Introduction to Machine Learning
HPC AI — 2019-20

3. Model evaluation & selection

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr

Inserm
Institut national
de la santé et de la recherche médicale

institutCurie

PSL
RESEARCH UNIVERSITY PARIS

MINES
ParisTech

No free lunch theorem

- Any two machine learning algorithms are **equivalent** if their performance are **averaged over all possible learning tasks.**
- For any algorithm, **there is a task for which it performs poorly.**

- Rules of thumbs will give you an idea of what to expect... but there is **no guarantee** that the first algorithm you choose gives you the best model.

Learning objectives

After this lecture you should be able to

- recognize **overfitting;**
- explain what **generalization** is;
- **design experiments to select and evaluate supervised machine learning models.**

Generalization

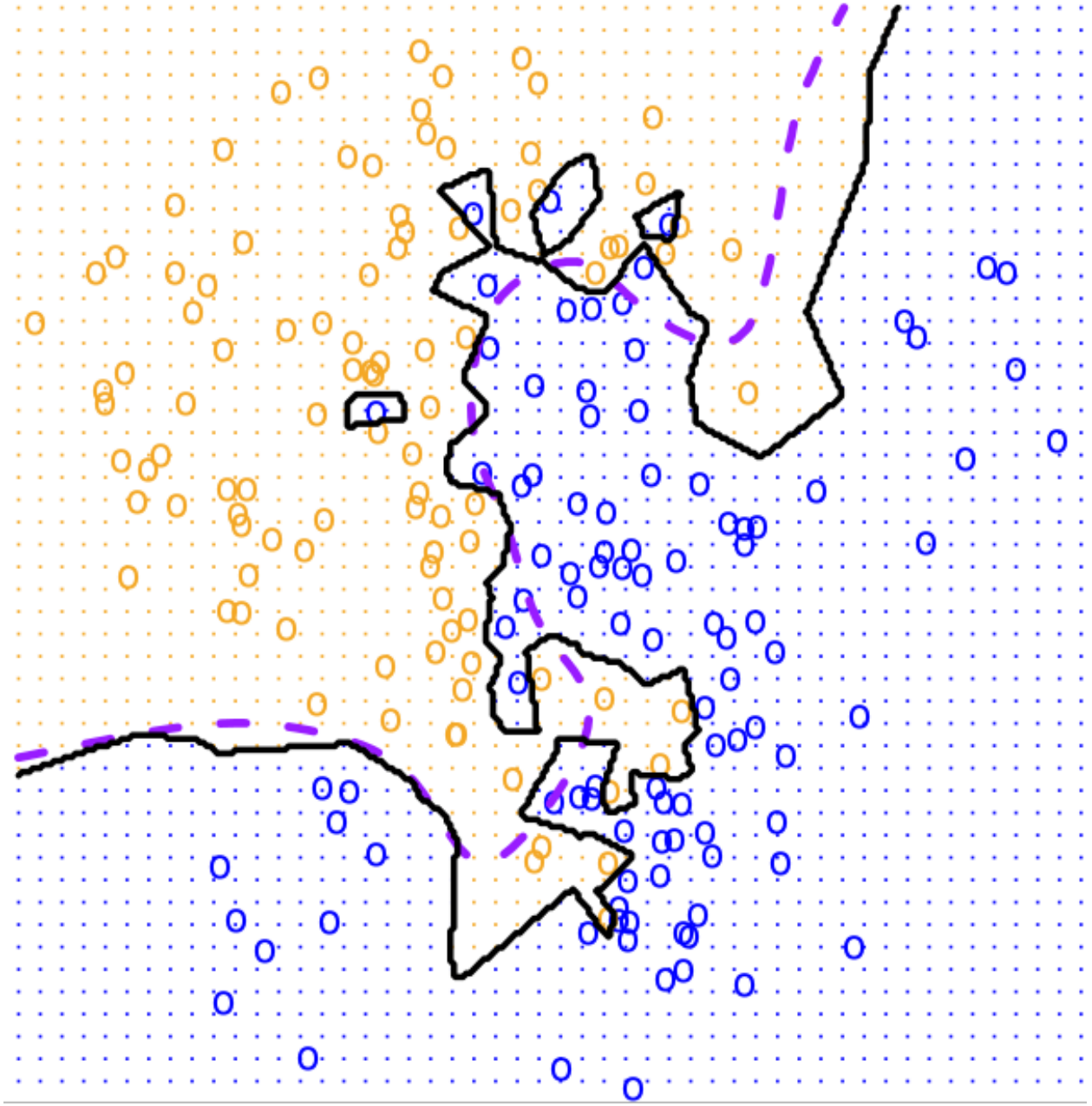
- **Goal of ML:** *Find a **good** and **useful** approximation of the data.*
- It's easy to build a model that performs well on the training data.
- But how well will it perform on **new data?**
- "Predictions are hard, especially about the future" — Niels Bohr.
- Challenges:
  - **Learn** models that **generalize** well.
  - **Evaluate** whether models generalize well.

Noise in the data

- Imprecision in **recording the features**
- **Errors in labeling** the data points (**teacher noise**)
- **Missing features** (**hidden** or **latent**)

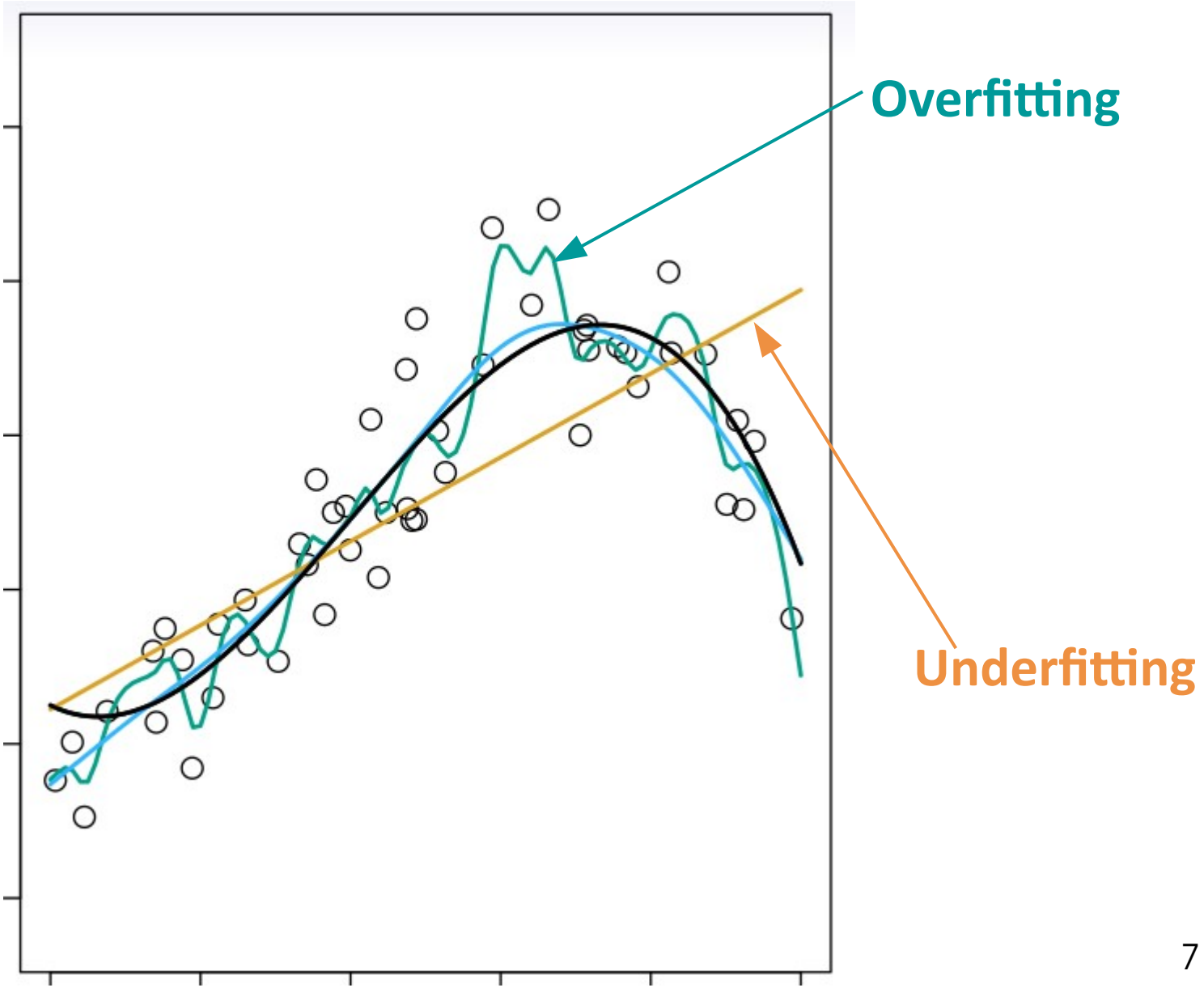
- Making no errors on the training set might not be possible.
- **Overfitting: You may learn the noise in addition to the signal.**

Overfitting

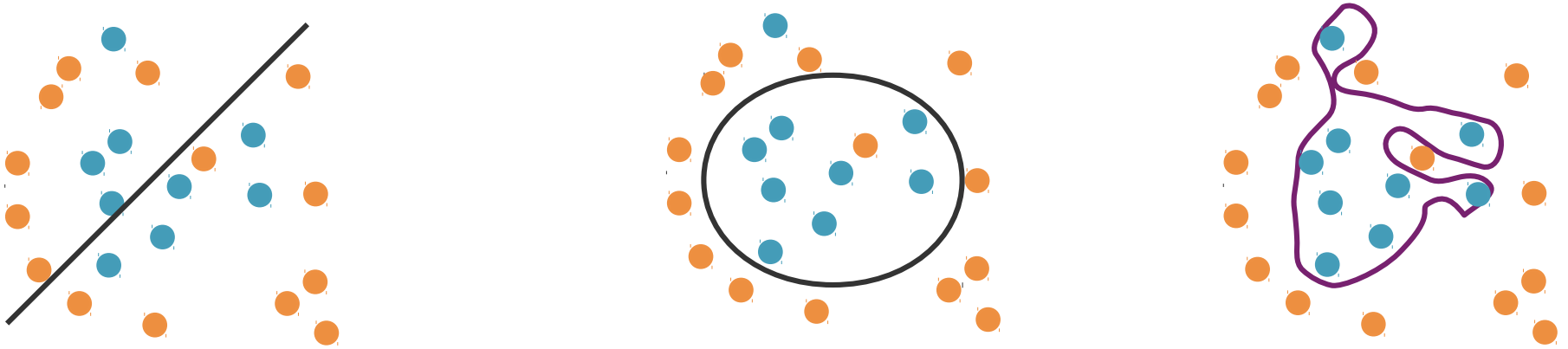


- What are the empirical errors of the black and purple classifiers?
- Which model seems more likely to be correct?

