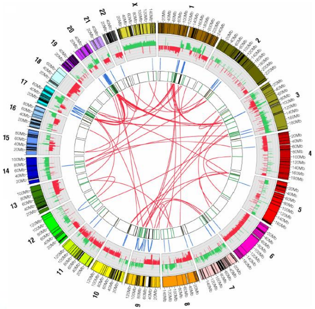
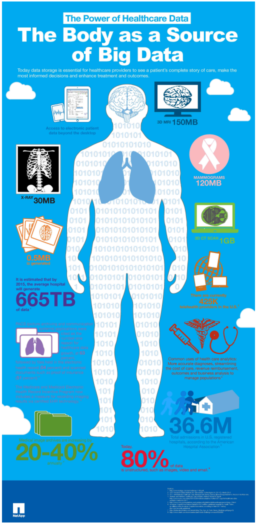
Machine learning for patient stratification from genomic information

Jean-Philippe Vert
jean-philippe.vert@ens.fr

Paris Sciences and Data, December 1, 2016

Big data for health




The Power of Healthcare Data
The Body as a Source of Big Data

Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.

Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 30MB
MAMMOGRAMS 120MB
3D CT SCAN 1GB
0.5MB is generated

It is estimated that by 2015, the average hospital will generate 665TB of data[1]

There are currently 425K telehealth providers in the U.S.[7]

PACS (picture archiving and communication systems) applications were cited as the number-one reason for healthcare data growth, at 63 percent, followed by files held in the electronic health record (54 percent) and scanned documents such as proof of insurance (51 percent).[3]

Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations[4]

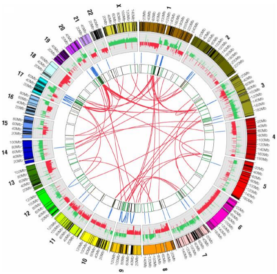
The Medicare and Medicaid Electronic Health Record Incentive Program now includes a measure for recording imaging results via certified EHR technology.[5]

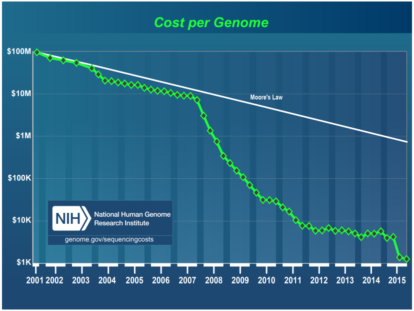
36.6M Total admissions in U.S. registered hospitals, according to the American Hospital Association[6]

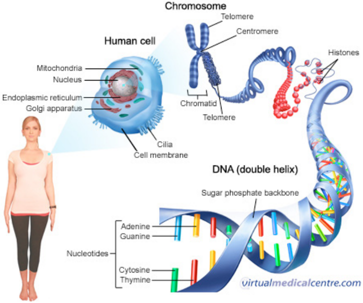
Medical image archives are increasing by 20-40% annually[1]

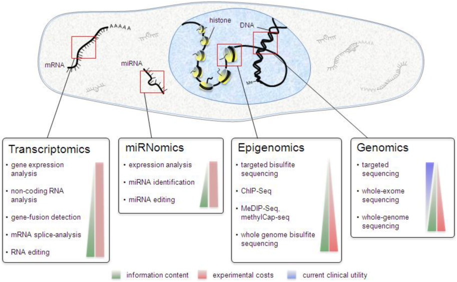
Today, 80% of data is unstructured, such as images, video and email.[8]

Molecular data









Transcriptomics
- gene expression analysis
- non-coding RNA analysis
- gene-fusion detection
- mRNA splice-analysis
- RNA editing

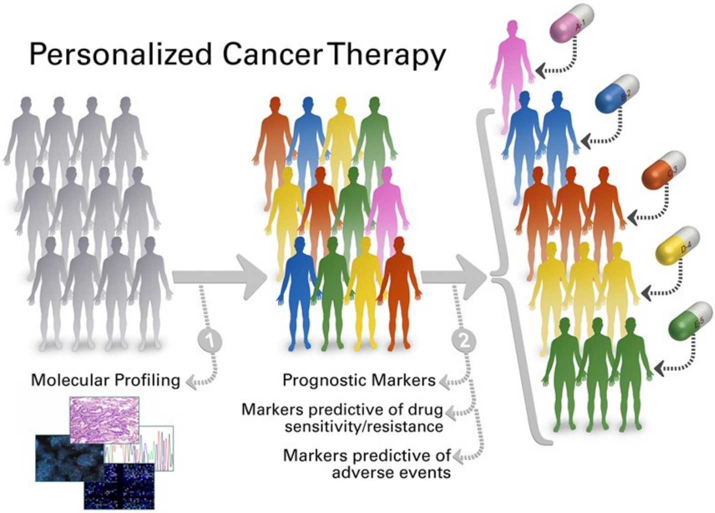
miRNomics
- expression analysis
- miRNA identification
- miRNA editing

