Edinburgh Imaging
www.ed.ac.uk/edinburgh-imaging

SINAPSE
Scottish Imaging Network - A Platform for Scientific Excellence

THE UNIVERSITY OF EDINBURGH

Computational Approaches to Multiple Comparisons Correction

Cyril Pernet, PhD

Edinburgh Imaging &
Centre for Clinical Brain Sciences

Stats in EEGLAB

- Many options: channels, IC, ERP, power, ERSP
- All type of stats: t-tests, ANOVA, etc
- No peaks ! Typically the whole time course is analysed to look for differences of amplitude/ intensity
- Difference in latency is anyway translated in the difference of amplitude/intensity
- Issue: many tests = many errors

Overview

1. Type I error rate
2. Multiple testing and the Family Wise Error rate
3. Correcting using the maximum under H0
4. Computational approaches to estimate H0
   - Bootstrap
5. Cluster Mass for MEEG
6. TFCE for MEEG

Type I error rate

Pearson-Newman hypothesis testing

- H0: no effect
- H1: there is an effect

	Results is null	Results is significant
H0 is true	True negative	False positive
H1 is true	False negative	True positive

The false positive rate is called type 1 error, and corresponds to the set alpha value (i.e. if you choose 0.05 then the test will 'fail' 5% of the time)

Multiple testing and the Family Wise Error rate

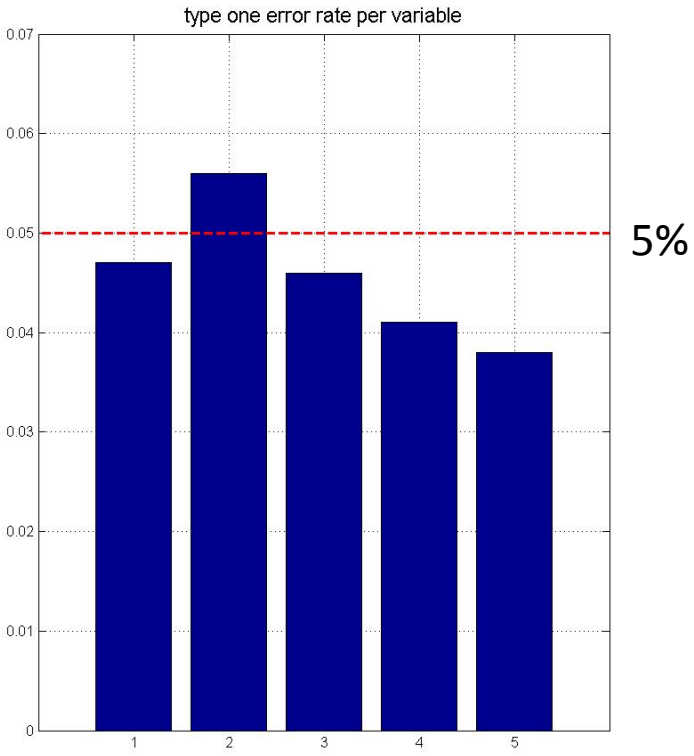
What is the problem?

- Assuming tests are independents from each other, the famillywise error rate FWER = 1 - (1 - alpha)^n

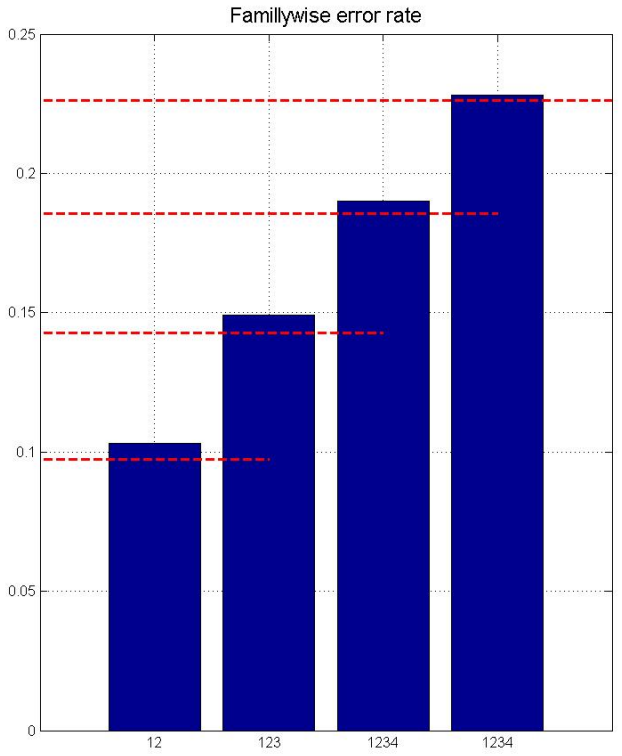
- for alpha =5/100, if we do 2 tests we should get about 1-(1-5/100)^2 ~ 9% false positives, if we do 126 electrodes * 150 time frames tests, we should get about 1-(1-5/100)^18900 ~ 100% false positives! i.e. you can't be certain of any of the statistical results you observe

What is the problem?

- Illustration with 5 independent variables from N(0,1)
- Repeat 1000 times and measures type 1 error rate



5%



22%

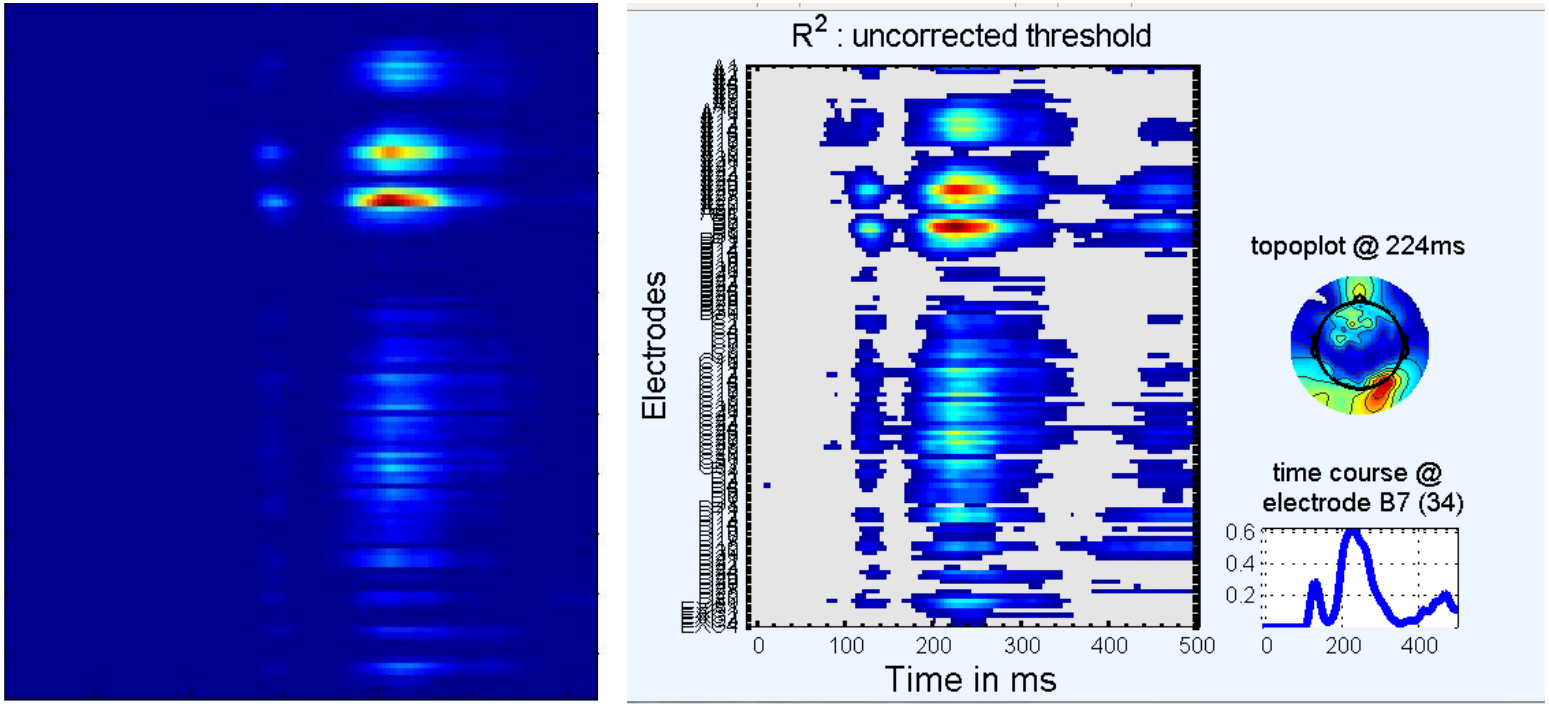
18%

14%

9%

What is the problem?

- Illustration with 18900 independent variables (126 electrodes and 150 time frames)



we know there are false positives – which ones is it?

