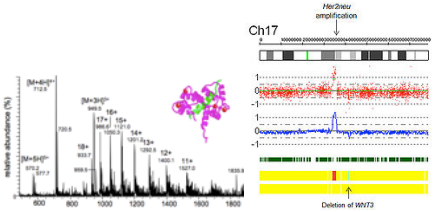
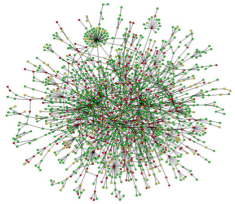
Machine Learning for cancer precision medicine

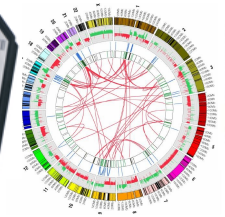
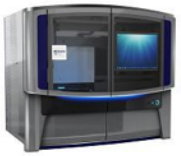
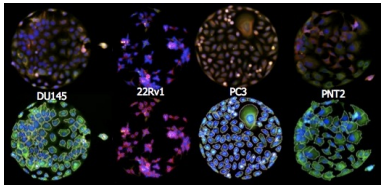
Jean-Philippe Vert

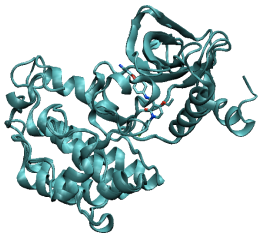
Académie de Médecine, Paris, July 4, 2017

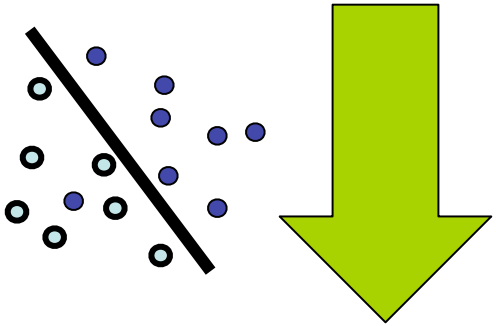
Overview



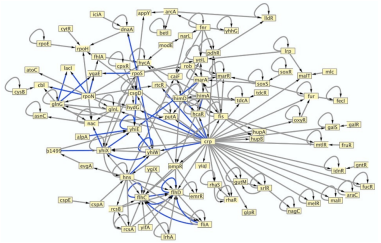






Machine learning

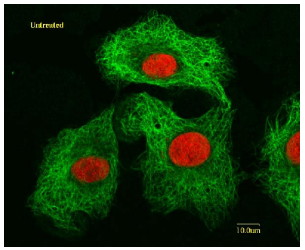


Molecular level

Gene regulation

Epigenetics

Structure/Function prediction

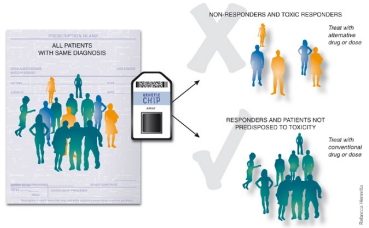


Cellular level

High-content screening

Chemo/Toxicogenomics

Tumour heterogeneity



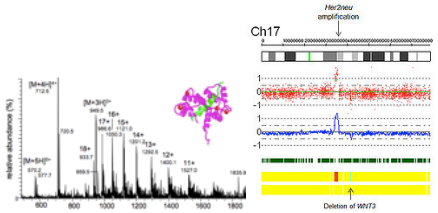
Precision medicine

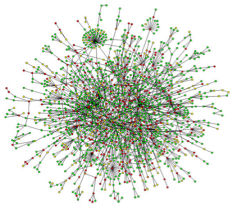
Patient stratification

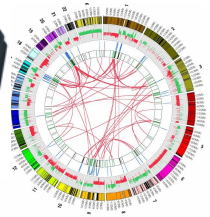
Prognostic / Predictive

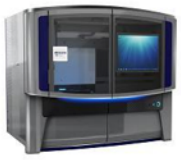
Side effect prediction

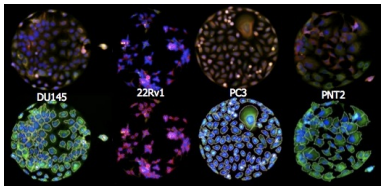
Machine Learning?

