

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

A PhD prospectus presentation for the Bioinnovation PhD Program at Tulane University.

ALEXEJ GOSSMANN

March 12, 2018

PRECISION MEDICINE

Inter-personal diversity in the patients' biology

- differences in disease susceptibility/progression
 - differences in treatment efficacy
 - "personalized" treatment plans.

PRECISION MEDICINE

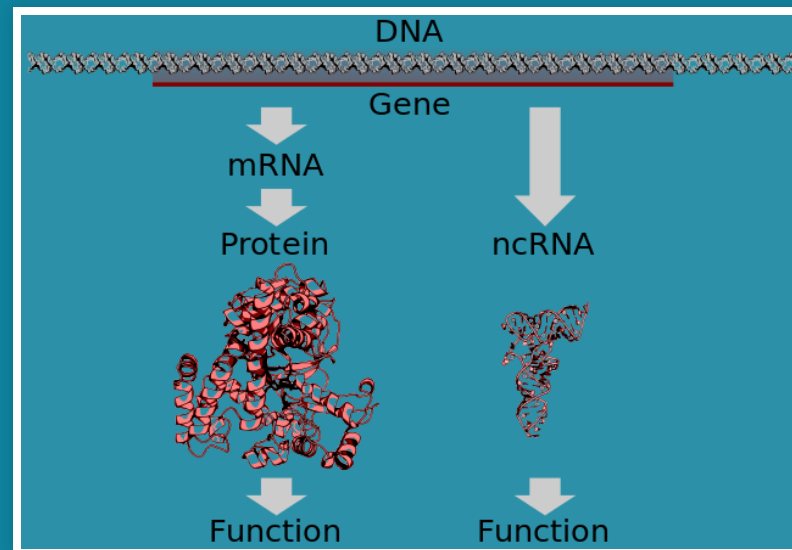
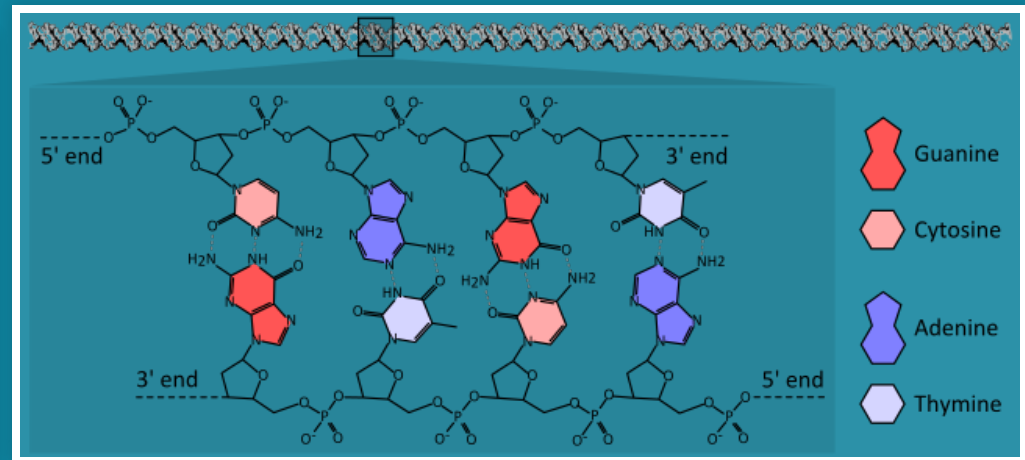
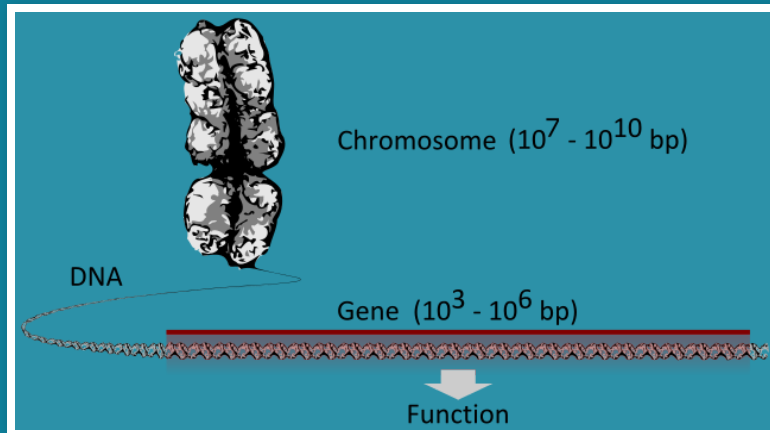
- New drugs and devices targeting specific subpopulations (or even individuals).
- No more treatment based on trial-and-error.
- ↑ Quality of healthcare
- ↓ Treatment time and cost.

