Neuroinformatic techniques for provenance & data sharing

Camille Maumet

GlaxoSmithKline - Neurophysics Workshop on Skeptical Neuroimaging

January 14th, 2014

THE UNIVERSITY OF WARWICK

Outline

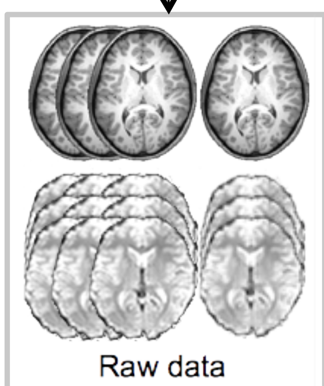
1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing

Outline

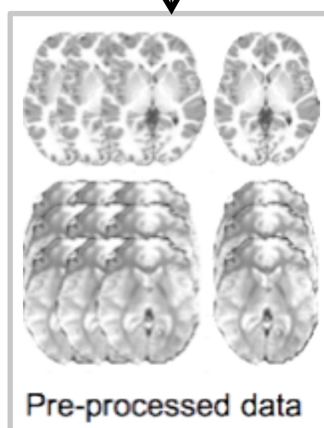
1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing

Overview of a neuroimaging study

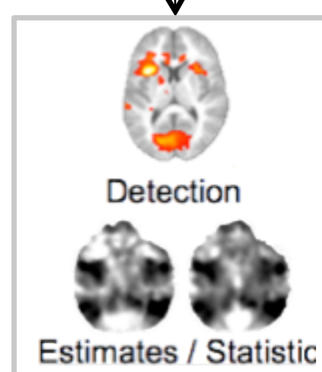
- MR scanner: 1.5T, 3T…
- Type of coils: 32-channels, 16-channels…
- Imaging sequence: TR, TE, FOV…
- fMRI paradigm



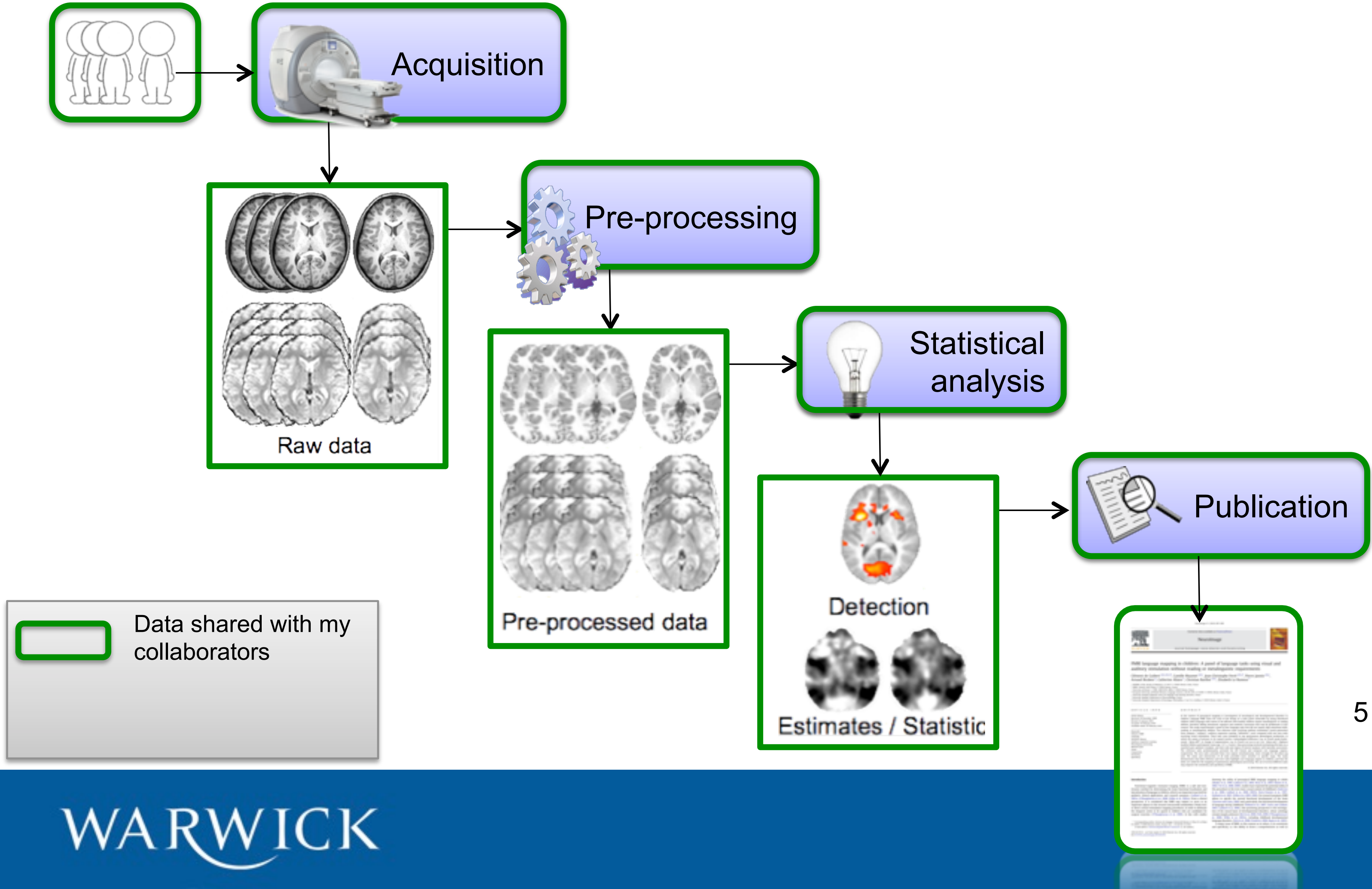
- Pre-processing pipeline
- Method employed for each processing
- Parameters for each method, software…



