REGAINING CONTROL OF FALSE FINDINGS IN FEATURE SELECTION, CLASSIFICATION, AND PREDICTION ON NEUROIMAGING AND GENOMICS DATA

Oral defense of a dissertation submitted to the **Bioinnovation PhD Program** of the **School of Science and Engineering** of **Tulane University** in partial fulfillment of the requirements for the PhD degree by

ALEXEJ GOSSMANN

July 11, 2018

PRECISION MEDICINE

Inter-personal diversity in the patients' biology

⇀ "personalized" treatment plans

↑ Quality of healthcare

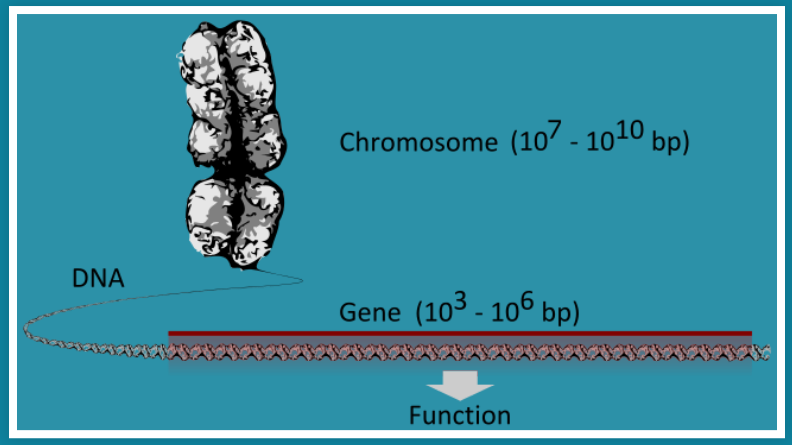
↓ Treatment time and cost

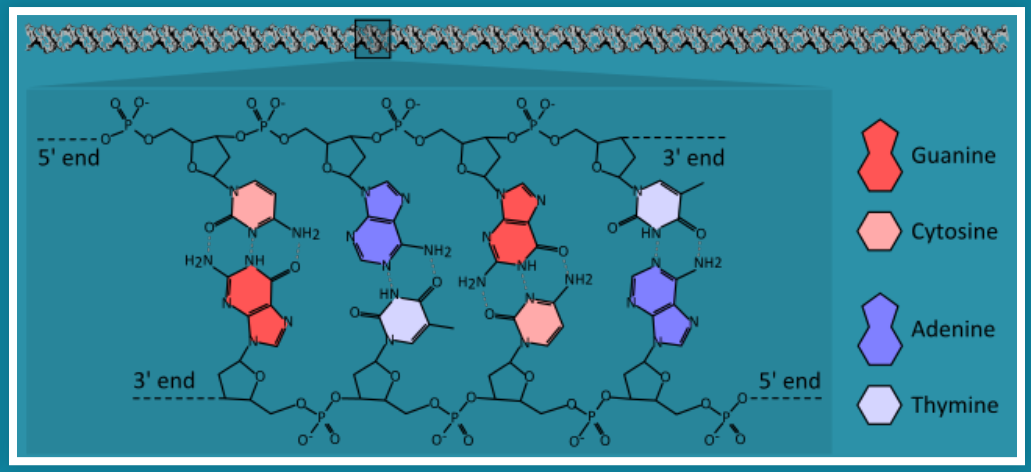
PRECISION MEDICINE

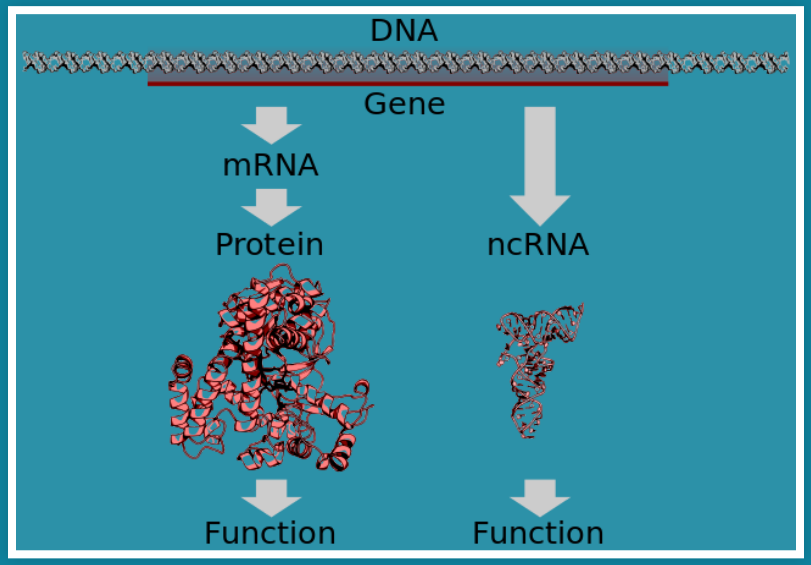
Made possible by:

1. Big data including **genomics** and **neuroimaging**.
2. Computational methods including **machine learning** and **modern statistics**.

From left to right: (1) gene region on a chromosome; (2) chemical structure of DNA; (3) transcription/translation of genes into ncRNA, mRNA, protein.

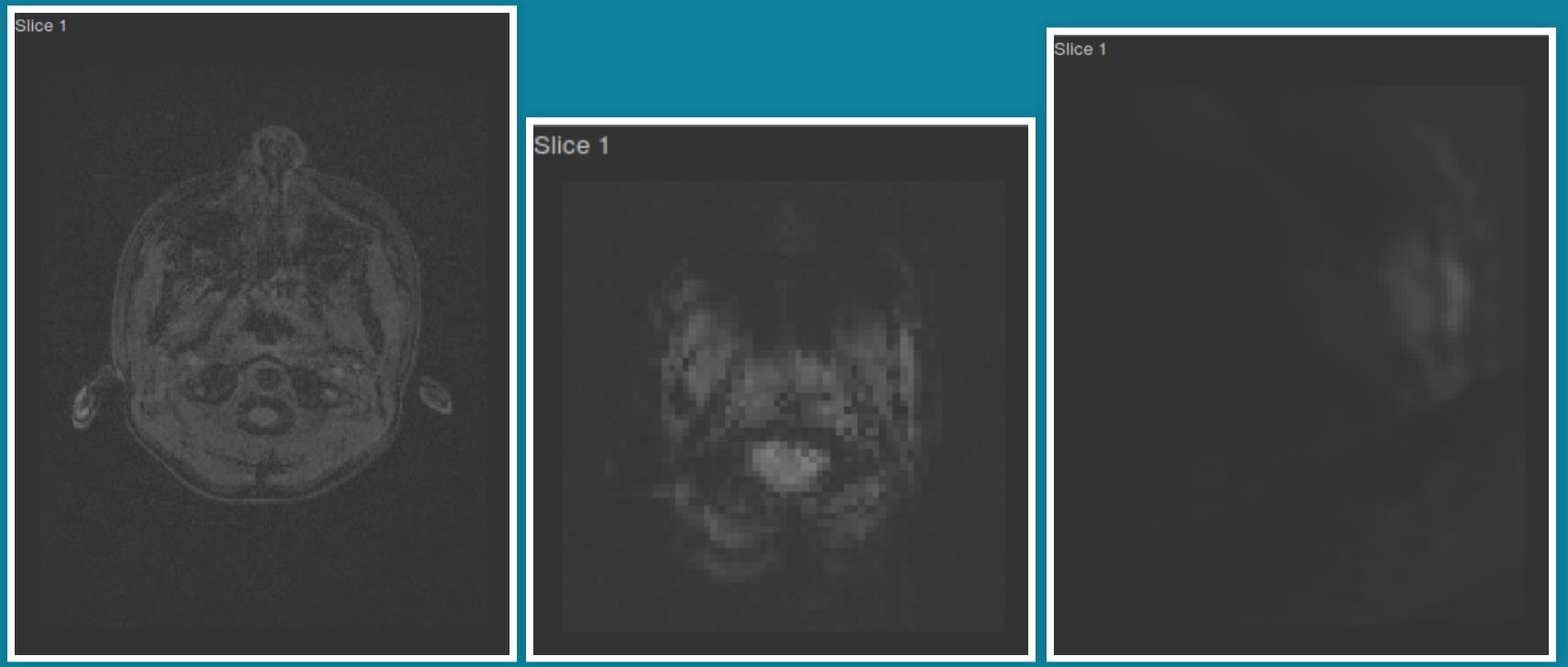






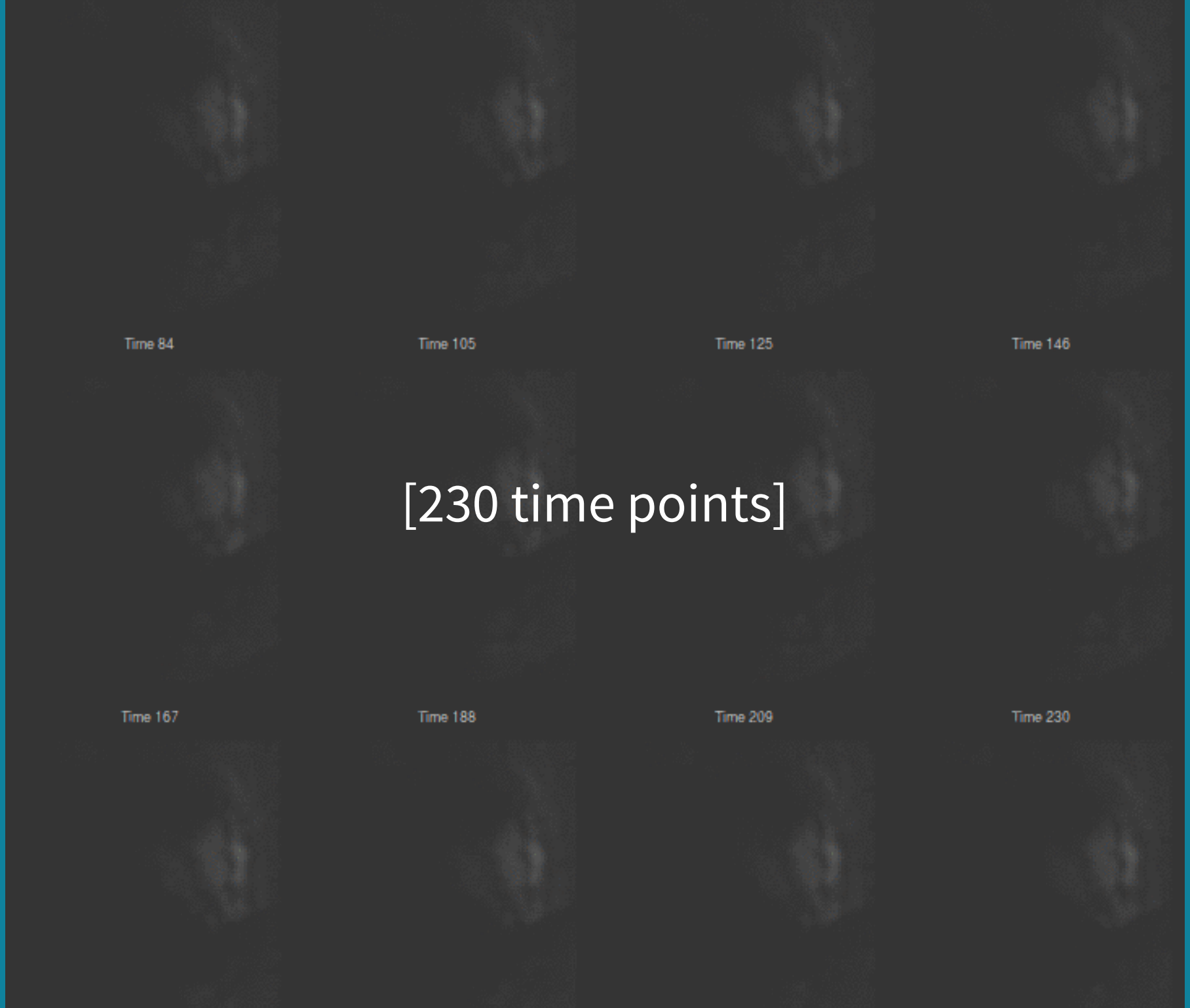
Source: Images by Thomas Shafee [CC BY 4.0] via Wikimedia Commons.

