Statistical data mining

Finn Årup Nielsen

Informatics and Mathematical Modelling
Technical University of Denmark

February 3, 2004

Introduction

- "Statistical data mining".
- The goal is "knowledge" discovery in databases.
- Classic example is co-occurence in market-backets: Beer and diapers.
- Heterogeneous data analysis on text, numbers, images, ...
- Examples from Neuroinformatics (Neuroscience + informatics).

Example: Neuroinformatics databases

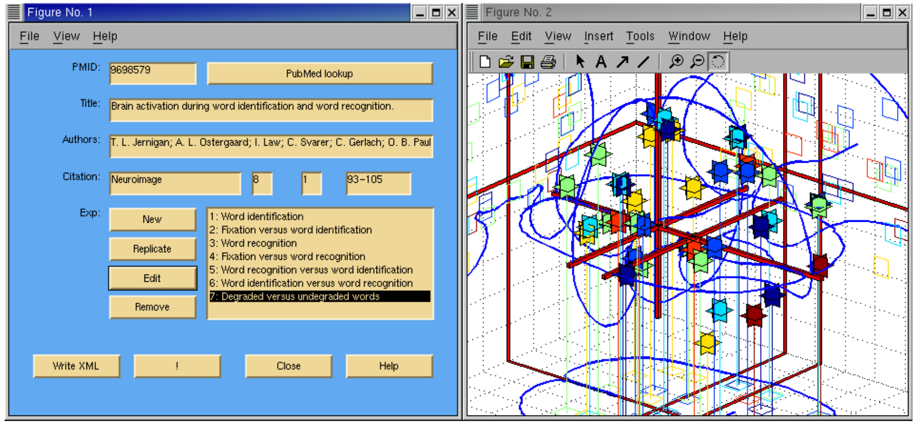


Figure 1: Screenshot of main window of Matlab program for data entry of scientific article, here (Jernigan et al., 1998).

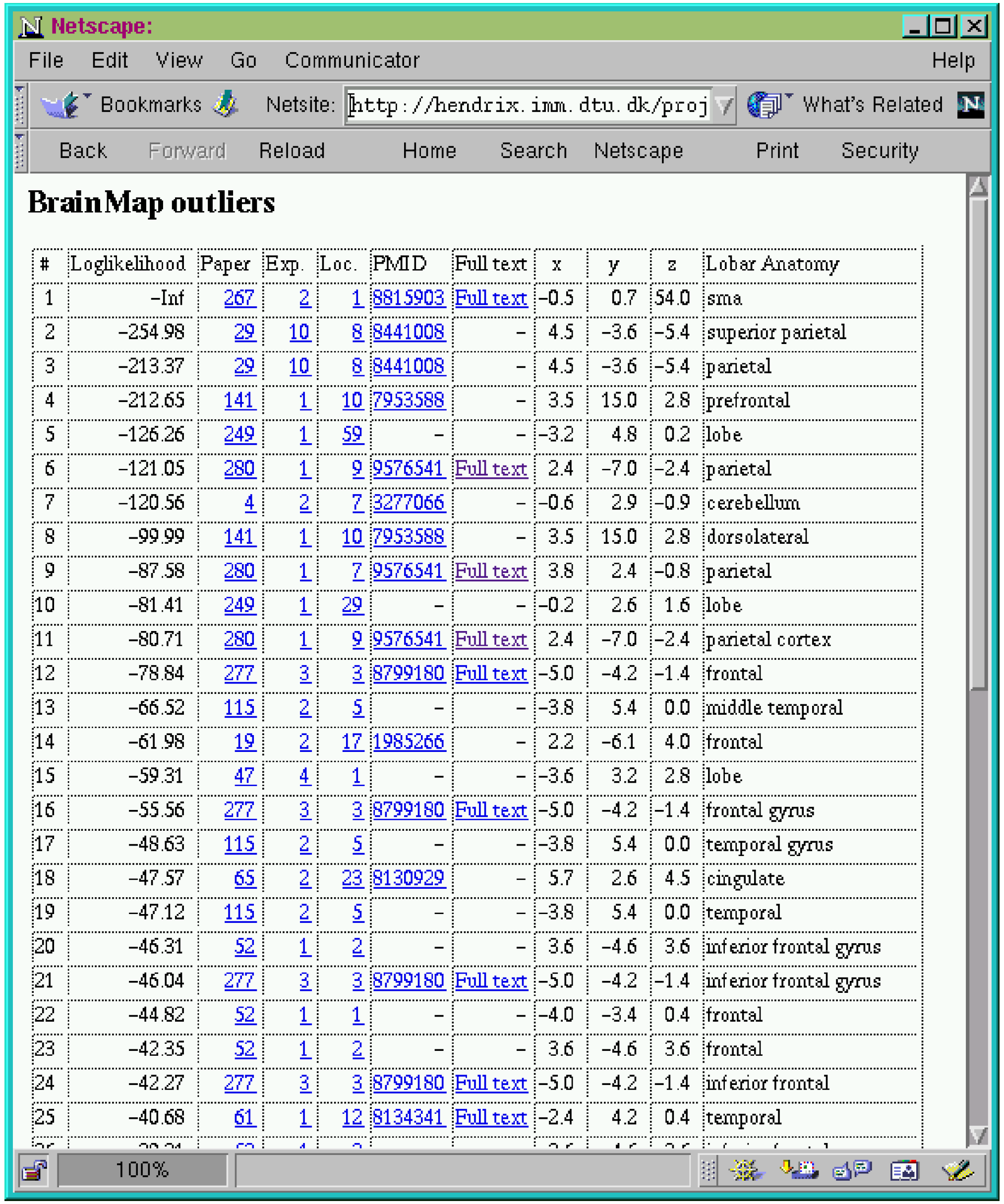
Database containing data from scientific articles in “human brain mapping”.

