Using structure to select features in high dimension

Chloé-Agathe Azencott

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

April 2, 2019 – IHP

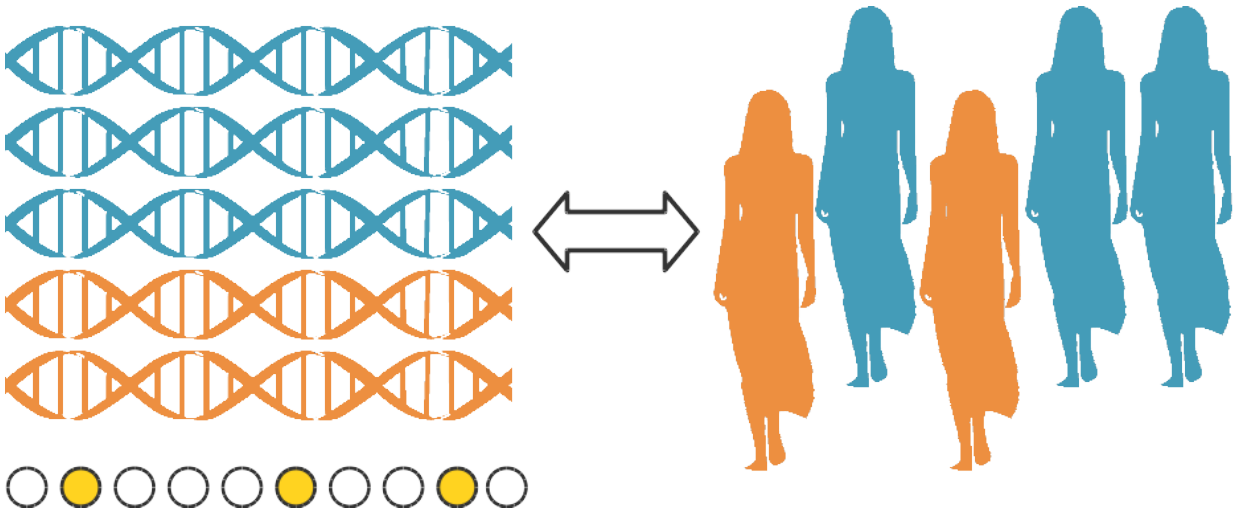
http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott

Precision Medicine

- The top highest-grossing drugs in the US only help 1/25 to 1/4 patients.
- **Differences in drug response** are partially due to **genetic** differences.
- **Adapt** treatment to the **(genetic) specificities** of the patient.

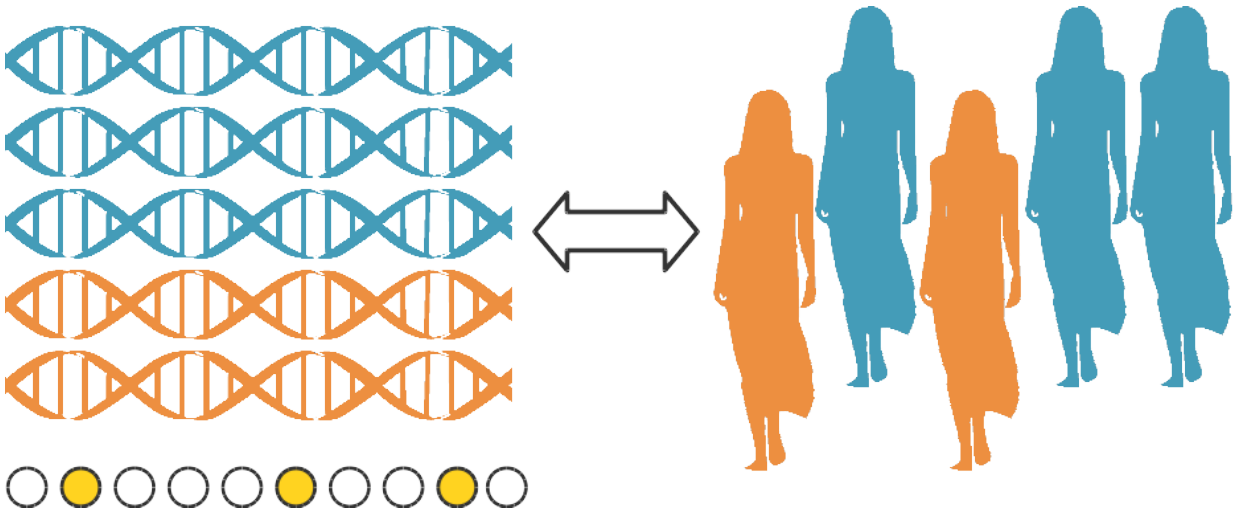
  E.g. Trastuzumab for HER2+ breast cancer.

From genotype to phenotype



Which genomic features explain the phenotype?

