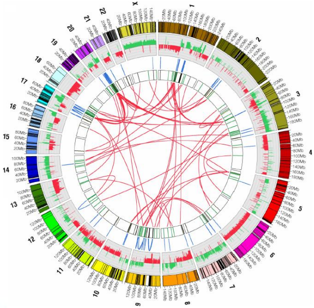
Machine learning for precision medicine

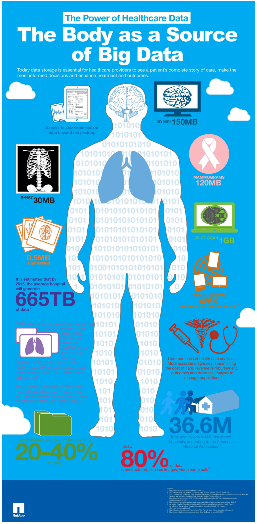
Jean-Philippe Vert
jean-philippe.vert@m4x.org

Health data

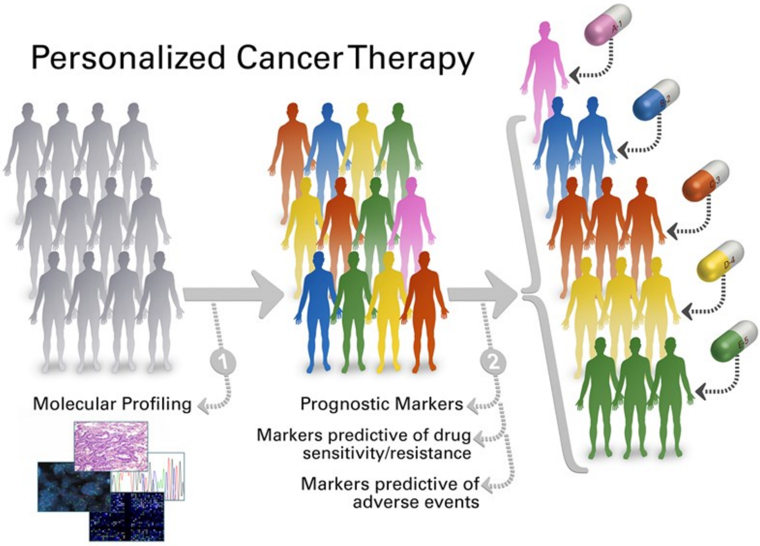

treatment
genes
clinical
drug
patients
genetic
response
gene
development
therapy
studies
study
polymorphisms
expression
human
analysis
cancer
drugs
disease
data




97

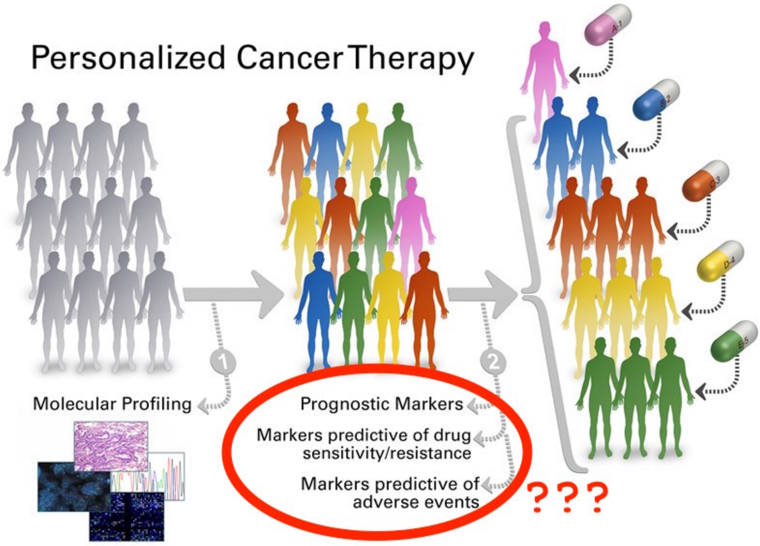

The Power of Healthcare Data
The Body as a Source of Big Data
Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.
Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 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
There are currently 425K telehealth providers in the U.S.
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)
The Medicare and Medicaid Electronic Health Record Incentive Program now includes a measure for recording imaging results via certified EHR technology.
Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations
36.6M
Total admissions in U.S. registered hospitals, according to the American Hospital Association
Medical image archives are increasing by
20-40%
annually
Today,
80%
of data
is unstructured, such as images, video and email.
NetApp

