



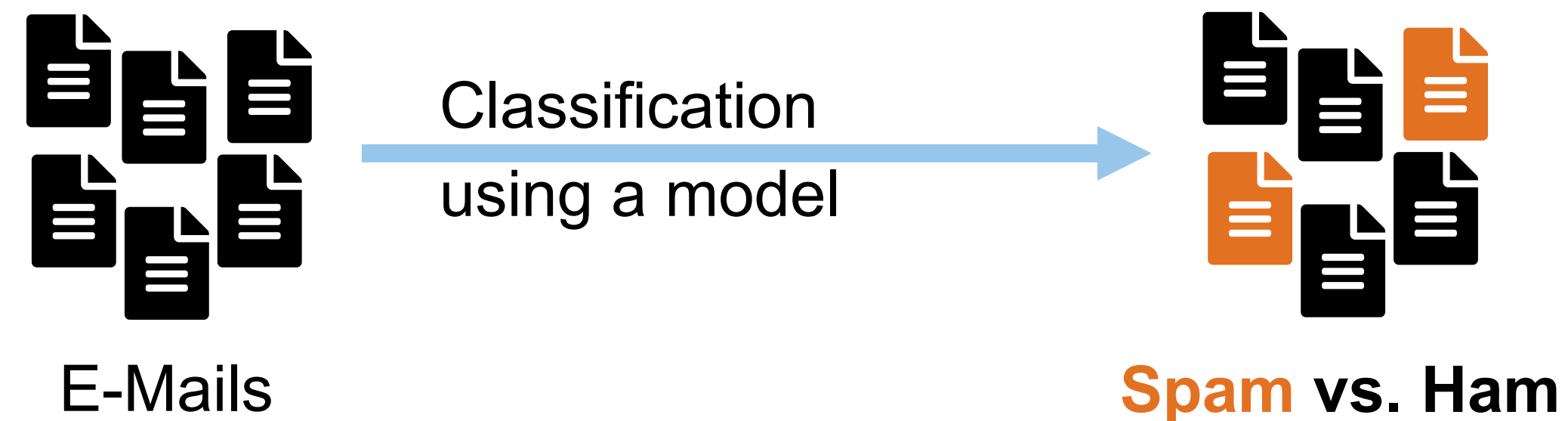
# Supervised Learning: Choosing the right model

Mirco Schönfeld  
[mirco.schoenfeld@uni-bayreuth.de](mailto:mirco.schoenfeld@uni-bayreuth.de)



# How to Design a Model: Feature Selection

Formulate characteristics that help distinguishing between classes.  
 For spam-detection: find words or combinations of words that indicate a mail being spam.



Spam: Wholesale Fashion Watches -57% today. Designer watches **for cheap** ...

Spam: **You can buy** **Viagra** Fr\$1.85 All Medications at unbeatable prices! ...

Spam: WE CAN TREAT ANYTHING YOU SUFFER FROM JUST TRUST US ...

Spam: Sta.rt earn\*ing the salary **yo,u d-eserve** by o'btaining the prope,r crede'ntials!

Ham: The practical significance of hypertree width in identifying more ...

Ham: Abstract: We will motivate the problem of social identity clustering: ...

Ham: Good to see you my friend. Hey Peter, It was good to hear from you. ...

