# Optimal Design for Discovery Science via Maximizing Discriminability: Applications in Neuroimaging

Shangsi Wang, Zhi Yang, Xi-Nian Zuo, Michael Milham, Cameron Craddock, Greg Kiar, William Gray Roncal, Eric Bridgeford, Consortium for Reliability and Reproducibility, Carey E. Priebe, Joshua T. Vogelstein

June 14, 2016

## Contents

# I Introduction

**Opportunity and Challenge** In this era of big data, many scientific, government, and corporate groups are collecting and processing massive datasets. To obtain optimal quantitative answers to any inquiry about data requires making two decisions: (i) how should the data be collected?, and (ii) how should the data be processed? When the downstream inference task is specified, a priori, we can collect and process data to optimize the performance of task. However, recently, across industry, governmental, and academic settings, certain datasets become benchmark or reference datasets. Such datasets are then used for a wide variety of different inferential problems. Collecting and processing these datasets requires massive institutional investments, and choices related to questions(i) and (ii) above have dramatic effects on all subsequent analyses. Optimally addressing experimental design decisions can yield significant savings in both the financial and human costs, and also improve accuracy of analytical results. Therefore, a theoretical framework to enable investigators to select from a set of possible design decisions in the absence of an explicit task or for multiple tasks could reap great rewards.

**Action** To this end, we have proposed and developed a formal definition of discriminability to guide data collection and processing. Discriminability is a non-parametric statistical property of a joint distribution in a hierarchical model, to differentiate between classes of objects. We prove that discriminability (which may be more aptly called reliability), provides a lower bound on predictive accuracy for any downstream inference task, even if we have never seen the covariates to predict in the processing. We then design an estimator of discriminability computed from test-retest data set, demonstrate that it is unbiased, and derive our estimators asymptotic distribution.

**Resolution** Numerical simulations are conducted to demonstrate the basic property of our discriminability estimator in a variety of settings. Then, we apply our approach to choose amongst a set of choices one must make when designing a neuroimaging study to investigate functional connectomics. We start by finding the maximally discriminable threshold for converting correlation matrices into binary graphs. Indeed, consistent with our theoretical and simulated results, maximizing the discriminability of our datasets also maximizes performance on a suite of different downstream inference tasks. We then ask about a series of pre-processing steps: should one motion correct or not, and should one implement global signal regression or not, etc. We determine the optimal choice for each pre-processing steps, and find the maximally discriminable pipelines amongst 64 pre-processing pipelines.

Thus, in total, our discriminability analysis is a powerful tool for making decisions about how to collect and analyze datasets designed for discovery science. We expect this method to be useful in a wide variety of applications, and therefore have made all the code open source and available from http://openconnecto.me.

# II Results

## II.A Theory

### II.A.1 Discriminability as a framework to guide processing

**Rigorously define discriminability** Discriminability measures the overall consistency and differentiability of observations. For example, if a subject is measured twice under the same conditions, two observations should be close to each other given the measure is consistent. In addition, one should be able to tell these two observations come from the same subject when compared to observations from other subjects given the measure is differentiable. We quantify this idea of consistency and differentiability through discriminability.

To formalize the definition of discriminability, consider the following generative process. For each sample $i$, there exists some true physical property $\boldsymbol{v}_i$. Unfortunately, we do not get to directly observed $\boldsymbol{v}_i$, rather, we measure it with some device, that transforms the truth from $\boldsymbol{v}_i$ to $\boldsymbol{w}_i$ via $f_{\boldsymbol{\phi}}$. The parameter $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ characterizes all options in the measurement, including, for example, which scanner to use, which resolution, the number of images, sampling rate, etc. The output of $f_{\boldsymbol{\phi}}$ is the "raw" observation data $\boldsymbol{w}_i$, but it is corrupt in various ways, including movement or intensity artifacts introduced by the measurement process.

Therefore, rather than operating directly on $\boldsymbol{w}_i$, we intentionally "pre-process" the data, in an effort to remove a number of nuisance variables. This pre-processing procedure further transforms the data from $\boldsymbol{w}_i$ to $\boldsymbol{x}_i$ via $g_{\boldsymbol{\psi}}$. The parameter $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ indexes all pre-processing options, including whether to perform motion correction, which motion correction, deconvolution, etc. More specifically, the entire code base, including dependencies, and even the hardware the pre-processing is running on, could count as $\boldsymbol{\psi}$. For brevity, we define $\boldsymbol{x}_i := g_{\boldsymbol{\psi}}\big(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)\big)$. We should notice that $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ by their natures are random functions which means even if we measure the same physical property $\boldsymbol{v}_i$ twice the results could be different.

Let $i$ denote the sample's unique *identity* (hereafter, referred to as the *subject*) and $t$ denote the trial number. Thus, there is a single $\boldsymbol{v}_i$ for subject $i$, but we have $\boldsymbol{x}_{i,t}$, which is the $t^{th}$ trial, implicitly also a function of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, which encodes all the details of the measurement and pre-processing. If both $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ together do not introduce too much noise, then we would expect that $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{i,t'}$ are *closer* to one another than either are to any other subject's data, $\boldsymbol{x}_{i',t}$. Define $\delta$ to be a metric computing the distance between two data points, $\delta\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Formally, we expect that $\delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'}) \leq \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$, for most combinations of $i, i' \neq i, t, t' \neq t, t''$. For brevity, let $\delta_{i,t,t'} := \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'})$ and $\delta_{i,i',t,t''} := \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$. This intuition leads to our definition of discriminability:

$$D(\boldsymbol{\psi}, \boldsymbol{\phi}) = \mathbb{P}[\delta_{i,t,t'} \leq \delta_{i,i',t,t''}] \tag{1}$$

In words, discriminability is the probability that within subject distance is smaller than across subject distance. Implicitly, $D(\boldsymbol{\psi}, \boldsymbol{\phi})$ also depends on the distribution of true physical property $\boldsymbol{v}$, that is

$$D(\boldsymbol{\psi}, \boldsymbol{\phi}) = E(\mathbb{P}[\delta_{i,t,t'} \leq \delta_{i,i',t,t''}] | \boldsymbol{v}_i, \boldsymbol{v}_{i'}) \tag{2}$$

Through out this paper, we will assume the distribution of $\boldsymbol{v}_i$ is fixed and independent of $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$.

