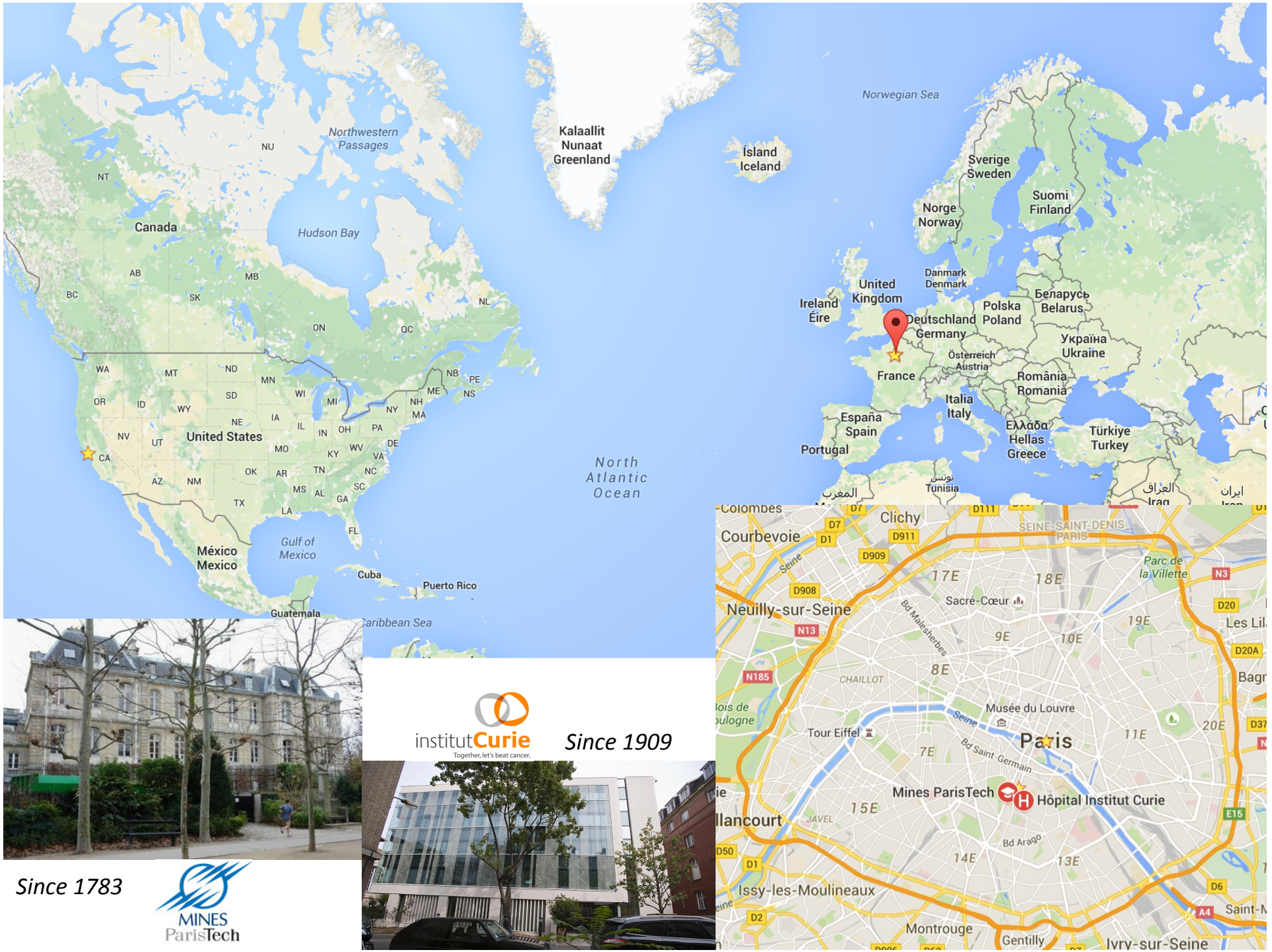
Can big data cure cancer?

Jean-Philippe Vert

Miller Institute, UC Berkeley, Nov 24, 2015



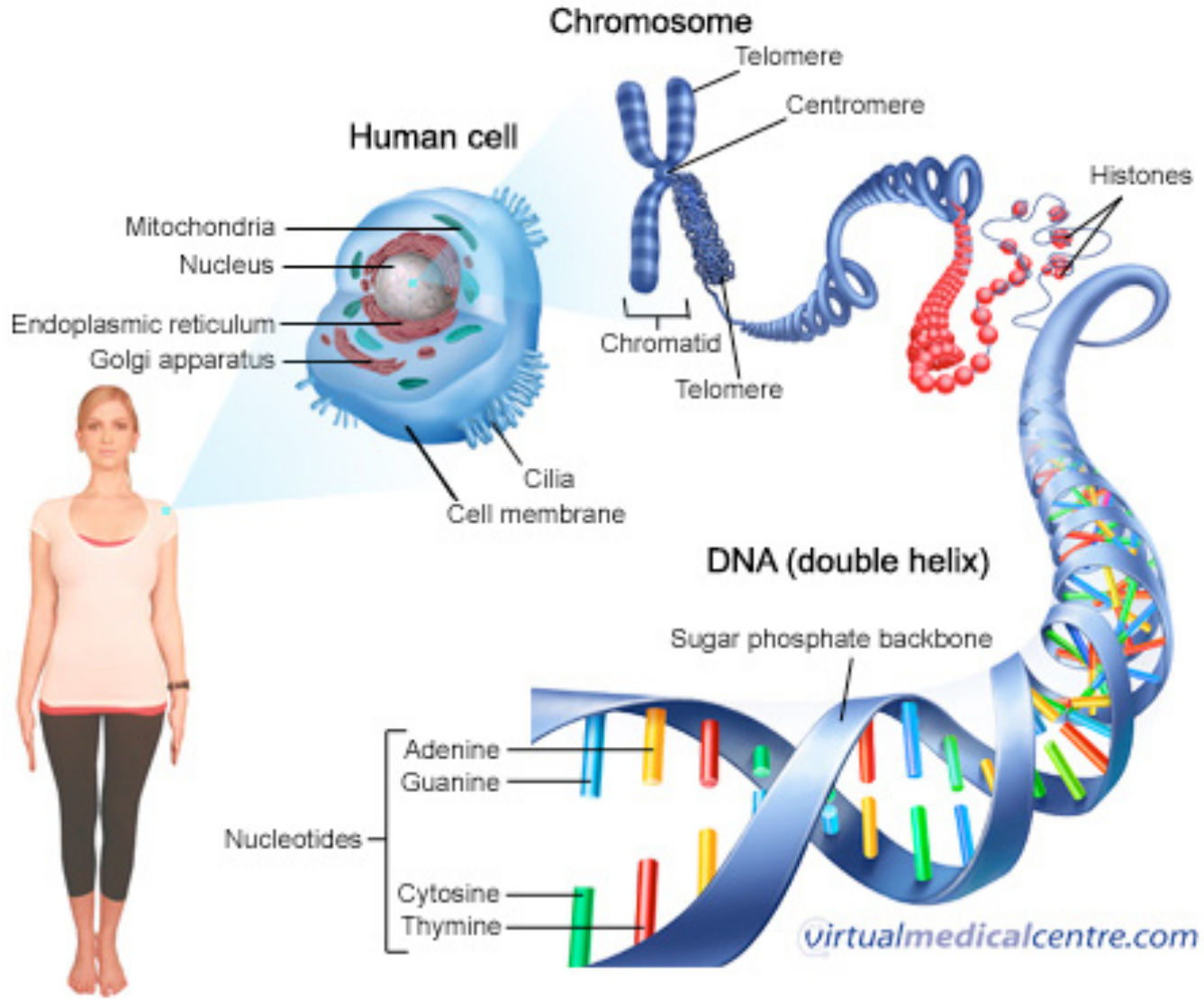
institutCurie
Together, let's beat cancer.

