Machine learning in cancer genomics

Jean-Philippe Vert
Jean-Philippe.Vert@mines.org

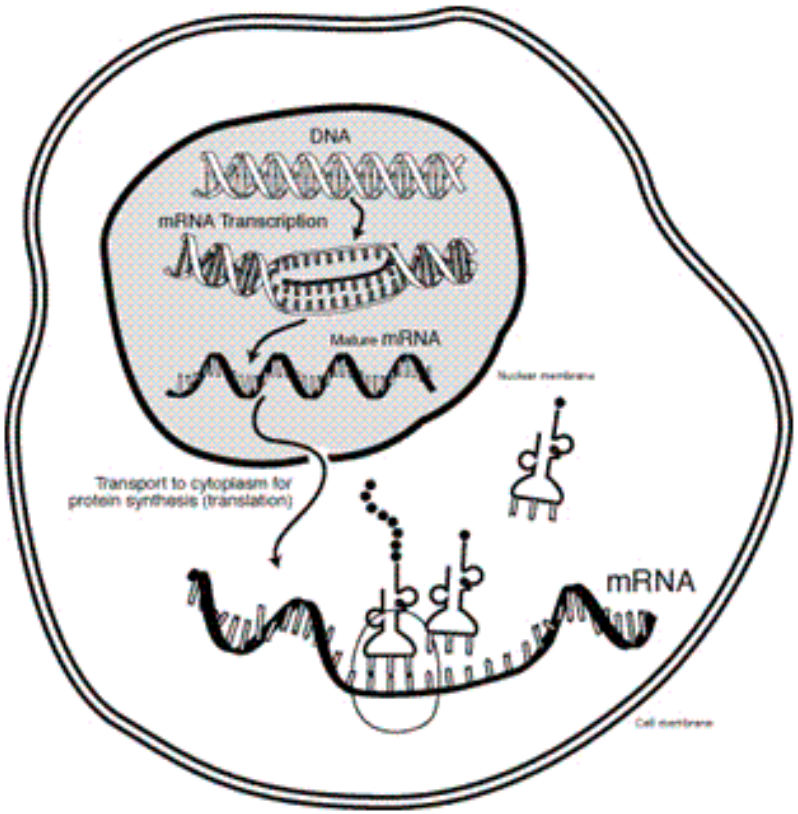
Mines ParisTech / Curie Institute / Inserm

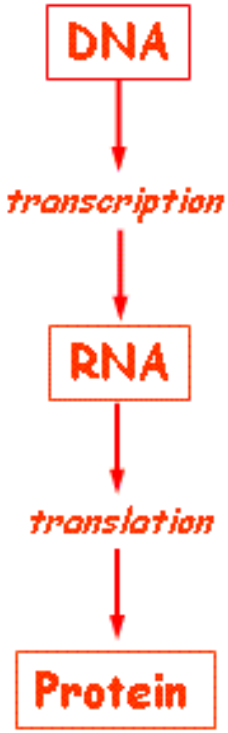
ReaDiLab conference, University of Tokyo, Nov 28, 2011.

Outline

Outline

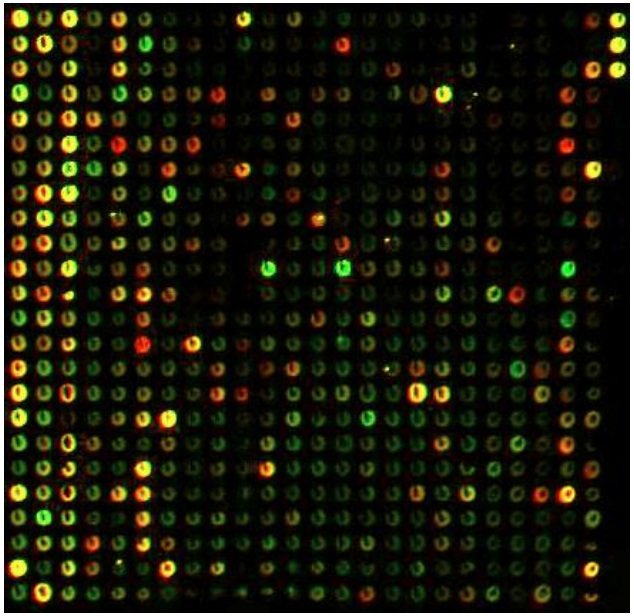
DNA $\rightarrow$ RNA $\rightarrow$ protein





- Cancer have abnormal genomes
- This leads to abnormal (dynamic) gene expression (RNA)

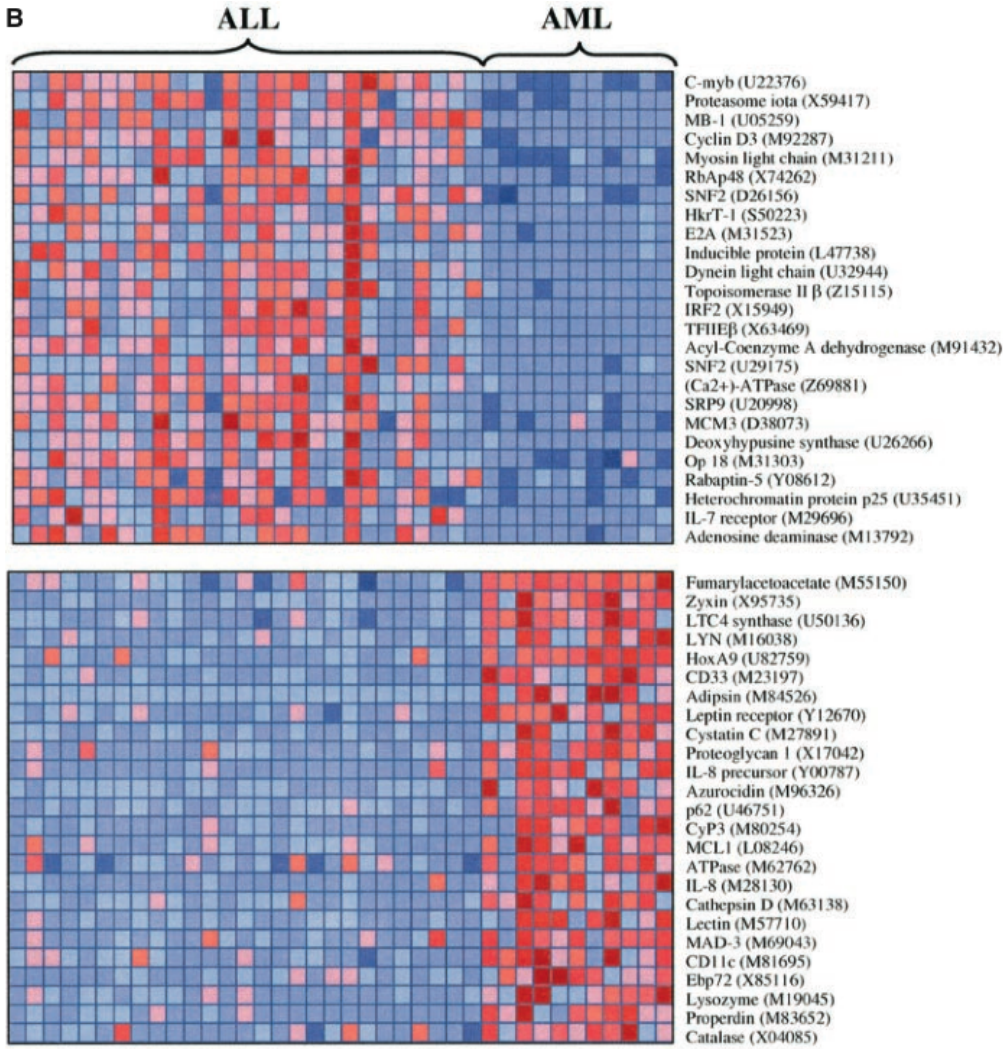
Tissue profiling with DNA chips



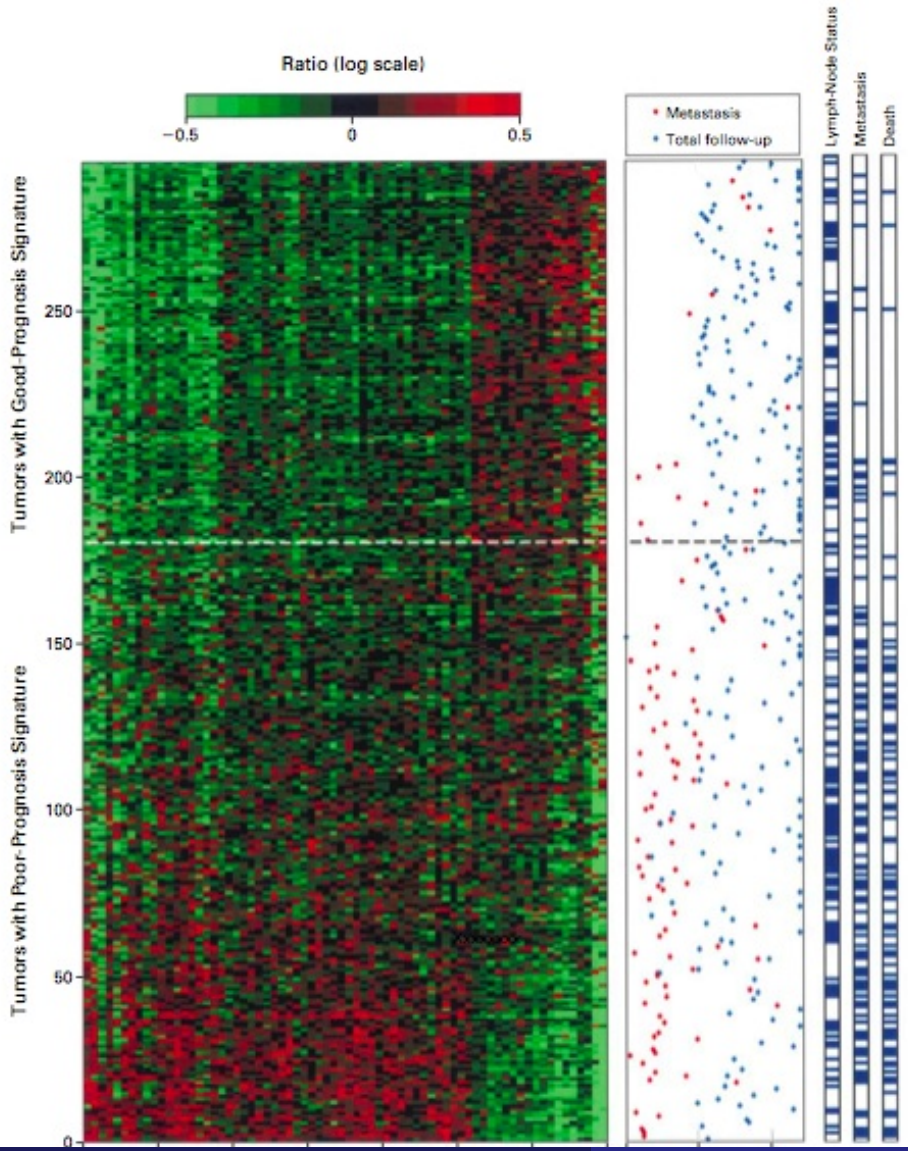
Data

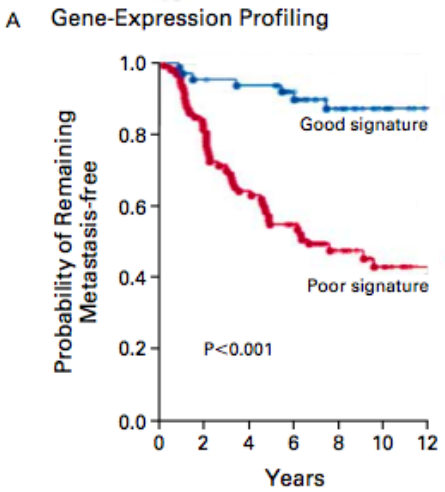
- Gene expression measures for more than $10k$ genes
- Measured typically on less than a few 100's samples

