High-dimensional feature selection in precision medicine

Chloé-Agathe Azencott

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

April 25, 2017 – ICLR

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.

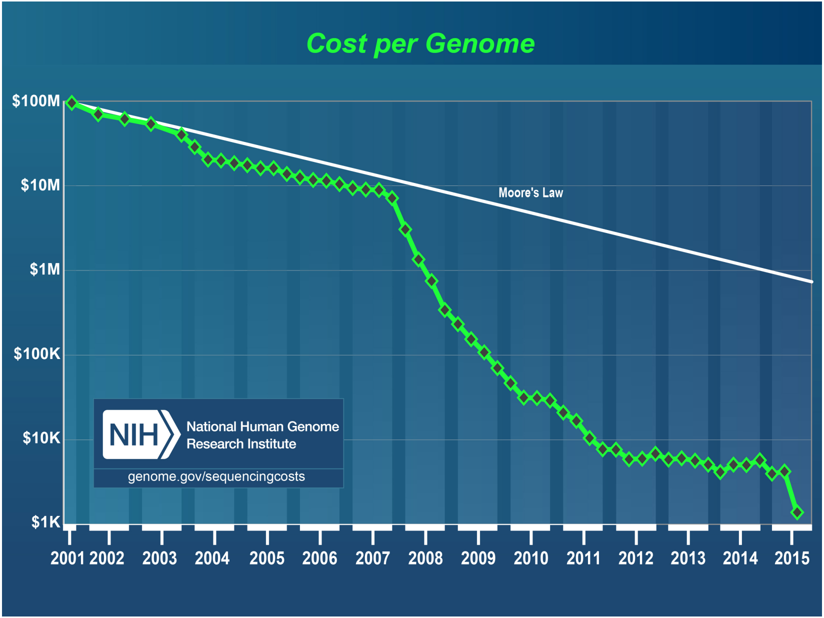
  E.g. Trastuzumab for HER2+ breast cancer.

- **Data-driven** biology/medicine

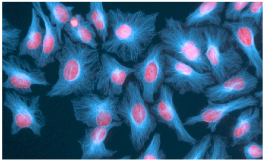
  Identify similarities between patients that exhibit similar phenotypes.

Data + **Feature Selection**

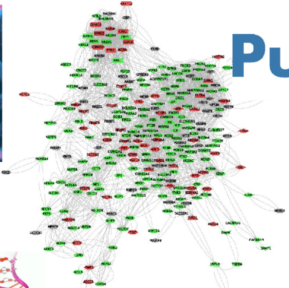
Sequencing costs



Big data!