The future of medicine



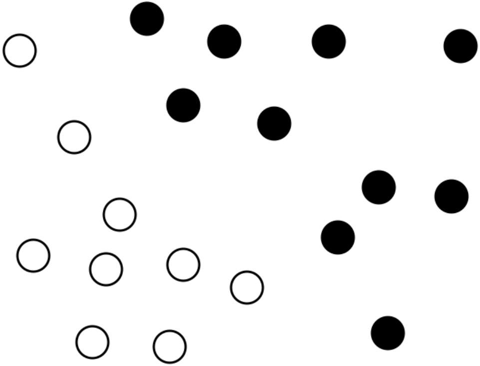
https://pct.mdanderson.org

The future of medicine

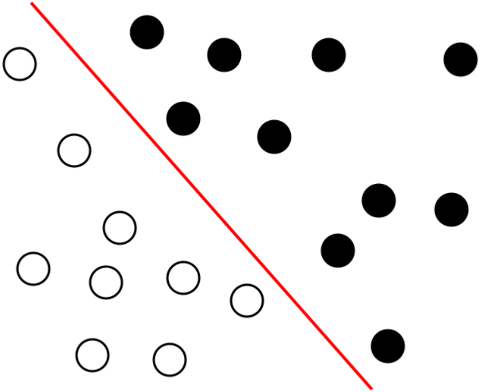


https://pct.mdanderson.org

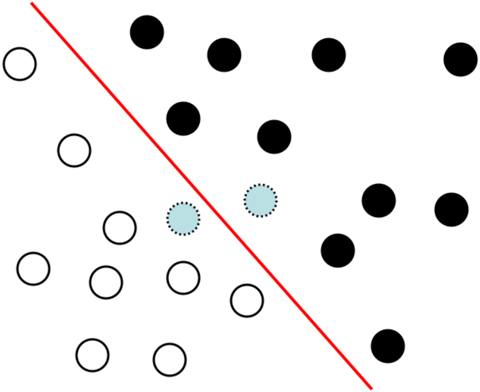
Machine learning



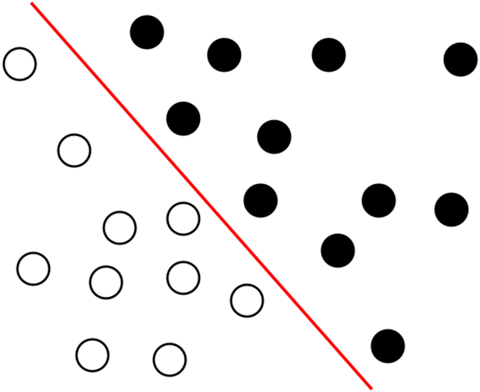
Machine learning



Machine learning



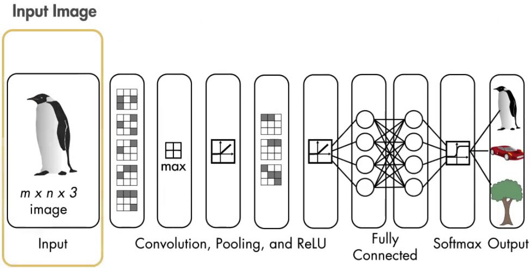
Machine learning



Modern ML works well!

Ingredients:

1. Collect big, labeled data (eg, 10M images)
2. Use a model well adapted to the data (eg, CNN)

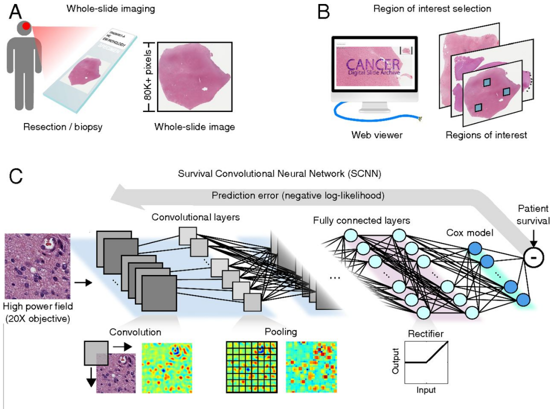


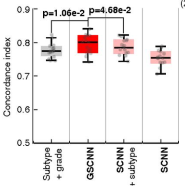
(from `https://www.youtube.com/watch?v=gjK70r0Rqzs`)

3. Large computational power + know-how ("alchemy"?)

Many applications: object/face recognition in images, machine translation, speech recognition, go, self-driving cars, trading, recommender systems, chemistry, material science...

