Machine learning applications to genomics

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)
Mines ParisTech – Institut Curie – INSERM U900
PSL Research University, Paris, France

BioComp Summer School 2017

http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott

Precision Medicine

- **Adapt** treatment to the **(genetic) specificities** of the patient.

  E.g. Trastuzumab for HER2+ breast cancer.

Precision Medicine

- **Adapt** treatment to the **(genetic) specificities** of the patient.

  E.g. Trastuzumab for HER2+ breast cancer.

- **Data-driven** biology/medicine

  Identify similarities between patients that exhibit similar phenotypes.

Precision Medicine

- **Adapt** treatment to the **(genetic) specificities** of the patient.

  E.g. Trastuzumab for HER2+ breast cancer.

- **Data-driven** biology/medicine

  Identify similarities between patients that exhibit similar phenotypes.

- **Biomarker discovery** = **Feature Selection**

Precision Medicine

- **Adapt** treatment to the **(genetic) specificities** of the patient.

  E.g. Trastuzumab for HER2+ breast cancer.

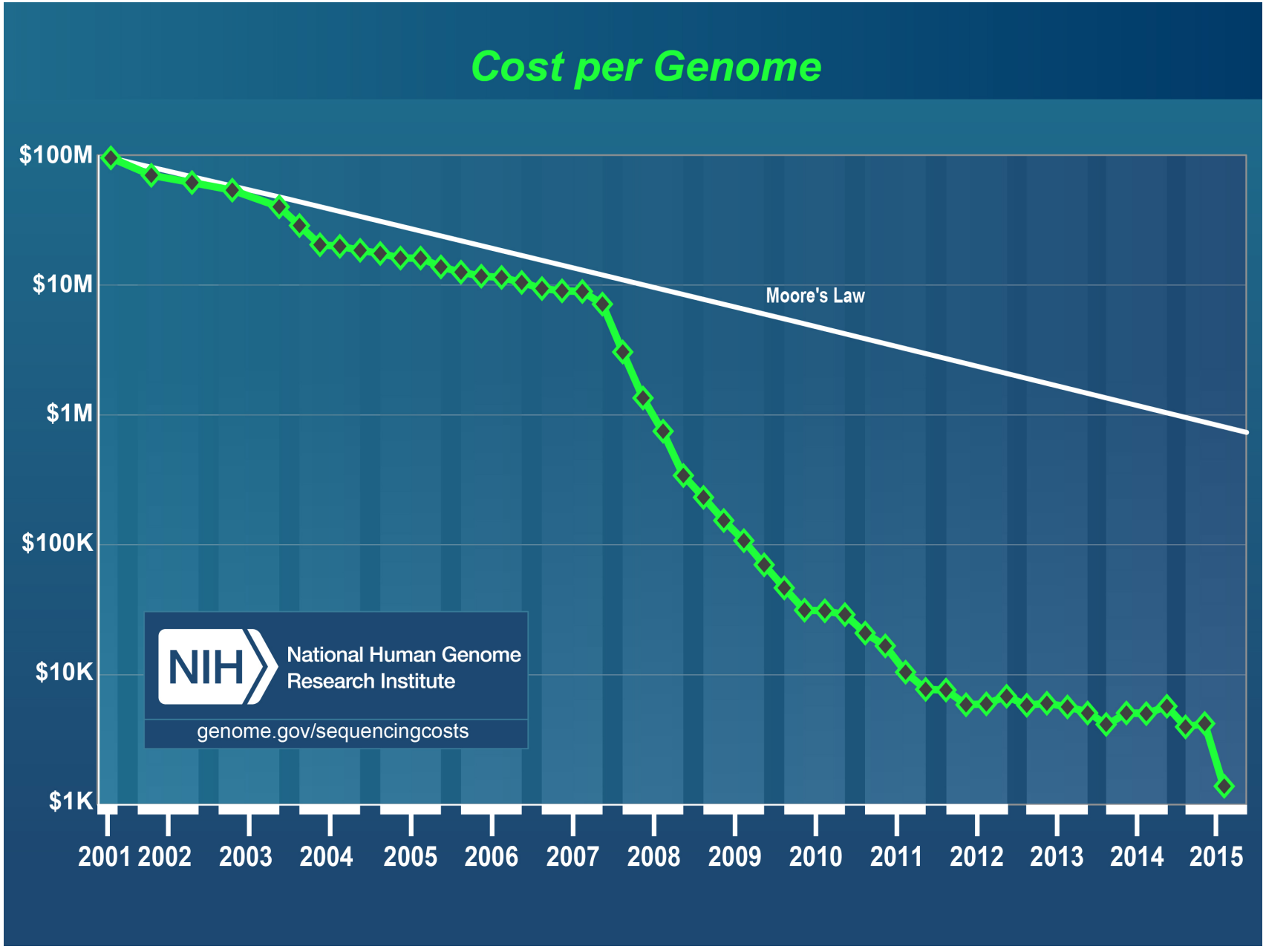
- **Data-driven** biology/medicine

  Identify similarities between patients that exhibit similar phenotypes.

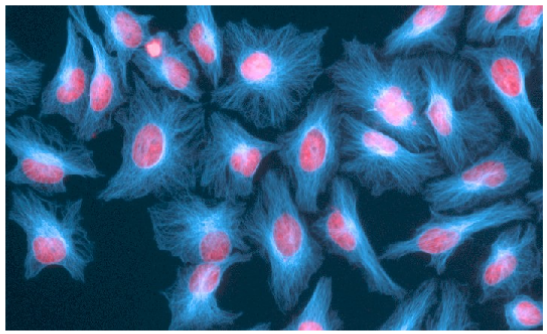
- **Biomarker discovery** = **Feature Selection**

- **Prediction**

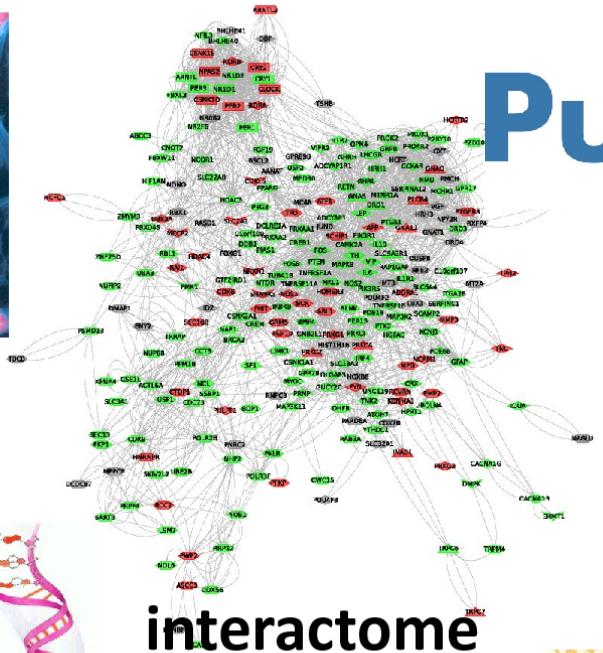
Sequencing costs



Big data!



phenome

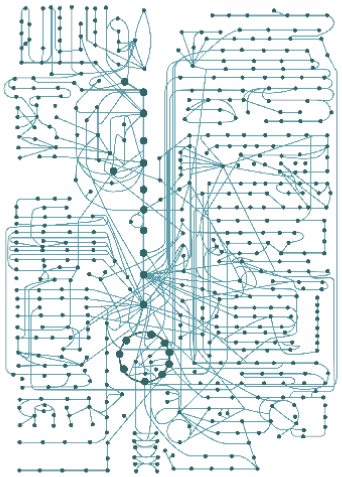


interactome

PubMed

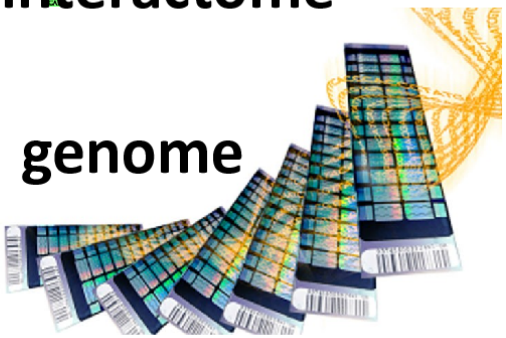
publications

transcriptome

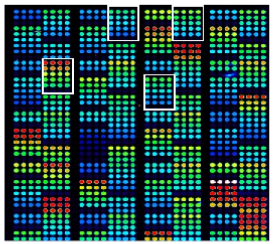


metabolome

methylome



genome



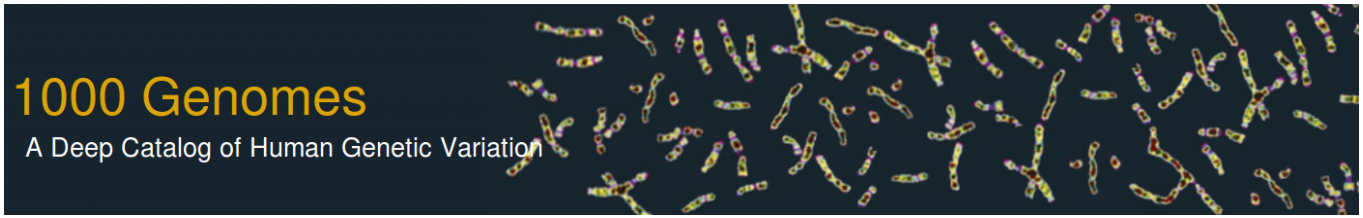
proteome

Image sources: ajc1@ flickr; Zlir'a@wikimedia

Big data!


NIH
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

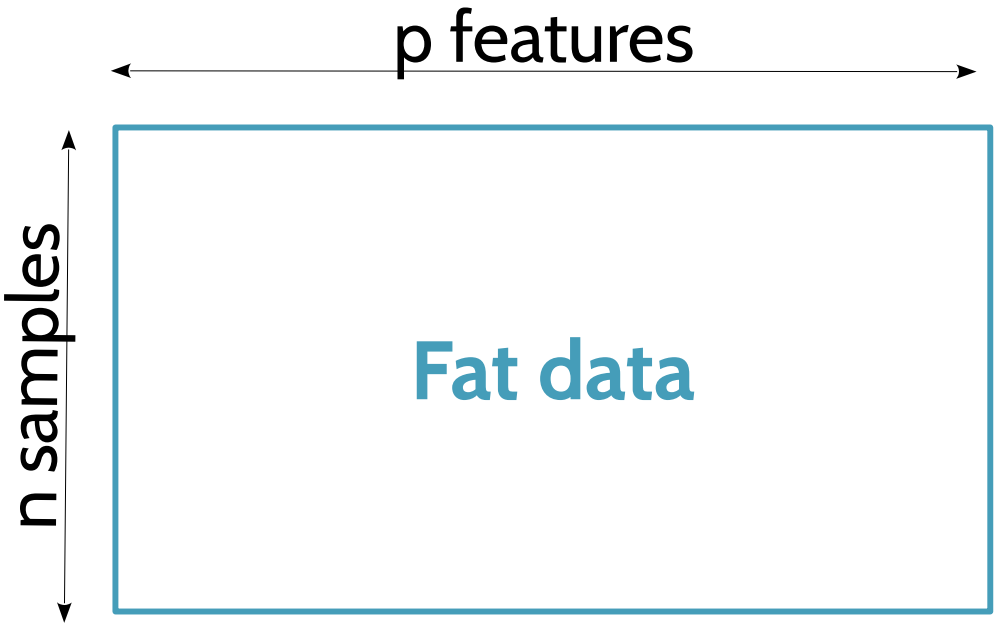

Cancer Genomics Hub
A resource of the National Cancer Institute


1000 Genomes
A Deep Catalog of Human Genetic Variation


DREAM Challenges


SAY BIG DATA
ONE MORE TIME
memegenera

Big data vs. fat data



Challenges of high-dimensional data

- **Computational challenges** for some algorithms

  Linear regression: inverting $X^\top X$ takes $\mathcal{O}(p^3)$ computations.