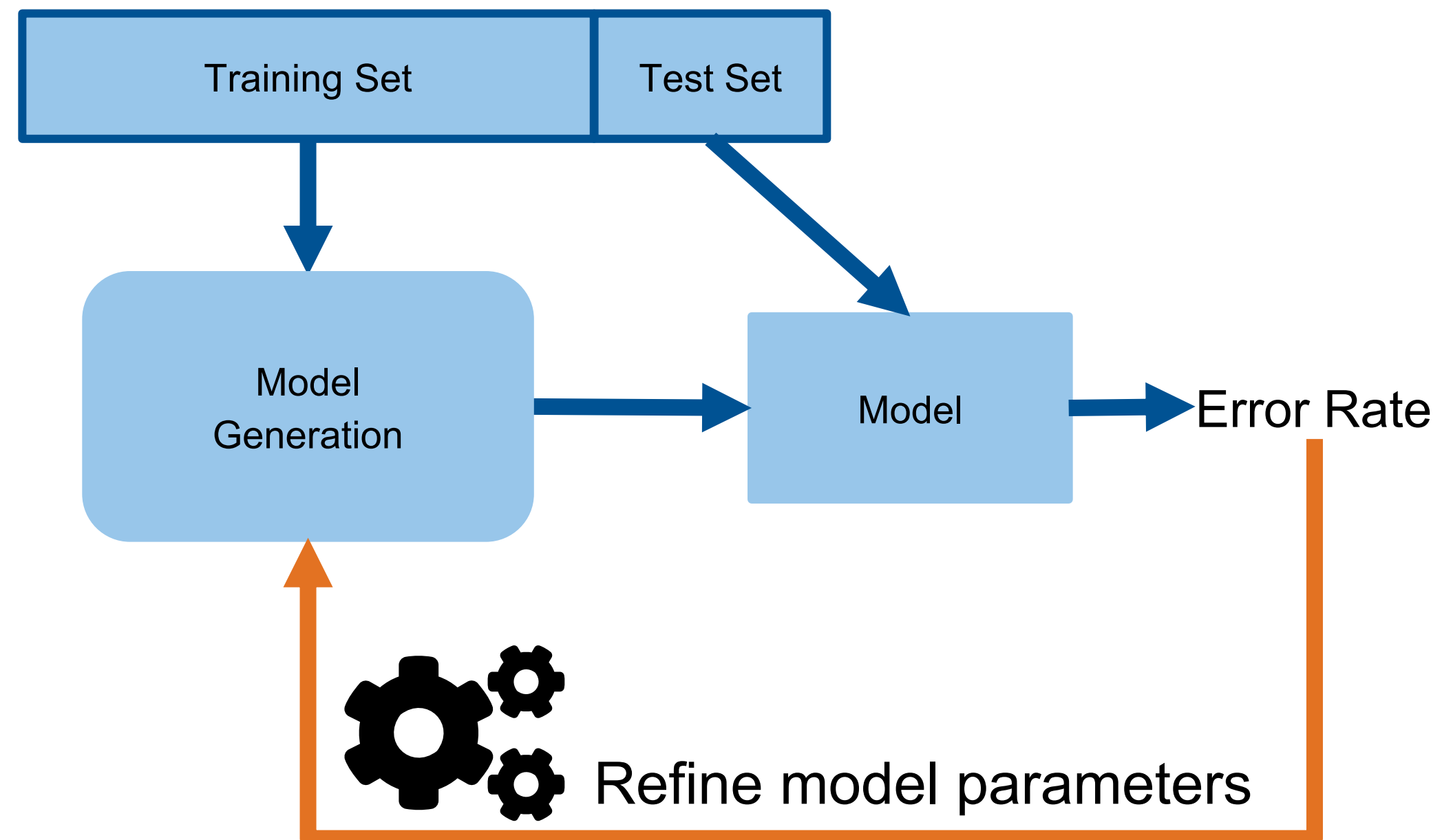
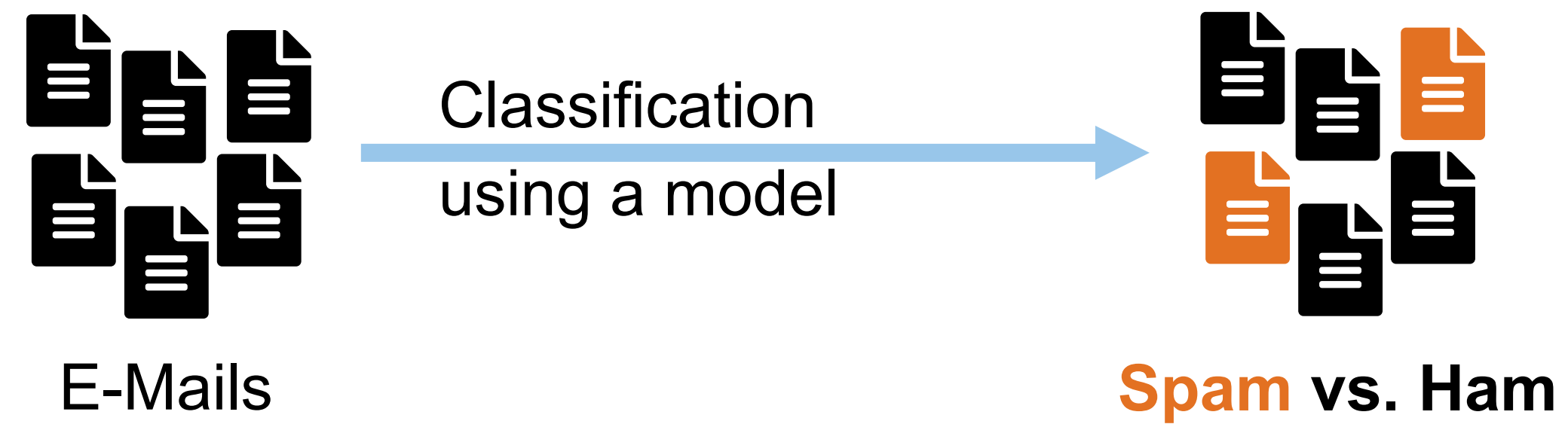
Ham: PDS implies convexity of the resulting optimization problem (Kernel Ridge ...

