Multiple comparisons corrections and the use of bootstrap

Cyril Pernet, PhD
Neurobiology Research Unit,
Copenhagen University Hospital, Rigshospitalet

30th EEGLAB workshop – 17th June 2021

cyril.pernet@ed.ac.uk @CyrilRPernet

Motivation for whole channel/IC analyses

- **Data collection** consists in recording electromagnetic events over the whole brain and for a relatively long period of time, with regards to neural spiking.
- In the majority of cases, **data analysis** consists in looking where we have signal and restrict our analysis to these channels and components.

➢ Are we missing the forest by choosing working on a single, or a few trees?

➢ By analysing where we see an effect, we increase the type 1 FWER because the effect is partly driven by random noise (*solved if chosen based on prior results or split the data*)

Rousselet & Pernet – It's time to up the Game Front. Psychol., 2011, 2, 107

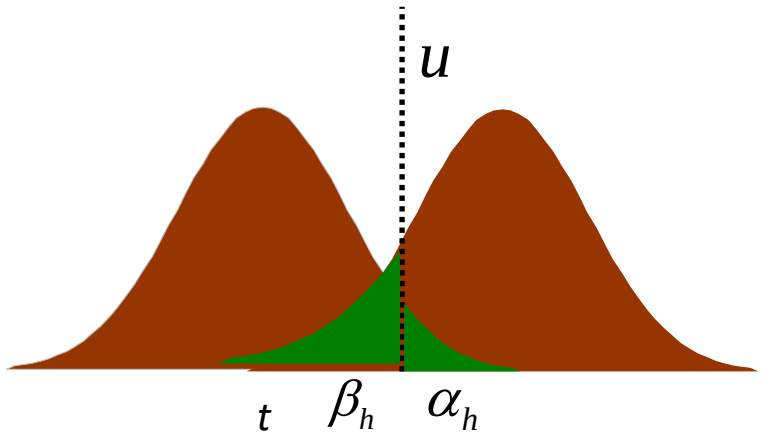
Motivation for whole channel/IC analyses

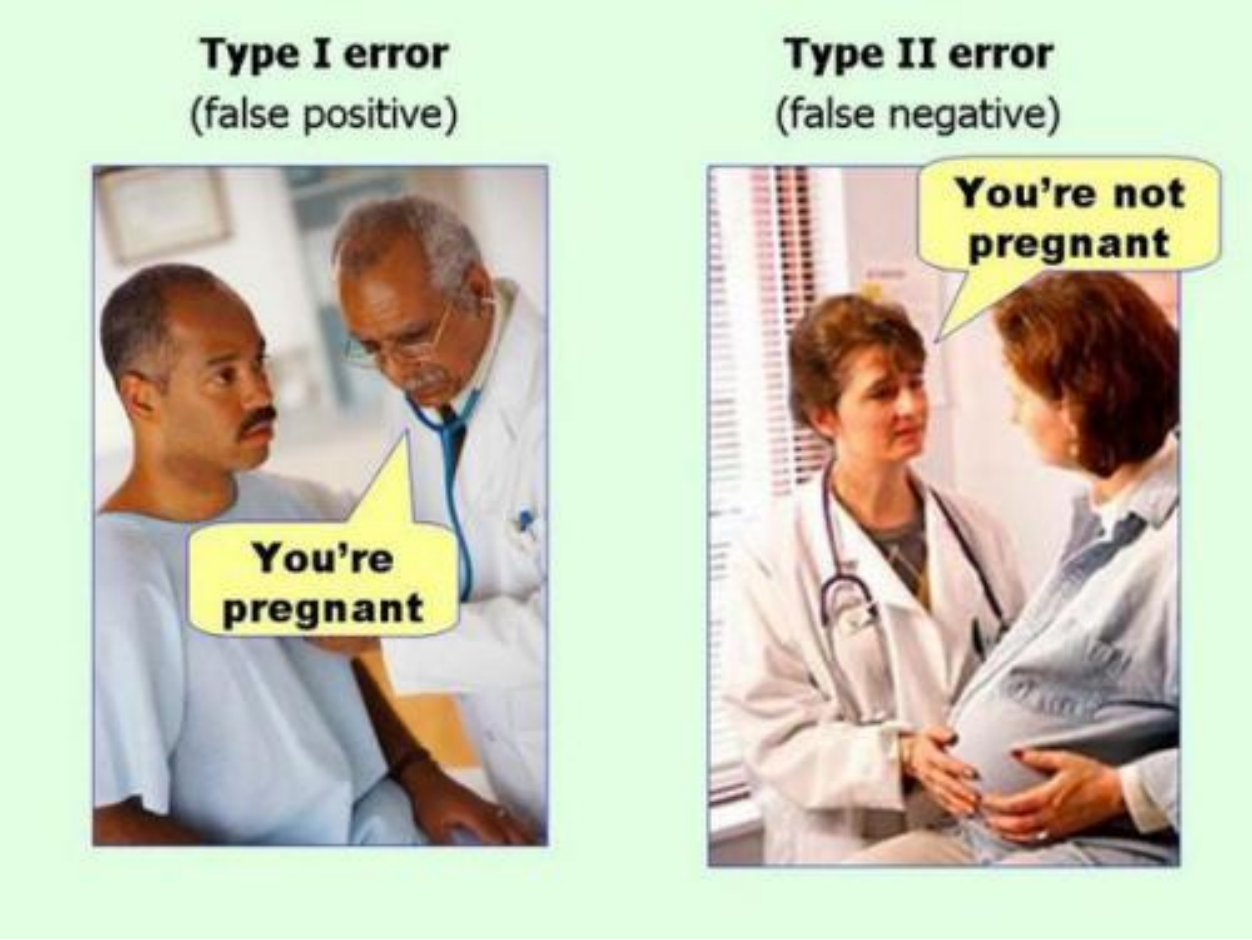
- Statistics on peak latencies and amplitudes? But several lines of evidence suggest that peaks mark the end of a process and therefore it is likely that most of the interesting effects lie in a component before a peak
- **Neurophysiology**: whether ERPs are due to additional signal or to phase resetting effects a peak will mark a transition such as neurons returning to baseline, a new population of neurons increasing their firing rate, a population of neurons getting on / off synchrony.
- **Neurocognition**: reverse correlation techniques showed that e.g. the N170 component reflects the integration of visual facial features relevant to a task at hand (Schyns and Smith) and that the peak marks the end of this process.

Rousselet & Pernet – It's time to up the Game Front. Psychol., 2011, 2, 107

Pearson-Newman hypothesis testing

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





https://chemicalstatistician.wordpress.com

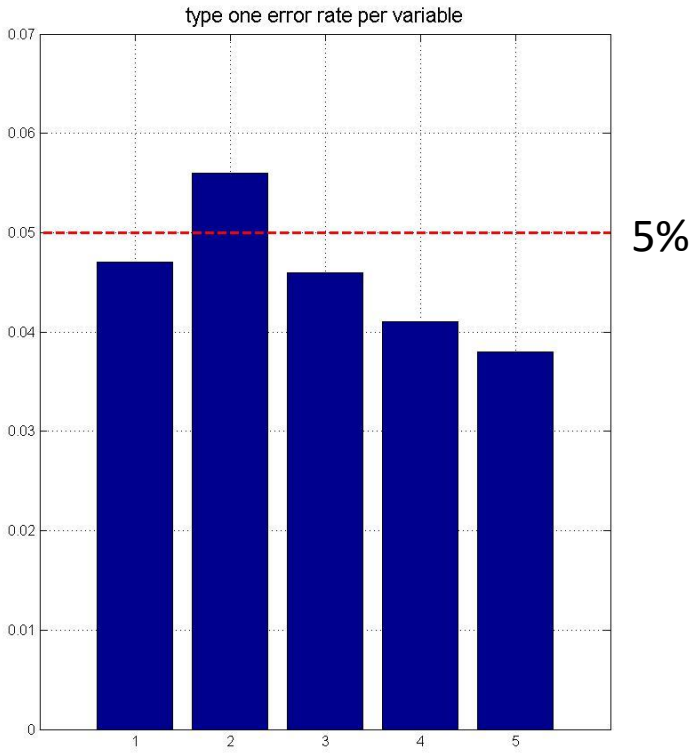
What is the problem?

- FWER is the probability of making one or more Type I errors (false positive) in a family of tests, under H0
- Assuming tests are independents from each other, the family-wise error rate FWER = 1 - (1 - alpha)^n

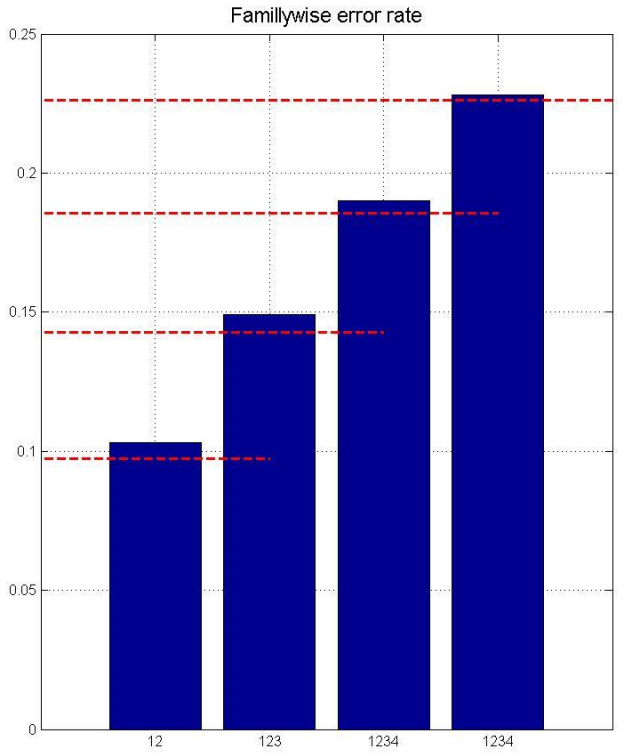
- for alpha =5/100, if we do 2 tests we should get about 1-(1-5/100)^2 ~ 9% false positives, if we do 126 electrodes * 150 time frames tests, we should get about 1-(1-5/100)^18900 ~ 100% false positives! i.e. you can’t be certain of any of the statistical results you observe

What is the problem?

