Multiple comparisons problem and solutions

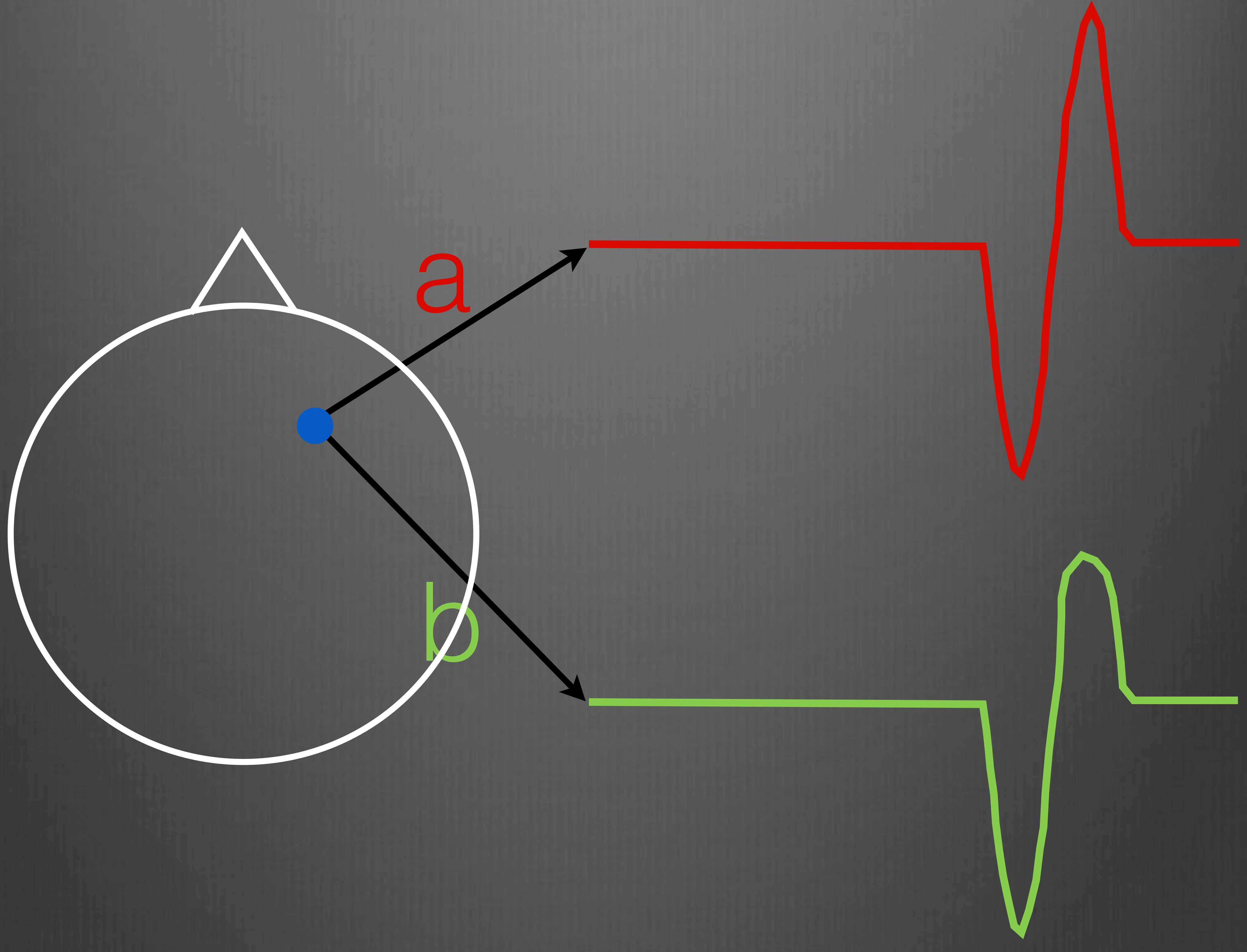
James M. Kilner

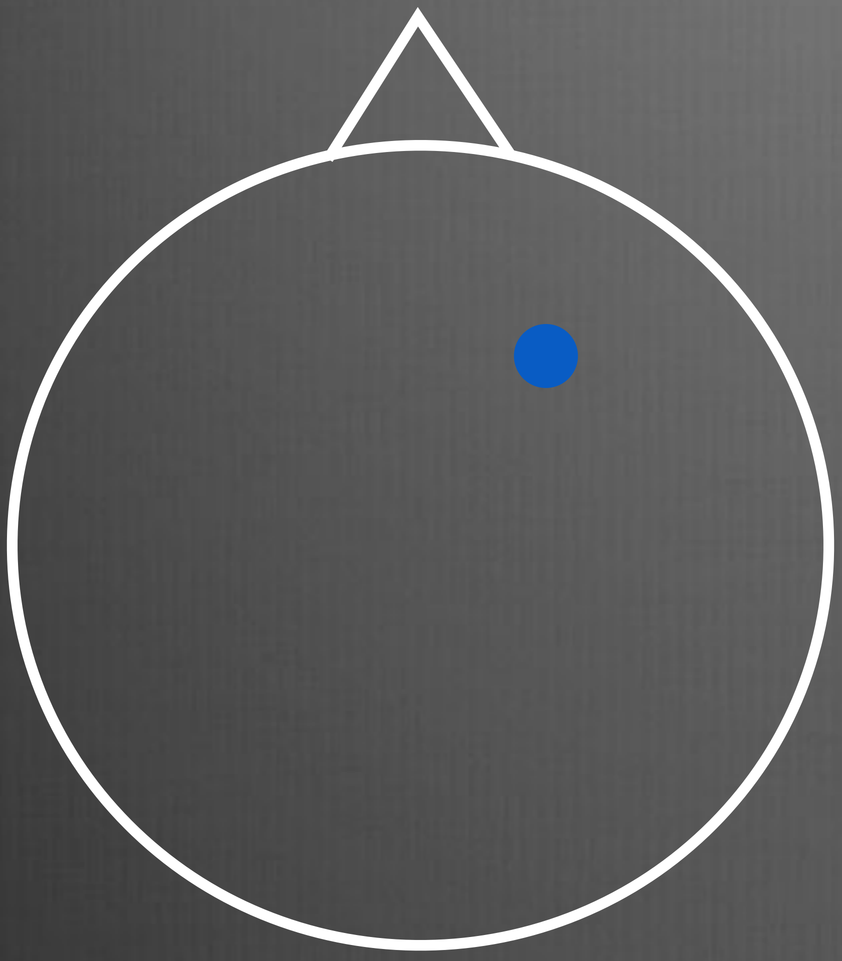
http://sites.google.com/site/kilnerlab/home

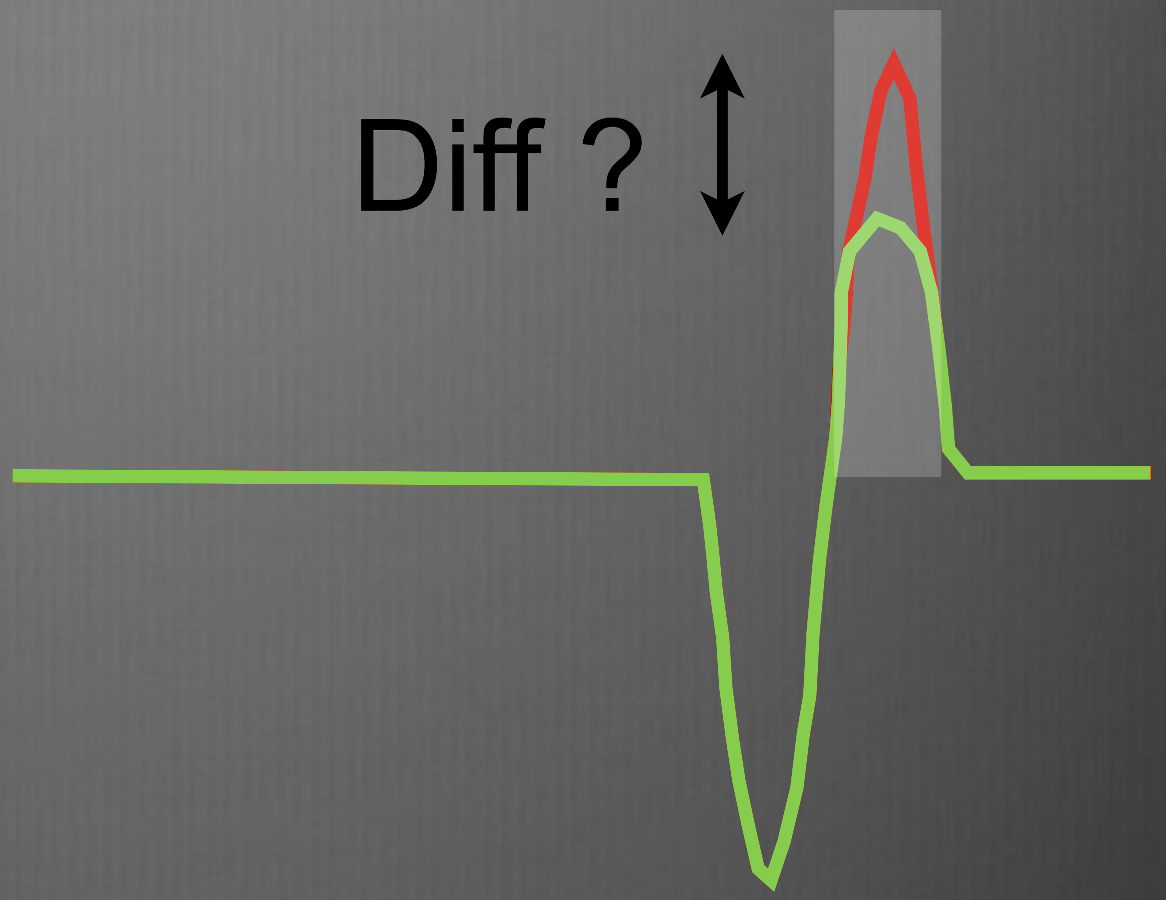
What is the multiple comparisons problem

