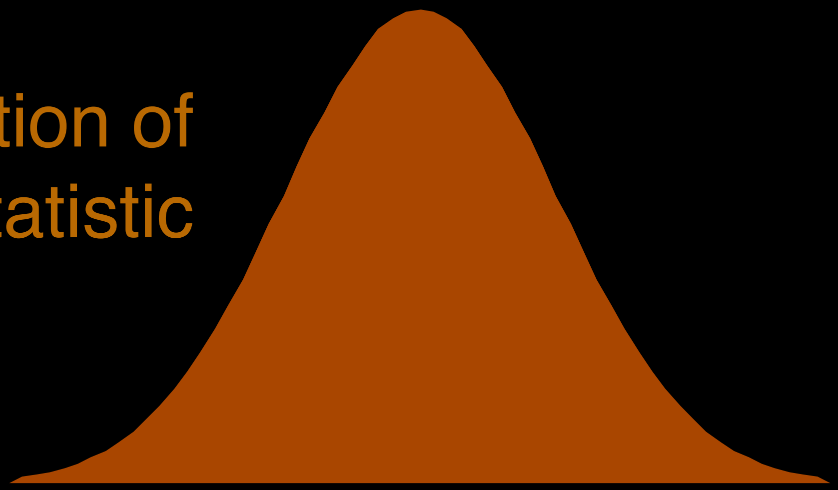
Multiple Comparison Correction

Typical Statistical Inference

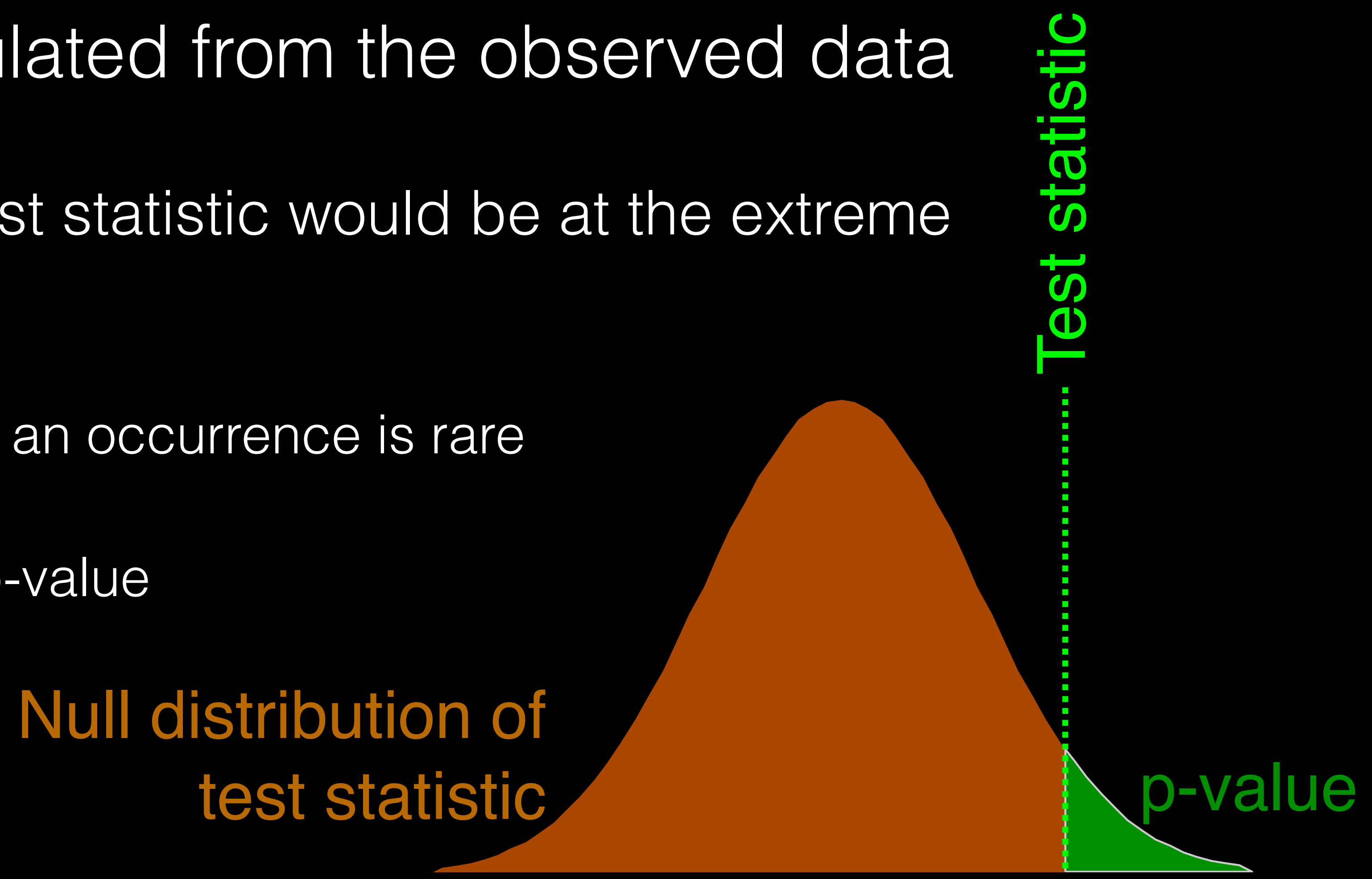
- Hypothesis to be tested
    - E.g., H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 > \mu_2$
- A test statistic is calculated from the observed data
    - If H_0 is true, then the test statistic should follow the null distribution

