Text and spatial data mining in the Brede Database

Finn Årup Nielsen

Neurobiology Research Unit, Rigshospitalet;
Informatics and Mathematical Modelling
Technical University of Denmark

March 15, 2005

Abstract

The Brede Database records information from published human brain mapping studies. With brain scanners these studies investigate the spatial distribution of neural activity as modulated by different kinds of mental processes. The information contained in the database, is, e.g., title and abstract of the paper describing the study, as well as 3-dimensional coordinates representing the important modes in the spatial distribution. Papers with coordinates present in a specified brain area can be extracted from the database and automatically grouped. The coordinates in the grouped papers are extracted and it is tested whether the coordinates are spatially clustered in subregions of the area based on the group structure. This allows for more or less automated data-mining for segregation in the human brain. The method uses kernel density estimation with cross-validation and Hotelling's T2 test on the coordinates, and non-negative matrix factorization on the texts in a vectorial "bag-of-words" representation.

Human brain mapping

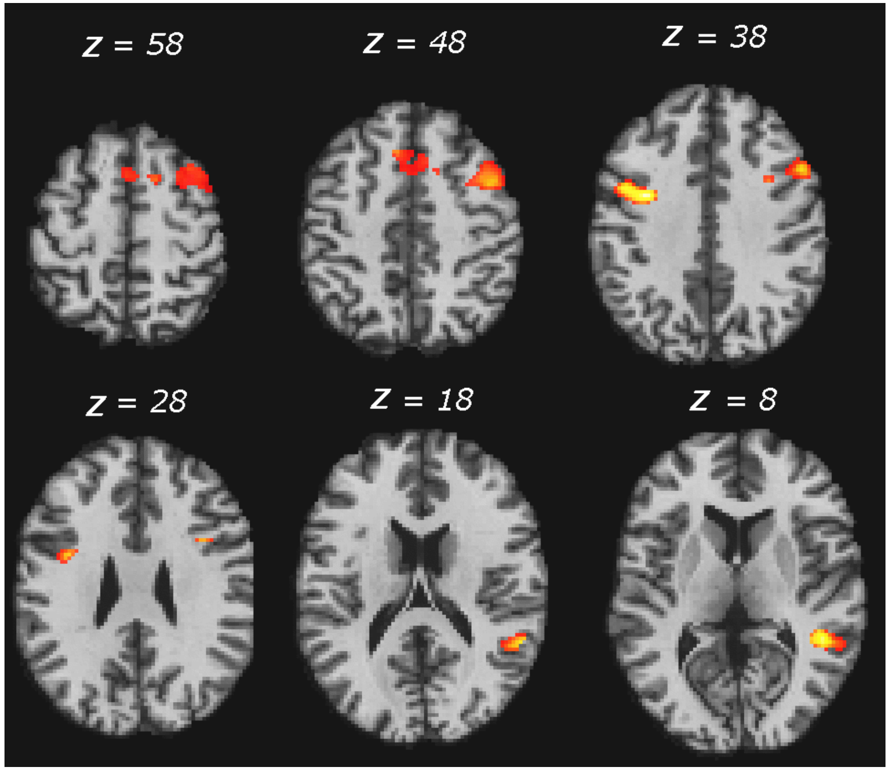


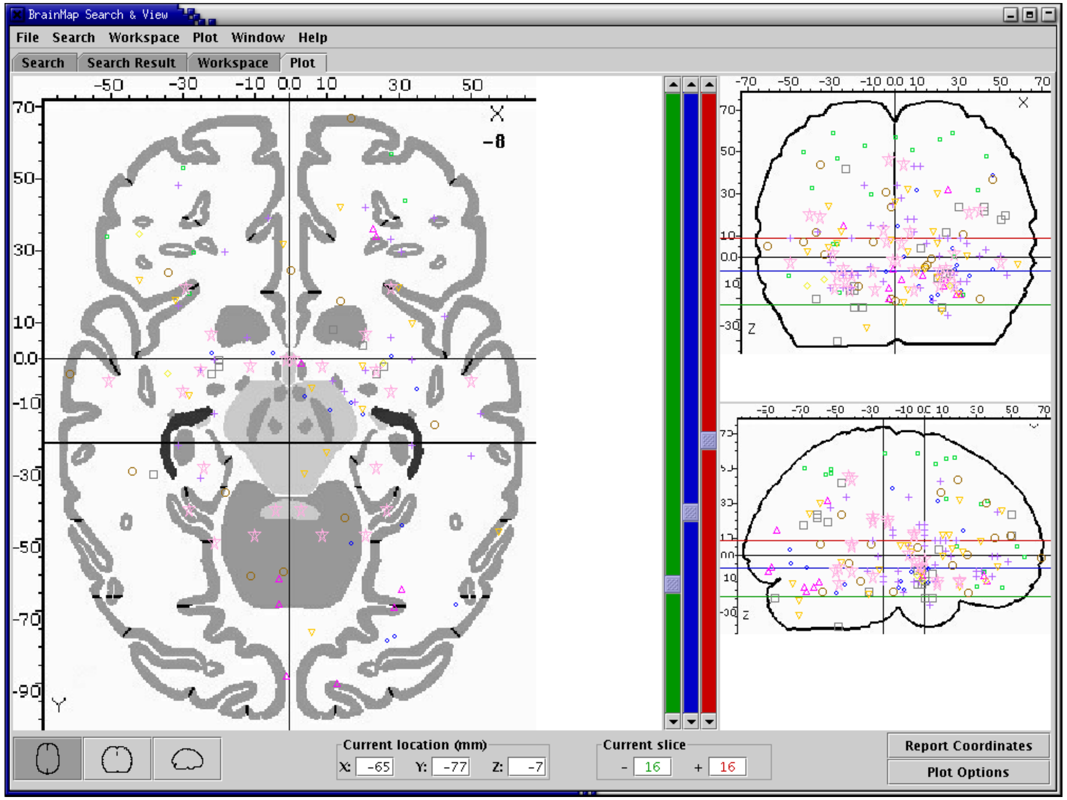
Figure 1: Results from a human brain mapping study (Balslev et al., 2005, figure 2). Hot color-mapped functional results from an fMRI study on top of a gray-scaled "single subject" reference brain scan.

Positron emission tomography (PET) or functional magnetic resonance (fMRI) brain scans of the human brain while subjects are engaged in different mental processes.

Result represented in the literature with lists of 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

BrainMap database



Current location (mm)
X: -65 Y: -77 Z: -7

Current slice
- 16 + 16

Figure 2: Screen shot of a graphical user interface to the BrainMap database with locations plotted after a search for experiments on olfaction.

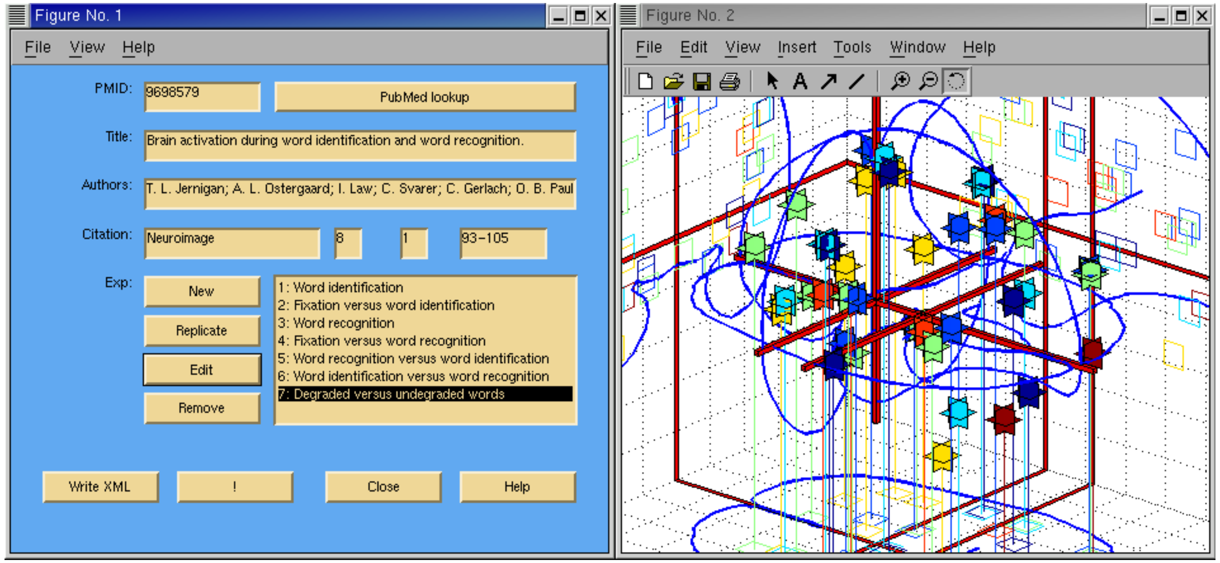
The BrainMap database (Fox and Lancaster, 2002; Fox and Lancaster, 1994).

On of the earliest dedicated neuroscience databases.

Contains over 500 studies from published scientific articles.

Programs for data entry, searching and visualization.

Brede database



PMID: 9698579

Title: Brain activation during word identification and word recognition.

