Data mining a functional neuroimaging database for functional segregation in brain regions

Finn Årup Nielsen
with Daniela Balslev and Lars Kai Hansen

Lundbeck Foundation Center for Integrated Molecular Brain Imaging;
Informatics and Mathematical Modelling,
Technical University of Denmark;
Neurobiology Research Unit,
Copenhagen University Hospital Rigshospitalet
August 30, 2006

Human brain mapping

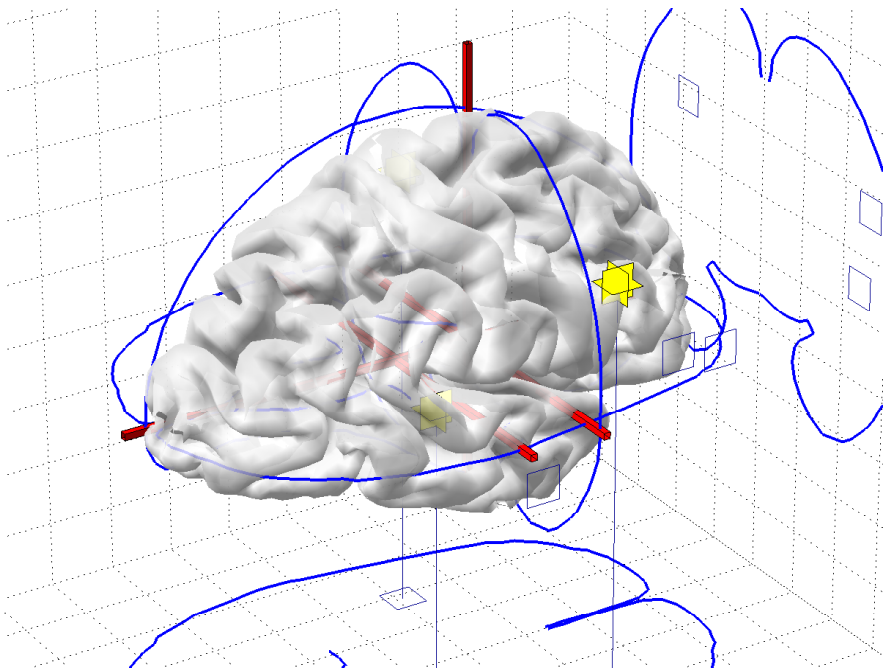


Figure 1: Results from a human brain mapping study (Balslev et al., 2005) with a "Visible Human" surface (Drury et al., 1996) displayed in a 3-dimensional corner cube environment. Two of three reported activations are visible.

Positron emission tomography or functional magnetic resonance brain scans of the human brain while subjects are engaged in the investigated mental processes.

Result represented in the literature with lists of "locations", i.e., three dimensional coordinates (in standardized "Talairach" brain space, of the hot spot activations, e.g.,

(x, y, z)	z-score
$-38, 0, 40$	4.91
$48, -42, 8$	4.66
$52, 14, 38$	4.07

Functional segregation

Two brain functions may involve different parts of a brain region, and meta-analyses can elucidate this, e.g.,

(Bush et al., 2000): *Cognitive* and *affective* division of *anterior cingulate* cortex (lower part "emotional", upper part "cognitive")

(Steel and Lawrie, 2004): *Emotion* and *cognition* in the *prefrontal* cortex.

(Poldrack et al., 1999): *Semantic* and *phonological* processing in *left inferior prefrontal* cortex

