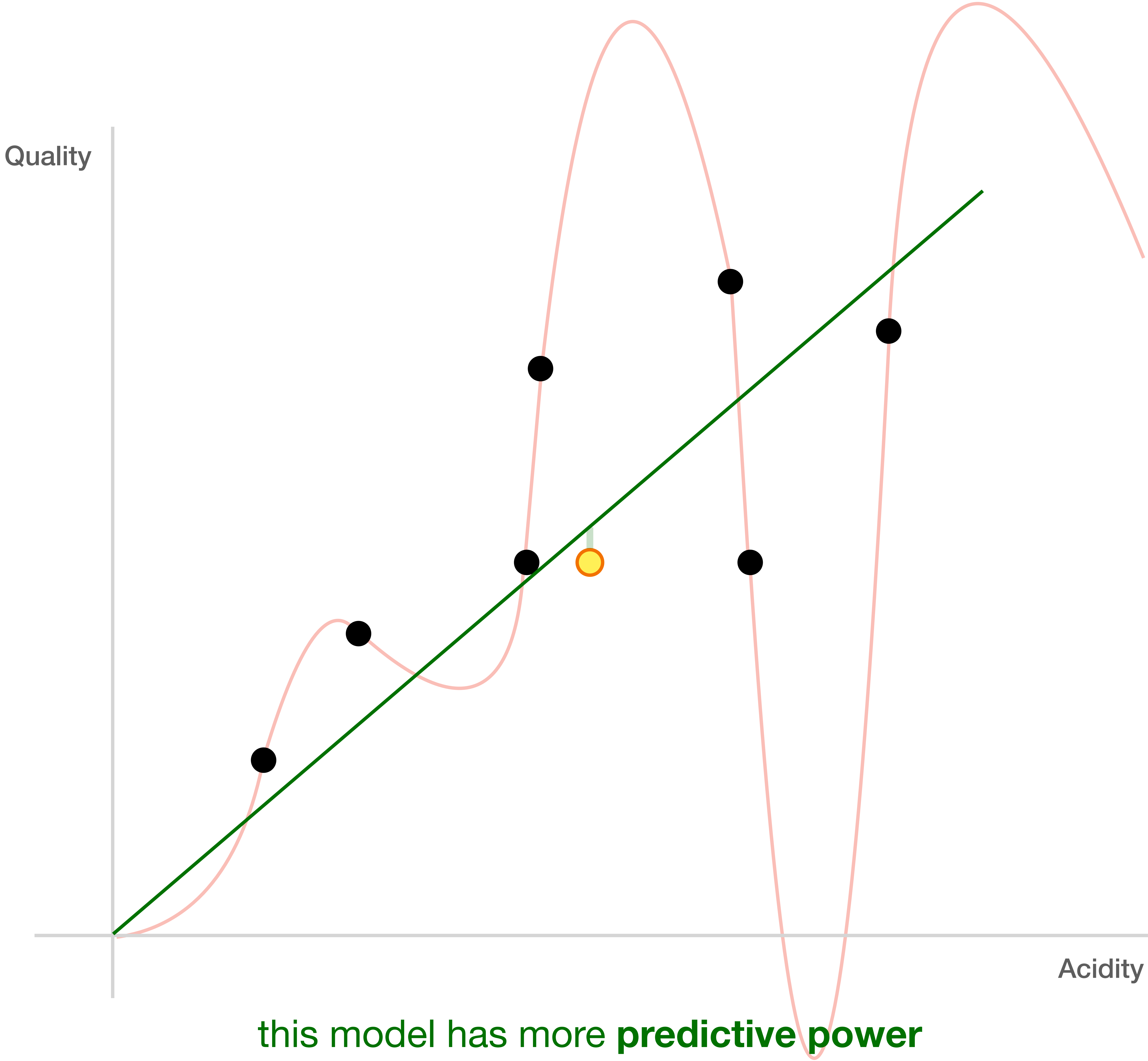
AIBridge

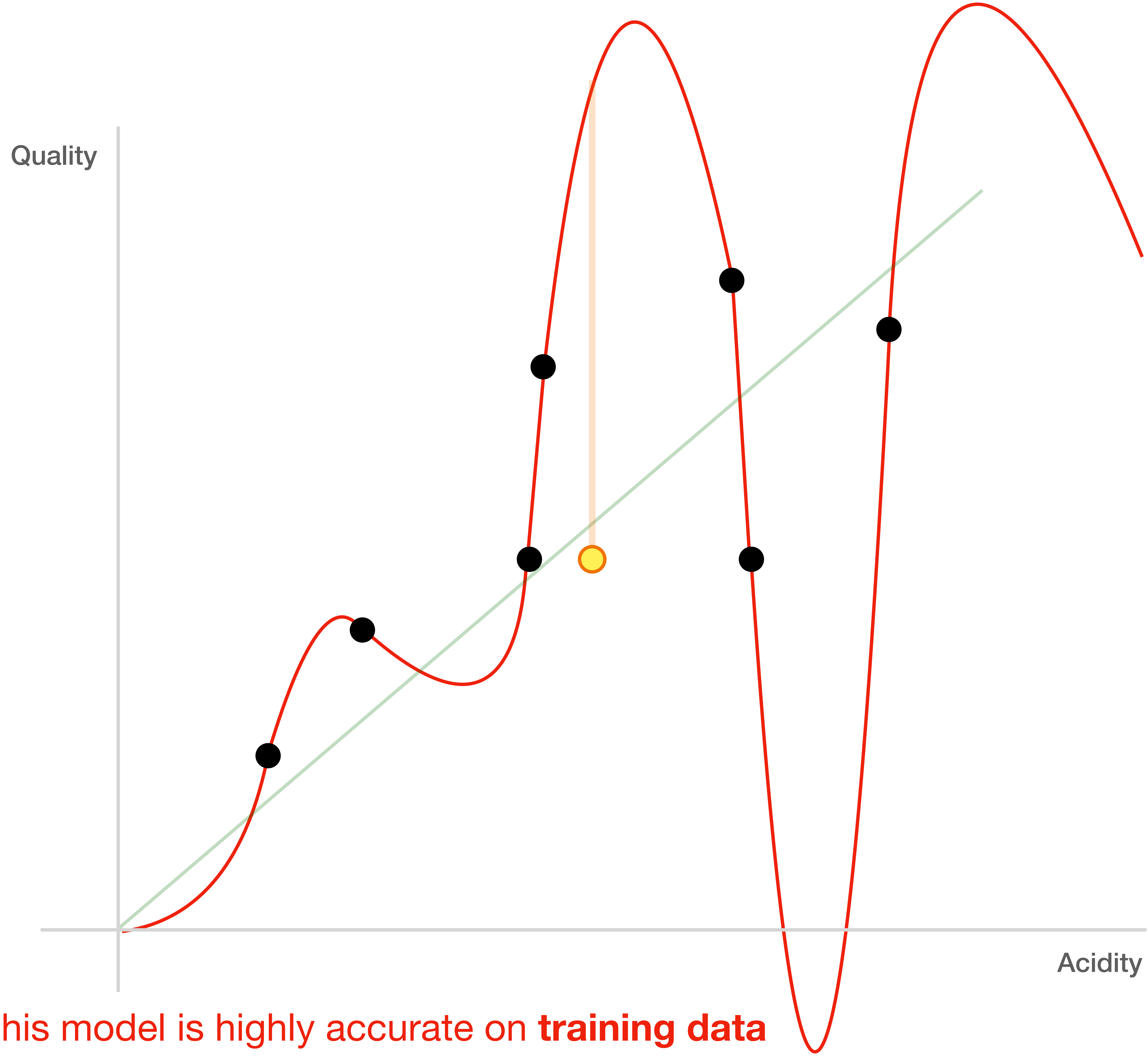
Lecture 7



Which one is a better line?