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



5%



22%

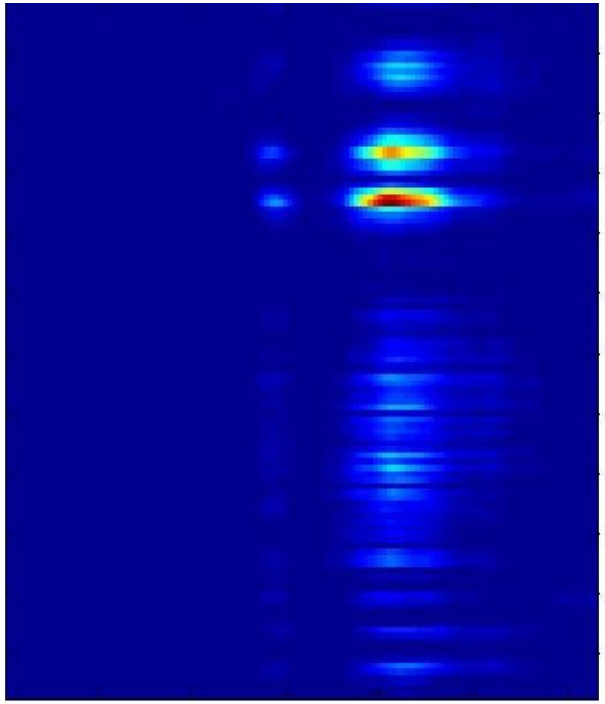
18%

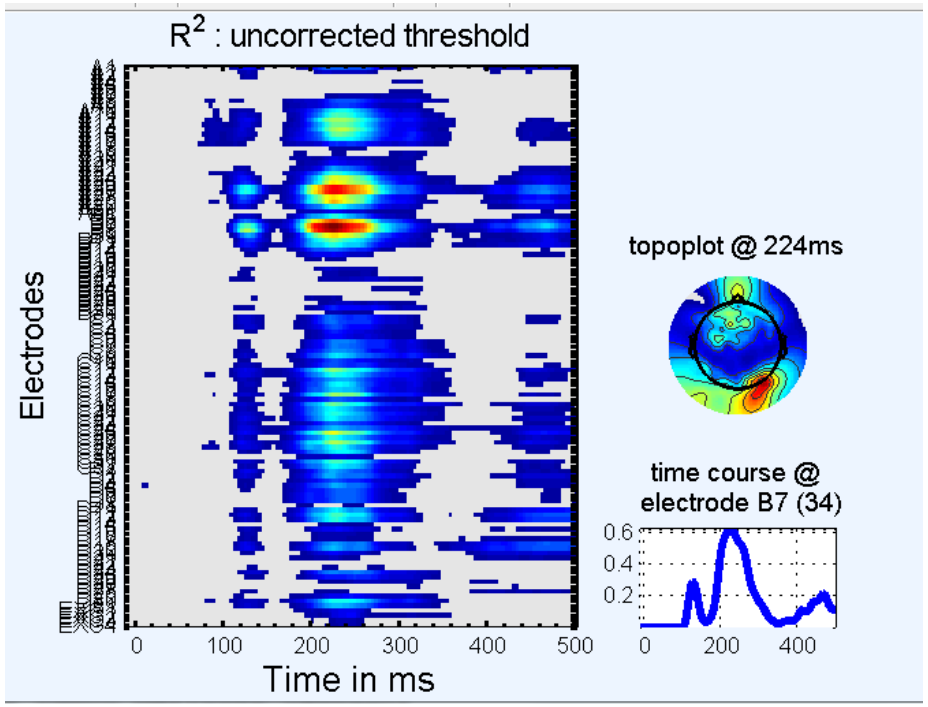
14%

9%

What is the problem?

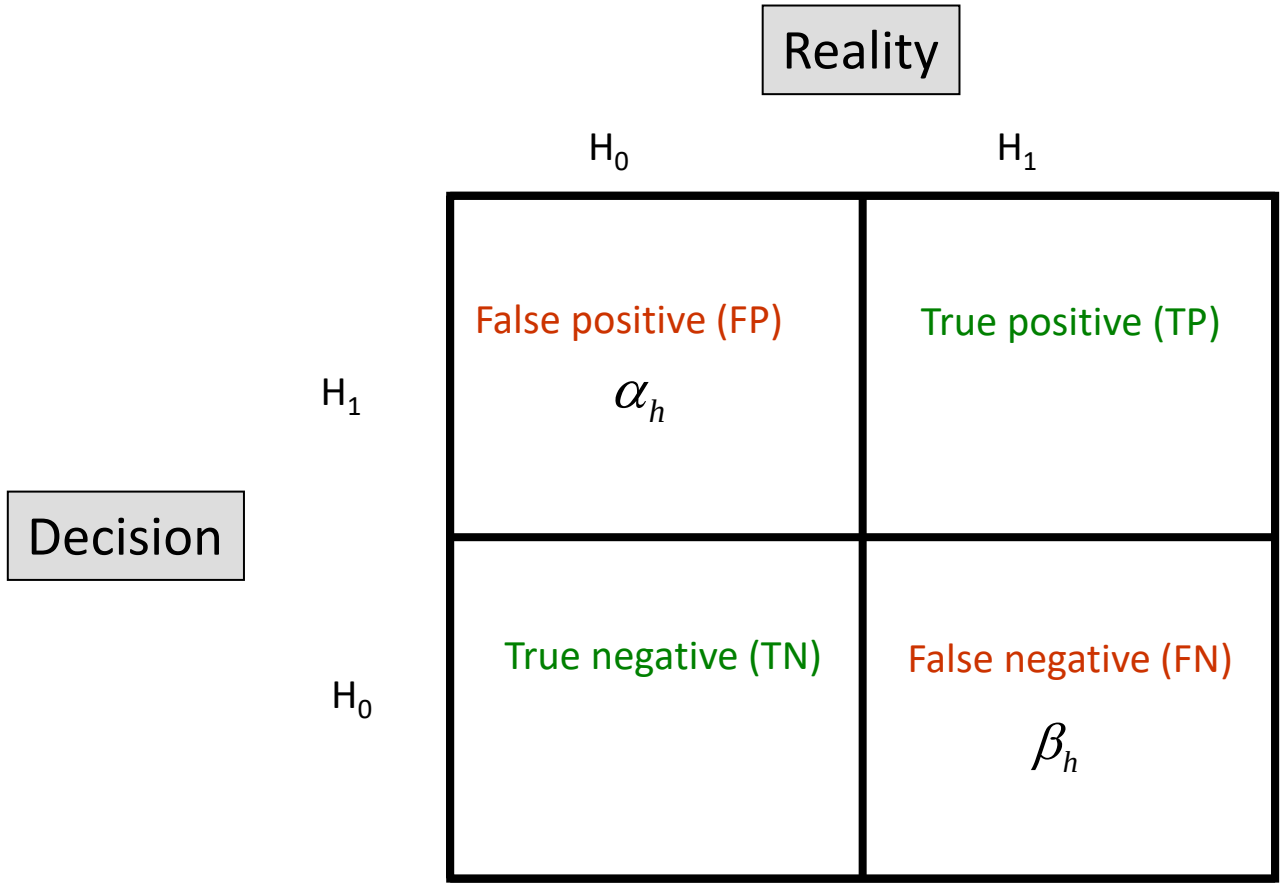
- Illustration with 18900 independent variables (126 electrodes and 150 time frames)





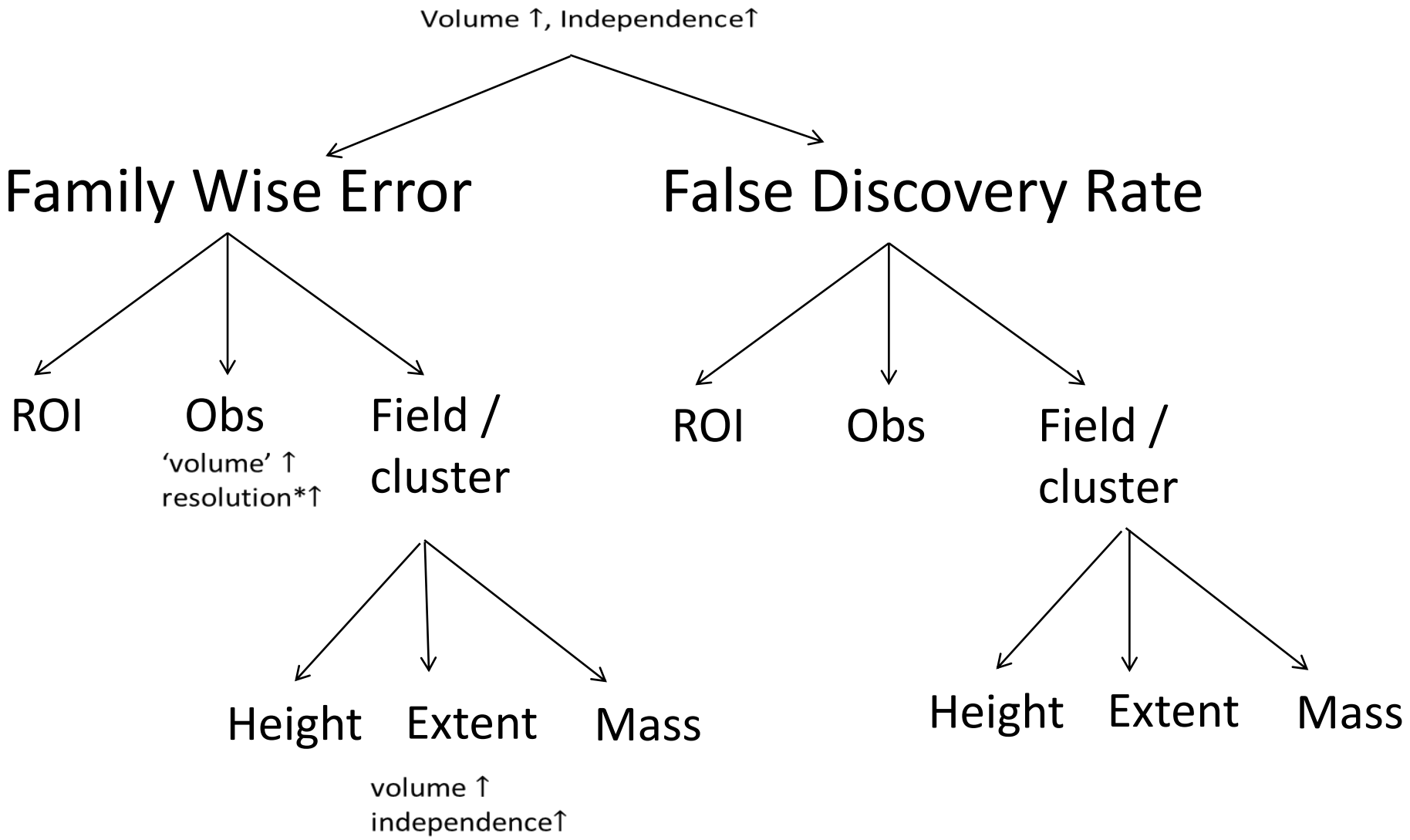
we know there are false positives – which ones is it?