Promising applications in health: images, texts, ..?



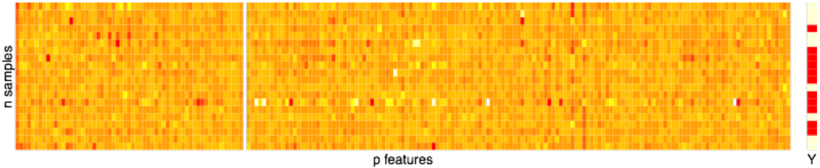


(Mobadersany et al., 2018)

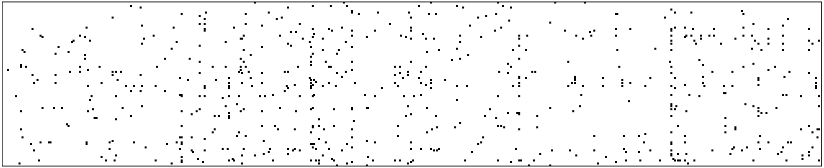
Also: high-content screening, digital pathology, radiomics, skin diagnosis, EHR, ...

More challenging data

- Gene expression



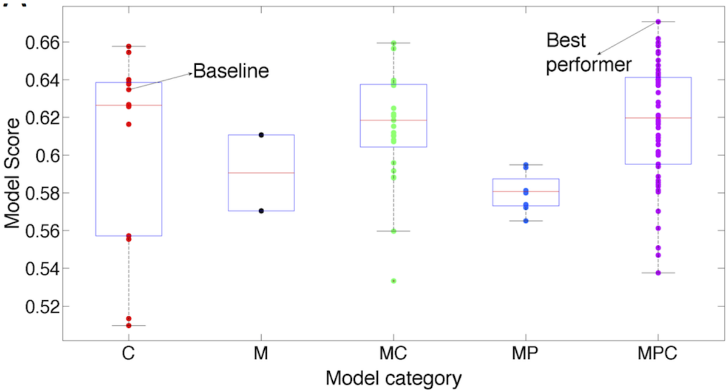
- Somatic mutations



- $n = 10^2 \sim 10^4$ (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

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,

70 genes (Nature, 2002)

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

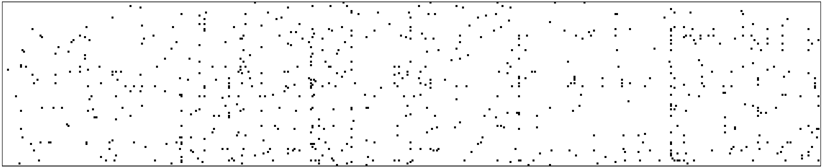
Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

76 genes (Lancet, 2005)

3 genes in common

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

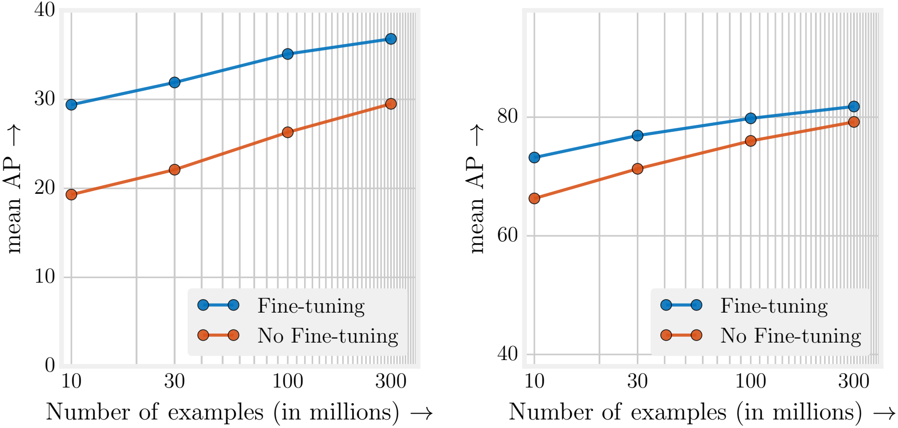
What to do?