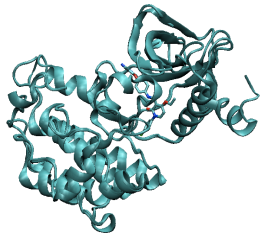


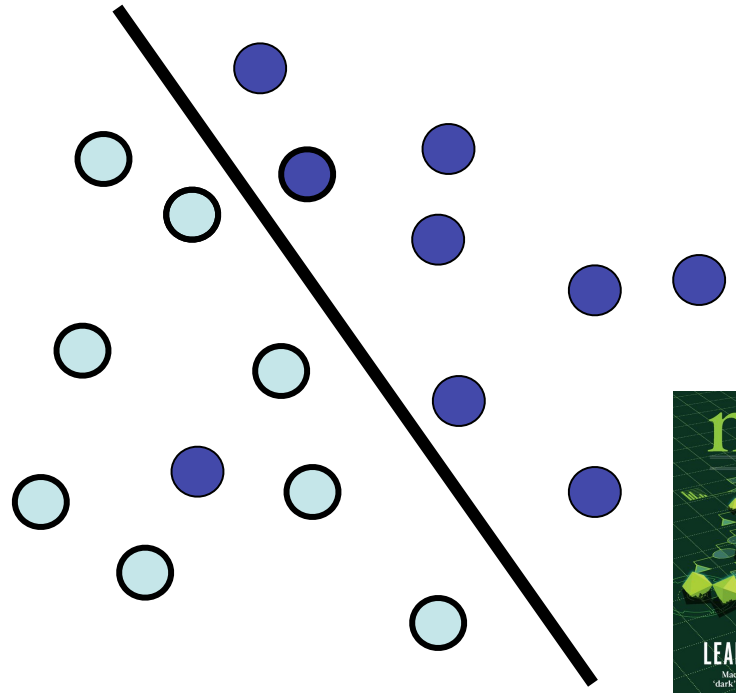




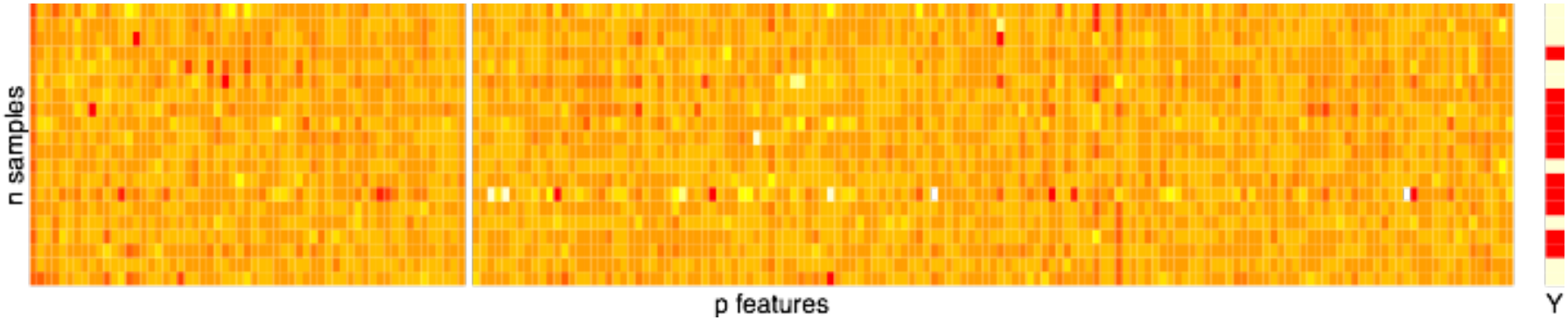




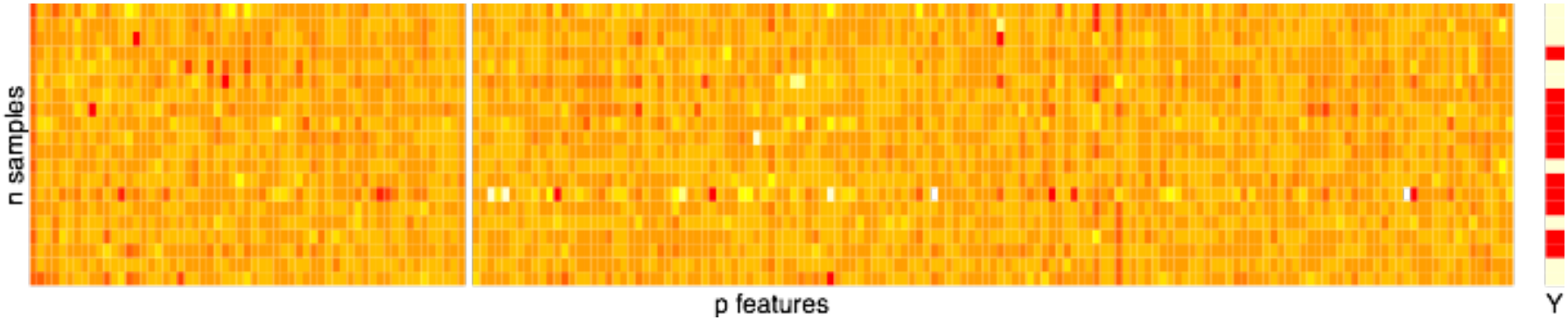




Example: Patient stratification



Problem : n << p



n = 1E2 ~ 1E4
(patients)

p = 1E4 ~ 1E7
(genes, mutations, copy numbers, …)