phenome

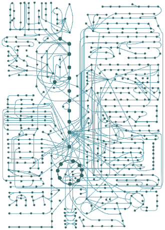


interactome

PubMed

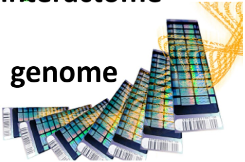
publications

transcriptome

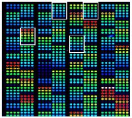


metabolome

methylome



genome



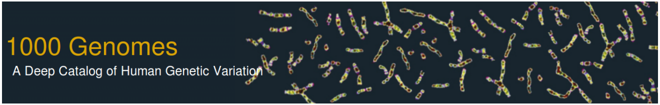
proteome

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

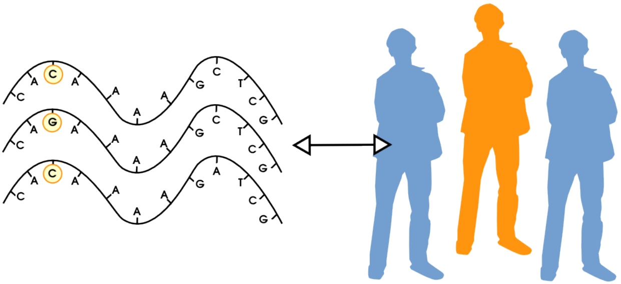


1000 Genomes
A Deep Catalog of Human Genetic Variation

DREAM Challenges

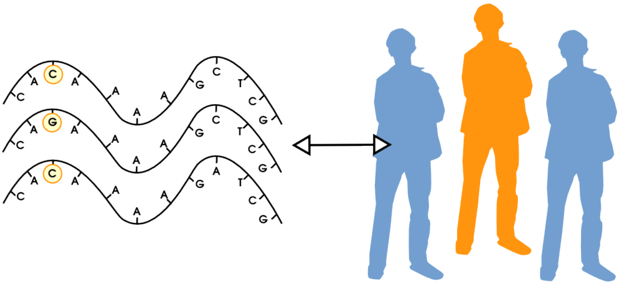
SAY BIG DATA
ONE MORE TIME

GWAS: Genome-Wide Association Studies



Which genomic features explain the phenotype?

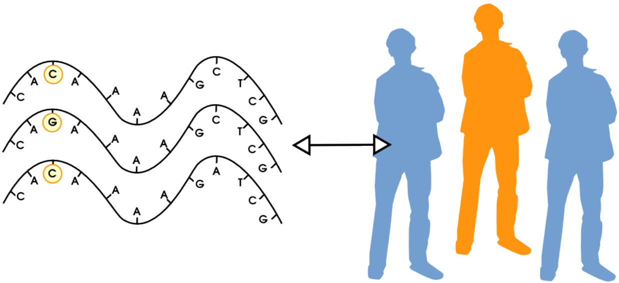
GWAS: Genome-Wide Association Studies



Which genomic features explain the phenotype?

p = 10^5 – 10^7 Single Nucleotide Polymorphisms (SNPs)
n = 10^2 – 10^4 samples

GWAS: Genome-Wide Association Studies

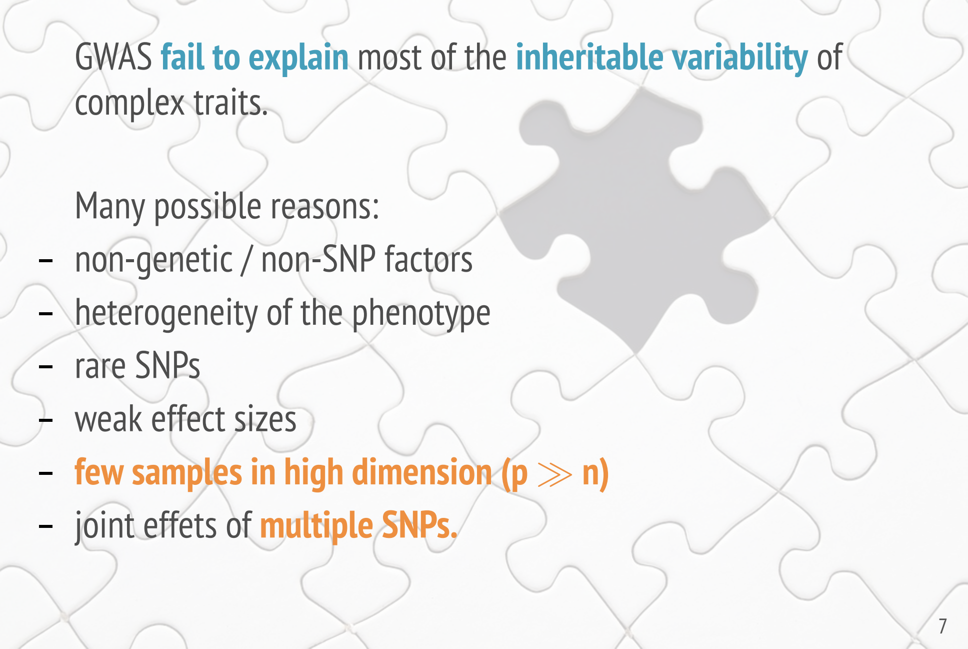


Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ Single Nucleotide Polymorphisms (SNPs)
$n = 10^2 - 10^4$ samples

High-dimensional (large p)
Low sample size (small n)

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.**

Is extracting knowledge from such data doomed from the start?