Brede database

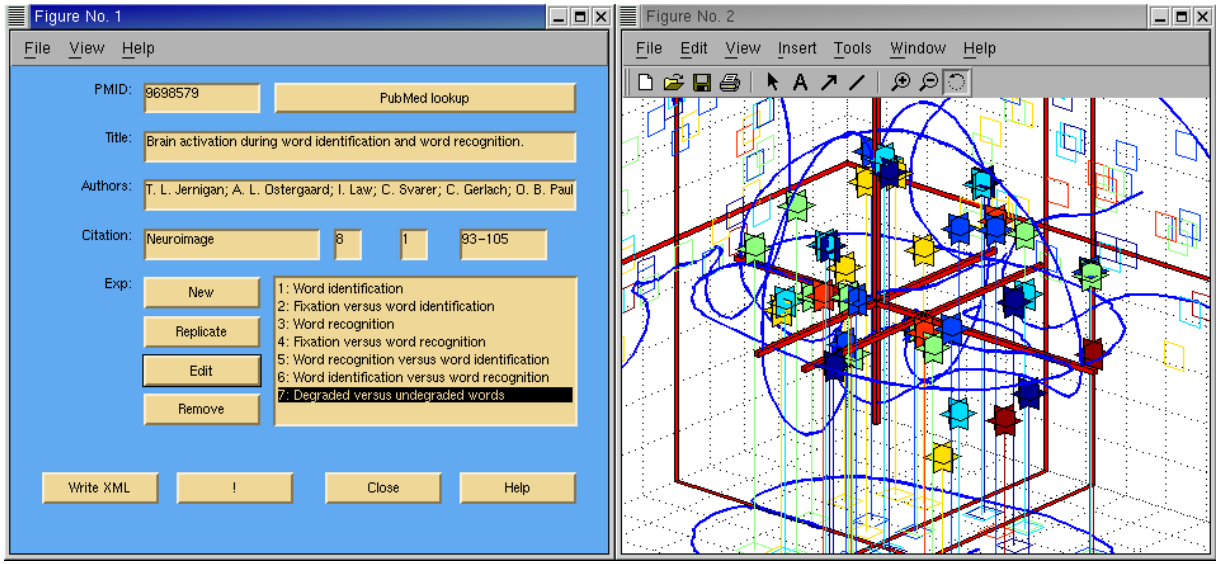


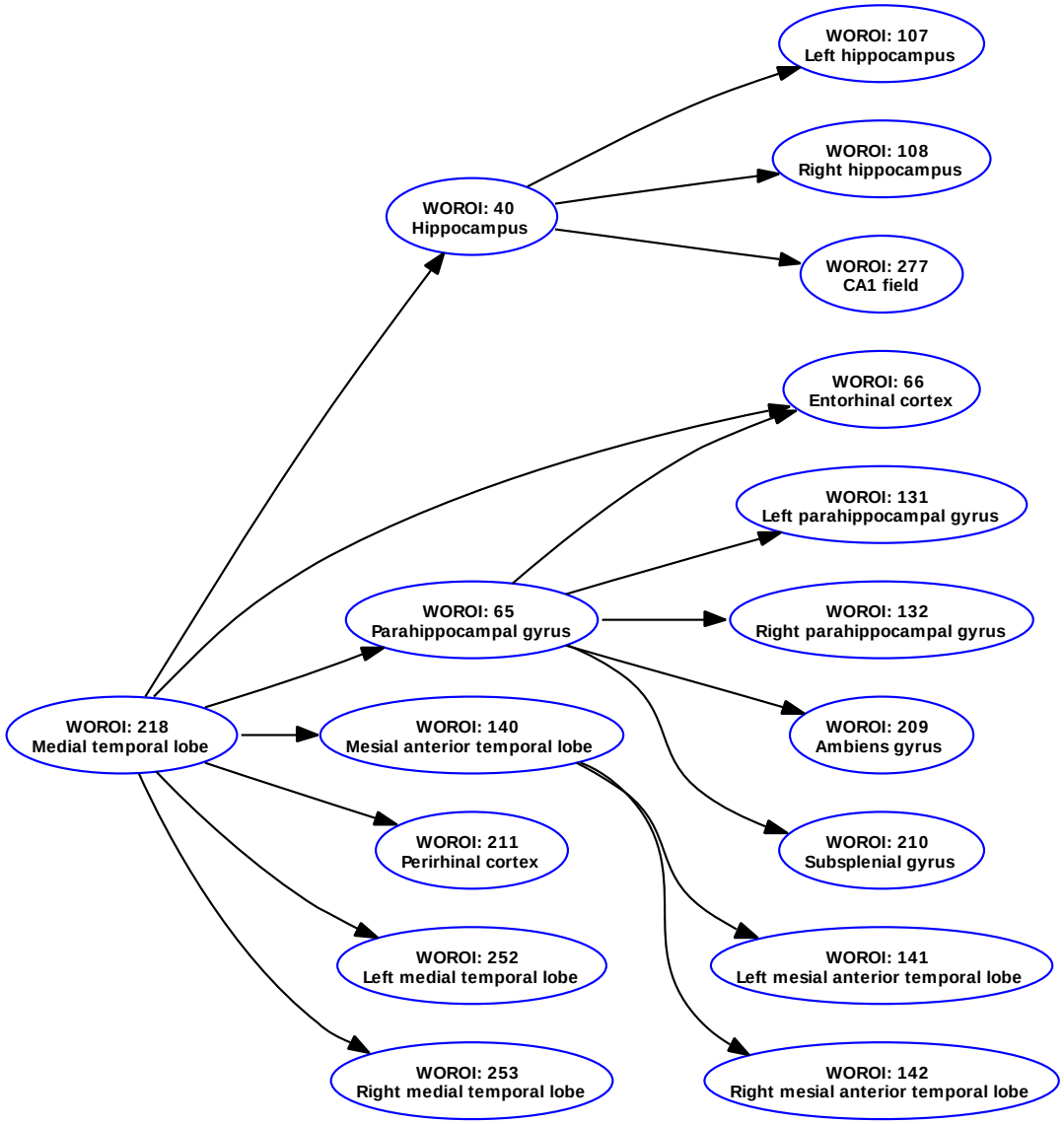
Figure 2: Screenshot of main window of Matlab program for data entry of one of the studies in the Brede database (Jernigan et al., 1998).

Brede Database contains, e.g., abstract, locations stored in XML (Nielsen, 2003).

Presently contains almost 4000 locations each with the 3-dimensional coordinates and many with anatomical annotation.

Abstract, the 3-dimensional coordinates and anatomical annotation are used in the following.

Brede Database neuroanatomy taxonomy



Hierarchy of brain regions.

Based on another neuroanatomical database "BrainInfo/NeuroNames" (Bowden and Martin, 1995) and atlases, e.g. "Mai atlas" (Mai et al., 1997).

Fields recorded: Canonical name, variation in names, abbreviations, links to NeuroNames and other databases.

Graph constructed with GraphViz (Gansner and North, 2000).

This study

For a brain region = 1 To 313 brain regions

Step 1: Get all coordinates for the specific area, build a density model, exclude coordinates that are outliers

Step 2: Determine themes of the brain area with text mining on abstracts that contain coordinates within the brain area

Step 3: Determine whether specific themes are spatially clustered in the brain area by testing whether two sets of coordinates are separated.

end

Step 4: Intertwine results from all brain regions

Example names for "medial temporal lobe"

```
'Medial temporal lobe'
'Hippocampus'
'Parahippocampal gyrus'
'Parahippocampal'
'Parahippocampus'
'Gyrus parahippocampi'
'Gyrus parahippocampalis'
'Entorhinal cortex'
'Cortex entorhinalis'
'Entorhinal area'
'Area entorhinalis'
'Left hippocampus'
                ⋮
```

