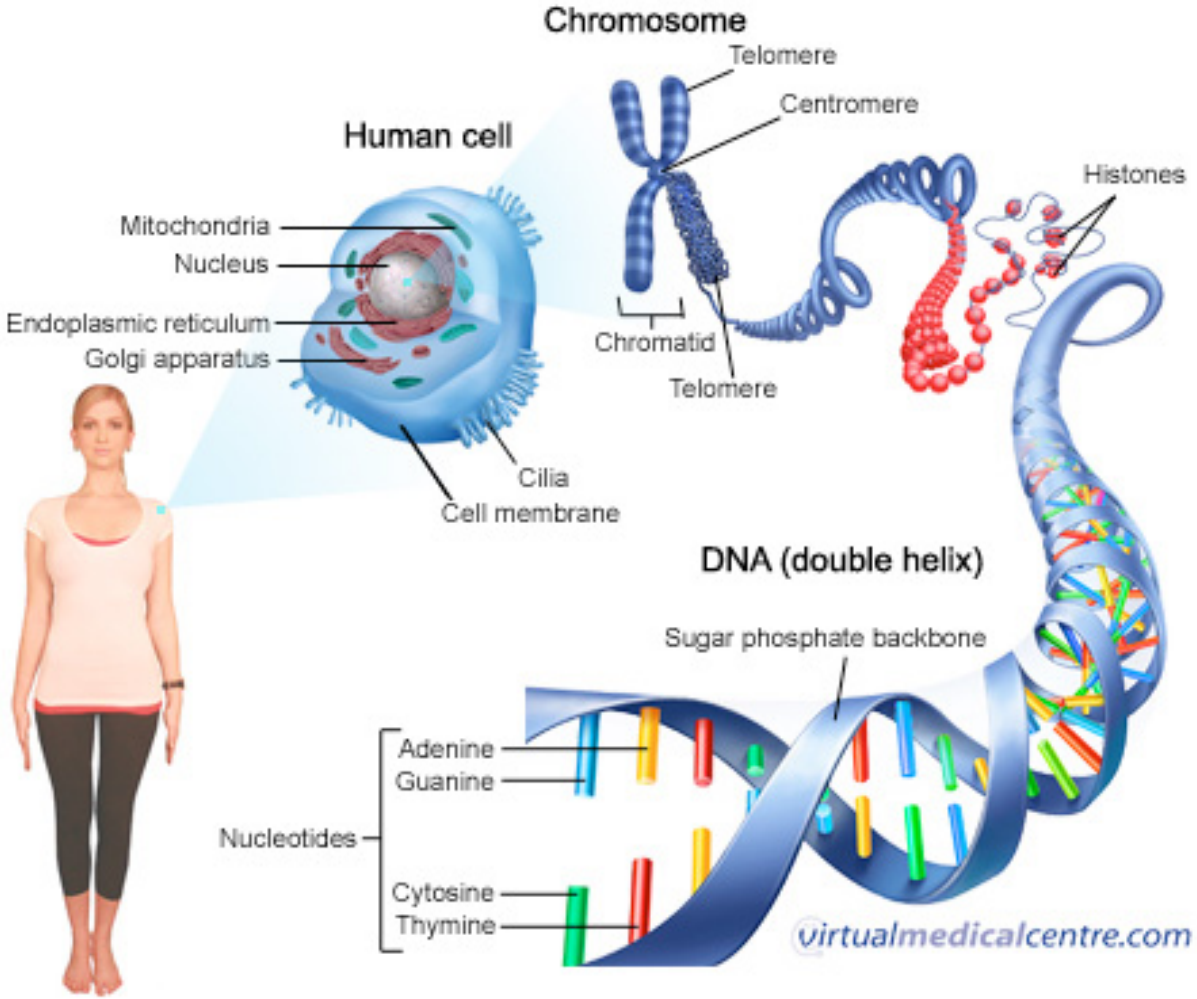
Fighting cancer with chinese lanterns

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

Séminaire "Les Mathémathiques", ENS Paris, November 2, 2016

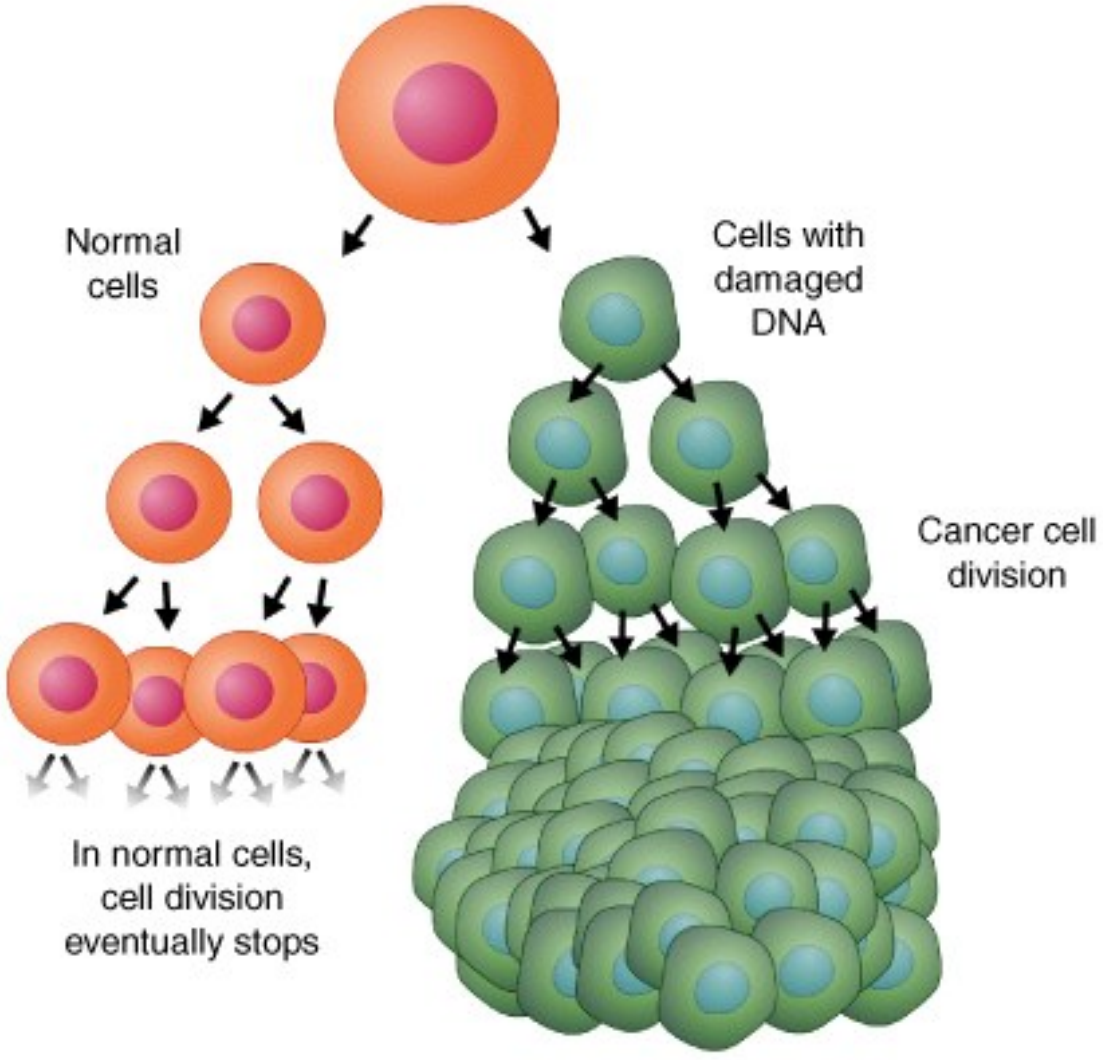
Biology in numbers



1 body = 10^{14} human cells (and 100x more non-human cells)

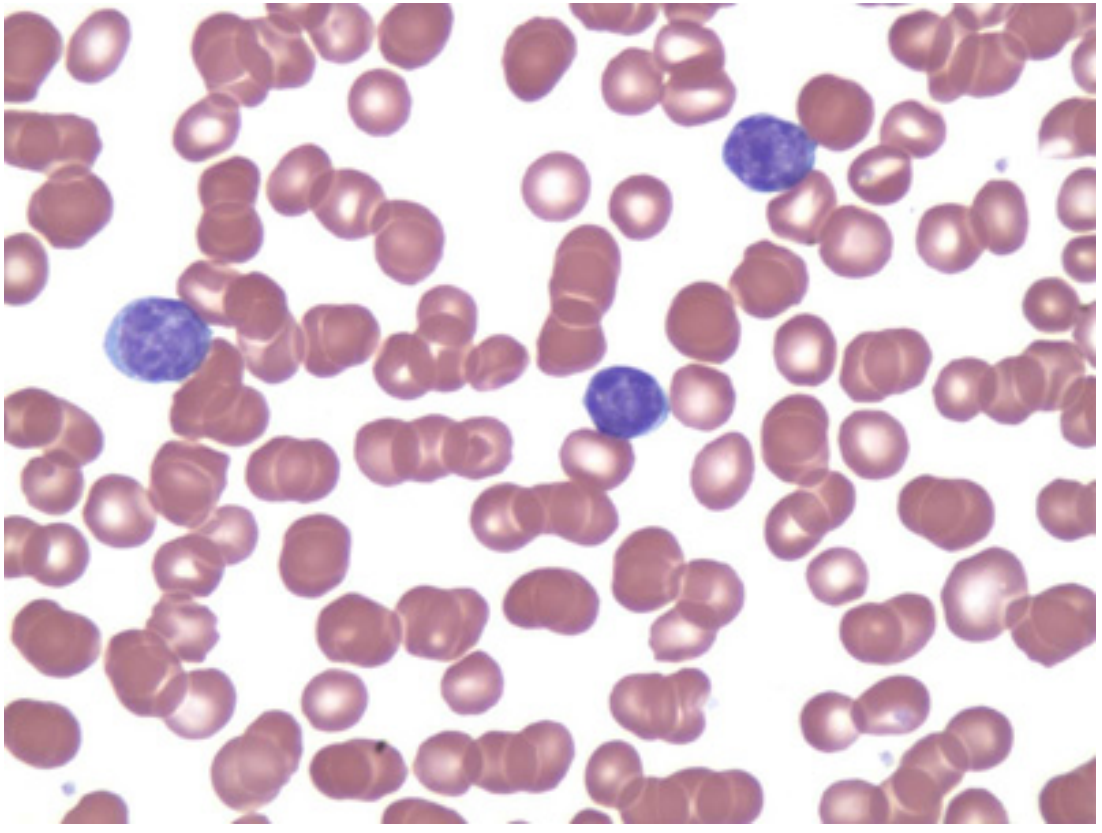
1 cell = 6×10^9 ACGT coding for $20,000+$ genes

