Analysis of Mass-Throughput Data: A Technical Commentary from the Perspective of Machine Learning

Constantin F. Aliferis, MD, PhD

Dpt. of Biomedical Informatics Seminar 12-07-05

Main Points…

- Sound data analysis is critical for research that involves mass-throughput data.
- Such data pose a series of non-trivial data analysis problems.
- This talk will discuss some platform-independent problems
- 2 Goals:
  - shed light on the nature and causes of these problems and
  - outline viable methodological approaches to overcome them,

This talk is based on joint work with

Dr. Ioannis Tsamardinos and Alexander Statnikov.

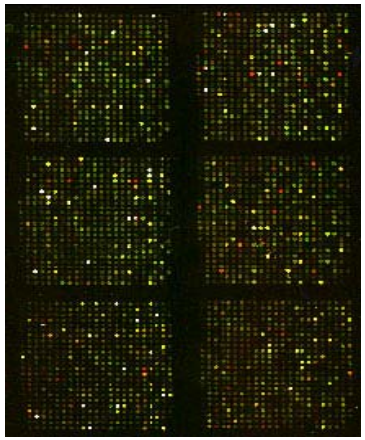
Disclaimers…

1. Not a complete review of literature
2. Heavily influenced by our own research
3. We are interested in proposing viable solutions not only to identify the problems

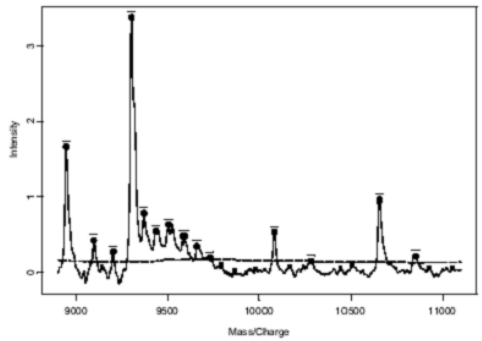
Mass-throughput technologies

- Gene expression microarrays,
- Mass spectrometry,
- SNP arrays,
- Tissue arrays,
- Array genomic hybridization,
- And newer variants such as tiled expression arrays, high-resolution mass spectrometry, liquid chromatography-mass spectrometry and others…

Promise of mass-throughput data



→ Metastatic AdenoCa of the lung



→ Early prostate cancer

Promise of mass-throughput data

- Radically improving medical prevention, diagnosis, development of novel targeted therapeutic agents, personalized treatments, and the ability to predict and monitor the course of the patient
- In the near future such techniques are expected to be combined with novel very-high resolution imaging methods as well as with traditional phenotypic, genetic, and environmental information

Problems

- Organizational, cultural
- Assay validity
- Data analysis, platform-specific
- Data analysis, platform-independent

Organizational & cultural Problems

- How to optimally deliver the results of the respective literature to physicians at the bedside,
- Regulating molecular medicine modalities for safety,
- Explaining complex models to practitioners,
- Exploring proper ways to build and maintain interdisciplinary teams,
- Storing, protecting, and retrieving patient data, etc.

Assay validity problems

- Appropriate protocols
- Dynamic range/coverage
- Marker identification
- Cross-laboratory reproducibility

Platform-specific data analysis problems

- Normalization
- Signal processing
- Feature extraction, etc.

Platform-independent data analysis problems (Our focus)

- Over-fitting & error estimation,
- Curse of dimensionality,
- Causal versus predictive modeling,
- Integration of heterogeneous types of data, and
- Lack of standard protocols for data analysis.

Perspective:
Statistical Machine Learning

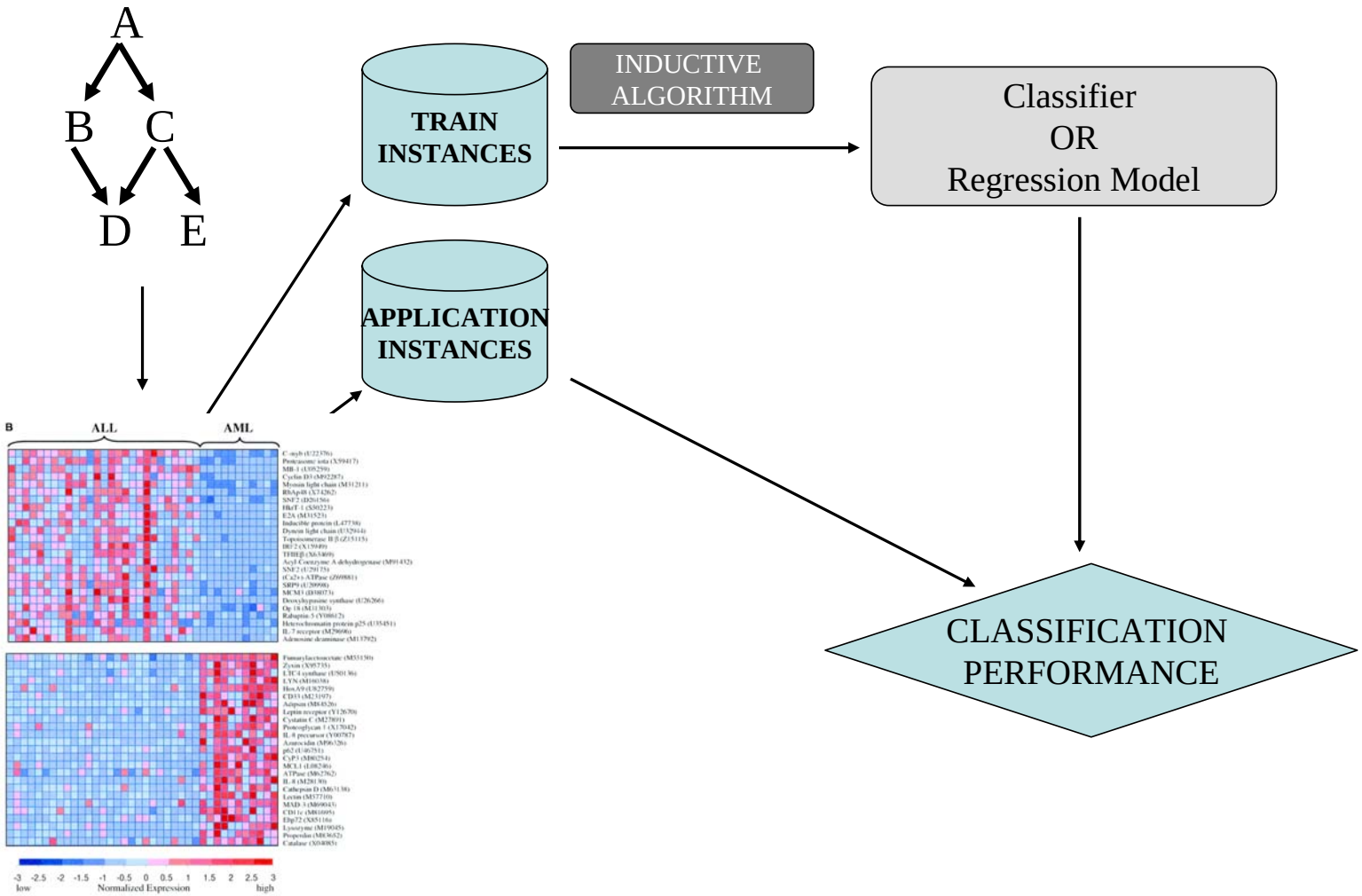
- Machine Learning (informally) studies algorithms and systems that learn from data how to approximate a data distribution (in part or in whole) or a decision function implicit in data, or what are characteristics of the process that generated the data.
- Learners may learn symbolic (e.g., first-order logic) or non-symbolic models of data. (hence: *Statistical Machine Learning)*

Perspective:
Statistical Machine Learning

- Examples of well-established learning methods with applications in biomedicine are:
  - Artificial Neural Networks, Bayesian Networks, restricted Bayesian Classifiers (e.g., simple Bayesian Classifier, Tree-Augmented Networks), Decision Tree Induction, Genetic Algorithms, Clustering methods, Rule Induction, Ensemble Classification (e.g., Voting, Bagging, Boosting, Bayesian model averaging), Support Vector Machines, Hidden Markov Models, Causal Inference methods, and many others

Perspective:
Statistical Machine Learning

- For example, a statistical machine learning diagnostic system may learn how to diagnose patients from the results of microarray profiles of past patients and the correct diagnosis. I.e., it learns how to approximate the decision function that assigns diagnoses to patients profiles.
- This is called *supervised learning*


A
B
C
D
E
TRAIN INSTANCES
INDUCTIVE ALGORITHM
Classifier OR Regression Model
APPLICATION INSTANCES
CLASSIFICATION PERFORMANCE
B
ALL
AML
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3
low
Normalized Expression
high

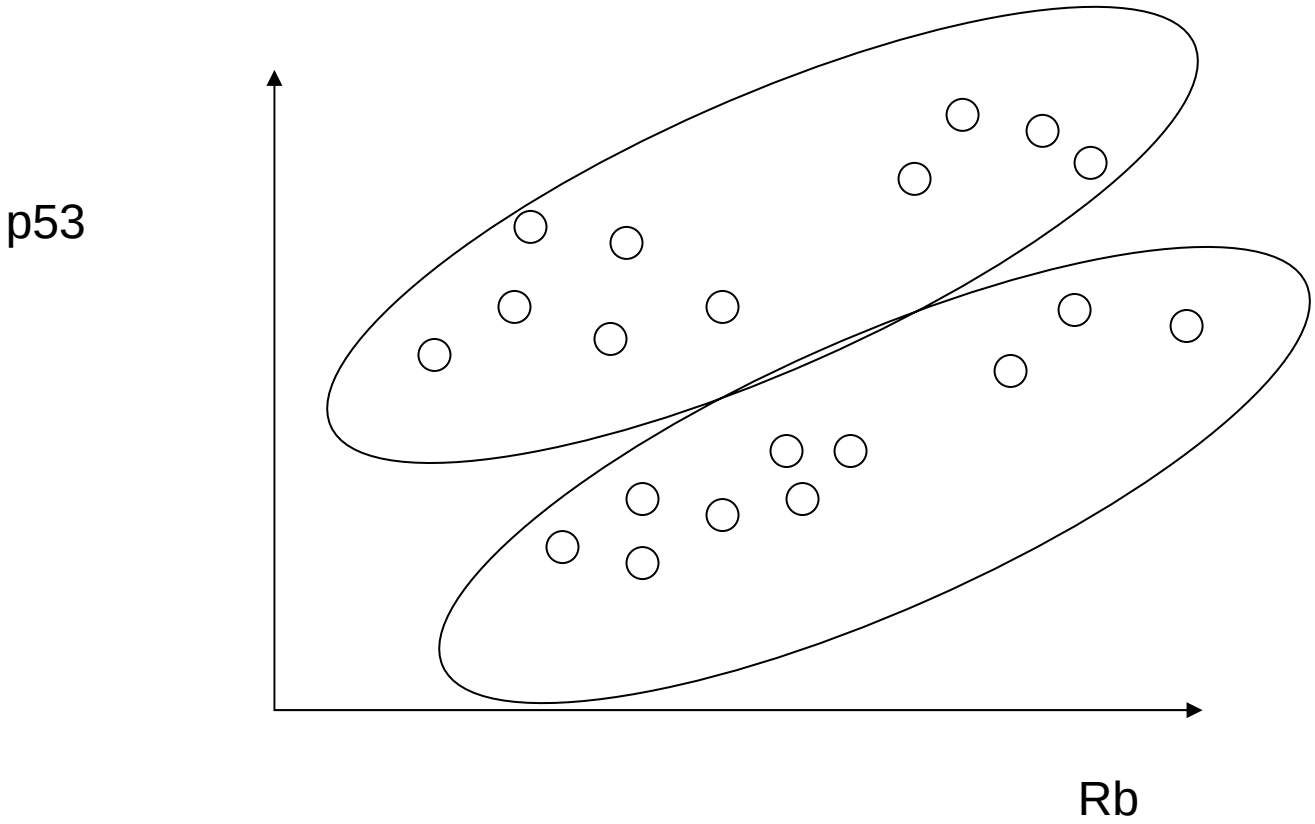
Perspective:
Statistical Machine Learning