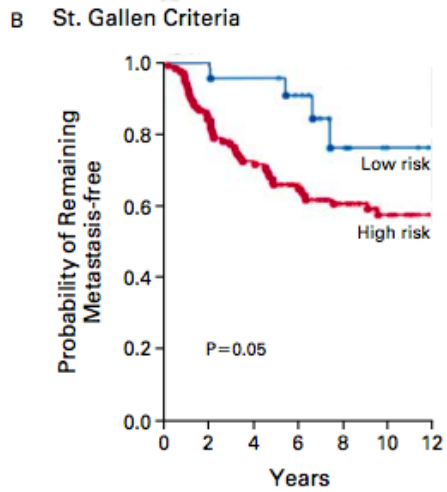
Can we identify the cancer subtype? (diagnosis)



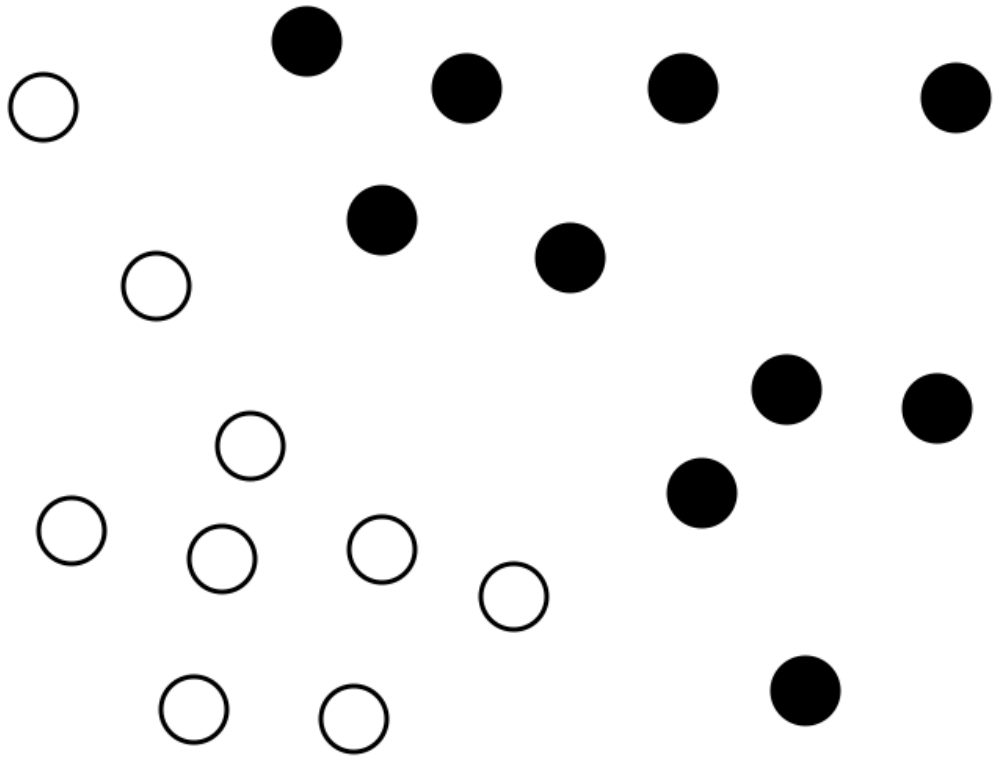
Can we predict the future evolution (prognosis), the response to drugs (theragnosis)?





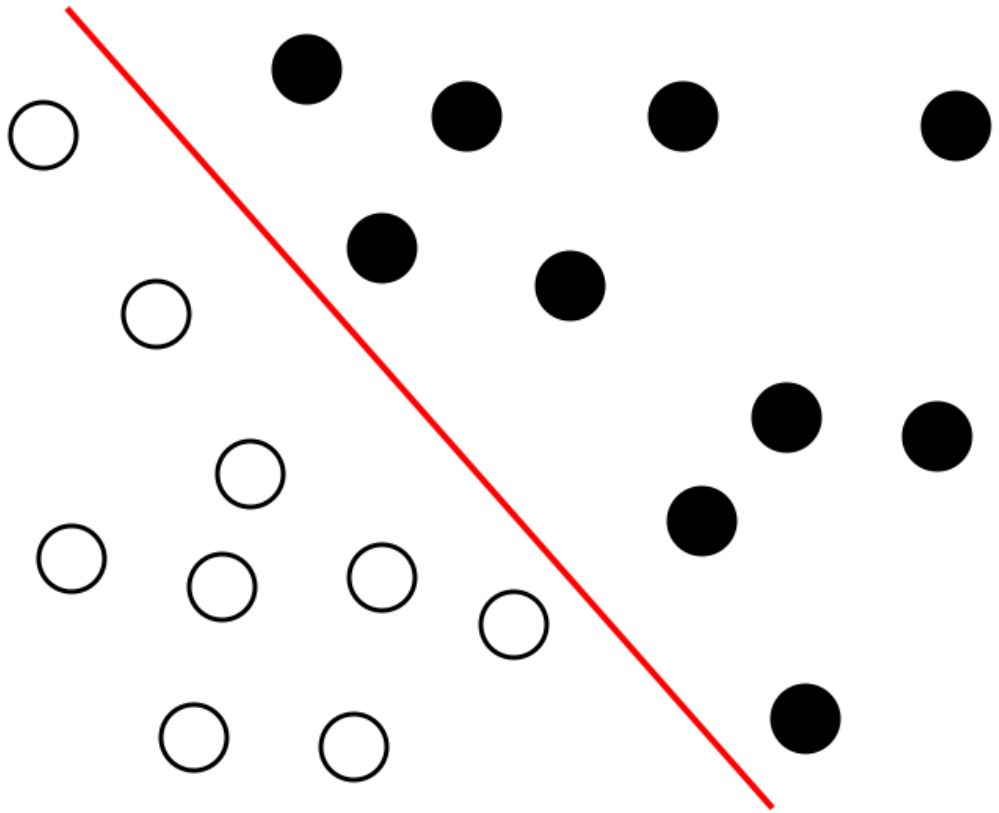


Machine learning (a.k.a. pattern recognition, supervised classification)



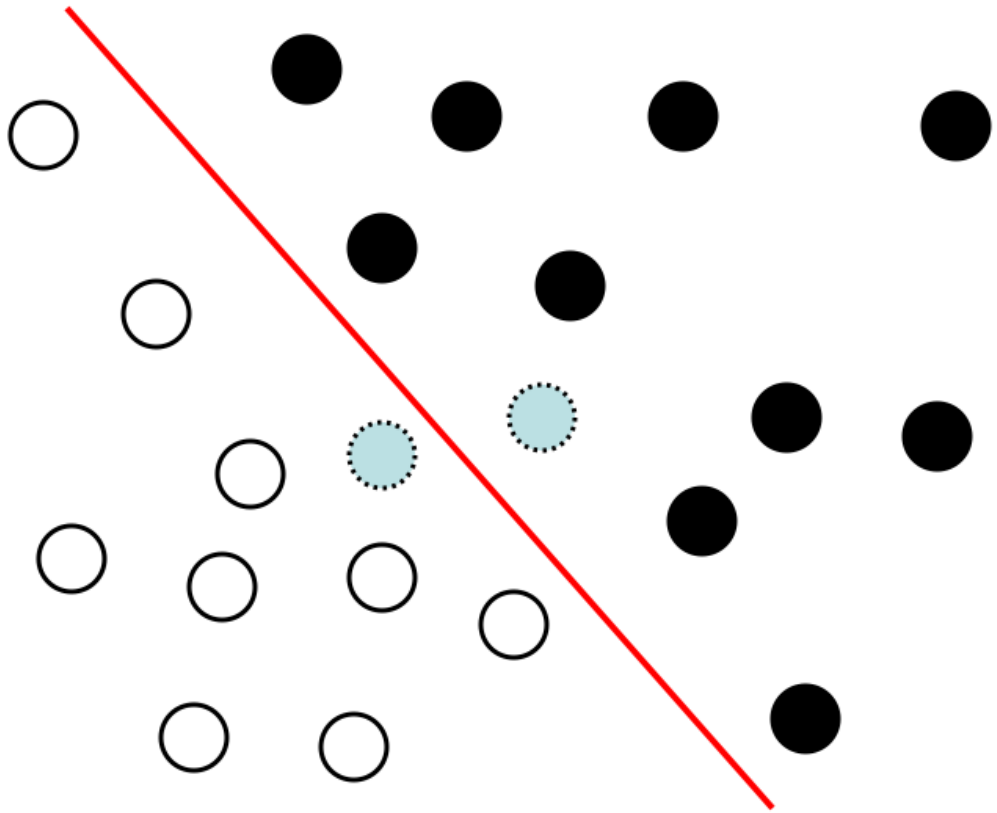
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning (a.k.a. pattern recognition, supervised classification)



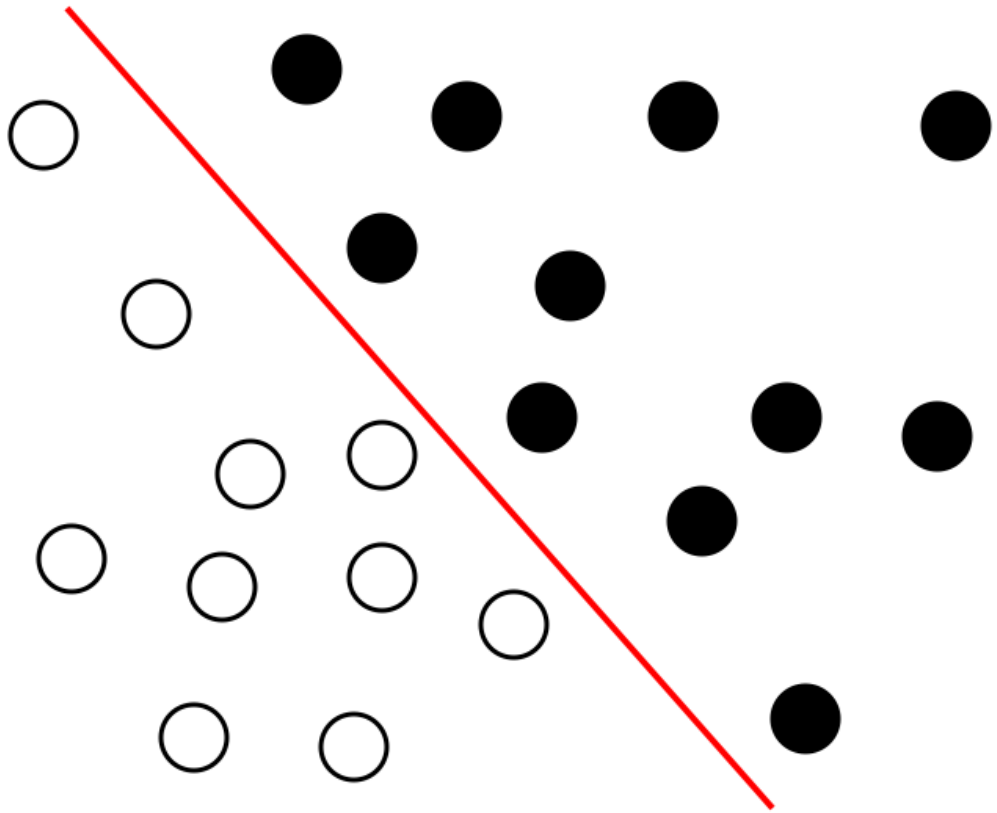
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning (a.k.a. pattern recognition, supervised classification)



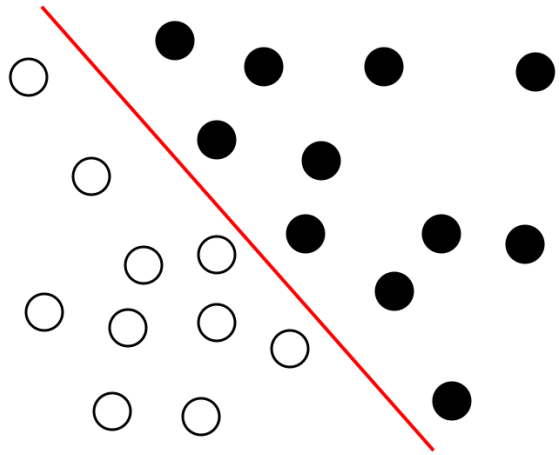
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning (a.k.a. pattern recognition, supervised classification)



1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning in bioinformatics



Genome annotation, systems biology, personalized medicine...

