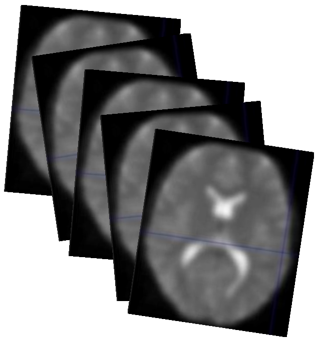
Multiple comparisons: problem and solutions

Gareth Barnes

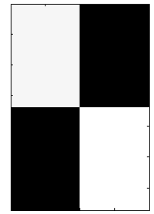
Wellcome Centre for Human Neuroimaging
University College London

By the end of this talk

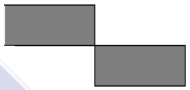
- Understand what the multiple comparisons problem is.
- Be familiar with some common approaches.
- Be able to explain Random field theory.



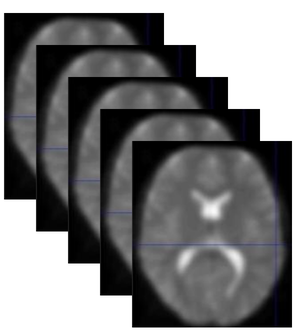
$$y = X\beta + \varepsilon$$



Contrast c

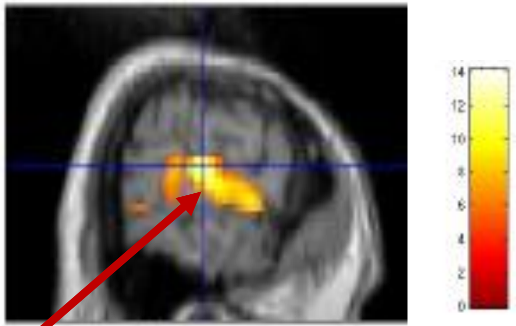


Pre-processing → **General Linear Model** → **Statistical Inference**



$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N - p}$$



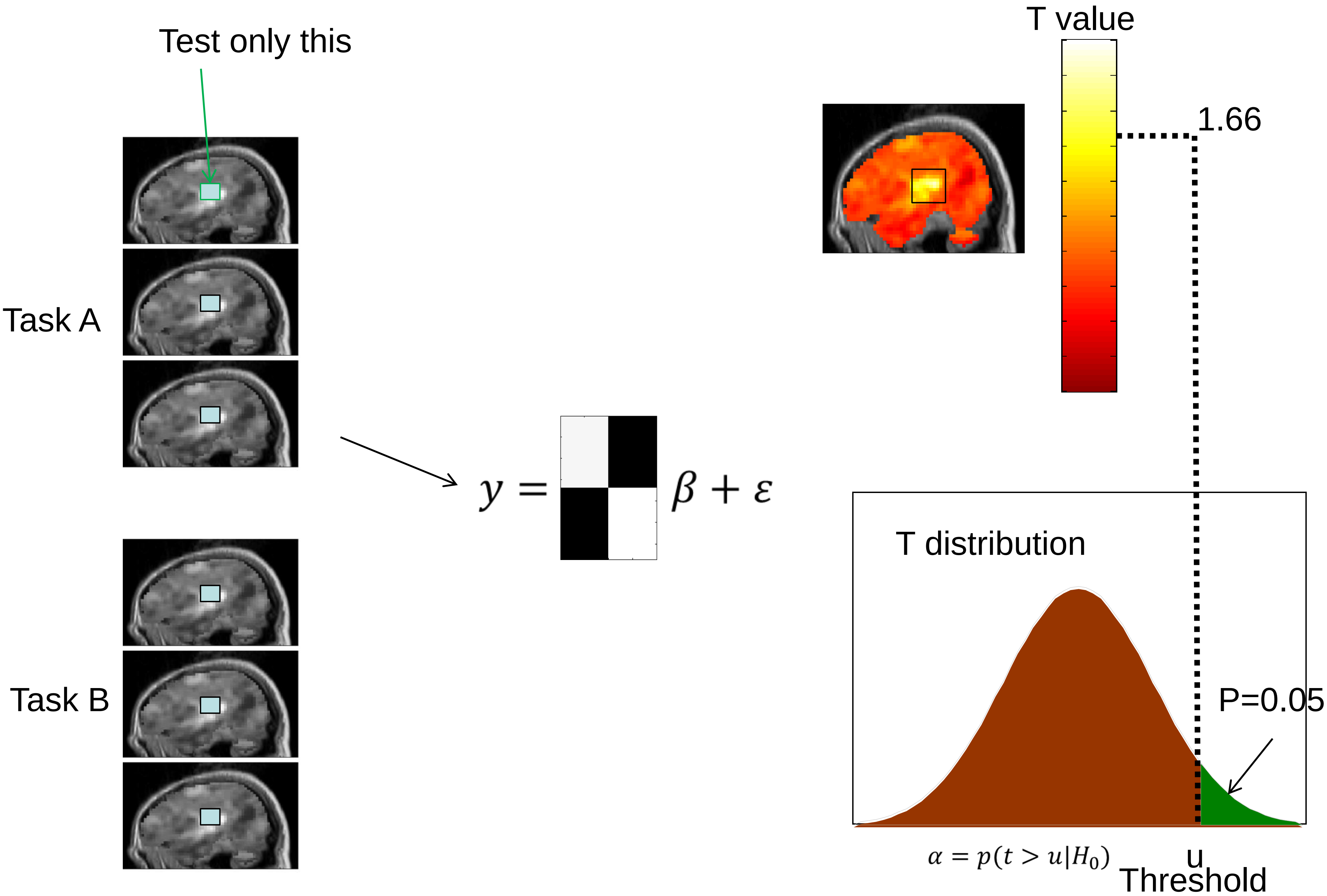
$SPM\{T,F\}$

Is this real ?

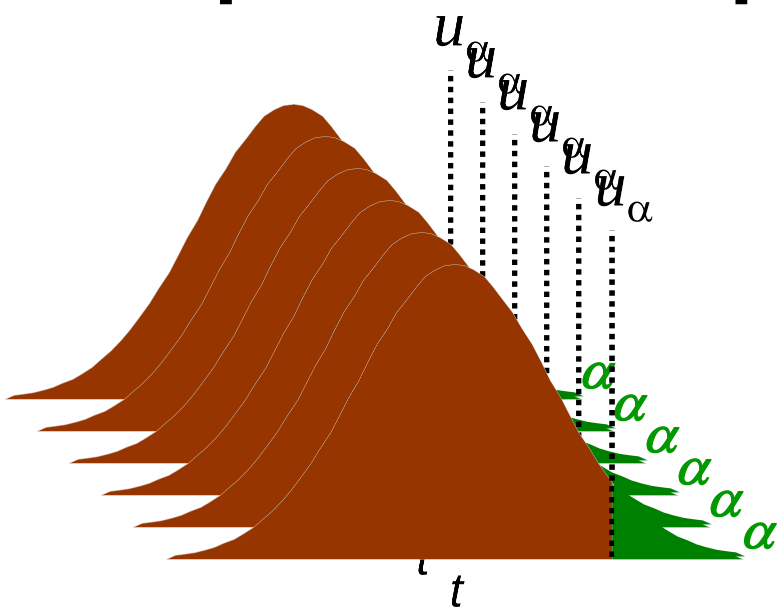
- Need to avoid cherry-picking. i.e. need an objective threshold.