Types of error

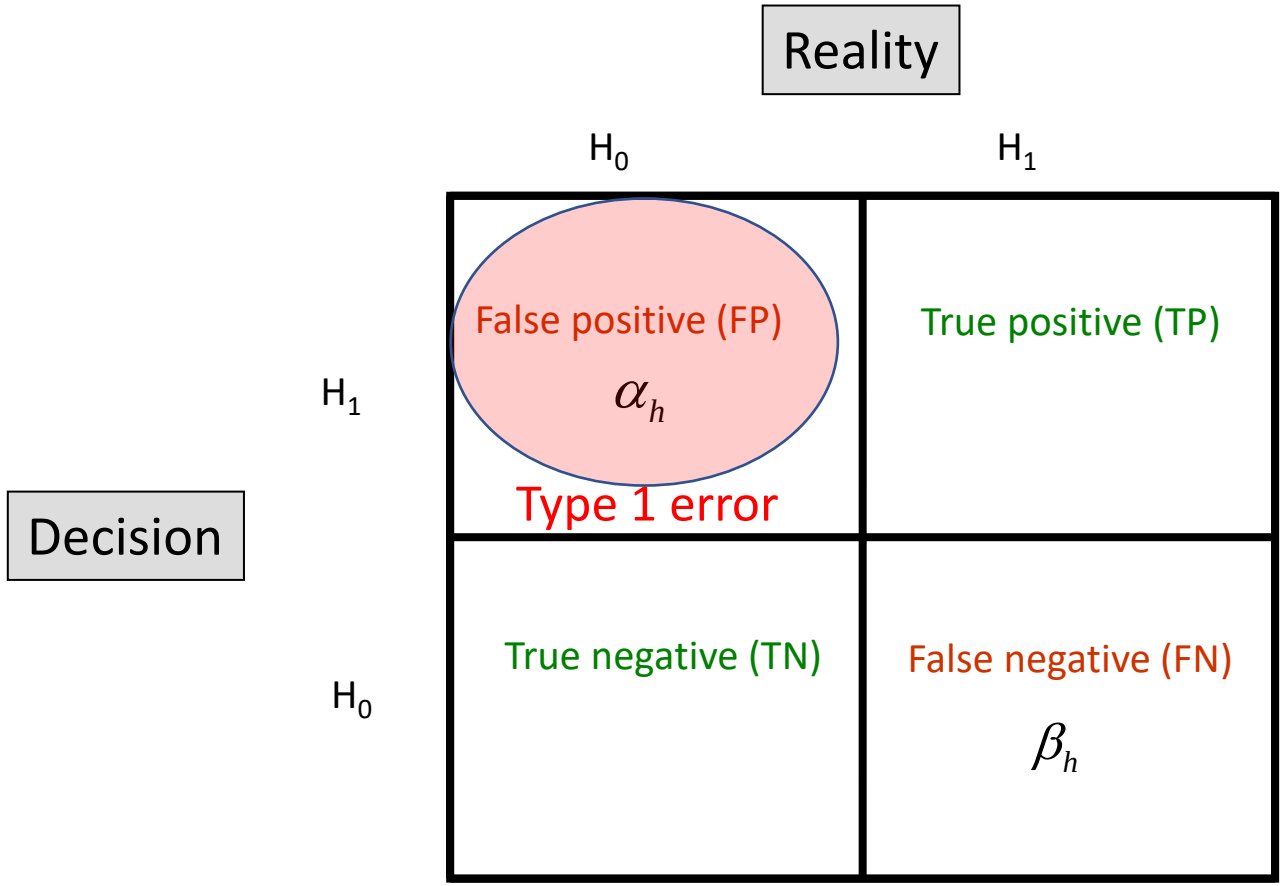


From J Chumbley's slide 2015

Detect an effect of *unknown* extent & location



Types of error and control



False discovery rate
Among all positives control the rate q
FDR = FP / (FP + TP)

specificity: 1- α_h
= TN / (TN + FP)
= proportion of actual negatives which are correctly identified

sensitivity (power): 1- β_h
= TP / (TP + FN)
= proportion of actual positives which are correctly identified

From J Chumbley's slide 2015

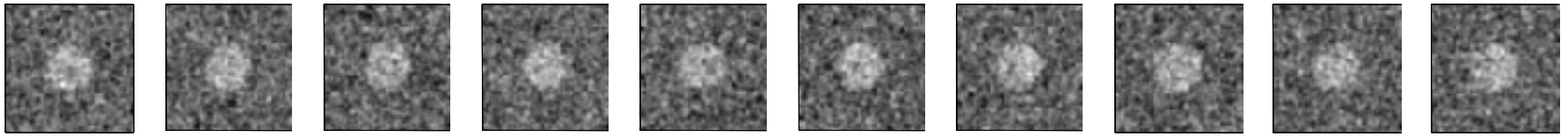
False Discovery Rate

- Whereas family wise approach corrects for any false positive, the FDR approach aim at correcting among positive results only.

1. Run an analysis with alpha = x%
2. Sort the resulting positive data
3. Threshold among all significant results

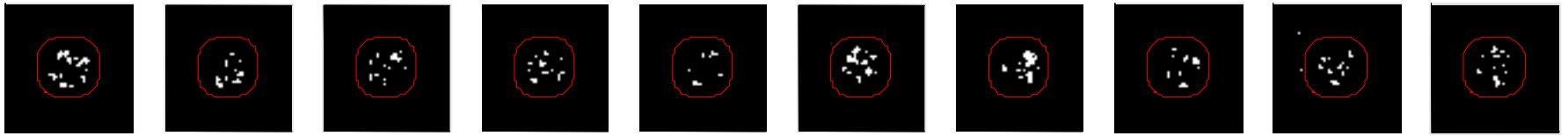
False Discovery Rate

Signal+Noise



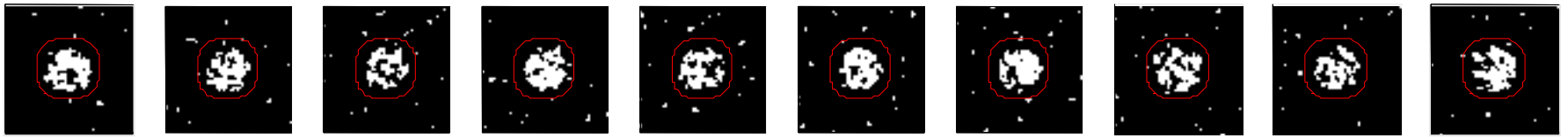
FWE (RFT) correction

As it models truth, it can control false positives only



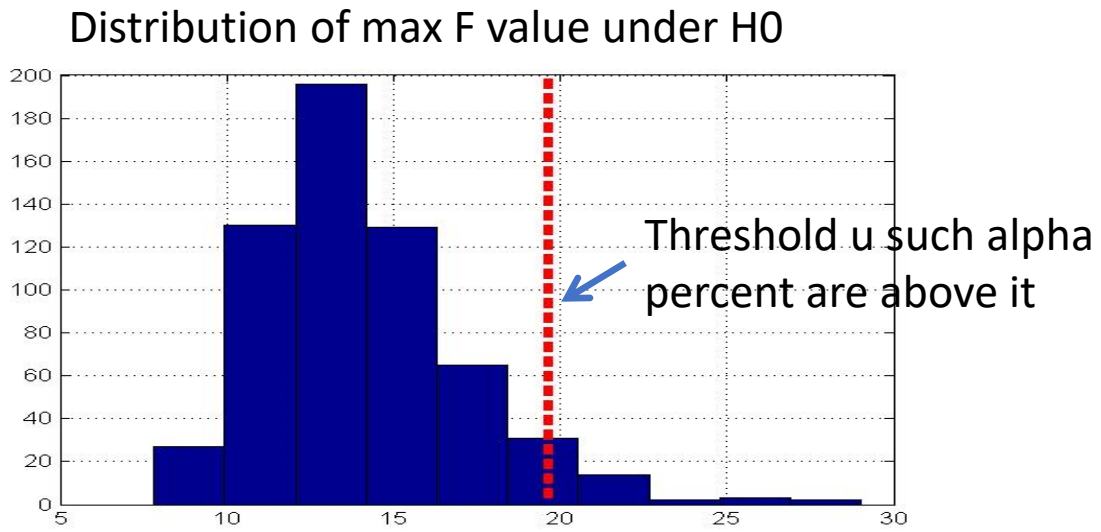
FDR correction

As it doesn't distinguish True from false, both remains



Family Wise Error Rate control

- Since the type 1 FWER is the prob that any stats > u, then it is also the prob. that the max stats > u (since max is one of any)
- All we have to do, is thus to find a threshold u such that the max only exceed u alpha percent of the time.



Bonferroni Correction

Bonferroni correction allows to keep the FWER at 5% by simply dividing alpha by the number of tests – it find the threshold u

$$P(T_i \geq u|H0) \leq \frac{\propto}{m}$$ *Find u to keep the FWER < α/m*

$$\text{FWER} = P(\cup_{i \in V}\{T_i \geq u\}|H_0) \leq \propto$$

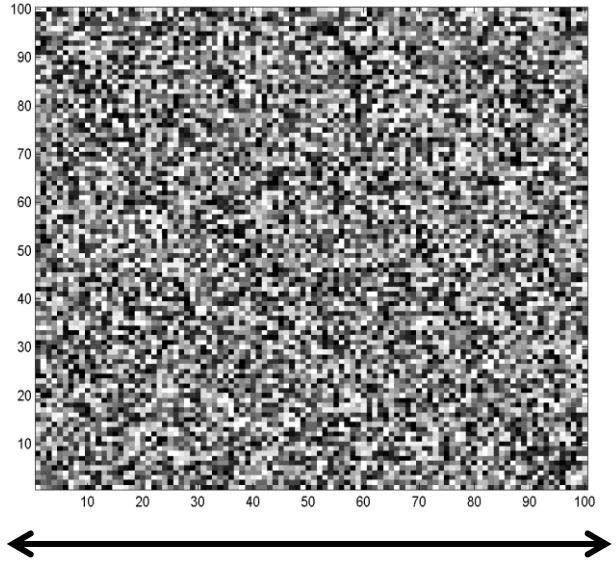
$$\leq \sum P(T_i \geq u|H0)$$ *Boole's inequality*

$$\leq \sum_i \frac{\propto}{m} = \propto$$

Bonferroni Correction

- 10000 independent Z-scores ;
- alpha corrected = .000005
- z-score = 4.42

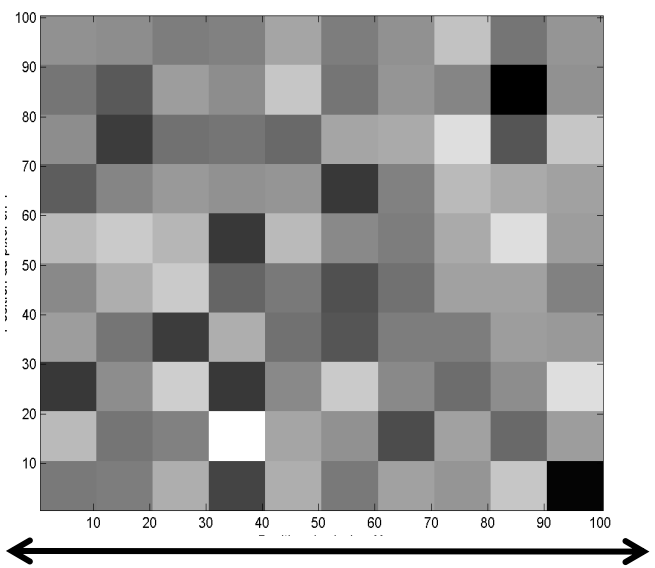
100 obs



100 obs

- 100 independent observations
- alpha corrected = .0005
- z-score = 3.29

100 obs



100 obs

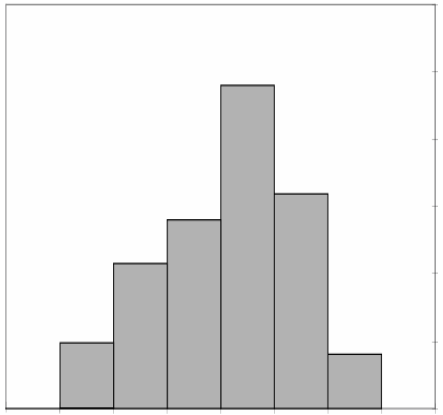
Finding the threshold u using resampling

- Estimate the distribution of max under H0 (bootstrap/permutation) and simply threshold the observed results with a threshold u like Bonferroni

