



האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

Artificial Intelligence in Medicine

Learning (2)

Nir Friedman and Tommy Kaplan 21/11/22

“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?

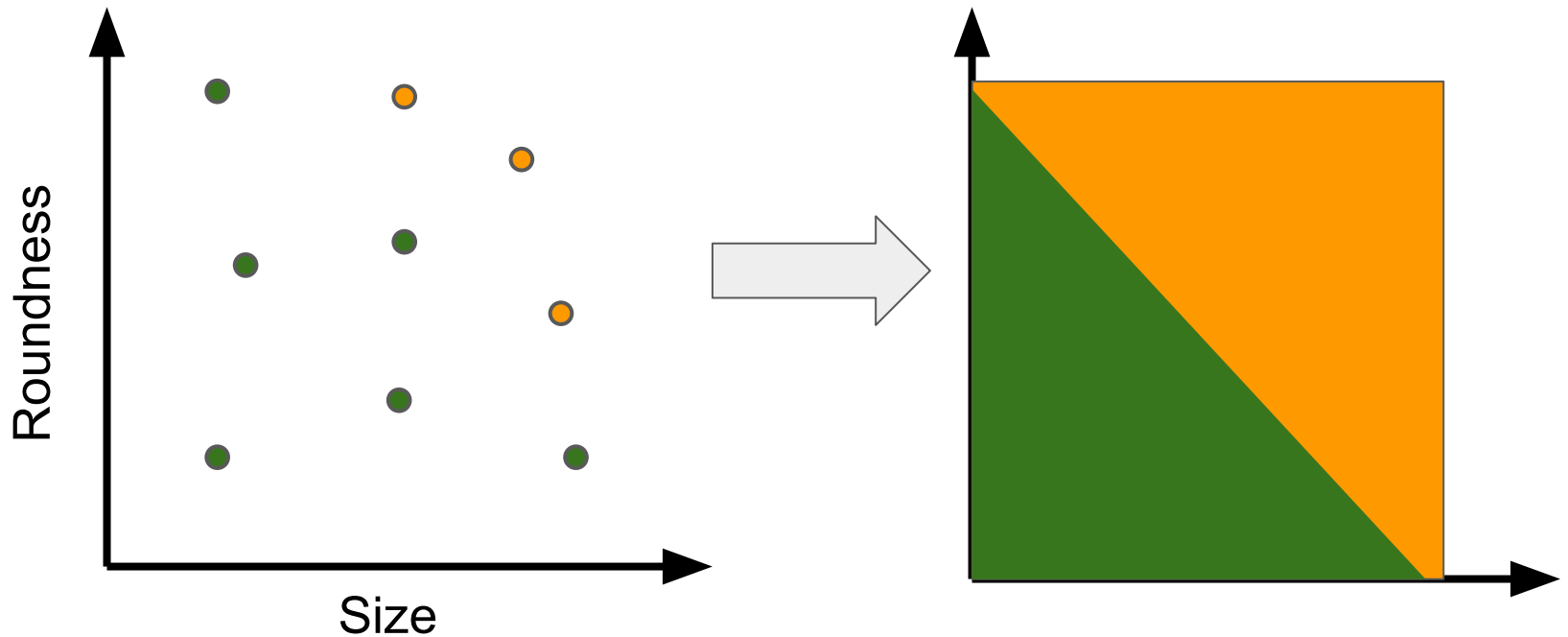
Virus?

- An elderly man, presenting stomach ache

Colon cancer?

Virus?

What is learning?



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)

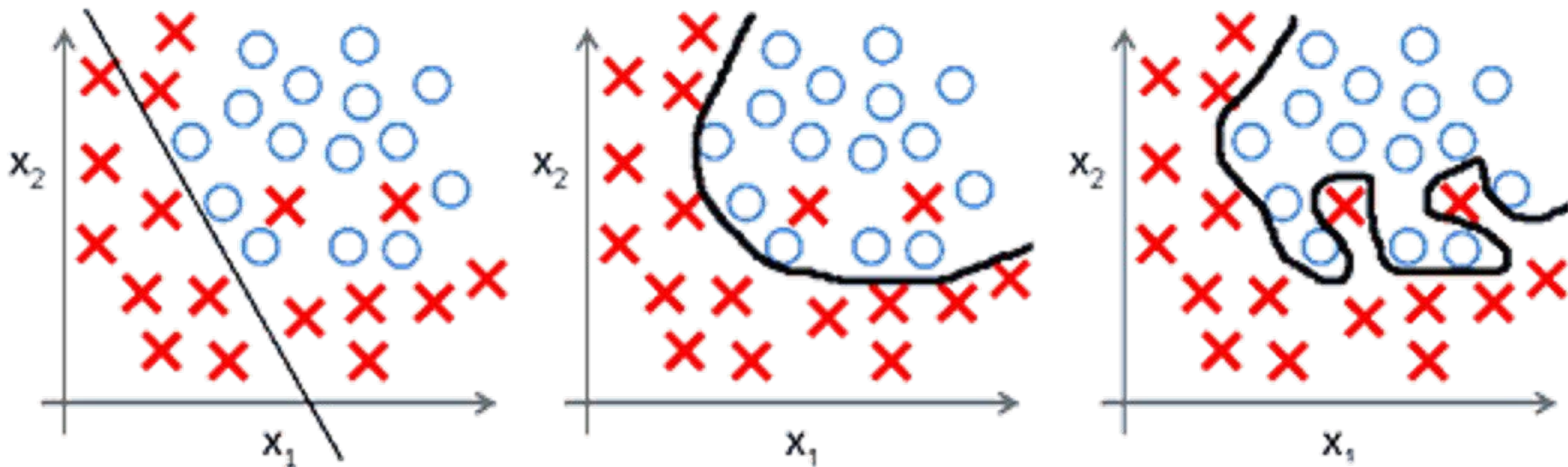
Test error - error on test set (yet unseen)

Are these related?