How can it be avoided

Ways to correct for the multiple comparisons problem

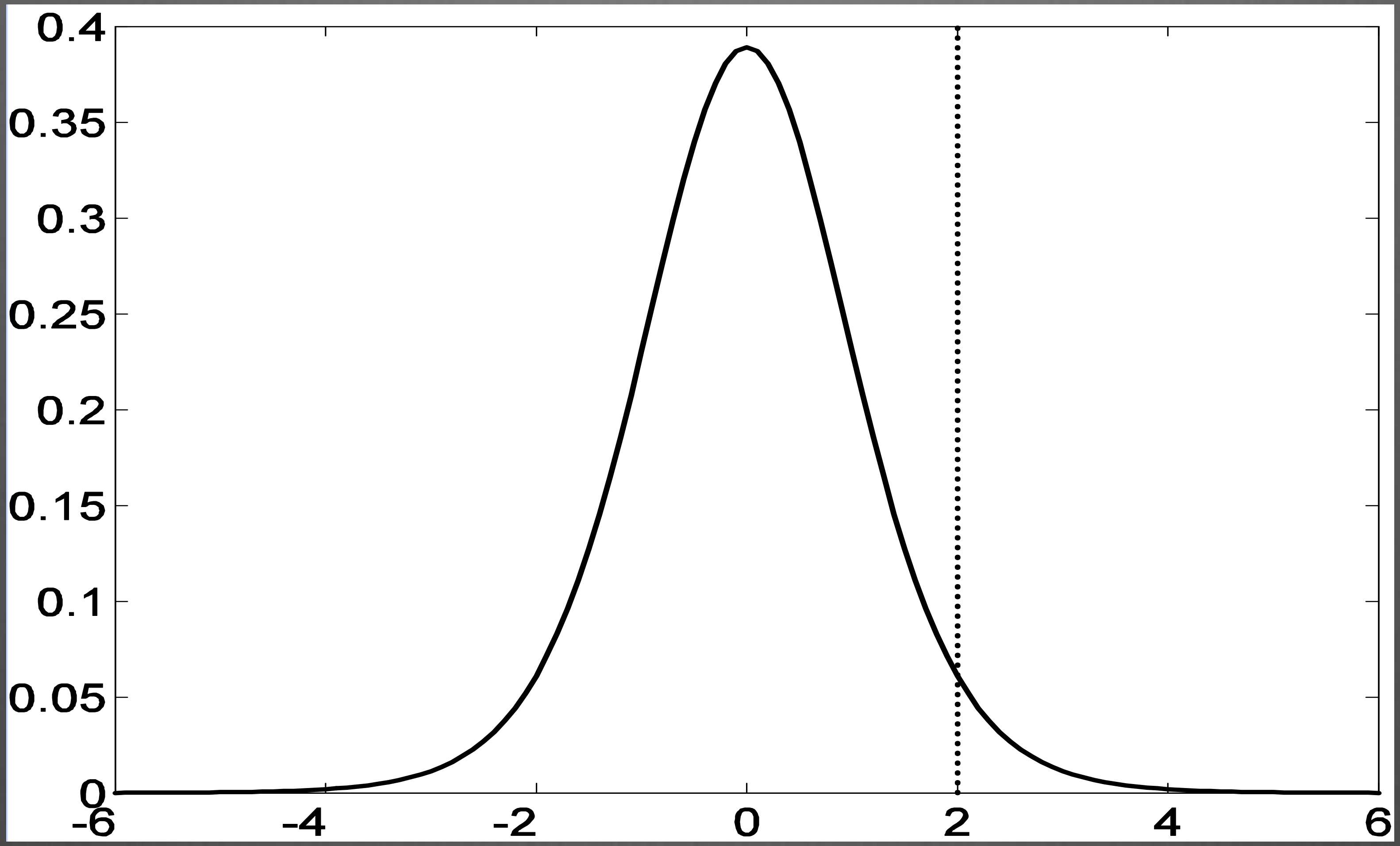

a
b





a - b > 0

t-distribution



NULL hypothesis, H: activation is zero

$$\alpha = p(t>u|H)$$

u=(effect size)/std(effect size)

Problems

MEG and EEG does not have zero dimensionality

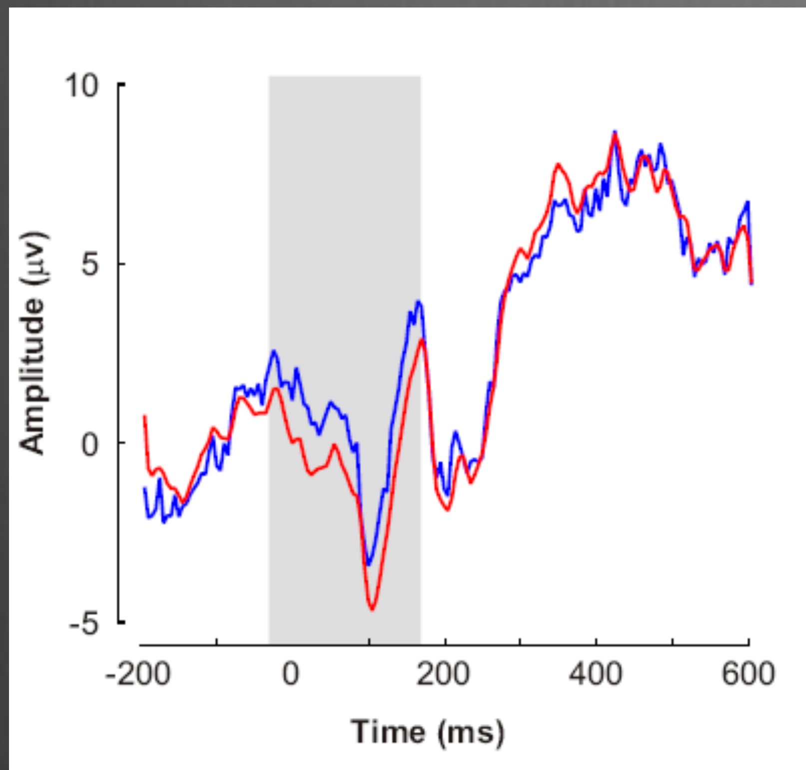
If one electrode data is at least one-dimensional

