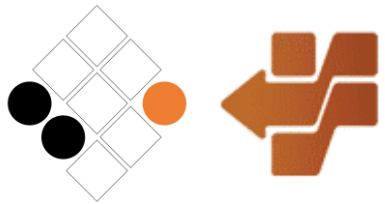


Investigation of hypothesis-driven approaches to managing scientific experiments in data intensive domains

Dmitry Kovalev
Institute of Informatics Problems
Russian Academy of Sciences
31 March 2015

Outline

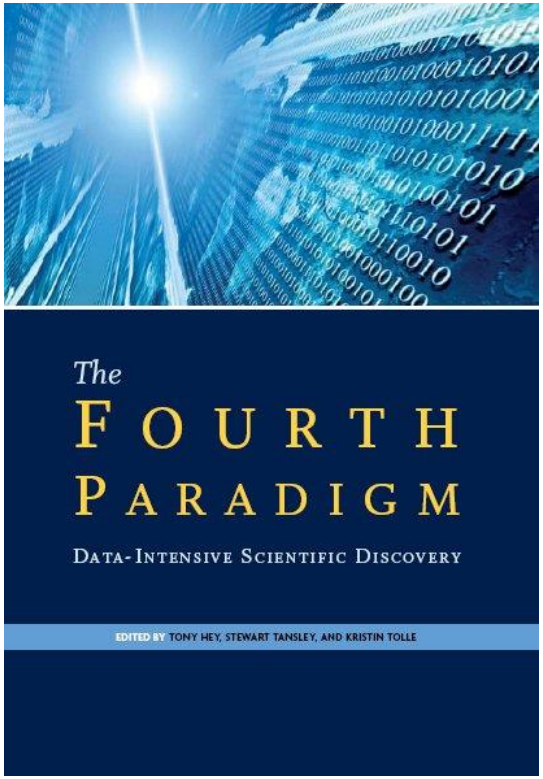
- Data Intensive Domains: Short Survey
- Hypothesis-driven experiment organization: examples
- Hypotheses-Models-Data, Hypotheses Lattice

Research areas

- Heterogeneous information resources specification, interoperation and integration in the interests of their compositional re-use for different applications.
- Currently our group is focused mostly on methods and tools for heterogeneous information resources integration applying subject mediation methodology.
- Generalized methods and infrastructures for data analytics and management in data intensive domains

Data Intensive Domains

- The emergence of Data-Intensive Sciences (the 4th paradigm of science)
- A complete data collection on any complex object (e.g., Earth, the Universe, or the Human Brain) encodes the knowledge possible to be mined and analyzed
- It is used to call such domains as X-informatics (X = astro , bio, geo, neuro, ...)

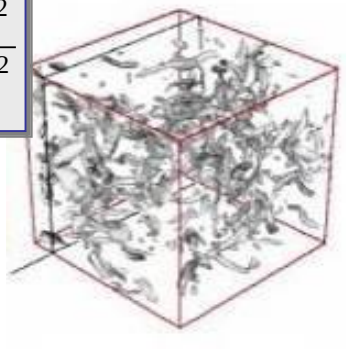


Science Paradigms

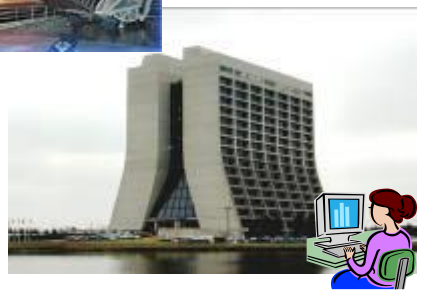
Jim Gray, eScience Talk at NRC-CSTB meeting, 2007

- Thousand years ago:
  science was **empirical**
    describing natural phenomena
- Last few hundred years:
  theoretical branch
    using models, generalizations
- Last few decades:
  a **computational** branch
    simulating complex phenomena
- Today:
  data exploration (eScience)

