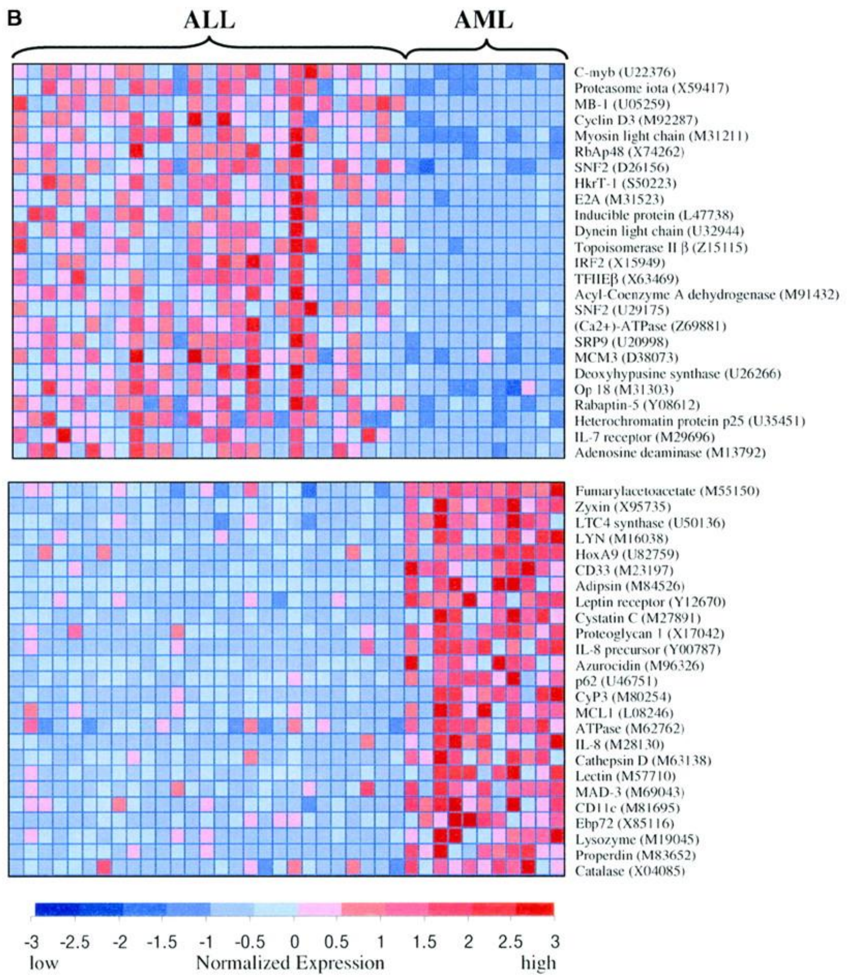
Proteomics Informatics – Molecular signatures (Week 11)

Definition of a molecular signature

A **molecular signature** is a computational or mathematical model that links high-dimensional molecular information to phenotype or other response variable of interest.



FDA calls them "in vitro diagnostic multivariate assays"

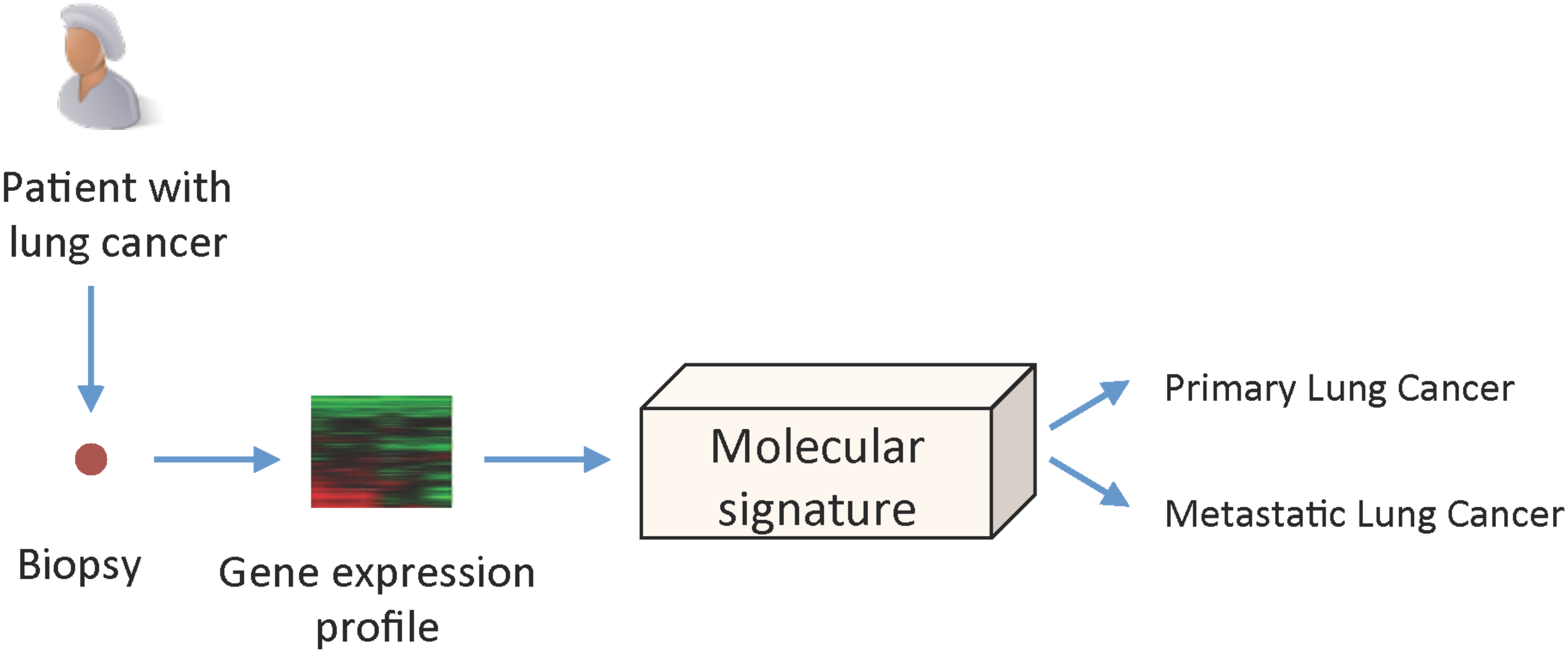
Uses of molecular signatures

1. Models of disease phenotype/clinical outcome
   - Diagnosis
   - Prognosis, long-term disease management
   - Personalized treatment (drug selection, titration)

2. Biomarkers for diagnosis, or outcome prediction
   - Make the above tasks resource efficient, and easy to use in clinical practice

3. Discovery of structure & mechanisms (regulatory/interaction networks, pathways, sub-types)
   - Leads for potential new drug candidates

Example of a molecular signature



Patient with lung cancer

Biopsy

Gene expression profile

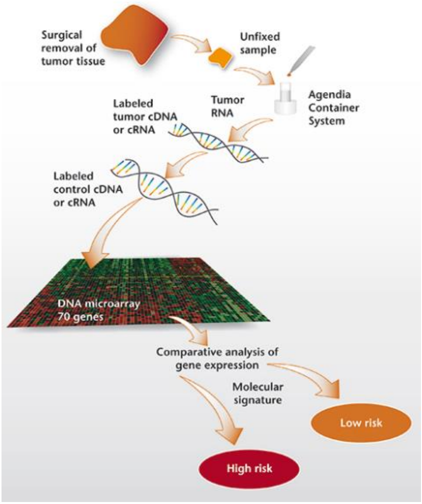
Molecular signature

Primary Lung Cancer

Metastatic Lung Cancer

MammaPrint

- Developed by Agendia (www.agendia.com)
- 70-gene signature to stratify women with breast cancer that hasn't spread into "low risk" and "high risk" for recurrence of the disease
- Independently validated in >1,000 patients
- So far performed >10,000 tests
- Cost of the test is ~$3,000
- In February, 2007 the FDA cleared the MammaPrint test for marketing in the U.S. for node negative women under 61 years of age with tumors of less than 5 cm.
- TIME Magazine's 2007 "medical invention of the year".



Surgical removal of tumor tissue

Unfixed sample

Agendia Container System

Tumor RNA

Labeled tumor cDNA or cRNA

Labeled control cDNA or cRNA

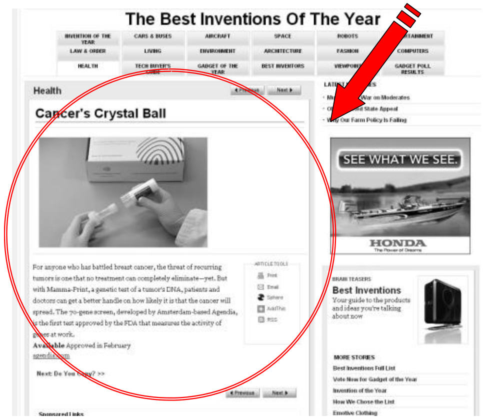
DNA microarray 70 genes

Comparative analysis of gene expression

Molecular signature

Low risk

High risk



The Best Inventions Of The Year

Health

Cancer's Crystal Ball

For anyone who has battled breast cancer, the threat of recurring tumors is one that no treatment can completely eliminate—yet. But with Mamma-Print, a genetic test of a tumor's DNA, patients and doctors can get a better handle on how likely it is that the cancer will spread. The 70-gene screen, developed by Amsterdam-based Agendia, is the first test approved by the FDA that measures the activity of genes at work.

Available Approved in February

agendia.com

Next: Do You Copy? >>

Sponsored Links

LATEST HEADLINES

- War on Moderates
- State Appeal
- Why Our Farm Policy Is Failing

BRAIN TEASERS

Best Inventions

Your guide to the products and ideas you're talking about now

MORE STORIES

Best Inventions Full List

Vote Now for Gadget of the Year

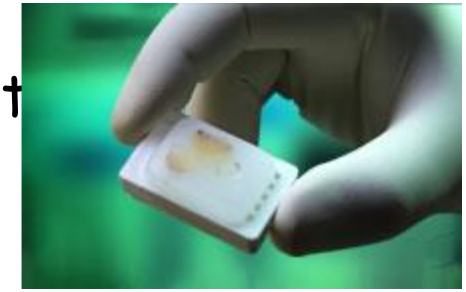
Invention of the Year

How We Chose the List

Emotive Clothing

Oncotype DX Breast Cancer Assay

- Developed by Genomic Health (www.genomichealth.com)
- 21-gene signature to predict whether a woman with localized, ER+ breast cancer is at risk of relapse
- Independently validated in thousands of pat



- So far performed >100,000 tests
- Price of the test is $4,175
- Not FDA approved but covered by most insurances including Medicare
- Its sales in 2010 reached $170M and with a compound annual growth rate is projected to hit $300M by 2015.