- "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

given that we have no other information about the population, the sample is our best single estimate of the population



Sample

Population

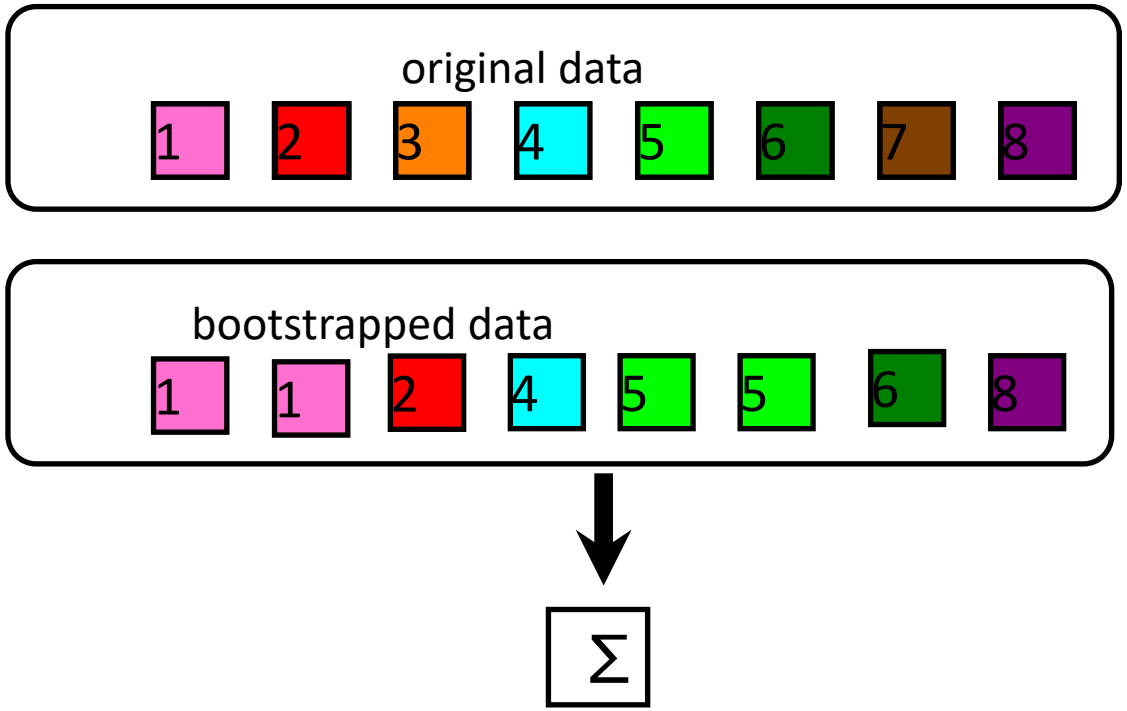
Bootstrap: central idea

- Parametric statistics rely on estimators (e.g. the mean) and measures of accuracy for those estimators (standard error and confidence intervals)
- The bootstrap is a type of resampling procedure along with jack-knife and permutations.

- “The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates.” Efron & Tibshirani, 1993
- Bootstrap is particularly effective at estimating accuracy (bias, SE, CI) but it can also be applied to many other problems – in particular to estimate distributions.

General recipe

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



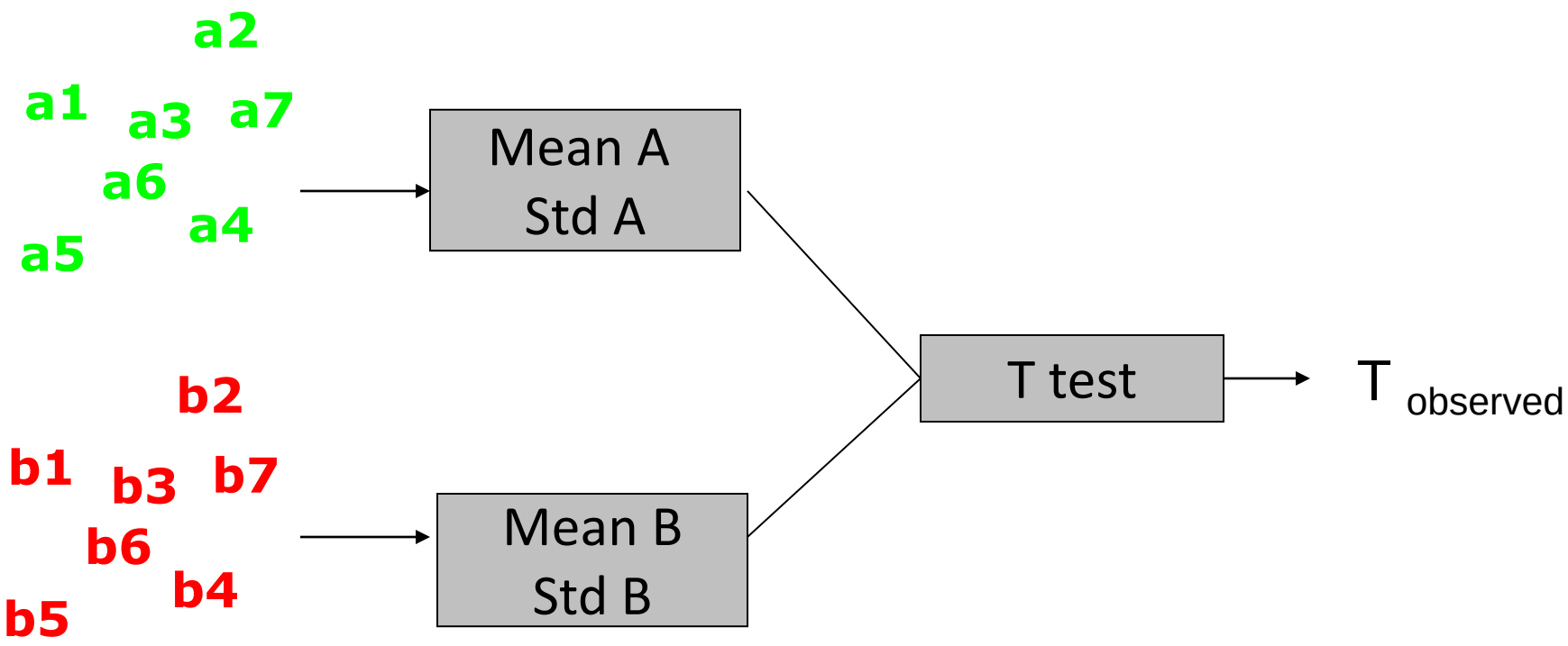
(2) compute estimate
e.g. sum, trimmed mean

