fMRI 3: Group analysis and statistical inference

Johan Carlin

MRC Cognition and Brain Sciences Unit, Methods Group

johan.carlin@mrc-cbu.cam.ac.uk

MRC | Cognition and Brain Sciences Unit

Outline

- How to build a second-level GLM for random-effects group analysis in SPM

- Statistical inference and multiple comparisons correction

- Workshop: SPM group analysis and a tour of the SPM results user interface

Summary statistics RFX

First level

Second level

fMRI data   Design Matrix   Contrast Images   One-sample t-test @ second level

Subject 1

...

Subject N

contrast(s)

Design matrix

$$t = \frac{c^T \hat{\beta}}{\sqrt{V\hat{a}r(c^T \hat{\beta})}}$$

SPM{T$_{15}$}
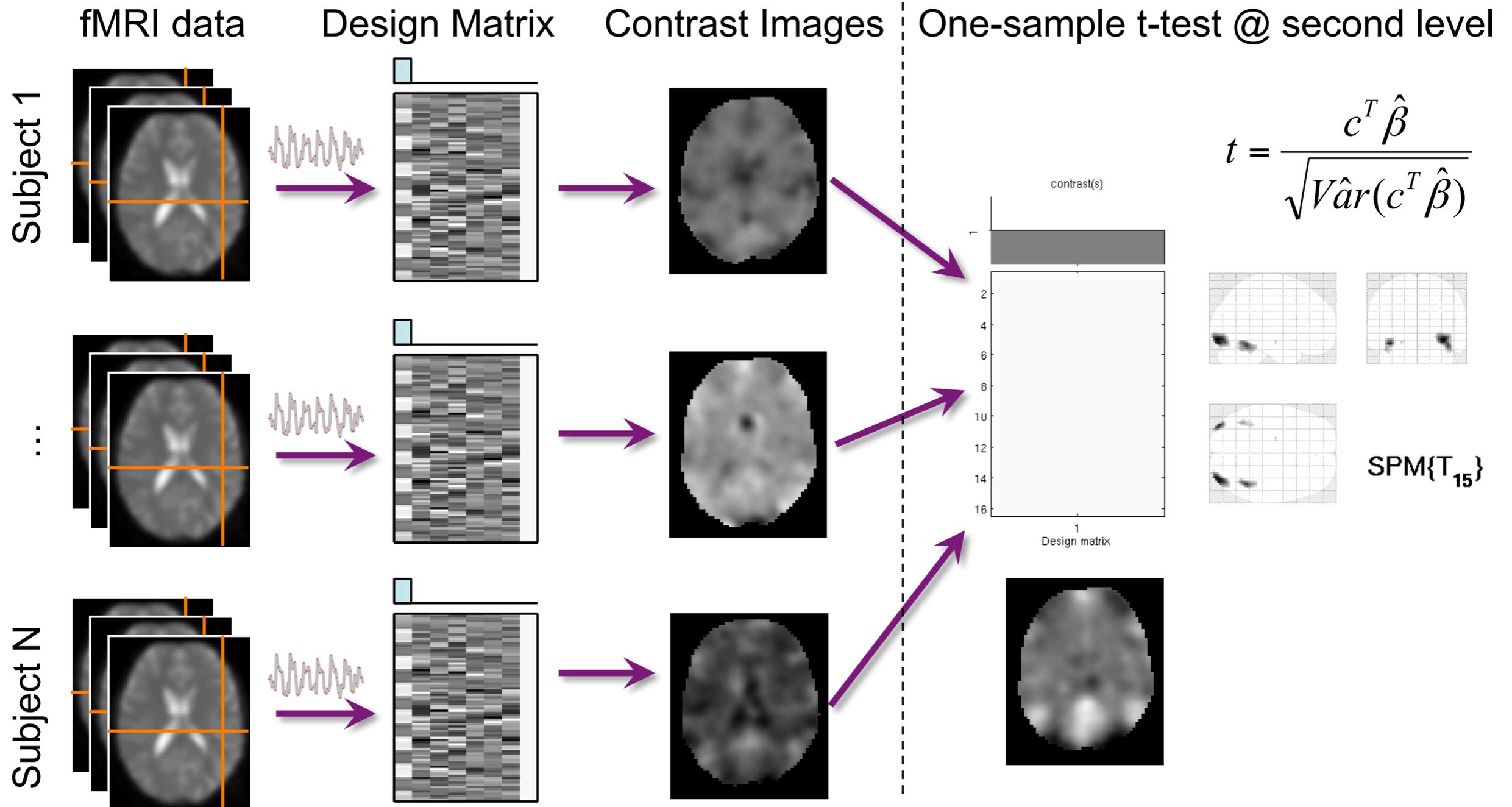
Summary-statistic RFX logic and assumptions

- The variance of the contrast estimates that go into the group model are a mixture of within-subject and between-subject variance

- Sounds weird, but this turns out to be equivalent to a 'full' RFX model where within- and between-subject variance is explicitly modelled, given the following assumptions:

  - Similar within-subject variance (e.g., each participant is broadly similar in head movement)

  - Similar first-level designs (e.g., same number of trials)

- Reasonably robust to violations of these assumptions - but for extreme cases you might have to fit the 'full' hierarchical model instead

- (Remember, if you ever ran a standard Psychology analysis in e.g. SPSS you probably used a very similar summary-stat group analysis approach, just on mean accuracies/RTs instead of fMRI contrast estimates)

Design matrices for popular analyses (careful, SPM terminology!)

