FORSCHUNGS CAMPUS

öffentlich-private Partnerschaft für Innovationen

SPONSORED BY THE

Federal Ministry of Education and Research

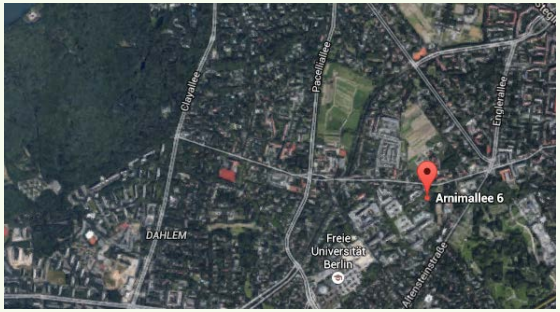
Omics-based Cancer Diagnostics

Tim Conrad, Freie Universität Berlin

Forschungs Campus MODAL, **MedLAB**

About me…

- Studied BioInformatics & Computer Science
- Ph.D. Mathematics
- Professor for Medical Bioinformatics
  @ Freie Universität Berlin
- Current research topics:
  - Network & compressed-sensing based analysis methods (Einstein Center for Mathematics)
  - Frameworks for mass-data processing (Berlin Big Data Center)
  - Analysis of bio-medical mass-data (Forschungscampus MODAL)



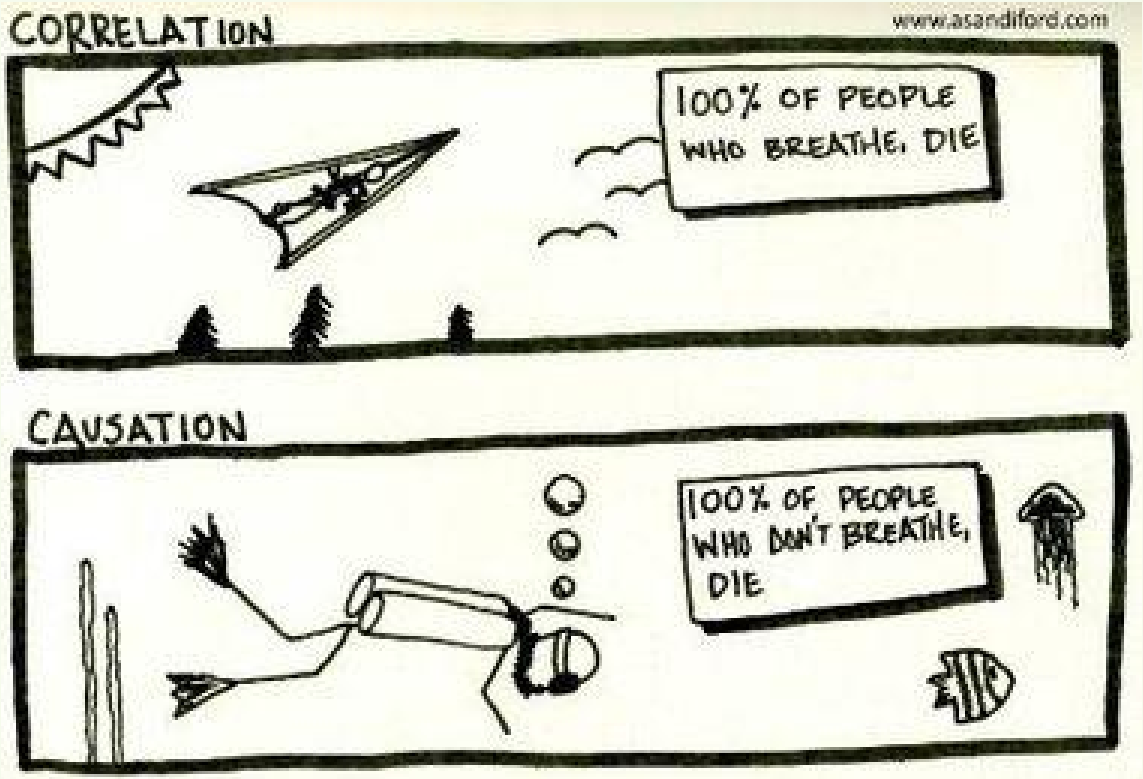
After this talk you should know…

- … what -omics data is
- … that -omics data can be very large
- … why it is useful to analyze -omics data
- … a method how to analyze large -omics data
- … a framework implementing this method

Spoiler: we will do this during the hands-on session


ACHTUNG

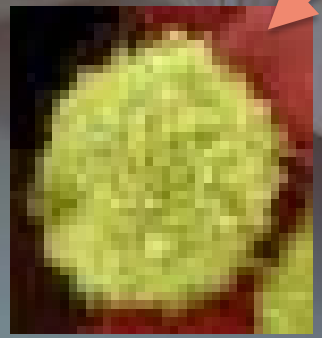
Correlation is not Causation!




example

TRAINING

APPLICATION



YES OR NO?

Introduction

Background & Goal

Omics Data Analysis

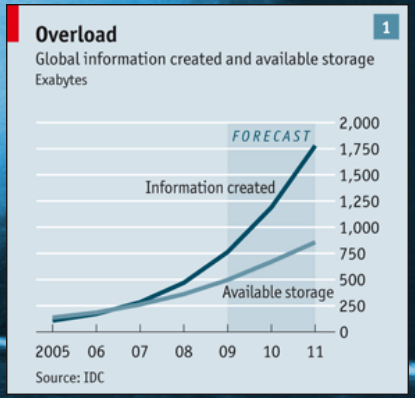
How can it be done?

Application

Reaching out
to the real world

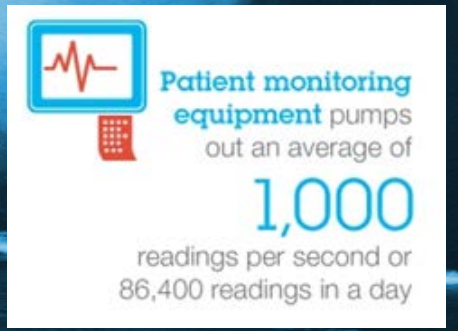

Part 0:
Big Data
in Healthcare

Healthcare is challenged by large amounts of data in motion that is diverse, unstructured and growing exponentially.



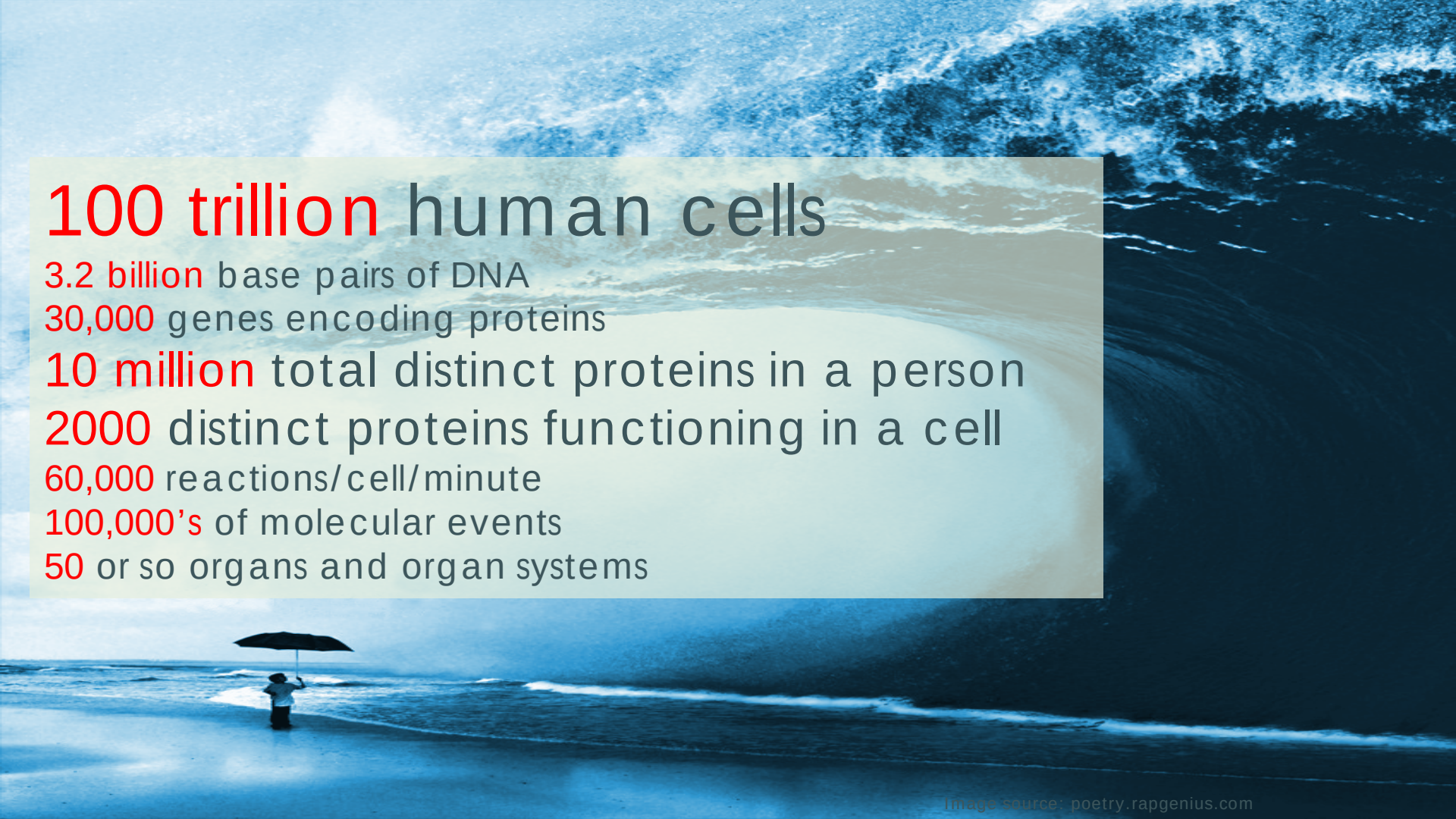
Source: iHT^2
Image source: Flickr

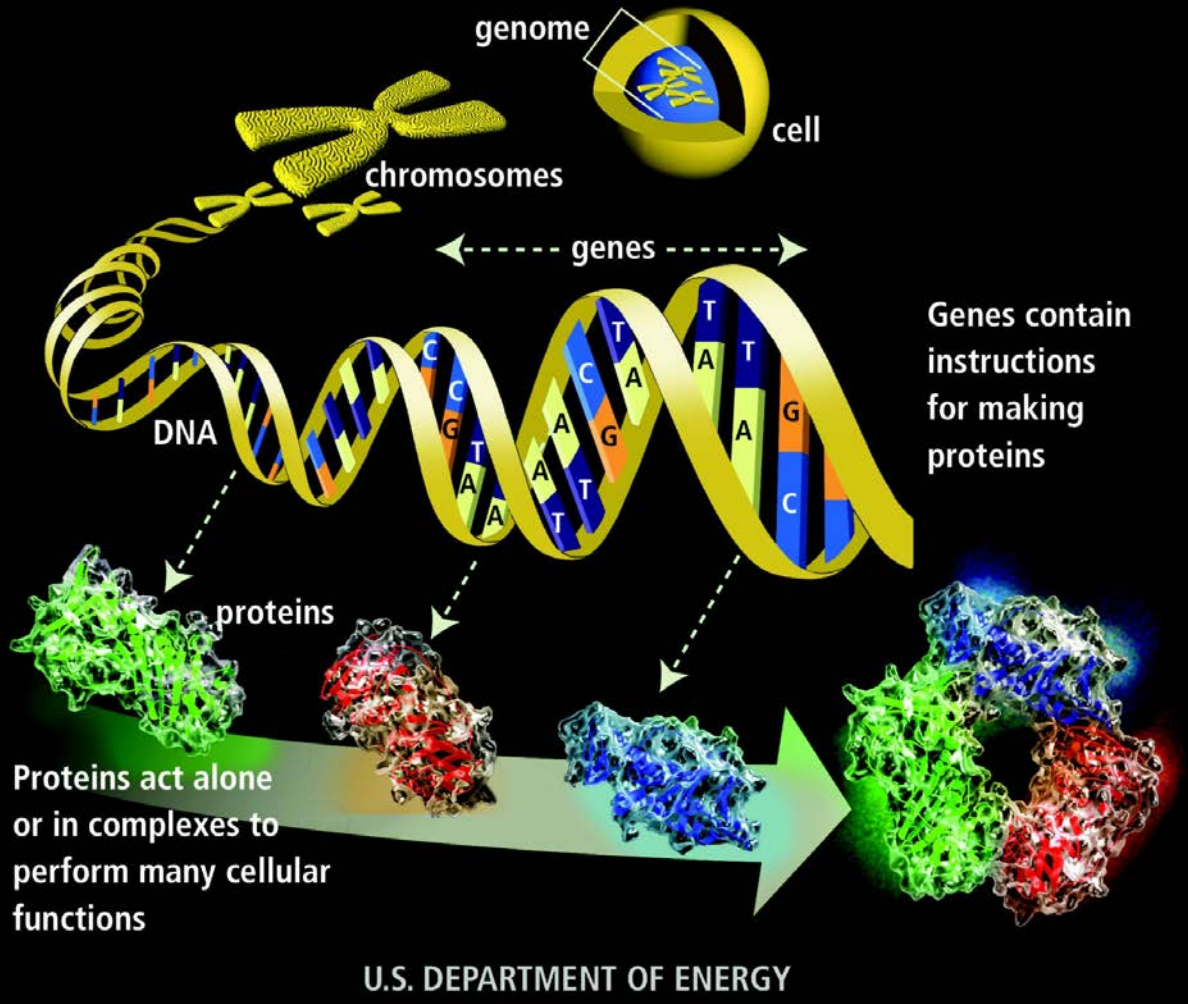
Data constantly streams in through interconnected sensors, monitors and instruments in real-time faster than a physician or nurse can keep up.

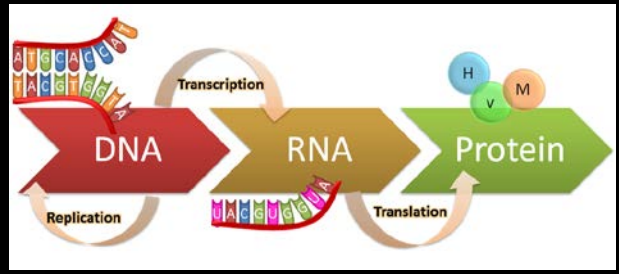


By 2011, U.S. healthcare organizations had generated 150 Exabytes – that's 150 billion GB.

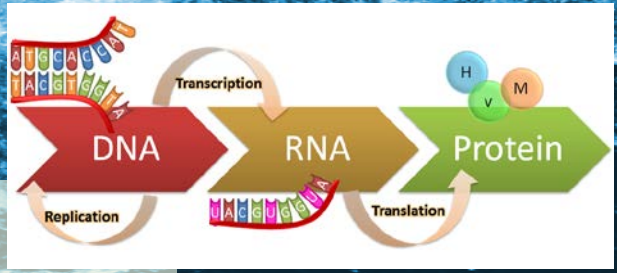
Source: iHT^2


100 trillion human cells
3.2 billion base pairs of DNA
30,000 genes encoding proteins
10 million total distinct proteins in a person
2000 distinct proteins functioning in a cell
60,000 reactions/cell/minute
100,000's of molecular events
50 or so organs and organ systems
Image source: poetry.rapgenius.com


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY


Transcription
DNA
RNA
Protein
Replication
Translation

WHY IS THIS INTERESTING?

