Data mining and neuroinformatics

Finn Årup Nielsen

Technical University of Denmark
Department of Informatics and Mathematical Modelling
Lyngby, Denmark

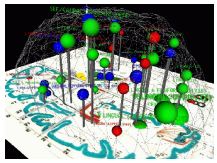
Cimbi
Center for integrated molecular brain imaging

The Brede Tools

Brede Toolbox: Matlab neuroinformatics toolbox with, e.g., analysis of neuroimages

Brede Database: A database with result data from neuroimaging

Brede Wiki: A wiki for structured neuroscience data – paper, brain regions, ”topics”, researchers, organizations



Cimbi Center for integrated molecular brain imaging

Finn Årup Nielsen
IMM, Technical University of Denmark

2

DTU

Neuroscience publishes an overwhelming number of studies and these are often inconsistent. To deal with that we need neuroinformatics databases and tools that can store the result data, integrate them and do quantitative analysis and present the results in a useful format.

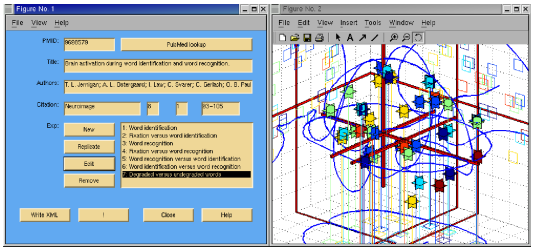
My neuroinformatics efforts are mostly centered around the Brede line of tools.

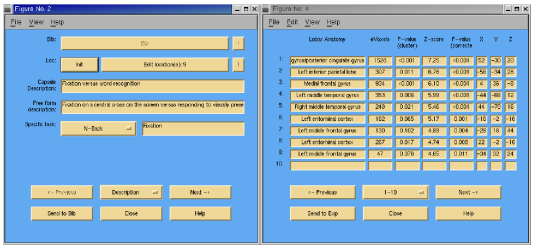
The first one, the Brede Toolbox, started out as a Matlab Toolbox for neuroinformatics visualization and data mining of neuroimaging result, but now have functionality for, e.g., analysis of neuroimages.

Together with the Brede Toolbox is the Brede Database which is a small database of neuroimaging papers and their result data as well as ontologies for brain regions and topics. The example of neuroinformatics data mining that I will show you are where Brede Toolbox is used to analyze data from the Brede Database.

A major bottleneck in neuroinformatics databasing is data entry. Database curators have a very hard time keeping up with the papers published, and to explore new means of data entry I have set up the Brede Wiki, which is a structured wiki with text and data from published papers as well as information about brain regions, topics, researchers and organizations.

Brede Toolbox

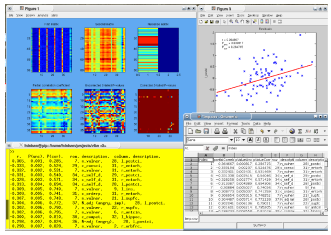




Matlab Toolbox for neuroinformatics

Data entry

Data mining



Cimbi Center for Integrated molecular brain imaging

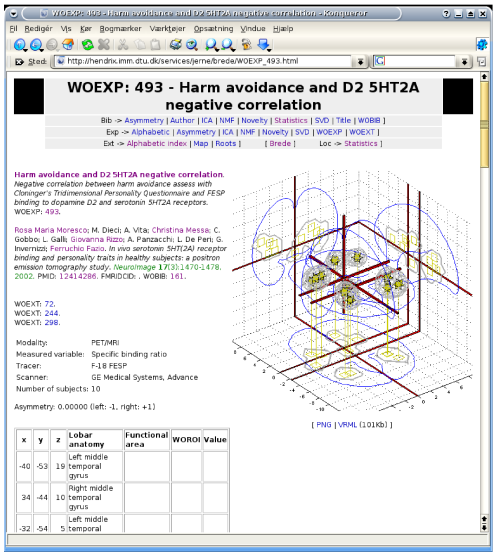
Finn Årup Nielsen
IMM, Technical University of Denmark

3

Taking a closer look on the Brede Toolbox that is available on the Web.

The toolbox enables data entry of results from published neuroimaging studies, analysis and visualization of brain coordinates. But its modular structure also allows data mining of neuroimages, region of interest data and text data.