Implicit assumption:

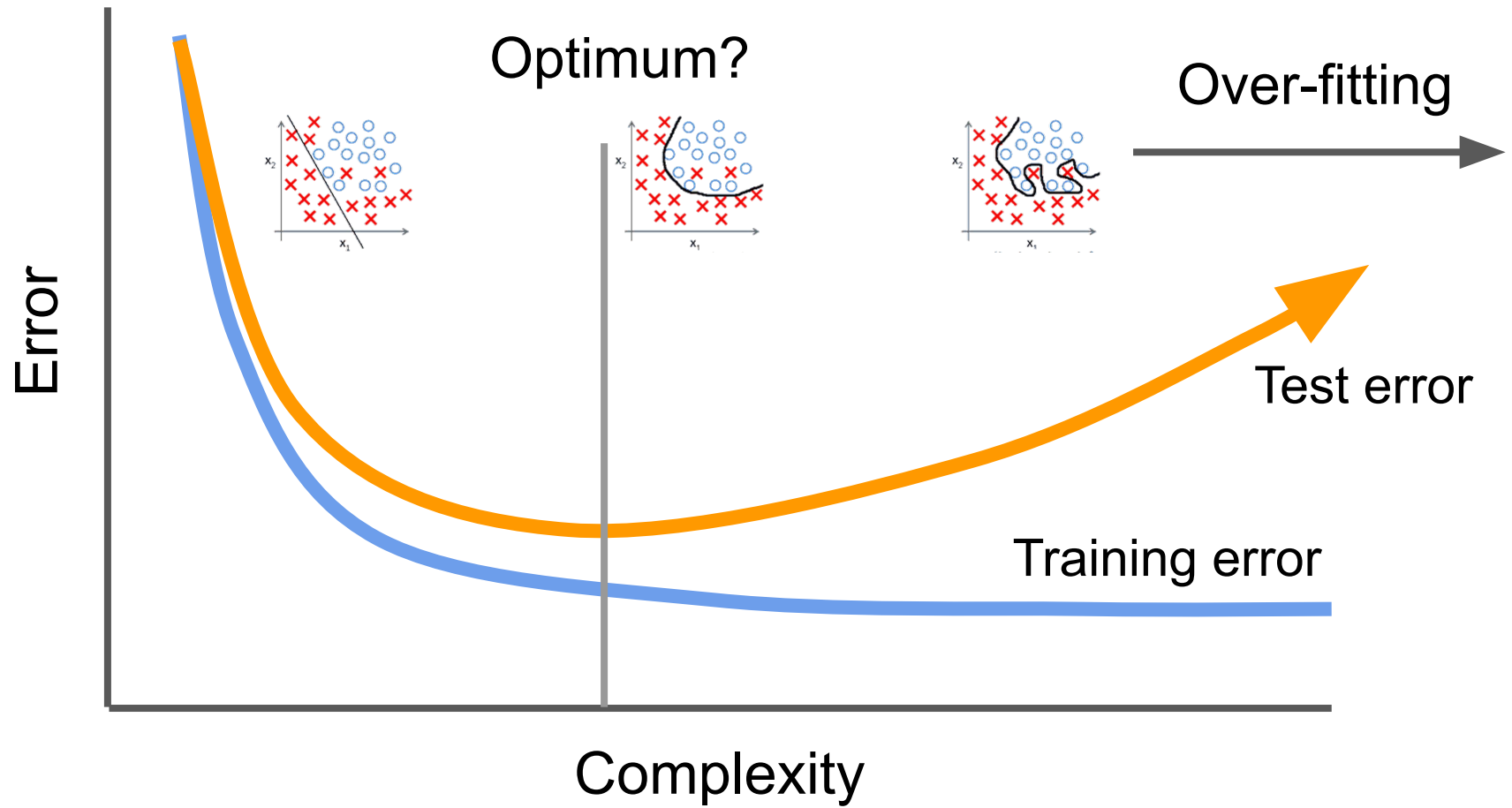
- Reducing training error will reduce also test error
- Is this reasonable?

