Large Scale Evaluation of Random Field Theory Inference in fMRI

Thomas Nichols, Ph.D.

Department of Statistics & WMG
University of Warwick

http://warwick.ac.uk/tenichols

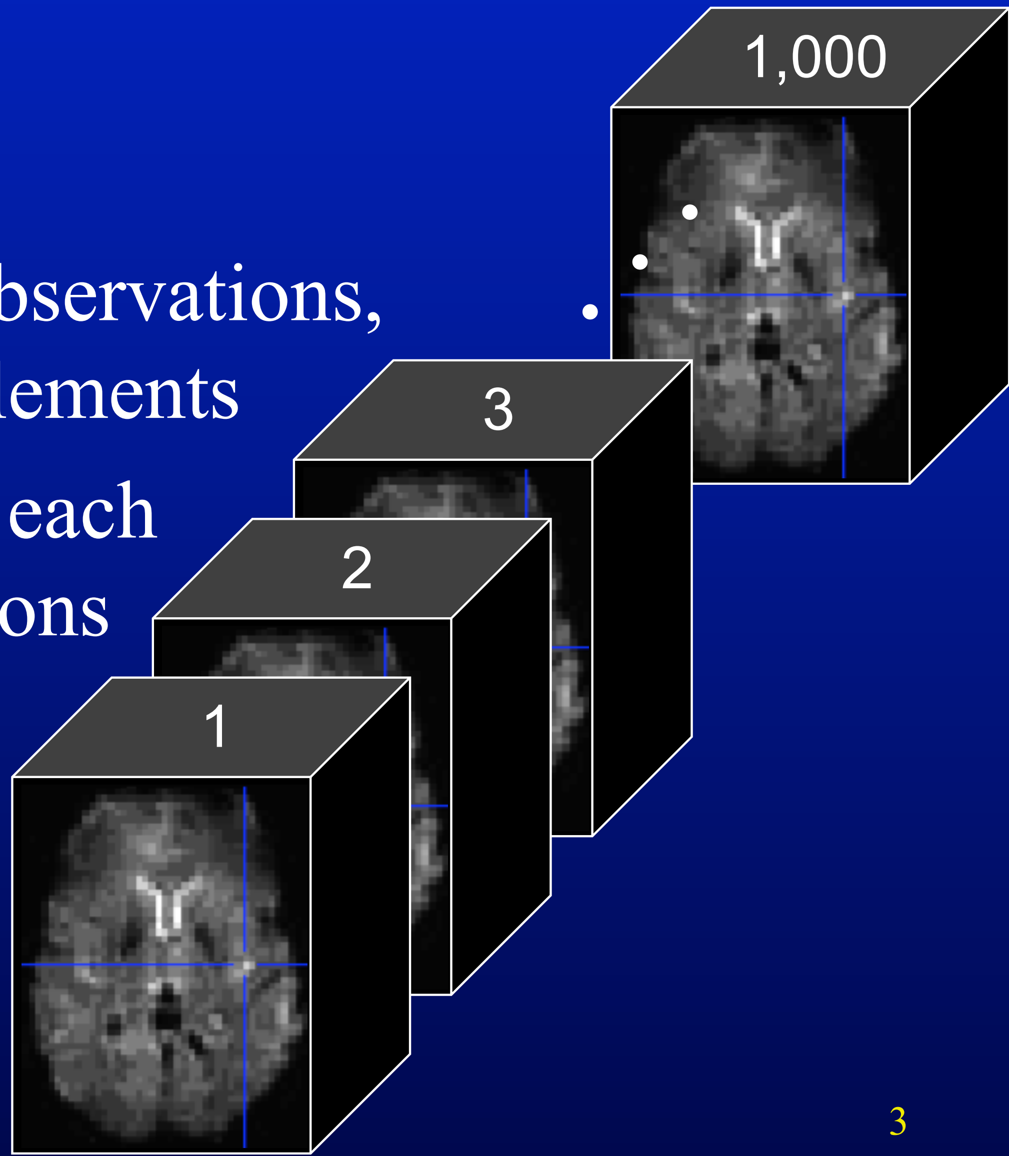
15 September, 2016

Overview

- fMRI Introduction
- Controlling MCP with FWE methods
  - Random Field Theory
  - Permutation
- Evaluations
  - Real data & simulations
- Conclusions

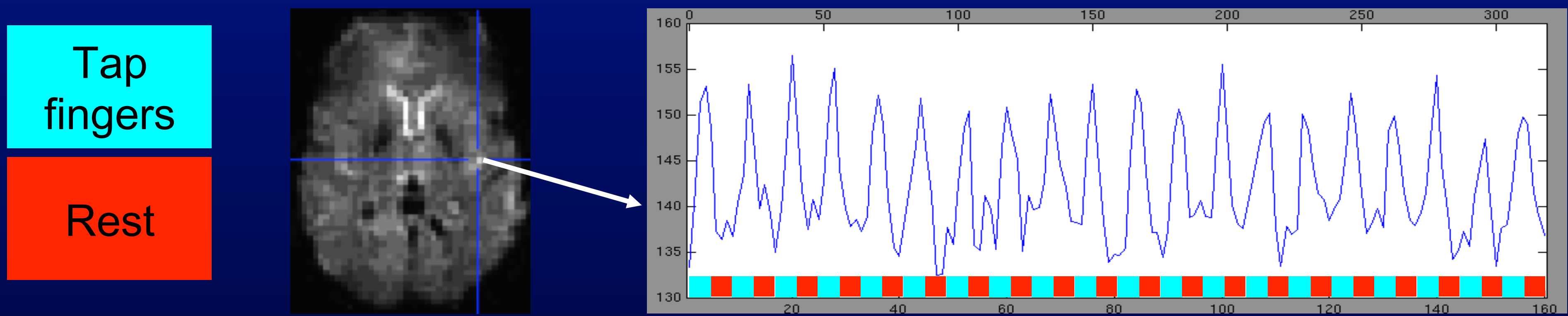
fMRI Perspective

- 4-Dimensional Data
  - 1,000 multivariate observations, each with 100,000 elements
  - 100,000 time series, each with 1,000 observations
- Usual approach is the time-series perspective

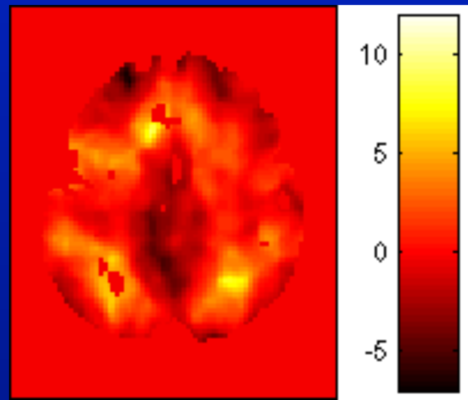


Functional Magnetic Resonance Imaging (fMRI)

- Magnetic properties of blood vary
  - Blue blood → Red blood
  - Paramagnetic → Diamagnetic
- BOLD
  - Blood Oxygenation Level Dependent effect
  - ↑ Blood flow ↑ fMRI Signal

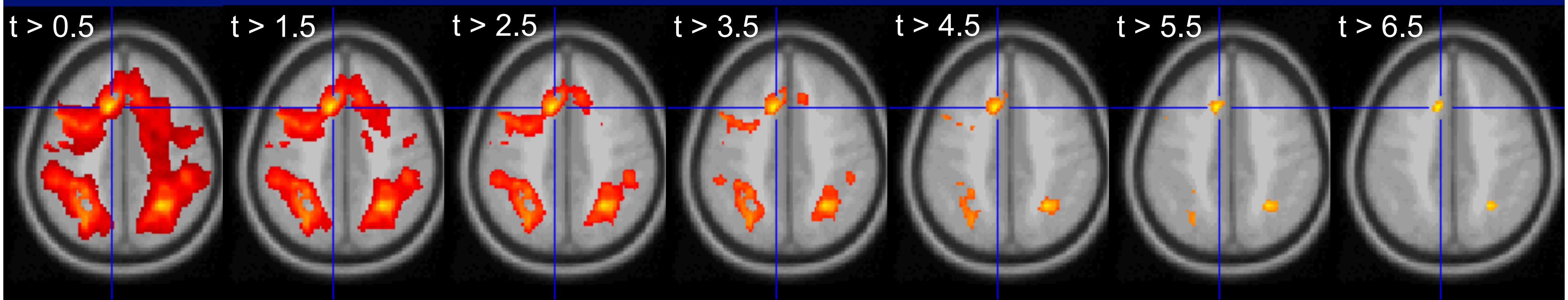


Hypothesis Testing in fMRI



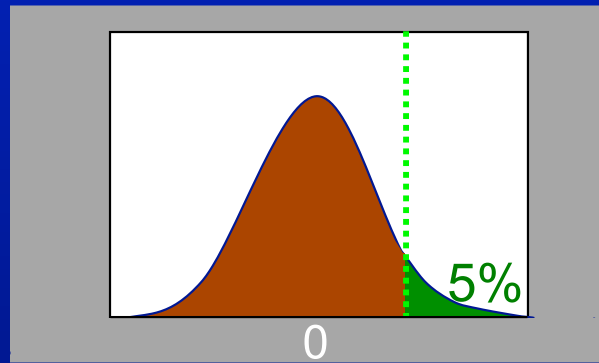
- Massively Univariate Modeling
  - Fit model at each voxel
  - Create statistic images of effect
- Which of 100,000 voxels are significant?
  - $\alpha=0.05 \Rightarrow$ 5,000 false positives!

Must we threshold?



Multiple Comparisons Problem (MCP)

- Standard Hypothesis Test
  - Controls Type I error of each test, at say 5%
  - "Type I Error" only defined for single test



- Must control false positive rate over image
  - What false positive rate?
  - Chance of 1 or more Type I errors
  - Chance of 50 or more?
  - Expected fraction of false positives?