Brede Database

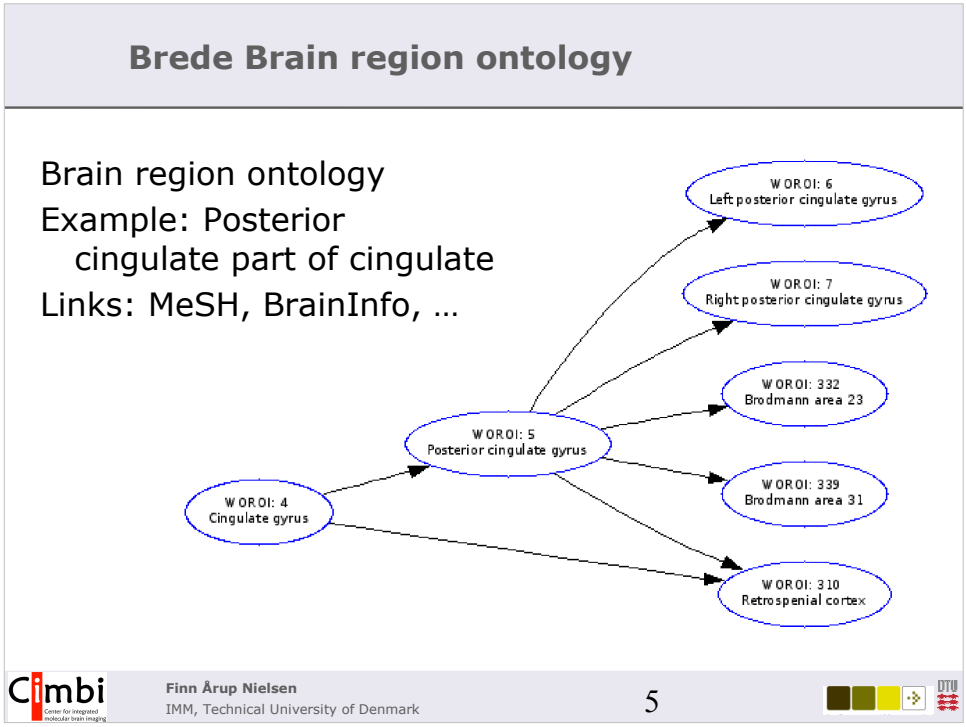


- Available on the Web
- Stereotaxic coordinates converted to volumes
- Bibliographic information, experiment description (scanner, paradigm), labels to ontologies

With the Brede Toolbox I type in data for the Brede Database. This database is also available on the Web and you see one of the Web pages here.

Presently it contains close to 4000 brain coordinates from 186 neuroimaging papers. With functions from the Brede Toolbox the brain coordinates can be converted to a volume via so-called kernel density estimation.

Apart from brain coordinates the Brede Database also records detailed bibliographic information, and description of the experiment such as scanner type and experimental paradigm. Some of this information is linked to items in ontologies.



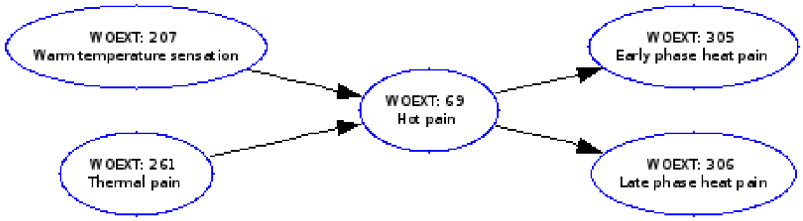
The Brede brain region ontology structures brain regions in a hierarchical graph, capturing which brain regions are part of a larger region, e.g., it tells that the cingulate gyrus has a part called posterior cingulate gyrus which in turn has the left posterior cingulate gyrus as a part.

The ontology also records the different naming variations, e.g., the cingulate gyrus may be called 'gyrus cinguli'.

Each brain region in Brede is linked to a number of other brain region ontologies such as MeSH of NIH, NeuroNames, NeuroLex and the CoCoMac brain connectivity database.

Also digital atlases are linked so brain regions can be associated with specific voxels in the neuroimage.

Brede Database topic ontology



Topics organized in a hierarchy

Examples: Memory, episodic memory, episodic memory retrieval, obsessive compulsive disorder, disgust, 5-HT2A receptor

Others: BrainMap taxonomy. MeSH. Under development: Cognitive Atlas (Poldrack), Cognitive Paradigm Ontology (Laird, Turner)

Another ontology captures topics in a hierarchy. These topics can be cognitive functions or mental disorders, e.g., hot pain as part of thermal pain.

