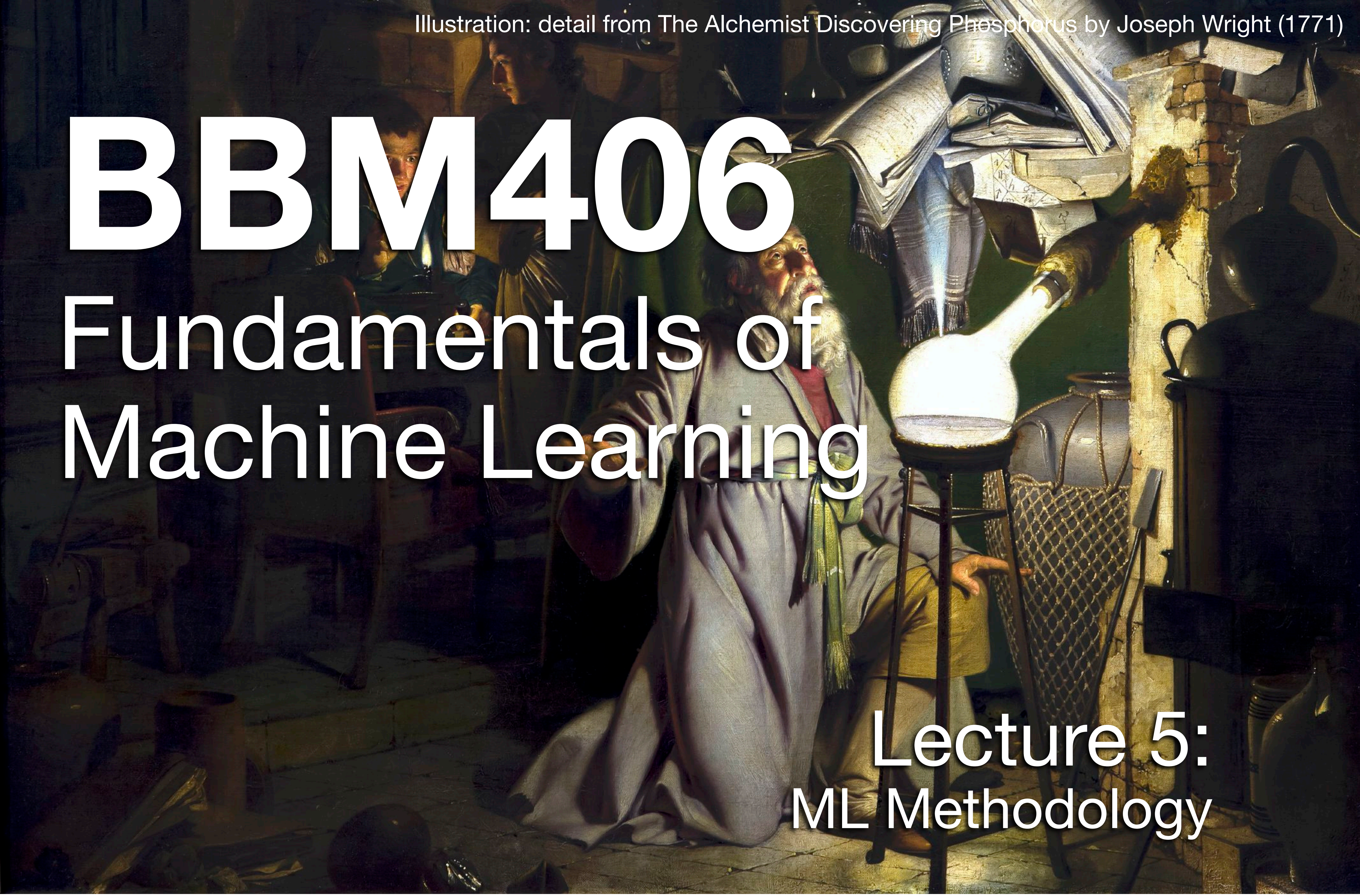


Illustration: detail from The Alchemist Discovering Phosphorus by Joseph Wright (1771)

BBM406

Fundamentals of Machine Learning

Lecture 5:
ML Methodology

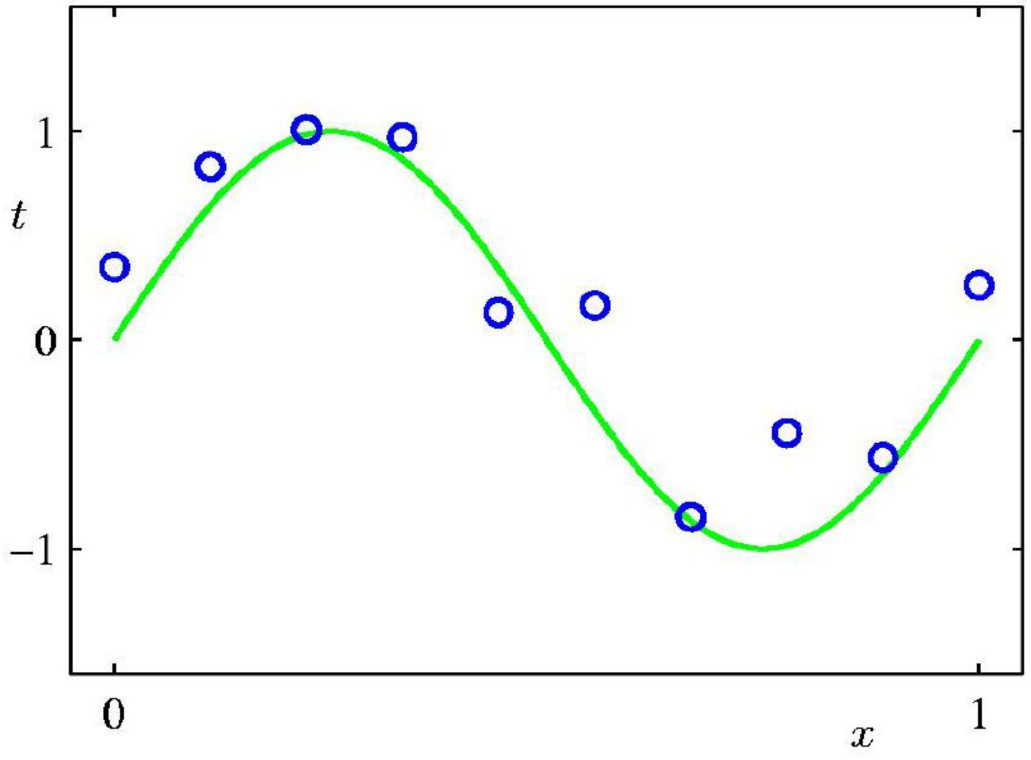
HACETTEPE
UNIVERSITY
COMPUTER
VISION LAB

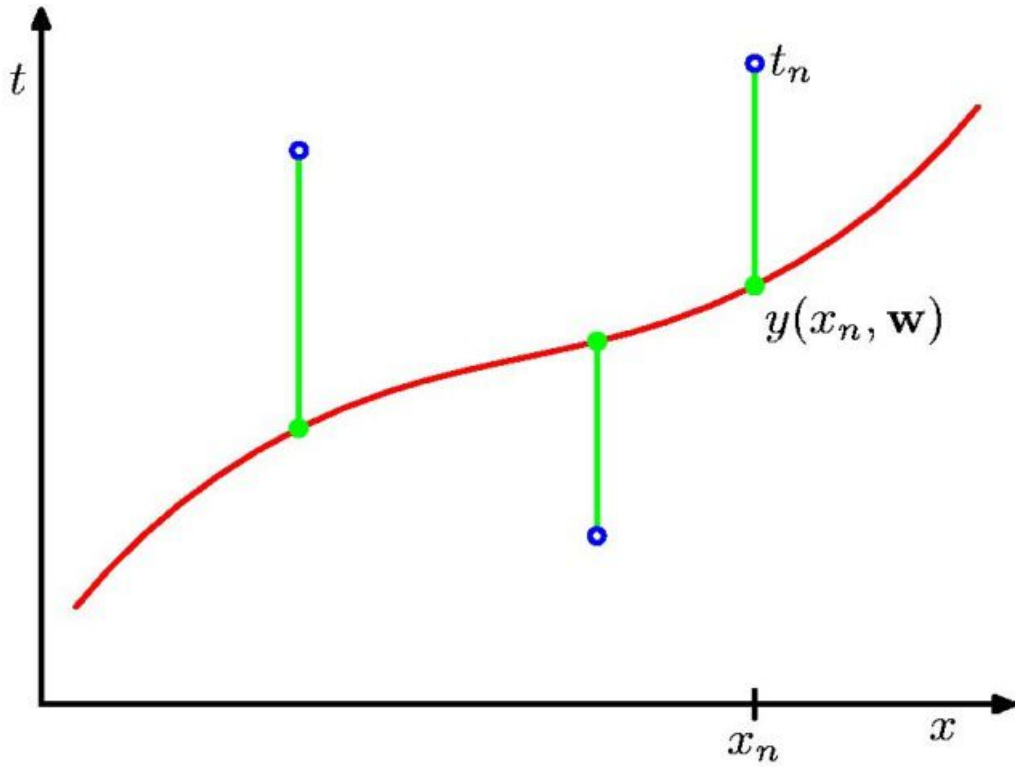
Aykut Erdem // Hacettepe University // Fall 2019

About class projects

- This semester the theme is **machine learning for good.**
- To be done in **groups of 3 people.**
- **Deliverables:** Proposal, blog posts, progress report, project presentations (classroom + video presentations), final report and code
- For more details please check the project webpage: http://web.cs.hacettepe.edu.tr/~aykut/classes/fall2019/bbm406/project.html.

Recall from last time… **Linear Regression**





$$y(x) = w_0 + w_1 x \qquad \mathbf{w} = (w_0, w_1)$$

$$\ell(\mathbf{w}) = \sum_{n=1}^{N} \left[t^{(n)} - (w_0 + w_1 x^{(n)}) \right]^2$$

Gradient Descent Update Rule:

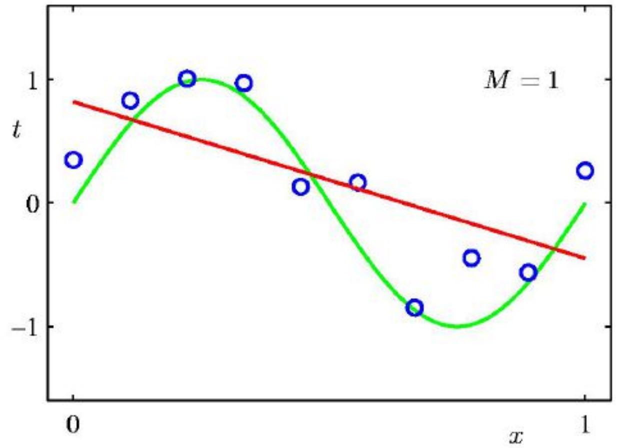
$$\mathbf{w} \leftarrow \mathbf{w} + 2\lambda \left(t^{(n)} - y(x^{(n)}) \right) x^{(n)}$$

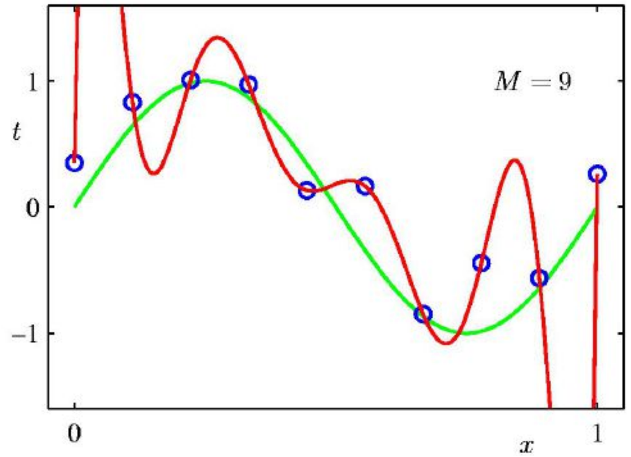
Closed Form Solution:

$$\mathbf{w} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{t}$$

Recall from last time… **Some key concepts**

- Data fits – is linear model best (**model selection**)?
  - Simplest models do not capture all the important variations (signal) in the data: **underfit**
  - More complex model may **overfit** the training data (fit not only the signal but also the **noise** in the data), especially if not enough data to constrain model



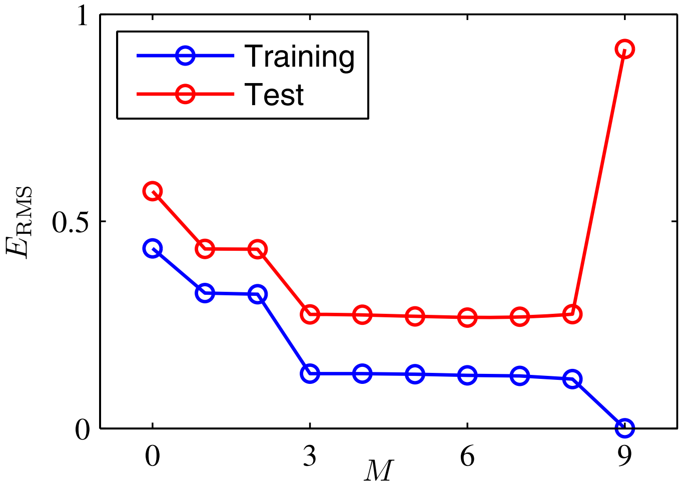


- One method of assessing fit:
  - test **generalization** = model's ability to predict the held out data
- **Regularization**

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 \boxed{+ \frac{\lambda}{2}\|\mathbf{w}\|^2}$$