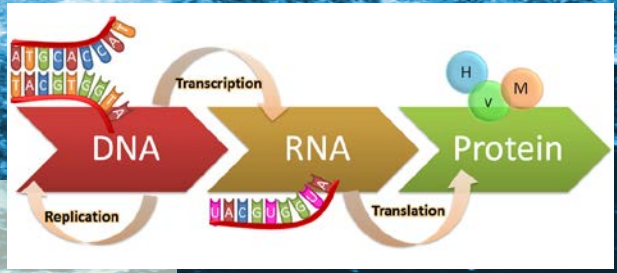


100 trillion human cells

3.2 billion base pairs of DNA
30,000 genes encoding proteins
10 million total distinct proteins in a person
2000 distinct proteins functioning in a cell
60,000 reactions/cell/minute
100,000's of molecular events
50 or so organs and organ systems

All processes follow a well „designed program" – changes are not good…

Image source: poetry.rapgenius.com



100 trillion human cells

3.2 billion base pairs of DNA
30,000 genes encoding proteins
10 million total distinct proteins in a person
2000 distinct proteins functioning in a cell
60,000 reactions/cell/minute
100,000's of molecular events
50 or so organs and organ systems

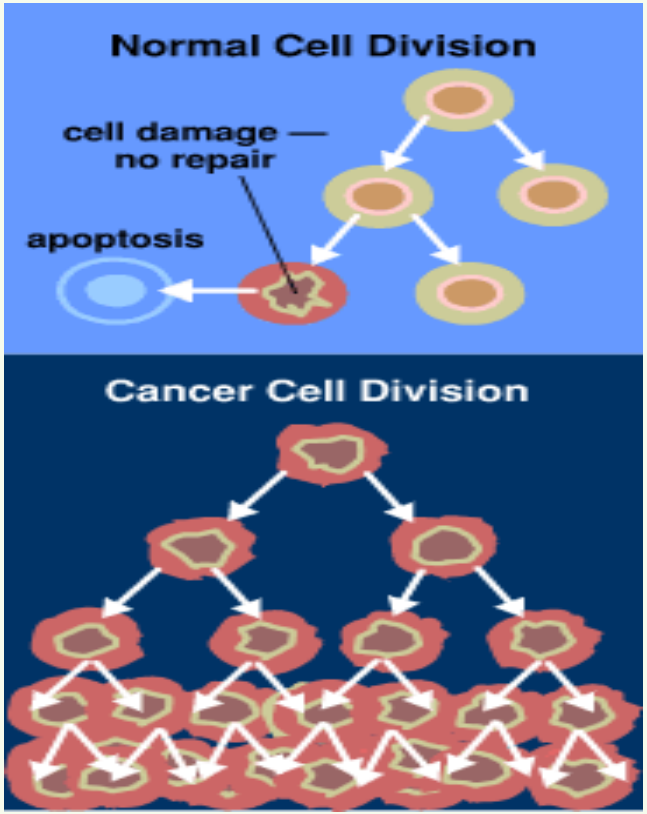
A disease introduces changes from the „norm" – and can be seen in data

Image source: poetry.rapgenius.com

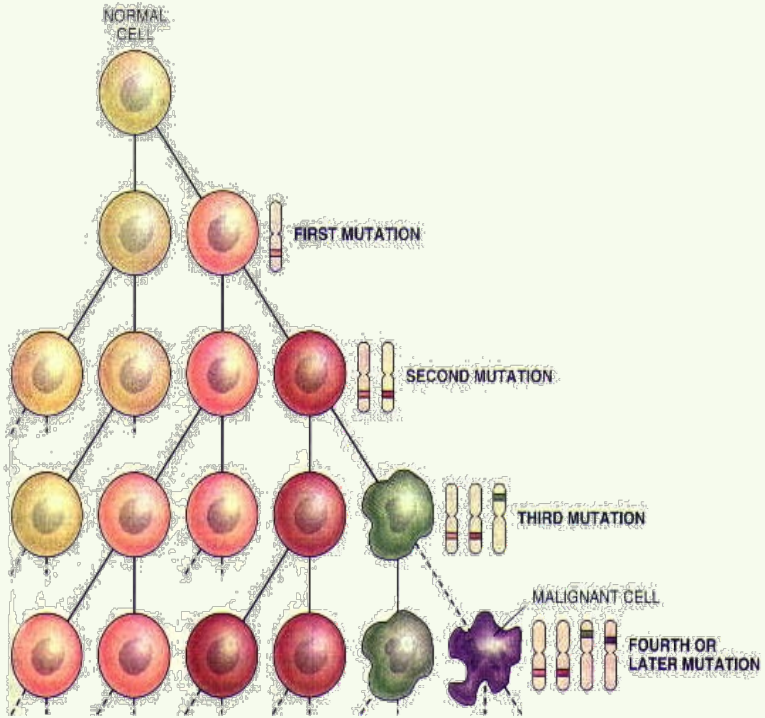
CANCER MOLECULAR BIOLOGY IN A NUTSHELL

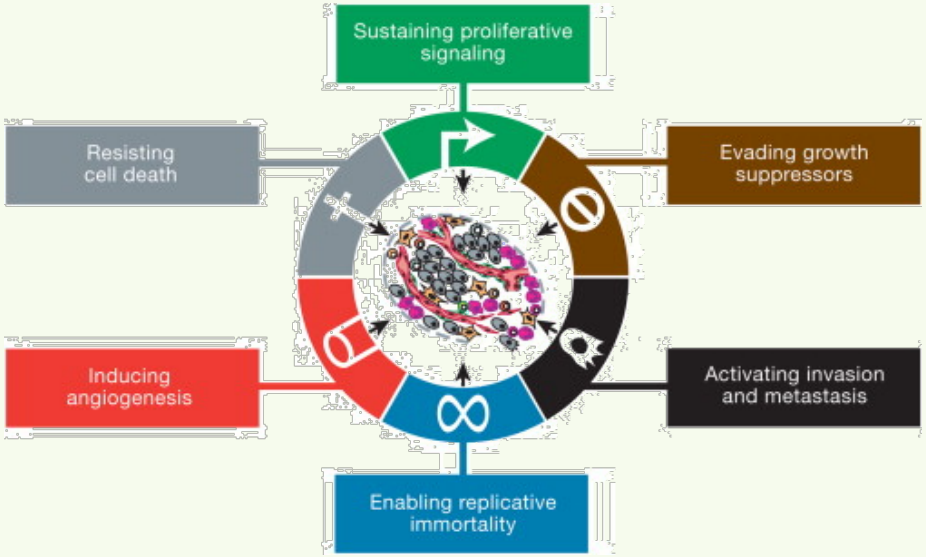
- All cancers derive from single cells that have **acquired the characteristics of continually dividing in an unrestrained manner** and invading surrounding tissues.

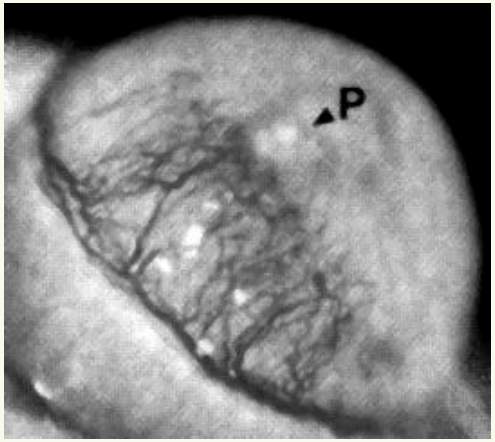
- Cancer cells behave in this abnormal manner because of **mutations in the DNA sequence of key genes**, which are known as cancer genes. Therefore all cancers are genetic diseases.



Mutations in multiple cancer genes are required for the development and progression of a single cancer.







Nucleic acids

- *Mutated DNA*
- *Methylated DNA*

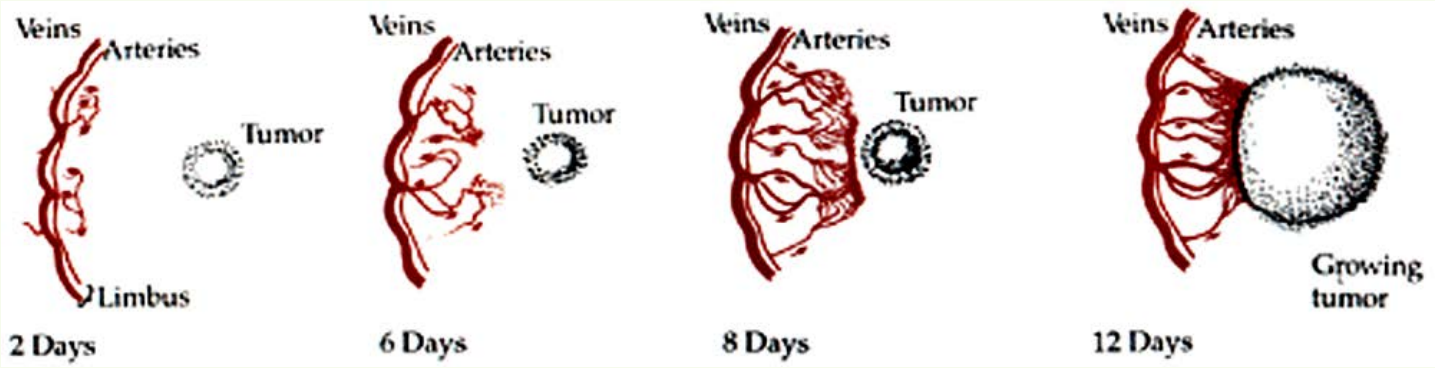
Altered protein and metabolic profiles

- *Tumor cell derived*
- *host response derived*

Immune response signatures

- *Immune cells*
- *Cytokines/chemokines*

Circulating tumor cells



FACT: Diseases leave fingerprints in the human body (e.g. blood proteins)

Image source: Flickr

Diseases leave fingerprints in the human body (e.g. blood proteins)

Image source: Flickr

BUT: 10 million total distinct proteins in a person

Image Source: Flickr

*Take home message

- All cancers result from changes in the DNA sequence of our genome.
- Occasionally, a series of these mutations alters the function of a set of critical genes, resulting in tumour formation.
- This data can be measured with –omics technology, e.g. from blood samples

Part I:

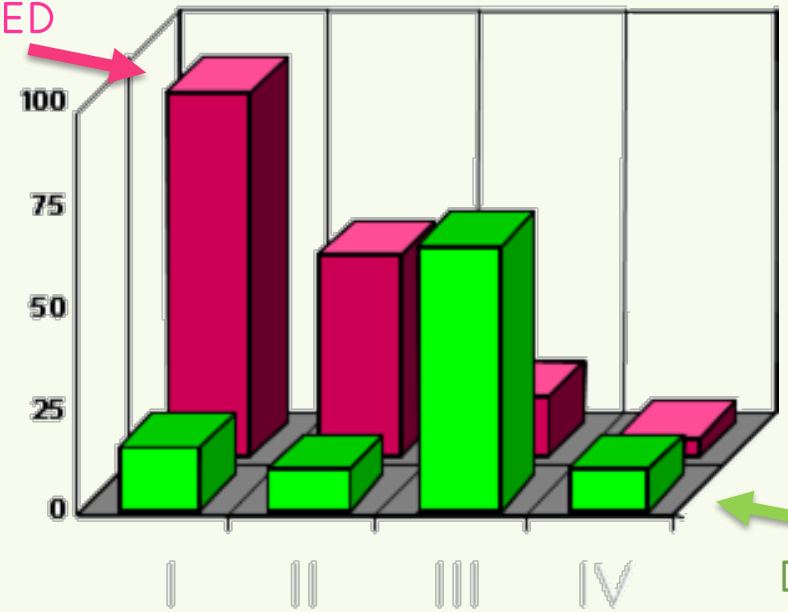
The Vision

- There are approximately 200 types of cancer, each with different causes, symptoms and treatments
- In 2007, ~470.000 people were newly diagnosed with cancer in Germany
- An individual's risk of developing cancer depends on many factors, including age, lifestyle and genetic make-up
- One in three people in the Western world develop cancer and one in five die of the disease

RKI

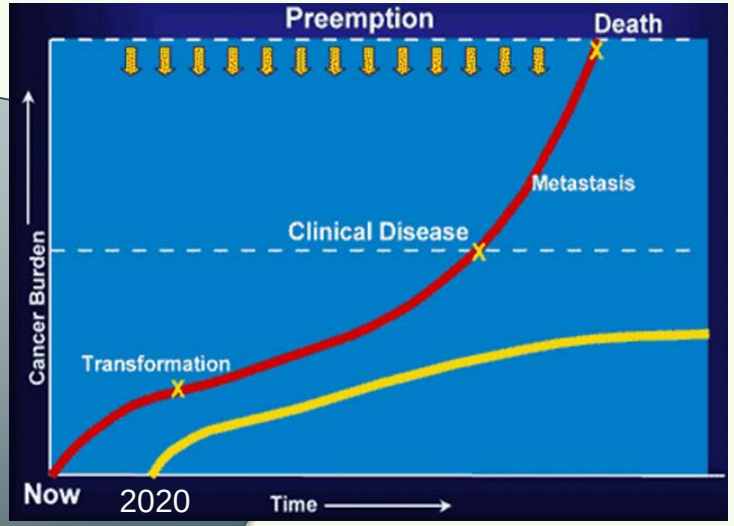
The main problem

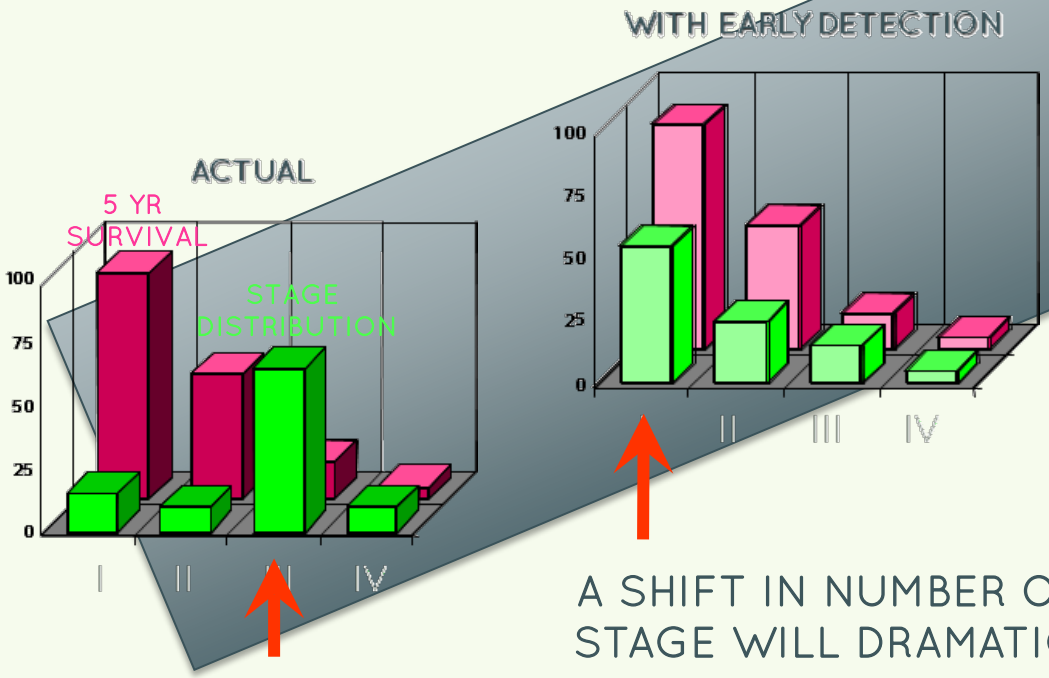
5 YR SURVIVAL WHEN DETECTED AT STAGE



STAGE DISTRIBUTION AT DIAGNOSIS

Develop a minimal-invasive method for early cancer detection allowing medical doctors to start treatment earlier!



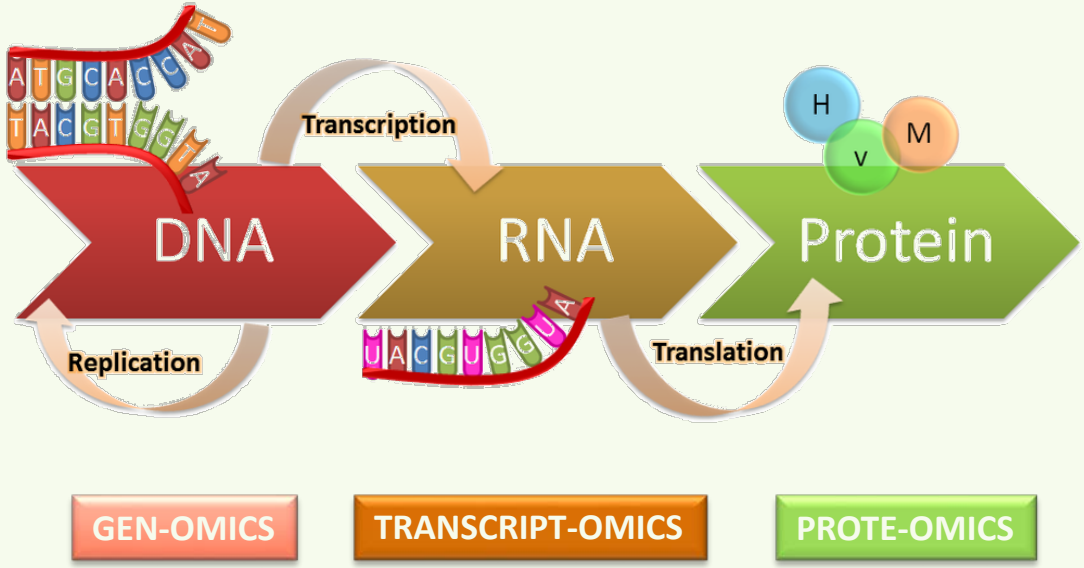


A SHIFT IN NUMBER OF PATIENTS DIAGNOSED AT EARLY STAGE WILL DRAMATICALLY EFFECT PATIENT SURVIVAL!