One-sample t-test

why is this empty?
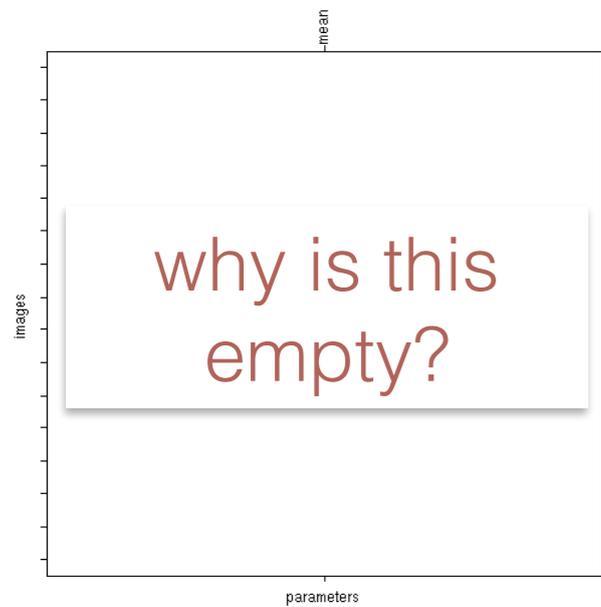
Two-sample t-test

Paired t-test

equivalent to one-sample t on difference score
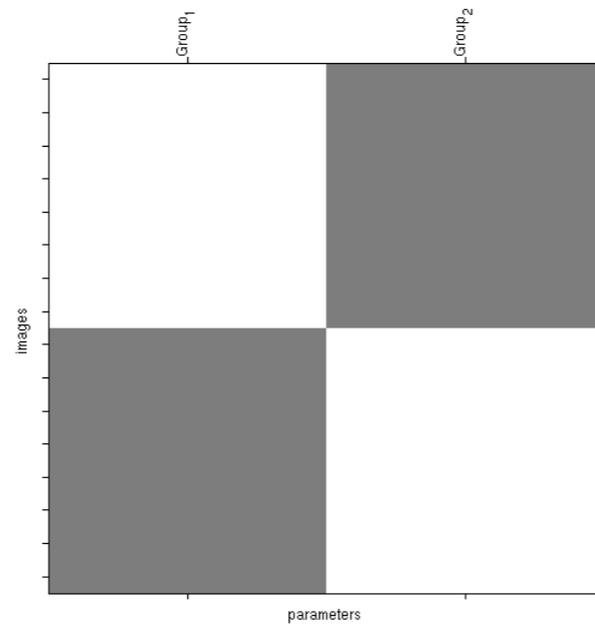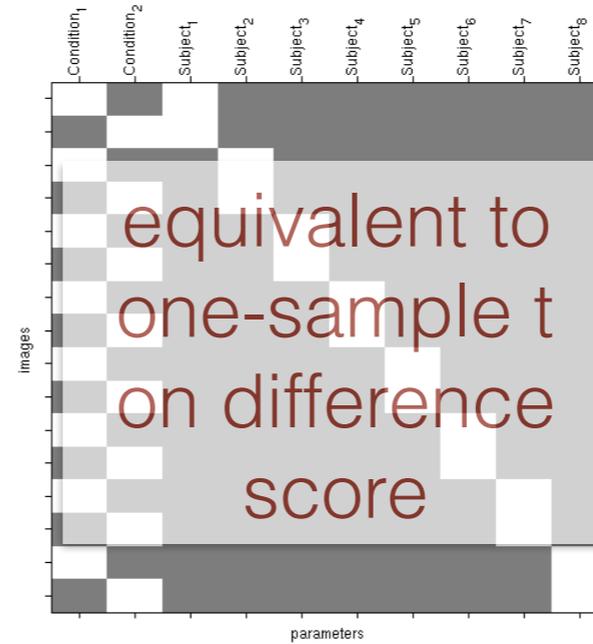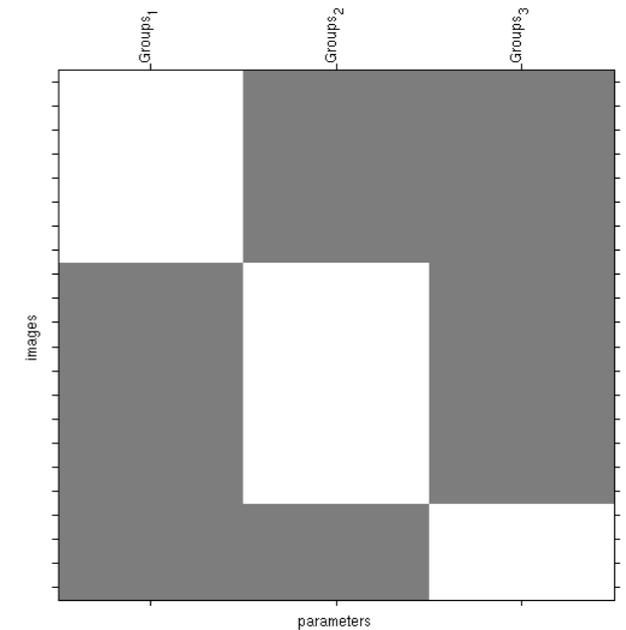
One-way ANOVA

One-way ANOVA within-subject

Full Factorial

between subjects only!
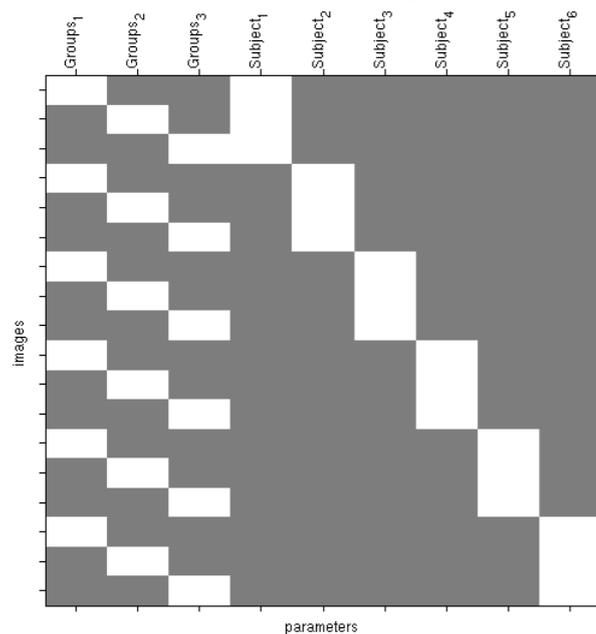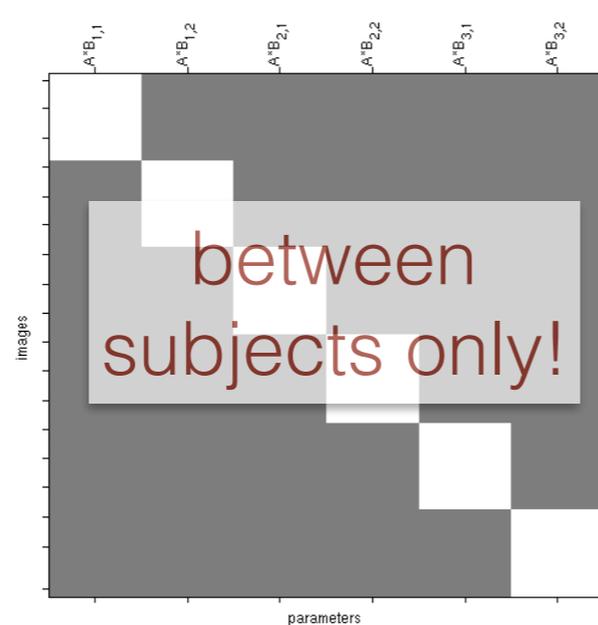
Flexible Factorial

Flexible Factorial

can accommodate within subjects factors
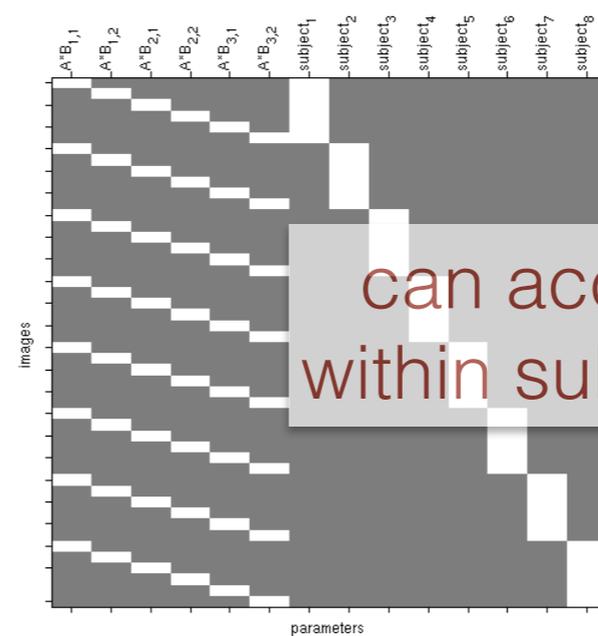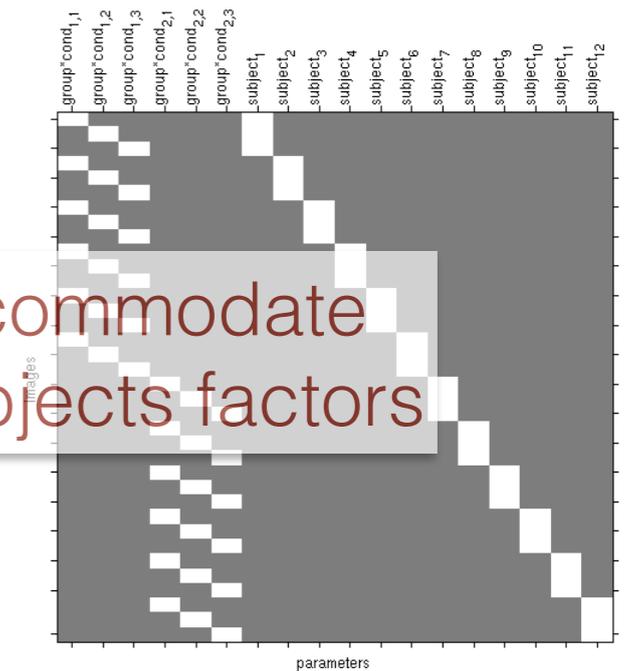
SPM GUI options for typical designs

SPM has two main options when specifying a factor: independence and variance. Here's how to set them:

- within-subjects factor: independence *no* (since the levels come from the same subject we expect correlated residuals), variance *equal*

- between-subjects factor: independence *yes*, variance *equal*

- between-groups factor: independence *yes*, variance *unequal* (just like in e.g. between-subjects T test, where variance is estimated separately per group)

- (if at all possible, avoid specifying this in the GUI - error prone, not reproducible. Use batch scripts instead, e.g., Rik's batch_spm_anova.m)