Null distribution of
test statistic



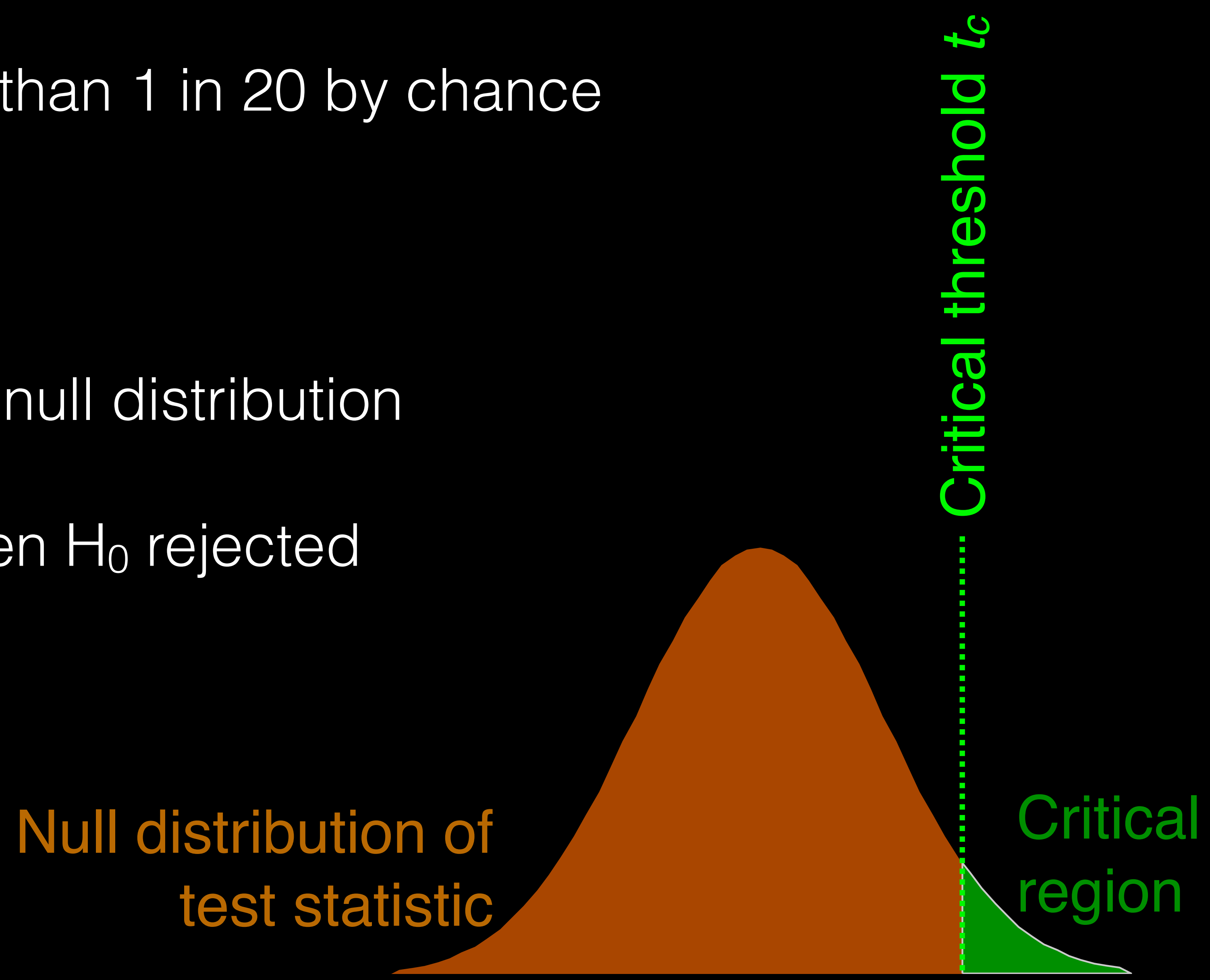
Typical Statistical Inference

- Hypothesis to be tested
  - E.g., H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 > \mu_2$
- A test statistic is calculated from the observed data
  - If H_A is true, then the test statistic would be at the extreme tail of the distribution
    - The probability of such an occurrence is rare
    - Quantified by a small p-value



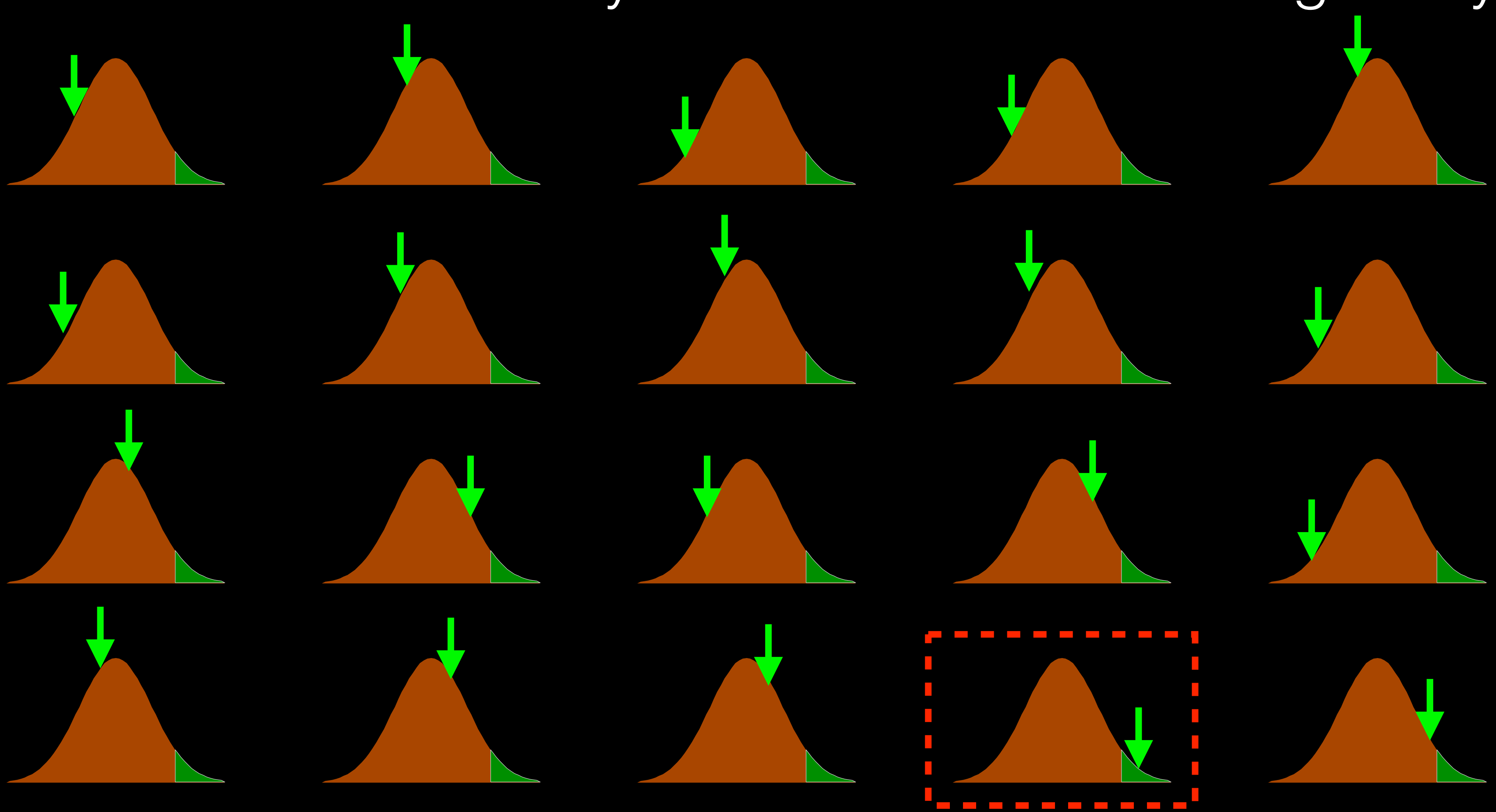
Typical Statistical Inference

- A small p-value — evidence for rejection of H_0
    - Often $p<0.05$ or, less than 1 in 20 by chance
- Critical threshold t_c
    - 95th percentile of the null distribution
    - If test statistic > t_c, then H_0 rejected



Multiple Comparisons

- A large number of statistical tests simultaneously
  - A large number of test statistics
  - Some of them likely fall inside the critical region by chance alone