- Note that the classifier model is not the learner, but the *output* of the learner.
- Also note that the algorithm learns to predict the correct output that corresponds to some inputs (not only previously seen ones but also previously unseen ones (we call this “generalization”)).

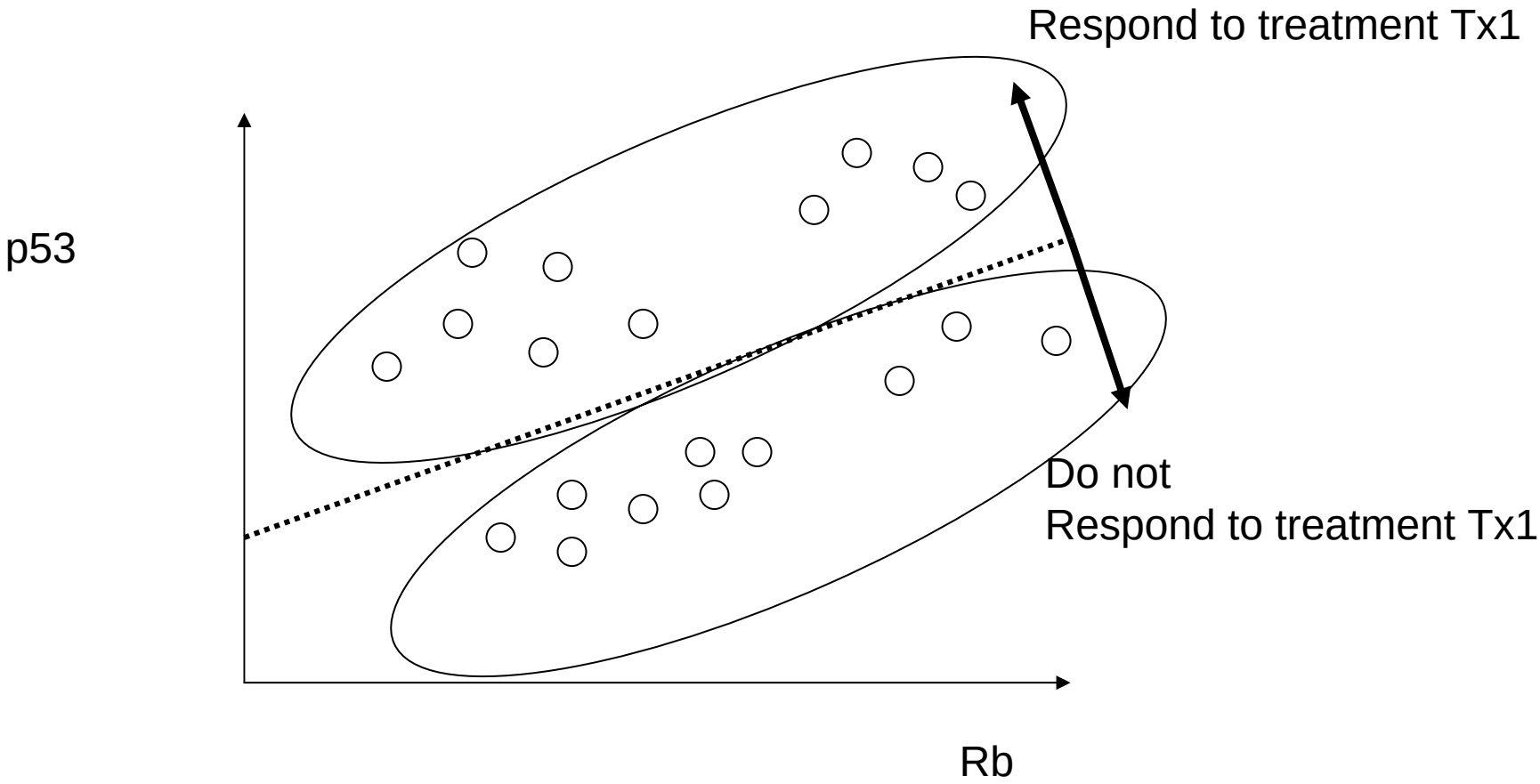
Statistical Machine Learning: Brief Introduction

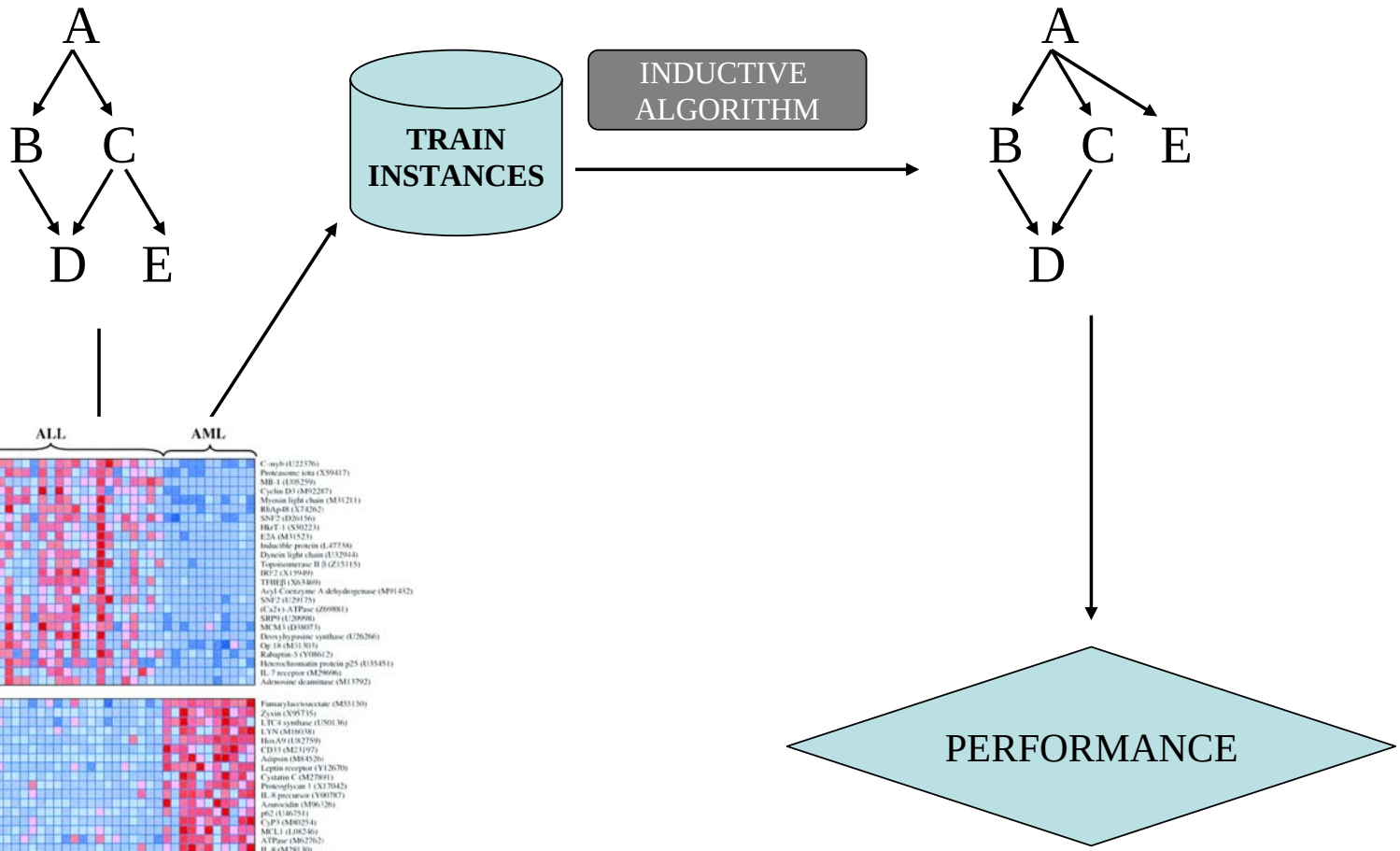
- In unsupervised learning, the algorithm seeks to discover categories or other structural properties of the domain.
- Two main flavors:
  - Learn groupings or subtypes (“clusters”)
  - Learn fine-grain structure

Sub-type learning: seeking ‘natural’ groupings & hoping that they will be useful…



E.g., for treatment




A
B C
D E
TRAIN INSTANCES
INDUCTIVE ALGORITHM
A
B C E
D
PERFORMANCE
B
ALL
AML
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3
low
Normalized Expression
high

Statistical Machine Learning vs Statistics

- Machine Learning > classical Statistics:
  - Learning with continuous feedback from the environment in autonomous robotic agents (reinforcement learning),
  - learning from relational databases (relational learning), or
  - learning interesting patterns or structure with no pre-specified outcome of interest (concept formation and causal discovery).
- Statistics > Machine Learning:
  - Constructing specialized research designs that minimize sample or maximize statistical power, or
  - Estimating the confidence limits of quantities of interest in specific sampling contexts.