Authors: T. L. Jernigan; A. L. Ostergaard; I. Law; C. Svarer; C. Gerlach; O. B. Paul

Citation: Neuroimage 8 1 93–105

Exp:
1: Word identification
2: Fixation versus word identification
3: Word recognition
4: Fixation versus word recognition
5: Word recognition versus word identification
6: Word identification versus word recognition
7: Degraded versus undegraded words

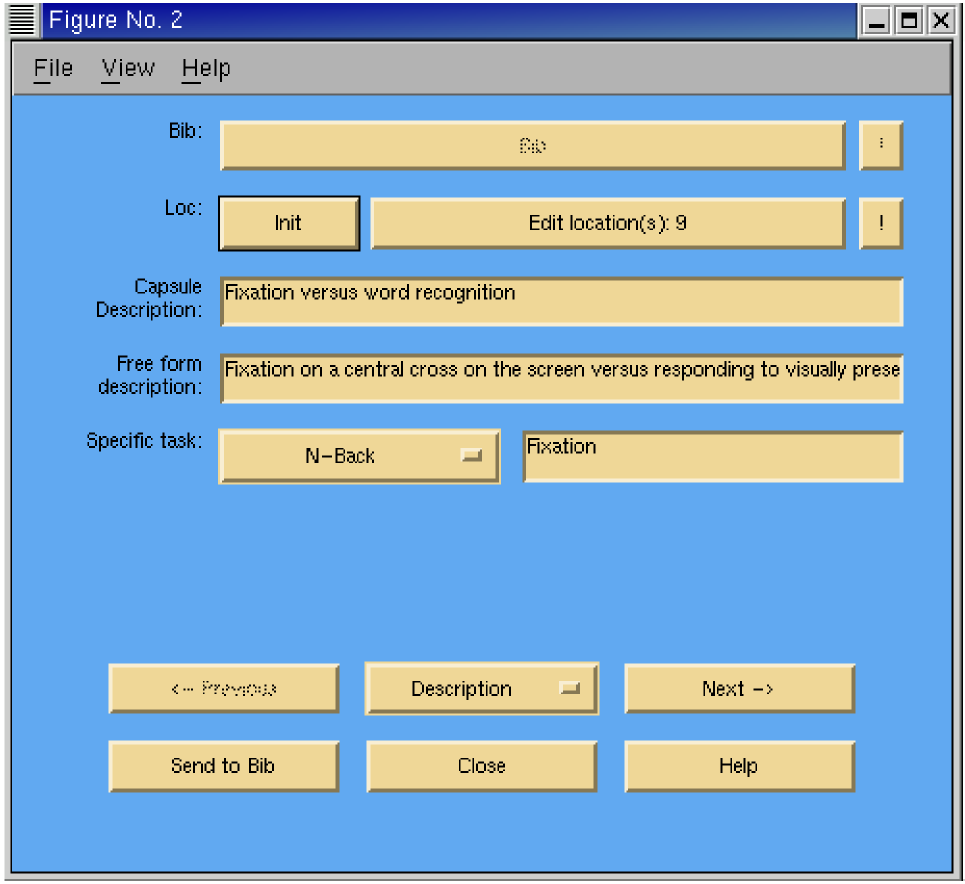
Figure 3: Screenshot of main window of Matlab program for data entry of one of the studies in the Brede database (Jernigan et al., 1998).

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

Presently contains 152 papers, 470 experiments and 3252 locations.

Each experiment is labeled with the specific function under investigation, e.g., response to “hot pain”, “cold pain” stimuli and “episodic memory retrieval”.

Entry of information in the Brede database



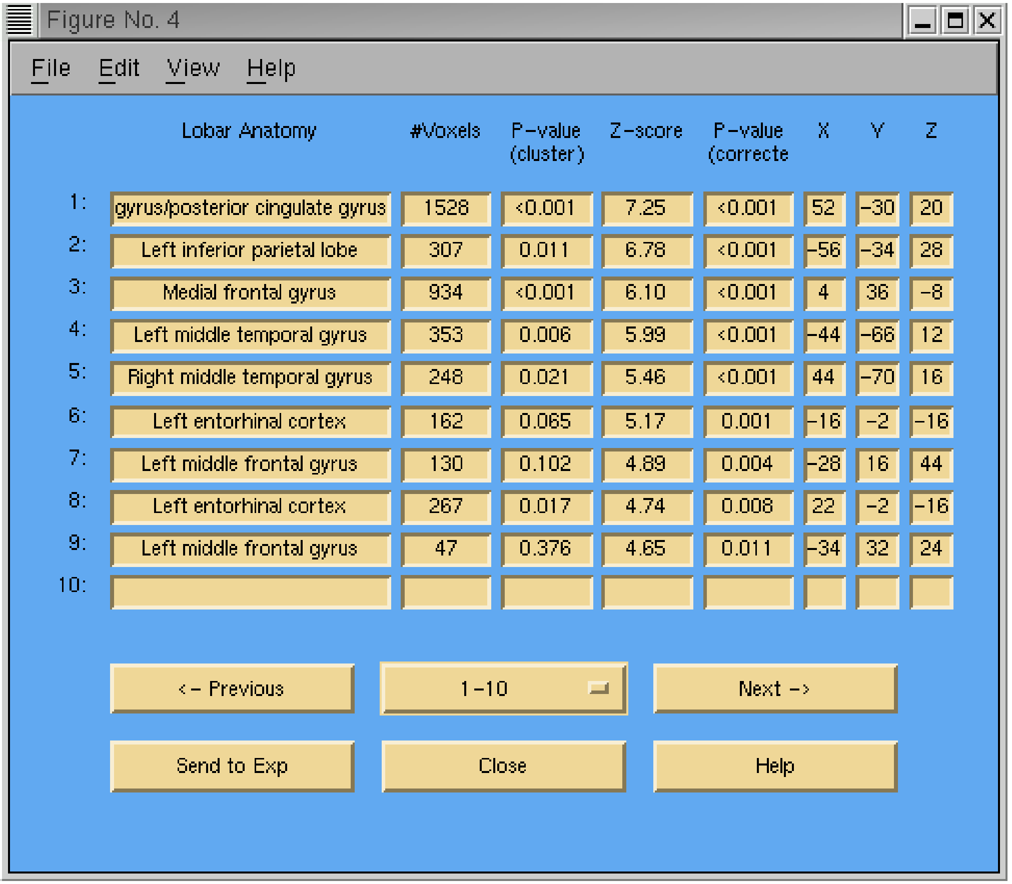
Bib:

Loc: Edit location(s): 9

Capsule Description: Fixation versus word recognition

Free form description: Fixation on a central cross on the screen versus responding to visually prese

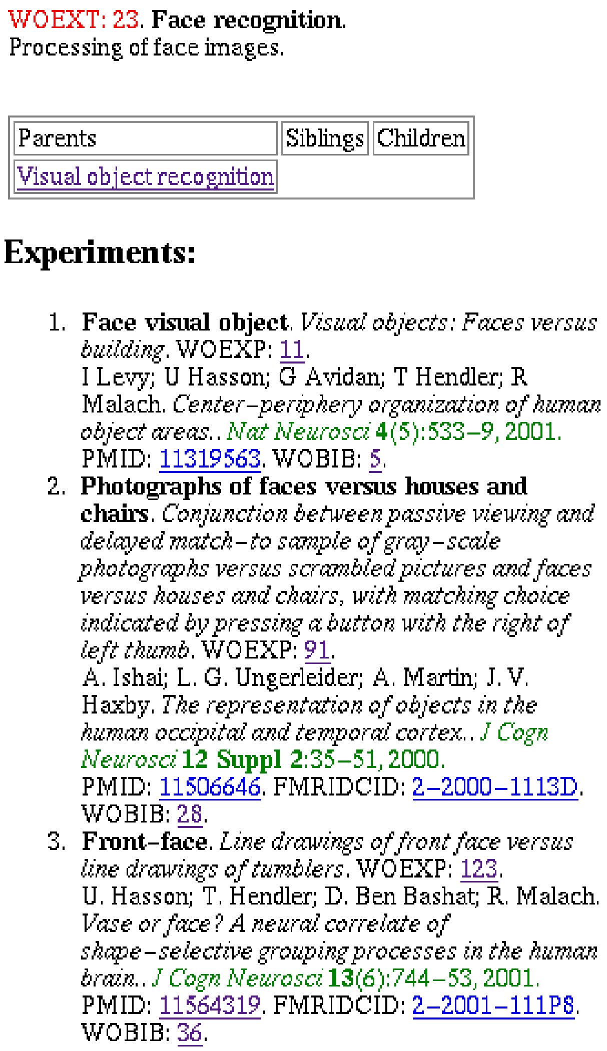
Specific task: N-Back Fixation



	Lobar Anatomy	#Voxels	P-value (cluster)	Z-score	P-value (correcte	X	Y	Z
1:	gyrus/posterior cingulate gyrus	1528	<0.001	7.25	<0.001	52	-30	20
2:	Left inferior parietal lobe	307	0.011	6.78	<0.001	-56	-34	28
3:	Medial frontal gyrus	934	<0.001	6.10	<0.001	4	36	-8
4:	Left middle temporal gyrus	353	0.006	5.99	<0.001	-44	-66	12
5:	Right middle temporal gyrus	248	0.021	5.46	<0.001	44	-70	16
6:	Left entorhinal cortex	162	0.065	5.17	0.001	-16	-2	-16
7:	Left middle frontal gyrus	130	0.102	4.89	0.004	-28	16	44
8:	Left entorhinal cortex	267	0.017	4.74	0.008	22	-2	-16
9:	Left middle frontal gyrus	47	0.376	4.65	0.011	-34	32	24
10:								

Each location is primarily represented by the 3D-coordinate and a textual field indicating the brain region

