

Theory in Practice: Modeling in Neuroimaging

How to model “big” MRI datasets

Outline of talk

- **Theory recap: modelling approaches can be reduced to two types: predictive and descriptive**
- “Big data” complicates our ability to apply both approaches
- Marginal Modelling is a good approach good for descriptive modelling
- Functional Random Forests is a good approach for predictive modelling
- Other approaches can also handle big data, but are beyond the scope of this workshop

Before even considering models, we need to know what question to ask

- How and where may cortical thickness be associated with working memory performance?











Before even considering models, we need to know what question to ask



- How and where may cortical thickness be associated with working memory performance?
- Can measures of functional brain organization predict an individual's working memory ability?

Each question requires a different modelling approach











- How and where may cortical thickness be associated with working memory performance? **Descriptive modelling**
- Can measures of functional brain organization predict an individual's working memory ability? **Predictive modelling**

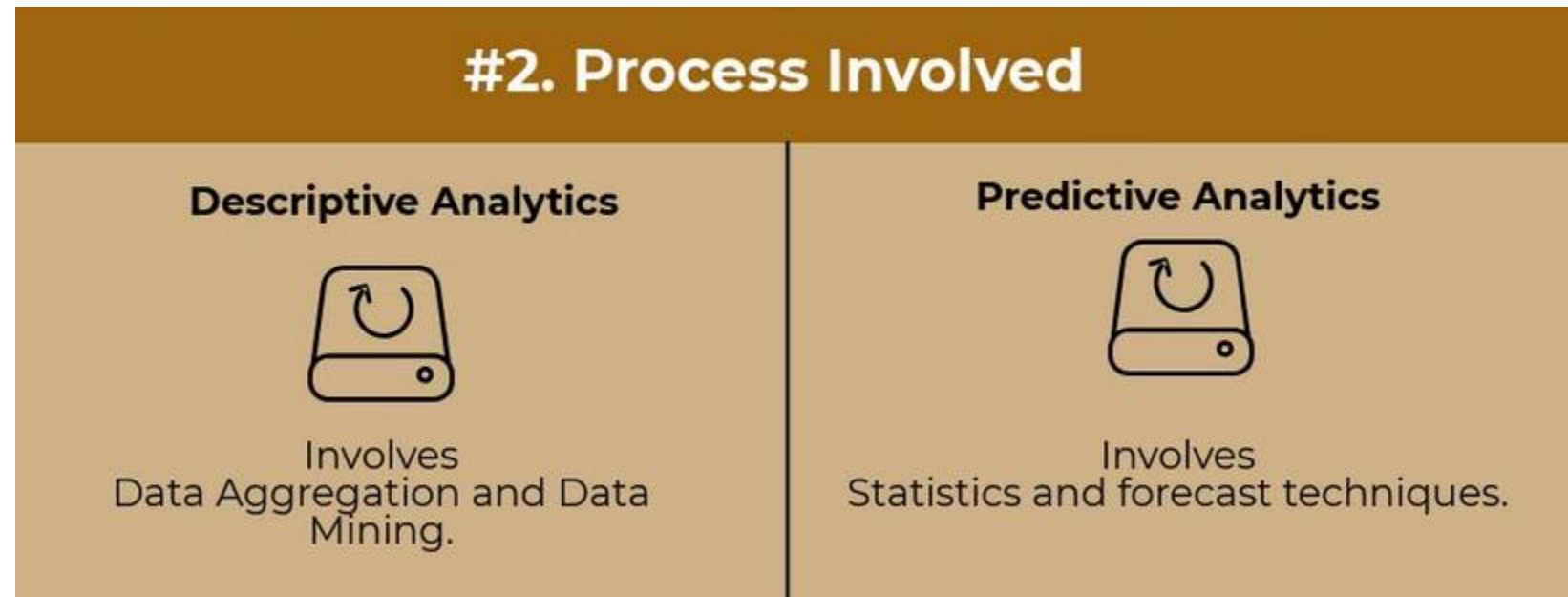
Descriptive models measure what one has collected predictive models measure what one will collect

1#. Describes	
Descriptive Analytics  What happened in the past? By using the stored data.	Predictive Analytics  What might happen in the future? By using the past data and analysing it.
#2. Process Involved	
Descriptive Analytics  Involves Data Aggregation and Data Mining.	Predictive Analytics  Involves Statistics and forecast techniques.
3#. Definition	
Descriptive Analytics  The process of finding the useful and important information by analysing the huge data.	Predictive Analytics  This process involves in forecasting the future of the company, which are very useful.
4#. Data Volume	
Descriptive Analytics  It involves in processing huge data that are stored in data warehouses. Limited to past data.	Predictive Analytics  It involves in analysing large past data and then predicts the future using advance techniques.
5#. Accuracy	
Descriptive Analytics  It provides accurate data in the reports using past data.	Predictive Analytics  Results are not accurate, it will not tell you exactly what will happen but it will tell you what might happen in the future.











1#. Describes	
Descriptive Analytics  What happened in the past? By using the stored data.	Predictive Analytics  What might happen in the future? By using the past data and analysing it.

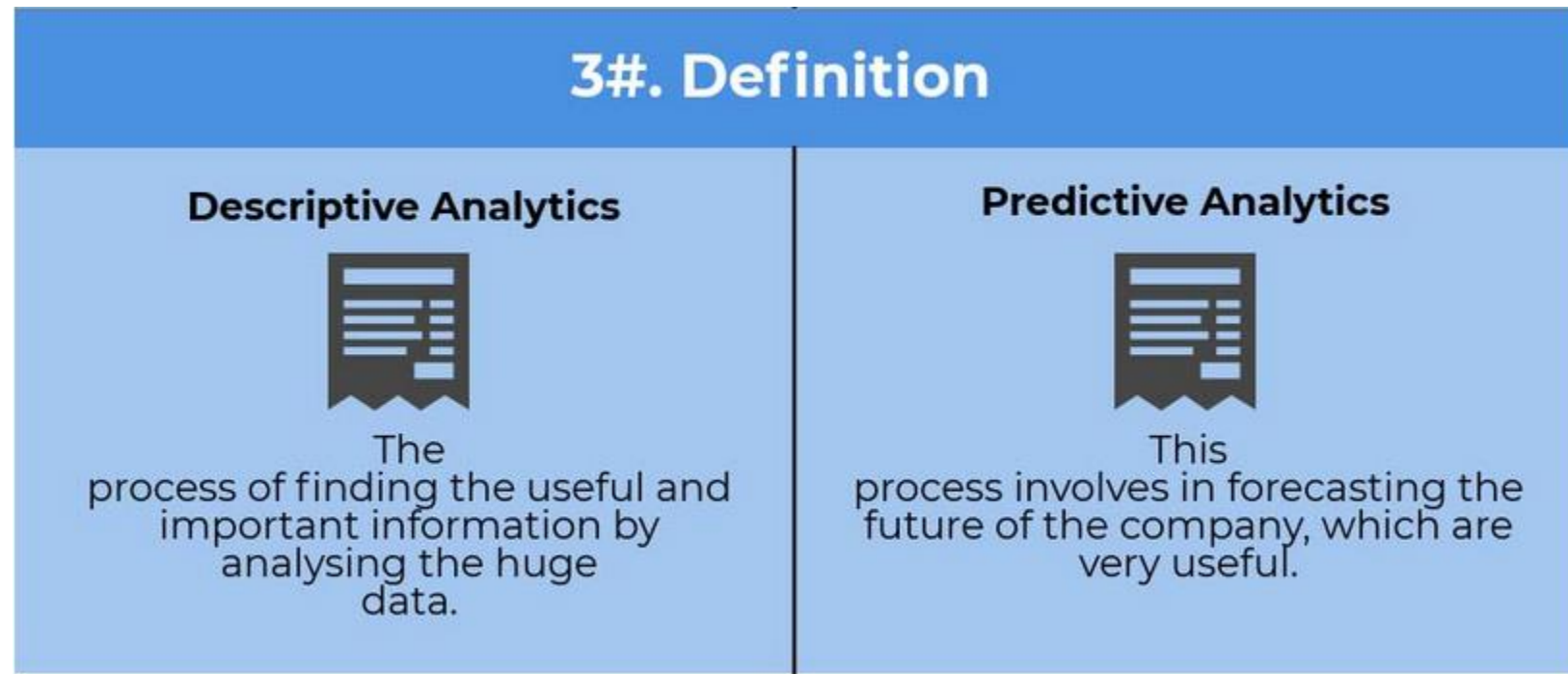
Descriptive models explore data, predictive models confirm properties of data

1#. Describes	
Descriptive Analytics  What happened in the past? By using the stored data.	Predictive Analytics  What might happen in the future? By using the past data and analysing it.
#2. Process Involved	
Descriptive Analytics  Involves Data Aggregation and Data Mining.	Predictive Analytics  Involves Statistics and forecast techniques.
3#. Definition	
Descriptive Analytics  The process of finding the useful and important information by analysing the huge data.	Predictive Analytics  This process involves in forecasting the future of the company, which are very useful.
4#. Data Volume	
Descriptive Analytics  It involves in processing huge data that are stored in data warehouses. Limited to past data.	Predictive Analytics  It involves in analysing large past data and then predicts the future using advance techniques.
5#. Accuracy	
Descriptive Analytics  It provides accurate data in the reports using past data.	Predictive Analytics  Results are not accurate, it will not tell you exactly what will happen but it will tell you what might happen in the future.













Descriptive models provide insight, predictive models apply insight



1#. Describes	
Descriptive Analytics  What happened in the past? By using the stored data.	Predictive Analytics  What might happen in the future? By using the past data and analysing it.
#2. Process Involved	
Descriptive Analytics  Involves Data Aggregation and Data Mining.	Predictive Analytics  Involves Statistics and forecast techniques.
3#. Definition	
Descriptive Analytics  The process of finding the useful and important information by analysing the huge data.	Predictive Analytics  This process involves in forecasting the future of the company, which are very useful.
4#. Data Volume	
Descriptive Analytics  It involves in processing huge data that are stored in data warehouses. Limited to past data.	Predictive Analytics  It involves in analysing large past data and then predicts the future using advance techniques.
5#. Accuracy	
Descriptive Analytics  It provides accurate data in the reports using past data.	Predictive Analytics  Results are not accurate. It will not tell you exactly what will happen but it will tell you what might happen in the future.













Descriptive models are limited to in-sample data, predictive models require out-of-sample data

1#. Describes	
Descriptive Analytics  What happened in the past? By using the stored data.	Predictive Analytics  What might happen in the future? By using the past data and analysing it.
#2. Process Involved	
Descriptive Analytics  Involves Data Aggregation and Data Mining.	Predictive Analytics  Involves Statistics and forecast techniques.
3#. Definition	
Descriptive Analytics  The process of finding the useful and important information by analysing the huge data.	Predictive Analytics  This process involves in forecasting the future of the company, which are very useful.
4#. Data Volume	
Descriptive Analytics  It involves in processing huge data that are stored in data warehouses. Limited to past data.	Predictive Analytics  It involves in analysing large past data and then predicts the future using advance techniques.
5#. Accuracy	
Descriptive Analytics  It provides accurate data in the reports using past data.	Predictive Analytics  Results are not accurate. It will not tell you exactly what will happen but it will tell you what might happen in the future.



4#. Data Volume

Descriptive Analytics	Predictive Analytics
 It involves in processing huge data that are stored in data warehouses. Limited to past data.	 It involves in analysing large past data and then predicts the future using advance techniques.

Descriptive models are assessed via theory and inference, predictive models are assessed by independent testing

1#. Describes	
Descriptive Analytics  What happened in the past? By using the stored data.	Predictive Analytics  What might happen in the future? By using the past data and analysing it.
#2. Process Involved	
Descriptive Analytics  Involves Data Aggregation and Data Mining.	Predictive Analytics  Involves Statistics and forecast techniques.
3#. Definition	
Descriptive Analytics  The process of finding the useful and important information by analysing the huge data.	Predictive Analytics  This process involves in forecasting the future of the company, which are very useful.
4#. Data Volume	
Descriptive Analytics  It involves in processing huge data that are stored in data warehouses. Limited to past data.	Predictive Analytics  It involves in analysing large past data and then predicts the future using advance techniques.
5#. Accuracy	
Descriptive Analytics  It provides accurate data in the reports using past data.	Predictive Analytics  Results are not accurate, it will not tell you exactly what will happen but it will tell you what might happen in the future.

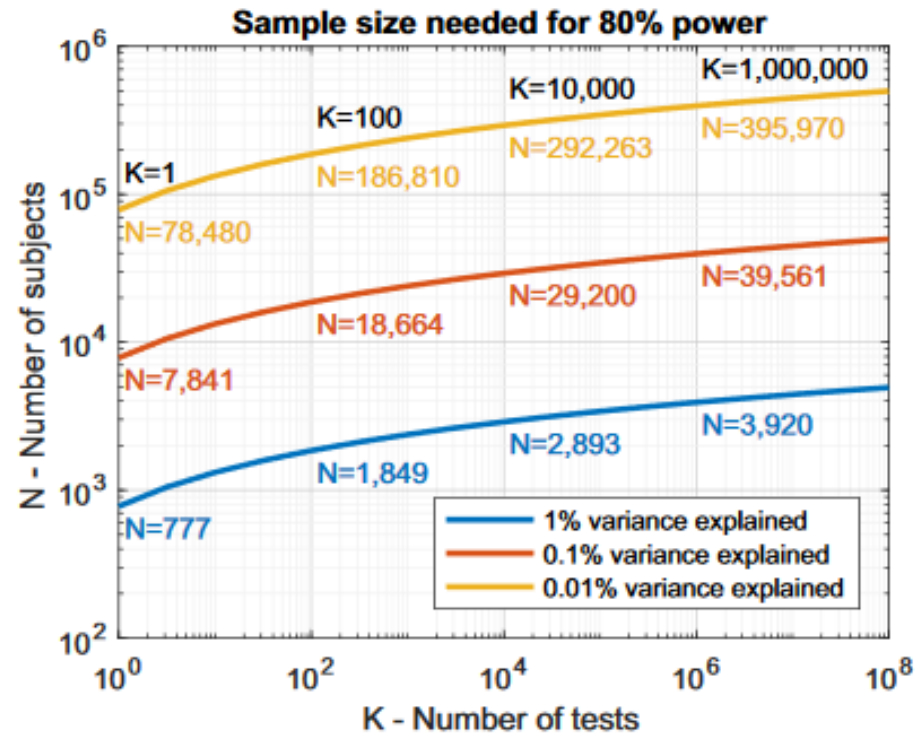
5 #. Accuracy

Descriptive Analytics	Predictive Analytics
	
It provides accurate data in the reports using past data.	Results are not accurate, it will not tell you exactly what will happen but it will tell you what might happen in the future.

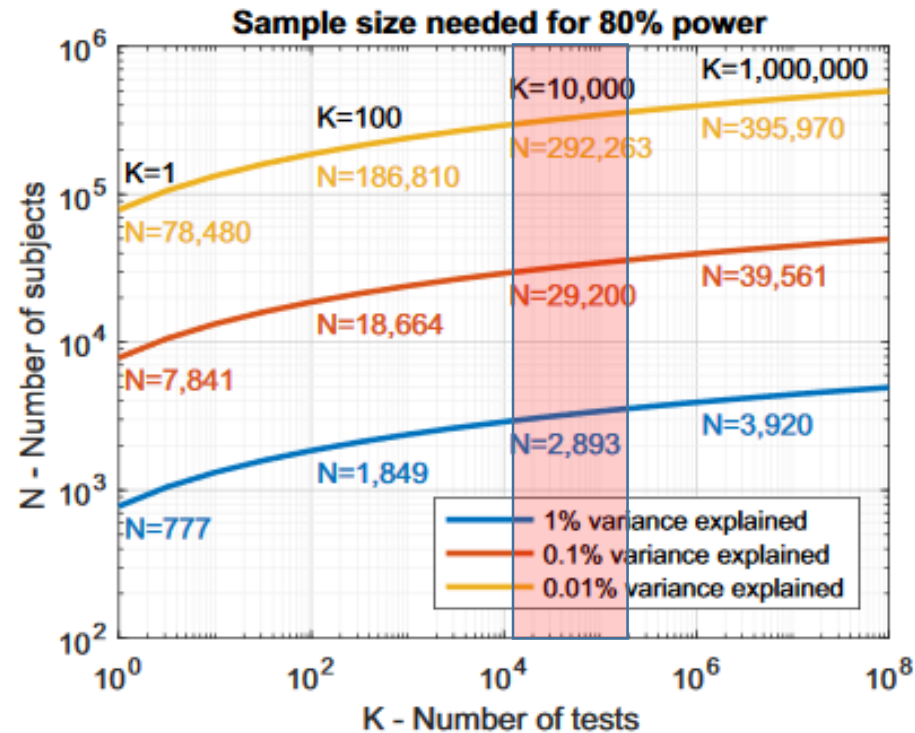
Outline of talk

- Theory recap: modelling approaches can be reduced to two types: predictive and descriptive
- **“Big data” complicates our ability to apply both approaches**
- Marginal Modelling is a good approach for descriptive modelling
- Functional Random Forests is a good approach for predictive modelling
- Other approaches can also handle big data, but are beyond the scope of this workshop

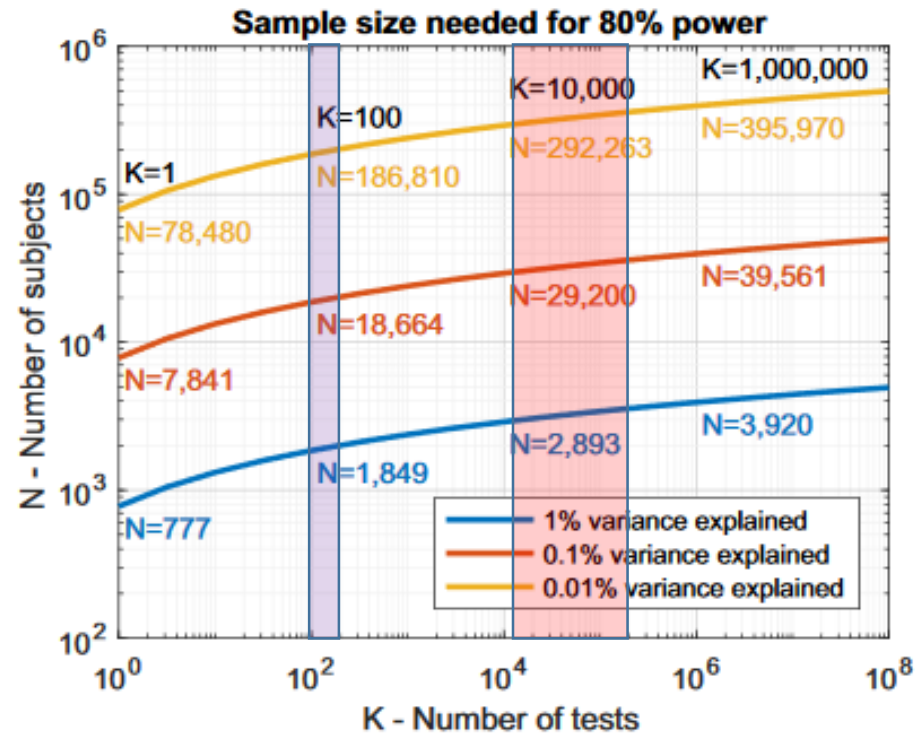
First, all health-focused imaging studies should probably be big data