**Define optimizing pipeline problem** When trying to find the best data collection and processing pipeline, we try to maximize the discriminability of processed data, that is

$$\underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}, \boldsymbol{\phi} \in \boldsymbol{\Phi}}{\text{maximize}} \quad D(\boldsymbol{\psi}, \boldsymbol{\phi}) \tag{3}$$

It is often the case that data collection is out of control of researchers, that is $\boldsymbol{\phi}$ is a fixed element in $\boldsymbol{\Phi}$. Therefore, we are only interested in finding the best pre-processing routine encoded by $\boldsymbol{\psi}$. This is also the focus of this paper, since we do not have opportunity to make decision on data collection choices. In this case, we drop $\boldsymbol{\phi}$ in our notation and only maximize the discriminability over set $\boldsymbol{\Psi}$

$$\underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}}{\text{maximize}} \quad D(\boldsymbol{\psi}) \tag{4}$$

This approach is intuitive and easy to understand. We will show in the theory section that maximizing discriminability leads to good prediction performance. In addition, an unbiased estimator is designed to compute discriminability from test-retest data set. In the simulation and application section, we will demonstrate the utility of discriminability through data experiments.

### II.A.2 Optimizing discriminability optimizes bound on performance for any task

Consider the situation that the downstream inference task is classification, that is in addition to $\boldsymbol{v}_i$, there are other properties of sample $i$ of interest; we call all of them $\boldsymbol{y}_i \in \mathcal{Y}$. These may include, for example, the phenotype of the subject, including personality tests, demographic information, and genetic data. In this paper, we focus on binary classification problem that is $\mathcal{Y} = \{0, 1\}$. The goal of experimental design, in this context, is to choose $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ to make prediction of $\boldsymbol{y}_i$ based on observation $\boldsymbol{x}_i$ easier. In this section, we will see that given two pipelines $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, the one with larger discriminability is more likely to have better prediction performance.

To quantify the performance of our choice, we introduce some assumptions. First, assume that each $(\boldsymbol{v}_i, \boldsymbol{y}_i)$ pair is sampled independently and identically from some distribution, $(\boldsymbol{v}_i, \boldsymbol{y}_i) \overset{iid}{\sim} F_{V,Y}$. The goal is to predict the binary-valued *target* variable $\boldsymbol{y}_i$, using $\boldsymbol{x}_i$ as the *predictor* variables. Given a classifier

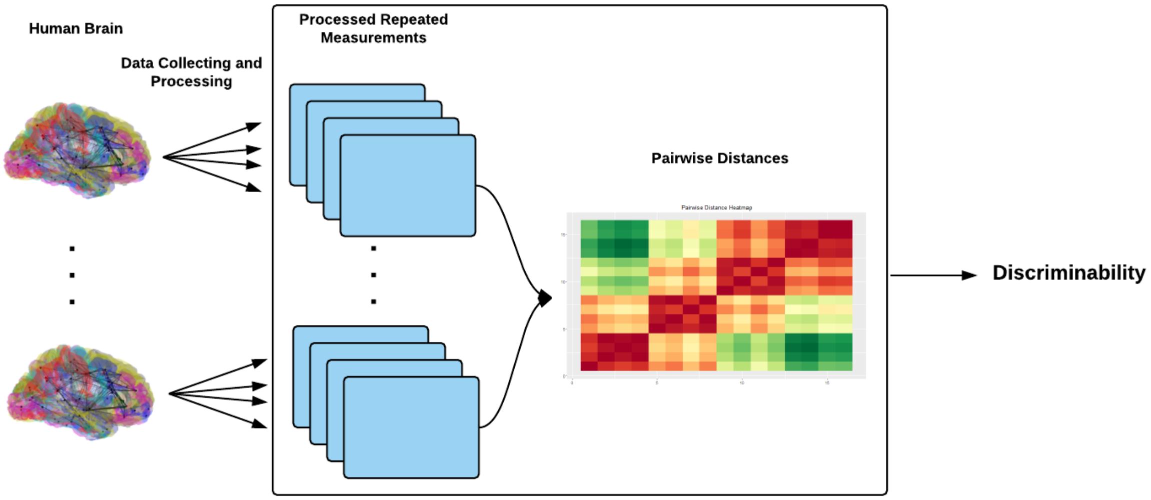
Human Brain
Data Collecting and Processing
Processed Repeated Measurements
Pairwise Distances
Pairwise Distance Heatmap
Discriminability

Figure 1: **Decision Making Through Discriminability Framework.** Test-retest data is collected from multiple subjects. The data is processed by a set of pipelines. For each processing pipeline, pairwise distances between observations are computed and discriminability is estimated.

$C : \mathcal{X} \to \mathcal{Y}$, to quantify the performance of classifier, we define the loss function $L(C)$ to be the probability of making error in prediction that is

$$L(C) = \mathbb{P}(C(\boldsymbol{x}_i) \neq \boldsymbol{y}_i)$$

It is well known the minimal prediction error is achieved by Bayes classification.

$$L^*(\boldsymbol{x}_i, \boldsymbol{y}_i) := L(C^B)$$

where $C^B$ is the Bayes classifier which is defined by

$$C^B(\boldsymbol{x}_i) := \underset{y \in \{0,1\}}{\text{argmax}}\, \mathbb{P}(\boldsymbol{y}_i = y | \boldsymbol{x}_i)$$

$L^*$ is determined by the distribution of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Since $\boldsymbol{x}_i$ depends on pipeline $\boldsymbol{\psi}$, we denote the loss of pipeline $\boldsymbol{\psi}$ by $\ell(\boldsymbol{\psi})$ which is the Bayes prediction error of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

$$\ell(\boldsymbol{\psi}) := L^*(\boldsymbol{x}_i, \boldsymbol{y}_i) = L^*(g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)), \boldsymbol{y})$$

The next theorem shows the relationship between Bayes classification error ad discriminability. Under assumptions that the noise is additive, we can prove theorem 1 which asserts that Bayes classification error is bounded by a decreasing function of discriminability.