MCP Solutions: Measuring False Positives

- Familywise Error Rate (FWER)
  - Familywise Error
    - Existence of one or more false positives
  - FWER is probability of familywise error
- False Discovery Rate (FDR)
  - R voxels declared active, V falsely so
    - Observed false discovery rate: V/R
  - $FDR = E(V/R)$

FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation

FWER MCP Solutions: Controlling FWER w/ Max

- FWER & distribution of maximum

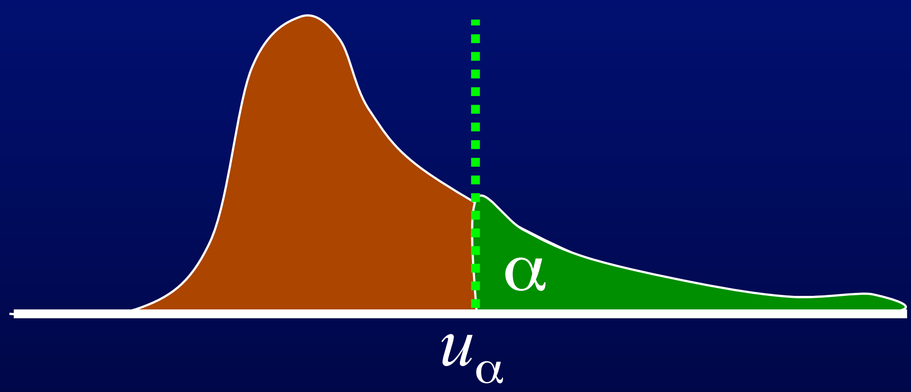
$$\begin{aligned} \text{FWER} &= \text{P(FWE)} \\ &= \text{P}(\cup_i \{T_i \geq u\} \mid H_o) \\ &= \text{P}(\max_i T_i \geq u \mid H_o) \end{aligned}$$

- 100(1-α)%ile of max dist^n controls FWER

$$\text{FWER} = \text{P}(\max_i T_i \geq u_\alpha \mid H_o) = \alpha$$

  – where

$$u_\alpha = F^{-1}_{\max} (1-\alpha)$$



FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation

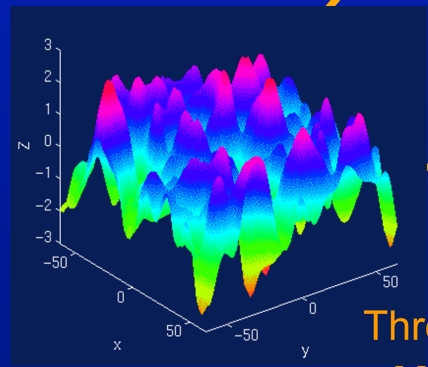
FWER MCP Solutions: Random Field Theory

- Euler Characteristic χ_u
  - Topological Measure
    - #blobs - #holes
  - At high thresholds, just counts blobs
  - FWER = P(Max voxel $\geq u \mid H_o$)

No holes

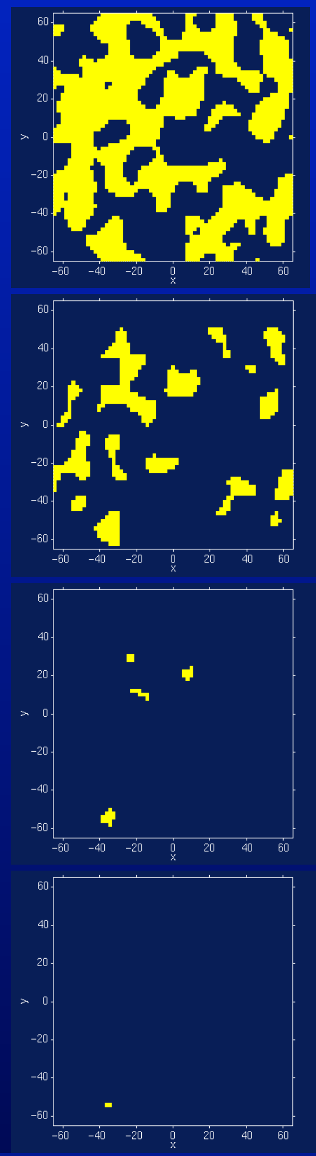
Never more than 1 blob

$$\begin{aligned} &= P(\text{One or more blobs} \mid H_o) \\ &\approx P(\chi_u \geq 1 \mid H_o) \\ &\approx E(\chi_u \mid H_o) \end{aligned}$$



Random Field

Threshold



Suprathreshold Sets

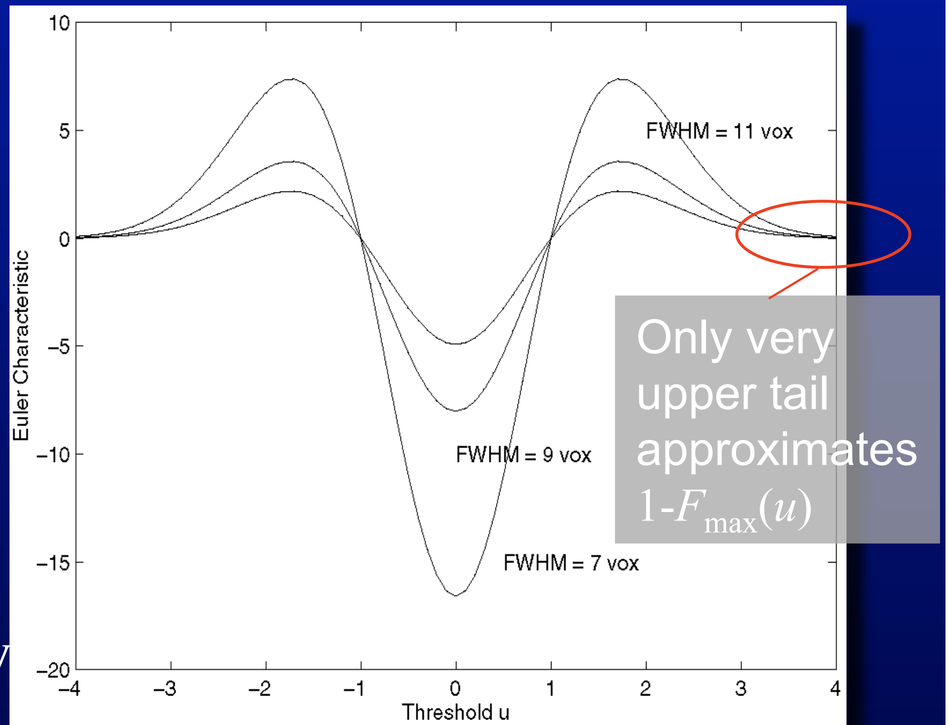
RFT Details: Expected Euler Characteristic

$$E(\chi_u) \approx \lambda(\Omega)\ |\Lambda|^{1/2}\ (u^2 - 1)\ \exp(-u^2/2)\ /\ (2\pi)^2$$

- Ω → Search region $\Omega \subset \mathcal{R}^3$
- $\lambda(\Omega)$ → volume
- $|\Lambda|^{1/2}$ → roughness

- Assumptions
  - Multivariate Normal
  - Stationary*
  - ACF twice differentiable at 0

* Stationary
  - Only cluster results need stationary
  - Most accurate when stat. holds



RFT Details: Super General Formula

- General form for expected Euler characteristic
  - χ^2, F, & t fields • restricted search regions • D dimensions •

$$E[\chi_u(\Omega)] = \Sigma_d \, R_d(\Omega) \, \rho_d(u)$$

$R_d(\Omega)$: d-dimensional Minkowski functional of Ω

– function of dimension, space Ω and smoothness:

$R_0(\Omega) = \chi(\Omega)$ Euler characteristic of Ω

$R_1(\Omega)$ = resel diameter

$R_2(\Omega)$ = resel surface area

$R_3(\Omega)$ = resel volume

$\rho_d(\Omega)$: d-dimensional EC density of $Z(\underline{x})$

– function of dimension and threshold, specific for RF type:

E.g. Gaussian RF:

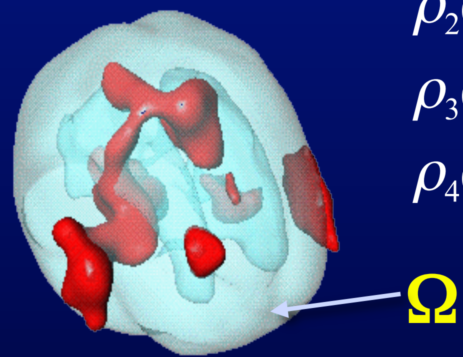
$\rho_0(u) = 1- \Phi(u)$

$\rho_1(u) = (4 \ln 2)^{1/2} \exp(-u^2/2) / (2\pi)$

$\rho_2(u) = (4 \ln 2) \quad \exp(-u^2/2) / (2\pi)^{3/2}$

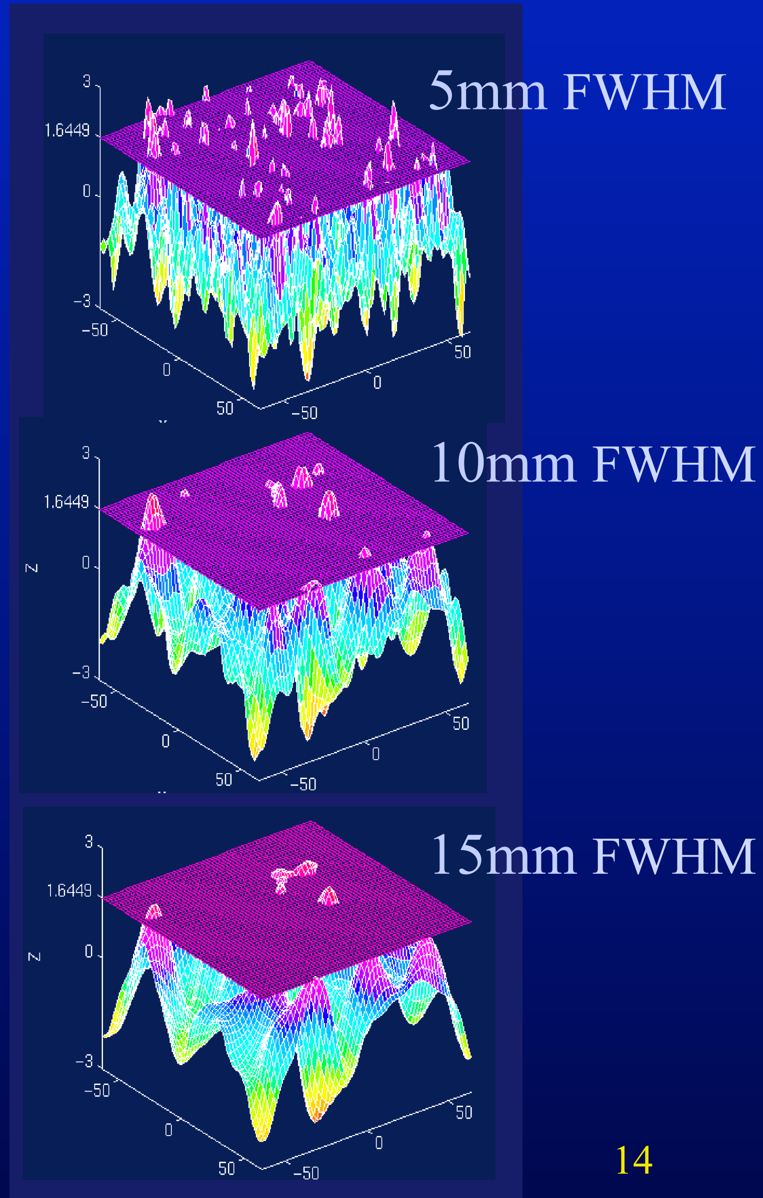
$\rho_3(u) = (4 \ln 2)^{3/2} (u^2 -1) \quad \exp(-u^2/2) / (2\pi)^2$