## Curse of Dimensionality:

Including more features will improve classification *conceptually* but will render computation increasingly difficult.

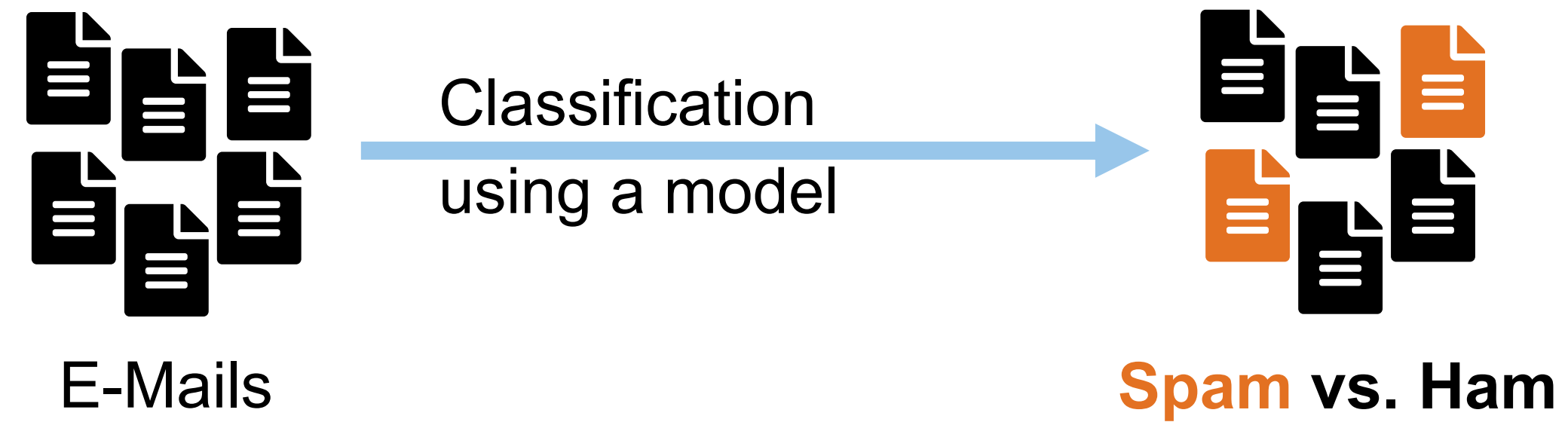


# Training the Model



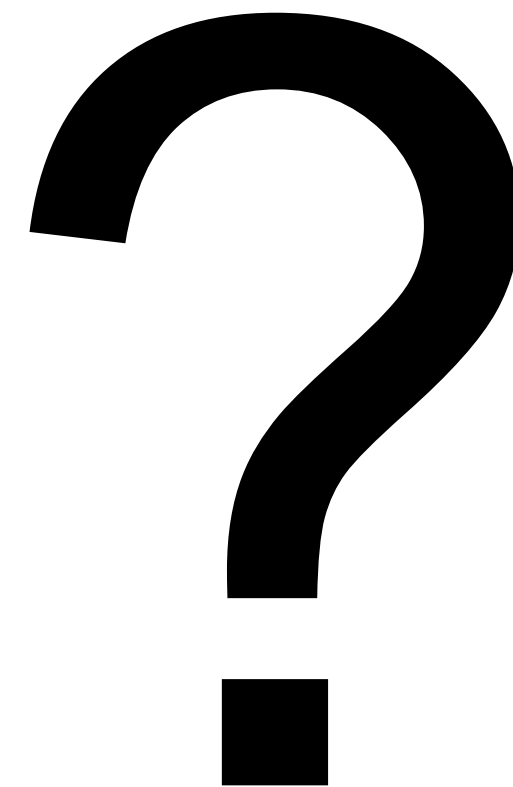


# How to Choose a Model



Model A:

		Correct class	
		Ham	Spam
Spam-Filter reports	Ham	189	1
	Spam	11	799

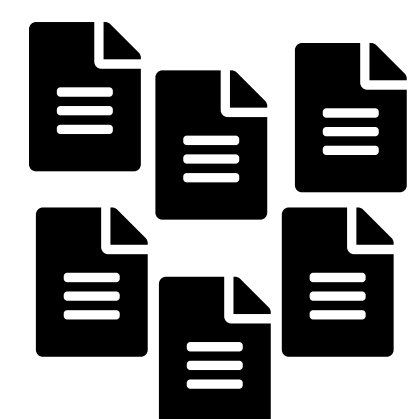


Model B:

		Correct class	
		Ham	Spam
Spam-Filter reports	Ham	200	38
	Spam	0	762



# Good Classifications: The Confusion Matrix



E-Mails

Classification  
using a model



Spam vs. Ham

Model A:

		Correct class	
		Ham	Spam
Spam-Filter reports	Ham	189	1
	Spam	11	799

Model B:

		Correct class	
		Ham	Spam
Spam-Filter reports	Ham	200	38
	Spam	0	762

Correct

Positive (P)

Negative (N)

