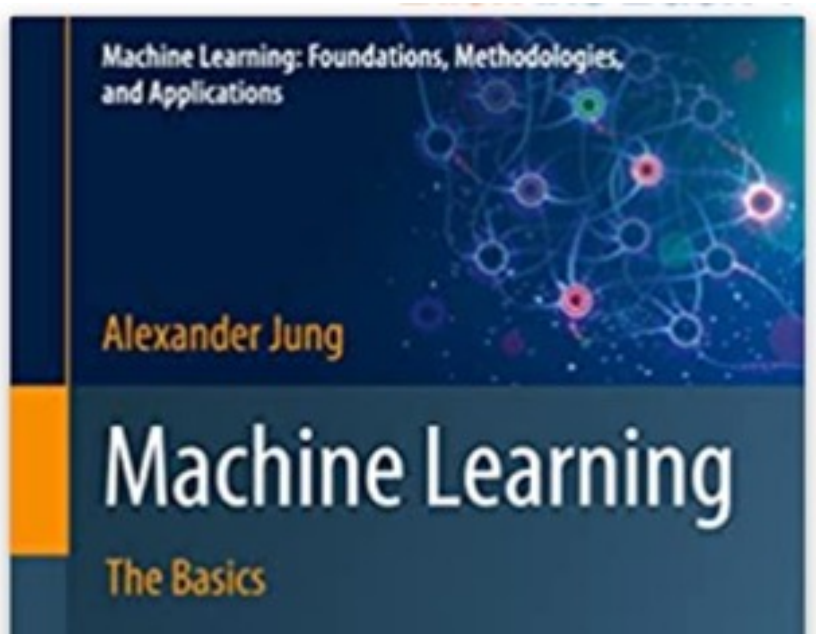
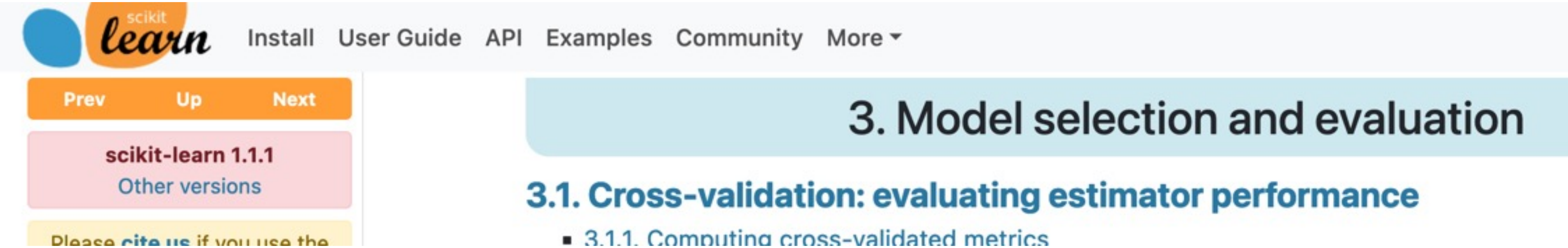
Model Validation and Selection

Alex(ander) Jung
Assistant Professor for Machine Learning
Department of Computer Science
Aalto University

Reading.

Ch. 6 of https://mlbook.cs.aalto.fi





https://scikit-learn.org/stable/model_selection.html

Model Validation

How do we know a ML method is any good ?

Model Selection

How to choose between different alternative methods?

Learning Goals

- know train err is bad quality measure for ML method

- val.err. is more useful as quality measure for a ML model

- basic idea of k-fold CV

- hyper-parameter tuning = model selection

- Python implementations of k-fold CV / gridsearch

”Model”

=

Hypothesis Space

What are three main components of machine learning?

1. Data

Data

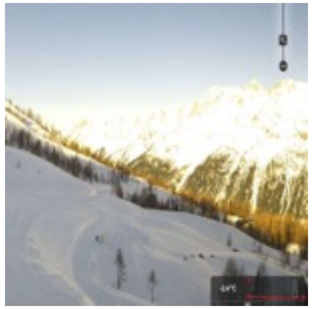
- set of “data points” (atomic unit of information)
- data point has features and labels
- features are properties that can measured easily
- labels =higher-level facts or quantities of interest

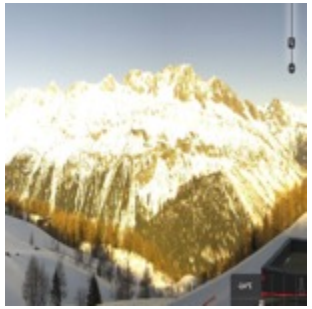
Data Point = ”Some Ski Day”

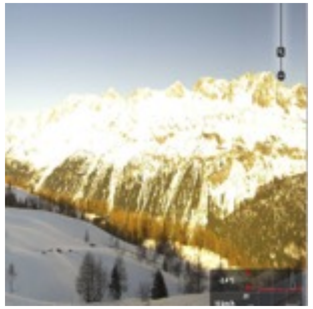
feature x : morning temperature

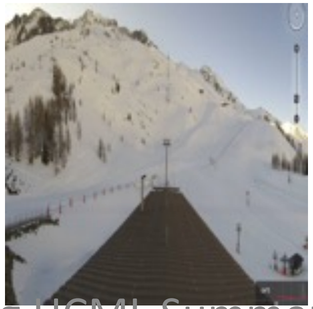
label y : maximum daytime temperature

Data = Bunch of Data Points









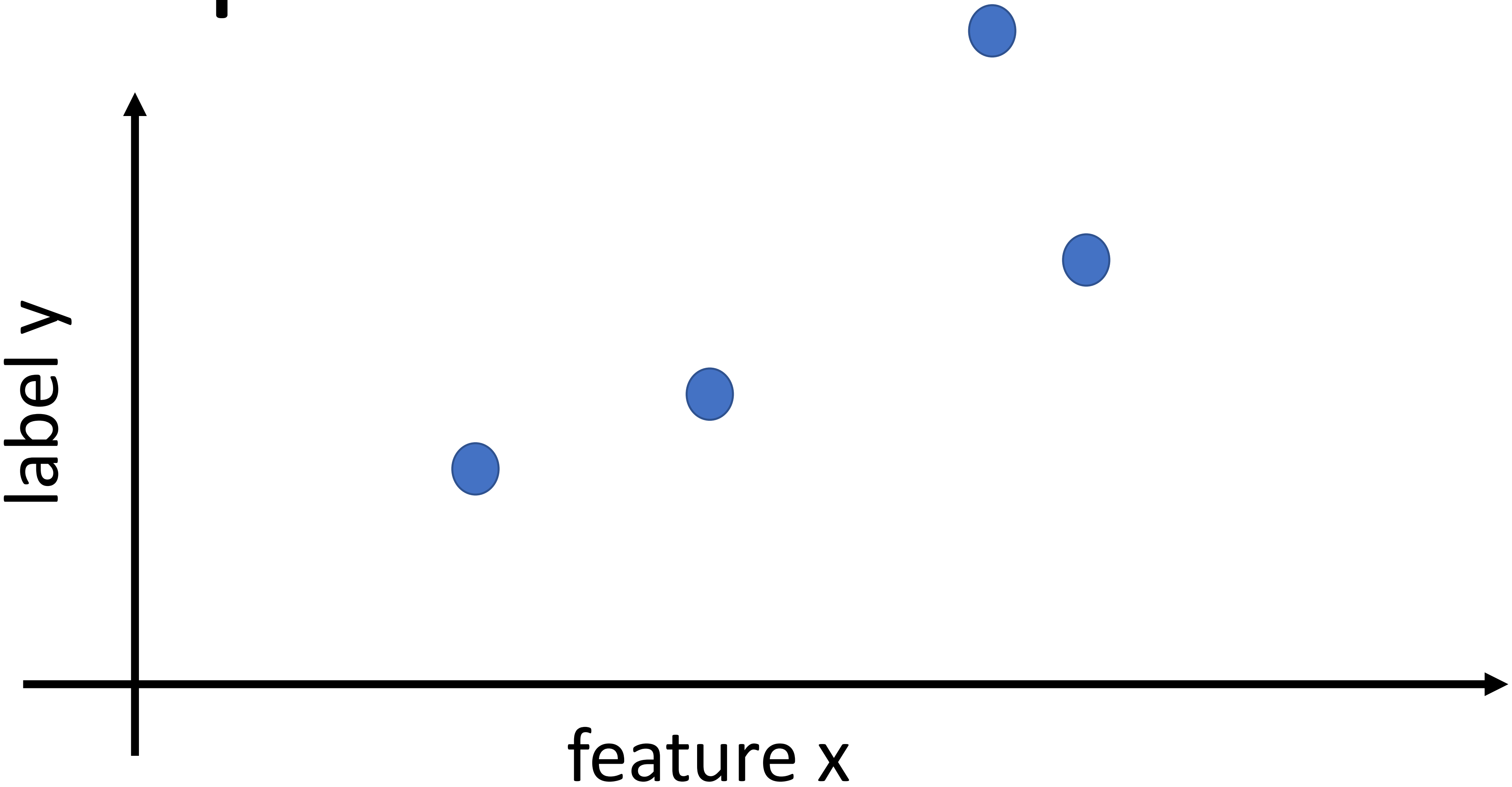
Sample Size

“sample size” m

=

number of (labeled) data points

Sample Size m = 4



2. Hypothesis Space

How Many Hypotheses Are There?

predict

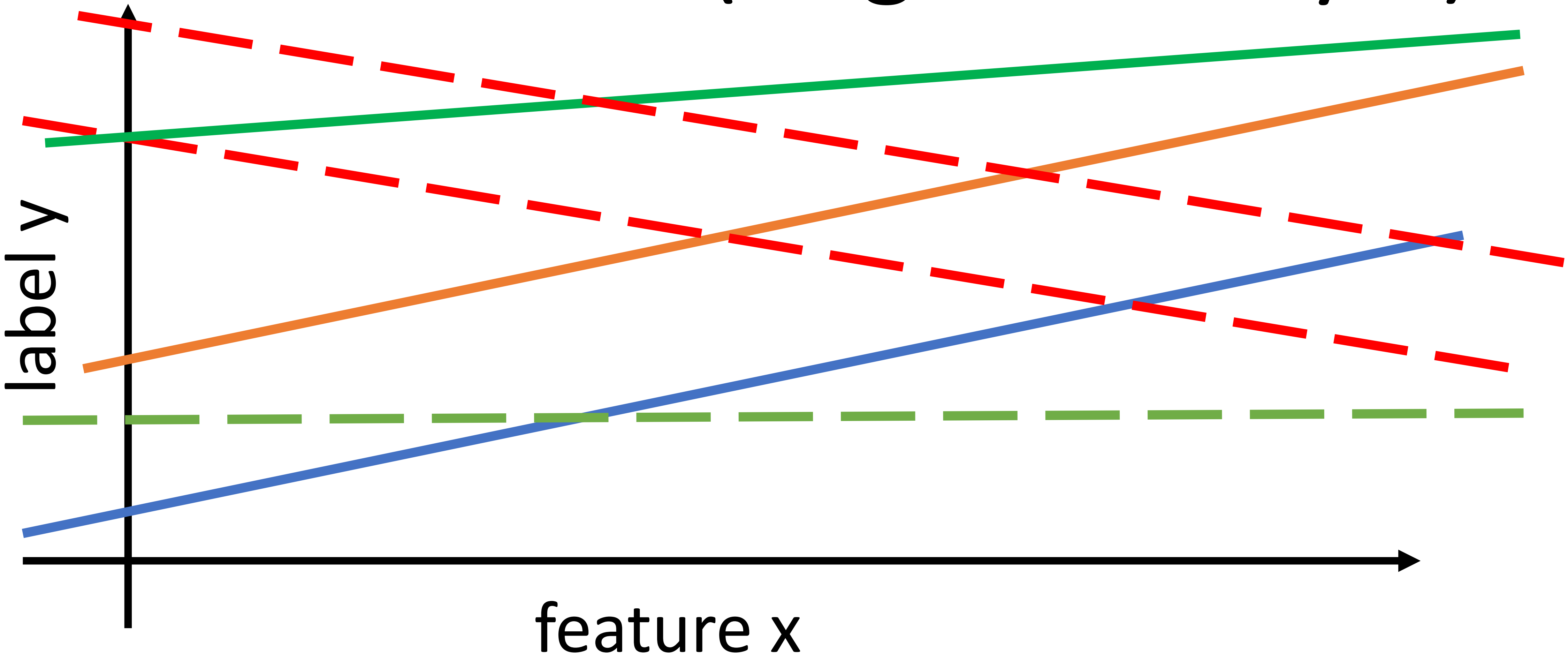
h(x)

predicted max temp.

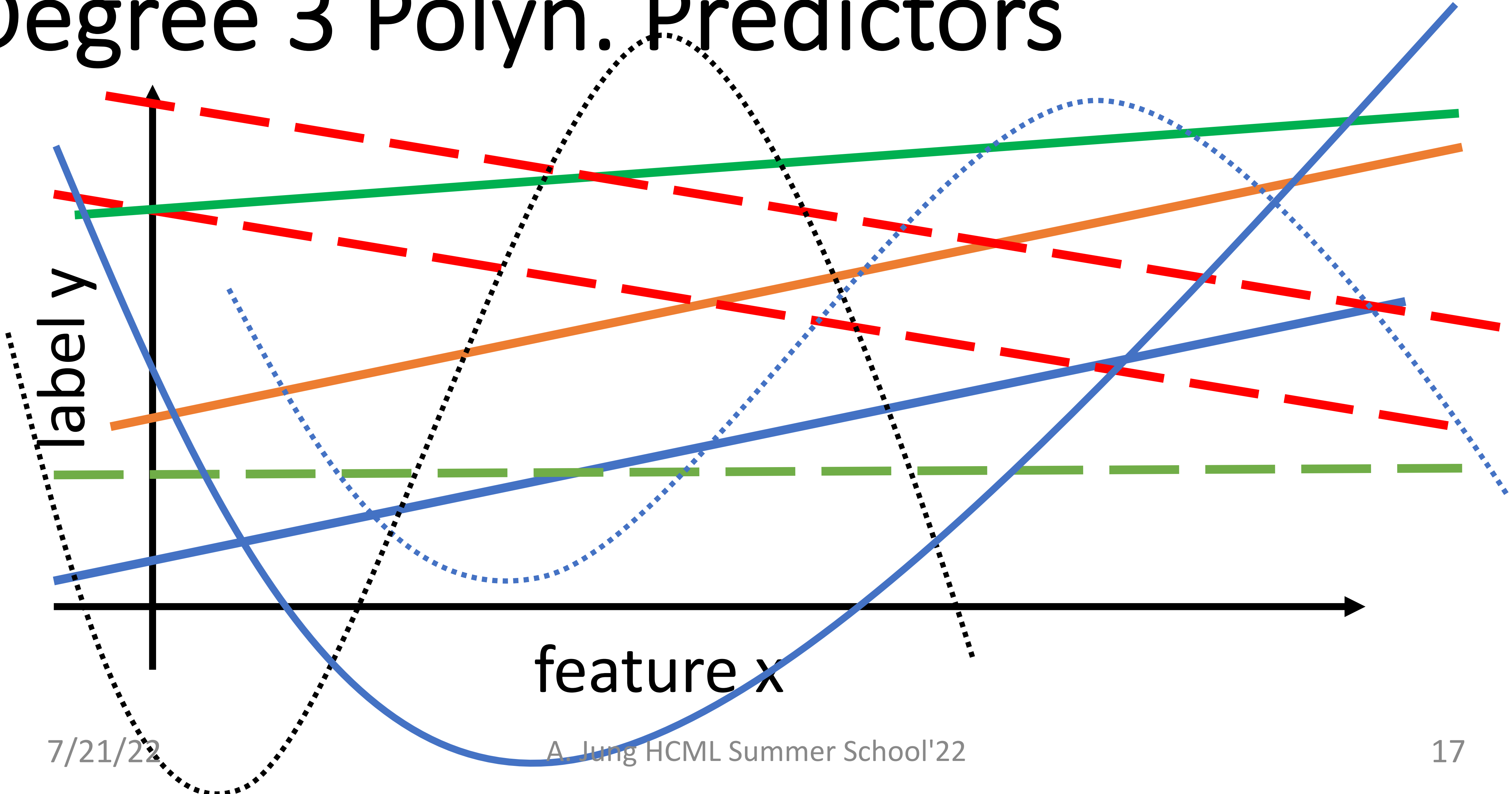
$\hat{y} = -2.3$

feature x = -10

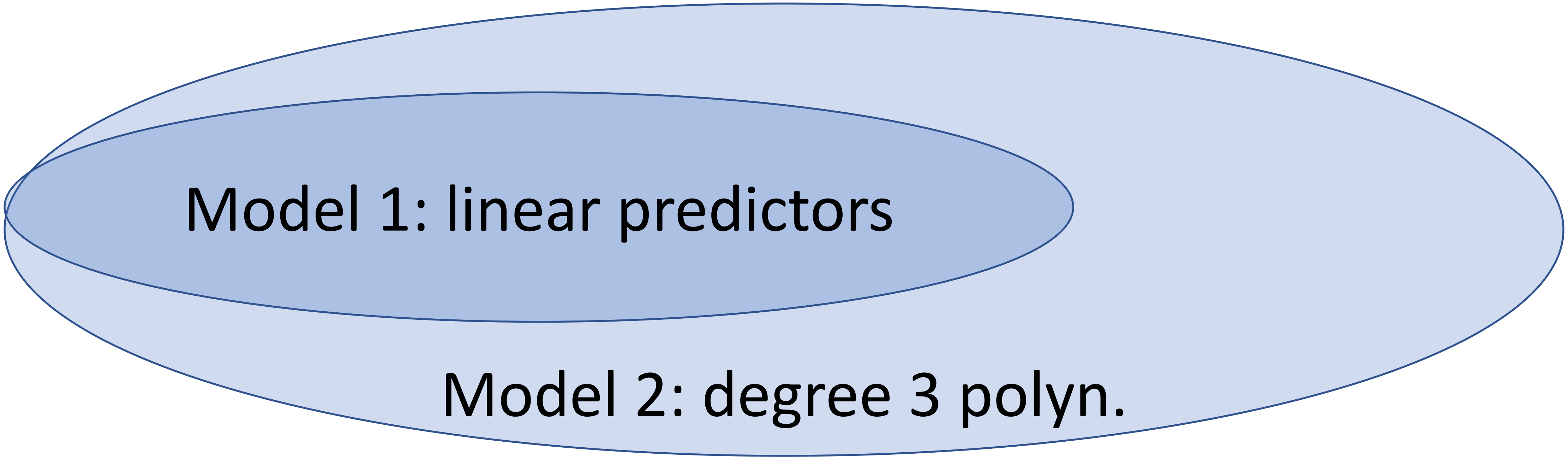
Model 1: Linear Predictors (Degree 1 Polyn.)



