Valid double-dipping for fMRI cluster analysis via permutation-based selective inference

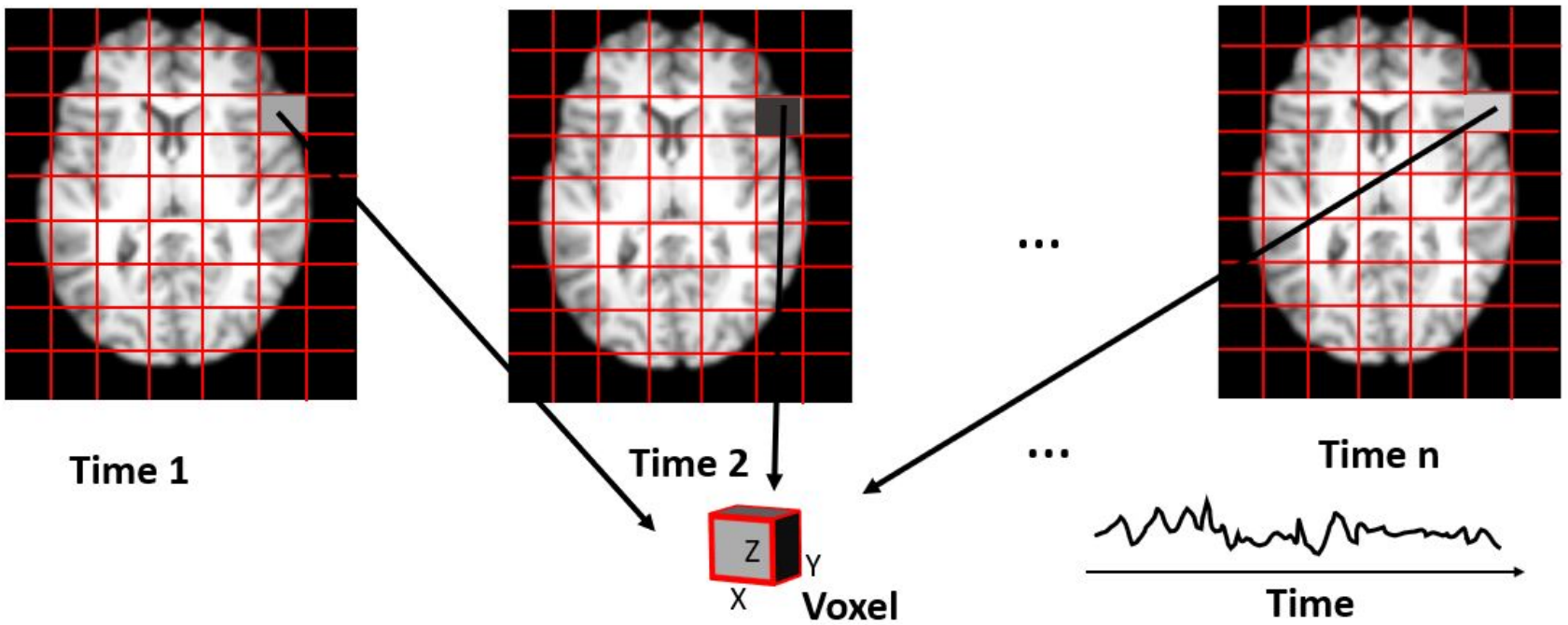
Angela Andreella[1], Jesse Hemerik[2], Livio Finos[3], Wouter Weeda[4], and Jelle Goeman[5]

[1]Department of Economics, Ca' Foscari University of Venice
[2]Biometris, Wageningen University and Research
[3]Department of Developmental Psychology and Socialisation, University of Padua
[4]Department of Psychology, Leiden University
[5]Biomedical Data Sciences, Leiden University Medical Center

EMPG - 7 September 2022

Functional Magnetic Resonance Imaging (fMRI) data

fMRI is the most used technology to study human cognition's neural underpinnings. The brain activation is measured as changes in blood flow (BOLD) under a sequence of **stimuli**.