$\rho_4(u) = (4 \ln 2)^2 \quad (u^3 -3u) \exp(-u^2/2) / (2\pi)^{5/2}$



Random Field Theory Cluster Size Tests

- Expected Cluster Size
  - $E(S) = E(N)/E(L)$
  - S cluster size
  - N suprathreshold volume $\lambda(\{T > u_{\text{clus}}\})$
  - L number of clusters
- $E(N) = \lambda(\Omega)\ P(\ T > u_{\text{clus}}\ )$
- $E(L) \approx E(\chi_u)$
  - Assuming no holes



Random Field Theory Cluster Size Distribution

- Gaussian Random Fields (Nosko, 1969)

$$S^{2/D} \sim Exp\left(\left[\frac{E(N)}{\Gamma(D/2+1)E(L)}\right]^{-2/D}\right)$$

  - D: Dimension of RF

- t Random Fields (Cao, 1999)
  - B: Beta dist^n
  - U' s: χ^2' s
  - c chosen s.t. E(S) = E(N) / E(L)

$$S \sim cB^{1/2}\left[\frac{U_0^D}{\Pi_{b=0}^D U_b}\right]^{2/D}$$

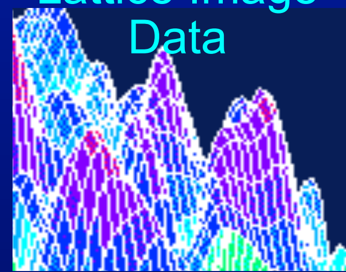
Random Field Theory
Cluster Size Corrected P-Values

- Previous results give uncorrected P-value
- Corrected P-value
  - Bonferroni
    - Correct for expected number of clusters
    - Corrected $P^c = \mathrm{E}(L)\, P^{\mathrm{uncorr}}$
  - Poisson Clumping Heuristic (Adler, 1980)
    - Corrected $P^c = 1 - \exp(-\mathrm{E}(L)\, P^{\mathrm{uncorr}})$

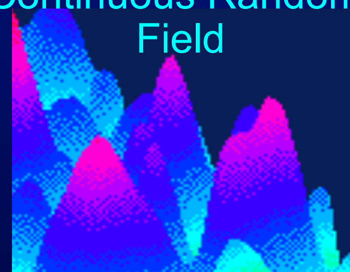
Random Field Theory Strengths & Weaknesses

- Closed form results for $E(\chi_u)$
  - Z, t, F, Chi-Squared Continuous RFs
- Results depend only on volume & smoothness

- Smoothness assumed known
- Sufficient smoothness required
  - Results are for *continuous* random fields
  - Smoothness estimate becomes biased
- Multivariate normality
- Several layers of approximations



≈

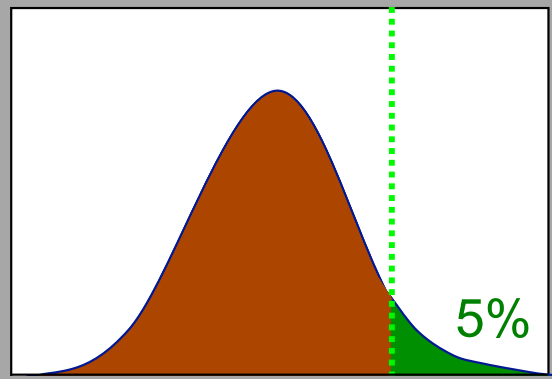


FWER MCP Solutions

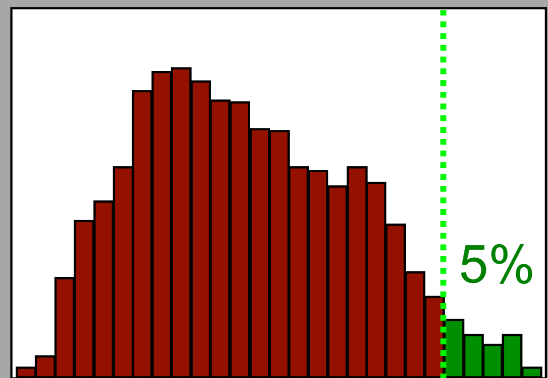
- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation

Nonparametric Permutation Test

- Parametric methods
  - Assume distribution of statistic under null hypothesis
- Nonparametric methods
  - Use *data* to find distribution of statistic under null hypothesis
  - Any statistic!



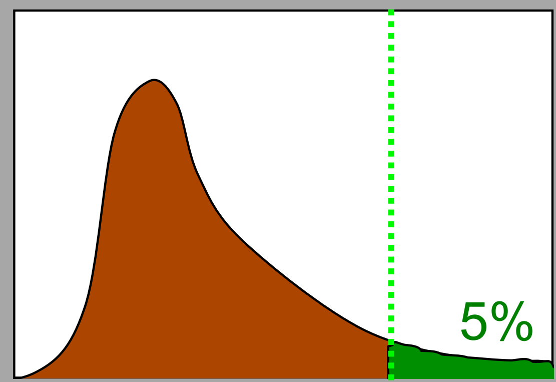
Parametric Null Distribution



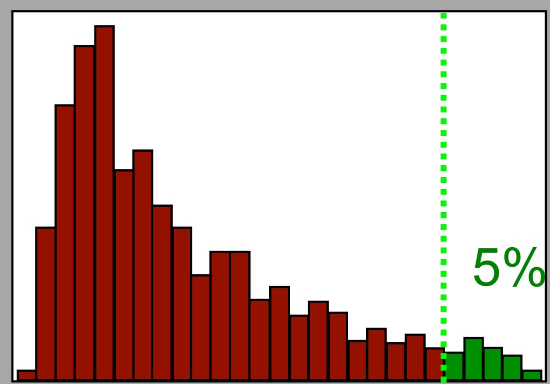
Nonparametric Null Distribution

Controlling FWER: Permutation Test

- Parametric methods
  – Assume distribution of *max* statistic under null hypothesis
- Nonparametric methods
  – Use *data* to find distribution of *max* statistic under null hypothesis
  – Again, any max statistic!



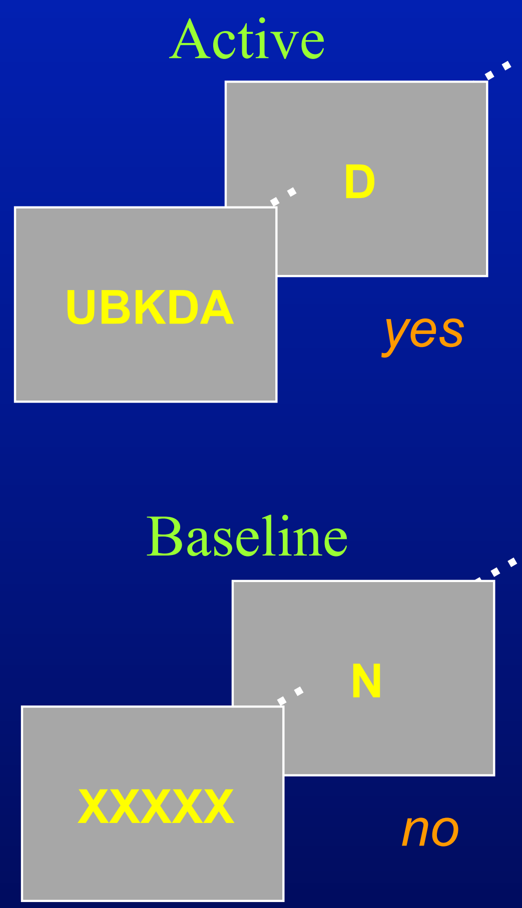
Parametric Null Max Distribution



Nonparametric Null Max Distribution

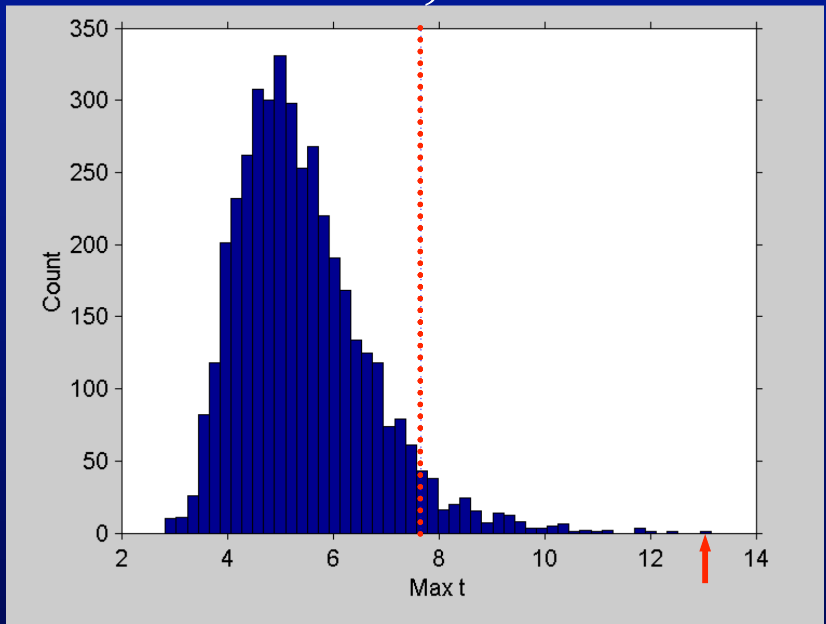
Real Data Example

- fMRI Study of Working Memory
  - 12 subjects, block design Marshuetz et al (2000)
  - Item Recognition
    - Active: View five letters, 2s pause, view probe letter, respond
    - Baseline: View XXXXX, 2s pause, view Y or N, respond
- Second Level RFX
  - Difference image, A-B constructed for each subject
  - One sample, smoothed variance t test