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01



slide by Richard Zemel

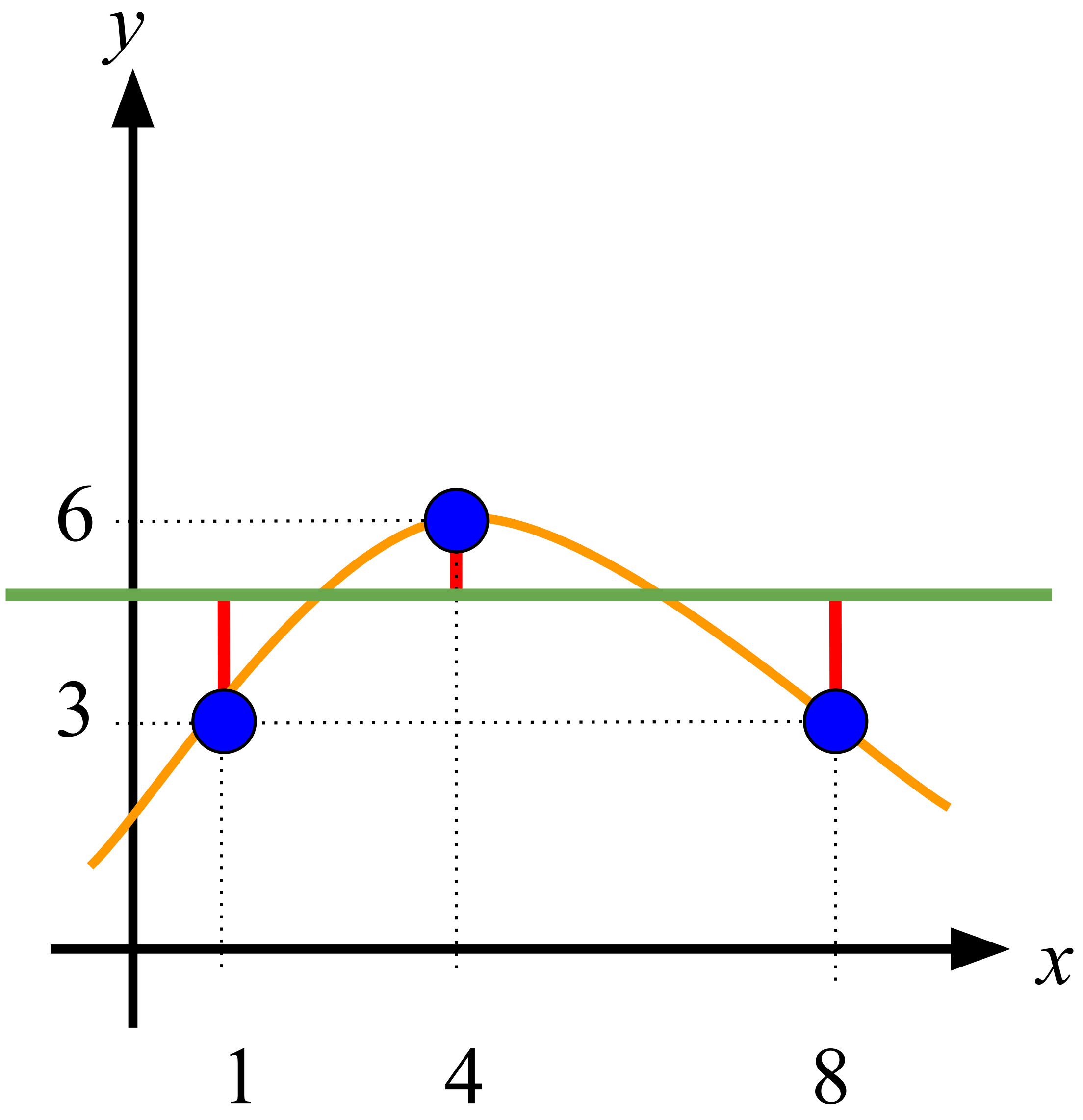
Today

- Machine Learning Methodology
  - validation
  - cross-validation (k-fold, leave-one-out)
  - model selection

Machine Learning Methodology

Recap: Regression

- In regression, labels y^i are continuous
- Classification/regression are solved very similarly
- Everything we have done so far transfers to classification with very minor changes
- Error: sum of distances from examples to the fitted model



slide by Olga Veksler

Training/Test Data Split

- Talked about splitting data in training/test sets
  - training data is used to fit parameters
  - test data is used to assess how classifier generalizes to new data
- What if classifier has “non-tunable” parameters?
  - a parameter is “non-tunable” if tuning (or training) it on the training data leads to overfitting
  - Examples:
    - k in kNN classifier
    - number of hidden units in MNN
    - number of hidden layers in MNN
    - etc …

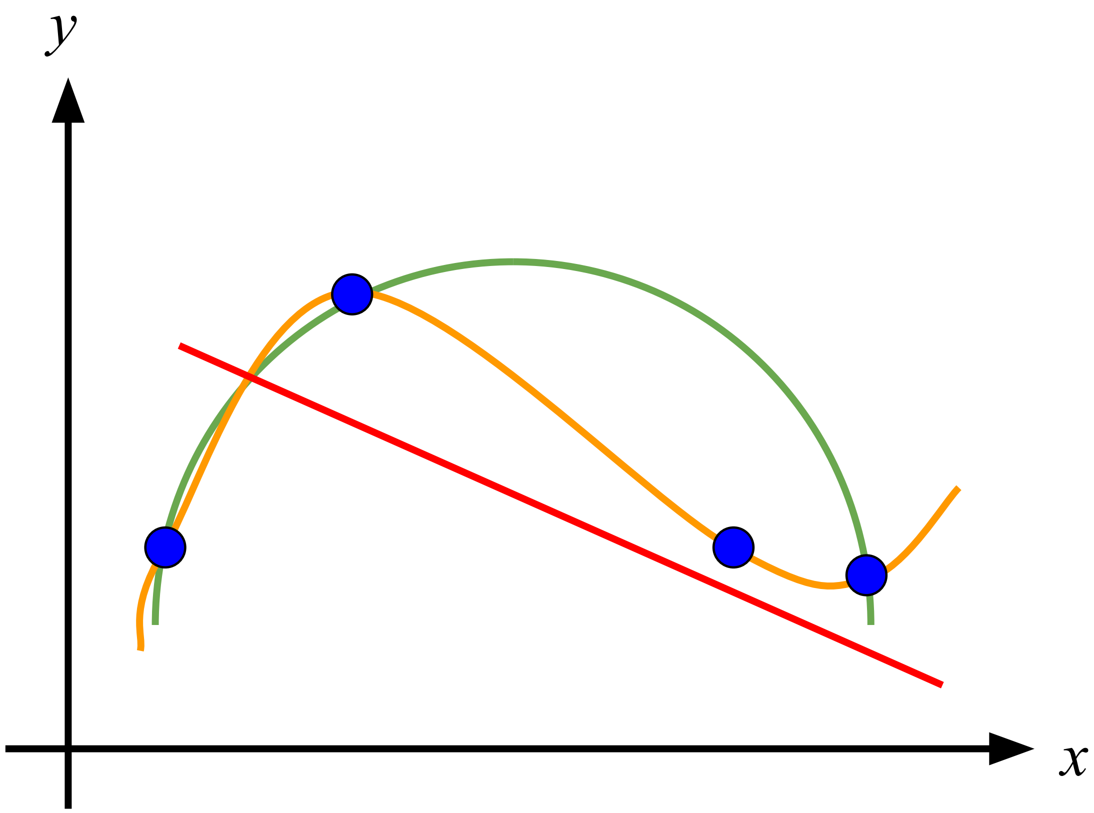
slide by Olga Veksler

Example of Overfitting

- Want to fit a polynomial machine $f(\mathbf{x},\mathbf{w})$

- Instead of fixing polynomial degree, make it parameter $\mathbf{d}$
  - learning machine $f(\mathbf{x},\mathbf{w},\mathbf{d})$

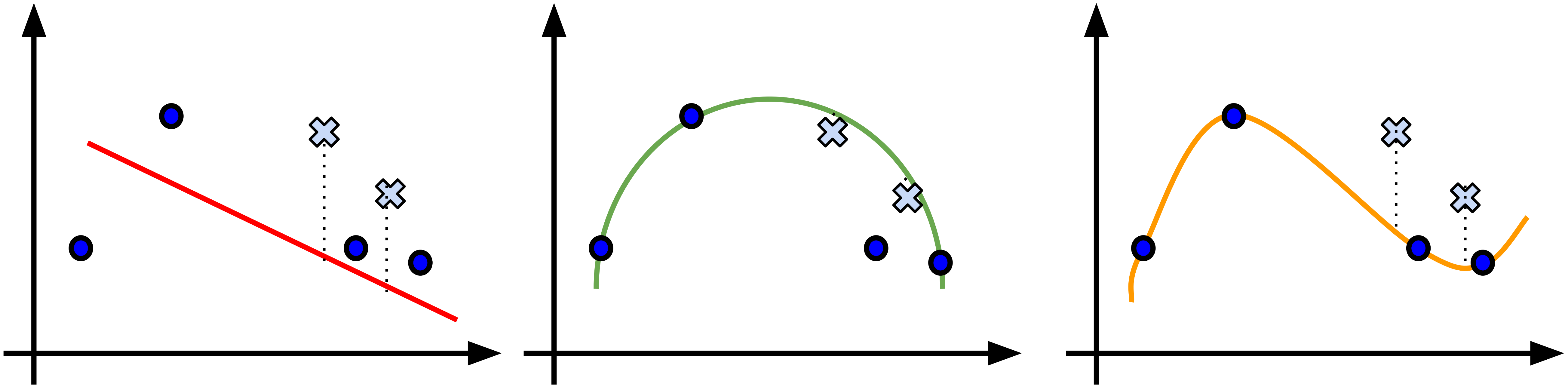
- Consider just three choices for $\mathbf{d}$
  - degree 1
  - degree 2
  - degree 3



- Training error is a bad measure to choose $\mathbf{d}$
  - degree 3 is the best according to the training error, but overfits the data

slide by Olga Veksler

Training/Test Data Split



- What about test error? Seems appropriate
  - degree 2 is the best model according to the test error
- Except what do we report as the test error now?
- Test error should be computed on data that was **not used for training at all!**
- Here used “test” data for training, *i.e.* choosing model

slide by Olga Veksler

Validation data

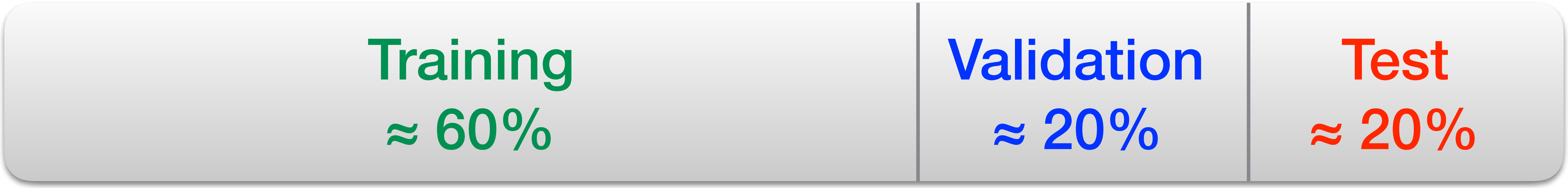
- Same question when choosing among several classifiers
  - our polynomial degree example can be looked at as choosing among 3 classifiers (degree 1, 2, or 3)

slide by Olga Veksler

Validation data

- Same question when choosing among several classifiers
  - our polynomial degree example can be looked at as choosing among 3 classifiers (degree 1, 2, or 3)
- Solution: split the labeled data into three parts

labeled data

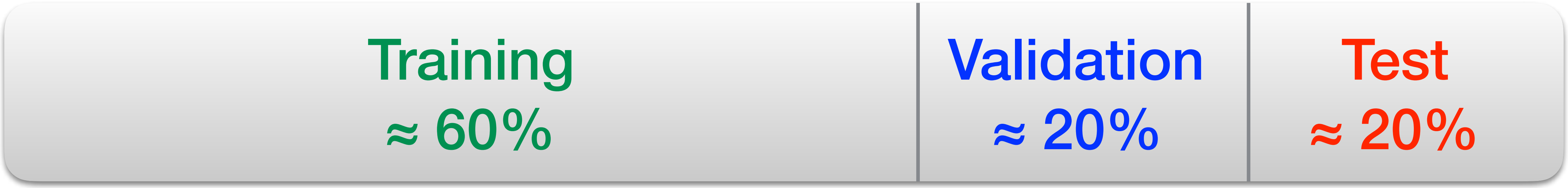


Training ≈ 60%	Validation ≈ 20%	Test ≈ 20%
train tunable parameters **w**	train other parameters, or to select classifier	use **only** to assess final performance

slide by Olga Veksler

Training/Validation

labeled data



Training
≈ 60%

Validation
≈ 20%

Test
≈ 20%

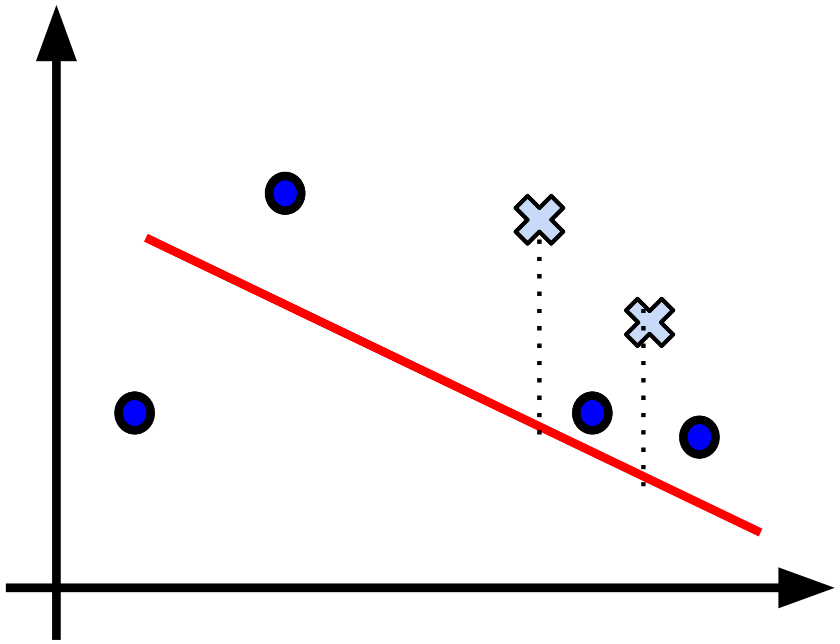
Training error: computed on training example