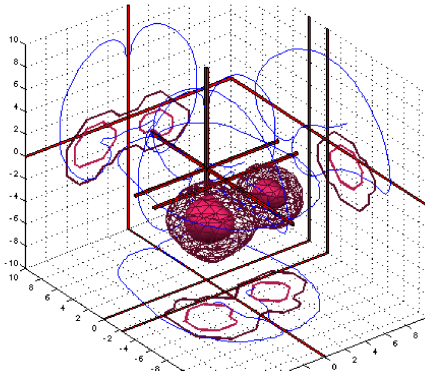
Other concrete examples are memory, episodic memory, episodic memory retrieval, OCD, disgust, 5-HT2A receptor

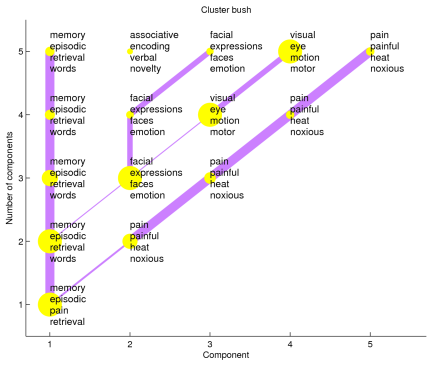
These topics are used to label each individual neuroimaging experiment result.

The items in the ontology are linked to MeSH terms. One may ask why its is necessary to develop a further ontology and not just rely on MeSH alone. I find the MeSH is not fine-grained enough for labeling.

Other research groups have also felt it necessary to develop ontologies for cognitive functions: There is the Cogntive Atlas and the Cognitive Paradigm Ontology.

Data mining without labels





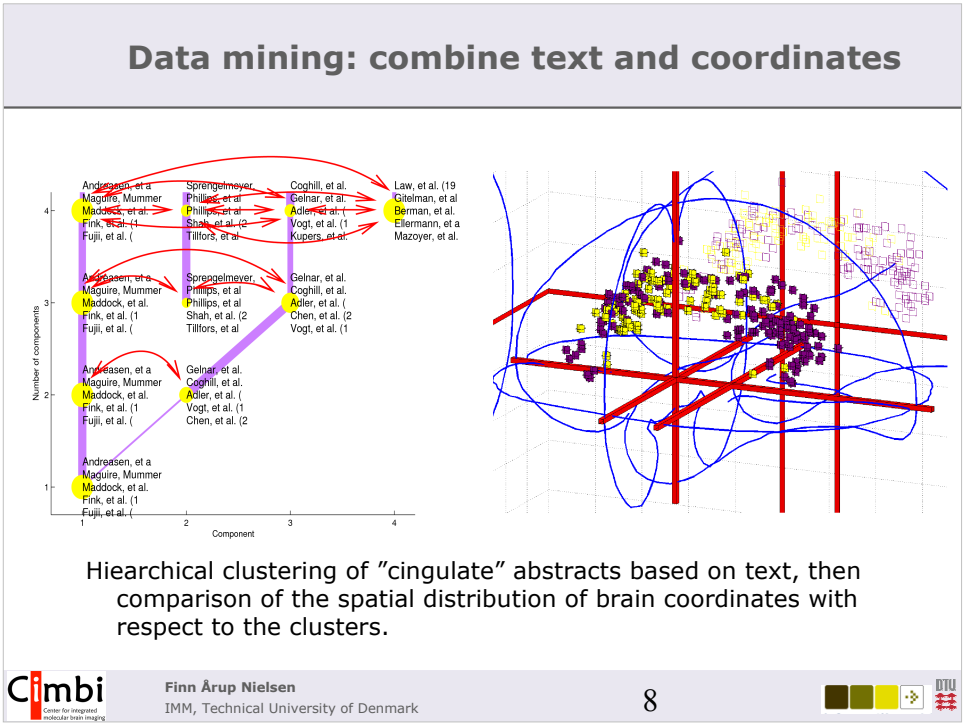
Non-negative matrix factorization (NMF) on volume from brain coordinates via kernel density estimation (KDE) and text from abstracts (”bag-of-words”)

When we finally have result data from neuroimaging experiments and ontologies setup in the Brede Database it will be possible to do large-scale data-mining across the database.

