

Welcome to Data Science for Biomedical Research

BMIN 7054

Co-directors with complimentary expertise representing several departments:

* Department of Environmental Health,
College of Medicine, University of Cincinnati

** Department of Biomedical Informatics,
Cincinnati Children's Hospital Medical Center

*** Dept. of Electrical Engineering & Computer
Science, College of Engineering, University of
Cincinnati



Jarek Meller *,**,***



Michal Kouril **

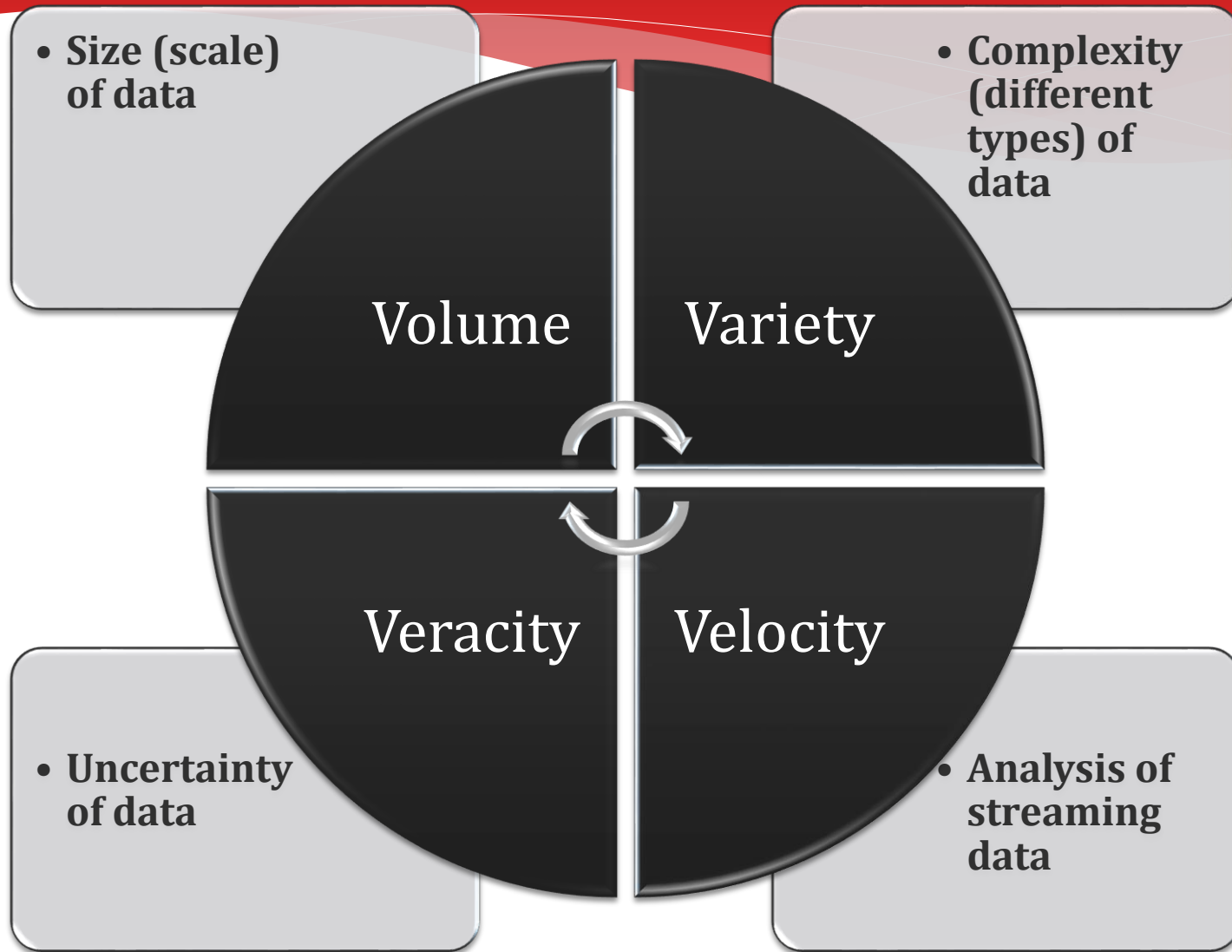


Raj Bhatnagar ***



MB Rao *

Big Biomedical Data

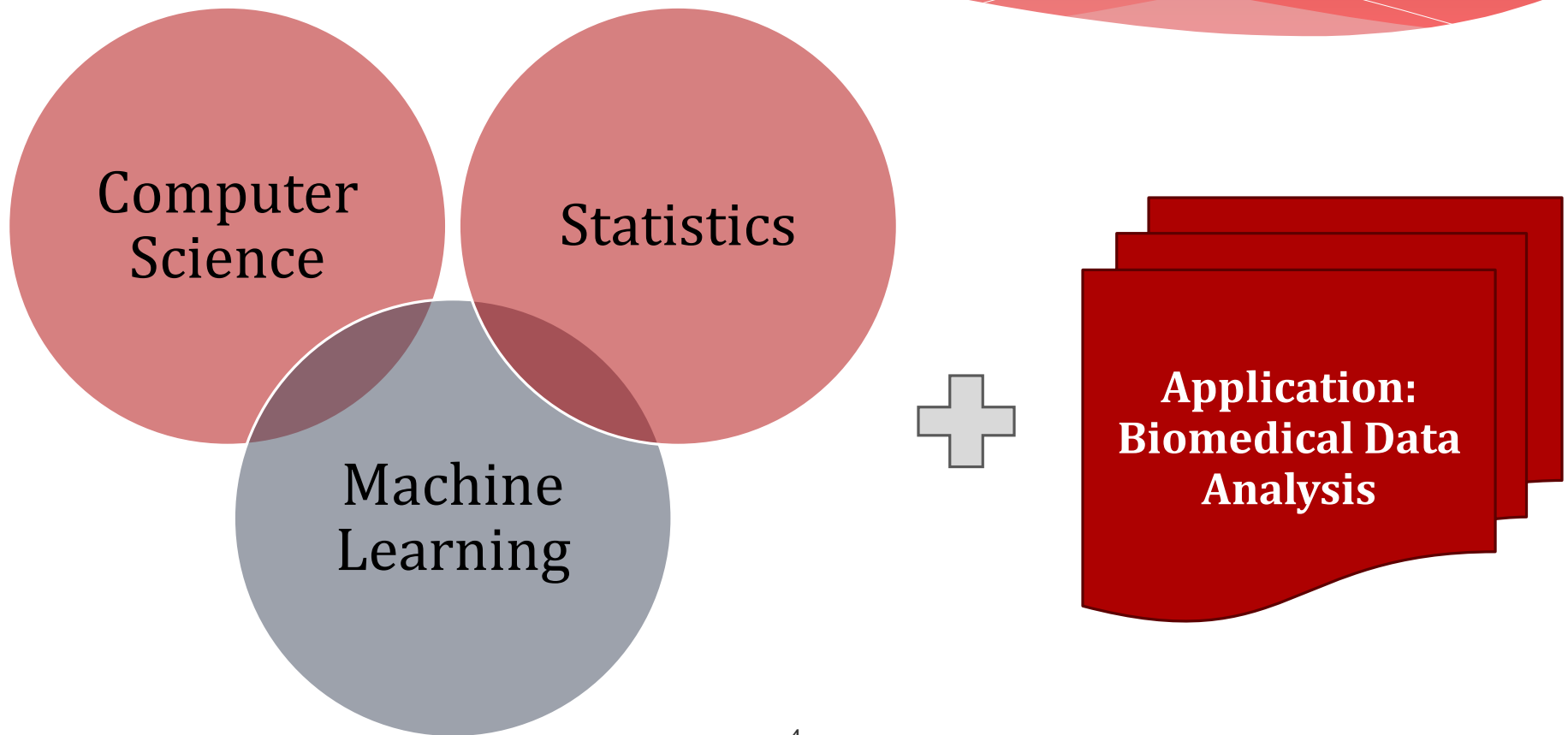


Big Data to Knowledge Initiative

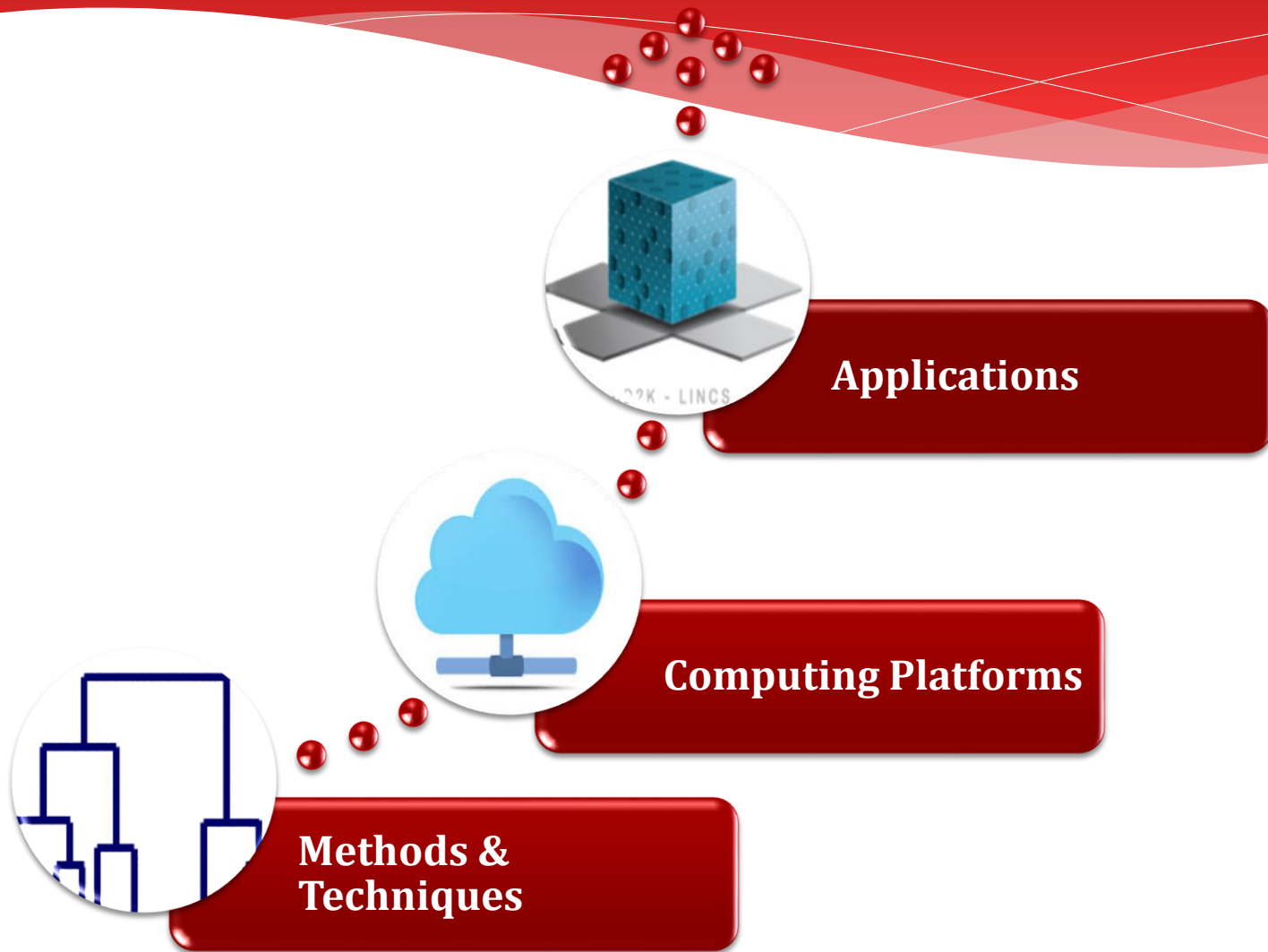
“The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and **insufficient training**, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.”



