Machine learning in cancer genomics

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

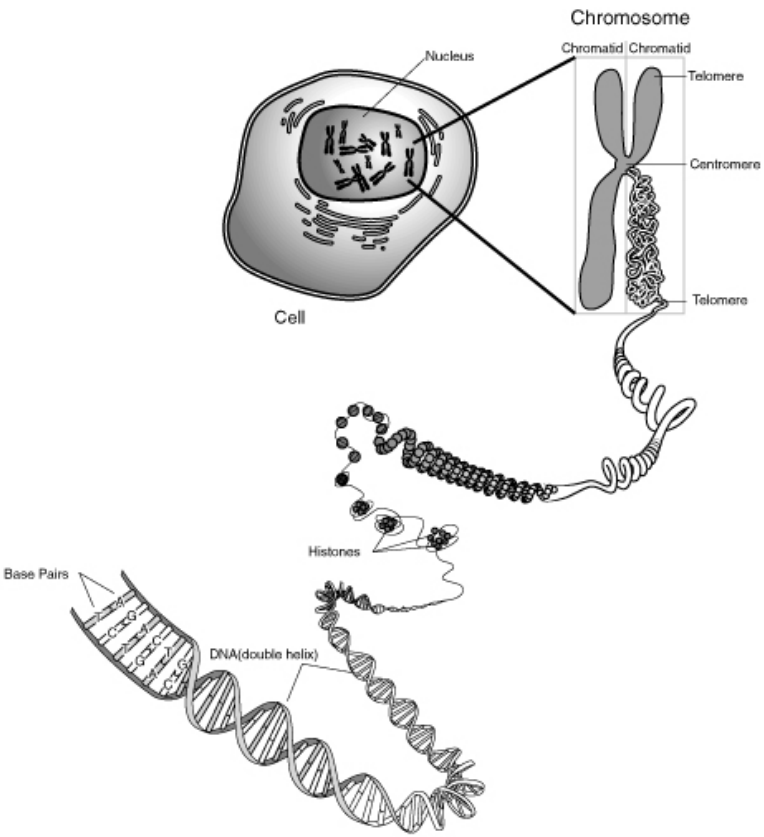
Mines ParisTech / Curie Institute / Inserm

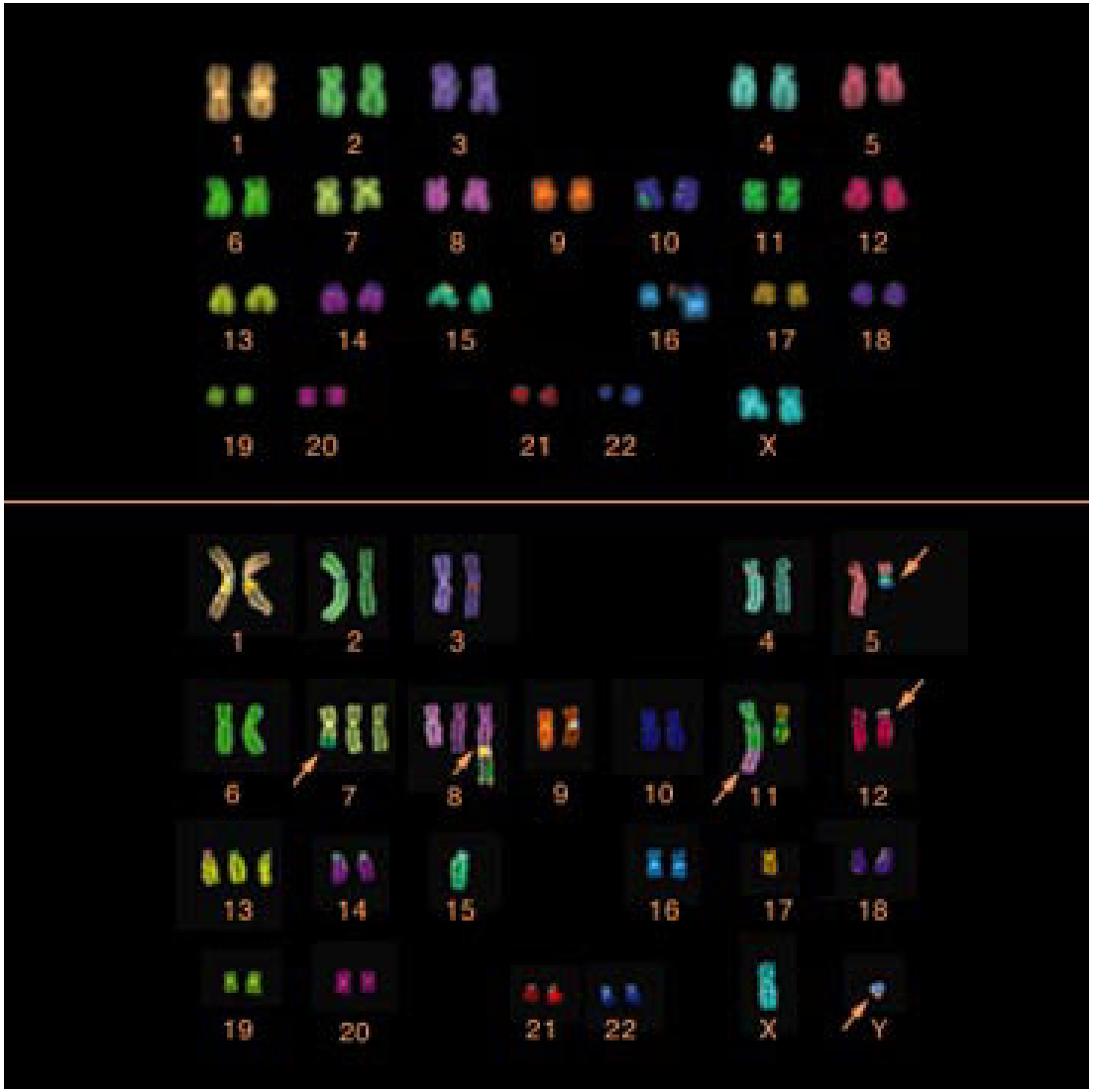
Université catholique de Louvain, Nov 14, 2011.

Outline

Outline

Chromosomic aberrations in cancer

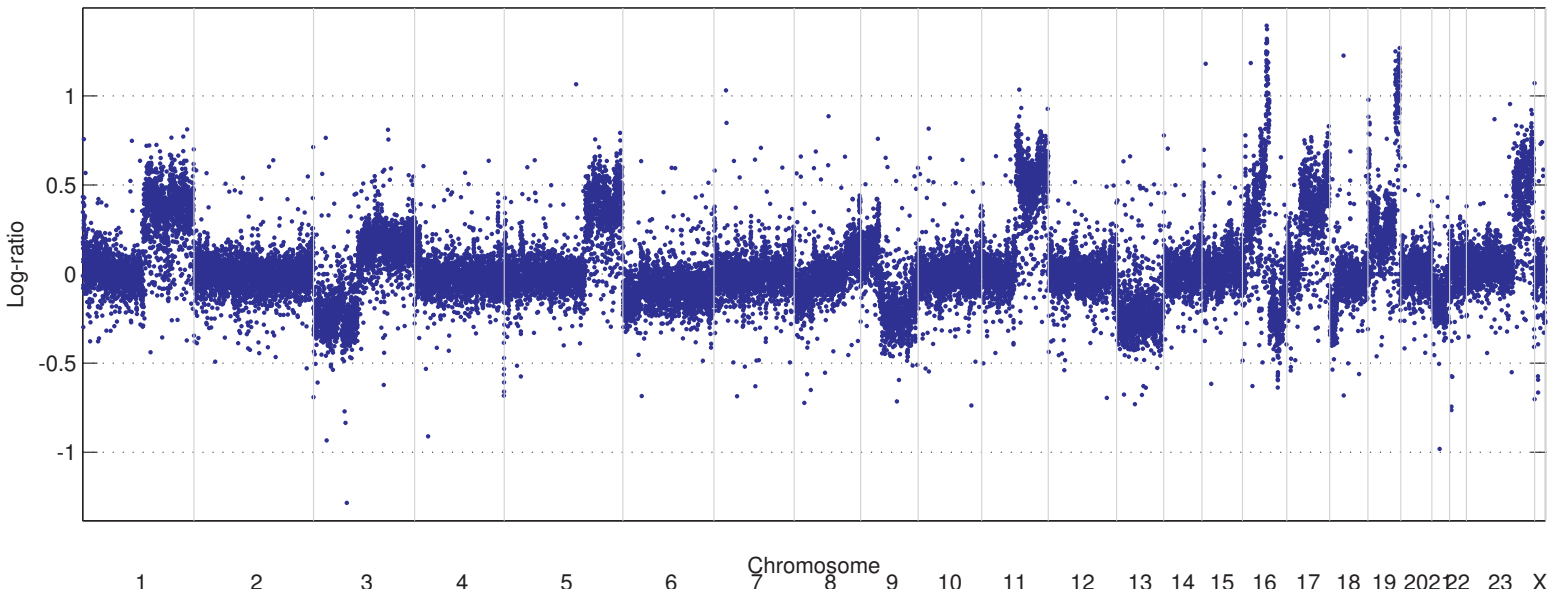




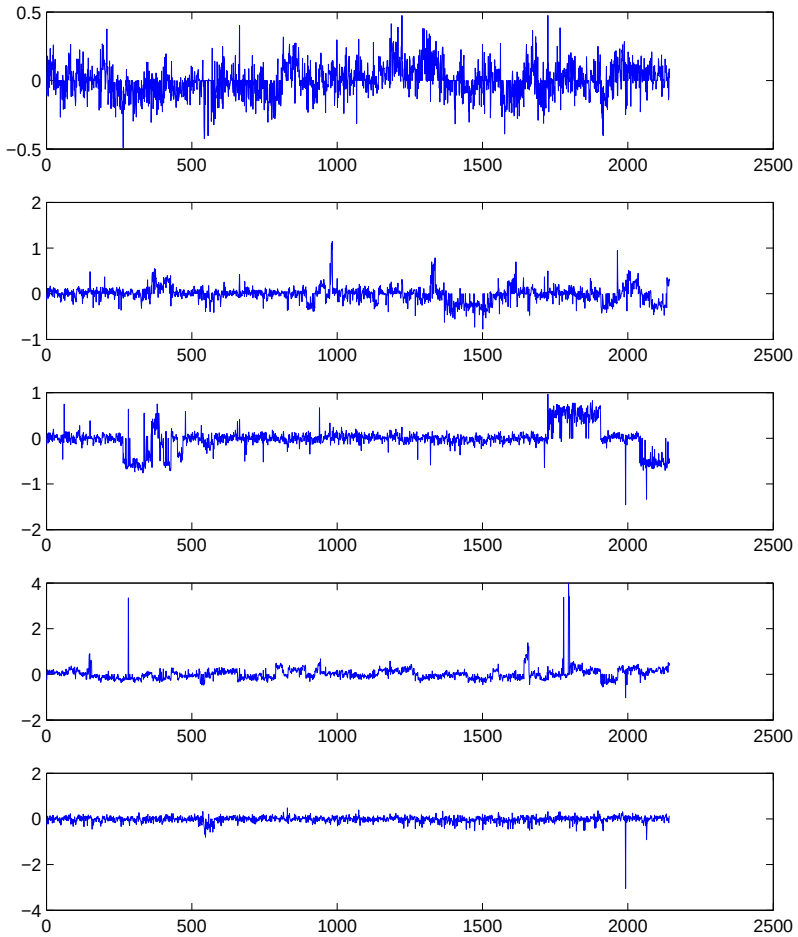
Comparative Genomic Hybridization (CGH)

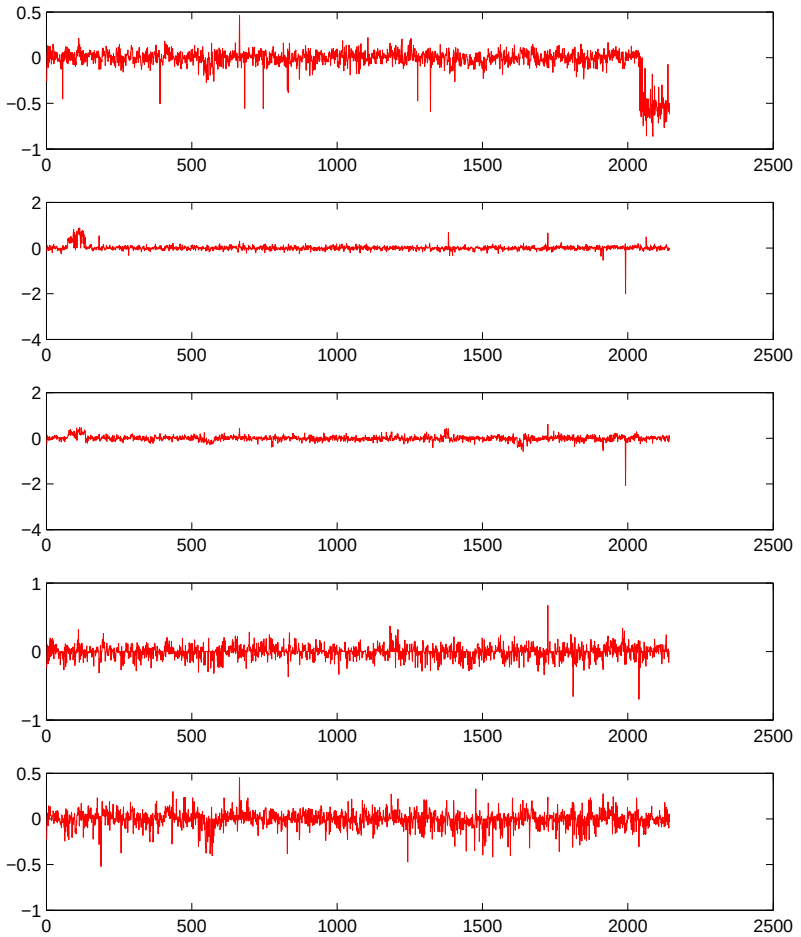
Motivation

- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research to observe systematically variants in DNA content



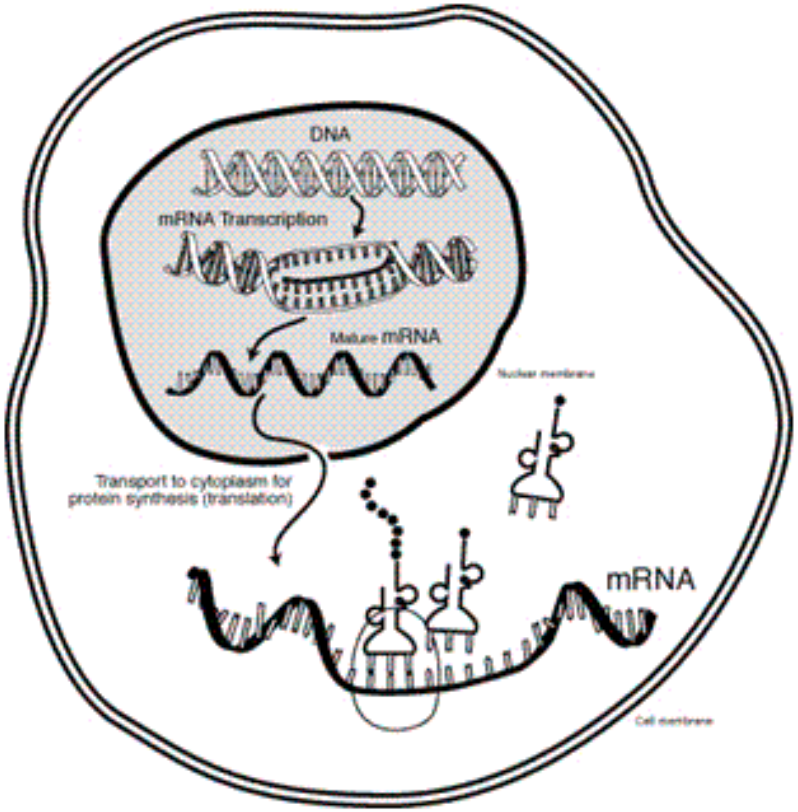
Cancer prognosis: can we predict the future evolution?

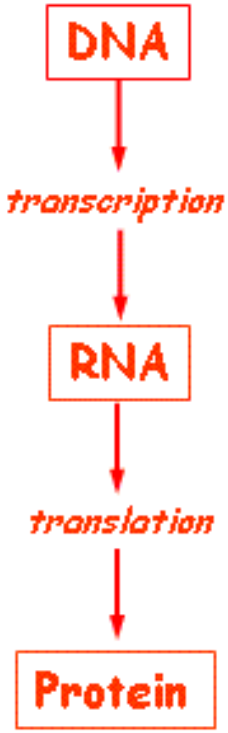




Aggressive (left) vs non-aggressive (right) melanoma

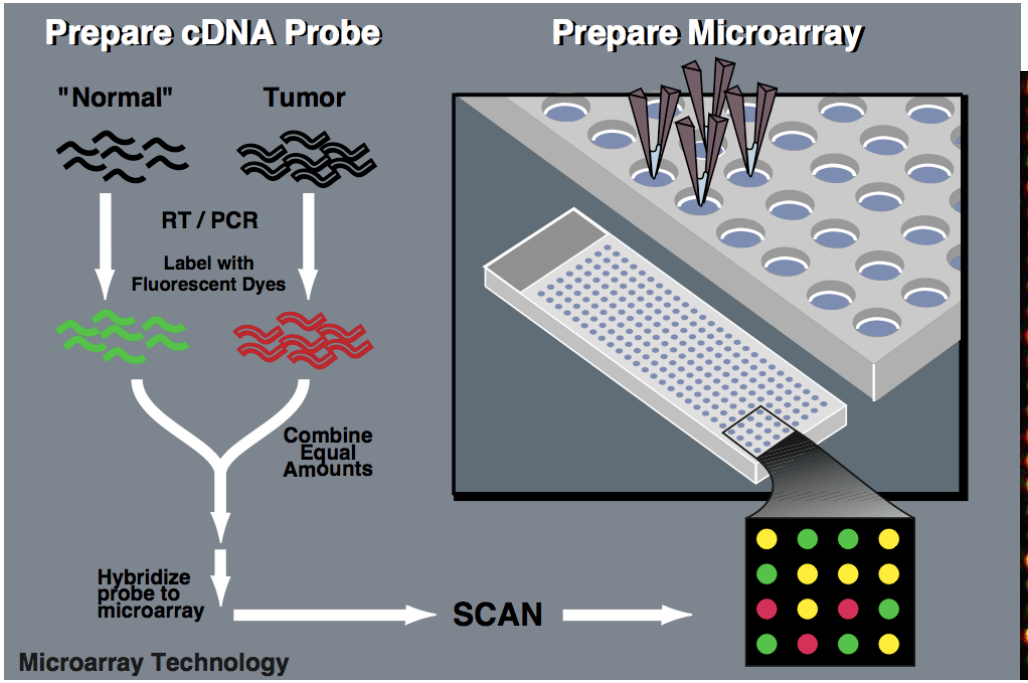
DNA $\rightarrow$ RNA $\rightarrow$ protein

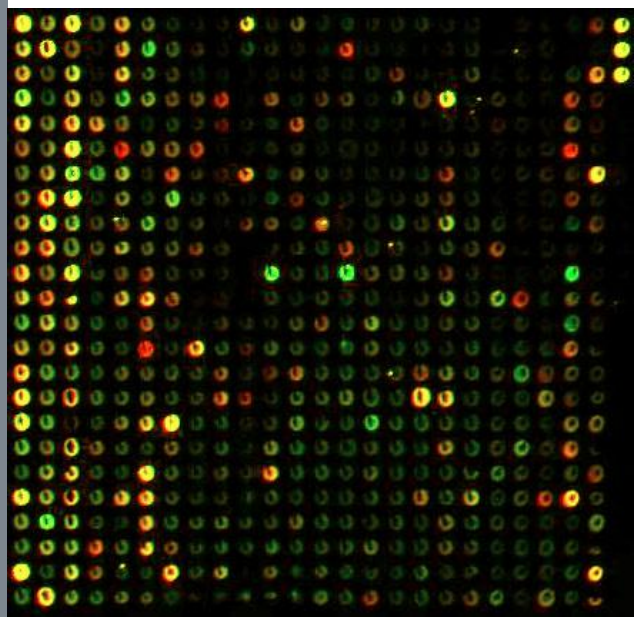




- CGH shows the (static) DNA
- Cancer cells have also abnormal (dynamic) gene expression (= transcription)