Predicted

Positive

True  
Positive

False  
Positive

Negative

False  
Negative

True  
Negative

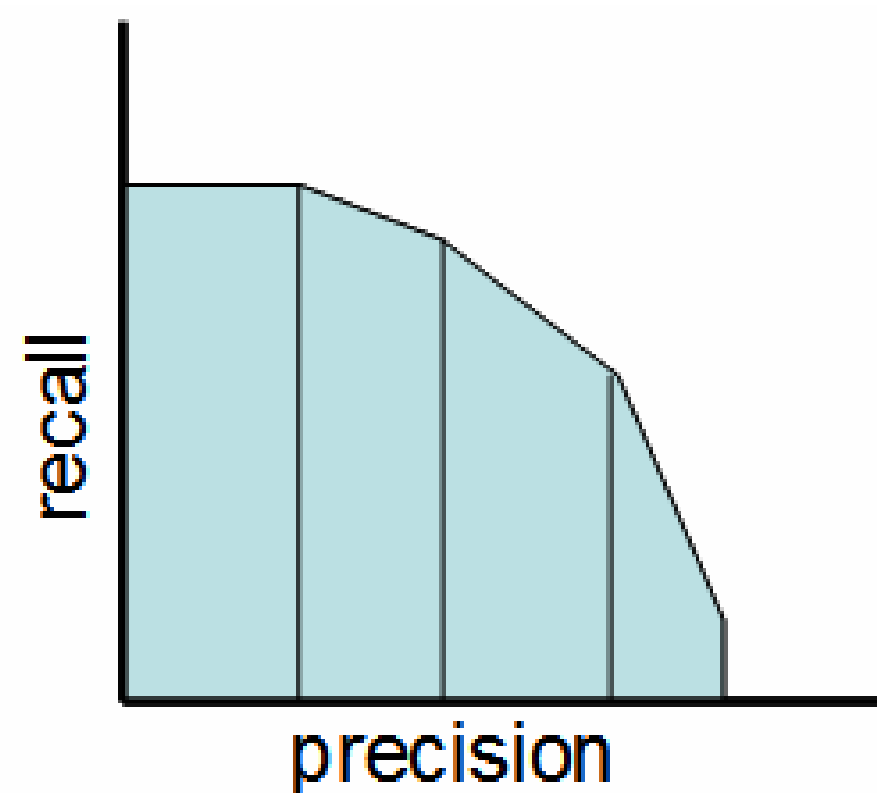


# Measuring Goodness

- **Precision:** Proportion of predicted positives that are truly positive  
good choice when we need to be very sure of prediction

$$\frac{TP}{TP + FP}$$

- **Recall:** Proportion of actual positives that are correctly classified  
good choice when as many positives as possible should be captured



$$\frac{TP}{TP + FN}$$

		Correct	
		Positive (P)	Negative (N)
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



# Measuring Goodness

- **Accuracy:** Proportion of true results among total number of cases  
good choice when classes are balanced

