Data Graduate Program N°2
Cdiscount Academy

Lundi 20/09 – Mercredi 23/10

SOMMAIRE

- ML – Predictive Modelling Pipeline

MACHINE LEARNING : OBJECTIVES

- build intuitions regarding an unknown dataset
- identify and differentiate numerical and categorical features
- create an advanced predictive pipeline with scikit-learn

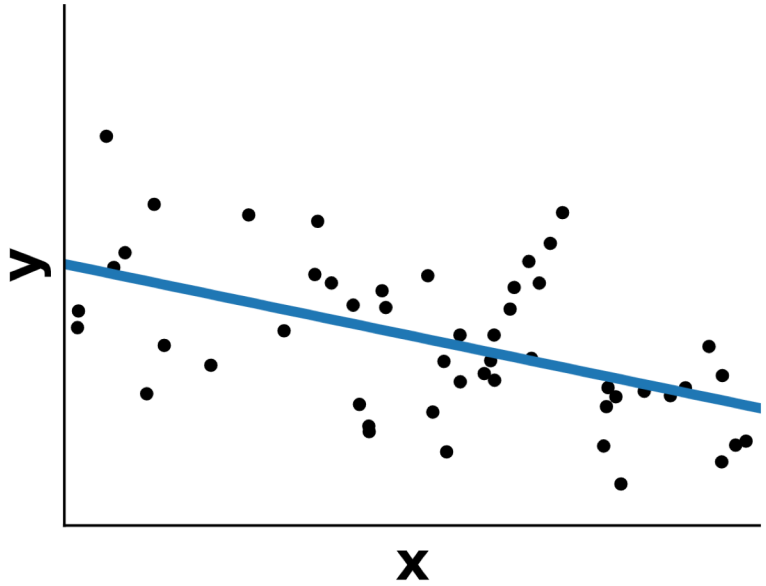
Overfitting and underfitting

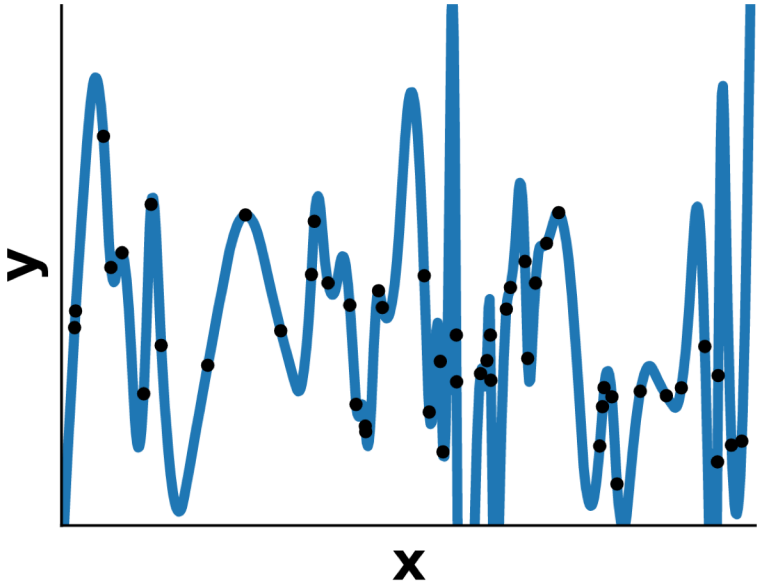
Understanding when and why a model does or does not geenralize well on unseen data

OVERFITTING - UNDERFITTING

Cdiscount

Which data fit do you prefer?



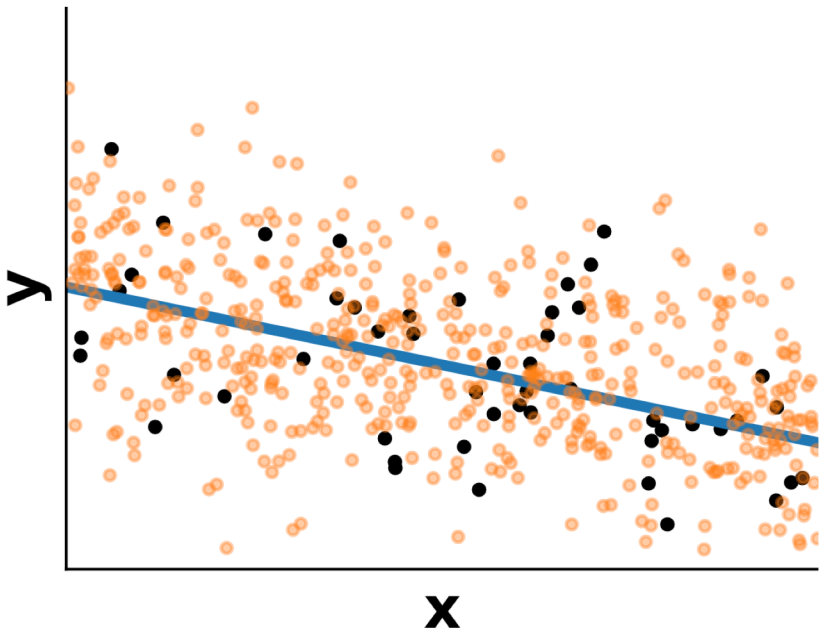


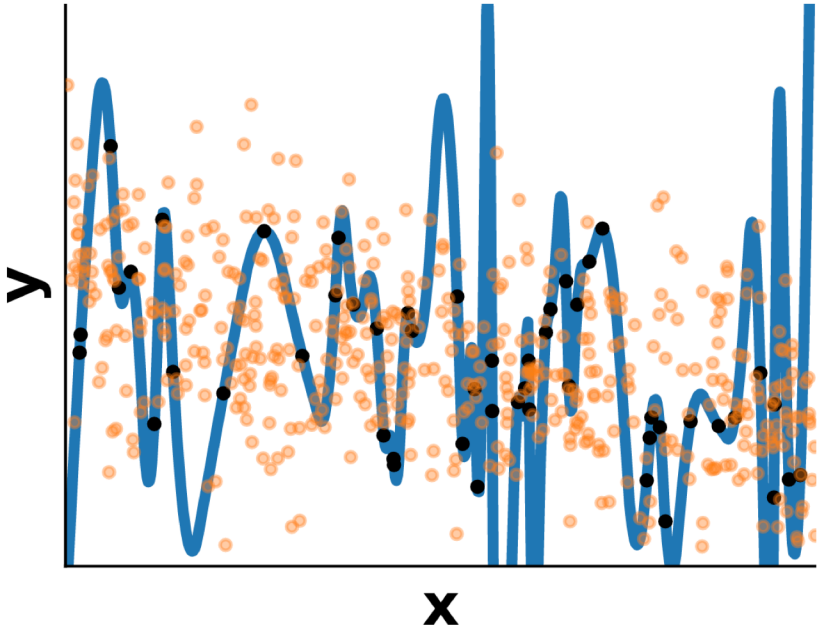
KEYRUS data

OVERFITTING - UNDERFITTING

Cdiscount

Which data fit do you prefer?