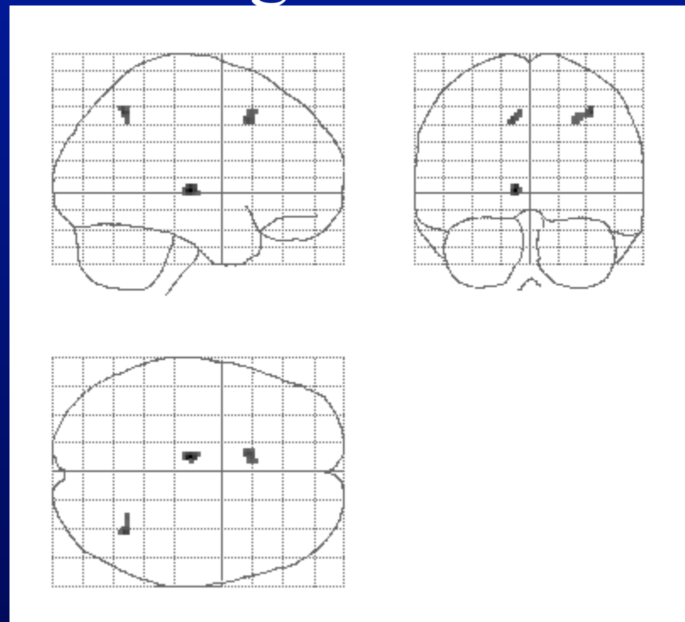


Permutation Test Example

- Permute!
  - $2^{12} = 4{,}096$ ways to flip 12 A/B labels
  - For each, note maximum of t image



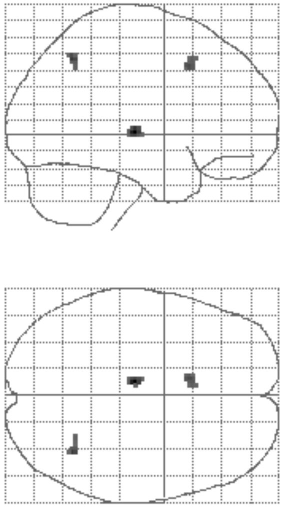
Permutation Distribution
Maximum t



Maximum Intensity Projection
Thresholded t

Permutation Test Example

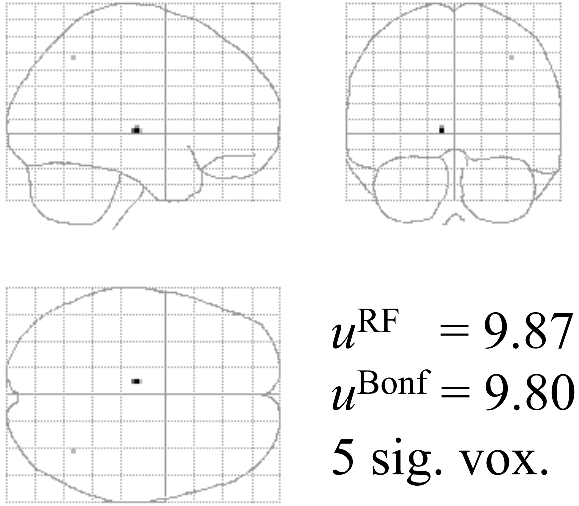
- Compare with Bonferroni
  - $\alpha = 0.05/110{,}776$
- Compare with parametric RFT
  - 110,776 $2\times2\times2$mm voxels
  - $5.1\times5.8\times6.9$mm FWHM smoothness
  - 462.9 RESELs



$u^{\mathrm{Perm}} = 7.67$

58 sig. vox.

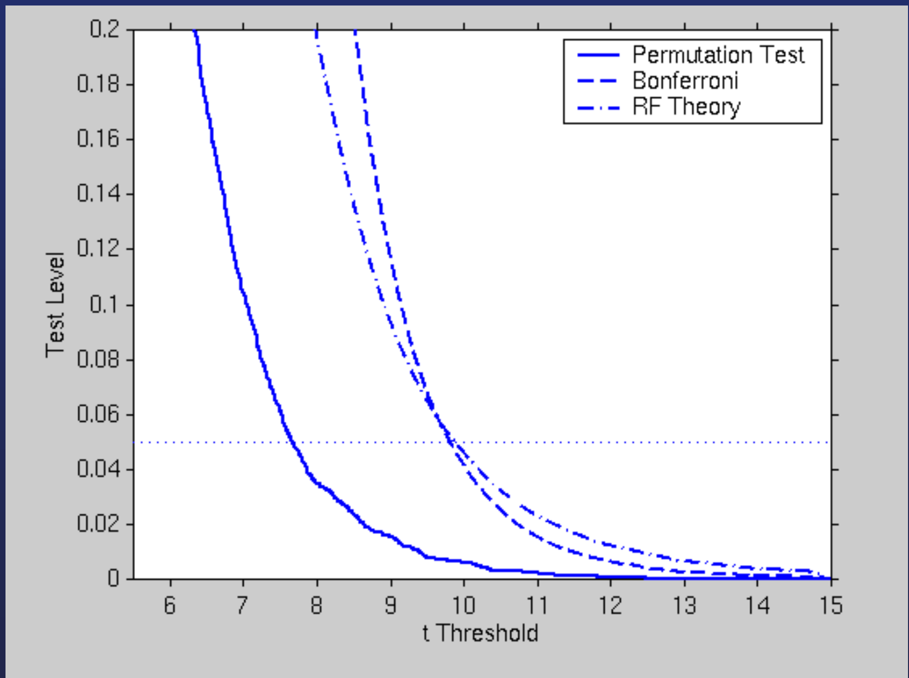
t_{11} Statistic, Nonparametric Threshold



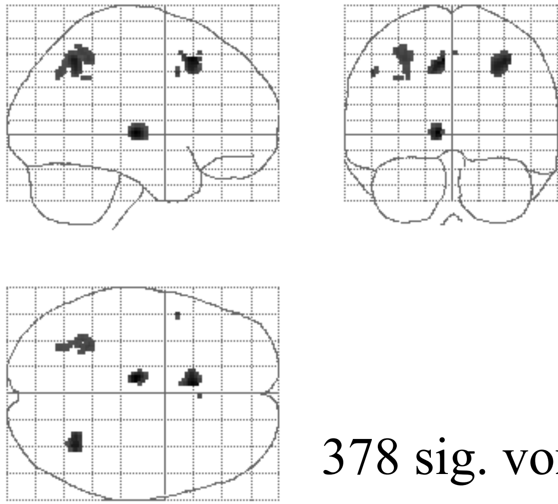
$u^{\mathrm{RF}} = 9.87$
$u^{\mathrm{Bonf}} = 9.80$

5 sig. vox.

t_{11} Statistic, RF & Bonf. Threshold



Test Level vs. t_{11} Threshold



378 sig. vox.

Smoothed Variance *t* Statistic, Nonparametric Threshold

Does this Generalize? RFT vs Bonf. vs Perm.

		t Threshold (0.05 Corrected)		
	df	RF	Bonf	Perm
Verbal Fluency	4	4701.32	42.59	10.14
Location Switching	9	11.17	9.07	5.83
Task Switching	9	10.79	10.35	5.10
Faces: Main Effect	11	10.43	9.07	7.92
Faces: Interaction	11	10.70	9.07	8.26
Item Recognition	11	9.87	9.80	7.67
Visual Motion	11	11.07	8.92	8.40
Emotional Pictures	12	8.48	8.41	7.15
Pain: Warn ing	22	5.93	6.05	4.99
Pain: Anticipation	22	5.87	6.05	5.05

Massive Empirical Evaluation

- Monte Carlo doesn’t capture weirdness of real data
- In last 5 years, explosion of open resting fMRI data repositories
  - Suddenly null (task) fMRI data is plentiful

FUNCTIONAL CONNECTOMES PROJECT

International Neuroimaging Data-Sharing Initiative

First-Level (single subject) fMRI

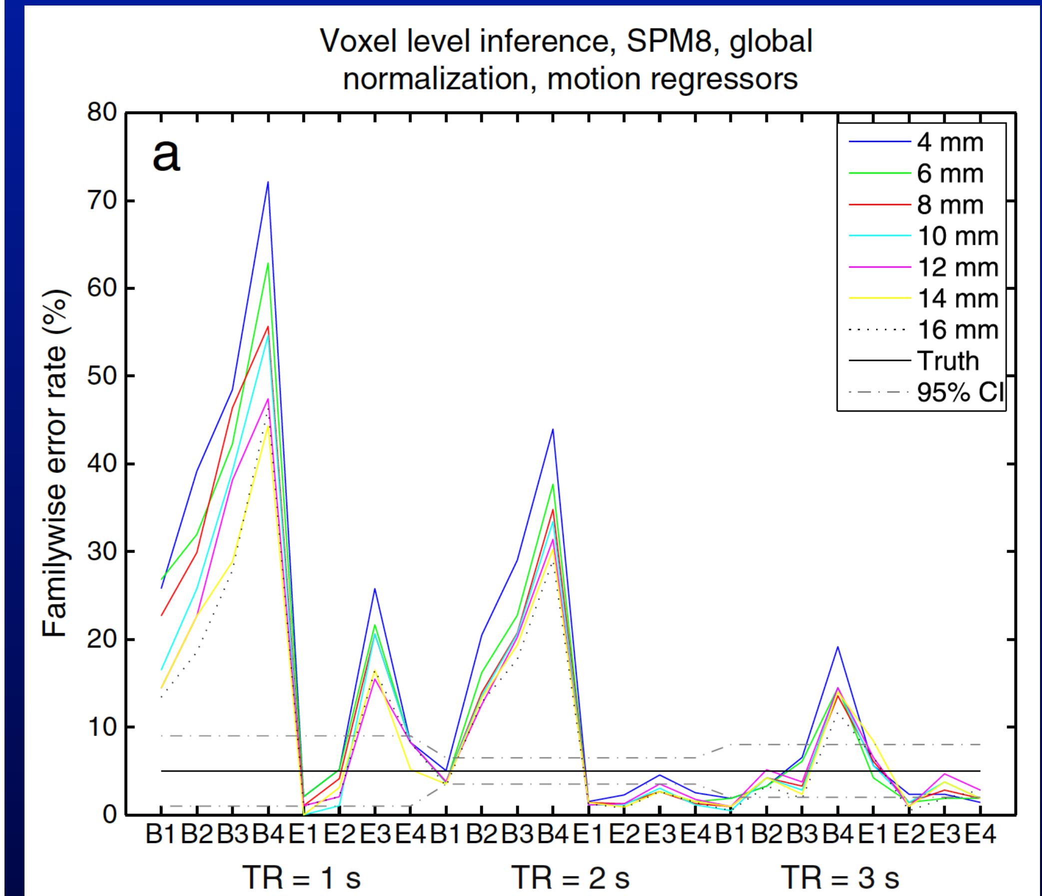
- Eklund (2012) analyzed **1,484** resting fMRI datasets from public repositories
- Fed through standard SPM pipeline, with 8 different "pretend" paradigms