- Need to adjust this threshold depending on how many independent tests you do. (e.g. if you do 20 tests with a false positive rate of 1/20 .. then expect one peak just due to chance).

T test on a single voxel / time point



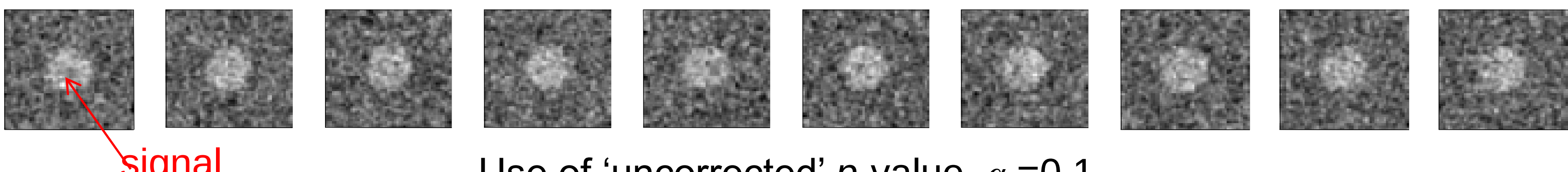
Multiple tests in space



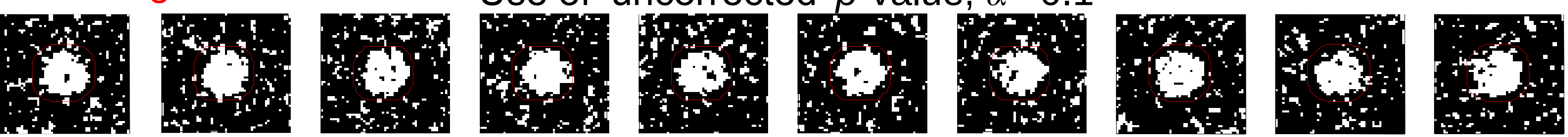
If we have 100,000 voxels,

α=0.05 $\Rightarrow$ 5,000 false positive voxels.

This is clearly undesirable; to correct for this we can define a null hypothesis for a collection of tests.



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

Bonferroni correction

Set the test-wise error rate (α) to be a the ideal Family-Wise Error rate (FWER) (α_{FWE}) divided by the number of tests.

$$\alpha = \frac{\alpha_{FWE}}{N}$$

e.g. for five tests choose $p<0.01$ for each test to control family at $p<0.05$

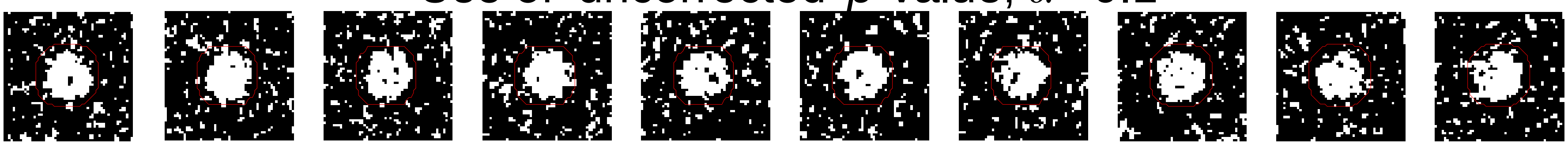
This correction does not require the tests to be independent but becomes very stringent if they are not.

Family-Wise Error Rate

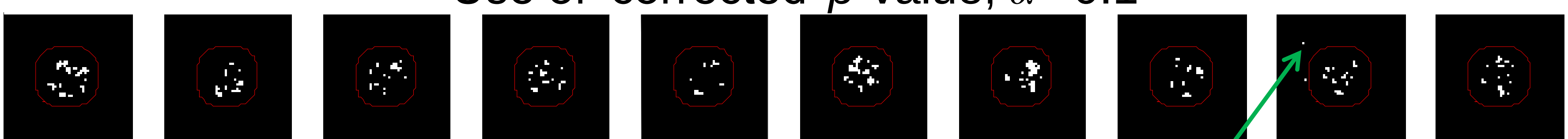
Family-Wise Error rate (FWER) = ‘*corrected*’ *p*-value

i.e. p<0.01 corrected, i.e. 1 false positive every 10 experiments

Use of ‘uncorrected’ *p*-value, α =0.1



Use of ‘corrected’ *p*-value, α =0.1



False positive

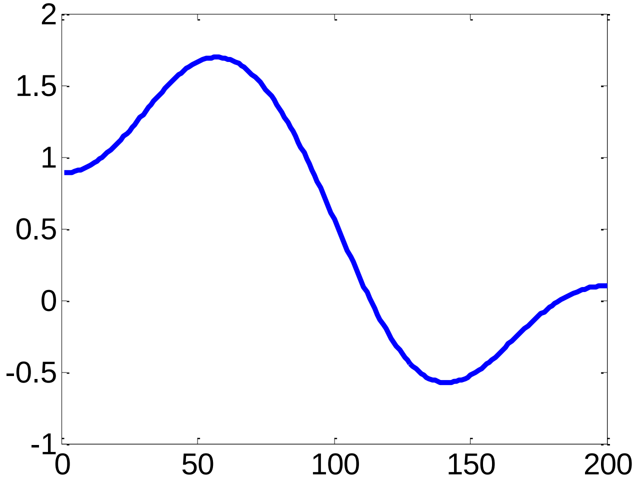
Summary

- Typically we control one test. Set threshold such that one in twenty tests we will get a false positive ($p<0.05$).

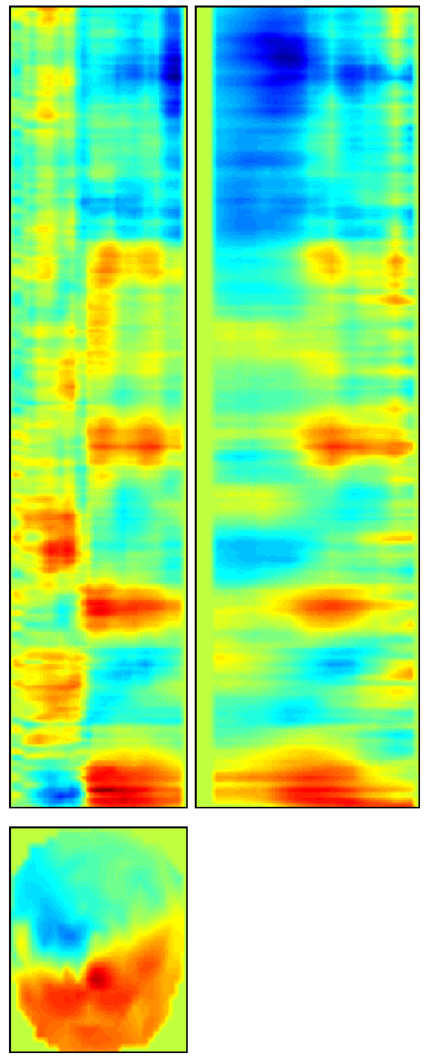
- Need to set family wise error rate so that in one in twenty experiments you will get a false positive ($p<0.05$, corrected).

- Do this by setting a much more conservative threshold.

What about data with different topologies and smoothness..