(3) repeat (1) & (2) b times

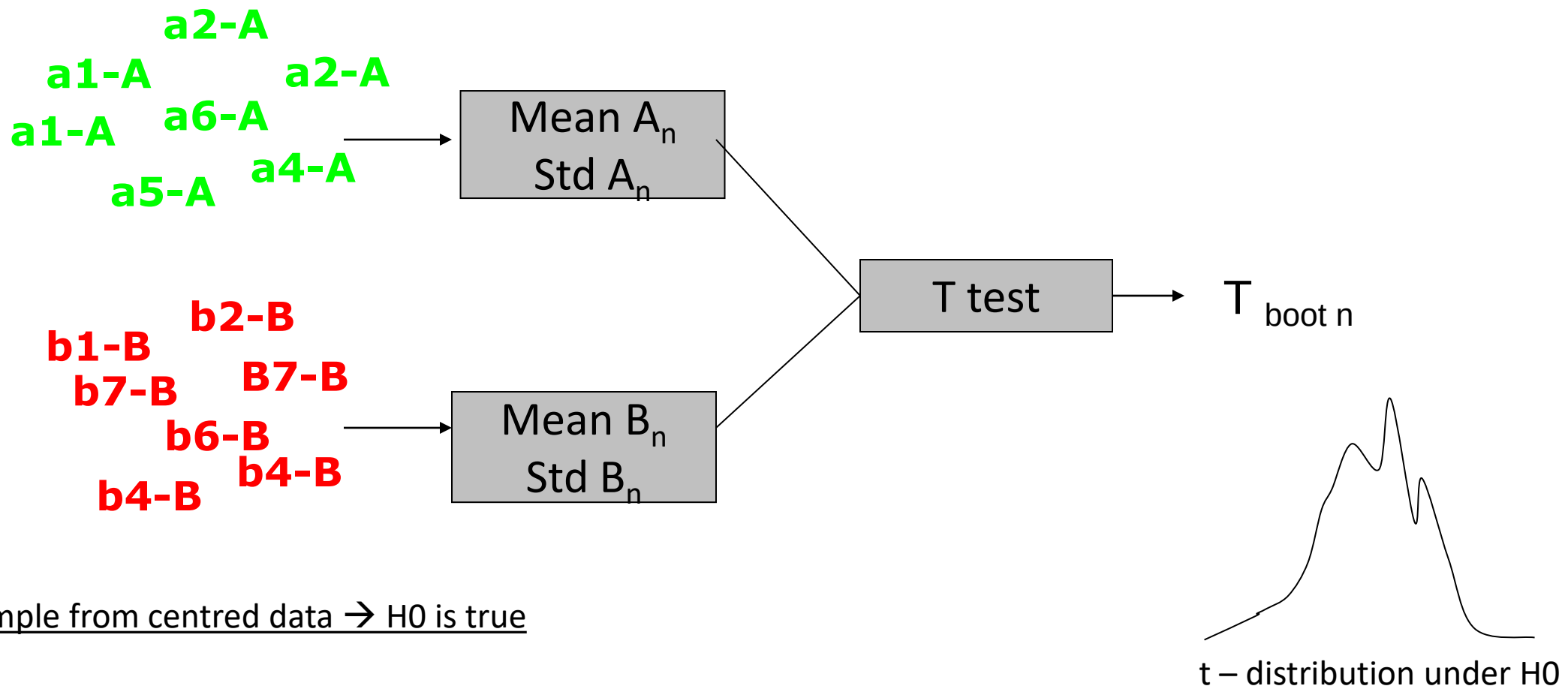
Σ_1 Σ_2 Σ_3 Σ_4 Σ_5 Σ_6 . . . Σ_b

(4) get bias, std, confidence interval, p-value

Application to a 2 samples t-test: Bootstrap under H0



Application to a 2 samples t-test: Bootstrap under H0

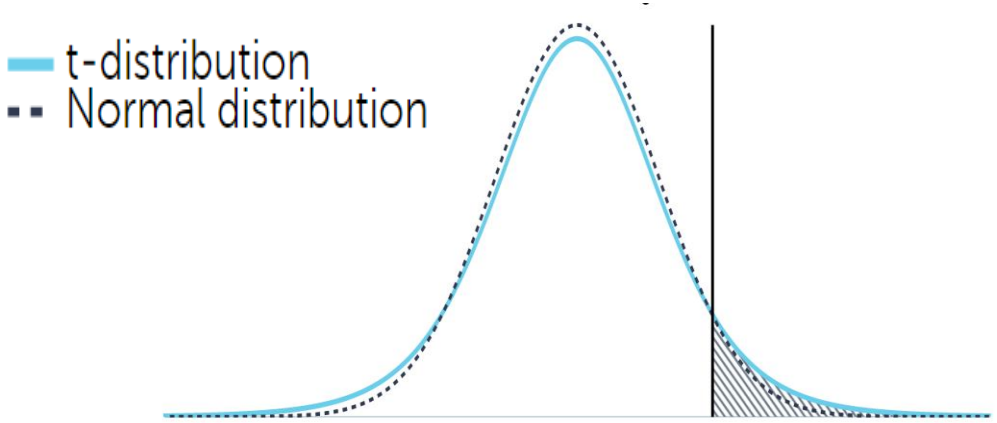


Resample from centred data → H0 is true

Application to a 2 samples t-test: Bootstrap under H0

What is the p value of the sample

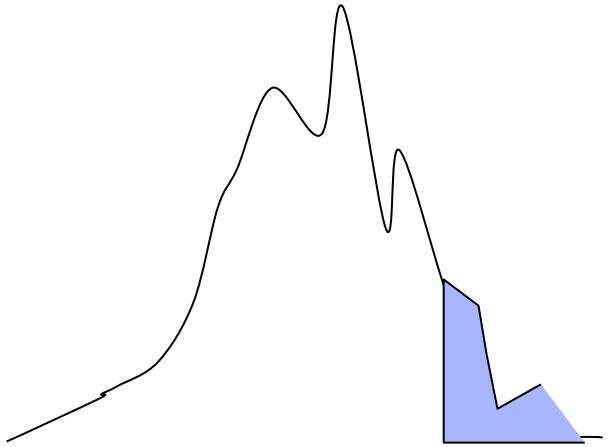
p(Obs≥t|H0) → cumulative probability



area under the curve for T obs = p value
Significance = point of T critical

What is the p value of the sample

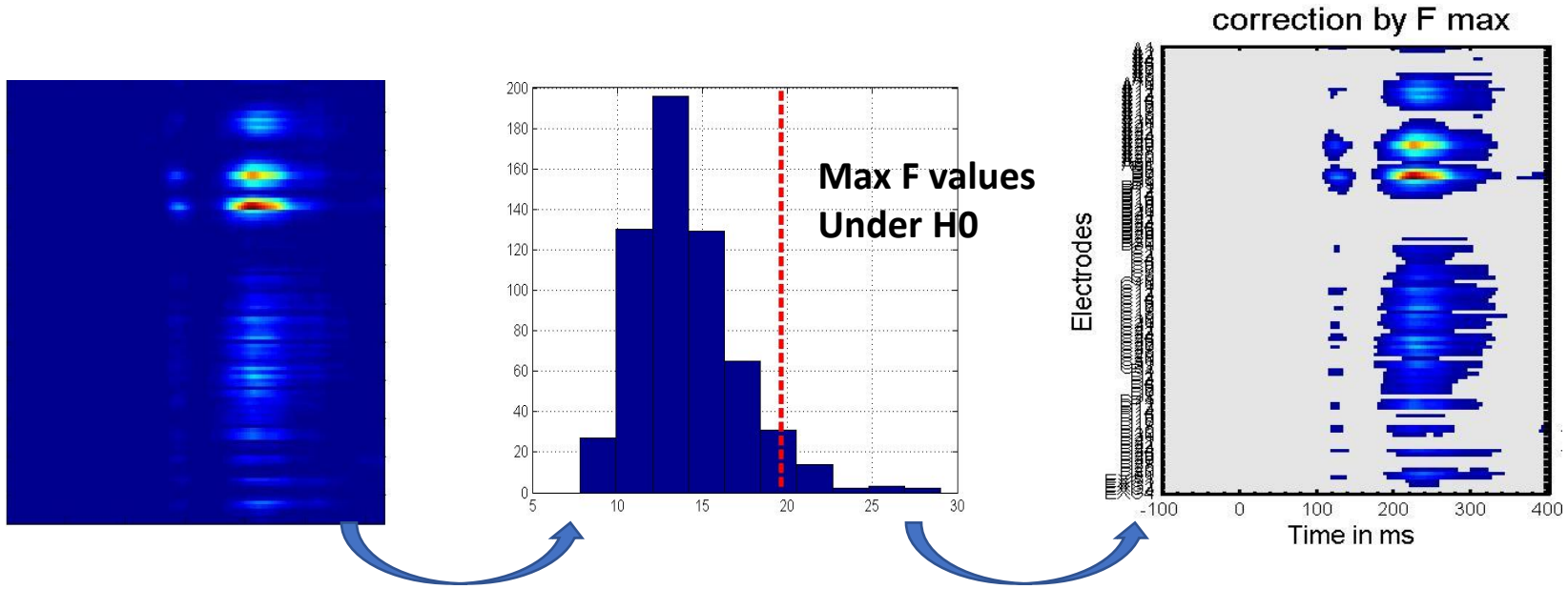
p(Obs≥t|H0) → cumulative probability



area under the curve for T obs = p value
Significance = percentile of the empirical t distribution
→ Theoretical T assumes data normality, we don't

Maximum Statistics based on resampling

- Do many whole brain analyses (t/F tests), and record the maximum value
- After N boot, we have the distribution of maxima under H0 and thus u
- Accounts inherently for smoothness but still assumes all tests are independent



Cluster inference

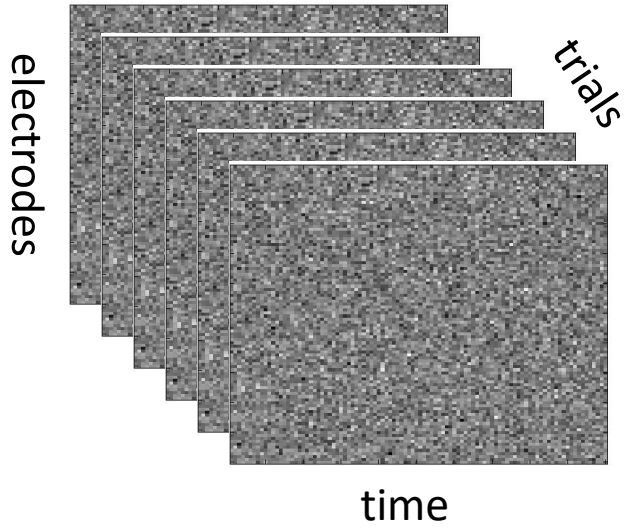
Solutions for imaging data

- An important feature of neuroimaging data is that we have a family of stat values that has topological features (Bonferroni for instance consider tests as independent) and we can thus considering data as a smooth lattice, i.e. based our inference on clusters
- fMRI/PET are projection methods of data points onto the whole space – MEEG forms continuous functions in time and are smooth by the scalp (space)
- Neural activity propagate locally through intrinsic/lateral connections and is distributed via extrinsic connections / Hemodynamic correlates are initiated by diffusing signals (e.g. NO)

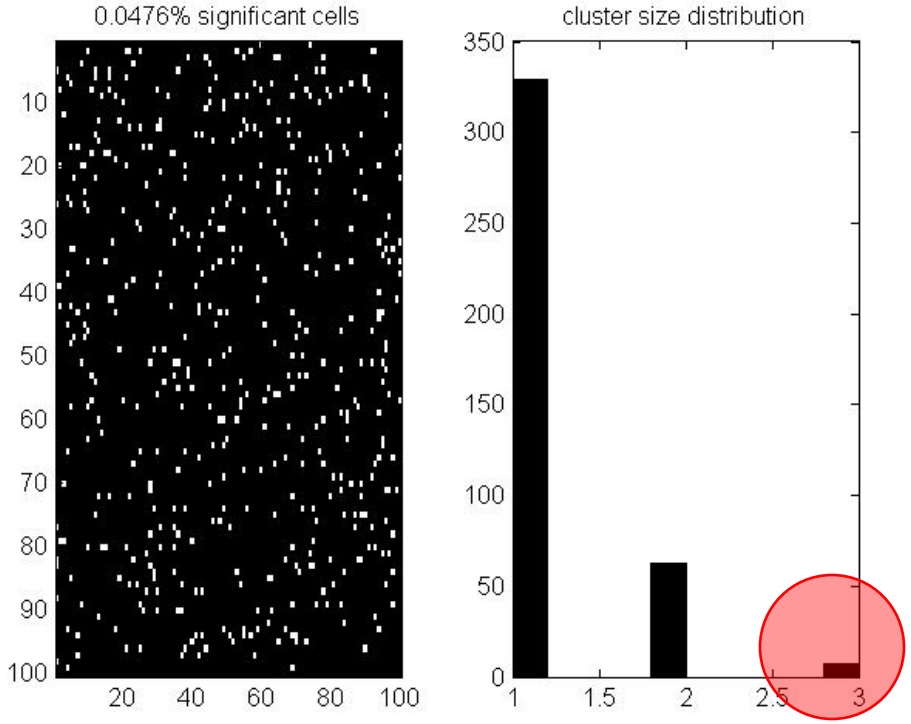
Chumbley et al., 2009 NeuroImage 44

Let’s analyse clusters

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

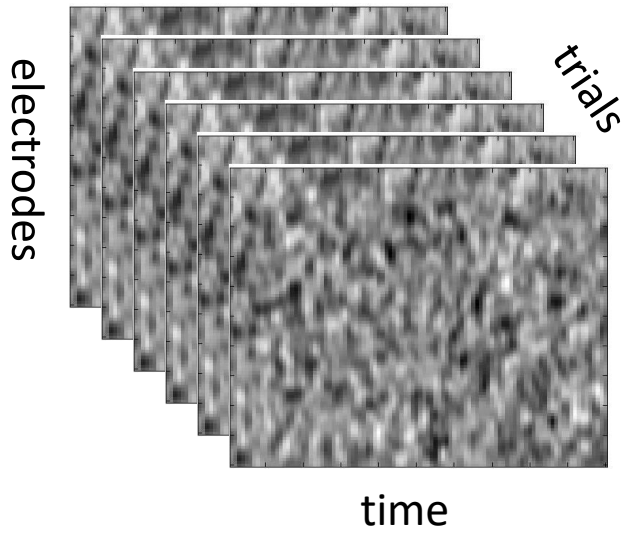


One sample t test > 0 ?

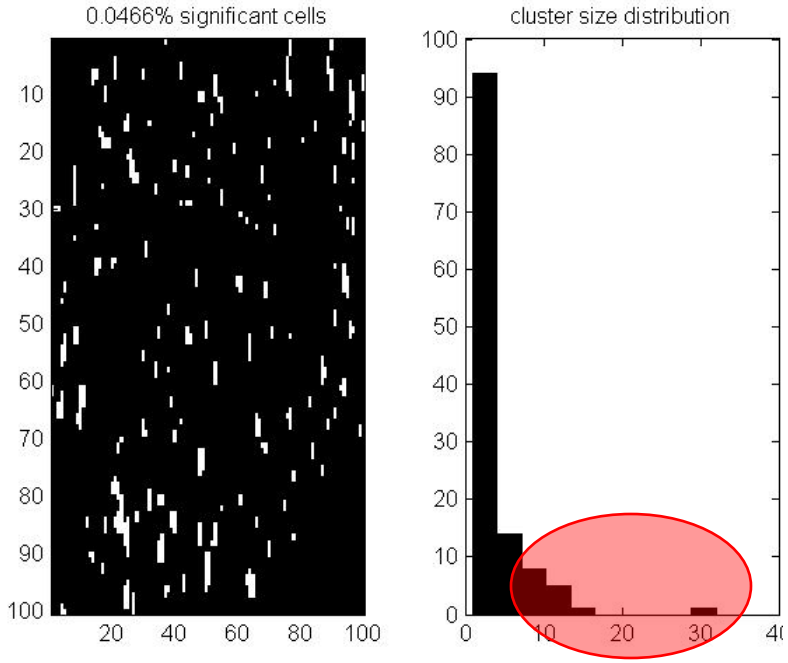


Let’s analyse clusters

- 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!