Tissue profiling with DNA chips

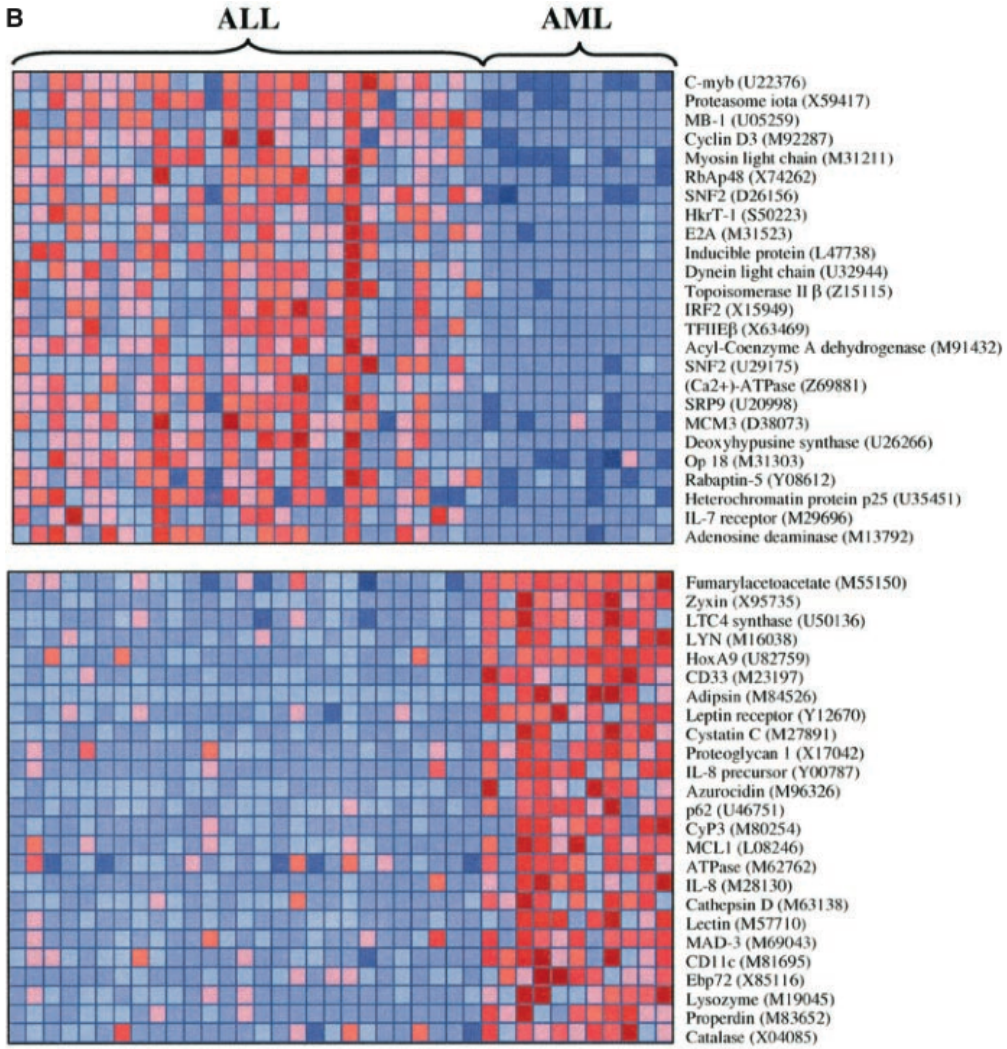




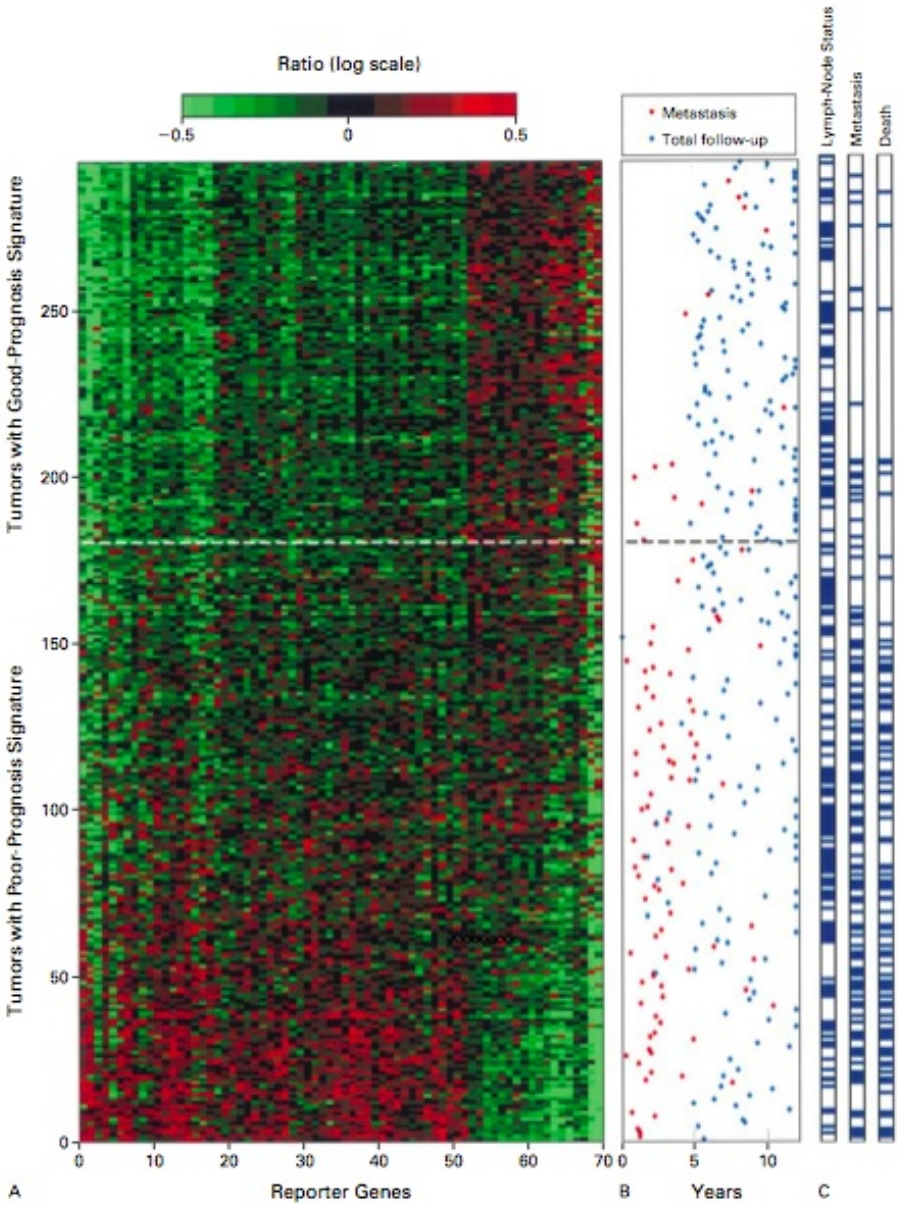
Data

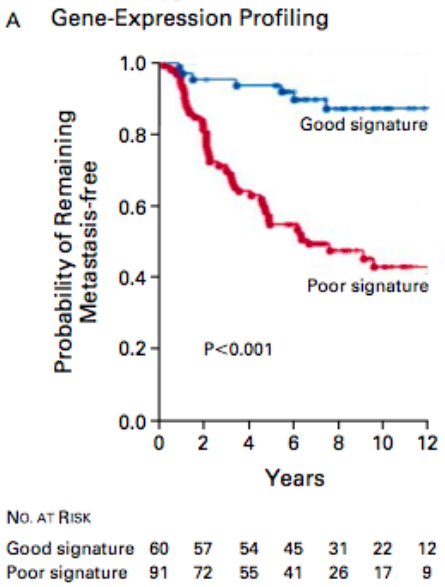
- Gene expression measures for more than $10k$ genes
- Measured typically on less than 100 samples of two (or more) different classes (e.g., different tumors)

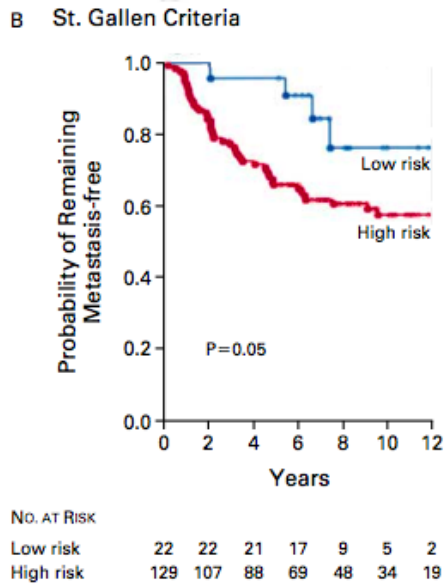
Can we identify the cancer subtype? (diagnosis)



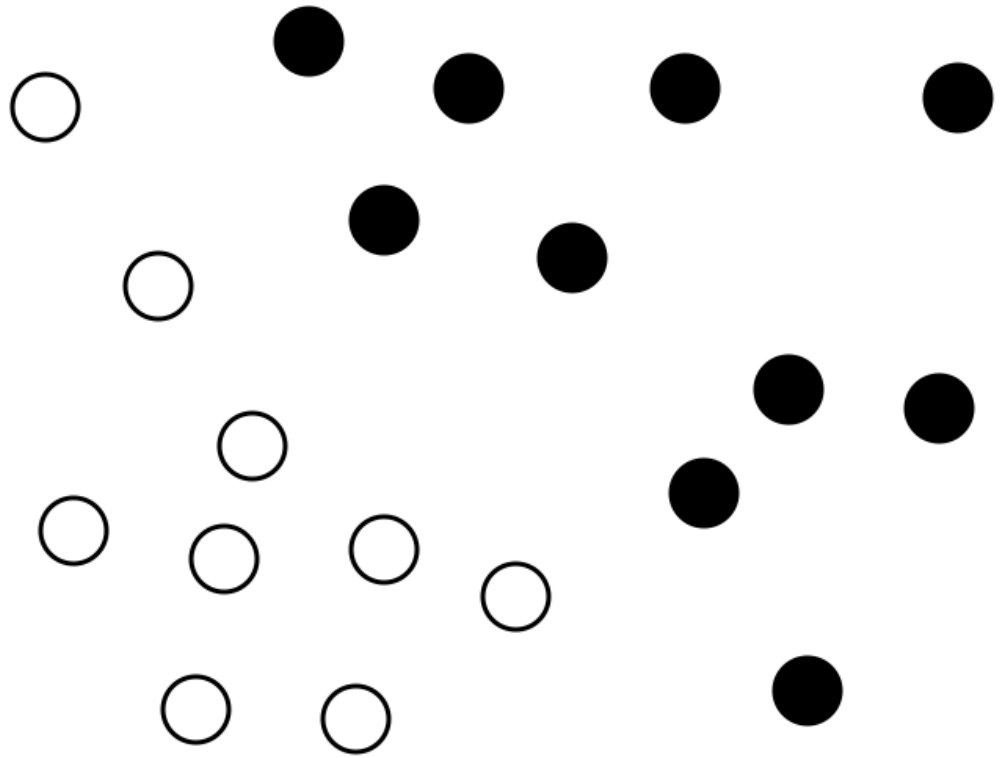
Can we predict the future evolution? (prognosis)





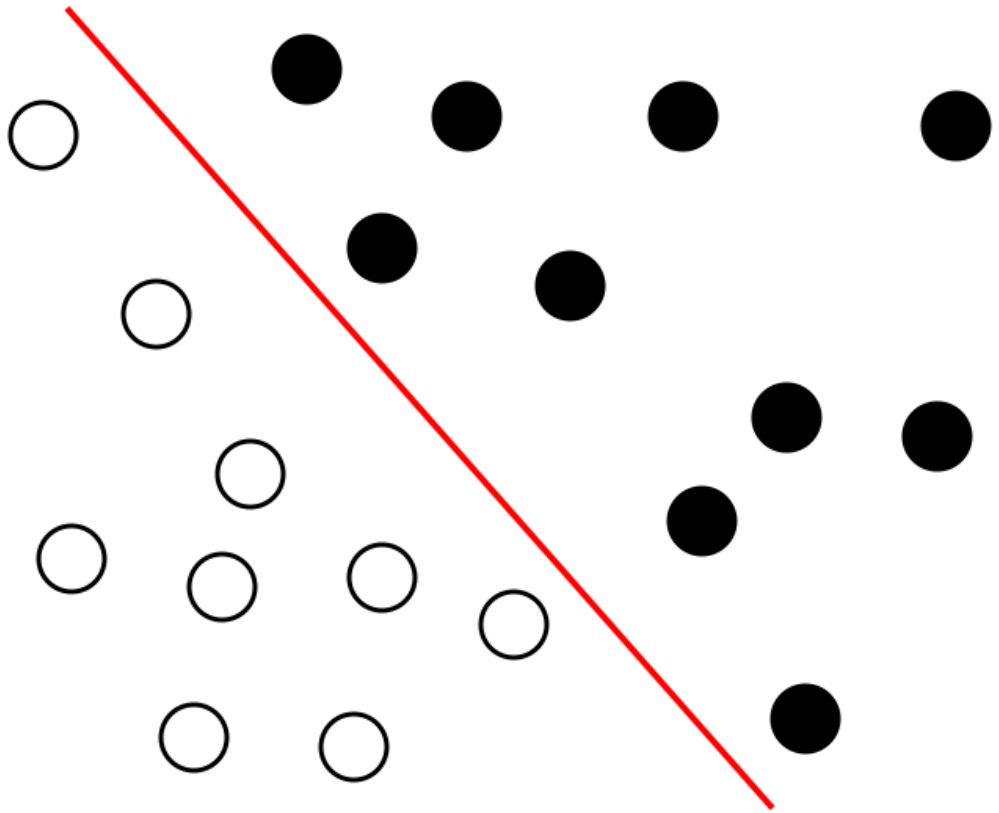


Machine learning (pattern recognition / supervised classification)



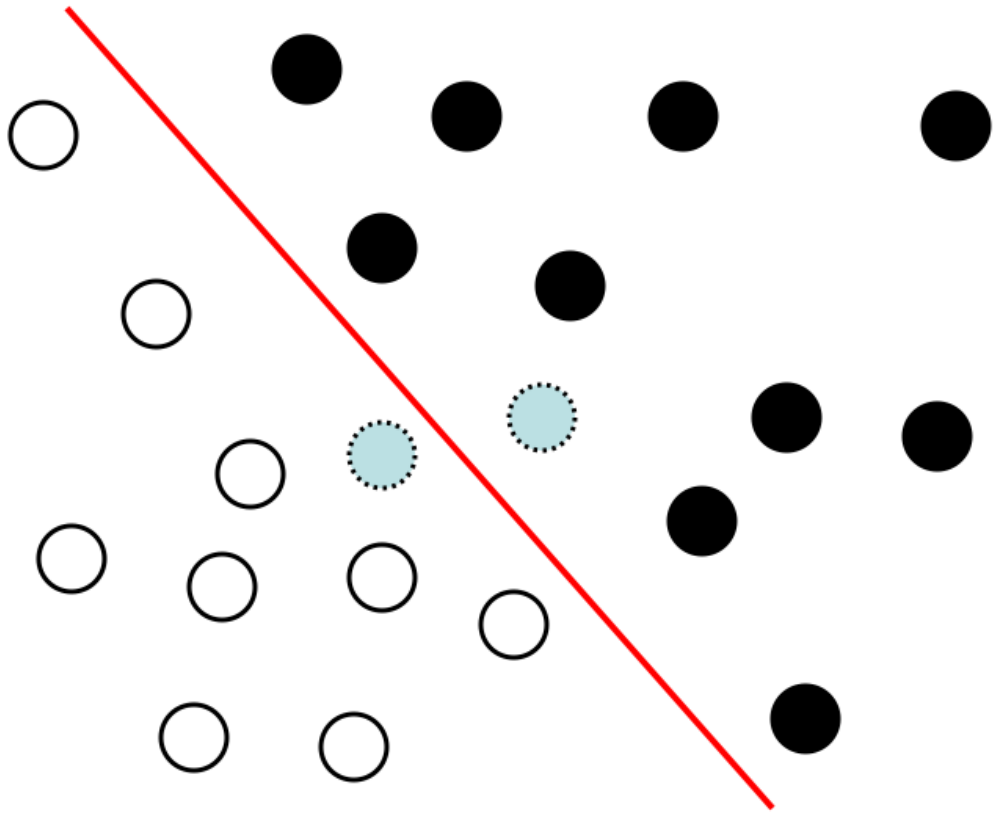
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning (pattern recognition / supervised classification)



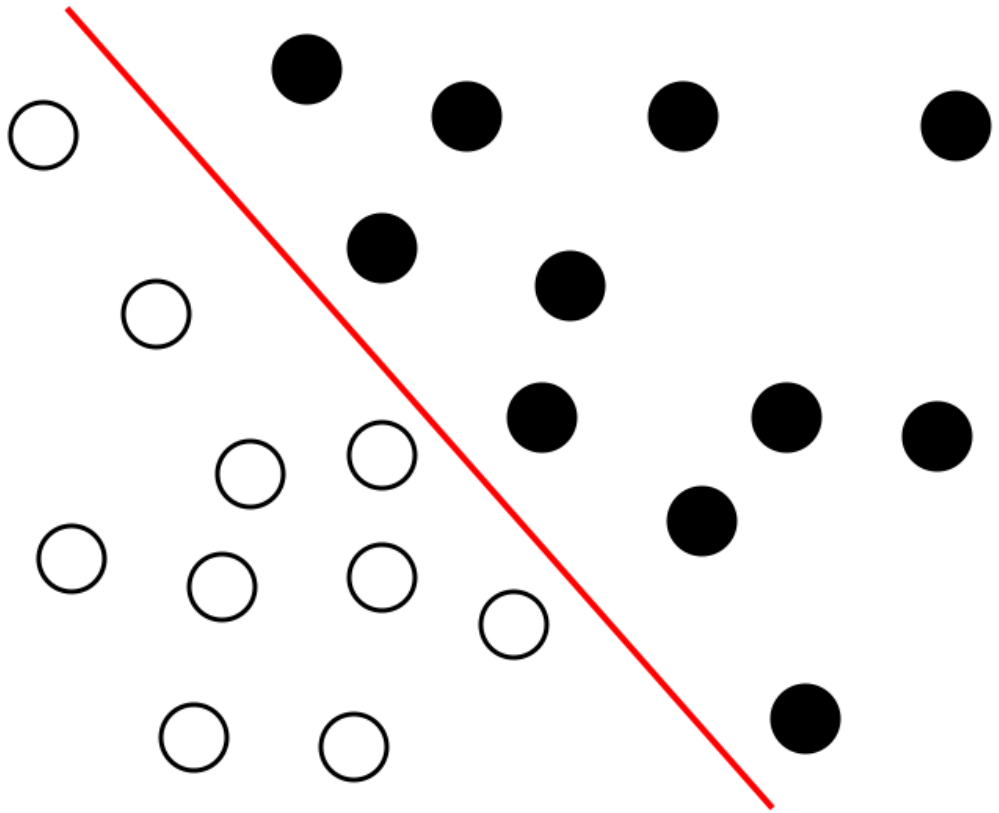
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning (pattern recognition / supervised classification)



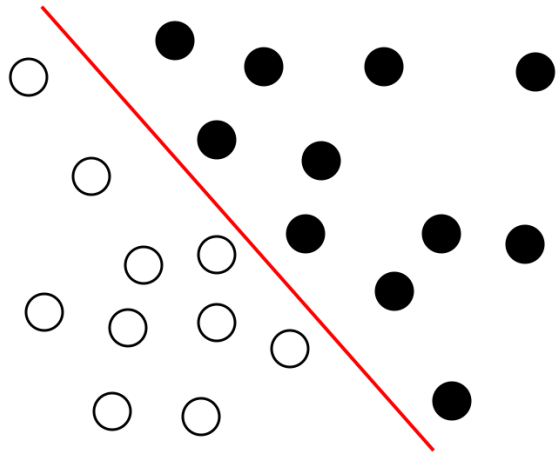
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning (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

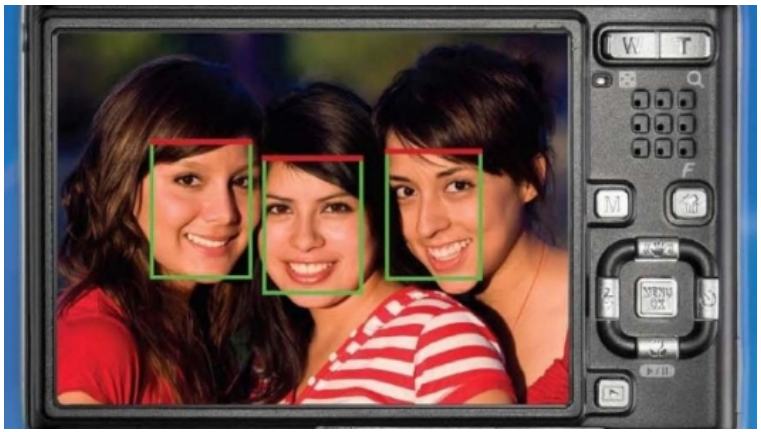


Genome annotation, systems biology, personalized medicine...

Challenges

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

Machine learning : tools and applications



Many applications

Multimedia, image, video, speech recognition, web, social network, online advertising, finance, biology, chemistry

Many tools

Linear discriminant analysis, logistic regression, decision trees, neural networks, support vector machines...