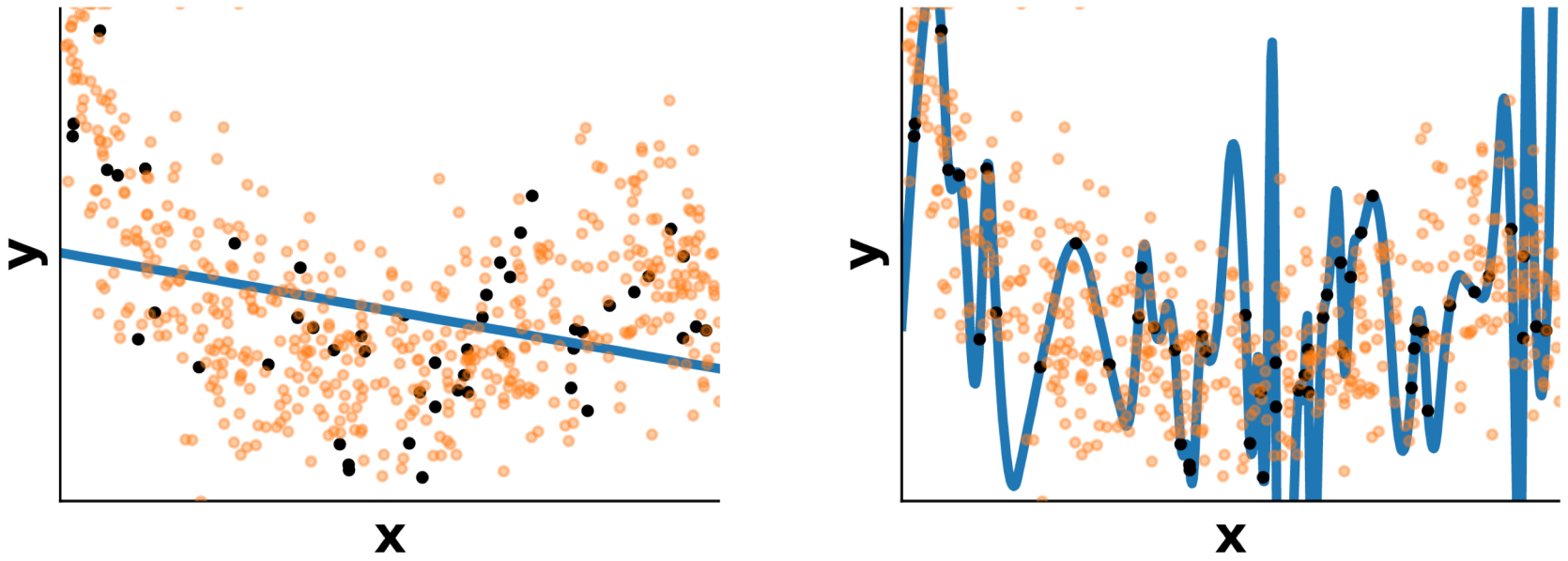
On new data

KEYRUS data

OVERFITTING - UNDERFITTING

Cdiscount

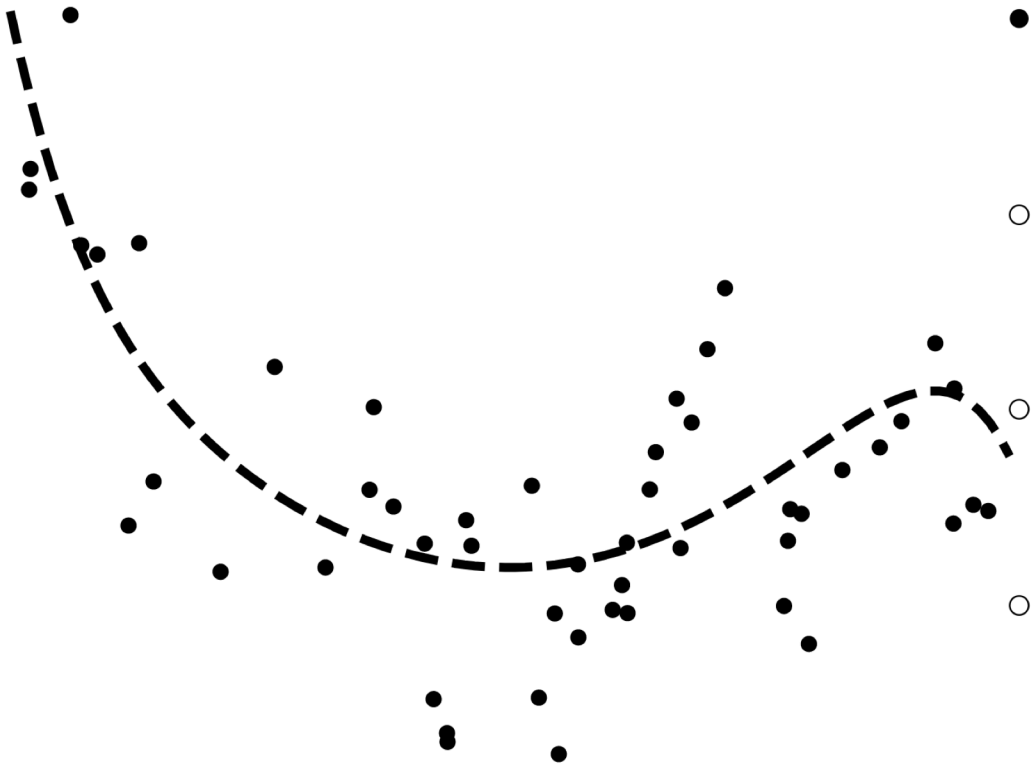
Which data fit do you prefer?



A harder example

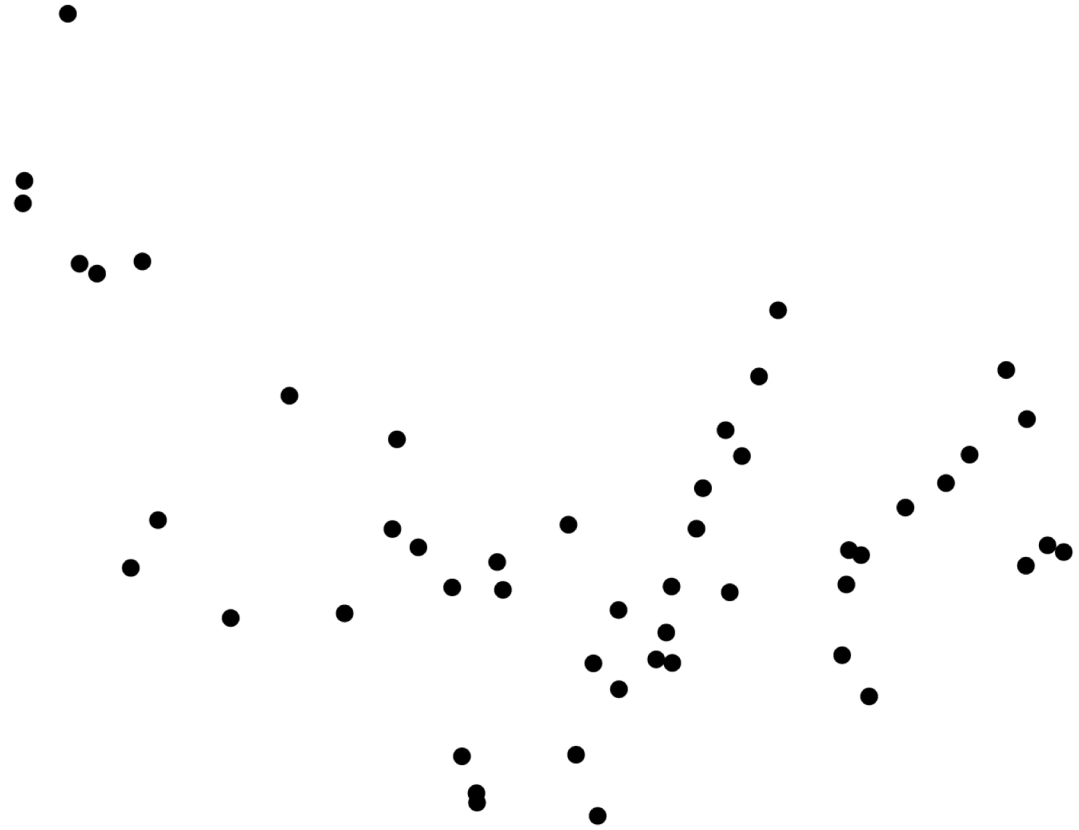
KEYRUS data

VARYING MODEL COMPLEXITY



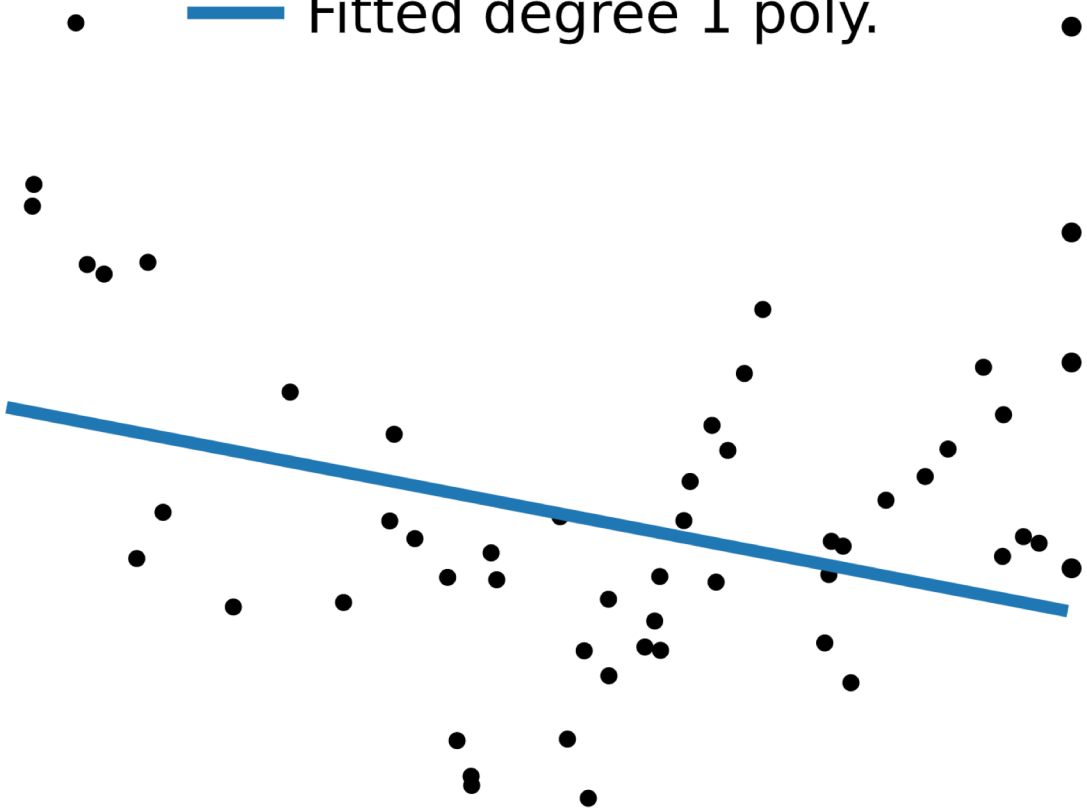
- Data generated by a random process
  - Sample a value of the input *x*
  - Transform it with a 9th-degree polynomial
  - Add some noise to get the output *y*