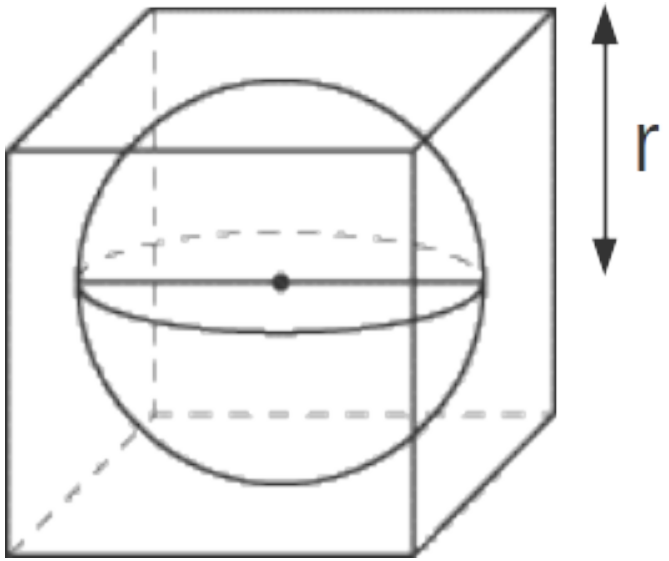
- The **curse of dimensionality** makes it hard to learn
- **Overfitting** is more likely
- **Ill-posed problems.**

The curse of dimensionality

- Methods and intuitions that work in low dimension **might not apply to higher dimensions.**
- **Hyperspace is very big and everything is far apart**

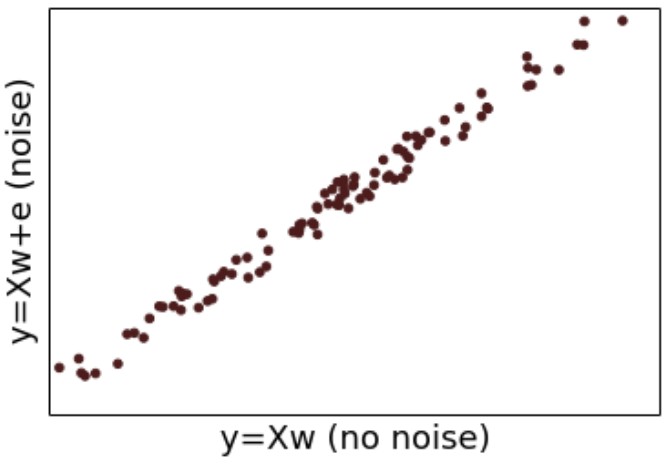
  Fraction of the points within a cube that fall outside the inscribed circle:

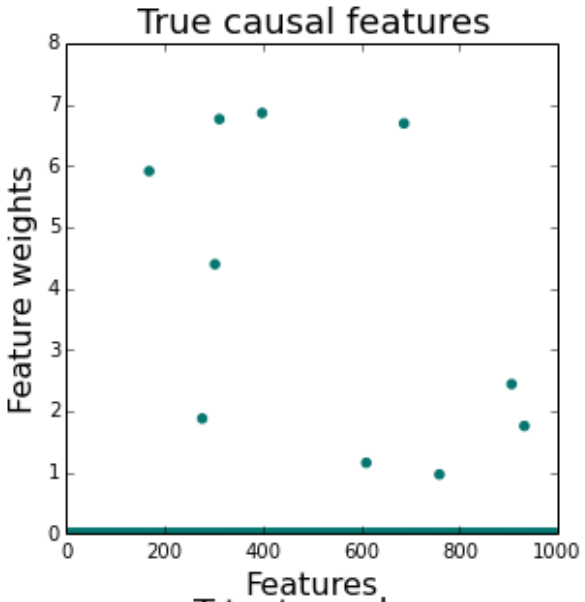
  - In two dimensions: $1 - \frac{\pi r^2}{4r^2} = 1 - \frac{\pi}{4}$
  - In three dimensions: $1 - \frac{4/3\pi r^3}{8r^3} = 1 - \frac{\pi}{6}$
  - In higher dimension: tends towards 1.

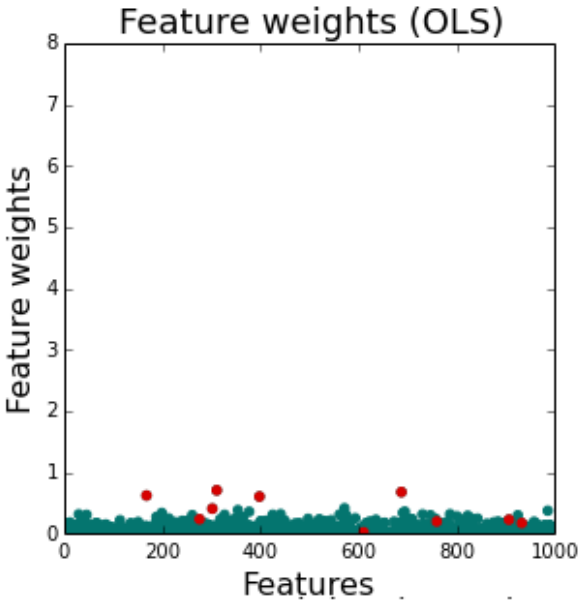


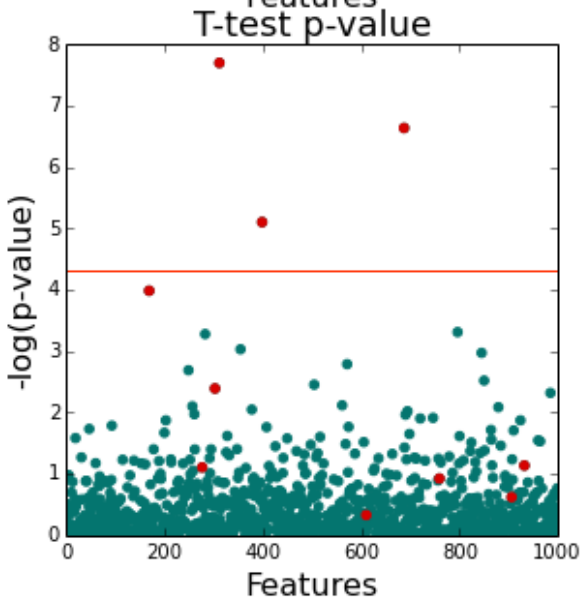
Large p, small n data

Simulation: n=100, p=1000, 10 causal features, y = Xw + e.





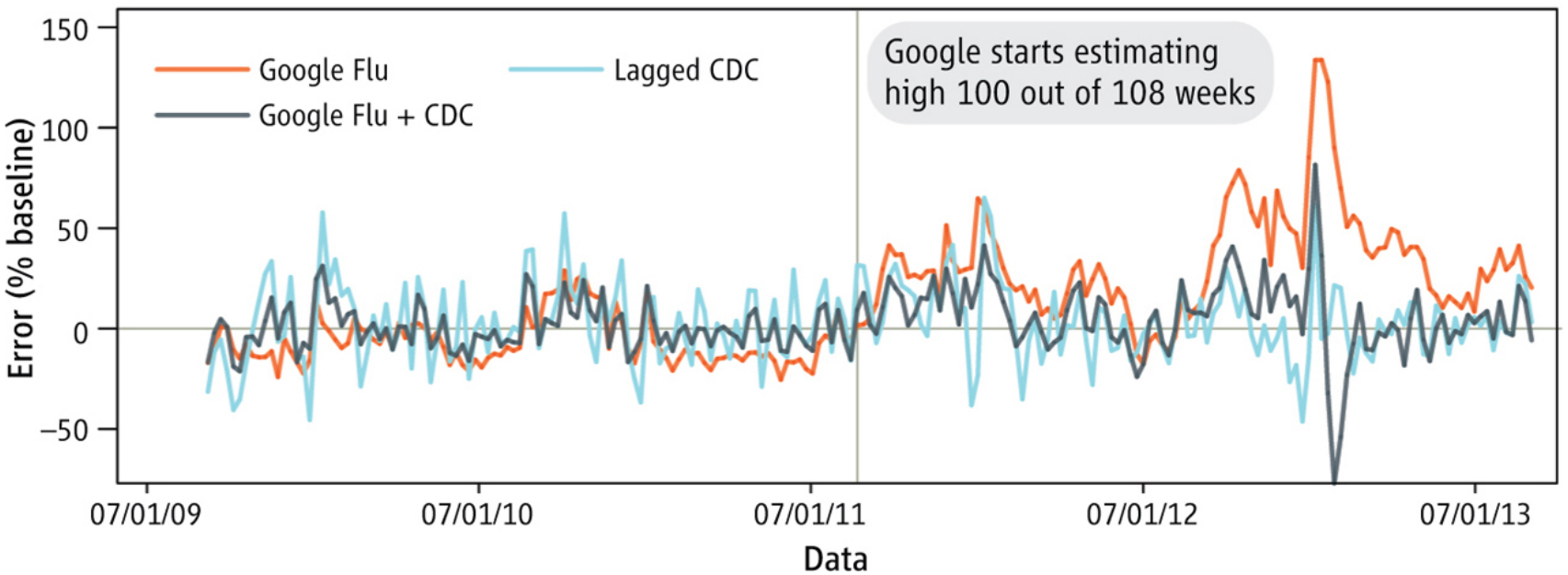




Google Flu Trends

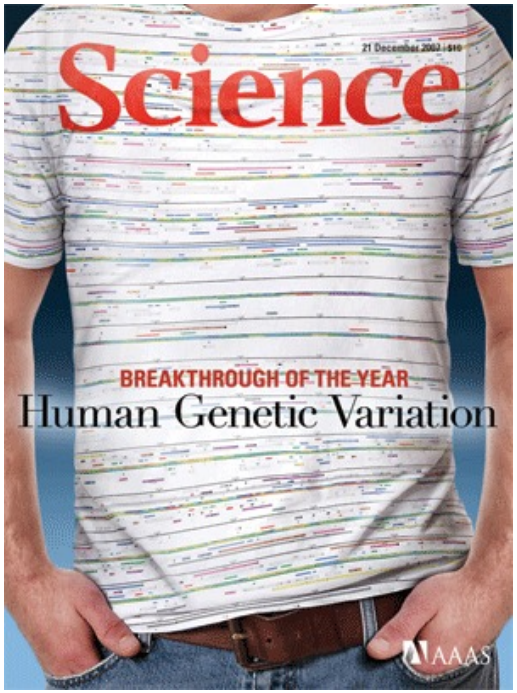
D. Lazer, R. Kennedy, G. King and A. Vespignani. **The Parable of Google Flu: Traps in Big Data Analysis.** Science 2014 [Lazer et al. 2014]

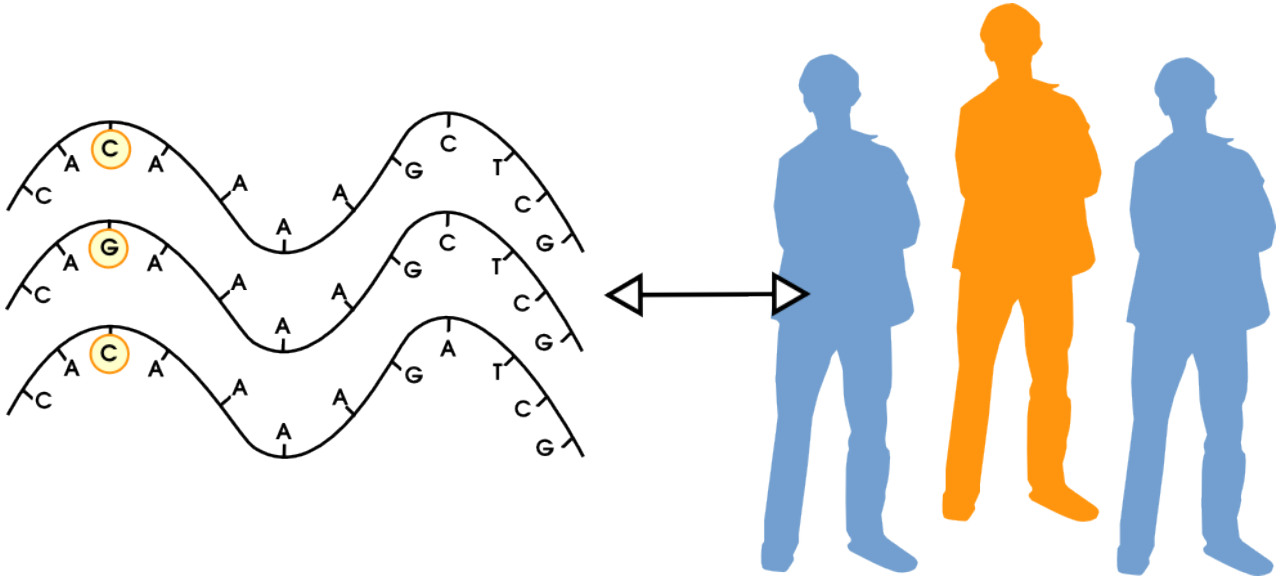
- **p = 50 million** search terms
- **n = 1152** data points



- Predictive search terms include terms related to high-school basketball.