Cancer

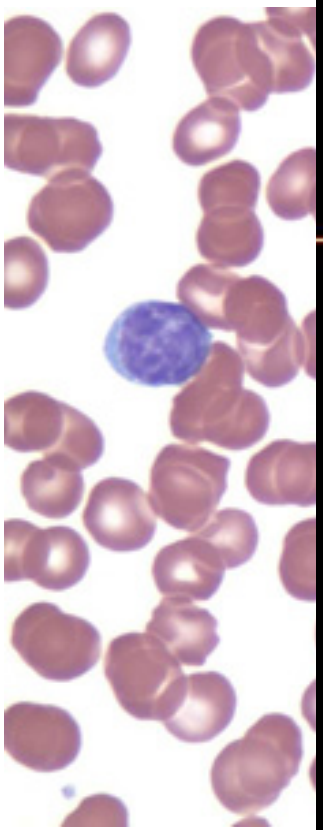


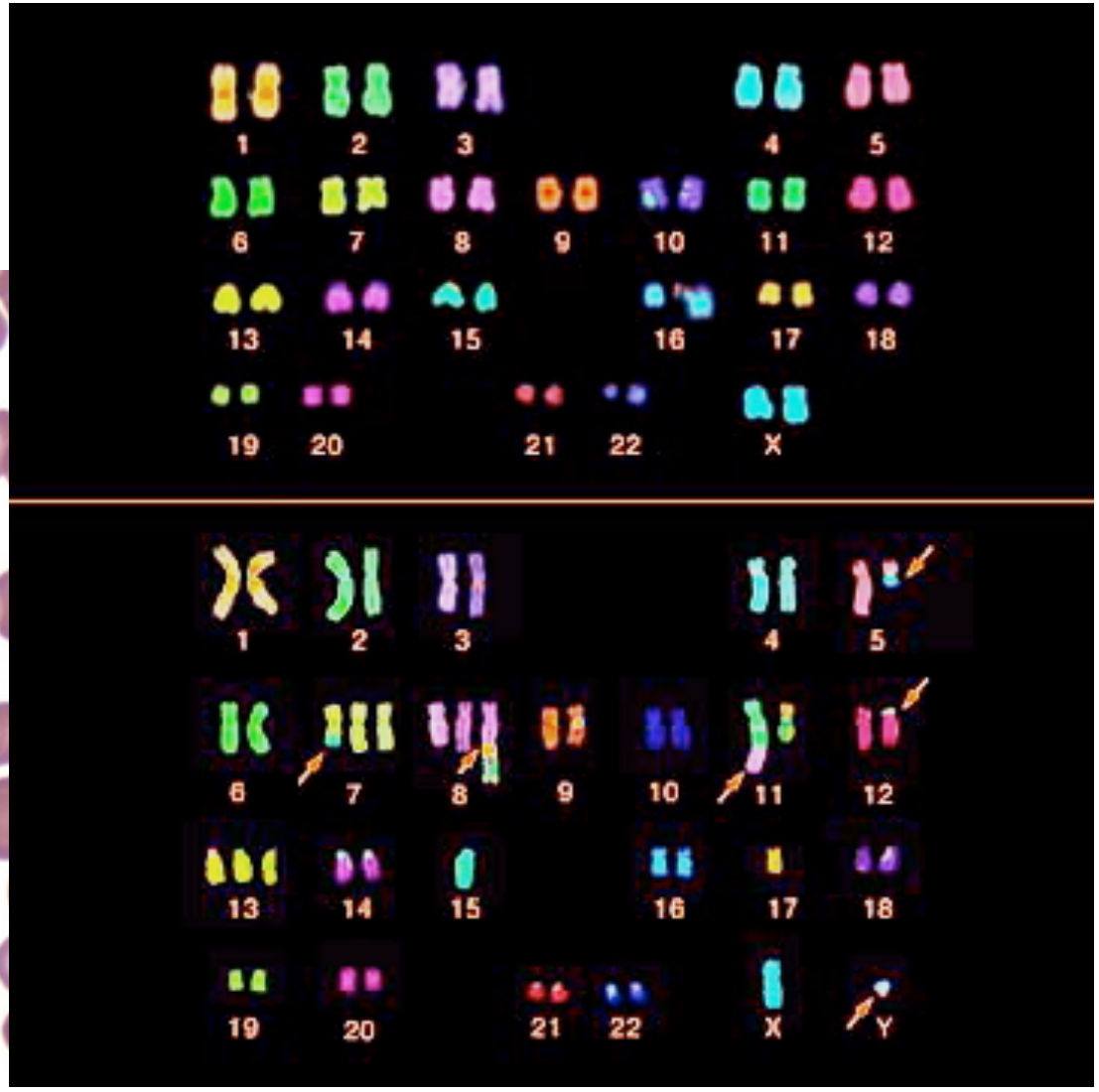
http://rise.duke.edu/seek/pages/page.html?0205

A cancer cell (1900)

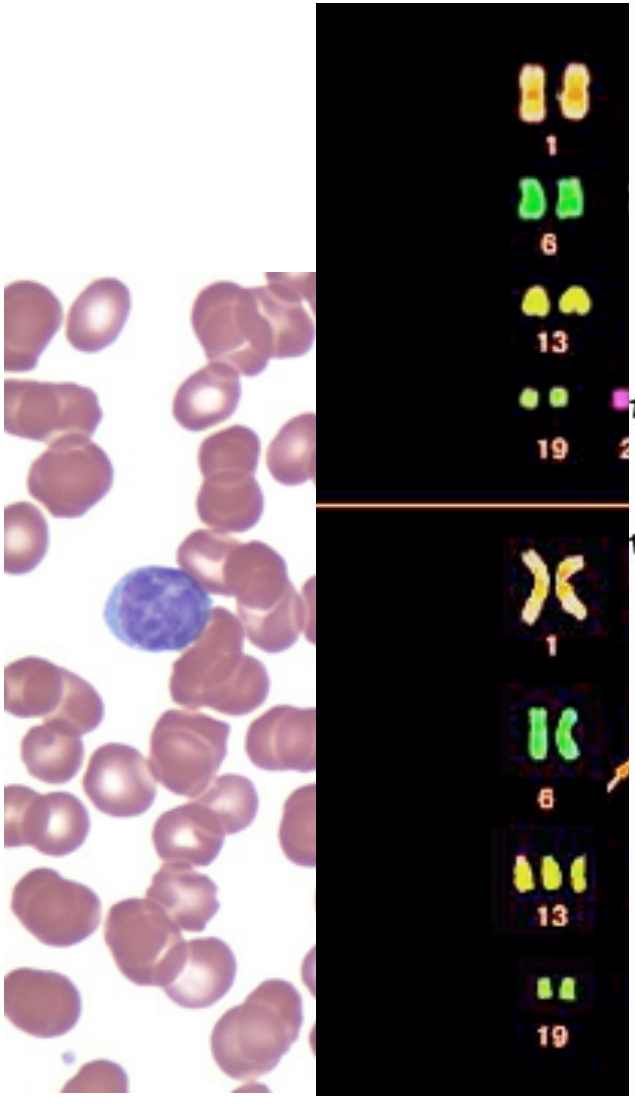


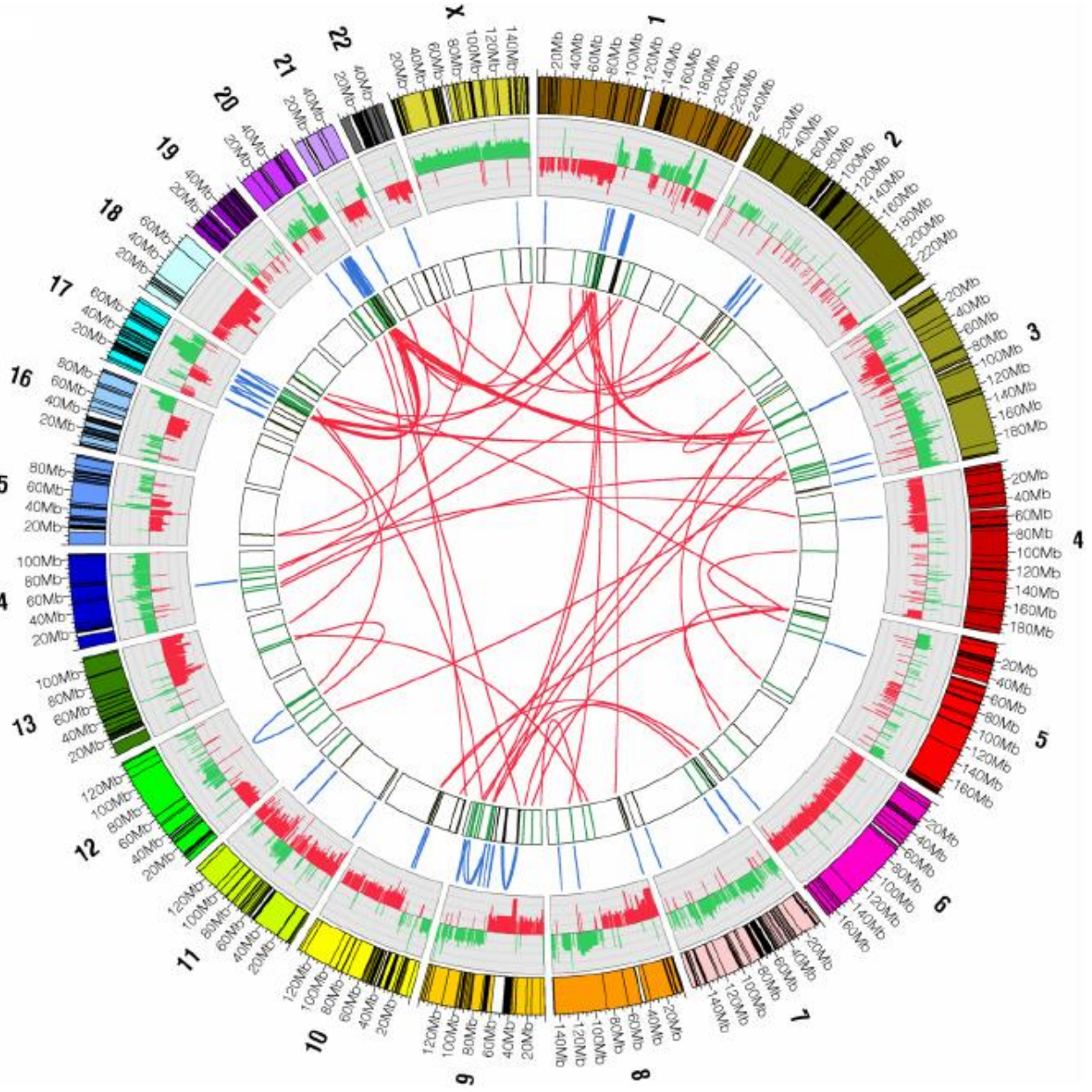
A cancer cell (1960)



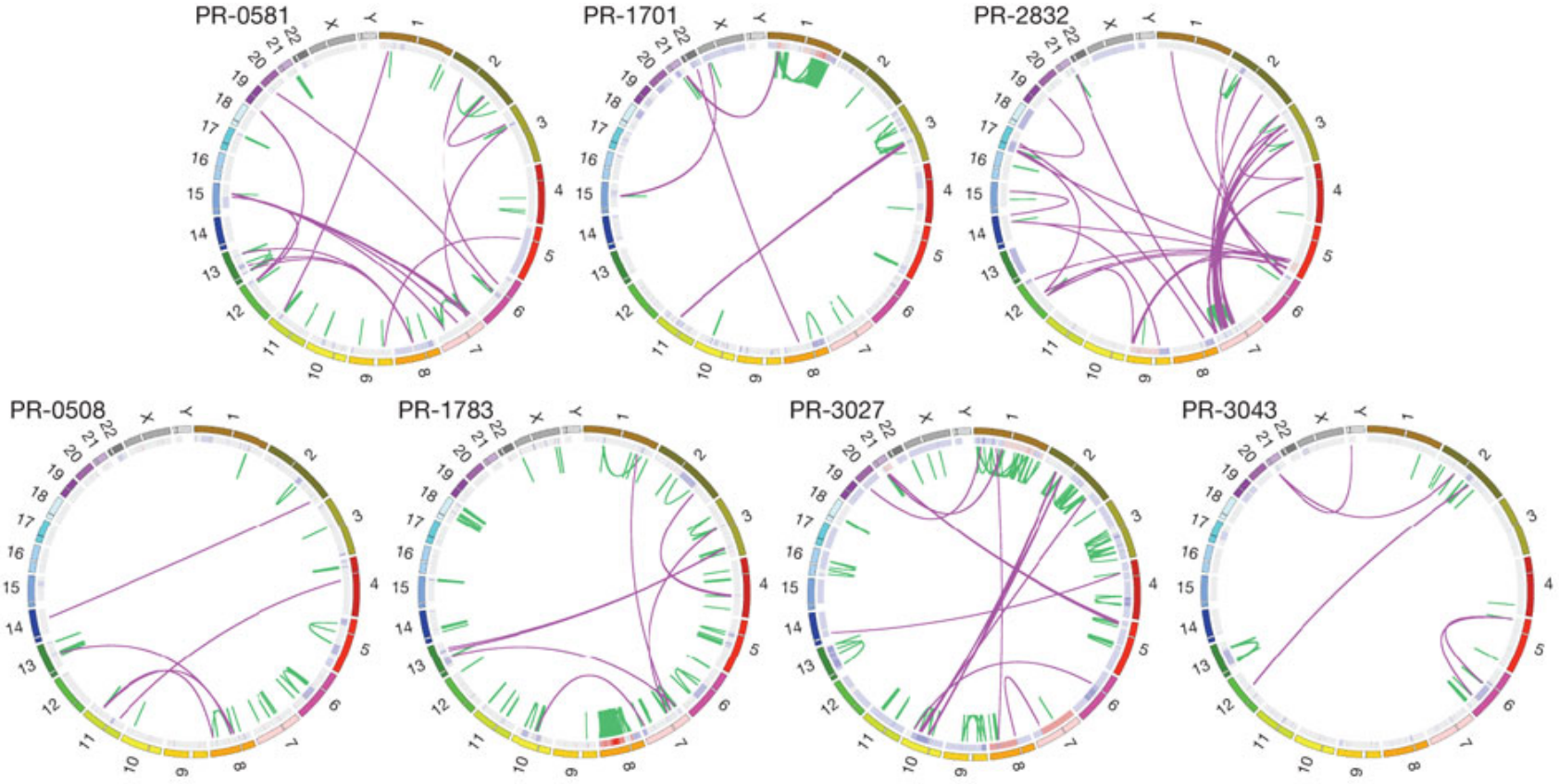


A cancer cell (2010)





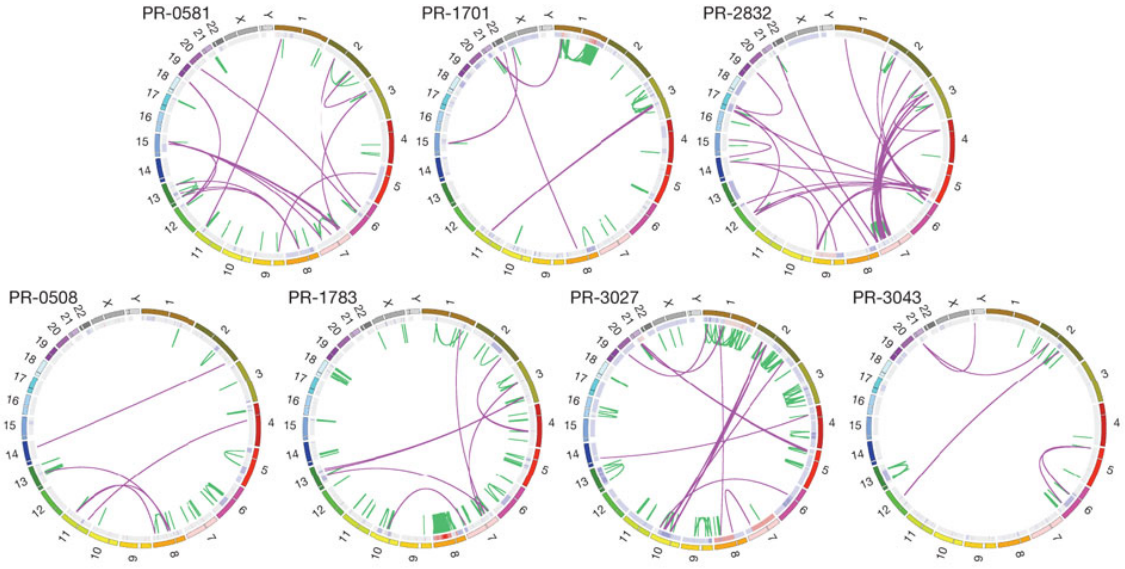
All cancers are different



All happy families are alike; each unhappy family is unhappy in its own way.