Bibliographic information: Title, author, abstract.

Three-dimensional coordinates, so-called Talairach coordinates, that are focal brain activations.

Experiment description: Brain scanner, stimulus, response.

Linked to other databases (PubMed, MeSH, fMRIDC, SenseLab)



BrainMap outliers

#	Loglikelihood	Paper	Exp.	Loc.	PMID	Full text	x	y	z	Lobar Anatomy
1	-Inf	267	2	1	8815903	Full text	-0.5	0.7	54.0	sma
2	-254.98	29	10	8	8441008	-	4.5	-3.6	-5.4	superior parietal
3	-213.37	29	10	8	8441008	-	4.5	-3.6	-5.4	parietal
4	-212.65	141	1	10	7953588	-	3.5	15.0	2.8	prefrontal
5	-126.26	249	1	59	-	-	-3.2	4.8	0.2	lobe
6	-121.05	280	1	9	9576541	Full text	2.4	-7.0	-2.4	parietal
7	-120.56	4	2	7	3277066	-	-0.6	2.9	-0.9	cerebellum
8	-99.99	141	1	10	7953588	-	3.5	15.0	2.8	dorsolateral
9	-87.58	280	1	7	9576541	Full text	3.8	2.4	-0.8	parietal
10	-81.41	249	1	29	-	-	-0.2	2.6	1.6	lobe
11	-80.71	280	1	9	9576541	Full text	2.4	-7.0	-2.4	parietal cortex
12	-78.84	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	frontal
13	-66.52	115	2	5	-	-	-3.8	5.4	0.0	middle temporal
14	-61.98	19	2	17	1985266	-	2.2	-6.1	4.0	frontal
15	-59.31	47	4	1	-	-	-3.6	3.2	2.8	lobe
16	-55.56	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	frontal gyrus
17	-48.63	115	2	5	-	-	-3.8	5.4	0.0	temporal gyrus
18	-47.57	65	2	23	8130929	-	5.7	2.6	4.5	cingulate
19	-47.12	115	2	5	-	-	-3.8	5.4	0.0	temporal
20	-46.31	52	1	2	-	-	3.6	-4.6	3.6	inferior frontal gyrus
21	-46.04	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	inferior frontal gyrus
22	-44.82	52	1	1	-	-	-4.0	-3.4	0.4	frontal
23	-42.35	52	1	2	-	-	3.6	-4.6	3.6	frontal
24	-42.27	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	inferior frontal
25	-40.68	61	1	12	8134341	Full text	-2.4	4.2	0.4	temporal

Mining for novelty:

Automatic generated list with entries sorted according to novelty (outlierness/interestingness).

Comparing the "lobar anatomy" field and Talairach coordinates.

By "manual investigation" one finds that some of the interesting are database entry errors.

How is this done?