Functional Magnetic Resonance Imaging (fMRI) data

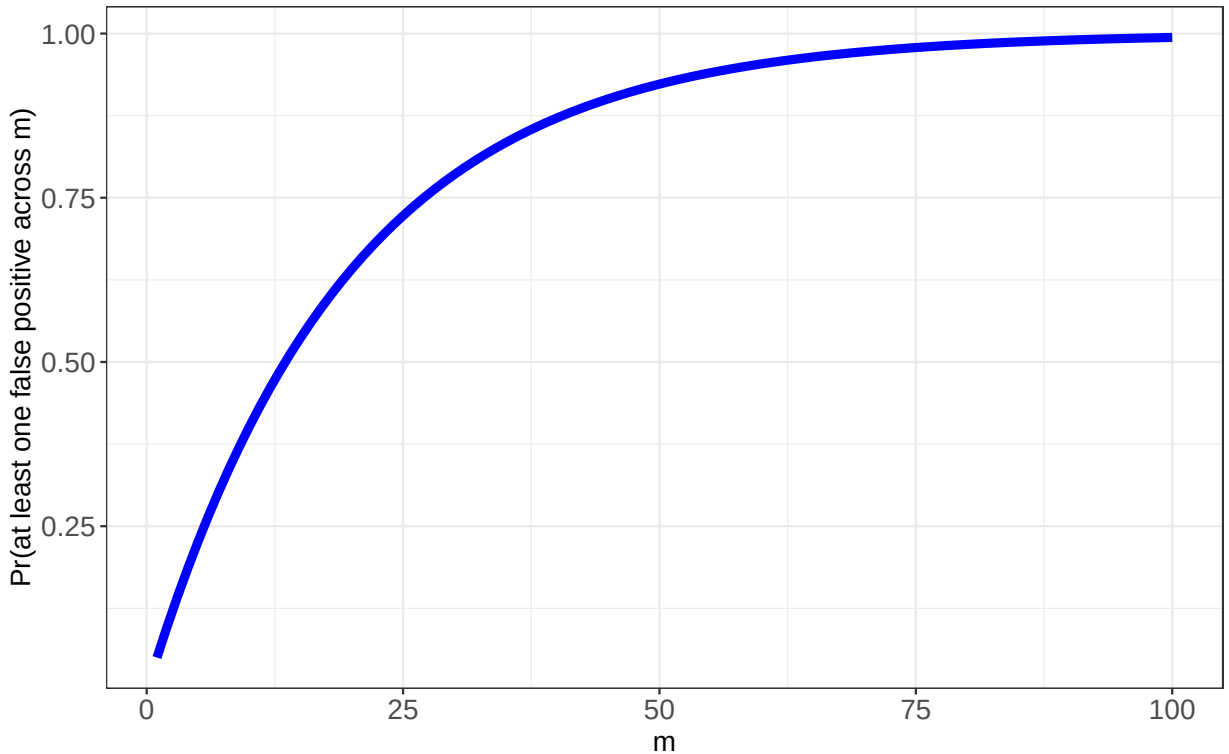
- **Response**: Brain activity measured for each **voxel** → about $200,000$ response activation to test.
- **Question**: Which regions respond differently to different **stimuli**? → is the region active?

So, We have a **test statistic**:

- for each voxel →
  H_v : voxel v is not active;
- for each set of voxels →
  H_S : the set of voxel S is not active .

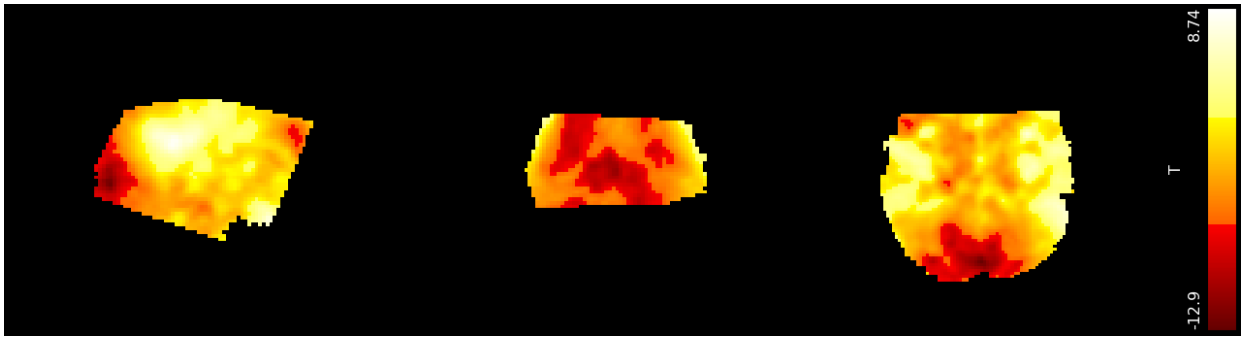
In a nutshell:

- When we test a single null hypothesis $H_0 \rightarrow$ the probability to reject H_0 when is true equals α (e.g., 0.05).
- If we test $m > 1$ null hypothesis, the probability to reject at least one hypothesis across m hypothesis equals $1 - (1 - \alpha)^m$