Learning is hard when n<<p

- Lack of robust biomarkers

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‡

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

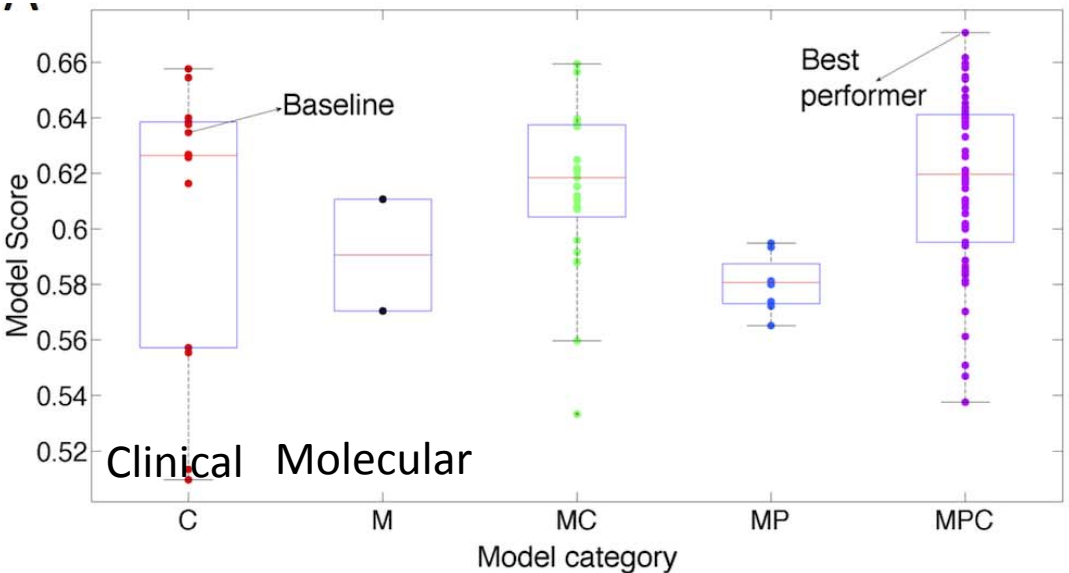
- Poor accuracy

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling

Erhan Bilal[1¶], Janusz Dutkowski[2¶], Justin Guinney[3¶], In Sock Jang[3¶], Benjamin A. Logsdon[3,4¶], Gaurav Pandey[5,6¶], Benjamin A. Sauerwine[3¶], Yishai Shimoni[7,8¶], Hans Kristian Moen Vollan[9,10,11,12,13¶], Brigham H. Mecham[3], Oscar M. Rueda[11,12], Jorg Tost[14], Christina Curtis[15], Mariano J. Alvarez[7,8], Vessela N. Kristensen[9,10,16], Samuel Aparicio[17,18], Anne-Lise Børresen-Dale[9,10], Carlos Caldas[11,12,19,20], Andrea Califano[7,8,21,22,23,24‡], Stephen H. Friend[3‡], Trey Ideker[2‡], Eric E. Schadt[5‡], Gustavo A. Stolovitzky[1‡], Adam A. Margolin[3‡*]



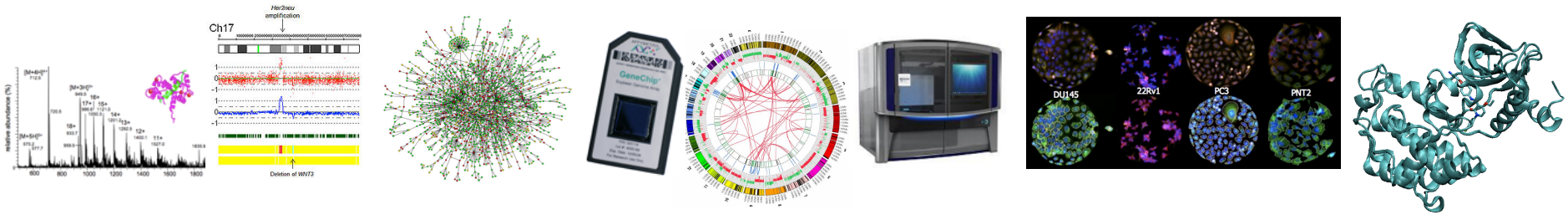
Why?

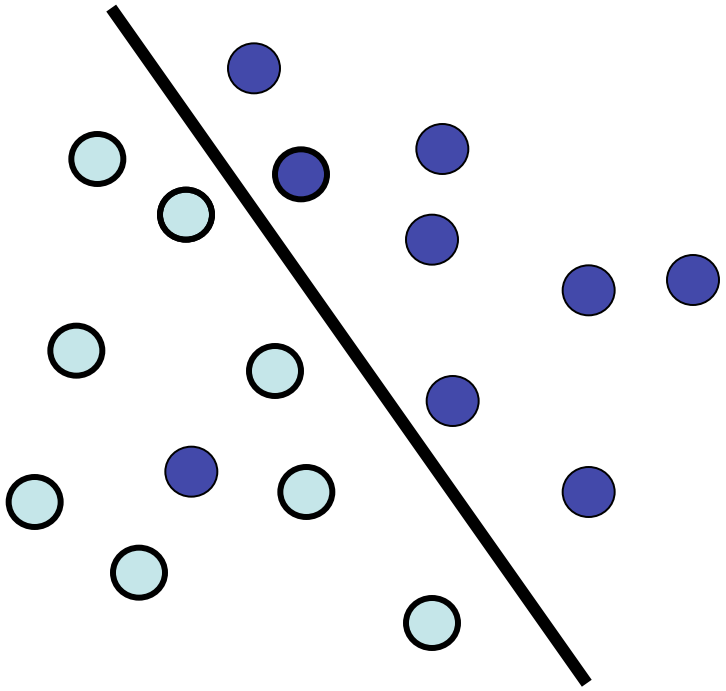
- Wrong data?
- Wrong method?
- Not enough data?
- …?

