Foundations of Machine Learning
CentraleSupélec — Fall 2016

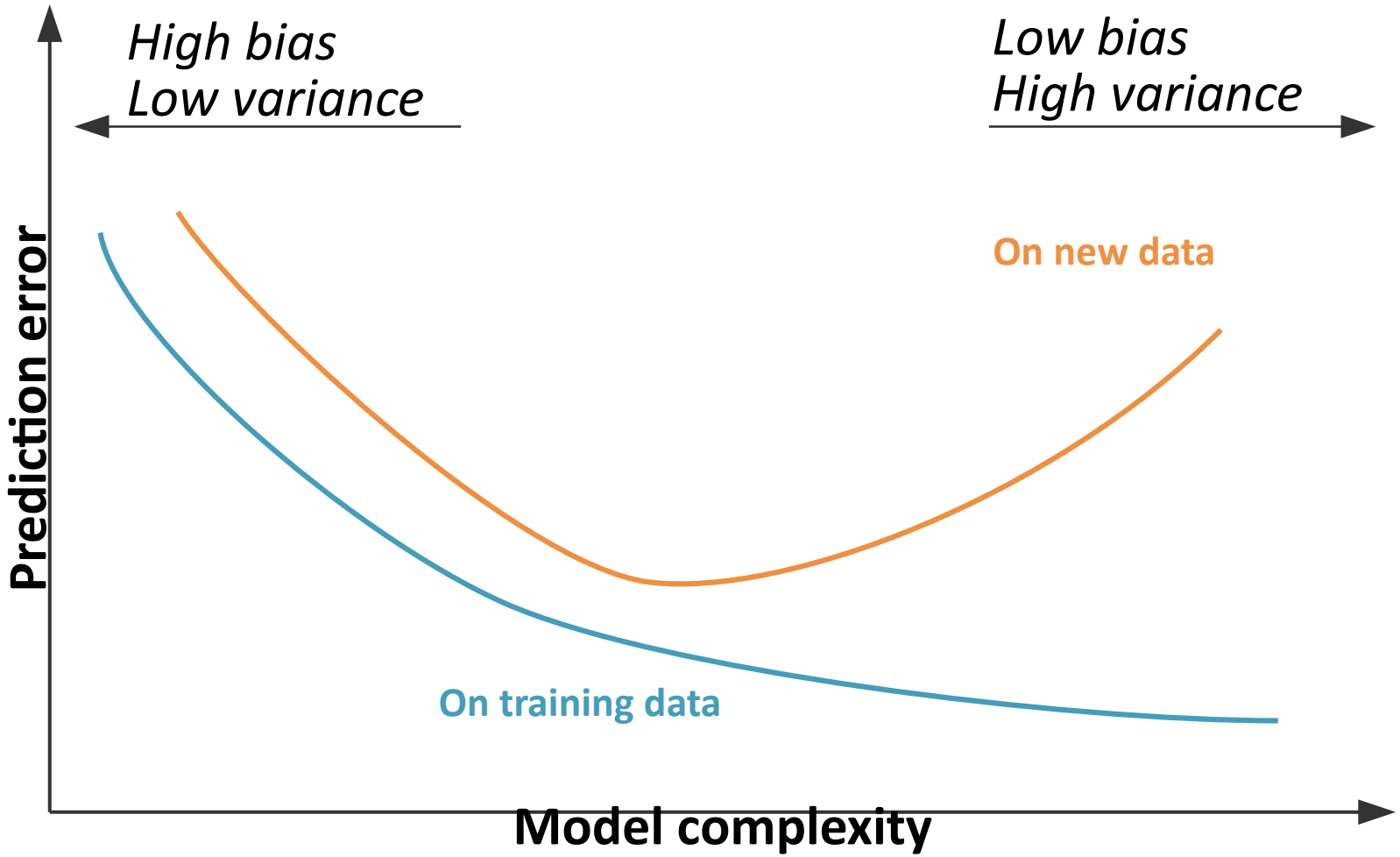
3. Model evaluation & selection

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

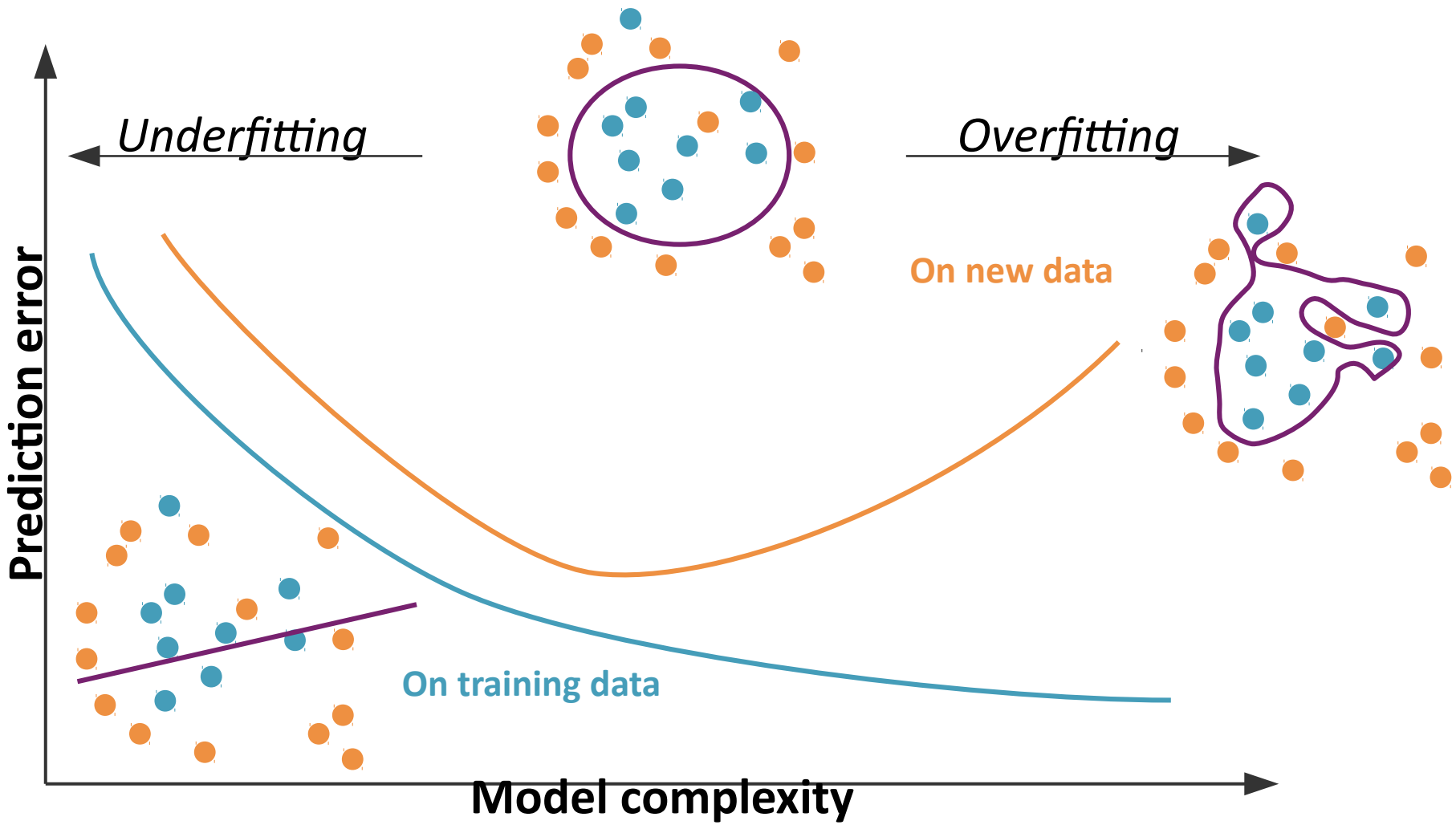
Practical matters

- **Scribes**
  - One person signed up for today. Anyone wants to assist her?
  - Two people signed up for next week. Congrats!
  - No one signed up after that.

Generalization error vs. model complexity



Generalization error vs. model complexity



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 H

  model selection: choose the "right" inductive bias?

How do we decide a model is good?

Learning objectives

After this lecture you should be able to

design experiments to select and evaluate supervised machine learning models.

Concepts:

- training and testing sets;
- cross-validation;
- bootstrap;
- measures of performance for classifiers and regressors;
- measures of model complexity.

Supervised learning setting

- **Training set:**

$$\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$$

- **Classification:** $y^i \in \{0, 1\}$
