האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

Artificial Intelligence in Medicine

Learning (2)

Nir Friedman and Tommy Kaplan 22/1/24

“I hate to generalize, but…”

Gloria Steinem

Generalization in medicine

- Family medicine clinic, Jerusalem, yesterday:
- Physician examines 30 patients
- First 28 with cough / runny nose / headaches

  Upper respiratory viral infection

- A young pregnant woman with runny nose:

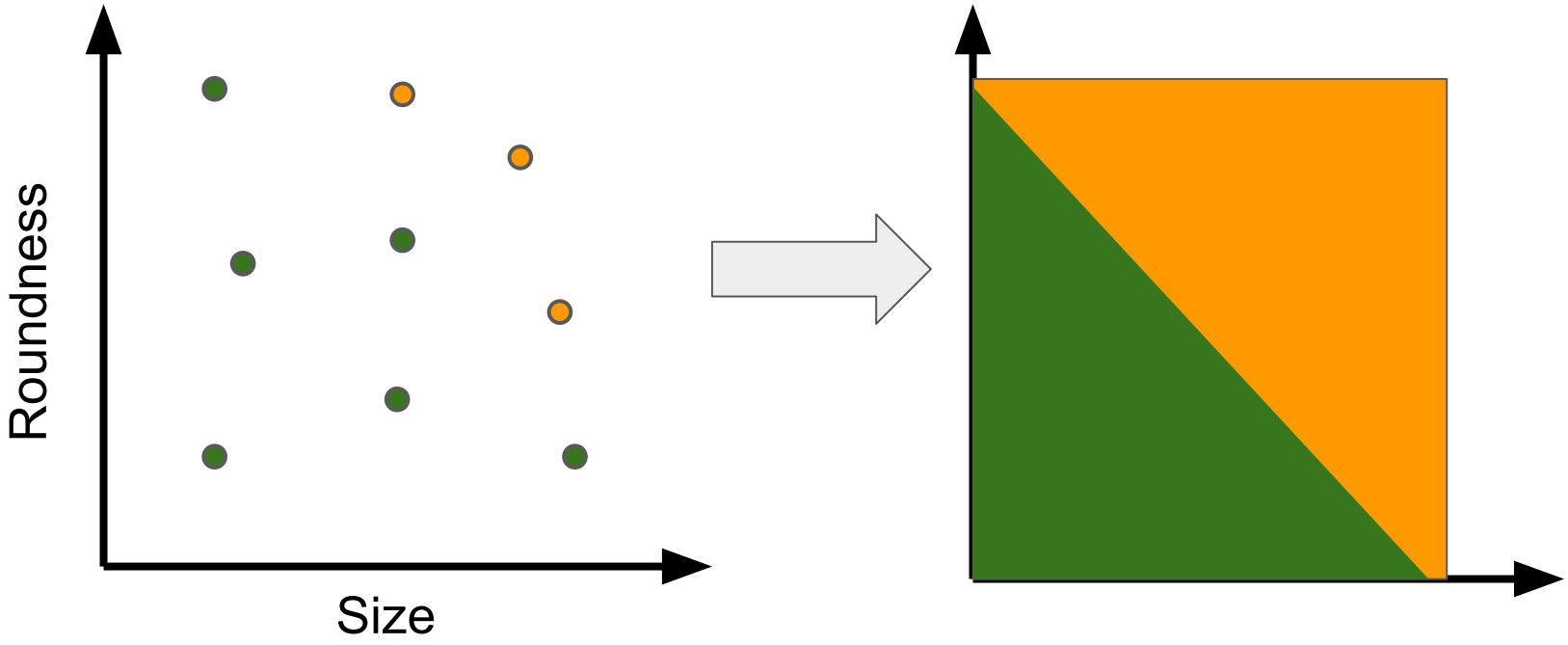
  Allergy? Virus?

- An elderly man, presenting stomach ache

  Colon cancer? Virus?

What is learning?

Examples → Classifier → Test



Generalization

Performance on heretofore unseen cases.

Why should it work?

Generalization

"Though this be madness, yet there is method in't"

Hamlet

Assumptions about the nature of examples:

- Samples are from the same "**population**"
- Actual concept has some regularity
  - Smoothness
  - Simplicity
  - …

Concepts

Train error - error on the training set (seen)

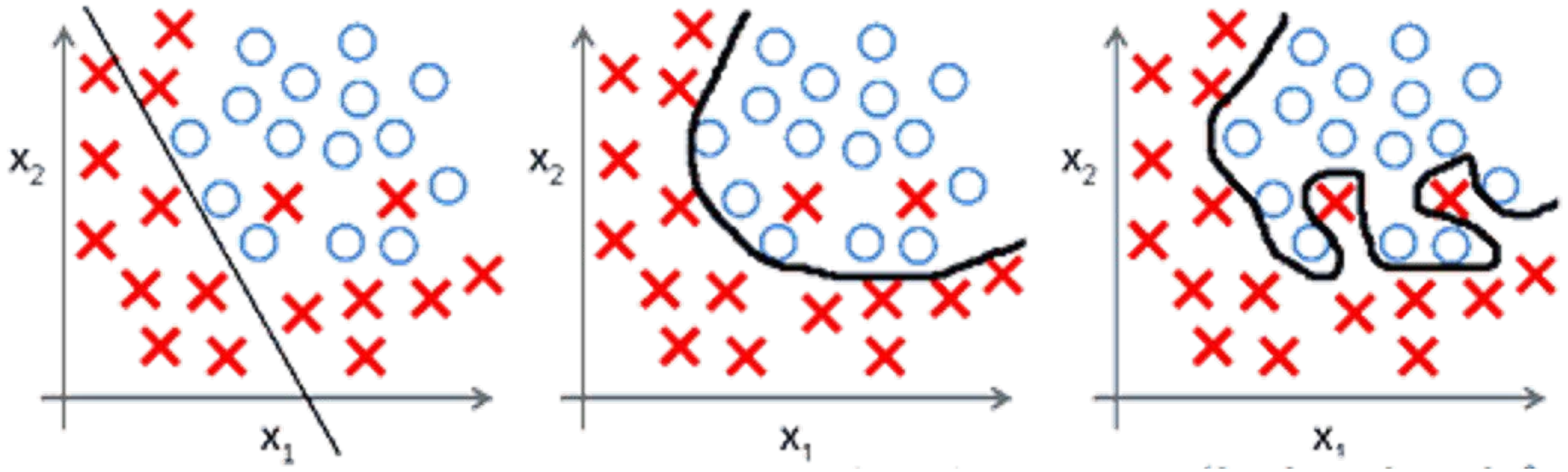
Test error - error on test set (yet unseen)

Are these related?

Implicit assumption:

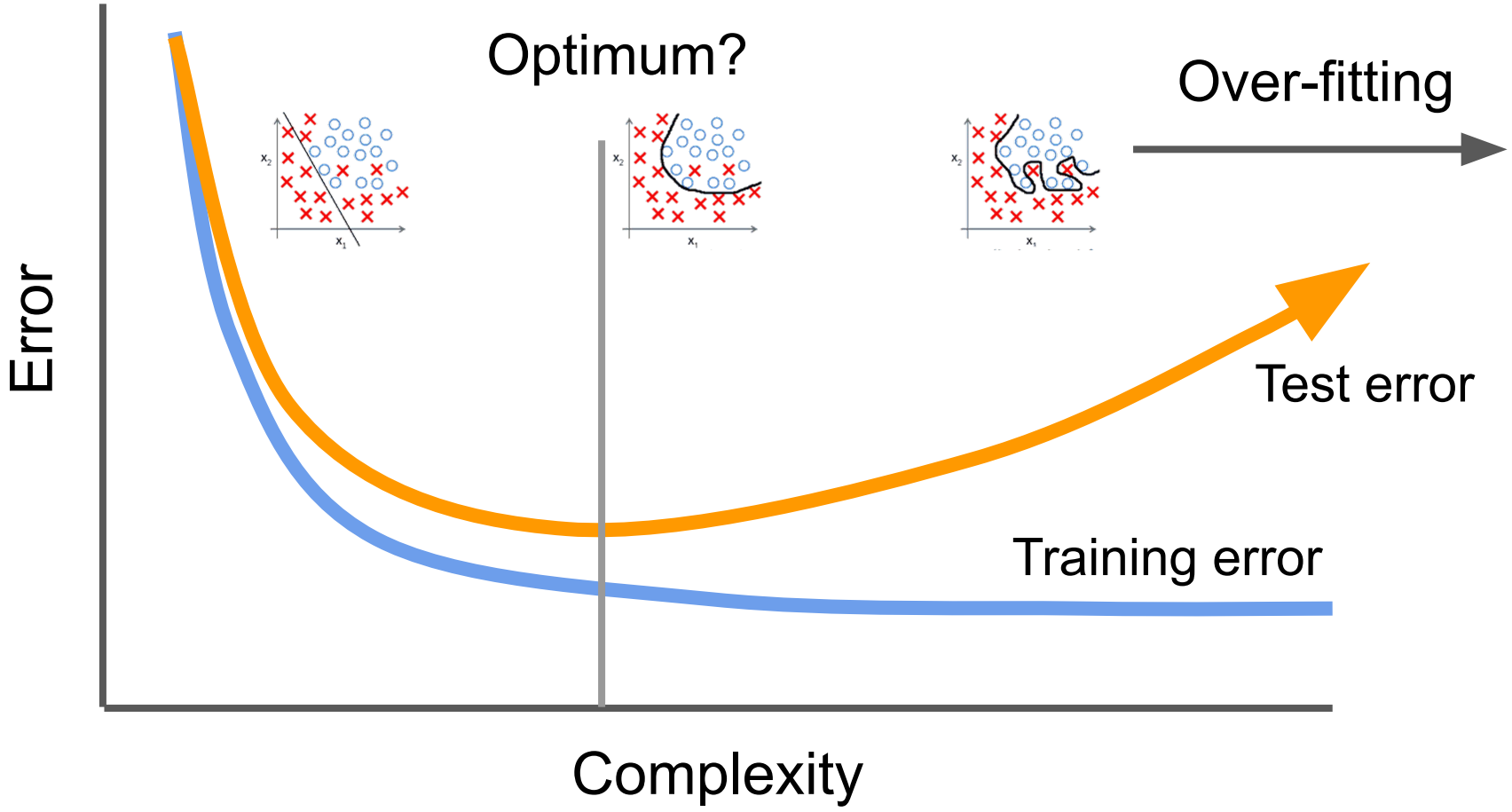
- Reducing training error will reduce also test error
- Is this reasonable?