Big Data & Emergence of Data Science



Three Main Modules



More about the Course

- * Textbook (suggested):
 - * The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani, Jerome Friedman
- * Accounts/devices:
 - * Laptops with R installed locally
 - * Cluster accounts
- * Grading:
 - * 20% Methods homework assignments + 30% Midterm
 - * 50% 'Big projects'

Methods & Algorithms: Topics & Assignments

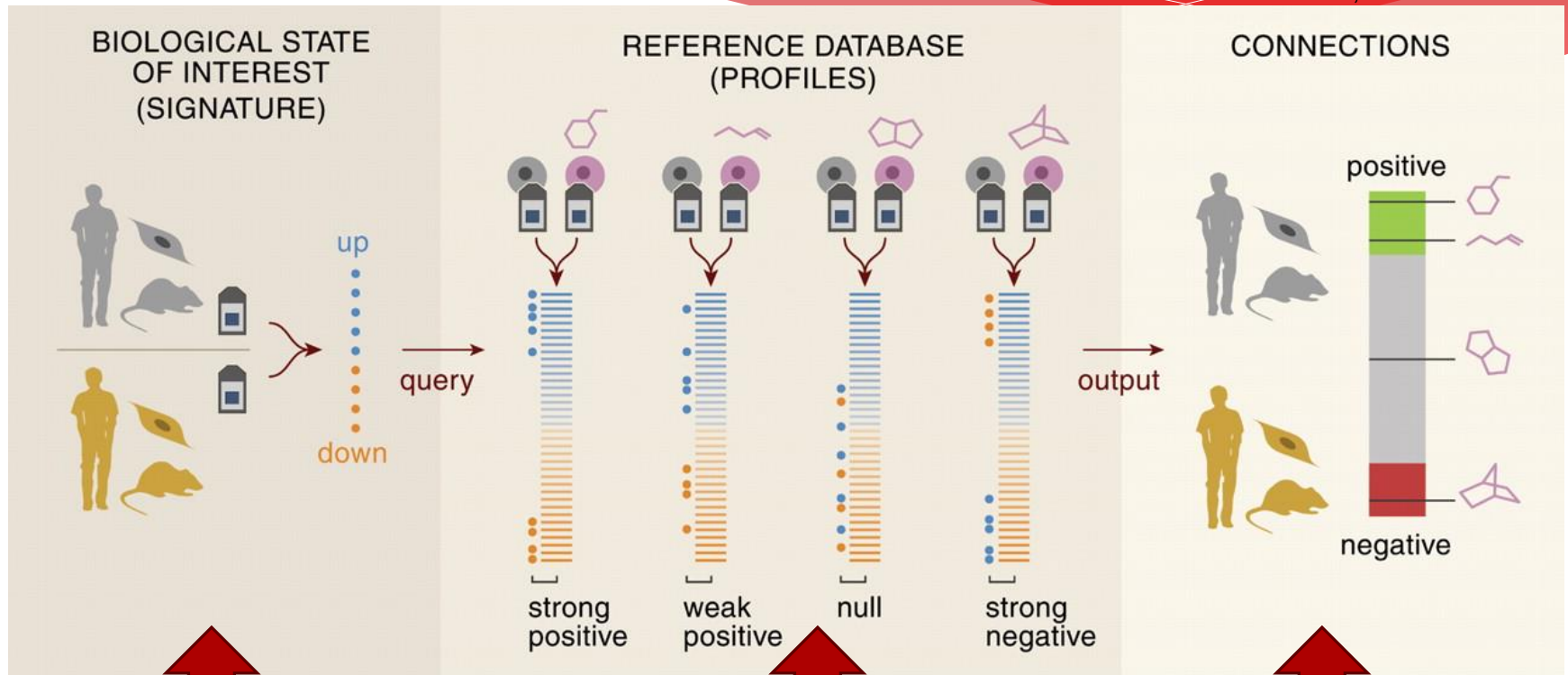
- * Logistic regression
- * Decision trees
- * Random forests
- * Clustering analysis
- * Principal component analysis
- * Midterm

Platforms & Applications: Topics & Projects

- * Cross-validation based assessment of tumor classification accuracy: project 1 (R Markdown)
- * K-means clustering of gene expression data using a Map-Reduce (Hadoop/Spark) framework: project 2 (R Spark)
- * Identification and retrieval of concordant/discordant molecular perturbation signatures using LINCS data: project 3 (R Shiny + iLINCS API)

Using CMAP Perturbation Profiles to Predict Loss/Gain of Function

Illustration from Lamb et al., Science 313




Mutant Expression Signature

KD or Inhibitor Signature

Gain vs. Loss

LINCS: Library of Integrated Network-based Cellular Signatures




NIH LINCS
PROGRAM

LIBRARY OF INTEGRATED NETWORK-BASED CELLULAR SIGNATURES

HOME CENTERS DATA COMMUNITY PUBLICATIONS NEWS

SEARCH

LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents



PTEN

MAP2K4

MAPK8

JUN

FOSL

MYC

CD19

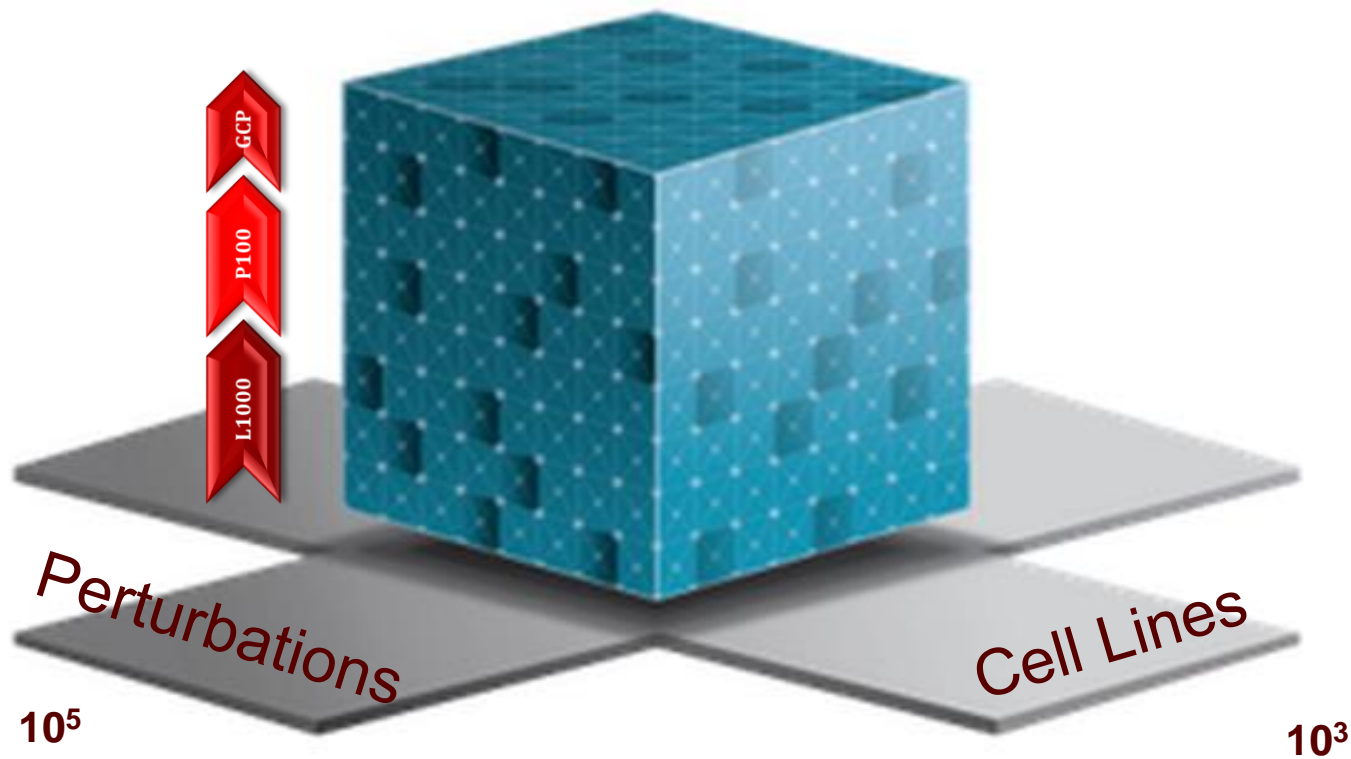
PIK3R1

PIK3CA

AKT1

NFKB1A

LINCS Data Cube

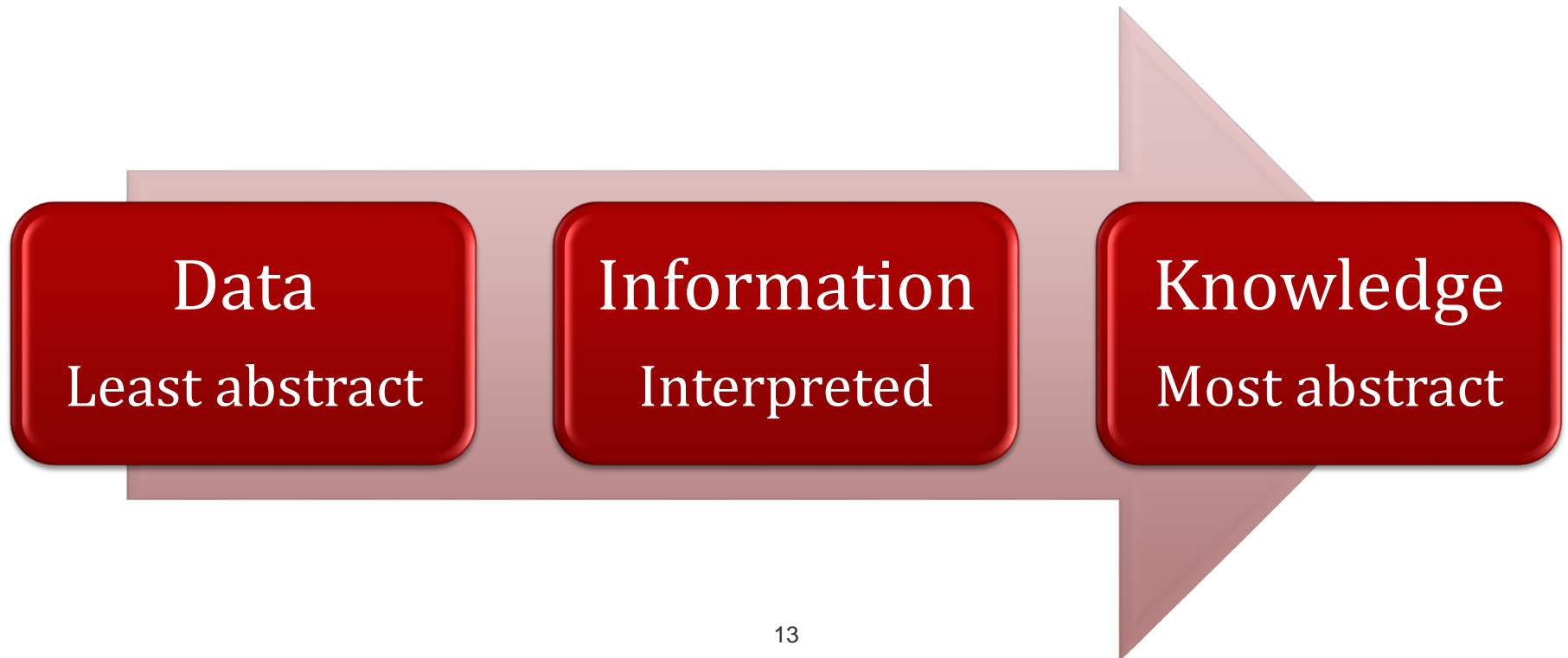


A Very SHORT and Highly INCOMPLETE Introduction to Data Science Topics

Jarek Meller

Departments of Environmental Health and Electrical Engineering &
Computing Systems, University of Cincinnati
Division of Biomedical Informatics, Cincinnati Children's Hospital
Research Foundation

What is data?



What is data?

Data

TCGA mRNA-seq
data sets

Information

PAM50 gene
signature

Knowledge

Molecular
models of cancer

What is data?

Data is a set of values of some qualitative or quantitative variables

Data is interpreted to generate actionable information

What is data?