PRECISION MEDICINE

Made possible by:

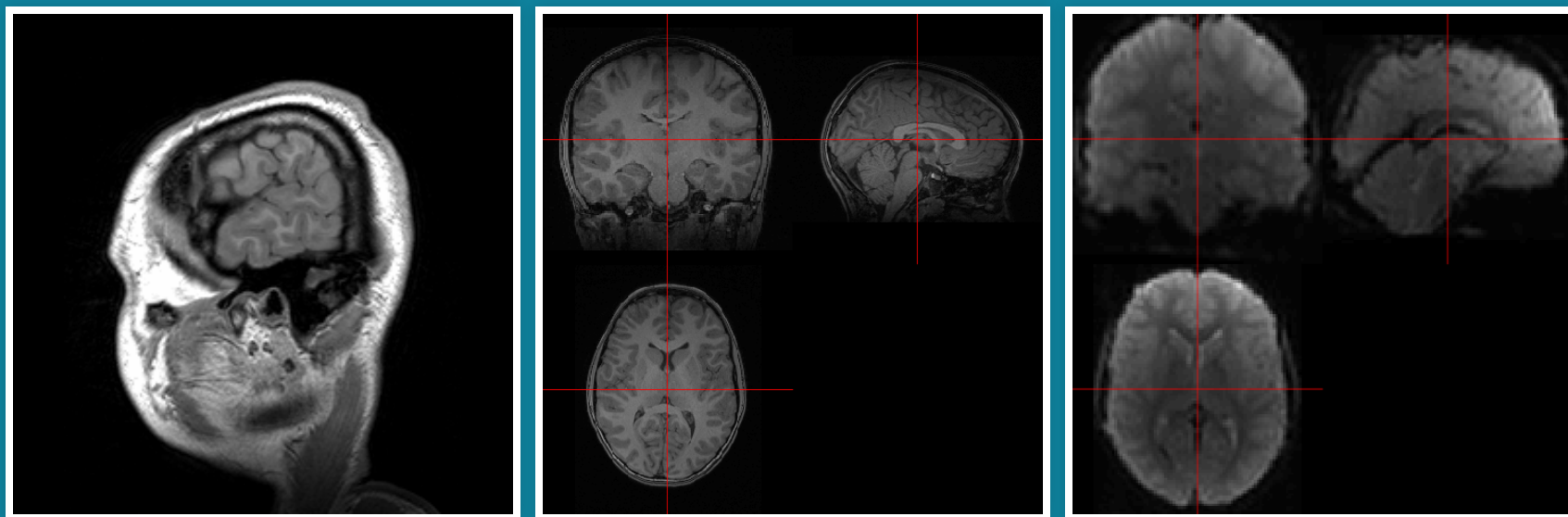
1. Big data including **genomics** and **neuroimaging**.
2. Computational methods including **machine learning** and **modern statistics**.

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



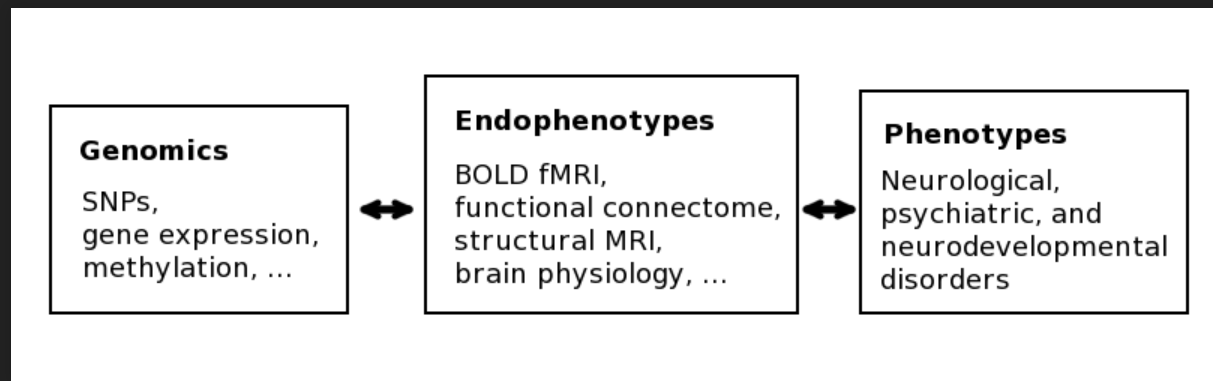
Source: Images by Thomas Shafee [CC BY 4.0] via Wikimedia Commons.

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



- Animation by Dwayne Reed at English Wikipedia [[CC BY-SA 3.0](#)] via [Wikimedia Commons](#).
- A randomly chosen subject from the Philadelphia Neurodevelopmental Cohort:
 - T1-weighted MRI before preprocessing ($192 \times 256 \times 160$ voxels).
 - fMRI after preprocessing ($79 \times 95 \times 79$ voxels at > 200 time points).

PRECISION MED. & MENTAL DISORDERS



- Neuroimaging as an endophenotype.^[1-2]
- Use of fMRI 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.

REPRODUCIBLE RESEARCH

Reproducibility = *re-performing the same analyses on the same data with the same code, while using a different data analyst* [Patil et al., 2016].

- Clear documentation of methods and analyses in peer-reviewed publications.
- Usage of publicly accessible datasets only.
- Publication of free & open-source software (through [CRAN](#), [Bioconductor](#), [PyPI](#), [RubyGems](#), [Github](#), etc.).
- One-off analysis scripts deposited on [Github](#).
- Open access, [arXiv](#), [bioRxiv](#), [Creative Commons](#).

MODELS

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon,$$

where x_1, x_2, \dots, x_p are predictor variables, ε is random noise, and y is the phenotype.

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

$$\Rightarrow y = f(x_{a_1}, x_{a_2}, \dots, x_{a_m}) + \varepsilon,$$

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

$$(Y \mid x_1, x_2, x_3, \dots) = (Y \mid x_5, x_8, x_{13})$$

x1	■	x1	■
x2	■	x2	■
x3	■	x3	■
x4	■	x4	■
x5	■	x5	■
x6	■	x6	■
x7	■	x7	■
x8	■	x8	■
x9	■	x9	■
x10	■	x10	■
x11	■	x11	■
x12	■	x12	■
x13	■	x13	■
x14	■	x14	■
x15	■	x15	■
x16	■	x16	■
...
...