Family Wise Error rate

- FWER is the probability of making one or more Type I errors in a family of tests, under H0
- H0 = no effect in any channel/time and/or frequency bins → implies that rejecting a single bin null hyp. is equal to rejecting H0

$$P(Ui \in V \uparrow \blacksquare \{T \downarrow i \geq u\} | H \downarrow 0 \ ) \leq \propto$$

We want to find the threshod u such the prob of any false positives under H0 is controlled at value alpha

Bonferroni Correction

Bonferroni correction allows to keep the FWER at 5% by simply dividing alpha by the number of tests

$$P(T_i \geq u \mid H_0) \leq \alpha/m$$ *Find u to keep the FWER < α/m*

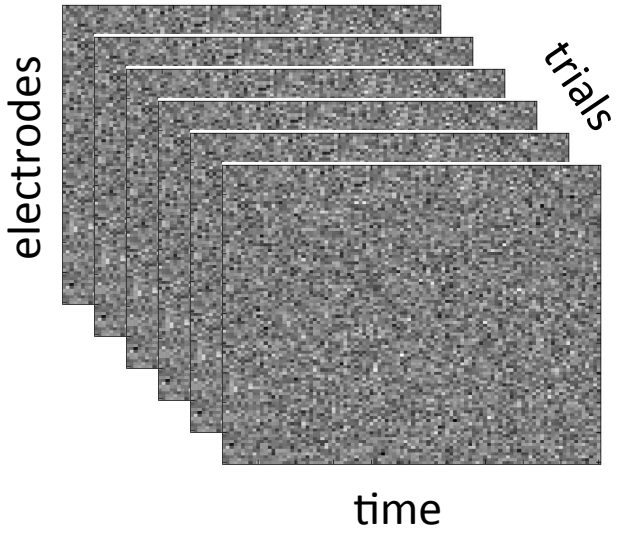
$$\text{FWER} = P(U_{i \in V} \{T_i \geq u\} \mid H_0) \leq \alpha$$

$$\leq \sum_i P(T_i \geq u \mid H_0)$$ *Boole's inequality*

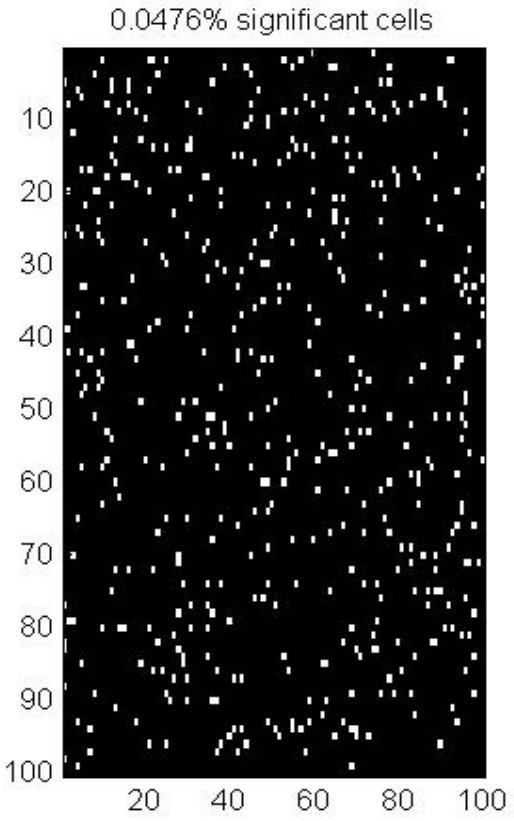
$$\leq \sum_i \alpha/m = \alpha$$

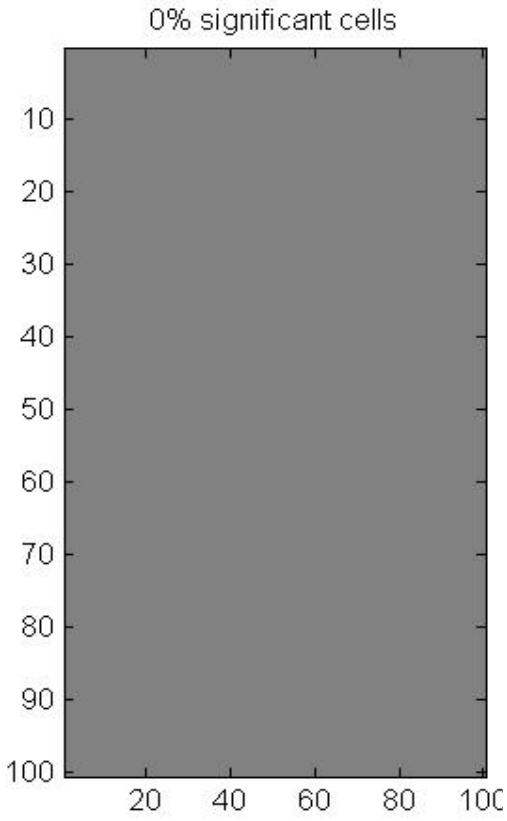
Bonferroni Correction

- Assumes all tests are independent
- Too conservative



One sample t test > 0 ?

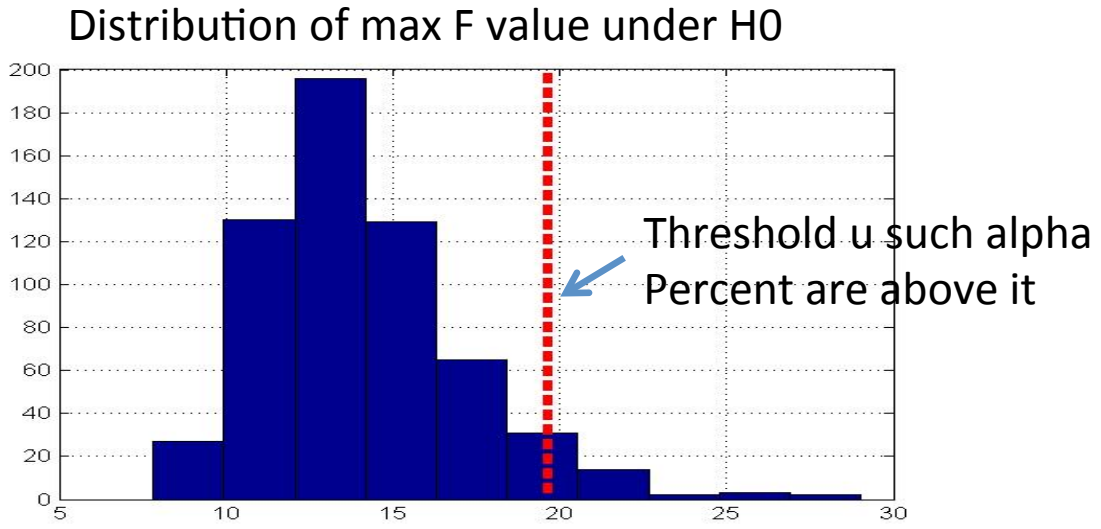




Correcting using the maximum under H0

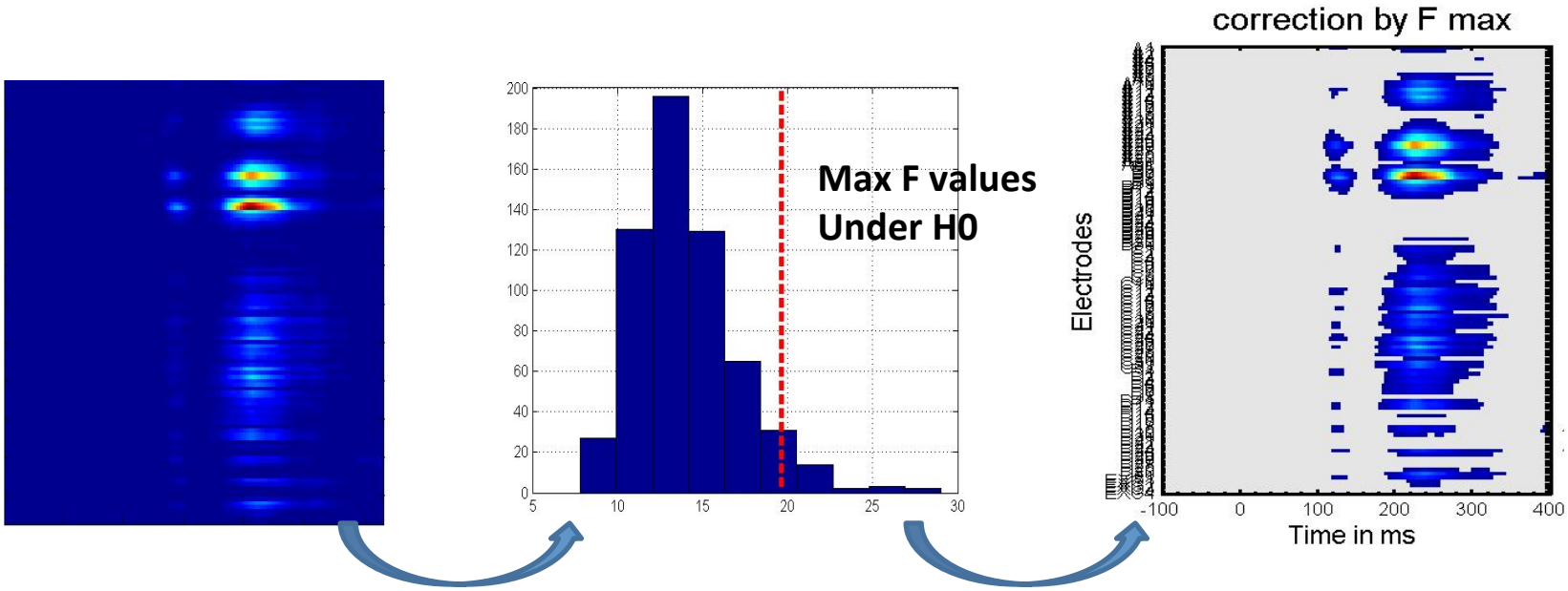
Maximum Statistics

- Since the FWER is the prob that any stats > u, then the FWER is also the prob. that the max stats > u
- All we have to do, is thus to find a threshold u such that the max only exceed u alpha percent of the time.