European Research Council

ERC SMAC (2012-2017)

Statistical Machine Learning for Complex Biological Data





General framework for learning:

$$\min_{f \in \mathcal{F}} R(f) \quad \text{such that} \quad \Omega(f) \leq \gamma$$

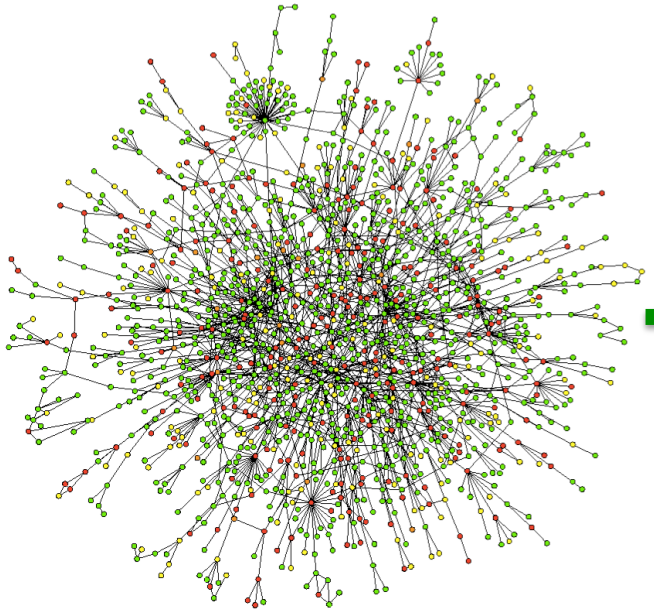
Class of models — *Data fitting term* — *Penalty*

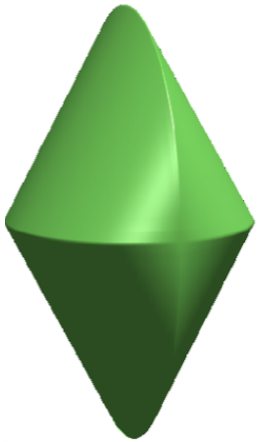
Challenges

- Prior knowledge
- Efficient algorithms
- Interpretability
- *Heterogeneous data integration*
- *Structured, complex data*
- *Few examples*
- *Need for efficient algorithms*

Structured feature selection

- Use a gene network as « prior knowledge »





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

(a convex body in p dimensions)

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

(convex optimization)

- Increases stability and accuracy

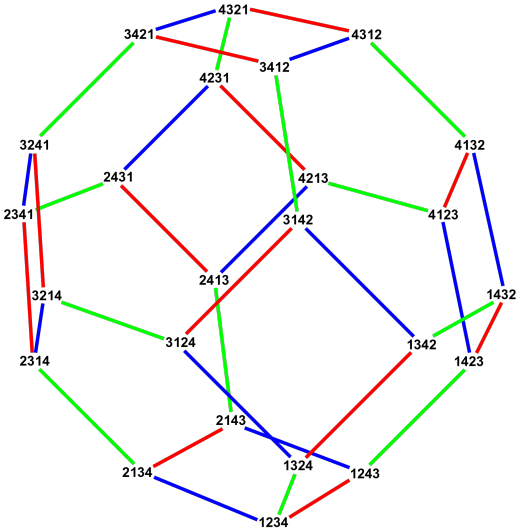
Lasso	Graph Lasso
0.61 %	0.64 %

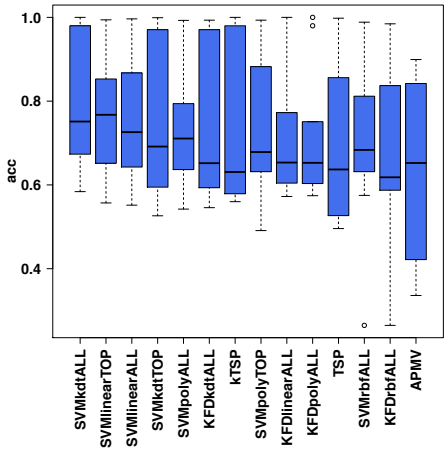
Breast cancer prognosis, accuracy

Change data representation

Replace $x \in \mathbb{R}^p$ by $\Phi(x) \in \{0,1\}^{p(p-1)/2}$:

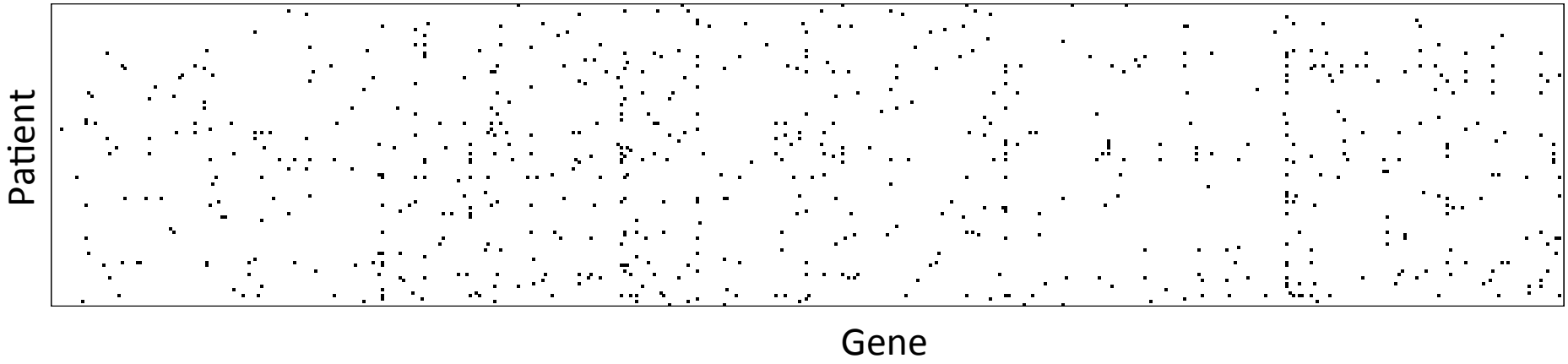
$$\Phi_{i,j}(x) = \begin{cases} 1 & \text{if } x_i \leq x_j\,, \\ 0 & \text{otherwise.} \end{cases}$$