Data Types	Examples
Numbers	{1.5, 3.1, -2.4, 0.3, 5.6}
Strings	{PTEN, HRAS, TP53, VHL, EGFR1}
Categorical data	{Basal, p53mut, ERneg}
Meta-data	{mRNA-seq assay ver. 1.01}

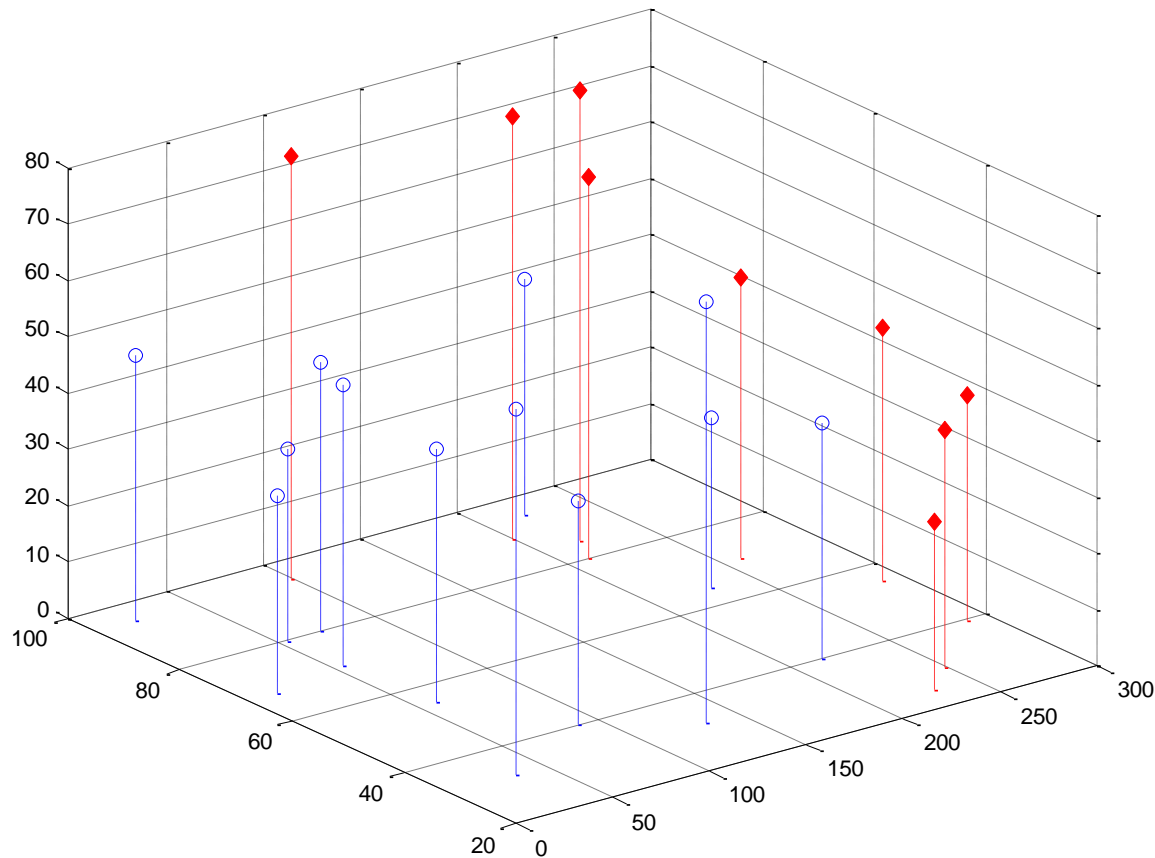
Data Sets Consist of Multiple Instances of Data: HDL vs. LDL Example

Here, for each patient, \mathbf{x}_i , $i=1, \dots, N$, a set of features (attributes and the corresponding measurements on these attributes) are given, including some categorical clinical outcome attributes that together define the vector (row) \mathbf{x}_i :

Age	LDL	HDL	Sex	Clinical outcome
41	230	60	F	no stroke or heart attack
32	120	50	M	stroke within 5 years
45	90	70	M	heart attack within 5 years

$\{ \mathbf{x}_i \}$ $i=1, \dots, N$

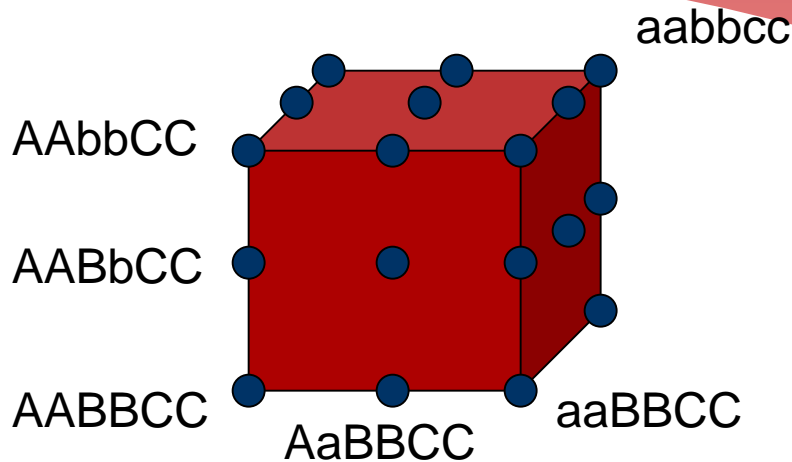
Plotting the Data: Each Feature Adds a Dimension



Healthy controls: blue; Heart attack or stroke within 5 years from the exam: red

• x – LDL; y – HDL; z – age, simulated data based on Westendorp et. al. 2003

Data Dimensionality: Genotypes as Features



In general, 3^n genotypes for n bi-allelic loci.

A – major allele

a – minor allele

AA, aa – homozygous genotypes

Aa – heterozygous genotype

Big Biomedical Data

Omics: 20,000 genes, 200,000 proteins, >10 mln variants

Deep phenotyping, EHRs

Exposures, biomedical surveillance

Personalized precision medicine

Data Science

Data is measured and collected, e.g., using high throughput omics assays (realm of basic science)

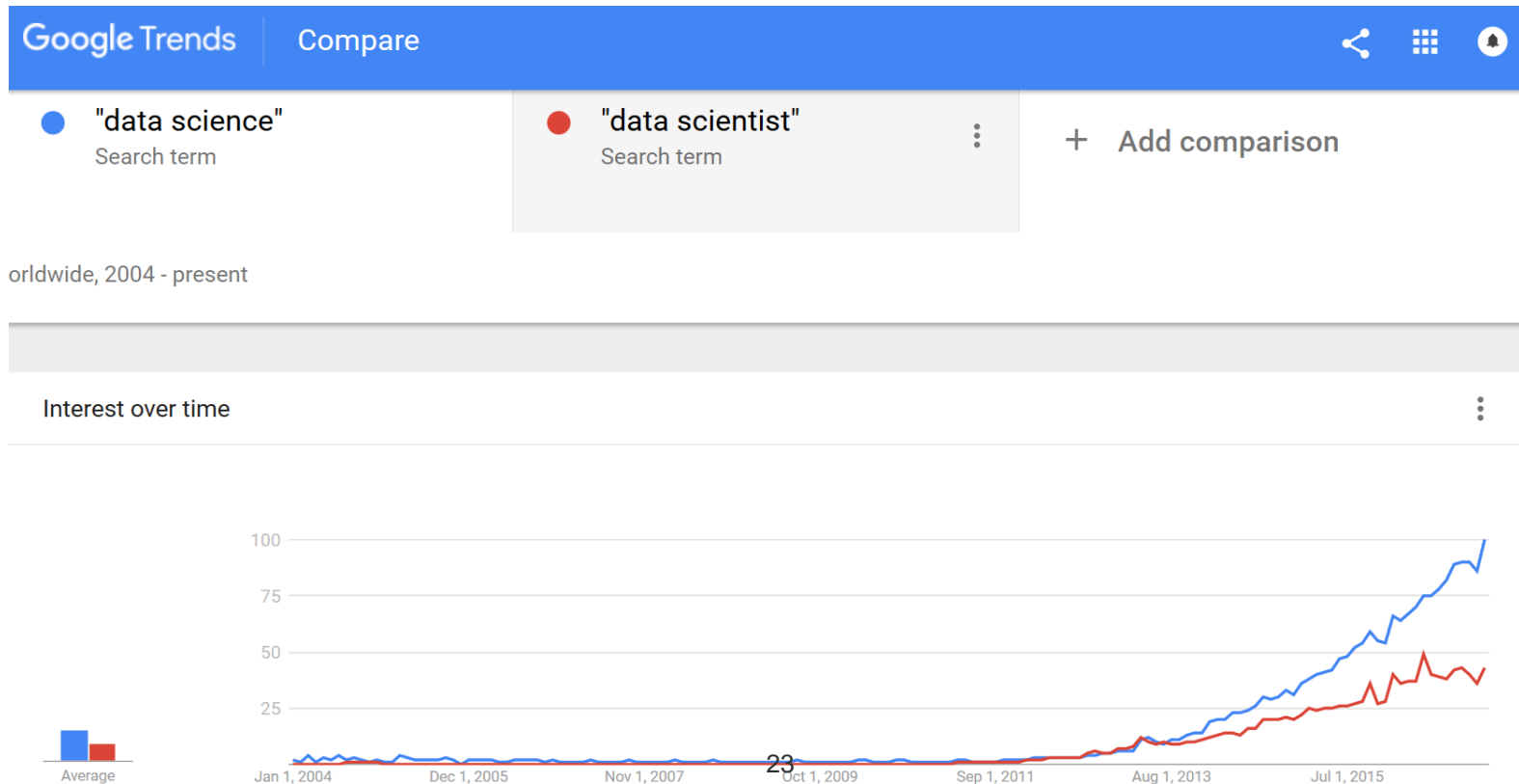
Data is captured, stored, pre-processed and reported (realm of informatics)

Data is analyzed, visualized and interpreted (realm of data science)

Data Science

- * Data science can be explained as a field that emerged at an intersection of machine learning, computer science and statistics in response to the rise of data (Chris Wiggins)
- * Data analysts + research scientist = data scientist (Jeff Hammerbacher, Facebook)
- * Importance of some specific field of applications and expertise in that field
- * Some founding fathers:
 - * John W. Tukey, box plot, FFT, S language, exploratory data analysis, Princeton (1960s)
 - * Edward R. Tufte, Princeton (1970s)
 - * William S. Cleveland, Bell Labs (2000s)

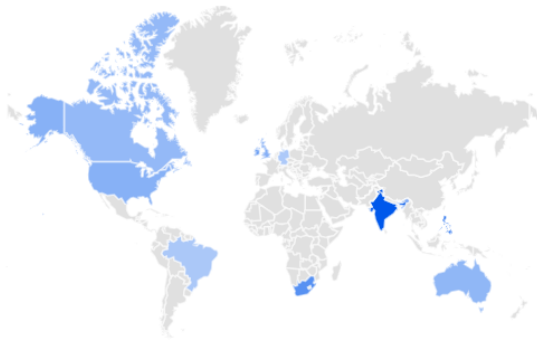
Emergence of Data Science



Emergence of Data Science

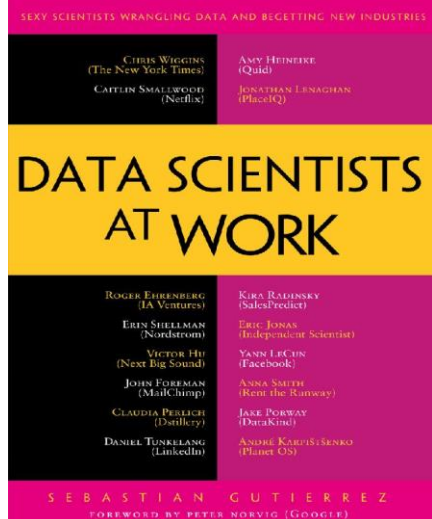
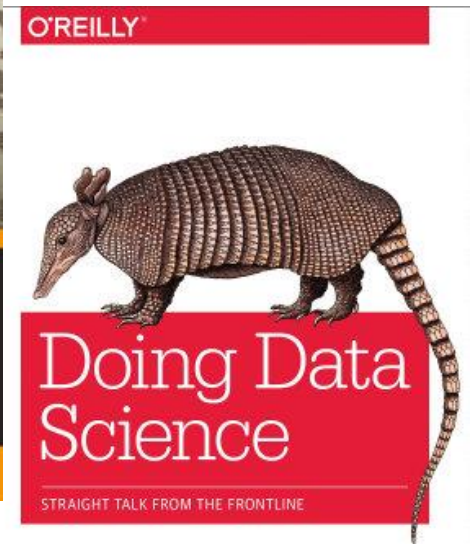
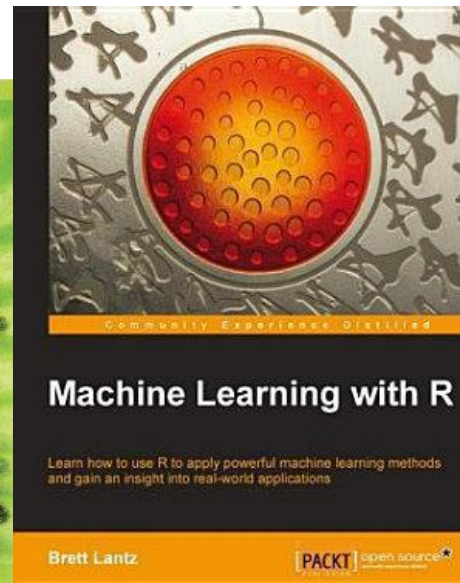
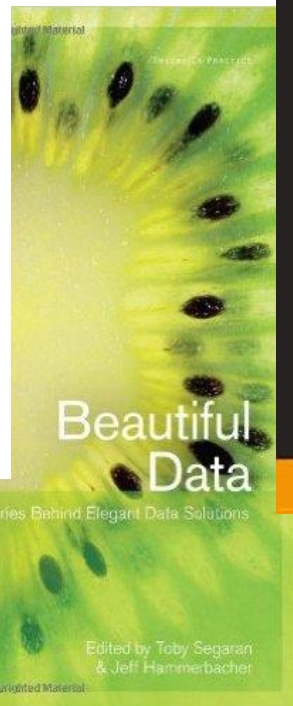
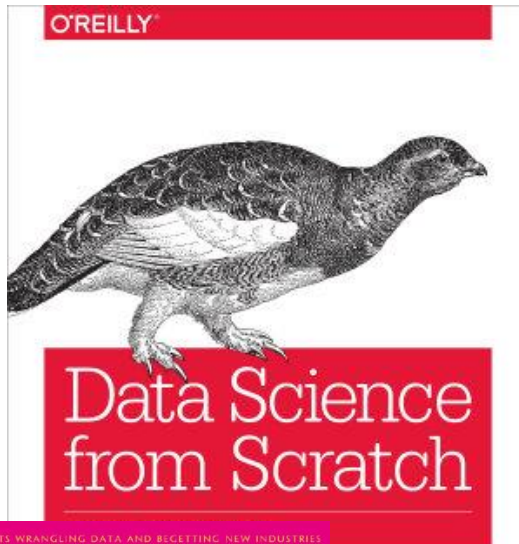
Interest by region