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.:

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

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.:

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

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

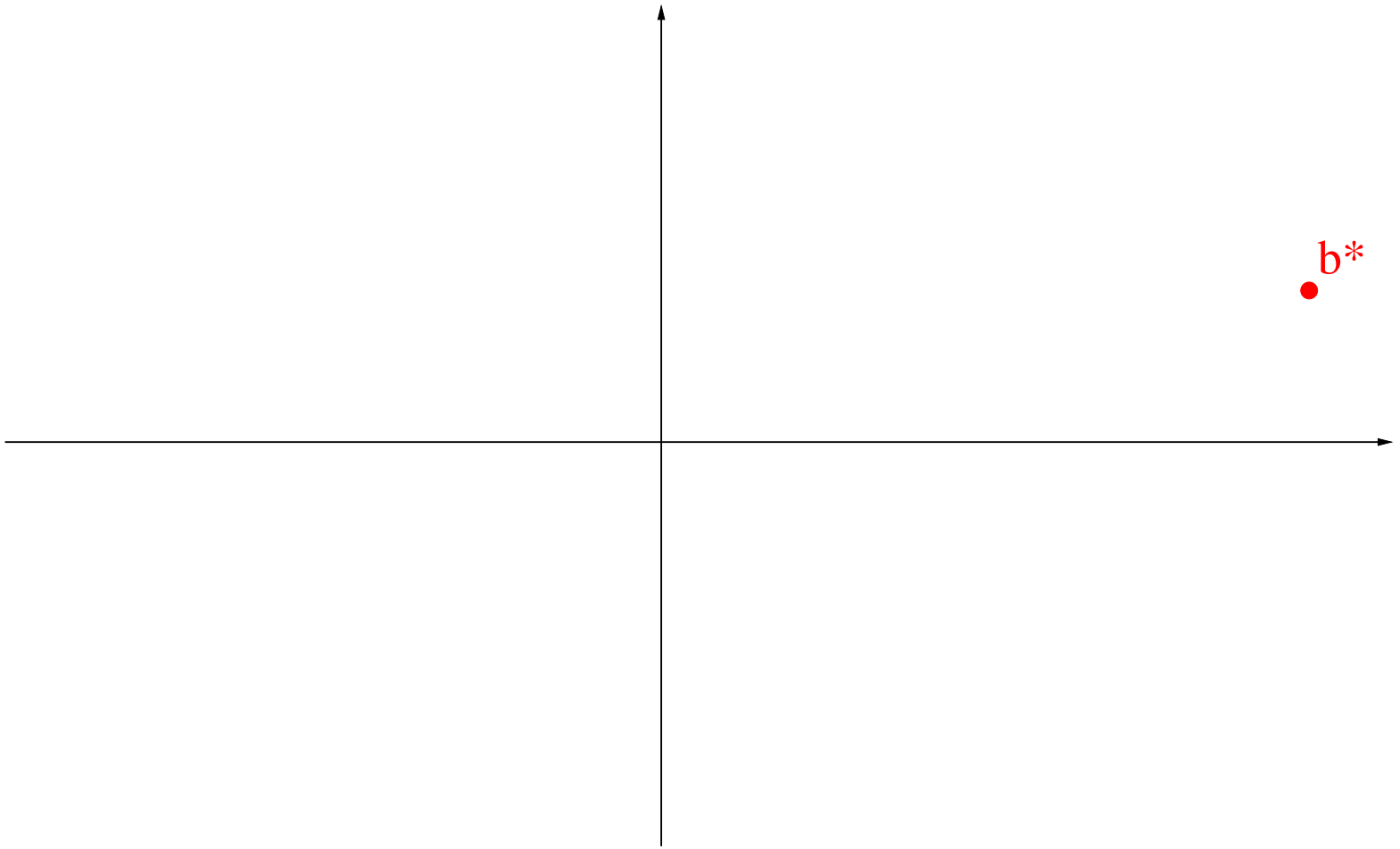
1. Define a large family of "candidate classifiers", e.g., linear predictors:
$$f_\beta(x) = \beta^\top x \quad \text{for } x \in \mathbb{R}^p$$

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

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

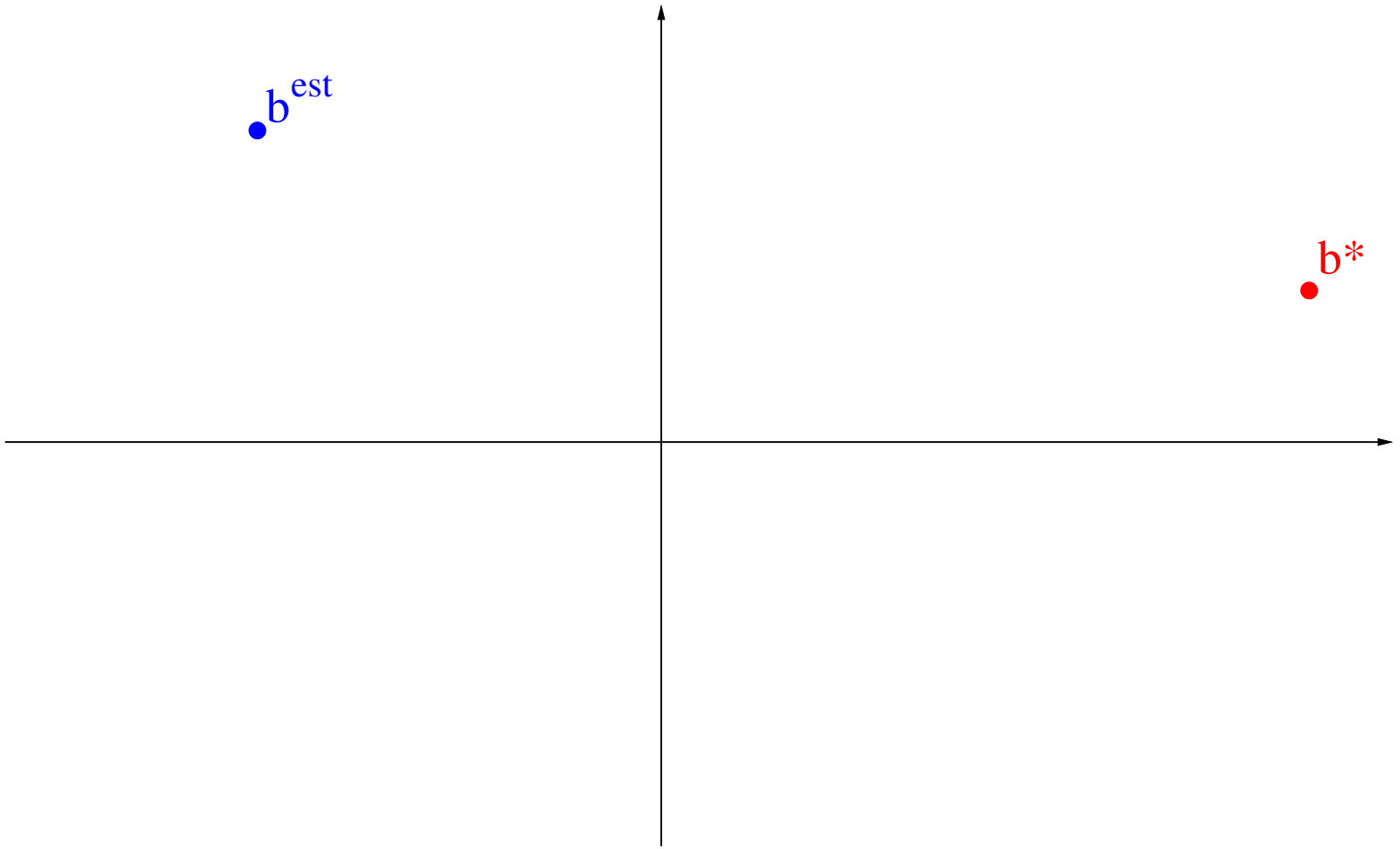
Why skrinkage classifiers?

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



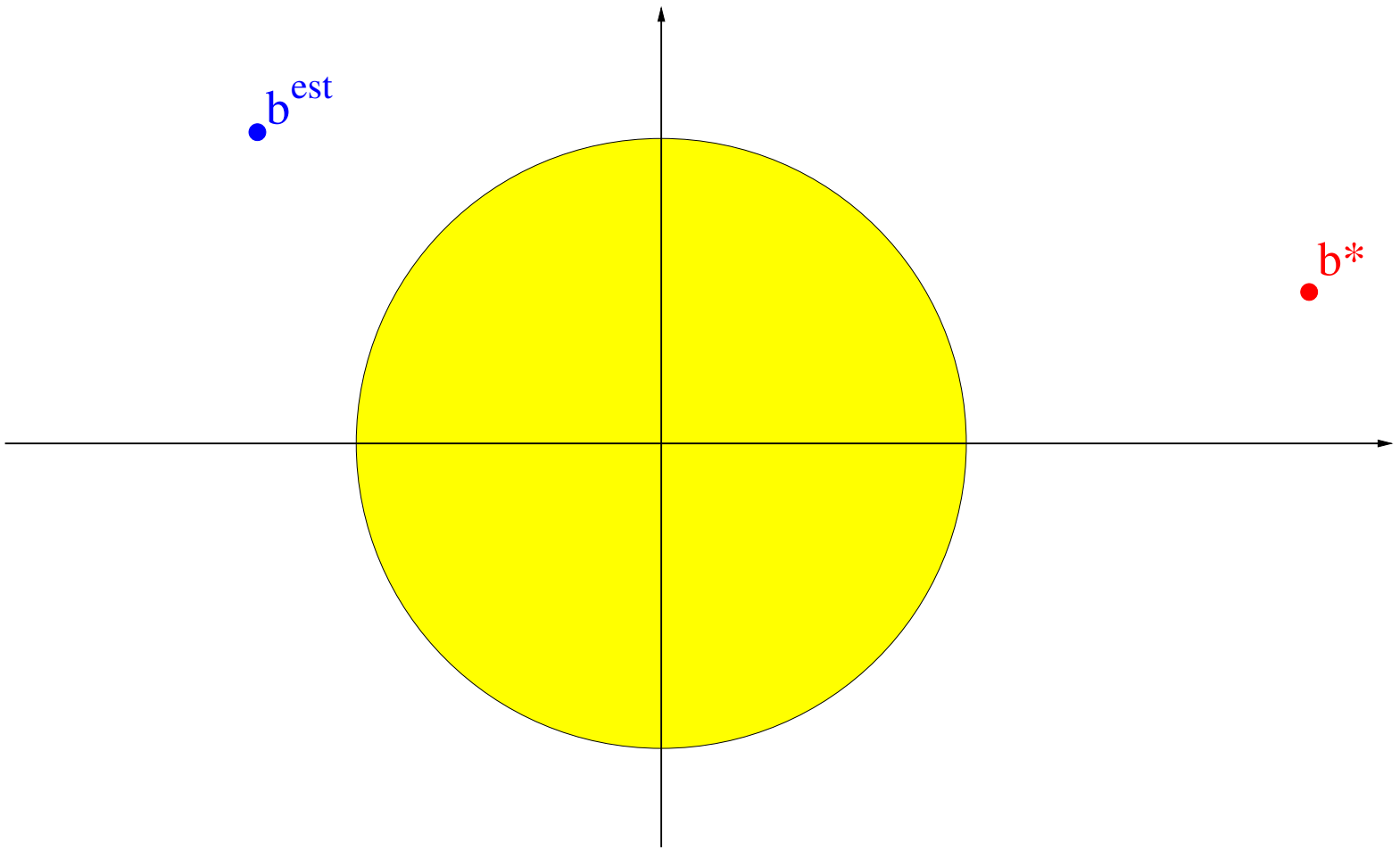
Why skrinkage classifiers?

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



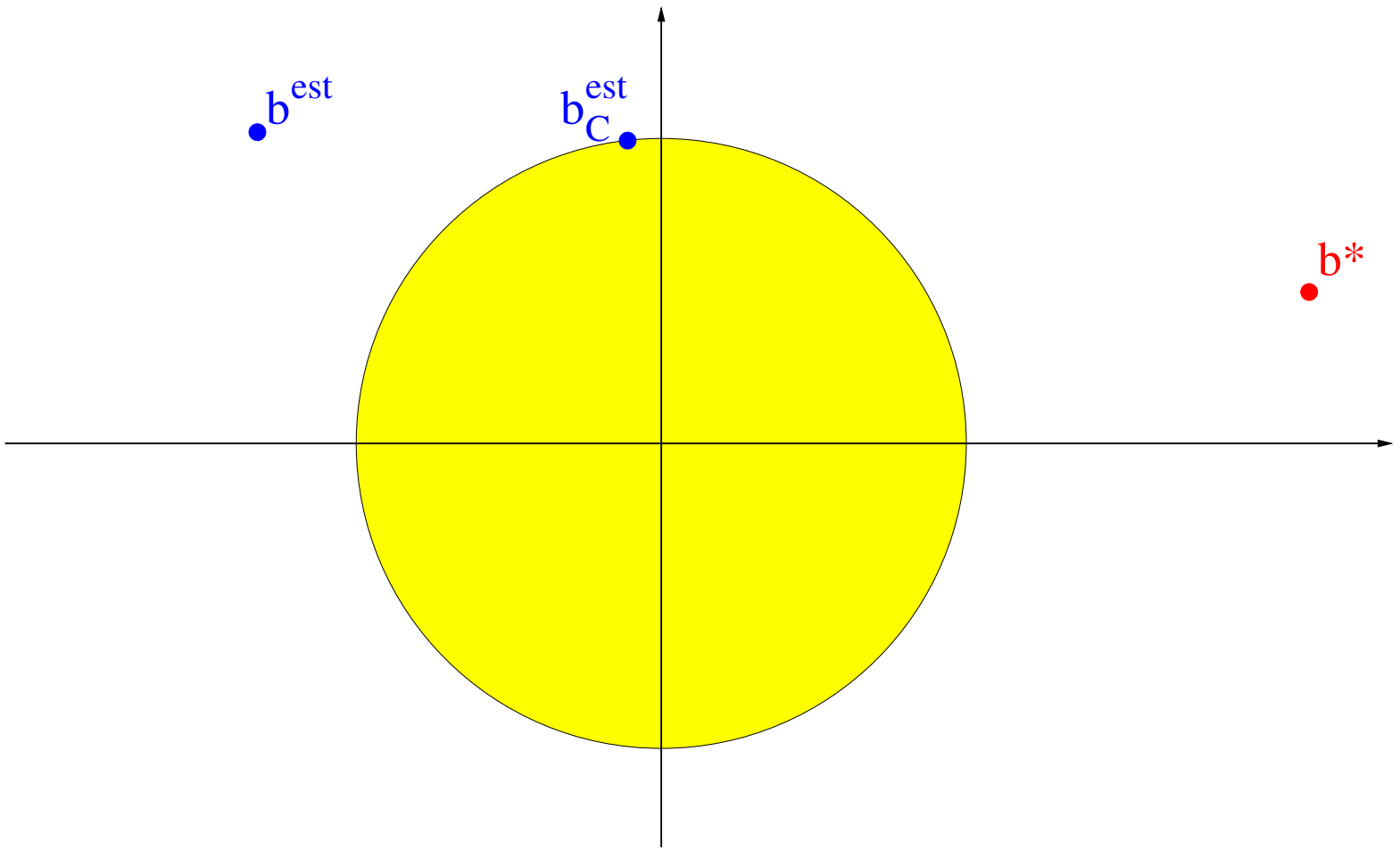
Why skrinkage classifiers?

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



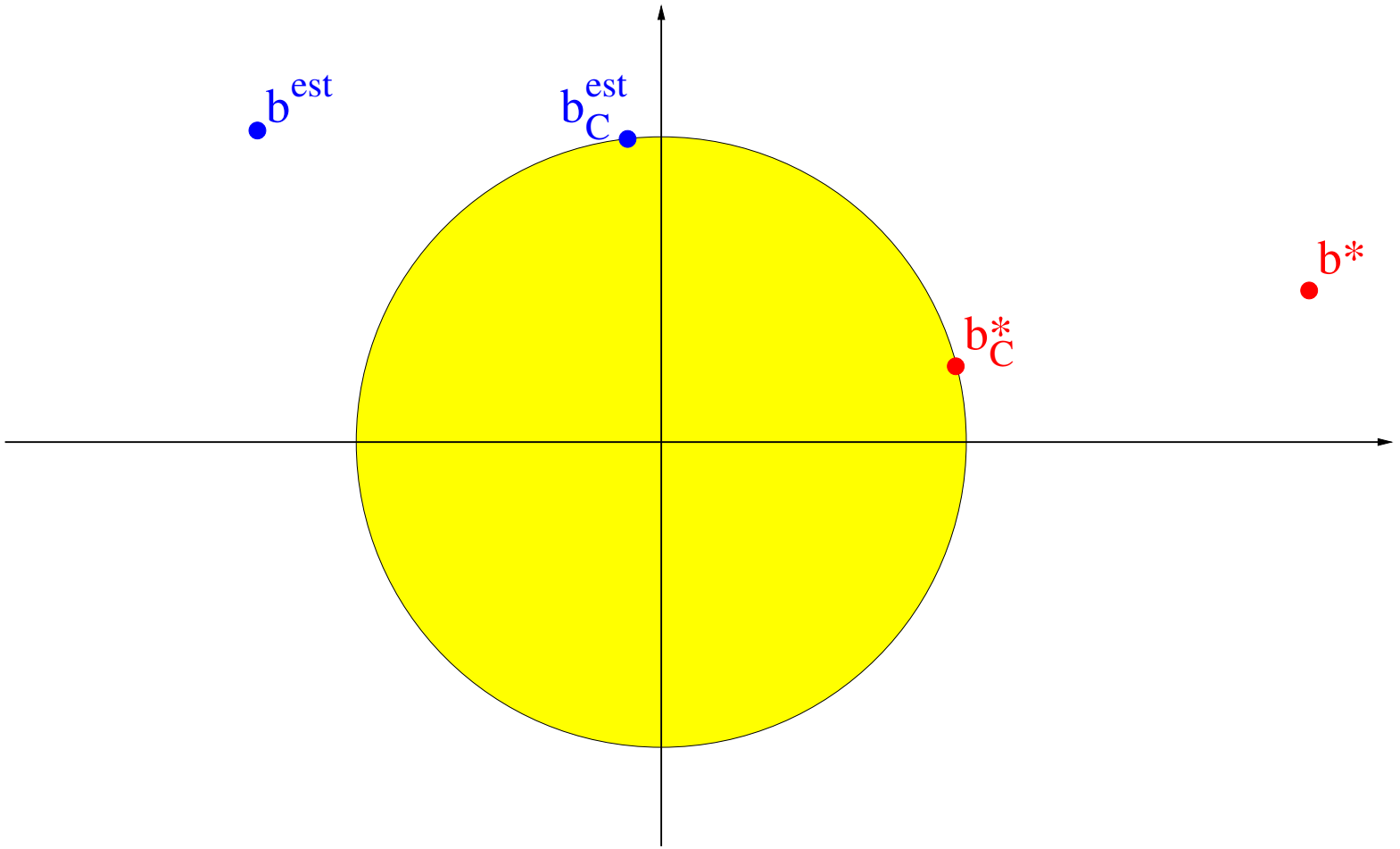
Why skrinkage classifiers?

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



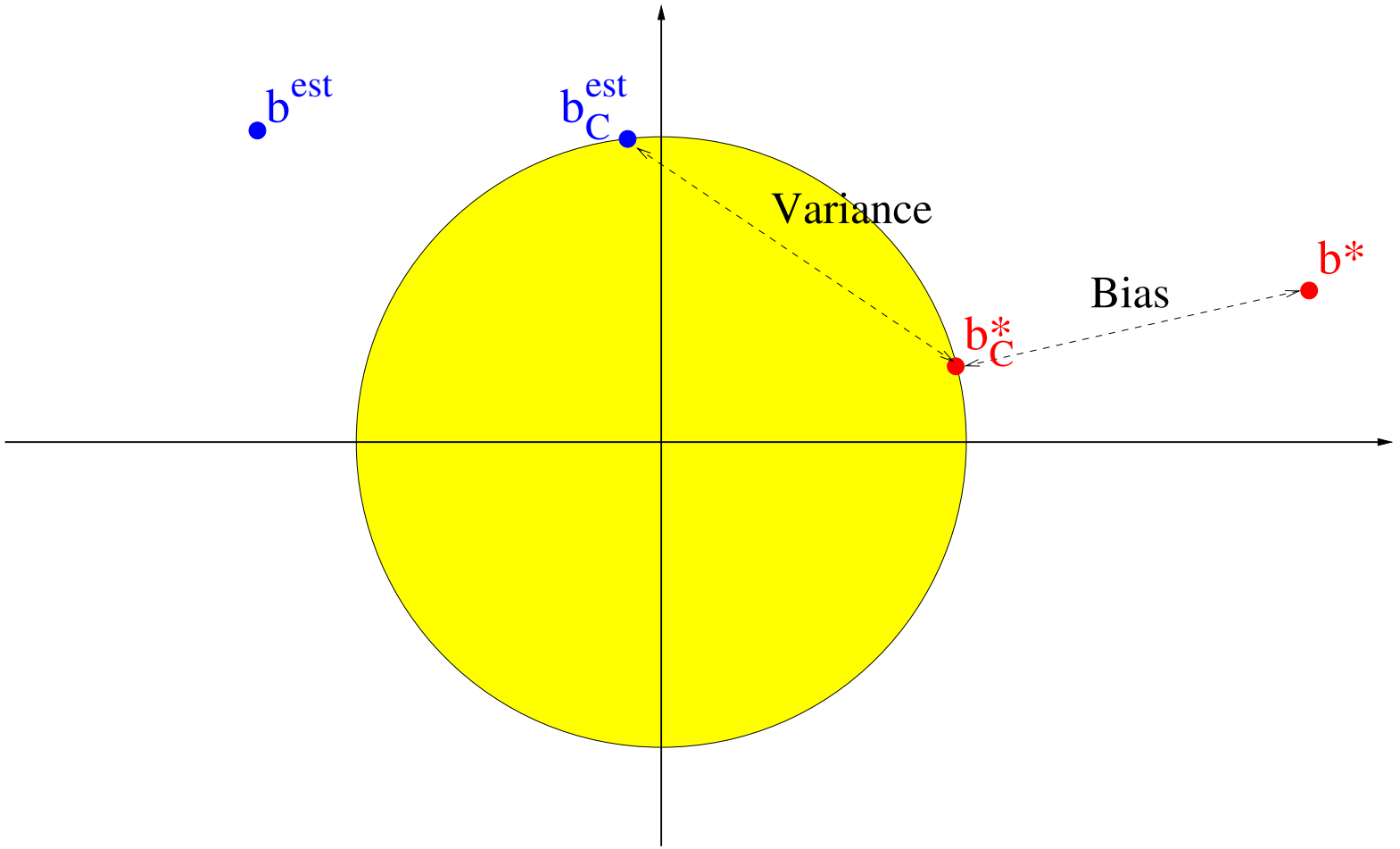
Why skrinkage classifiers?