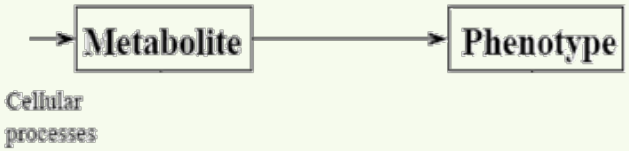
The Strategy

- Identify changes (signals) in the human body that only occur if a disease is present
- Use this information to test a patient whether the signal(s) are present
- Use (easily accessible) data, e.g. derived from blood, that allows to track very early changes

THE DATA



Manifestation of interactions between "-omics" and the environment



What might happen

What can happen

What makes it happen

What happens/happened

Size of OMICS data

- Genome : tens of GBs raw data
- Transcriptome : tens of GBs raw data
- Proteome : several GBs raw data
- Metabolome : hundreds of MBs raw data

- Phenome : potentially TBs

- Epigenome, lipidome, glycome, interactome, spliceome, mechanome, exposome, etc...

*Take home message

- All cancers result from changes in the DNA sequence of our genome.
- The changes are “transported” through the -omics “layers”
- Consequence: cancer cells leave traces (fingerprints), e.g. in the blood proteome.
  This can be measured by -omics technologies.

THE (BASIC) METHODS

Problem Statement

In biomedical mass-data analysis we are often interested in three main things:

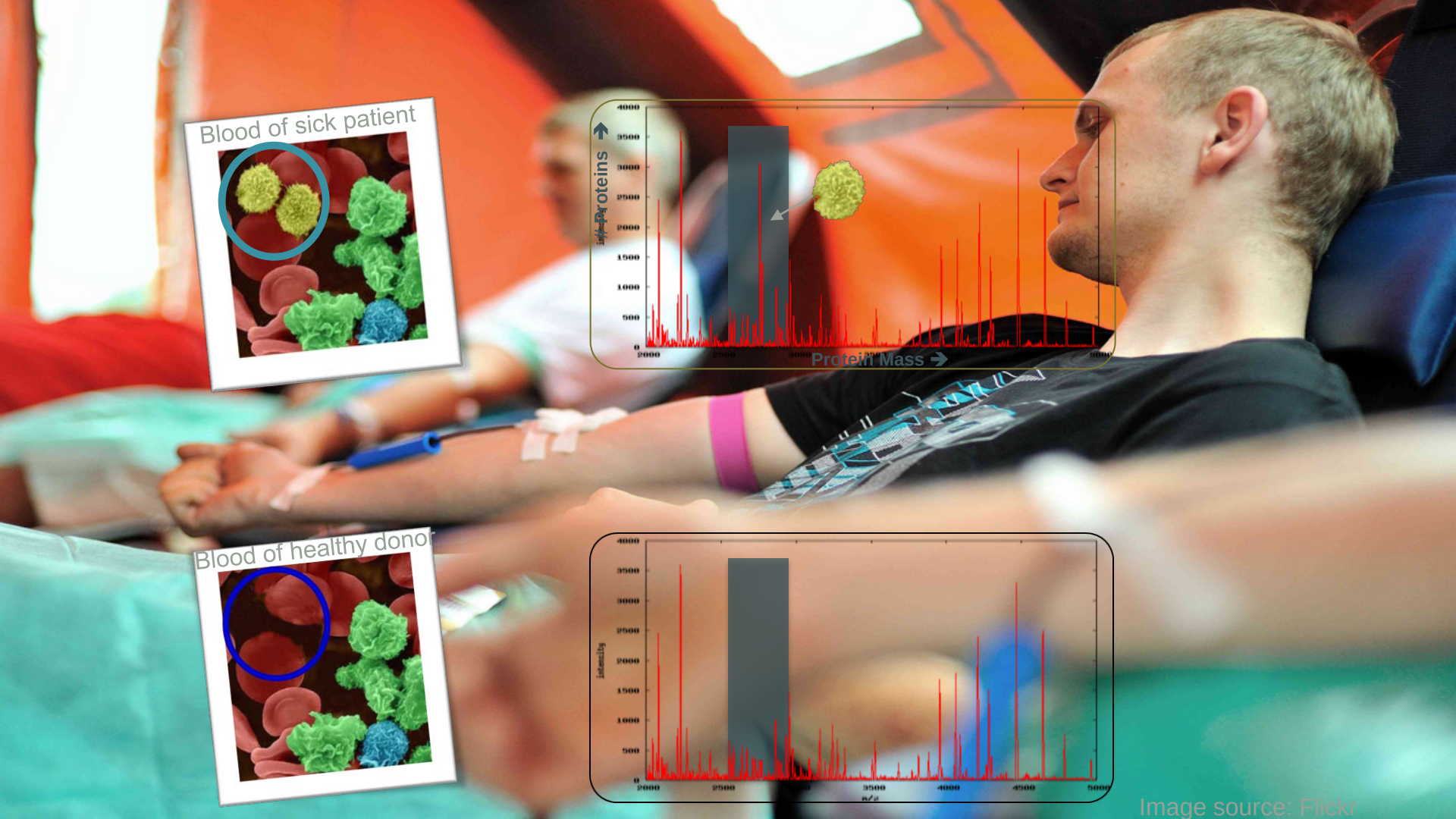
- (1) **CLASSIFICATION** - Given the data: **which class** (e.g. “healthy” vs. “diseased”) does the patient belong to?
- (2) **REGRESSION** - Given the data: **which score** (e.g. “survival estimation”) does a patient get?
- (3) **FEATURE SELECTION** - Given the data: **which features** (e.g. expression level of genes A and D) are most relevant (e.g. for classification or regression)?

Find significantly changed protein concentrations to predict class

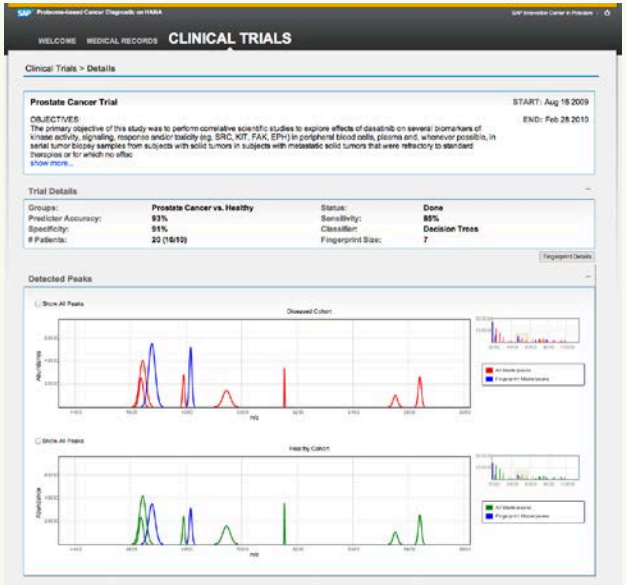
CLASSIFICATION & PROTEOMICS

Diseases leave fingerprints in the human body (e.g. blood proteins)

Image source: Flickr


Blood of sick patient
Proteins
Protein Mass
Blood of healthy donor
Intensity
Image source: Flickr

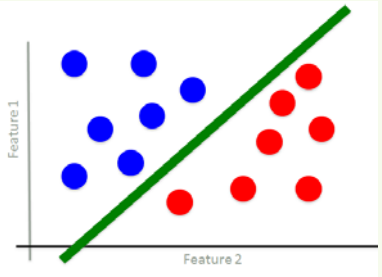
- It is possible to identify cancer specific fingerprints for early diagnosis based on mass spectrometry-based proteomics



1. Leichtle, A. and Nuoffer, J.-M. and Ceglarek, U. and Kase, J. and Conrad, T. O. F. and Witzigmann, H. and Thiery, J. and Fiedler, G. M. (2011) Serum amino acid profiles and their alterations in colorectal cancer. Metabolomics . ISSN 1573-3890 (In Press)
2. Leichtle, A. and Ceglarek, U. and Kase, J. and Conrad, T. O. F. and Hauss, J. and Witzigmann, H. and Thiery, J. and Fiedler, G. M. (2010) Metabolome Alterations in Pancreatic Cancer. Clinical Cancer Research . ISSN 1078-0432 (Submitted)
3. Strenziok, R. and Hinz, S. and Wolf, C. and Conrad, T. O. F. and Krause, H. and Miller, K. and Schrader, M. (2009) Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry: serum protein profiling in seminoma patients. World J of Urology, 28 (2). pp. 193-197.
4. Fiedler, G. M. and Leichtle, A. and Kase, J. and Baumann, S. and Ceglarek, U. and Felix, K. and Conrad, T. O. F. and Witzigmann, H. and Weimann, A. and Schütte, Ch. and Hauss, J. and Büchler, M. and Thiery, J. (2009) Serum Peptidome Profiling Revealed Platelet Factor 4 as a Potential Discriminating Peptide Associated With Pancreatic Cancer. Clinical Cancer Research, 15 (11). pp. 3812-3819. ISSN 1078-0432
5. Strenziok, R. and Hinz, S. and Wolf, C. and Conrad, T. O. F. and Krause, H. and Lingnau, A. and Lein, M. and Miller, K. and Schrader, M. (2008) Serum proteomic profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry in testicular germ cell cancer patients. European Urology Supplements, 7 (3). p. 83.
6. Unpublished: lung cancer (2012)

Preliminary results: fingerprints for 6 different cancer types two of which have proven successful in first clinical studies.

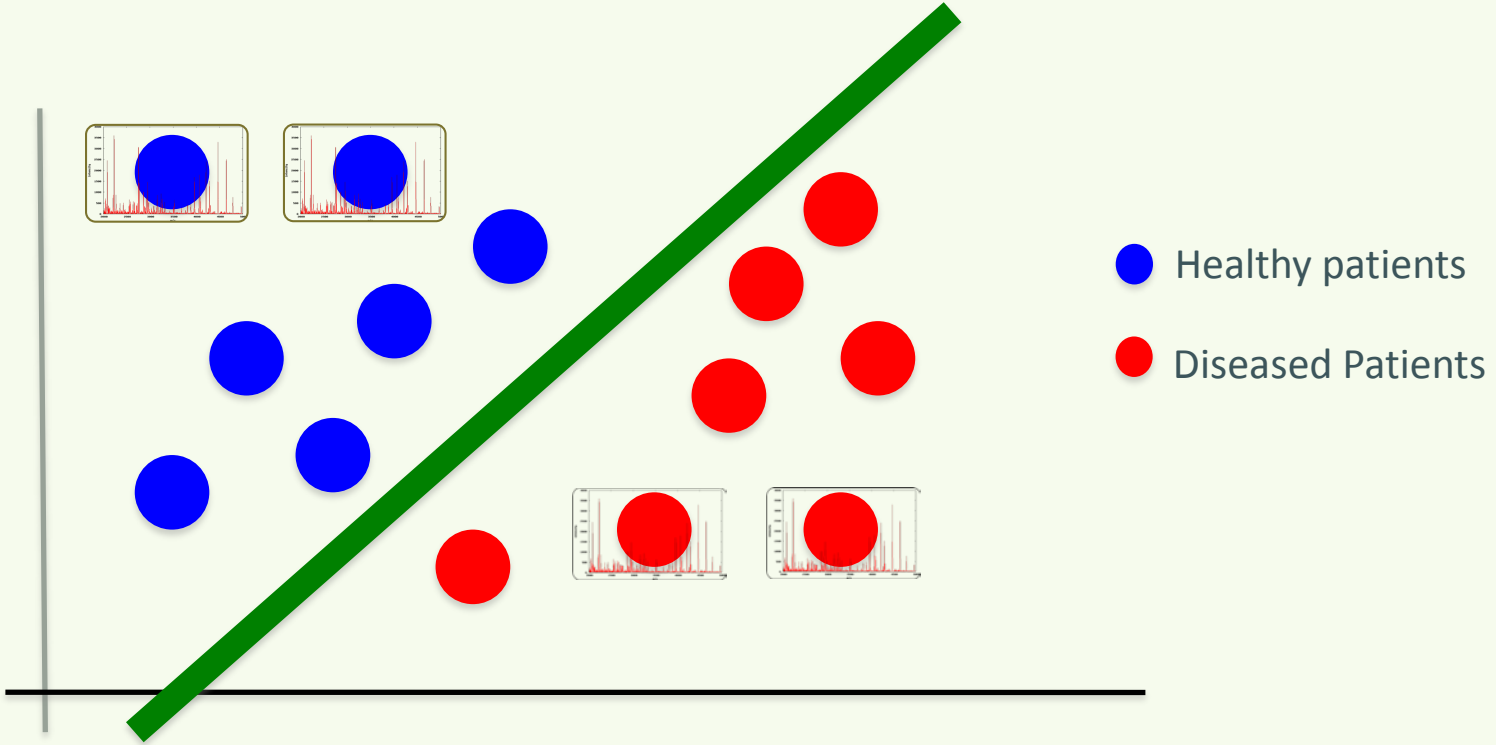
Classification in General



- Classification is a supervised learning problem
- Preliminary task is to construct classification rule (some functional form) from the training data
- For $p<<n$, many methods are available in classical statistics,
  - ♦ Linear (LDA, LR)
  - ♦ Non-Linear (QDA, KLR)
- However when $n<<p$ we face estimability problem (“Curse of dimensionality”)
- Some kind of data transformation is inevitable
- Well known techniques for $n<<p$: PCA , SVM etc.