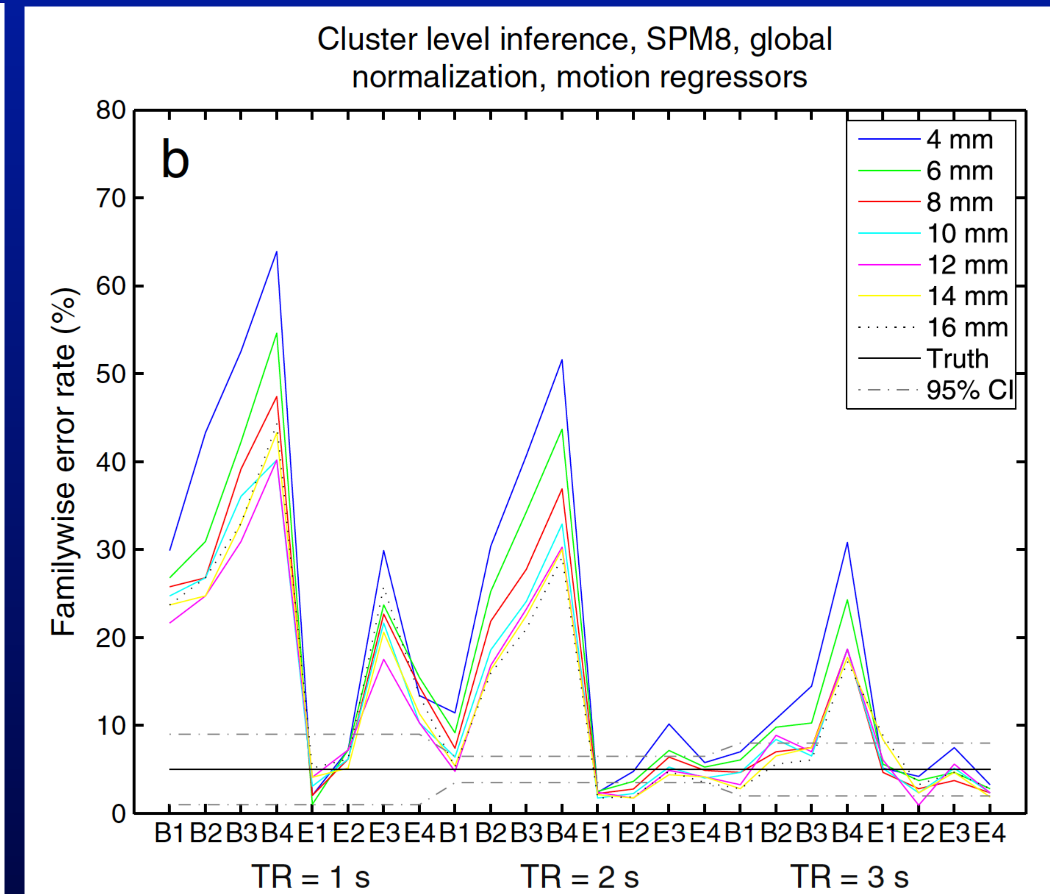
Paradigm	Activity periods (s)	Rest periods (s)
B1	10	10
B2	15	15
B3	20	20
B4	30	30
E1	2	6
E2	4	8
E3	1–4 (R)	3–6 (R)
E4	3–6 (R)	4–8 (R)

Eklund et al. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, 61(3), 565–78.

Computed Familywise Error (FWE) Rates

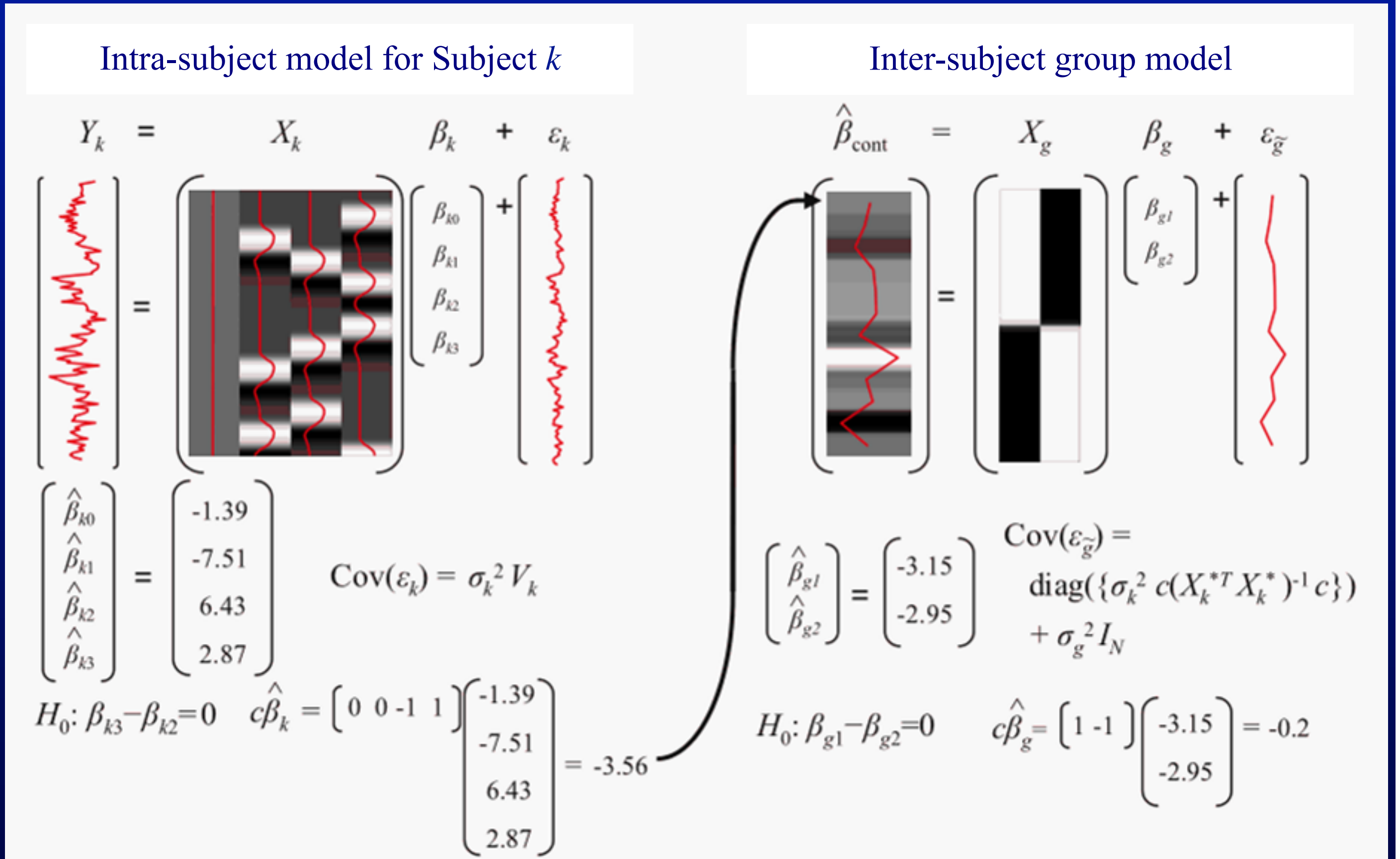
- Many settings had awful FWE!
  - Block worse than event; fast TR worse that slow





Massive Empirical Evaluation – Take II

- Previous result only for first level fMRI
- 2nd level fMRI doesn't depend on 1st level P-values
- Data quality also an issue