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

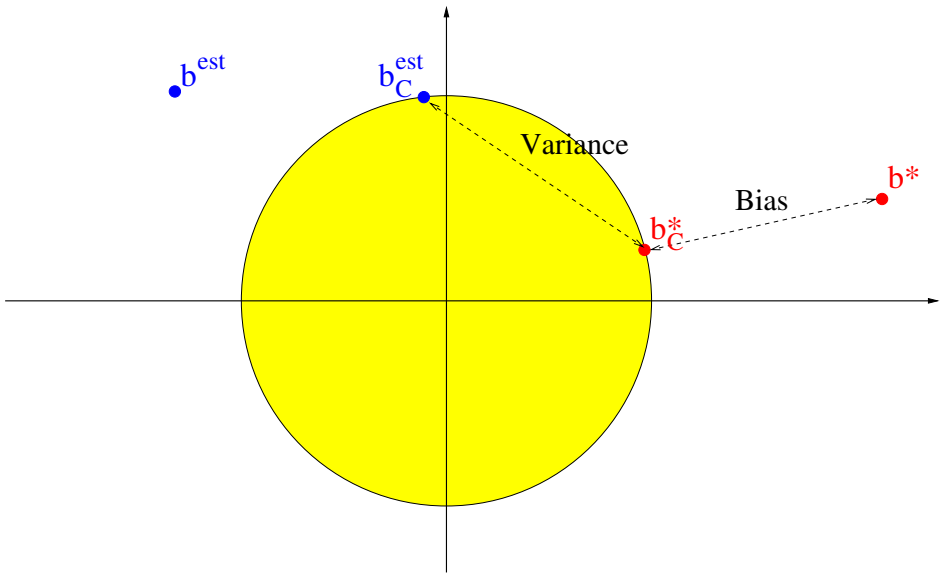


Why skrinkage classifiers?

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



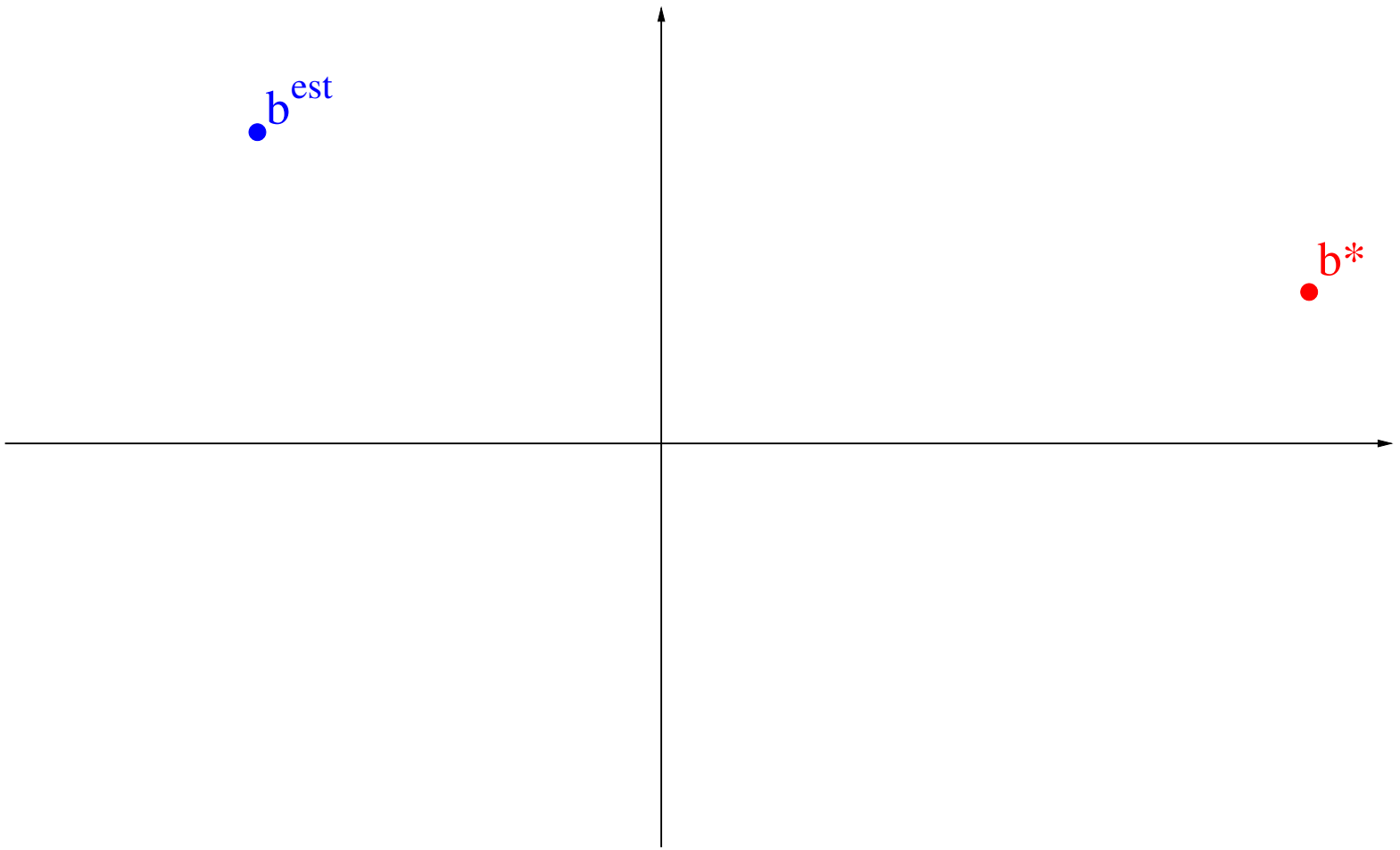
Why skrinkage classifiers?



- "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, ...)

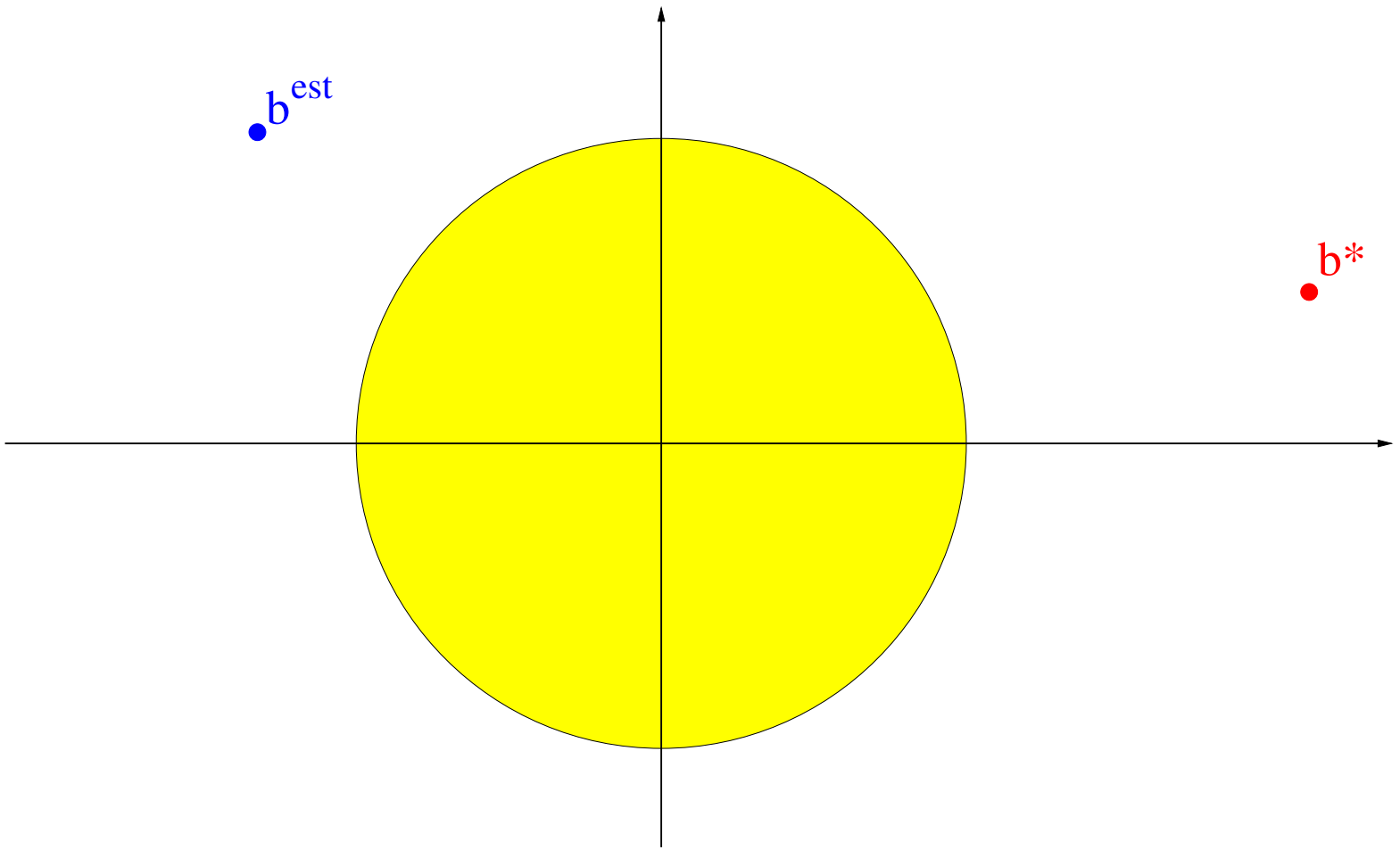
Including prior knowledge in the penalty?

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



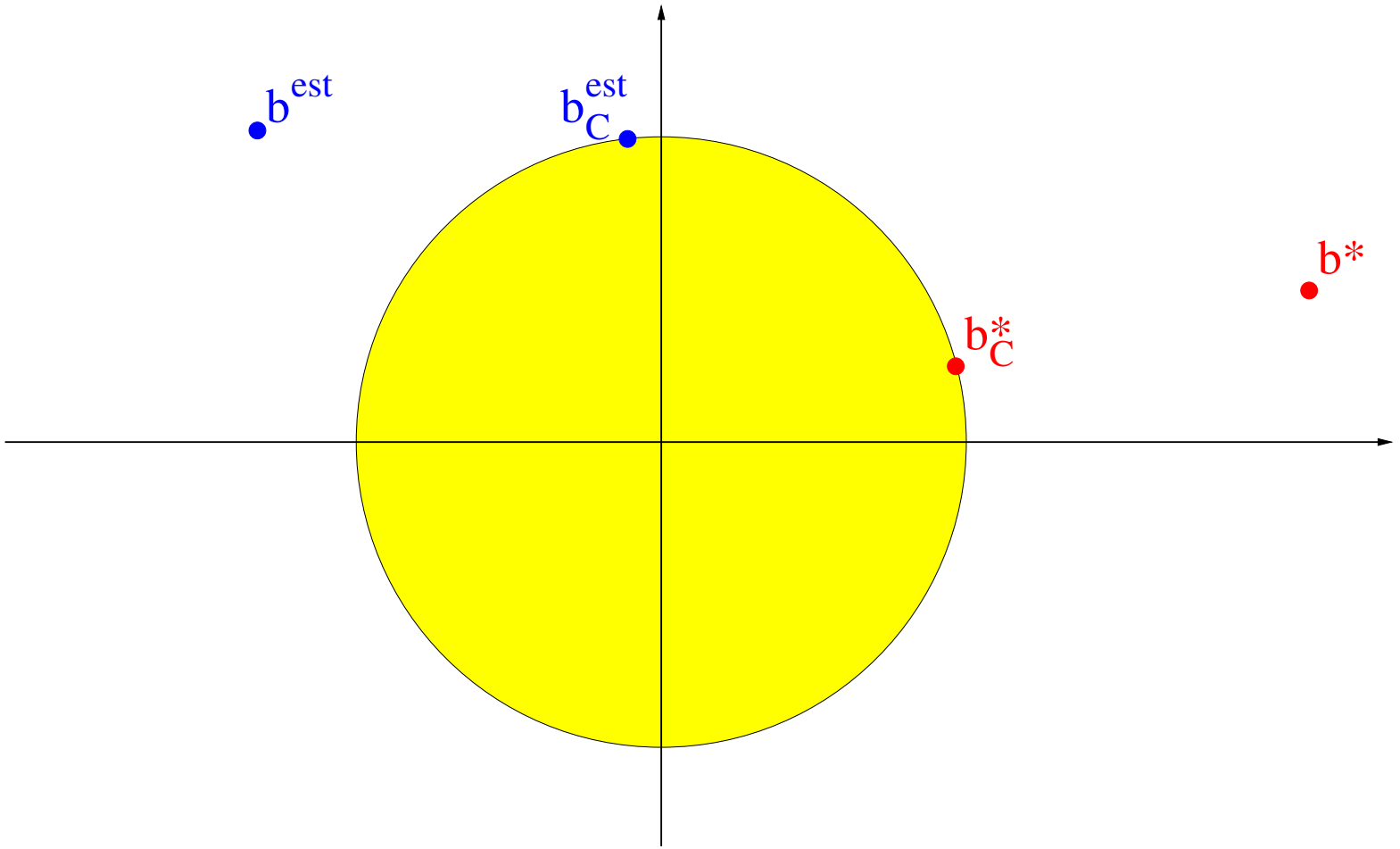
Including prior knowledge in the penalty?

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



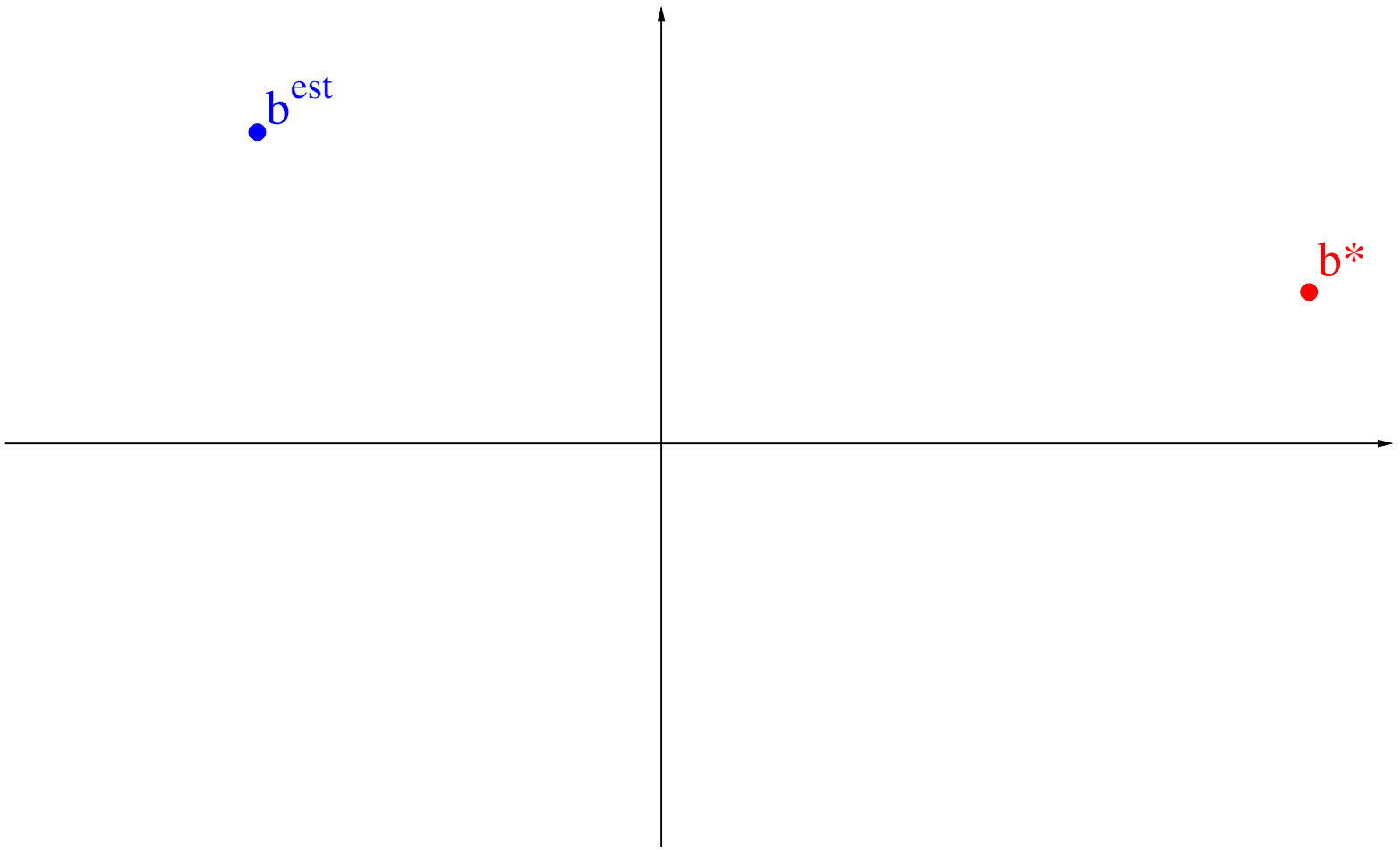
Including prior knowledge in the penalty?

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



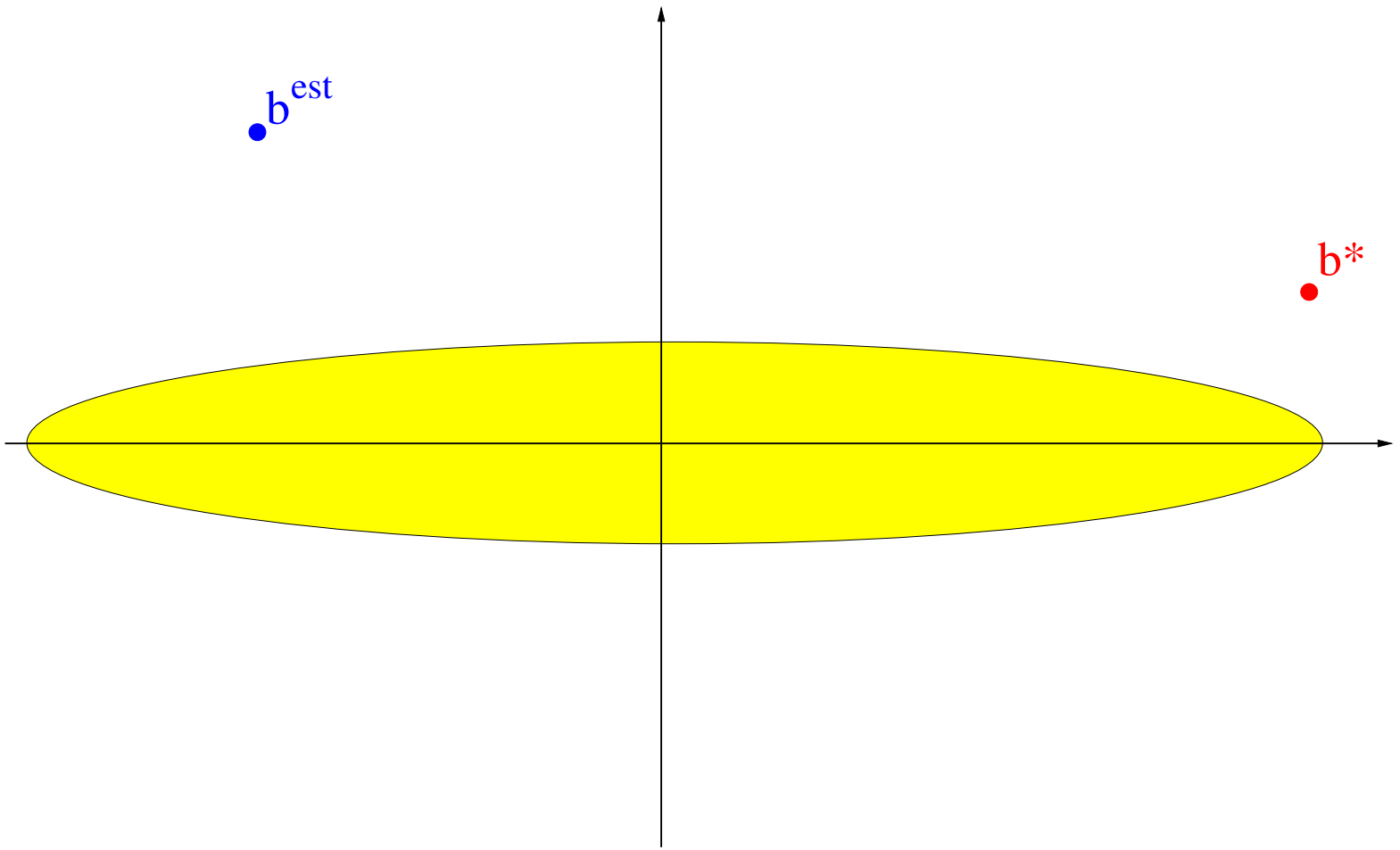
Including prior knowledge in the penalty?