Different classifiers



Complexity

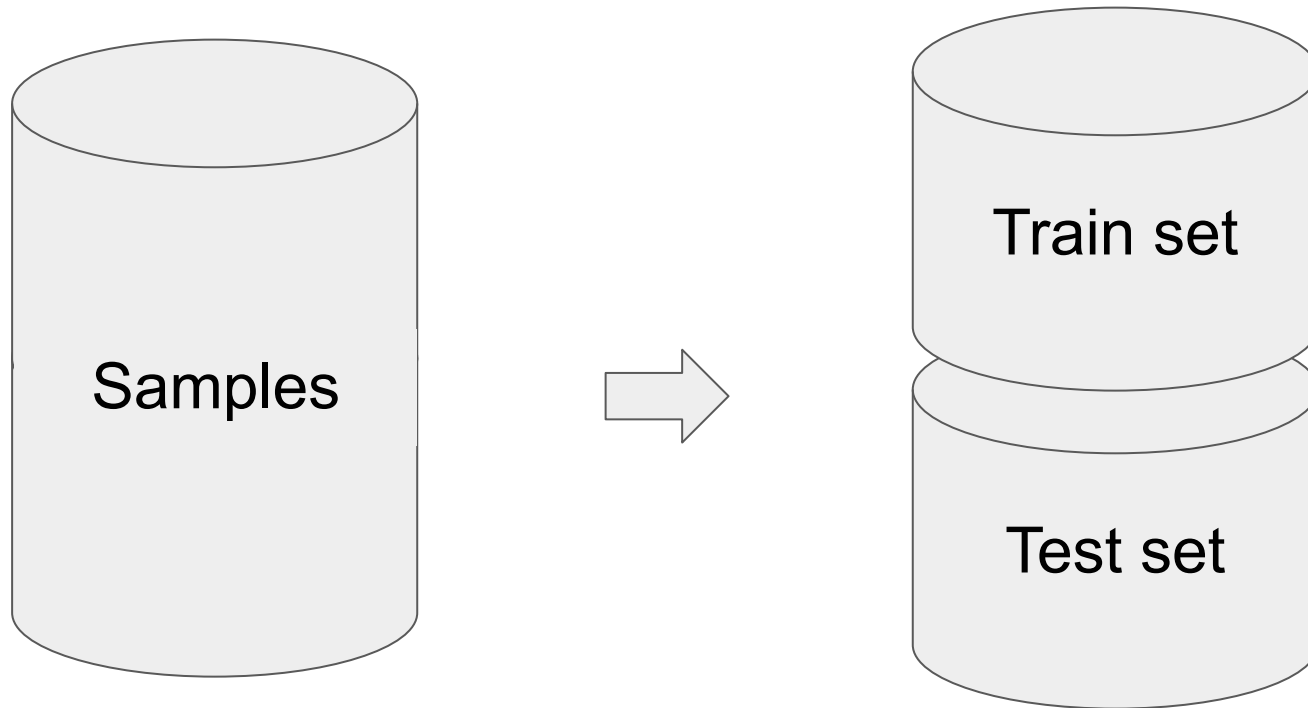
Train vs. test errors



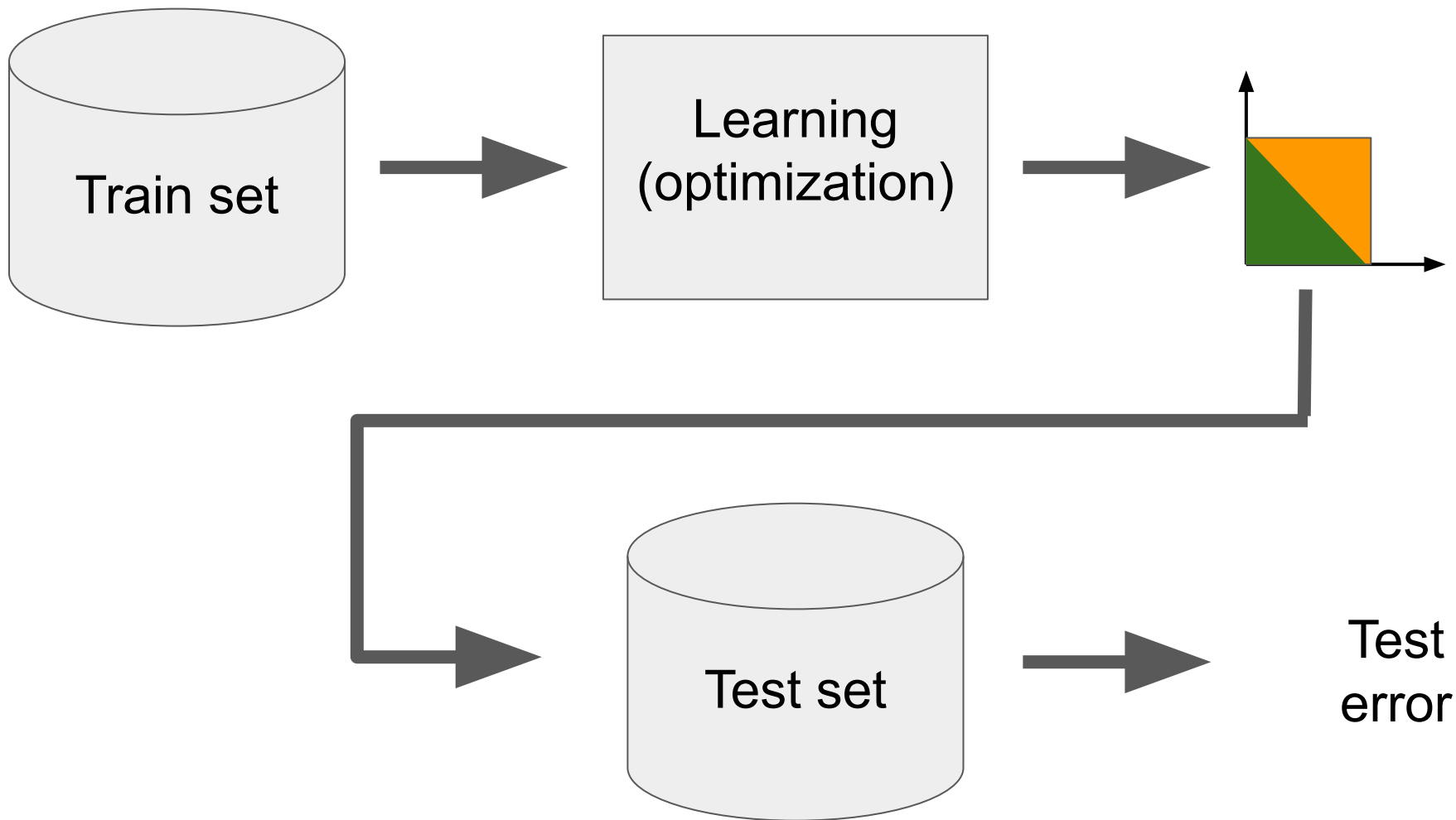
Measures of complexity

- 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

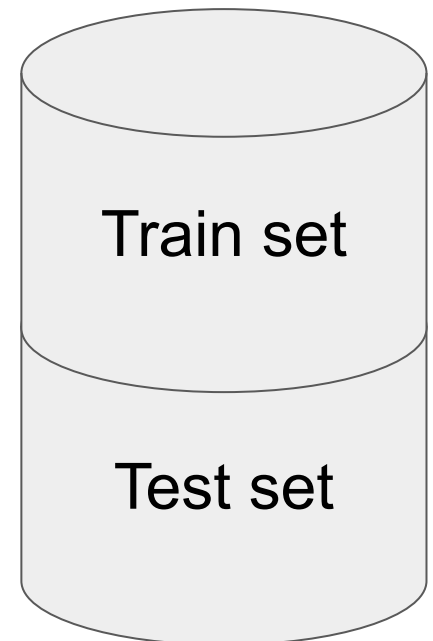
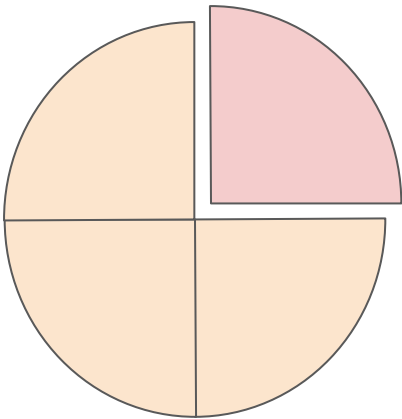