Representing text

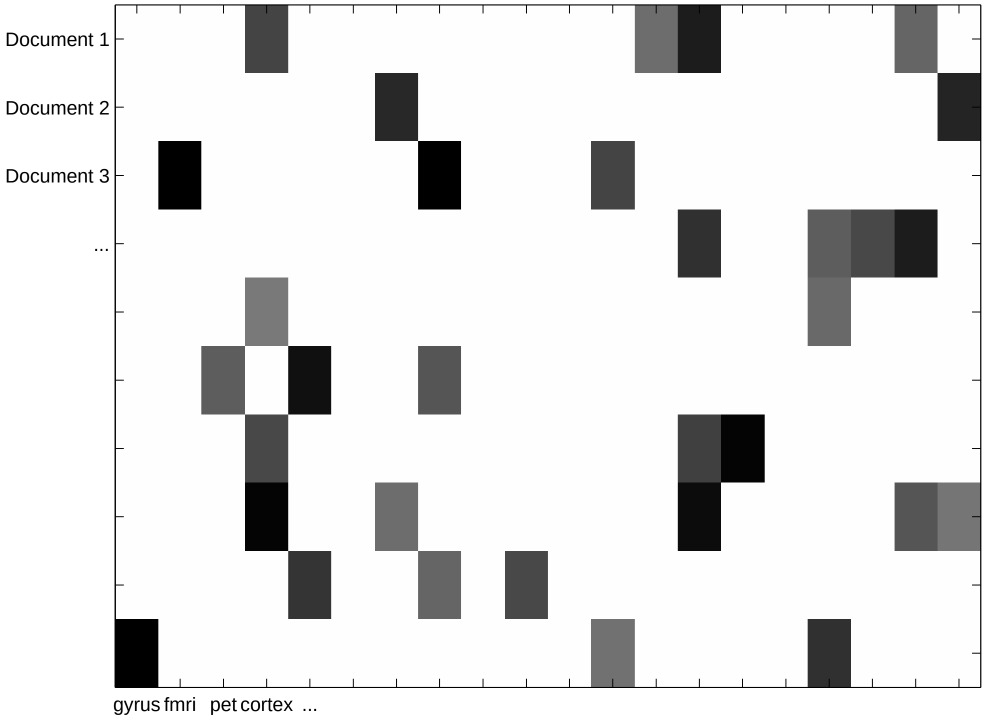


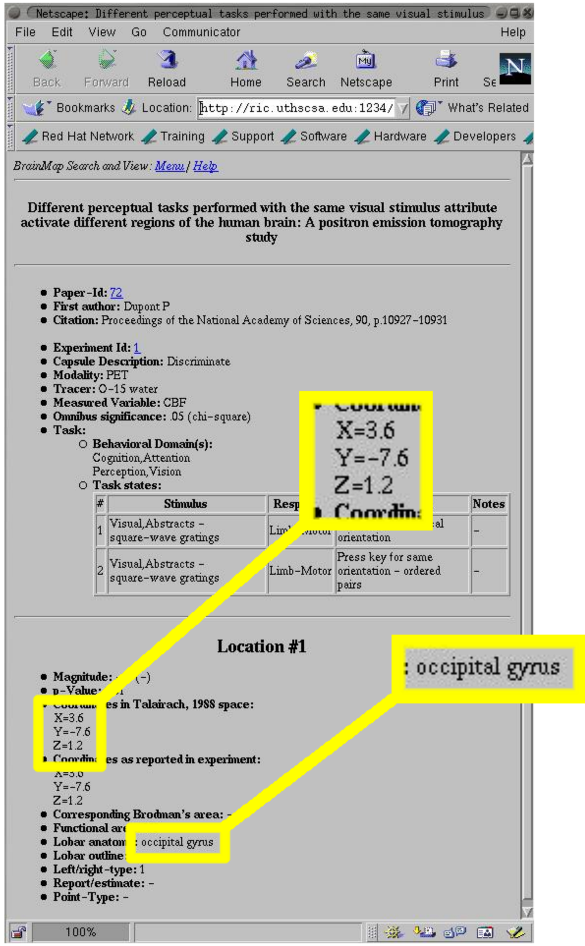
Figure 2: Bag-of-words matrix.

"Vector space model" or "bag-of-words". A matrix $\mathbf{X}(N \times Q)$ with N documents and Q words/terms. Represented in hash array.

A vector for each document containing the presence or frequency of words in the document.

The ordering of words is not relevant.