Overfitting & Underfitting (Regression)



Models of increasing complexity



Noise and model complexity

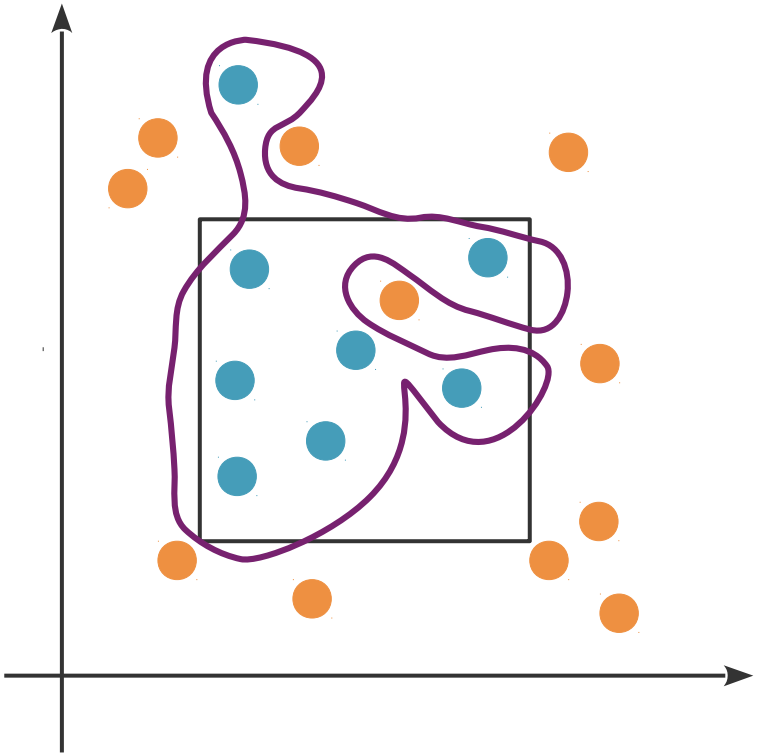
- **Use simple models!**
  - Easier to **use**

    lower computational complexity
  - Easier to **train**

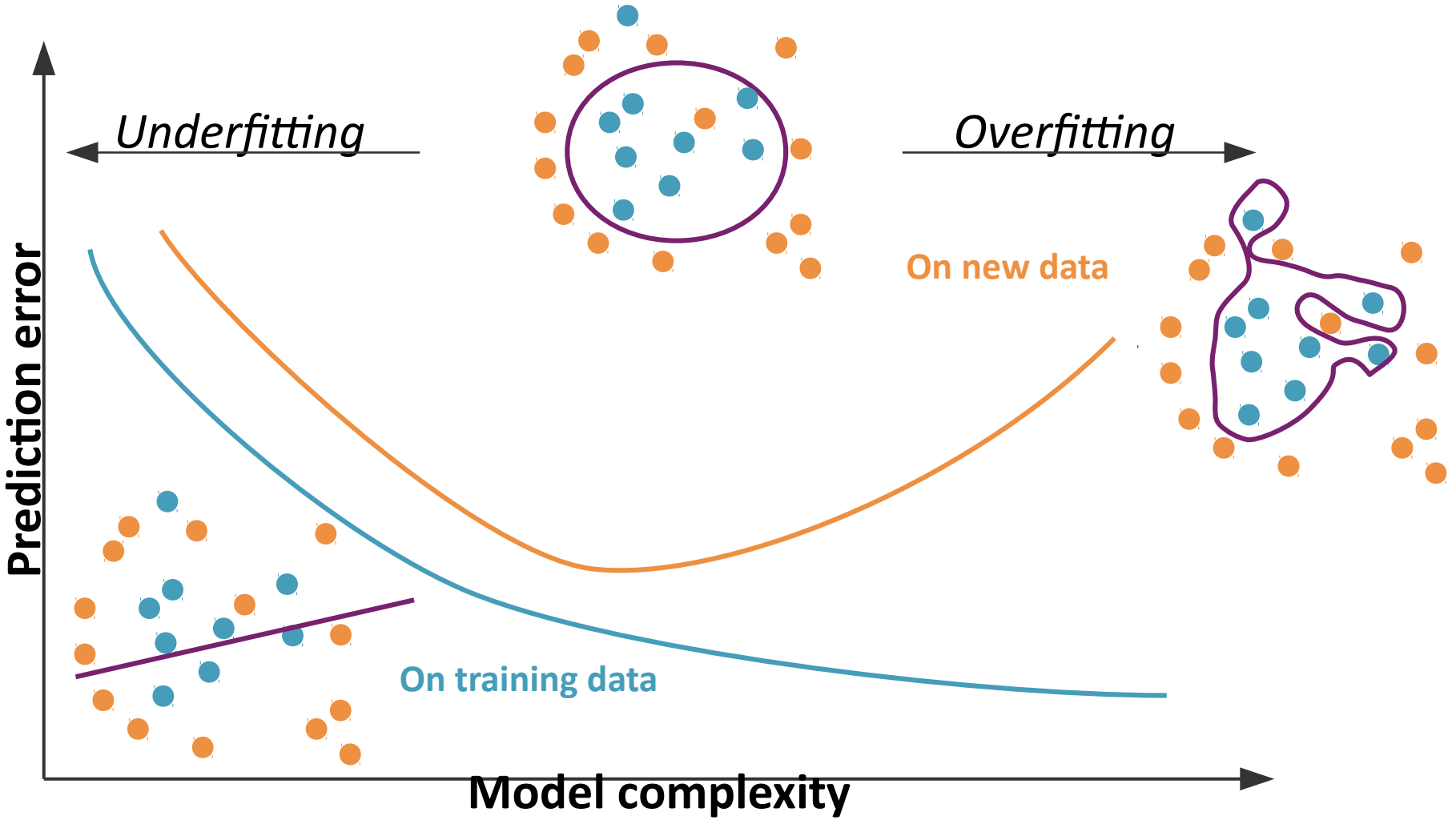
    lower space complexity
  - Easier to **explain**

    more interpretable
  - **Generalize better**

    Occam's razor: simpler explanations are more plausible.



Generalization error vs. model complexity



OPTIONAL

Bias-variance tradeoff

- **Bias:** difference between the expected value of the estimator and the true value being estimated.

$$\text{Bias}(f(\boldsymbol{x})) = \mathbb{E}[f(\boldsymbol{x}) - y]$$

  - A simpler model has a higher bias.
  - **High bias can cause underfitting.**

- **Variance:** deviation from the expected value of the estimates.

$$\text{Var}(f(\boldsymbol{x})) = \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}(f(\boldsymbol{x}))^2]$$

  - A more complex model has a higher variance.
  - **High variance can cause overfitting.**

OPTIONAL

Bias-variance decomposition

- $\mathrm{Bias}(f(\boldsymbol{x})) = \mathbb{E}[f(\boldsymbol{x}) - y]$
- $\mathrm{Var}(f(\boldsymbol{x})) = \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}(f(\boldsymbol{x}))^2]$
- **Mean squared error:**

$$\begin{aligned} \mathrm{MSE}(f(\boldsymbol{x})) &= \mathbb{E}[(f(\boldsymbol{x}) - y)^2] \\ &= \mathrm{Var}(f(\boldsymbol{x})) + \mathrm{Bias}^2(f(\boldsymbol{x})) \end{aligned}$$

- Proof 

OPTIONAL

Bias-variance decomposition

- $\text{Bias}(f(\boldsymbol{x})) = \mathbb{E}[f(\boldsymbol{x}) - y]$
- $\text{Var}(f(\boldsymbol{x})) = \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}(f(\boldsymbol{x}))^2]$
- **Mean squared error:**