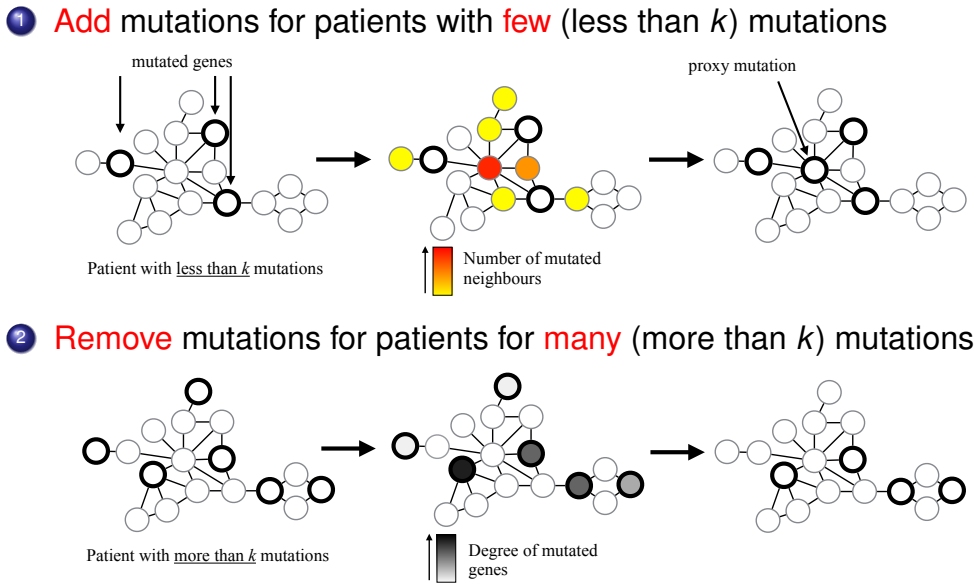


Dataset	No. of features	No. of samples (training/test)	
Breast Cancer 1	23624	44/7 (Non-relapse)	32/12 (Relapse)
Breast Cancer 2	22283	142 (Non-relapse)	56 (Relapse)
Breast Cancer 3	22283	71 (Poor Prognosis)	138 (Good Prognosis)
Colon Tumor	2000	40 (Tumor)	22 (Normal)
Lung Cancer 1	7129	24 (Poor Prognosis)	62 (Good Prognosis)