Group analysis mask (mask.hdr)

- Intersection of all the single-participant masks

  - which are defined as voxels with mean signal > 80% of the global mean

- Always inspect to make sure a region you care about wasn't eliminated (interpret null finding)

- To improve coverage, you can disable implicit masking (it's defaults.mask.threshold in spm_defaults) and use a custom explicit mask in single-subject models (e.g., FSL BET)

  - But careful with group mask smoothness if using RFT correction (e.g., GM mask is inappropriate)

  - And consider why this happened - excessive dropout? Bad slice plan? If so consider dropping problem subject instead
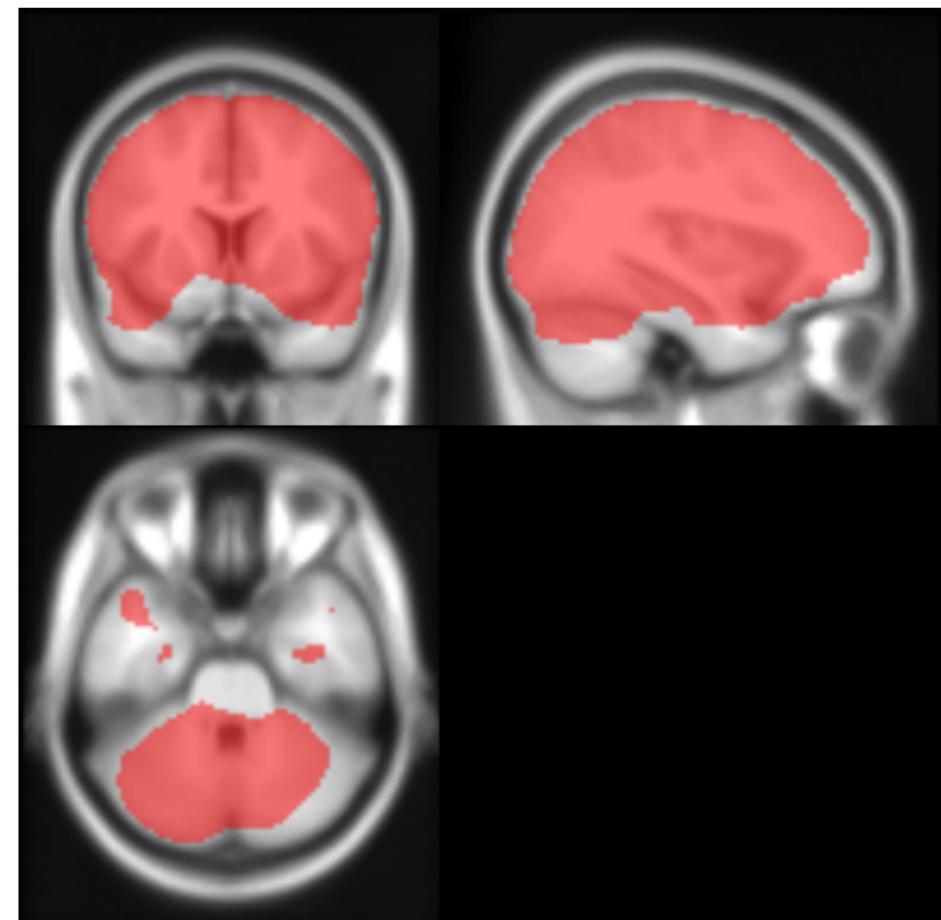
Multiple comparisons correction in fMRI

- Just Bonferroni? With 176000 voxels inside the brain in a typical SPM group analysis, p(FWE)<0.05 corresponds to an uncorrected threshold of p<0.0000002

- Massively conservative since the data is smooth - the number of independent observations is much less than the number of voxels

- NB smoothness comes from the images themselves and from reslicing during preprocessing, not just from spatial smoothing we apply directly. So we can't fix this by skipping explicit smoothing

- So the appropriate correction is not 0.05 / numvoxel, but then what is it?

Voxel and cluster-extent thresholds

Height-based

Voxel threshold

u_α

Significant Voxels

No significant Voxels

space
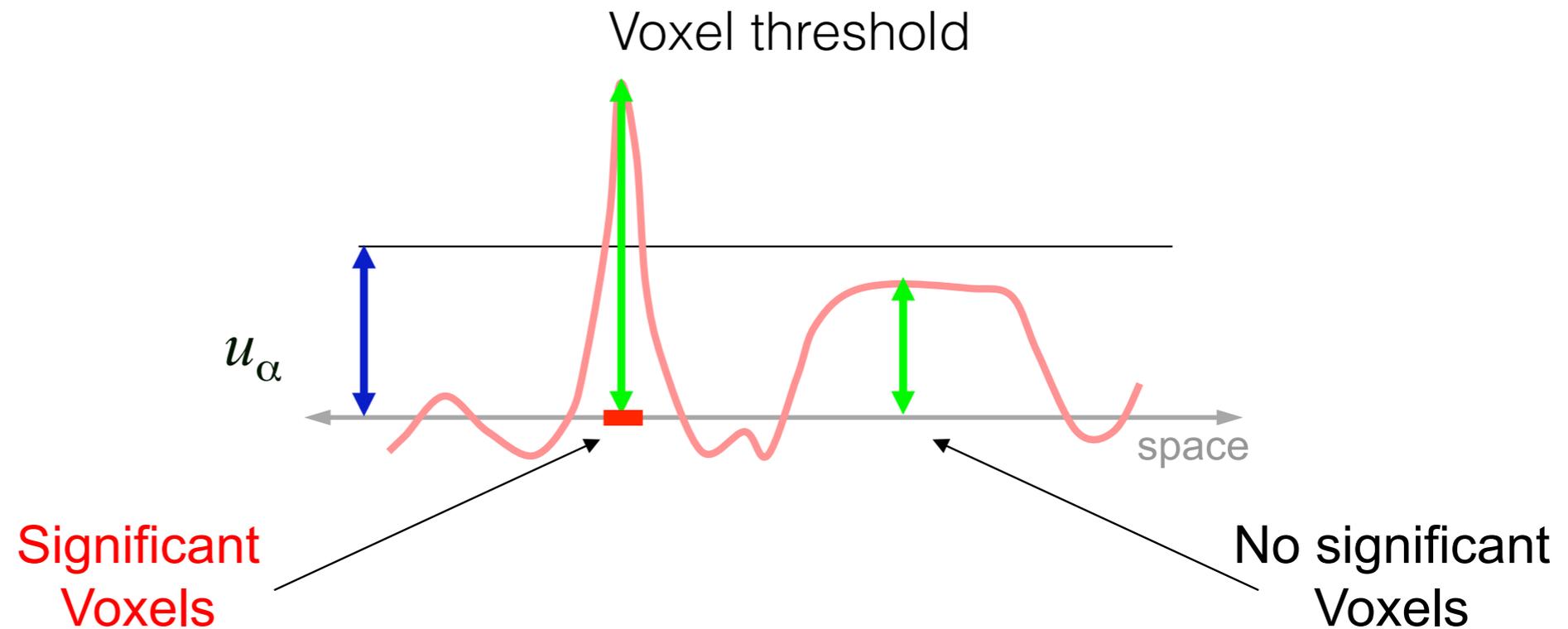
Extent-based
(but NB, depends
on setting an initial
height threshold)