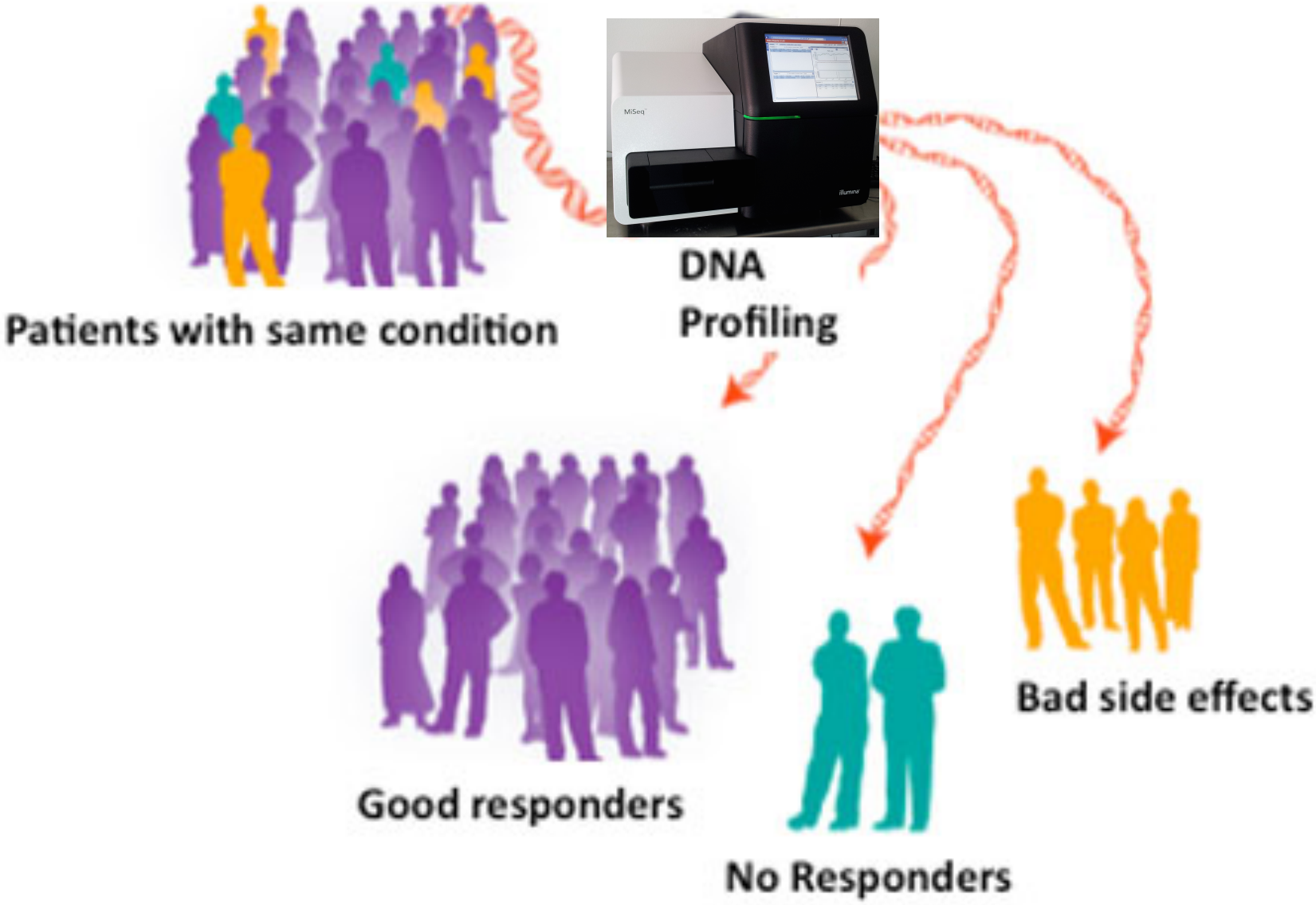
- Leon Tolstoy, Anna Karenina.

Opportunities



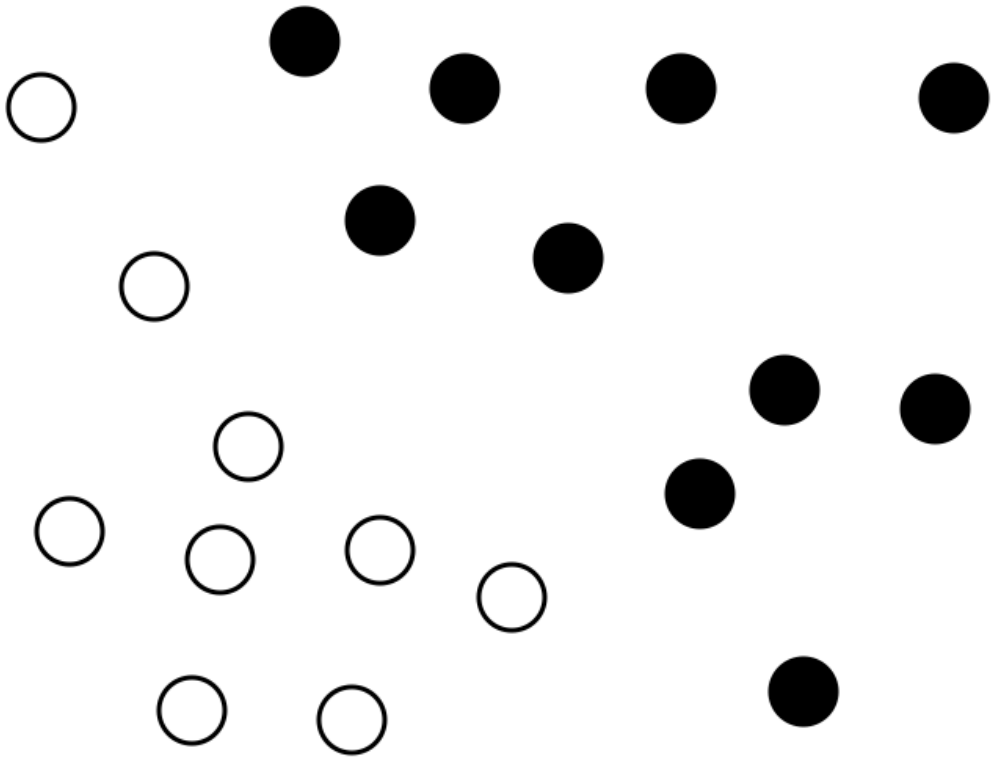
- What is your risk of developing a cancer? (*prevention*)
- Once detected, what precisely is your cancer? (*diagnosis*)
- After treatment, are you cured? (*prognosis*)
- What is the best way to treat your cancer? (*precision medicine*)

Example: precision medicine



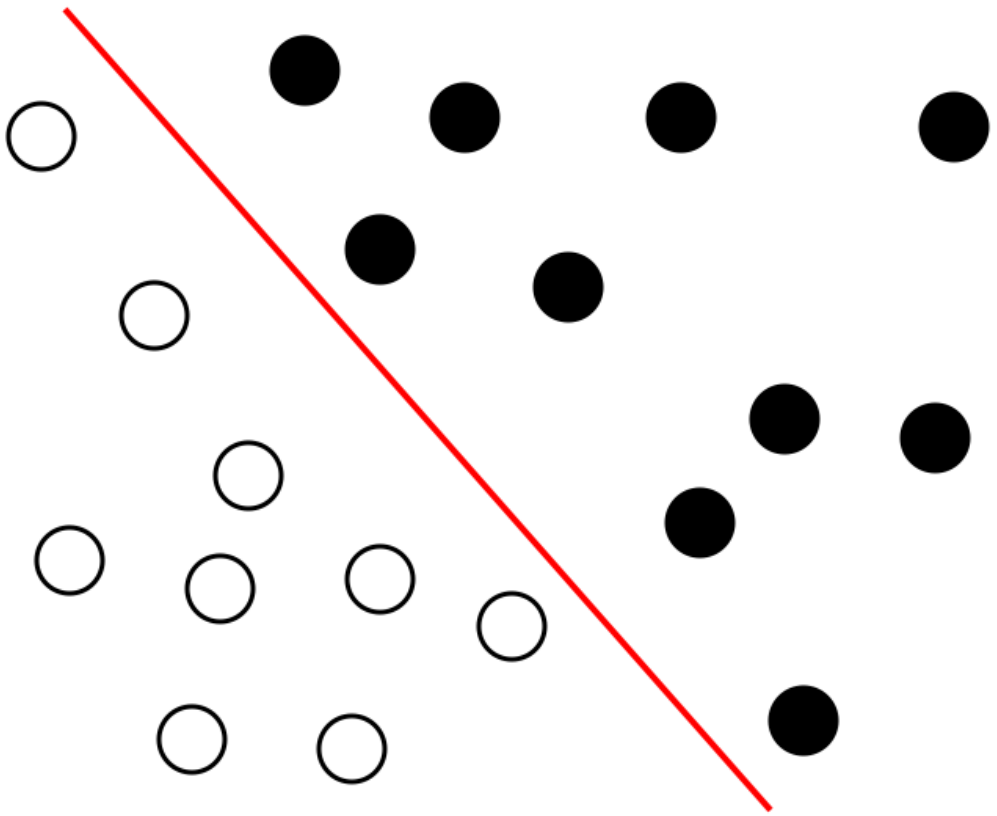
Supervised classification

- Each point is a patient
- Color is the response: good (black) vs bad (white) responder



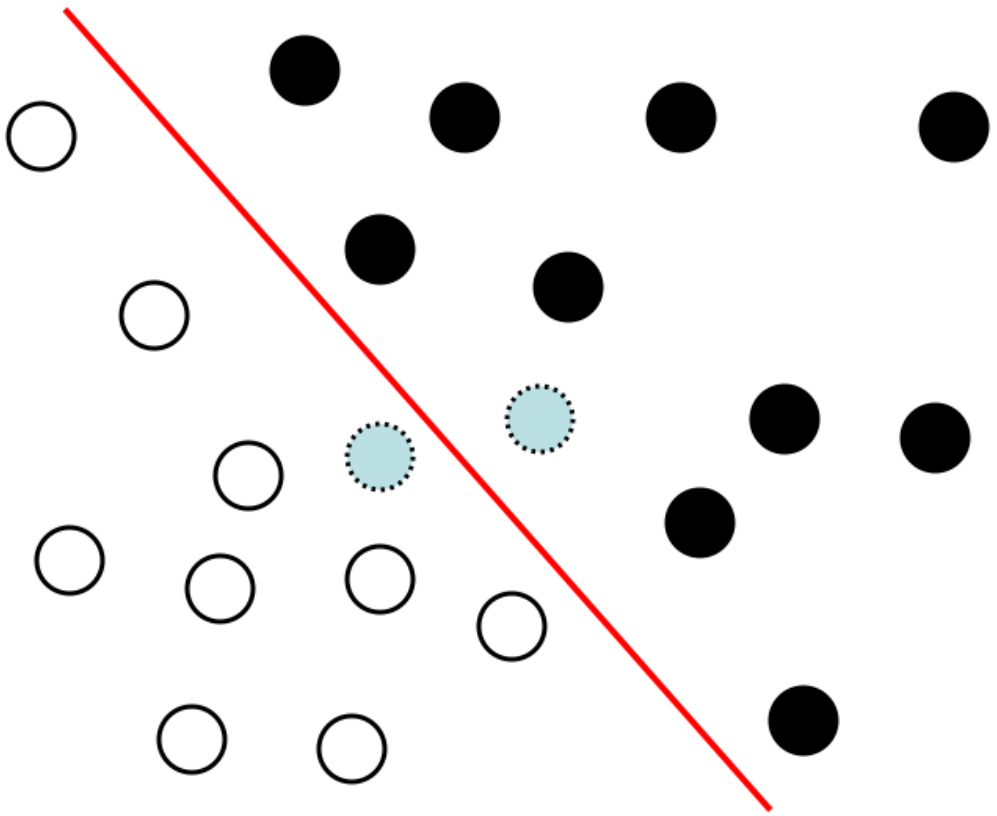
Supervised classification

- Each point is a patient
- Color is the response: good (black) vs bad (white) responder