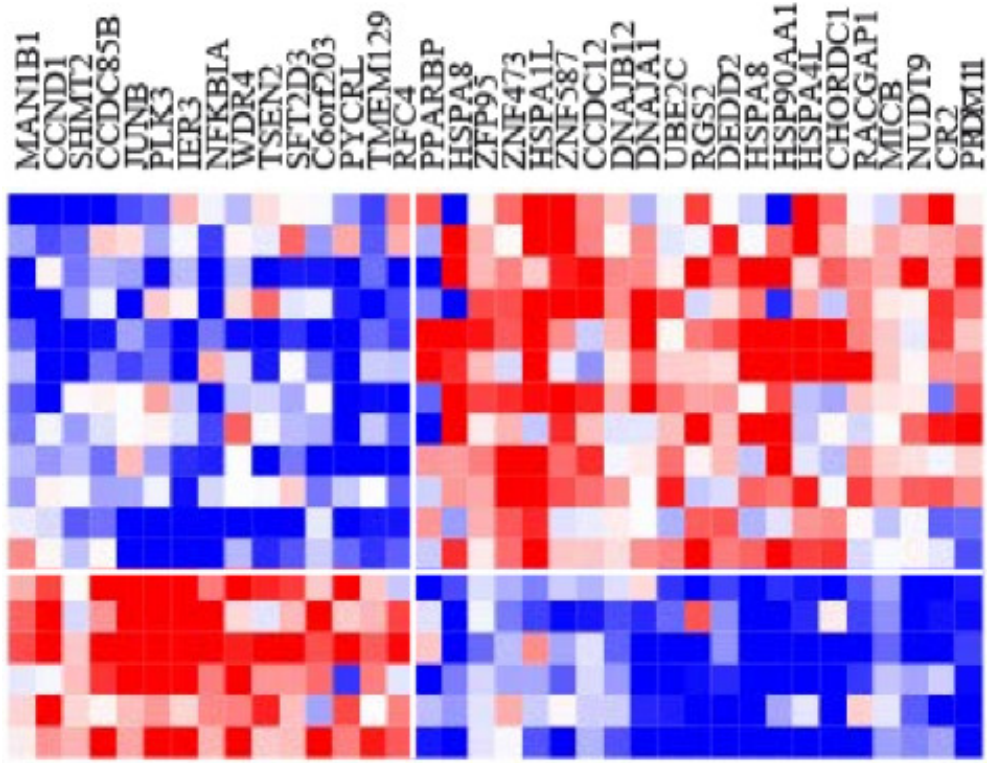
Challenges

- Few samples
- High dimension
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models

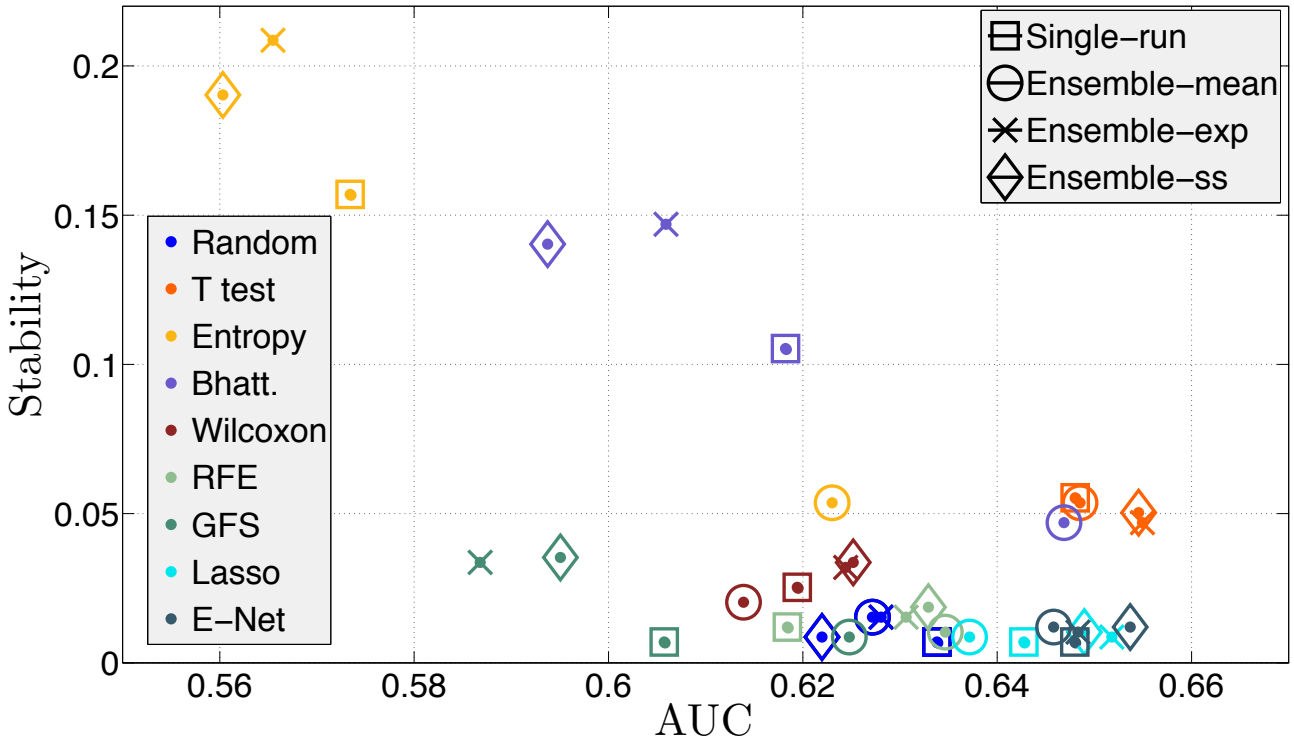
Gene selection, molecular signature

The idea

- We look for a limited set of genes that are sufficient for prediction.
- This should improve predictive accuracy (for statistical reasons)
- Selected genes should inform us about the underlying biology



But... unstability of selected features



- Can we go beyond generic methods, and design new methods better adapted to this scenario (10^4 genes, 10^2 samples)?
- How to include prior knowledge in the inference process?

Outline

ML with shrinkage estimators

1. Define a large family of "candidate classifiers", e.g., linear predictors:

$$f_\beta(x) = \beta^\top x \quad \text{for } x \in \mathbb{R}^p$$

2. For any candidate classifier f_β, quantify how "good" it is on the training set with some empirical risk, e.g.:

$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n}(f_\beta(x_i) - y_i)^2 \,.$$

3. Choose β that achieves the minimium empirical risk, subject to some constraint:

$$\min_\beta R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$

ML with shrinkage estimators

1. Define a large family of "candidate classifiers", e.g., linear predictors:

$$f_\beta(x) = \beta^\top x \quad \text{for } x \in \mathbb{R}^p$$

2. For any candidate classifier f_β, quantify how "good" it is on the training set with some empirical risk, e.g.:

$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n}(f_\beta(x_i) - y_i)^2 \,.$$

3. Choose β that achieves the minimium empirical risk, subject to some constraint:

$$\min_\beta R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$

ML with shrinkage estimators

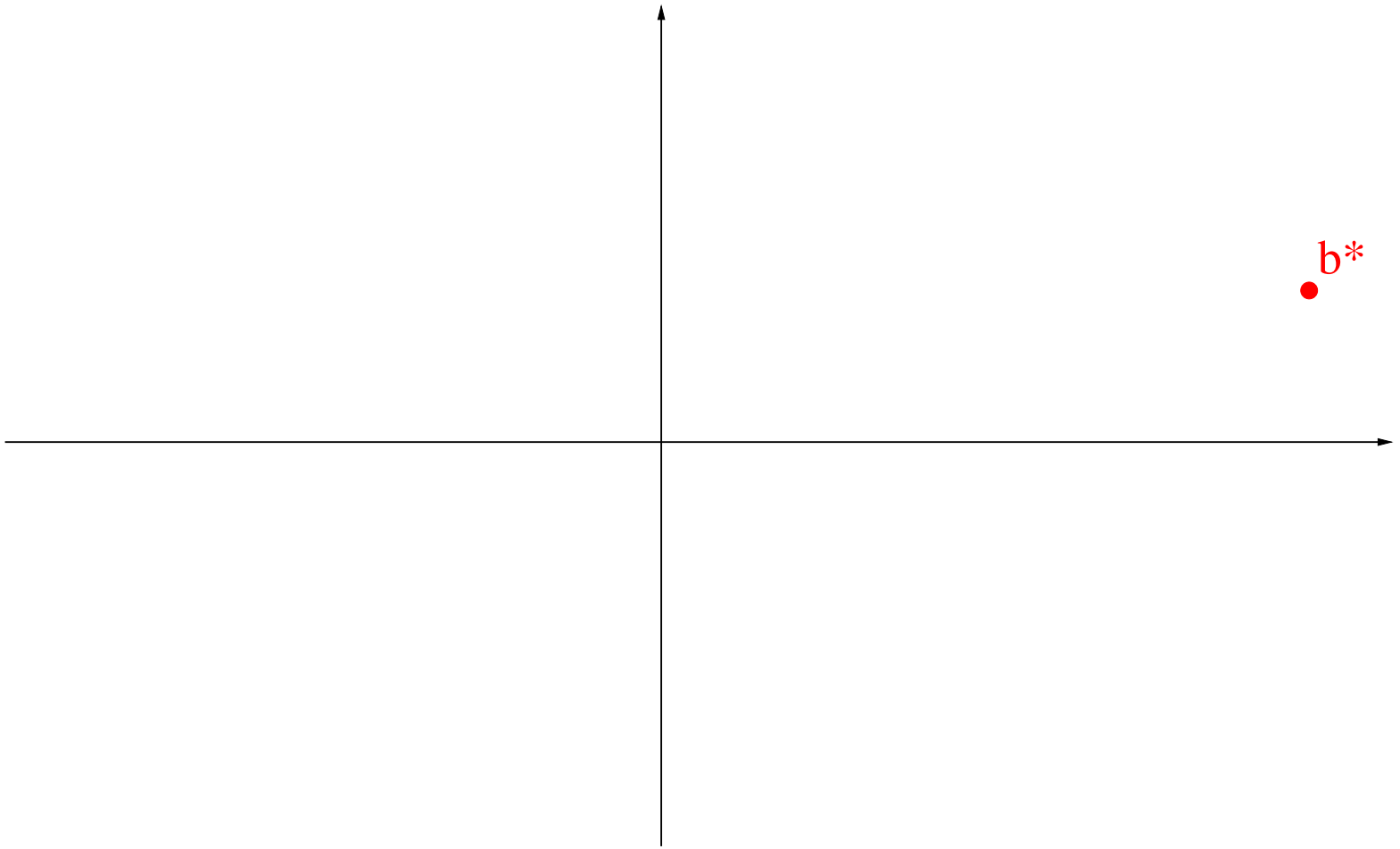
1. Define a large family of "candidate classifiers", e.g., linear predictors:
$$f_\beta(x) = \beta^\top x \quad \text{for } x \in \mathbb{R}^p$$

2. For any candidate classifier f_β, quantify how "good" it is on the training set with some empirical risk, e.g.:
$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n}(f_\beta(x_i) - y_i)^2 \,.$$