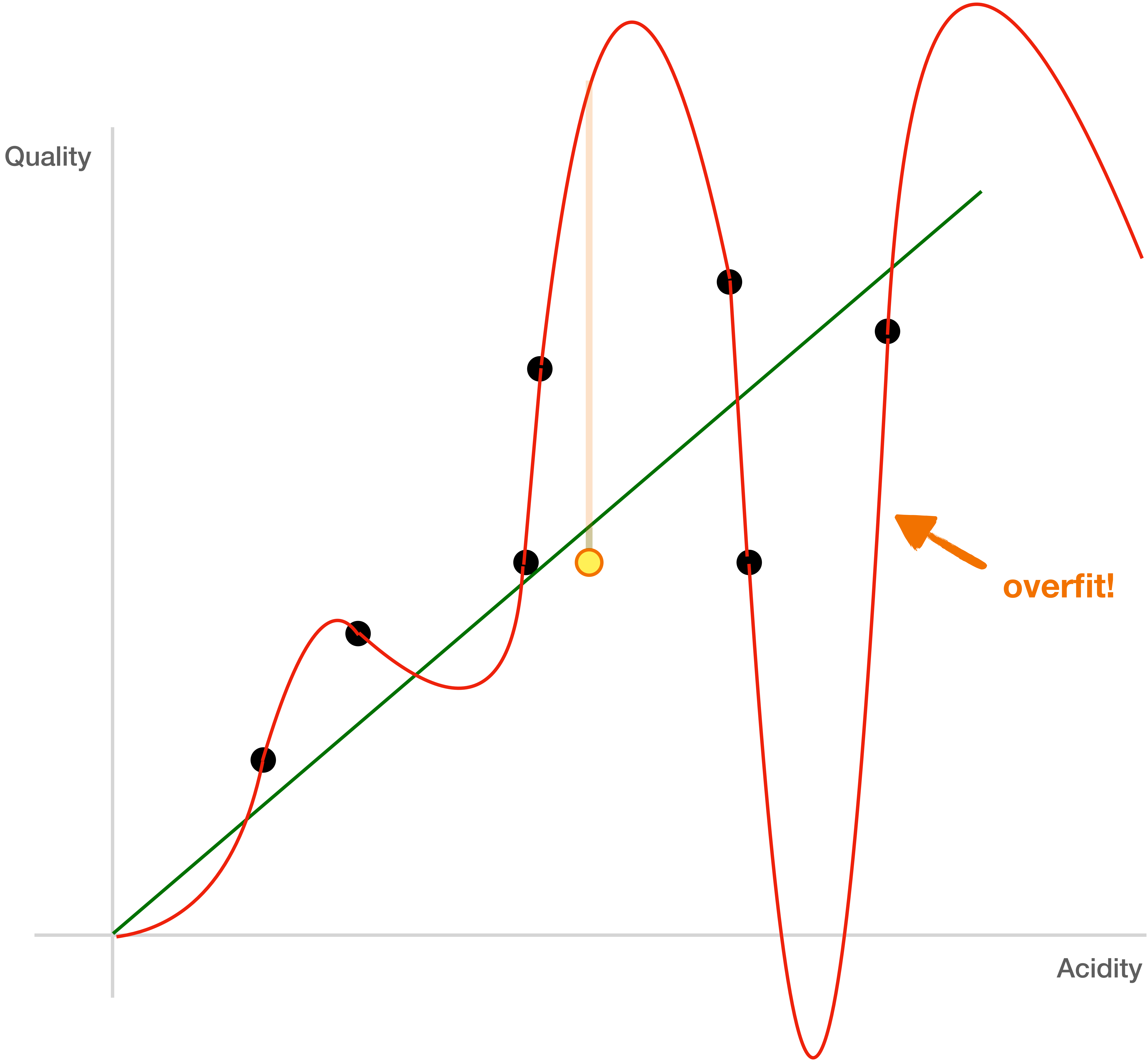


this model has more **predictive power**



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

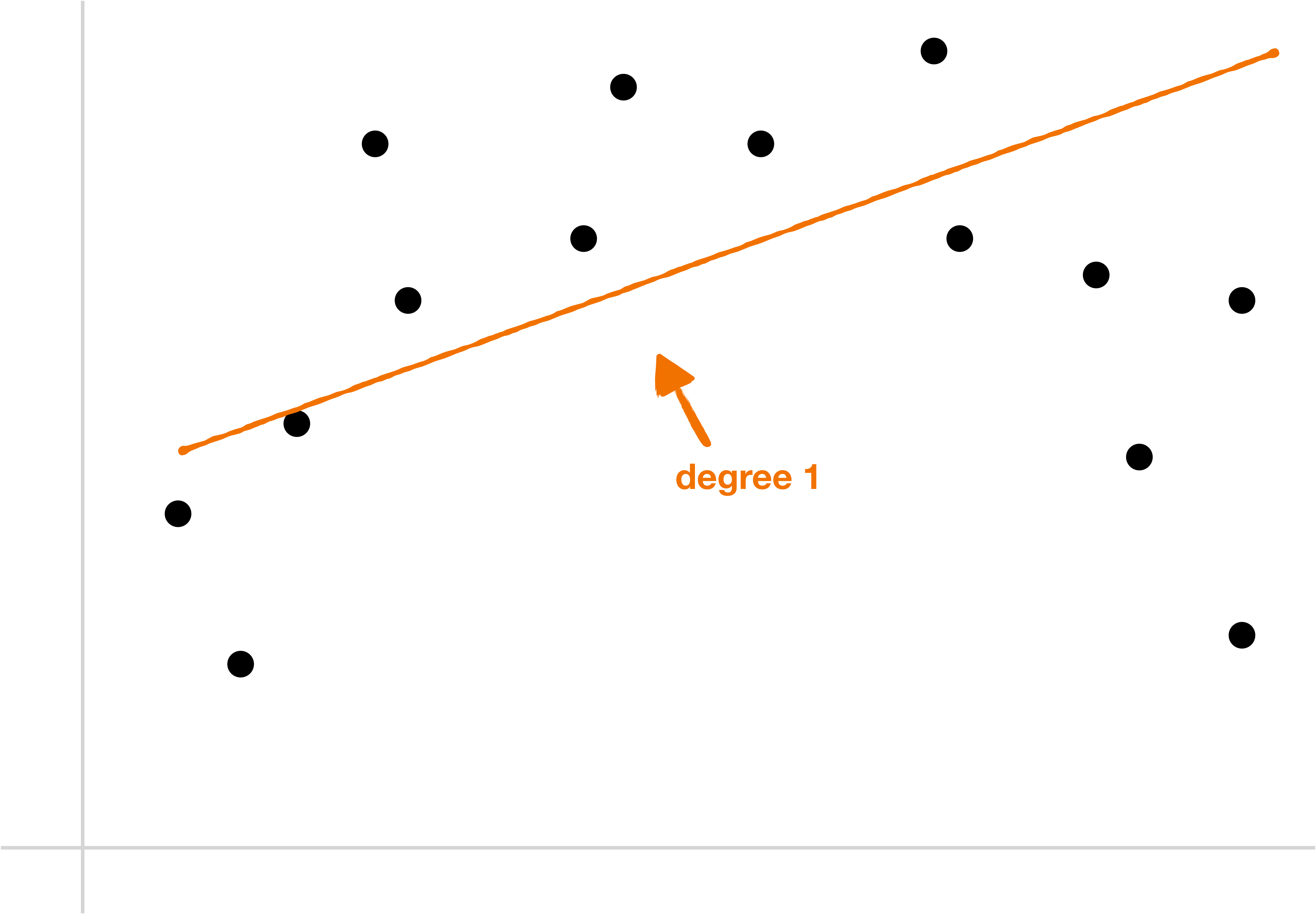
but bad at predictions anywhere else



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

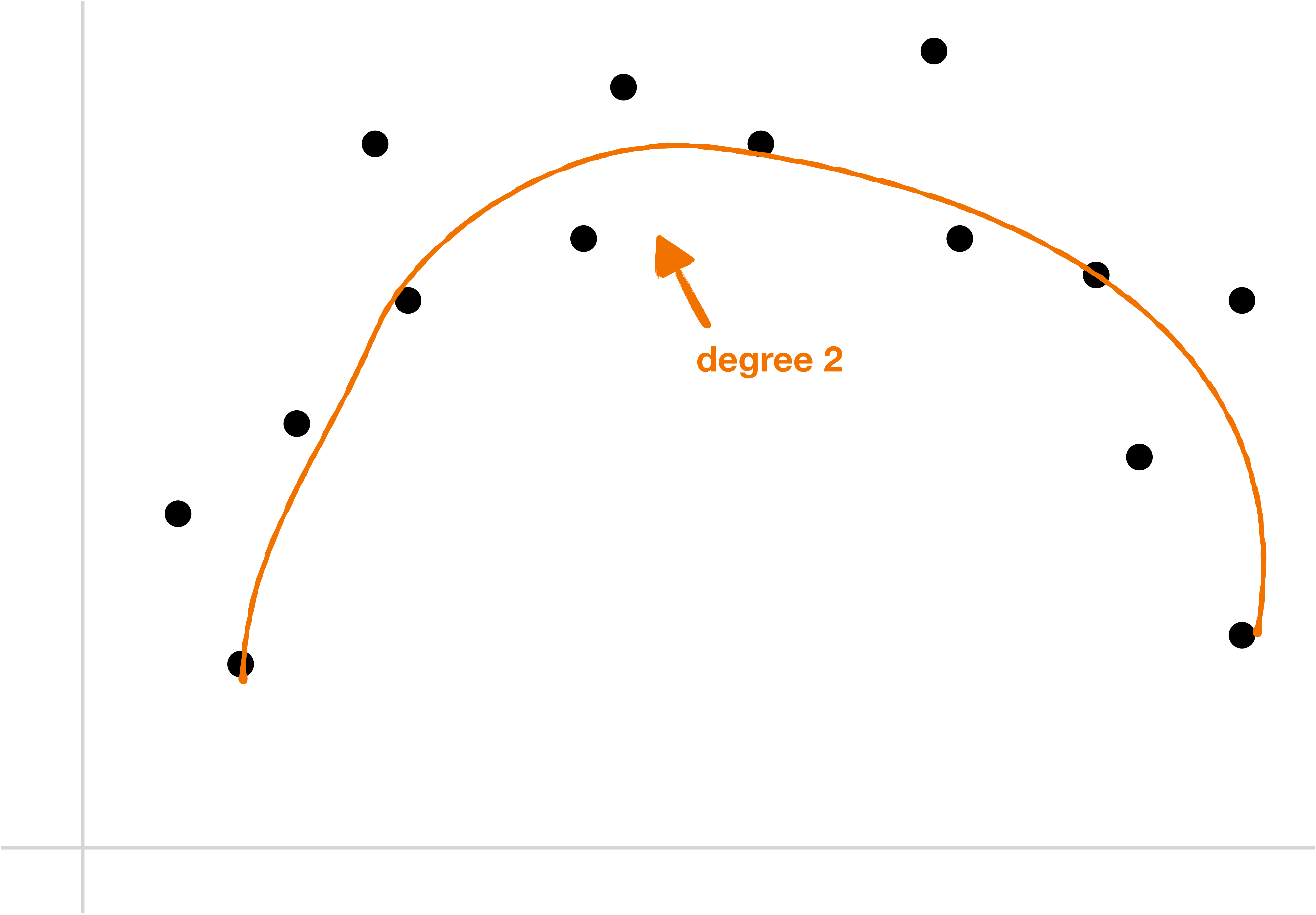
overfitting

underfit!

