BIOINNOVATION PROGRAM MEETING

ALEXEJ GOSSMANN

2017/05/09

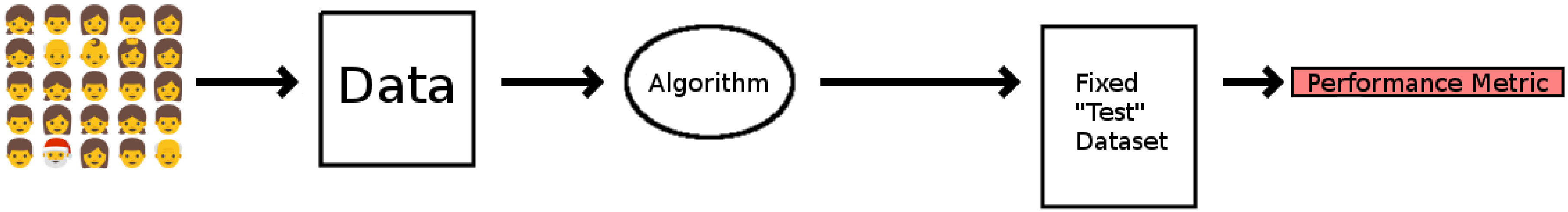
I. MY PROJECT AT THE FDA

CENTER FOR DEVICES AND RADIOLOGICAL HEALTH

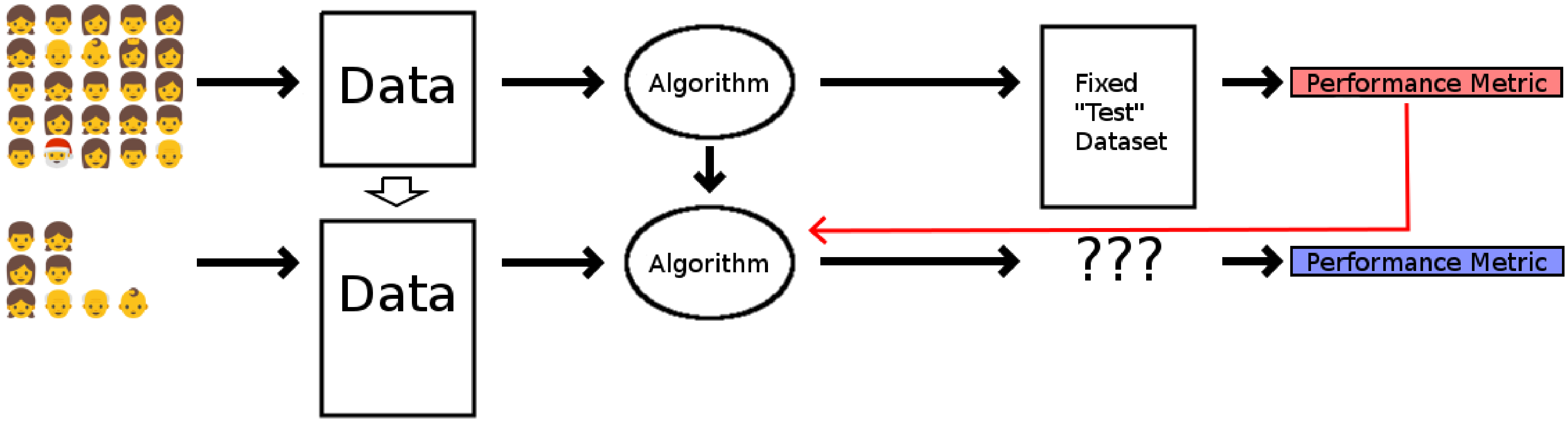
OFFICE OF SCIENCE AND ENGINEERING LABORATORIES

DIVISION OF IMAGING, DIAGNOSTICS, AND SOFTWARE RELIABILITY

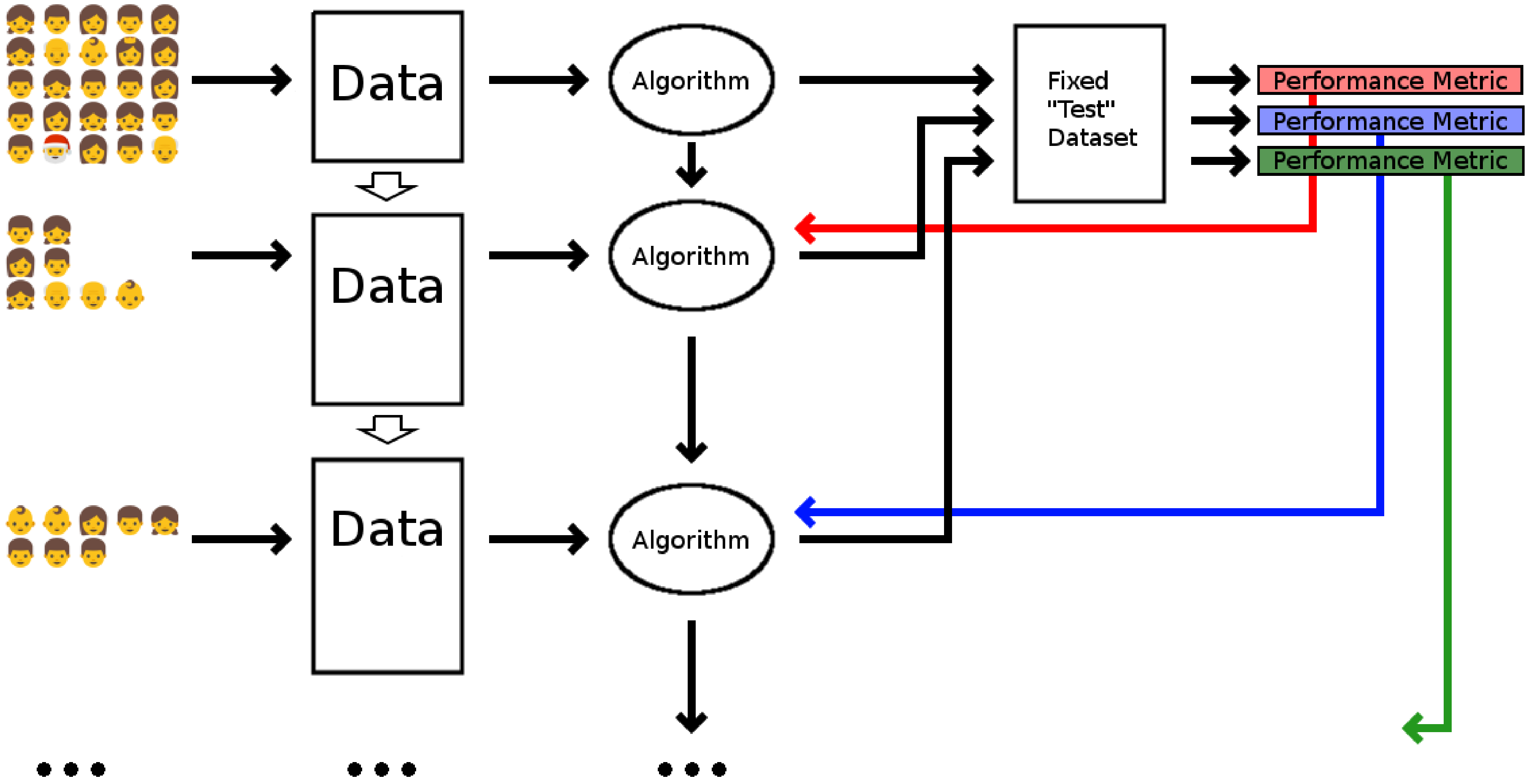
General machine learning process

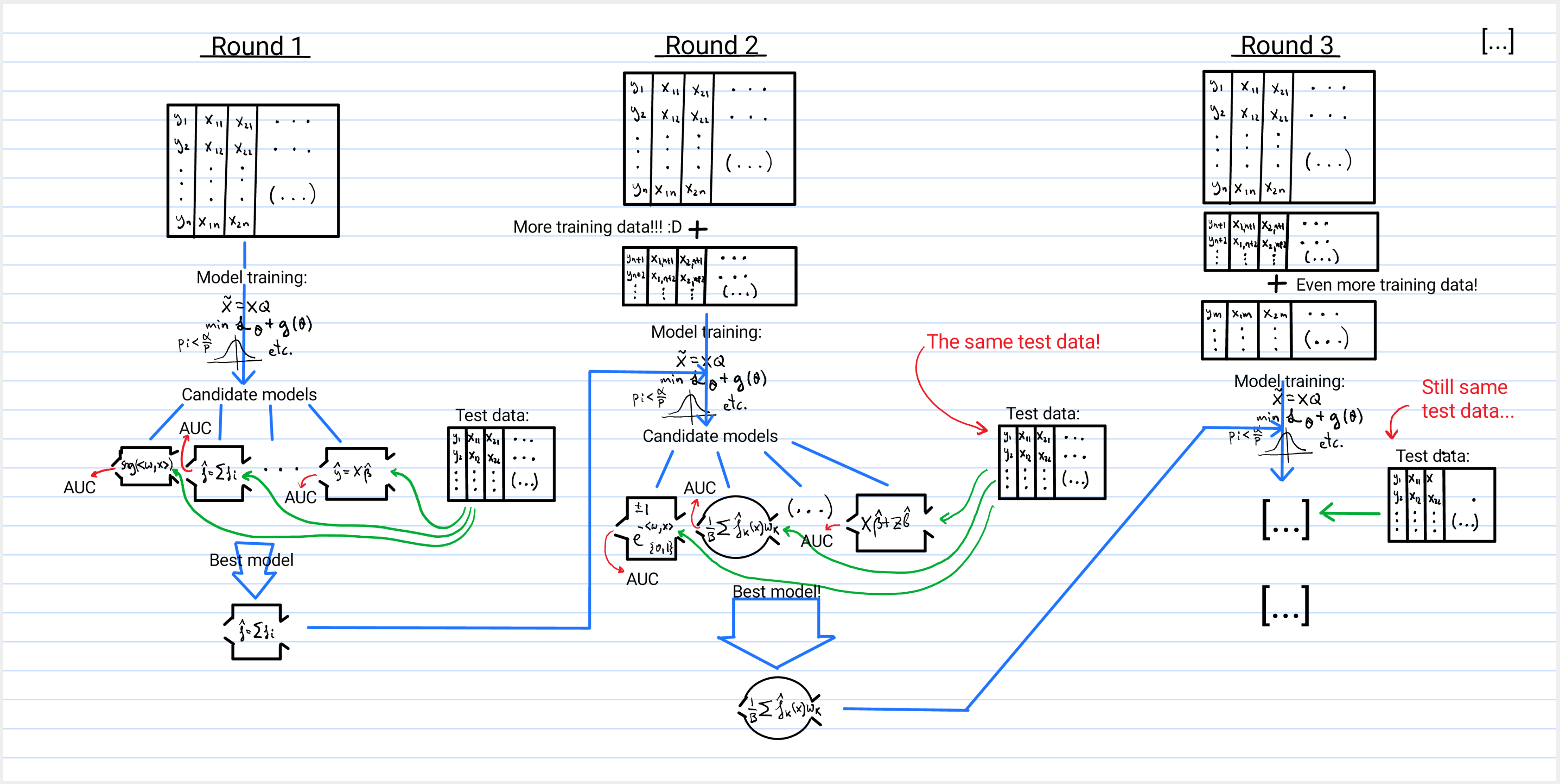


"Adaptive" machine learning



"Adaptive" machine learning with test data reuse





Round 1

Model training:

Candidate models

AUC

AUC

AUC

Best model

Test data:

Round 2

More training data!!! :D +

Model training:

Candidate models

AUC

AUC

AUC

Best model!

The same test data!

Test data:

Round 3

[...]

+ Even more training data!

Model training:

Still same test data...

Test data:

[...]

