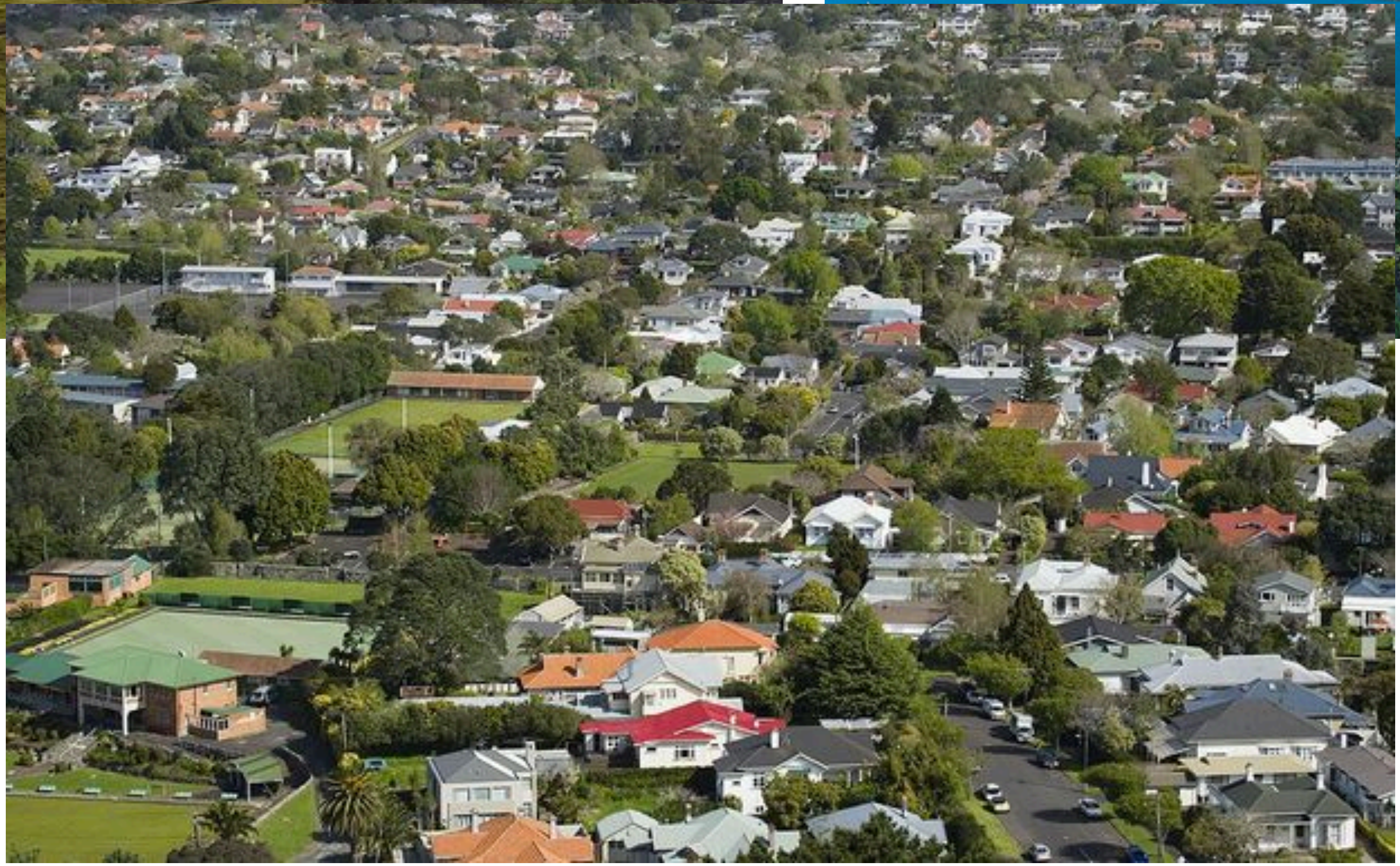
Categories are in flux, but their computational representations are static and isolated: that's a problem

Mark Gahegan & Prashant Gupta
Computer Science and Centre for eResearch
The University of Auckland

AUCKLAND



Ontology: definition

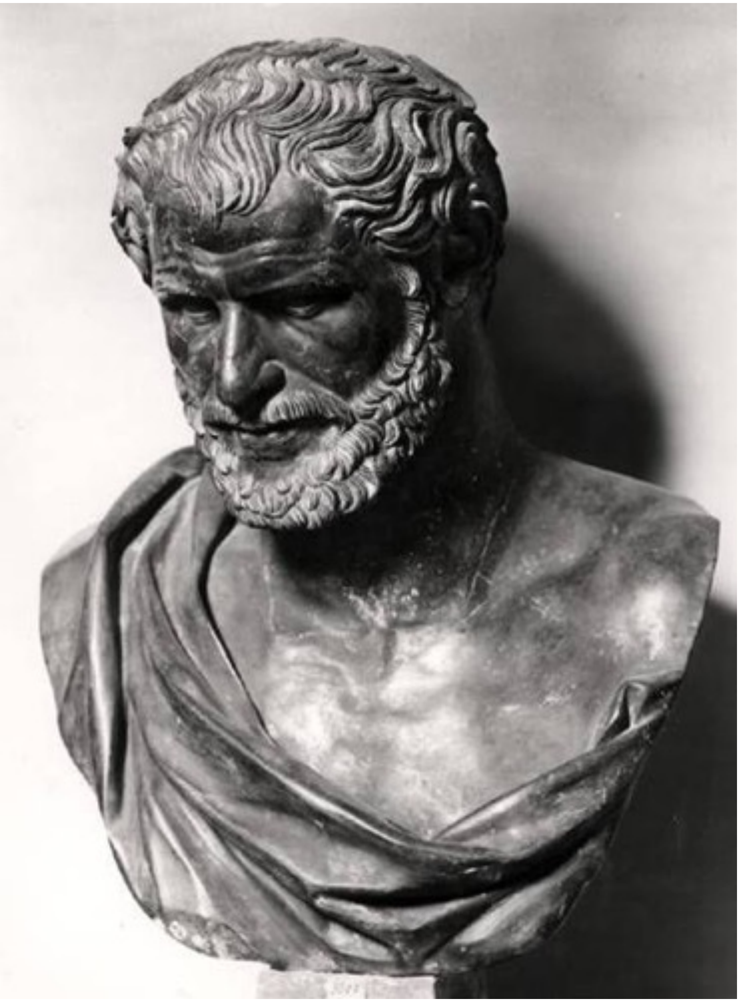
Asking the questions only a child would ask…

…And answering them in the language only a lawyer would use!

Everything changes…

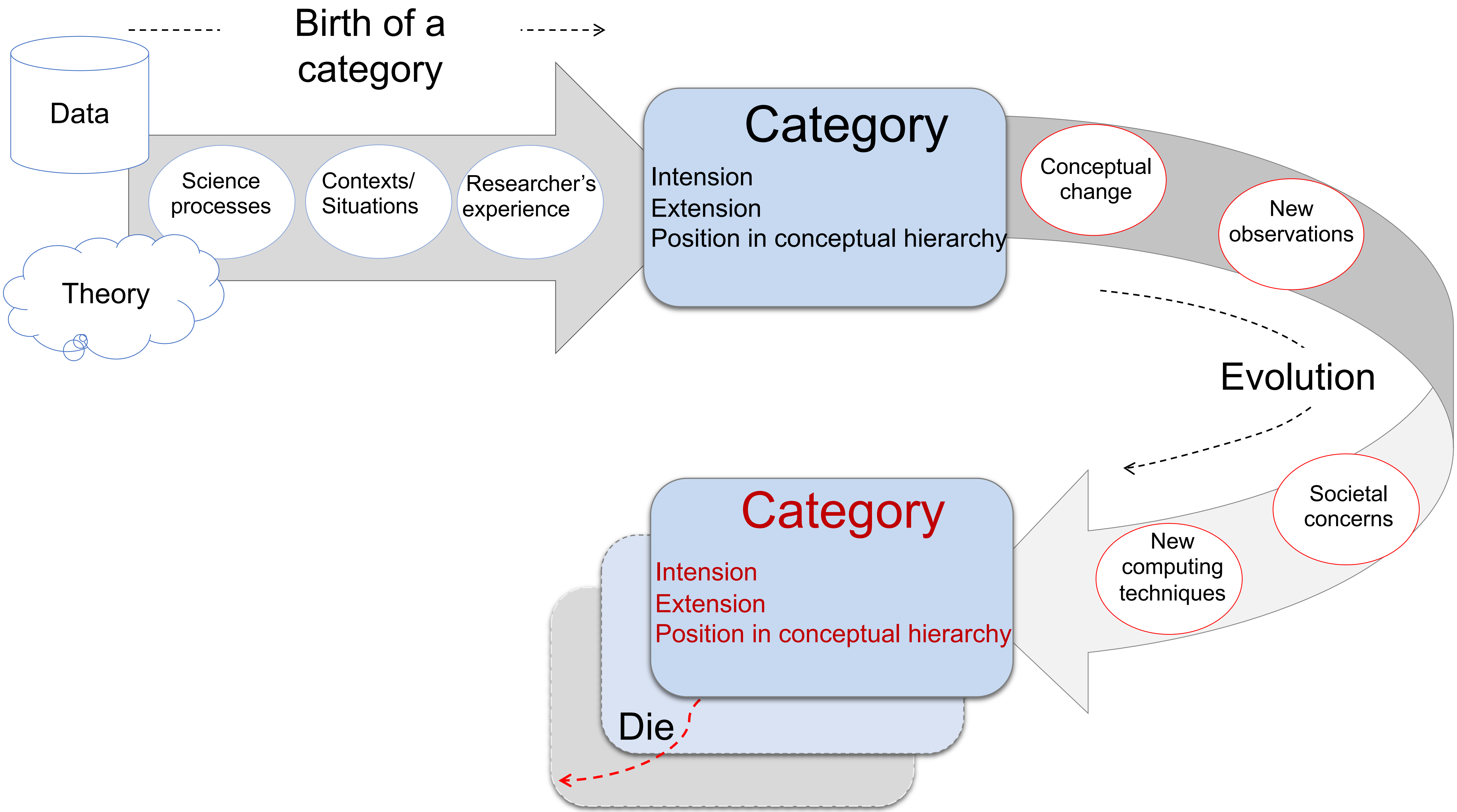
"According to Heraclitus, *panta rhei* — everything is in flux. But what gives that flux its form is the logos — the words or signs that enable us to perceive patterns in the flux, remember them, talk about them, and take action upon them even while we ourselves are part of the flux we are acting in and on…