3. Choose β that achieves the minimium empirical risk, subject to some constraint:
$$\min_\beta R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$

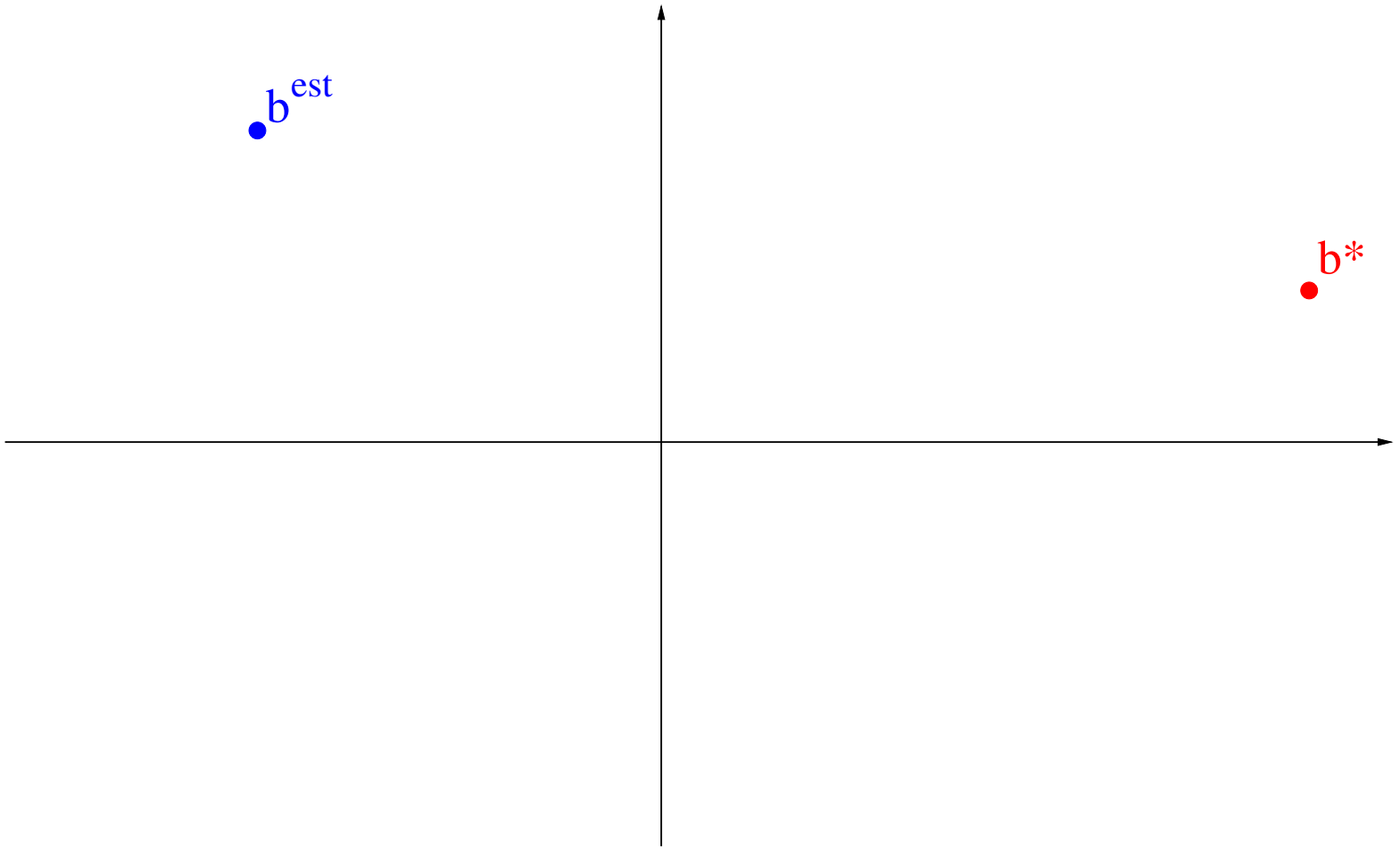
Why skrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$



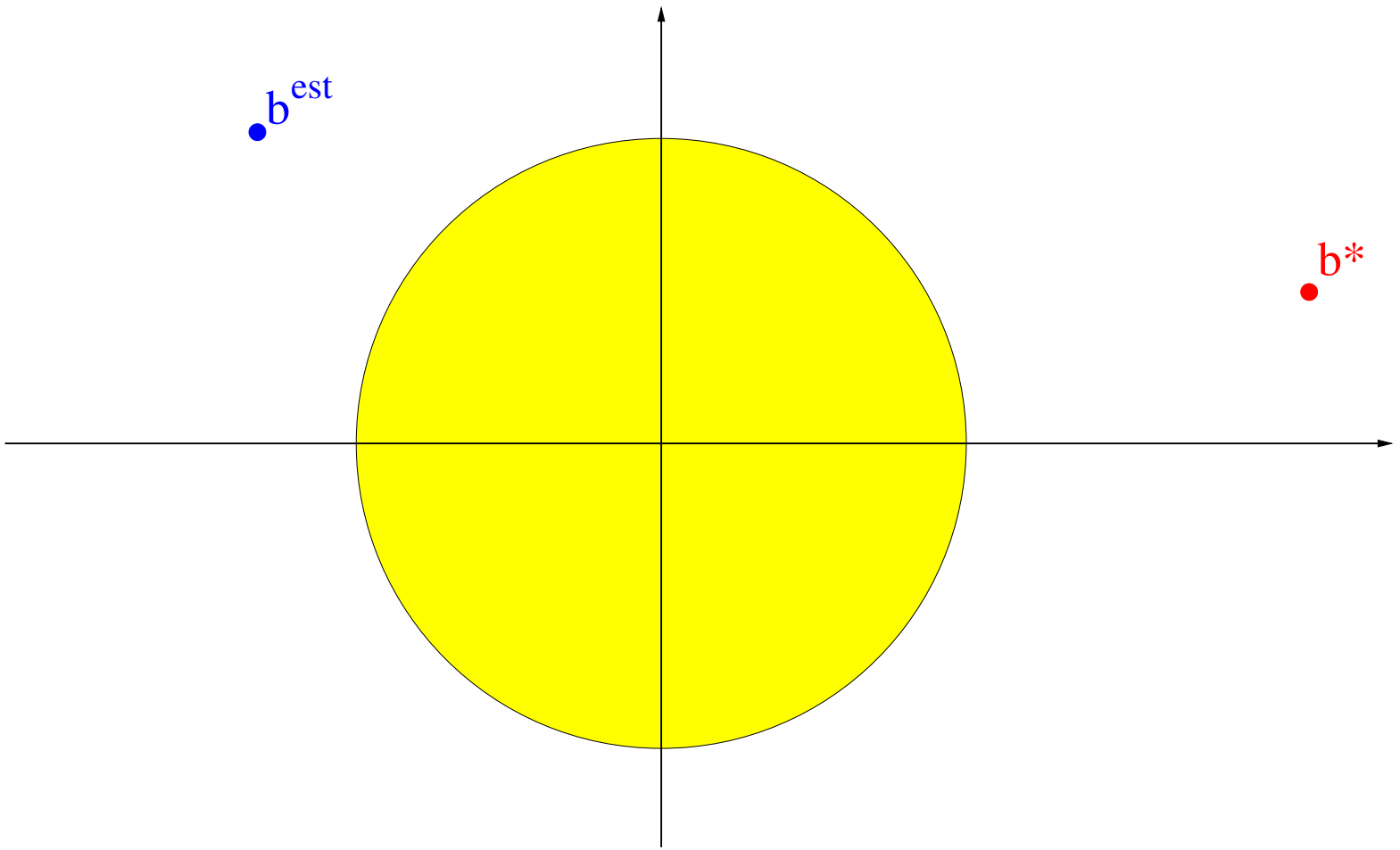
Why skrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C\,.$$



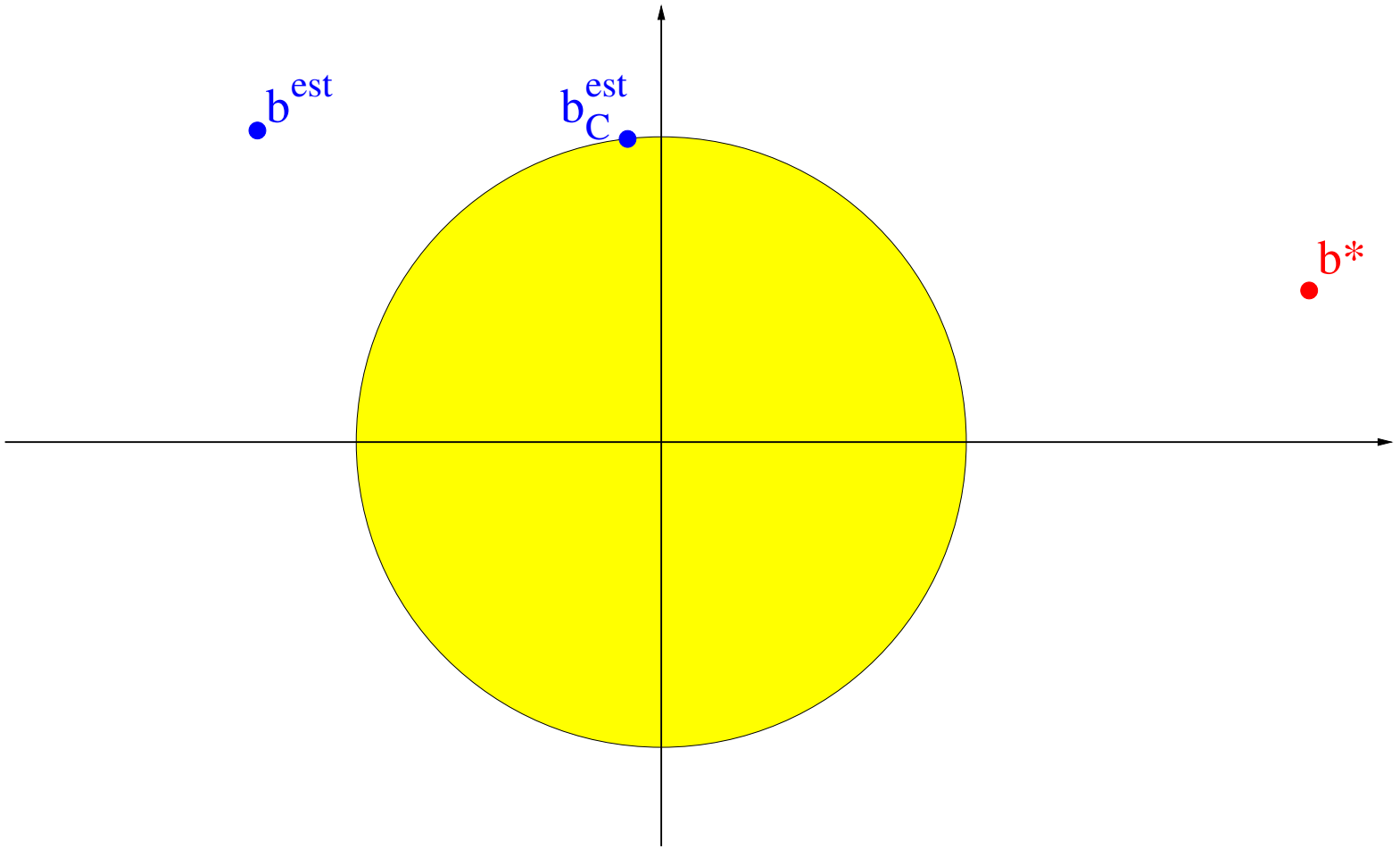
Why skrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$