GWAS: Genome-Wide Association Studies





Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ Single Nucleotide Polymorphisms (SNPs)

$n = 10^2 - 10^4$ samples

[Pennisi 2007]

Qualitative GWAS

Binary phenotype, i.e. case/controls encoded as 1/0.

- **Contingency table**

	AA	Aa	aa
Cases			
Ctrls			

	0	1
Cases	a	b
Ctrls	c	d

Statistical tests: χ^2, Cochran-Armitage trend test, etc.

- **Logistic regression**

$$\text{logit}(p(y|X)) = \beta_0 + \beta_1 X$$

- **Odds-ratio**

$$\underbrace{\frac{\text{P(0|case)}}{\text{P(0|ctrl)}}}_{\text{odds of 0 in cases}} \Big/ \underbrace{\frac{\text{P(1|case)}}{\text{P(1|ctrl)}}}_{\text{odds of 1 in cases}} = \frac{ad}{bc}$$

Quantitative GWAS

- **Linear regression**

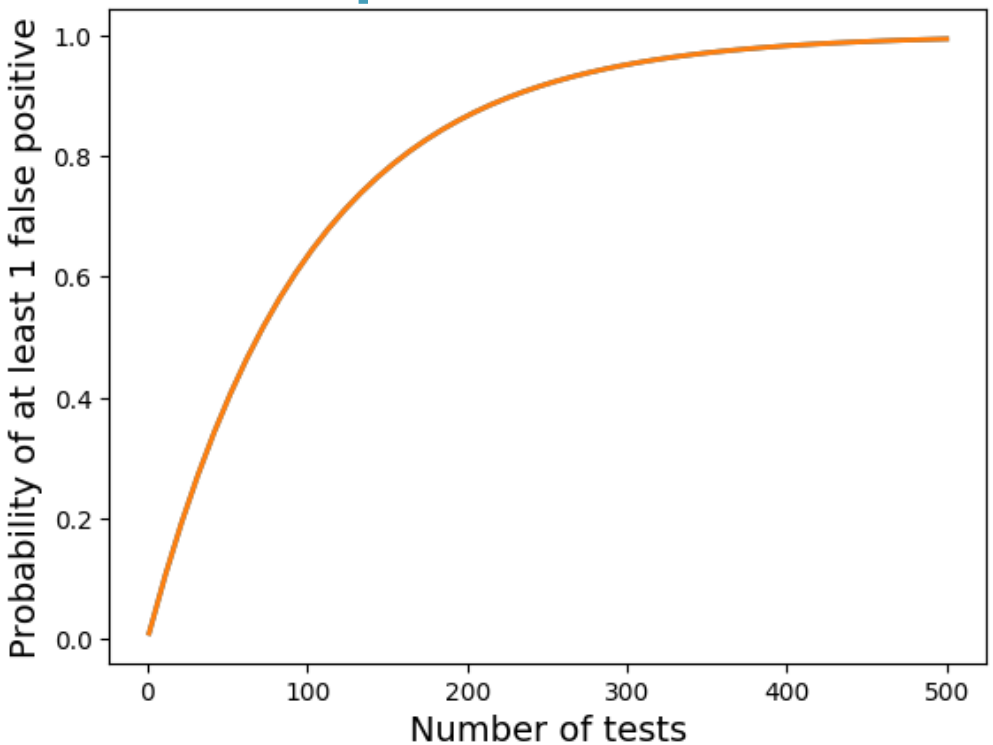
$$y = \beta_0 + \beta_1 X$$

- p-value: Is $\hat{\beta}_1$ significantly different from 0?

  Wald test: compare $\frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)}$ to a χ^2 distribution.
- **Effect size:** β_1.

Multiple Hypothesis Testing

- Probability of having **at least one false positive:**
  - For **one** test: α
  - For **p** tests: $1 - (1 - \alpha)^p$



- Controlling **Family-Wise Error Rate** (FWER)

  $\text{FWER} = P(|\text{FP}| \geq 1)$

  FP = number of false positives (Type I errors)

- **Bonferroni** correction: $\alpha \rightarrow \frac{\alpha}{p}$

GWAS discoveries

- **The GWAS Catalog** `https://www.ebi.ac.uk/gwas/`
- **Clinical benefits:** Ankylosing spondylitis
  - Role of interleukine 17 pathway
  - Consentyx (**secukinumab**), approved January 15, 2016.
- SNPs associated with **drug resistance**

  in e.g. *Myobacterium tuberculosis*, *Staphylococcus aureus*, *Streptococcus pneumoniae* or HIV.

Ref: [Visscher et al. 2012; Manolio 2013; Power, Parkhill, and Oliveira 2017]

Missing heritability

GWAS **fail to explain** most of the **inheritable variability** of complex traits.

Many possible reasons:

- non-genetic / non-SNP factors
- heterogeneity of the phenotype
- rare SNPs
- weak effect sizes
- **few samples in high dimension ($p \gg n$)**
- **joint effets of multiple SNPs.**

Ref: [Manolio et al. 2009]

What solution does machine learning provide?

Reducing p

Dimensionality reduction

- **Feature selection:** Keep relevant features only
  - **Filter approaches:** Apply a statistical test to assign a score to each feature.
  - **Wrapper approaches:** Use a greedy search to find the "best" set of features for a given predictive model.
  - **Embedded approaches:** Fit a **sparse** model, i.e. that is encouraged to not use all the features.
- **Feature extraction:** Project the data on a new space
  - Creates new features, which makes interpretability harder.
  - **Matrice factorization techniques:** PCA, factorial analysis, NMF, kPCA.
  - **Manifold learning:** Multidimensional scaling, t-SNE.
  - **Autoencoders.**

Regularization

Control the number of parameters

- limit **model complexity**
- avoid **overfitting**

- Least-squares regression

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2}_{\text{loss}}$$

- **Lasso**  ani 1994]

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2}_{\text{loss}} + \lambda \underbrace{||\mathbf{w}||_1}_{\text{sparsity}}$$

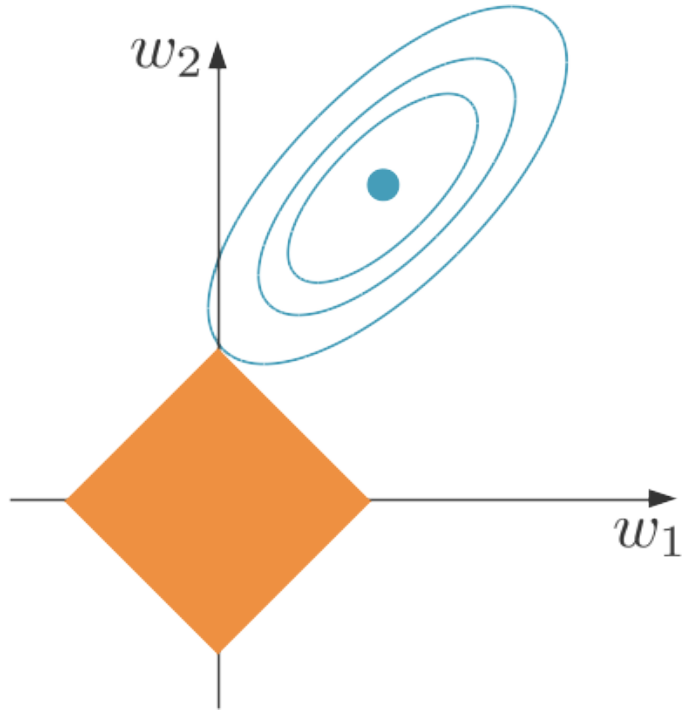
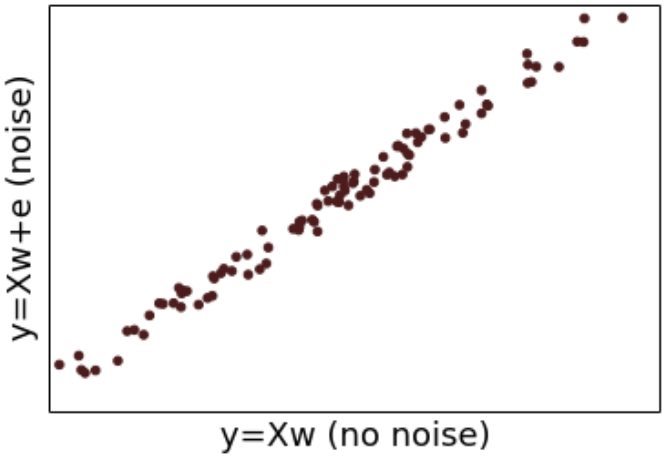
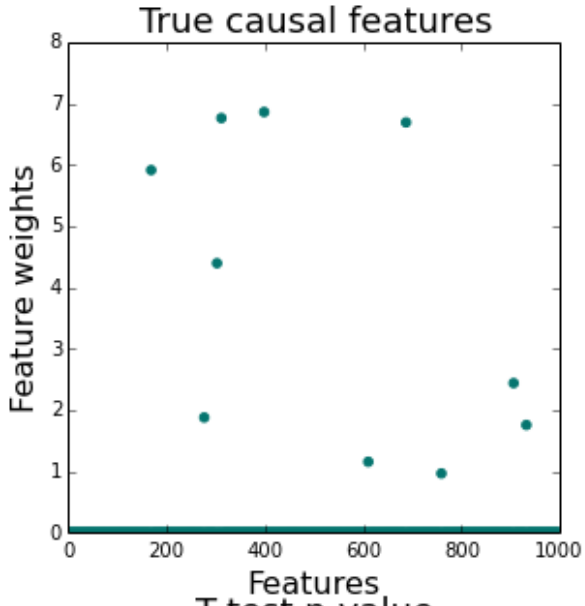


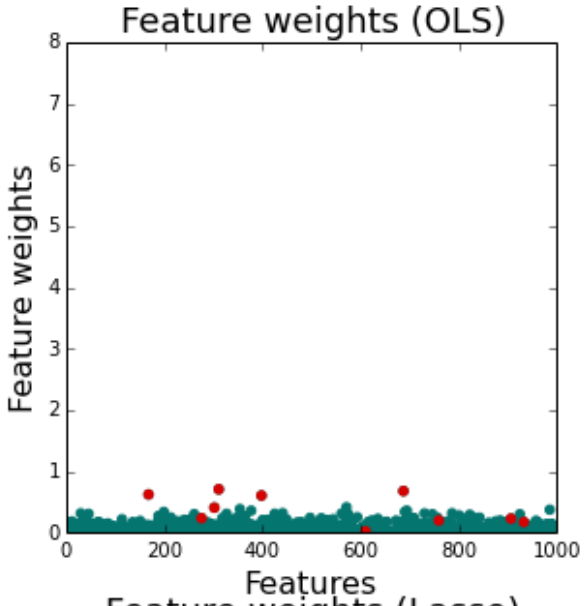
Image credits: Harold M. Walker via Wikimedia Commons

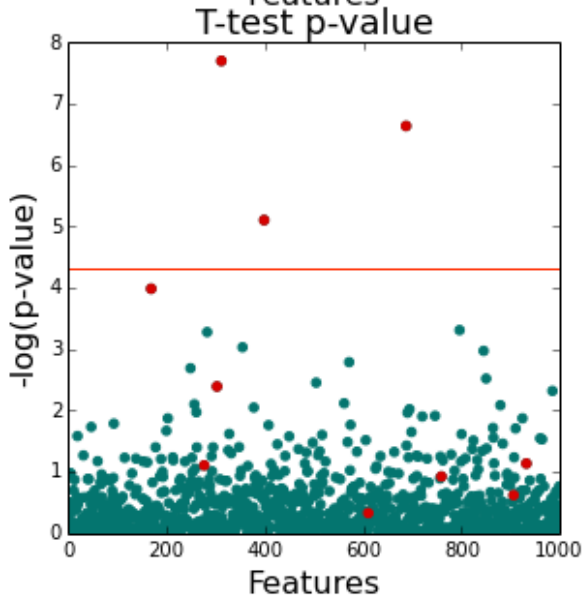
Large p, small n data

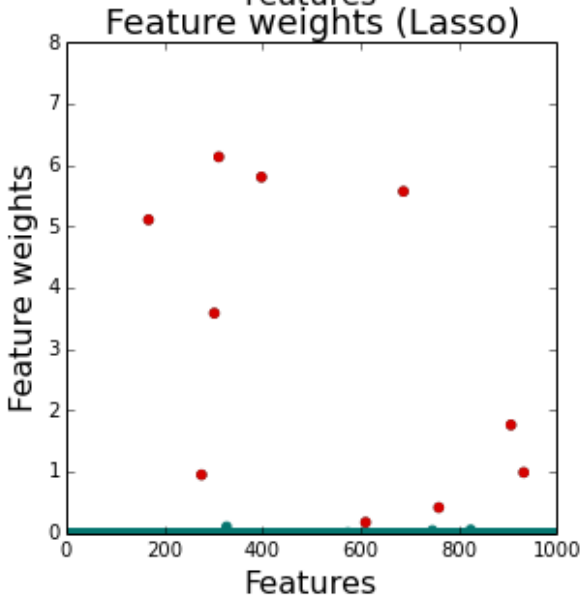
Simulation: n=100, p=1000, 10 causal features, y = Xw + e.