Region ▼



1	India	100	<div></div>
2	Philippines	96	<div></div>
3	South Africa	51	<div></div>
4	Singapore	46	<div></div>
5	Ireland	41	<div></div>

Emergence of Data Science



Machine Learning

- * Unsupervised learning: explore & “discover” patterns, descriptive modeling
- * Supervised learning: learn & test a predictor, predictive modeling
- * Reinforcement learning: optimize & reconfigure, prescriptive modeling

Machine Learning

- * Unsupervised learning: what and why happened?
- * Supervised learning: what will happen?
- * Reinforcement learning: how can we make it happen?

Learnings:

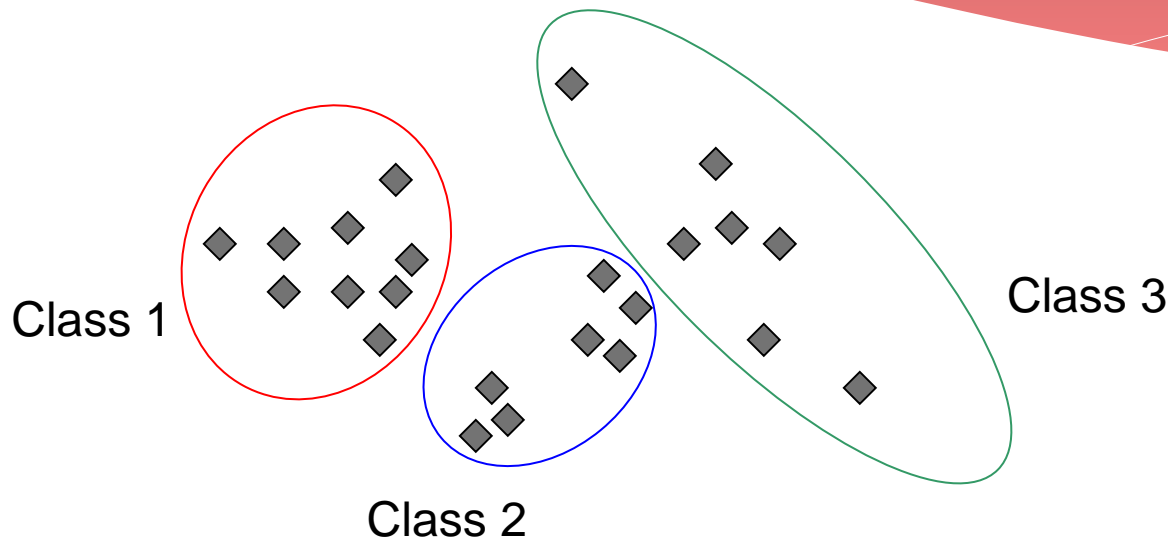
After Chris Wiggins

- * Descriptive modeling: specify x , learn $z(x)$ or $p(z/x)$ where z is “simpler” than x
- * Predictive modeling: specify x and y , learn to predict y from z , or in other words learn a predictor $z(x)$ such that $y=z(x)$
- * Prescriptive modeling: specify x, y and a , learn to prescribe a to maximize y given x

Unsupervised vs. Supervised Learning

- * Concepts and some reminders
- * Models and algorithms
- * Examples of applications

Pattern Recognition through Unsupervised Learning

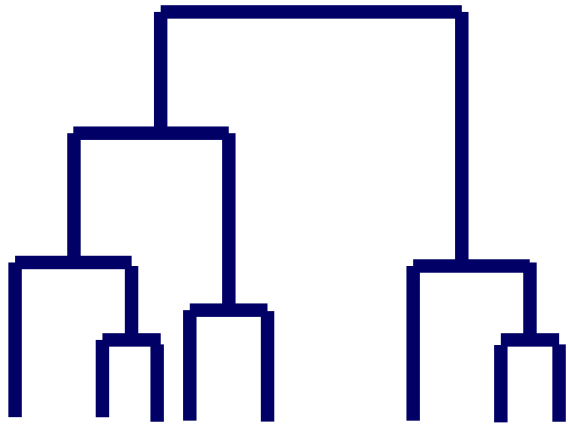


The goal of unsupervised learning is to “discover” the structure (patterns) in the data and group (cluster) similar objects, given a suitable similarity measure.

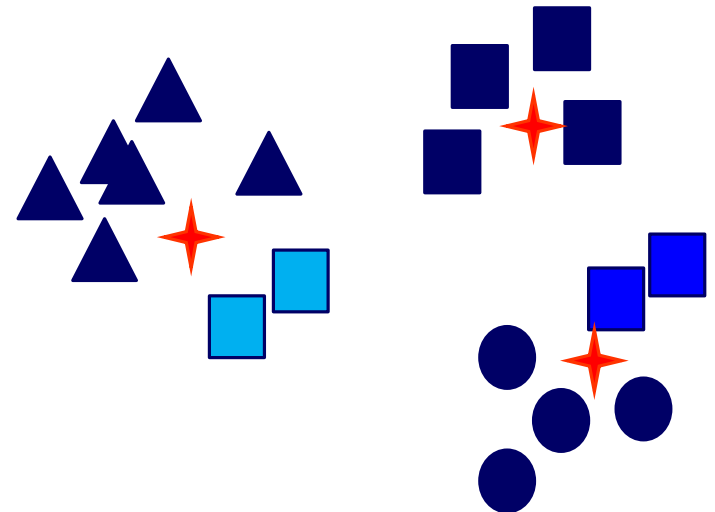
Hierarchical Clustering vs. K-means

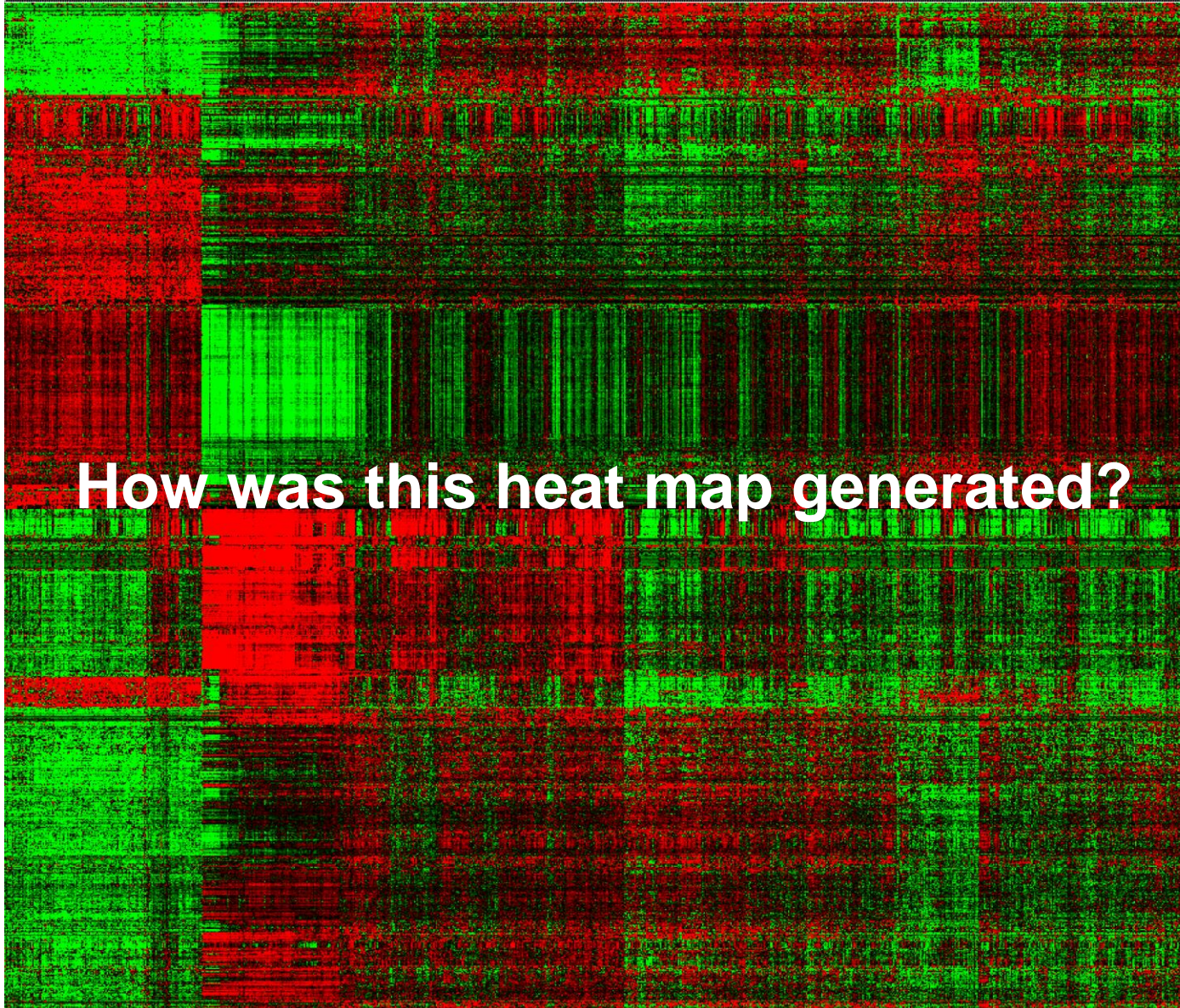
D_{ij}

Iterate over putative K centroids;
How many clusters are there?



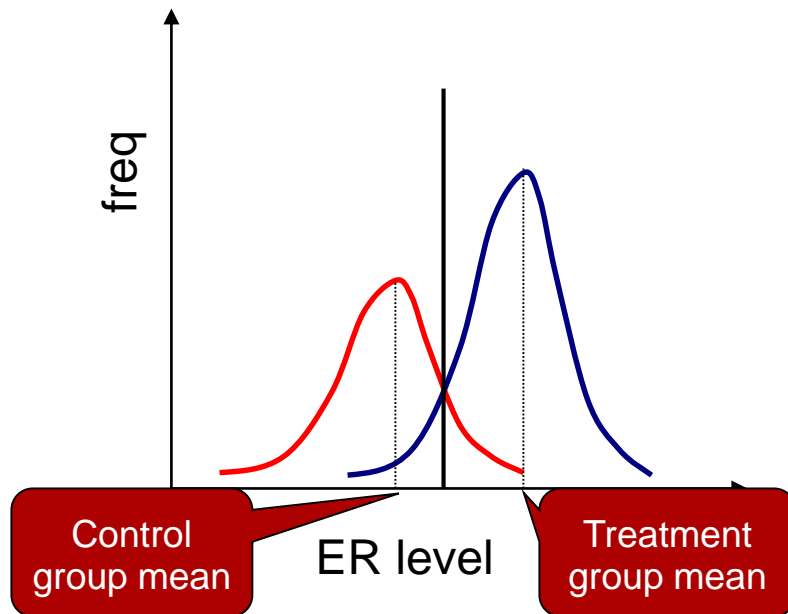
Pairwise distance matrix as
memory and time bottleneck





How was this heat map generated?