Since 1909

Since 1783

MINES ParisTech

Your body contains 100 trillions cells
Each cell contains a copy of the genome



The genome (DNA) differs:

- Between **species**

  - *>1 nucleotide / 100*

- Between **individuals**

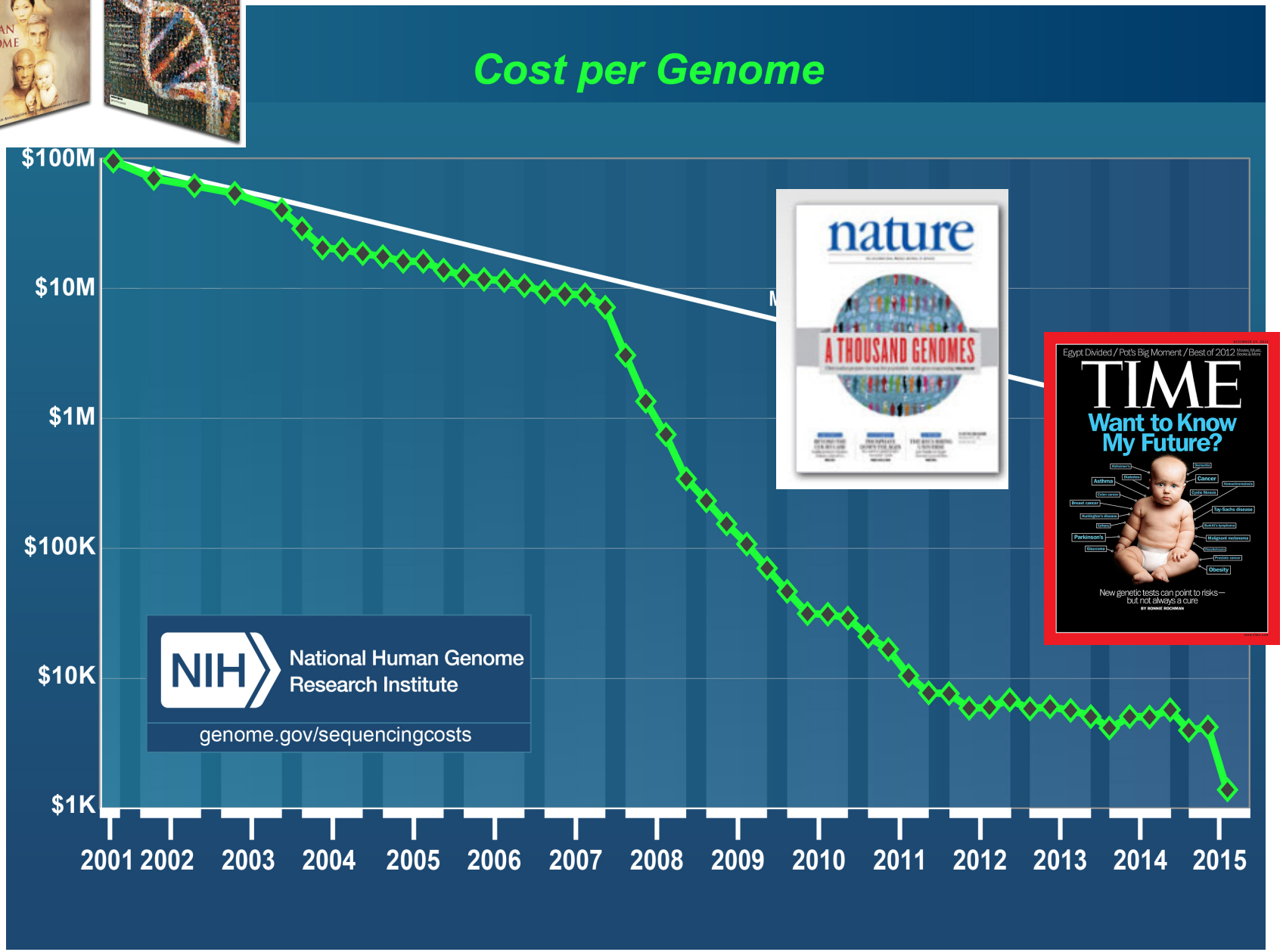
  - *1 nucleotide / 1,000*

- Between **cells**

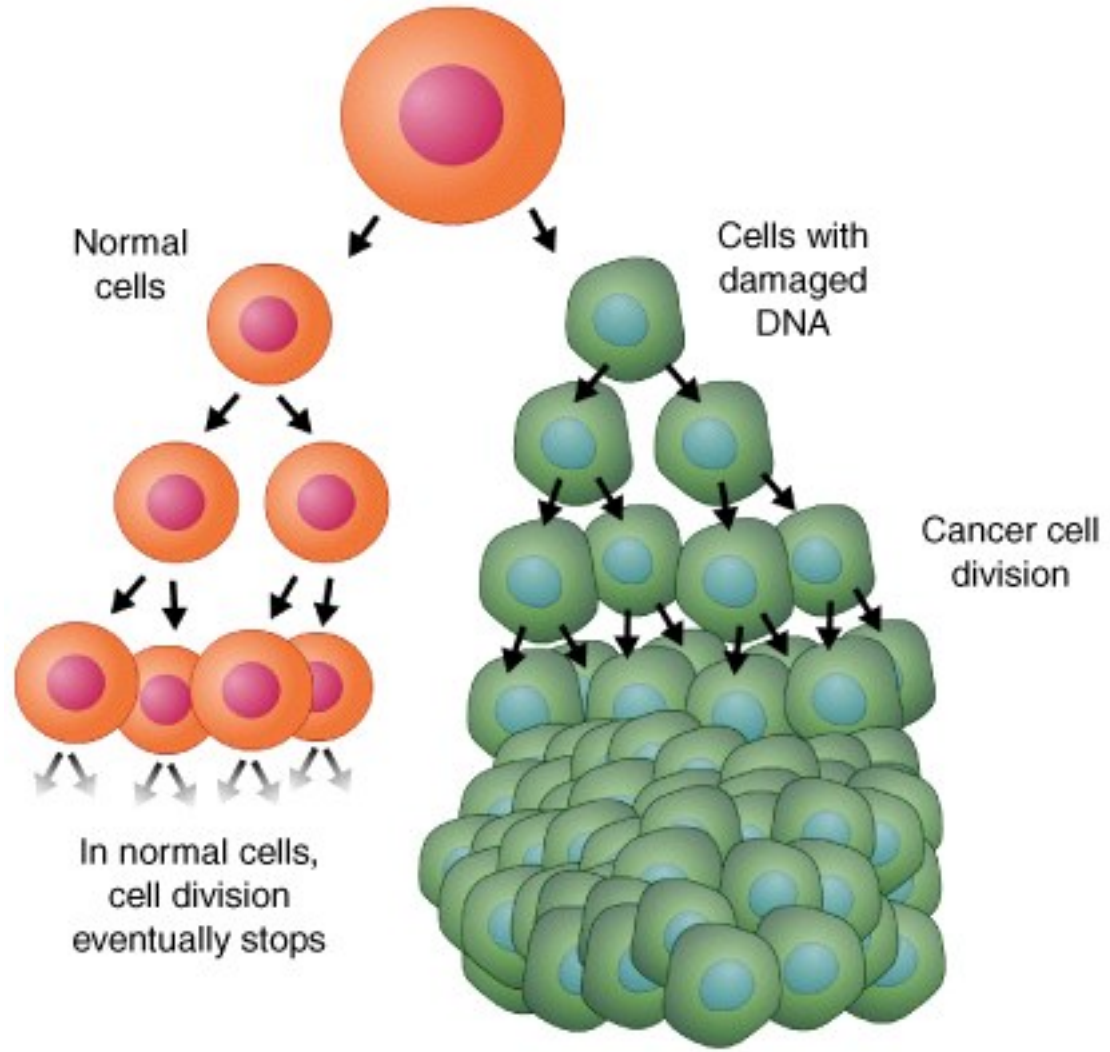
  - *1 nucleotide / 100,000,000*

  (~10 mutations per cell division)