Integrating prior knowledge

Use prior knowledge as a constraint on the selected features

Prior knowledge can be represented as **structure:**

- Linear structure of DNA
- **Groups:** e.g. pathways
- **Networks** (molecular, 3D structure).

Original feature space

Constrained feature space

Elephant image by Danny Chapman @ Flickr.

Structured sparsity

- **Lasso:** Selecting variables in high dimension.

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \quad \underbrace{||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2}_{\text{loss}} + \lambda \underbrace{||\mathbf{w}||_1}_{\text{sparsity}}$$

- **Lasso+:** Structured regularizer.

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \quad \underbrace{||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(\mathbf{w})}_{\text{structure}}$$

Image by Harold M. Walker via Wikimedia Commons

Structured regularizers

- **Group Lasso** [Yuan and Lin 2006]

$$\Omega_{\text{group}}(\mathbf{w}) = \sum_{g \in \mathcal{G}} ||\mathbf{w}_g||$$

- **Overlapping Group Lasso** [Jacob, Obozinski, and Vert 2009]

$$\Omega_{\text{overlap}}(\mathbf{w}) = \sum_{\mathbf{v} \in \mathcal{V}_{\mathcal{G}}, \sum_{g \in \mathcal{G}} \mathbf{v}_g = \mathbf{w}} ||\mathbf{v}_g||$$

  $\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}_g$ and $\text{supp}(\mathbf{v}_g) \subset g$.

  Graph Lasso: groups = edges.

- **ncLasso / Grace** [Li and Li 2008; Li and Li 2010]

$$\Omega_{\text{ncLasso}}(\mathbf{w}) = \mathbf{w}^\top L \mathbf{w}$$

  $L = D - W$: **Laplacian** of the graph of adjacency matrix W and degree matrix D.

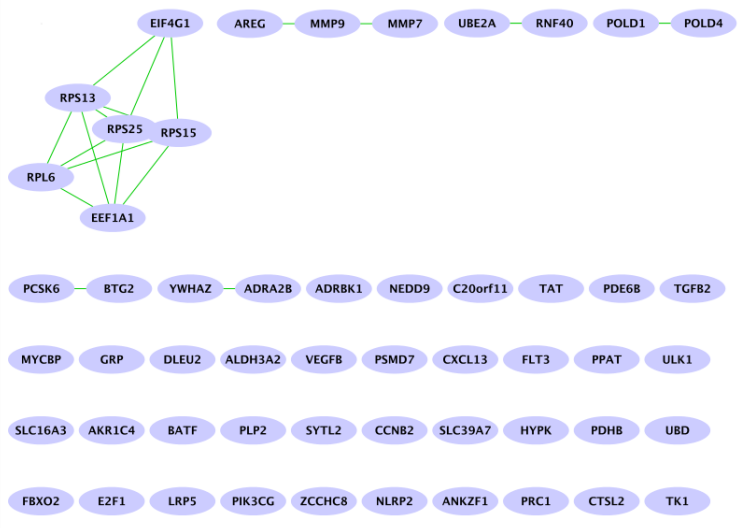
Gene selection with the Graph lasso

icml2009

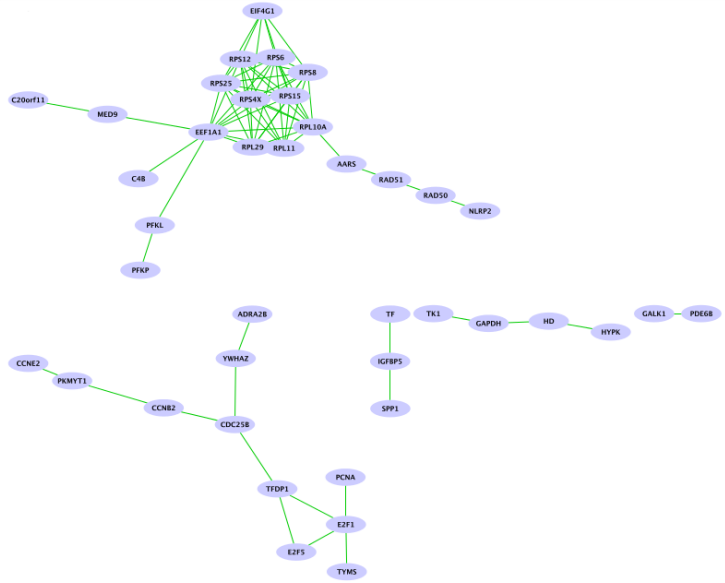
Group Lasso with Overlaps and Graph Lasso

Laurent Jacob, Guillaume Obozinski and Jean-Philippe Vert

Lasso



Graph Lasso



Ref: [Jacob, Obozinski, and Vert 2009]

Regularized relevance

Set $\mathcal{V}$ of p variables.

- **Relevance score** $R : 2^{\mathcal{V}} \to \mathbb{R}$

  Quantifies the importance of any subset of variables for the question under consideration.

  Ex : correlation, HSIC, statistical test of association.

- **Structured regularizer** $\Omega : 2^{\mathcal{V}} \to \mathbb{R}$

  Promotes a sparsity pattern that is compatible with the constraint on the feature space.

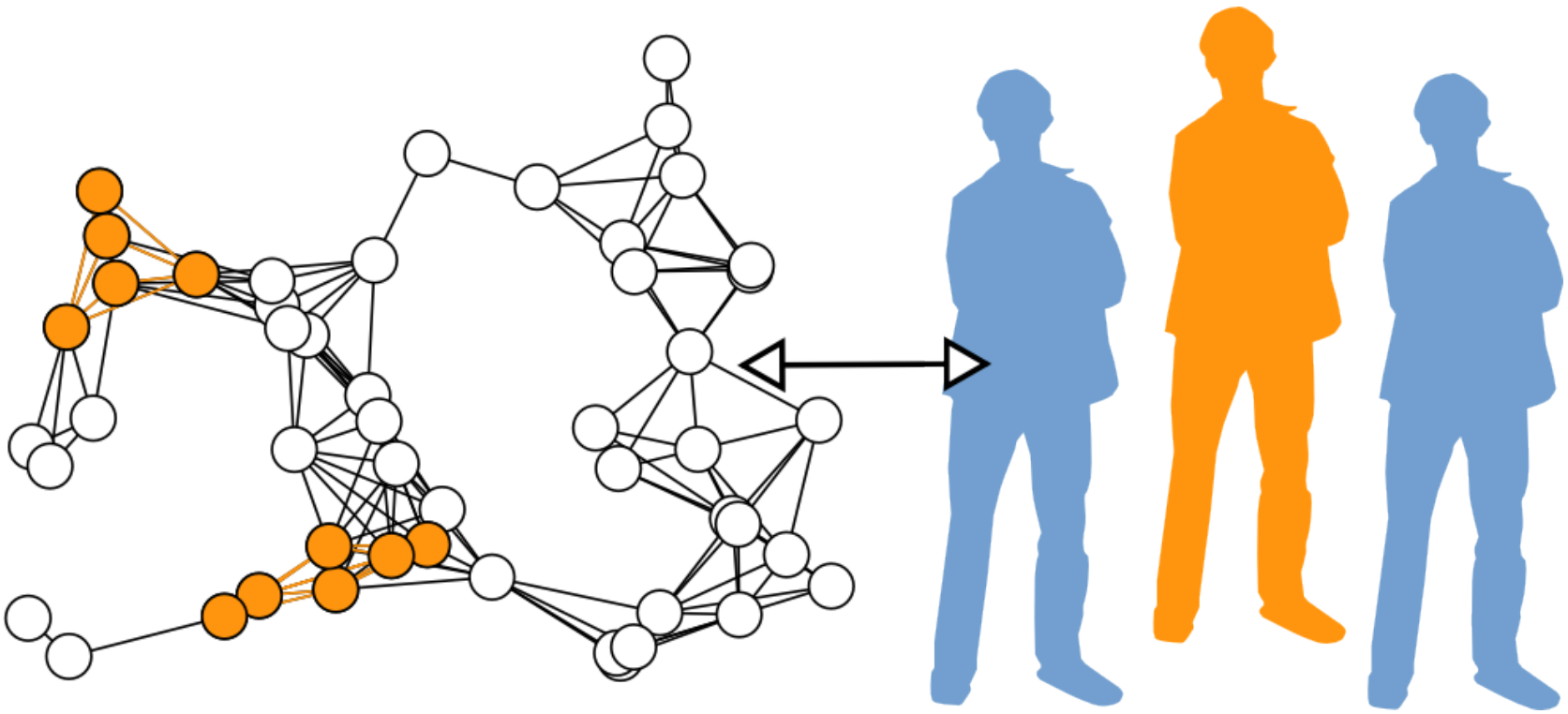
  Ex : cardinality $\Omega : \mathcal{S} \mapsto |\mathcal{S}|$.

- **Regularized relevance**

$$\arg\max_{\mathcal{S} \subseteq \mathcal{V}} R(\mathcal{S}) - \lambda \Omega(\mathcal{S})$$

Network-guided multi-locus GWAS

Goal: Find a **set of explanatory SNPs** compatible with a **given network** structure.



Network-guided GWAS

- **Additive test of association** SKAT [Wu, Lee, et al. 2011]

$$R(\mathcal{S}) = \sum_{i \in \mathcal{S}} c_i \qquad c_i = (\mathbf{X}^\top (\mathbf{y} - \mu))_i^2$$

- **Sparse Laplacian regularization**

$$\Omega : \mathcal{S} \mapsto \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij} + \alpha |\mathcal{S}|$$

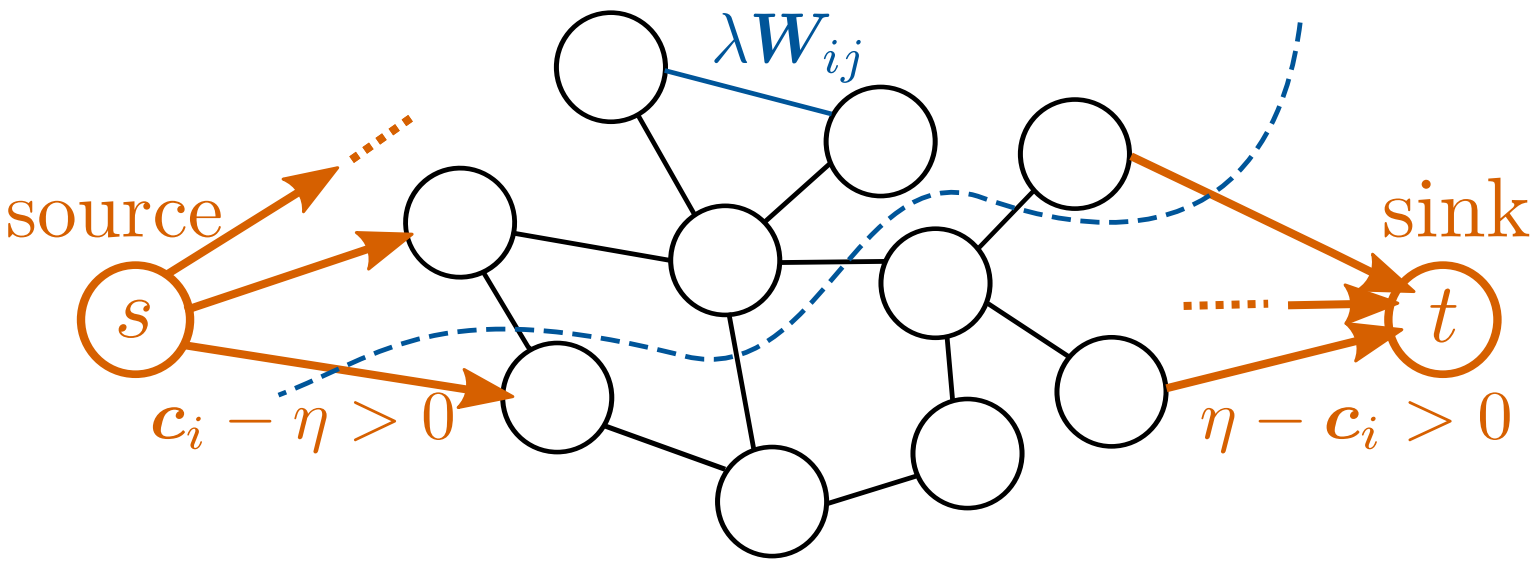
- **Regularized maximization of** R

$$\underset{\mathcal{S} \subseteq \mathcal{V}}{\arg\max} \underbrace{\sum_{i \in \mathcal{S}} c_i}_{\text{association}} - \underbrace{\eta\,|\mathcal{S}|}_{\text{sparsity}} - \underbrace{\lambda \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij}}_{\text{connectivity}}$$

Minimum cut reformulation

The graph-regularized maximization of score $Q(*)$ is equivalent to a s/t-min-cut for a graph with adjacency matrix $\mathbf{A}$ and two additional nodes s and t, where $\mathbf{A}_{ij} = \lambda \mathbf{W}_{ij}$ for $1 \leq i, j \leq p$ and the weights of the edges adjacent to nodes s and t are defined as

$$\mathbf{A}_{si} = \begin{cases} c_i - \eta & \text{if } c_i > \eta \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{A}_{it} = \begin{cases} \eta - c_i & \text{if } c_i < \eta \\ 0 & \text{otherwise} \end{cases}.$$



SConES: Selecting **Con**nected **E**xplanatory **S**NPs.