Reducing p

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.

Regularized relevance

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

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

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

  Ex : correlation, HSIC, statistical test of association.

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

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

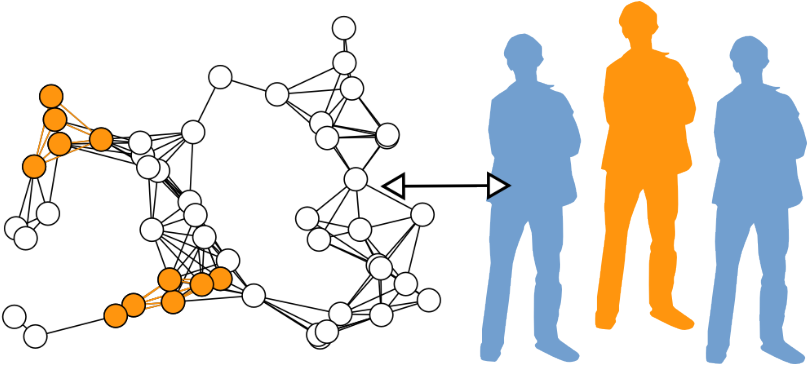
  Ex : cardinality $\Omega : \mathcal{S} \mapsto |\mathcal{S}|$.

- **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 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**

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

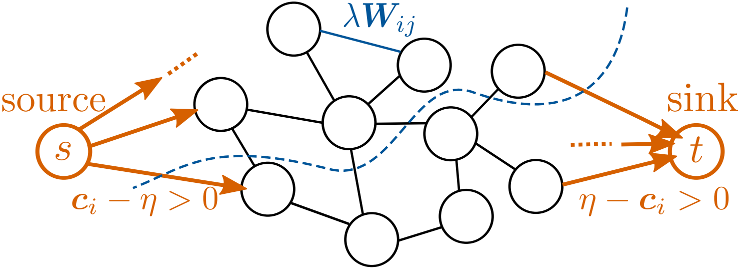
- **Regularized maximization of R**

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

Minimum cut reformulation

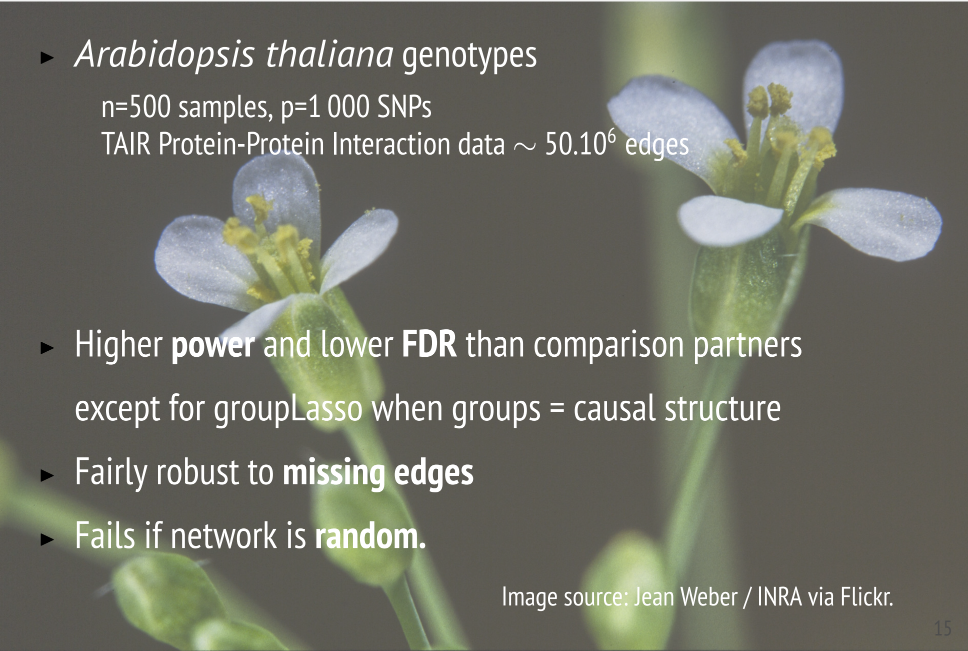
The graph-regularized maximization of score $Q(*)$ is equivalent to a s/t-min-cut for a graph with adjacency matrix $\mathbf{A}$ and two additional nodes s and t, where $\mathbf{A}_{ij} = \lambda \mathbf{W}_{ij}$ for $1 \leq i, j \leq p$ and the weights of the edges adjacent to nodes s and t are defined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