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.

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.

FEATURE SELECTION IN GENOMICS AND NEUROIMAGING

- Prediction of a phenotype based on few features.
 - ↳ Inexpensive diagnosis.
- Elimination of noisy or redundant features.
 - ↳ More accurate prediction.
- Data-generated hypotheses.
 - ↳ Biological insights.

MULTIPLE HYPOTHESES TESTING

Feature selection as testing of hypotheses:

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

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

^(*): E.g., Bonferroni, Holm (1979), Hommel (1988).

^(**): E.g., Benjamini-Hochberg (1995), Benjamini-Yekutieli (2001).

SPARSE REGRESSION

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

- Yields a sparse solution $\hat{\beta}$.
- Computationally efficient (convex).
- Very useful in practice.
- **Problem:** how sparse should $\hat{\beta}$ be?
- **Problem:** how to do statistical inference on $\hat{\beta}$?

SORTED L-ONE PENALIZED ESTIMATION^[1]

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

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$; and

$|\mathbf{b}|_{(1)} \geq |\mathbf{b}|_{(2)} \geq \dots \geq |\mathbf{b}|_{(p)}$ 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 λ s.t. $\text{FDR}(\hat{\beta}_{\text{SLOPE}}) \leq q$ is guaranteed. *...if the explanatory variables have very small pair-wise correlations.*

[1]: Bogdan et. al., Annals Appl Stat, 2015.

GROUP SLOPE MOTIVATION

- Divide the data into groups by correlation.
- Then select/drop entire groups rather than individual variables.
- Redefine FDR w.r.t. groups: $gFDR$.

GROUP SLOPE

- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, X \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$
- $\boldsymbol{\beta}$ divided into J groups of sizes p_1, p_2, \dots, p_J , i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \dots, \boldsymbol{\beta}_J^T)^T$ with $\boldsymbol{\beta}_i \in \mathbb{R}^{p_i}$.



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

where


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

GROUP SLOPE - THEORETICAL GUARANTEES

- Given a user-specified $q \in (0, 1)$, we show how to construct λ , such that $\text{gFDR} \leq q$.^[1-3]
 - ➔ Confirmed with extensive simulation studies on synthetic and real data.^[1-3]
- Asymptotically minimax estimation.^[2]

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

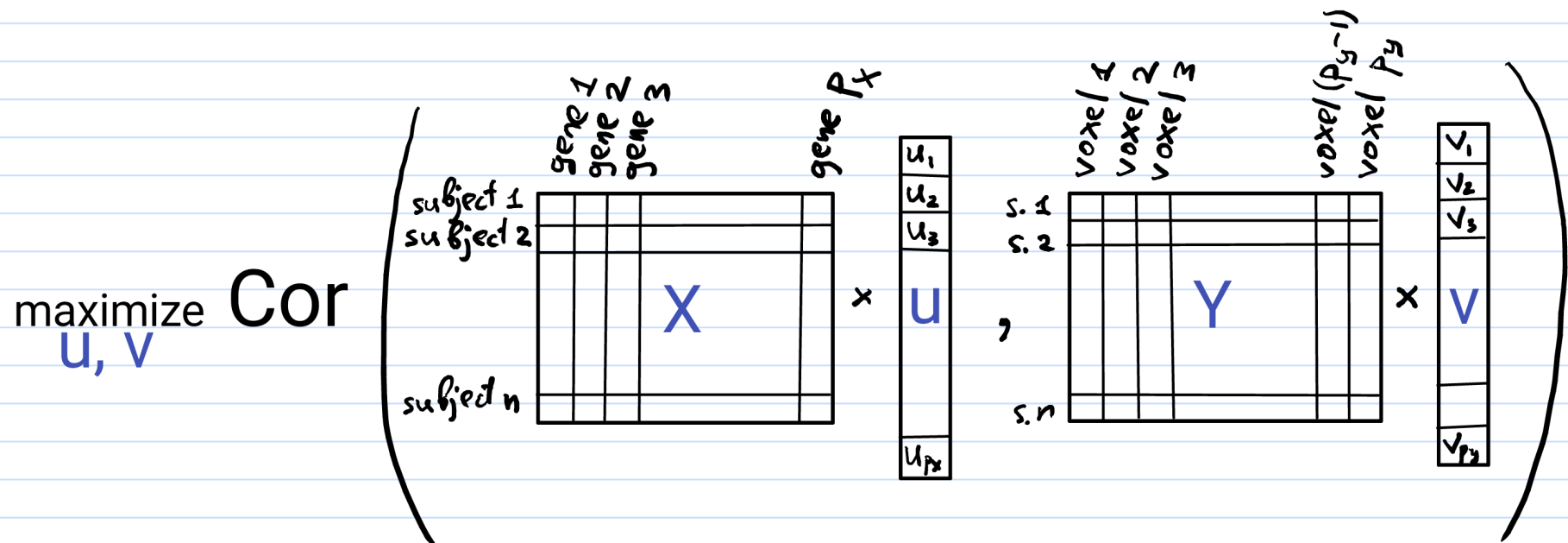
APPLICATION - FRAMINGHAM COHORT

- 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).

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).

