Knowledge discovery in neuroimaging databases

Lars Kai Hansen

DTU Informatics
Technical University of Denmark

Co-workers:

Finn Å. Nielsen, Daniela Balslev, Bart Wilkowski, Marcin Szewczyk



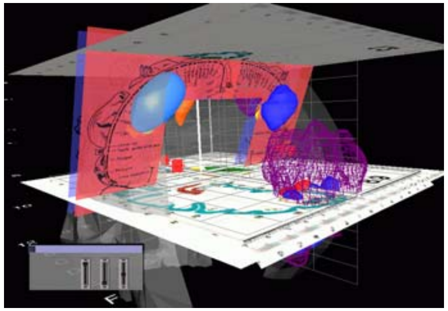
Cimbi Center for integrated molecular brain imaging

DTU

OUTLINE

- Knowledge Discovery?
- Nature of neuroimaging databases
- BrainMap,
- Brede demo
- Machine learning for meta analysis
- Search for similar volumes
- A Science 2.0 interface to populate imaging databases

What is Knowledge discovery?



Knowledge discovery can broadly be defined as *'extraction of implicit and potentially useful information from data'*.

A distinction can be made between data mining and knowledge discovery: The knowledge discovery process takes the raw results from data mining and place it in context.

The useful information – cf. search – is uncovered through the use of machine learning techniques and often involves visualization.

Fayyad et al.: From Data Mining To Knowledge Discovery: An Overview. In Advances In Knowledge Discovery And Data Mining , eds. U.M. Fayyad et al. AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 1-34.

Knowledge discovery in bio-medicine

- Neuroimaging and other bio-medical areas are characterized by an extremely heterogenous, chaotic, and noisy data collection process.
- Biological variability is ~infinite compared to other database applications such e.g. “market basket analysis”
- Increasing specialization in the bio-sciences calls for knowledge discovery for inventions to transcend from anecdote to science, i.e., for efficient division of labour

- The traditional “review/meta analysis” mechanism is hampered by the exponentially increasing volume of scattered results across a vast set of journals/conferences/e-lists

James A. Evans Electronic Publication and the Narrowing of Science and Scholarship Science 18 July 2008:
Vol. 321. no. 5887, pp. 395 - 399

Why is the databasing progress so slow in neuroimaging?

- Hints
  - Neuroimaging findings are more ill-defined than e.g. sequence information in bioinformatics?
  - The rewards from sharing are less obvious in neuroimaging?
  - Angst ...imaging is so hard thus... my competitors will find errors / spot a Nobel prize ... in my data

Nature of neuroimaging databases

- Database
  - Neuroimaging results: Data sets (fMRIDC, Neurogenerator)
  - Neuroimaging results: Activation foci/coordinates
  - Neuroimaging results: Metadata paradigm subject social networks etc
  - Database model / Ontologies
  - Links to neuroscience databases, behavior
  - Inclusion criteria
  - Search functionality

- Challenges
  - Lack of interoperability at the group level, consortia level, between sub-disciplines (e.g. PET vs fMRI vs EEG)
  - Poor specification of targets, biology and behavior

- Knowledge discovery tools
  - Machine learning tools for meta-analysis can operate in noise

Fox and Lancaster's BrainMap®

BrainMap

The Social Evolution of a Human Brain Mapping Database

Angela R. Laird, Jack L. Lancaster, and Peter T. Fox*

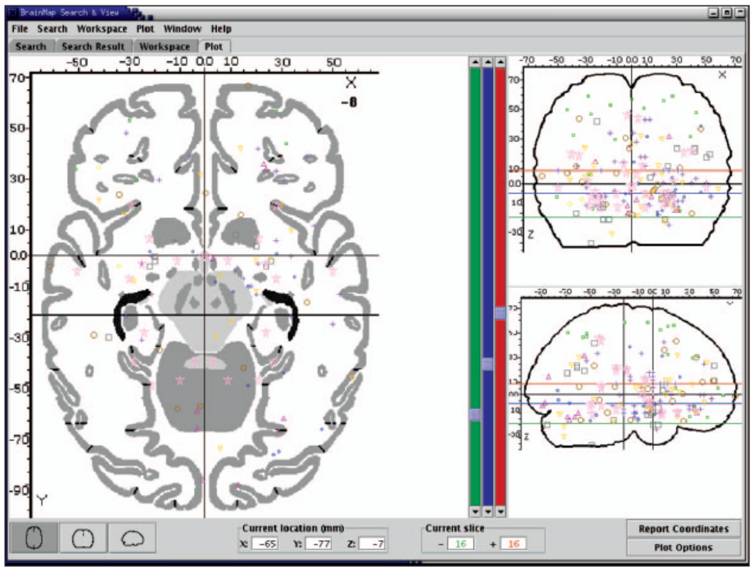


Fig. 5. A screen shot of a graphical user interface to the BrainMap database with Talairach coordinates plotted after a search for experiments on olfaction

Volume

Papers: 1515

Experiments: 6943

Locations: 55549

Paradigm Classes: 77

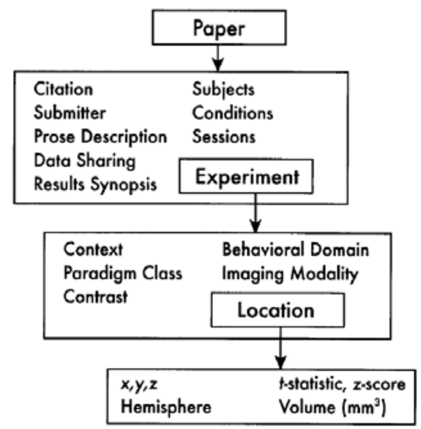


Fig. 2. The BrainMap coding scheme. Each paper is coded according to three levels of information: *Paper* (a set of one or more experiments reported in a single publication), *Experiment* (comparisons or contrasts that are generated when comparing different behavioral conditions), and *Location* (*x*, *y*, *z*-coordinates of activation).