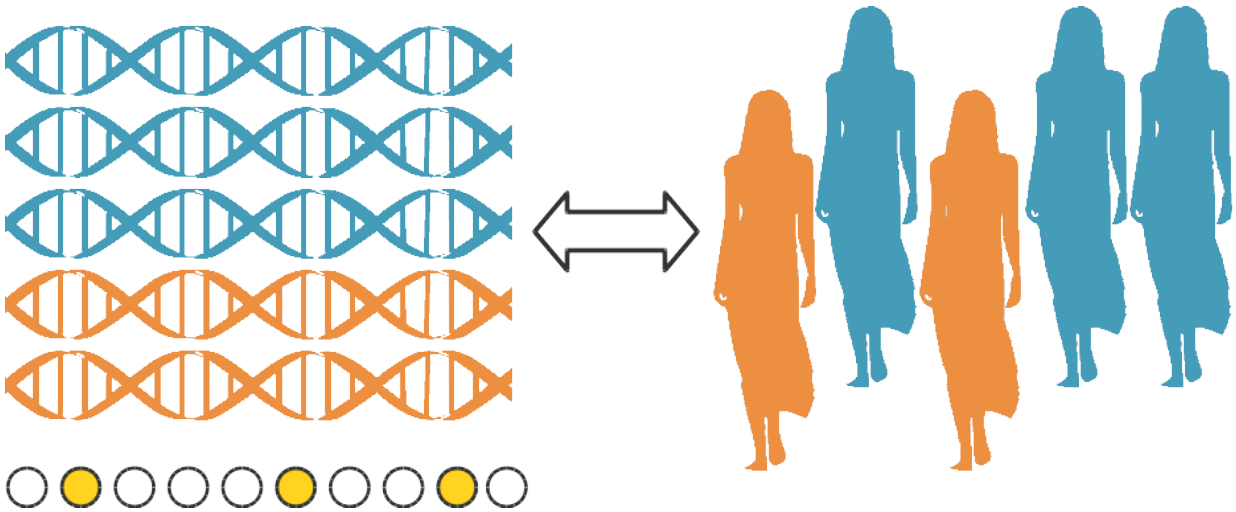
From genotype to phenotype



Which genomic features explain the phenotype?

- 80 000 proteins;
- 200 000 mRNA;
- 10 million SNPs;
- 28 million CpG islands.

From genotype to phenotype

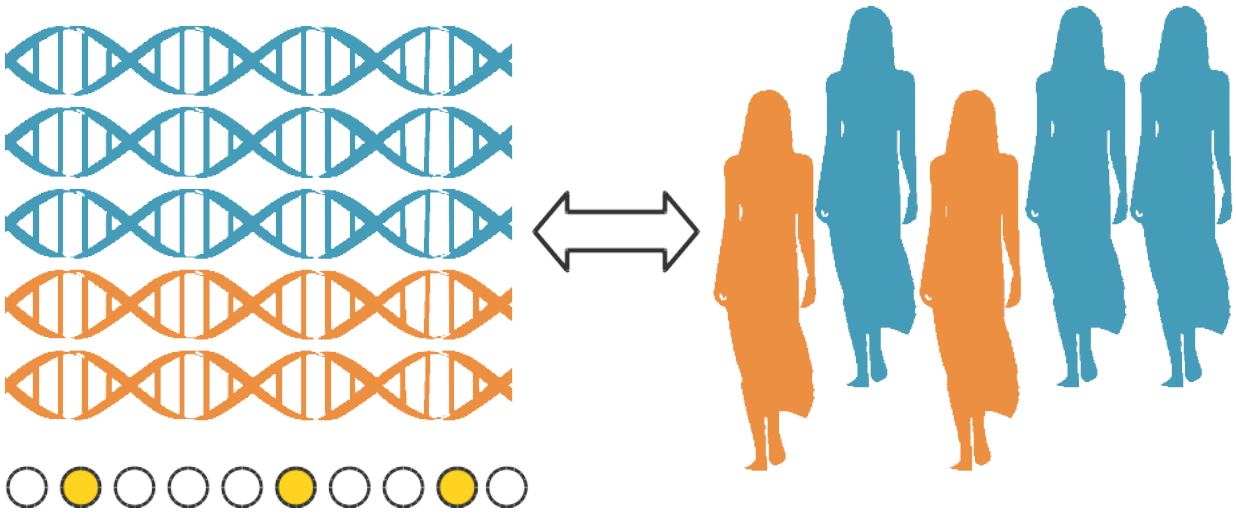


Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ **genomic features** $n = 10^3 - 10^5$ **samples**.

- 80 000 proteins;
- 200 000 mRNA;
- 10 million SNPs;
- 28 million CpG islands.

From genotype to phenotype



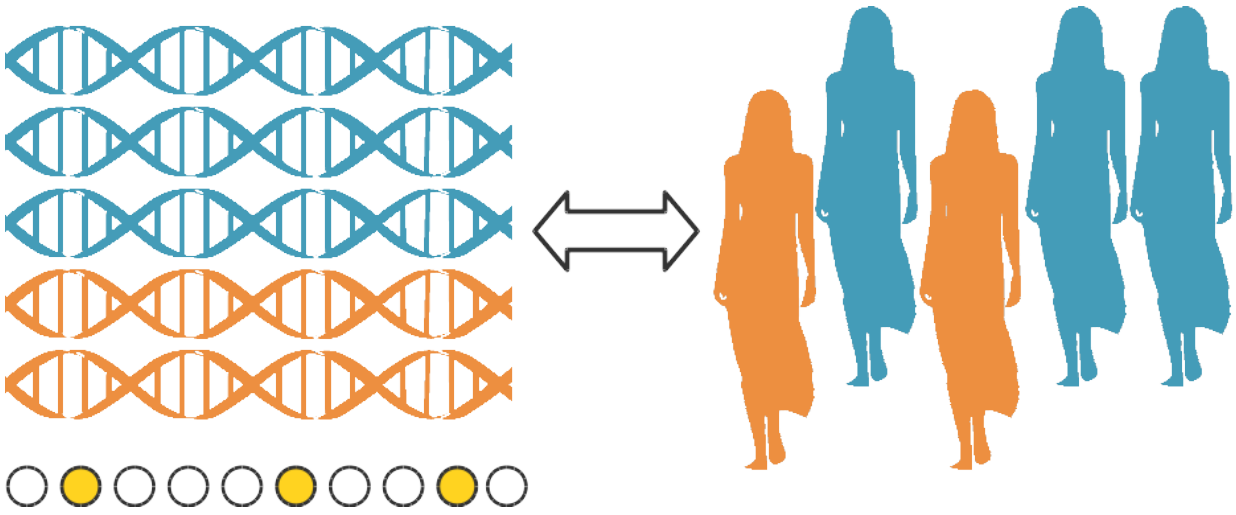
Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ **genomic features** $n = 10^3 - 10^5$ **samples**.