VARYING MODEL COMPLEXITY



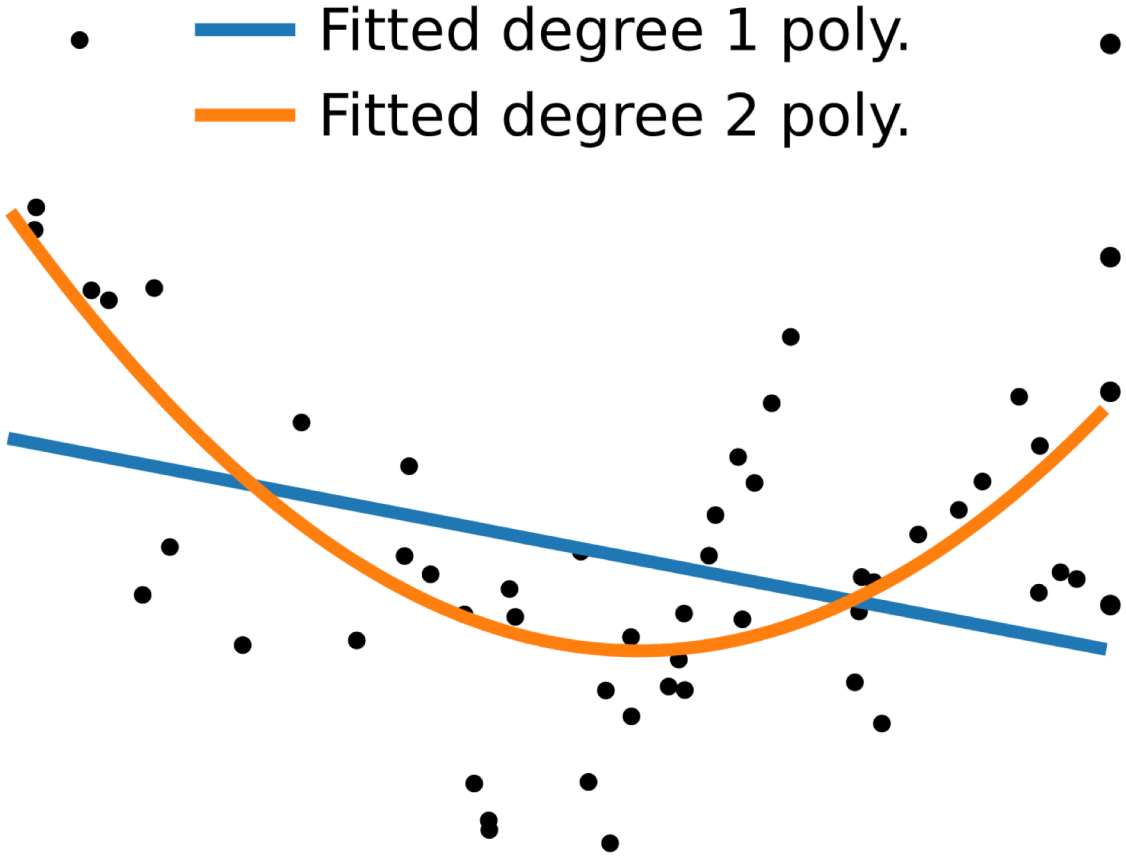
- Data generated by a random process
- This process is unknown
- We can only access the observations

VARYING MODEL COMPLEXITY



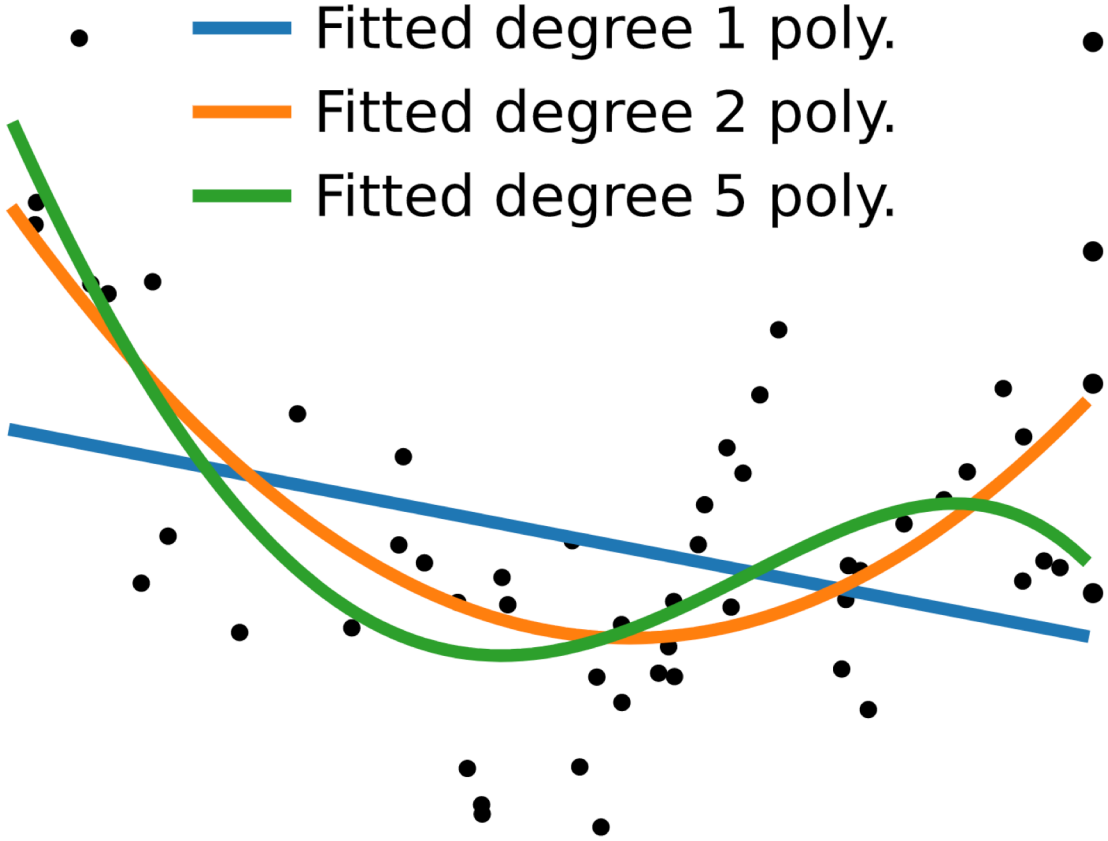
- Data generated by a random process
- This process is unknown
- We can only access the observations
- Fit polynomials of various degrees

VARYING MODEL COMPLEXITY



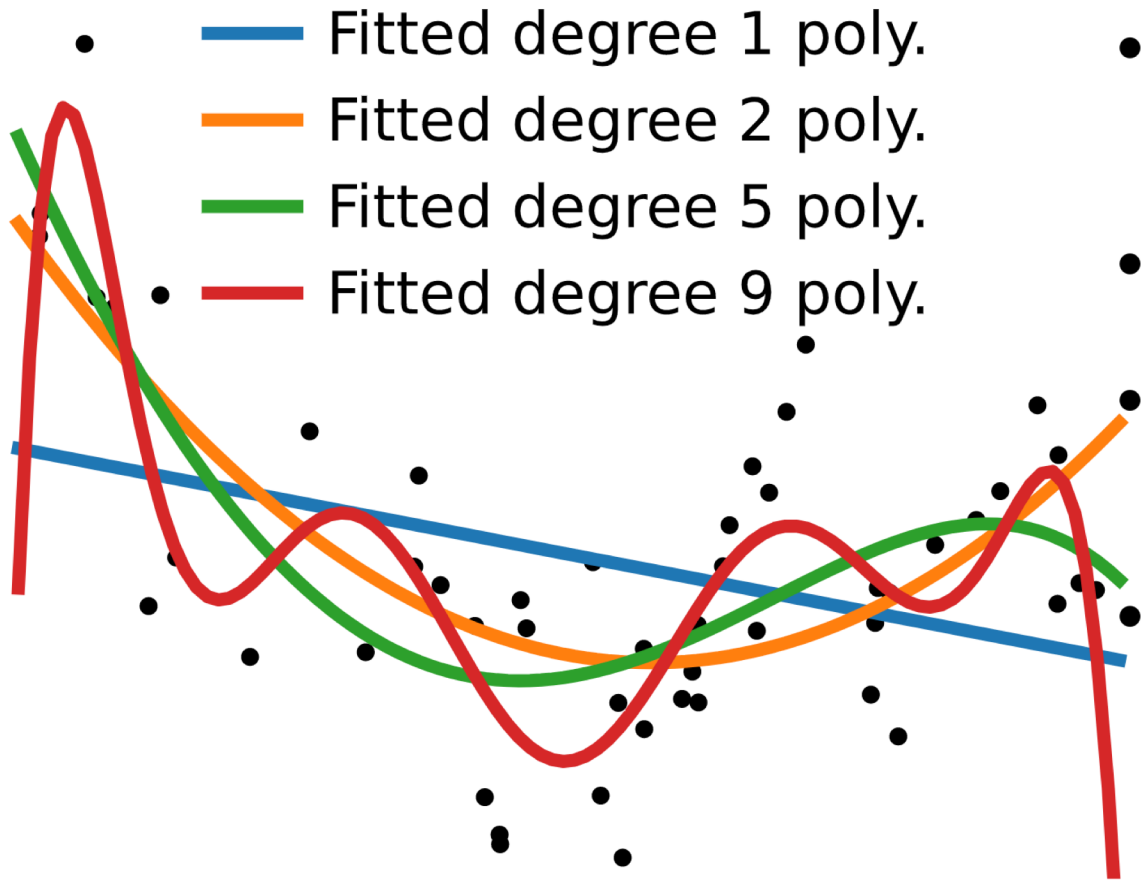
- Data generated by a random process
- This process is unknown
- We can only access the observations
- Fit polynomials of various degrees