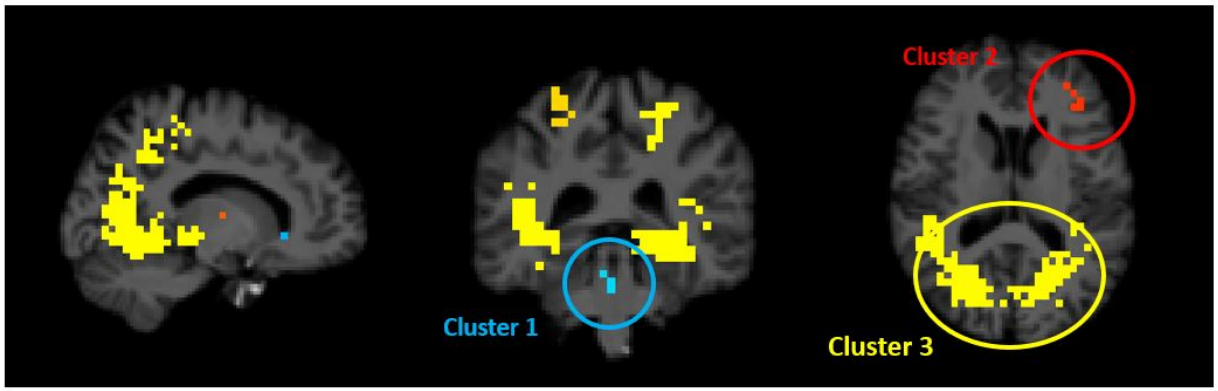


Multiple Testing problem

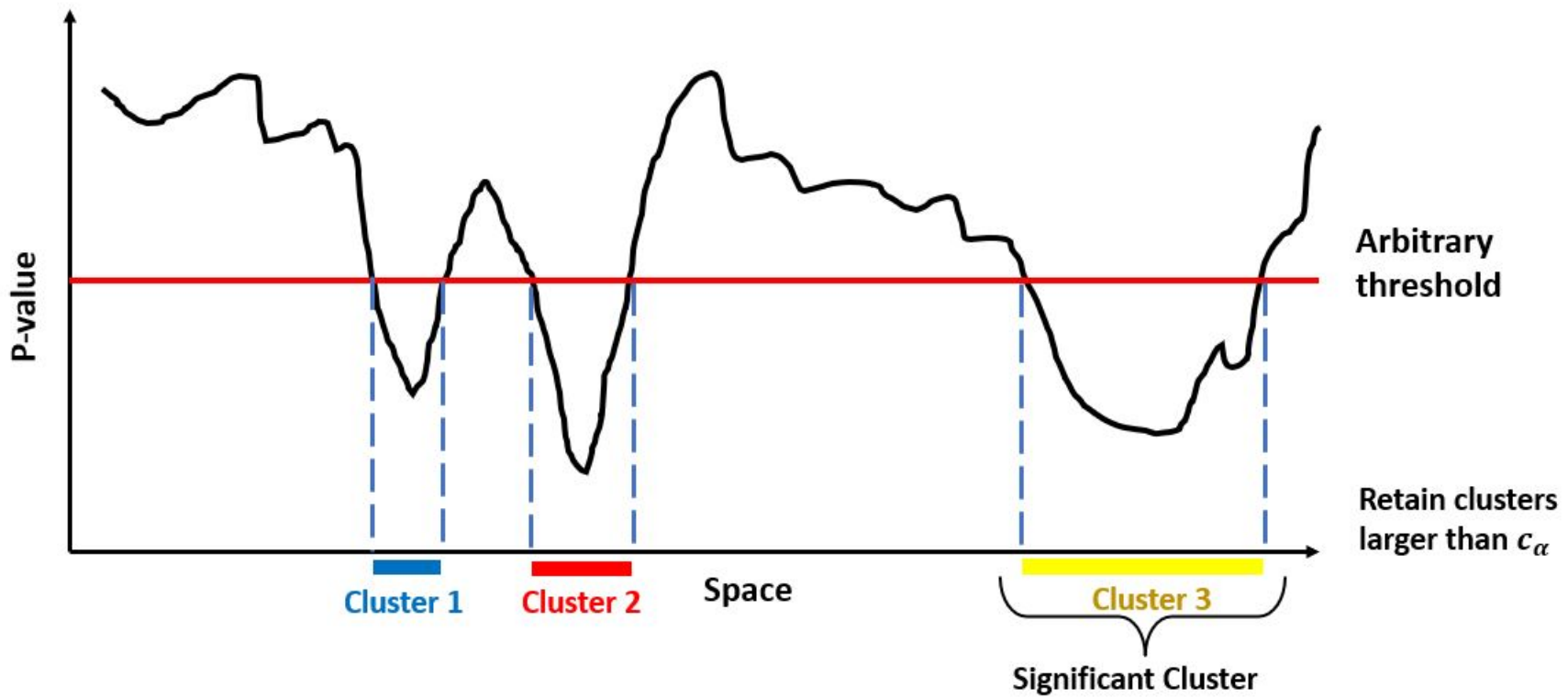
If we want to infer at voxel level $\rightarrow$ 200,000 statistical tests $\rightarrow$ $\alpha = 0.05/200,000$ (i.e., Bonferroni correction) $\rightarrow$ very low power to detect activation.