- 80 000 proteins;
- 200 000 mRNA;
- 10 million SNPs;
- 28 million CpG islands.

High-dimensional (large p), **low sample size** (small n) data.

From genotype to phenotype



Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ **genomic features** $n = 10^3 - 10^5$ **samples**.

- 10 million **S**ingle **N**ucleotide **P**olymorphisms.

Genome-**W**ide **A**ssociation **S**tudies.

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.**

Integrating prior knowledge: Network-guided GWAS

Joint work with Dominik Grimm, Yoshinobu Kawahara, Karsten Borgwardt, and Héctor Climente González.

Integrating prior knowledge

Use additional data and prior knowledge to constrain the feature selection procedure.

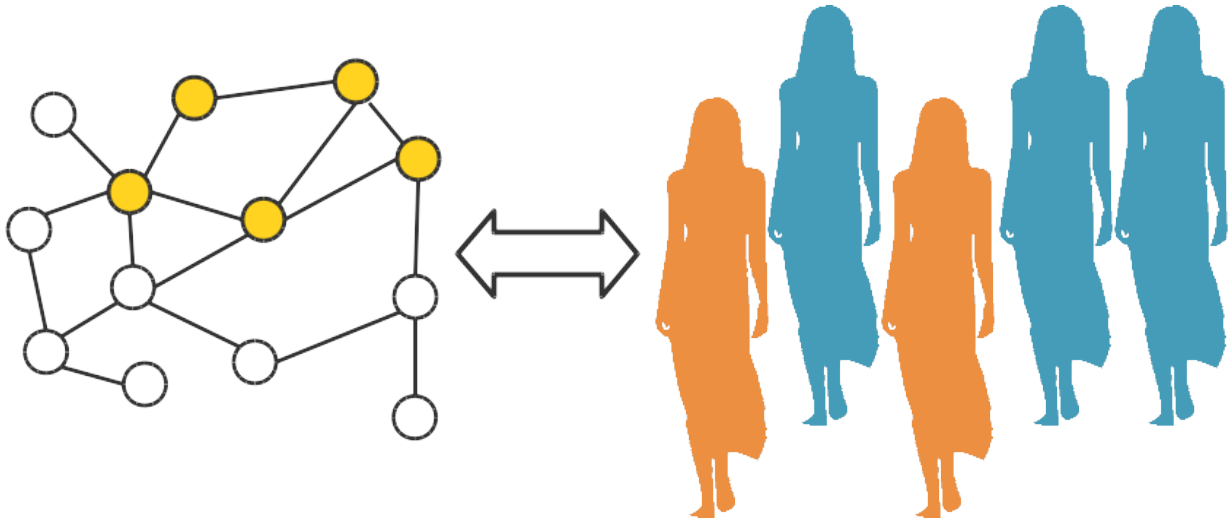
- **Consistant** with previously established knowledge;
- More easily **interpretable**;
- **Statistical power.**

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

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

Network-guided biomarker discovery

- **Biological networks** help understanding disease.
- Goal: Find a **set of explanatory features** compatible with a **given network** structure.



C.-A. Azencott (2016). **Network-guided biomarker discovery,** LNCS.

Integrating prior network knowledge

- **Network-constrained lasso:**

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{2}\sum_{i=1}^{n}\left(y^i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{sparsity}} + \eta \underbrace{\sum_{j=1}^{p}\sum_{k=1}^{p}\beta_j L_{jk}\beta_k}_{\text{connectivity}}$$

- **Graph Laplacian** $L \to \boldsymbol{\beta}$ varies **smoothly** on the network.

$$L_{jk} = \begin{cases} 1 & \text{if } j = k \\ -W_{jk}/\sqrt{d_j d_j} & \text{if } j \sim k \\ 0 & \text{otherwise.} \end{cases}$$

C. Li and H. Li (2008). **Network-constrained regularization and variable selection for analysis of genomic data,** Bioinformatics, 24, 1175–1182.