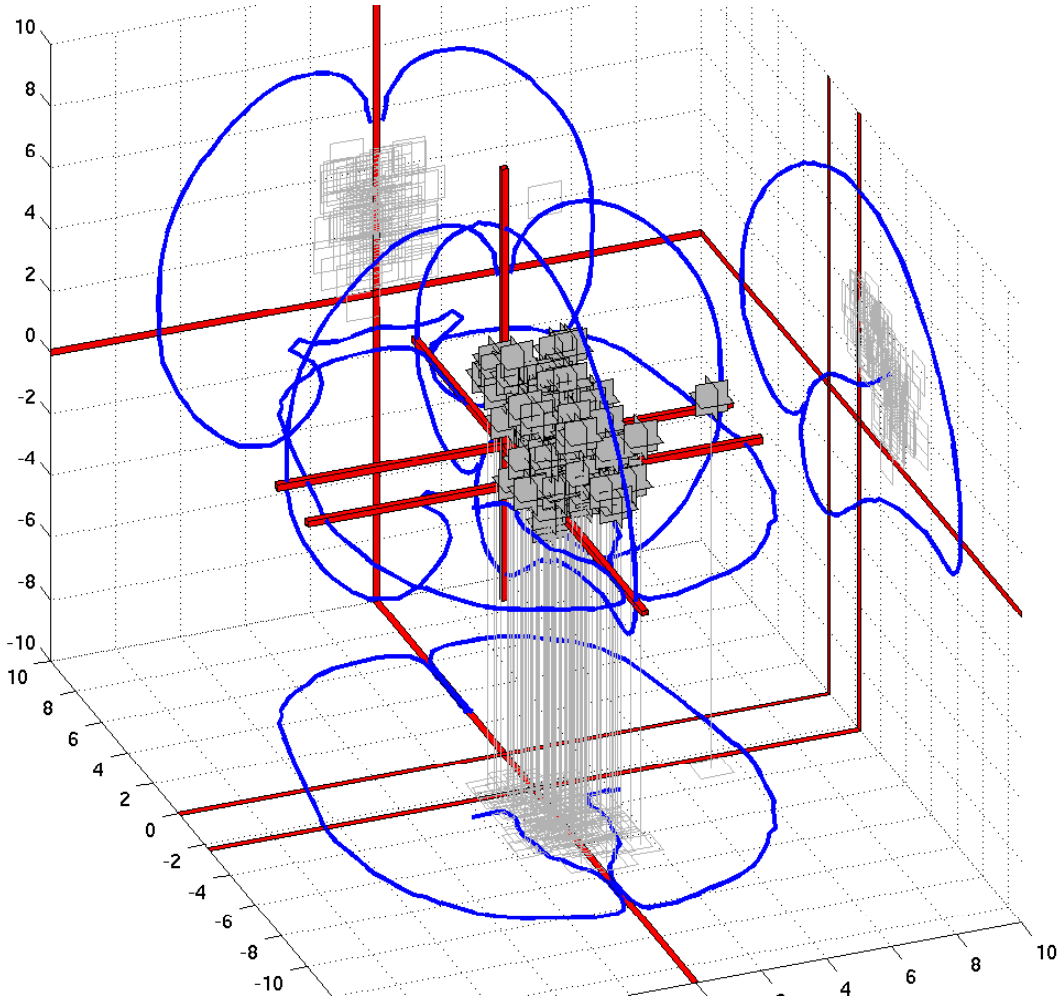
Use of brain region taxonomy.

Example of expansion from "medial temporal lobe"

Only one location matches on "medial temporal lobe"

After expansion with 32 names for subareas (and the region itself) there are 67 locations.

Step 1: Identify coordinates



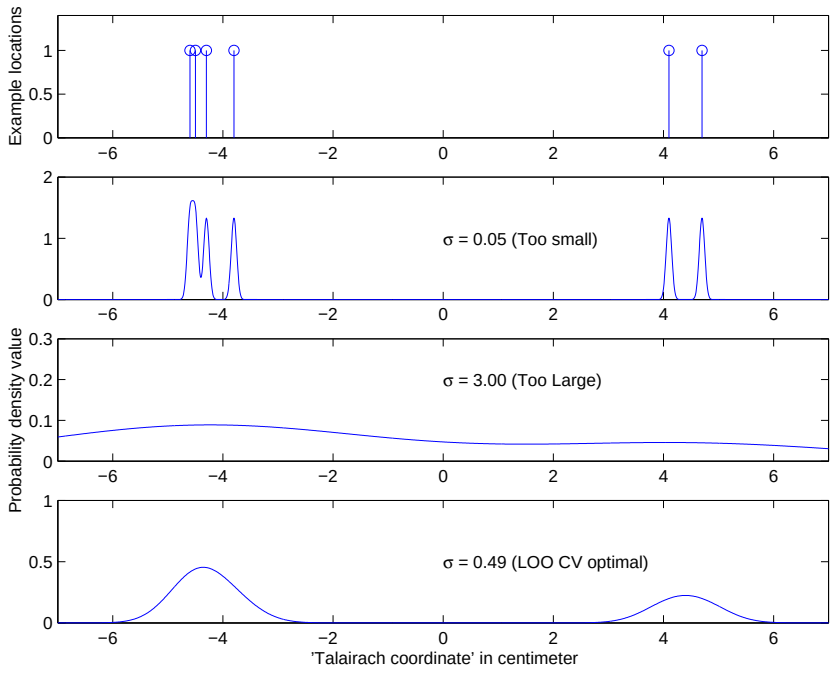
Simple SQL-like command in Matlab to find locations

Corner cube visualization of 116 "posterior cingulate" coordinates found

An outlier: "Right postcentral gyrus/posterior cingulate gyrus" from (Jernigan et al., 1998).

Build kernel density estimate of the coordinates.

Step 1: Spatial outlier elimination



Throw away the 5% most extreme coordinates (111 locations back).

Find a threshold as the lowest probability density estimate for a location with leave-one-out kernel density estimate.

Search in the entire database for all location above the threshold (184 locations). This should find coordinates that are not labeled.

Step 2: Bag-of words matrix

	'memory'	'visual'	'motor'	'time'	'retrieval'	...
Fujii	6	0	1	0	4	...
Maddock	5	0	0	0	0	...
Tsukiura	0	0	4	0	0	...
Belin	0	0	0	0	0	...
Ellerman	0	0	0	5	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

For the further analysis: Include all papers that contain one or more of coordinates found.

Representation of the abstracts of the papers in a bag-of-words matrix: (abstract $\times$ words)-matrix $\equiv \mathbf{X}(N \times P)$.

Step 2: Elimination of stop words and scaling

Common words: a, a's, able, about, above, accordingly . . . (571 words)

Common "scientific" words (from MEDLINE): accordingly, affected, affecting, affects, . . . (243 words)

Brain anatomy: amygdala, amygdaloid, angular, anterior, area, basal, bilateral, brain, brainstem . . . (148 words)

Words not associated with mental function: aberrant, aberrations, abilities, . . . (2534 words)

Element-wise square root scaling of the elements in the bag-of-words matrix . . . (Penrose, 1946).