Any system of ontology that is adequate for defining the concepts used in natural languages must be at least as flexible as the languages themselves" John Sowa: Signs, Processes and Language Games

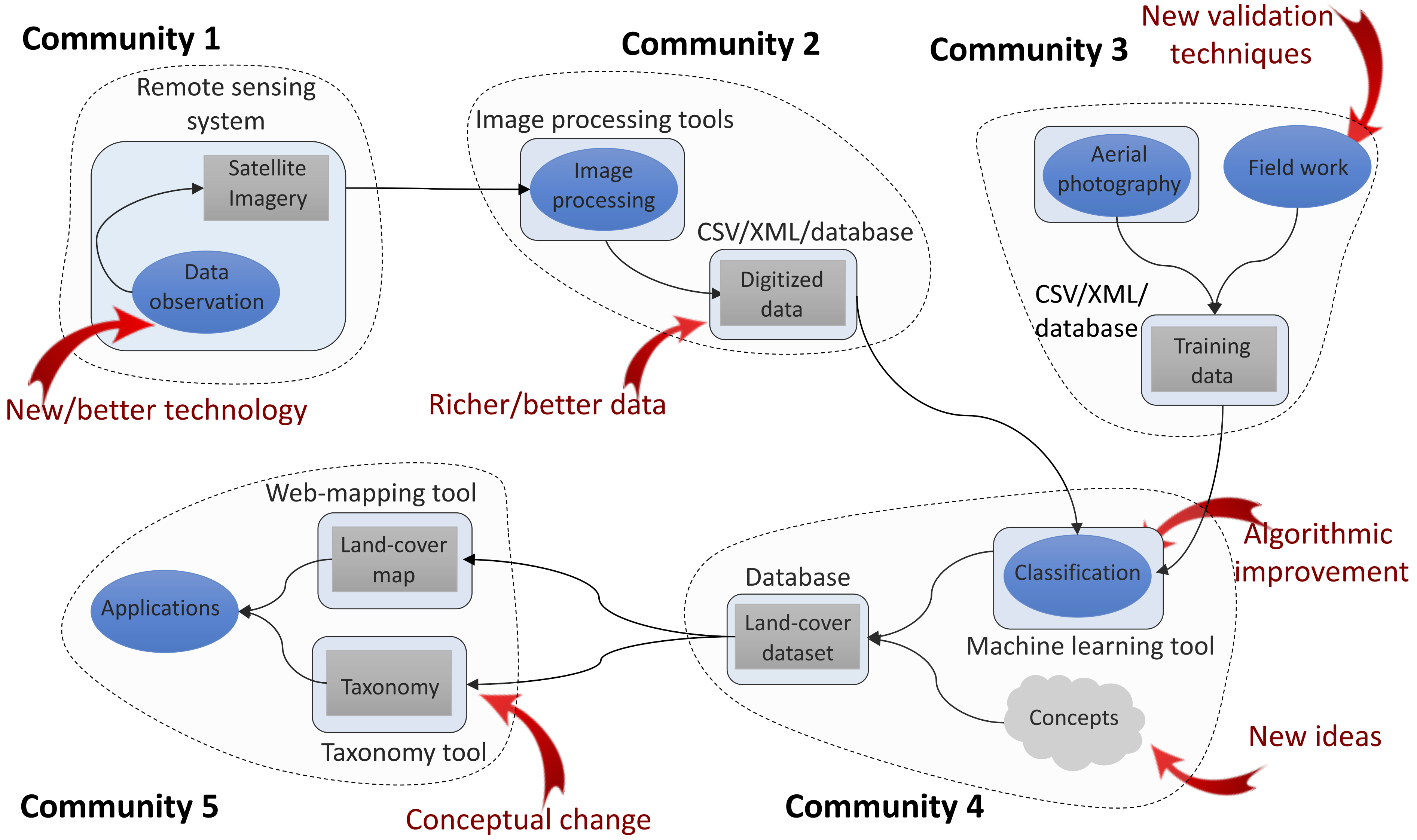


Categories

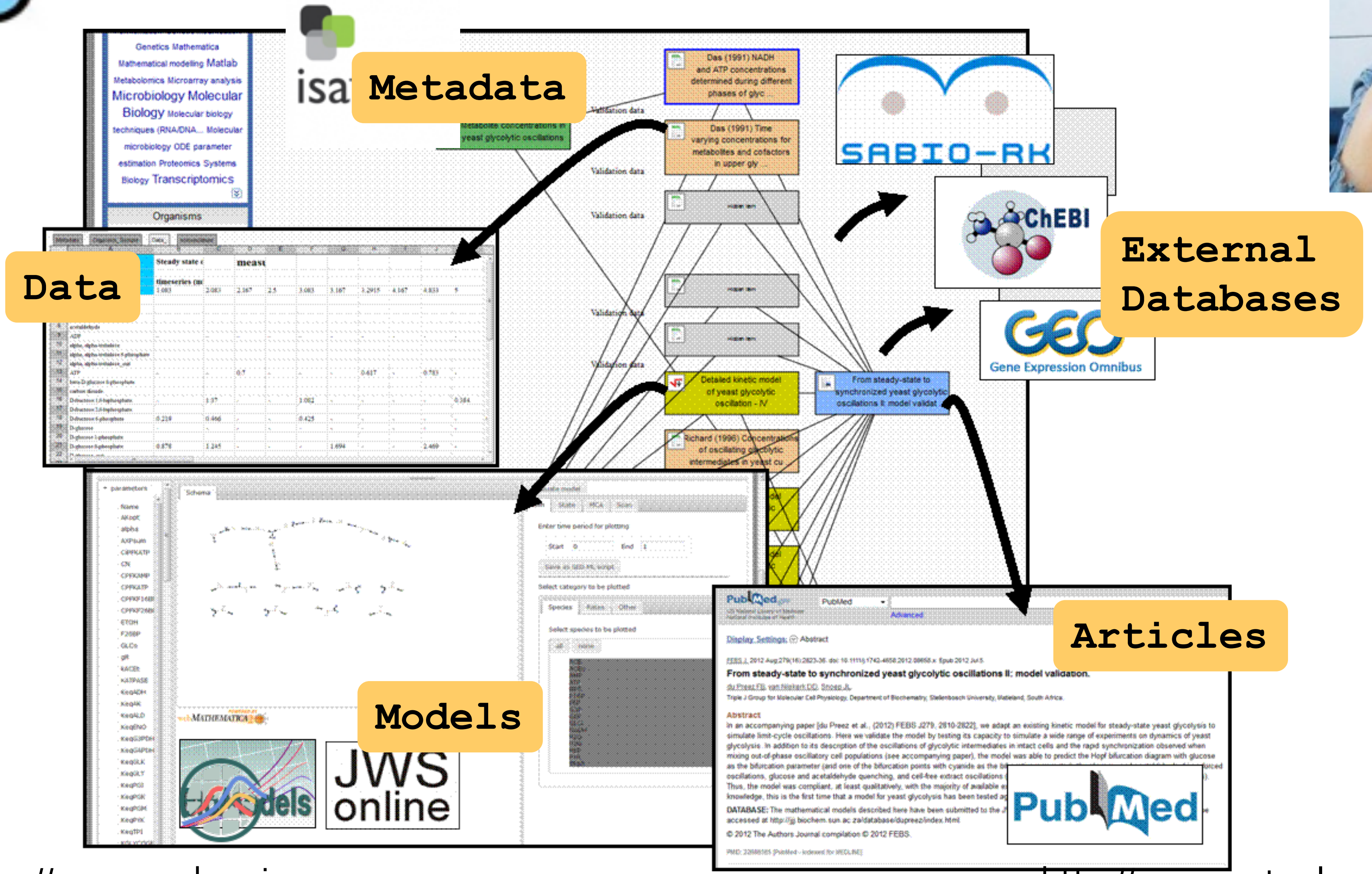
- Like Heraclitus, I think of objects and their categories as simply the instantaneous snapshot of the outworking of processes.
- These processes occur at different levels of abstraction, for example:
  - there are natural processes (and societal processes) occurring in our world
  - We see them through observation processes
  - We interpret them through analytical processes
  - We gain understanding of them through experiential processes
  - We agree how and what we will name and describe via social processes