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


$\lambda \boldsymbol{W}_{ij}$
source
s
$\boldsymbol{c}_i - \eta > 0$
sink
t
$\eta - \boldsymbol{c}_i > 0$

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

Experiments: Performance on simulated data



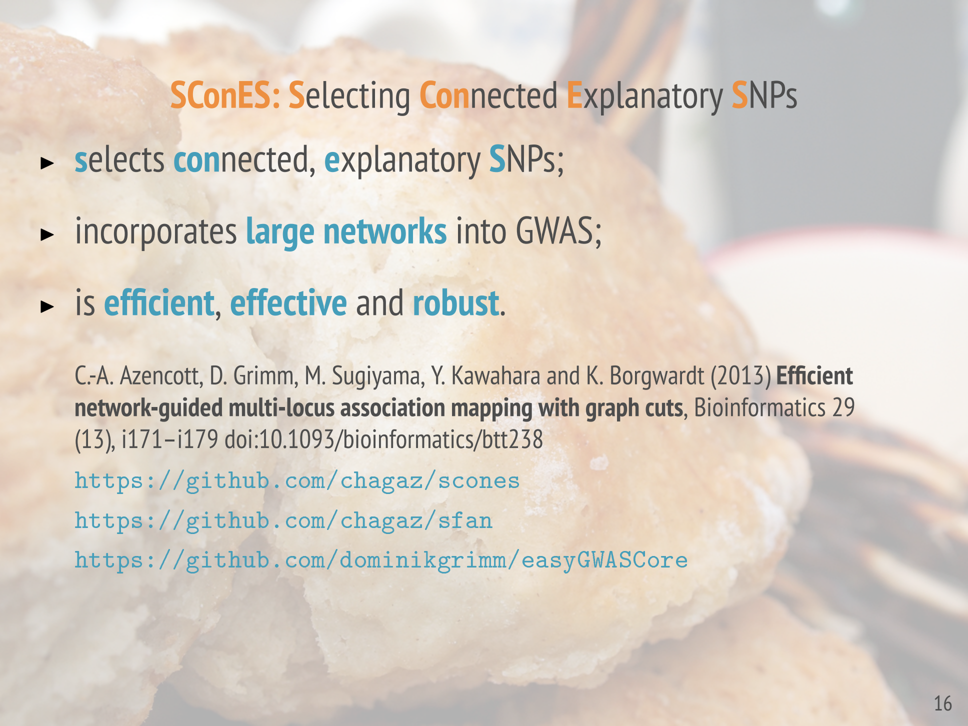
- *Arabidopsis thaliana* genotypes

  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.