One way is to use so-called 'unsupervized' multivariate analysis, such as principal component analysis and cluster analysis. What we often use is the method called 'non-negative matrix factorization' or NMF, since this method is particularly useful for non-negative data which often arise with the data we have at hand, e.g., when the brain coordinates are converted to a volume via kernel density estimation the volume becomes non-negative.

One result from an application of NMF across all the 586 experiments in the Brede Database is shown here at the left. It is a 3-dimensional visualization of the voxels that are loaded highly on a specific NMF component, and they appear in the fusiform and parahippocampal gyrus. When examining the associated experiments loaded highly on this component one finds they often deal with visual objects processing such as faces processing.

Another type of data where NMF can extract useful information is text. For each abstract in the database we count the frequency of words and let the NMF work on the set of counts. With hierarchical NMF we can draw a graph of the text clusters we get out – such as the graph on the right. In the subset of text analyzed here 'memory' and 'pain' text clusters are two prominent groupings of the text, leading us to say that memory and pain are important for this particular text corpus.



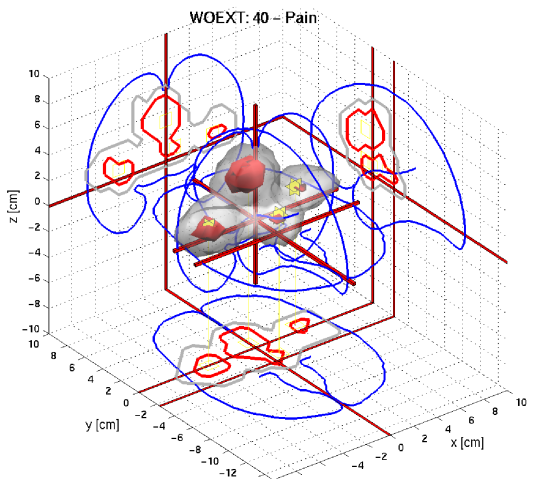
Another type of large-scale data mining is combining text and brain coordinates for exploring functional segregation across all brain regions.

Here we select a particular brain area from the brain region ontology, say 'cingulate gyrus', and extract the naming variations of the region itself and all its subregions. With these names we extract all coordinates labeled with the name and find the associated papers that contain the coordinates. We cluster the papers with hierarchical NMF, so we can draw a graph like displayed on the left. Finally we compare the spatial distribution of brain coordinates in the clustered papers. This comparison is between each text cluster.

On the right side is a 3-dimensional sagittal plot seen from the left sideof the brain where cingulate brain coordinates have been colored according to the text cluster they are assigned to. In this case yellow coordinates are from papers clustered as 'pain' while the magenta is for coordinates clustered as 'memory', and they tend to be grouped in different regions in the cingulate gyrus.

In these cases of data mining we haven't used the fact the experiments have been labeled with the Brede topic ontology. The label of 'memory' arises from the automated text mining.

Data mining with labels



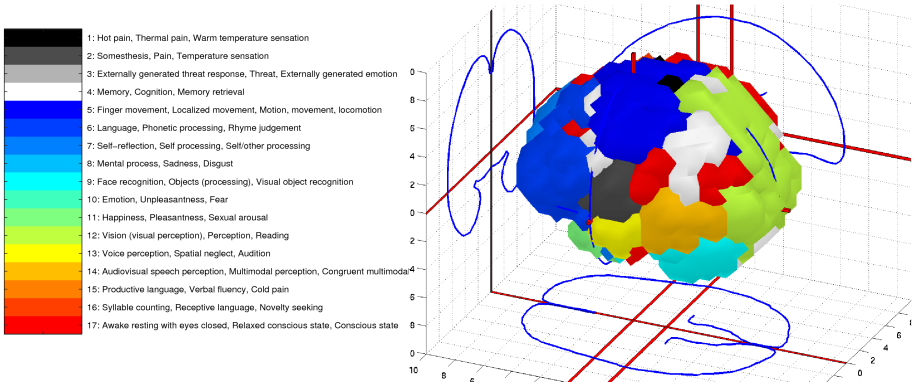
- Collect all experiments labeled with a specific ”topic” (here: pain)
- Extract brain coordinates
- Model the distribution with KDE to get volume
- Perform for all ”topics” in the ontology
- ”ALE”: Bullmore, Laird, Salimi-Khorshidi, Turkeltaub, Wager, Zacks, …

But it is also possible to utilize the experiment labeling, e.g, we can collect all the neuroimaging experiment labeled with the 'pain' item in the ontology or its subtopics and extract their coordinates. Then model the distribution of coordinates with kernel density estimation to get a volume, and apply a threshold based on a resampling method, so we get the 3-dimensional plot on the left. It shows areas of importance for pain, in this case anterior cingulate, insula and thalamus.