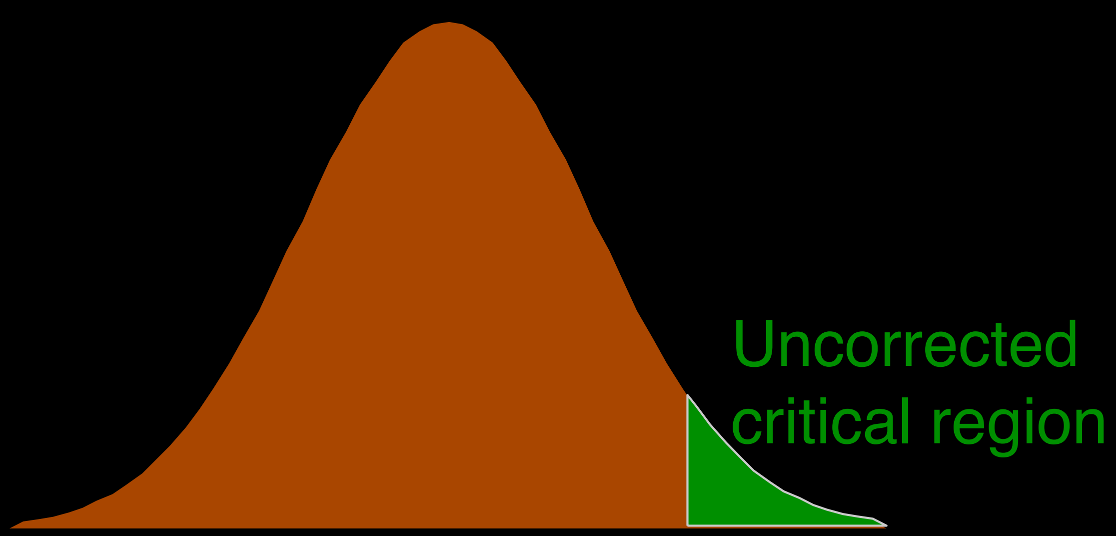


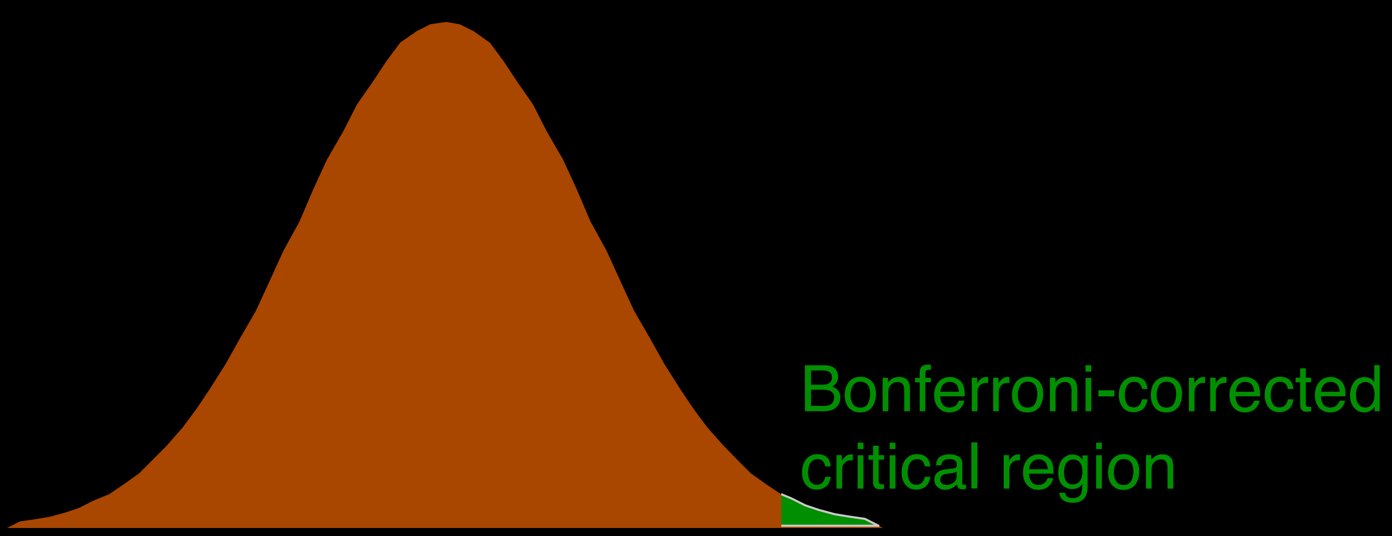
Multiple Comparisons

- In fMRI data analysis,
  - Separate statistical test at each voxel
  - 20,000 voxels ➔ 20,000 T-tests
  - Significant effect at $p<0.05$ level
    - 20,000 x 0.05 = 100 voxels show significant effect by chance alone
    - Even without any true effect

Multiple Comparison Correction

- Traditional multiple comparison correction — Bonferroni
  - Shrink the critical region
  - Critical p-value divided by the number of tests
  - E.g., Critical p= 0.05, 20 tests ➔ Bonferroni critical p= 0.05/20 = 0.0025





Multiple Comparison Correction

- Bonferroni on fMRI data analysis
  - 20,000 voxels ➔ 20,000 simultaneous T-tests
  - Bonferroni corrected $p<0.05$
    - $p<0.05$ / 20,000 ➔ $p<0.0000025$
    - Perhaps too stringent?

Multiple Comparison Correction

- A Bonferroni correction assumes statistical tests to be independent

- In fMRI analysis

  - Neighboring voxel values are often correlated

  - *Multiple comparison methods specialized for neuroimaging data*

Multiple Comparison Correction, Neuroimaging Style

Typical verbiage:

"*The statistical analysis is corrected for multiple comparisons controlling the _______ at _______ level.*"

Multiple Comparison Correction, Neuroimaging Style

Typical verbiage:

"*The statistical analysis is corrected for multiple comparisons controlling the ________ at ________ level.*"

(Correction method)

- FWE — Family-wise error
- FDR — False discovery rate

Multiple Comparison Correction, Neuroimaging Style

Typical verbiage:

"The statistical analysis is corrected for multiple comparisons controlling the ________ at ________ level."

(Correction method)

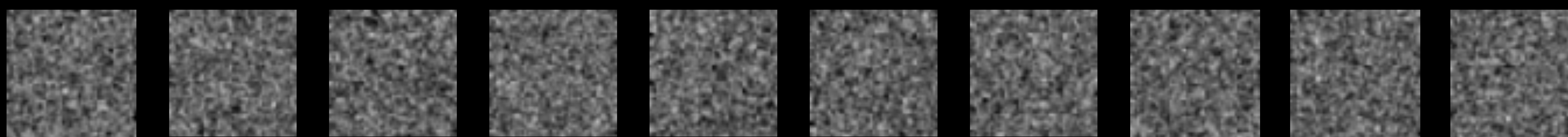
- FWE — Family-wise error
- FDR — False discovery rate

(Significance level)

- $p = 0.05$ (FWE-correction)
- $q = 0.05$ (FDR-correction)

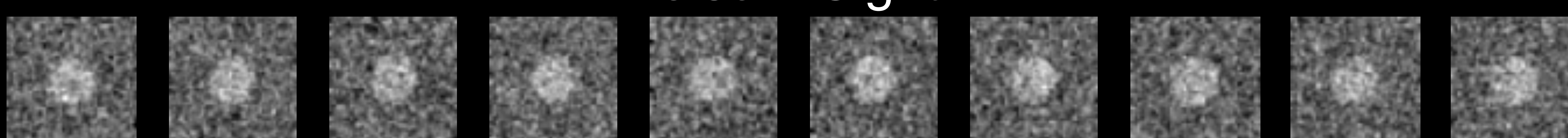
Uncorrected, FWE, FDR

Noise



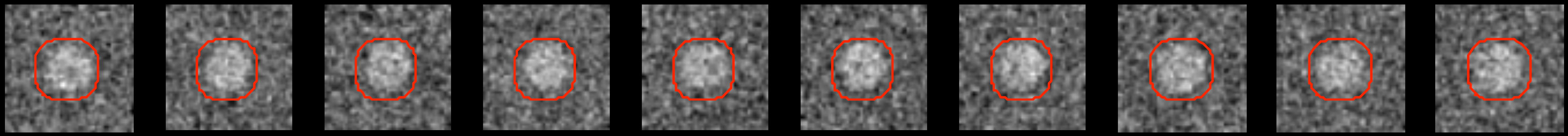
Signal

Noise + Signal

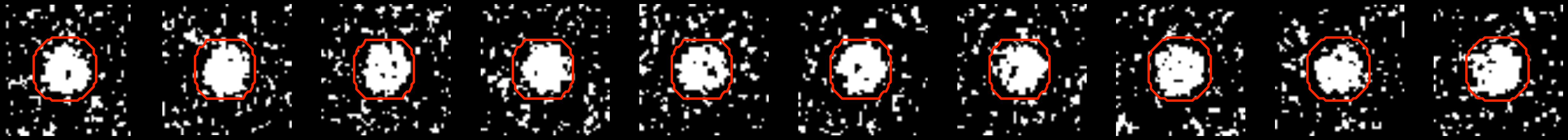


Uncorrected, FWE, FDR

Noise + Signal



Thresholded, uncorrected at p=0.10



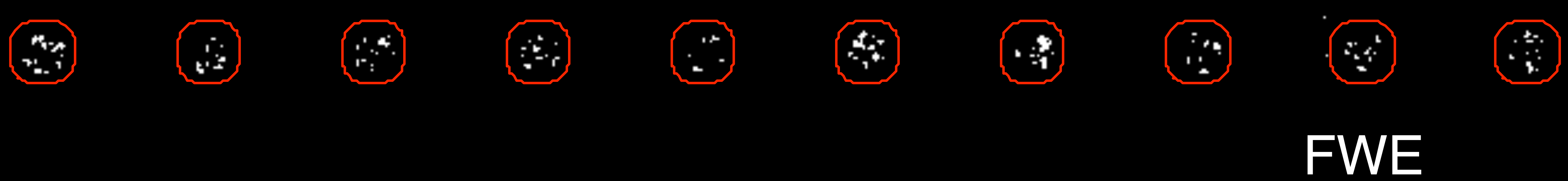
10% of noise voxels are false positives
(erroneously identified as significant)

➔ Needless to say, this is very bad!