- Model
- (Non-)parametric
- Parameters

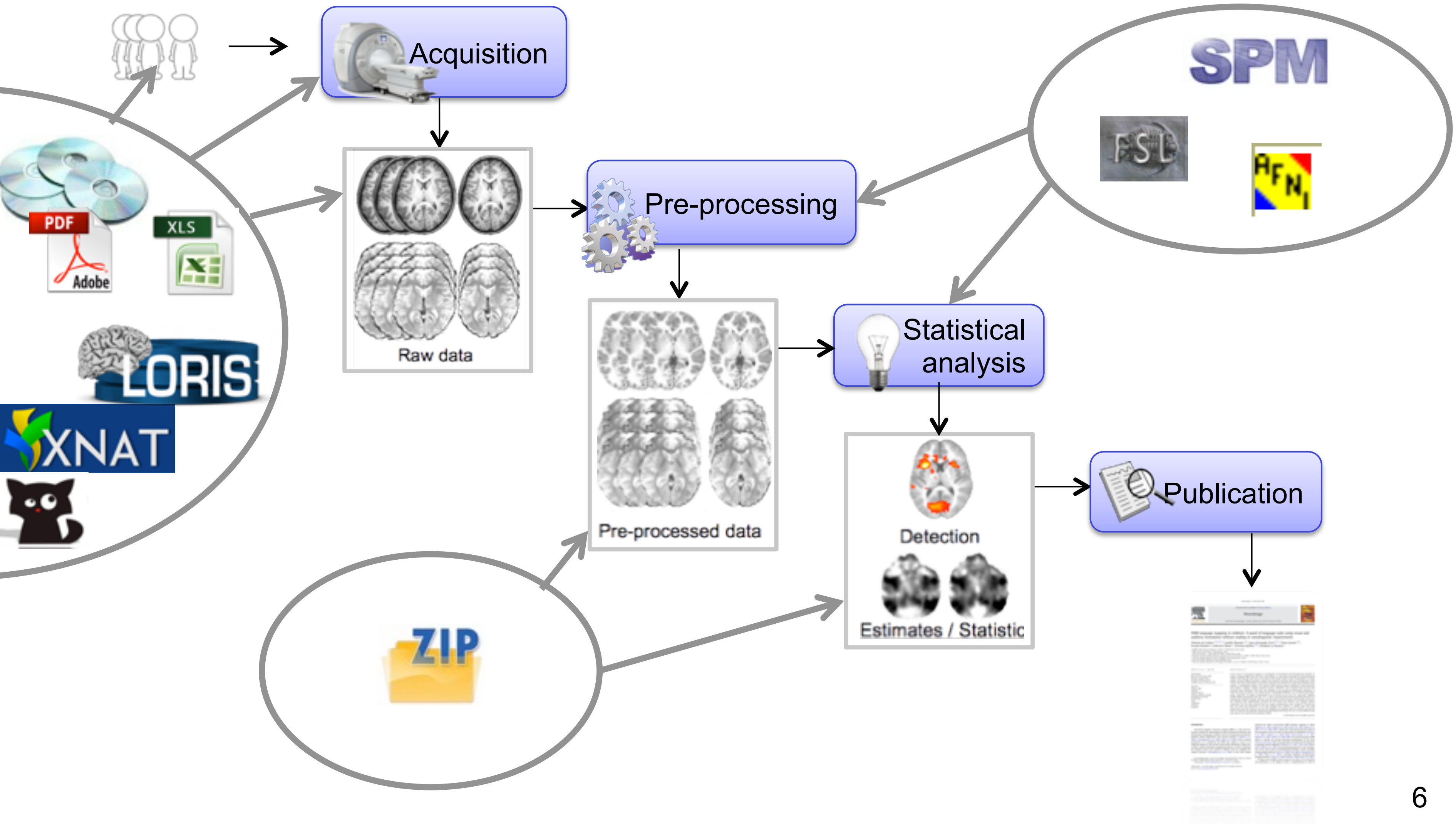


Neuroimaging and data sharing

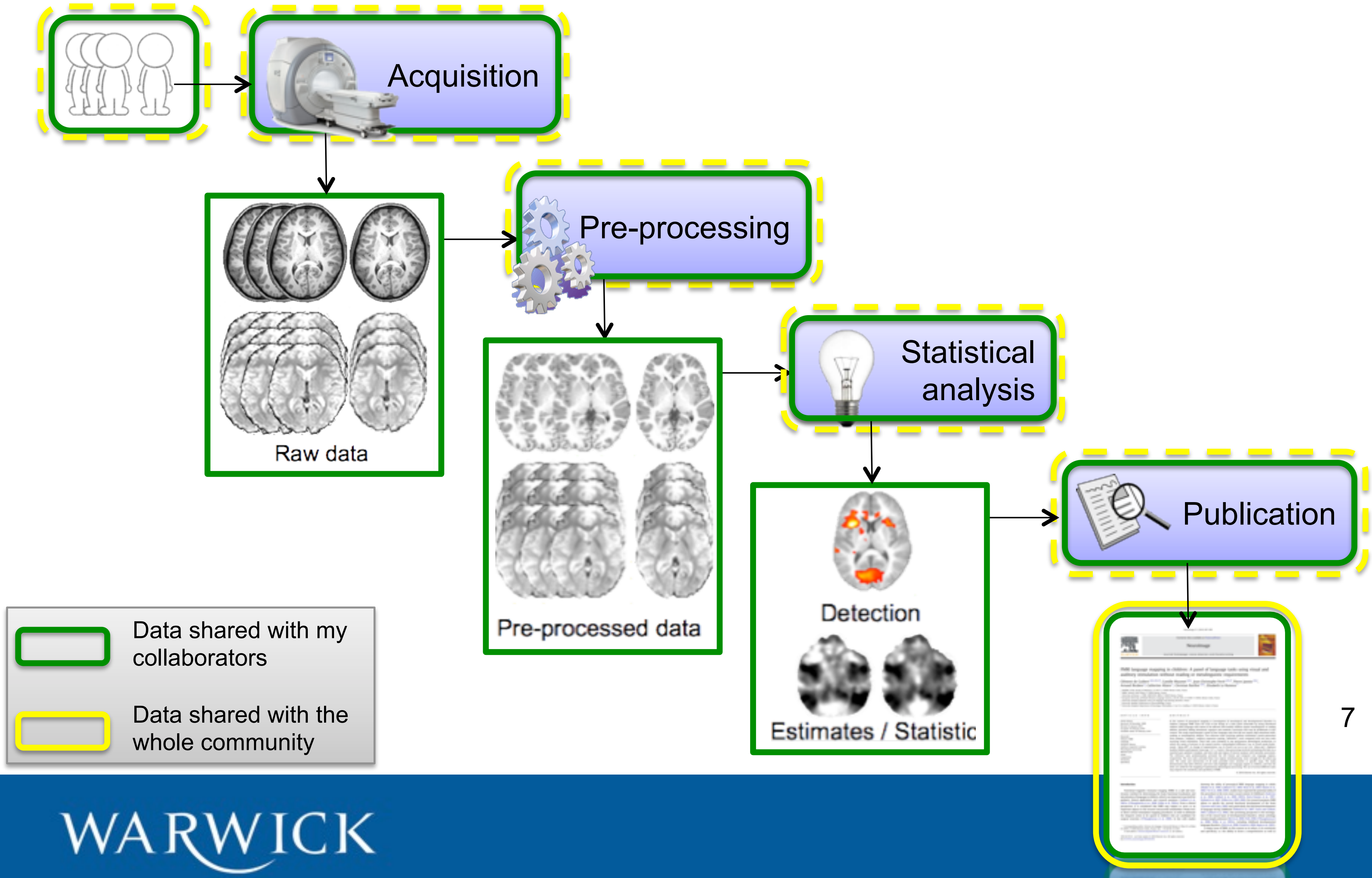


WARWICK

Sharing data with my collaborators



Neuroimaging and data sharing



A neuroimaging publication

- *Methods* section: metadata in free-form text.

General technical implementation

A single scanner session included the four paradigms separately implemented with the same parameters: a simple block design alternated a rest condition as control and the language task, starting with rest, with a preliminary period of signal acquisition for MRI signal stabilization which was later discarded during data processing. Each paradigm included three 27-s blocks of each condition and had a total duration of 2 min 48 s. The scanner session, including the anatomical acquisition and the four language paradigms, had a duration of about 30-35 min. All subjects performed the tasks in the same order, as during the preparation step, in order to avoid the mix of auditory and visual tasks and the resulting complication for the child. Words required by the tasks were one-to-three-syllable words highly frequent in the lexicon of French 8 years old children (Lambert and Chesnet, 2001).

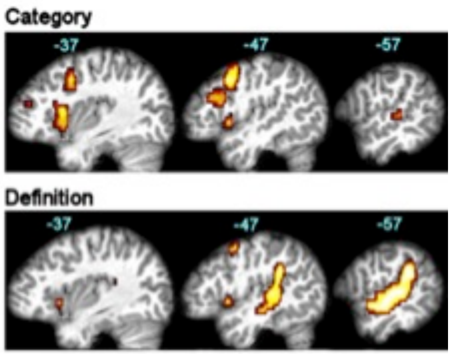
During the rest condition, a red cross was displayed on the projection screen and children were asked "not to work", to "think

Table 2
Task comparisons (>) and conjunctions (C).

	Auditory language
	Categ > Def
Left Hemisphere	
Inf frontal-Oper	---
Precentral	18–3.38[5]
Mid frontal	33–3.66
SMA	---
Cingulate	---
Med sup frontal	174–4.69
Rol operculum	---
Insula	---
Sup temporal	---
Mid temporal	---
Inf parietal	---
Sup parietal	---
Postcentral	---
Sup occipital	---

Table

- *Results* section:



2D plot(s) of the detections

General technical implementation

A single scanner session included the four paradigms separately implemented with the same parameters: a simple block design alternated a rest condition as control and the language task, starting with rest, with a preliminary period of signal acquisition for MRI signal stabilization which was later discarded during data processing. Each paradigm included three 27-s blocks of each condition and had a total duration of 2 min 48 s. The scanner session, including the anatomical acquisition and the four language paradigms, had a duration of about 30-35 min. All subjects performed the tasks in the same order, as during the preparation step, in order to avoid the mix of auditory and visual tasks and the resulting complication for the child. Words required by the tasks were one-to-three-syllable words highly frequent in the lexicon of French 8 years old children (Lambert and Chesnet, 2001).

During the rest condition, a red cross was displayed on the projection screen and children were asked "not to work", to "think about nothing" and, because of the complexity of this instruction, to listen to the noise of the scanner and fix attention on the red cross

Description of the detections

Table 2
Task comparisons (>) and conjunctions (C). Peak locations, cluster extent–Z-score (p<0.001 unc.; k = 10).