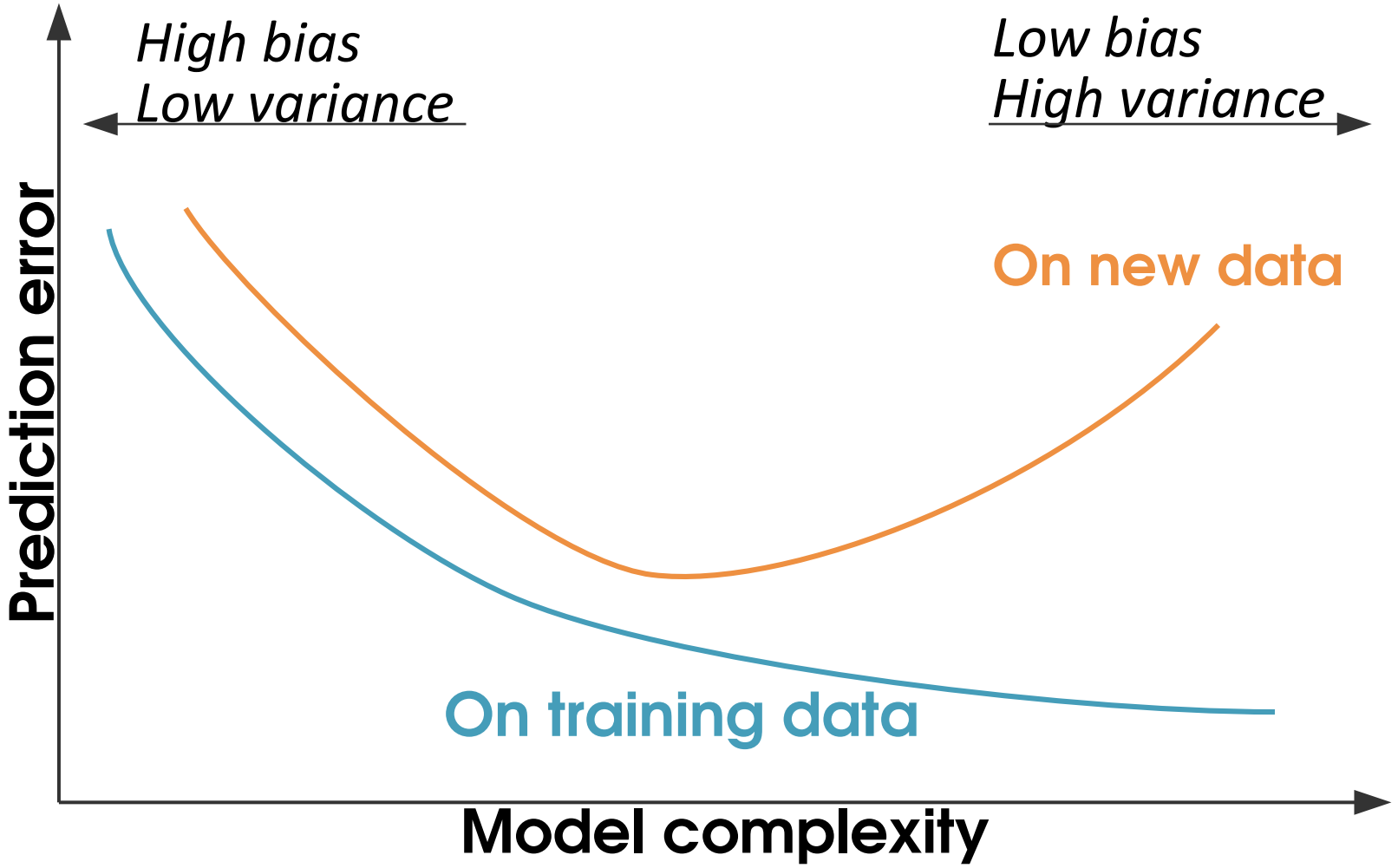
$$\begin{aligned} \text{MSE}(f(\boldsymbol{x})) &= \mathbb{E}[(f(\boldsymbol{x}) - y)^2] \\ &= \text{Var}(f(\boldsymbol{x})) + \text{Bias}^2(f(\boldsymbol{x})) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[(f(\boldsymbol{x}) - y)^2] &= \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})] + \mathbb{E}[f(\boldsymbol{x})] - y)^2] \\ &= \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})])^2] + \mathbb{E}[(\mathbb{E}[f(\boldsymbol{x})] - y)^2] + 2\mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})])(\mathbb{E}[f(\boldsymbol{x})] - y)] \end{aligned}$$

$\mathbb{E}[f(\boldsymbol{x})]$ and y are deterministic.

OPTIONAL

Generalization error vs. model complexity



OPTIONAL

Model selection & generalization

- **Well-posed problems:**
  - a solution exists;
  - it is unique;
  - the solution changes continuously with the initial conditions

Hadamard, on the mathematical modelisation of physical phenomena.

- Learning is an **ill-posed problem**:

  data helps carve out the hypothesis space

  but data is not sufficient to find a unique solution.

- Need for **inductive bias**

  assumptions about the hypothesis space

  model selection: choose the "right" inductive bias?

How do we decide a model is good?

Empirical Risk Minimization

- **Empirical risk** on dataset $\mathcal{D}$

$$E_{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y, f(\boldsymbol{x}))$$

- **Empirical Risk Minimization (ERM):** find f that minimizes the empirical risk

$$f^* = \arg\min_{f\in\mathcal{F}} E_{\mathcal{D}}(f).$$

Empirical Risk Minimization

- **Empirical risk** on dataset $\mathcal{D}$

$$E_{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y, f(\boldsymbol{x}))$$

- **Empirical Risk Minimization (ERM):** find f that minimizes the empirical risk

$$f^* = \arg\min_{f\in\mathcal{F}} E_{\mathcal{D}}(f).$$

- What we would actually like is to minimize the **expected risk:**

$$\mathbb{E}_{(X,Y)}\left[\mathcal{L}(Y, f(X))\right]$$

but we cannot access it.

ERM and generalization

- The empirical risk is a **poor estimate** of the **expected risk.**

  A model that learns by heart has an empirical risk of 0 and can have an arbitrarily large expected risk.

- The empirical risk is a **poor estimate** of the ability of a model to **generalize.**
- **Generalization error:** expected risk – empirical risk.
- **The expected risk is better estimated on data that has never been seen for training.**
- This means it has also **never been seen for exploratory data analyses steps** (feature engineering, dimensionality reduction, etc.) **nor for model selection.**

Validation sets

- Choose the model that performs best on a **validation set separate from the training set**.

- Because we have not used the validation data at any point during training, the validation set can be considered “new data” and **the error on the validation set is an estimation of the generalization error.**

Model selection

- What if we want to choose among k models?
  - Train each model on the train set
  - Compute the prediction error of each model on the validation set
  - Pick the model with the smallest prediction error on the validation set.
- What is the expected risk?
  - We don't know!
  - Validation data was used to select the model
  - We have "cheated" and looked at the validation data: it is not a good proxy for new, unseen data any more.

Validation sets

- Hence we need to set aside part of the data, the test set, that remains untouched during the entire procedure and on which we'll estimate the generalization error.

- Model **selection:** pick the best model.
- Model **assessment:** estimate its prediction error on new data.

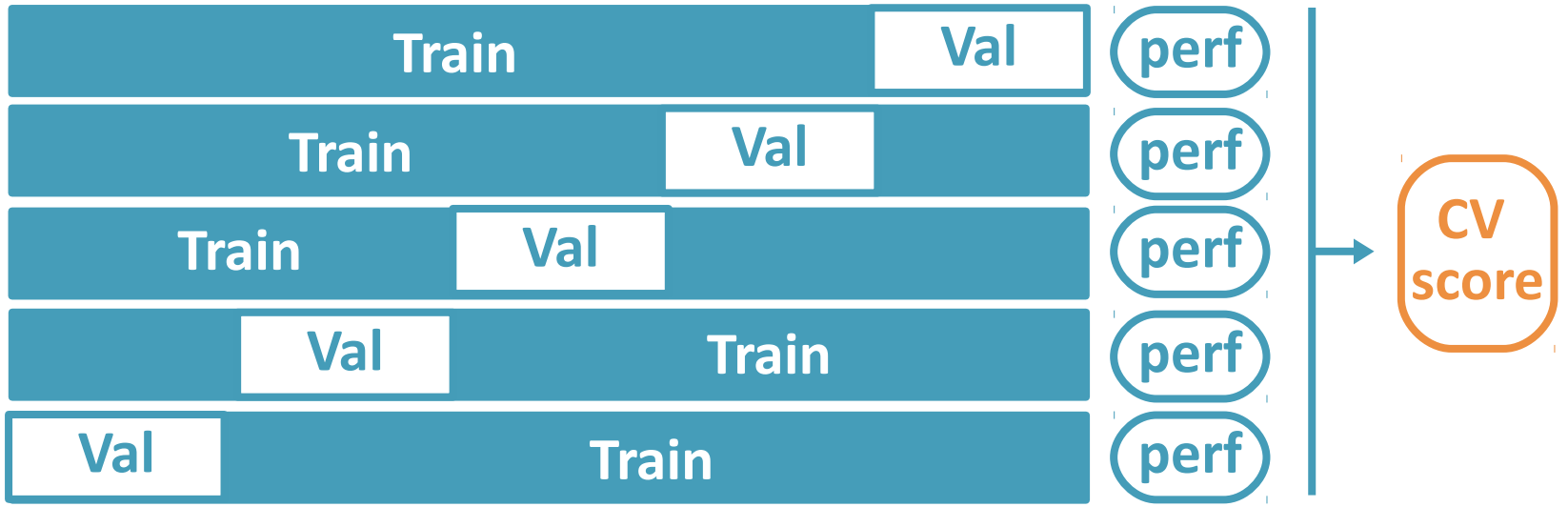
- **How much data** should go in each of the training, validation and test sets?
- How do we know we have **enough data** to evaluate the prediction and generalization errors?

- **Empirical evaluation with sample re-use**
  - cross-validation
  - bootstrap
- **Analytical tools**
  - Mallow's Cp, AIC, BIC
  - MDL.

Sample re-use

Cross-validation

- Cut the training set in k separate **folds.**
- For each fold,
  - **train** on the (k-1) remaining folds.
  - **estimate performance** on this fold.
- Average the performance to obtain a **CV score.**



Issues with cross-validation

- **Training set size** becomes (K-1)n/K

  Why is this a problem? 

Issues with cross-validation

- **Training set size** becomes (K-1)n/K
  - small training set ⇒ biased estimator of the error
- **Leave-one-out cross-validation:** K = n
  - approximately **unbiased estimator** of the expected prediction error
  - potential **high variance** (the training sets are very similar to each other)
  - **computation** can become burdensome (n repeats)
- In practice: set **K = 5 or K = 10.**

OPTIONAL

Bootstrap

- **Randomly draw datasets with replacement** from the training data
- **Repeat B times** (typically, B=100) ⇒ B models
- **Leave-one-out bootstrap error:**
  - For each training point i, predict with the $b_i < B$ models that did not have i in their training set
  - Average prediction errors
- Each training set contains 

OPTIONAL

Bootstrap

- **Randomly draw datasets with replacement** from the training data
- **Repeat B times** (typically, B=100) ⇒ B models
- **Leave-one-out bootstrap error:**
  - For each training point i, predict with the $b_i < B$ models that did not have i in their training set
  - Average prediction errors
- Each training set contains **0.632.n distinct examples**

  ⇒ same issue as with cross-validation

$$\begin{aligned} Pr(i \in X_k) &= 1 - (1 - \tfrac{1}{n})^n \\ &\sim 1 - e^{-1} \\ &= 0.632 \end{aligned} \qquad e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Evaluating model performance

Classification model evaluation

- **Confusion matrix**

		True class	
		-1	+1
Predicted class	-1	True Negatives	False Negatives
	+1	False Positives	True Positives

- False positives (false alarms) are also called **type I errors**
- False negatives (misses) are also called **type II errors**

- **Sensitivity** = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

 # positives

- **Specificity** = True negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

- **Precision** = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

 # predicted positives

- **False discovery rate** (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}$$

- **Accuracy**

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$$

- **F1-score** = harmonic mean of precision and sensitivity.