Comparison partners

- **Univariate linear regression** $y_k = \alpha_0 + \beta \mathbf{G}_k^i$
- **Lasso**

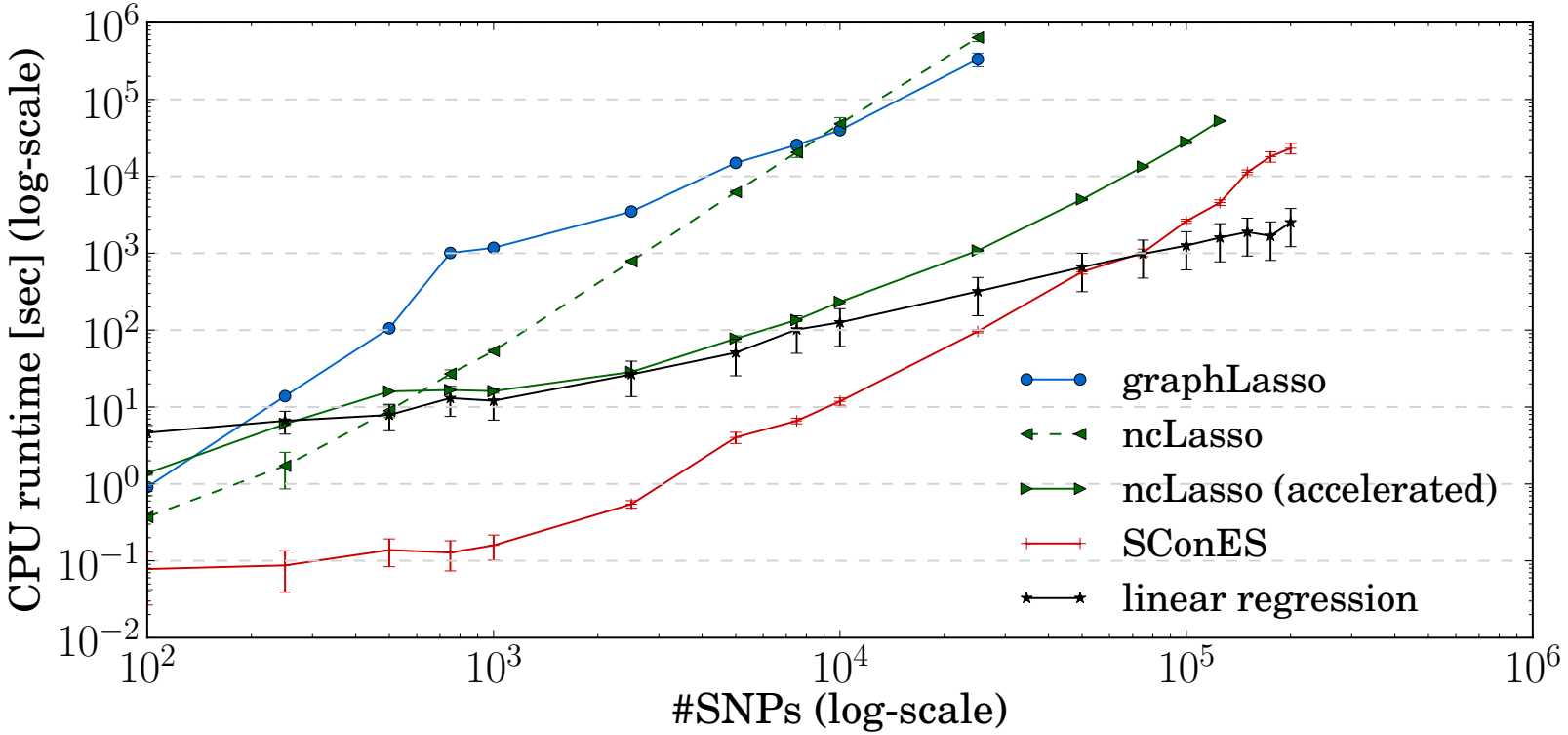
$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \underbrace{\frac{1}{2} ||\mathbf{y} - \mathbf{G}\beta||_2^2}_{\text{loss}} + \underbrace{\eta \ ||\beta||_1}_{\text{sparsity}}$$

- **Feature selection with sparsity and connectivity constraints**

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \underbrace{\mathcal{L}(\mathbf{y}, \mathbf{G}\beta)}_{\text{loss}} + \underbrace{\eta \ ||\beta||_1}_{\text{sparsity}} + \underbrace{\lambda \, \Omega(\beta)}_{\text{connectivity}}$$

  - **ncLasso**: network connected Lasso [Li and Li 2008]
  - Overlapping group Lasso [Jacob, Obozinski, and Vert 2009]
    - **groupLasso**: E.g. SNPs near the same gene grouped together
    - **graphLasso**: 1 edge = 1 group.

Runtime



$n = 200$ exponential random network (2 % density)

Data simulation

Arabidopsis thaliana genotypes
[Atwell et al. 2010; Segura et al. 2012]

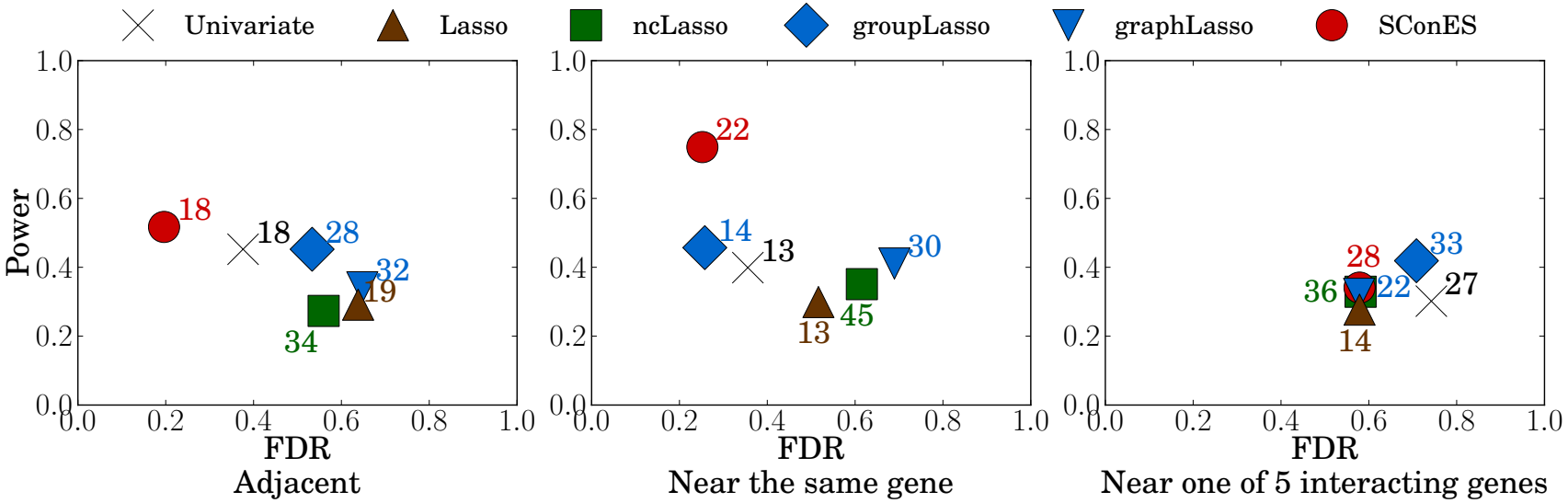
$n = 500$ samples, $p = 1,000$ SNPs

TAIR **Protein-Protein Interaction data**
$\rightarrow \quad \sim 50 \times 10^6$ edges

20 causal SNPs: $y = \omega^\top x + \epsilon$

- Causal SNPs **adjacent in the genomic sequence**
- Causal SNPs **near the same gene**
- Causal SNPs **near any of 2–5 interacting genes**

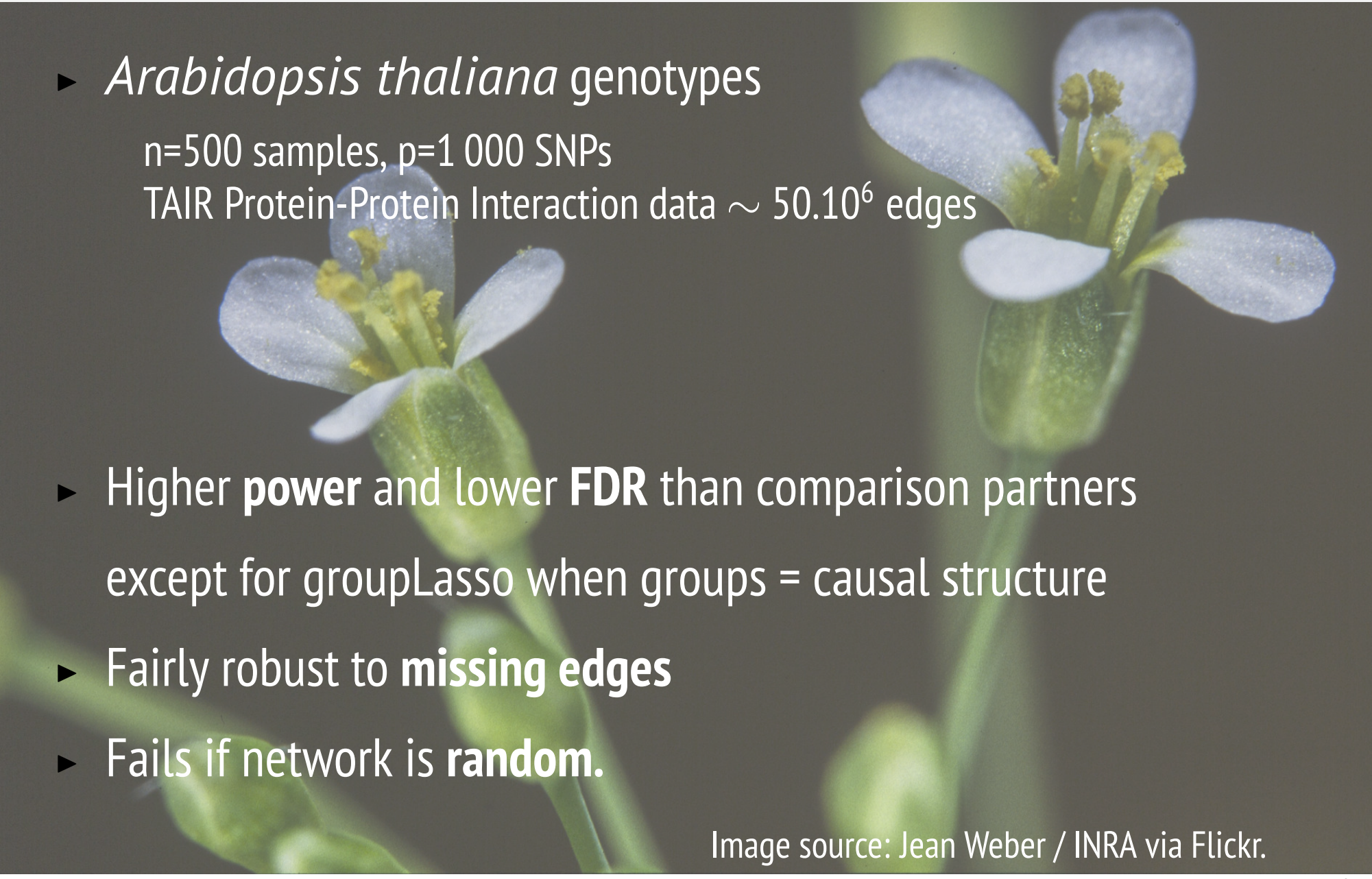
Performance on simulated data



$$\text{Power} = \text{precision} = \frac{\text{SNPs selected and causal}}{\text{causal SNPs}}$$

$$\text{False Discovery Rate (FDR)} = 1 - \text{recall} = \frac{\text{SNPs selected and non-causal}}{\text{selected SNPs}}$$