Maximum Statistics

- Estimate the distribution of max under H0 and simply threshold the observed results a threshold u
- Still assumes all tests are independent



Computational approaches to estimate H0

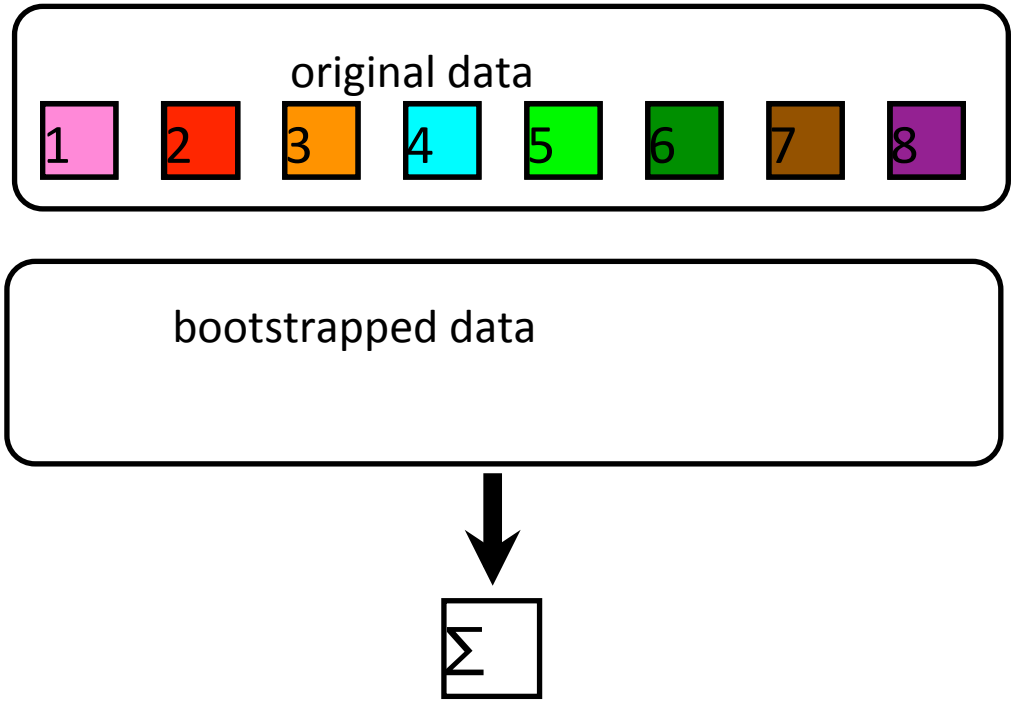
Slides provided by G. Rousselet – University of Glasgow

Bootstrap: central idea

- "The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates." Efron & Tibshirani, 1993
- "The central idea is that it may sometimes be better to draw conclusions about the characteristics of a population strictly from the sample at hand, rather than by making perhaps unrealistic assumptions about the population." Mooney & Duval, 1993

Percentile bootstrap: general recipe

(1) sample WITH replacement n observations (under H1 for CI of an estimate, under H0 for the null distribution)

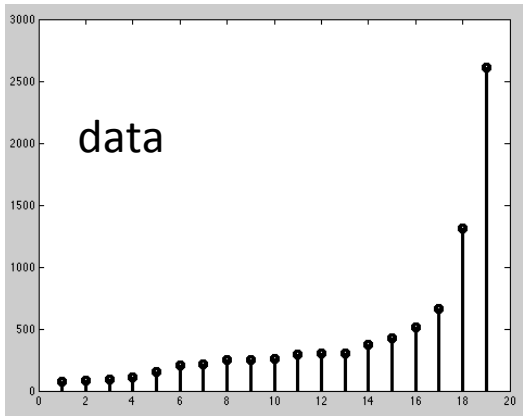


(2) compute estimate
e.g. sum, trimmed mean

(3) repeat (1) & (2) b times $\Sigma_1\ \Sigma_2\ \Sigma_3\ \Sigma_4\ \Sigma_5\ \Sigma_6\ \ldots\ \Sigma_b$
(4) sort the b estimates*
(5) get 1-alpha confidence interval