Laird, Lancaster, Fox: Neuroinformatics (2005)

Brede database and search engine

- Main component
  - data from functional neuroimaging papers reporting activation foci as Talairach coordinates.
- Structure initially inspired by BrainMap
- Sandbox for knowledge discovery in neuroimaging
- Distributed with the Brede neuroinformatics toolbox.

- Designed, programmed, and maintained by Finn Årup Nielsen 

Brede database identifiers

WOBIB identifier for a 'bib' structure, i.e., a published paper. This contains one of more 'exp' structures.

WOEXP identifier for a 'exp' structure, i.e., an experiment in a scientific paper. This can, e.g, correspond to a 'contrast'. An 'exp' structure might have one or more 'loc' structures.

WOEXT identifier for a 'ext' structure: An 'external component', e.g., a cognitive component. Each 'exp' structure in the Brede database will usually have one or more WOEXT associated with it.

WOPER identifier for a 'per' structure: a person, usually an author

WOROI identifier for a 'roi' structure: a region of interest, i.e., a brain area referred to by lobar anatomy, or a functional or Brodmann area.

Database inter-operability

The Brede database is hyper-linked to other neuroscience databases

- Each 'bib' item is linked to Entrez-PubMed. E.g WOBIB: 52 is linked to PMID: 12507950.
- Some Brede items are linked to items in fMRIDC.
- Some 'ext' items are linked to MeSH terms of the U. S. National Library of Medicine. These are linked to the MeSH Browser, e.g., The Brede database 'Pain' (WOEXT: 40) is linked to MeSH 'Pain' (MeSH UID: D010146).
- Some of the 'ext' items that are associated with genes are linked to standard bioinformatics databases such as Ensembl, Entrez-Protein, Genecards, GenomeNet, GENSAT, PubGene, see, e.g., the 5-HT2A receptor. Other 'ext' items are linked to SenseLab (e.g., WOEXT: 233 - GABA-A receptor), OMIM (e.g., WOEXT: 346 - Apolipoprotein E gene) and the English version Wikipedia.
- Some 'Roi' items (brain regions) are linked to items in the BrainInfo/NeuroNames database, see, e.g., WOROI: 5 - Posterior cingulate gyrus, or items in the CoCoMac database, e.g,. WOROI: 96 - Posterior insula; or the Internet Brain Volume Database (IBVD), e.g., Insula.

Brede demo

- hendrix.imm.dtu.dk/services/jerne/brede/

Brede database

Jerne > Brede database

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

Experiments (Exp): Alphabetic | Asymmetry | ICA | NMF | Novelty | SVD | WOEXP | WOEXT

External Components (Ext): Alphabetic index | Map | Roots

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

[ ] Search Brede Database

0 0 0 Location Search

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).