Epigenomics
- targeted bisulfite sequencing
- ChIP-Seq
- MeDIP-Seq, methylCap-seq
- whole genome bisulfite sequencing

Genomics
- targeted sequencing
- whole-exome sequencing
- whole-genome sequencing

🟩 information content 🟥 experimental costs 🟦 current clinical utility

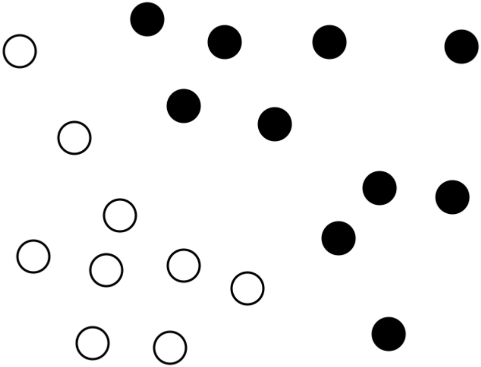
Opportunity


Personalized Cancer Therapy
1
Molecular Profiling
2
Prognostic Markers
Markers predictive of drug sensitivity/resistance
Markers predictive of adverse events
A-1
B-2
C-3
D-4
E-5

https://pct.mdanderson.org

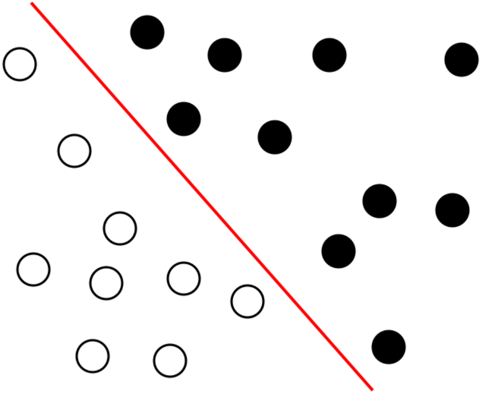
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where n (=19) patients $>>$ p (=2) markers



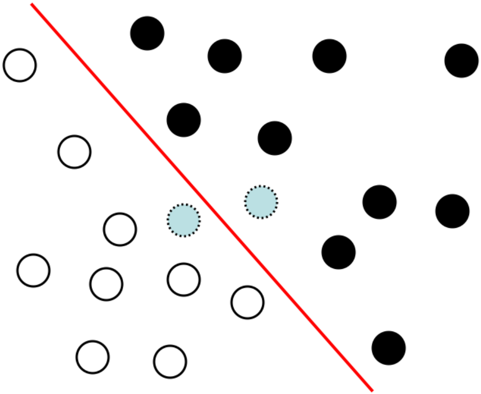
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where n (=19) patients $>>$ p (=2) markers



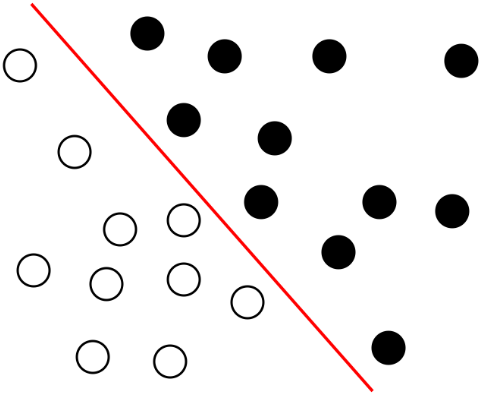
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where *n* (=19) patients >> *p* (=2) markers



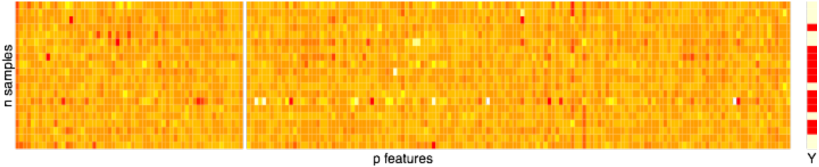
Learning from data: supervised classification/regression

- Patients with VS without relapse in 5 years
- Case where n (=19) patients $>>$ p (=2) markers