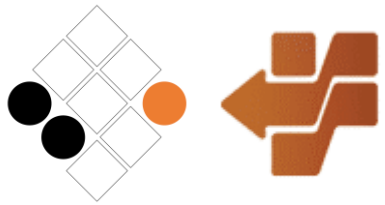
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



unify theory, experiment, and simulation
- Data captured by instruments
  Or generated by simulator
- Processed by software
- Information/Knowledge stored in computer
- Scientist analyzes database / files
  using data management and statistics



Data Intensive Domains



We have to do better at producing tools to support the whole research cycle – from data capture and data curation to data analysis and data visualization.

— Jim Gray, 2007

New tools are needed to bring humans into the data-analysis loop at all stages, recognizing that knowledge is often subjective and context-dependent and that some aspects of human intelligence will not be replaced anytime soon by machines.

— Frontiers in Massive Data Analysis, 2013

Characteristics of DID

Science is increasingly dependent on data as the core source for discovery:

- scientific instruments,
- sensors,
- simulations,
- Web or social nets

The basic objective of Data-Intensive Domains (DID) is to infer knowledge from the integrated data organized in networked infrastructures such as warehouses, grids, clouds.

Characterizing and Exposing the Big Data Hype: 3 V's or ?

http://bit.ly/1hH6sB9

- If the only distinguishing characteristic was that we have lots of data, we would call it **"Lots of Data"** (or a ***Tonnabytes!***)
- Big Data characteristics: **the 3+n V's** =

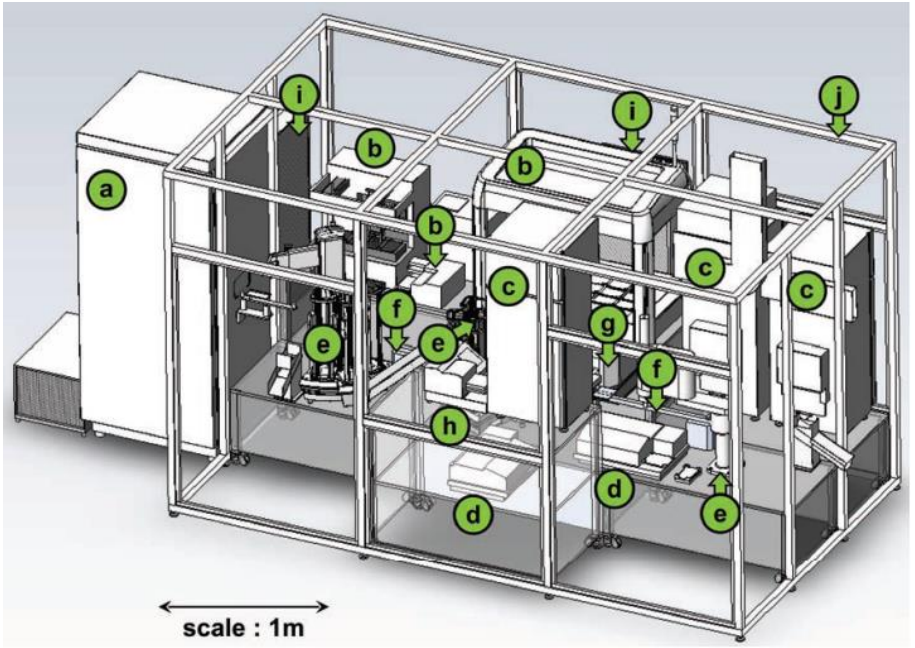
1. **Volume** *(lots of data = "Tonnabytes")*
2. **Variety** *(complexity, curse of dimensionality, many formats)*
3. **Velocity** *(high rate of data and information flow, real-time, incoming!)*
4. **Veracity** *(necessary & sufficient data to test many hypotheses)*
5. **Validity** *(data quality, governance, master data management)*
6. **Value** *(= the all-important V!)*
7. **Variability** *(dynamic, evolving, spatiotemporal data, time series)*
8. **Venue** *(distributed, heterogeneous, multiple platforms/owners)*
9. **Vocabulary** *(ontologies, semantics, schema, data models,...)*
10. **Vagueness** *(confusion over the meaning of Big Data, tools, methods,...)*