[...]

y_1 x_{11} x_{21}
y_2 x_{12} x_{22}
y_n x_{1n} x_{2n}
$(\ldots)$
y_{n+1} $x_{1,n+1}$ $x_{2,n+1}$
y_{n+2} $x_{1,n+2}$ $x_{2,n+2}$
y_m x_{1m} x_{2m}
$\tilde{X} = XQ$
$\min \mathcal{L}_\theta + g(\theta)$
$p_i < \frac{\alpha}{p}$
etc.
$\mathrm{sog}(\langle \omega, x \rangle)$
$\hat{f} = \sum f_i$
$\hat{y} = X\hat{\beta}$
$e^{-\langle \omega, x \rangle}$
± 1
$\{0,1\}$
$\frac{1}{B} \sum \hat{f}_k(x) \omega_k$
$X\hat{\beta} + Z\hat{b}$

PERFORMANCE ASSESSMENT IN ADAPTIVE MACHINE LEARNING WITH TEST DATA REUSE

Repeated usage of the same test data inadvertently leads to:

- **Overly optimistic performance assessments** 😰 (sometimes substantially so).
- **Loss of generalization** 😱: A machine learning system that performs much better on the available test cases than on the general population; i.e., overfitting to the test dataset.

POSSIBLE SOLUTION: DIFFERENTIALLY PRIVATE ACCESS TO TEST DATA

- Differential privacy is a mathematically rigorous definition of data privacy (see work of Cynthia Dwork and her collaborators).
- **Intuition:** If the test dataset can be accessed only via a differentially private mechanism, then the machine learning algorithm will have *no way to extract information about individual dataset records, but will only learn characteristics of the population as a whole.* 👉 Algo will adapt to the underlying distribution rather than to records in the specific dataset. 😃
- Under certain theoretical conditions this works even if the test dataset is reused thousands of times (Dwork et. al., Science, 2015). 👍
- In practice the reported performance metrics are much more accurate even when the theoretical conditions are not met (our work). 😸