Modeling database items



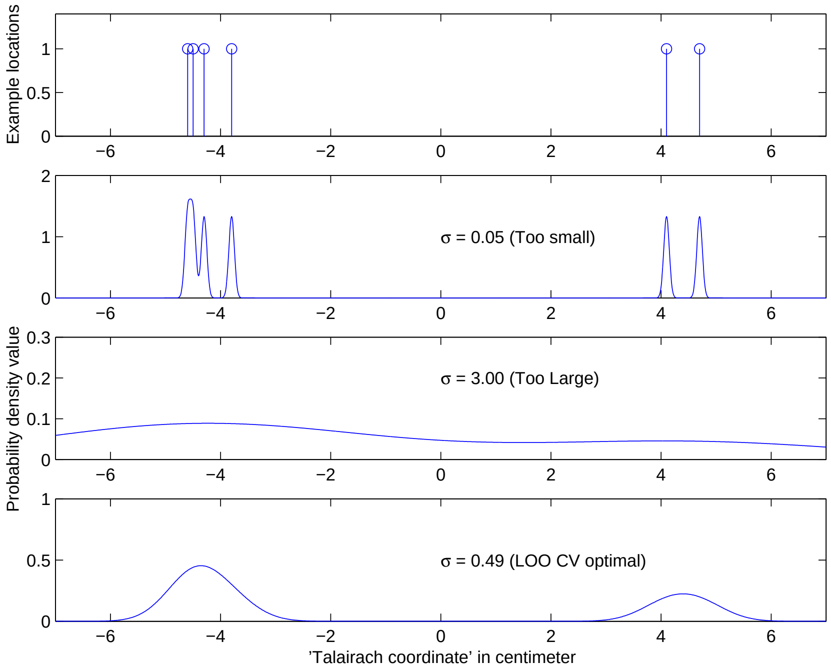
Extraction of Talairach coordinate. Example: $\mathbf{x} = (3.6, -7.6, 1.2)$.

Extraction of each word and phrase from the field "Lobar anatomy".

Example "lateral superior parietal" $\rightarrow$ $c \in$ {"lateral", "superior", "parietal", "lateral superior", "superior parietal", "lateral superior parietal"}.

Multiple data generated for one location.

Modeling Talairach coordinates



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

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

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

σ^2 fixed or optimized with leave-one-out cross-validation.

Condition on, e.g., anatomical label, behavioral domain c: $p(\mathbf{x}|c)$

Probability density for “cerebellum”

