Combining text mining and coordinate-based meta-analysis

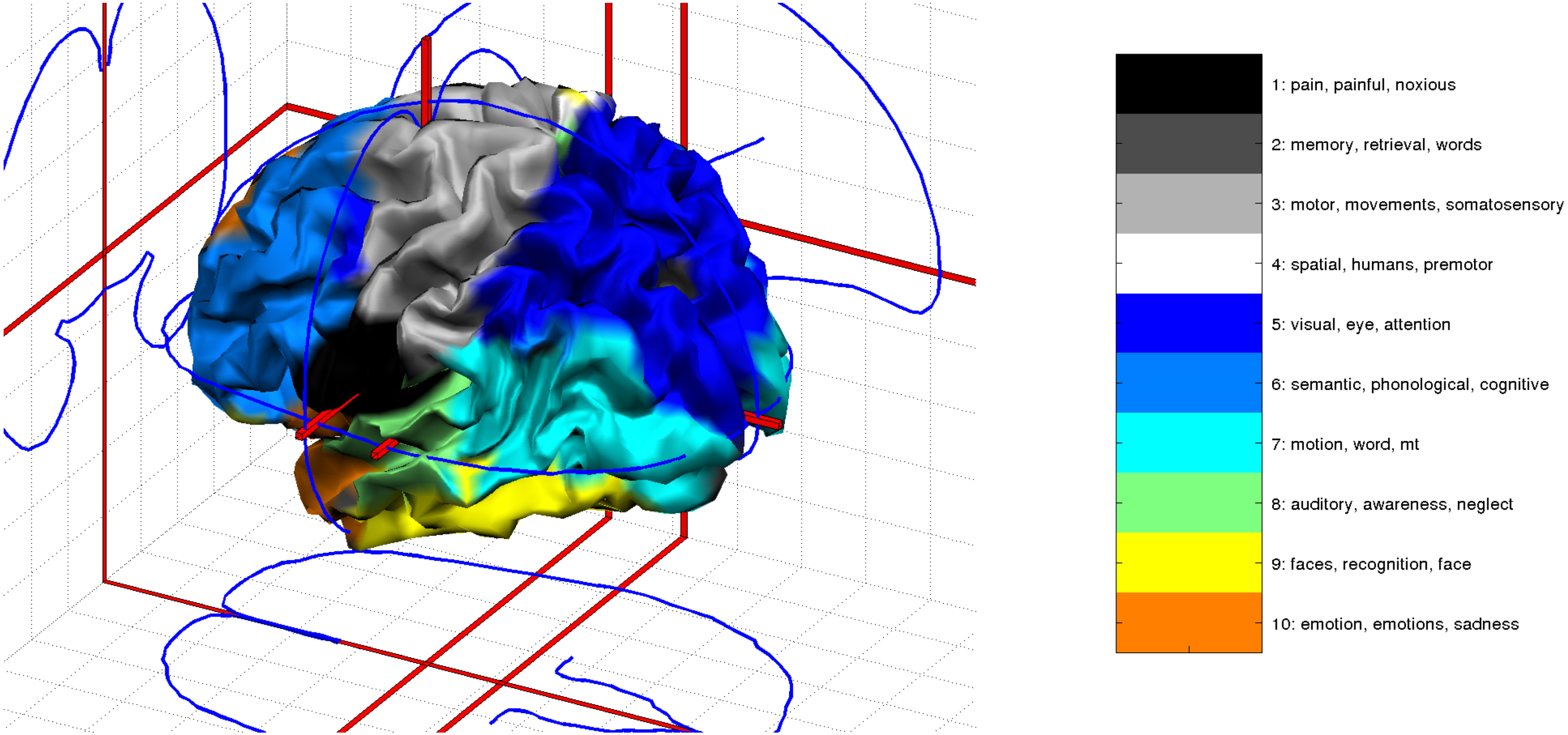
Finn Årup Nielsen

DTU Compute
Technical University of Denmark

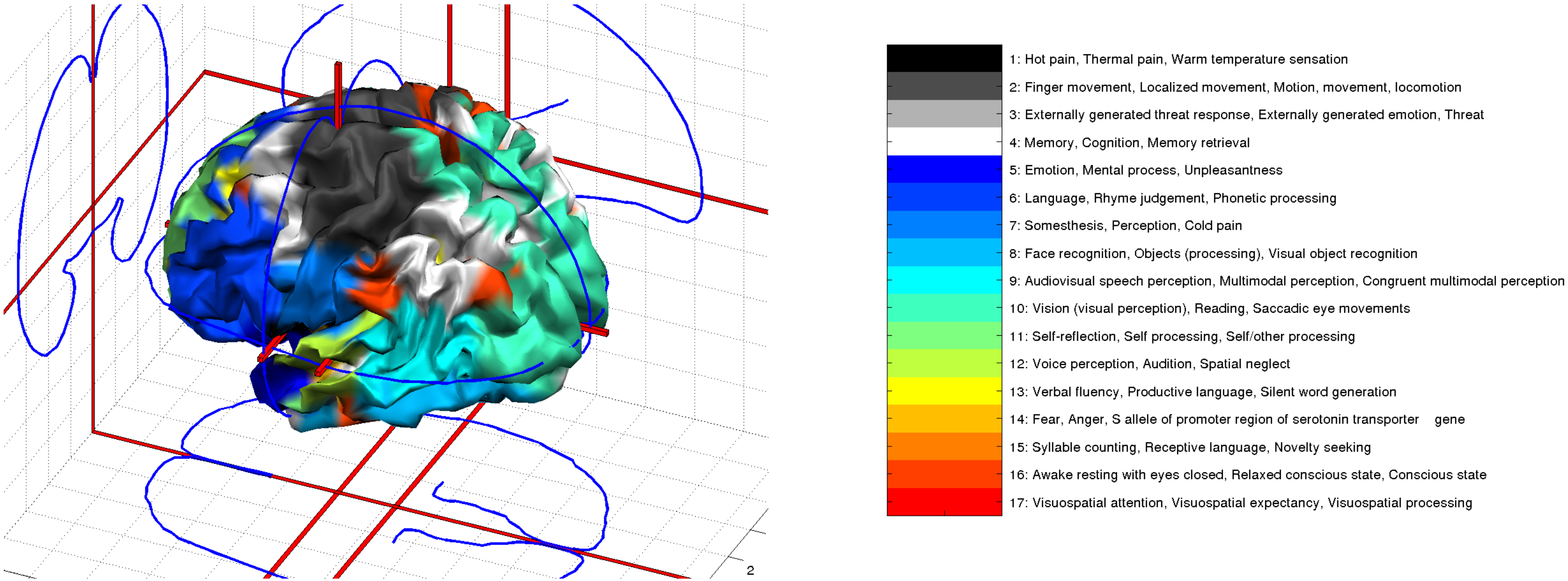
Workshop: Neuroimaging meta-analysis methods