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

OPTIONAL

Example: Pap smear

- 4,000 apparently healthy women of age 40+
- Tested for cervical cancer through pap smear and histology (gold standard)

	Cancer	No cancer	Total
Positive test	190	210	400
Negative test	10	3590	3600
Total	200	3800	4000

- **What are the sensitivity, specificity, and PPV of the test?**

OPTIONAL

- **Sensitivity** = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- **Specificity** = True negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

- **Precision** = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

	Cancer	No cancer	Total
Positive test	190	210	400
Negative test	10	3590	3600
Total	200	3800	4000

OPTIONAL

- In this population:

Sensitivity = 95.0 % Specificity = 94.5 % PPV = 47.5 %

	Cancer	No cancer	Total
Positive test	190	210	400
Negative test	10	3590	3600
Total	200	3800	4000

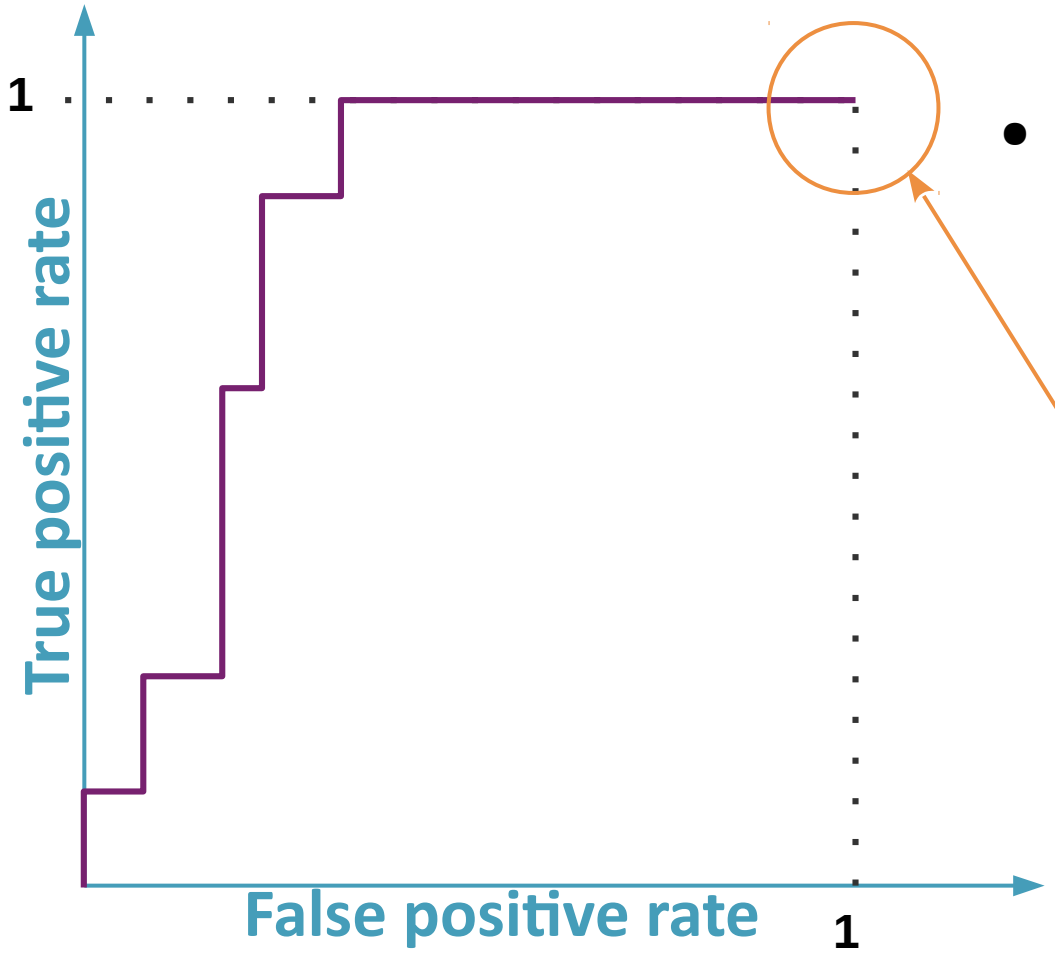
- **Prevalence** of the disease = 200/4000 = 0.05
- P(cancer|positive test) = PPV = **47.5 %**
- P(no cancer|negative test) = 3590/3600 = **99.7 %**

- Poor **diagnosis** tool
- Good **screening** tool

OPTIONAL

ROC curves

- ROC = Receiver-Operator Characteristic.
- Summarized by the area under the curve (AUROC).



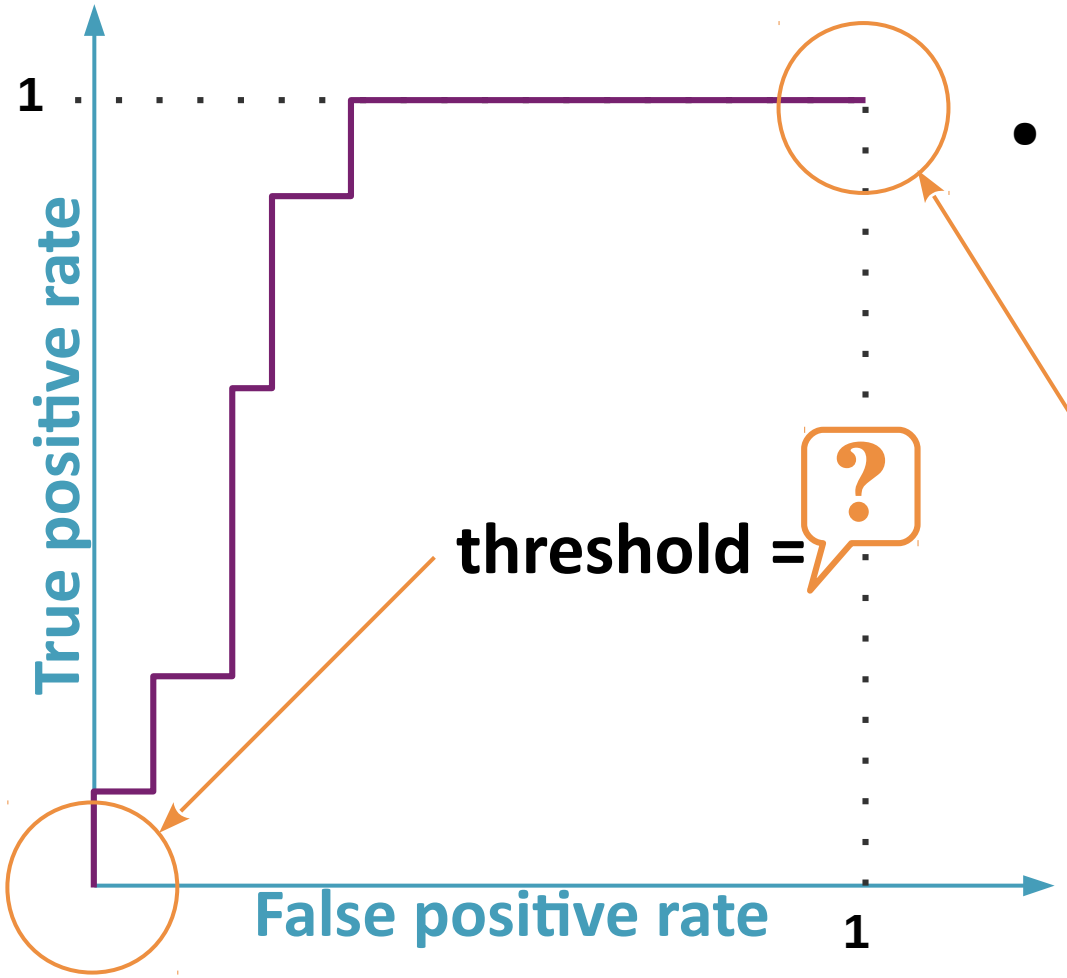
- Plot TPR vs FPR for all possible thresholds.

threshold = ?

OPTIONAL

ROC curves

- ROC = Receiver-Operator Characteristic.
- Summarized by the area under the curve (AUROC).



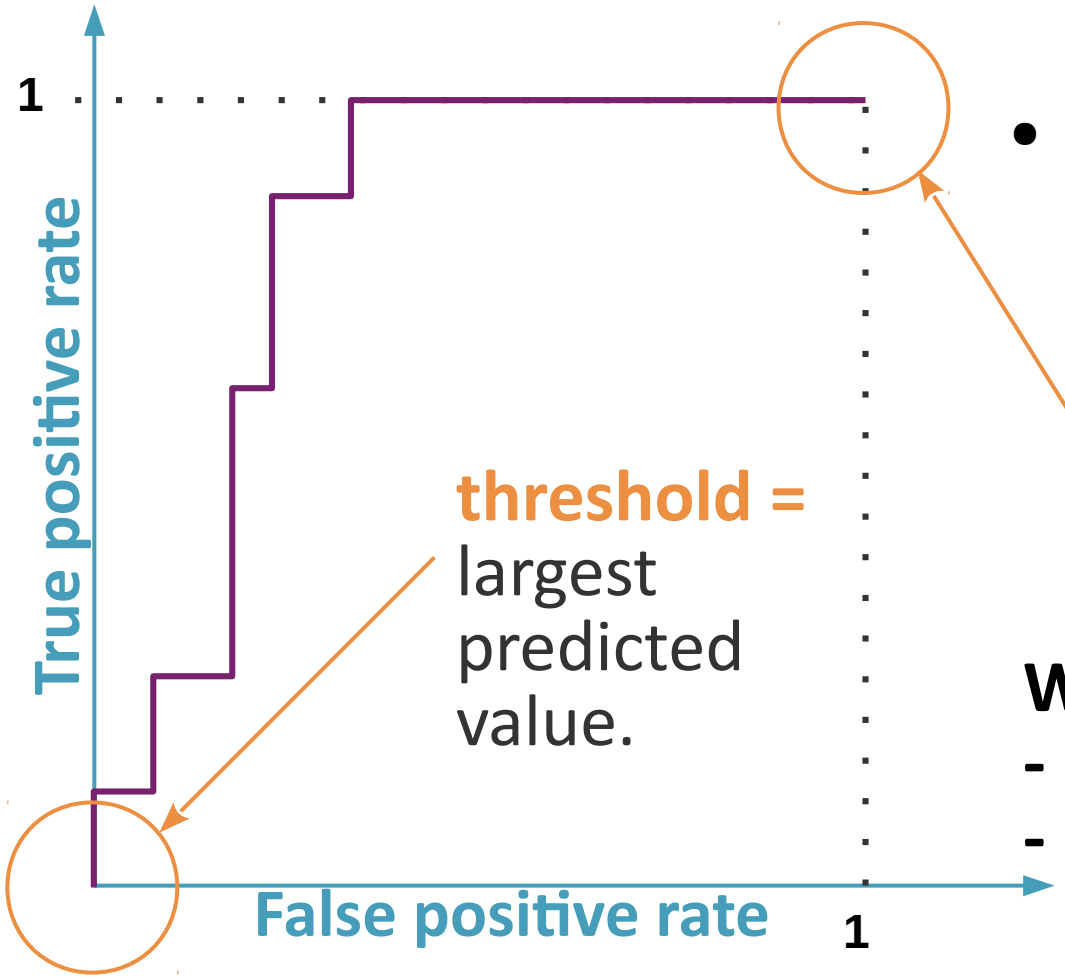
- Plot TPR vs FPR for all possible thresholds.

threshold = smallest predicted value.