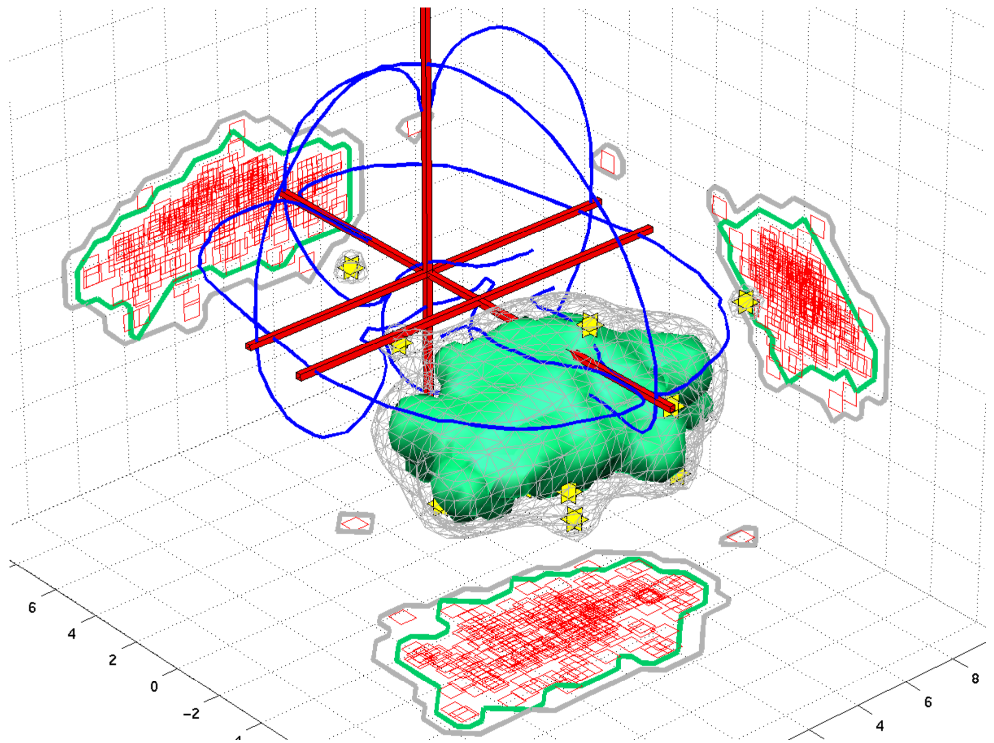


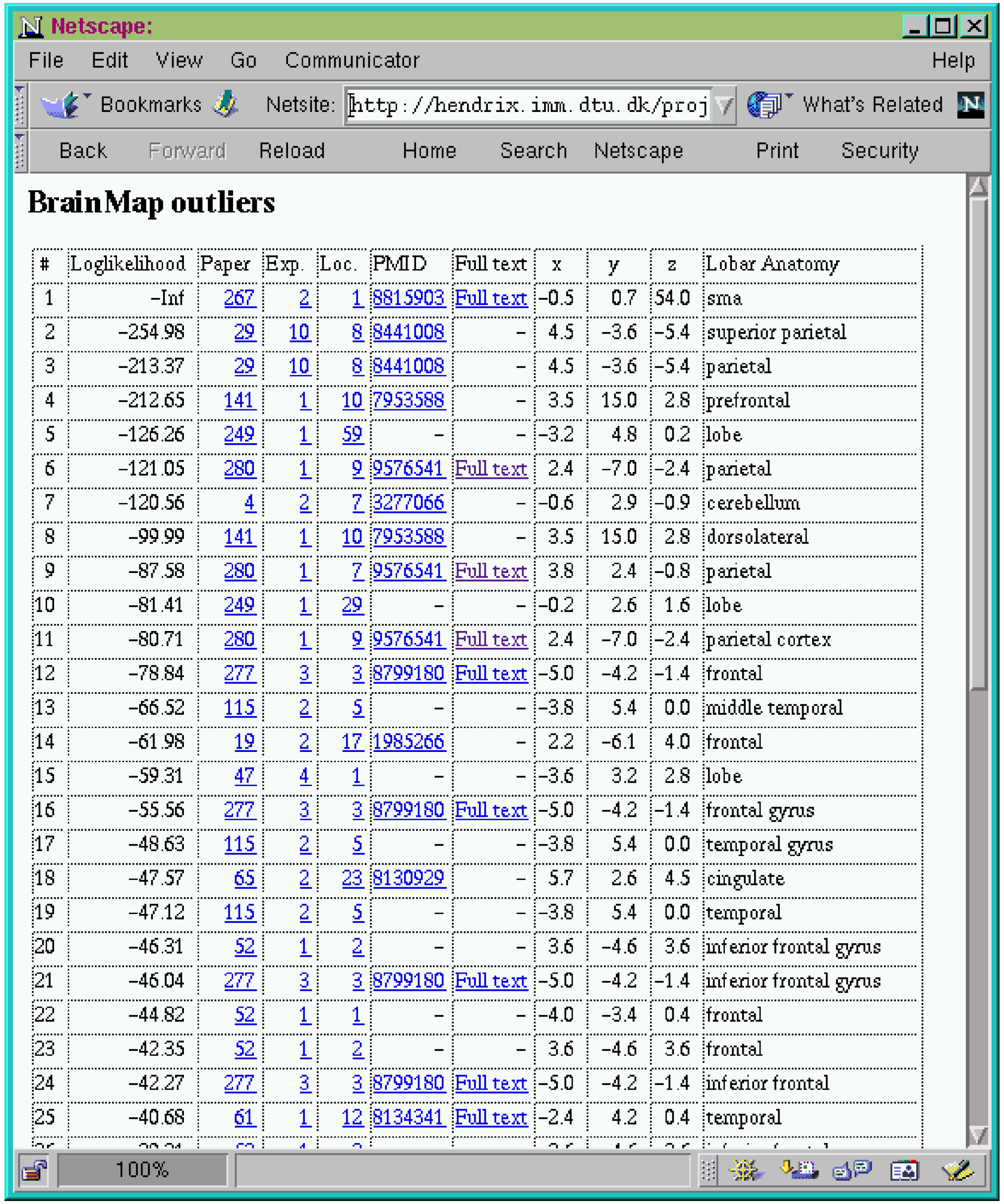
Figure 3: Densities from cerebellum locations. Yellow glyphs are the original BrainMap locations. Grey wireframe: Isosurface in the first level probability density estimate. Green surface: Isosurface in the second level.

Condition on anatomical label: $p(\mathbf{x}|c = \text{cerebellum})$.

Evaluate each location with respect to its probability densities: its “novelty”.

Robust estimate of $p(\mathbf{x})$ by excluding the 5% most extreme locations in a two-stage scheme.

Novelty detection by comparing all Talairach coordinates $\mathbf{x}_n$ with their associated $p(\mathbf{x}|c)$.