April 17, 2015

Brain atlas constructed from abstract



Brain atlas constructed from experiment labels



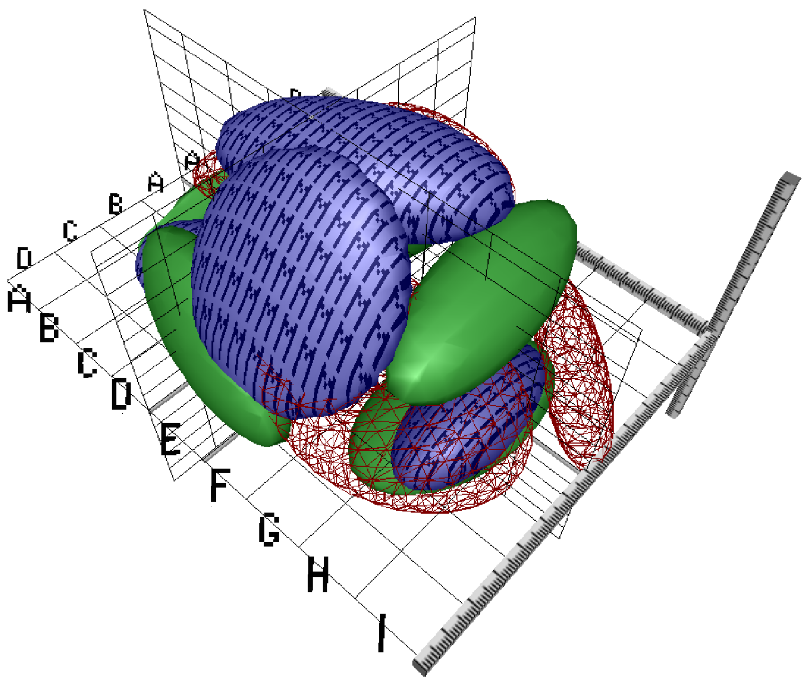
How is this done?

Coordinate-based database

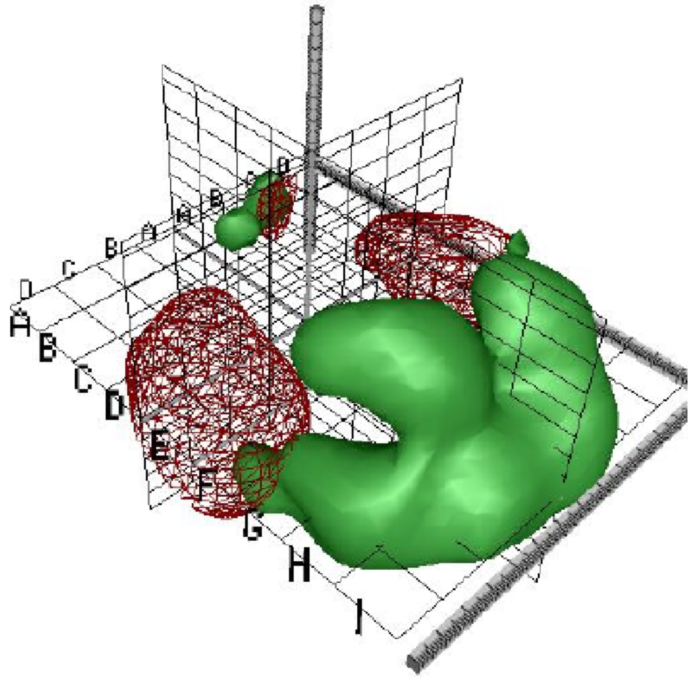
Name	'Old' BrainMap	Brede	Neurosynth
Entry	Manual	Manual	Automatic
Collection	Web scraping	XML	TSV
Papers	224	186	$\approx 10,000$
Experiments	771	586	$> 18,000$
Locations	7,263	3,912	$>$150,000
Abstracts	✓	✓	✓
Loc. labels	✓	✓	
Exp. labels	✓	✓	
Reference	(Fox and Lancaster, 1994)	(Nielsen, 2003)	(Yarkoni et al., 2011)

Other databases: AMAT (Hamilton, 2009), SumsDB (Van Essen, 2009). Part of Brede Database is in AMAT. 'New' BrainMap.

Modeling of data in BrainMap



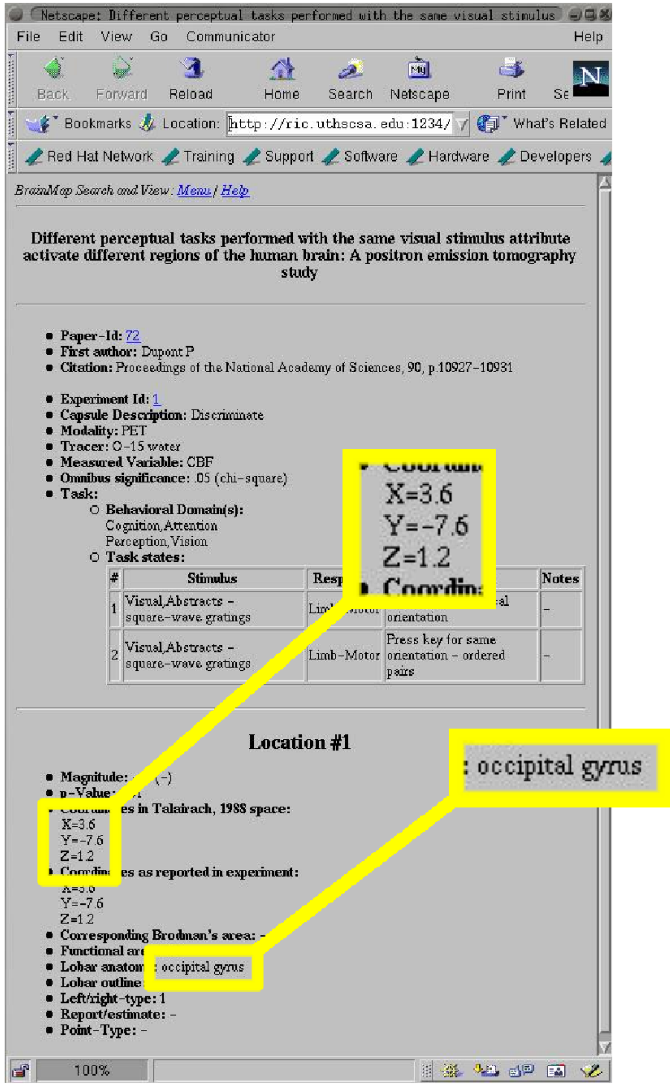
Gaussian mixture model for perception, cognition, motion (Nielsen and Hansen, 1999)



Kernel density estimation for audition, vision (Nielsen and Hansen, 2000)

$p(\mathbf{x}|l)$: $\mathbf{x}$ is Talairach space, l is a BrainMap label ('behavioral domain').

Modeling anatomical labels in BrainMap



"We downloaded 7,263 location web-pages and 3,935 of these locations had an associated anatomical label" (Nielsen and Hansen, 2002b).

$p(\mathbf{x}|l)$: $\mathbf{x}$ is Talairach space, l is the anatomical label (word or phrase) associated with each coordinate, e.g., "occipital gyrus", "gyrus", "occipital".

1,231 word/phrases.

Volume available from: neuro.compute.dtu.dk. as Analyze files: "Meta-analytic region of interest atlas" (Nielsen and Hansen, 2002a).