Algorithm: **Thresholdout**$_{\text{AUC}}$

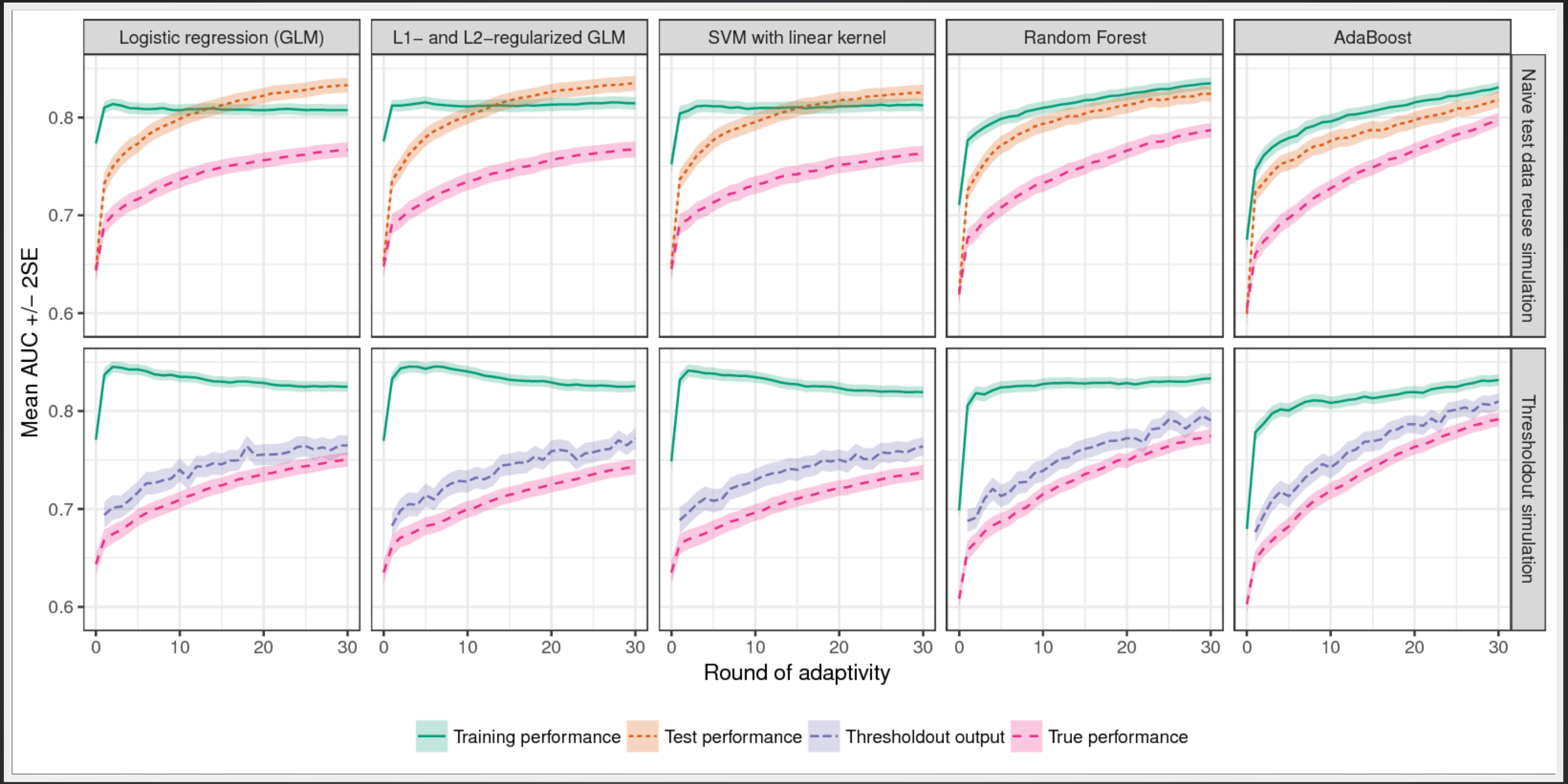
Input:

- Training dataset S_{train} and test dataset S_{test}.
- Noise rate σ, budget B, threshold T.
- Set $\hat{T} \leftarrow T + \gamma$ for $\gamma \sim \text{Lap}(2\sigma)$, where $\text{Lap}(2\sigma)$ denotes the Laplace distribution with mean 0 and scale parameter 2σ.

Query step:

Given a function ϕ that assigns a score between 0 and 1 to each observation, **do**:

- If $B < 1$ output $\perp$ (i.e., the test data access budget is exhausted).
- Else sample $\xi \sim \text{Lap}(\sigma), \gamma \sim \text{Lap}(2\sigma)$, and $\eta \sim \text{Lap}(4\sigma)$:
    - If $\left|\widehat{\text{AUC}}_{S_{\text{test}}}(\phi) - \widehat{\text{AUC}}_{S_{\text{train}}}(\phi)\right| > \hat{T} + \eta$, output $\widehat{\text{AUC}}_{S_{\text{test}}}(\phi) + \xi$ and set $B \leftarrow B - 1$ and $\hat{T} \leftarrow T + \gamma$.
    - Otherwise output $\widehat{\text{AUC}}_{S_{\text{train}}}(\phi)$.



This work has been submitted to SPIE Medical Imaging 2018 for publication in the Proceeding of SPIE and a conference presentation.

II. MY TULANE WORK

CERTAIN METHODS FOR FDR CONTROL IN SPARSE REGRESSION AND SPARSE CCA

THE MODEL SELECTION PROBLEM

- The simplest example: 👍 linear model.
- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where
  - $\mathbf{y} \in \mathbb{R}^n$ dependent variable (e.g., disease status/severity for n 😷),
  - $X \in \mathbb{R}^{n \times p}$ explanatory variables (for each 😷 record p features: 🚬, 💉, 💊, ...),
  - *Unknowns:* $\boldsymbol{\beta} \in \mathbb{R}^p$ (want to estimate), $\boldsymbol{\varepsilon}$ noise.
- **Prediction**: Find best predictions for $\mathbf{y}$.
- **Feature selection**: Find which β_i are non-zero.

...IN GENOMICS AND BRAIN IMAGING

..WHY WE CARE

🔬 + 💰

- Prediction of a disease phenotype based on a handful of features is needed for inexpensive diagnosis.
- Elimination of noisy or redundant features leads to more accurate prediction.
- "Data-generated hypotheses" lead to a better understanding of the underlying biology.

...IN GENOMICS AND BRAIN IMAGING

..CHALLENGES

- *~~Possibly~~ Only slightly less often than always* $n \ll p$. 😓
- Curse of dimensionality (when $n < p$). 😰
- Overfitting. 😱
- Underfitting. 🙀

ℓ_1 REGULARIZATION (E.G. LASSO BY TIBSHIRANI, 1994)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1$$

- Yields a sparse $\hat{\boldsymbol{\beta}}$.
- Computationally efficient and very useful in practice.
- Problem 1: unclear how to select λ.
- Problem 2: unclear how to do statistical inference on $\hat{\boldsymbol{\beta}}$.