CANONICAL CORRELATION ANALYSIS



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 CANONICAL CORRELATION ANALYSIS

$$\begin{aligned} &\text{maximize}_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \widehat{\text{Cov}}(Xu, Yv) = \frac{1}{n} u^T X^T Y v, \\ &\text{subject to} \quad \widehat{\text{Var}}(Xu) = 1, \widehat{\text{Var}}(Yv) = 1. \\ &\quad \quad \quad [\text{Hotelling, 1936}] \end{aligned}$$

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

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 $p_X, p_Y \gg n$.
- Selection of the sparsity parameters remains a challenging problem.

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


SPARSE CCA

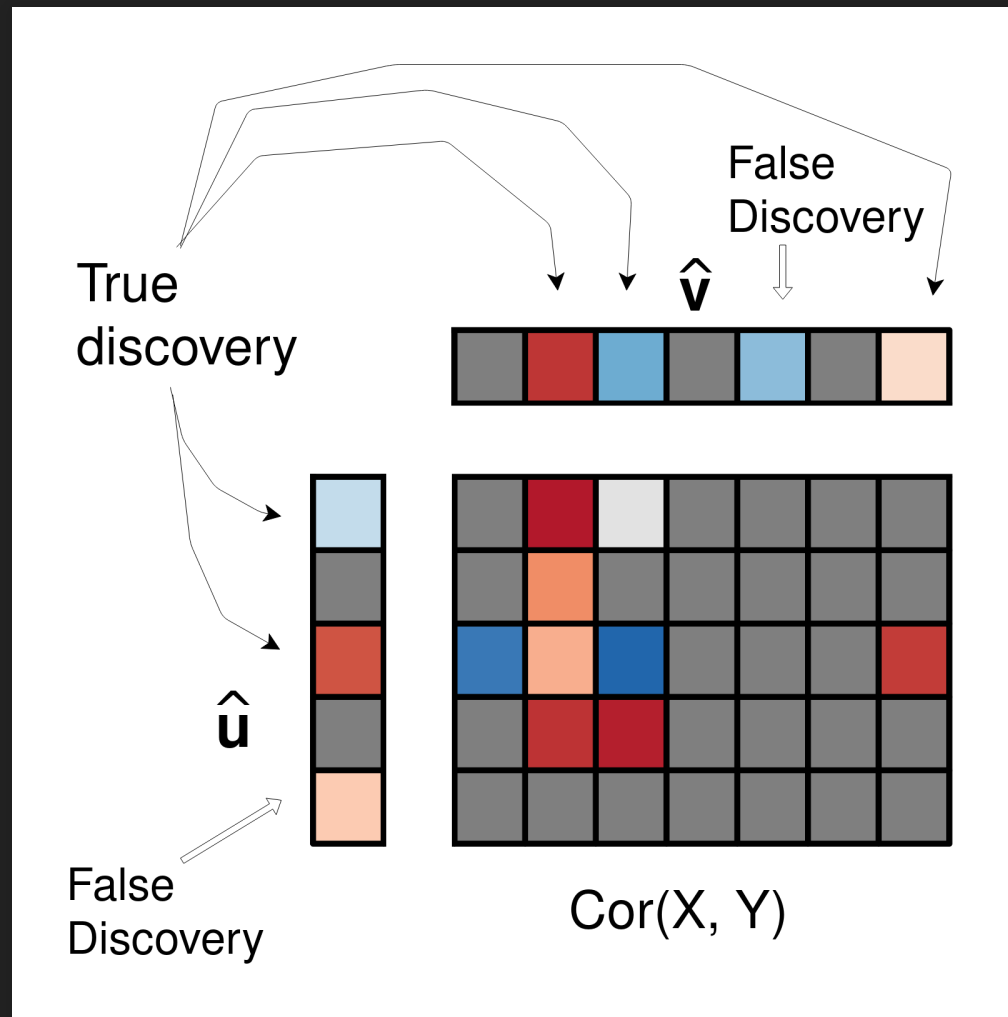


Select sparsity parameters in a data-driven fashion, such that FDR is controlled.

DEFINING FDR FOR SPARSE CCA

- Consider FDR in \mathbf{u} and in \mathbf{v} separately.
- Consider hypotheses tests $H_i : u_i = 0$.
- The null hypothesis H_i is true if the i th feature in X is uncorrelated with all features in Y , i.e., if
$$(\forall j \in \{1, 2, \dots, p_Y\}) : \rho_{i,j}^{XY} = 0.$$
- Let $R_{\hat{\mathbf{u}}}$ be the number of the rejected H_i , and $V_{\hat{\mathbf{u}}}$ the number of false rejections (i.e., when $\hat{u}_i \neq 0$ but $\rho_{i,j}^{XY} = 0$ for all j).

 Define: $\text{FDR}(\hat{\mathbf{u}}) := \mathbb{E} \left(\frac{V_{\hat{\mathbf{u}}}}{\max\{R_{\hat{\mathbf{u}}}, 1\}} \right) .$



→ False discovery proportions: $\text{FDP}(\hat{u}) = 1/3$ and $\text{FDP}(\hat{v}) = 1/4$. → $\text{FDR} = E(\text{FDP})$.