Supervised labeling

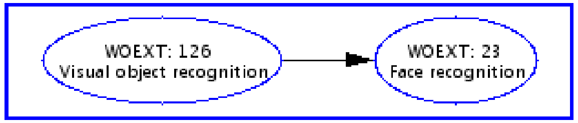


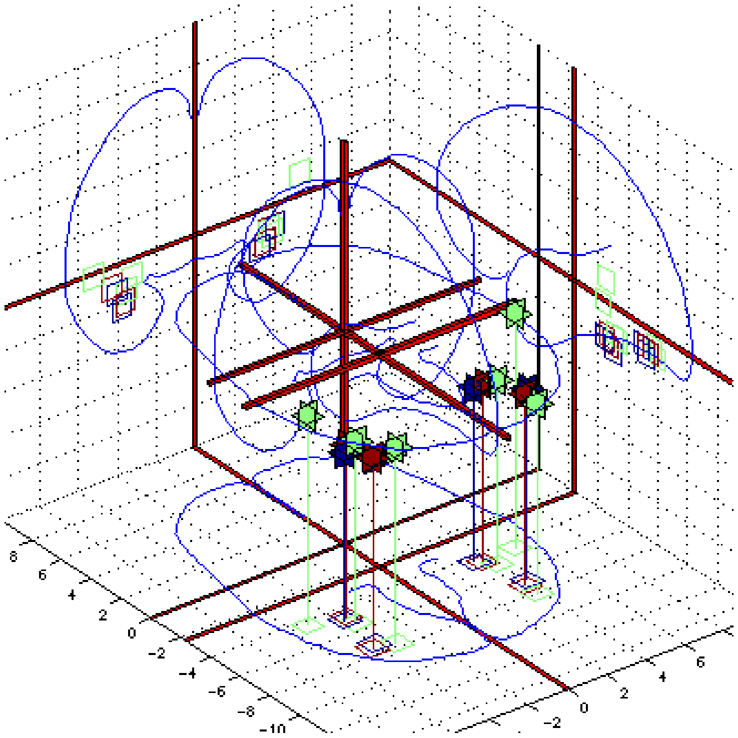
WOEXT: 23. **Face recognition**.
Processing of face images.

Parents	Siblings	Children
Visual object recognition		

Experiments:

1. **Face visual object**. *Visual objects: Faces versus building*. WOEXP: 11.
I Levy; U Hasson; G Avidan; T Hendler; R Malach. *Center–periphery organization of human object areas.*. *Nat Neurosci* **4**(5):533–9, 2001.
PMID: 11319563. WOBIB: 5.
2. **Photographs of faces versus houses and chairs**. *Conjunction between passive viewing and delayed match–to sample of gray–scale photographs versus scrambled pictures and faces versus houses and chairs, with matching choice indicated by pressing a button with the right of left thumb*. WOEXP: 91.
A. Ishai; L. G. Ungerleider; A. Martin; J. V. Haxby. *The representation of objects in the human occipital and temporal cortex.*. *J Cogn Neurosci* **12 Suppl 2**:35–51, 2000.
PMID: 11506646. FMRIDCID: 2–2000–1113D.
WOBIB: 28.
3. **Front–face**. *Line drawings of front face versus line drawings of tumblers*. WOEXP: 123.
U. Hasson; T. Hendler; D. Ben Bashat; R. Malach. *Vase or face? A neural correlate of shape–selective grouping processes in the human brain.*. *J Cogn Neurosci* **13**(6):744–53, 2001.
PMID: 11564319. FMRIDCID: 2–2001–111P8.
WOBIB: 36.





Example with “Face recognition” studies in a “corner cube” visualization.

The “expert” label added during database entry can provide the grouping structure.

Statistical tests can be constructed to measure whether the spatial distribution is “clustered” (Turkeltaub et al., 2002; Nielsen, 2004).

This study

Focus on specific brain area.

No expert label: Get context from abstract text

Determine themes of the brain area

Determine whether specific themes are spatially clustered in the brain area.

Identifying studies

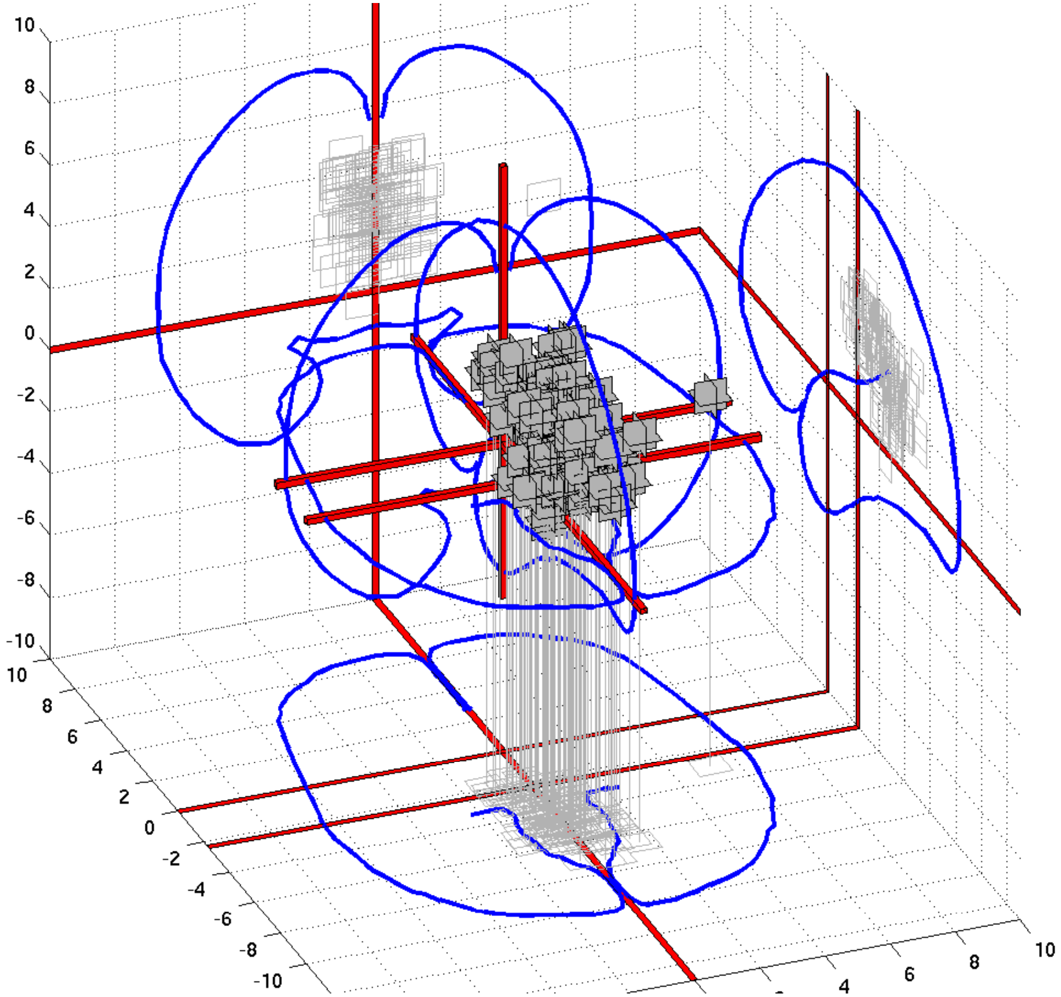
Simple SQL-like command in Matlab to find locations:

```
Lpc = brede_struct_select(L, 'where', { 'lobarAnatomy' 'findstri', ...
      'posterior cingulate' });
```

It finds locations where the “lobarAnatomy” field matches the string “posterior cingulate”.

Presently 116 locations are identified.

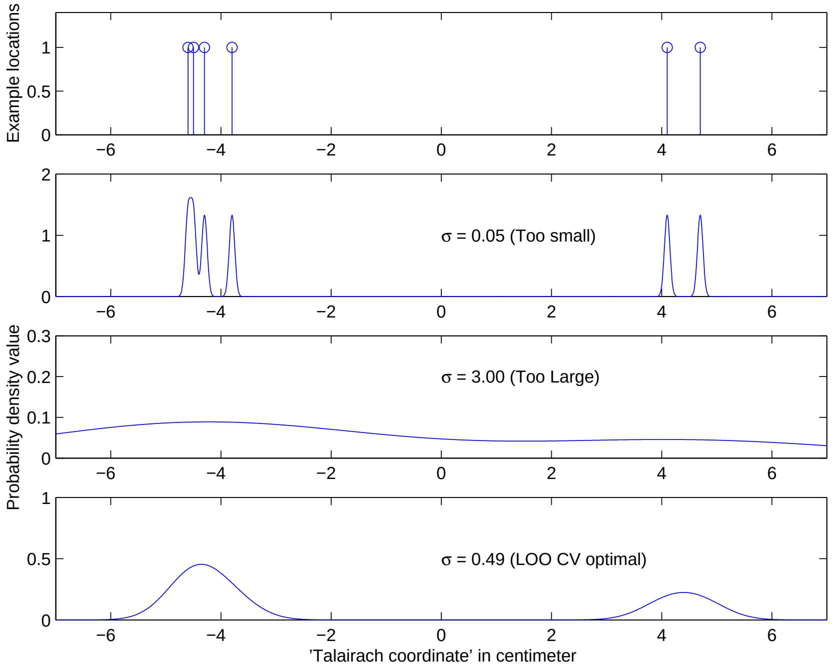
Identified locations



Corner cube visualization of 116 “posterior cingulate” locations

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

Kernel density estimators for locations



Regard the "locations" as being generated from a distribution $p(\mathbf{x})$, where $\mathbf{x}$ is in 3D Talairach space (Fox et al., 1997).

Kernel methods (N kernels centered on each location: $\boldsymbol{\mu}_n$) with homogeneous Gaussian kernel in 3D Talairach space $\mathbf{x}$

$$\widehat{p}(\mathbf{x}) = \frac{(2\pi\sigma^2)^{-3/2}}{N} \sum_n^N e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_n)^2}$$

σ^2 fixed ($\sigma = 1\text{cm}$) or optimized with leave-one-out cross-validation (Nielsen and Hansen, 2002).

Handling outliers

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.