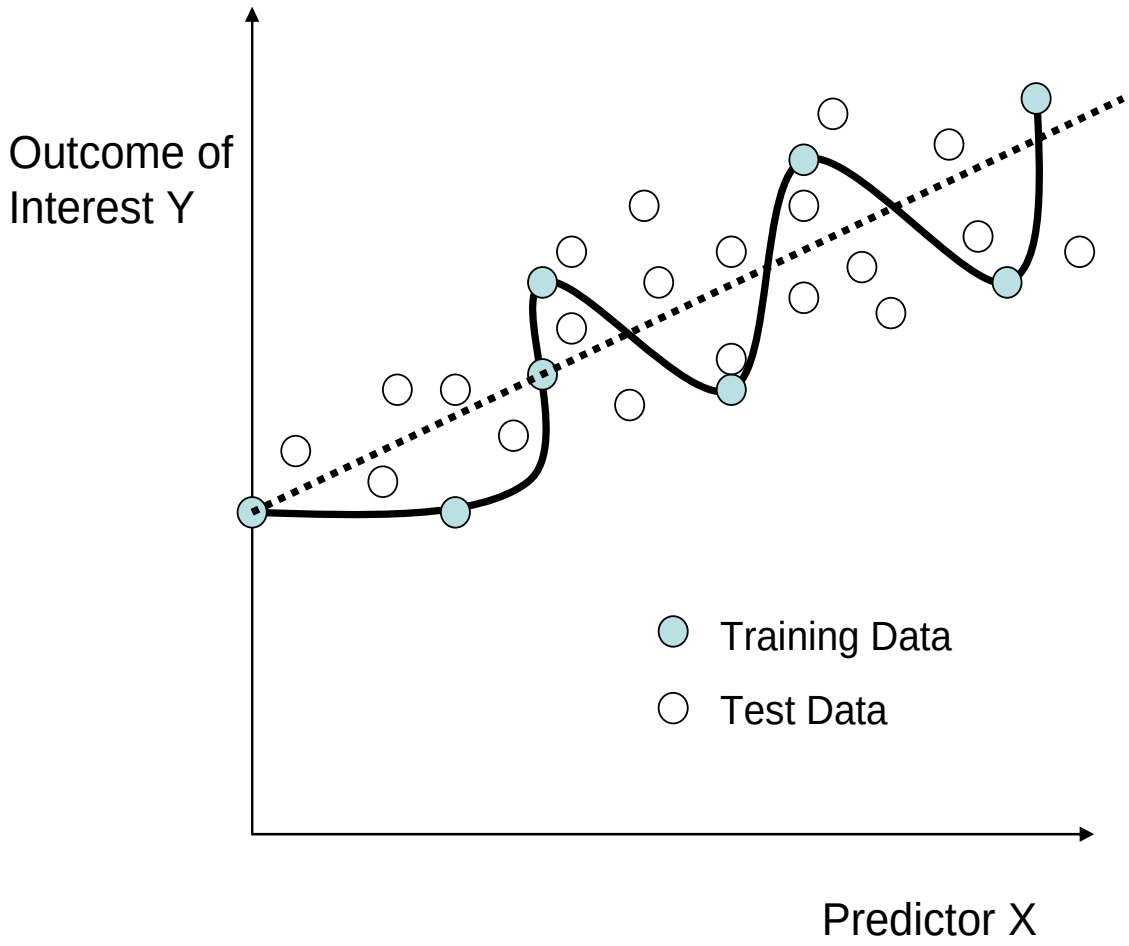
Statistical Machine Learning vs Statistics

- Statistics ~ Machine Learning:
  - Deriving classification models using a generative model approach (the statistical approach, as embodied in logistic regression), or finding a highly predictive decision surface (the machine learning approach, as embodied in Support Vector Machines).
- In addition, numerous Machine Learning methods build upon statistical theory and techniques.
- Thus Machine Learning and Statistical approaches are quite often synergistic for attacking hard modeling and discovery data analysis problems.

Problem #1:
Over-fitting & estimation of a model's generalization error

What is Over-fitting?

- The phenomenon in which a classification or a regression model is exhibiting small prediction error in the training data but much larger generalization error in unseen future data.

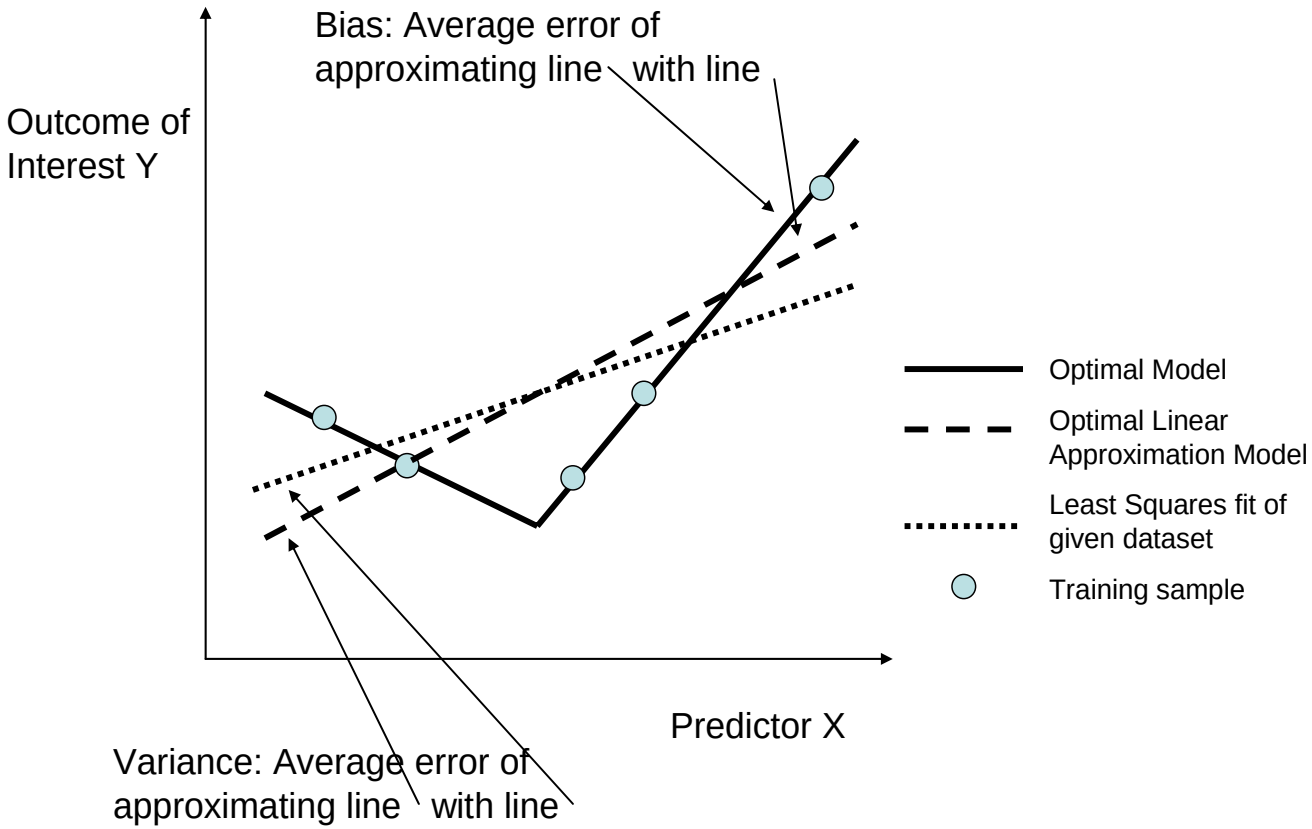

Outcome of
Interest Y
Training Data
Test Data
Predictor X

What Over-fitting *Is Not*

- Some authors loosely define over-fitting as synonymous or equivalent to having too many free parameters, or “over-specifying” the model space by having many more predictors than samples.
- However these factors are *merely facilitating factors*, and strictly speaking they are neither necessary nor sufficient for over-fitting, as we will show below.

What causes over-fitting?

- The Bias-Variance Error Decomposition Analysis View



What causes over-fitting?

- The Computational Learning Theory perspective:

  Computational Learning Theory (COLT) formally studies under which conditions learning is feasible and provides several bounds for the generalization error depending on the classifier used, the definition of error to be minimized (e.g., number of misclassification), and other assumptions.

COLT Error Bounds

- The VC (Vapnik-Chervonenkis) dimension is a measure of decision function complexity
- VC dimension is (informally) defined as the maximum number of training examples that can be correctly classified by a learner for any possible assignment of class labels.
- The VC dimension frequently appears in error bounds such that higher VC dimension leads to increased generalization error.
- An example of VC bound follows: if VC dimension *h* is smaller than *l*, then with probability of at least 1-η, the generalization error of a learner will be bounded by the sum of its empirical error and a confidence term defined as