OPTIONAL

ROC curves

- ROC = Receiver-Operator Characteristic.
- Summarized by the area under the curve (AUROC).



- Plot TPR vs FPR for all possible thresholds.

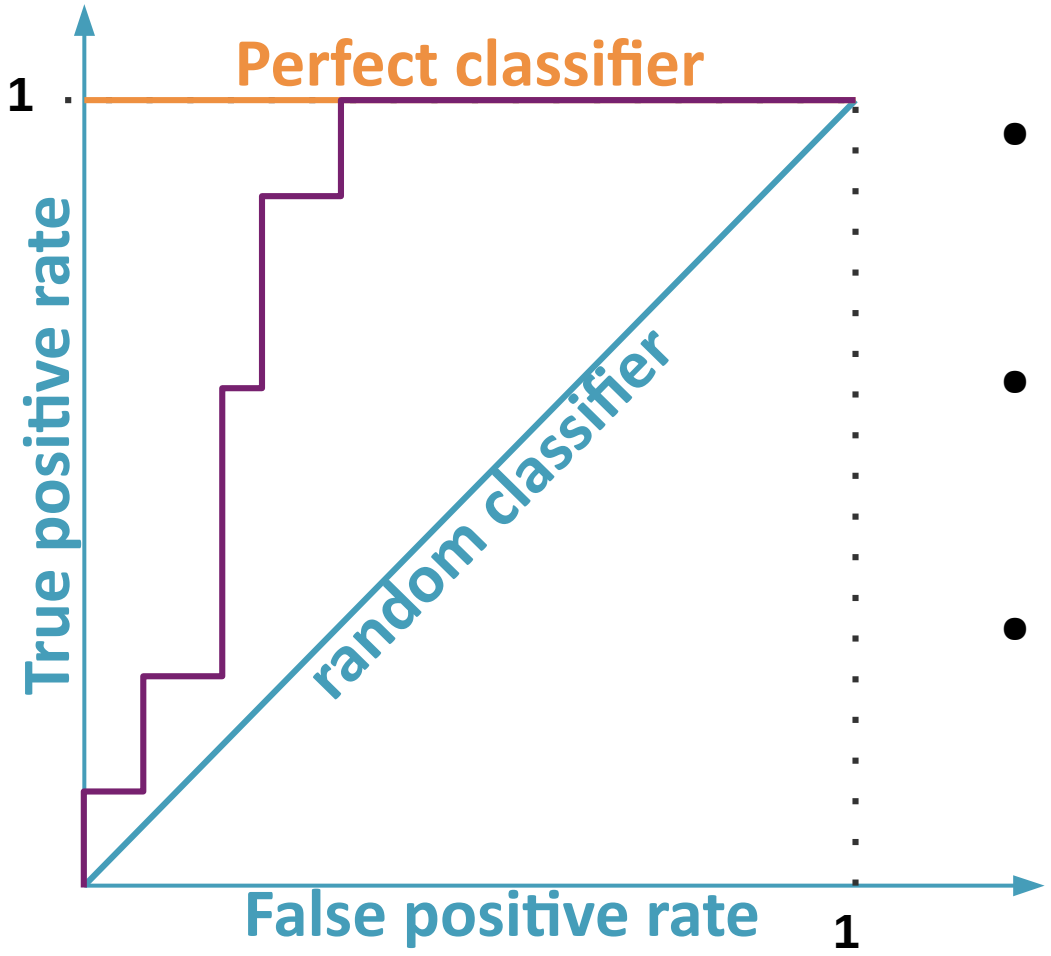
threshold = largest predicted value.

threshold = smallest predicted value.

What is the ROC curve of:
- a random classifier?
- a perfect classifier?

ROC curves

- ROC = Receiver-Operator Characteristic.
- Summarized by the area under the curve (AUROC).



- **Perfect classifier:**

  AUROC = 1.0
- **Random classifier:**

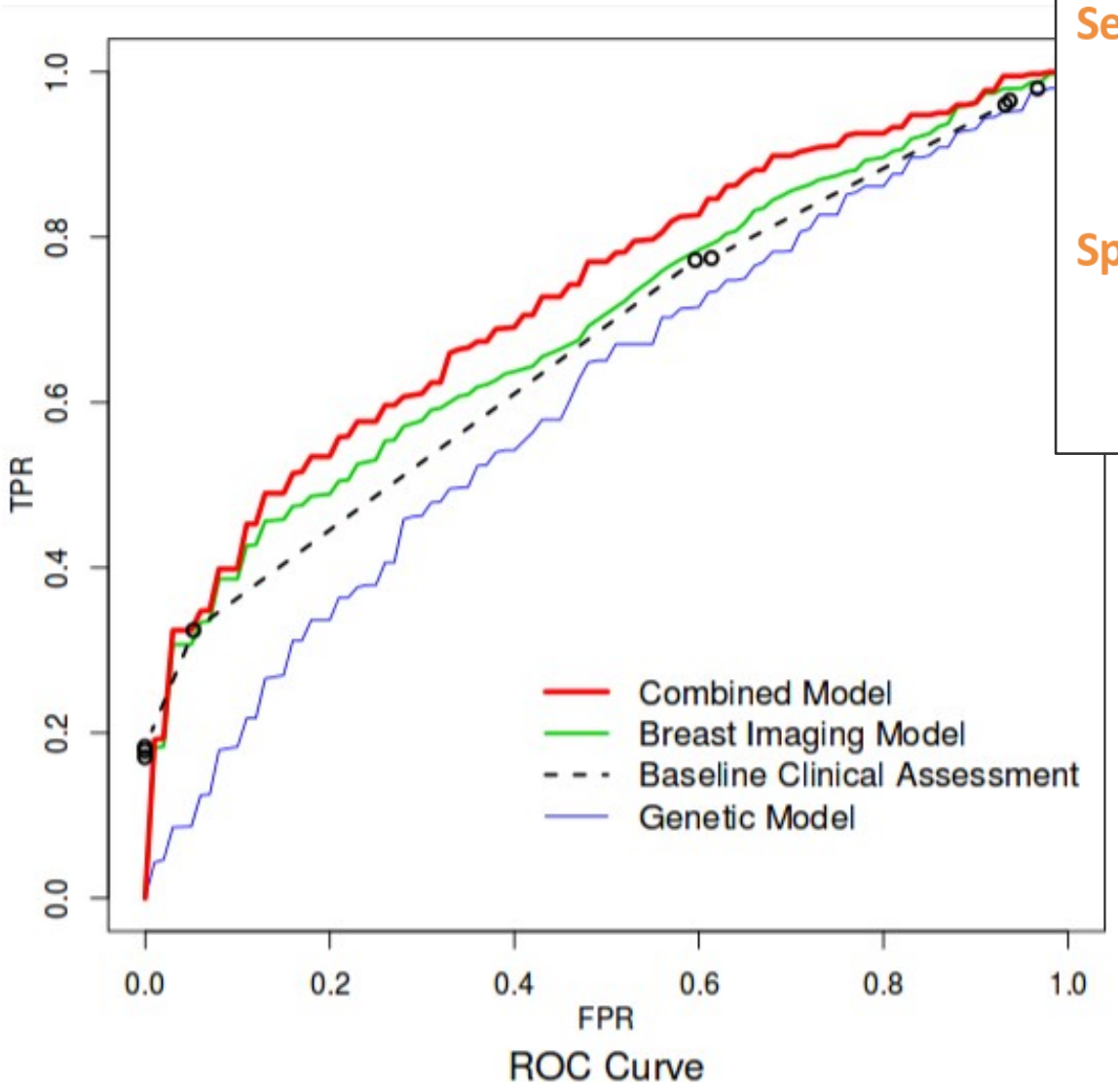
  AUROC = 0.5
- **Our classifier:**

  0.5 < AUROC < 1.0

OPTIONAL

Predicting breast cancer risk based on mammography images, SNPs, or both.

Liu J, Page D, Nassif H, et al. (2013). **Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms.** *AMIA Annual Symposium Proceedings.* 876-885.