Differentially Expressed Genes



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

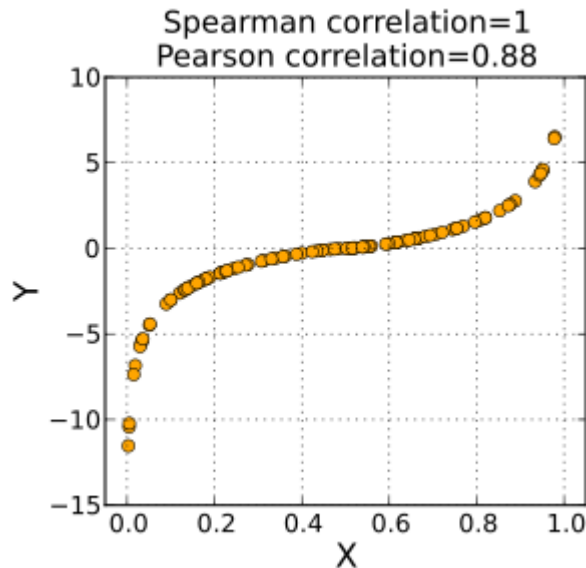
$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Choice of features (genes) based on a test of statistical significance of observed expression levels in 2 (or more) groups

Similarity Measures for Omics Data

PCC

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



SCC

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

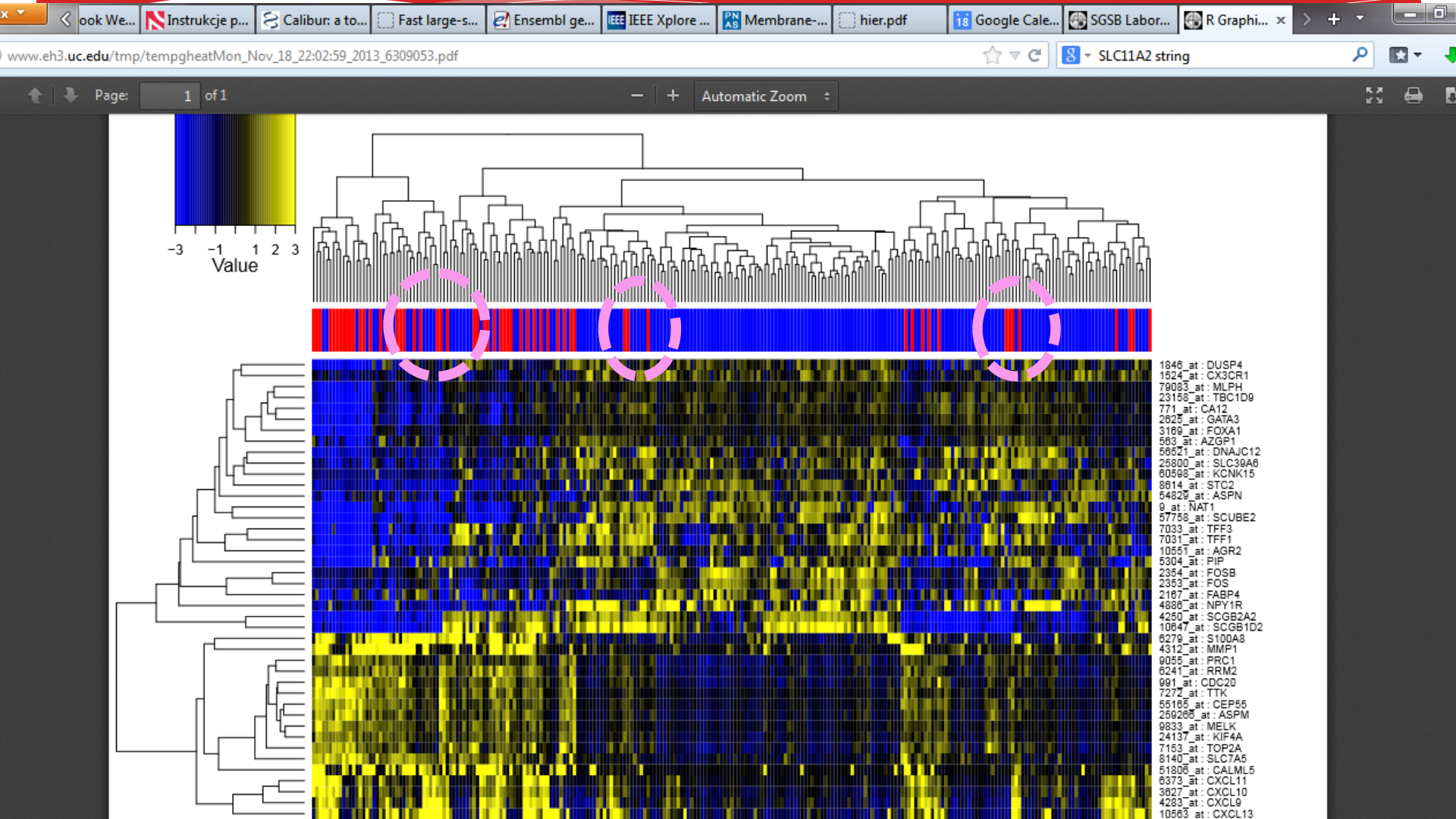
Clustering Algorithms

- * Choice of data pre-processing and normalization
- * Choice of the actual distance (or dissimilarity) measure, e.g., Euclidean vs. Pearson vs. Cosine
- * Choice of the agglomerative heuristic, e.g., average vs. maximum (complete) vs. minimum (single) distance linkage

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

$$d_{X,Y} = 1 - \rho_{X,Y}.$$

Discovering Interesting Patterns: P53 Signature



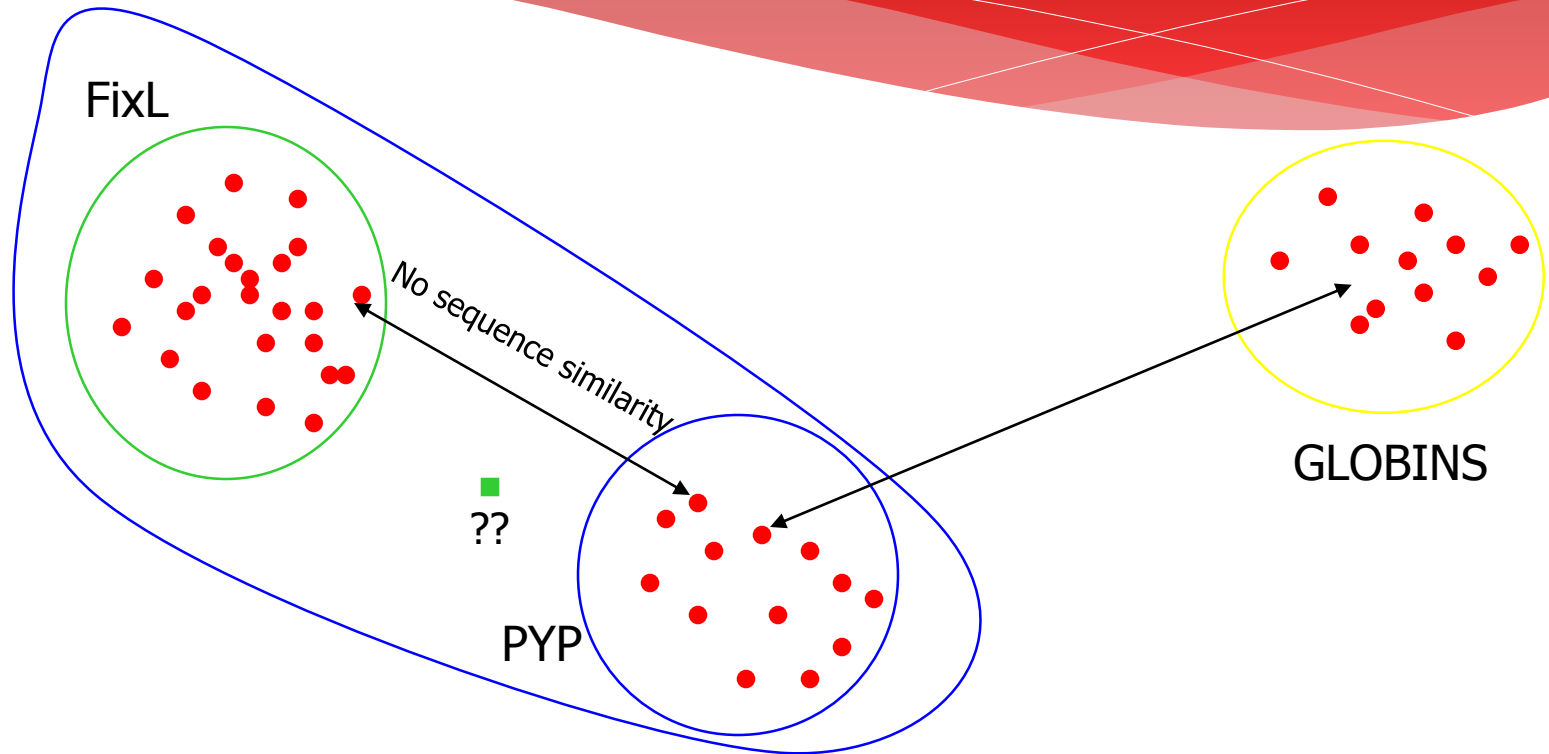
Supervised Learning with Labeled Data Points: Prediction and Decision Making

Supervised learning (or learning with a teacher) extrapolates from a set of examples with class assignments (e.g. healthy vs. diseased).

The goal is to find a suitable *representation of the problem* in some feature (attribute) space that can separate the imposed classes.

Such obtained classifiers with the resulting decision boundaries may be subsequently used to make *prediction* for new data points.

Advantages and Pitfalls of Prior Knowledge: Supervised Learning with Class Assignment



Structural data suggests the same class despite low sequence similarity; How to find a “good model” question again ...

Three Stages of Supervised Learning

Training data:

- Examples with class assignment are given

Learning:

- Appropriate model (or representation) of the problem needs to be selected in terms of attributes (features), distance measure and classifier type;
- Adaptive parameters in the model need to be optimized to provide correct classification of training examples (e.g. minimizing the number of misclassified training vectors)

Validation:

- Cross-validation, independent control sets and other measure of “real” accuracy and generalization should be used to assess the success of the model and the training phase (*finding trade off between accuracy and generalization is not trivial*)

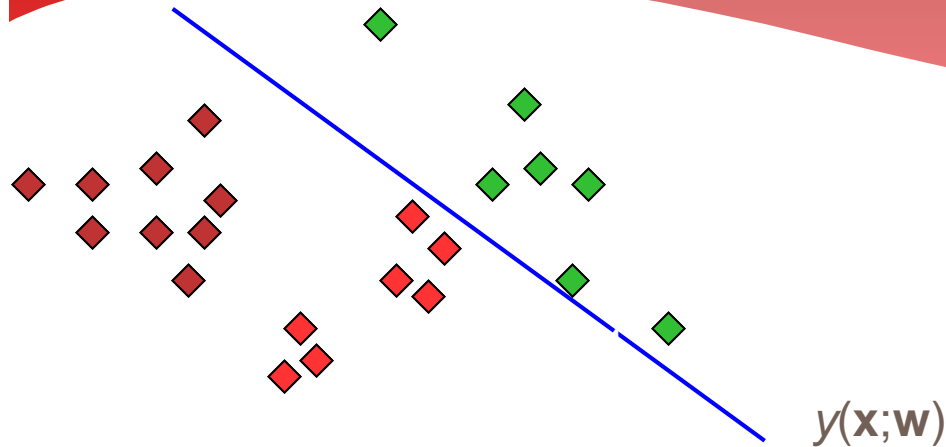
Training Set and Class Assignment: HDL/LDL vs. Cardiovascular Risk

A set of objects (here patients) \mathbf{x}_i , $i=1, \dots, N$ is given. For each patient a set of features (attributes and the corresponding measurements on these attributes) are given too. Finally, each patient is assigned to one of the classes C_k , $k=1, \dots, K$.