BrainMap outliers

#	Loglikelihood	Paper	Exp.	Loc.	PMID	Full text	x	y	z	Lobar Anatomy
1	-Inf	267	2	1	8815903	Full text	-0.5	0.7	54.0	sma
2	-254.98	29	10	8	8441008	–	4.5	-3.6	-5.4	superior parietal
3	-213.37	29	10	8	8441008	–	4.5	-3.6	-5.4	parietal
4	-212.65	141	1	10	7953588	–	3.5	15.0	2.8	prefrontal
5	-126.26	249	1	59	–	–	-3.2	4.8	0.2	lobe
6	-121.05	280	1	9	9576541	Full text	2.4	-7.0	-2.4	parietal
7	-120.56	4	2	7	3277066	–	-0.6	2.9	-0.9	cerebellum
8	-99.99	141	1	10	7953588	–	3.5	15.0	2.8	dorsolateral
9	-87.58	280	1	7	9576541	Full text	3.8	2.4	-0.8	parietal
10	-81.41	249	1	29	–	–	-0.2	2.6	1.6	lobe
11	-80.71	280	1	9	9576541	Full text	2.4	-7.0	-2.4	parietal cortex
12	-78.84	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	frontal
13	-66.52	115	2	5	–	–	-3.8	5.4	0.0	middle temporal
14	-61.98	19	2	17	1985266	–	2.2	-6.1	4.0	frontal
15	-59.31	47	4	1	–	–	-3.6	3.2	2.8	lobe
16	-55.56	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	frontal gyrus
17	-48.63	115	2	5	–	–	-3.8	5.4	0.0	temporal gyrus
18	-47.57	65	2	23	8130929	–	5.7	2.6	4.5	cingulate
19	-47.12	115	2	5	–	–	-3.8	5.4	0.0	temporal
20	-46.31	52	1	2	–	–	3.6	-4.6	3.6	inferior frontal gyrus
21	-46.04	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	inferior frontal gyrus
22	-44.82	52	1	1	–	–	-4.0	-3.4	0.4	frontal
23	-42.35	52	1	2	–	–	3.6	-4.6	3.6	frontal
24	-42.27	277	3	3	8799180	Full text	-5.0	-4.2	-1.4	inferior frontal
25	-40.68	61	1	12	8134341	Full text	-2.4	4.2	0.4	temporal

Automatic generated list.

Entries sorted according to novelty.

2nd and 3rd entry: More information in a phrase than in a word.

By "manual investigation" one finds that some of the interesting are database entry errors.

Finding similar items

[WOEXP 89] **Passively viewed scenes**. *Passive viewing of outdoor scenes, furnished rooms, landscapes and landmarks*. WOEXP: 89.

R. Epstein; N. Kanwisher. *A cortical representation of the local visual environment*. . *Nature* **392**(6676):598–601, 1998. PMID: 9560155. DOI: 10.1038/33402. WOBIB: 27.

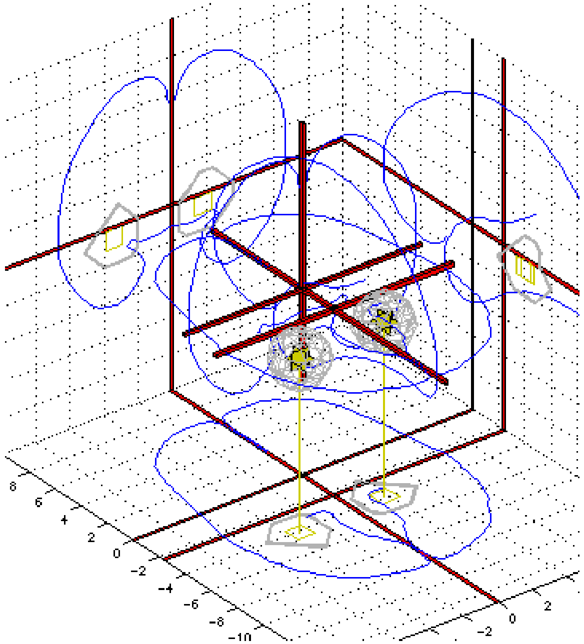
Perception,Vision – Places

Modality: fMRI

Asymmetry: 0.00000 (left: –1, right: +1)

VRML97 file (61 Kb)

x	y	z	Lobar anatomy	Functional area
18	–39	–6		Parahippocampal place area
–34	–42	–6		Parahippocampal place area



Related – positive correlated volumes

+2: 0.80010 (12) **Buildings visual objects**. *Visual object stimuli: Building versus faces*. WOEXP: 12.
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. DOI: 10.1038/87490. WOBIB: 5.