For the further analysis: Include all papers with contain on or more of these 184 locations.

Presently 79 papers are found.

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	. . .
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Representation of the abstracts of the papers in a bag-of-words matrix: (abstract × words)-matrix.

Each element counts of the frequency of a word occurring in an abstract text.

Eliminated stop words

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)

Scaling

Element-wise square root scaling . . . (Penrose, 1946)

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)

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

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

Results from abstract grouping

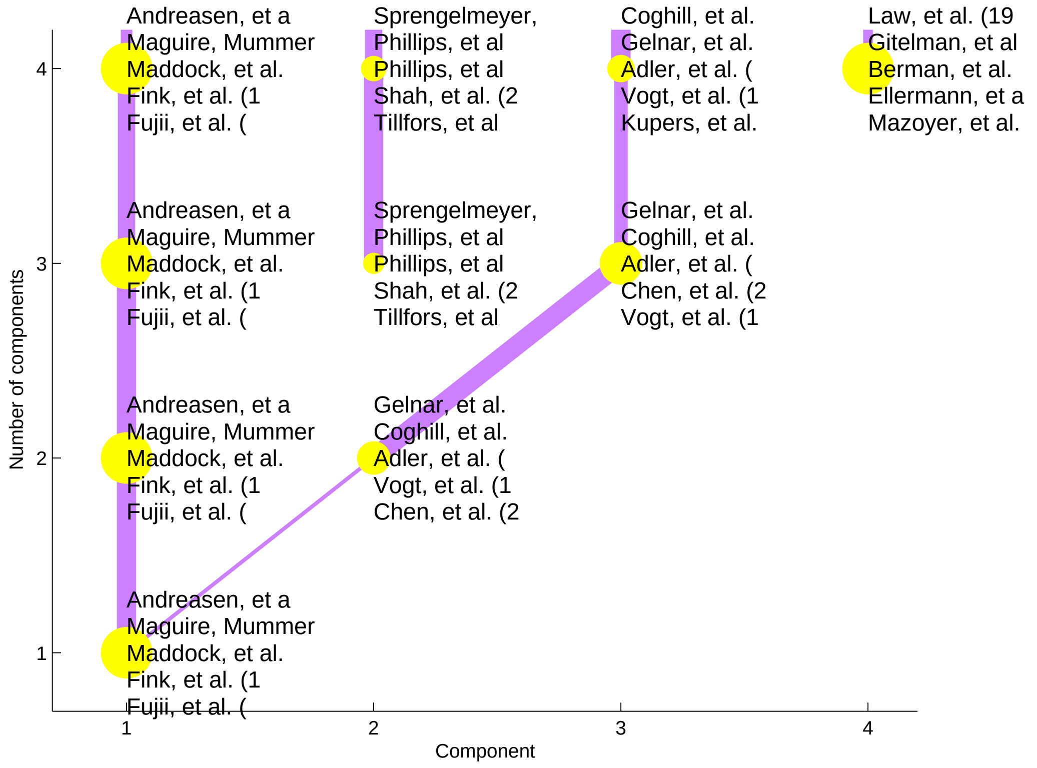


Figure 4: Closeup of "clustered" abstracts.

Hierarchical NMF with varying $K = 1 \ldots \left\lceil \sqrt{\min(N, P)} \right\rceil$

Abstracts with highest score on $\mathbf{w}_k$ shown

Winner-takes-all function applied on $\mathbf{W}$

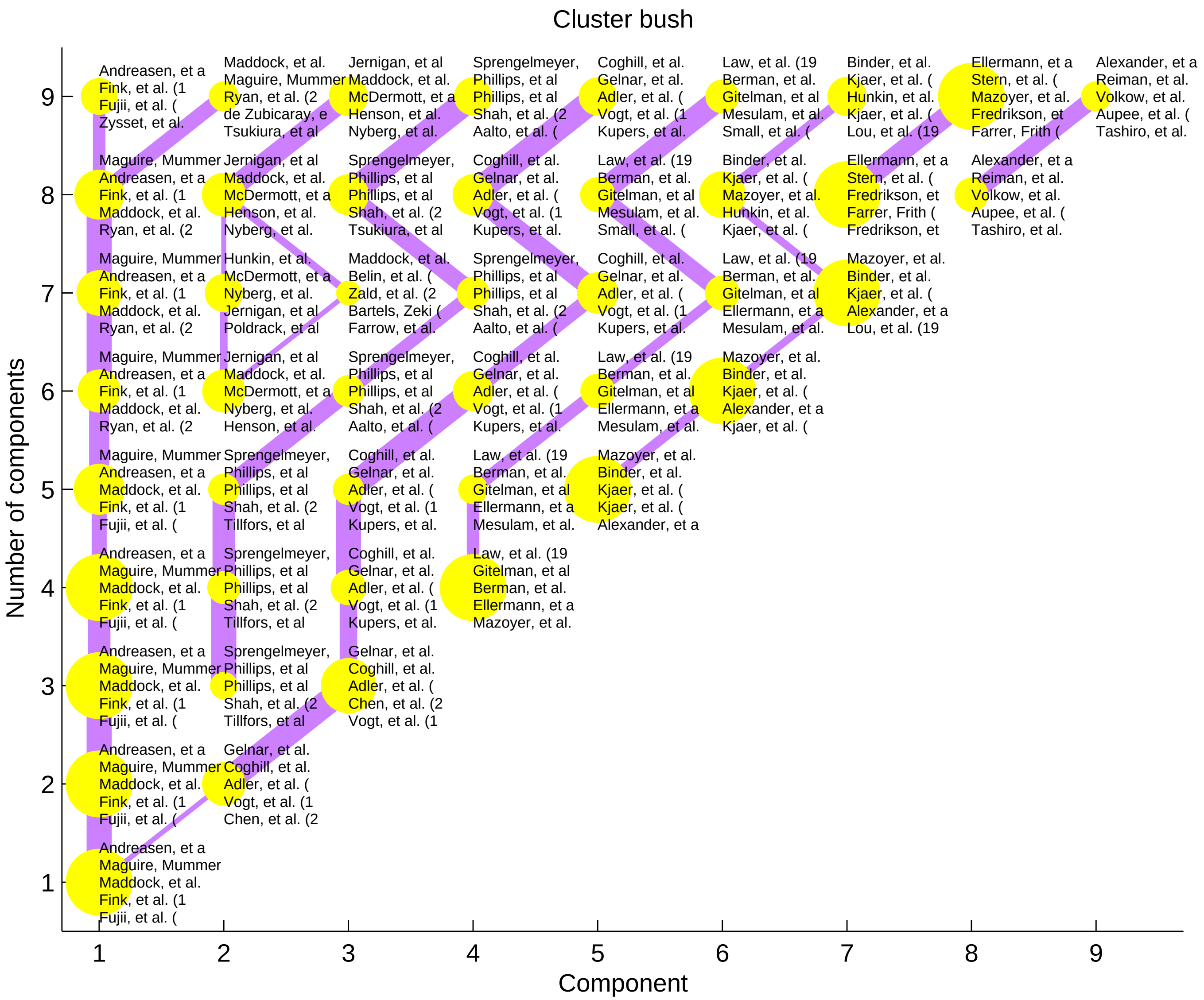
Nodes are ordered with

$$\mathbf{C} = \mathbf{H}_K \mathbf{H}_{K+1}^{\mathsf{T}} \qquad (5)$$

Line width determined by

$$\text{width} \propto c_{ij}^2 / \max_{ij}(\mathbf{C}) \qquad (6)$$

Results from abstract grouping



Results from word grouping

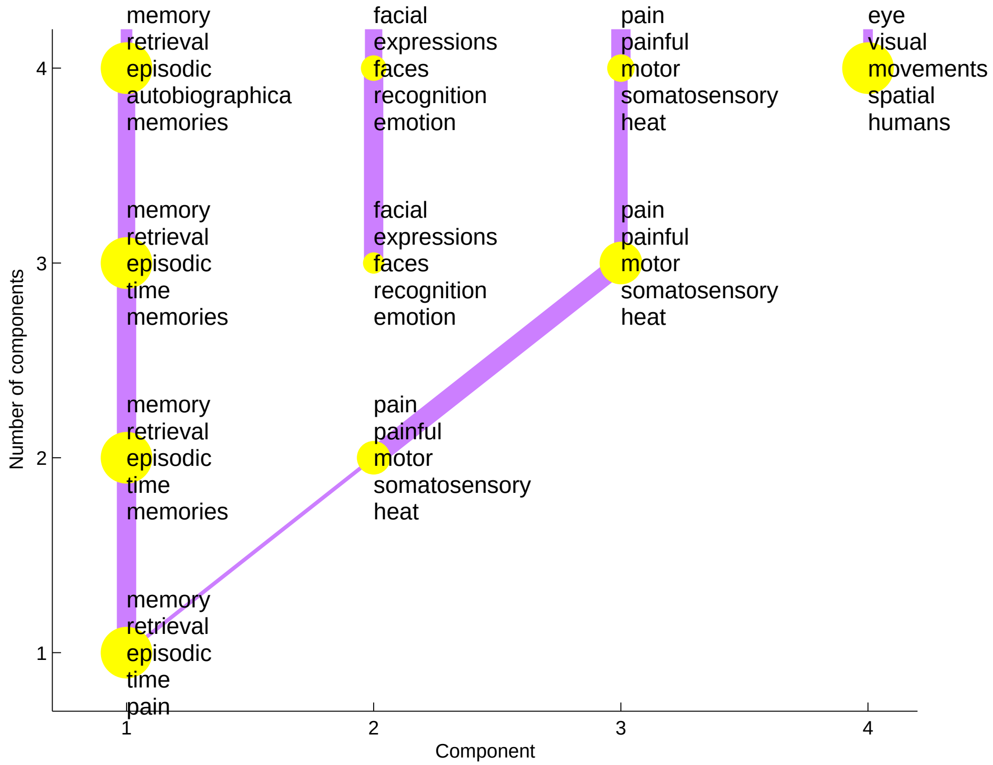


Figure 5: Closeup of "clustered" words.