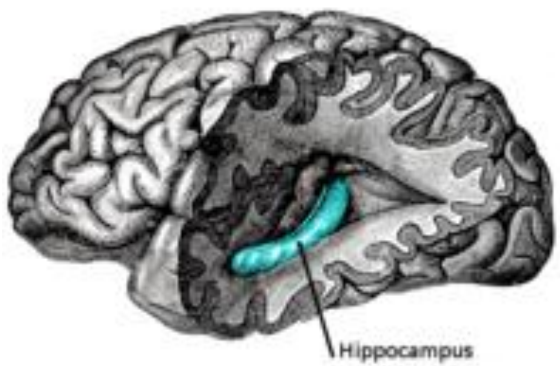


Smooth time

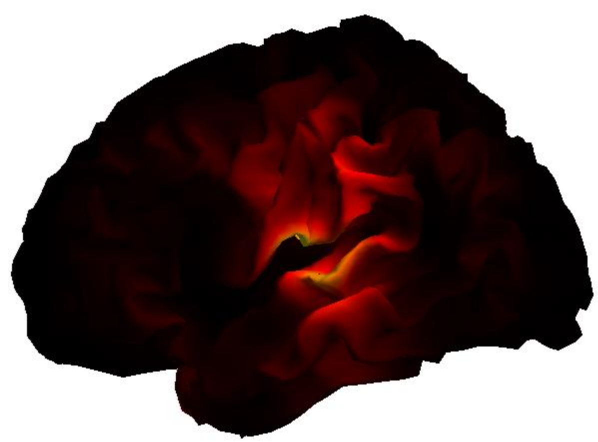


Different smoothness
In time and space

Volumetric ROIs



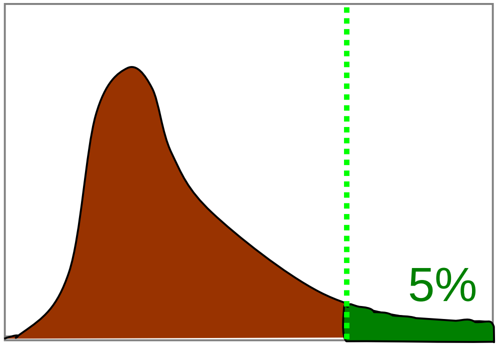
Surfaces



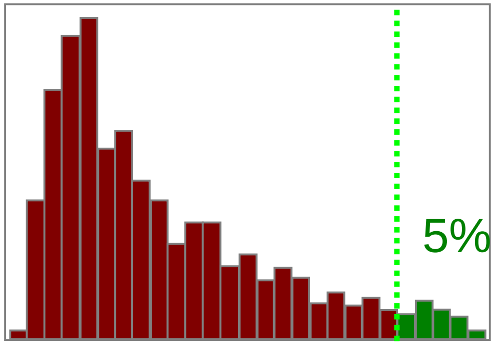
Non-parametric inference: permutation tests

to control FWER

- ❑ Parametric methods
  - – Assume distribution of *max* statistic under null hypothesis.

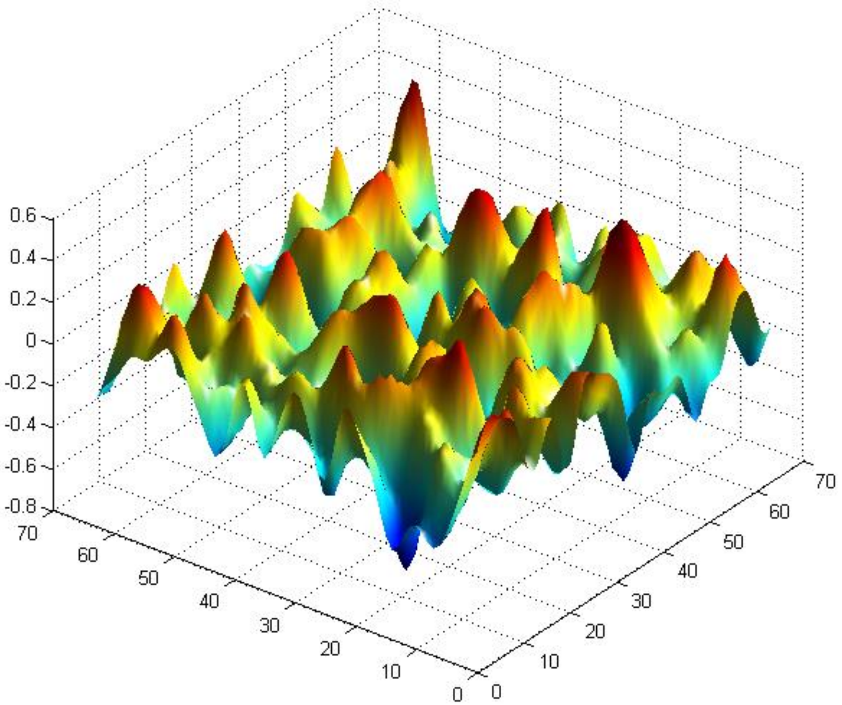


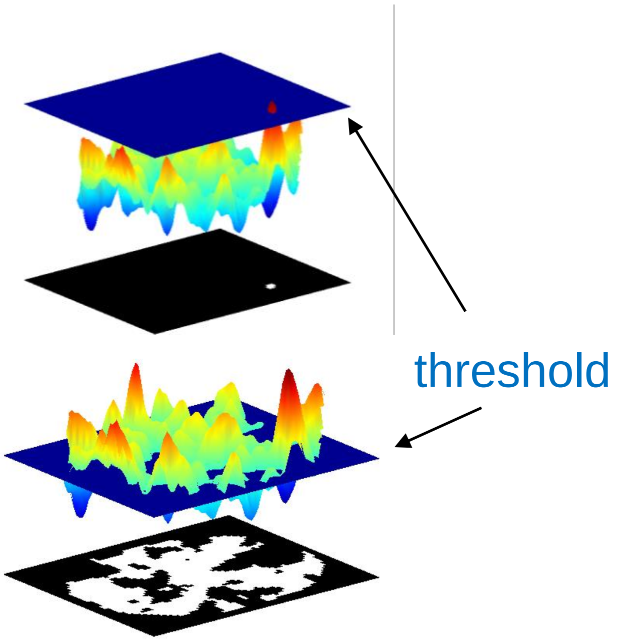
- ❑ Nonparametric methods
  - – Use *data* to find distribution of *max* statistic under null hypothesis.



Random Field Theory

A random field : an array of smoothly varying test statistics.
e.g. a slice through a t-statistic brain image.