Birth of a category
Data
Theory
Science processes
Contexts/ Situations
Researcher's experience
Category
Intension
Extension
Position in conceptual hierarchy
Conceptual change
New observations
Evolution
Societal concerns
New computing techniques
Category
Intension
Extension
Position in conceptual hierarchy
Die

Fragmentation of scientific artifacts and processes among communities



Interlinking methods, models, data, samples..



http://www.seek4science.org

http://www.isatools.org

Other knowledge integration models..

- Research Ob

S. Bechhofer, D. De ards exchange and reuse of digital know nce, NC, USA, 2010.

- Reproducib

J. P. Mesirov, "Acces

- Linked Scie

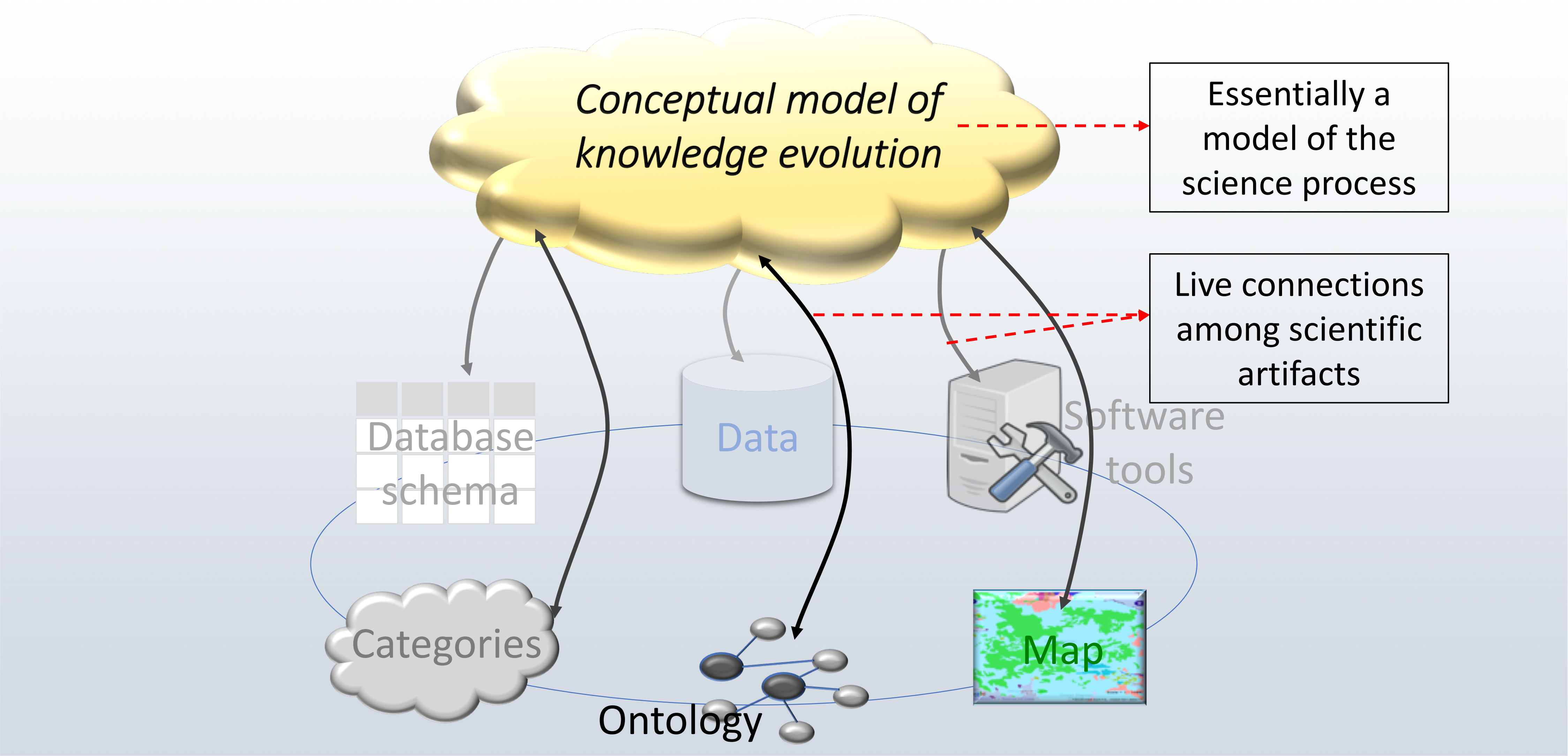
T. Kauppinen and G. valuating data, methods and results Science (ICCS), 2011.

- Workflows

What are the shortcomings?

- Focus on a single experiment of science, rather than science as an ongoing and evolving process
- Provide a linear view of science, but science is instead exploratory, dynamic and cyclic
- Focus typically on data and not on conceptual structures