Step 2: Non-negative matrix factorization

Non-negative matrix factorization (NMF) decomposes a non-negative data matrix $\mathbf{X}(N \times P)$ (Lee and Seung, 1999)

$$\mathbf{X} = \mathbf{W}\mathbf{H} + \mathbf{U}, \tag{1}$$

where $\mathbf{W}(N \times K)$ and $\mathbf{H}(K \times P)$ are also non-negative matrices.

"Euclidean" cost function for

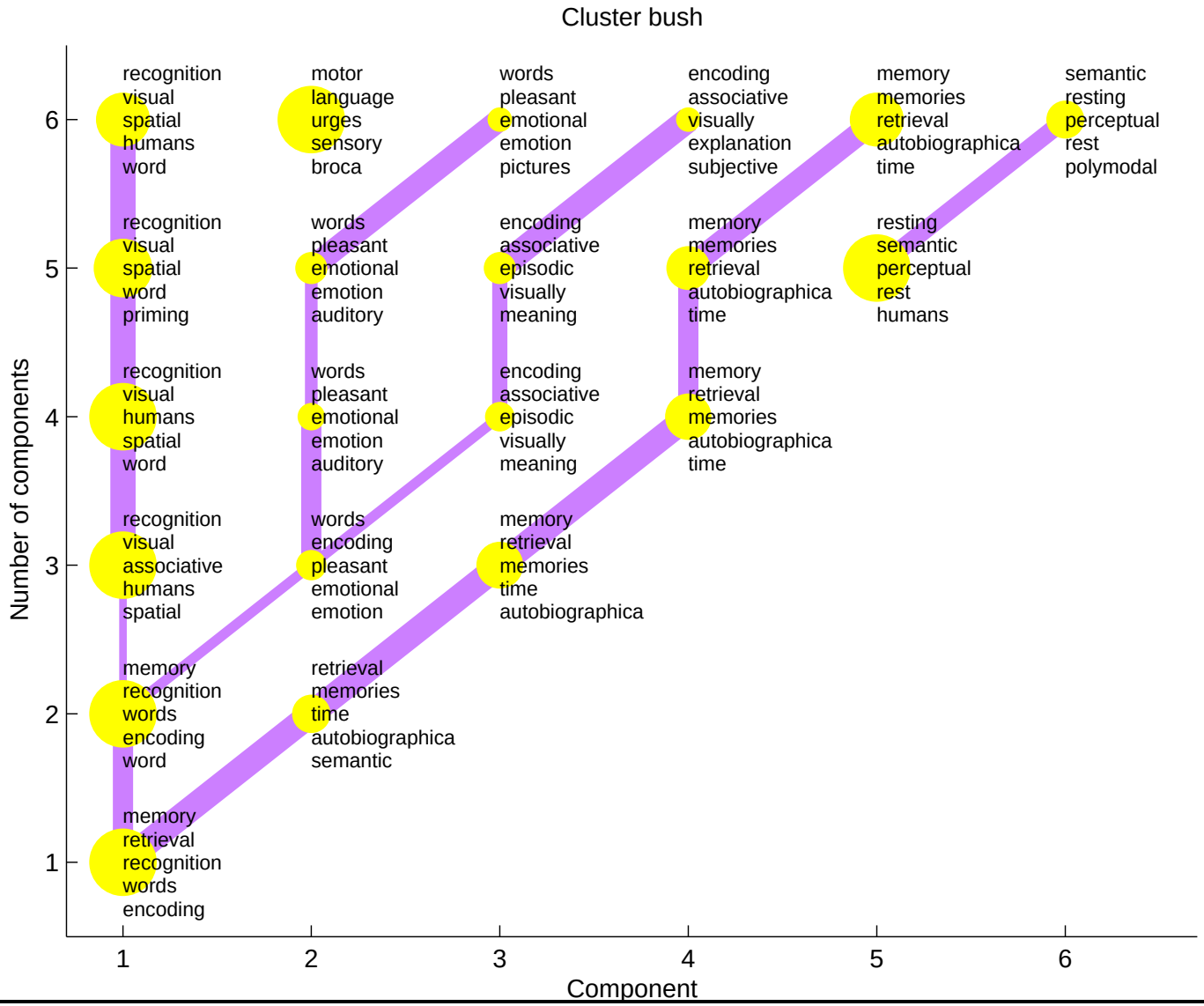
$$E_{\text{"eucl"}} = ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 \tag{2}$$

Iterative algorithm (Lee and Seung, 2001)