Cimbi Center for integrated molecular brain imaging

Lars Kai Hansen
IMM, Technical University of Denmark

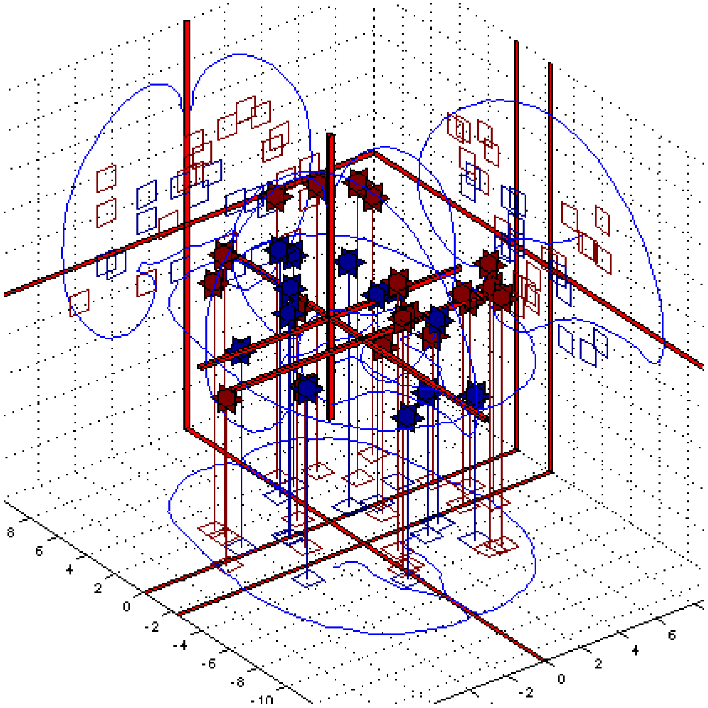
Brede database and search engine

Volume

- Papers: 180
- Experiments: 586
- Coordinates: 3912

Features

- Google based search
- Ontology
- On-line visualization



Machine learning tools in Brede

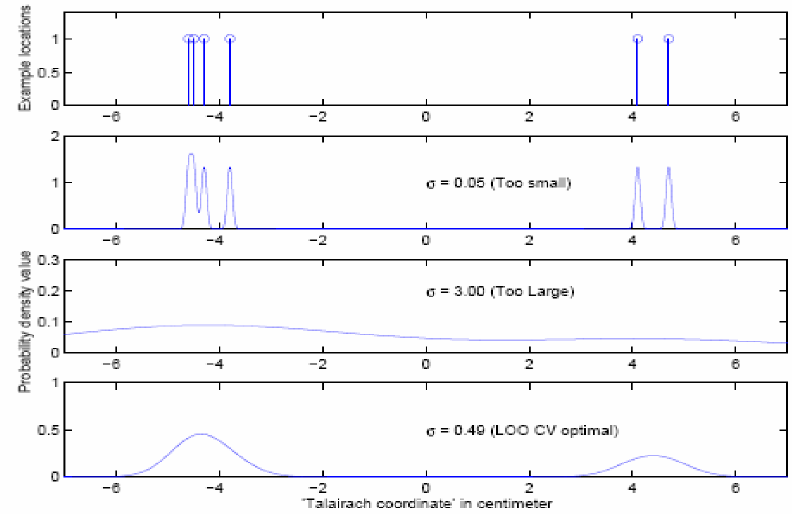
- SPM reconstruction
- Conditional density modeling for novelty detection
- Conditional density modeling for paper/exp similarity measure
- Combined text and foci mining

Cimbi
Center for integrated molecular brain imaging
Lars Kai Hansen
IMM, Technical University of Denmark
DTU

Reconstruction of SPM's from foci

Estimate label/metadata conditional densities P(foci(x,y,z)|term) with Parzen windows, estimated with LOO cross-validation

$$p(x\,|\,D) = \frac{1}{N}\sum_{n=1}^{N} \frac{e^{-\frac{(x-x_n)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{3/2}}$$



Densities are sampled (if necessary) in voxelated space

Peter E. Turkeltaub et al. Meta-Analysis of the Functional Neuroanatomy of Single-Word Reading: Method and Validation. NeuroImage 16, 765–780 (2002)
Nielsen & Hansen: Modeling of Activation Data in the BrainMap™ Database: Detection of Outliers. Human Brain Mapping 15:146–156(2002)

Statistical modeling of foci & meta-data

- Density modeling of foci distribution reveals novelty/outliers
- Outliers defined as foci with relative low probability density
- We model *conditional densities*, to increase outlier sensitivity

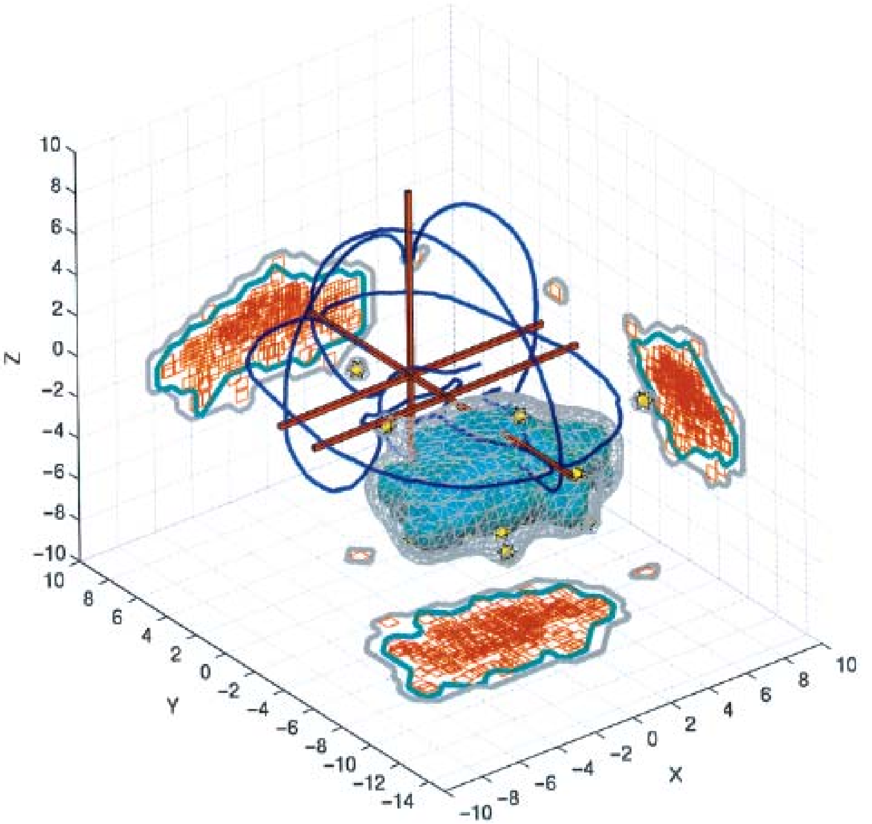
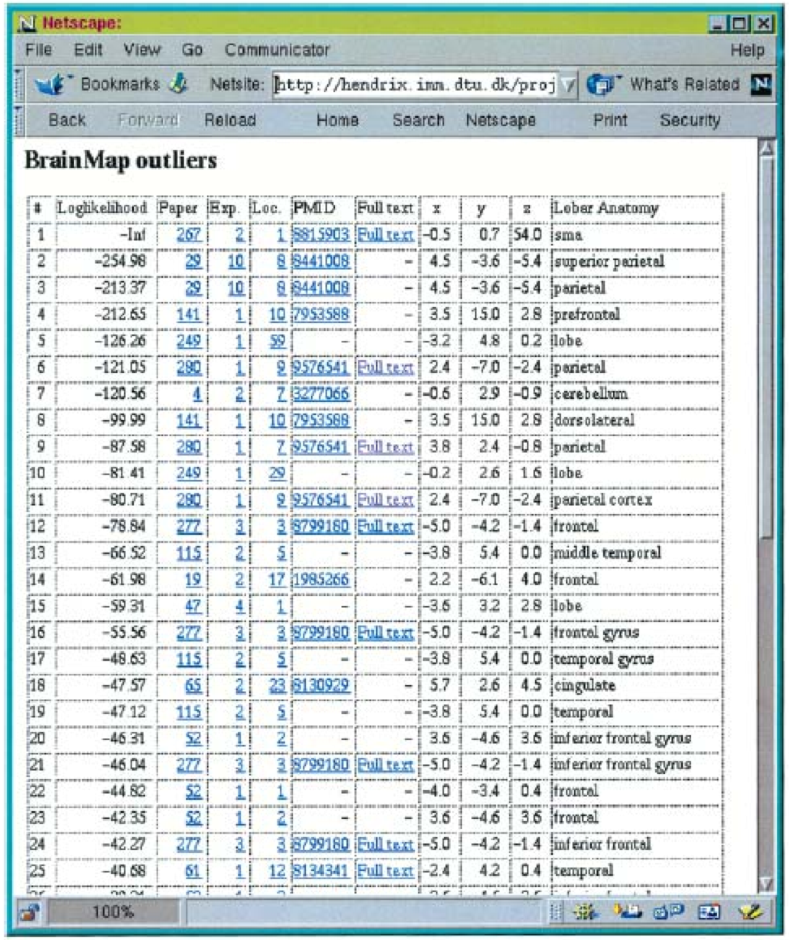