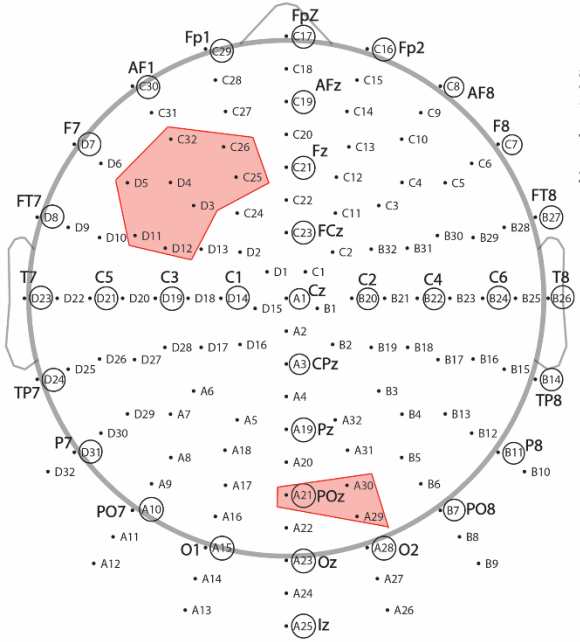
One sample t test > 0 ?

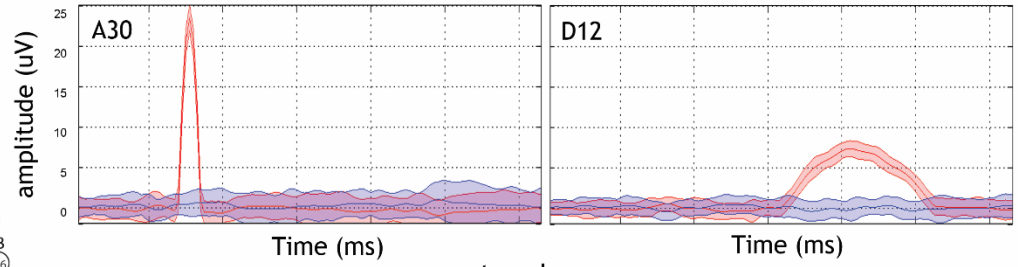


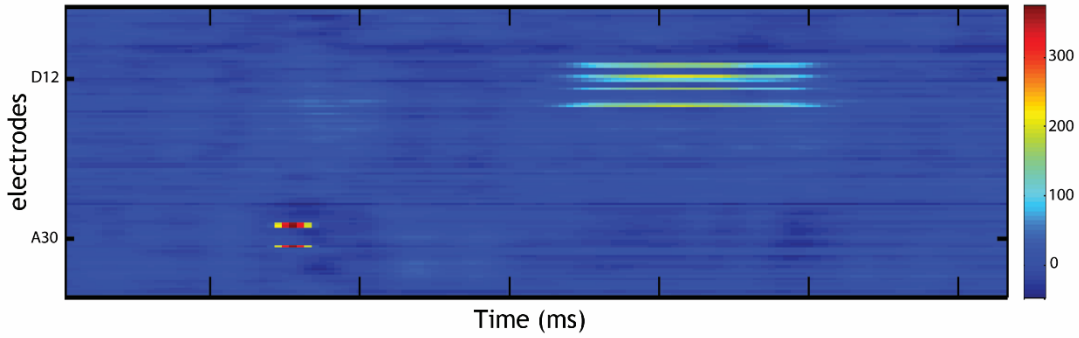
The clustering solution

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

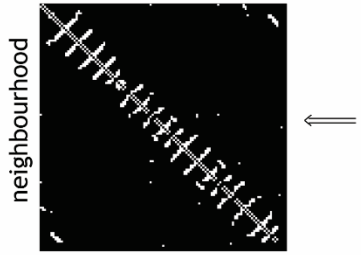
- To use clustering, we need to consider cluster statistics rather than individual statistics
- Cluster statistics depend on (i) the cluster size, which depends on the data at hand (how correlated data are in space and in time/frequency), and (ii) the strength of the signal (how strong are the t, F values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





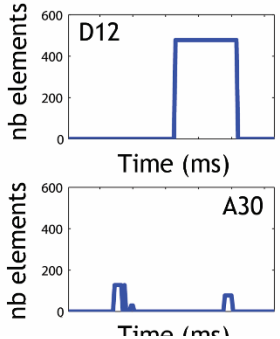


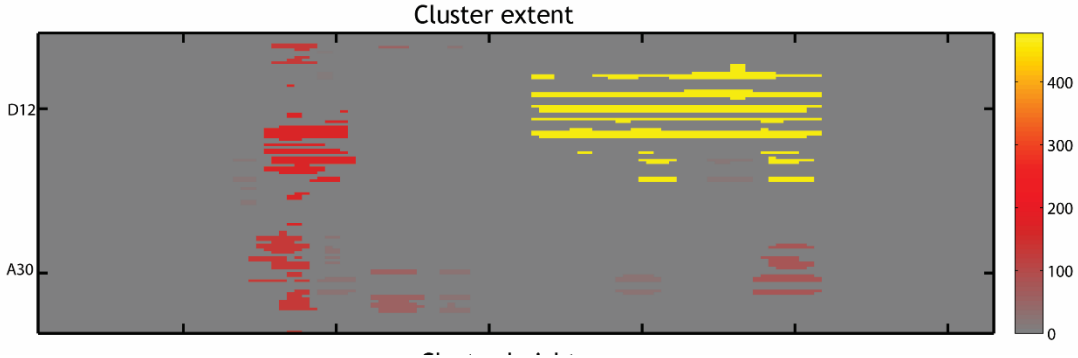
Spatial - Temporal clustering

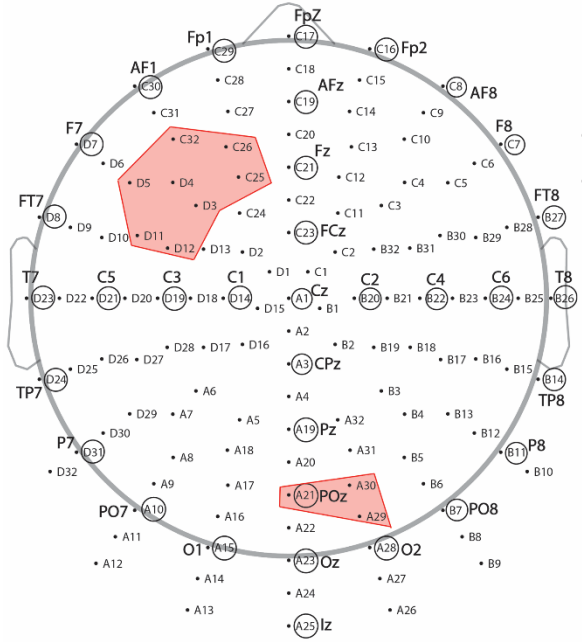


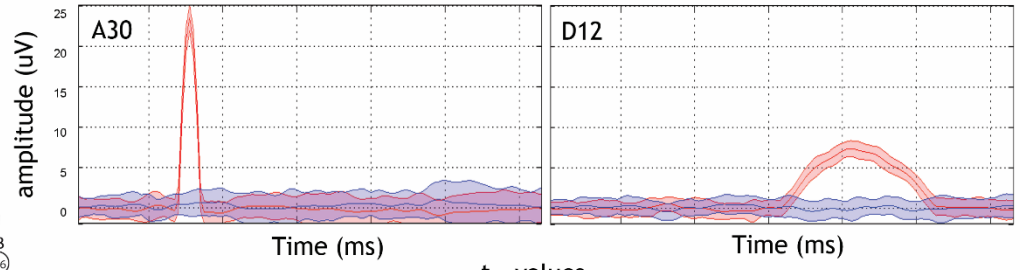
maximum extent
= number of
electrodes and
time points