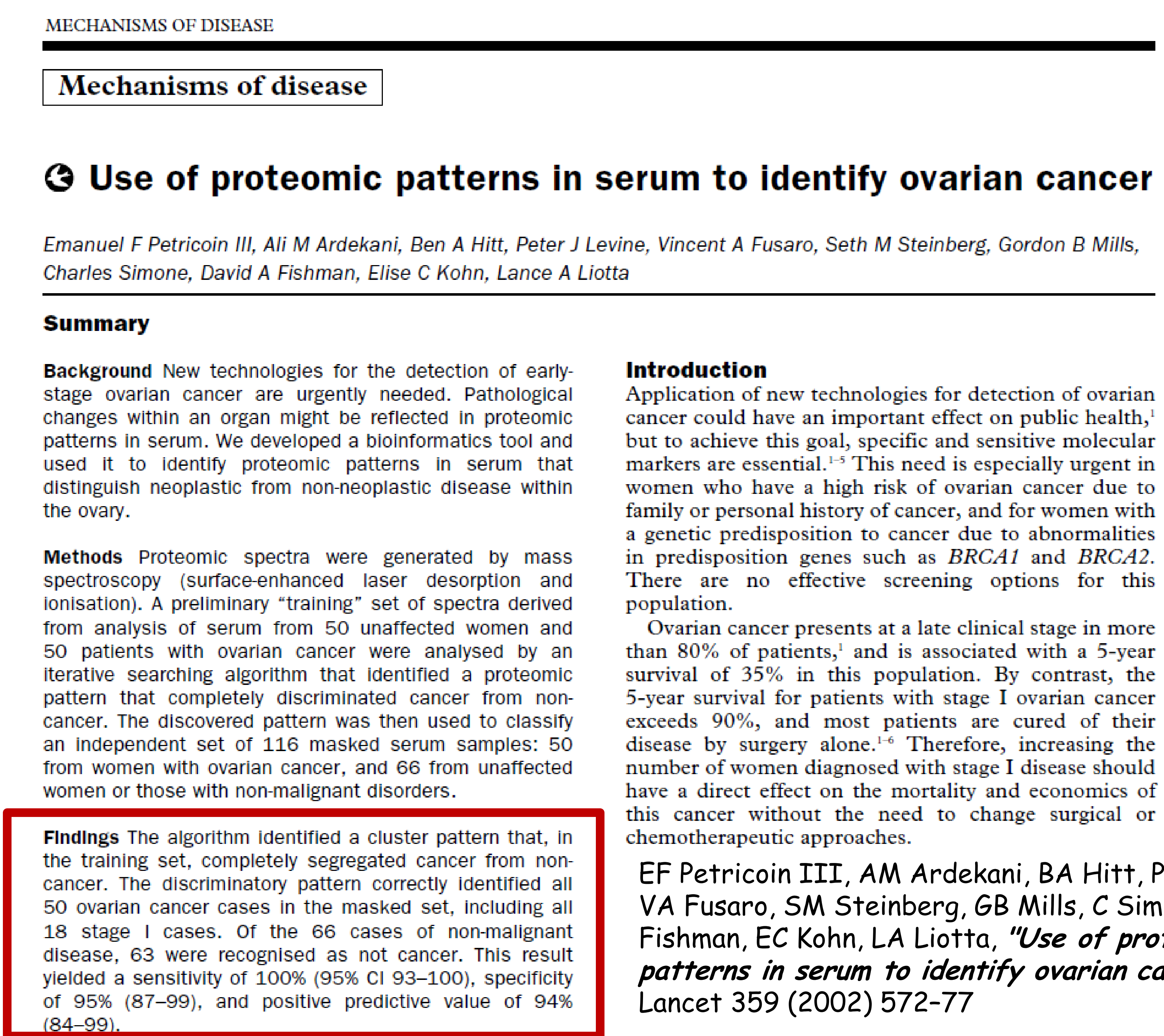
Improved Survival and Cost Savings

In a 2005 economic analysis of recurrence in LN-,ER+ patients receiving tamoxifen, Hornberger et al. performed a cost-utility analysis using a decision analytic model. Using a model, recurrence Score result was predicted on average to increase quality-adjusted survival by 16.3 years and reduce overall costs by $155,128.

In a 2 million member plan, approximately 773 women are eligible for the test. If half receive the test, given the high and increasing cost of adjuvant chemotherapy, supportive care and management of adverse events, the use of the Oncotype DX assay is estimated to save approximately $1,930 per woman tested (given an aggregate 34% reduction in chemotherapy use).

Molecular signatures on the market

Company	Product name	Disease/phenotype	Purpose
Agendia	***MammaPrint***	Breast cancer	Risk assessment for the recurrence of distant metastasis in a breast cancer patient.
Agendia	***TargetPrint***	Breast cancer	Quantitative determination of the expression level of estrogen receptor, progesteron receptor and HER2 genes. *This product is supplemental to MammaPrint*.
Agendia	***CupPrint***	Cancer	Determination of the origin of the primary tumor.
University Genomics	***Breast Bioclassifier***	Breast cancer	Classification of ER-positive and ER-negative breast cancers into expression-based subtypes that more accurately predict patient outcome.
Clarient	***Insight Dx Breast Cancer Profile*** (formely ***GeneRx Breast Cancer Profile*** by *Prediction Sciences*)	Breast cancer	Prediction of disease recurrence risk.
Clarient	***Prostate Gene Expression Profile***	Prostate cancer	Diagnosis of grade 3 or higher prostate cancer.
Prediction Sciences	***RapidResponse c-Fn Test***	Stroke	Identification of the patients that are safe to receive tPA and those at high risk for HT, to help guide the physician's treatment decision.
Genomic Health	***OncotypeDx***	Breast cancer	Individualized prediction of chemotherapy benefit and 10-year distant recurrence to inform adjuvant treatment decisions in certain women with early-stage breast cancer.
bioTheranostics (previously *AviaraDx*)	***CancerTYPE ID***	Cancer	Classification of 39 types of cancer.
bioTheranostics (previously *AviaraDx*)	***Breast Cancer Index***	Breast cancer	Risk assessment and identification of patients likely to benefit from endocrine therapy, and whose tumors are likely to be sensitive or resistant to chemotherapy.
Applied Genomics	***MammaStrat***	Breast cander	Risk assessment of cancer recurrence.
Applied Genomics	***PulmoType***	Non-small cell lung cancer	Classification of non-small cell lung cancer into adenocarcinoma versus squamous cell carcinoma subtypes.
Applied Genomics	***PulmoStrat***	Lung cancer	Assessment of an individual's risk of lung cancer recurrence following surgery for helping with adjuvant therapy decisions.
Correlogic	***OvaCheck***	Ovarian cancer	Early detection of epithelial ovarian cancer.
LabCorp	***OvaSure***	Ovarian cancer	Assessment of the presence of early stage ovarian cancer in high-risk women.
Veridex	***GeneSearch BLN Assay***	Breast cancer	Determination of whether breast cancer has spread to the lymph nodes.
Power3	***BC-SeraPro***	Breast cancer	Differentiation between breast cancer patients and control subjects.



Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,[1] but to achieve this goal, specific and sensitive molecular markers are essential.[1–5] This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,[1] and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.[1–6] Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

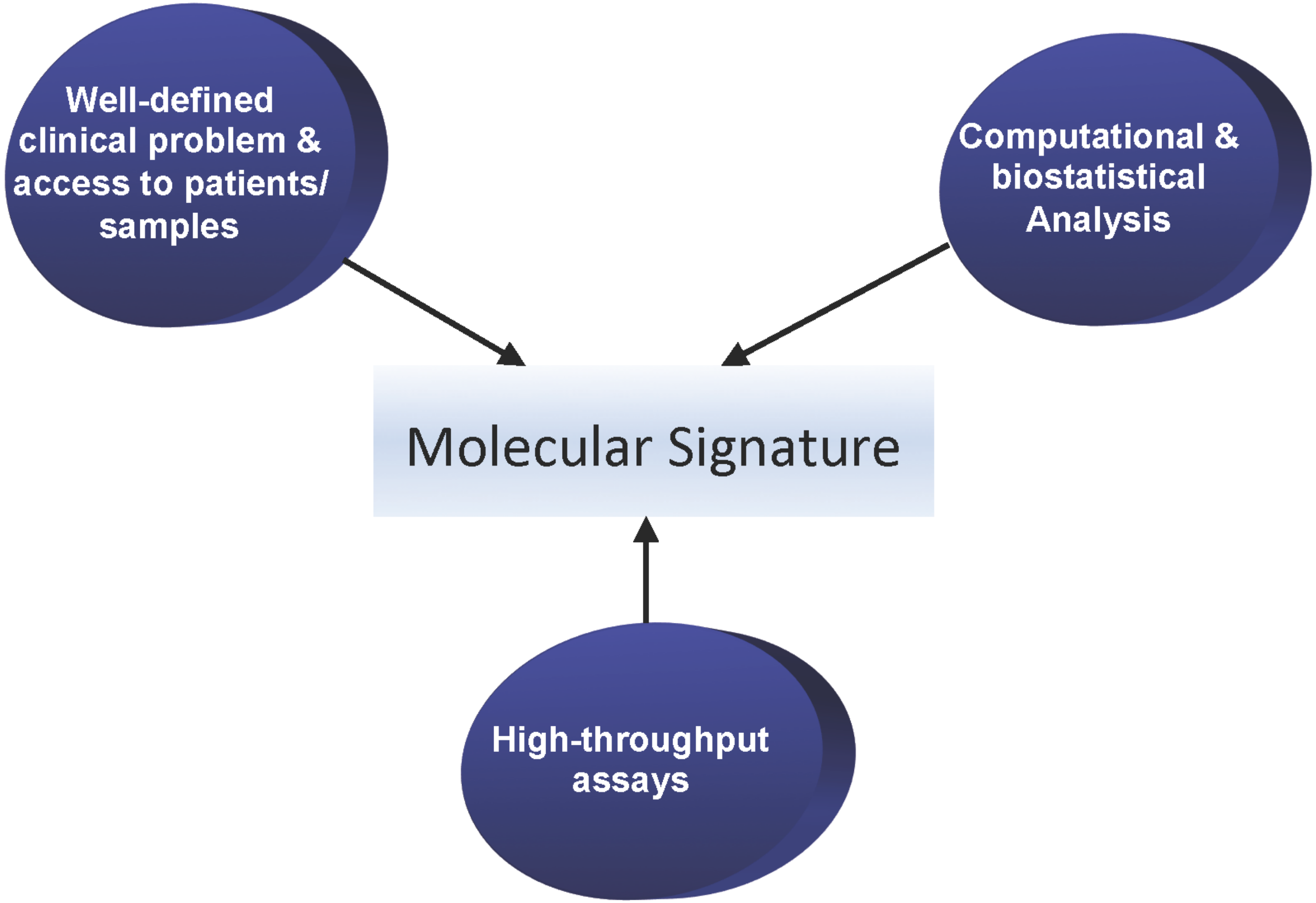
EF Petricoin III, AM Ardekani, BA Hitt, PJ Levine, VA Fusaro, SM Steinberg, GB Mills, C Simone, DA Fishman, EC Kohn, LA Liotta, ***"Use of proteomic patterns in serum to identify ovarian cancer"***, Lancet 359 (2002) 572–77