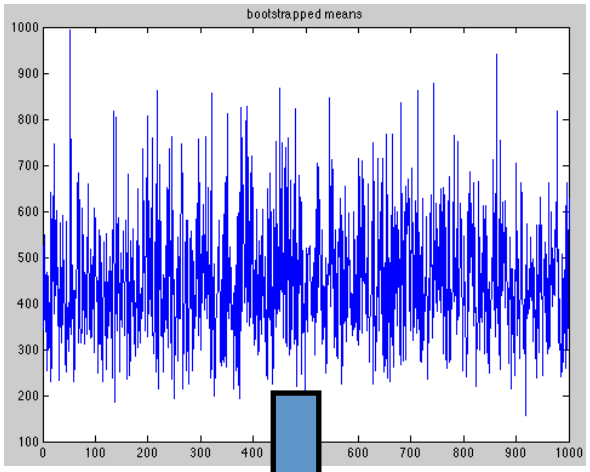
Percentile bootstrap estimate of mean

% self-awarness data, Wilcox, 2005, p58



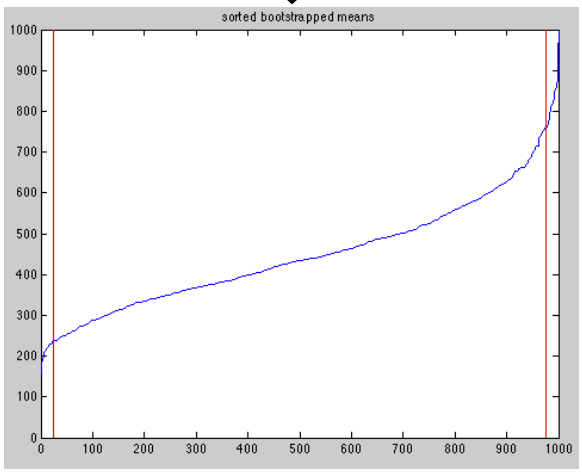
Sample with replacement b times

Bootstrapped estimates



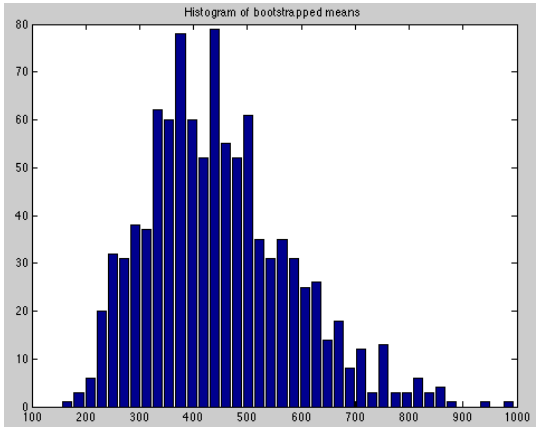
compute estimate

Sort & get CI

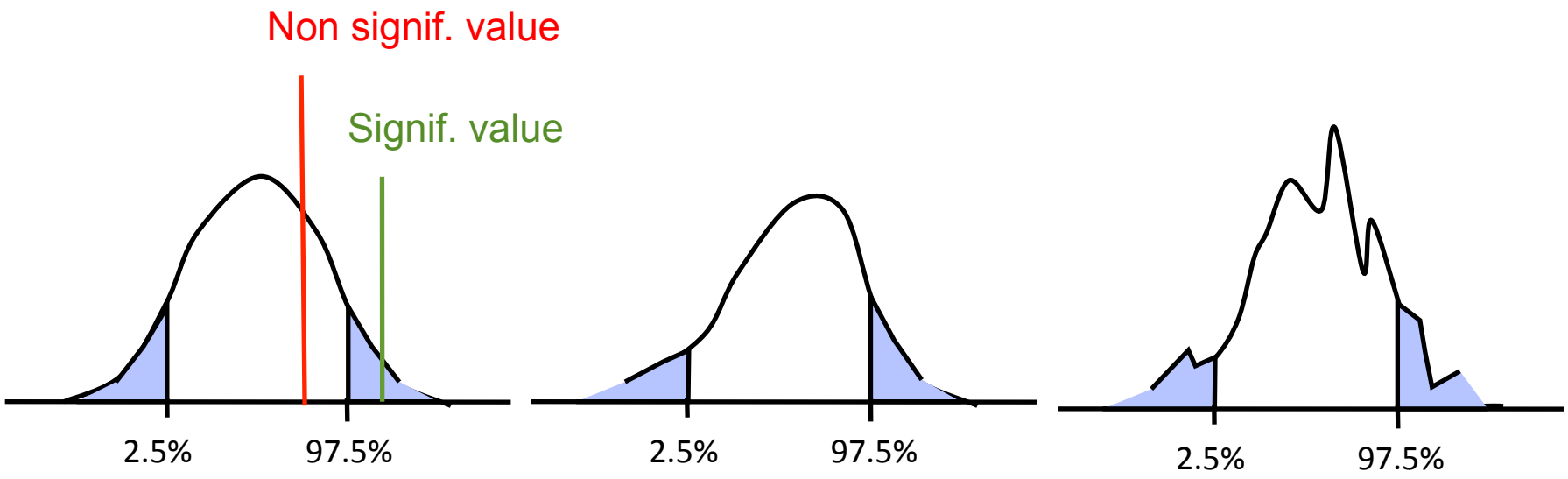


get PDF

Distribution of bootstrapped estimates of the mean



Distributions can take any shape

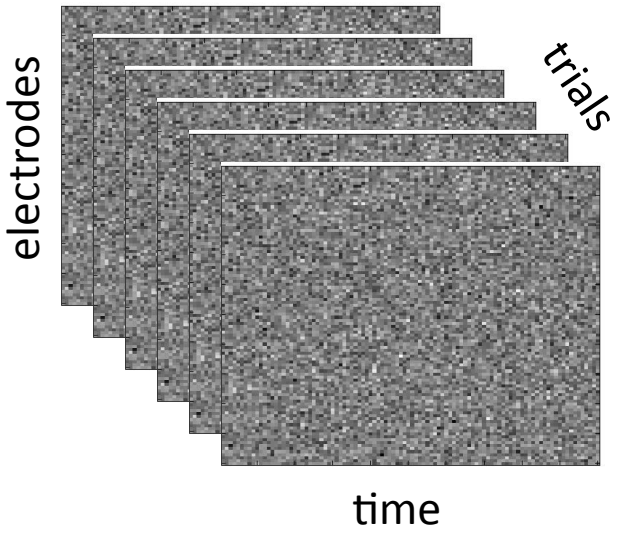


The percentile bootstrap method allows the bootstrap estimate of the sampling distribution to conform to any shape the data suggest, taking into account the variance and the skewness of the sample. This can be the distribution of all T/F values, and thus we can estimate the distribution of the maximum T/F value under the null.

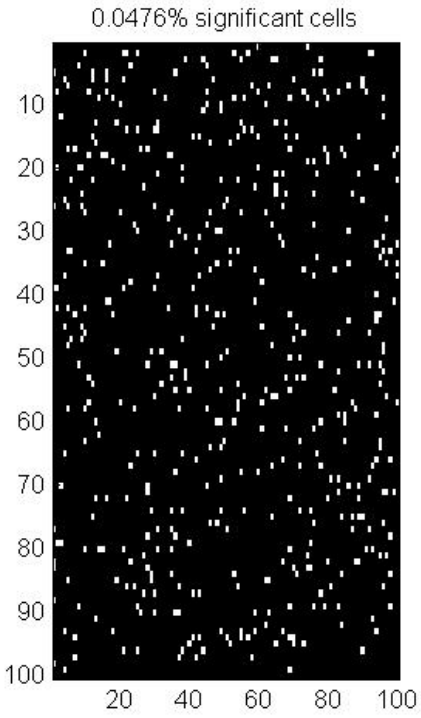
Cluster Mass for MEEG

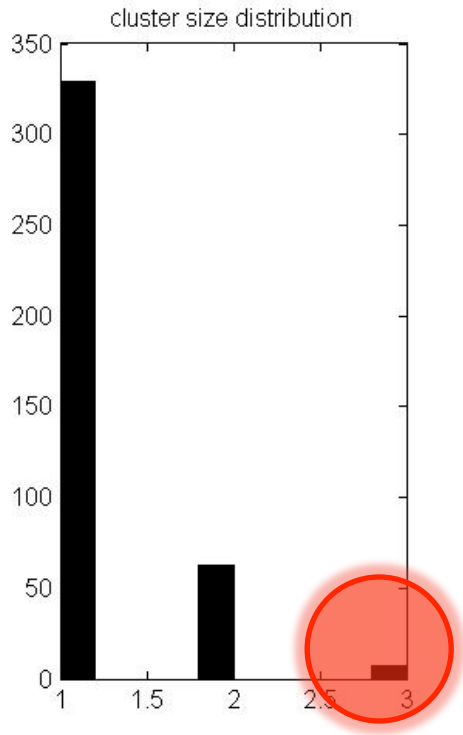
Let's analyse clusters

- In MEEG, instead of the max, we consider clusters as it is much less likely that statistics are significant in groups



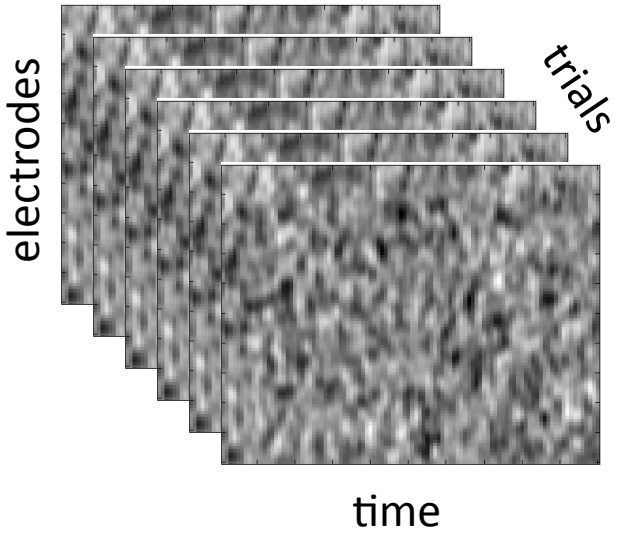
One sample t test > 0 ?



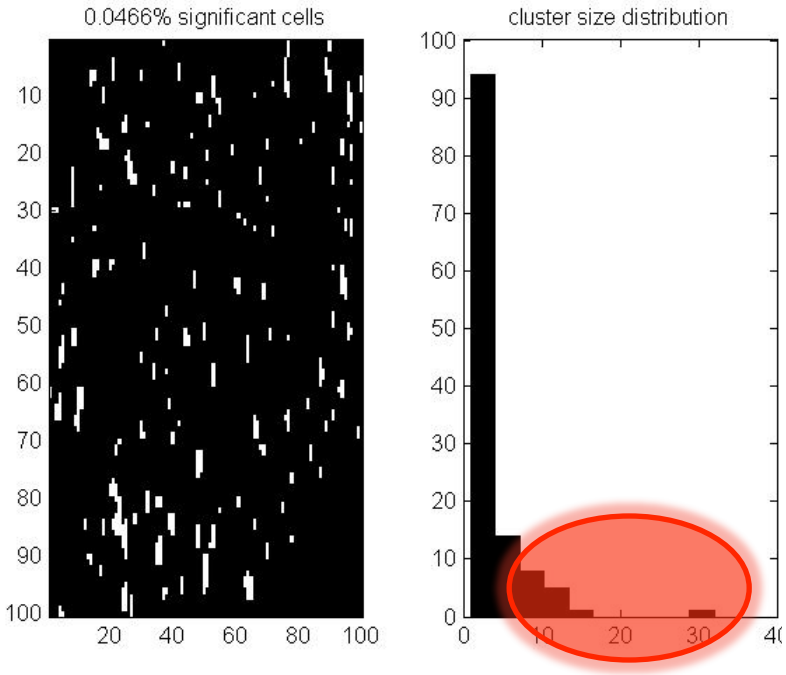


Let's analyse clusters

- In MEEG, instead of the max, we consider clusters as it is much less likely that statistics are significant in groups because data are smooth in space and time!