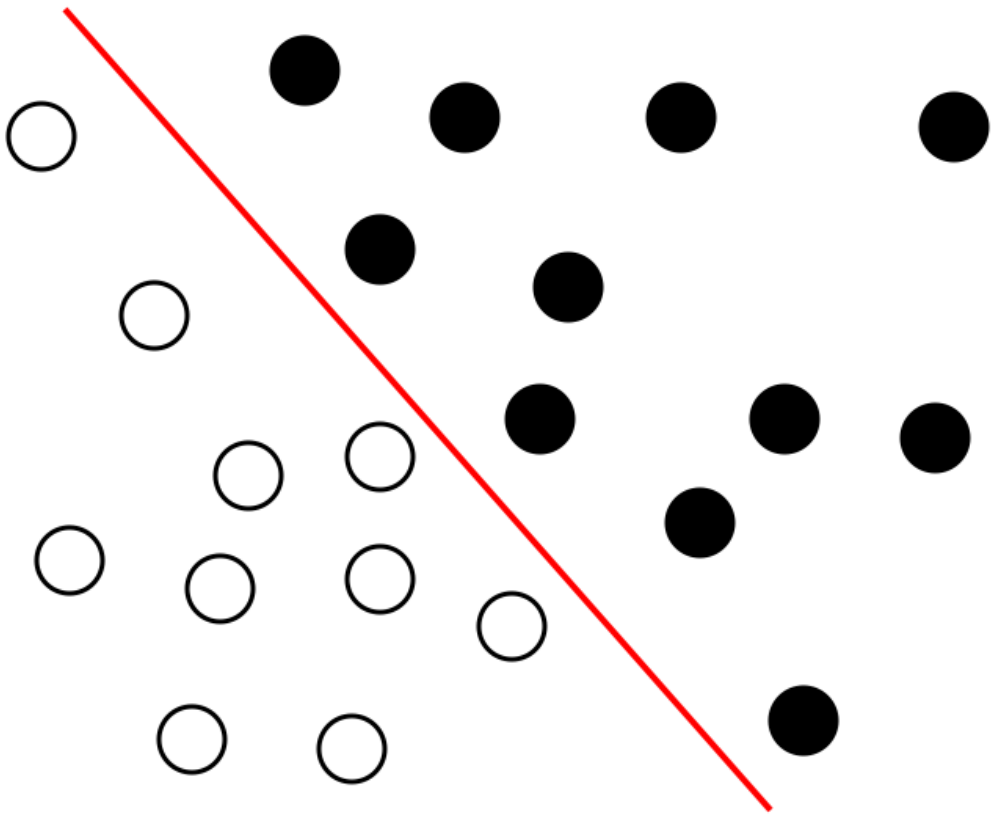
Supervised classification

- Each point is a patient
- Color is the response: good (black) vs bad (white) responder

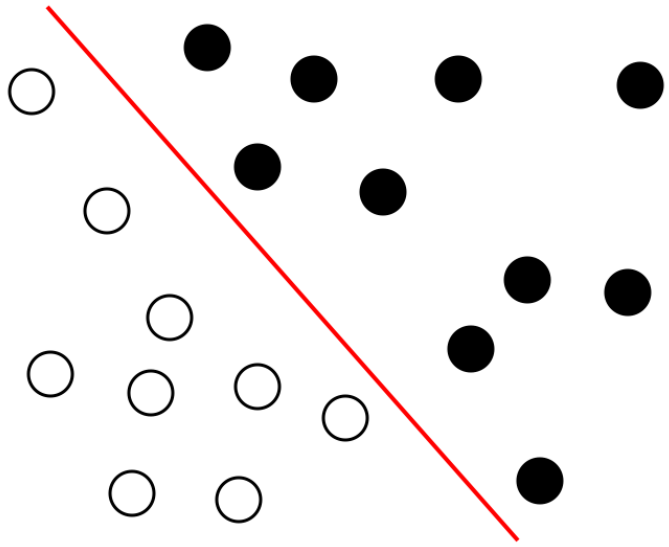


Supervised classification

- Each point is a patient
- Color is the response: good (black) vs bad (white) responder



Example: logistic regression (Berkson, 1944)



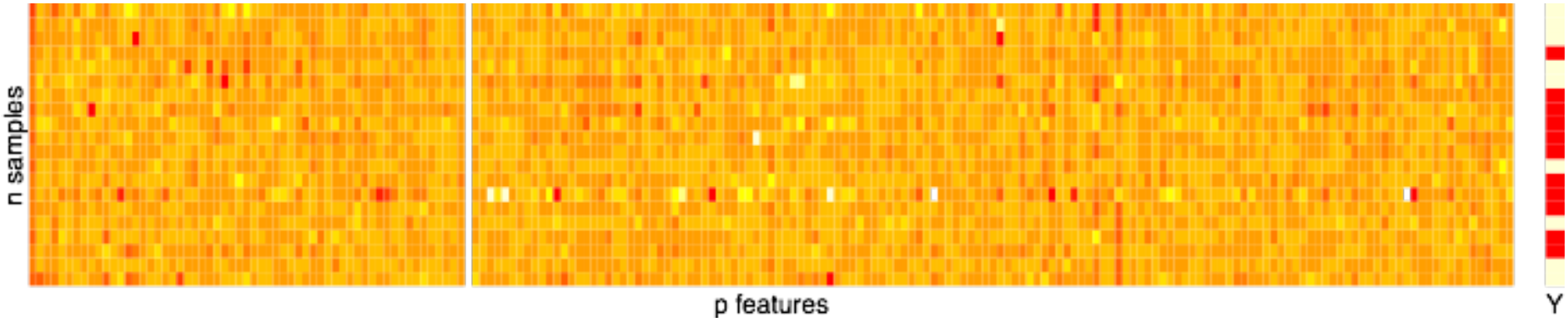
- Given a training set: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where
  - $x_i \in \mathbb{R}^p$ (sample)
  - $y_i \in \{-1, 1\}$ (label)
- Fit a linear model

$$f_\beta(x) = \beta^\top x$$

by solving:

$$\min_{\beta \in \mathbb{R}^p} R(\beta) := \sum_{i=1}^{n} \ln\left(1 + e^{-y_i f_\beta(x_i)}\right)$$

Challenge: $n << p$

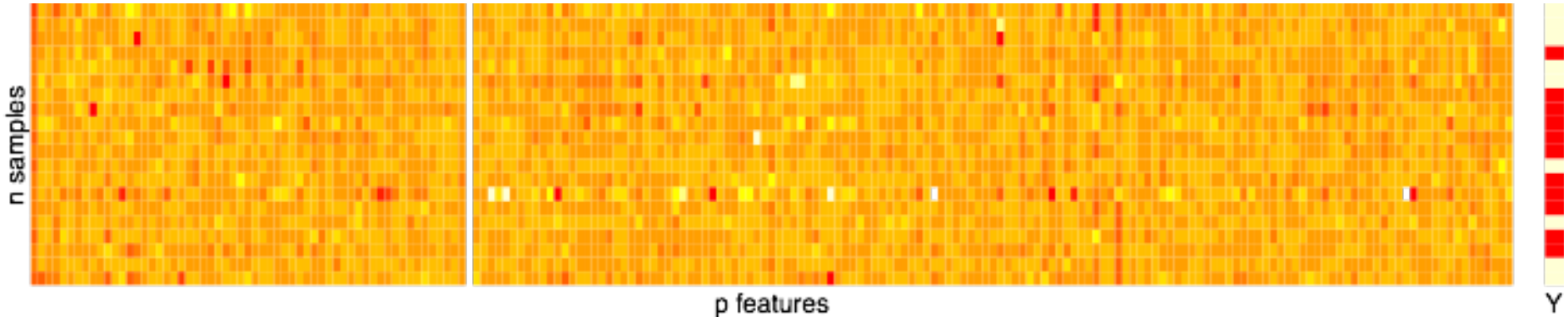


- $n = 10^2 \sim 10^4$ (patients)
- $p = 10^4 \sim 10^7$ (genes, mutations, copy number, ...)

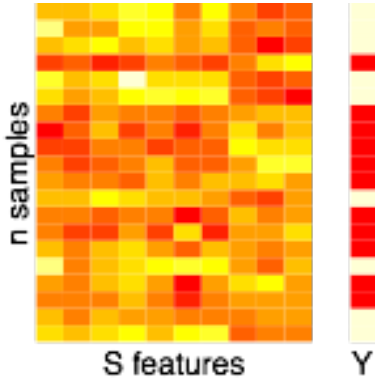
Consequences:

- Problem ill-posed
- Overfitting
- Prediction accuracy drops
- Features selection unstable

Feature selection



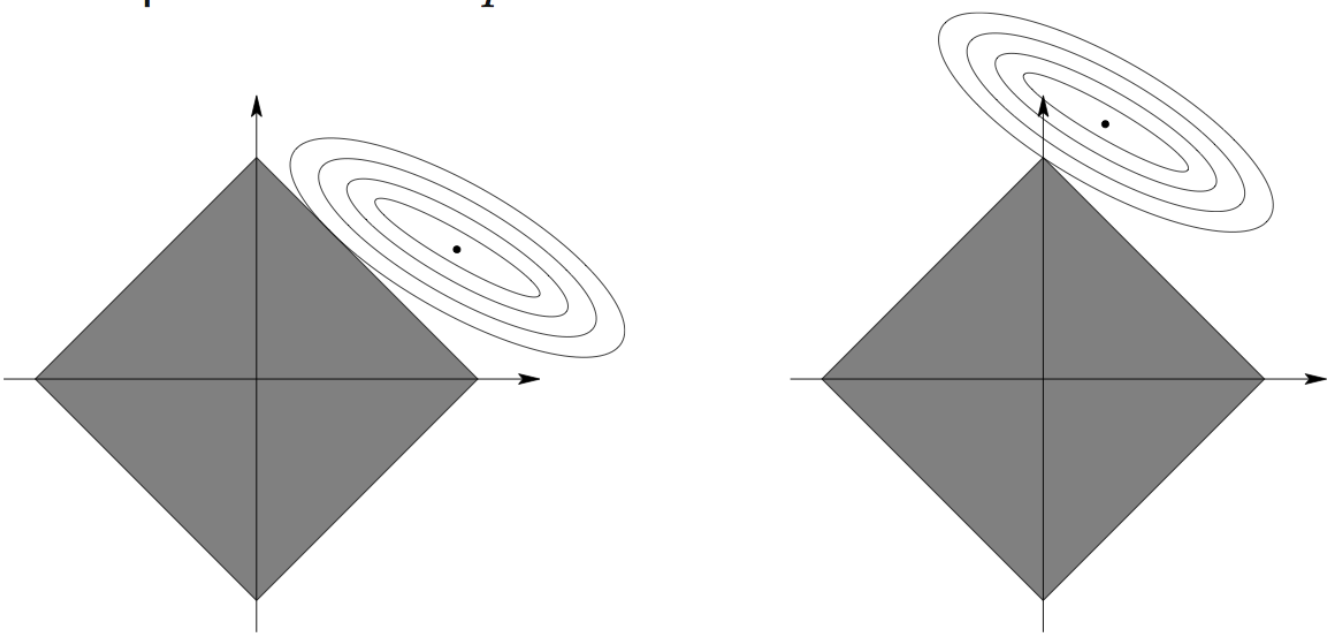
- Filter methods
- Wrapper methods
- Embedded methods