- **Regression:** $y^i \in \mathbb{R}$
- Goal: Find f,θ such that $f(\boldsymbol{x}^i|\theta)$ approximates y^i.
- **Empirical error** of f on the training set, given a **loss**:

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

  - E.g. (classification)

$$L(\{y^i\}, \{f(\boldsymbol{x}^i|\theta)\}) = 1_{y^i \neq f(\boldsymbol{x}^i|\theta)}$$

  - E.g. (regression)

$$L(\{y^i\}, \{f(\boldsymbol{x}^i|\theta)\}) = \left(y^i - f(\boldsymbol{x}^i|\theta)\right)^2$$

Validation sets

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

Training	Validation

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

Training	Validation	Test

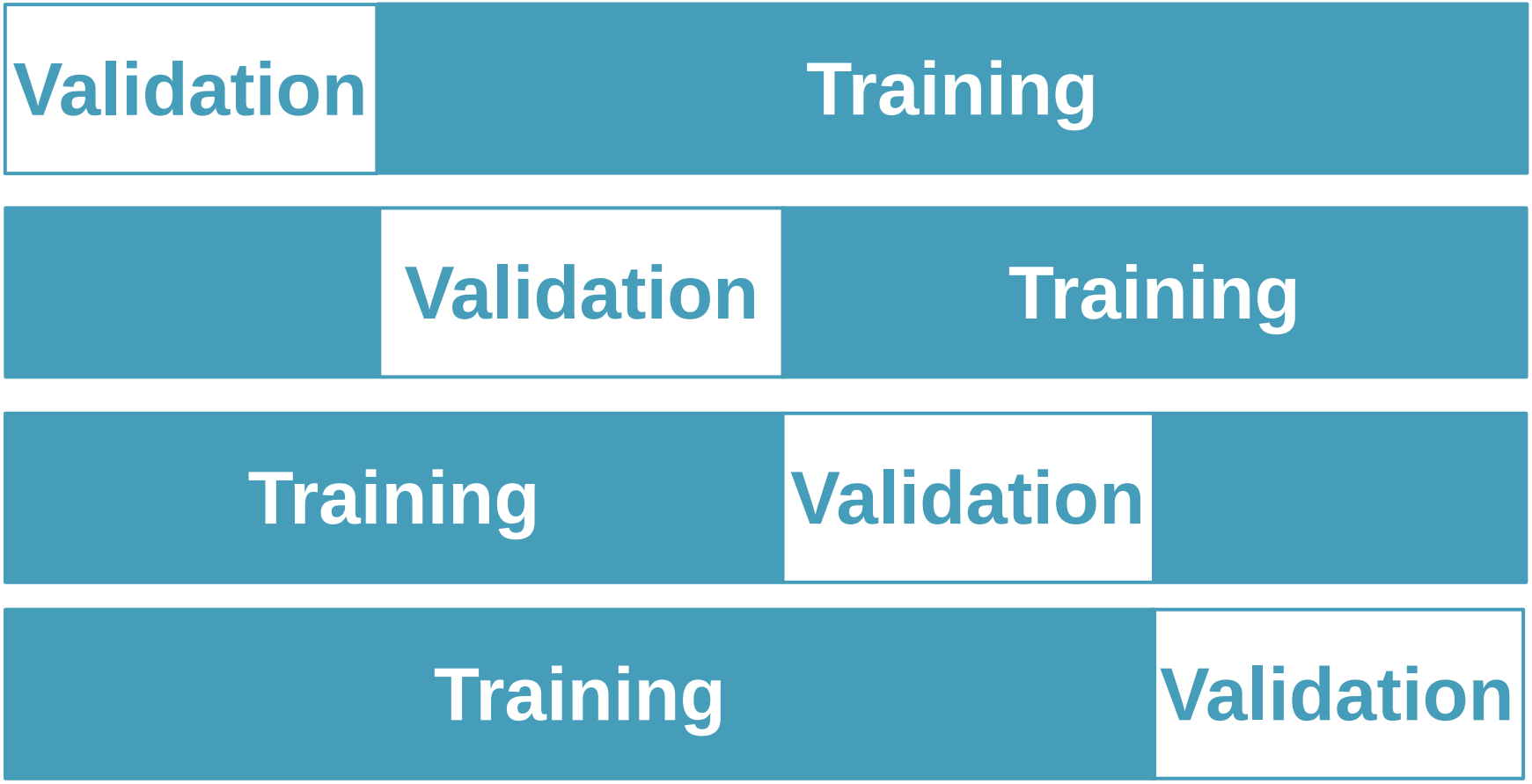
- **How much data** should go in each of the training, validation and test sets?
- How do we know we have **enough data** to evaluate the prediction and generalization errors?

- **Sample re-use**
  - cross-validation
  - bootstrap
- **Analytical tools**
  - Mallow's Cp, AIC, BIC
  - MDL
  - SRM.

Sample re-use

Cross-validation

- Cut the training set in k separate **folds.**
- For each fold, train on the (k-1) remaining folds.