+3: 0.49922 (42) **Attention to musical instruments versus attention to consonant–vowels**. *Attend to sound and press a button when the target stimulus appeared*. WOEXP: 42.
K. Hugdahl; I. Law; S. Kyllingsbaek; K. Bronnick; A. Gade; O. B. Paulson. *Effects of attention on dichotic listening: an 15O-PET study*. . *Hum Brain Mapp* **10**(2):87–97, 2000. PMID: 10864233. WOBIB: 14.

+4: 0.45377 (97) **Visual object decision**. *Visual object decision with novel and chimeric, natural and artefact line drawings versus pattern discrimination*. WOEXP: 96.
C. Gerlach; I. Law; A. Gade; O. B. Paulson. *Perceptual differentiation and category effects in normal object recognition: a PET study*. . *Brain* **122 (Pt 11)**:2159–70, 1999. PMID: 10545400. WOBIB: 29.

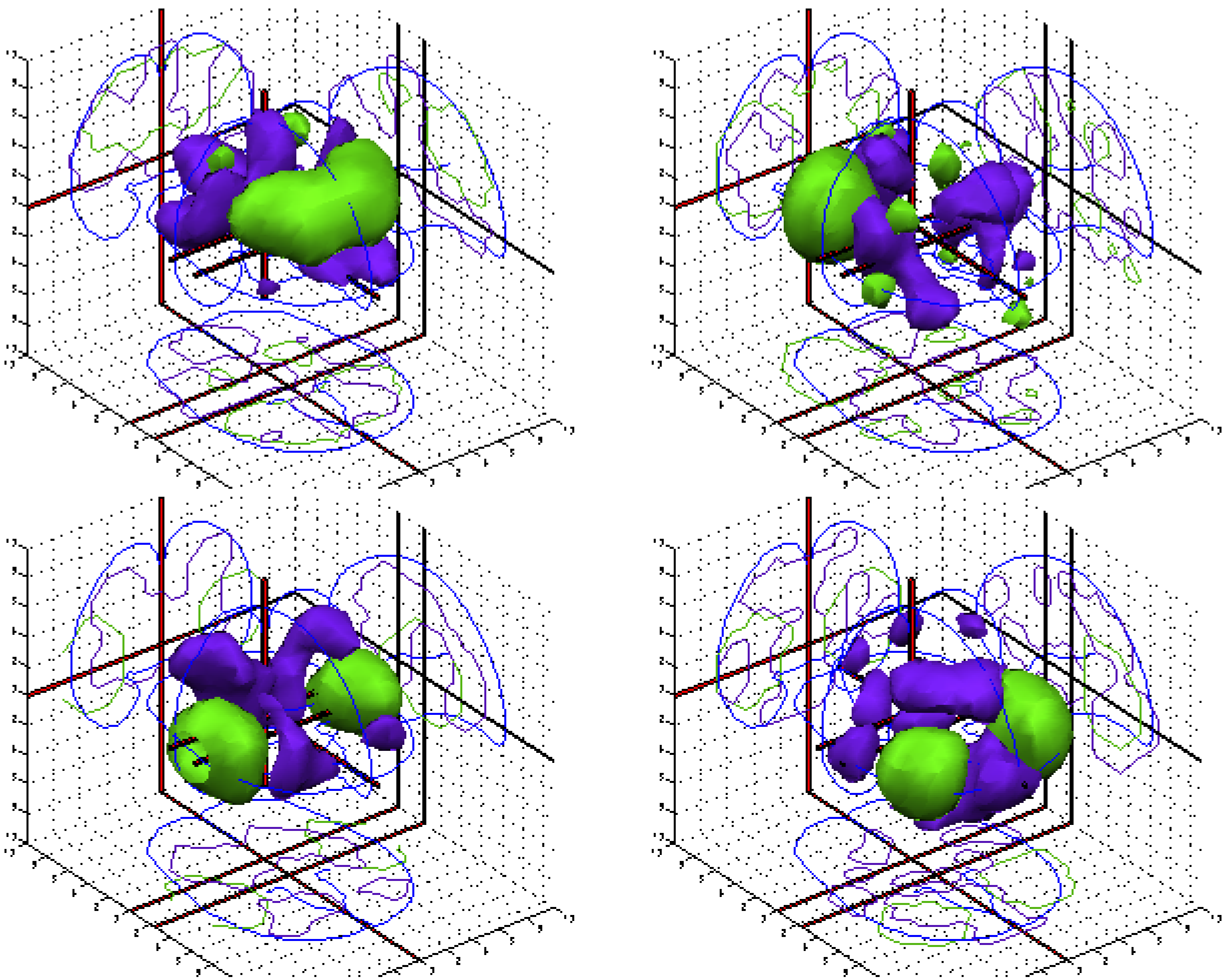
Each experiment a volume: $p(\mathbf{x}|\text{experiment} = \text{“89”}) \equiv \mathbf{z}_{89}$ sampled on a fixed 8mm grid