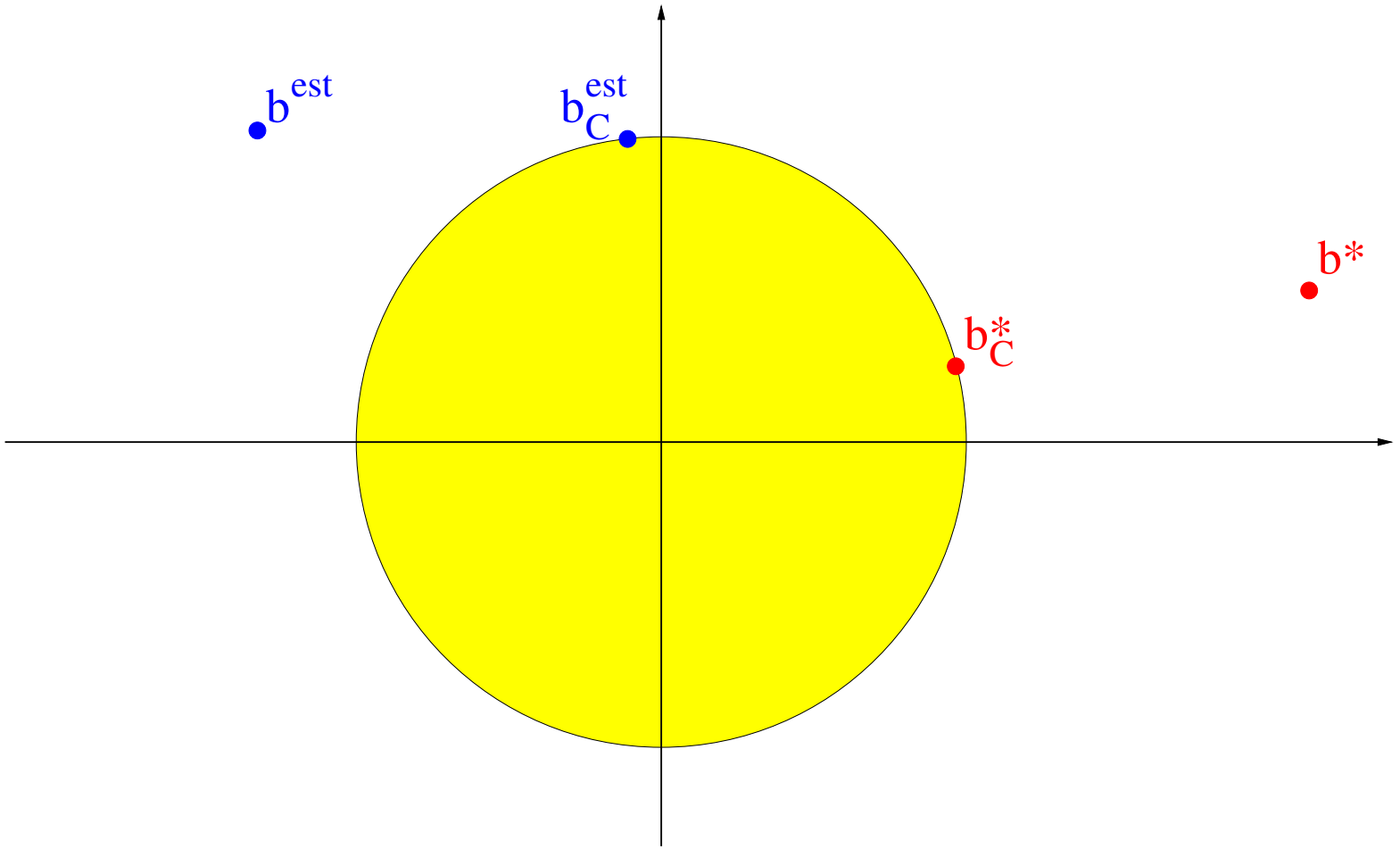
Why skrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$



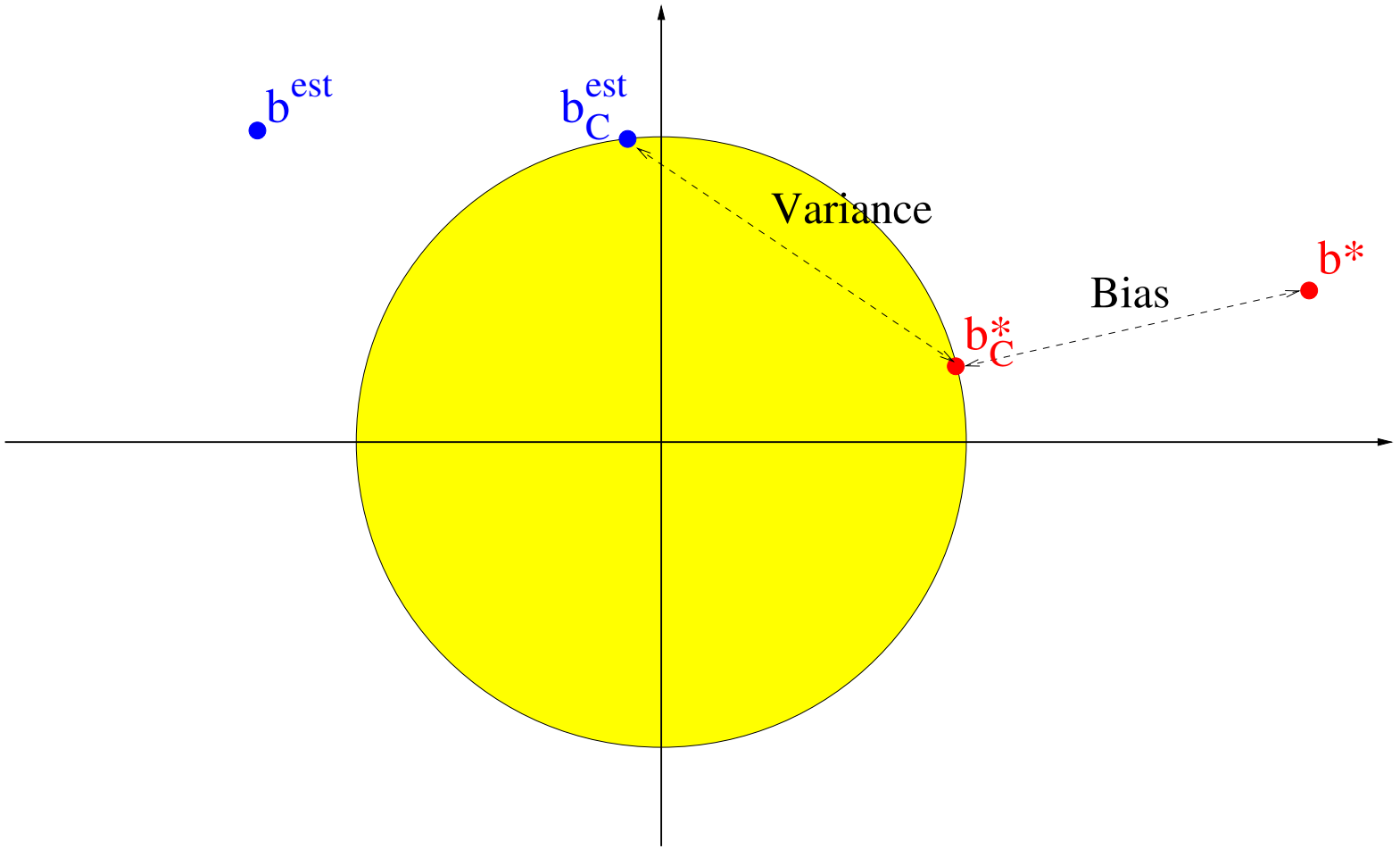
Why skrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C \,.$$

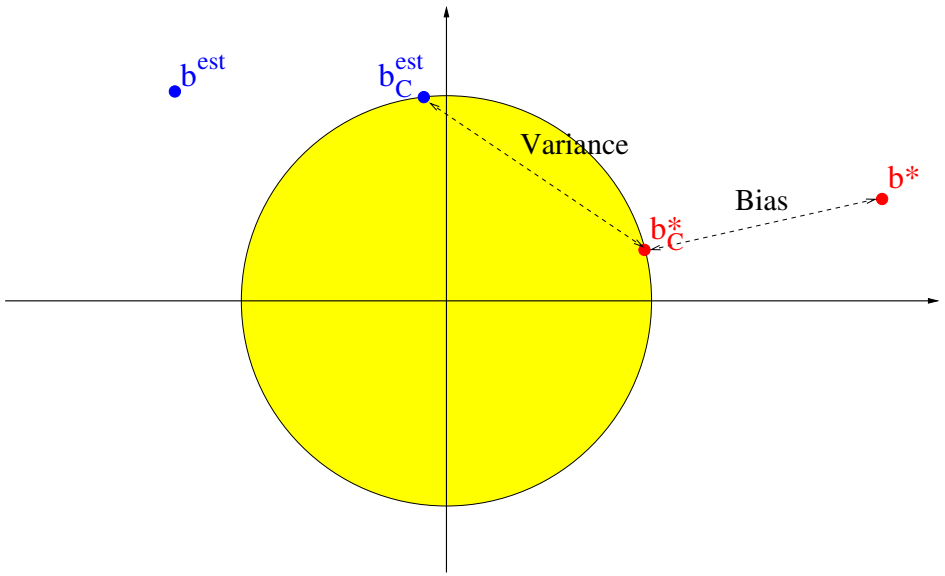


Why skrinkage classifiers?

$$\min_{\beta} R(\beta) \quad \text{subject to} \quad \Omega(\beta) \leq C\,.$$



Why skrinkage classifiers?



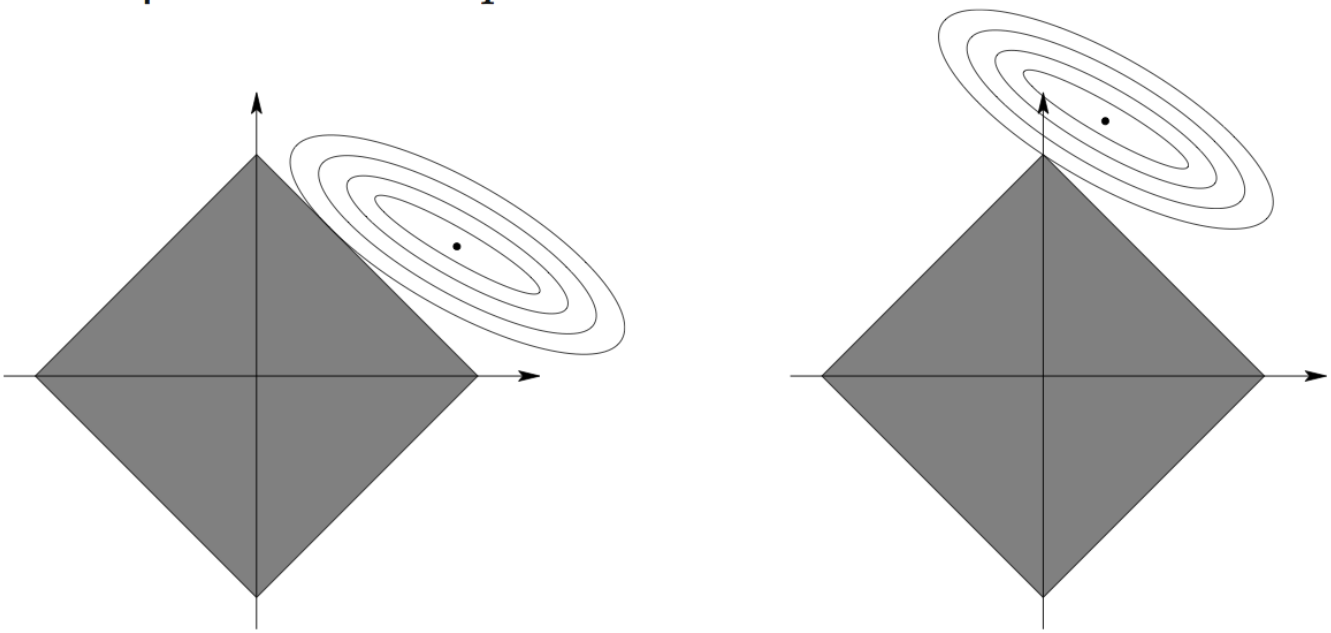
- "Increases bias and decreases variance"
- Common choices are
  - $\Omega(\beta) = \sum_{i=1}^{p} \beta_i^2$ (ridge regression, SVM, ...)
  - $\Omega(\beta) = \sum_{i=1}^{p} |\beta_i|$ (lasso, boosting, ...)