Different classifiers



Complexity

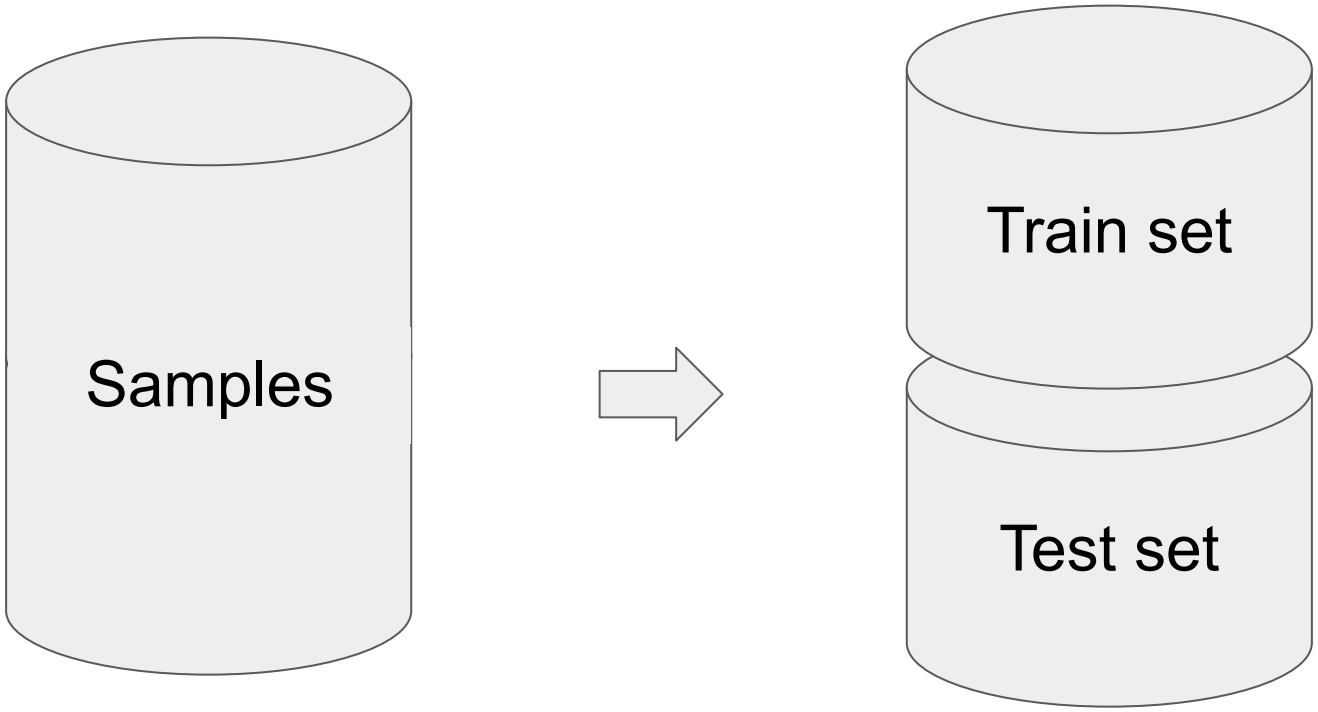
Train vs. test errors



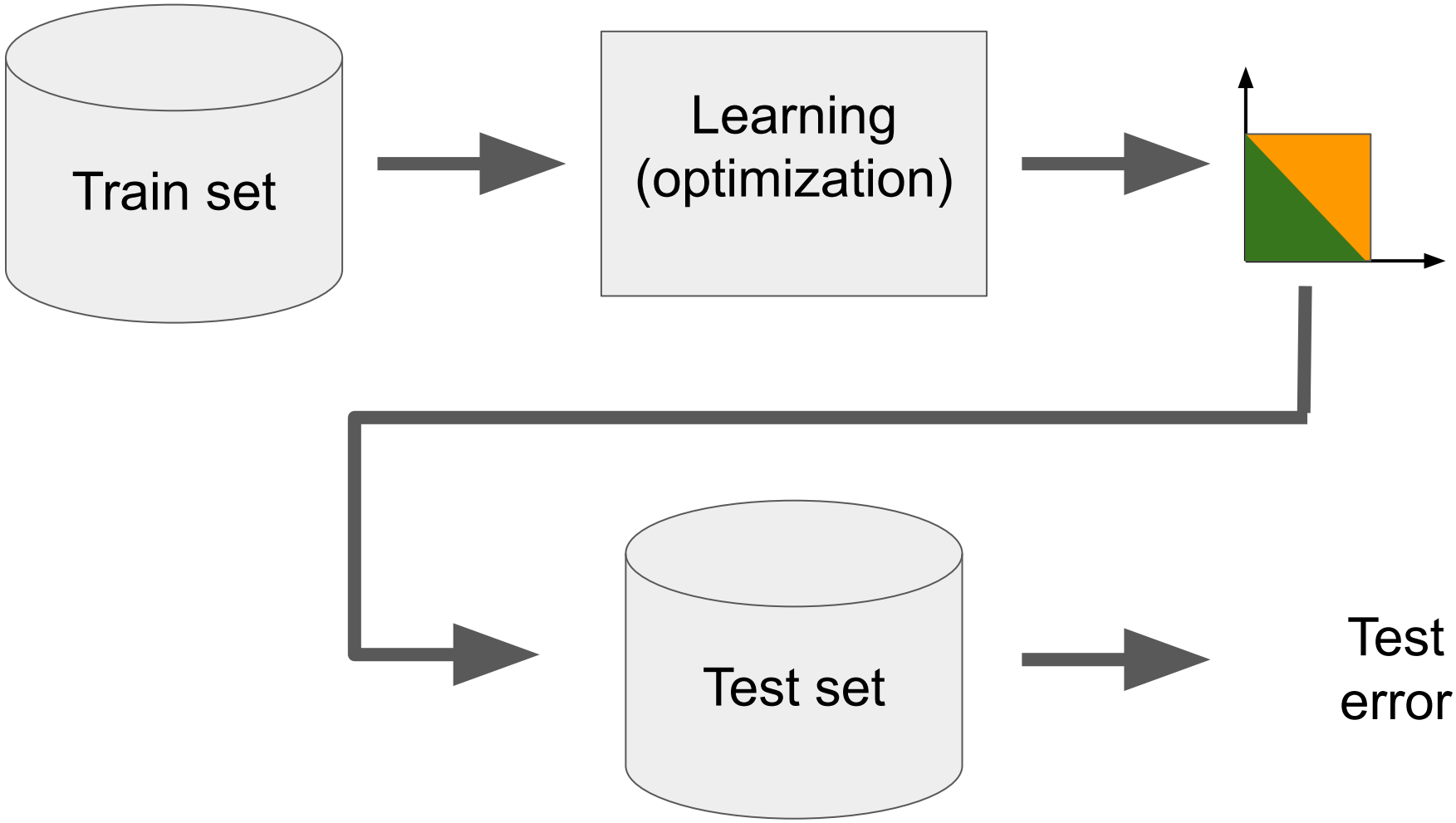
Measures of complexity

- Degree of polynomial
- Number of questions in decision tree
- Number of free parameters
- Magnitude of parameters
- Curvature of decision surface
- Neighborhood size in K-nearest neighbors
- …