Validation error: computed on validation examples

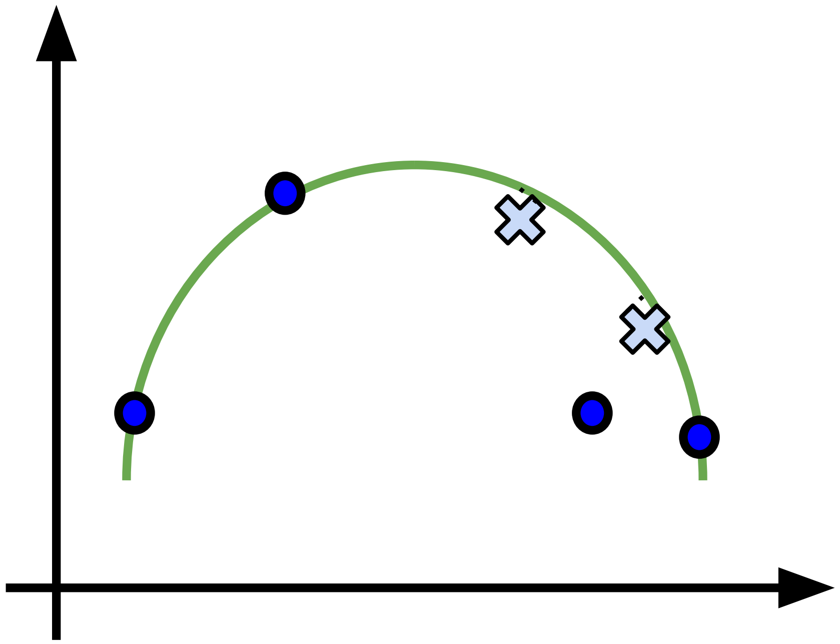
Test error: computed on test examples

slide by Olga Veksler

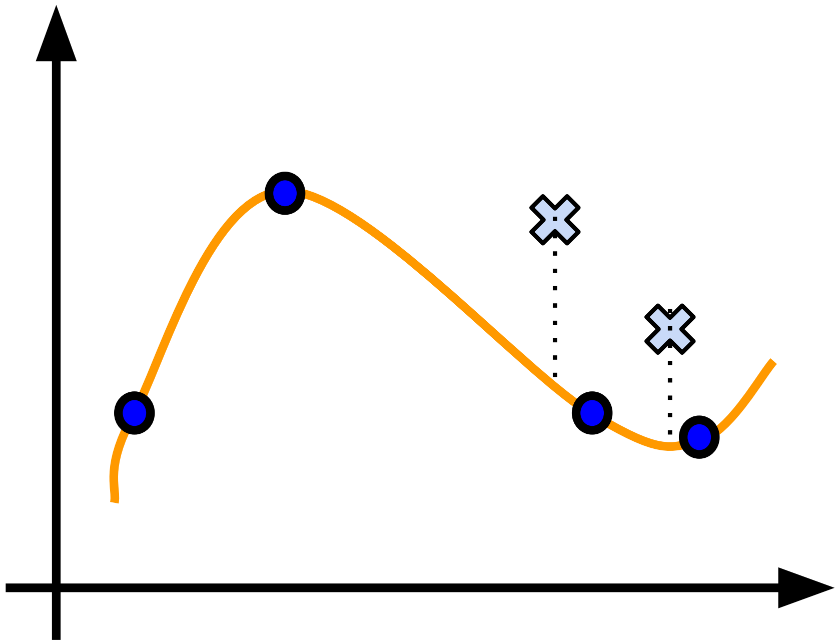
Training/Validation/Test Data



validation error: 3.3



validation error: 1.8

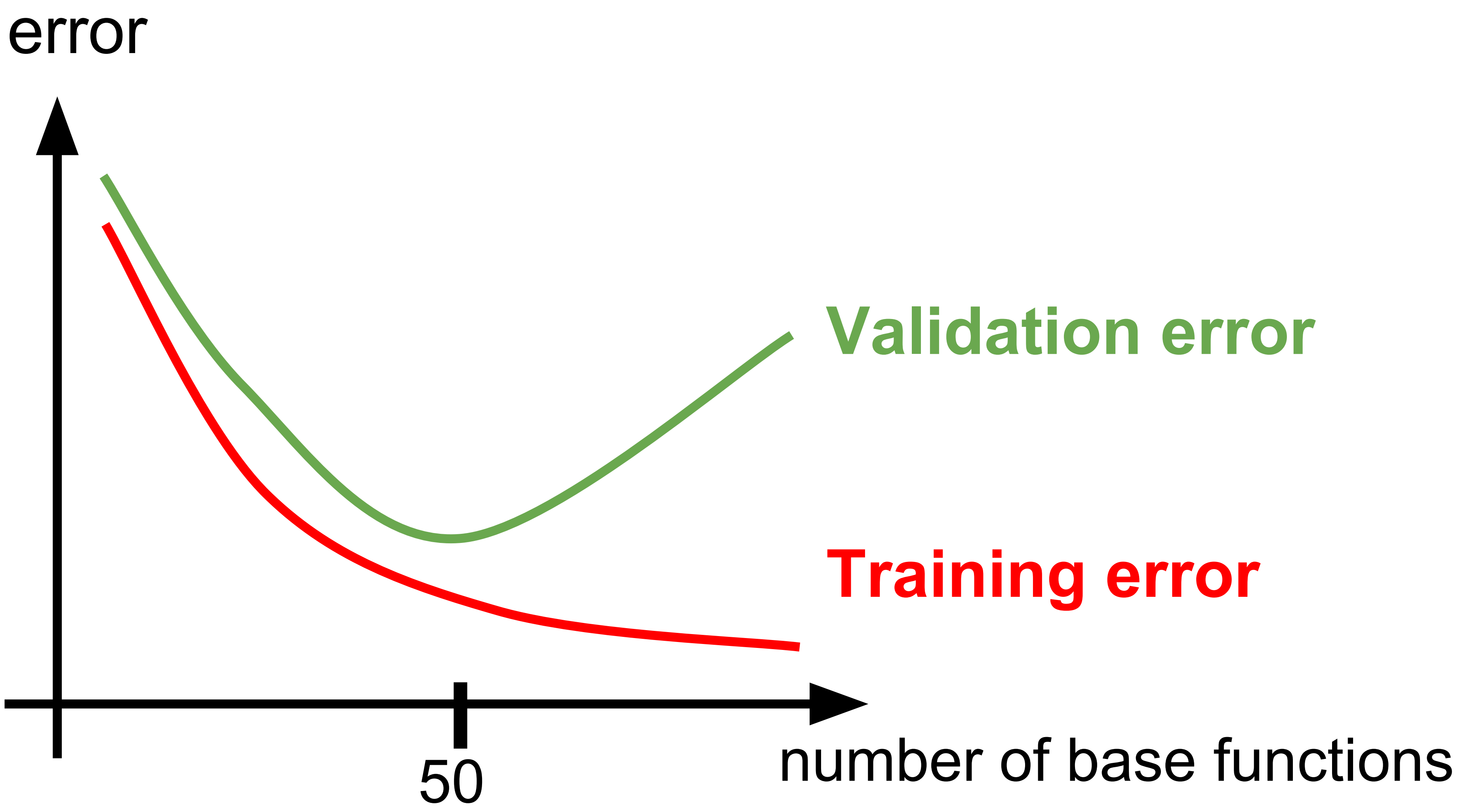


validation error: 3.4

- Training Data
- Validation Data
  - $\mathbf{d} = 2$ is chosen
- Test Data
  - 1.3 test error computed for $\mathbf{d} = 2$

slide by Olga Veksler

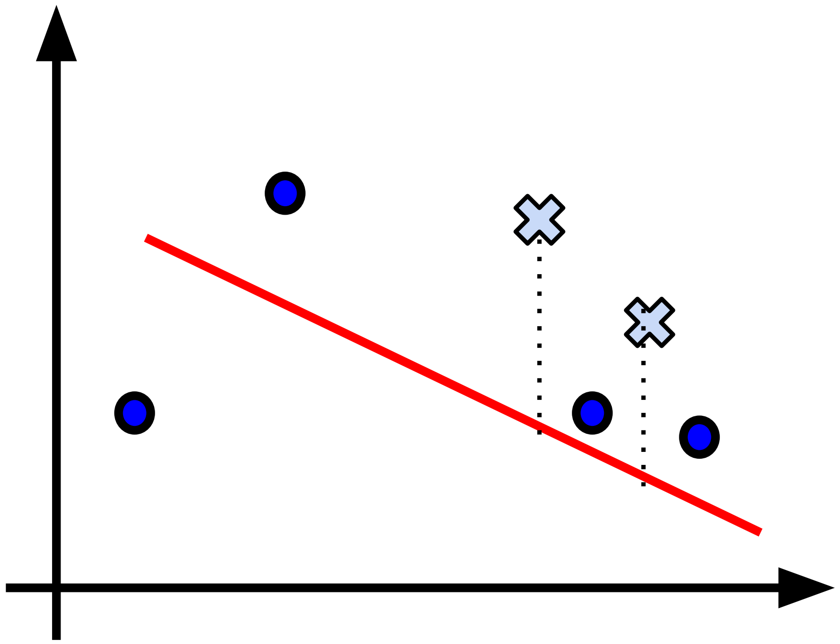
Choosing Parameters: Example



- Need to choose number of hidden units for a MNN
  - The more hidden units, the better can fit training data
  - But at some point we overfit the data

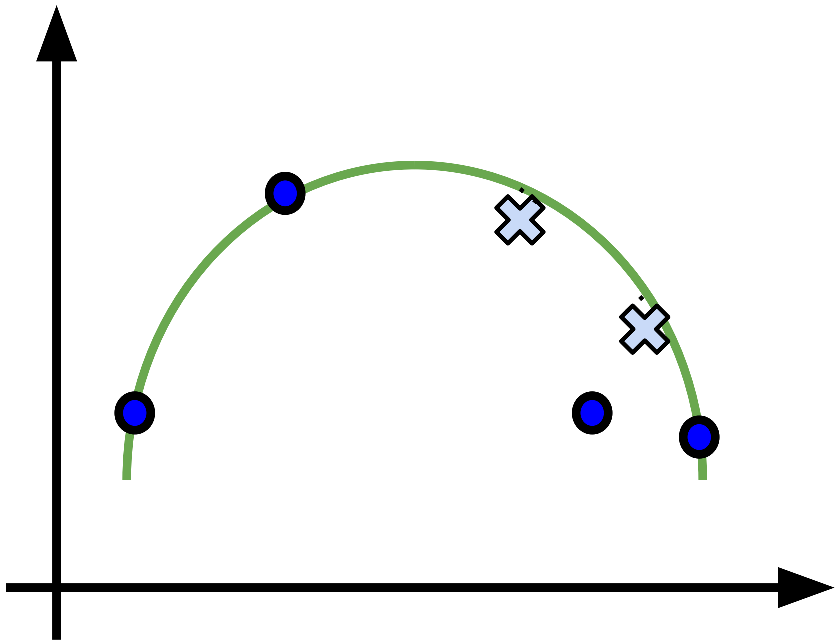
slide by Olga Veksler

Diagnosing Underfitting/Overfitting



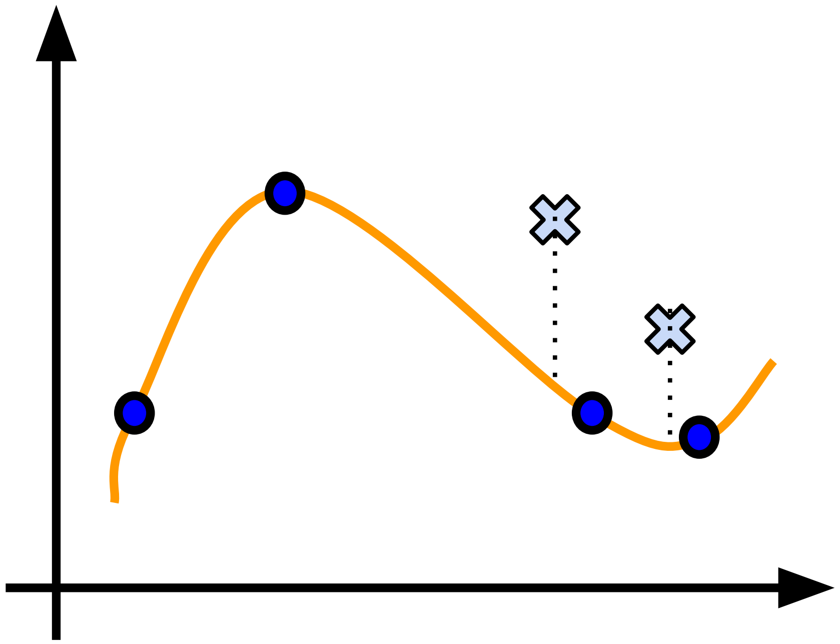
Underfitting

- large training error
- large validation error



Just Right

- small training error
- small validation error



Overfitting

- small training error
- large validation error

slide by Olga Veksler

Fixing Underfitting/Overfitting

- Fixing Underfitting
  - getting more training examples will not help
  - get more features
  - try more complex classifier
    - if using MLP, try more hidden units

- Fixing Overfitting
  - getting more training examples might help
  - try smaller set of features
  - Try less complex classifier
    - If using MLP, try less hidden units

slide by Olga Veksler

Train/Test/Validation Method

- Good news
  - Very simple

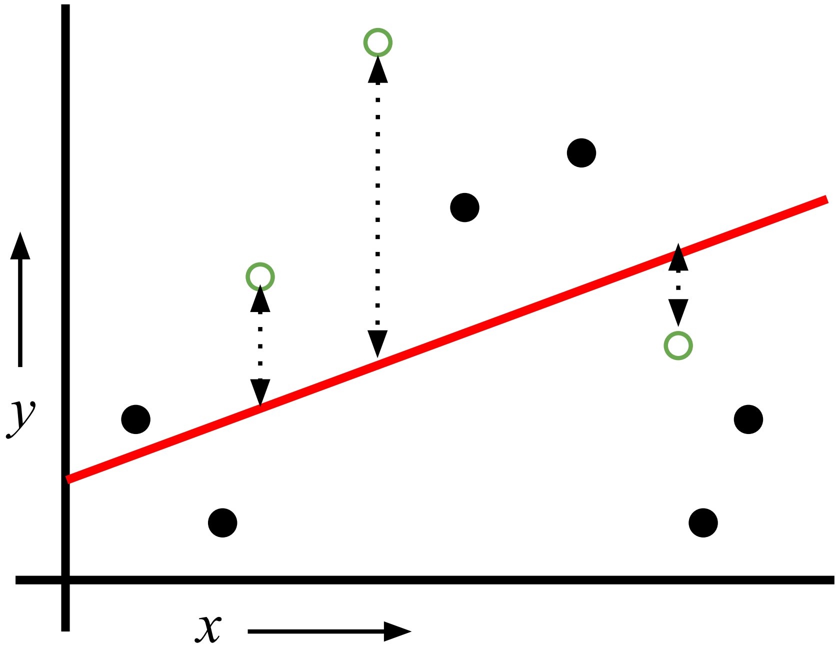
- Bad news:
  - Wastes data
    - in general, the more data we have, the better are the estimated parameters
    - we estimate parameters on 40% less data, since 20% removed for test and 20% for validation data
  - If we have a small dataset our test (validation) set might just be lucky or unlucky