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



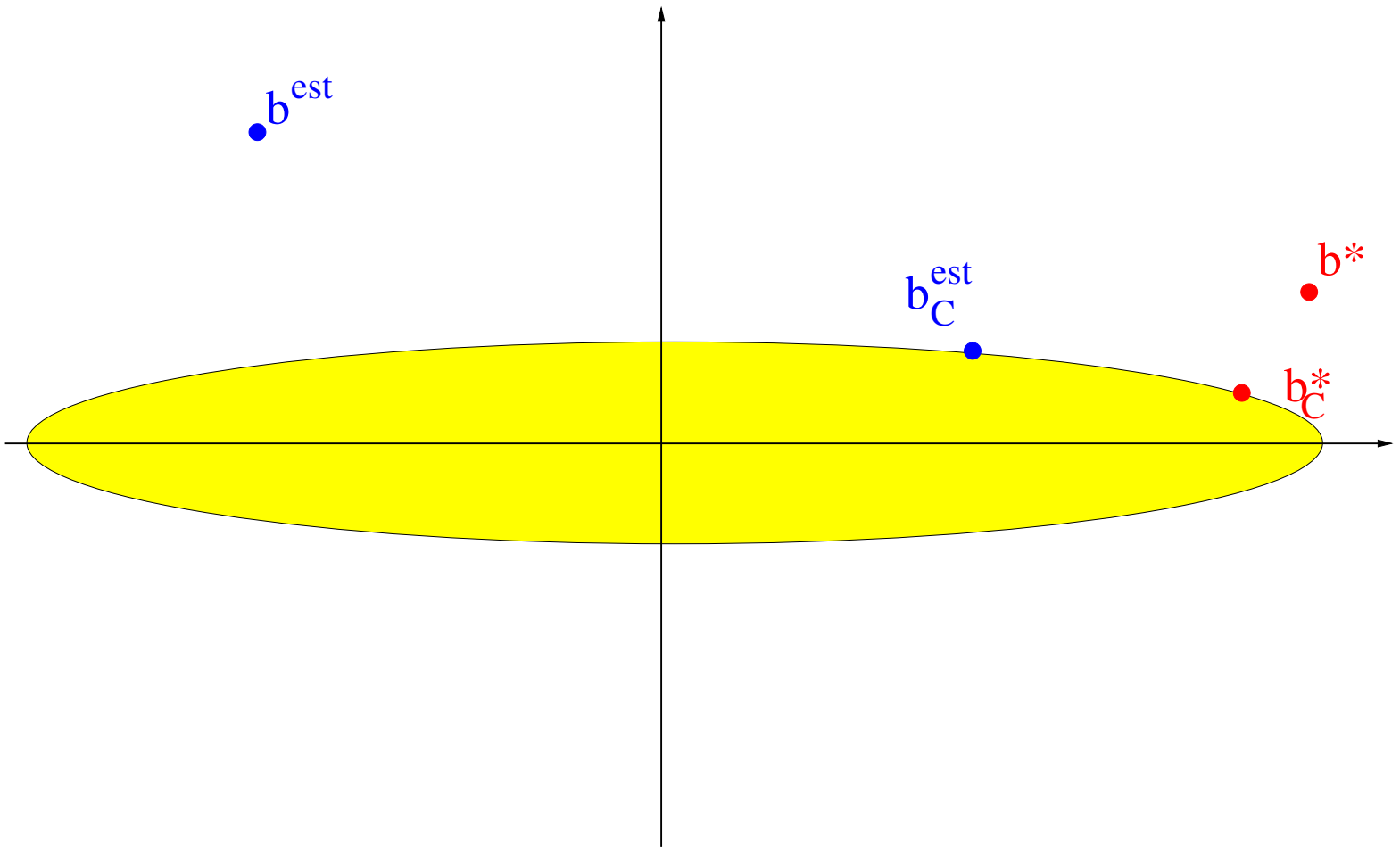
Including prior knowledge in the penalty?

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



Including prior knowledge in the penalty?

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

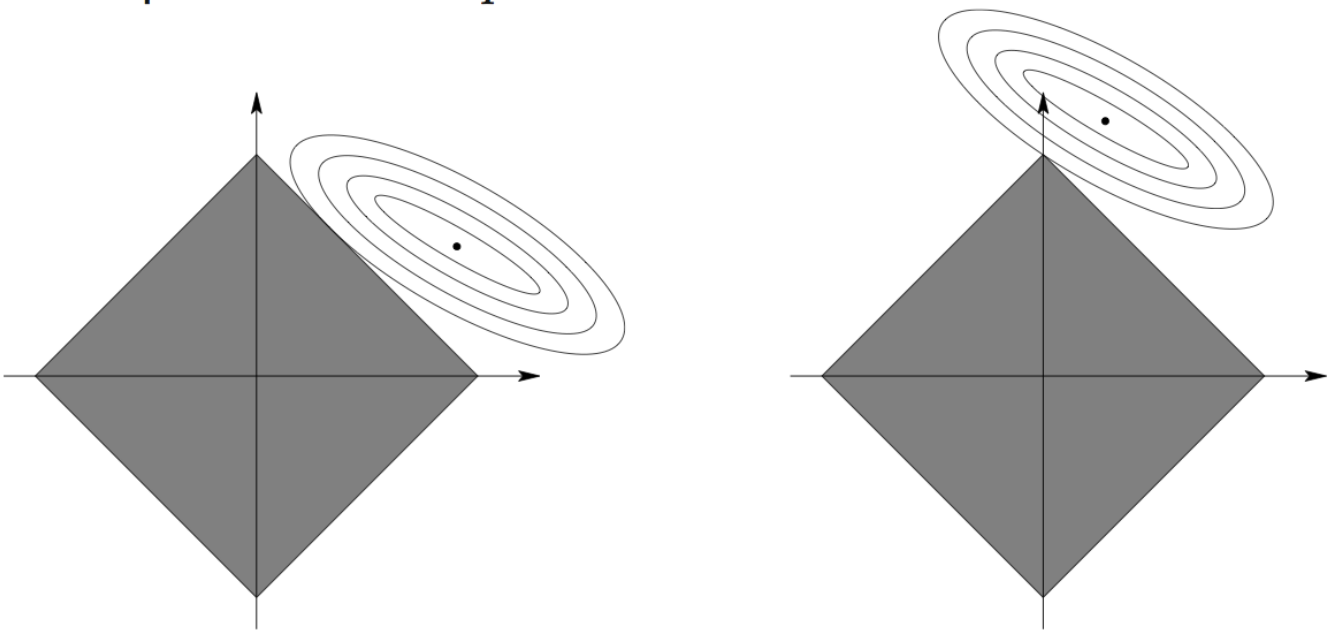


Further benefit: sparsity-inducing penalties

(Lasso)

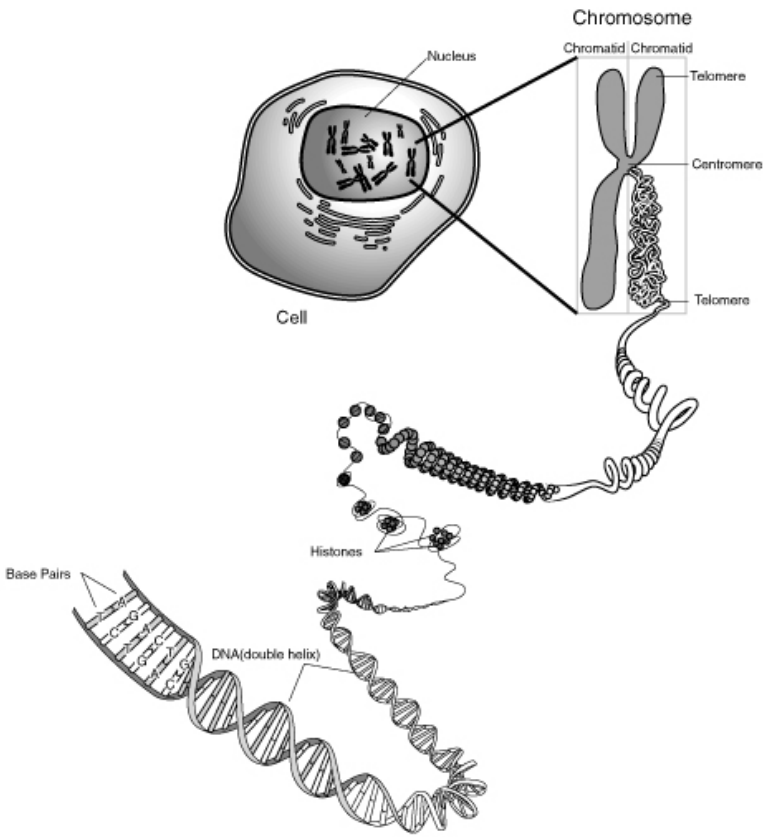
$$\min_{\beta} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

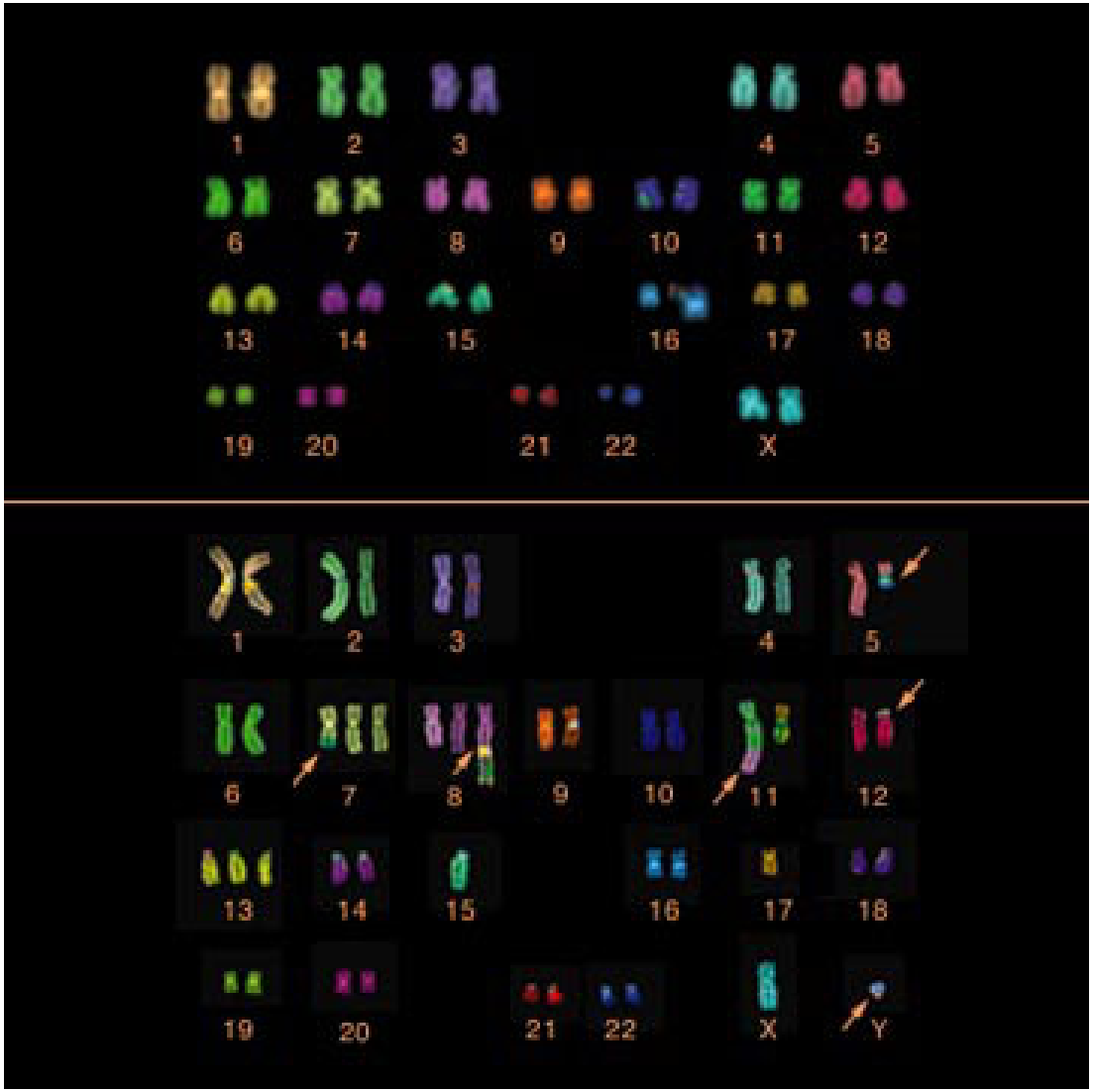
Geometric interpretation with $p = 2$



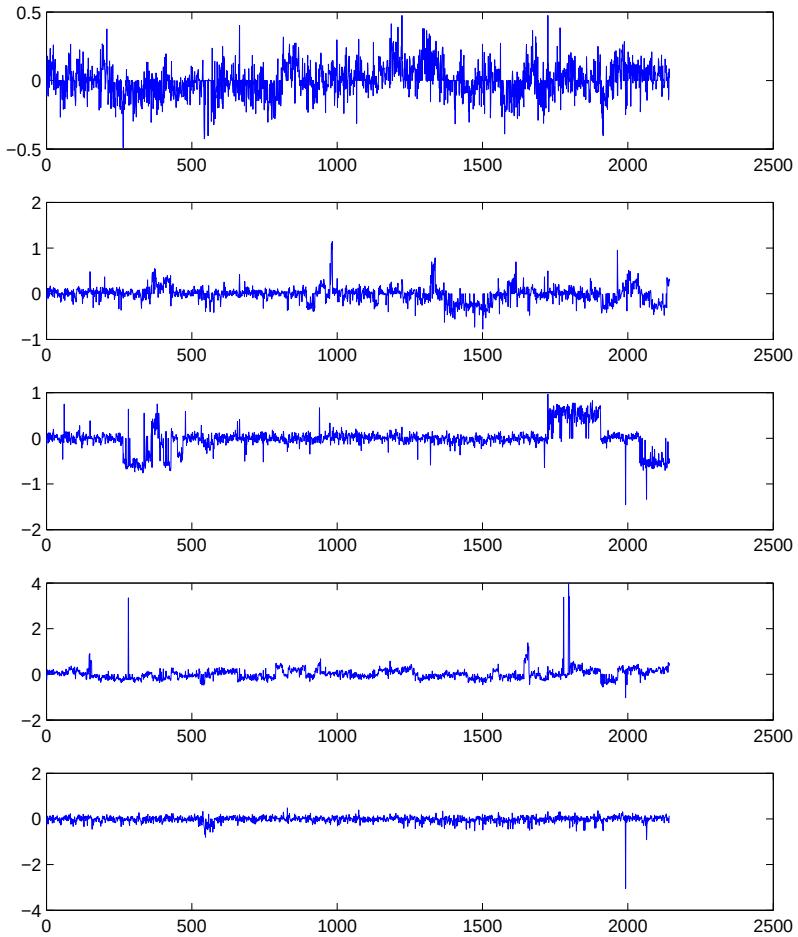
Outline

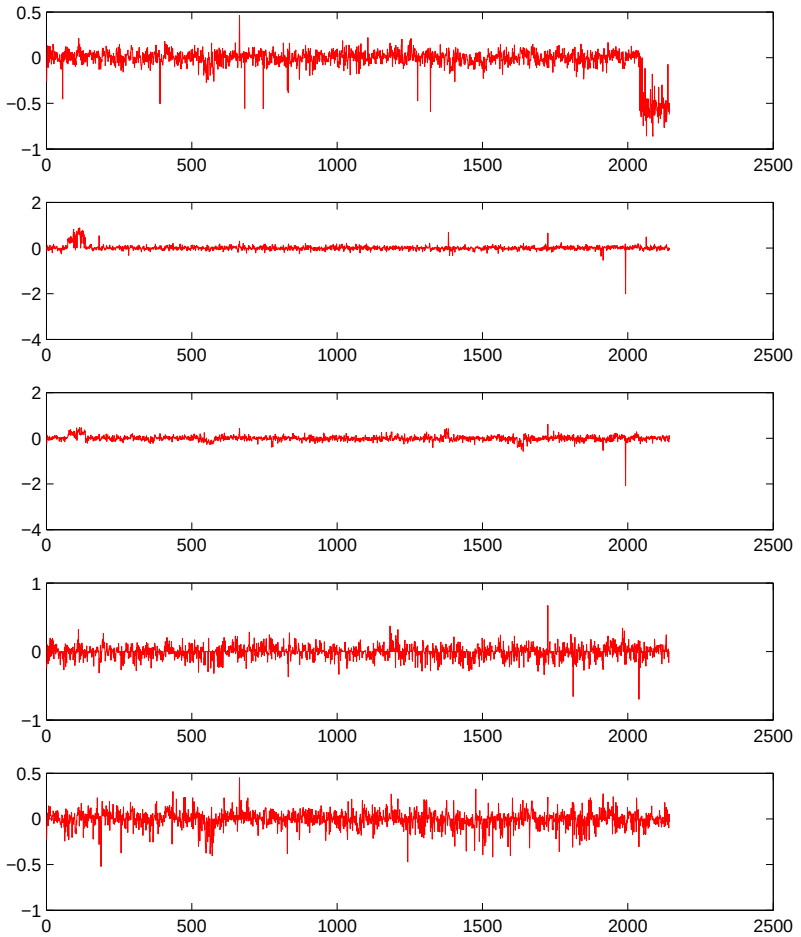
Chromosomic aberrations in cancer





Cancer prognosis: can we predict the future evolution?





Aggressive (left) vs non-aggressive (right) melanoma

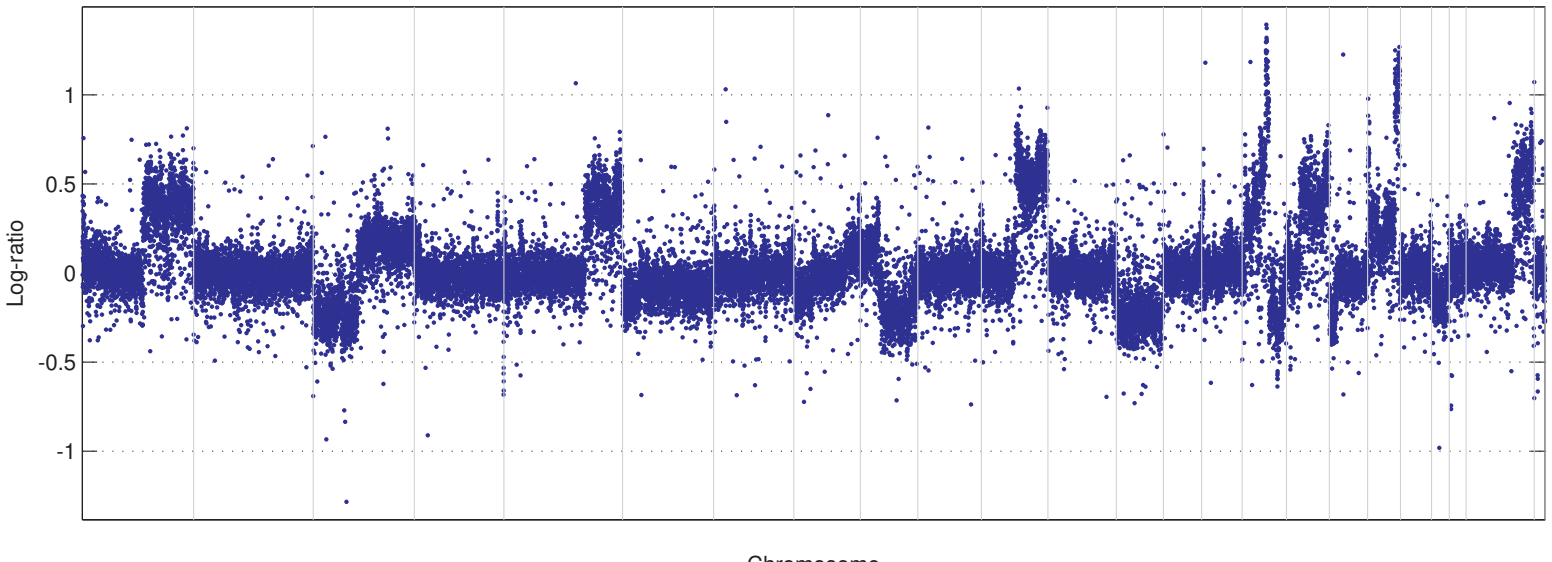
CGH array classification

Prior knowledge

- For a CGH profile $x \in \mathbb{R}^p$, we focus on linear classifiers, i.e., the sign of :

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

- We expect β to be
  - sparse : not all positions should be discriminative
  - piecewise constant : within a selected region, all probes should contribute equally



Promoting sparsity with the ℓ_1 penalty

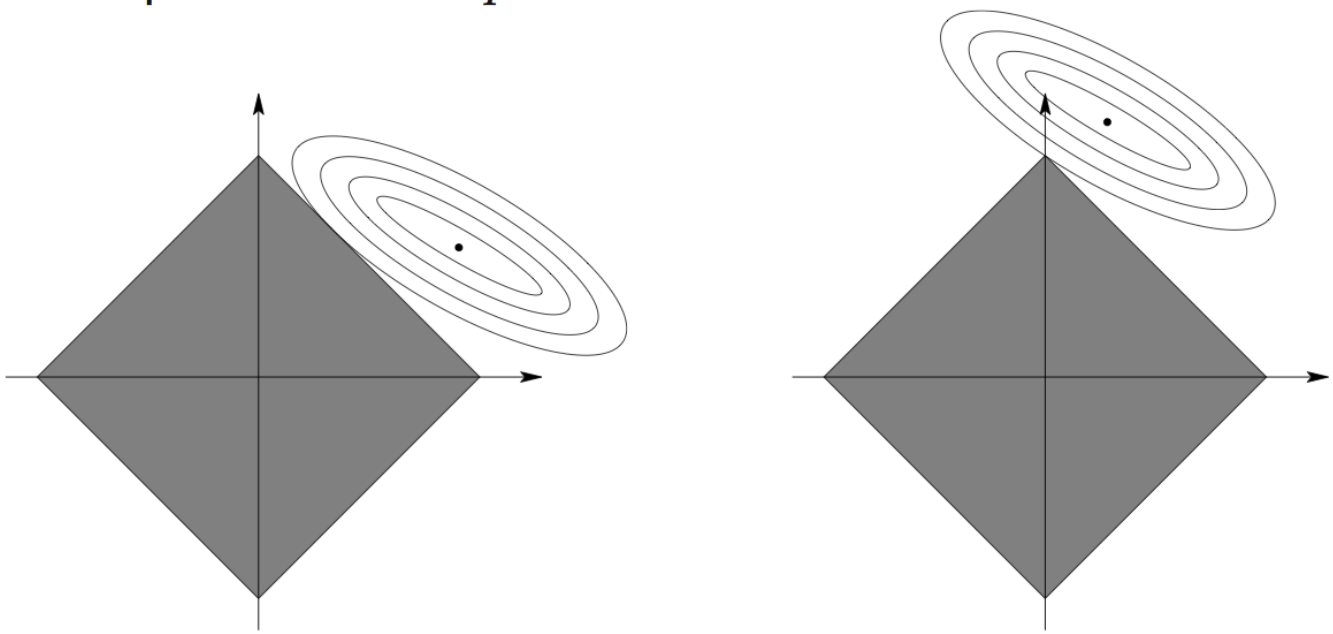
The ℓ_1 penalty (Tibshirani, 1996; Chen et al., 1998)

The solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

is usually sparse.