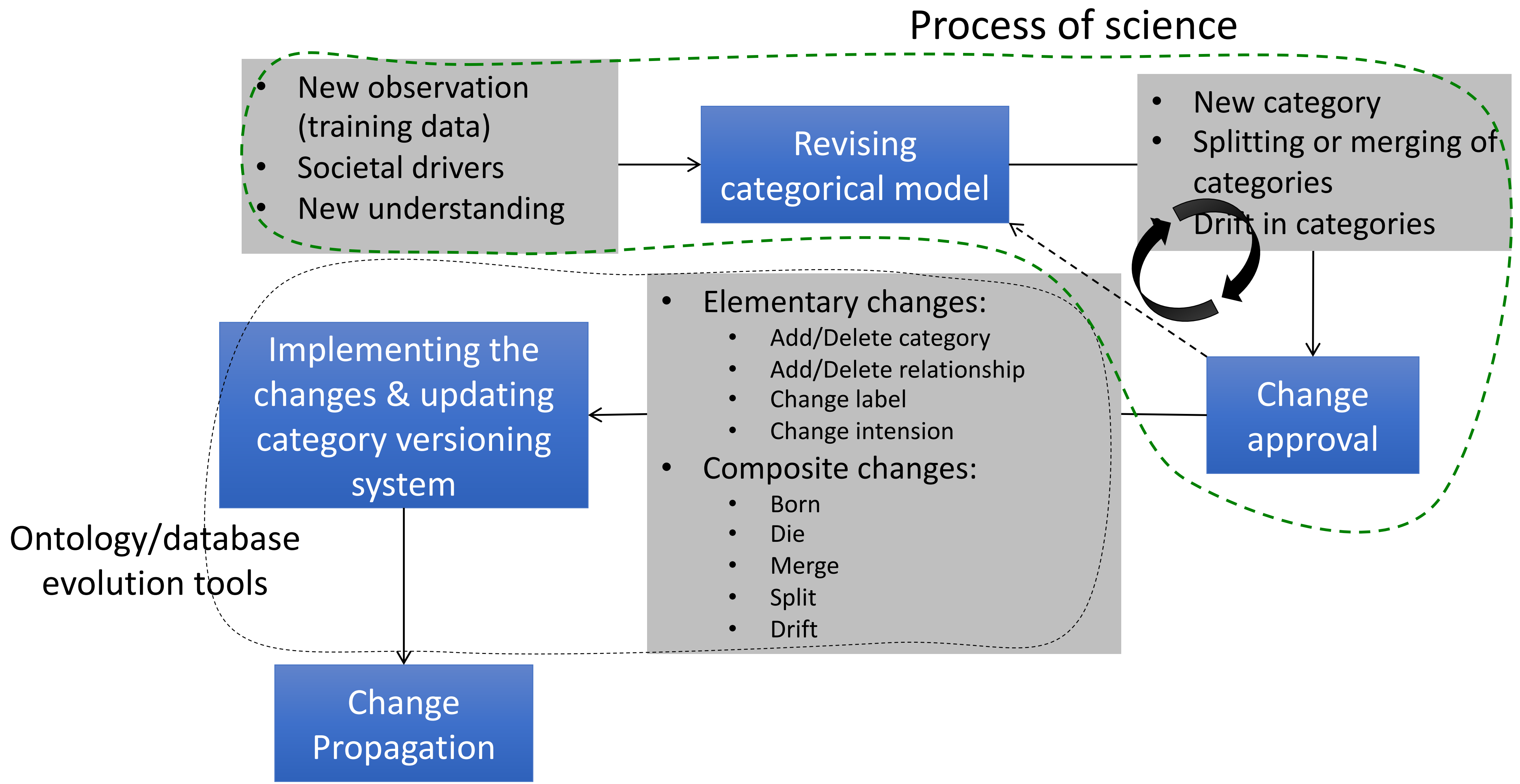
Connecting scientific artifacts



Adventures of Categories (AdvoCate)

- An integrated system for managing categories in action, based on a process model of category evolution
- Captures changes in categories, via the process of category evolution and maintains a category-versioning system
- Allows the entire evolution process to be replayed, questioned, communicated
- Can compare versions of categories based on intension as well as extension

Process model of category evolution



Change language

Change Algebra

Typology	Operations
Birth	$\circledcirc C$
Death	$\otimes C$
Split	$C \ominus C_1, C_2$
Merge	$C \oplus C_1, C_2$
Drift	$C \leadsto C'$

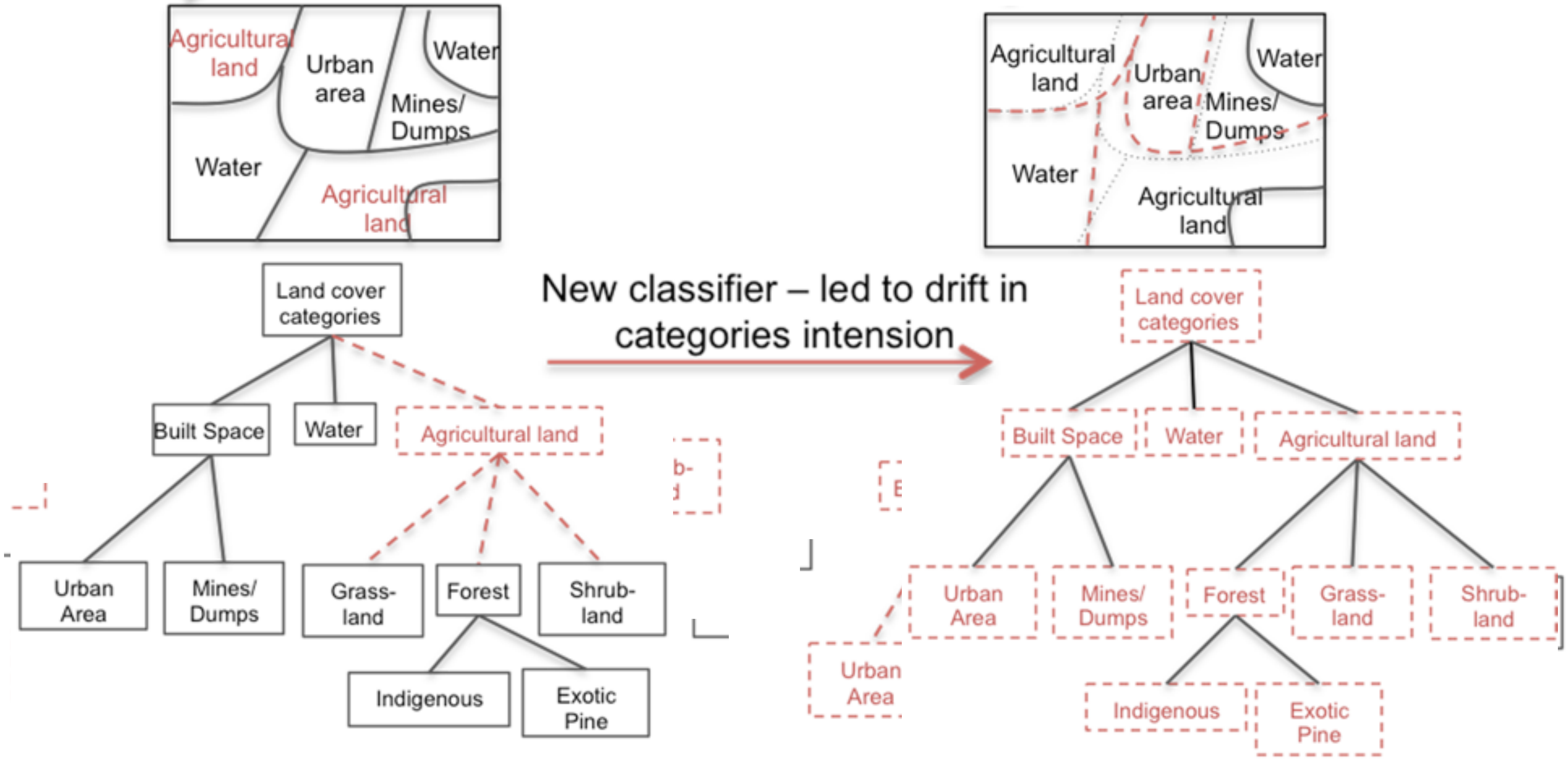
Relational Operations

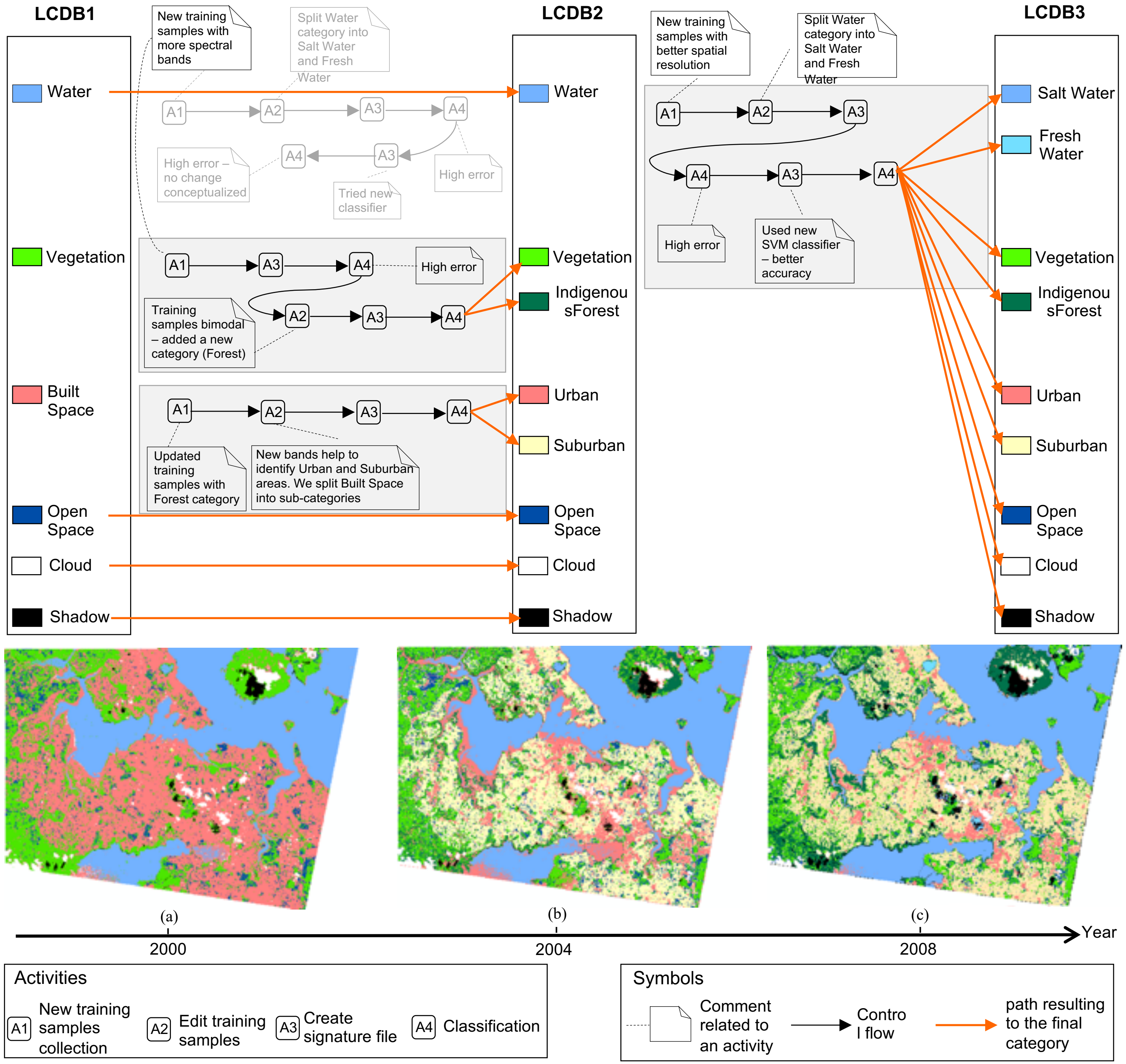
Relationship	Symbol
Is same as	$\asymp$
contains	$\Supset$
Is contained by	$\Subset$
Is confused with	$\parallel$
Is independent of	$\nparallel$