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/scones

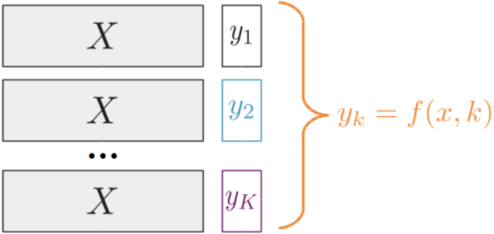
https://github.com/chagaz/sfan

https://github.com/dominikgrimm/easyGWASCore

Increasing n

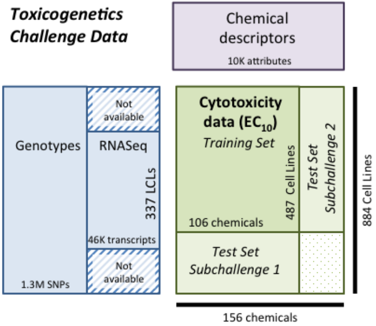
Multi-trait GWAS

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



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

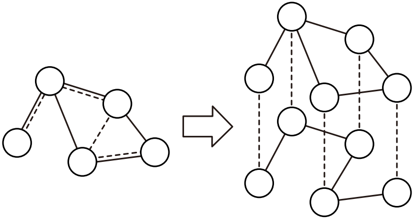
T **related phenotypes.**

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

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

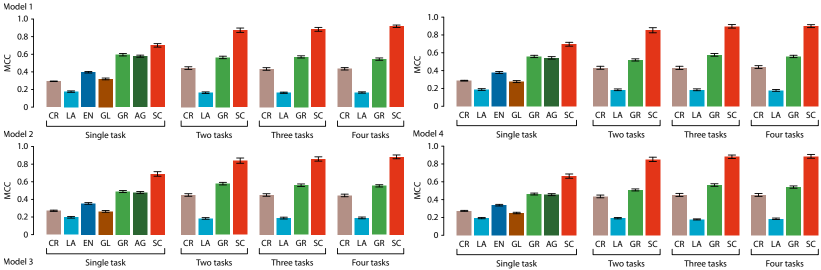
$\mathcal{S} \,\Delta\, \mathcal{S}' = (\mathcal{S} \cup \mathcal{S}') \setminus (\mathcal{S} \cap \mathcal{S}')$ (symmetric difference)

- 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

Using task similarity

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

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

Can also be mapped to a meta-network.

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

Using task descriptors

PhD thesis of Víctor Bellón.

Multiplicative Multitask Lasso with Task Descriptors

- **Multitask Lasso** [Obozinski et al. 2006]

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

- **Multilevel Multitask Lasso** [Lozano and Swirszczw, 2012]

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

- **Multiplicative Multitask Lasso with Task Descriptors**

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

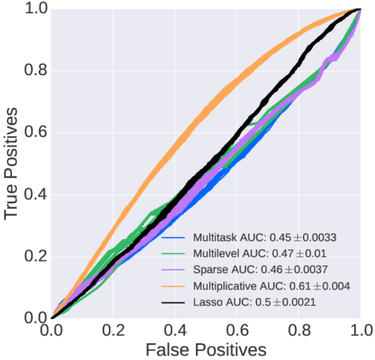
Multiplicative Multitask Lasso with Task Descriptors

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

- 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**

  Recovering the same SNPs when the data changes slightly.

- **Complex interaction patterns**
  - Limited to additive or quadrative effects
  - Some work on e.g. random forests + importance score.

- **Statistical significance**
  - Computing p-values
  - Correcting for multiple hypotheses.

Further challenges

Privacy

- More data → Data sharing → **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

L. Gay et al. (2016), F1000Research

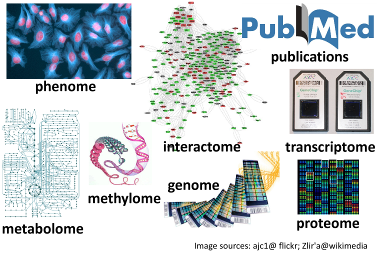


Image sources: ajc1@ flickr; Zlir'a@wikimedia

Heterogeneous data sources

Further challenges

Risk prediction

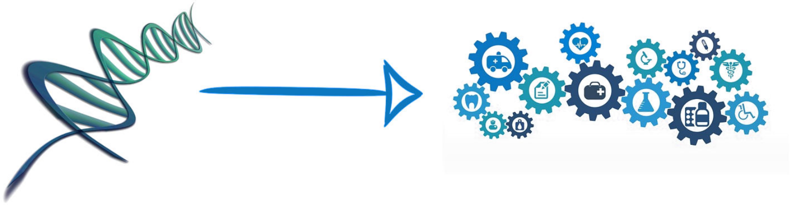
- State of the art: **Polygenic Risk Scores**

  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?**

  H.-C. So and P. C. Sham (2017) **Improving polygenic risk prediction from summary statistics by an empirical Bayes approach.** Scientific Reports 7.