Example: ℓ_1 regularization

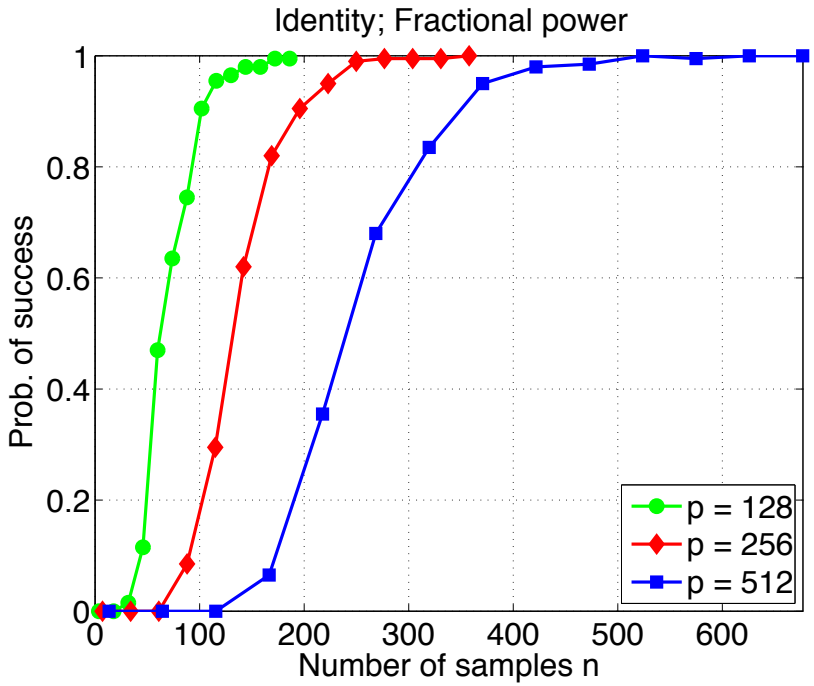
$$\min_{\beta \in \mathbb{R}^p} R(\beta) \quad \text{such that} \quad \sum_{i=1}^{p} |\beta_i| \leq C$$

Geometric interpretation with $p = 2$

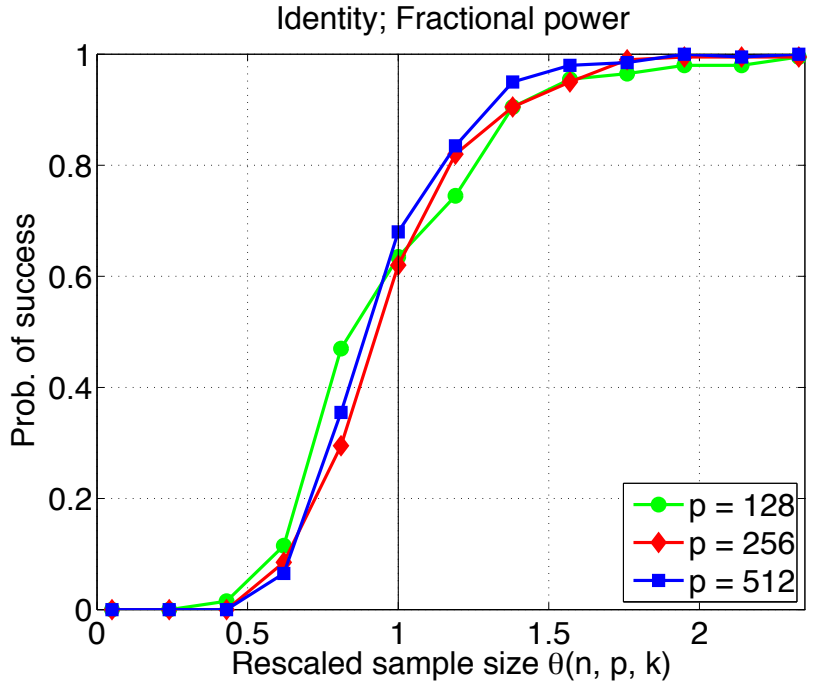


Leads to sparse models (feature selection)

ℓ_1 regularization works well in theory



(a)

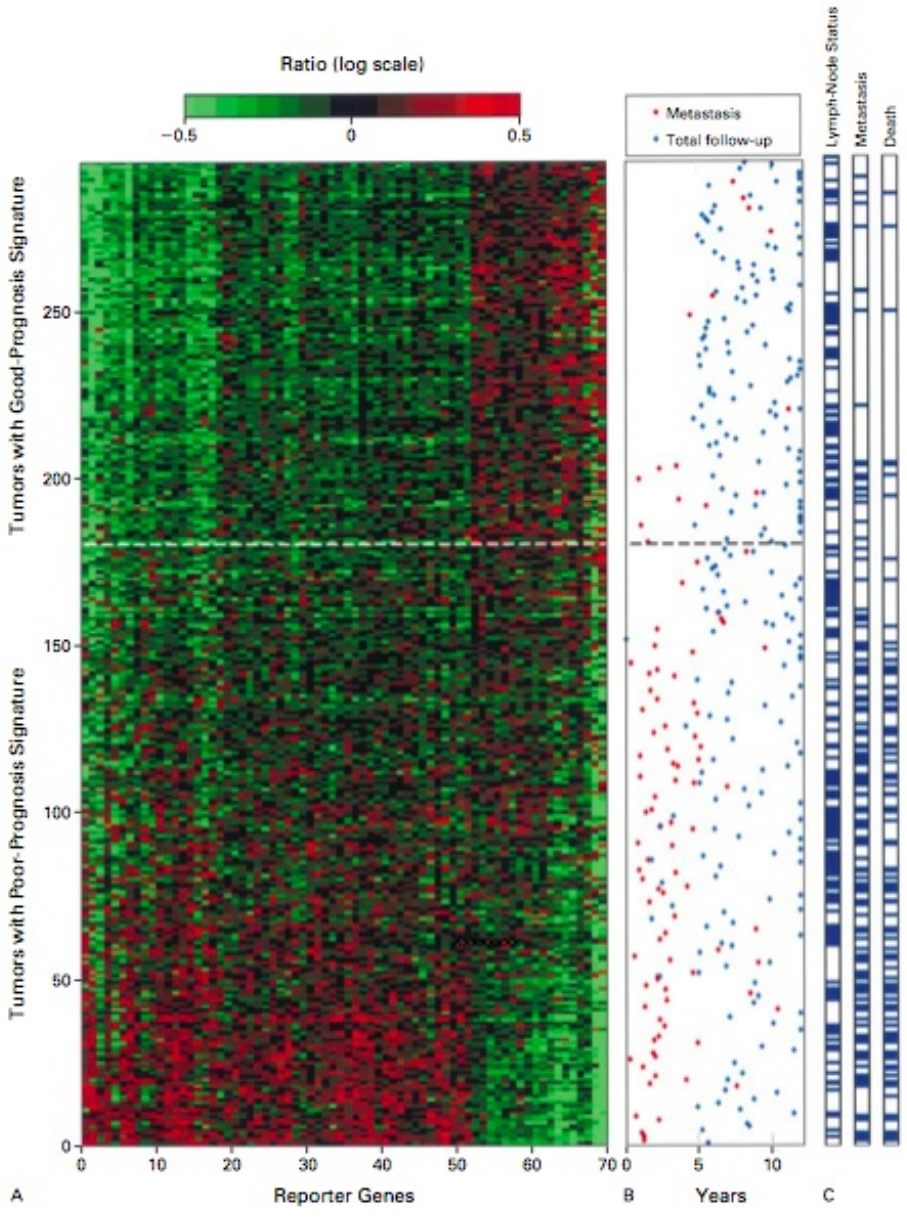


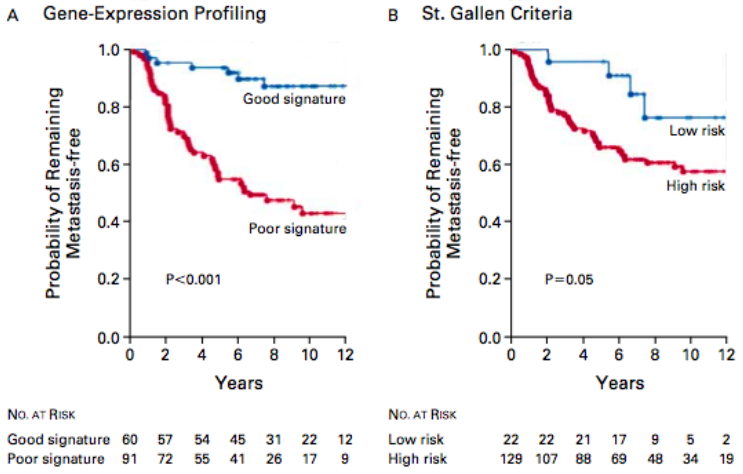
(b)

Fig. 1. (a) Plots of the success probability $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)]$ of obtaining the correct signed support versus the sample size n for three different problem sizes p, in all cases with sparsity $k = \lceil 0.40 p^{0.75} \rceil$. (b) Same simulation results with success probability plotted versus the rescaled sample size $\theta(n, p, k) = n/[2k \log(p - k)]$. As predicted by Theorems 3 and 4, all the curves now lie on top of one another. See Section VII for further simulation results.

- $n \sim s \ln(p - s)$, see e.g. Wainwright (2009) and many more
- If features are not "too correlated"

Example: 70-gene breast cancer prognostic signature





van ’t Veer et al. (2002);
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)

3 genes in common

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

No feature selection method seems to work well

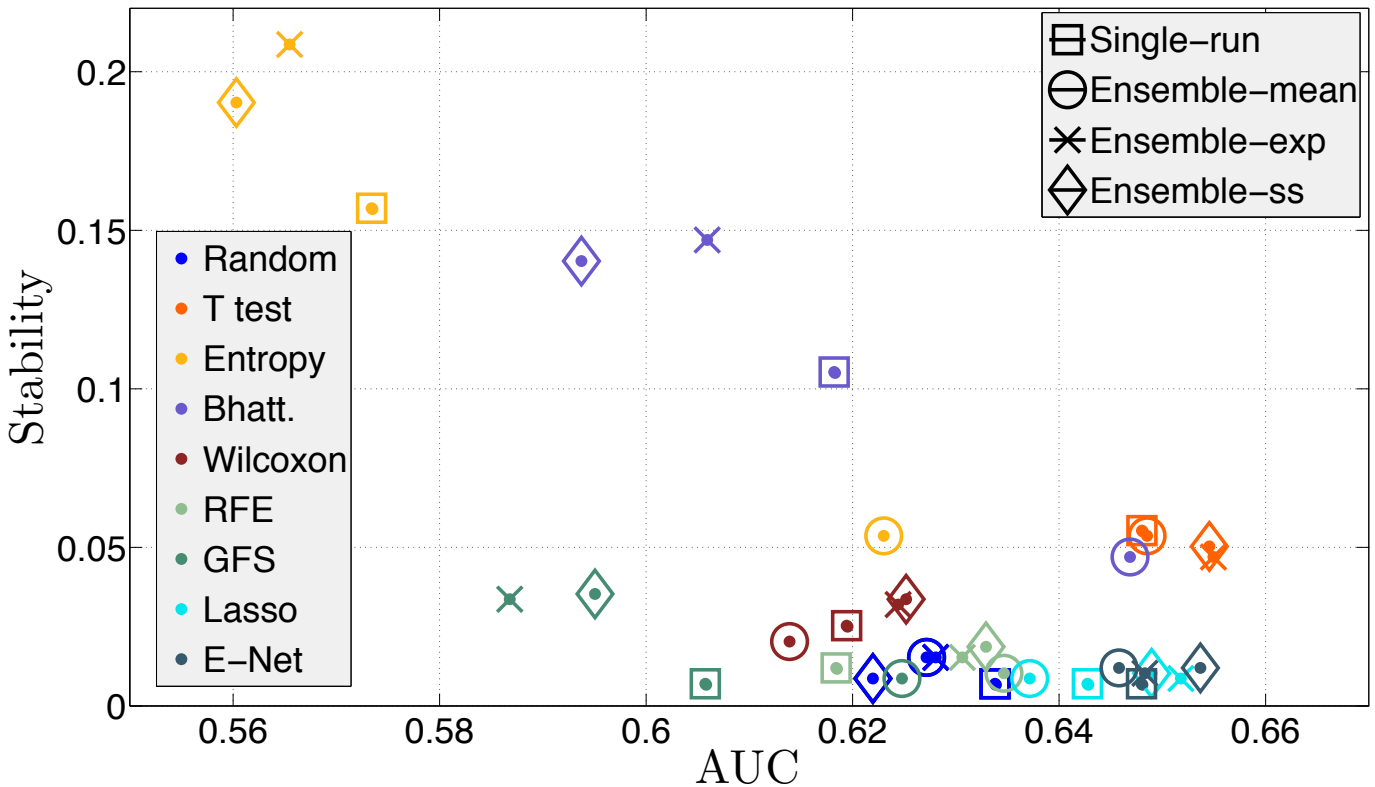
OPEN ACCESS Freely available online

PLoS one

The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures

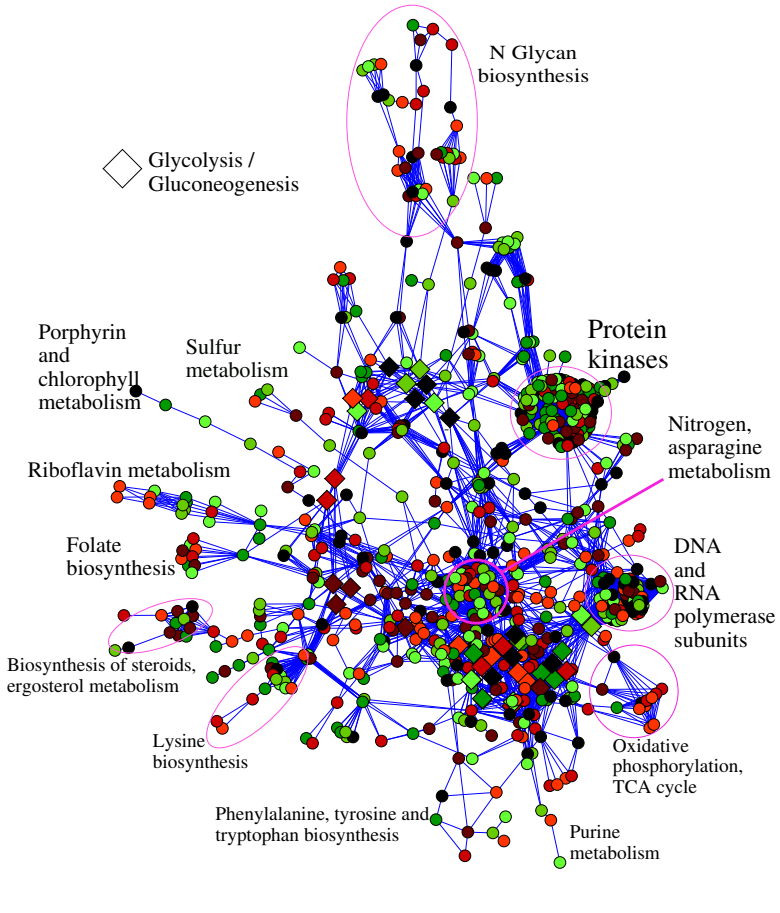
Anne-Claire Haury[1,2,3]*, Pierre Gestraud[1,2,3], Jean-Philippe Vert[1,2,3]