Uncorrected, FWE, FDR

- Family-wise error (FWE)
  - Occurrence of ANY false positive
- FWE-Correction:
  - Controlling the FWE-rate to a small proportion

Thresholded, FWE-corrected at p=0.10

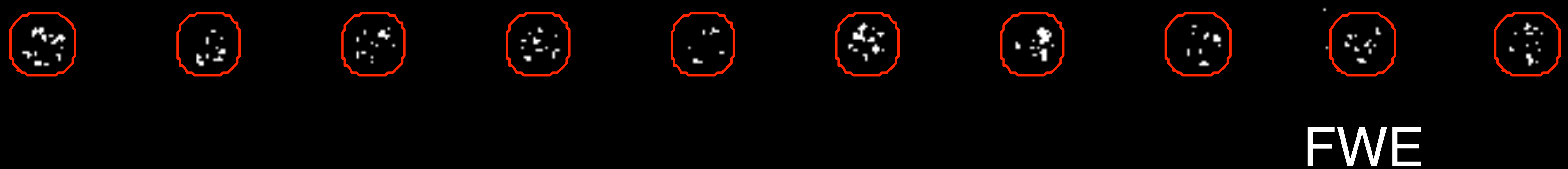


On average, FWE occurs with 10% probability

Uncorrected, FWE, FDR

- FWE-Correction:
  - False positives must be avoided at any cost!!
  - Even with diminished ability to detect true signal

Thresholded, FWE-corrected at p=0.10



On average, FWE occurs with 10% probability

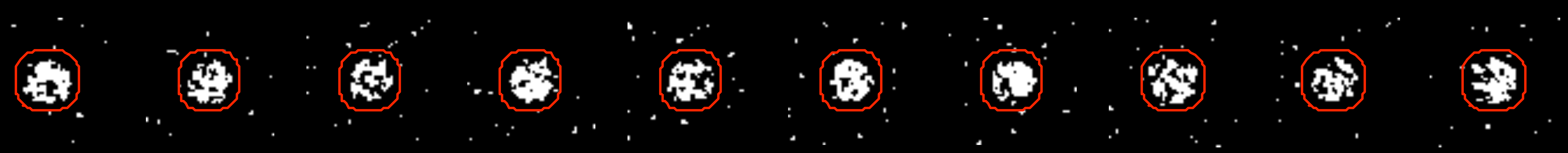
Uncorrected, FWE, FDR

- FWE-correction is too stringent
- If a large number of tests are performed…
  - Tolerating a small number of false positive — more realistic
- False discovery rate (FDR)
  - Proportion of false positives among all positives
    (true & false)

Uncorrected, FWE, FDR

- FDR-Correction:
  - A small proportion of false positives (among all positives) is tolerated
  - FDR is often denoted by q (as opposed to p)

Thresholded, FDR-corrected at q=0.10

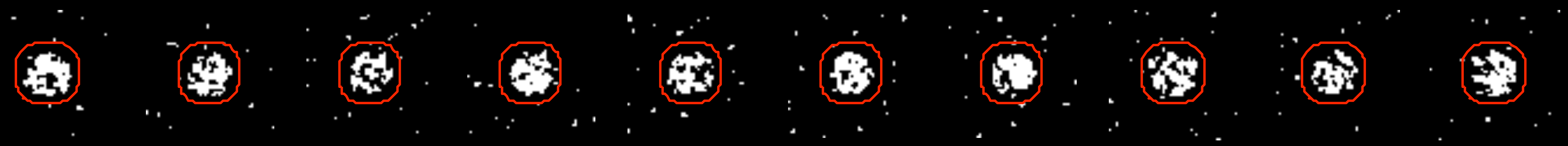


On average, 10% of positives are false positives

Uncorrected, FWE, FDR

- But wait?!!
    - Shouldn't we need to know which positives are true / false in order to control FDR?
    - We try to control the average FDR

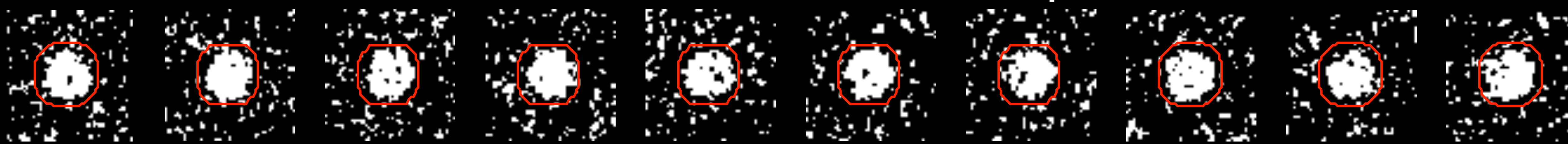
Thresholded, FDR-corrected at q=0.10



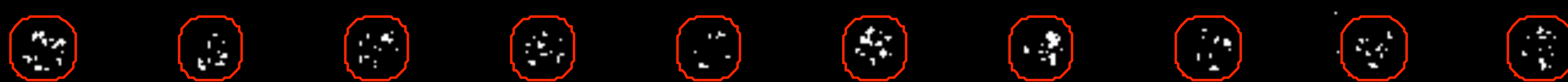
On average, 10% of positives are false positives

Uncorrected, FWE, FDR

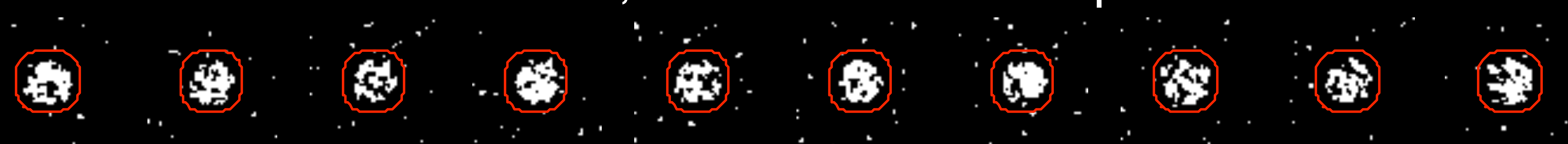
Thresholded, uncorrected at p=0.10



Thresholded, FWE-corrected at p=0.10



Thresholded, FDR-corrected at q=0.10



FDR-Correction

- Implemented by Benjamini-Hochberg procedure

  Genovese et al., NeuroImage (2002)

  - Distribution of signal + noise from data
  - Spatial correlation can be implicitly corrected
  - Details are beyond the scope of this course
- Threshold controlling FDR at the desired q-level
  - Highly specific to data / contrast

FDR-Correction

- Extent of signal can be very large
  - May not be ideal if you are interested in localizing activation
- NOT part of FEAT in FSL
  - However, FDR correction available
    - Requires some programming

FWE-Correction

<u>Typical verbiage:</u>

“*The statistical analysis is FWE (family-wise error)-corrected ($p<0.05$) at the _______ level based on ______.*”

FWE-Correction

<u>Typical verbiage:</u>

"*The statistical analysis is FWE (family-wise error)-corrected ($p<0.05$) at the _______ level based on ______.*"

<u>(Topology)</u>

- Voxel
- Cluster

FWE-Correction

Typical verbiage:

“The statistical analysis is FWE (family-wise error)-corrected ($p<0.05$) at the ________ level based on _______.”

(Topology)