There are several other efforts that perform this kind of analysis. It usually goes under the name ALE. In a typical meta-analysis of this kind researchers select just one specific brain function or mental disorder and extract coordinates from papers and model their distribution.

In our case once we have the data entered in the database we can perform the meta-analysis automatically across all topics in the ontology.

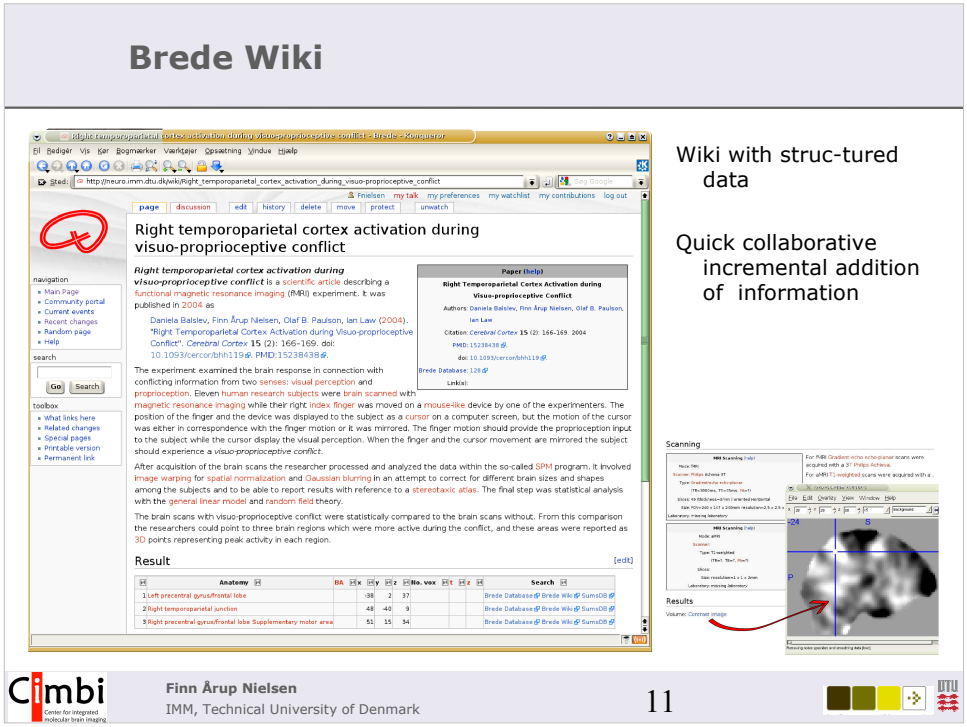
Combining coordinates & experiment labels



Automated functional labeling of voxels:

Conversion of coordinates and ontology labels to matrices and perform NMF

Yet another database-wide data mining method combines the topic ontology and the brain coordinates information, so the clustering from NMF allows us to label each voxel in the brain with a label from the ontology. In this case we get specific area labeled as, e.g., pain, voice perception and emotion.