Lung Cancer 2	12533	16/134 (ADCA)	16/15 (MPM)
Medulloblastoma	7129	39 (Failure)	21 (Survivor)
Ovarian Cancer	15154	162 (Cancer)	91 (Normal)
Prostate Cancer 1	12600	50/9 (Normal)	52/25 (Tumor)
Prostate Cancer 2	12600	13 (Non-relapse)	8 (Relapse)

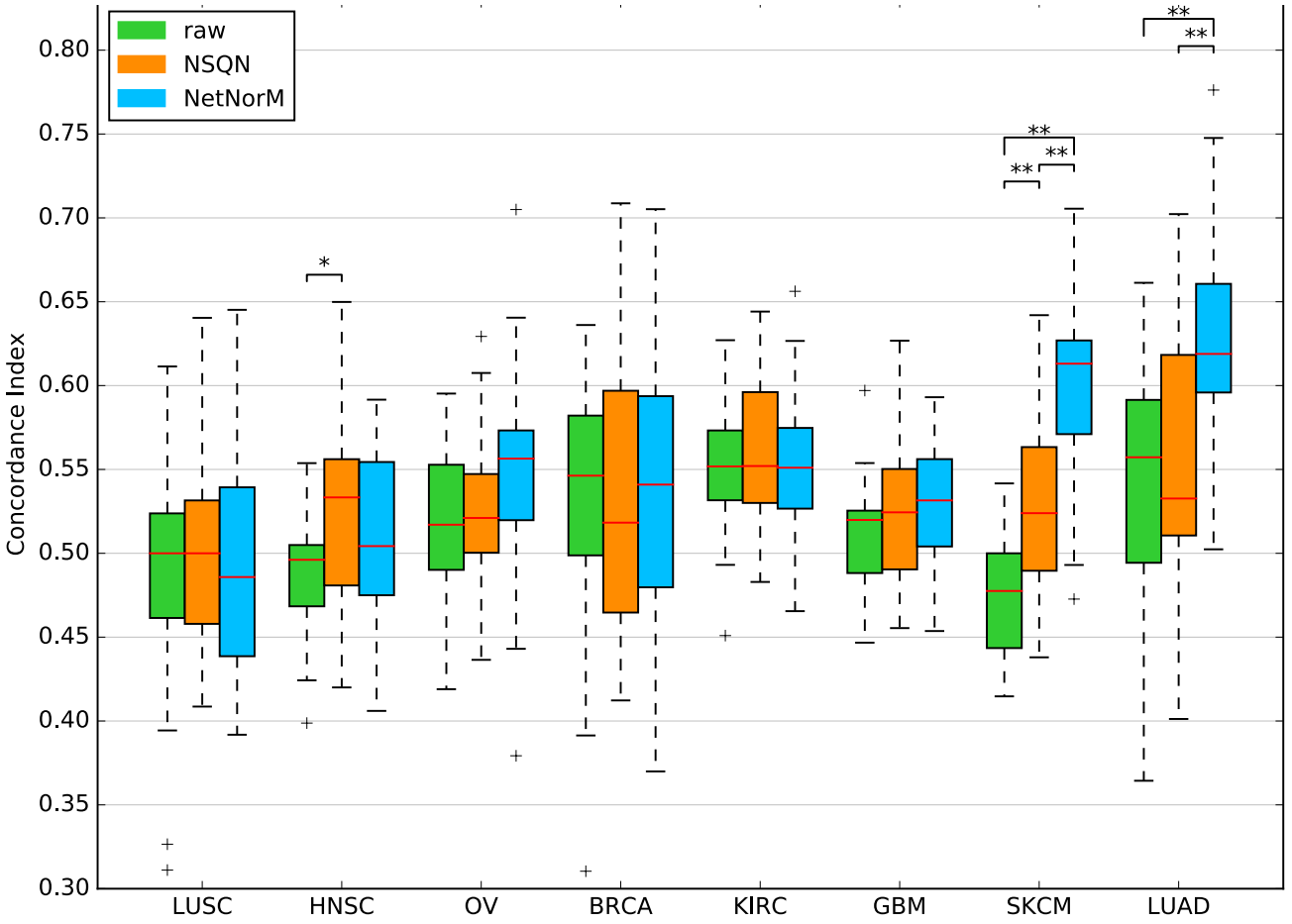
Survival prediction from Whole-exome somatic mutations