- **Cross Validation is a method for performance evaluation that wastes less data**

slide by Olga Veksler

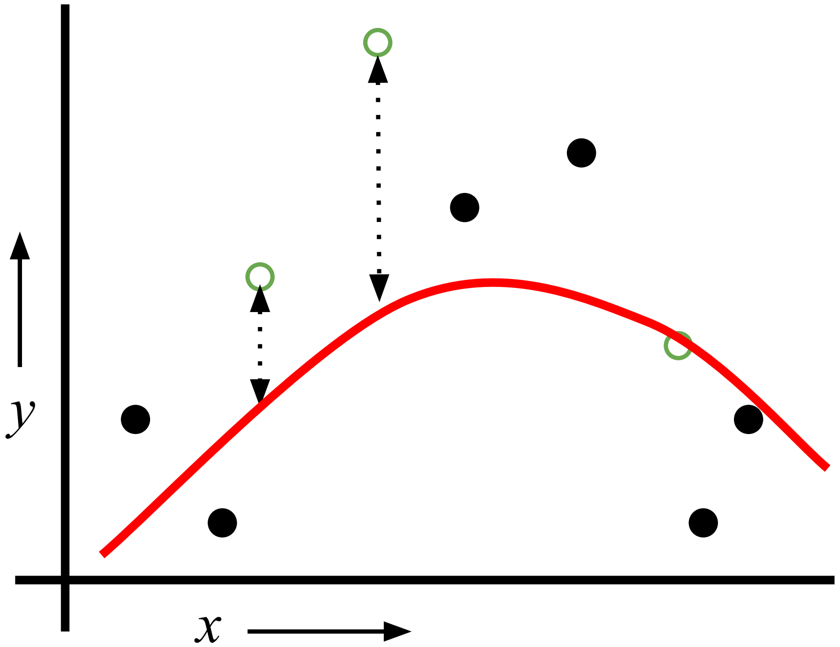
Small Dataset

Linear Model:



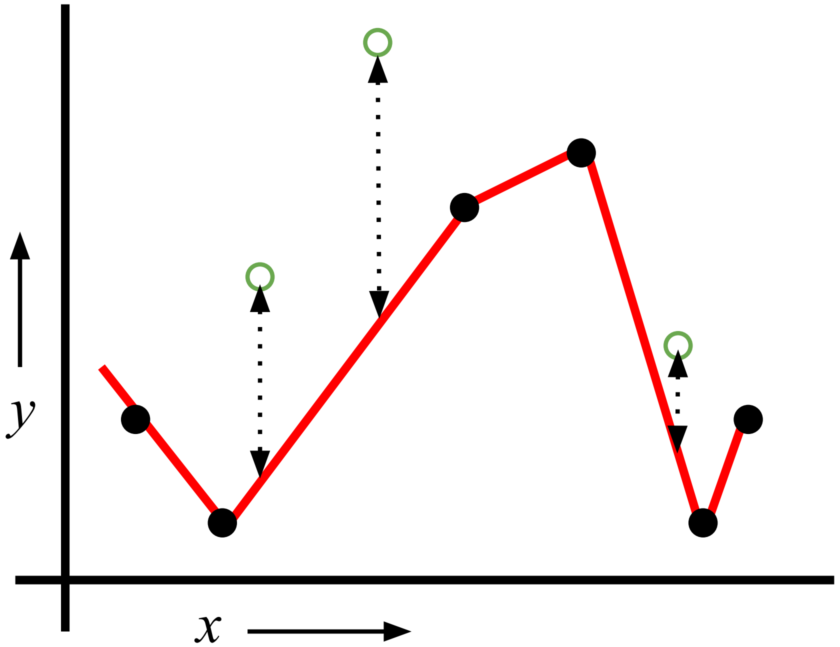
Mean Squared Error = 2.4

Quadratic Model:



Mean Squared Error = 0.9

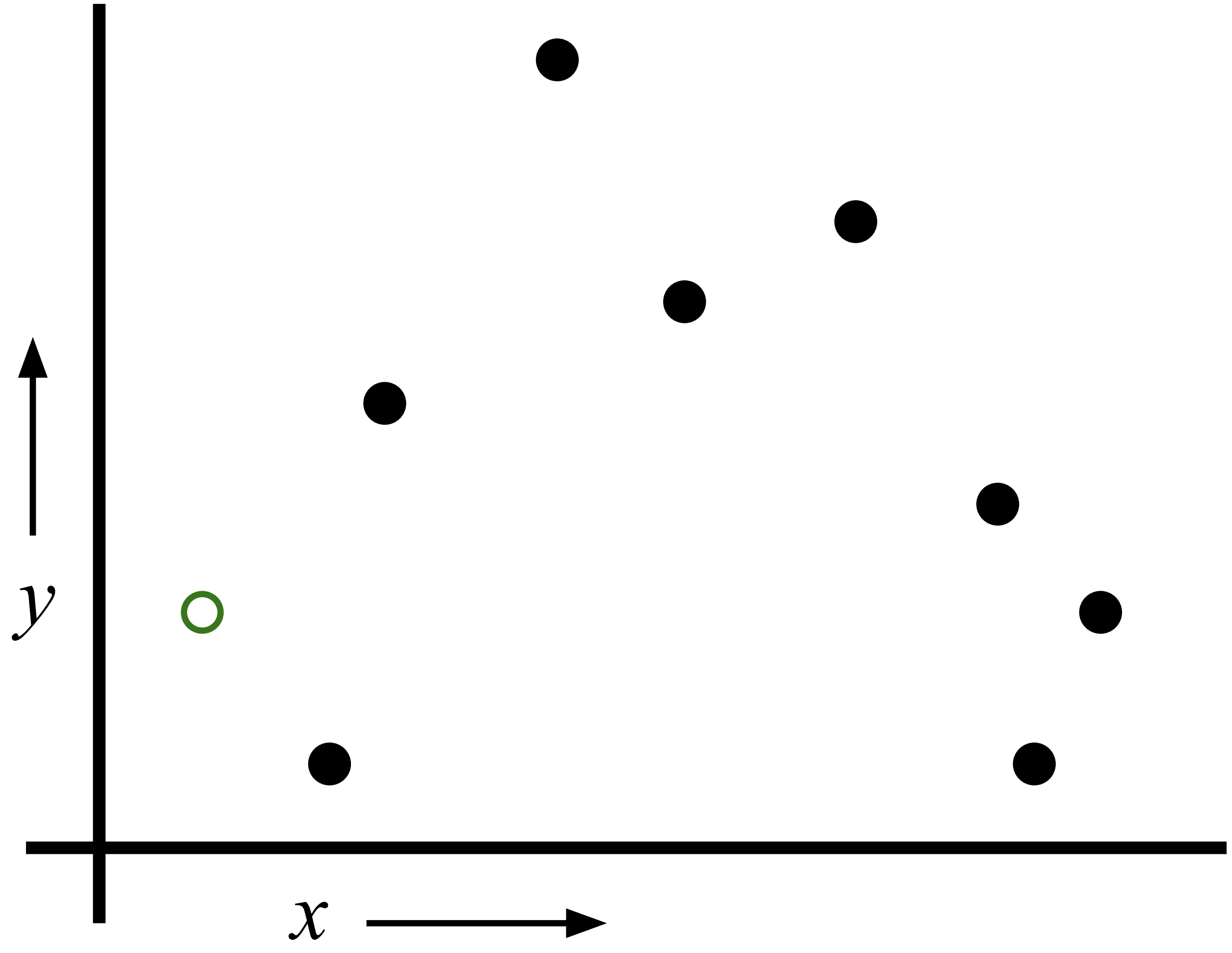
Join the dots Model:



Mean Squared Error = 2.2

slide by Olga Veksler

LOOCV (Leave-one-out Cross Validation)

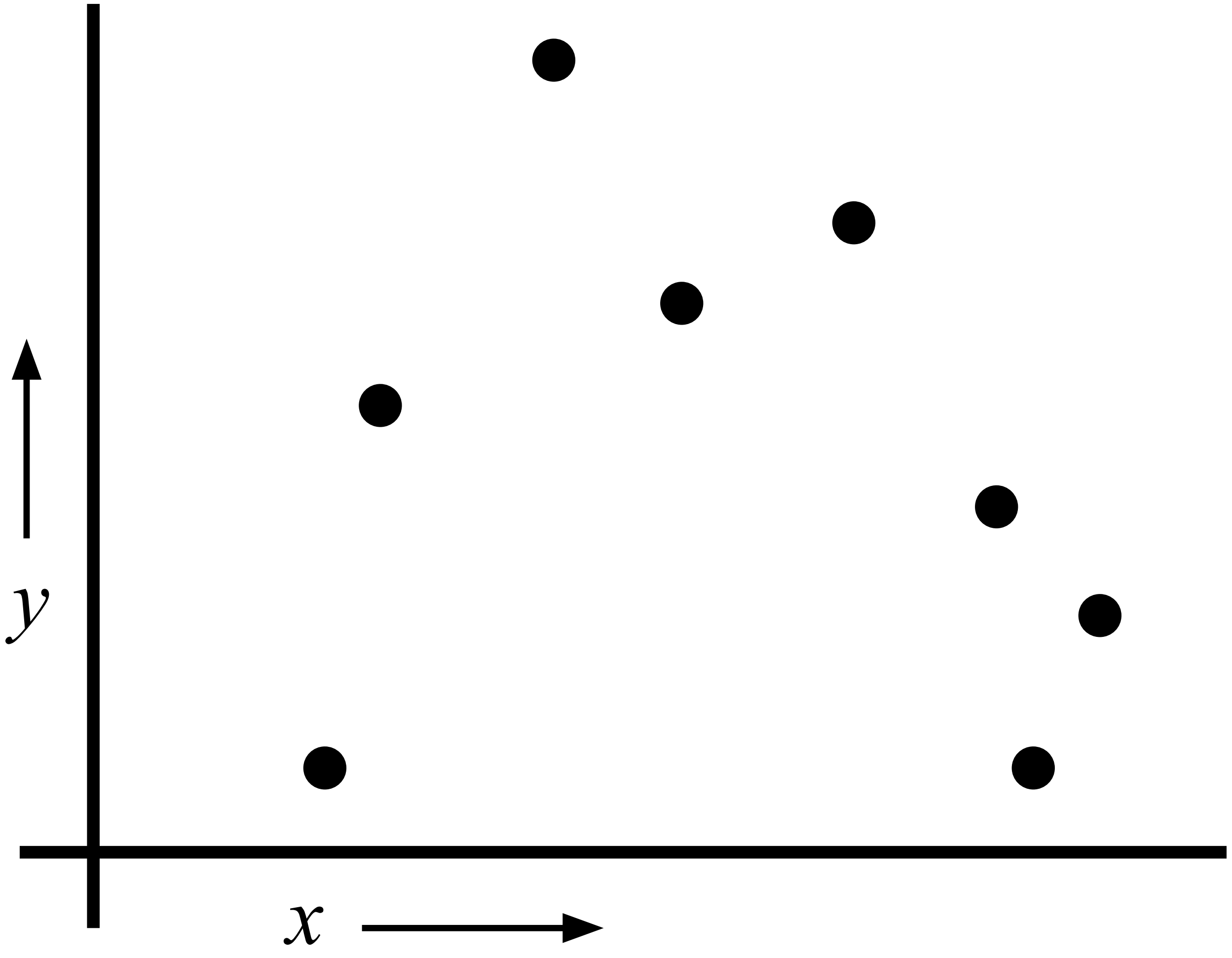


For k=1 to n

1. Let $(\mathbf{x}^k,\mathbf{y}^k)$ be the kth example

slide by Olga Veksler

LOOCV (Leave-one-out Cross Validation)