If multiple electrodes data is at least three dimensions

If time-frequency analysis then data can be four dimensional

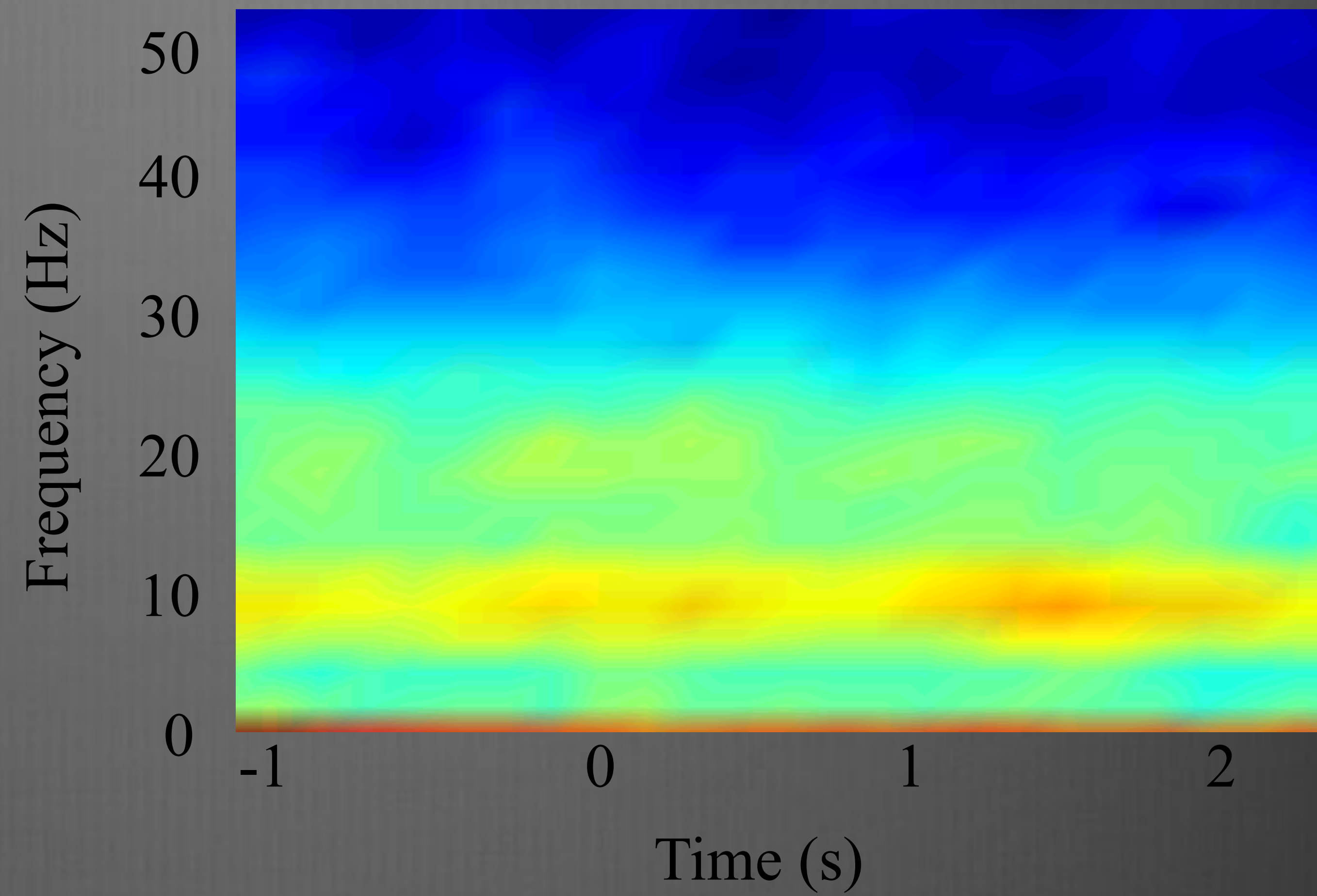
Massive multiple comparison problem

1-D



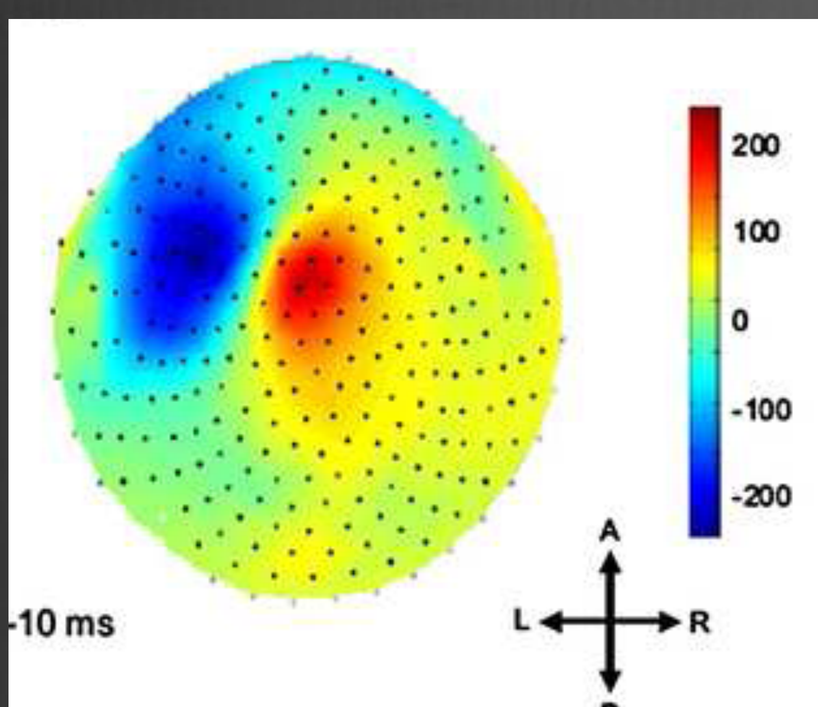
10
5
0
-5
Amplitude (µv)
-200
0
200
400
600
Time (ms)

2-D



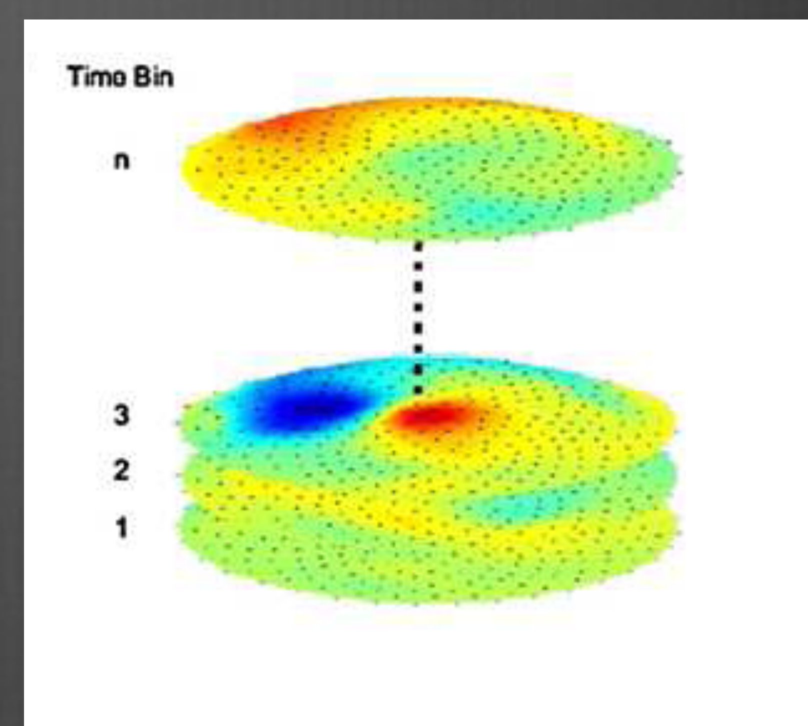
50
40
30
20
10
0
Frequency (Hz)
-1
0
1
2
Time (s)

2-D

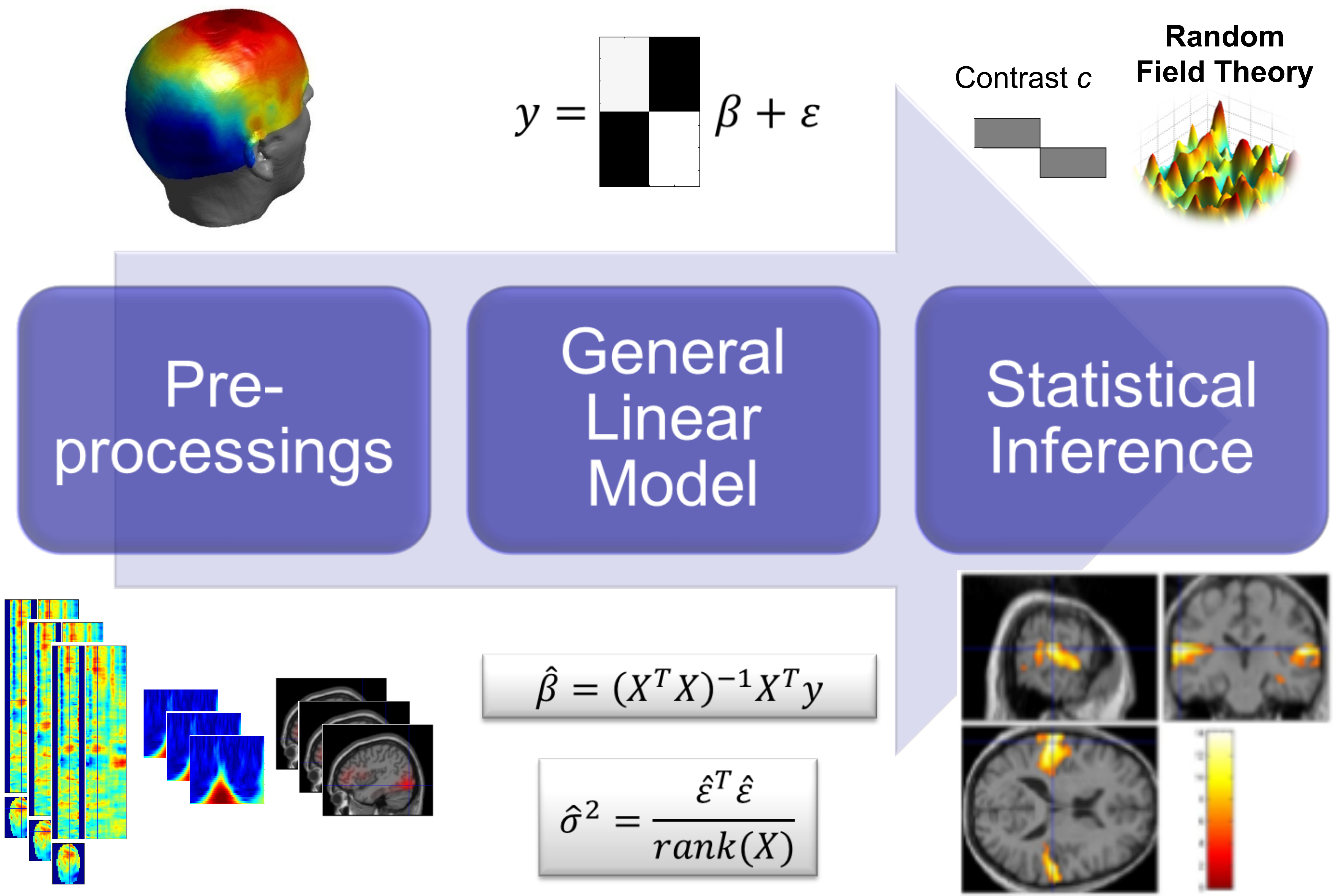


200
100
0
-100
-200
-10 ms
A
L
R
P

3-D



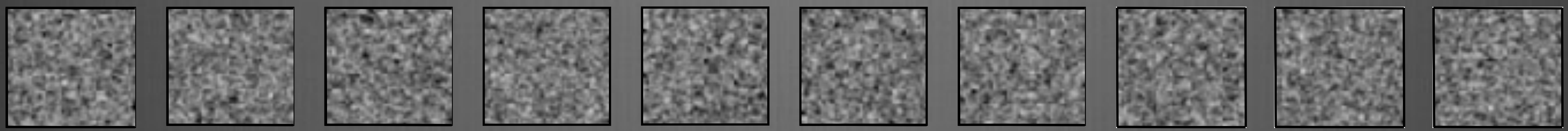
Time Bin
n
3
2
1



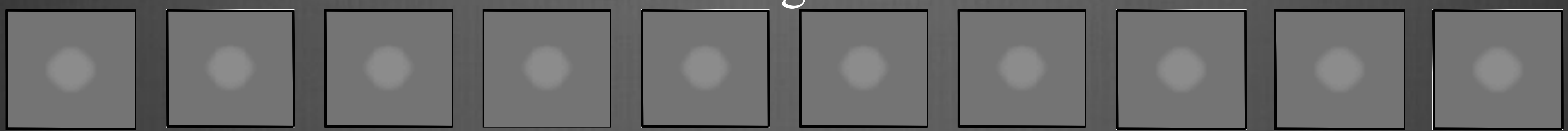
Random Field Theory
Contrast c
$y = X\beta + \varepsilon$
Pre-processings
General Linear Model
Statistical Inference
$\hat{\beta} = (X^T X)^{-1} X^T y$
$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{rank(X)}$

MCP example

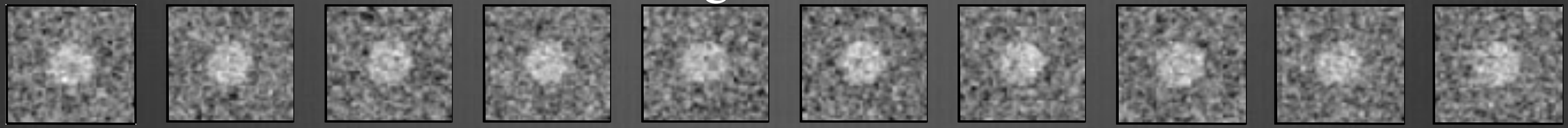
Noise



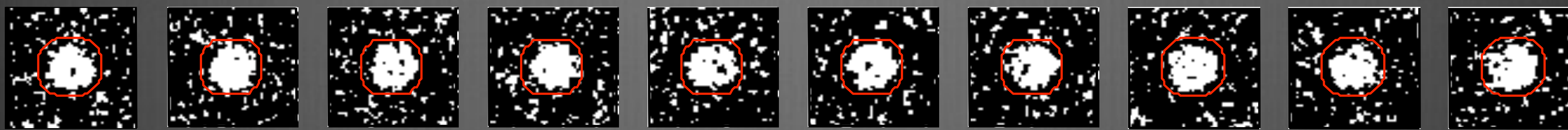
Signal



Signal+Noise



Use of ‘uncorrected’ p-value, α=0.1



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

Using an ‘uncorrected’ p-value of 0.1 will lead us to conclude on average that 10% of voxels are active when they are not.