DNA sequencing

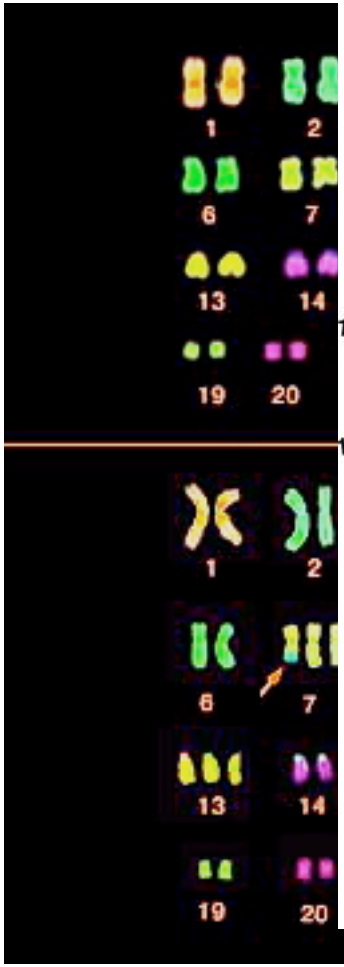


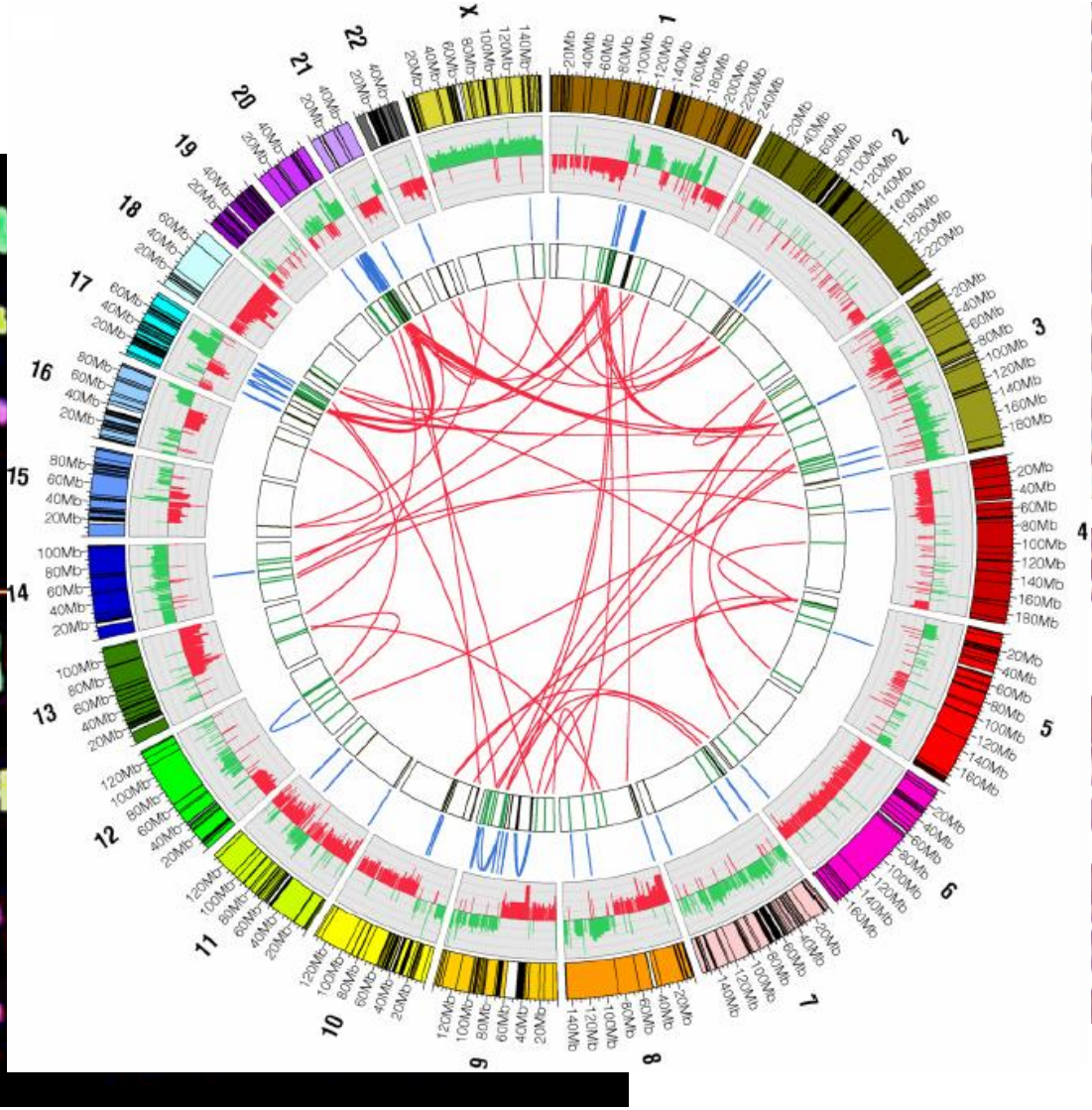
What is cancer(s)?

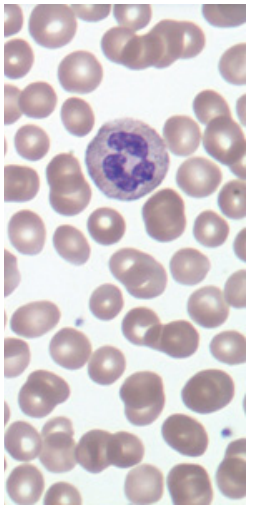


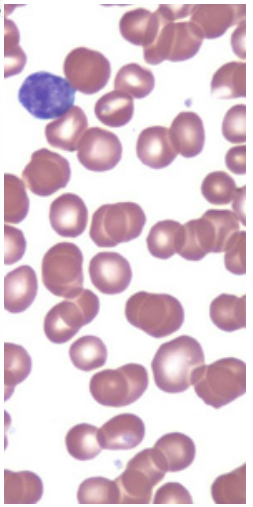
http://rise.duke.edu/seek/pages/page.html?0205

New view of cancer









Towards precision medicine

« strategy »

?

1) Human-designed « strategy », or
2) Computer-designed « strategy » ?

1) Human-designed strategy: Example of the SHIVA clinical trial

Table 1 SHIVA treatment algorithm established to select molecularly targeted agents based on the molecular profile

Targets	Targeted therapies	Molecular alterations
KIT, ABL1/2, RET	Imatinib	Activating mutations/amplification
PI3K		
AKT		
PTE		
STK		
INPP		
BRA		
PDG		
EGF		
HER		
SRC		
EPH		
ER,		
AR		
Com		

defi

amplification with an amplicon size of maximum 1 Mb were directly validated by the MBB. If amplicon size >1 and <10 Mb, IHC is required. Comments for tumor suppressor genes, inactivation of tumor suppressor genes implies that the 2 alleles that code for a particular protein are affected: (I) homozygous deletion (loss of 2 alleles); (II) heterozygous deletion: Loss of one allele if the second hold an inactivation mutation or can be validated by loss of expression using IHC; (III) loss is defined by 1 copy for diploid tumors and 1 or 2 copies for tetraploid tumors; (IV) deletion corresponds to 0 copy.

Limits of human-designed strategies:

- Limited to **what we know** (or believe)
- Limited to a **few alterations**, and a **few drugs**
- No **combinatorial** rule
- No **weighting** of evidences
- No **combination** of drugs
- … and **did not succeed** in the clinical trial

institutCurie
Together, let's beat cancer.