SlopeCCA:

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

gSlopeCCA:

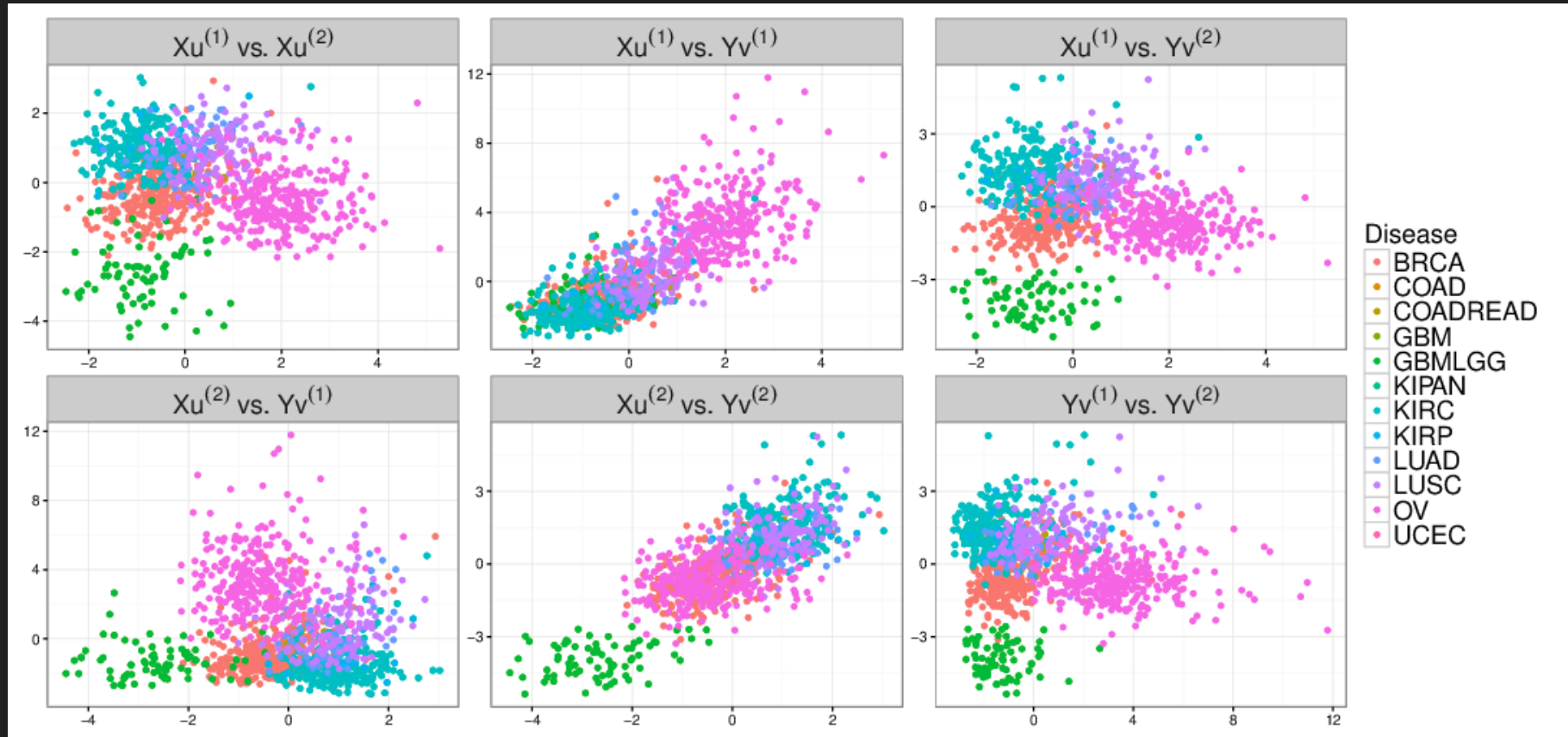
$$\begin{aligned} \text{minimize}_{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q} \quad & \left\{ -\mathbf{u}^T X^T Y \mathbf{v} + \sqrt{n} J_{\lambda^u} \left((\|\mathbf{u}_1\|_2, \dots)^T \right) + \sqrt{n} J_{\lambda^v} \left((\|\mathbf{v}_1\|_2, \dots)^T \right) \right\}, \\ \text{subject to} \quad & \|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1. \end{aligned}$$

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

SLOPECCA AND GSLOPECCA - THEORETICAL GUARANTEES

Asymptotic FDR guarantees if $Cov(X)$ and $Cov(Y)$
are block-diagonal.

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



Application example: Results of gSlopeCCA applied to methylation and mRNA data from 12 diseases available from the Cancer Genome Atlas data.

FDR-CORRECTED SPARSE CCA

A split-sample, two-step procedure:

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,

$$H_i^{(u)} : u_i^{(0)} = 0, \quad H_j^{(v)} : v_j^{(0)} = 0,$$

and adjust for multiple comparisons to control the FDR.

FDR-CORRECTED SPARSE CCA - THEORETICAL GUARANTEES

1. Equivalence to hypotheses:

$$H_i^{(u)} : \left(X^T Y \hat{\mathbf{v}}^{(0)} \right)_i = 0, \quad H_j^{(v)} : \left(Y^T X \hat{\mathbf{u}}^{(0)} \right)_j = 0$$

2. After approximating the distribution of $X^T Y \hat{\mathbf{v}}^{(0)}$, we can use Benjamini-Hochberg to control FDR.



Confirmed with extensive simulation studies on
synthetic and real data.^[1]

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

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.