Cross-validated performance

- Cross-validation estimate of the prediction error

$$CV(f) = \frac{1}{n}\sum_{i=1}^{n} L(y^i, f_{k(i)}(\boldsymbol{x}^i))$$

Computed with the k(i)-th part of the data removed. k(i) = fold in which i is.

- Estimates the **expected prediction error**

$$\mathrm{Err} = \mathbb{E}[L(Y, f(X))]$$

Y, X: (independent) test sample

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.**

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
- **What is the size of the training sets?**

Bootstrap

- **Randomly draw datasets with replacement** from the training data
- **Repeat B times** (typically, B=100) ⇒ B models
- **Leave-one-out boostrap 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 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.

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

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?**

- **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

- In this population:

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

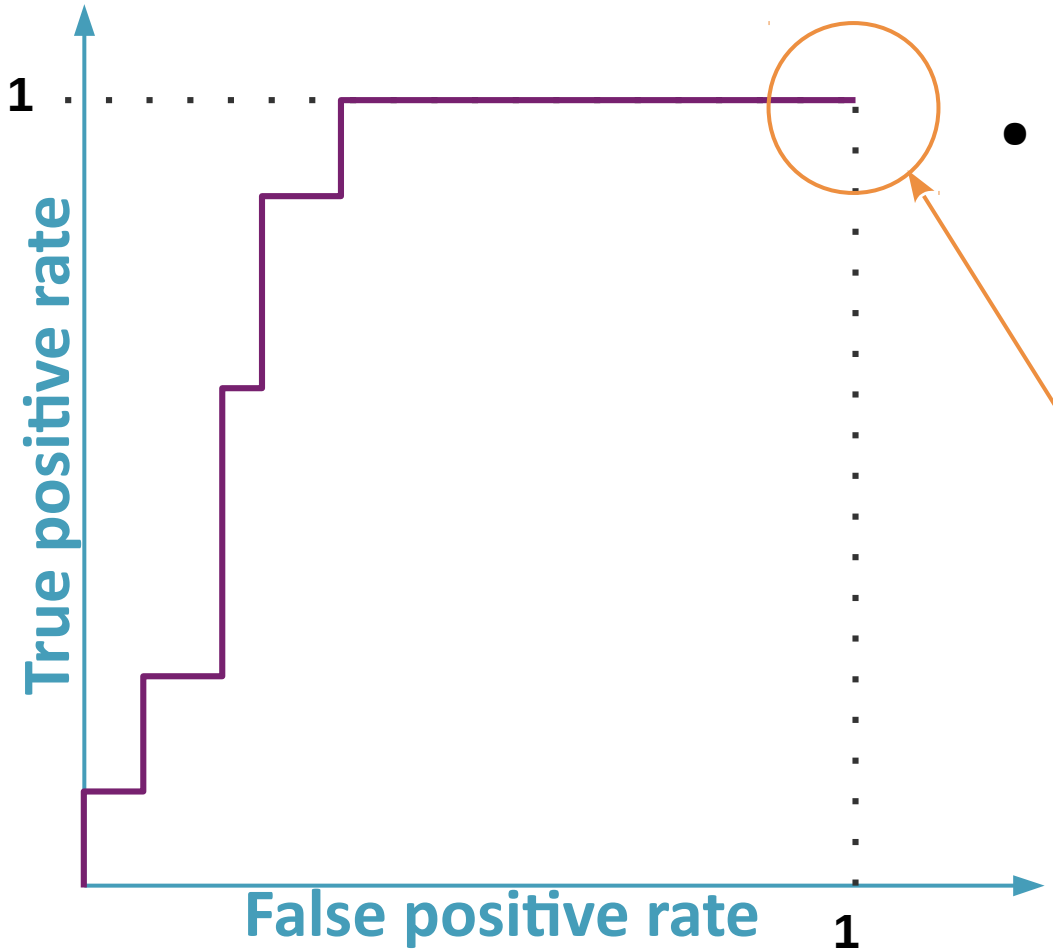
	Cancer	No cancer	Total
Positive test	190	210	400
Negative test	10	3590	3600
Total	200	3800	4000

- **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

ROC curves

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

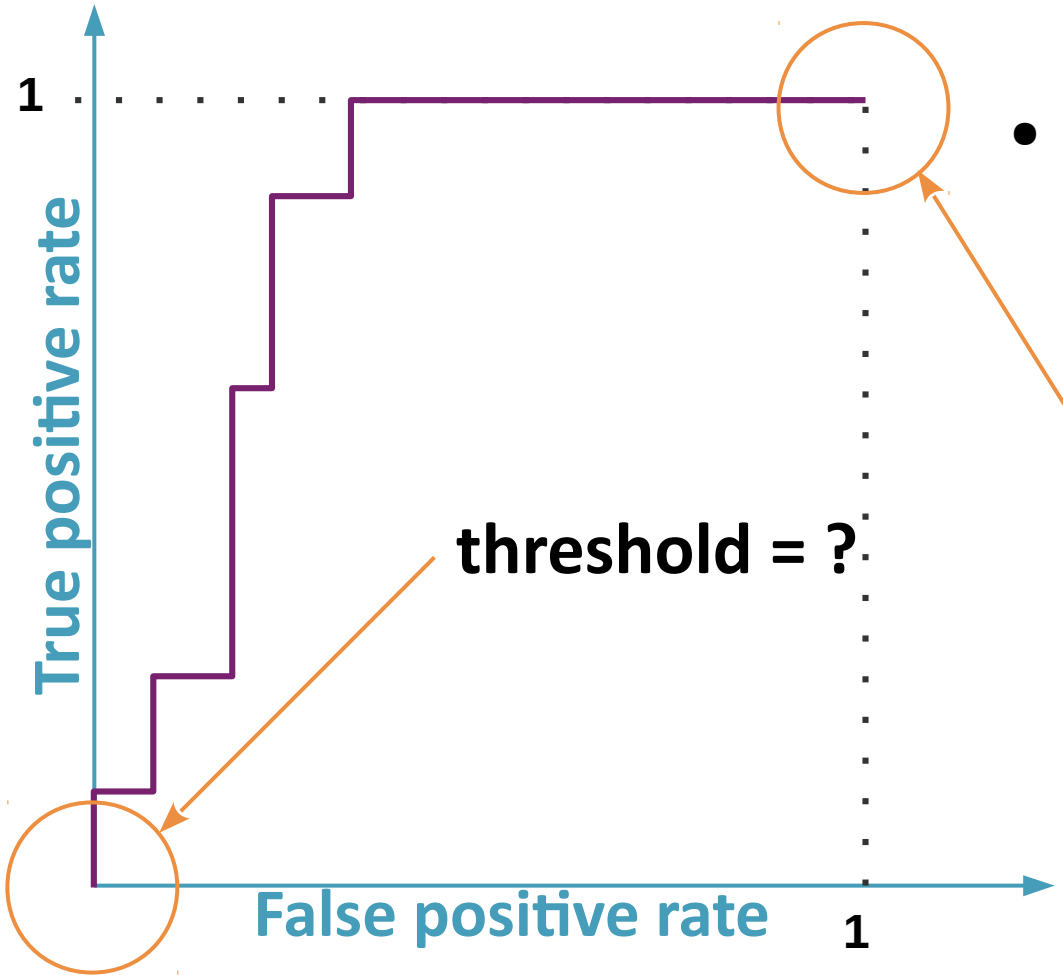


- Plot TPR vs FPR for all possible thresholds.

threshold = ?