Example: OvaCheck

- Developed by Correlogic (www.correlogic.com)
- Blood test for the early detection of epithelial ovarian cancer
- **Failed to obtain FDA approval**
- Looks for subtle changes in patterns among the tens of thousands of proteins, protein fragments and metabolites in the blood
- Signature developed by genetic algorithm
- Significant artifacts in data collection & analysis questioned validity of the signature:
  - Results are not reproducible
  - Data collected differently for different groups of patients

http://www.nature.com/nature/journal/v429/n6991/full/429496a.html

Main ingredients for developing a molecular signature



Well-defined clinical problem & access to patients/ samples

Computational & biostatistical Analysis

Molecular Signature

High-throughput assays

Base-Line Characteristics

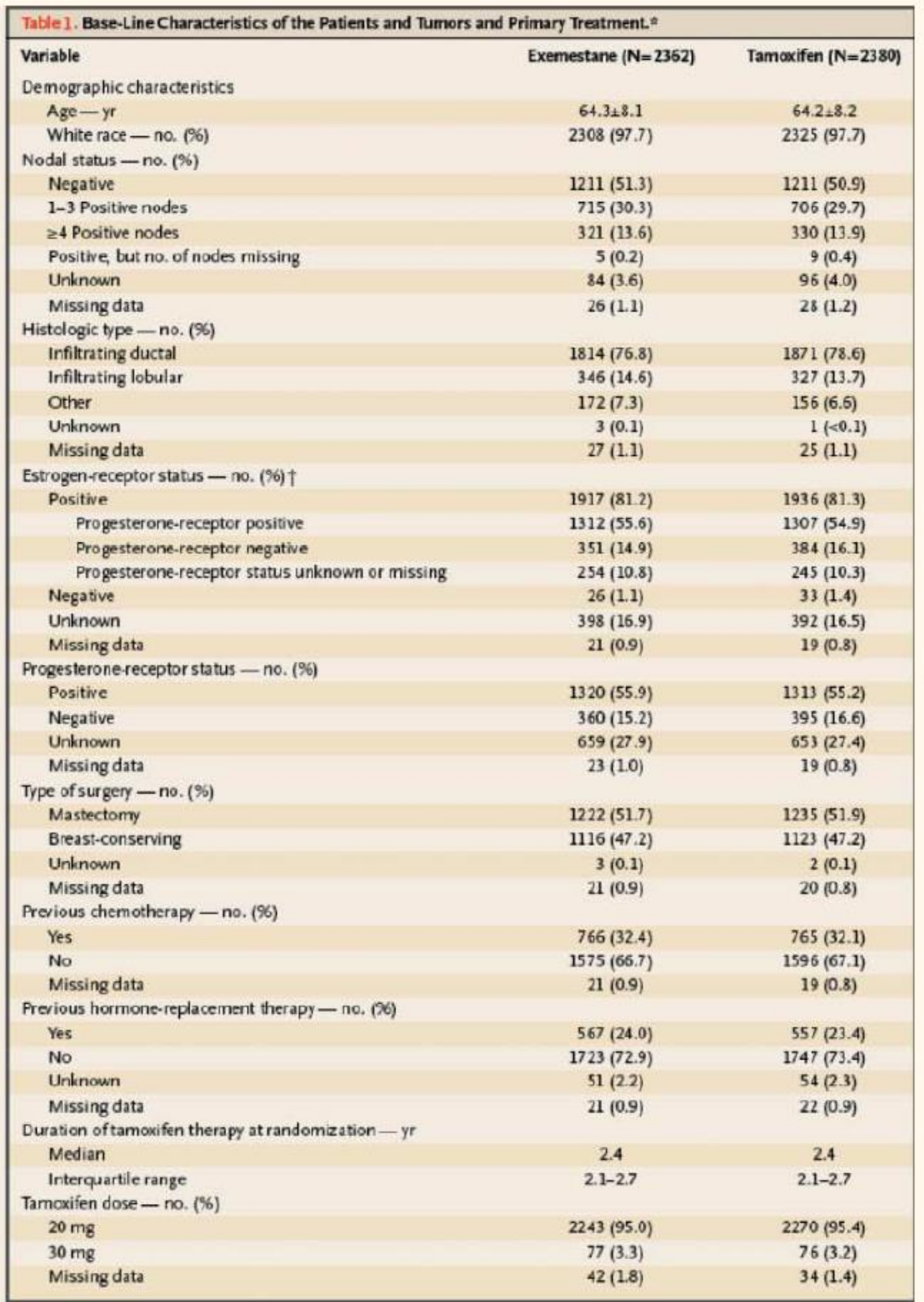


Table 1. Base-Line Characteristics of the Patients and Tumors and Primary Treatment.*

Variable	Exemestane (N=2362)	Tamoxifen (N=2380)
Demographic characteristics		
Age — yr	64.3±8.1	64.2±8.2
White race — no. (%)	2308 (97.7)	2325 (97.7)
Nodal status — no. (%)		
Negative	1211 (51.3)	1211 (50.9)
1–3 Positive nodes	715 (30.3)	706 (29.7)
≥4 Positive nodes	321 (13.6)	330 (13.9)
Positive, but no. of nodes missing	5 (0.2)	9 (0.4)
Unknown	84 (3.6)	96 (4.0)
Missing data	26 (1.1)	28 (1.2)
Histologic type — no. (%)		
Infiltrating ductal	1814 (76.8)	1871 (78.6)
Infiltrating lobular	346 (14.6)	327 (13.7)
Other	172 (7.3)	156 (6.6)
Unknown	3 (0.1)	1 (<0.1)
Missing data	27 (1.1)	25 (1.1)
Estrogen-receptor status — no. (%)†		
Positive	1917 (81.2)	1936 (81.3)
Progesterone-receptor positive	1312 (55.6)	1307 (54.9)
Progesterone-receptor negative	351 (14.9)	384 (16.1)
Progesterone-receptor status unknown or missing	254 (10.8)	245 (10.3)
Negative	26 (1.1)	33 (1.4)
Unknown	398 (16.9)	392 (16.5)
Missing data	21 (0.9)	19 (0.8)
Progesterone-receptor status — no. (%)		
Positive	1320 (55.9)	1313 (55.2)
Negative	360 (15.2)	395 (16.6)
Unknown	659 (27.9)	653 (27.4)
Missing data	23 (1.0)	19 (0.8)
Type of surgery — no. (%)		
Mastectomy	1222 (51.7)	1235 (51.9)
Breast-conserving	1116 (47.2)	1123 (47.2)
Unknown	3 (0.1)	2 (0.1)
Missing data	21 (0.9)	20 (0.8)
Previous chemotherapy — no. (%)		
Yes	766 (32.4)	765 (32.1)
No	1575 (66.7)	1596 (67.1)
Missing data	21 (0.9)	19 (0.8)
Previous hormone-replacement therapy — no. (%)		
Yes	567 (24.0)	557 (23.4)
No	1723 (72.9)	1747 (73.4)
Unknown	51 (2.2)	54 (2.3)
Missing data	21 (0.9)	22 (0.9)
Duration of tamoxifen therapy at randomization — yr		
Median	2.4	2.4
Interquartile range	2.1–2.7	2.1–2.7
Tamoxifen dose — no. (%)		
20 mg	2243 (95.0)	2270 (95.4)
30 mg	77 (3.3)	76 (3.2)
Missing data	42 (1.8)	34 (1.4)

DF Ransohoff, "Bias as a threat to the validity of cancer molecular-marker research", Nat Rev Cancer 5 (2005) 142-9.