Modeling anatomical labels in BrainMap

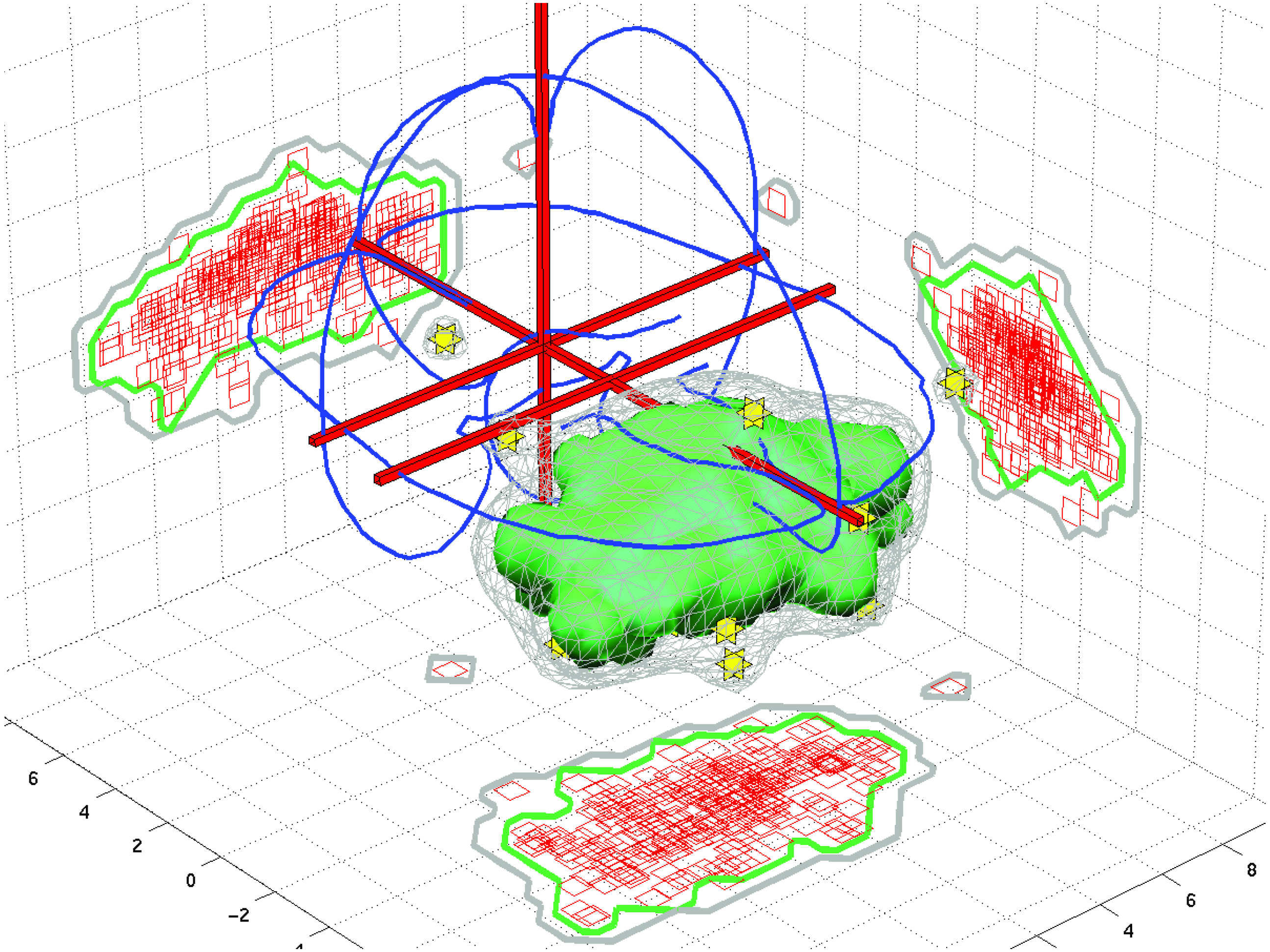
Kernel density estimation with leave-one-out cross-validation for determining the width of the kernel (Nielsen and Hansen, 2002b).

$$E(\sigma^2, l) = -\sum_{n=1}^{N_l} \log p_{-n}(\mathbf{x}_n|\sigma^2, l)$$

$$p_{-n}(\mathbf{x}|\sigma^2, l) = \frac{1}{N_l - 1} \sum_{n' \neq n}^{N_l} (2\pi\sigma^2)^{-3/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{x}_{n'})^2\right)$$

Lars Kai Hansen's algorithm with Newton optimization for optimizing the kernel width. Available in Brede Toolbox (Matlab).

Modeling anatomical labels in BrainMap



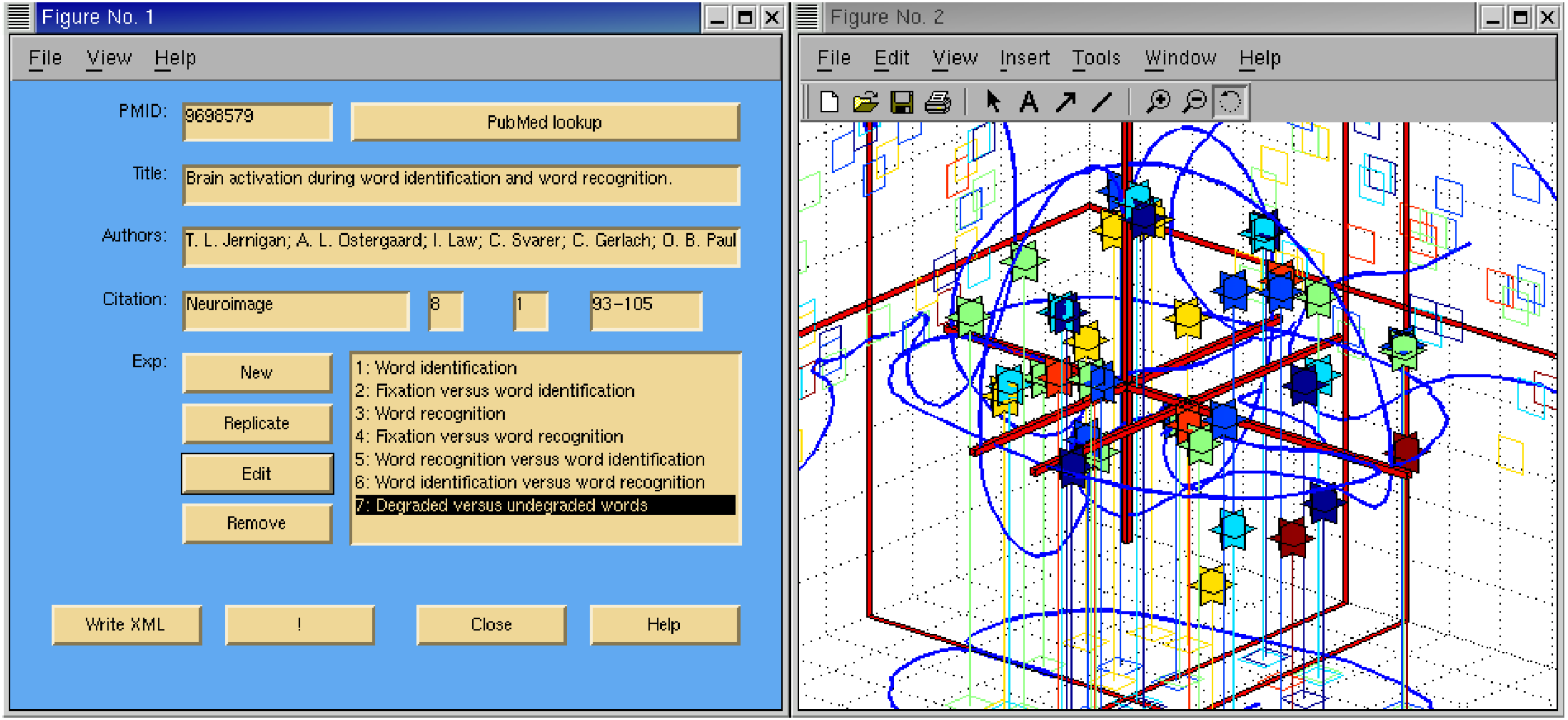
Finding outliers in BrainMap by looking at coordinates located in areas with low probability density.

Here $p(\mathbf{x}|l = \text{cerebellum})$ with a coordinate in the anterior part of the brain

Typos: centimeter/millimeter, left/right, . . .