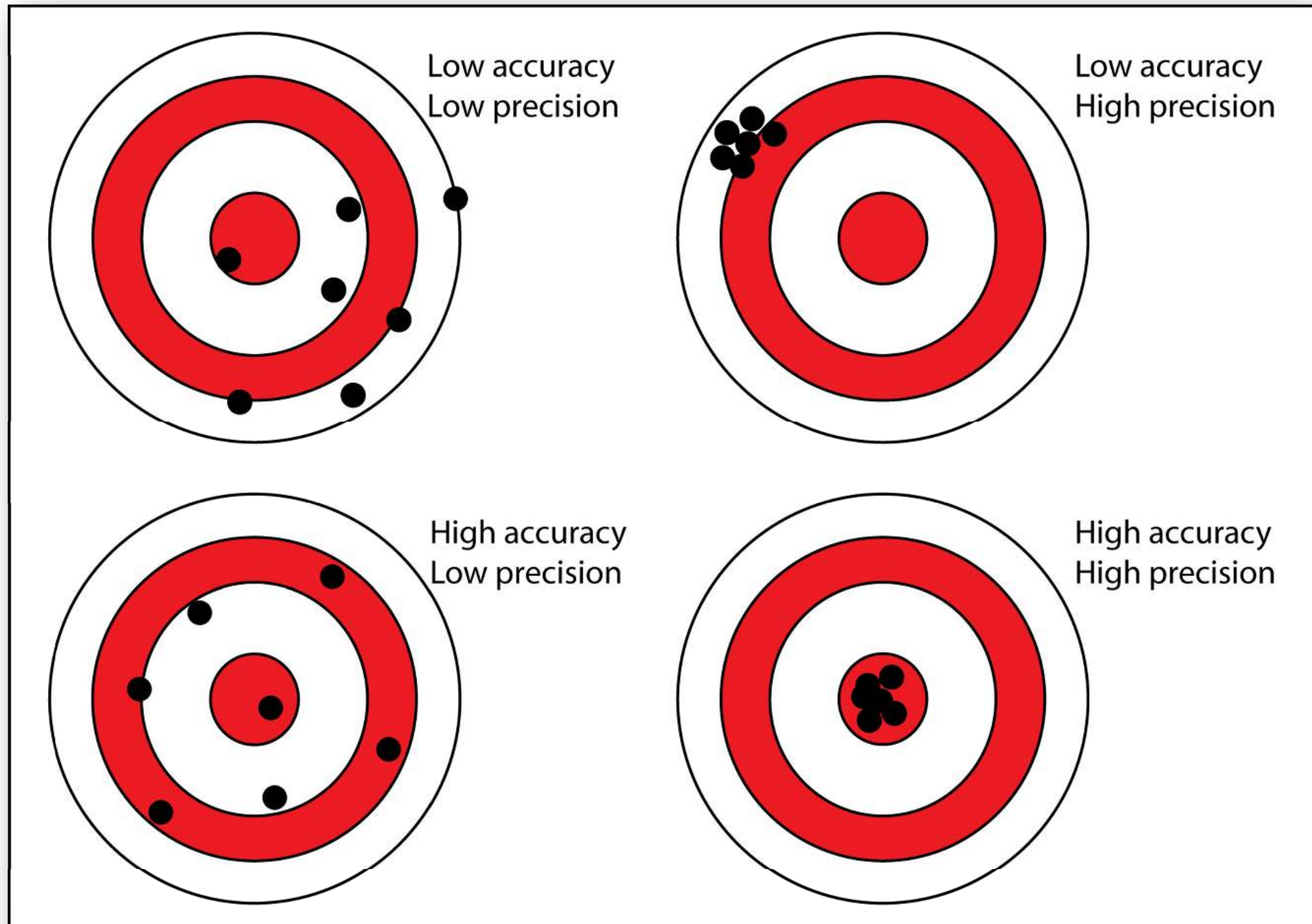
$$\frac{TP + TN}{TP + FP + FN + FN}$$

- **$F_1$  Score:** harmonic mean between precision & recall – a number between 0 and 1  
good choice when we want a model with both good precision and recall

$$2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

Important variant  $F_\beta$  allows to apply a custom weight to precision & recall

		Correct	
		Positive (P)	Negative (N)
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative





# Measuring Goodness & more

## Prevalence

$$\frac{P}{P + N}$$

## accuracy (ACC)

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

## balanced accuracy (BA)

$$BA = \frac{TPR + TNR}{2}$$

## F1 score

is the harmonic mean of precision and sensitivity:

$$F_1 = 2 \times \frac{PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

## phi coefficient ( $\phi$ or $r_\phi$ ) or Matthews correlation coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

## Fowlkes-Mallows index (FM)

$$FM = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} = \sqrt{PPV \times TPR}$$

## informedness or bookmaker informedness (BM)

$$BM = TPR + TNR - 1$$

## markedness (MK) or deltaP ( $\Delta p$ )

$$MK = PPV + NPV - 1$$

## Diagnostic odds ratio (DOR)

$$DOR = \frac{LR+}{LR-}$$

## sensitivity, recall, hit rate, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

## specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

## precision or positive predictive value (PPV)

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

## negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN} = 1 - FOR$$

## miss rate or false negative rate (FNR)

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$$

## fall-out or false positive rate (FPR)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

## false discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

## false omission rate (FOR)

$$FOR = \frac{FN}{FN + TN} = 1 - NPV$$

## Positive likelihood ratio (LR+)

$$LR+ = \frac{TPR}{FPR}$$

## Negative likelihood ratio (LR-)

$$LR- = \frac{FNR}{TNR}$$

## prevalence threshold (PT)

$$PT = \frac{\sqrt{TPR(-TNR + 1)} + TNR - 1}{(TPR + TNR - 1)} = \frac{\sqrt{FPR}}{\sqrt{TPR} + \sqrt{FPR}}$$