Empirical Test Error

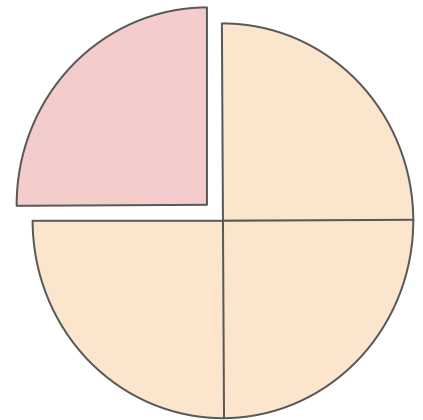
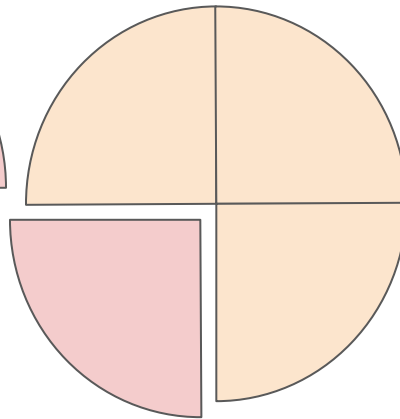
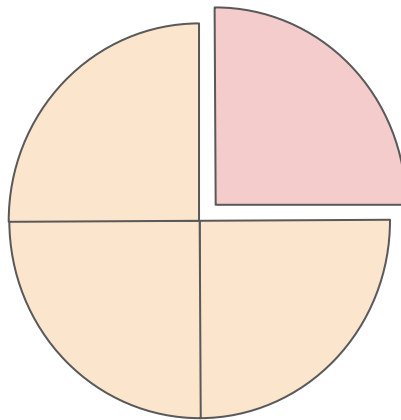
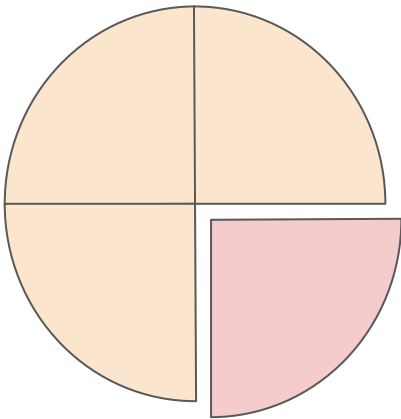
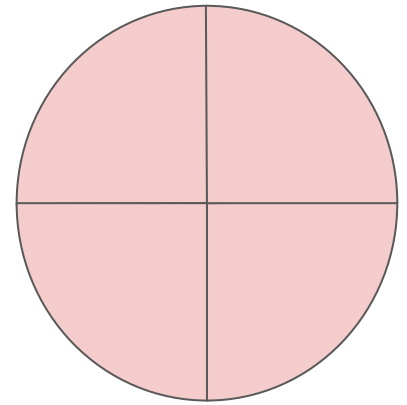
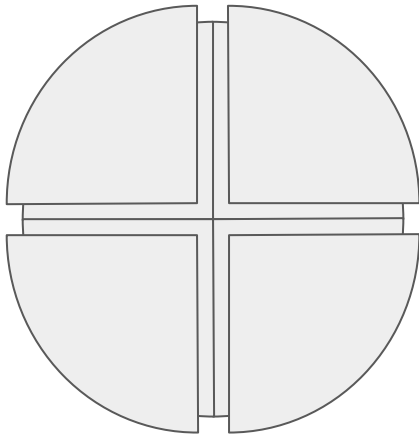
Issues:

Train/Test allocation

- Small train set $\rightarrow$ not enough for learning
- Small test set $\rightarrow$ noisy error estimation

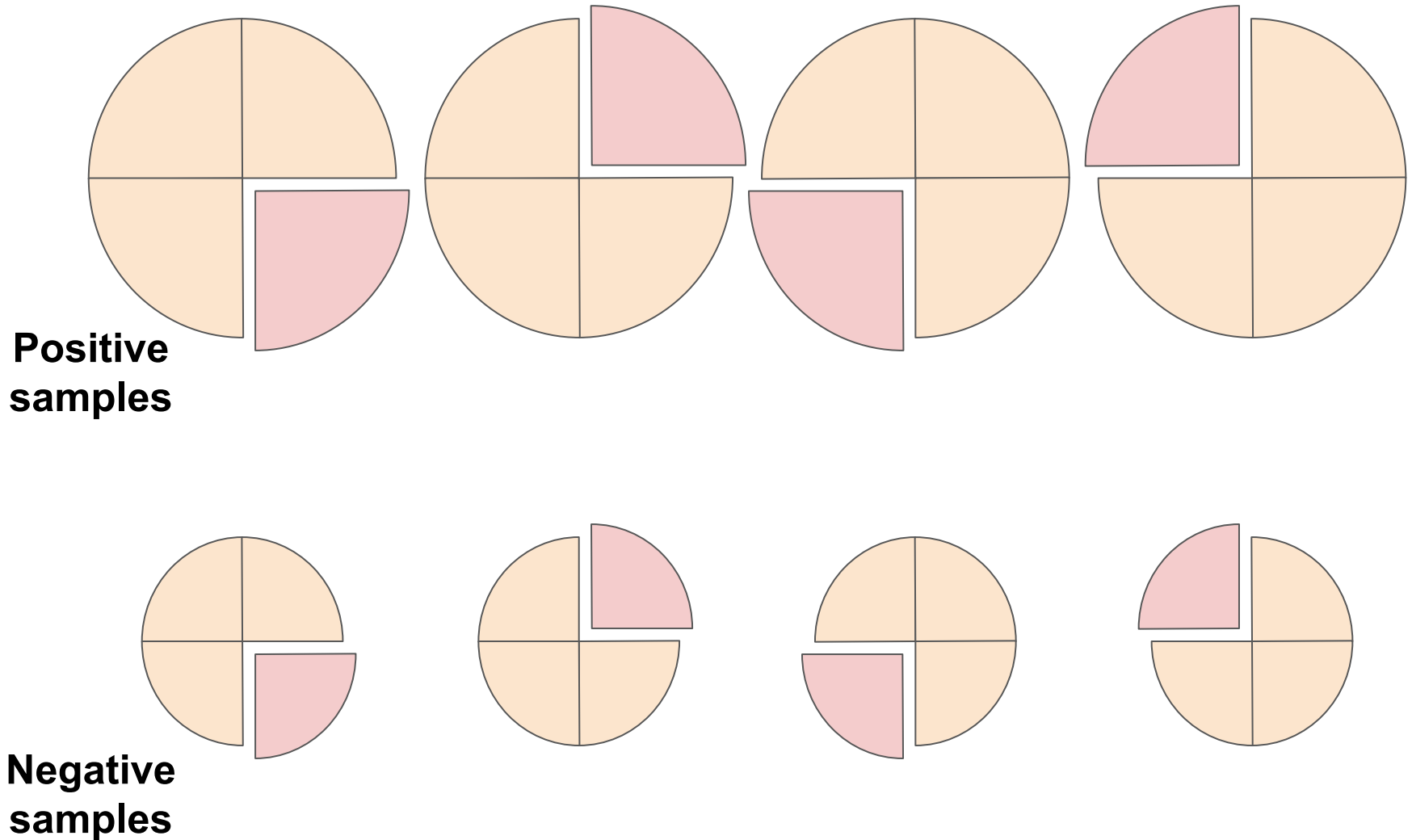


(4-fold) cross validation



Train
Test

Stratified cross validation

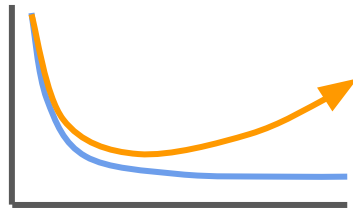


Empirical Test Error

Issues:

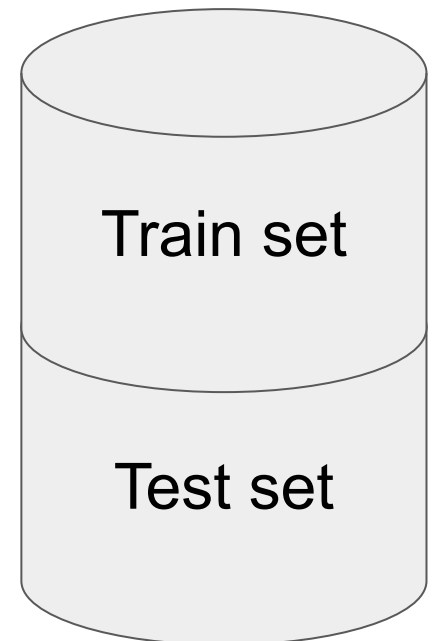
Train/Test allocation

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

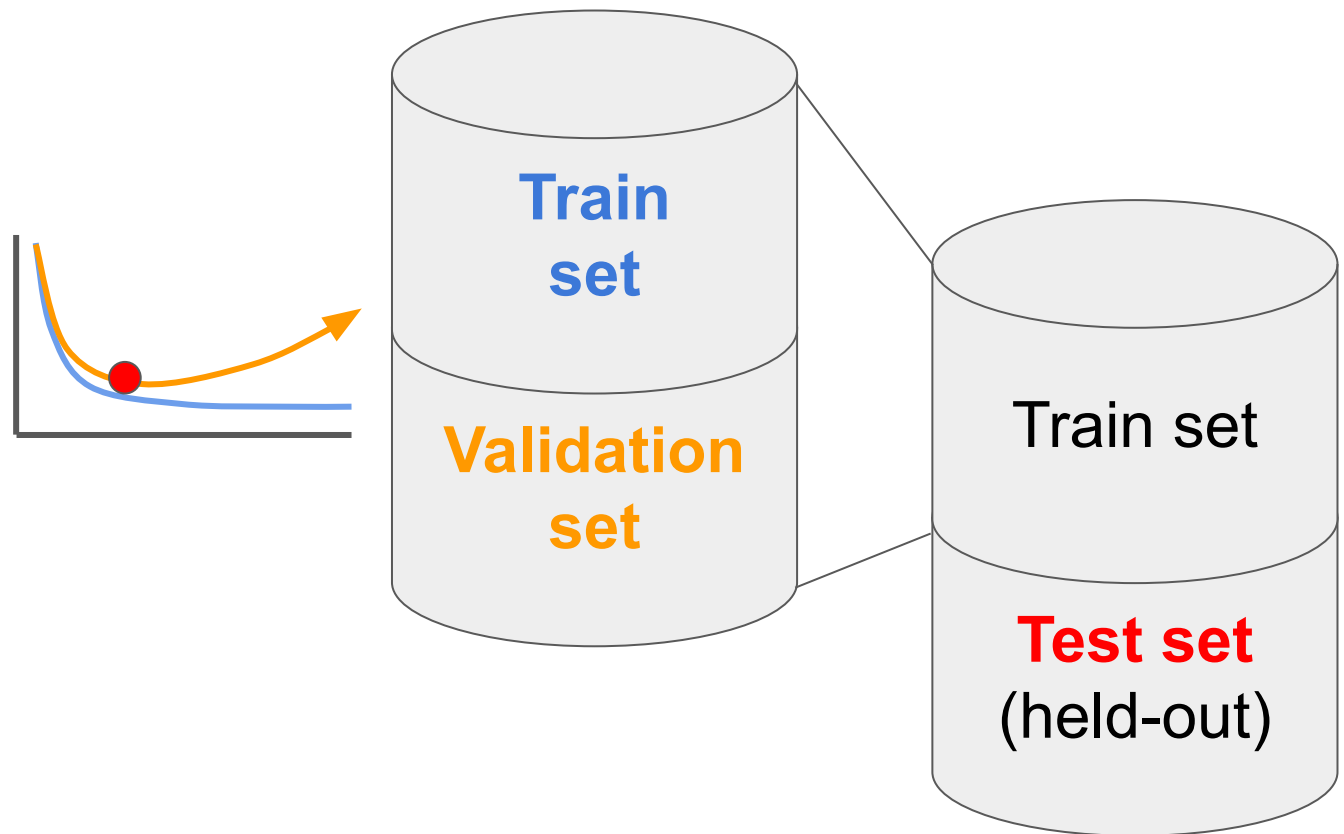


Use of Test set

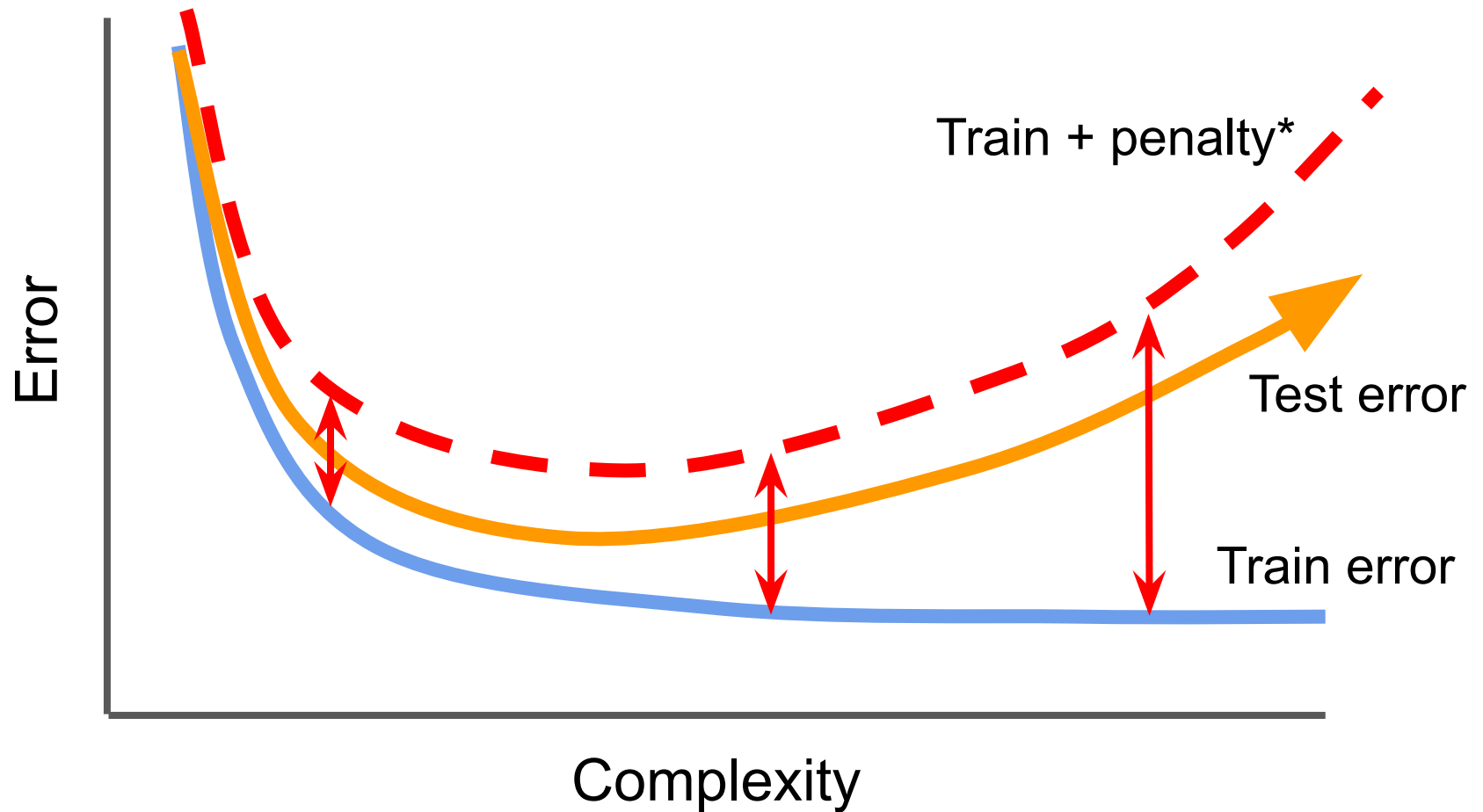
- Multiple evaluations (of classifiers)
- **Test samples used for learning**
- **Over-fitting** the test data!!



Empirical Test Error



Theoretical bound of test error



Penalty* = function of model complexity

Summary so far

Choose complexity (hyper-parameters)

Estimate generalization performance

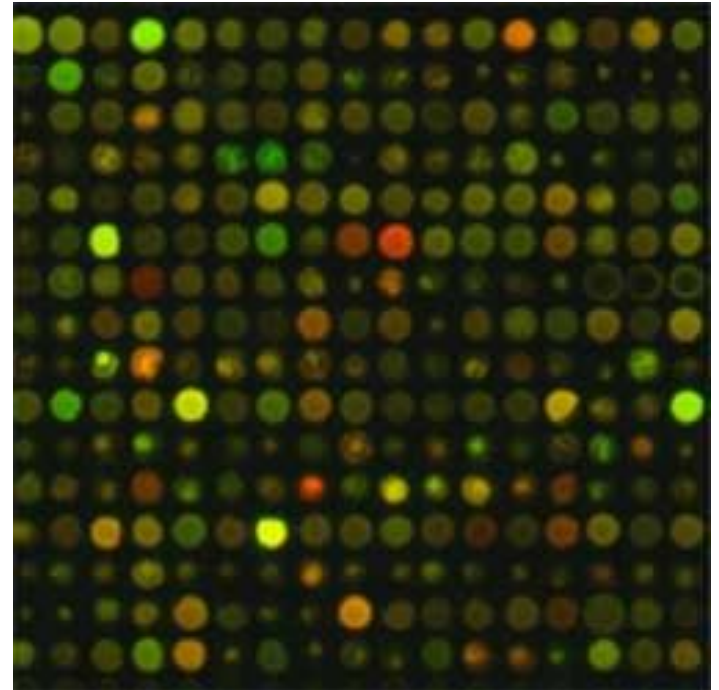
- Train/test split
- Cross-validation
- Theoretical limits

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 Rijnⁱ, 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 ^bGenetics and ^lSurgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; ^dDepartment of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of ^eHealth Research and Policy and Statistics, ^cGenetics, ⁱPathology, ^jSurgery, and ^mBiochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of ^gMedicine (Section of Oncology), ^fSurgery, and ^kBiochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and ^hLife Sciences Division, Lawrence Orloand 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 Davis⁵, Chi Ma⁵, Izidore S. Lossos⁶, Andreas