Feature Selection

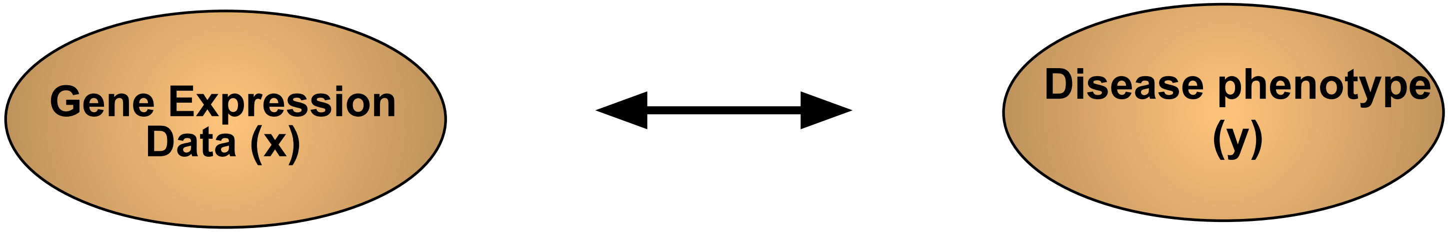
April 27, 2017

Karsten Borgwardt, ETH-Department BSSE in Basel

Content:

- What is feature selection?
- How can feature selection algorithms be used to gain insights into biological systems?
- What are typical problems in feature selection in practice?

What is Feature Selection?



Gene Expression Data (x)

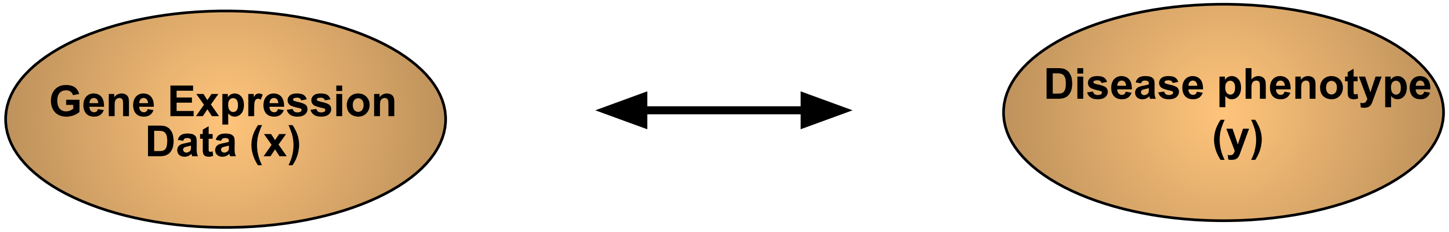
Disease phenotype (y)

In Systems Biology, the fundamental feature selection problem is to find those **components of a large biological system that affect a particular output/function/phenotype** of this system.

Some examples:

- Which gene expression levels are indicative of a particular disease?
- Which de novo mutations in the genome correlate with increased disease risk?

Why Feature Selection?



Gene Expression Data (x)

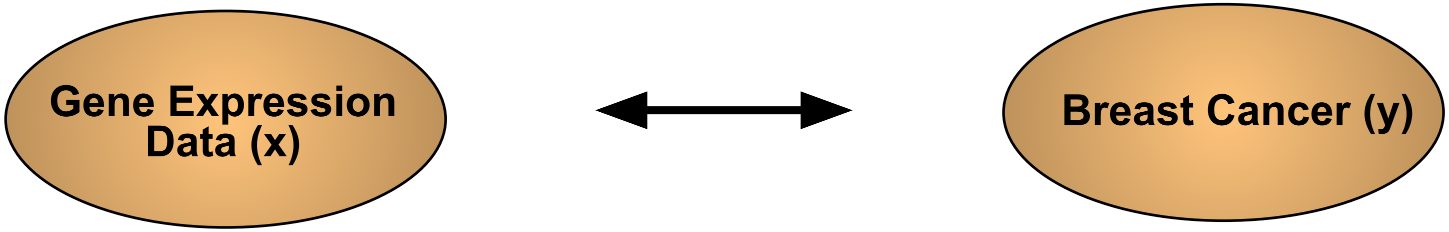
Disease phenotype (y)

In the omics age, we have plenty of high-dimensional molecular data to describe the state of a system/individual.

Why is feature selection necessary?

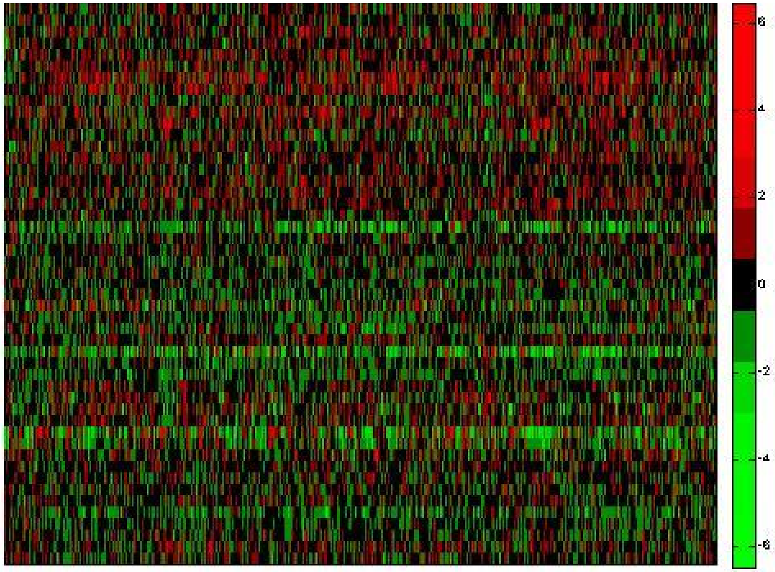
- To get a **better understanding of the molecular mechanisms** that correlated or even causal for a particular phenotype.
- To get **lower-dimensional models** that are easier and cheaper to observe, and study.
- To **remove noisy features**.

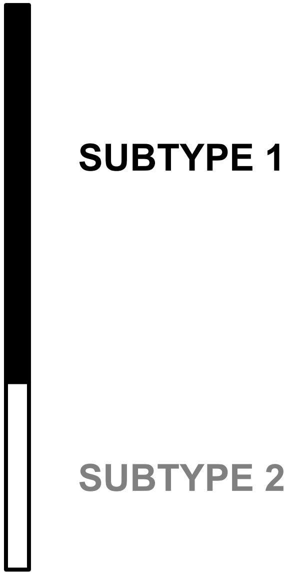
Example Application: Gene Selection



Gene Expression Data (x)

Breast Cancer (y)

