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

Speaker notes

I will present to you my dissertation work on machine learning methodology for neuroimaging genomics.

In the interest of time, I'll skip some of the details, so just let me know if you have questions.

PRECISION MEDICINE

Inter-personal diversity in the patients' biology

⇾ "personalized" treatment plans

Speaker notes

The motivation and the ultimate purpose of this research is precision medicine.

Differences in disease susceptibility/progression between subpopulations/individual

...lead to differences in treatment efficacy

Identifying/utilizing the differences enables approaches for "personalized" treatment plans.

Including, new drugs and devices targeting specific subpopulations (or even individuals).

Avoidance of treatment based on trial-and-error.

PRECISION MEDICINE

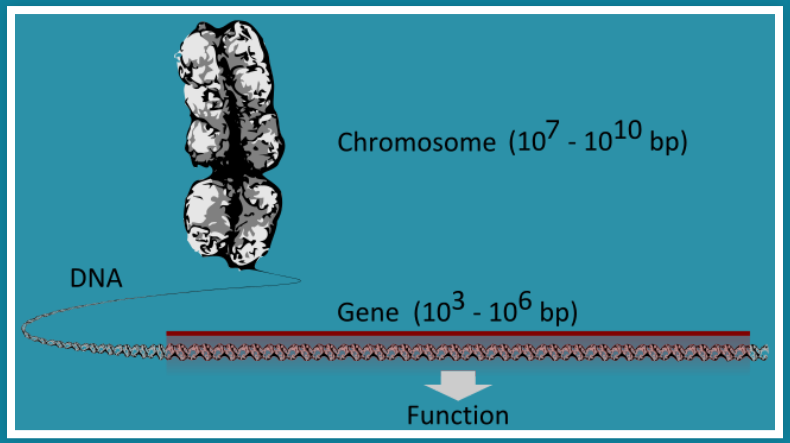
Made possible by:

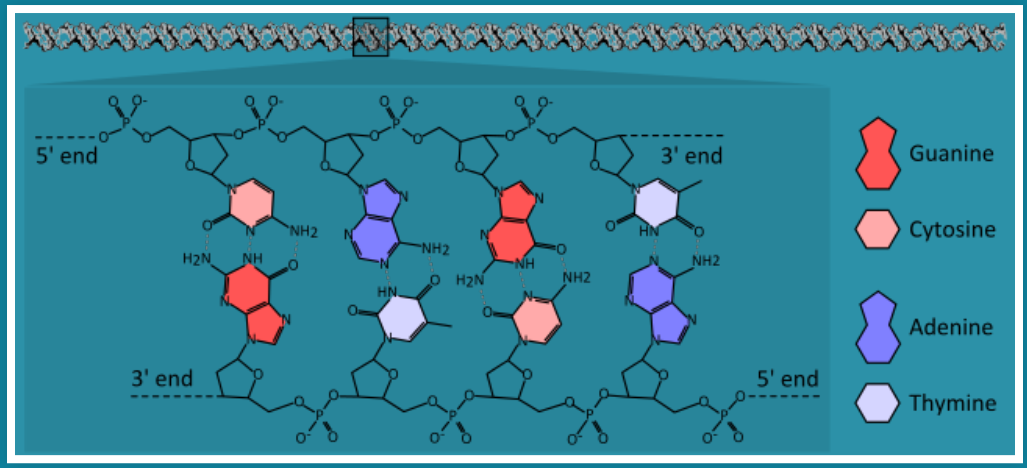
1. Big data including **genomics** and **neuroimaging**.
2. Computational methods including **machine learning** and **modern statistics**.

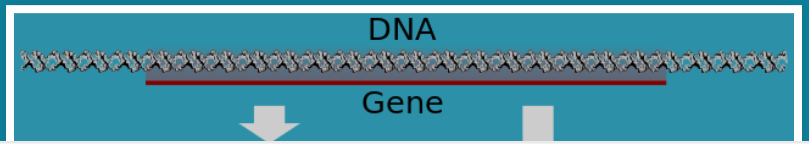
Speaker notes

Precision medicine approaches have been largely made possible with the last 10 or 20 years by...

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







Speaker notes

First let me give you some background on where genomic and neuroimagning data comes from.

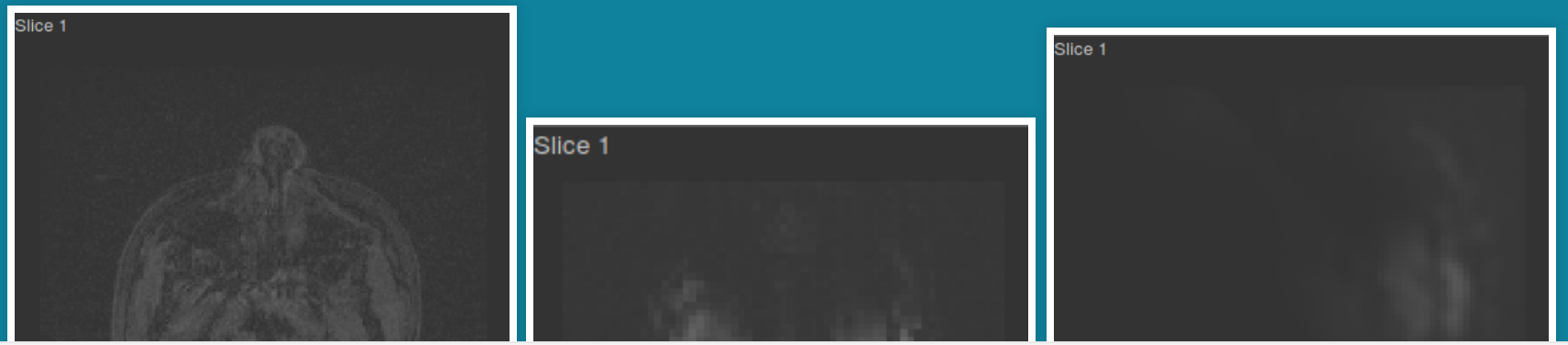
(top left) A gene is a section of DNA.

(top right) DNA is physically encoded in four molecules (nucleobases).

(bottom image) Broadly speaking, genes encode proteins or non-coding RNA, which have a biological function.

Genomic data include information on the structure of the DNA, and amount of protein produced, which can all be expressed/stored in numeric form.

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



Speaker notes

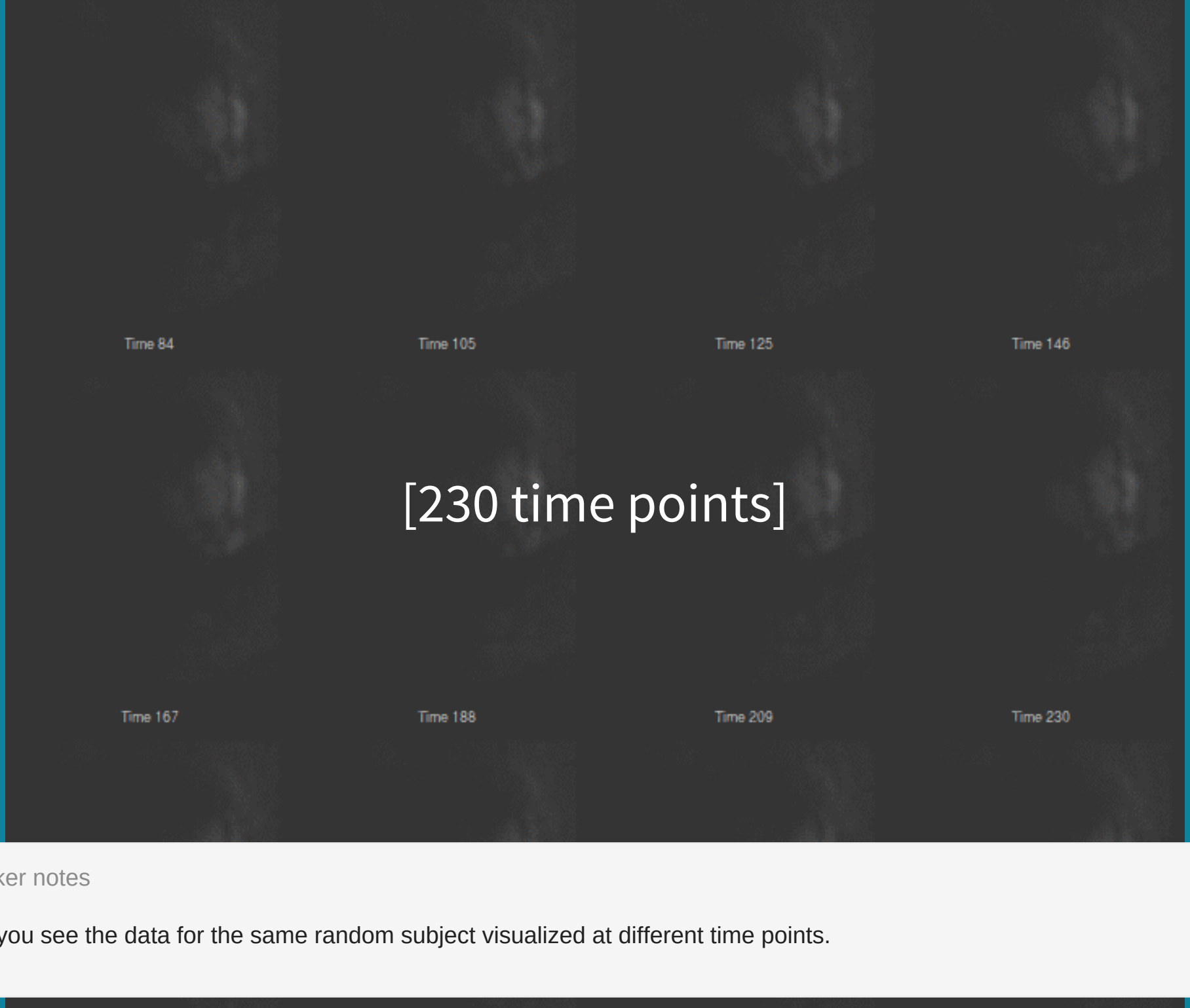
These are some MRIs of the human brain.

They either supply us with anatomical structure, or brain activity over time.

MRI 3D volumes are collected as 2D axial slices, as I have visualized in the shown animations for a randomly chosen PNC subject (a 9 years old female).

I mostly use blood oxygen level dependent, or BOLD, fMRI data which gives you brain activity measures in brain voxels, which are 3d pixels, at discrete points in time.