$$\sqrt{\frac{h\left(\log\frac{2l}{h}+1\right)-\log(\eta/4)}{l}}$$

- Notice that this bound is independent of dimensionality of the problem

Complexity vs Error

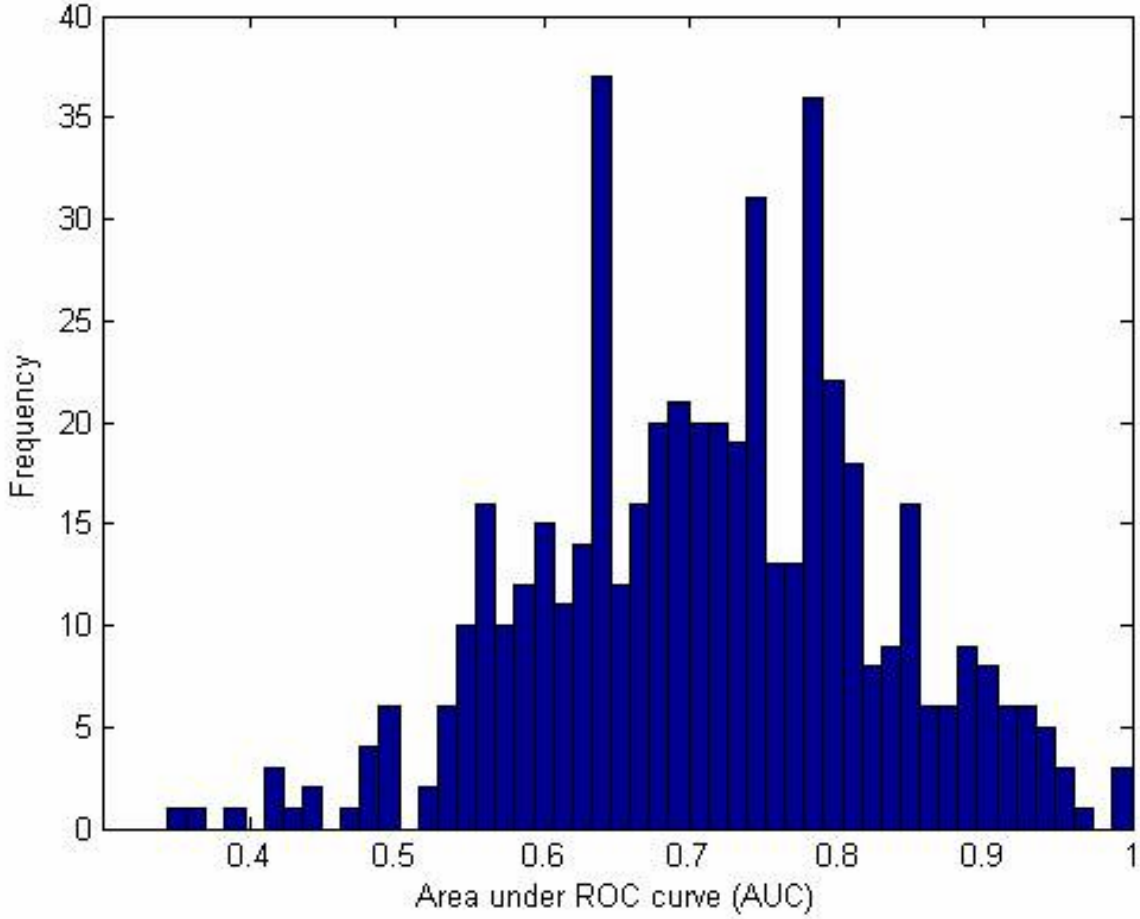
- The number of parameters of a classifier *does not necessary correspond to its complexity!*
  - There exist classifiers with a single parameter and with infinite VC dimension (i.e., high complexity)
  - There exist classifiers with an unbounded number of parameters but with VC dimension of 1 (i.e., low complexity).
- Thus, a classifier with a large number of parameters can still have low complexity which implies low error estimates (i.e., does not over-fit).
- Some of the theoretical error bounds are small even when the number of dimensions is much higher than the number of training cases. Some are independent of number of dimensions.

Complexity vs Error

- These COLT results prove that *learning is possible in the situation common in mass-throughput data where the number of observed variables is much higher than the number of available training sample*.
- The mentioned COLT results also justify our prior assertion that *over-fitting is not equivalent to a high number of parameters*.

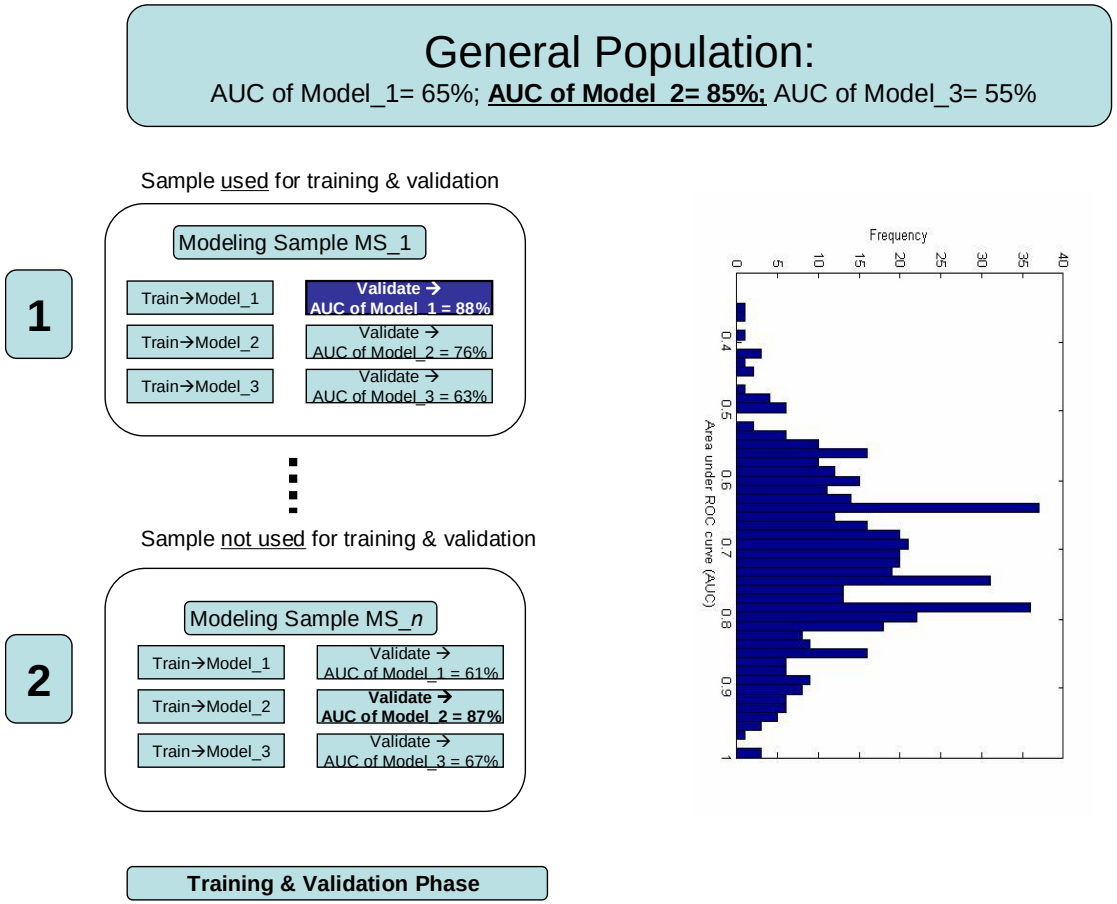
Multiple validation as cause of over-fitting

A Fundamental Observation: In Small Samples Empirical Error Estimates Have High Variance



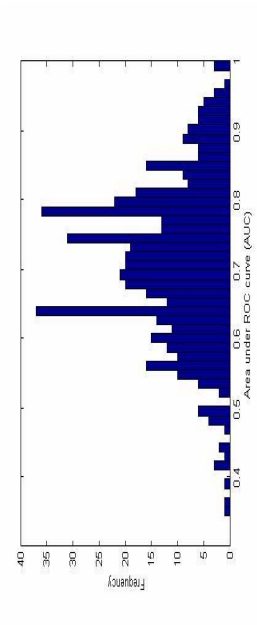
What causes over-fitting?

Multiple Validation (in small samples with high variance error estimation)



What causes over-fitting?

Can independent prospective validation help?



General Population:
AUC of Model_1= 65%; **AUC of Model_2= 85%;** AUC of Model_3= 55%

Sample used for training & validation

1

Modeling Sample MS_1

Train→Model_1	**Validate → AUC of Model_1 = 88%**
Train→Model_2	Validate → AUC of Model_2 = 76%
Train→Model_3	Validate → AUC of Model_3 = 63%

Sample not used for training & validation

2

Modeling Sample MS_*n*

Train→Model_1	Validate → AUC of Model_1 = 61%
Train→Model_2	**Validate → AUC of Model_2 = 87%**
Train→Model_3	Validate → AUC of Model_3 = 67%

Training & Validation Phase

A sample in which over-fitting is detected

3

Independent Evaluation Sample ES_1

Evaluate → AUC of Model_1 = 65%

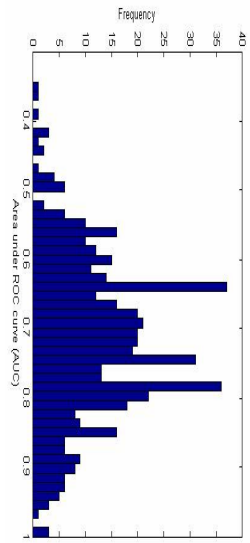
A sample in which over-fitting is not detected

4

Independent Evaluation Sample ES_2

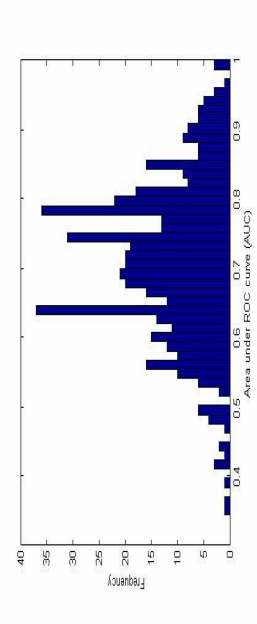
Evaluate → AUC of Model_1 =84%

Validation With Independent Dataset



What causes over-fitting?

Can independent prospective validation help?



General Population:
AUC of Model_1= 65%; **AUC of Model_2= 85%;** AUC of Model_3= 55%

Sample not used for training & validation

1

Modeling Sample MS_1

Train→Model_1	**Validate → AUC of Model_1 = 88%**
Train→Model_2	Validate → AUC of Model_2 = 76%
Train→Model_3	Validate → AUC of Model_3 = 63%

Sample used for training & validation

2

Modeling Sample MS_*n*

Train→Model_1	Validate → AUC of Model_1 = 61%
Train→Model_2	**Validate → AUC of Model_2 = 87%**
Train→Model_3	Validate → AUC of Model_3 = 67%