How to Address Bias

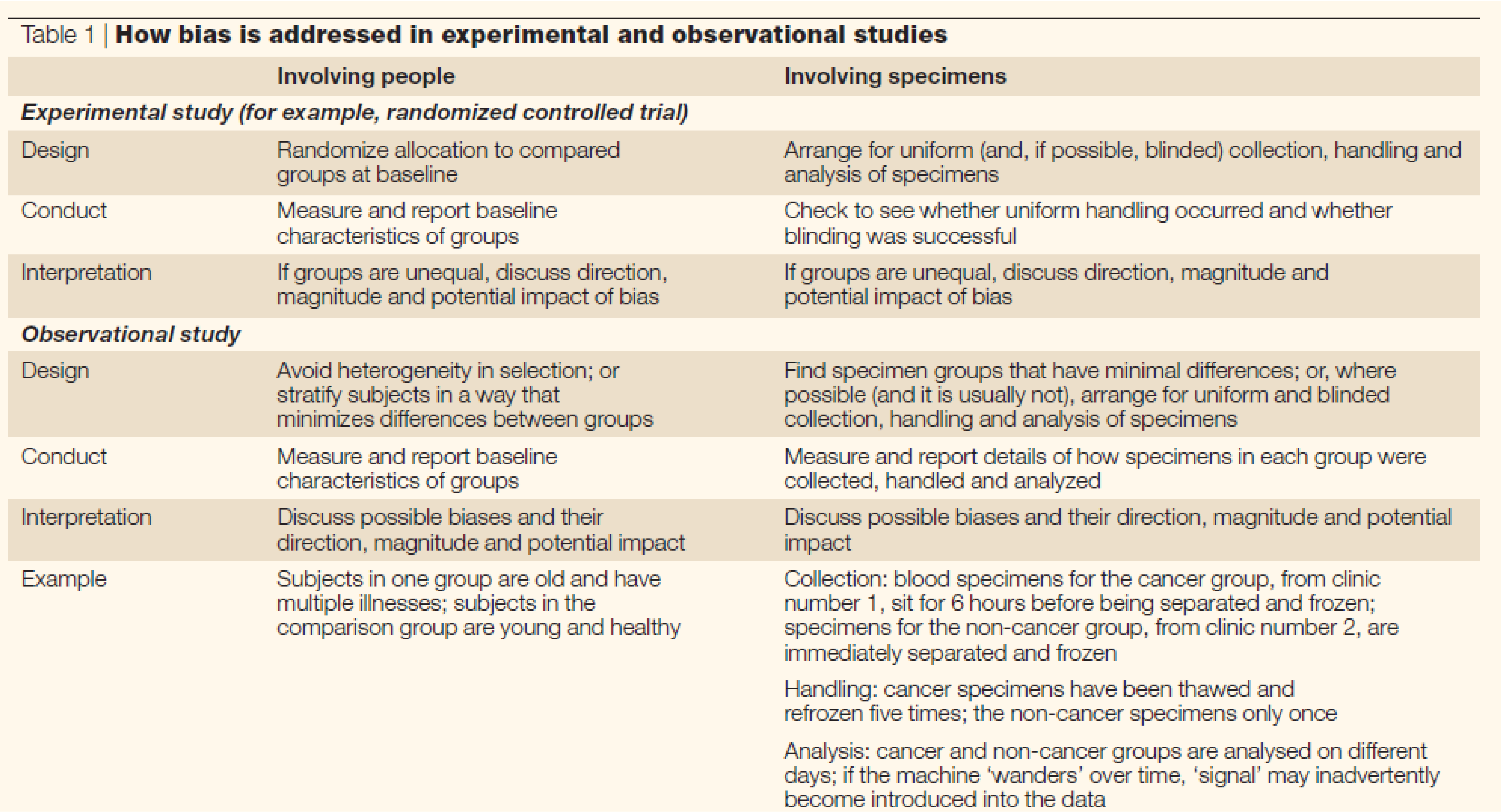
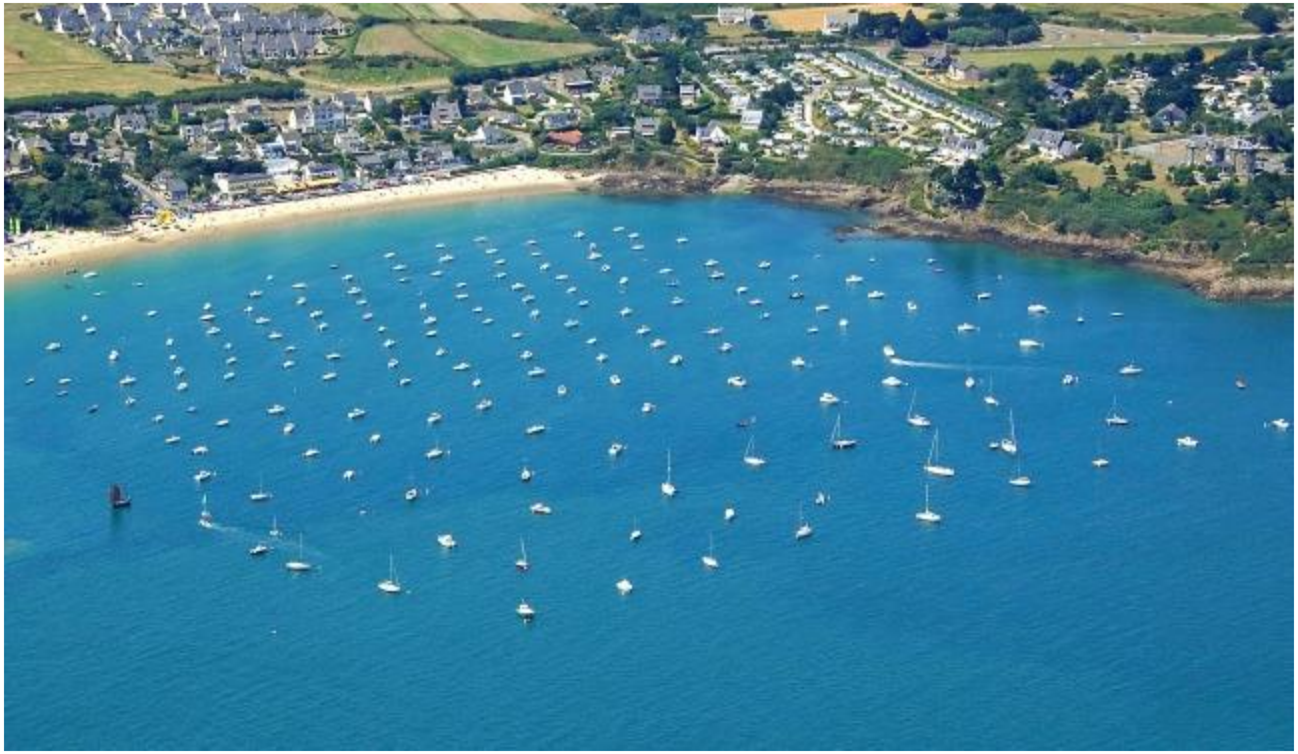


Table 1 | **How bias is addressed in experimental and observational studies**

	Involving people	**Involving specimens**
Experimental study (for example, randomized controlled trial)		
Design	Randomize allocation to compared groups at baseline	Arrange for uniform (and, if possible, blinded) collection, handling and analysis of specimens
Conduct	Measure and report baseline characteristics of groups	Check to see whether uniform handling occurred and whether blinding was successful
Interpretation	If groups are unequal, discuss direction, magnitude and potential impact of bias	If groups are unequal, discuss direction, magnitude and potential impact of bias
Observational study		
Design	Avoid heterogeneity in selection; or stratify subjects in a way that minimizes differences between groups	Find specimen groups that have minimal differences; or, where possible (and it is usually not), arrange for uniform and blinded collection, handling and analysis of specimens
Conduct	Measure and report baseline characteristics of groups	Measure and report details of how specimens in each group were collected, handled and analyzed
Interpretation	Discuss possible biases and their direction, magnitude and potential impact	Discuss possible biases and their direction, magnitude and potential impact
Example	Subjects in one group are old and have multiple illnesses; subjects in the comparison group are young and healthy	Collection: blood specimens for the cancer group, from clinic number 1, sit for 6 hours before being separated and frozen; specimens for the non-cancer group, from clinic number 2, are immediately separated and frozen Handling: cancer specimens have been thawed and refrozen five times; the non-cancer specimens only once Analysis: cancer and non-cancer groups are analysed on different days; if the machine 'wanders' over time, 'signal' may inadvertently become introduced into the data

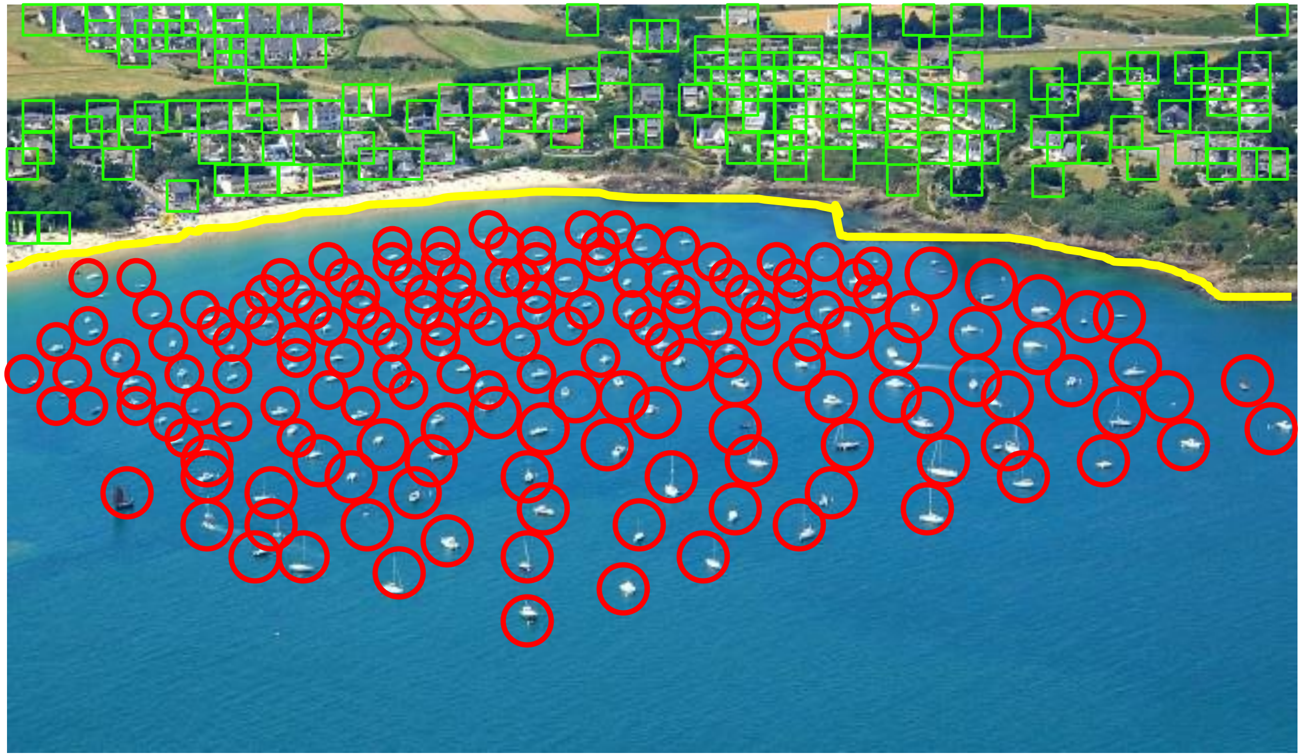
DF Ransohoff, "Bias as a threat to the validity of cancer molecular-marker research", Nat Rev Cancer 5 (2005) 142-9.