Cluster threshold

u_clus

Cluster not significant

k_α

k_α

Cluster significant

space
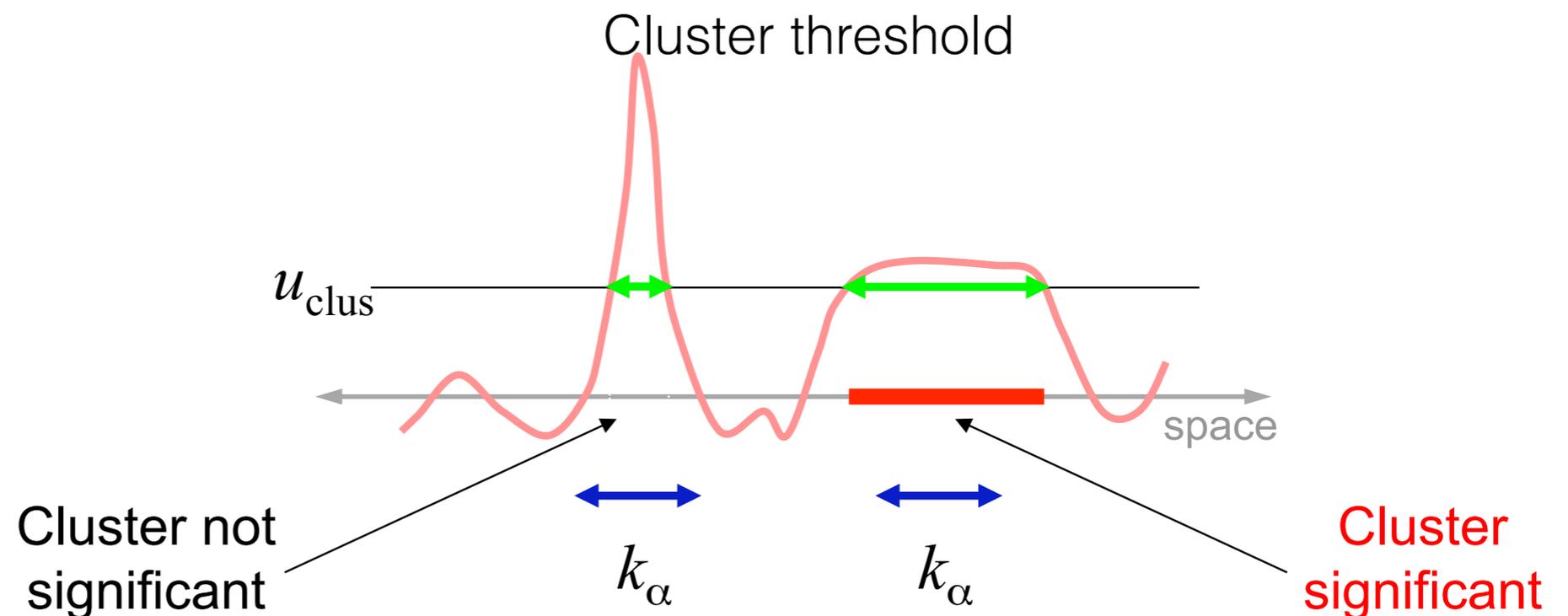
Pros and cons: cluster

+ More sensitive to extended activations (compared to voxelwise)

- Depends *critically* on cluster-forming threshold, which must be set a priori (resist temptation to drop it when first analysis doesn't work as intended) - unless using TFCE and permutation tests

- *Less* sensitive to focal effects (but this is unusual in group-level, smooth fMRI data)

- Rejects null for entire cluster: interpret as 'one or more voxels' in the cluster are active. Can complicate things if the cluster is large and spans multiple anatomical regions

- Stationarity assumption: RFT cluster inference is invalid if smoothness varies over the image (big clusters are more likely in smoother regions…) - this may be the case with fMRI data

Cluster-forming threshold - choose your poison

Z > 2.0	Z > 3.0	Z > 4.0	Z > 5.0
7 clusters largest: 98,246 voxels	32 clusters largest: 57,500 voxels	54 clusters largest: 10,960 voxels	50 clusters largest: 783 voxels
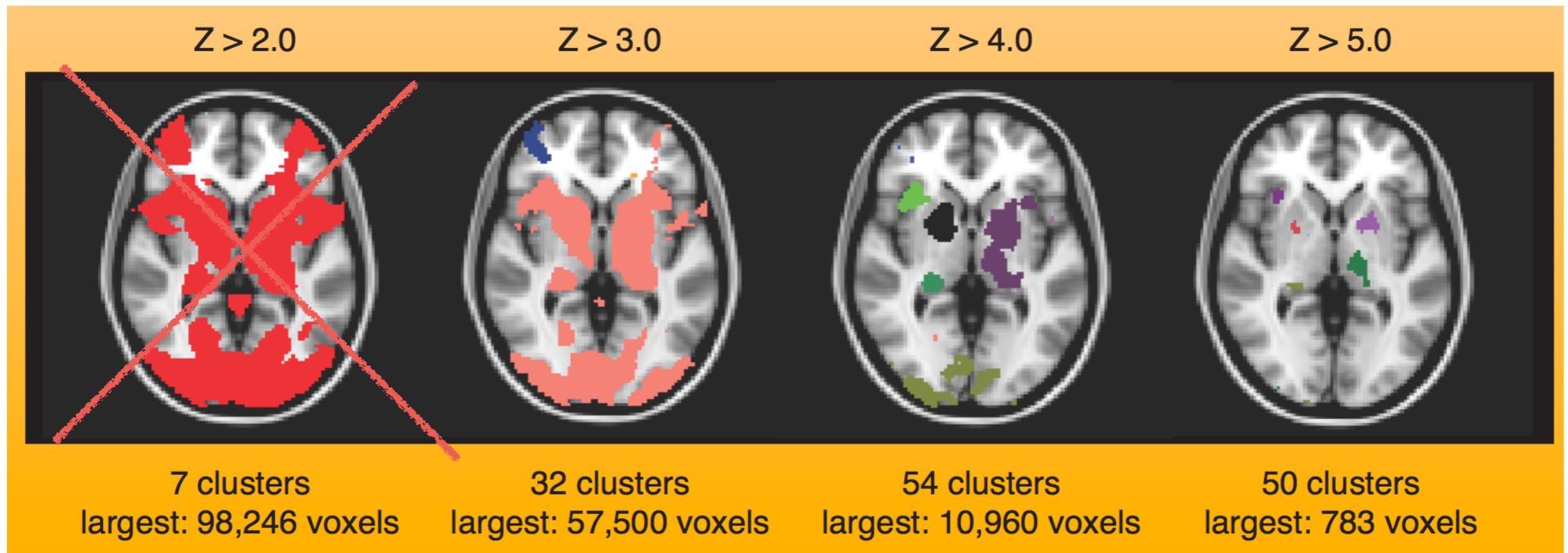
too low thresholds lead to liberal corrections

… But no real a priori reason to prefer any of the remaining thresholds

SPM parametric inference with random field theory

- Estimate smoothness in the data from the residuals

- Summarise as resolution elements (resels), which is similar to but not quite the same as the number of independent observations in the data

  - (You can inspect RPV.hdr image - resels per voxel. Big value means local *roughness*)

- Intuitively, really smooth data needs to be corrected less severely for multiple comparisons

- Parametric inferential method - can be used for any flavour of correction (FWE/FDR, voxel/cluster)

Random field theory pros and cons

+ Analytical solution: Instant results, parametric inference if that's your sort of thing

+ Easy to use: Well supported by SPM results user interface

+ Standard practice: By far the most widely used correction. If there was a problem, we'd know about it by now (we'll return to this point)

- Smoothness assumption: Requires 'sufficient' smoothness for validity. But a) we don't know how much that is, b) it will depend on degrees of freedom, c) it's probably more than you'd like (e.g., 10x voxel for small N - 30mm FWHM for 3mm voxels!)