NetNorm

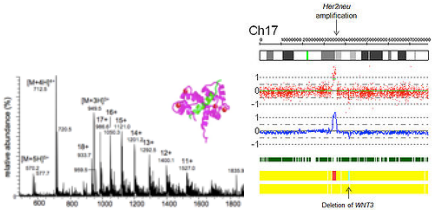
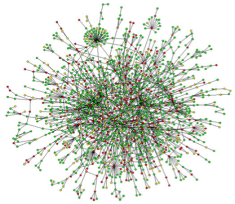
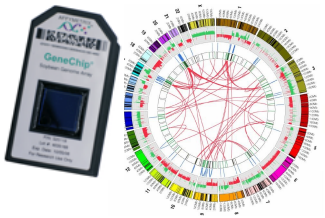
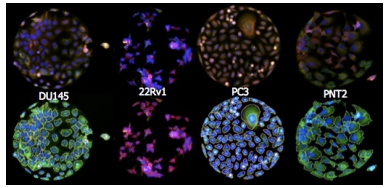


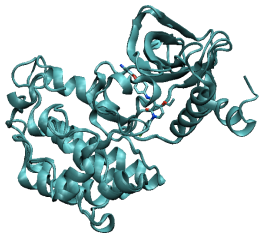
Survival prediction from Whole-exome somatic mutations



Use Pathway Commons as gene network.
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)

Challenges

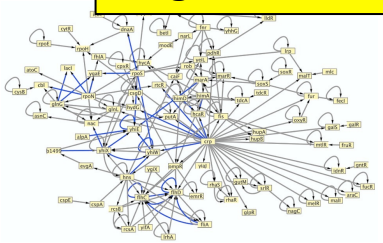


Causality VS Association:
Predictive markers, gene regulation

New data:
Single-cell, Hi-C, medical records, images…

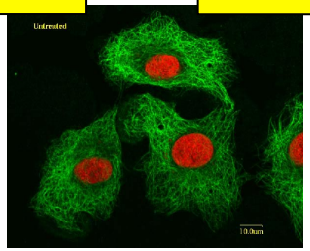
Models / Algorithms:
large-scale ML algorithms

Complex prediction tasks:
Combinatorial therapeutic strategies



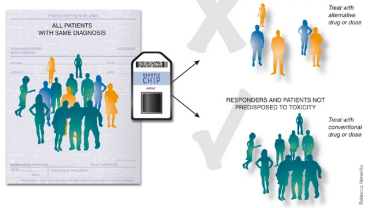
Molecular level

Gene regulation
Epigenetics
Structure/Function prediction



Cellular level

High-content screening
Chemo/Toxicogenomics
Tumour heterogeneity



Precision medicine

Patient stratification
Prognostic / Predictive
Side effect prediction

Thanks!