Empirical Test Error



Empirical Test Error

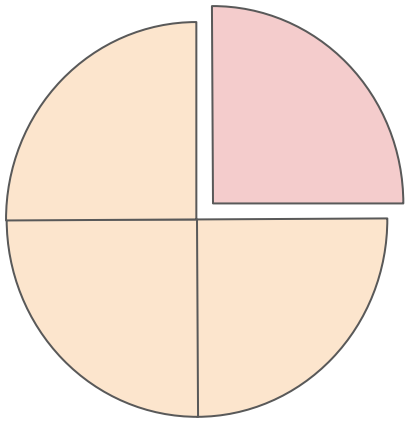


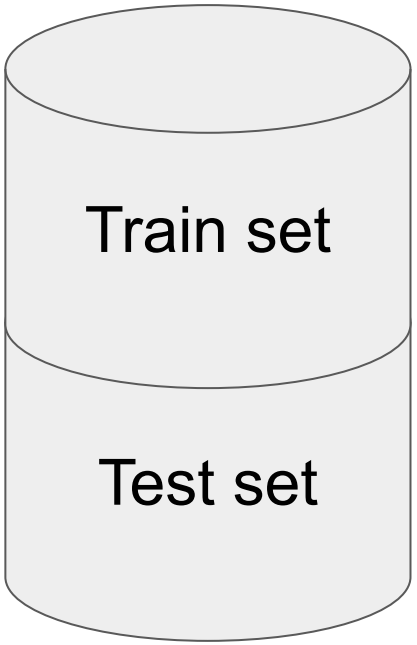
Empirical Test Error

Issues:

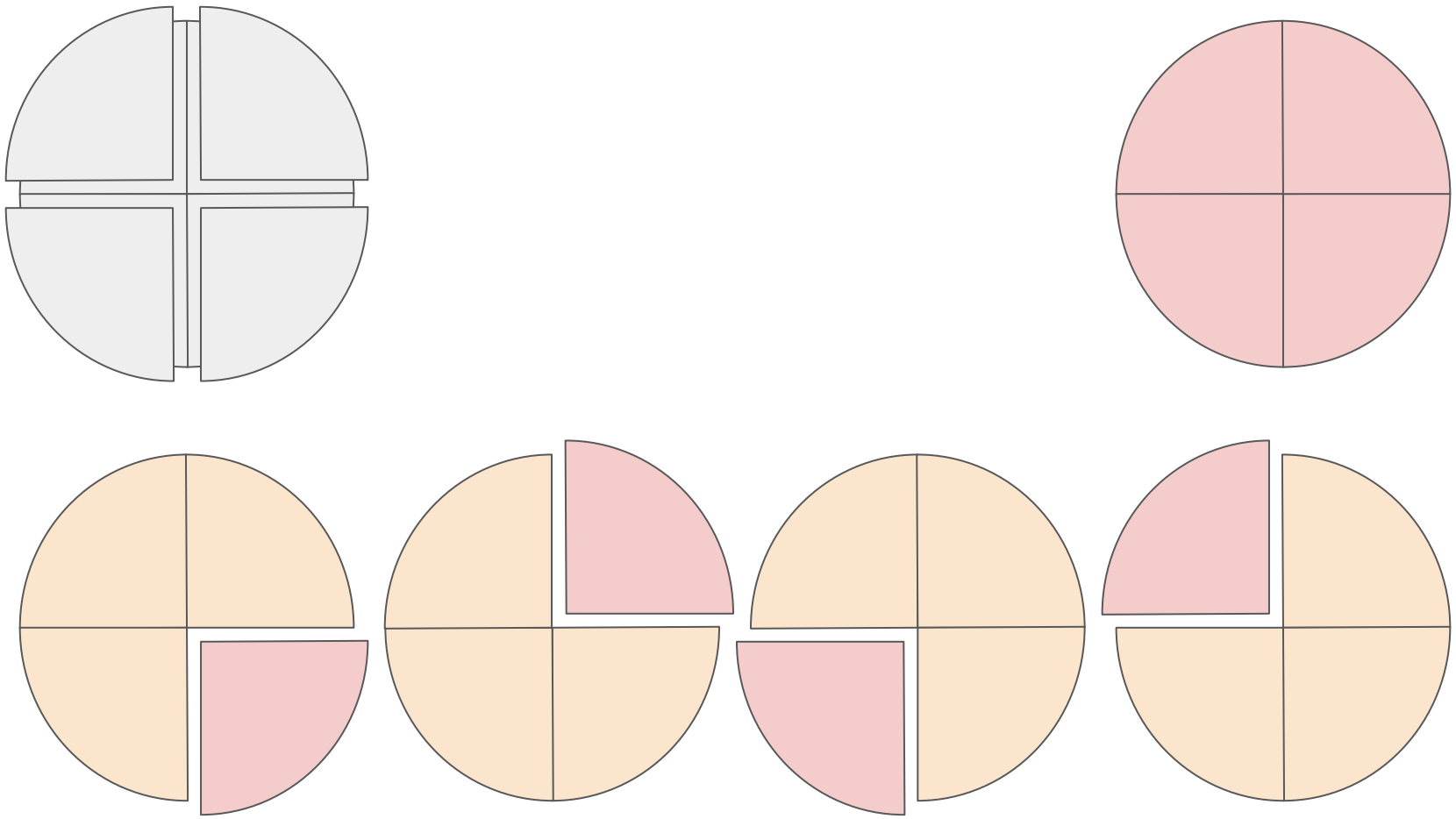
Train/Test allocation

- Small train set → not enough for learning
- Small test set → noisy error estimation



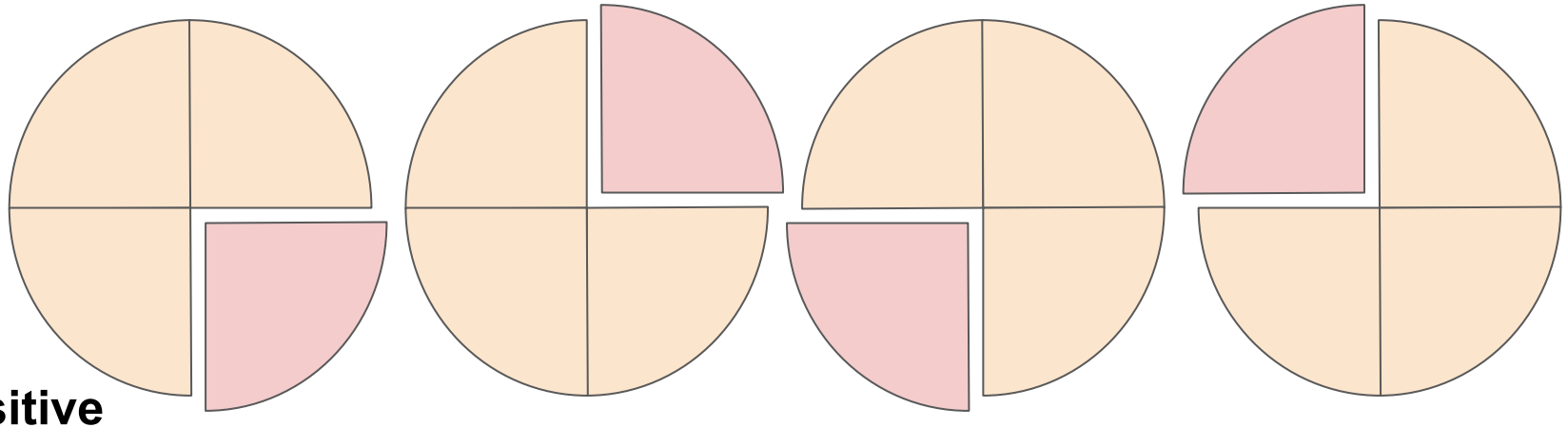


(4-fold) cross validation

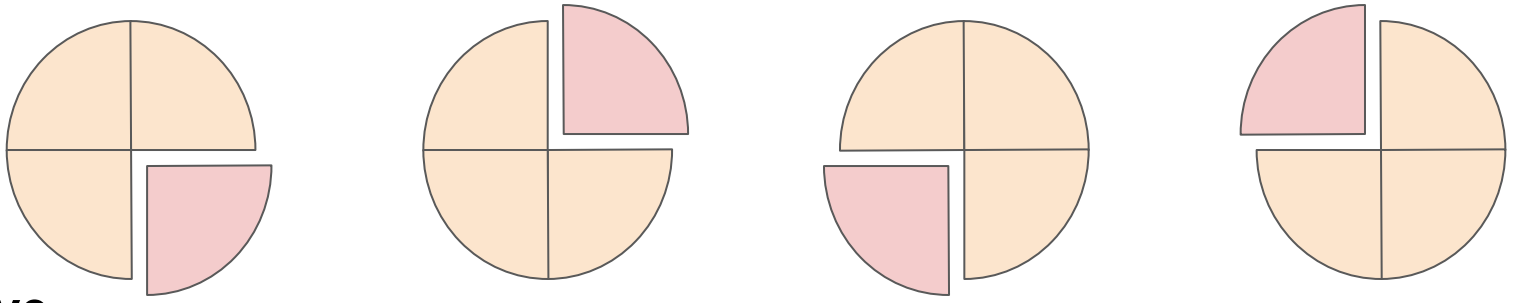


Train
Test

Stratified cross validation



Positive samples



Negative samples

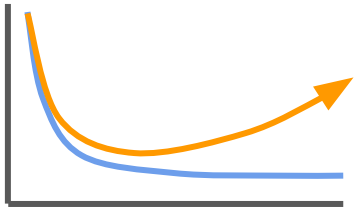
Empirical Test Error

Issues:

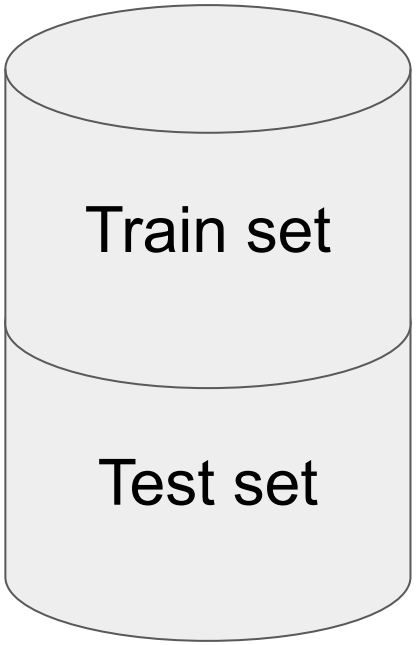
Train/Test allocation

- Small train set → not enough for learning
- Small test set → noisy error estimation

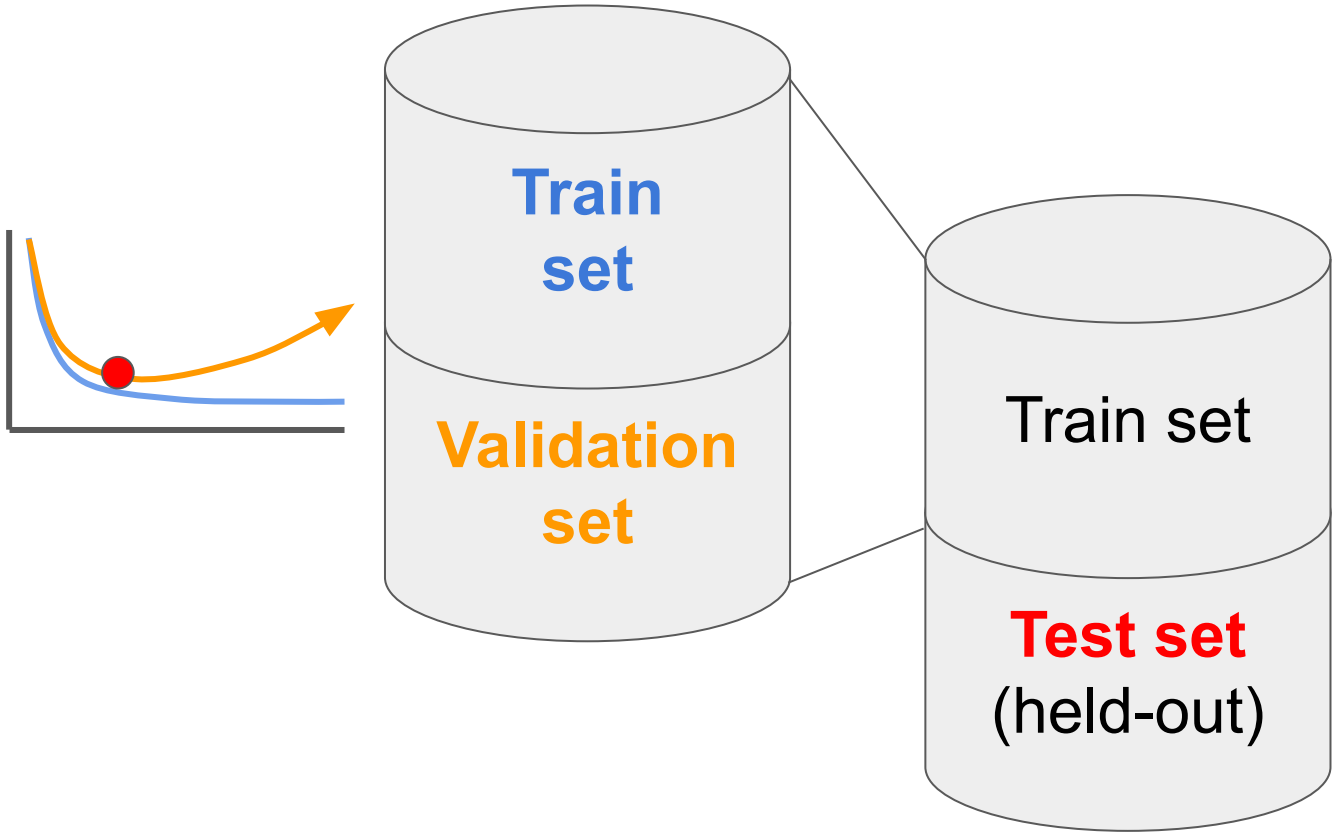
Use of Test set



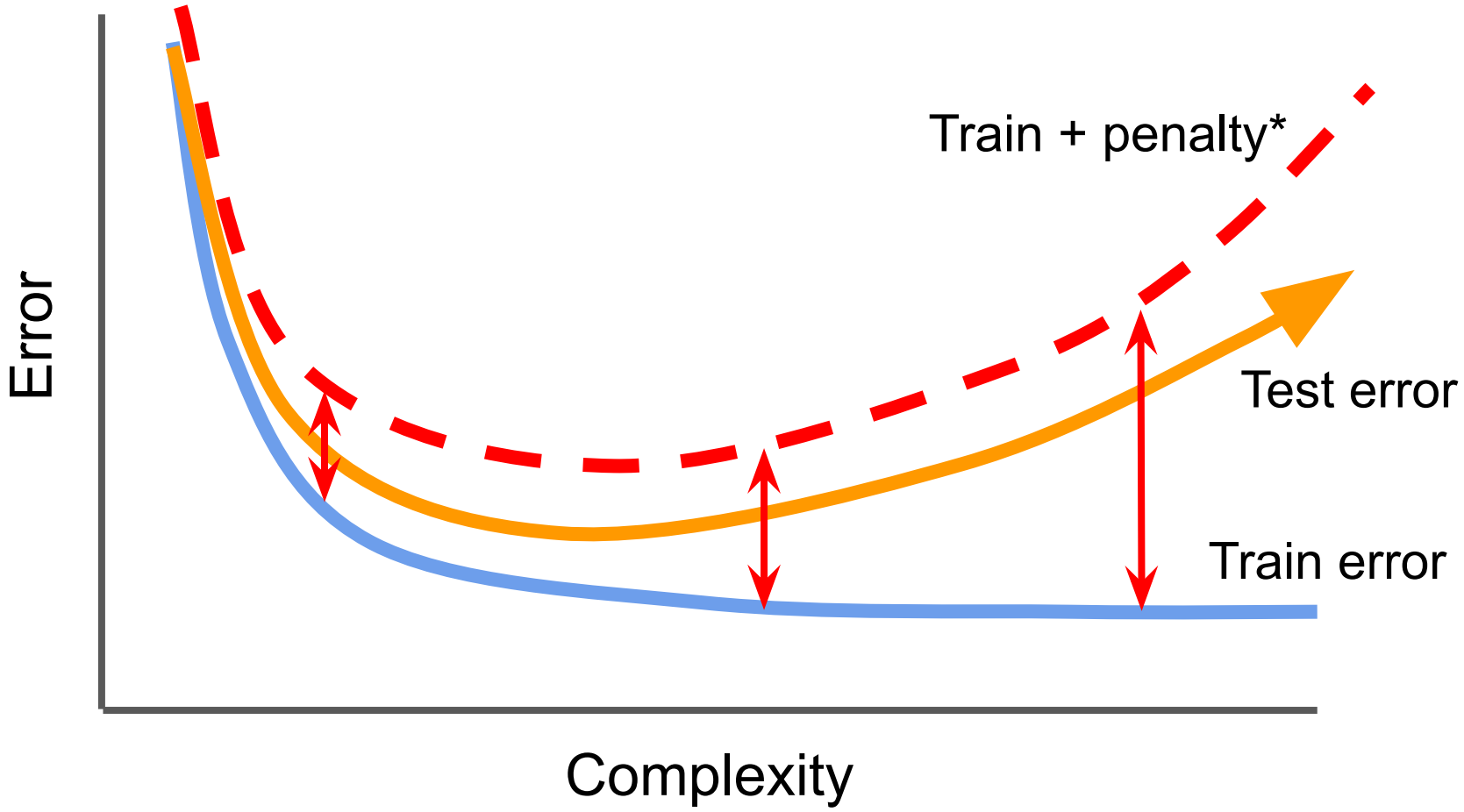
- Multiple evaluations (of classifiers)
- Test samples used for learning
- **Over-fitting** the test data!!