ROC curves

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

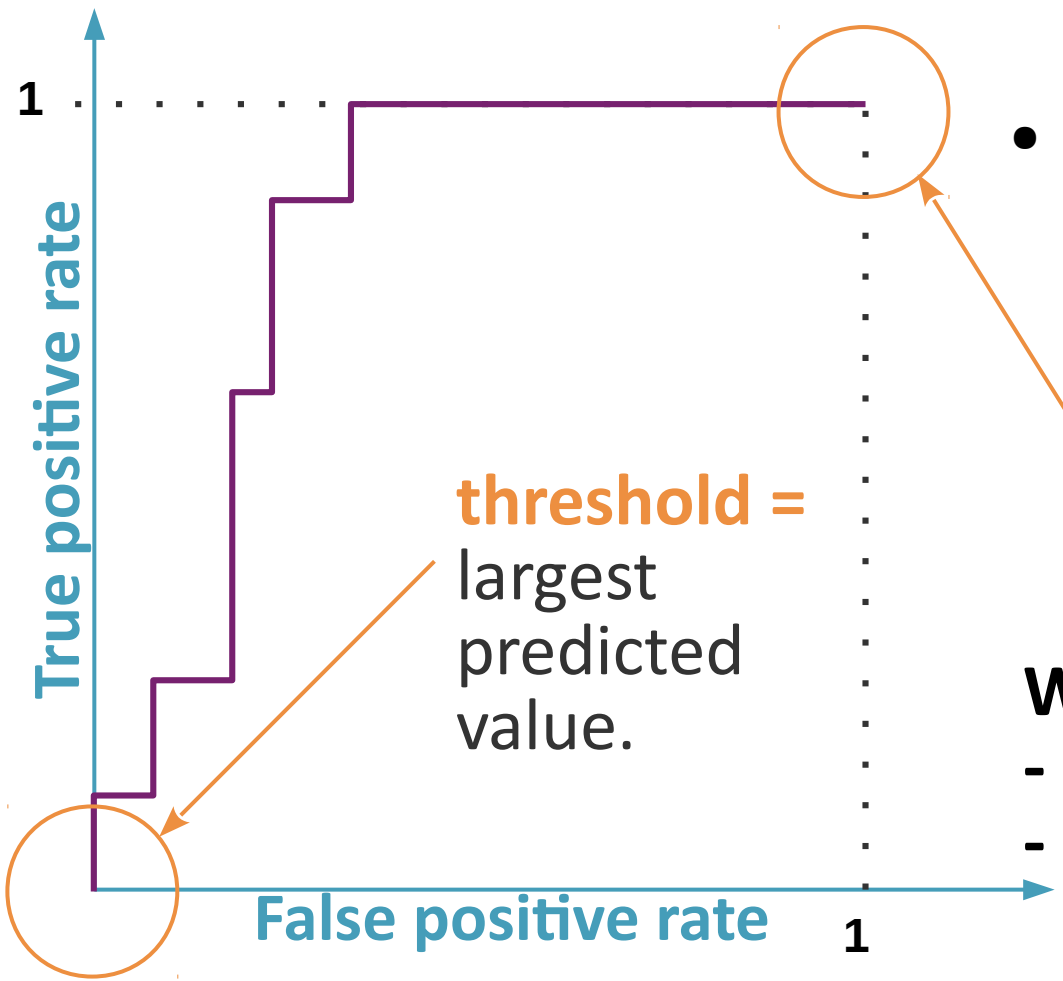


- Plot TPR vs FPR for all possible thresholds.

threshold = smallest predicted value.

ROC curves

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



- Plot TPR vs FPR for all possible thresholds.

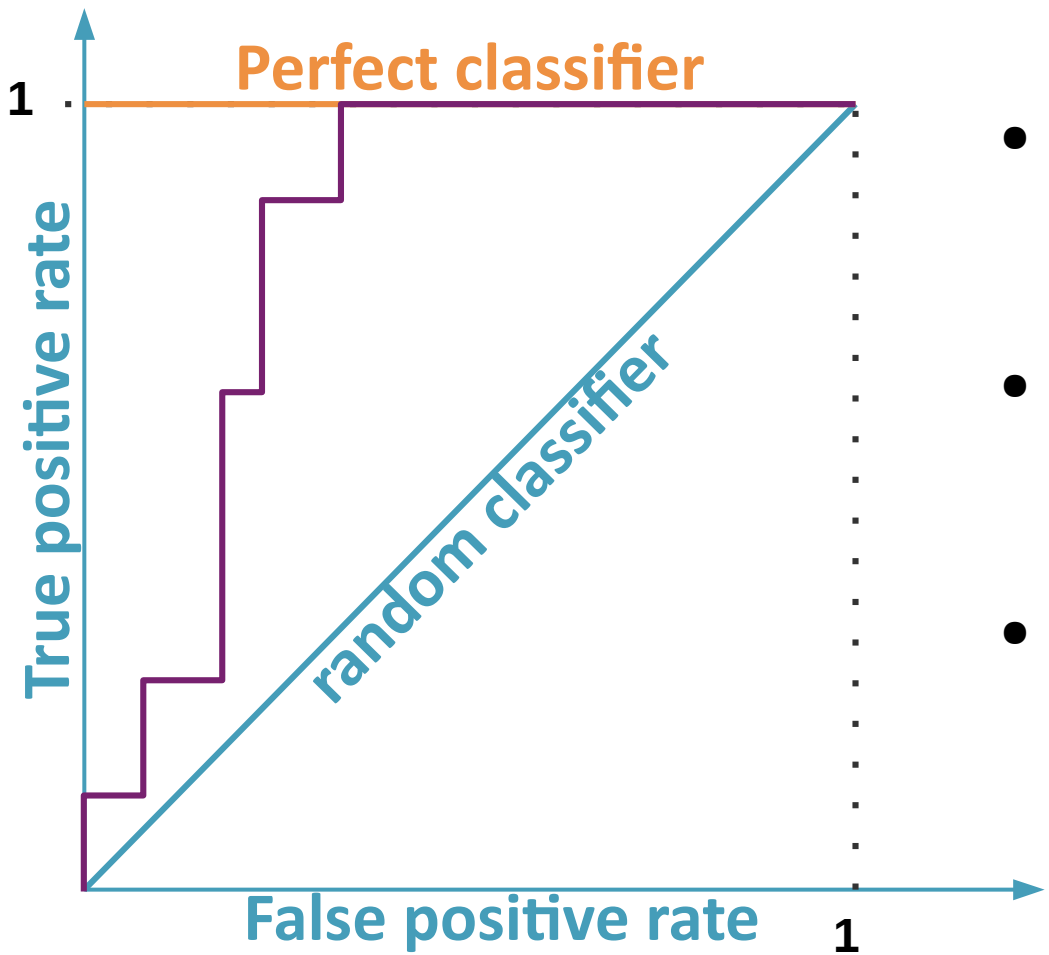
threshold = largest predicted value.