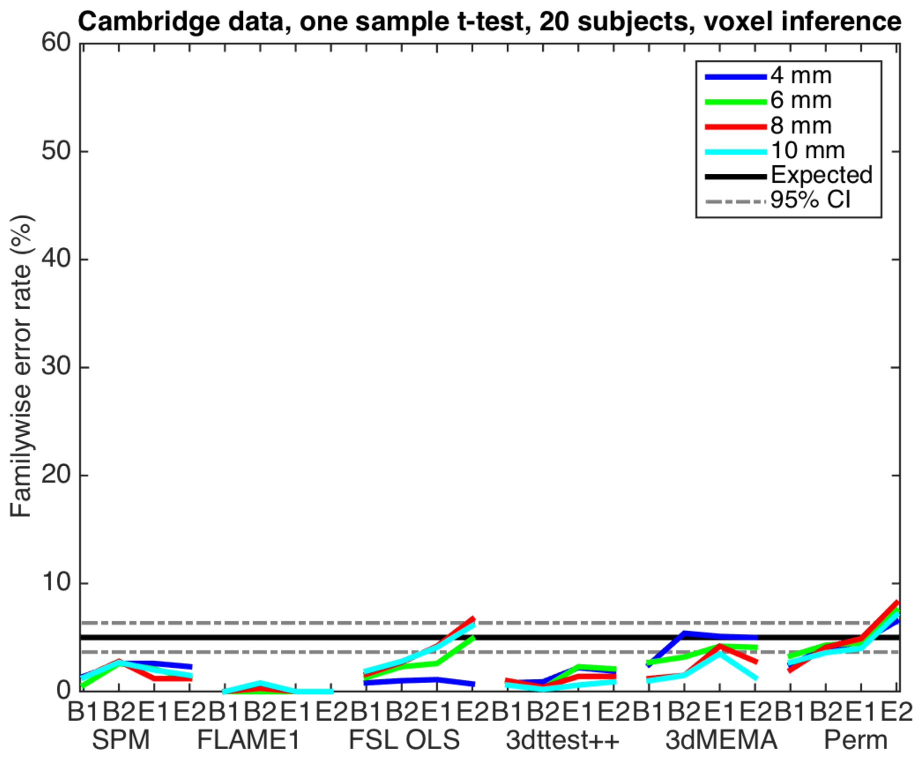
Massive Empirical Evaluation – Take II

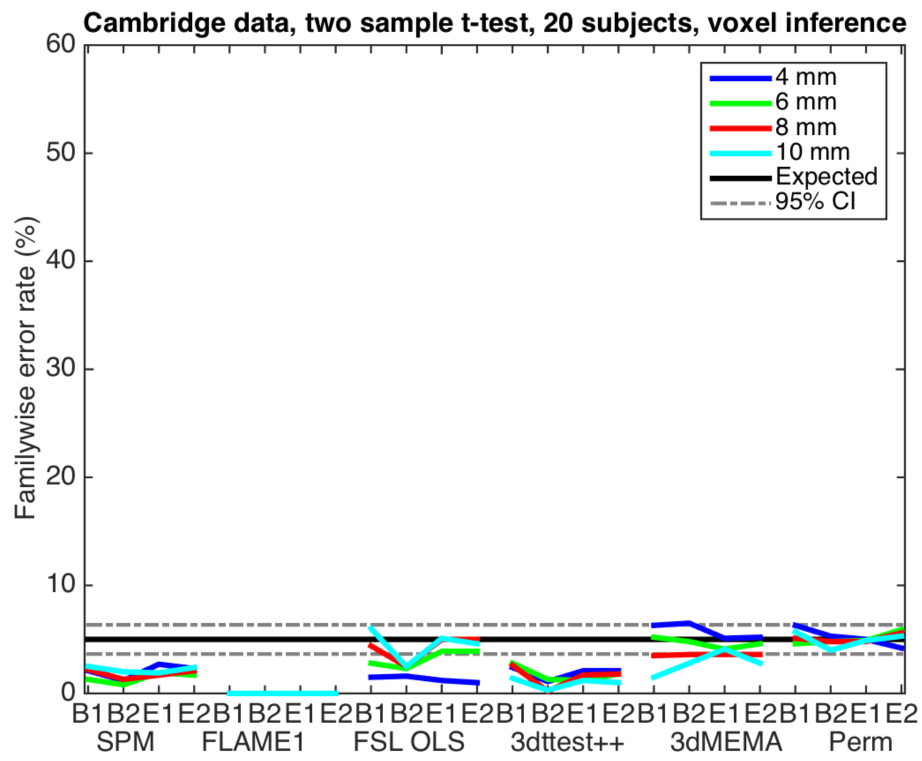
- Same fcon1000 repository, just 2 largest sites: Beijing & Cambridge
- Second level analyses
  - 1-sample t-test: $n = 20, 40$
  - 2-sample t-test: $n_1 = n_2 = 10, 20$

Parameter	Values used
fMRI data	Beijing (198 subjects), Cambridge (198 subjects)
Block activity paradigms	B1 (10 s on off), B2 (30 s on off)
Event activity paradigms	E1 (2 s activation, 6 s rest), E2 (1 - 4 s activation, 3 - 6 s rest, randomized)
Smoothing	4, 6, 8, 10 mm FWHM
Analysis type	One sample t-test (group activation), two sample t-test (group difference)
Number of subjects	20, 40
Inference level	Voxel, cluster
Cluster defining threshold	$p = 0.01$ ($z = 2.3$), $p = 0.001$ ($z = 3.1$)

Massive Group fMRI Evaluation Voxel-wise

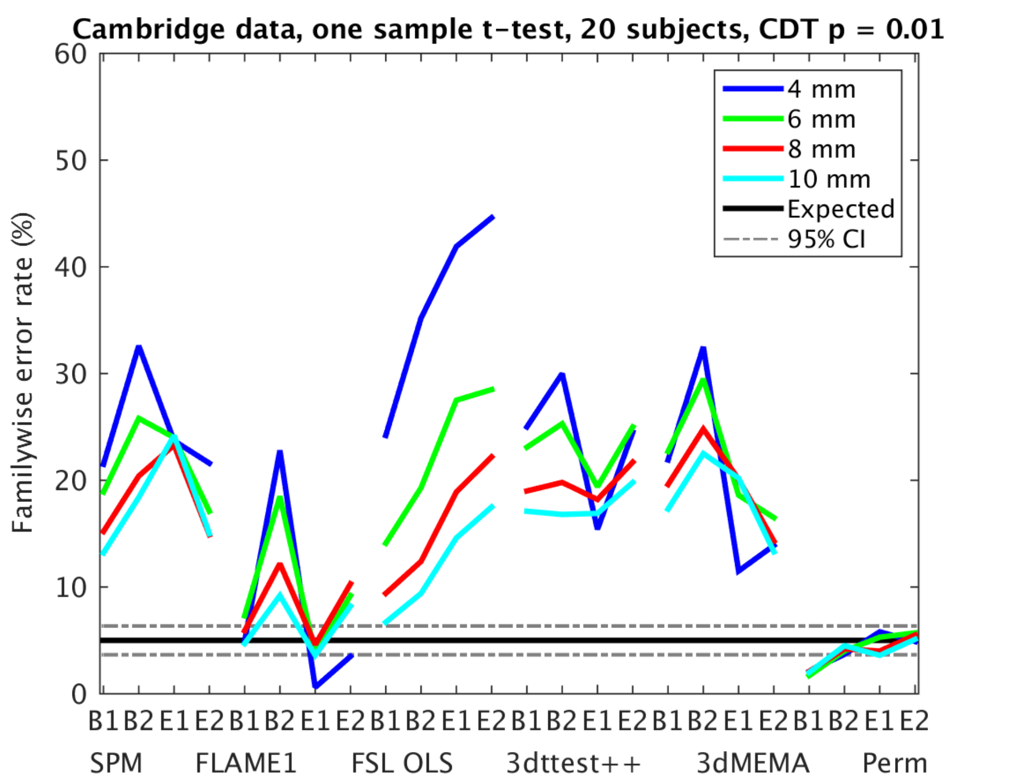
- Voxel-wise inference OK
  - Sometimes very conservative!

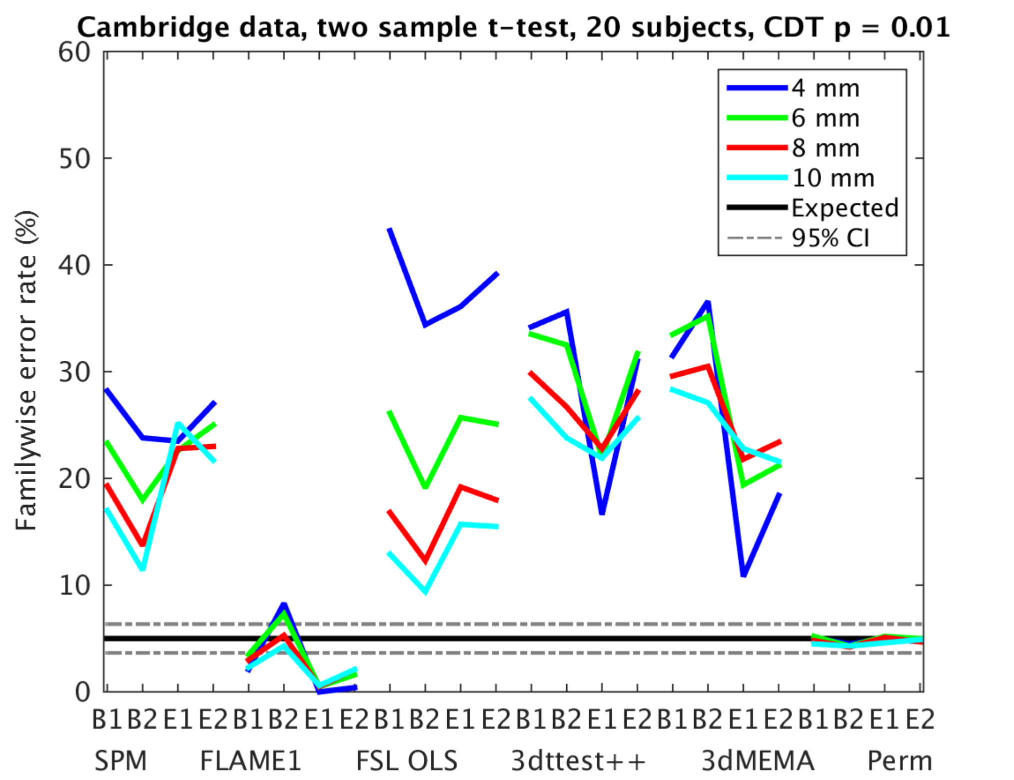




Massive Group fMRI Evaluation
Cluster-wise CFT p=0.01

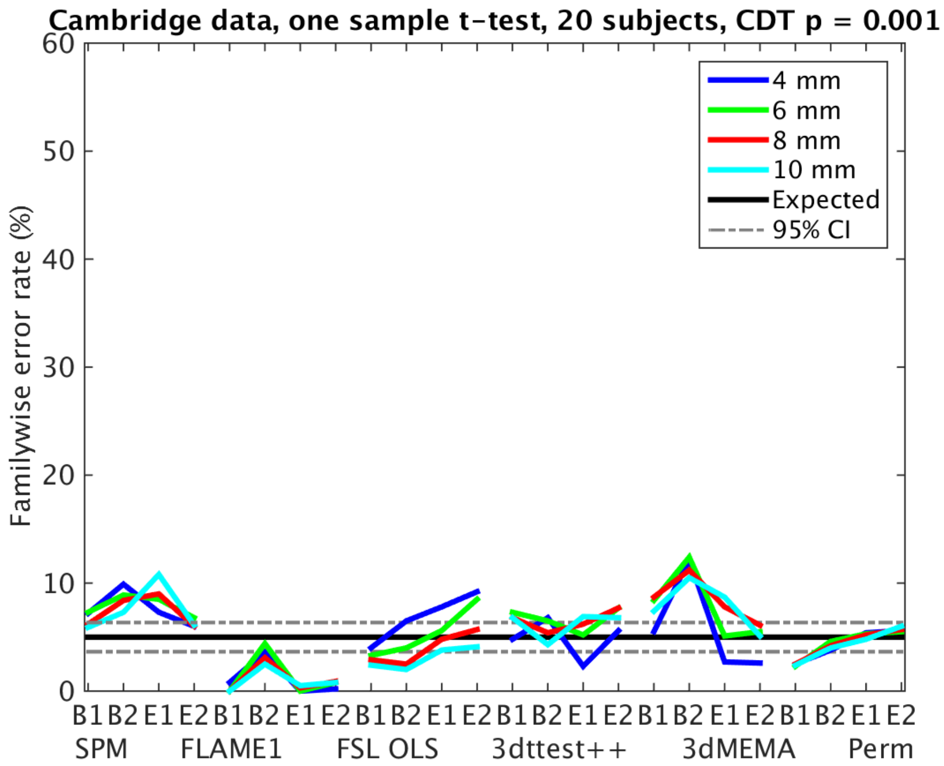
- Cluster-wise a catastrophe!
  - Rarely valid at cluster forming threshold (CFT) p=0.01 – default CFT in FSL