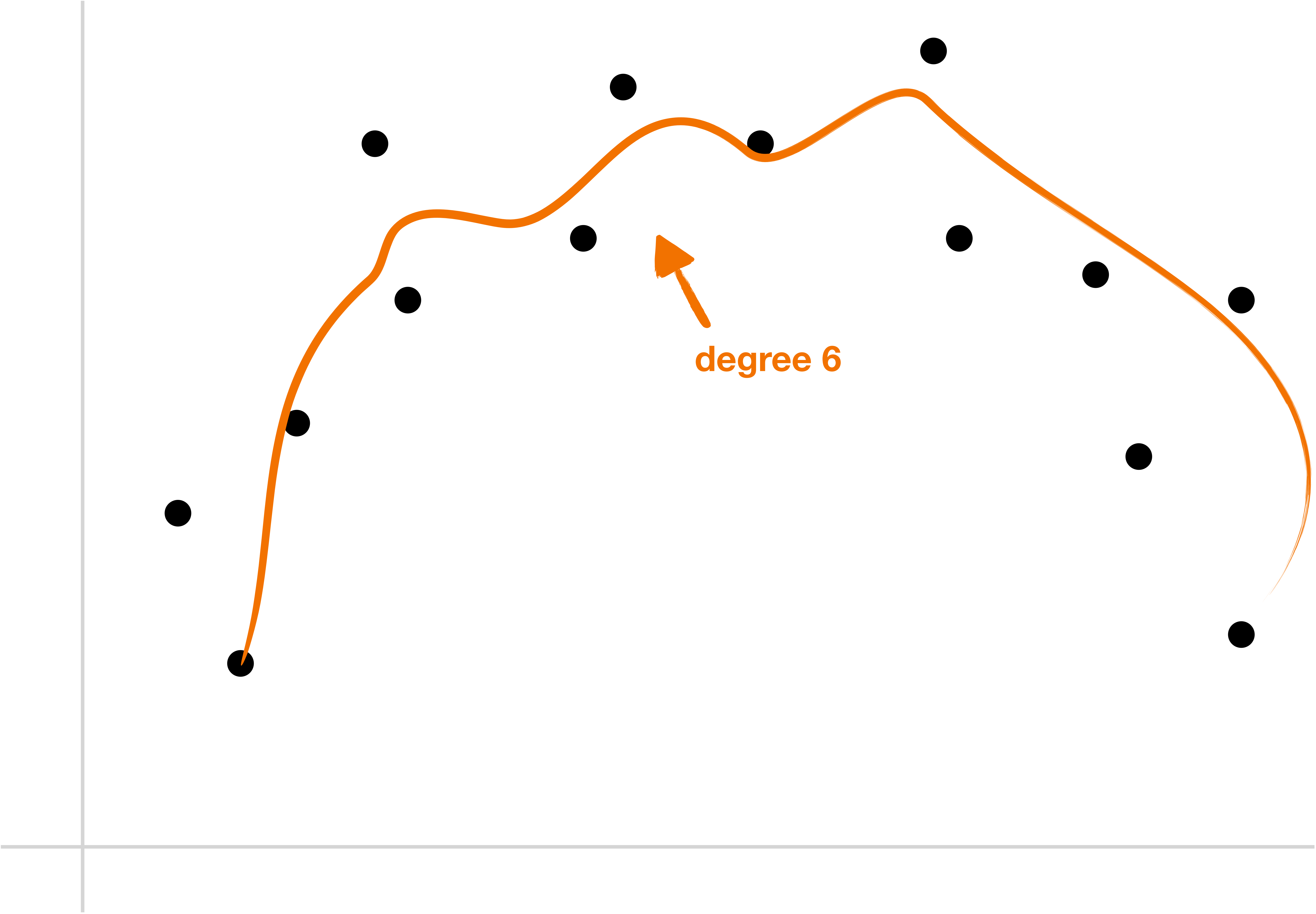


- model is unable to capture relationship between variables

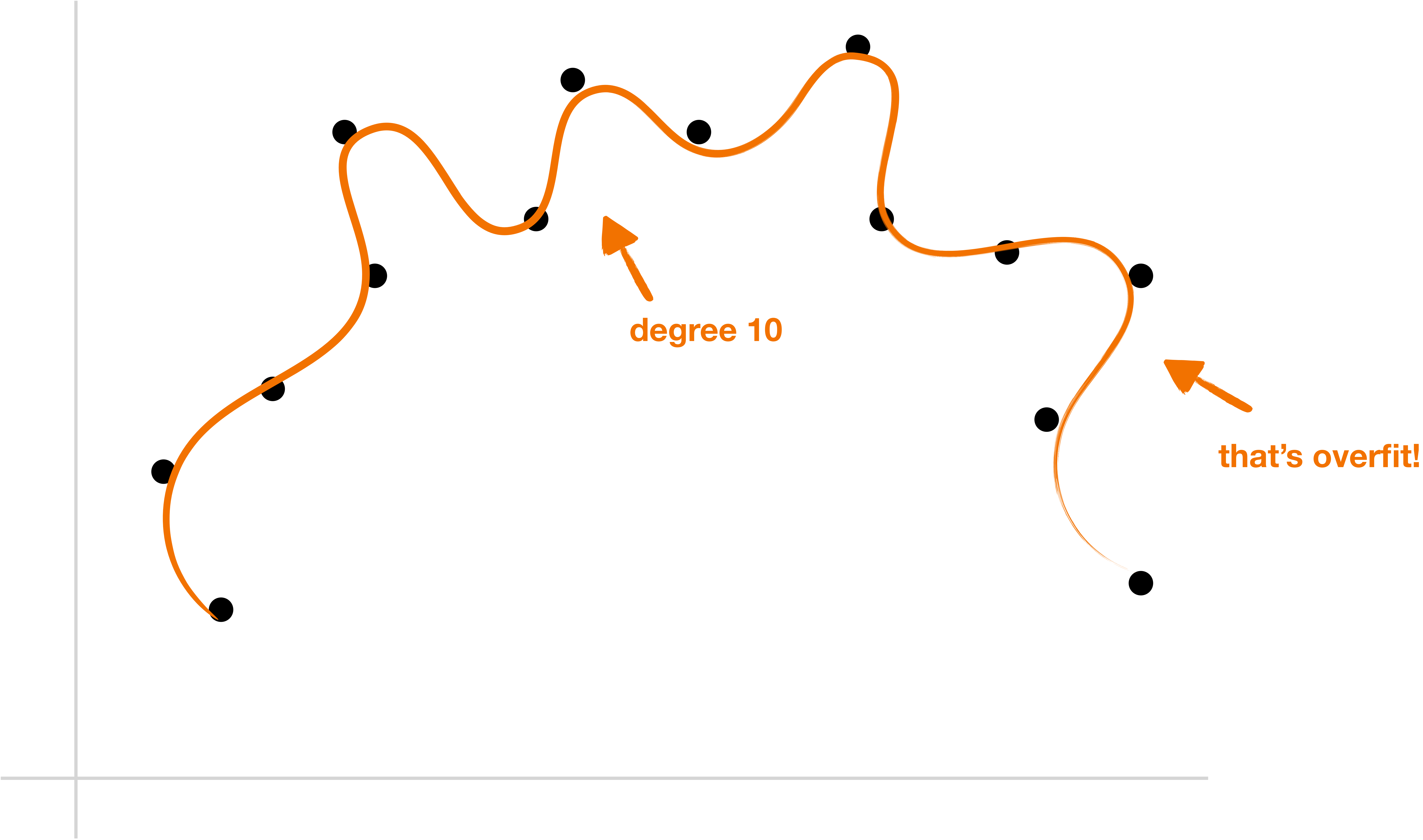
overfitting



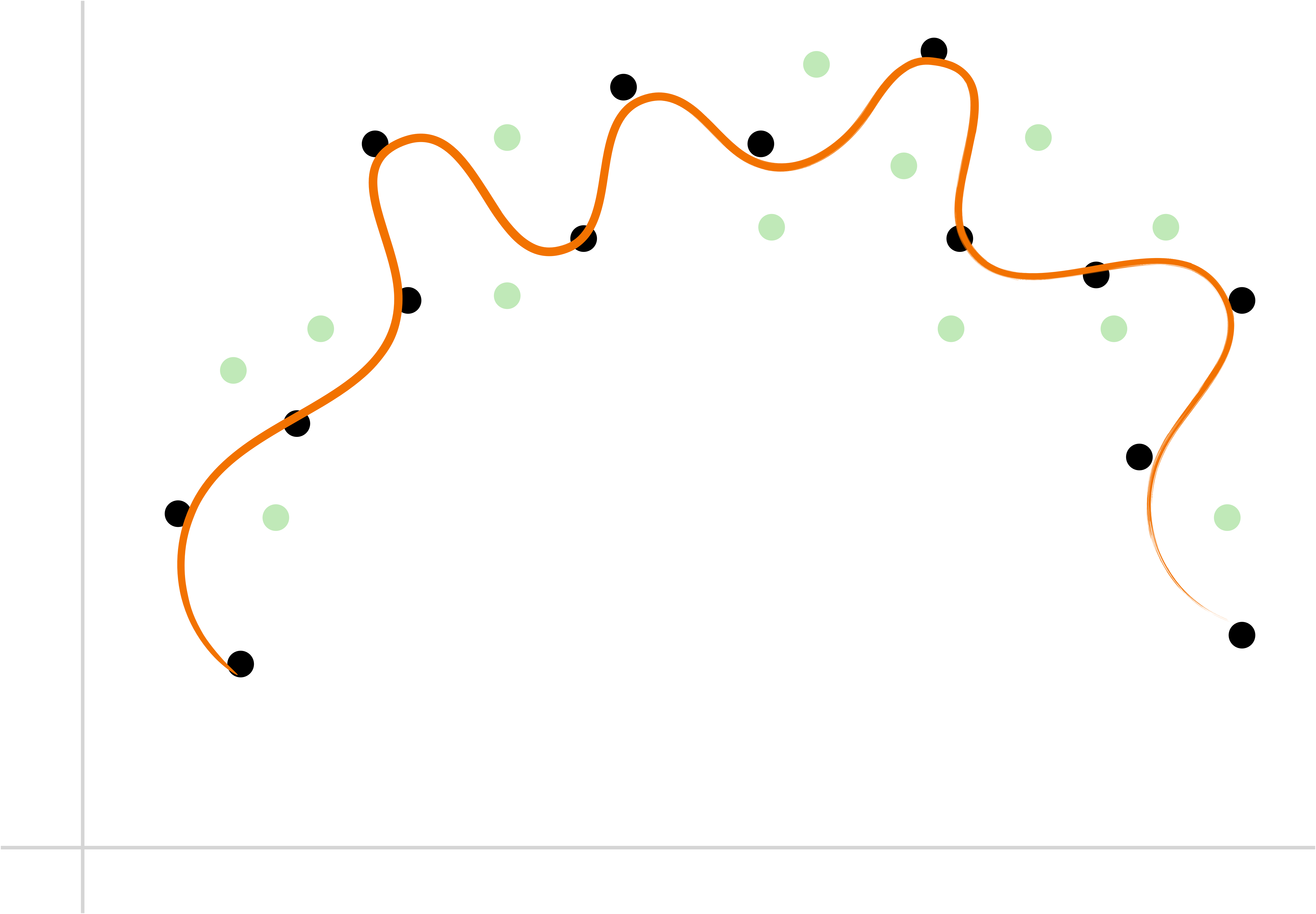