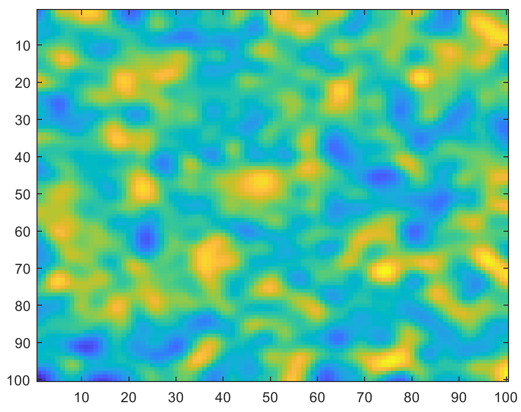


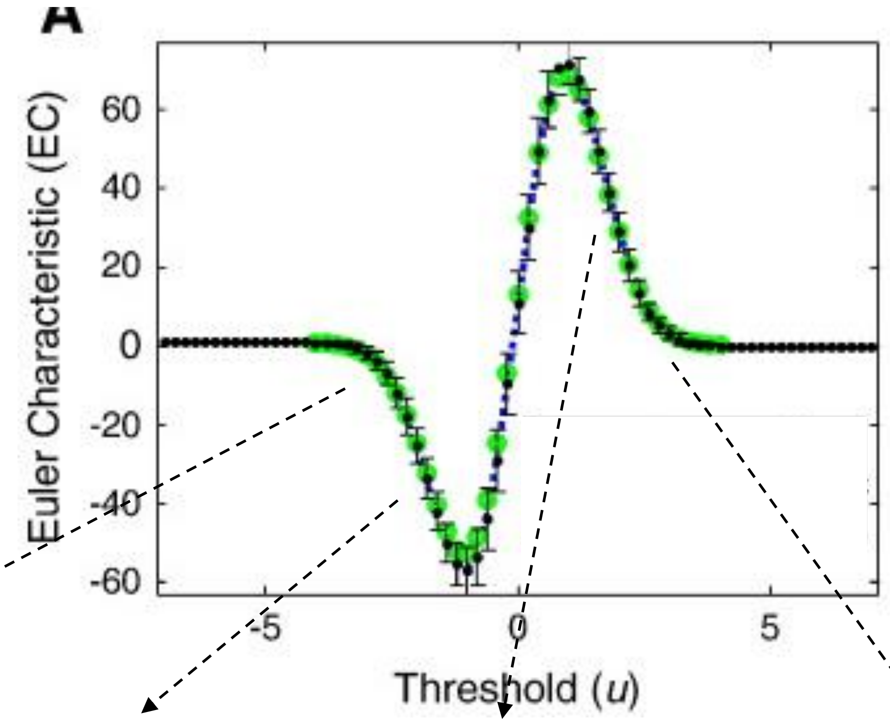


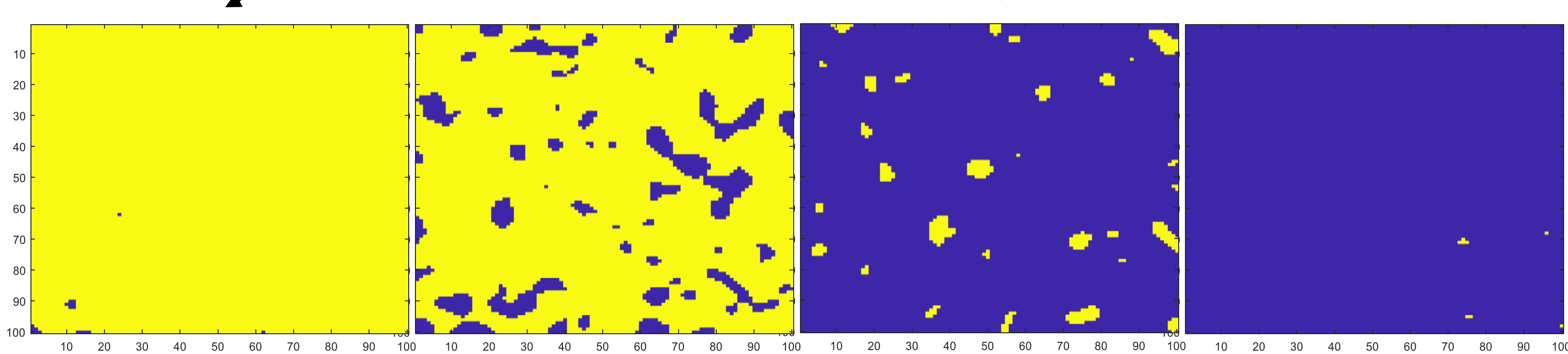
Keith Worsley, Karl Friston, Jonathan Taylor, Robert Adler and colleagues

Euler characteristic (EC) at threshold (u) = Number blobs- Number holes

Smooth Random field





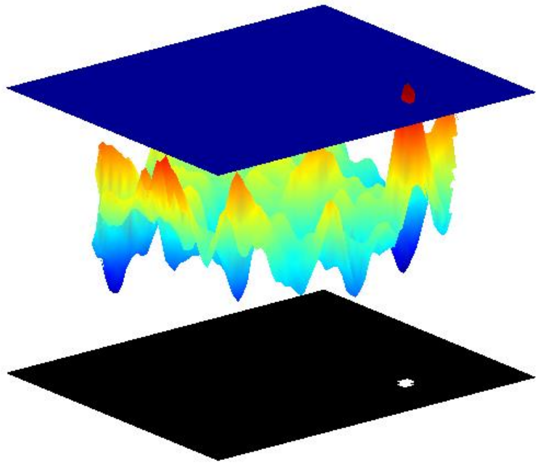


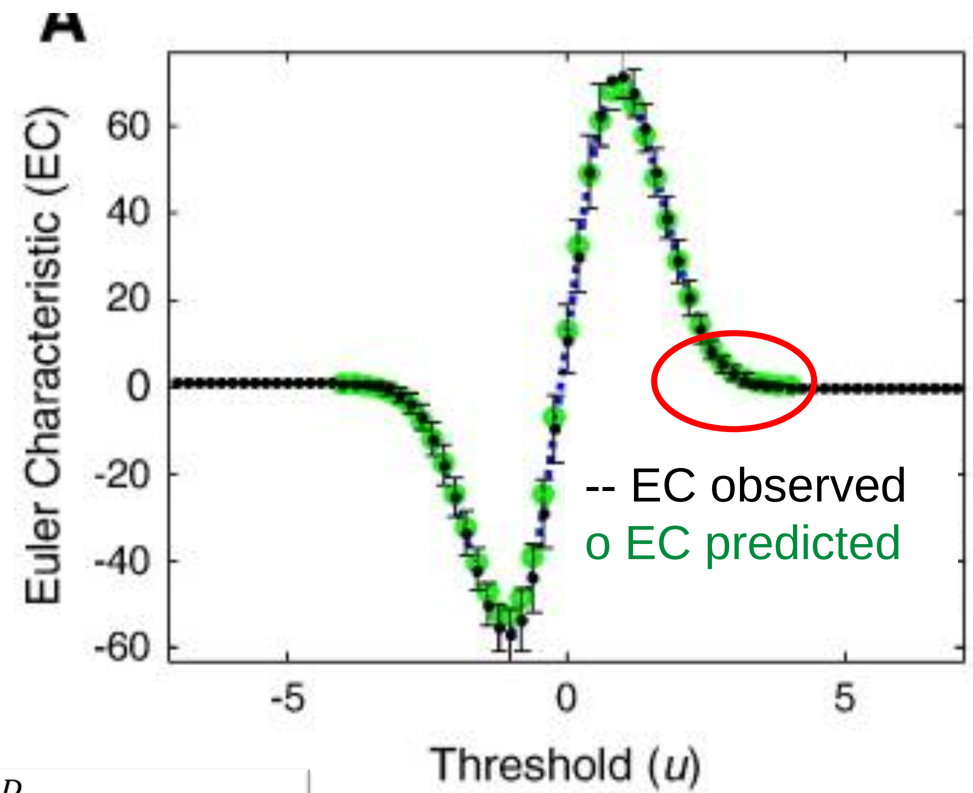
A few holes

Many holes

Many peaks

A few peaks





At high threshold,
EC = number of peaks

Expected Euler Characteristic

$$E[\chi_u(\Omega)] = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

Intrinsic volume (depends on shape and smoothness of space)