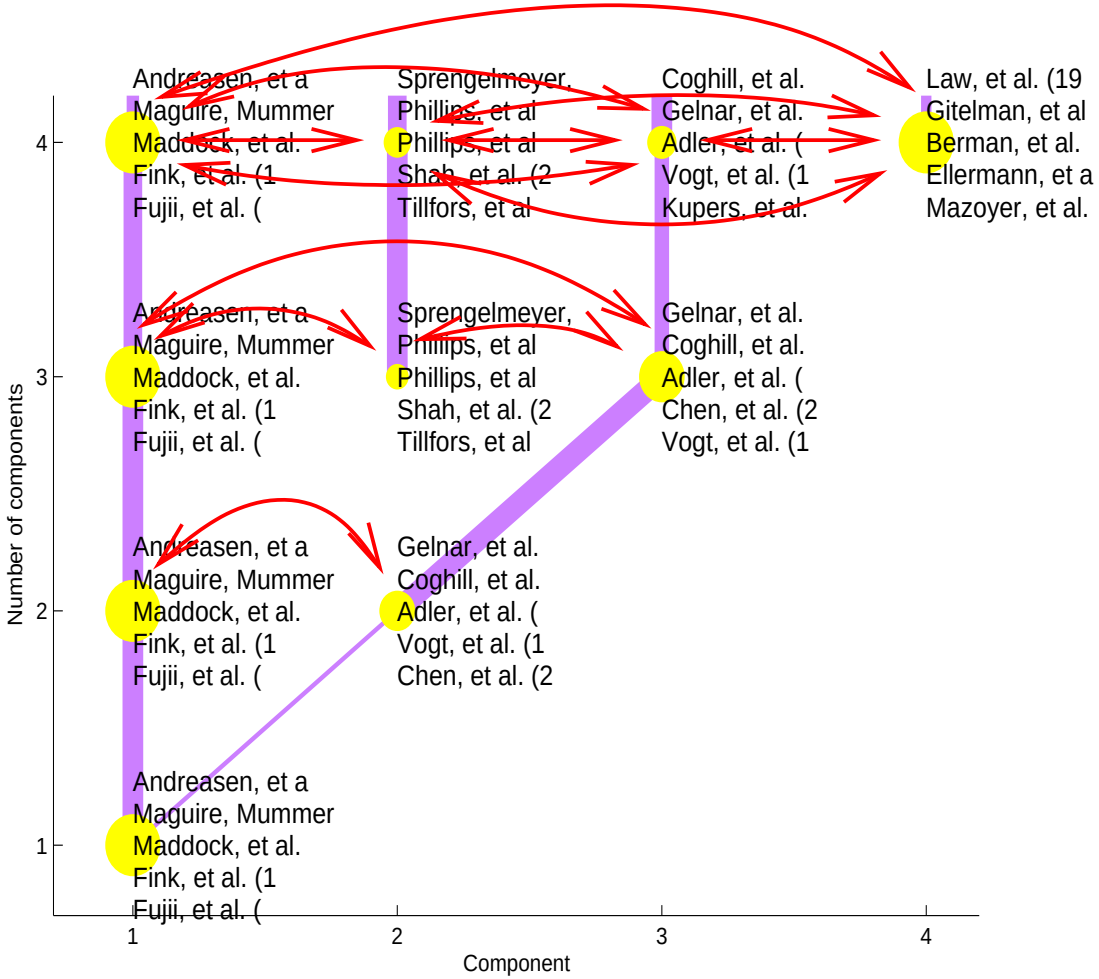
$$\mathbf{H}_{kp} \leftarrow \mathbf{H}_{kp} \frac{\left(\mathbf{W}^\mathsf{T}\mathbf{X}\right)_{kp}}{\left(\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{H}\right)_{kp}} \tag{3}$$

$$\mathbf{W}_{nk} \leftarrow \mathbf{W}_{nk} \frac{\left(\mathbf{X}\mathbf{H}^\mathsf{T}\right)_{nk}}{\left(\mathbf{W}\mathbf{H}\mathbf{H}^\mathsf{T}\right)_{nk}}. \tag{4}$$

Step 2: “Medial temporal lobe” NMF result



Step 3: Test spatial distribution



Extract locations from grouped papers.

Test if the spatial distribution of locations for a group is different from the distribution from an other group.

All possible tests within a level of non-negative matrix factorization are performed.

Step 3: Tests on "segregation"

Two-sample Hotelling's T^2 test follows an F-distribution if multivariate Gaussian distributions are assumed

$$\frac{M_1 M_2 (M-P-1)}{M(M-2)P} D^2 \sim F_{P,M-P-1}. \tag{5}$$

The Mahalanobis distance is computed as

$$D^2 = (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)^{\mathsf{T}} \mathbf{S}_u^{-1} (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)\,, \tag{6}$$

with the covariance $\mathbf{S}_u$ found as

$$\mathbf{S}_u = (M_1 \mathbf{S}_1 + M_2 \mathbf{S}_2)/(M-2), \tag{7}$$

$\bar{\mathbf{z}}_1$ and $\mathbf{S}_1$ are the mean and covariance for one set of Talairach coordinates

Step 3: Convex hull peeling

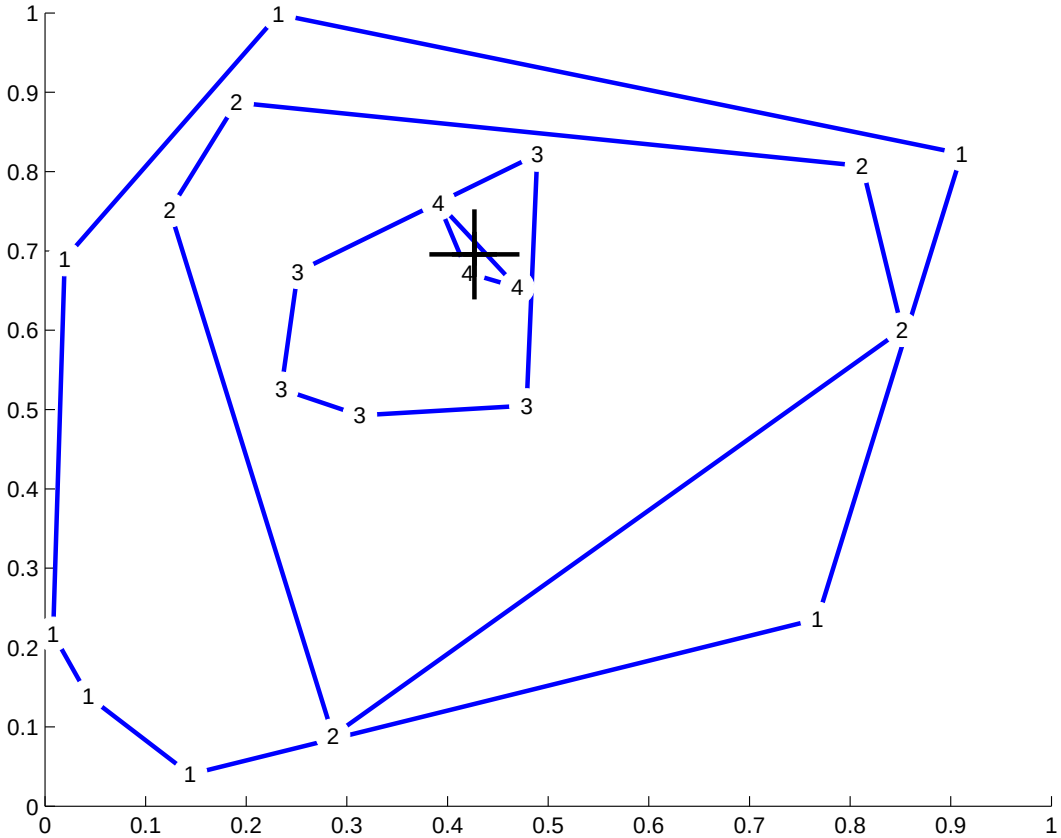


Figure 3: Convex hull peeling

Perhaps the Gaussian assumptions are not appropriate for sets of locations.