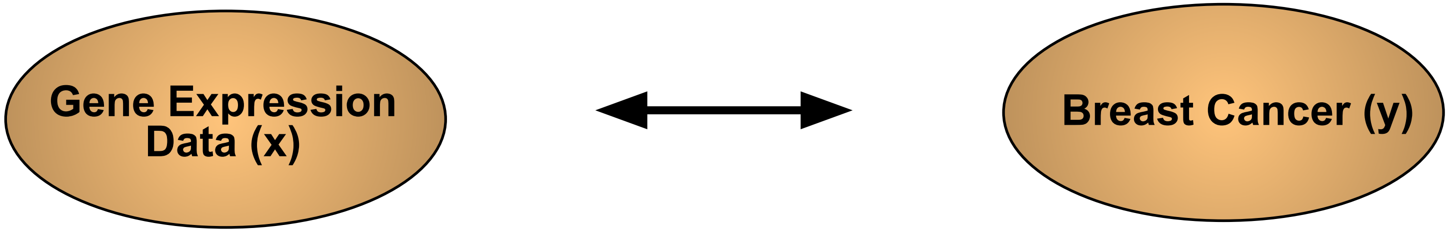




SUBTYPE 1

SUBTYPE 2

Example Application: Gene Selection



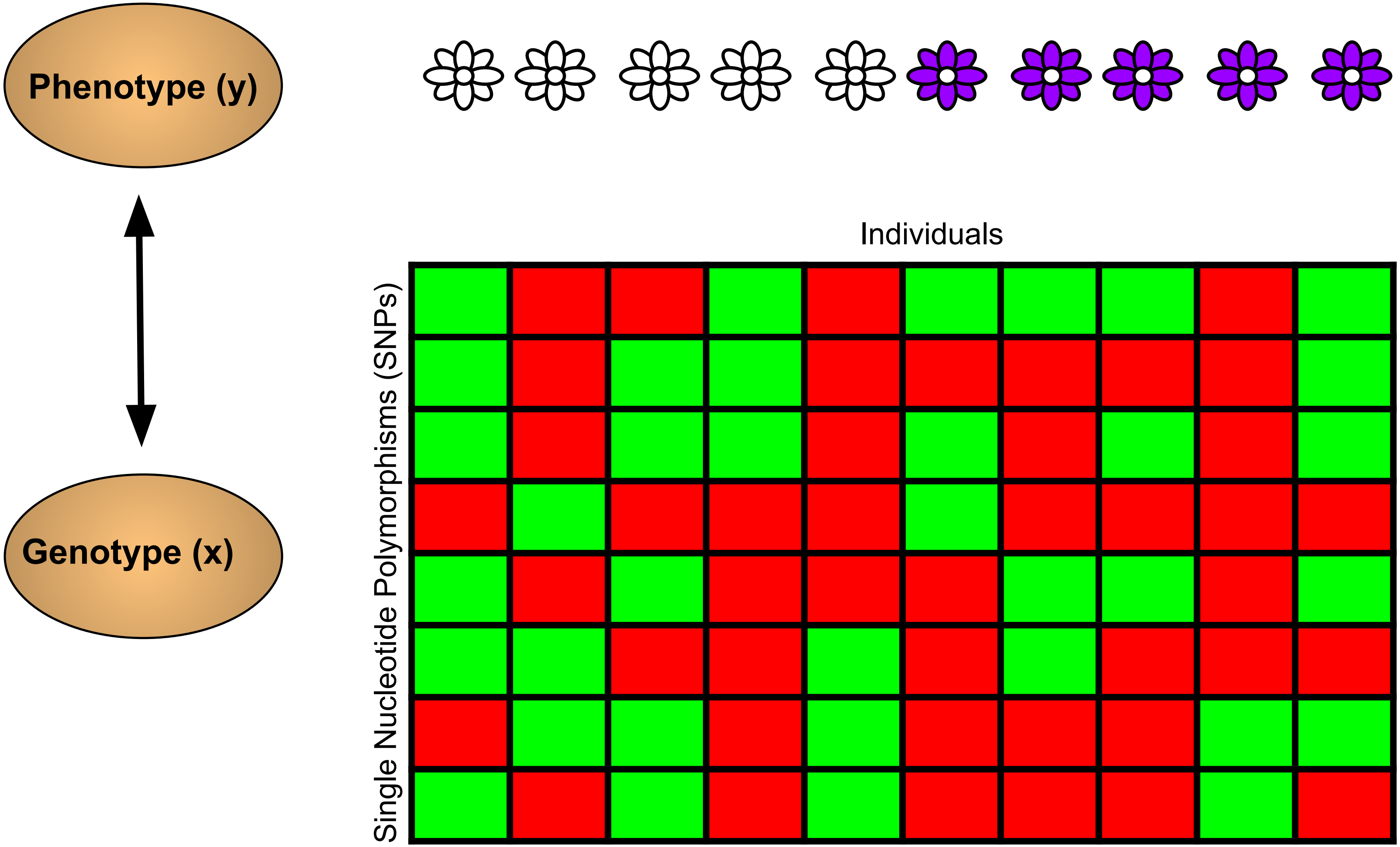
For each of *n* patients, we are given a vector **x** that includes the expression levels of all d genes and a phenotype **y**.

First question to ask:

Which (single) gene is most associated with variation in the phenotype?

This is often referred to as univariate feature selection, as we are considering the effect of each gene in isolation.

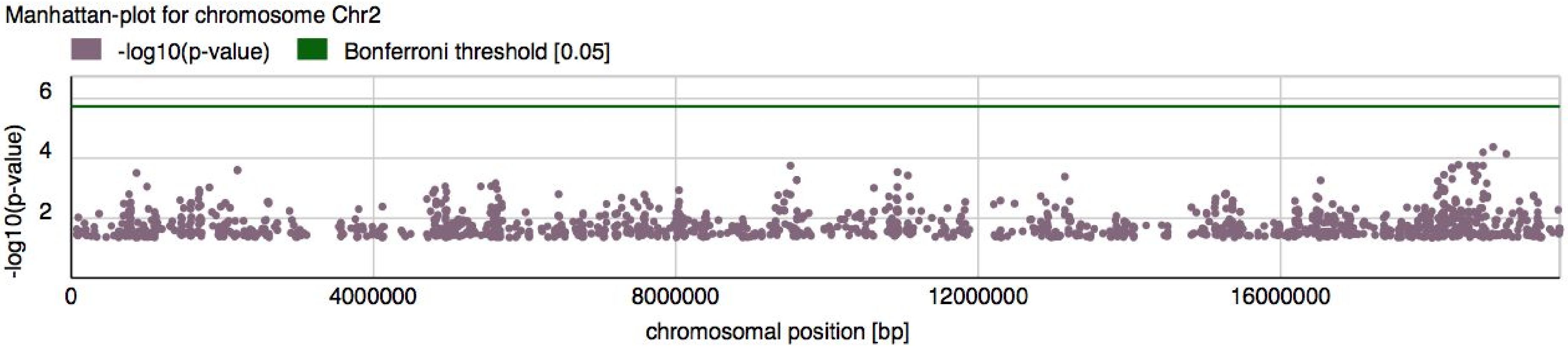
Example Application: Genome-wide association studies (GWAS)



🟩	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟥	🟩
🟩	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩
🟩	🟥	🟩	🟩	🟥	🟩	🟥	🟩	🟥	🟩
🟥	🟩	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥
🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟩
🟩	🟩	🟥	🟥	🟩	🟥	🟩	🟥	🟥	🟥
🟥	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟩
🟩	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟥

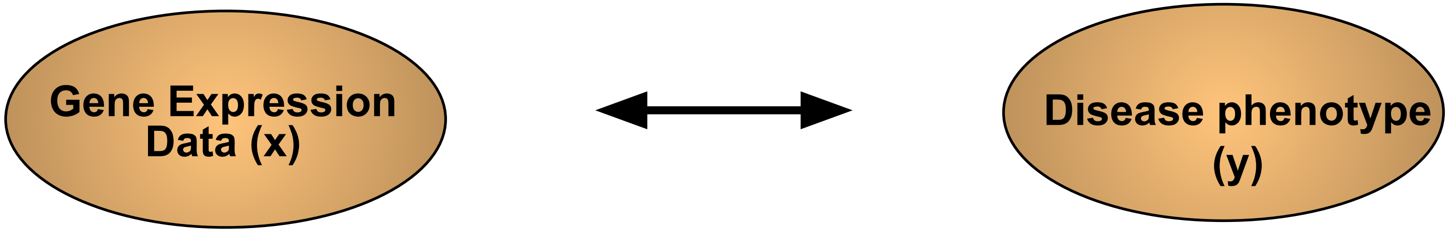
Example Application: GWAS

Manhattan Plot (SNPs vs. p-value)



Example: Anthocyanin, Chromosome 2, *Arabidopsis thaliana* from easygwas.org

Univariate Feature Selection



Univariate feature selection corresponds to the following optimization problem:

$$\arg\max_{j}(r(\mathbf{x}(:,j),\mathbf{y}))$$

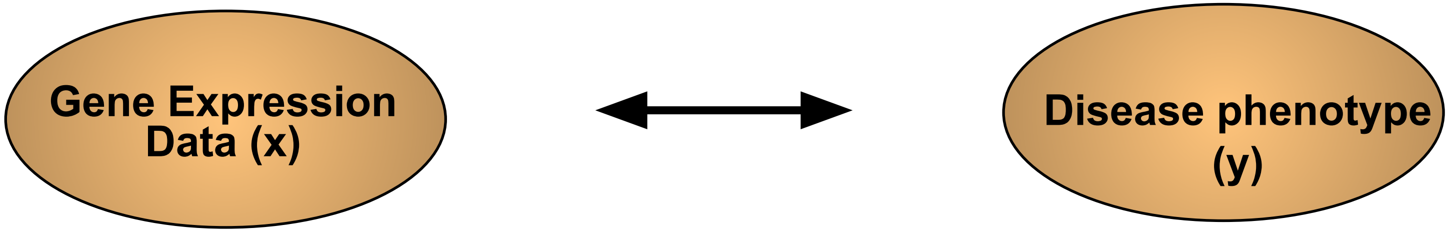
where

$r : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an information criterion (quality function, feature score),

x(:,j) is the vector of gene expression levels of gene j across all n patients,

y is the vector of phenotypes for all patients.

Univariate Feature Selection



Popular information criteria

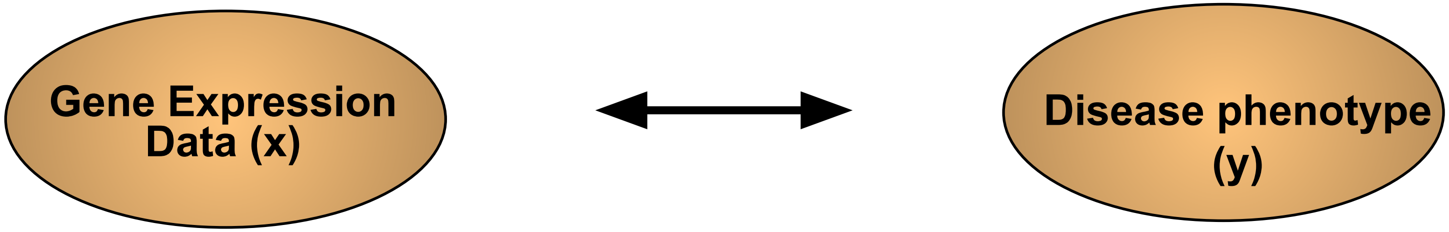
Pearson’s correlation coefficient (sample correlation coefficient)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Mutual Information (definition for discrete data)

$$I(\mathbf{x},\mathbf{y}) = \sum_{z_x \in Z_x} \sum_{z_y \in Z_y} p(z_x, z_y) \log(\frac{p(z_x, z_y)}{p(z_x)p(z_y)})$$

Univariate Feature Selection



General framework for univariate feature selection

1. For each feature *j*, compute its feature score *r(j)*

2. Sort features according to their score *r(j)* and return this sorted list as output

Common Pitfalls: Size of Solution

What is the ultimate goal?

1. **Determine the most relevant genes**

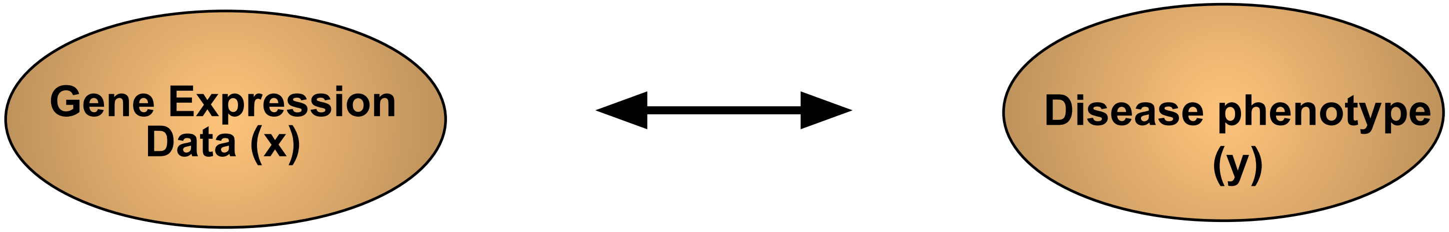
Problem: How to choose the number genes to be selected?

Approaches:

1.1 Probe method (Bi et al., JMLR 2003): Randomly generate a noise feature *z*. Select all features *j* with *r(j)* > *r(z).*

1.2 Significance method: Compute a *p*-value for the association between each feature and the phenotype. Select all features whose association is statistically significant.

Common Pitfalls: Multiple Hypotheses Testing



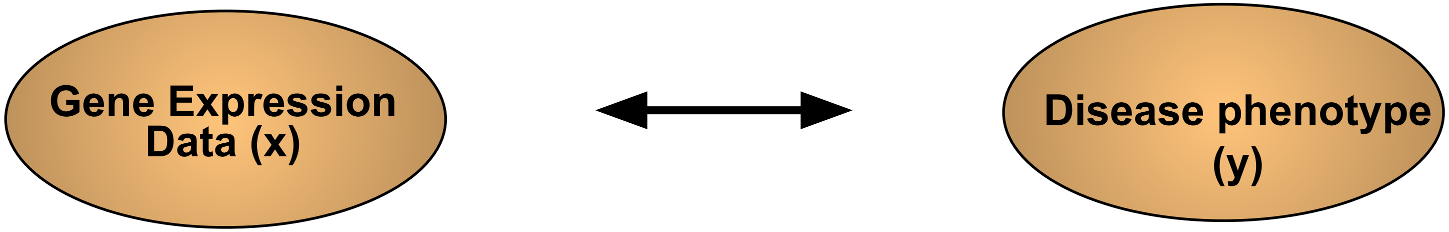
We are testing thousands of genes for association with the phenotype.