- Structural MRI: anatomical structure of the brain.
- Functional MRI: brain activity associated with blood flow related to energy use by brain cells across time.

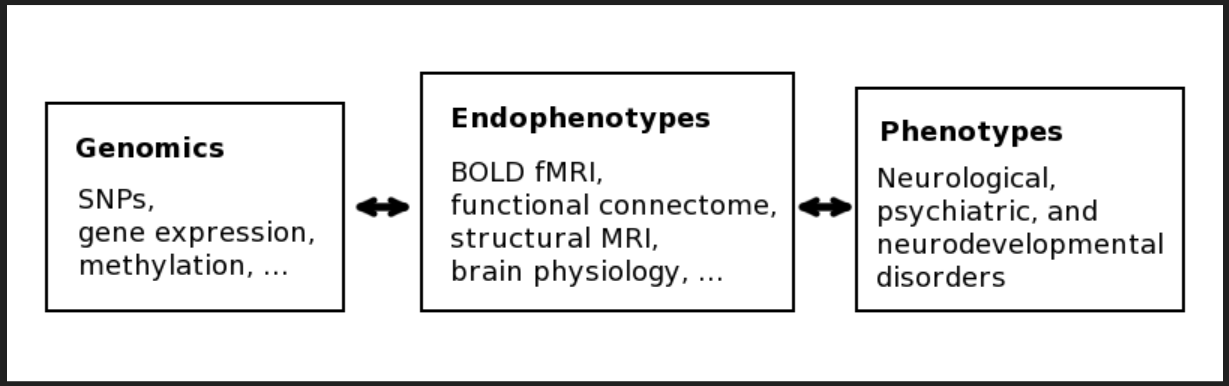


A randomly chosen subject from the Philadelphia Neurodevelopmental Cohort:

- T1-weighted MRI before preprocessing ($192 \times 256 \times 160$ voxels).
- n-back task BOLD fMRI (i.e., T2*-weighted) before preprocessing ($64 \times 64 \times 49$ voxels), and after preprocessing ($79 \times 95 \times 79$ voxels).

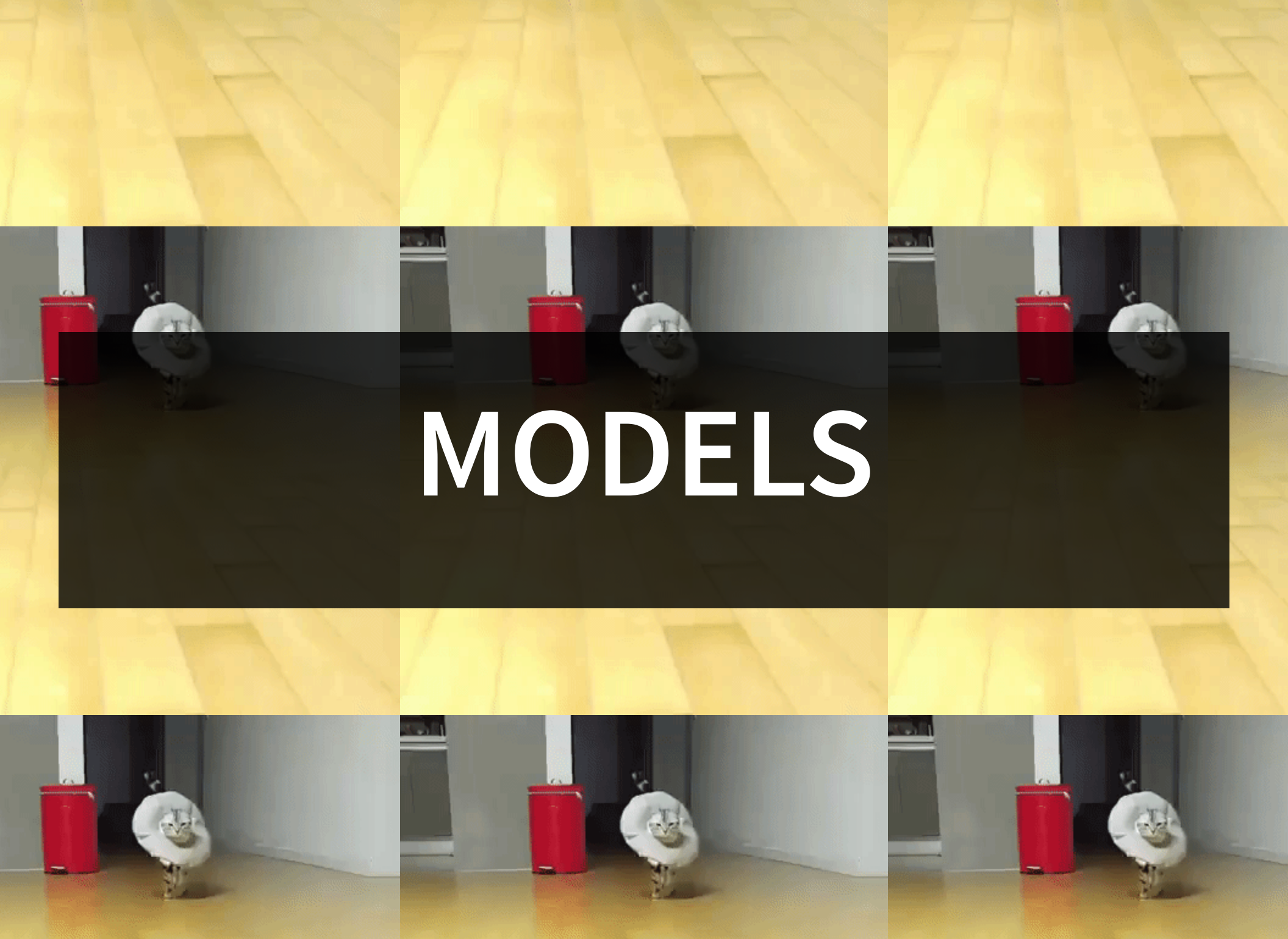

Time 84
Time 105
Time 125
Time 146
[230 time points]
Time 167
Time 188
Time 209
Time 230

PRECISION MED. & MENTAL DISORDERS



- Neuroimaging as an endophenotype.[1-2]
- Use of fMRI to aid diagnosis, or to monitor & guide drug treatment.[3-5]

[1]: Hashimoto et. al., 2015, [2]: Poline et. al., 2015, [3]: Weickert et al., 2004, [4]: Apud et al., 2007, [5]: Goldstein-Piekarski et al., 2016.


MODELS

MODELS

$$y = f_{\boldsymbol{\theta}}(x_1, x_2, \ldots, x_p) + \varepsilon,$$

where $x_1, x_2, \ldots, x_p$ are predictor variables, ε is random noise, y is the phenotype, and $\boldsymbol{\theta}$ is a vector of parameters to be estimated.

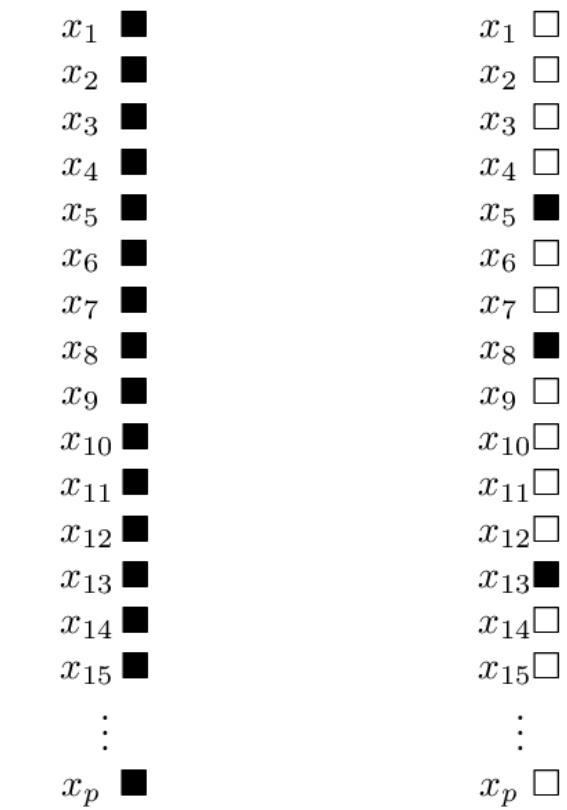
Human DNA $\approx 3 \cdot 10^9$ base pairs ⇝ vast majority not related to phenotype of interest ⇝ *sparse models*

$$\Rightarrow y = f_{\tilde{\boldsymbol{\theta}}}(x_{a_1}, x_{a_2}, \ldots, x_{a_m}) + \varepsilon,$$

where $\{a_1, a_2, \ldots, a_m\} \subset \{1, 2, \ldots, p\}$ is a small subset ($m \ll p$).

SPARSE MODELS

$$(y \,|\, x_1, x_2, x_3, \ldots, x_p) \quad = \quad (y \,|\, x_5, x_8, x_{13})$$



THE TWO-FACED MODEL SELECTION PROBLEM

➡️ **Prediction:**

Find best predictions for y.

➡️ **Feature selection:**

Which x_j are predictive?

TWO TYPES OF FALSE FINDINGS

😱 **False positives.**

Overfitting.

🙀 **False discoveries.**

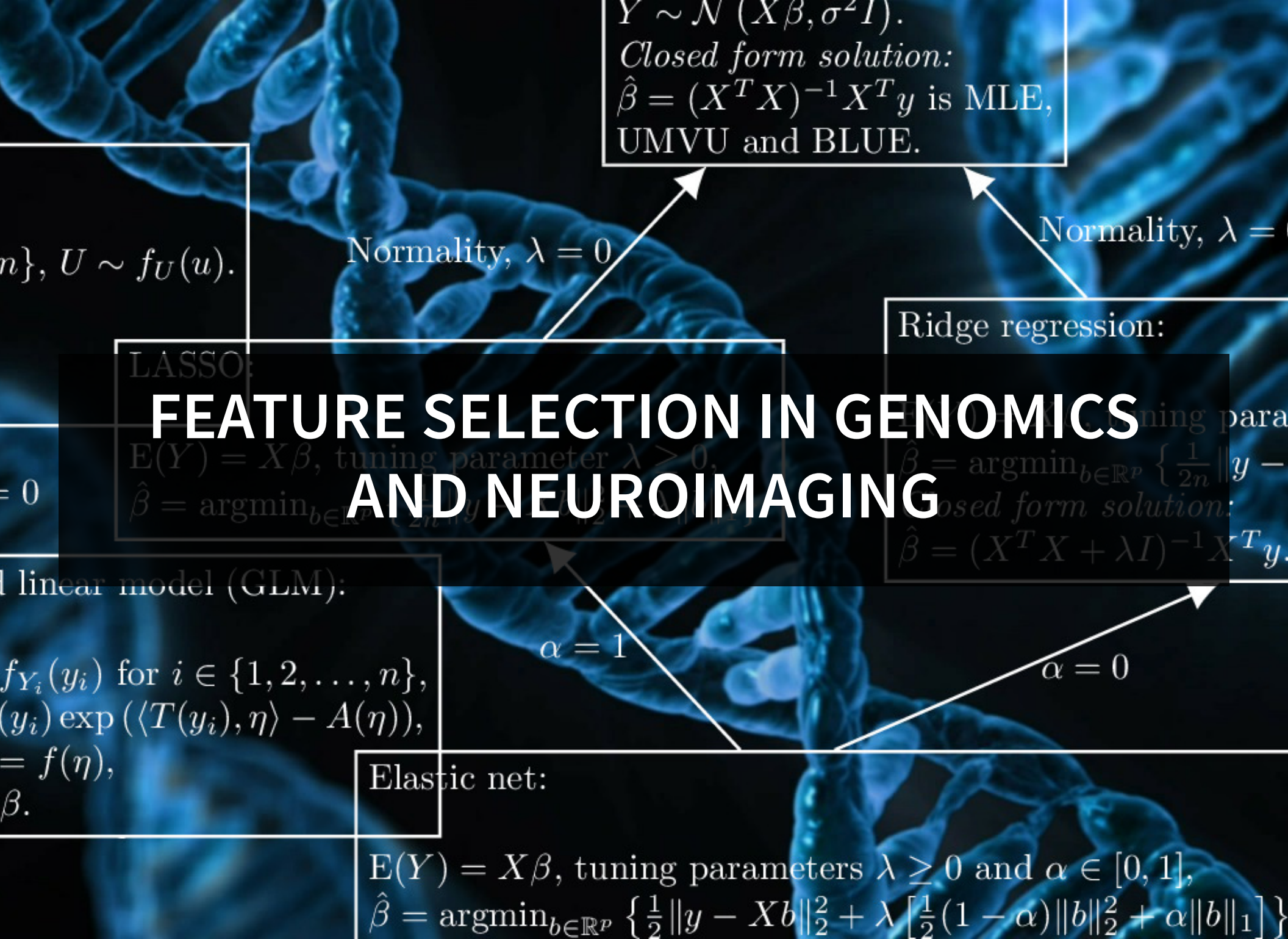
Curse of dimensionality.