Age	LDL	HDL	Sex	Class
41	230	60	F	healthy (0)
32	120	50	M	stroke within 5 years (1)
45	90	70	M	heart attack within 5 years (1)


$$\{\mathbf{x}_i, C_k\} \quad i=1, \dots, N$$

Learning from Data: Optimizing Adaptable Parameters in the Model

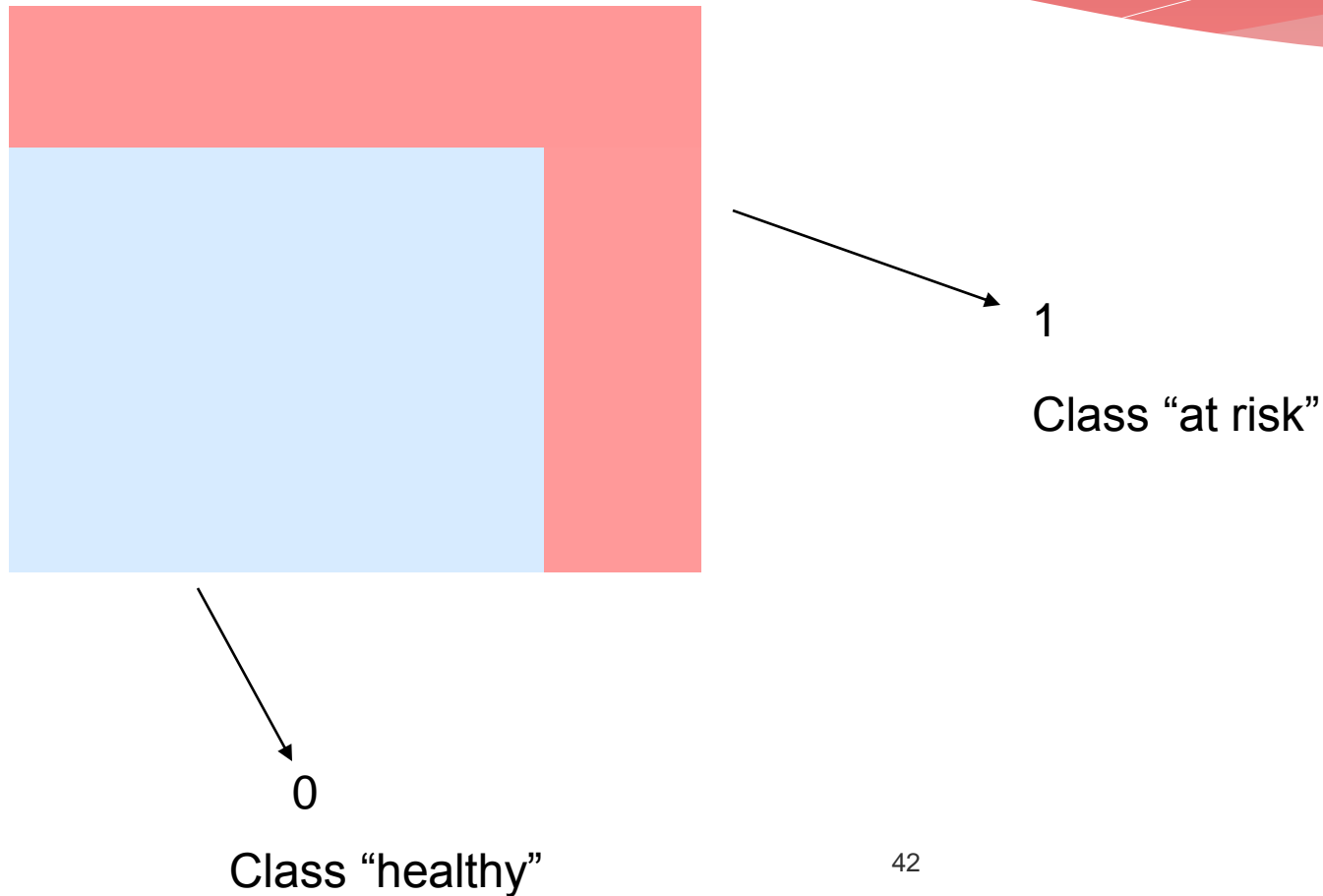


Find a model $y(\mathbf{x}; \mathbf{w})$ that describes the objects of each class as a function of features \mathbf{x} and adaptive parameters (weights) \mathbf{w} .

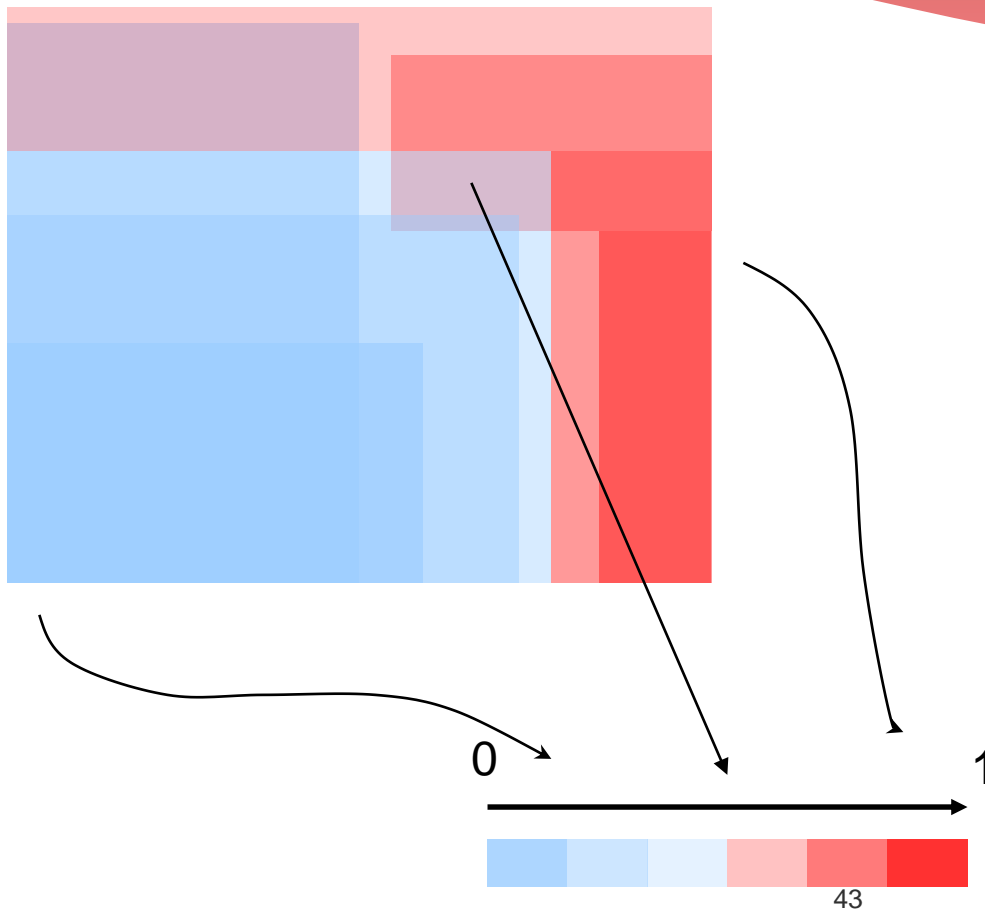
Prediction: given \mathbf{x} (e.g. LDL=240, age=52, sex=male) assign the class $C=?$

For example, if $y(\mathbf{x}, \mathbf{w}) > 0.5$ then $C=1$; here likely to suffer from a stroke or heart attack in the next 5 years.

Categorical Outcomes: Classification Approach



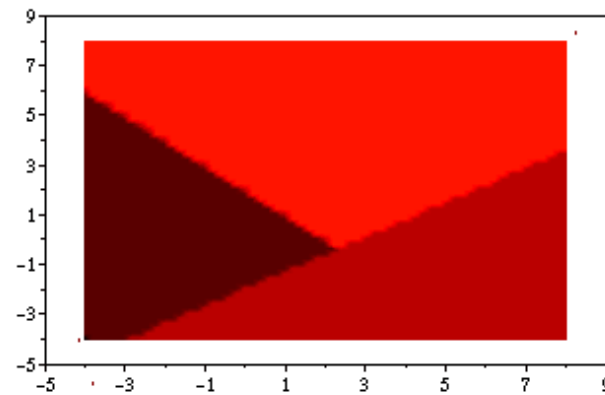
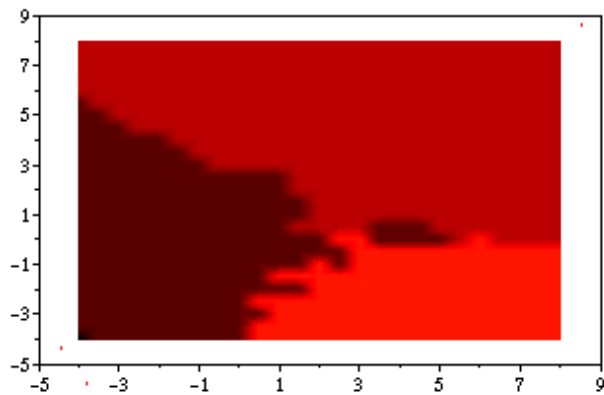
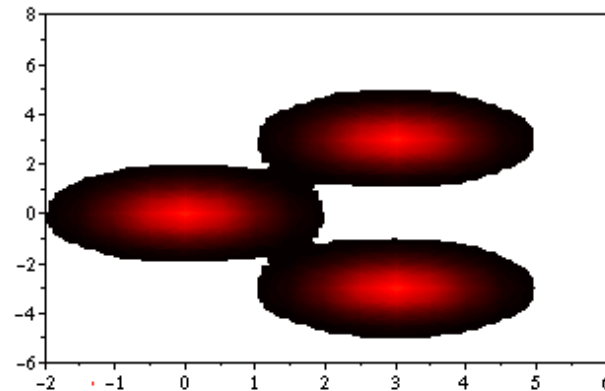
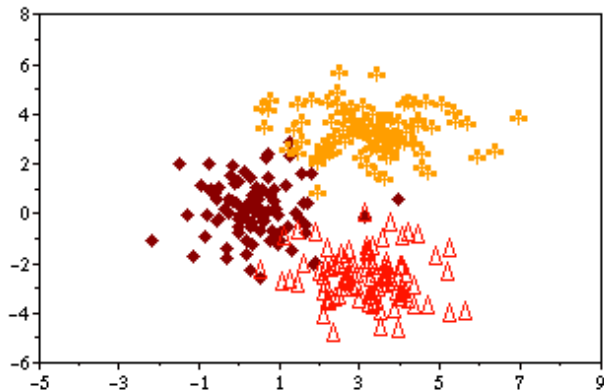
Continuous Outcomes: Regression Approach



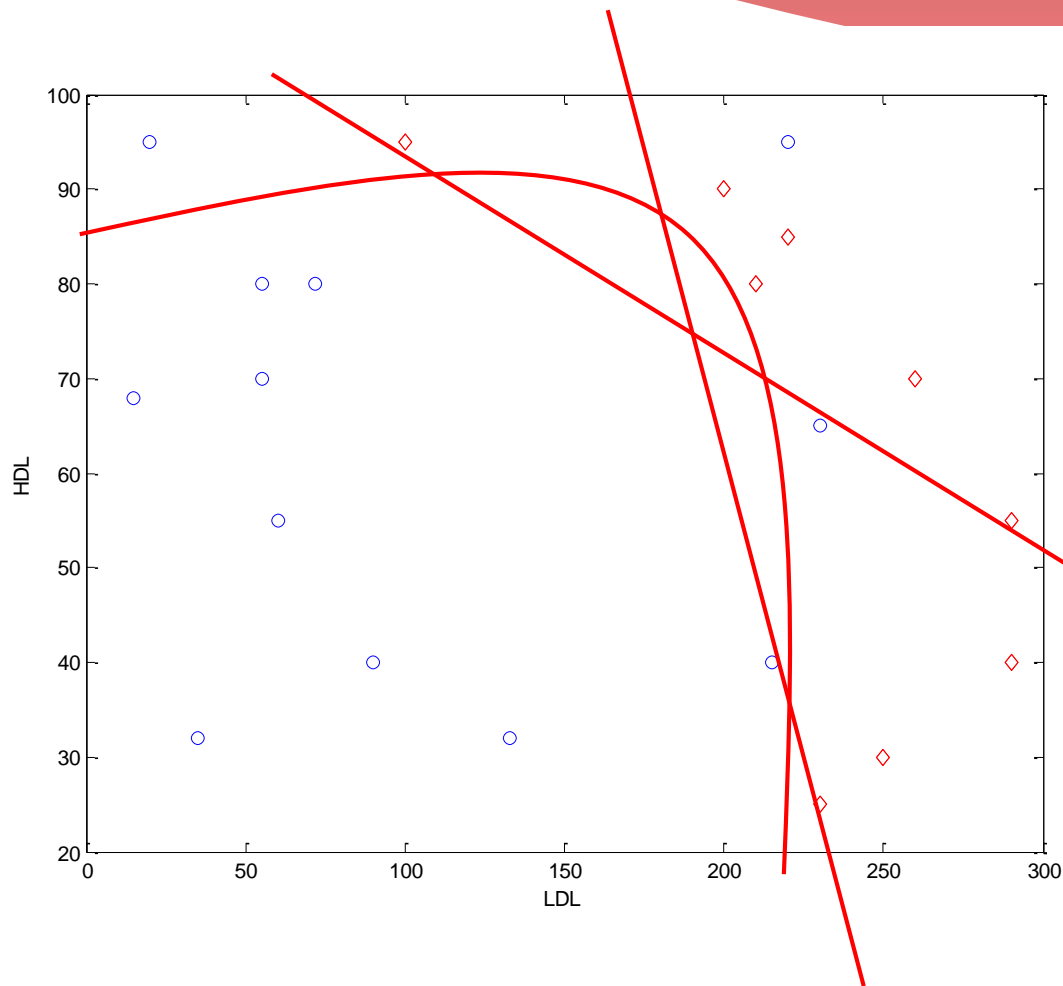
Machine Learning Algorithms for Classification and Regression Problems