We typically compute a p-value that measures the probability of observing such a strong (or even stronger) association ***if gene expression levels and phenotype are independent***.

It is the probability of a significant finding given that there is no true association.

A finding is called significant if its p-value is below a predefined significance threshold α (usually 0.05 or 0.01).

Common Pitfalls: Multiple Hypotheses Testing



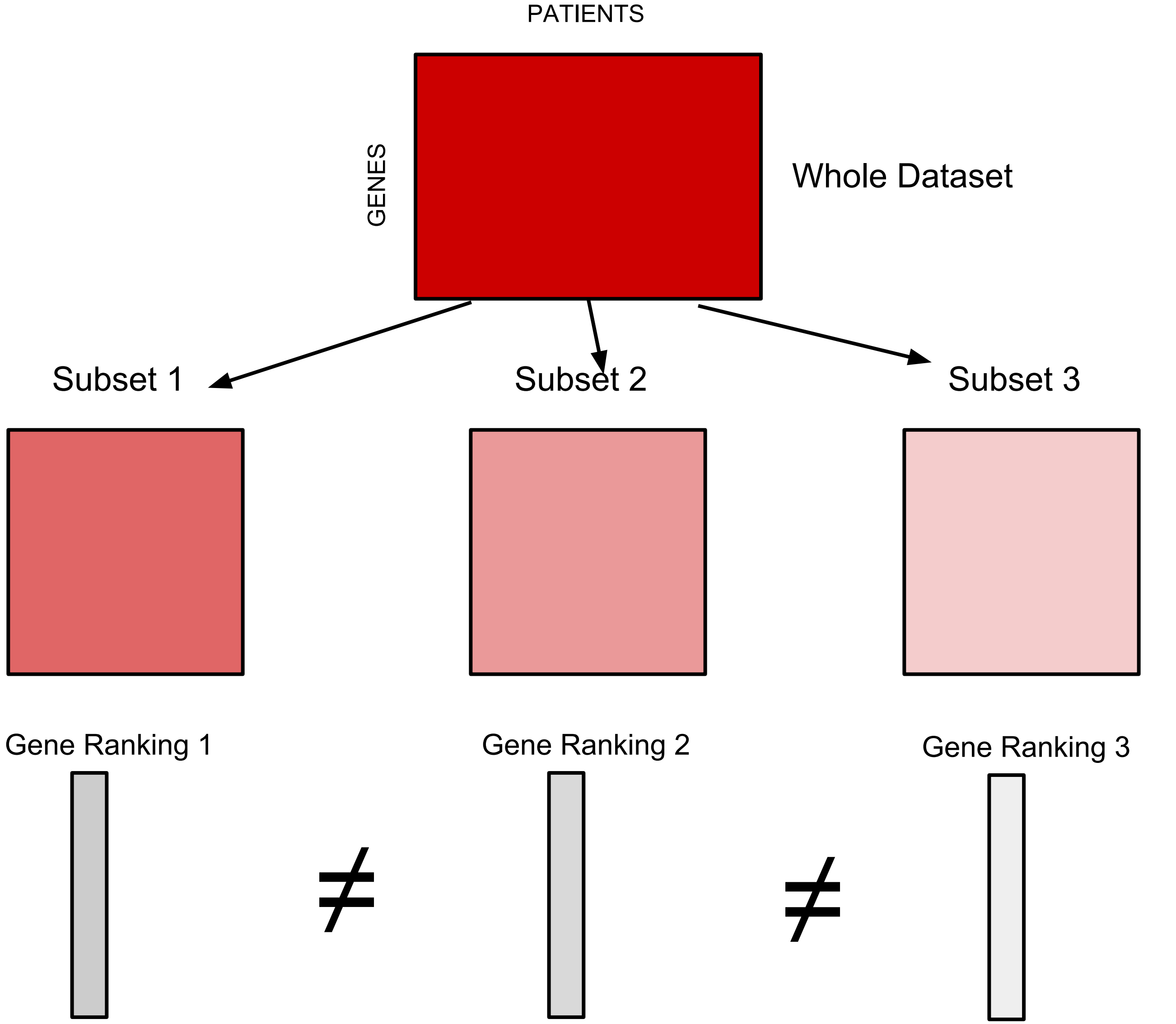
Problematic consequence: We will deem many genes significantly associated by mistake if we test thousands of them.

Way out: Correction for multiple hypothesis testing to control the Family-Wise Error Rate (FWER), the probability of “making at least one mistake”, i.e. selecting at least one false positive.

Most popular approaches: Bonferroni correction (1936), which divides the significance level α by the number of tests performed.

Problem: Very conservative, will find few significant genes
-> less conservative alternatives such as **False Discovery Rate** (last week’s lecture: matlab command **mafdr**)

Common Pitfalls: Instability



Common Pitfalls: Instability

Instability of Results: When performing feature selection on subsets of the same dataset or with different ranking criteria, the ranking of features often varies greatly.

Boulesteix & Slawski (Briefings in Bioinformatics 2009):

- Univariate analyses in the small *n* large *d* scenario are well known to produce highly instable outputs, in the sense that a small change in the data or a minimal modification of the ranking criterion often results in a fully different ordering of the features.

- A ranked gene list should not be considered as a unique definitive result. It makes sense to study the stability of a list by considering alternative ranking criteria and/or slightly modified versions of the data set.

Common Pitfalls: Instability

- How to aggregate results from different rankings?

Two popular strategies:

Strategy 1: Compute the average rank of a gene across all experiments

Strategy 2: Compute the probability of gene to be ranked among the top-k genes in each experiment

Common Pitfalls: Instability

- How to aggregate p-values from different experiments?

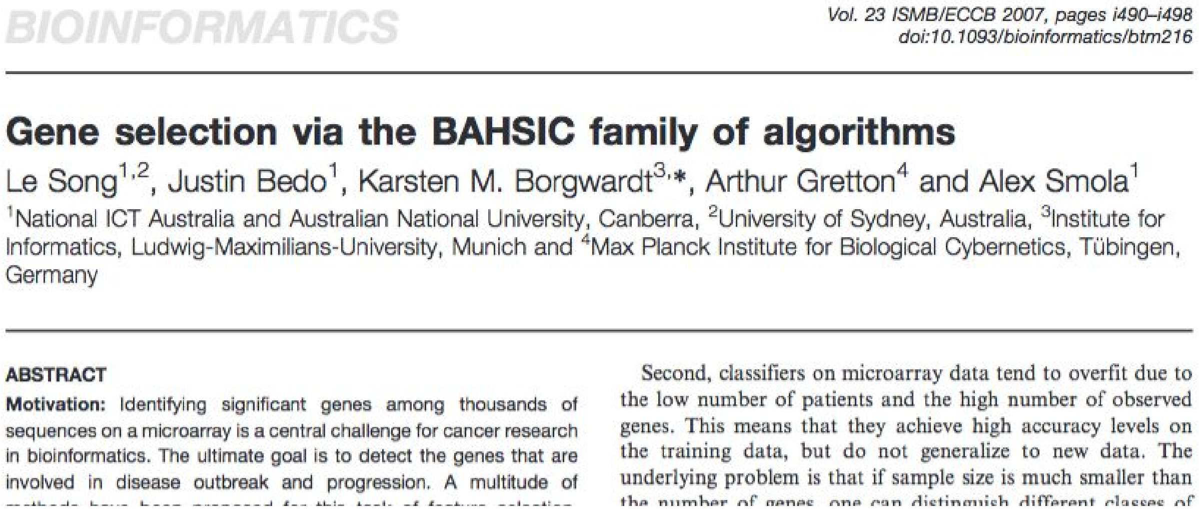
Classic strategy for this kind of meta-analysis (Hong & Breitling, Bioinformatics 2007):