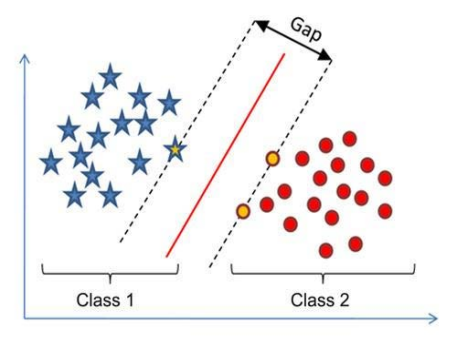
Classification (e.g. SVM)

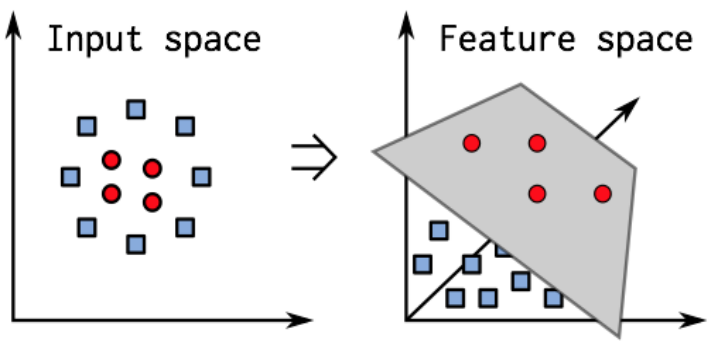
Determine separating line



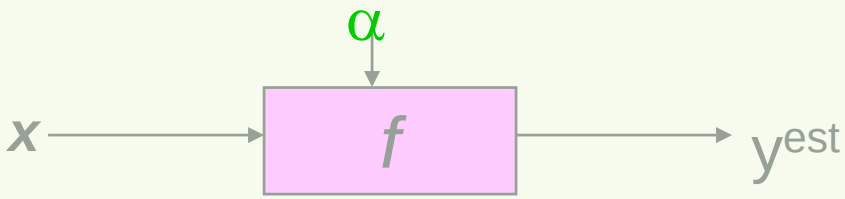
Support Vector Machines (SVM)

- Goal of SVM method: create a hyperplane separating data from 2 classes that maximizes margin between the hyperplane and the classes
- Formulation: $\min \|\beta\| \quad \text{subject to} \begin{cases} y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i \ \forall i, \\ \xi_i \geq 0, \ \sum \xi_i \leq \text{constant}. \end{cases}$
- Kernels
- Extensions: Post-processing with ROC variable importance to reduce the number of features

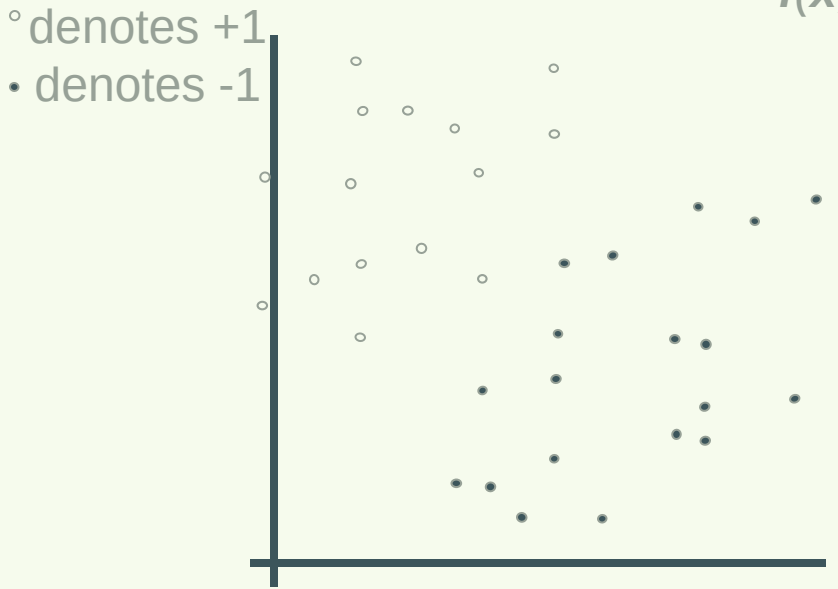




Linear Classifiers

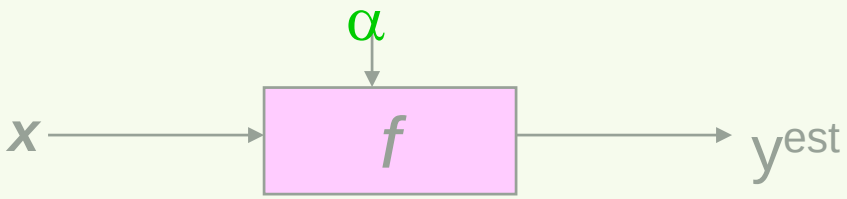


$f(x, w, b) = sign(w \cdot x - b)$

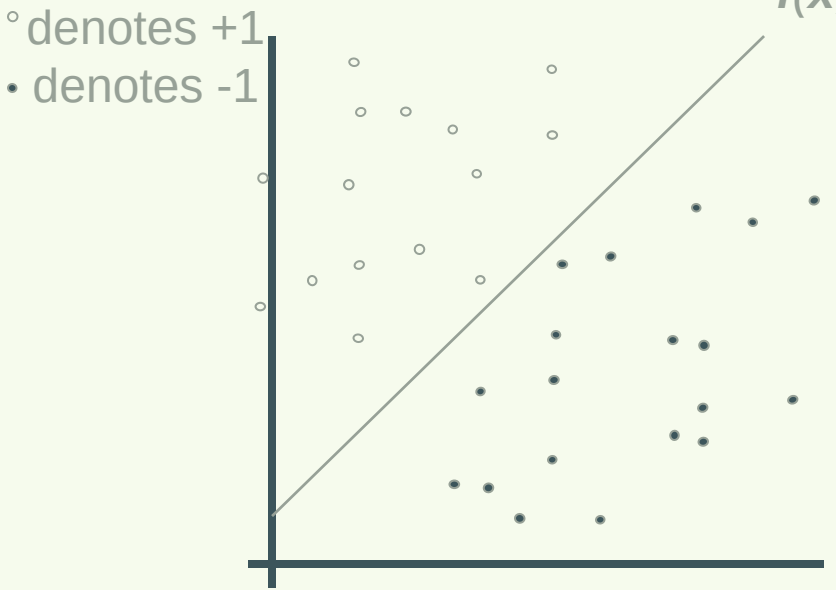


How would you classify this data?

Linear Classifiers

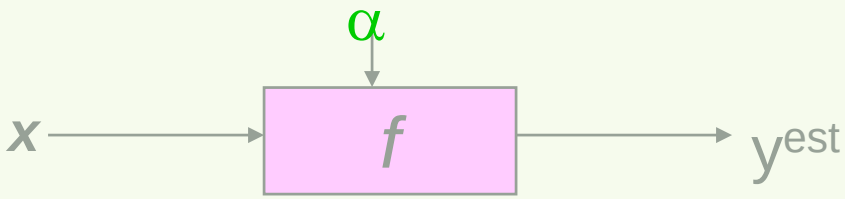


$f(x,w,b) = sign(w.x - b)$



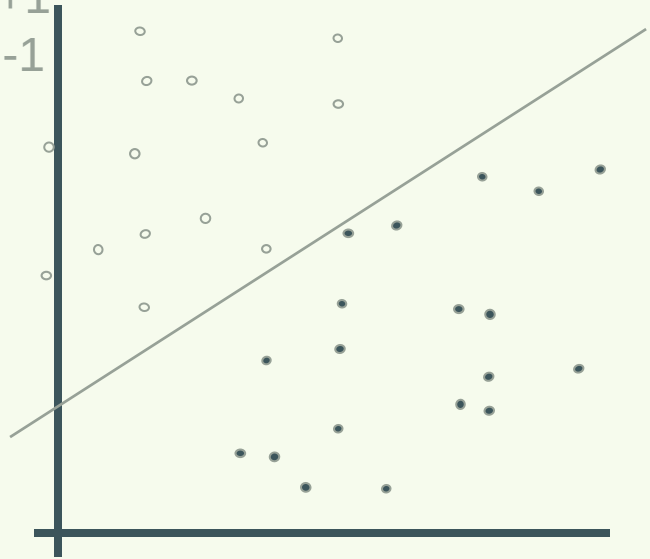
How would you classify this data?

Linear Classifiers



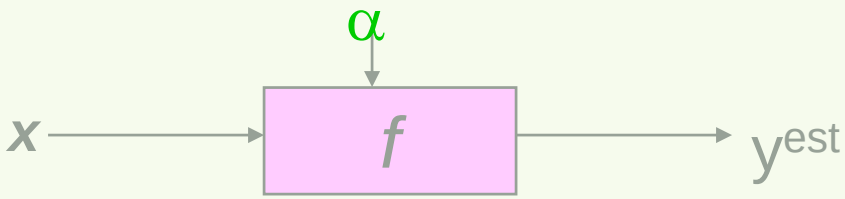
$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\ \mathbf{x} - b)$

° denotes +1
• denotes -1

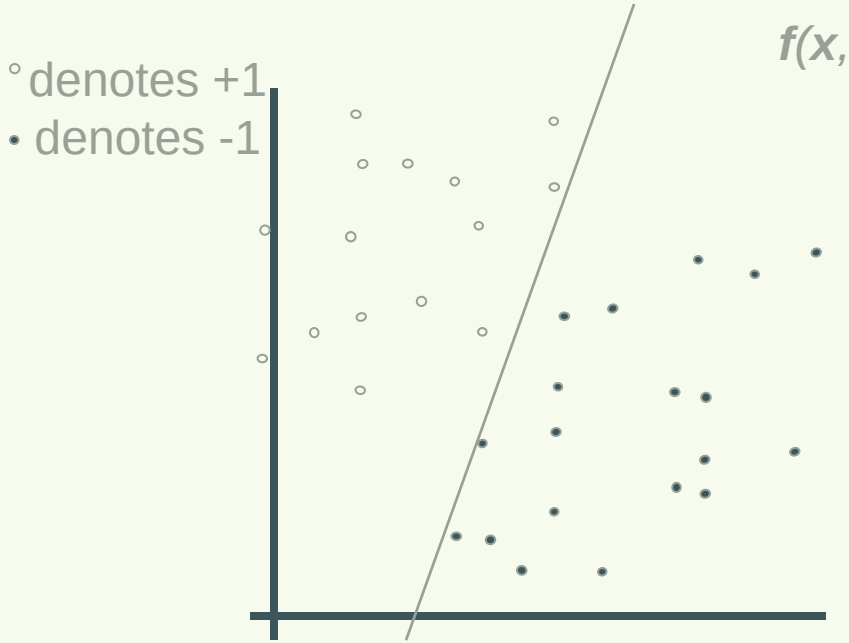


How would you classify this data?

Linear Classifiers

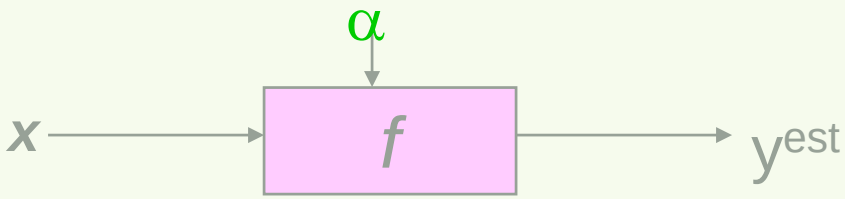


$f(x,w,b) = sign(w.\ x - b)$



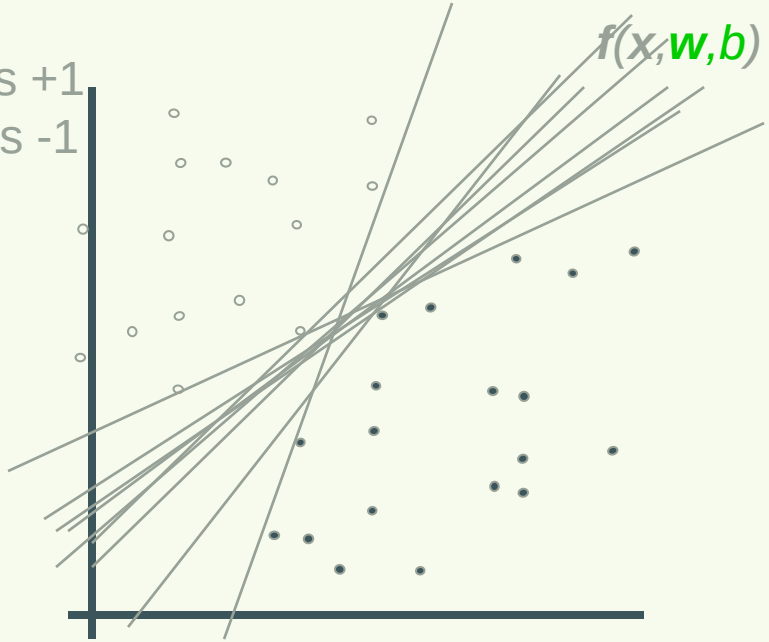
How would you classify this data?

Linear Classifiers



$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\ \mathbf{x} - b)$

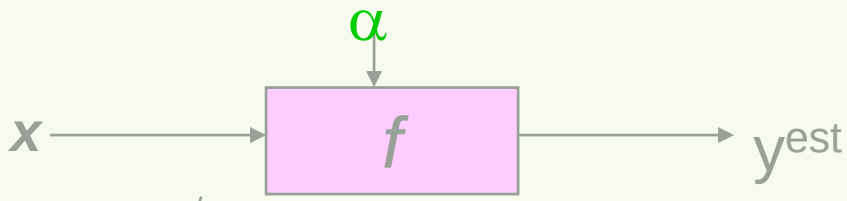
° denotes +1
• denotes -1



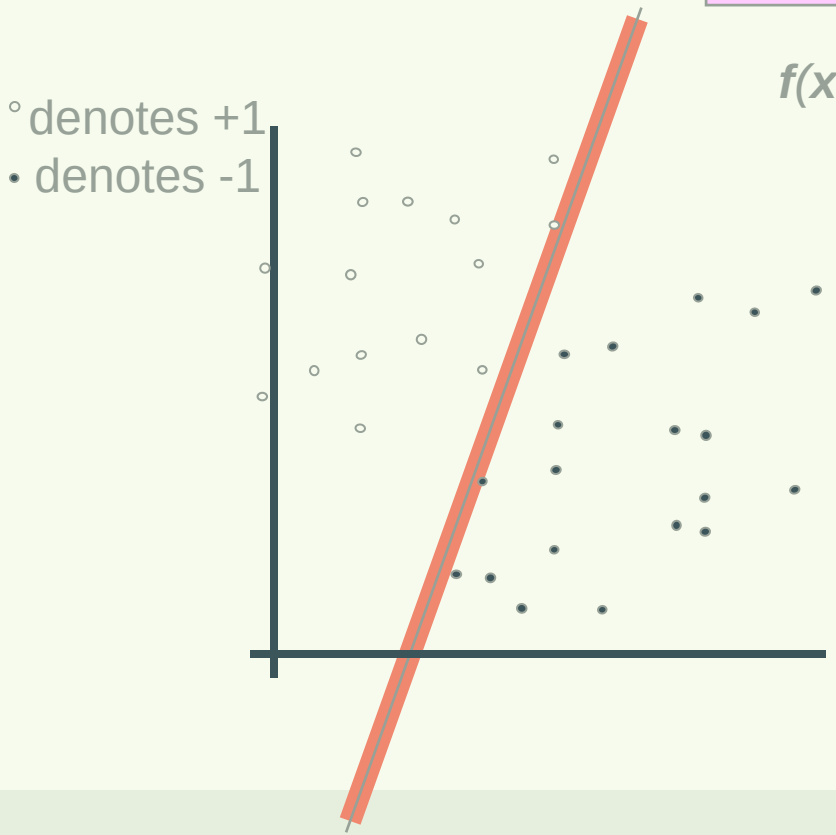
Any of these would be fine..

..but which is best?