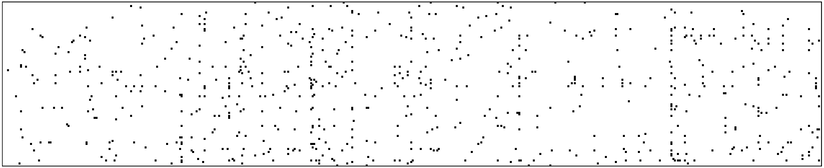


Real data: $n << p$

- Gene expression



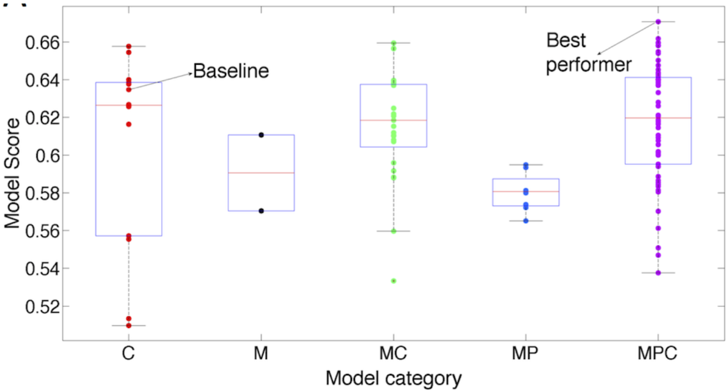
- Somatic mutations



- $n = 10^2 \sim 10^4$ (patients)
- $p = 10^4 \sim 10^7$ (genes, mutations, copy number, ...)
- Data of various nature (continuous, discrete, structured, ...)
- Data of variable quality (technical/batch variations, noise, ...)

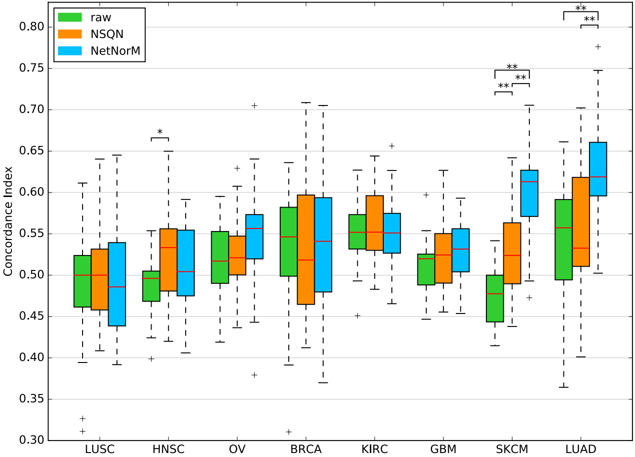
Consequence: limited accuracy

Breast cancer prognosis competition, $n = 2000$, Bilal et al (2013)



- C: 16 standard clinical data (age, tumor size, ...)
- M: 80k molecular features (gene expression, DNA copy number)
- P: incorporate prior knowledge

Example: survival prediction from somatic mutations



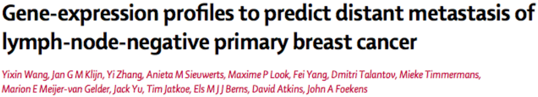
- Data from TCGA (3.3k samples, 8 cancer types, >10k genes)
- Survival SVM on raw binary data, or processed by NSQN (Hofree et al., 2013) or NetNorm (Le Morvan et al., 2016).

Consequence: unstable biomarker selection

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

*Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands
‡ Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034,



Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

70 genes (Nature, 2002)

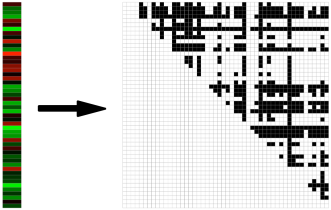
76 genes (Lancet, 2005)

3 genes in common

van 't Veer et al. (2002); Wang et al. (2005)