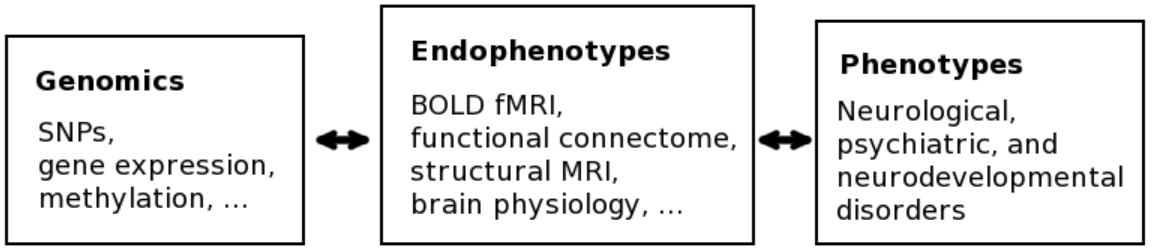
You can see a raw structural or T1-weighted MRI on the left, — a raw BOLD fMRI in the middle, — and the same 3D fMRI image after standard preprocessing steps have been performed (MRI preprocessing is a huge research area in its own right).



Speaker notes

Here you see the data for the same random subject visualized at different time points.

PRECISION MED. & MENTAL DISORDERS



Speaker notes

To bring this back to the topic of precision medicine:

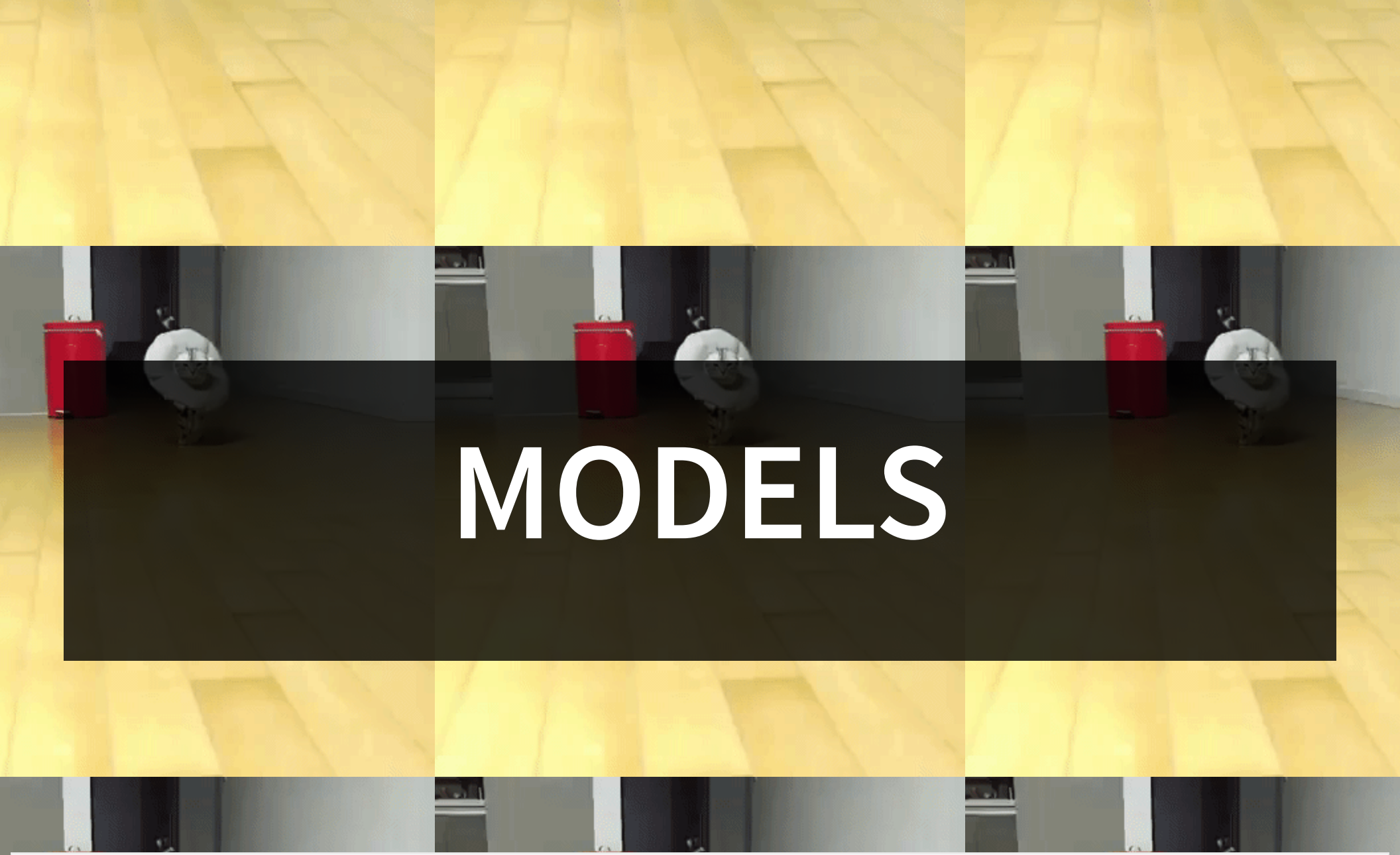
Mental disorders are known to be heritable, but it has proven to be extremely difficult to identify the causal genes of these complex diseases.

It became clear in the last few years that measures of brain function and structure can be used, as an intermediate phenotype between genomics and clinical traits or behavioral measures.

This has led to more powerful methods in the identification of genomic associations.

Helpful in the interpretation of the function of the identified genes, because gene expression directly affects cellular metabolism, which directly influences the neural circuity.

Recent papers show that joint analysis genomics and fMRI can be used for diagnosis, and, more importantly, for monitoring and guidance of drug treatment (use fMRI to see whether drug has the intended effect).



MODELS

Speaker notes

Before I present my methods and results, let me give you some brief background on statistical 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 ($m \ll p$).

Speaker notes

Here you see a general definition of a statistical model; more precisely, you may call this type of model a regression model.

It is essentially a randomized function that maps a set of predictor variables x1, x2, ..., xp to an outcome variables y, whereby the randomness is captured by the additive noise term epsilon.

However, consider for example your favorite phenotype (like certain disease or trait) and the human DNA.

The human DNA consists of 3 billion base pairs, and one cannot expect that changes at all bases will have an effect on this specific phenotype.

In fact only the changes/mutations within a relatively small number of genes will have an effect.

This brings us to the topic of sparse models, where the effect size of most predictor variables is exactly zero.

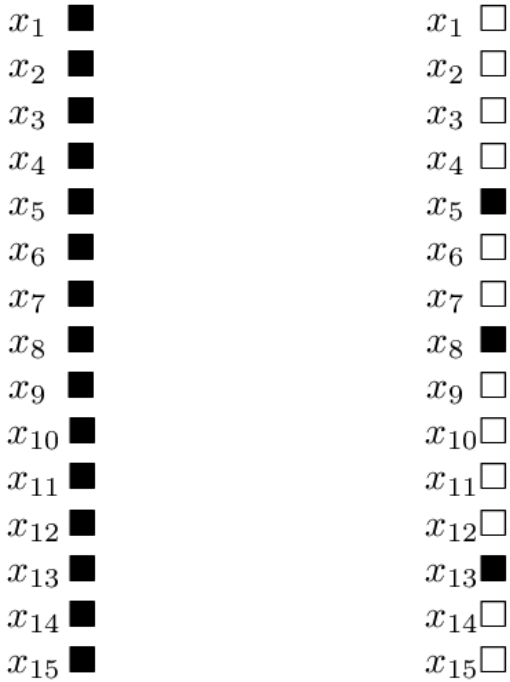
"Sparsity" refers to the fact that, the entire set of predictor variables is "sparse" with respect to non-zero effects.

The number of predictor variables which are actully significant to the outcome variable, is a relatively small subset of the entire set of variables.

Although sparsity can be assumed in many biological applications based on prior biological knowledge, a priori we know neither which variables fall within the significant subset, nor how small this subset is.

SPARSE MODELS

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



Speaker notes

Here is an illustration of what it means to have a sparse regression model, where the distribution of the phenotype or outcome variable y conditional on the predictor variables x5, x8, and x13 is the same as the distribution conditional on all potential predictor variables x1, x2, and so on.

THE TWO-FACED MODEL SELECTION PROBLEM

Speaker notes

When choosing the best model for a particular dataset, then, this involves solving two related but fundamentally different problems at the same time.

Obtain the best possible predictions of the outcome variables y based on the given predictor variables x1, x2, and so on;

Find the subset of predictor variables that are associated with the outcome variable y, and dispose of the irrelevant variables.

Both problems are essential in biomedical applications. The first for example for disease diagnosis or risk prediction, and the second to help understand the cause of disease, e.g., for the identification of targets for treatment.

Depending on which of the two modeling problems you consider, you are dealing with a conceptionally different type of false findings.

In prediction you may have FP (or FN), e.g., a misdiagnosis; while in feature selection we have false discoveries, e.g., a gene identified as associated to disease when it is not.

Unfortunately, many ML methods, and also many statistical models, which are commonly used in genomics and neuroimaging, provide no way to obtain an idea of how many false findings to expect.

AIMS

Establish **guarantees** on...