Internal & external change triggers

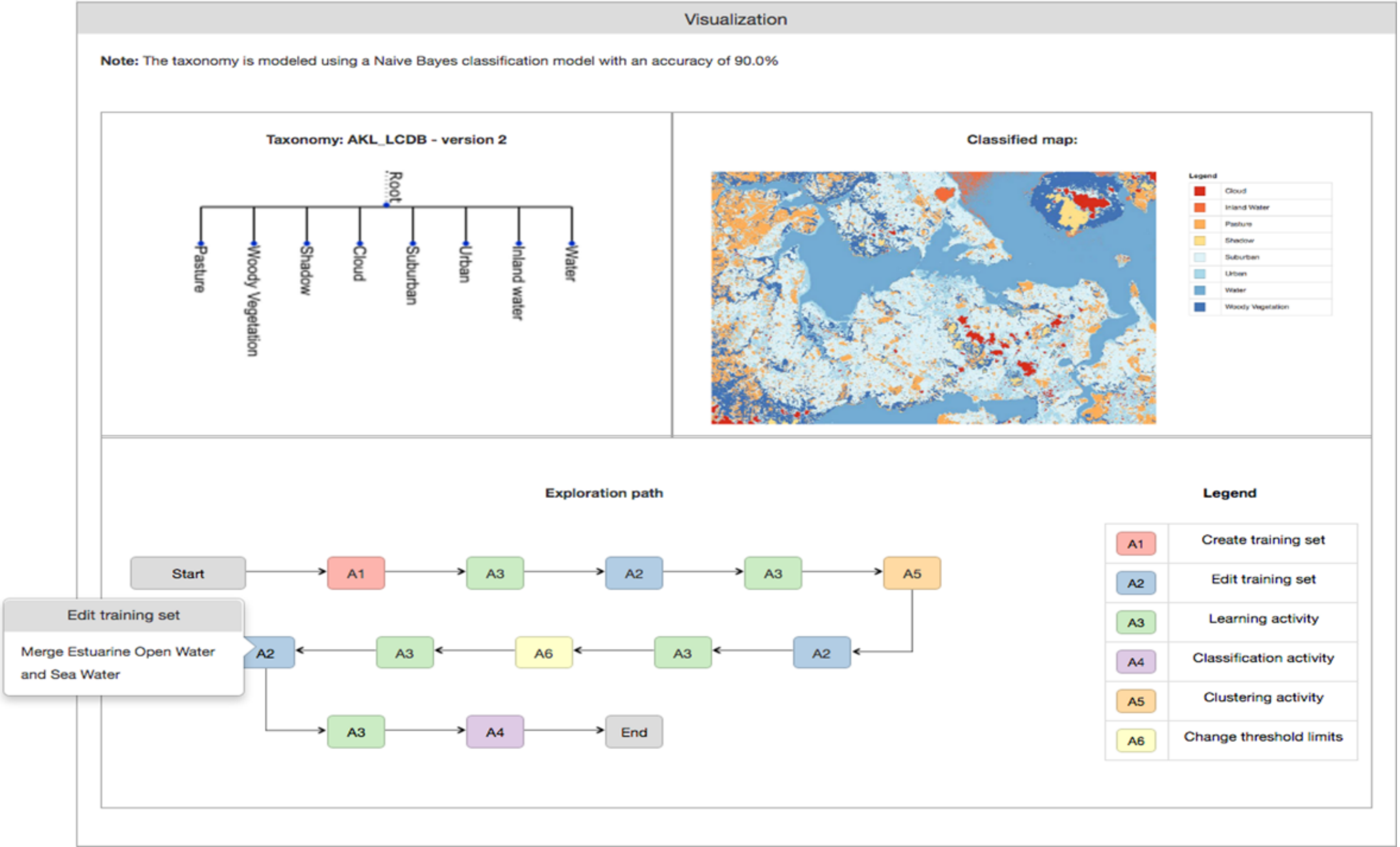
Change trigger (external)	Change trigger (internal)
Social requirement (addition of a new category)	Change in training sample
Social requirement (descriptive categories)	Change in training sample
Scientific requirement (accuracy)	New classifier
Scientific requirement (accuracy)	Change in training sample
Scientific requirement (accuracy)	Change in parameters
Conceptual change	Change in training sample
Error in data collection activity	Change in training sample