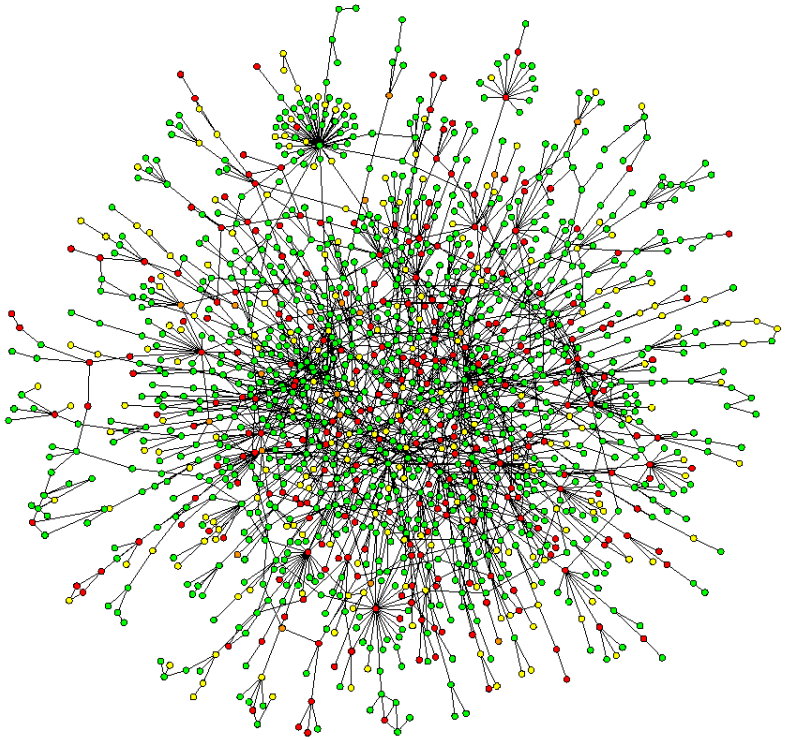
1 Mines ParisTech, Centre for Computational Biology, Fontainebleau, France, **2** Institut Curie, Paris, France, **3** Institut National de la Santé et de la Recherche Médicale, Paris, France



Haury et al. (2011)

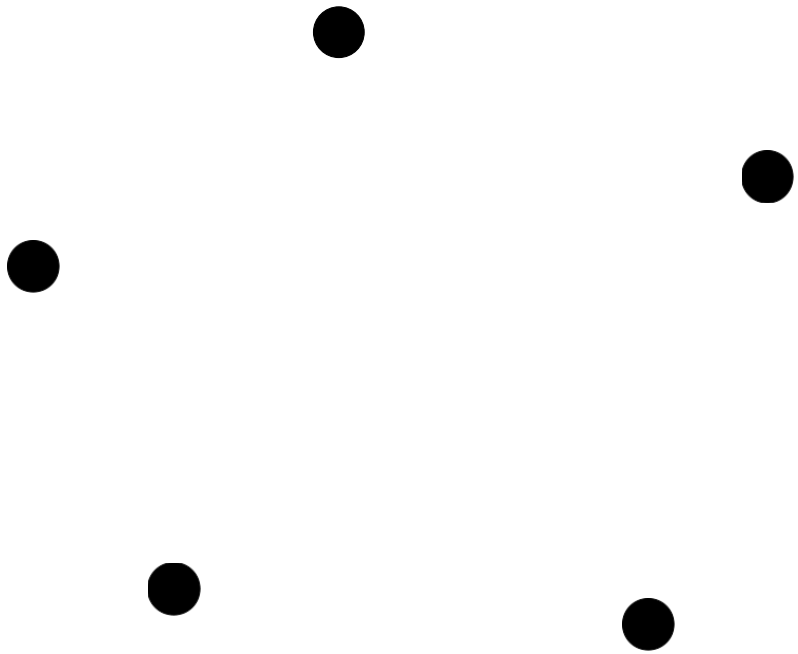
Adding prior knowledge





- Genes (=features) are known to interact with each other
- Predictive features are likely to interact
- Can we "bias" the set of selected features towards sets of interacting genes?

Atomic Norm (Chandrasekaran et al., 2012)



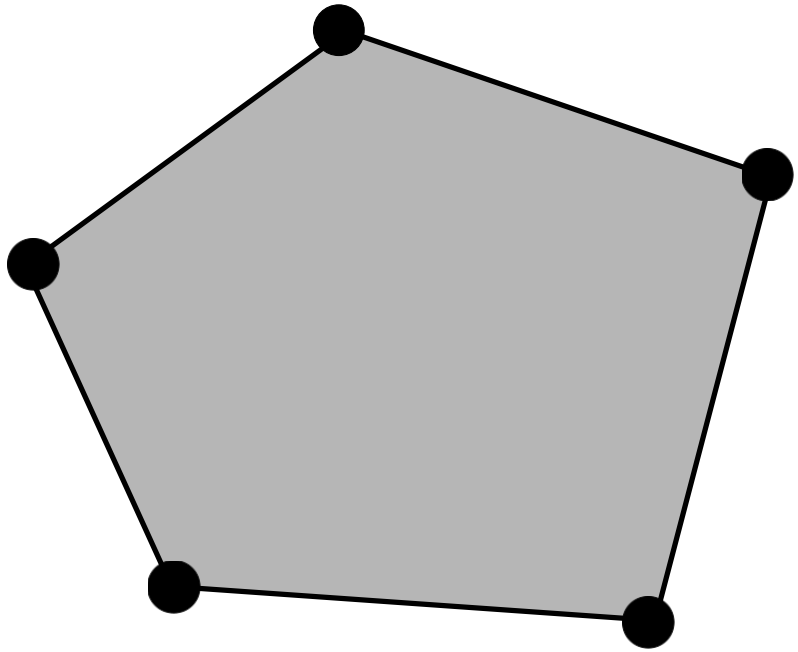
Definition

Given a set of atoms $\mathcal{A}$, the associated atomic norm is

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

$\mathcal{A}$ should be centrally symmetric and span $\mathbb{R}^p$

Atomic Norm (Chandrasekaran et al., 2012)



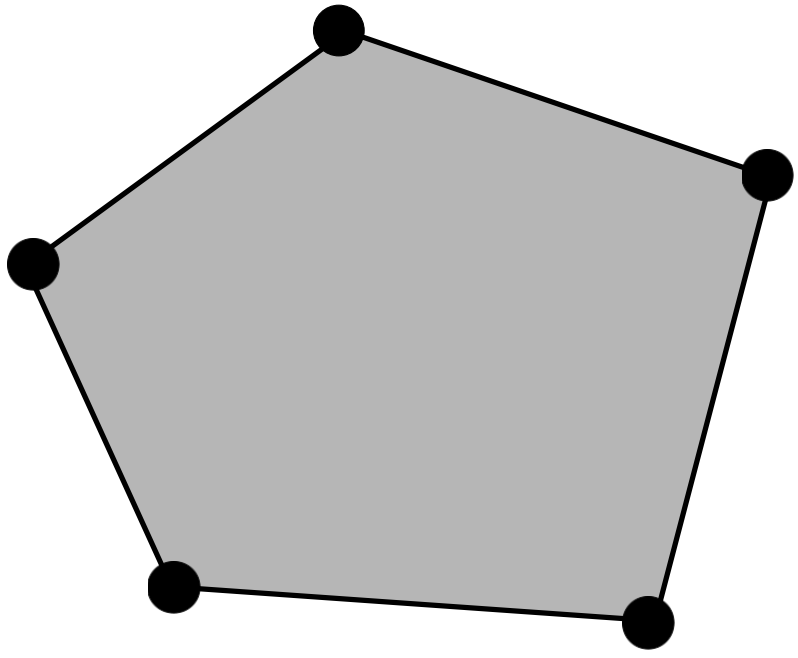
Definition

Given a set of atoms $\mathcal{A}$, the associated atomic norm is

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

$\mathcal{A}$ should be centrally symmetric and span $\mathbb{R}^p$

Equivalent formulations



$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}$$

$$\|x\|_{\mathcal{A}} = \inf\left\{ \sum_{a\in\mathcal{A}} c_a \;\mid\; x = \sum_{a\in\mathcal{A}} c_a\, a, \quad c_a > 0,\ \forall a \in \mathcal{A} \right\}$$

$$\|x\|^*_{\mathcal{A}} = \sup_{a\in\mathcal{A}} \langle a, x\rangle$$

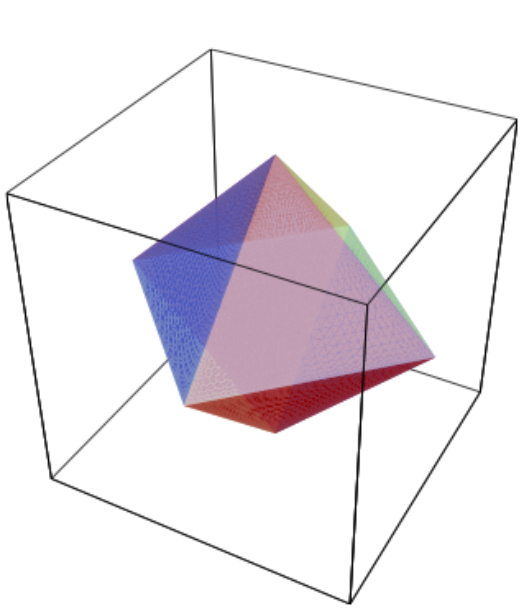
Examples

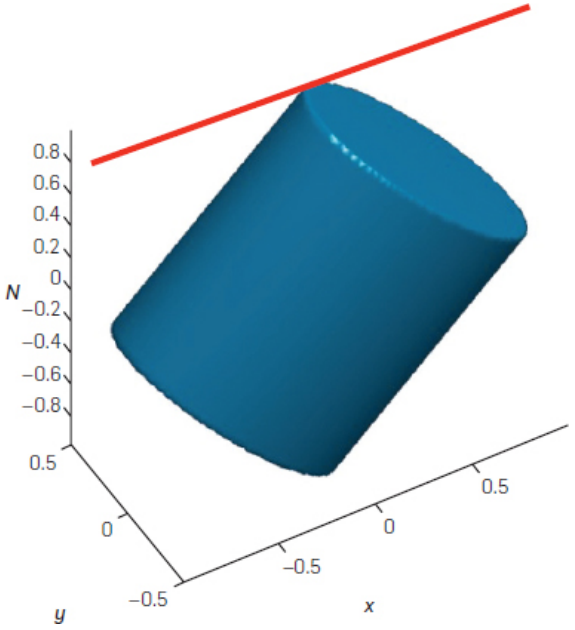
- Vector ℓ_1-norm: $x \in \mathbb{R}^p \mapsto \|x\|_1$

$$\mathcal{A} = \left\{ \pm e_k \;\middle|\; 1 \leq k \leq p \right\}$$

- Matrix trace norm: $Z \in \mathbb{R}^{m_1 \times m_2} \mapsto \|Z\|_*$ (sum of singular value)