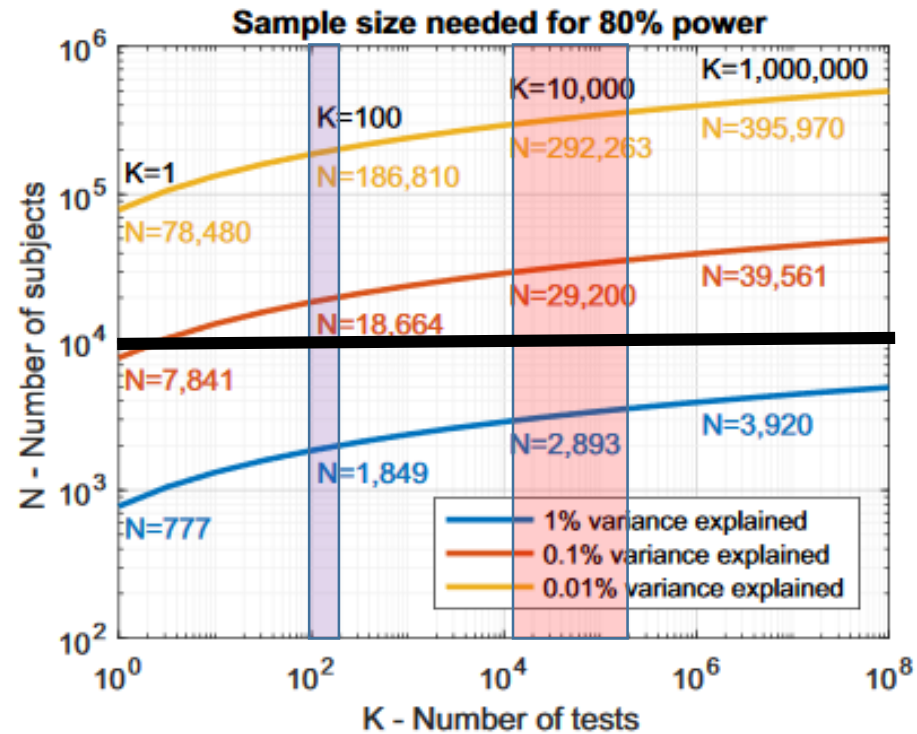
Our ABCD pipeline generates anywhere from 10 to 90 thousand tests



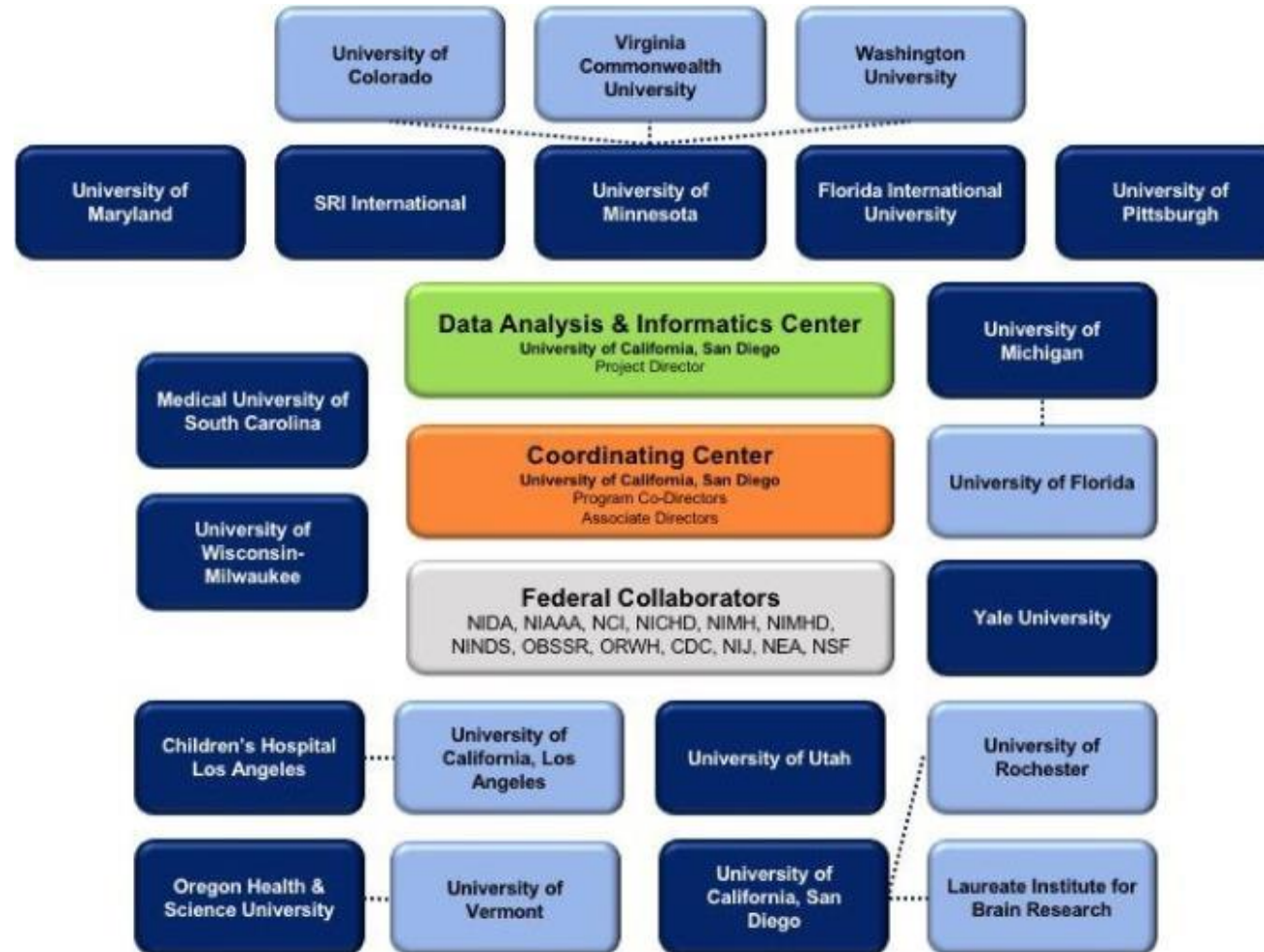
Our ABCD pipeline generates anywhere from 10 to 90 thousand tests (some special cases are in hundreds)



We've collected about 10,000 cases



ABCD needed a lot of coordination and data aggregation to collect over 10,000 participants



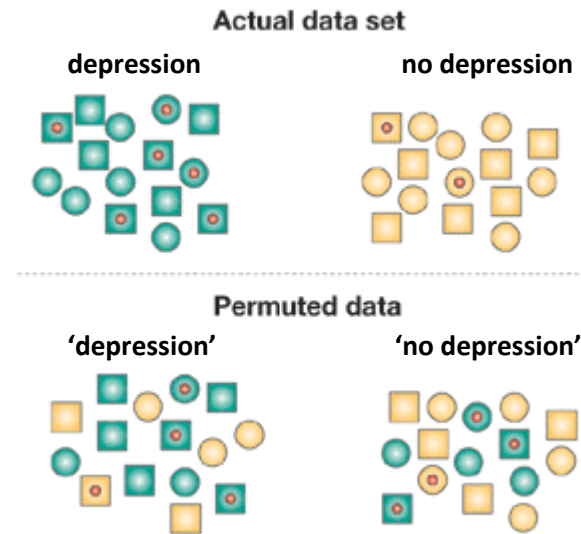
Descriptive models must take into account this nested structure

- Complex models may be slow to calculate when analyzing ~4500 participants
 - Permutation tests may take days or even weeks
- Permutation tests lack exchangeability for complex questions

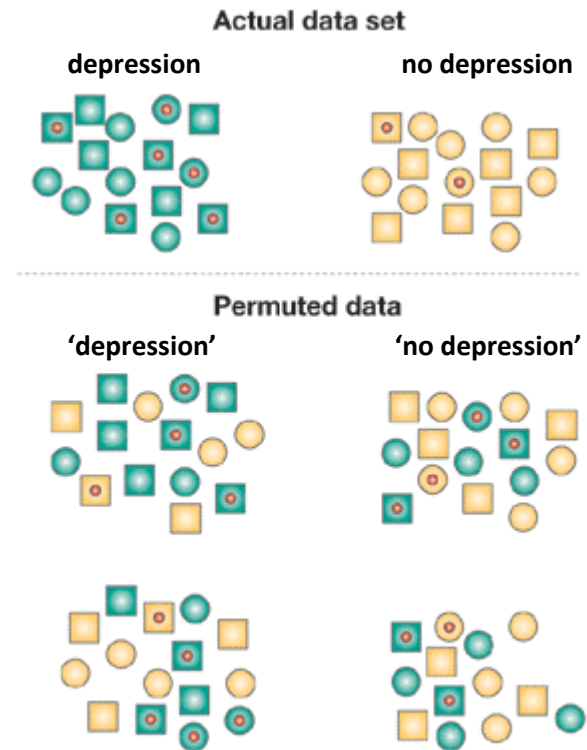
Permutation testing can reveal whether differences in community structure are significantly different



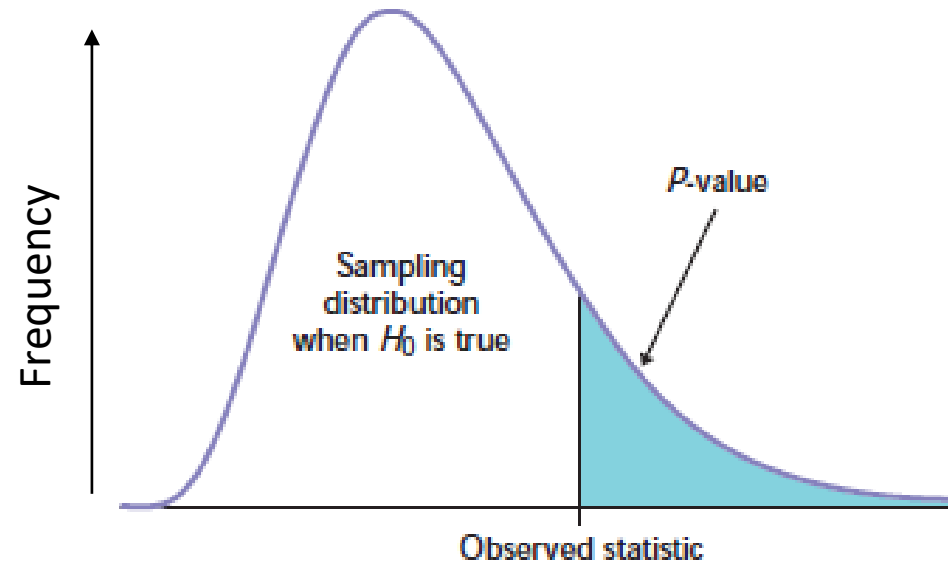
Permute group assignment and calculate statistic



Do so for multiple permutations and construct a distribution of the statistic for permuted groups



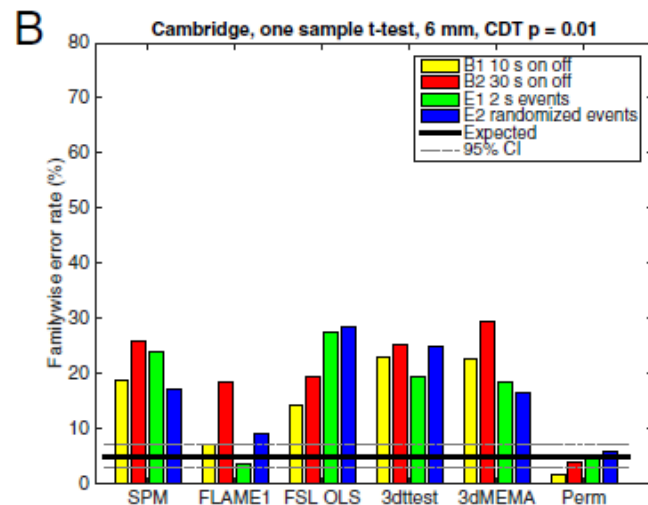
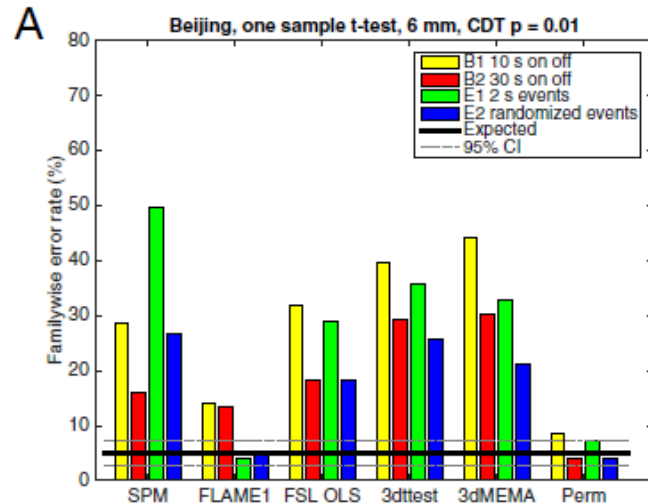
P value is determined by the proportional rank of the observed statistic compared to the permuted distribution



Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund^{a,b,c,1}, Thomas E. Nichols^{d,e}, and Hans Knutsson^{a,c}

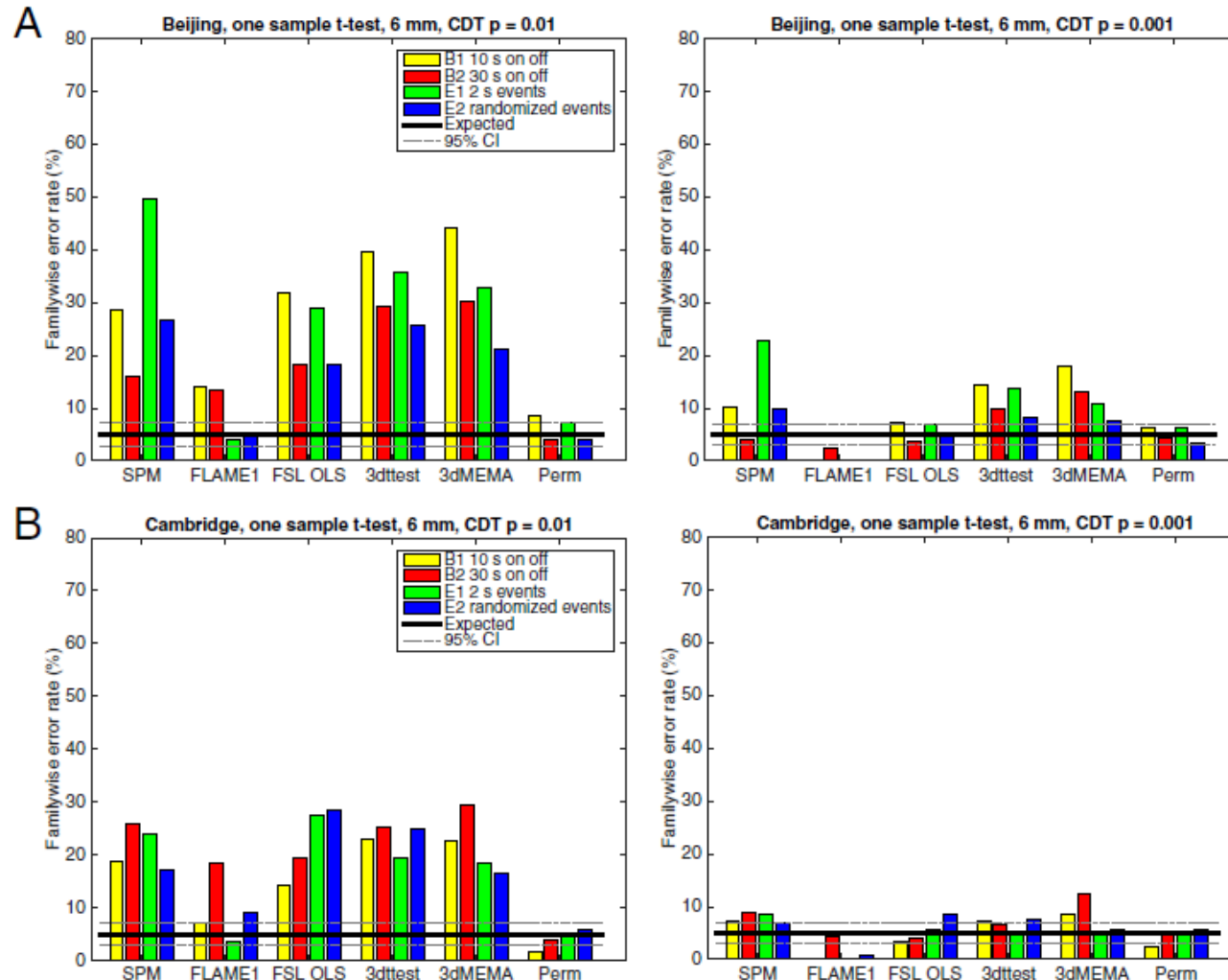
At a $Z=2.3$, false positive rates are high when not using permutation testing



Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund^{a,b,c,1}, Thomas E. Nichols^{d,e}, and Hans Knutsson^{a,c}

At a $Z=3.1$, false positive rates are generally better and in-line with the true FP rate

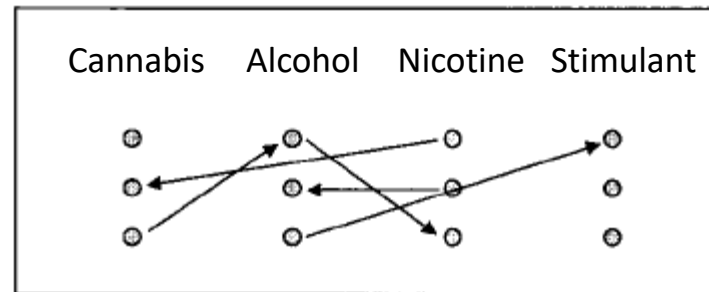


This all works because each individual is independently acquired from one another – the data are **exchangeable**

Independence gets more complicated when you have more complicated designs – but even here we can exchange every individual