2) Computer-designed strategy

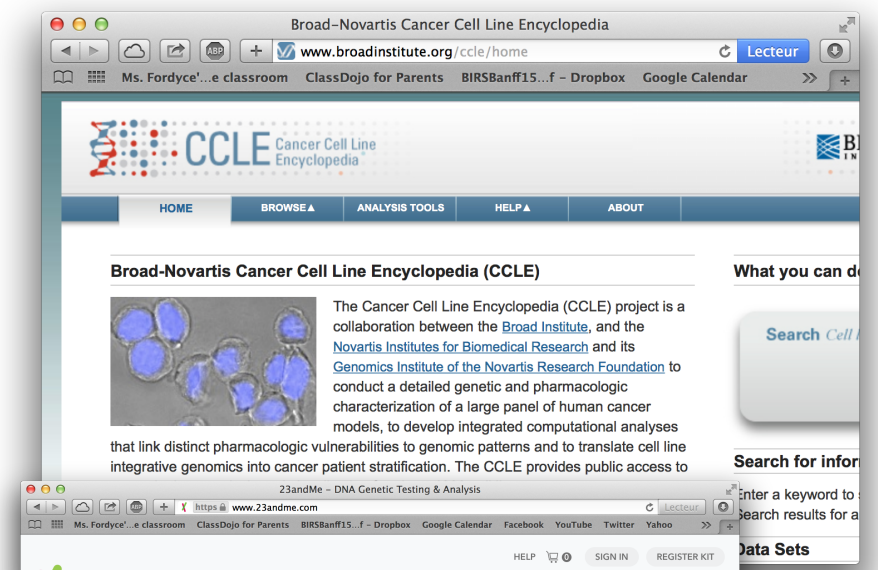
1. **Collect molecular data** about many individuals

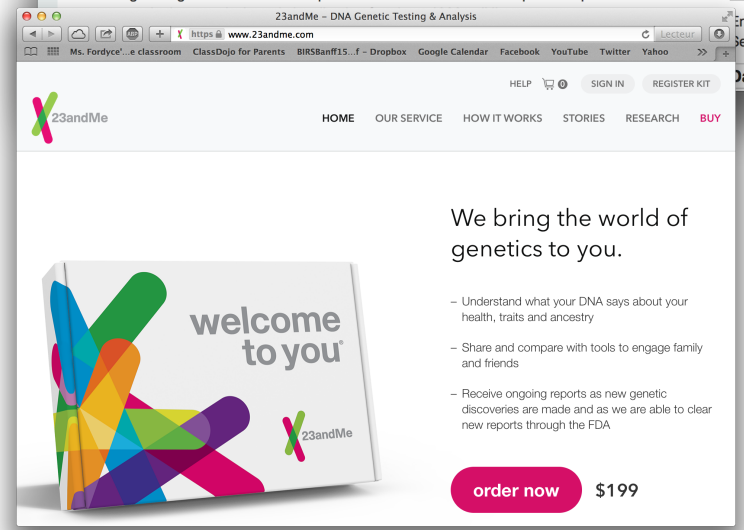
2. **Collect the response** to treatment

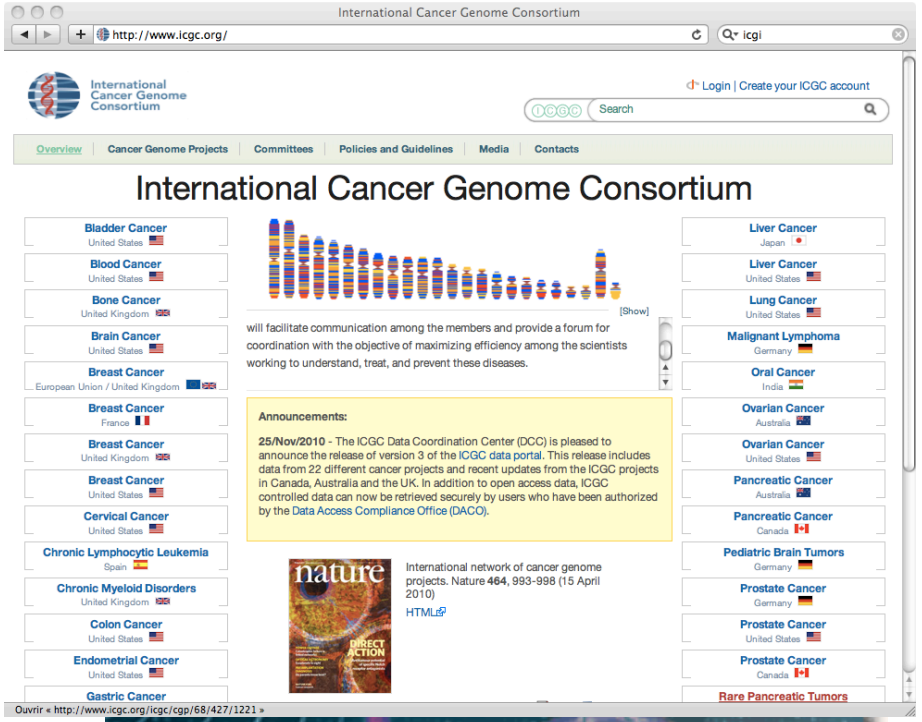
3. **Let the computer** figure out how to **predict** the response from the molecular data

Collecting data: ongoing

- http://aws.amazon.com/1000genomes/

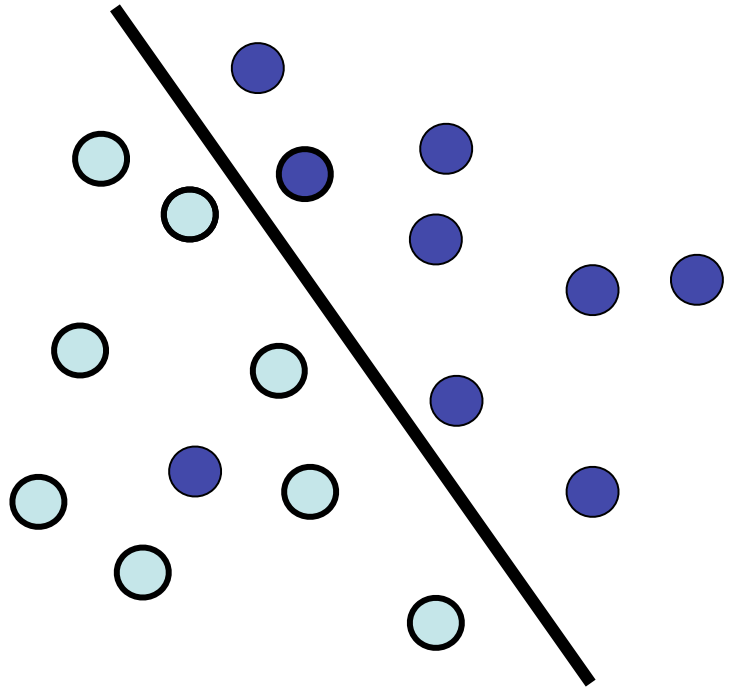






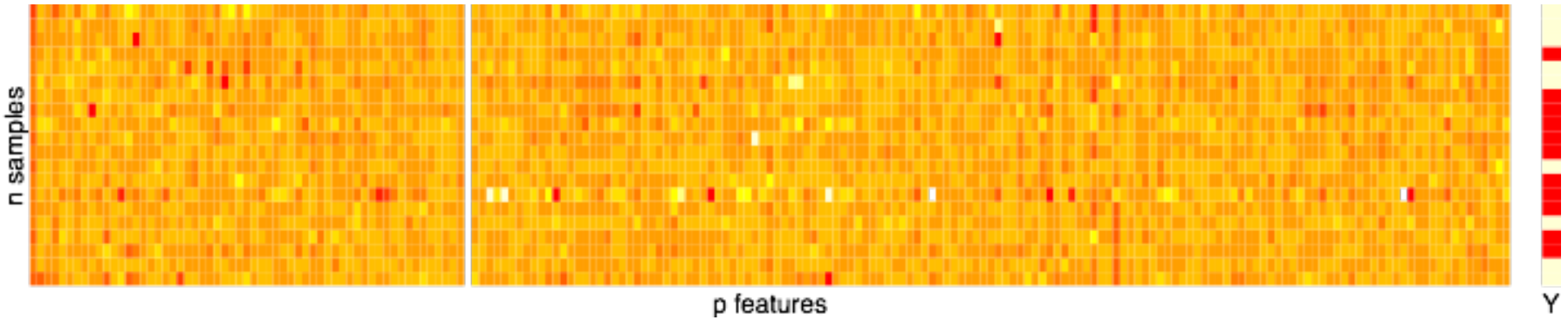
Let the computer « learn » the rule

(a.k.a. machine learning)



n=15 samples >> p=2 descriptors (easy)