AIMS

Establish **guarantees** on...

- false discoveries in feature selection,
- false predictions on new data (generalization)

...for types of methods commonly used in the analysis of genomic and neuroimaging data.



FEATURE SELECTION IN GENOMICS AND NEUROIMAGING

MULTIPLE HYPOTHESES TESTING

Feature selection as testing of hypotheses:

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

- $\beta_i :=$ effect of ith feature.
- $R :=$ number of rejected hypotheses.
- $V :=$ number of false rejections (i.e., Type I errors).
- **Family-wise error rate:** $\mathrm{FWER} = \mathbb{P}(V \geq 1)$.[1]
- **False discovery rate:** $\mathrm{FDR} = \mathbb{E}\left(\frac{V}{\min\{R,1\}}\right)$.[2]

[1]: E.g., Bonferroni, Holm (1979), Hommel (1988).
[2]: E.g., Benjamini-Hochberg (1995), Benjamini-Yukutieli (2001).

SPARSE REGRESSION

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

- $\mathbf{y} = f(X) + \boldsymbol{\varepsilon} \approx f(X) = X\boldsymbol{\beta} \approx X\hat{\boldsymbol{\beta}}$.
- Yields a sparse solution $\hat{\boldsymbol{\beta}}$.
- Computationally efficient (convex).
- Very useful in practice.

Problem: how to do statistical inference on $\hat{\boldsymbol{\beta}}$?

Problem: how sparse should $\hat{\boldsymbol{\beta}}$ be?

[1]: Tibshirani, JRSSB, 1996.

SORTED L-ONE PENALIZED ESTIMATION[1]

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

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. $\mathrm{FDR}(\hat{\boldsymbol{\beta}}_{\mathrm{SLOPE}}) \leq q$ is guaranteed. *...if the explanatory variables have very small pair-wise correlations.* ← *typically not the case in genomics.*[2]

[1]: Bogdan et. al., Annals Appl Stat, 2015. [2]: Gossmann et. al., ACM BCB, 2015.

GROUP SLOPE MOTIVATION

- Divide the data into groups by correlation. ← *Often possible for biological data.*
- Then select/drop entire groups rather than individual variables.
- Redefine FDR w.r.t. groups: **gFDR**.

GROUP-WISE FALSE DISCOVERY RATE

- $\mathrm{Rg} :=$ "total # discovered groups"
- $\mathrm{Vg} :=$ "# falsely discovered groups"

We define

$$\mathrm{gFDR} := \mathrm{E}\left(\frac{\mathrm{Vg}}{\max(\mathrm{Rg},1)}\right).$$

GROUP SLOPE

- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, X \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 I)$
  .
- $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}_i \in \mathbb{R}^{p_i}$.

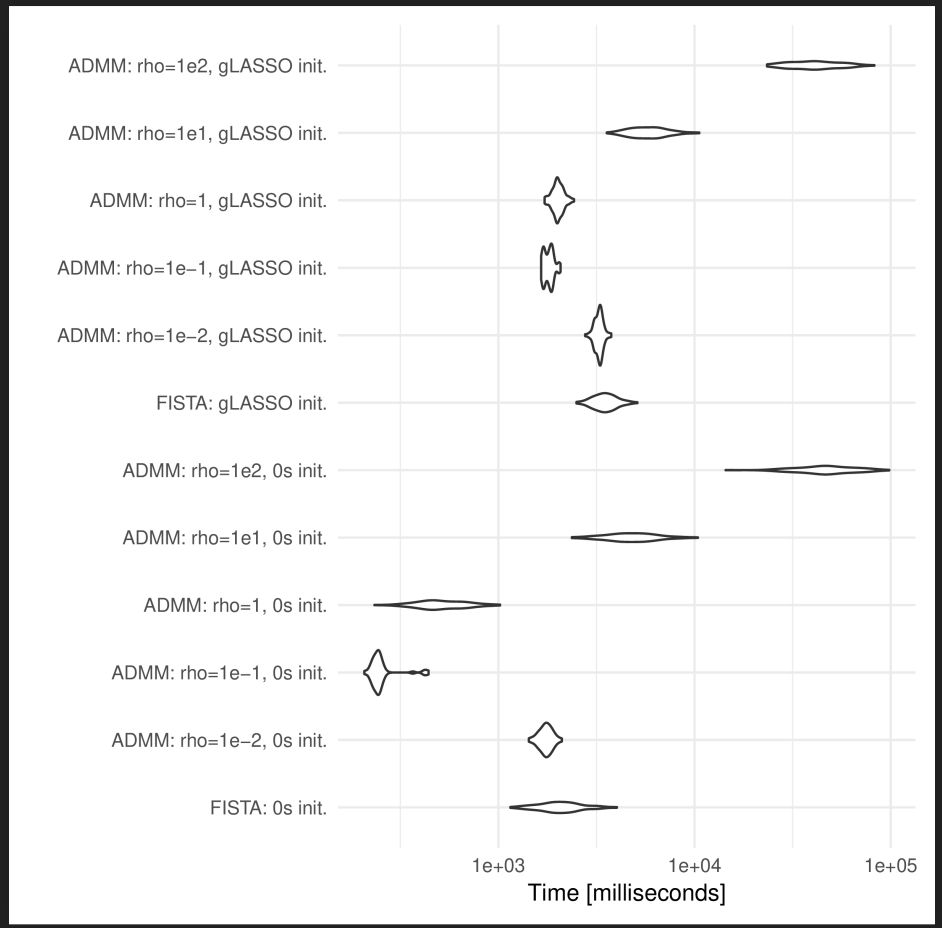
$$\min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{J} \lambda_i \sqrt{p_{(i)}} \left\|X_{(i)} \mathbf{b}_{(i)}\right\|_2,$$

where

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

Different ways to find the global solution:

- *Fast iterative shrinkage-thresholding algorithm (FISTA)* — a proximal gradient method used in [1-3].
- *Alternating direction method of multipliers (ADMM)* — derived in the thesis.



[Figure 3.2 in the thesis]

[1]: Gossmann et. al., 2015. [2]: Brzyski, Gossmann, et. al., 2018. [3]: Gossmann et. al., 2017.

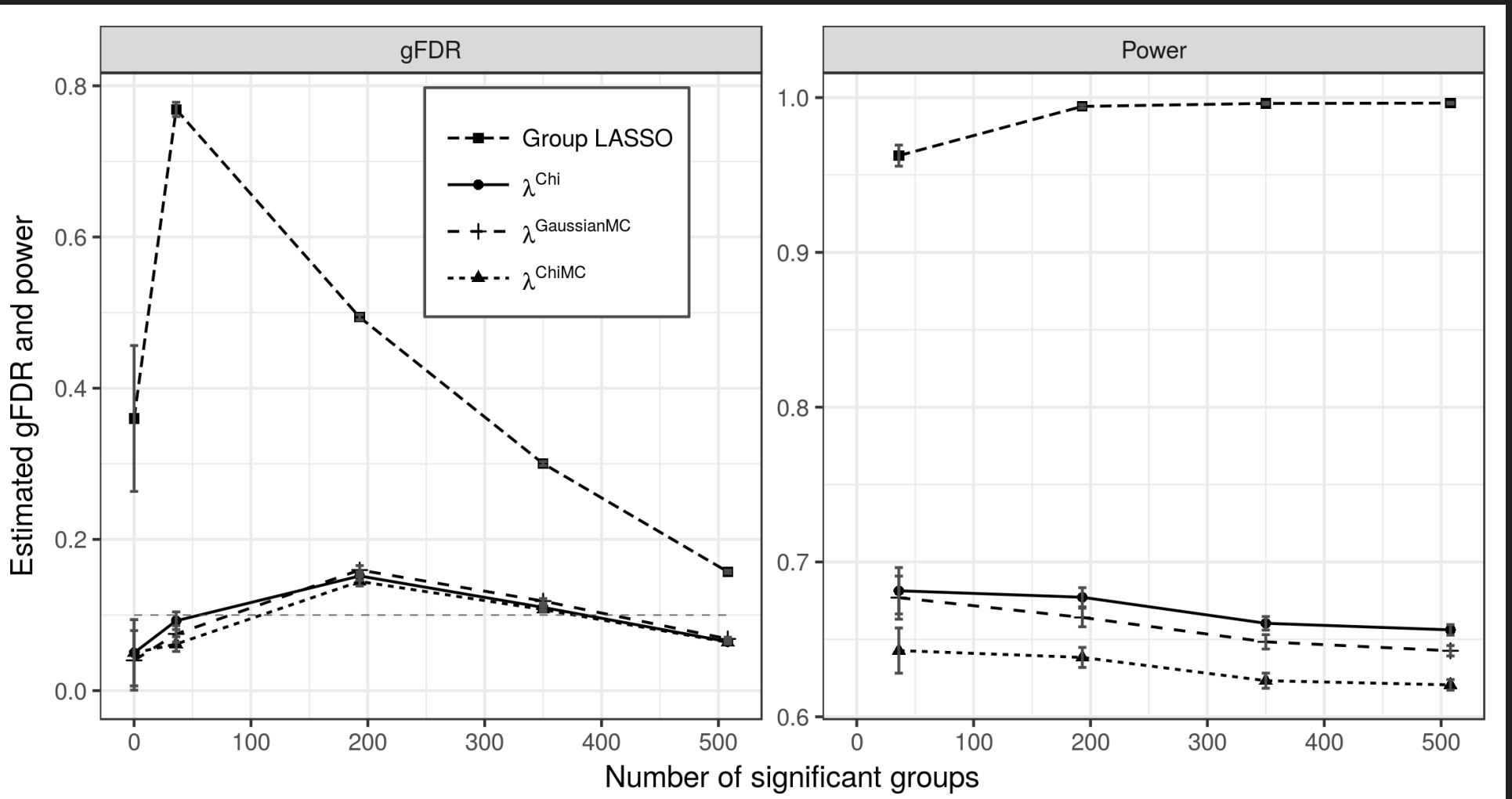
GROUP SLOPE - THEORETICAL GUARANTEES

Given a user-specified $q \in (0, 1)$, we show how to choose $\boldsymbol{\lambda}$ such that $\mathrm{gFDR} \leq q$.[1-3]

Different approaches: theoretical for orthogonal designs[2-3], heuristic based on theory for general designs[1-2], Monte Carlo based for general designs[3].

➡ Confirmed with extensive simulation studies on synthetic and real data.[1-3]

[1]: Gossmann et. al., 2015. [2]: Brzyski, Gossmann, et. al., 2018. [3]: Gossmann et. al., 2017.



$X \in \mathbb{R}^{8915 \times 5976}$ contains real DNA sequence data of chromosome 22; 726 groups (lengths from 1 to 1657, mean=8.23, median=1); between-group correlation <0.3; *simulated response* $\mathbf{y} = X\beta + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, I)$. [Figure 3.5 in the thesis]

APPLICATION - FRAMINGHAM COHORT ANALYSIS[1]

- SNP data for 8915 subjects.
- 1771 subjects have corresponding spine BMD measurements.
- The remaining ~7000 subjects used to group SNPs.

X with dimensions 1771×117933, consisting of 6403 groups of average size 18.42 (median size 2).

[1]: Gossmann et. al., 2017.