Further benefit: sparsity-inducing penalties

(Lasso)

$$\min_{\beta} R(\beta) \text{ s.t. } \sum_{i=1}^{p} |\beta_i| \leq C$$

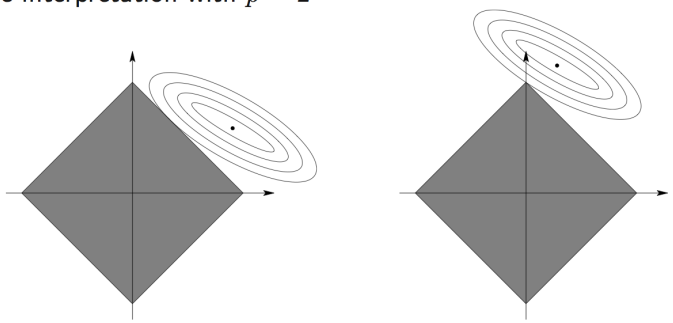
Geometric interpretation with $p = 2$



Summary

$$\min_{\beta} R(\beta) \text{ s.t. } \Omega(\beta) \leq C$$

Geometric interpretation with $p = 2$



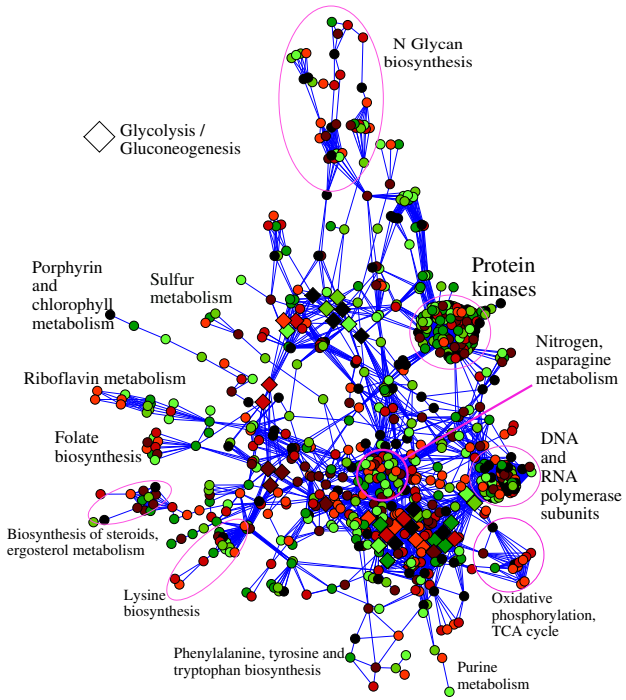
Shrinkage methods can:

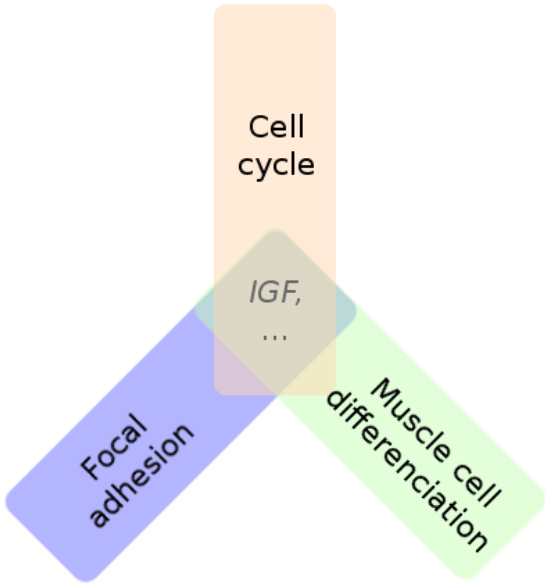
1. Improve the accuracy of the model by better controlling the bias/variance trade-off
2. Further decrease the bias by including prior knowledge in the penalty $\Omega(\beta)$
3. Perform feature selection with non-smooth penalties
4. Be efficiently implement with convex risk and penalty

Outline

Prior knowledge

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- We know these functional groups and gene networks





Shrinkage estimators for gene expression data

$$\min_{\beta} R(\beta) \text{ s.t. } \Omega(\beta) \leq C$$

How to design penalties $\Omega(\beta)$ to encode the following hypotheses:

1. Connected genes on a network should have similar weights (with or without gene selection)
2. Select few genes that are connected or belong to same predefined functional groups (without constraint on the weights)

Hypothesis 1: connected genes on a network should have similar weights

- Smooth weights on the graph (or more generally graph kernels)

$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2$$

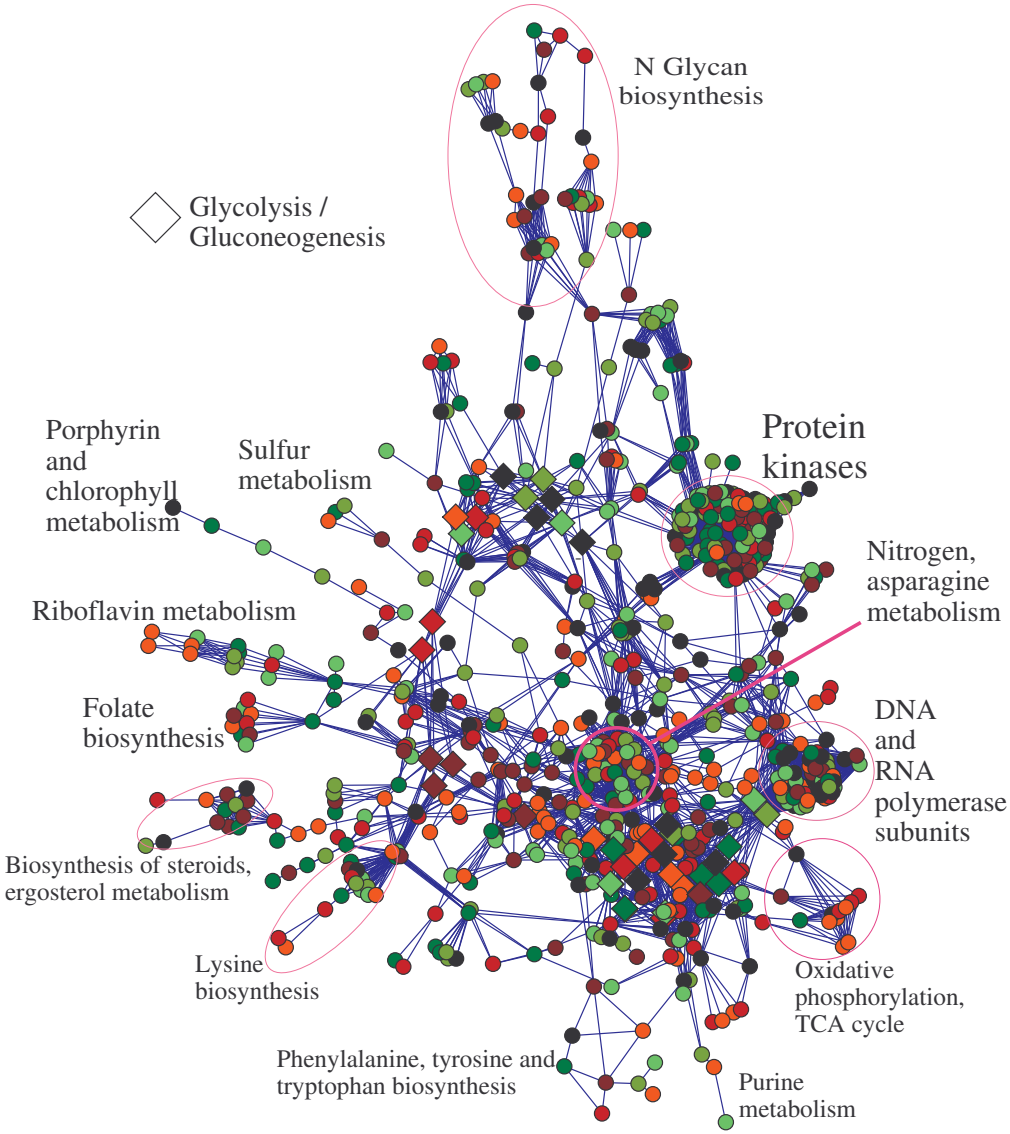
- Gene selection + smooth on the graph

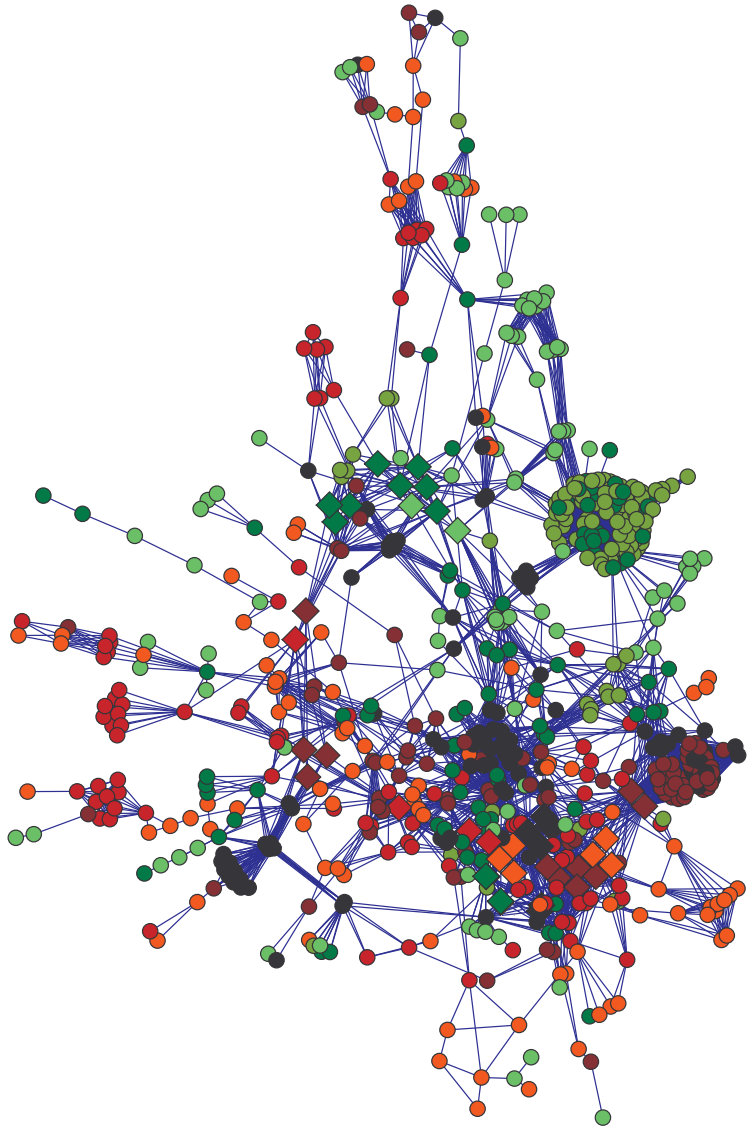
$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\beta_i|$$

- Gene selection + Piecewise constant on the graph (total variation)

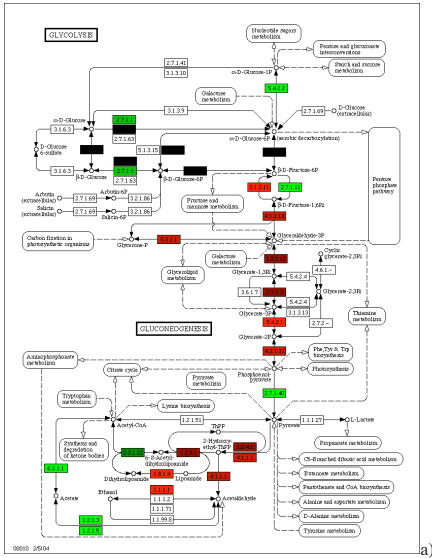
$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i|$$