Training & Validation Phase

A sample falsely detecting over-fitting

3

Independent Evaluation Sample ES_1

Evaluate → AUC of Model_2 = 74%

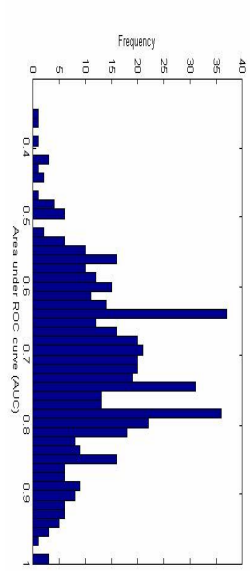
A sample not detecting over-fitting

4

Independent Evaluation Sample ES_2

Evaluate → AUC of Model_2 =90%

Validation With Independent Dataset



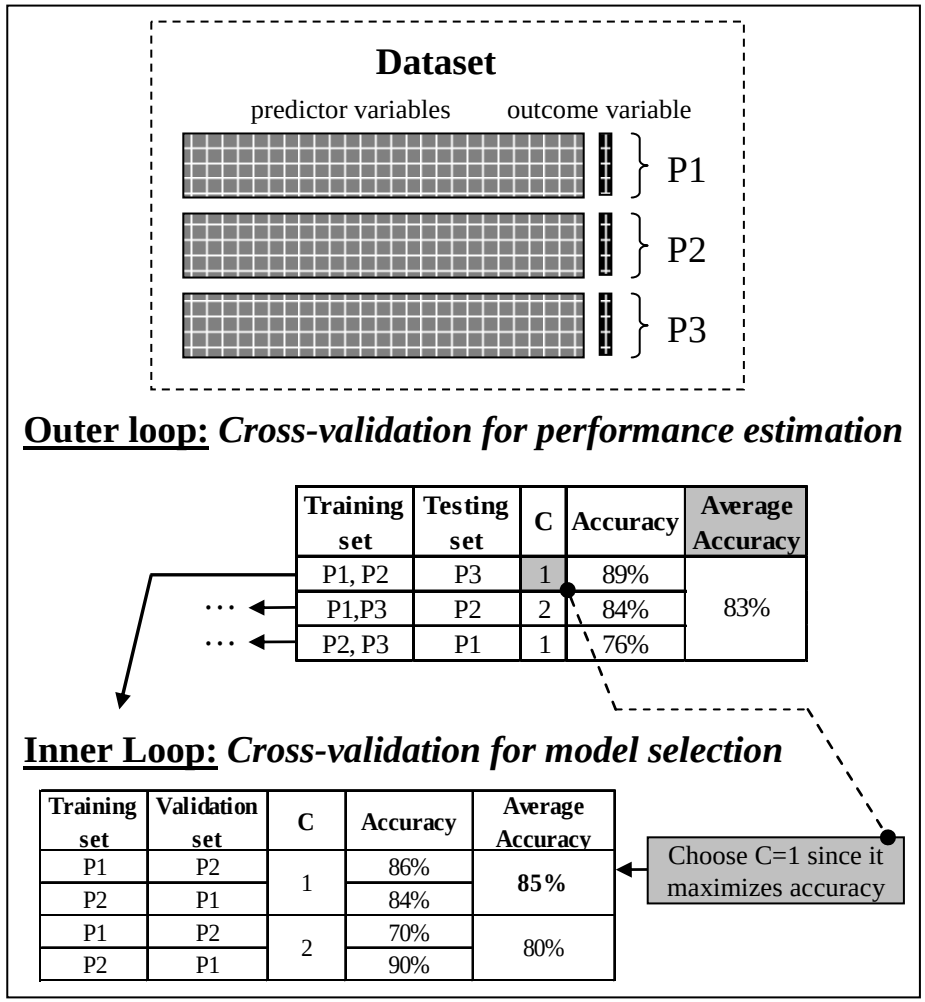
Independent testing is not a panacea

- The two previous sets of scenarios taken together show that validation by independent prospective testing is neither sufficient nor necessary for either preventing or detecting over-fitting when training and/or testing sample size is small.

Methods to prevent or detect over-fitting and to estimate error accurately

Prevent or detect over-fitting and estimate error accurately

- Lower-variance cross-validation schemes → estimate error and detect overfitting

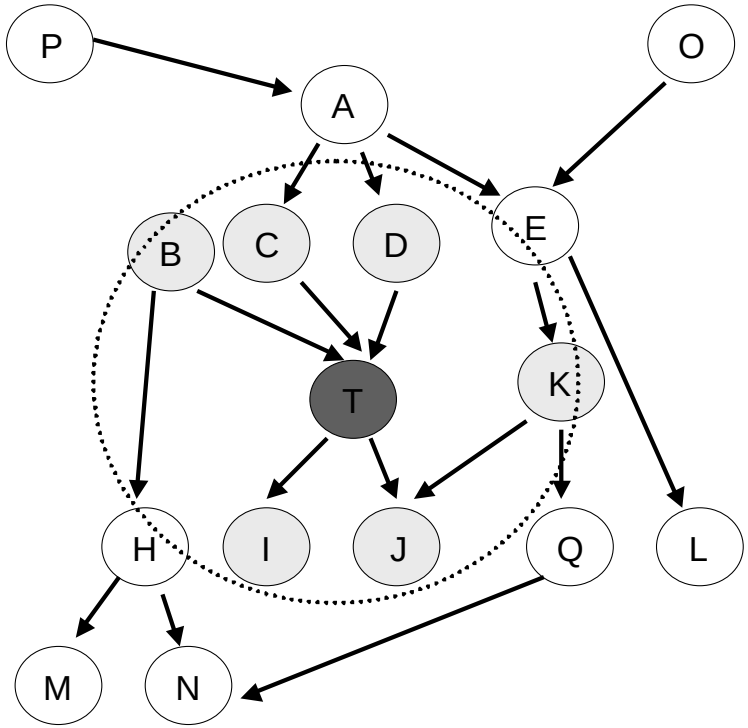


Prevent or detect over-fitting and estimate error accurately

- Theoretical bounds on error → estimate error

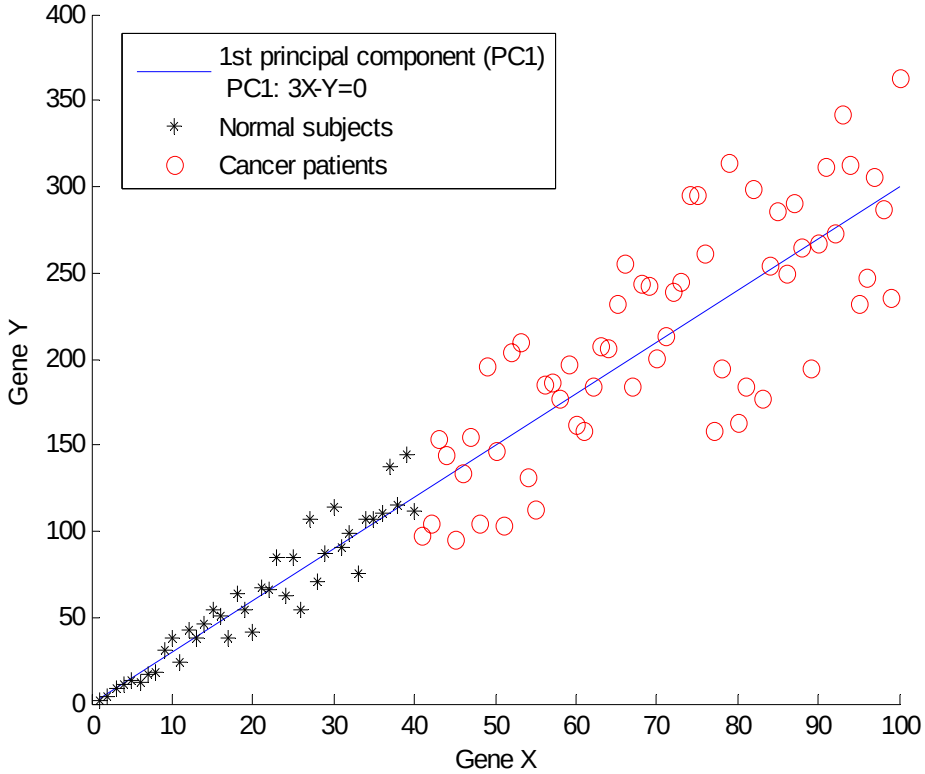
Prevent or detect over-fitting and estimate error accurately

- Feature Selection → prevent over-fitting by reducing complexity



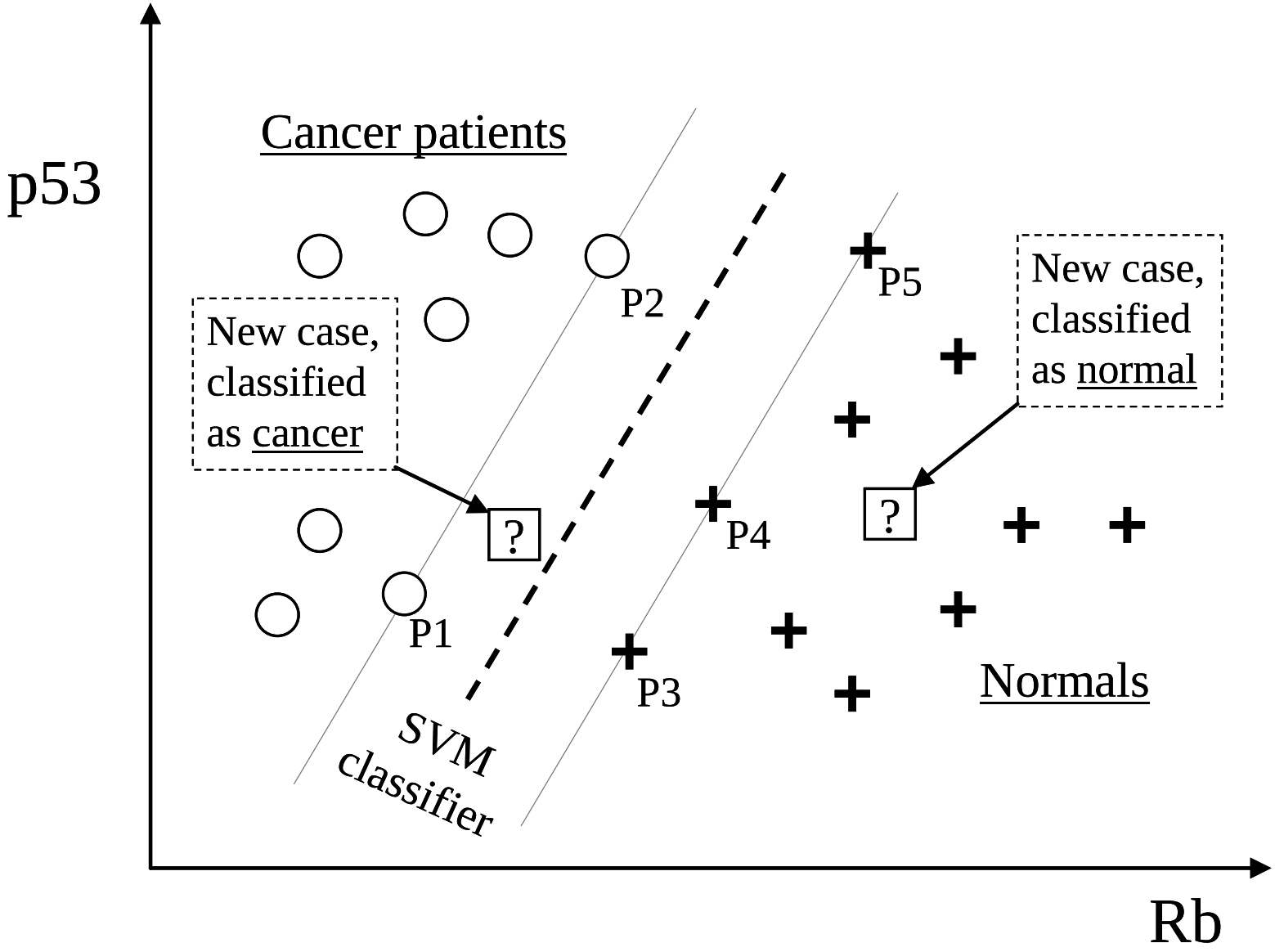
Prevent or detect over-fitting and estimate error accurately