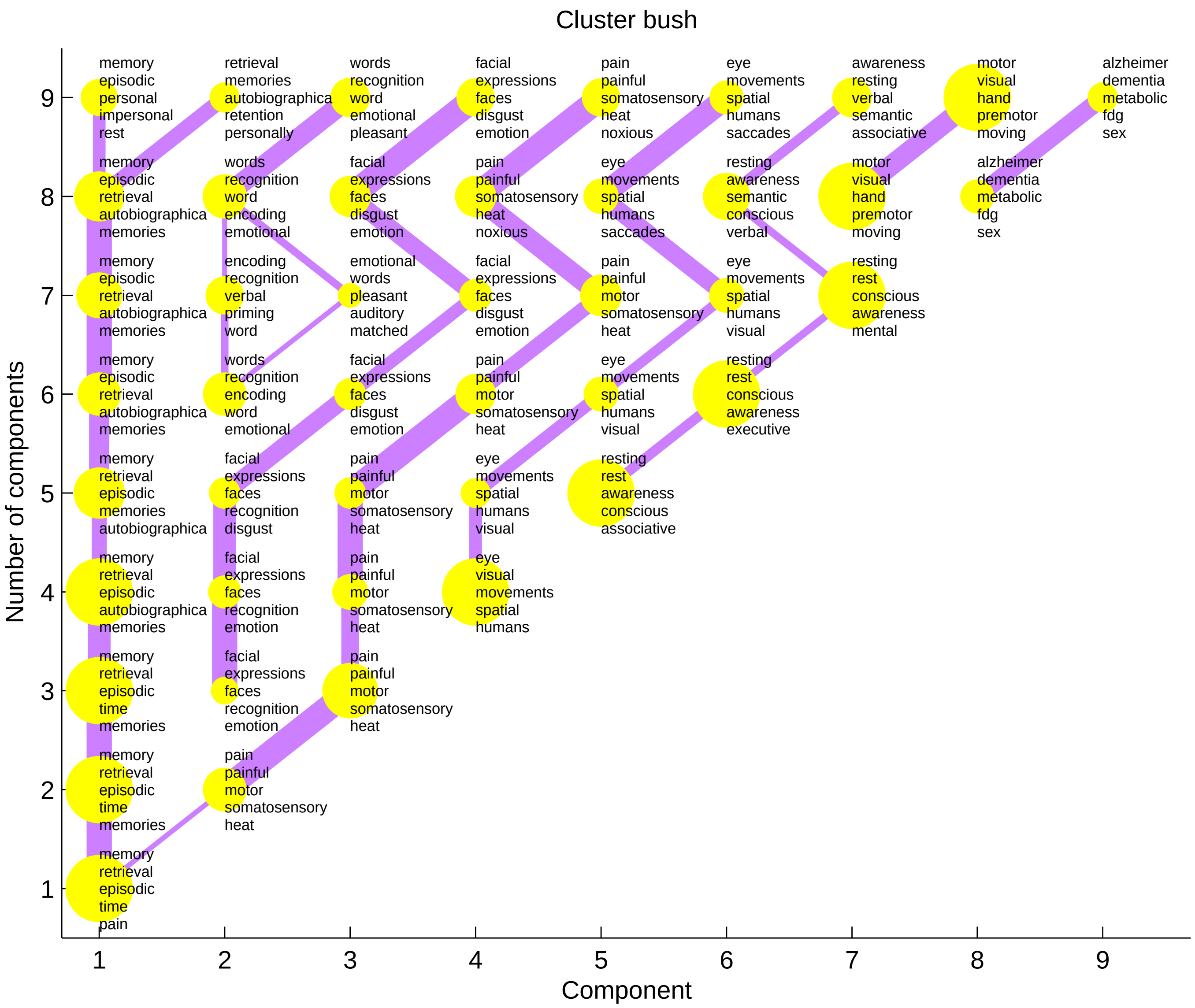
Hierarchical NMF with the $\mathbf{H}_\mathbf{K}$s.

Words with highest score on $\mathbf{h}_k$ shown

Winner-takes-all function applied on each $\mathbf{H}_K$

Major themes: *Memory*, *pain*, emotional *facial* expressions, (*visual*) eye movements

Results from word grouping



Relation to “manual reviews”

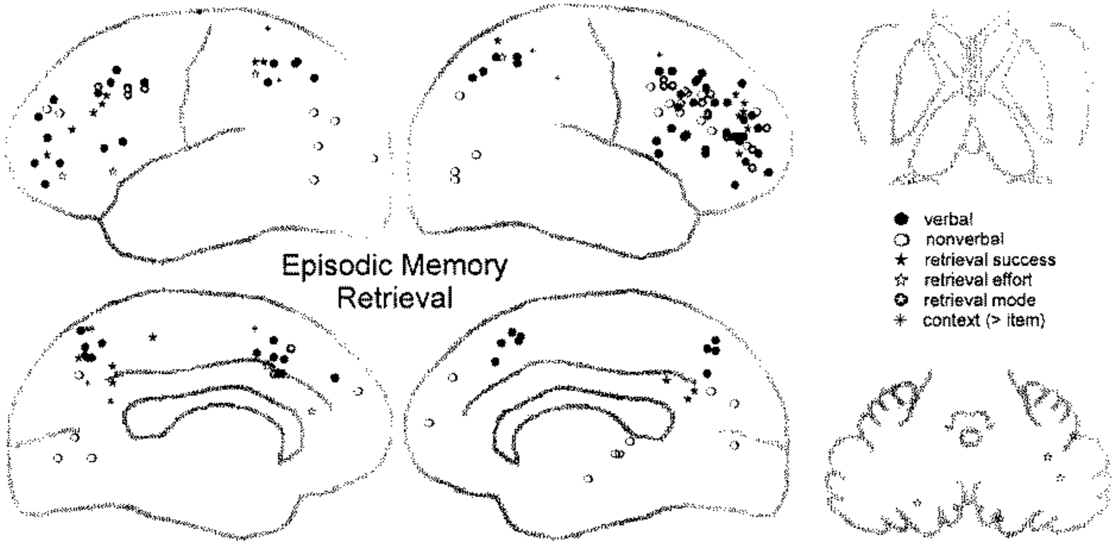
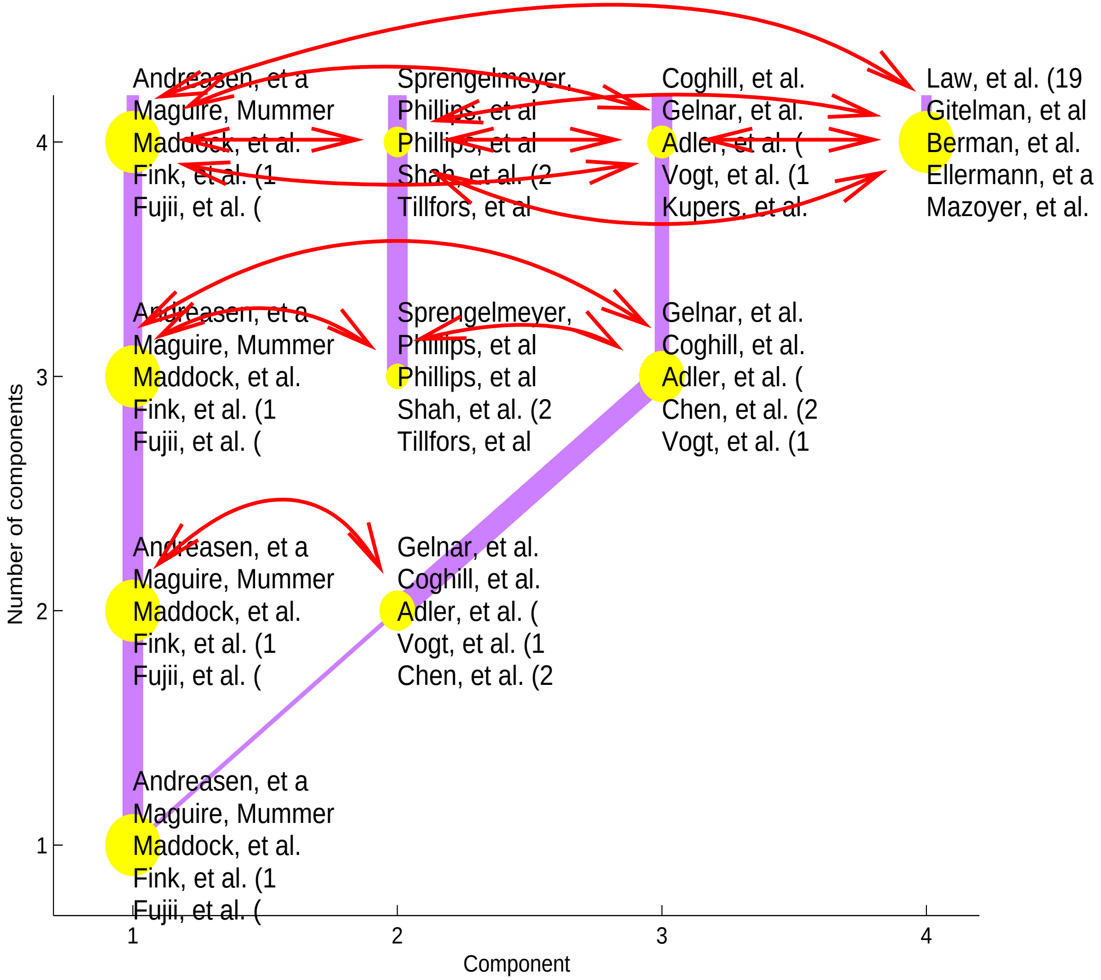


Figure 6: (Cabeza and Nyberg, 2000, figure 10).