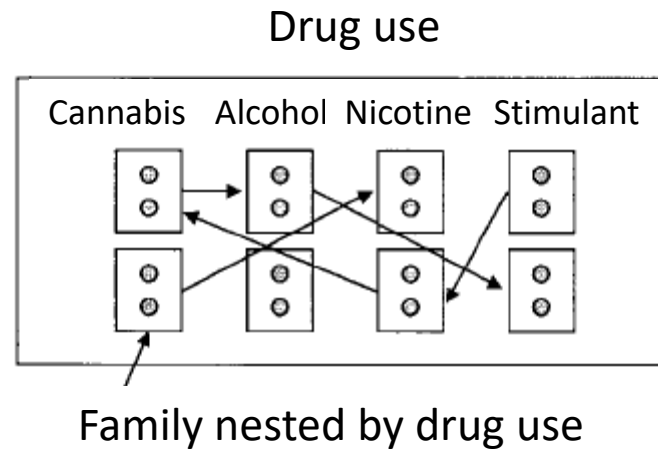
a) One-way model

Drug use



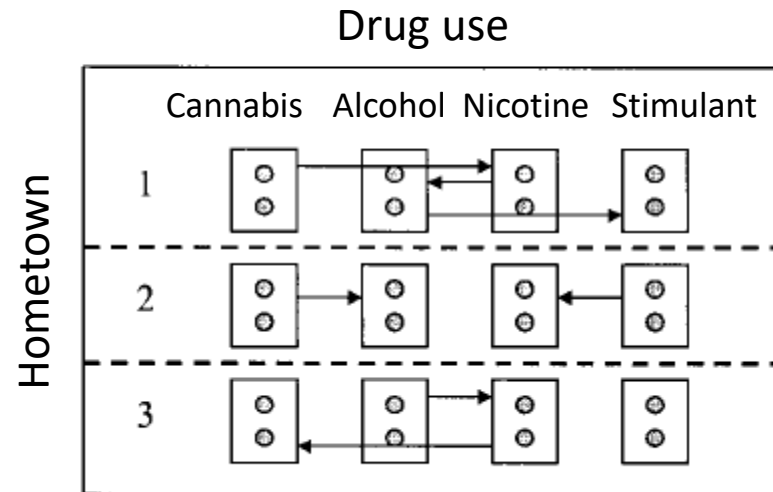
However, if a second factor is nested, our permutations are limited to the nested pairs, restricting our permutations

b) Two-way nested model

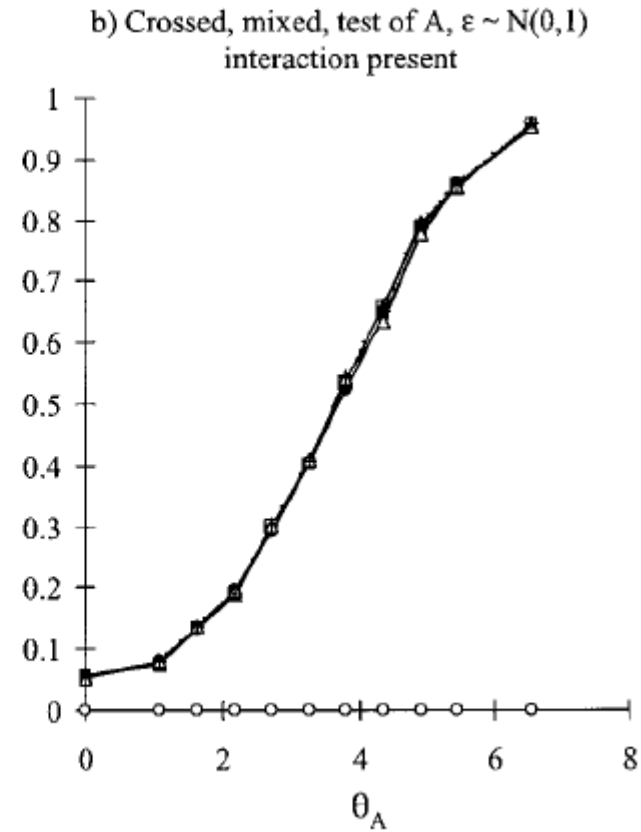
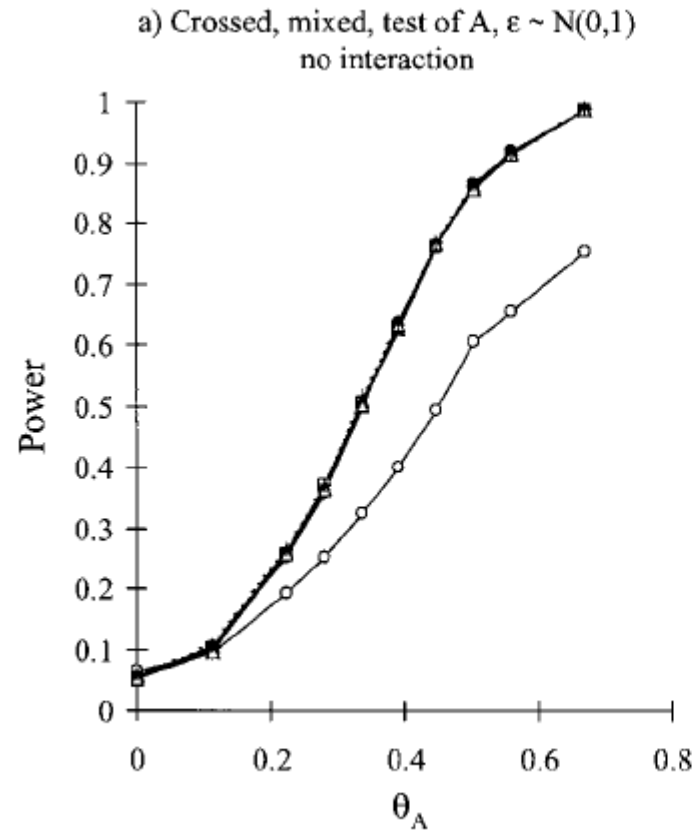


More complex designs have even more restrictions, relative to the total number of permutations

c) Two-way crossed, mixed model

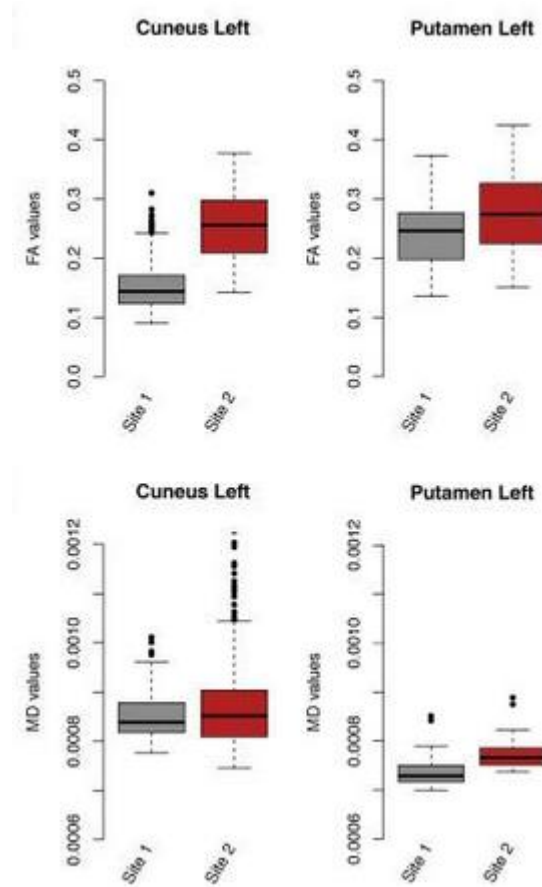


In turn, restricted permutations have reduced power when controlling for the false positive rate

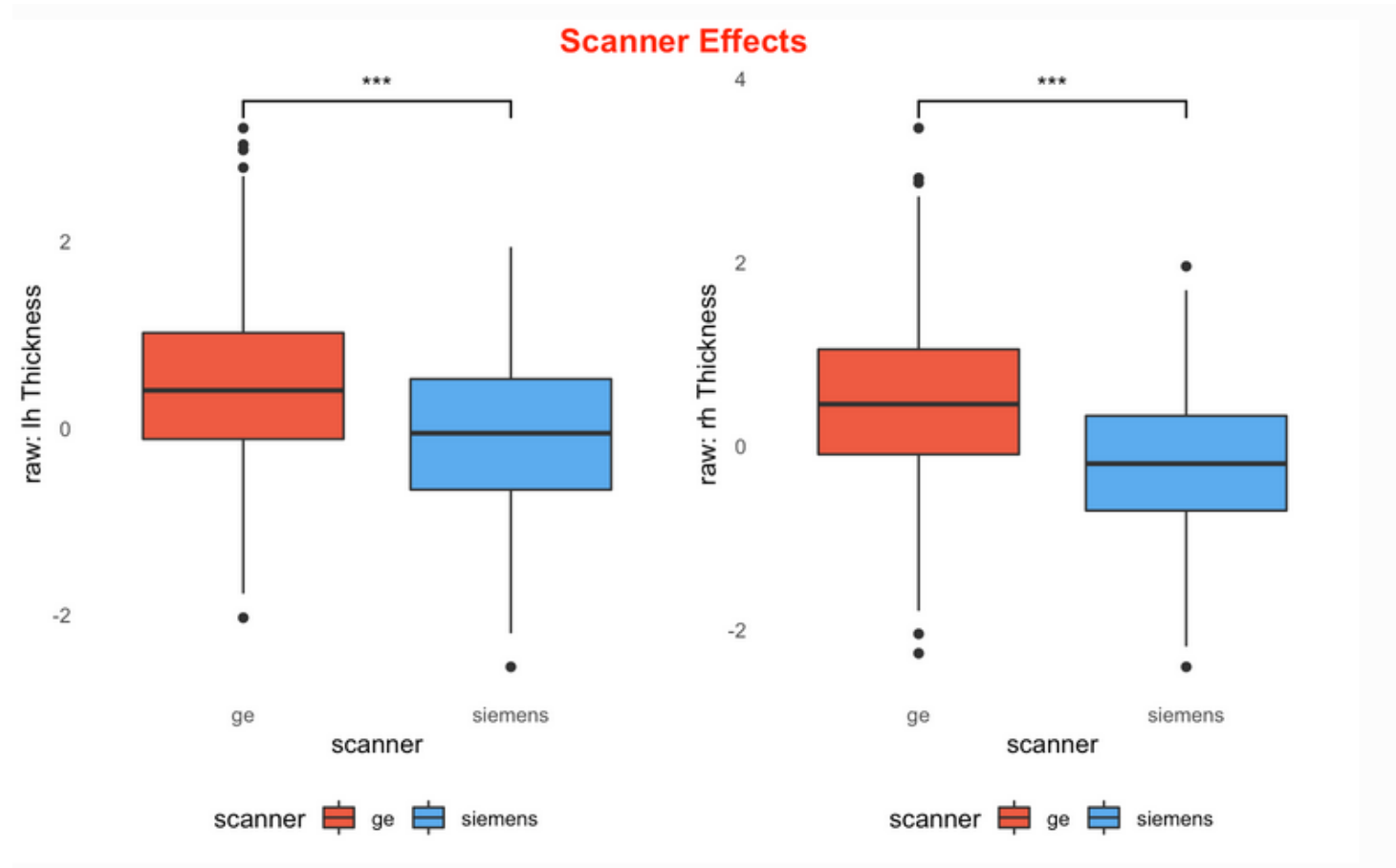


- = Y (permutation of raw data),
- ▲ = R (permutation of residuals),
- = Y(B) (permutation of raw data restricted within levels of B),
- = Rab (permutation of residuals as *ab* units),
- △ = Y(B)ab (permutation of raw data as *ab* units, restricted within levels of B),
- + = normal-theory *F*-test.

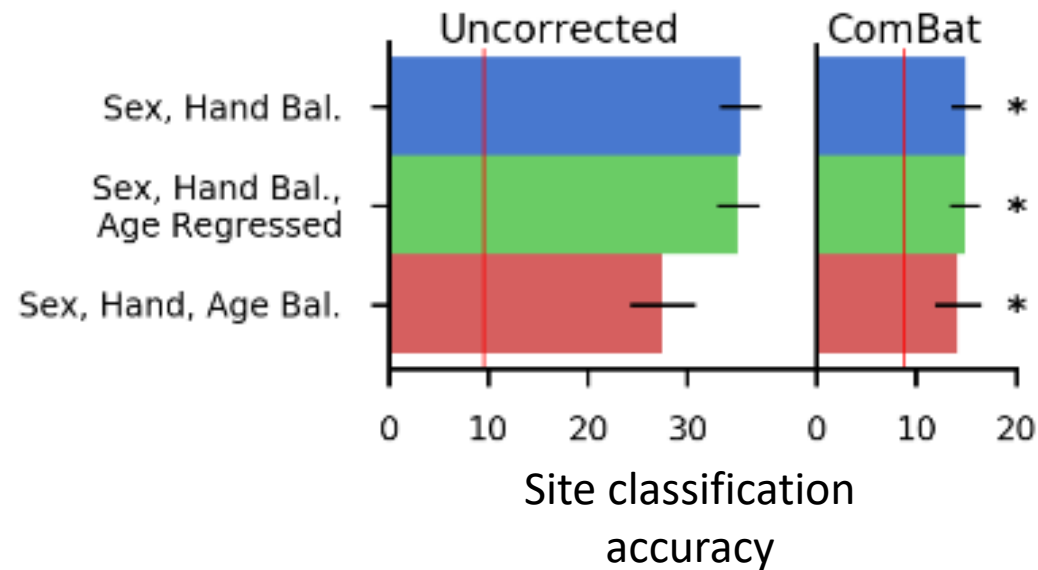
Predictive models must also take into account nested structure



Scanner effects can be common, independent of site



ComBat has also been used to correct for ABCD data, which can be predicted by site



Cross-validation strategies can mitigate known but not unknown effects

- Stratified validation is possible via independent stratified groups
- Leave-one-site-out validation can help catch site effects
- But what about effects of scanner upgrades, software maintenance, or even changes in personnel?

Outline of talk

- Theory recap: modelling approaches can be reduced to two types: predictive and descriptive
- “Big data” complicates our ability to apply both approaches
- **Marginal Modelling is a good approach for descriptive modelling**
- Functional Random Forests is a good approach for predictive modelling
- Other approaches can also handle big data, but are beyond the scope of this workshop

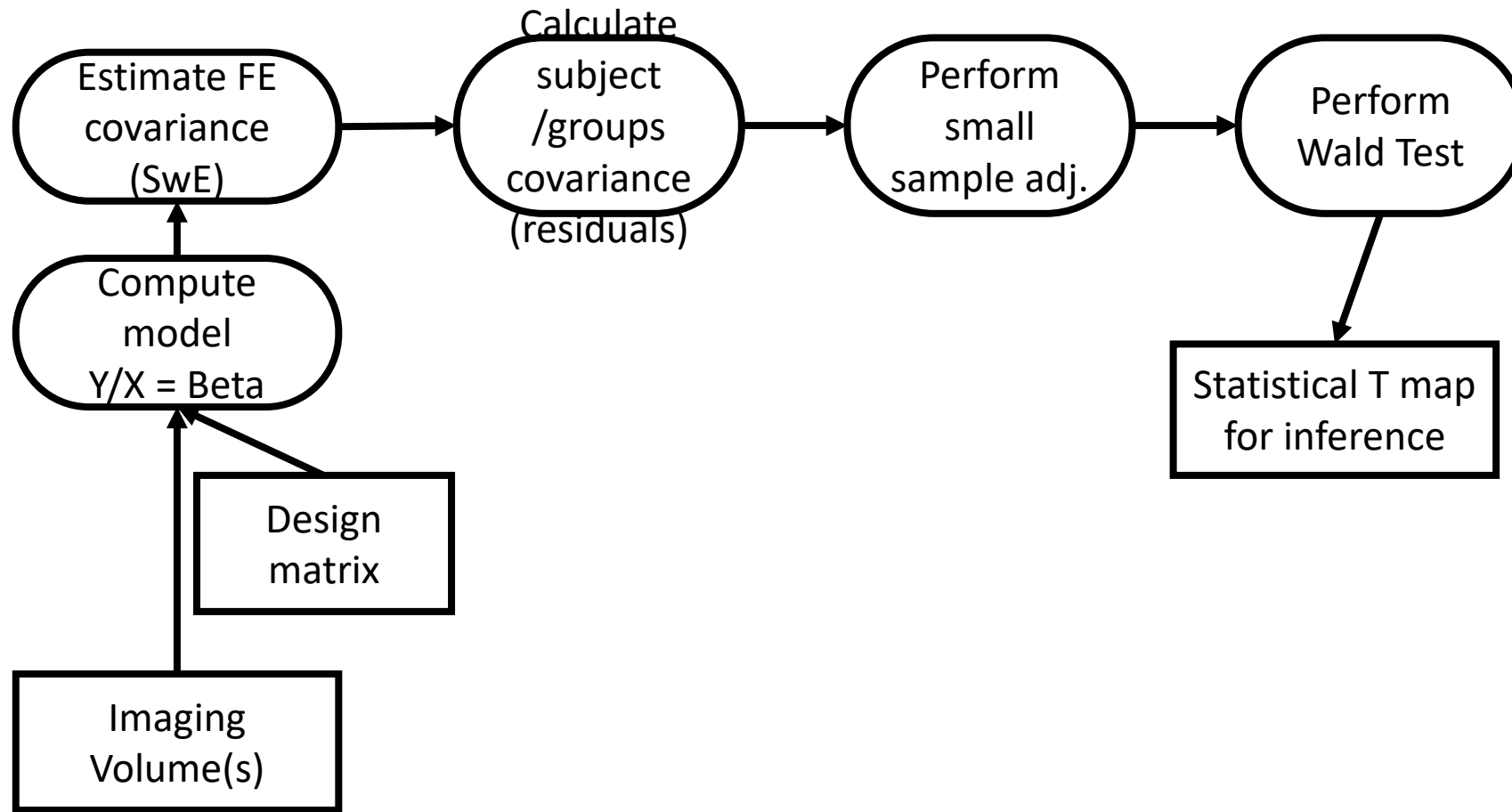
The marginal model may be a more feasible solution for modeling ABCD populations

- Strengths:
 - Marginal model makes few assumptions with respect to the data
 - Nested-designs can be modeled or unmodeled, and left to the error term (hopefully)
 - Individual cases can be incomplete or missing for a marginal model
 - Longitudinal designs are feasible within the marginal model framework
 - Marginal model has a closed-form solution to the equation via a Sandwich Estimator (SwE)
 - It's fast, and can be feasibly run with limited resources on lots of data
 - Use of a wild bootstrap (WB) provides an NHST framework for complex questions

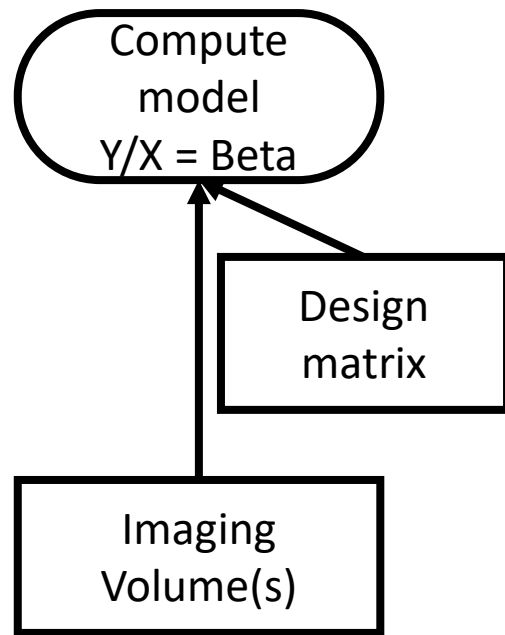
Critical limitations

- The marginal model cannot be used to draw inferences about individuals within a population
- It is an exploratory approach, which can be verified using subsequent confirmatory approaches
 - DEAP can help conform such analyses to best standards and practices through pre-registered reports, reproducibility, and independent validation

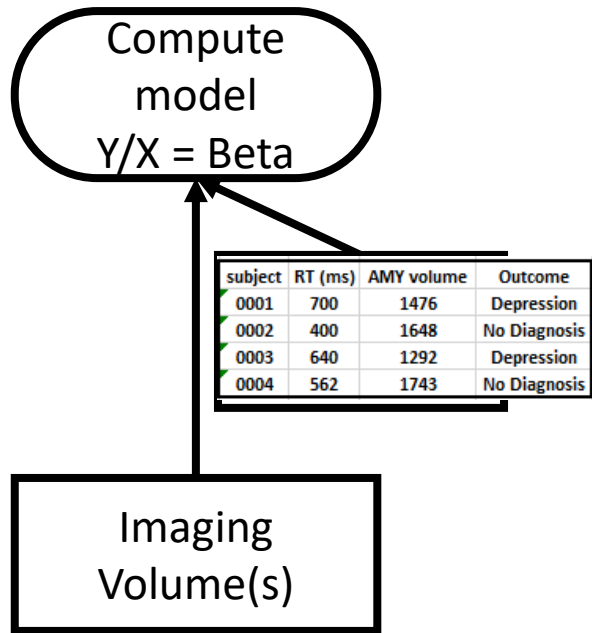
Bryan Gillaume's and Tom Nichols implemented an approach that uses a sandwich estimator to solve a marginal model



Marginal models are effectively linear, so we first estimate the parameters for our design matrix by dividing the imaging measure (Y) by the design (X)