Fisher's Inverse χ^2 test (Fisher, 1925) computes a combined statistic from the *p*-values obtained from the analysis of the *k* individual datasets, $s = -2 \Sigma_i \log(p_i)$, where *s* follows a χ^2 distribution with *2k* degrees of freedom under the joint null hypothesis.

Common Pitfalls: Instability

- Another source of instability is the vast number of methods that have been proposed for feature selection

- We could show that the different methods differ mostly in the (1) preprocessing of that data and (2) in the similarity measures used to compare gene expression levels and phenotypes to each other.



BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i490–i498
doi:10.1093/bioinformatics/btm216

Gene selection via the BAHSIC family of algorithms

Le Song[1,2], Justin Bedo[1], Karsten M. Borgwardt[3,*], Arthur Gretton[4] and Alex Smola[1]

[1]National ICT Australia and Australian National University, Canberra, [2]University of Sydney, Australia, [3]Institute for Informatics, Ludwig-Maximilians-University, Munich and [4]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

ABSTRACT

Motivation: Identifying significant genes among thousands of sequences on a microarray is a central challenge for cancer research in bioinformatics. The ultimate goal is to detect the genes that are involved in disease outbreak and progression. A multitude of methods have been proposed for this task of feature selection

Second, classifiers on microarray data tend to overfit due to the low number of patients and the high number of observed genes. This means that they achieve high accuracy levels on the training data, but do not generalize to new data. The underlying problem is that if sample size is much smaller than the number of genes, one can distinguish different classes of

Multivariate Feature Selection: Additive Models

The results of univariate feature selection are limited in the following ways:

- Captures effect of single genes only
- Does not consider correlations between genes
- Does not consider additive effects between genes
- Does not consider interactions between genes

In short: Univariate feature selection ignores the *systems biology* character of the problem!

Multivariate Feature Selection: Additive Models

Large class of methods for **linear regression**: $y = \sum_{i=1}^{d} x_i \beta_i + \epsilon.$

- Try to predict y from **x**
- Features receive weights **β**

Mathematical formulation:

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2$$

The approach to feature selection is “indirect” here: Relevant features get a non-zero weight in **β**

Problem: In the above formulation, almost all entries of **β** tend to be non-zero.

Multivariate Feature Selection: Lasso Model

Lasso Model (Tibshirani, 1996)

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda_1 ||\beta||_1$$

Idea: Reward solutions ($\boldsymbol{\beta}$) in which few entries of $\boldsymbol{\beta}$ are non-zero.

Concept: This is achieved by minimizing the L1-norm of $\boldsymbol{\beta}$:

$$||\beta||_1 = \sum_{i=1}^{d} |\beta_i|$$

Disadvantage: If there are groups of correlated features, the Lasso often picks just one feature from a group.

Multivariate Feature Selection: Ridge Regression

Ridge Regression

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda_2 ||\beta||_2^2$$

Idea: Reward solutions ($\boldsymbol{\beta}$) in which correlated features get similar weights.

Concept: This is achieved by minimizing the L2-norm of $\boldsymbol{\beta}$:

$$||\beta||_2 = \sqrt{\sum_{i=1}^{d} \beta_i^2}$$

Disadvantage: The solution is often not sparse.

Multivariate Feature Selection: Different Norms

Example: Effect of different norms

Scenario 1 - Higher values: We compare two solutions β = 0.5 and β*=0.4.

- The L1-norm of β is 0.5, the L1-norm of β* is 0.4 (Difference is 0.1).
- The squared L2-norm of β is 0.25, the squared L2-norm of β* is 0.16.
- Changing the solution from β to β* leads to an **improvement of 0.1 in terms of L1-norm, and of 0.09 in terms of squared L2-norm**.

Scenario 2 - Lower values: We compare two solutions β = 0.2 and β*=0.1.

- The L1-norm of β is 0.2, the L1-norm of β* is 0.1 (Difference is again 0.1).
- The squared L2-norm of β is 0.04, the squared L2-norm of β* is 0.01.
- Changing the solution from β to β* leads to an **improvement of 0.1 in terms of L1-norm, and of 0.03 in terms of squared L2-norm**.

Insight: The L2-norm rewards reducing larger values more than reducing lower values. The L1-norm rewards both identically.

Multivariate Feature Selection: Elastic Net

Elastic Net (Zou and Hastie, 2005)

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$$

Idea: Reward solutions (**β**) in which groups of correlated features get similar weights and few weights are non-zero.

Concept: This is achieved by simultaneous minimization of the L1-norm and the L2-norm of **β**.

Disadvantage: Two parameters have be to set.

Common Pitfalls: Overfitting

What is the ultimate goal?

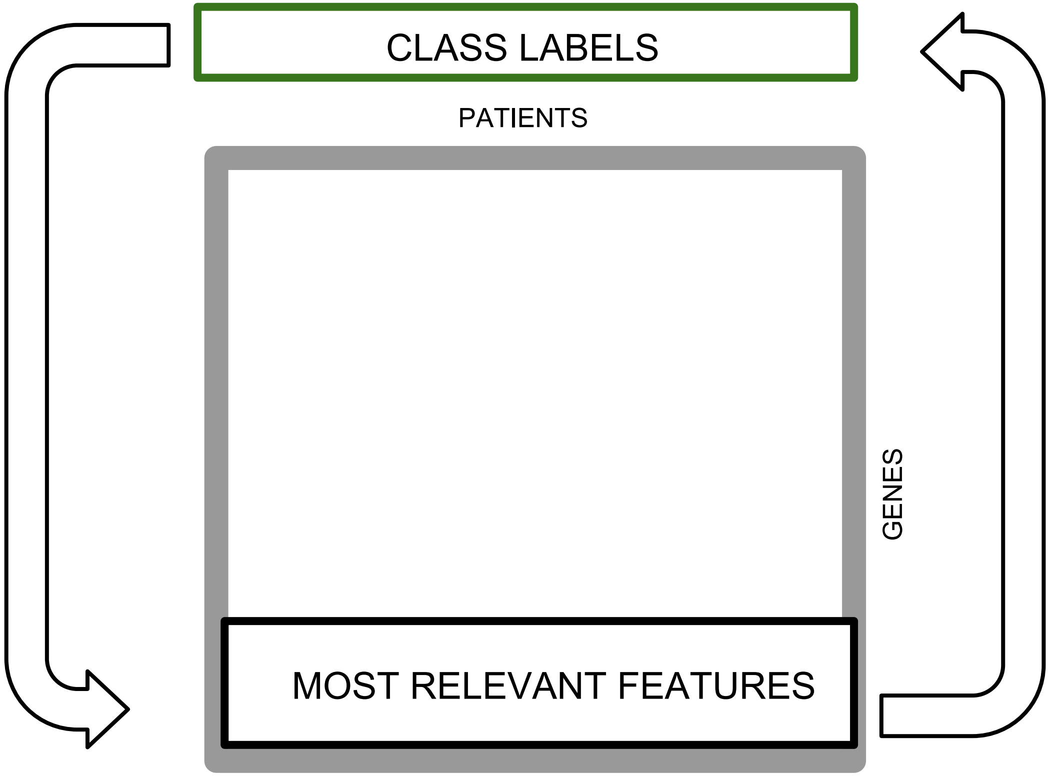
Use the most relevant genes to predict the phenotype based on gene expression levels