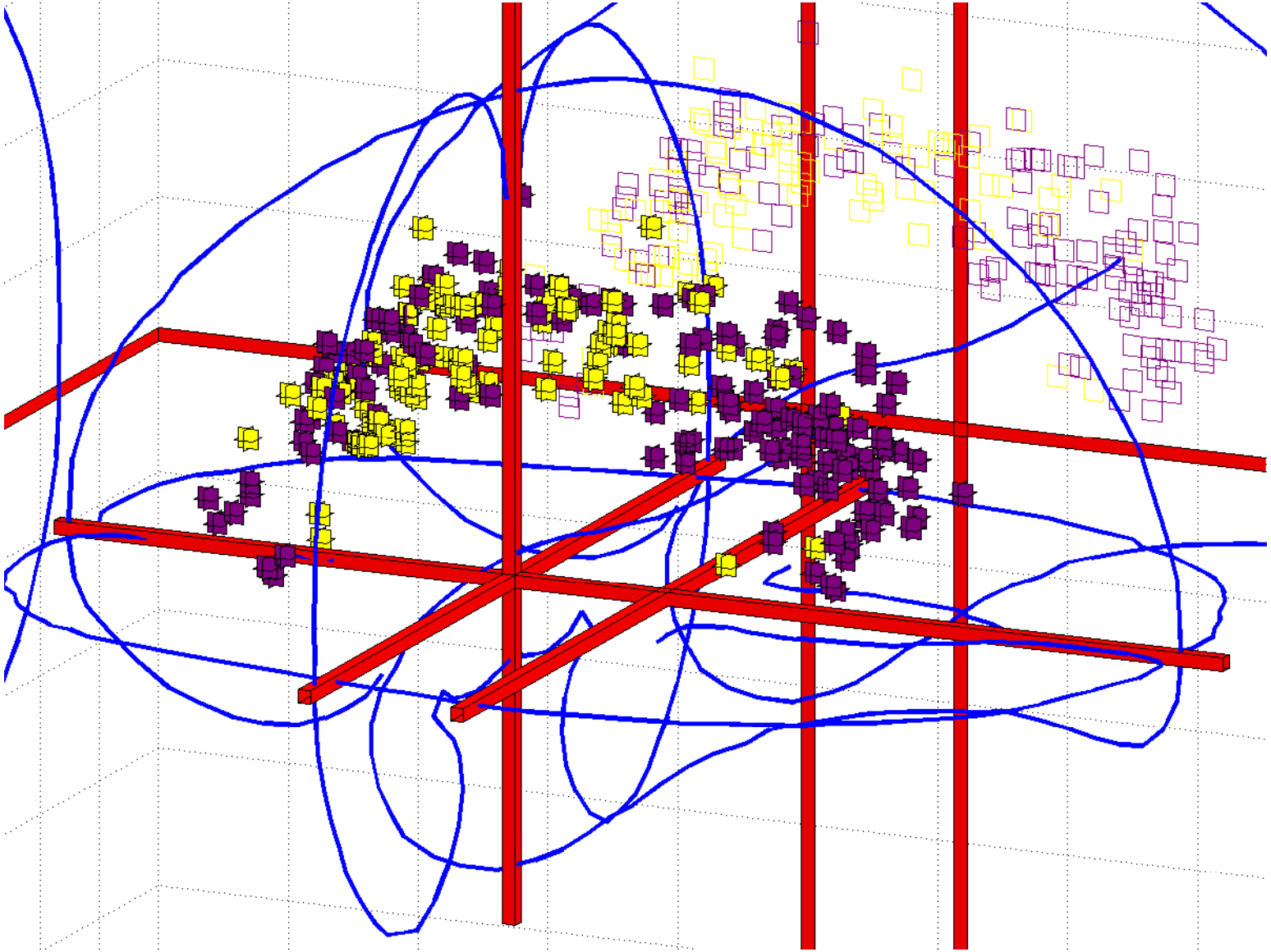
Convex hull peeling centroid (Barnett, 1976) is a robust multivariate estimate of the centroid.

Monte Carlo permutation test on the distance between centroids.