Principles and geometric representation for supervised learning



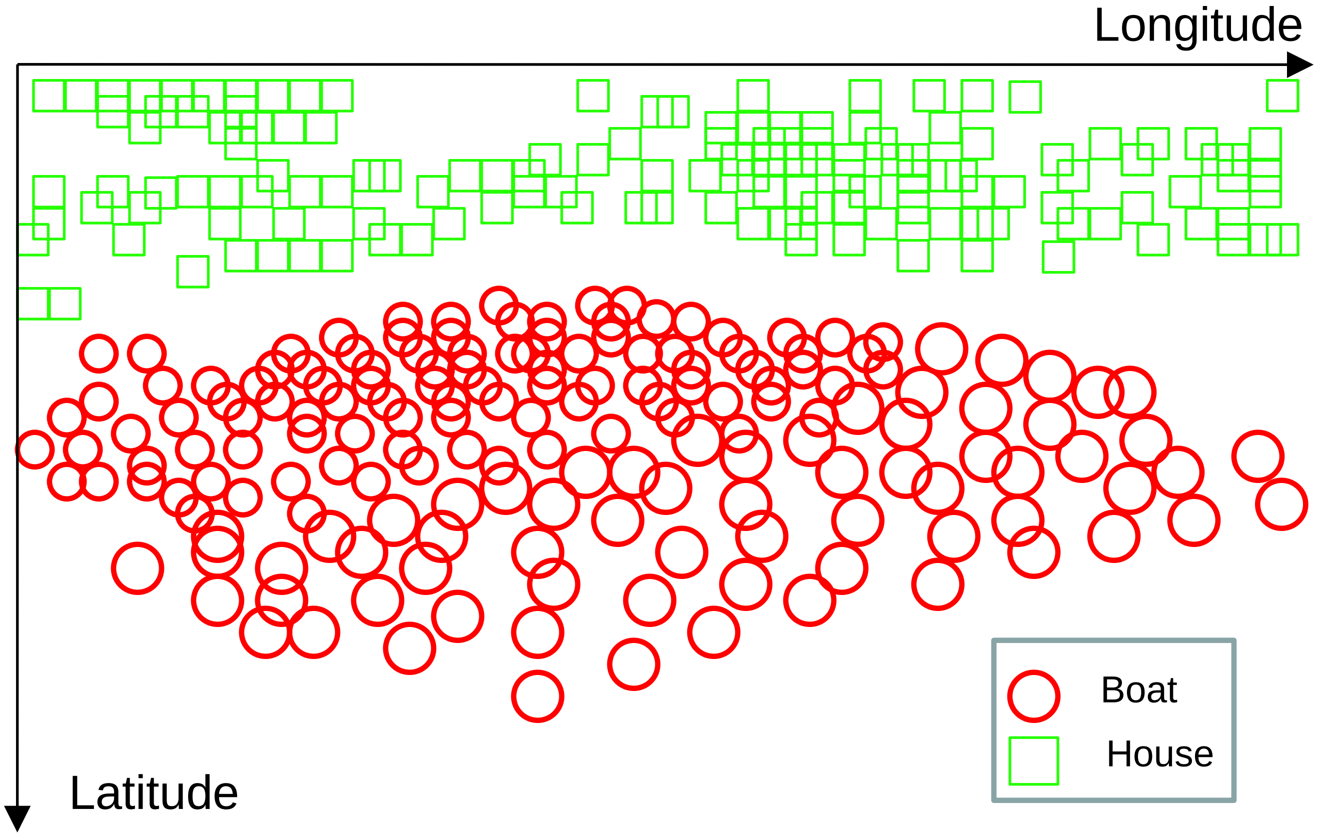
- Want to classify objects as boats and houses.

Principles and geometric representation for supervised learning

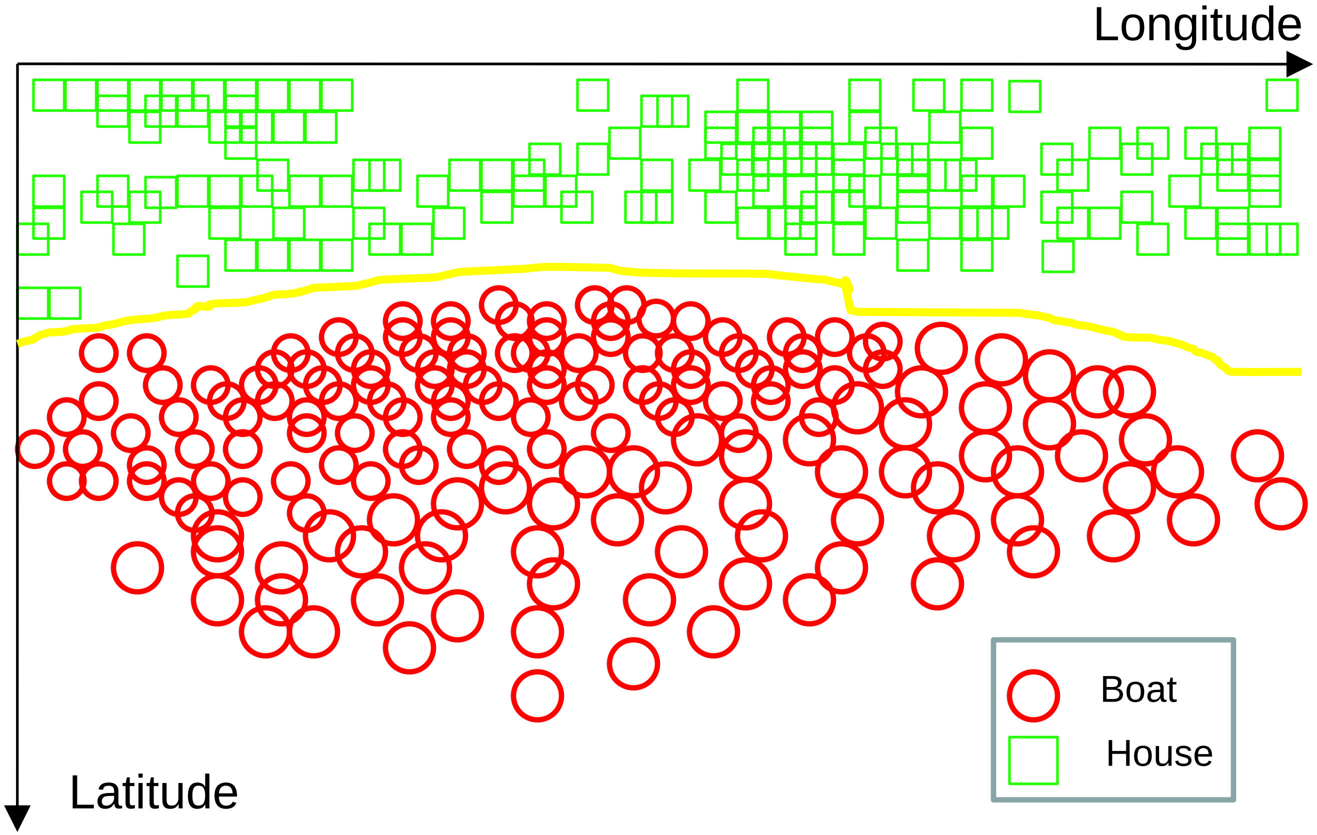


- All objects before the coast line are boats and all objects after the coast line are houses.
- Coast line serves as a decision surface that separates two classes.