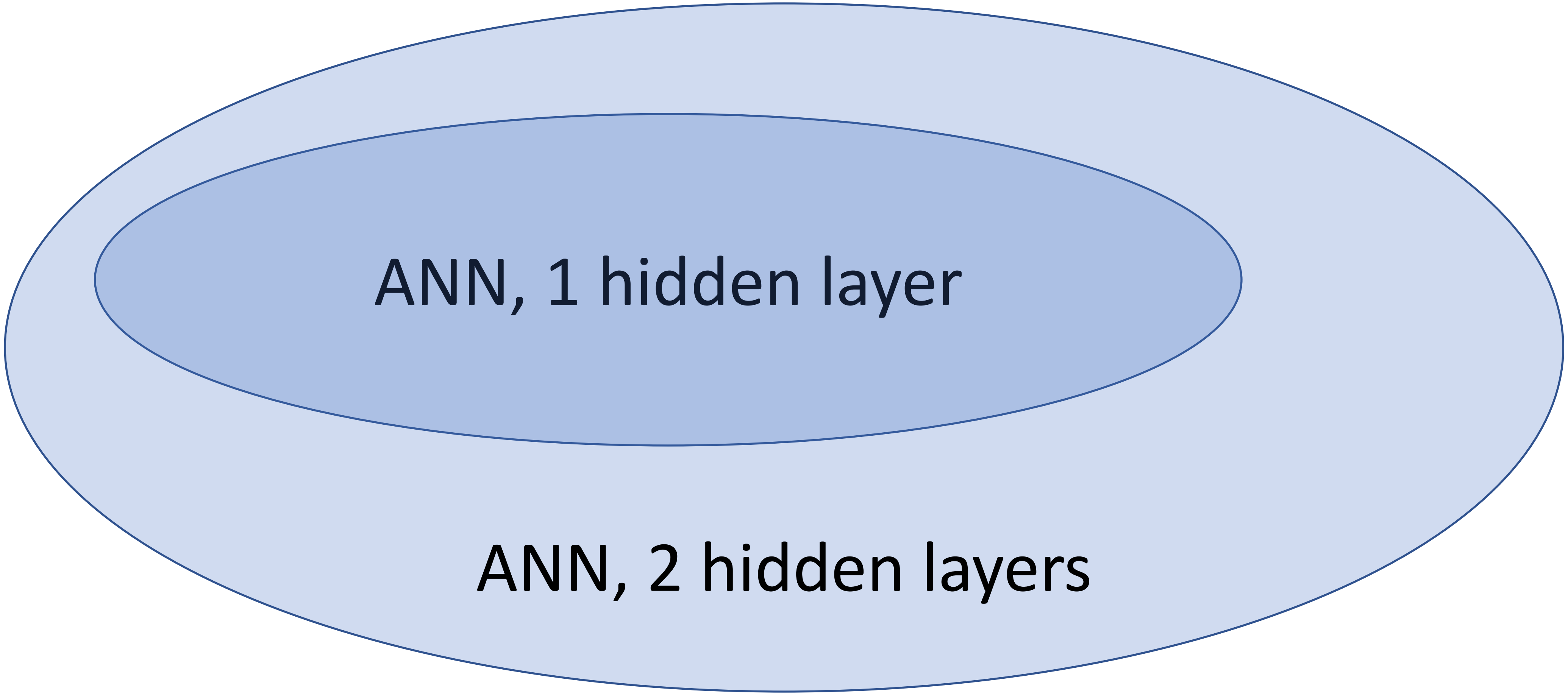
Model 2:
Degree 3 Polyn. Predictors



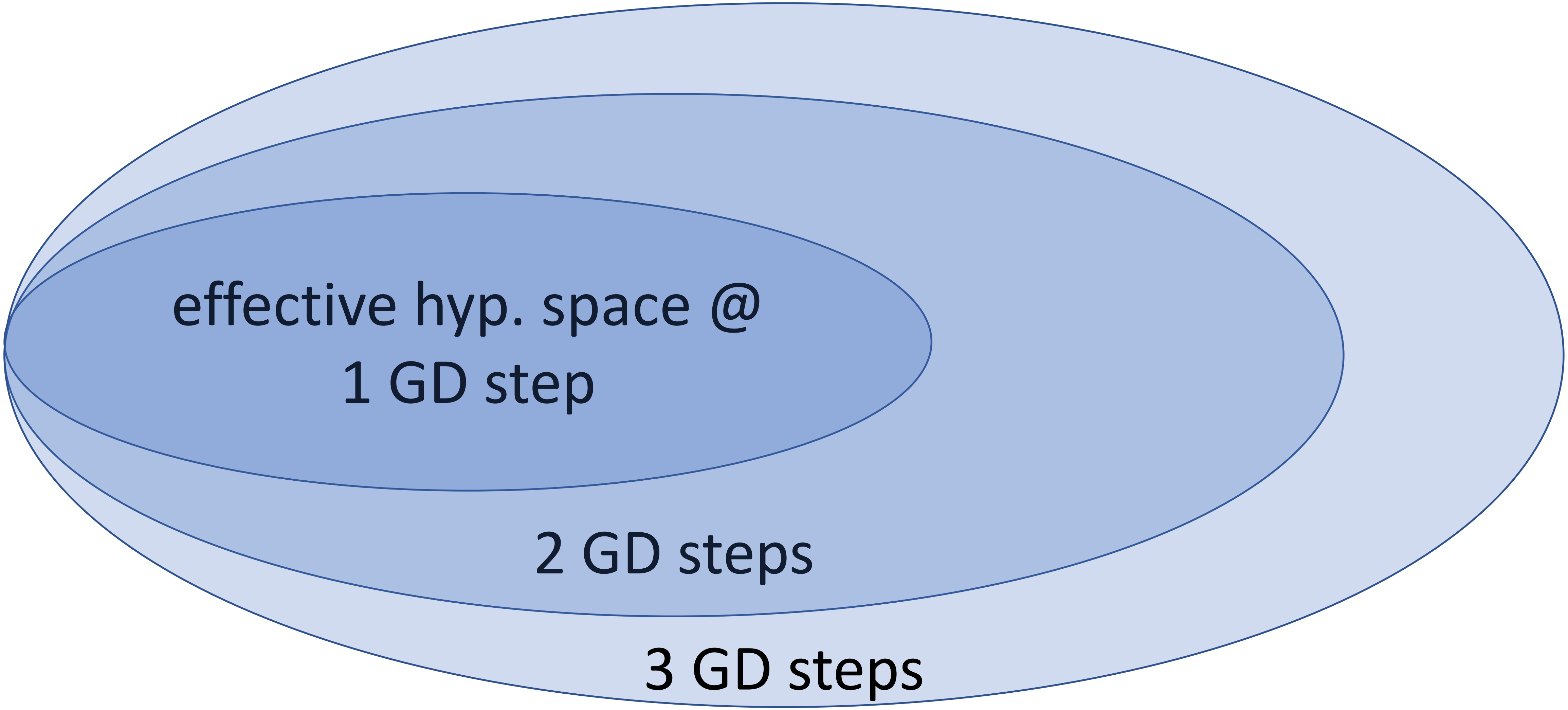
Nested Models – I



Nested Models - II



Nested Models - III



Math Notation

$$\mathcal{H}^{(n)} = \left\{ h(x) = \sum_{l=0}^{n-1} w_l x^l \; with \; some \; w_l \right\}$$

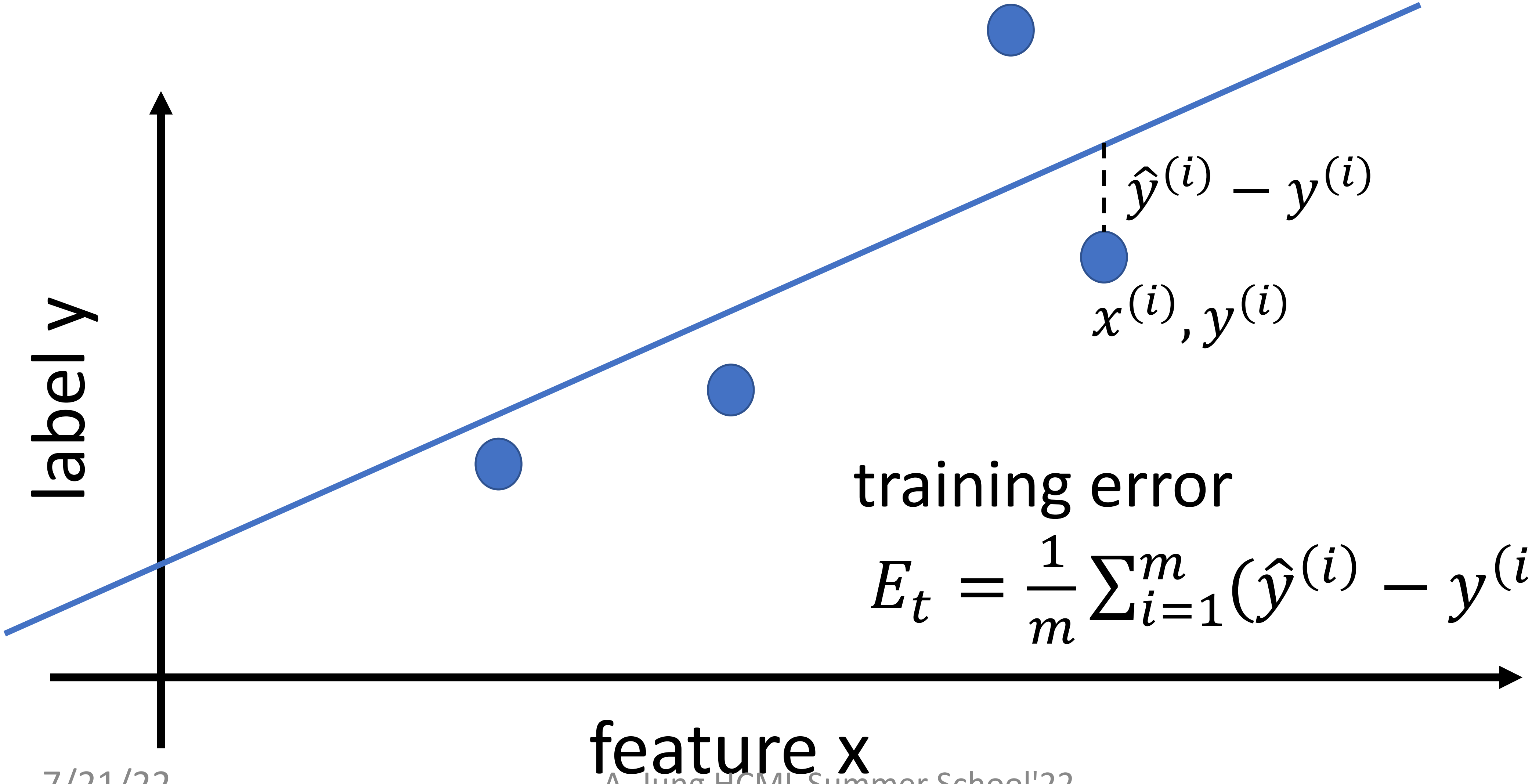
$\mathcal{H}^{(2)}$ … linear hypotheses

$\mathcal{H}^{(4)}$ … degree 3 polyn.

$$\mathcal{H}^{(1)} \subseteq \mathcal{H}^{(2)} \subseteq \mathcal{H}^{(3)} \subseteq \mathcal{H}^{(4)} \subseteq \ldots$$

3. Loss Function

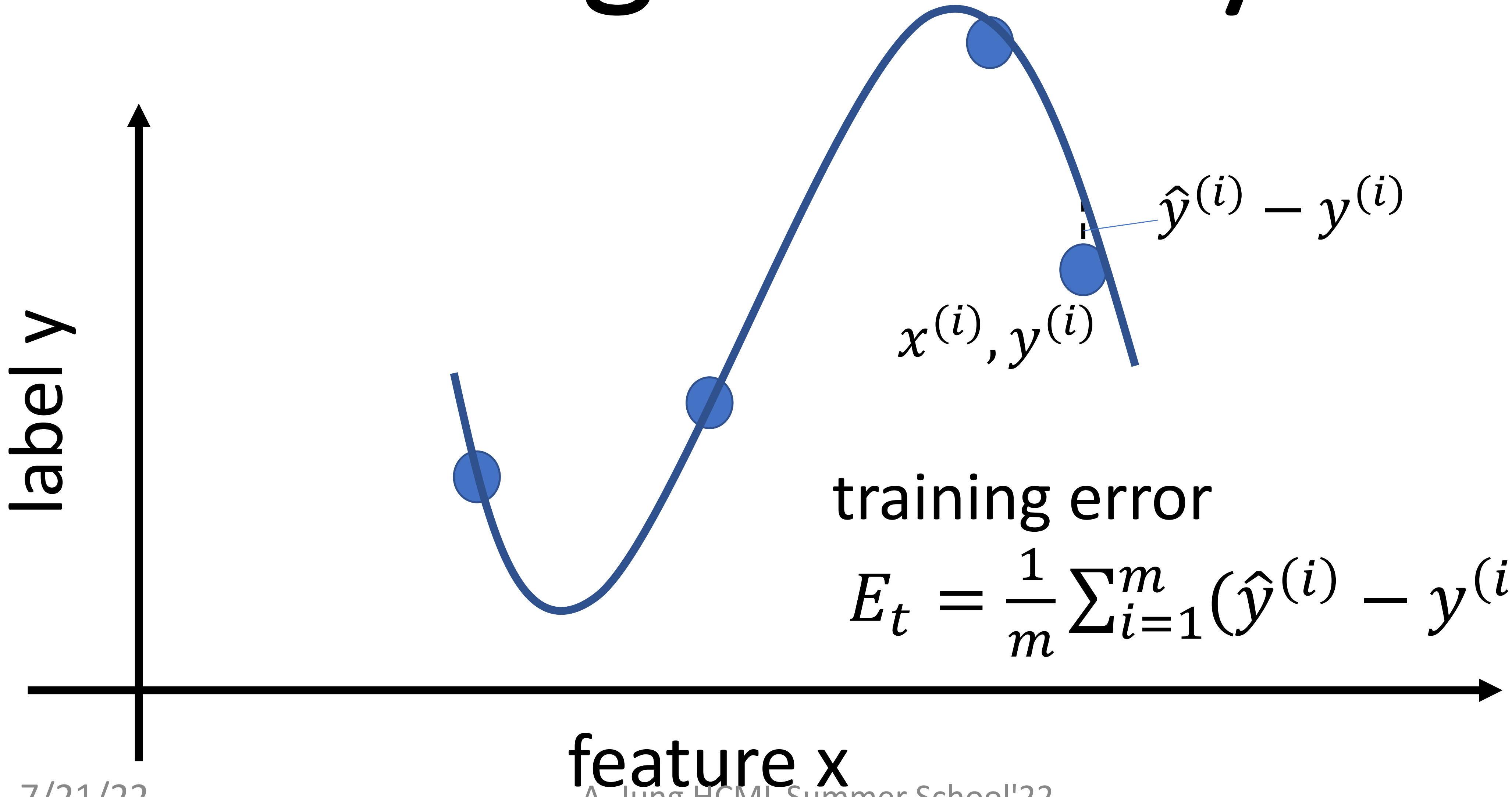
Learn Linear Predictor



training error

$$E_t = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

Learn Degree 3 Polyn.