Problem of Selection Bias: Feature selection and prediction must not happen on the same dataset.

Ignoring Selection Bias leads to overly optimistic results (Ambroise and McLachlan, PNAS 2002)

Common Pitfalls: Overfitting

UNCLEAN EVALUATION STRATEGY THAT LEADS TO OVERFITTING:



Step 1:
Feature Selection

Step 2:
Class Label
Prediction

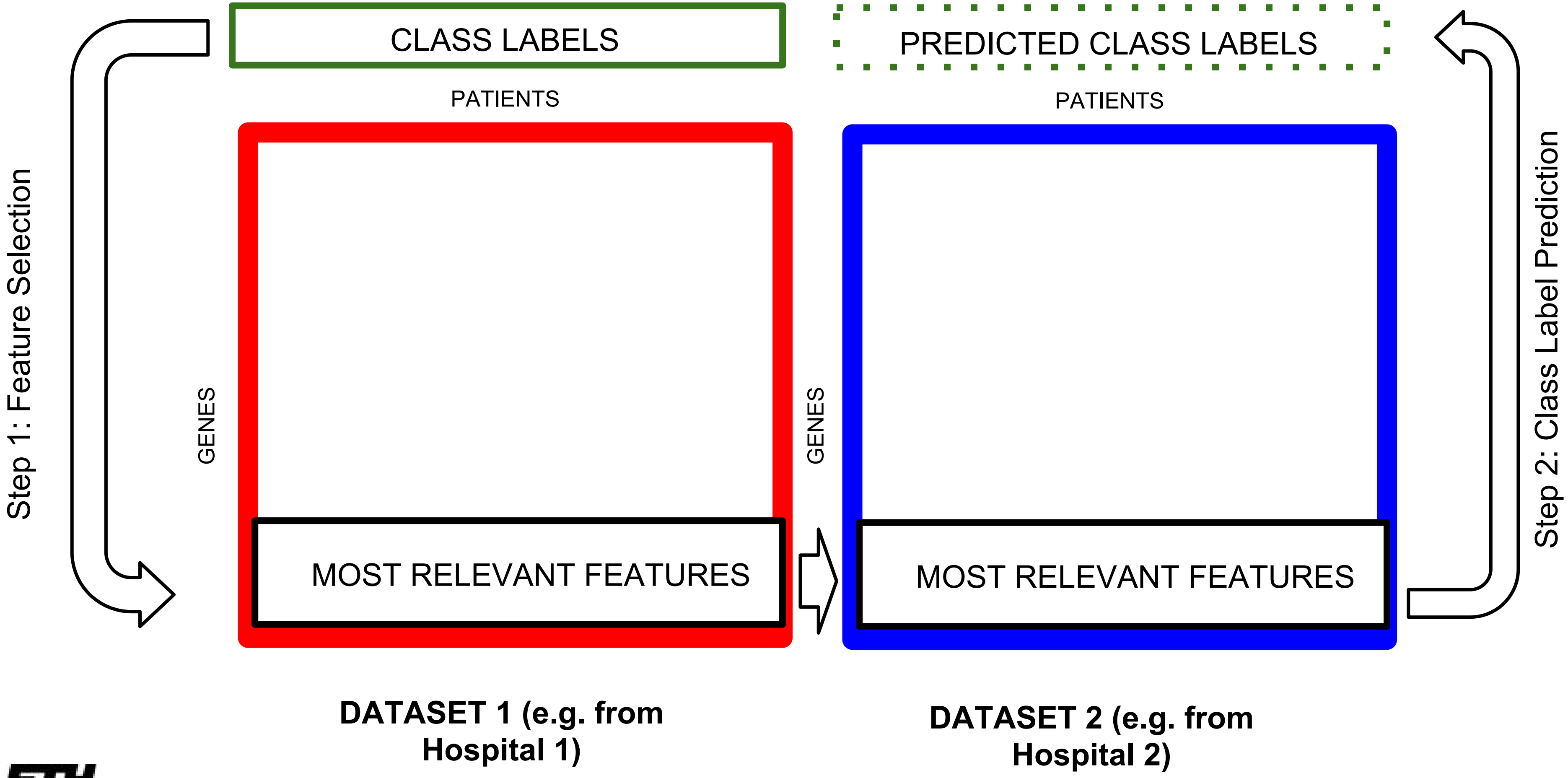
Common Pitfalls: Overfitting

How to avoid the selection bias?

- Select the features on the training dataset only
- Use these features for prediction on the separate test dataset

Common Pitfalls: Overfitting

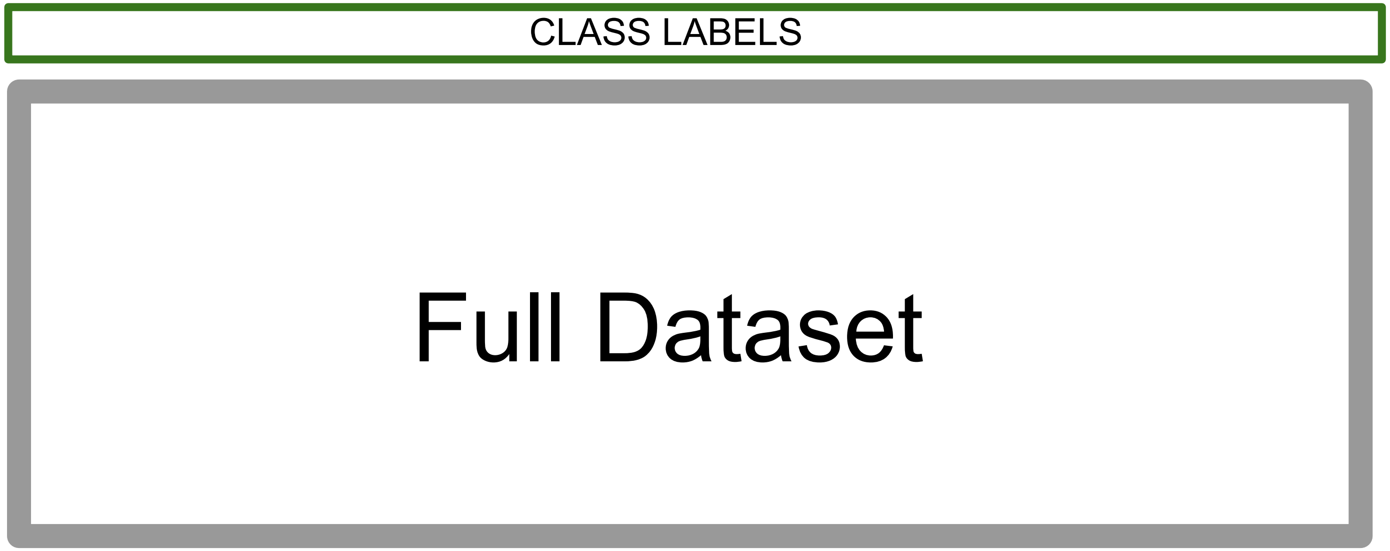
CLEAN EVALUATION STRATEGY THAT AVOIDS OVERFITTING:


CLASS LABELS
PREDICTED CLASS LABELS
PATIENTS
PATIENTS
Step 1: Feature Selection
GENES
GENES
Step 2: Class Label Prediction
MOST RELEVANT FEATURES
MOST RELEVANT FEATURES
DATASET 1 (e.g. from Hospital 1)
DATASET 2 (e.g. from Hospital 2)

Common Pitfalls: Overfitting

How to avoid the selection bias?

- Select the features on the training dataset only
- Use these features for prediction on the separate test dataset



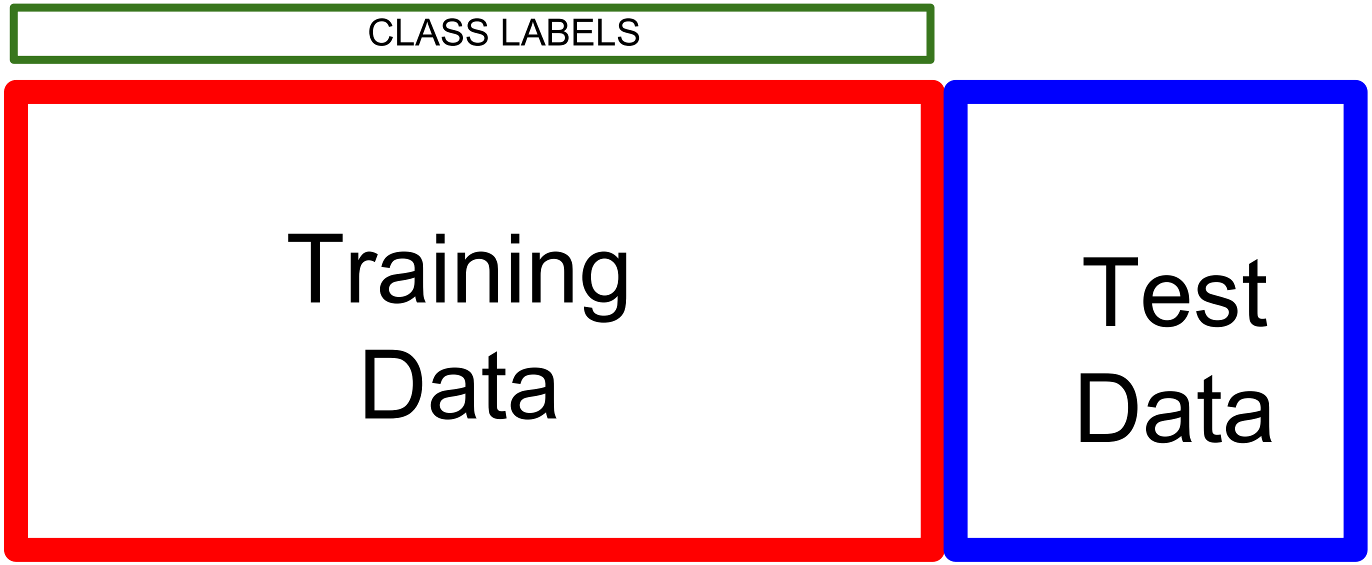
CLASS LABELS

Full Dataset

Common Pitfalls: Overfitting

How to avoid the selection bias?

- Select the features on the training dataset only
- Use these features for prediction on the separate test dataset



CLASS LABELS

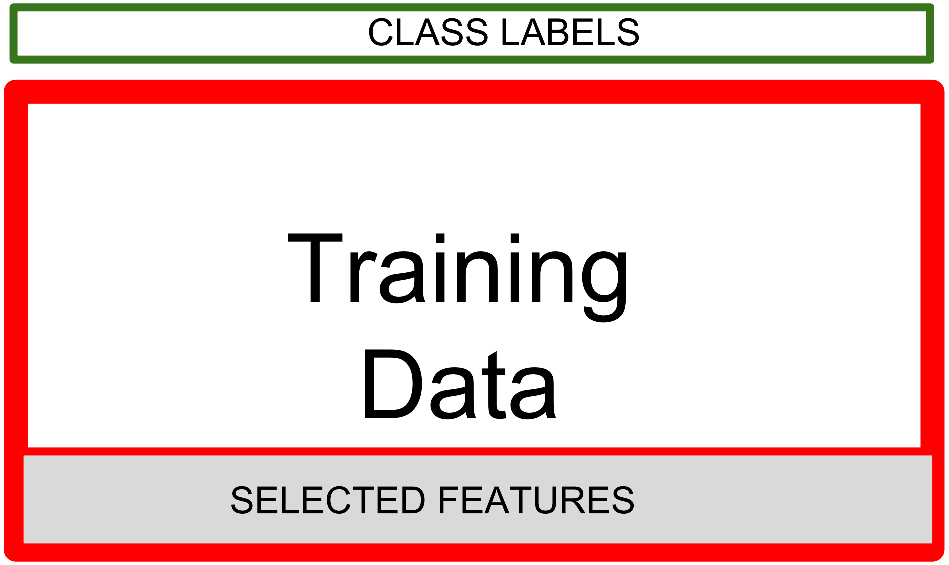
Training Data

Test Data

Common Pitfalls: Overfitting

How to avoid the selection bias?