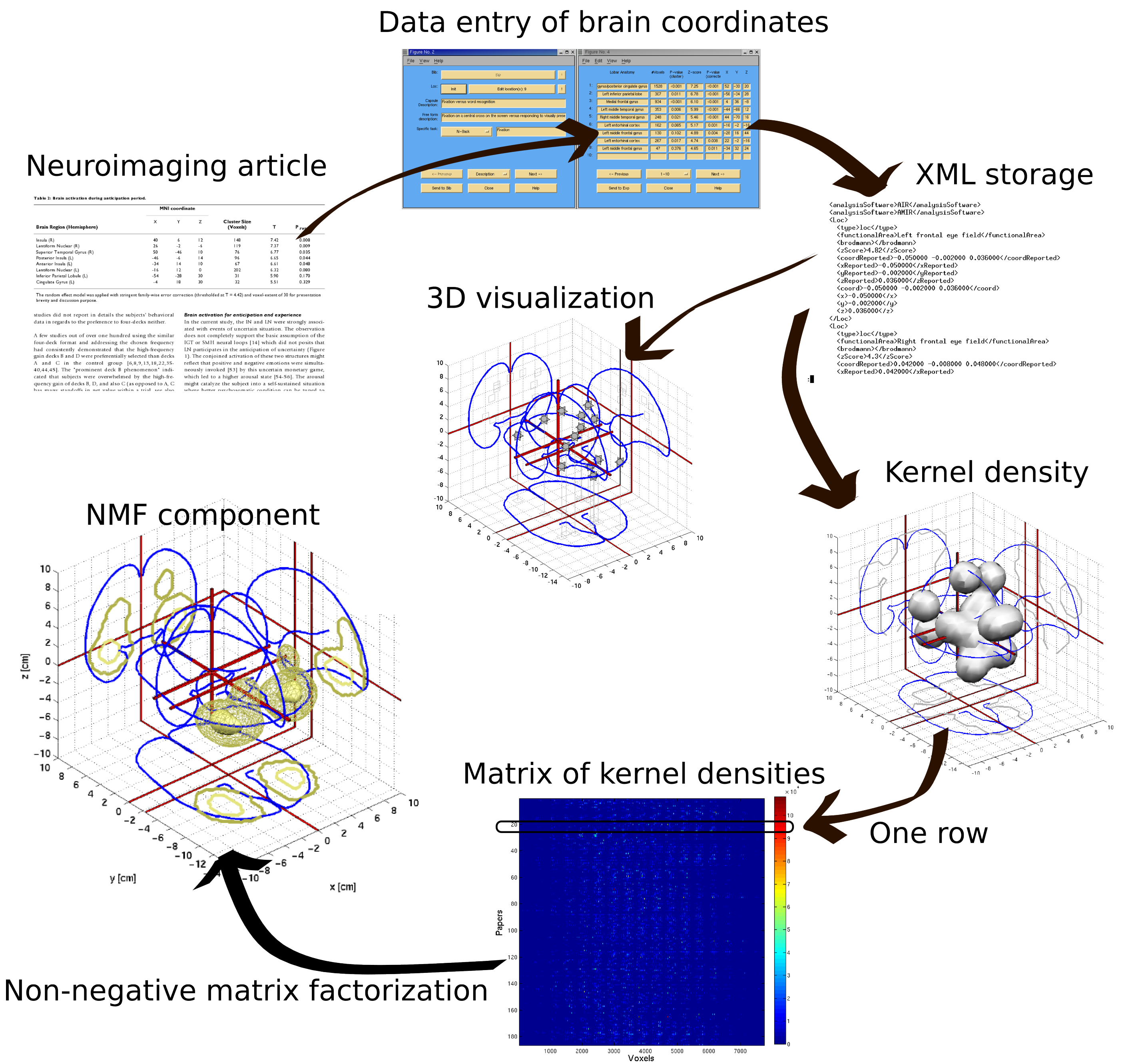
Algorithm also run for the Brede Database.

Brede Database

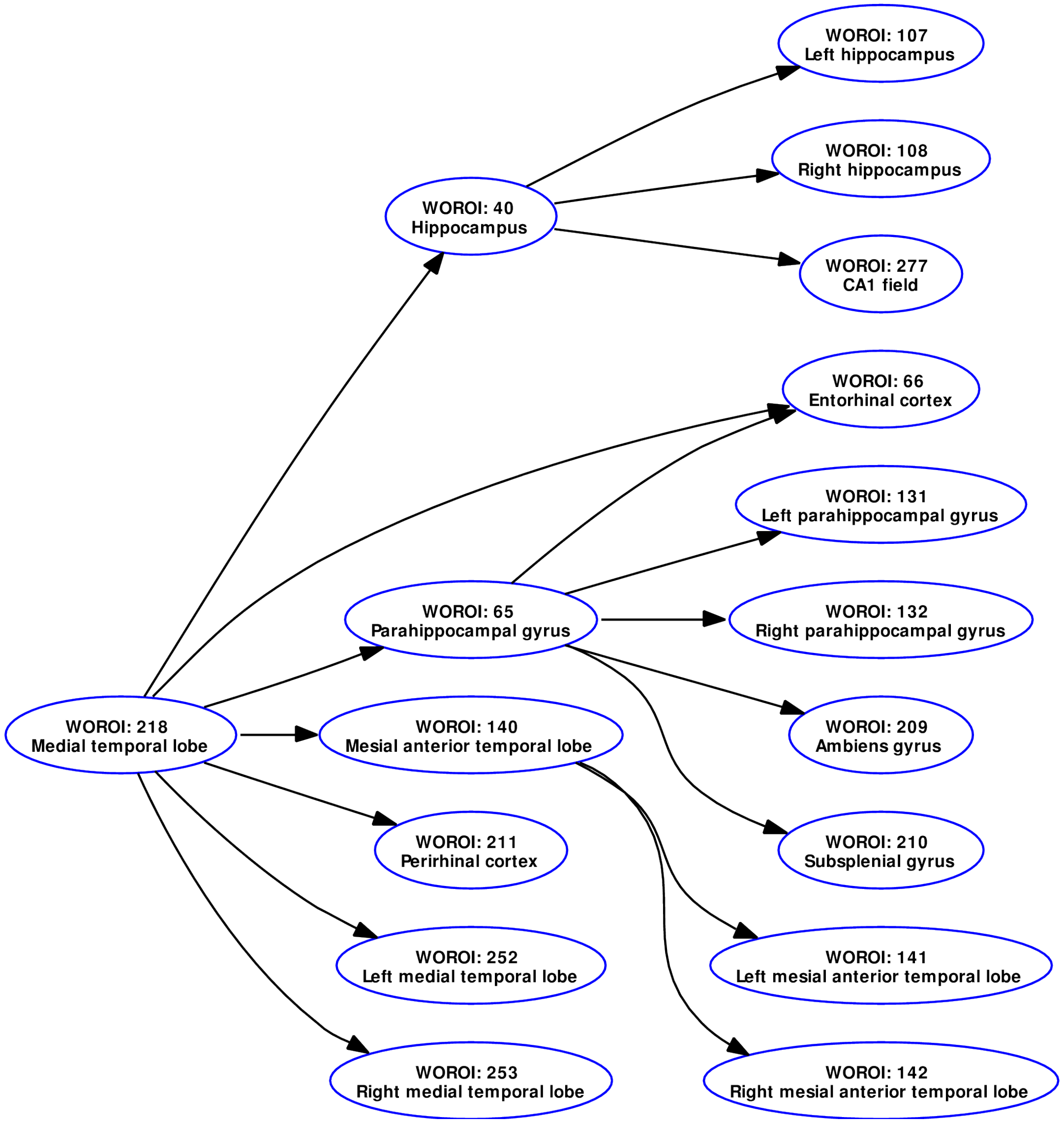


Matlab graphical user interfaction program for data entry and visualization of studies with Talairach coordinates.