Empirical Test Error



Theoretical bound of test error



Penalty* = function of model complexity

Summary so far

Choose complexity (hyper-parameters)

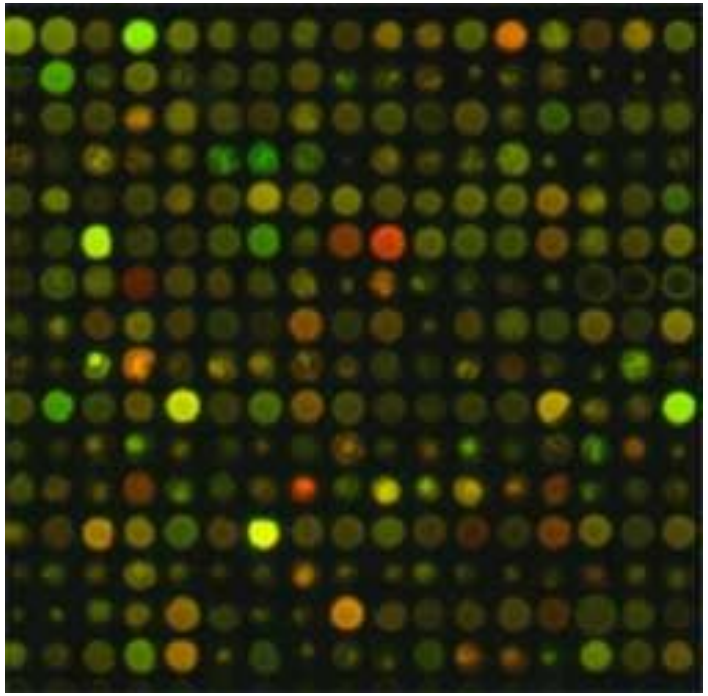
Estimate generalization performance

- Train/test split
- Cross-validation
- Theoretical limits

Learn model with chosen hyper-parameter value

Case study

June 2000, the human genome sequenced



Microarrays measure mRNA levels of 23K genes

Gene expression classifications and predictions of human cancers

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

Departments of [b]Genetics and [l]Surgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; [d]Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of [e]Health Research and Policy and Statistics, [c]Genetics, [i]Pathology, [j]Surgery, and [m]Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of [g]Medicine (Section of Oncology), [f]Surgery, and [k]Biochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and [h]Life Sciences Division, Lawrence Orlando Berkeley National Laboratories, and Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720

Contributed by David Botstein, July 17, 2001

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

Ash A. Alizadeh[1,2], Michael B. Eisen[2,3,4], R. Eric Davis[5], Chi Ma[5], Izidore S. Lossos[6], Andreas Rosenwald[5], Jennifer C. Boldrick[1], Hajeer Sabet[5], Truc Tran[5], Xin Yu[5], John I. Powell[7], Liming Yang[7], Gerald E. Marti[8], Troy Moore[9], James Hudson Jr[9], Lisheng Lu[10], David B. Lewis[10], Robert Tibshirani[11], Gavin Sherlock[4], Wing C. Chan[12], Timothy C. Greiner[12], Dennis D. Weisenburger[12], James O. Armitage[13], Roger Warnke[14], Ronald Levy[6], Wyndham Wilson[15], Michael R. Grever[16], John C. Byrd[17], David Botstein[4], Patrick O. Brown[1,18] & Louis M. Staudt[5]

Departments of [1]Biochemistry, [3]Genetics, [14]Pathology, [6]Medicine, [10]Pediatrics and [11]Health Research & Policy and Statistics, and [18]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA
[5] Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[7] Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA
[8] CBER, FDA, Bethesda, Maryland 20892, USA
[9] Research Genetics, Huntsville, Alabama 35801, USA
Departments of [12]Pathology and Microbiology, and [13]Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA
[15] Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[16] Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA
[17] Walter Reed Army Medical Center, Washington, DC 20307, USA

Gene expression profiling predicts clinical outcome of breast cancer

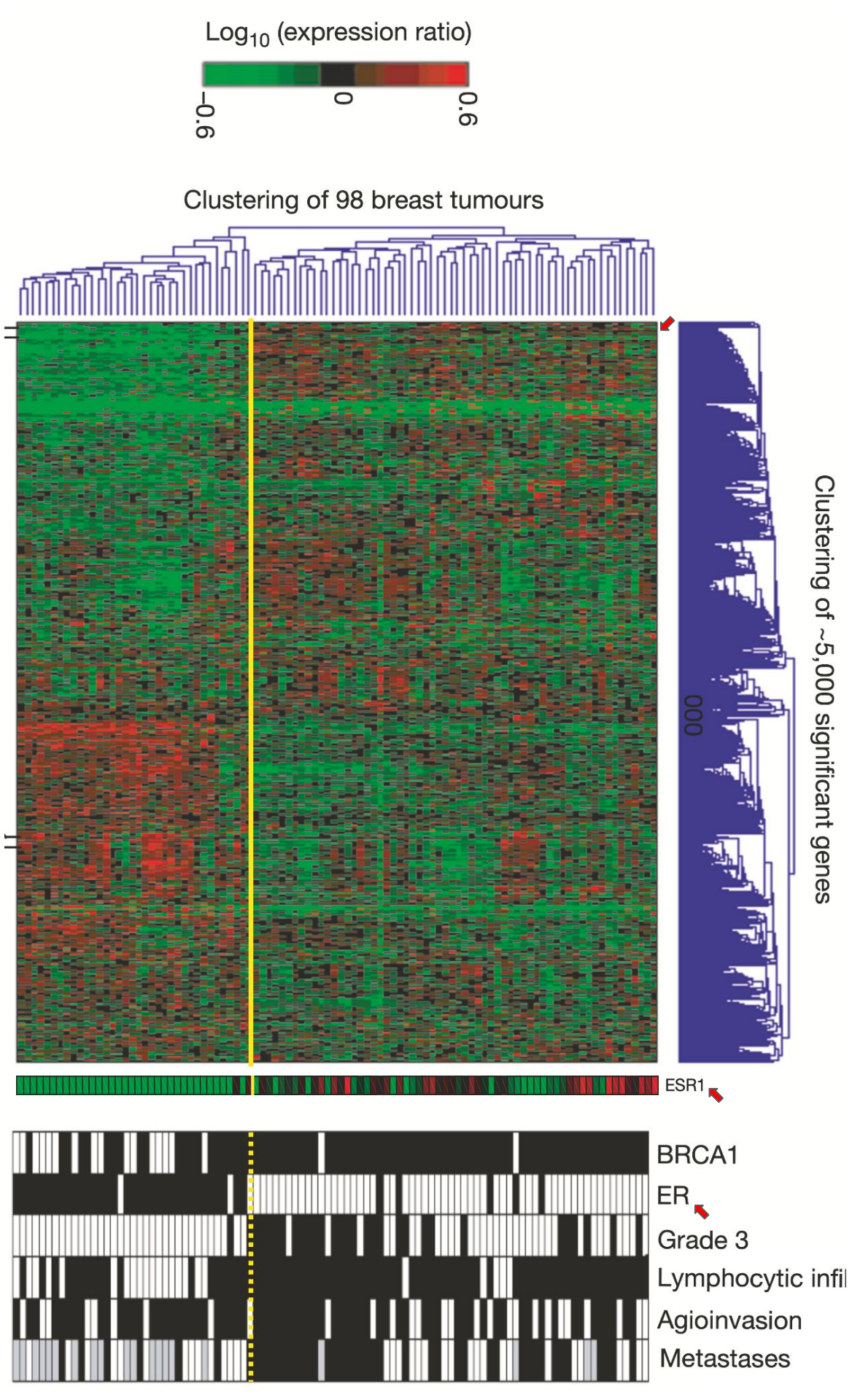
Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

** Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*
‡ Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034,

Could gene expression predict clinical outcout?

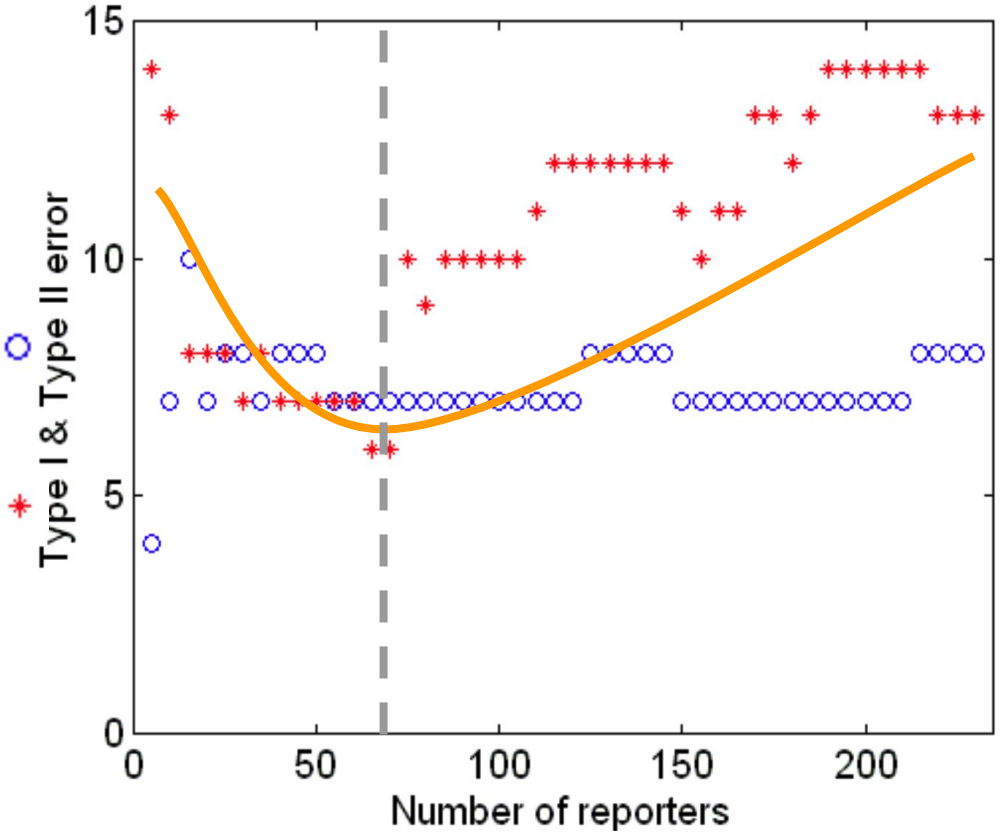
- Early breast cancer, at young patients.
- Chemotherapy reduces risk of metastases by ⅓
- 70-80% patients receiving it, would have survived without.
- How to identify which patients would likely need chemo, and which won't?