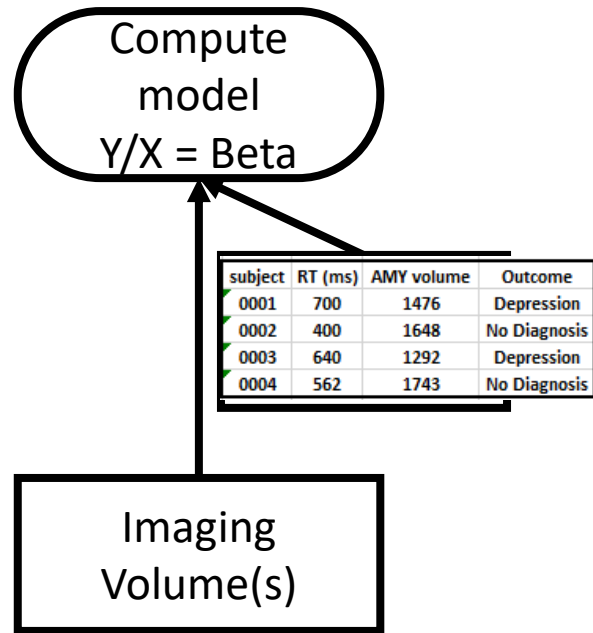


For our software, the design matrix is just your non-imaging data

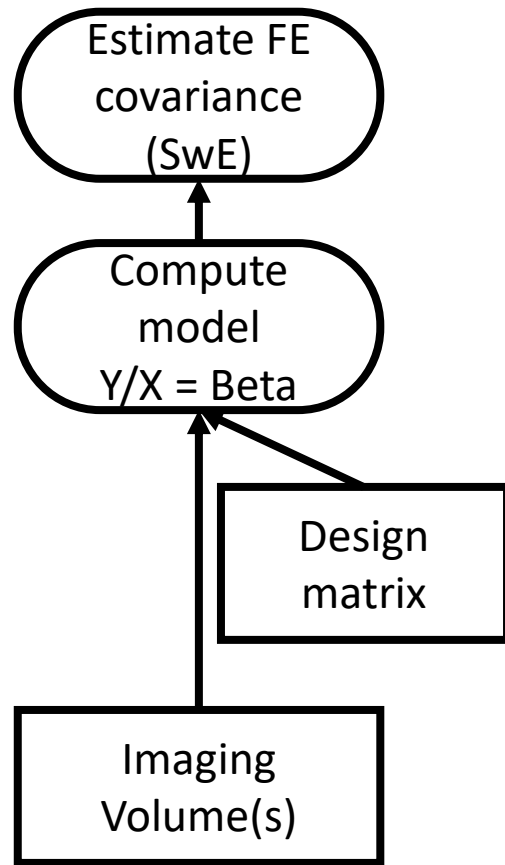


So for example, with the ABCD data we can input measures and test a model

Marginal model: $y \sim RT$

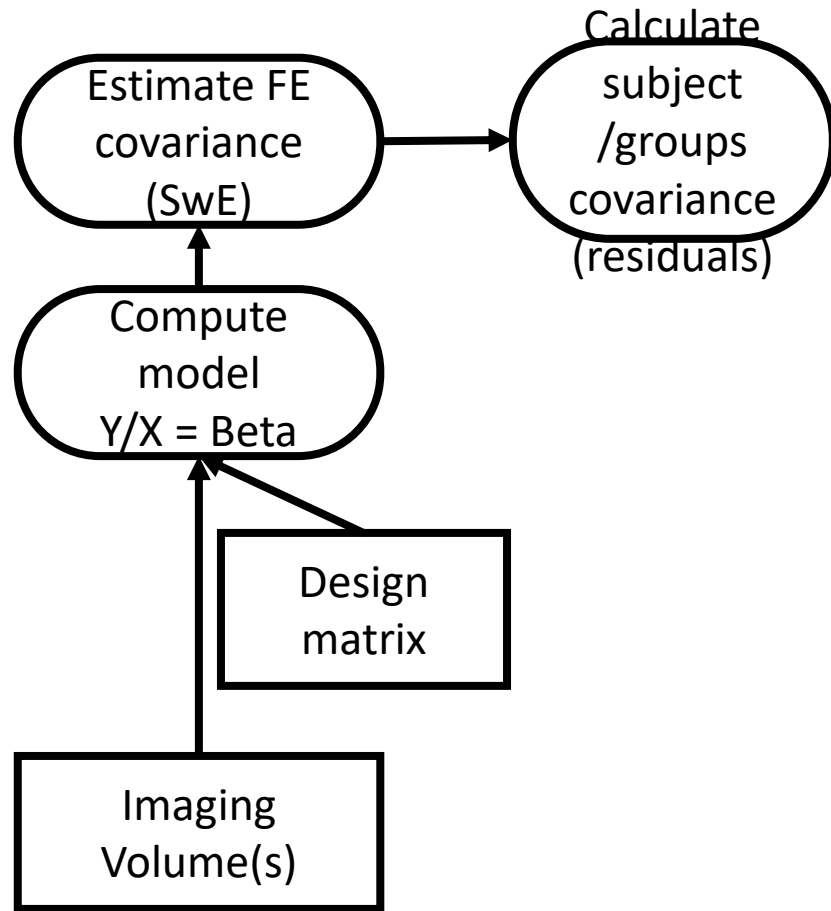


A sandwich estimator is used to estimate covariance and determine the fixed effects parameters

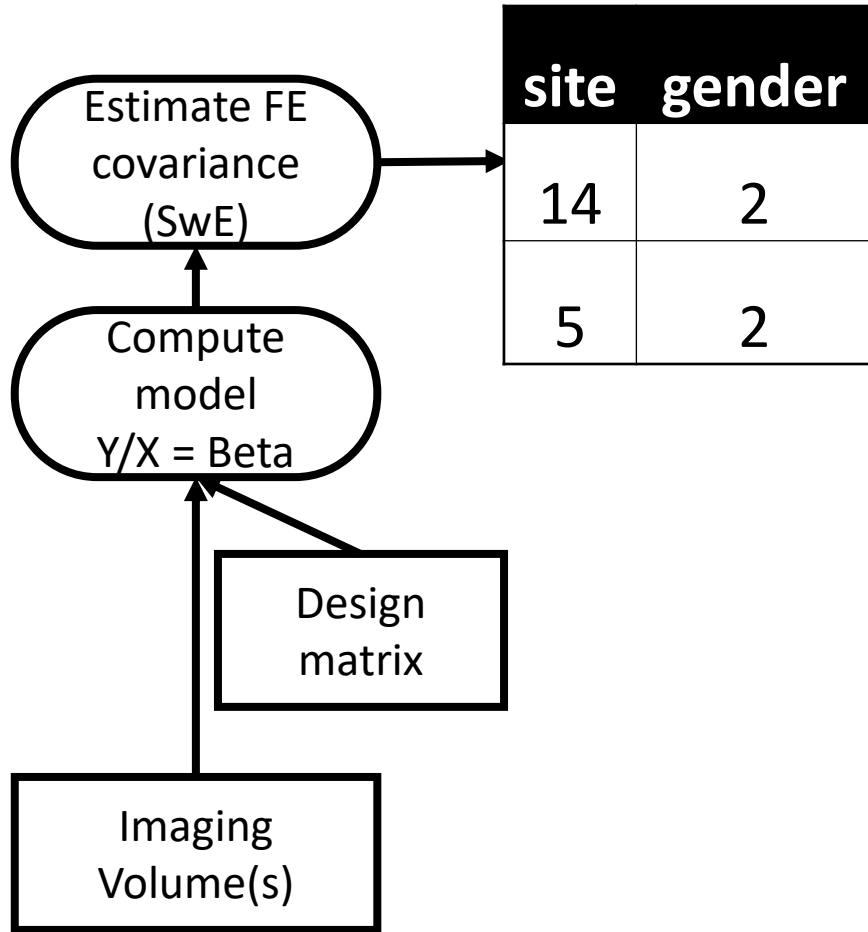


$$S = \underbrace{\left(\sum_{i=1}^m X_i' W_i X_i \right)^{-1}}_{\text{Bread}} \underbrace{\left(\sum_{i=1}^m X_i' W_i \hat{V}_i W_i X_i \right)}_{\text{Meat}} \underbrace{\left(\sum_{i=1}^m X_i' W_i X_i \right)^{-1}}_{\text{Bread}},$$

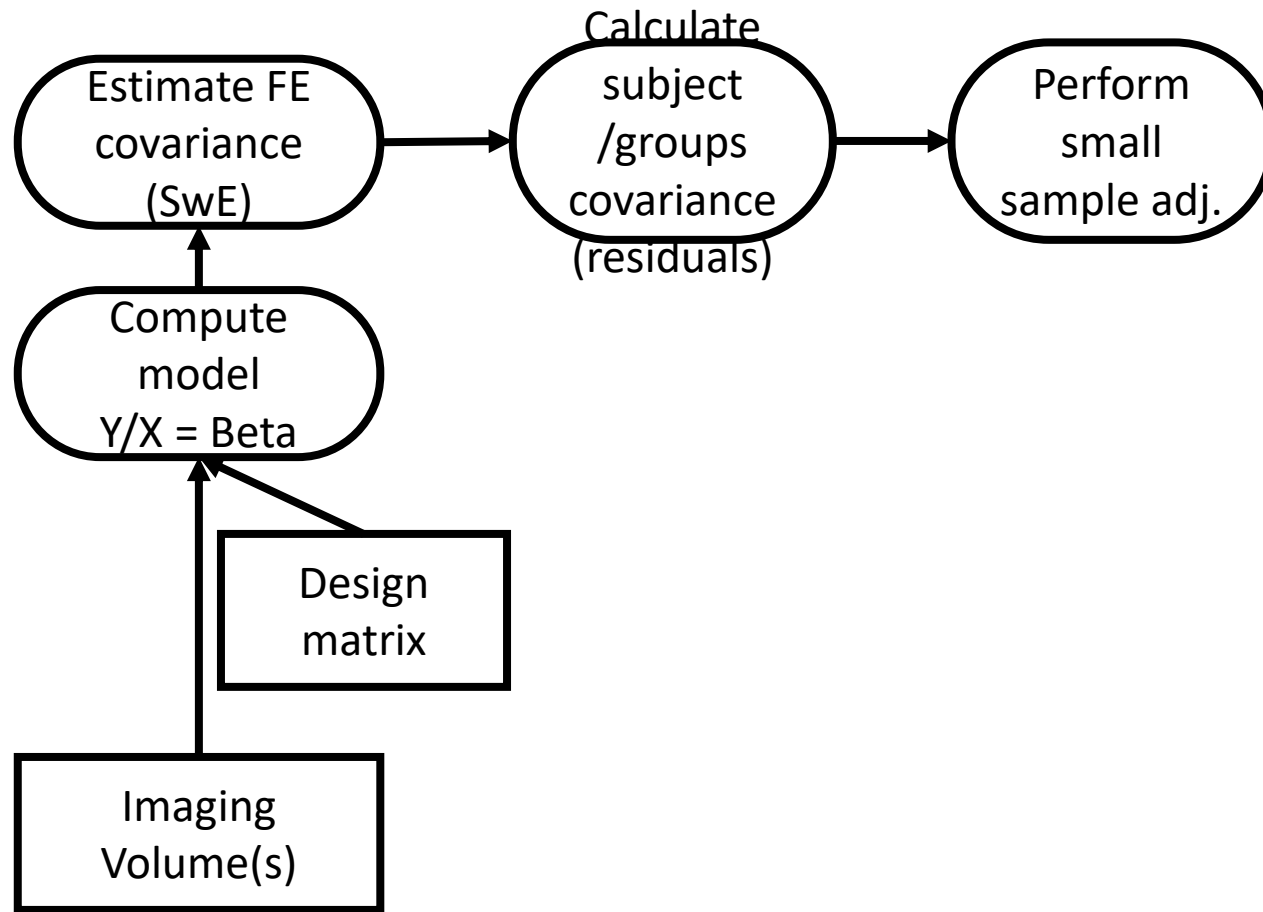
To handle nested structure, group covariance can be calculated separately (CRITICAL FOR ABCD)



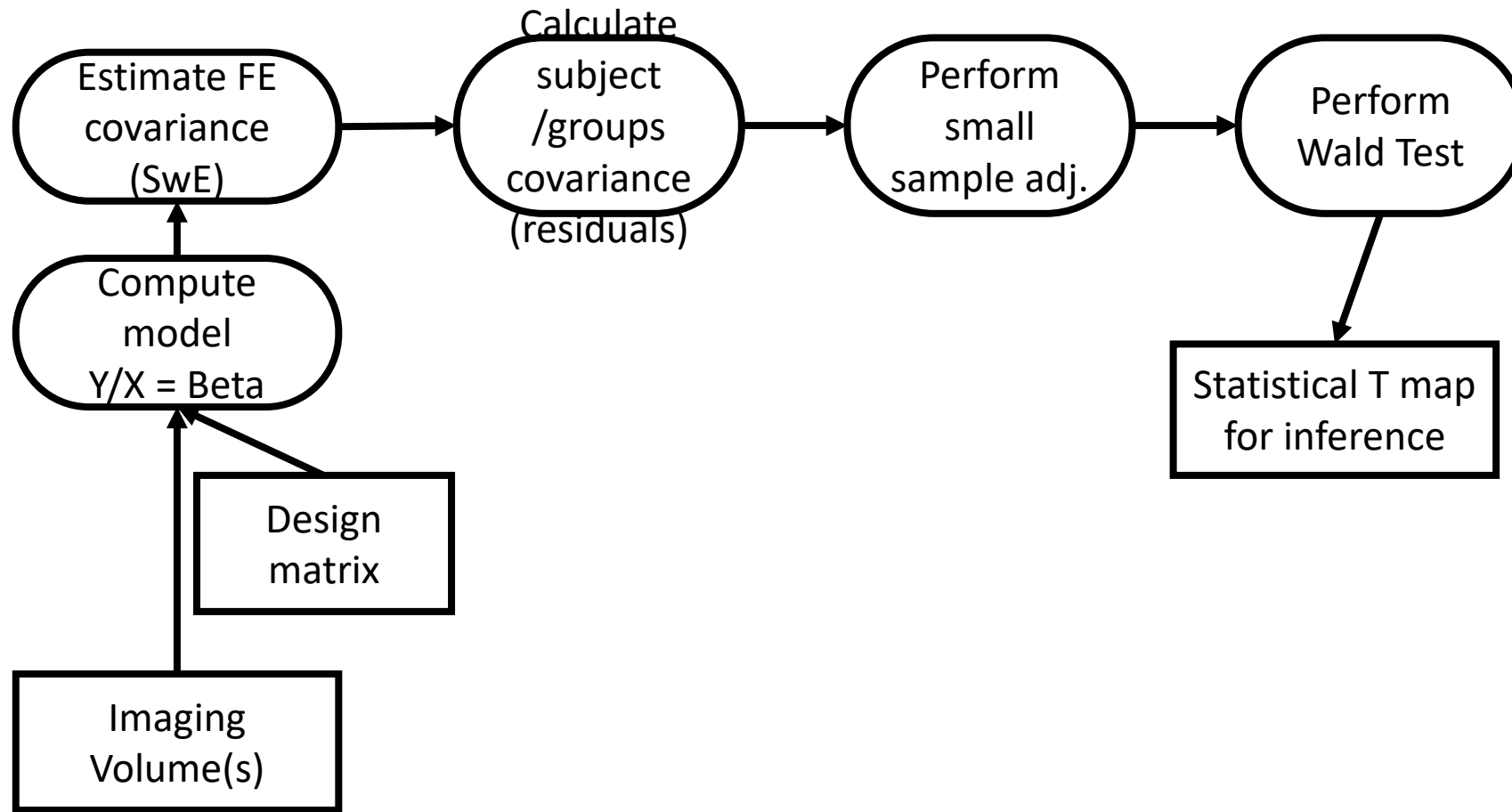
For ABCD, it is good to control for site and gender



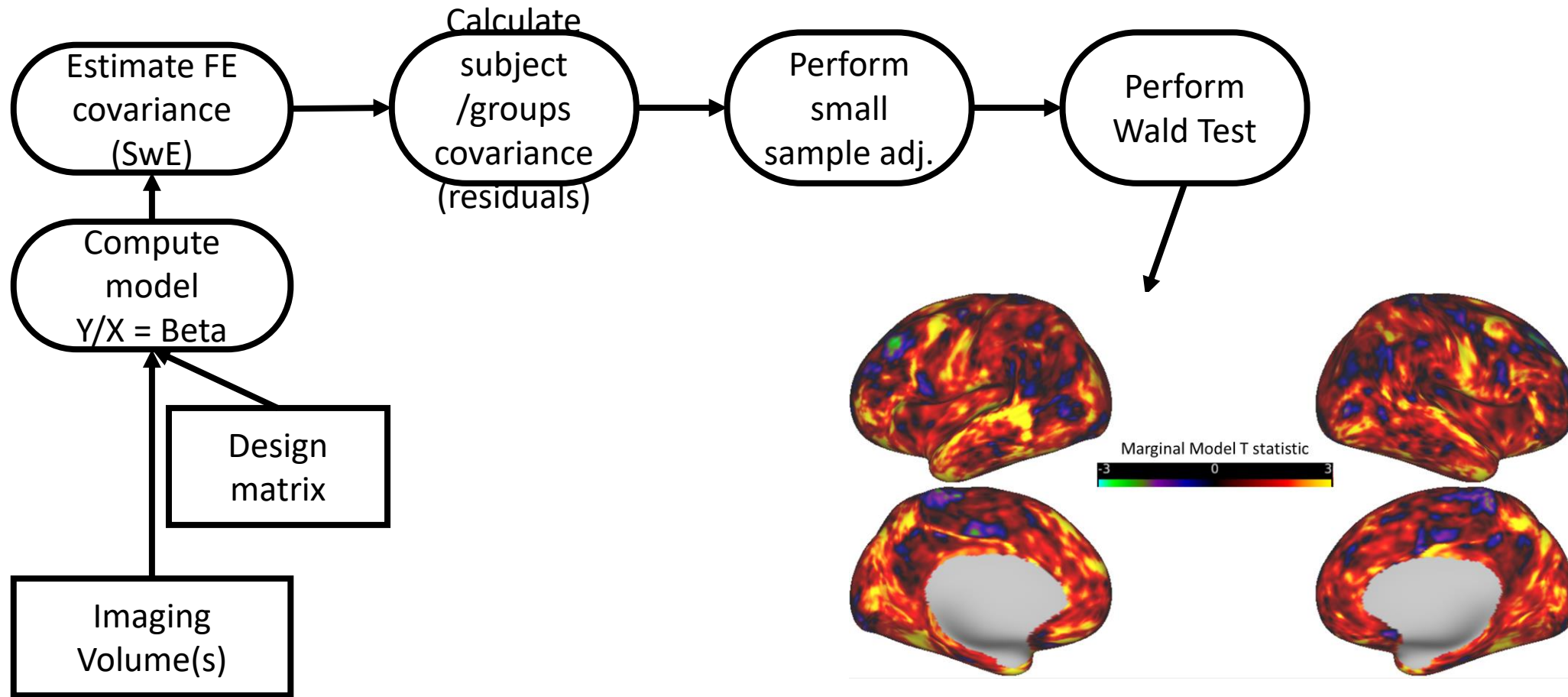
If needed we can perform a small sample size adjustment – this may be important if we used family as a nesting variable