Sensitivity = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Specificity = True negative rate (TNR) = 1 - FPR

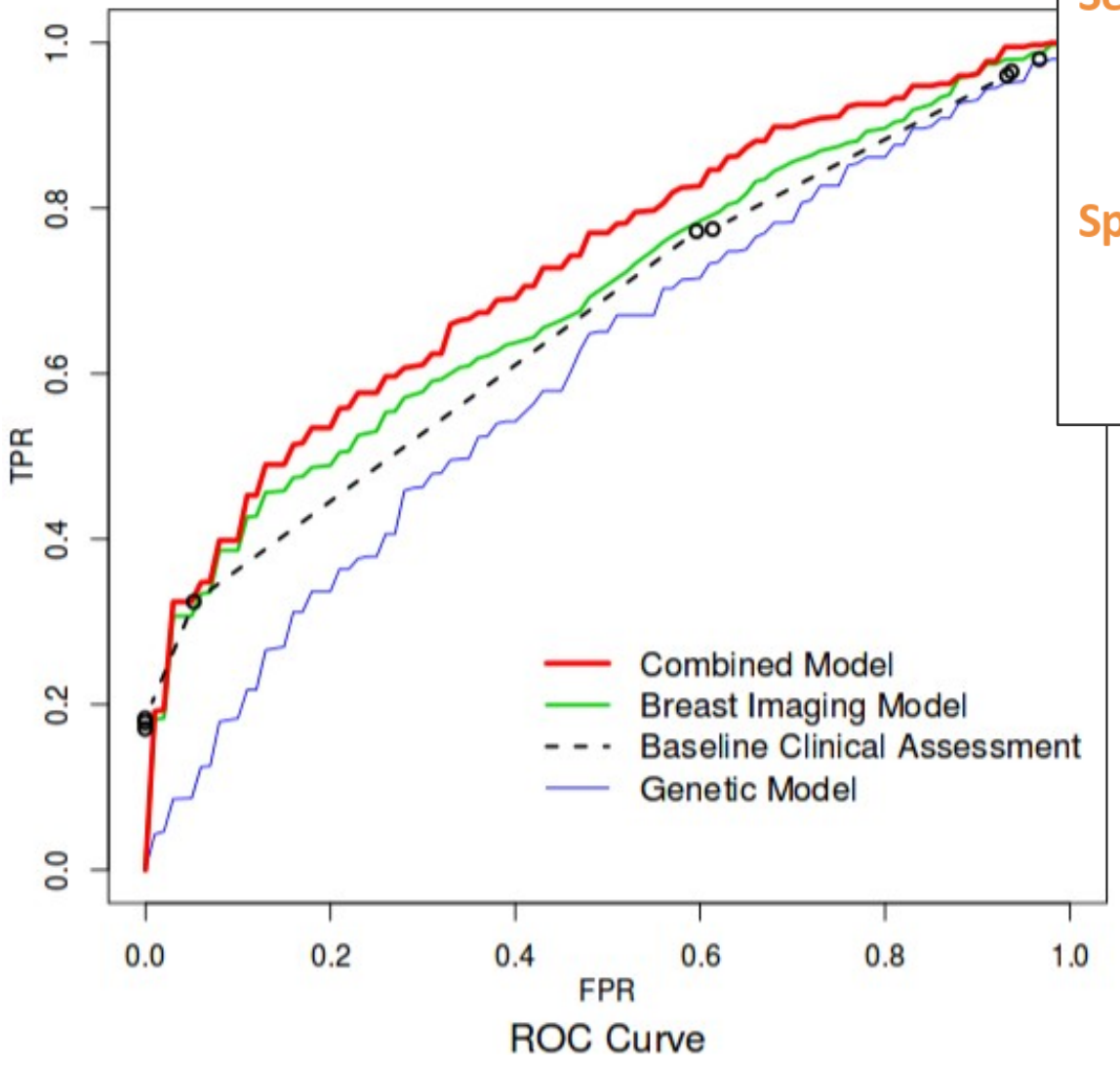
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

- **Which method outperforms the others?**
- **Is a low FPR or high TPR preferable in a clinical setting?**

OPTIONAL

Predicting breast cancer risk based on mammography images, SNPs, or both.

Liu J, Page D, Nassif H, et al. (2013). **Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms.** *AMIA Annual Symposium Proceedings.* 876-885.



Sensitivity = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

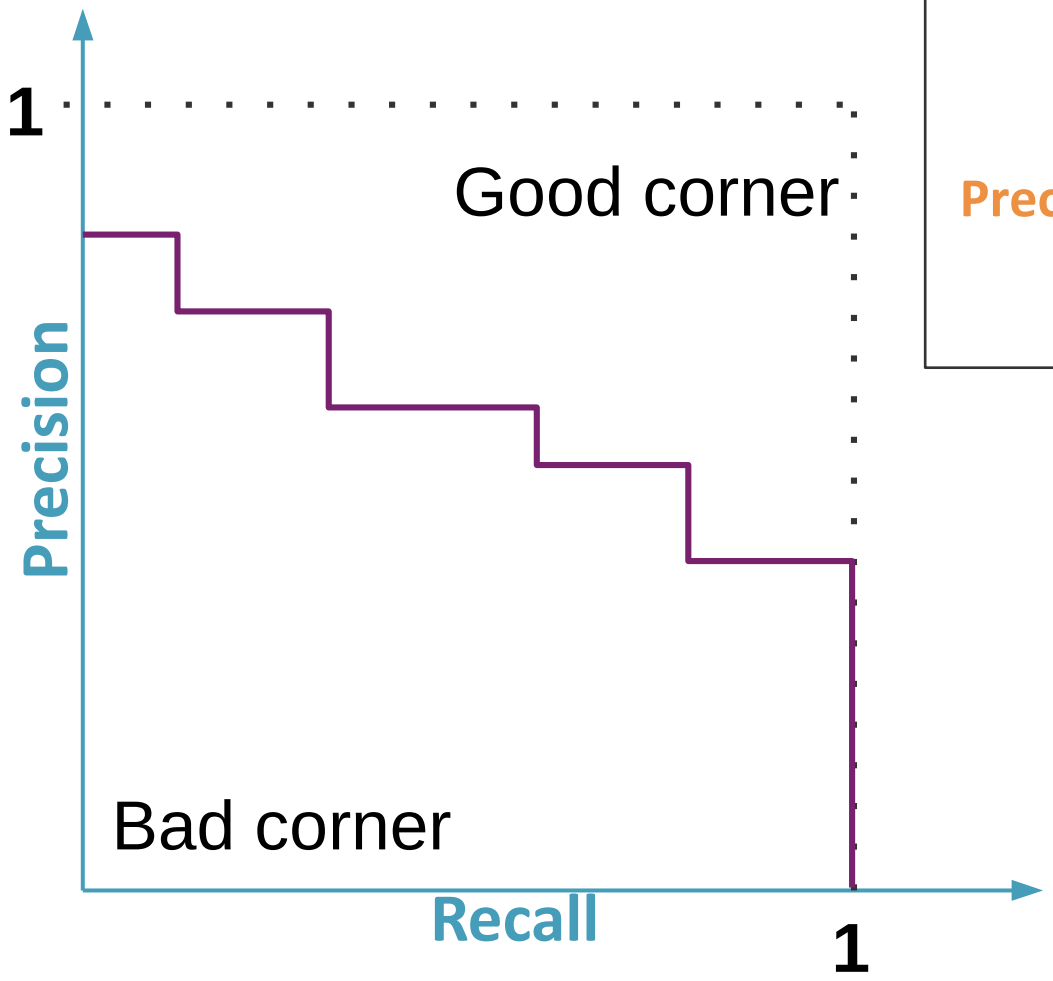
Specificity = True negative rate (TNR) = 1 - FPR

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

High recall = fewer chances to miss a case

High specificity / low FPR = fewer false alarms

Precision-Recall curves



Sensitivity = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

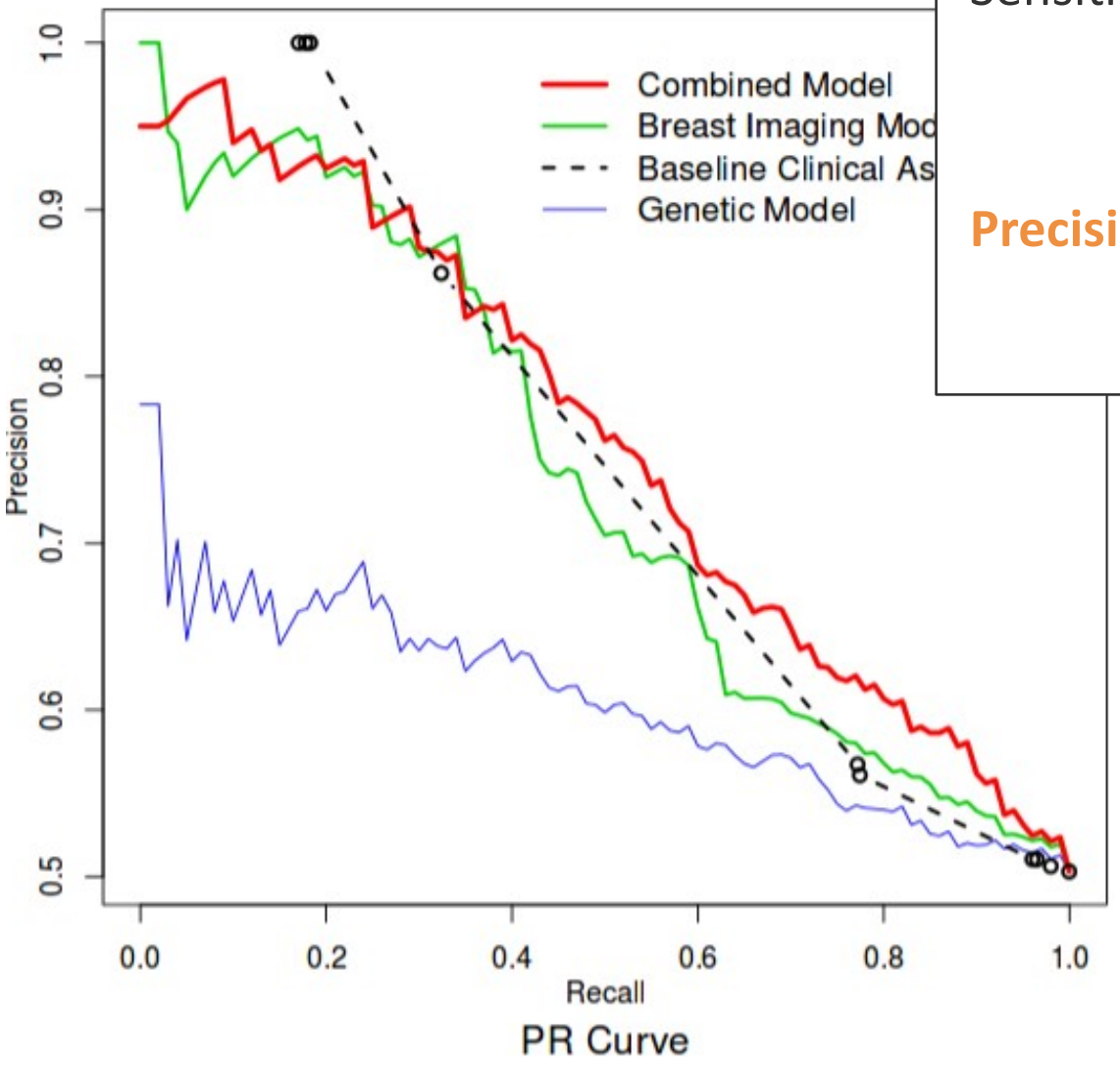
Precision = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

OPTIONAL

Predicting breast cancer risk based on mammography images, SNPs, or both.