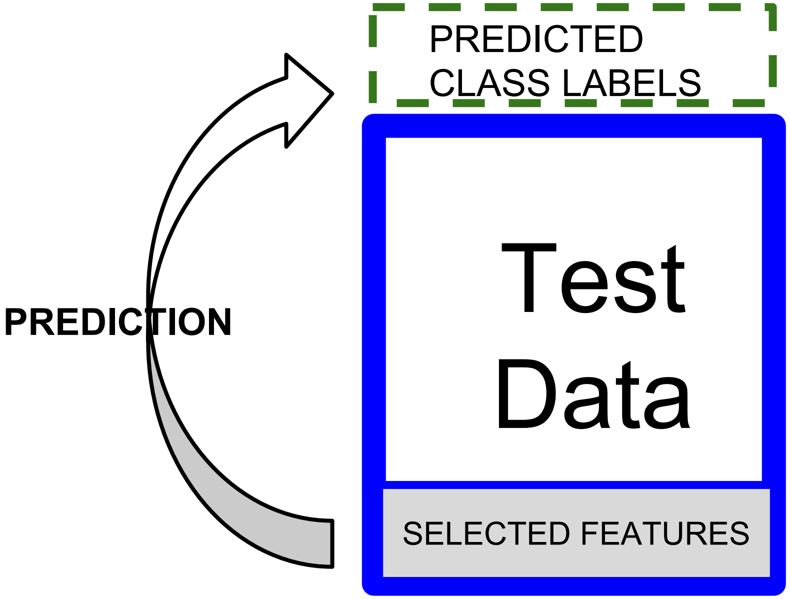
- Select the features on the training dataset only
- Use these features for prediction on the separate test dataset



Common Pitfalls: Overfitting

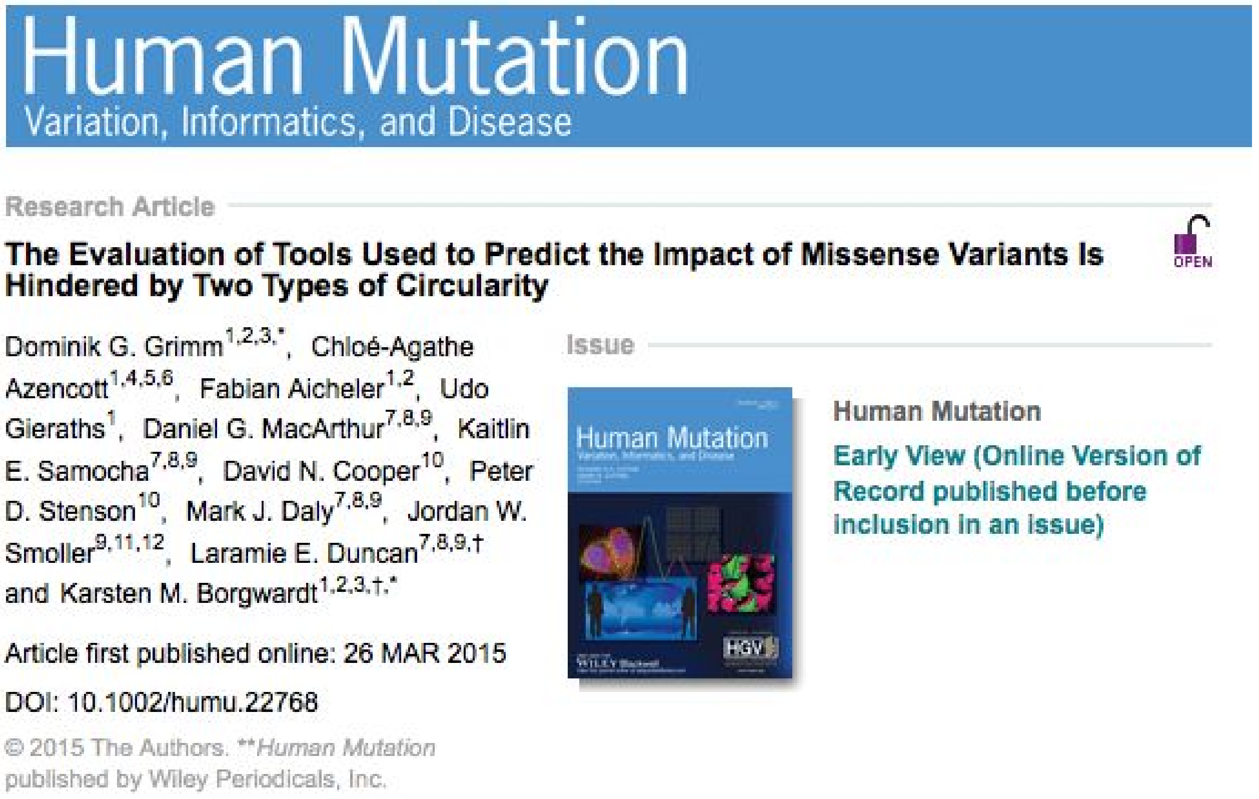
How to avoid the selection bias?

- Select the features on the training dataset only
- Use these features for prediction on the separate test dataset



Common Pitfalls: Overfitting

Overfitting is a very common problem in systems biology and computational biology in general



Human Mutation

Variation, Informatics, and Disease

Research Article

The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity

OPEN

Dominik G. Grimm[1,2,3,*], Chloé-Agathe Azencott[1,4,5,6], Fabian Aicheler[1,2], Udo Gieraths[1], Daniel G. MacArthur[7,8,9], Kaitlin E. Samocha[7,8,9], David N. Cooper[10], Peter D. Stenson[10], Mark J. Daly[7,8,9], Jordan W. Smoller[9,11,12], Laramie E. Duncan[7,8,9,†] and Karsten M. Borgwardt[1,2,3,†,*]

Article first published online: 26 MAR 2015

DOI: 10.1002/humu.22768

© 2015 The Authors. *Human Mutation* published by Wiley Periodicals, Inc.

Issue

Human Mutation

Early View (Online Version of Record published before inclusion in an issue)

Feature Selection: Summary

When performing feature selection in systems biology, be aware of:

- the **multiple hypothesis testing problem**. Appropriately correct for multiple testing, at least using Bonferroni correction.

- the **instability** of most methods on sample datasets. Aggregate rankings from several subsamples of the data. Aggregate rankings for different methods.

- the relative advantages and disadvantages of **univariate and multivariate feature selection**, when choosing your method.

- the **problem of overfitting**. Avoid it by cleanly separating the training dataset from the test dataset. Do not use the test set for feature selection.

References

Ambroise, Christophe, and Geoffrey J. McLachlan. “Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data.” *Proceedings of the National Academy of Sciences of the United States of America* 99, no. 10 (May 14, 2002): 6562–66.

Bi, Jinbo, Kristin Bennett, Mark Embrechts, Curt Breneman, and Minghu Song. “Dimensionality Reduction via Sparse Support Vector Machines.” *J. Mach. Learn. Res.* 3 (March 2003): 1229–43.

Bonferroni, Carlo E. 1936. *Teoria statistica delle classi e calcolo delle probabilita*.

Boulesteix, Anne-Laure, and Martin Slawski. “Stability and Aggregation of Ranked Gene Lists.” *Briefings in Bioinformatics* 10, no. 5 (September 1, 2009): 556–68.

Grimm, Dominik G., Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G. MacArthur, Kaitlin E. Samocha, David N. Cooper, M.J. Daly, J.W. Smoller, L.E. Duncan and K.M. Borgwardt. “The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity.” *Human Mutation*, February 14, 2015. doi:10.1002/humu.22768.

Song, Le, Justin Bedo, Karsten M Borgwardt, Arthur Gretton, and Alex Smola. “Gene Selection via the BAHSIC Family of Algorithms.” *Bioinformatics (Oxford, England)* 23, no. 13 (July 1, 2007): i490–98.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).