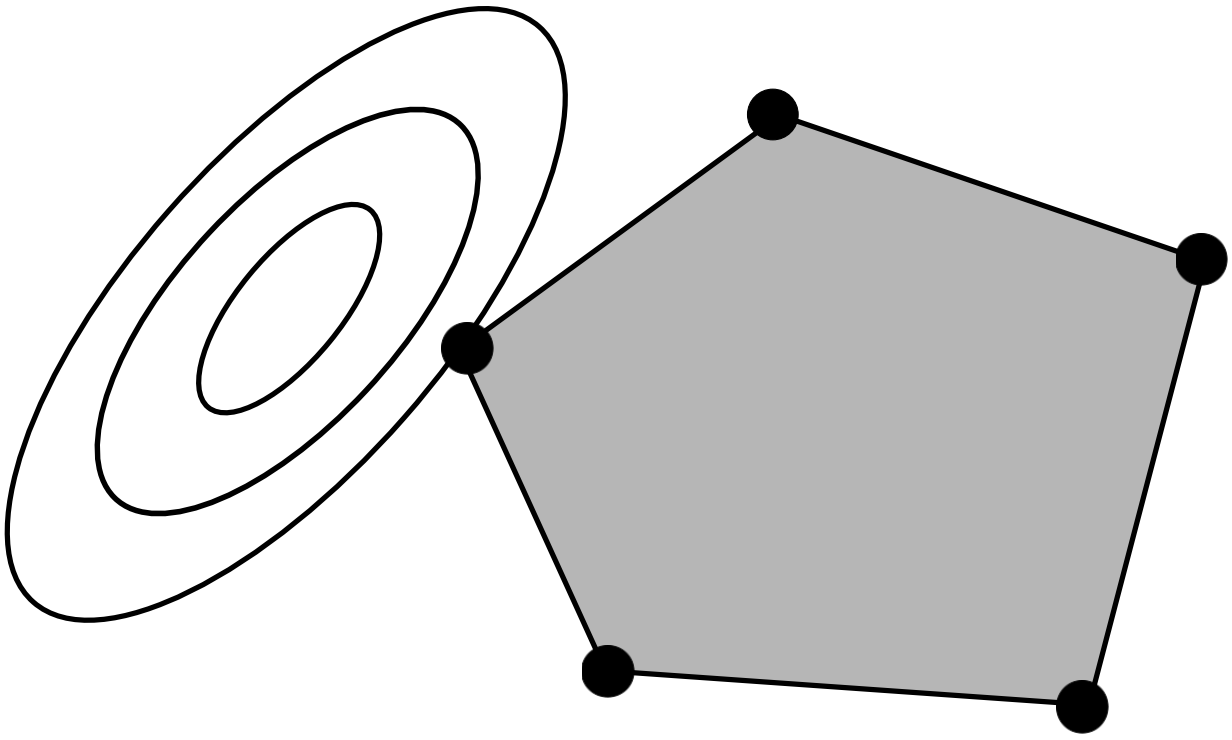
$$\mathcal{A} = \left\{ ab^\top \;:\; a \in \mathbb{R}^{m_1}, b \in \mathbb{R}^{m_2}, \| a \|_2 = \| b \|_2 = 1 \right\}$$





Learning with an Atomic Norm

$$\min_{\beta} R(\beta) \quad \text{such that} \quad \|\beta\|_{\mathcal{A}} \leq C$$



- Property: the solution β^* is a sparse combination of atoms
- More precisely, how "easy" is it to learn such a β^*?

Statistical dimension (Amelunxen et al., 2013)

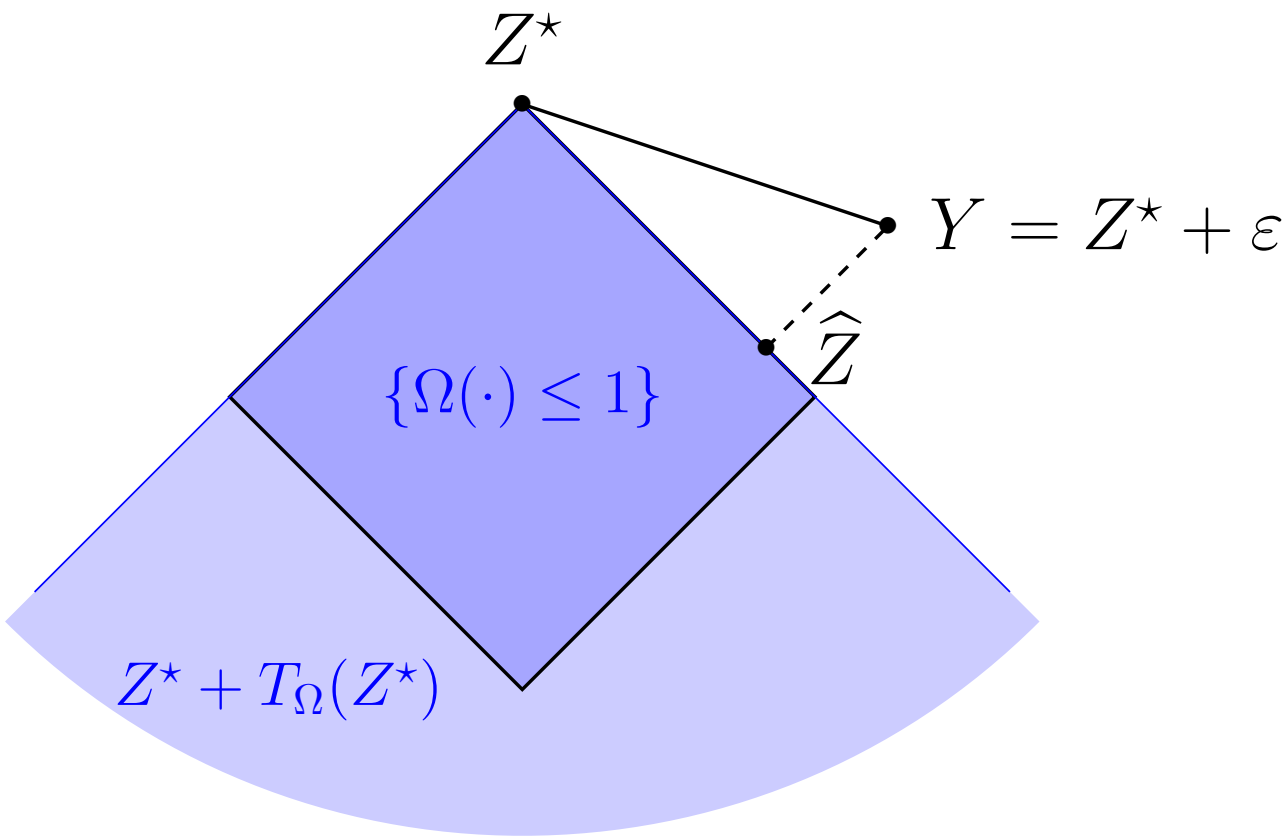


figure inspired by Amelunxen et al. (2013)

$$\mathfrak{S}(Z, \Omega) := \mathbb{E}\left[\left\| \Pi_{T_\Omega(Z)}(G) \right\|_{\text{Fro}}^2 \right] ,$$

Nullspace property and $\mathfrak{S}$ (Chandrasekaran et al., 2012)

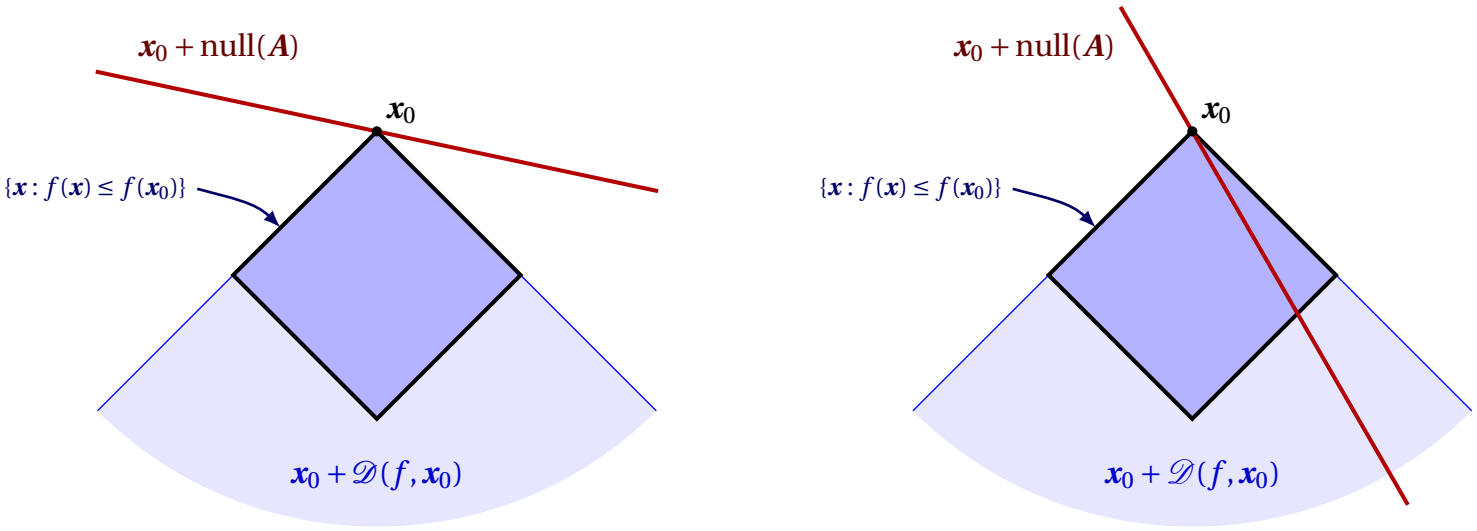


Figure from Amelunxen et al. (2013)

- With $X : \mathbb{R}^p \to \mathbb{R}^n$ random Gaussian matrix,

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\ \Omega(\beta) \quad \text{such that} \quad X\beta = y$$

  is equal to $\beta^\star$ w.h.p. as soon as $n \geq \mathfrak{S}(\beta^\star, \Omega)$.
- Similar results with noisy outputs etc..

Statistical dimensions of a few standard norms

Matrix norm	$\mathfrak{S}$	Vector norm	$\mathfrak{S}$
ℓ_1	$\Theta(kq \log \frac{m_1 m_2}{kq})$	ℓ_1	$\Theta(k \log \frac{p}{k})$
trace-norm	$\Theta(m_1 + m_2)$	ℓ_2	p
$\ell_1 +$ trace	$\Omega\big(kq \wedge (m_1 + m_2)\big)$	elastic net	$\Theta(k \log \frac{p}{k})$
(k, q)-trace	$\mathcal{O}((k \vee q) \log (m_1 \vee m_2))$	k-support	$\Theta(k \log \frac{p}{k})$

Lower bound for ℓ_1+ trace norm based on a result of Oymak et al. (2012)
$f = \Theta(g)$ means $(f = \mathcal{O}(g) \& g = \mathcal{O}(f))$

$f = \Omega(g)$ means $g = \mathcal{O}(f)$

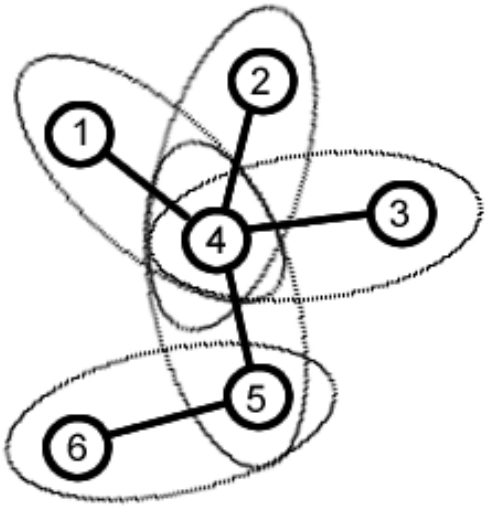
See Richard et al. (2014)

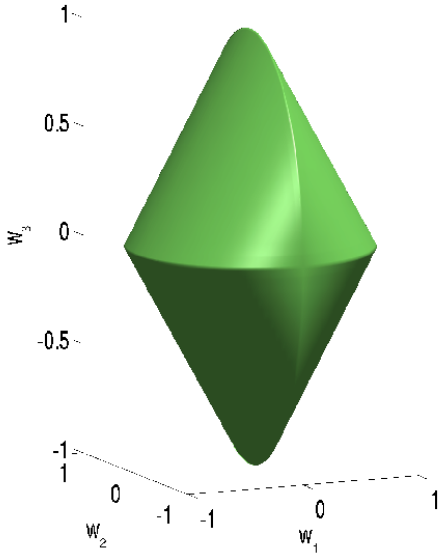
Making atomic norms

http://www.homemade-gifts-made-easy.com/make-paper-lanterns.html

- Choose atoms and make a chinese lantern
- Enforce statistical dimensions to solutions you expect
- Think of algorithms for constrained convex optimization

Graph lasso (Jacob et al., 2009)