GROUP SLOPE RESULTS

- 40 SNPs were selected by Group SLOPE with target gFDR $q = 0.1$, and mapped to nearby genes.
- 15 genes reported in previous studies:
  - BMD (SMOC1, RPS6KA5, FGFR2, GAA, SCN1A, RAB5A, SOX1, and A2BP1),
  - osteoarthritis (A2BP1, ADAM12, MATN1),
  - lumbar disc herniation (KIAA1217),
  - osteopetrosis (VAV3),
  - biology of osteoclasts, osteoblasts and osteogenesis (VAV3, SLC7A7, ADAM12, PPARD, FGFR2, PTPRU, SMOC1).

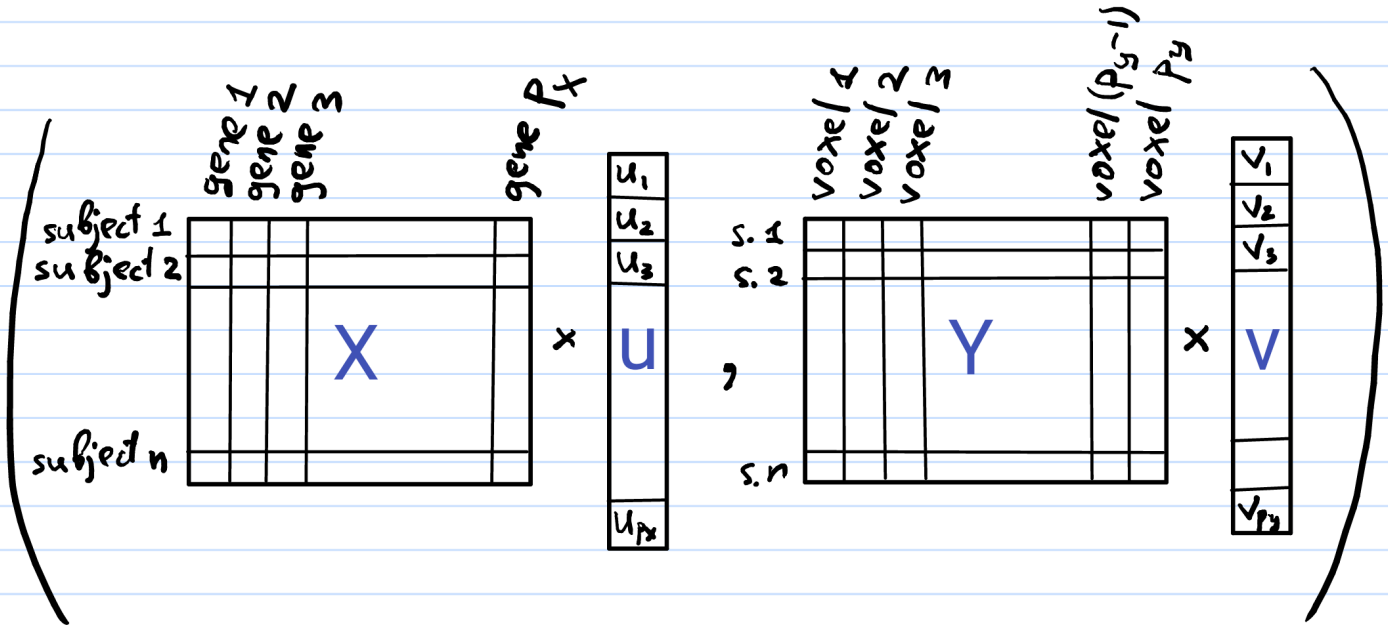
Group SLOPE — Some further topics:

- Performance of SLOPE on DNA seq data.[1]
- An alternative formulation of Group SLOPE.[1, 3]
- Theoretical gFDR control under orthogonal groups.[2-3]
- Error variance estimation: *Scaled sparse linear regression* vs. *EigenPrism*.
- Analysis of runtime on large DNA seq datasets.[3]
- Genomic data preprocessing.[3]
- Group SLOPE is asymptotically minimax w.r.t. estimation.[2]

[1]: Gossmann et. al., 2015. [2]: Brzyski, Gossmann, et. al., 2018. [3]: Gossmann et. al., 2017.

CANONICAL CORRELATION ANALYSIS

maximize Cor
u, v



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. 👍

CLASSICAL CCA [HOTELLING, 1936]

Consider two matrices (i.e., datasets)
$X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ (with columns centered).

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

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

The problem is degenerate if $n \leq \max(p, q)$.

+ Sparsity assumption on biological data

SPARSE CCA[1-2]

$$\text{maximize}_{\mathbf{u}\in\mathbb{R}^p,\mathbf{v}\in\mathbb{R}^q}\ \frac{1}{n}\mathbf{u}^T X^T Y \mathbf{v},$$

subject to

$$\|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1, \|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2.$$

- Unique solution even when $p_X, p_Y \gg n$.

election of the sparsity parameters remains a
allenging problem.

[1]: Witten et. al., 2009, [2]: Parkhomenko et. al., 2009.

SPARSE CCA

 Select sparsity parameters in a data-driven fashion, such that FDR is controlled.

DEFINING FDR FOR SPARSE CCA

- Adapt the widely-used FDR concept from multiple hypotheses testing to CCA.[1]

> *The expected proportion of "discoveries" that are false.*

- Consider FDR in $\mathbf{u}$ and in $\mathbf{v}$ separately.

[1]: Benjamini & Hochberg, JRSSB, 1995.

DEFINING FDR FOR SPARSE CCA

The coefficient estimate $\hat{u}_i \neq 0$ represents a ***false discovery*** *of the ith feature of X, if u_i doesn't affect the value of $Cov(\boldsymbol{x} \cdot \boldsymbol{u}, \boldsymbol{y} \cdot \boldsymbol{v})$, or equivalently if*

$$\hat{u}_i \neq 0 \quad \textbf{and} \quad E(X^T Y \boldsymbol{v})_i = 0.$$

DEFINING FDR FOR SPARSE CCA

- $\mathrm{R}_{\hat{\mathbf{u}}} =$ the number of non-zero elements in $\hat{\mathbf{u}}$,
- $\mathrm{V}_{\hat{\mathbf{u}}} =$ the number of false discoveries.

Define

$$\mathbf{FDR}(\hat{\mathbf{u}}) := \mathbb{E}\left(\frac{\mathrm{V}_{\hat{\mathbf{u}}}}{\max\{\mathrm{R}_{\hat{\mathbf{u}}}, 1\}}\right).$$

DEFINING GFDR FOR SPARSE CCA

Analogously we define the group-wise false discovery rate (gFDR).

- $\mathrm{Rg}_{\hat{\mathbf{u}}} =$ the number of non-zero groups of elements in $\hat{\mathbf{u}}$,
- $\mathrm{Vg}_{\hat{\mathbf{u}}} =$ the number of falsely discovered groups.

Define

$$\mathrm{gFDR}(\hat{\mathbf{u}}) := \mathbb{E}\left(\frac{\mathrm{Vg}_{\hat{\mathbf{u}}}}{\max\{\mathrm{Rg}_{\hat{\mathbf{u}}}, 1\}}\right).$$

SLOPECCA AND GSLOPECCA

slopeCCA:

$$\text{minimize}_{\mathbf{u}\in\mathbb{R}^p,\mathbf{v}\in\mathbb{R}^q} \ \left\{-\mathbf{u}^T X^T Y \mathbf{v} + \sqrt{n}\, J_{\boldsymbol{\lambda}^u}(\mathbf{u}) + \sqrt{n}\, J_{\boldsymbol{\lambda}^v}(\mathbf{v})\right\},$$
$$\text{subject to } \|\mathbf{u}\|_2^2 \le 1, \|\mathbf{v}\|_2^2 \le 1.$$

gslopeCCA:

$$\text{minimize}_{\mathbf{u}\in\mathbb{R}^p,\mathbf{v}\in\mathbb{R}^q} \left\{-\mathbf{u}^T X^T Y \mathbf{v} + \sqrt{n}\, J_{\boldsymbol{\lambda}^u}\left((\|\mathbf{u}_1\|_2, \ldots)^T\right) + \sqrt{n}\, J_{\boldsymbol{\lambda}^v}\left((\|\mathbf{v}_1\|_2, \ldots)^T\right)\right.$$
$$\text{subject to } \|\mathbf{u}\|_2^2 \le 1, \|\mathbf{v}\|_2^2 \le 1.$$

Where $J_{\boldsymbol{\lambda}}(\mathbf{u}) = \Sigma_{i=1}^p \lambda_i |u|_{(i)}$ is the Sorted L1 Norm.

SLOPECCA AND GSLOPECCA - COMPUTATIONAL METHODS

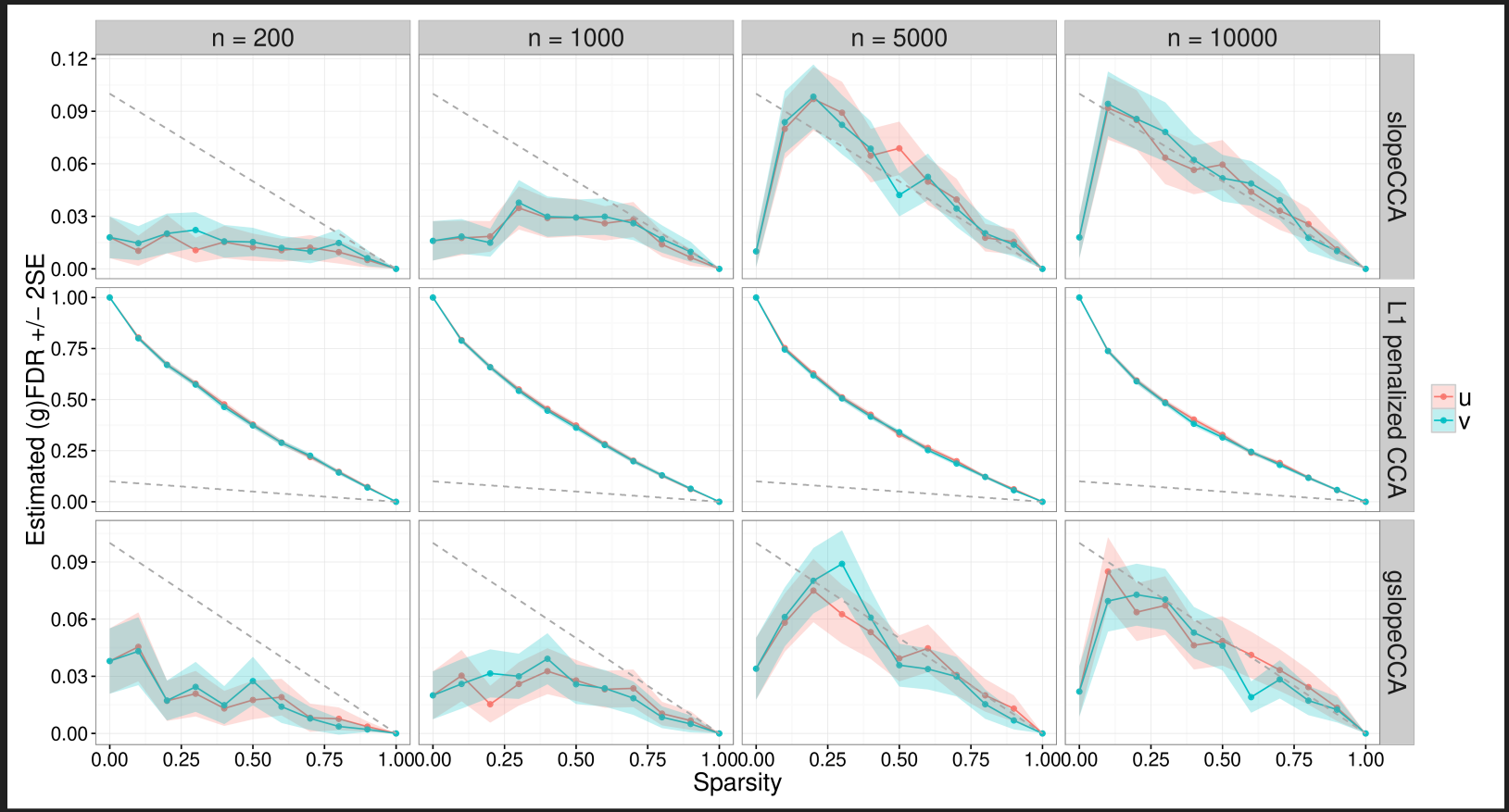
Both are **biconvex optimization problems**.