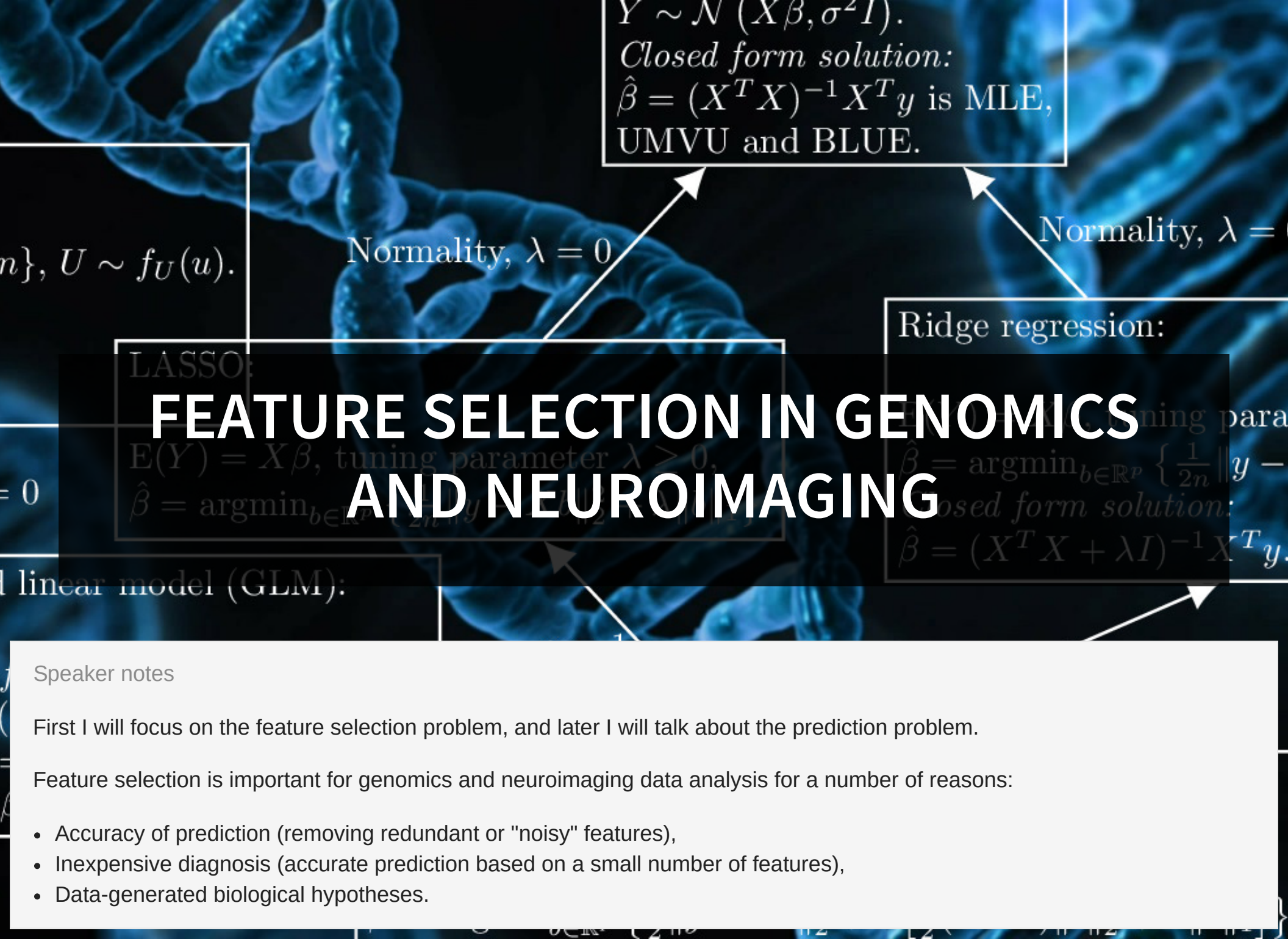
- false discoveries in feature selection,
- false predictions on new data (generalization)

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

Speaker notes

In my thesis, I proposed new methods that build on ones widely used in genomics and neuroimaging.

However, my methods focus on explicitly establishing guarantees...



FEATURE SELECTION IN GENOMICS AND NEUROIMAGING

Speaker notes

First I will focus on the feature selection problem, and later I will talk about the prediction problem.

Feature selection is important for genomics and neuroimaging data analysis for a number of reasons:

- Accuracy of prediction (removing redundant or "noisy" features),
- Inexpensive diagnosis (accurate prediction based on a small number of features),
- Data-generated biological hypotheses.

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).

Speaker notes

The classical way to avoid false discoveries is found in the theory on multiple hypotheses testing.

It involves performing a separate HT for each variable/feature.

...and then correcting the tests to account for multiple comparisons.

There are two important error rates that one aims to control - FWER and FDR.

FWER which is P(...), i.e., which demands no FD at all; and FDR which is defined as E(...), i.e., which states that it is OK to have FD as long as ...

Existing procedures control the FDR or FWER. E.g, widely-used in GWAS.

Nowadays, most people in genomics and neuroimaging, including myself focus on FDR - more powerful in a very high-dimensional setting.

But: Doing a hypothesis test on every possible genetic variant separately is like trying to understand a sentence by analysing each letter separately, one at a time (ignoring the sequence in which they are written).

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).

Speaker notes

All features are analysed jointly (possible to include interaction effects, non-linear effects).

It essentially amounts to the estimation of a coefficient vector beta.

By looking at which elements of beta are 0 and which are non-zero, we see which variables are associated to the outcome.

A subset of significant features is selected from the whole set of features **in one step**, rather than "one letter at a time".

Problem: no reliable way to tune the sparsity of the solution...

...implies that: **Cannot control for the error rates mentioned on the last slide.**

(Feature selection and prediction are performed simultaneously.)

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.

Speaker notes

To my knowledge, the first sparse regression method that can control the FDR is SLOPE.

It provides an explicit way of setting lambda so that the FDR will be under some q.

BUT this requires...

Usually this isn't the case with biological data.

GROUP SLOPE MOTIVATION

- Divide the data into groups by correlation. ← *Often possible for biological data.*
- Then select/drop entire groups rather than individual variables

Speaker notes

This motivated us to develop the Group SLOPE method, in which we divide the data intro groups according to the correlation between variables.

This is often possible for biological data, and in particular DNA sequence data.

Then we select or drop entire groups of variables rather than individual variables, ...

...and we redefine the FDR with respect to groups.

GROUP-WISE FALSE DISCOVERY RATE

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

We define

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

Speaker notes

We define the group-wise FDR as the expected value of the number of falsely discovered groups over the total number of discovered groups.

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 $\boldsymbol{\beta}_i \in \mathbb{R}^{p_i}$.

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

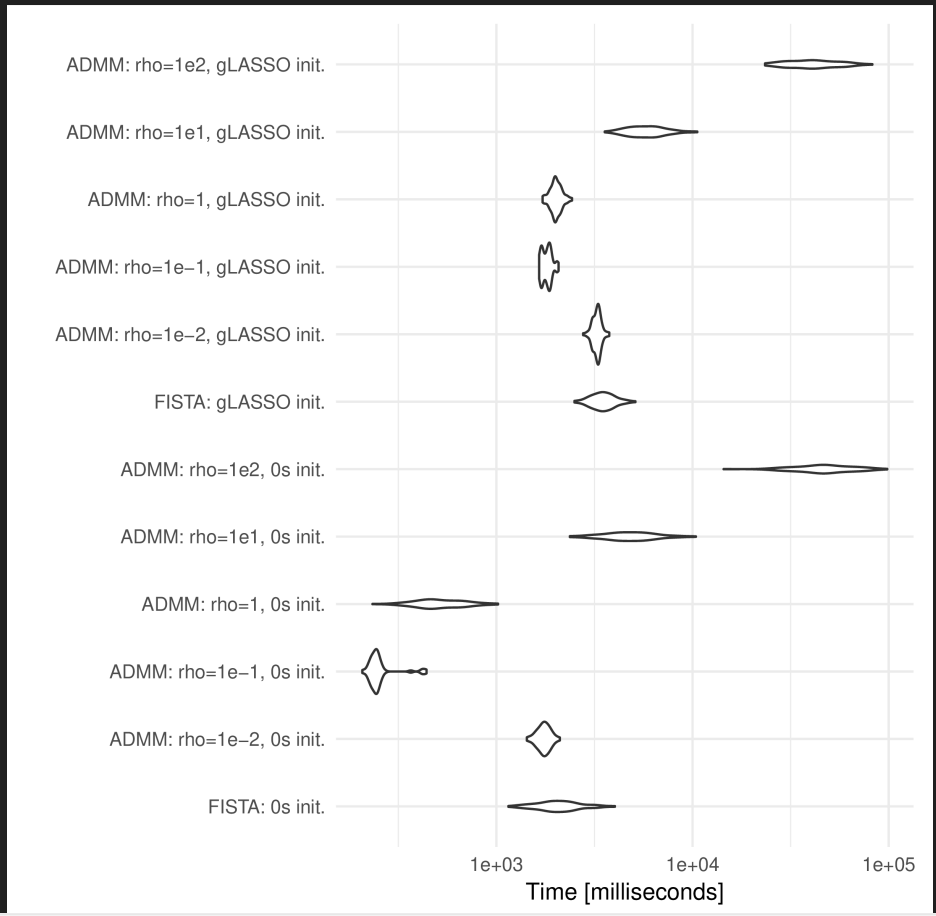
Speaker notes

The Groups SLOPE method is formulated in this way, where the coefficient vector beta is divided into J groups of length p1,p2,...and so on variables.

The parametric formulation of the model is very similar to SLOPE, but with the penalty term operating at a group level rather than on individual variables.

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.



Speaker notes

The method would be useless without a way to obtain the solution efficiently.

In the thesis I derive two types of optimization routines for this model based on...

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].

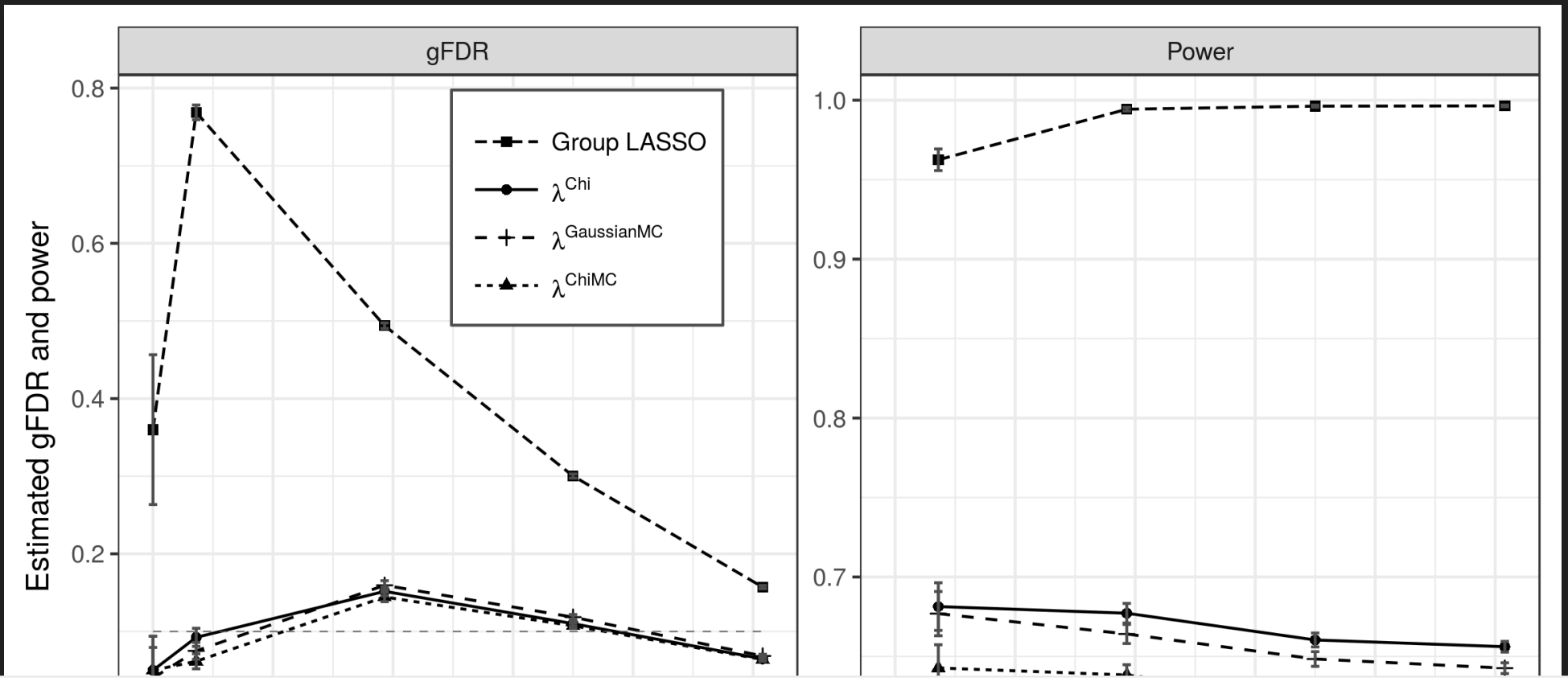
Speaker notes

The Group SLOPE method also provides theoretical guarantees on FDR control, which was our main goal.