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:**

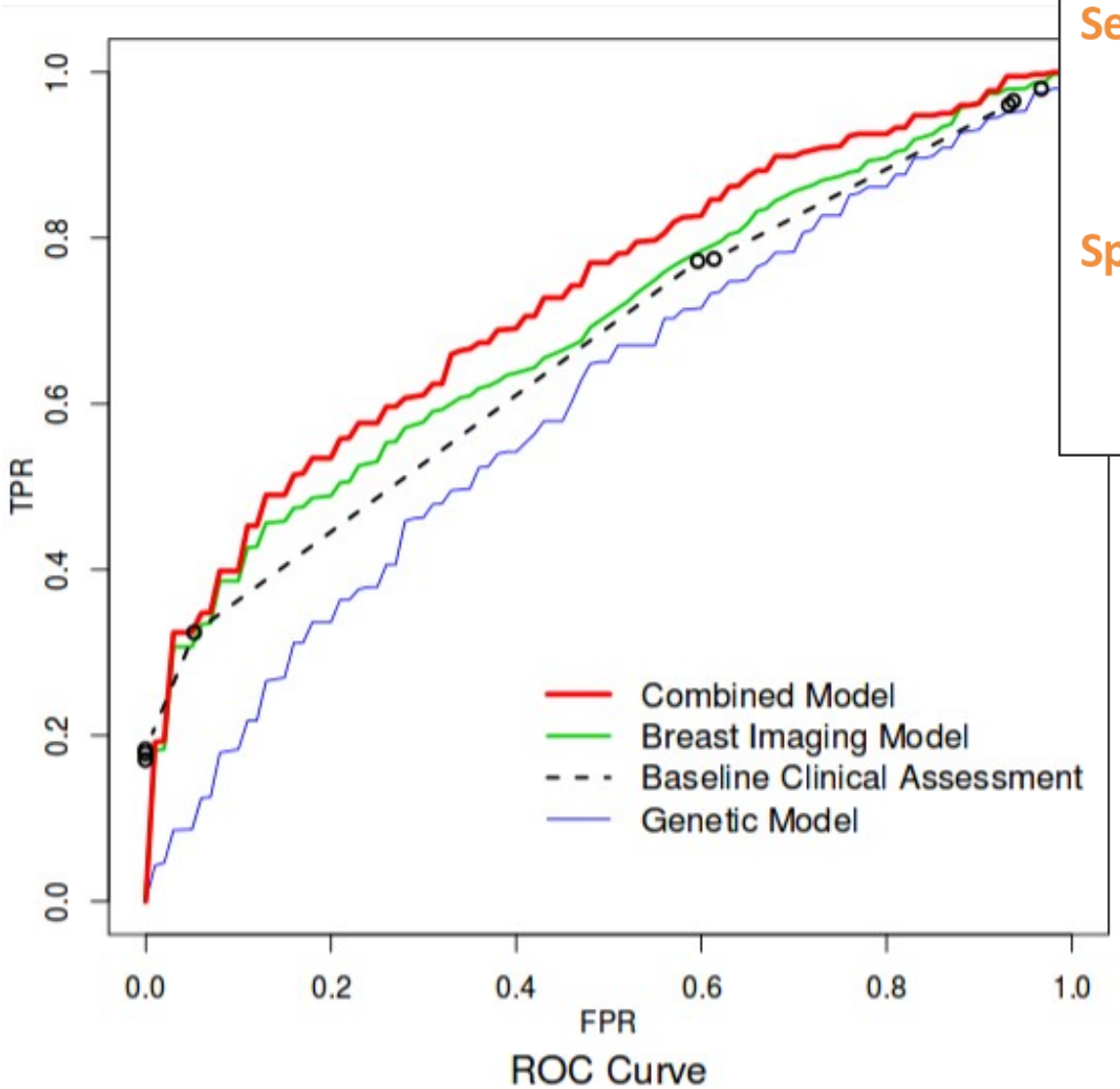
  AUROC = 1.0
- **Random classifier:**

  AUROC = 0.5
- **Our classifier:**

  0.5 < AUROC < 1.0

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)

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

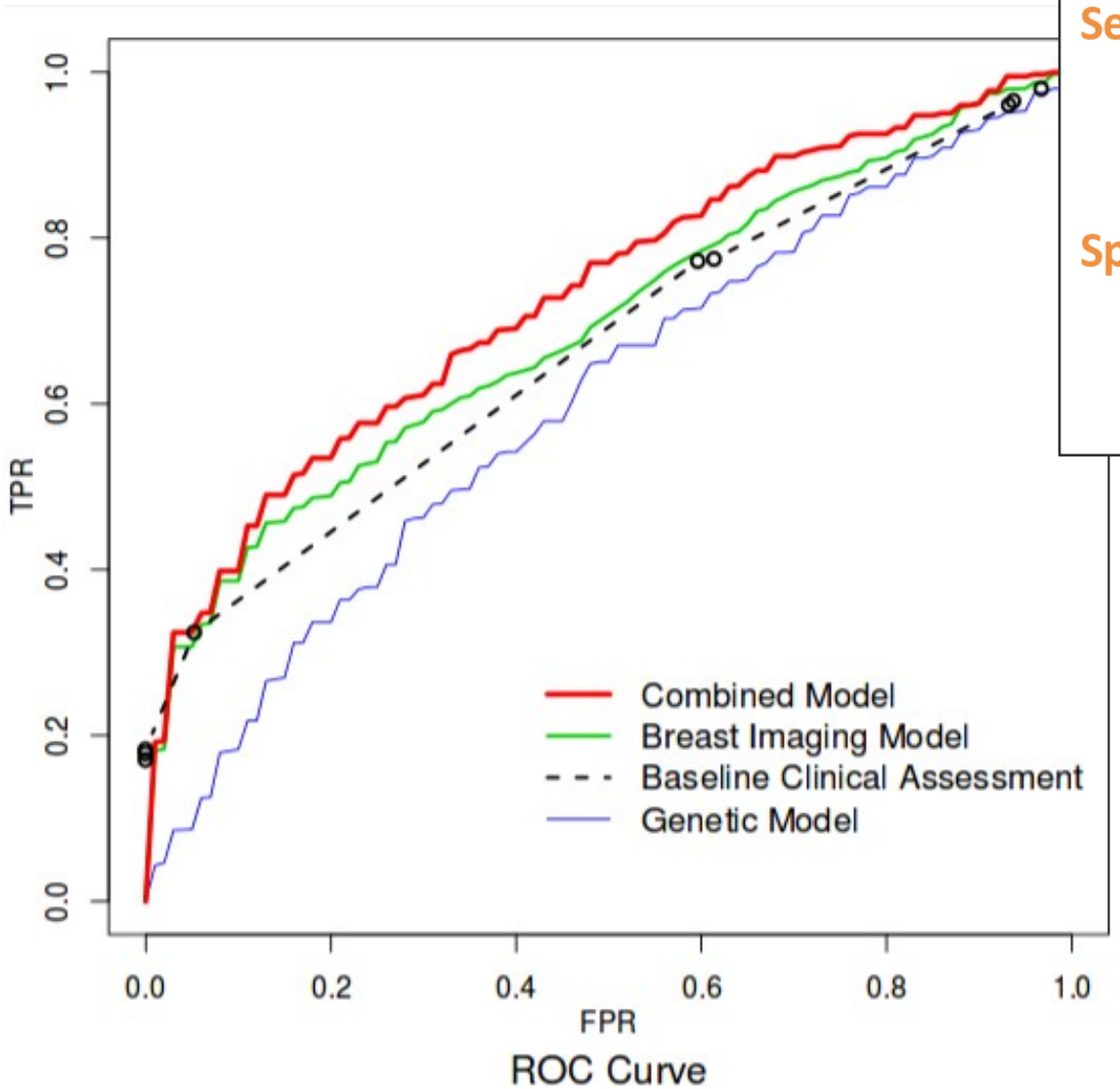
Specificity = True negative rate (TNR)= 1 - FPR