Machine learning is hard when n << p

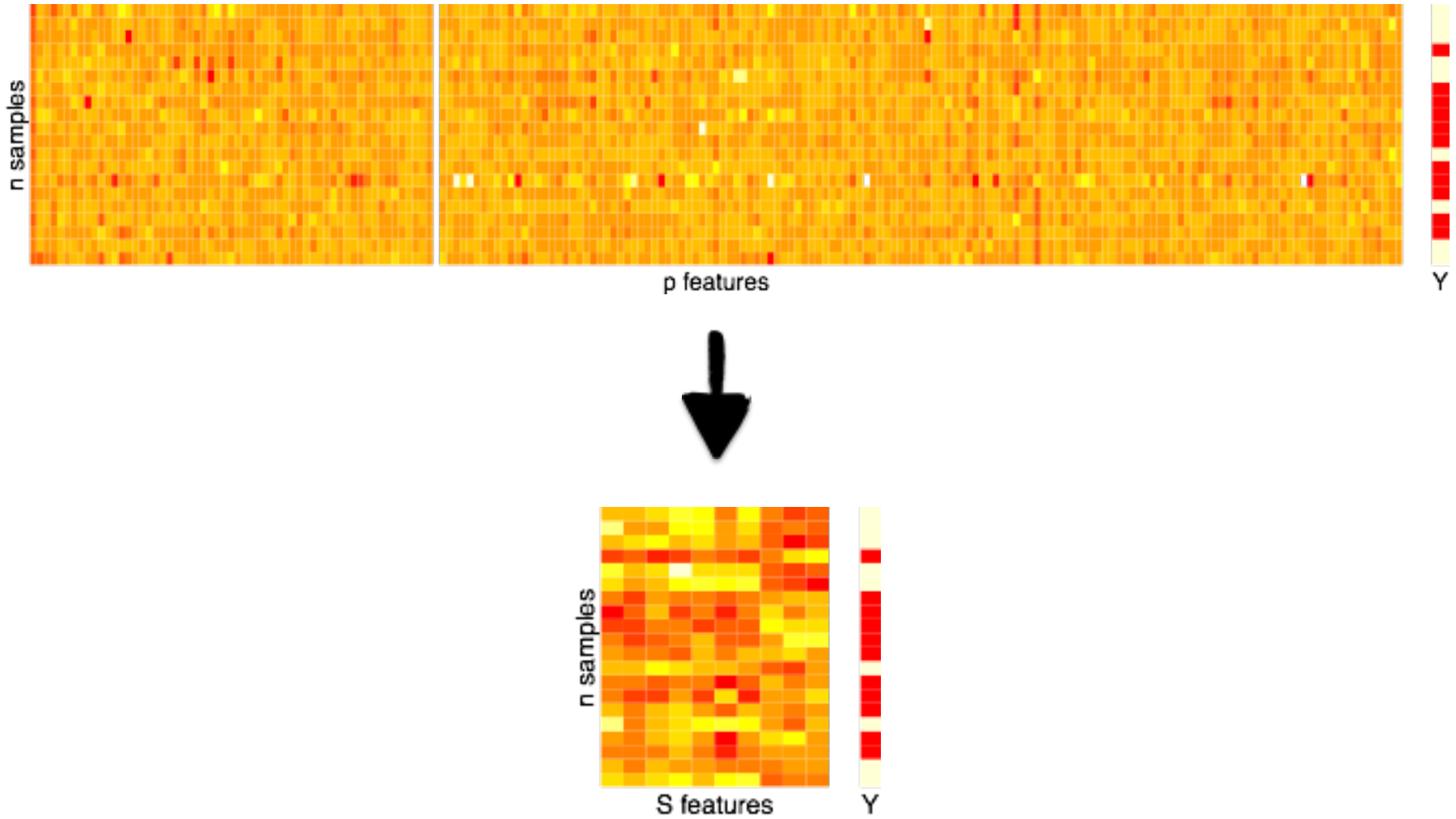


n = 1e2 ~ 1e4
(patients)

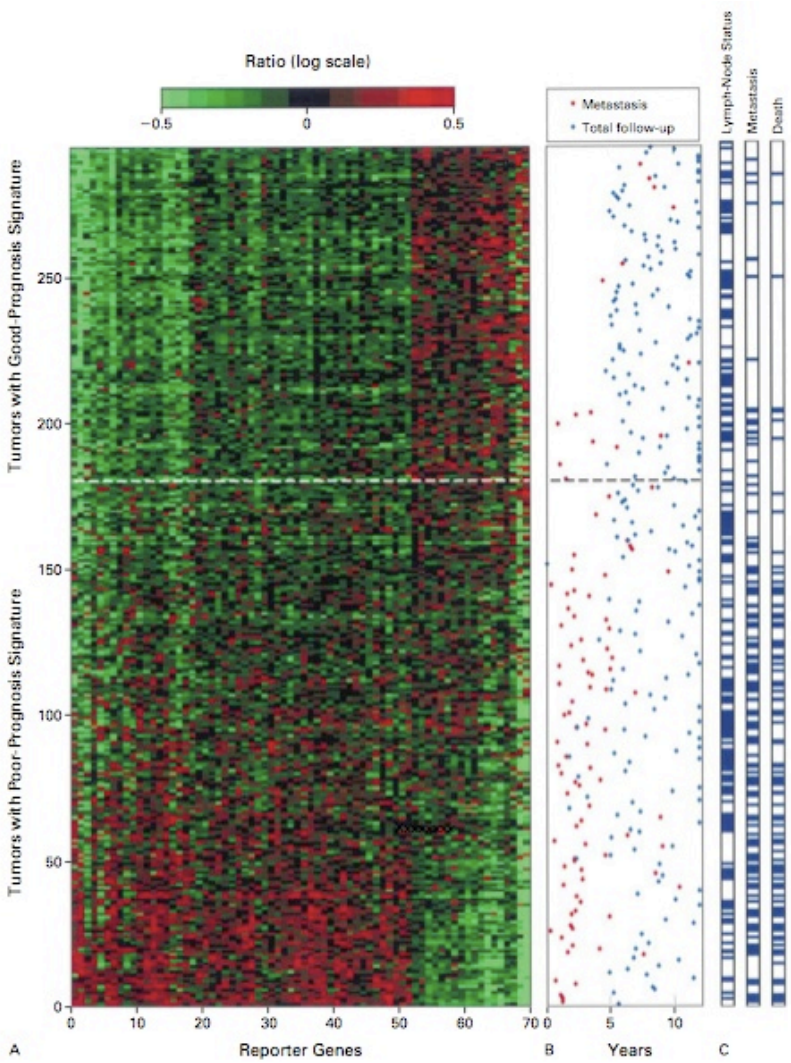
p = 1e4 ~ 1e7
(genes, mutations,
copy numbers, …)

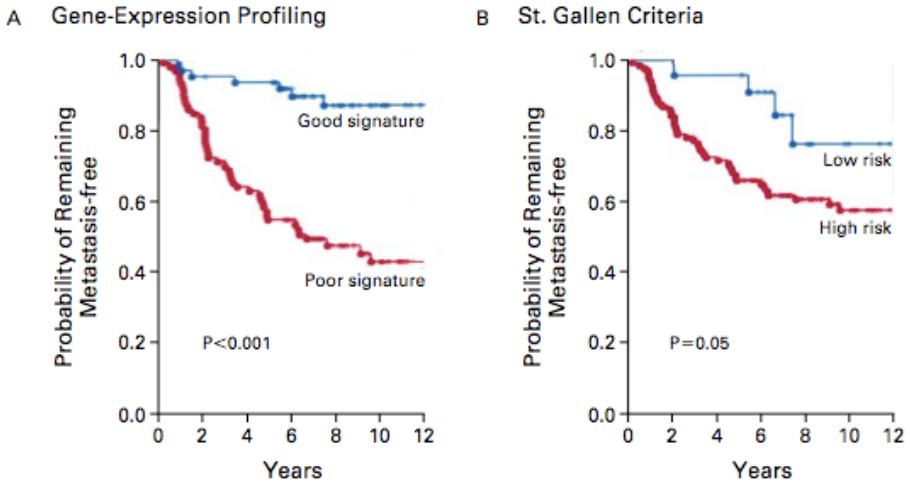
One solution: reduce dimension

(a.k.a. « feature selection », « molecular signature »)



Example: Mammaprint, the 70-gene Breast cancer prognostic signature





(Van de Vijver et al 2002)

But…

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,

70 genes (Nature, 2002)