An example of category evolution from land cover mapping

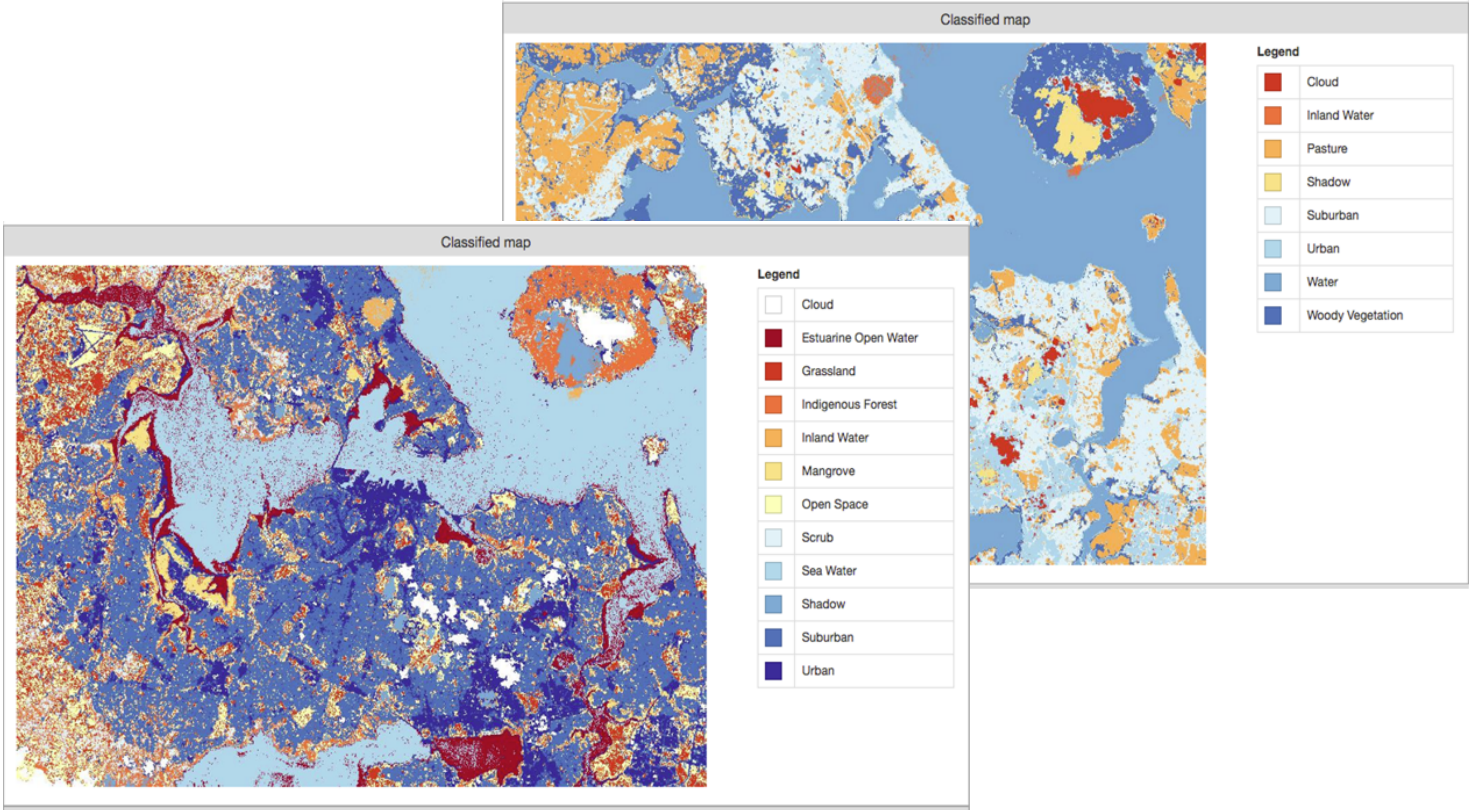




LCDB1
LCDB2
LCDB3
New training samples with more spectral bands
Split Water category into Salt Water and Fresh Water
High error – no change conceptualized
Tried new classifier
High error
High error
Training samples bimodal – added a new category (Forest)
Updated training samples with Forest category
New bands help to identify Urban and Suburban areas. We split Built Space into sub-categories
New training samples with better spatial resolution
Split Water category into Salt Water and Fresh Water
High error
Used new SVM classifier – better accuracy
Water
Vegetation
Built Space
Open Space
Cloud
Shadow
Water
Vegetation
Indigenou sForest
Urban
Suburban
Open Space
Cloud
Shadow
Salt Water
Fresh Water
Vegetation
Indigenou sForest
Urban
Suburban
Open Space
Cloud
Shadow
A1
A2
A3
A4
(a)
(b)
(c)
2000
2004
2008
Year
Activities
A1 New training samples collection
A2 Edit training samples
A3 Create signature file
A4 Classification
Symbols
Comment related to an activity
Contro l flow
path resulting to the final category



Visualization
Note: The taxonomy is modeled using a Naive Bayes classification model with an accuracy of 90.0%
Taxonomy: AKL_LCDB - version 2
Root
Pasture
Woody Vegetation
Shadow
Cloud
Suburban
Urban
Inland water
Water
Classified map:
Exploration path
Start
A1
A3
A2
A3
A5
A2
A3
A6
A3
A2
Edit training set
Merge Estuarine Open Water and Sea Water
A3
A4
End
Legend
A1 Create training set
A2 Edit training set
A3 Learning activity
A4 Classification activity
A5 Clustering activity
A6 Change threshold limits



Classified map
Legend
Cloud
Inland Water
Pasture
Shadow
Suburban
Urban
Water
Woody Vegetation
Classified map
Legend
Cloud
Estuarine Open Water
Grassland
Indigenous Forest
Inland Water
Mangrove
Open Space
Scrub
Sea Water
Shadow
Suburban
Urban

Summary of current exploration

The current exploration process, resulting from an external trigger of ' ***Conceptual change*** ', models changes in existing taxonomy ***AKL LCDB*** and leads to the following changes.

The new classification model is conceptualized as a Support Vector Machine with an accuracy of 86.0% as compared to the existing classification model stored in AdvoCate as Naive Bayes with an accuracy of 90.0%.

Below is given the different lists of categories and the changes that occurred:

Comparison between categories corresponding to the common concepts:

Concept	User accuracy (new)	Producer accuracy (new)	User accuracy (existing)	Producer accuracy (existing)	Extensional similarity	J M Distance
Cloud	0.93	1.00	0.95	0.99	0.59	N/A
Inland Water	0.86	0.88	0.73, 0.62	0.98, 0.94	0.26, 0.36	N/A
Shadow	0.94	0.96	0.9, 0.83	0.95, 0.96	0.68, 0.68	N/A
Suburban	0.94	0.94	0.85	0.9	0.72	N/A
Urban	0.96	0.92	0.86	0.83	0.48	N/A

Details of the categories corresponding to the concepts resulted from splitting an existing concept:

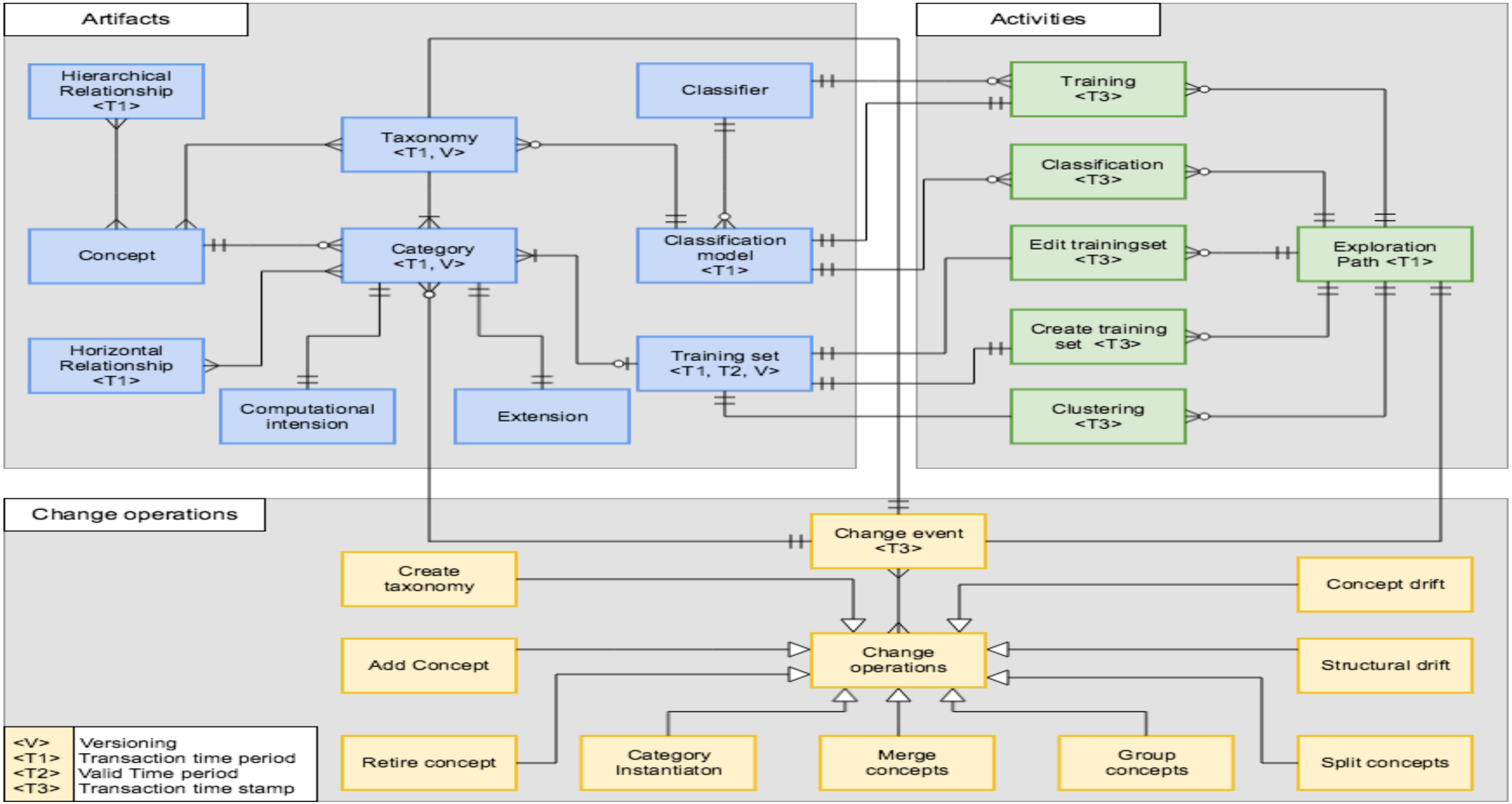
New concept	User accuracy	Producer accuracy	Split from	Extensional containment	J-M distance
Grassland	0.74	0.80	Pasture	0.86	N/A
Open Space	0.82	0.76	Pasture	0.52	N/A
Sea Water	0.87	0.92	Water	0.97, 0.99	N/A
Estuarine Open Water	0.74	0.60	Water	0.86, 0.84	N/A
Scrub	0.56	0.63	Woody Vegetation	0.98, 0.97	N/A
Indigenous Forest	0.76	0.75	Woody Vegetation	0.92, 0.92	N/A
Mangrove	0.98	0.96	Woody Vegetation	0.49, 0.63	N/A

Grouped concepts:

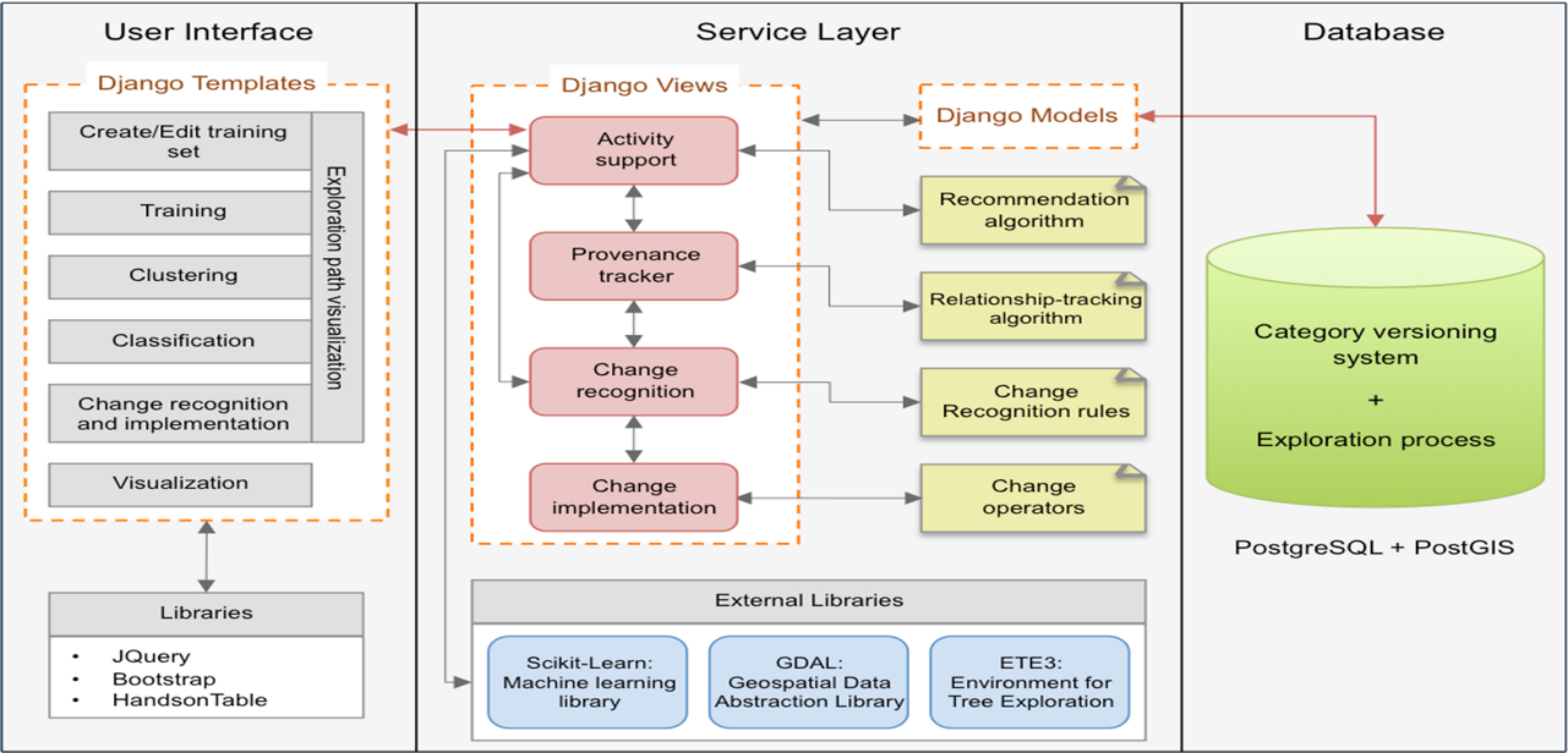
Parent concept	Concepts that are grouped
Water Bodies	Estuarine Open Water, Inland Water, Sea Water
Forest	Indigenous Forest, Mangrove
Vegetation	Forest, Scrub, Grassland
Built-up Area	Suburban, Urban
Artificial Surface	Built-up Area, Open Space

Do you wish to implement these changes in AdvoCate: Yes No

Logical design



Implementation



Conclusions

- I believe it is possible to engineer systems that contain both the methods for doing science AND a meta-model of the science process, so we can explicitly see how these two worlds connect
- This bridges the gap between the process and products of science – revealing the dynamic and evolving aspects of knowledge
- It also connects and synchronizes all of the research artefacts through the process of evolution

End

Questions, comments?

Sources of ontological confusion

(Gahegan & Brodaric, 2014)

Ontologic

Situated

Conceptual: geoscientists are using different concepts and categories; mapping of new areas, or scientific evolution, often requires existing concepts to be revised or supplanted in the field

Theoretic: geoscientists may use different theories with the same evidence and categories.

Inferential: geoscientists may use different reasoning mechanisms.

Intentional: different purposes or goals, including choice of conceptual or geographic scale and level of detail of observation, may naturally lead to diverse results.

Evidential: geoscientists are considering different data.

Model-based: given the same evidence, concepts, theories, and reasoning techniques open systems such as the Earth's may still lead to the generation of diverse valid -process models

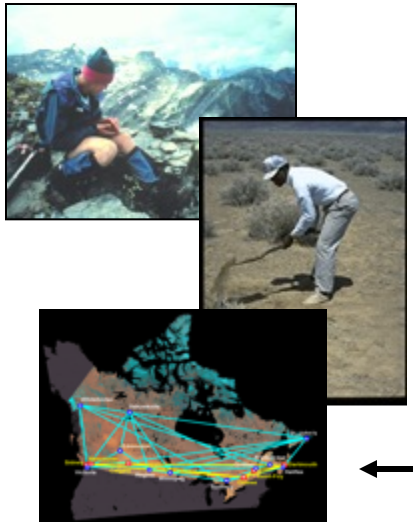
Methodological: geoscientists may use different methods and instruments or perform different actions leading to diverse models.

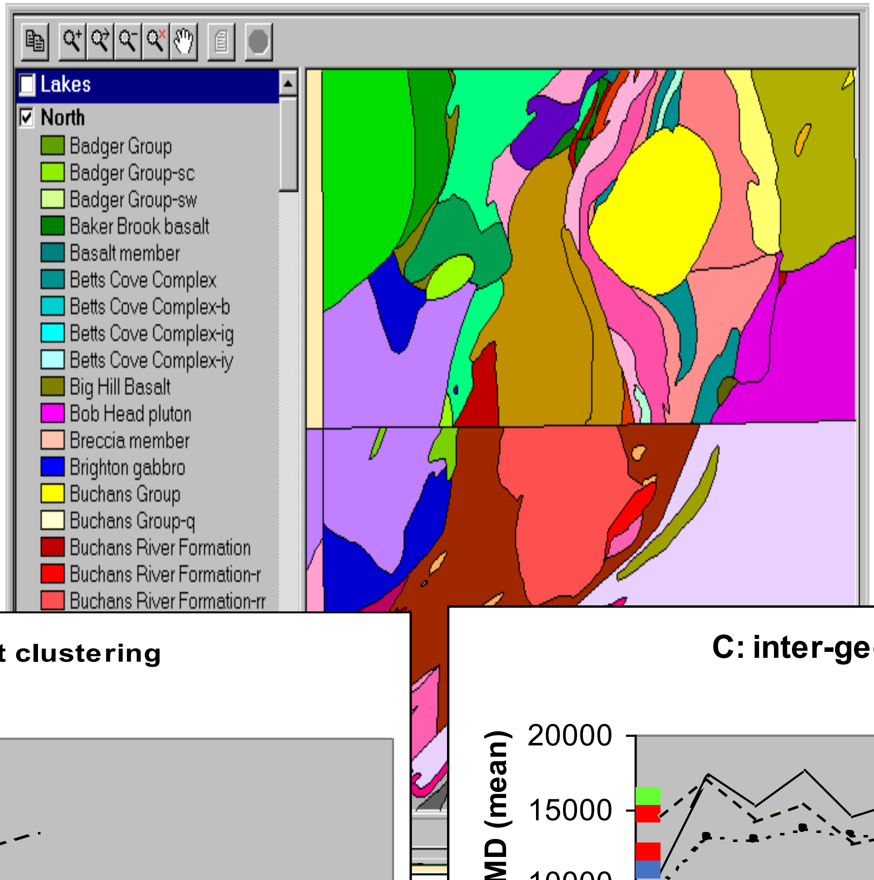
Tacit: geoscientists' implicit understanding of the region, developed in concordance with unconscious predispositions, may differ and lead to model variability.