	Auditory language tasks			Visual language tasks	
	Categ > Def	Def > Categ	Categ C Def	Ph-s > Ph-d	Ph-
Left Hemisphere					
Inf frontal-Oper	---	---	348–4.10[4]	---	82
Precentral	18–3.38[5]	---	348–5.09	---	82
Mid frontal	33–3.66	---	---	---	---
SMA	---	---	1433–5.48	---	35
Cingulate	---	---	1433–5.08[3]	---	---
Med sup frontal	174–4.69	---	---	---	---
Rol operculum	---	---	---	36 – 4.31	---
Insula	---	---	396–4.87[8]	---	58-
Sup temporal	---	---	351–3.81[1]	---	91-
Mid temporal	---	1658–4.67[3]	351–5.61[2]	---	10-
Inf parietal	---	1658–5.18[4]	---	---	---
Sup parietal	---	---	---	---	97
Postcentral	---	---	---	---	97
Sup occipital	---	---	---	---	---
Mid occipital	---	---	---	146–4.43	14
Inf occipital	---	---	---	---	14
Fusiform	---	---	---	397–5.44	14

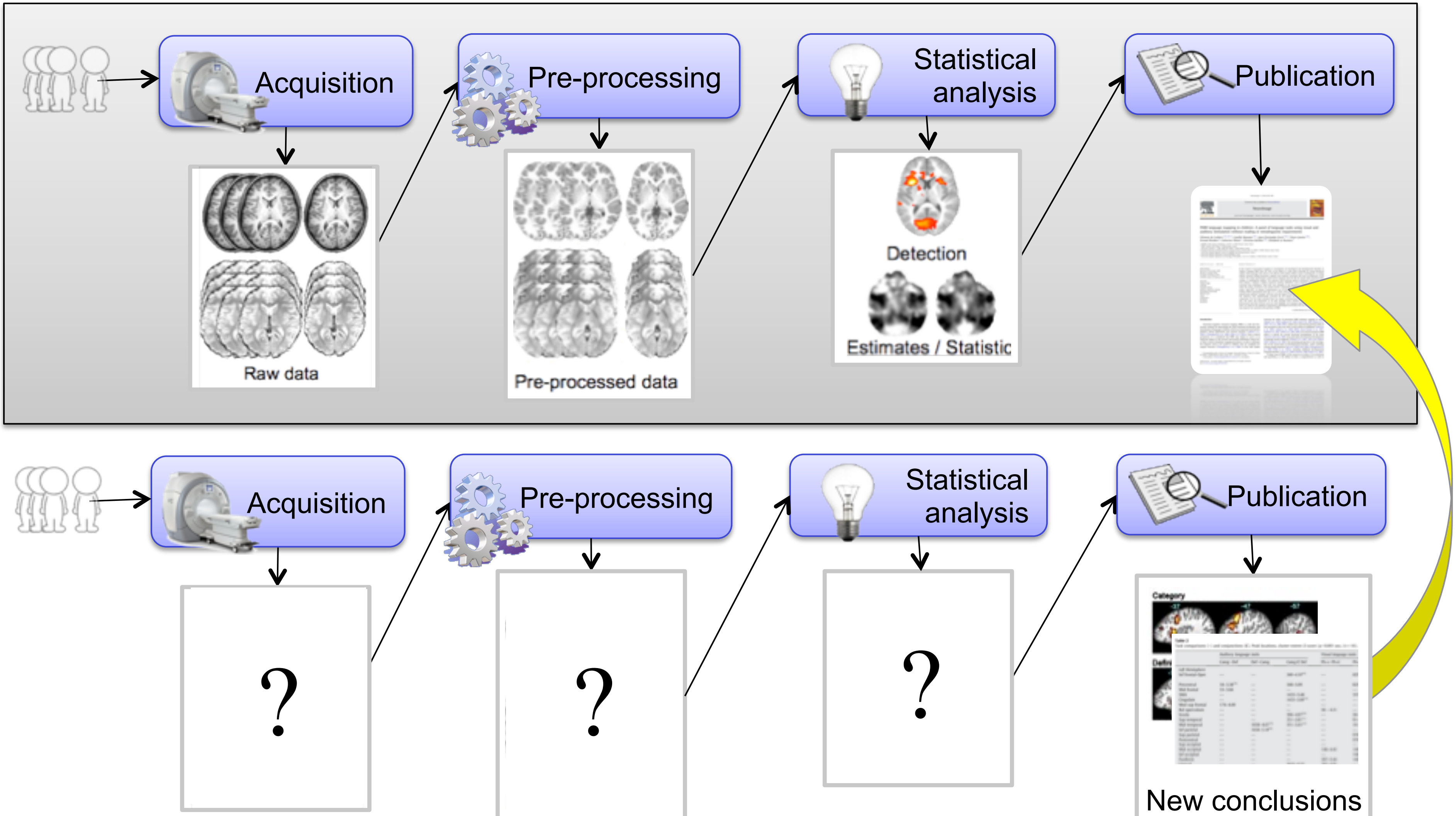
Table of local maxima

WARWICK

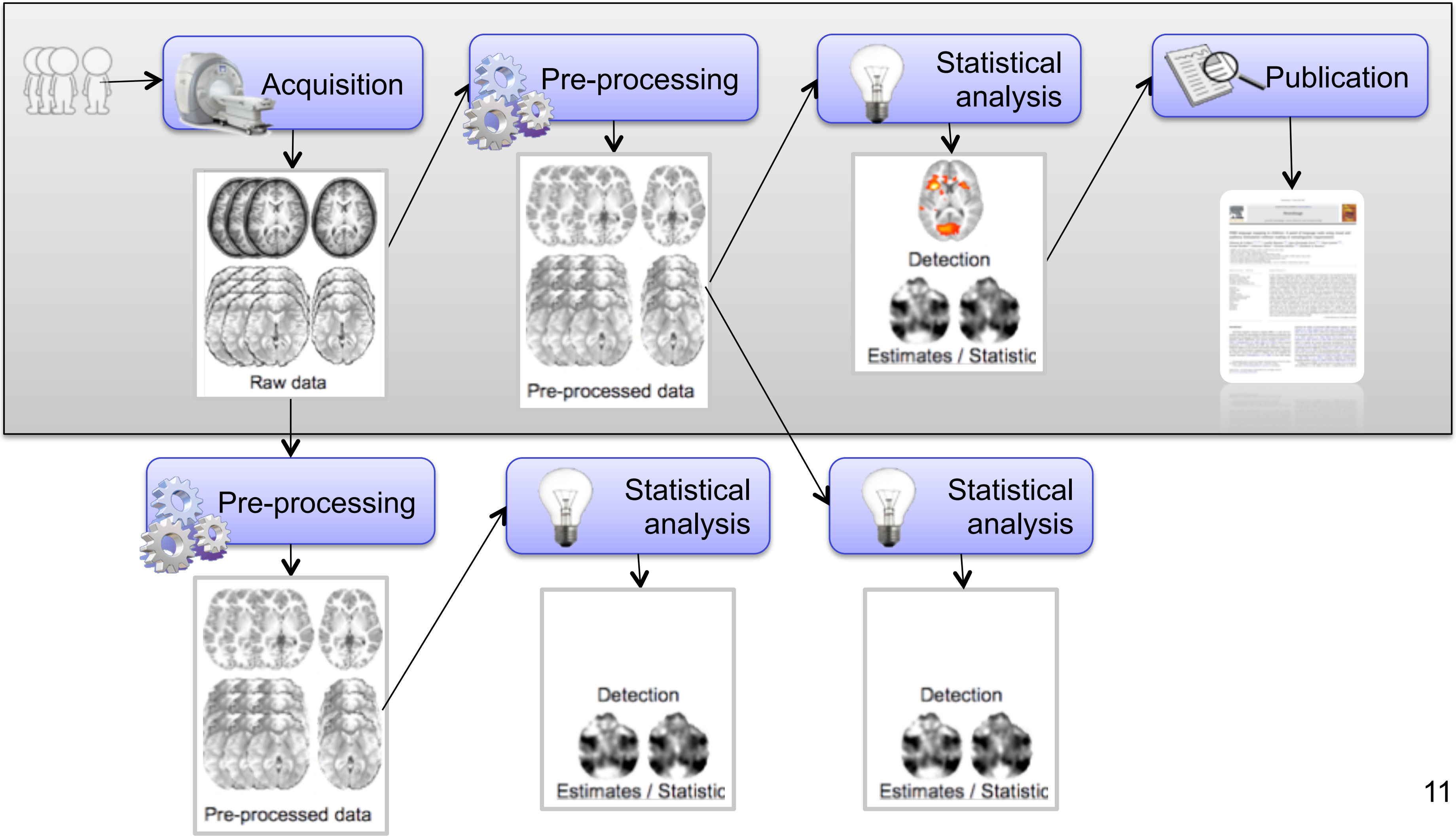