Alternating optimization algorithms are guaranteed to converge.

SLOPECCA AND GSLOPECCA - THEORETICAL GUARANTEES

Asymptotic (i.e., as $n \to \infty$) FDR and gFDR guarantees if $Cov(X)$ and $Cov(Y)$ are diagonal or block-diagonal.

($Cov(X, Y)$ can be of arbitrary shape)

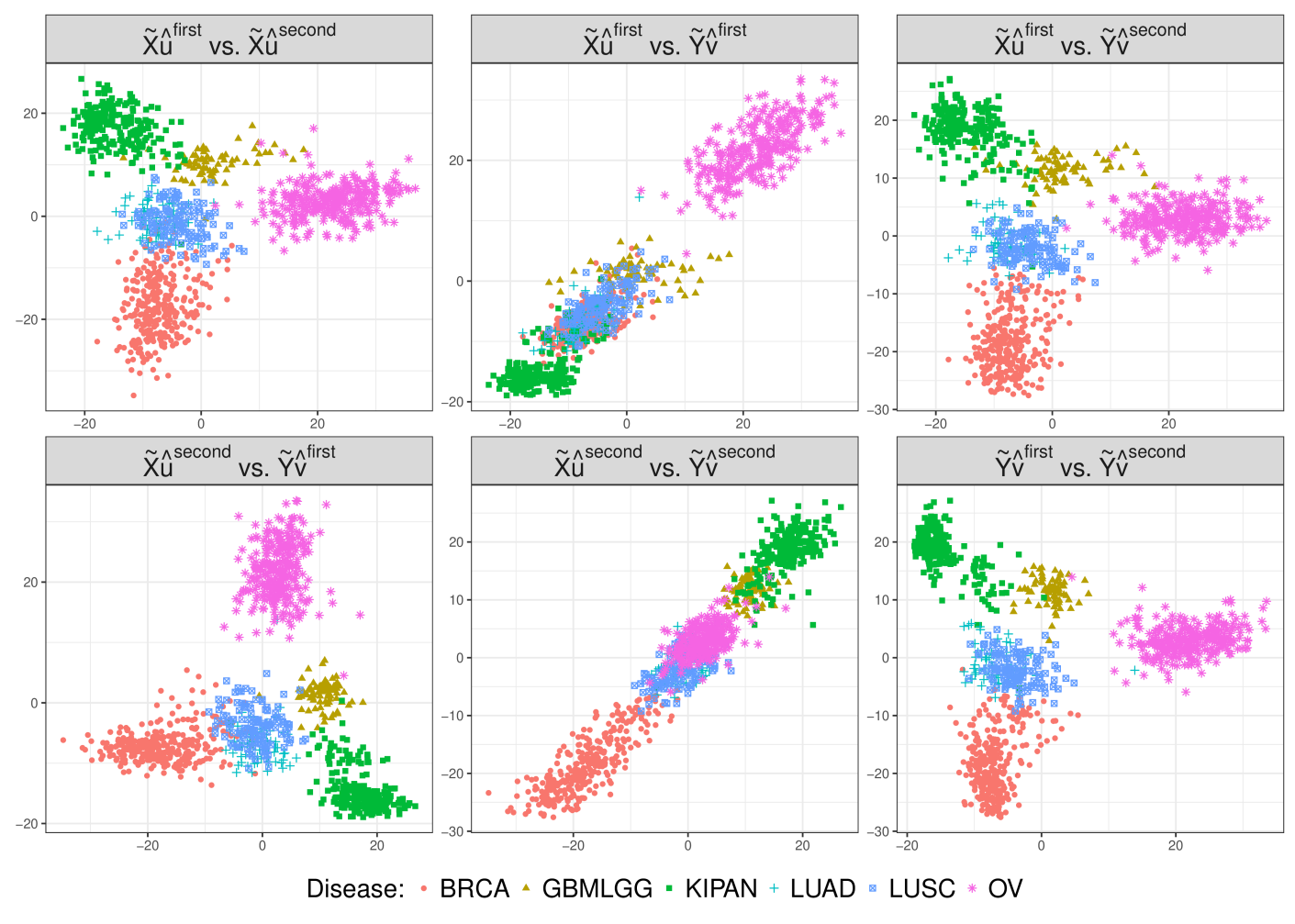


Simulation studies: FDR and gFDR of slopeCCA, gslopeCCA, and ℓ_1-penalized CCA solutions; $n = 200, 1000, 5000, 10000$ and $p_X = p_Y = 300$; 11 evenly spaced sparsity levels between 0 (i.e., X is uncorrelated with Y) and 1 (i.e., every feature of X is correlated to some feature of Y, and vice versa); 500 independent simulation runs; dashed line represents theoretical bounds for slopeCCA and gslopeCCA.
[Figure 4.1 in the thesis]

APPLICATION TO TCGA

The Cancer Genome Atlas (TCGA):

- NIH initiative since 2005.
- Coordinated data collection at 20 collaborating institutions in U.S. and Canada.
- Genomic samples from tumor cells of different cancers and matched normal cells.
- Data include: gene expression, methylation, CNV, SNP, microRNA, and whole genome, exon, or transcriptome sequencing.
- cancergenome.nih.gov/



Canonical variates estimated by gslopeCCA on methylation and mRNA data from TCGA ($n = 1109, p_X = 24981, p_Y = 20255$). Observations are colored according to cancer type, revealing clear differences.
[Figure 4.2 in the thesis]

TCGA — CONCLUSIONS

- Canonical variates reveal differences between cancer types.
- Using the four canonical variates as the only predictors to classify cancer type yields classification accuracy over 93% on test data.

SLOPECCA AND GSLOPECCA

Some further topics covered in the thesis:

- Optimization algorithms.
- Asymptotic normality results.
- Genomic data pre-processing.


Time 84
Time 105
Time 125
Time 146
FDR-CORRECTED
SPARSE CCA
Time 167
Time 188
Time 209
Time 230

FDR-CORRECTED SPARSE CCA

Motivation: The assumption of slopeCCA and gslopeCCA that $Cov(X)$ and $Cov(Y)$ are diagonal or block-diagonal is too restrictive in many cases.

FDR-CORRECTED SPARSE CCA

A split-sample, two-step procedure:[1]

1. Split the data in two parts.
2. **Using the first subsample:** obtain initial estimates $\hat{\mathbf{u}}^{(0)}$ and $\hat{\mathbf{v}}^{(0)}$ using conventional sparse CCA.
3. **Using the second subsample:** test hypotheses of the form,
   $$\mathrm{H}_i^{(u)} : \mathrm{E}\big(X^T Y \mathbf{v}\big)_i = 0, \quad \mathrm{H}_j^{(v)} : \mathrm{E}\big(Y^T X \mathbf{u}\big)_j = 0,$$
   and adjust for multiple comparisons to control FDR.

[1]: Gossmann et. al., IEEE TMI, 2018.

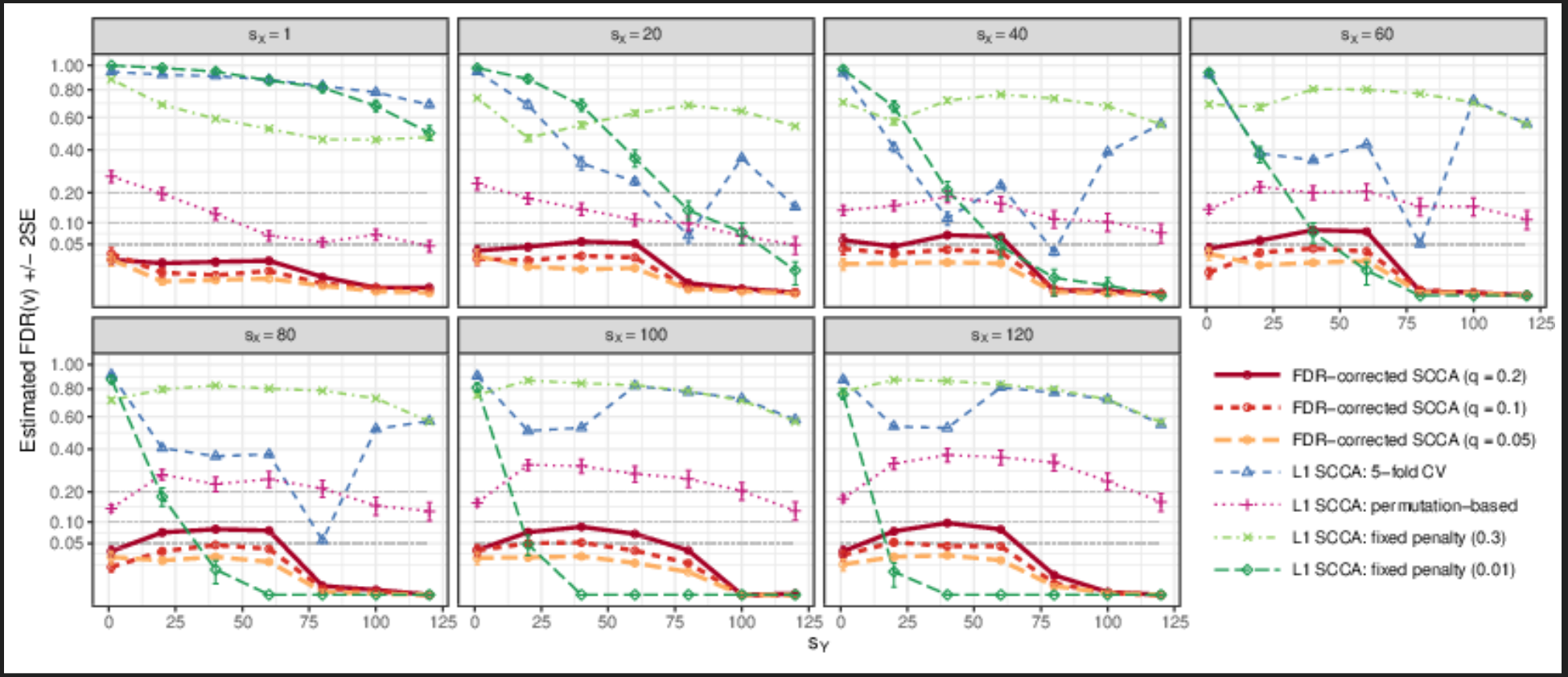
FDR-CORRECTED SPARSE CCA - THEORETICAL FDR GUARANTEES

1. Asymptotic theory allows us to approximate the distributions of $X^T Y \hat{\mathbf{v}}^{(0)}$ and $Y^T X \hat{\mathbf{u}}^{(0)}$.
2. We can use the Benjamini-Hochberg procedure to control the FDR.