Experiments: Performance on simulated data



- *Arabidopsis thaliana* genotypes

  n=500 samples, p=1 000 SNPs
  TAIR Protein-Protein Interaction data $\sim 50.10^6$ edges

- Higher **power** and lower **FDR** than comparison partners

  except for groupLasso when groups = causal structure

- Fairly robust to **missing edges**
- Fails if network is **random.**

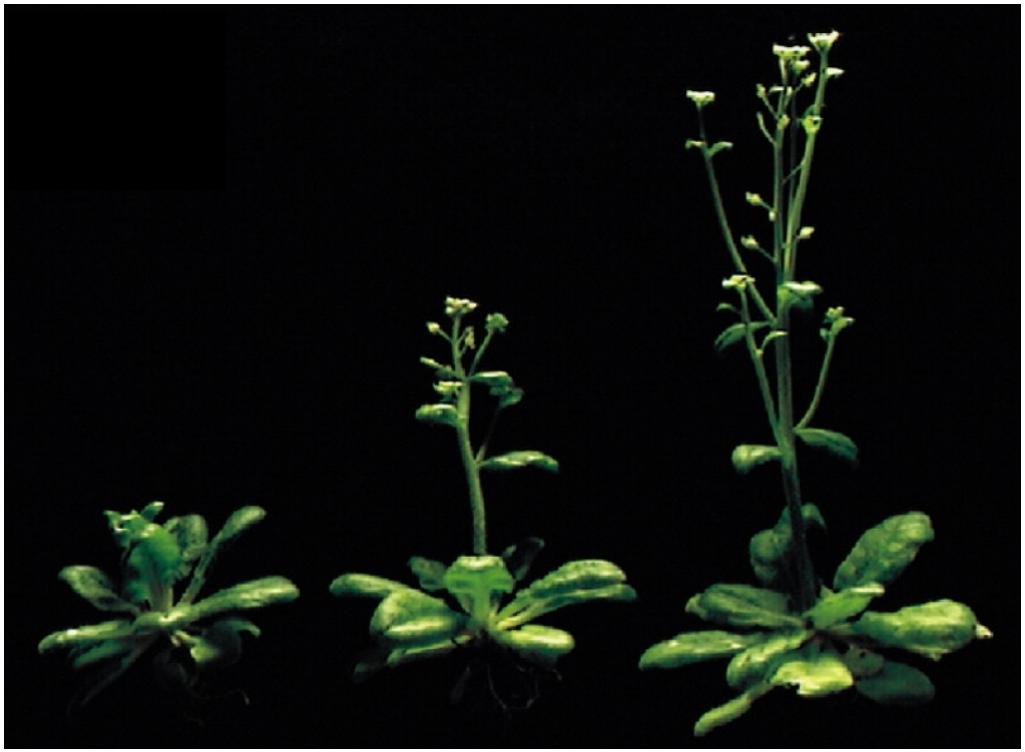
Image source: Jean Weber / INRA via Flickr.

Arabidopsis thaliana flowering time

17 flowering time phenotypes
[Atwell et al., Nature, 2010]

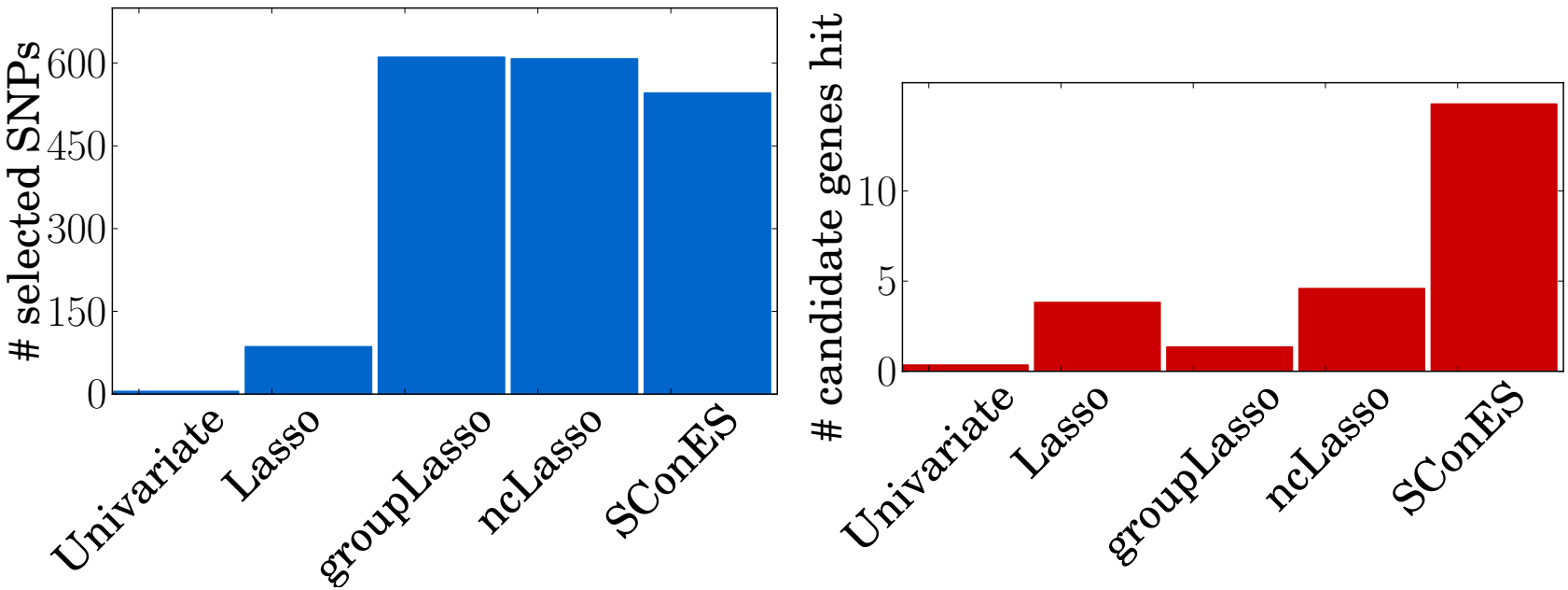
$p \sim 170\,000$ SNPs
(after MAF filtering)
$n \sim 150$ samples

165 **candidate genes**
[Segura et al., Nat Genet 2012]



Correction for **population structure**: regress out PCs.

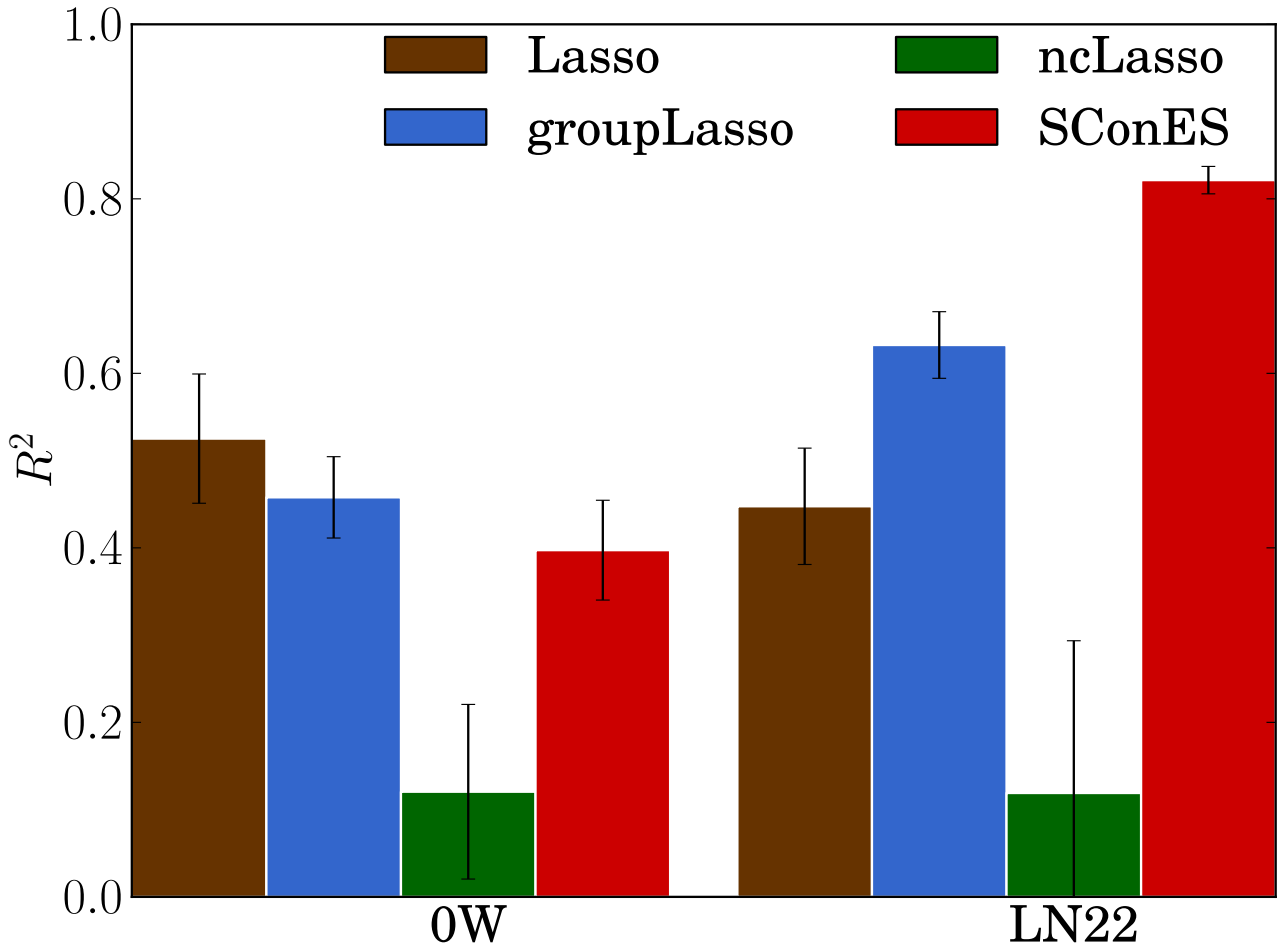
Arabidopsis thaliana flowering time



- SConES selects **about as many SNPs** as other network-guided approaches but **detects more candidates.**

Arabidopsis thaliana flowering time

Predictivity of selected SNPs



SConES: Selecting Connected Explanatory SNPs

- selects connected, explanatory SNPs;
- incorporates large networks into GWAS;
- is efficient, effective and robust.