Outline

1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing

Reproducibility

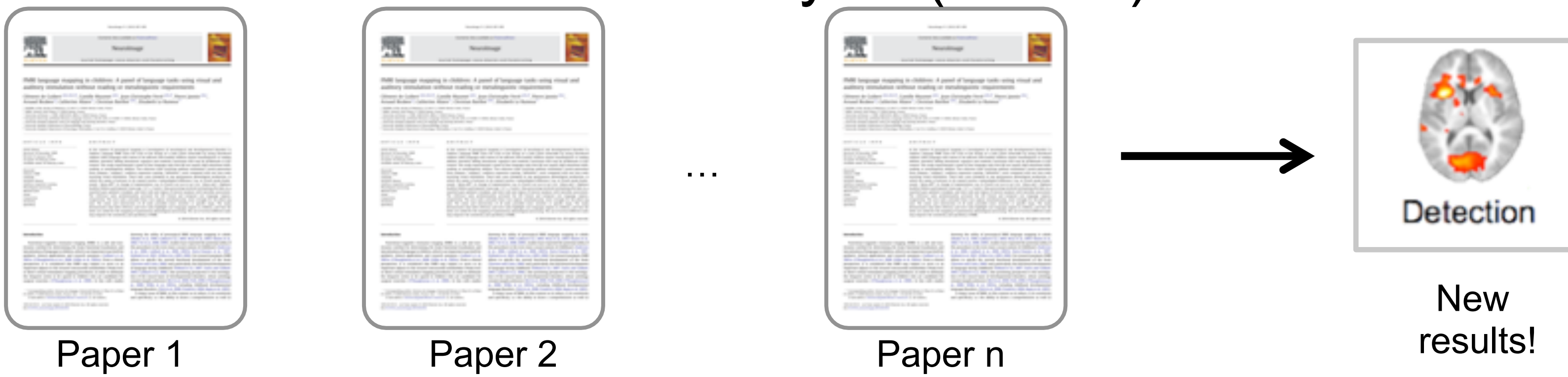


Full provenance

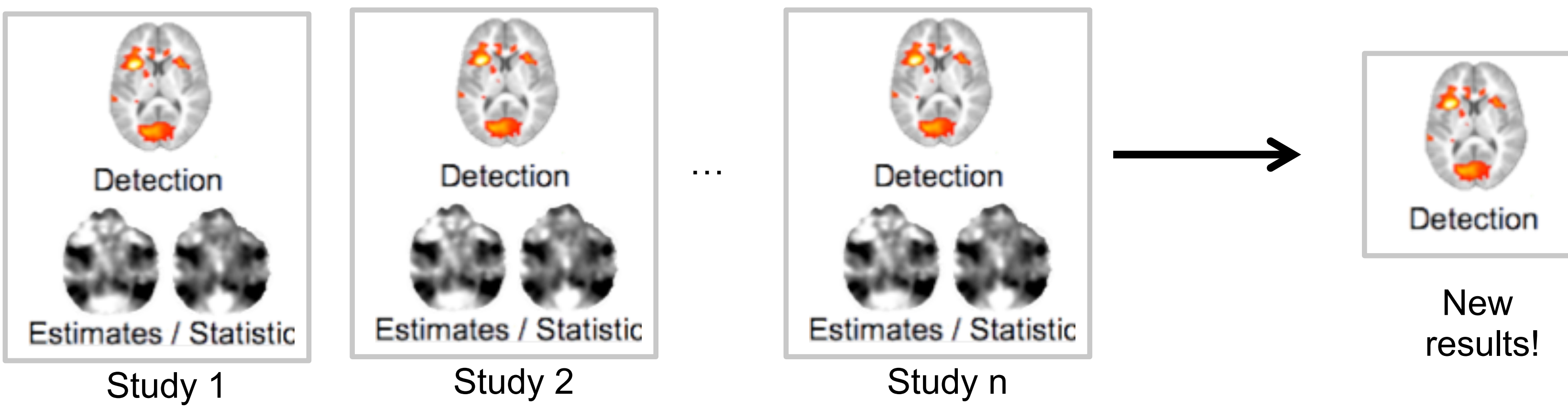


Meta-analysis: analyzing the analyses

- Coordinate-Based Meta-Analysis (CBMA)



- Image-Based Meta-Analysis (IBMA).



How to become less skeptical?

- Reproducibility
  - Confirm results by re-running an analysis
- Provenance
  - Needed for reproducibility
  - Avoid selection bias.
- Meta-analysis
  - Strengthen results by combining studies.

- What do we need?
  - Sharing data, meta-data and provenance.

Data sharing: obstacles

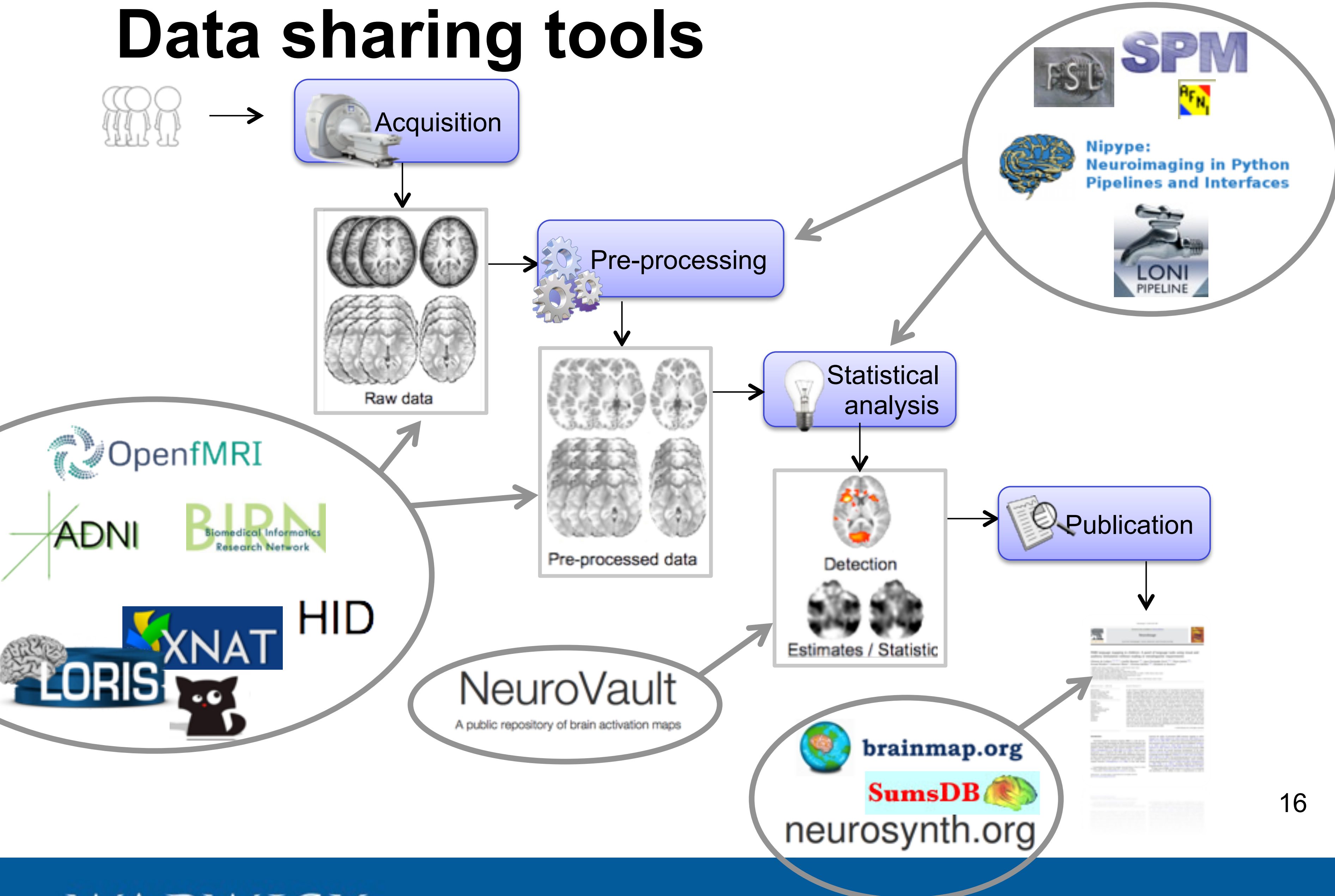
- Psychological
  - "My" data
- Ethical constraints
- Technical: difficulties to share data with enough metadata to be really useful
  - *Available* data versus *usable* data.