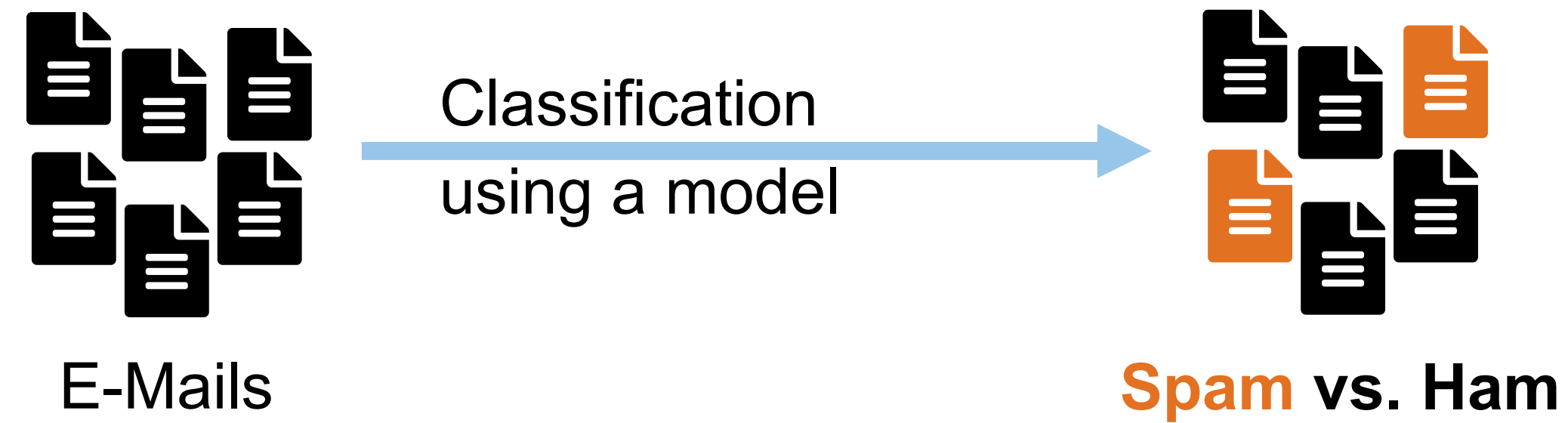
## threat score (TS) or critical success index (CSI)


$$TS = \frac{TP}{TP + FN + FP}$$

		Correct	
		Positive (P)	Negative (N)
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative




# How to Choose a Model



 Minimize error rate


Model A:

		Correct class	
		Ham	Spam
Spam-Filter reports	Ham	189	1
	Spam	11	799

 Maximize *utility*

Model B:

		Correct class	
		Ham	Spam
Spam-Filter reports	Ham	200	38
	Spam	0	762





# Interpretability

Large project set out to evaluate ML application to problems in healthcare

Scenario: predicting pneumonia risk

Goal: predict probability of death for patients with pneumonia

The most accurate model of the study was a multitask neural net.

Outperformed other models by wide margin but was still dropped. Why?

One rule-based system learned the rule „patient has asthma → lower risk“

Reflected a true pattern in training data

The best model was the least intelligible one – was deemed to risky

No way of checking the features that were picked up



# Goals of Interpretable Models

- **Trust: Identify and mitigate *bias***  
Recognizing bias in a black-box algorithm is *very* hard
- **Causality: Account for context**  
Helps you understand how the factors included in the model led to the prediction
- **Informativeness: Extract knowledge**  
Helps you determine if patterns that appear to be present in the model are really there.  
Rather learning from the model than evaluating it (compared to identifying bias)
- **Transferability: Generalize**  
Models are trained on carefully collected datasets to solve narrowly defined problems. Interpretable models should help you determine if and how they can be generalized
- **Fair and Ethical Decision-Making**  
algorithmic decision-making mediates more and more of our interactions. Need a way to make sure that decisions conform to ethical standards