Rosenwald⁵, Jennifer C. Boldrick¹, Hajeer Sabet¹, Truc Tran⁵, Xin Yu⁵, John I. Powell⁷, Liming Yang⁷, Gerald E. Marti⁸, Troy Moore⁹, James Hudson Jr², Lisheng Lu¹⁰, David B. Lewis¹⁰, Robert Tibshirani¹¹, Gavin Sherlock⁴, Wing C. Chan¹², Timothy C. Greiner¹², Dennis D. Welschberger¹², James O. Armitage¹³, Roger Warnke¹⁴, Ronald Levy⁵, Wyndham Wilson¹⁵, Michael R. Grever¹⁶, John C. Byrd¹⁷, David Botstein⁴, Patrick O. Brown^{1,18} & Louis M. Staudt⁵

Departments of ¹Biochemistry, ²Genetics, ¹⁴Pathology, ⁶Medicine, ¹⁰Pediatrics and ¹¹Health Research & Policy and Statistics, and ¹⁸Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA

⁵Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

⁷Bioinformatics and Molecular Analysis Section, CBER, CIT, NIH, Bethesda, Maryland 20892, USA

⁸CBER, FDA, Bethesda, Maryland 20892, USA

⁹Research Genetics, Huntsville, Alabama 35801, USA

Departments of ¹²Pathology and Microbiology, and ¹³Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA

¹⁵Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

¹⁶Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA

¹⁷Walter 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[‡]

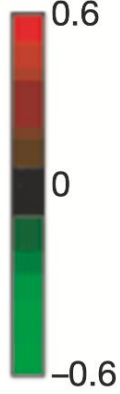
^{*} Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands

[‡] Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034,

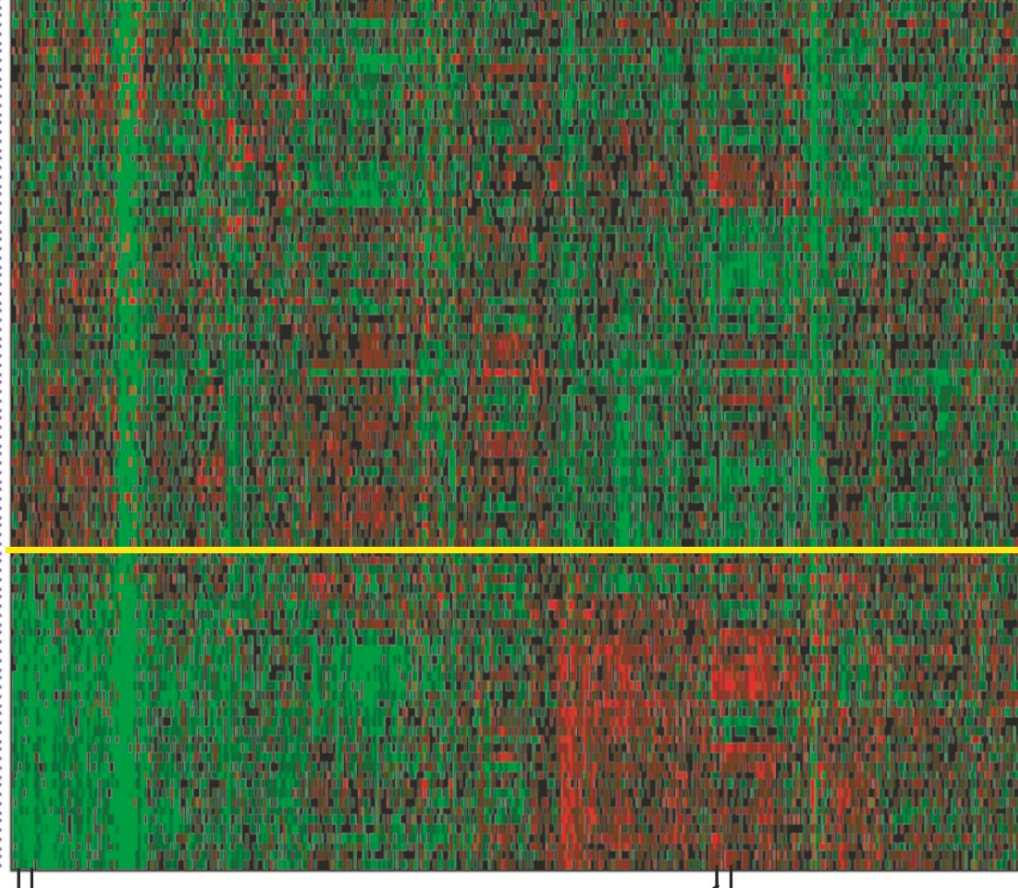
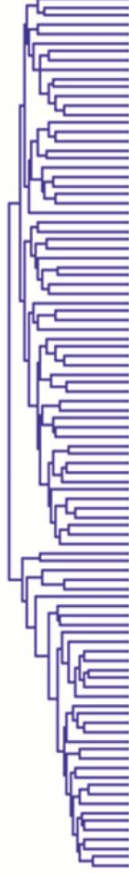
Could gene expression predict clinical outcome?

- Early breast cancer, at young patients.
 - Chemotherapy reduces risk of metastases by $\frac{1}{3}$
 - 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)

Log_{10} (expression ratio)



Clustering of 98 breast tumours



Clustering of ~5,000 significant genes



ESR1



BRCA1

ER

Grade 3

Lymphocytic infil

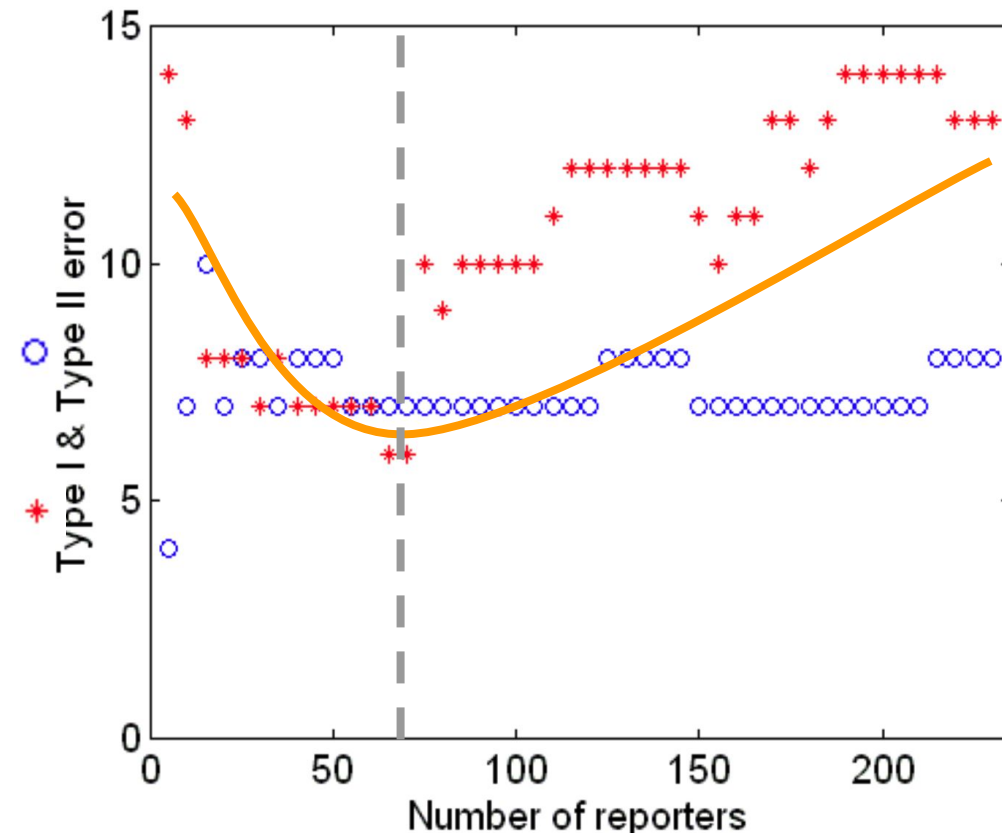
Angioinvasion

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

15 years later...