- Voxel
- Cluster

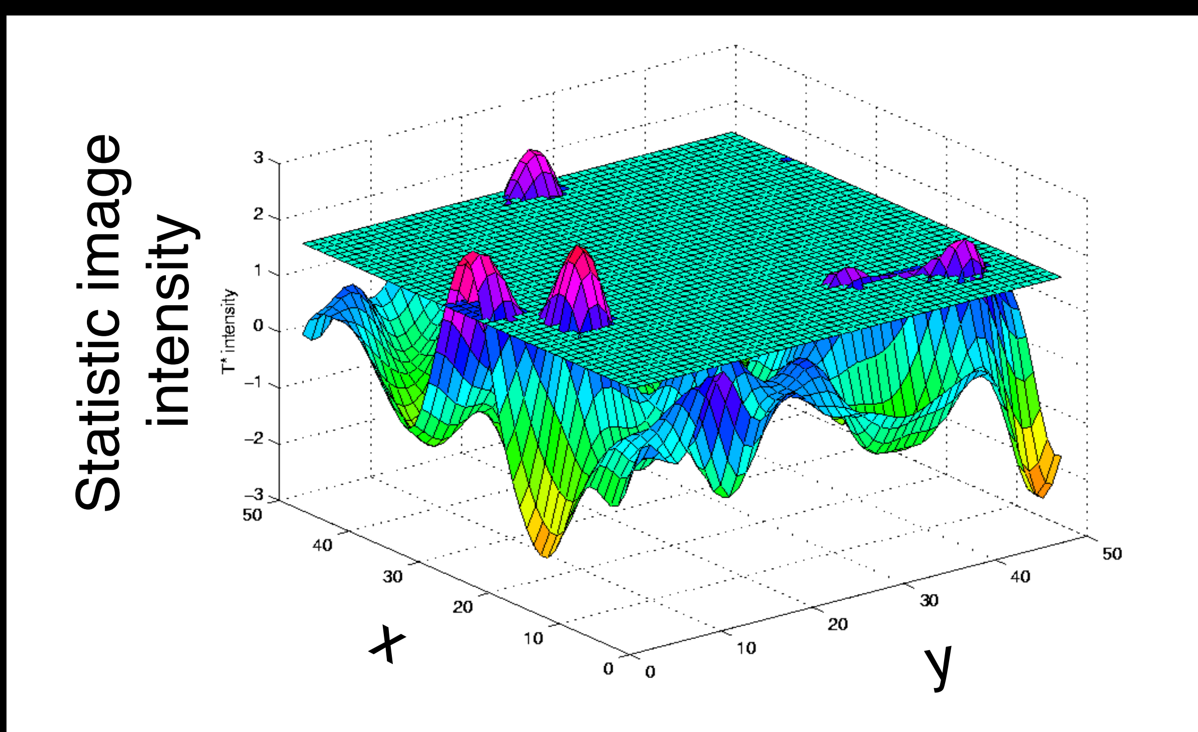
(Distribution modeling)

- Random field theory (RFT)
- Permutations

Voxel vs. Cluster-Level

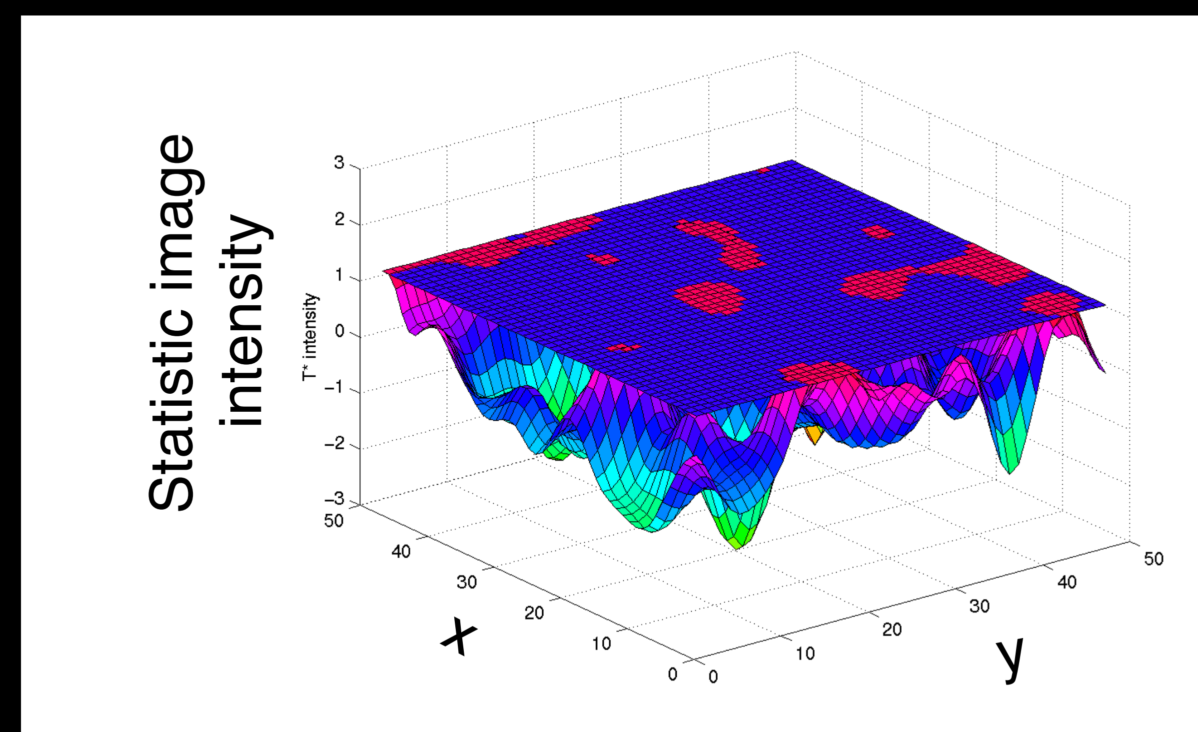
Voxel-level

Activation: high-intensity voxels



Cluster-level

Activation: signals with large spatial extent



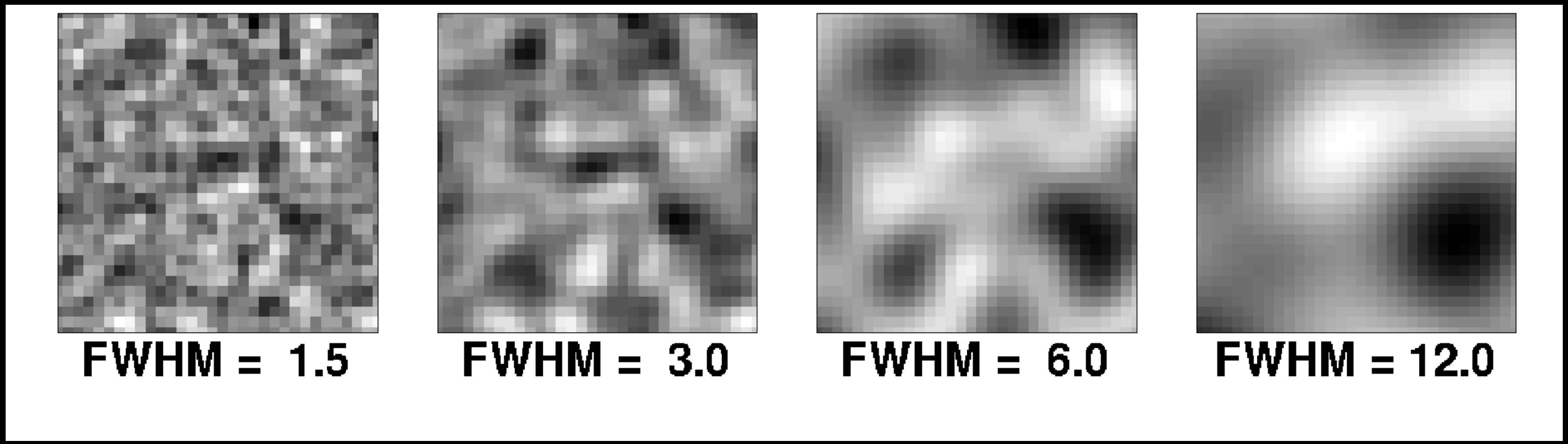
Voxel-Level FWE-Correction

- Threshold is determined from the max distribution
  - Distribution of the global maximum of a statistic image
    - 95th percentile of the max distribution
      ➔ FWE-corrected threshold at $p=0.05$ level
- Controls multiple comparison among all voxels in a statistic image