"Less than a few percents of acquired neuroimaging data is available in public repositories" [Poline 2012]

Outline

1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing

Data sharing tools

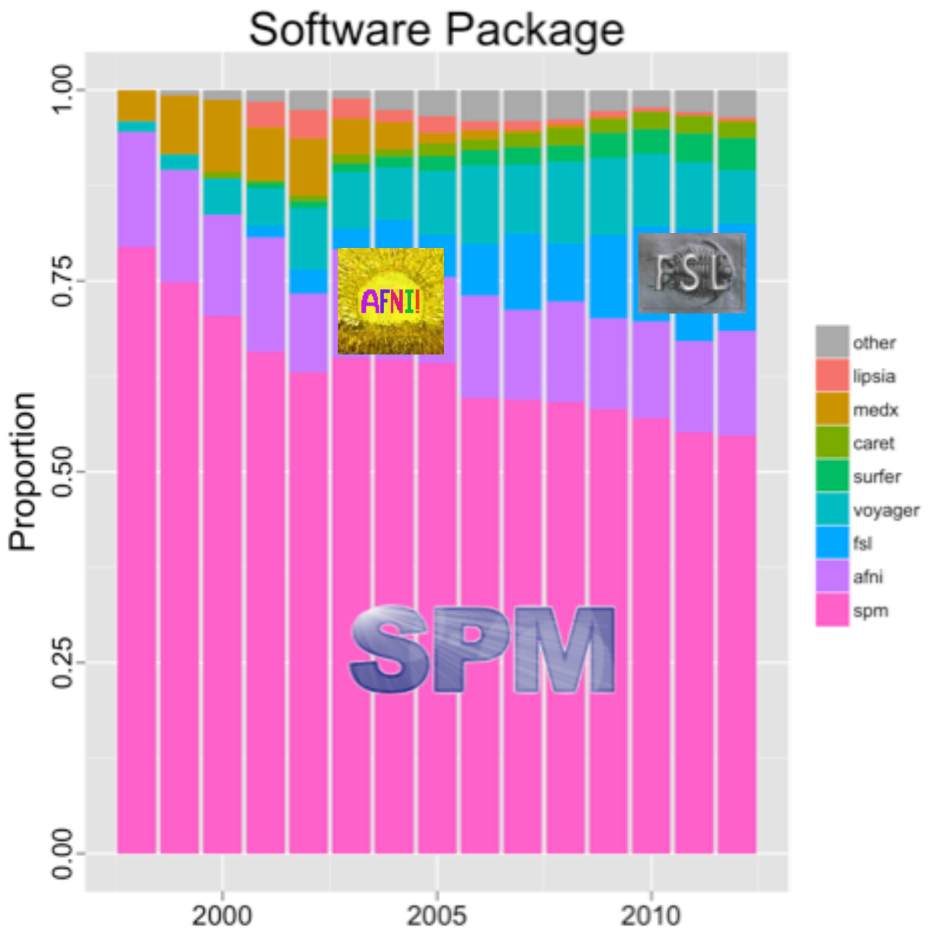


A standard format for meta-data

- Sharing data across the data sharing tools…
- First attempt of an agnostic format: **X**ML-Based **C**linical **E**xperiment **D**ata **E**xchange Schema (XCEDE): www.xcede.org
  - Describes subject, study, activation
  - Limited provenance encoding
  - Initiative of the BIRN
- **N**euro**I**maging **D**ata **M**odel NI-DM: www.nidm.nidash.org
  - Based on web-semantic tools.
  - Initiative of the BIRN and INCF

Three major players

- Bottom-up approach.
- Lean on **existing analysis software (SPM, FSL, AFNI)** to disseminate the standard.



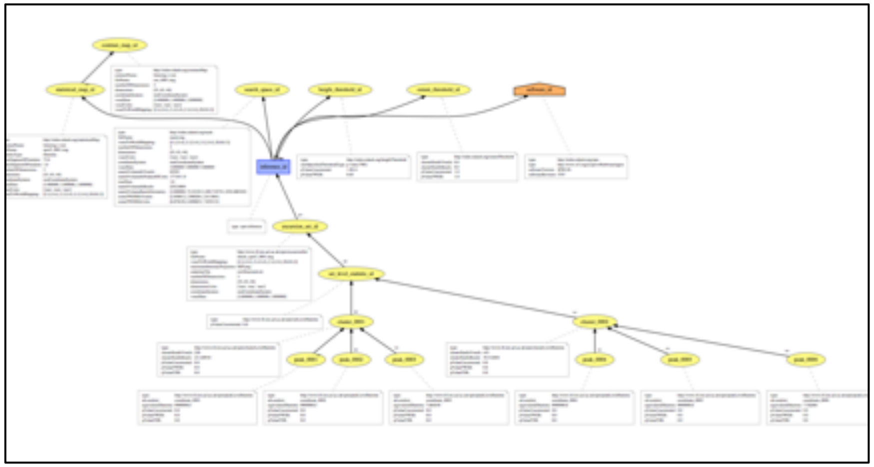
Automatically created with Neurotrends based on over 16 000 journal articles

Work in progress

- Define a format to represent the results of a neuroimaging study with a focus on meta-analysis.

Term name	Definition	Example	BIRNLex or NIDM Concept ID
BonferroniCorrection	Bonferroni correction for multiple statistical tests		nidm:nidm_80
chi-squareStatistic	A statistical parameter drawn from a chi-square statistic		nidm:nidm_81
FDR	False Discovery Rate correction		nidm:nidm_82
FWER	Family-wise Error Rate correction		nidm:nidm_83
Scan Image	An image that is the ouput of an MRI or CT or PET scan		nidm:nidm_84
SliceOrder	The temporal order in which the 2D slices were acquired by the imaging systems		nidm:nidm_85
Voxel	volumetric pixel		nidm:nidm_86
Z-Statistic	A statistical parameter drawn from a normal or z distribution		nidm:nidm_87
extentThresh	Minimum cluster size used when thresholding a statistic image	5voxels	nidm:nidm_88
errorDegreesOfFreedom	Degrees of freedom of the error.	73	nidm:nidm_89
effectDegreesOfFreedom	Degrees of freedom of the effect.	1	nidm:nidm_90
StatisticMap	A map (2D or 3D structured dataset) whose value at each location is a statistic.		nidm:nidm_91
voxelSize	3D size of a voxel measured in voxelUnits.	[2 2 4]	nidm:nidm_92
cluster	A group of neighboring image elements (voxels or vertices)		nidm:nidm_93
qValueFDR	p-value adjusted for the search volume, controlling for the False Discovery Rate	0.000154	nidm:nidm_94
pValueFWE	p-value adjusted for the search volume, controlling for the Familywise Error Rate	0.00554	nidm:nidm_95
pValueUncorrected	Uncorrected p-value	0.0542	nidm:nidm_96