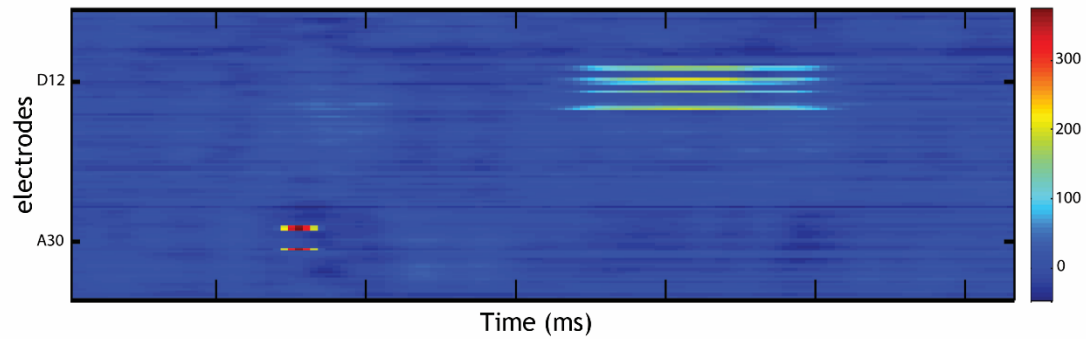
cluster 1 = 478
cluster 2 = 127



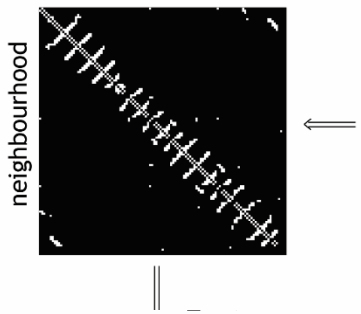


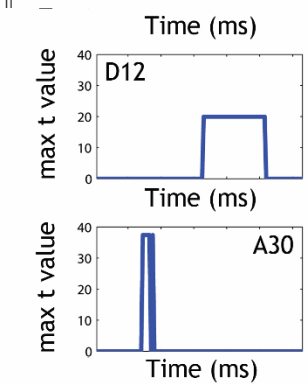






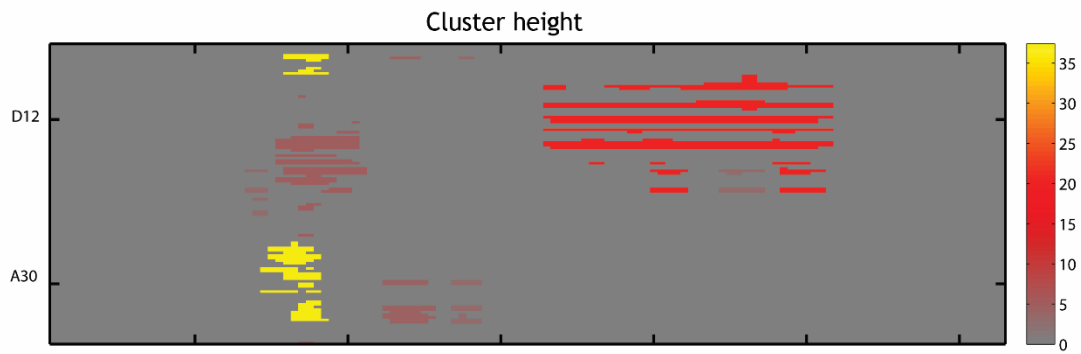
Spatial - Temporal clustering

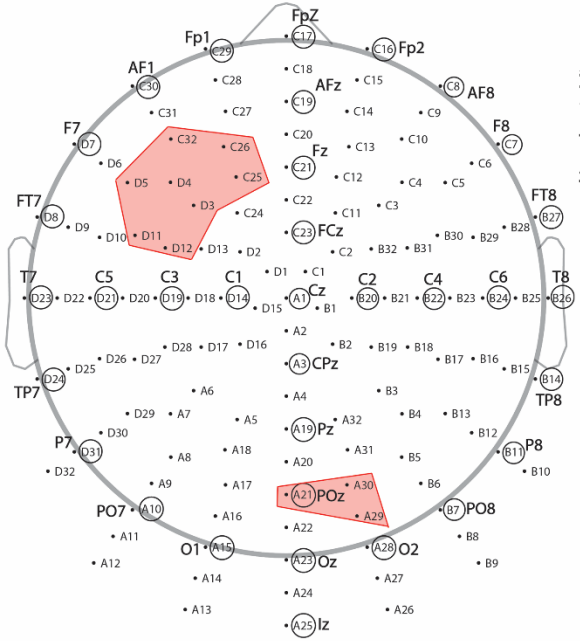


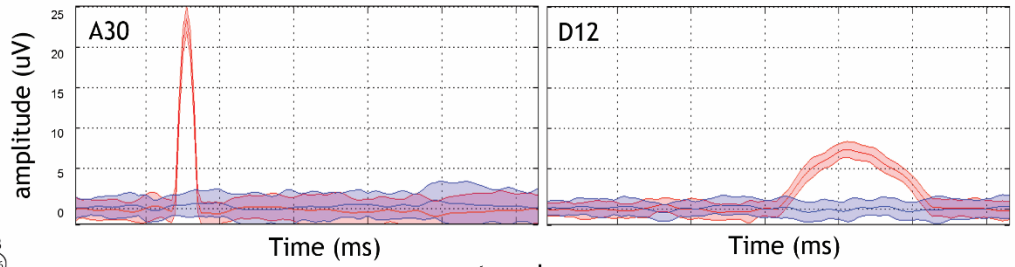


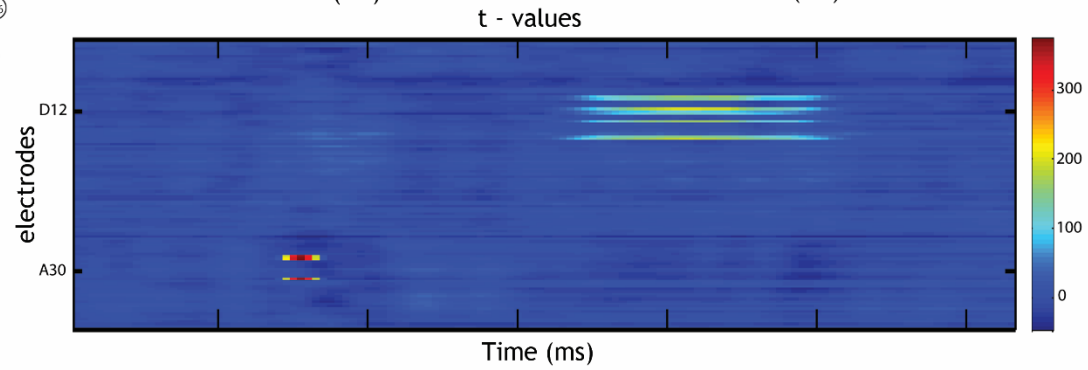
maximum height
within a cluster
of electrodes
and time points

cluster 1 = 19.7
cluster 2 = 37.4

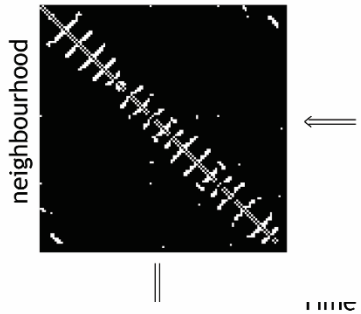






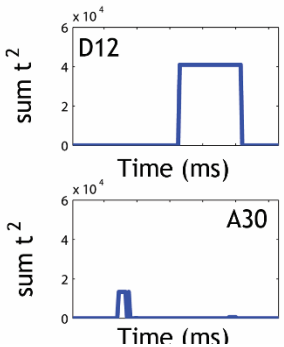


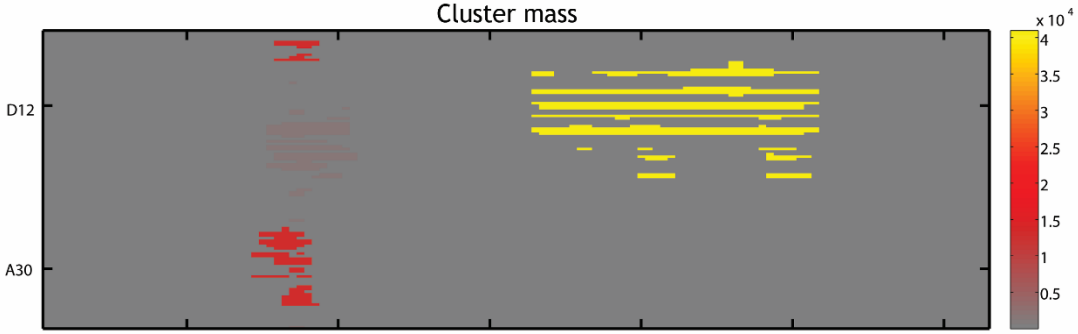
Spatial - Temporal clustering



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

cluster 1 = 40984
cluster 2 = 13386

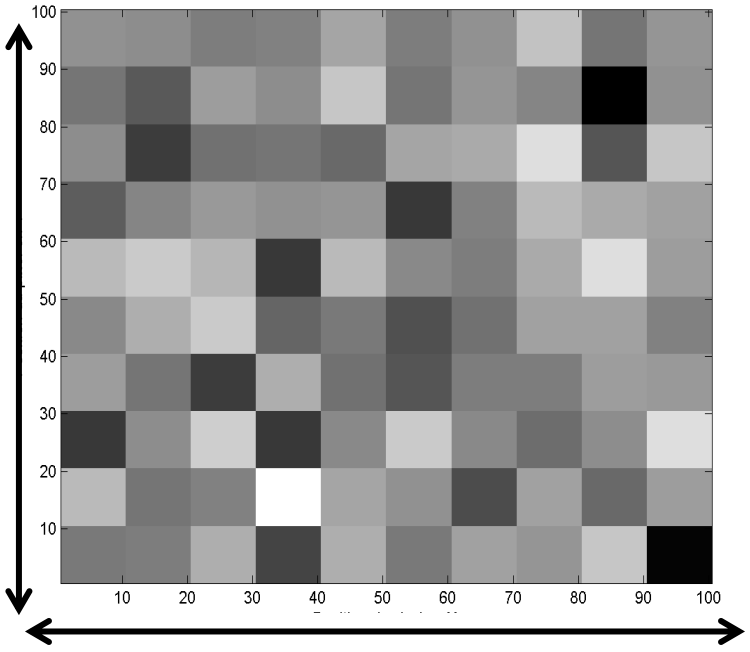


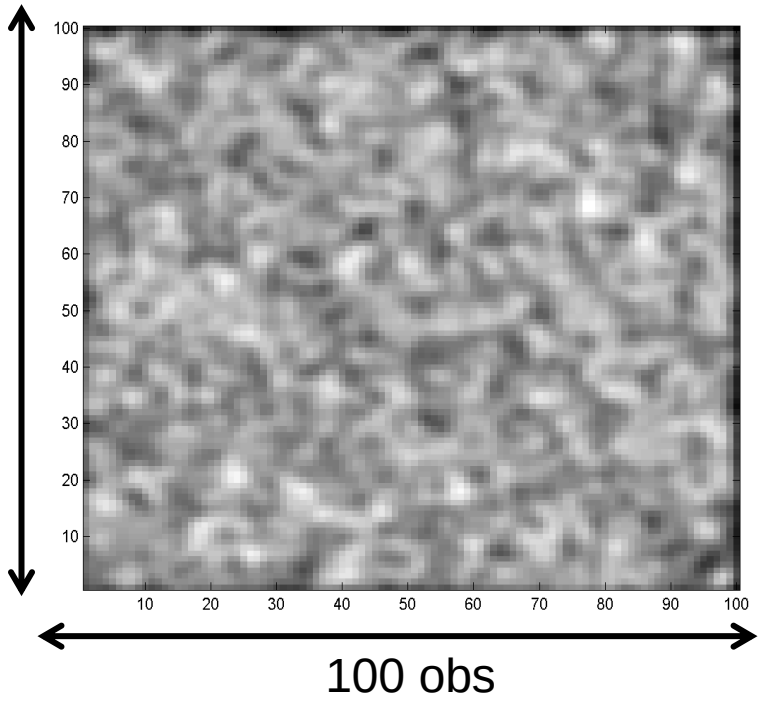


Random Field Theory

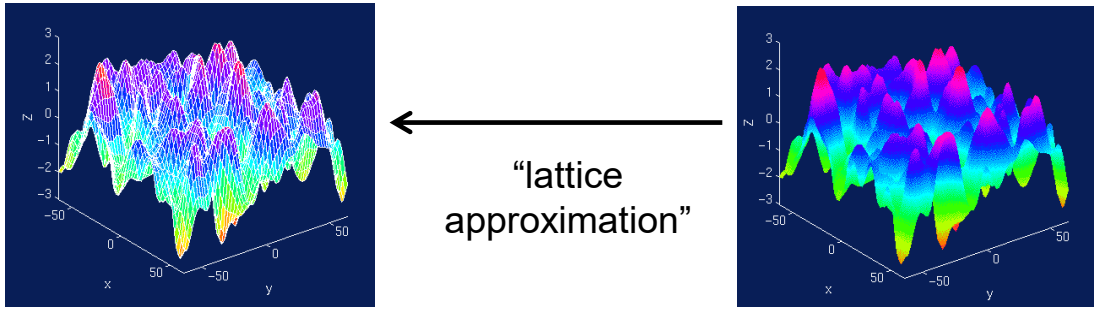
- 10000 Z-scores ; alpha = 5%
- How many independent observations ?