Given a user-specified target FDR threshold q, we give explicit procedures for selecting lambda so that the FDR will be below q.

We have three types of approaches to set lambda - a fully theoretically derived lambda sequence for the special case when the groups are orthogonal to each other - a heuristic based on rigorous theory leading to FDR control when groups of variables are nearly uncorrelated - and Monte Carlo based approaches for more general designs.

We have confirmed with extensive simulations studies that all of these are successful in controlling the gFDR.



Speaker notes

We compared the gFDR of the group SLOPE method on real DNA sequence data to the gFDR of the most commonly used group-wise feature selection method called group LASSO.

Here, the x-axis shows the number of truly relevant groups, and the y-axis the gFDR of each method. Regardless of the number of truly relevant groups group SLOPE always keeps the FDR near the target level of 10% while group LASSO clearly does not.

As you see on the right, even while keeping the gFDR low, group SLOPE also keeps the true discovery proportion relatively high.

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.

Speaker notes

We applied group SLOPE to the Framingham Heart Study data, which contains single nucleotide polymorphism (or SNP) data for almost 9000 subjects,...

A longitudinal study that has been collecting data since 1948 and is still ongoing.

A SNP is a variation or mutation of a single nucleotide that occurs at a specific position on the DNA strand.

We aimed to select SNPs associated with differences in BMD.

...and we apply group SLOPE to this data matrix X for subjects with BMD measurements.

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),

Speaker notes

15 genes reported as associated to BMD or other aspects of bone health in previous studies.

This validates our biological results, and implies that the other selected genomic features represent new candidate genes for association with BMD, which should be followed up with further studies.

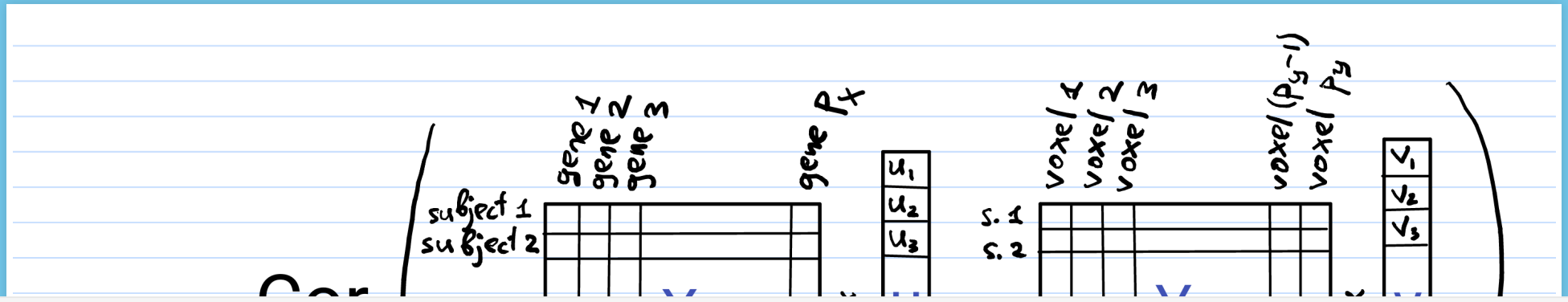
Group SLOPE — Some further topics:

- 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]

Speaker notes

We have many other results on the group SLOPE method, included in these three papers that do not overlap aside from the basic definition of the method, that are in my thesis but I don't have time to go into today, including its performance on DNA sequence data and mathematical results related to gFDR control, error variance estimation, and estimation properties.

CANONICAL CORRELATION ANALYSIS



Speaker notes

In the previous model we had an outcome variables y and many predictor variables represented by a high-dimensional vector x.

What if we y is high-dimensional too?

For example y may represent a subject's fMRI activations per brain voxel, while x represents expression values per gene, and we would like to know which genes are related to which brain voxels.

A popular technique to address this is CCA.

CCA finds a linear combination of the fMRI features and a linear combination of the genetic features as two new variables with as large a correlation as possible. This amounts to estimating two coefficient vectors u and v, which are referred to as canonical vectors. Xu and Yv are called canonical variates.

In practice, we apply further constraints on u and v, including sparsity.

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

Speaker notes

This is the mathematical definition of the classical CCA method, where we aim to maximize a covariance term and put constraints on u and on v.

This formulation is degenerate when n<p, and in addition it does not account for the sparsity assumption, which is often true for 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

Speaker notes

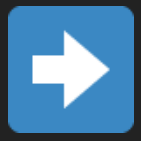
Sparse CCA methods have been proposed to provide a way to account for the sparsity assumption, as well as to ensure a unique solution even when the sample size is smaller than the number of variables.

So that CCA can be used for the analysis of genomic data.

The model is similar to the classical CCA, with an L1 penalty put onto u and v.

BUT it's unclear how to set the sparsity of the solution, which is controlled by the penalty parameters c1 and c2.

SPARSE CCA



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

Speaker notes

...

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.

Speaker notes

In order to define the FDR, which is the expected proportion of discoveries that are false, for sparse CCA, we consider the FDR for u and v spearately since they come from two different types of data, and we need to consider what it means to have a false discovery in the context of CCA.

DEFINING FDR FOR SPARSE CCA

The coefficient estimate $\hat{u}_i \neq 0$ *represents a* ***false discovery*** *of the* i*th 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*

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

Speaker notes

Since CCA aims to maximize the covariance between Xu and Yv, we consider u_i to be a false discovery if it's nonzero and does not affect the valu eof the covariance. You can show that this is equivalent to this expected value being equal to zero.

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

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

Speaker notes

Then we define the FDR similarly to before, as the expected value of the number of false discoveries over the number of non-zero elements in u.

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

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

Speaker notes

We can then analogously define the groupwise FDR.

SLOPECCA AND GSLOPECCA

Speaker notes

Having the FDR and gFDR definition in place, we propose three different sparse CCA methods that keep the FDR or gFDR below a user-specified level.

The first two work on data with block diagonal covariance structures. The third put no assumptions on the covariance structures and is intended for sparser settings, but has a lower detection power.

slopeCCA:

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

gslopeCCA:

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

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

Speaker notes

...

SLOPECCA AND GSLOPECCA - COMPUTATIONAL METHODS

Both are **biconvex optimization problems**.

Alternating optimization algorithms are guaranteed to converge.

Speaker notes

Again, for the methods to be useful in practice, we need to know how to obtain the solution efficiently.

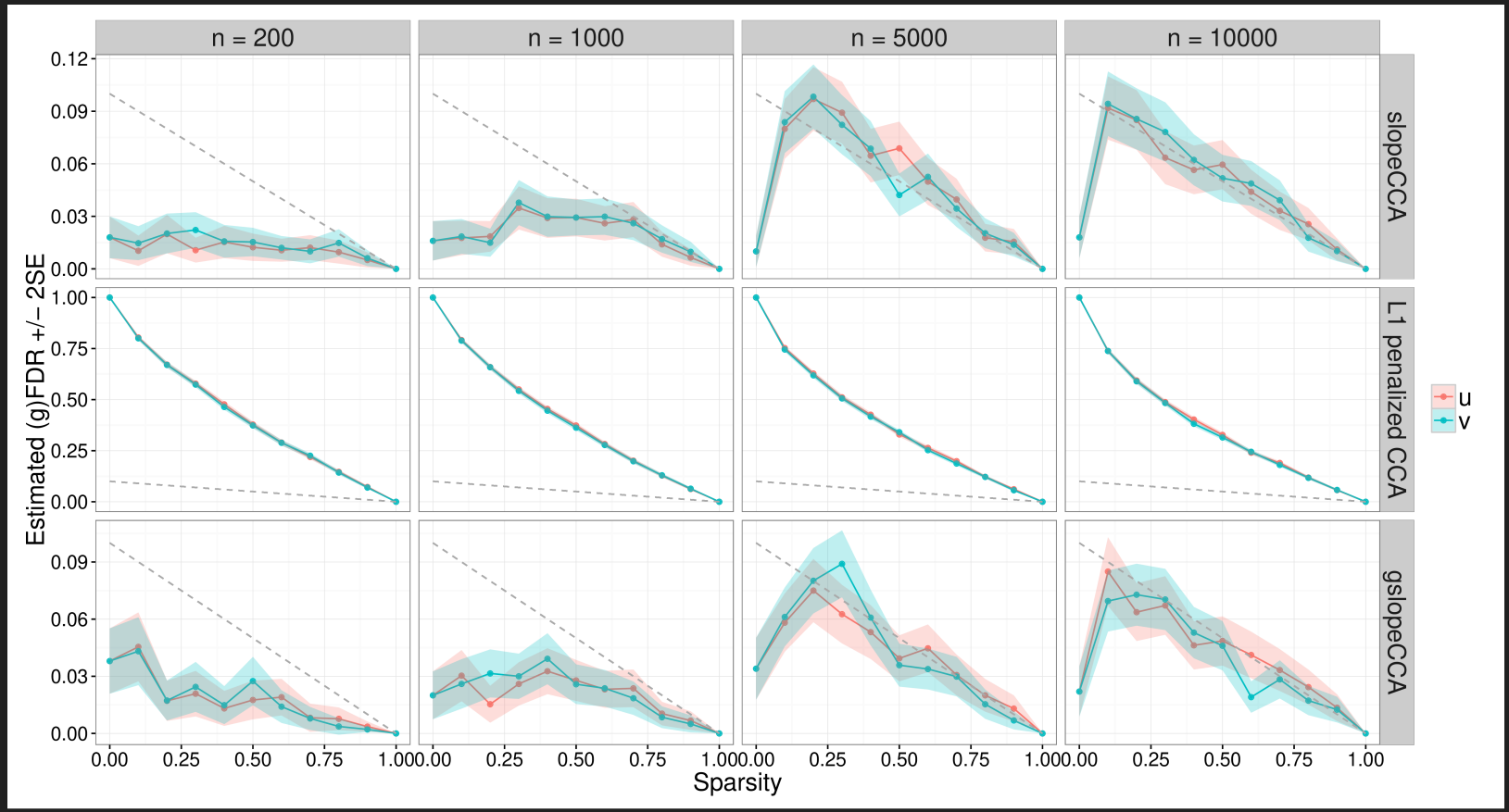
SLOPECCA AND GSLOPECCA - THEORETICAL GUARANTEES

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)

Speaker notes

We provide theoretical guarantees for the FDR as the sample size gets large, as long as Cov(x) and Cov(y) are block-diagonal.