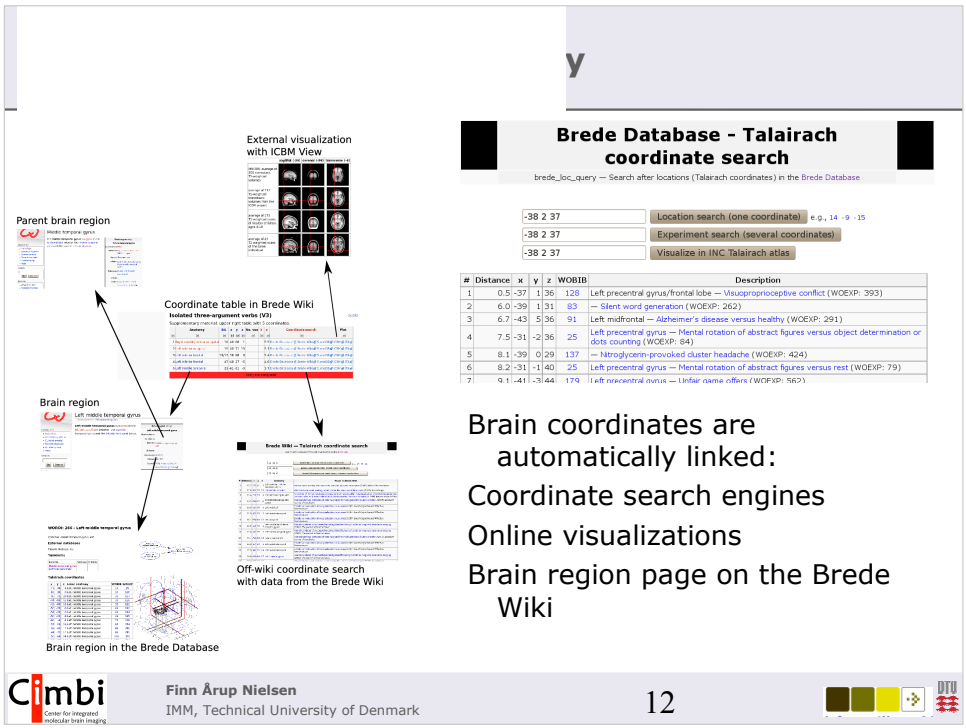
A major bottleneck in large-scale integration of data sets is data entry. Neuroinformatics databases cannot keep up with the generated results published in papers.

We are currently exploring a wiki-based approach for data entry, so we have setup the so-called Brede Wiki. It is based on the MediaWiki engine that Wikipedia also runs. The wiki approach allows quick collaborative and incremental addition of information.

The Brede Wiki is a structured wiki. It uses the template functionality of MediaWiki to structure information, so that when a wiki editor, e.g., writes a number for a brain coordinate the wiki interprets that number as a brain coordinate.

As the Brede Database the Brede Wiki contains information from neuroimaging papers.

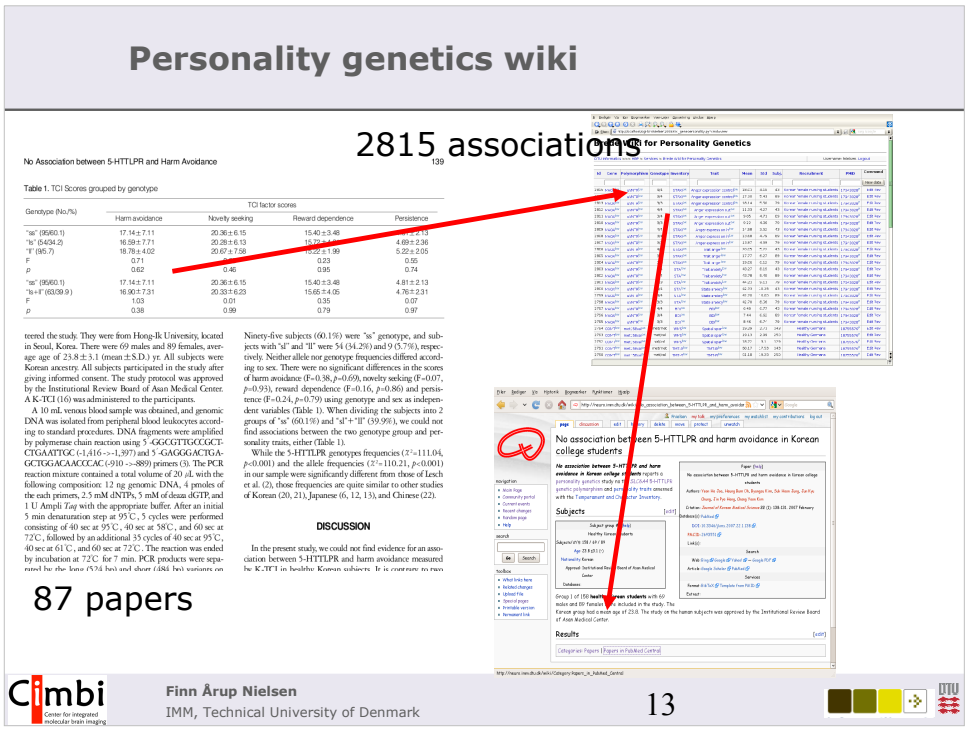
Neuroimaging volumes in the form of NIFTI files can be uploaded to the wiki.