- Transparency: Unless you have a strong maths background you will not be able to understand how RFT works in detail

Nonparametric inference with permutation tests

- Instead of assuming a null distribution, estimate it from data

- Basically a big loop where on each iteration:

  1. randomise the rows of the design matrix (For a one-sample test, we flip the sign of the contrast estimate instead)

  2. re-calculate the test statistic (e.g., T, Z)

  3. store the full null volume (for voxelwise) and the max stat over all voxels (for FWE correction)

- Voxel-wise uncorrected p value: proportion of null tests>true test at this voxel

- Voxel-wise FWE correction: proportion of *maximal* statistics>true test

- Very few assumptions for validity:

  - *exchangeability of labels* under the null (ie, we can permute without expecting anything to change, if the null hypothesis is true)

  - *symmetry about zero* under the null (for sign flip test)

- Software: SnPM toolbox for SPM, randomise for FSL, PALM toolbox for Matlab
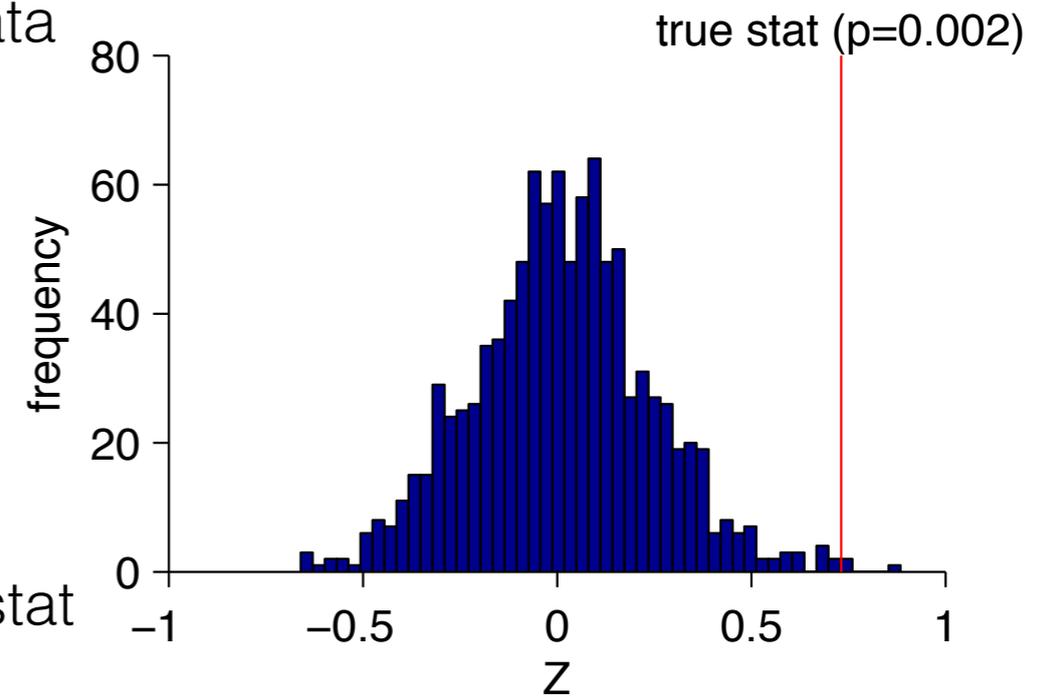
see permutationdemo.m
in workshop

Nichols & Holmes, 2001, *HBM*

Pros and cons: permutation tests

+ Minimal assumptions: No major failure points for typical fMRI group analysis cases (unlike RFT)

+ Sensitivity: Maximal stat FWE-correction is generally more powerful than parametric corrections (especially if using variance-smoothed voxelwise stats, or threshold-free cluster enhancement, TFCE)

+ Transparency: the maths are ridiculously simple

+ Threshold-free: Cluster correction without setting that annoying cluster-defining threshold (if using TFCE)

- Harder to use without writing your own analysis code (SnPM has an ok GUI for basic analyses)

- Computationally expensive (but this is not really a problem anymore)

- Single-participant inference is more complicated (exchangeability assumption gets tricky with temporally-dependent fMRI volumes)

Small volume correction

- Given a priori hypotheses about localisation, we can reduce the pain of whole-brain correction to a local, small-volume correction

- Can be image (e.g., thresholded stat map from previous study, anatomical region - should be reasonably smooth for RFT, for permutation test it doesn't matter), or sphere centred on MNI coordinate from previous study

- VERY sensitive to fishing expeditions (e.g., almost any MNI coordinate has been implicated in almost any task over the history of fMRI). Don't fool yourself!

- If the 'small volume' is really small and focal, perhaps consider trying a straight ROI analysis instead? More on this later

False discovery rate (FDR) correction

- Analyse the distribution of uncorrected p values to find a threshold where the expected FDR is at specified threshold

- *Not* (strong) familywise error correction:

  - FWE: 5% of analyses produce *any* false positive clusters/voxels

  - FDR: Control the proportion of significant voxels/ clusters that are expected to be false positives (conventionally at 5%)

- 'Weak' FWE control - if the null is true for all voxels (ie, the null distribution is uniform), we only expect FDR to reject the null anywhere in 5% of analyses, just like with FWE. Bit confusing…

- Voxelwise FDR tends to to work 'better' than FWE - you are making a less stringent correction

- But probably not the right test if you want to make a big claim about a tiny, borderline blob in a map with much more extensive primary sensory activations…
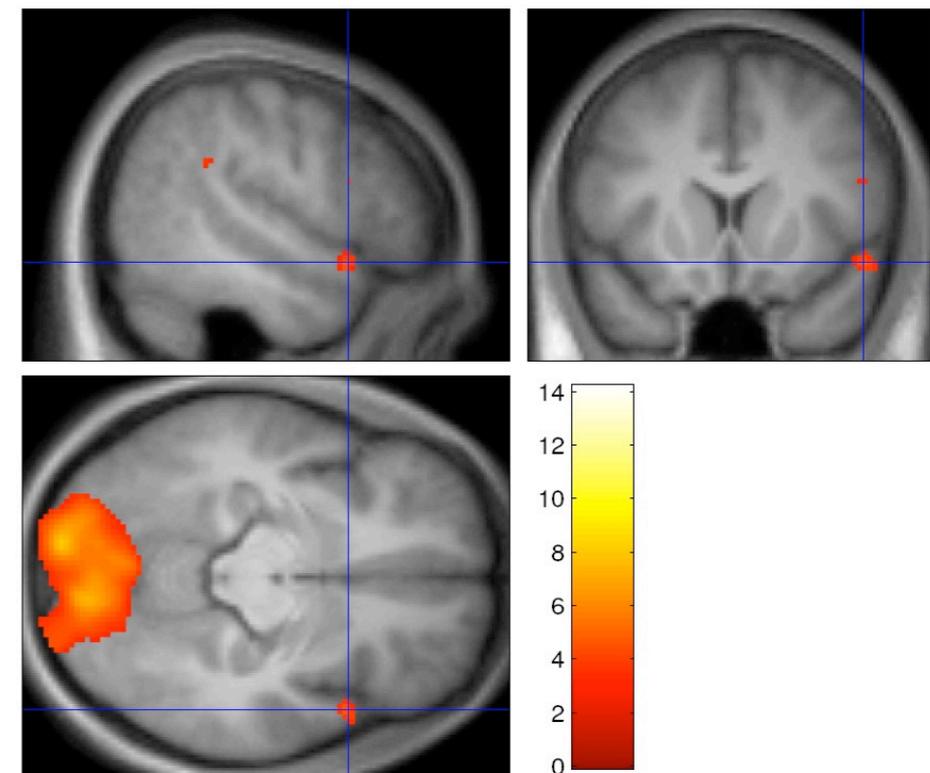
Understanding FWE/FDR corrections

Noise
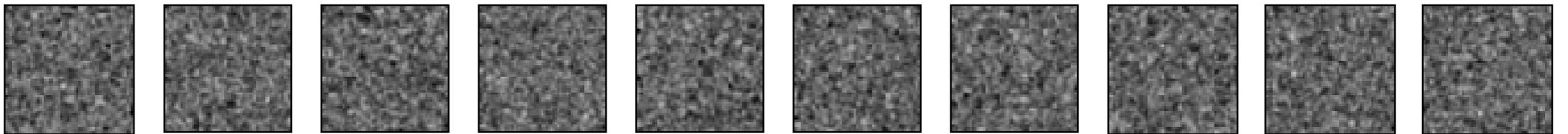
Signal
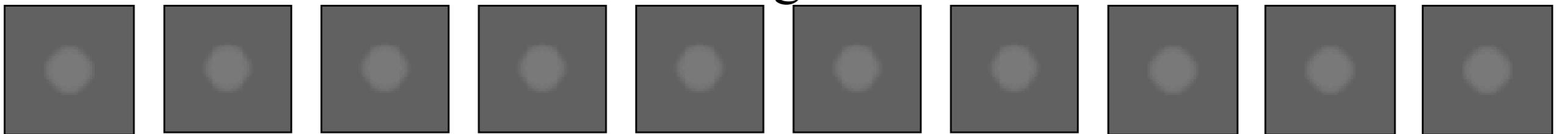
Signal+Noise
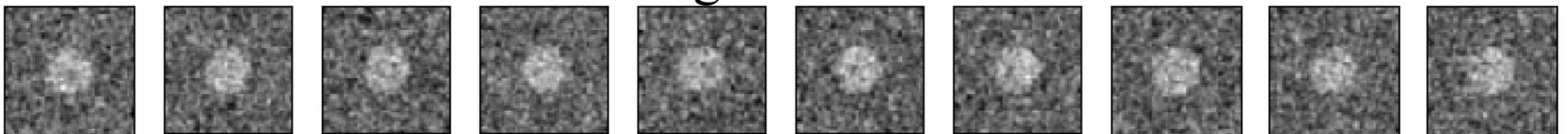
Tom Nichols