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

Speaker notes

...

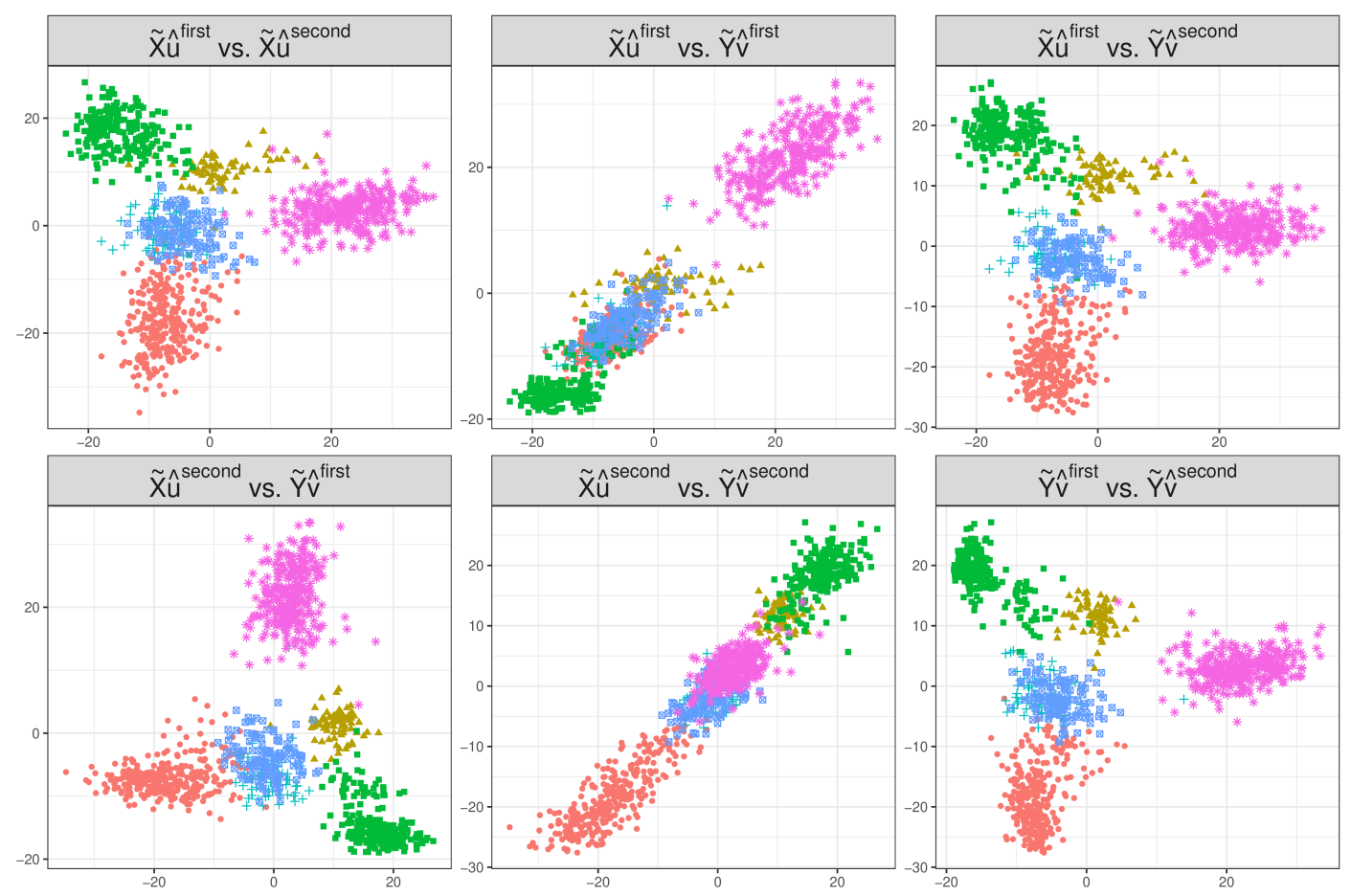
APPLICATION TO TCGA

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.

Speaker notes

...



Speaker notes

We can use gslopeCCA to look at the canonical variates, reducing the number of variables we need to consider from over 40,000 to just 4.

Plotting these 4 against each other, and coloring the points based on the cancer type of the corresponding observation, we see that these four derived variables in fact separate the difference cancer types very well.

This implies that this method can be used in diagnosis, and to reveal new information about disease phenotypes.

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.

Speaker notes

...

SLOPECCA AND GSLOPECCA

Some further topics covered in the thesis:

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

Speaker notes

...

Time 84 Time 105 Time 125 Time 146

FDR-CORRECTED SPARSE CCA

Time 167 Time 188 Time 209 Time 230

Speaker notes

...

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.

Speaker notes

...

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.

Speaker notes

...

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.

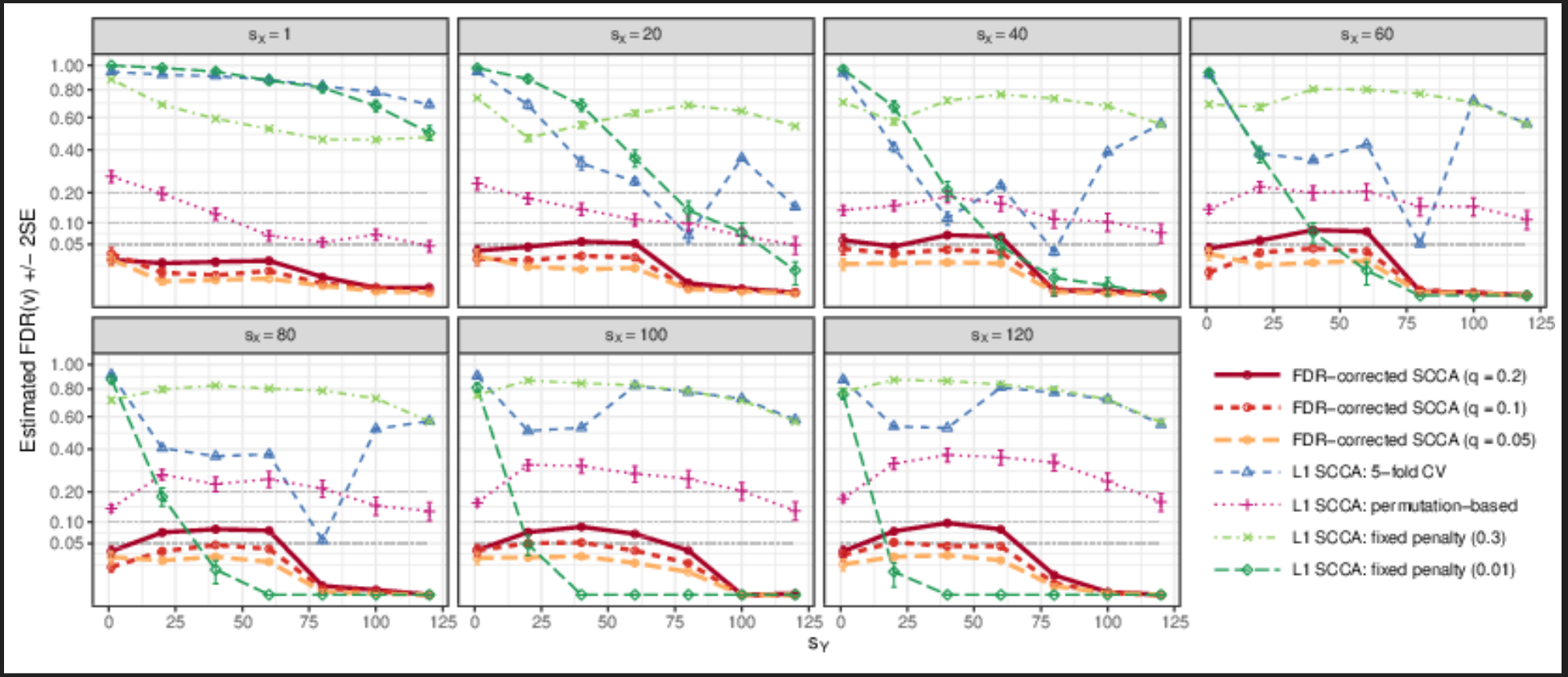
➡ FDR control confirmed with extensive simulation studies on synthetic and real data.[1]

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

Speaker notes

...

SIMULATION STUDIES

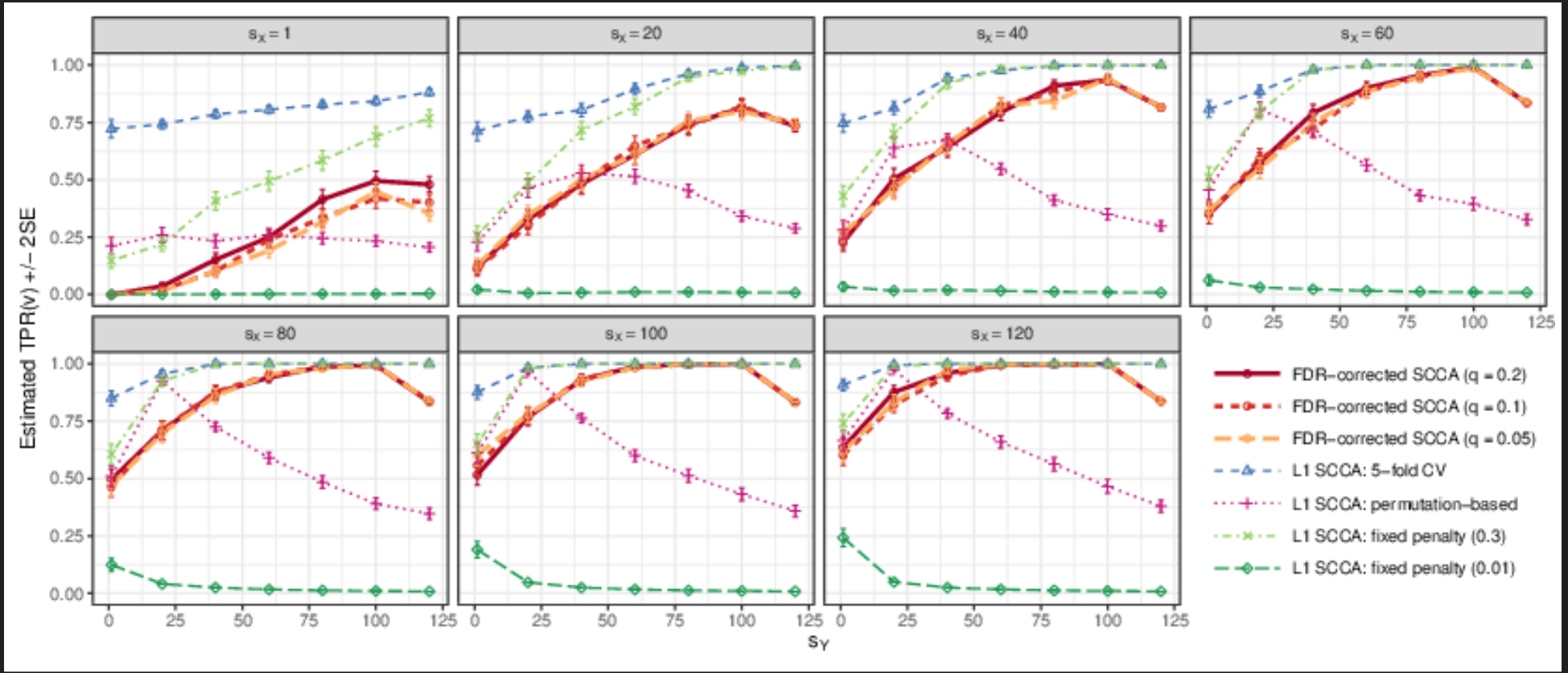


$X, Y \in \mathbb{R}^{600 \times 1500}$. FDR is controlled regardless of the

Speaker notes

...