- Get more data
  - with labels
  - sharing data (or models) is crucial
  - of good quality
- Improve the models
  - include prior knowledge (biology, structure of noise, invariants...)
  - balance model complexity vs data available

More data helps

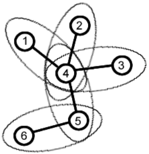
...but performance increases slowly. How much can be afford?

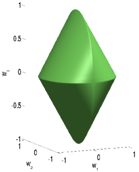


Object detection performance on two benchmarks (COCO minimal, left, and PASCAL VOC 2007, right) as a function of the number of labeled images used to train the model (Sun et al., 2017).

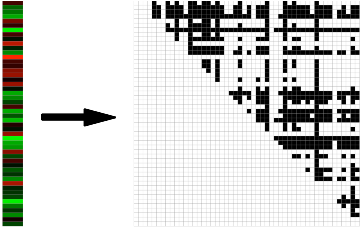
Some research directions

- Regularize and incorporate prior knowledge



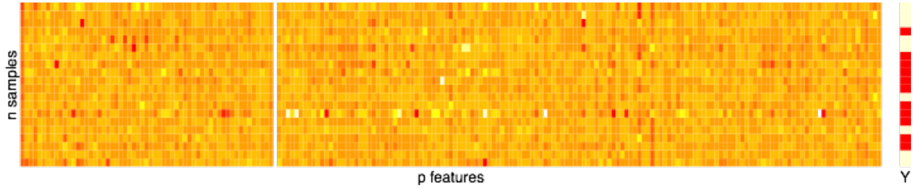


- Find a better representation



Outline

Typical problem

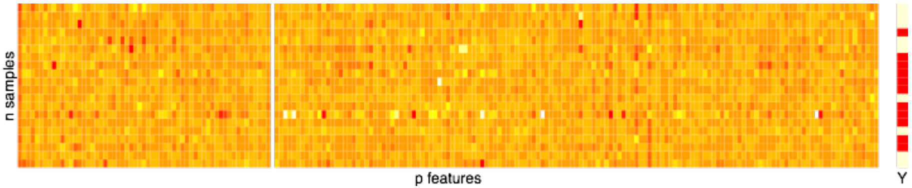


- n samples (patients), p features (genes)
- $X \in \mathbb{R}^{n \times p}$ gene expression profile of each patient
- $Y \in \mathcal{Y}^n$ survival information of each patient
- Fit a linear model for a sample $x \in \mathbb{R}^p$:

$$f(x) = \beta^\top x = \sum_{i=1}^{p} \beta_i x_i$$

- Standard methods (least squares or logistic regression) won't work because $n < p$

Regularized linear models



In high dimension, estimate β by solving

$$\min_{\beta \in \mathbb{R}^p} R(Y, X\beta) + \lambda J(\beta)\,,$$

where

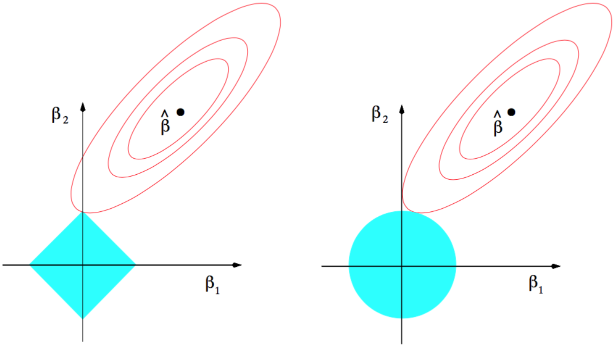
- $R(Y, X\beta)$ is an empirical risk to measures the fit to the training data
- $J(\beta)$ is a penalty to control the complexity of the model
- $\lambda > 0$ is a regularization parameter