Figure 2.
Probability density estimate of the "cerebellum" class in Talairach space in a Corner Cube Environment. The wireframe-model is the first stage probability density estimation where all the locations are included and the polygon model is the second stage probability density estimate where the 5% most extreme are excluded. Note that two isolated "blobs" created by isolated, outlying locations were eliminated going from the first to the second level density. This figure as well as Figures 4 and 5 are made with the Brede Matlab toolbox available at http://hendrix.imm.dtu.dk/software/brede.

Result

- A list of the globally least probable entries was inspected
- Authors contacted to check for mistakes or novelty of interest
- Worst case: 50 cm outside brain volume
- Mirroring errors in entry of coordinates
- Conditional density allows detection of outliers wrt label



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

Figure 3.

An automatically generated list of those locations estimated to have the highest novelty. "Paper," "Exp.," and "Loc." correspond to the identifiers used in the BrainMap database. X, y, z, and "Lobar anatomy" are the associated fields in the database with the coordinates in centimeter and the "loglikelihood" is our novelty measure. The "Full text" column indicates whenever it is possible to extract a link from the Entrez-PubMed to the electronic full text of the articles.

Confer with other volume definitions

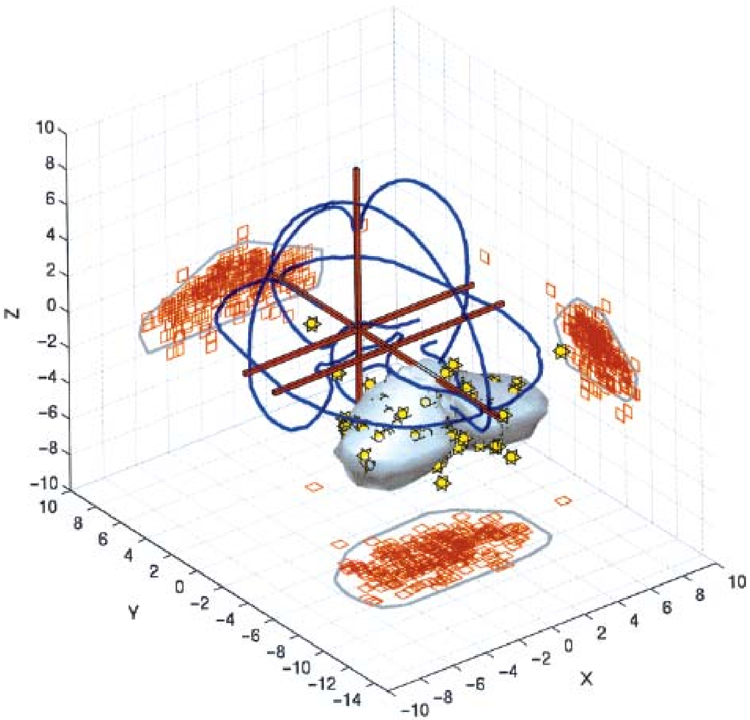


Figure 4.
Surface of the cerebellum from the Talairach Atlas with the "cerebellum" locations. The inferior part of the cerebellum is not in the atlas, thus not in the visualization. The contour shadows are the convex hull of the digitized Talairach cerebellum.

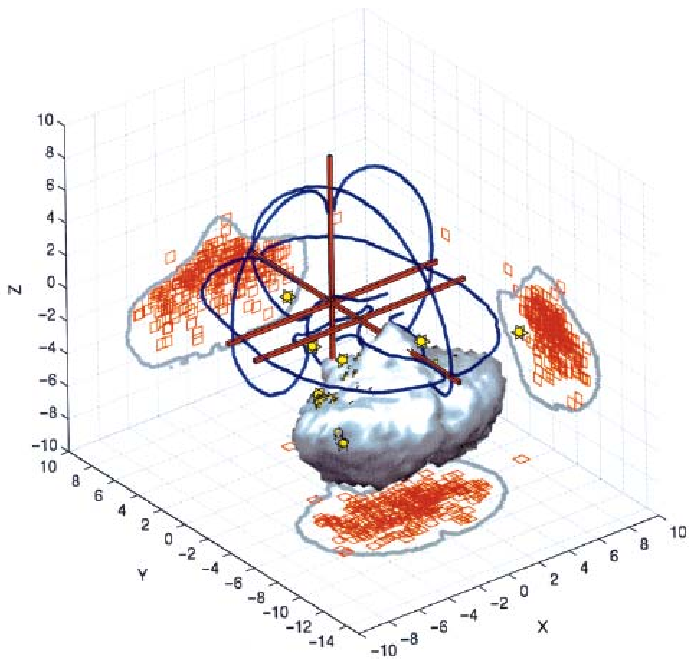


Figure 5.
The ICBM cerebellum with all the cerebellum locations from BrainMap. The locations have been transformed by the inverse operation of Brett's [1999] nonlinear transformation (see text).