VARYING MODEL COMPLEXITY



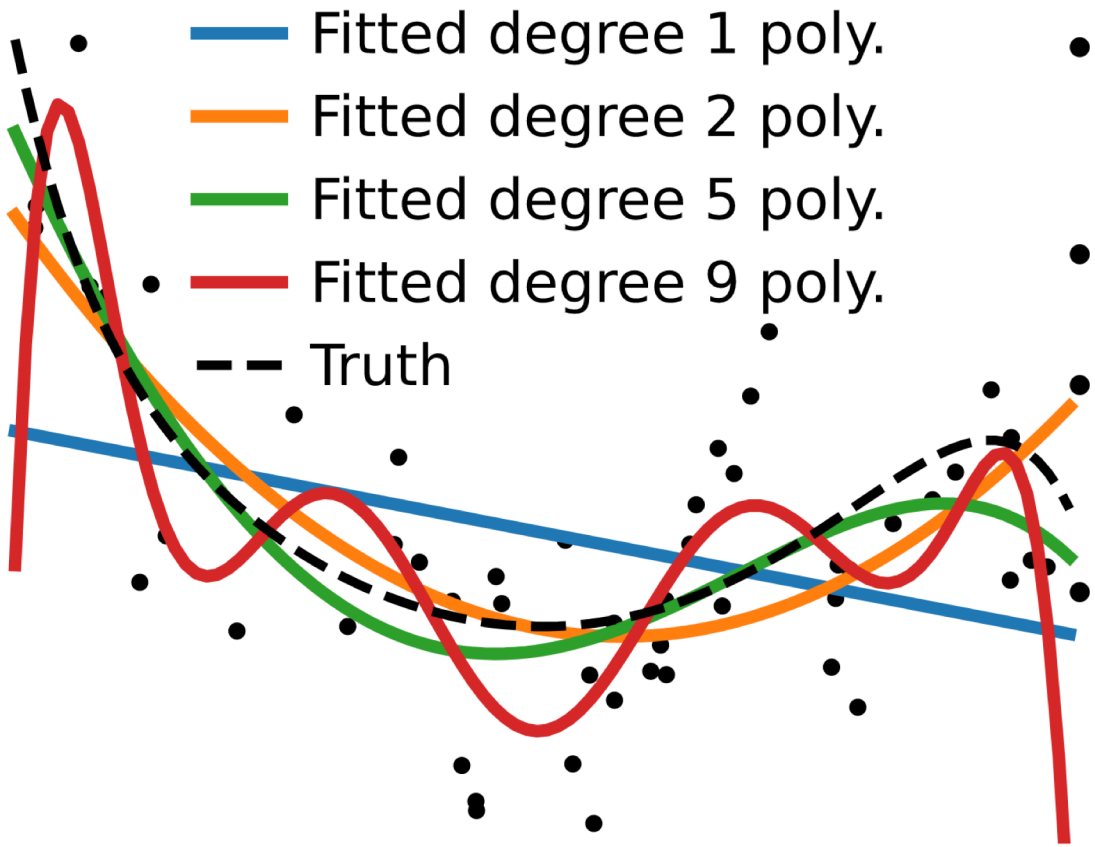
- Data generated by a random process
- This process is unknown
- We can only access the observations
- Fit polynomials of various degrees

VARYING MODEL COMPLEXITY



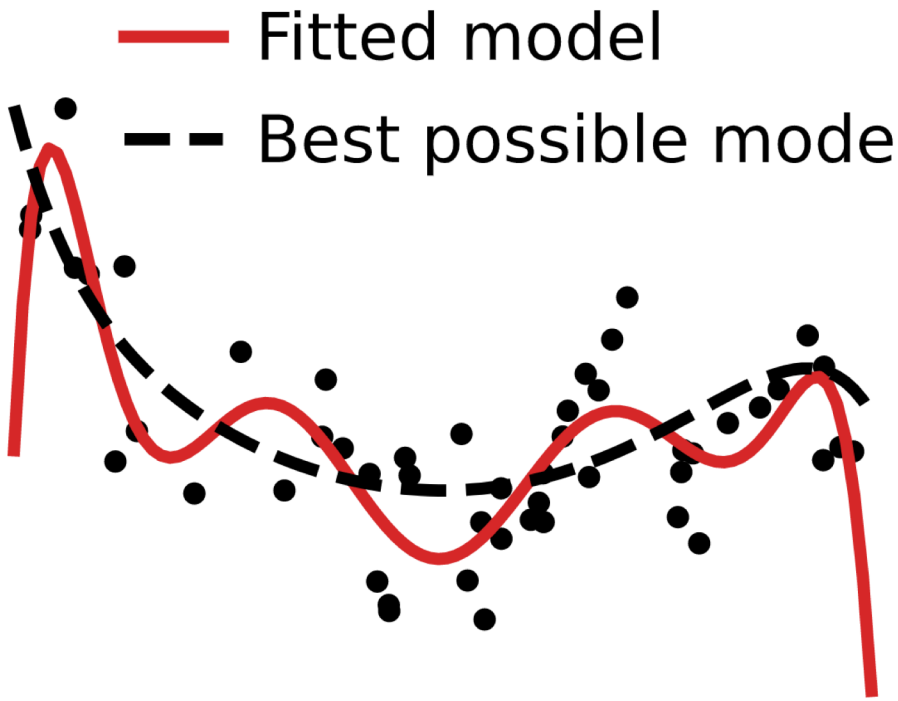
- Data generated by a random process
- This process is unknown
- We can only access the observations
- Fit polynomials of various degrees

VARYING MODEL COMPLEXITY



- Data generated by a random process
- This process is unknown
- We can only access the observations
- Fit polynomials of various degrees

OVERFIT: MODEL TOO COMPLEX

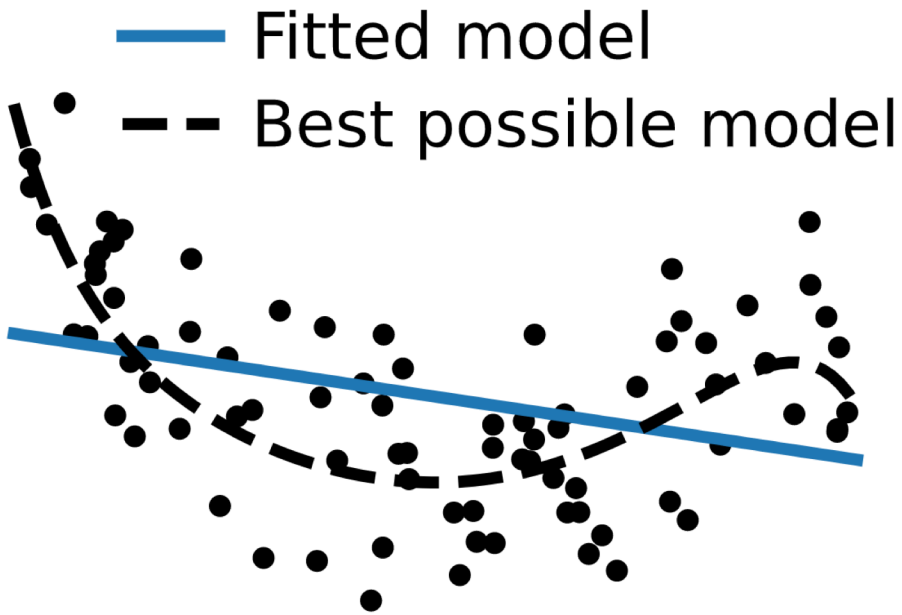


Model too complex for the data:

- Its best possible fit would approximate well the generative process
- However, its flexibility captures noise

Not enough data **Too much noise**

UNDERFIT: MODEL TOO SIMPLE



Plenty of data Low noise

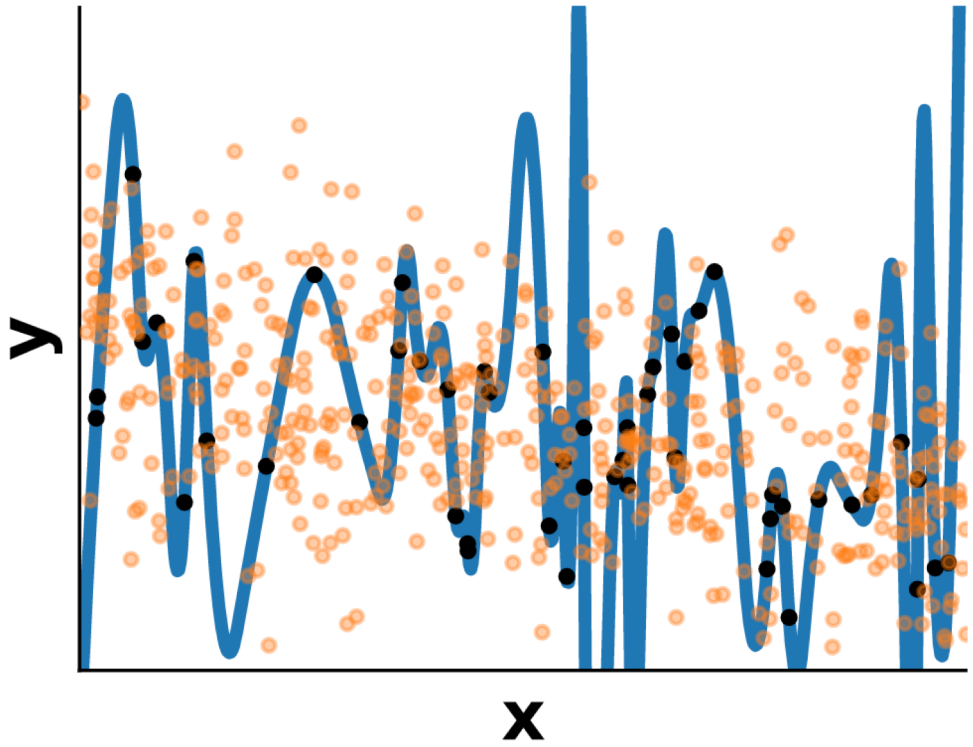
Model too simple for the data:

- Its best fit does not approximate well the generative process
- Yet it captures little noise

Comparing train and test erros

- Varying complexity: validation curves
- Varying the sample size: learning curves
- Goal: understand the overfitting / underfitting trade-off

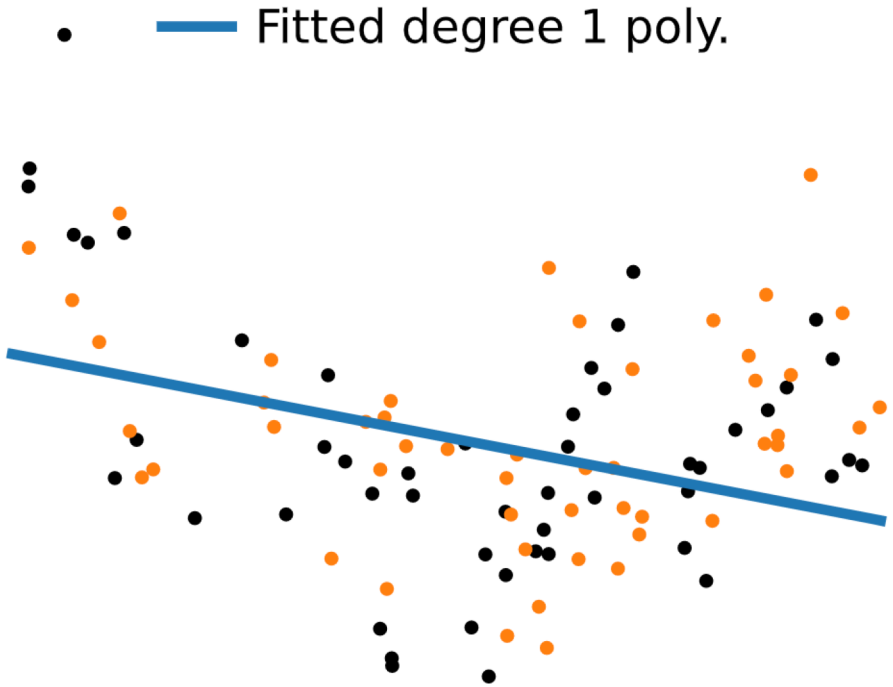
TRAIN V/S TEST ERROR

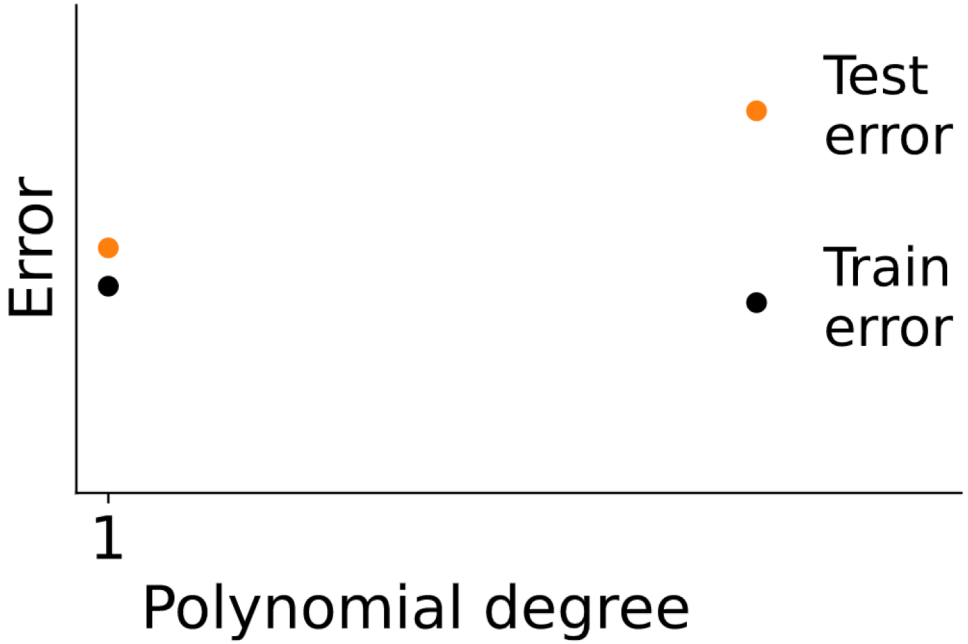


Measure:

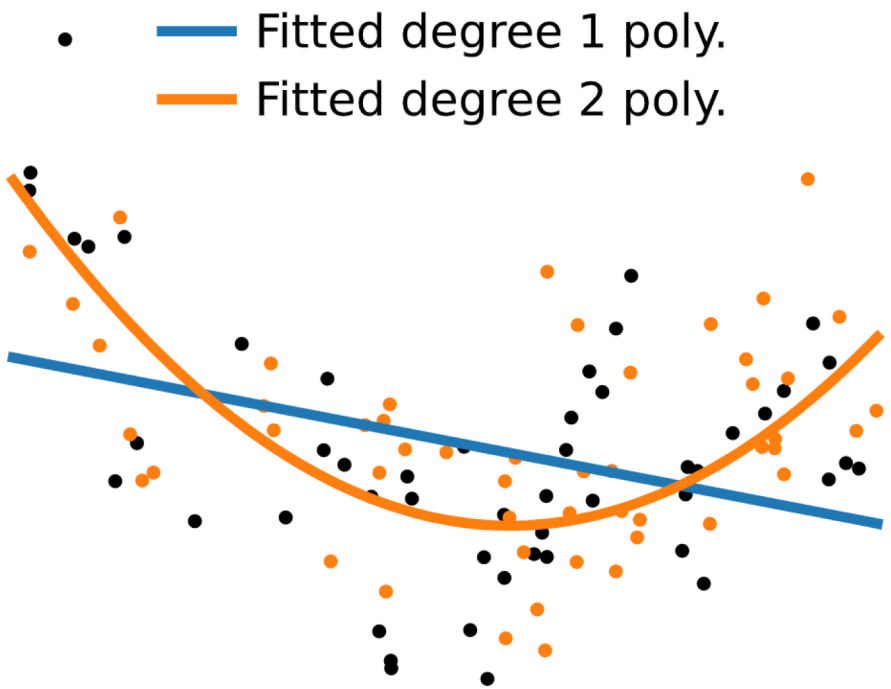
- errors on test data (generalization)
- errors on the train data