FDR control confirmed with extensive simulation studies on synthetic and real data.[1]

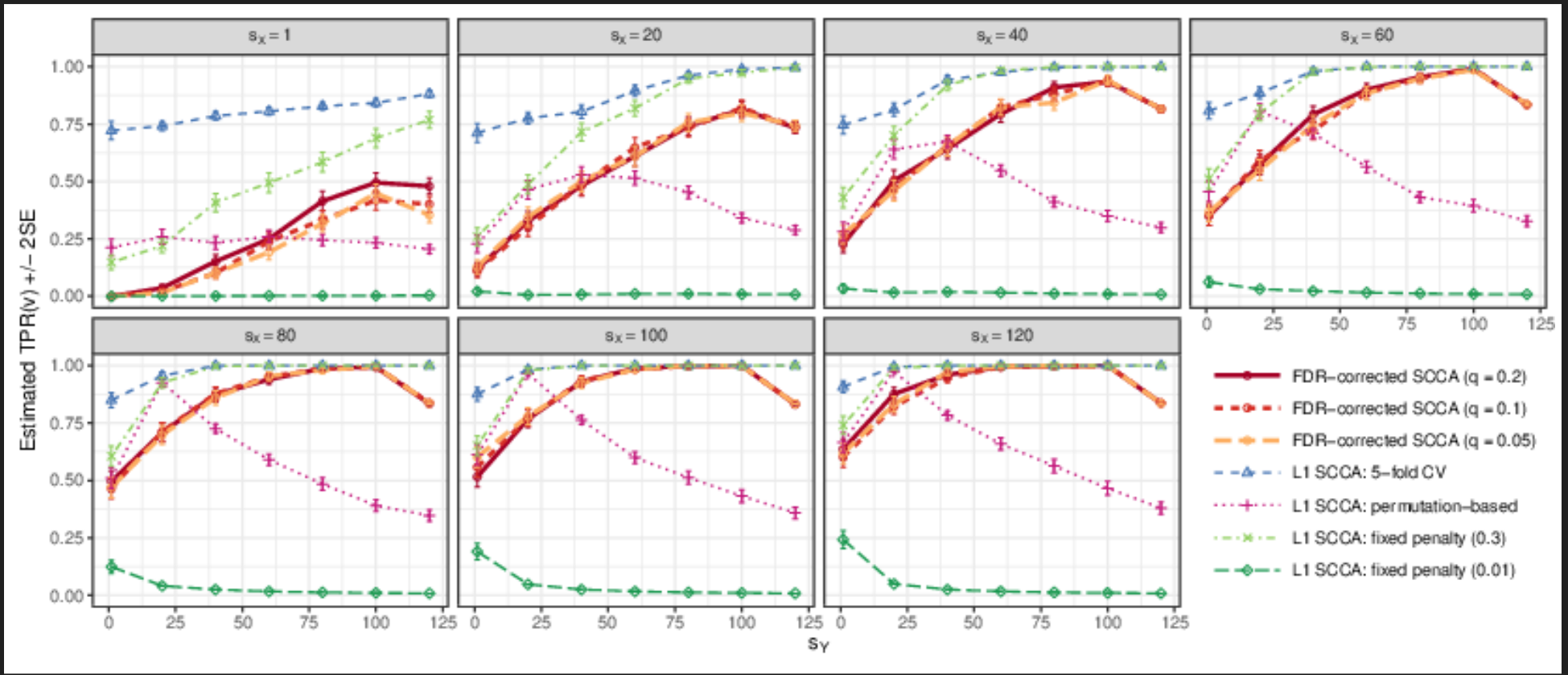
[1]: Gossmann et. al., IEEE TMI, 2018.

SIMULATION STUDIES



$X, Y \in \mathbb{R}^{600 \times 1500}$. FDR is controlled regardless of the sparsity of the true solution.

SIMULATION STUDIES



True positive rate is on par with competing methods.

FDR-CORRECTED SCCA - APPLICATION

- Diversity in brain activity and brain connectivity in children and adolescents.
- What are the driver genes?
- Relationship to neurodevelopmental and psychiatric disorders.

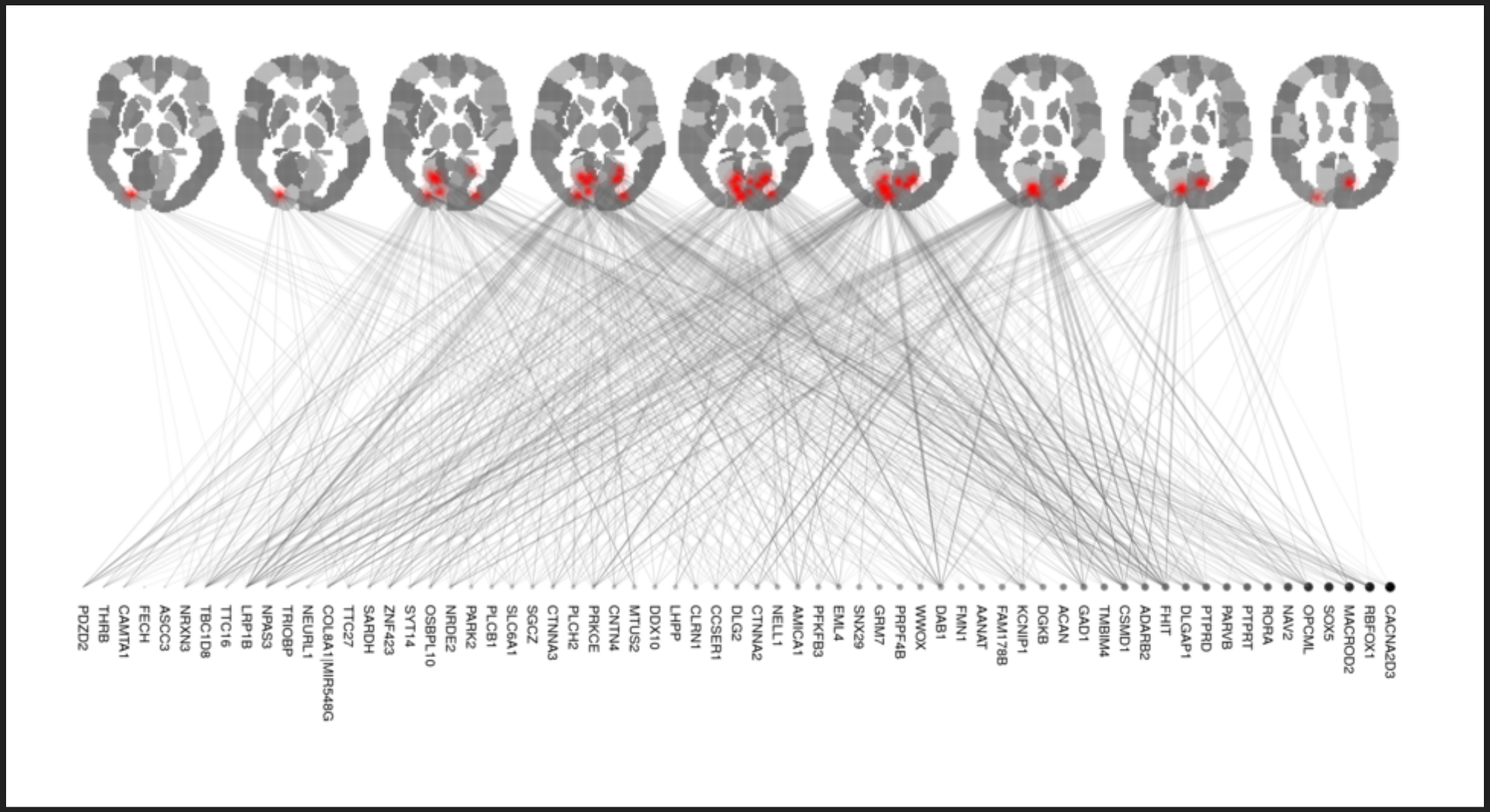
DATASET

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 a fractal n-back fMRI task, an emotion identification fMRI task, SNP arrays, and questionnaire data for over 900 adolescents.

OBJECTIVE

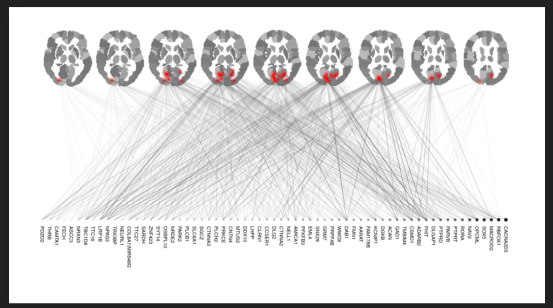
Use sparse CCA to identify the relationships between brain activity, brain connectivity, and genomics.

APPLICATION 1: N-BACK FMRI VS. SNPS



Selection from 85796 brain voxels and 60372 genomic features. [Figure 5.5 in the thesis]

RESULTS VALIDATION - N-BACK FMRI VS. SNPS



1. Similar brain regions have been found in other fMRI studies of working memory.
2. At least 34 out of the 65 identified genes have been previously associated with various aspects of human cognitive function.

APPLICATION 2: FUNCTIONAL CONNECTIVITY (FC) VS. SNPS

1. Emotion identification task fMRI data transformed to FC measures.
2. FDR-corrected sparse CCA solution includes 129 genomic features and 107 FC features.

FC VS. SNPS - TOP 10 SELECTED GENES

Gene	Previously studies in association with...
DAB1	Autism, schizophrenia, brain development
NAV2	Brain development
WWOX	Cognitive ability, brain development
CNTNAP2	Autism, brain connectivity, brain development, schizophrenia, major depression, cognitive ability (linguistic processing)
NELL1	Brain development
PTPRT	Brain development
FHIT	Cognitive ability, autism, ADHD
MACROD2	Autism
LRP1B	Cognitive function
DGKB	Brain development, bipolar disorder

(for detail see [Gossmann et. al., TMI, 2018])

FDR-CORRECTED SPARSE CCA

Some further topics:

- Further simulation studies with other underlying covariance structures.[1]
- Further simulation studies on real DNA sequence data.[1]
- Preprocessing steps for the genomic and the fMRI data.[1]
- Exploratory analysis, and analysis of confounding factors in the data.[1]

[1]: Gossmann et. al., IEEE TMI, 2018.

ANOTHER TYPE OF FALSE FINDINGS

✅ Feature selection with FDR control.

➡️ Features can be used to fit a predictive model.

Danger of over-fitting to the local noise in the given dataset, resulting in false predictions on new data.

What to do?

In Machine Learning practice, generally, usage of two independent datasets — *"training"* and *"test"* data.

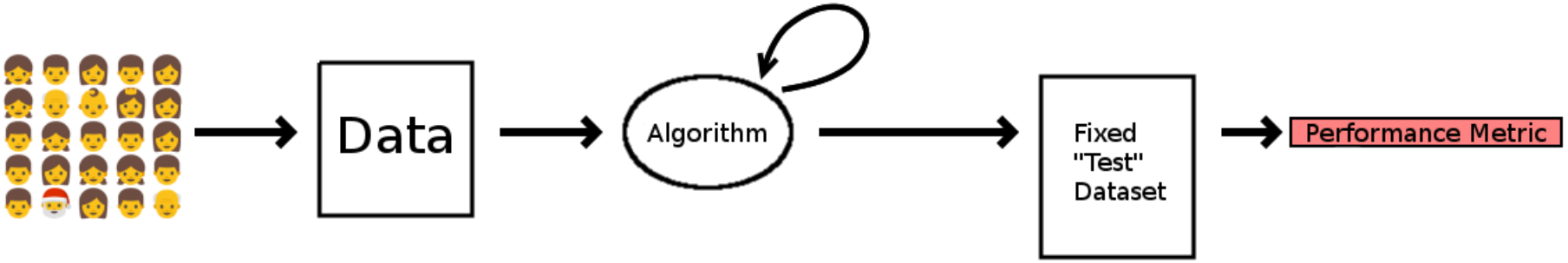
Training data: exploratory analysis, model fitting, parameter tuning, comparison of different machine learning algorithms, feature selection, etc.

$\Longrightarrow$ Adaptive machine learning, risk of overfitting.

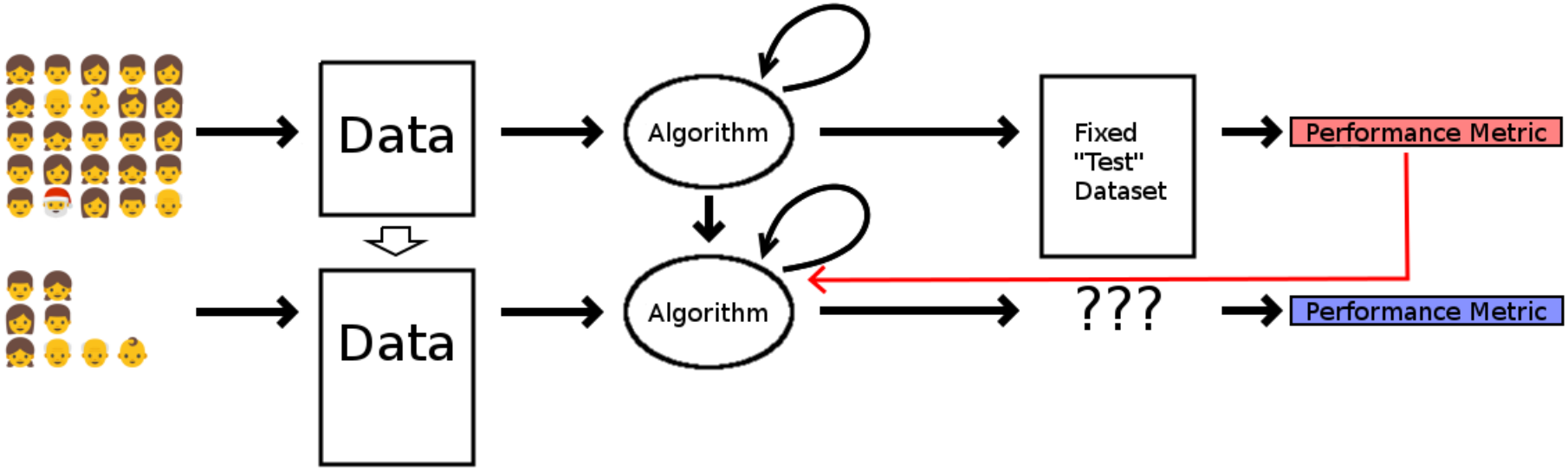
Test data: Performance evaluation *after the trained machine learning algorithm has been "frozen"*.