Common Solutions

One solution is to reduce the multi-dimensional data to zero-dimensional data by averaging over a window of interest

This must be specified a priori or derived from an independent contrast.

One can not base this window on where the effect is largest!

‘BUT the basic question remains - why would one do all this, and search for some odd effects this way, when it is all visible in the sensor level’

u=(effect size)/std(effect size)

Other Solutions

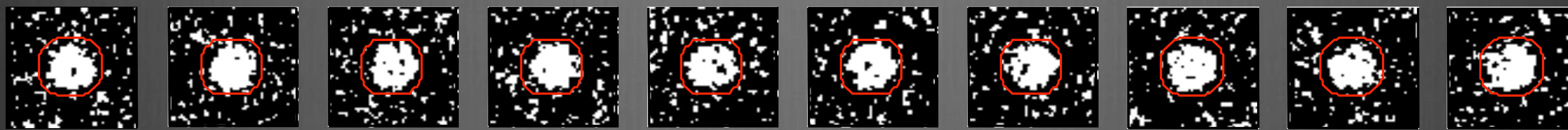
FAMILY-WISE NULL HYPOTHESIS:
Activation is zero everywhere

If we reject a voxel null hypothesis
at *any* voxel, we reject the family-wise
Null hypothesis

A FP **anywhere** in the image
gives a Family Wise Error (FWE)

Family-Wise Error (FWE) rate = ‘corrected’ p-value

Use of 'uncorrected' p-value, $\alpha=0.1$



Use of 'corrected' p-value, $\alpha=0.1$

FWE

Bonferroni correction

The Family-Wise Error rate (FWE), α, for a family of N **independent** voxels is

$$\alpha = Nv$$

where v is the voxel-wise error rate. Therefore, to ensure a particular FWE set

$$v = \alpha / N$$

However, the data points in M/EEG data are not independent
They are correlated either temporally, spatially or in frequency space

Random Field Theory

Statistical parametric maps (e.g., t-maps) are fields with values that are, under the null hypothesis, distributed according to a known probability distribution.

RFT is used to resolve the multiple comparisons problem that occurs when making inferences over the search-space:

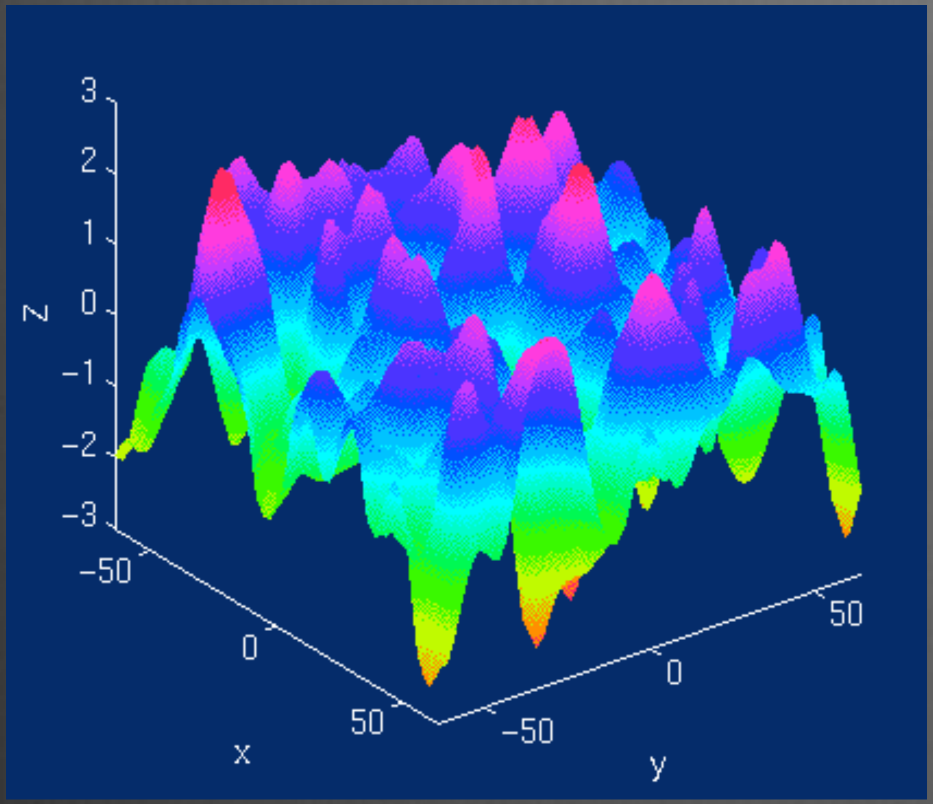
Adjusted p-values are obtained by using results for the expected Euler characteristic.

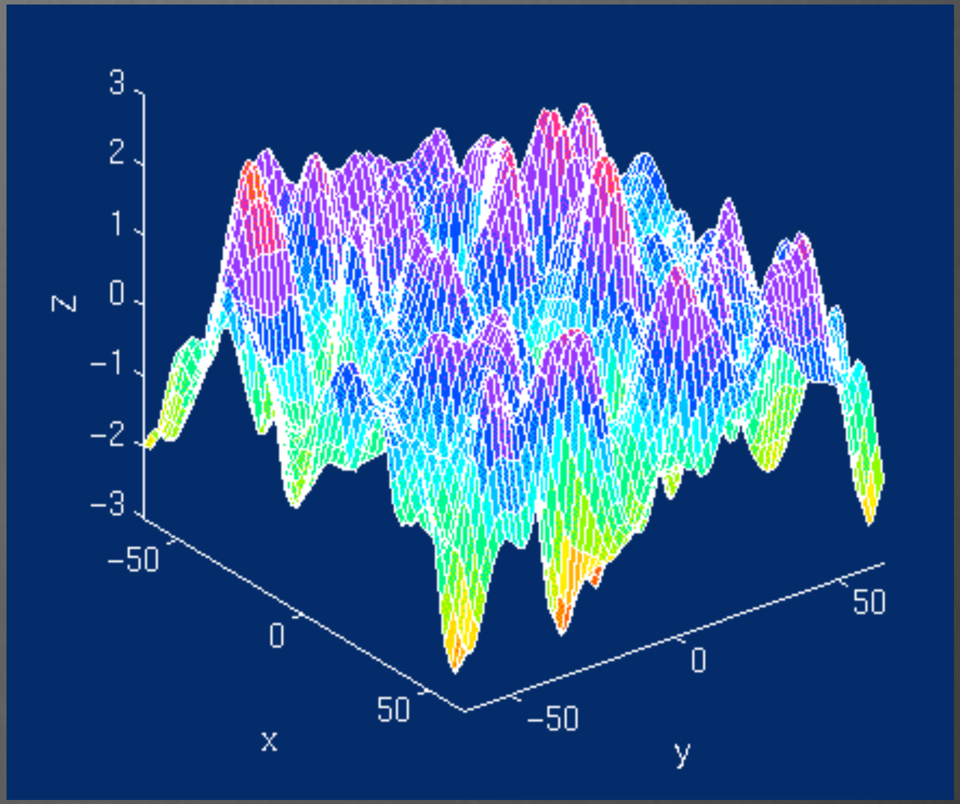
At very high thresholds the Euler characteristic reduces to the number of suprathreshold peaks and the expected EC becomes the probability of getting a peak above threshold by chance.

The expected EC therefore approximates the probability that the SPM exceeds some height by chance.

The ensuing p-values can be used to find a corrected height threshold or assign a corrected p-value to any observed peak in the SPM.

Good lattice approximation?