Memory is the main component in the automated analysis

Successful episodic memory retrieval found as the most important cognitive function for PCC in a large review (Cabeza and Nyberg, 2000).

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.

Tests on “segregation”

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

$$\frac{M_1 M_2 (M - P - 1)}{M(M-2)P} D^2 \sim F_{P, M-P-1}. \tag{7}$$

The Mahalanobis distance is computed as

$$D^2 = (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)^{\mathsf{T}} \mathbf{S}_u^{-1} (\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2), \tag{8}$$

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

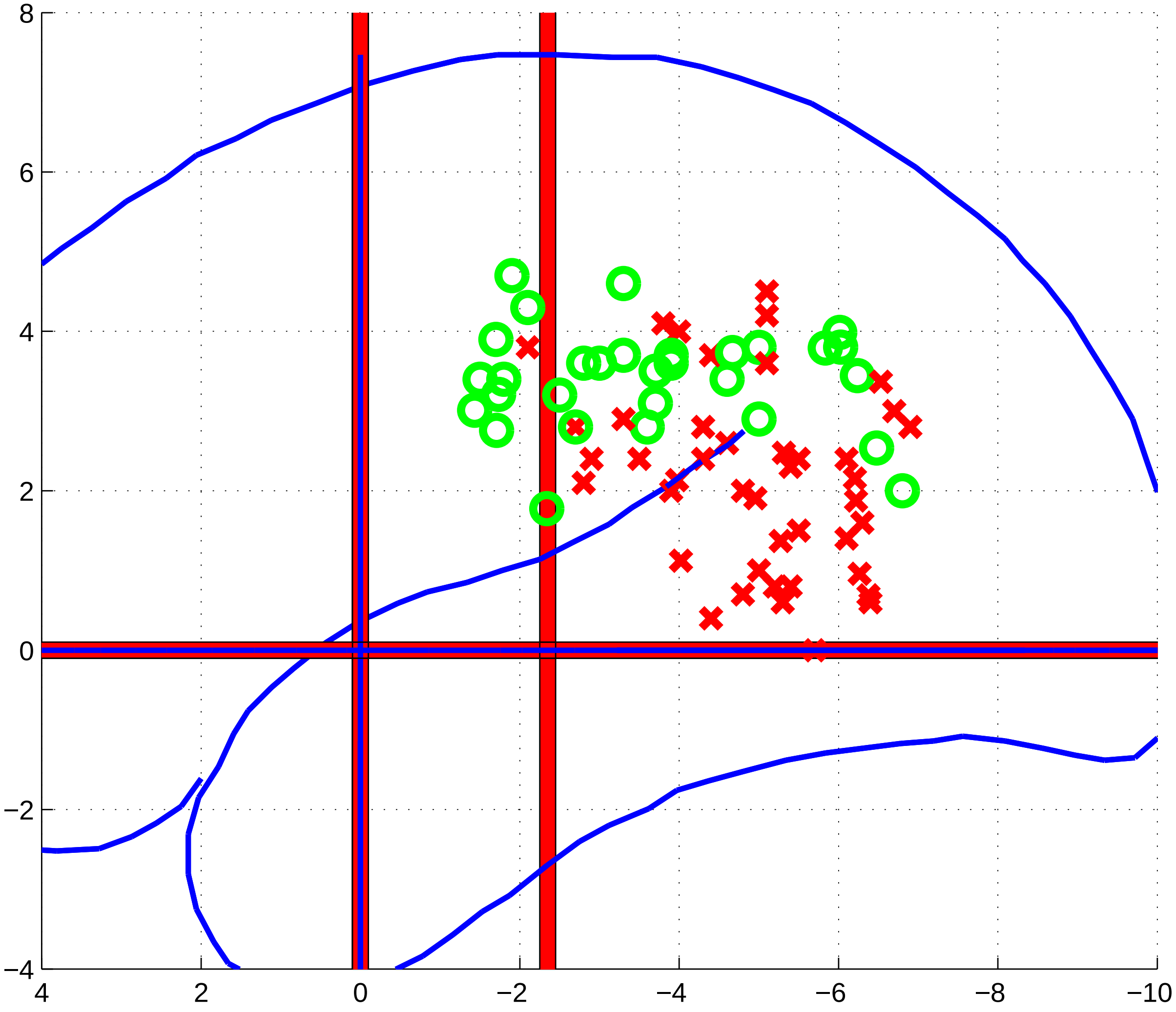
$$\mathbf{S}_u = (M_1 \mathbf{S}_1 + M_2 \mathbf{S}_2)/(M-2), \tag{9}$$

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

Results from spatial tests

#Comp	C1	C2	P-values		C1 - C2
5	1	2	0.000001	0.000023	memory - pain
2	1	2	0.000001	0.000663	memory - pain
4	1	2	0.000009	0.000102	pain - memory
8	6	8	0.000014	0.000231	pain - memory
3	1	2	0.000023	0.004385	memory - pain
7	1	7	0.000026	0.001388	pain - encoding
5	2	5	0.000030	0.009020	pain - facial
7	1	2	0.000030	0.027949	pain - facial
6	5	6	0.000030	0.027949	pain - facial
8	5	6	0.000050	0.000872	words - pain
9	1	7	0.000084	0.001452	pain - words
8	1	6	0.000112	0.005228	alzheimer - pain

Plot of pain and memory



Plot of memory (red x) and pain (green circles) locations viewed from the side (sagittal).

The locations are picked from the most separated nodes: $K = 5$, $k_1 = 1$ and $k_2 = 2$ with a P-value of 0.000001 (Hotelling's test).

Further modeling . . .