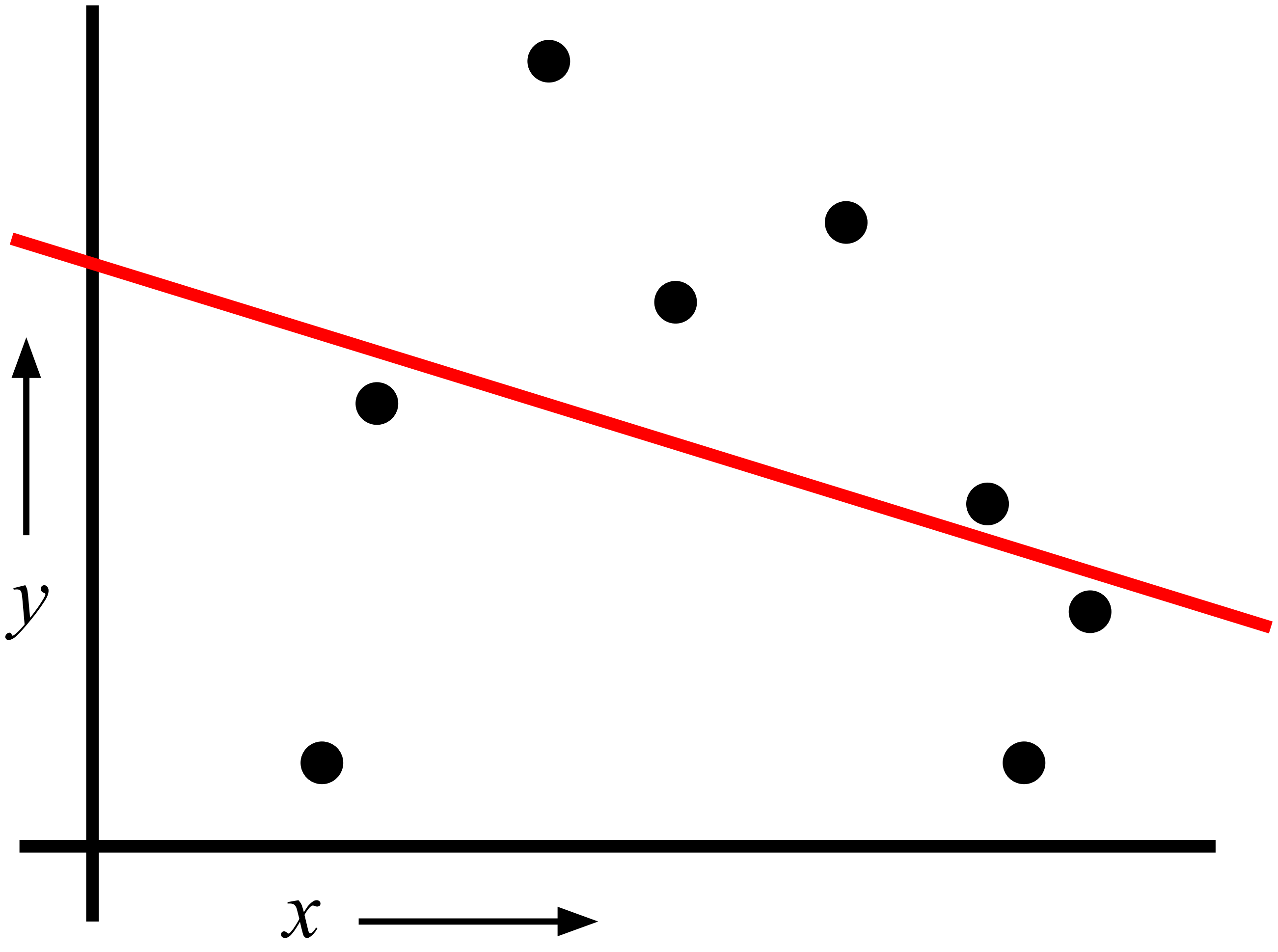


For k=1 to n

1. Let $(\mathbf{x}^k,\mathbf{y}^k)$ be the kth example
2. Temporarily remove $(\mathbf{x}^k,\mathbf{y}^k)$ from the dataset

slide by Olga Veksler

LOOCV (Leave-one-out Cross Validation)

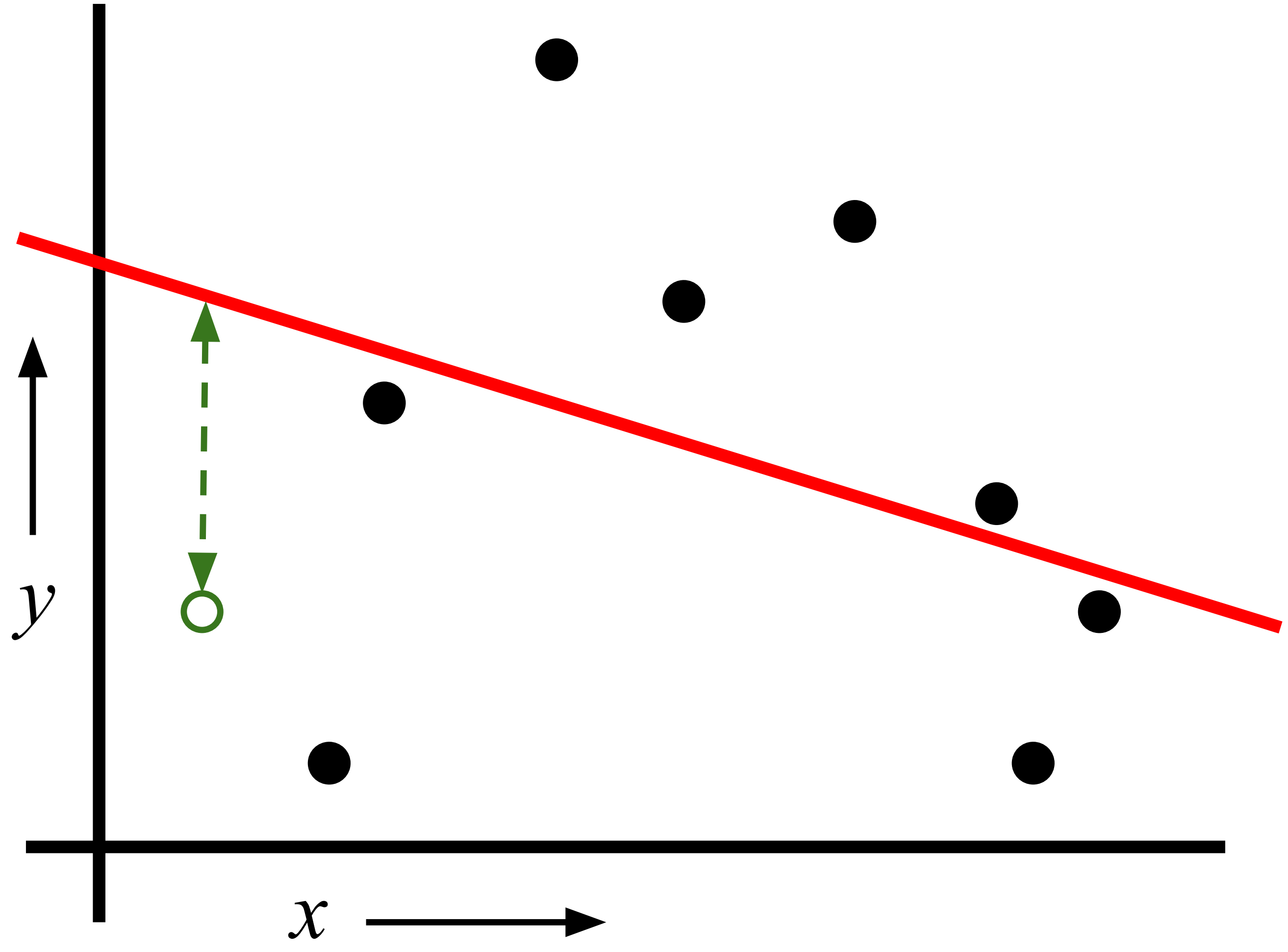


For k=1 to n

1. Let $(\mathbf{x}^k,\mathbf{y}^k)$ be the kth example
2. Temporarily remove $(\mathbf{x}^k,\mathbf{y}^k)$ from the dataset
3. Train on the remaining n-1 examples

slide by Olga Veksler

LOOCV (Leave-one-out Cross Validation)

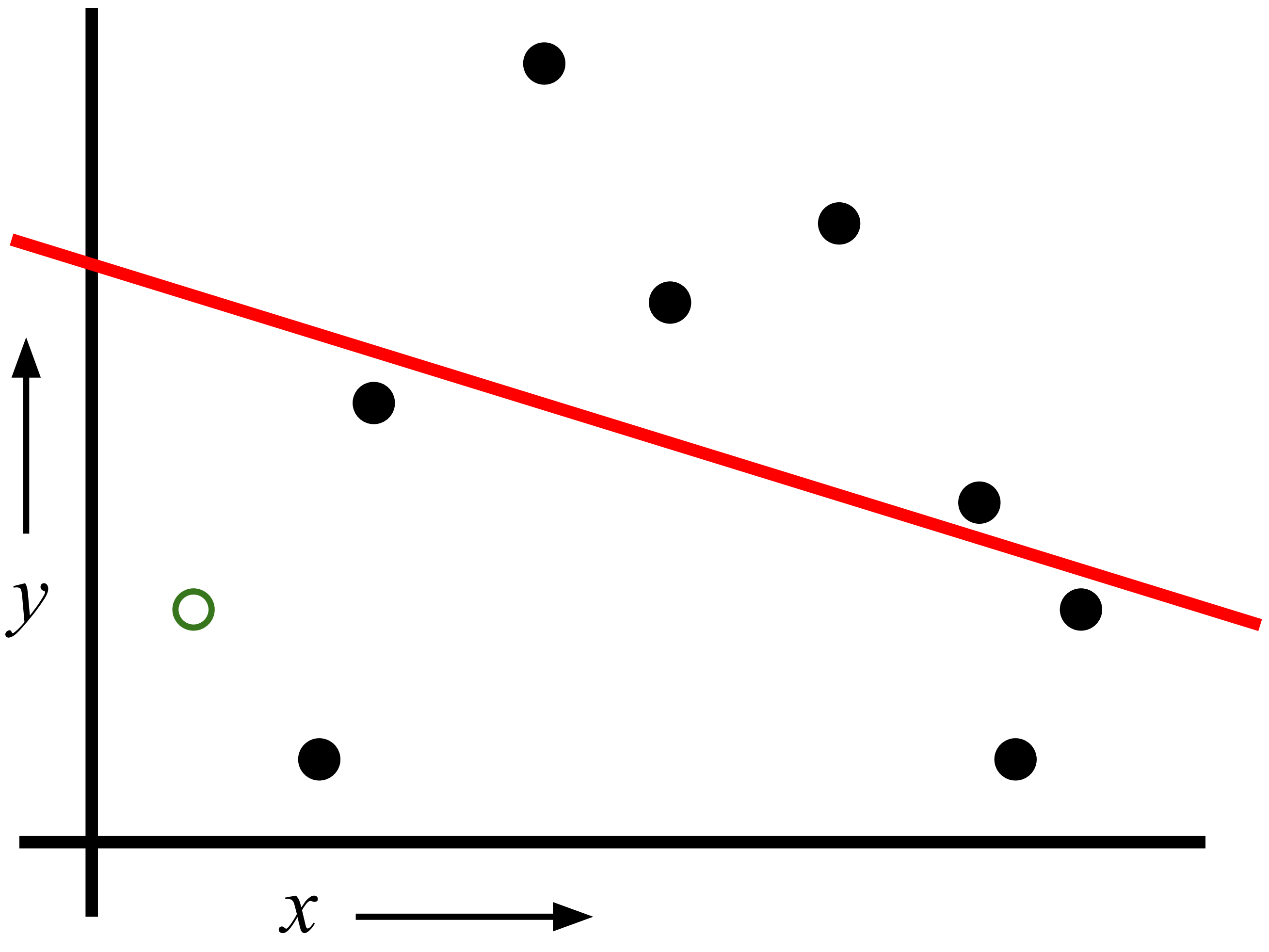


For k=1 to n

1. Let $(\mathbf{x}^k,\mathbf{y}^k)$ be the k^{th} example
2. Temporarily remove $(\mathbf{x}^k,\mathbf{y}^k)$ from the dataset
3. Train on the remaining n-1 examples
4. Note your error on $(\mathbf{x}^k,\mathbf{y}^k)$

slide by Olga Veksler

LOOCV (Leave-one-out Cross Validation)



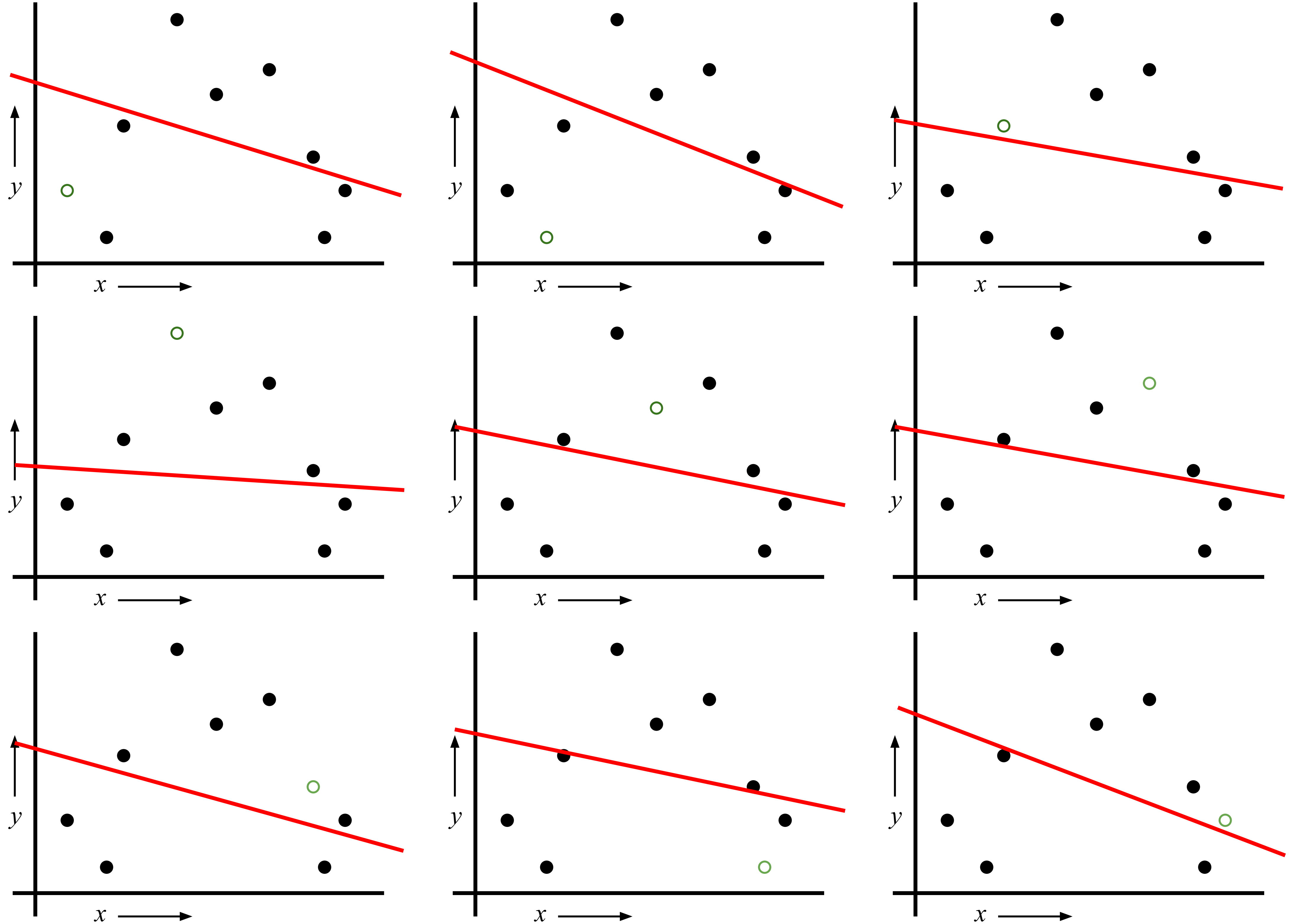
For k=1 to n

1. Let $(\mathbf{x}^k, \mathbf{y}^k)$ be the k^th example
2. Temporarily remove $(\mathbf{x}^k, \mathbf{y}^k)$ from the dataset
3. Train on the remaining n-1 examples
4. Note your error on $(\mathbf{x}^k, \mathbf{y}^k)$