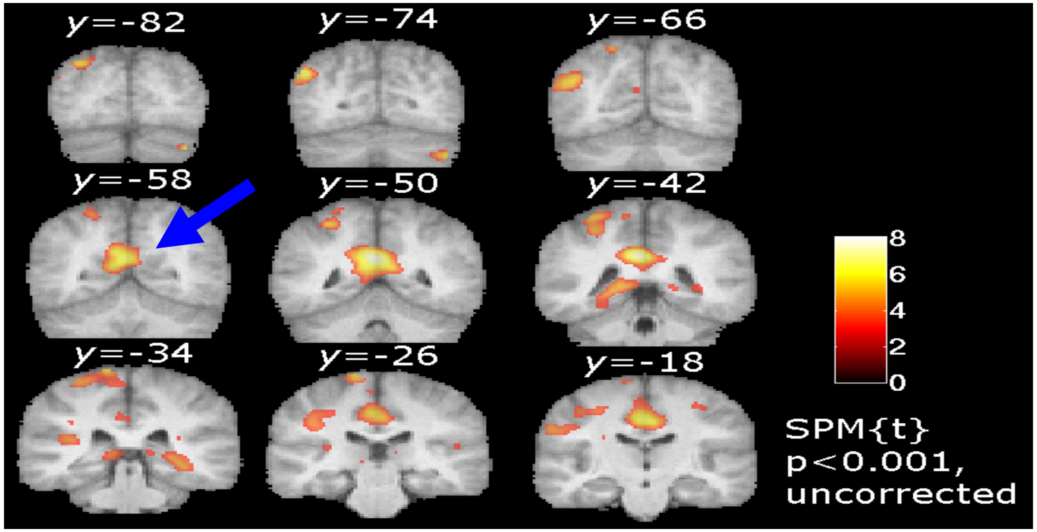
Nielsen & Hansen: Modeling of Activation Data in the BrainMap™ Database: Detection of Outliers. Human Brain Mapping 15:146–156(2002)

Meta-analysis example

Mining the posterior cingulate cortex

- PCC is a cyto-architecturally well defined brain region (Vogt et al, 2001)

- However, no textbook consensus about its function!



Finn Årup Nielsen, Daniela Balslev, Lars Kai Hansen, "Mining the Posterior Cingulate: Segregation between memory and pain components. NeuroImage, 27(3):520-532, (2005)

Text modeling: ”Bag of words”

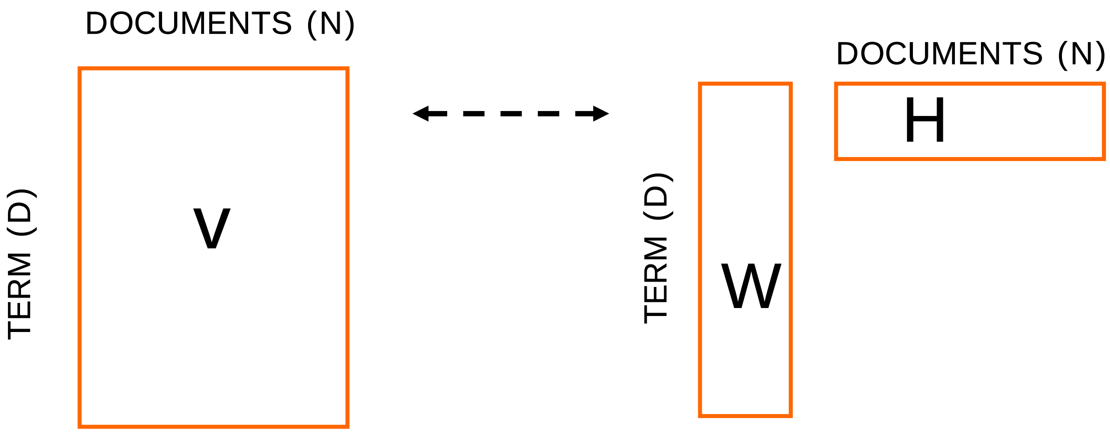
- 271 abstracts of functional imaging studies that responded to PubMed query (march 13th, 2003) :
  - *("posterior cingulate" OR "posterior cingulum" OR "retrosplenial" OR "retrosplenium")*
  - *AND*
  - *("magnetic resonance imaging" OR "positron emission tomography")*

- Create term list from abstracts starting from all words (D_0 =4792) screened with stop word lists to eliminate irrelevant words, PubMed stopwords, and an in-house manually created list including anatomical terms irrelevant for ”cognitive” tasks (D=549)

- Form a term x document frequency of occurrence matrix V of dimension (549x271)