- Classical approaches: Linear Perceptron, Least Squares and Logistic Regression
- LDA/FDA (Linear/Fisher Discriminate Analysis): simple linear cuts, kernel non-linear generalizations
- SVM (Support Vector Machines): optimality guarantees, wide margin linear cuts, kernel non-linear generalizations
- Decision trees: recursive partitioning, intuitive logical rules
- k-NN (k-Nearest Neighbors): simple and non-parametric
- Neural Networks: general non-linear models, flexibility and adaptivity, “artificial brain”

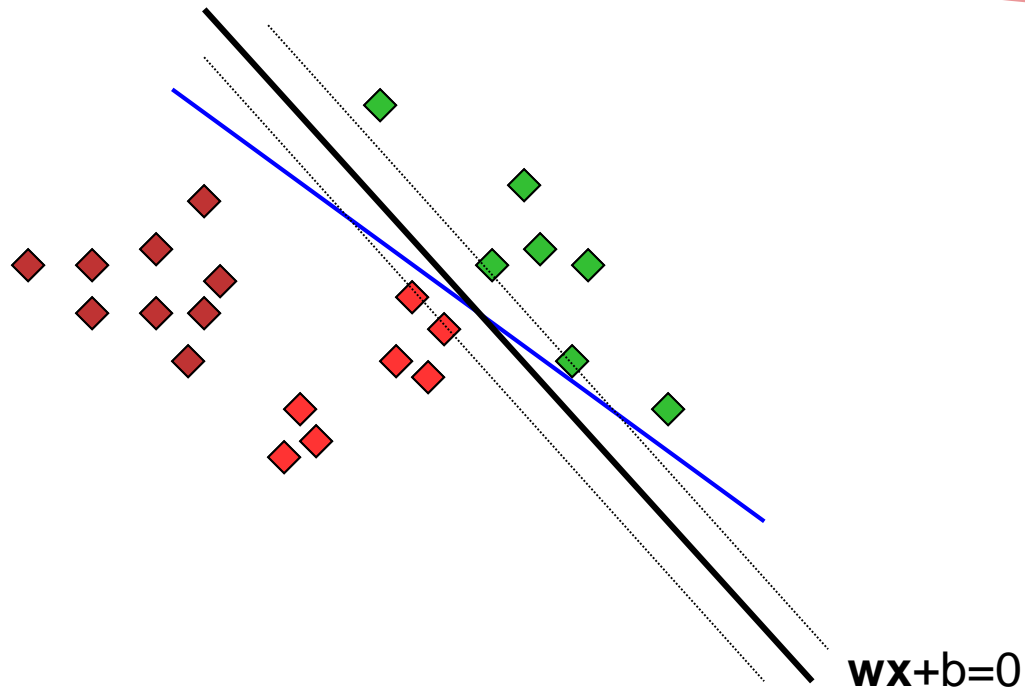
Validation: Training Accuracy vs. Generalization



LDL Example: Finding Decision Boundaries to Make Prediction for New Data Points

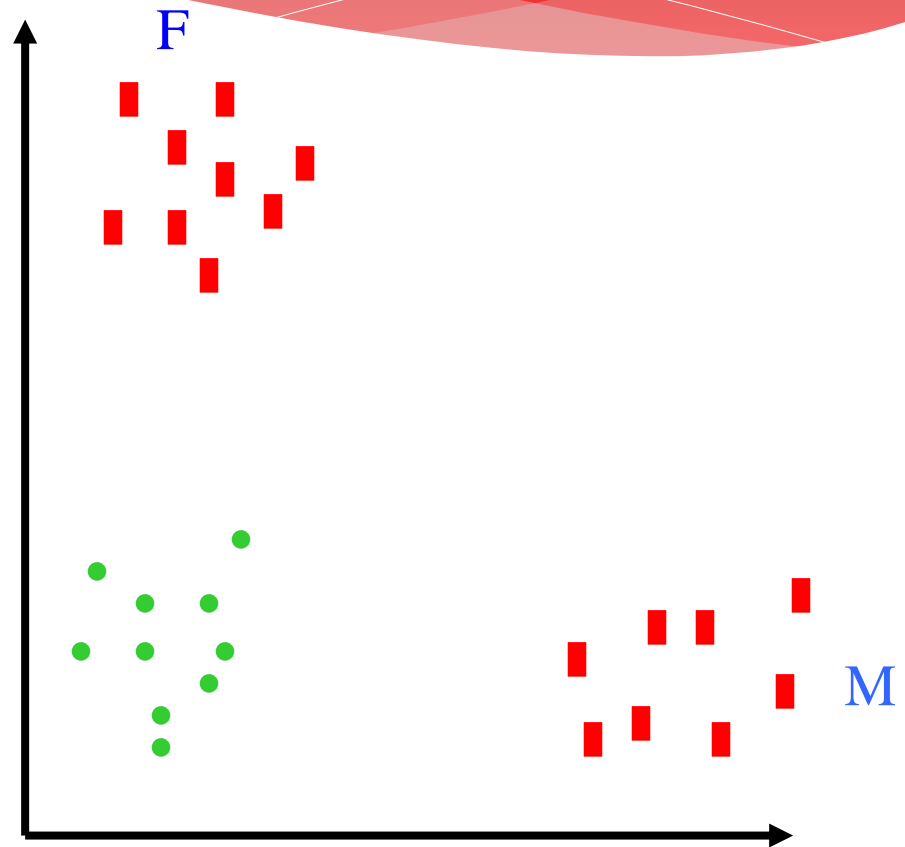
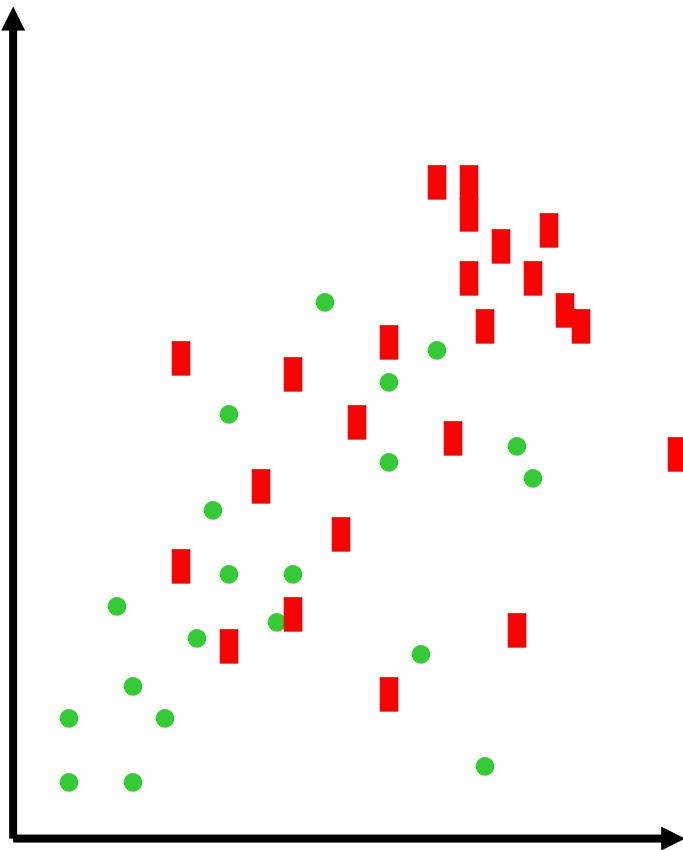


Support Vector Machines: Good Generalization with a Wide Margin Solution

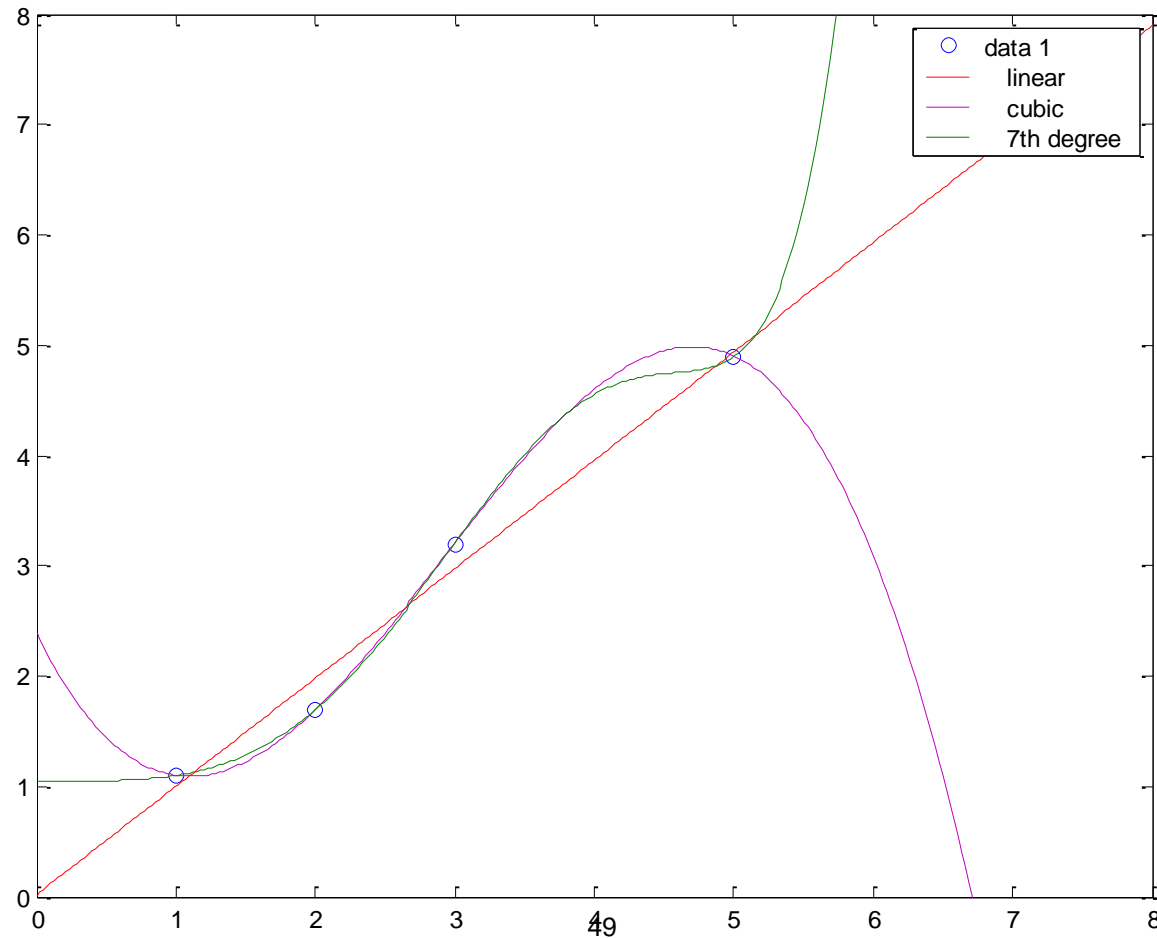


Choice of the Model: Representation of the Problem and Feature Selection

■ adults ● children



Model Complexity and Generalization : Training Set Size vs. Number of Parameters



Curse of Dimensionality

20,000 genes

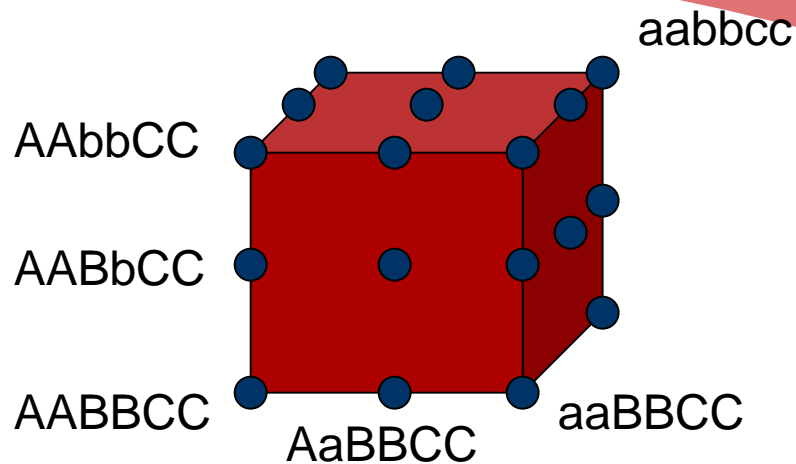
200,000 proteins

10,000,000 variants

Pairs, triplets and other combinations

$$\begin{pmatrix} N \\ K \end{pmatrix}$$

Curse of Dimensionality: Combination of Genotypes



In general, 3^n genotypes for n bi-allelic loci.