Standard regularizations

$$\min_{\beta \in \mathbb{R}^p} R(Y, X\beta) + \lambda J(\beta)$$

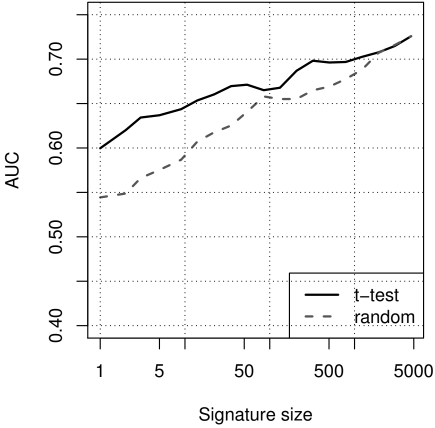
where

- Lasso: $J(\beta) = \|\beta\|_1$ for gene selection.
- Ridge: $J(\beta) = \|\beta\|_2^2$ to address $n \gg m$.
- Elastic net: $J(\beta) = \alpha\|\beta\|_2^2 + (1 - \alpha)\|\beta\|_1$

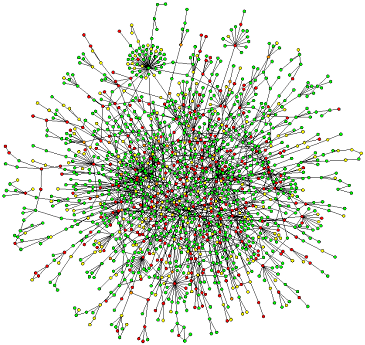


Which regularization is the best?

- Feature selection (lasso, t-tests, ...) is popular, it leads to a limited set of genes that form a molecular signatures
- Ridge is less interpretable but often leads to better performance... e.g., breast cancer prognosis ($n = 286$):

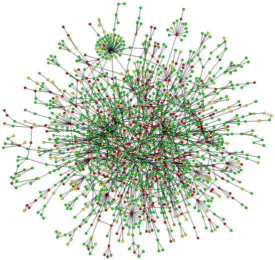


Adding prior knowledge: network-based regularizations



- $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a graph of genes (PPI, metabolic, signaling, regulatory network...)
- Prior knowledge:
    - β should be "smooth" on the graph?
    - Selected genes should be connected?

Examples of network-based regularizations



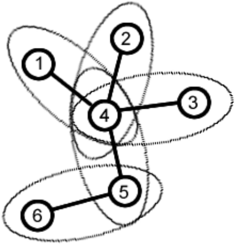
$$J_{\mathcal{G}}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \qquad \text{(Rapaport et al., 2007)}$$

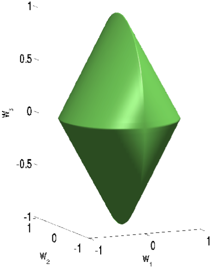
$$J_{\mathcal{G}}(\beta) = a\|\beta\|_1 + (1-a) \sum_{i \sim j} (\beta_i - \beta_j)^2 \qquad \text{(Li and Li, 2008)}$$

$$J_{\mathcal{G}}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j \ \alpha_i^2 + \alpha_j^2 \leq 1} \alpha^\top \beta \qquad \text{(Jacob et al., 2009)}$$

$$J_{\mathcal{G}}(\beta) = a\|\beta\|_1 + (1-a) \sum_{i \sim j} |\beta_i - \beta_j| \qquad \text{(Hoefling, 2010)}$$