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**

$$\underset{\mathcal{S} \subseteq \mathcal{V}}{\arg\max}\, R(\mathcal{S}) - \lambda \Omega(\mathcal{S})$$

Network-guided GWAS

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

$$R(\mathcal{S}) = \sum_{j \in \mathcal{S}} c_j \qquad c_j = (\mathbf{X}^\top (\mathbf{y} - \mu))_j^2.$$

- **Sparse Laplacian regularization:**

$$\Omega : \mathcal{S} \mapsto \sum_{j \in \mathcal{S}} \sum_{k \notin \mathcal{S}} W_{jk} + \alpha |\mathcal{S}|.$$

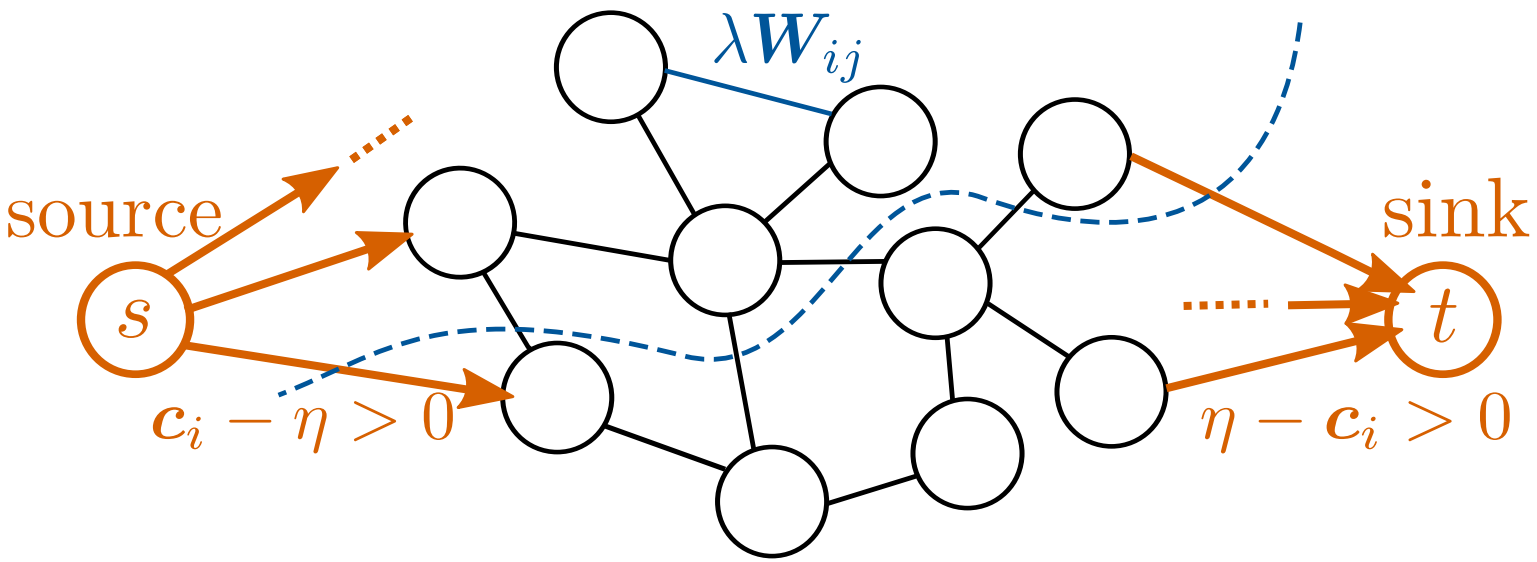
- **Regularized maximization of R:**

$$\underset{\mathcal{S} \subseteq \mathcal{V}}{\arg\max} \underbrace{\sum_{j \in \mathcal{S}} c_j}_{\text{association}} - \underbrace{\eta\,|\mathcal{S}|}_{\text{sparsity}} \underbrace{- \lambda \sum_{j \in \mathcal{S}} \sum_{k \notin \mathcal{S}} W_{jk}}_{\text{connectivity}}.$$

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**

$$\arg\min_{\beta_j \in \mathbb{R}} \frac{1}{2} ||\mathbf{y} - \beta_j \mathbf{x}_j||_2^2 .$$