Some research directions

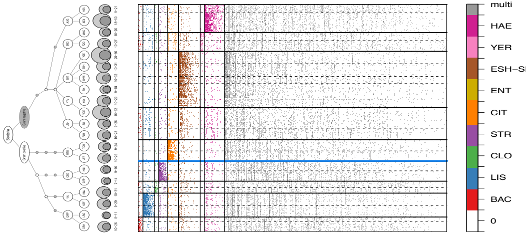
- Find a better representation



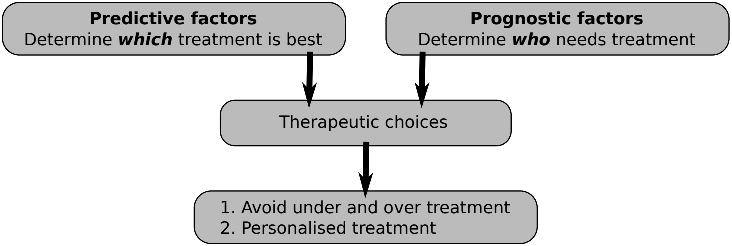
One sample x
p features

Mapping f(x)
p(p-1)/2 bits

- Incorporate prior knowledge



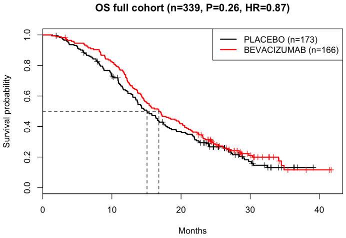
From prognostic to predictive models

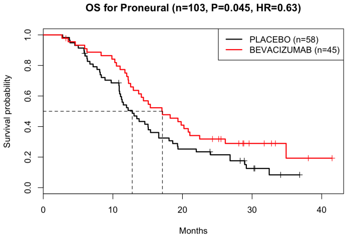

Predictive factors
Determine ***which*** treatment is best
Prognostic factors
Determine ***who*** needs treatment
Therapeutic choices
1. Avoid under and over treatment
2. Personalised treatment

- Prognostic:
  - Predict outcome Y of a disease on an untreated individual X
  - Standard supervised learning: model $Y = f(X)$ from observations of (X_i, Y_i) pairs
- Predictive:
  - Predict the benefit in outcome Y of a treatment A on an individual X
  - We observe (X_i, A_i, Y_i) but want to model $Y = f(X, A_1) - f(X, A_2)$
  - For each X we only observe the outcome Y under one treatment A (cf e-marketing)

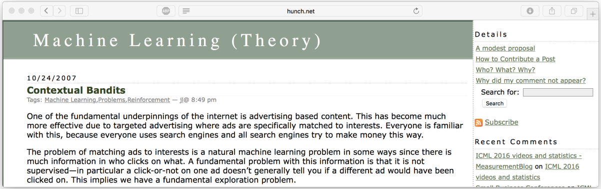
Clinical trials for precision medicine?

1. Meta-analysis of clinical trials (typically *A/B testing*) to estimate predictive models


OS full cohort (n=339, P=0.26, HR=0.87)


OS for Proneural (n=103, P=0.045, HR=0.63)

2. Dynamic trial to jointly optimize the predictive model and its performance (contextual multi-armed bandit problem)



Machine Learning (Theory)

10/24/2007

Contextual Bandits

Tags: Machine Learning,Problems,Reinforcement — jl@ 8:49 pm

One of the fundamental underpinnings of the internet is advertising based content. This has become much more effective due to targeted advertising where ads are specifically matched to interests. Everyone is familiar with this, because everyone uses search engines and all search engines try to make money this way.

The problem of matching ads to interests is a natural machine learning problem in some ways since there is much information in who clicks on what. A fundamental problem with this information is that it is not supervised—in particular a click-or-not on one ad doesn't generally tell you if a different ad would have been clicked on. This implies we have a fundamental exploration problem.

Details

A modest proposal

How to Contribute a Post

Who? What? Why?

Why did my comment not appear?

Search for:

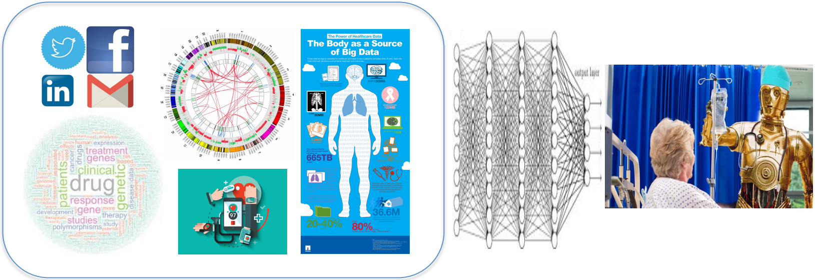
Search

Subscribe

Recent Comments

ICML 2016 videos and statistics - MeasurementBlog on ICML 2016 videos and statistics

Conclusion


The Power of Healthcare Data
The Body as a Source of Big Data
665TB
36.6M
20-40%
80%
output layer

- Lots of data
- $n << p$ is the rule, more and more...
- Limited impact so far for patients
- Active research
  - new representations $x \to \Phi(x)$
  - new learning techniques (structured sparsity, regularization, ...)
  - new experimental design strategies (contextual bandit)

References

M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker. Network-based stratification of tumor mutations. *Nat Methods*, 10(11):1108–1115, Nov 2013. doi: 10.1038/nmeth.2651. URL `http://dx.doi.org/10.1038/nmeth.2651`.

L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, 415(6871):530–536, Jan 2002. doi: 10.1038/415530a. URL `http://dx.doi.org/10.1038/415530a`.

Y. Wang, J. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talantov, M. Timmermans, M. Meijer-van Gelder, J. Yu, T. Jatkoe, E. Berns, D. Atkins, and J. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet*, 365(9460):671–679, 2005. doi: 10.1016/S0140-6736(05)17947-1. URL `http://dx.doi.org/10.1016/S0140-6736(05)17947-1`.