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

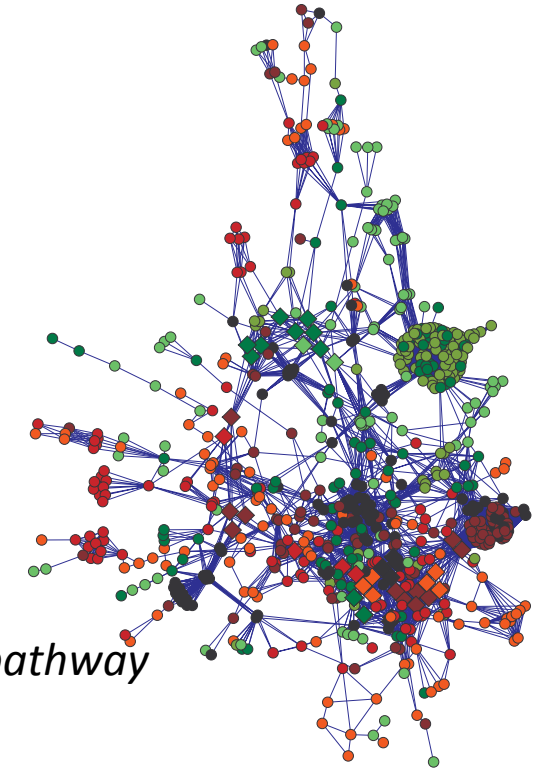
Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

76 genes (Lancet, 2005)

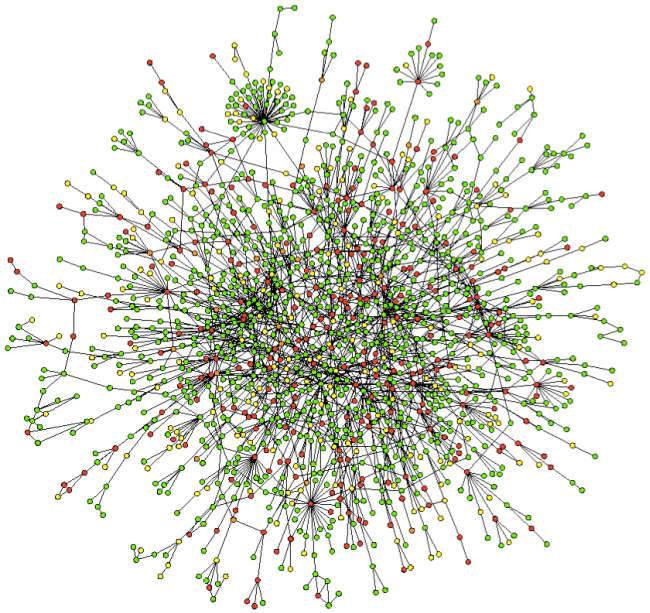
Only 3 genes in common

… and not really better than choosing 70 genes at random!
(Haury et al., PLoS One 2011)

Improving feature selection with prior knowledge



Metabolic pathway

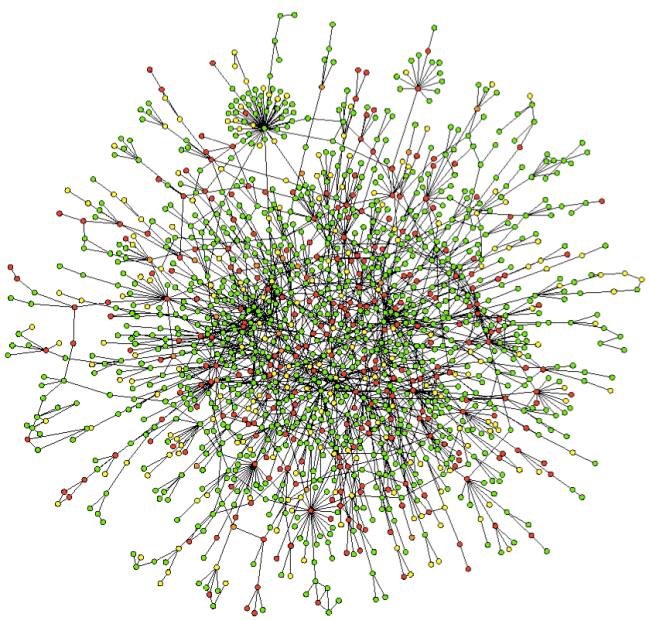


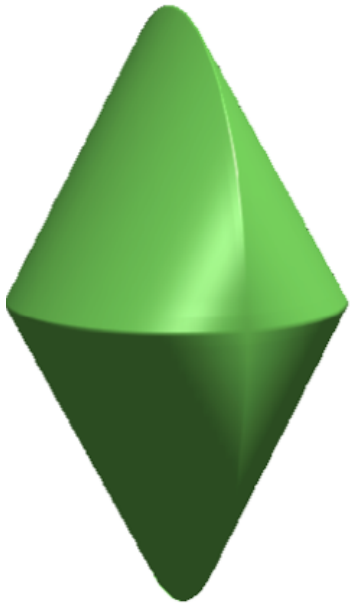
Protein-Protein interaction network

Can we « force » the signature to be « coherent » with a known gene network?

Example: the graph lasso

- **Step 1**: Using the network, define a subset of « candidate » signatures





$$\Omega(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta$$

(a convex body in p dimensions)

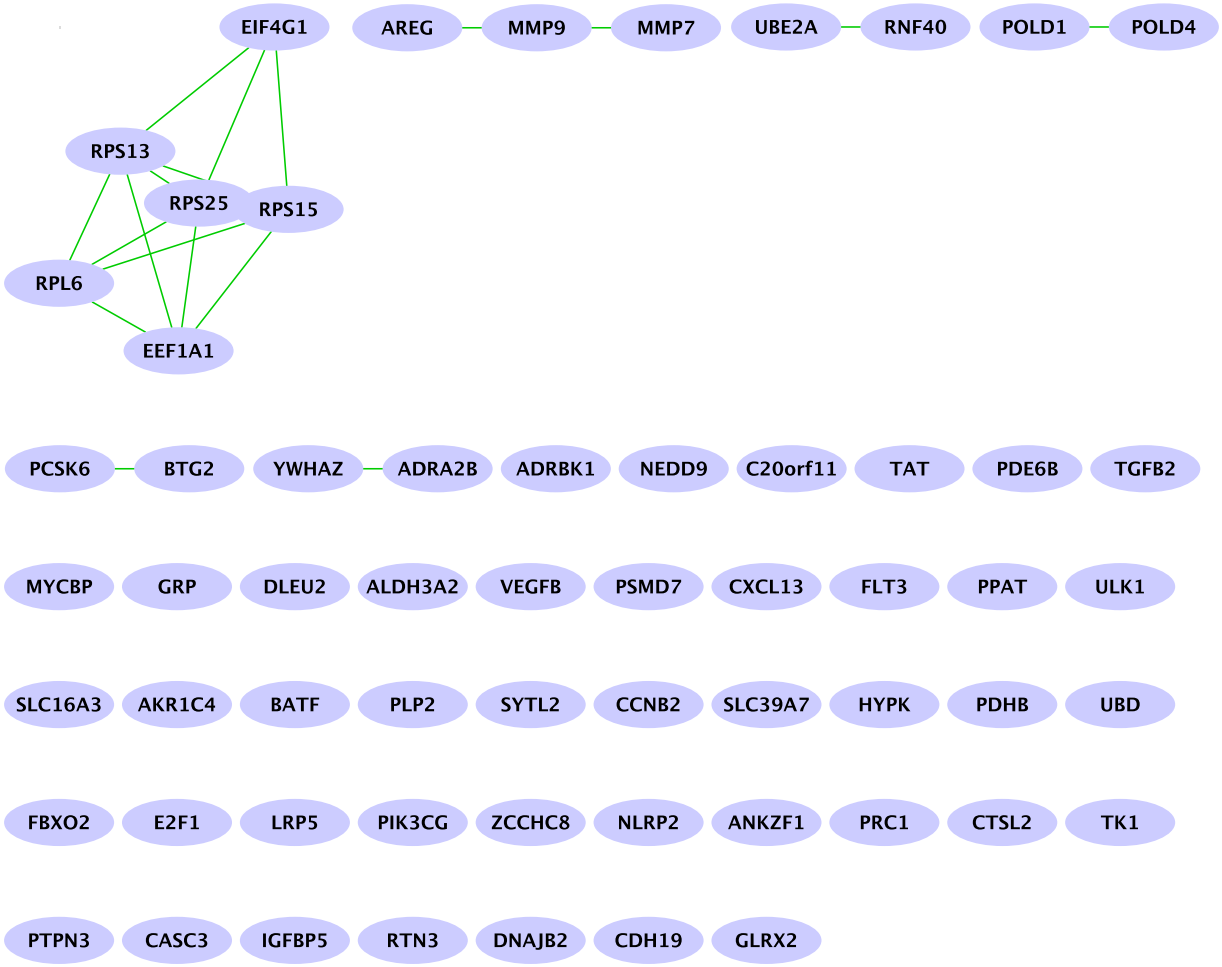
- **Step 2**: Among the candidates, find the best signature to explain the data