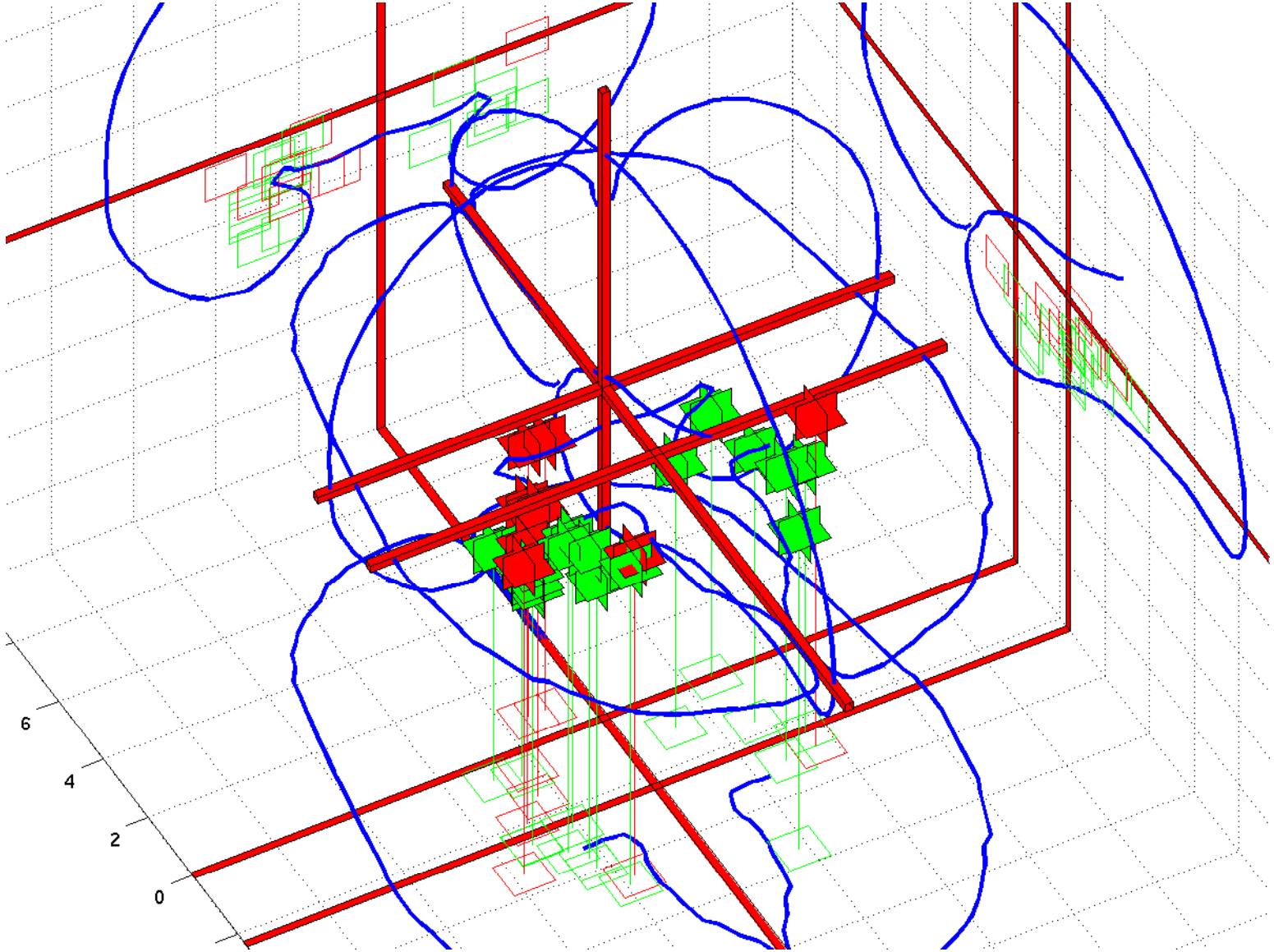
Step 4: Combined results

```
  #        P-values      (First set) - (Second set) - Brain region
-----------------------------------------------------------------------
  1   0.000 0.000 0.000  (pain, painful, 211) - (visual, eye, 565) - Cerebral Cortex (14)
  2   0.000 0.000 0.000  (pain, painful, 230) - (visual, eye, 587) - Telencephalon (13)
  3   0.000 0.000 0.002  (pain, painful, 97) - (memory, retrieval, 141) - Cingulate gyrus (4)
  4   0.000 0.002 0.003  (pain, painful, 269) - (visual, eye, 607) - Forebrain (12)
  5   0.000 0.005 0.000  (expressions, facial, 15) - (recognition, humans, 10) - Amygdala and Hippocampus (
  6   0.000 0.004 0.005  (memory, retrieval, 22) - (pain, painful, 5) - Anterior cingulate gyrus (8)
  7   0.000 0.004 0.005  (memory, retrieval, 22) - (pain, painful, 5) - Posterior medial prefrontal cortex
  8   0.000 0.006 0.000  (ear, musical, 5) - (retrieval, faces, 13) - Right frontal lobe (82)
  9   0.000 0.000 0.006  (pain, painful, 100) - (memory, retrieval, 159) - Limbic gyrus (125)
 10   0.009 0.002 0.000  (memory, episodic, 27) - (motor, sensorimotor, 20) - Cerebellum (32)
 11   0.001 0.004 0.011  (artefacts, categorization, 2) - (memory, word, 28) - Precentral gyrus (68)
 12   0.000 0.001 0.015  (pain, painful, 71) - (words, memory, 45) - Limbic lobe (2)
 13   0.000 0.000 0.016  (pain, painful, 79) - (memory, episodic, 72) - Prefrontal cortex (22)
 14   0.000 0.000 0.024  (artefacts, categorization, 7) - (verbal, visual, 16) - Middle frontal gyrus (148)
 15   0.000 0.002 0.029  (memory, episodic, 26) - (pain, painful, 5) - Medial prefrontal cortex (55)
 16   0.000 0.031 0.002  (musical, ear, 6) - (artefacts, decision, 10) - Right temporal lobe (86)
 17   0.002 0.037 0.009  (pain, noxious, 25) - (motor, visual, 20) - Insula (67)
 18   0.000 0.042 0.000  (memory, retrieval, 34) - (pain, painful, 25) - Posterior cingulate gyrus (5)
 ...
```

Step 4: “Cingulate gyrus”



Step 4: “Medial temporal lobe”