Depends on type of test, dimension and threshold

Number peaks= intrinsic volume * peak density

Currant bun analogy

$$E[\chi_u(\Omega)] = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

Number peaks= intrinsic volume * peak density

Number of currants = volume of bun * currant density

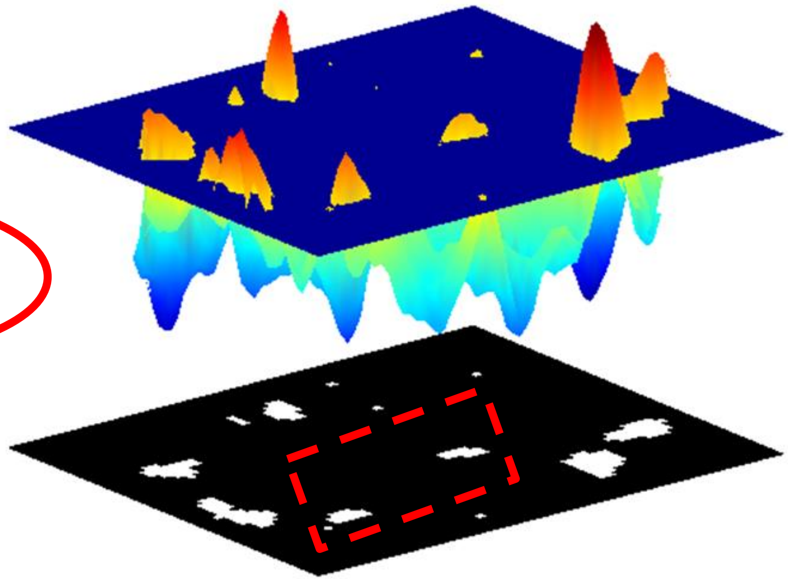
How do we specify peak (or EC) density

Number peaks= intrinsic volume * peak density

$$E[\chi_u(\Omega)] = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

EC density, $\rho_d(u)$

Number peaks= intrinsic volume * peak density



The EC density - depends on the type of random field (t, F, Chi etc), the dimension of the test (2D,3D), and the threshold. i.e. it is **data independent**