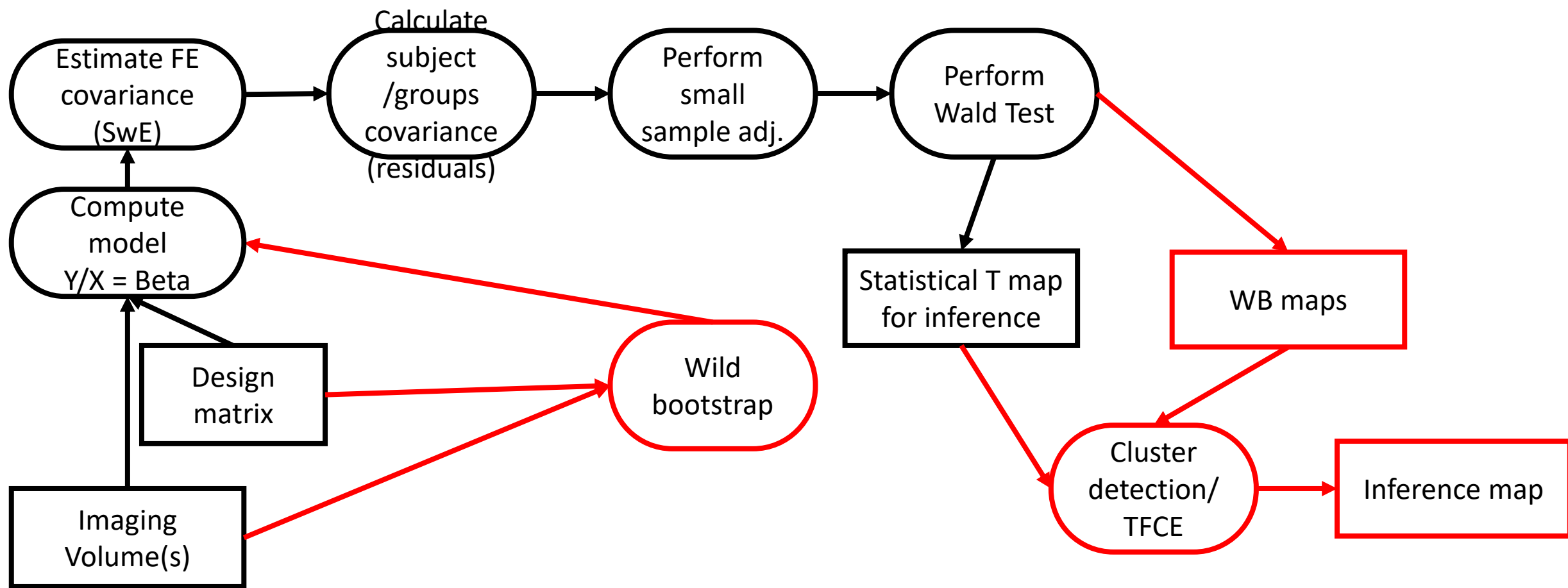
Finally, a Wald test extracts a t-map for statistical inference



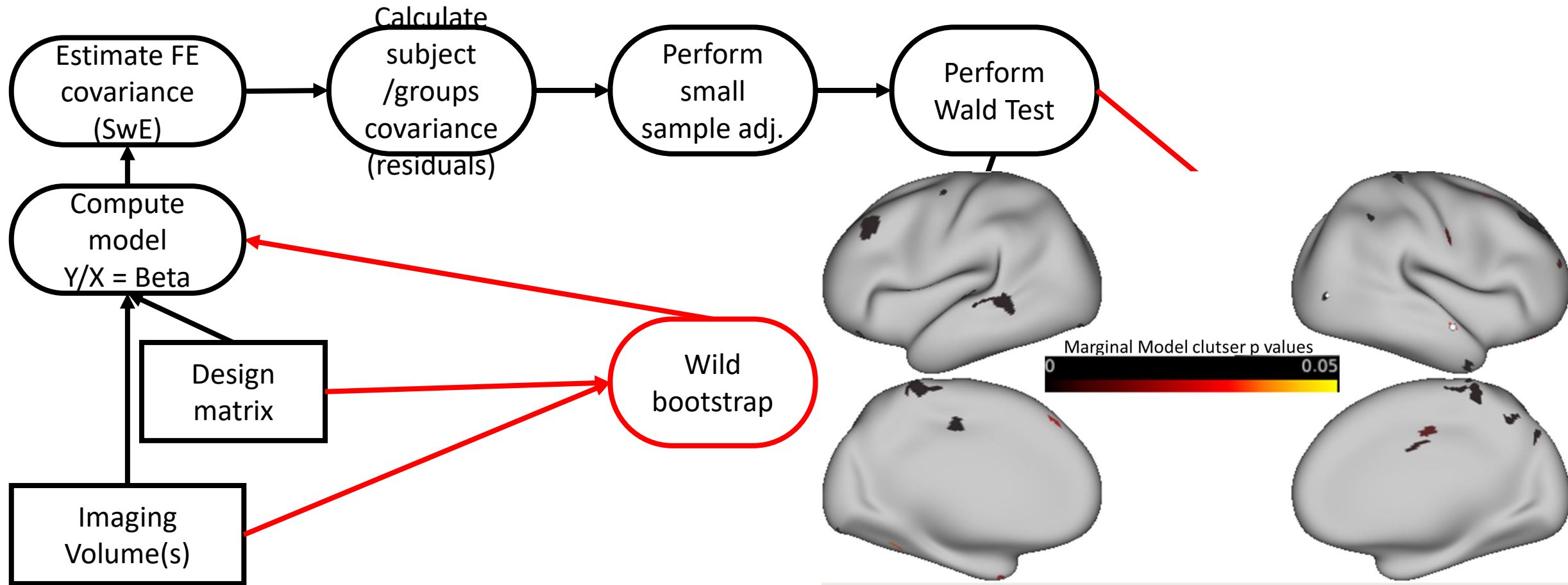
The statistical map looks like this



Use of a wild bootstrap enables inference similar to a permutation test – so we can control for the FWER



Such a test allows us to detect significant clusters



Wild bootstrap

- $WB_value = fitted_value + residual_value * sample_value$
- Sample with replacement can be from simple or complex distributions:
 - Radenbacher (-1, 1) would mean we either:
 - $WB_value = fitted_value - residual_value$
 - $WB_value = fitted_value + residual_value$
- However, LOTS of possible distributions, so choice of distribution is important.

We have begun to implement a standalone MarginalModelCifti package in R

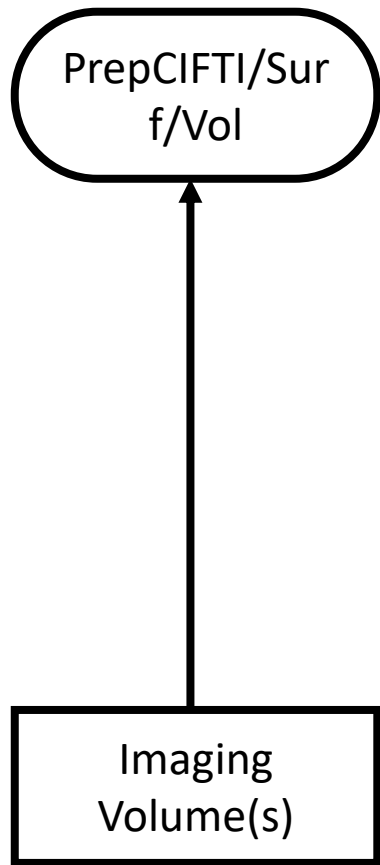
 README.md

MarginalModelCifti

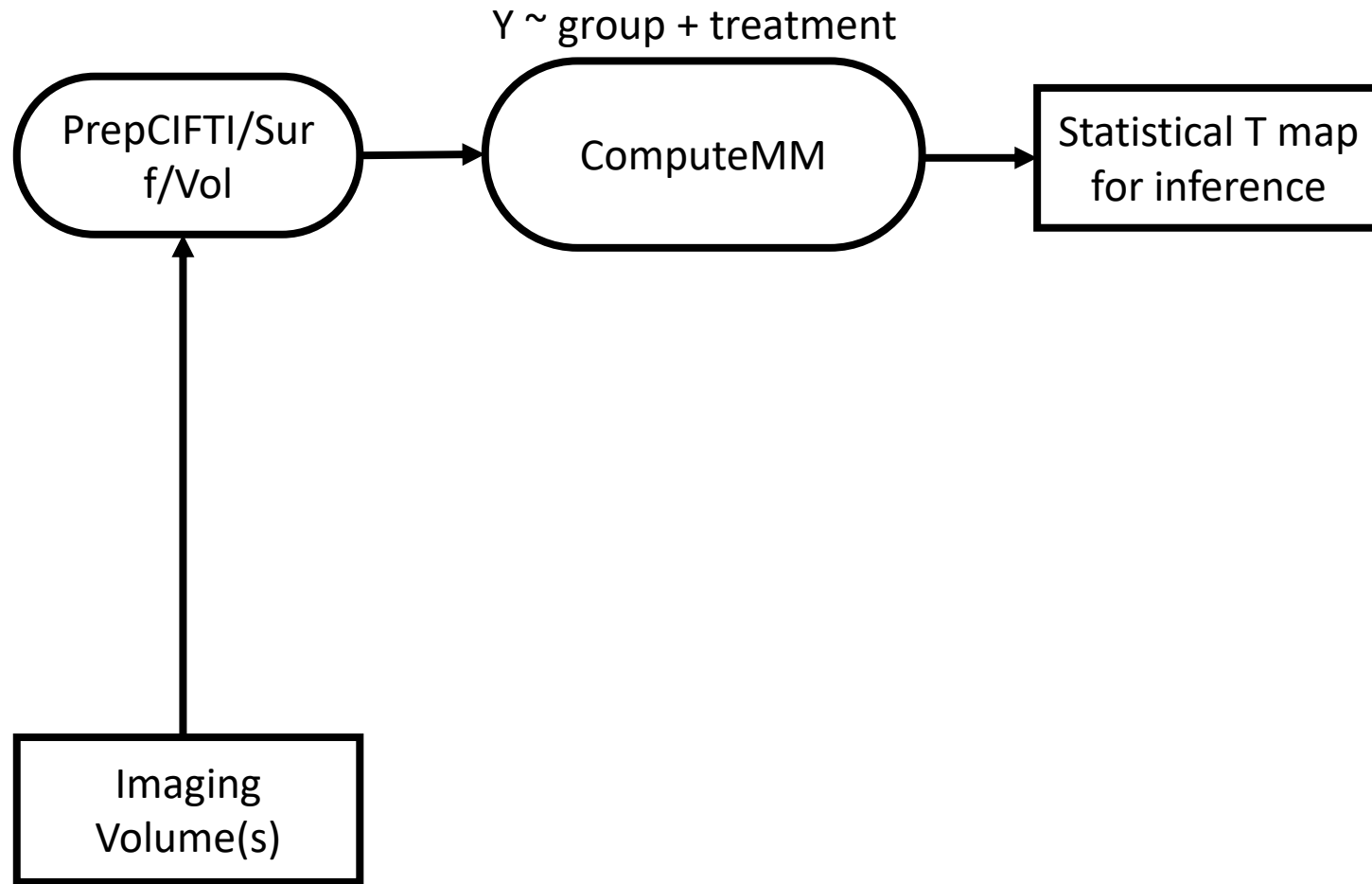
The goal of MarginalModelCifti is to perform marginal models on CIFTI processed datasets. The package contains a single main function that runs multiple subfunctions. Advanced users can use the subfunctions to construct their own analytic pipeline. However, this is not recommend for beginning users.

Alpha version will be released at -- <http://github.com/dcan-labs/MarginalModelCifti>

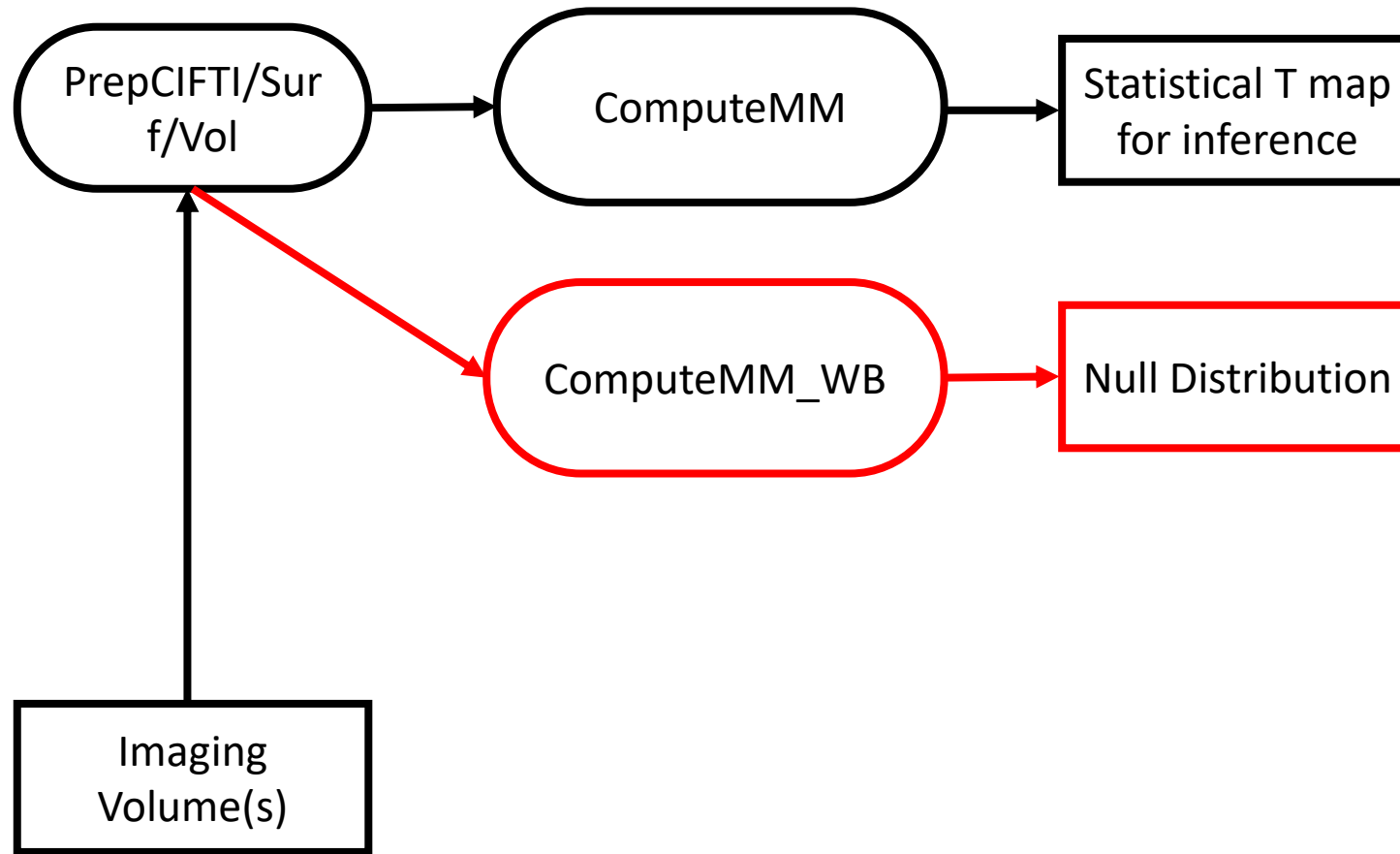
The main wrapper for MarginalModelCifti takes in imaging volumes and prepares them for analysis



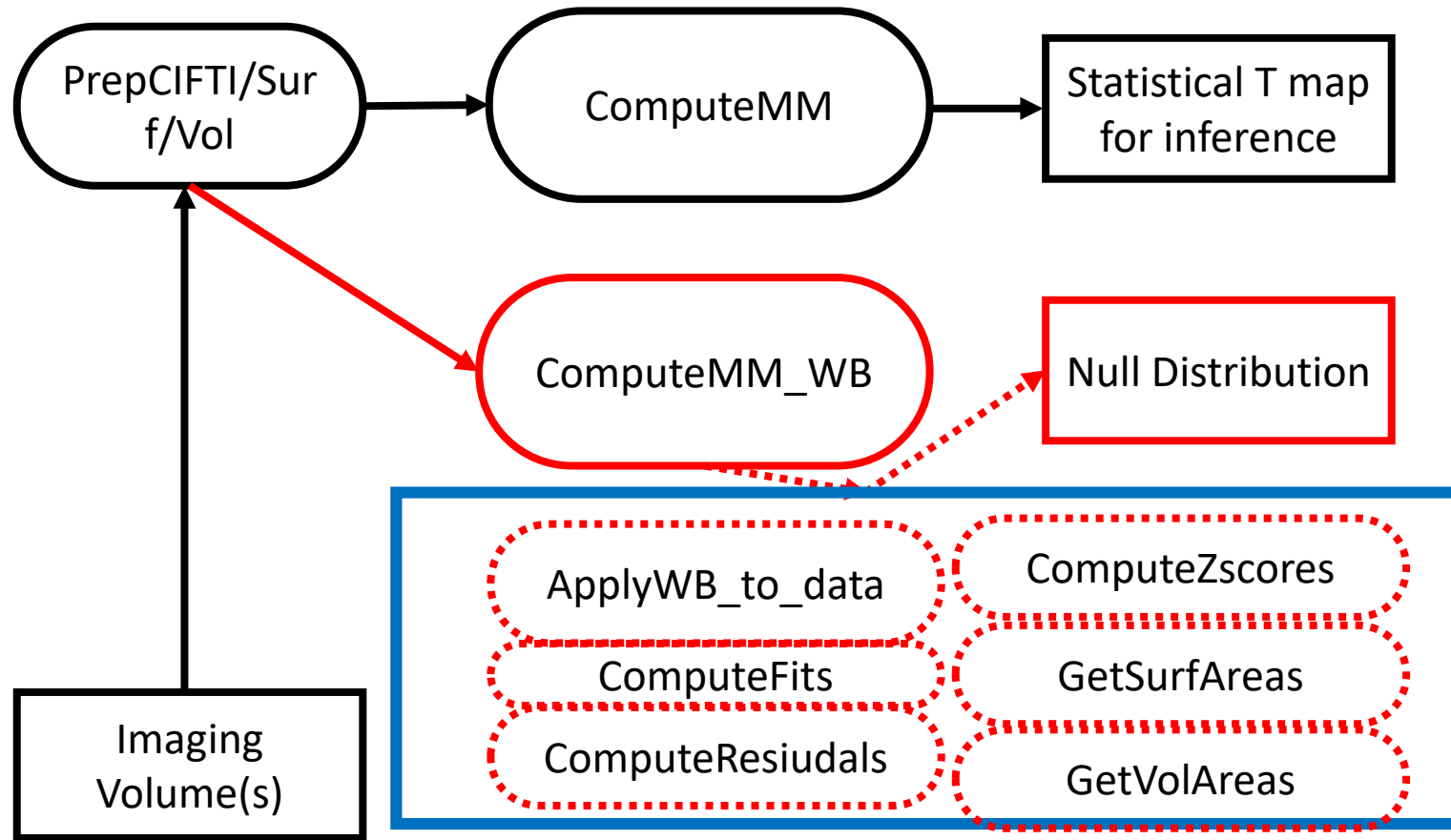
ComputeMM is applied to the prepared data; user specifies the model using Wilkinson notation and wraps the SwE and Wald Test using Geepack