Factor model

- Represent the datamatrix by a low-dimensional approximation



$$V(i,j) \approx \sum_{k=1}^{K} W(i,k) H(k,j)$$

Component/topic vocabulary

Component topic expressions

Matrix factorization: SVD / PCA, NMF, Clustering

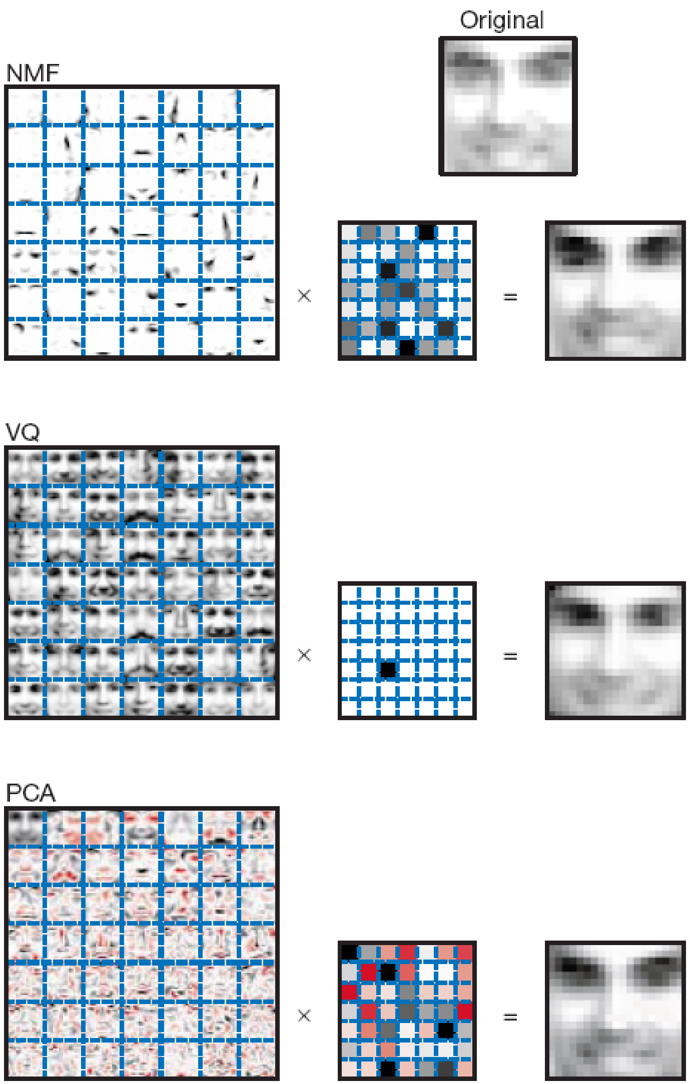


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting of $n = 19 \times 19$ pixels, and constituting an $n \times m$ matrix V. All three find approximate factorizations of the form $V \approx WH$, but with three different types of constraints on W and H, as described more fully in the main text and methods. As shown in the 7×7 montages, each method has learned a set of $r = 49$ basis images. Positive values are illustrated with black pixels and negative values with red pixels. A particular instance of a face, shown at top right, is approximately represented by a linear superposition of basis images. The coefficients of the linear superposition are shown next to each montage, in a 7×7 grid, and the resulting superpositions are shown on the other side of the equality sign. Unlike VQ and PCA, NMF learns to represent faces with a set of basis images resembling parts of faces.

Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung*†

** Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA*
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

NATURE | VOL 401 | 21 OCTOBER 1999 | www.nature.com

Multinomial mixture model, V is a matrix of 'counts'

$$\frac{V(i,j)}{\sum_{i',j'} V(i',j')} \approx \sum_{k=1}^{K} W(i,k) H(k,j)$$

$$\frac{V(i,j)}{\sum_{i',j'} V(i',j')} \approx P(i,j) \approx \sum_{k=1}^{K} P(i,k) P(j,k) P(k)$$

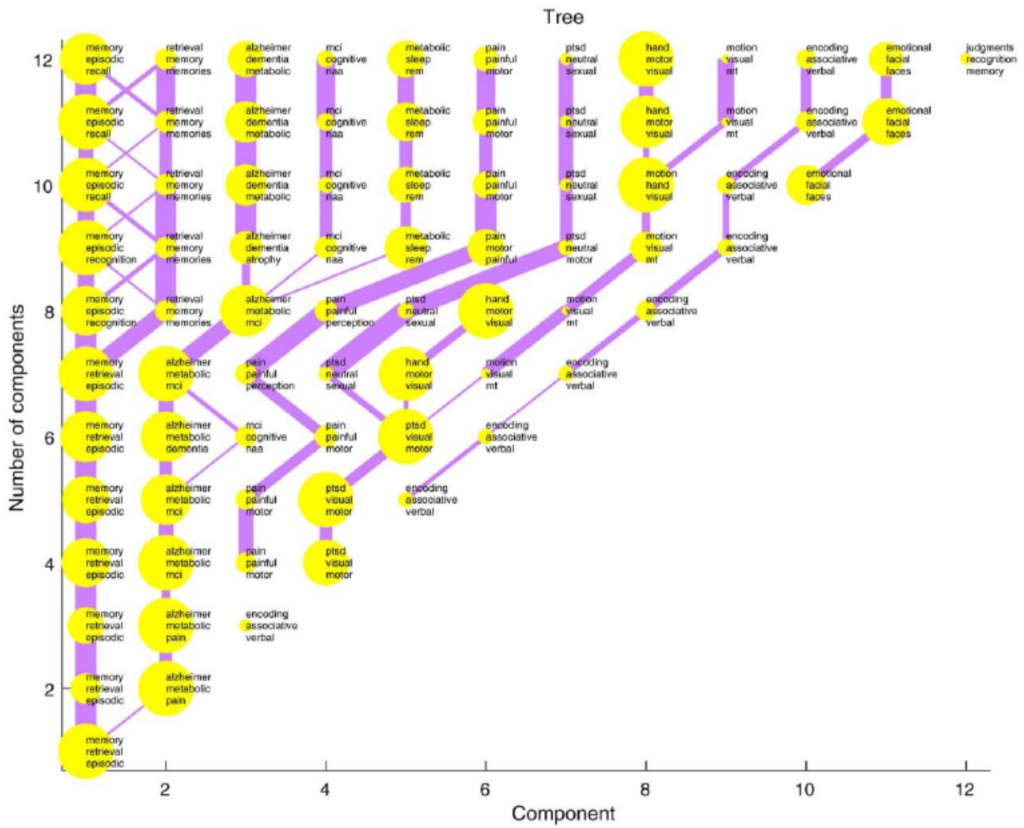
Terms — Documents

Text mining result: “The topic tree”

We created a tree of solutions to investigate the stability of vocabularies estimated by the model for increasing K

Links measure similarity of vocabularies

Finds independent components and locate: ”memory” and ”pain” components.



