



# Neuroinformatics techniques for provenance & data sharing

Camille Maumet

## ► To cite this version:

Camille Maumet. Neuroinformatics techniques for provenance & data sharing. GlaxoSmithKline-Neurophysics Workshop on Skeptical Neuroimaging, Jan 2014, London, United Kingdom. . inserm-01887730

**HAL Id: inserm-01887730**

**<https://inria.hal.science/inserm-01887730>**

Submitted on 4 Oct 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Neuroinformatic techniques for provenance & data sharing

Camille Maumet

GlaxoSmithKline - Neurophysics Workshop on  
Skeptical Neuroimaging

January 14<sup>th</sup>, 2014

THE UNIVERSITY OF  
WARWICK

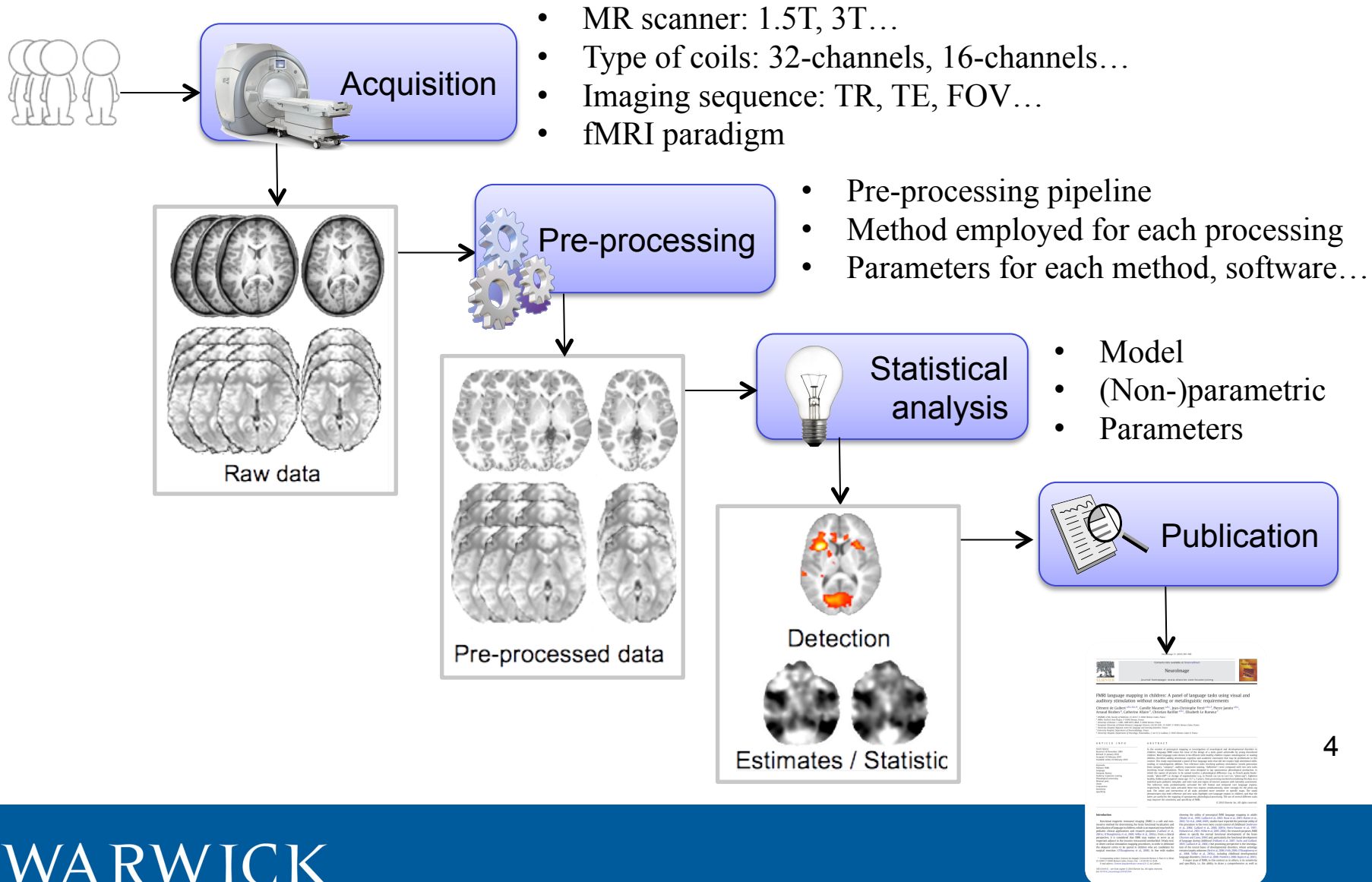
# Outline

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

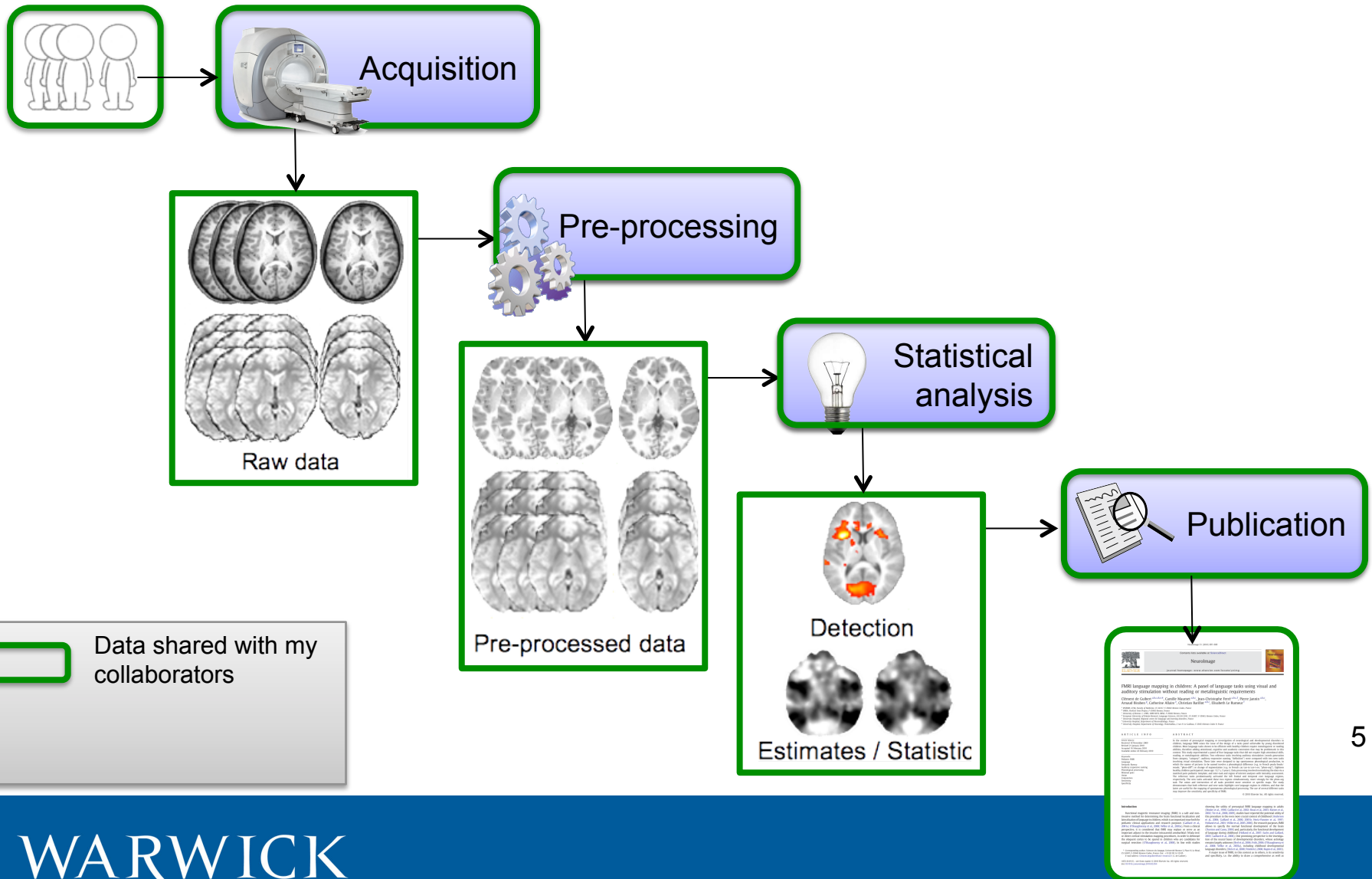
# Outline

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

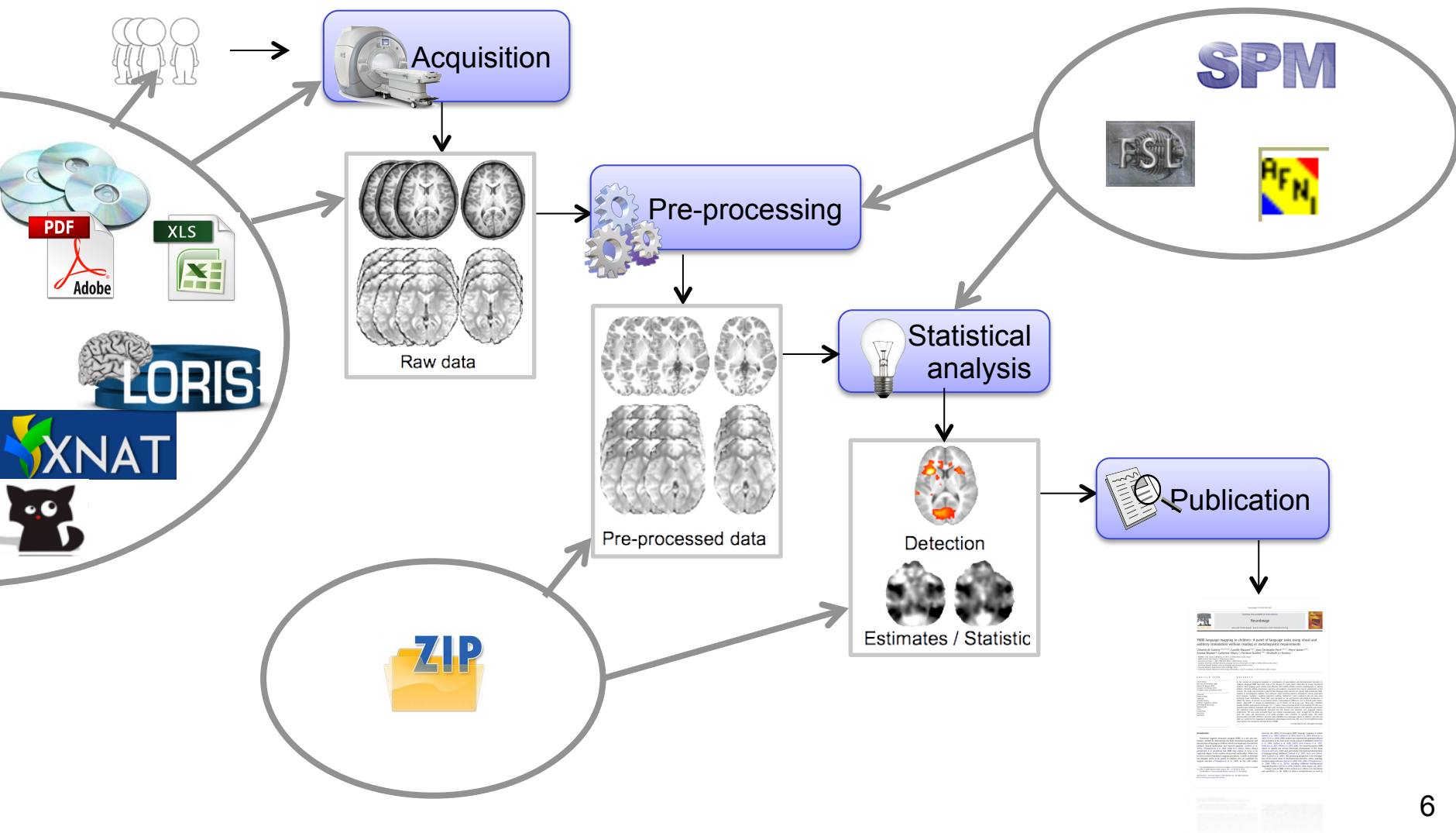
# Overview of a neuroimaging study



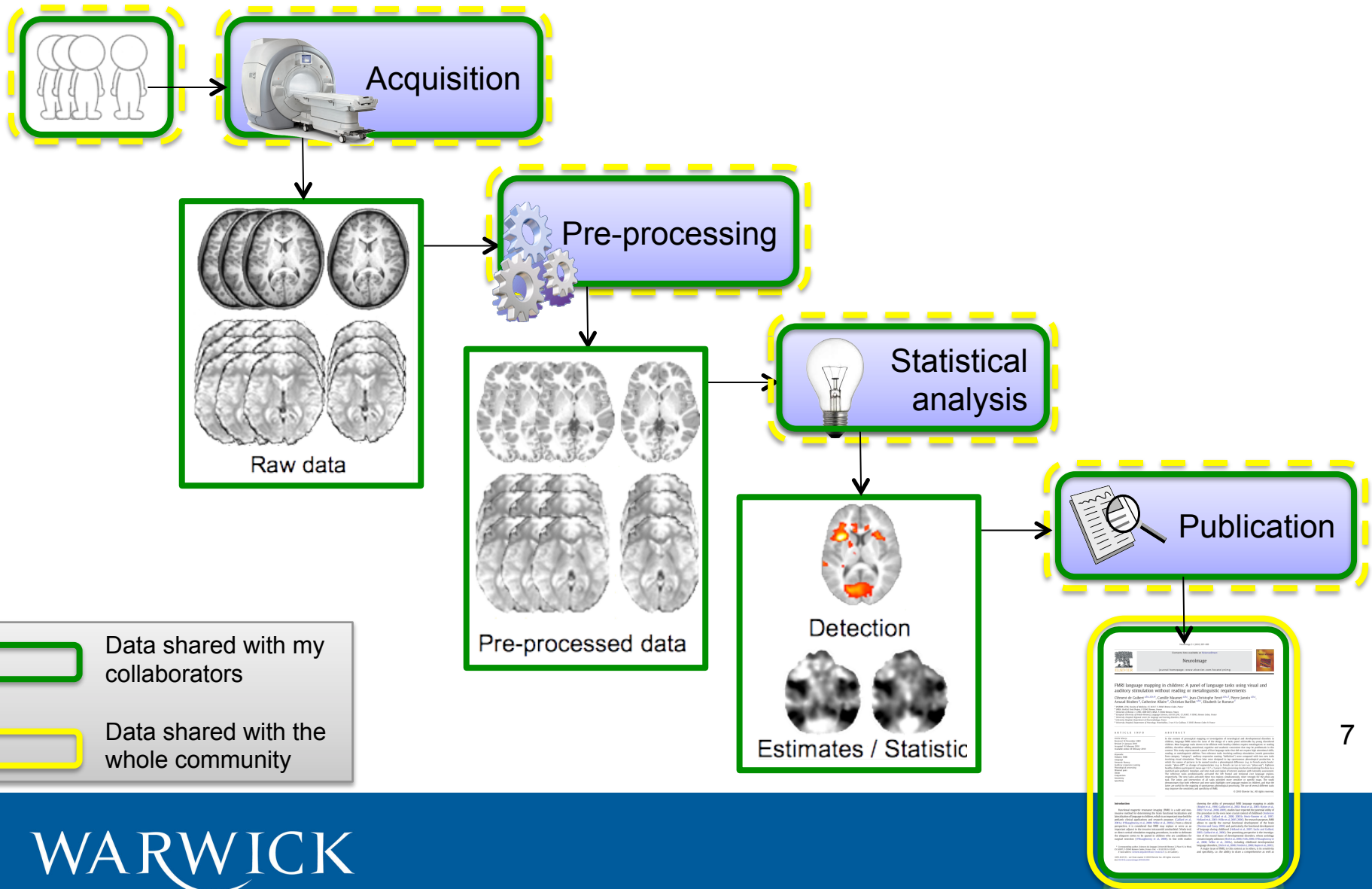
# Neuroimaging and data sharing



# Sharing data with my collaborators



# Neuroimaging and data sharing





# A neuroimaging publication

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

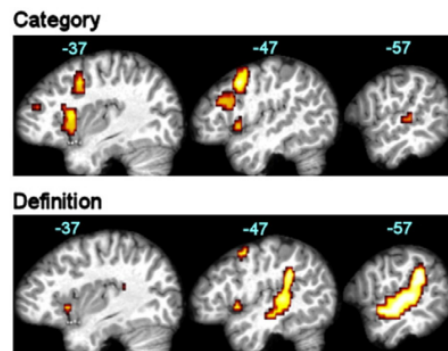
**General technical implementation**  
A single scanner session included the four paradigms separately implemented with the same parameters: a simple block design alternated a rest condition as control and the language task, starting with rest, with a preliminary period of signal acquisition for MRI 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 tasks	
Categ->Def	Def->Categ
Left Hemisphere	---
Inf frontal-Oper	---
Precentral	18-3.38 <sup>(3)</sup>
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;  $k = 10$ ).

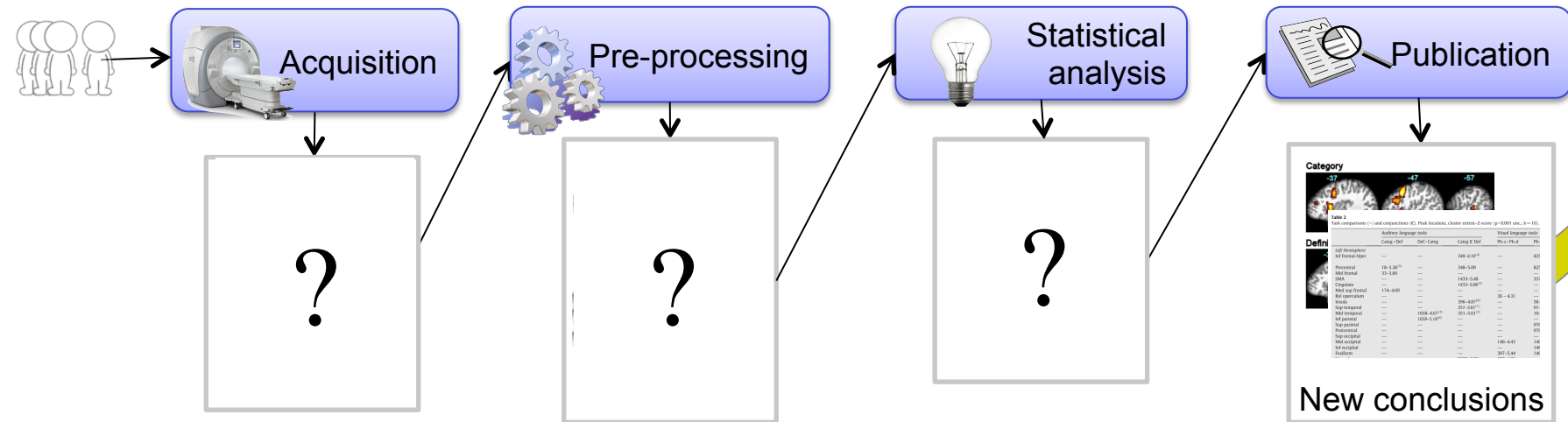
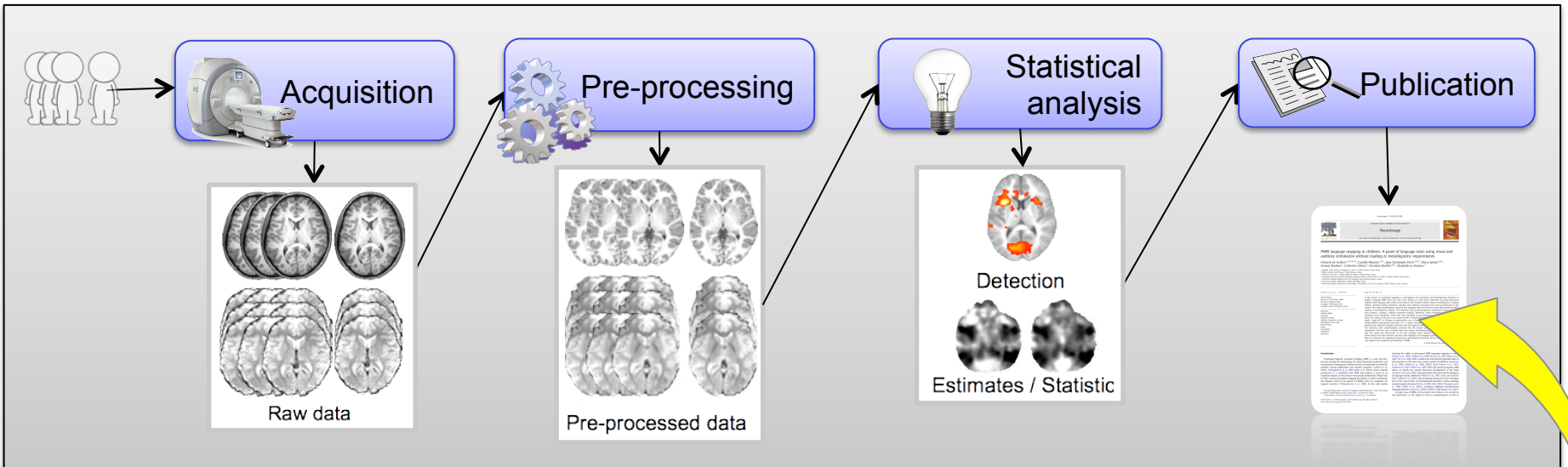
Auditory language tasks		Visual language tasks	
Categ->Def	Def->Categ	Categ C Def	Ph-s->Ph-d
Left Hemisphere	---	---	---
Inf frontal-Oper	---	348-4.10 <sup>(4)</sup>	---
Precentral	18-3.38 <sup>(3)</sup>	---	---
Mid frontal	33-3.66	---	---
SMA	---	1433-5.48	---
Cingulate	---	1433-5.08 <sup>(3)</sup>	---
Med sup frontal	174-4.69	---	---
Rol operculum	---	---	36-4.31
Insula	---	396-4.87 <sup>(8)</sup>	---
Sup temporal	---	351-3.81 <sup>(1)</sup>	---
Mid temporal	---	1658-4.67 <sup>(3)</sup>	---
Inf parietal	---	351-5.61 <sup>(2)</sup>	---
Sup parietal	---	---	---
Postcentral	---	---	---
Sup occipital	---	---	---
Mid occipital	---	---	---
Inf occipital	---	---	---
Fusiform	---	---	---

Table of local maxima

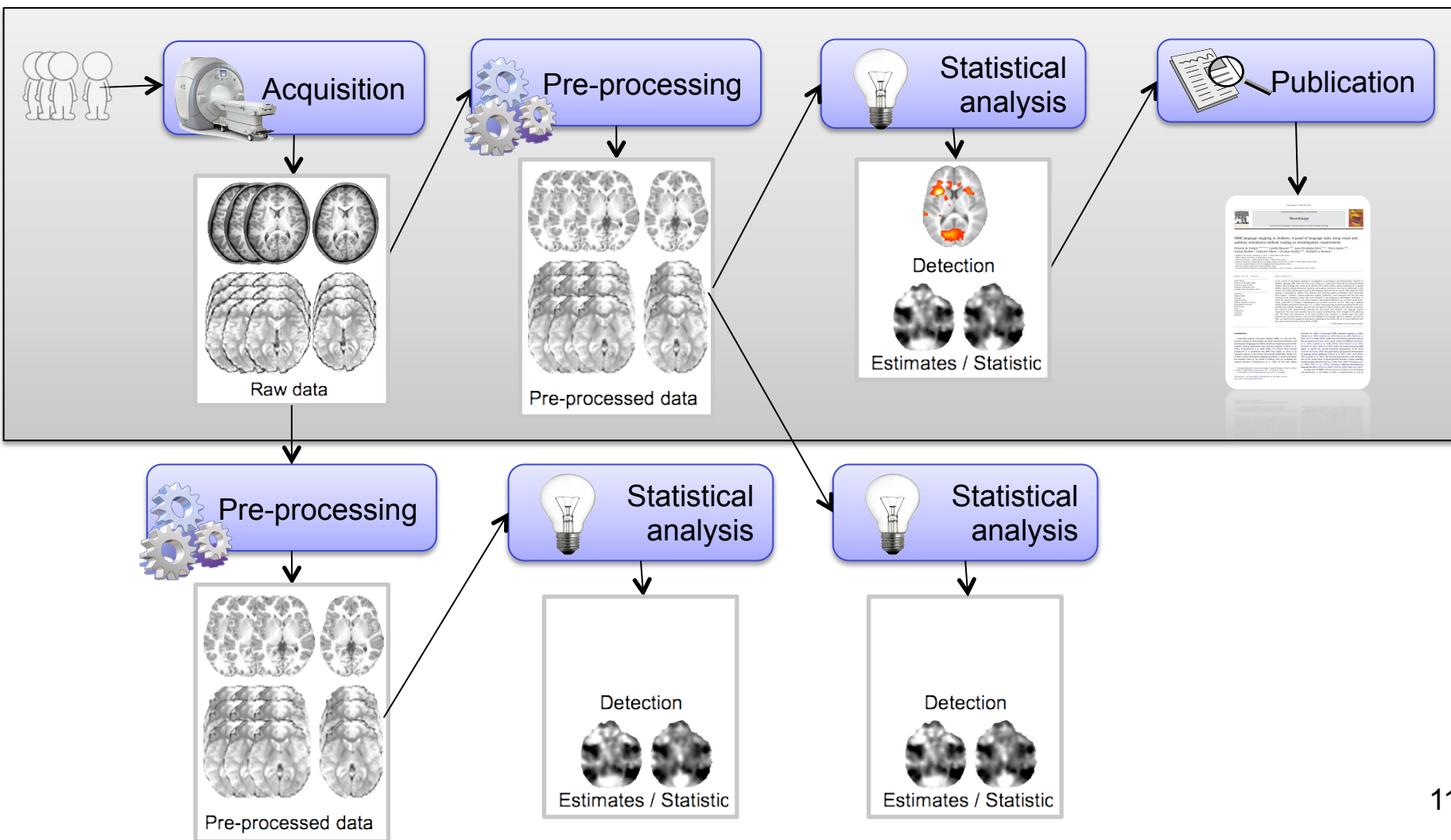
# Outline

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

# Reproducibility

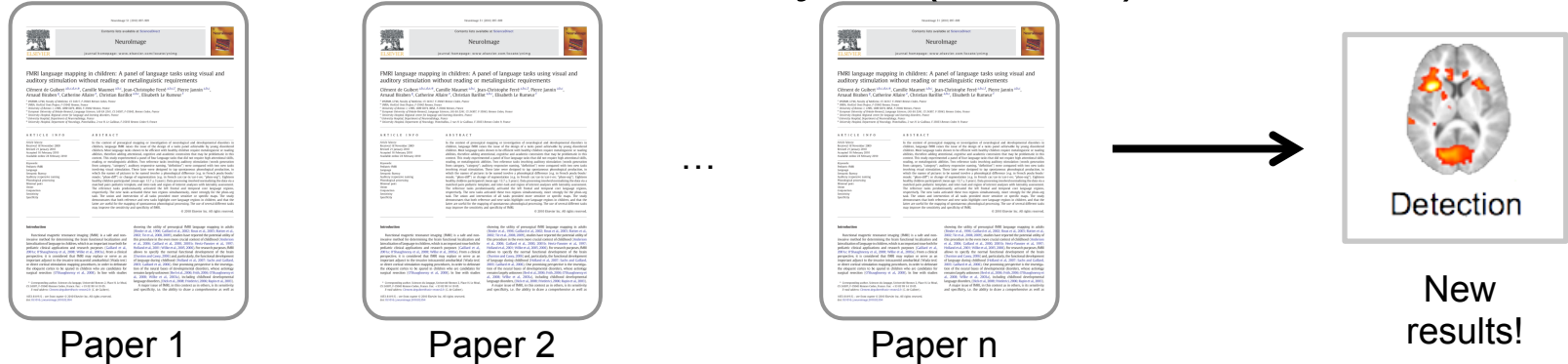


# Full provenance

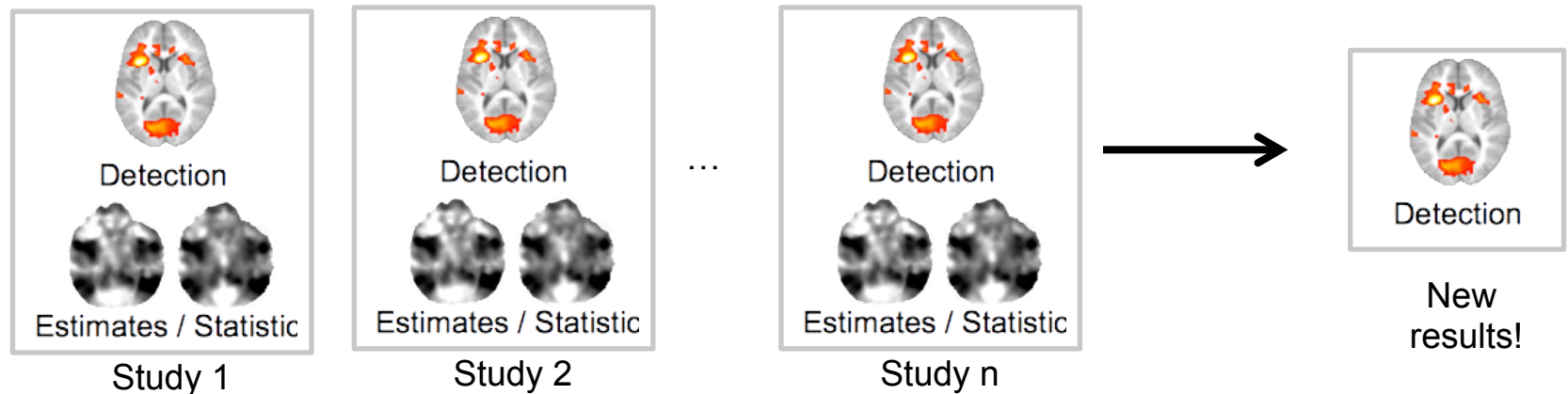


# Meta-analysis: analyzing the analyses

- 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.

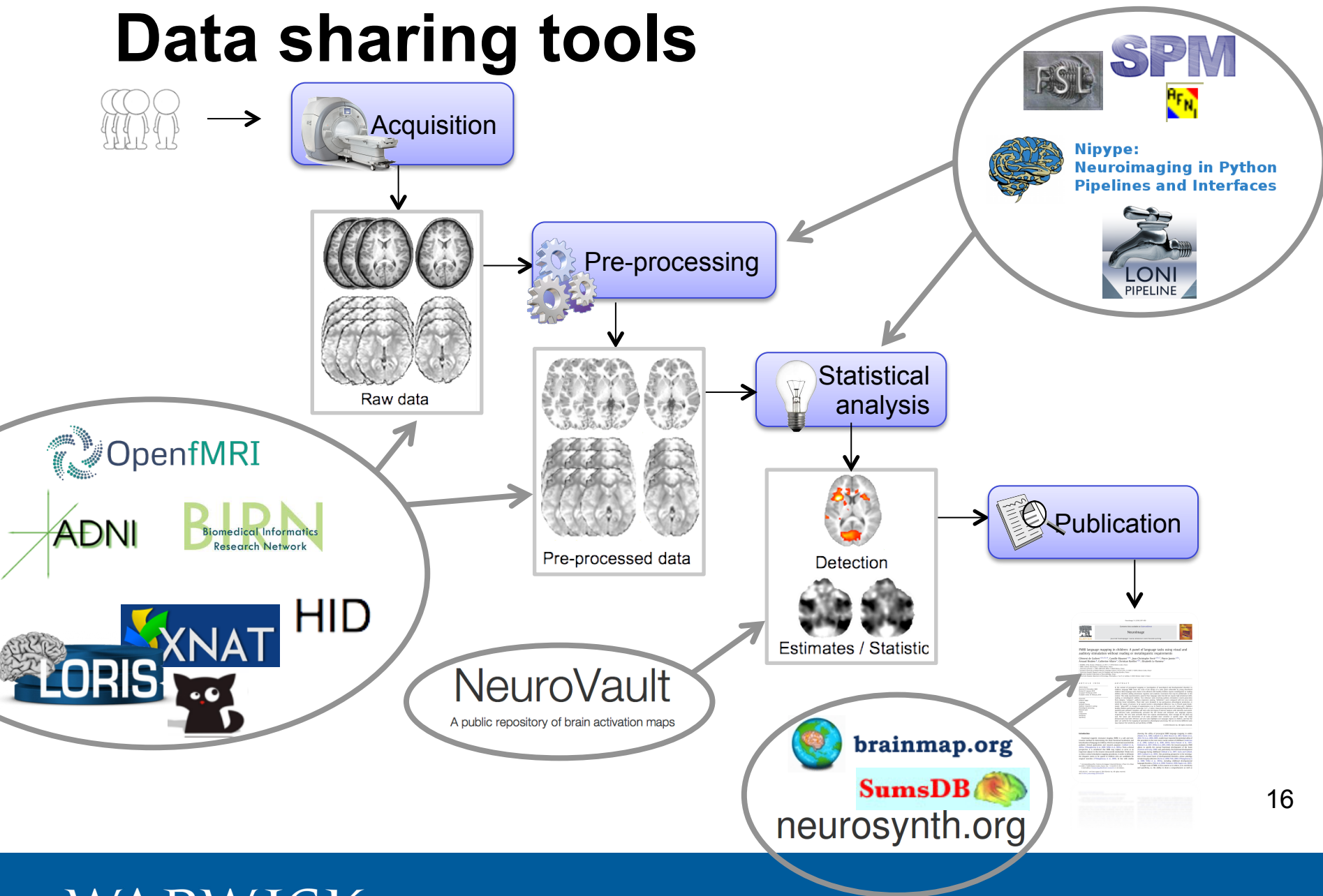
“Less than a few percents of acquired neuroimaging data is available in public repositories” [Poline 2012]

# 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

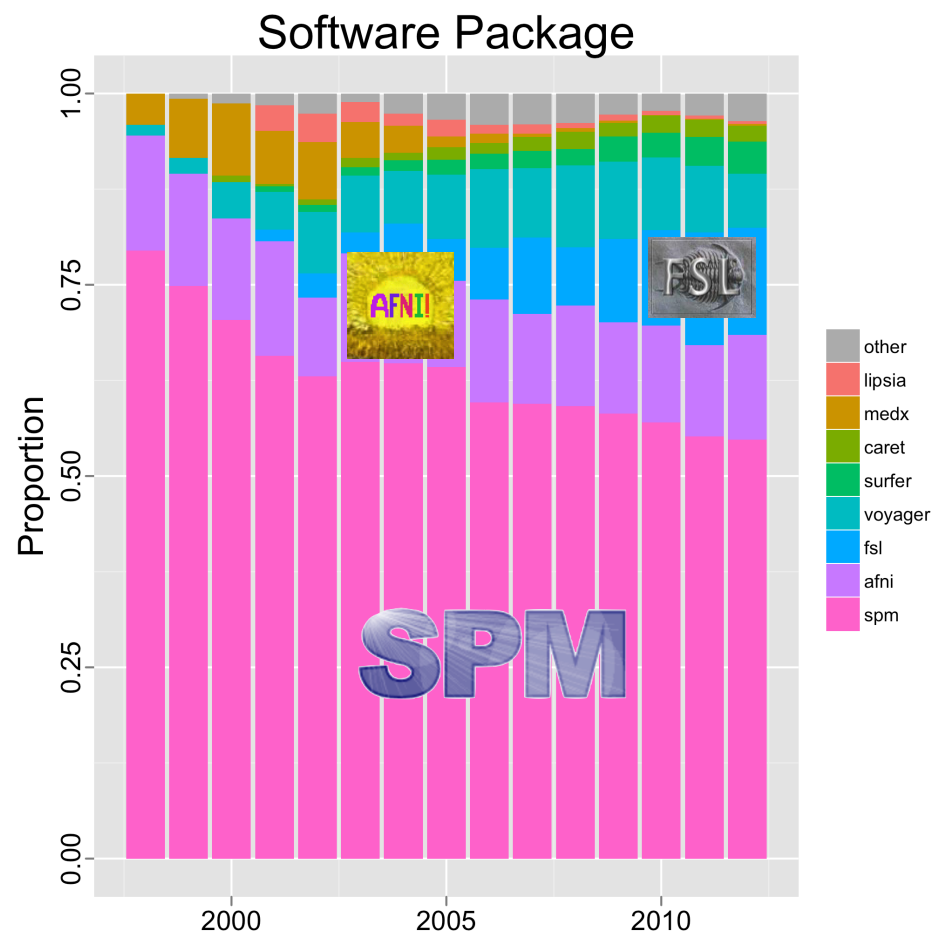


# A standard format for meta-data

- Sharing data across the data sharing tools...
- First attempt of an agnostic format: **XML-Based Clinical Experiment Data Exchange Schema (XCEDE)**: [www.xcede.org](http://www.xcede.org)
  - Describes subject, study, activation
  - Limited provenance encoding
  - Initiative of the BIRN
- **NeuroImaging Data Model NI-DM**: [www.nidm.nidash.org](http://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.



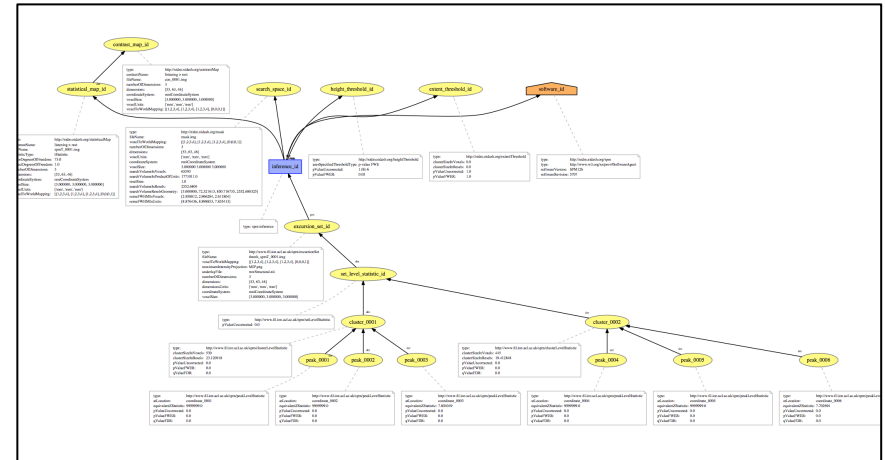
Automatically created with [Neurotrends](#) based on over 16 000 journal articles

# 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 output 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		0.0054 nidm:nidm_95
pValueUncorrected	Uncorrected p-value		0.0542 nidm:nidm_96

# Vocabulary



## Data model

# Neuroimaging terms

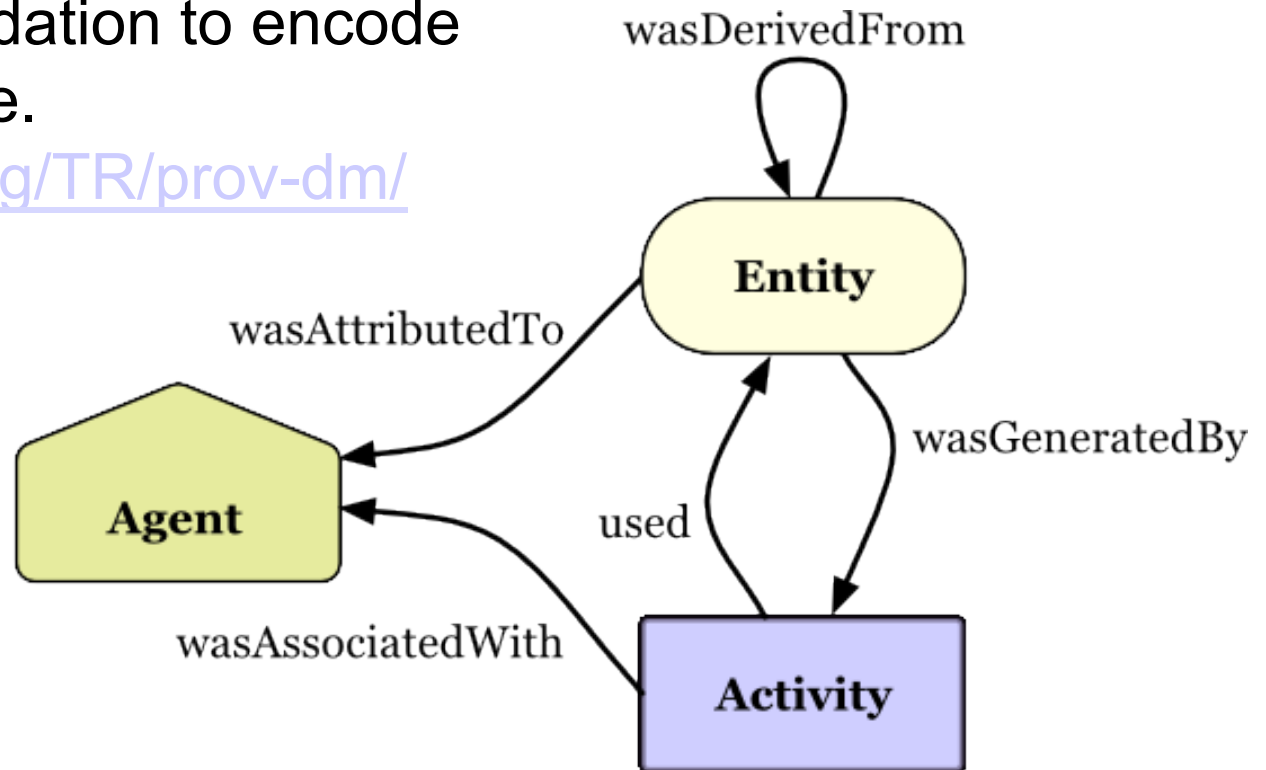
- Define a vocabulary to support the format.

Term name	Definition	Example	BIRN Lex or NIDM Concept ID	synonyms and related urls	Parent term
cluster	A group of neighboring image elements (voxels or vertices)		nidm:nidm_93	<a href="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43">http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43</a> or <a href="http://purl.obolibrary.org/obo/OBI_0000251">http://purl.obolibrary.org/obo/OBI_0000251</a>	
qValueFDR	p-value adjusted for the search volume, controlling for the False Discovery Rate	0.000154	nidm:nidm_94	<a href="http://purl.obolibrary.org/obo/OBI_0001442">http://purl.obolibrary.org/obo/OBI_0001442</a>	p-value i.e. nidm:nidm_0011
pValueUncorrected	Uncorrected p-value	0.0542	nidm:nidm_96	<a href="http://purl.obolibrary.org/obo/OBI_0001442">http://purl.obolibrary.org/obo/OBI_0001442</a>	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-dimensional 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	<a href="http://en.wikipedia.org/wiki/Resel">http://en.wikipedia.org/wiki/Resel</a>	
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-dimensional data.	3	nidm:nidm_108		
dimensions	Dimensions of some N-dimensional 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	

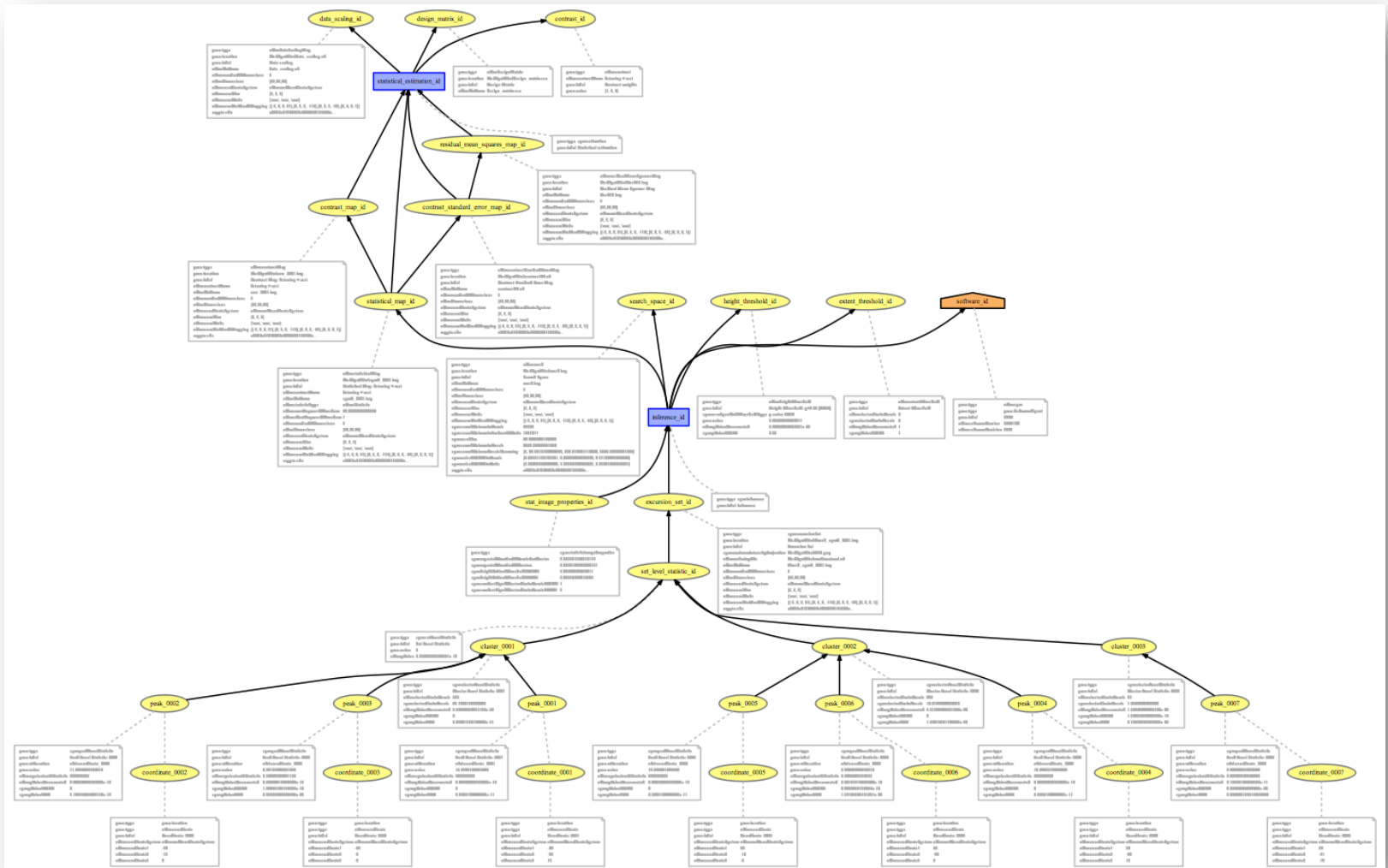
# Data model

- Based on PROV-DM a W3C recommendation to encode provenance.

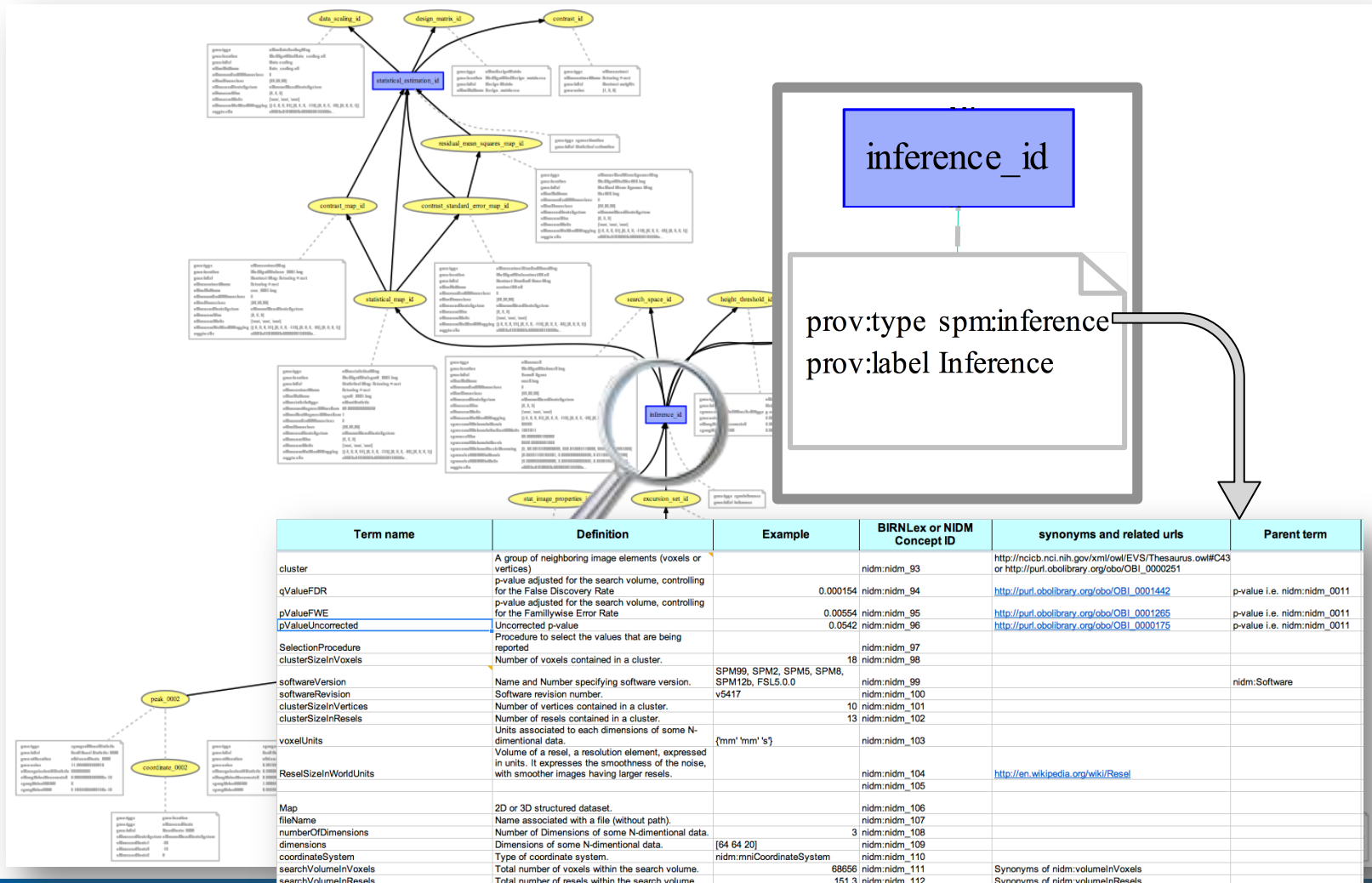
[www.w3.org/TR/prov-dm/](http://www.w3.org/TR/prov-dm/)



# Data model



# Data model: activities

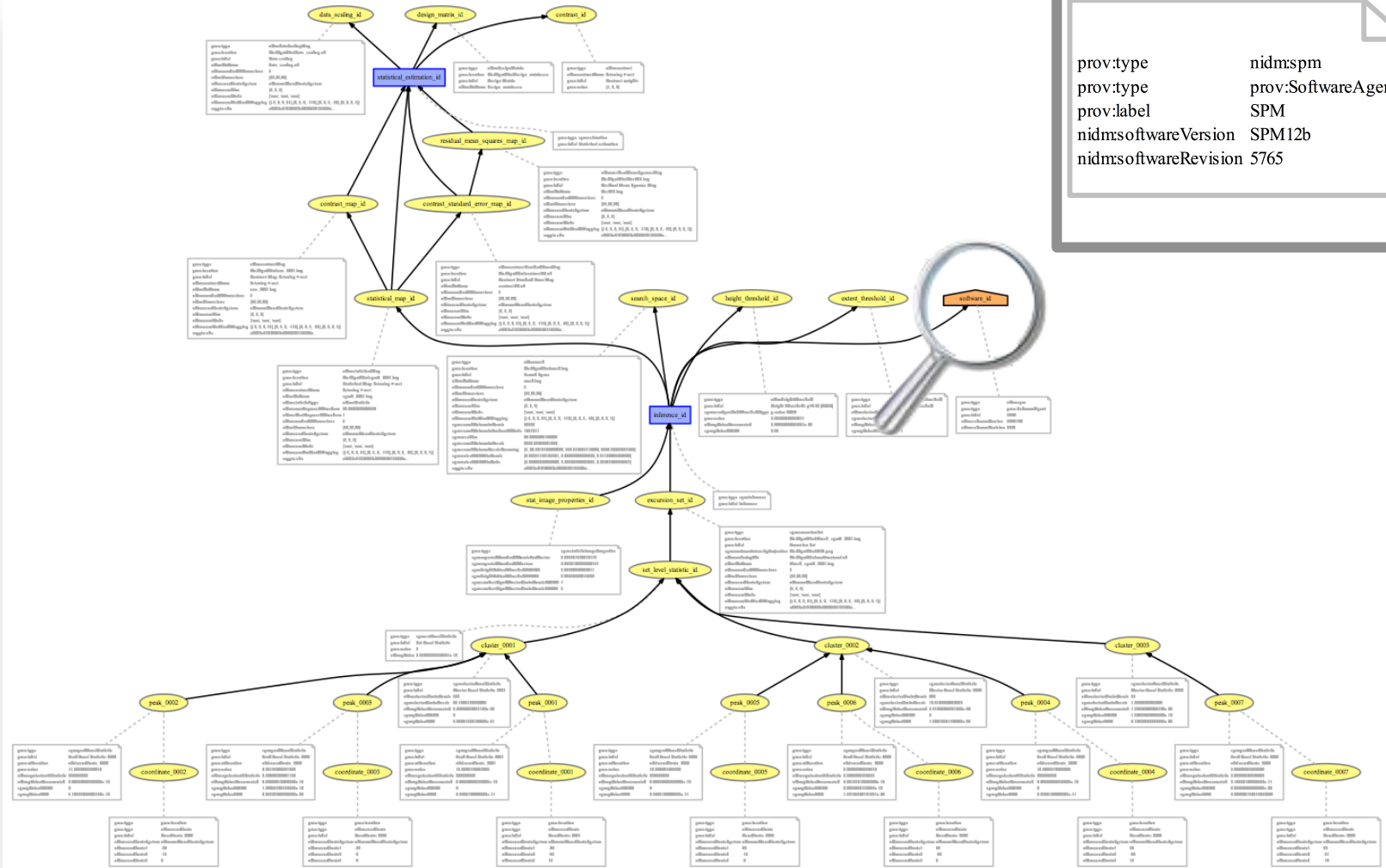




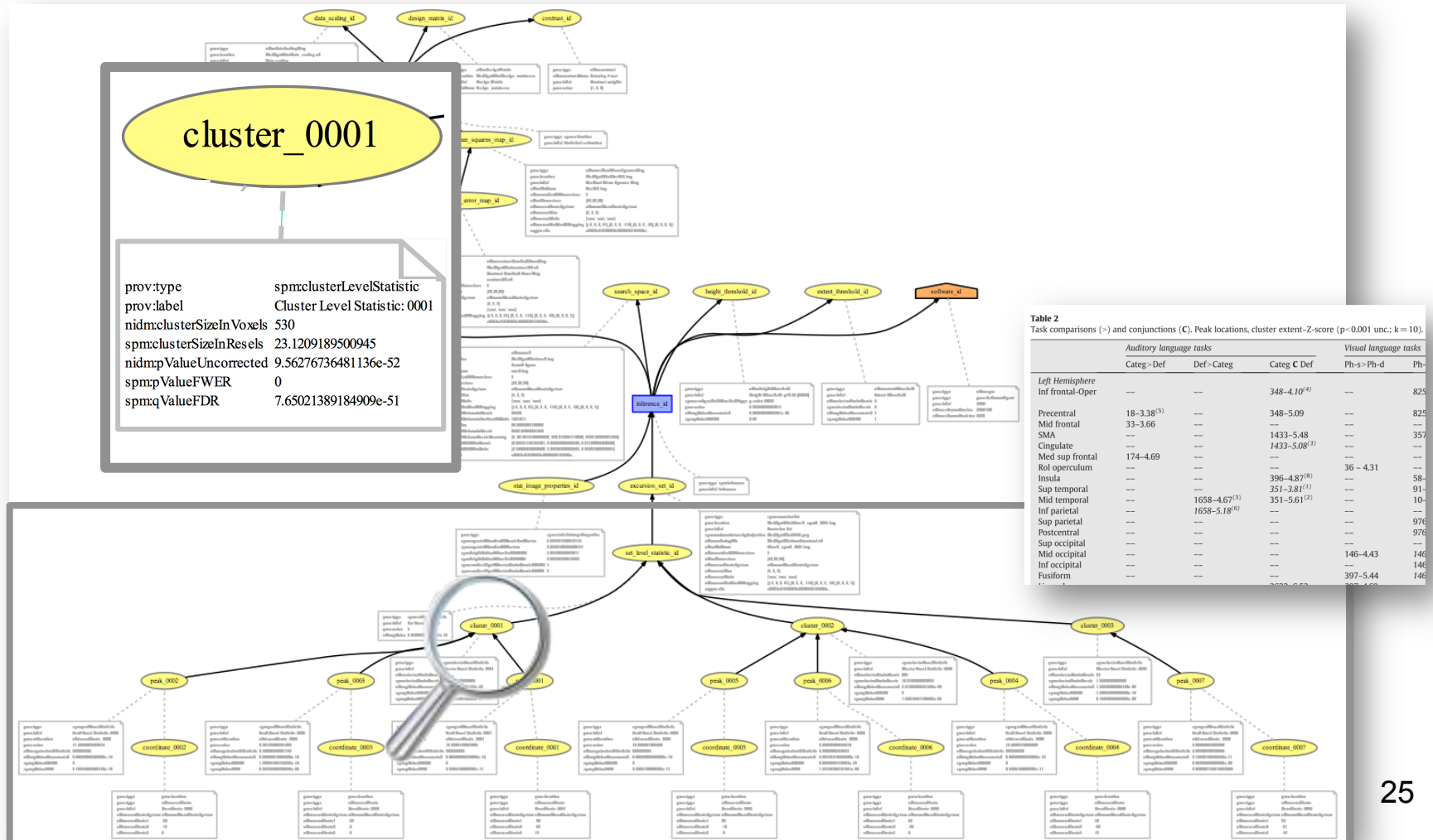
# Data model: agent

software\_id

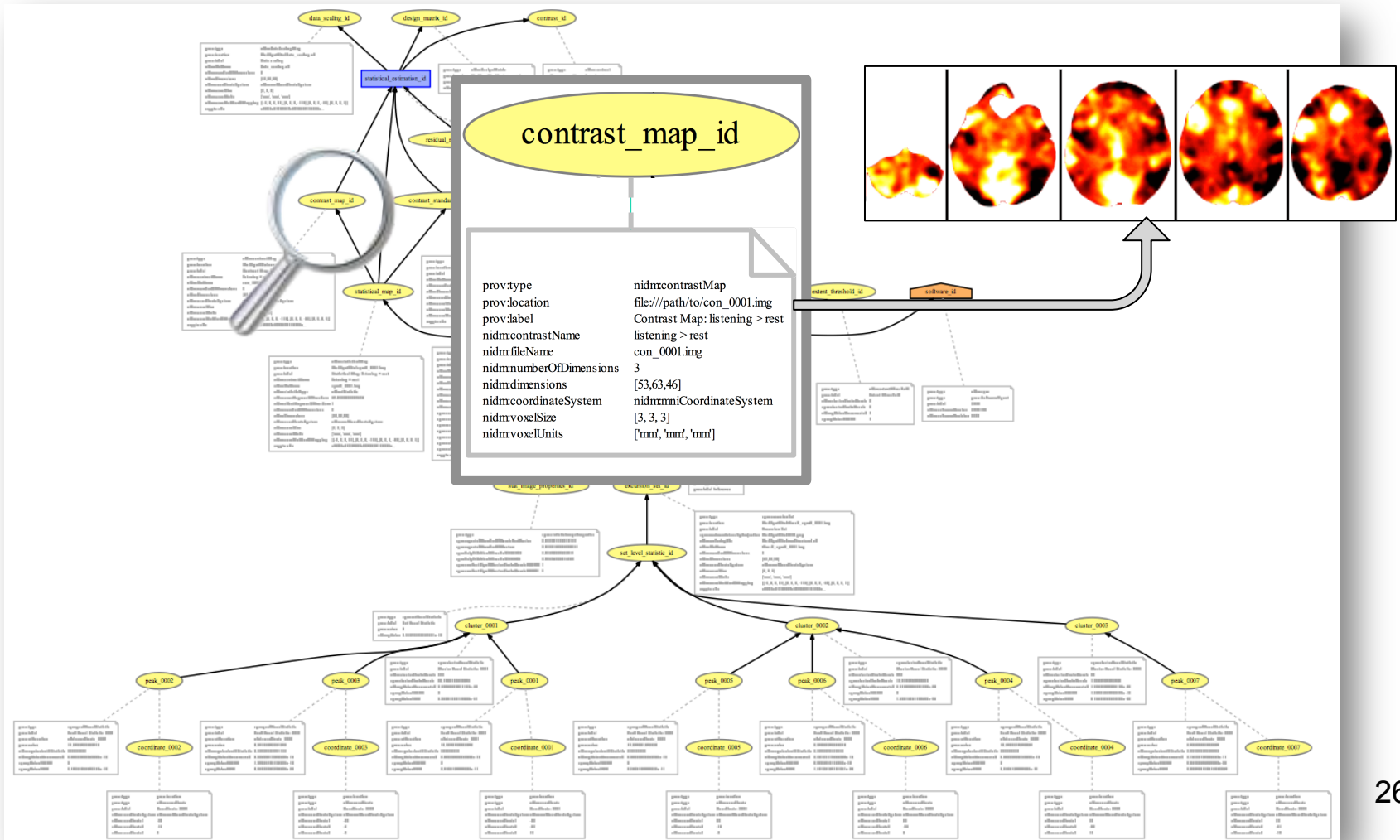
prov:type nidm:spm  
 prov:type prov:SoftwareAgent  
 prov:label SPM  
 nidm:softwareVersion SPM12b  
 nidm:softwareRevision 5765



# 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 benefit to all of these efforts
  - NI-DM: <http://nidm.nidash.org>

# Q & A

This work is supported by the **wellcome**trust