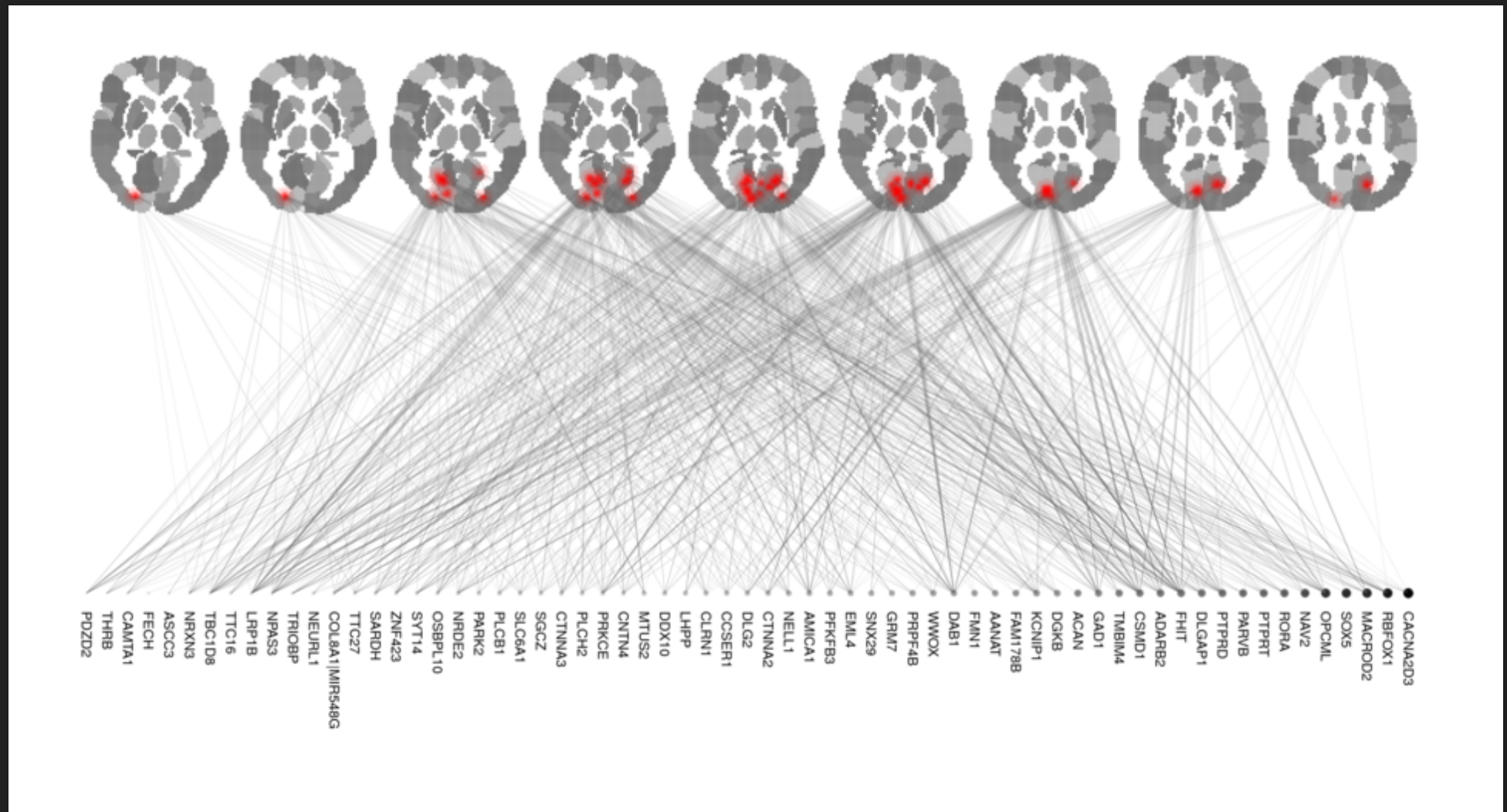
DATASET

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

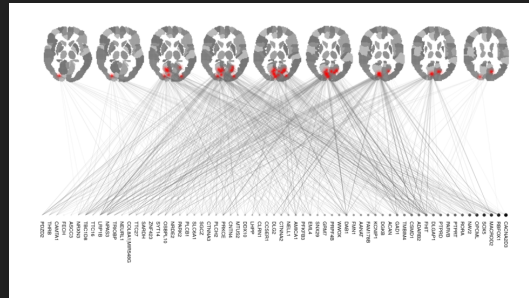
Use sparse CCA to identify the relationships between brain activity, brain connectivity, and genomics.

N-BACK FMRI VS. SNPS



Selection from 85796 brain voxels and 60372 genomic features.

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.

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])

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.

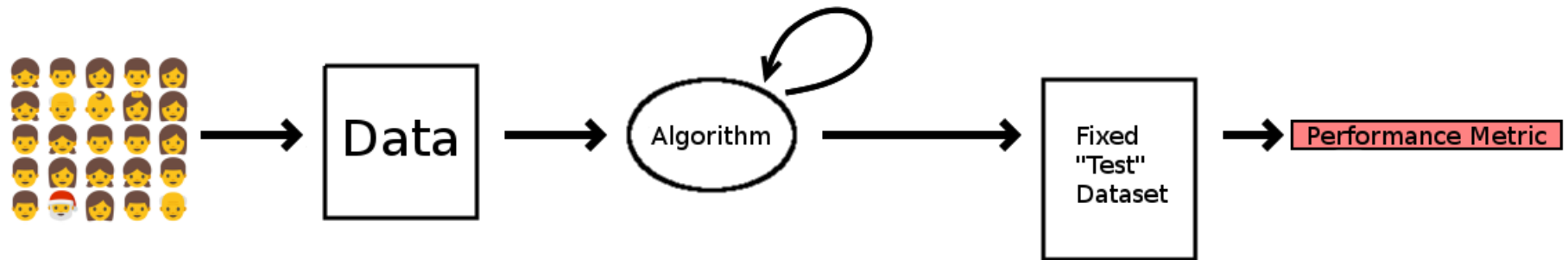
Training data: exploratory analysis, model fitting, parameter tuning, comparison of different machine learning algorithms, feature selection, etc.