Finn Årup Nielsen, Daniela Balslev, Lars Kai Hansen, ”Mining the Posterior Cingulate: Segregation between memory and pain components”. NeuroImage, 27(3):520-532, (2005)

Hypothesis testing in retrieved components

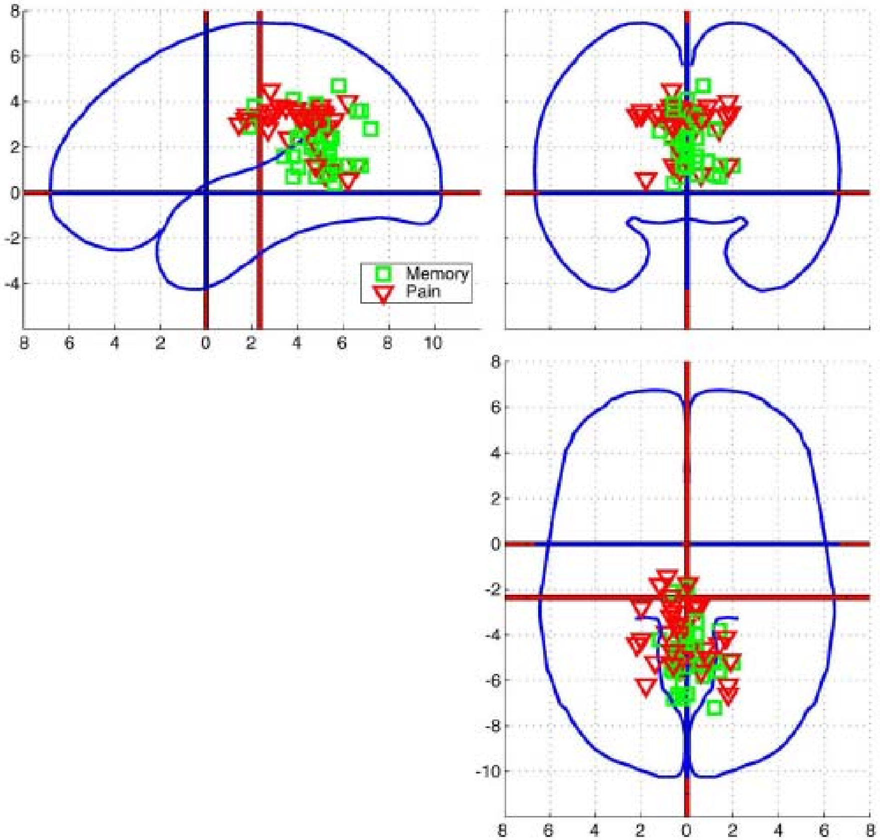


Fig. 3. Distribution of memory and pain brain activations in the posterior cingulate cortex shown on a sagittal plot: y is the AP axis with posterior as negative. The blue outline follows that of the Talairach atlas. The gray outline is an isocurvature in a probability volume for posterior cingulate cortex based on modeling of coordinates from the Brede database. Green squares are associated with "memory" articles and red triangles with "pain" articles.

Extract coordinates for abstracts associated with dominant components:
"Pain" and "Memory"

Significant difference of the coordinate sets

Not in conflict with major reviews

Similarity metric for search

- Use the correlation between reconstructed volumes as a distance metric

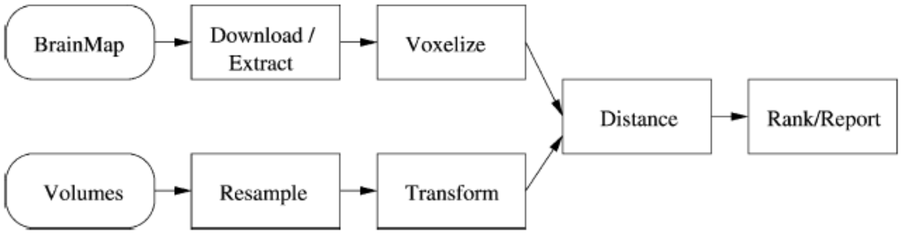


Fig. 1. Pipeline for finding related volumes for data from the BrainMap database.

$$v(z) = \frac{1}{J}(2\pi\sigma^2)^{-3/2}\sum_{j}^{J}\mathrm{sgn}(z_j)\exp\left(-\frac{1}{2\sigma^2}(z-z_j)^{\mathrm{T}}(z-z_j)\right). \quad (1)$$

Nielsen, Hansen:" Finding related functional neuroimaging volumes"
Artificial Intelligence in Medicine 30 (2004) 141–151

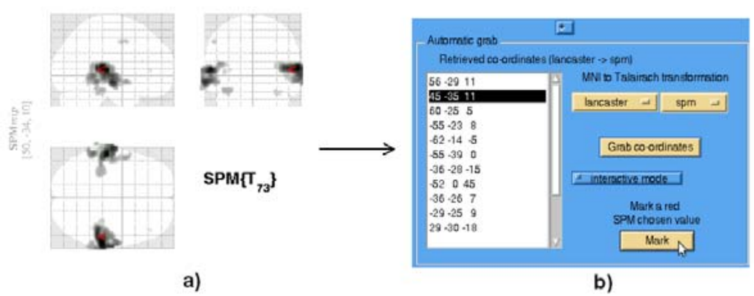
Vision / Wishlist for a neuroimaging search engine

- A useful neuroimaging search engine will
  1. Index 'published results' from neuroimaging
  2. Index published results from cognitive and behavioral psychology
  3. Have a query interface that allows query for typical situations that arise in writing a neuroimaging paper

- An *extremely* useful neuroimaging search engine will
  1. Index 'published results' from fields relevant to neuroimaging
  2. Index raw data from many disciplines
  3. Index workflows from many disciplines
  4. Have a proactive query interface, e.g., an interface that intercepts the (e.g., SPM) workflow and suggest steps based on 2-3.