Geometric interpretation with $p = 2$



Promoting piecewise constant profiles penalty

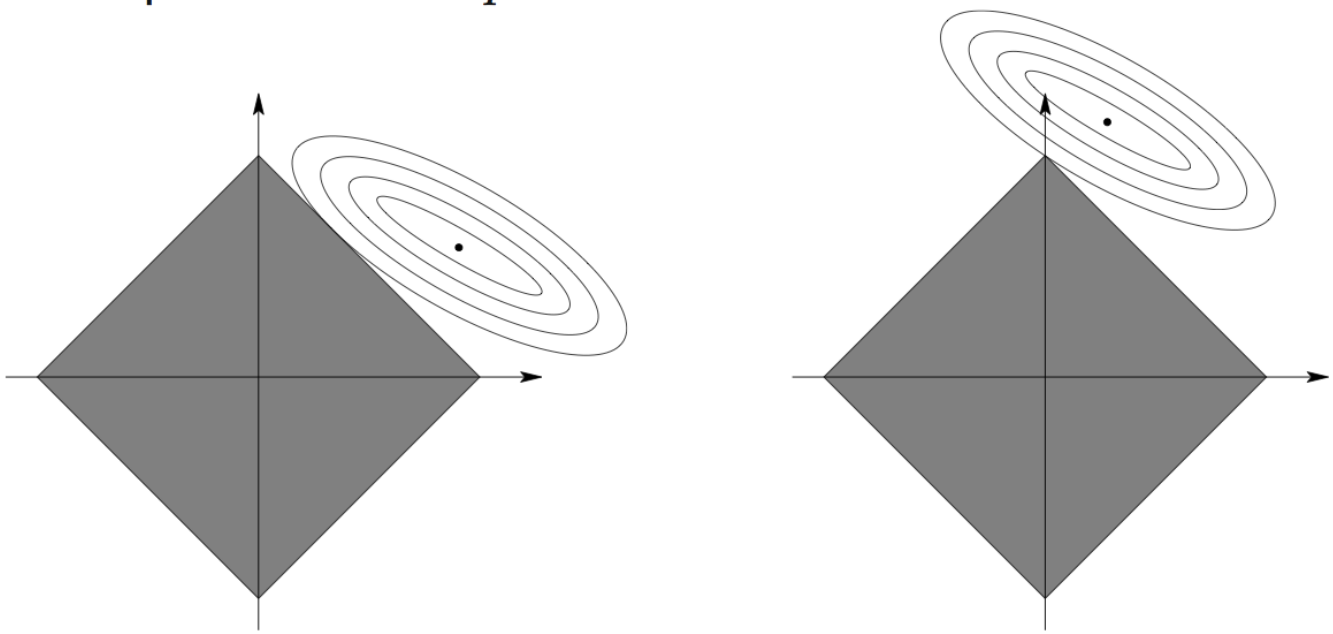
The variable fusion penalty (Land and Friedman, 1996)

The solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

is usually piecewise constant.

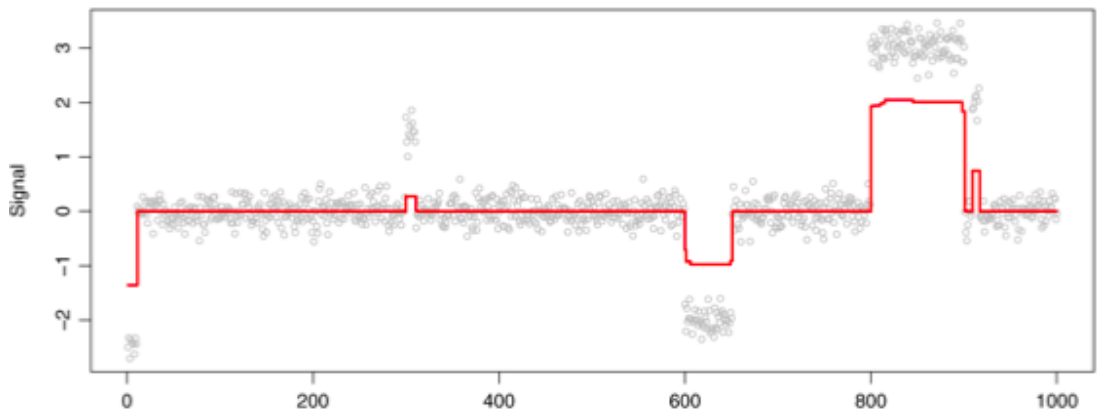
Geometric interpretation with $p = 2$



Fused Lasso signal approximator (Tibshirani et al., 2005)

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{p} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

- First term leads to sparse solutions
- Second term leads to piecewise constant solutions



Fused lasso for supervised classification (Rapaport et al., 2008)

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

where ℓ is, e.g., the hinge loss $\ell(y,t) = max(1 - yt, 0)$.

Implementation

- When ℓ is the hinge loss (fused SVM), this is a linear program -> up to $p = 10^3 \sim 10^4$
- When ℓ is convex and smooth (logistic, quadratic), efficient implementation with proximal methods -> up to $p = 10^8 \sim 10^9$

Fused lasso for supervised classification (Rapaport et al., 2008)

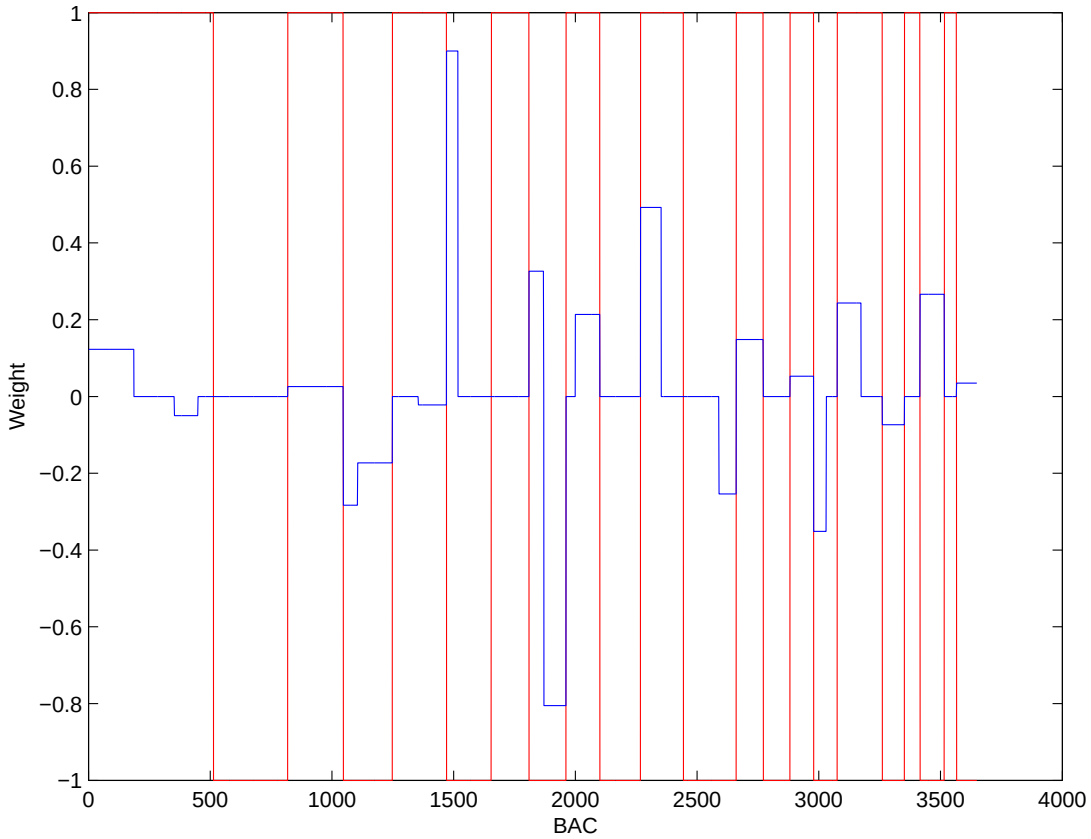
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \beta^\top x_i\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \,.$$

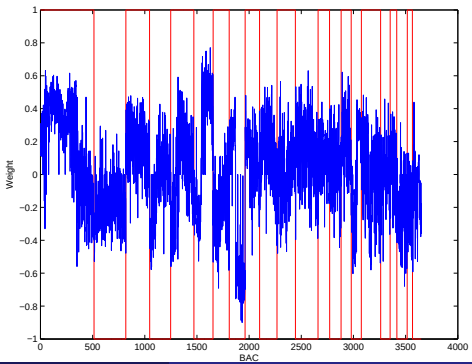
where ℓ is, e.g., the hinge loss $\ell(y,t) = max(1 - yt, 0)$.

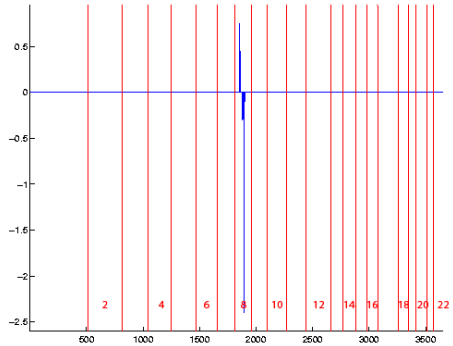
Implementation

- When ℓ is the hinge loss (fused SVM), this is a linear program -> up to $p = 10^3 \sim 10^4$
- When ℓ is convex and smooth (logistic, quadratic), efficient implementation with proximal methods -> up to $p = 10^8 \sim 10^9$

Example: predicting metastasis in melanoma

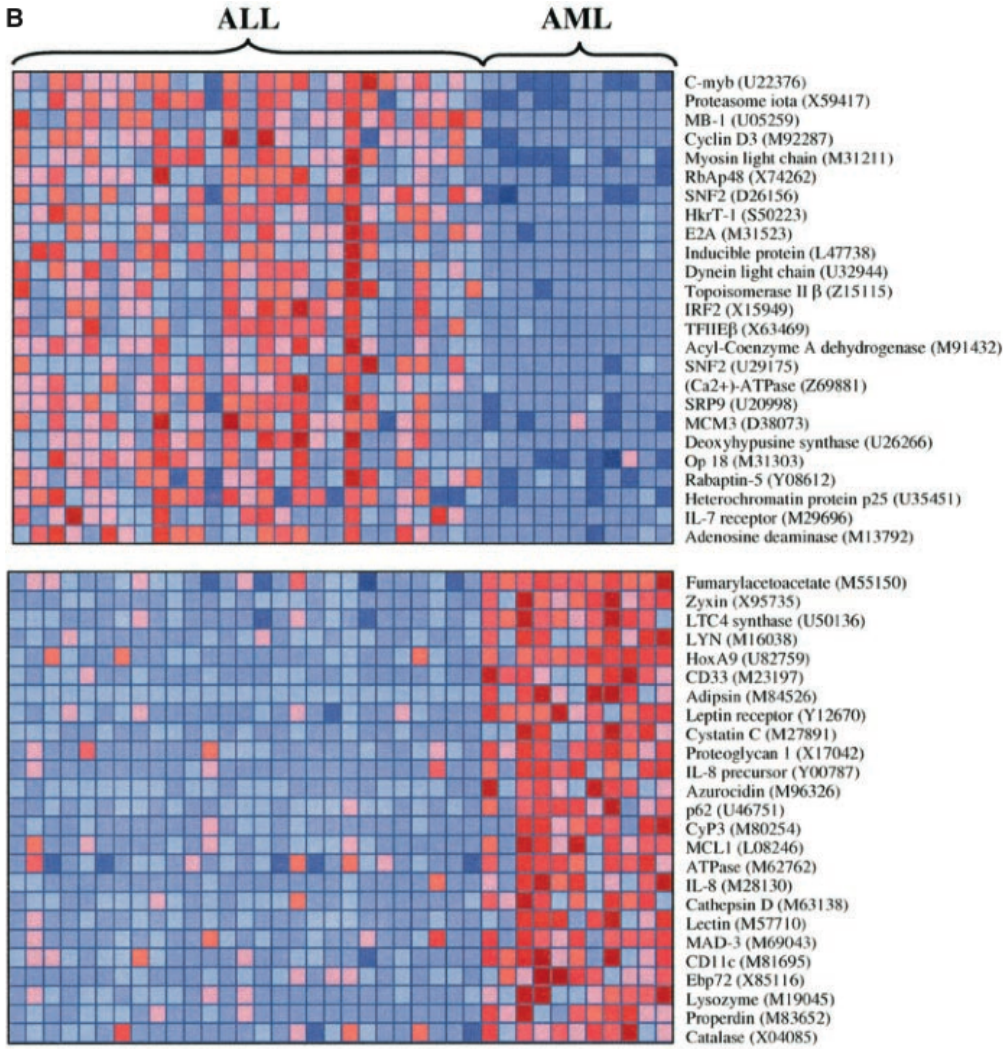




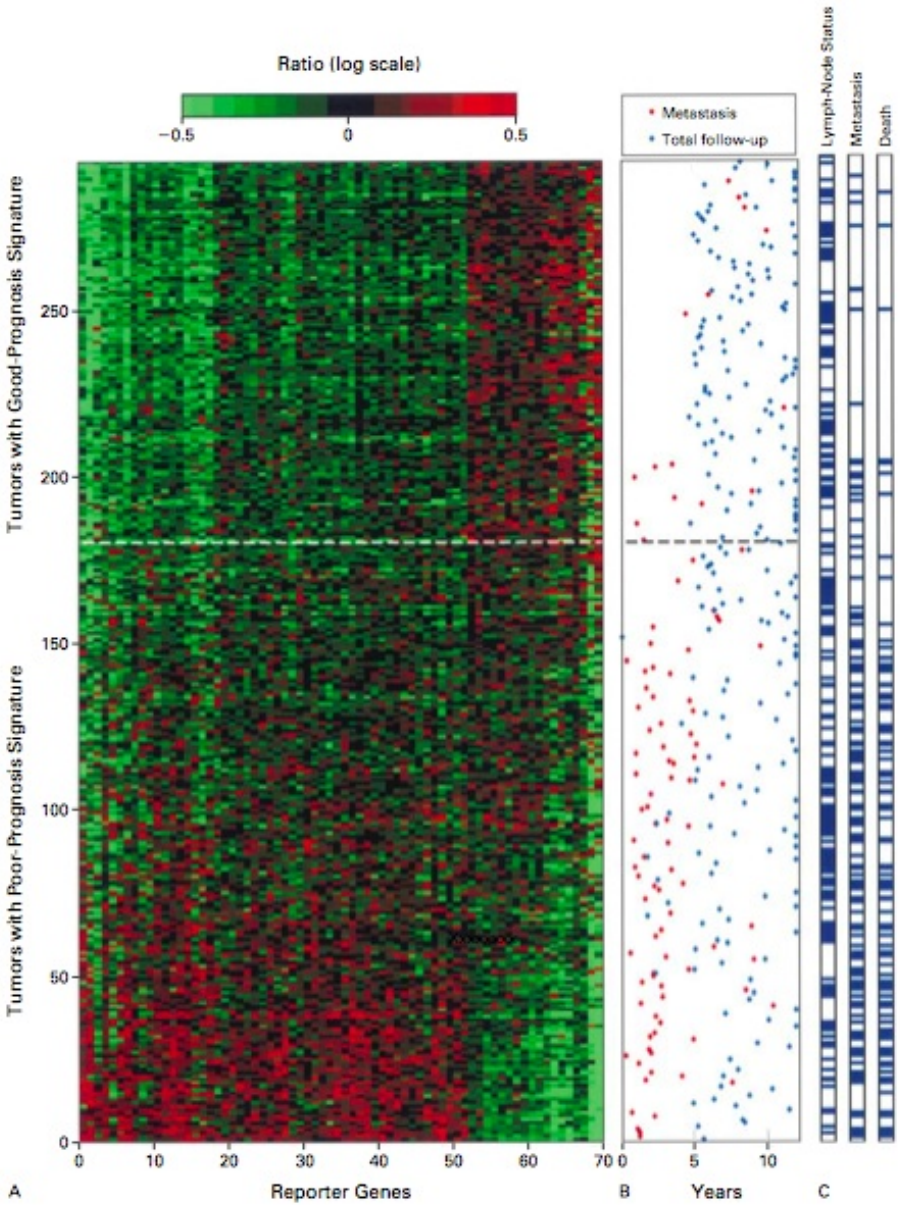


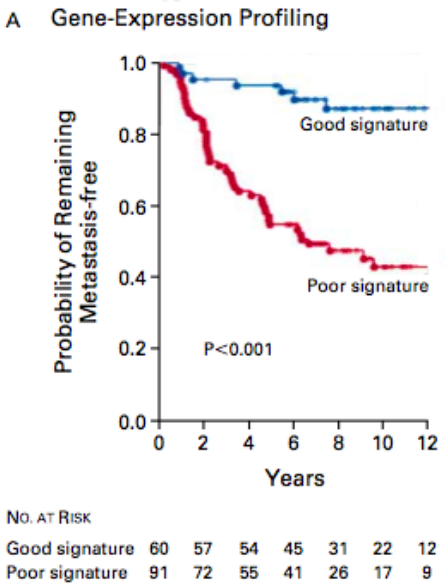
Outline

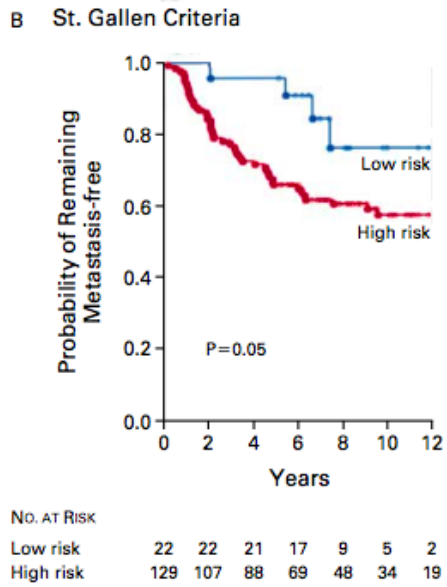
Diagnosis


B
ALL
AML
C-myb (U22376)
Proteasome iota (X59417)
MB-1 (U05259)
Cyclin D3 (M92287)
Myosin light chain (M31211)
RbAp48 (X74262)
SNF2 (D26156)
HkrT-1 (S50223)
E2A (M31523)
Inducible protein (L47738)
Dynein light chain (U32944)
Topoisomerase II β (Z15115)
IRF2 (X15949)
TFIIEβ (X63469)
Acyl-Coenzyme A dehydrogenase (M91432)
SNF2 (U29175)
(Ca2+)-ATPase (Z69881)
SRP9 (U20998)
MCM3 (D38073)
Deoxyhypusine synthase (U26266)
Op 18 (M31303)
Rabaptin-5 (Y08612)
Heterochromatin protein p25 (U35451)
IL-7 receptor (M29696)
Adenosine deaminase (M13792)
Fumarylacetoacetate (M55150)
Zyxin (X95735)
LTC4 synthase (U50136)
LYN (M16038)
HoxA9 (U82759)
CD33 (M23197)
Adipsin (M84526)
Leptin receptor (Y12670)
Cystatin C (M27891)
Proteoglycan 1 (X17042)
IL-8 precursor (Y00787)
Azurocidin (M96326)
p62 (U46751)
CyP3 (M80254)
MCL1 (L08246)
ATPase (M62762)
IL-8 (M28130)
Cathepsin D (M63138)
Lectin (M57710)
MAD-3 (M69043)
CD11c (M81695)
Ebp72 (X85116)
Lysozyme (M19045)
Properdin (M83652)
Catalase (X04085)

Prognosis


Ratio (log scale)
−0.5
0
0.5
Metastasis
Total follow-up
Lymph-Node Status
Metastasis
Death
Tumors with Good Prognosis Signature
Tumors with Poor Prognosis Signature
0
50
100
150
200
250
0
10
20
30
40
50
60
70
A
Reporter Genes
0
5
10
B
Years
C