Text mining and CBMA



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).

Functional segregation in brain regions

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

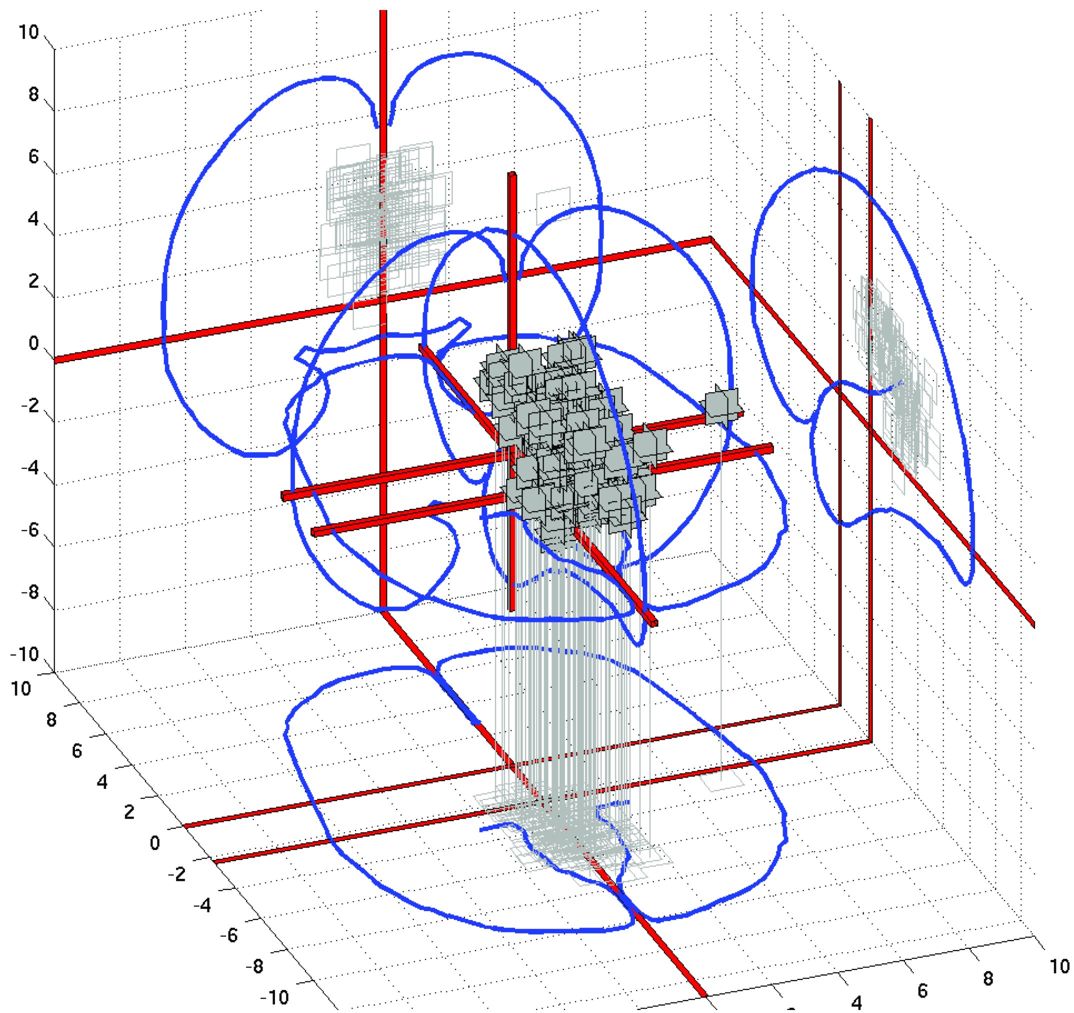
Step 3: Determine whether specific themes are spatially clustered in the brain area by testing whether two sets of coordinates are separated.

end

Step 4: Intertwine results from all brain regions

(Nielsen et al., 2006; Nielsen et al., 2005).

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 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	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

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 $\equiv \mathbf{X}(N \times P)$.

Exclude a large list of word: Anatomical, "stop words", ...

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

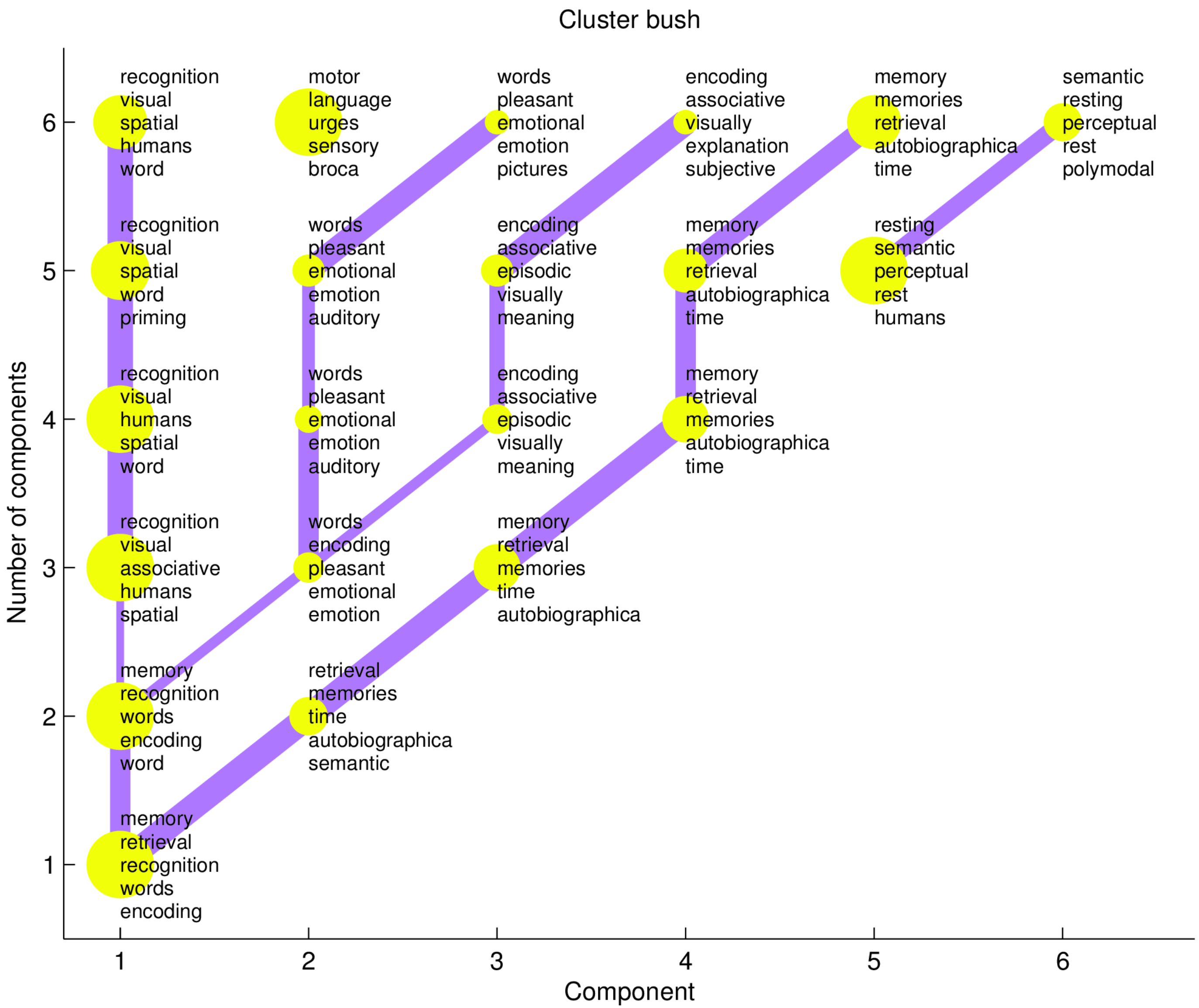
$$E_{\text{“eucl”}} = ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 \tag{2}$$