MULTIPLE HYPOTHESES TESTING PERSPECTIVE

Alternatively, feature selection can be regarded as testing the p hypotheses

$$H_i : \beta_i = 0, \quad i = 1, \dots, p.$$

- Denote R := number of rejected hypotheses, and V := number of false rejections (i.e., Type I errors).
- *Family-wise error rate*:

  $$\mathrm{FWER} = \mathbb{P}\left(\text{At least one false rejection}\right) = \mathbb{P}(V \geq 1).$$

  E.g. Bonferroni correction (60ies?):

  $$\mathbb{P}(V \geq 1) \leq \mathbb{P}\left(\bigcup_{i=1}^{n}\{H_i \text{ falsely rejected}\}\right) \leq \sum_{i=1}^{n} \underbrace{\mathbb{P}\left(\{H_i \text{ falsely rejected}\}\right)}_{\leq \alpha/n} \leq \alpha.$$

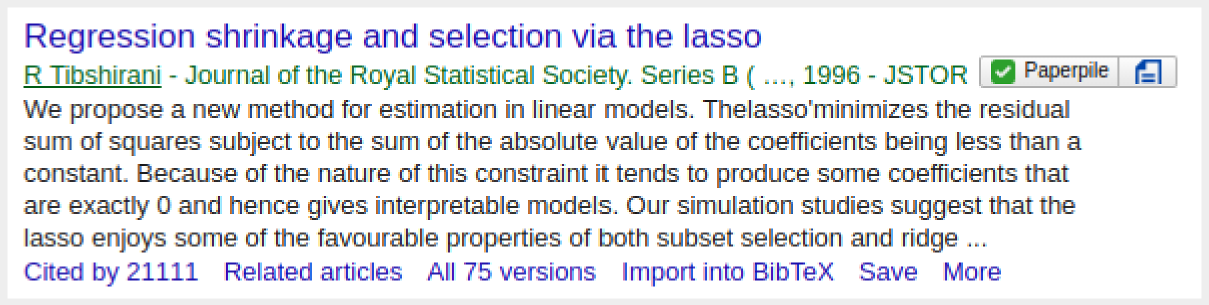
- *False discovery rate* ('95):

  $$\mathrm{FDR} = \mathbb{E}\left(\frac{\#\text{False rejections}}{\#\text{Rejections}}\right) = \mathbb{E}\left(\frac{V}{\min\{R, 1\}}\right).$$

  E.g. Benjamini-Hochberg:

  1. Sort the p-values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$.
  2. Find the largest k such that $p_{(k)} \leq \frac{k}{n}\alpha$.
  3. Reject the null hypothesis for all $H_{(i)}$ for $i = 1, \dots, k$.

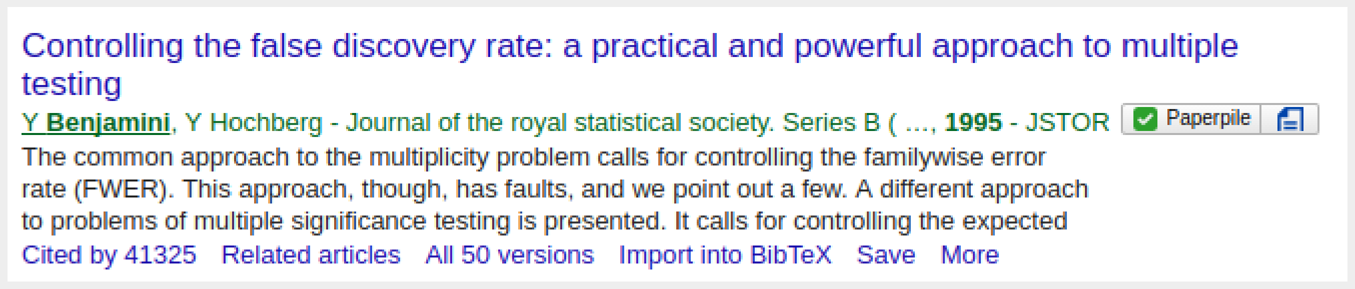
THE MODEL SELECTION PROBLEM



Regression shrinkage and selection via the lasso

R Tibshirani - Journal of the Royal Statistical Society. Series B (..., 1996 - JSTOR

We propose a new method for estimation in linear models. Thelasso'minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge ...



Controlling the false discovery rate: a practical and powerful approach to multiple testing

Y **Benjamini**, Y Hochberg - Journal of the royal statistical society. Series B (..., **1995** - JSTOR

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected

SORTED L-ONE PENALIZED ESTIMATION (SLOPE, BOGDAN ET. AL., ANNALS APPL STAT, 2015)

$$\hat{\beta}_{\mathrm{SLOPE}} = \mathrm{argmin}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{p} \lambda_i |\mathbf{b}|_{(i)},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$; and $|b|_{(1)} \geq |b|_{(2)} \geq \ldots \geq |b|_{(p)}$ denotes the order statistic of the magnitudes of the vector $\mathbf{b} \in \mathbb{R}^p$.

😆 Given $q \in (0, 1)$, there is a procedure to choose $\boldsymbol{\lambda}$ s.t. $\mathrm{FDR}(\hat{\beta}_{\mathrm{SLOPE}}) \leq q$,... 😩 *if the explanatory variables have very small pair-wise correlations.*

GROUP SLOPE MOTIVATION

- Typically, genomic data are highly correlated.
- Often the data can be subdivided into groups with possibly a high within group correlation but a low between group correlation. ~~(Oh really?)~~
- *In case of biomedical data available prior knowledge often provides grouping structures naturally*. E.g., Genomic data: genes or genetic pathways; brain MRI data: anatomical atlases of brain regions; etc.