**Theorem 1.** *There is a decreasing function $h$ which only depends on $\boldsymbol{v}$ and $\boldsymbol{y}$, such that*

$$\ell(\boldsymbol{\psi}) \leq h(D(\boldsymbol{\psi}))$$

As a consequence, we expect the classification performance to be good when the discriminability is large. An immediate corollary justifies using discriminability to select the optimal processing pipeline.

**Corollary 2.** *Given two processing pipelines $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, suppose $\boldsymbol{\psi}_1$ is more discriminable than $\boldsymbol{\psi}_2$, that is $D(\boldsymbol{\psi}_1) > D(\boldsymbol{\psi}_2)$. If $\ell(\boldsymbol{\psi}_2) \geq h(D(\boldsymbol{\psi}_1))$, then*

$$\ell(\boldsymbol{\psi}_1) \leq \ell(\boldsymbol{\psi}_2)$$

*Also, we must have*

$$\ell(\boldsymbol{\psi}_1) \leq h(D(\boldsymbol{\psi}_2))$$

It tells us for any distribution of $\boldsymbol{y}$, we have a tighter bound on Bayes error using the more discriminable pipeline. When choosing from two processing pipelines $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, we should first compute $D(\boldsymbol{\psi}_1)$ and $D(\boldsymbol{\psi}_2)$. We then select the pipeline which yields larger discriminability to have lower bound on the Bayes classification error. This theorem justifies maximizing discriminability for subsequent classification tasks.

### II.A.3 Estimating discriminability

In real world, distribution of $\boldsymbol{x}_{i,t}$ may never known to us; hence, it is not possible to compute discriminability $D(\boldsymbol{\psi})$ or $D$ in short when there is no ambiguity in processing pipelines under consideration. However, samples $x_{i,t}$ are observed, and we can approximate true discriminability $D$ using an estimator $\hat{D}$ which is a function of observed samples. For each pair of observations $x_{i,t}$ and $x_{i,t'}$ from the same subject $i$, we first define

$$\hat{D}_{i,t,t'} = \frac{\sum\limits_{i' \neq i}^{n} \sum\limits_{t''=1}^{s} \mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}}{(n-1)s}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, $n$ is the number of subjects, and $s$ denotes the number of observations per subject. $\hat{D}_{i,t,t'}$ approximate the probability that distances from observations of other subjects to the $t^{th}$ observation of subject $i$ is larger than the distance between $t^{th}$ and $t'^{th}$ trial of subject $i$. Then, we define the discriminability estimator $\hat{D}$ to be the mean of $\hat{D}_{i,t,t'}$.

$$\hat{D} := \frac{\sum\limits_{i=1}^{n} \sum\limits_{t=1}^{s} \sum\limits_{t' \neq t}^{s} \hat{D}_{i,t,t'}}{ns(s-1)}$$

$\hat{D}$ serves as the sample approximated discriminability. The next two lemmas asserts that the discriminability estimator $\hat{D}$ is unbiased and converges to $D$ as $n$ goes to infinity.

**Lemma 1.** *$\hat{D}$ is an unbiased estimator of $D$, that is*

$$E(\hat{D}) = D$$

**Lemma 2.** *As $n \to \infty$, $\hat{D}$ converges to $D$ in probability, that is*

$$\hat{D} \xrightarrow{p} D$$

## II.B Simulations

### II.B.1 Convergence of discriminability estimator

In Lemma 1 and 2, we claim discriminability $\hat{D}$ is unbiased and converges to the true population discriminability in probability. We demonstrate this idea with simulation. We consider a simple case that $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ together introduces independent Gaussian noise $\epsilon$, that is

$$\boldsymbol{x}_{i,t} = g_{\boldsymbol{\psi}}\big(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)\big) = \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t} \tag{5}$$

We sample $\boldsymbol{v}_i$ and $\boldsymbol{\epsilon}_{i,t}$ independently from standard Gaussian distribution. That is $\boldsymbol{v}_i \overset{i.i.d.}{\sim} \mathbb{G}(0,1)$ and $\boldsymbol{\epsilon}_{i,t} \overset{i.i.d.}{\sim} \mathbb{G}(0,1)$. For each subject, we sample two observations and let the number of subjects $n$ vary from $10$ to $200$. For each value of $n$, we repeatedly generate data and compute discriminability $100$ times to estimate the distribution of $\hat{D}$. With this data generation scheme, we can compute the population discriminability $D$ from numerical integration $0.6150$. The figure 1 shows the difference $\hat{D}$ and $D$. We can see from the figure that sample discriminabiity $\hat{D}$ converges to $D$ as the number of subject increases.