- **34** patients developed distant metastases within 5 years
- **44** were disease-free after 5 years
- (18 with BRCA1, 2 BRCA2 mutations)


Log10 (expression ratio)
−0.6
0
0.6
Clustering of 98 breast tumours
Clustering of ~5,000 significant genes
ESR1
BRCA1
ER
Grade 3
Lymphocytic infil
Agioinvasion
Metastases

How they learned a classification model?

- **34** “poor prognosis”, **44** “good prognosis” samples
- From 23K genes, found top 231 correlated genes (“features”)



- Leave-one-out CV
- Correlation-based classifier (“poor” or “good” prognosis)
- “Optimal set” of 70 genes

AGENDIA®
MAMMAPRINT+BLUEPRINT

15 years later…

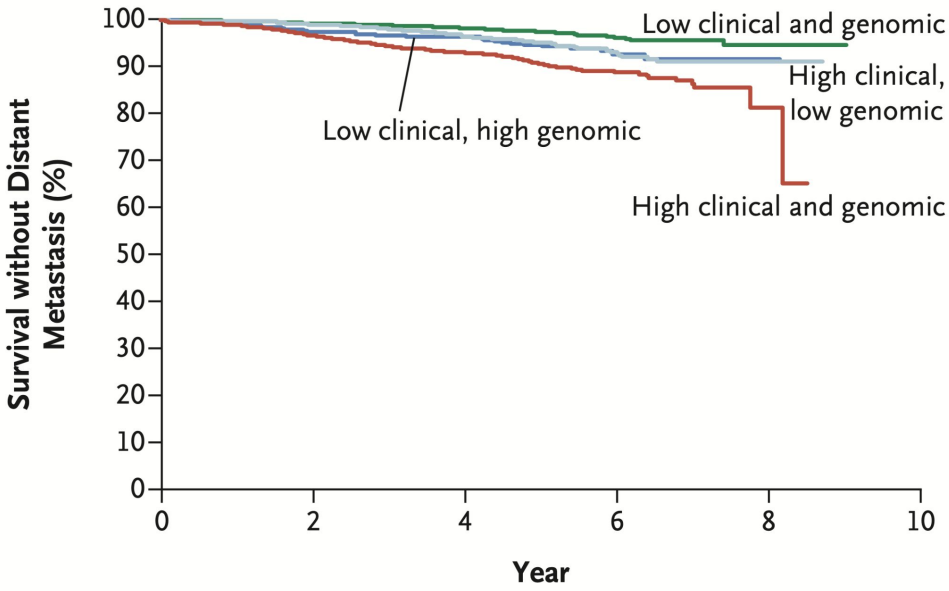
The NEW ENGLAND JOURNAL of MEDICINE

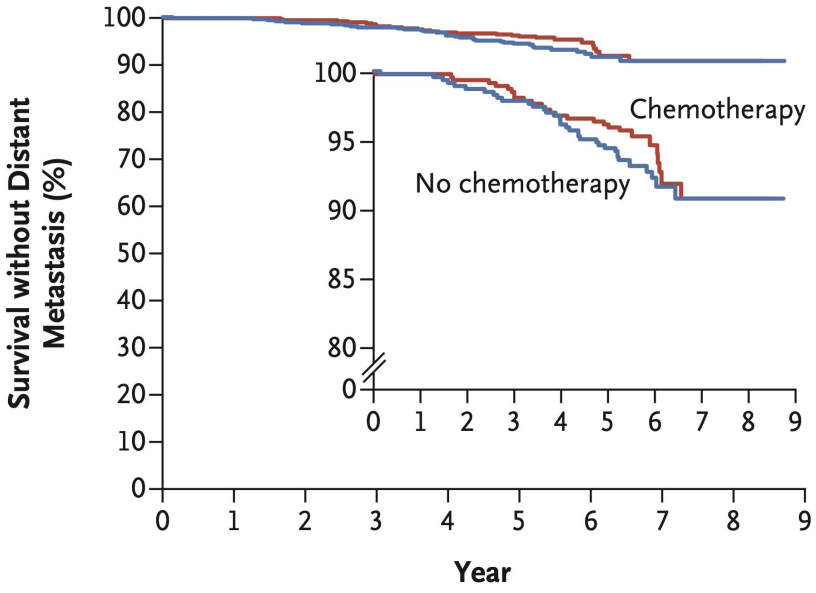
ESTABLISHED IN 1812 AUGUST 25, 2016 VOL. 375 NO. 8

- 6693 women
- 1550: high clinical risk, low genomic risk

70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer

F. Cardoso, L.J. van't Veer, J. Bogaerts, L. Slaets, G. Viale, S. Delaloge, J.-Y. Pierga, E. Brain, S. Causeret, M. DeLorenzi, A.M. Glas, V. Golfinopoulos, T. Goulioti, S. Knox, E. Matos, B. Meulemans, P.A. Neijenhuis, U. Nitz, R. Passalacqua, P. Ravdin, I.T. Rubio, M. Saghatchian, T.J. Smilde, C. Sotiriou, L. Stork, C. Straehle, G. Thomas, A.M. Thompson, J.M. van der Hoeven, P. Vuylsteke, R. Bernards, K. Tryfonidis, E. Rutgers, and M. Piccart, for the MINDACT Investigators*





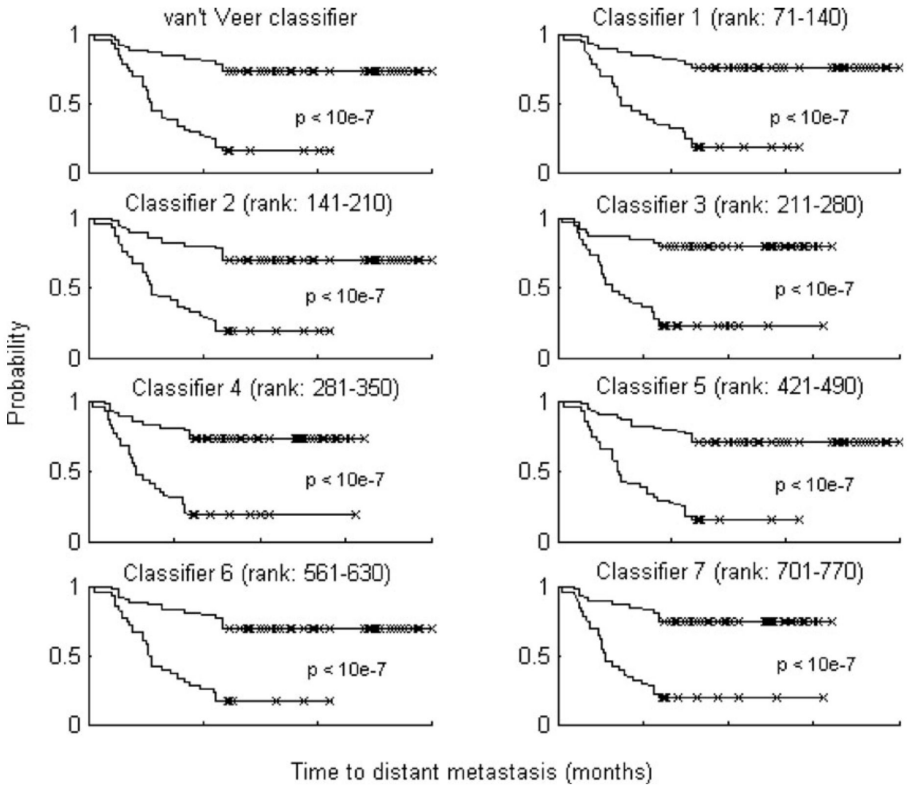
Is this a special set of genes?

Outcome signature genes in breast cancer: is there a unique set?