👍 Select or drop entire groups rather than individual variables. Redefine FDR w.r.t. groups (**gFDR**).

MODEL FORMULATION

- Let $X \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 I)$.
- The predictor variables $\boldsymbol{\beta}$ are divided into J groups of sizes $p_1, p_2, \cdots, p_J$, i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_J^T)^T$ with $\boldsymbol{\beta}_i \in \mathbb{R}^{p_i}$.
- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{i=1}^J X_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}$.

GROUP SLOPE MODEL

FORMULATION 1 (GOSSMANN ET. AL. 2015)

$$\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{J} \lambda_i \sqrt{p_{(i)}}\,\|\mathbf{b}_{(i)}\|_2,$$

where

$$\sqrt{p_{(1)}}\,\|\mathbf{b}_{(1)}\|_2 \geq \sqrt{p_{(2)}}\,\|\mathbf{b}_{(2)}\|_2 \geq \ldots \geq \sqrt{p_{(J)}}\,\|\mathbf{b}_{(J)}\|_2.$$

GROUP SLOPE MODEL

FORMULATION 2 (BRZYSKI ET. AL. 2016)

$$\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{J} \lambda_i \sqrt{p_{(i)}}\,\|X_{(i)}\mathbf{b}_{(i)}\|_2,$$

where

$$\sqrt{p_{(1)}}\|X_{(1)}\mathbf{b}_{(1)}\|_2 \geq \sqrt{p_{(2)}}\|X_{(2)}\mathbf{b}_{(2)}\|_2 \geq \ldots \geq \sqrt{p_{(J)}}\|$$

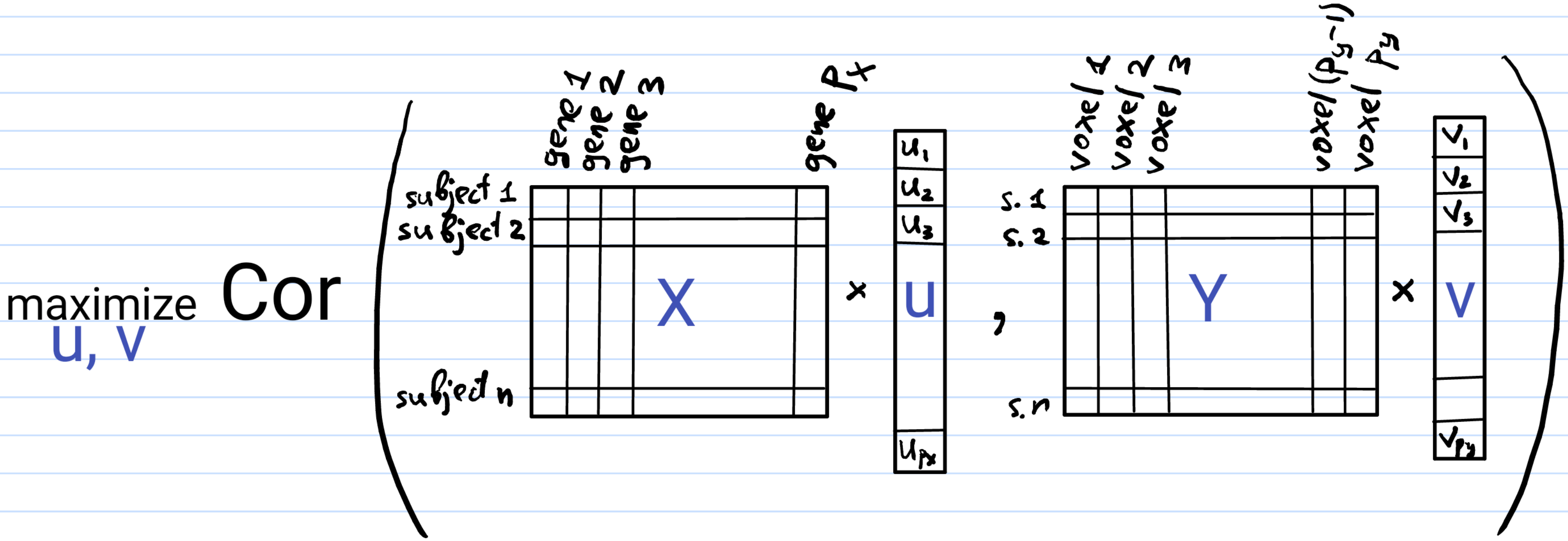
GROUP SLOPE

- Given a user-specified $q \in (0, 1)$, we came up with a procedure to select λ, such that we get $\mathrm{gFDR} \leq q$, if any two variables *from different groups* are nearly uncorrelated (Brzyski, Gossmann, et. al., 2016; Gossmann et. al., 2016).
- The method was applied to DNA sequence data from the Framingham Heart Study, in order to predict bone mineral density and identify genes that influence it (Gossmann et. al., 2016).