Will be true for high density recordings

Euler Characteristic χ

Euler Characteristic χ_u:

- Topological measure

  χ_u = # blobs - # holes

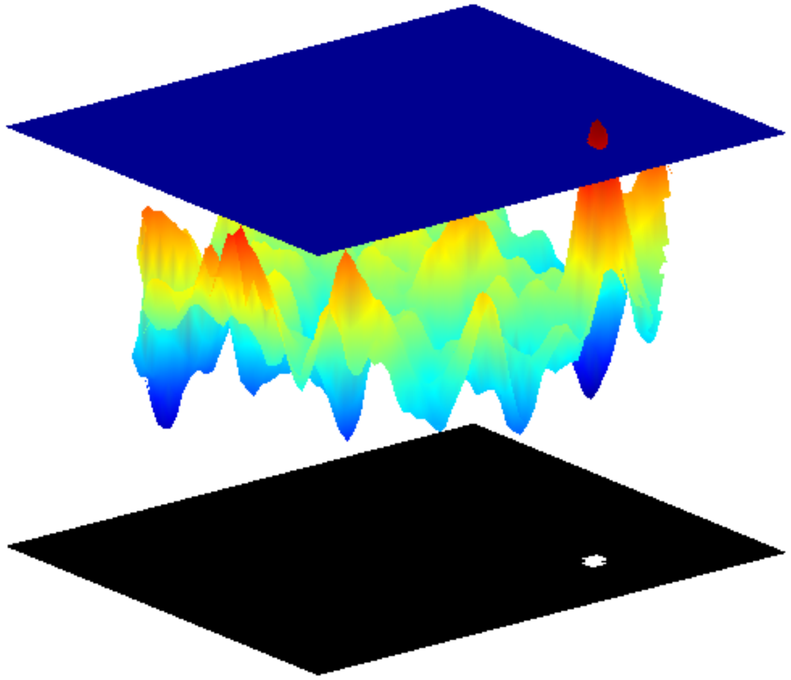
- at high threshold u:

  χ_u = # blobs

$$
\begin{aligned}
FWER &= p(FWE) \\
&= p\left(\bigcup_i \{T_i \geq u\} \middle| H_0\right) \\
&= p\left(\max_i T_i \geq u \middle| H_0\right) \\
&= p(one\ or\ more\ blobs\ |H_0) \\
&\approx p(\chi_u \geq 1|H_0) \\
&\approx E[\chi_u|H_0] \approx \boldsymbol{\alpha_{FWE}}
\end{aligned}
$$

No holes

Zero or one blob



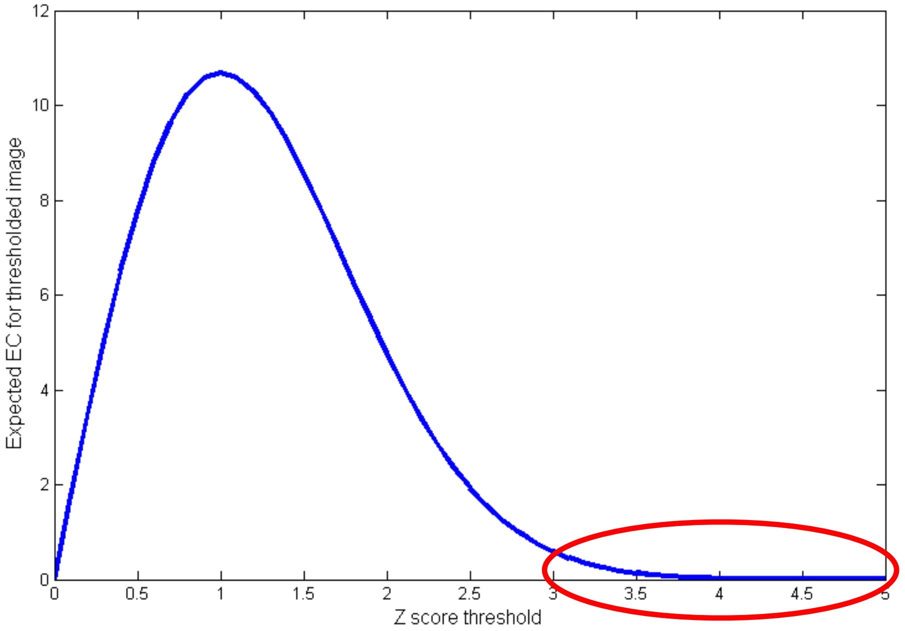
Expected Euler Characteristic

$$E[\chi_u] = \lambda(\Omega)|\Lambda|^{1/2}\, u \exp(-u^2/2)/(2\pi)^{3/2}$$

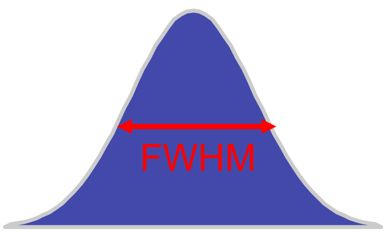
2D Gaussian Random Field

- Ω : search region
- $\lambda(\Omega)$: volume
- $|\Lambda|^{1/2}$: roughness (1 / smoothness)

100 x 100 Gaussian Random Field
with FWHM=10 smoothing
$\alpha_{FWE} = 0.05 \Rightarrow u_{RFT} = 3.8$
$(u_{BONF} = 4.42, u_{uncorr} = 1.64)$

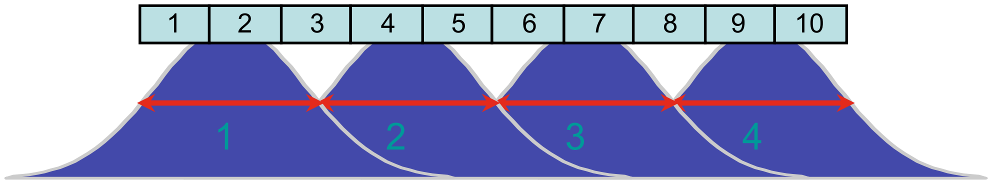


Smoothness



Smoothness parameterised in terms of FWHM:
Size of Gaussian kernel required to smooth i.i.d. noise to have same smoothness as observed null (standardized) data.

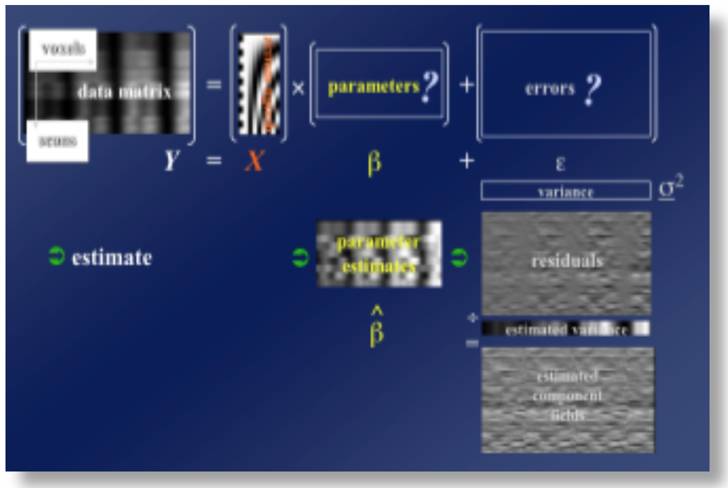
RESELS (Resolution Elements):
1 RESEL = $FWHM_x FWHM_y FWHM_z$
RESEL Count *R* = volume of search region in units of smoothness



Eg: 10 voxels, 2.5 FWHM, 4 RESELS

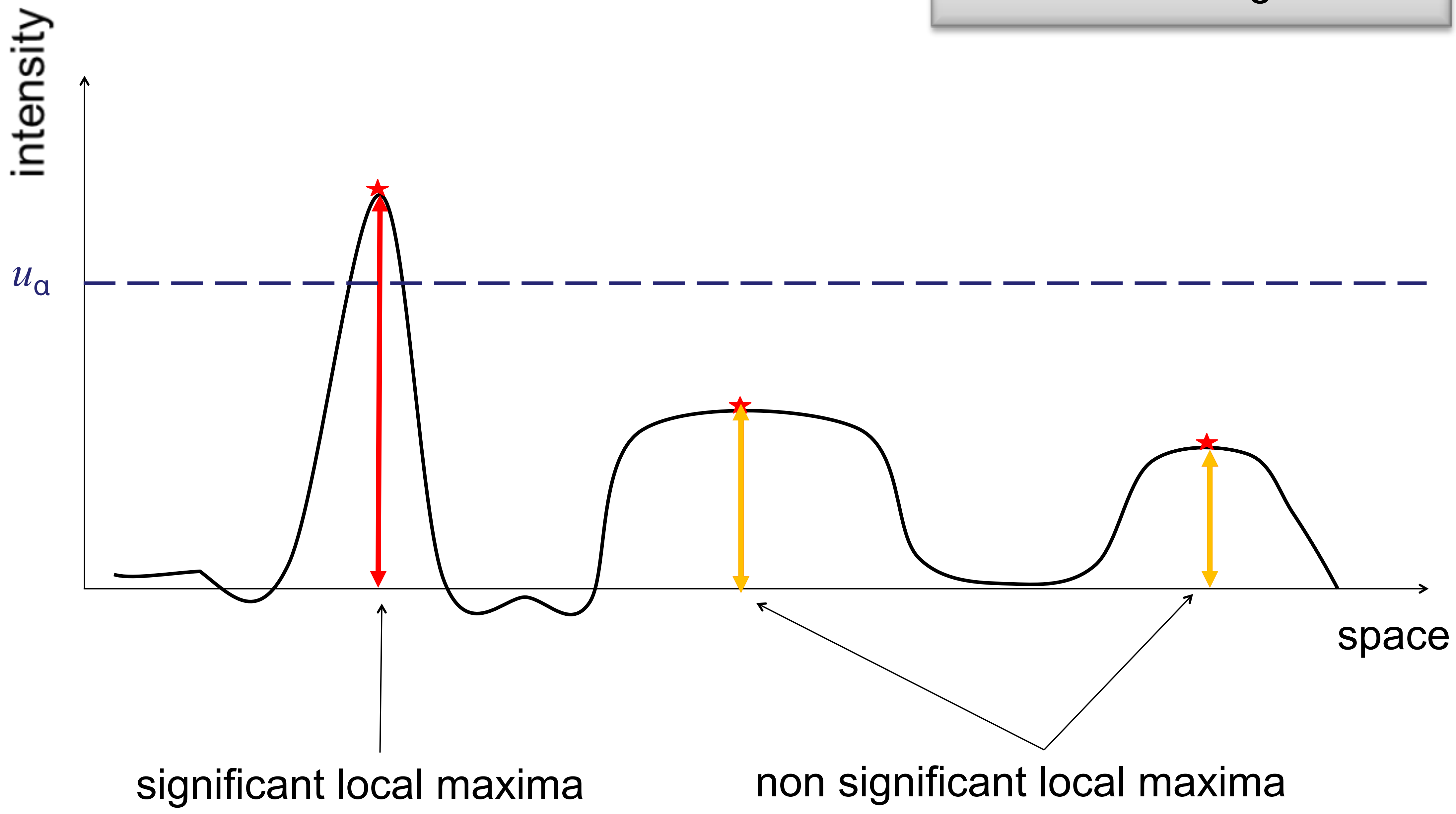
The number of resels is similar, but not identical to the number independent observations.