overfitting



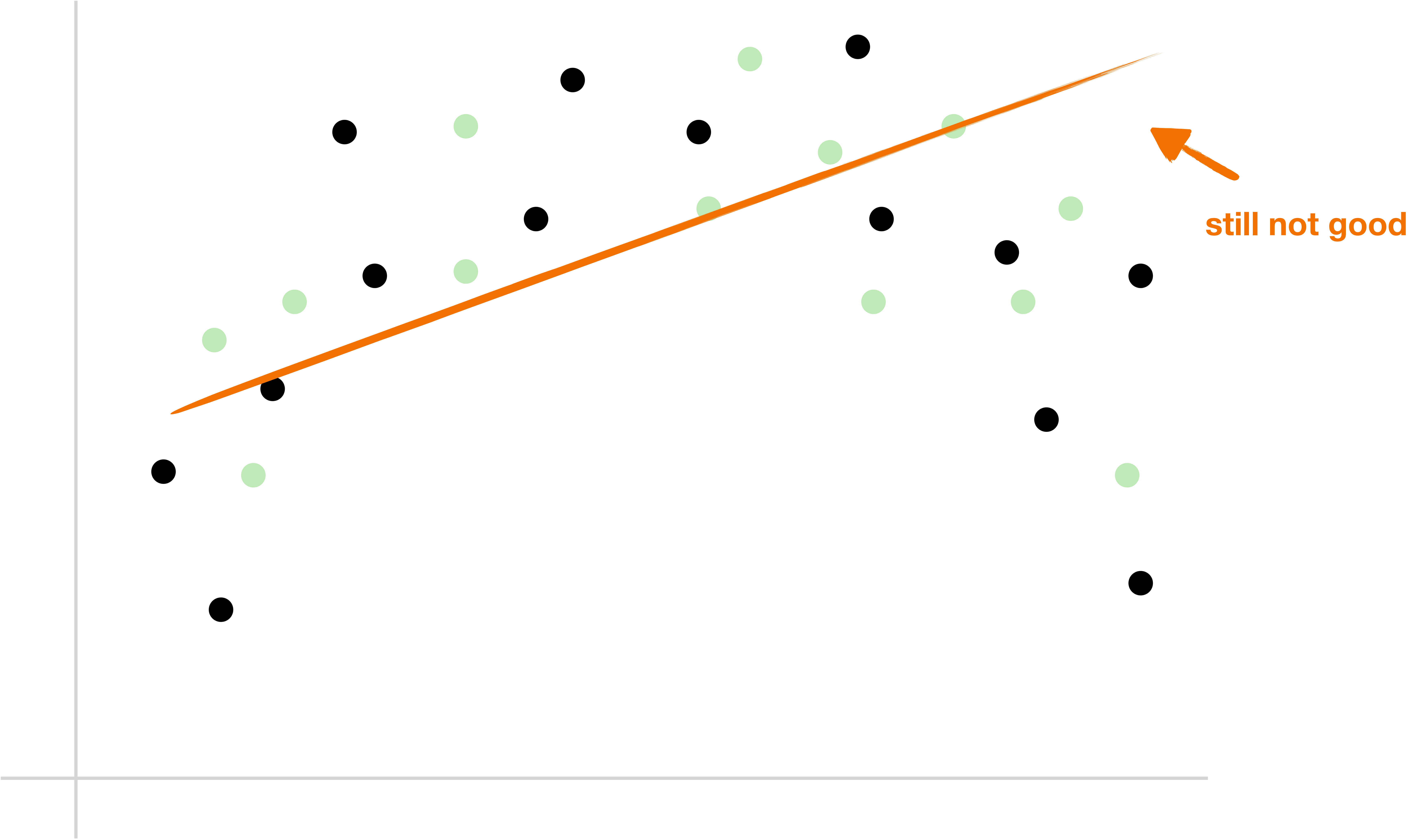
overfitting



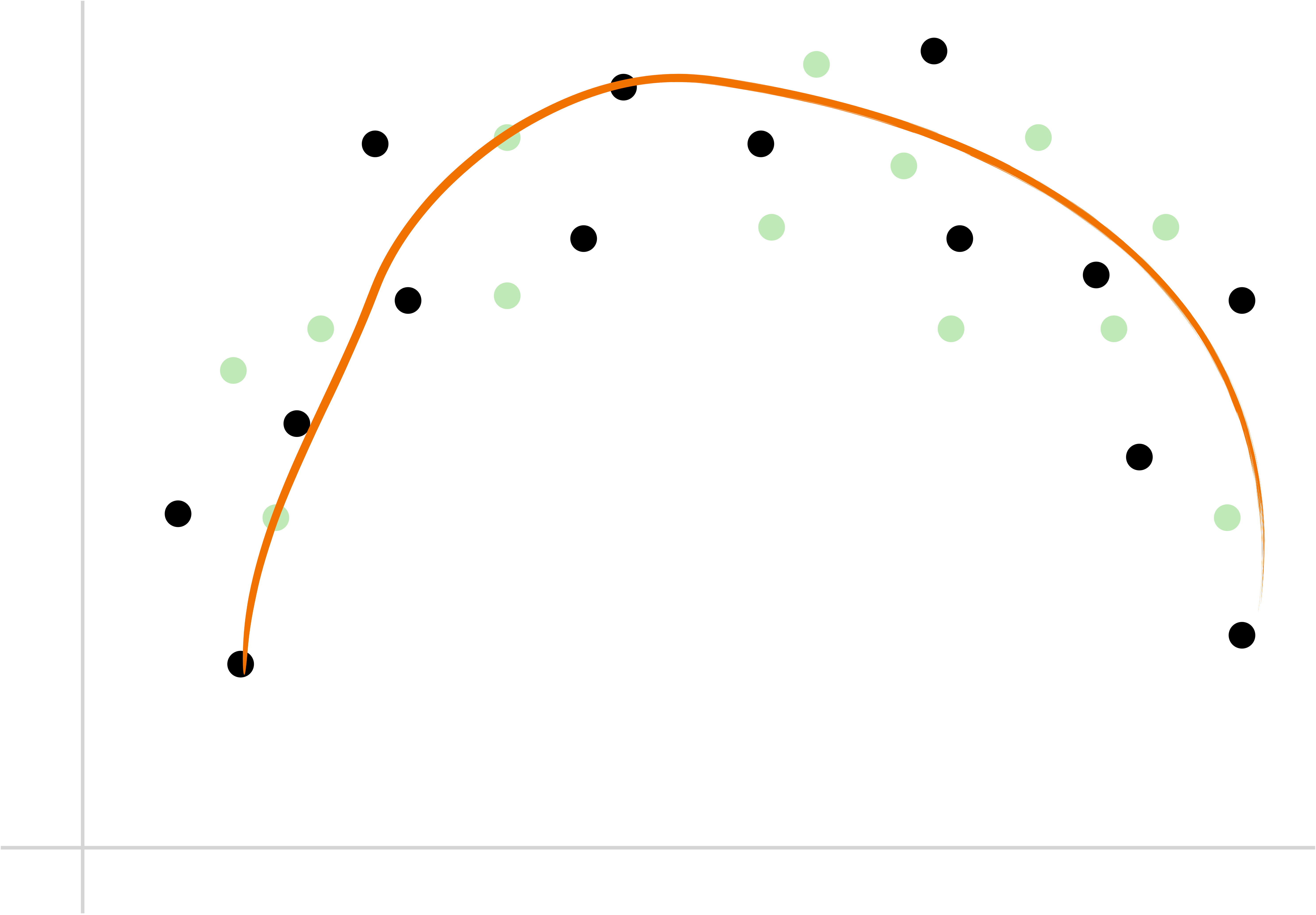
overfitting



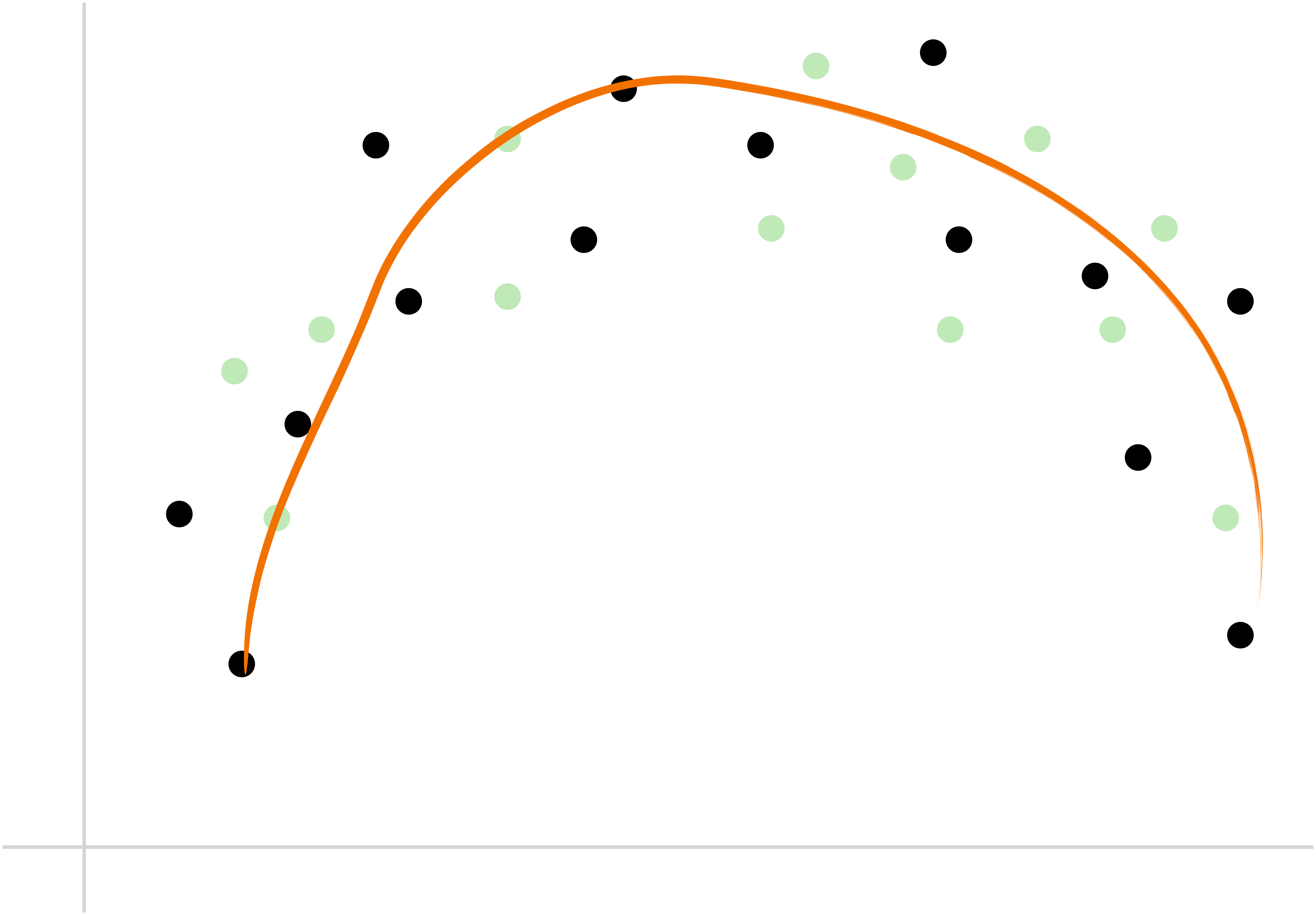
overfitting



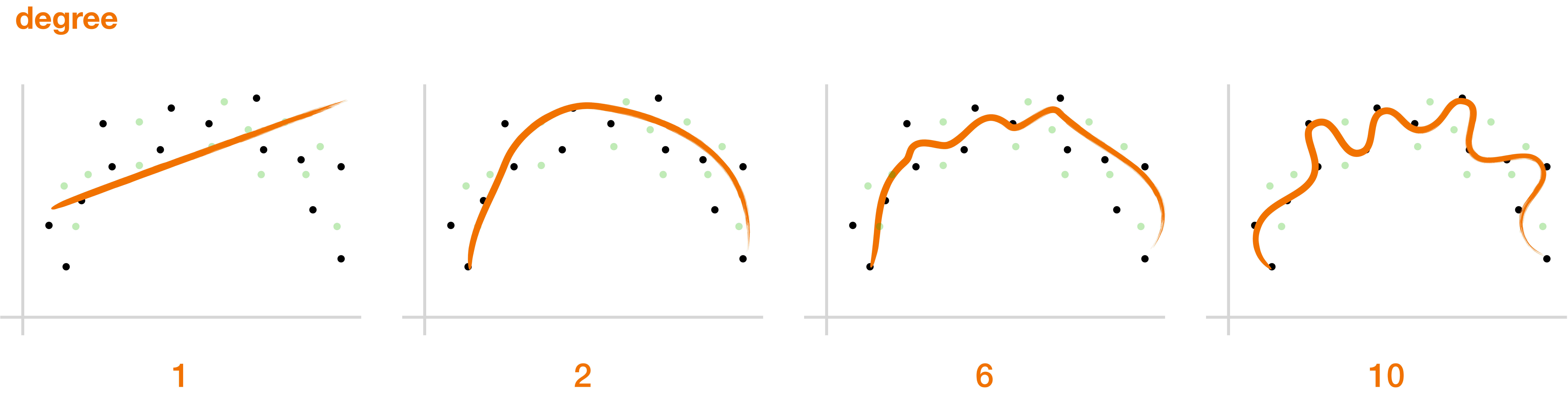