Iterative algorithm (Lee and Seung, 2001)

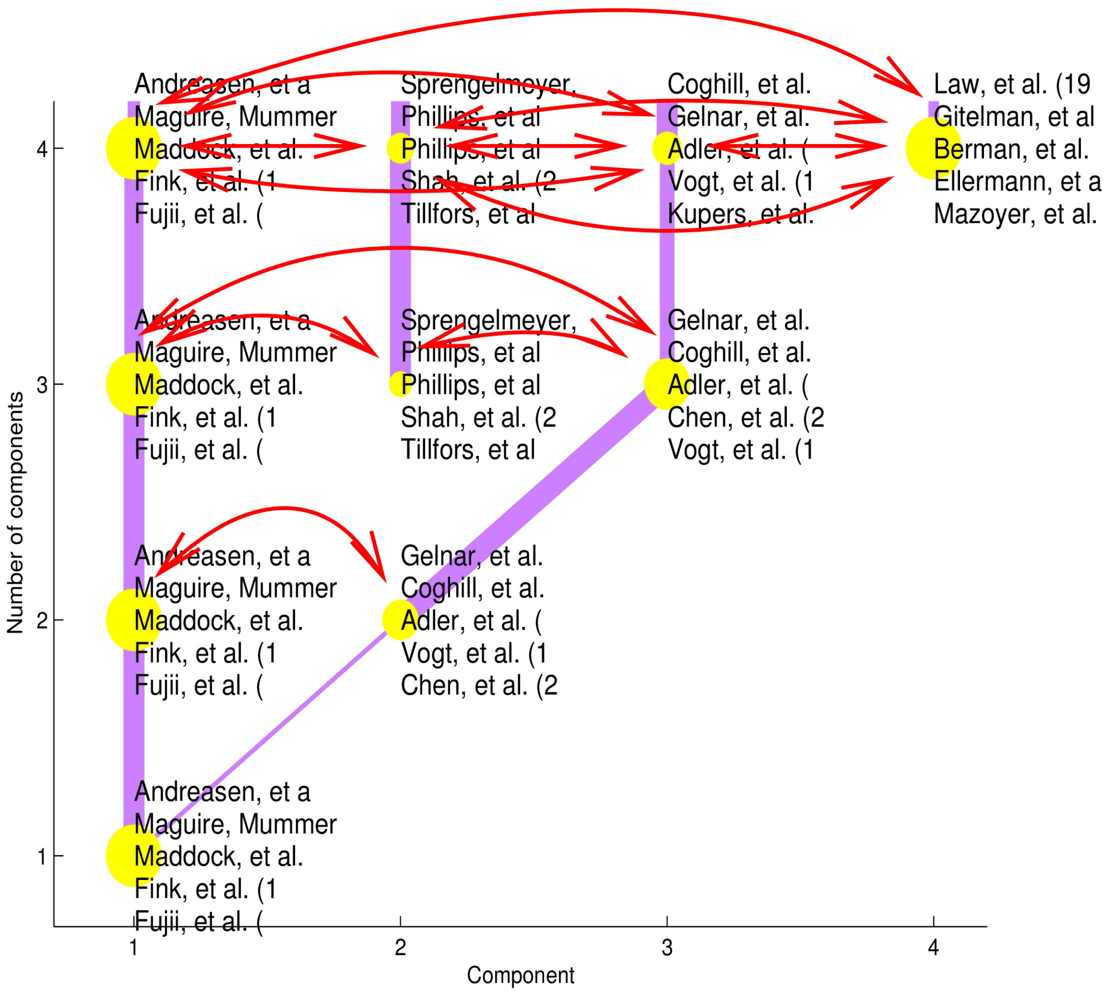
$$\mathbf{H}_{kp} \leftarrow \mathbf{H}_{kp} \frac{\left(\mathbf{W}^\mathsf{T}\mathbf{X}\right)_{kp}}{\left(\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{H}\right)_{kp}} \tag{3}$$

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

Step 2: ‘Medial temporal lobe’ NMF result



Step 3: Test spatial distribution

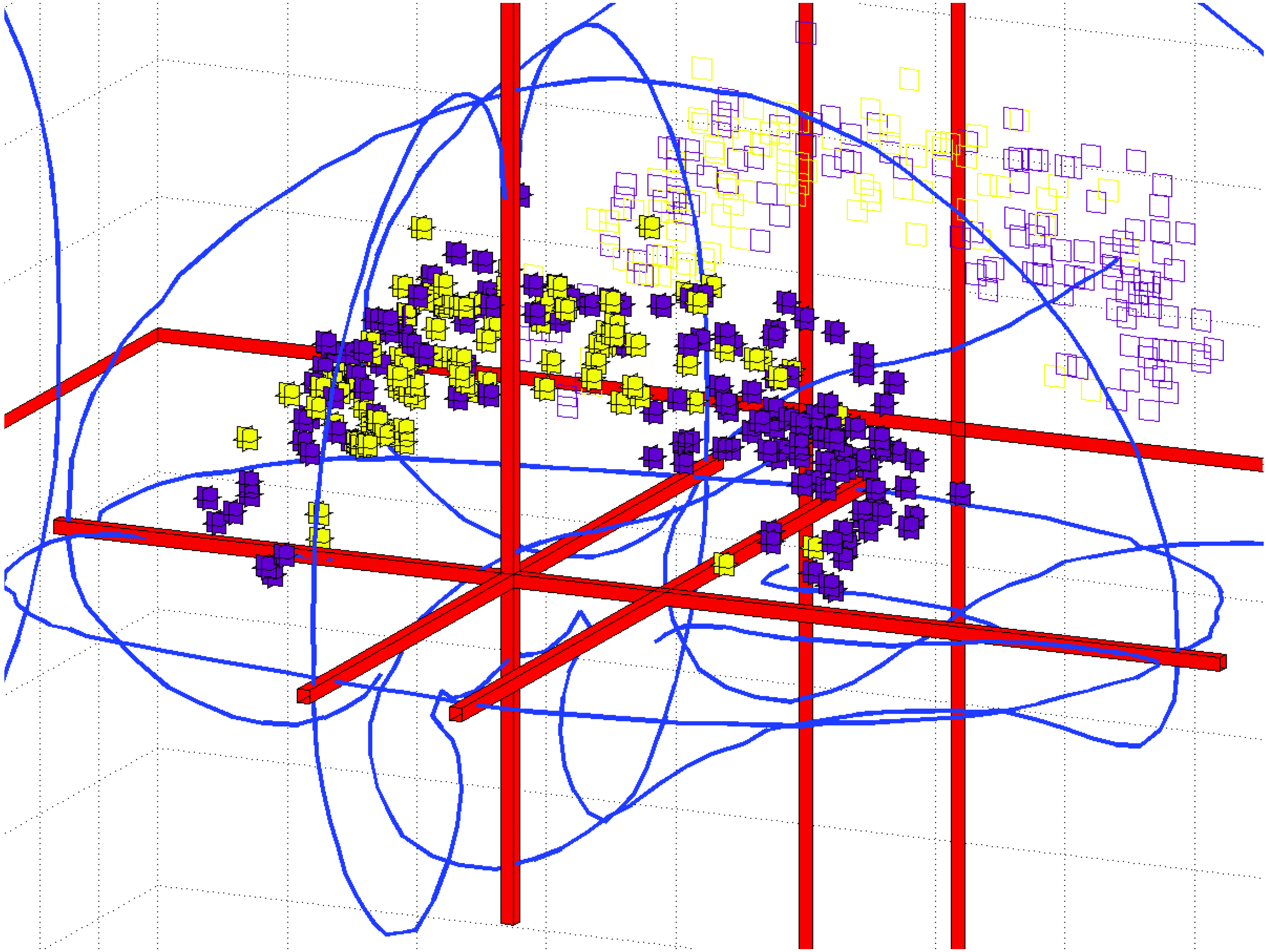


Extract locations from grouped papers.