$$\min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \Omega(\beta)$$

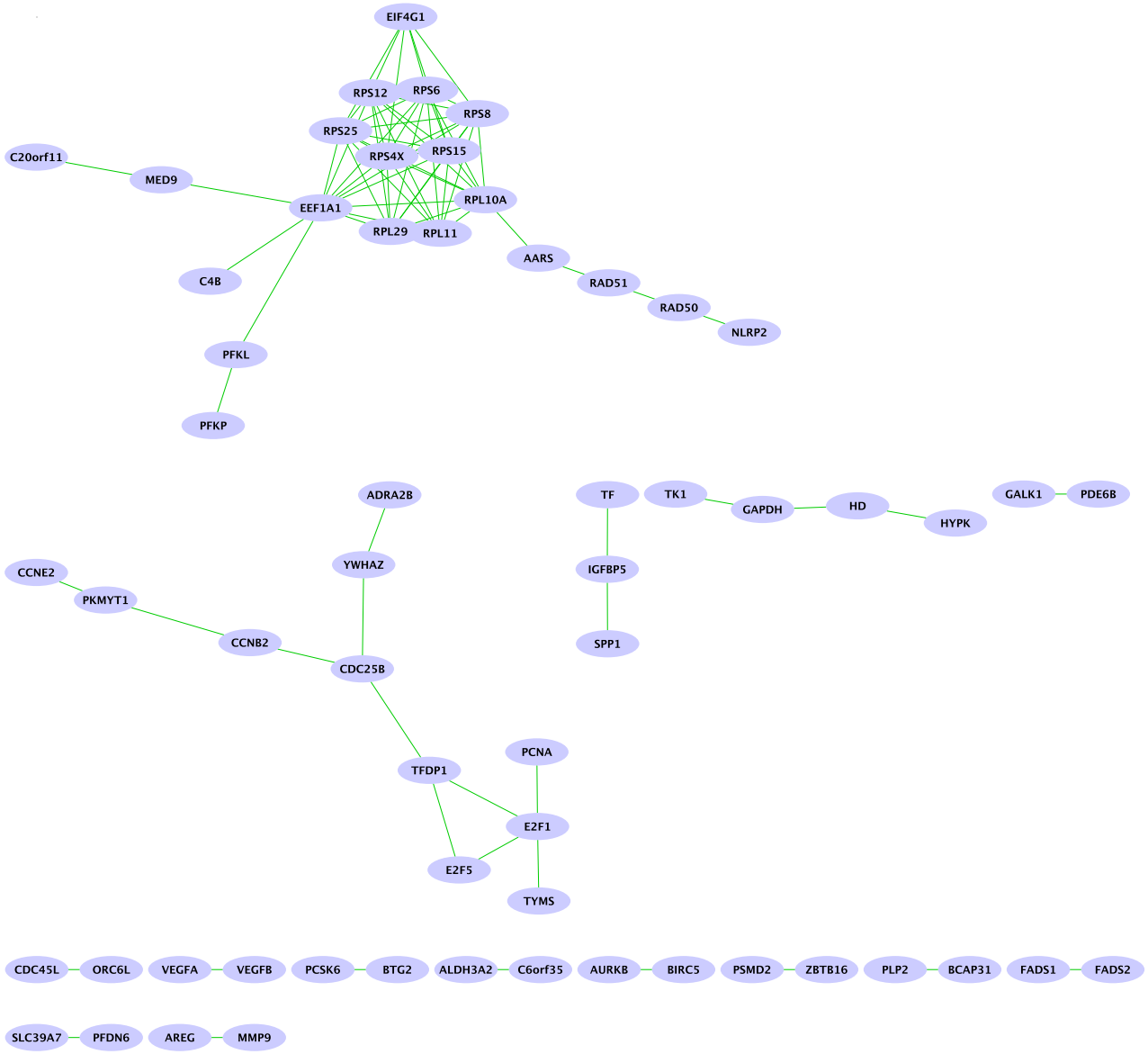
(convex optimization)

(Jacob et al 2009)

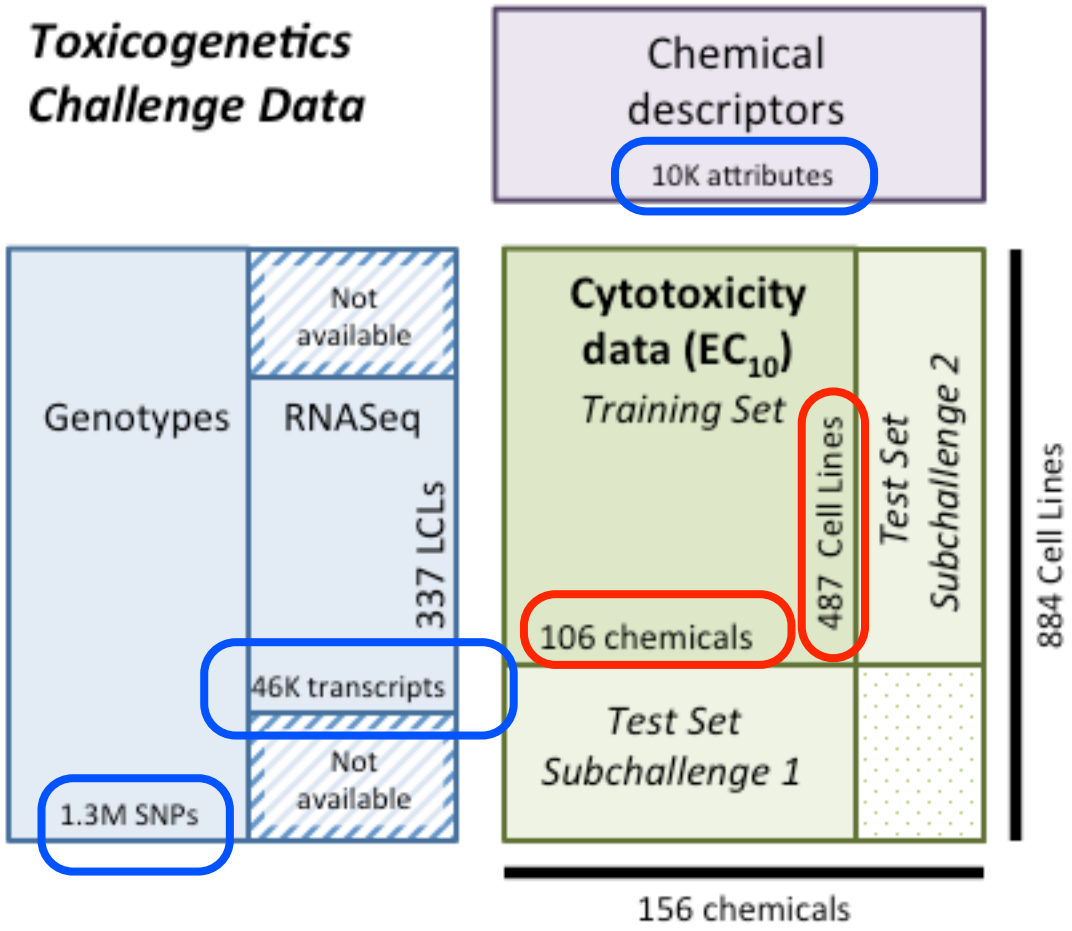
Classical signature (accuracy = 0.61)



Graph lasso signature (accuracy=0.64)