SIMULATION STUDIES



Speaker notes

...

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.

Speaker notes

One application of FDR corrected SCCA is studying the diversity in brain activity and connectivity in children and adolescents in relationship to their genomic information to answer questions like "What are the driver genes behind certain 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,

Speaker notes

Including subjects with increased (prodromal) symptoms of ADD (107 PNC subjects), schizophrenia (103 PNC subjects), and depression (85 PNC subjects) [Kaufmann et. al., 2017].

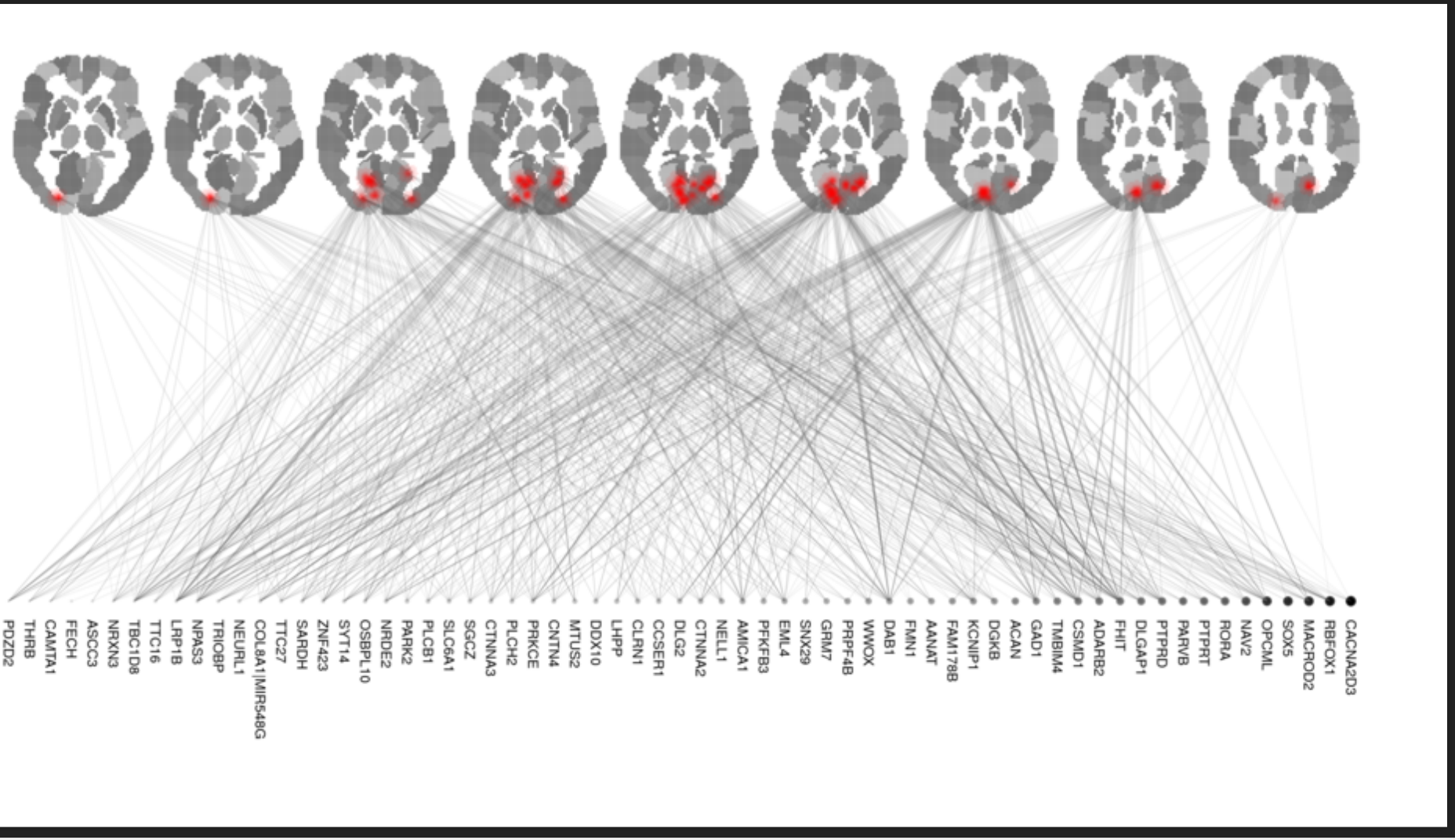
OBJECTIVE

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

Speaker notes

...

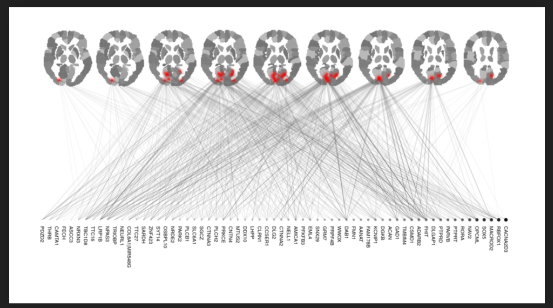
APPLICATION 1: N-BACK FMRI VS. SNPS



Speaker notes

...

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

Speaker notes

More than half of the genes we identified have been previously associated with human cognitive function, supporting the biological validity of our results, and indicating that the remaining identified genes may provide valuable biological hypotheses that should be followed up with additional studies.

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.

Speaker notes

...

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

Speaker notes

...

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.

Speaker notes

...

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

Speaker notes

Now that we know how to identify truly relevant predictive features, we can come back to the problem of prediction.

Whenever one does prediction based on data, there is the danger of obtaining a model that is extremely accurate on the data that you are using, but which does not generalize to the population from which the data were drawn.

This is called over-fitting in statistics and machine learning.

A model that is overfit to a specific dataset will yield many FP on new data when the reported performance of the model is much better.

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.

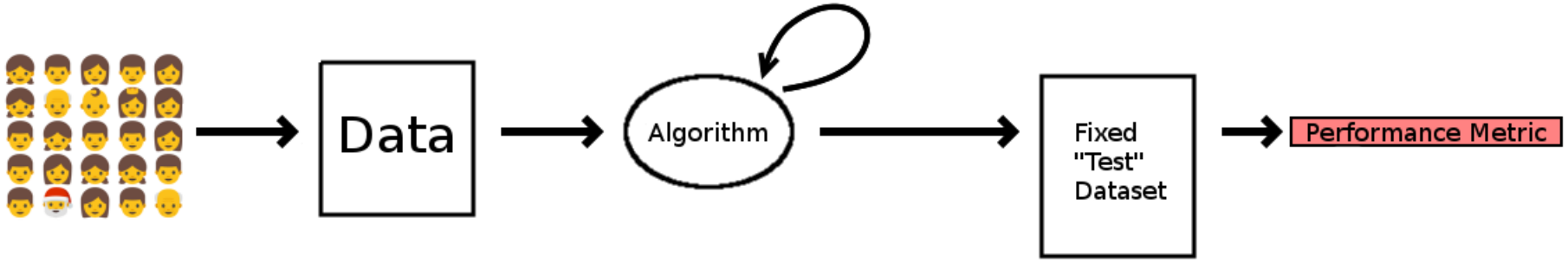
Test data: Performance evaluation *after the trained machine learning algorithm has been "frozen"*.

Speaker notes

The performance metric obtained on test data will make it evident whether any overfitting has occurred during model training.

Other techniques (multiple testing, cross-validation, bootstrap) can be used to avoid or reduce the overfitting on the training data.

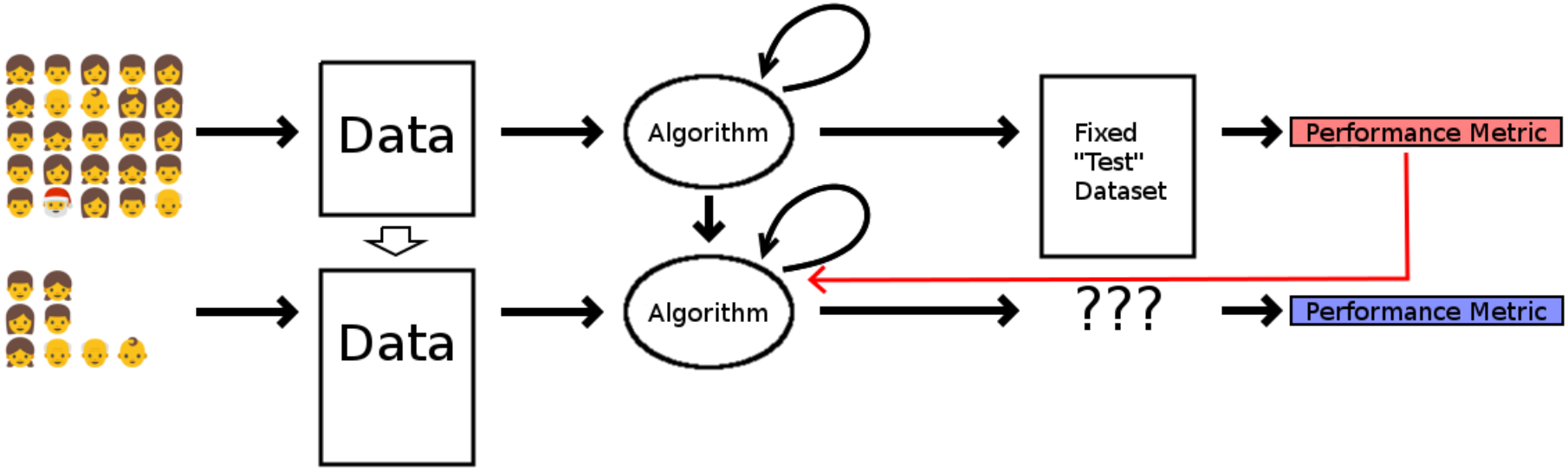
General machine learning process



Speaker notes

...