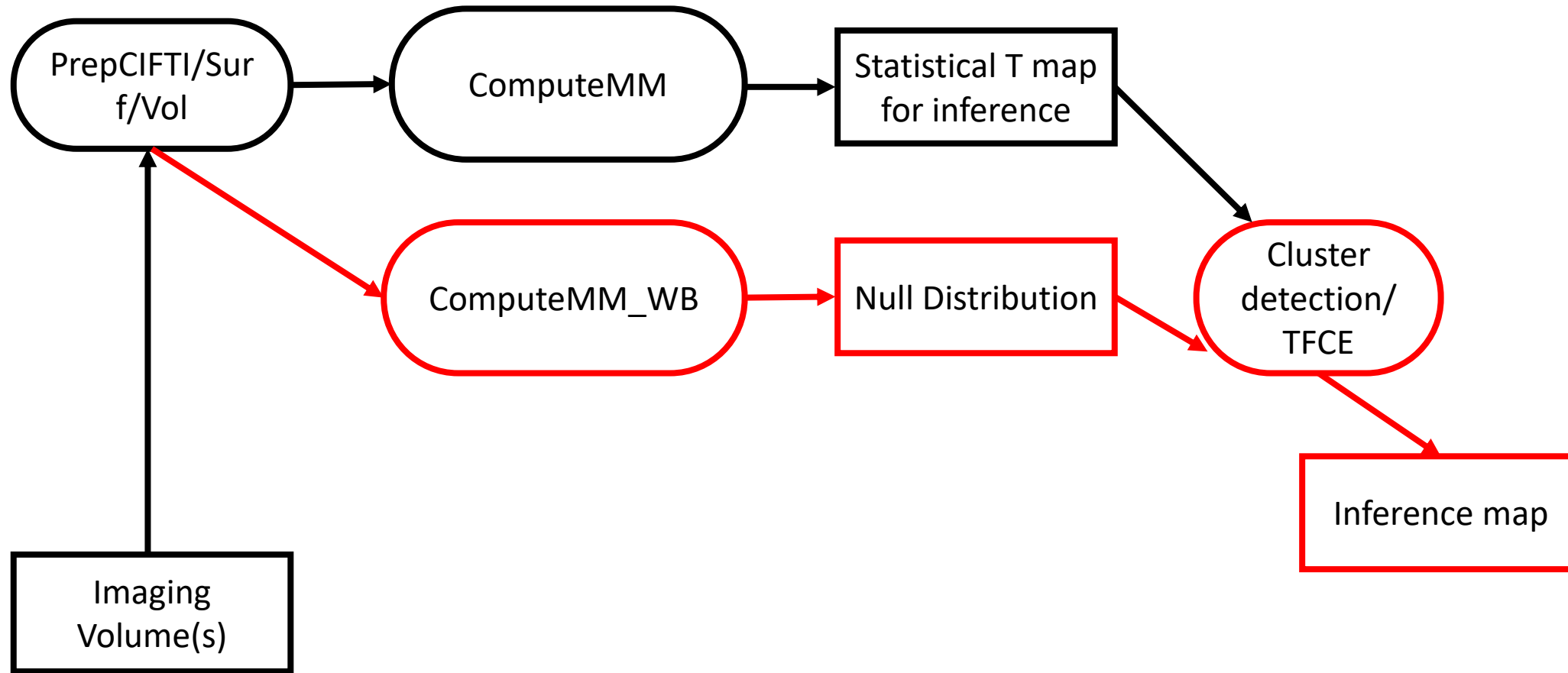
ComputeMM_WB generates the WB maps used to draw inferences about the T map



In turn a family of functions are used to parallelize ComputeMM_WB



Cluster detection is performed within the main wrapper, using information from both processes



ApplyWB_to_data.R

ComputeFits.R

ComputeMM.R

ComputeMM_WB.R

ComputeResiduals.R

ComputeZscores.R

GetSurfAreas.R

GetVolAreas.R

PrepCIFTI.R

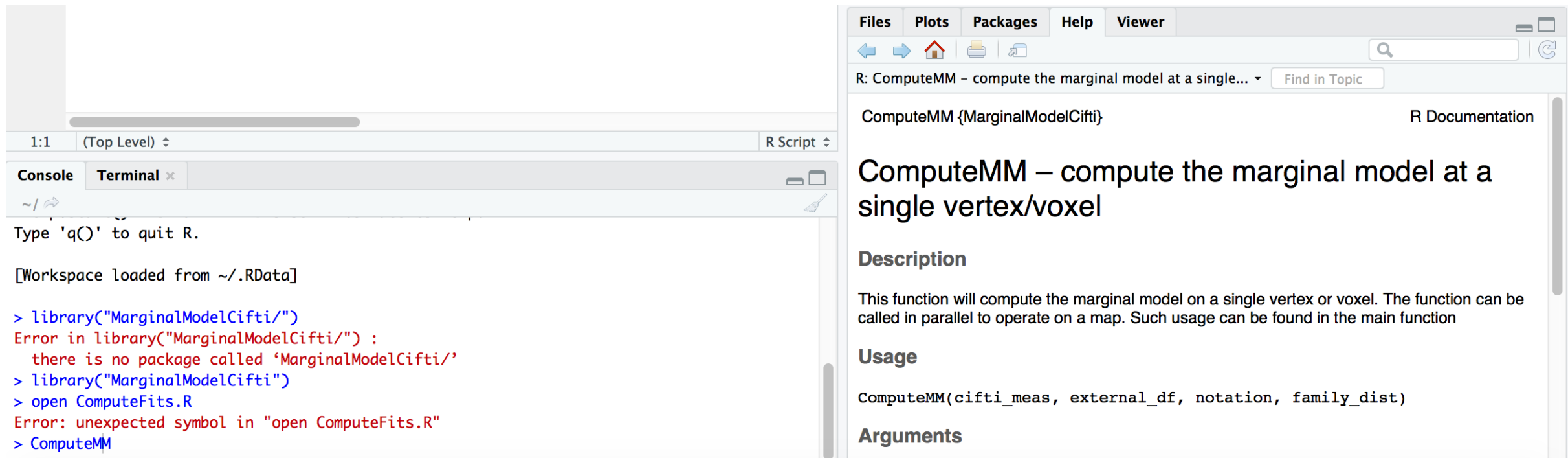
PrepSurf.R

PrepSurfMetric.R

PrepVolMetric.R

The MarginalModelCifti package
comprises multiple functions that can
be accessed by anyone

Functions are documented in accordance with CRAN guidelines



The image shows two windows from the R Studio environment. The left window is the R console, and the right window is the R Documentation viewer.

R Console:

```
1:1 (Top Level) R Script
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> library("MarginalModelCifti/")
Error in library("MarginalModelCifti/") :
  there is no package called 'MarginalModelCifti/'
> library("MarginalModelCifti")
> open ComputeFits.R
Error: unexpected symbol in "open ComputeFits.R"
> ComputeMM
```

R Documentation Viewer:

Files Plots Packages Help Viewer

R: ComputeMM – compute the marginal model at a single... Find in Topic

ComputeMM {MarginalModelCifti} R Documentation

ComputeMM – compute the marginal model at a single vertex/voxel

Description

This function will compute the marginal model on a single vertex or voxel. The function can be called in parallel to operate on a map. Such usage can be found in the main function

Usage

```
ComputeMM(cifti_meas, external_df, notation, family_dist)
```

Arguments

Here are all the parameters for ConstructMarginalModel()

```
1 external_df="/mnt/rose/shared/projects/ABCD/avg_pconn_maker/cordova_analysis_margmod_pcs/gp1_10min_pconn.csv"
2 concfile="/mnt/rose/shared/projects/ABCD/avg_pconn_maker/cordova_analysis_margmod_pcs/group1_10min.conc"
3 structtype="pconn"
4 structfile=NULL
5 matlab_path="/mnt/max/shared/code/external/utilities/Matlab2016bRuntime/v91"
6 surf_command="/mnt/max/shared/projects/FAIR_users/Feczko/code_in_dev/SurfConnectivity/"
7 wave = "/mnt/rose/shared/projects/ABCD/avg_pconn_maker/cordova_analysis_margmod_pcs/gp1_marg_nested.csv"
8 notation = formula(y~pc2_new)
9 constr="independence"
10 family_dist="gaussian"
11 dist_type="radenbacher"
12 z_thresh = 2.3
13 nboot=4
14 p_thresh=0.05|
15 sigtype="enrichment"
16 id_subjects="subjectkey"
17 output_directory="/mnt/rose/shared/projects/ABCD/avg_pconn_maker/cordova_analysis_margmod_pcs/pc2_gp1_test"
18 fastSwE=TRUE
19 adjustment=NULL
20 ncores=4
21 norm_external_data=TRUE
22 norm_internal_data=TRUE
23 marginal_outputs = FALSE
24 marginal_matrix = NULL
25 enrichment_path = "/mnt/max/shared/projects/FAIR_users/Feczko/code_in_dev/CommunityChisquaredAnalysis/"
26 modules = "/mnt/max/shared/projects/FAIR_users/Feczko/code_in_dev/CommunityChisquaredAnalysis/gordon_modules.csv"
27 wb_command = "/usr/local/bin/wb_command"
28
```

To make things easier – we've made a jupyter notebook that can be used as a reference

jupyter MarginalModelCifti_LH_analysis Last Checkpoint: 02/01/2019 (autosaved)



Logout

Control Panel

File Edit View Insert Cell Kernel Help

Trusted

R

File Edit View Insert Cell Kernel Help Run Markdown

```
2) make a directory for the MarginalModelCifti package mkdir MarginalModelCifti
3) enter the directory cd MarginalModelCifti
4) clone the MarginalModelCifti repository git clone https://gitlab.com/Fair_lab/marginalmodelcifti.git ./
5) return to your initial home directory cd ..
5) Type R
6) After a prompt appears, make sure devtools is installed by typing install.packages("devtools")
7) Load devtools library(devtools)
8) install the MarginalModelCifti package install("MarginalModelCifti/")

NOTE: You may also want to clone the SurfConnectivity package, in case you do not have access to it.
a) open a new terminal on exacloud
b) make a directory for SurfConnectivity mkdir SurfConnectivity
c) go into SurfConnectivity folder cd SurfConnectivity
d) clone the SurfConnectivity repository here git clone https://gitlab.com/Fair_lab/surfconnectivity.git ./
```

Call the MarginalModelCifti library -- if this errors you will need to install it using devtools

```
In [1]: library(MarginalModelCifti)
```

Set your projects folder, which is where you plan to run the analysis, then go to the folder

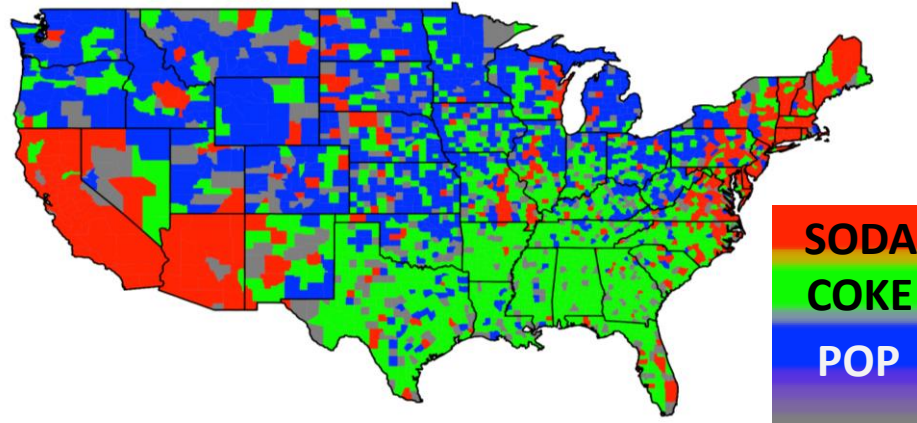
```
In [2]: projectsfolder="/home/exacloud/lustrel/fnl_lab/projects/marginalmodelciftitest"
```

Outline of talk

- Theory recap: modelling approaches can be reduced to two types: predictive and descriptive
- “Big data” complicates our ability to apply both approaches
- Marginal Modelling is a good approach for descriptive modelling
- **Functional Random Forests is a good approach for predictive modelling**
- Other approaches can also handle big data, but are beyond the scope of this workshop

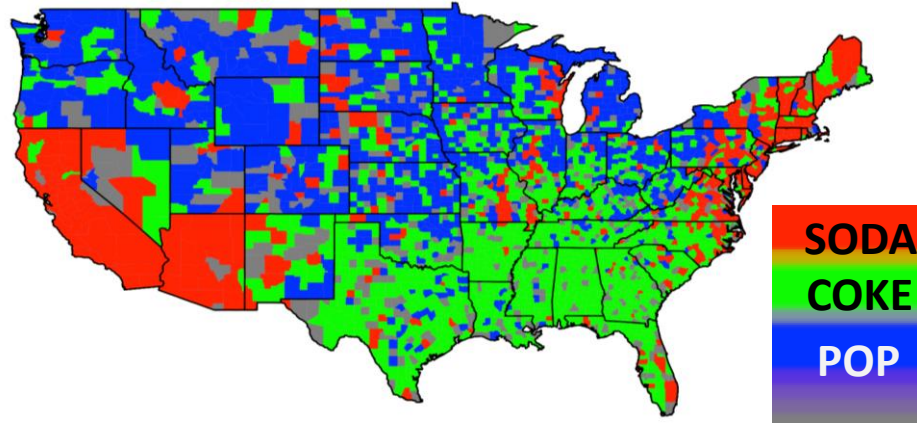
Nested structures -- people belong to multiple subtypes

Dialect preferences: soda, coke or pop?

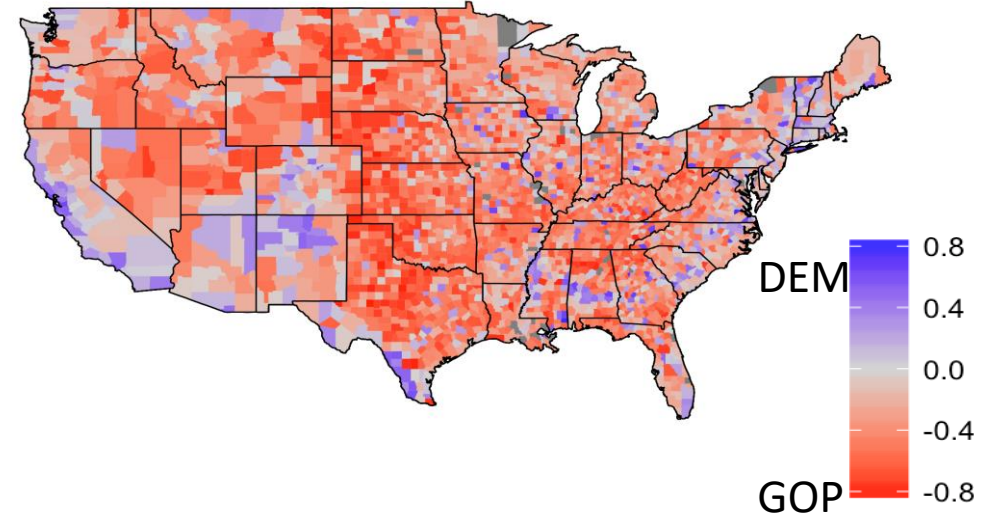


Nested structures -- people belong to multiple subtypes

Dialect preferences: soda, coke or pop?

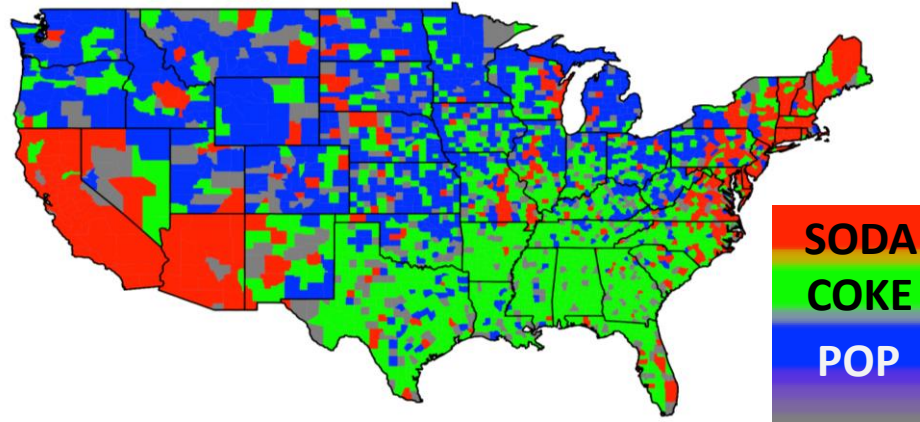


U.S. 2016 presidential election voting preferences

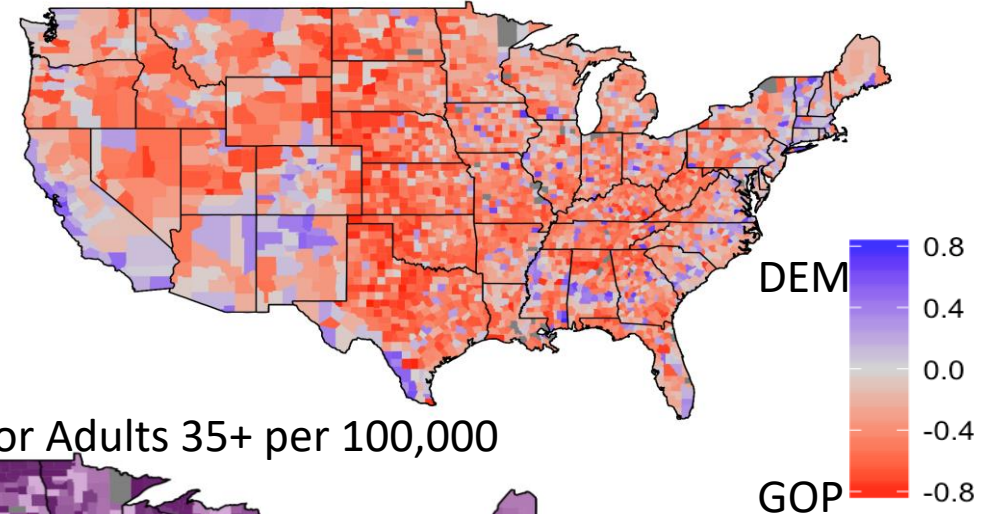


Nested structures -- people belong to multiple subtypes

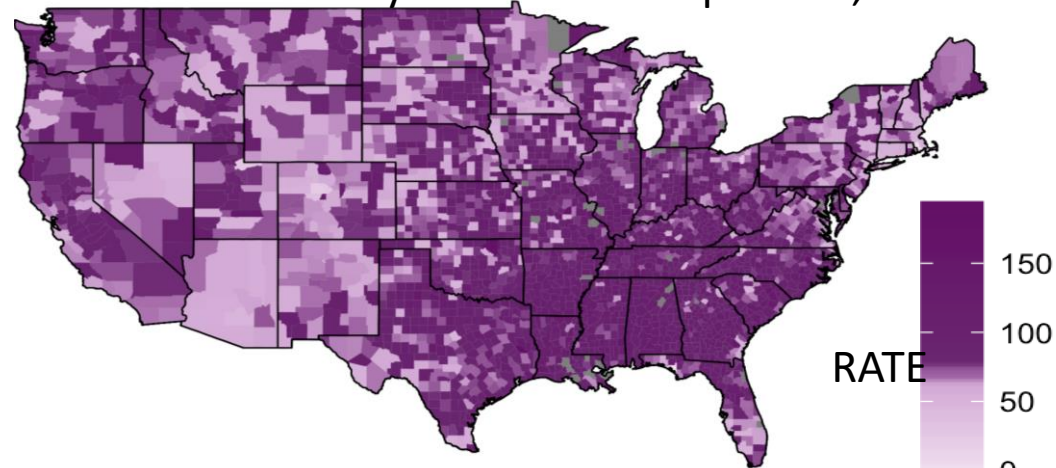
Dialect preferences: soda, coke or pop?



U.S. 2016 presidential election voting preferences



Stroke mortality for Adults 35+ per 100,000



But what about effects of scanner upgrades, software maintenance, or even changes in personnel?

If we want to control for unknown structure, we need to identify subtypes tied to an outcome

- **Supervised** approaches can confirm known subtypes but not discover unknown subtypes tied to an outcome

If we want to control for unknown structure, we need to identify subtypes tied to an outcome

- **Supervised** approaches can confirm known subtypes but not discover unknown subtypes tied to an outcome
- **Unsupervised** approaches can discover unknown subtypes, but not tied to any outcome

How does the **Functional Random Forest** work?

Supervised component

Ask a question: can we predict depression diagnosis?