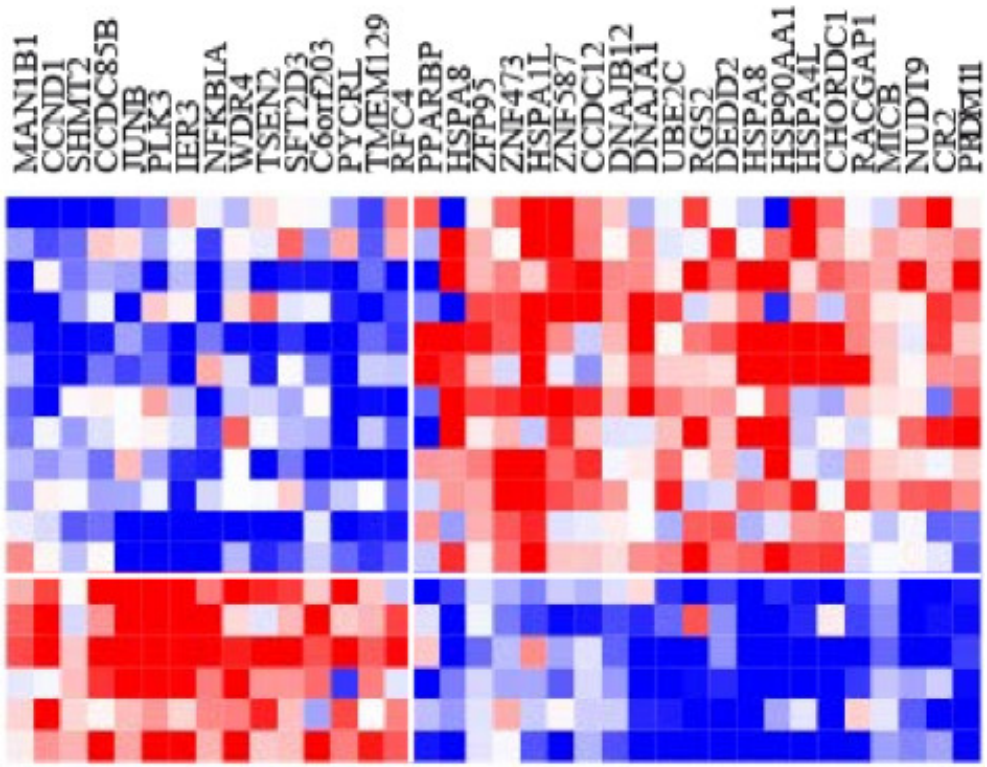

A
Gene-Expression Profiling
Probability of Remaining Metastasis-free
1.0
0.8
0.6
0.4
0.2
0.0
Good signature
Poor signature
P<0.001
0
2
4
6
8
10
12
Years
No. at Risk
Good signature 60 57 54 45 31 22 12
Poor signature 91 72 55 41 26 17 9


B
St. Gallen Criteria
Probability of Remaining Metastasis-free
1.0
0.8
0.6
0.4
0.2
0.0
Low risk
High risk
P=0.05
0
2
4
6
8
10
12
Years
No. at Risk
Low risk 22 22 21 17 9 5 2
High risk 129 107 88 69 48 34 19

Gene selection, molecular signature

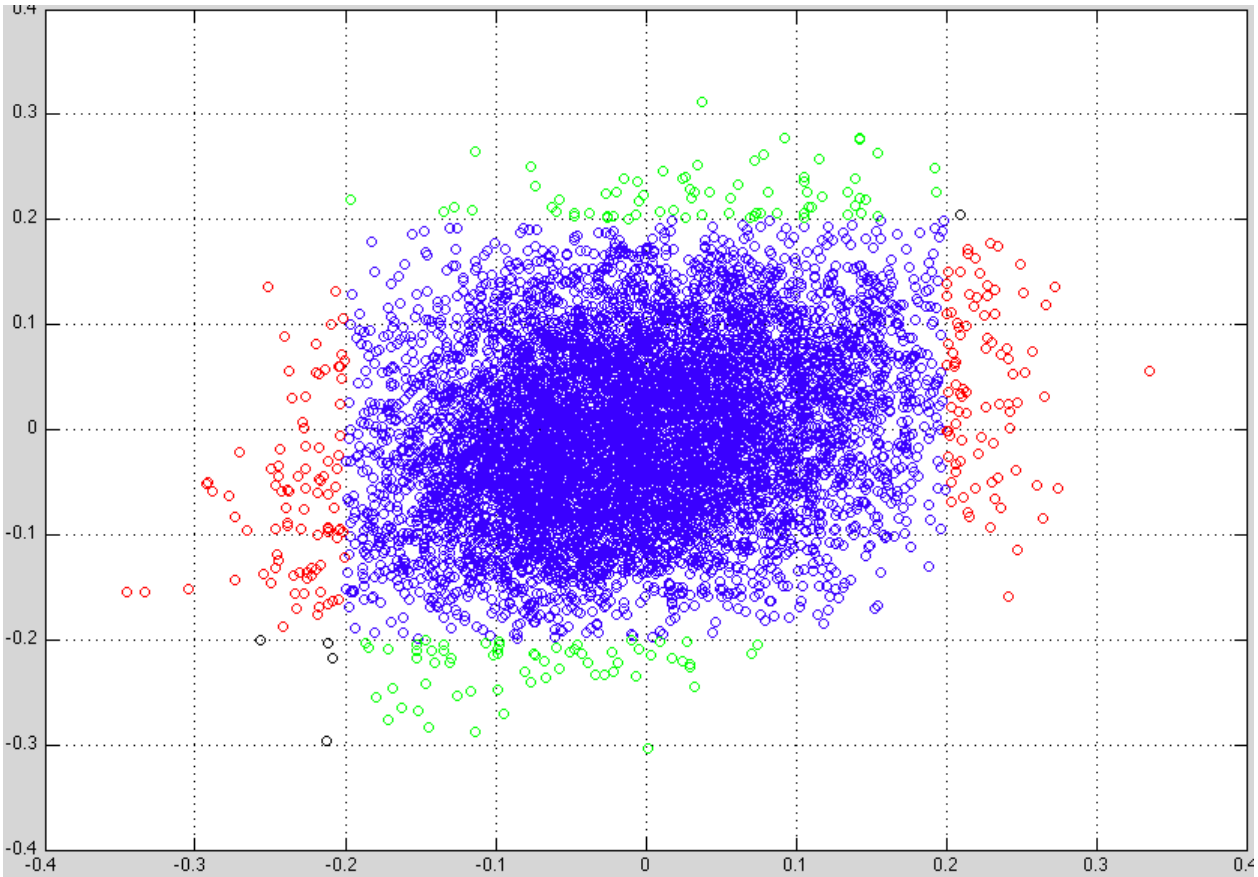
The idea

- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology

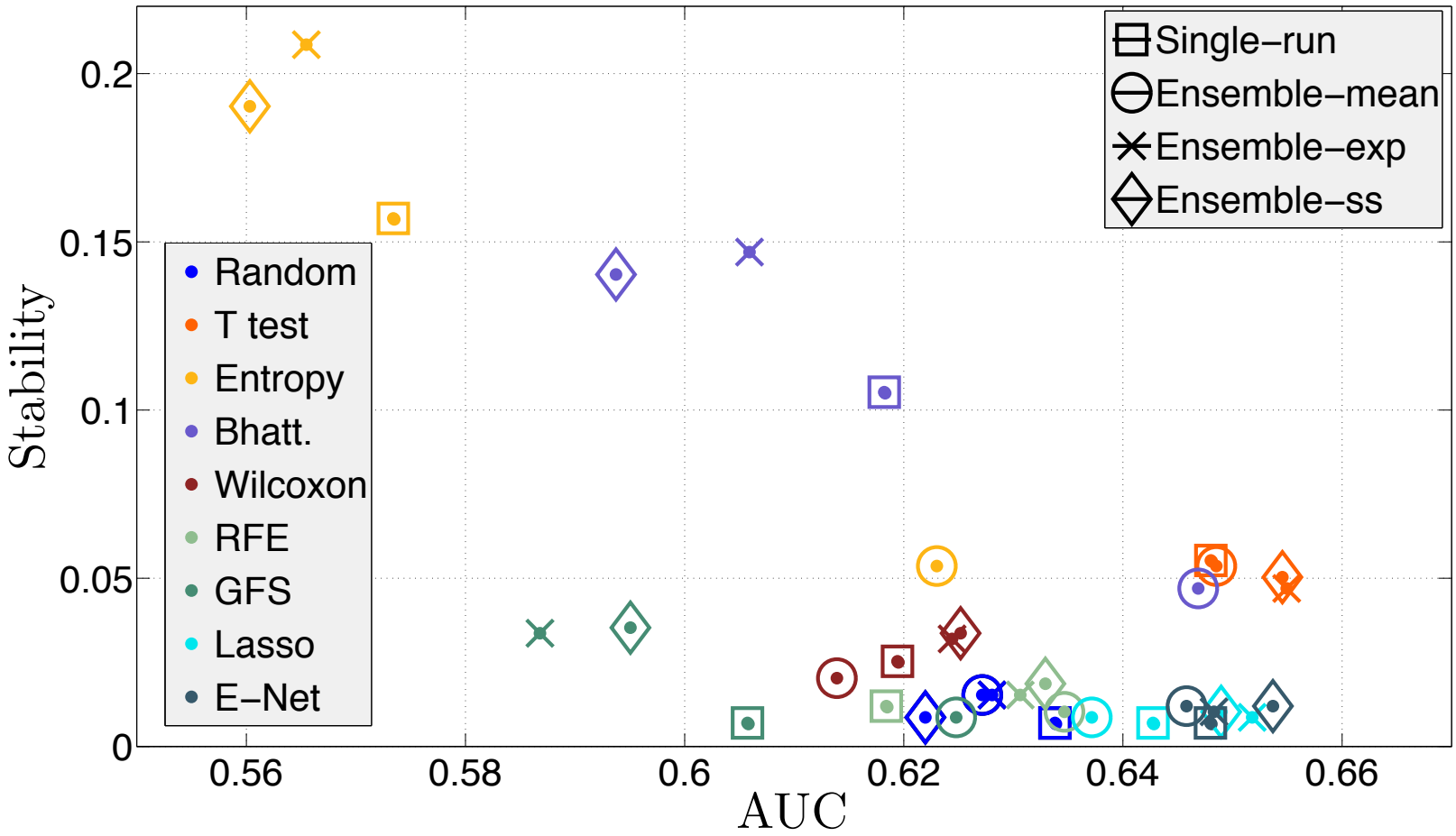


But... unstability of selected features

- Wang dataset: $n = 286$, $p = 8141$
- Pearson correlation with the output on 2 random subsamples of 143 samples:

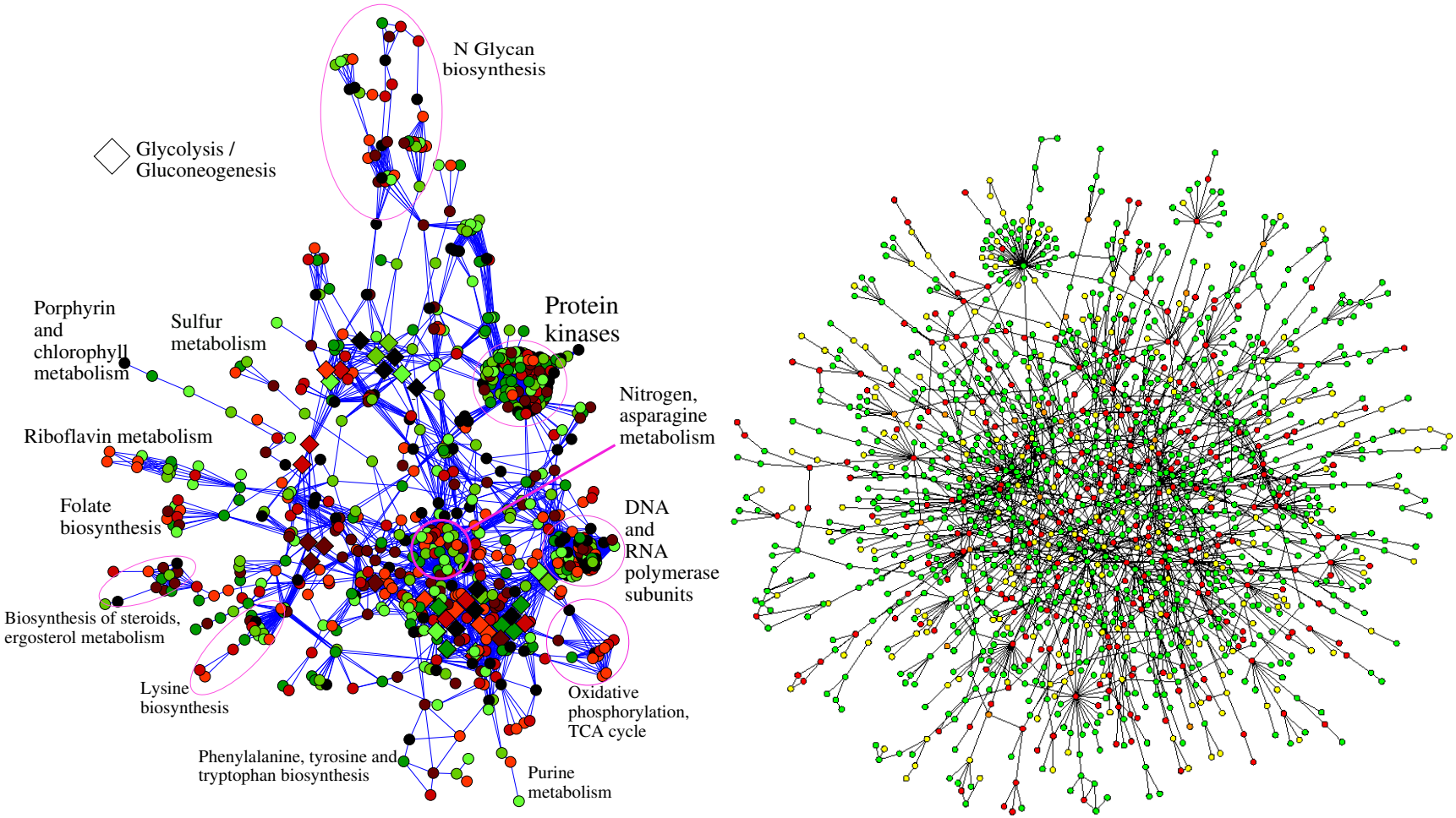


Comparison of feature selection methods...



Haury et al. (2011)

Gene networks



Gene networks and expression data

Motivation

- 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 groups through functional groups and protein networks

Shrinkage estimators with prior knowledge

$$\min_{\beta} R(\beta) + \lambda \Omega(\beta)$$

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

1. Connected genes on a network should have similar weights
2. Select few genes that are connected or belong to same predefined functional groups

Gene networks and expression data

Motivation

- 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 groups through functional groups and protein networks

Shrinkage estimators with prior knowledge

$$\min_{\beta} R(\beta) + \lambda \Omega(\beta)$$

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

1. Connected genes on a network should have similar weights
2. Select few genes that are connected or belong to same predefined functional groups

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

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

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

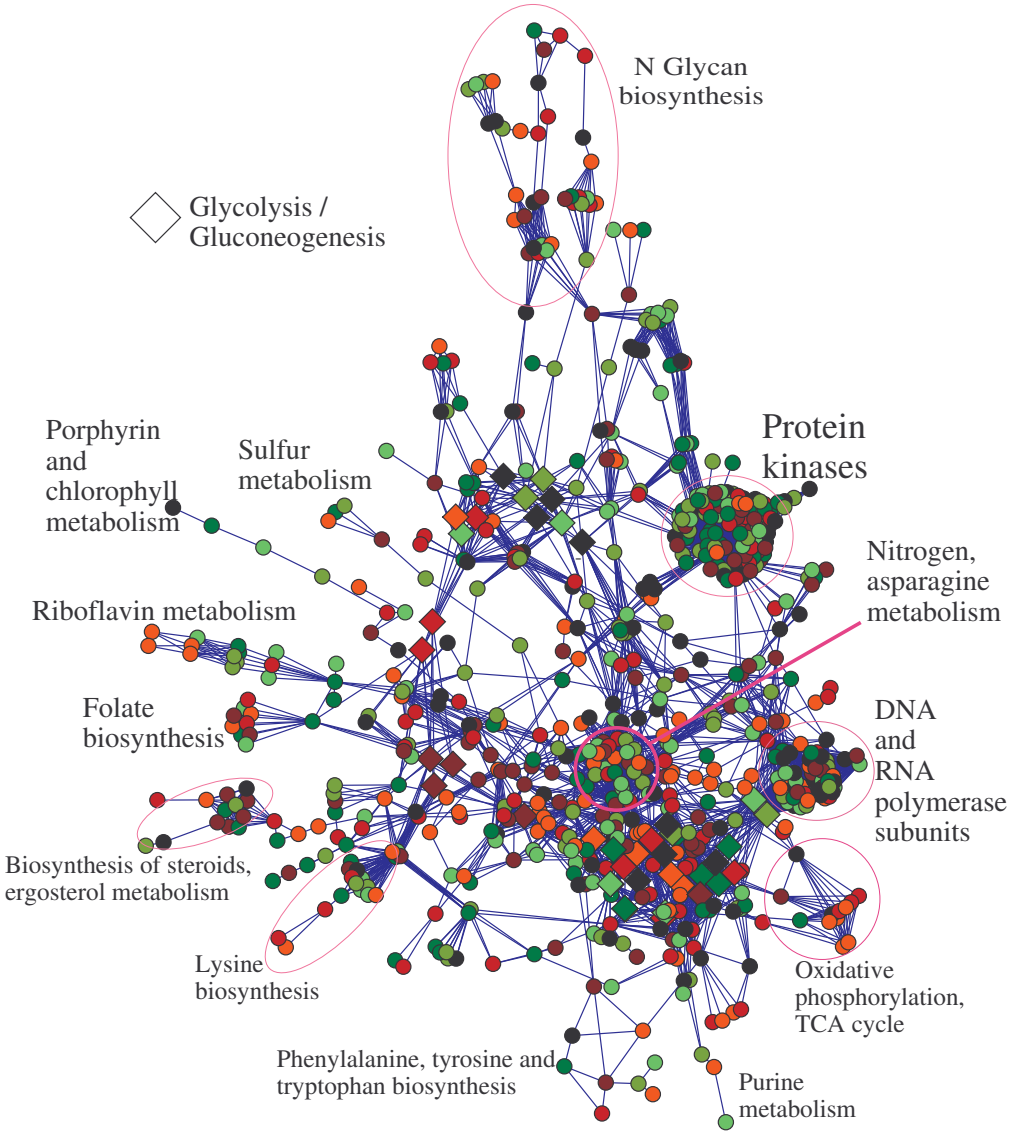
- Gene selection + smooth on the graph

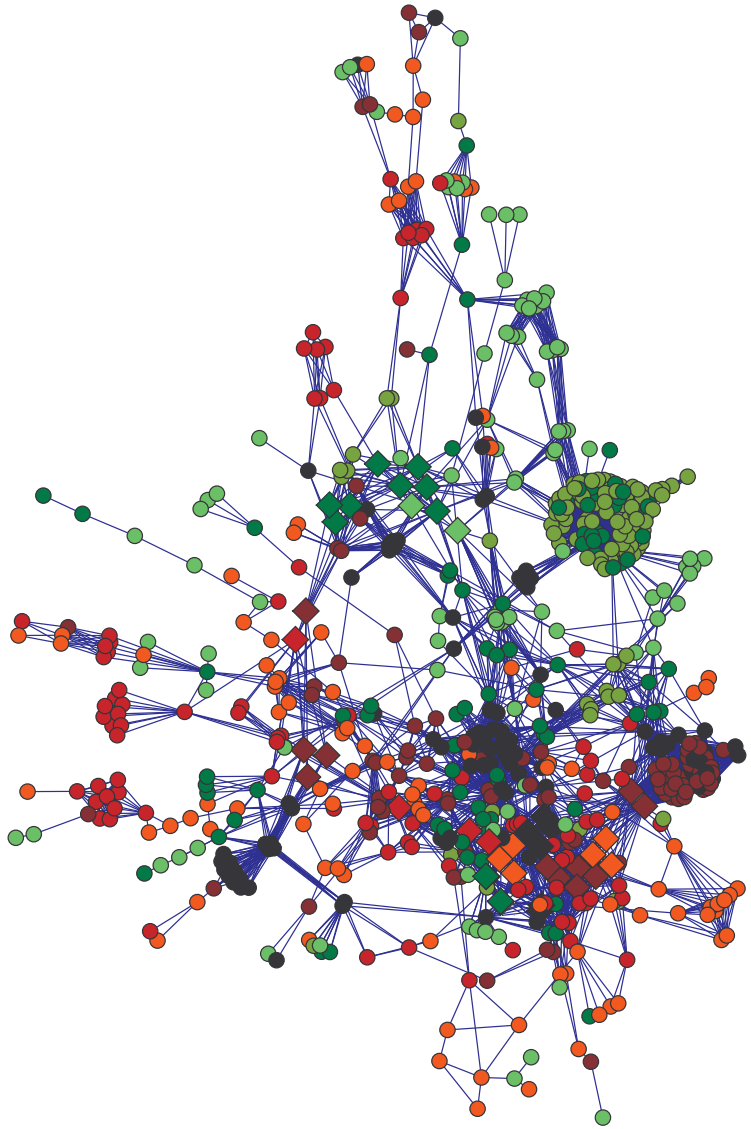
$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\beta_i|$$

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

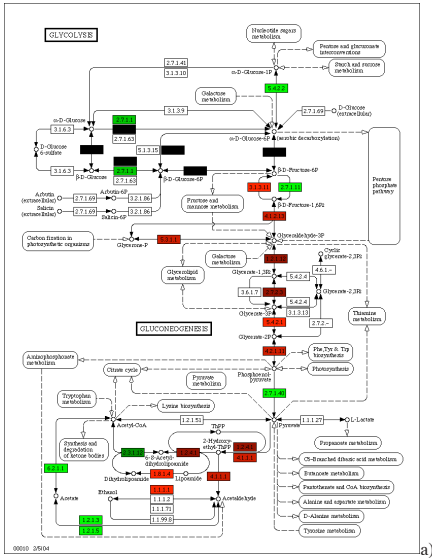
$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i|$$

Illustration

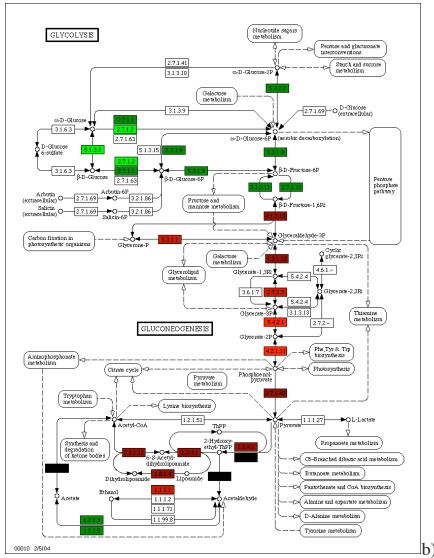




Limits



a)



b)

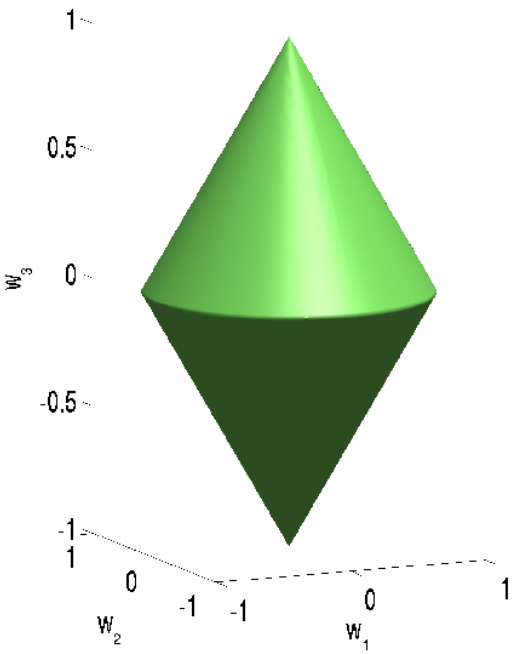
- We are happy to see pathways appear.
- However, in some cases, connected genes should have "opposite" weights (inhibition, pathway branching, etc...)
- How to capture pathways without constraints on the weight similarities?