100 obs





Random Field Theory

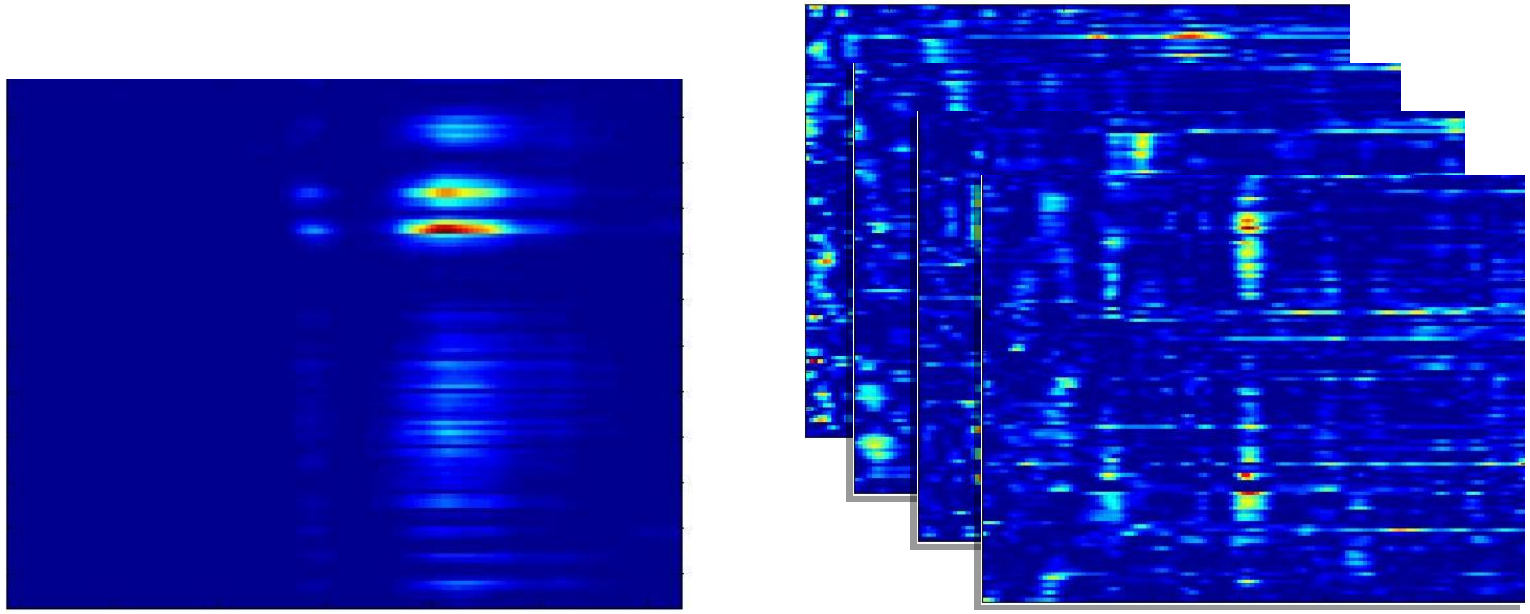


- RFT relies on theoretical results for smooth statistical maps (hence the need for smoothing), allowing to find a threshold in a set of data where it's not easy to find the number of independent variables. Uses the expected Euler characteristic (EC density)

- 1 Estimation of the smoothness = number of resel (resolution element) = f(nb voxels, FWHM)
- 2 expected Euler characteristic = number of clusters above the threshold
- 3 Calculation of the threshold(s) – set(cluster(obs))

Computing cluster distributions

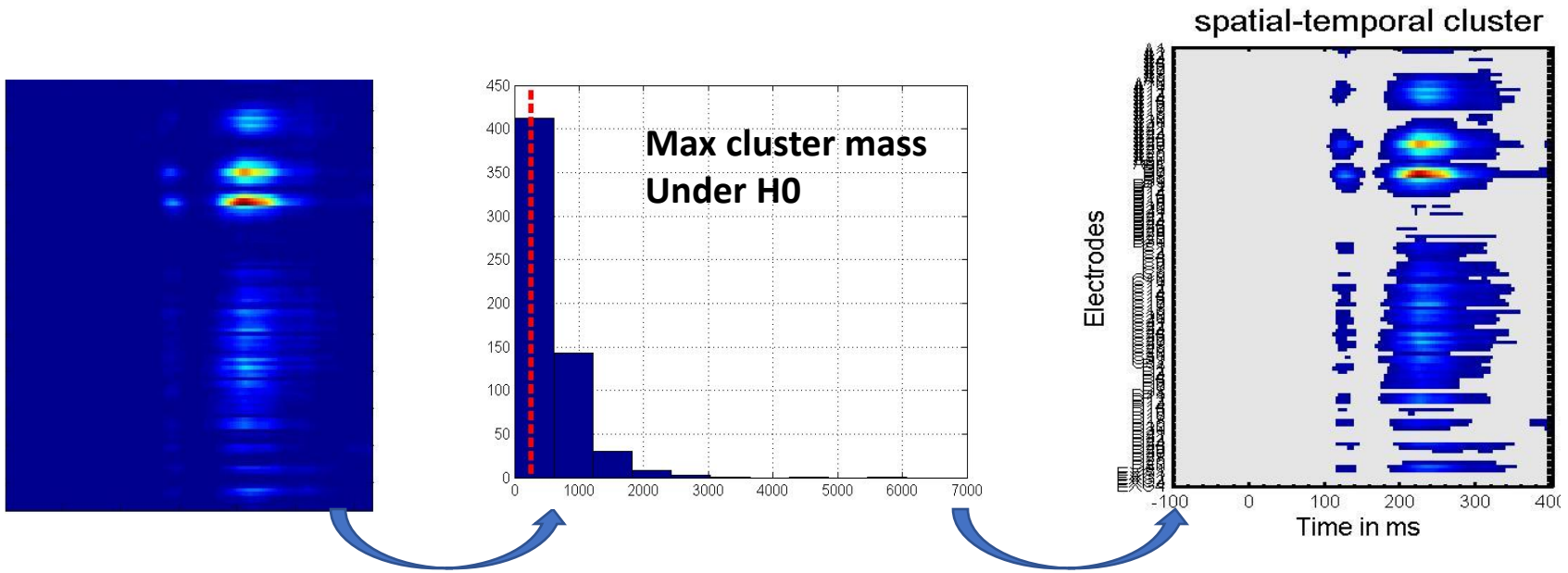
- In LIMO EEG, we bootstrap the data under H0: center the data or break the link between the design matrix and the data and then resample and test. This way we can find u for a single bin, the the whole space, or for clusters.



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

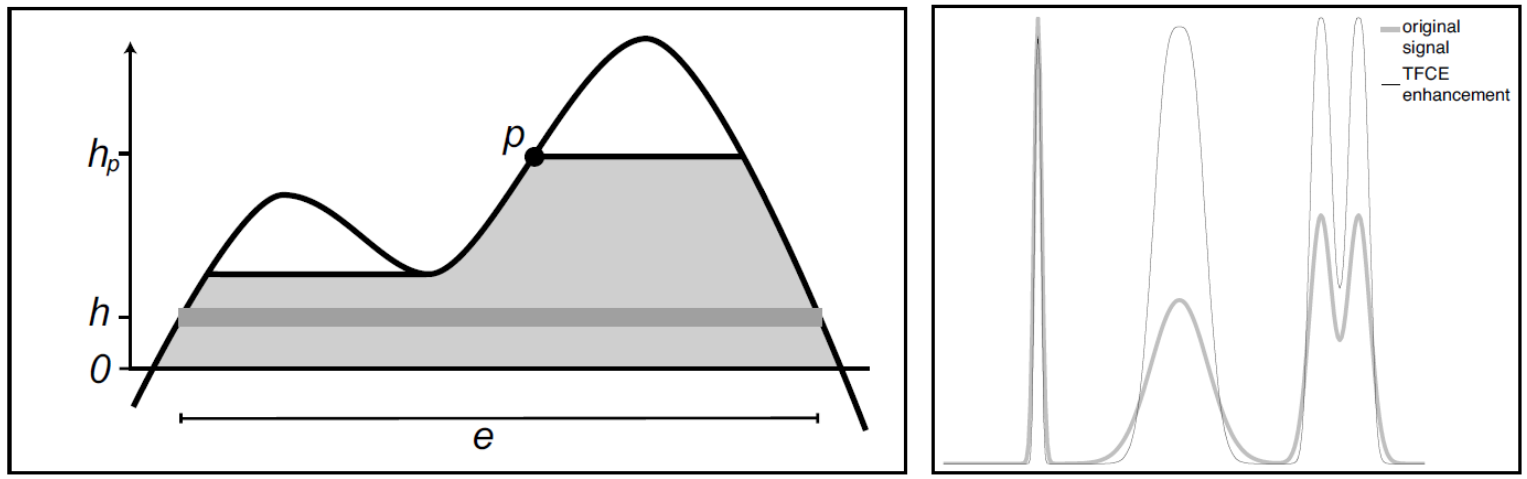
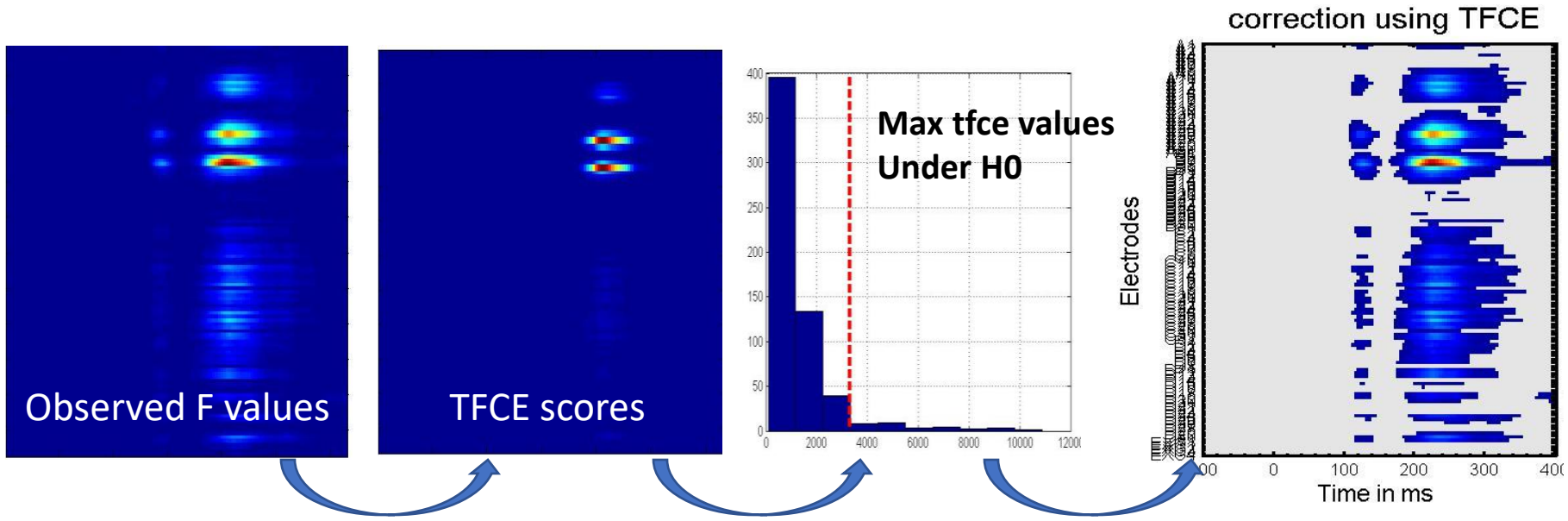


Figure 1: Illustration of the TFCE approach. Left: The TFCE score at voxel p is given by the sum of the scores of all incremental supporting sections (one such is shown as the dark grey band) within the area of "support" of p (light grey). The score for each section is a simple function of its height 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

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
- Asymptotically = RFT

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., Chauveau, N., Gaspar, C. & Rousselet, G (2011).** Linear Modelling of MEEG. Comp. Intel. Neurosc. Article ID 831409
- **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