training error

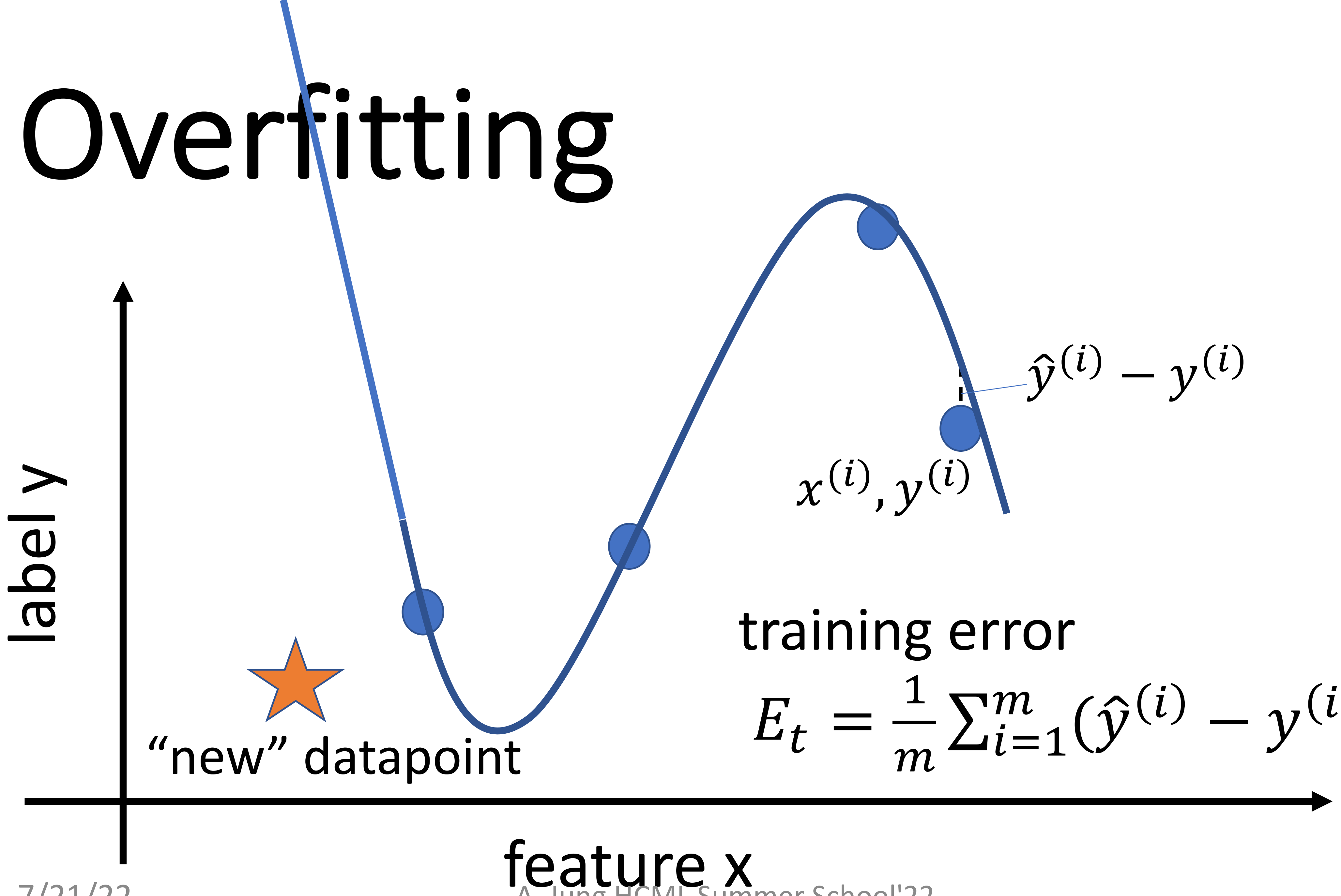
$$E_t = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

Training Errors

model 1
linear predictors

model 2:
degree 3 polyn.

Overfitting



training error

$$E_t = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

Small Training Error Does Not Imply Good Performance on New Data Points!

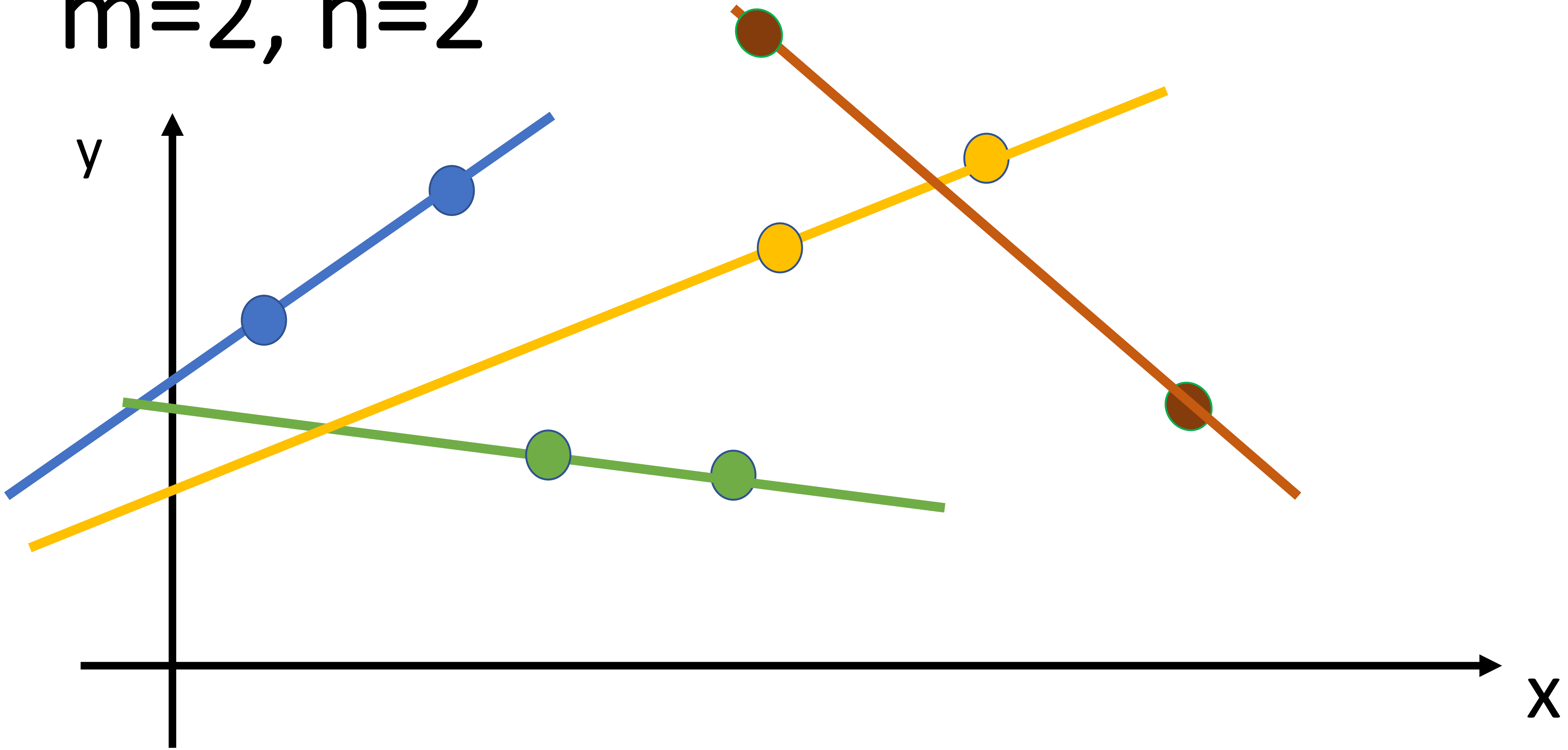
Small Training Error Merely Indicates That Optimization/Training Algorithm Works

A Case in Point

we can perfectly fit (almost) any m data points using polynomials of degree n-1 as soon as

$$n \geq m$$

m=2, n=2



Reminder: Probabilistic Model

- data points are realizations of RVs
- joint pdf p(x,y) of features and label
- training set is a RV
- learnt hypothesis h(.) is a RV
- prediction h(x) is a RV

Why is Train. Err. Misleading?

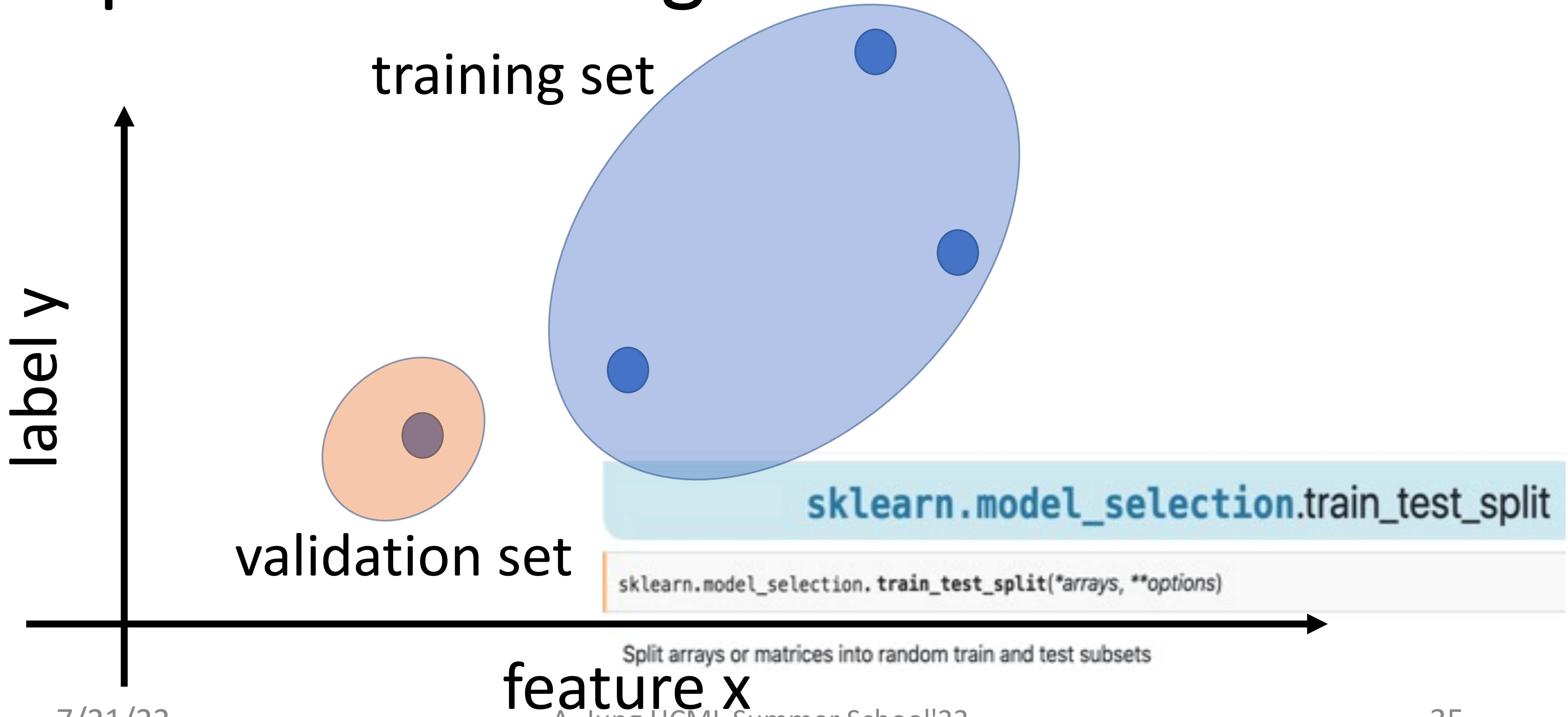
- consider expected loss of hypothesis
- estimate expectation using sample average
- this only works if hypothesis does not depends on data points used in average
- does not hold for training error

Model Validation and Selection

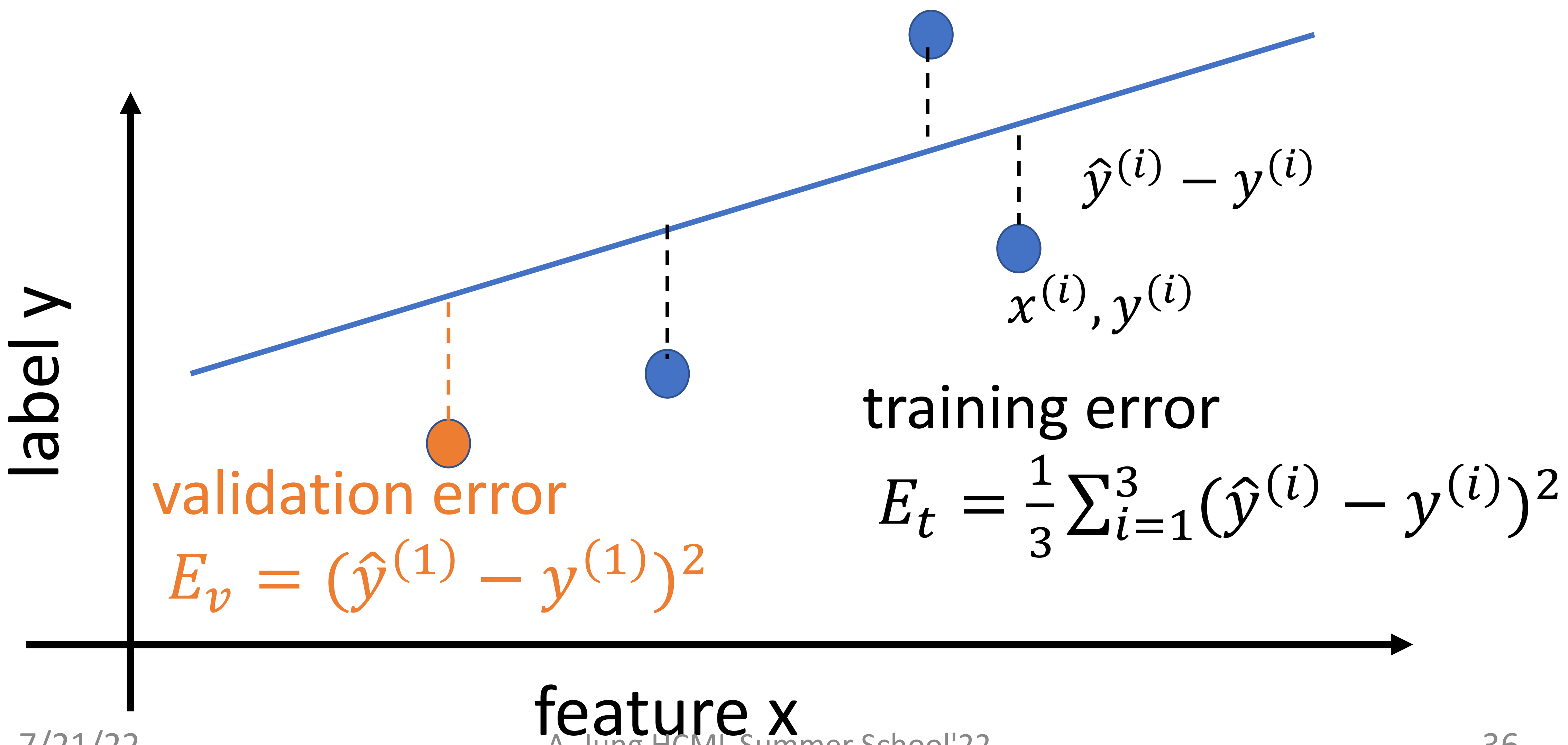
Basic Idea of Validation

- divide data points into two subsets
- use training set to learn predictor
- use validation set to estimate loss