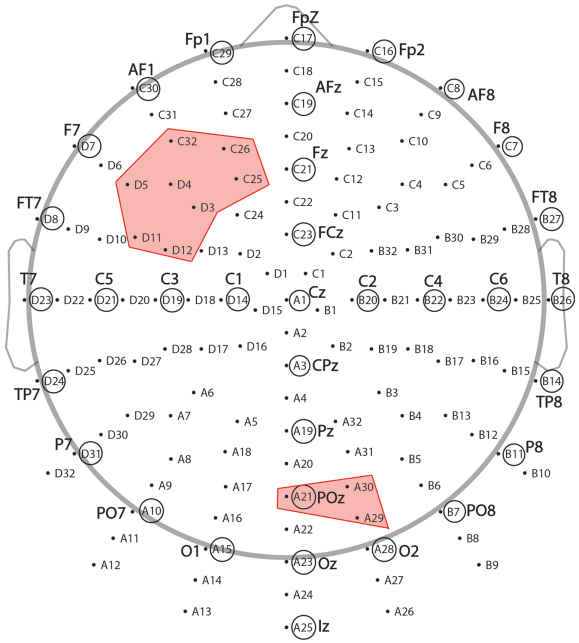
One sample t test > 0 ?

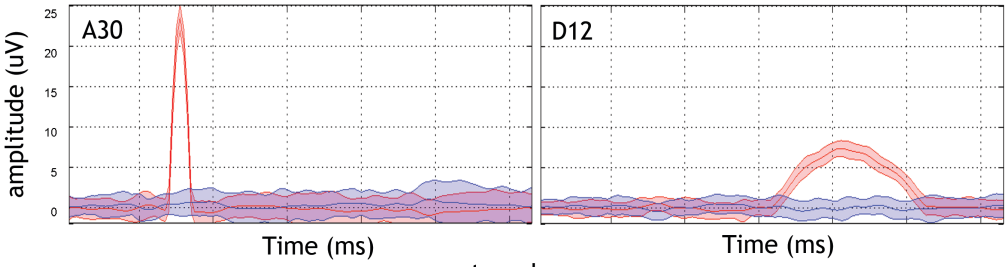


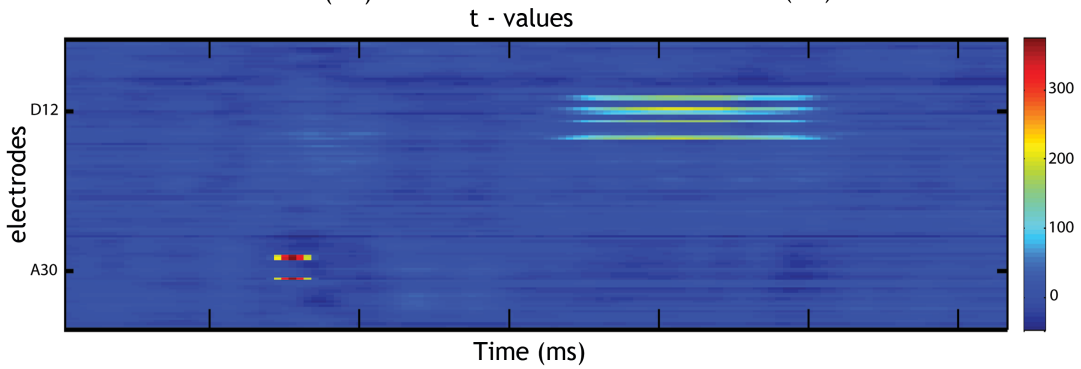
The clustering solution

- Clustering is a good option because it accounts for topological features in the data. Techniques like Bonferroni, FDR, max(stats) control the FWER but independently of the correlation between tests.

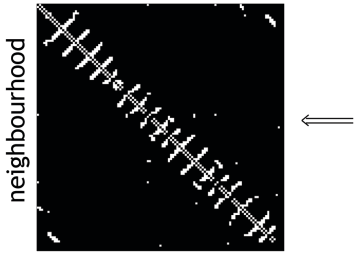
- To use clustering we need to consider cluster statistics rather than individual statistics
- Cluster statistics depend on (i) the cluster size, which depends on the data at hand (how correlated data are in space and in time/frequency), and (ii) the strength of the signal (how strong are the t, F values in a cluster) or (iii) a combination of both.







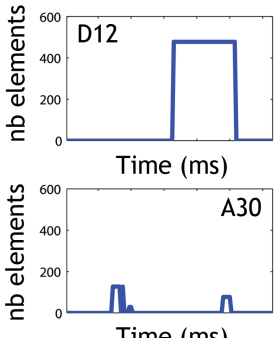
Spatial - Temporal clustering



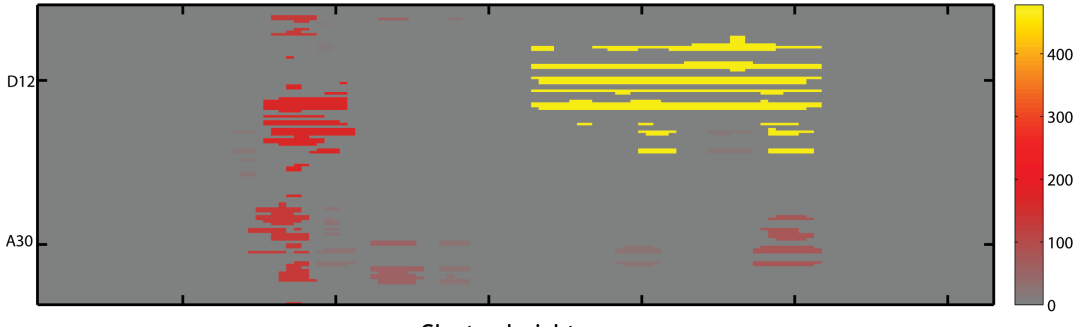
p values < 0.05

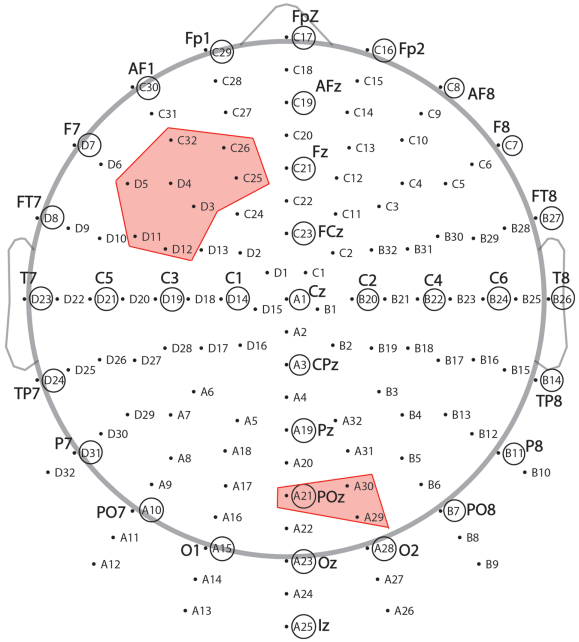
maximum extent
= number of
electrodes and
time points