- **Lasso**

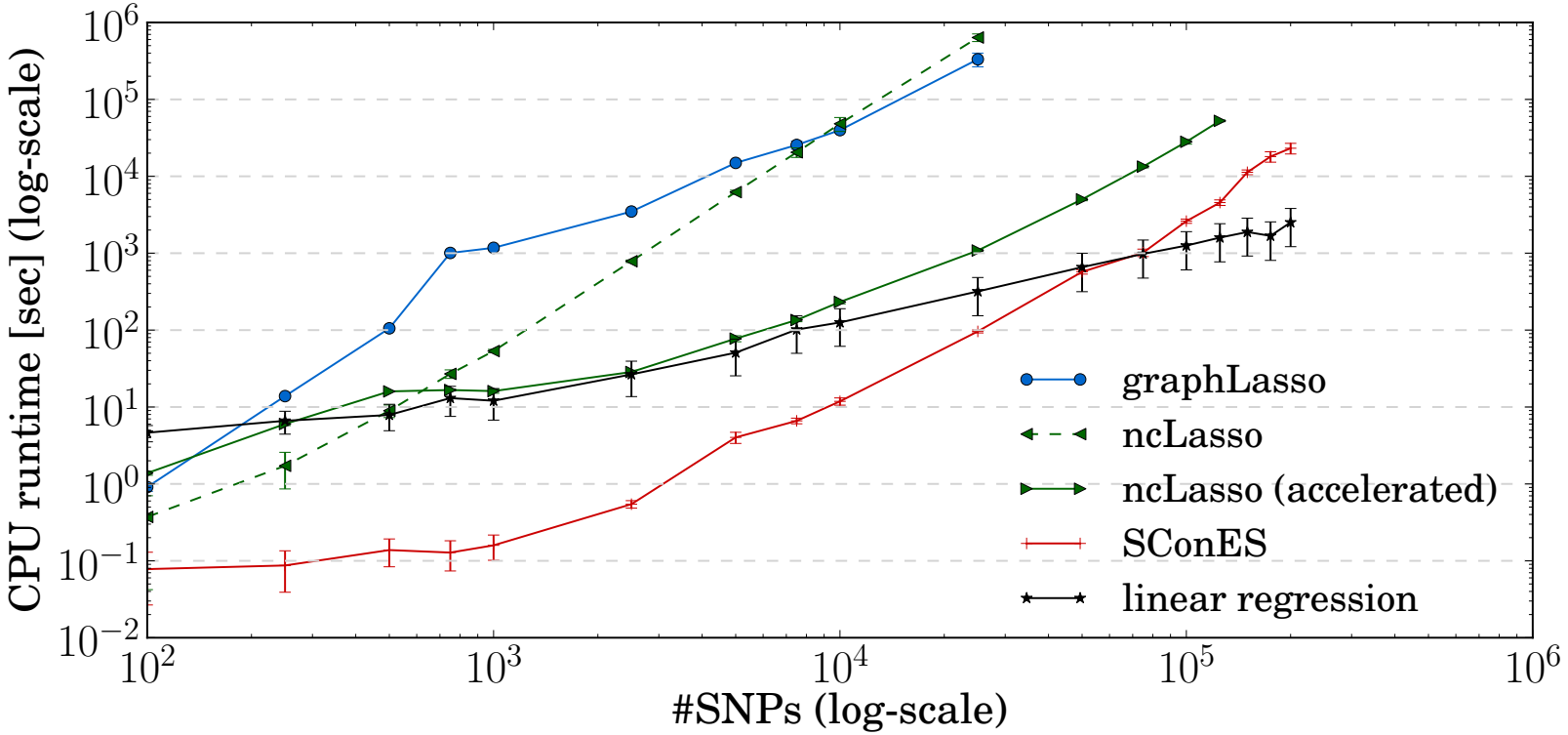
$$\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \eta \, ||\boldsymbol{\beta}||_1 .$$

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

$$\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||_2^2 + \eta \, ||\boldsymbol{\beta}||_1 + \lambda \, \Omega(\boldsymbol{\beta}).$$

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

Runtime

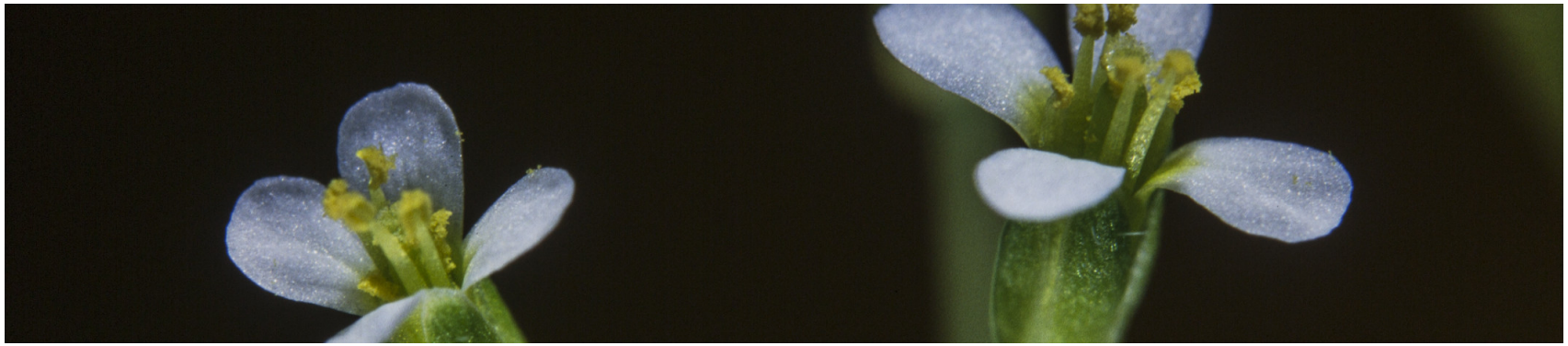


$n = 200$ exponential random network (2 % density)

Experiments: Performance on simulated data

- *Arabidopsis thaliana* genotypes:

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



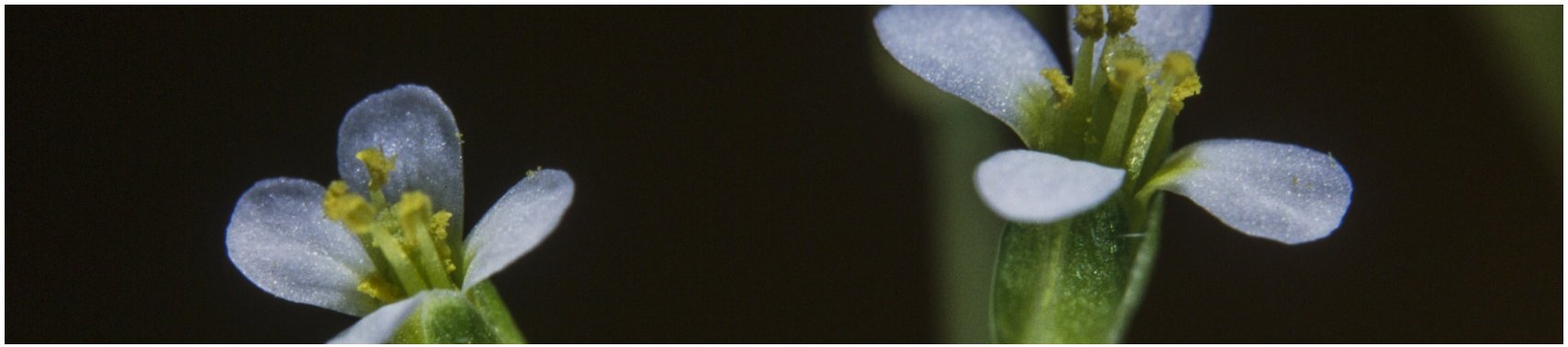
- Higher **power** and lower **FDR** than comparison partners except for groupLasso when groups = causal structure.
- Systematically **better than relaxed** version (ncLasso).
- Fairly robust to **missing edges.**
- Fails if network is **random.**

Image source: Jean Weber / INRA via Flickr.

Experiments: Performance on real data

- *Arabidopsis thaliana* genotypes:

  n ≈ 150 samples, p ≈ 170 000 SNPs,
  165 **candidate genes** [Segura et al., Nat Genet 2012].



- SConES selects **about as many SNPs** as other network-guided approaches but **they tag more candidate genes.**
- **Predictivity** of the selected SNPs:
  - In half the cases, **lasso** outperforms all other approaches;
  - In the remaining cases, **SConES** outperforms all other approaches.

Image source: Jean Weber / INRA via Flickr.

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.

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

- H. Climente, C.-A. Azencott (2017). **martini: GWAS incorporating networks in R,** doi:10.18129/B9.bioc.martini.

  Bioconductor/martini

Finding interactions between a target SNP and the rest of the genome.

Joint work with Lotfi Slim, Jean-Philippe Vert, and Clément Chatelain.

- p **variables** $X_1, X_2, \ldots, X_p \in \{0, 1, 2\}$;
- one **target variable** $A \in \{-1, 1\}$;
- **outcome** Y.

Which of the p variables interact with A towards Y?

- p **variables** $X_1, X_2, \ldots, X_p \in \{0, 1, 2\}$;
- one **target variable** $A \in \{-1, 1\}$;
- **outcome** Y.

Which of the p variables interact with A towards Y?

- **GBOOST:** For each $j = 1, \ldots, p$, LRT between
  - a full logistic regression model on $(X_j, A, A.X_j)$;
  - a main-effect logistic regression model on (X_j, A).

- p **variables** $X_1, X_2, \ldots, X_p \in \{0, 1, 2\}$;
- one **target variable** $A \in \{-1, 1\}$;
- **outcome** Y.

Which of the p variables interact with A towards Y?

- **GBOOST:** For each $j = 1, \ldots, p$, LRT between
  - a full logistic regression model on $(X_j, A, A.X_j)$;
  - a main-effect logistic regression model on (X_j, A).
- **product Lasso:** Lasso on $(X_1, X_2, \ldots, X_p, A, A.X_1, A.X_2, \ldots, A.X_p)$.

Modeling epistasis

- p **variables** $X_1, X_2, \ldots, X_p \in \{0, 1, 2\}$;
- one **target variable** $A \in \{-1, 1\}$;
- **outcome** Y.

Which of the p variables interact with A towards Y?

- $$Y = \mu(X) + A.\delta(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Modeling epistasis

- p **variables** $X_1, X_2, \ldots, X_p \in \{0, 1, 2\}$;
- one **target variable** $A \in \{-1, 1\}$;
- **outcome** Y.

Which of the p variables interact with A towards Y?

- $$Y = \mu(X) + A.\delta(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$
- SNPs in **epsitasis** with A = **support** of $\delta(X)$.

Clinical trials

$$Y = \mu(X) + A.\delta(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

- Which of the **SNPs** in X interact with **target SNP** A towards **phenotype** Y?
- Which of the **clinical covariates** X interact with **treatment** A towards **outcome** Y?

L. Tian et al. (2014). **A simple method for estimating interactions between a treatment and a large number of covariates.** JASA 109, 1517–1532.

Clinical trials

$$Y = \mu(X) + A.\delta(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

- Which of the **SNPs** in X interact with **target SNP** A towards **phenotype** Y?

- Which of the **clinical covariates** X interact with **treatment** A towards **outcome** Y?

  L. Tian et al. (2014). **A simple method for estimating interactions between a treatment and a large number of covariates.** JASA 109, 1517–1532.

  Modified outcome method to model δ:

$$Y' = 2\,Y\,A,$$

$$\delta(X) = \tfrac{1}{2}\,\mathbb{E}\left[Y'|X\right].$$

Clinical trials

$$Y = \mu(X) + A.\delta(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

- Which of the **SNPs** in X interact with **target SNP** A towards **phenotype** Y?

- Which of the **clinical covariates** X interact with **treatment** A towards **outcome** Y?