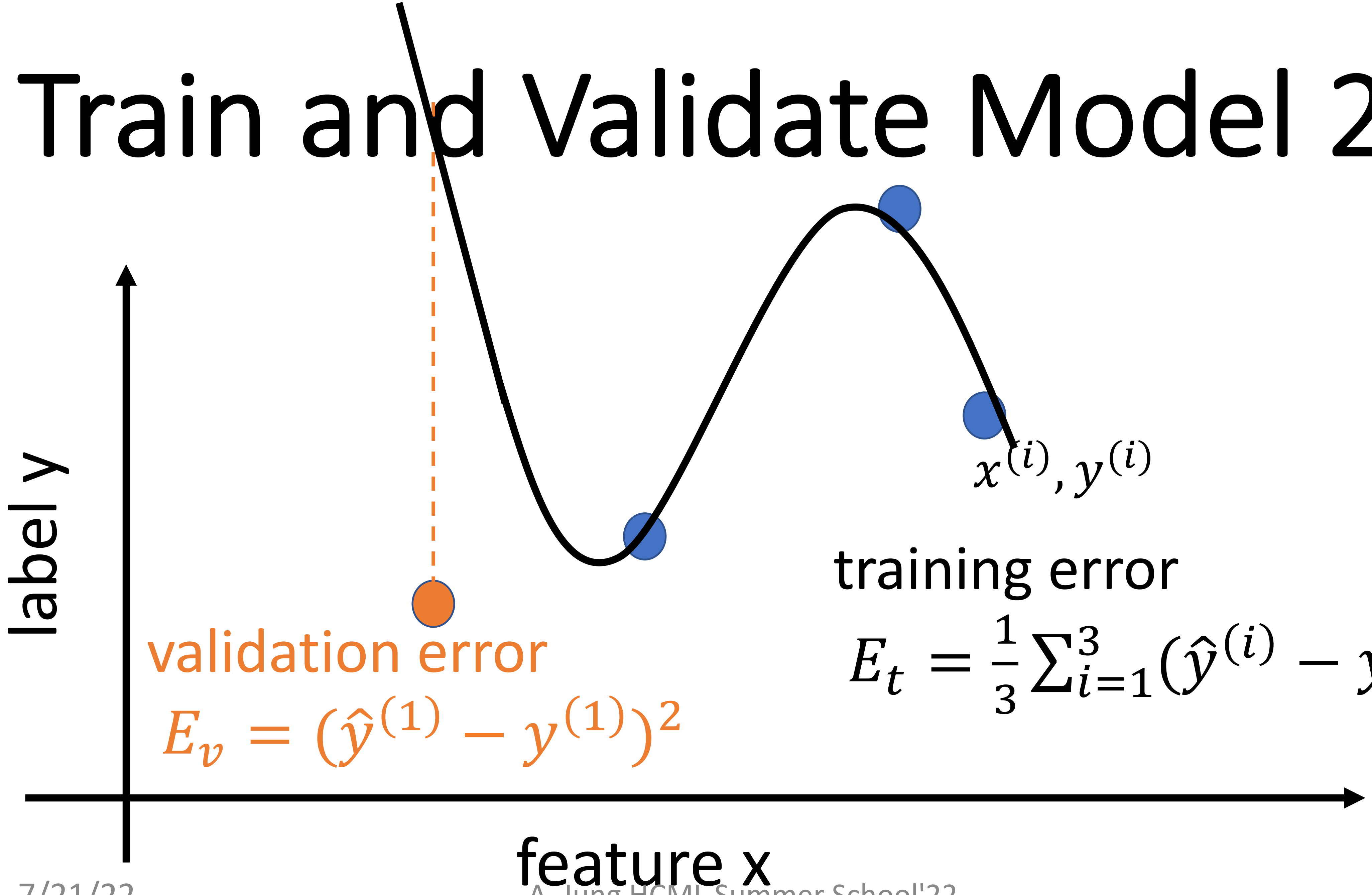
Split into Training and Validation Set



Train and Validate Model 1



Train and Validate Model 2



validation error

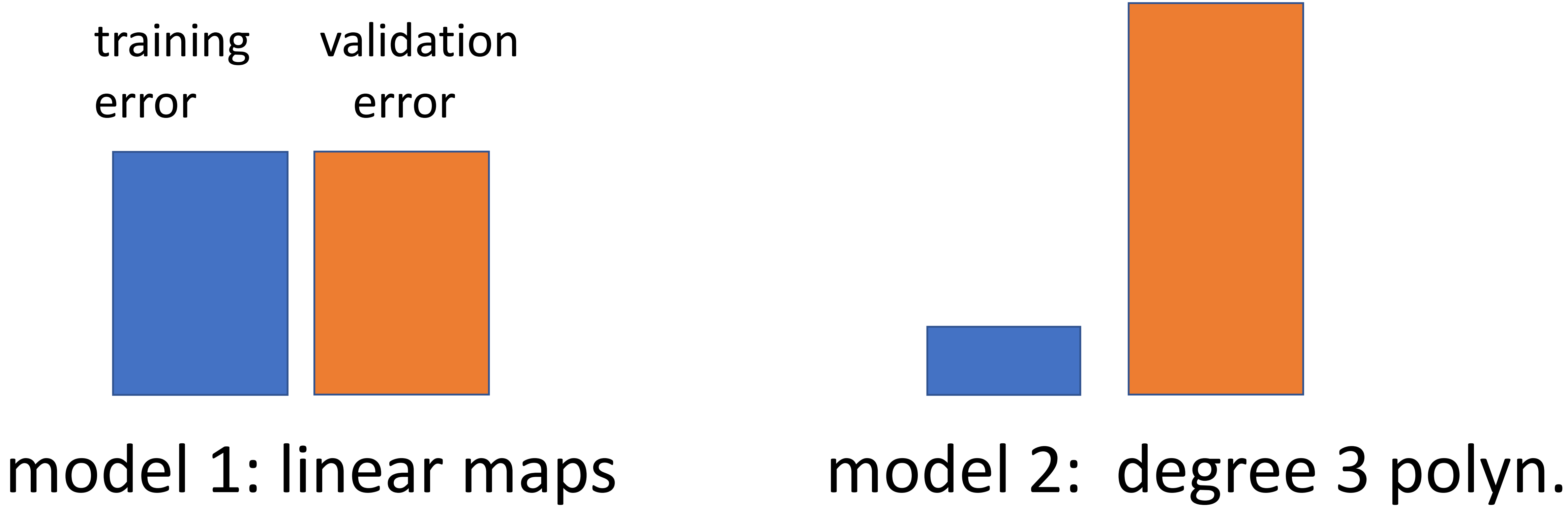
$E_v = (\hat{y}^{(1)} - y^{(1)})^2$

training error

$E_t = \frac{1}{3}\sum_{i=1}^{3}(\hat{y}^{(i)} - y^{(i)})^2$

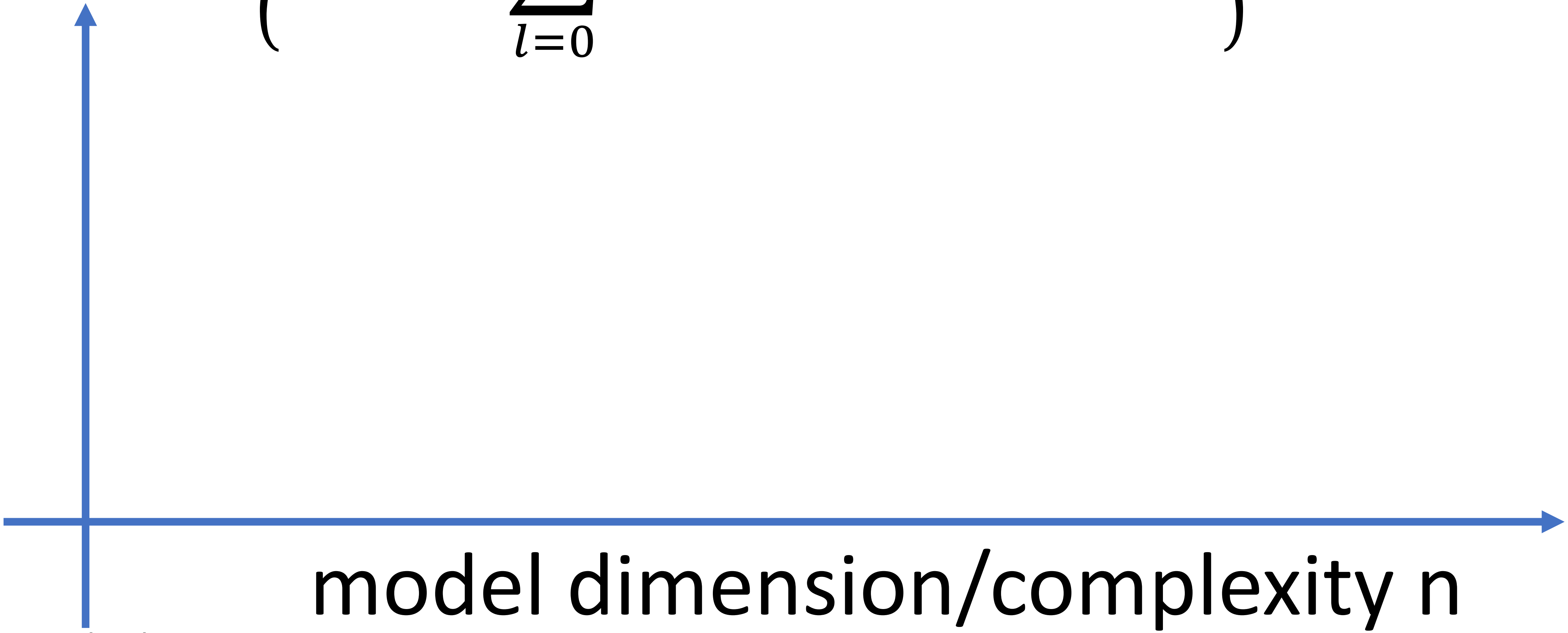
Basic Idea of Model Selection

choose model via validation error



Train/Val Error vs Model Complexity

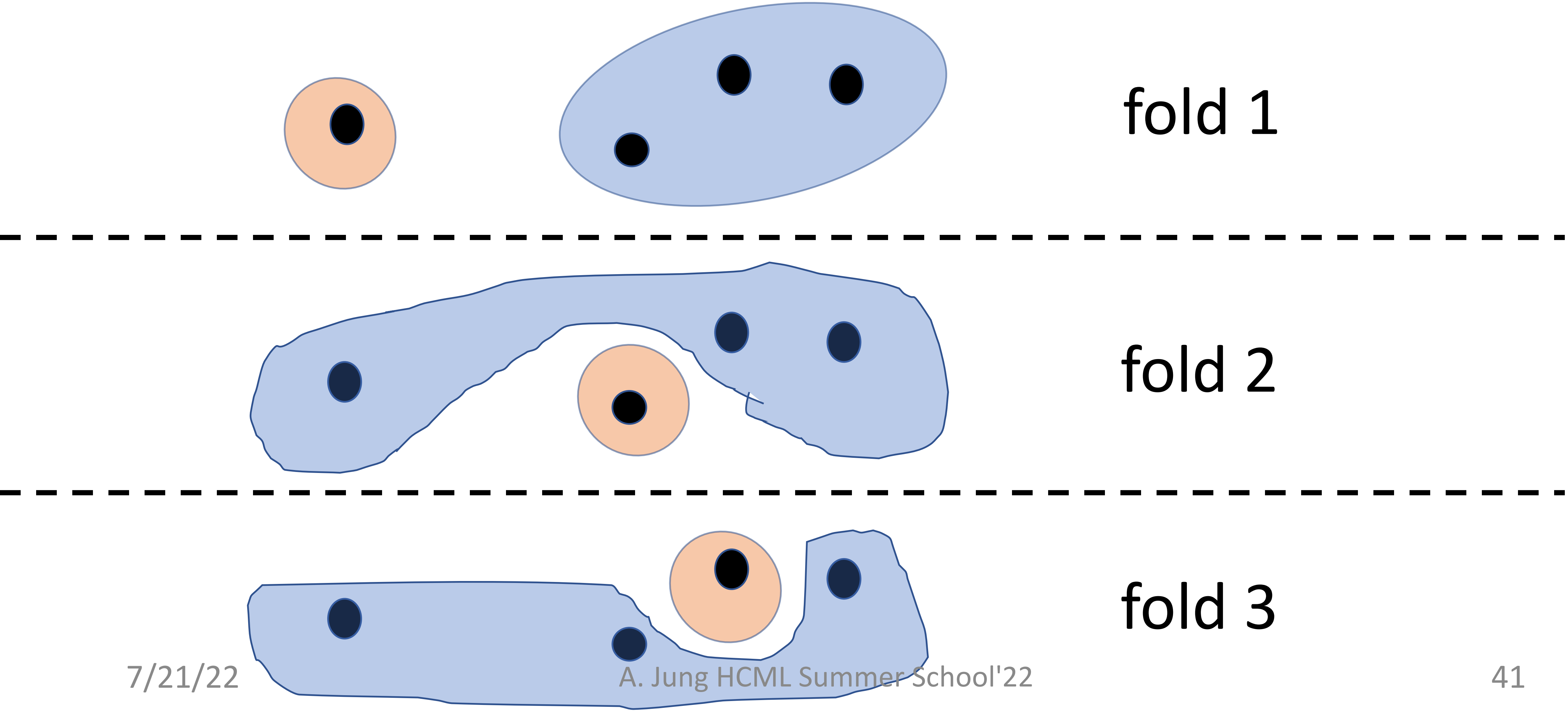
$$\mathcal{H}^{(n)} = \left\{ h(x) = \sum_{l=0}^{n-1} w_l x^l \text{ with } \textcolor{red}{\text{weights}}\ w_l \right\}$$



k-Fold Cross Validation

- might be unlucky with train/val split
- problematic for small datasets
- IDEA: randomly split several times
- “average out” unlucky splits

K-Fold Cross Validation



k-Fold Cross Validation

how to choose nr of folds (the “k” in k-fold CV) ?

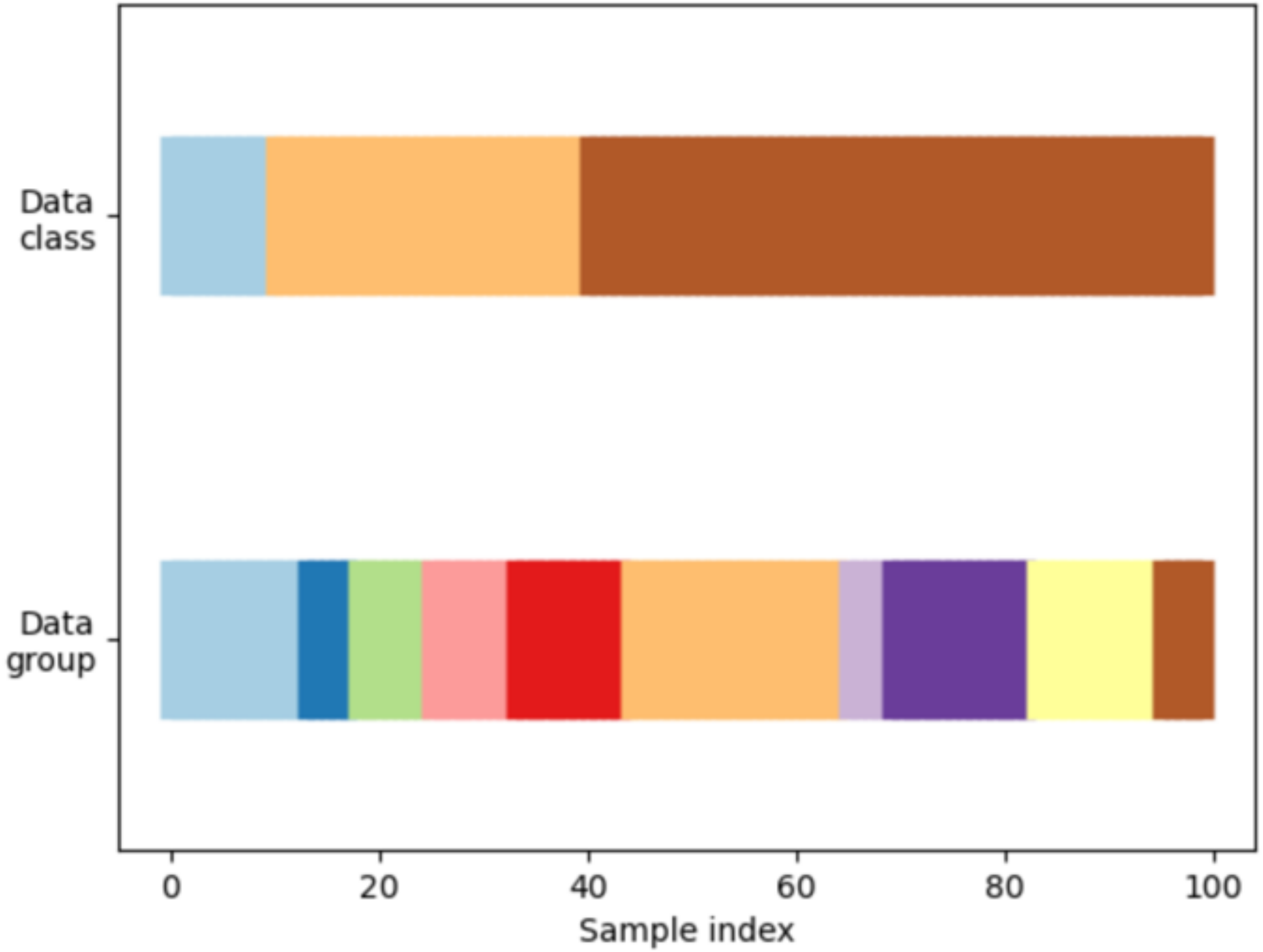
- train fold should be sufficiently large (avoid overfitting)

- val folds should sufficiently large (to get reliable estimate of generalization)

CAUTION!