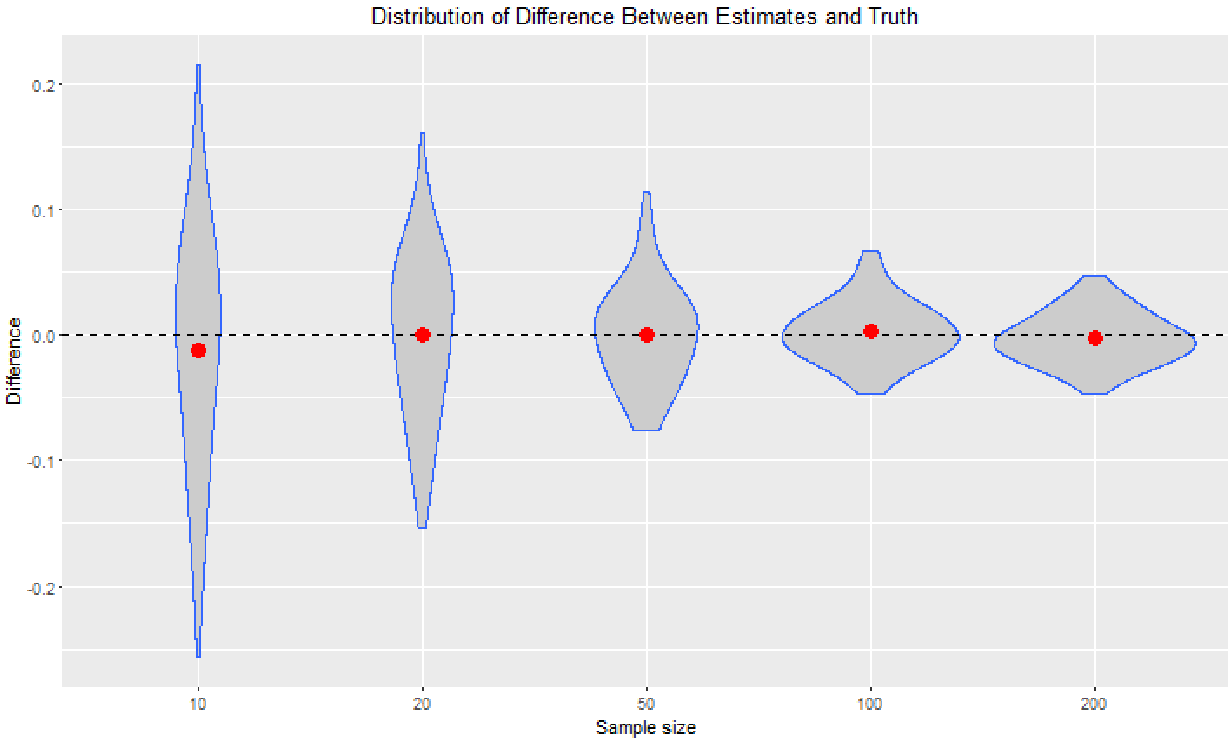

Figure 2: **Convergence of sample discriminability.** Distribution of sample discriminability is estimated. The red dots indicate the mean over $100$ repeats. As the number of subjects increases, the sample discriminability converges to the true population discriminability.

### II.B.2 Parameter selection through discriminability

In this simulation, we consider the task of projecting 2-dimensional observations linearly into 1-dimensional space. Again, we assume additive noise. In addition to $\boldsymbol{x}_{i,t}$, there is a binary class label $\boldsymbol{y}_i$ associated with subject $i$. The true physical property is Gaussian distributed conditioned on $\boldsymbol{y}_i$,

$$\boldsymbol{v}_i|\boldsymbol{y}_i = 1 \overset{i.i.d.}{\sim} \mathbb{G}(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}) \text{ and, } \boldsymbol{v}_i|\boldsymbol{y}_i = 0 \overset{i.i.d.}{\sim} \mathbb{G}(\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix})$$

The optimal linear projection should keep two classes separated which is just keep the first dimension of the observations. We consider two cases for the distribution $\boldsymbol{\epsilon}_{i,t}$. The first case is that $\boldsymbol{\epsilon}_{i,t}$ has larger variance in the first dimension; the other case is that $\boldsymbol{\epsilon}_{i,t}$ has larger variance in the second dimension. We use both discriminability and principal component analysis to find the optimal linear projection. The result of two cases are provided in two columns of figure 3. In the first case, both methods find the linear projection which separate two classes. However, in the second case only discriminability recovers the projection which separate two classes.

## II.C Connectome Applications

### II.C.1 optimal Discriminability yields optimal predictive accuracy

**real experiment** Describe the data and threshold experiment.

**real experiment** Emphasize that discriminability selects threshold which is close to optimal for multiple tasks.

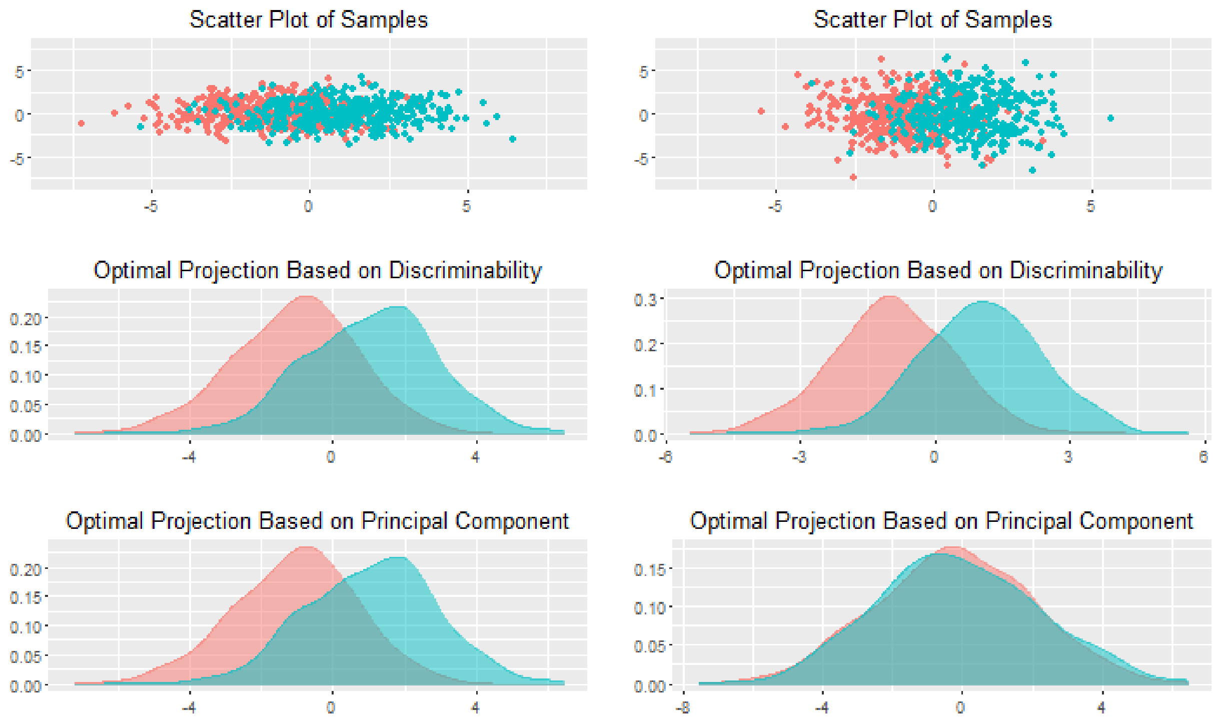

Figure 3: **Linear projection based on by PCA and Discriminability.** Linear projections are computed using PCA and optimizing Discriminability. Maximizing discriminability yield separated samples which have Bayes optimal classification error.