overfitting



overfitting



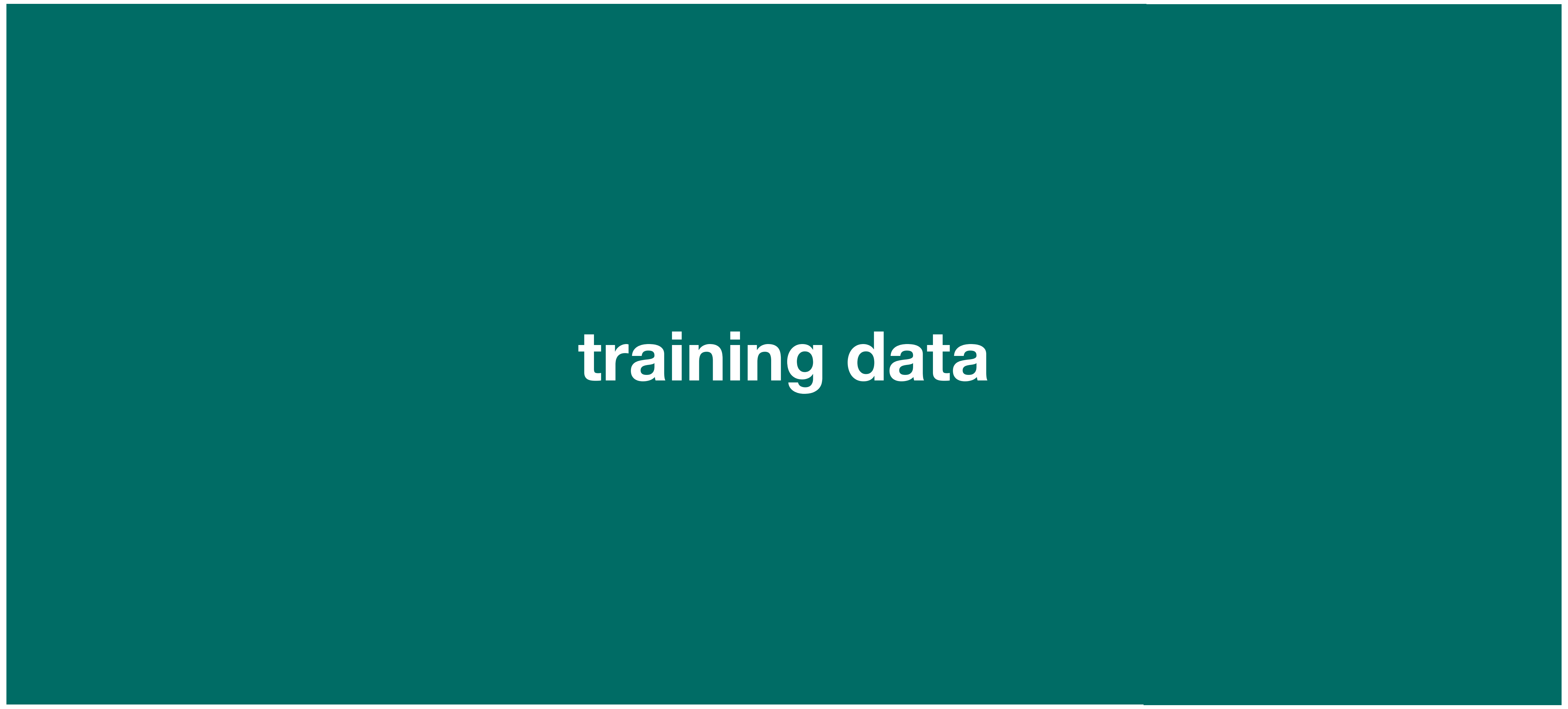
overfitting



- **overfitting** frequently takes place when the degree of a regression model is set too high