Convex hull peeling

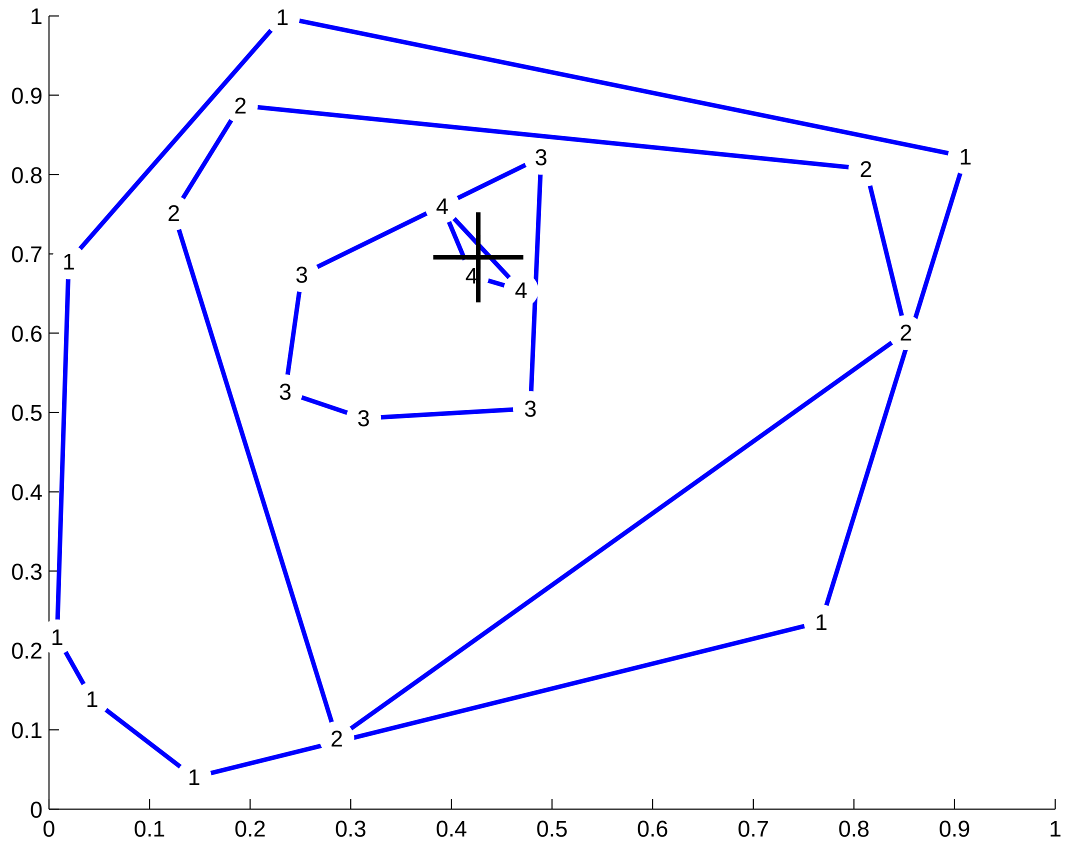


Figure 7: Convex hull peeling

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

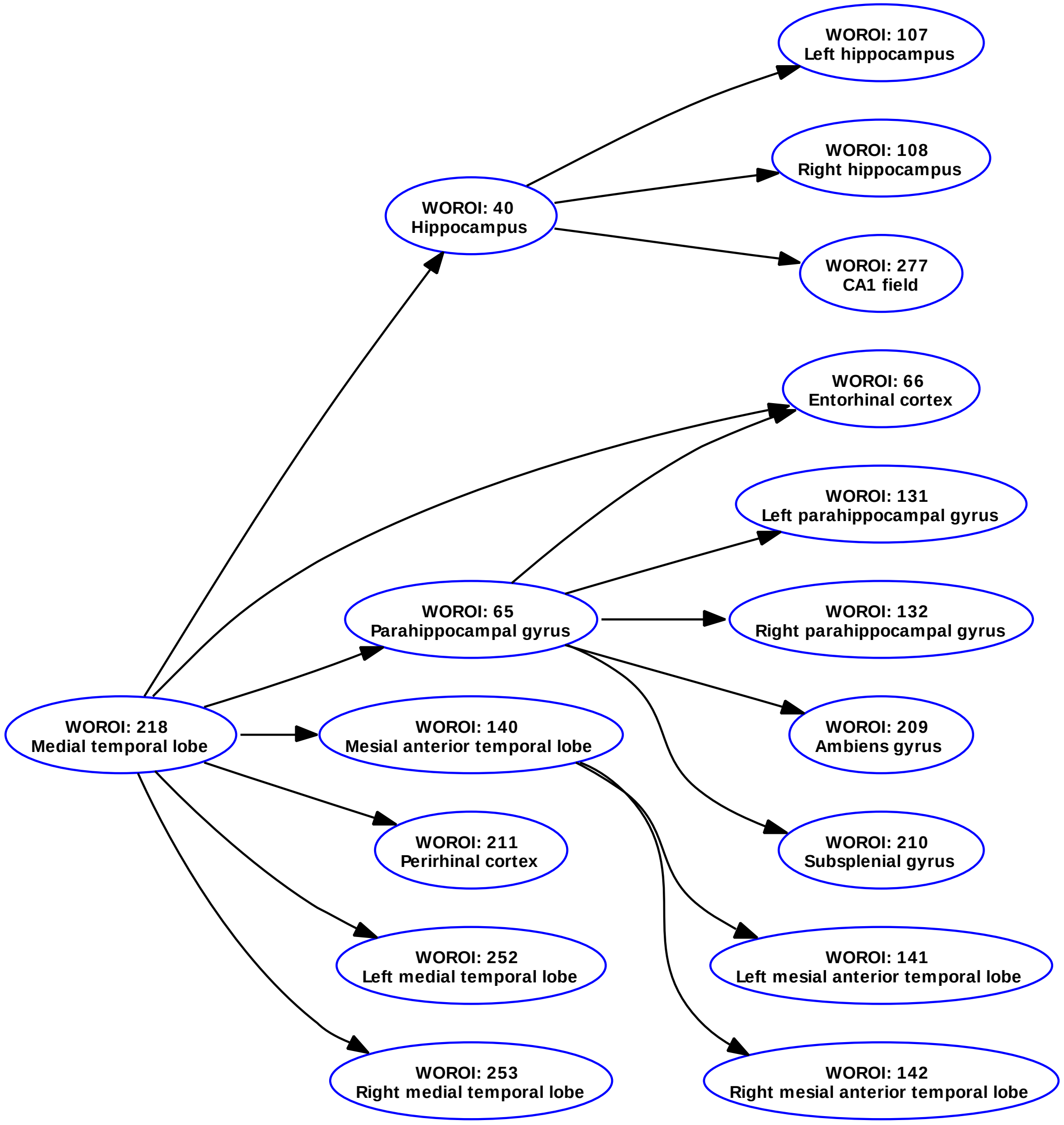
Convex hull peeling centroid (Barnett, 1976) is a robust multivariate estimate of the centroid.

Monte Carlo permutation test on the distance between centroids.

Hotelling's P-value: 0.0000013

Peeling permutation P-value: $\approx$ 0.0057

Neuroanatomy taxonomy



Taxonomy of neuroanatomical areas.

Items linked in a hierarchy with “Brain” in the top root and smaller areas in the leafs.

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.

Examples names for “medial temporal lobe”

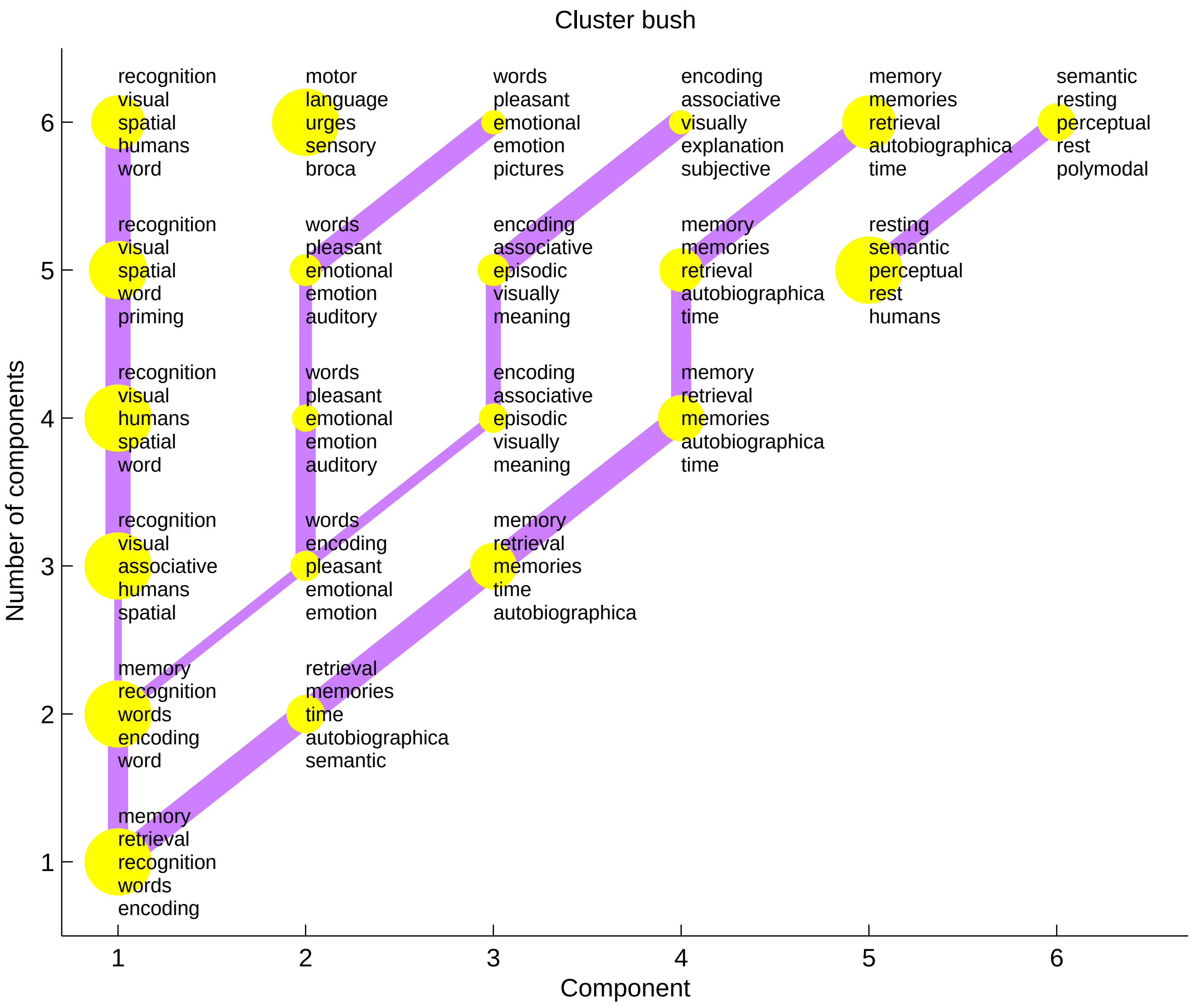
```
'Medial temporal lobe'
'Hippocampus'
'Parahippocampal gyrus'
'Parahippocampal'
'Parahippocampus'
'Gyrus parahippocampi'
'Gyrus parahippocampalis'
'Entorhinal cortex'
'Cortex entorhinalis'
'Entorhinal area'
'Area entorhinalis'
'Left hippocampus'
              ⋮
```

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.