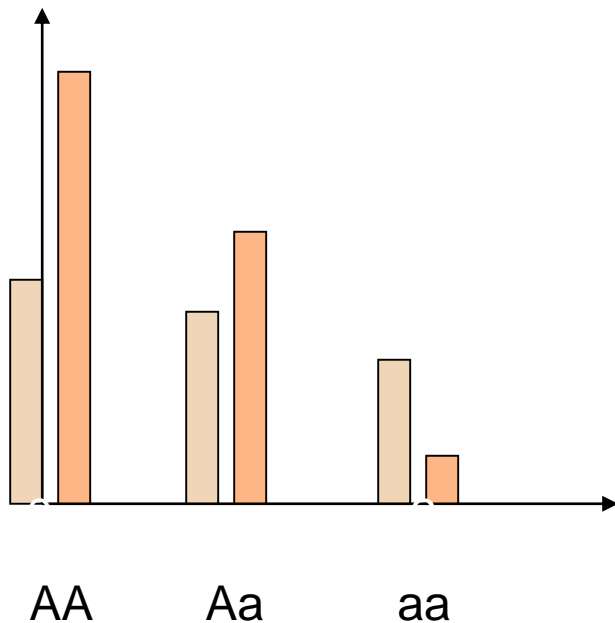
A – major allele

a – minor allele

AA, aa – homozygous genotypes

Aa – heterozygous genotype

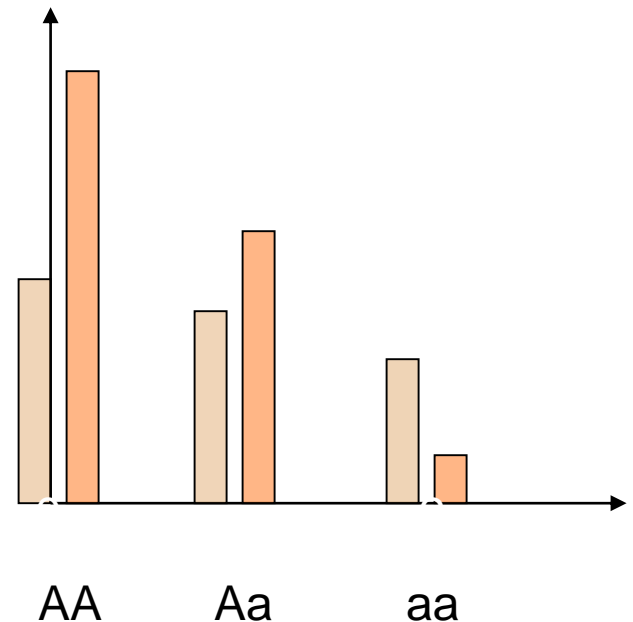
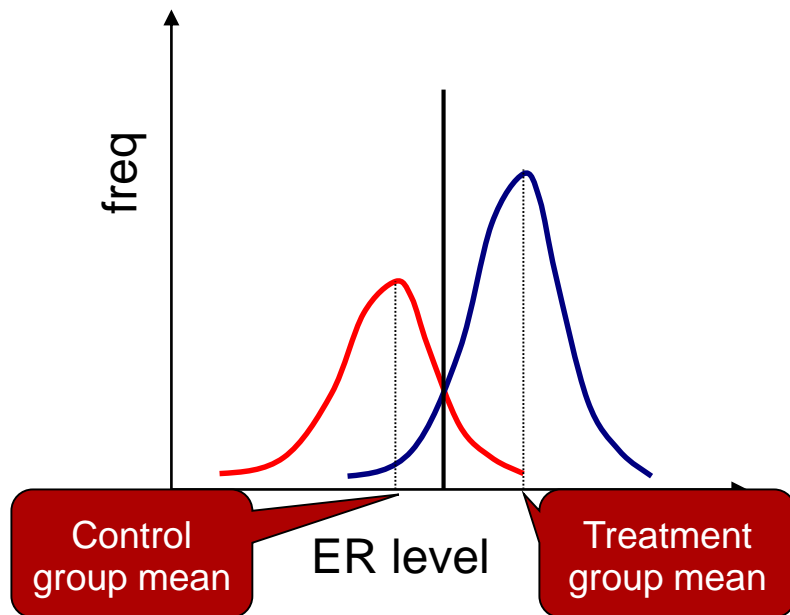
Testing for Association with Phenotypes



$$\Sigma (O-E)^2 / E$$

Testing for the effect of individual SNPs using χ^2 test.

Numerical vs. Categorical Variables



Differences in means vs. differences in counts:

• t-Test vs. Chi²-Test

Testing for Association with Phenotypes

	High Risk	Low Risk	<i>total</i>
AA	27	24	51
Aa	36	38	74
aa	21	24	45
<i>total</i>	84	86	170

$$\Sigma (O-E)^2 / E$$

No association here and small value of Chi2.

Testing for Association with Phenotypes

	High Risk	Low Risk	<i>total</i>
AA	47	14	61
Aa	26	48	74
aa	11	24	35
<i>total</i>	84	86	170

$$\Sigma (O-E)^2 / E$$

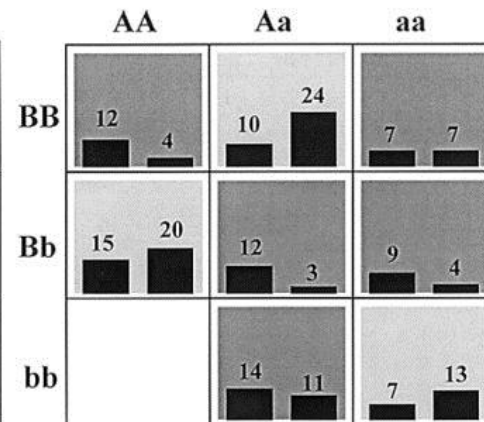
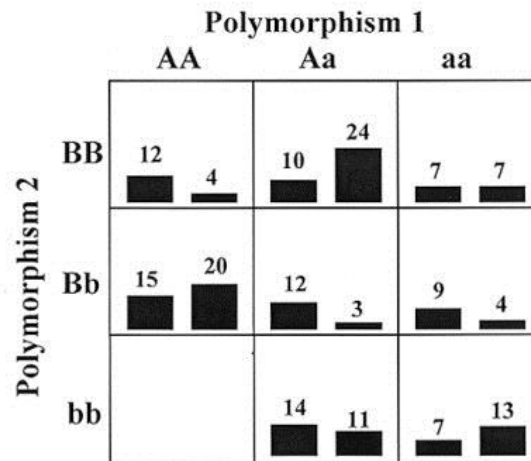
“Strong” effect of an individual SNP in this case.

Identification of Predictive Loci Using MDR

Ritchie et al.

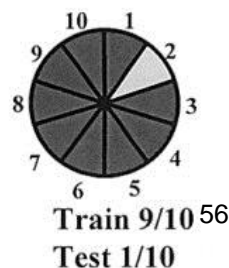
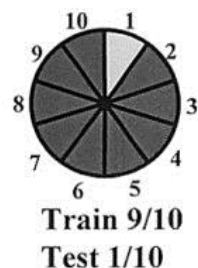
STEP 1 : Select Polymorphisms → STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype → STEP 3 : Identify High-Risk Multilocus Genotypes

Polymorphism 1
Polymorphism 2
Polymorphism 3
Polymorphism 4
...
...
Polymorphism 10

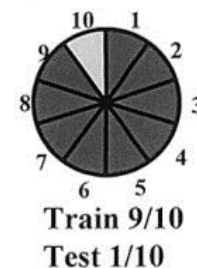


High-Risk
Low-Risk
Empty Cell

STEP 4 : Cross Validation

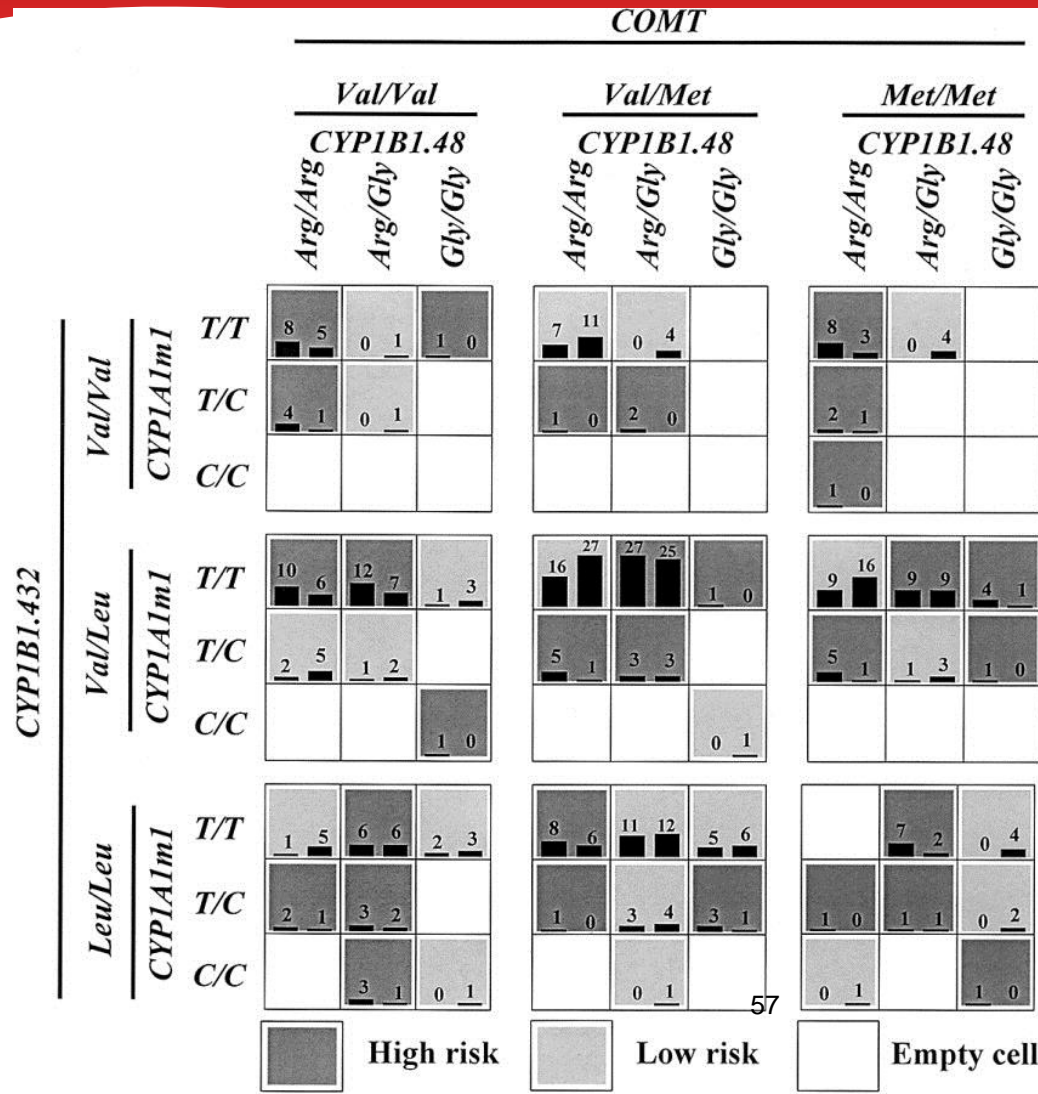


...



Multiple Loci and More Complex Fingerprints

Ritchie et al.



Complexity of the Model and Power Calculations

Ritchie et al.

In logistic regression, as each additional main effect is included in the model, the number of possible interaction terms grows exponentially.

On the other hand, simulation studies by Peduzzi et al. ([1996](#)) suggest that having fewer than 10 outcome events per independent variable can lead to biased estimates of the regression coefficients.

Hosmer and Lemeshow ([2000](#)) suggest that logistic-regression models should contain no more than $P < \min(n_1, n_0)/10$ parameters, where n_1 is the number of events of type 1 and n_0 is the number of events of type 0.

For the 200 cases and the 200 controls evaluated in the present study, this formula suggests that no more than 19 parameters should be estimated in a logistic-regression model.