Control of Per Comparison Rate at 10%
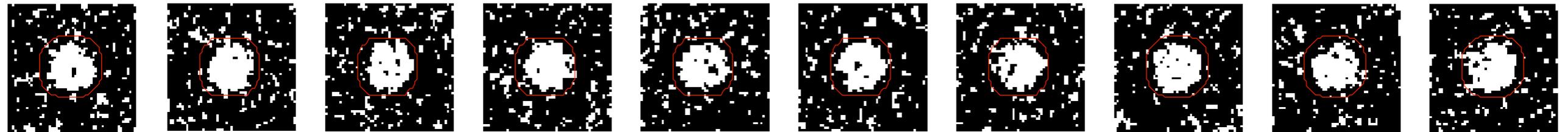
11.3% 11.3% 12.5% 10.8% 11.5% 10.0% 10.7% 11.2% 10.2% 9.5%

Percentage of Null Pixels that are False Positives

ie, uncorrected, p<0.1
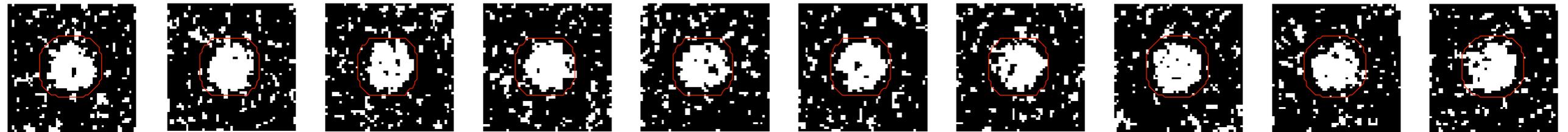
Tom Nichols

Control of Per Comparison Rate at 10%

11.3%   11.3%   12.5%   10.8%   11.5%   10.0%   10.7%   11.2%   10.2%   9.5%

Percentage of Null Pixels that are False Positives

Control of Familywise Error Rate at 10%

FWE
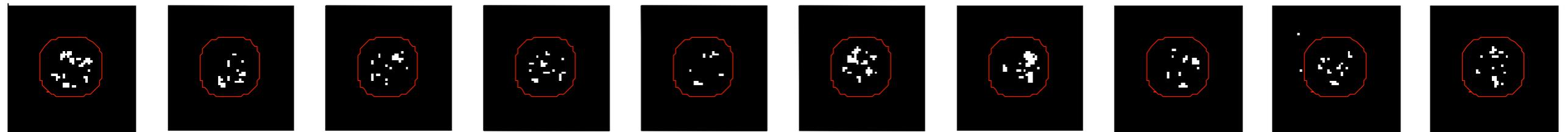
Occurrence of Familywise Error

Great false positive control
But so many false negatives…
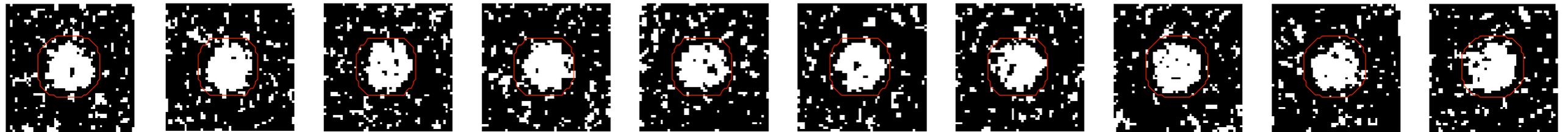
Tom Nichols

Control of Per Comparison Rate at 10%

11.3% 11.3% 12.5% 10.8% 11.5% 10.0% 10.7% 11.2% 10.2% 9.5%

Percentage of Null Pixels that are False Positives

Control of Familywise Error Rate at 10%

FWE
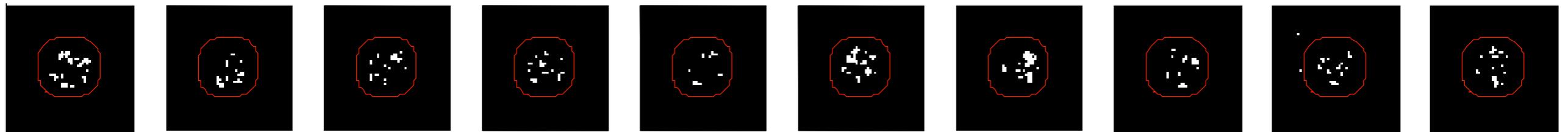
Occurrence of Familywise Error

Control of False Discovery Rate at 10%
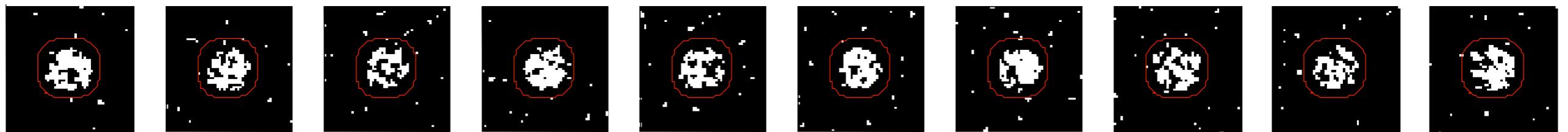
6.7% 10.4% 14.9% 9.3% 16.2% 13.8% 14.0% 10.5% 12.2% 8.7%

Percentage of Activated Pixels that are False Positives

Tom Nichols

Control of Per Comparison Rate at 10%
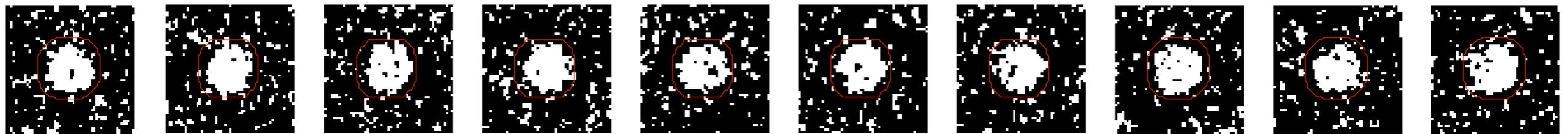
11.3% 11.3% 12.5% 10.8% 11.5% 10.0% 10.7% 11.2% 10.2% 9.5%

Percentage of Null Pixels that are False Positives

Control of Familywise Error Rate at 10%
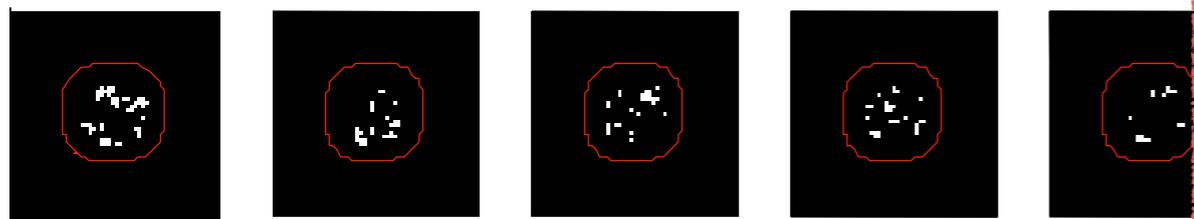
Occurrence

What's the FDR for *clusters* with voxelwise FDR? Huge in this example! (Chumbley & Friston, 2009 , *NeuroImage*)

Control of False Discovery Rate at 10%
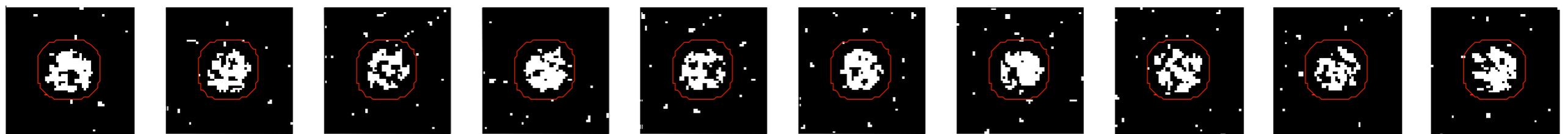
6.7% 10.4% 14.9% 9.3% 16.2% 13.8% 14.0% 10.5% 12.2% 8.7%

Percentage of Activated Pixels that are False Positives

Tom Nichols

FDR updates in SPM

- SPM8 and later defaults to a 'topological' FDR that depends on the cluster-forming threshold (so a voxelwise test that depends on a cluster-forming threshold? Afraid so)

- In practice it tends to provide a similar correction to standard FWE cluster-extent correction, so you don't often see topo-FDR reported in the literature

- If you want the old-style voxelwise FDR you can turn it on by setting defaults.stats.topoFDR = 0 in spm_defaults

Do neuroimaging corrections work?