Summary

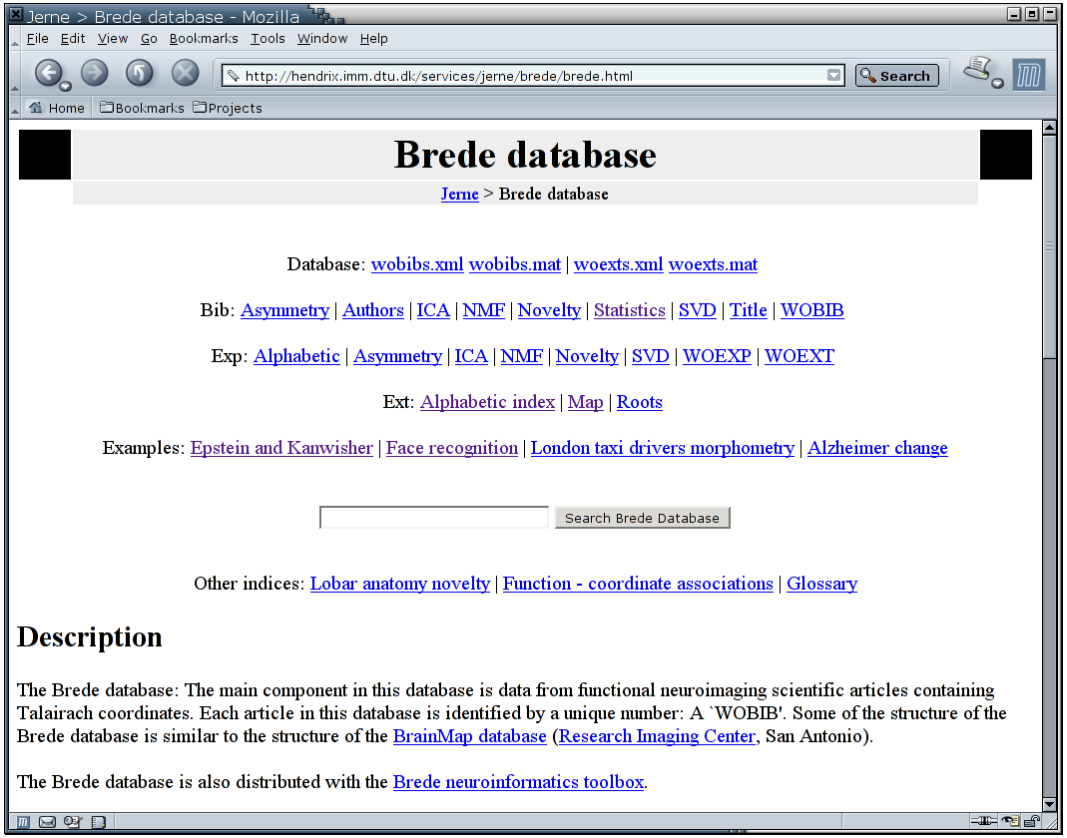


Figure 4: Brede Database on the Internet

Neuroinformatics database with brain region taxonomy.

Automated analysis combining: Kernel density estimation, non-negative matrix factorization, multivariate test.

313× upscaling of previous study on just a single region (Nielsen et al., 2005).

References

Balslev, D., Nielsen, F. Å., Paulson, O. B., and Law, I. (2005). Right temporoparietal cortex activation during visuo-proprioceptive conflict. *Cerebral Cortex*, 15(2):166–169. PMID: 152384438. WOBIB: 128. http://cercor.oupjournals.org/cgi/content/abstract/15/2/166?etoc.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A*, 139:319–354.

Bowden, D. M. and Martin, R. F. (1995). NeuroNames brain hierarchy. *NeuroImage*, 2(1):63–84. PMID: 9410576. ISSN 1053-8119.

Bush, G., Luu, P., and Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6):215–222. PMID: 10827444. http://www.nmr.mgh.harvard.edu/BushLab/Bush_2000_TICS_CingReview.pdf.

Drury, H. A., ad C. H. Anderson, D. C. V., Lee, C. W., Coogan, T. A., and Lewis, J. W. (1996). Computerized mappings of the cerebral cortex: A multiresolution flattening method and a surface-based coordinate system. *Journal of Cognitive Neuroscience*, 8(1):1–28. PMID: 11539144.

Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software — Practice and Experience*, 30(11):1203–1234. http://www.graphviz.org-/Documentation/GN99.pdf. ISSN 00380644.

Jernigan, T. L., Ostergaard, A. L., Law, I., Svarer, C., Gerlach, C., and Paulson, O. B. (1998). Brain activation during word identification and word recognition. *NeuroImage*, 8(1):93–105. PMID: 9698579. WOBIB: 35.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791. PMID: 10548103.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems*

13: Proceedings of the 2000 Conference, pages 556–562, Cambridge, Massachusetts. MIT Press. http://hebb.mit.edu/people/seung/papers/nmfconverge.pdf. CiteSeer: http://citeseer.ist.psu.edu/-lee00algorithms.html.

Mai, J. K., Assheuer, J., and Paxinos, G. (1997). *Atlas of the Human Brain*. Academic Press, San Diego, California. ISBN 0124653618.

Nielsen, F. Å. (2003). The Brede database: a small database for functional neuroimaging. *NeuroImage*, 19(2). http://208.164.121.55/hbm2003/abstract/abstract906.htm. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY. Available on CD-Rom.

Nielsen, F. Å., Balslev, D., and Hansen, L. K. (2005). Mining the posterior cingulate: Segregation between memory and pain component. *NeuroImage*, 27(3):520–532. DOI: 10.1016/j.neuroimage.2005.04.034.

Penrose, L. S. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109:53–57.

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., and Gabrieli, J. D. E. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage*, 10(1):15–35. PMID: 10385578. DOI: 10.1006/nimg.1999.0441. WOBIB: 178. http://www.sciencedirect.com/science/article/B6WNP-45FCP4S-19/2/22ce9903da63d3db7f801d23303f08d6. An fMRI study and a review of studies with Talairach coordinates of the activation distribution in the left inferior prefrontal cortex. Abstract/concrete judgement, syllable counting and case judgment are used as tasks.

Steel, J. D. and Lawrie, S. M. (2004). Segregation of cognitive and emotional function in the prefrontal cortex: a stereotactic meta-analysis. *NeuroImage*, 21(3):868–875. PMID: 15006653. DOI: 10.1016/j.neuroimage.2003.09.066. A mathematical meta-analysis of Talairach coordinates in emotion and cognitive studies in the prefrontal cortex. Only one coordinate is extracted from each study. A spline-based warp transformation is applied on the coordinates so new axes correspond ot the shape of corpus callosum. A variety of tests are used on the coordinates: 2D Kolmogorov-Smirnov,

MANOVA/canonical correlation analysis with weighted covariance matrix, one-dimension tests, and significance is determined by resampling. It is found that emotion coordinates tend to the inferior anterior part of the medial prefrontal cortex while cognitive tend to the posterior superior part in this region. On the lateral surface of the prefrontal cortex the emotion coordinates appear predominately in the inferior part.