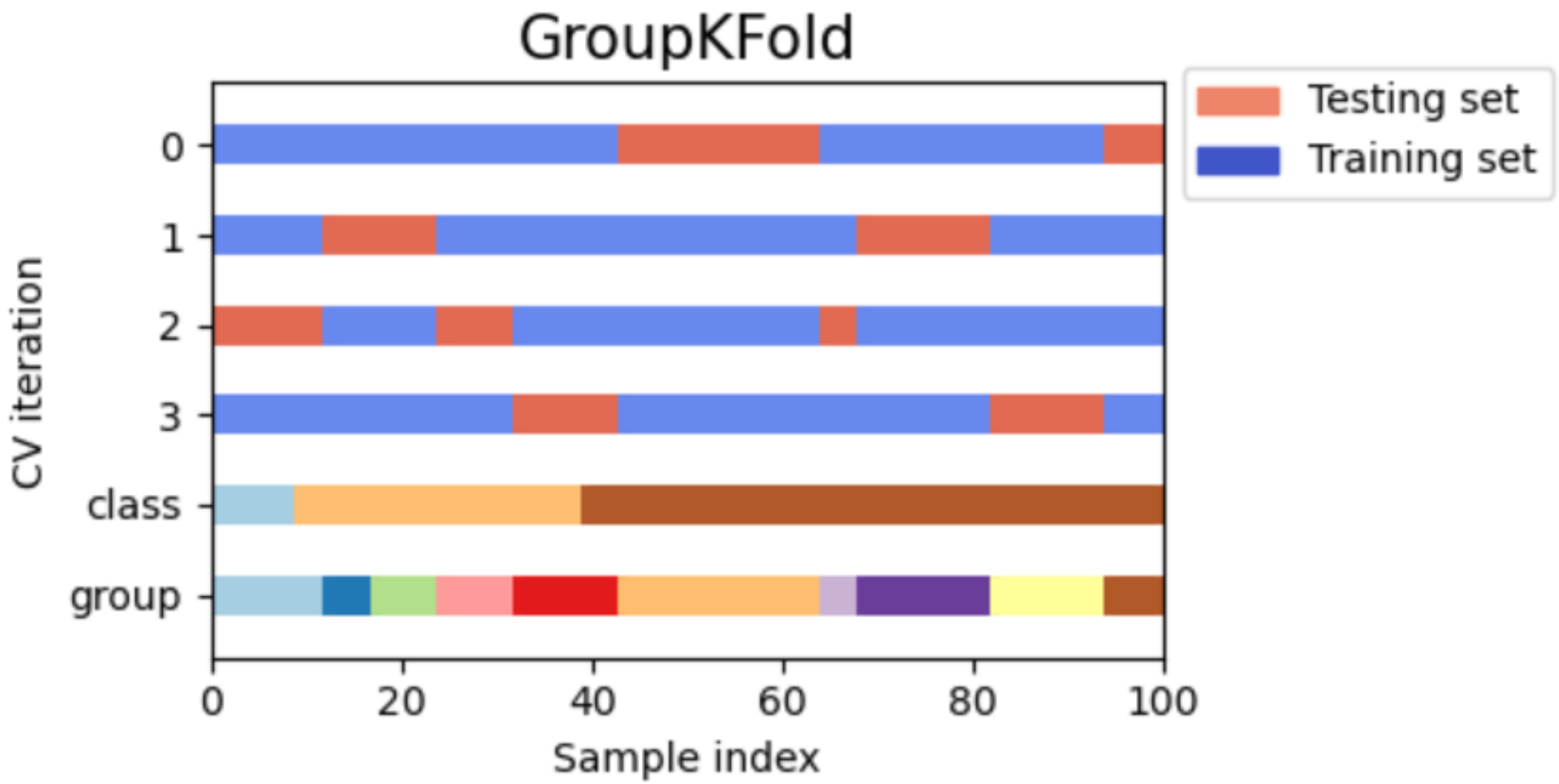
- k-fold CV requires a method to split into folds
- most basic method: evenly divide into k folds
- works if data is i.i.d. (“order of data points is arbitrary”)
- fails if data points are grouped or ordered

Imbalanced Classes and Group Structure



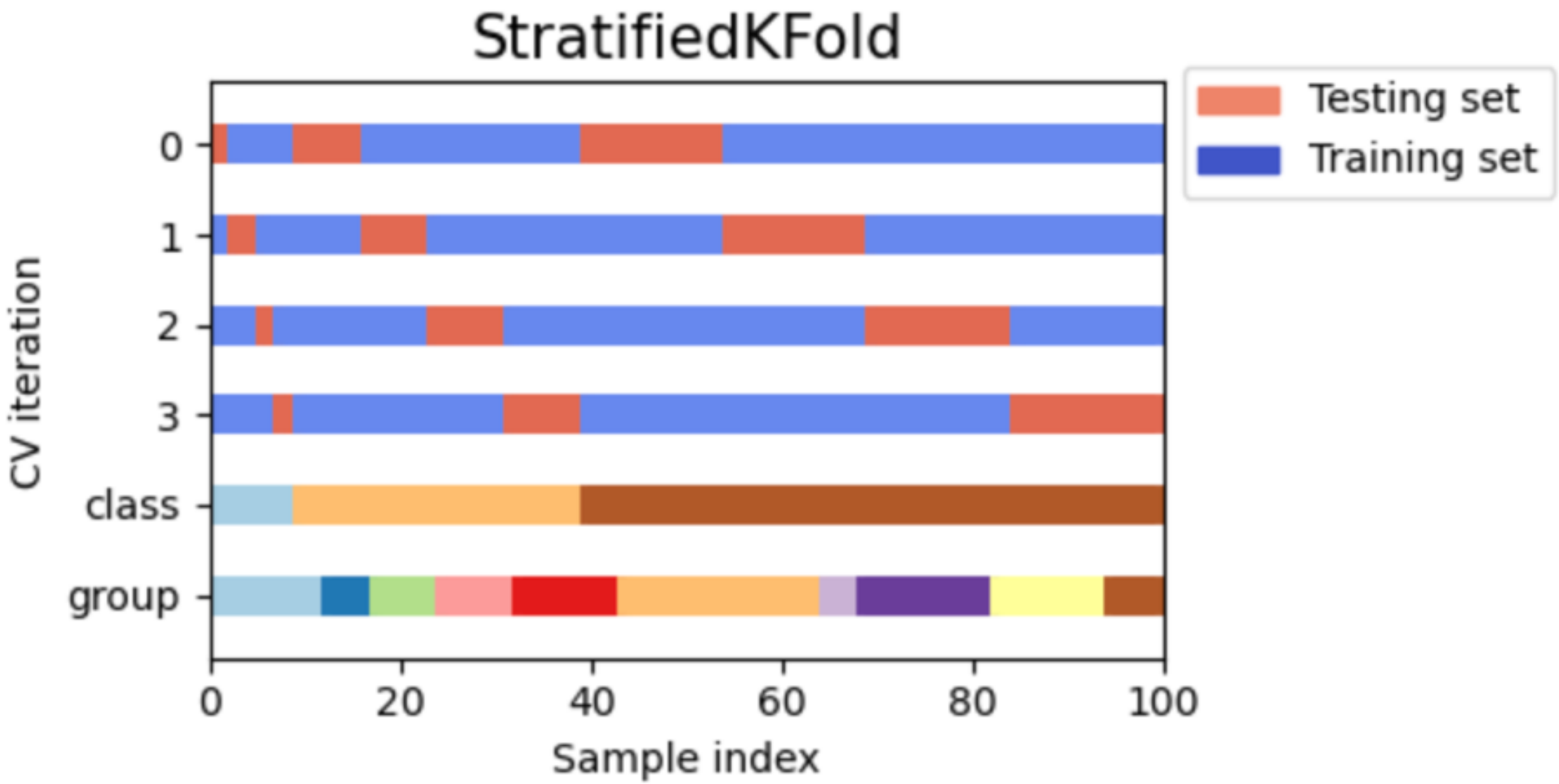
- e.g. data points with same label are contiguous blocks
- or data points are obtained at consecutive time instants (→ correlations)

Group-Preserving Splitting

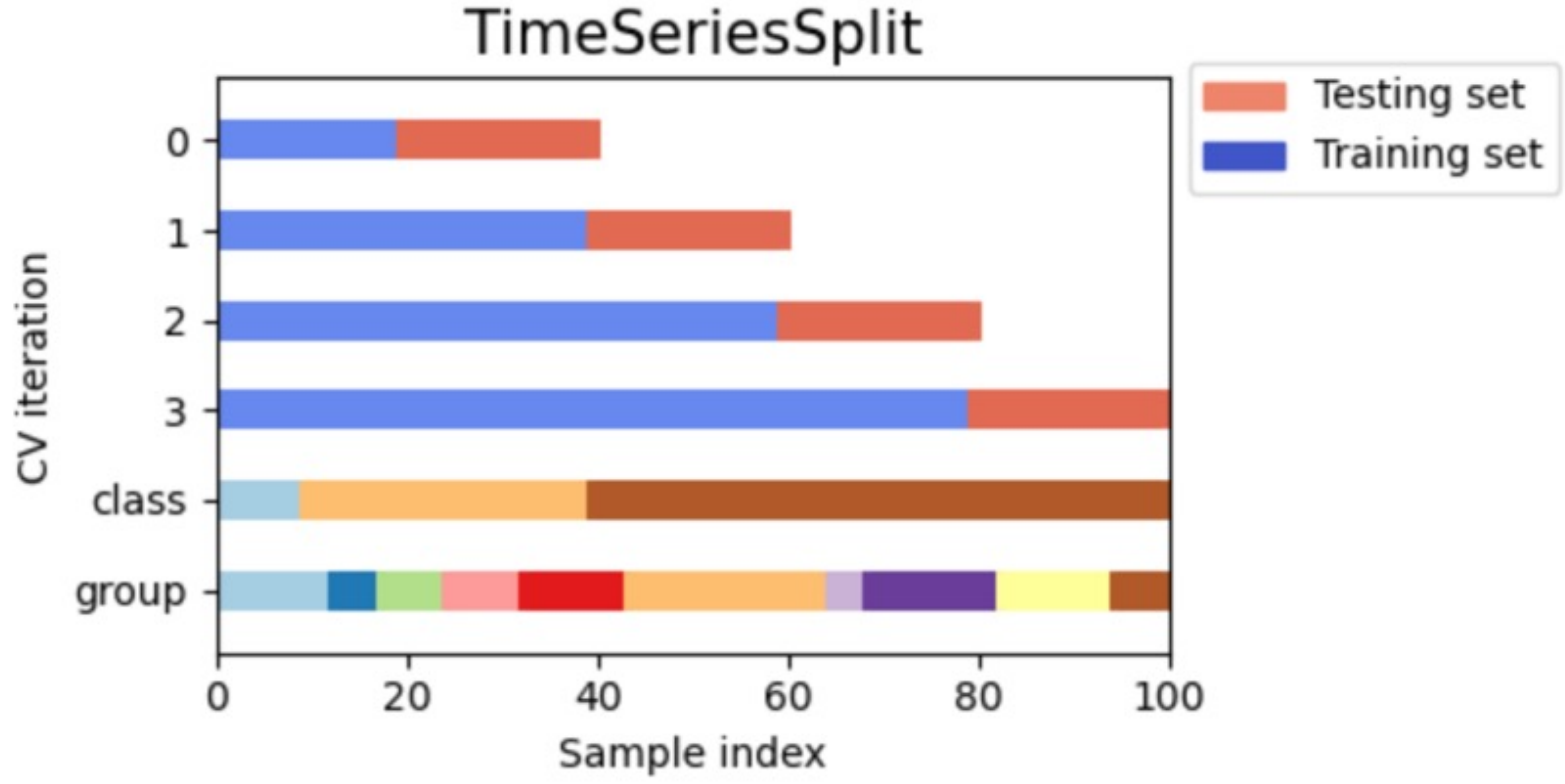


https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html

Class-Ratio Preserving Splitting



Temporal Successive Splitting

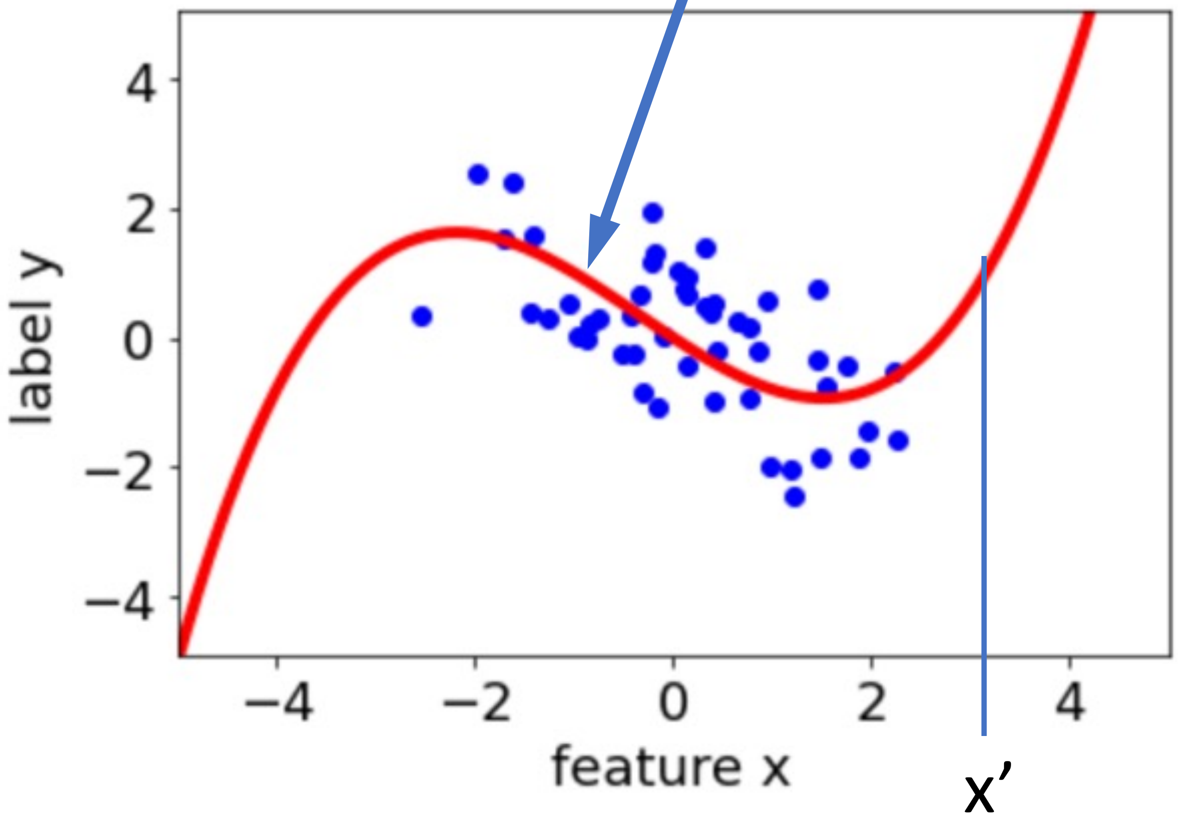


source: https://scikit-learn.org/stable/

Bias and Variance Decomposition

Toy Data

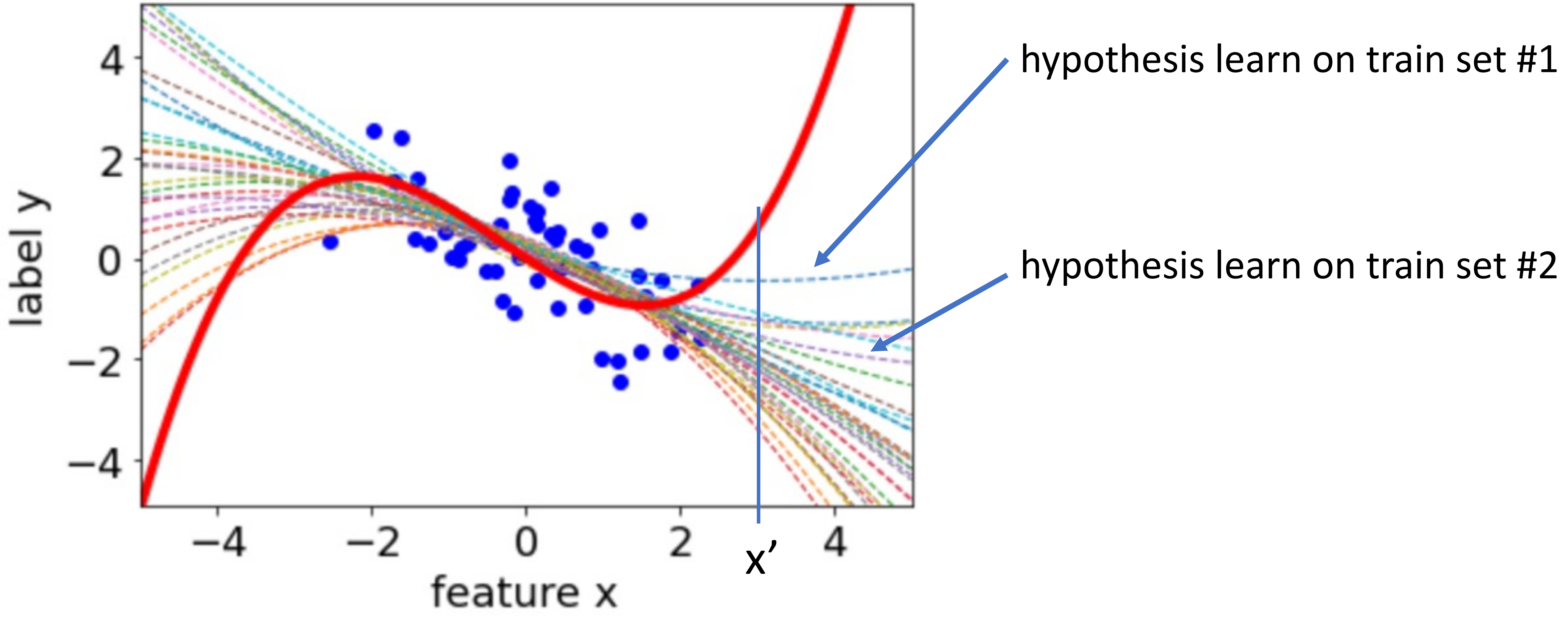
y = g(x) + ”noise”