C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara and K. Borgwardt (2013) **Efficient network-guided multi-locus association mapping with graph cuts**, Bioinformatics 29 (13), i171–i179 doi:10.1093/bioinformatics/btt238 [Azencott et al. 2013]

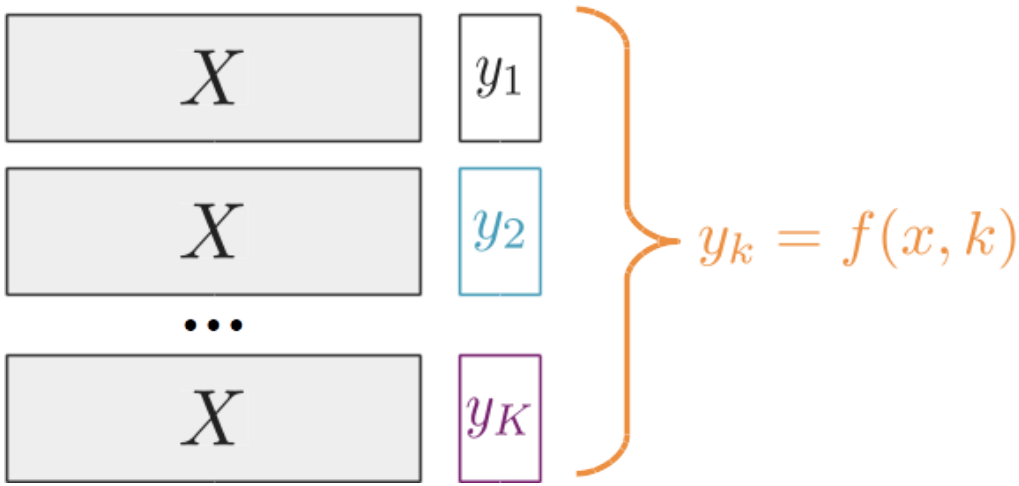
`https://github.com/chagaz/scones`

`https://github.com/chagaz/sfan`

Increasing n

Multi-phenotype GWAS

Increase sample size by **jointly** performing GWAS for **multiple related phenotypes**

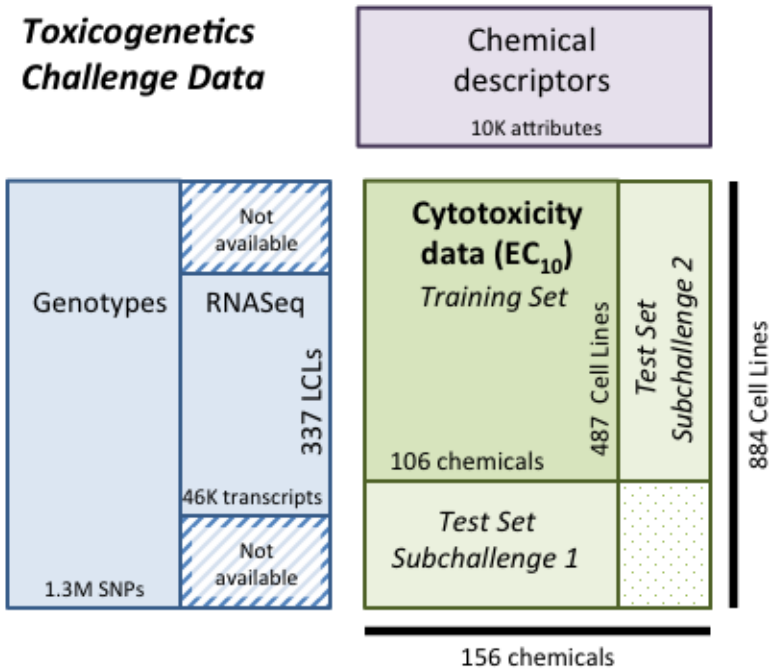


Multi-task feature selection

Transfer learning

Toxicogenetics / Pharmacogenomics

Tasks (phenotypes) = chemical compounds



F. Eduati, L. Mangravite, et al. (2015) **Prediction of human population responses to toxic compounds by a collaborative competition.** Nature Biotechnology, 33 (9), 933–940 doi: 10.1038/nbt.3299

Multi-SConES

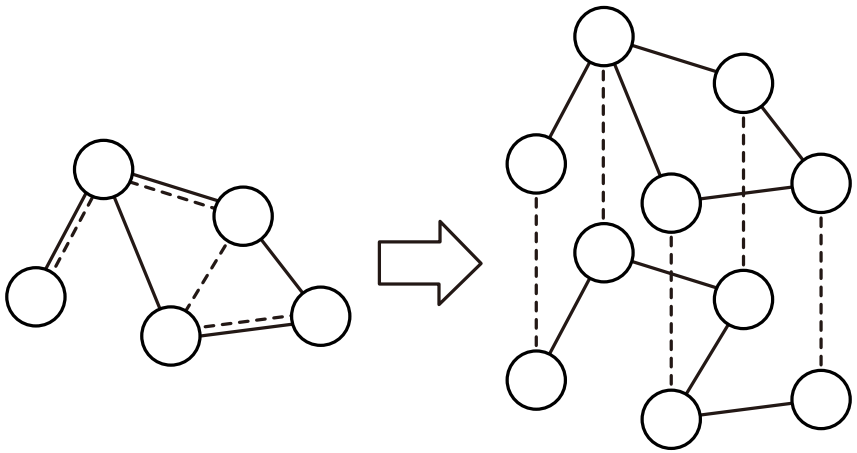
T **related phenotypes.**

- Goal: obtain **similar sets of features** on related tasks.

$$\underset{\mathcal{S}_1,\ldots,\mathcal{S}_T \subseteq \mathcal{V}}{\arg\max} \sum_{t=1}^{T} \left(\sum_{i \in \mathcal{S}} c_i - \eta \left| \mathcal{S} \right| - \lambda \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij} - \sum_{u > t} \underbrace{\mu \left| \mathcal{S}_u \,\Delta\, \mathcal{S}_t \right|}_{\text{task sharing}} \right)$$

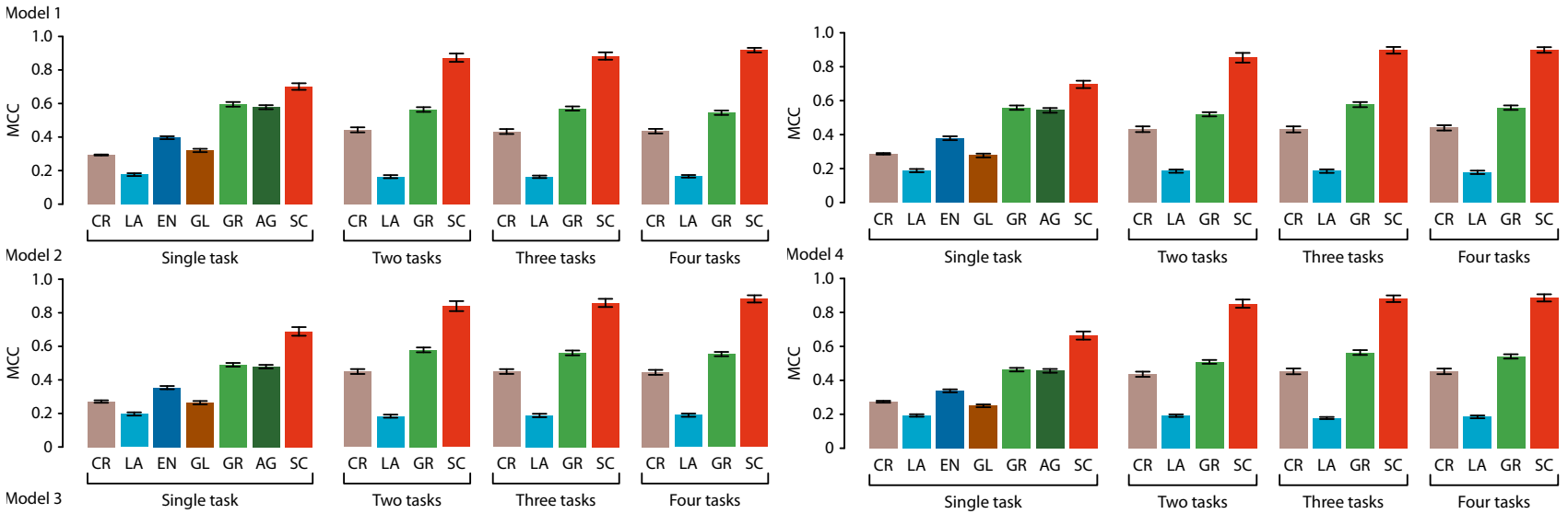
$$\mathcal{S} \,\Delta\, \mathcal{S}' = (\mathcal{S} \cup \mathcal{S}') \setminus (\mathcal{S} \cap \mathcal{S}') \qquad \text{(symmetric difference)}$$

- Can be reduced to single-task by building a **meta-network.**



Multi-SConES: Multiple related tasks

Simulations: retrieving causal features



M. Sugiyama, C.-A. Azencott, D. Grimm, Y. Kawahara and K. Borgwardt (2014) **Multi-task feature selection on multiple networks via maximum flows**, SIAM ICDM, 199–207 doi:10.1137/1.9781611973440.23

https://github.com/mahito-sugiyama/Multi-SConES

https://github.com/chagaz/sfan

Using task similarity

Use **prior knowledge** about the **relationship** between the tasks: $\Omega \in \mathbb{R}^{T \times T}$

$$\underset{\mathcal{S}_1,\ldots,\mathcal{S}_T \subseteq \mathcal{V}}{\arg\max} \sum_{t=1}^{T} \left(\sum_{i \in \mathcal{S}} c_i - \eta \left| \mathcal{S} \right| - \lambda \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij} - \underbrace{\mu \sum_{u=1}^{T} \sum_{i \in \mathcal{S}_t \cap \mathcal{S}_u} \Omega_{tu}^{-1}}_{\text{task sharing}} \right)$$

Can also be mapped to a meta-network.

Code: `http://github.com/chagaz/sfan`

Multiplicative Multitask Lasso with Task Descriptors

- **Multitask Lasso** [Obozinski, Taskar, and Jordan 2006]

$$\underset{\beta \in \mathbb{R}^{T \times p}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \beta_i g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda \sum_{i=1}^{p} ||\beta_i||_2}_{\text{task sharing}}$$

- **Multilevel Multitask Lasso** [Swirszcz and Lozano 2012]

$$\underset{\theta \in \mathbb{R}_+^p, \gamma \in \mathbb{R}^{T \times p}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \theta_i \gamma_i^t g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{i=1}^{p} \sum_{t=1}^{T} |\gamma_i^t|}_{\text{task sharing}}$$

- **Multiplicative Multitask Lasso with Task Descriptors** [Bellon, Stoven, and Azencott 2016]

$$\underset{\theta \in \mathbb{R}_+^p, \alpha \in \mathbb{R}^{p \times L}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \theta_i \left(\sum_{l=1}^{L} \alpha_{il} d_l^t\right) g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{i=1}^{p} \sum_{l=1}^{L} |\alpha_{il}|}_{\text{task sharing}}$$

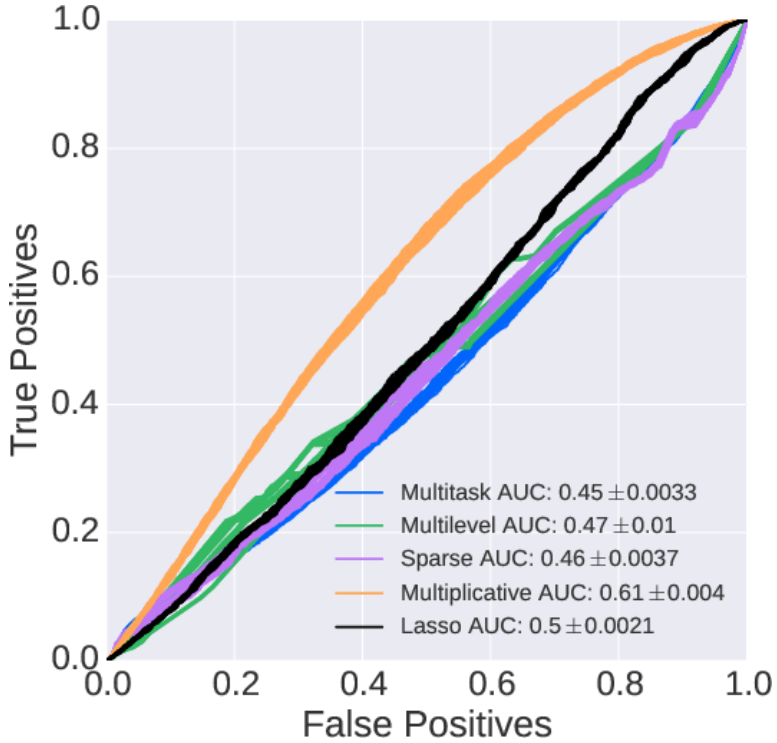
Multiplicative Multitask Lasso with Task Descriptors

$$\underset{\theta \in \mathbb{R}_+^p, \alpha \in \mathbb{R}^{p \times L}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \theta_i \left(\sum_{l=1}^{L} \alpha_{il} d_l^t\right) g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{i=1}^{p} \sum_{l=1}^{L} |\alpha_{il}|}_{\text{task sharing}}$$

- On **simulations:**
  - **Sparser** solution
  - Better **recovery of true features** (higher PPV)
  - Improved **stability**
  - Better **predictivity** (RMSE).