GROUP SLOPE REFERENCES

1. Gossmann, A., Cao, S., & Wang, Y.-P. (2015). Identification of Significant Genetic Variants via SLOPE, and Its Extension to Group SLOPE. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, ACM BCB ’15. DOI: 10.1145/2808719.2808743.
2. Gossmann, A., Cao., S., Brzyski, D., Zhao, L.-J., Deng, H.-W., & Wang, Y.-P. (2016). A sparse regression method for group-wise feature selection with false discovery rate control. *(Under review in IEEE/TCBB)*
3. Brzyski, D., Gossmann, A., Su, W., & Bogdan, M. (2016). Group SLOPE — adaptive selection of groups of predictors. arXiv:1610.04960. *(Under review in JASA)*
4. R packages:
    - cran.r-project.org/package=grpSLOPE
    - github.com/agisga/grpSLOPEMC

SPARSE CANONICAL CORRELATION ANALYSIS



$$\underset{u,\,v}{\text{maximize}}\ \text{Cor}\left(X \times u,\ Y \times v \right)$$

subject to sparsity (and other) conditions on u and v.

👉 Find a subset of genes and a subset of brain voxels that are related to each other. 👍

CANONICAL CORRELATION ANALYSIS

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be independent $\mathcal{N}(0, \Sigma_X)$, $y_1, \ldots, y_n \in \mathbb{R}^q$ be independent $\mathcal{N}(0, \Sigma_Y)$, $\mathrm{Cov}(x_k, y_k) = \Sigma_{XY} \in \mathbb{R}^{p \times q}$ for all $k \in \{1, \ldots, n\}$, and that $\mathrm{Cov}(x_k, y_j) = 0$ whenever $k \neq j$.

$$X := \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}, \qquad Y := \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{bmatrix} \in \mathbb{R}^{n \times q}.$$

CLASSICAL CANONICAL CORRELATION ANALYSIS

$$\text{maximize}_{u\in\mathbb{R}^p, v\in\mathbb{R}^q} \widehat{\text{Cov}}(Xu, Yv) = \frac{1}{n}u^T X^T Yv,$$

$$\text{subject to} \quad \widehat{\text{Var}}(Xu) = 1, \widehat{\text{Var}}(Yv) = 1.$$

- Due to Hotelling, 1936.
- The solution is called first pair of canonical vectors.
- Subsequent pairs of canonical vectors are restricted to be uncorrelated with the previous ones.
- The problem is degenerate if $n \leq \max(p, q)$.

SPARSE CCA

- Sparsity in the CCA solution can be achieved by utilizing penalty terms such as the ℓ_1-norm. Unique solution even when $p_X, p_Y \gg n$.
- Witten et. al. (2009):

$$\begin{aligned} \text{maximize}_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \quad & \frac{1}{n} u^T X^T Y v, \\ \text{subject to} \quad & \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, \\ \text{and} \quad & \|u\|_1 \leq c_1, \|v\|_1 \leq c_2. \end{aligned}$$

- Selection of c_1 and c_2 remains a challenging problem.
- Higher-order pairs of canonical vectors can be found by applying sparse CCA to a residual matrix, obtained from $X^T Y$ and the previously found canonical variates.

DEFINING FALSE DISCOVERY RATE (FDR) FOR SPARSE CCA

- Consider the FDR in u and in v separately.
- Consider p_X hypotheses tests $H_i : u_i = 0$.
- The null hypothesis H_i is true if the ith feature in X is uncorrelated with all features in Y, i.e., if
$$(\forall j \in \{1, 2, \ldots, p_Y\}) : \rho_{i,j}^{XY} = 0.$$
- Let $R_{\hat{u}}$ be the number of the rejected H_i, and $V_{\hat{u}}$ the number of false rejections (i.e., when $\hat{u}_i \neq 0$ but $\rho_{i,j}^{XY} = 0$ for all j).
- Define the false discovery rate in u as
$$\mathrm{FDR}(\hat{u}) := \mathbb{E}\left(\frac{V_{\hat{u}}}{\max\{R_{\hat{u}}, 1\}}\right).$$