“Medial temporal lobe” abstract grouping

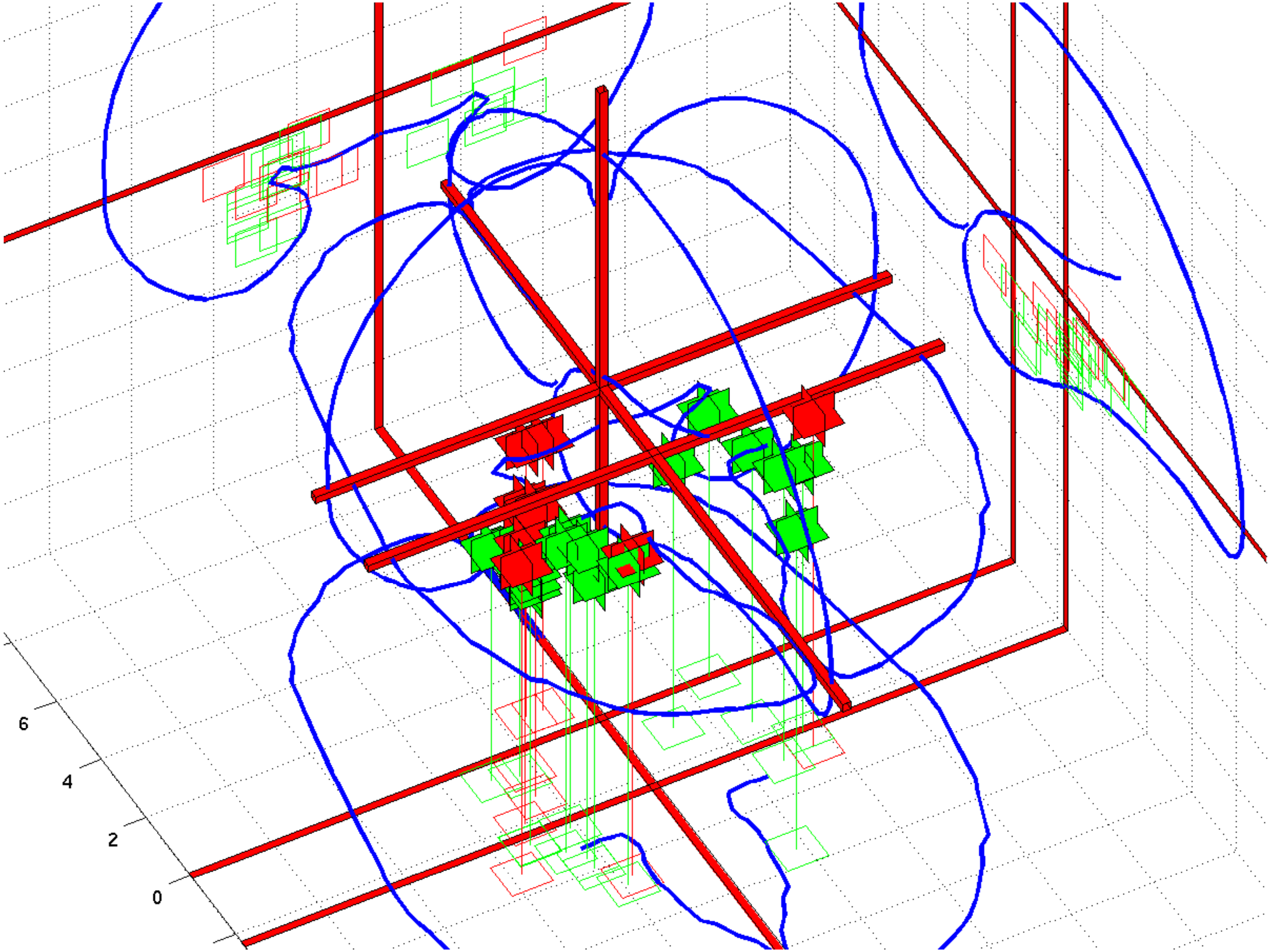


Spatial test "medial temporal lobe"

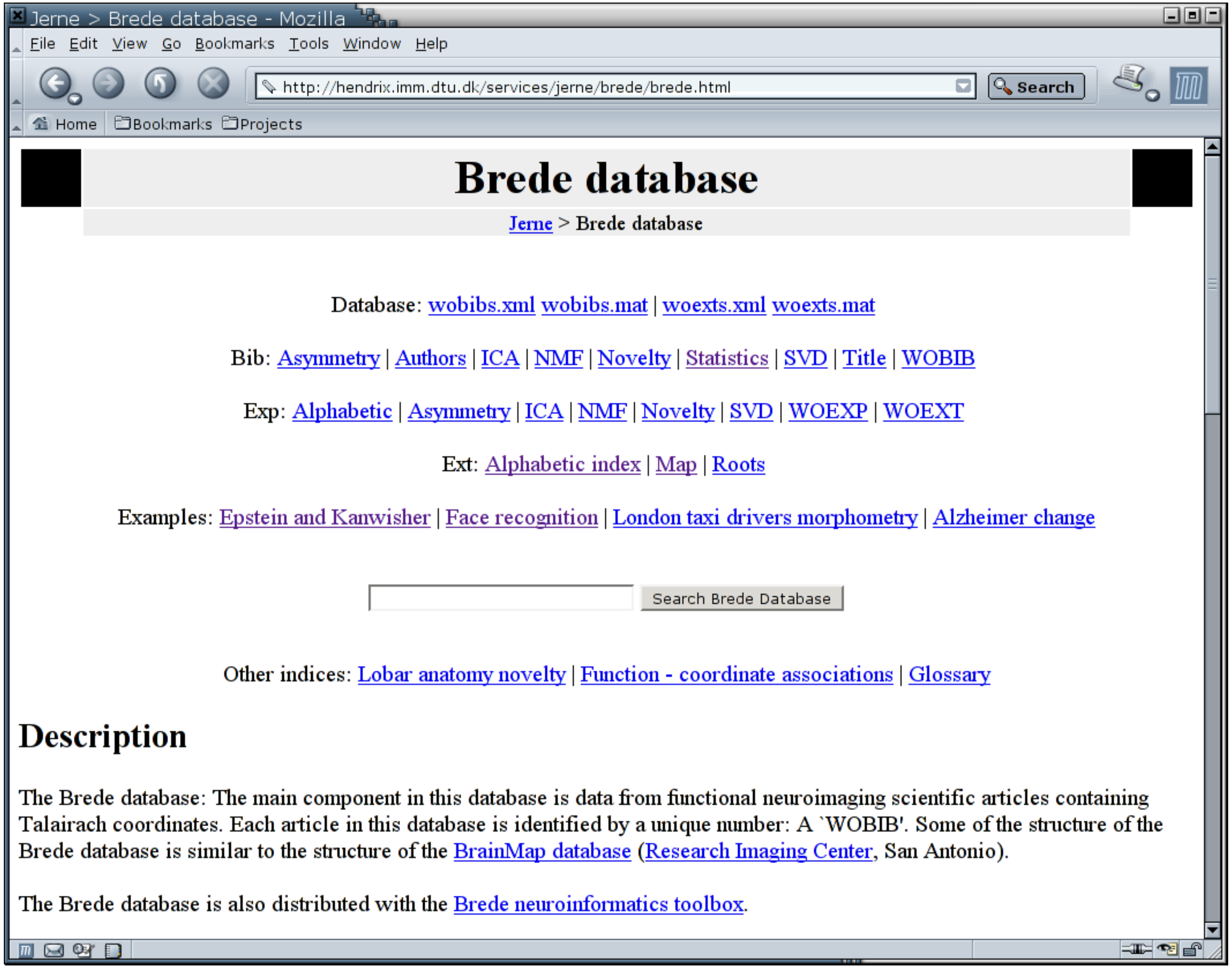
#Comp	C1	C2	P-values		C1 - C2
4	1	3	0.000663	0.057905	words - encoding
6	2	3	0.021106	0.023796	recognition - words
4	2	3	0.032204	0.430667	recognition - encoding
6	1	3	0.041044	0.167343	encoding - words
5	3	5	0.041044	0.167343	encoding - words
6	1	5	0.074015	0.089068	encoding - memory
6	2	5	0.120152	0.010952	recognition - memory

Peeling permutation test $P \approx 0.5128$

Spatial test "medial temporal lobe"



Brede database on the web


Jerne > Brede database - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://hendrix.imm.dtu.dk/services/jerne/brede/brede.html
Search
Home
Bookmarks
Projects

Brede database

Jerne > Brede database

Database: wobibs.xml wobibs.mat | woexts.xml woexts.mat

Bib: Asymmetry | Authors | ICA | NMF | Novelty | Statistics | SVD | Title | WOBIB

Exp: Alphabetic | Asymmetry | ICA | NMF | Novelty | SVD | WOEXP | WOEXT

Ext: Alphabetic index | Map | Roots

Examples: Epstein and Kanwisher | Face recognition | London taxi drivers morphometry | Alzheimer change

Search Brede Database

Other indices: Lobar anatomy novelty | Function - coordinate associations | Glossary

Description

The Brede database: The main component in this database is data from functional neuroimaging scientific articles containing Talairach coordinates. Each article in this database is identified by a unique number: A `WOBIB'. Some of the structure of the Brede database is similar to the structure of the BrainMap database (Research Imaging Center, San Antonio).

The Brede database is also distributed with the Brede neuroinformatics toolbox.

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. 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, [defkat.dk — bibliotek.dk].

Cabeza, R. and Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience*, 12(1):1–47. PMID: 10769304. http://jocn.mitpress.org/cgi/content/abstract/12/1/1.

Fox, P. T. and Lancaster, J. L. (1994). Neuroscience on the net. *Science*, 266(5187):994–996. PMID: 7973682.

Fox, P. T. and Lancaster, J. L. (2002). Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience*, 3(4):319–321. http://www.brainmapdbj.org/Fox01context.pdf. Describes the philosophy behind the (new) BrainMap functional brain imaging database with "BrainMap Experiment Coding Scheme" and tables of activation foci. Furthermore discusses financial problems and quality control of data.

Fox, P. T., Lancaster, J. L., Parsons, L. M., Xiong, J.-H., and Zamarripa, F. (1997). Functional volumes modeling: Theory and preliminary assessment. *Human Brain Mapping*, 5(4):306–311. http://www3.interscience.wiley.com/cgi-bin/abstract/56435/START.

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.

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.nj.nec.com/lee00algorithms.html.

Mai, J. K., Assheuer, J., and Paxinos, G. (1997). *Atlas of the Human Brain*. Academic Press, San Diego, California. ISBN 0124653618, [defkat.dk — bibliotek.dk — amazon.com — bn.com — isbn.nu].

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London. ISBN 0124712525, [defkat.dk — bibliotek.dk — amazon.com — bn.com — isbn.nu].

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. (2004). Mass meta-analysis in Talairach space. Accepted for the *Neural Information Processing Systems* conference in 2004.

Nielsen, F. Å. and Hansen, L. K. (2002). Modeling of activation data in the BrainMap^{TM} database: Detection of outliers. *Human Brain Mapping*, 15(3):146–156. http://www3.interscience.wiley.com/cgi-bin/abstract/89013001/. CiteSeer: http://citeseer.nj.nec.com/nielsen02modeling.html.

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

Turkeltaub, P. E., Eden, G. F., Jones, K. M., and Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *NeuroImage*, 16(3 part 1):765–780. PMID: 12169260. http://www.sciencedirect.com/science/article/B6WNP-46HDMPV-N/2/87ce95b60732a8f0c917e288efe59004.