Multiplicative Multitask Lasso with Task Descriptors

- Making predictions for tasks for which you have **no data**.



V. Bellón, V. Stoven, and C.-A. Azencott (2016) **Multitask feature selection with task descriptors**, PSB.

https://github.com/vmolina/MultitaskDescriptor

Limitations of current approaches

- **Robustness/stability**
  - Doo we recover the same SNPs when the data changes slightly?
  - Mixed ℓ_1/ℓ_2 norms: Elastic Net [Zou and Hastie 2005].
- **Complex interaction patterns**
  - Limited to additive or quadrative effects (limited power) [Niel et al. 2015]
  - A few papers on "higher-order epistasis".
- **Statistical significance**
  - How do we compute p-values?
  - How do we correct for multiple hypotheses?

Further challenges

Privacy

- More data $\longrightarrow$ Data sharing $\longrightarrow$ **ethical** concerns
- How to learn from **privacy-protected** patient data?

  S. Simmons and B. Berger (2016) **Realizing privacy preserving genome-wide association studies**, Bioinformatics 32 (9), 1293–1300

Further challenges

Heterogeneity

- Multiple relevant **data sources** and **types**
- Multiple (unknown) **populations** of samples.

Tumor heterogeneity

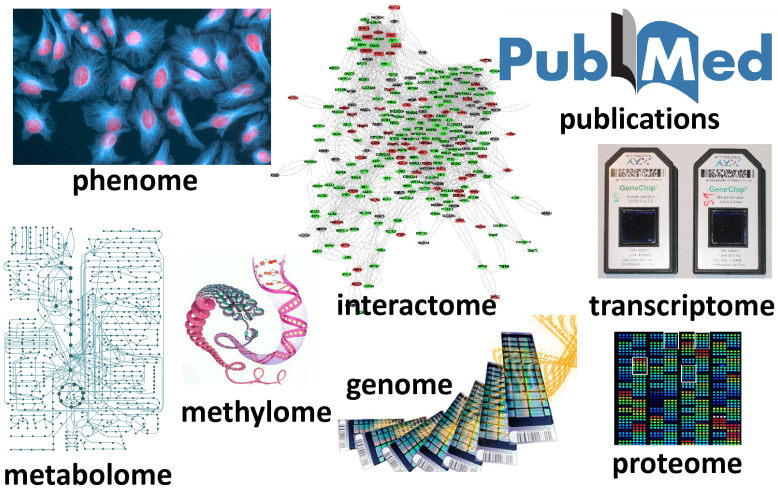
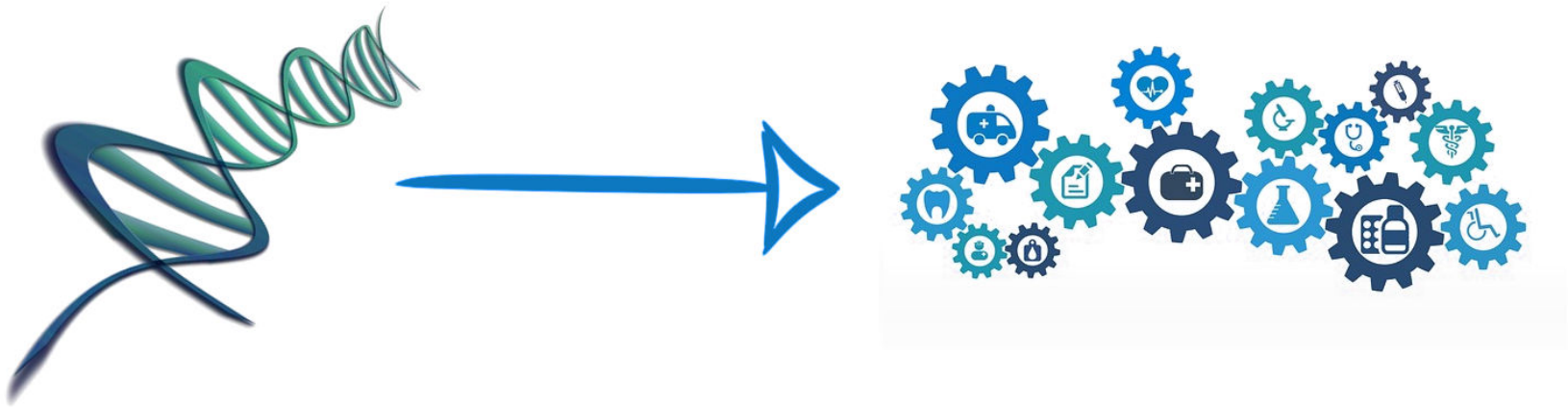


Image sources: ajc1@ flickr; Zlir'a@wikimedia

Heterogeneous data sources

L. Gay et al. (2016), F1000Research

Further challenges



Risk prediction

- State of the art: **Polygenic Risk Scores**

  Linear combination of SNPs with high p-values (**summary statistics**)
  Weighted by log odd ratios / univariate linear regression coefficients.

- **More complex models** slow to be adopted – **reliability?**

Interpretability

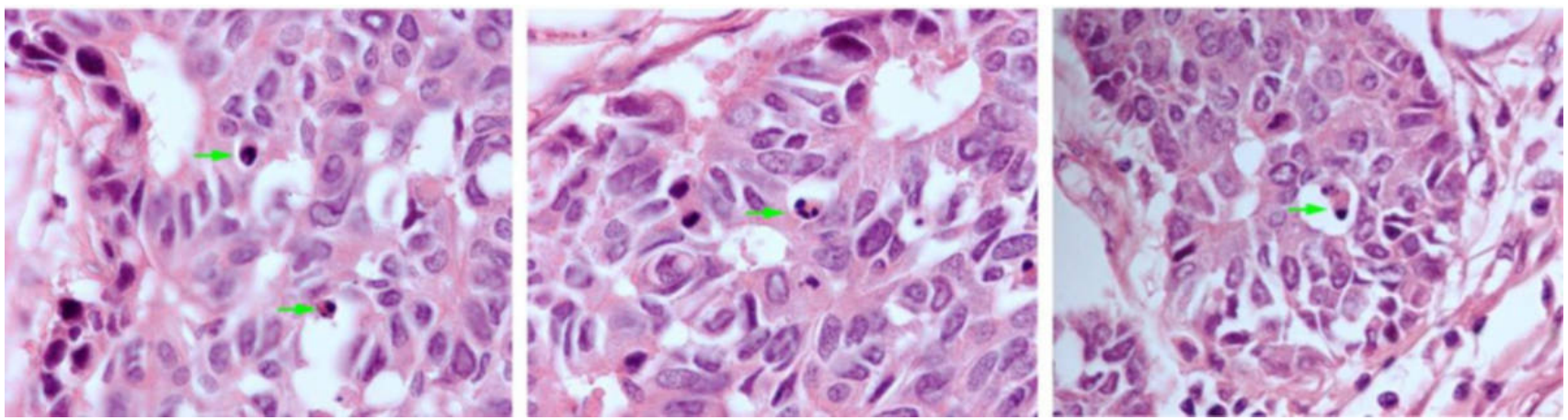
Predictivity

Further challenges

Bioimage informatics

High-throughput **molecular** and **cellular** images

BioImage Informatics `http://bioimageinformatics.org/`



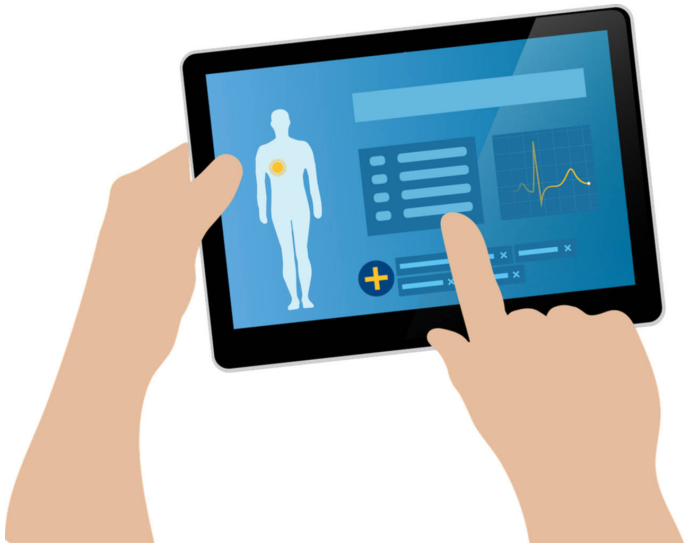
Detecting cells undergoing apoptosis

Further challenges

Electronic health records

- **Clinical notes:** incomplete, imbalanced, time series
- Combine **text** + **images** + **genetics**
- Assisting **evidence-based medicine**

R. Miotto et al. (2016) **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records** Scientific Reports 6.



https://github.com/chagaz/

CBIO: Víctor Bellón, Yunlong Jiao, Véronique Stoven, Athénaïs Vaginay, Nelle Varoquaux, Jean-Philippe Vert, Thomas Walter.

MLCB Tübingen: Karsten Borgwardt, Aasa Feragen, Dominik Grimm, Theofanis Karaletsos, Niklas Kasenburg, Christoph Lippert, Barbara Rakitsch, Damian Roqueiro, Nino Shervashidze, Oliver Stegle, Mahito Sugiyama.

MPI for Intelligent Systems: Lawrence Cayton, Bernhard Schölkopf.

MPI for Developmental Biology: Detlef Weigel.

MPI for Psychiatry: André Altmann, Tony Kam-Thong, Bertram Müller-Myhsok, Benno Pütz.

Osaka University: Yoshinobu Kawahara.

source: http://www.flickr.com/photos/wwworks/

References I

- Atwell, S. et al. (2010). "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines". In: Nature 465.7298, pp. 627–631. ISSN: 0028-0836. DOI: 10.1038/nature08800.
- Azencott, Chloé-Agathe et al. (2013). "Efficient network-guided multi-locus association mapping with graph cuts". In: Bioinformatics 29.13, pp. i171–i179.
- Bellon, Victor, VÃľronique Stoven, and ChloÃĽ-Agathe Azencott (2016). "Multitask feature selection with task descriptors". In: Pacific Symposium on Biocomputing. Vol. 21, pp. 261–272.
- Jacob, L., G. Obozinski, and J.-P. Vert (2009). "Group lasso with overlap and graph lasso". In: ICML, pp. 433–440.
- Lazer, David et al. (2014). "The parable of Google Flu: traps in big data analysis". In: Science 343.6176, pp. 1203–1205.
- Li, C. and H. Li (2008). "Network-constrained regularization and variable selection for analysis of genomic data". In: Bioinformatics 24.9, pp. 1175–1182.

References II

- Li, Caiyan and Hongzhe Li (2010). "Variable selection and regression analysis for graph-structured covariates with an application to genomics". In: The annals of applied statistics 4.3, pp. 1498–1516.
- Manolio, Teri A (2013). "Bringing genome-wide association findings into clinical use". In: Nature Reviews Genetics 14.8, pp. 549–558.
- Manolio, Teri A. et al. (2009). "Finding the missing heritability of complex diseases". In: Nature 461.7265, pp. 747–753.
- Niel, Clément et al. (2015). "A survey about methods dedicated to epistasis detection". In: Bioinformatics and Computational Biology, p. 285.
- Obozinski, Guillaume, Ben Taskar, and Michael I. Jordan (2006). Multi-task feature selection. Tech. rep. UC Berkeley.
- Pennisi, Elizabeth (2007). "Human Genetic Variation". In: Science 318.5858, pp. 1842–1843.
- Power, Robert A., Julian Parkhill, and Tulio de Oliveira (2017). "Microbial genome-wide association studies: lessons from human GWAS". In: Nature Reviews Genetics 18.1, pp. 41–50.

References III

- Segura, V. et al. (2012). "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations". In: Nat Genet 44.7, pp. 825–830.
- Swirszcz, Grzegorz and Aurelie C. Lozano (2012). "Multi-level Lasso for Sparse Multi-task Regression". In: Proceedings of the 29th International Conference on Machine Learning (ICML-12), pp. 361–368.
- Tibshirani, Robert (1994). "Regression shrinkage and selection via the lasso". In: J. R. Stat. Soc. 58, pp. 267–288.
- Visscher, Peter M. et al. (2012). "Five years of GWAS discovery". In: Am. J. Hum. Genet. 90.1, pp. 7–24.
- Wu, M. C., S. Lee, et al. (2011). "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test". In: AJHG 89.1, pp. 82–93.
- Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1, pp. 49–67.
- Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2, pp. 301–320.