- Dimensionality reduction → prevent over-fitting by reducing complexity



Prevent or detect over-fitting and estimate error accurately

- Prevent over-fitting by penalizing complexity/parameters
- Support Vector Machines [Vapnik, 1998] as exemplars of learners that penalize complexity. The term used for penalization of complexity in the literature for such methods is *regularization*.


p53
Cancer patients
New case, classified as cancer
New case, classified as normal
P1
P2
P3
P4
P5
?
?
SVM classifier
Normals
Rb

Original SVM formulation

- *n* inequality constraints
- *n positivity constraints*
- *n* number of ξ variables

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$s.t.\ \ y_i(w \cdot \Phi(x) + b) \geq 1 - \xi_i, \forall x_i$$

$$\xi_i \geq 0$$

The (Wolfe) dual of this problem
- one equality constraint
- *n* positivity constraints
- *n* number of α variables (Lagrange multipliers)
- Objective function more complicated

NOTICE: Data only appear as $\Phi(xi) \cdot \Phi(xj)$

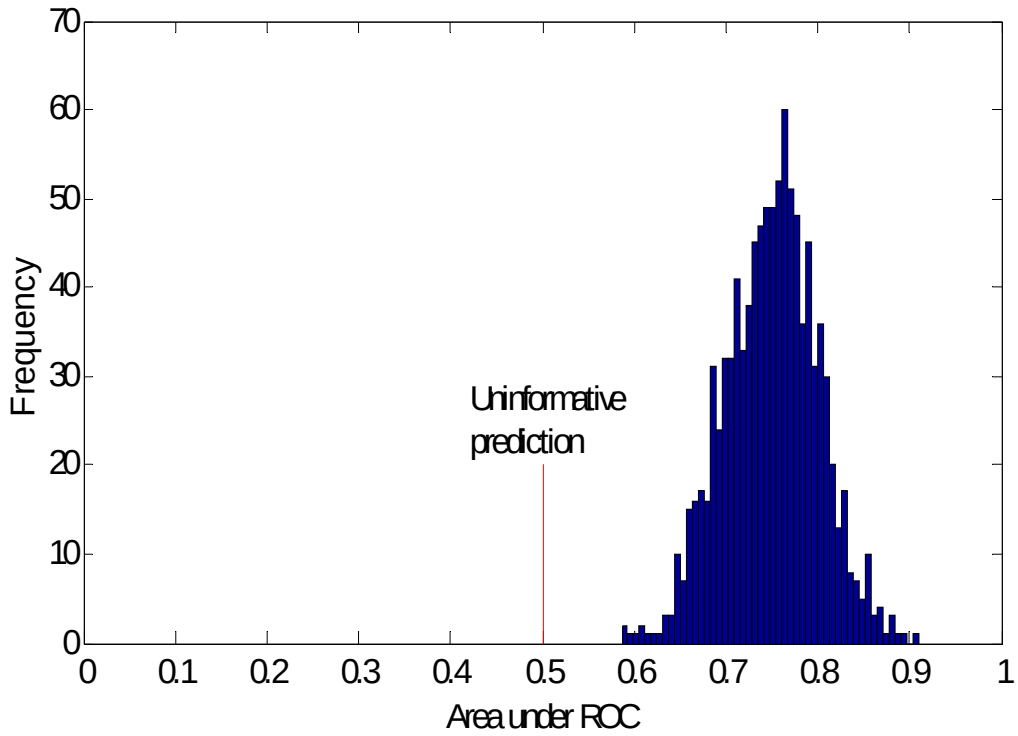
$$\min_{a_i} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\Phi(x_i) \cdot \Phi(x_j)) - \sum_i \alpha_i$$

$$s.t.\ \mathrm{C} \geq \alpha_{\mathrm{i}} \geq 0, \forall x_i$$

$$\sum_i \alpha_i y_i = 0$$

Prevent or detect over-fitting and estimate error accurately

- Label re-shuffling → detects over-fitting



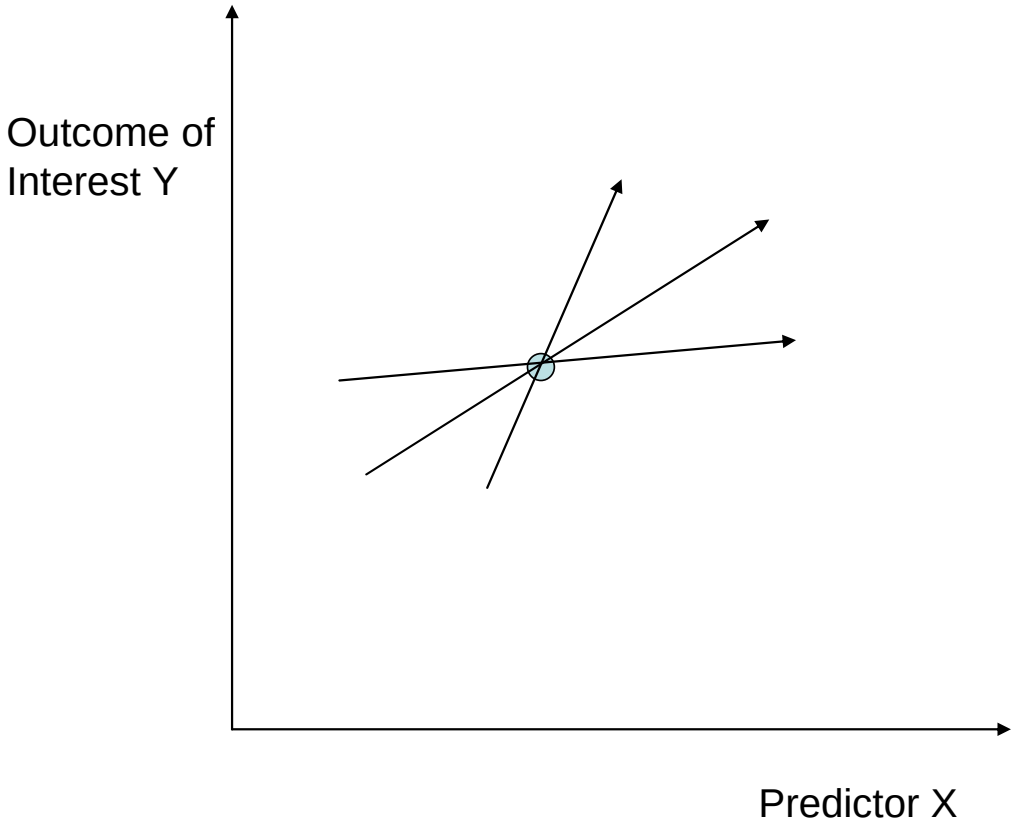
Problem #2:
Curse of Dimensionality

What is The Curse of Dimentionality?

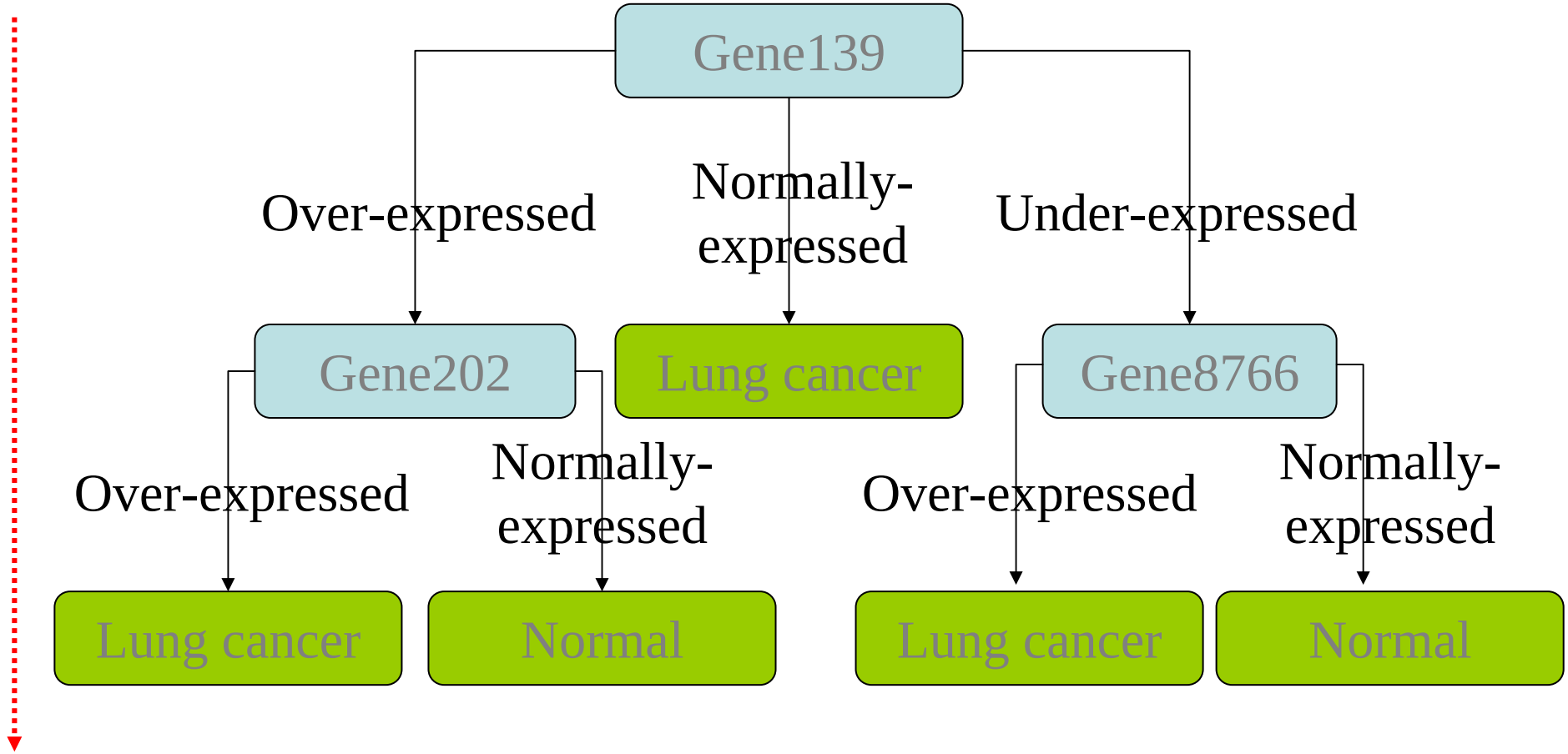
- The term “Curse of Dimensionality” refers to the increased difficulty (or even, in the worst case, *collapse*) of applying statistical or machine learning methods due to very large number of predictors.
- Large predictor numbers may facilitate not only over-fitting but also several additional problems as described below:

What are the effects of The Curse of dimensionality?