Gene selection with the graph lasso

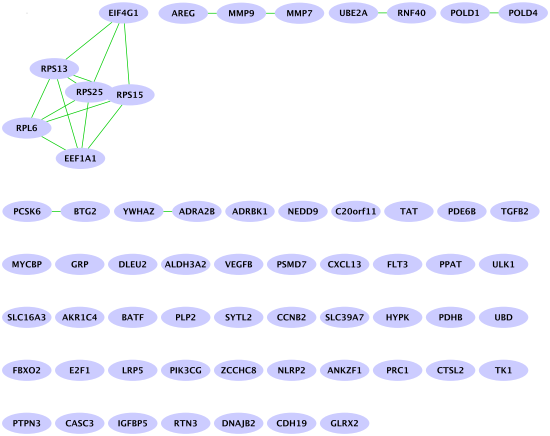




$$J_{\mathcal{G}}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta$$

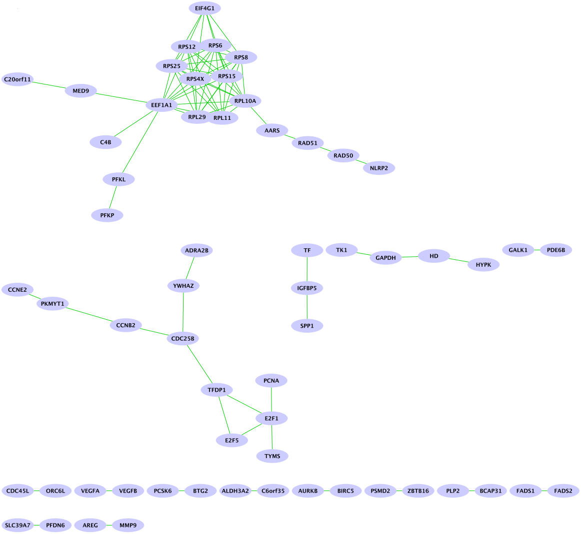
Jacob et al. (2009)

BC prognosis: Lasso signature (accuracy 0.61)



Jacob et al. (2009)

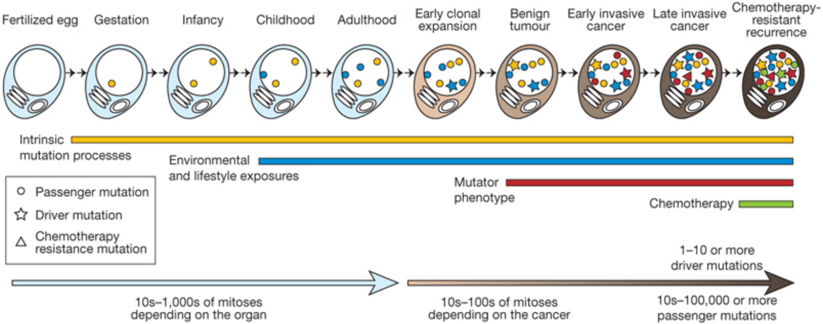
BC prognosis: Graph Lasso signature (accuracy 0.64)

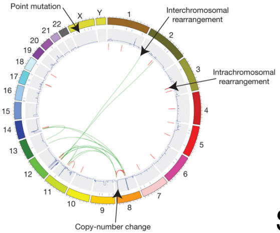


Jacob et al. (2009)

Outline

Somatic mutations in cancer

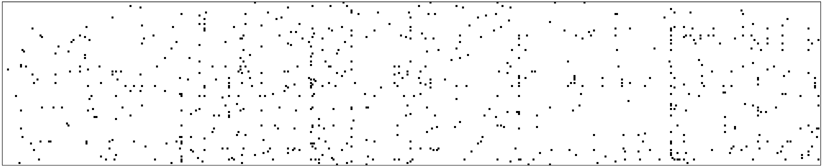




Stratton et al. (2009)

Large-scale efforts to collect somatic mutations

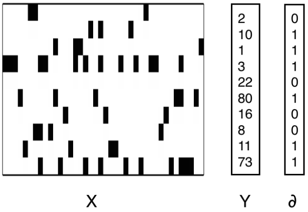
- $3,378$ samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.

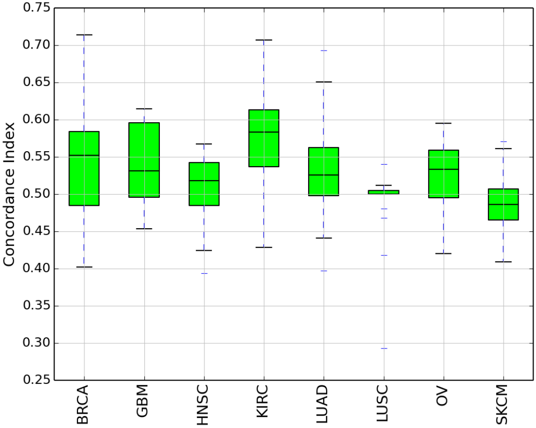


Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC (Lung squamous cell carcinoma)	169	13 590
OV (Ovarian serous cystadenocarcinoma)	363	10 195

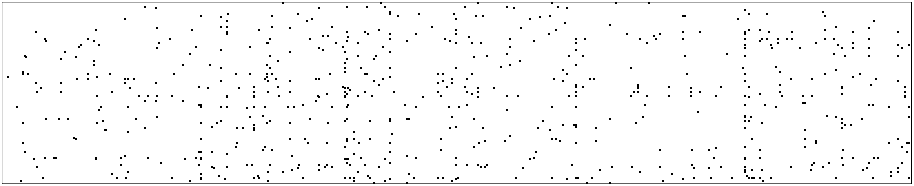
Survival prediction from raw mutation profiles

- Each patient is a binary vector: each gene is mutated (1) or not (2)
- Silent mutations are removed
- Survival model estimated with sparse survival SVM
- Results on 5-fold cross-validation repeated 4 times





Approach: change representation?