  S. Okser et al (2014) **Regularized machine learning in the genetic prediction of complex traits.** PLoS Genet 10.11: e1004754.



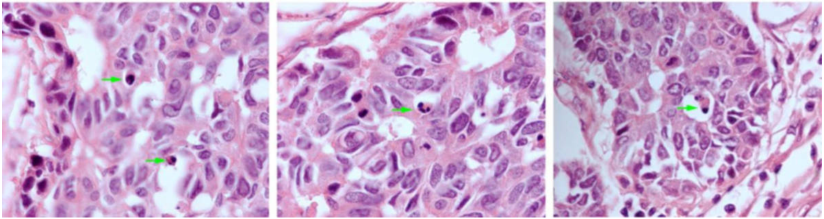
Further challenges

Bioimage informatics

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

- **Subcellular location** analysis
- **High-content screening**
- Segmentation, tracking, registration.

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



Detecting cells undergoing apoptosis

Further challenges

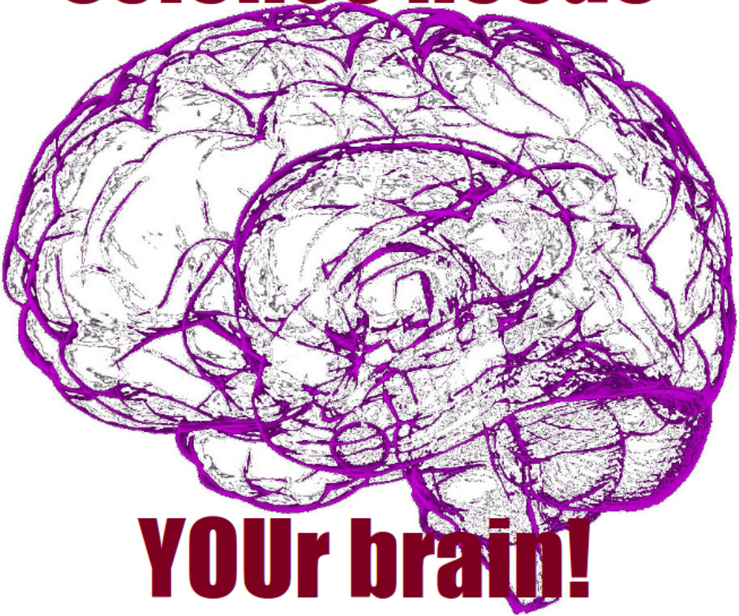
Electronic health records

- **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.

Machine Learning in Health Care `http://mucmd.org/`
Previously known as Meaningful Use of Complex Medical Data

Science needs



YOUr brain!

A few starting places

Data and Challenges

- **DREAM Challenges:** Crowdsourcing challenges for biology and medicine `http://dreamchallenges.org/`
- **Epidemium:** Cancer research through data challenges `http://www.epidemium.cc/`
- **MIMIC:** Deidentified electronic health records `https://mimic.physionet.org/`
- **BioImage Informatics Challenges** `https://bii.eecs.wsu.edu/challenges/`

A few starting places

Workshops

- **Machine Learning in Healthcare** at NIPS
  `http://www.nipsml4hc.ws/`
- **Machine Learing in Computational Biology**
  `https://mlcb.github.io/`
- **Machine Learning in Systems Biology** `http://mlsb.cc`

A few starting places

Basics in molecular biology

- Talk to **specialists!**
- **The DNA Learning Center**
  `https://www.dnalc.org/resources/`
- **Scitable eBooks**
  `https://www.nature.com/scitable/ebooks`

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/