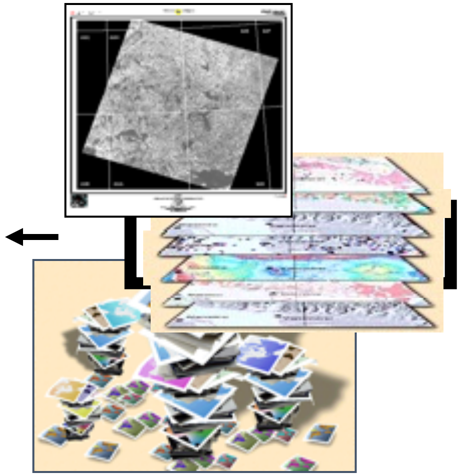
Social: knowledge transfer between geoscientists may vary according to the degree of scientific interaction as facilitated or impeded by political, cultural, and institutional or other structures.

Historical: geoscientists might still develop diverse models due to the order of presentation of evidence and the sequence of decisions made at each stage of observation and reasoning

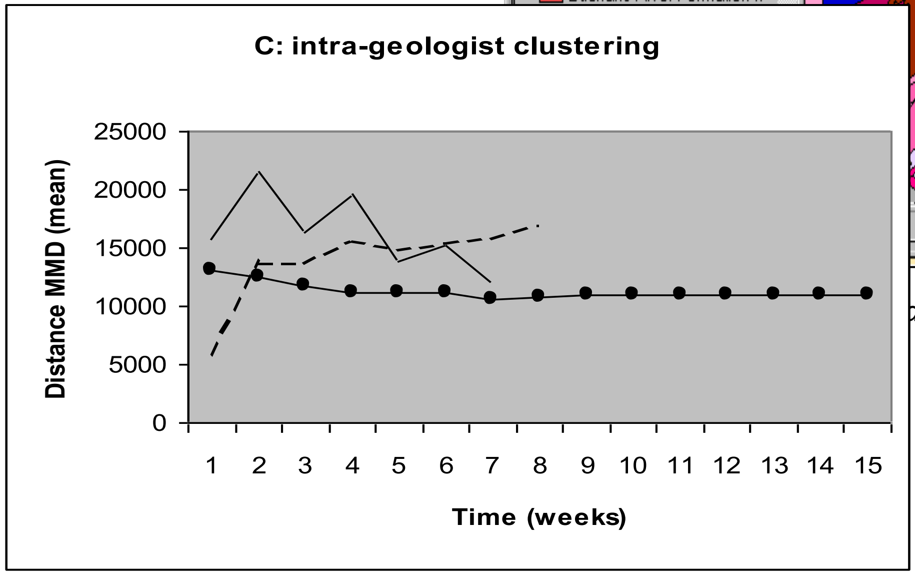
Another MOTIVATING EXAMPLE: Map construction and semantic conflict

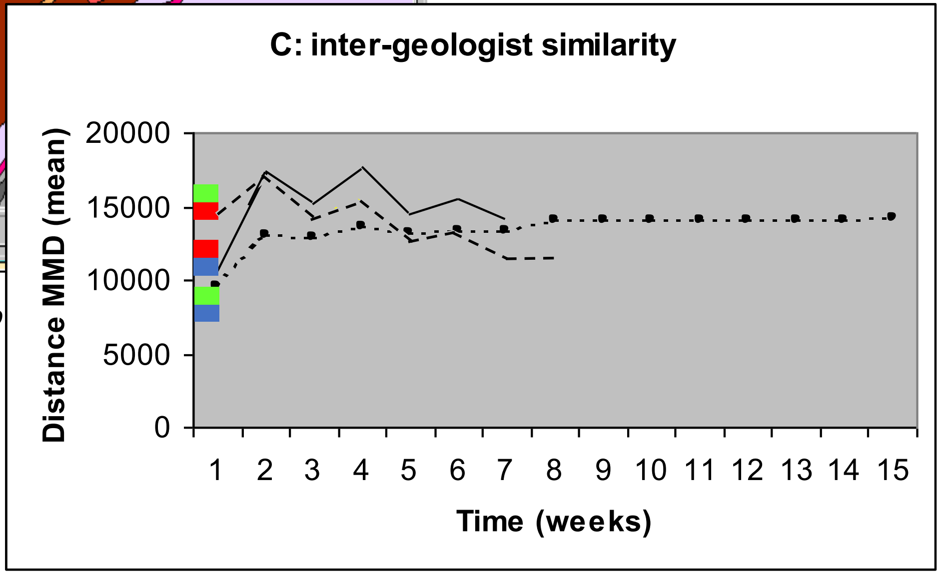






Gahegan & Brodaric, 2002, Brodaric & Gahegan 2006

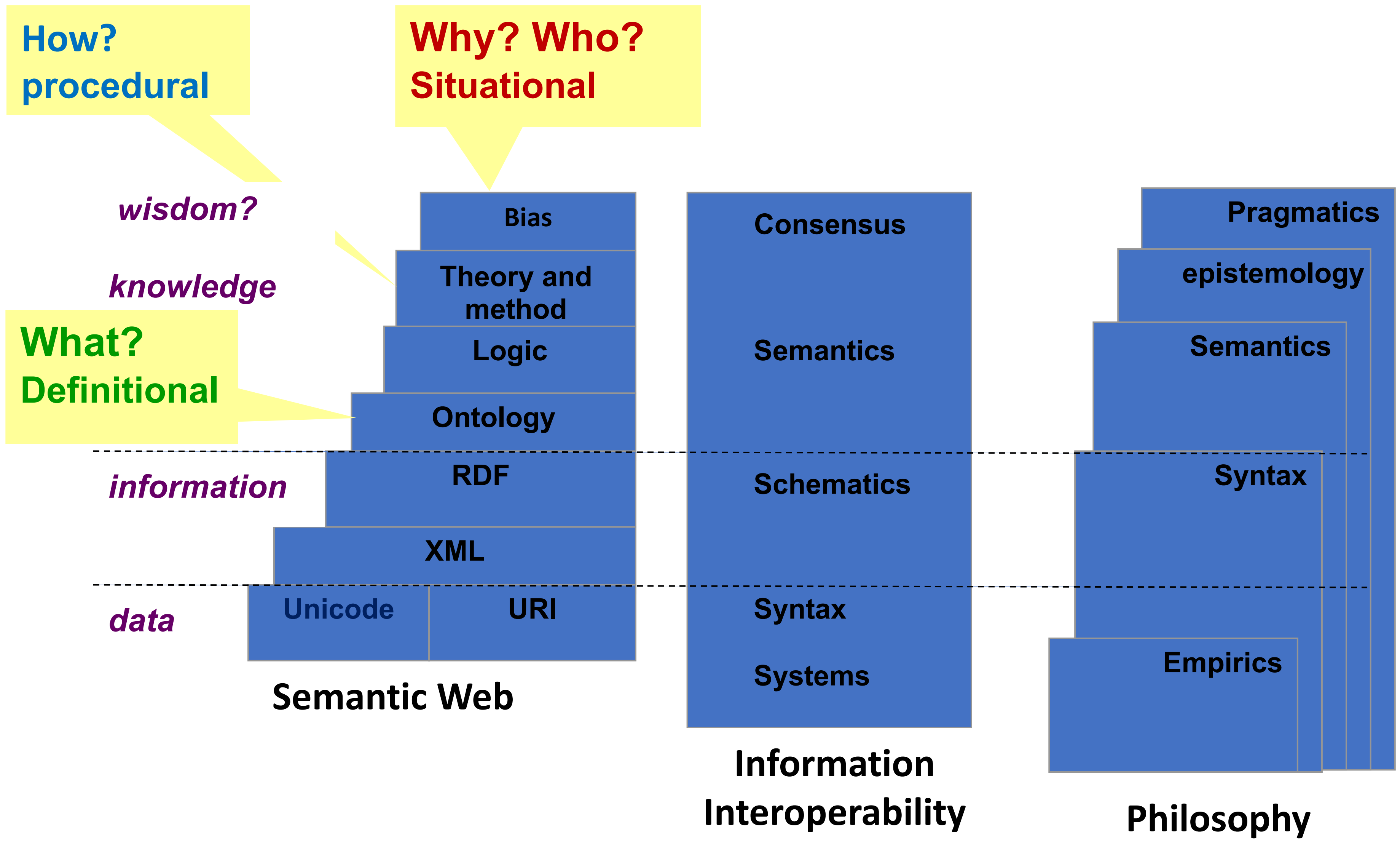




Aside: are we relying too much on ontology as our 'carrier of meaning'?

Ontology tells us what is known, but *epistemology* considers how it is known, how it came to be, and why it came to be the way it is (and not some other way); *pragmatics* addresses how it is understood, who understands it…

Levels of Meaning in systems



Knowledge: Ontology, Epistemology, Pragmatics

- Classically, Ontology describes what we know, or what is true, via description logics
- Epistemology describes how we know something is true, via methodology, research paradigms
- Pragmatics describes the process of interpretation, how and why humans construct and communicate meaning. It is experiential.