- Under-specified models

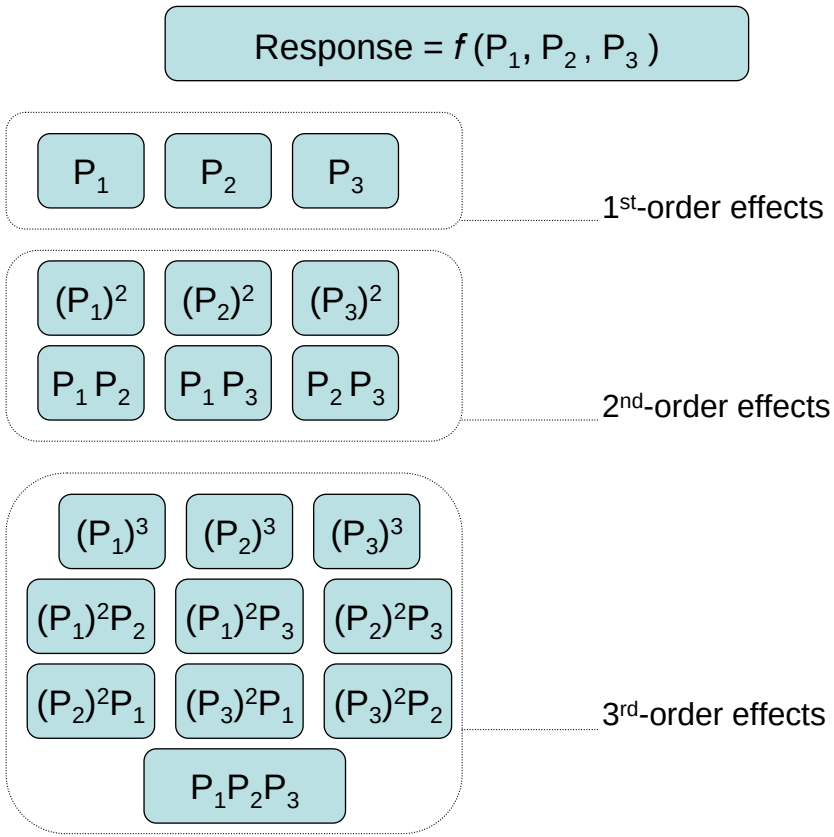


What are the effects of The Curse of dimensionality?



What are the effects of The Curse of dimensionality?

- Computational intractability to fully explore interaction effects/non-linearities



What are the effects of The Curse of dimensionality?

- Irrelevant of superfluous dimensions may severely affect methods that calculate distances in the input space. E.g., KNN:
- Example problem classify subject3 as likely to develop cancer of the lung or not by similarity to known cancer status of subjects 1 & 2:

Subject1=[smoking+, eyes-blue, likes-Rock, name-of spouse-John | Ca+]

Subject3=[smoking-, eyes-blue, likes-Rock, name-of spouse-Peter]

Subject2=[smoking-, eyes-brown, likes-Jazz, name-of spouse-Adam | Ca-]

Methods to deal with the curse of dimensionality

How to circumvent the curse of dimensionality

- Feature selection → reduces dimensions to minimum required by learning task
- Dimensionality reduction → projects many original dimensions to few new ones sufficient for the learning task
- Methods robust to high dimensionality → inherently resist COD (e.g., SVMs)
- Expert filtering based on domain knowledge → i.e., domain knowledge-based feature selection
- Newer algorithms to fit regression models → address over-specificity of models

Problem #3:
Causality versus Predictiveness

Causality versus Predictiveness

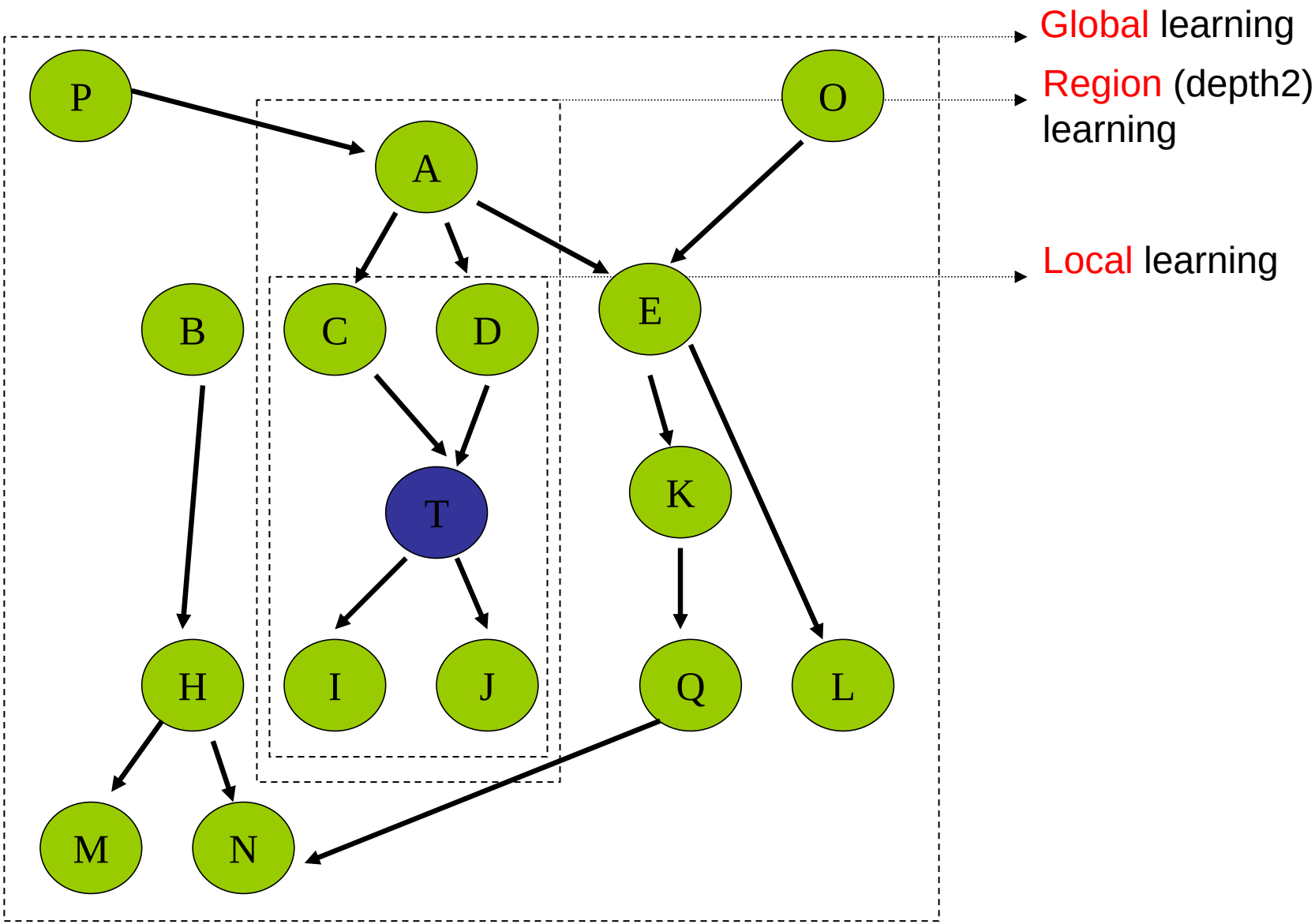
- In the biomedical and biostatistical communities it is widely accepted that the final judge of causal (as opposed to predictive or associational) knowledge is the randomized controlled experiment.
- However, quite often experimentation is impossible, impractical, and/or unethical.

Causality versus Predictiveness

- For these reasons some researchers have been using a number of heuristic methods for causal discovery.
- The four most prominent heuristics for causal discovery in biomedicine are:
  - Heuristic 1 = “If they cluster together they have similar or related function”.
  - Heuristic 2 = “If *A* is a robust and strong predictor of *T* then *A* is likely a cause of *T”*.
  - Heuristic 3 = “The closer *A* and *T* are in a causal sense, the stronger their correlation”.
  - Heuristic 4 = Surgeon’s General’s Epidemiological Criteria for Causality [U. S. Department of Health, Education, and Welfare, 1964]: “*A* is causing *B* with high likelihood if: (i) *A* precedes *B*; (ii) *A* is strongly associated with *B*; (iii) *A* is consistently associated with *B* in a variety of research studies, populations, and settings; (iv) *A* is the only available explanation for *B* (“coherence”); (v) *A* is specifically associated with *B* (but with few other factors)”.