Liu J, Page D, Nassif H, et al. (2013). **Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms.** *AMIA Annual Symposium Proceedings.* 876-885.



Sensitivity = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision = Positive predictive value (PPV)

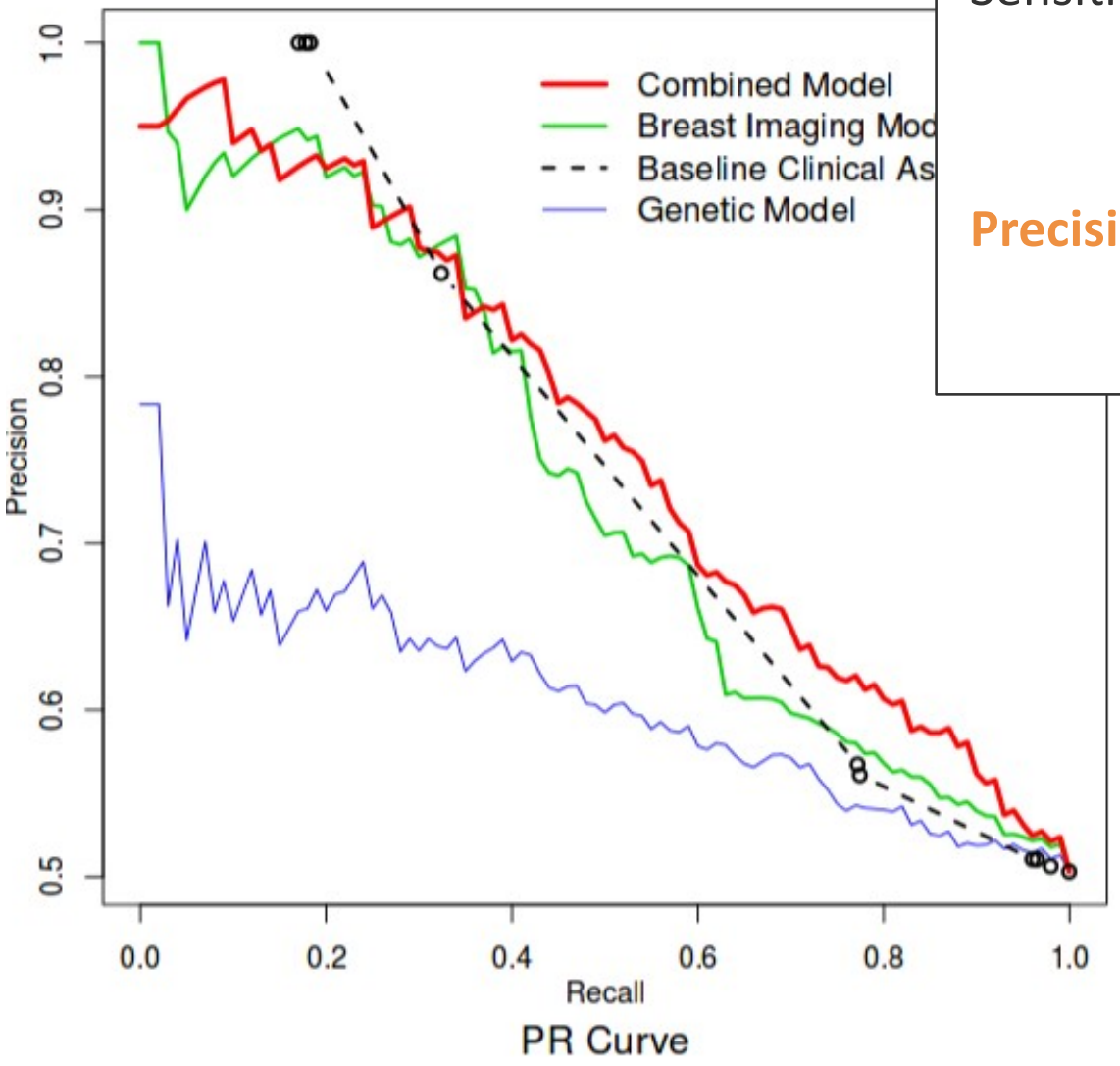
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- **Which method has the highest area under the PR curve?**
- **Is a high recall or high precision preferable in a clinical setting?**

OPTIONAL

Predicting breast cancer risk based on mammography images, SNPs, or both.

Liu J, Page D, Nassif H, et al. (2013). **Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms.** *AMIA Annual Symposium Proceedings.* 876-885.



Sensitivity = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

High recall = fewer chances to miss a case

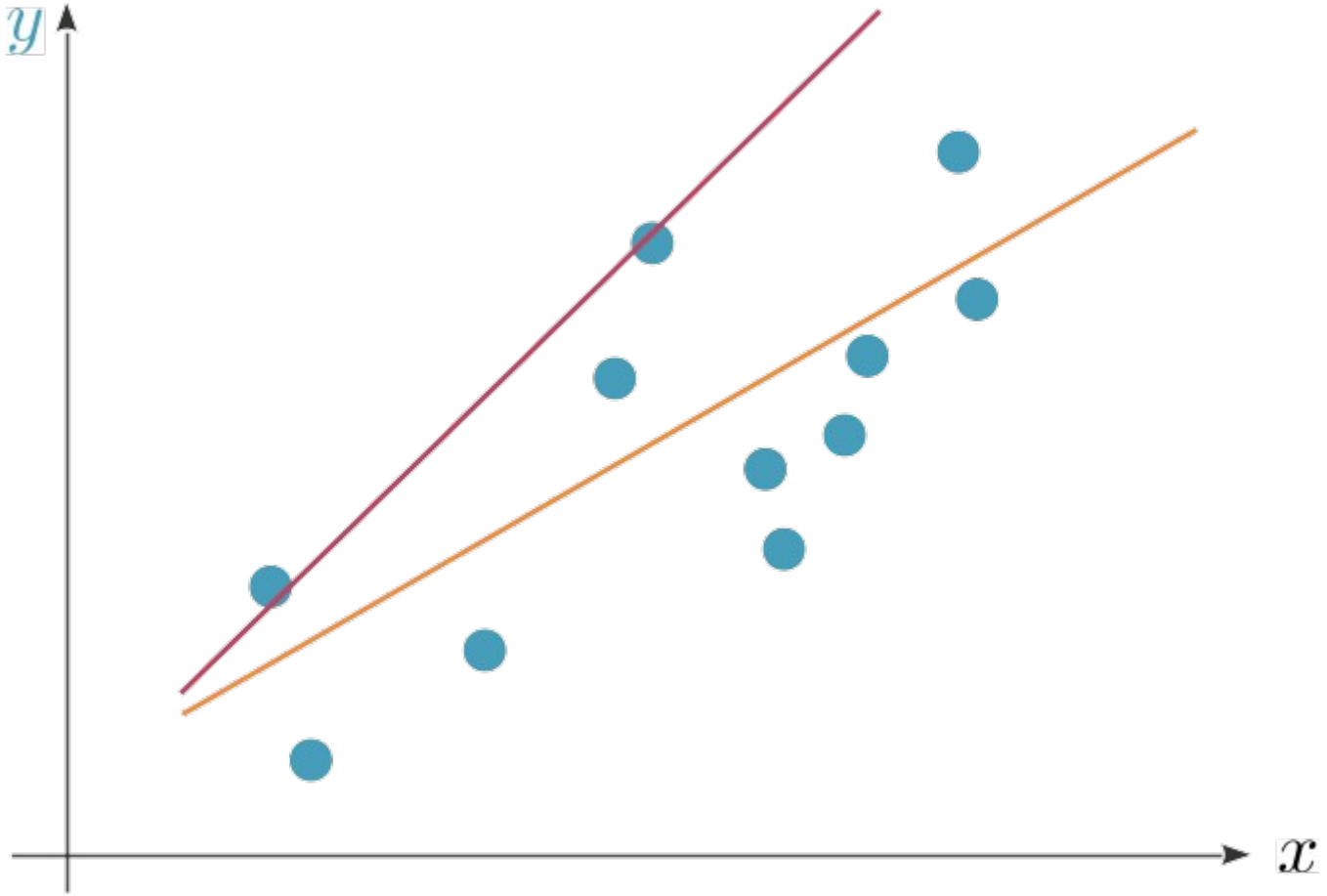
High precision = substantially more true diagnoses than false alarms

Regression model evaluation

- Counting the number of errors is not reasonable

Regression model evaluation

- Counting the number of errors is not reasonable
  - What does error even mean for numerical values?
  - Not all errors are created equal.