- Fit made-up event (EX) and block (BX) design matrices to resting state fMRI data

- Run through group analysis with standard neuroimaging package multiple comparisons corrections

- Expect 5% of group analyses to come out significant if the tests are well calibrated

Eklund et al., 2016, PNAS

Beijing, one sample t-test, 6 mm, CDT p = 0.01

Legend:
- B1 10 s on off
- B2 30 s on off
- E1 2 s events
- E2 randomized events
- Expected
- 95% CI

X-axis: SPM FLAME1 FSL OLS 3dttest 3dMEMA Perm
Y-axis: Familywise error rate (%)

Cambridge, one sample t-test, 6 mm, CDT p = 0.01

Legend:
- B1 10 s on off
- B2 30 s on off
- E1 2 s events
- E2 randomized events
- Expected
- 95% CI

X-axis: SPM FLAME1 FSL OLS 3dttest 3dMEMA Perm
Y-axis: Familywise error rate (%)

Do these corrections work - p<0.01 cluster-forming threshold

- parametric (RFT) cluster correction is extremely liberal at a p<.01 cluster-forming threshold - up to 50% probability of rejecting null hypothesis when it is true

- permutation tests preserve intended false positive rate

Eklund et al., 2016, PNAS
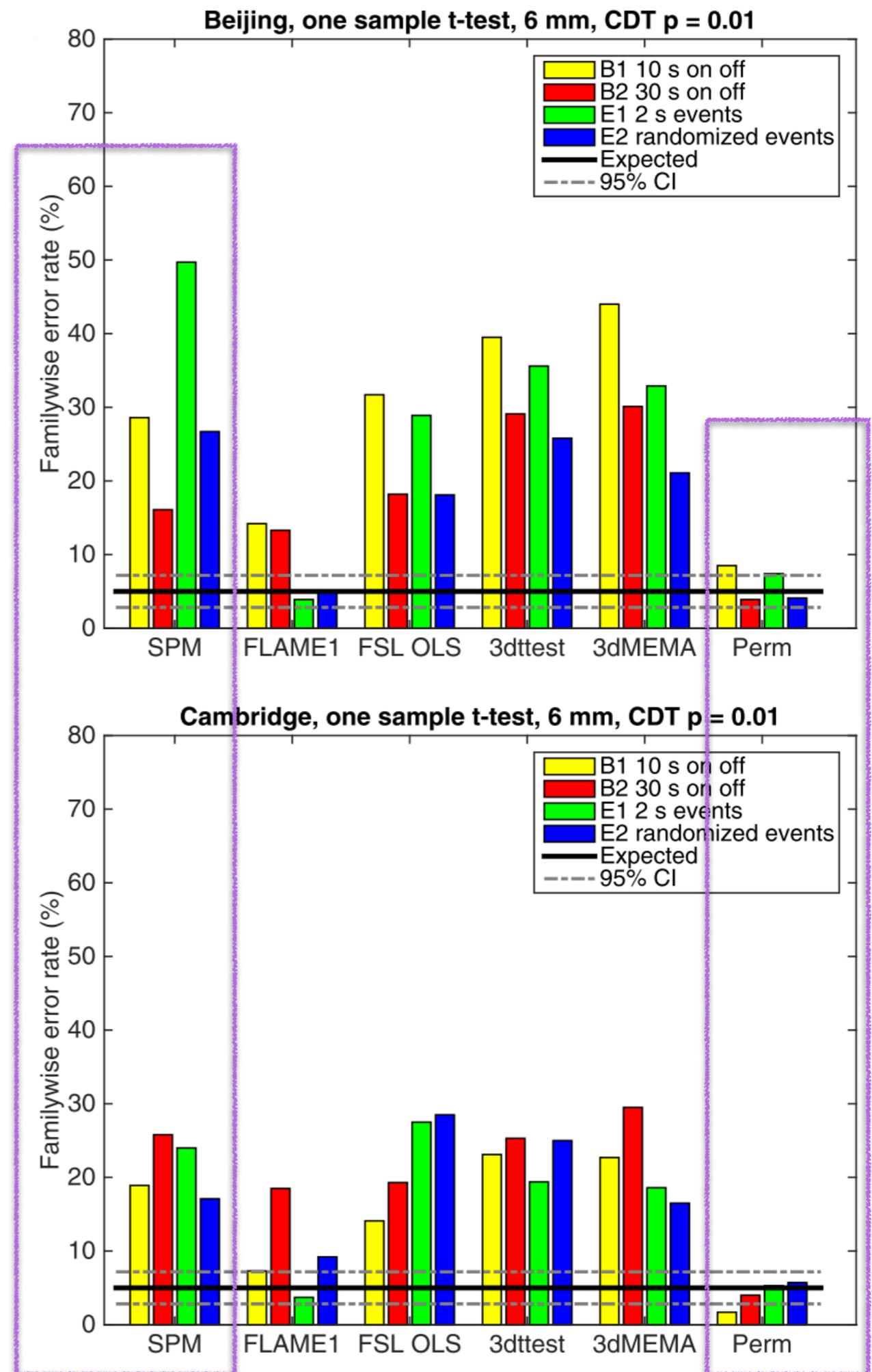
Do these corrections work - p<0.001 cluster-forming threshold

- Parametric cluster correction (RFT) is still liberal, but better than p<.01 case

- Permutation tests are still closer to the intended false positive rate

- (This is the default cluster-forming threshold in SPM)