So, we can infer at level of set of voxels $\rightarrow$ 3 statistical tests $\rightarrow$ $\alpha = 0.05/3$ (i.e., Bonferroni correction) $\rightarrow$ we gain power.



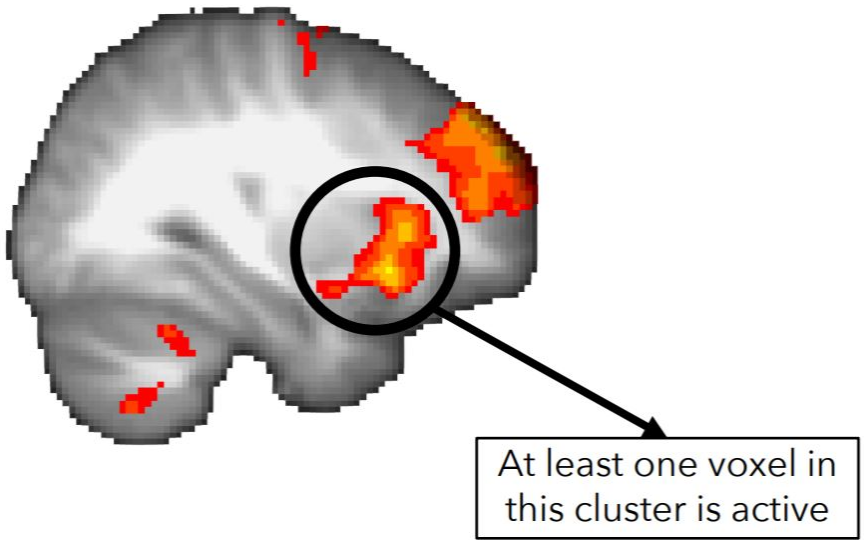
In fMRI data analysis, the most widely used method for locating brain activity is **cluster-extent based thresholding** [1].



[1] Woo, C. et al. (2014).

Problems:

- If we reject the null hypothesis we can only say:



Spatial Specificity paradox [2] →
The larger the cluster, the weaker the finding!

[2] Woo, C. et al. (2014).

- Inferring inside data-driven clusters leads to biased results!

  →

  Double-dipping

Solution:

- **Selective inference** approach: you can make post-hoc selection and follow-up inference (i.e., double-dipping);
- **Permutation-based** approach: account for the unknown dependence structure of voxels

Permutation-based All-Resolution Inference (**pARI**) → computes the lower bounds for the number of true active voxels inside each clusters (after seeing the data, changing the cluster-wise threshold, etc.).

Let consider B collection of m voxel activation to test.

- **Hypothesis**: $H_1, \ldots, H_m$;
- **True Discoveries**: $TD \subset \{1, \ldots, m\}$;
- **Cluster**: $S \subset \{1, \ldots, m\}$ selected hypothesis;
- **True Discoveries in S**: $|TD \cap S| \subset \{1, \ldots, m\}$.

We want to infer on $TD(S) = |TD \cap S|$ $\rightarrow$ compute the lower bound $\bar{TD}(S)$ $\rightarrow$ we can say "in S at least $\bar{TD}(S)$ are truly active".

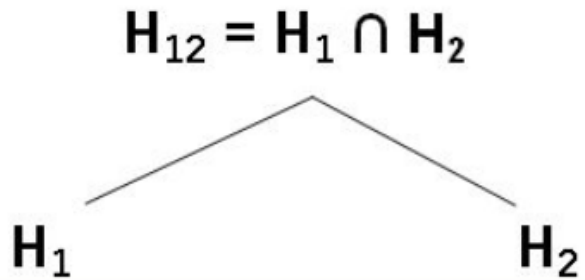
Two ingredients:

- **Closed testing procedure** for controlling the familywise error rate[3]:

$$\Pr(TD(S) \geq \bar{TD}(S) \text{ for each } S \in B) \geq 1 - \alpha$$

- **Local test** used in the closed testing procedure → permutation-based tests.