⇒ Adaptive machine learning, risk of overfitting.

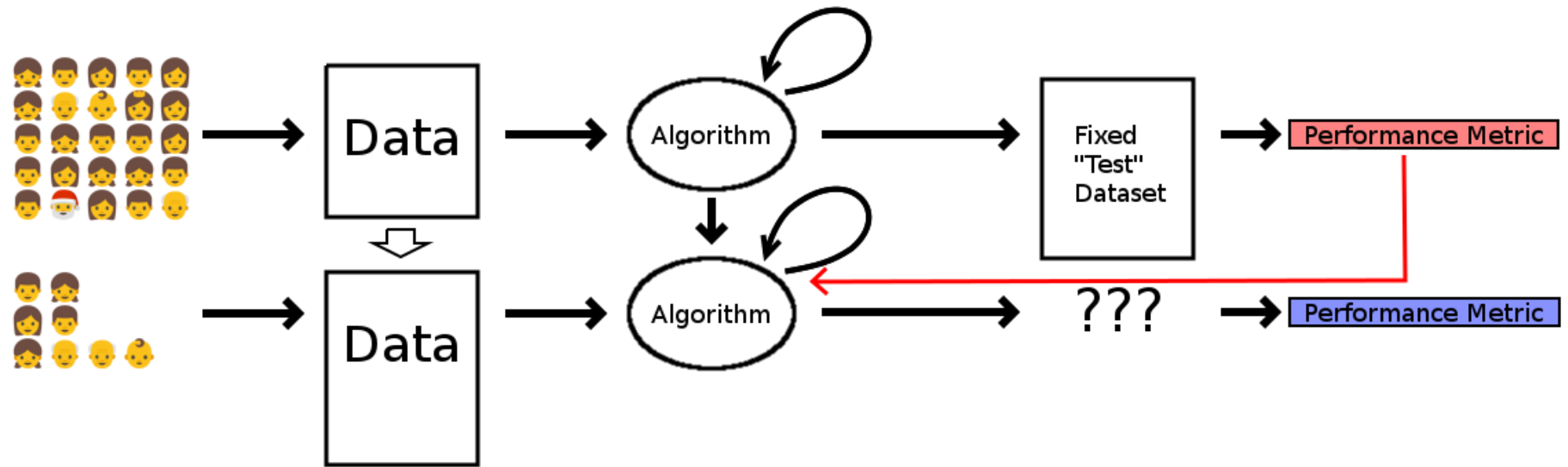
Test data: Performance evaluation *after the trained machine learning algorithm has been "frozen"*.

⇒ Accurate performance measures of the final model, if the test data is used only once.