Eklund et al., 2016, PNAS
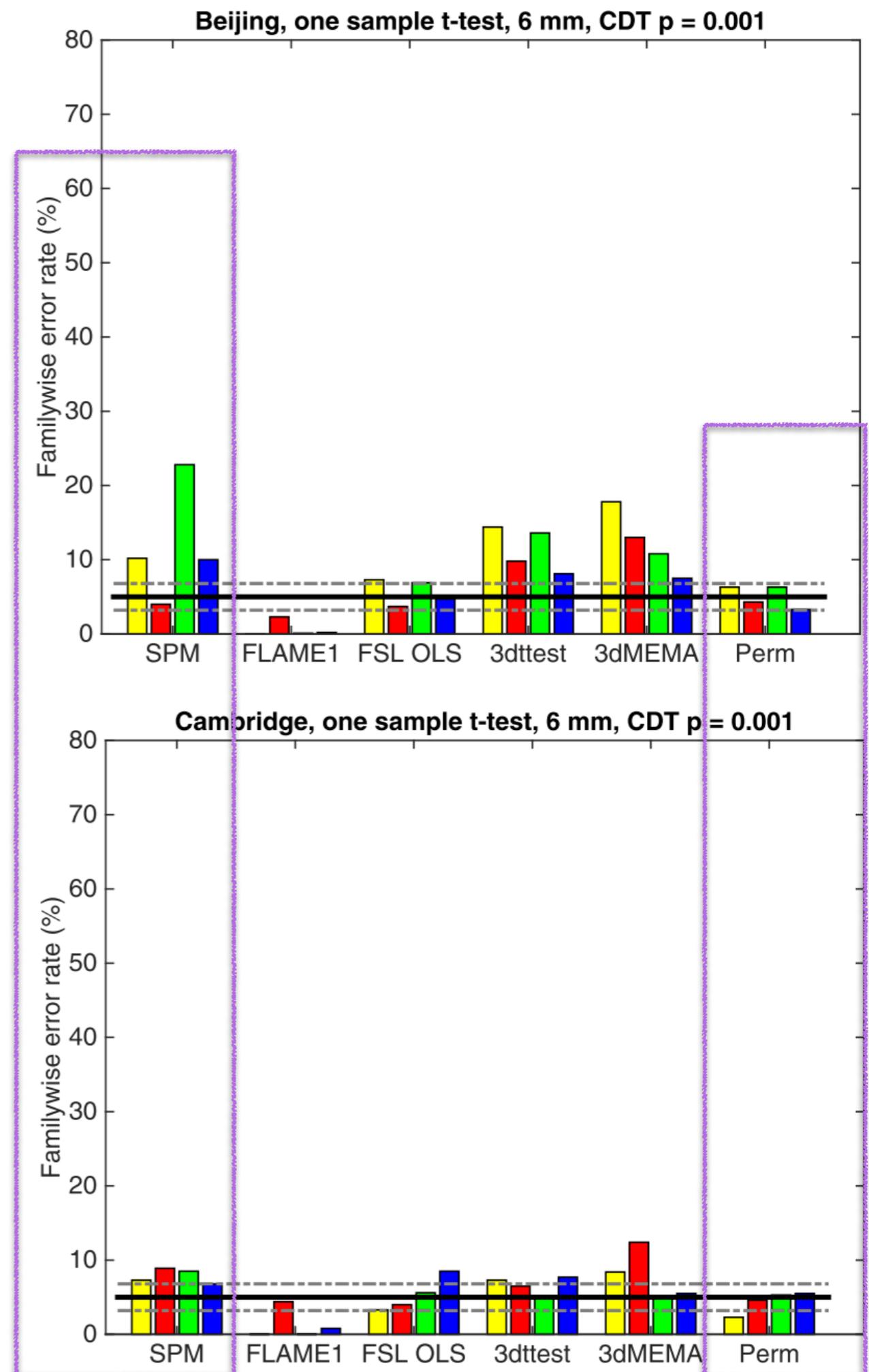
Do these corrections work - voxelwise correction

- Voxel-wise correction with parametric (RFT) methods is *conservative*

- Permutation tests are closer to intended false positive rate

Eklund et al., 2016, PNAS
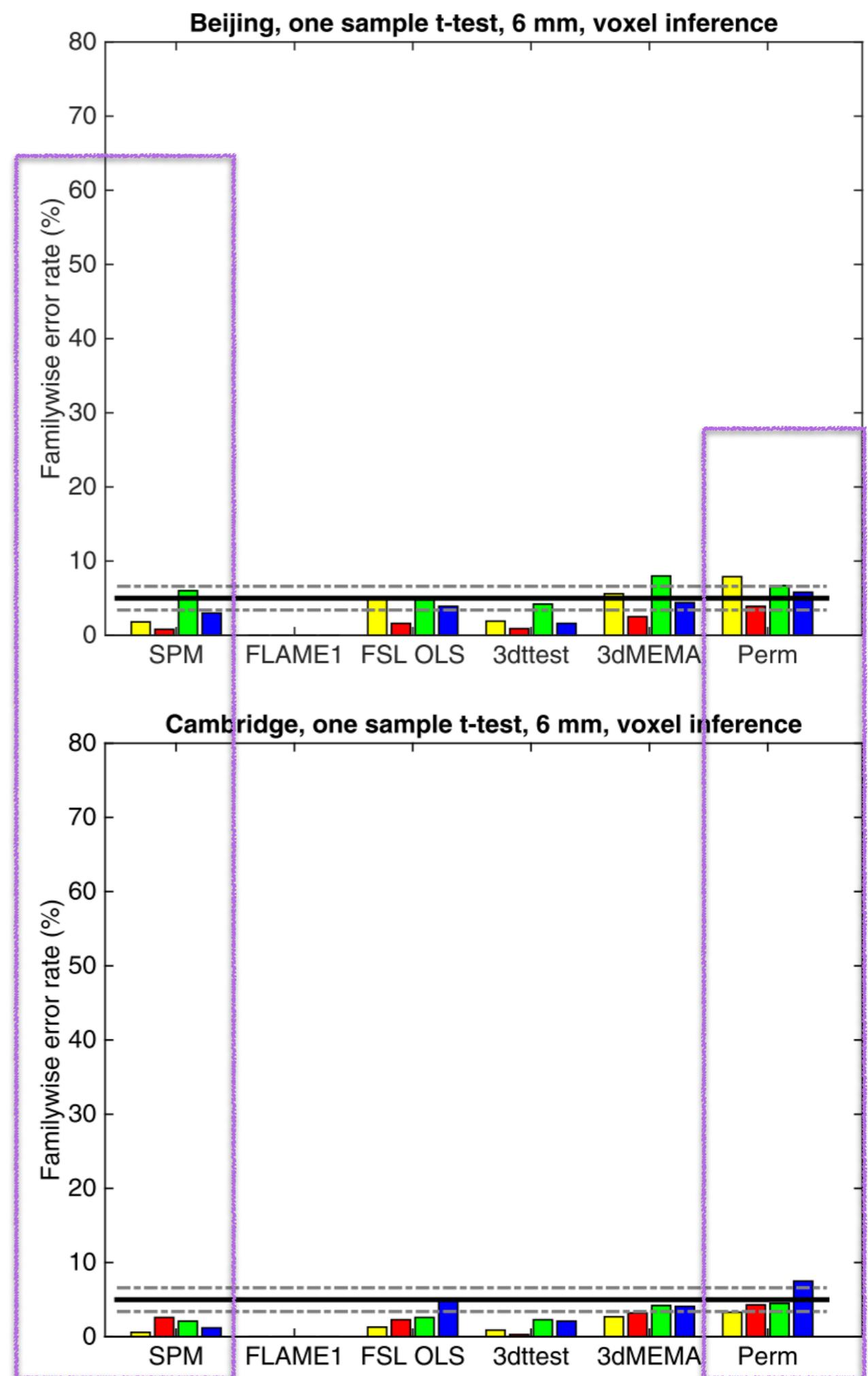
Brain mapping is hard - can't we do something more fun?

- If you can define a region of interest (ROI) by *independent* functional or anatomical criteria, your life will be easier:

  - Massively greater detection power (due to reduced multiple comparisons correction)

  - A more sensible concept of replication, which enables cumulative science (what is a replication in brain mapping? Same voxel? How close?)

- Region-mean ROI analysis: see marsbar SPM toolbox

- Multivariate: see e.g. RSA toolbox

Workshop time

Workshop plan

1. The workshop directory contains multiple example scripts for running various types of group-level analyses

2. Let's take a look at one of these scripts: AA_LD_OsT_Cov, which is a one-sample T test with a covariate. Use the Matlab editor to look at the code then run it (e.g. by pressing F5).

3. If it runs successfully, you will find yourself transported to the results directory (er, thanks SPM?). If you don't already have the SPM user interface open, type spm FMRI into the command window.

4. Click "Review", select SPM.mat. Inspect the design matrix. Try Design >> Design Orthogonality

5. Click "Results", select SPM.mat. Choose a contrast, threshold it at uncorrected, p<0.001.

6. Inspect table. note cluster extent threshold (k) for FWE - it's FWEc at the bottom left. Reset extent threshold to this (shortcut: Contrasts >> Significance level >> Change…)

7. Choose overlays… >> Sections. Use previous overlay to find the SPM dir. Inside 'canonicals' there are MNI template brains. You can also overlay on mask.hdr from the analysis directory to check this.

8. Right-click in the glass brain >> go to global max. Click around, try out goto nearest local maximum. Switch back to table by clicking one of the 'p-values' options. The cluster you're on is highlighted in the table.

9. Plot the parameter estimates for the peak. Go to Results, define new contrast. Call it 'eye', set it to F, and evaluate it to eye(2) (this will not work for rank-deficient designs, where contrasts must sum to zero - Glascher & Gitelman have a nice technical note explaining how to handle this case).

10. Click on Display >> plot >> Contrast estimates >> eye. Parameter estimates+CIs!

11. Go back to the original directory so we can try another model: cd(cwd)