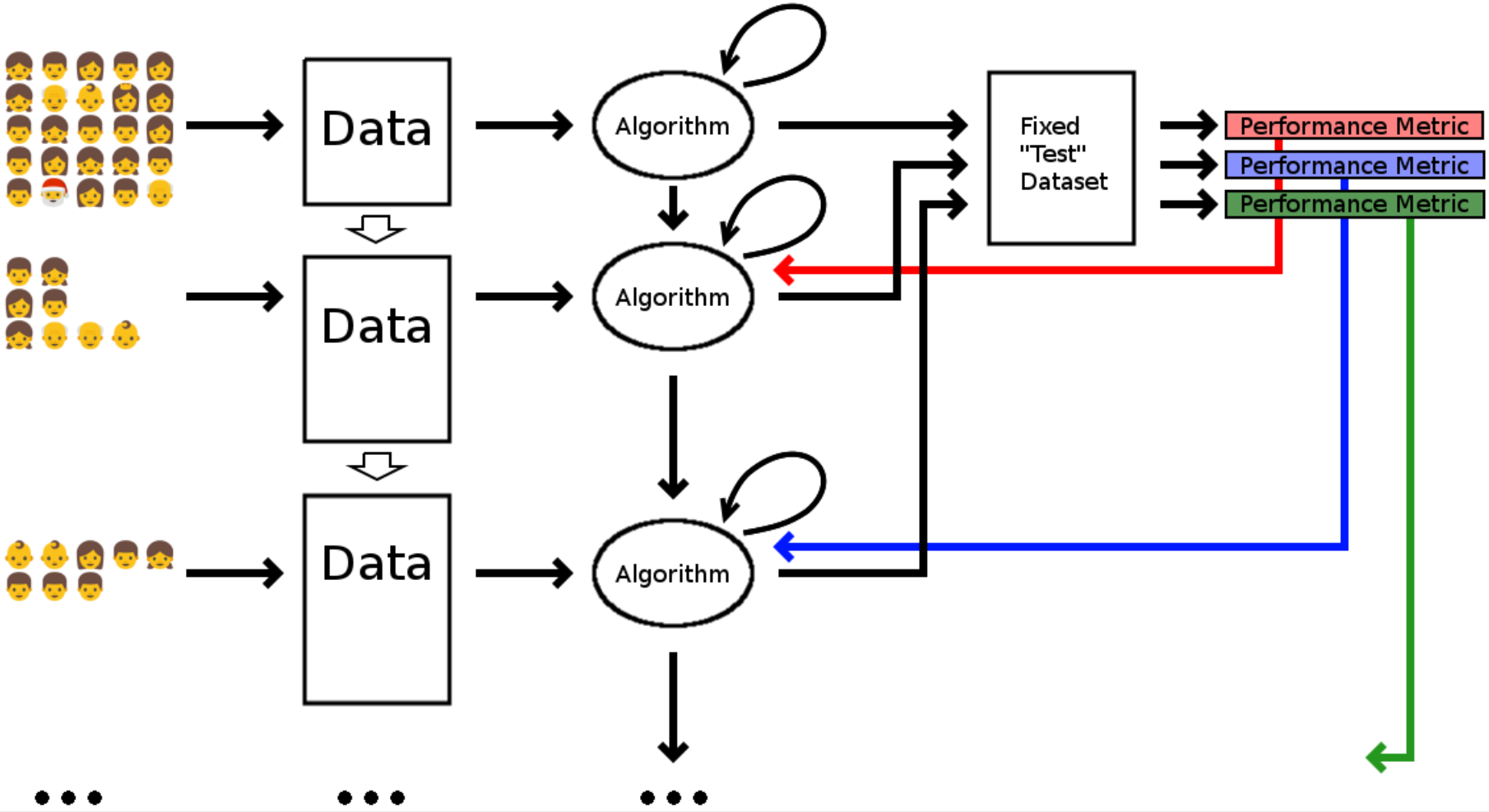
"Adaptive" machine learning



Speaker notes

...

"Adaptive" machine learning with test data reuse



Speaker notes

...

IDEA

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

➡️ Differential privacy.[1]

Promising simulation results.[2-3]

Speaker notes

We would like to be able to reuse the test data while avoiding overfitting by somehow obfuscating the test data.

The idea of concealing the data is intuitively related to data privacy.

One way to do this is through the use of differential privacy...

...which has shown promising simulation results.

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

Speaker notes

Differential privacy is a beautiful mathematically rigorous definition of privacy, that has gained increasing attention in about the last 5 years (in fact, Cynthia Dwork and co-authors have received the Goedel price for the invention of DP last year, which is the highest price in theoretical computer science, named after the logician Kurt Goedel, who is known for his incompleteness theorems).

The idea behind DP is ... so that an adversary cannot learn an individual data point. We can apply this idea to test data reuse where a machine learning algorithm will not be able to learn individual records within the test data set and so will not be able to overfit to them.

DP is preserved under post-processing and adaptive composition, which means that it can be applied to the type of adaptive data analysis process that we are dealing with.

DIFFERENTIALLY PRIVATE ACCESS TO TEST DATA

Currently available literature:

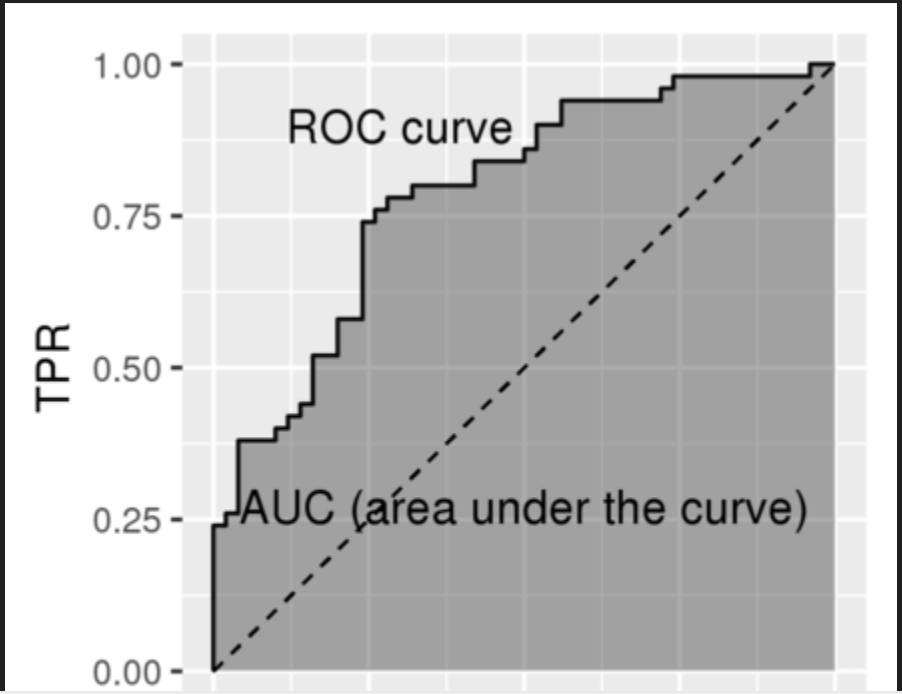
- Focuses on theory.
- Theoretical assumptions are too restrictive for most of applied data analysis and machine learning.
- Computational experiments are rather simplistic.

Speaker notes

...

Thresholdout + AUC = ❤️

Thresholdout$_{\mathrm{AUC}}$ [2] combines the Thresholdout technique[1] with AUC as the reported performance metric



Speaker notes

In order to report the performance on the test data in a differentially private way, we consider the Thresholdout technique which can return some simple kinds of performance metrics, and modify it in order to obtain the area under the ROC curve, which relates the true positive and false positive rates. The AUC is the recommended performance metric in many fields of medical research.

- Invariance to prevalence.
- Independence of the decision threshold (can be chosen later).
- Probabilistic meaning.
- Extensively used in the medical field, including medical imaging.

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:

Speaker notes

Essentially, our ThresholdoutAUC mechanism will return the AUC on the test data plus some noise if this performance metric on the test data and training data is significantly different, and otherwise will return the AUC on the training data and not provide any information on the test data at all.

THEOREM — ROUGH SUMMARY

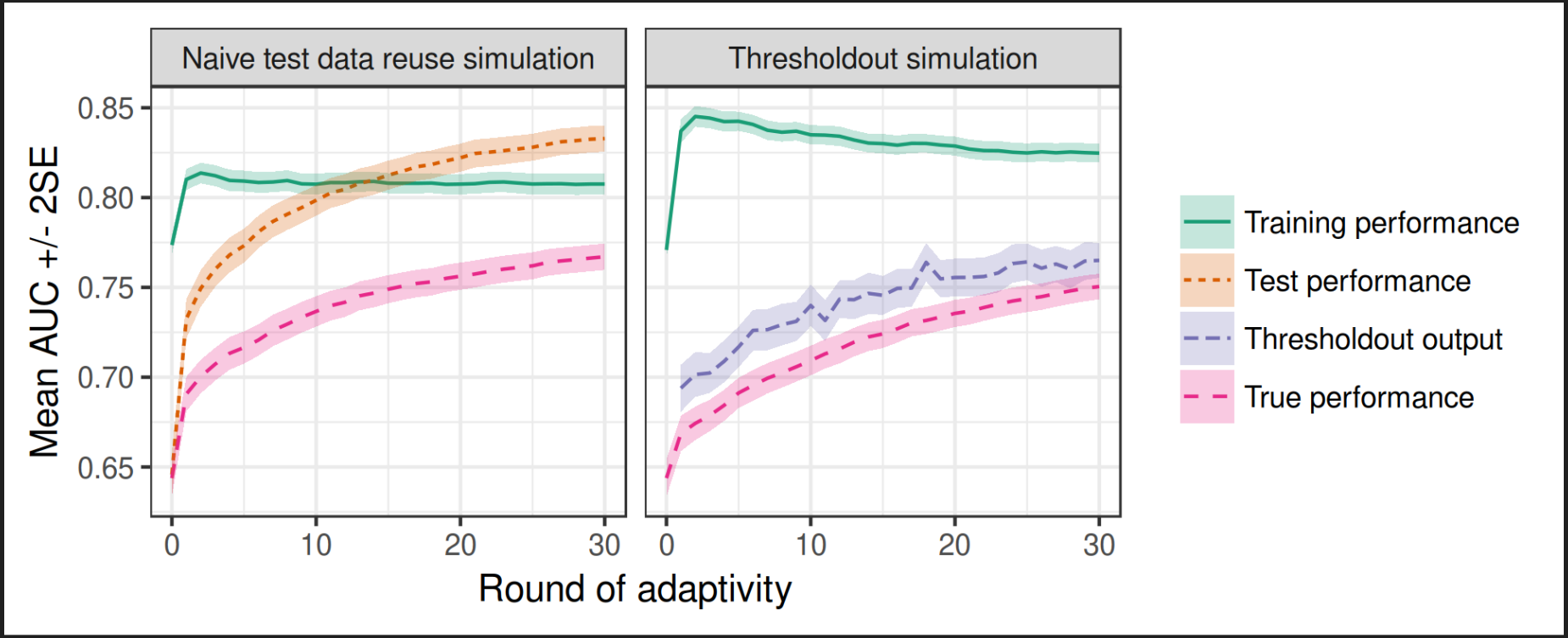
1. **If** a test dataset, which is used for performance evaluation repeatedly, is only accessed via $\text{Thresholdout}_{\text{AUC}}$. $\Longrightarrow$ **Then** with a high probability $(1 - \beta)$ the reported AUC estimates will be correct up to a small tolerance τ.
2. **Restriction**: Test data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

Speaker notes

Beta and tau are pre-specified by the analyst.