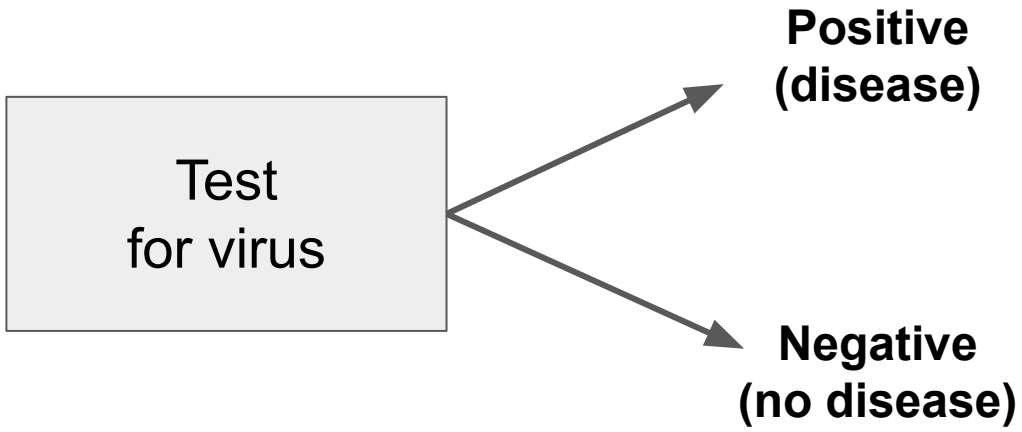
Liat Ein-Dor[1,†], Itai Kela[1,3,†], Gad Getz[1,†], David Givol[2] and Eytan Domany[1,]*

[1]Department of Physics of Complex Systems, [2]Department of Molecular Cell Biology and [3]Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel

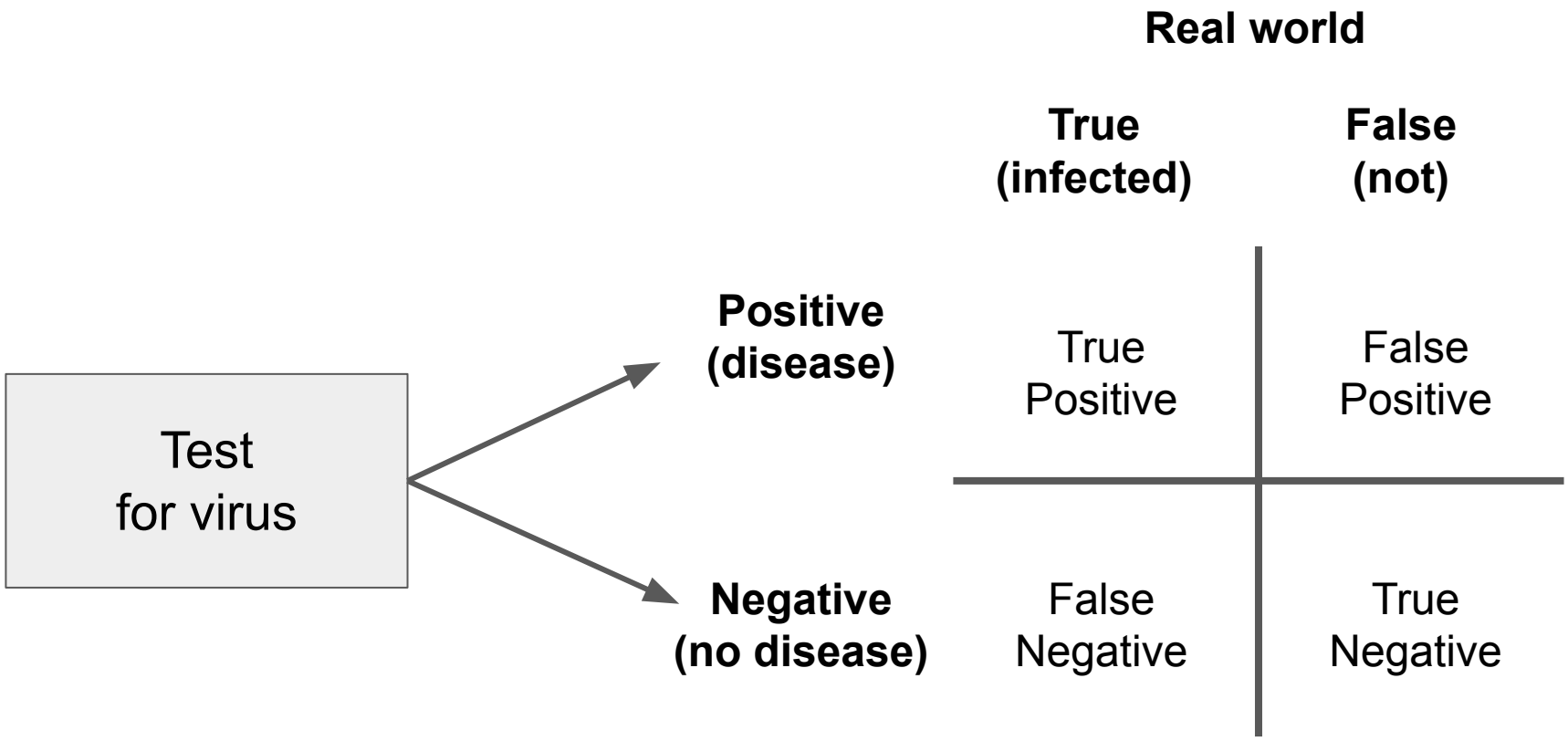
Received on June 4, 2004; revised on August 2, 2004; accepted on August 3, 2004



More on Error



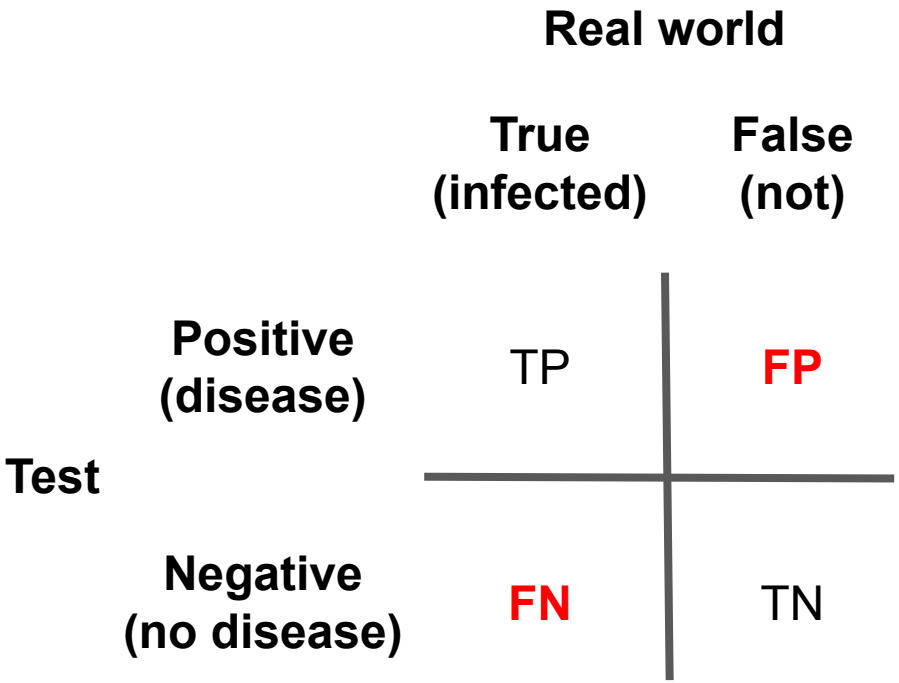
More on Error



More on Error

		Real world: True (infected)	Real world: False (not)
Test	Positive (disease)	TP	FP
	Negative (no disease)	FN	TN

More on Error



FP = Type I error

FN = Type II error

Cost of error

Different costs depend on context

Infectious diseases

Type I error (FP):

- Unneeded treatment (cost, side effects etc)
- Burden on system
- Risk to a healthy patient

Type II error (FN):

- Untreated condition (worse outcome)
- Increase disease propagation

Cost of error

Different costs depend on context

Genetic test (e.g. BRCA1)

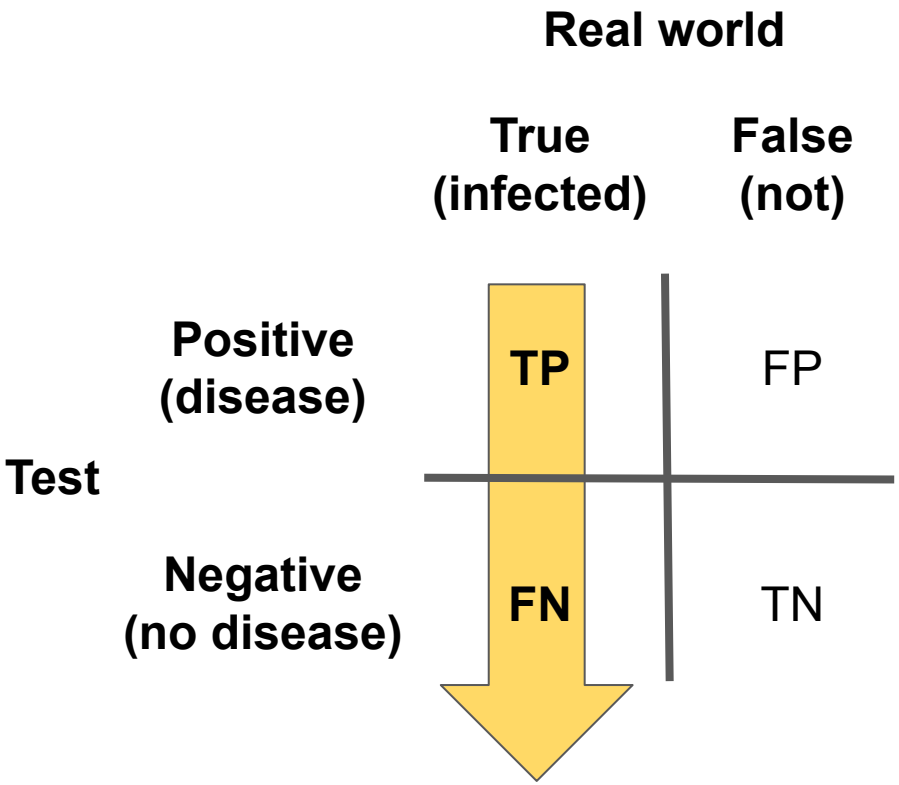
Type I error (FP):