Principles and geometric representation for supervised learning

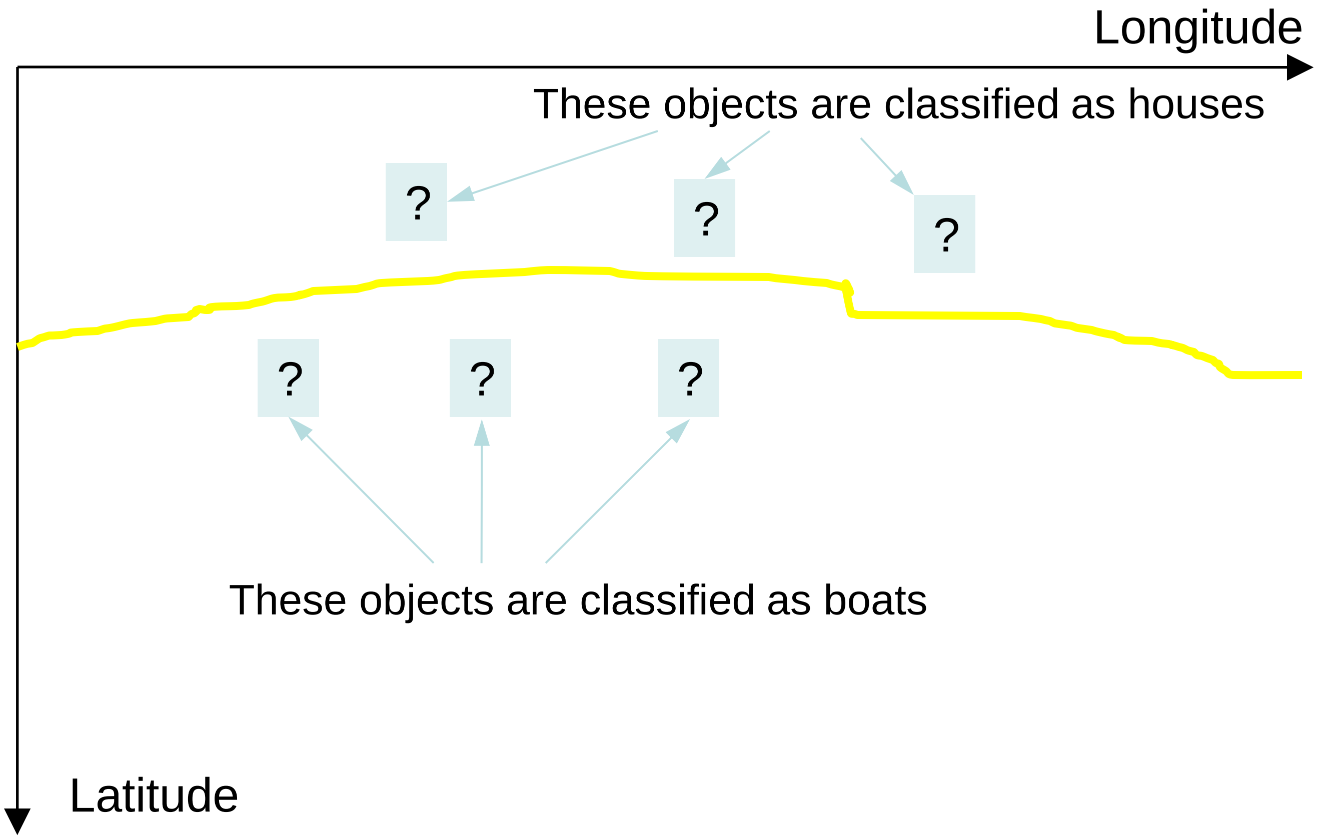


Principles and geometric representation for supervised learning



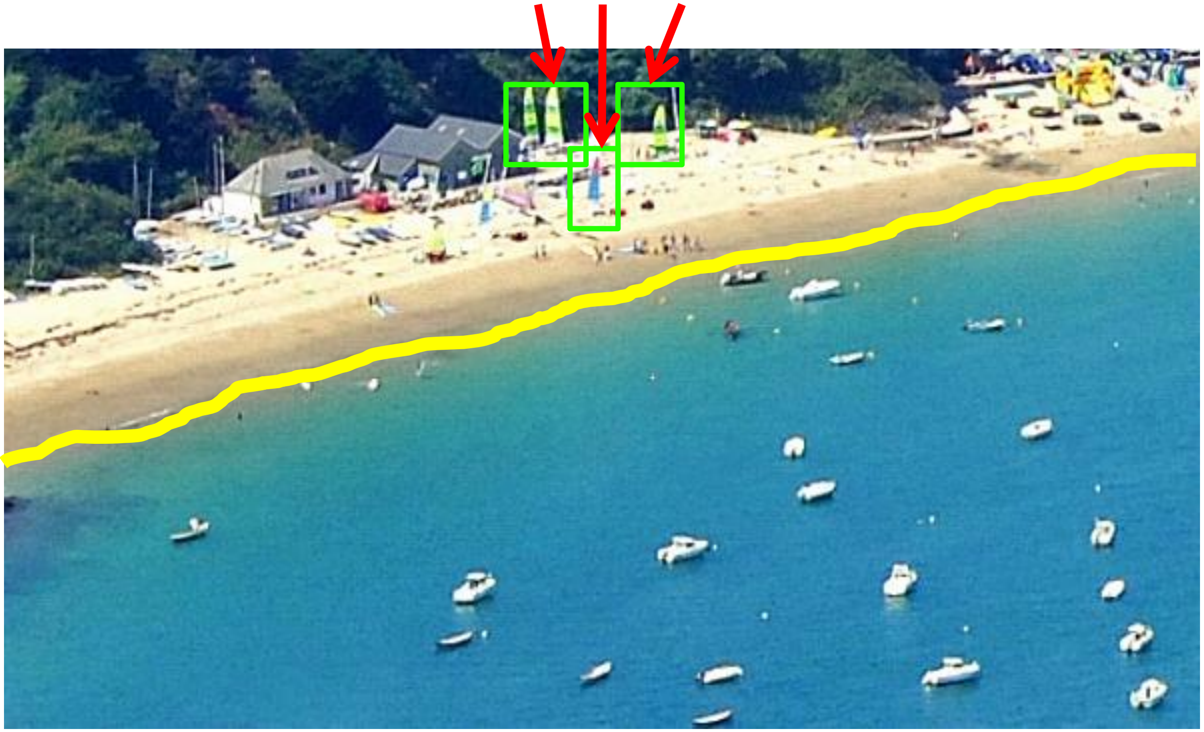
Then the algorithm seeks to find a decision surface that separates classes of objects

Principles and geometric representation for supervised learning

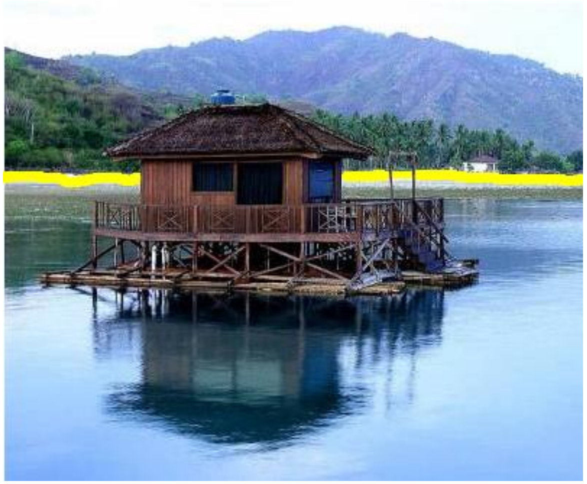

Longitude
These objects are classified as houses
?
?
?
?
?
?
These objects are classified as boats
Latitude

Unseen (new) objects are classified as "boats" if they fall below the decision surface and as "houses" if the fall above it

Principles and geometric representation for supervised learning



These boats will be misclassified as houses



This house will be misclassified as boat

In 2-D this looks simple but what happens in higher dimensional data…

- 10,000-50,000 (gene expression microarrays, aCGH, and early SNP arrays)
- >500,000 (tiled microarrays, SNP arrays)
- 10,000-1,000,000 (MS based proteomics)
- >100,000,000 (next-generation sequencing)

This is the 'curse of dimensionality'

High-dimensionality (especially with small samples) causes:

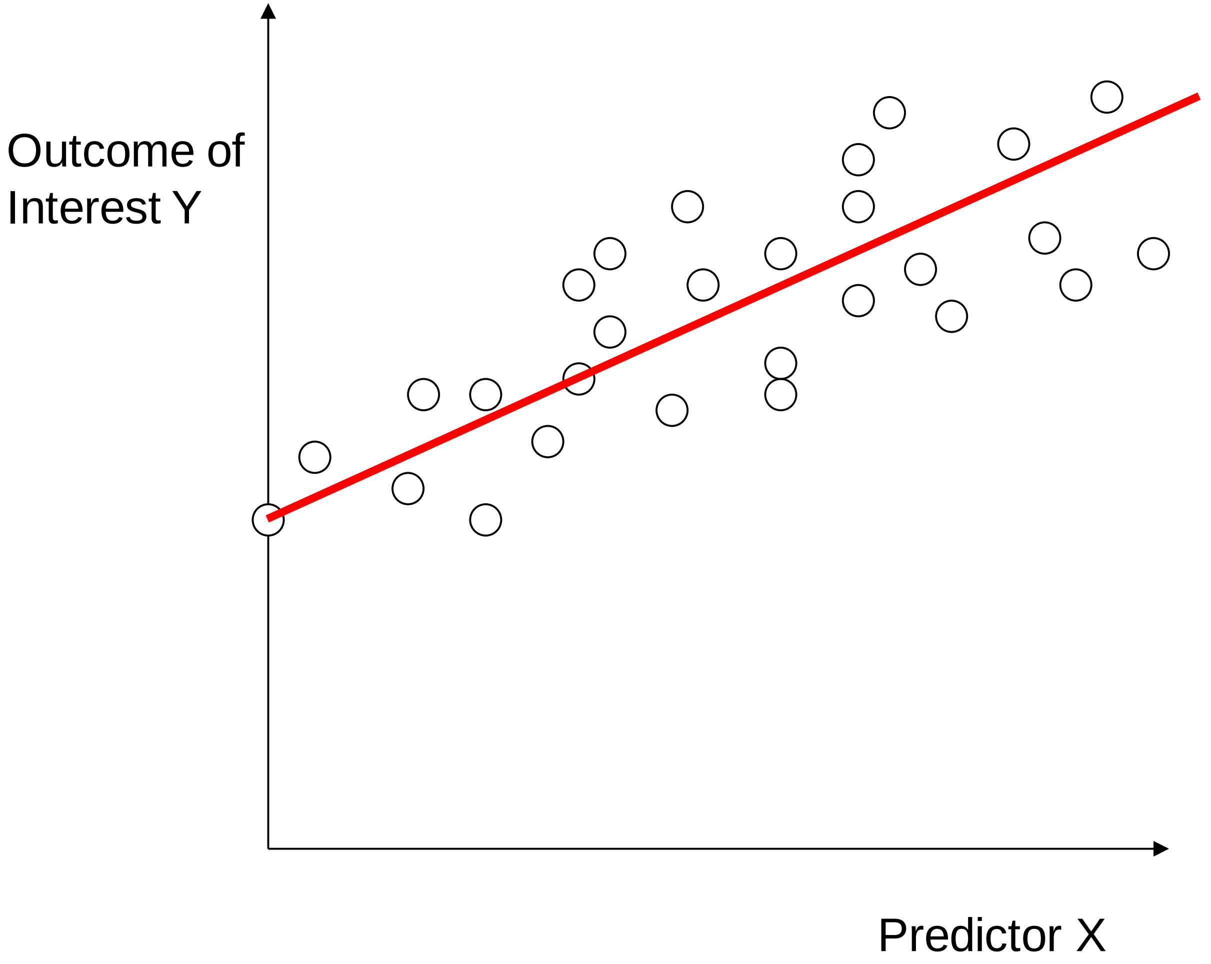
- Some methods do not run at all (classical regression)
- Some methods give bad results (KNN, Decision trees)
- Very slow analysis
- Very expensive/cumbersome clinical application
- Tends to “overfit”