$\Longrightarrow$ Accurate performance measures of the final model, **if the test data is used only once**.

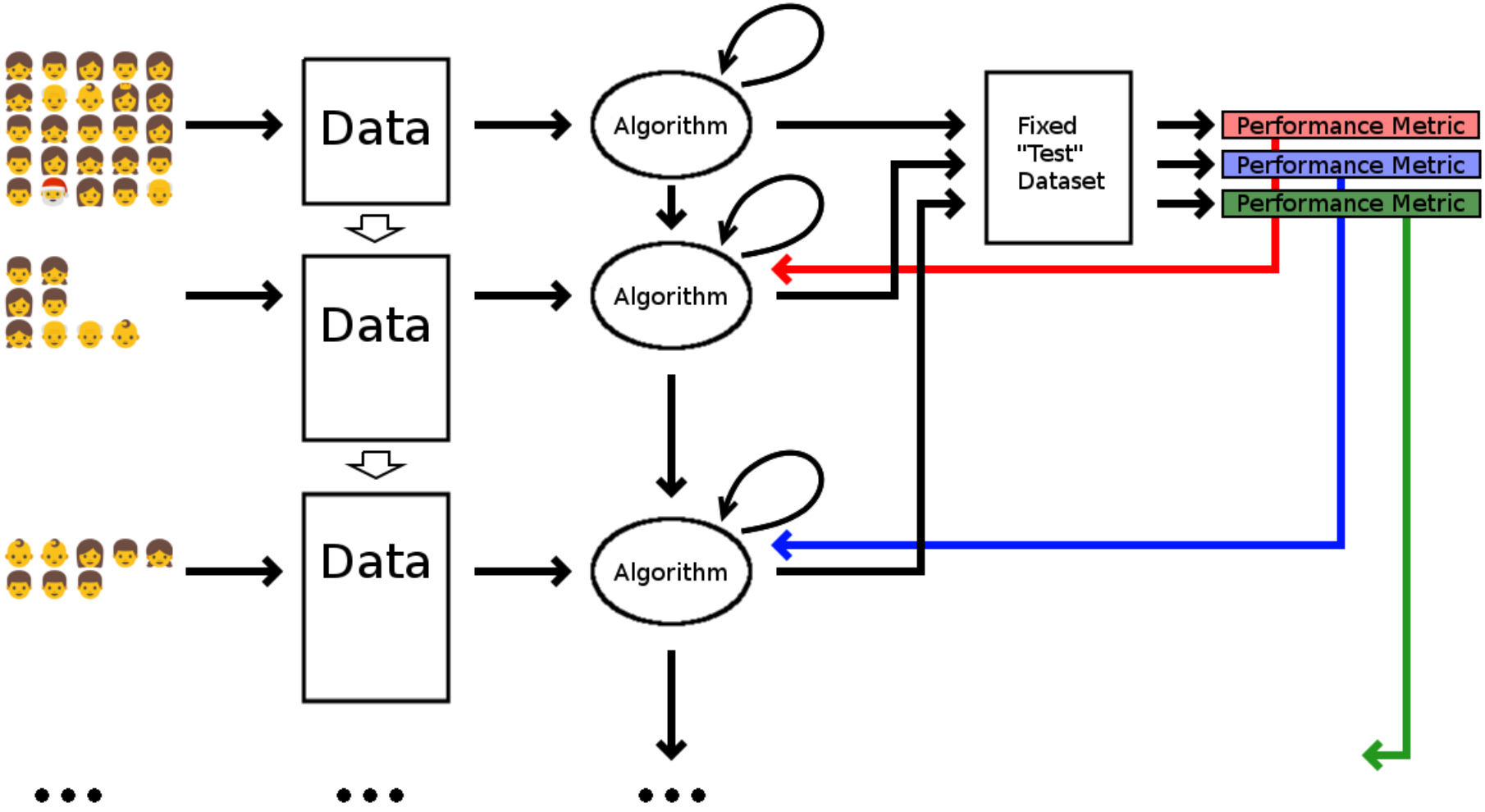
General machine learning process



"Adaptive" machine learning



"Adaptive" machine learning with test data reuse



IDEA

Can we **obfuscate** the test data to avoid overfitting?

➡️ Differential privacy.[1]

Promising simulation results.[2-3]

[1]: Dwork, McSherry, Nissim, Smith, 2006.
[2]: Dwork et. al., Science, 2015.
[3]: Gossmann et. al., SPIE 2018.

DIFFERENTIAL PRIVACY (DWORK, MCSHERRY, NISSIM, SMITH, 2006)

- A mathematically rigorous definition of data privacy.

$$P[\mathcal{M}(D) \in S] \leq e^{\varepsilon} P[\mathcal{M}(D') \in S] + \delta.$$

- **Idea:** An individual data point has little impact on the value reported by a DP mechanism.
- **Purpose:** An adversary cannot learn an individual data point from querying a DP mechanism.
- **Properties:** DP is *preserved* under *post-processing* and under *adaptive composition*.

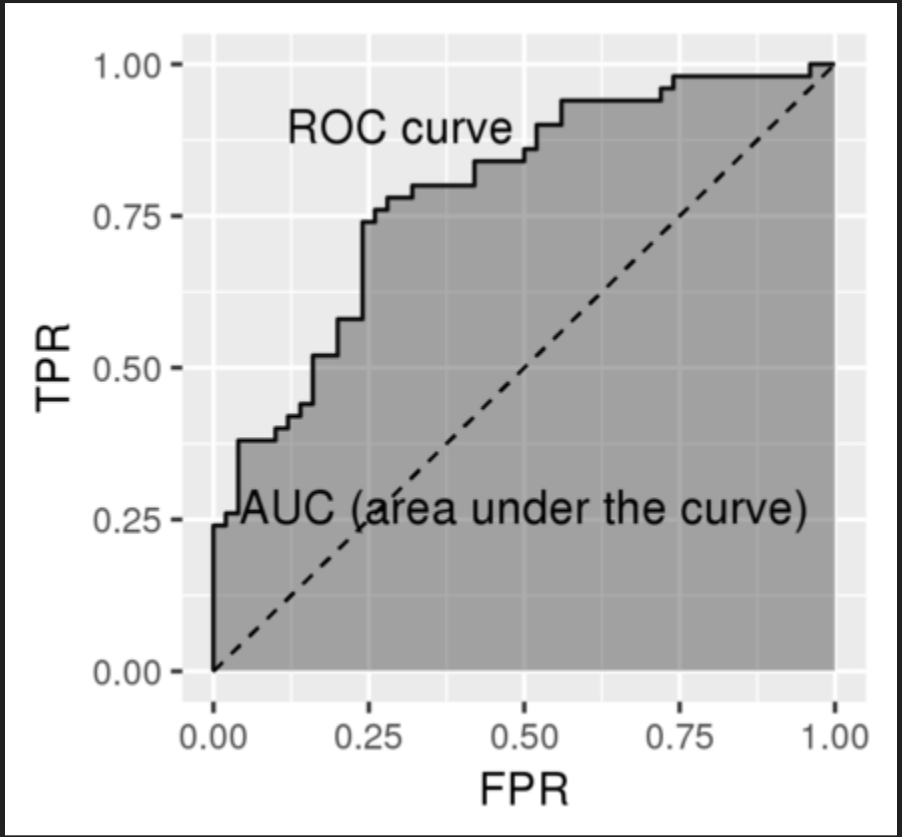
DIFFERENTIALLY PRIVATE ACCESS TO TEST DATA

Currently available literature:

- Focuses on theory.
- Theoretical assumptions are too restrictive for most of applied data analysis and machine learning.
- Computational experiments are rather simplistic.

Thresholdout + AUC = ❤️

$\text{Thresholdout}_{\text{AUC}}$[2] combines the Thresholdout technique[1] with AUC as the reported performance metric on test data.



[1]: Dwork et. al., Science, 2015.
[2]: Gossmann et. al., SPIE 2018.

THRESHOLDOUT$_{\mathrm{AUC}}$ — ROUGH SUMMARY

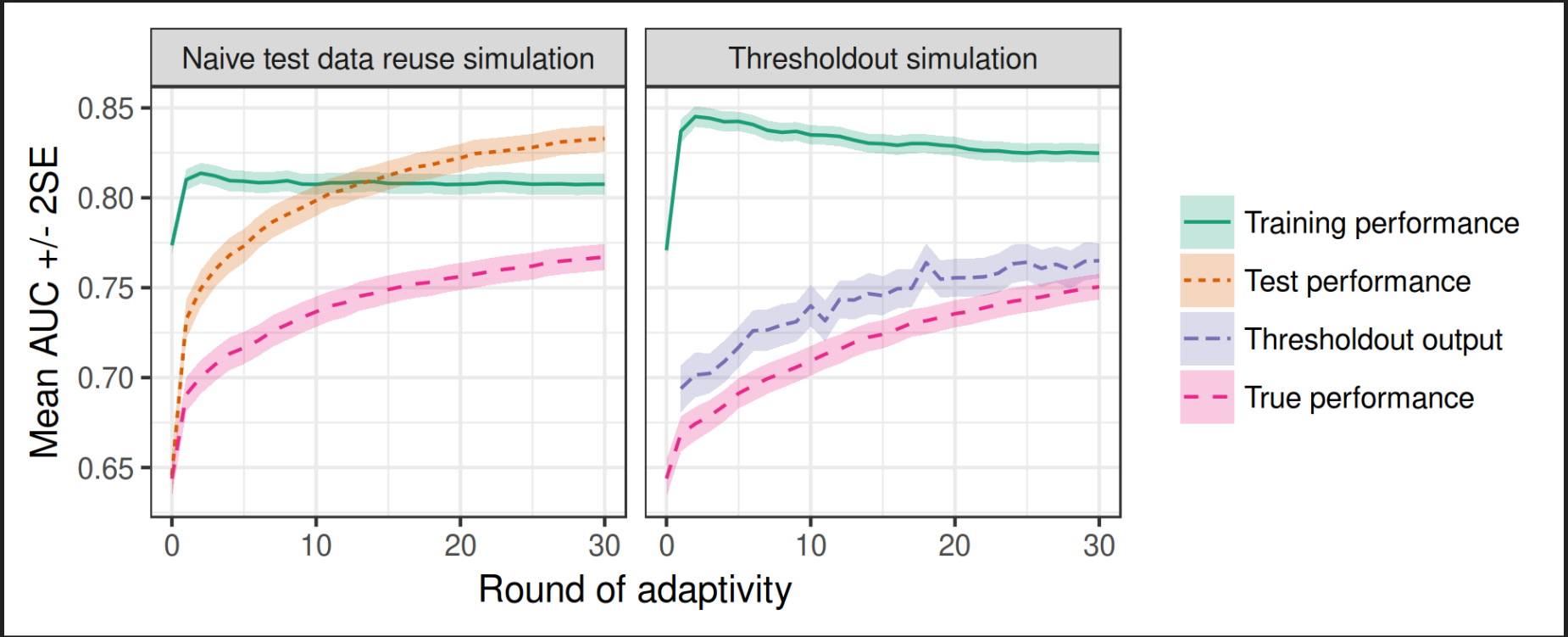
Trained classifier $\phi(x) \in [0, 1]$

If $|\mathrm{AUC}_{\mathrm{training}}(\phi) - \mathrm{AUC}_{\mathrm{test}}(\phi)| > \tilde{T}$:
output $\mathrm{AUC}_{\mathrm{test}}(\phi)+$ "a little noise"
Else:
output $\mathrm{AUC}_{\mathrm{training}}(\phi)$

THEOREM — ROUGH SUMMARY

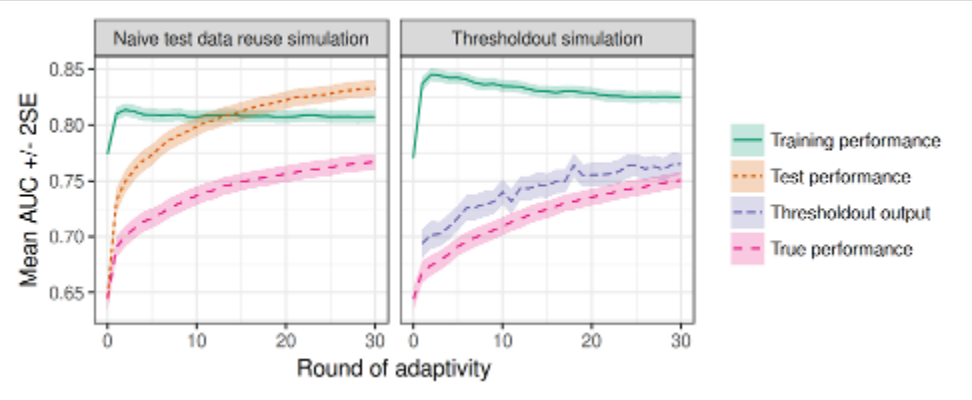
1. **If** a test dataset, which is used for performance evaluation repeatedly, is only accessed via $\text{Thresholdout}_{\text{AUC}}$. $\Longrightarrow$ **Then** with a high probability $(1 - \beta)$ the reported AUC estimates will be correct up to a small tolerance τ.
2. **Restriction**: Test data access "budget" B, which is linear in the size of the test data n, and also depends on β, τ, and the class balance.