L. Tian et al. (2014). **A simple method for estimating interactions between a treatment and a large number of covariates.** JASA 109, 1517–1532.

Modified outcome method to model δ:

$$Y' = 2\,Y\,A,$$

$$\delta(X) = \tfrac{1}{2}\,\mathbb{E}\left[Y'|X\right].$$

No need to model the main effects μ!

Modified outcome

$$Y = \mu(X) + A.\delta(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

$$Y = \mathbb{E}\left[Y|A = a, X\right] + \epsilon.$$

- $\mu(X) = \frac{1}{2}\left(\mathbb{E}\left[Y|A = 1, X\right] + \mathbb{E}\left[Y|A = -1, X\right]\right)$
- $\delta(X) = \frac{1}{2}\left(\mathbb{E}\left[Y|A = 1, X\right] - \mathbb{E}\left[Y|A = -1, X\right]\right).$

- Introduce $\widetilde{A} = \frac{1}{2}(A + 1) \in \{0, 1\}$:

$$\delta(X) = \frac{1}{2}\,\mathbb{E}\left[Y\left(\frac{\widetilde{A}}{\pi(\widetilde{A} = 1|X)} - \frac{1 - \widetilde{A}}{\pi(\widetilde{A} = 0|X)}\right)\middle|X\right].$$

- **Modified outcome:**

$$Y' = Y\left(\frac{\widetilde{A}}{\pi(\widetilde{A} = 1|X)} - \frac{1 - \widetilde{A}}{\pi(\widetilde{A} = 0|X)}\right).$$

Propensity scores

- In GWAS, the target SNP is **not independent** from the rest of the genome because of **linkage disequilibrium.**
- Estimate **propensity scores** $\pi(\widetilde{A}|X)$
- Use **genomic structure** $\Rightarrow$ **Hidden Markov Model**.
  - **Hidden states:** contiguous clusters of phased haplotypes;
  - **Emission states:** SNPs.
- Typically used for
  - imputing **missing values;**
    P. Scheet and M. Stephens (2006). **A fast and flexible statistical model for large-scale population genotype data,** AJHG 78, 629–44.
  - constructing **knockoffs** for **FDR control.**
    M. Sesia, C. Sabatti and E. J. Candès (2018). **Gene hunting with hidden markov model knockoffs,** Biometrika.

Modified outcome variants

$$Y' = Y \left(\frac{\widetilde{A}}{\pi(\widetilde{A} = 1|X)} - \frac{1 - \widetilde{A}}{\pi(\widetilde{A} = 0|X)} \right).$$

- Propensity scores tend to be close to 0.
- **Shifted modified outcome:** $\pi(\widetilde{A}|X) \leftarrow \pi(\widetilde{A}|X) + \xi$.
- **Robust modified outcome.**

J. M. Robins, A. Rotnitzky, and L. P. Zhao (1994). **Estimation of regression coefficients when some regressors are not always observed,** J. Am. Stat. Ass., 427 (89), 846–866.

Evaluating the support of δ

- $\delta(X) = \frac{1}{2}\,\mathbb{E}[Y'|X]$.
- Use an **elastic net** regression to relate Y' and X:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}\left(Y_i' - \boldsymbol{\beta}^\top X_i\right)^2 + \lambda\left((1-\alpha)\,||\boldsymbol{\beta}||_1 + \alpha\,||\boldsymbol{\beta}||_2^2\right).$$

  α small $\rightarrow$ **sparsity**.

- Add **stability selection**
  - B bootstrap samples;
  - rank features based on the **area under the stability path**.

    A.-C. Haury et al. (2012), **TIGRESS: Trustful Inference of Gene REgulation using Stability Selection,** BMC Sys. Bio. 6.