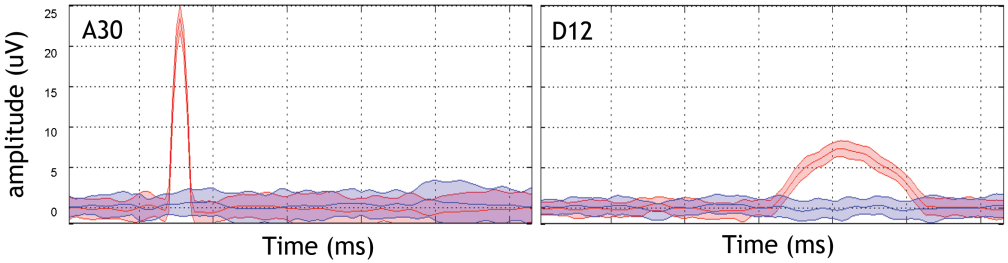
cluster 1 = 478
cluster 2 = 127

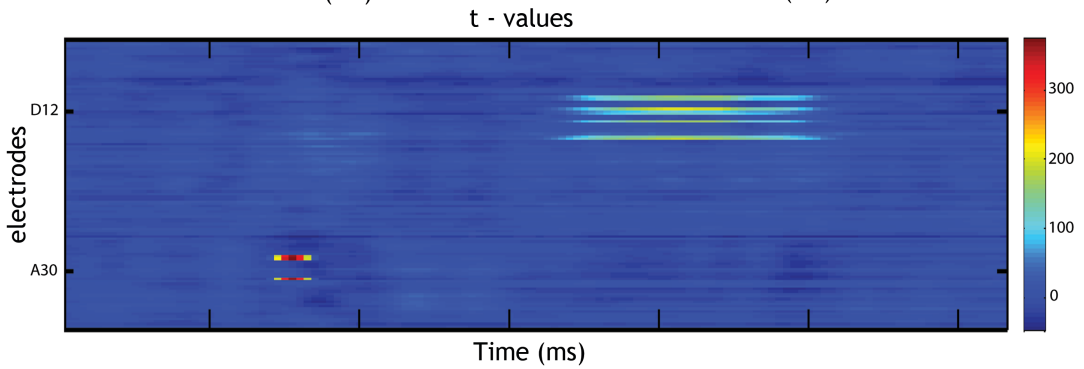


Cluster extent

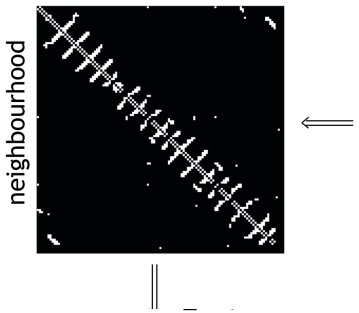






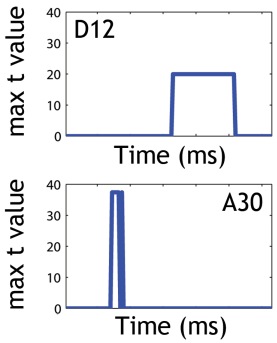


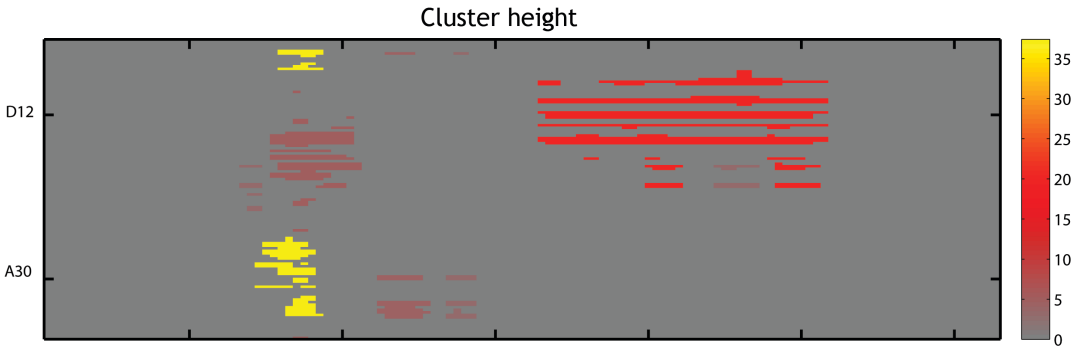
Spatial - Temporal clustering

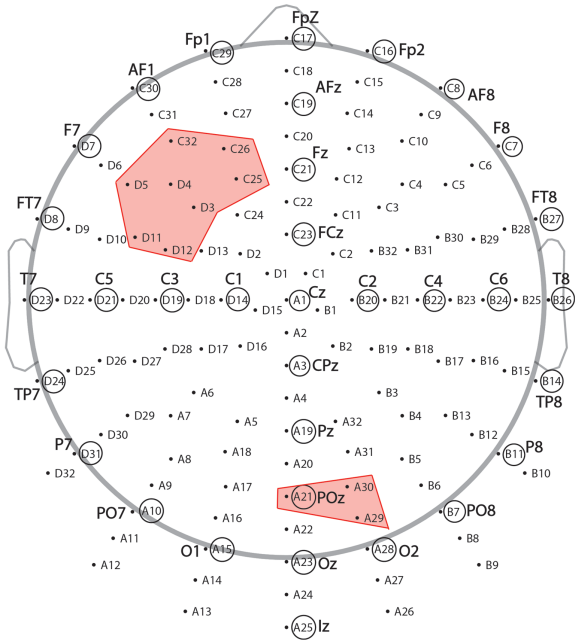


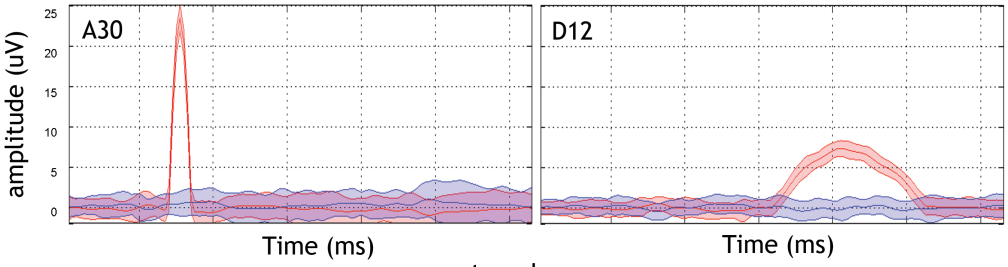
maximum height
within a cluster
of electrodes
and time points