Classifier Margin

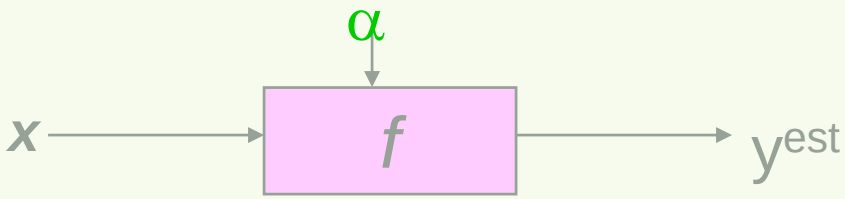


$f(x,w,b) = sign(w.\ x - b)$

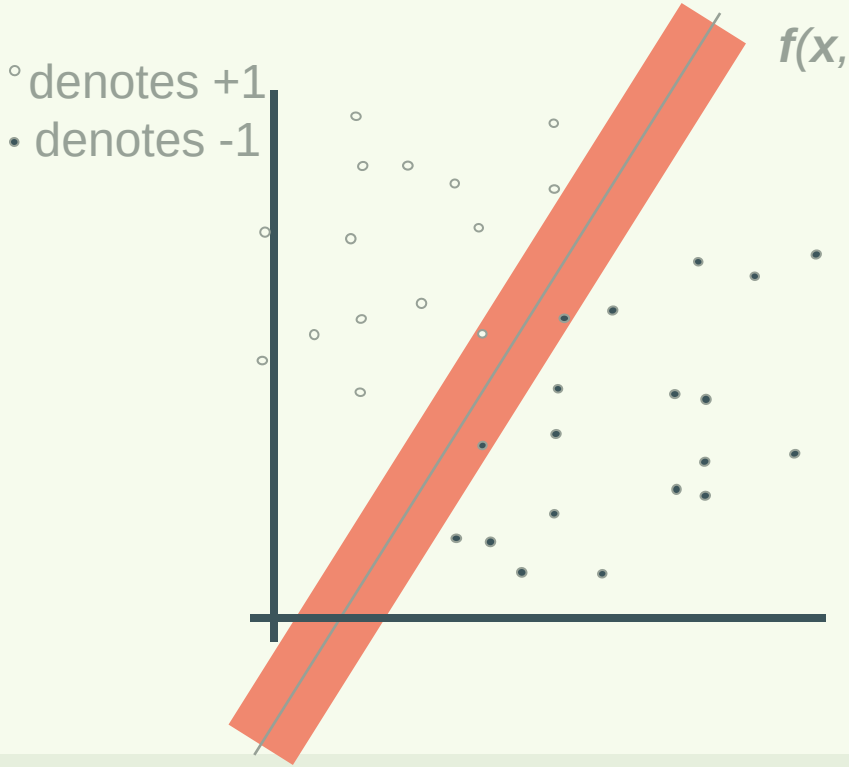


Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Maximum Margin



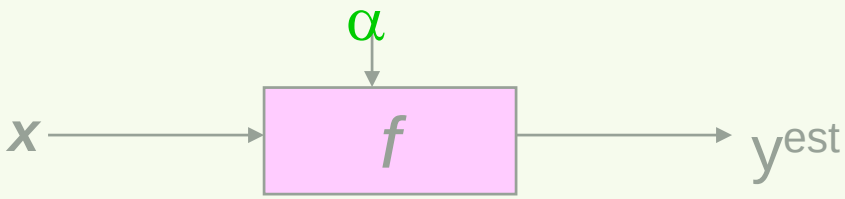
$f(x,w,b) = sign(w.\ x - b)$



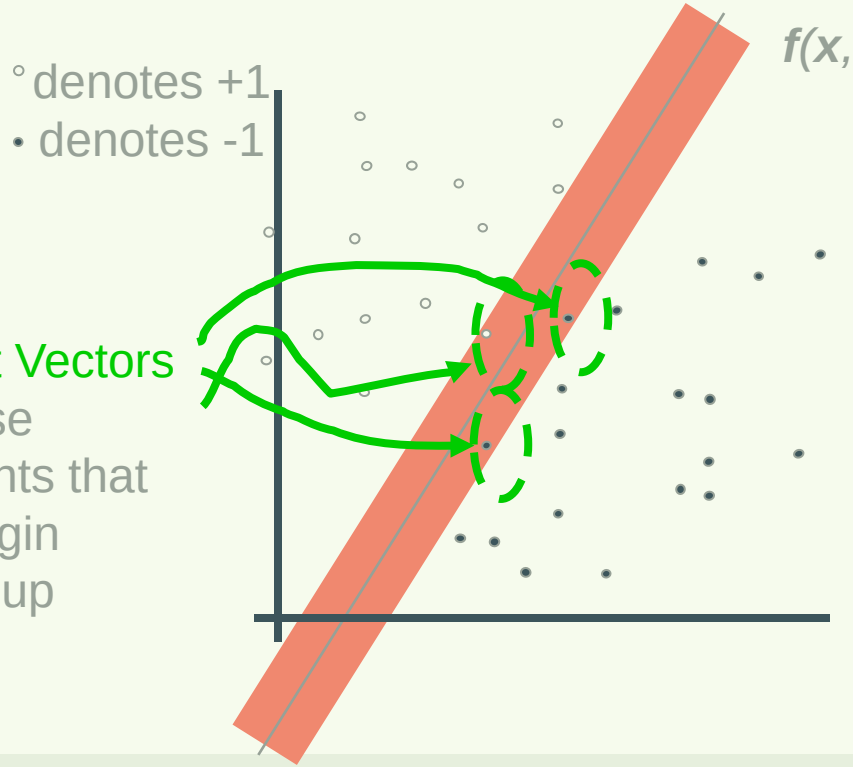
The maximum margin linear classifier is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

Linear SVM

Maximum Margin



$f(x,w,b) = sign(w.\ x - b)$



Support Vectors are those datapoints that the margin pushes up against

The maximum margin linear classifier is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

Linear SVM

Linear SVMs Mathematically

- We can formulate the *quadratic optimization problem:*

Find $\mathbf{w}$ and b such that

$$\rho = \frac{2}{\|\mathbf{w}\|}$$ is maximized; and for all $\{(\mathbf{x_i}, y_i)\}$

$\mathbf{w^T x_i} + b \geq 1$ if $y_i = 1$; $\mathbf{w^T x_i} + b \leq -1$ if $y_i = -1$

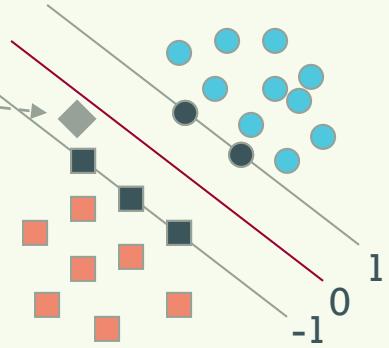
Classification with SVMs

- Given a new point **x**, we can score its projection onto the hyperplane normal:
  - I.e., compute score: $\mathbf{w}^T\mathbf{x} + b = \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$
    - Decide class based on whether < or > 0
  - Can set confidence threshold t.

Score > t: yes

Score < -t: no

Else: don't know



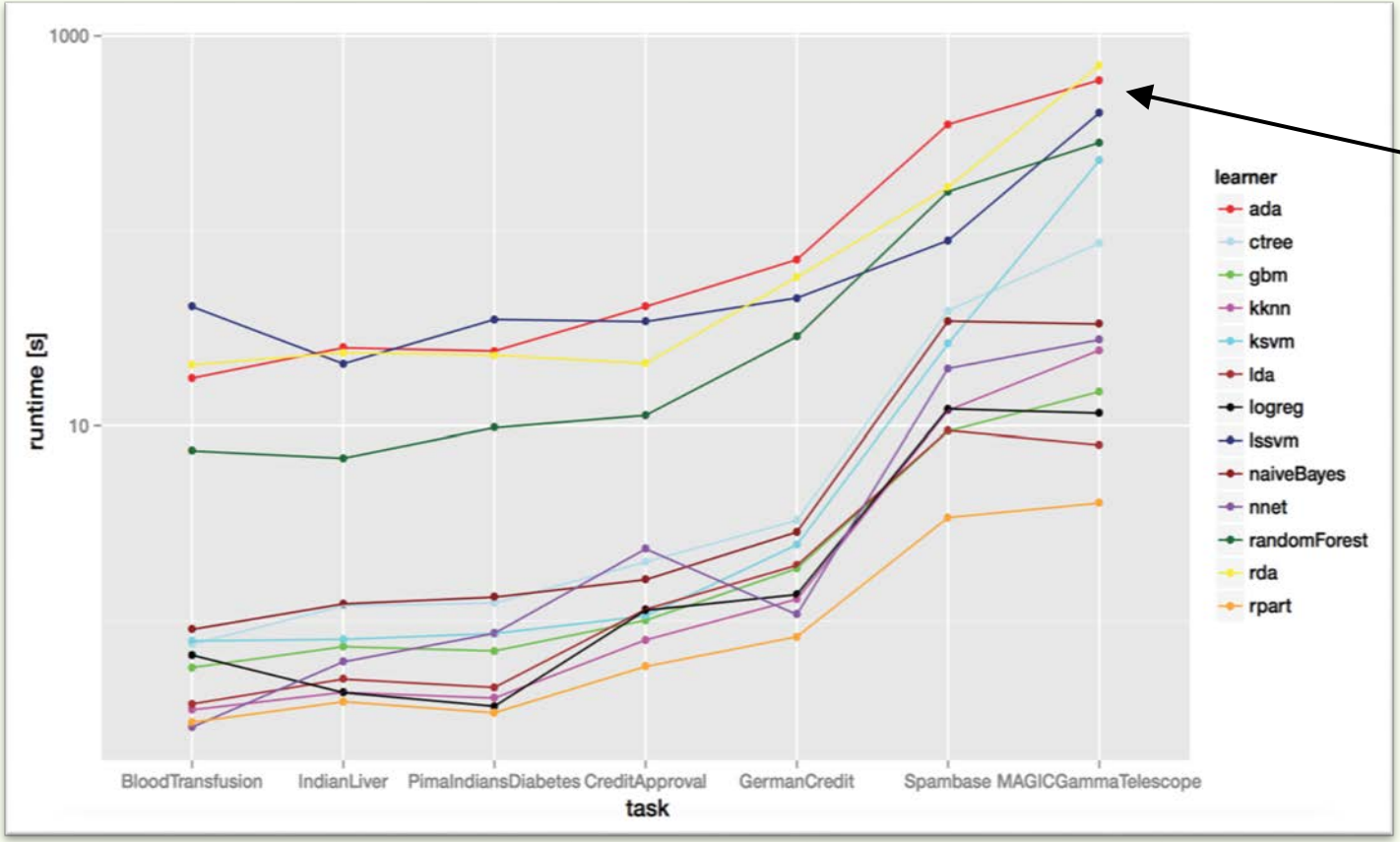
Linear Regularized Loss Minimization

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}^T\mathbf{x}_i)$$

- support vector machines
- logistic regression
- lasso regression
- ridge regression
- etc…

PRACTICAL ISSUE(S)

Runtime analyses for machine learning R programs



ML on a
19,020 x 10
matrix:

takes up to
15 minutes

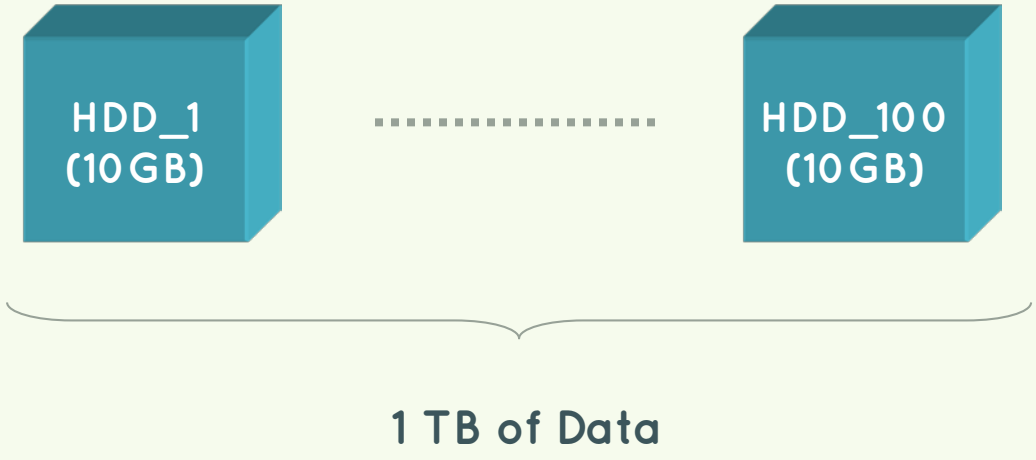
Kotthaus et al. "Runtime and memory consumption analyses for machine learning R programs" Journal of Statistical Computation and Simulation, 2015

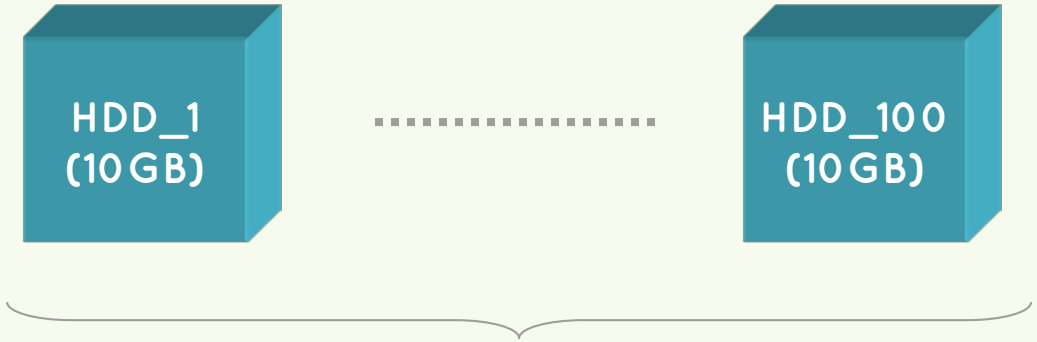
Time & Size Problem

- What if data does not fit into memory any more?

- Goal: efficient solving of objective when data is distributed

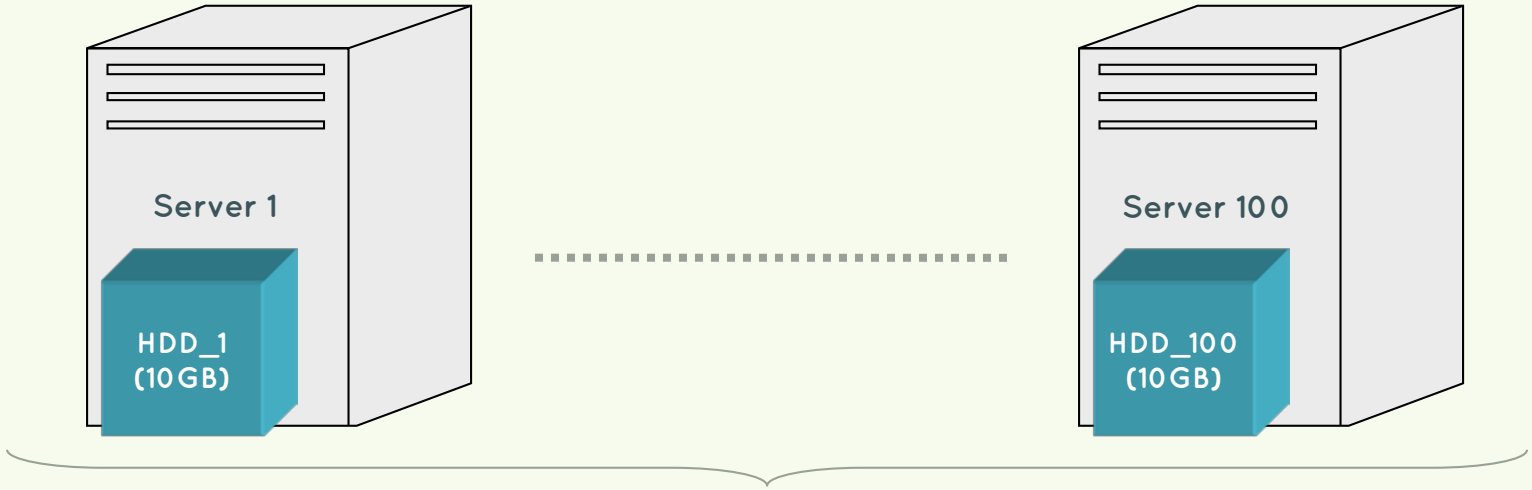
Data Distribution





1 TB of Data in 100 HDDs, but how many computers should there be?
When N = 1, reading 1 TB requires 2.5 HOURS.

What should N be in order to give us appreciable speed-up on reads?



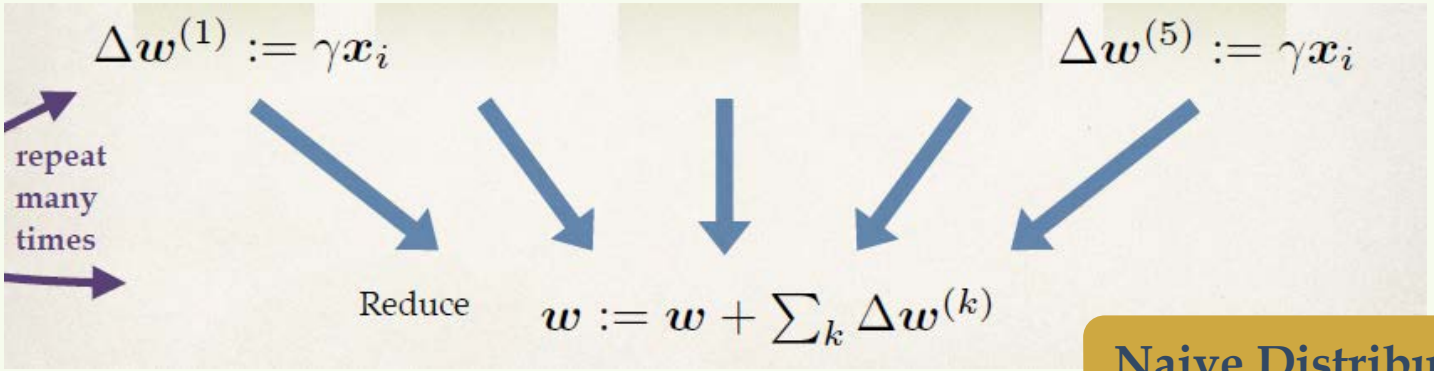
Given: 10 GB per drive
100x 10 GB drives = 1 TB
Read rate is 100 MB/second

Full 1 TB of data can be read in 100 seconds :

10 GB / 100 MB per second = 100 seconds to read one drive.
We read all 100 drives in parallel, and the computers can process the data read in parallel.