Vocabulary



Data model

WARWICK

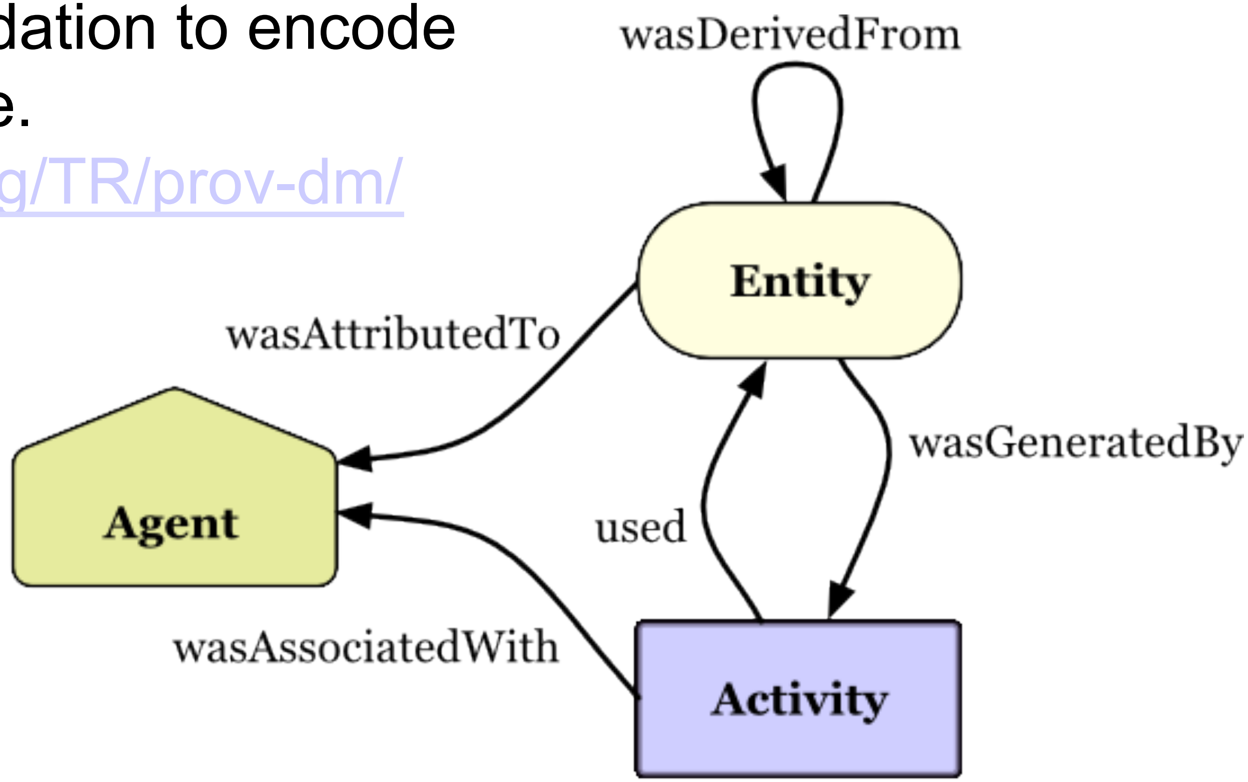
Neuroimaging terms

- Define a vocabulary to support the format.

Term name	Definition	Example	BIRNLex or NIDM Concept ID	synonyms and related urls	Parent term
cluster	A group of neighboring image elements (voxels or vertices)		nidm:nidm_93	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43 or http://purl.obolibrary.org/obo/OBI_0000251	
qValueFDR	p-value adjusted for the search volume, controlling for the False Discovery Rate	0.000154	nidm:nidm_94	http://purl.obolibrary.org/obo/OBI_0001442	p-value i.e. nidm:nidm_0011
[illegible]	p-value adjusted for the search volume, controlling [illegible]	[illegible]	[illegible]	[illegible]	[illegible]
pValueUncorrected	Uncorrected p-value	0.0542	nidm:nidm_96	http://purl.obolib	p-value i.e. nidm:nidm_0011
SelectionProcedure	Procedure to select the values that are being reported		nidm:nidm_97		
clusterSizeInVoxels	Number of voxels contained in a cluster.	18	nidm:nidm_98		
softwareVersion	Name and Number specifying software version.	SPM99, SPM2, SPM5, SPM8, SPM12b, FSL5.0.0	nidm:nidm_99		nidm:Software
softwareRevision	Software revision number.	v5417	nidm:nidm_100		
clusterSizeInVertices	Number of vertices contained in a cluster.	10	nidm:nidm_101		
clusterSizeInResels	Number of resels contained in a cluster.	13	nidm:nidm_102		
voxelUnits	Units associated to each dimensions of some N-dimentional data.	{'mm' 'mm' 's'}	nidm:nidm_103		
ReselSizeInWorldUnits	Volume of a resel, a resolution element, expressed in units. It expresses the smoothness of the noise, with smoother images having larger resels.		nidm:nidm_104	http://en.wikipedia.org/wiki/Resel	
			nidm:nidm_105		
Map	2D or 3D structured dataset.		nidm:nidm_106		
fileName	Name associated with a file (without path).		nidm:nidm_107		
numberOfDimensions	Number of Dimensions of some N-dimentional data.	3	nidm:nidm_108		
dimensions	Dimensions of some N-dimentional data.	[64 64 20]	nidm:nidm_109		
coordinateSystem	Type of coordinate system.	nidm:mniCoordinateSystem	nidm:nidm_110		
searchVolumeInVoxels	Total number of voxels within the search volume.	68656	nidm:nidm_111	Synonyms of nidm:volumeInVoxels	
searchVolumeInResels	Total number of resels within the search volume.	151.3	nidm:nidm_112	Synonyms of nidm:volumeInResels	

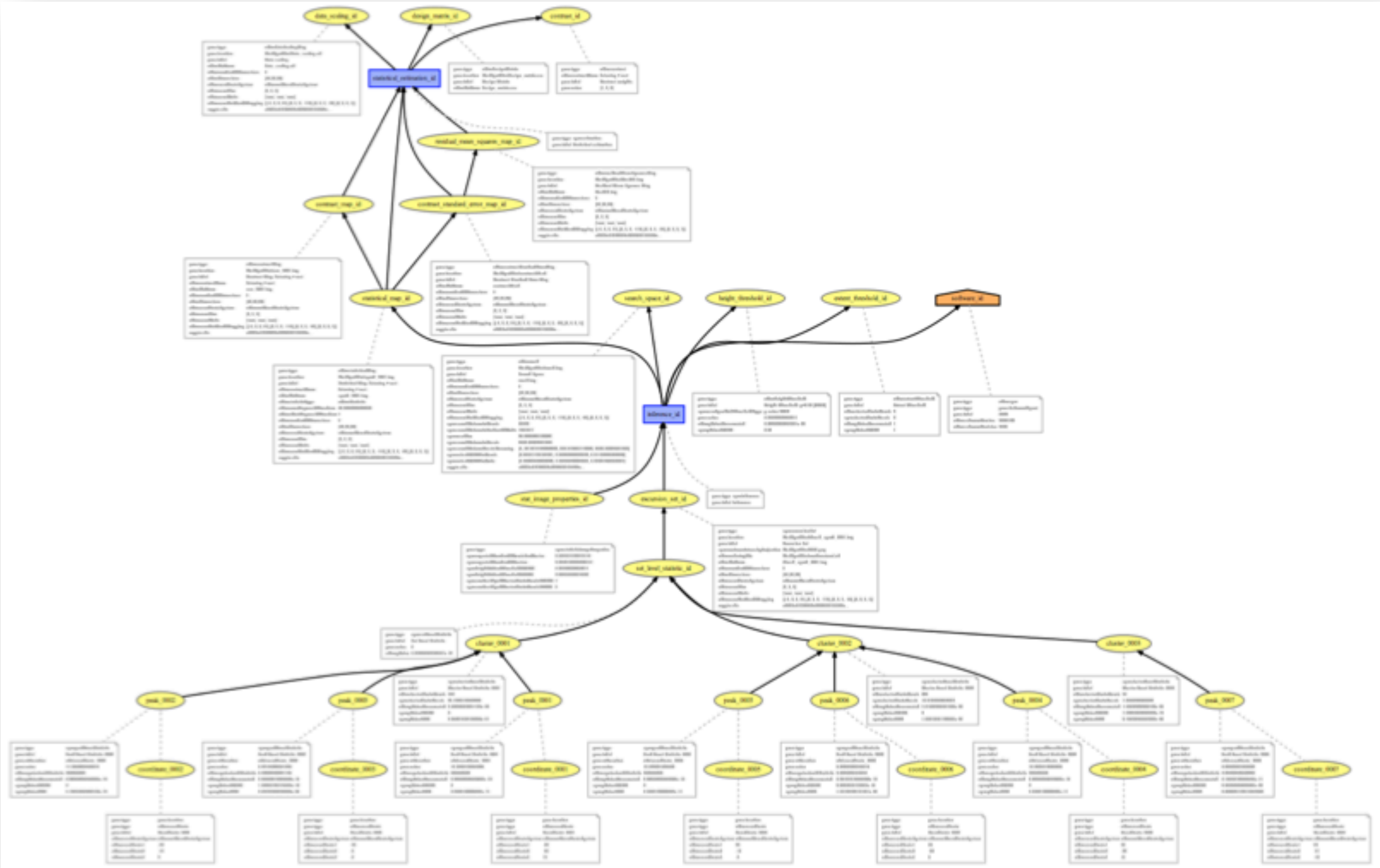
Data model

- Based on PROV-DM a W3C recommendation to encode provenance. www.w3.org/TR/prov-dm/