Causal Discovery Algorithms

- Newer formal frameworks allow discovery of:
  - causal relationships (existence and/or directionality) among all variables: learning the full causal graph or the skeleton
  - direct causal relationships between a variable of interest and the remaining variables: learning the local causal neighborhood
  - indirect causal relationships between a variable of interest and the remaining variables: learning the local causal region

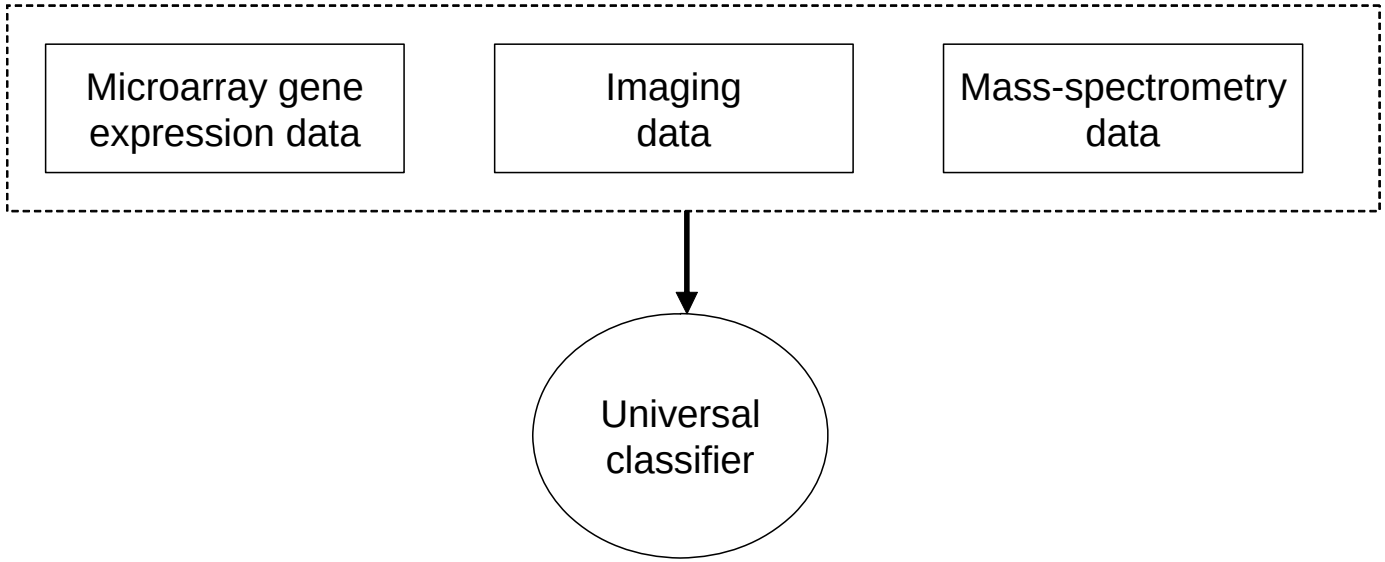

P
A
O
B
C
D
E
T
K
H
I
J
Q
L
M
N
Global learning
Region (depth2) learning
Local learning

Causal Discovery Algorithms

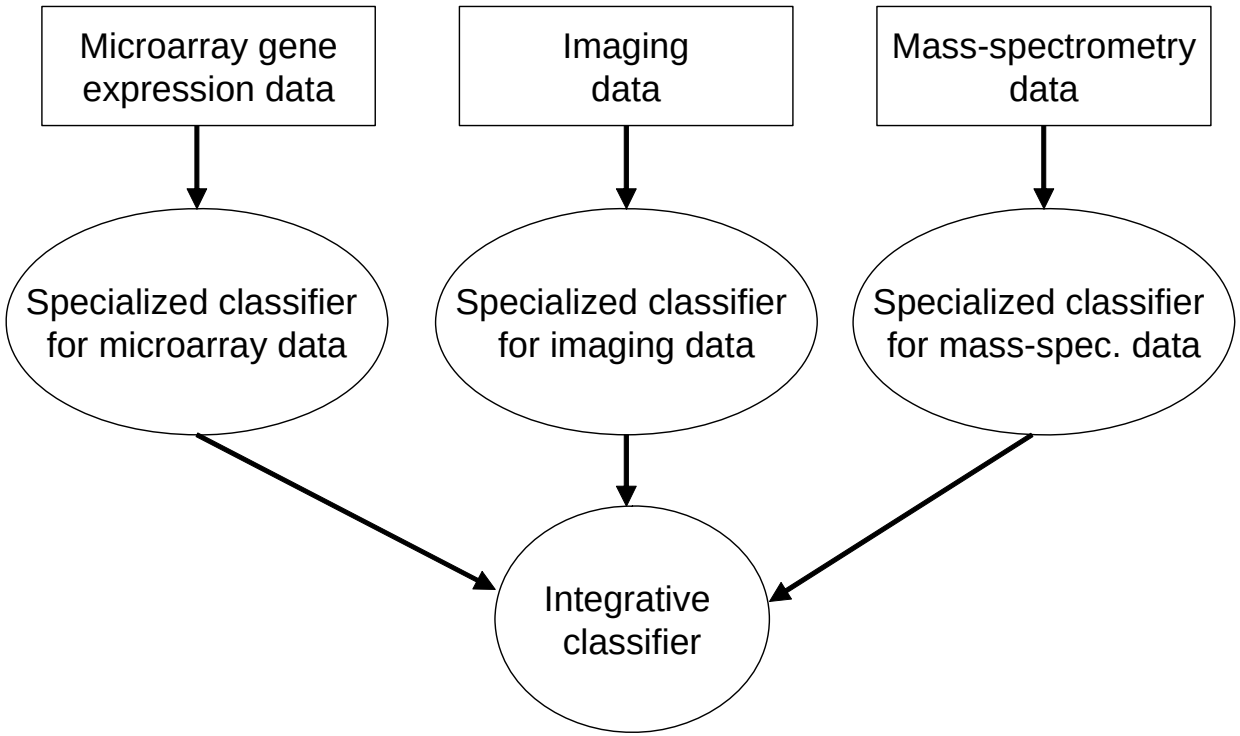
- Typical assumptions:
  1. Causal sufficiency (no hidden variables, local or global)
  2. Dependences and independences in the data are behaved well (i.e., strictly according to the "Causal Markov Property")
  3. There is enough sample relative to the functional forms among variables and the size of the causal neighbourhood
- While these assumptions are sufficient, there are not necessary and departures from them can be tolerated to various degrees depending on the nature of the data analysed & the algorithm employed

Problem #4:
Integrating heterogeneous data

Simultaneous (tightly-integrated) analysis



Sequential (loosely-integrated) analysis



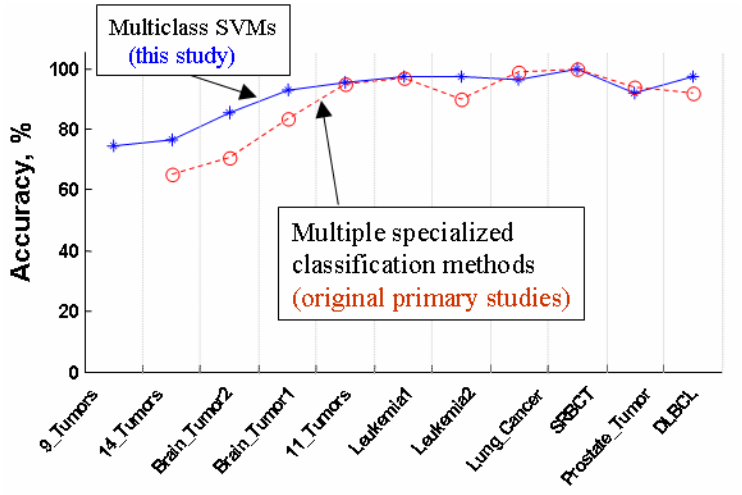
Problem #5:
Lack of standard protocols for data analysis

Consider for example mass spectrometry

- Data analysis may involve (but not limited to) the following steps:
  - M/Z range restriction,
  - baseline subtraction,
  - normalization,
  - peak detection,
  - peak alignment,
  - binning,
  - feature selection,
  - classifier construction, and
  - classifier evaluation.
- Thre is no agreement on methods for each step or on the specific sequence of steps
- For example, while building classification models from mass spectrometry data some studies apply Decision Trees while others apply Support Vector Machines; some studies perform baseline subtraction and peak detection as a single integrated step, others perform these operations sequentially, yet others do not perform baseline subtraction at all. Etc. etc.

A proposed process for protocol development

- Step 1 → thorough evaluation of component algorithms, model selection schemes and error estimation procedures across many representative datasets for a specific problem area (e.g., diagnosis of cancer from microarray or mass spectrometry data).
- Step 2 → validation of the protocols in independent datasets and compared to published analyses and new ones, including cross-dataset experiments.
- Step 3 → automated software & usability testing of the resulting software.



Evaluation Using New Datasets

Datasets

Dataset name	Number of			Reference
	Samples	Variables (genes)	Categories	
6_Tumors	353	7069	6	Shedden, 2003
Leukemia3	248	12135	6	Yeoh, 2002
Lung_Cancer2	96	7129	2	Beer, 2002
Lung_Cancer3	181	12533	2	Gordon, 2003
DLBCL2	210	32404	2	Savage, 2003

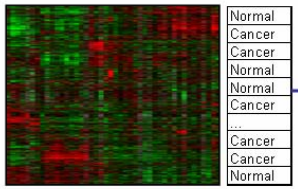
Comparison with literature

Dataset name	GEMS classification accuracy	Published classification accuracy
6_Tumors	97.2%	96.0%
Leukemia3	98.4%	98.4%
Lung_Cancer2	100.0%	100.0%
Lung_Cancer3	99.4%	99.3%
DLBCL2	87.1%	83.9%

Analyzes were completed within 10-30 minutes with *GEMS*.

GEMS

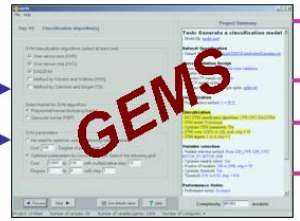
Gene expression data and outcome variable



Normal
Cancer
Cancer
Normal
Normal
Cancer
...
Cancer
Cancer
Normal

***Optional:* Gene names & IDs**

ring finger protein 1
tubulin, beta, 5
glucose-6-phosphate dehydrogenase
glutathione S-transferase M 5
carnitine acetyltransferase
Rho GTPase activating protein 4
SMA3
mannose phosphate isomerase
mitogen-activated protein kinase 3
leukotriene A 4 hydrolase
chromosome 21 open reading frame 1
dihydropyrimidinase-like 2
beta-2-microglobulin
discs, large (Drosophila) homolog 4



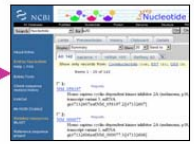
Classification model

Cross-validation performance estimate

Reduced set of genes

Rho GTPase activating protein 4
SMA3
mannose phosphate isomerase
mitogen-activated protein kinase 3

Links to literature



(model generation & performance estimation mode)

Conclusions

- All current error estimation procedures are sensitive to extremely small samples.
- Given a fixed sample size, avoiding over-fitting relies on techniques that control the complexity of the model and match it well to the available sample and complexity of modeling task. Also model search must be restrained.
- Contrary to wide-spread belief, when sample size is very small, "Independent dataset validation" is **not** a thorough solution to the over-fitting problem: it neither prevents over-fitting nor does it detect it always.

- For all the above reasons, having sufficiently large sample is an important factor in designing mass-throughput studies.

Conclusions CONT'D

- Machine learning offers robust and computationally scaleable solutions often not currently available via traditional statistical modeling. Useful to combine approaches and to have collaborating experts.
- Separating mechanistic (causal) from predictive modeling is necessary but quite undeveloped currently.
- Two fundamentally distinct approaches with different strengths and weaknesses for integration: the tightly-integrated and the loosely-integrated frameworks.
- A three-step process for the development, validation and automation of analysis protocols.