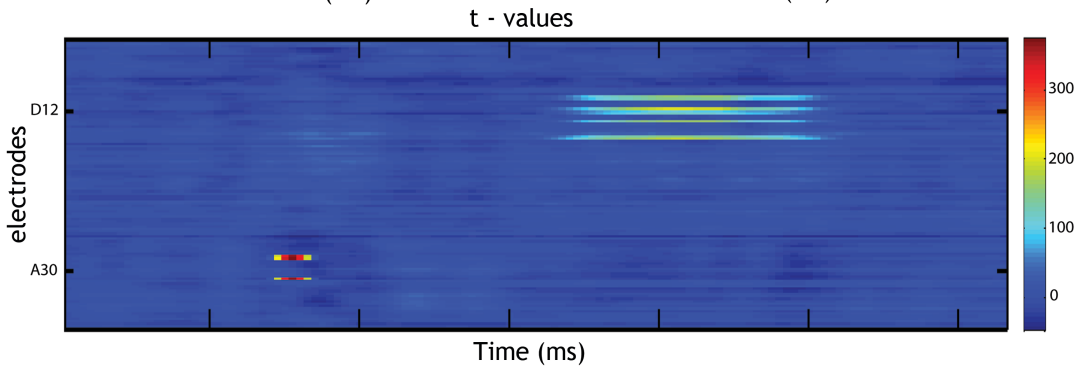
cluster 1 = 19.7
cluster 2 = 37.4



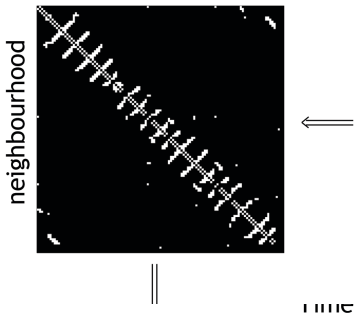








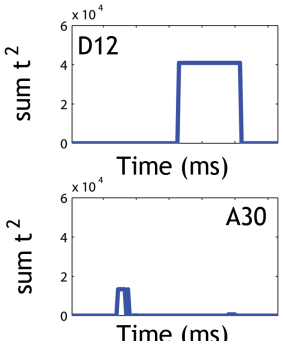
Spatial - Temporal clustering



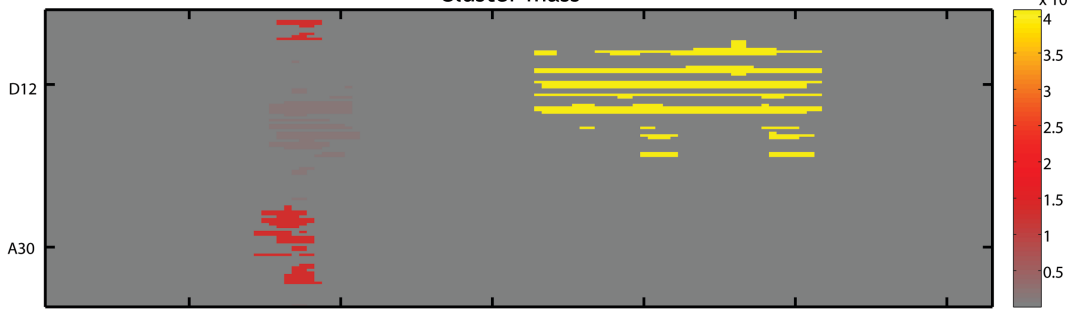
p values < 0.05

mass (sum t^2) of values within a cluster of electrodes and time points

cluster 1 = 40984
cluster 2 = 13386

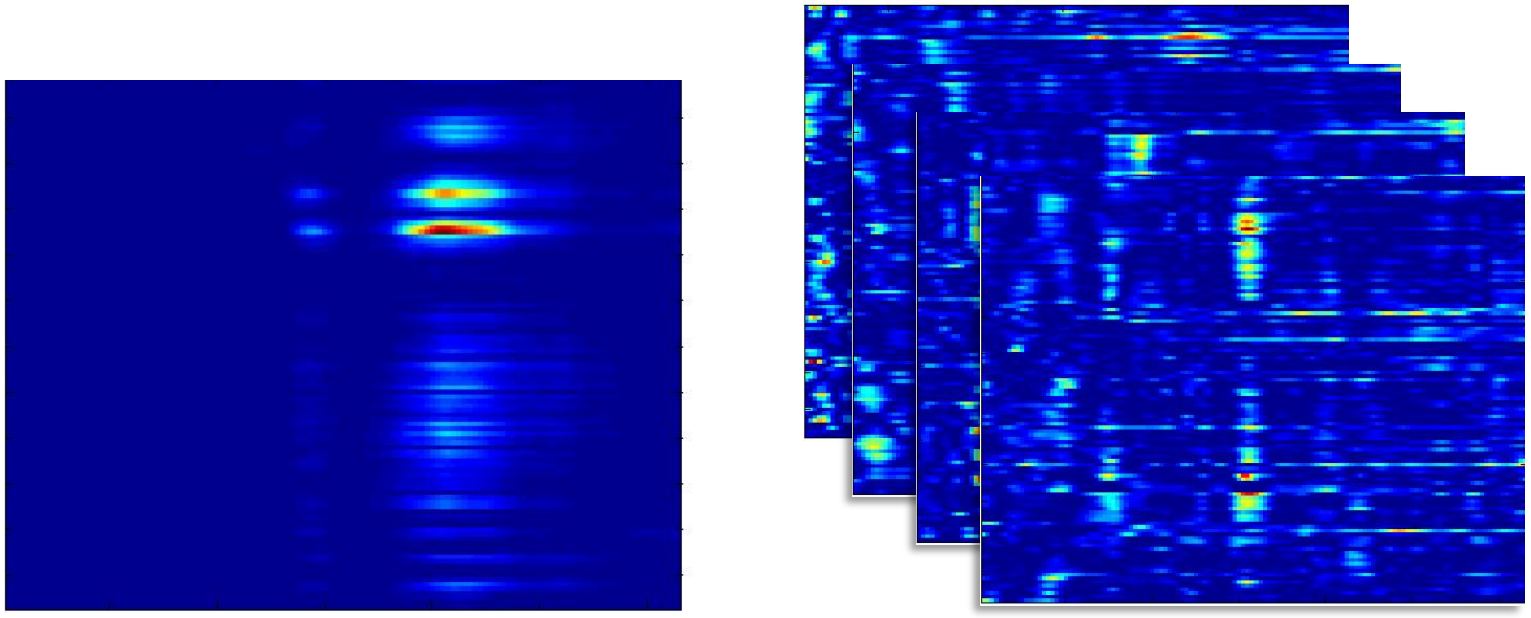


Cluster mass



The clustering solution

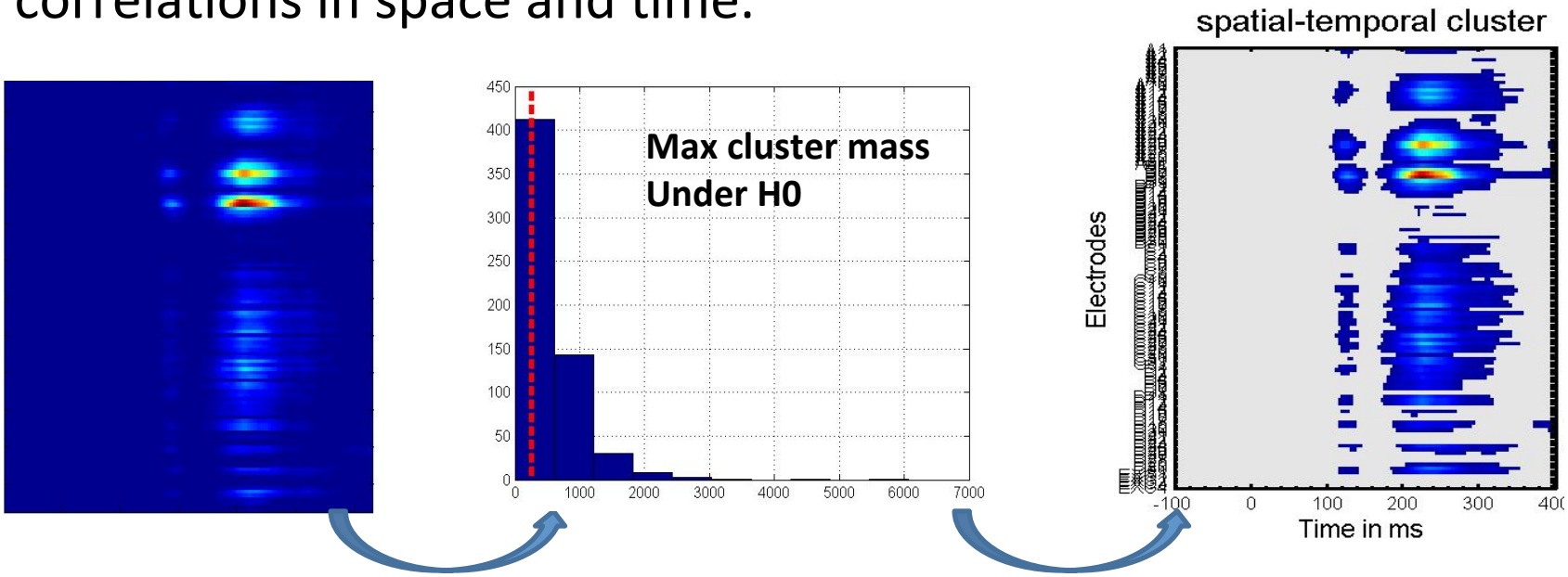
- In LIMO EEG, we bootstrap the data under H0: center the data or break the link between the design matrix and the data and then resample and test. This way we can find u for a single bin, the the whole space, or for clusters.



Observed F values → F values under H0

The clustering solution

- Spatial-Temporal clustering: for each bootstrap, threshold at alpha and record the max(cluster mass), i.e. sum of F values within a cluster. Then threshold the observed clusters based on there mass using this distribution → accounts for correlations in space and time.



Loss of resolution: inference is about the cluster, not max in time or a specific electrode !

TFCE for MEEG

Threshold Free Cluster Enhancement

- Threshold Free Cluster Enhancement (TFCE): Integrate the cluster mass at multiple thresholds. A TFCE score is thus obtain per cell but the value is a weighted function of the statistics by it's belonging to a cluster.

Smith & Nichols 2009 NeuroImage 44

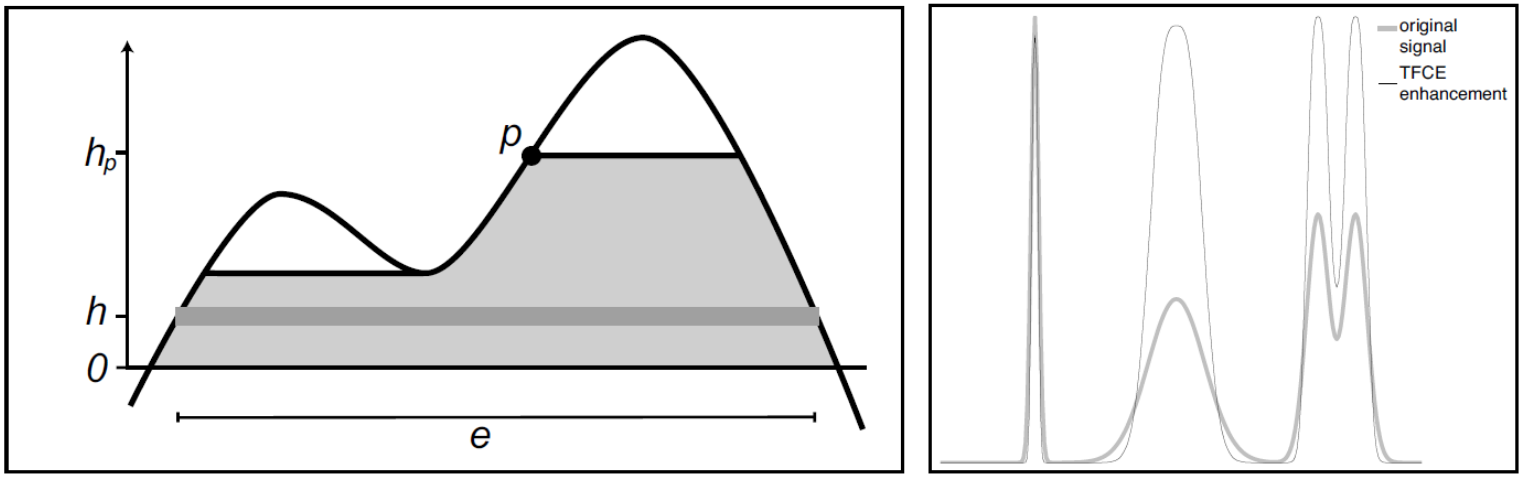
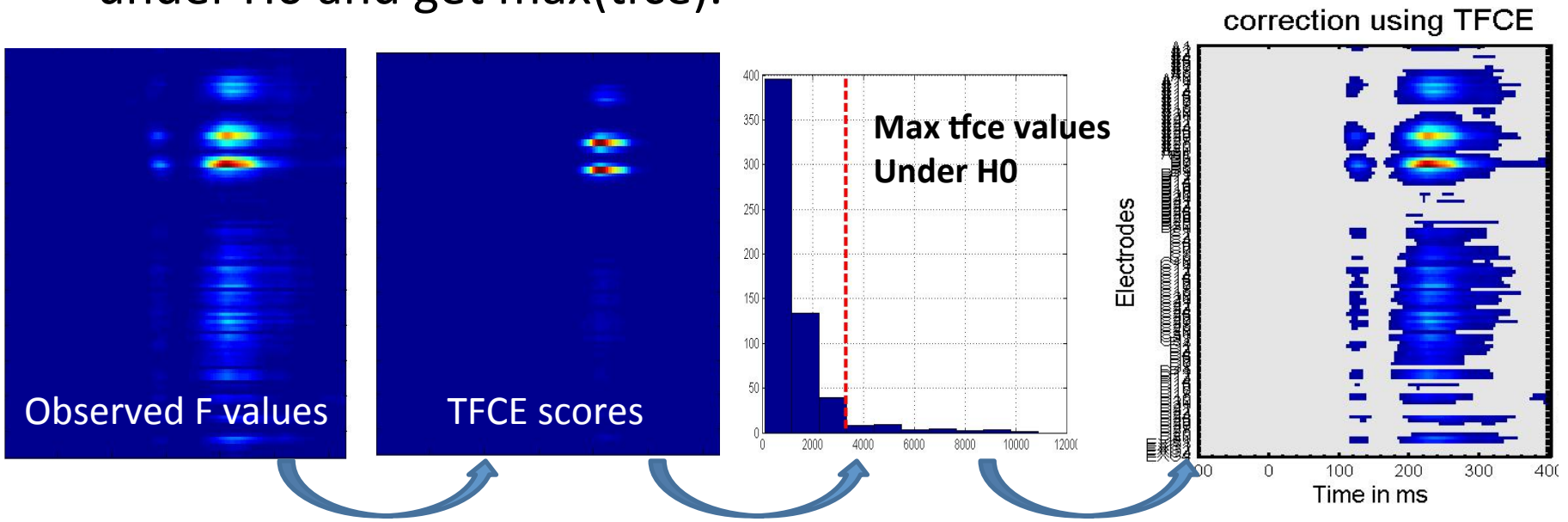