Voxel-Level FWE-Correction

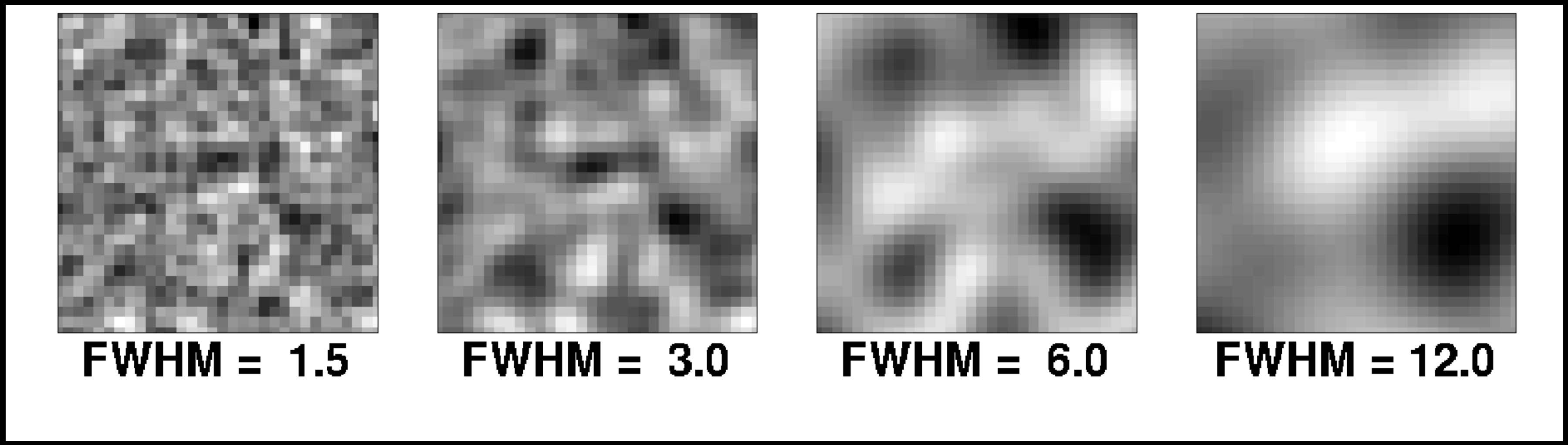
- The max distributions can be determined
  - Theoretically by RFT (random field theory)
  - Empirically by permutations

RFT, Voxel-Level



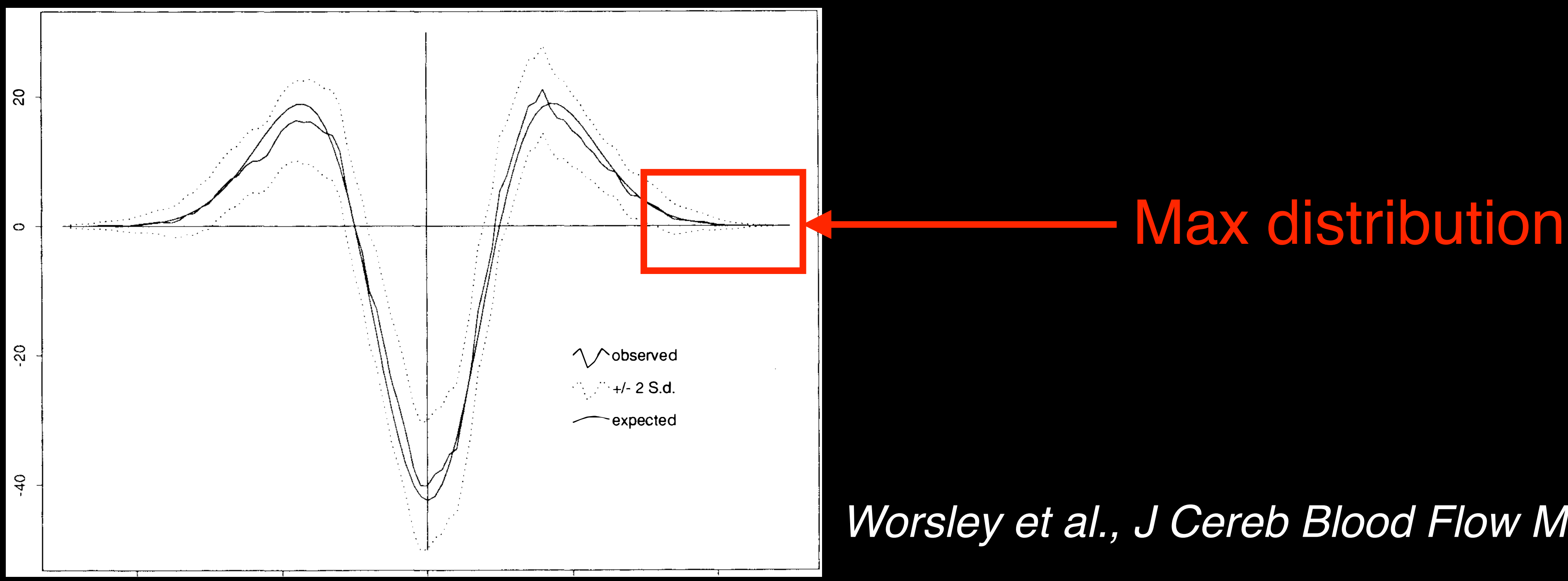
- A random field — random image
  - Intensity at each point — known distribution
  - Smooth — known spatial correlation

RFT, Voxel-Level



- Distribution of the brightest point (max distribution)
  - Theoretically approximated *Worsley et al., Hum Brain Map (1996)*
    - Depending on the volume and smoothness
  - FWE-corrected threshold, p-values can be determined

RFT, Voxel-Level



Worsley et al., J Cereb Blood Flow Metab (1992)

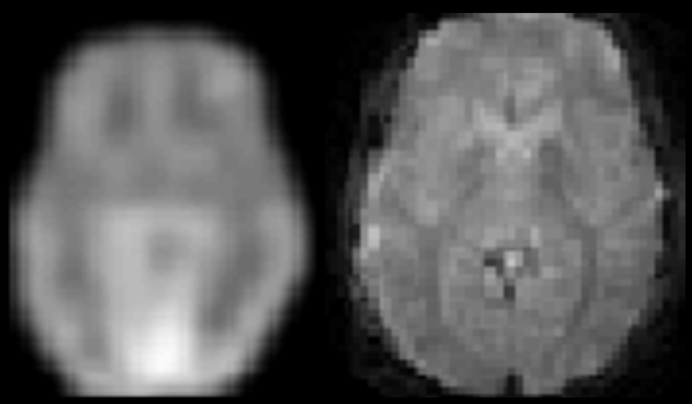
- Distribution of the brightest point (max distribution)
  - Theoretically approximated *Worsley et al., Hum Brain Map (1996)*
    - Depending on the volume and smoothness
  - FWE-corrected threshold, p-values can be determined

Notes: RFT, Voxel-Level

- Statistic image has to be smooth (FWHM>3 voxels)
- SPM can handle T-statistic images
- FSL converts T-statistic image to Z-statistic image
- Unnecessarily conservative
    - Developed for PET data analyses c. 1992
        - Much smoother than fMRI data
    - The actual FWE rate is much less than p=0.05

Nichols & Hayasaka, Stat Meth in Med Res (2003)

PET fMRI

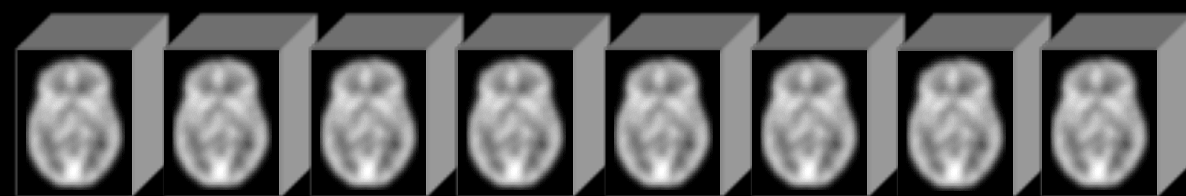


Permutations, Voxel-Level

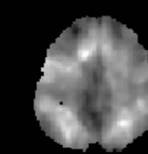
- The max distribution can be determined empirically from the data
  - Assuming that H_0 is true
    - Zero activation (one-sample) or zero difference (two-sample)
- How? Random shuffling of data labels (a.k.a., permutations)