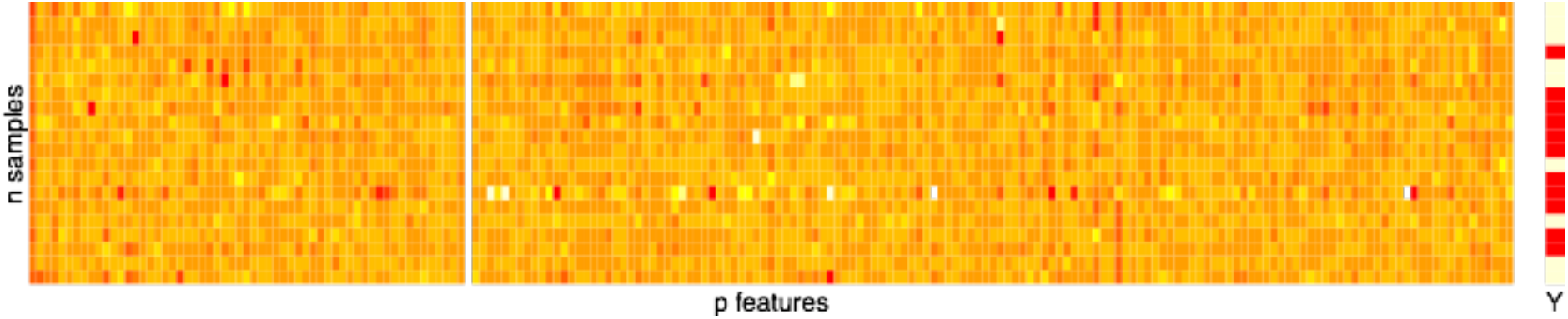




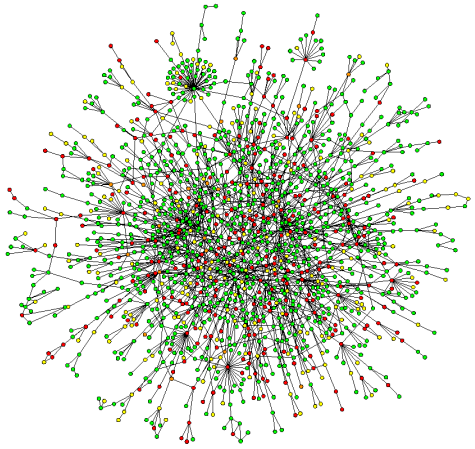
$$\Omega(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta$$

Application: breast cancer survival prediction

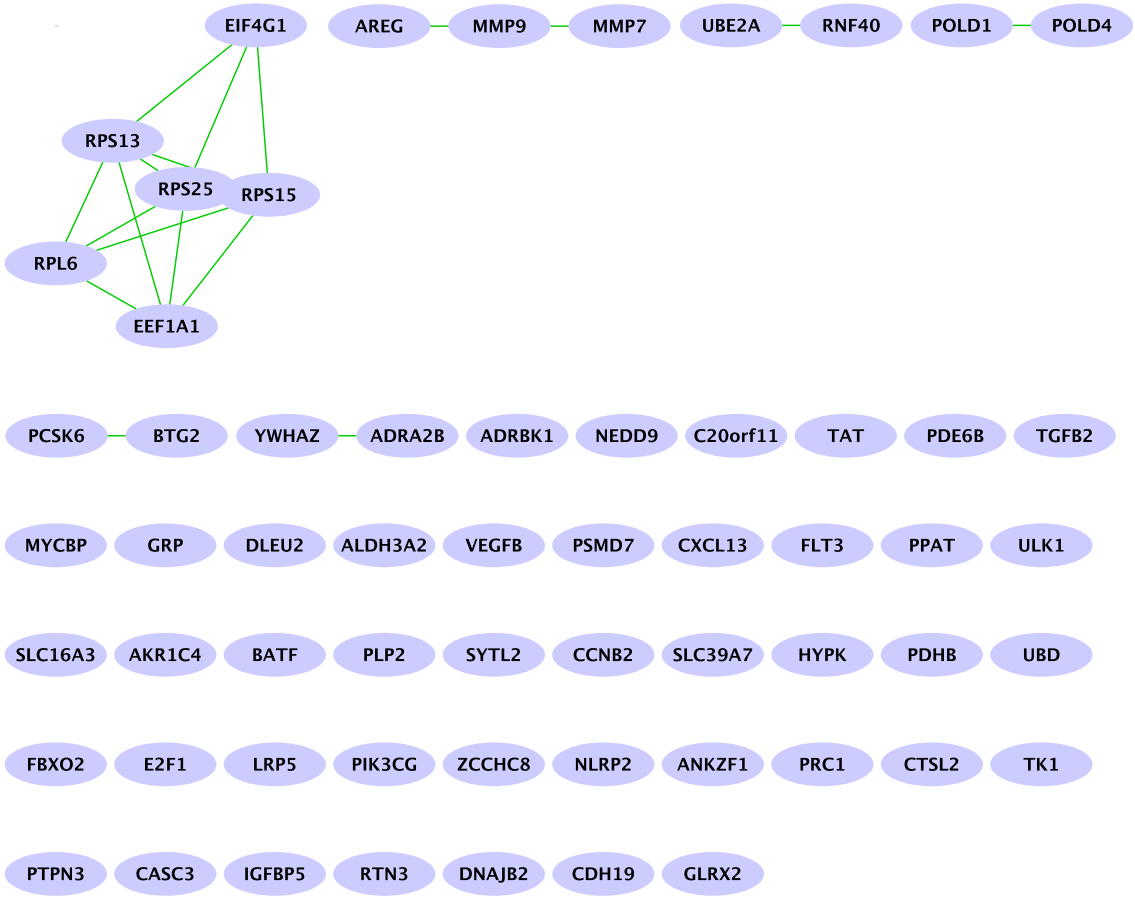
- $n = 295$ breast cancers, 78 metastatic vs 217 non-metastatic
- $p = 8,141$ gene expression measures (van de Vijver et al., 2002)



- Gene network compiled by Chuang et al. (2007)
- 57,235 interactions among 11,203 proteins

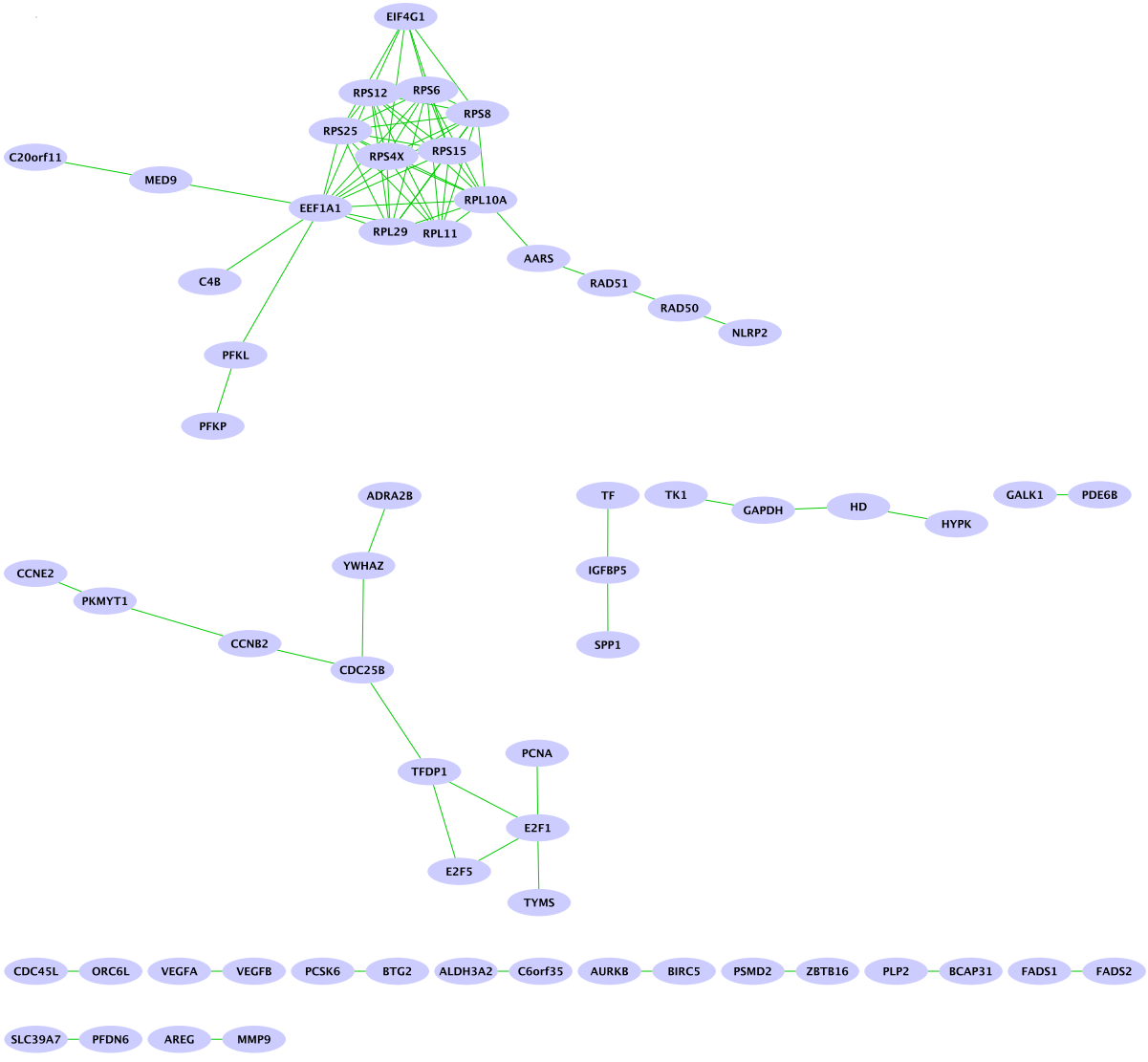


Lasso signature (accuracy 0.61)



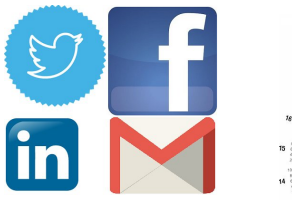
Jacob et al. (2009)

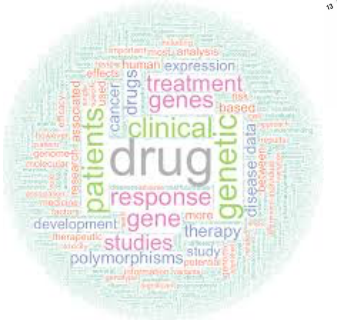
Graph Lasso signature (accuracy 0.64)

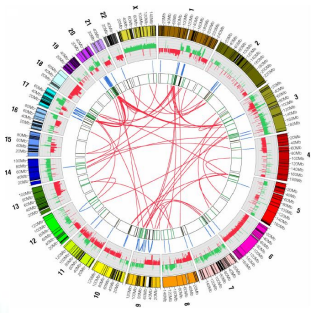


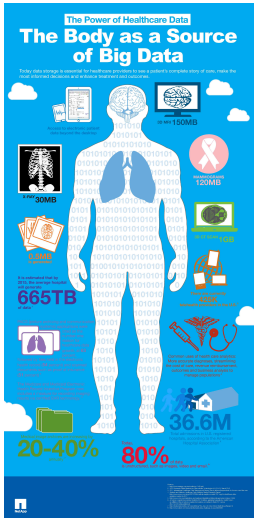
Jacob et al. (2009)

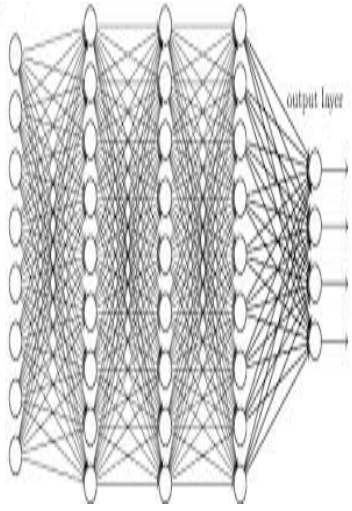
Conclusion











- Many new exciting problems and lots of data in computational genomics and precision medicine
- Machine learning tempting but sometimes challenging ($n << p$)
- Very active field of research at the interface of math / CS / biology

Thanks

References

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. Technical Report 1303.6672, arXiv, Mar 2013. URL http://arxiv.org/abs/1303.6672.

J. Berkson. Application of the logistic function to bio-assay. *J. R. Stat. Soc.*, 39(227):357–365, 1944. doi: 10.2307/2280041. URL http://dx.doi.org/10.2307/2280041.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012. doi: 10.1007/s10208-012-9135-7. URL http://dx.doi.org/10.1007/s10208-012-9135-7.

H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007. doi: 10.1038/msb4100180. URL http://dx.doi.org/10.1038/msb4100180.

A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*, 6(12):e28210, 2011. doi: 10.1371/journal.pone.0028210. URL http://dx.doi.org/10.1371/journal.pone.0028210.

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.

References (cont.)

S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. Technical Report 1212.3753, arXiv, 2012. URL http://arxiv.org/abs/1212.3753.

E. Richard, G. Obozinski, and J.-P. Vert. Tight convex relaxations for sparse matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, volume 27, pages 3284–3292. Curran Associates, Inc., 2014. URL https://papers.nips.cc/paper/5408-tight-convex-relaxations-for-sparse-matrix-factorization.

M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347(25):1999–2009, Dec 2002. doi: 10.1056/NEJMoa021967. URL http://dx.doi.org/10.1056/NEJMoa021967.

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.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE T. Inform. Theory.*, 55(5):2183–2202, 2009. doi: 10.1109/TIT.2009.2016018. URL http://dx.doi.org/10.1109/TIT.2009.2016018.

References (cont.)

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.