Number of currants = bun volume * currant density

Peak densities (as a function of threshold) for the different fields are known

See J.E. Taylor, K.J. Worsley. J. Am. Stat. Assoc., 102 (2007), pp. 913–928

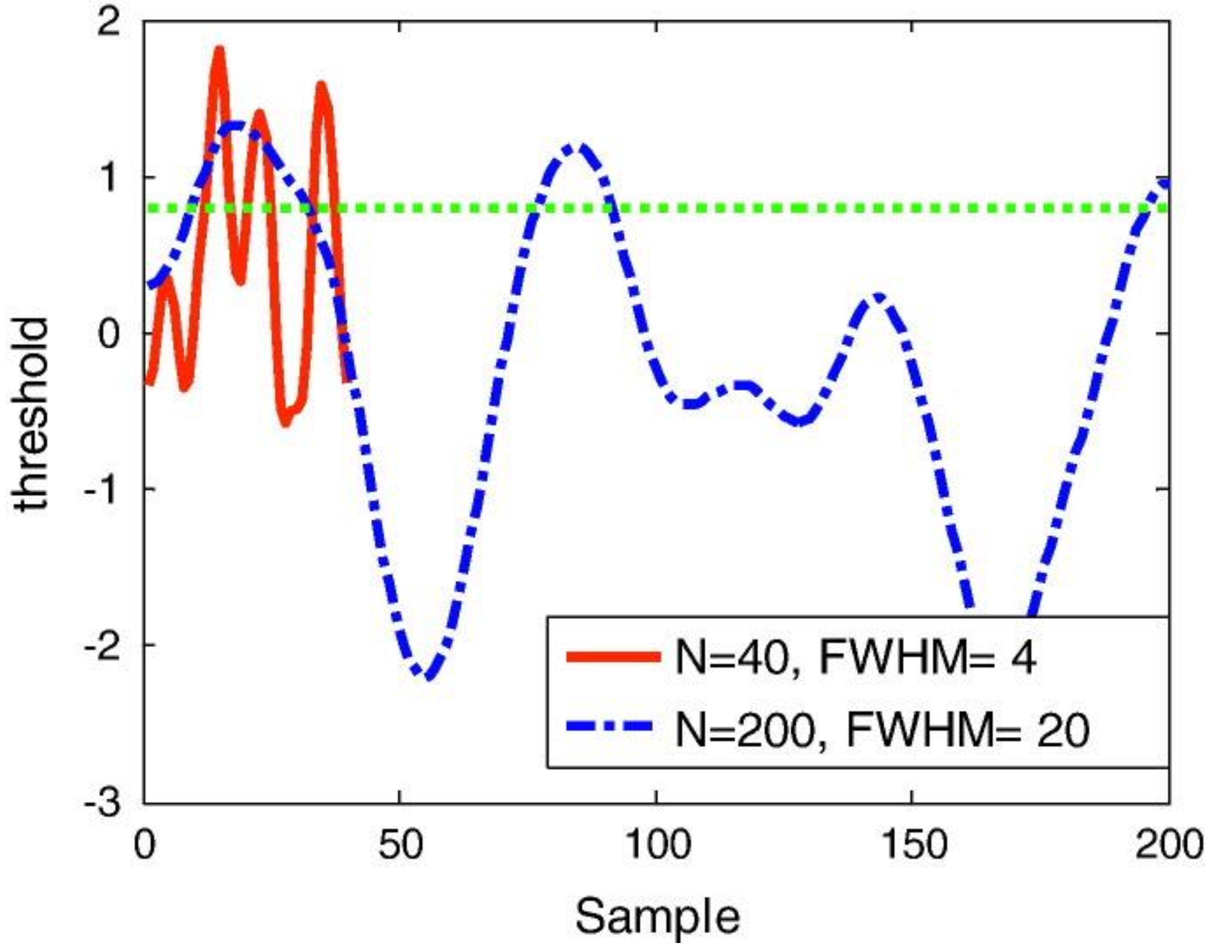
Currant bun analogy

$$E[\chi_u(\Omega)] = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

Number peaks= intrinsic volume * peak density

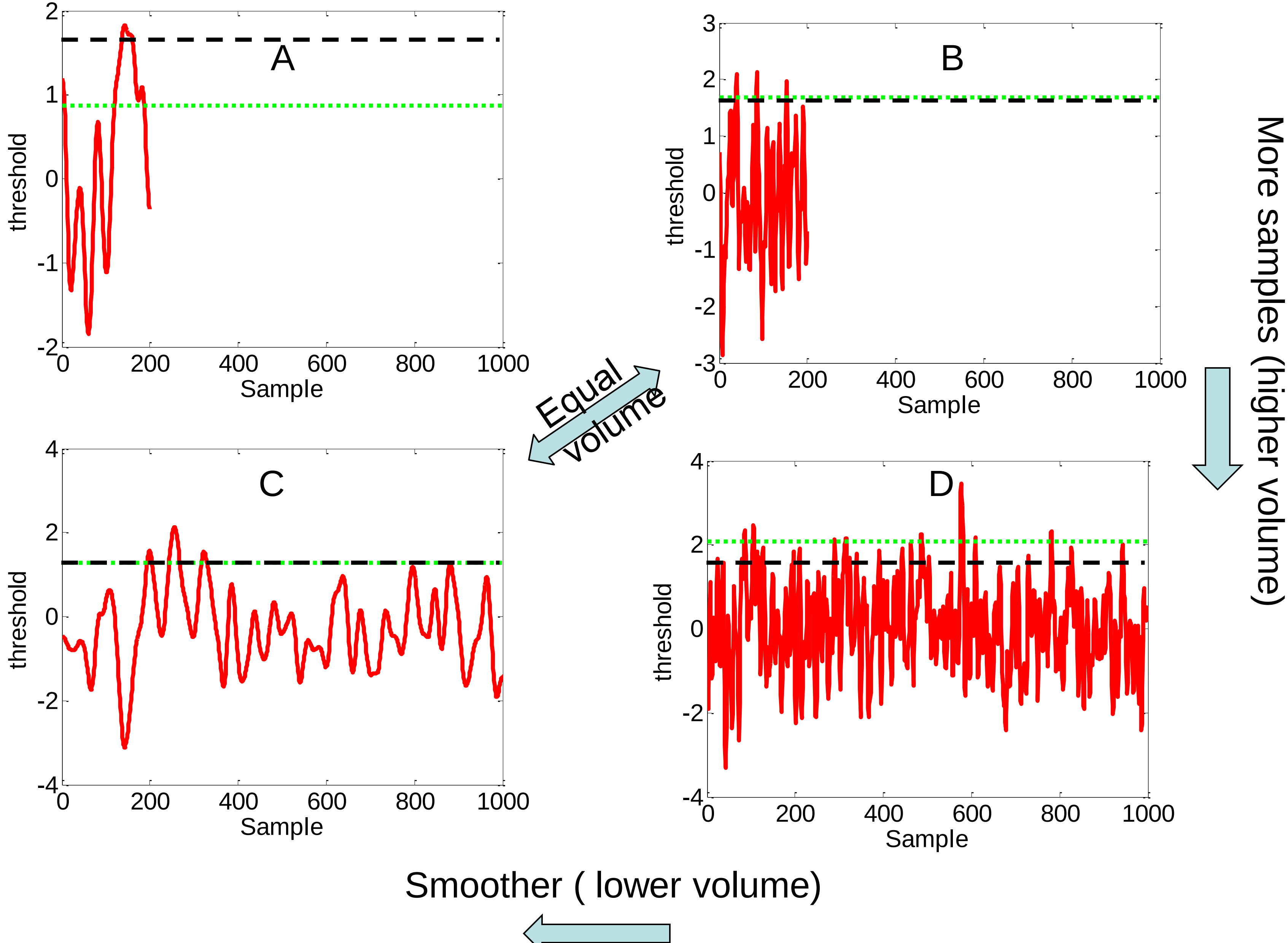
Number of currants = bun volume * currant density

LKC or resel estimates normalize volume

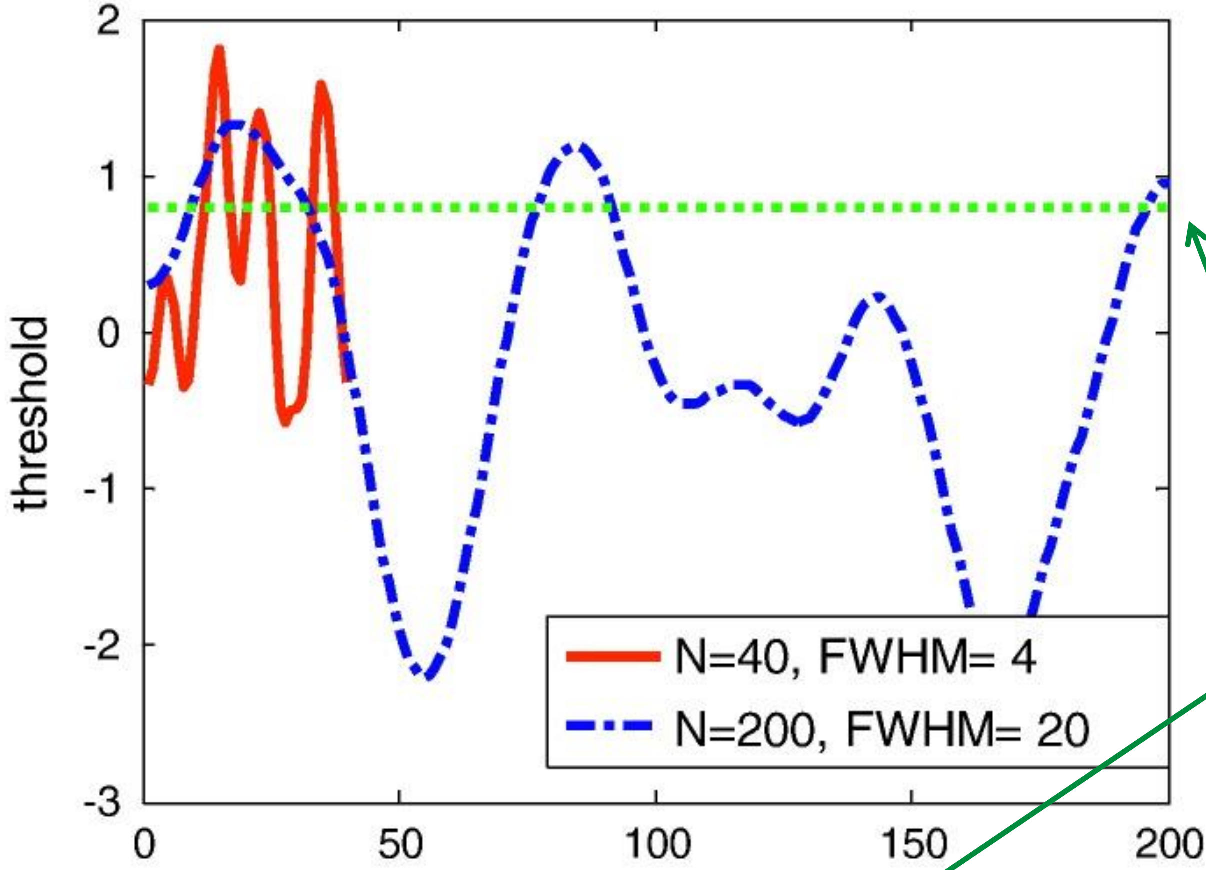


The intrinsic volume (or the number of resels or the Lipschitz-Killing Curvature) of the two fields is identical

Which field has highest intrinsic volume ?