Permutations, Voxel-Level

Contrast images

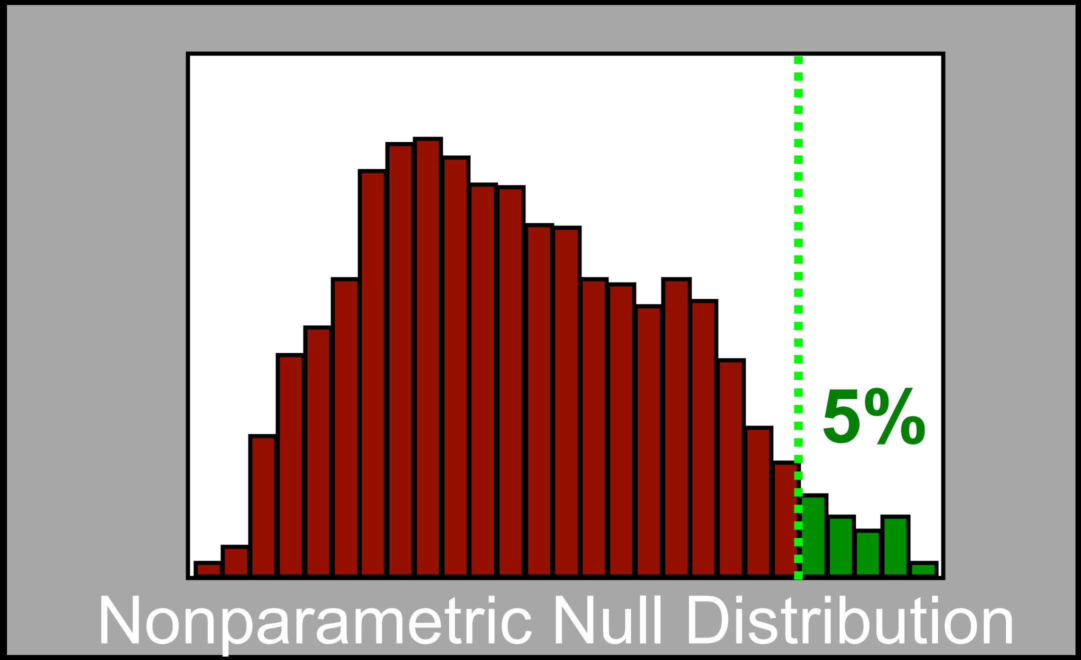


Statistic image



									Max intensity
Original	A	A	A	A	B	B	B	B	6.23
Permutation 1	A	B	A	B	B	A	B	A	3.42
Permutation 2	A	B	B	A	A	B	B	A	2.96
				⋮					⋮
Permutation 1000	B	A	A	B	B	A	B	A	3.12

Permutations, Voxel-Level



- From all permutations, max distribution can be generated
    - FWE-corrected threshold and p-values can be determined

Notes: Permutations, Voxel-Level

- Permutation schemes
  - Unpaired two-sample: shuffle group labels
  - Paired two-sample: flip 1st and 2nd images
  - One-sample: flip the sign (+/-) of stat images

Notes: Permutations, Voxel-Level

Pros

- More sensitive than RFT-based test
  - FWE rate is very close to 0.05
  - Especially useful when sample size is small
- Statistic image does not have to be smooth
- Can be used for “unusual” test statistic

Notes: Permutations, Voxel-Level

Cons

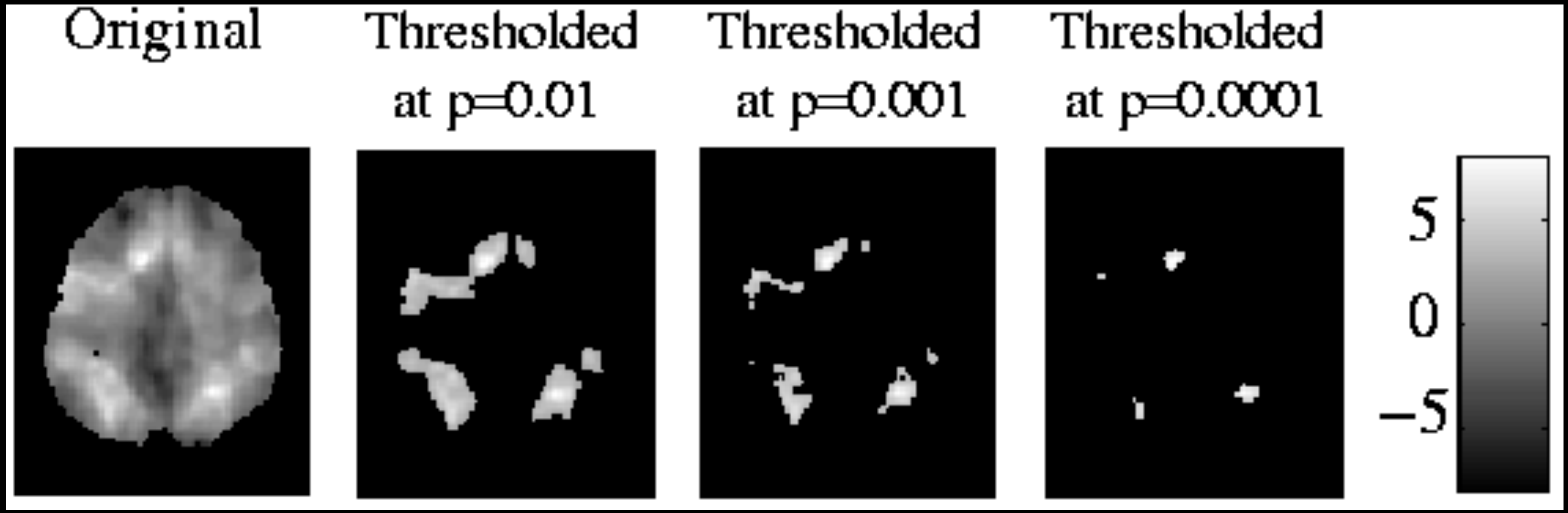
- Simple study designs only
  - One-sample, paired / unpaired two-sample
  - Cannot be used in the 1st level analysis
    - Temporal correlation cannot be preserved when permuted
- Permutations can be time consuming
  - Repeating the 2nd level analysis many times
  - Recommended # of permutations: at least 1,000

Notes: Permutations, Voxel-Level

- Permutation test — optional feature
  - FSL — **Randomise**
  - SPM — **SnPM**

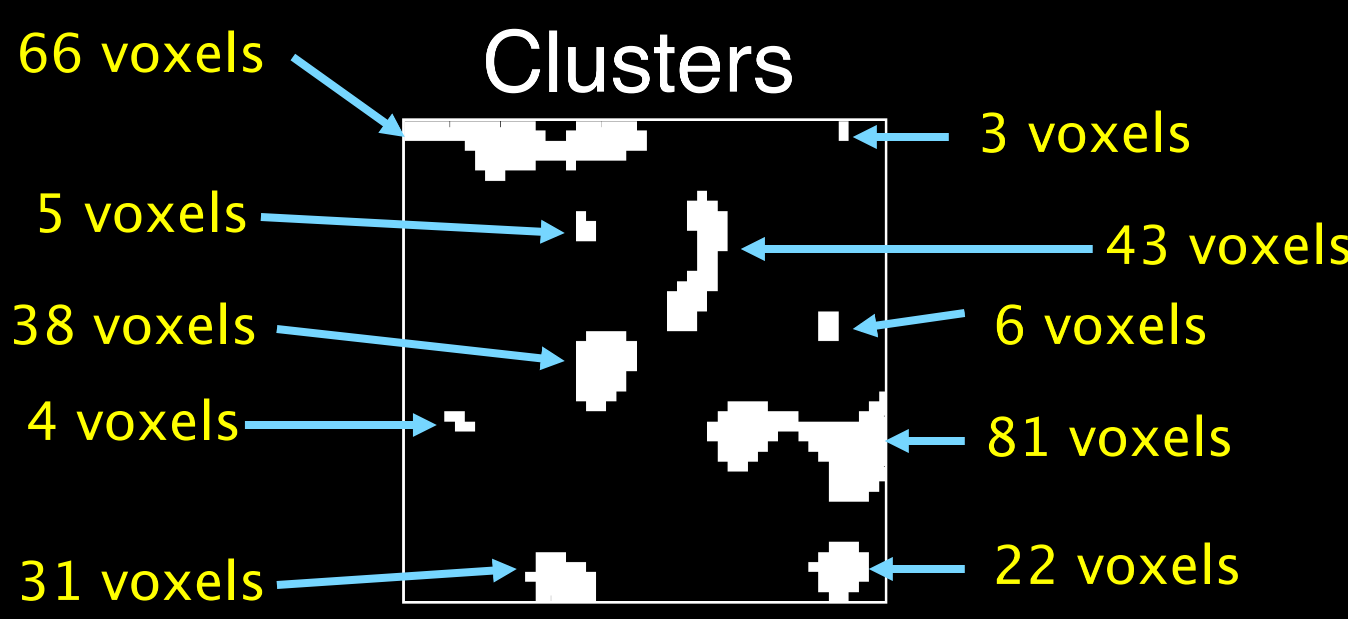
Cluster-Level FWE-Correction

- Cluster formation
  - Statistic image is thresholded with an uncorrected threshold
    - Cluster-forming threshold
  - Contiguous above-threshold voxels ➔ clusters