Test if the spatial distribution of locations for a group is different from the distribution from an other group with Hotelling's T^2 or convex hull pelling permutation test.

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

Step 4: ‘Cingulate gyrus’



Coordinates associated with topics on: pain, painful vs. memory, retrieval

Combining text and coordinate directly

Construct a functional parcellation of the brain based on combined text and coordinate analysis (Nielsen et al., 2004; Nielsen, 2009).

Brede Database with neuroimaging papers.

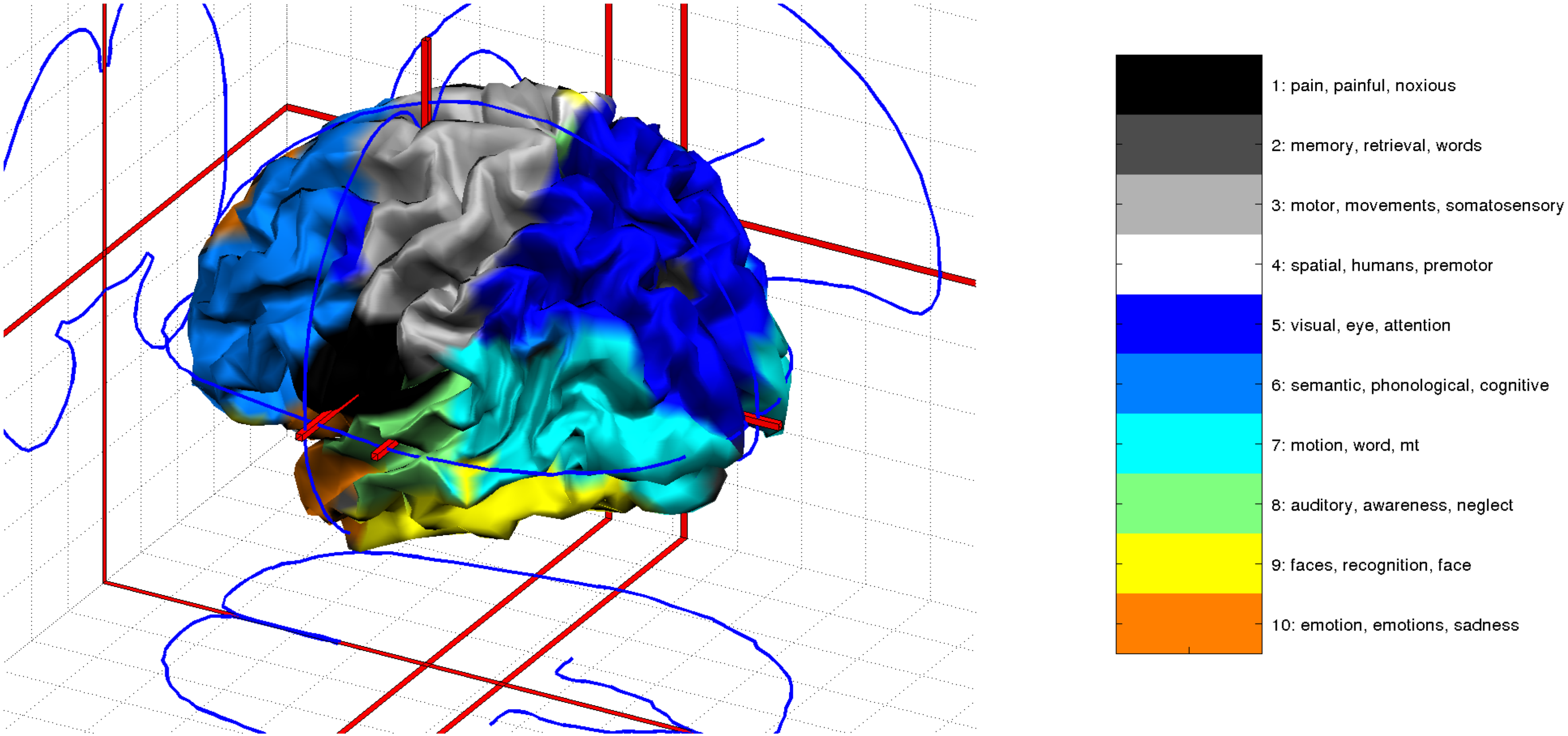
Bag-of-words matrix from abstracts, scaling, and stop words elemination getting a (paper $\times$ words)-matrix, $\mathbf{X}_1$

Kernel density estimates from coordinates contained in the papers getting a (papers $\times$ vertices)-matrix or (papers $\times$ voxels), $\mathbf{Y}$.

Product matrix: $\mathbf{Z}_1 = \mathbf{Y}'\mathbf{X}_1$ getting a (vertices $\times$ labels)-matrix.

Approximative non-negative matrix factorization: $\mathbf{W}\mathbf{H} \approx \mathbf{Z}_1$ getting a (vertices $\times$ topics)-matrix, $\mathbf{W}$ and a (topic $\times$ words)-matrix, $\mathbf{H}$.

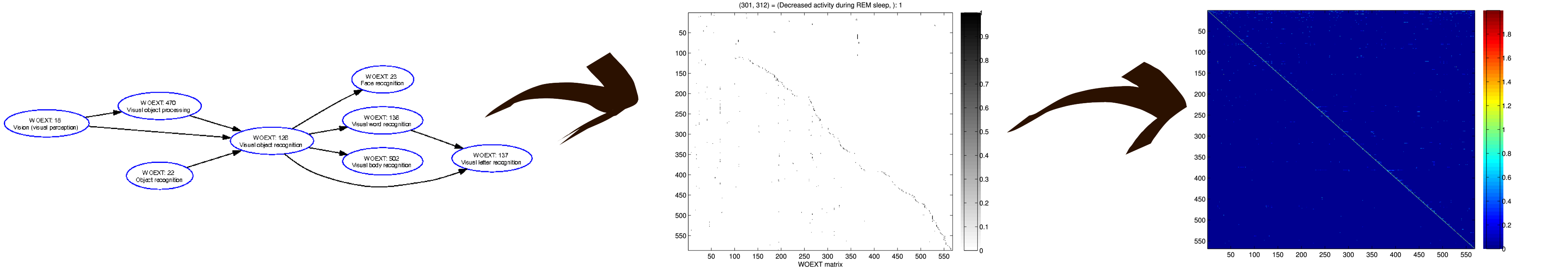
Functional parcellation of the cortex



Winner-take-all function on $\mathbf{W}$ (the surface) and $\mathbf{H}$ (the legend).