Figure 1: Illustration of the TFCE approach. Left: The TFCE score at voxel p is given by the sum of the scores of all incremental supporting sections (one such is shown as the dark grey band) within the area of "support" of p (light grey). The score for each section is a simple function of its height h and extent e. Right: Example input image and TFCE-enhanced output. The input contains a focal, high signal, a much more spatially extended, lower, signal and a pair of overlapping signals of intermediate extent and height. The TFCE output has the same maximal values for all three cases, and preserves the distinct local maxima in the third case.

Threshold Free Cluster Enhancement

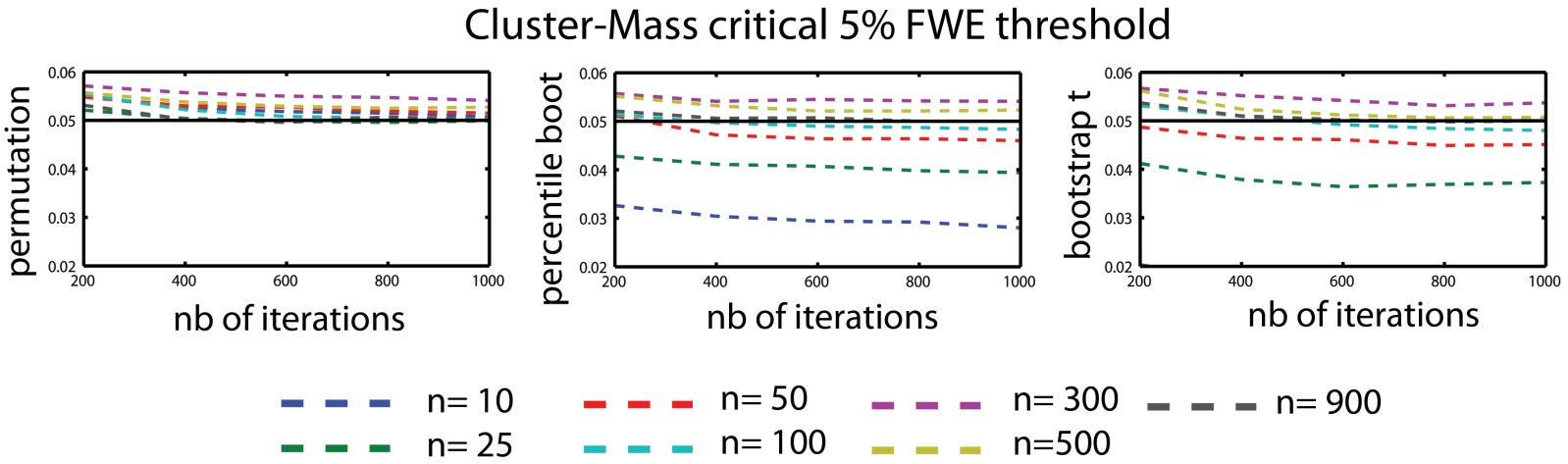
- Threshold Free Cluster Enhancement (TFCE): Integrate the cluster mass at multiple thresholds. A TFCE score is thus obtain per cell but the value is a weighted function of the statistics by it's belonging to a cluster. As before, bootstrap under H0 and get max(tfce).



Excellent resolution: inference is about cells, but we accounted for space/time dependence

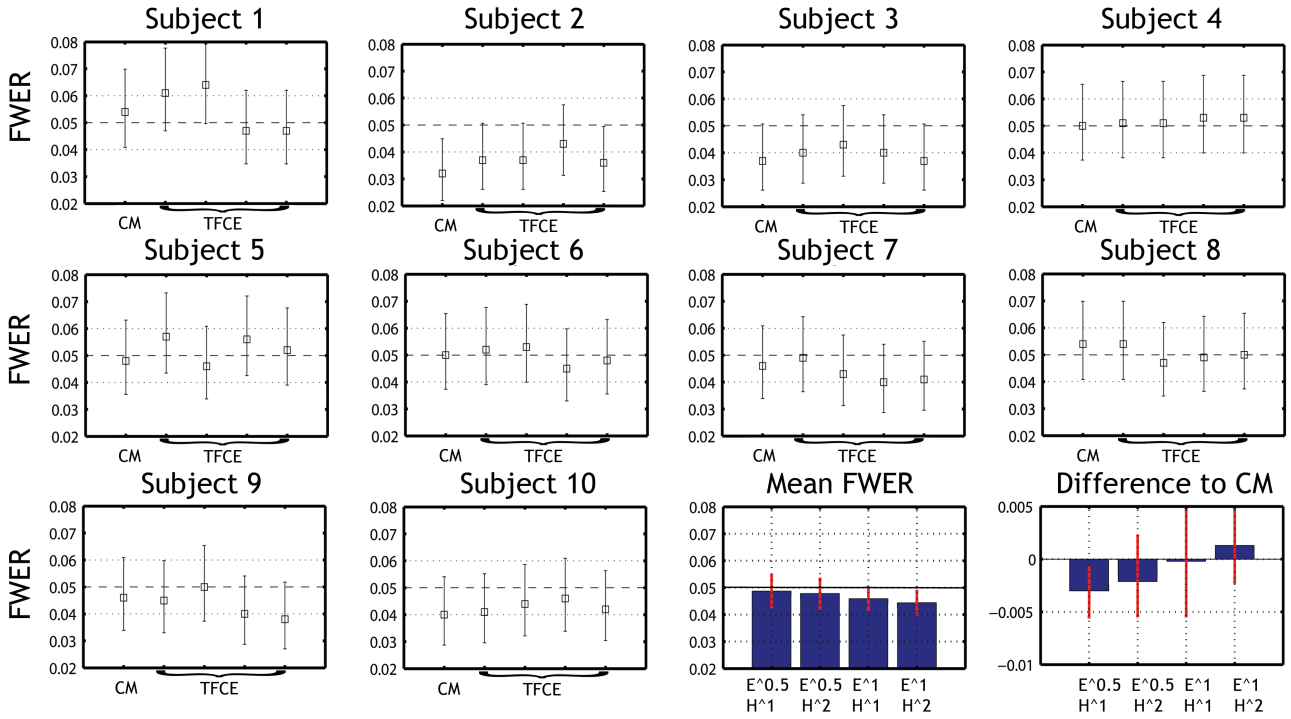
Review of techniques

- All techniques (including permutation not shown here) control well the FWER under H0 with some limitations for small sample sizes



Review of techniques

- All techniques (including permutation not shown here) control well the FWER under H0 with some limitations for small sample sizes



MCC summary

- Simulation work show that overall permutation / bootstrap / cluster-mass / TFCE control well the type 1 FWER.
- a minimum of 800 iterations are necessary to obtain stable results
- for low critical family-wise error rates (e.g. p = 1%), permutations can be too liberal;
- For within subject bootstrap, a min of 50 trials per condition is requested at the risk to be too conservative

Conclusions

- When performing multiple tests, statistical correction MUST be applied.

- All techniques provide a FWER at the specified level but not all techniques have the same power.
- Spatial-temporal clustering and TFCE seem to provide good estimates, with TFCE giving higher spatio-temporal inference resolution, but at the cost of long computing time.

References

- **Maris, E. & Oostenveld, R. (2007).** Nonparametric statistical testing of EEG- and MEG-data. Journal of *Neuroscience Methods, 164*, 177-190

- **Pernet, C., Latinus, M., Nichols, T. & Rousselet, G.A. (2015).** Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. Journal of *Neuroscience Methods, 250*, 85-93