Supervised component

Unsupervised component

We start with an input dataset

Input dataset			
subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis

Supervised component

Unsupervised component

We start with an input dataset

Supervised component

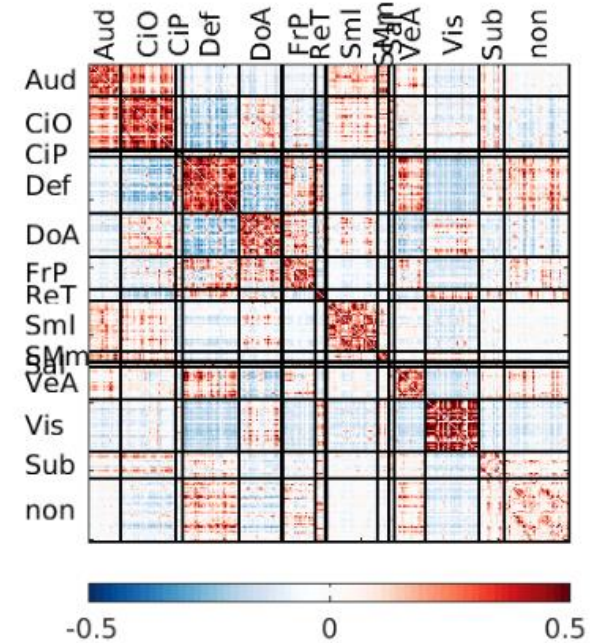
Input dataset			
subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis

Unsupervised component

This dataset can be a functional connectivity matrix

Supervised component

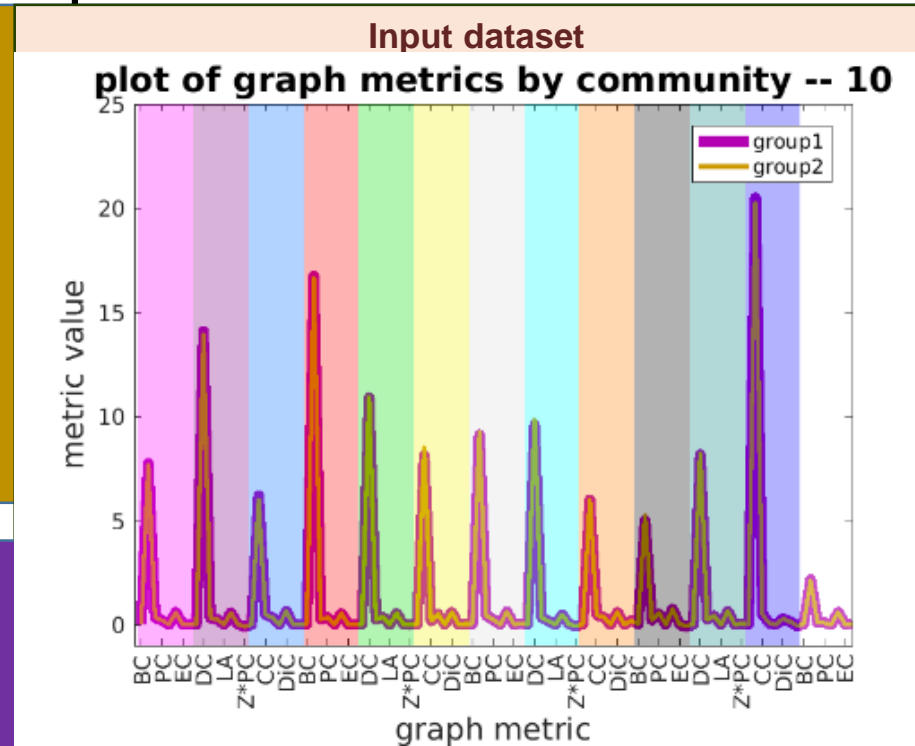
Input dataset



Unsupervised component

This dataset can be a functional connectivity matrix – which gets reduced to either graph metrics or principal components

Supervised component

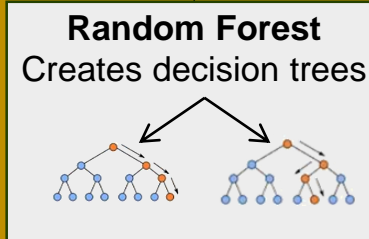


Unsupervised component

Input data are modeled via a random forest via validation/testing

Supervised component

Input dataset			
subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis

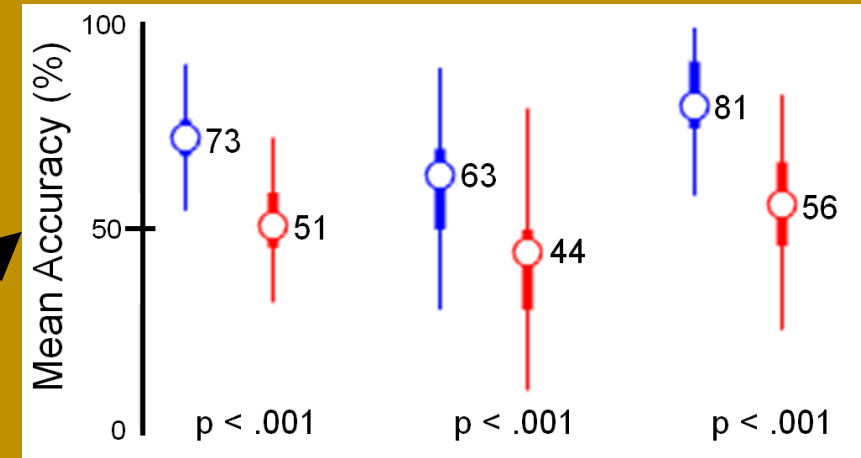
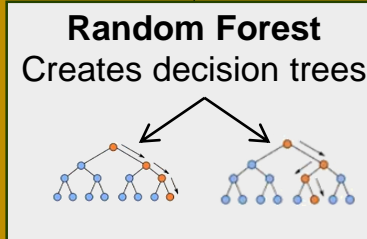


Unsupervised component

Model is supervised because it attempts to predict the outcome of interest

Supervised component

Input dataset			
subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis

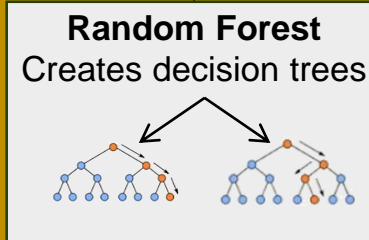


Unsupervised component

If the random forest performs well on independent test data, a similarity matrix is produced from the RFs

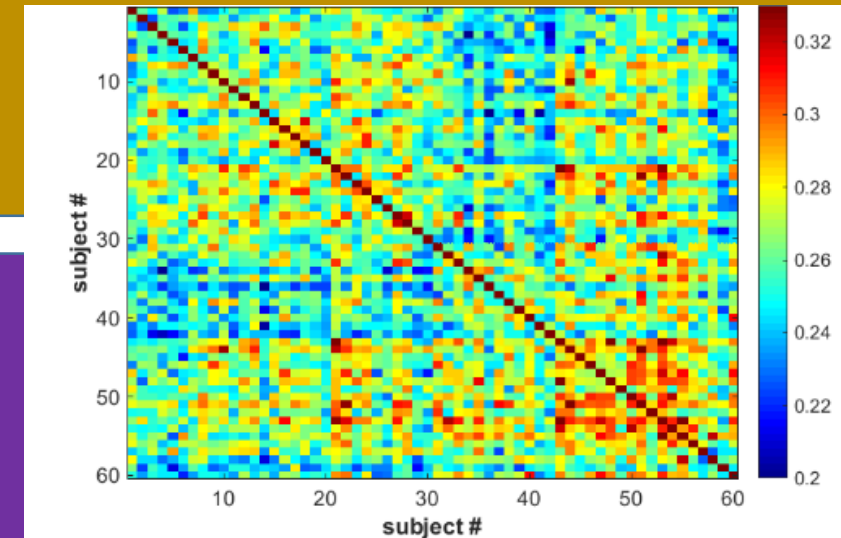
Supervised component

Input dataset			
subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis



Similarity matrix				
subject	0001	0002	0003	0004
0001	1000	291	756	151
0002	291	1000	133	628
0003	756	133	1000	172
0004	151	628	172	1000

=

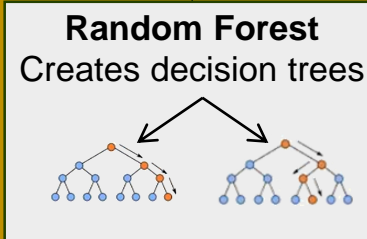


Unsupervised component

Subgroups are identified from this matrix via Infomap

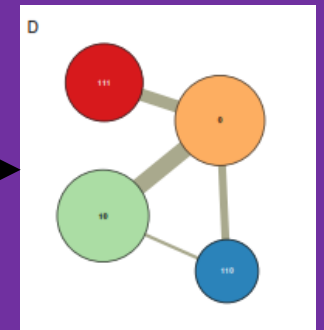
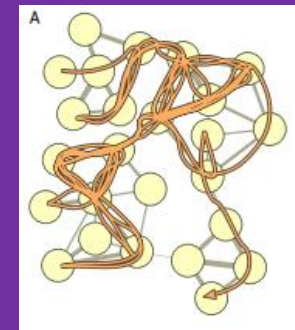
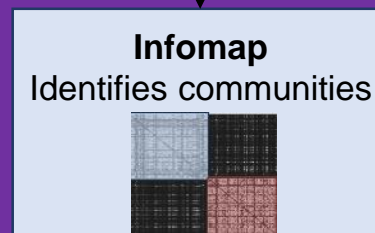
Supervised component

Input dataset			
subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis



Unsupervised component

Similarity matrix				
subject	0001	0002	0003	0004
0001	1000	291	756	151
0002	291	1000	133	628
0003	756	133	1000	172
0004	151	628	172	1000

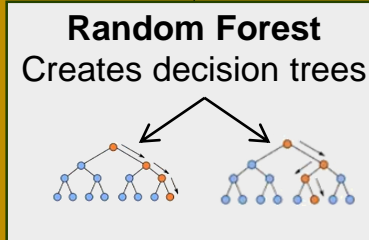


Subtypes arise from the model that are tied to the outcome

Supervised component

Input dataset

subject	RT (ms)	AMY volume	Outcome
0001	700	1476	Depression
0002	400	1648	No Diagnosis
0003	640	1292	Depression
0004	562	1743	No Diagnosis



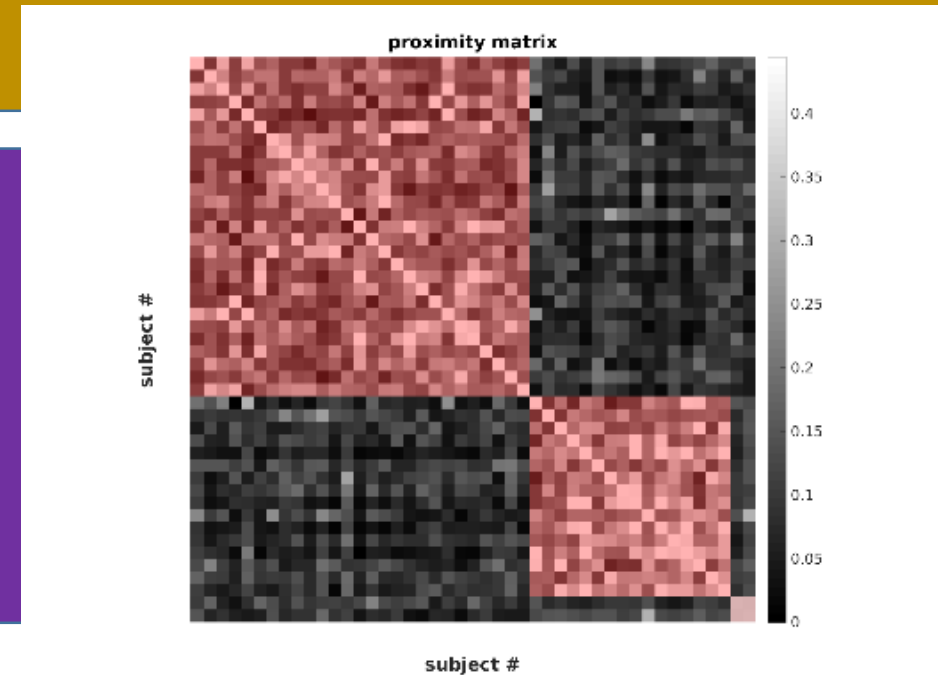
Unsupervised component

Similarity matrix

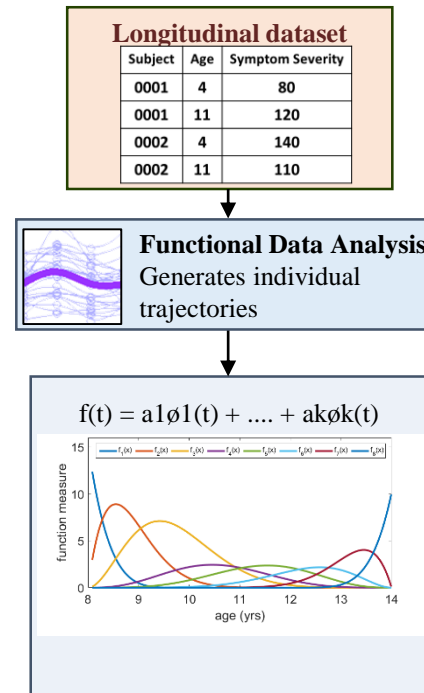
subject	0001	0002	0003	0004
0001	1000	291	756	151
0002	291	1000	133	628
0003	756	133	1000	172
0004	151	628	172	1000



Subpopulations



The FRF can be used to identify trajectories in longitudinal data

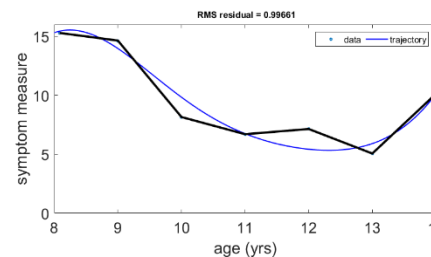
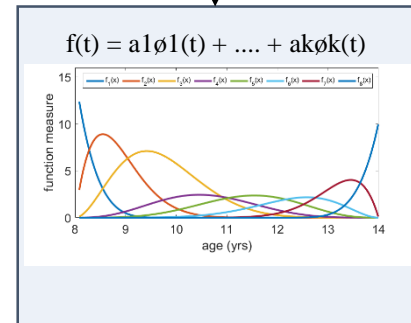
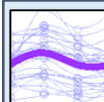


Combining the set of functions estimates a smooth trajectory for an individual's symptoms

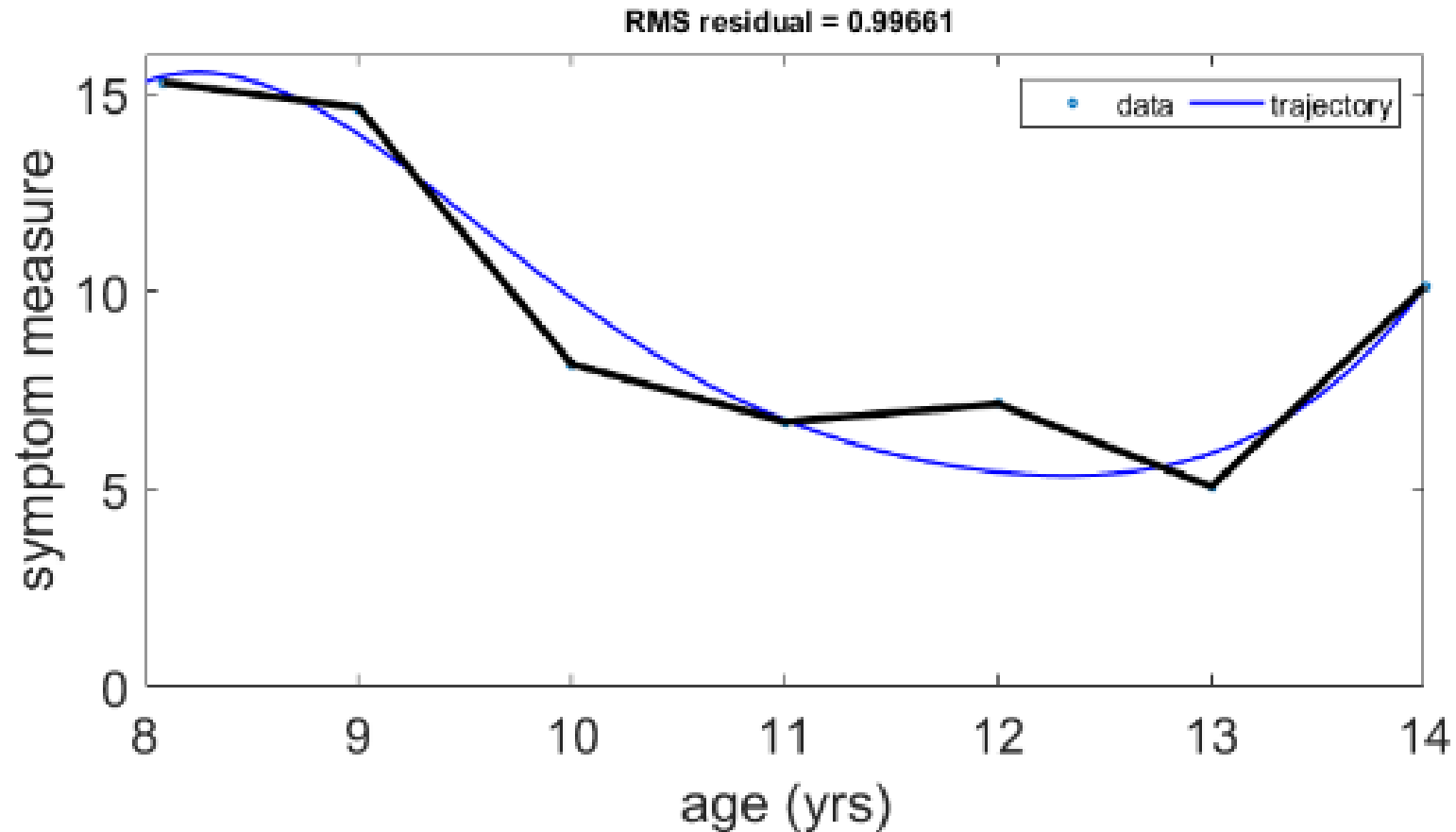
Longitudinal dataset

Subject	Age	Symptom Severity
0001	4	80
0001	11	120
0002	4	140
0002	11	110

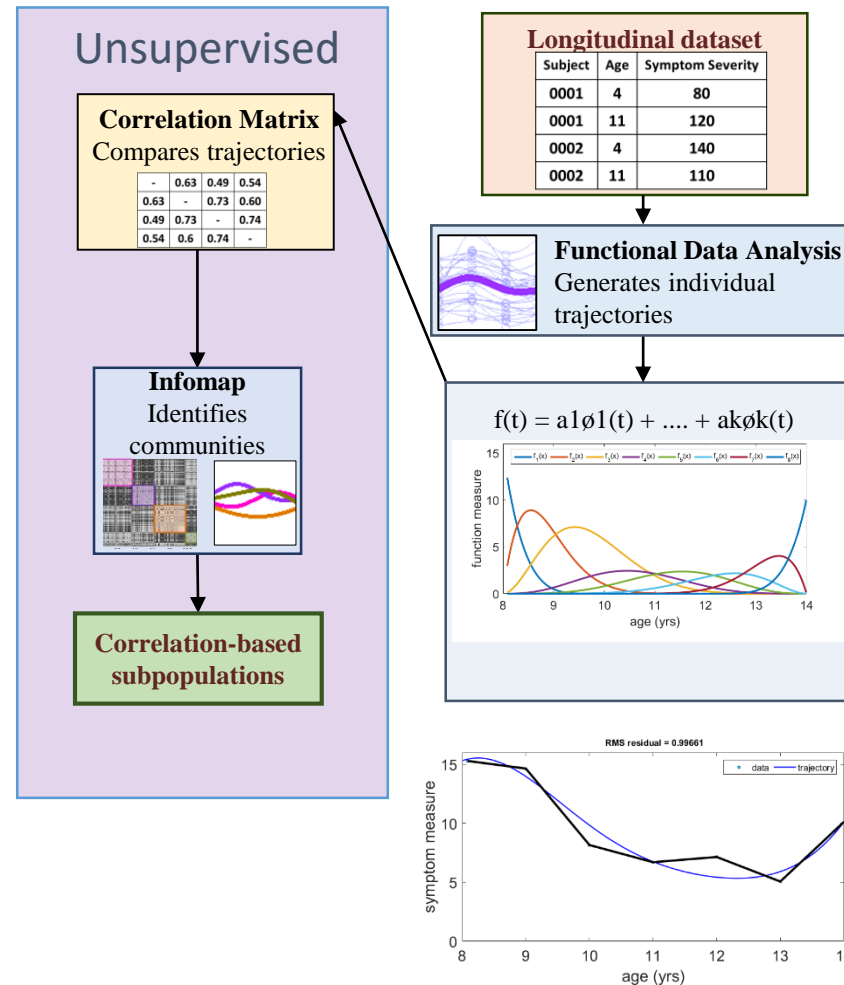
Functional Data Analysis
Generates individual trajectories