When you've done all points, report the mean error

slide by Olga Veksler

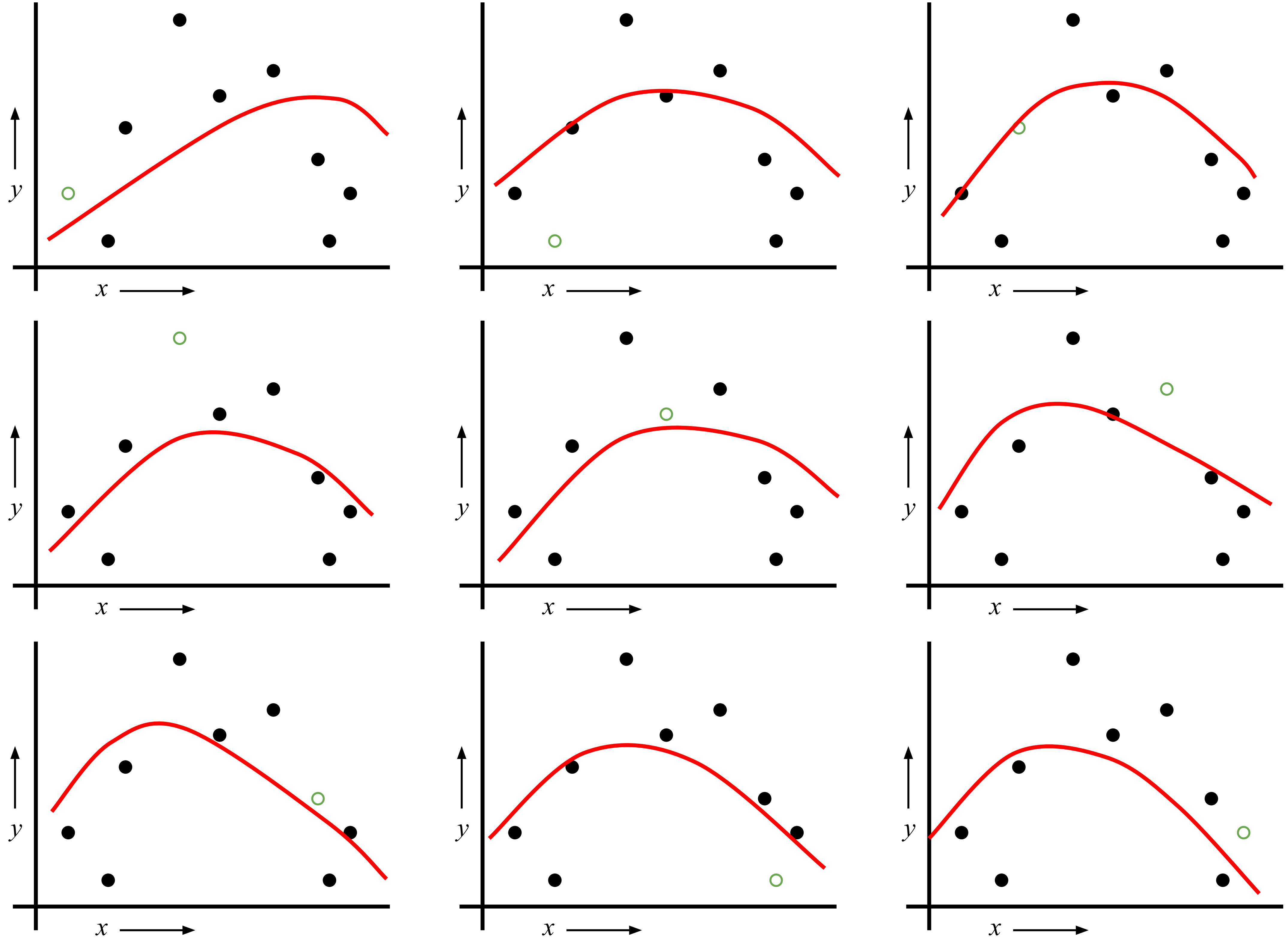
LOOCV (Leave-one-out Cross Validation)



MSE$_{\text{LOOCV}}$ = 2.12

slide by Olga Veksler

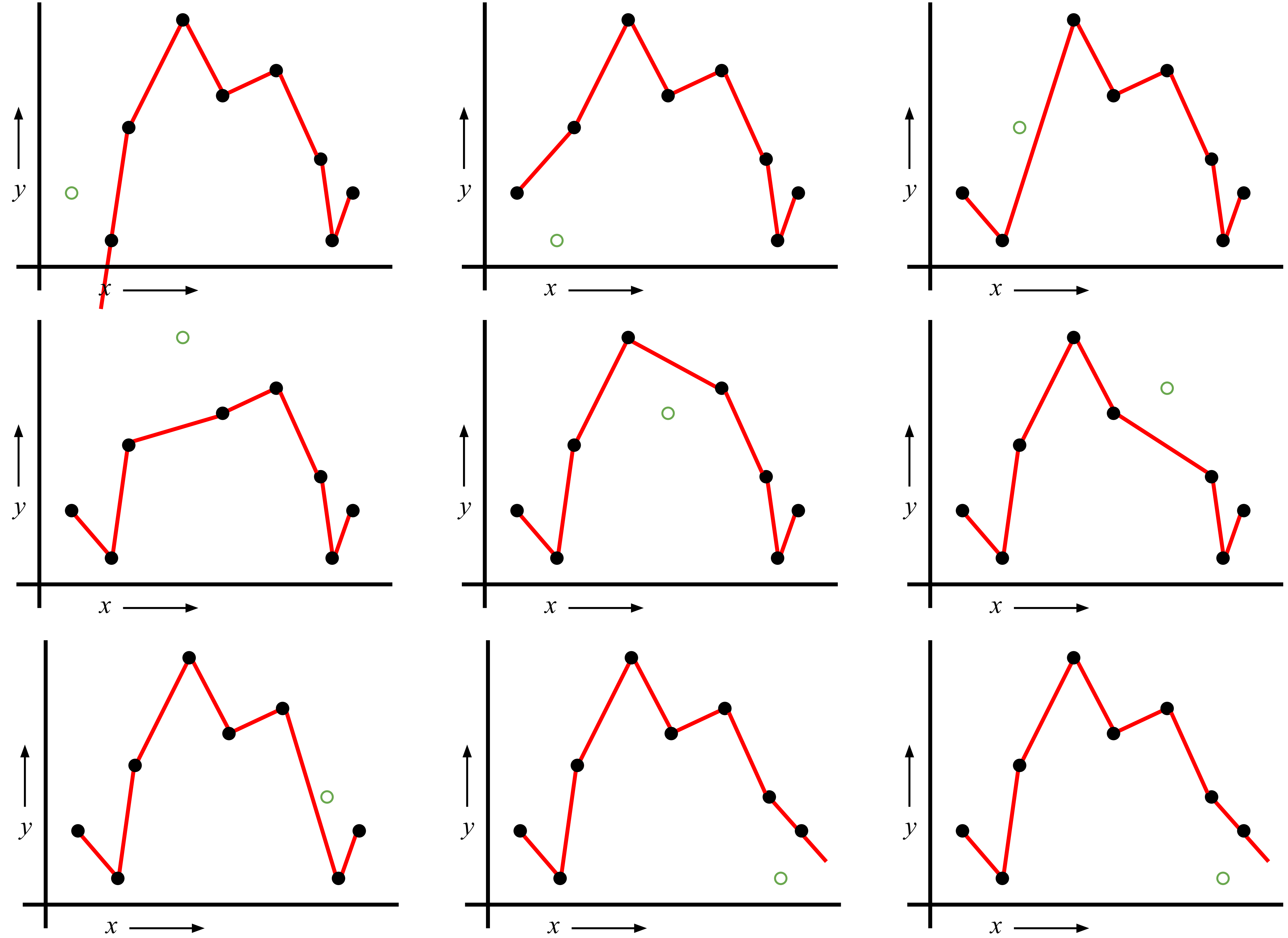
LOOCV for Quadratic Regression



$\text{MSE}_{\text{LOOCV}}$
$= 0.96$

slide by Olga Veksler

LOOCV for Joint The Dots



$\text{MSE}_{\text{LOOCV}}$ = 3.33

slide by Olga Veksler

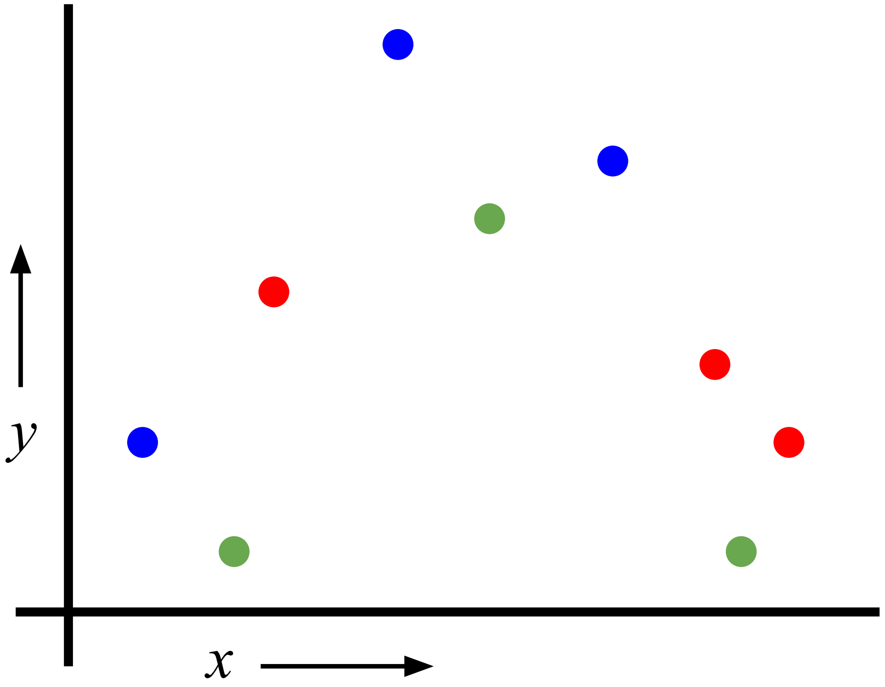
Which kind of Cross Validation?

	Downside	**Upside**
Test-set	may give unreliable estimate of future performance	cheap
Leave-one-out	expensive	doesn’t waste data

- Can we get the best of both worlds?

slide by Olga Veksler

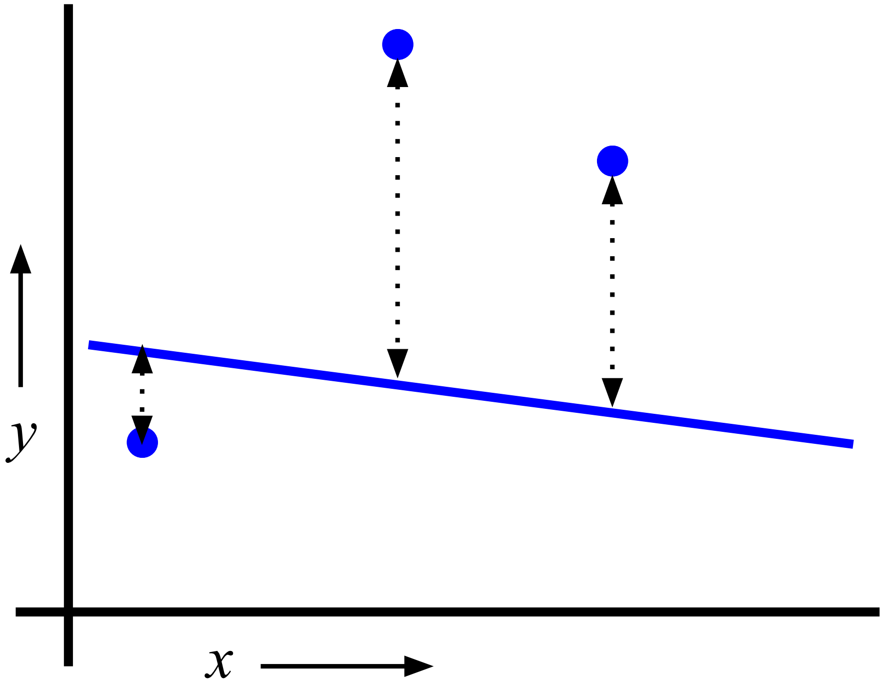
K-Fold Cross Validation



- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue

slide by Olga Veksler

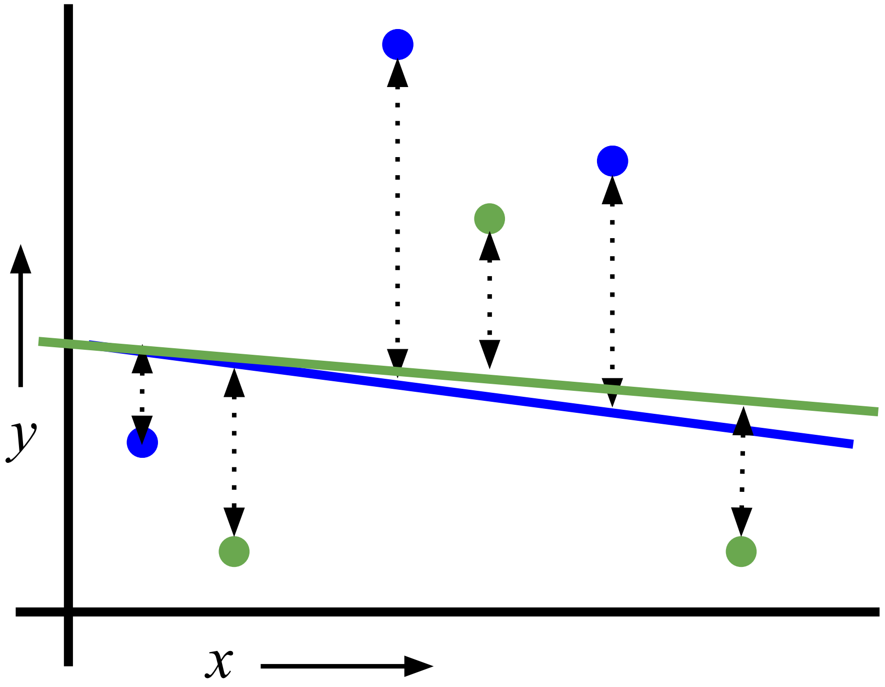
K-Fold Cross Validation



- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue
- For the blue partition: train on all points not in the blue partition. Find test-set sum of errors on blue points