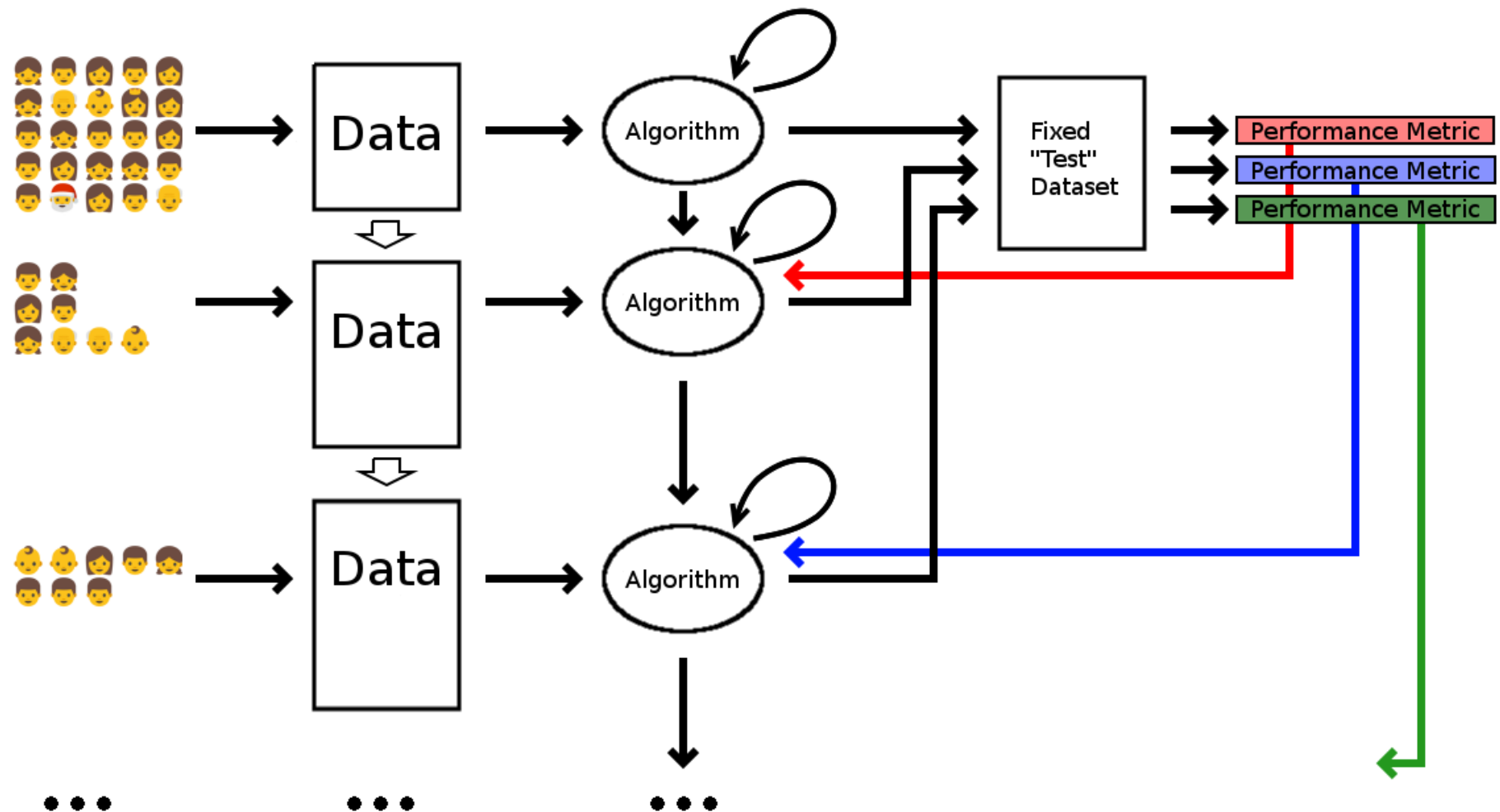
General machine learning process



"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]: Gossman et. al., SPIE 2018.

NEXT

- ➔ Investigation of the overfitting behavior resulting from the use of modern "black box" machine learning algorithms.
- ➔ Apply our test data reuse method to real neuroimaging and genomic data.

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.

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

Closed form solution:

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

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

LASSO:

$$E(Y) = X\beta, \text{ tuning parameter } \lambda \geq 0$$

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - Xb\|_2^2 + \lambda \|b\|_1 \right\}$$

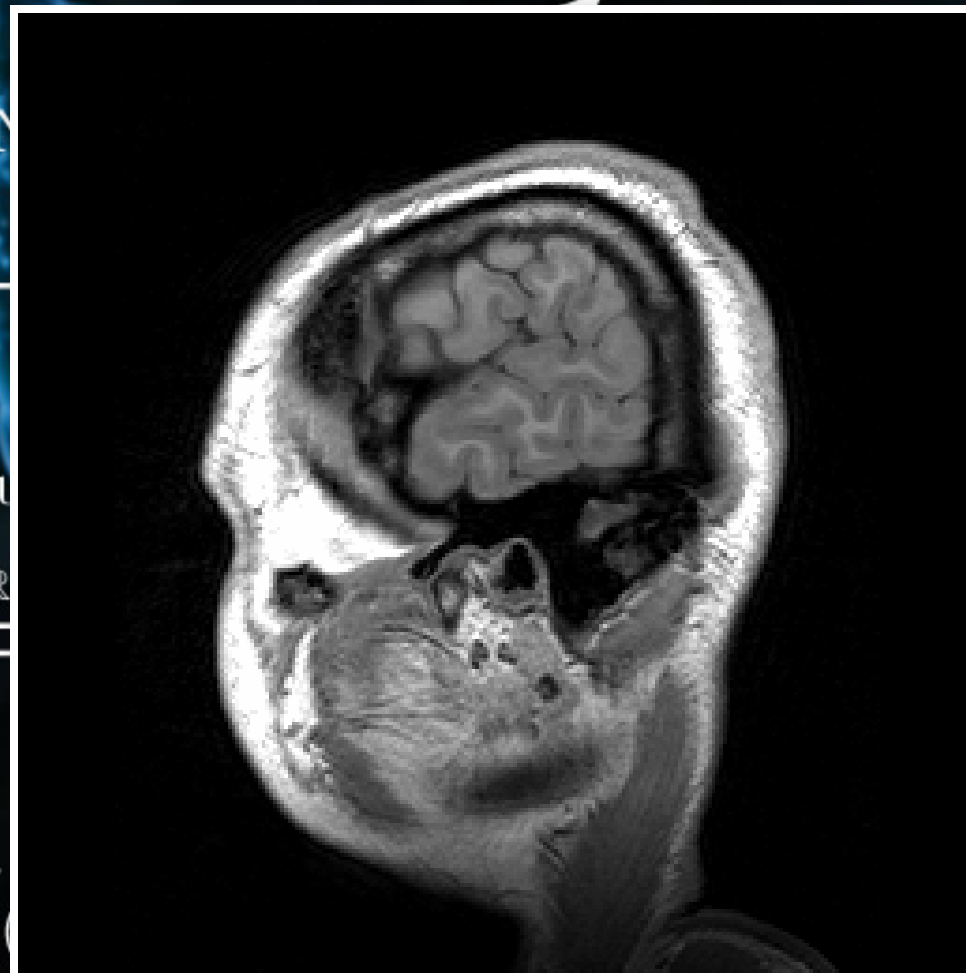
Generalized linear model (GLM):

$$f_{Y_i}(y_i) \text{ for } i \in \{1, 2, \dots, n\}$$

$$E(Y_i) = \exp(\langle T(y_i), \eta \rangle - A(\eta))$$

$$= f(\eta),$$

$$\beta.$$



Normality, $\lambda = 0$

Ridge regression:

$$E(Y) = X\beta, \text{ tuning parameter } \lambda \geq 0$$

$$= \operatorname{argmin}_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Closed form solution:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y.$$

$\alpha = 0$

Elastic net:

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

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