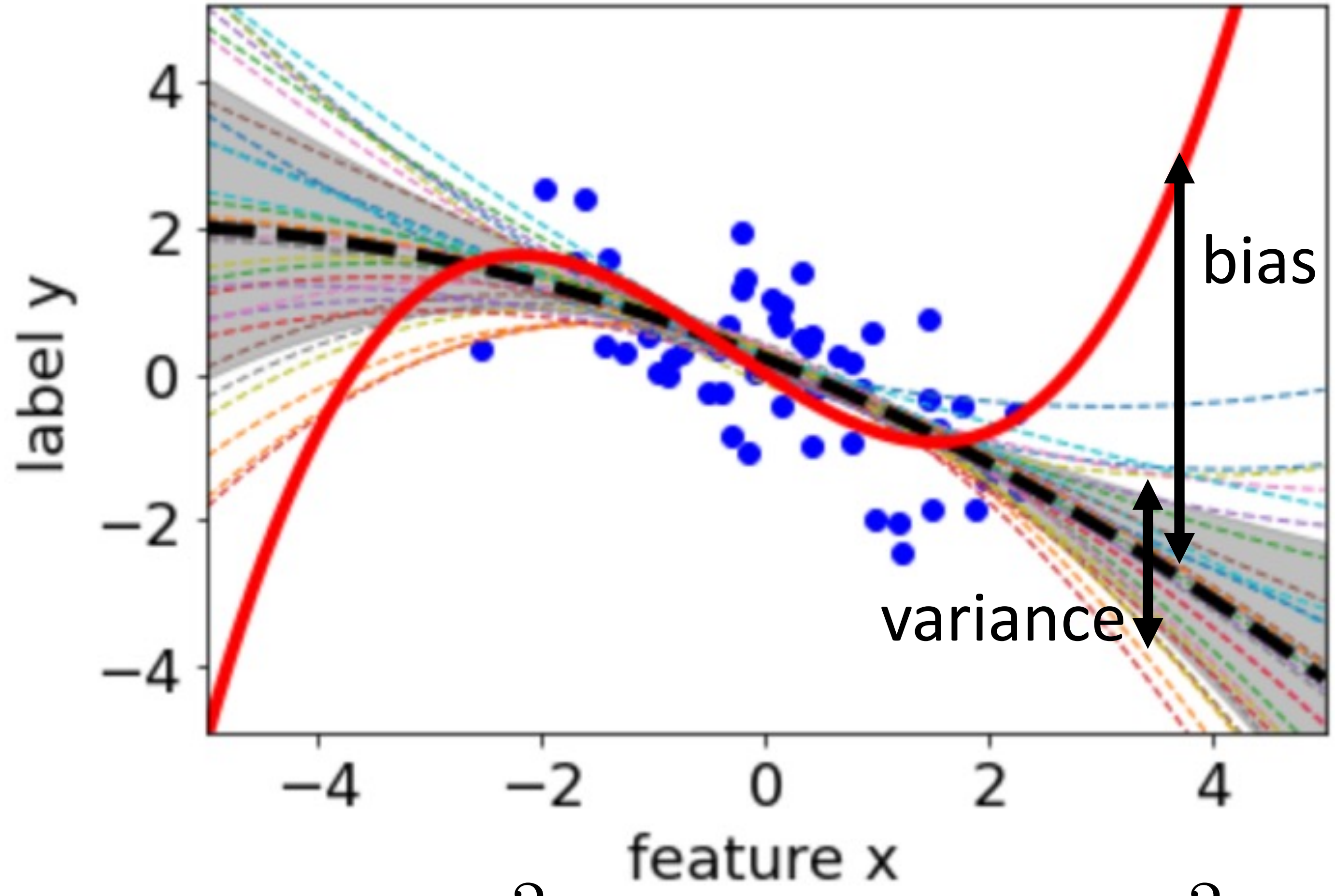
learn hypothesis h(.) using a randomly selected training set

compute prediction h(x’) for a fixed feature value x’

Ensemble of Learnt Hypotheses



Bias and Variance

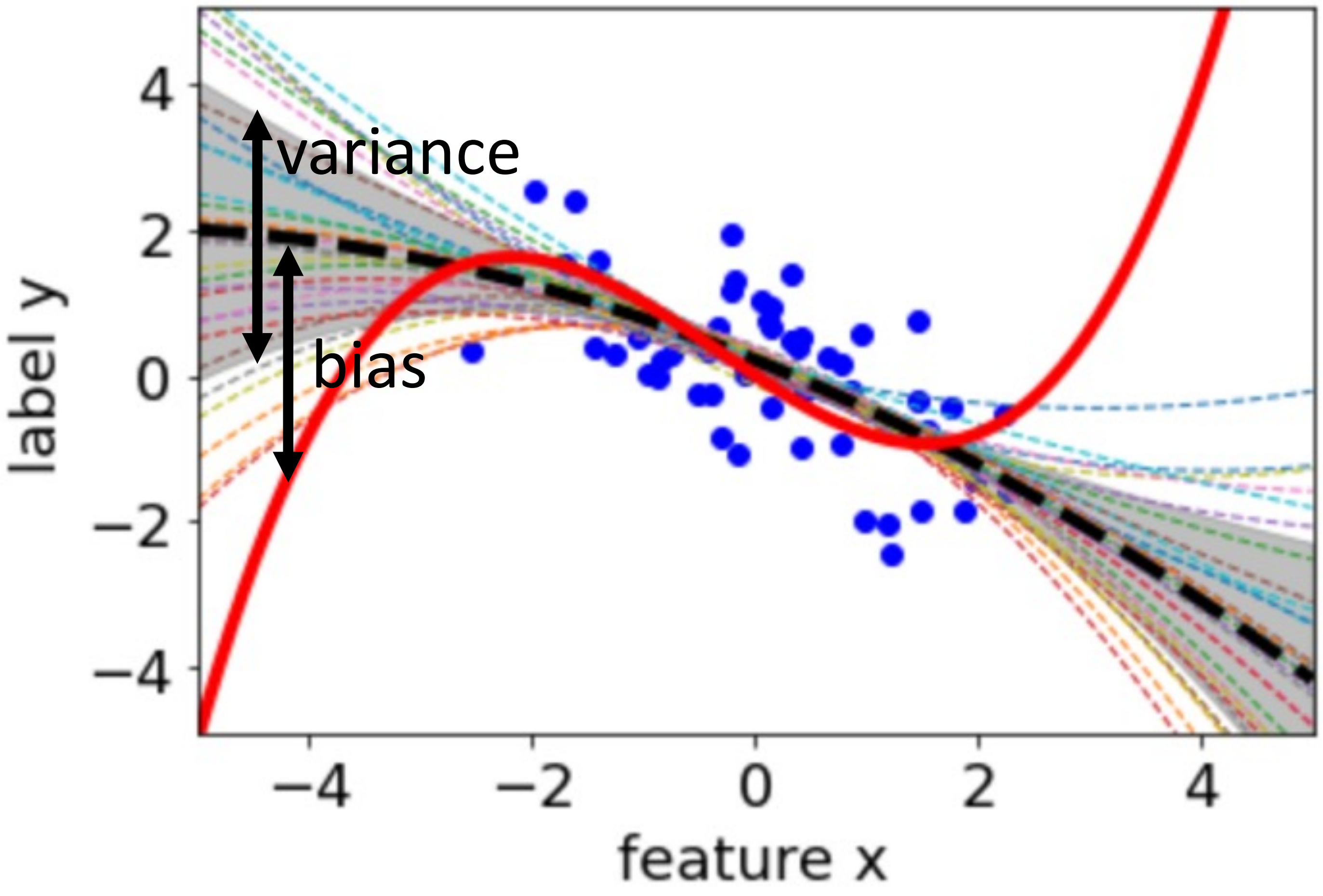


$$\hat{y} = h(x')$$

RV since obtained from a randomly selected training set

$$\mathsf{E}\{(\hat{y}-y)^2\} = (\mathsf{E}\{\hat{y}\}-y)^2 + \mathsf{E}\{(\hat{y}-\mathsf{E}\{\hat{y}\})^2\}$$

Bias and Variance Tradeoff

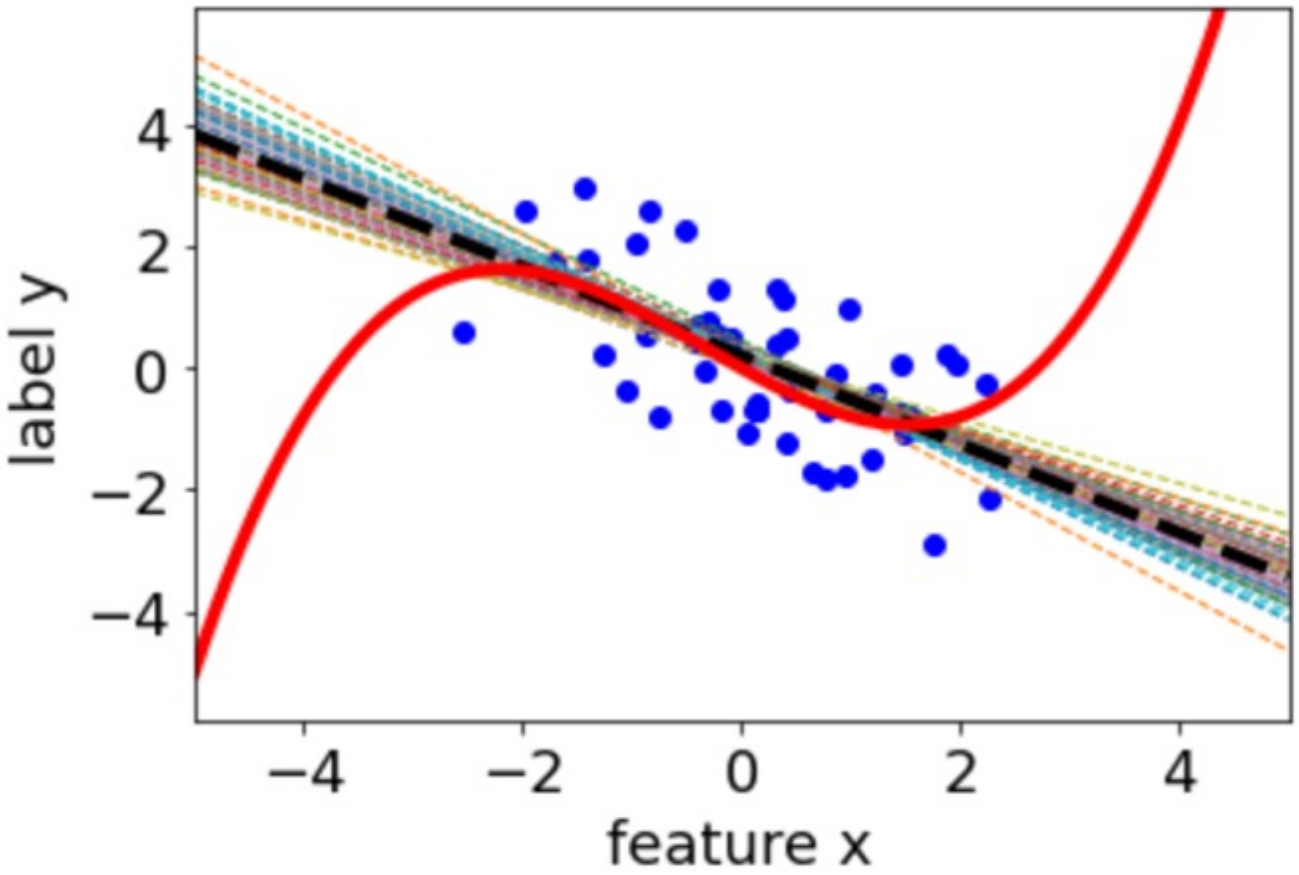


“Prediction Error = Bias + Variance"

bias reduction typically incurs variance increase and vice versa

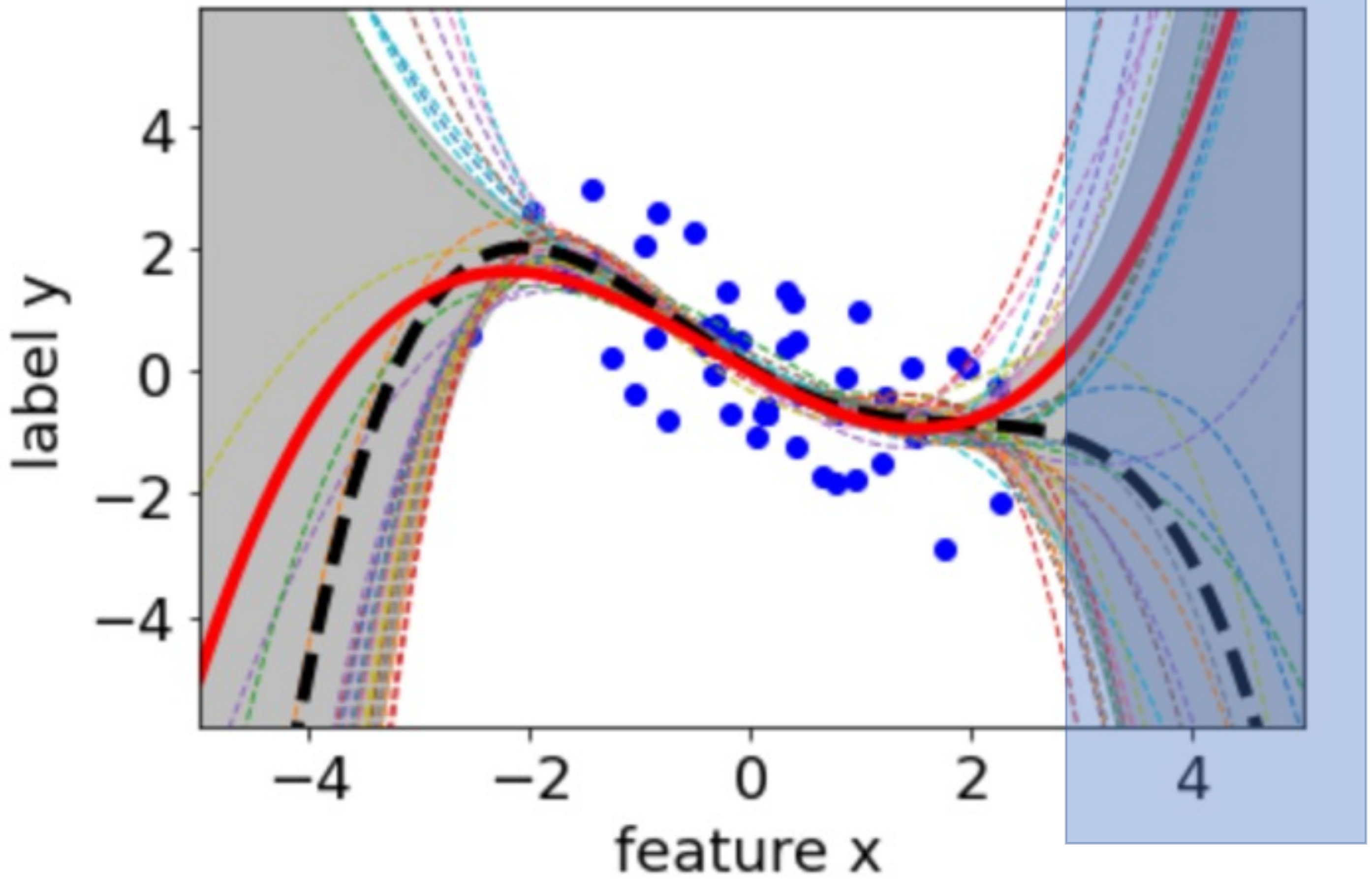
Smaller Model (Poly.Degree)