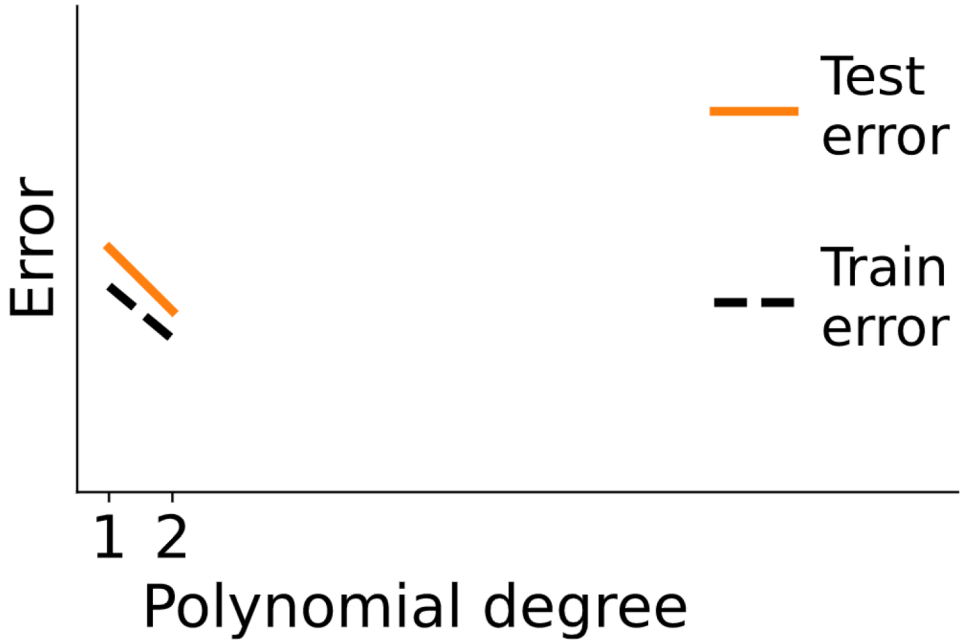
TRAIN V/S TEST ERROR : INCREASING COMPLEXITY



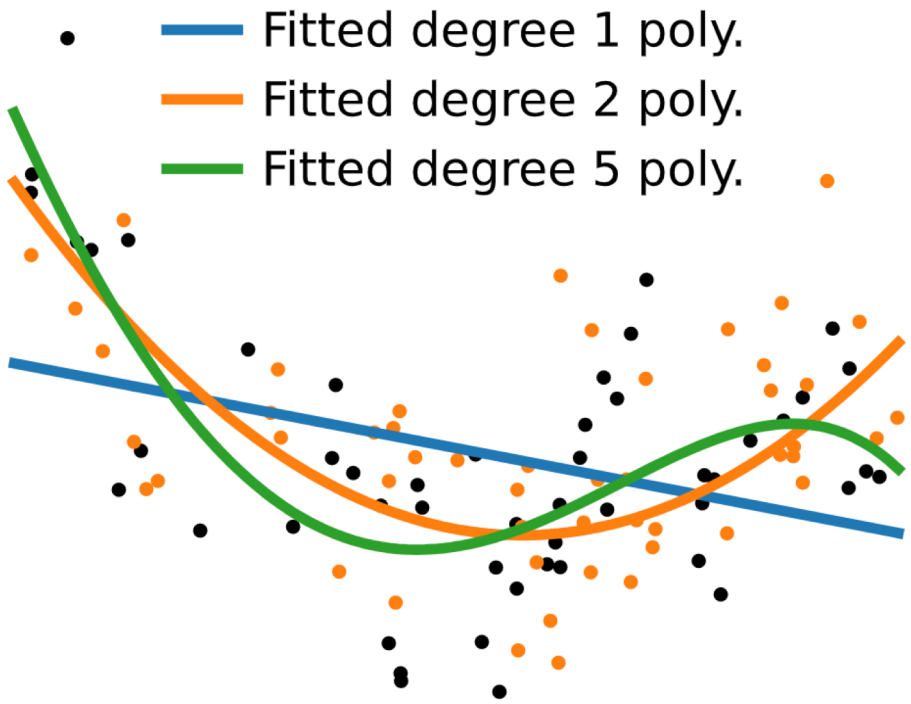


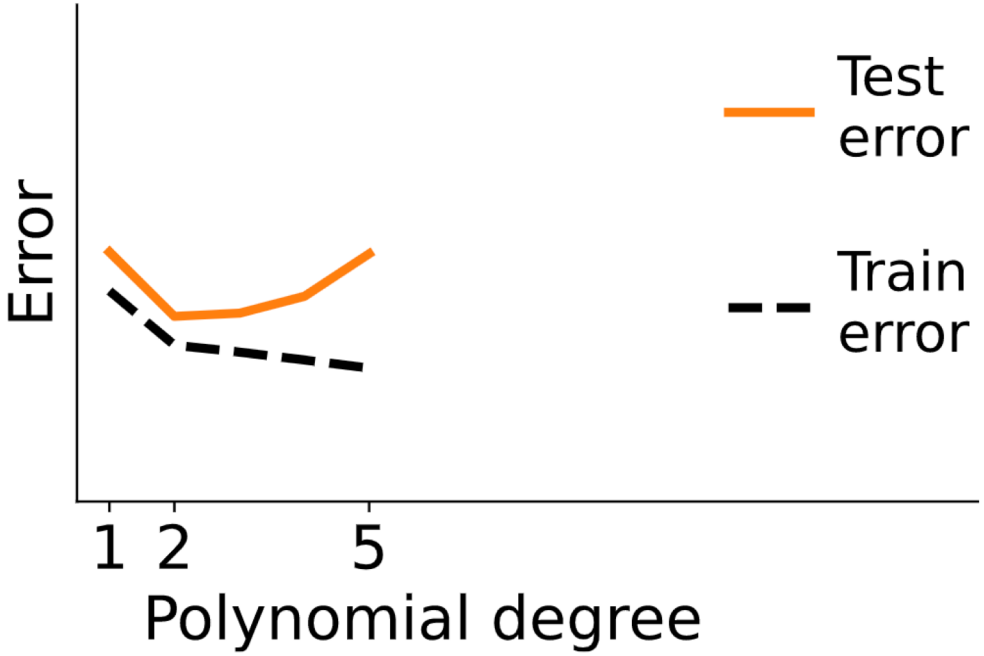
TRAIN V/S TEST ERROR : INCREASING COMPLEXITY



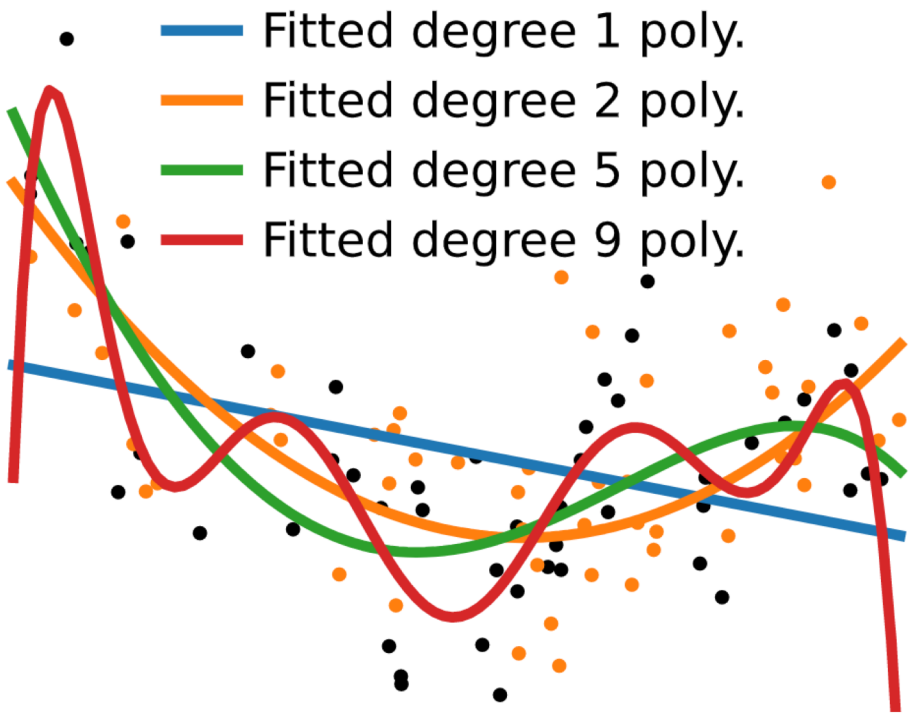


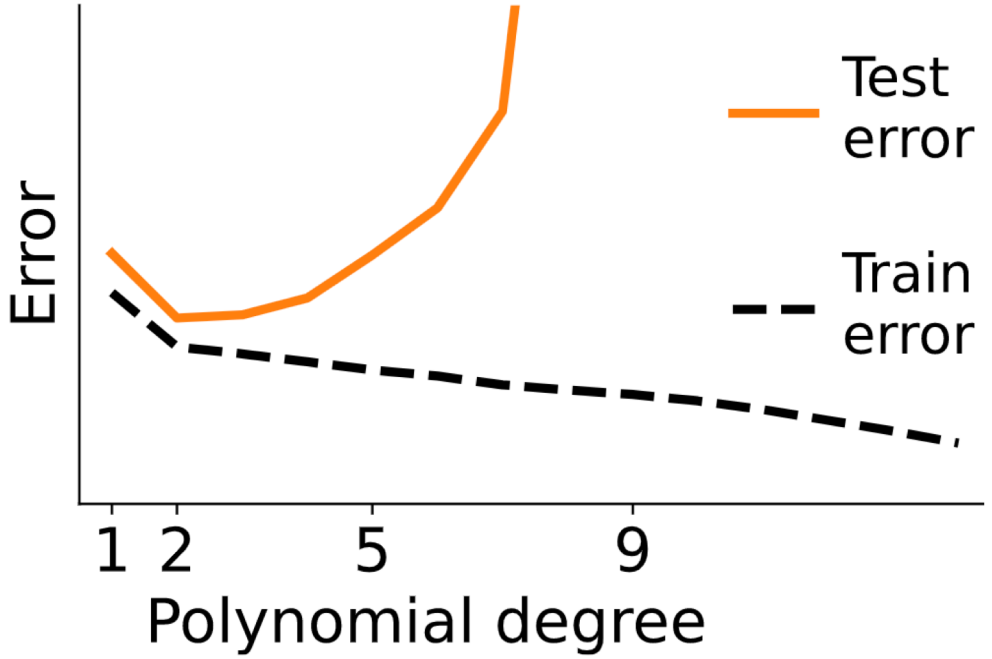
TRAIN V/S TEST ERROR : INCREASING COMPLEXITY