Smoothness estimated from spatial derivatives of standardised residuals:
Yields an RPV image containing local roughness estimation.



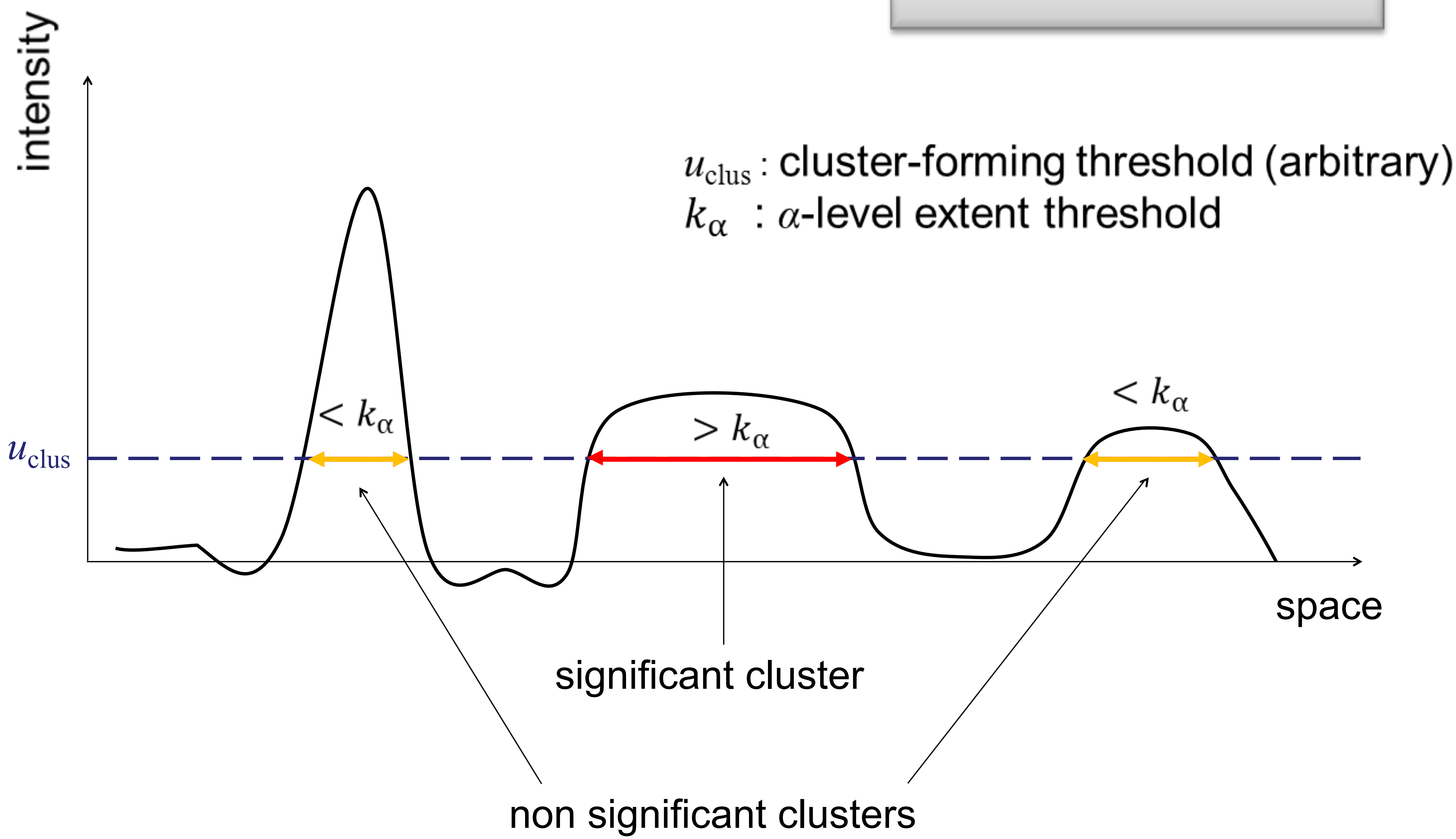
Topological inference

Topological feature:
Peak height



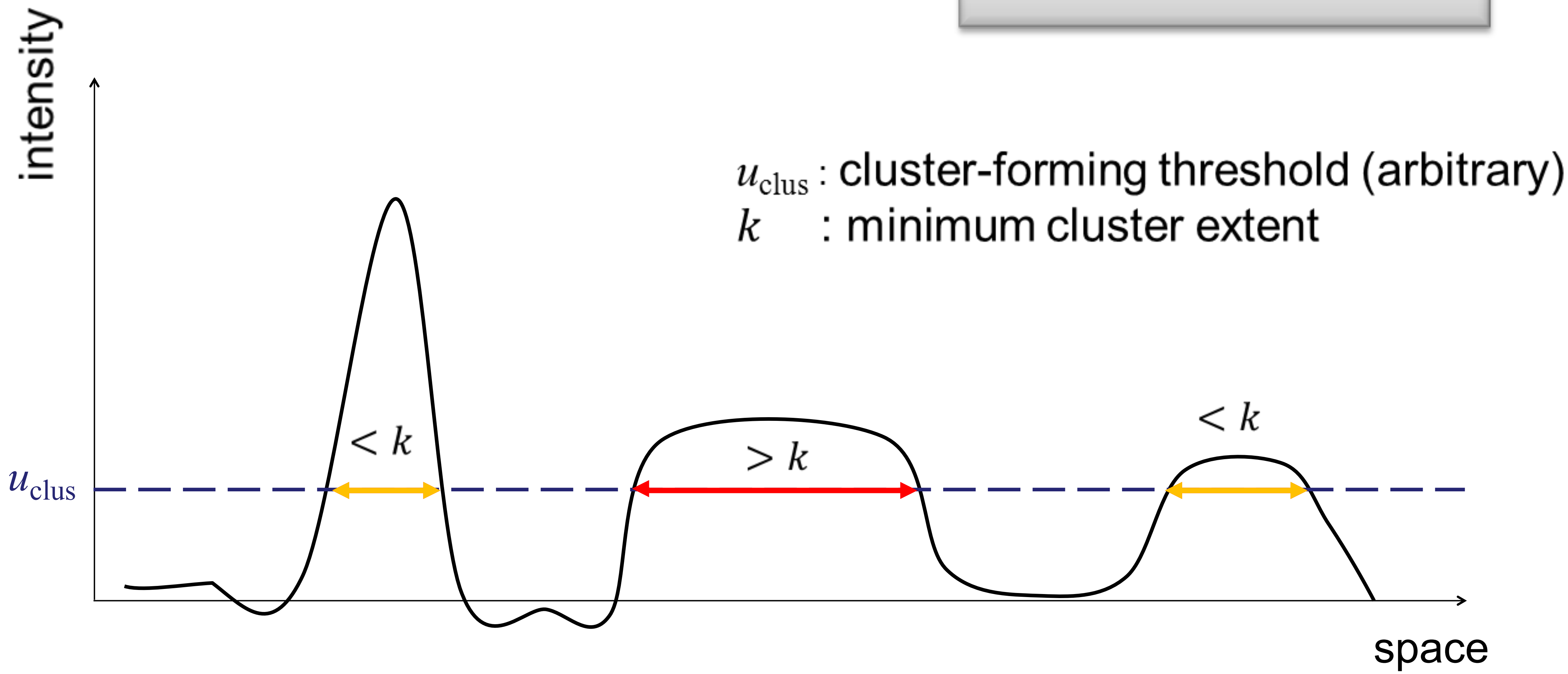
Topological inference

Topological feature:
Cluster extent



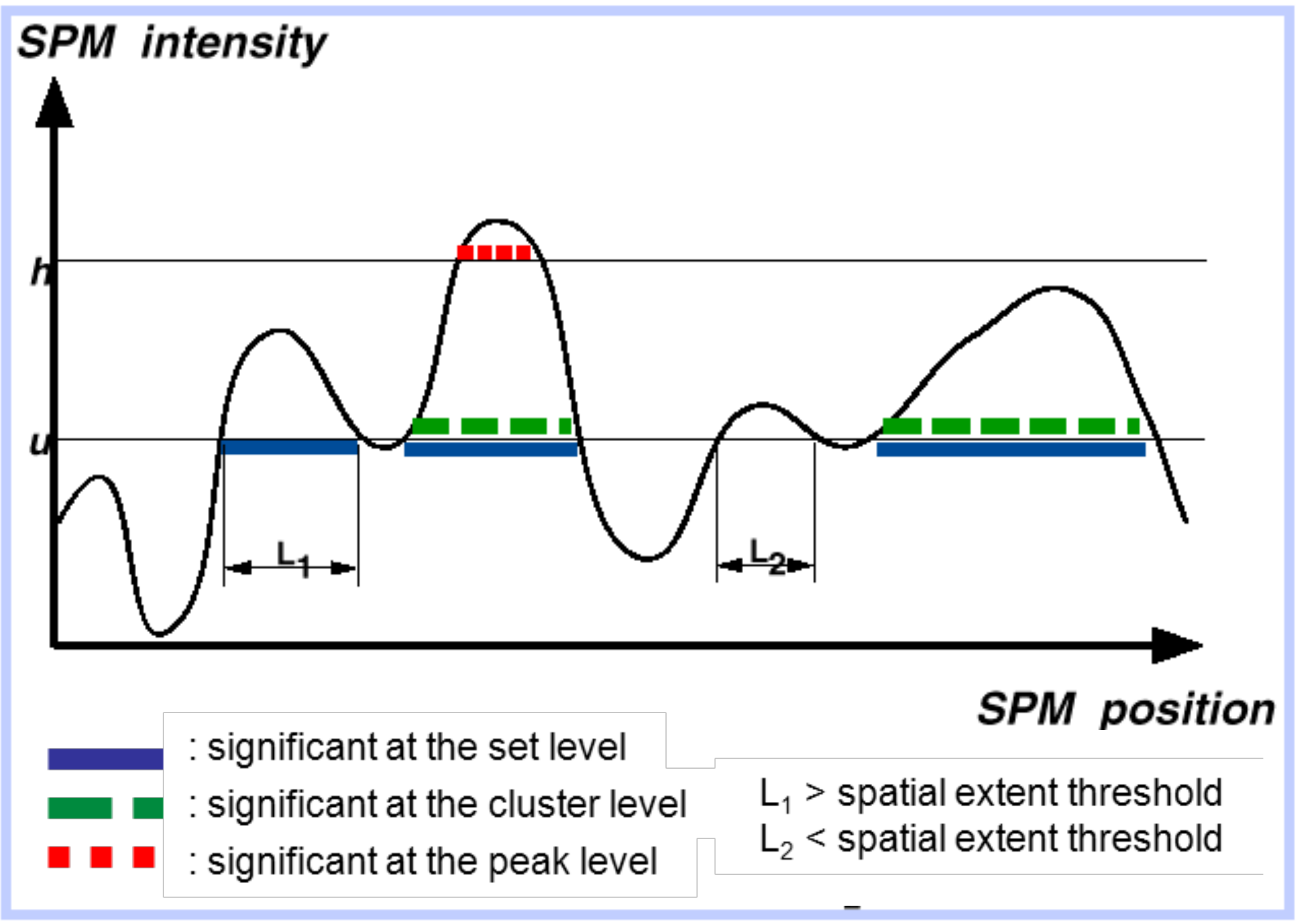
Topological inference

Topological feature:
Number of clusters



Here, $c=1$, only one cluster larger than k.

Peak, cluster and set level inference



Sensitivity

Regional specificity

Peak level test:
height of local maxima

Cluster level test:
spatial extent above u

Set level test:
number of clusters above u

Random Field Theory

- The statistic image is assumed to be a good lattice representation of an underlying continuous stationary random field.