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

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

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)

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

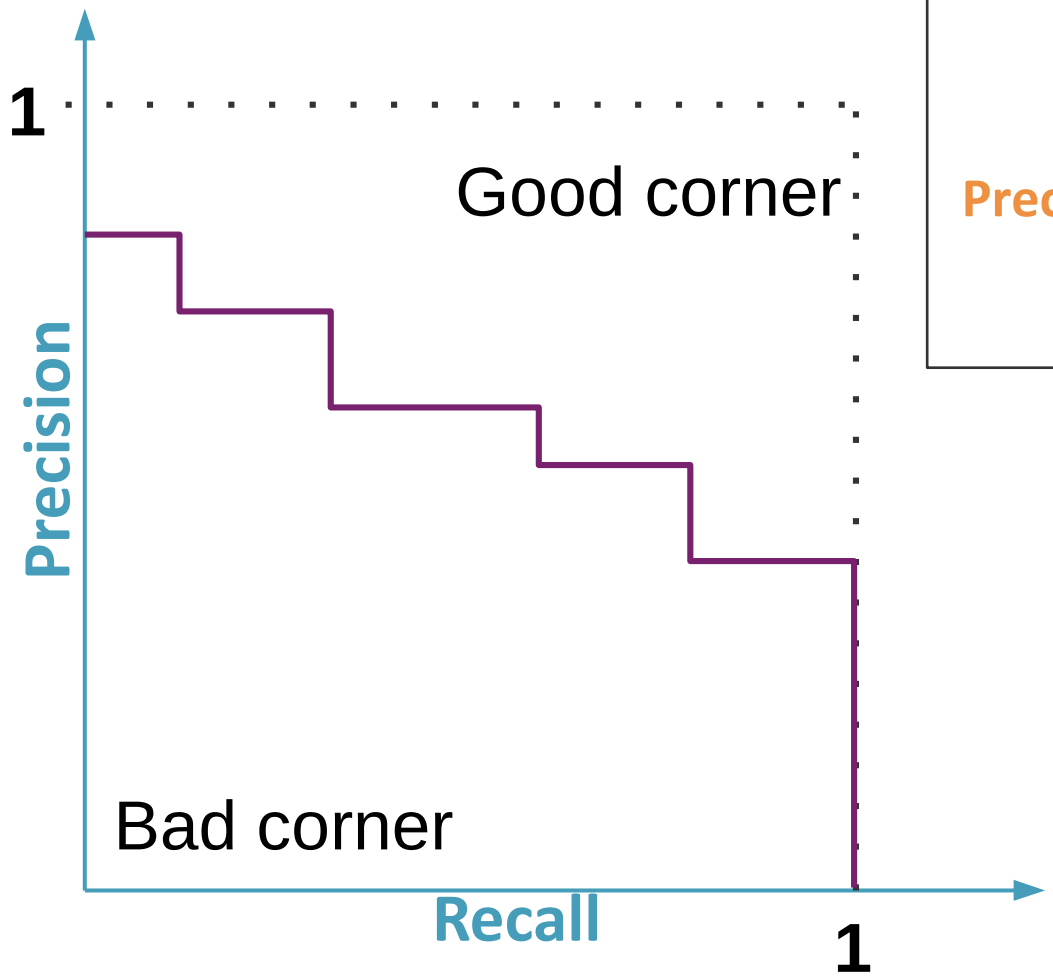
Specificity = True negative rate (TNR)= 1 - FPR