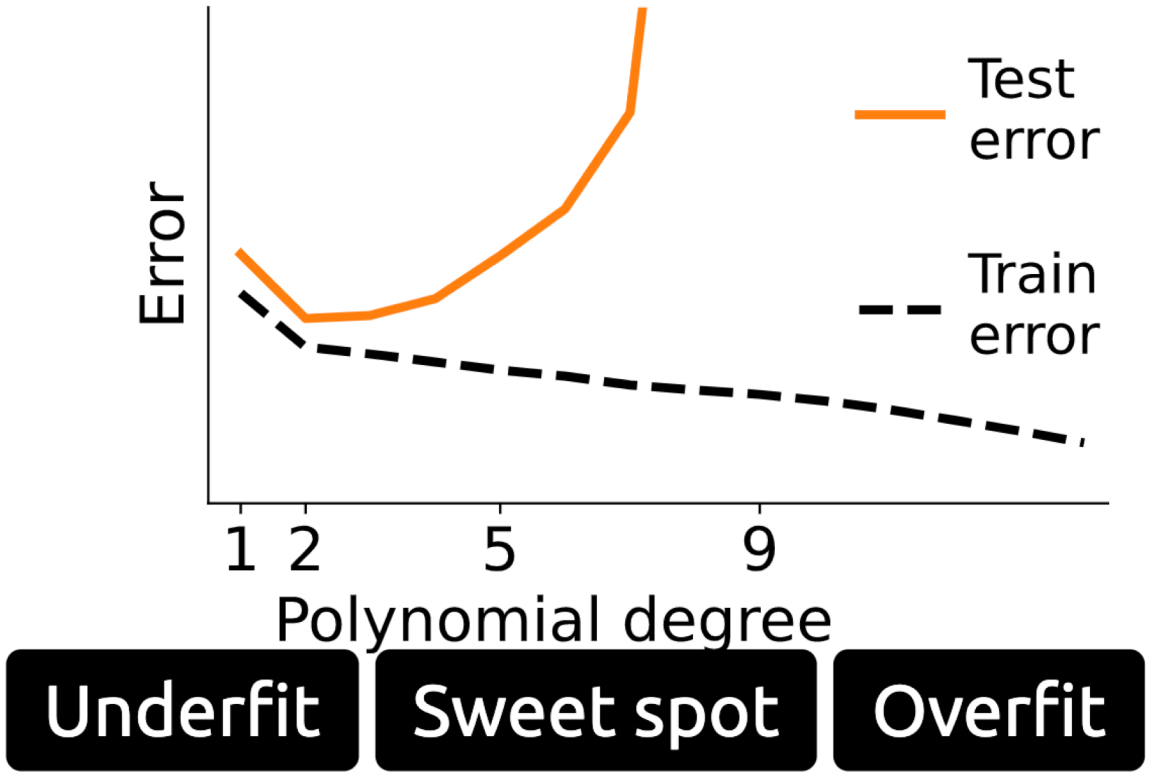


TRAIN V/S TEST ERROR : INCREASING COMPLEXITY

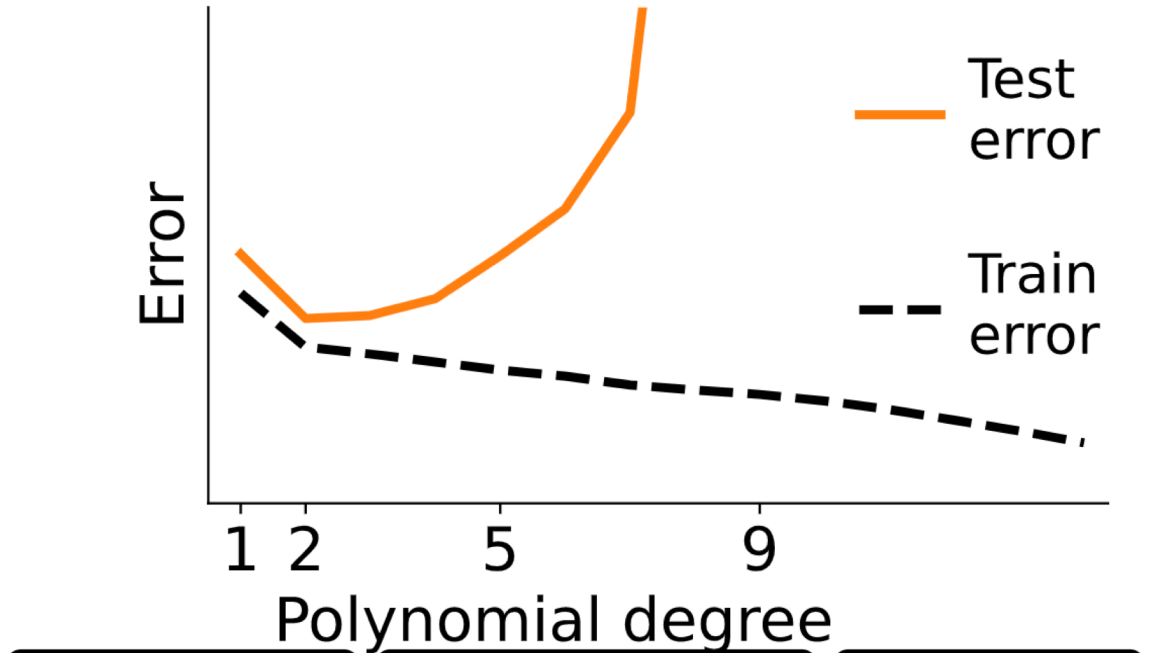




TRAIN V/S TEST ERROR : INCREASING COMPLEXITY

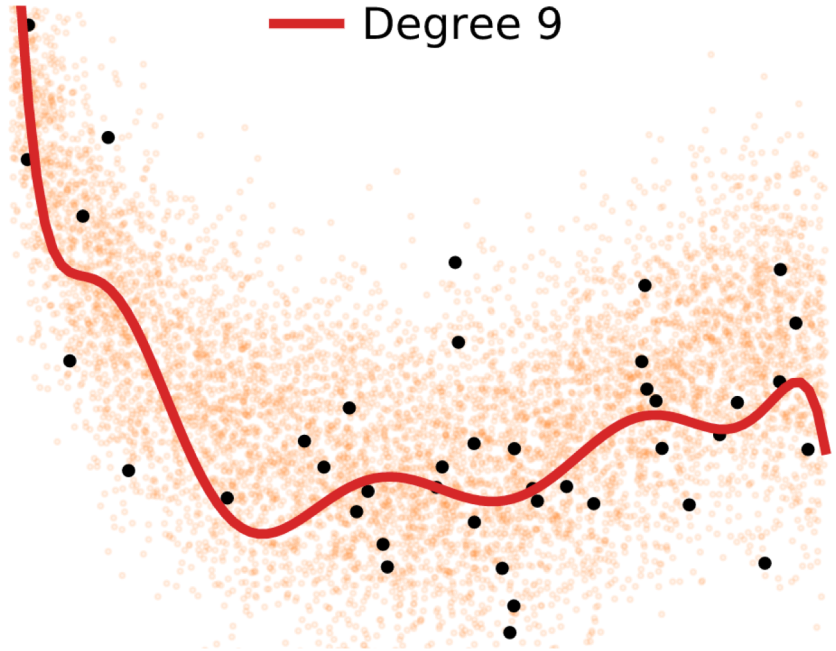


TRAIN V/S TEST ERROR : INCREASING COMPLEXITY

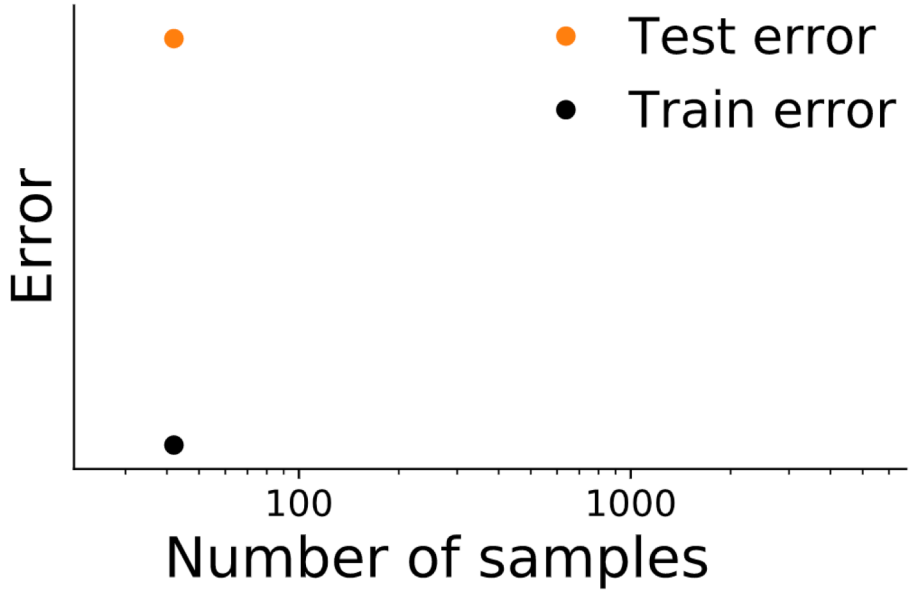


Varying the sample size: learning curves

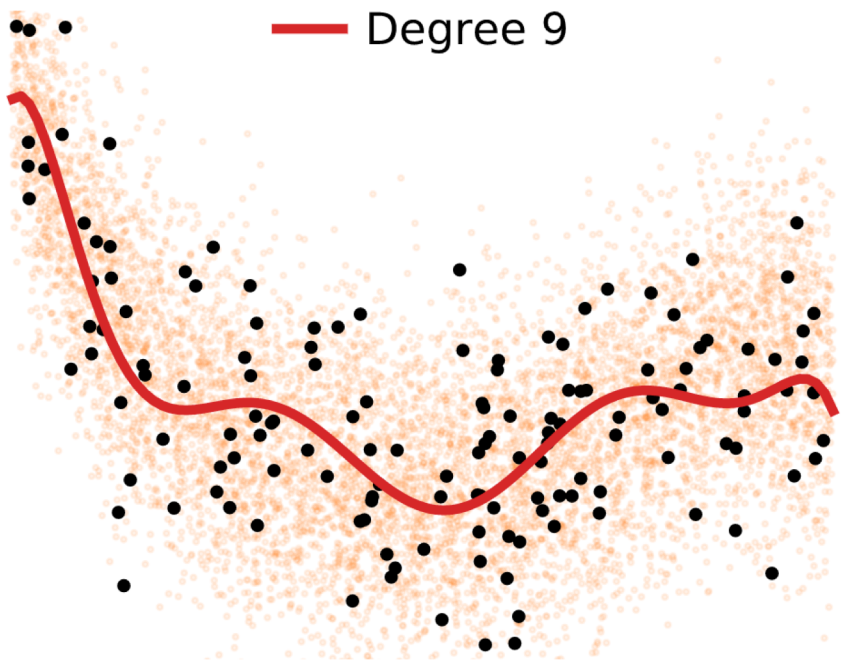
VARYING SAMPLE SIZE