Since the Brede Wiki is structured with templates it is possible to automatically format the wiki page so that brain coordinates are linked to coordinate search engines and online visualization service. Also anatomical labels are automatically linked to pages on the wiki describing brain regions.

We are also able to extract the information from the templates and add each template field value to a database. It means that we can search for nearby brain coordinates in the wiki based on a query coordinate.

We also have the BredeQuery plugin that allows researcher to query the Brede Database for nearby brain coordinates from individual coordinates within the SPM program.



The Brede Wiki cannot make computational and advanced visualizations directly. To explore a system for doing that I have setup a dedicated structured wiki for personality genetics.

The field of personality genetics determines the association between genetic polymorphism and personality scores from personality tests.

The field has an advantage over that of neuroimaging in that papers typically report all results, – not just the results that are statistically significant. It means that we can use standard meta-analytic statistics.

From information usually presented in tables in published papers I type in the result data in a table-like web-based form in the wiki.

I can also export the data for inclusion in the Brede Wiki.

There are presently 2815 personality scores with 39 different polymorphisms from 87 papers. This is probably below a third of all personality genetics studies.

Personality genetics wiki: Meta-analysis

	Effect	Std	P	Studies	Subjects	Gene	Polymorphism	Trait
1	0.854	0.223	0.00013	2	107	ESR1	TA repeat	Harm avoidance
2	-1.102	0.289	0.00014	2	245	HTR3A	C178T	Harm avoidance
3	-0.779	0.220	0.00039	1	90	ESR1	TA repeat	Anxiety
4	-0.445	0.135	0.00098	1	247	TH	TCAT repeat	Extraversion
5	-0.401	0.123	0.00108	1	315	DRD4	Exon 3 VNTR	Positive emotions
6	0.165	0.051	0.00118	13	1747	MAOA	uVNTR	Reward dependence
7	-0.393	0.123	0.00135	1	315	DRD4	Exon 3 VNTR	Extraversion
8	-1.355	0.427	0.00152	1	125	HTR3A	C178T	Nonconformity
9	-0.758	0.240	0.00161	1	122	SLC6A4	5-HTTLPR	Activity
10	-0.174	0.055	0.00163	16	1791	SLC6A4	5-HTTLPR	Agreeableness

Top 10 polymorphims/traits correlations.

With data added to the database the wiki can perform a meta-analysis across genetic polymorphisms and personality traits.

First an effect size is computed for each association and then a meta-analytic effect size is found for all polymorphisms and all personality traits.

One of the pages on the wiki site displays the result of such a large-scale meta-analysis and when it is sorted according to p-value the table here is generated.

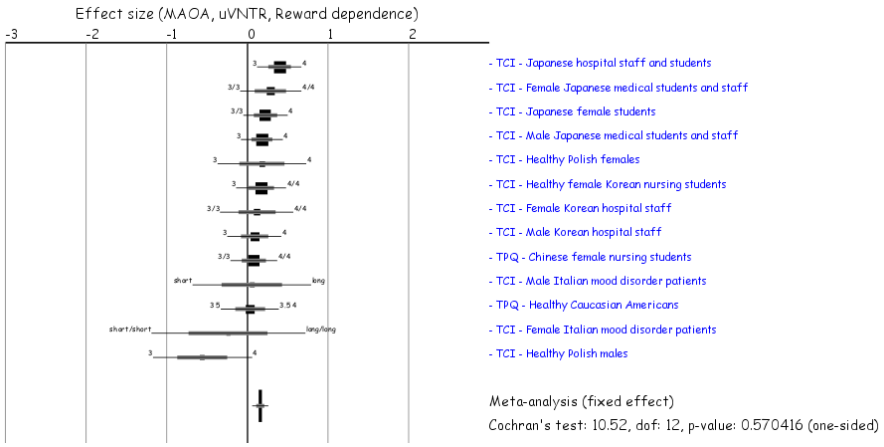
Note that these p-values are not corrected for multiple comparisons.