Similarity as a raw correlation coefficient between two volumes

$$s = \frac{\mathbf{z}_1'\mathbf{z}_2}{\sqrt{\mathbf{z}_1'\mathbf{z}_1}\sqrt{\mathbf{z}_2'\mathbf{z}_2}}$$

Sorted list of similar volumes.

Image-based indices: ICA



Independent component analysis of the $\mathbf{X}$(experiment $\times$ voxel) data matrix: $\mathbf{X} = \mathbf{AS} + \mathbf{U}$. $\mathbf{A}$ is the mixing matrix, $\mathbf{S}$ the sources.

ICA components: hand movement, visuospatial, words/verbs, audition, visual motion.

Figure shows both ends of the third to sixth source images $\mathbf{s}_3, \ldots, \mathbf{s}_6$. Data from Brede.

Image-based indices: Asymmetry

Left dominate	Asymmetry	Right dominate
	0.99902	[WOEXP 185] **Spatial neglect**. *Patients with spatial neglect and right brain damage from infarct or haemorrhage versus right brain damage patients without spatial neglect*. WOEXP: 185.
[WOEXP 5] **Visual artefact object**. *Decision or categorization of visual artefact*. WOEXP: 5.	−0.99219	
[WOEXP 114] **Categorization of artefacts**. *Categorization of visually presented artefacts versus categorization of natural objects, naming of artefacts and pattern discrimination*. WOEXP: 114.	−0.99219	
[WOEXP 137] **Names versus occupation**. *Retrieval and whispering of names from presented photographs of faces. Conjunction between newly learned face and famous face*. WOEXP: 137.	−0.99219	

"Experiment" left/right asymmetry: Count the number of locations in the left side X

$$P_{\text{Bin}} = \sum_{0}^{X} \binom{N}{X} 0.5^{N}. \qquad (1)$$

Normalize the value to $[-1;+1]$ range with $a = 1 - 2P_{\text{Bin}}$

When conditioning on anatomical labels:

- Left dominate (-1): 'motor', 'area', . . . , 'broca s area'.
- Right dominate (+1): 'anterior cerebellum',

Summary

Statistical data mining.

Heterogeneous data: text, and point sets (Talairach coordinates).

Transform the data to vectorial form.

Use statistical method to mine for knowledge.

References

Allison, T., McCarthy, G., Nobre, A., Puce, A., and Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex*, 4(5):544–554. PMID: 7833655.

Balslev, D., Nielsen, F. Å., Frutiger, S. A., Sidtis, J. J., Christiansen, T. B., Svarer, C., Strother, S. C., Rottenberg, D. A., Hansen, L. K., Paulson, O. B., and Law, I. (2002). Cluster analysis of activity-time series in motor learning. *Human Brain Mapping*, 15(3):135–145. http://www3.interscience.wiley.com/cgi-bin/abstract/89011762/. ISSN 1097-0193 [bibliotek.dk].

Drevets, W. C., Videen, T. O., MacLeod, A. K., Haller, J. W., and Raichle, M. E. (1992). PET images of blood flow changes during anxiety: Correction. *Science*, 256(5064):1696. PMID: 1609283. A previous functional neuroimaging study found correlation between anxiety and the temporopolar region. This study finds that it is more likely muscle signal from teeth-clenching.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601. PMID: 9560155. DOI: 10.1038/33402. ISSN 0028-0836 [bibliotek.dk].

Inoue, K., Kawashima, R., Sugiura, M., Ogawa, A., Schormann, T., Zilles, K., and Fukuda, H. (2001). Activation in the ipsilateral posterior parietal cortex during tool use: a PET study. *NeuroImage*. PMID: 11707103. DOI: 10.1006/nimg.2001.0942. WOBIB: 48.

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.

Reiman, E. M., Fusselman, M. J., Fox, P. T., and Raichle, M. E. (1989). Neuroanatomical correlates of anticipatory anxiety. *Science*, 243(4894 Part 1):1071–1074. PMID: 2784226.