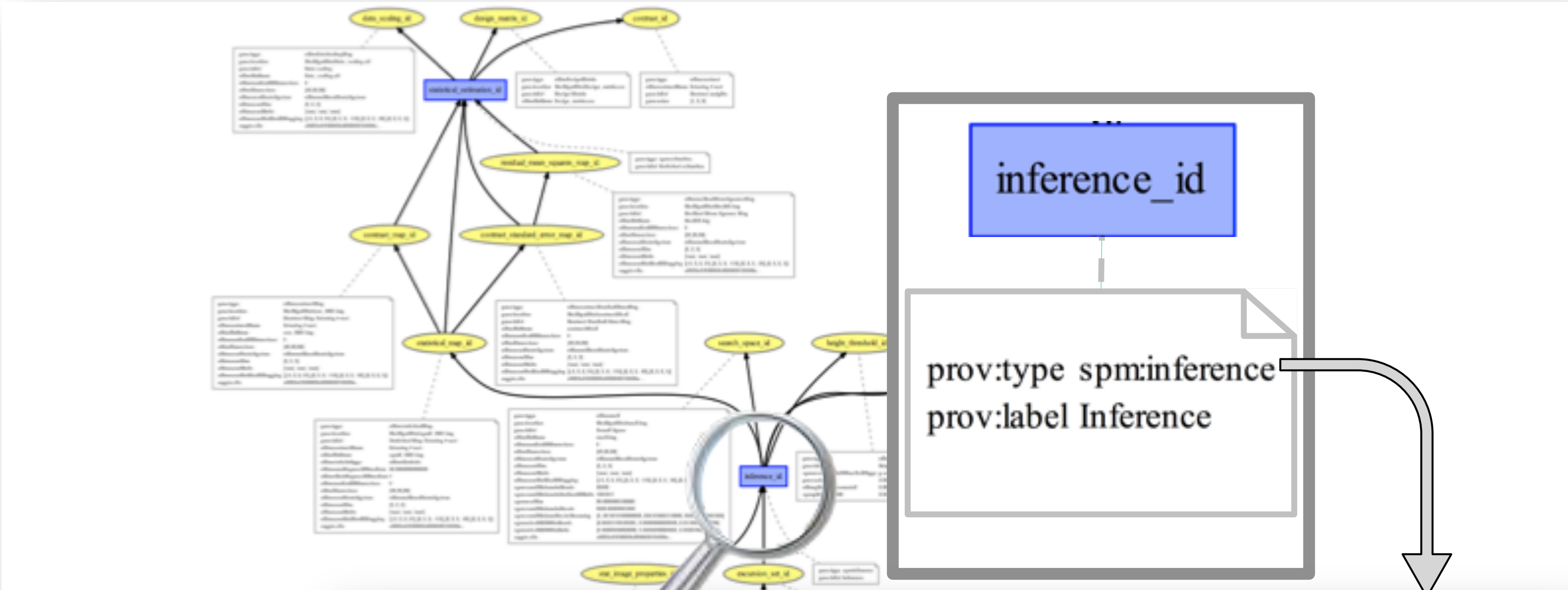


WARWICK

Data model



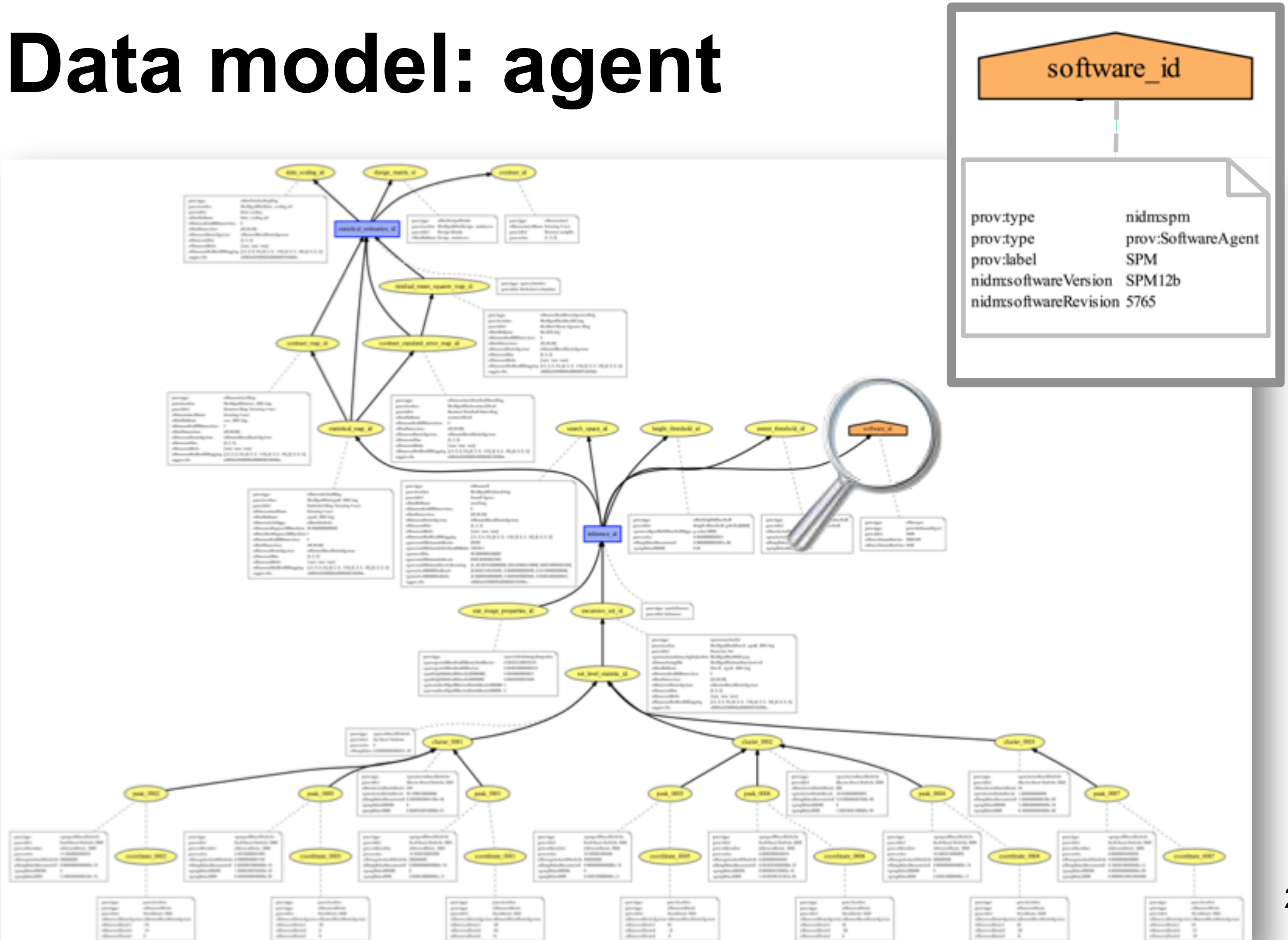
Data model: activities



Term name	Definition	Example	BIRNLex or NIDM Concept ID	synonyms and related urls	Parent term
cluster	A group of neighboring image elements (voxels or vertices)		nidm:nidm_93	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43 or http://purl.obolibrary.org/obo/OBI_0000251	
qValueFDR	p-value adjusted for the search volume, controlling for the False Discovery Rate	0.000154	nidm:nidm_94	http://purl.obolibrary.org/obo/OBI_0001442	p-value i.e. nidm:nidm_0011
pValueFWE	p-value adjusted for the search volume, controlling for the Familywise Error Rate	0.00554	nidm:nidm_95	http://purl.obolibrary.org/obo/OBI_0001265	p-value i.e. nidm:nidm_0011
pValueUncorrected	Uncorrected p-value	0.0542	nidm:nidm_96	http://purl.obolibrary.org/obo/OBI_0000175	p-value i.e. nidm:nidm_0011
SelectionProcedure	Procedure to select the values that are being reported		nidm:nidm_97		
clusterSizeInVoxels	Number of voxels contained in a cluster.	18	nidm:nidm_98		
softwareVersion	Name and Number specifying software version.	SPM99, SPM2, SPM5, SPM8, SPM12b, FSL5.0.0	nidm:nidm_99		nidm:Software
softwareRevision	Software revision number.	v5417	nidm:nidm_100		
clusterSizeInVertices	Number of vertices contained in a cluster.	10	nidm:nidm_101		
clusterSizeInResels	Number of resels contained in a cluster.	13	nidm:nidm_102		
voxelUnits	Units associated to each dimensions of some N-dimentional data.	{'mm' 'mm' 's'}	nidm:nidm_103		
ReselSizeInWorldUnits	Volume of a resel, a resolution element, expressed in units. It expresses the smoothness of the noise, with smoother images having larger resels.		nidm:nidm_104	http://en.wikipedia.org/wiki/Resel	
			nidm:nidm_105		
Map	2D or 3D structured dataset.		nidm:nidm_106		
fileName	Name associated with a file (without path).		nidm:nidm_107		
numberOfDimensions	Number of Dimensions of some N-dimentional data.	3	nidm:nidm_108		
dimensions	Dimensions of some N-dimentional data.	[64 64 20]	nidm:nidm_109		
coordinateSystem	Type of coordinate system.	nidm:mniCoordinateSystem	nidm:nidm_110		
searchVolumeInVoxels	Total number of voxels within the search volume.	68656	nidm:nidm_111	Synonyms of nidm:volumeInVoxels	
searchVolumeInResels	Total number of resels within the search volume.	151.3	nidm:nidm_112	Synonyms of nidm:volumeInResels	