Brain function ontology

Brain function ontology | Adjacency matrix | "Recursed" adjacency matrix



Experiment label ('external components', brain function ontology, cognitive ontology) organized in a directed graph (Nielsen, 2005).

This graph can be converted to an adjacency matrix.

An experiment (i.e. contrast) in Brede Database may be labeled with an item from the ontology.

Functional parcellation with cognitive ontology

Brede Database with neuroimaging paper.

Adjacency matrix of experiment label ('external components', brain function ontology) graph ($\mathbf{C}$) propagated $\mathbf{D} = \sum_i (\lambda \mathbf{C})^i$.

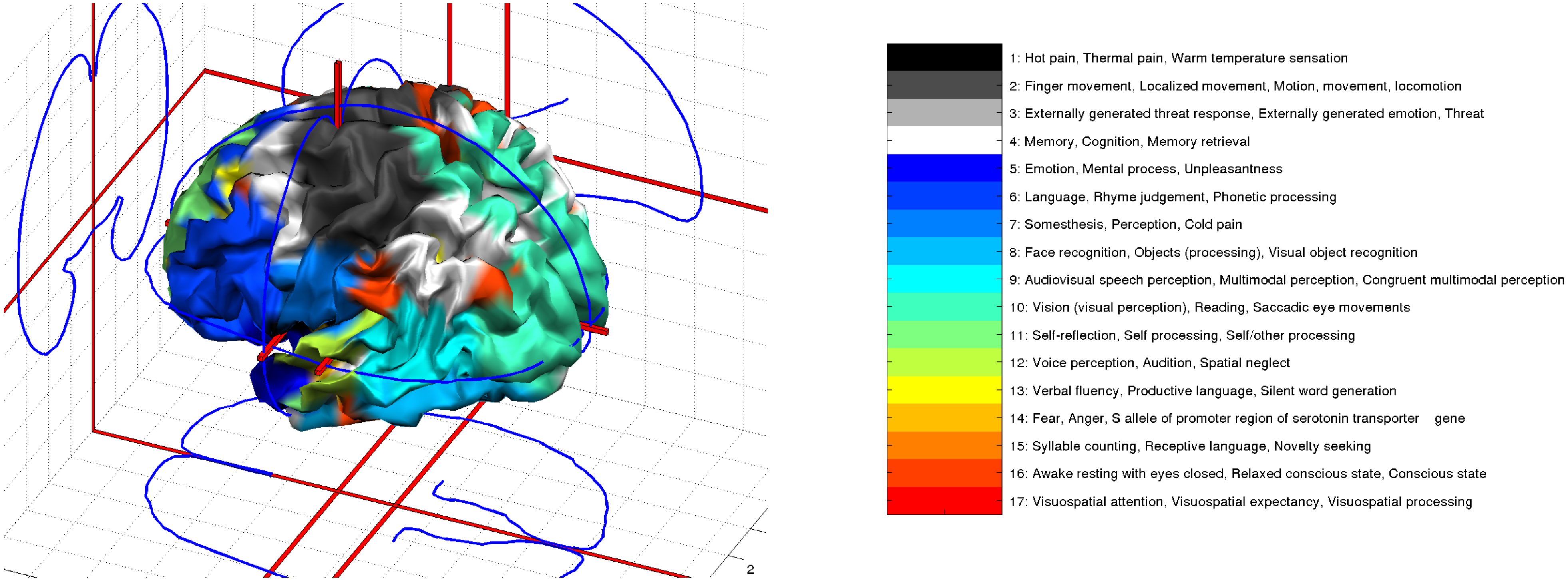
Propagated experiment label (experiments $\times$ labels)-matrix $\mathbf{X}_2 = \mathbf{TD}$, where the (experiments $\times$ experiment labels)-matrix $\mathbf{T}$ represents the (manual) label of each experiment.

Kernel density estimates from coordinates contained in the experiment getting a (experiments $\times$ vertices)-matrix or (experiments $\times$ voxel)-matrix ($\mathbf{Y}$).

Product matrix: $\mathbf{Z}_2 = \mathbf{Y}'\mathbf{X}_2$ getting a (vertices $\times$ experiment labels)-matrix.

Approximative non-negative matrix factorization: $\mathbf{WH} \approx \mathbf{Z}_2$ getting a (vertices $\times$ topics)-matrix ($\mathbf{W}$) and a (topic $\times$ words)-matrix ($\mathbf{H}$)

Brain atlas constructed from experiment labels



Winner-take-all function on $\mathbf{W}$ (the surface) and $\mathbf{H}$ (the legend with items from the experiment labels/ontology).

Brede tools

Presently developing a Python Brede Package with features such as Features: Handling of data from Brede Wiki, Neurosynth, word lists, data sets, surfaces, . . . : https://github.com/fnielsen/brede

Brede Database: http://neuro.compute.dtu.dk/services/brededatabase/

Brede Toolbox (matlab): http://neuro.compute.dtu.dk/software/brede/

Brede Wiki: http://neuro.compute.dtu.dk/wiki/.

References

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

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

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. ISSN 00380644.

Hamilton, A. F. (2009). Lost in localization: a minimal middle way. *NeuroImage*, 48(1):8–10.

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. 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.

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.

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). Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY.

Nielsen, F. Å. (2005). Mass meta-analysis in Talairach space. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 985–992, Cambridge, MA. MIT Press. *Coordinate-based meta-analysis on data from a neuroinformatics database across many different brain functions.*

Nielsen, F. Å. (2009). Visualizing data mining results with the Brede tools. *Frontiers in Neuroinformatics*, 3:26. DOI: 10.3389/neuro.11.026.2009.

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. *Text mining of PubMed abstracts for detection of topics in neuroimaging studies mentioning posterior cingulate. Subsequent analysis of the spatial distribution of the Talairach coordinates in the clustered papers.*

Nielsen, F. Å., Balslev, D., and Hansen, L. K. (2006). Data mining a functional neuroimaging database for functional segregation in brain regions. In Olsen, S. I., editor, *Den 15. Danske Konference i Mønstergenkendelse og Billedanalyse*, Copenhagen, Denmark. The Department of Computer Science, University of Copenhagen.

Nielsen, F. Å. and Hansen, L. K. (1999). Modeling of BrainMap data. Online at http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4711/pdf/imm4711.pdf.

Nielsen, F. Å. and Hansen, L. K. (2000). Functional volumes modeling using kernel density estimation. Online at http://www.imm.dtu.dk/pubdb/views/edoc_download.php/4688/pdf/imm4688.pdf. *Description of kernel density modeling of Talairach coordinates. Examples are shown with modeling the spatial distribution of Talairach coordinates from the BrainMap database based on the database labels.*

Nielsen, F. Å. and Hansen, L. K. (2002a). Automatic anatomical labeling of Talairach coordinates and generation of volumes of interest via the BrainMap database. *NeuroImage*, 16(2). Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan.

Nielsen, F. Å. and Hansen, L. K. (2002b). Modeling of activation data in the BrainMap^{TM} database: Detection of outliers. *Human Brain Mapping*, 15(3):146–156. DOI: 10.1002/hbm.10012.

Nielsen, F. Å., Hansen, L. K., and Balslev, D. (2004). Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics*, 2(4):369–380.

Van Essen, D. C. (2009). Lost in localization—but found with foci?! *NeuroImage*, 48(1):14–17. DOI: 10.1016/j.neuroimage.2009.05.050.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670. DOI: 10.1038/NMETH.1635. *Describes the Neurosynth system for collected, extracting and analysis of stereotaxic coordinates in human brain mapping publication.*