slide by Olga Veksler

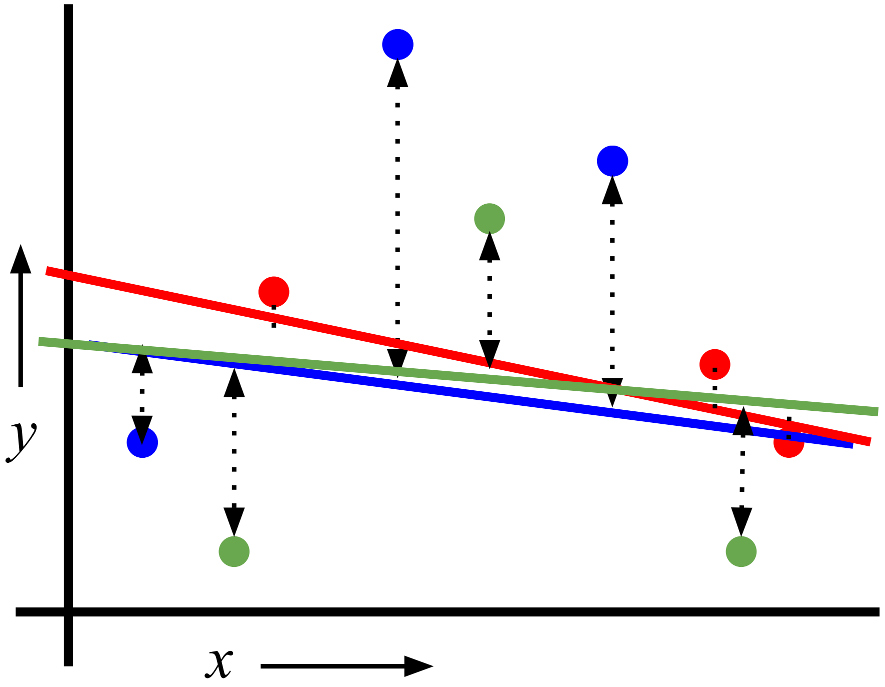
K-Fold Cross Validation



- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue
- For the blue partition: train on all points not in the blue partition. Find test-set sum of errors on blue points
- For the green partition: train on all points not in green partition. Find test-set sum of errors on green points

slide by Olga Veksler

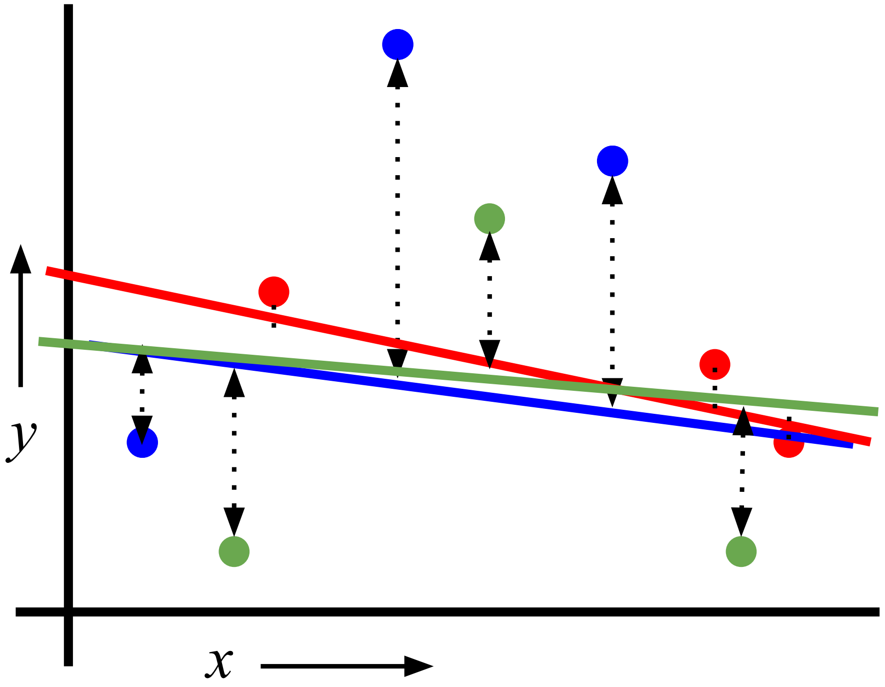
K-Fold Cross Validation



- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue
- For the blue partition: train on all points not in the blue partition. Find test-set sum of errors on blue points
- For the green partition: train on all points not in green partition. Find test-set sum of errors on green points
- For the red partition: train on all points not in red partition. Find the test-set sum of errors on red points

slide by Olga Veksler

K-Fold Cross Validation

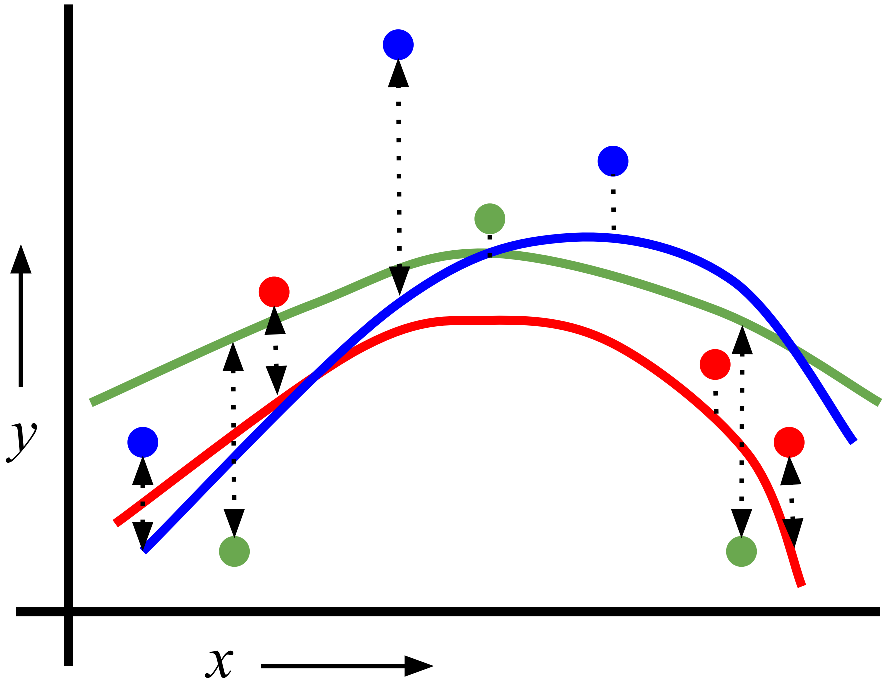


Linear Regression
MSE_{3FOLD} = 2.05

- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue
- For the blue partition: train on all points not in the blue partition. Find test-set sum of errors on blue points
- For the green partition: train on all points not in green partition. Find test-set sum of errors on green points
- For the red partition: train on all points not in red partition. Find the test-set sum of errors on red points
- Report the mean error

slide by Olga Veksler

K-Fold Cross Validation

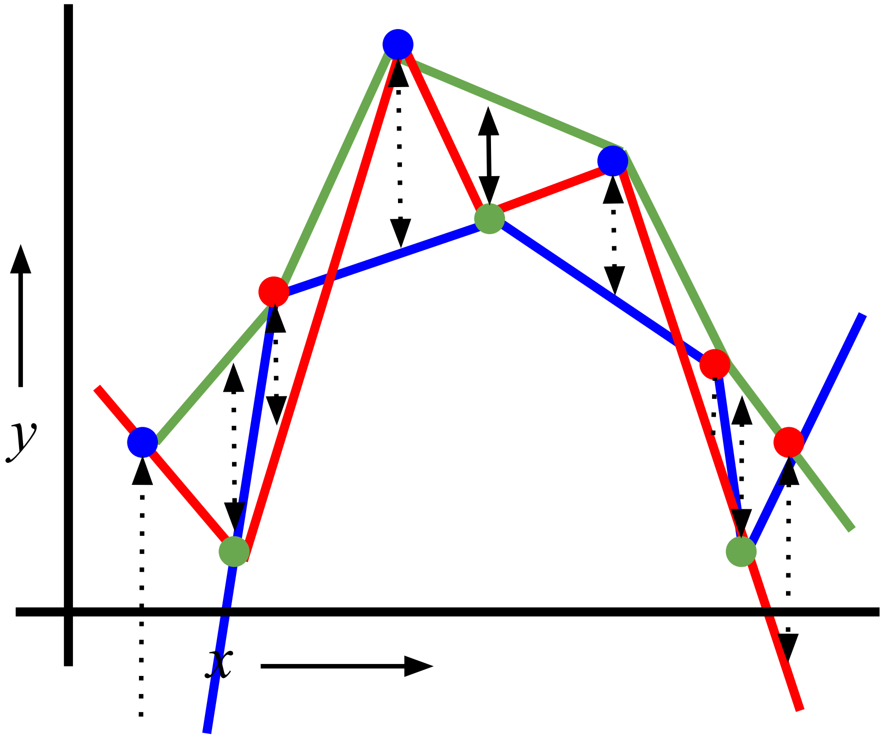


Quadratic Regression
MSE$_{\text{3FOLD}}$ = 1.1

- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue
- For the blue partition: train on all points not in the blue partition. Find test-set sum of errors on blue points
- For the green partition: train on all points not in green partition. Find test-set sum of errors on green points
- For the red partition: train on all points not in red partition. Find the test-set sum of errors on red points
- Report the mean error

slide by Olga Veksler

K-Fold Cross Validation



Join the dots
MSE$_{3FOLD}$ = 2.93

- Randomly break the dataset into k partitions
- In this example, we have k=3 partitions colored red green and blue
- For the blue partition: train on all points not in the blue partition. Find test-set sum of errors on blue points
- For the green partition: train on all points not in green partition. Find test-set sum of errors on green points
- For the red partition: train on all points not in red partition. Find the test-set sum of errors on red points
- Report the mean error

slide by Olga Veksler

Which kind of Cross Validation?

	Downside	Upside
Test-set	may give unreliable estimate of future performance	cheap
Leave-one-out	expensive	doesn't waste data
10-fold	wastes 10% of the data,10 times more expensive than test set	only wastes 10%, only 10 times more expensive instead of **n** times
3-fold	wastes more data than 10-fold, more expensive than test set	slightly better than test-set
N-fold	Identical to Leave-one-out	

slide by Olga Veksler

Cross-validation for classification

- Instead of computing the sum squared errors on a test set, you should compute...

slide by Andrew Moore

Cross-validation for classification

- Instead of computing the sum squared errors on a test set, you should compute…

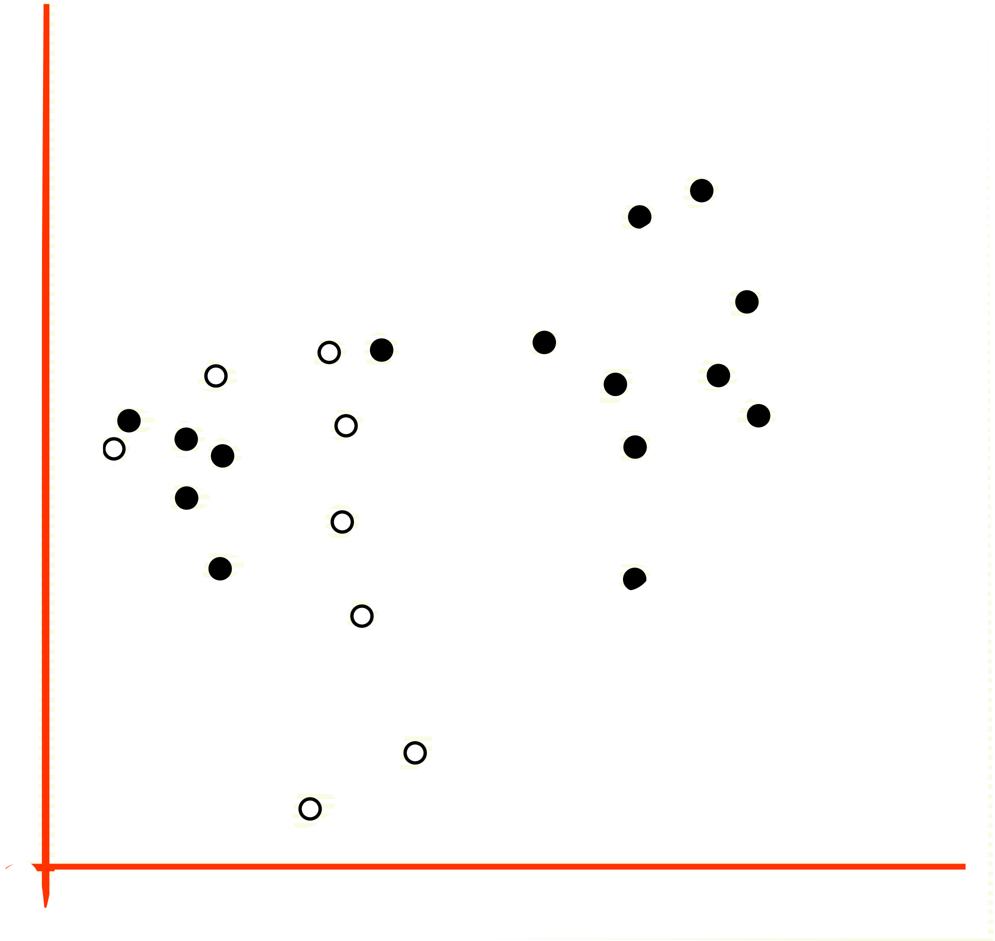
The total number of misclassifications on a test set

slide by Andrew Moore

Cross-validation for classification

- Instead of computing the sum squared errors on a test set, you should compute…

The total number of misclassifications on a test set



- What's LOOCV of 1-NN?
- What's LOOCV of 3-NN?
- What's LOOCV of 22-NN?