### II.C.2 fMRI Processing pipelines

**real experiment** Describe the 12 data sets and 64 pipelines.

**real experiment** Decide the best among 64 pipelines.

**real experiment** Decide the optimal for each decision (atlas, nff vs frf, ant vs fsl, nsc vs scr, gsr vs ngs) using anova test.

**real experiment** Describe how to convert raw graphs to rank graphs

**real experiment** Decide rank is better, especially when global signal regression is not performed.

### II.C.3 DTI processing pipelines

**real experiment** Describe we process a dti data data set with rank, raw and log for 15 atlases.

**real experiment** We see a trend that large roi are better. Since discriminability after removing outliers is close to one, more experiments need to be done.

### II.C.4 DTI vs. fMRI

**real experiment** Describe the 4 fmri and dti data sets.

**real experiment** After removing outliers, dti is more discriminable than fmri.

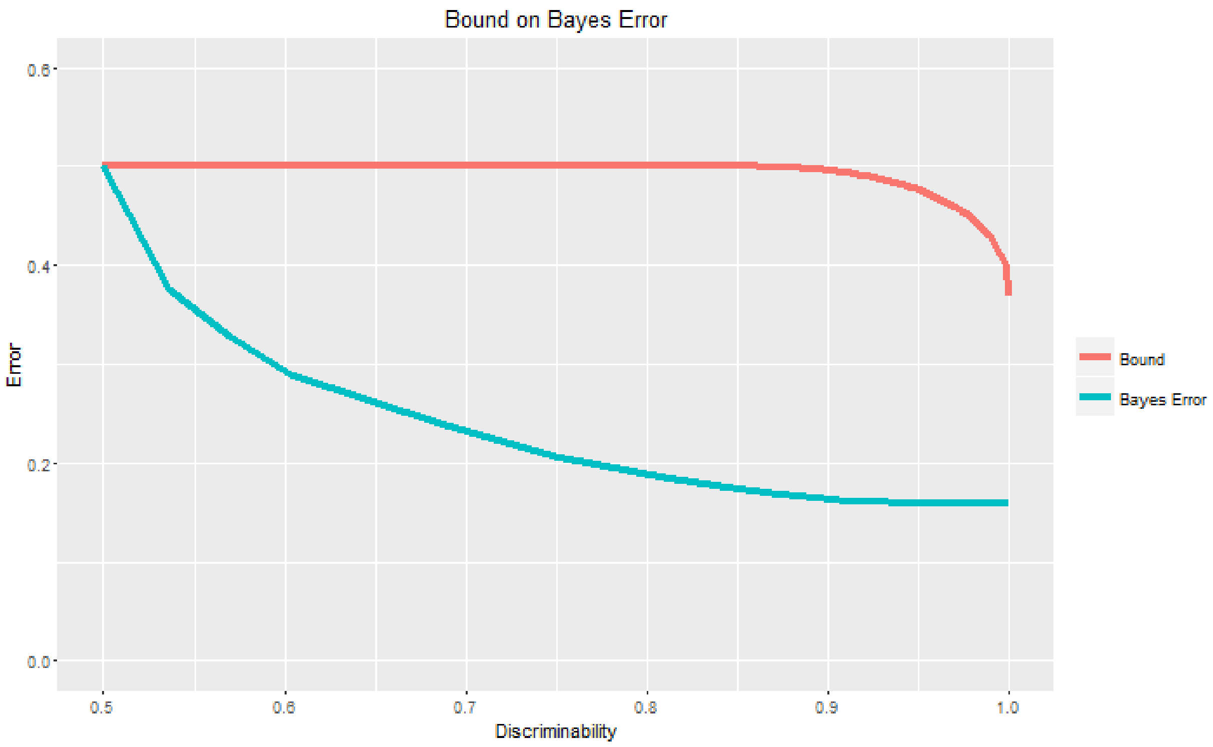

Figure 4: **Prediction error vs discriminability.** Samples are drawn from a mixture of two Gaussian distribution with additive noise. The red line is the theoretical bound derived from the theorem 1. The green line indicates the true Bayes error.