Smoother (lower volume)



Threshold u

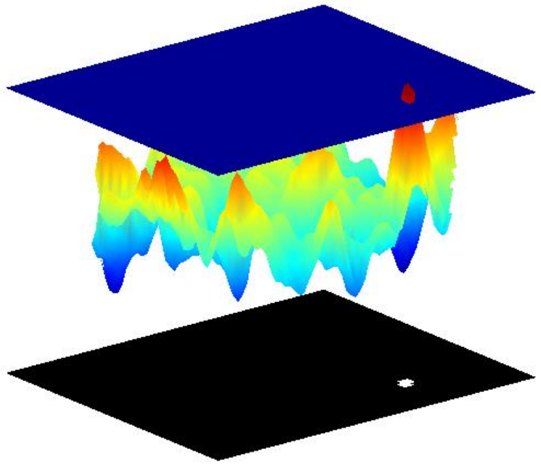
$$E[\chi_u(\Omega)] = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

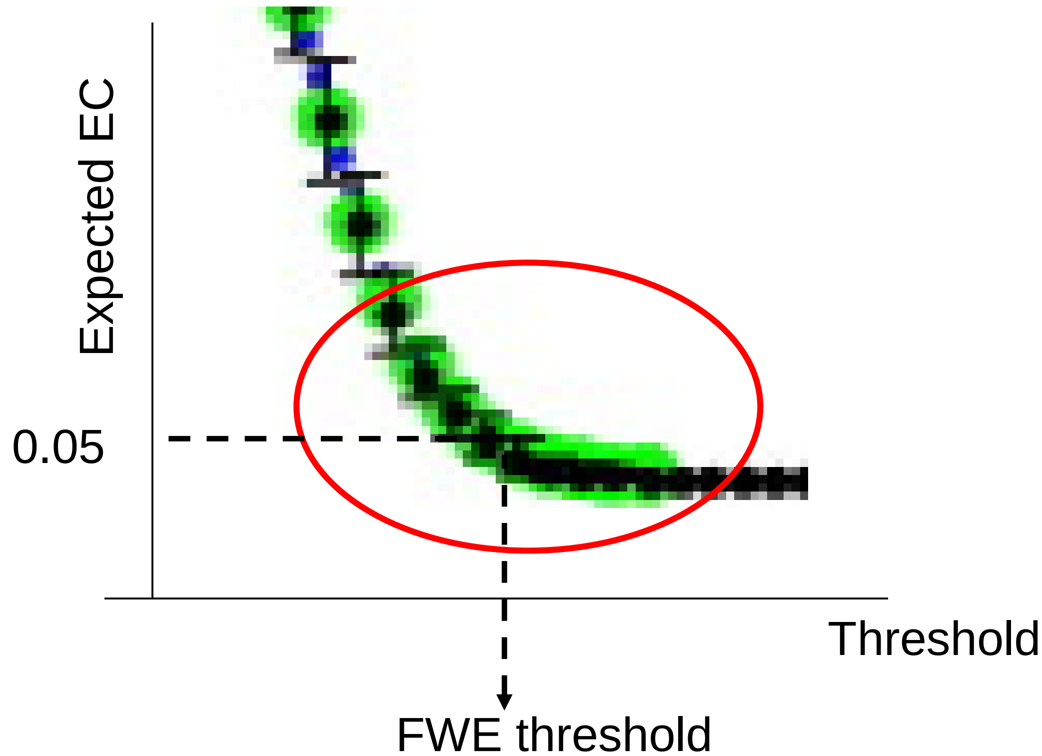
Expected Euler Characteristic

=2.9 (in this example)

Intrinsic volume (depends on shape and smoothness of space)

Depends only on type of test and dimension





i.e. we want one false positive every 20 realisations (FWE=0.05)

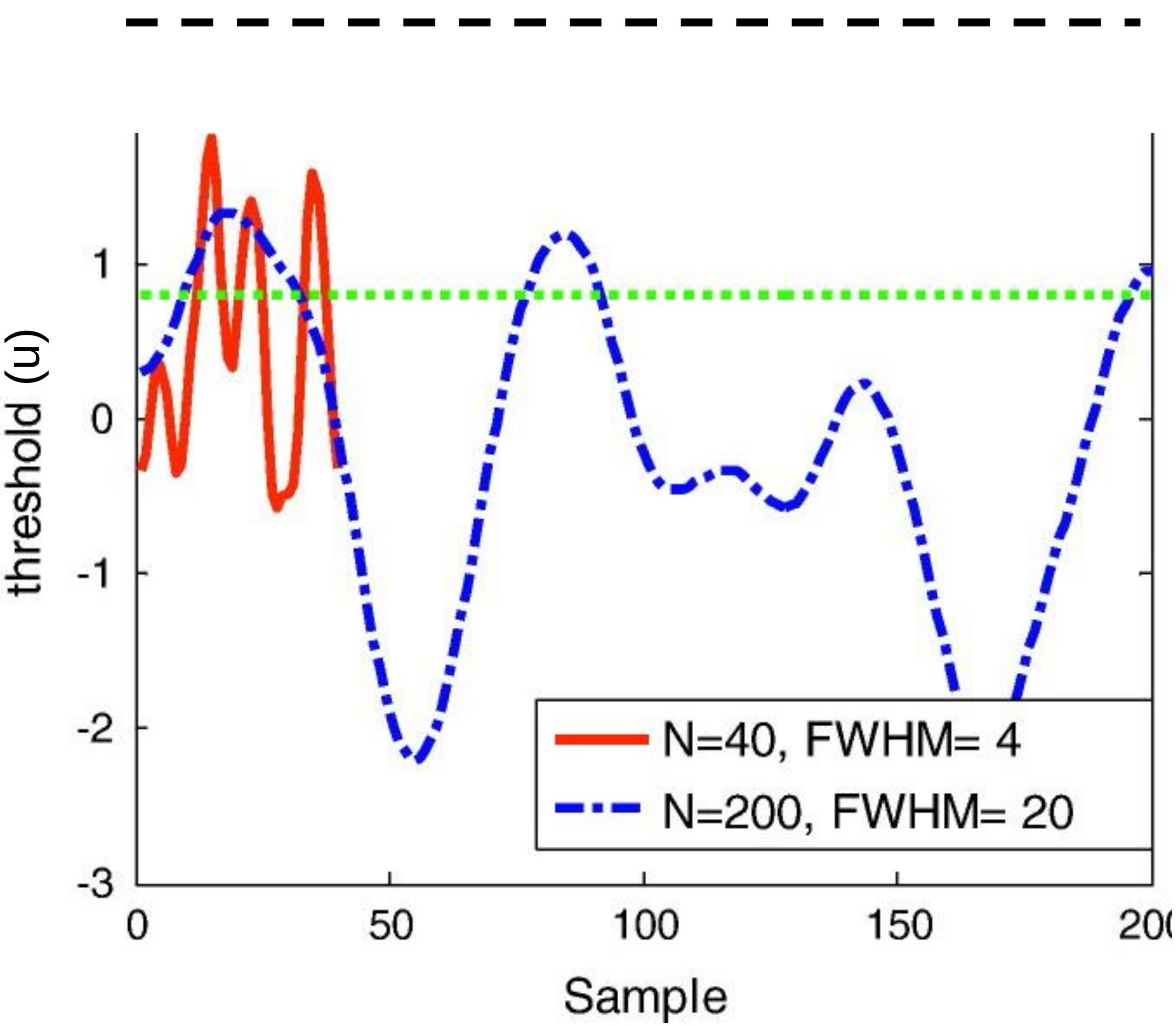
Expected Euler Characteristic

$$E[\chi_u(\Omega)] = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

Intrinsic volume (depends on shape and smoothness of space)