slide by Andrew Moore

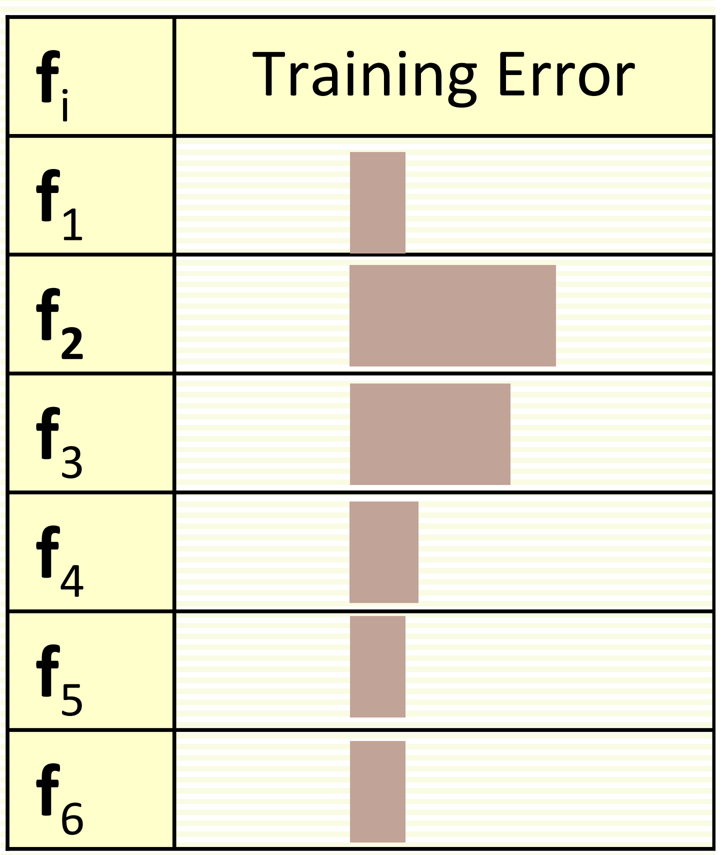
Cross-validation for classification

- Choosing k for k-nearest neighbors
- Choosing Kernel parameters for SVM
- Any other “free” parameter of a classifier
- Choosing Features to use
- Choosing which classifier to use

slide by Andrew Moore

CV-based Model Selection

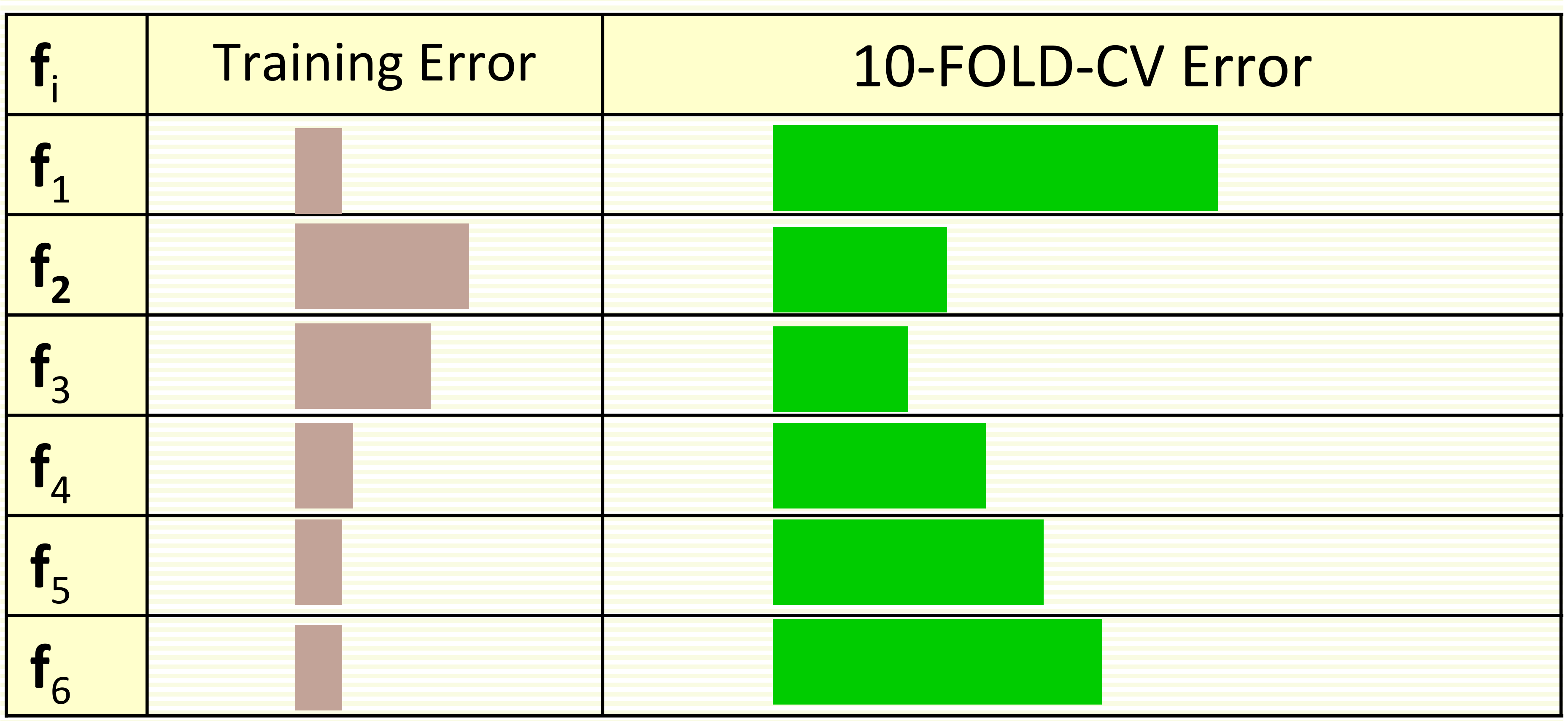
- We’re trying to decide which algorithm to use.
- We train each machine and make a table...



slide by Olga Veksler

CV-based Model Selection

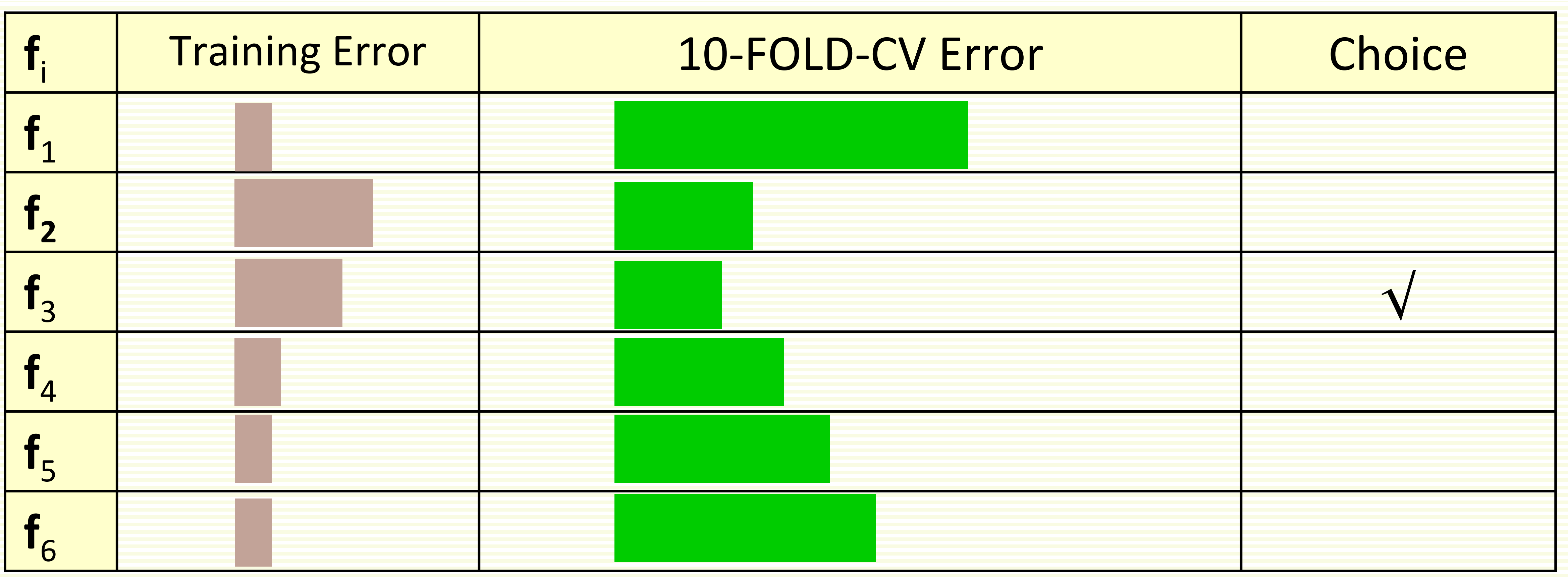
- We’re trying to decide which algorithm to use.
- We train each machine and make a table...



slide by Olga Veksler

CV-based Model Selection

- We're trying to decide which algorithm to use.
- We train each machine and make a table...

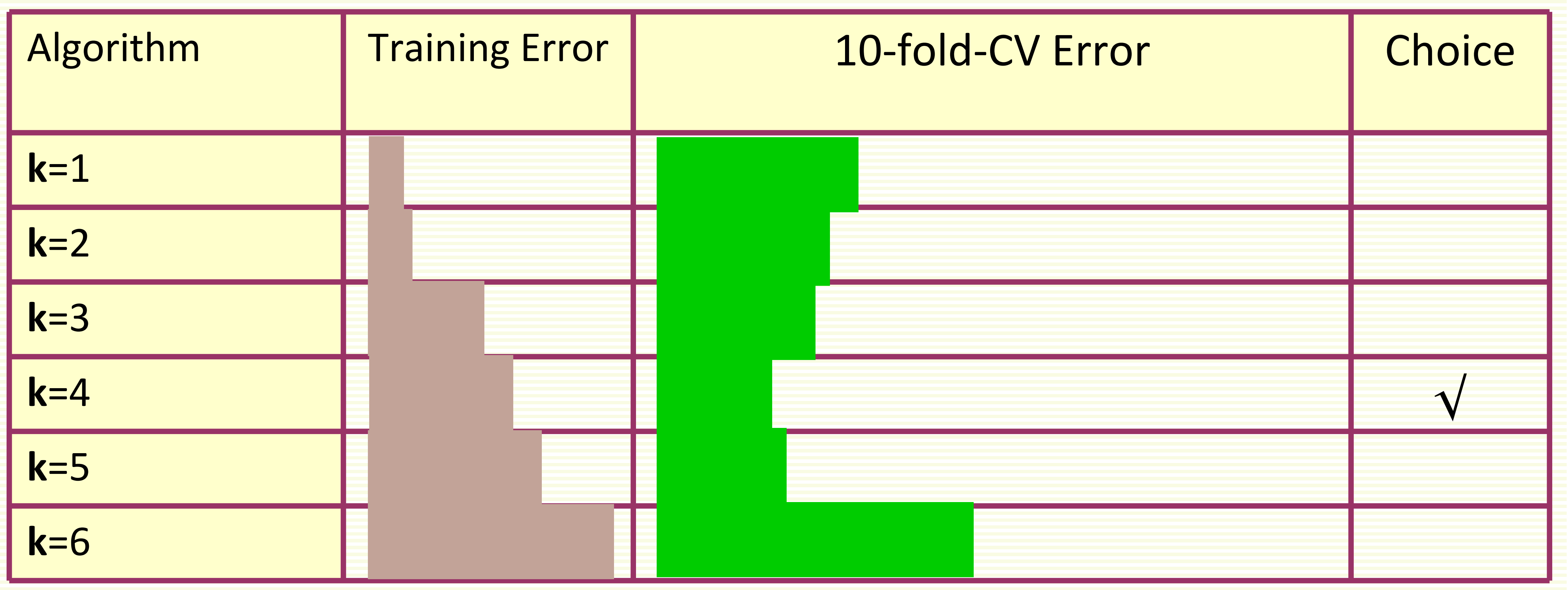


f_i	Training Error	10-FOLD-CV Error	Choice
f_1			
f_2			
f_3			√
f_4			
f_5			
f_6			

slide by Olga Veksler

CV-based Model Selection

- Example: Choosing “k” for a k-nearest-neighbor regression.
- Step 1: Compute LOOCV error for six different model classes:



Algorithm	Training Error	10-fold-CV Error	Choice
k=1			
k=2			
k=3			
k=4			√
k=5			
k=6			

- Step 2: Choose model that gave the best CV score
- Train with all the data, and that’s the final model you’ll use

slide by Olga Veksler

CV-based Model Selection

- Why stop at k=6?
    - No good reason, except it looked like things were getting worse as K was increasing

- Are we guaranteed that a local optimum of K vs LOOCV will be the global optimum?
    - No, in fact the relationship can be very bumpy

- What should we do if we are depressed at the expense of doing LOOCV for k = 1 through 1000?
    - Try: k=1, 2, 4, 8, 16, 32, 64, ... ,1024
    - Then do hillclimbing from an initial guess at k

slide by Olga Veksler

Next Lecture:

Learning Theory &
Probability Review