Ongoing project: multiple drugs

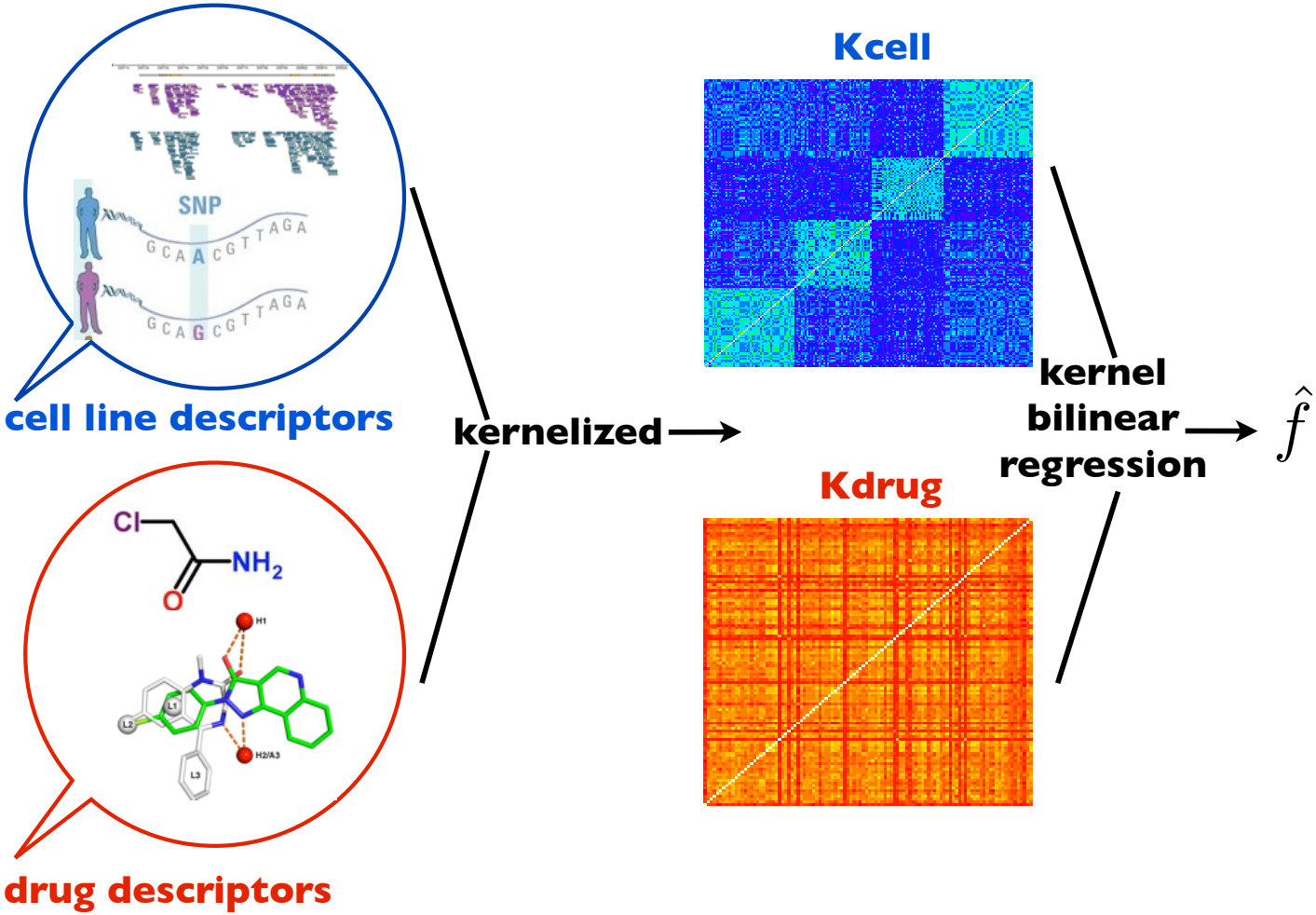


n = 5e4

p = 1e10

Collaboration S. Dudoit (UC Berkeley), R. Bourgon (Genentech)

Our approach



Somehow it worked…

NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567

https www.synapse.org/#!Synapse:syn1761567/wiki/60497

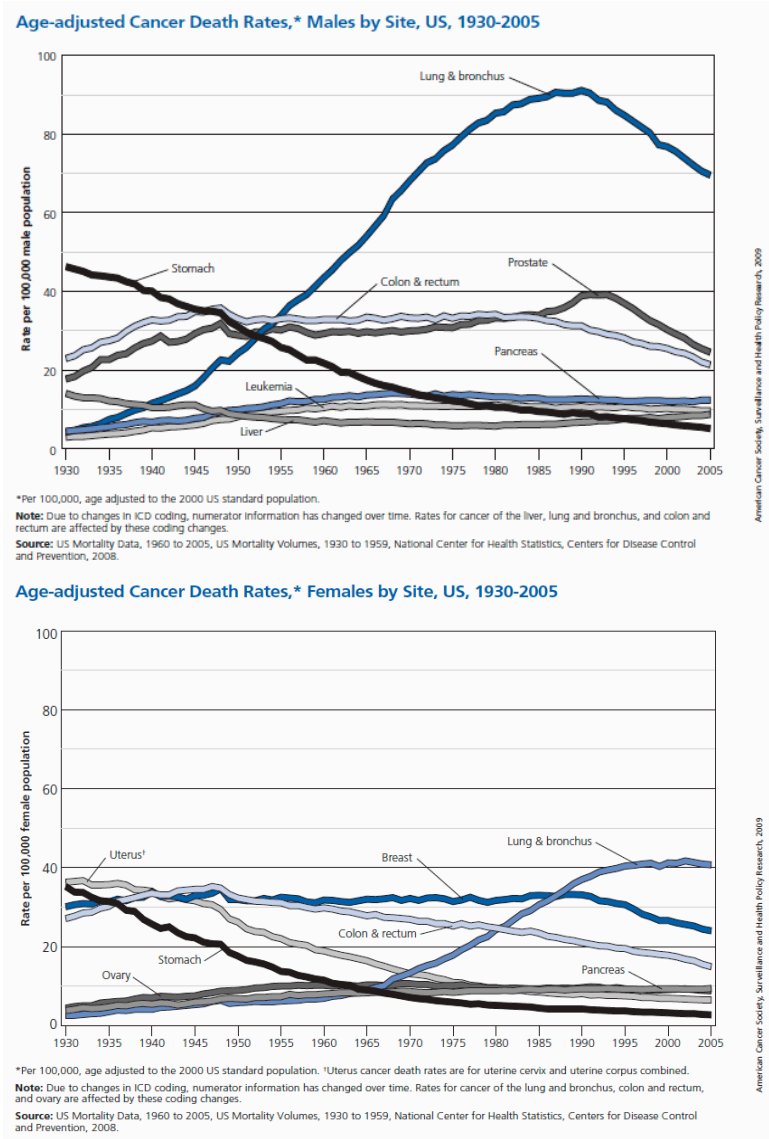
Sage Synapse: Contri... | NIEHS-NCATS-UNC D... | NIEHS-NCATS-UNC... | NIEHS-NCATS-UNC... | NIEHS-NCATS-UNC...

Team	Submission	SynapseID	Mean ranking PCI	Rank PCI	Mean ranking PC	Rank PC	Mean ranking	Rank
Yang_Lab	UTSW_QBRC_kmb310.txt	syn2219079	27.2198	1	31.8681	2	1.5	1.0
CASSIS	Final_prediction_KRR_int_empiric...	syn2224212	31.5714	2	34.3516	4	3.0	2.0
amss2012								
UT_CCB								
Yang_Lab								
O6d0A								
Yang_Lab								
CQB								
Yang_Lab	UTSW_QBRC_kmb49.txt	syn2218923	36.2747					
UT_CCB	Prediction_Result_3.txt	syn2227281	41.0659					
D-Tox	ToxSubchallenge_1_prediction_mat...	syn2223065	39.0549					
Yang_Lab	UTSW_QBRC_lm4.txt	syn2223153	38.8132					
CASSIS	Final_prediction_KRR_int_dirac_b...	syn2224209	38.2857					
WarwickDataScience	predictions_subChallenge1_submis...	syn2211154	38.9011					
Kaiju								
CQB								
UTSW_QBRC								

But the best performance is barely better than random

Conclusion

- New opportunities to exploit big data in precision medicine
- Challenging machine learning problems
- Still a long way to go before curing cancer…



Thanks!

The Adolph C. and Mary Sprague Miller Institute for Basic Research in Science
University of California, Berkeley