- small variance
- large bias

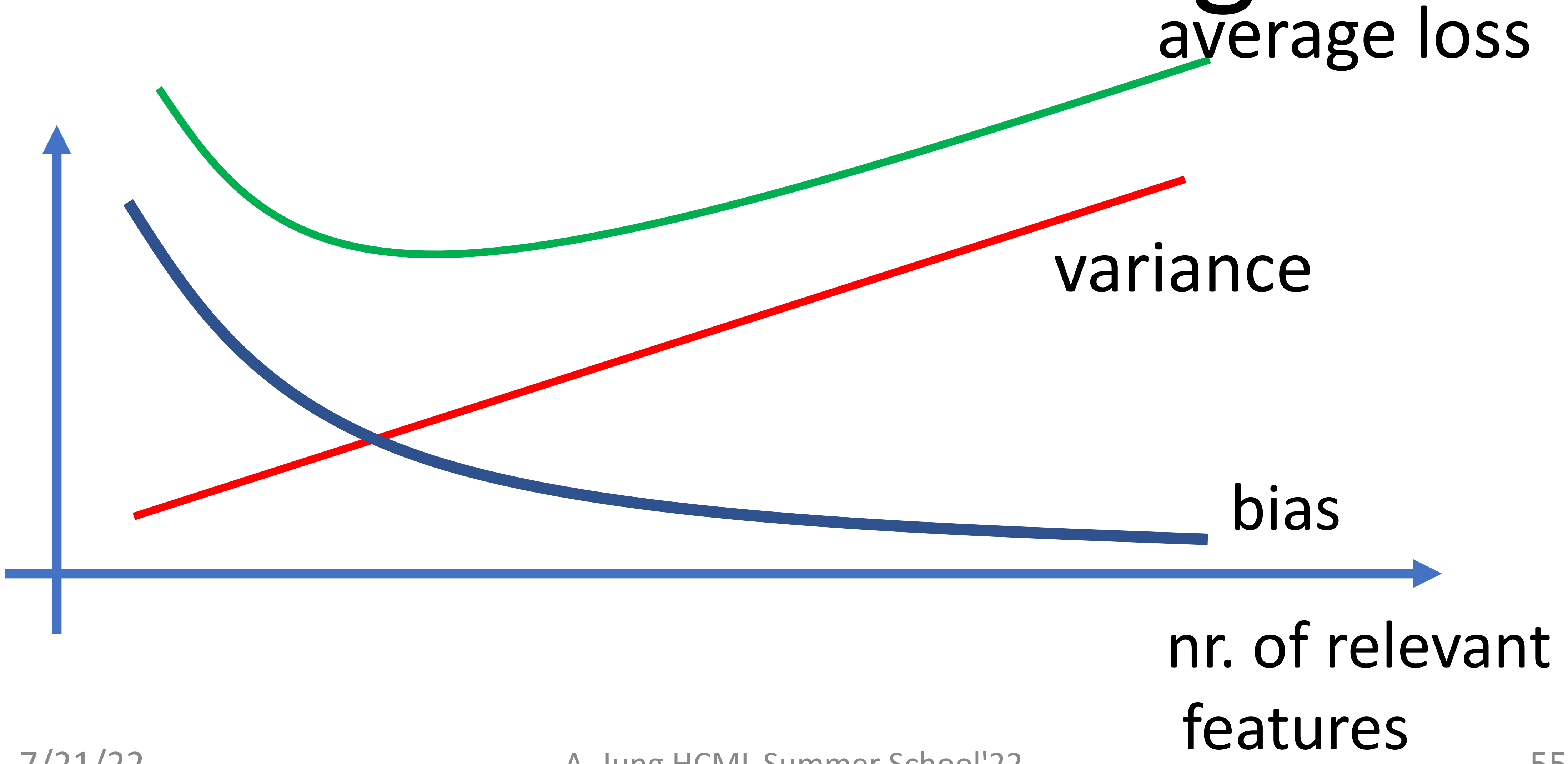


Larger Model (Poly. Degree)

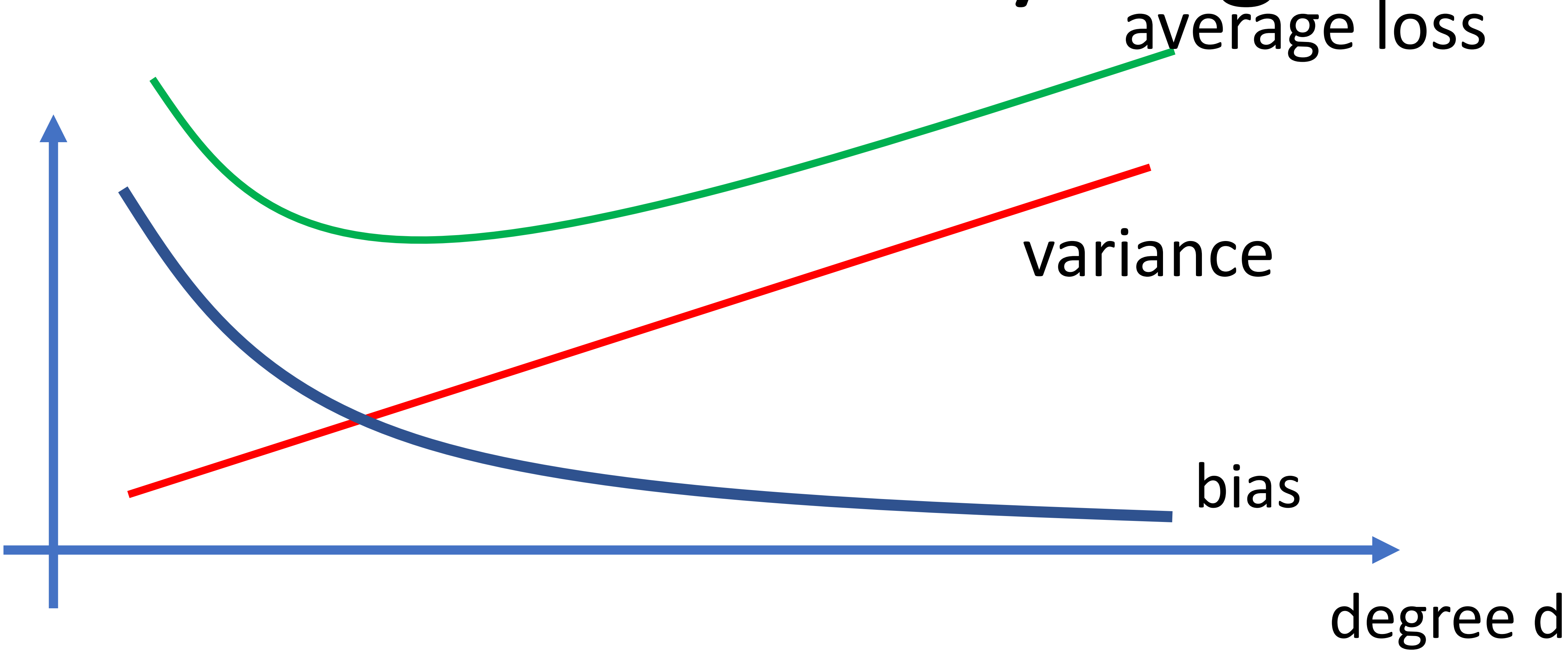
- large variance
- small bias



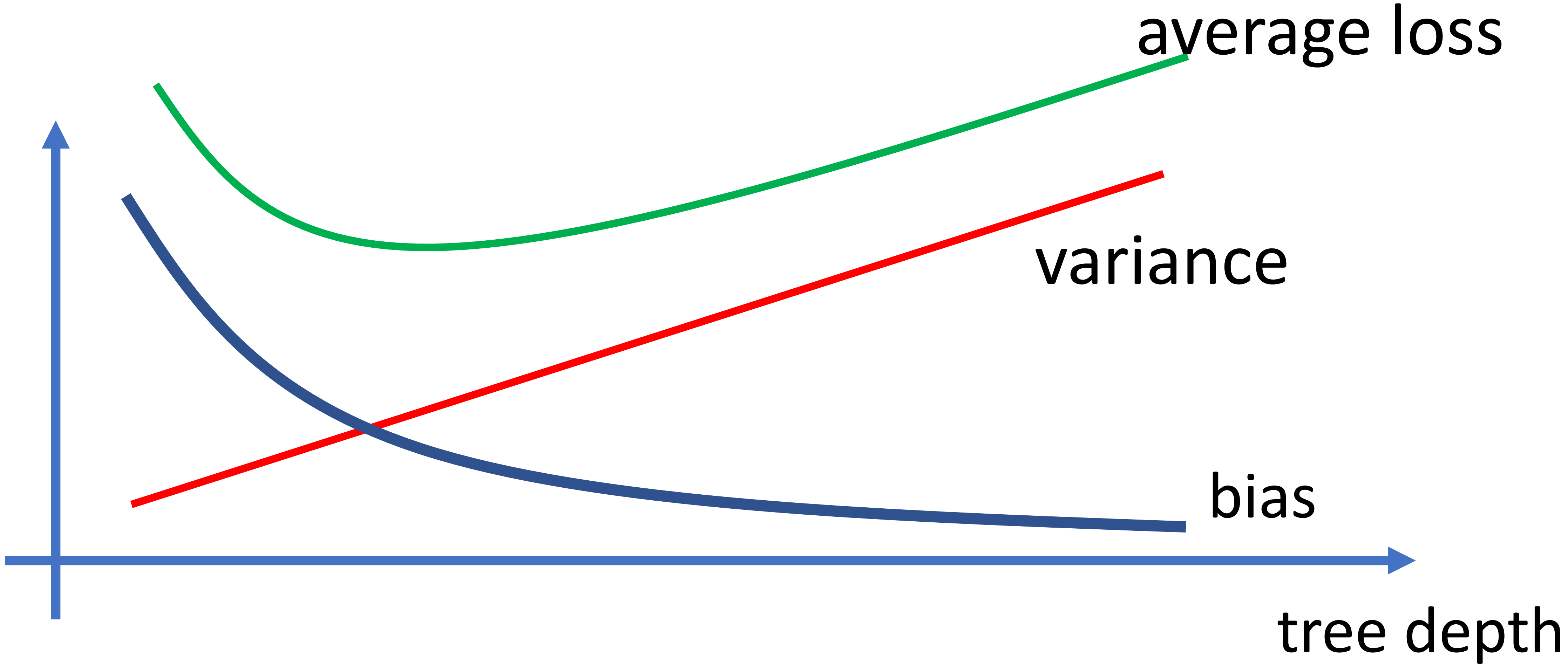
Bias vs. Variance Lin.Reg.



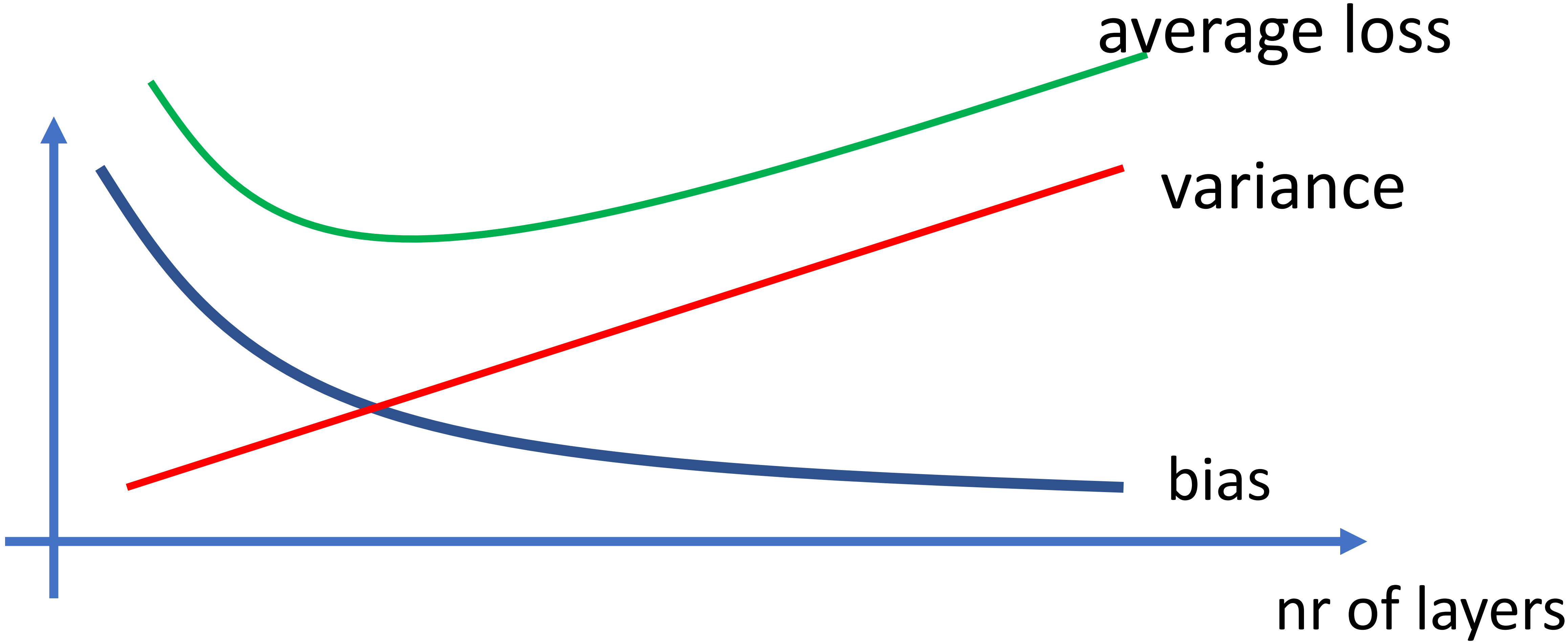
Bias vs. Variance Poly.Reg.



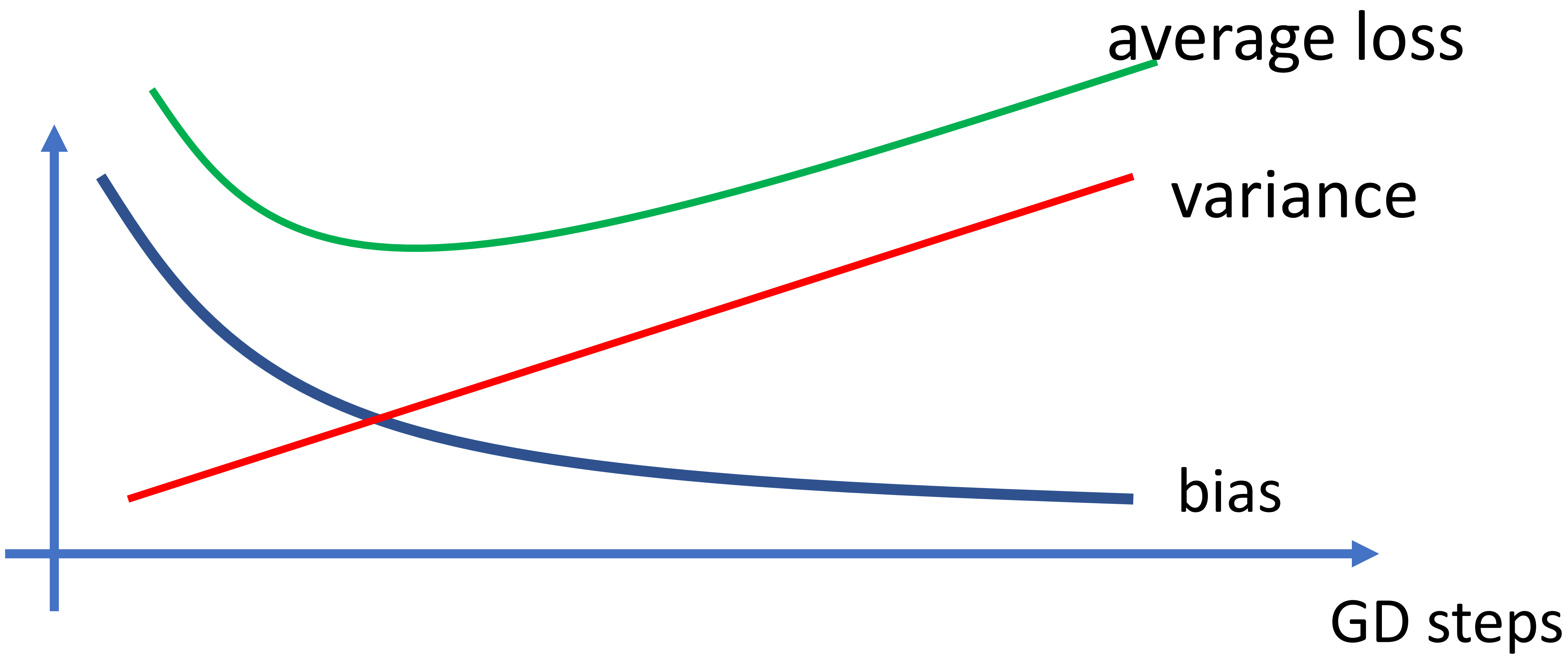
Bias vs. Variance Dec. Tree.