- Unneeded monitoring (cost, patient time)
- Unneeded pre-emptive treatment (cost, patient health)
- Mental cost (stress)

Type II error (FN):

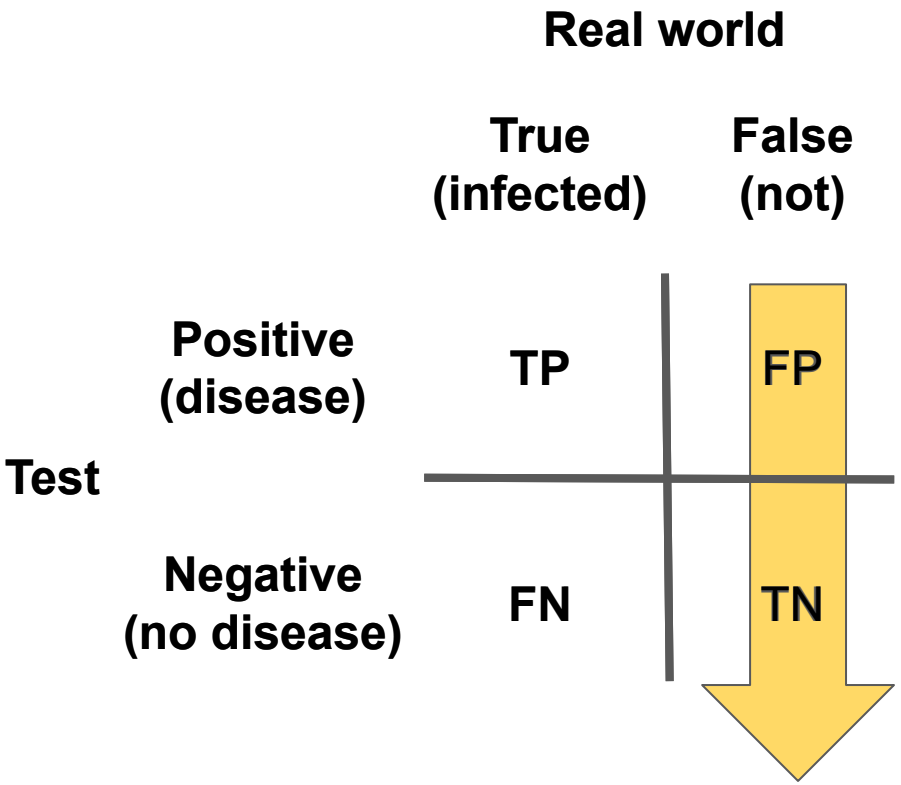
- Missed chance for early monitoring (worse outcome)

Evaluating errors



$$\textbf{sensitivity} \quad \frac{TP}{TP+FN}$$

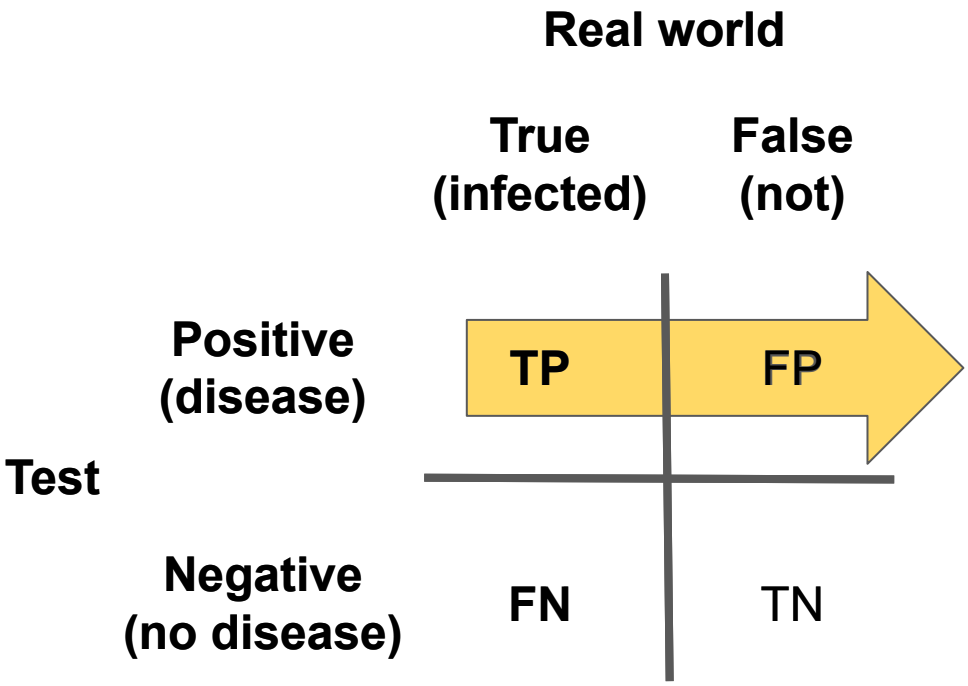
Evaluating errors



$$\text{sensitivity} \quad \frac{TP}{TP+FN}$$

$$\textbf{specificity} \quad \frac{TN}{TN+FP}$$

Evaluating errors

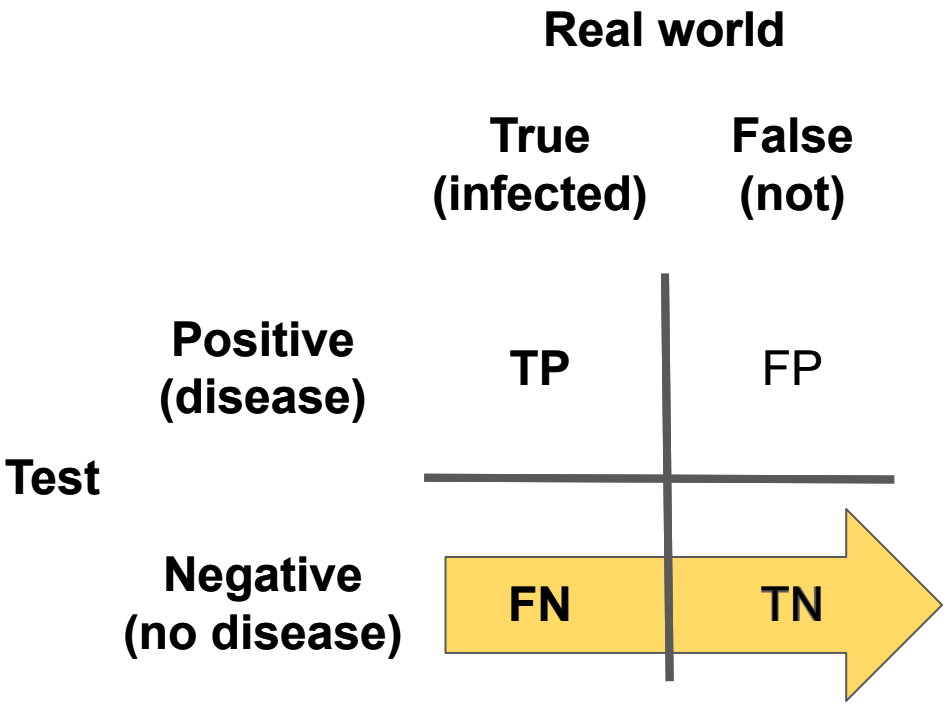


$$\text{sensitivity} = \frac{TP}{TP+FN}$$

$$\text{specificity} = \frac{TN}{TN+FP}$$

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

Evaluating errors



$$\text{sensitivity} = \frac{TP}{TP+FN}$$

$$\text{specificity} = \frac{TN}{TN+FP}$$

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

$$\textbf{NPV} = \frac{TN}{TN+FN}$$

Evaluating errors

Sensitivity/specificity: test quality

$$\text{sensitivity} = \frac{TP}{TP+FN}$$

$$\text{specificity} = \frac{TN}{TN+FP}$$

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

$$\text{NPV} = \frac{TN}{TN+FN}$$

Evaluating errors

Sensitivity/specificity: test quality

PPV/NPV:

Disease prevalence & test quality

Same test can have different PPV/NPV

- General population

vs

- Population at risk

$$\text{sensitivity} = \frac{TP}{TP+FN}$$

$$\text{specificity} = \frac{TN}{TN+FP}$$

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

$$\text{NPV} = \frac{TN}{TN+FN}$$

Syllabus

1	1/1	AI in ophthalmology (Prof. Itay Chowers)
2	8/1	Classification
3	15/1	Learning 1
4	22/1	Learning 2
5	7/2	Regression (Wed.)
6	12/2	Deep learning in image analysis (Prof. Leo Joskowicz)
7	19/2	Clustering
8	26/2	Dimensionality reduction and visualization
9	28/2	Deep learning, Missing data (Wed.)
10	4/3	Natural language in medicine (Dr. Gabi Stanovsky)
11	11/3	?