# Properties of Interpretable Models

## 1. Transparency

- **Simulatability**

Transparency at the level of the entire model

- **Decomposability / Intelligibility**

Transparency at the level of the individual components, e.g. parameters

- **Algorithmic Transparency**

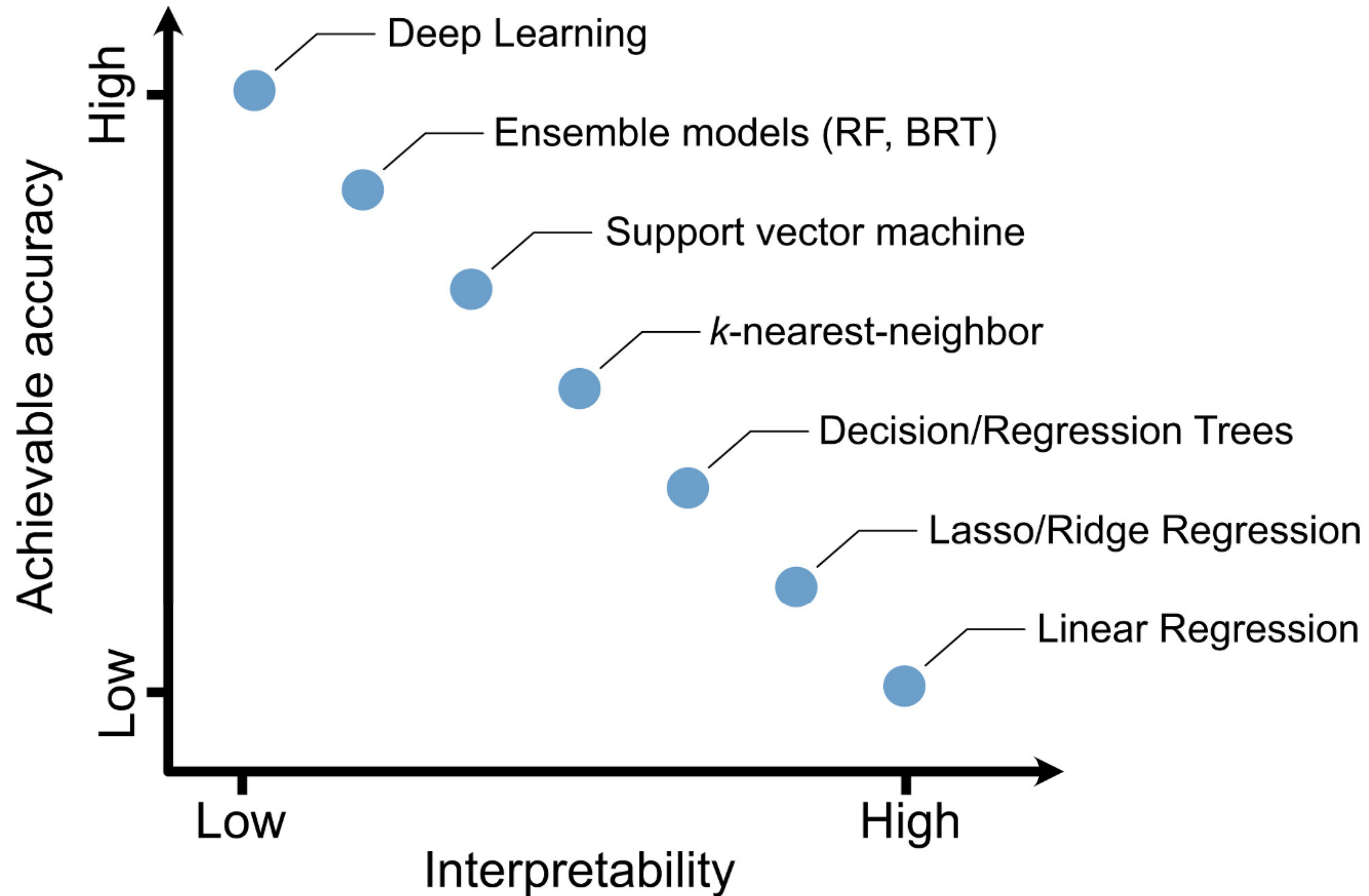
Transparency at the level of the training algorithm



# Properties of Interpretable Models

## 2. Post-hoc Interpretability

- Text Explanations
- Visualization
- Local Explanations
- Explanation by Example





# How to select a model?

- **Quality of predictions**  
i.e. performance in terms of a quality metric
- **Speed**  
i.e. training time, prediction time
- **Robustness**  
i.e. handling noise or missing values and still classify correctly
- **Scalability**  
i.e. computational efficiency
- **Interpretability**  
subjective means
- **Other**



Thanks.

[mirco.schoenfeld@uni-bayreuth.de](mailto:mirco.schoenfeld@uni-bayreuth.de)

<https://xkcd.com/1838/>