## III Discussion

**Summary** We propose a definition of discriminability and apply it to a variety of set ups.

**Related Work** I2C2, DISCO and GICC

**Next Steps**

## IV Appendix

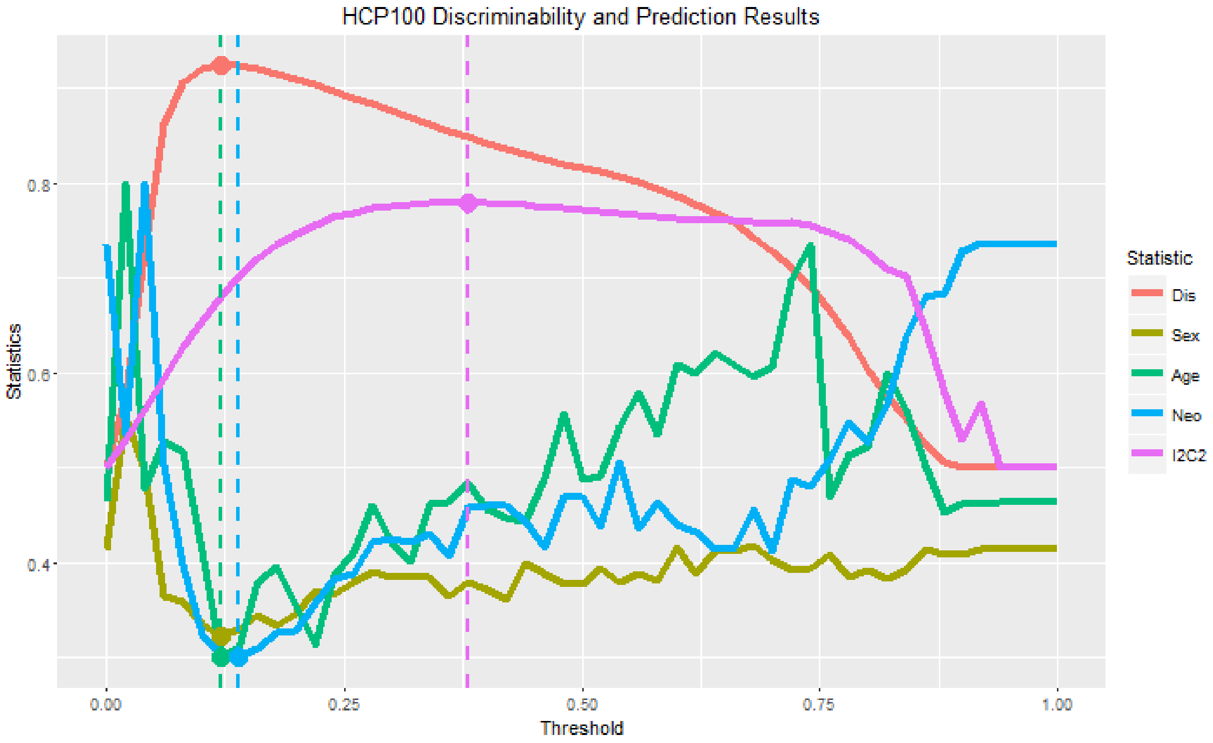

Figure 5: **Optimizing Discriminability yields optimal prediction accuracy for multiple covariates.** HCP100 is used to investigate optimal threshold to convert correlation graphs into binary graphs. The threshold is varied from 0 to 1. For each value of threshold, the discriminability is computed; sex, age and a neuro factor are predicted using k-NN. The threshold maximizing discriminability is close to optimal thresholds for predicting three covariates.

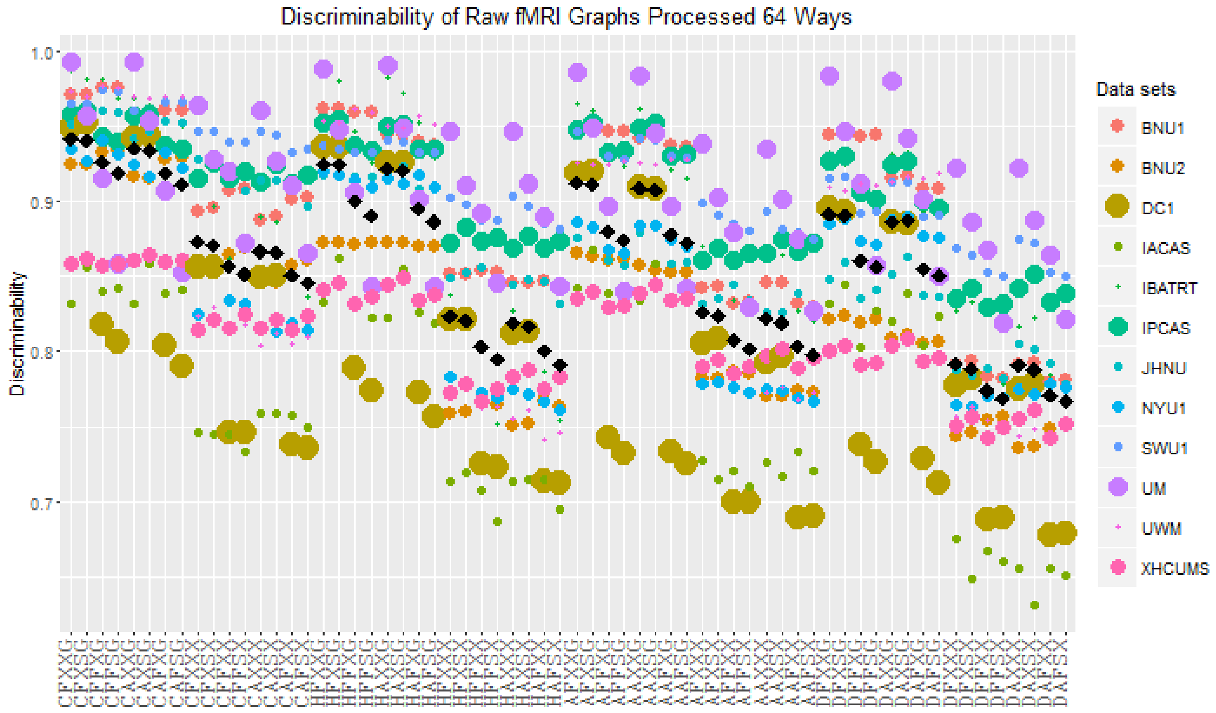

Figure 6: **Discriminability of raw fmri graphs from 12 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS processed by 64 pipelines are computed. Color of dot indicates data set and size indicates the number of observations in data set. The black square indicates the mean discriminability across 12 data sets. CFXXG pipeline has the best mean discriminability across data sets.

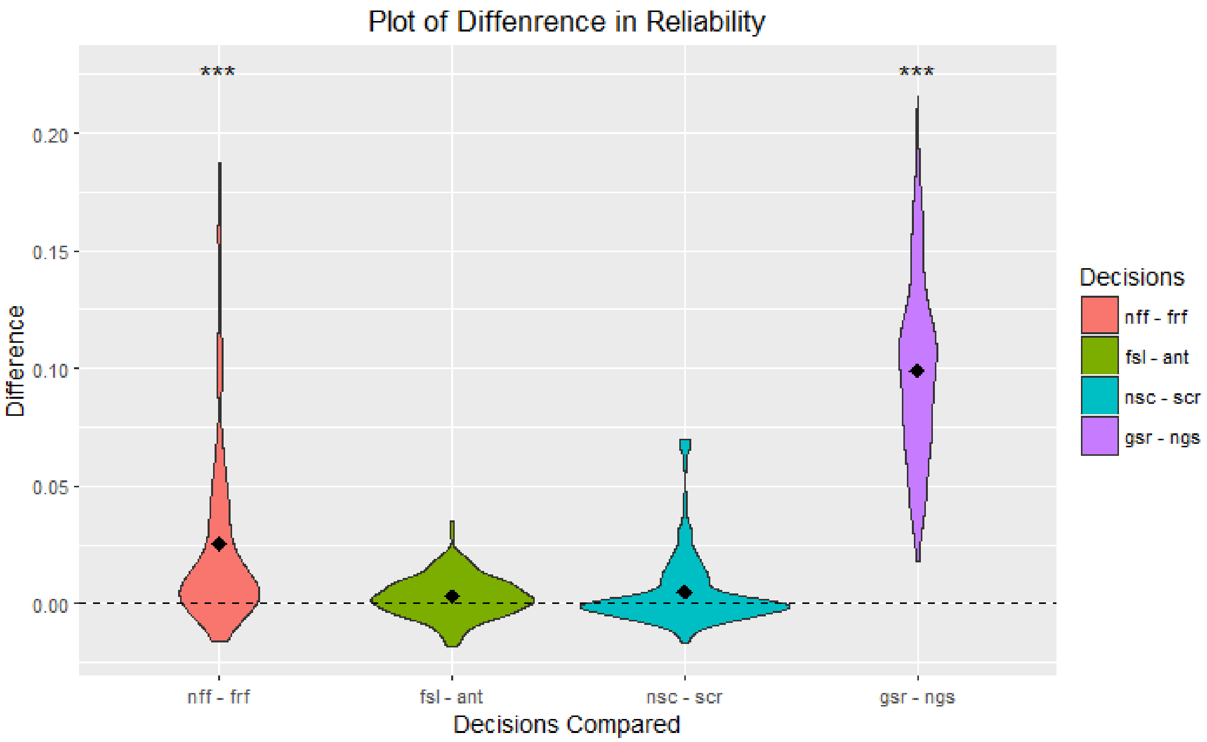

Figure 7: **Paired difference in discriminability of decisions.** Difference in discriminability for each decision in pipeline is compared by fixing other decisions and data sets. nff and gsr are statistical significantly better than the alternatives. fsl and nsc are not significantly better than the alternatives.

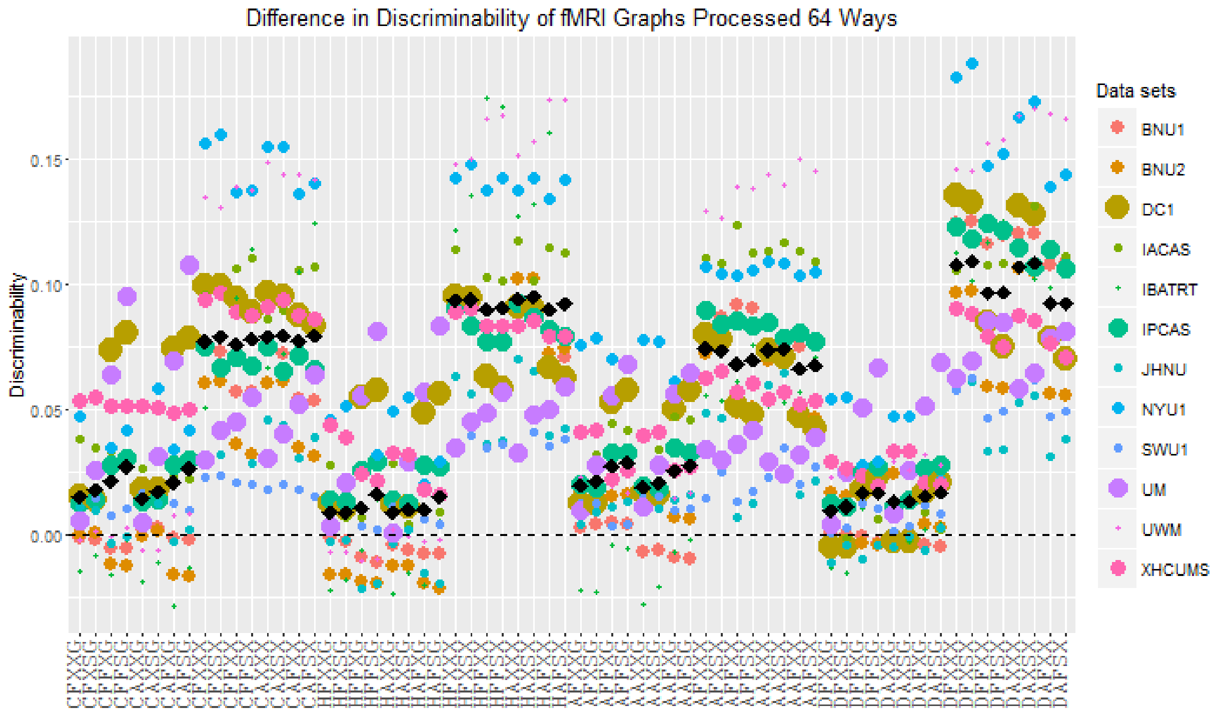

Figure 8: **Paired difference in discriminability between rank and raw graphs.** Difference in discriminability for rank and raw fmri graphs are computed for 12 data sets processed using 64 pipelines. Rank fmri graphs are more discriminable than raw fmri graphs.

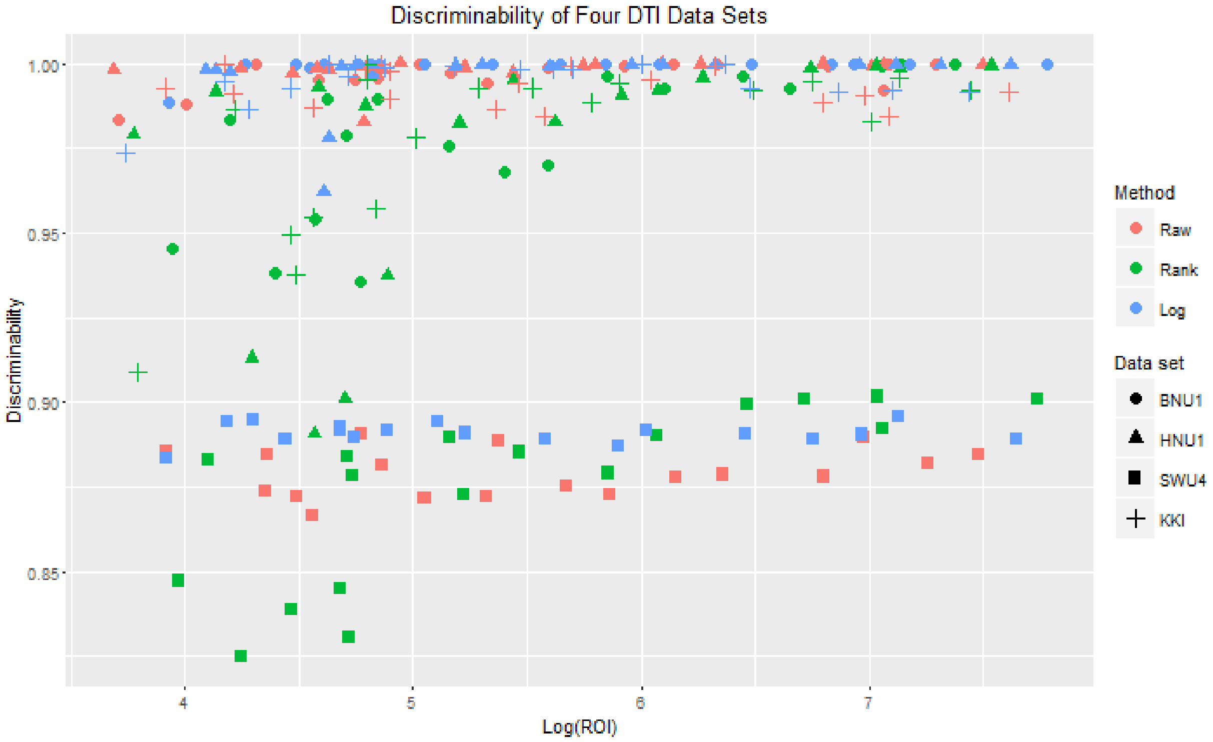

Figure 9: **Discriminability of 15 atlases.** Discriminability of SWU4 DTI registered with 15 atlases are computed. Raw, rank and log edges weights are considered. Atlases with a larger number of ROIs tend to be more discriminable.

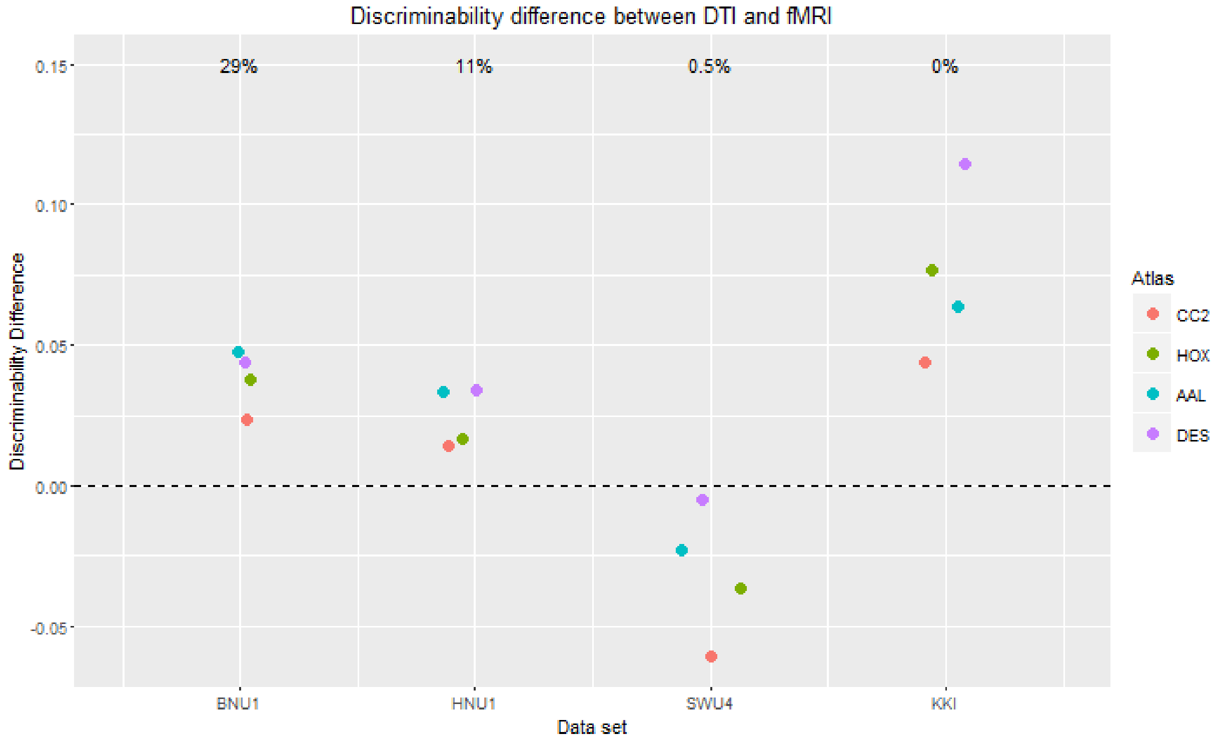

Figure 10: **Paired difference in discriminability between dti and fmri data sets.** Discriminability of DTI and fMRI graphs are computed for BNU1, HNU1, SWU4 and KKI data set. The number at the top indicates the percentage of outliers in DTI data sets. After removing outliers, DTI data sets are more discriminable than fMRI data sets.