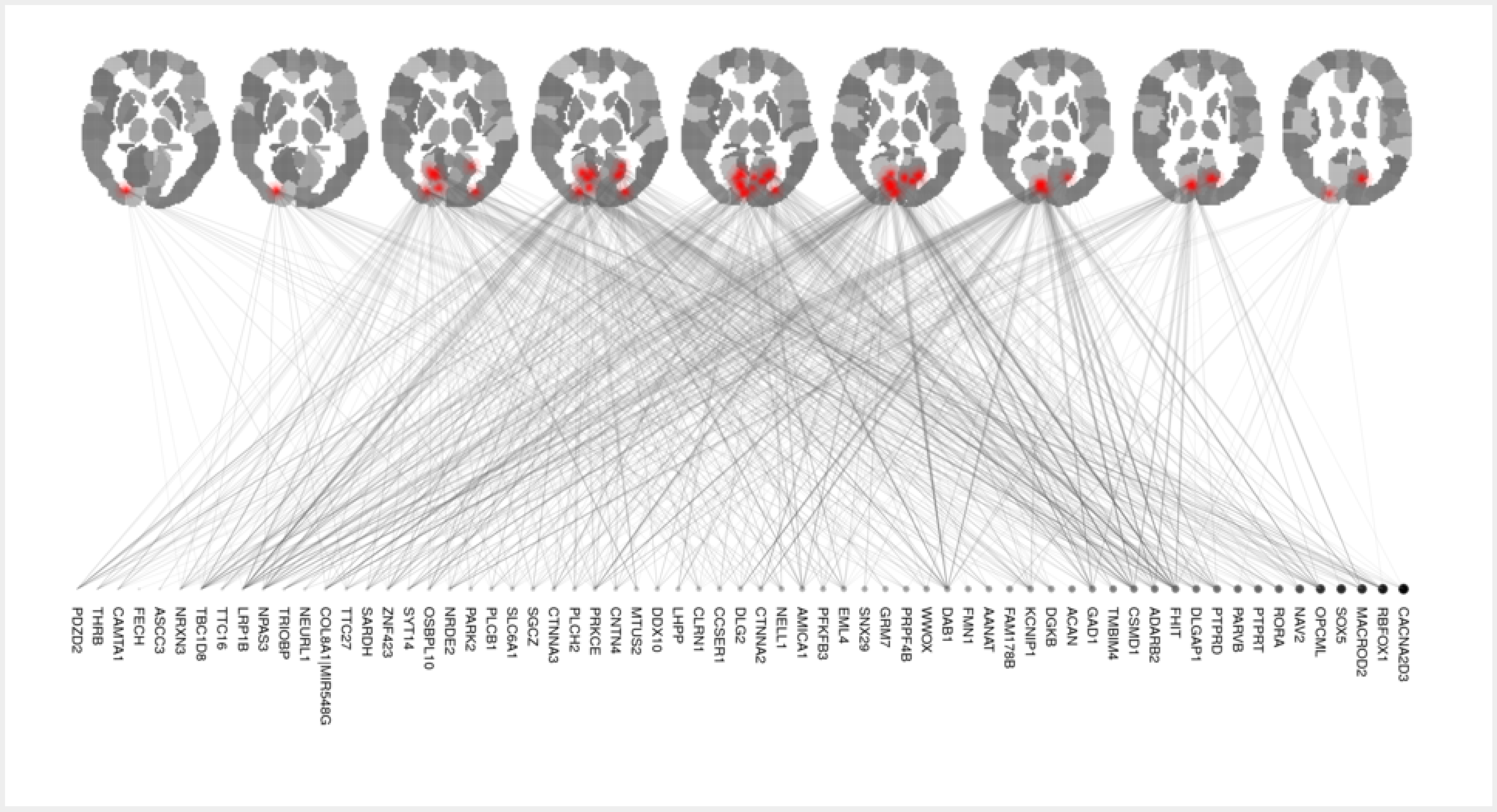
THE FDR-CORRECTED SPARSE CCA PROCEDURE

1. Divide each of X and Y into two subsets of sizes n_0 and n_1:
$$X = \begin{bmatrix} X^{(0)} \\ X^{(1)} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y^{(0)} \\ Y^{(1)} \end{bmatrix}.$$
2. Obtain preliminary sparse CCA estimates $\hat{u}^{(0)}$ and $\hat{v}^{(0)}$ on $X^{(0)}$ and $Y^{(0)}$. Additionally, use $X^{(0)}$ and $Y^{(0)}$ to obtain $\widehat{\Sigma}^{(0)}$, the ML estimate of $\mathrm{Cov}\left(\begin{bmatrix} X & Y \end{bmatrix}\right)$.
3. Obtain p-values using the asymptotic approximation (under the null)
$$\left(\frac{1}{\sqrt{n}}\left(X^{(1)}\right)^T Y^{(1)}\hat{v}^{(0)} \middle| \Sigma = \widehat{\Sigma}^{(0)}\right) \sim \mathcal{N}\left(0, \widehat{\Omega}^{(0)}\right),$$
where $\hat{\mu}^{(0)}$ and $\widehat{\Omega}^{(0)}$ are available in explicit form ($\hat{\mu}^{(0)} = 0$ under the null hypothesis).
4. Apply an FDR correcting procedure (such as BHq), and obtain the FDR-corrected estimates:
$$\hat{u}_i^{(1)} := \begin{cases} \left(X^T Y \hat{v}^{(0)}\right)_i, & \text{for any rejected } H_i^{(u)}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$
$$\hat{v}_j^{(1)} := \begin{cases} \left(Y^T X \hat{u}^{(0)}\right)_j, & \text{for any rejected } H_j^{(v)}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

PREPRINT

- Gossmann, A., Zille, P., Calhoun, V., & Wang, Y.-P. (2017). FDR-Corrected Sparse Canonical Correlation Analysis with Applications to Imaging Genomics. arXiv:1705.04312 [pdf] *(under review in IEEE/TMI)*
- Associated code: https://github.com/agisga/FDRcorrectedSCCA

APPLICATION TO IMAGING GENOMICS



Data: The Philadelphia Neurodevelopmental Cohort (PNC) is a large-scale collaborative study between the Brain Behaviour Laboratory at the University of Pennsylvania and the Children's Hospital of Philadelphia. It contains, among other modalities, a fractal n-back fMRI task, and SNP arrays for over 900 adolescents.

IMAGING GENOMICS RESULTS

- We group the selected voxels using the region of interest (ROI) definitions of the AAL parcellation. The findings correspond to the *middle occipital gyri*, *left and right calcarine sulcus*, and *left cuneus* (3 voxels). Similar brain regions have been found in other fMRI studies of working memory.
- A literature search confirmed that a majority of the identified genes (at least 34 out of the 65) have been previously associated with various aspects of human cognitive function.

THE END

THANK YOU