The most significant association is for an estrogen receptor and 'harm avoidance'. However, this results is based on only two groups of subjects reported in the same paper and with few subjects.

If we look further down the list for associations supported by multiple papers we find monoamine oxidase A variable number of tandem repeats polymorphism and 'reward dependence'.

Also on the top ten is the well-known serotonin transporter polymorphism and not neuroticism, - but rather agreeableness.

Personality genetics wiki: Forest plot



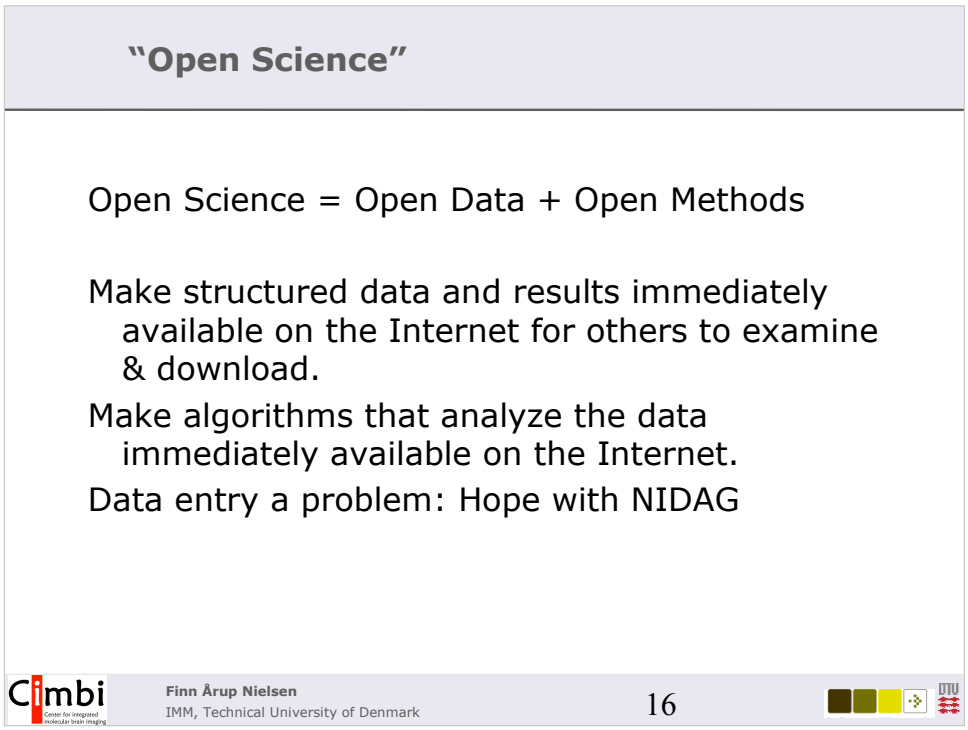
Forest plot for effect sizes for MAOA/reward dependence

There are many results in the wiki that can be examined, but if we look at the individual studies that make up the association with one of the highest correlation it is possible to draw a so-called forest plot with the wiki.

In the case with MAOA and the personality trait 'reward dependence' as measured with the Cloninger personality inventories, TCI and TPQ, there are presently 13 studies in the database that make up the meta-analytic result. Many of these studies are Asian, which all tend to contribute positively to the results.

It may be worth to note that only a single of these studies is significant in itself.

Before attributing to much significance to this finding we should remember that this particular result is selected among multiple comparisons. However, it is the most promising association when large-scale data mining is done in this personality genetics wiki.



I view the efforts here as part of the notion of 'Open Science' where the data and methods are openly available for others to examine and further develop.

Where data and results are immediately available online, distributed in a structured and standardized format for easy inclusion in other databases. Parts of Brede Database has been included in the AMAT and SumsDB coordinate databases and also federated into the NIF web service.

My main bottleneck remains data entry. Neuroinformatics databases with result data are far from complete. I think more collaborative and automated approaches are needed.

In a group called NIDAG we are working towards a common format for brain coordinates reporting and automated extractions of coordinates from journal papers. Hopefully this effort will result in neuroinformatics databases with more coverage.

Thanks!

Thanks to the Lundbeck Foundation for funding.

http://www.imm.dtu.dk/~fn/ – http://neuro.imm.dtu.dk/wiki/ – http://neuro.imm.dtu.dk/services/brededatabase/

Thanks for you attention.