Regression model evaluation

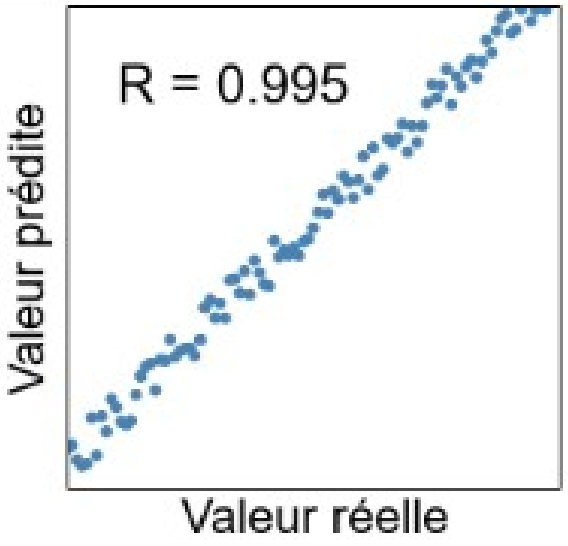
- **Residual sum of squares** $\mathrm{RSS} = \sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2$
- **Root-mean squared error**

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2}$$

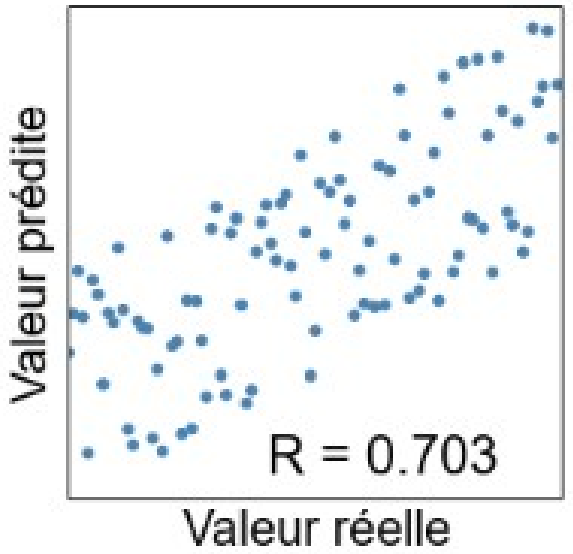
- **Relative squared error** $\mathrm{RSE} = \frac{\sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2}{\sum_{i=1}^{n} (y^i - \bar{y})^2}$
- **Coefficient of determination**

$$R^2 = 1 - \mathrm{RSE} = \frac{\sum_{i=1}^{n} (y^i - \bar{y})(f(\boldsymbol{x}^i) - \overline{f(\boldsymbol{x})})}{\sqrt{\sum_{i=1}^{n} (y^i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (f(\boldsymbol{x}^i) - \overline{f(\boldsymbol{x})})}}$$

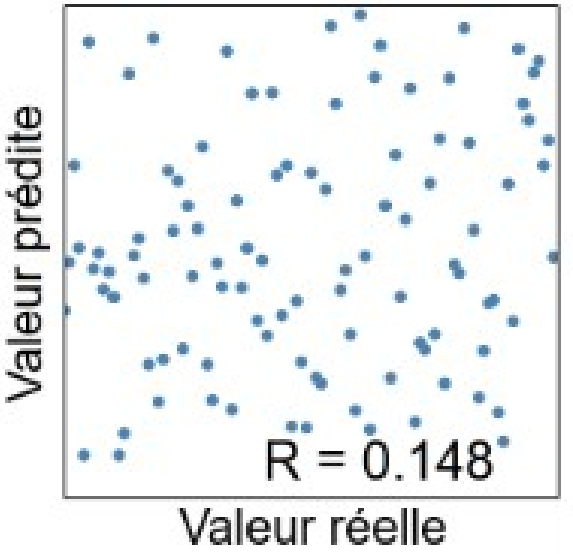
Correlation between true and predicted values



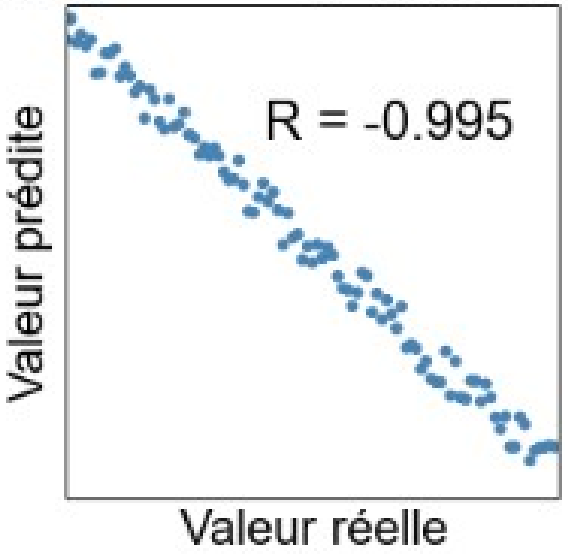
$R^2 = 0.990$

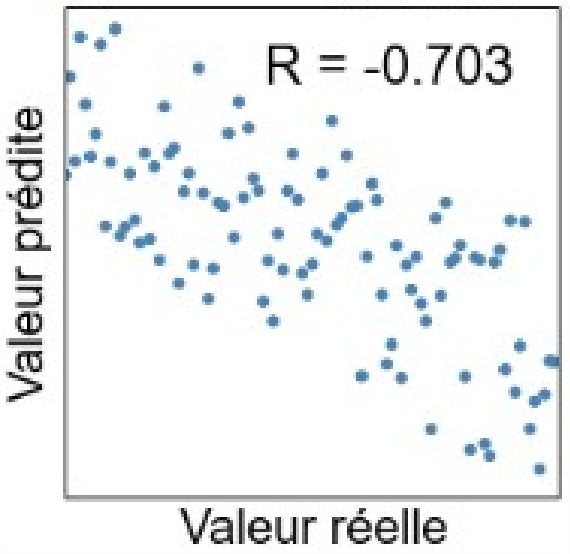


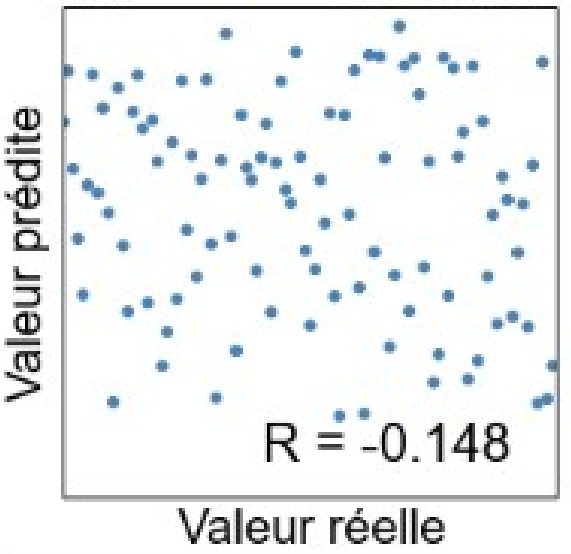
$R^2 = 0.494$



$R^2 = 0.022$







Analytical tools and model complexity

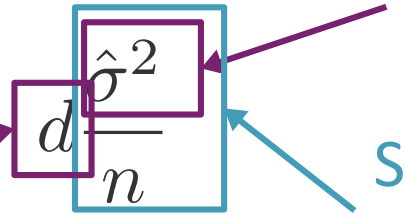
OPTIONAL

Optimism terms

- Correct the empirical error with an **optimism term**
- Theoretical estimate of the **discrepancy between training and test error**

Augmented error = empirical error + optimism term

- For **linear models**, optimism terms proportional to:
  - **Mallow's Cp:**



Variance of the residuals on the train set

Squared standard error of the mean of the residuals

parameters
= # non-zero coefficients

  - **Akaike Information Criterion (AIC):** d
  - **Bayesian Information Criterion (BIC):** $d \ln(n)$

OPTIONAL

Minimum description length (MDL)

Shortest code to transmit a random variable z:

$-\log_2 P(z)$ [Shannon's source coding theorem]

Consider discrete variable **z**

- Equiprobable case: use a **fixed-length code**

$$a \mapsto 00 \qquad b \mapsto 01 \qquad c \mapsto 10 \qquad d \mapsto 11$$

- Otherwise: use a **variable-length prefix code** in which frequent values get shorter codes

$$a \mapsto 1 \qquad b \mapsto 10 \qquad c \mapsto 110 \qquad d \mapsto 111$$

The prefix separates codes

OPTIONAL

Minimum description length (MDL)

- **Shortest code to transmit a random variable z:**

$$-\log_2 P(\boldsymbol{z})$$ [Shannon's source coding theorem]

- Assume
  - Parametric model f_θ
  - receiver knows inputs X, model family f.
- To transmit outputs y, need

$$\underbrace{-\log_2 P(\boldsymbol{y}|\theta, f, X)}_{\text{average code length to transmit the difference between model prediction and true outputs.}} \quad \underbrace{-\log_2 P(\theta)}_{\text{average code length to transmit } \theta.}$$

- Choose the model with smallest Kolmogorov complexity (=MDL)

Summary: model selection techniques

- **Empirical:**

  Estimate quality of generalization with

  – **cross-validation**

  – **bootstrap**

- **Theoretical:**

  – Estimate the difference between train error and generalization error with an optimism term

    E.g. Mallow's Cp, Akaike's / Bayesian Information Criteria

  – **Minimum description length (MDL)**

    Choose simplest model (according to Kolmogorov complexity)

OPTIONAL

References

- *A Course in Machine Learning.*
  http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf
  - **Noise:** Chap 2.3
  - **Overfitting:** Chap 2.4
  - **Bias-variance tradeoff:** Chap 5.9
  - **Train and test sets:** Chap 2.5
  - **Cross-validation:** Chap 5.6
  - **Performance measures:** Chap 5.5
- *The Elements of Statistical Learning.*
  http://web.stanford.edu/~hastie/ElemStatLearn/
  - **Overfitting:** Chap 7.1
  - **Bias-variance tradeoff:** Chap 2.9, 7.2–7.3
  - **Cross-validation:** Chap 7.10
  - **Bootstrap:** Chap 7.11
  - **Mallow's Cp, AIC, BIC:** Chap 7.7
  - **MDL:** Chap 7.8
- **Entropy encoding:**
  http://lesswrong.com/lw/o1/entropy_and_short_codes/

References for prerequisites

- **Linear algebra:**

  http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/

- **Statistics & probabilities:**
  - **Probability theory: A primer (Jeremy Kun)**
    http://jeremykun.com/2013/01/04/probability-theory-a-primer/
  - **Probability Primer (Jeffrey Miller)**
    https://www.youtube.com/playlist?list=PL17567A1A3F5DB5E4

kaggle challenge project

- Work **alone or in pairs**
  - Engineer features (see Lab 4)
  - **Model selection** for several approaches
  - Predict with selected models and submit to leaderboard
  - Choose **2 final models**
- Deadline: **February 2nd, 2020 23:59**
  - **Report** (2 pages + figures/tables)
  - **Leaderboard position**
- **Full instructions:**

`http://cazencott.info/dotclear/public/lectures/hpcai_2019-2020/kaggle-project.pdf`

kaggle challenge project

How Many Shares? Challenge

`https://www.kaggle.com/c/how-many-shares-1920`

- **Predict the number of shares on social media** for articles from the same media site
  - Regression
  - From article length, topics, subjectivity and much more.
- **Evaluation on**
  - Insights learned
  - Prediction performance.

Kaggle leaderboard setup

- The data is divided into:
  - Training data
  - Public validation data
  - Test validation data
- You only have the labels of the training data
- You make predictions for the **whole validation set**
- The **public part** is used to rank you on the **public leaderboard** throughout the challenge
- The **private part** is used to determine your **final ranking** at the end.