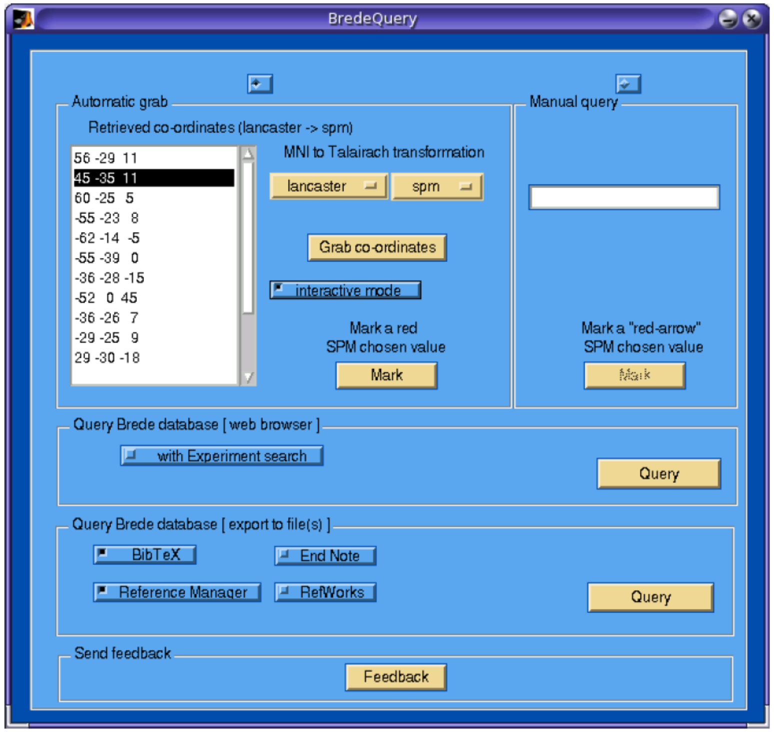
A Science 2.0 interface

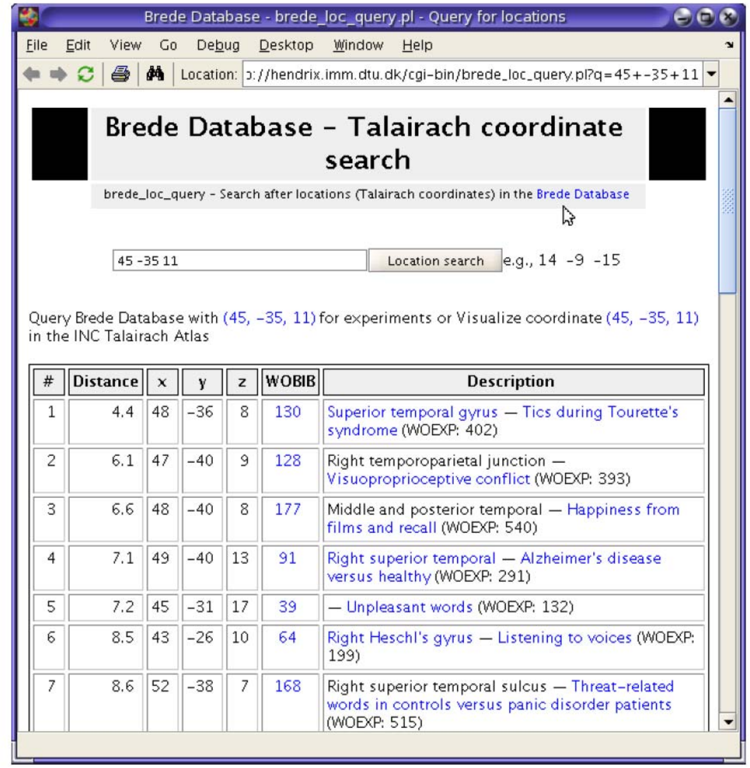
- How can we make data entry so attractive that user generated content can replace our current misery?
- BredePlugin for SPM
  - Intercepts the SPM pipeline and extract coordinates
  - Submit coordinates to Brede
  - Obtain list of relevant paper…proactive in relation to paper writing



Bart Wilkowski et al.: Coordinate-based meta-analytic search for the SPM pipeline, (Submitted)

BredePlugin





Envision a two level (Flickr-like) upload of coordinates

“Private”- shared with designated collaborators

“Public”- broadcast after publication & edit

Conclusions

- Knowledge discovery increasingly important for neuroimagers?!
- Knowledge discovery is hampered by the slow growth of database. An attractive Science 2.0 interface may assist (...let's vote...).
- Machine learning is a platform for producing generalizable models and visualizations in complex, noisy database.
- Major current advances are based on activation foci mining....we need to support the upload of more informative data structures...raw data!

Acknowledgments

Lundbeck Foundation (www.cimbi.org)
NIH Human Brain Project grant (P20 MH57180)
MAPAWAMO / EU Commission
Danish Research Councils

www.imm.dtu.dk/cisp
hendrix.imm.dtu.dk