This is the architecture in which distributed computing frameworks shine, because not only is the data read in parallel, it is processed in parallel as well.

Distributed Stochastic Optimization



Naive Distributed SGD

The Cost of Communication

$$v \in \mathbb{R}^{100}$$

- Reading v from Memory (RAM)

  100 ns

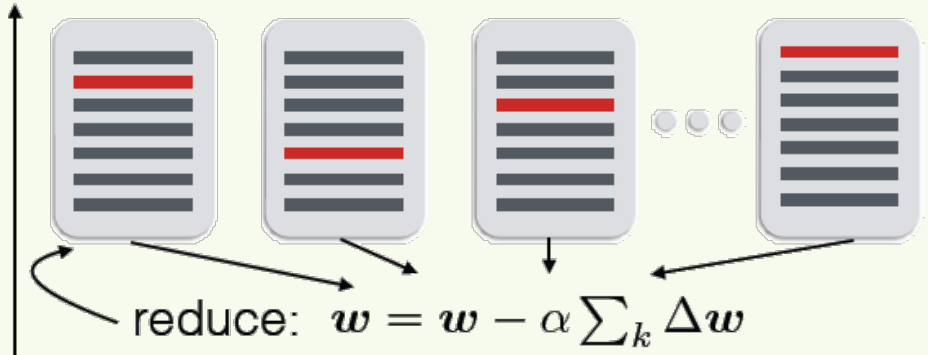
- Sending v to another Machine

  500’000 ns

- One Typical Map-Reduce Iteration *(Hadoop)*

  10’000’000’000 ns

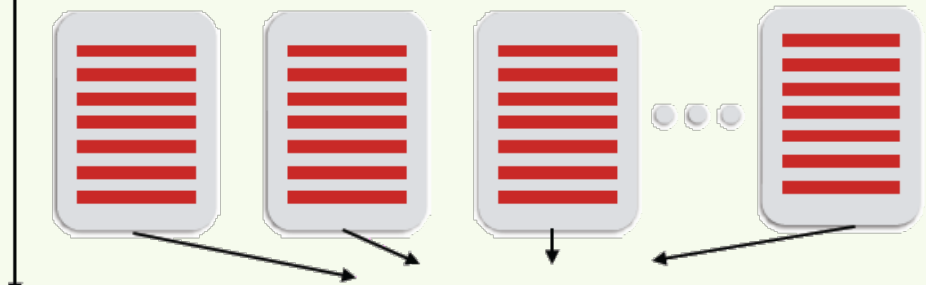
Distributed Optimization



reduce: $\boldsymbol{w} = \boldsymbol{w} - \alpha \sum_k \Delta \boldsymbol{w}$

✔ convergence guarantees
✗ high communication

"always communicate"

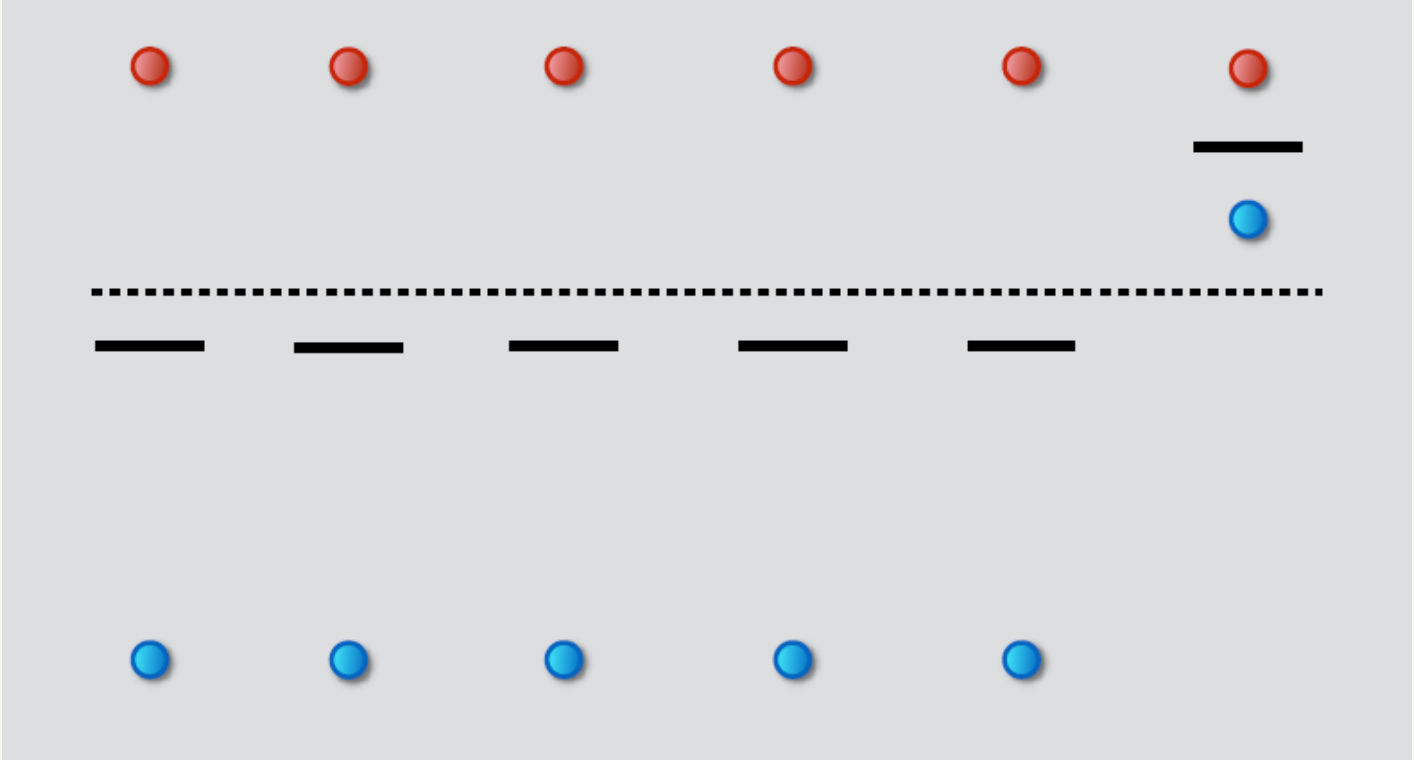


average: $\mathbf{w} := \frac{1}{K} \sum_k \mathbf{w}_k$

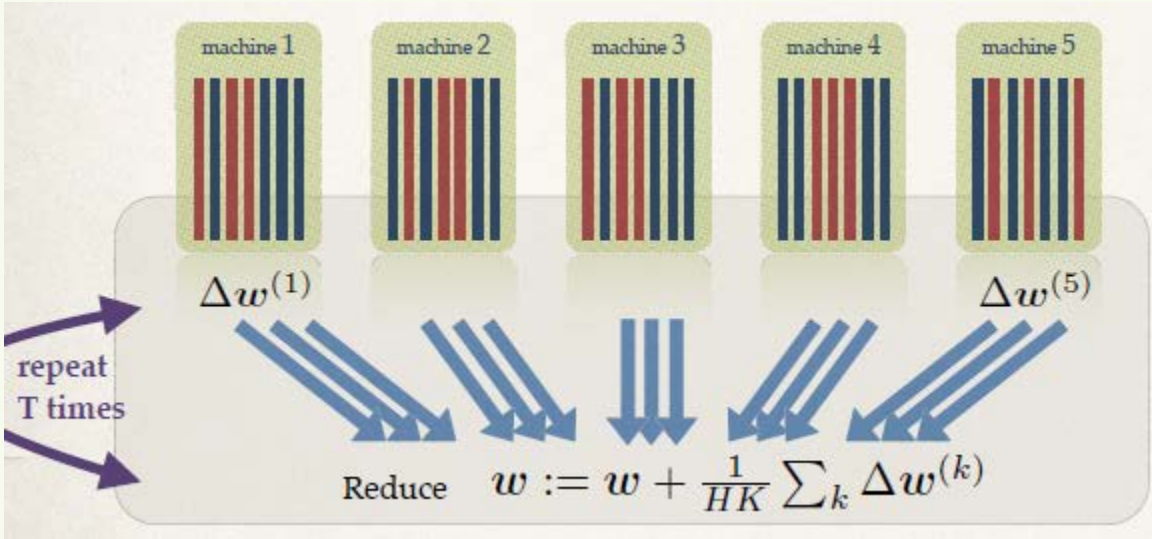
✔ low communication
✗ convergence not guaranteed

"never communicate"

One-shot strategy



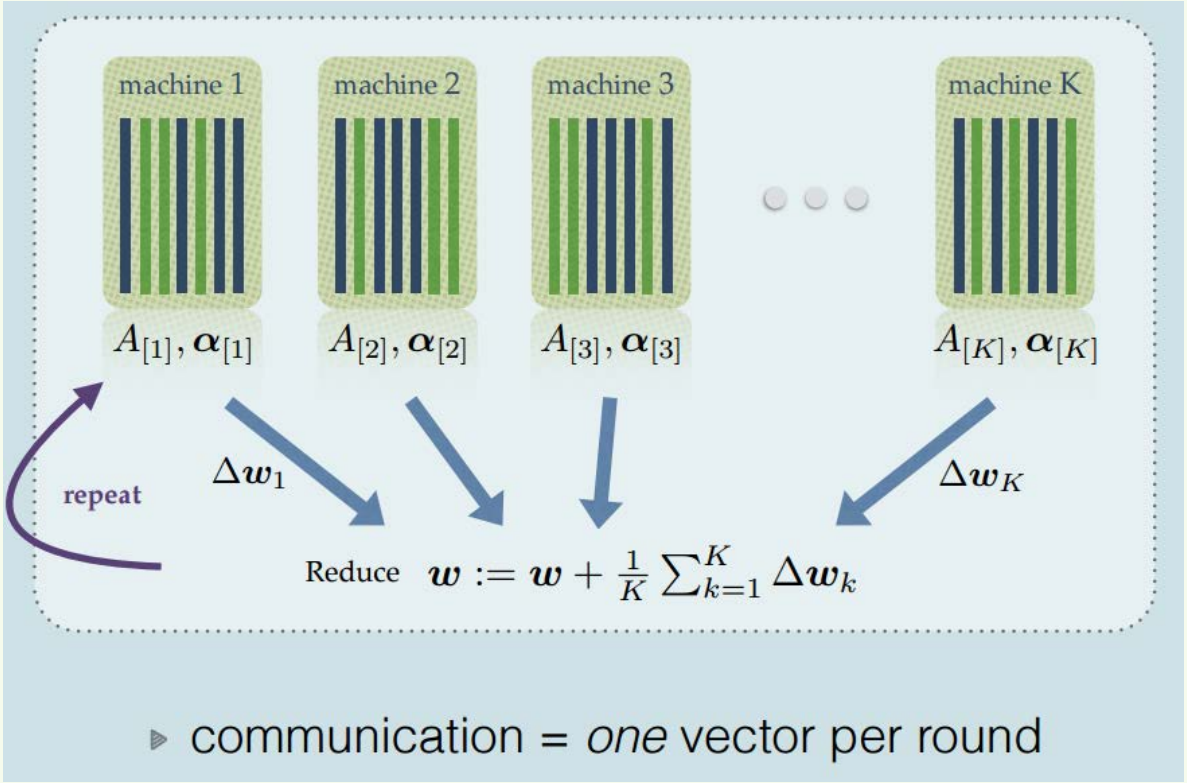
Something in the middle?

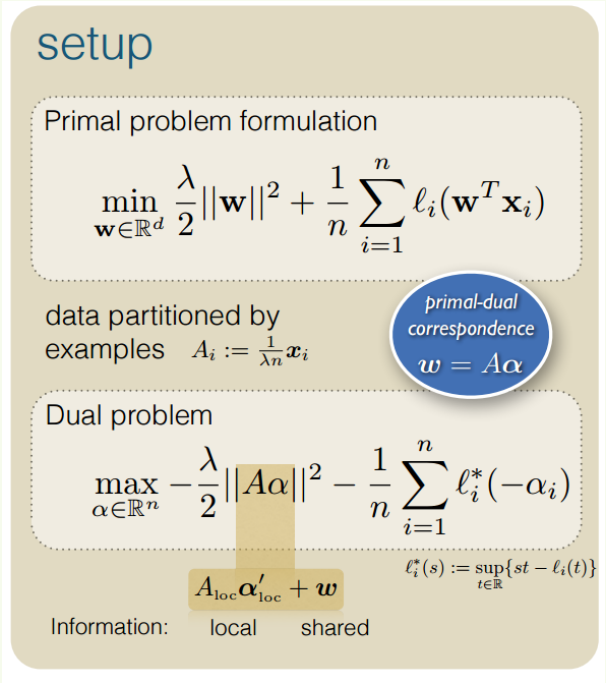


Mini-Batch SGD / CD

# local datapoints read:	TH
# communications:	T
convergence:	✓

Something in the middle?





Source: http://m8j.net/data/List/Files-172/cocoa-NIPS-poster-final.pdf

Communication Efficient Distributed *Dual* **Co**ordinate **A**scent

Algorithm 1: CoCoA

Input: $T \geq 1$, scaling parameter $1 \leq \beta_K \leq K$ (default: $\beta_K := 1$).

Data: $\{(x_i, y_i)\}_{i=1}^n$ distributed over K machines

Initialize: $\alpha_{[k]}^{(0)} \leftarrow 0$ for all machines k, and $w^{(0)} \leftarrow 0$

for $t = 1, 2, \ldots, T$

  for *all machines* $k = 1, 2, \ldots, K$ *in parallel*

    $(\Delta\alpha_{[k]}, \Delta w_k) \leftarrow \textsc{LocalDualMethod}(\alpha_{[k]}^{(t-1)}, w^{(t-1)})$

    $\alpha_{[k]}^{(t)} \leftarrow \alpha_{[k]}^{(t-1)} + \frac{\beta_K}{K} \Delta\alpha_{[k]}$

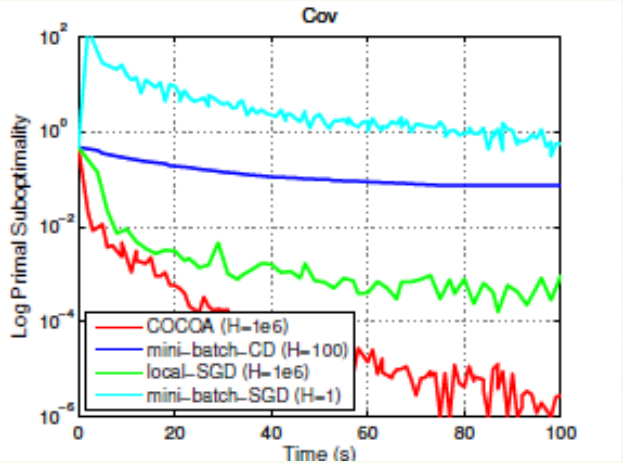
  end

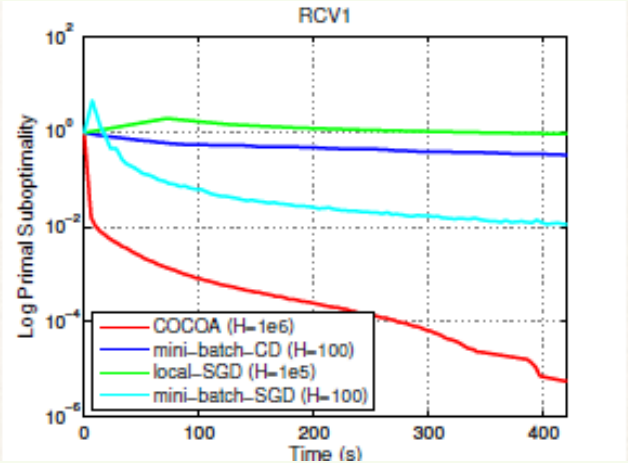
  reduce $w^{(t)} \leftarrow w^{(t-1)} + \frac{\beta_K}{K} \sum_{k=1}^K \Delta w_k$