This result holds even when each analysis step is adaptively chosen based on the reported AUC estimates obtained from previous analyses on the same test data using $\text{Thresholdout}_{\text{AUC}}$.

LOGISTIC REGRESSION

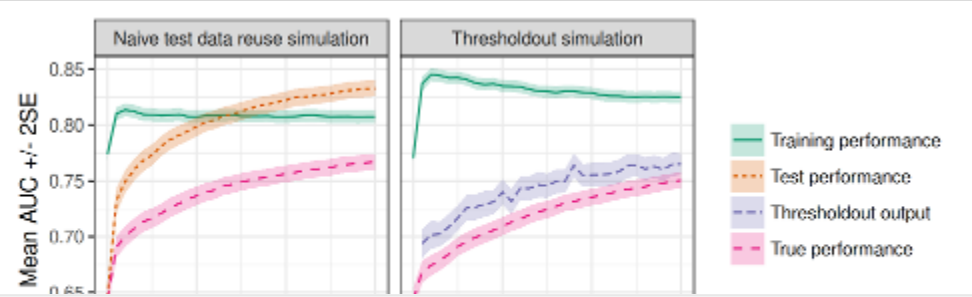


Speaker notes

...

LOGISTIC REGRESSION

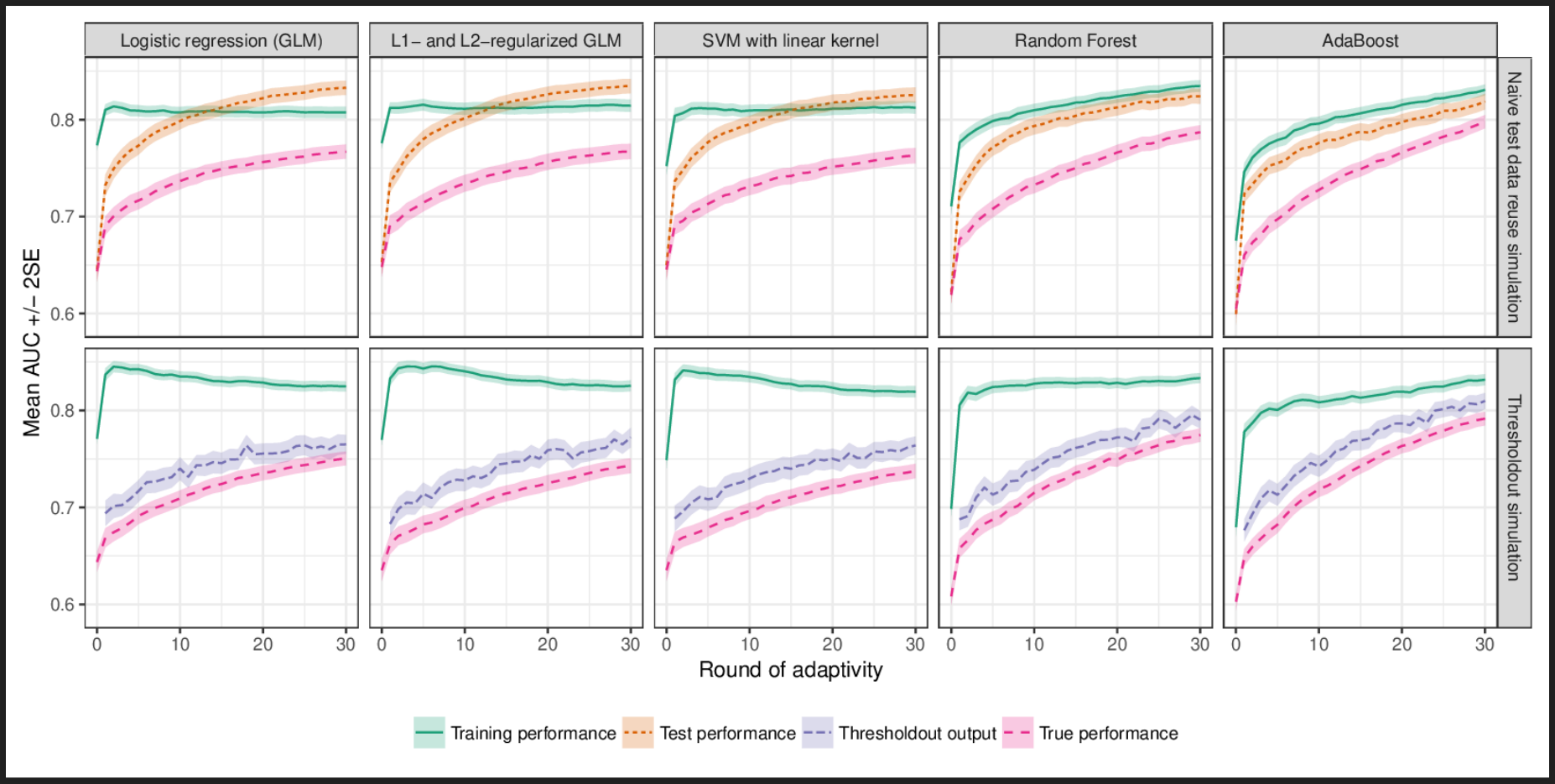
- **Naive approach**: Classifier learns the effect of local noise in the test data (overfitting).
- **Thresholdout approach**: The gap between the reported and the true AUC is much narrower!



Speaker notes

We may reuse the test data as often as we like, with the constraint that only AUC values can be reported back from the test data.

What model to try next is determined by the AUC values obtained on the test data for the previous models.



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]

Speaker notes

...

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$_{AUC}$.

Speaker notes

...

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

Speaker notes

At this point I would like to acknowledge my collaborators at Tulane at beyond...

Most importantly Dr. Yu-Ping Wang's MBB lab at Tulane.

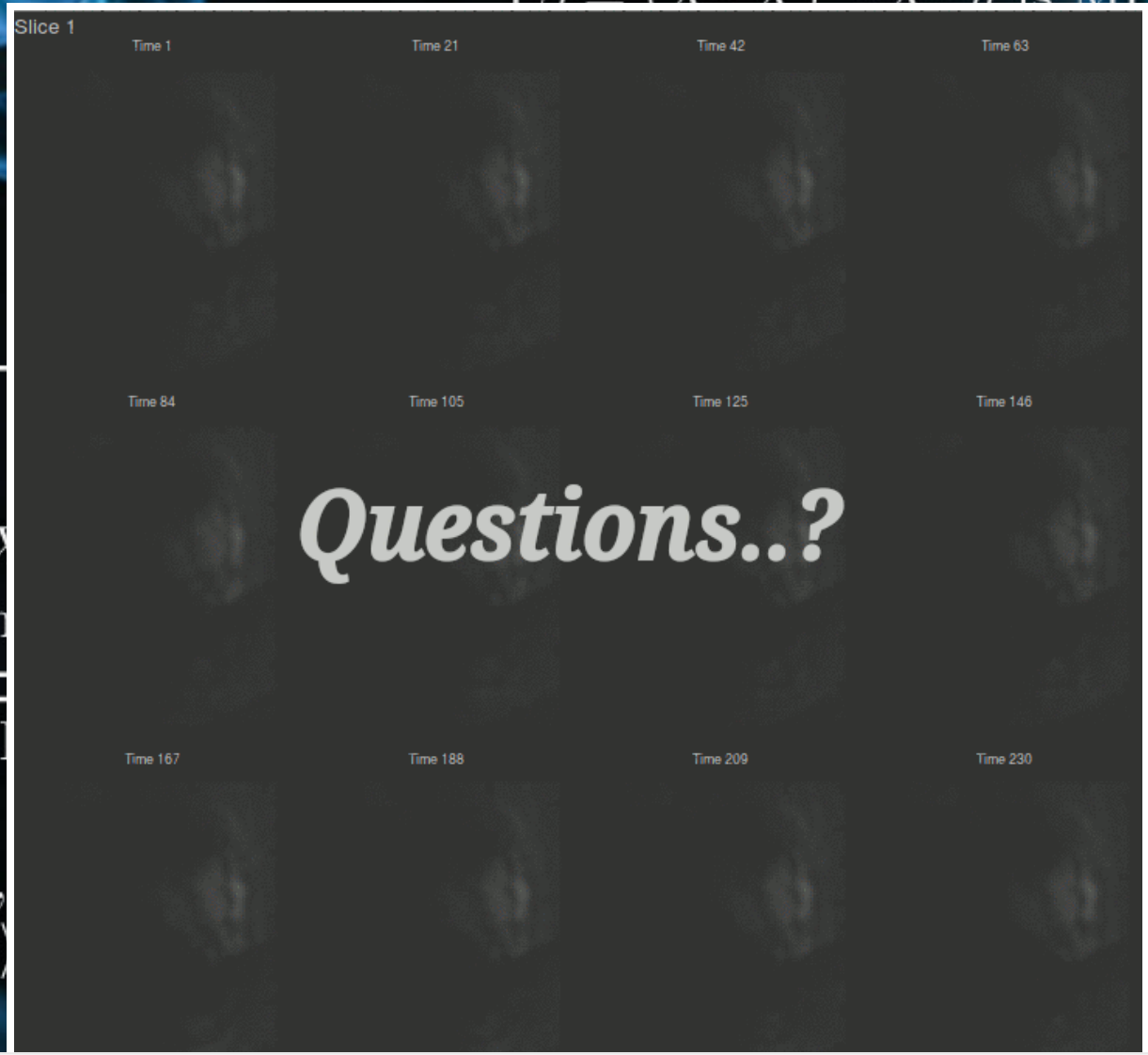
And a few other places where people who have helped me with my research work.

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

Speaker notes

...



Speaker notes

...