  Typically, FWHM > 3 voxels
  (combination of intrinsic and extrinsic smoothing)

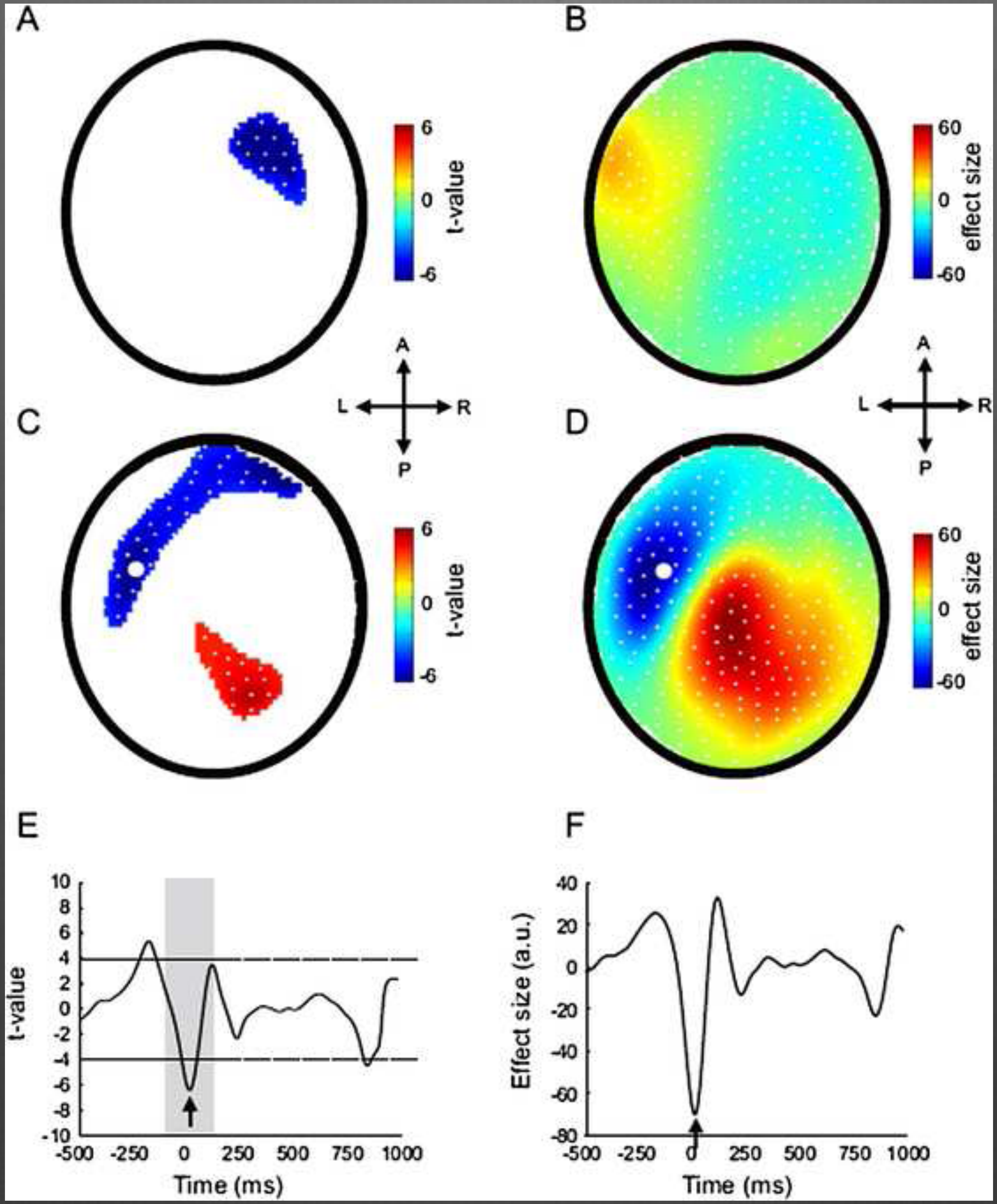
- Smoothness of the data is unknown and estimated:
  very precise estimate by pooling over voxels ⇒ stationarity assumptions (esp. relevant for cluster size results).

- RFT conservative for low degrees of freedom
  (always compare with Bonferroni correction).
  Afford littles power for group studies with small sample size.

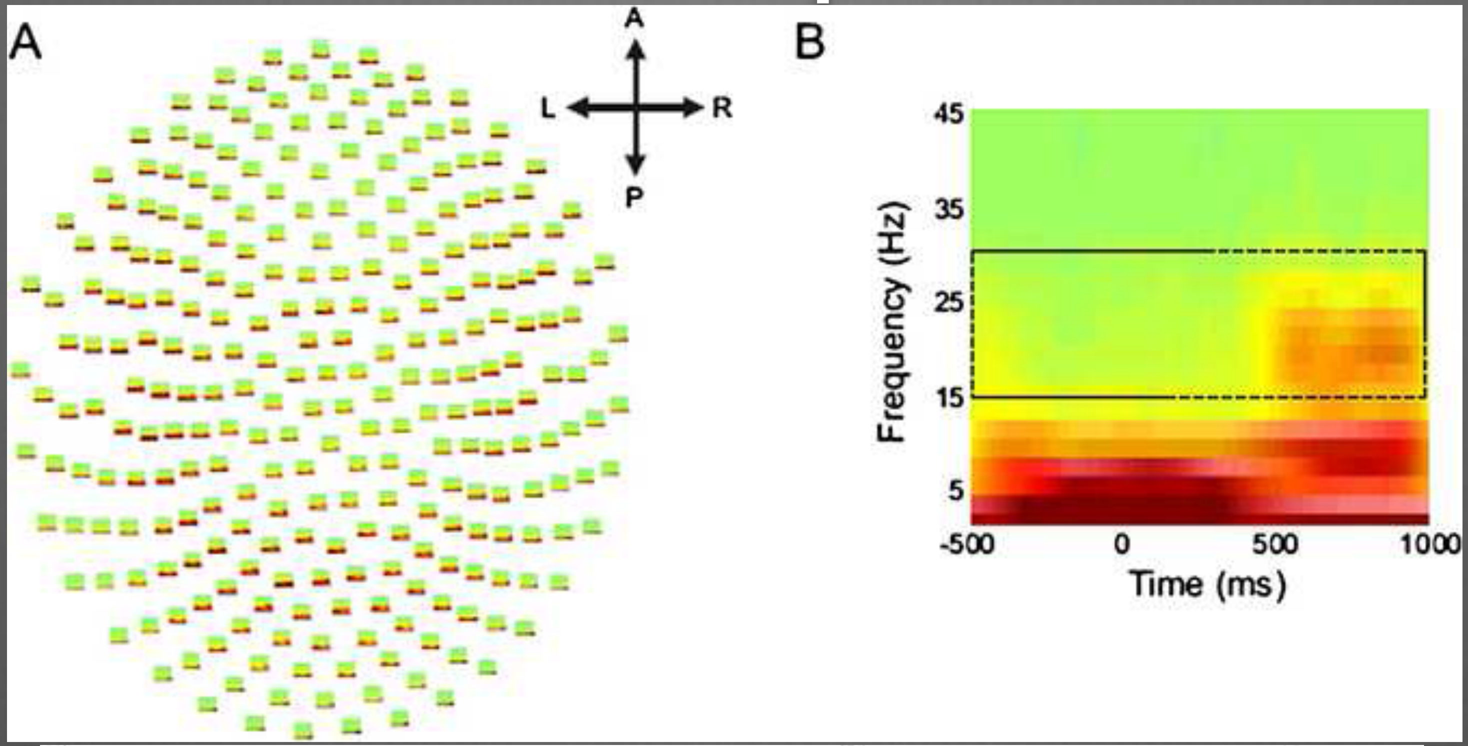
- *A priori* hypothesis about where an activation should be,
  reduce search volume ⇒ Small Volume Correction:

  - mask defined by (probabilistic) anatomical atlases
  - mask defined by separate "functional localisers"
  - mask defined by orthogonal contrasts
  - (spherical) search volume around previously reported coordinates