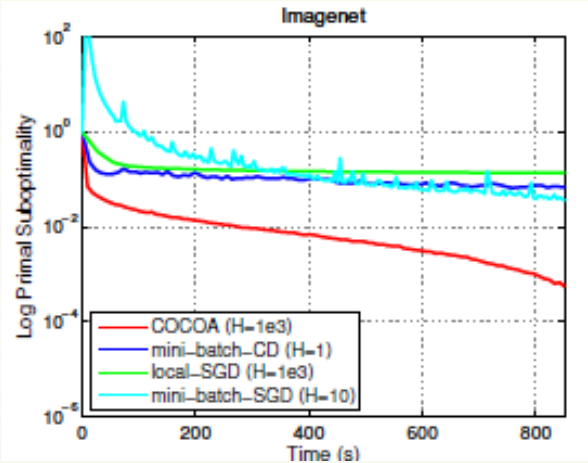
end

Benchmarks

Dataset	Training n	Features d	Sparsity	λ	Workers K
cov	522,911	54	22.22%	1e-6	4
rcv1	677,399	47,236	0.16%	1e-6	8
imagenet	32,751	160,000	100%	1e-5	32

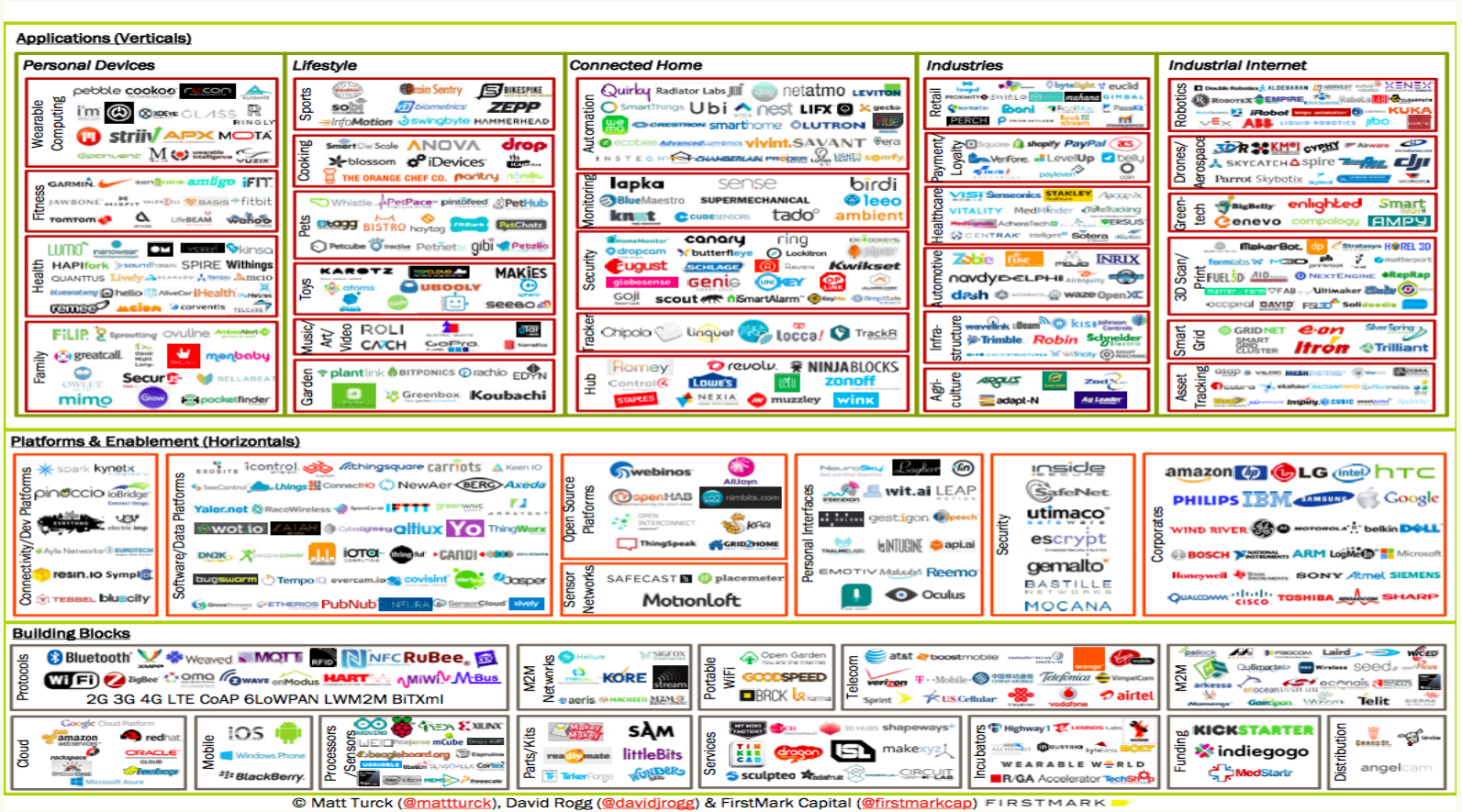






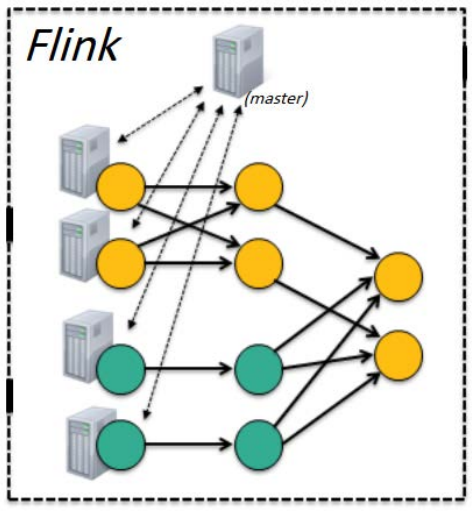

Part II:
Real World
Application

Big Data Frameworks

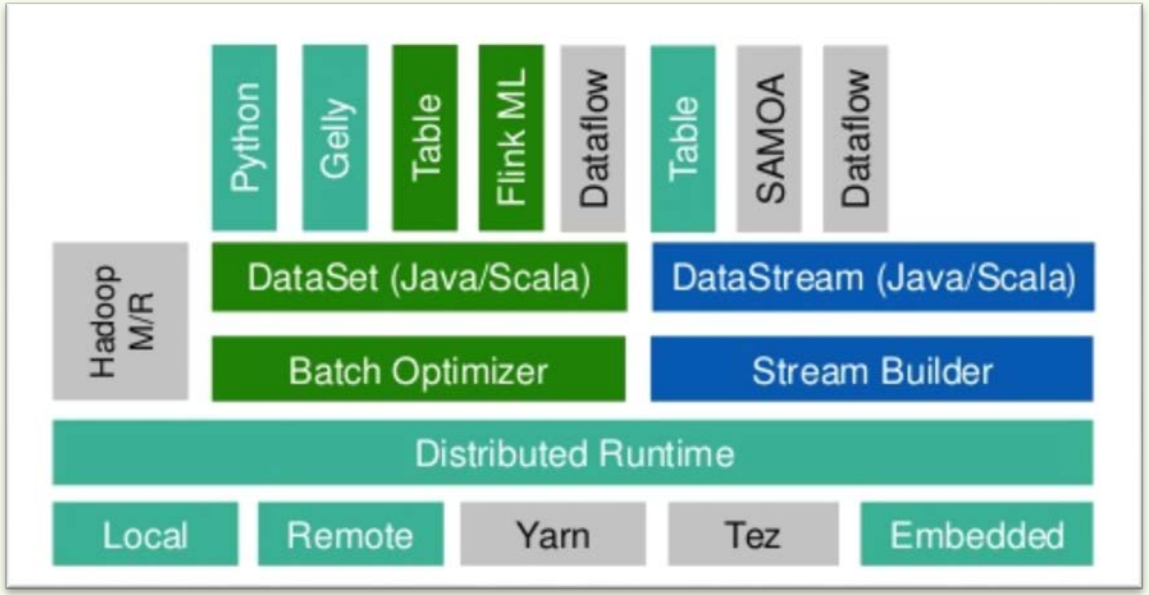


(2014)

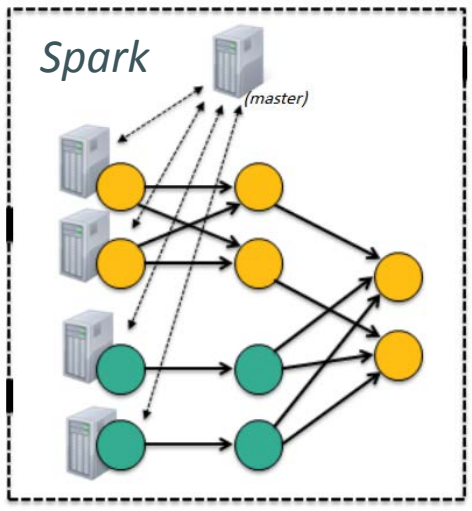
Apache FLINK



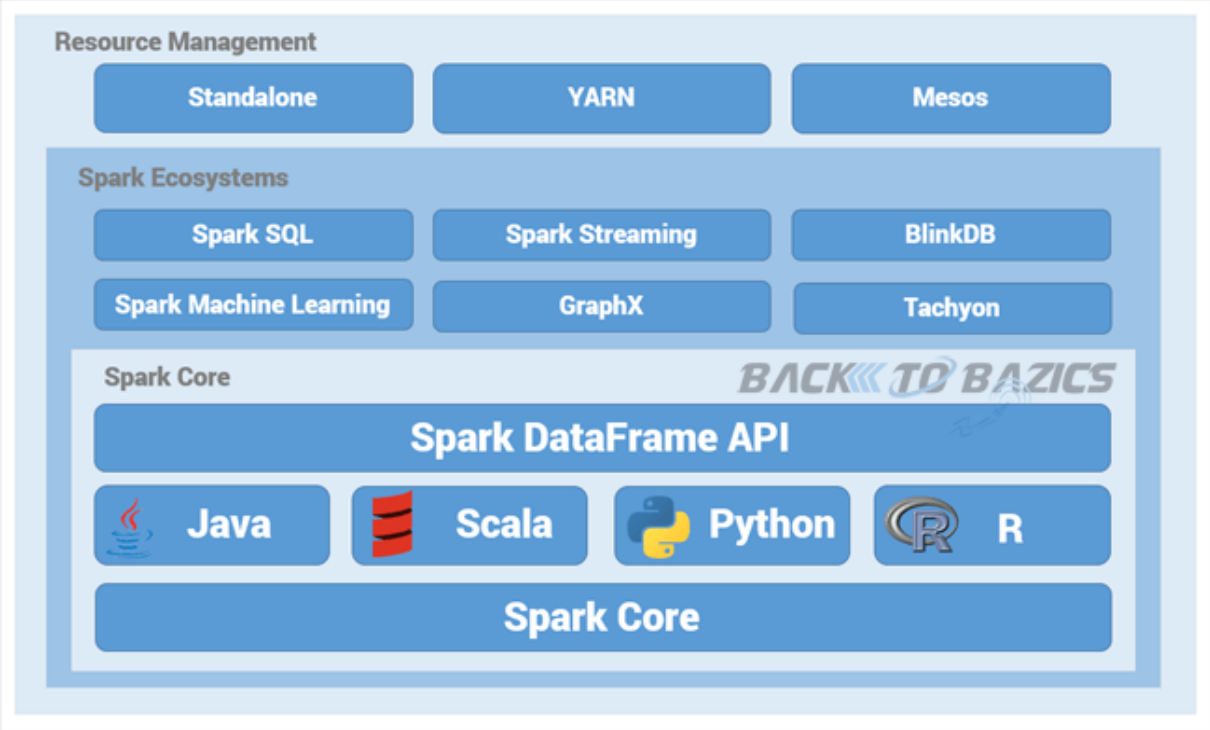
- Allows to execute same code locally and on a compute cluster (distributed computing)
- Automatically distributes data over cluster nodes and allows real-time queries



Apache SPARK

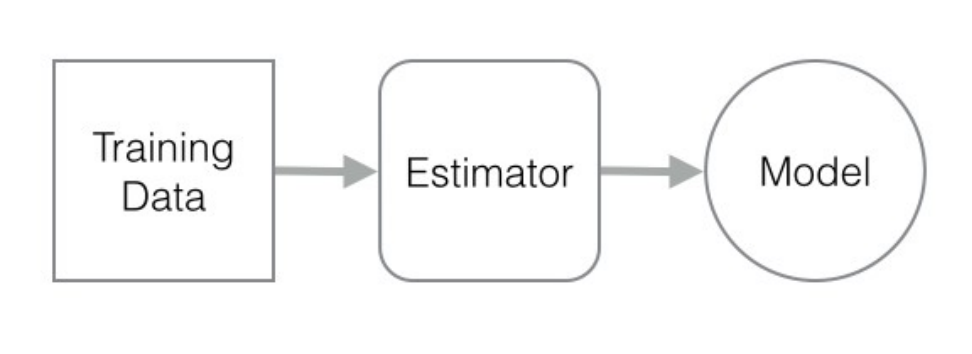


- Allows to execute same code locally and on a compute cluster (distributed computing)
- Automatically distributes data over cluster nodes and allows real-time queries

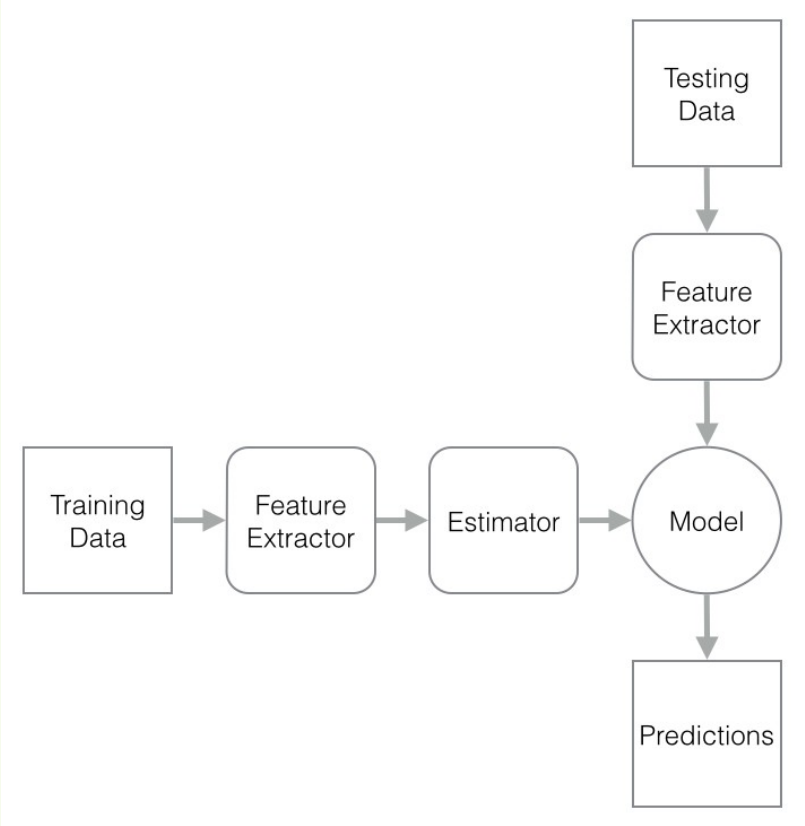


Machine learning using a framework

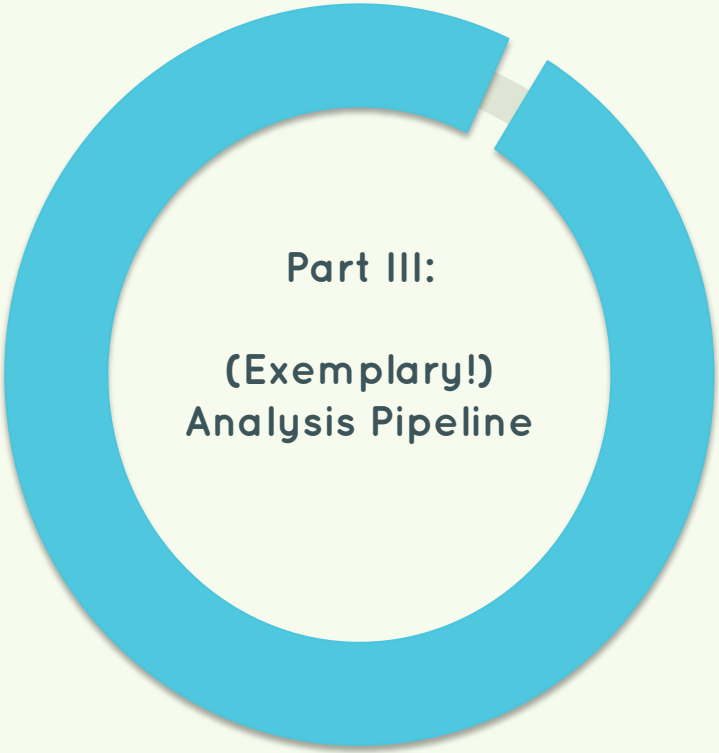
Does ML work like that?