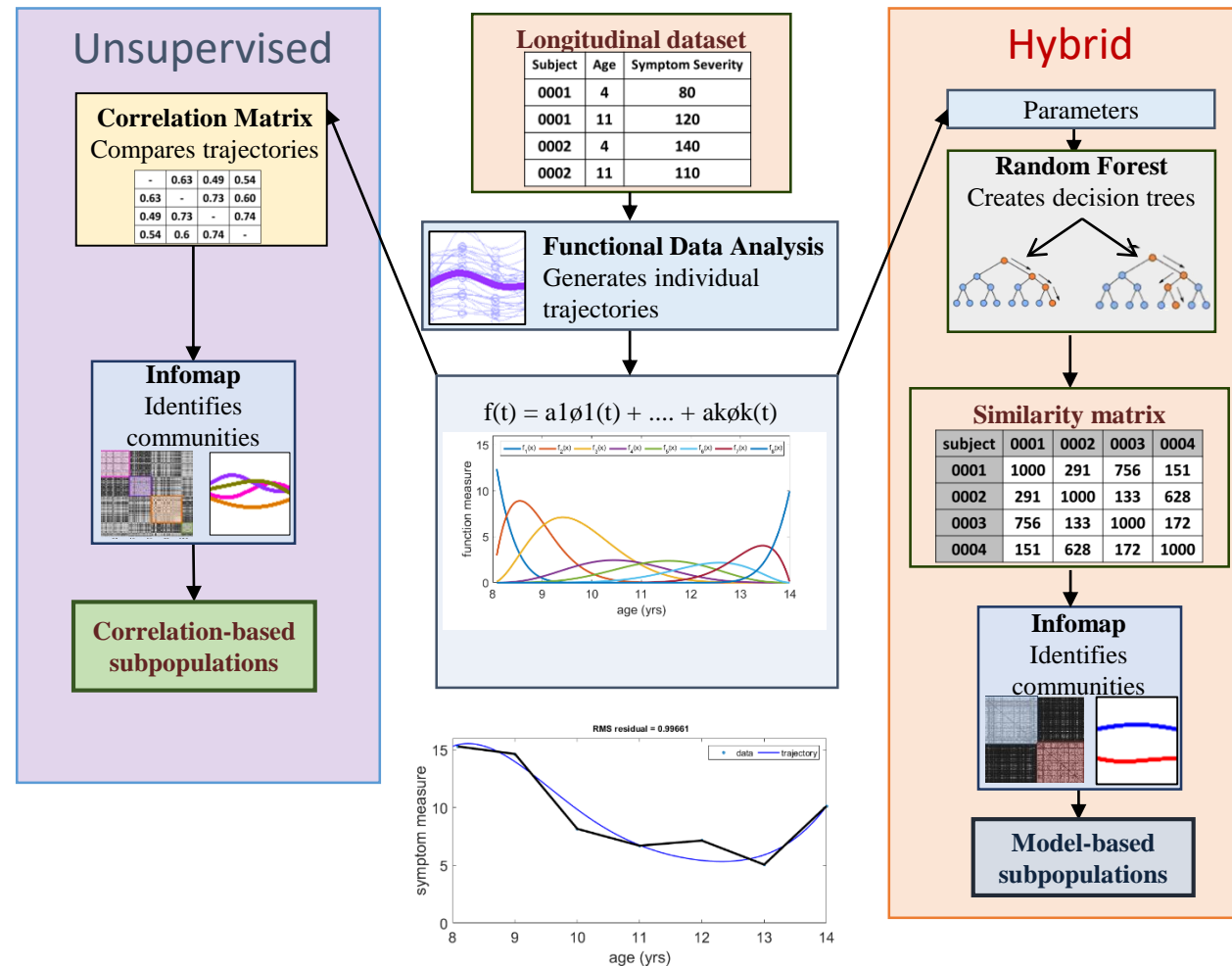
Combining the set of functions estimates a smooth trajectory for an individual's symptoms



We can use an **unsupervised** approach to identify trajectories



Or use a “hybrid” approach that identifies trajectory subtypes tied to an outcome of interest

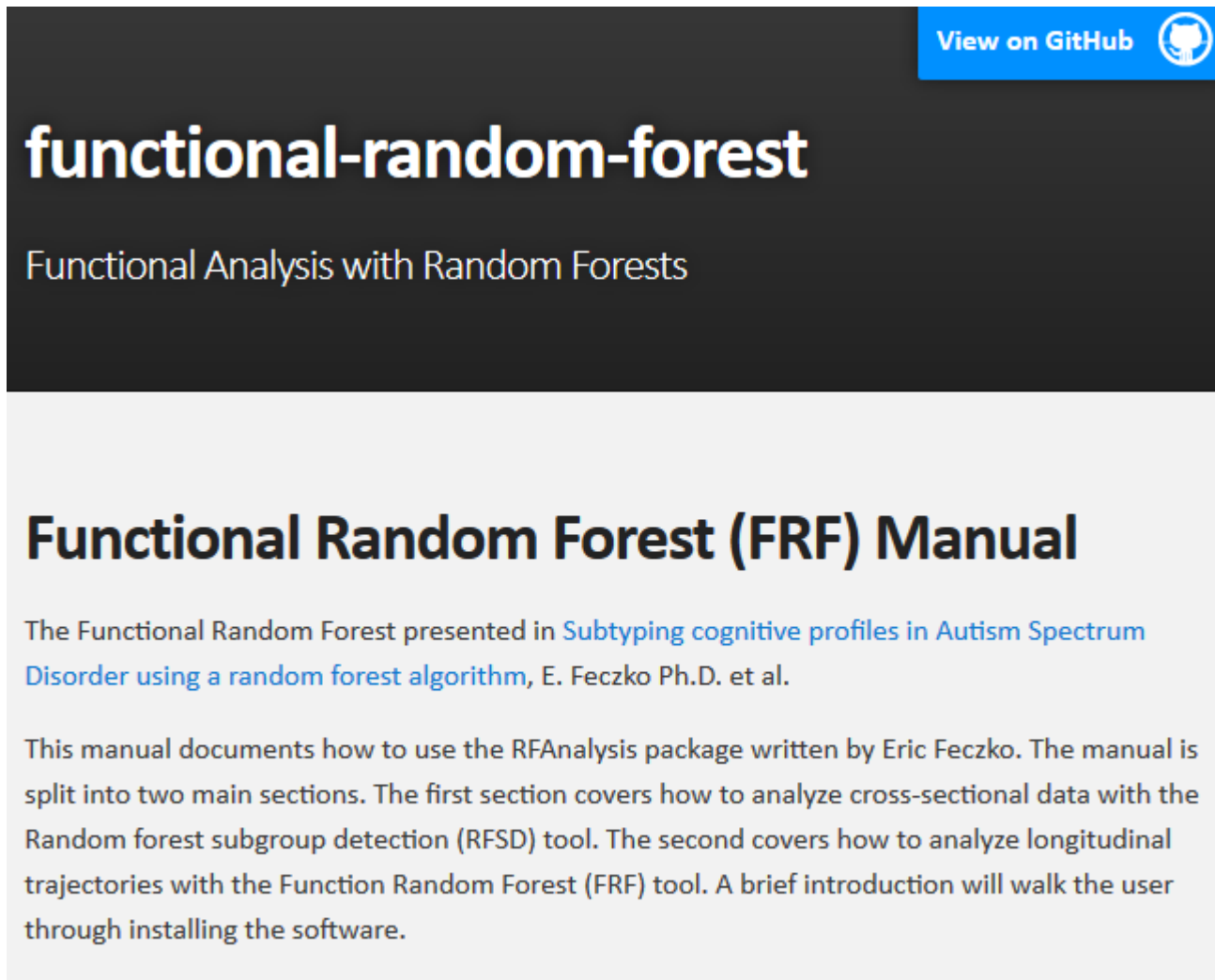


A manual for using the FRF exists online (<https://dcan-labs.github.io/functional-random-forest/>)




The image is a screenshot of the GitHub repository page for 'functional-random-forest'. At the top right, there is a blue button that says 'View on GitHub' with the GitHub logo. Below this, the repository name 'functional-random-forest' is displayed in a large, bold, white font. Underneath the name, the description 'Functional Analysis with Random Forests' is written in a smaller, white font. The main content area has a light gray background and features a large, bold, black heading: 'Functional Random Forest (FRF) Manual'. Below the heading, there is a paragraph of text: 'The Functional Random Forest presented in [Subtyping cognitive profiles in Autism Spectrum Disorder using a random forest algorithm](#), E. Feczko Ph.D. et al.' This is followed by another paragraph: 'This manual documents how to use the RFAAnalysis package written by Eric Feczko. The manual is split into two main sections. The first section covers how to analyze cross-sectional data with the Random forest subgroup detection (RFSD) tool. The second covers how to analyze longitudinal trajectories with the Function Random Forest (FRF) tool. A brief introduction will walk the user through installing the software.'

A new release is available at:



The image shows a screenshot of the GitHub repository page for 'functional-random-forest'. At the top right, there is a blue button with the text 'View on GitHub' and the GitHub logo. Below this, the repository name 'functional-random-forest' is displayed in a large, bold, white font on a dark background. Underneath the name, the subtitle 'Functional Analysis with Random Forests' is written in a smaller, white font. The main content area has a light gray background and features the title 'Functional Random Forest (FRF) Manual' in a large, bold, black font. Below the title, there is a paragraph of text in a smaller black font, followed by another paragraph of text in a smaller black font.

[View on GitHub](#) 

functional-random-forest

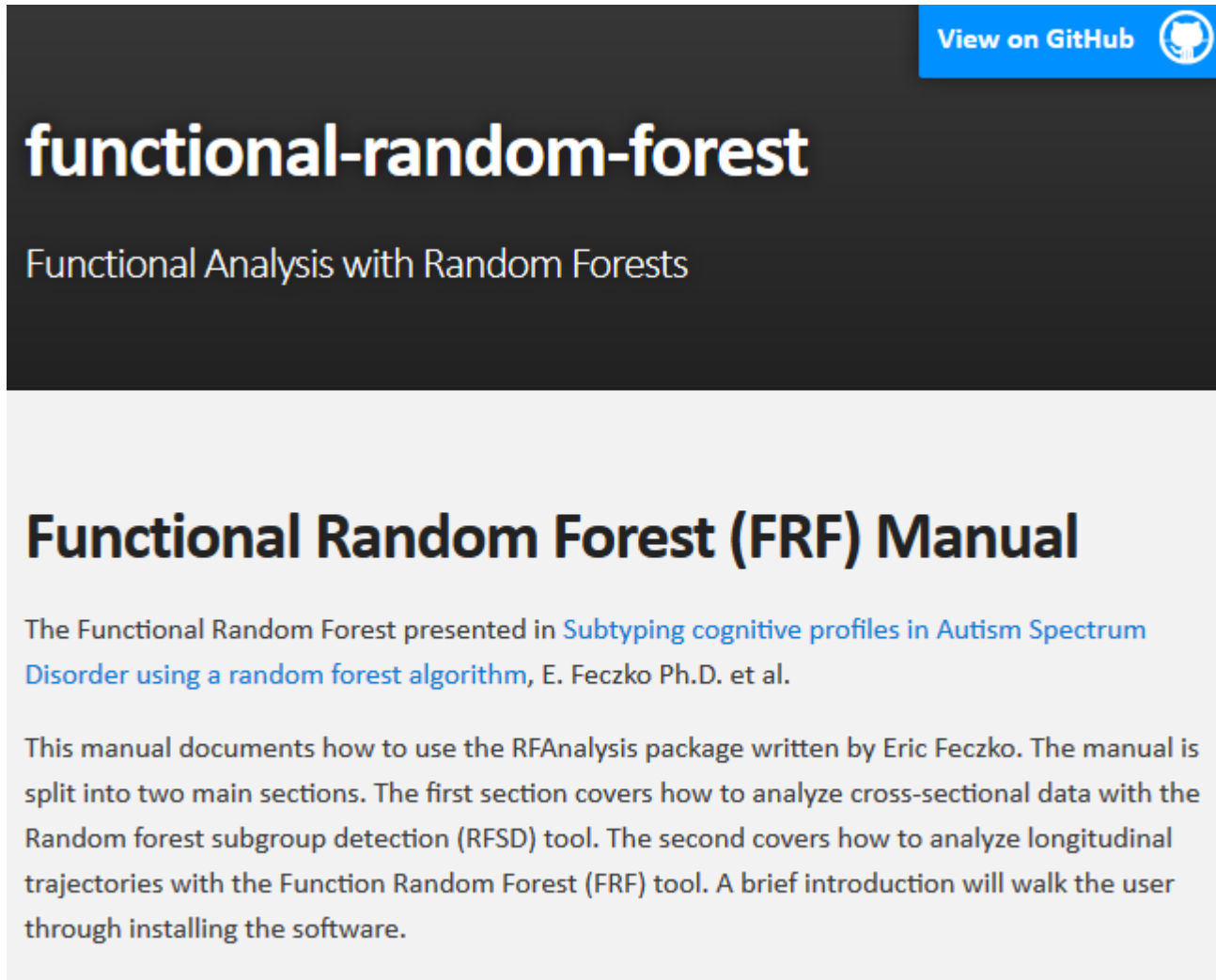
Functional Analysis with Random Forests

Functional Random Forest (FRF) Manual

The Functional Random Forest presented in [Subtyping cognitive profiles in Autism Spectrum Disorder using a random forest algorithm](#), E. Feczko Ph.D. et al.

This manual documents how to use the RFAAnalysis package written by Eric Feczko. The manual is split into two main sections. The first section covers how to analyze cross-sectional data with the Random forest subgroup detection (RFSD) tool. The second covers how to analyze longitudinal trajectories with the Function Random Forest (FRF) tool. A brief introduction will walk the user through installing the software.

A manual for using the FRF exists online (<https://dcan-labs.github.io/functional-random-forest/>)



The screenshot shows the GitHub repository page for 'functional-random-forest'. At the top right, there is a blue button that says 'View on GitHub' with the GitHub logo. Below this, the repository name 'functional-random-forest' is displayed in a large, bold, white font. Underneath the name, the description 'Functional Analysis with Random Forests' is written in a smaller white font. The main content area has a light gray background and features the title 'Functional Random Forest (FRF) Manual' in a large, bold, black font. Below the title, there is a paragraph of text: 'The Functional Random Forest presented in [Subtyping cognitive profiles in Autism Spectrum Disorder using a random forest algorithm](#), E. Feczko Ph.D. et al.' This is followed by another paragraph: 'This manual documents how to use the RFAAnalysis package written by Eric Feczko. The manual is split into two main sections. The first section covers how to analyze cross-sectional data with the Random forest subgroup detection (RFSD) tool. The second covers how to analyze longitudinal trajectories with the Function Random Forest (FRF) tool. A brief introduction will walk the user through installing the software.'

Outline of talk

- Theory recap: modelling approaches can be reduced to two types: predictive and descriptive
- “Big data” complicates our ability to apply both approaches
- Marginal Modelling is a good approach for descriptive modelling
- Functional Random Forests is a good approach for predictive modelling
- **Other approaches can also handle big data, but are beyond the scope of this workshop**

New approaches within statistics and machine learning can also accommodate problems with big data

- Many of these approaches have been developed in genomics
 - comBat is a Bayesian approach to handle known site effects in data
 - Surrogate Variable Analysis
- Such approaches need to be examined in the context of neuroimaging data to evaluate where each is most useful
- Knowing how to use these tools requires considerable skill in data science, which has been relatively untaught in mental health fields
- Hopefully, the workshop tomorrow should get you excited about applying these new tools and on your path towards doing “big data” science right.

Acknowledgments

Fair Lab

- Damien Fair
- Oscar Miranda-Dominguez
- Alice Graham



Alpha Testers

- Bene Ramirez
- Jennifer Zhu
- Robert Hermsillo
- Mollie Marr
- Oliva Doyle
- Michaela Cordova
- AJ Mitchell

Computing Team

- Darrick Sturgeon
- Eric Earl
- Anders Perrone
- Emma Schifsky
- Anthony Galassi
- Kathy Snider
- David Ball
- Lucille Moore

Acknowledgments

- The mentors
 - Damien Fair
 - Joel Nigg
 - Eric Fombonne
 - Shannon McWeeney
- The databasors
 - Lourdes Irwin
 - Darrick Sturgeon
 - Rachel Klein
- The developers
 - Eric Earl
 - Anders Perrone
 - Darrick Sturgeon

- The assessors:
 - Beth Langhorst
 - Michaela Cordova
 - Bene Ramirez
 - Brian Mills
 - Olivia Doyle

- Other Labs:
 - Nigg Lab
 - McWeeney Lab

- The collaborators:
 - Sarah Karalunas
 - Alison Hill
 - Jan Van Santen

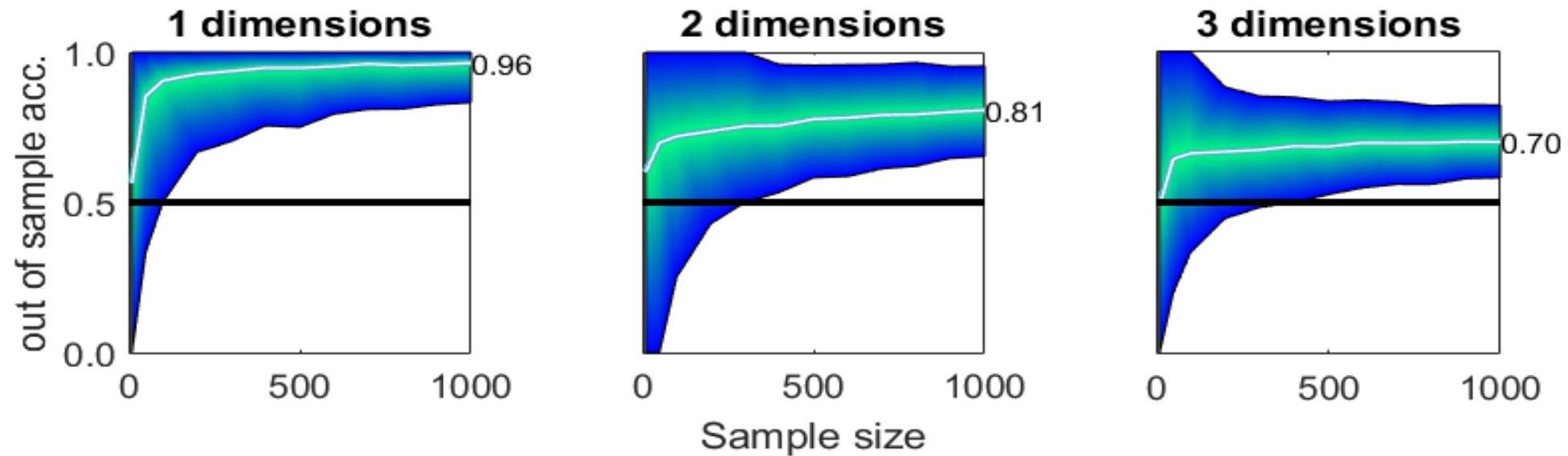


- Everyone I forgot, which is many 😊

- The docs:
 - Alice Graham
 - Oscar Miranda-Dominguez
 - Binyam Nardos

Questions?

High dimensionality is bad for predictive modelling



Predictive models must also take into account nested structure