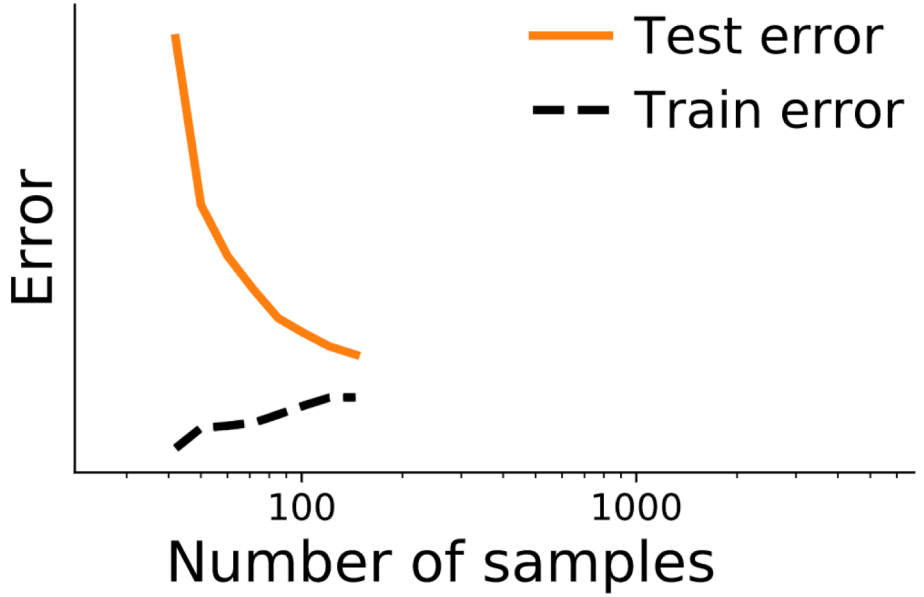


Overfit

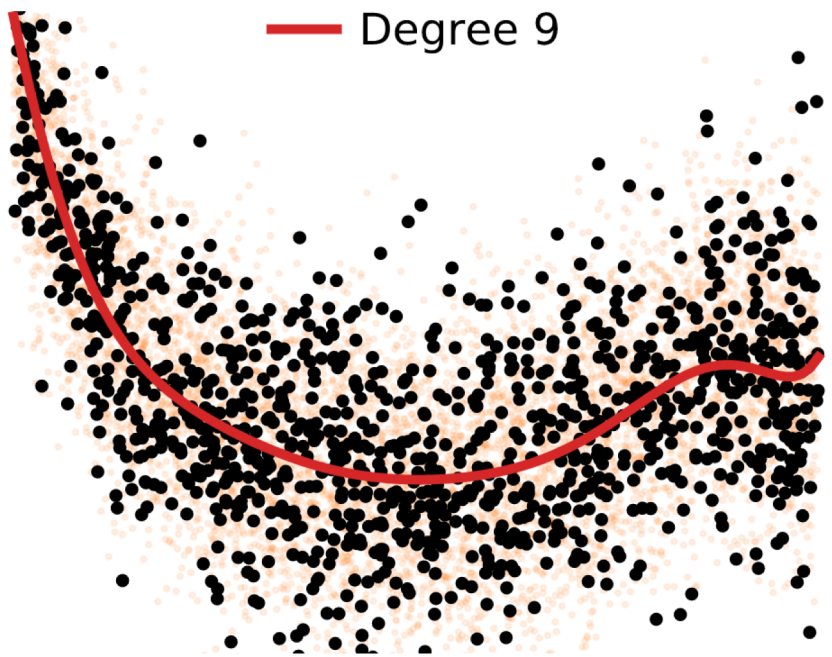


VARYING SAMPLE SIZE

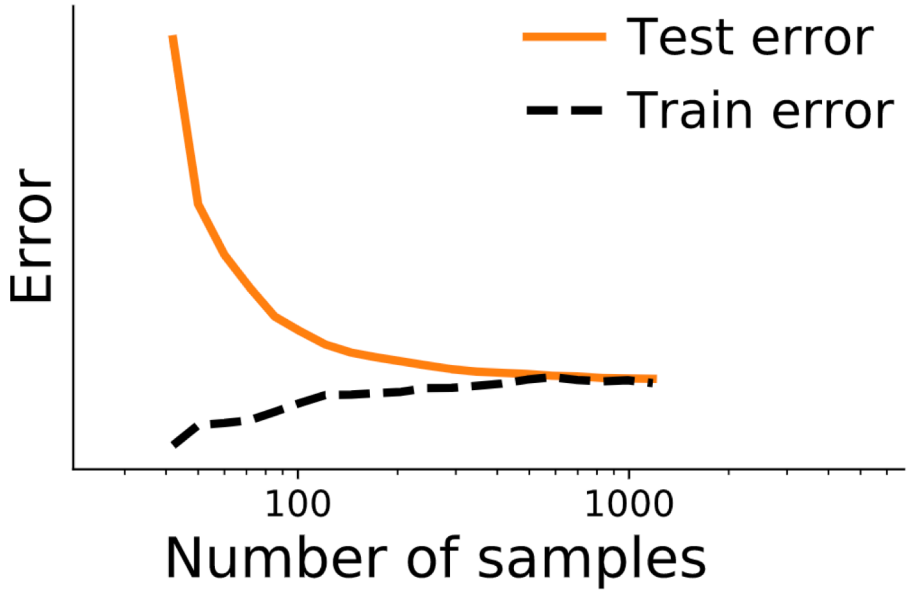




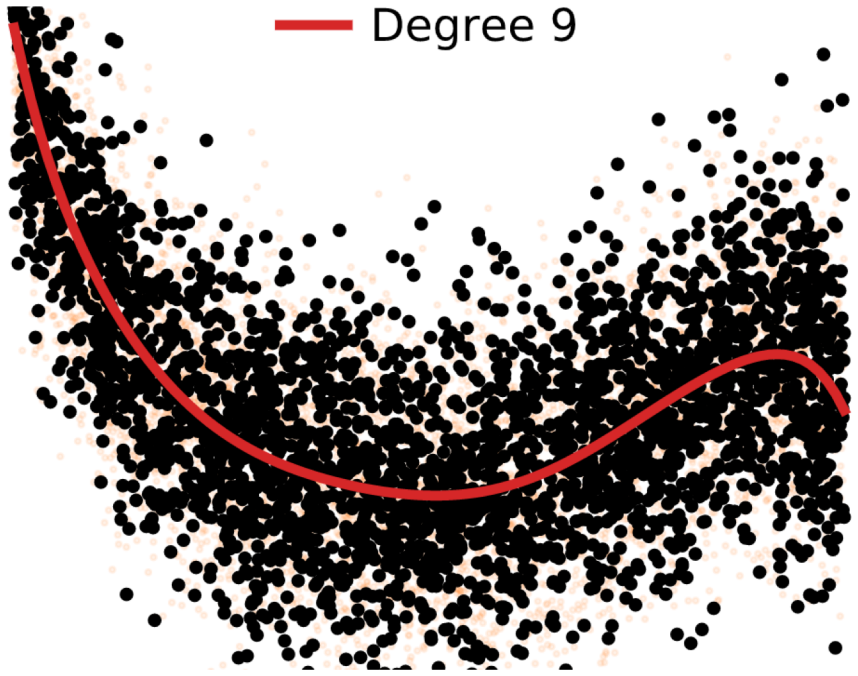
VARYING SAMPLE SIZE



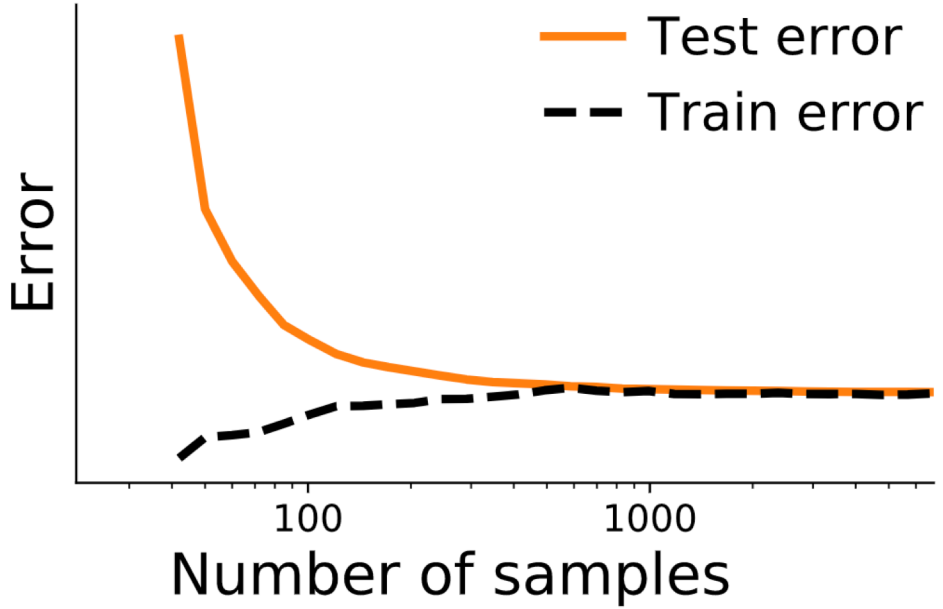
Sweet spot?



VARYING SAMPLE SIZE



Diminishing returns



Shriman TIWARI

Tech Lead/Manager Data Science

Mobile: +33 (0)6 49 71 80 68
shriman.tiwari@keyrus.com

www.keyrus.com

KEYRUS