Two problems: Over-fitting & Under-fitting

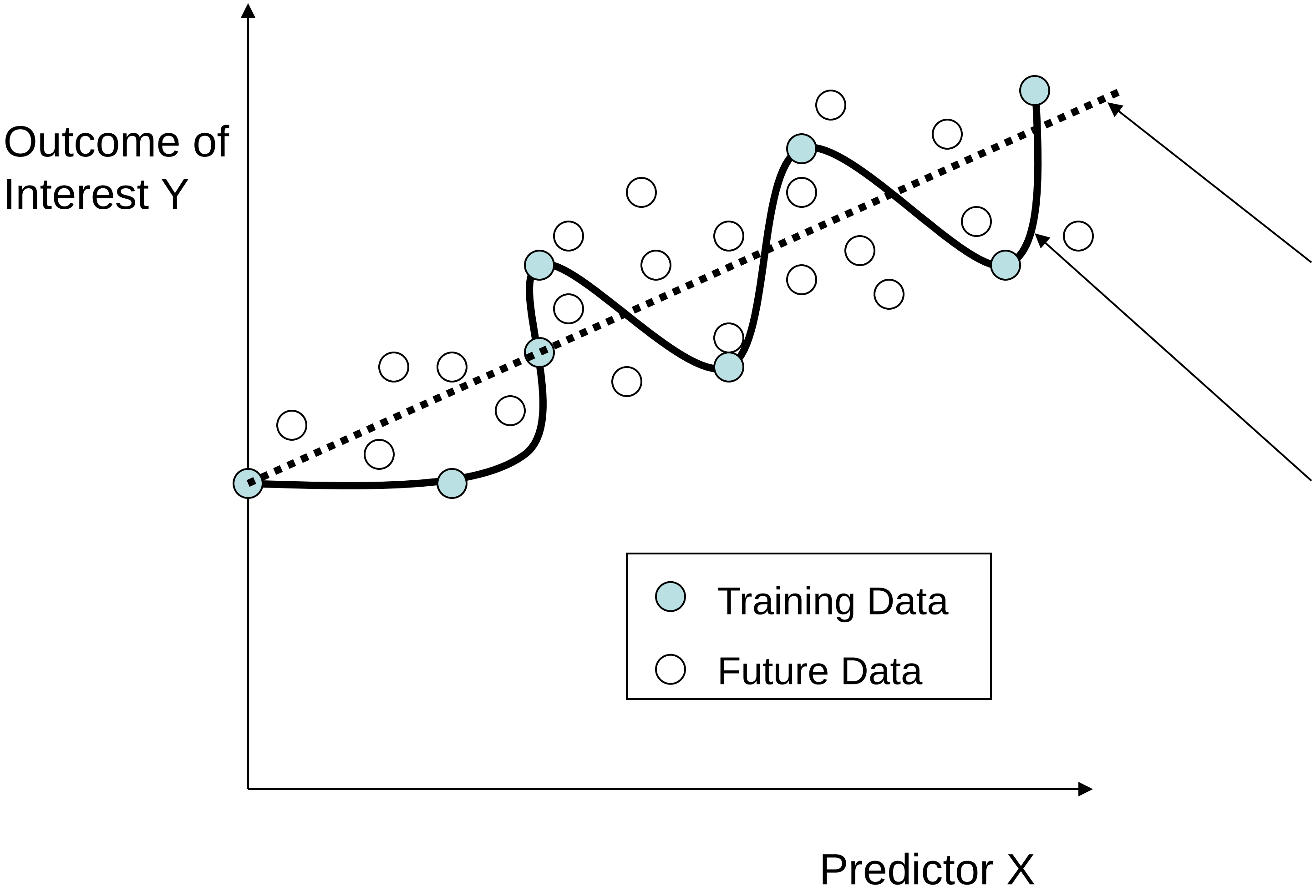
- **Over-fitting** (a model to your data) = building a model that is good in original data but fails to generalize well to new/unseen data.

- **Under-fitting** (a model to your data) = building a model that is poor in both original data and new/unseen data.

Over/under-fitting are related to complexity of the decision surface and how well the training data is fit



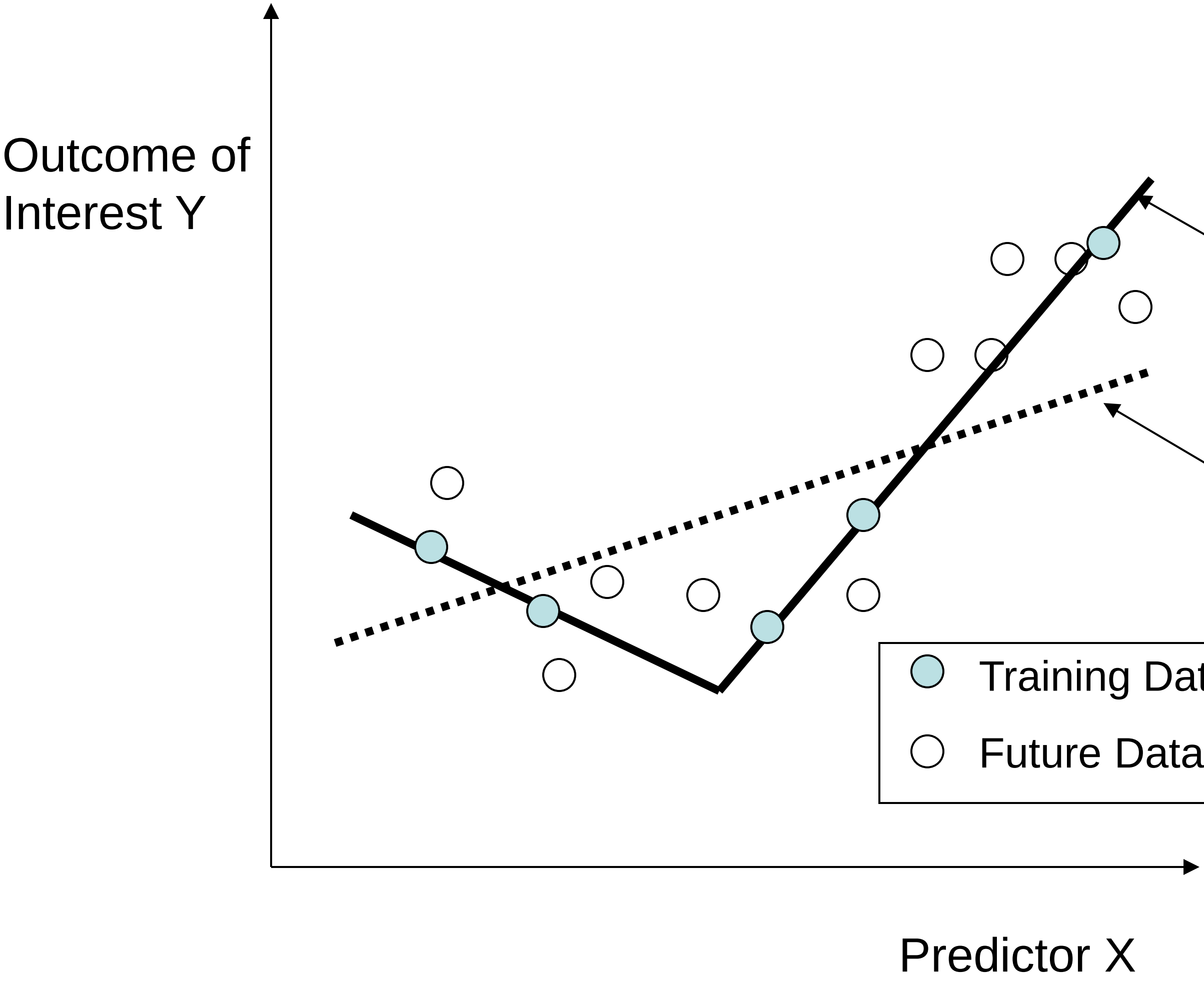
Over/under-fitting are related to complexity of the decision surface and how well the training data is fit



This line is good!

This line overfits!

Over/under-fitting are related to complexity of the decision surface and how well the training data is fit



This line is good!

This line underfits!

Successful data analysis methods balance training data fit with complexity

- Too complex signature (to fit training data well) ➔overfitting (i.e., signature does not generalize)

- Too simplistic signature (to avoid overfitting) ➔ underfitting (will generalize but the fit to both the training and future data will be low and predictive performance small).

Challenges in computational analysis of omics data

Relatively easy to develop a predictive model and even easier to believe that a model is when it is not.

There are both practical and theoretical problems.

Omics data has many special characteristics and it is difficult to analyze.

Example: OvaCheck, a blood test for early detection of epithelial ovarian cance, failed FDA approval.

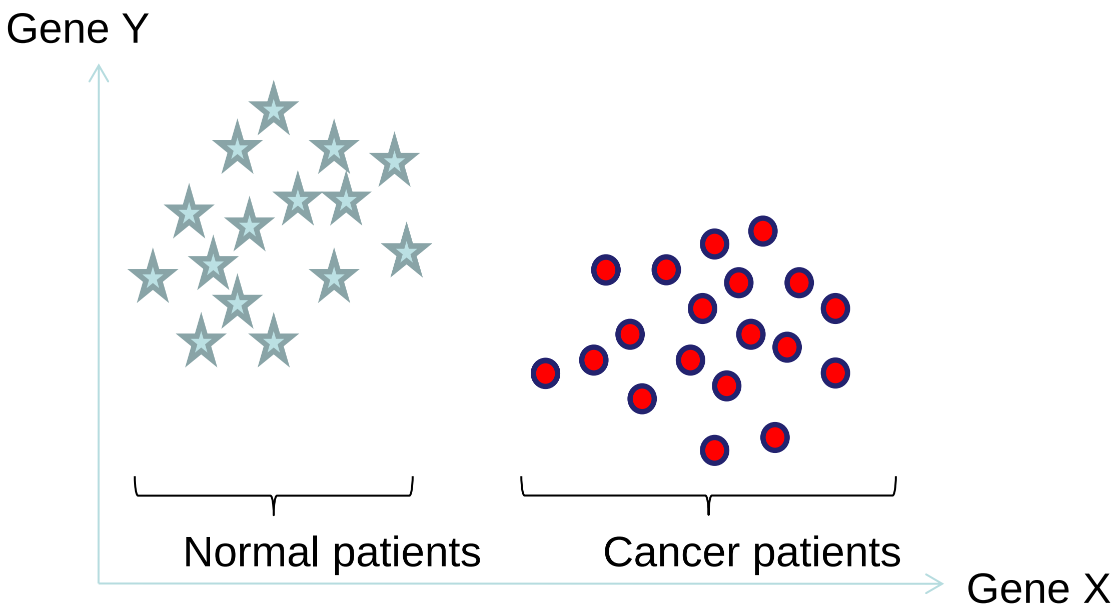
- Looks for subtle changes in patterns of proteins levels
- Signature developed by genetic algorithm
- Data collected differently for the different patient groups

The Support Vector Machine (SVM) approach for building molecular signatures

- Support vector machines (SVMs) is a binary classification algorithm.
- SVMs are important because of (a) theoretical reasons:
  - Robust to very large number of variables and small samples
  - Can learn both simple and highly complex classification models
  - Employ sophisticated mathematical principles to avoid overfitting