Bias vs. Variance Deep Learning

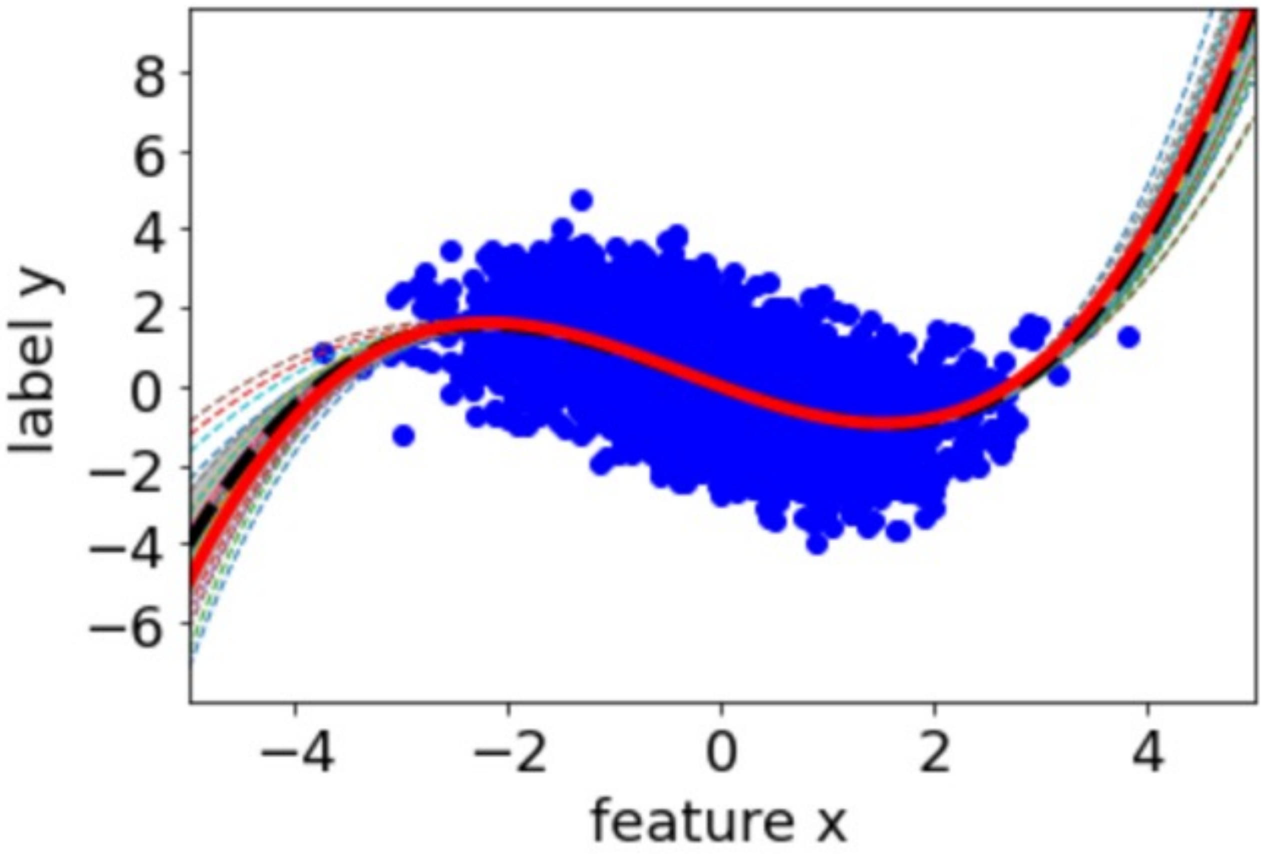


Bias vs. Variance Grad. Desc.



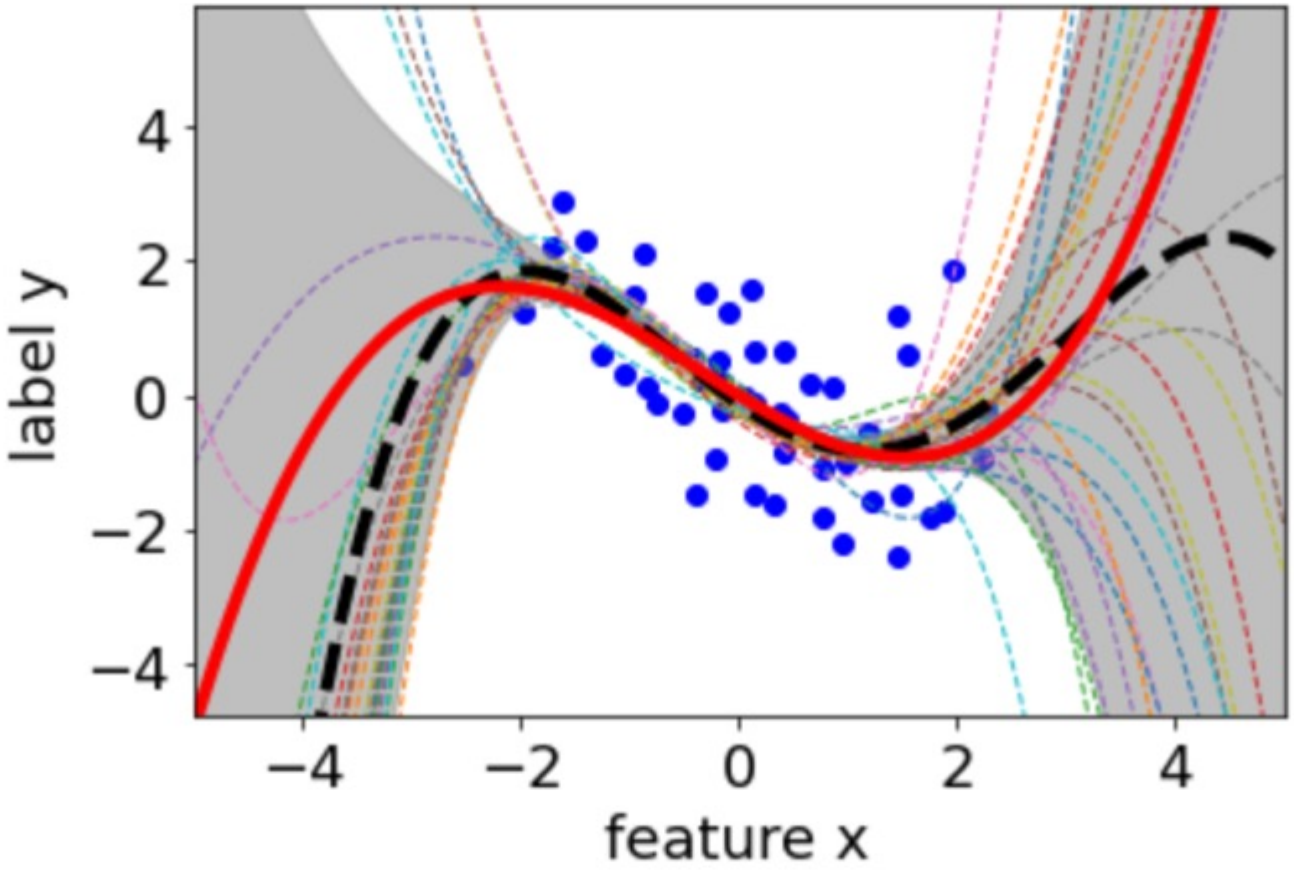
More Data

- small variance

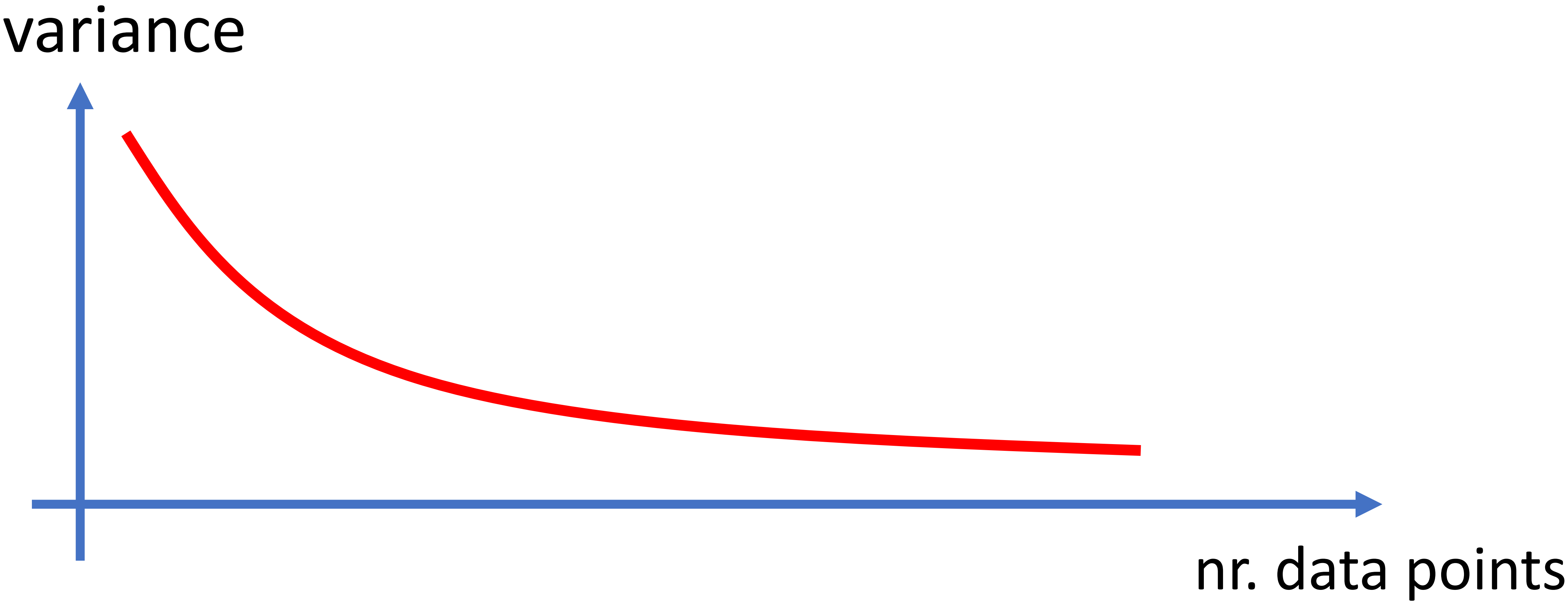


Less Data

- large variance



Learning Curve



Alex' Rule of Thumb

effective number of training data points

>

10 * nr. tunable effective model parameters

stretch the term "effective" as much as possible !

ML Diagnosis

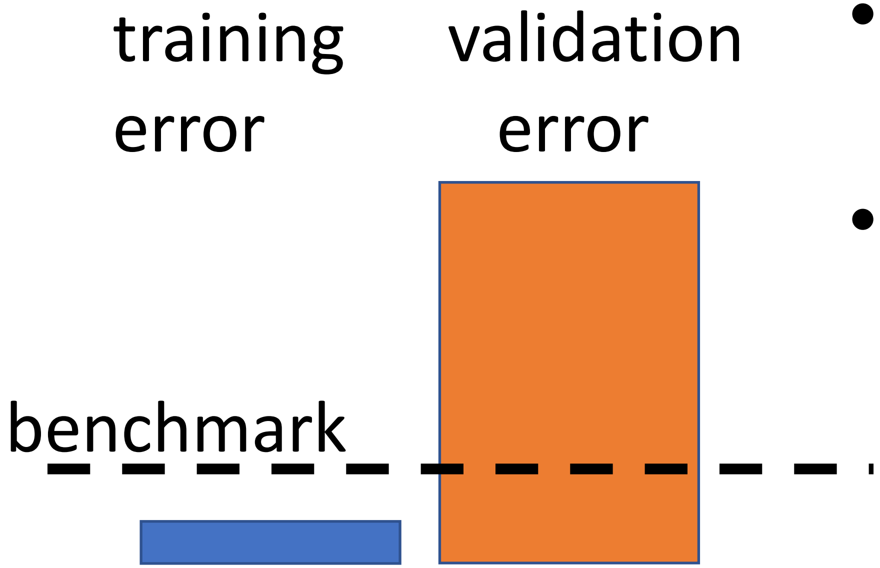
Simple Recipe

- consider ML method with some hypothesis space
- learn hypothesis by min. average loss on train.set
- training error = average loss of learnt hypothesis
- compute validation error
- compare val err, train err with a baseline

Benchmark/Baseline

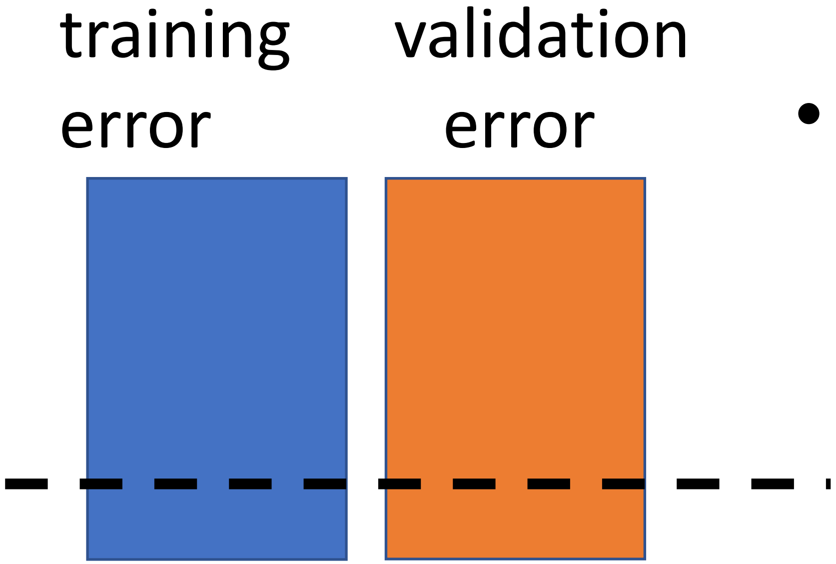
could be obtained from

- probabilistic models
- domain expertise
- existing ML methods
- human performance
- …

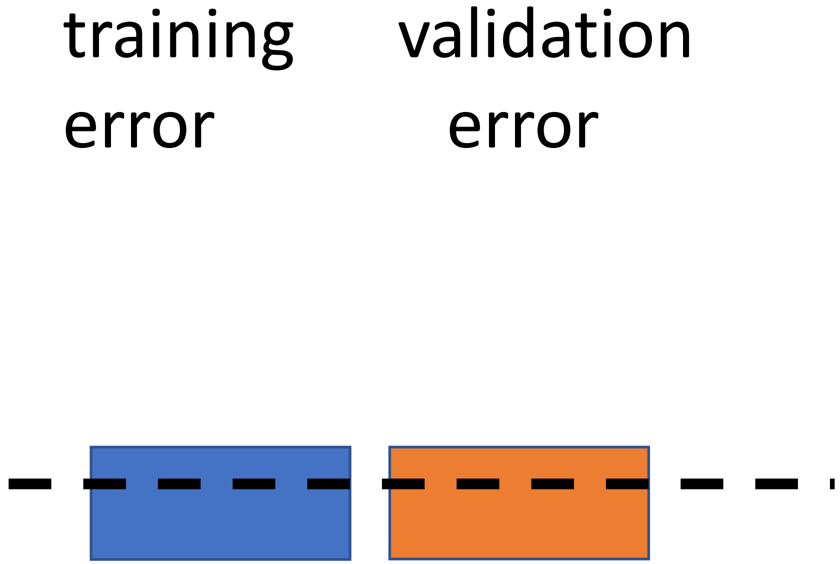


- small train error -> hypothesis space is large
- large val err -> overfitting

- Workaround ?



- large train error -> no good hypothesis found
- Workaround ?



- Case Solved !

Take Home Messages

- large models (e.g. deep nets) often overfit
- small training error does not mean much!
- diagnosis by comparing train/val err
- bias/variance analysis can guide model improvement

Thank You !