Simulations

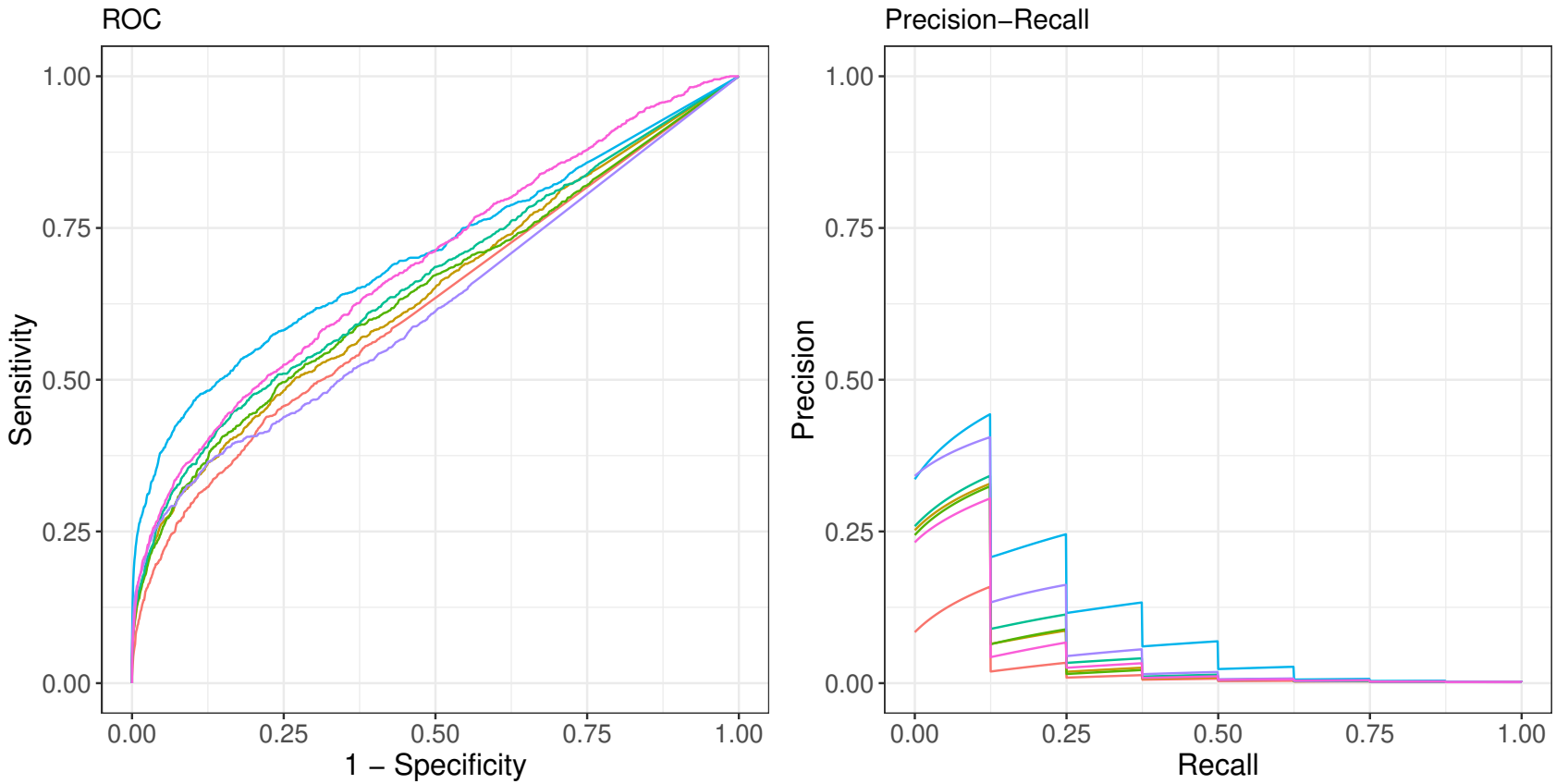
$$\pi_Y = \underbrace{\beta_{i,V}^\top X_V}_{\text{synergy with A}} + \underbrace{\beta_W^\top X_W}_{\text{marginal effects}} + \underbrace{X_{Z_1}^\top \text{diag}(\beta_{Z_1,Z_2} X_{Z_2})}_{\text{quadratic effects}}.$$

$$\pi_Y = \text{logit}(P(Y = 1 | \widetilde{A} = i, X)).$$

- p = 5 000, n = 500.
- $|V| = |W| = |Z_1| = |Z_2| = 8$
- $|V \cap W| = 2, |V \cap Z_1| = 2.$

Simulations

$$\pi_Y = \underbrace{\beta_{i,V}^\top X_V}_{\text{synergy with A}} + \underbrace{\beta_W^\top X_W}_{\text{marginal effects}} + \underbrace{X_{Z_1}^\top \text{diag}(\beta_{Z_1,Z_2} X_{Z_2})}_{\text{quadratic effects}}.$$



epiGWAS: Detecting epistasis with a target SNP.

- searches for a **sum of quadratic effects** with the **target SNP;**
- accounts for **main effects;**
- models **linkage disequilibrium.**

L. Slim, C. Chatelain, C.-A. Azencott, J.-P. Vert. (2018) **Novel methods for epistasis detection in genome-wide association studies**, BioRXiv.

CRAN/epiGWAS

Looking ahead

- **Robustness/stability**

  Stability selection is time consuming.

- **Complex interaction patterns**

  epiGWAS is limited to a sum of quadratic interactions between one target SNP and the rest of the genome.

- **Statistical significance**
  - **Significant pattern mining** [Llinares-López et al, Bioinformatics 2018].
  - **Post-selection inference**
    - For the **lasso** [Lee et al., AoS 2016].
    - For **higher-order interactions** [Suzumura et al., ICML 2017].
    - Ongoing work with L. Slim on **kernel PSI**.
  - Controlling **FDR** with **knockoffs** [Sesia et al., Biometrika 2018].

CBIO:
Héctor Climente González, Lotfi Slim, Jean-Philippe Vert (Google Brain).

Formerly MLCB Tübingen:
Karsten Borgwardt (ETH Zürich, Switzerland), Dominik Grimm (Weihenstephan, Germany), Mahito Sugiyama (National Institute of Informatics, Japan).

Osaka University & RIKEN AIP:
Yoshinobu Kawahara.

Sanofi:
Clément Chatelain.

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

WiMLDS Paris

meetup

Paris Women in Machine Learning and Data Science.

- March 12, 19:30

  Human body extraction from images – Gül Varol (INRIA Willow).
  Data is beautiful, please don't ruin it – Anne-Marie Tousch (Criteo Lab).
  Salary negociation workshop – Natalie Cernecka.

- March 28, 19:00 – **Femmes, sciences et société**

  Femmes, probabilités et finances – Nicole El Karoui.
  La féministe, l'économiste et la cité – Hélène Périvier.
  Discussion ouverte.