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

High recall = fewer chances to miss a case

High specificity / low FPR = fewer false alarms

Precision-Recall curves



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

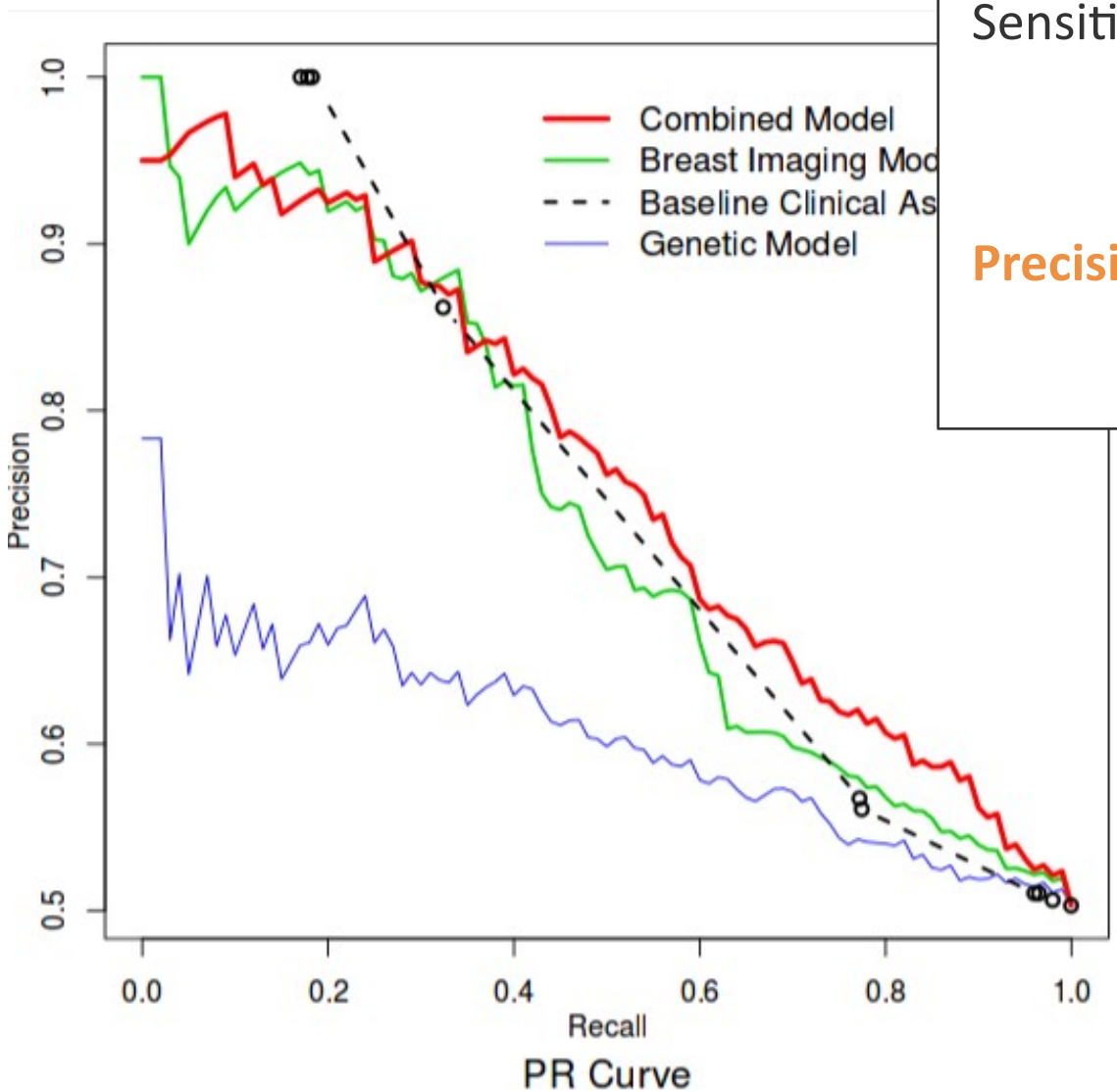
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision = Positive predictive value (PPV)

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

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)

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

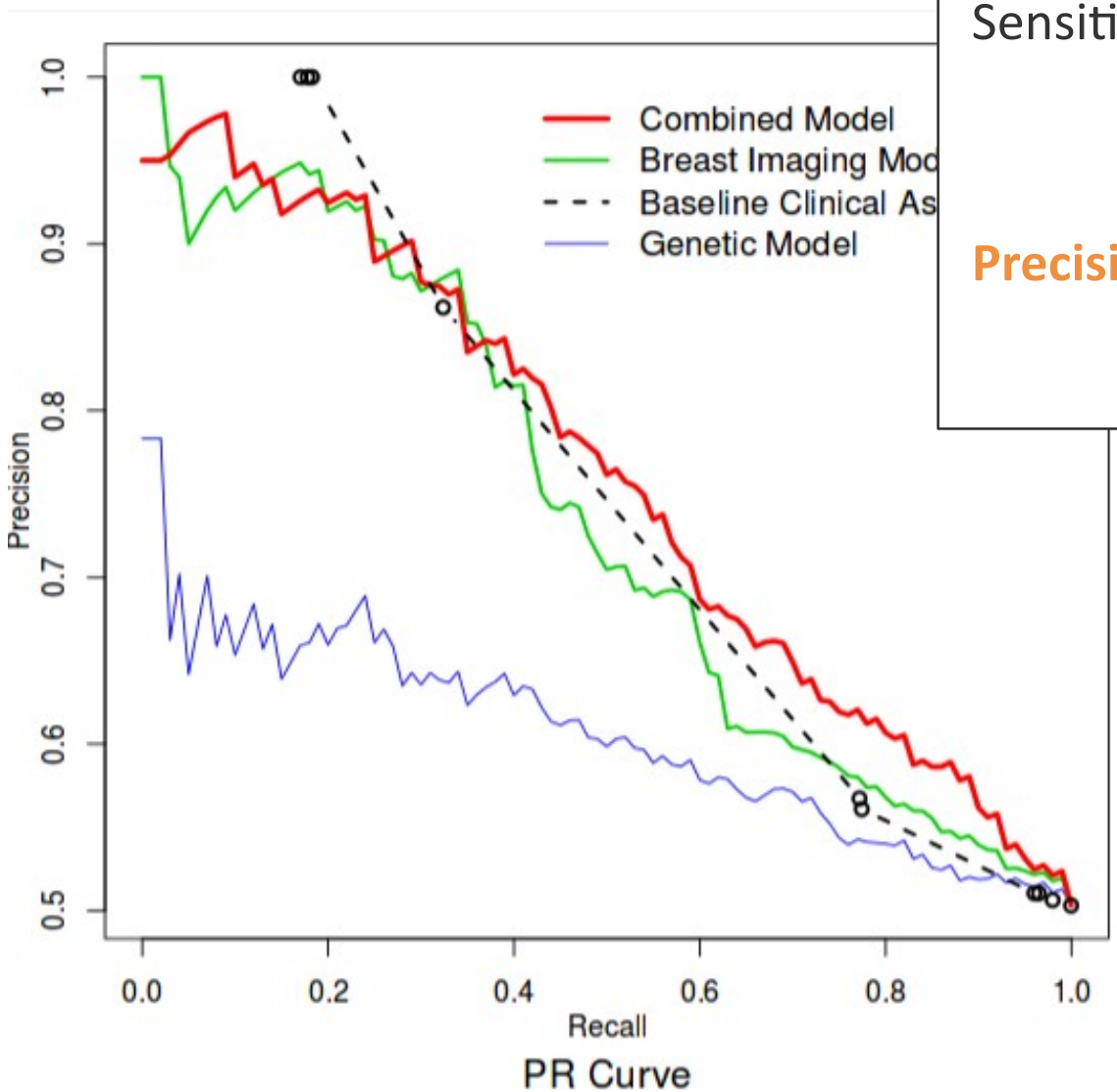
Precision = Positive predictive value (PPV)

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

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

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

- **Root-mean squared error**

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

- **Relative squared error**

$$E_{RSE} = \frac{\sum_{i=1}^{n}(y^i - f(\boldsymbol{x}^i|\theta))^2}{\sum_{i=1}^{n}(y^i - \bar{y})^2}$$