Illustration

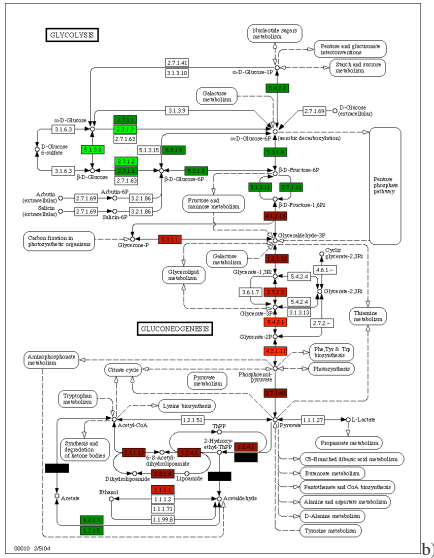




Limits



a)



b)

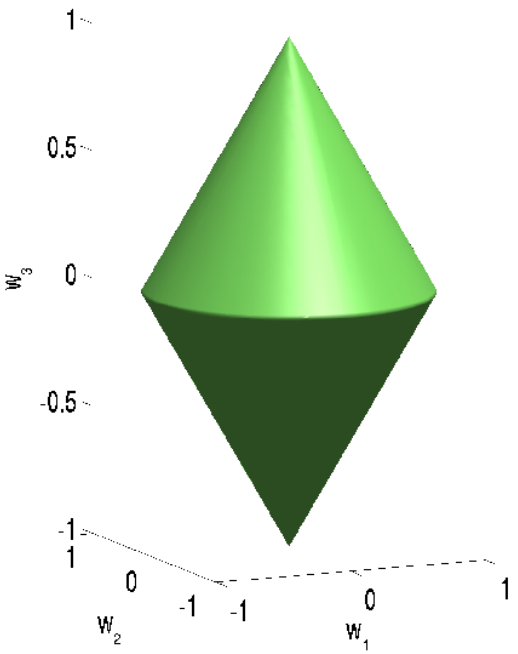
- We are happy to see pathways appear.
- However, in some cases, connected genes should have "opposite" weights (inhibition, pathway branching, etc...)
- How to capture pathways without constraints on the weight similarities?

Selecting pre-defined groups of variables

Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the ℓ_1/ℓ_2-norm induces sparse solutions *at the group level*:

$$\Omega_{group}(\beta) = \sum_g \|\beta_g\|_2$$



Groups $\{1,2\}$ and $\{3\}$:

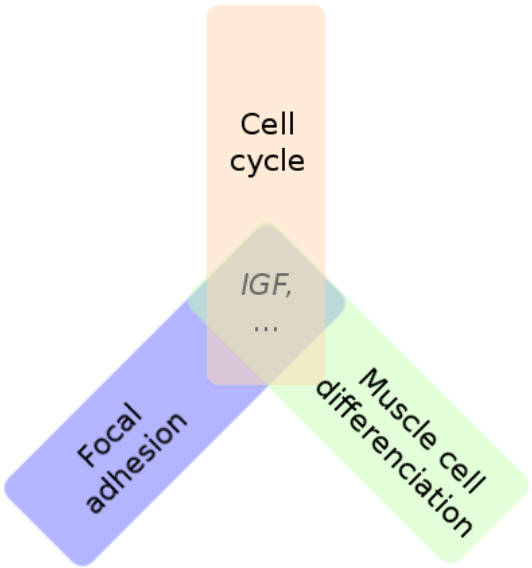
$$\Omega_{group}(\beta_1,\beta_2,\beta_3) = \|(\beta_1,\beta_2)\|_2 + \|\beta_3\|_2$$
$$= \sqrt{\beta_1^2+\beta_2^2} + |\beta_3|$$

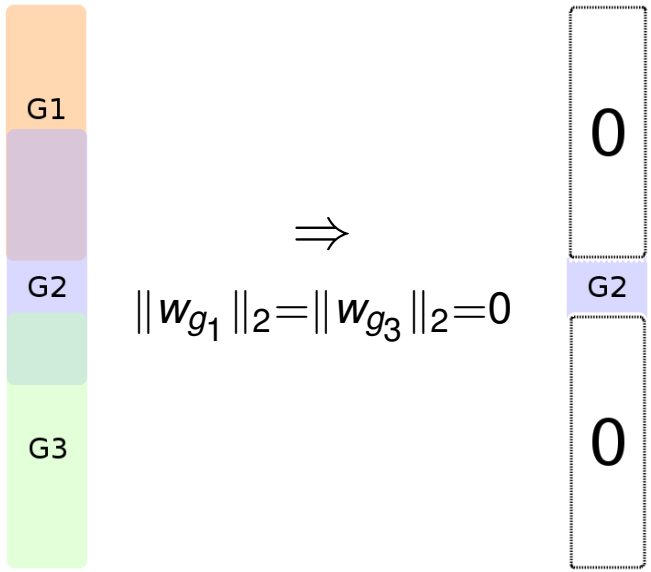
Group Lasso when groups overlap

When groups overlap, the group Lasso

$$\Omega_{group}(\beta) = \sum_g \| \beta_g \|$$

sets groups to 0 $\Longrightarrow$ the support of the solution is the complement of a union of groups





IGF selection $\Rightarrow$ selection of unwanted groups

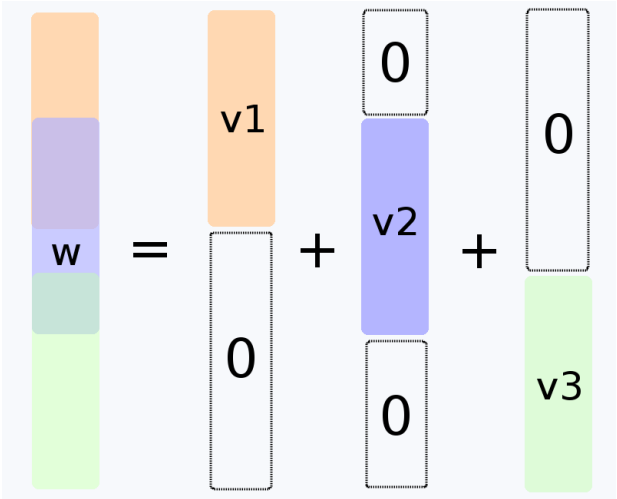
Removal of *any* group containing a gene $\Rightarrow$ the weight of the gene is 0.

The latent group Lasso (Jacob et al., 2009)

$$\Omega_{latent}(\beta) = \sup_{\alpha \in \mathbb{R}^p \,:\, \forall g, \| \alpha_g \| \leq 1} \alpha^\top \beta$$

or, equivalently:

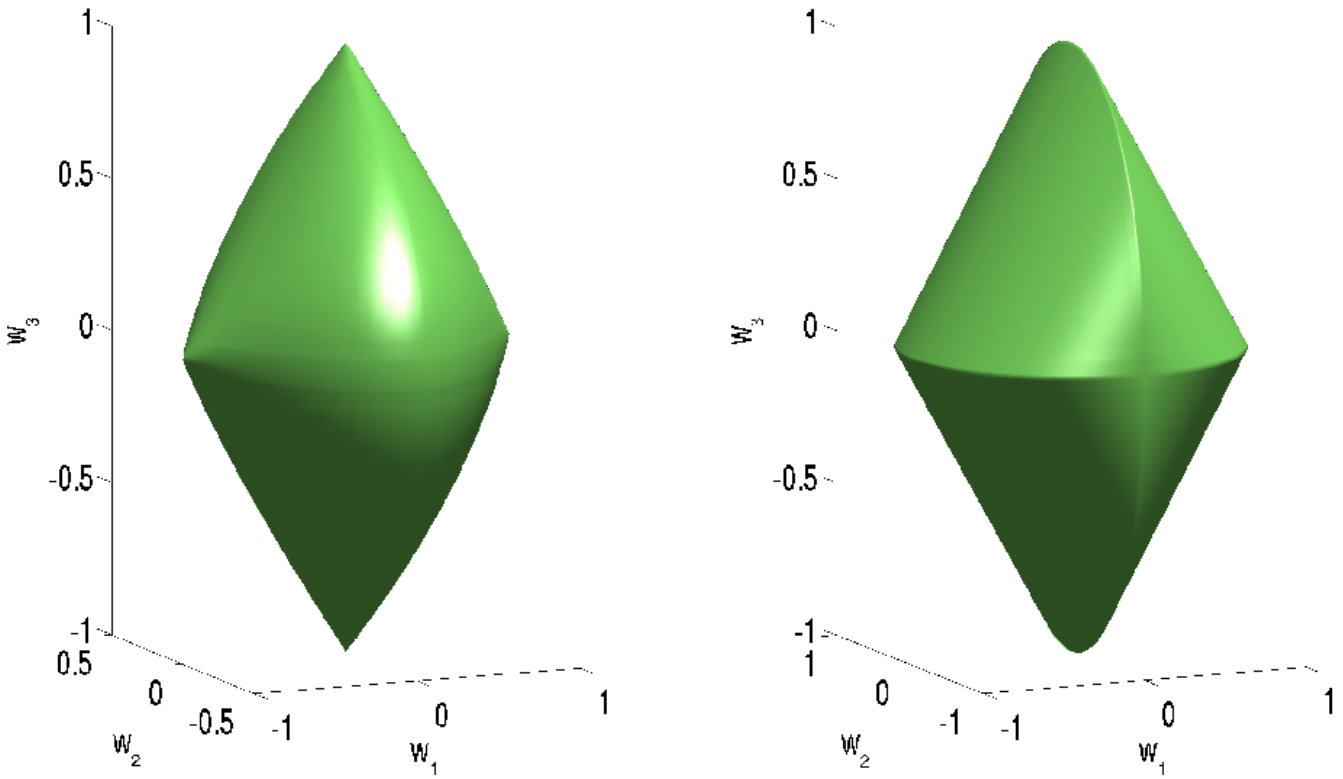
$$\Omega_{latent}(\beta) \triangleq \begin{cases} \min_v \sum_{g \in \mathcal{G}} \| v_g \|_2 \\ \beta = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases}$$



Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.
- Equivalent to group lasso when there is no overlap

Group Lasso vs latent group Lasso



Balls for $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ (middle) and $\Omega_{\textit{latent}}\cdot$ (right) for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ where w_2 is represented as the vertical coordinate.