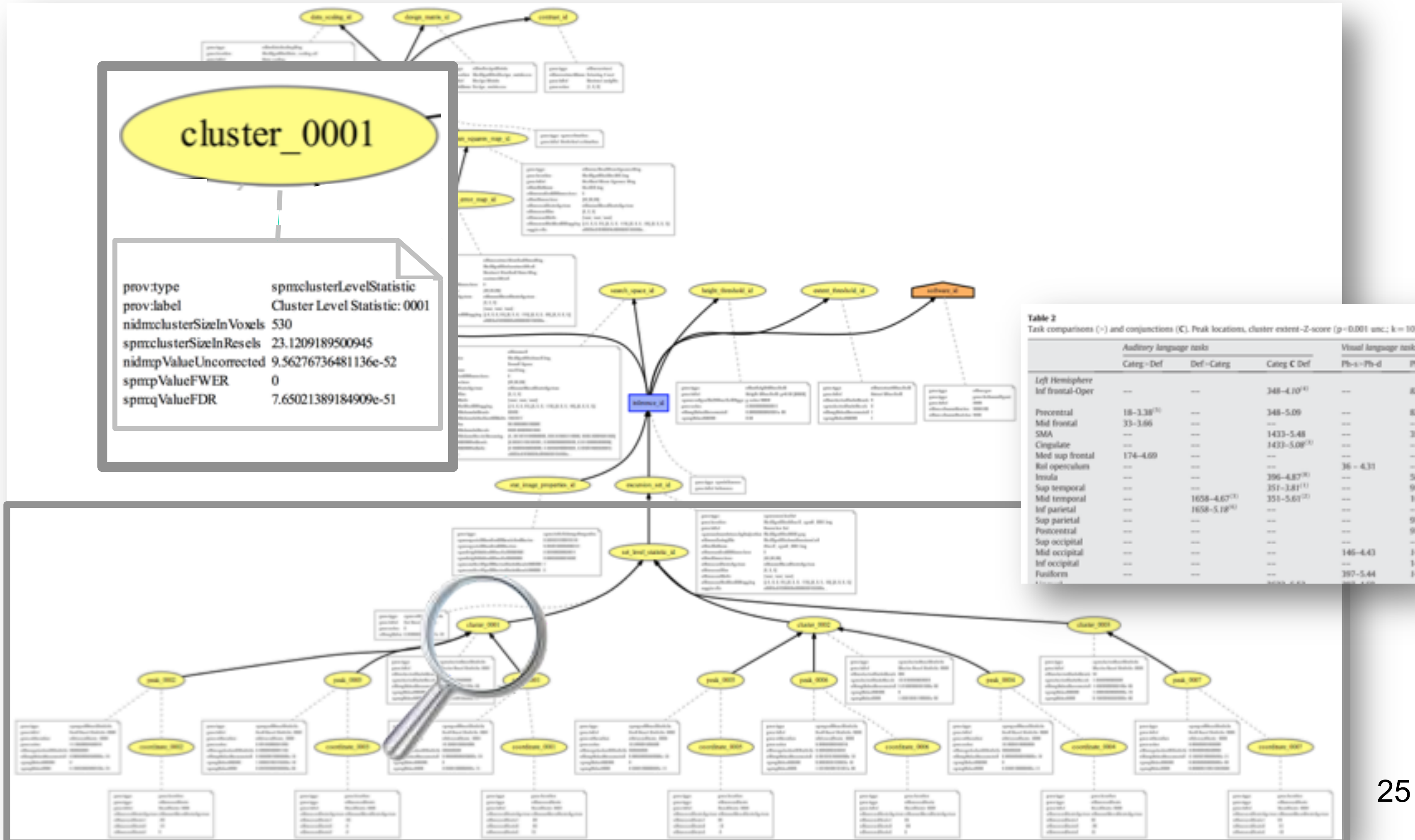
WARWICK

Data model: agent

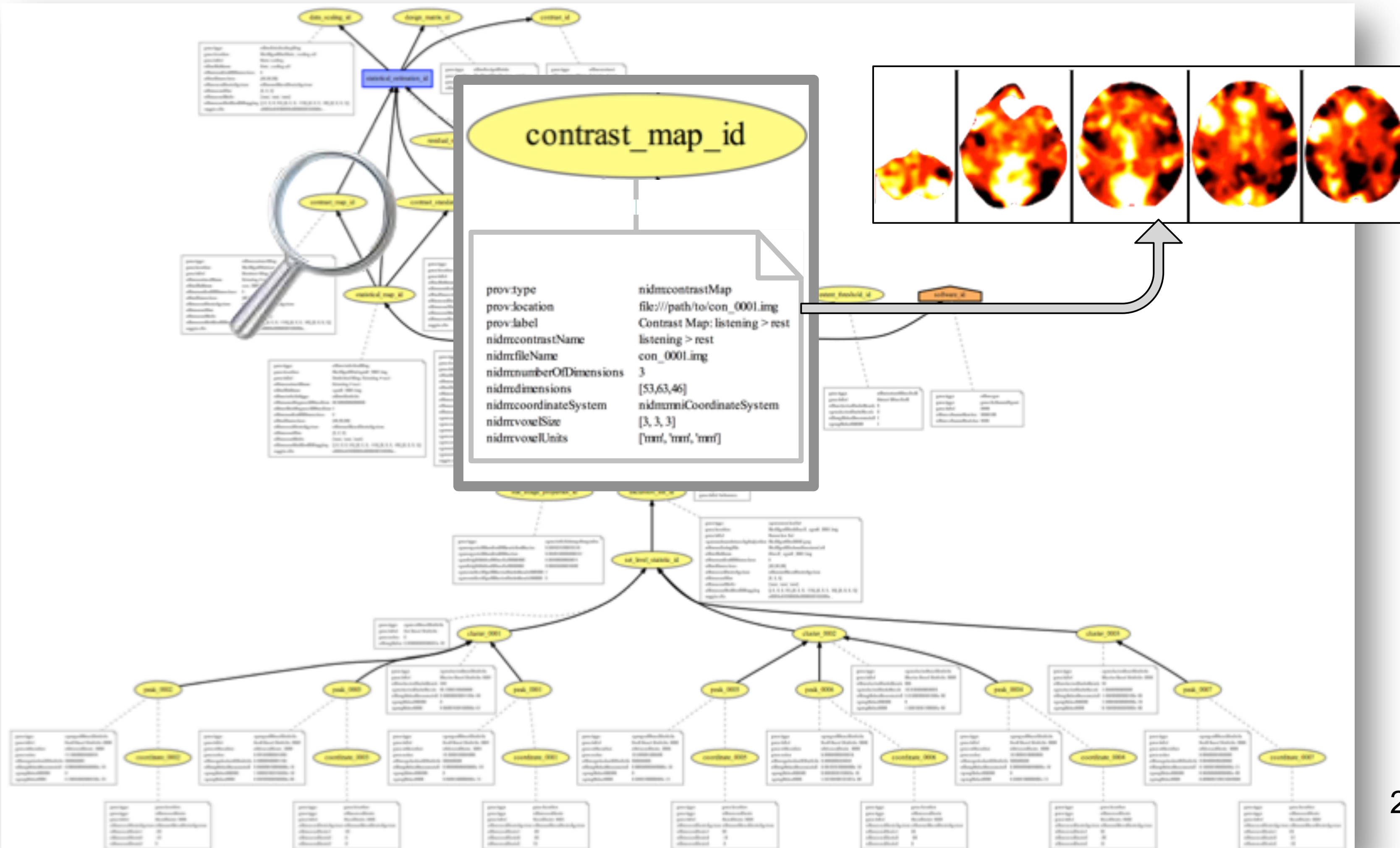


WARWICK

Data model: entities



Data model: entities



Conclusion

- Data sharing is one key to reduce skepticism.
- There is already a number of technical solutions for data sharing in neuroimaging.
- A meta-data standard would beneficiate to all of these efforts
  - NI-DM: http://nidm.nidash.org

Q & A

This work is supported by the wellcome trust