- **Coefficient of determination** $R^2 = 1 - E_{RSE}$

- **Residual sum of squares** $RSS = \sum_{i=1}^{n}(y^i - f(\boldsymbol{x}^i|\theta))^2$

Analytical tools and model complexity

Penalizing model complexity

augmented error:

$$E' = \text{empirical error} + \boldsymbol{\lambda}\ \text{model complexity}$$

- **If λ is small,** models that fit the training data well are encouraged (risk of introducing variance).
- **If λ is large,** simpler models are encouraged (risk of introducing bias).
- λ can be set by cross-validation
- in some cases (cf Chap. 6), it is possible to estimate E' for all values of λ

Cp, AIC and BIC

augmented error:

E' = empirical error + **optimism term**

The optimism term estimates the discrepancy between training and test error without any need for cross-validation:

– **Mallow's Cp**

(Linear regression

+ squared error)

$$Cp = \boxed{\frac{1}{n}(RSS} + 2d\hat{\sigma}^2)$$

empirical error

 # parameters used

estimate of the error variance

– **Akaike's Information Criterion (AIC)**

– **Bayesian Information Criterion (BIC)**

Minimum description length (MDL)

- **Shortest code to transmit a random variable z**
  - $-\log P(z)$ [Shannon's information theory]
- Assume receiver knows inputs X, model f.

  To transmit outputs Y, need

  $$-\log P(y \mid \theta, f, X) - \log P(\theta \mid f)$$

  average code length to transmit θ.

  average code length to transmit the difference between model prediction and true outputs.

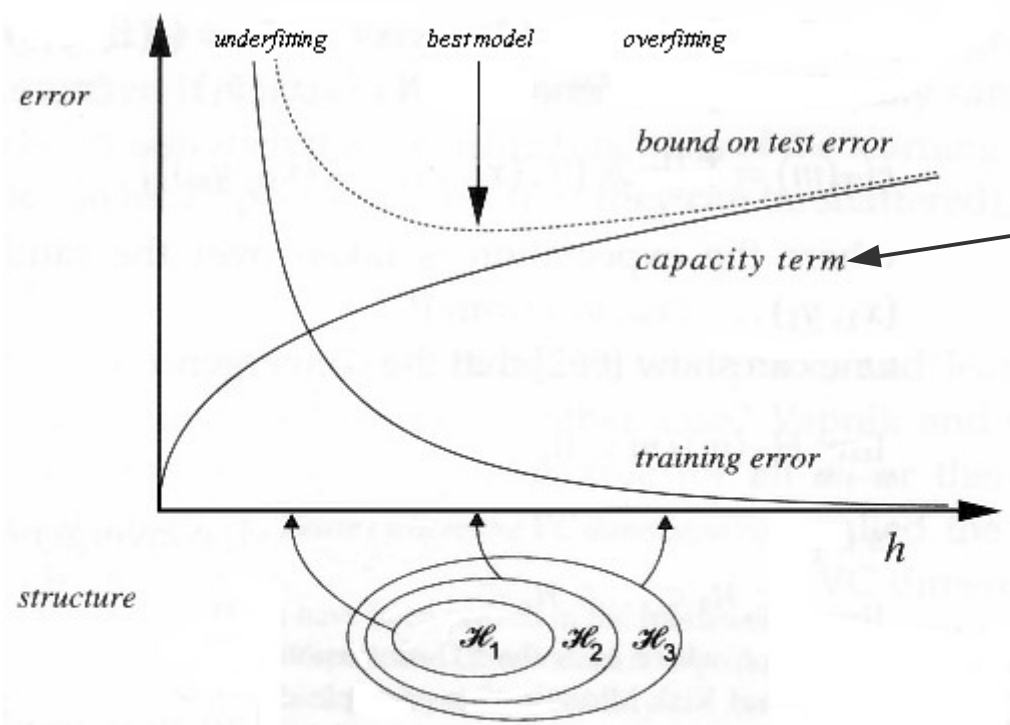
- Choose model with **smallest length.**

Structural risk minimization (SRM)

- Fit a nested sequence of **models of increasing VC dimensions** h1 < h2 < …
- Pick the one with **lower bound on test error**

E.g. Regression: with probability at least (1 – η),

$$\mathrm{Err} \leq \frac{\mathrm{RSS}_{\mathrm{train}}}{n(1-\sqrt{\epsilon})_+}$$



$$\epsilon = \frac{1}{n}\left(h(\log(n/h)+1) - \log(\eta/4)\right)$$

VC-dimension

What happens when n gets larger?

Summary: model selection techniques

- **Cross-validation:** estimate generalization accuracy empirically
- **Regularization:** Penalize complex models

  E' = empirical error + λ model complexity

  Mallow's Cp, Akaike's / Bayesian Information Criteria
- **Minimum description length (MDL)**

  Kolmogorov complexity = shortest description of data

  [Information theory]
- **Structural risk minimization (SRM)**

  Order models by complexity

  polynomes of ↗ degree; ↗ values of λ
- **Bayesian model selection**

ML Toolboxes

- **Python: scikit-learn**

  `http://scikit-learn.org`

- **R: Machine Learning Task View**

  `http://cran.r-project.org/web/views/MachineLearning.html`

- **Matlab™: Machine Learning with MATLAB**

  `http://fr.mathworks.com/machine-learning/index.html`

  - Statistics and Machine Learning Toolbox
  - Neural Network Toolbox

Getting started with Python

I highly recommend `http://scipy-lectures.github.io/`

This document

Tutorial material on the scientific Python ecosystem, a quick introduction to central tools and techniques. The different chapters each correspond to a 1 to 2 hours course with increasing level of expertise, from beginner to expert.

Authors What's new Scipy-Lecture-Notes License

Download

- PDF, 2 pages per side
- PDF, 1 page per side
- HTML and example files
- Source code (github)

1. Getting started with Python for science

▶ 1.1. Scientific computing with tools and workflow

▶ 1.2. The Python language

▶ 1.3. NumPy: creating and manipulating numerical data

▶ 1.4. Matplotlib: plotting

▶ 1.5. Scipy : high-level scientific computing

- 1.6. Getting help and finding documentation

2. Advanced topics

▶ 2.1. Advanced Python Constructs

▶ 2.2. Advanced Numpy

▶ 2.3. Debugging code

▶ 2.4. Optimizing code

▶ 2.5. Sparse Matrices in SciPy

▶ 2.6. Image manipulation and processing using Numpy and Scipy

▶ 2.7. Mathematical optimization: finding minima of functions

References

- **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

- More on **entropy encoding:**
  http://lesswrong.com/lw/o1/entropy_and_short_codes/

- **Textbook:**

  The Elements of Statistical Learning
  Hastie, Tibshirani, Friedman (2009)
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html

Practical matters

- Make sure you have handed in **HW1**
- **HW2** is online, due Sep. 21

- **Lab**

  `https://github.com/chagaz/ma2823_2016`