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

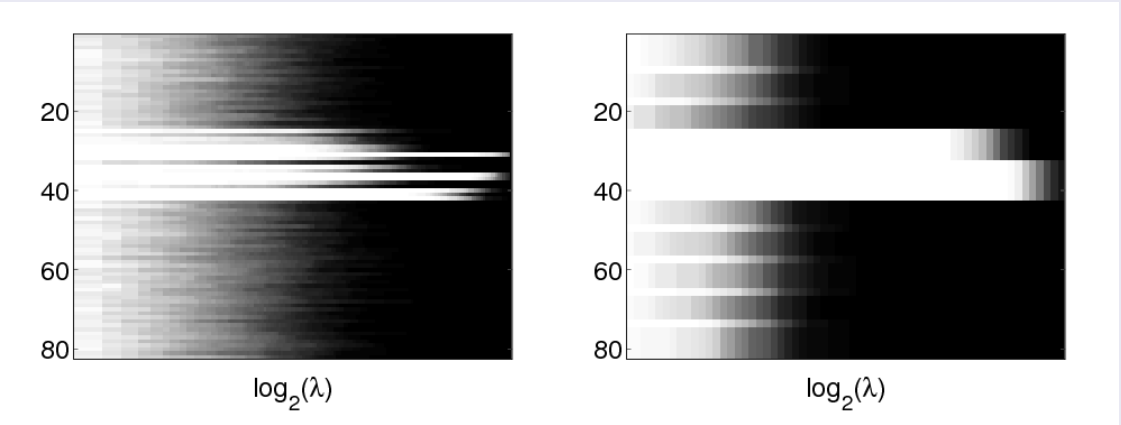
the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

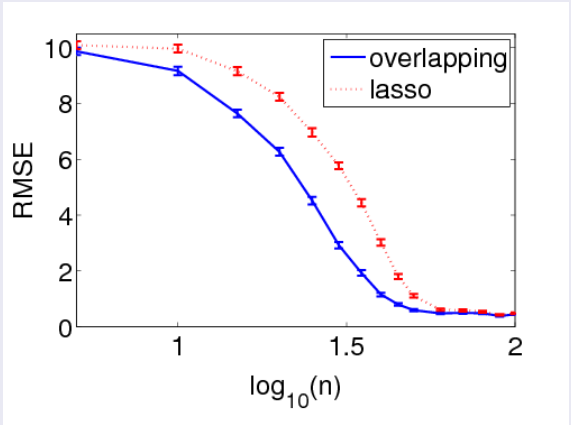
$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

Experiments

Synthetic data: overlapping groups

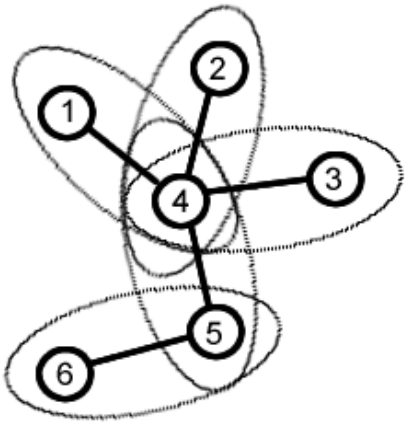
- 10 groups of 10 variables with 2 variables of overlap between two successive groups :$\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$.
- Support: union of $4th$ and $5th$ groups.
- Learn from 100 training points.





Frequency of selection of each variable with the lasso (left) and Ω_{latent} (.) (middle), comparison of the RMSE of both methods (right).

Graph lasso vs kernel on graph



- Graph lasso:

$$\Omega_{group}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2} \quad \text{or} \quad \Omega_{latent}(\beta) = \sup_{\alpha \in \mathbb{R}^p \,:\, \forall i \sim j, \sqrt{\alpha_i^2 + \alpha_j^2} \leq 1} \alpha^\top \beta$$

  constrains the sparsity, not the values

- Graph kernel

$$\Omega_{\text{graph kernel}}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \,.$$

  constrains the values (smoothness), not the sparsity

Preliminary results

Breast cancer data

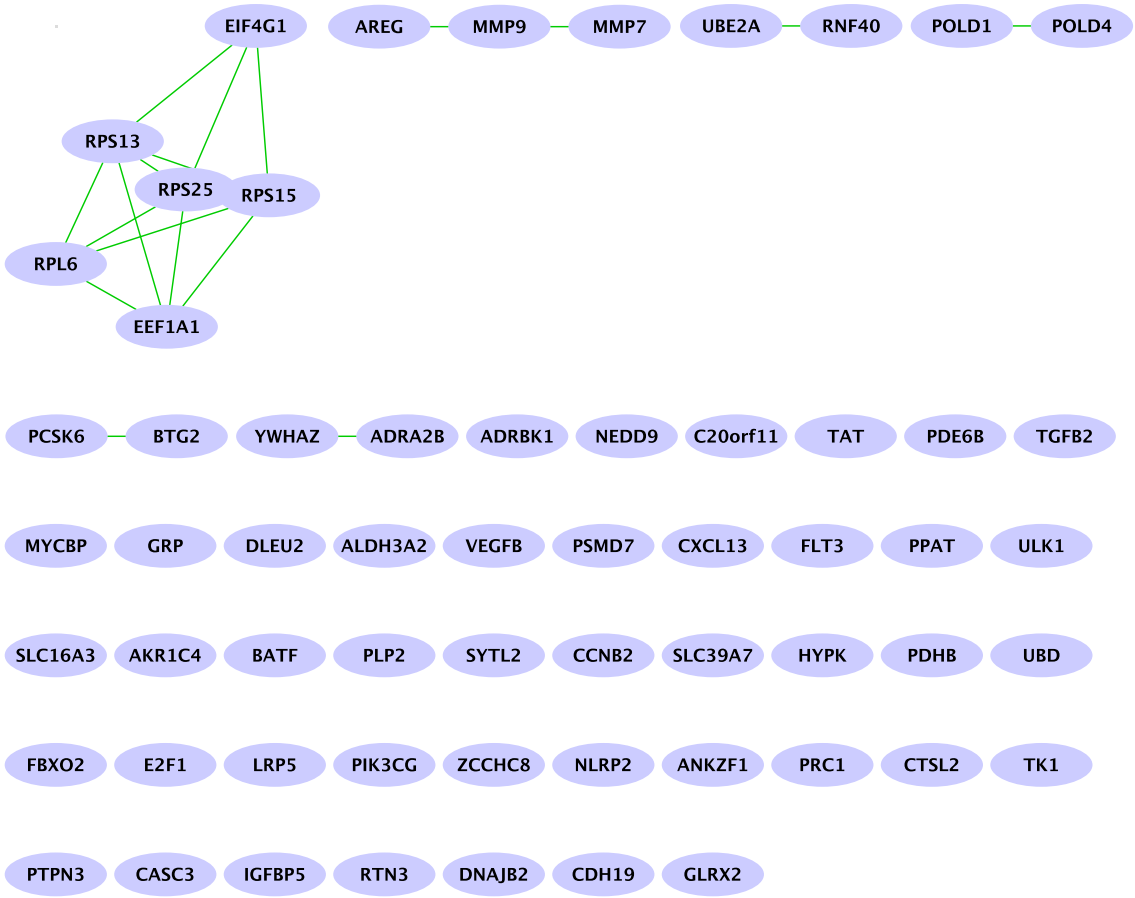
- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

METHOD	ℓ_1	$\Omega_{\text{LATENT}}(.)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
MEAN ♯ PATH.	130	30

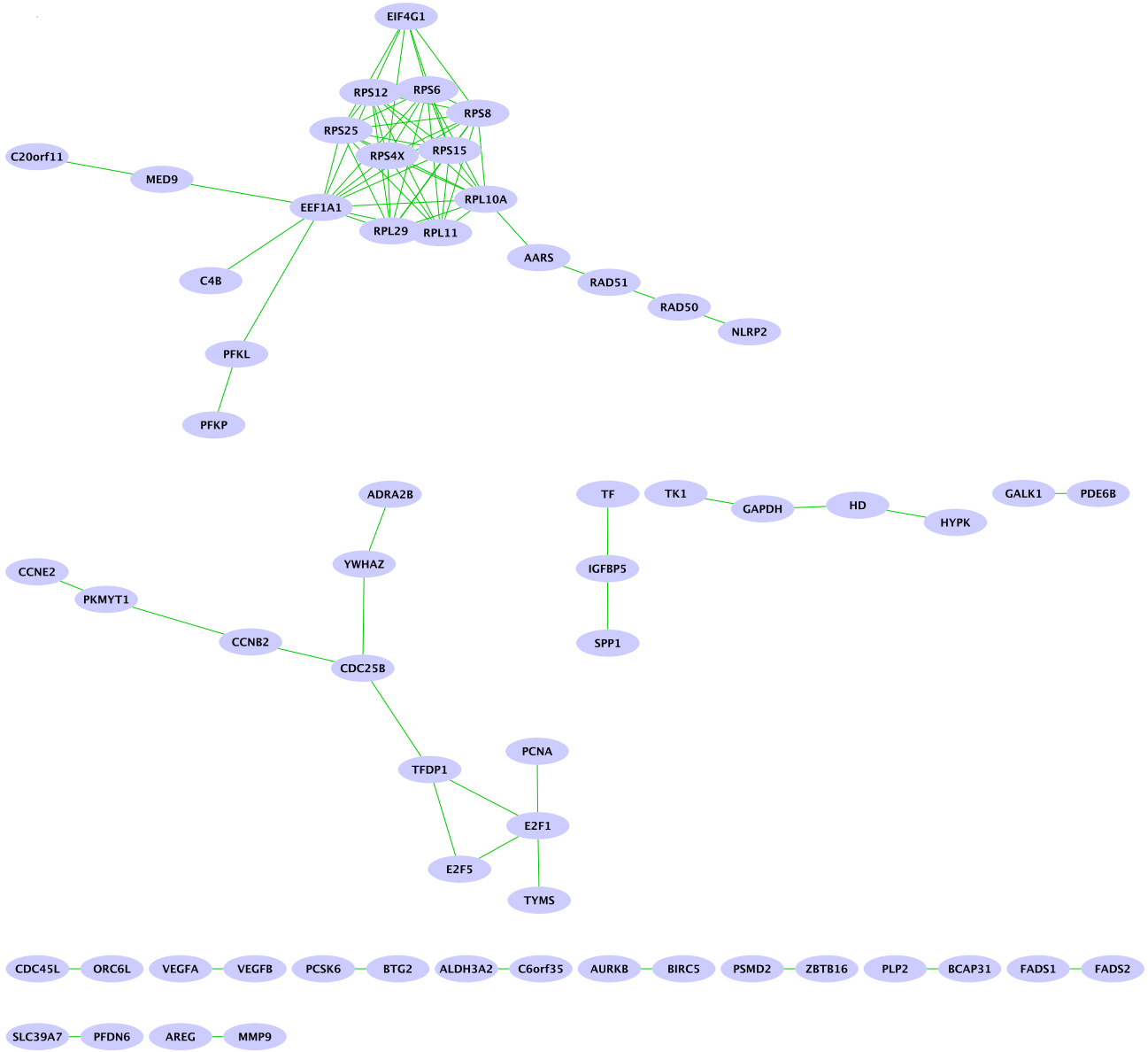
- Graph on the genes.

METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01
AV. SIZE C.C.	1.03	1.30

Classical lasso signature



Graph Lasso signature

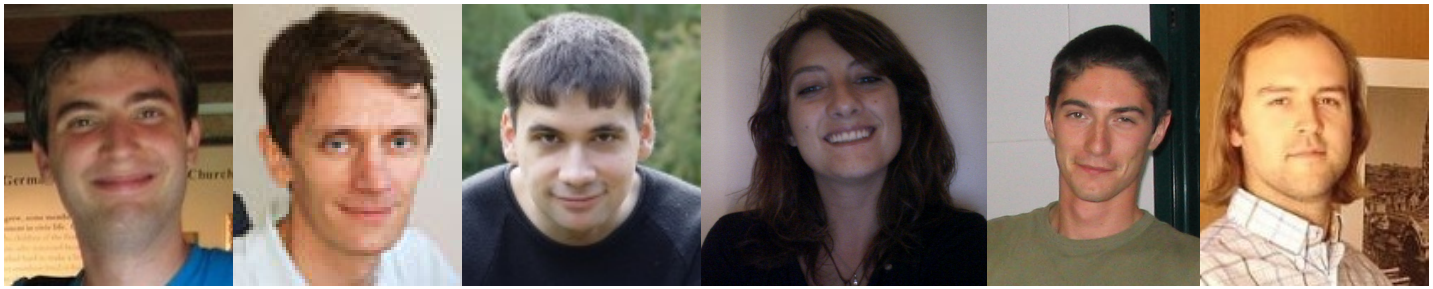


Outline

Conclusion

- Integration of prior knowledge in the penalization / regularization function is an efficient approach to fight the curse of dimension
- Structured sparsity can be obtained with particular non-smooth convex penalties
- How to include more knowledge, e.g., dynamics of the systems?

Acknowledgements!



Franck Rapaport (MSKCC), Emmanuel Barillot, Andrei Zynoviev, Anne-Claire Haury, Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA)