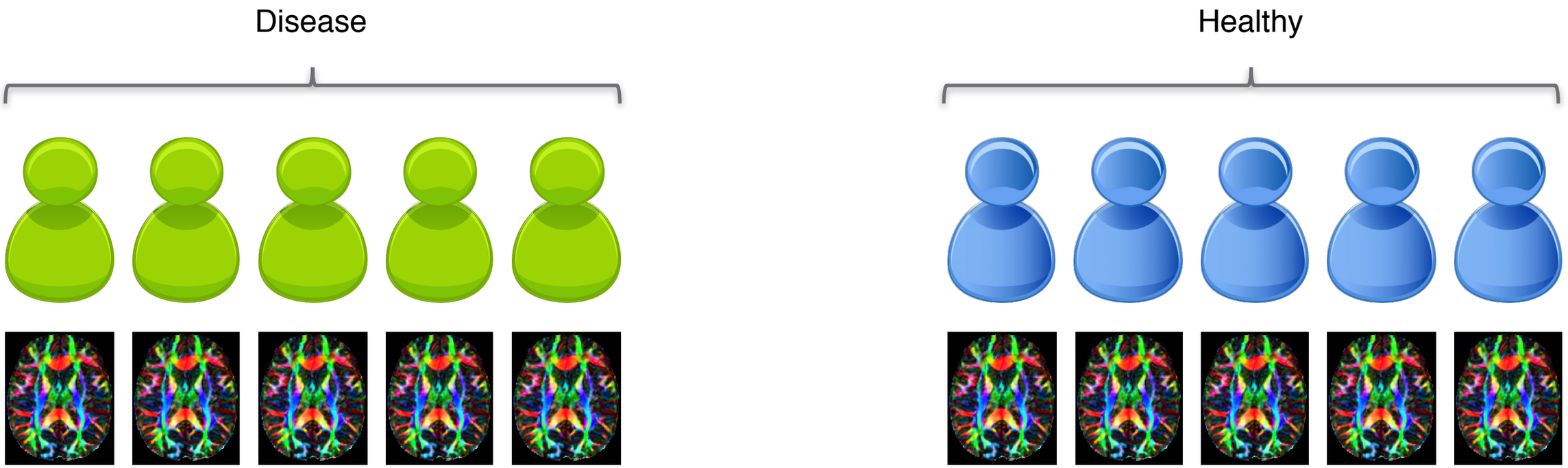
Riemannian Variance Filtering: An Independent Filtering Scheme for Statistical Tests on Manifold-valued Data

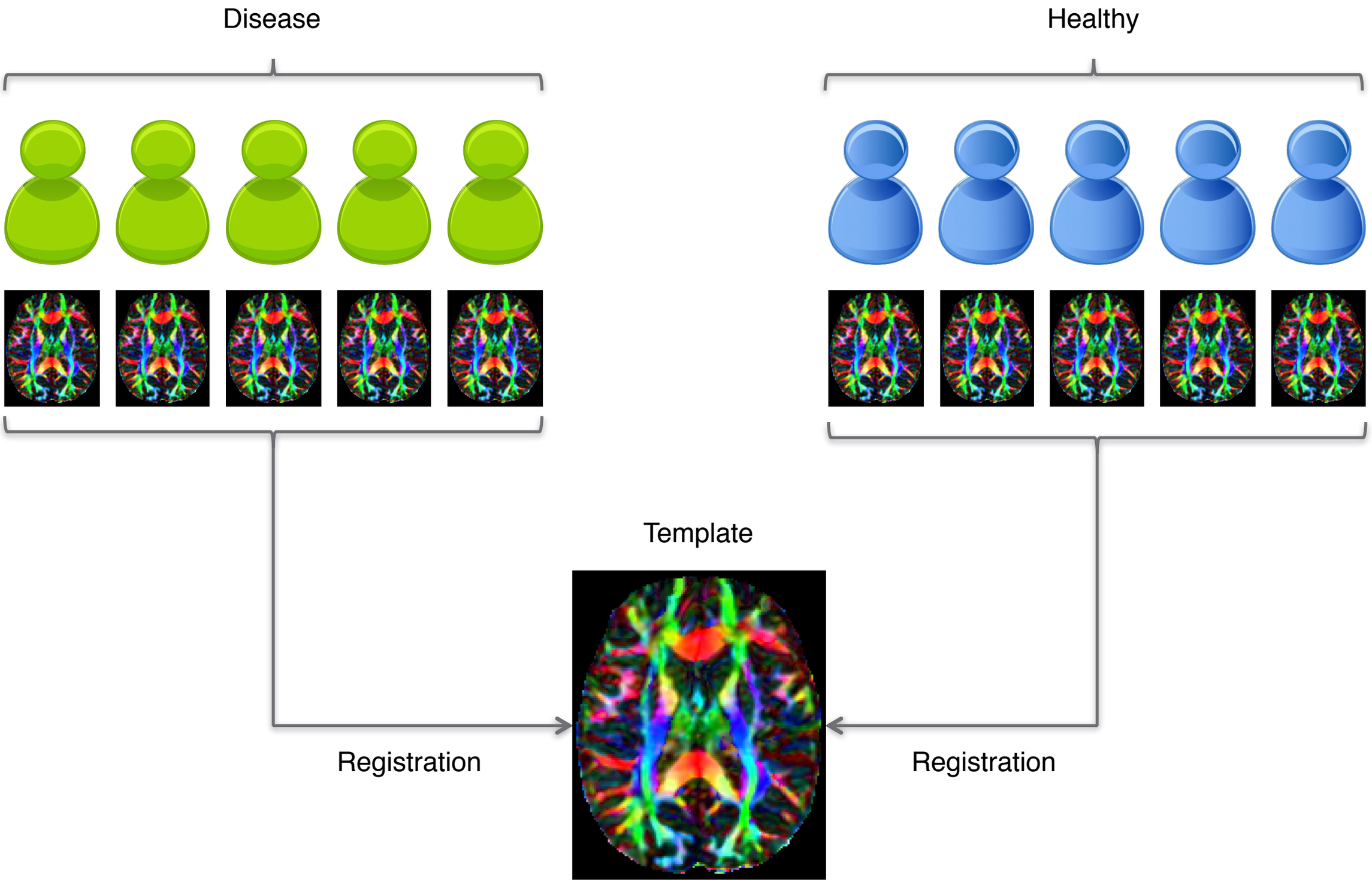
2017.7.21, Honolulu, Hawaii

Ligang Zheng, Hyunwoo J. Kim, Nagesh Adluru
Michael A.Newton, Vikas Singh

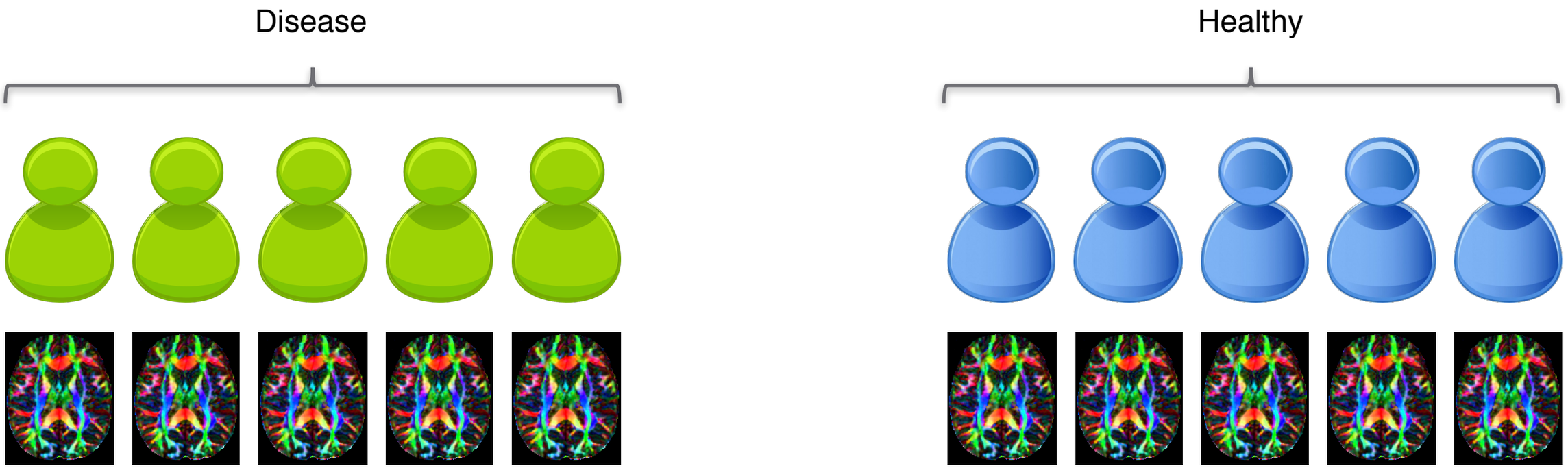
Motivating problem

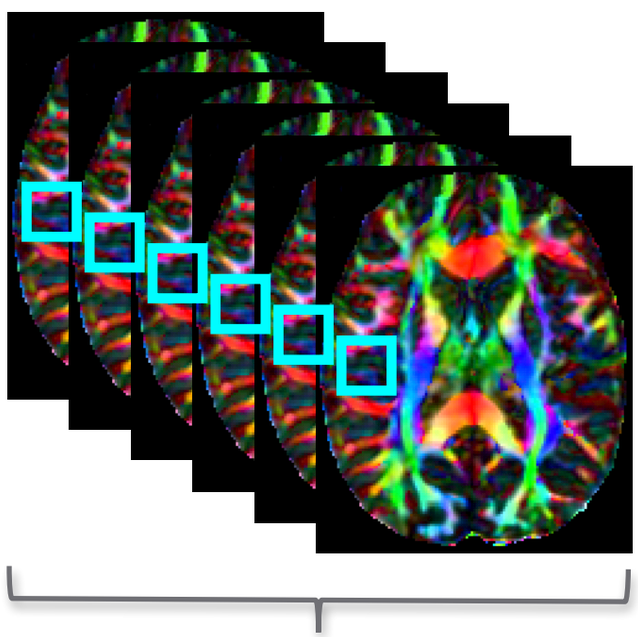


Motivating problem



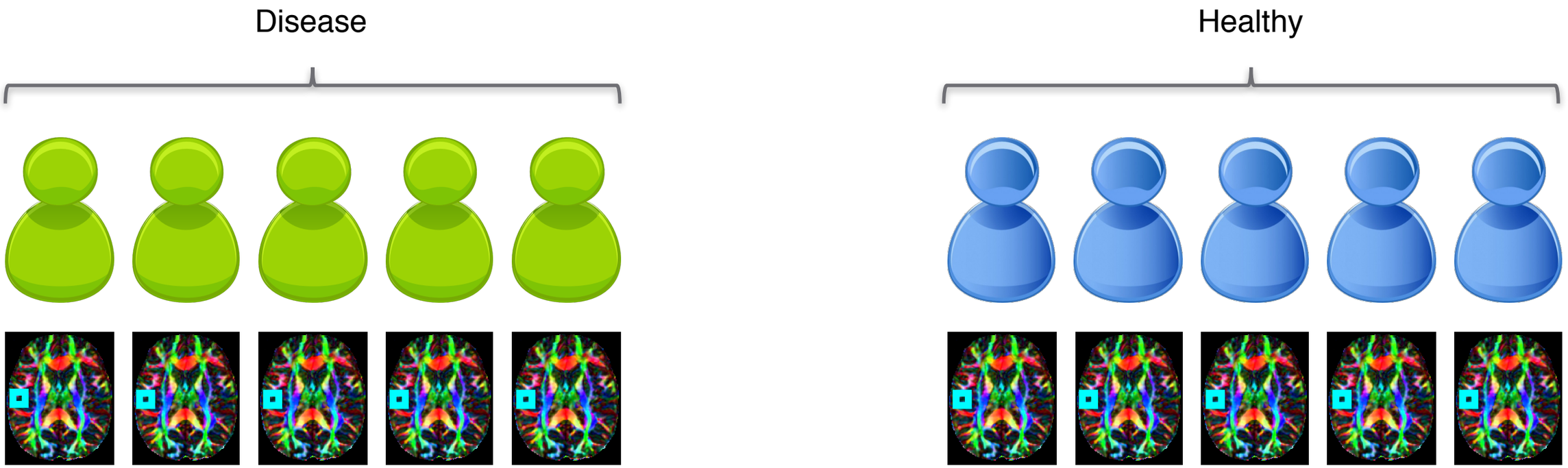
Motivating problem

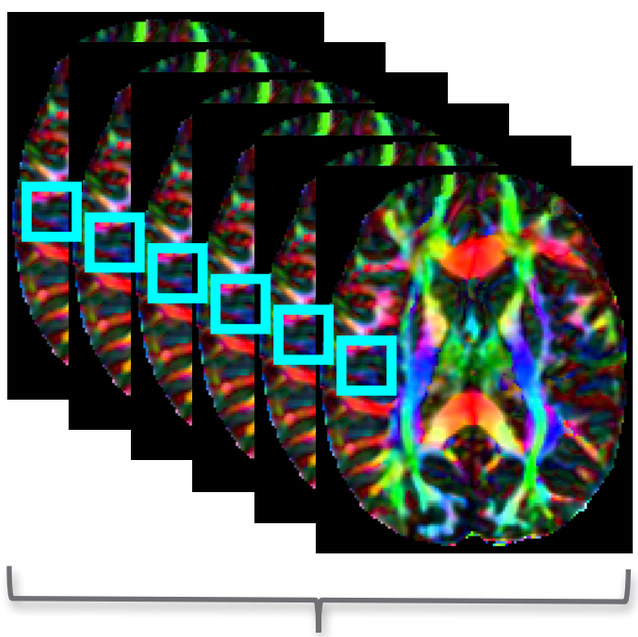




All subjects

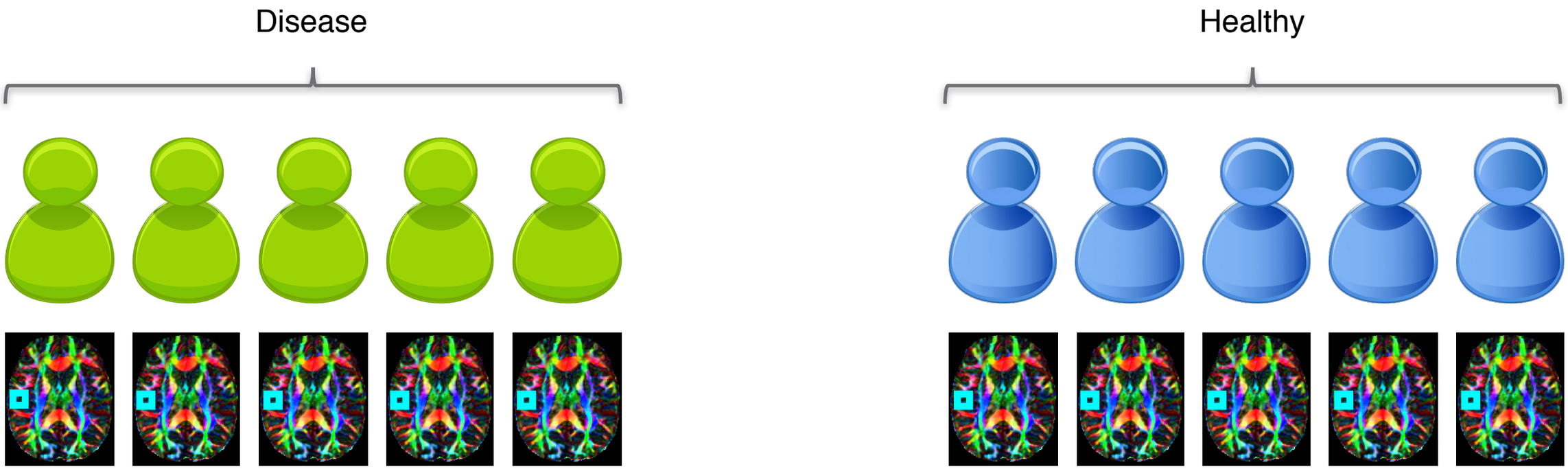
Motivating problem

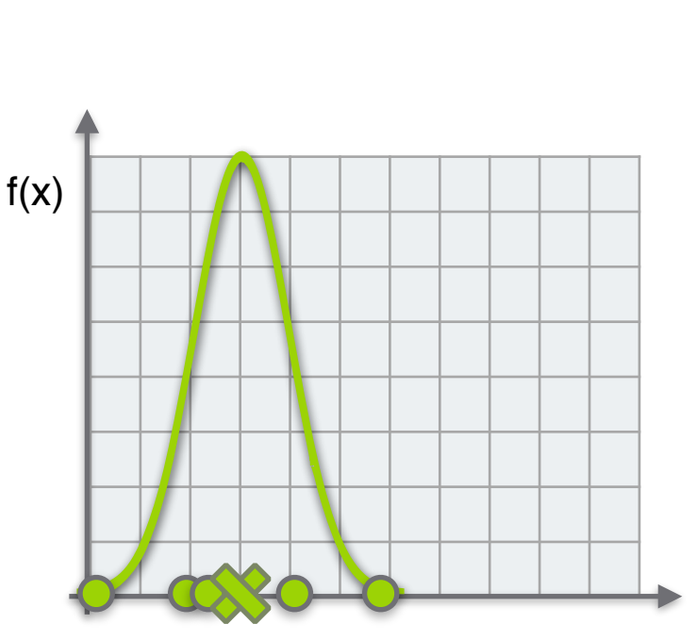


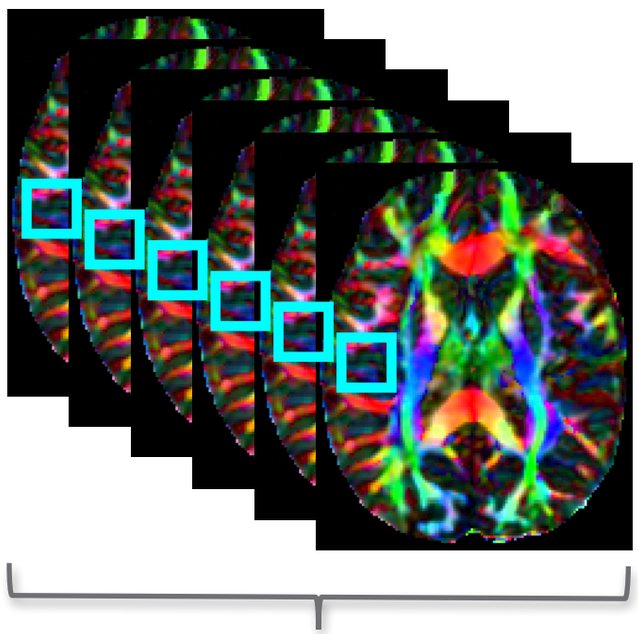


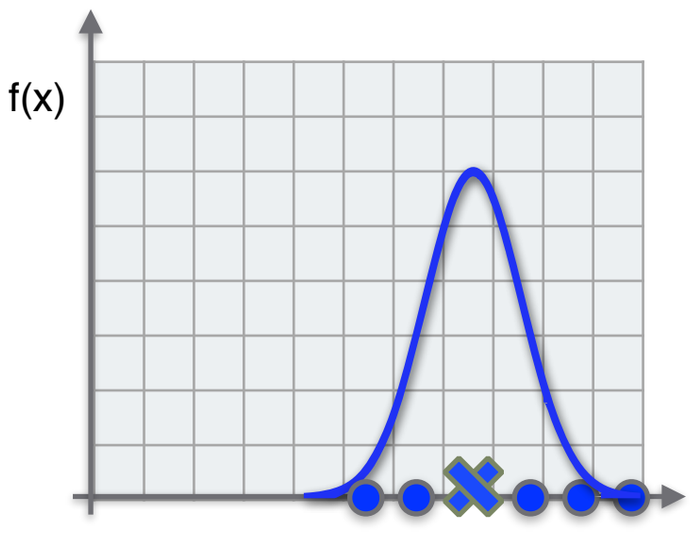
All subjects

Motivating problem

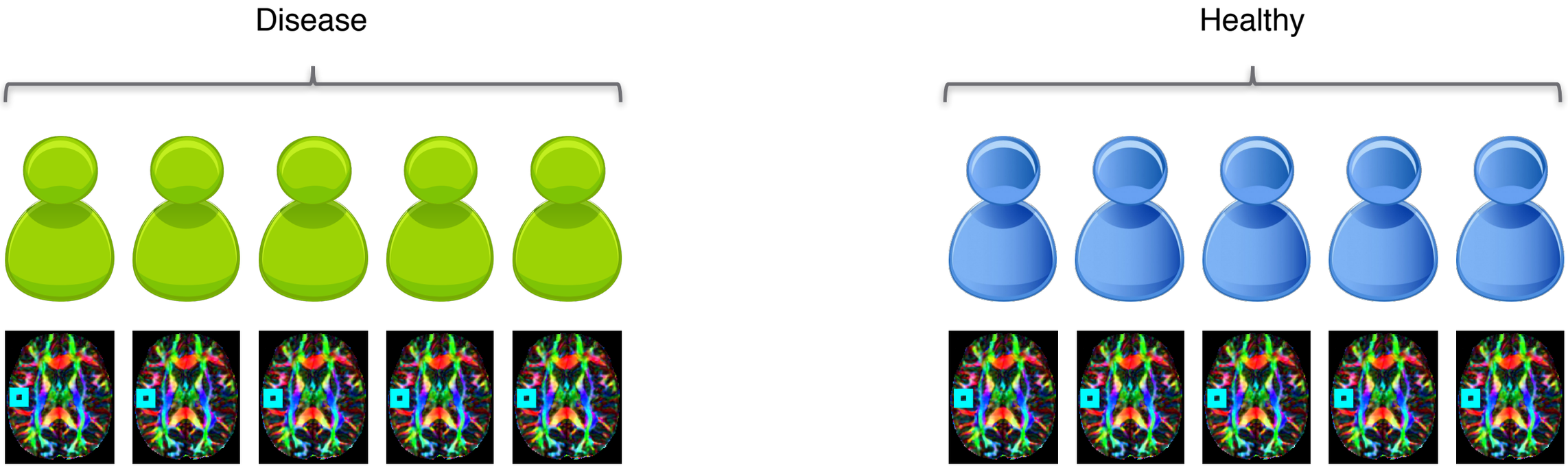


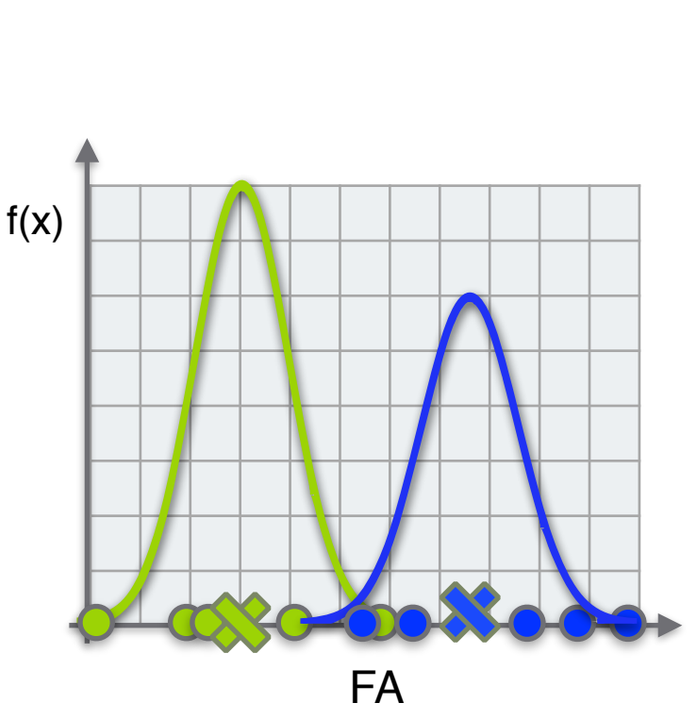


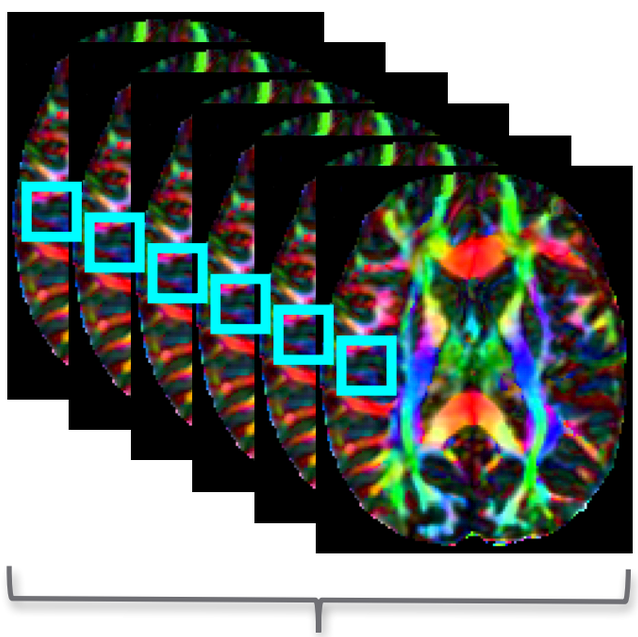


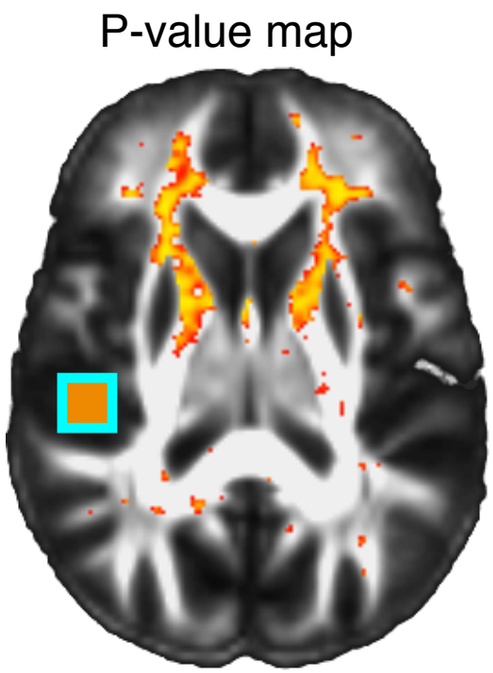


Motivating problem

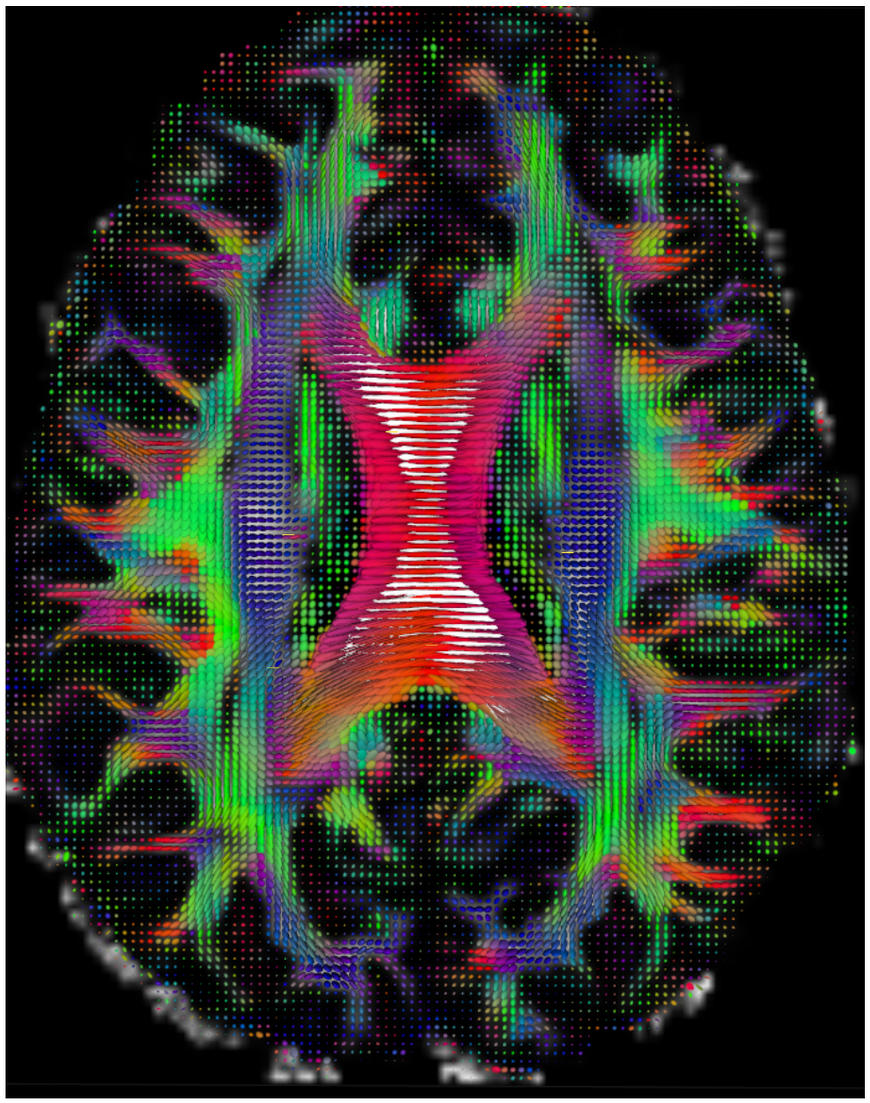








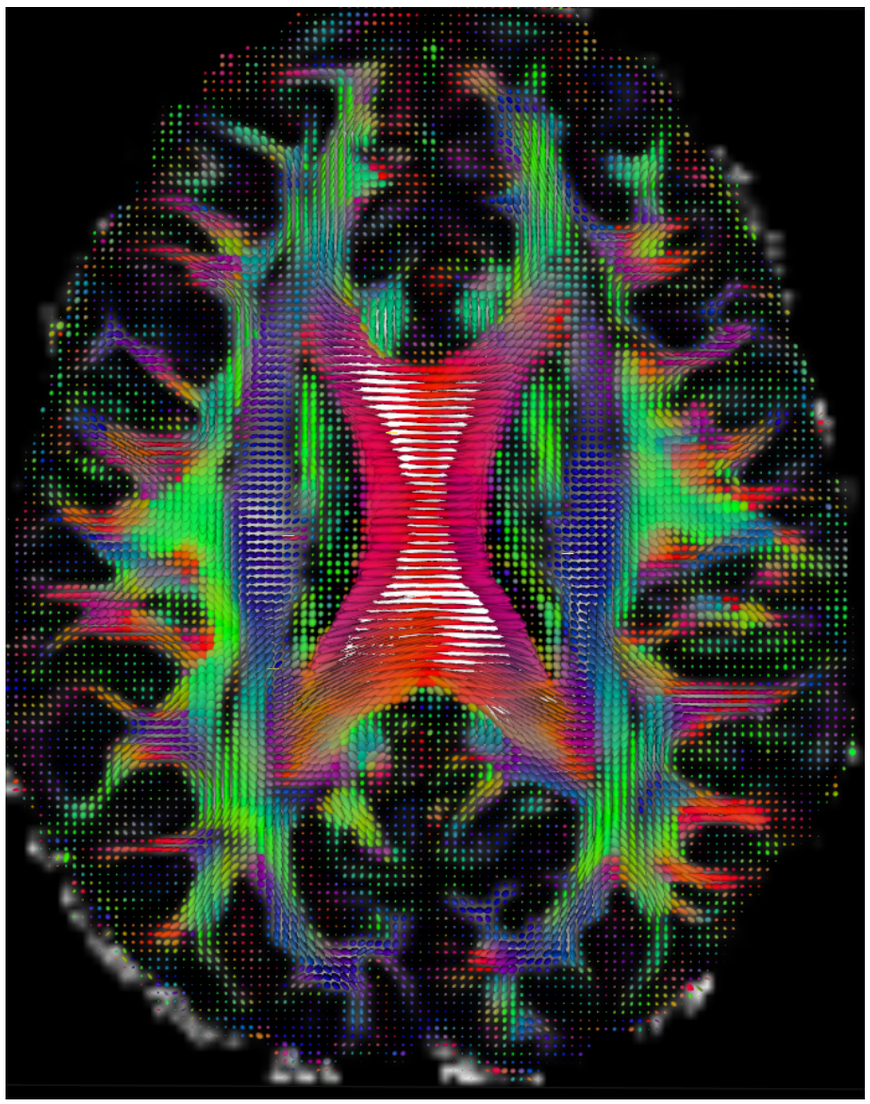
Multiple comparison problem



Hypothesis test at $\alpha = 0.05$

5 % chance of mistake

Multiple comparison problem



Hypothesis test at $\alpha = 0.05$

5 % chance of mistake

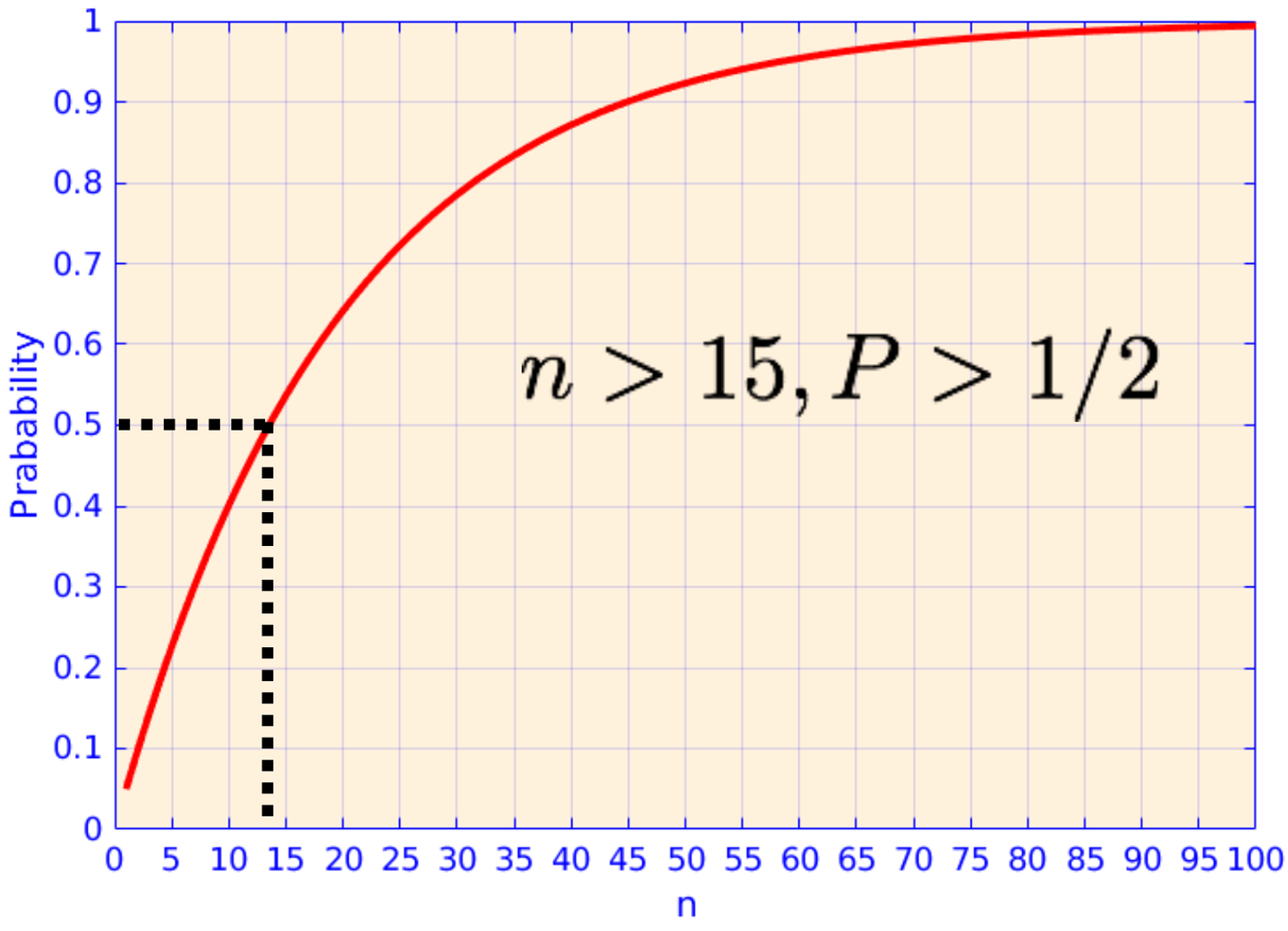
# of errors =

154 x 180 x 154 x 5% ≈ **200 k**

Multiple comparison problem

$$P(\text{at least one significant result}) = 1 - P(\text{no significant results})$$

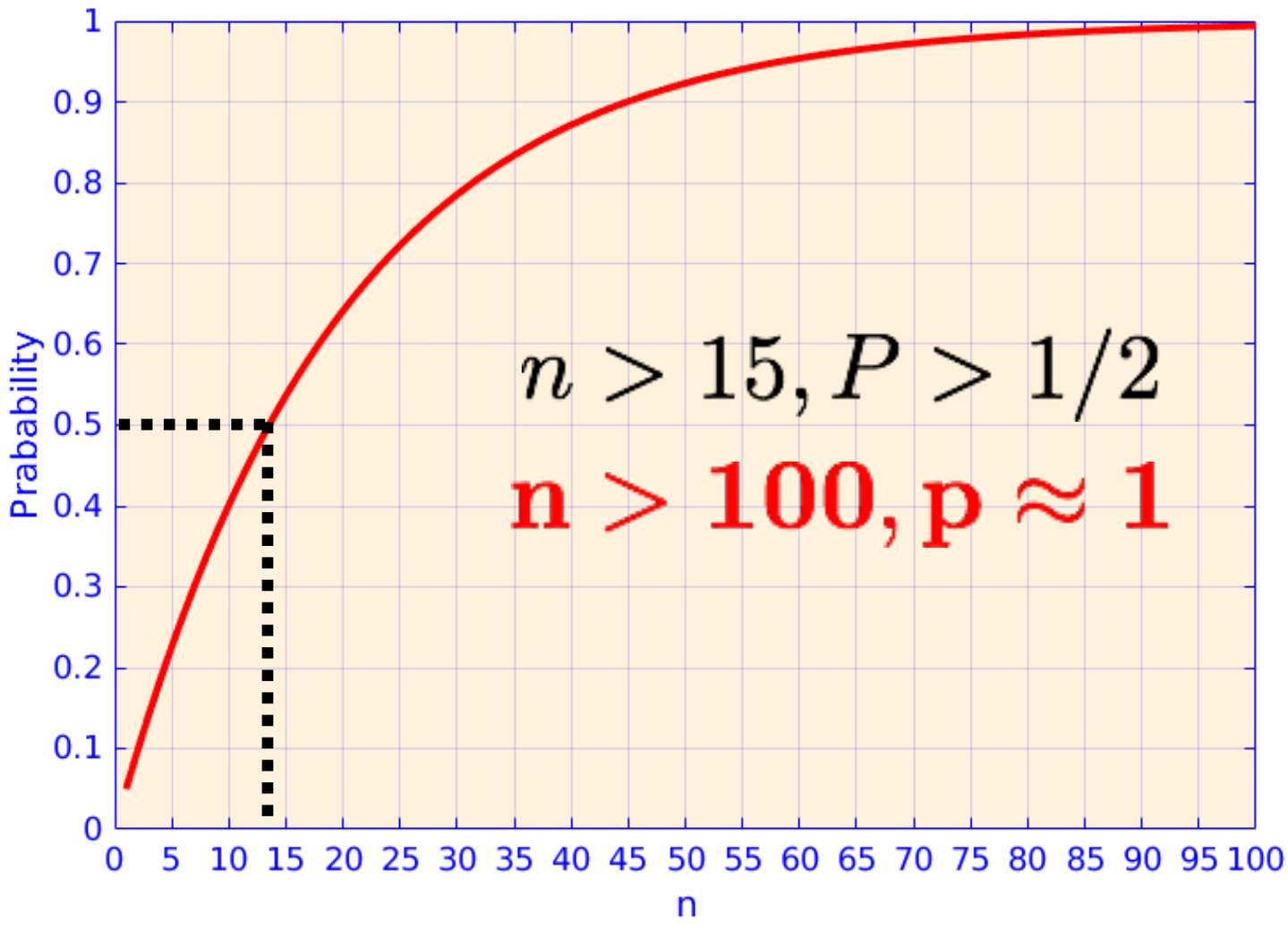
$$= 1 - (1 - 0.05)^n$$



Multiple comparison problem

$$P(\text{at least one significant result}) = 1 - P(\text{no significant results})$$

$$= 1 - (1 - 0.05)^n$$



Multiple correction

- Family-Wise Error Rate (FWER)

$$P(V \geq 0) \leq \alpha$$

- False discovery rate (FDR)

$$FDR = \mathbb{E}[V/R] \leq \alpha$$

Reduce statistical power (recall) !!

Bonferroni correction (FWER)

Controls Type I Error

$$P(V \geq 0) \leq \alpha$$

When there are n comparisons:

$$p_i \leq \frac{\alpha}{n}$$

Reduce the # tests, can achieve more rejections

Bonferroni correction (FWER)

Controls Type I Error

$$P(V \geq 0) \leq \alpha$$

When there are n comparisons:

$$p_i \leq \frac{\alpha}{n}$$

Reduce the # tests, can achieve more rejections

Bonferroni correction (FWER)

Controls Type I Error

$$P(V \geq 0) \leq \alpha$$

When there are n comparisons:

$$p_i \leq \frac{\alpha}{n}$$

Reduce the # tests, can achieve more rejections

Bonferroni correction (FWER)

Controls Type I Error

$$P(V \geq 0) \leq \alpha$$

When there are n comparisons:

$$p_i \leq \frac{\alpha}{n}$$

Reduce the # tests, can achieve more rejections

Bonferroni correction (FWER)

Controls Type I Error

$$P(V \geq 0) \leq \alpha$$

When there are n comparisons:

$$p_i \leq \frac{\alpha}{n}$$

Reduce the # tests, can achieve more rejections

Benjamini & Hochberg (FDR)

$$FDR = \mathbb{E}[Q] \ (Q = V/R)$$

Step 1: order the unadjusted p-value $p_1 \leq p_2 \leq \cdots \leq p_n$

Step 2: find the test with highest rank j, for which

$$p_i \leq \alpha \times \frac{i}{n}$$

Step 3: declare the tests of rank from 1 to j as significant

Reduce the # tests, can achieve more rejections

Benjamini & Hochberg (FDR)

$$FDR = \mathbb{E}[Q] \ (Q = V/R)$$

Step 1: order the unadjusted p-value $p_1 \leq p_2 \leq \cdots \leq p_n$

Step 2: find the test with highest rank j, for which

$$p_i \leq \alpha \times \frac{i}{n}$$

Step 3: declare the tests of rank from 1 to j as significant

Reduce the # tests, can achieve more rejections

Benjamini & Hochberg (FDR)

$$FDR = \mathbb{E}[Q] \ (Q = V/R)$$

Step 1: order the unadjusted p-value $p_1 \leq p_2 \leq \cdots \leq p_n$

Step 2: find the test with highest rank j, for which

$$p_i \leq \alpha \times \frac{i}{n}$$

Step 3: declare the tests of rank from 1 to j as significant

Reduce the # tests, can achieve more rejections

Benjamini & Hochberg (FDR)

$$FDR = \mathbb{E}[Q] \ (Q = V/R)$$

Step 1: order the unadjusted p-value $p_1 \leq p_2 \leq \cdots \leq p_n$

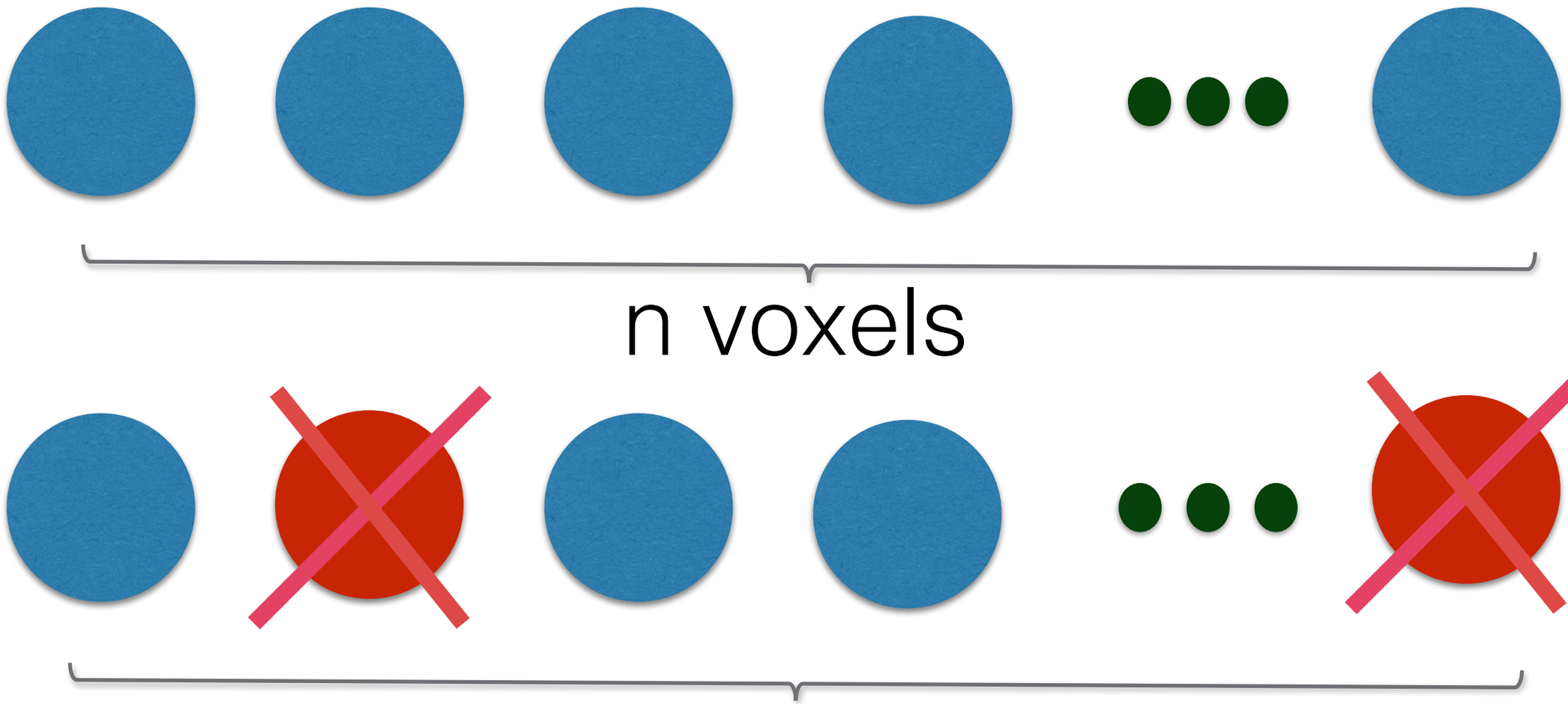
Step 2: find the test with highest rank j, for which

$$p_i \leq \alpha \times \frac{i}{n}$$

Step 3: declare the tests of rank from 1 to j as significant

Reduce the # tests, can achieve more rejections

Filtering



$n' < n$

Filtering

Improve the statistic power

n voxels

Control the type I error!

n' voxels

n' < n

Filtering methods

Random Filtering

Mask Filtering

Independent Filtering

Filtering methods

Random Filtering

Mask Filtering

Independent Filtering

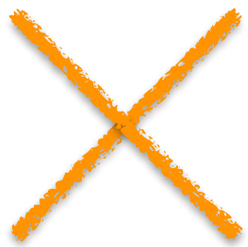
Filtering methods

Random Filtering

Mask Filtering

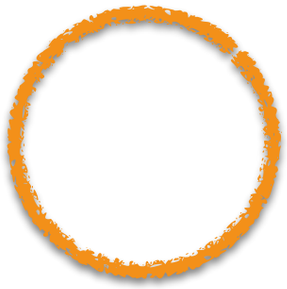
Independent Filtering

Filtering methods



Random Filtering

Mask Filtering



Independent Filtering

Independent filtering

Independent filtering increases detection power for high-throughput experiments

Richard Bourgon[a], Robert Gentleman[b], and Wolfgang Huber[c,1]

[a]European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; [b]Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080-4990; and [c]European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 22, 2010 (received for review December 3, 2009)

With high-dimensional data, variable-by-variable statistical testing is often used to select variables whose behavior differs across conditions. Such an approach requires adjustment for multiple testing, which can result in low statistical power. A two-stage approach that first filters variables by a criterion independent of the test statistic, and then only tests variables which pass the filter, can provide higher power. We show that use of some filter/test statistics pairs presented in the literature may, however, lead to loss of type I error control. We describe other pairs which avoid this problem. In an application to microarray data, we found that gene-by-gene filtering by overall variance followed by a *t*-test increased the number of discoveries by 50%. We also show that this particular statistic pair induces a lower bound on fold-change among the set of discoveries. Independent filtering—using filter/test pairs that are independent under the null hypothesis but correlated under the alternative—is a general approach that can substantially increase the efficiency of experiments.

gene expression | multiple testing

In many experimental contexts which generate high-dimensional data, variable-by-variable statistical testing is used to select variables whose behavior differs across the set of studied conditions. Each variable is associated with a null hypothesis which asserts that behavior for that variable does not differ across conditions. A null hypothesis is rejected when observed data, summarized into a per-variable *p*-value, are deemed to be inconsistent with the hypothesis. In biology, for example, microarrays or high-throughput sequencing may be used to identify genes (variables) whose expression level shows systematic covariation with a treat-

few dozen or hundred. As a consequence, the power of an experiment to detect a given differentially expressed gene could potentially be quite low.

In the microarray literature, several authors have suggested *filtering* to reduce the impact that multiple testing adjustment has on detection power (7–12). Conceptually similar screening approaches have also been proposed for variable selection in high-dimensional regression models (13, 14). In filtering for microarray applications, the data are first used to identify and remove a set of genes which seem to generate uninformative signal. Second, formal statistical testing is applied only to genes which pass the filter. An effective filter will enrich for true differential expression while simultaneously reducing the number of hypotheses tested at stage two—making multiple testing adjustment less severe. Such filtering is further motivated by the observation that the set of genes which are not differentially expressed can be partitioned into two groups: (*i*) genes that are not expressed in any of the conditions of the experiment or whose reporters on the array lack sensitivity to detect their expression; and (*ii*) genes that are expressed and detectable, but not differentially expressed across conditions.

This two-stage approach, the use of which need not be restricted to gene expression applications, assesses each variable on the basis of both a filter statistic (U^I) and a test statistic (U^{II}). Both statistics are required to exceed their respective cutoffs. Note, however, that the two-stage approach is not equivalent to standard hypothesis testing based on the joint distribution of the filter and test statistics: the latter uses a joint null distribution to compute type I error rate, while the former only considers the null distribution of the stage-two test statistic.

Independent filtering

PNAS

Independent filtering increases detection power for high-throughput experiments

Richard Bourgon[a], Robert Gentleman[b], and Wolfgang Huber[c,1]

[a]European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; [b]Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080-4990; and [c]European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 22, 2010 (received for review December 3, 2009)

With high-dimensional data, variable-by-variable statistical testing is often used to select variables whose behavior differs across conditions. Such an approach requires adjustment for multiple testing, which can result in low statistical power. A two-stage approach that first filters variables by a criterion independent of the test statistic, and then only tes… provide higher power. We s… tics pairs presented in the li… type I error control. We d… problem. In an application… gene-by-gene filtering by … increased the number of di… this particular statistic pair induces a lower bound on fold-change among the set of discoveries. Independent filtering—using filter/test pairs that are independent under the null hypothesis but correlated under the alternative—is a general approach that can substantially increase the efficiency of experiments.

gene expression | multiple testing

In many experimental contexts which generate high-dimensional data, variable-by-variable statistical testing is used to select variables whose behavior differs across the set of studied conditions. Each variable is associated with a null hypothesis which asserts that behavior for that variable does not differ across conditions. A null hypothesis is rejected when observed data, summarized into a per-variable *p*-value, are deemed to be inconsistent with the hypothesis. In biology, for example, microarrays or high-throughput sequencing may be used to identify genes (variables) whose expression level shows systematic covariation with a treat-

few dozen or hundred. As a consequence, the power of an experiment to detect a given differentially expressed gene could potentially be quite low.

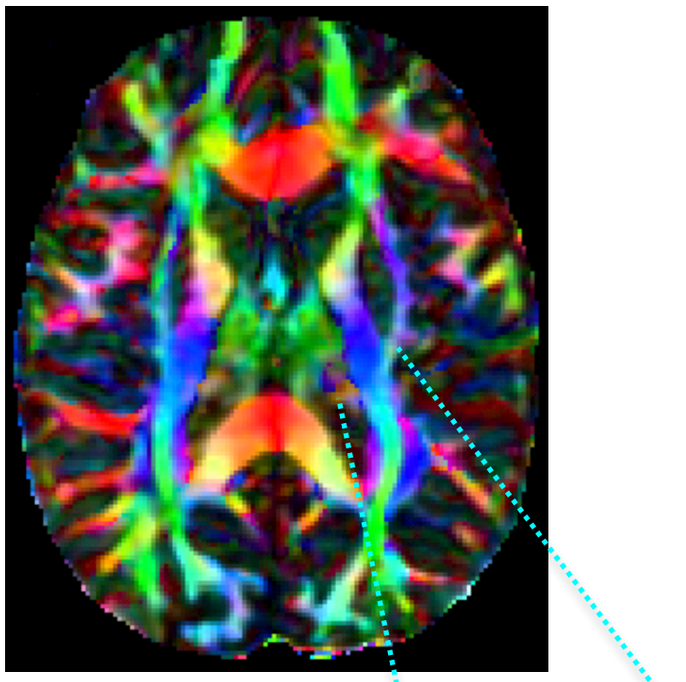
In the microarray literature, several authors have suggested *filtering* to reduce the impact that multiple testing adjustment … Conceptually similar screening … osed for variable selection in … dels (13, 14). In filtering for … are first used to identify and … em to generate uninformative … esting is applied only to genes … filter will enrich for true differential expression while simultaneously reducing the number of hypotheses tested at stage two—making multiple testing adjustment less severe. Such filtering is further motivated by the observation that the set of genes which are not differentially expressed can be partitioned into two groups: (*i*) genes that are not expressed in any of the conditions of the experiment or whose reporters on the array lack sensitivity to detect their expression; and (*ii*) genes that are expressed and detectable, but not differentially expressed across conditions.

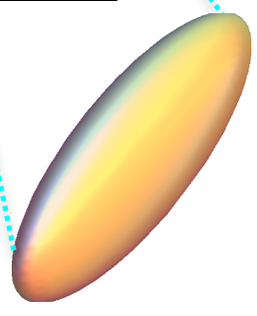
This two-stage approach, the use of which need not be restricted to gene expression applications, assesses each variable on the basis of both a filter statistic (U^I) and a test statistic (U^{II}). Both statistics are required to exceed their respective cutoffs. Note, however, that the two-stage approach is not equivalent to standard hypothesis testing based on the joint distribution of the filter and test statistics: the latter uses a joint null distribution to compute type I error rate, while the former only considers the null distribution of the stage-two test statistic.

Scalar variables

Manifold-valued data

DTI





NOT a vector space!

Manifold-valued data

$$D = \begin{pmatrix} 1.53 & 1.38 & 0.65 \\ 1.38 & 1.33 & 0.70 \\ 0.65 & 0.70 & 1.06 \end{pmatrix} \qquad x^T D x > 0, x \neq 0$$

Independent filtering

Two steps:

- Filter out the some variables using ***filter statistic***
- Testing on variables passing the filter using ***test statistic***

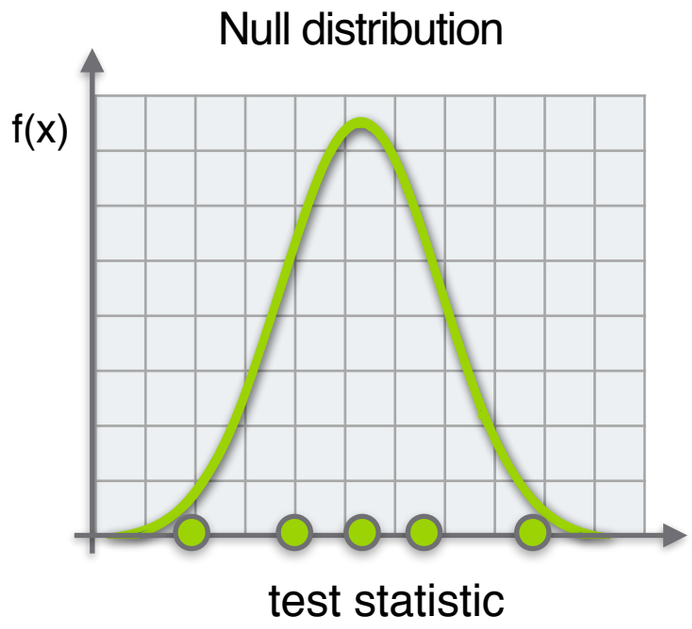
Independent filtering

Two steps:

- Filter out the some variables using ***filter statistic***
- Testing on variables passing the filter using ***test statistic***

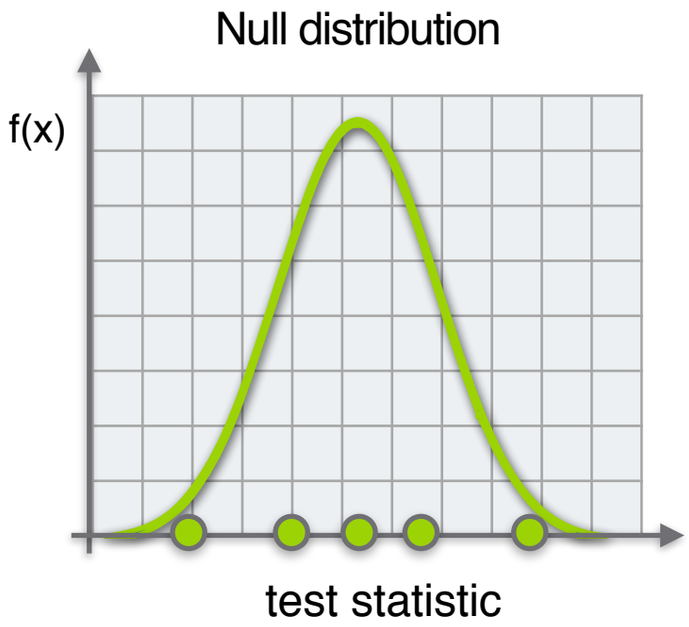
Key: filter statistic and test statistic are marginally independent. Independent under null hypothesis and dependent under alternative hypothesis.

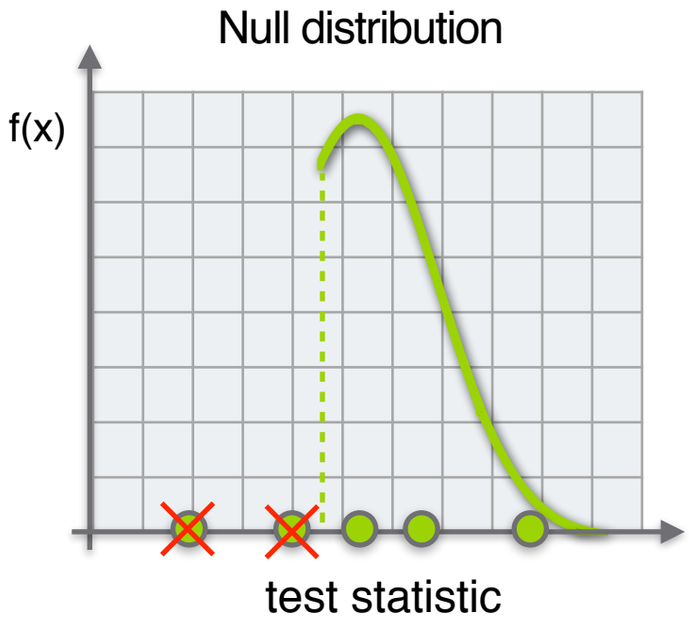
Independent filtering



voxels for hypothesis tests

Independent filtering

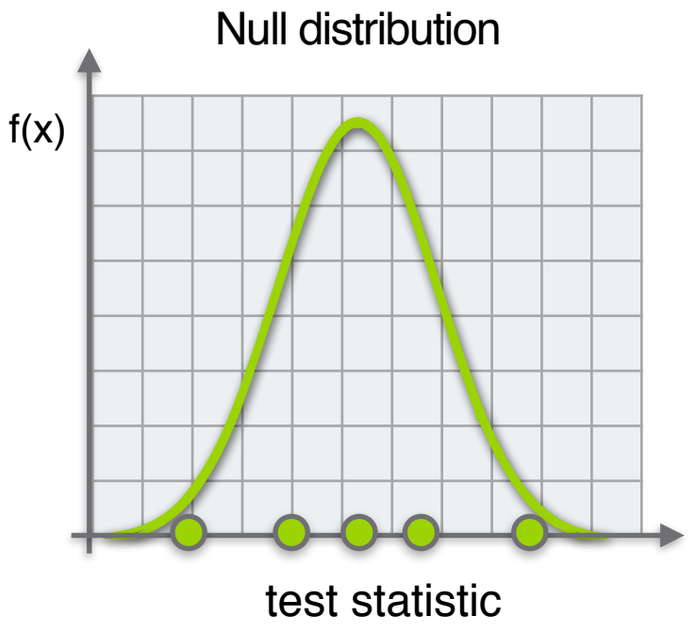


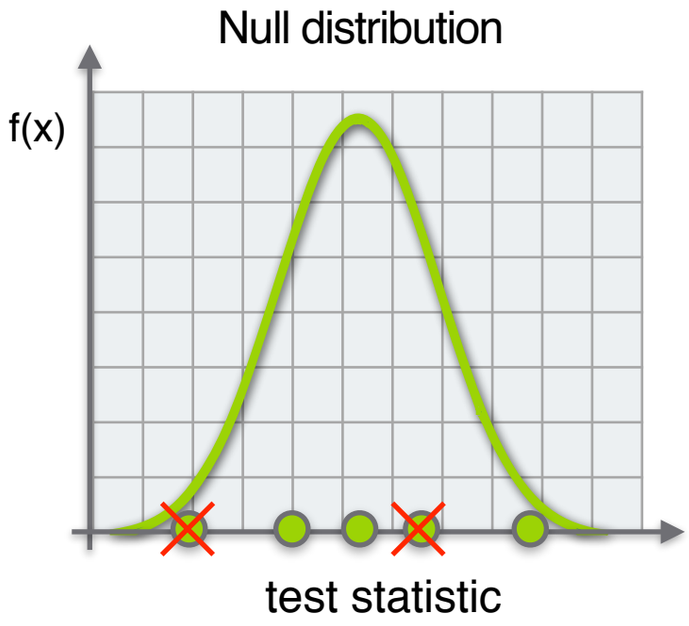


voxels for hypothesis tests

Filter by test statistic

Independent filtering

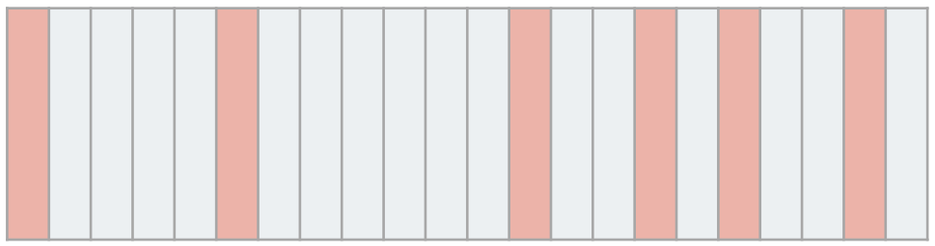




voxels for hypothesis tests

Filter by ~~test statistic~~

Independent filtering statistic



Our scheme

Riemannian Variance Filter + Riemannian manifold statistics

Our scheme

Riemannian Variance Filter

+

Riemannian manifold statistics

Riemannian Gaussian distribution

$$f(X;\mu,\sigma) = \frac{1}{\zeta(\sigma)} \exp\left(-\frac{\mathrm{d}(X,\mu)^2}{2\sigma^2}\right)$$

where

$$\zeta(\sigma) = \int_{\mathcal{M}} \exp\left(-\frac{d(X,\mu)^2}{2\sigma^2}\right) dX.$$

X is manifold-valued

d is the geodesic distance

Riemannian Gaussian distribution

$$f(X;\mu,\sigma) = \frac{1}{\zeta(\sigma)} \exp\left(-\frac{\mathrm{d}(X,\mu)^2}{2\sigma^2}\right)$$

where

$$\zeta(\sigma) = \int_{\mathcal{M}} \exp\left(-\frac{d(X,\mu)^2}{2\sigma^2}\right) dX.$$

X is manifold-valued

d is the geodesic distance

Parametric estimation

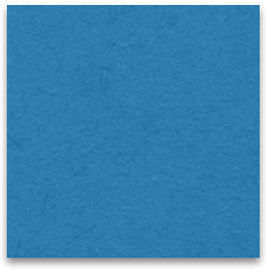
Energy function (second moment)

$$\mathcal{E}_n(\bar{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} d(\bar{X}, X_i)^2$$

MLE:

$$\sigma^3 \frac{d}{d\sigma} \log \zeta(\sigma) = \mathcal{E}_n(\overline{X})$$

The solution is:



$$\widehat{\sigma} = \phi(\mathcal{E}_n(\overline{X})) = \phi(\frac{1}{n}\sum_{i=1}^{n} d^2(\overline{X}, X_i))$$

ϕ is the inverse function of $\sigma \mapsto \sigma^3 \times \frac{d}{d\sigma}\log\zeta(\sigma)$.

[Salem Said et al. 2016]

Parametric estimation

Energy function (second moment)

$$\mathcal{E}_n(\bar{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} d(\bar{X}, X_i)^2$$

MLE:

$$\sigma^3 \frac{d}{d\sigma} \log \zeta(\sigma) = \mathcal{E}_n(\overline{X})$$

The solution is:

$$\widehat{\sigma} = \phi(\mathcal{E}_n(\overline{X})) = \phi(\frac{1}{n}\sum_{i=1}^{n} d^2(\overline{X}, X_i))$$

ϕ is the inverse function of $\sigma \mapsto \sigma^3 \times \frac{d}{d\sigma}\log\zeta(\sigma)$.
is a strictly increasing function

[Salem Said et al. 2016]

Our scheme

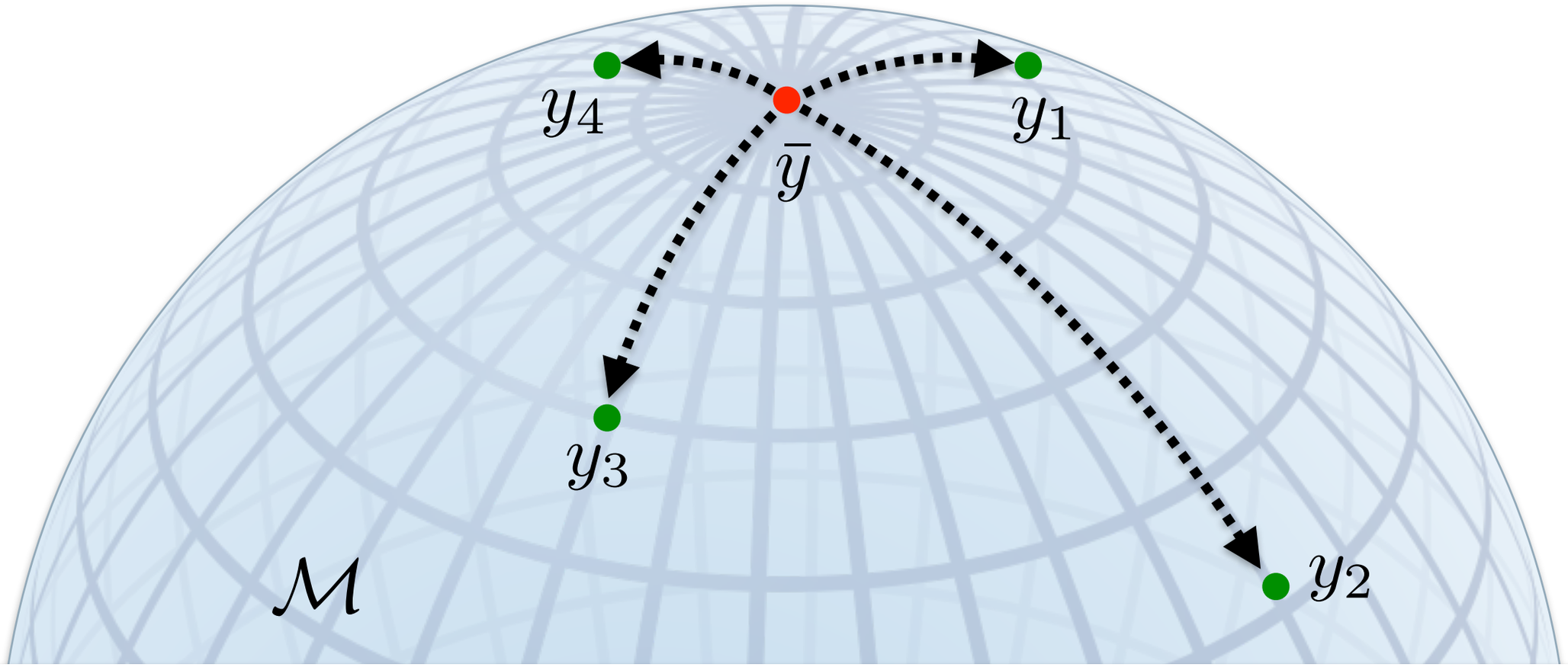
Riemannian Variance Filter

+

Riemannian manifold statistics

Intrinsic mean

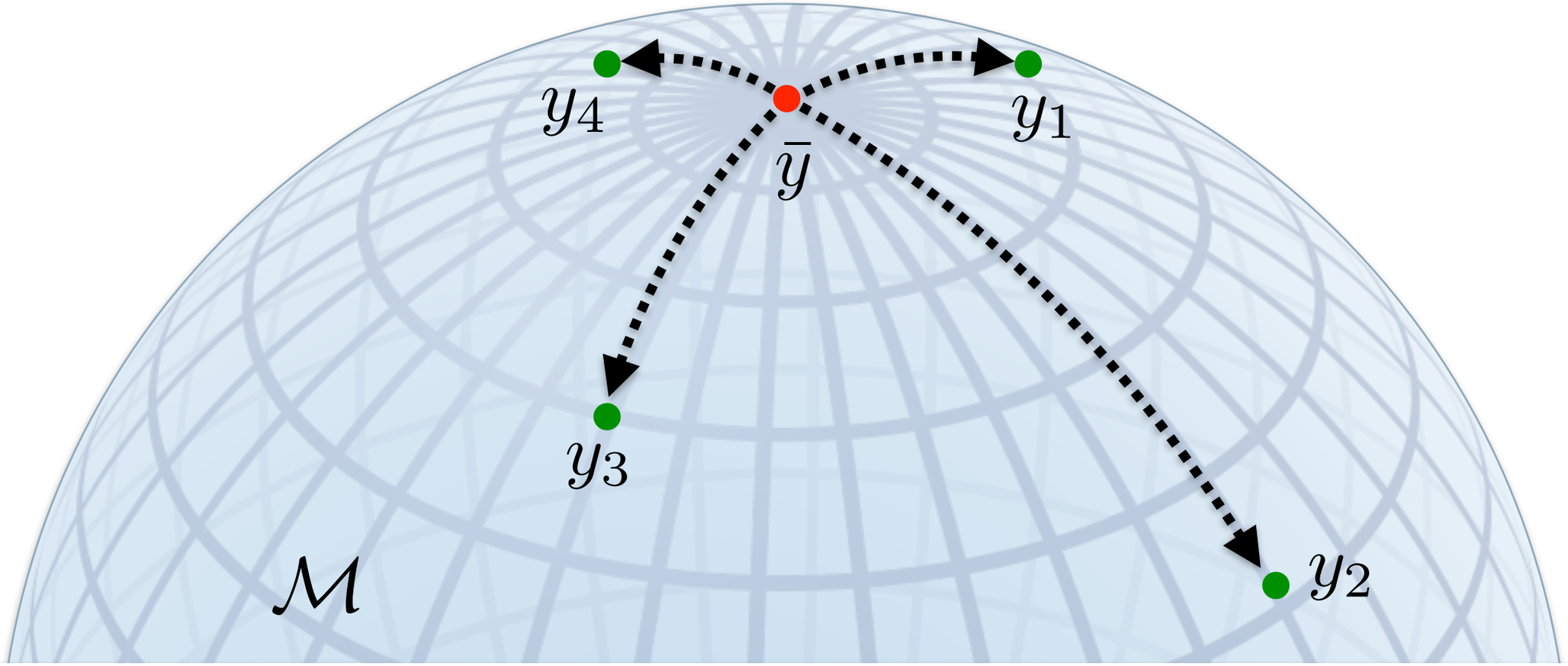
$$\underline{X} = \arg\min_{X \in \mathcal{M}} \sum_{i=1}^{N} d(X, \underline{\bar{X}_i})^2$$



Intrinsic mean

$$\bar{X} = \arg\min_{X \in \mathcal{M}} \sum_{i=1}^{N} d(X, \bar{X}_i)^2$$

$$\overline{X} \approx \exp\left(\frac{1}{n} \sum_{i=1}^{n} \log X_i\right)$$



Test statistic

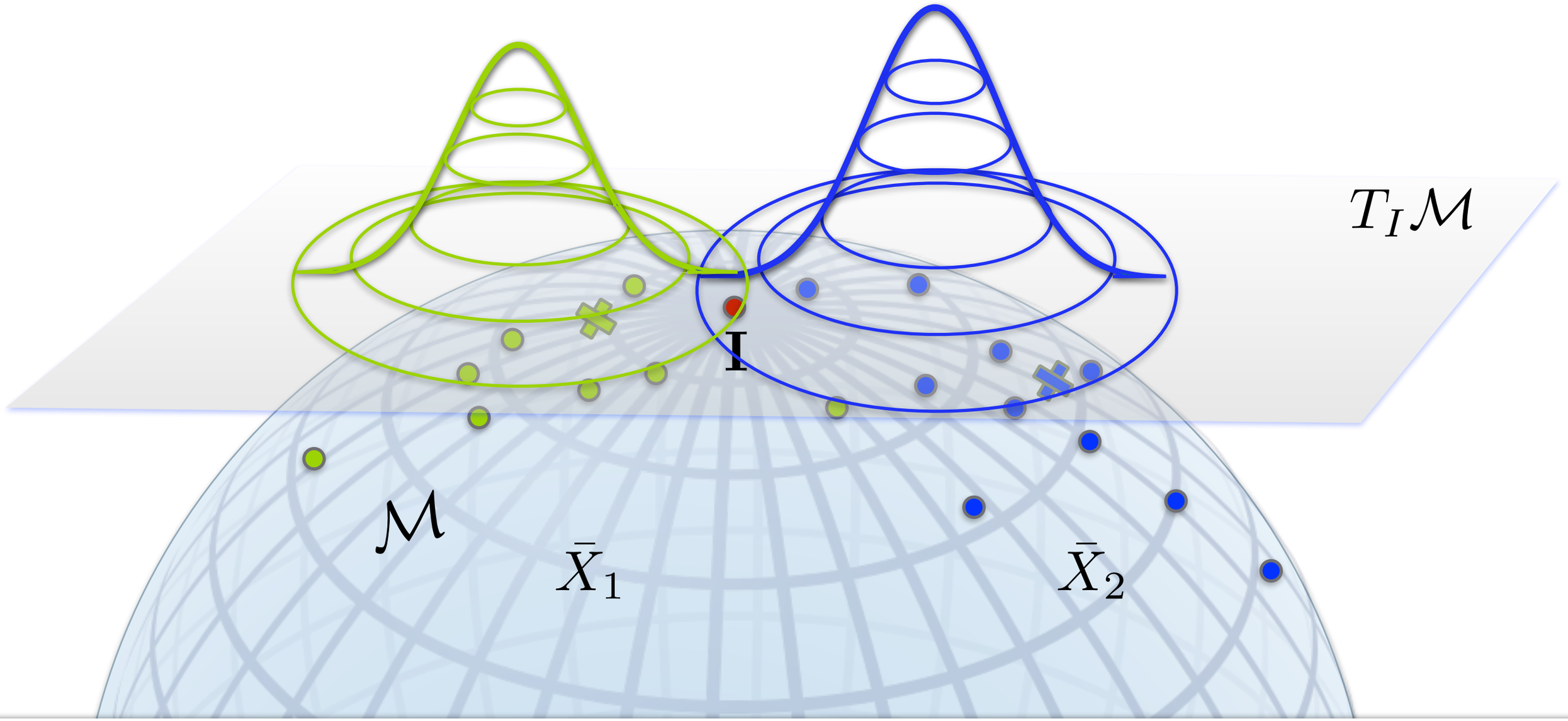
LeMean: Log-Euclidean mean-based permutation Test

$$u = d_{geo}(\bar{X}_1, \bar{X}_2)$$

Test statistic

LeMean: Log-Euclidean mean-based permutation Test

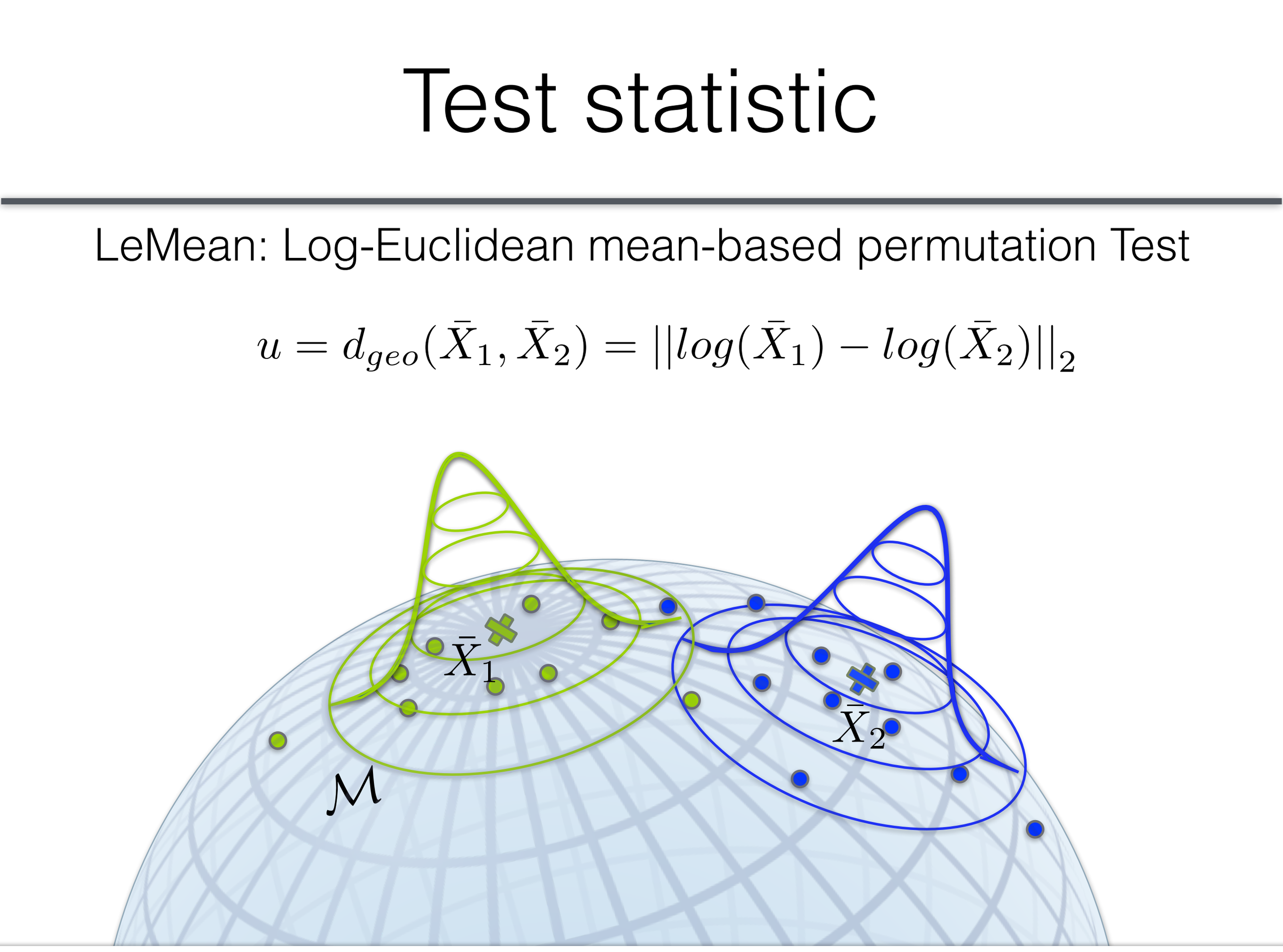
$$u = d_{geo}(\bar{X}_1, \bar{X}_2) = ||log(\bar{X}_1) - log(\bar{X}_2)||_2$$



Test statistic

LeMean: Log-Euclidean mean-based permutation Test

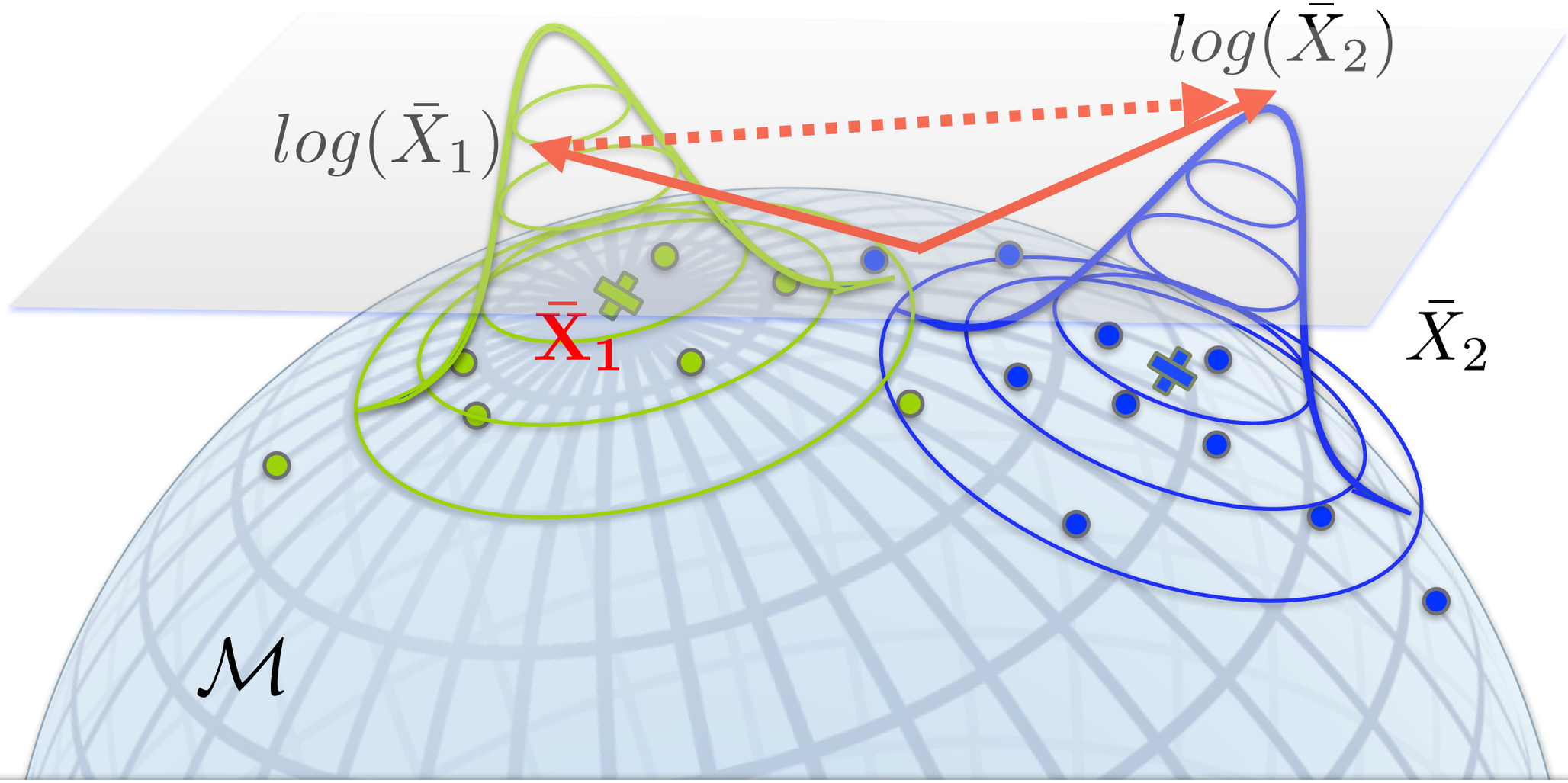
$$u = d_{geo}(\bar{X}_1, \bar{X}_2) = ||log(\bar{X}_1) - log(\bar{X}_2)||_2$$



Test statistic

LeMean: Log-Euclidean mean-based permutation Test

$$u = d_{geo}(\bar{X}_1, \bar{X}_2) = ||log(\bar{X}_1) - log(\bar{X}_2)||_2$$



Test statistic

LeMean: Log-Euclidean mean-based permutation Test

$$u = d_{geo}(\bar{X}_1, \bar{X}_2)$$

Cramer Test: with intrinsic metric or Log-Euclidean

$$\delta_{n_1,n_2} = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathrm{d}(X_i, Y_j) - \frac{1}{2n_1^2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_1} \mathrm{d}(X_i, X_{i'}) - \frac{1}{2n_2^2} \sum_{j=1}^{n_2} \sum_{j'=1}^{n_2} \mathrm{d}(Y_j, Y_{j'}) \right.$$

No distribution assumption

Experiment setup

Benchmark:

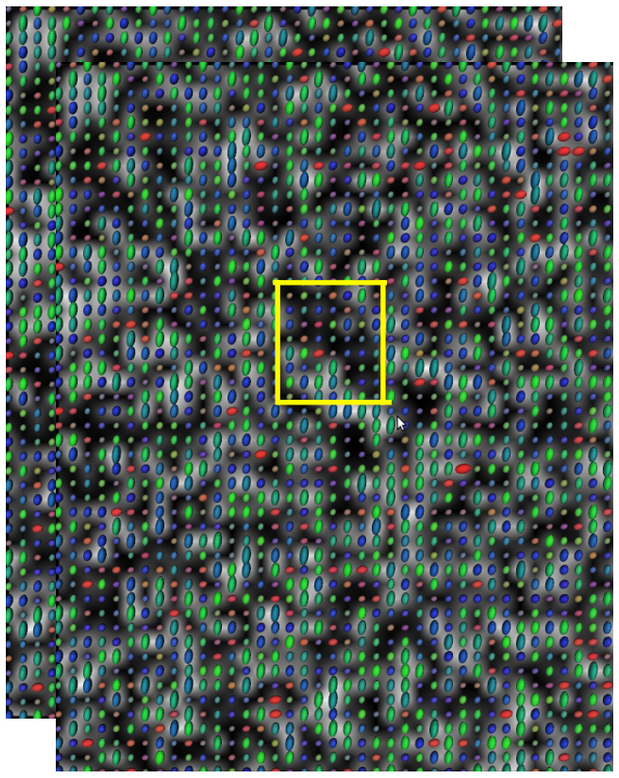
t-test + SVF

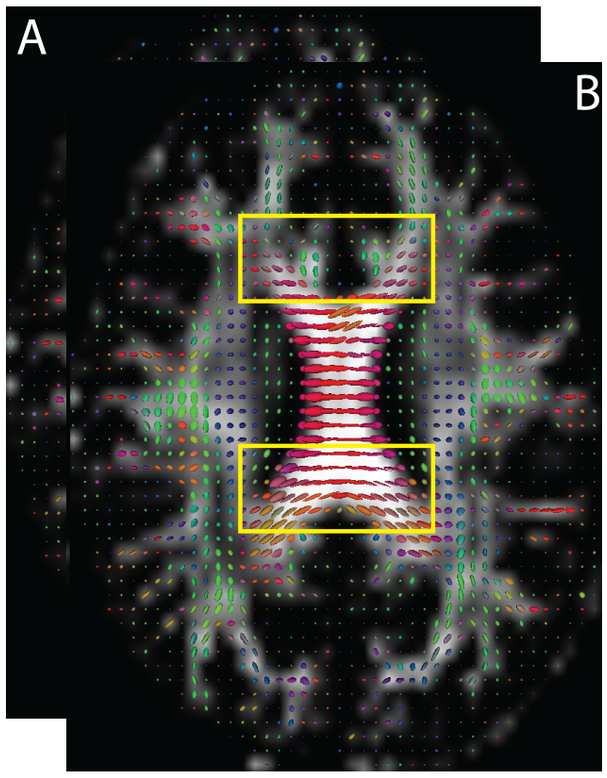
Comparison:

t-test + RVF LeMean+ RVF Cramer+ RVF

Data

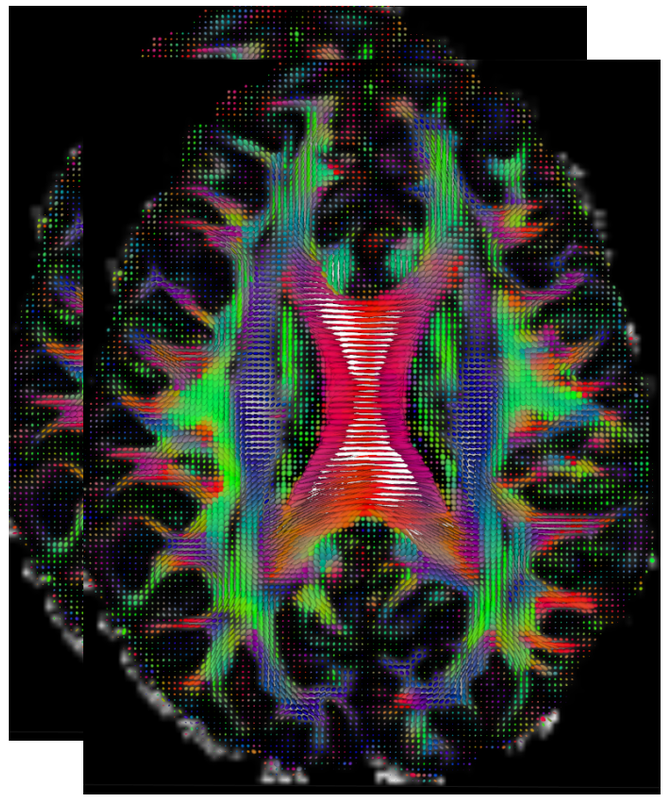
Synthetic data





30 subjects:
15 with effect, 15 without effect

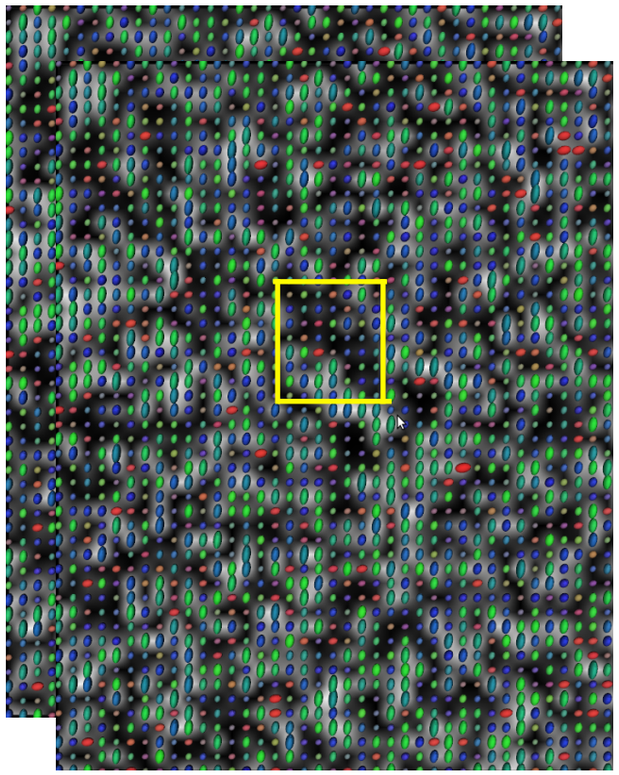
HCP

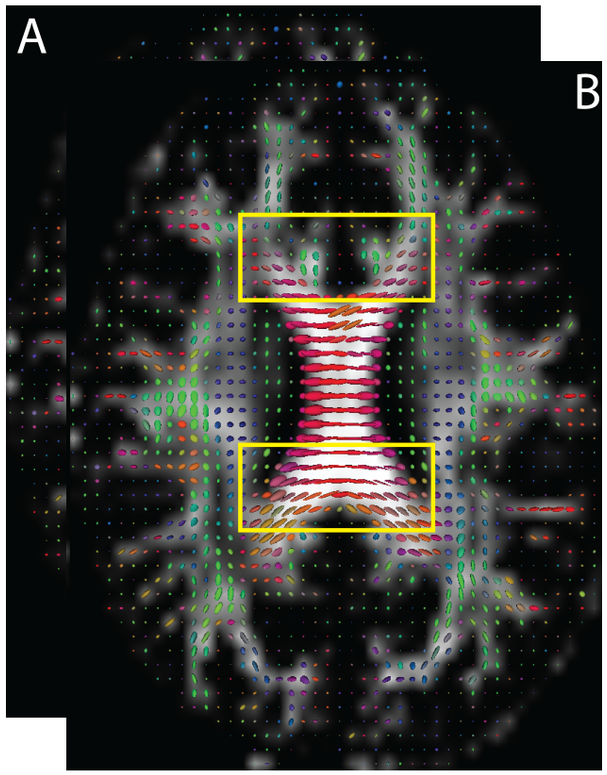


400 subjects:
200 male and 200 female

Data

Synthetic data





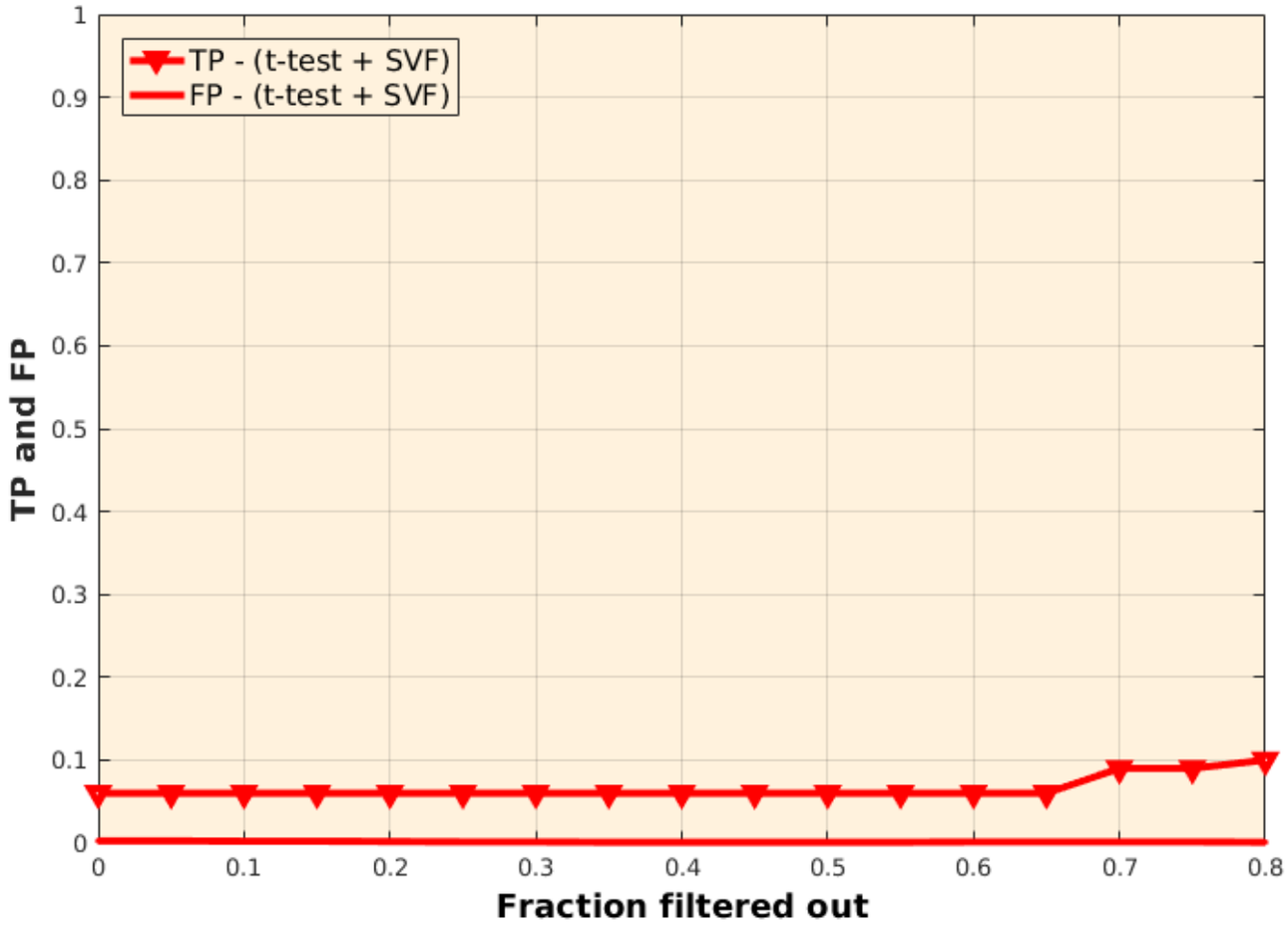
30 subjects:
15 with effect, 15 without effect

HCP

400 subjects:
200 male and 200 female

Experiment results - 1

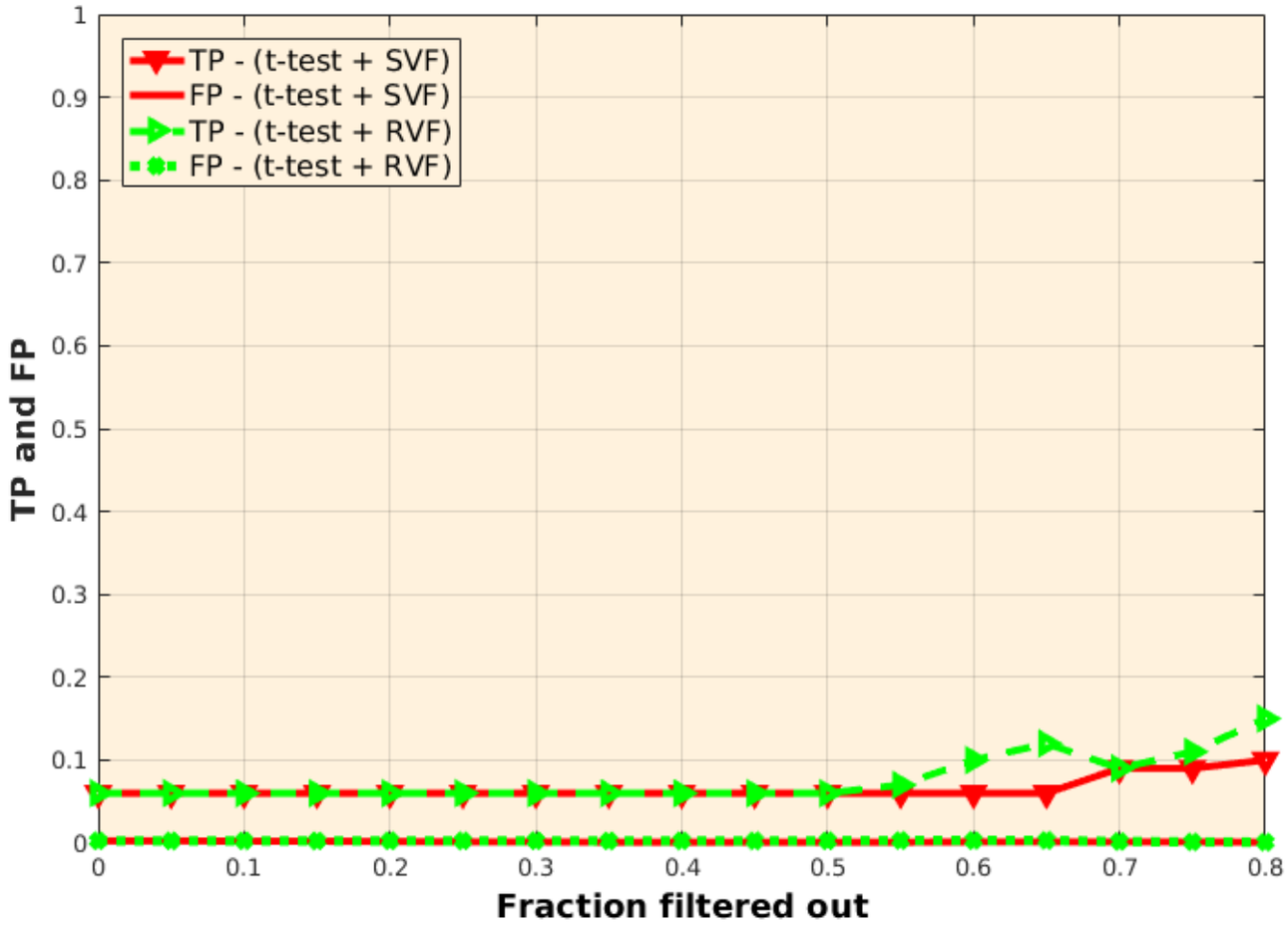
both change the eigenvalues and orientations



$$\alpha = 0.05$$

Experiment results - 1

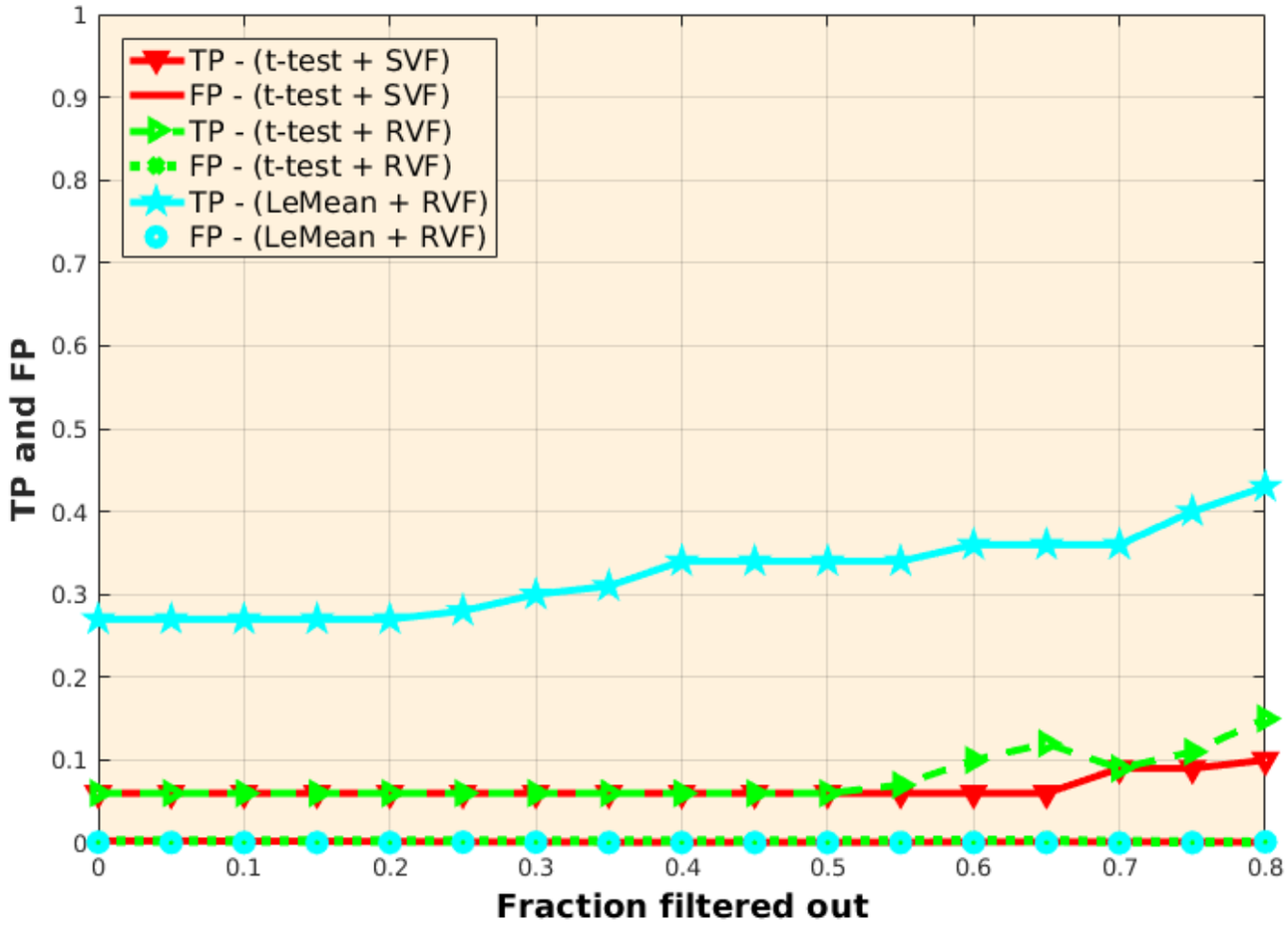
both change the eigenvalues and orientations



$$\alpha = 0.05$$

Experiment results - 1

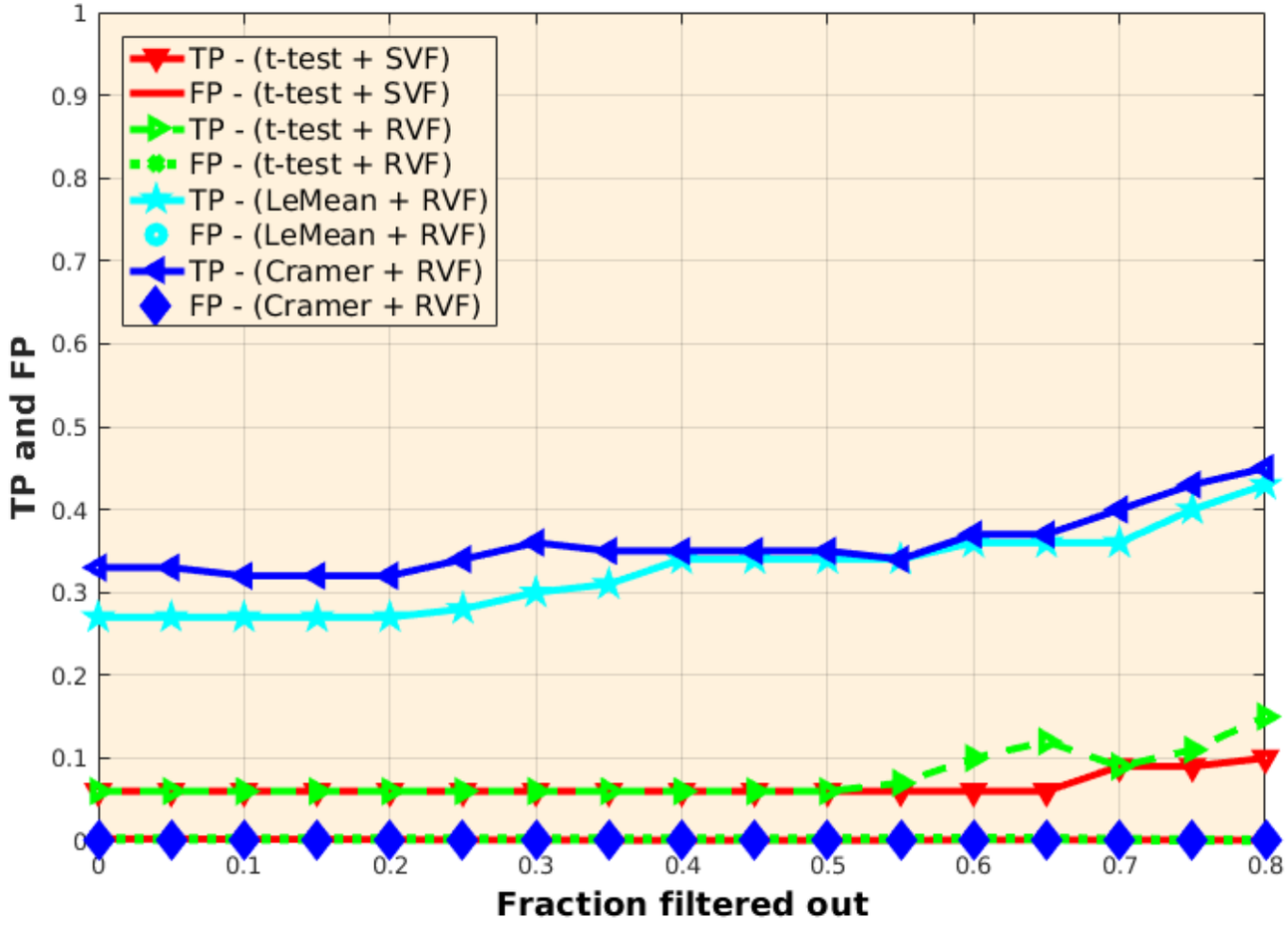
both change the eigenvalues and orientations



$$\alpha = 0.05$$

Experiment results - 1

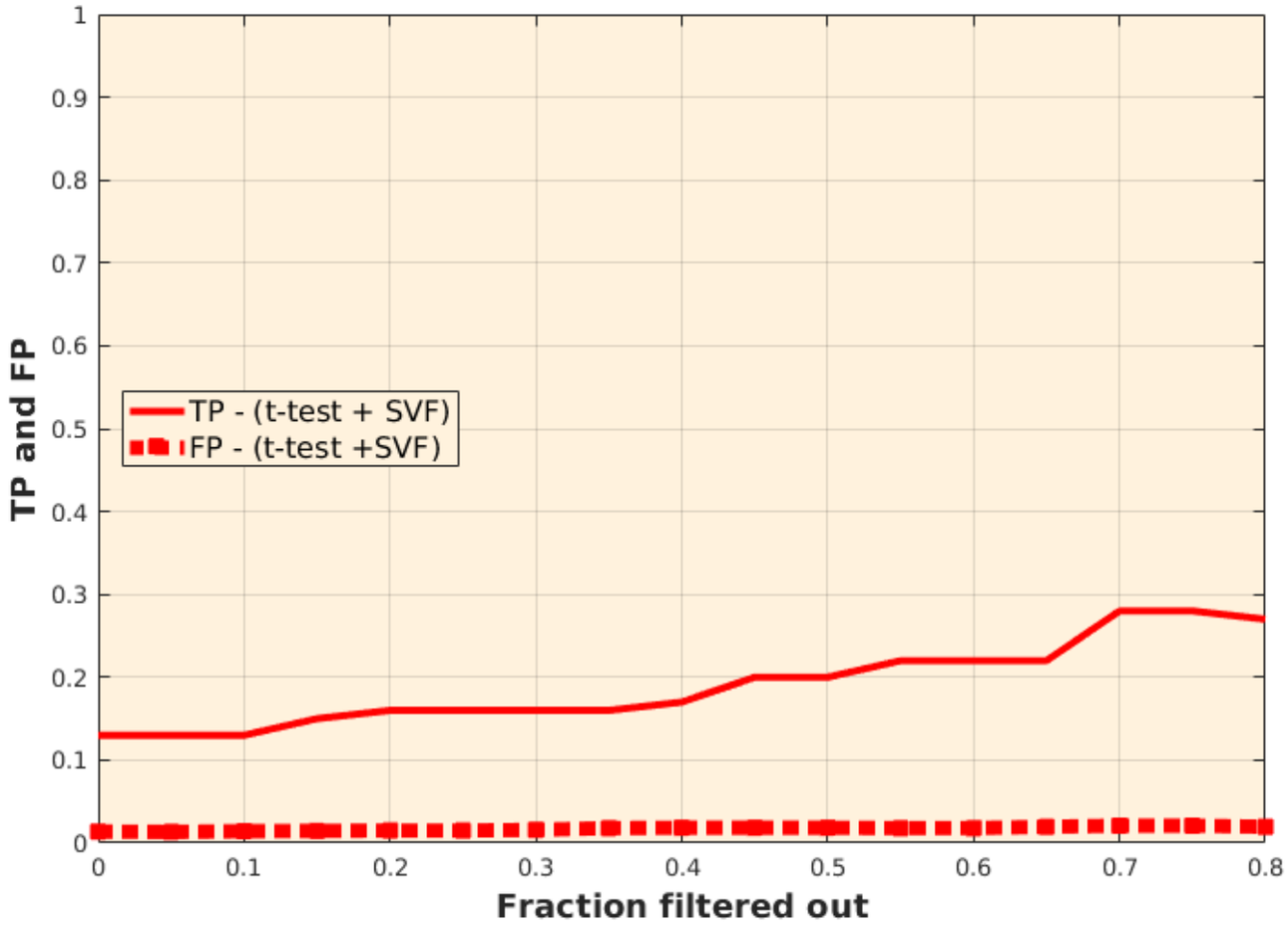
both change the eigenvalues and orientations



$$\alpha = 0.05$$

Experiment results - 1

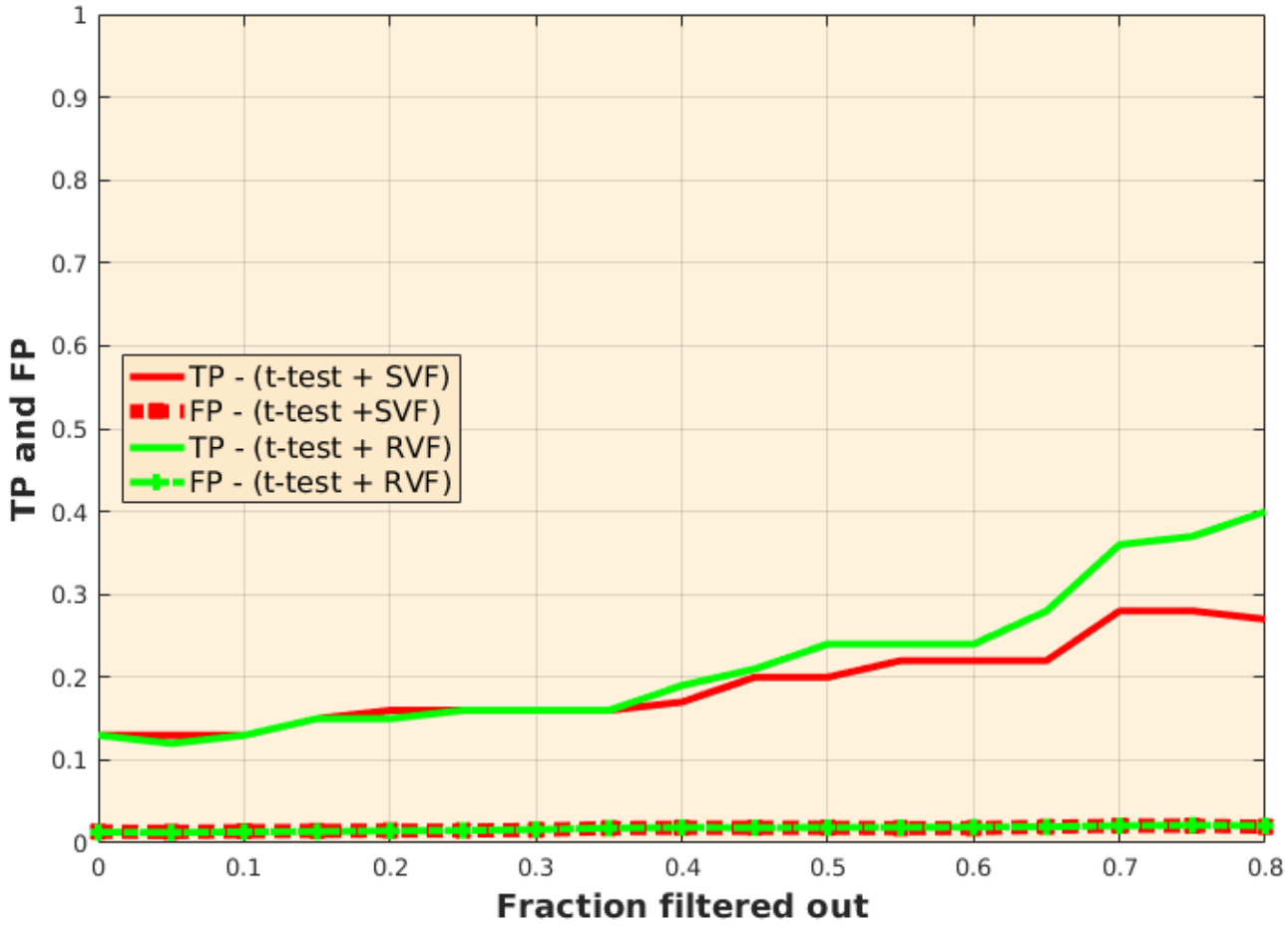
change eigenvalues



$$\alpha = 0.05$$

Experiment results - 1

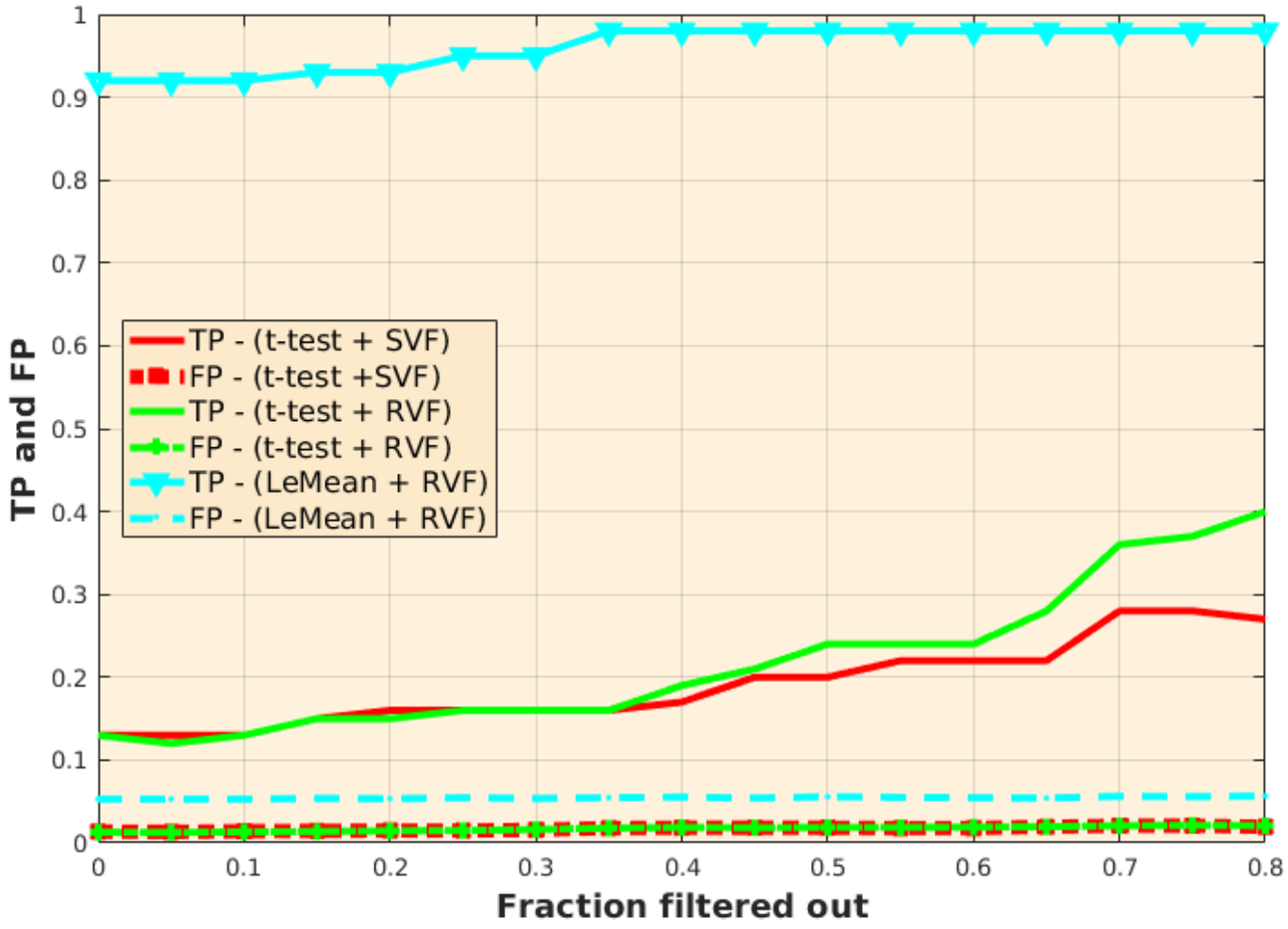
change eigenvalues



$$\alpha = 0.05$$

Experiment results - 1

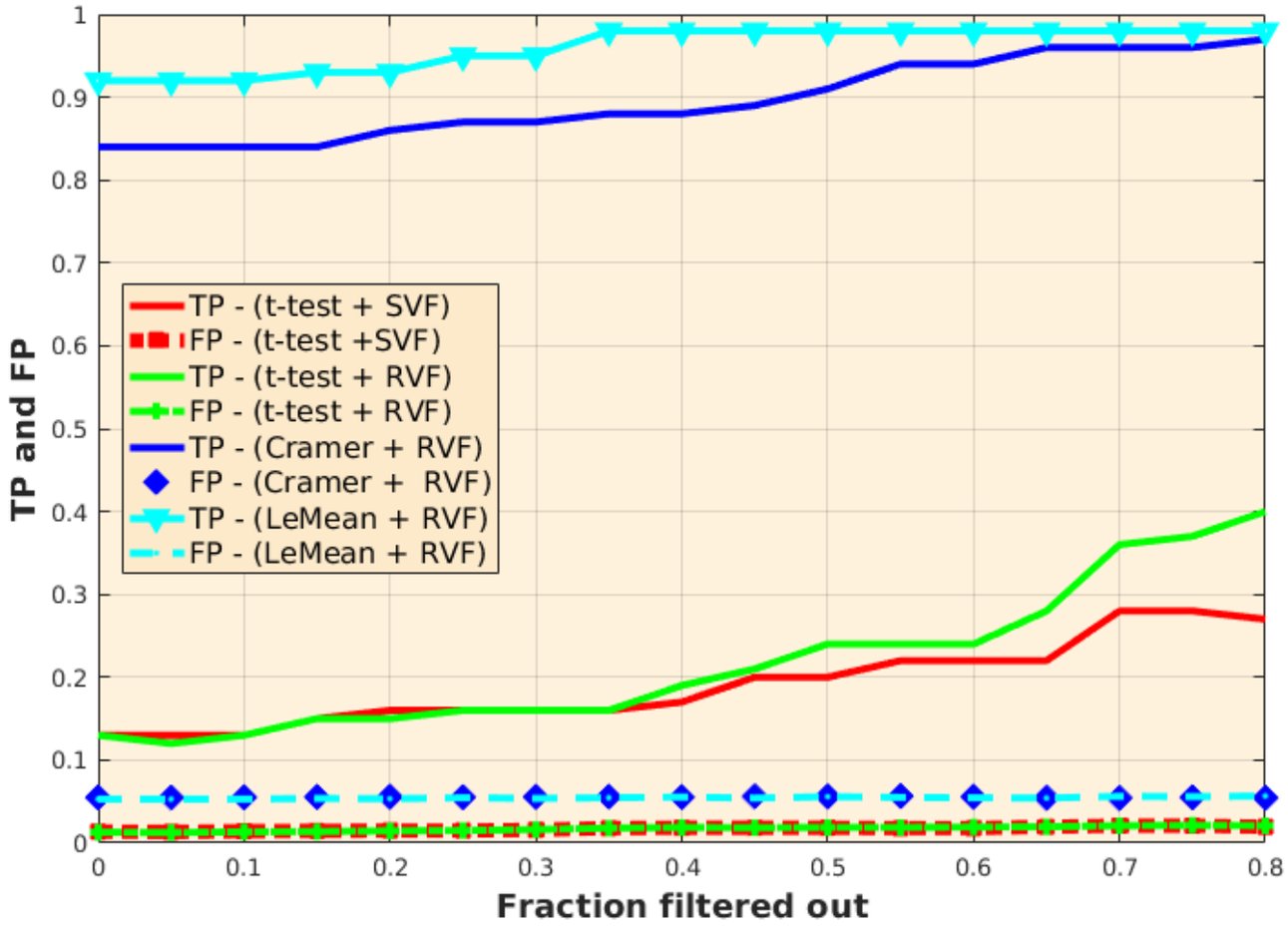
change eigenvalues



$$\alpha = 0.05$$

Experiment results - 1

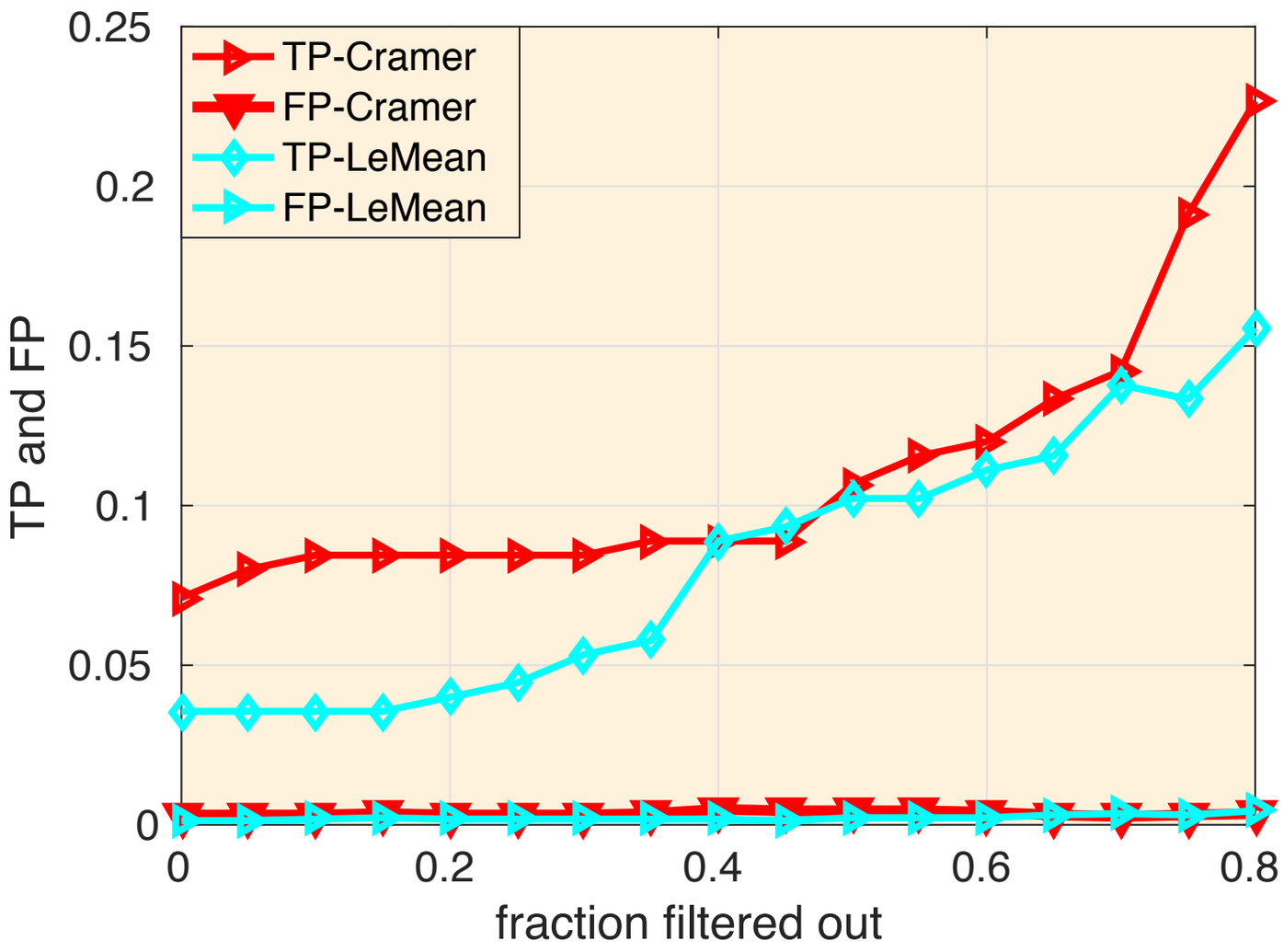
change eigenvalues



$$\alpha = 0.05$$

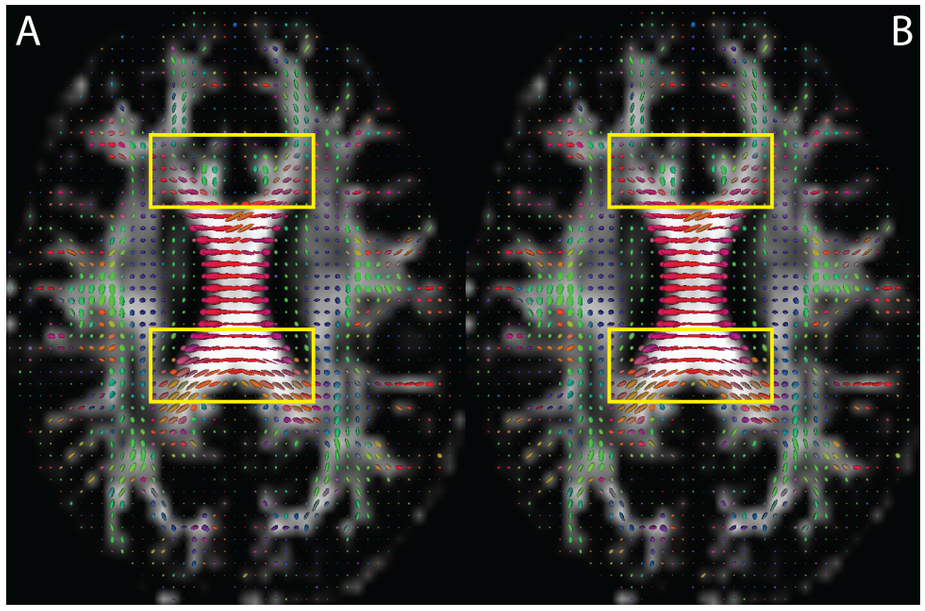
Experiment results - 1

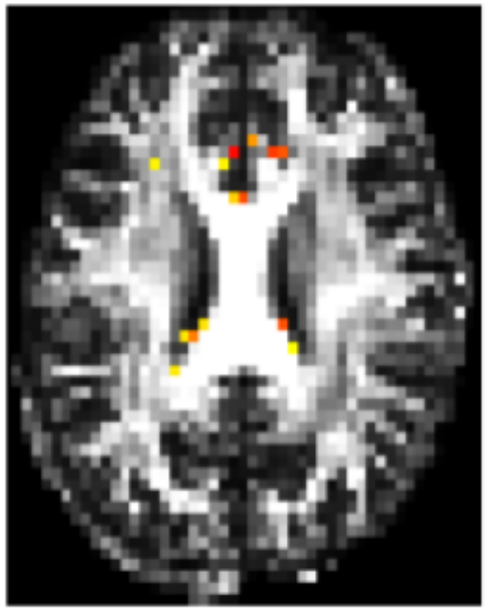
change orientations



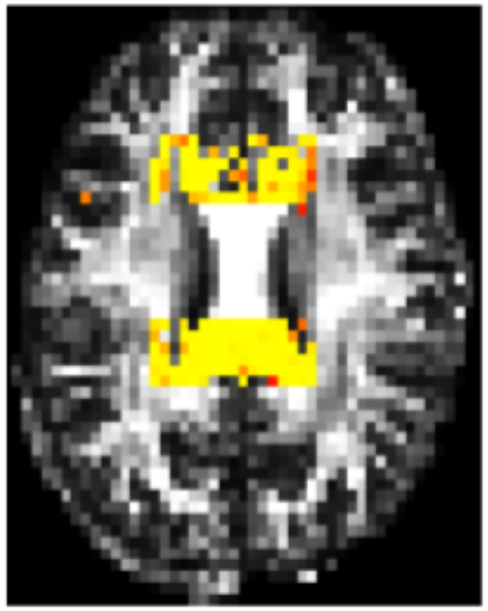
FA based methods failed! $\alpha = 0.05$

Experiment results - 2

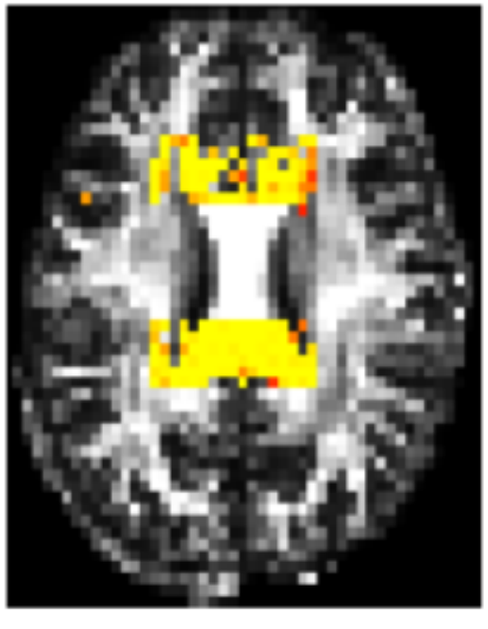




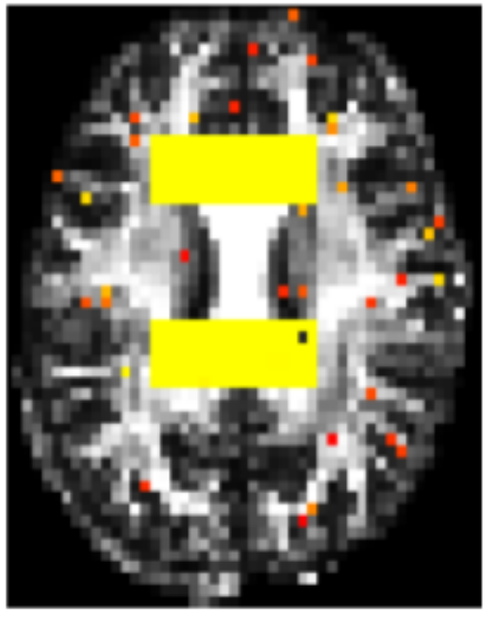
(a) SVF + t-test



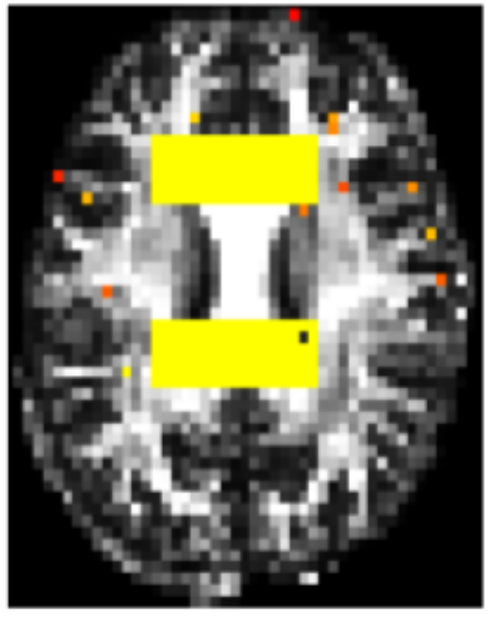
(b) RVF + t-test



(c) RVF + MFA



(d) RFV + Cramér



(e) RVF + LeMean

changes the orientation and eigenvalues

$\alpha = 0.05$

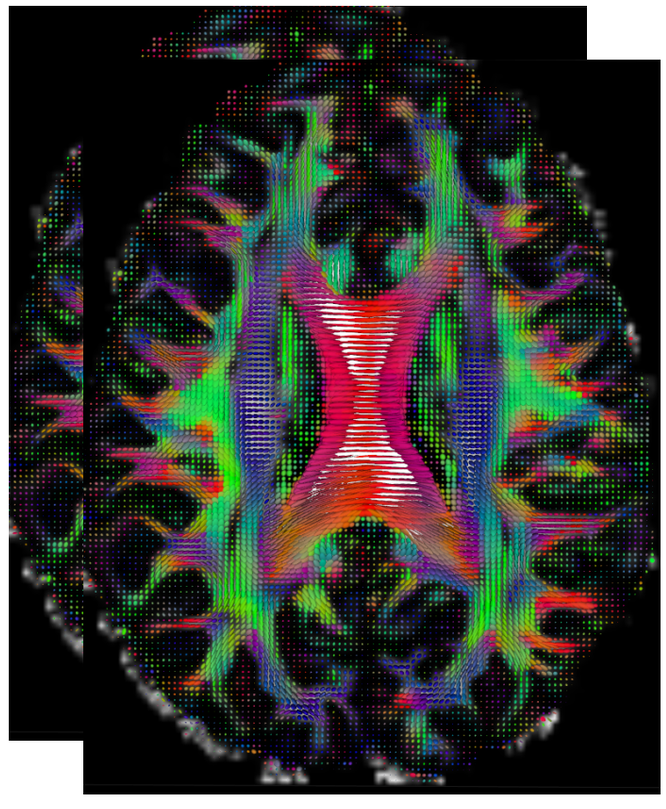
60% filtered out

Data

Simulated data

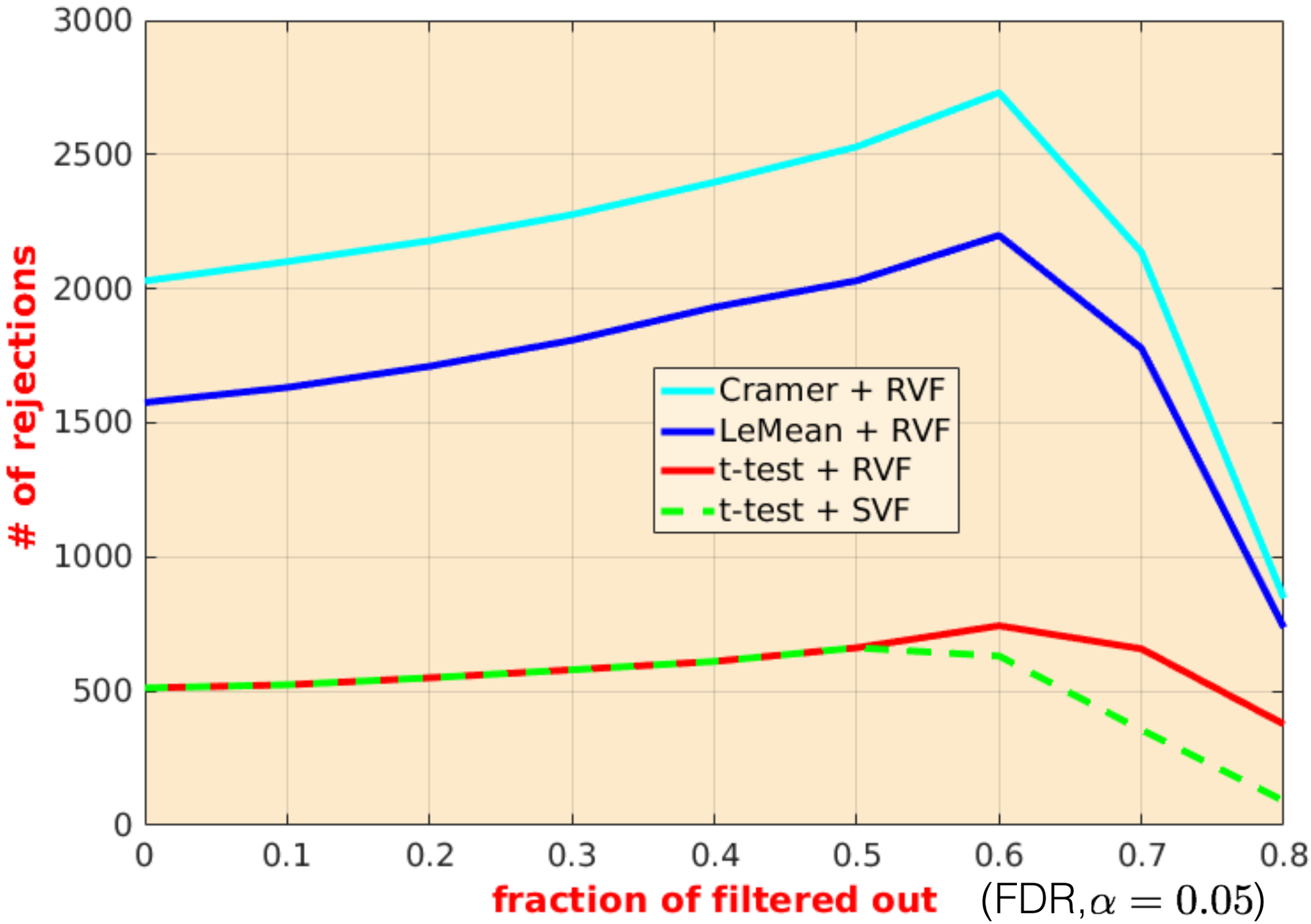
30 subjects:
15 with effect, 15 without effect

HCP



400 subjects:
200 male and 200 female

Experiment results - 3



(FDR, $\alpha = 0.05$)

Conclusion

- Filtering (feature selection) is ubiquitous in data science and it may change the null distribution of downstream analysis.

- Independent filtering does not change the null distribution (p-values in downstream analysis remain valid) while improving statistical power.

- We studied independent filterings for manifold-valued data.

Question