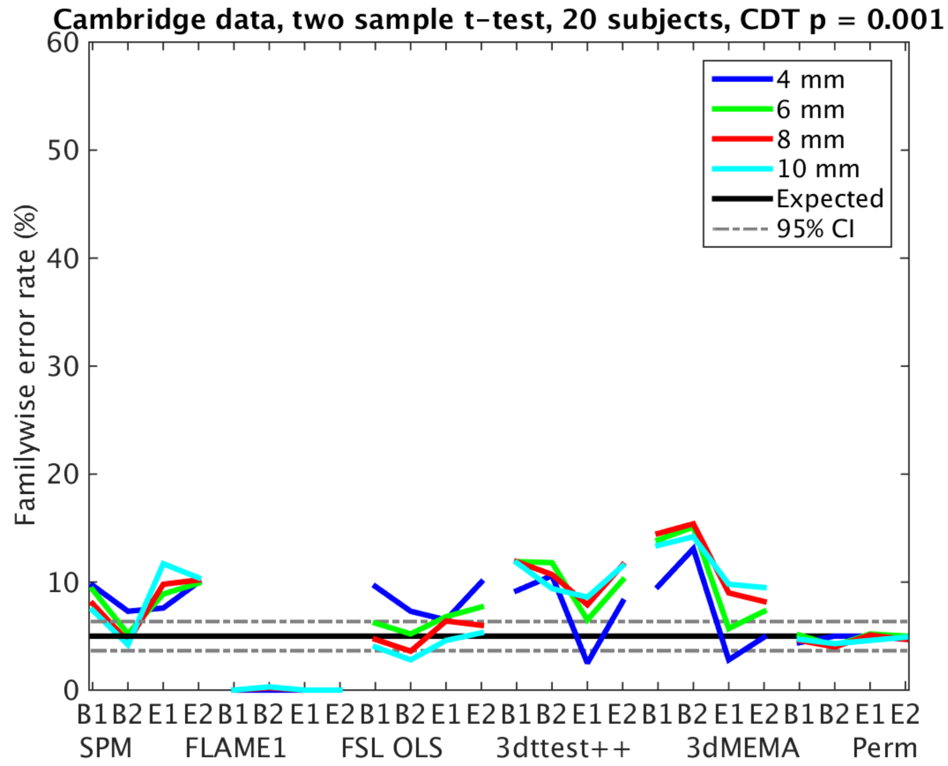




Massive Group fMRI Evaluation Cluster-wise CFT p=0.001

- Cluster-wise CFT p=0.001 better
  - Valid ≈ 50% time, depending on design



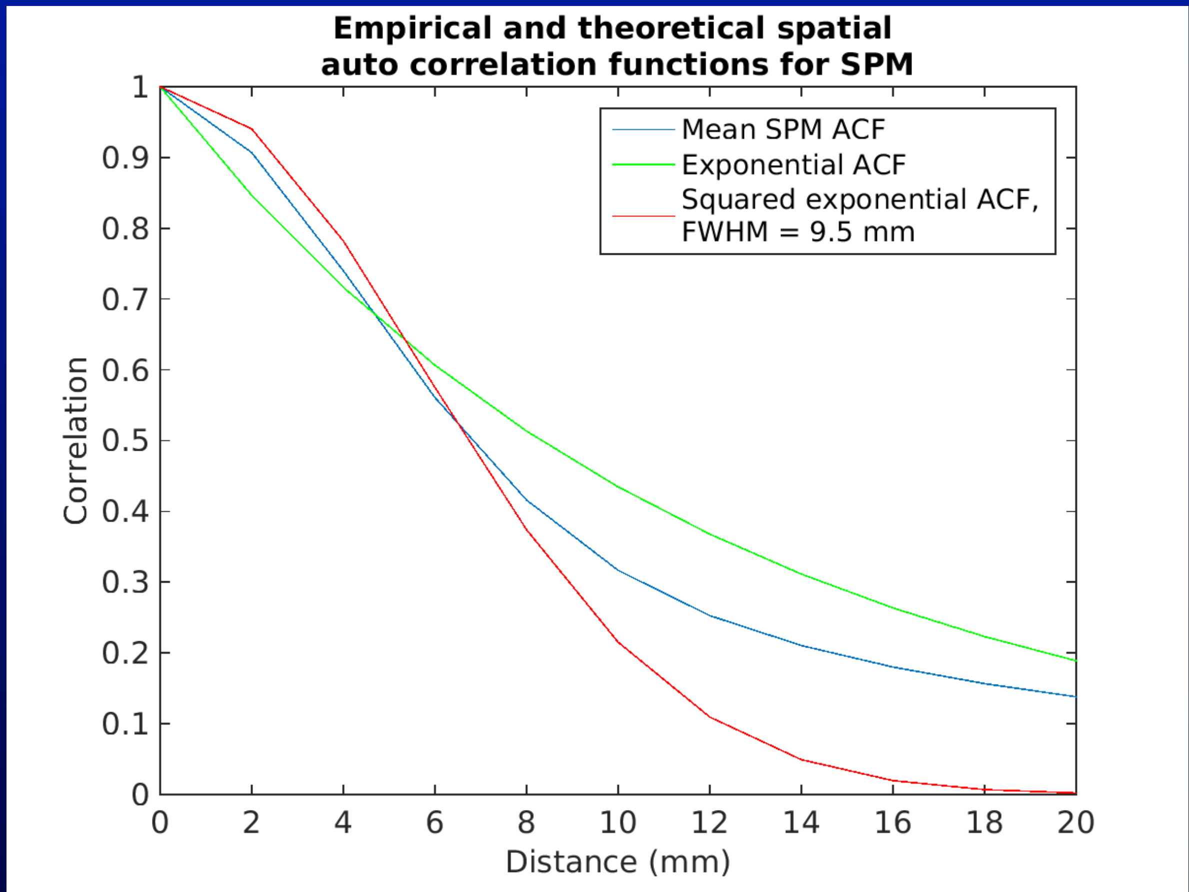


Massive Group fMRI Eval: What’s going wrong?

- RFT Assumptions
  - Gaussian errors
  - Spatial ACF has 2 derivatives at origin
  - For cluster-size only
    - Spatial ACF has Gaussian shape
    - CFT “sufficient” high
    - Stationary (spatially homogeneous smoothness)

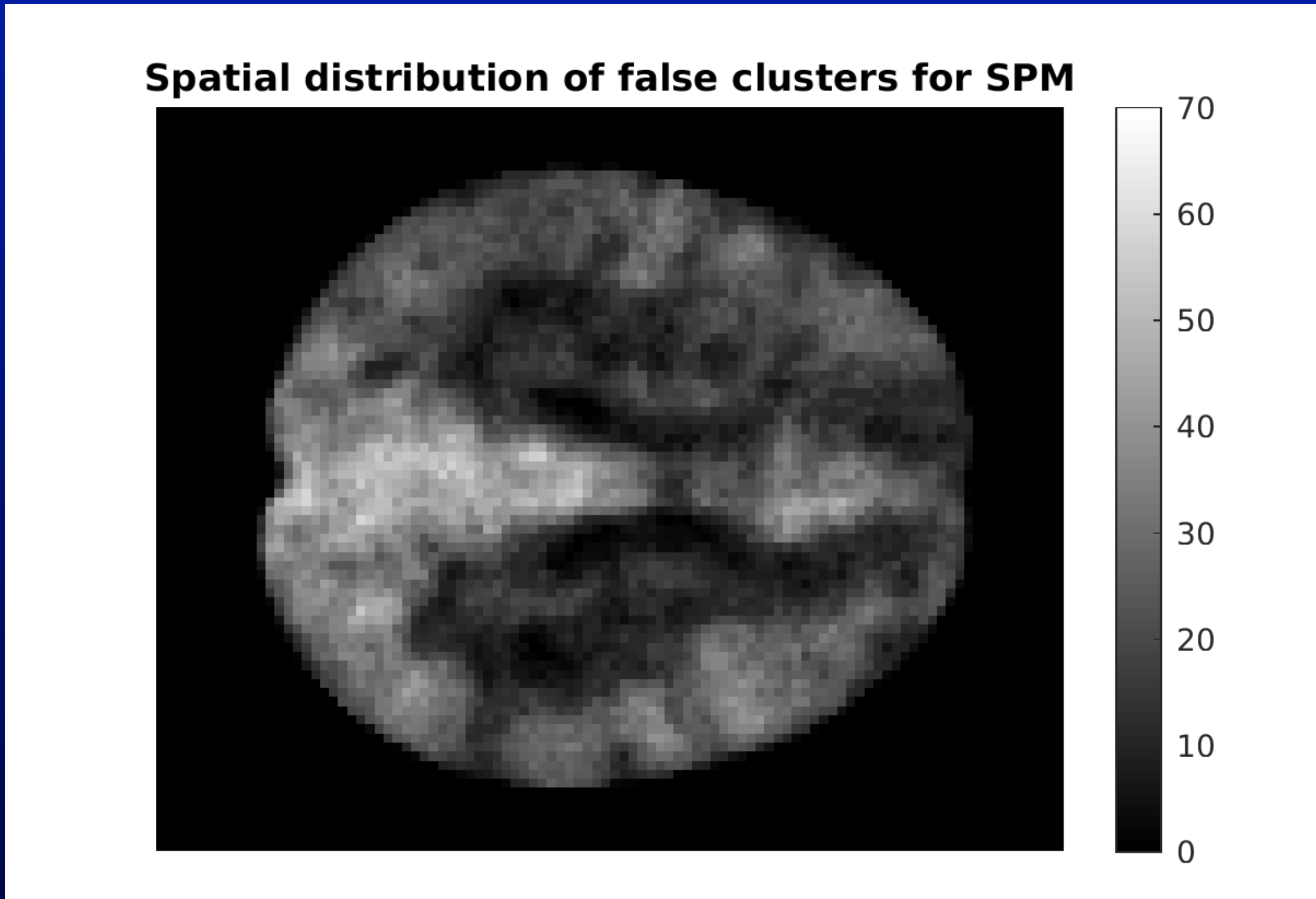
Massive Group fMRI Eval: Spatial ACF

- Much heavier tails than Gaussian pdf!