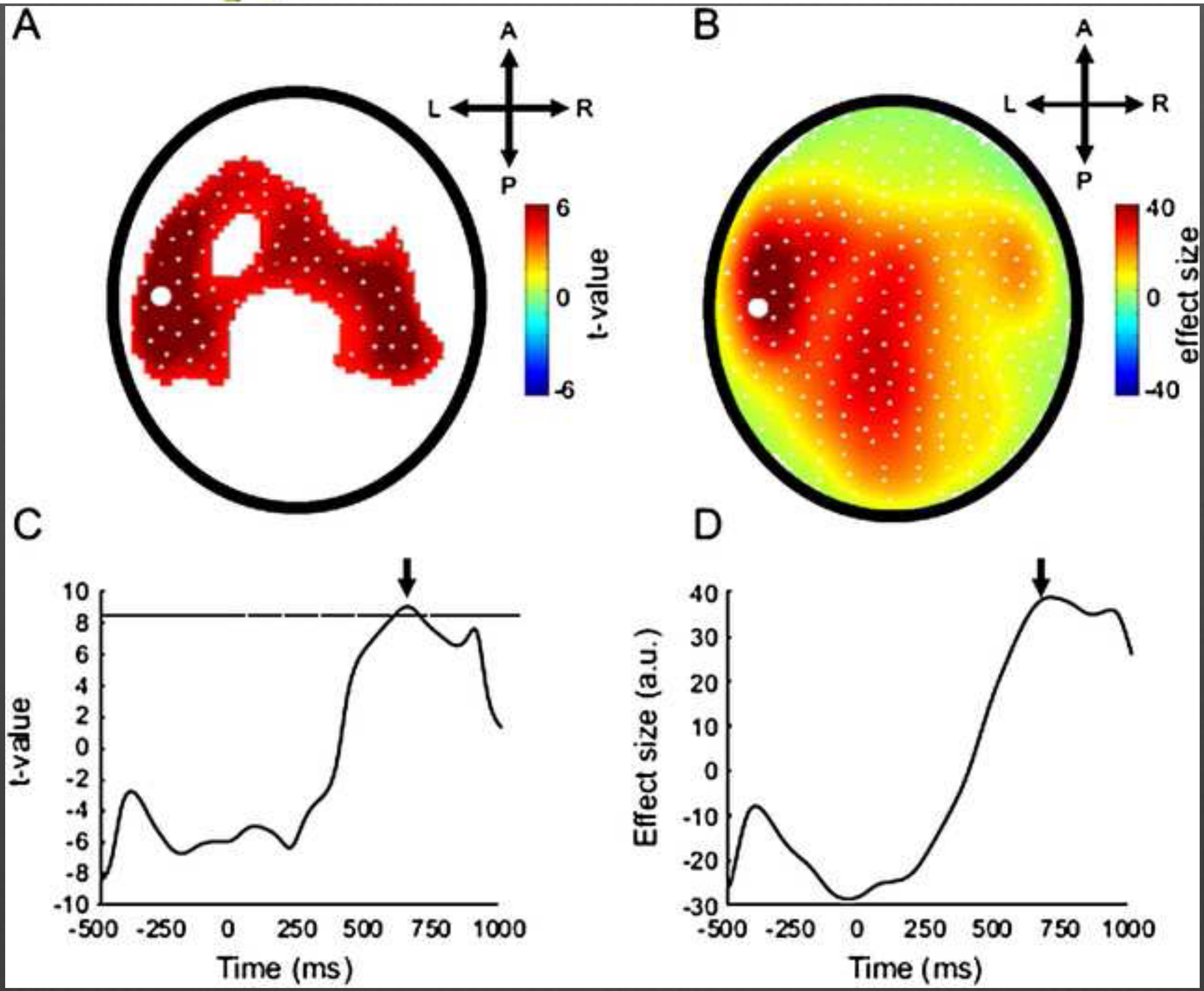
Examples


A
B
C
D
E
F
t-value
effect size
6
0
-6
60
0
-60
A
L
R
P
Time (ms)
Effect size (a.u.)
-500 -250 0 250 500 750 1000

Examples

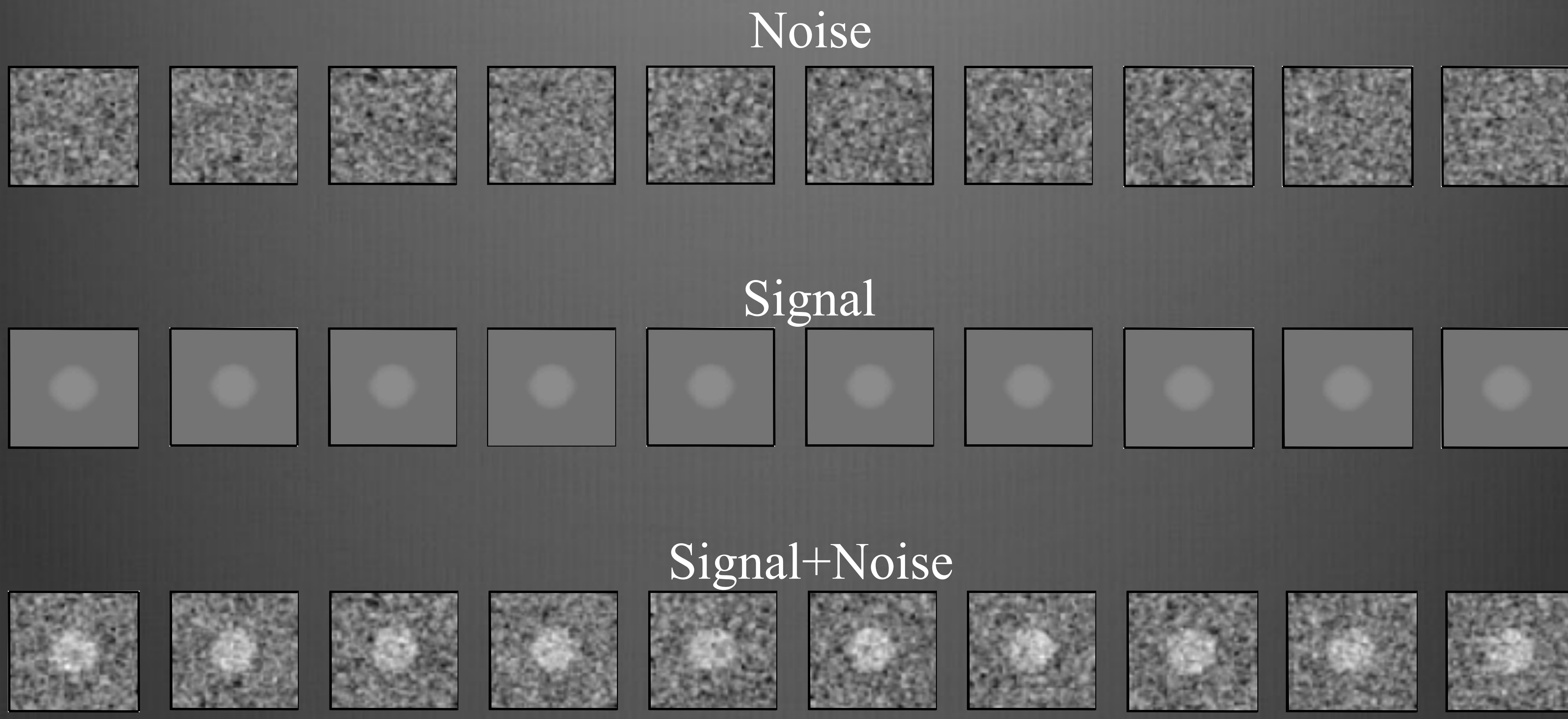


A
B
A
L
R
P
Frequency (Hz)
45
35
25
15
5
-500
0
500
1000
Time (ms)



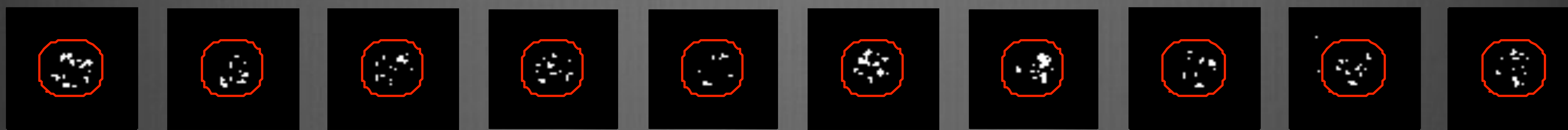
A
B
C
D
A
L
R
P
6
0
-6
t-value
40
0
-40
effect size
t-value
10
8
6
4
2
0
-2
-4
-6
-8
-10
-500
-250
0
250
500
750
1000
Time (ms)
Effect size (a.u.)
40
30
20
10
0
-10
-20
-30
-500
-250
0
250
500
750
1000
Time (ms)

MCP example



False discovery rate

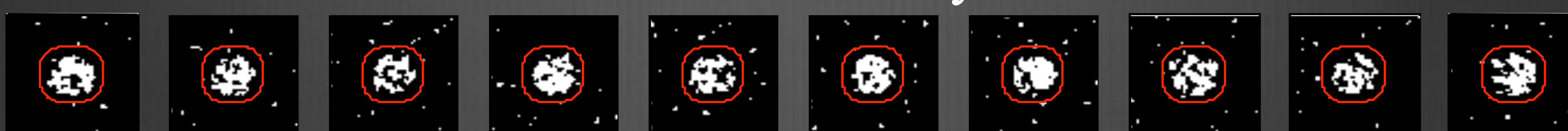
Control of Familywise Error Rate at 10%



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

Conclusion

- There is a ***multiple testing problem*** and *corrections* have to be applied on *p*-values (for the volume of interest only (see SVC)).

- Inference is made about ***topological features*** (peak height, spatial extent, number of clusters).
  Use results from the ***Random Field Theory***.

- **Control of *FWER*** (probability of a false positive anywhere in the image): very specific, not so sensitive.

- **Control of FDR** (expected proportion of false positives amongst those features declared positive (the discoveries)): less specific, more sensitive.