The NEW ENGLAND JOURNAL *of* MEDICINE

ESTABLISHED IN 1812

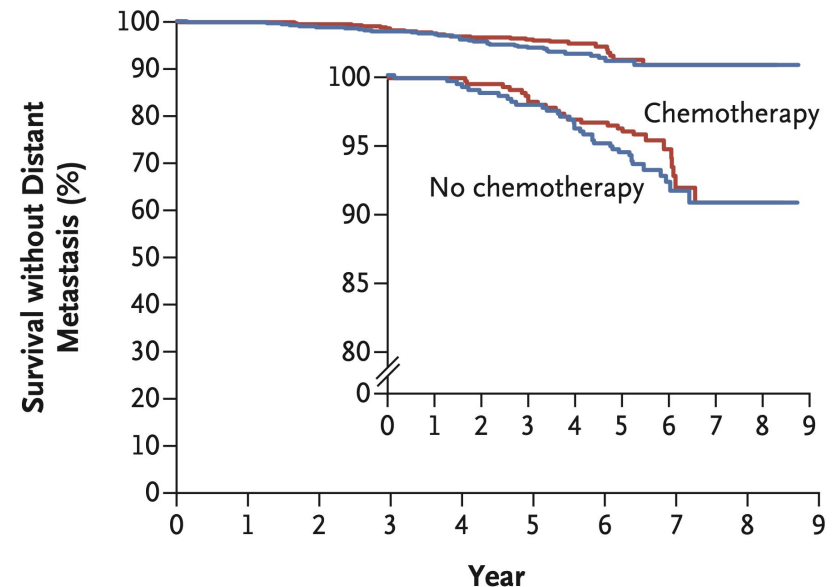
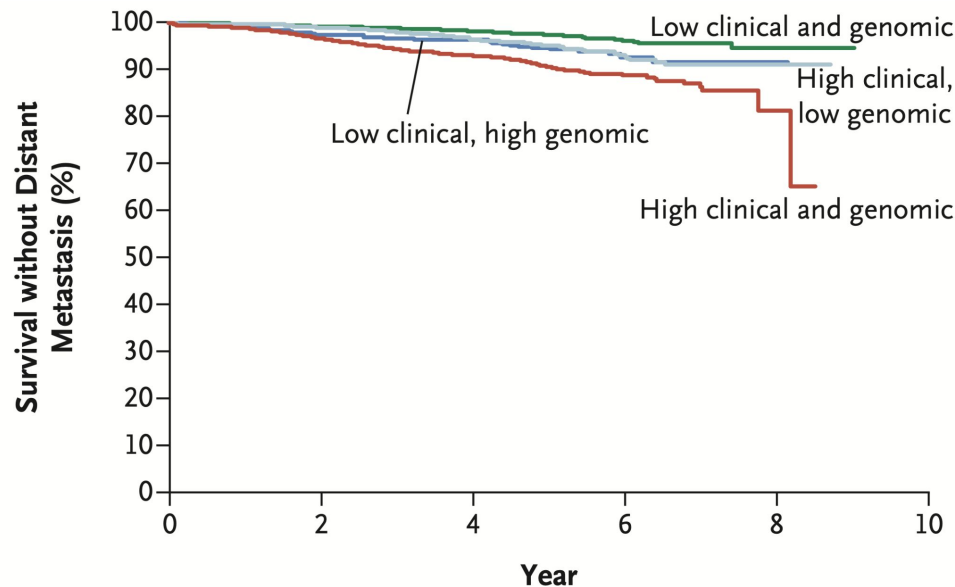
AUGUST 25, 2016

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. Goulinopoulos, 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*



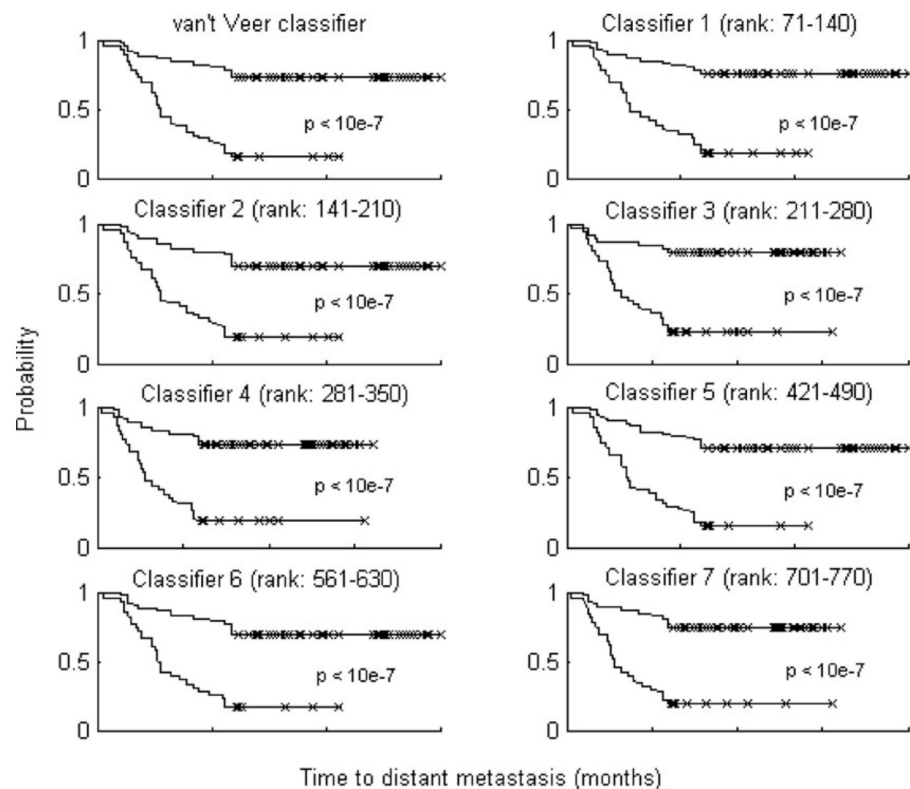
Is this a special set of genes?

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

Liat Ein-Dor^{1,†}, Itai Kela^{1,3,†}, Gad Getz^{1,†}, David Givol² and
Eytan Domany^{1,*}

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

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



Syllabus

1. Introduction
2. Classification
3. Learning 1
4. AI in ophthalmology (Prof. Itay Chowers)
5. Learning 2
6. Regression
7. Clustering
8. Visualization (and dimensionality reduction)
9. Deep learning in image analysis (Prof. Leo Joskowicz)
10. Missing data, statistical dependencies
11. Natural language in medicine (Dr. Gabi Stanovsky)
12. Decisions (utility)
13. Longitudinal Data / Project