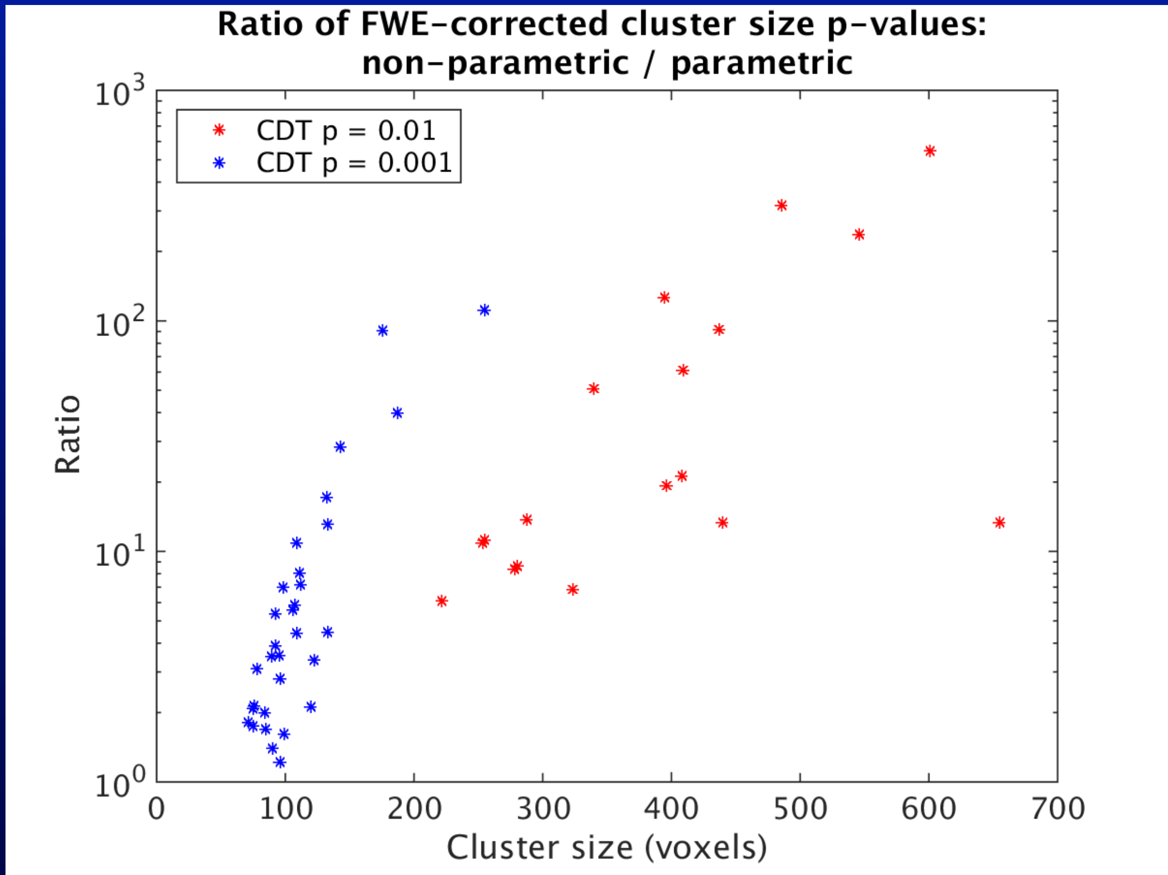
Massive Group fMRI Eval: Spatial Dist^n of False Clusters

- Great smoothness in “default mode” areas



What always works? Permutation!

- How does this compare on real (non-null) data?

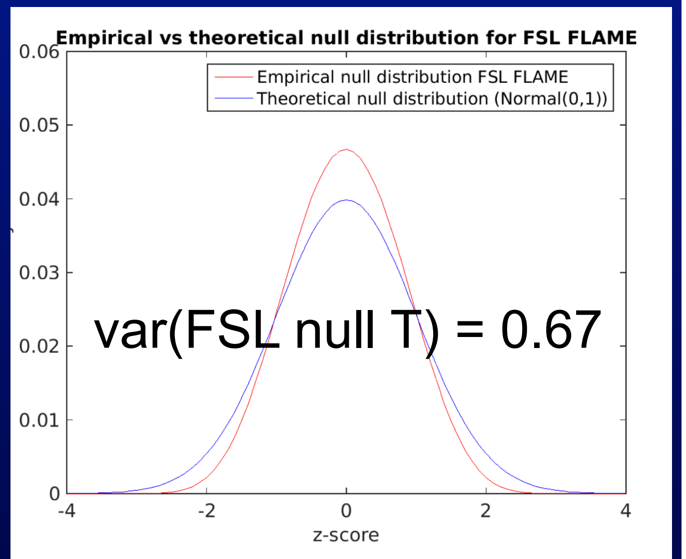


Usually, would say "non-parametric so much less powerful"

In light of evaluations, "non-parametric valid, parametric inflated signficance"

Other Findings

- AFNI software
  - Discovered 15 year-old bug
    - Failure to account for edge effects in MC simulation of smooth images
  - Inflated FWE slightly
    - CDT P=0.01: 31.0% before fix, 27.1% after
    - CDT P=0.001: 11.5% before, 8.6% after
- FSL software
  - When no effect, overestimates SE's, counteracts liberal RFT performance
  - But when $\sigma_{BTW}>0$ but null true, same bad performance
    - E.g. two-sample t-test; $\mu_1=\mu_2>0$



Proceedings of the National Academy of Sciences of the United States of America

> Current Issue > vol. 113 no. 28 > Anders Eklund, 7900–7905, doi: 10.1073/pnas.1602413113

CrossMark click for updates

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

Online Impact

1788

See more details

- Picked up by **41** news outlets
- Blogged by **38**
- Tweeted by **1522**
- On **49** Facebook pages
- Mentioned in **23** Google+ posts
- Reddited by **6**
- Mentioned in **1** Q&A threads

- **1042** readers on Mendeley
- **16** readers on CiteULike

This Altmetric score means that the article is:

in the 99 percentile … of a similar age in all journals

in the 99 percentile (ranked 2) of … a similar age in PNAS

Google

fmri 40000

All Images Videos News Shopping More Search tools

About 47,200 results (0.34 seconds)

A bug in fMRI software could invalidate 15 years of brain rese...
www.sciencealert.com/a-bug-in-**fmri**-software-could-invalidate-decades-...
6 Jul 2016 - "These results question the validity of some **40,000 fMRI** studies and may have a large impact on the interpretation of neuroimaging results," the ...

Cluster failure: Why fMRI inferences for spatial extent have inf...
www.pnas.org/content/113/28/7900.full
by A Eklund - 2016 - Cited by 21
12 Jul 2016 - Functional MRI (**fMRI**) is 25 years old, yet surprisingly its most common ... brain, with some **40,000** published papers according to PubMed.

fMRI software bug could invalidate 15 years of brain scans | ...
www.wired.co.uk/article/**fmri**-bug-brain-scans-results
6 Jul 2016 - "It is not feasible to redo **40,000 fMRI** studies, and lamentable archiving and data-sharing practices mean most could not be reanalysed either," ...

False-Positive fMRI Hits The Mainstream - Neuroskeptic
blogs.discovermagazine.com/.../2016/07/.../false-positive-**fmri**-mainstrea...
7 Jul 2016 - The article, called Cluster failure: Why **fMRI** inferences for spatial extent ..... These results question the validity of some **40,000 fMRI** studies and ...

fMRI bugs could upend years of research • The Register
www.theregister.co.uk/.../mri_software_bugs_could_upend_years_of_re...

Correction for Eklund et al., Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Extract | Full Text | Authors & Info | Metrics | Related Content | PDF

NEUROSCIENCE, STATISTICS Correction for "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," by Anders Eklund, Thomas E. Nichols, and Hans Knutsson, which appeared in issue 28, July 12, 2016, of *Proc Natl Acad Sci USA* (113:7900–7905; first published June 28, 2016; 10.1073/pnas.1602413113).

The authors note that on page 7900, in the Significance Statement, lines 9–11, "These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results" should instead appear as "These results question the validity of a number of fMRI studies and may have a large impact on the interpretation of weakly significant neuroimaging results."

Additionally, the authors note that on page 7904, left column, fifth full paragraph, lines 1–3, "It is not feasible to redo 40,000 fMRI studies, and lamentable archiving and data-sharing practices mean most could not be reanalyzed either" should instead appear as "Due to lamentable archiving and data-sharing practices, it is unlikely that problematic analyses can be redone."

These errors do not affect the conclusions of the article. The online version has been corrected.

Conclusions

- Gaussian Monte Carlo results only go so far
- Real data evaluations
  - RFT Voxel-wise OK, but conservative
  - Cluster-wise P=0.01 invalid **danger danger danger**
  - Cluster-wise P=0.001 – sometimes OK, sometimes invalid
- Permutation embarrassingly parallelizable, GPU friendly
- Pre-print publication (on arXiv) is the way
  - Received voluminous feedback that improved paper, much instigated as Twitter conversations
- When publishing in PNAS, think carefully about non-technical readers

References

- Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research, 12(5), 419–446.
- Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. NeuroImage, 20(4), 2343–56. doi: j.neuroimage.2003.08.003
- Eklund, A., Nichols, T., & Knutsson, H. (2015). Can parametric statistical methods be trusted for fMRI based group studies? http://arxiv.org/abs/1511.01863
- Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. PNAS, 201602413. http://doi.org/10.1073/pnas.1602413113