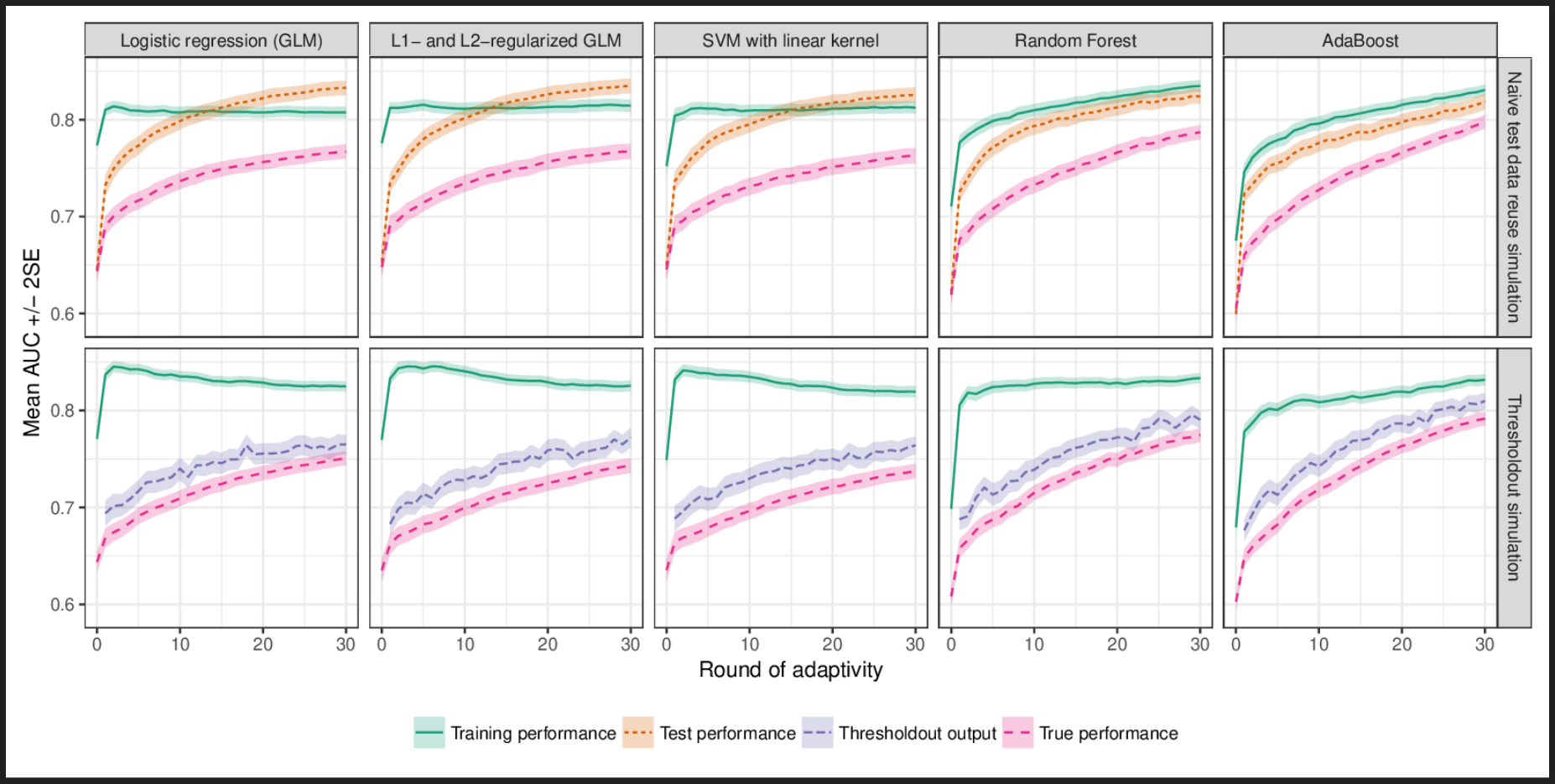
LOGISTIC REGRESSION



LOGISTIC REGRESSION

- **Naive approach**: Classifier learns the effect of local noise in the test data (overfitting).
- **Thresholdout approach**: The gap between the reported and the true AUC is much narrower!





Accuracy of reported AUC values is improved, at the cost of slightly higher uncertainty in the reported AUC, and slightly worse predictive performance.
[Figure 6.1 (a) in the thesis]

THESHOLDOUT$_{AUC}$

Some further topics covered in the thesis:

- AUC — discussion and benefits.
- Complete statement of the Thesholdout$_{AUC}$ procedure.
- Data generation.
- Simulation procedure detail.
- Choice of the tuning parameters within Thesholdout$_{AUC}$.
- Analysis of the AUC estimation error.

RESOURCES AND COLLABORATORS

- The Multiscale Bioimaging and Bioinformatics Laboratory (MBB) at Tulane University.
- Tulane Center for Bioinformatics and Genomics (CBG).
- FDA, Office of Science and Engineering, Division of Imaging, Diagnostics, and Software Reliability.
- Other: The Mind Research Network, University of Wrocław, Indiana University Bloomington, University of Tennessee Health Science Center.

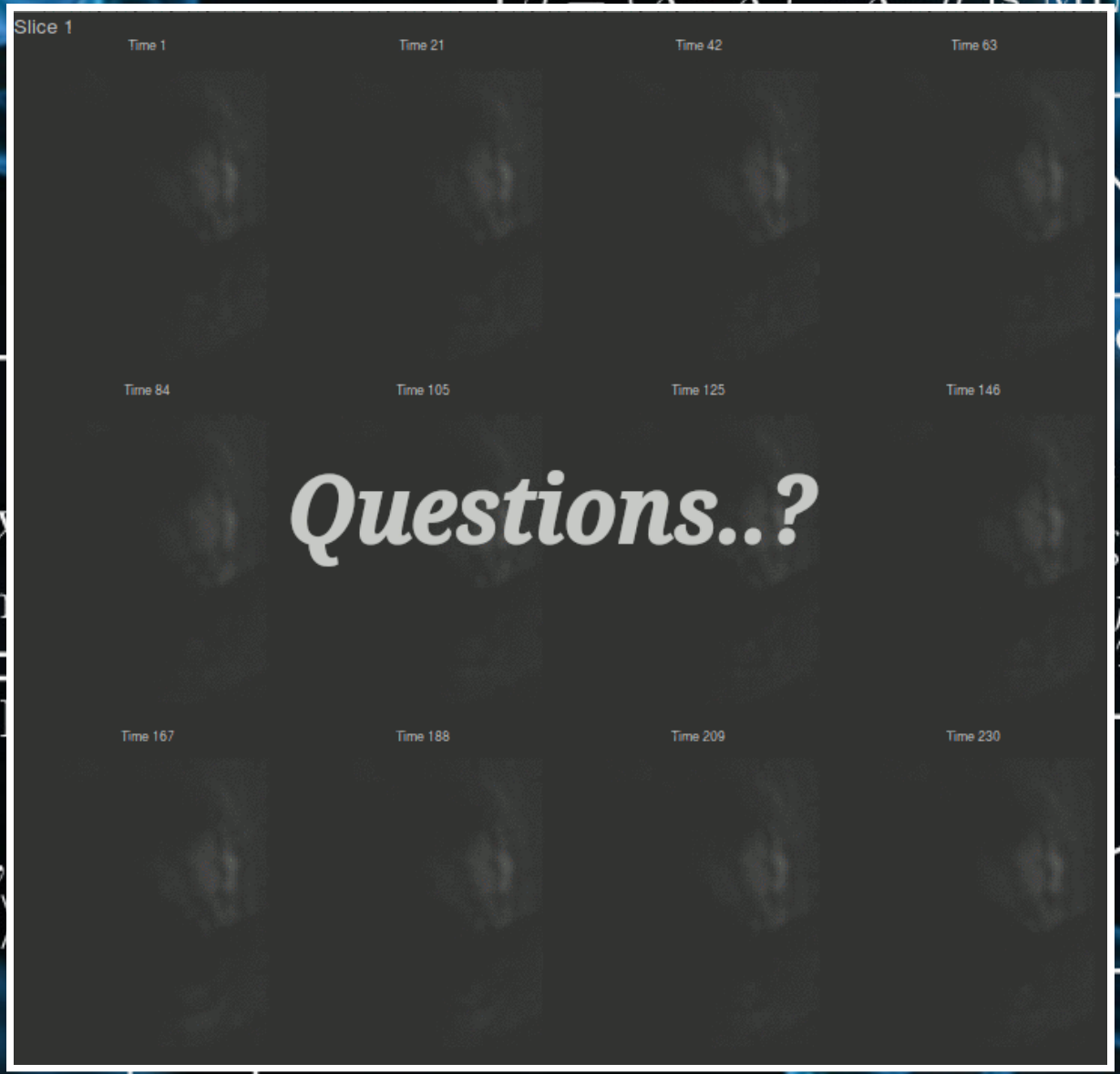
Parts of this work appear in:

1. **G.A.**, Cao, S., & Wang, Y.-P. In proceedings of ACM BCB '15. 2015.
2. **G.A.**, Cao, S., Brzyski, D., Zhao, L. J., Deng, H. W., & Wang, Y. P. IEEE/ACM TCBB. 2017.
3. Brzyski, D., **G.A.**, Su, W., & Bogdan, M. JASA. 2018.
4. **G.A.**, Zille, P., Calhoun, V., & Wang, Y.-P. IEEE TMI. 2018.
5. **G.A.**, Pezeshk, A., & Sahiner, B. In proceedings of SPIE Medical Imaging '18. 2018.

$Y \sim N(X\beta, \sigma^2 I)$.

Closed form solution:

$\hat{\beta} = (X^T X)^{-1} X^T y$ is MLE,



$n\}, U \sim f_U(u)$.

LASSO:

$\mathrm{E}(Y) = X$

$\hat{\beta} = \mathrm{argm}$

$= 0$

l linear model (G

$f_{Y_i}(y_i)$ for $i \in \{1,$

$(y_i) \exp(\langle T(y_i), \eta)$

$= f(\eta)$,

β.

Normality, $\lambda =$

egression:

$X\beta$, tuning para

$\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y -$

form solution:

$^T X + \lambda I)^{-1} X^T y$.

$\alpha = 0$

$\mathrm{E}(Y) = X\beta$, tuning parameters $\lambda \geq 0$ and $\alpha \in [0, 1]$,

$\hat{\beta} = \mathrm{argmin}_{b \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda \left[\frac{1}{2}(1-\alpha)\|b\|_2^2 + \alpha \|b\|_1 \right] \right\}$