Can we replace

$$x \in \{0,1\}^p \quad \text{with } p \text{ very large, very sparse}$$

by a representation with more information shared between samples

$$\Phi(x) \in \mathcal{H}$$

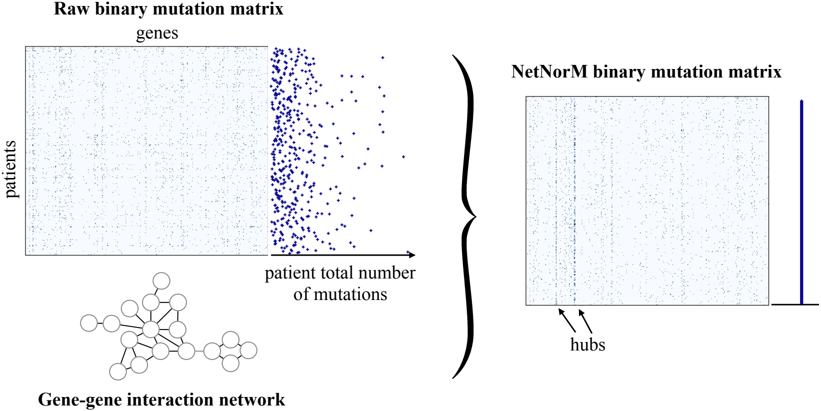
that would allow better supervised and unsupervised classification?

NetNorm Overview (Le Morvan et al., 2017)

Take

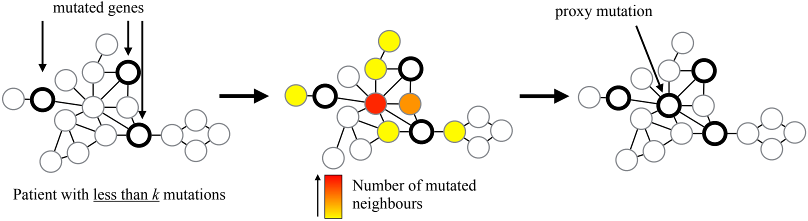
$$\mathcal{H} = \left\{ x \in \{0,1\}^p : \sum_{i=1}^{p} x_i = K \right\}$$

and use a gene network to transform x to $\phi(x) \in \mathcal{H}$ by adding/removing mutations

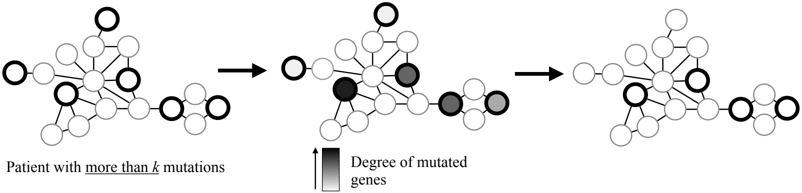


NetNorm detail (k=4)

1. Add mutations for patients with few (less than K) mutations



2. Remove mutations for patients for many (more than K) mutations



In practice, K is a free parameter optimized on the training set, typically a few 100's.

Related work (Hofree et al., 2013)

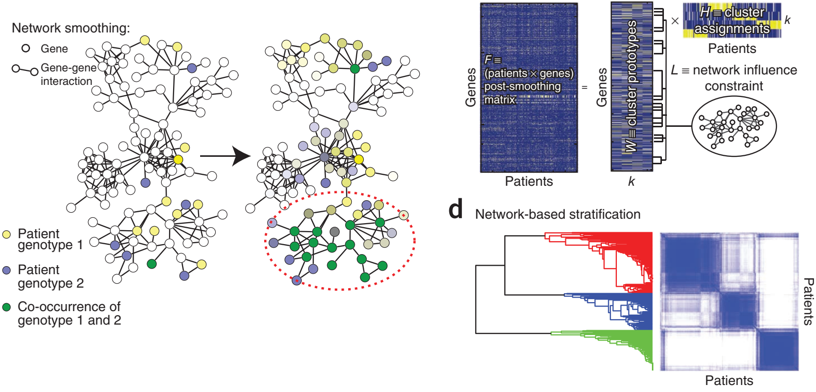
Network-based stratification of tumor mutations

Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

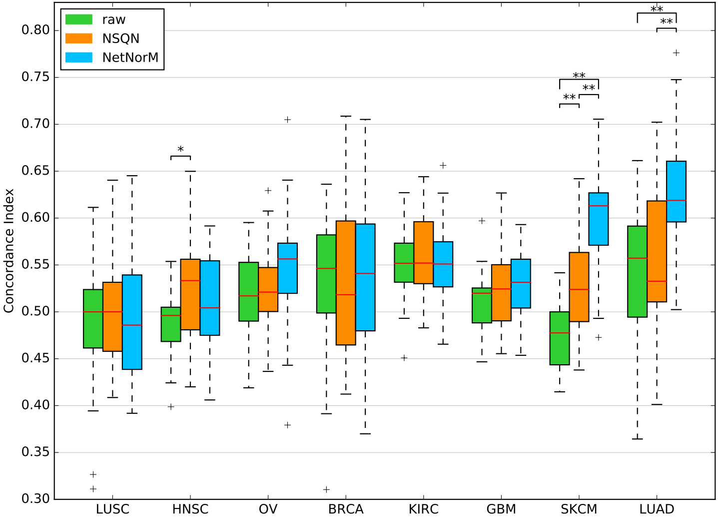
[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [3]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu).

RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651

1108 | VOL.10 NO.11 | NOVEMBER 2013 | NATURE METHODS

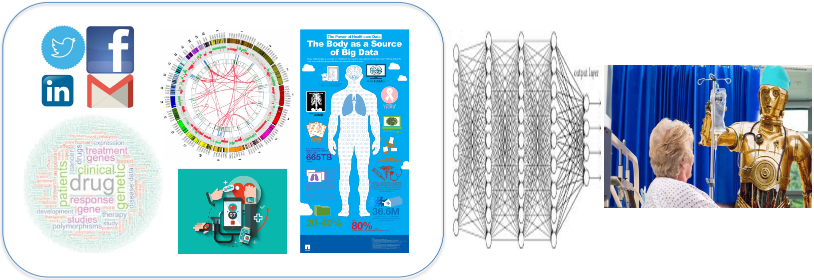


Results



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

Conclusion



- Lots of data, increasing role of ML (particularly with images, texts)
- Omics data is more challenging
- Getting more data is important, but unlikely to allow ML-based methods to reach their best
- Active research
  - allowing data sharing (federated learning, differential privacy, ...)
  - new representations $x \to \Phi(x)$
  - new learning techniques (structured sparsity, regularization, ...)
  - new experimental design strategies (contextual bandit, ...)

References

H. Hoefling. A path algorithm for the Fused Lasso Signal Approximator. *J. Comput. Graph. Stat.*, 19(4):984–1006, 2010. doi: 10.1198/jcgs.2010.09208. URL http://dx.doi.org/10.1198/jcgs.2010.09208.

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. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553431. URL http://dx.doi.org/10.1145/1553374.1553431.

M. Le Morvan, A. Zinovyev, and J.-P. Vert. NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comp. Bio.*, 13(6):e1005573, 2017. URL http://hal.archives-ouvertes.fr/hal-01341856.

C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24:1175–1182, May 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn081.

P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velézquez Vega, D. J. Brat, and L. A. D. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U.S.A.*, 115: E2970–E2979, Mar. 2018. ISSN 1091-6490. doi: 10.1073/pnas.1717139115.

References (cont.)

F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007. doi: 10.1186/1471-2105-8-35. URL http://dx.doi.org/10.1186/1471-2105-8-35.

M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239): 719–724, Apr 2009. doi: 10.1038/nature07943. URL http://dx.doi.org/10.1038/nature07943.

C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. doi: 10.1109/ICCV.2017.97.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1): 267–288, 1996. URL http://www.jstor.org/stable/2346178.

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.