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*:

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



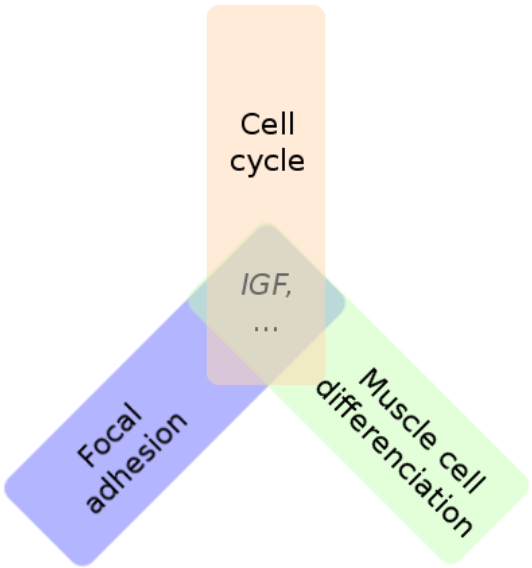
$$\Omega(\beta_1, \beta_2, \beta_3) = \|(\beta_1, \beta_2)\|_2 + \|\beta_3\|_2$$
$$= \sqrt{\beta_1^2 + \beta_2^2} + |\beta_3|$$

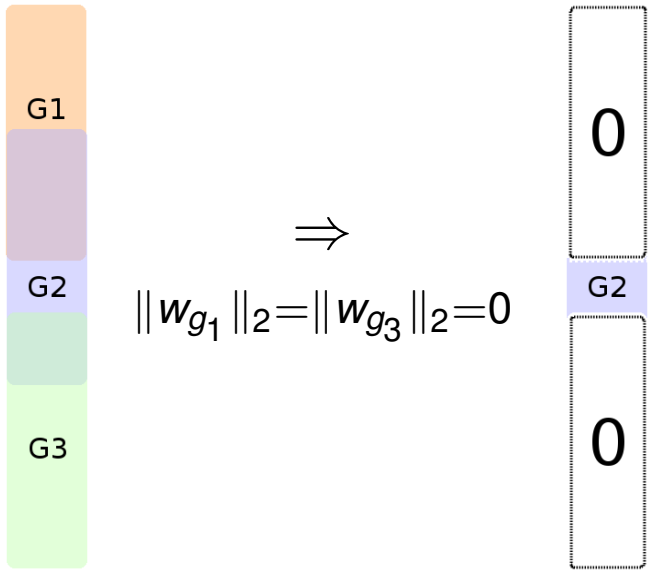
Group Lasso when groups overlap

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

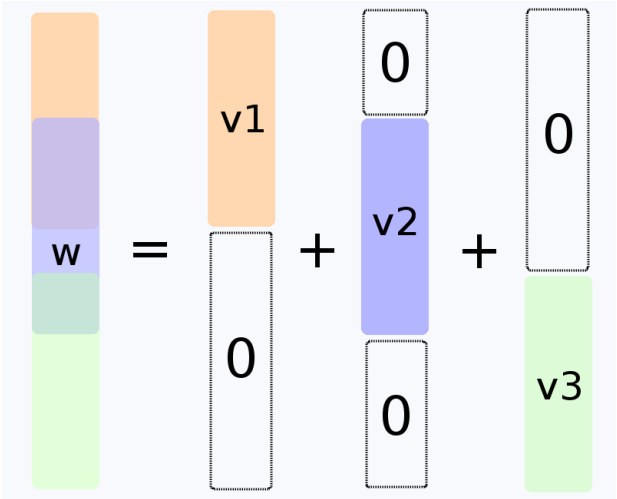
Removal of *any* group containing a gene $\Rightarrow$ the weight of the gene is 0.

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:

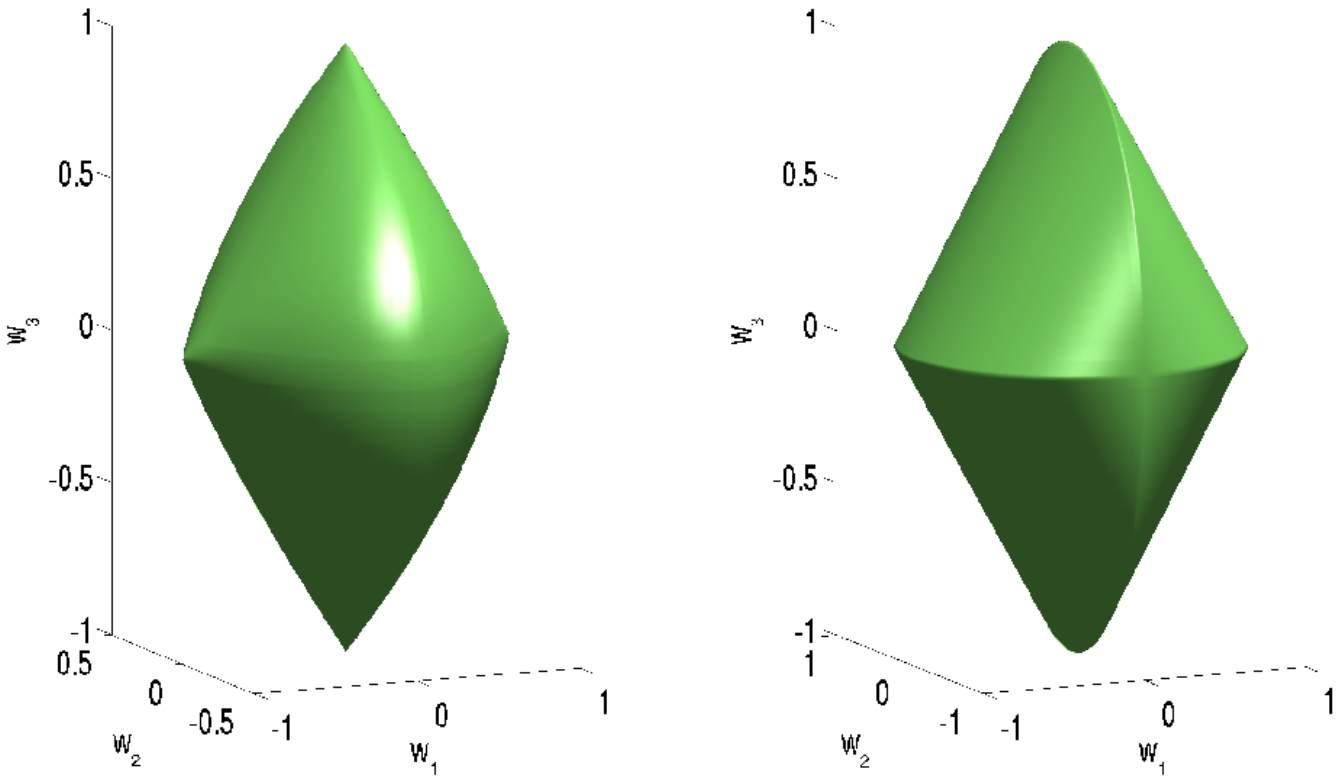
$$\Omega_{latent}(\beta) \triangleq \begin{cases} \min_v \sum_{g \in \mathcal{G}} \| v_g \|_2 \\ \beta = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases}$$



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,

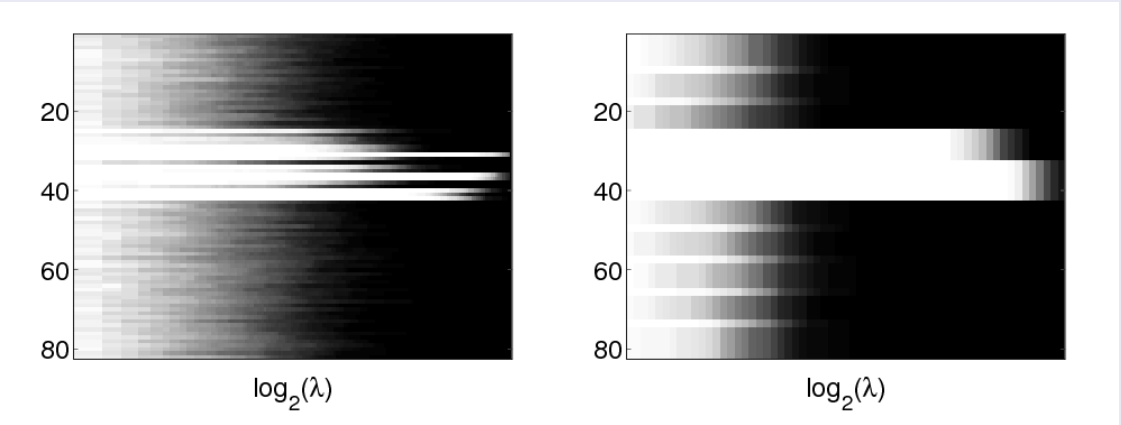
the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

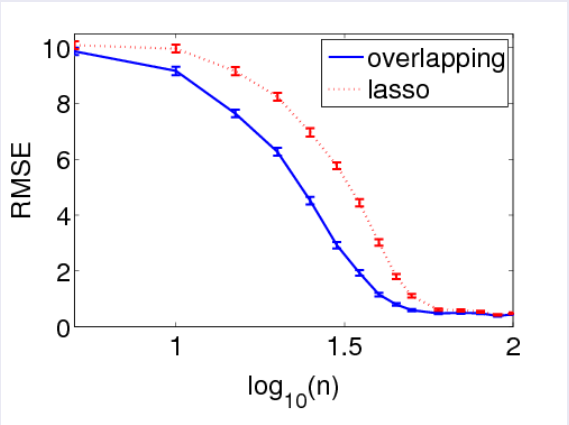
$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

Experiments

Synthetic data: overlapping groups

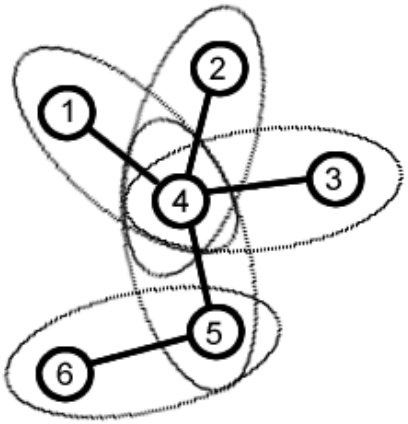
- 10 groups of 10 variables with 2 variables of overlap between two successive groups :$\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$.
- Support: union of $4th$ and $5th$ groups.
- Learn from 100 training points.





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

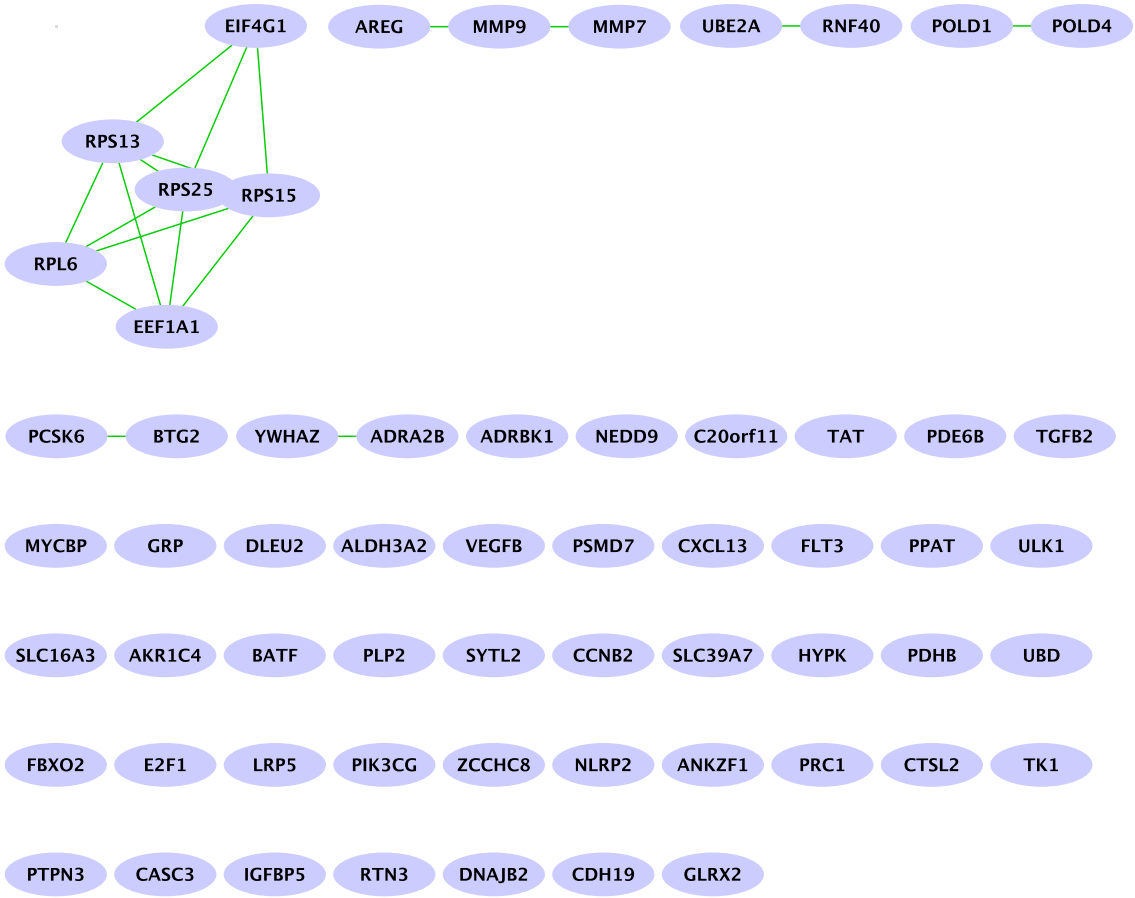
- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

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

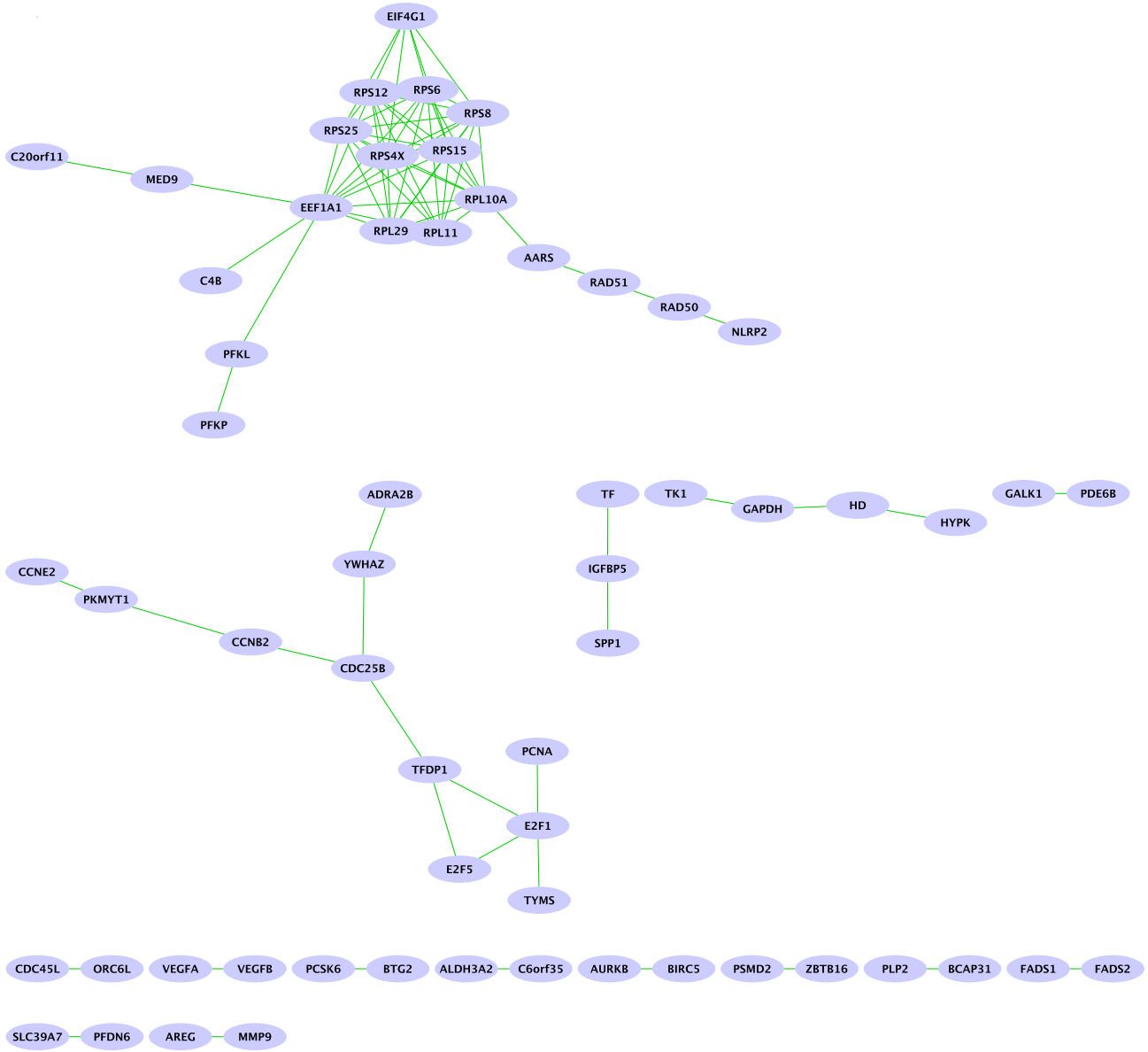
- Graph on the genes.

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

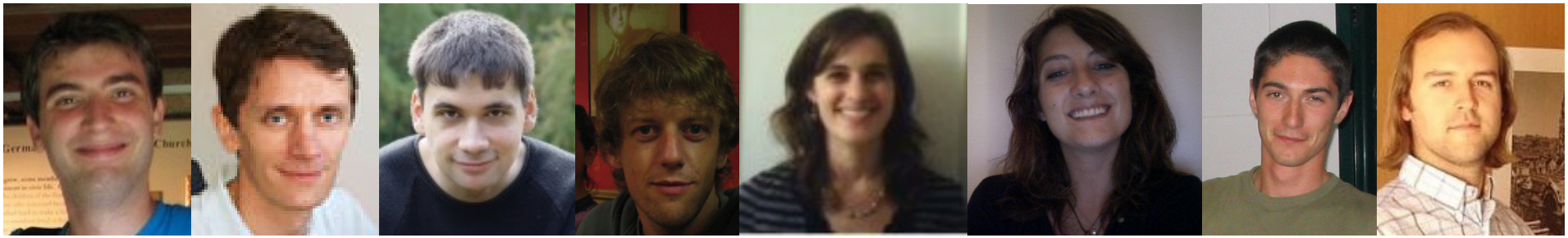


Outline

Conclusion

- Machine learning offers many powerful tools to learn predictive models from large sets of complex data
- Specific developments are required to solve complex problems that arise in bio-informatics
- Integration of prior knowledge in the penalization / regularization function is an efficient approach to fight the curse of dimension
- Requires interdisciplinary collaborations to incorporate expert knowledge at the heart of learning algorithms
- Many other applications not covered in this presentation!

Acknowledgements!



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