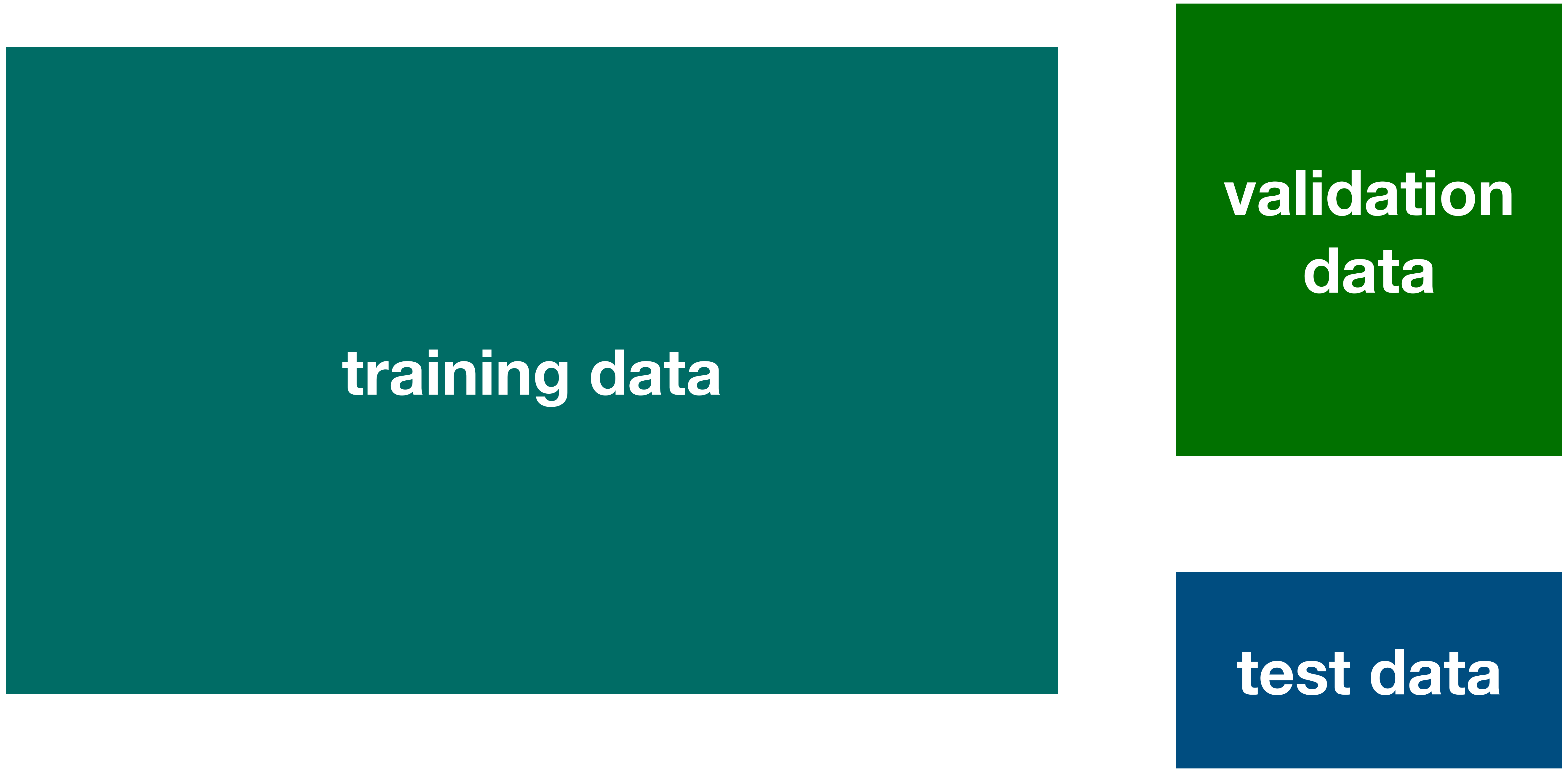
How do we address overfitting?

address overfitting



training data

address overfitting

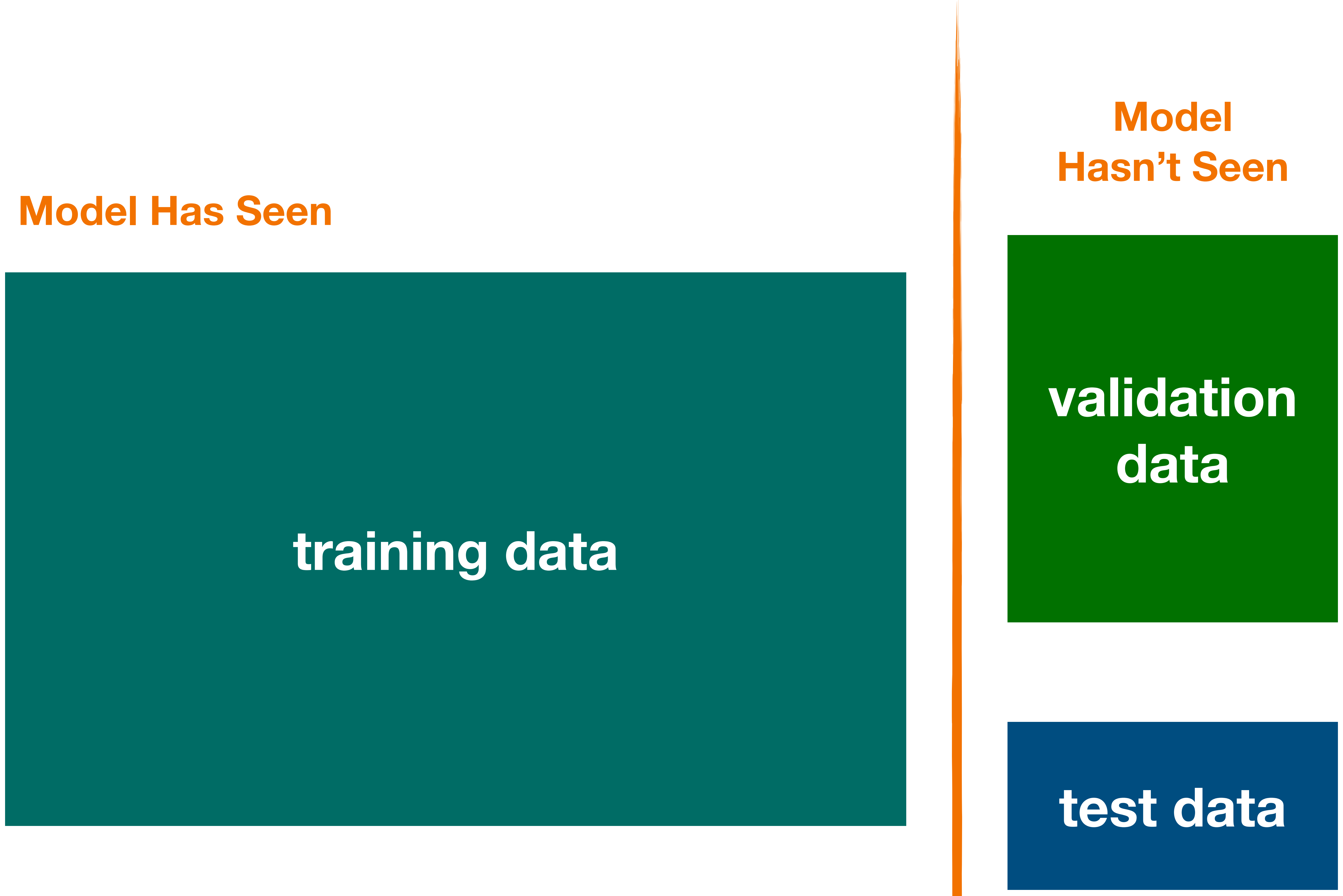


training data

validation data

test data

address overfitting



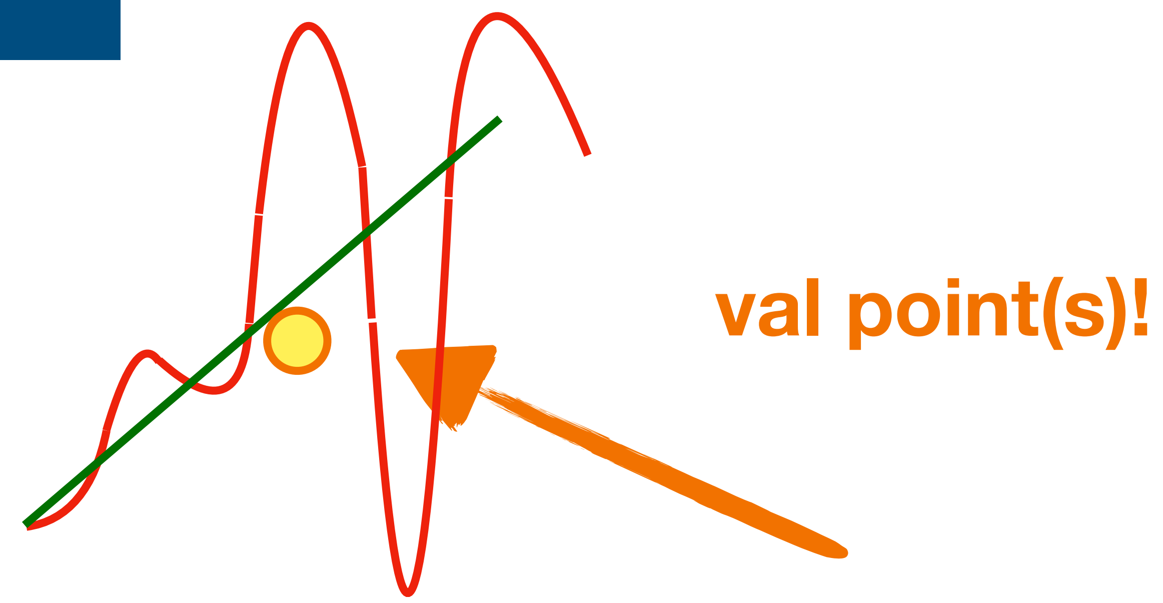
Model Has Seen

training data

Model Hasn’t Seen

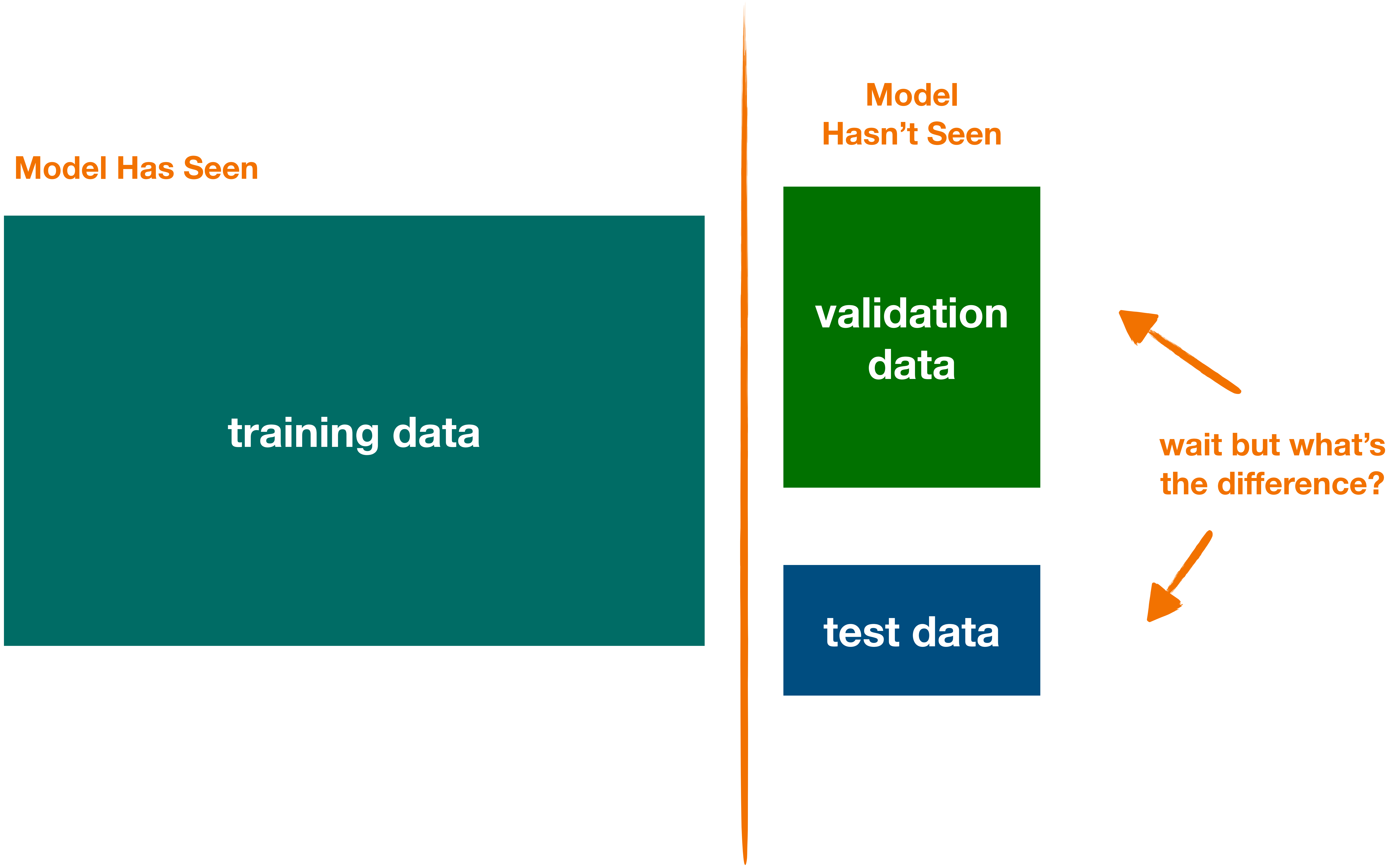
validation data

test data



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

address overfitting



Model Has Seen

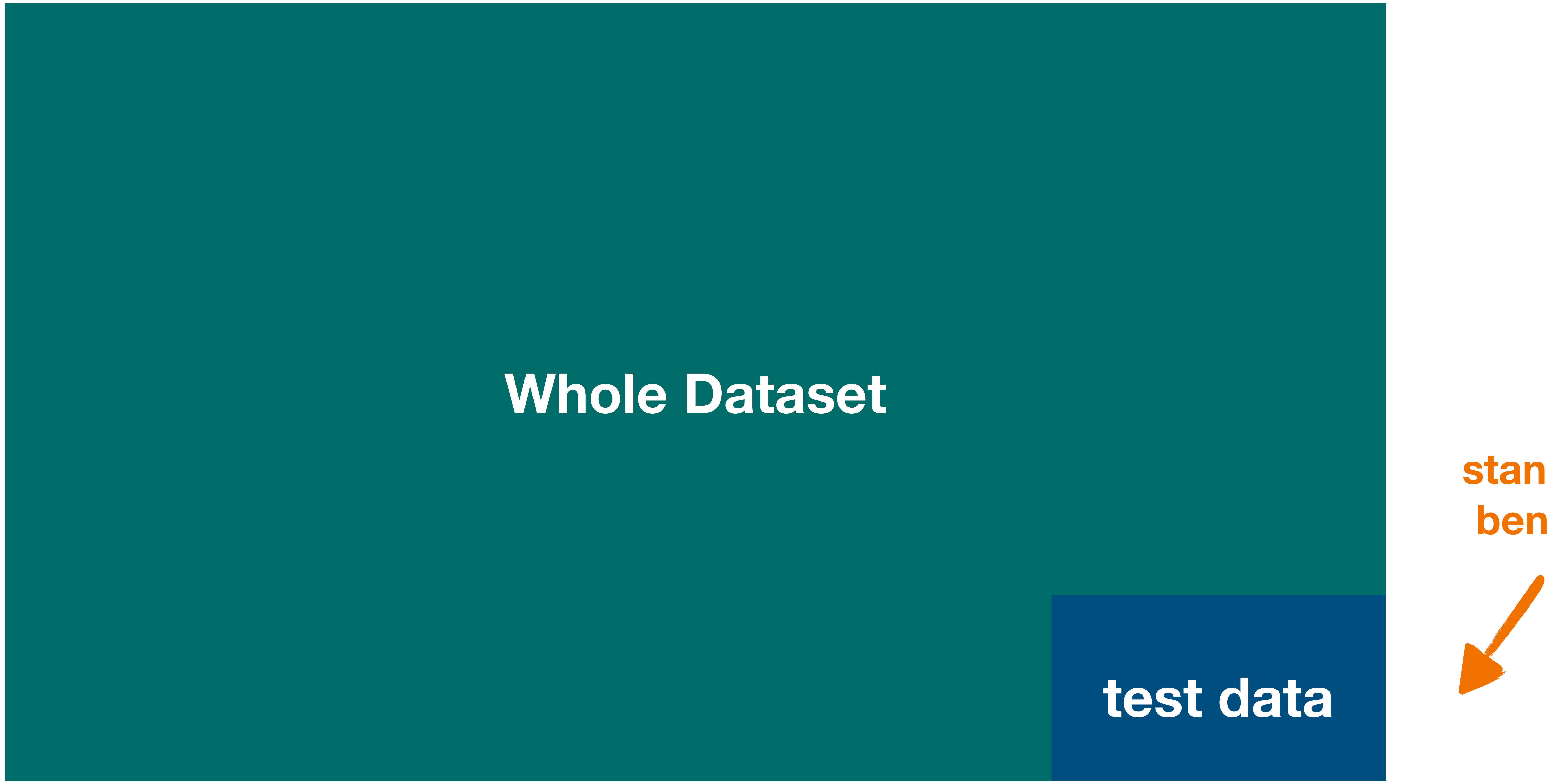
training data

Model Hasn't Seen

validation data

test data

wait but what's the difference?



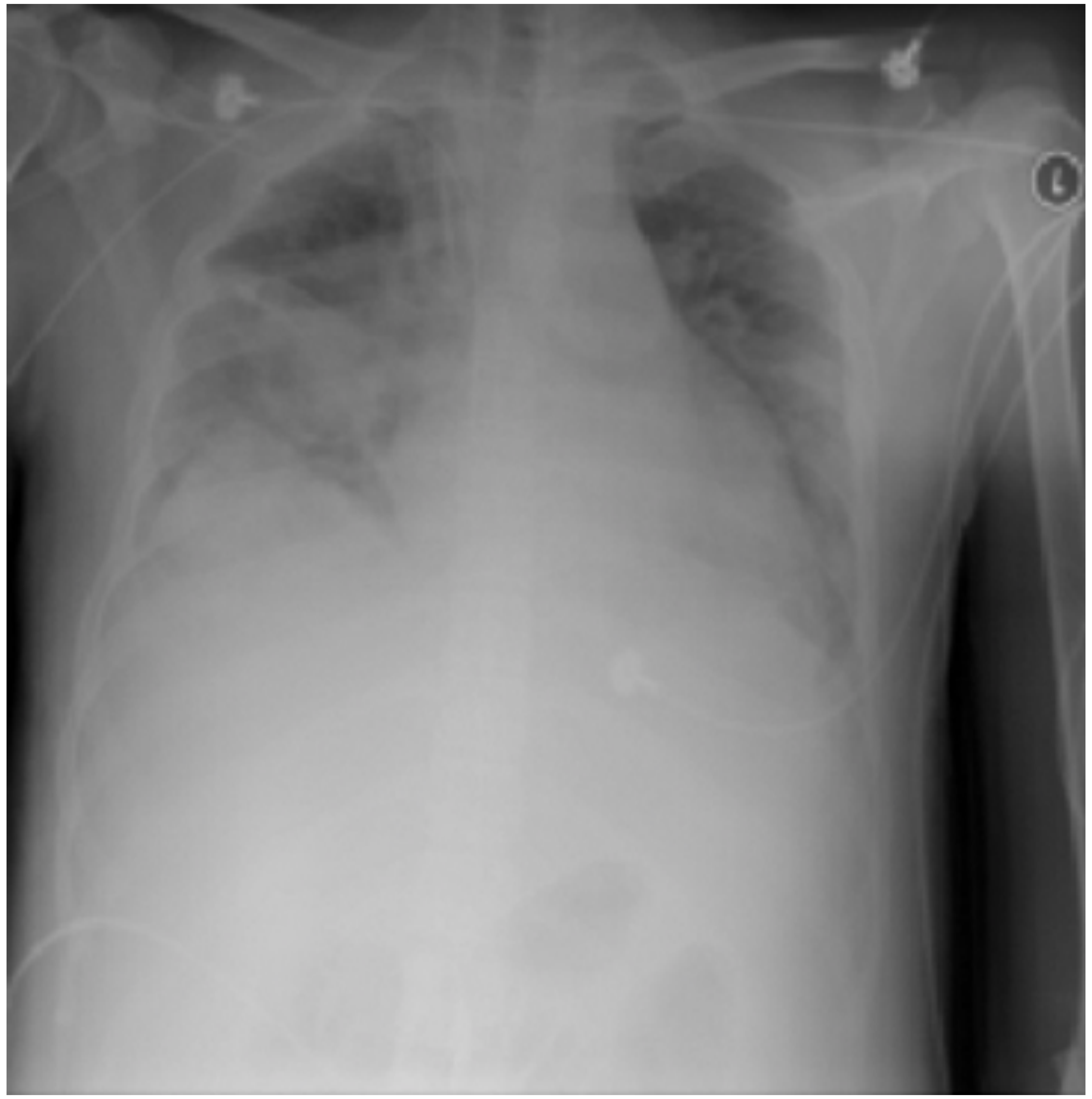
standardized for benchmarking!

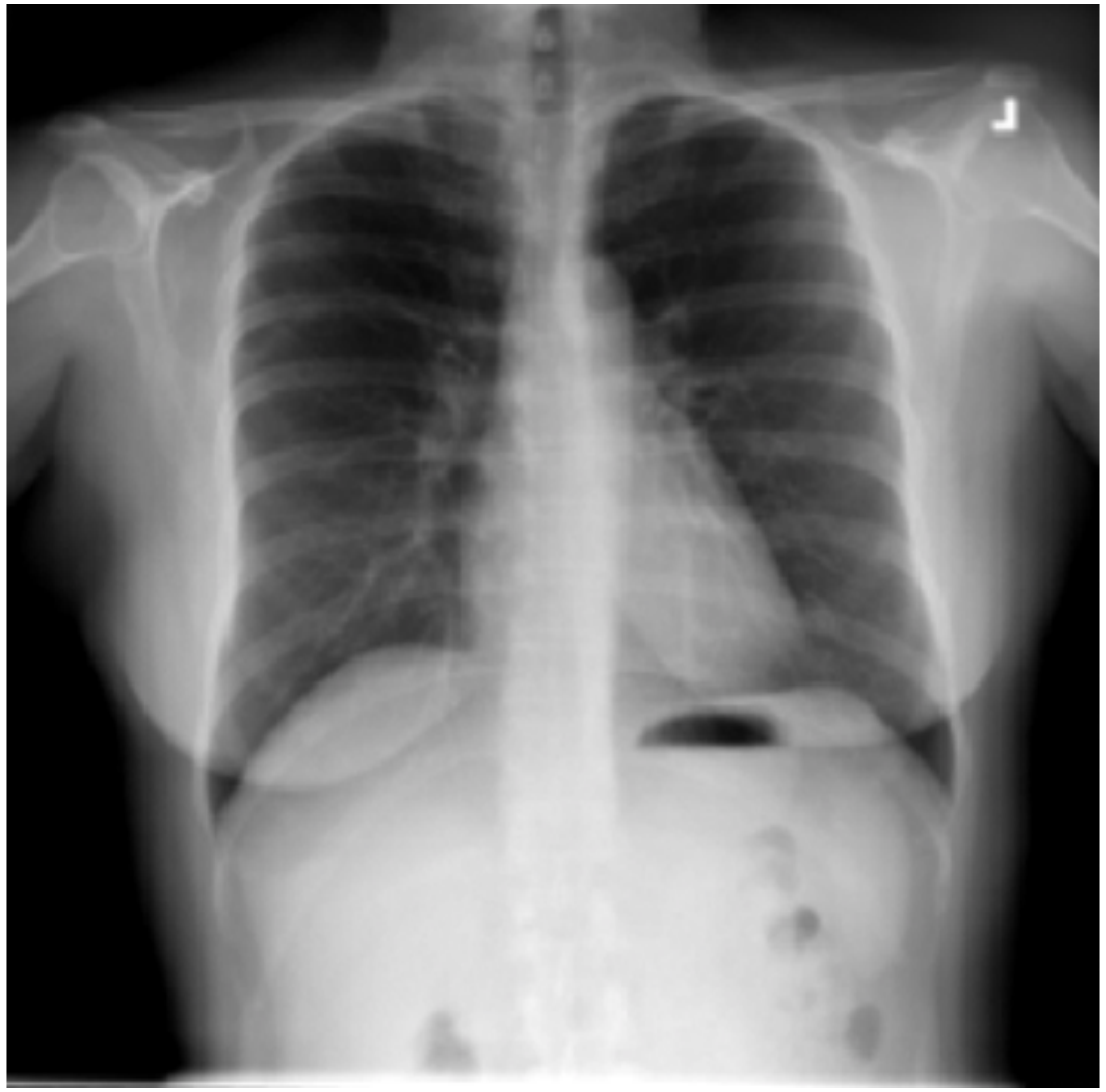
- **test sets** are, unlike validation sets, usually set by the data creator as common, unseen benchmark data.

overfitting can be dangerous

data ethics

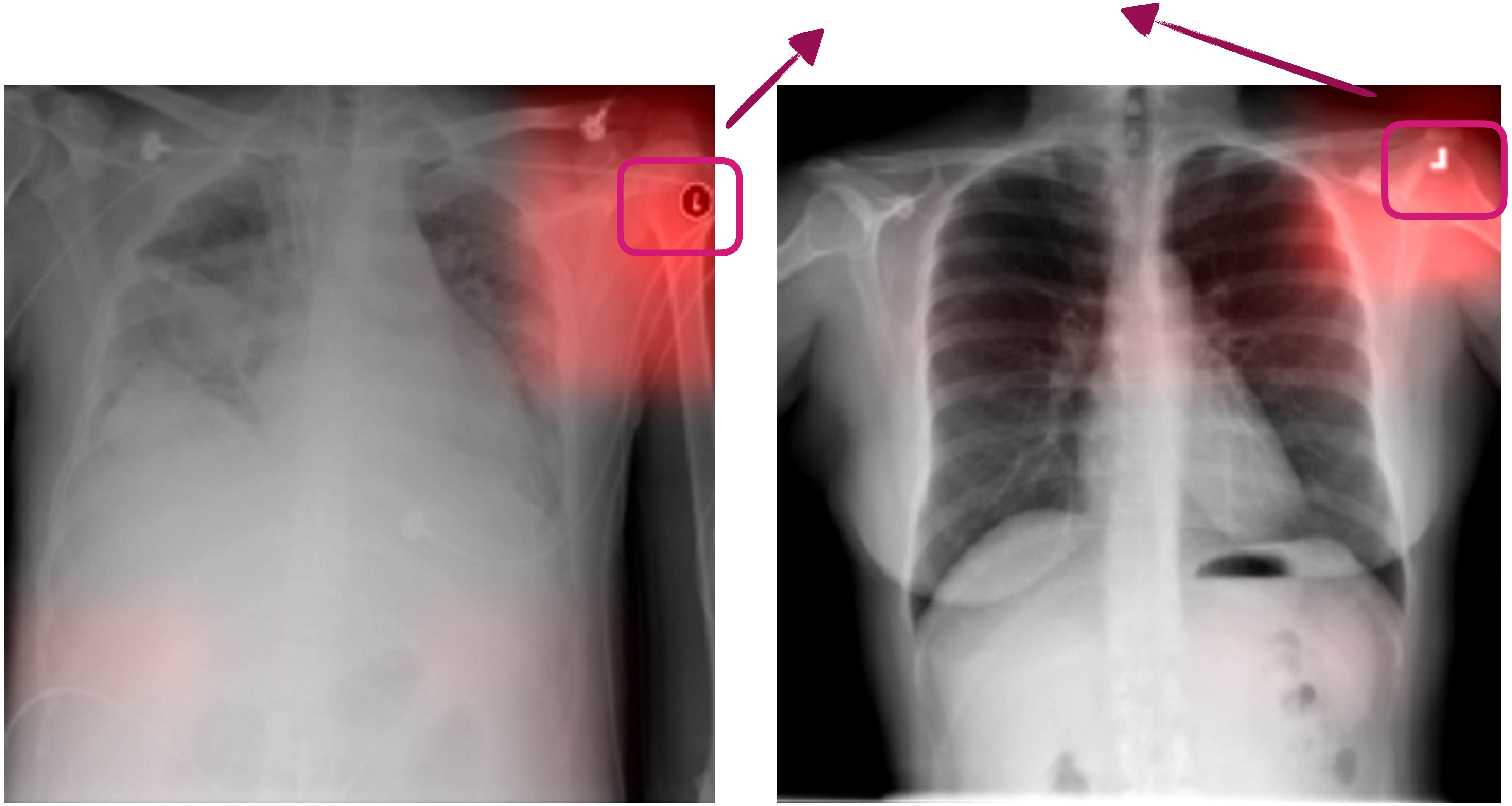
data ethics





which one has pneumonia?

data ethics



different hospitals used different markers!

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

Onto Feature Selection!