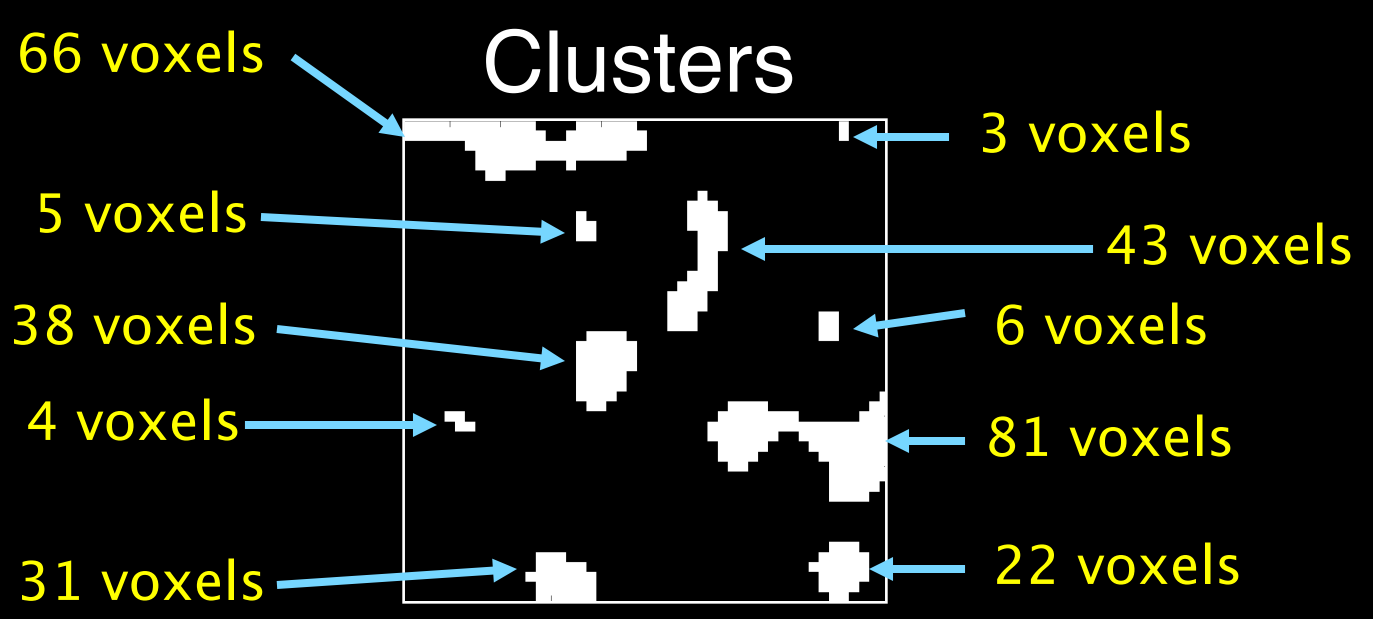
Cluster-Level FWE-Correction

- Large cluster ➔ likely true signal
- Multiple comparison correction among clusters
  - A few dozen clusters at most
  - As opposed to 20,000 multiple comparisons at voxel-level

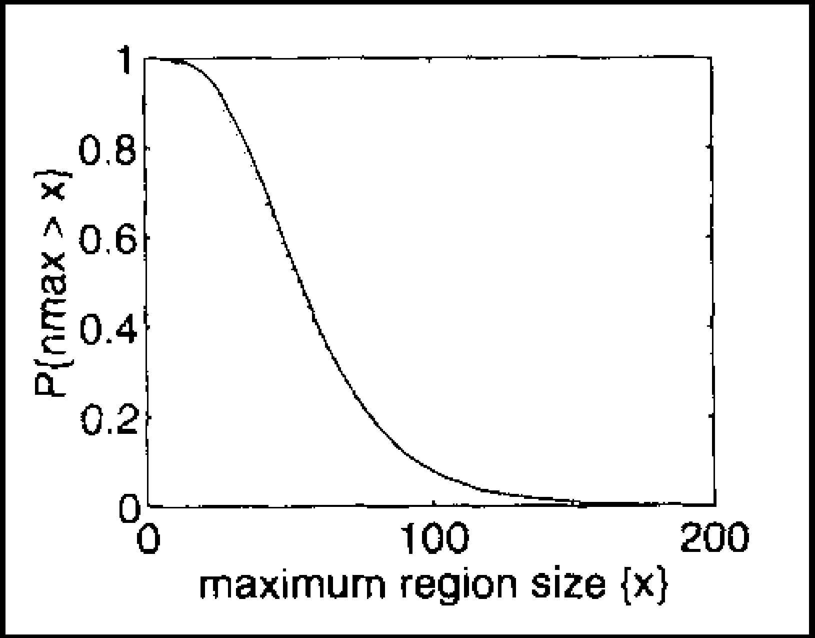


Cluster-Level FWE-Correction

- Max cluster size distribution
  - Way to control multiple comparisons among clusters
- Max distribution can be determined
  - Theoretically by RFT
  - Empirically by permutations



RFT, Cluster-Level



Friston et al., Hum Brain Map (1994)

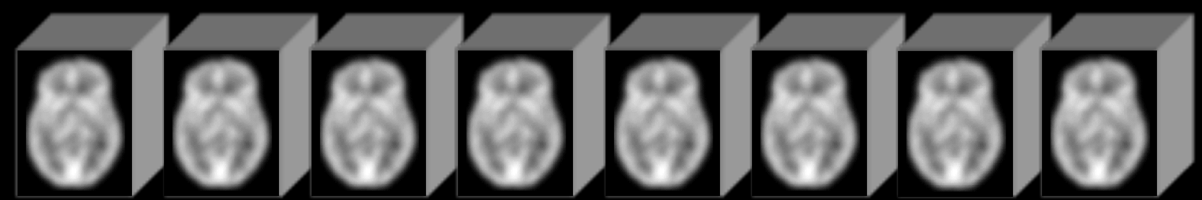
- Distribution of the largest cluster size (max distribution)
  - Theoretically approximated *Friston et al., Hum Brain Map (1994)* *Worsley et al., Hum Brain Map (1996)*
    - Depending on the volume, smoothness, and cluster-forming threshold
  - FWE-corrected threshold, p-values can be determined

Permutations, Cluster-Level

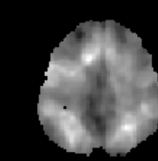
- The max distribution can be determined empirically from the data
    - Assuming that H_0 is true
        - Zero activation (one-sample) or zero difference (two-sample)
- Random shuffling of data labels as in voxel-level permutations

Permutations, Cluster-Level

Contrast images



Statistic image



									Max cluster size
Original	A	A	A	A	B	B	B	B	232
Permutation 1	A	B	A	B	B	A	B	A	18
Permutation 2	A	B	B	A	A	B	B	A	23
				⋮					⋮
Permutation 1000	B	A	A	B	B	A	B	A	17

Notes: Cluster-Level FWE-Correction

- Good for spatially extended signals
  Friston et al., NeuroImage (1996) *Poline et al., NeuroImage (1997)*
- Better sensitivity compared to voxel-level FWE
  - Smaller number of multiple comparisons
  - Use of lower threshold
- Limited spatial localization power
  - Cannot localize within a cluster

Notes: Cluster-Level FWE-Correction

- Permutation-based — more sensitive than RFT-based *Hayasaka & Nichols, NeuroImage (2003)*

Notes: Cluster-Level FWE-Correction

- The validity of cluster-level correction is disputed

  Eklund et al., PNAS (2016)

  - Validated with real data
  - Inflated false positives, up to 70%
  - Commonly used packages (FSL, SPM, AFNI)

- Significance of significant clusters may be challenged