Closed testing rejects any H_i, if all possible intersection hypotheses involving H_i are rejected by using valid local level α tests.



Reject H_1 at level 0.05 if H_{12} is also rejected at level 0.05.

[3] Marcus, E. et al. (1976).

```
install.packages("pARI")
library(pARI)
```

- fMRI framework:

```
pARIbrain(copes, thr, mask, alpha, ...)
```

- General framework:

```
pARI(data, ix, alpha, test.type, ...)
```

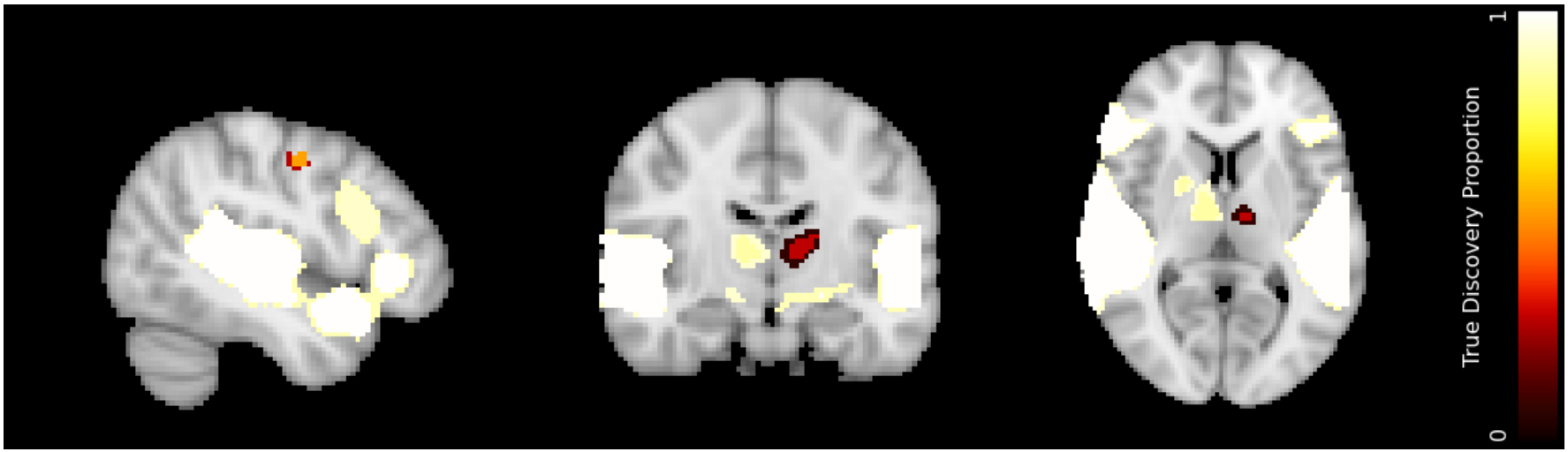
where `ix` is the **set** S of interest (i.e., vector of hypotheses indices).

We analyzed the **Auditory data** collected by Pernet et al. (2015), i.e, people listening vocal and non-vocal sounds.

Group analysis on 140 subjects of the Vocal > Non-vocal **contrast** by the one sample t-test flipping the sign of 140 voxel-wise contrasts maps.

Cluster S	**Threshold** t	**Size** $\lvert S\rvert$	**% active** $\bar{TD}(S)$ Perm. Simes	Param. Simes	**P-Values** p_{FWER}
Right STG/PT HG/IFG/T	3.2	11683	92.36%	84.98%	< 0.0001
Right STG/PT HG/IFG/T	4	8875	99.54%	98.5%	–
Right IFG	4	422	91.47%	83.18%	–
Right T	4	292	85.96%	64.04%	–
Right T	4	15	13.33%	0%	–

fMRI data



The **permutation-based ARI**:

- **flexible**, **mild** and **post-hoc** → You can choose the features set as many time as you want!
- resolves the **spatial specificity paradox** of cluster fMRI data analysis;
- accounts for the **dependence** structure of the data;
- makes **no assumptions** on the null distribution of the p-values → only **exchangeability**;
- the **power** does not depend on the number of features sets tested.