More realistic scenario!

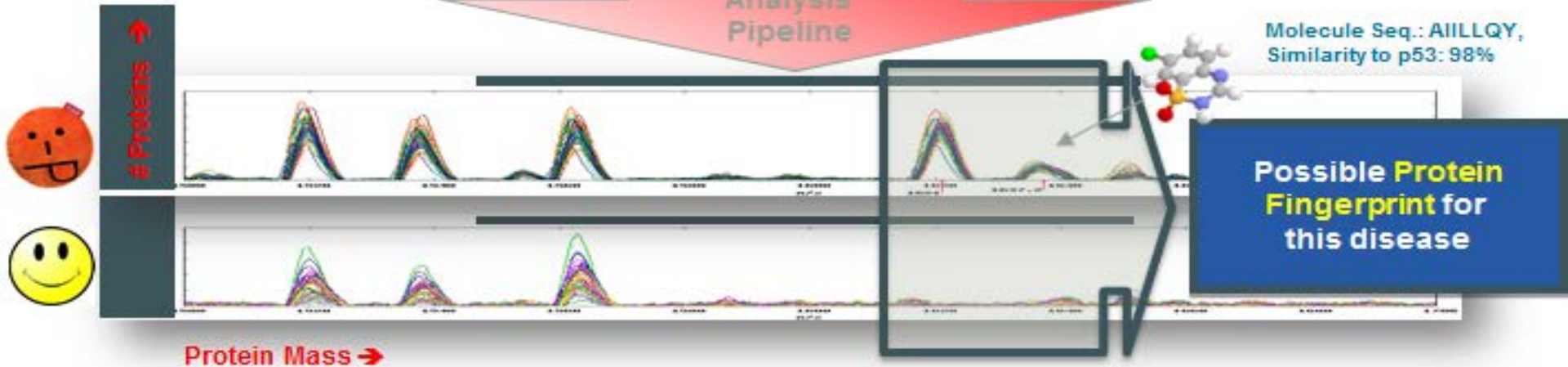


```scala
// Read the training data set, from a LibSVM formatted file
val trainingDS: DataSet[LabeledVector] = env.readLibSVM(pathToTrainingFile)

// Create the SVM learner
val svm = SVM()
.setBlocks(10)
.setIterations(10)
.setLocalIterations(10)
.setRegularization(0.5)
.setStepsize(0.5)

// Learn the SVM model
svm.fit(trainingDS)

// Read the testing data set
val testingDS: DataSet[Vector] = env.readVectorFile(pathToTestingFile)

// Calculate the predictions for the testing data set
val predictionDS: DataSet[LabeledVector] = svm.predict(testingDS)
```


Part III:
(Exemplary!)
Analysis Pipeline

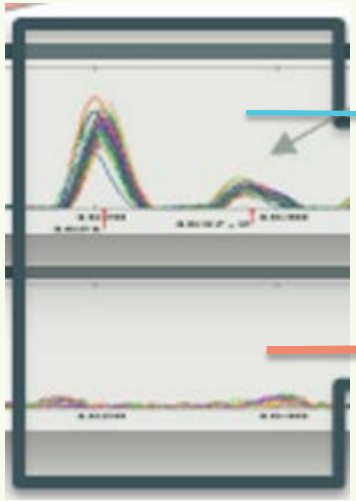

GOAL

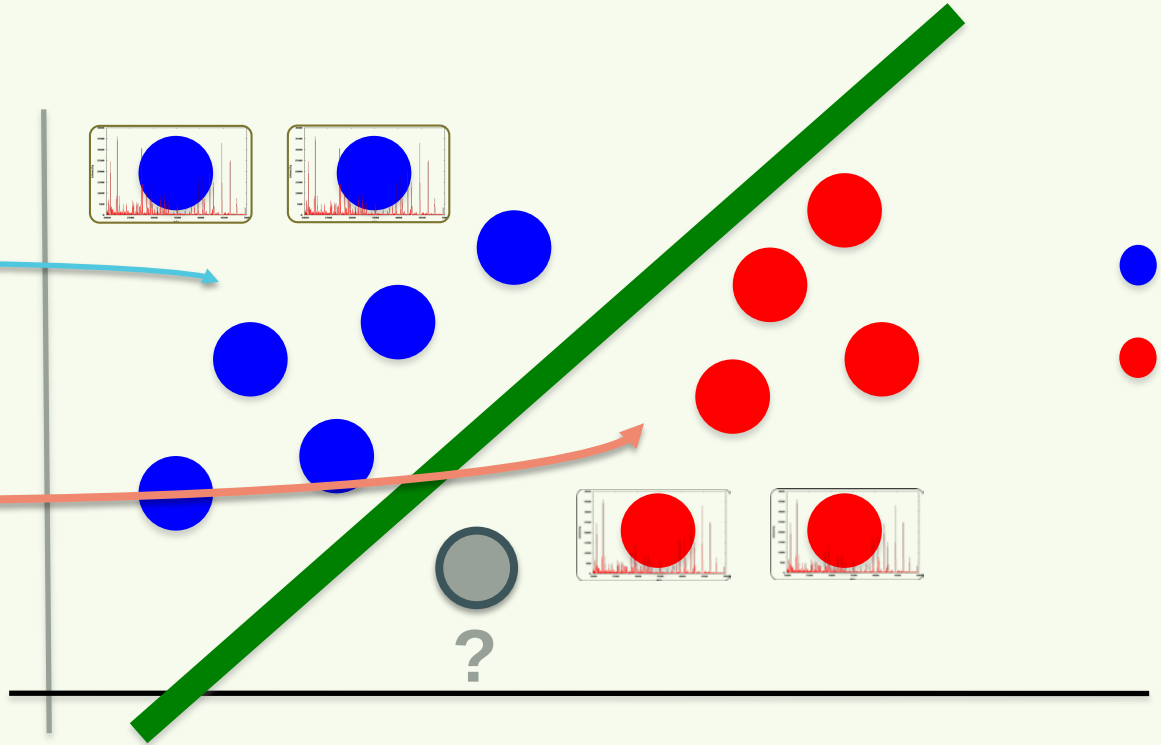
Analysis Pipeline



Classification (e.g. SVM)

Determine separating line

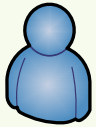


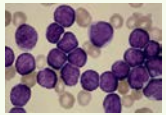


Healthy patients

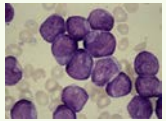
Diseased Patients

Biomarkers Discovery Workflow

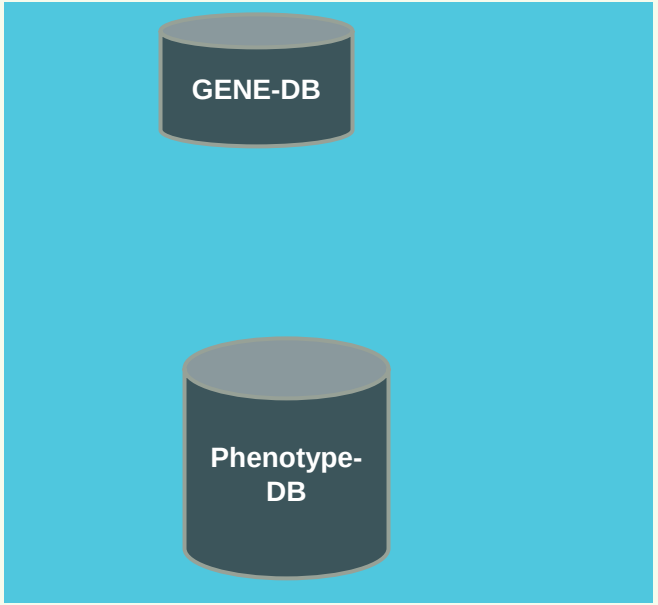




Disease



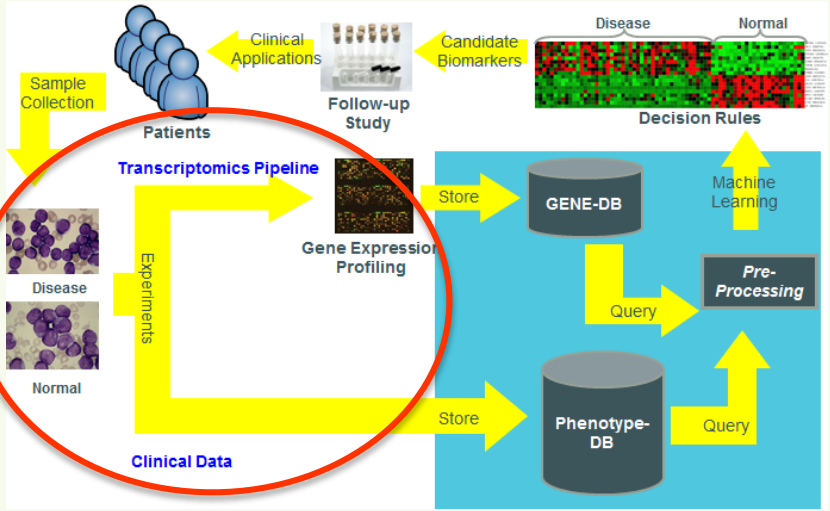
Normal



Workflow „Parts“

Data

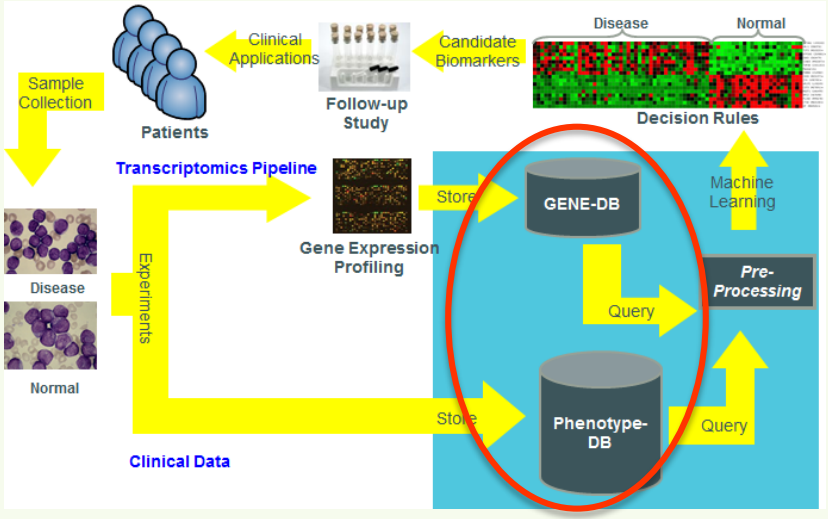
- The goal is to find differences between patients with different “properties”
- Thus: we need data from patients having these properties, e.g.
  - a „healthy“ group and a „diseased“ groups or
  - a group with different survival estimations



Workflow „Parts“

Storage

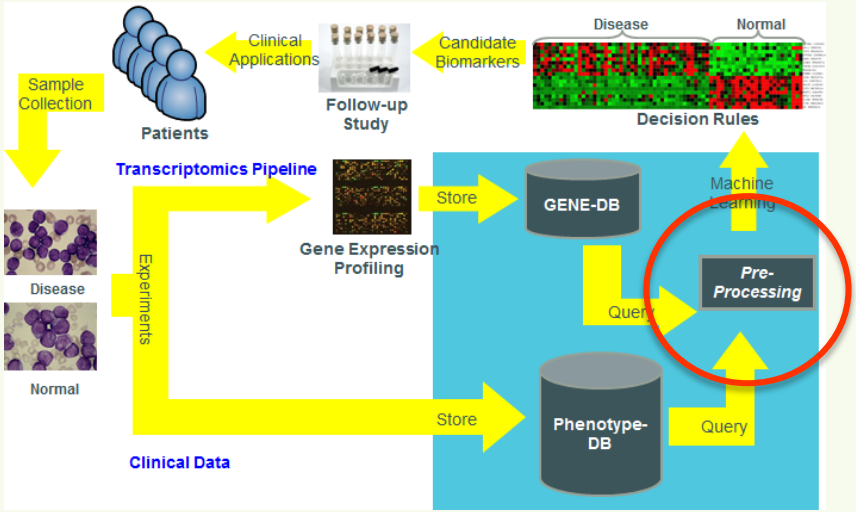
- The data is large, i.e. tens of Gigabytes per patient
- A storage system is needed that allows for fast access
- Hadoop‘s DFS (HDFS) distributes data on available machines



Workflow „Parts“

Pre-Processing

- Depending on the data-type, pre-processing might be needed
- E.g. noise reduction, filtering of „bad“ data, data-imputation, etc.

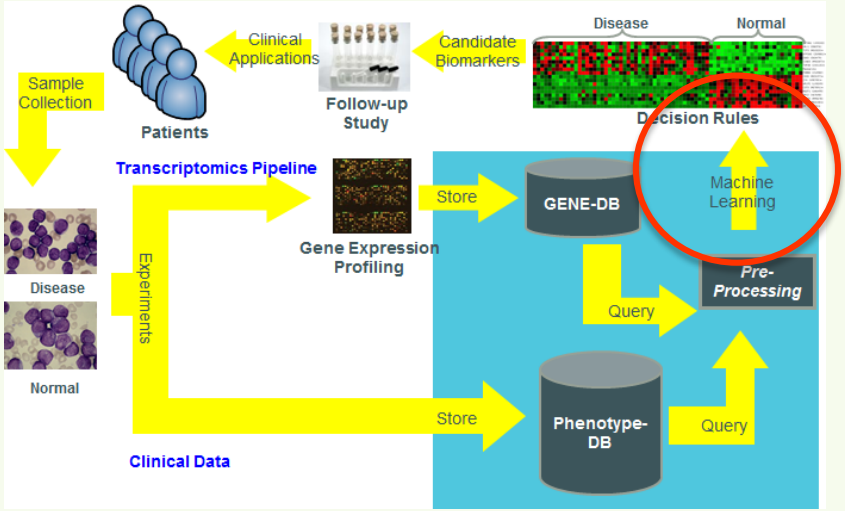


Workflow „Parts“

Machine Learning

- Build model for
  - classification or
  - Regression

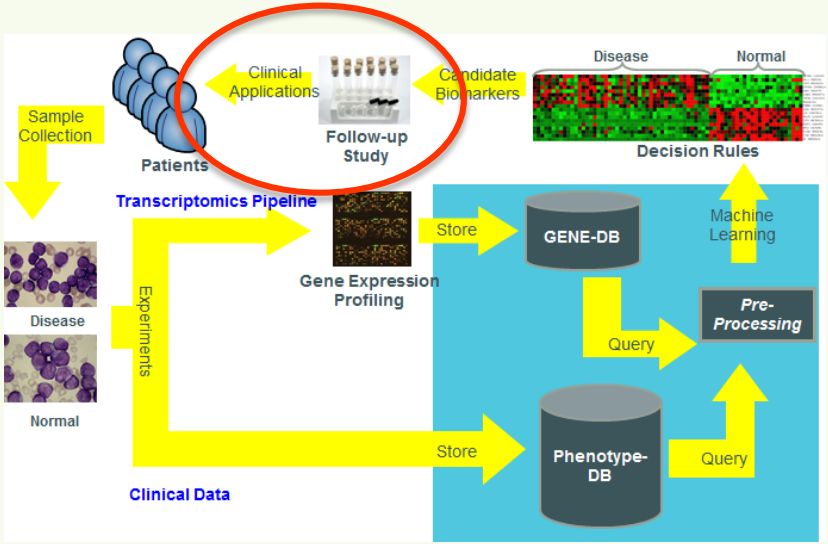
- Also interesting:
  - which features are important?
  - Causality?
  - Model?
  - Prediction?
  - …



Workflow „Parts“

Validation

- Results should always be validated, e.g. in an independent lab



After this talk you now should know…

- … what -omics data is
- … that -omics data can be very large
- … why it is useful to analyze -omics data
- … a method how to analyze large -omics data
- … a framework implementing this method

Spoiler: we will do this during the hands-on session

Thank you for your attention!

Questions?

Tim Conrad

Freie Universität Berlin
conrad@math.fu-berlin.de

Link

www.forschungscampus-modal.de