Depends only on type of test and dimension

Getting FWE threshold



Know test (t) and dimension (1) so can get threshold u

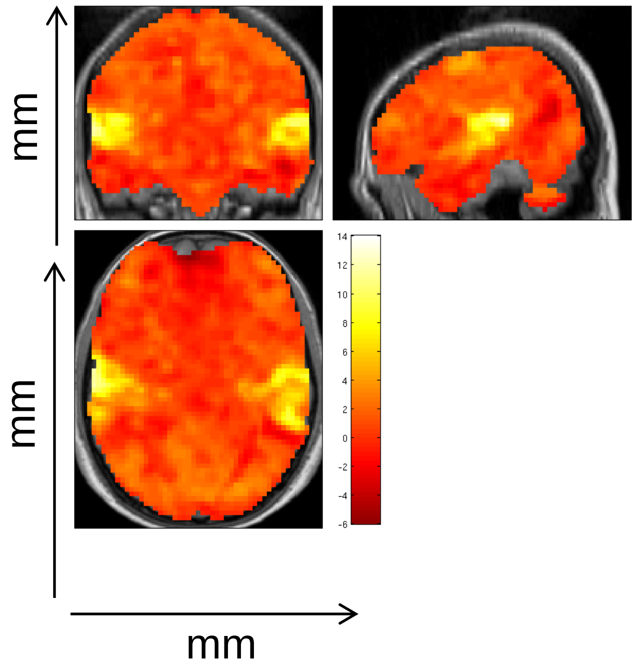
$$0.05 = \sum_{d=0}^{D} R_d(\Omega)\rho_d(u)$$

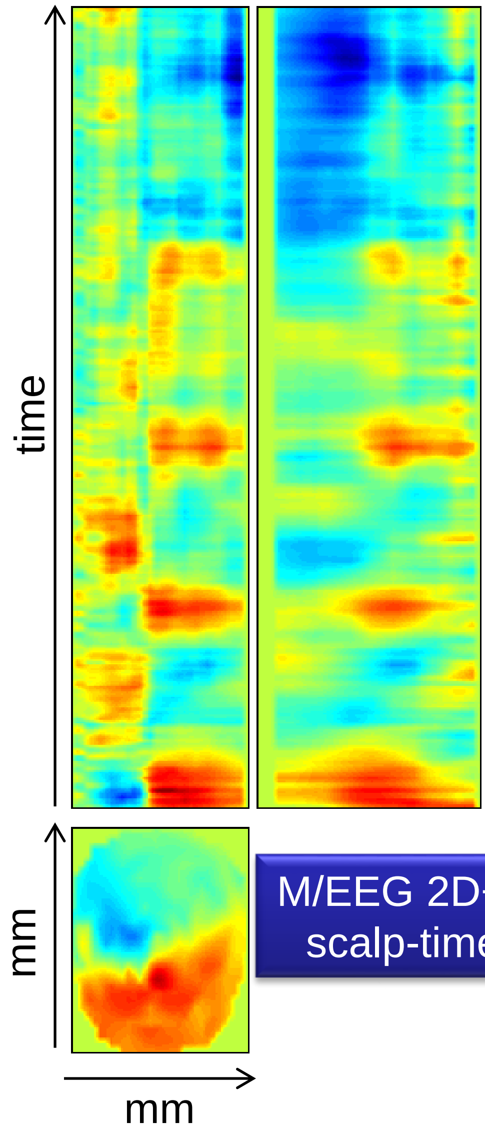
Know intrinsic volume (10 resels)

Want only a 1 in 20 Chance of a false positive

Can get correct FWE for any of these..

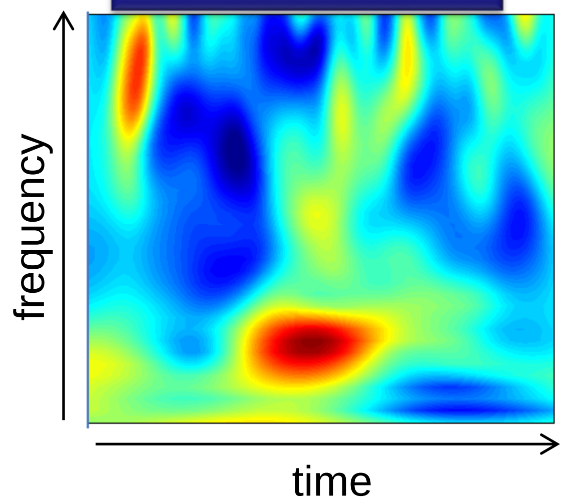
fMRI, VBM,
M/EEG source reconstruction



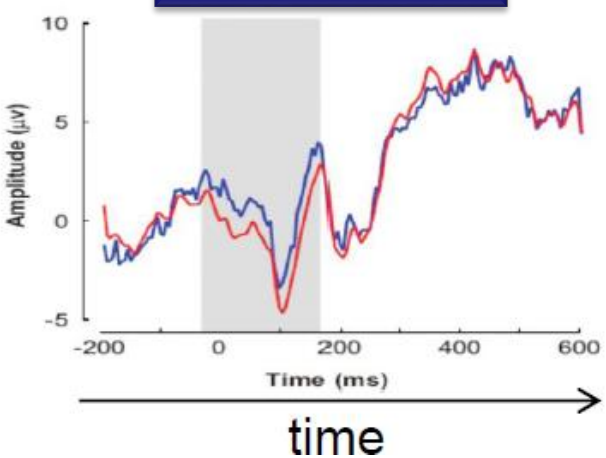


M/EEG 2D+t
scalp-time

M/EEG
2D time-frequency



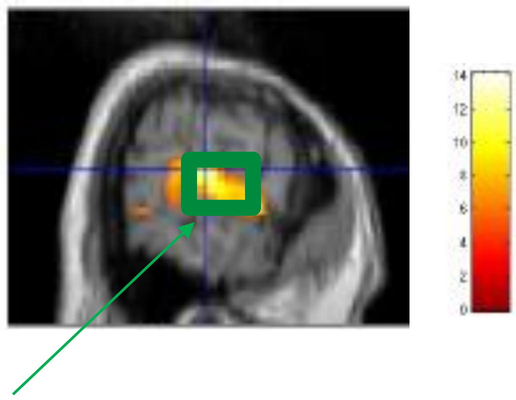
M/EEG
1D channel-time



If you already know something

If you know where or when (or both) (eg left auditory cortex at t=100ms) then you can avoid the multiple comparisons problem.

More powerful inference, simpler message.



Region of interest (ROI)
Defined before you do the experiment

To summarize

- Multiple comparisons problem- more tests, more false positives.
- Bonferroni correction – simple but conservative
- Random field theory- just like cooking- number of currants you would expect by chance.

Conclusions

- Strong prior hypotheses can reduce the multiple comparisons problem.

- Random field theory is a way of predicting the number of peaks you would expect in a purely random field (without signal).

- Can control FWE rate for a space of any dimension and shape.

Acknowledgments

Guillaume Flandin

Vladimir Litvak

James Kilner

Stefan Kiebel

Rik Henson

Karl Friston

Methods group at the FIL

References

- Friston KJ, Frith CD, Liddle PF, Frackowiak RS. *Comparing functional (PET) images: the assessment of significant change*. J Cereb Blood Flow Metab. 1991 Jul;11(4):690-9.
- Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. *A unified statistical approach for determining significant signals in images of cerebral activation*. Human Brain Mapping 1996;4:58-73.
- Chumbley J, Worsley KJ , Flandin G, and Friston KJ. *Topological FDR for neuroimaging*. NeuroImage, 49(4):3057-3064, 2010.
- Chumbley J and Friston KJ. *False Discovery Rate Revisited: FDR and Topological Inference Using Gaussian Random Fields*. NeuroImage, 2008.
- Kilner J and Friston KJ. *Topological inference for EEG and MEG data*. Annals of Applied Statistics, in press.
- Bias in a common EEG and MEG statistical analysis and how to avoid it. Kilner J. Clinical Neurophysiology 2013.