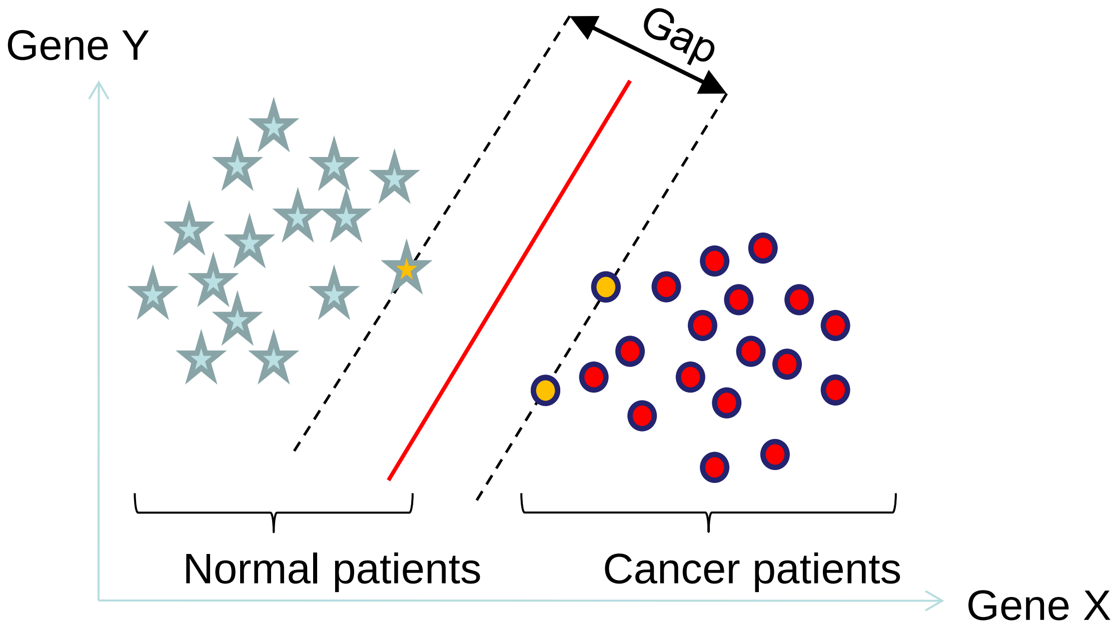
  and (b) superior empirical results.

The Support Vector Machine (SVM) approach for building molecular signatures



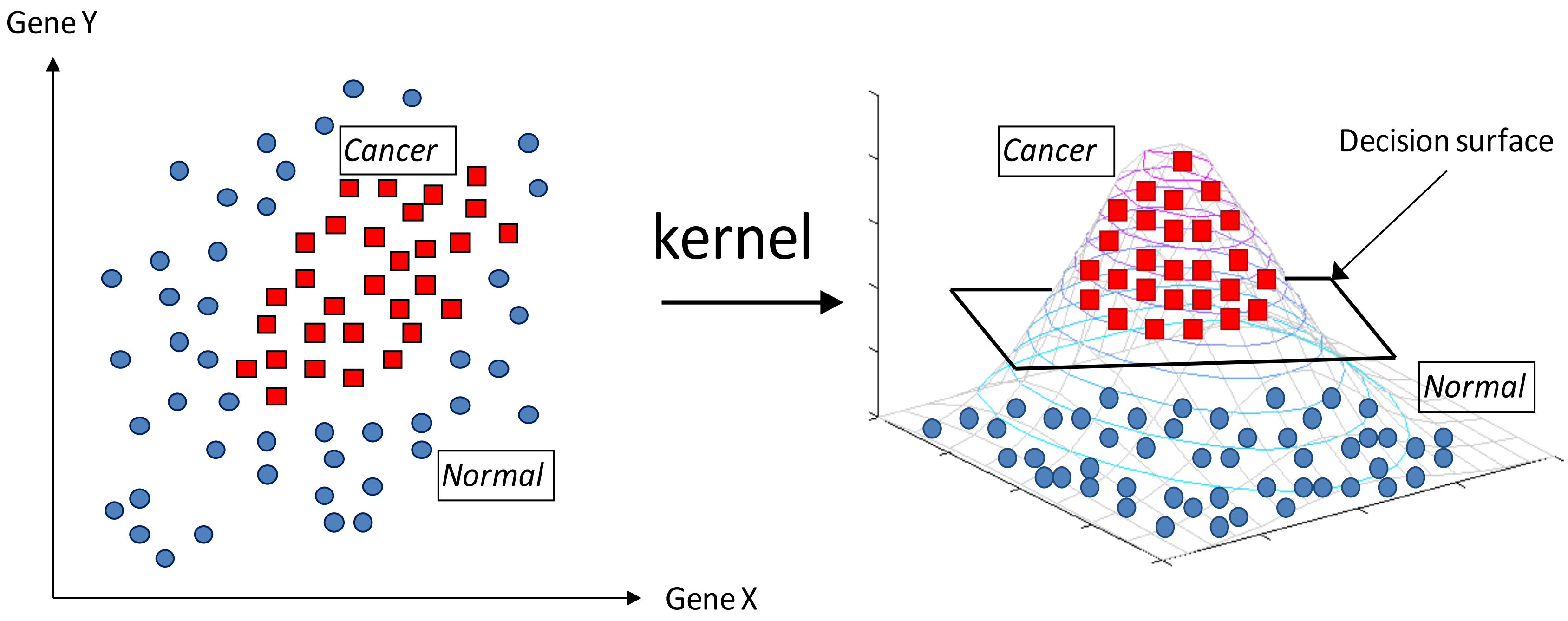
- Consider example dataset described by 2 genes, gene X and gene Y
- Represent patients geometrically (by "vectors")

The Support Vector Machine (SVM) approach for building molecular signatures



Find a linear decision surface ("hyperplane") that can separate patient classes and has the largest distance (i.e., largest "gap" or "margin") between border-line patients (i.e., "support vectors");

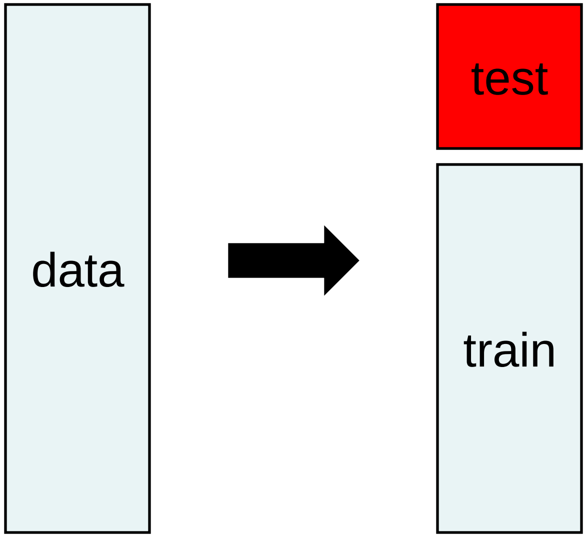
The Support Vector Machine (SVM) approach for building molecular signatures



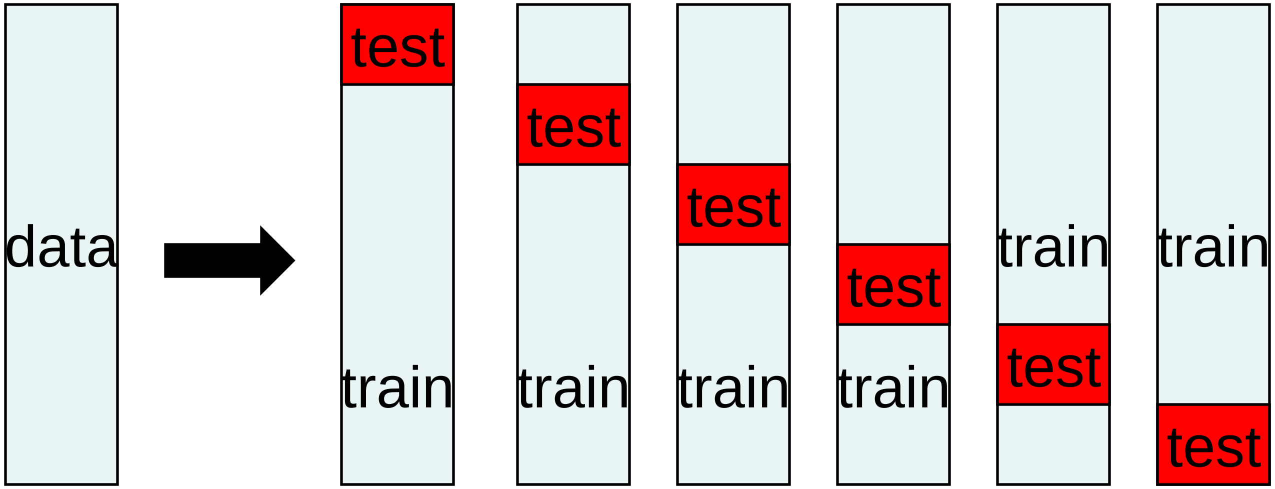
- If such linear decision surface does not exist, the data is mapped into a much higher dimensional space ("feature space") where the separating decision surface is found;
- The feature space is constructed via very clever mathematical projection ("kernel trick").

Estimation of signature accuracy

Large sample case:
use hold-out validation



Small sample case: use **N-fold cross-validation**



Challenges in computational analysis of omics data for development of molecular signatures

- Signature multiplicity (Rashomon effect)
- Poor experimental design
- Is there predictive signal?
- Assay validity/reproducibility
- Efficiency (Statistical and computational)
- Causality vs predictivness
- Methods development (reinventing the wheel)
- Many variables, few samples, noise, artifacts
- Editorialization/Over-simplification/Sensationalism

General conclusions

1. Molecular signatures play a crucial role in personalized medicine and translational bioinformatics.

2. Molecular signatures are being used to treat patients today, not in the future.

3. Development of accurate molecular signature should rely on use of supervised methods.

4. In general, there are many challenges for computational analysis of omics data for development of molecular signatures.

5. One of these challenges is molecular signature multiplicity.

6. There exist an algorithm that can extract the set of maximally predictive and non-redundant molecular signatures from high-throughput data.

Proteomics Informatics – Molecular signatures (Week 11)