Data Intensive Domains:
The Automation of Systems Biology

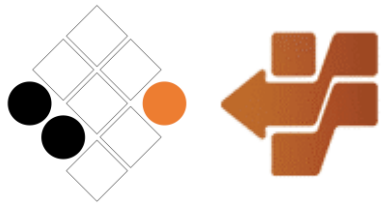
This is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence to execute cycles of scientific experiment

Adam formulated
and tested 20 hypotheses
concerning genes encoding
13 orphan enzymes . . . 12
hypotheses with no previous
evidence were confirmed.

Abductive Logic Programming
(PrologICA) was used to specify
the domain

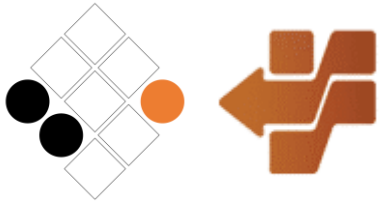


Data Intensive Domains:
The Automation of Systems Biology



- Complex research statements, where basic (atomic) statements like predicate(entity$_i$; entity$_j$) are combined by logical operators $\wedge$, $\vee$, $\neg$, $\rightarrow$, $\leftrightarrow$
- “If all genes with lactase activity are deleted from a yeast strain and if this strain is grown in medium with lactose as the sole carbon source, then the phenotype will be no growth.”

Data Intensive Domains:
The Automation of Systems Biology



$$
\begin{gathered}
(((\forall gene, \forall yeast\_strain, \forall x \mid \\
has - function(gene;\ lactase\ activity) \wedge \\
has - part(yeast\_strain;\ gene) \wedge \\
\text{is} - a(process, deletion) \wedge \\
has - paticipant(gene; deletion) \wedge \\
has - output(deletion; yeast\_strain) \wedge \\
has - part(growth\_medium; lactose) \wedge \\
has - function(lactose; carbon\_source) \wedge \\
has - part(growth\_medium; x) \wedge \\
\text{is}\ - a(phenotype; no\_growth) \wedge \\
\neg has - function(x; carbon_{source})))
\end{gathered}
$$

In combination with a logical model of metabolism these statements would enable deduction of the fact:

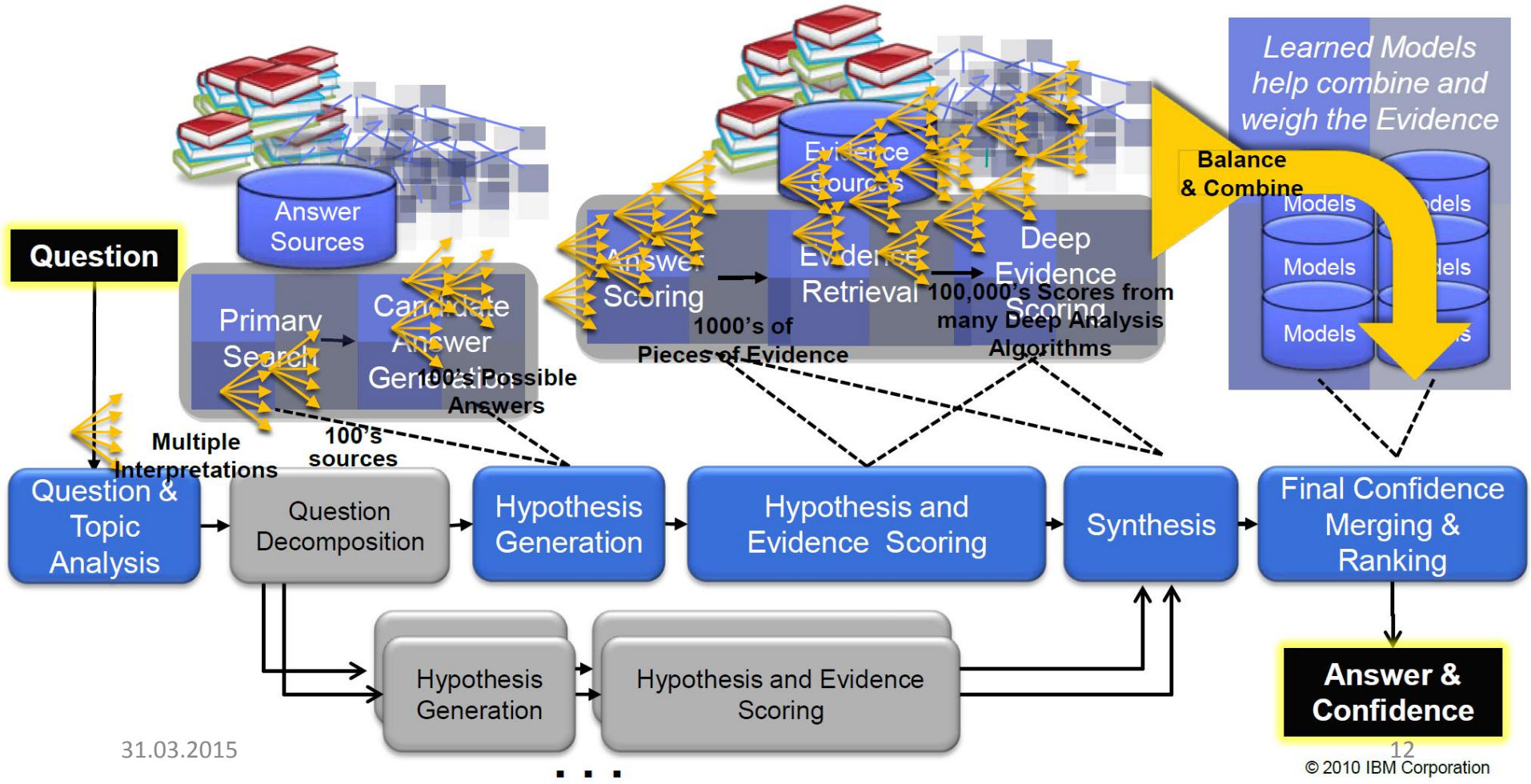
$$\rightarrow has - quality(yeast\_strain; no\_growth)$$

David Ferrucci. Building Watson Beyond Jeopardy, 2010

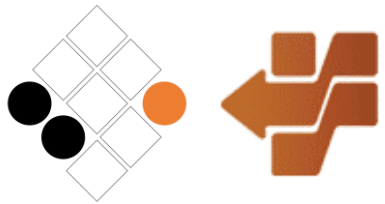
IBM

DeepQA: The architecture underlying Inside Watson

*Generates many hypotheses, **collects a wide range of evidence** and balances the combined confidences of **over 100 different analytics that analyze the evidence form different dimensions***



Data Intensive Domains:
IBM WATSON DeepQA



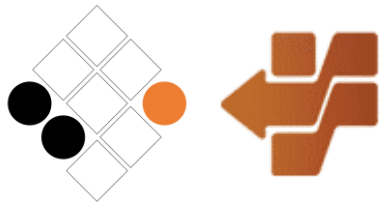
Hypotheses - answer-sized snippets from the search results, which the system has to prove correct with some degree of confidence.

Q: “He was presidentially pardoned on September 8, 1974”

Hypotheses:

- “Nixon,”,
- “Ford pardoned Nixon on Sept. 8, 1974.”

Data Intensive Domains:
IBM WATSON DeepQA



Reasoning: subsumption and disjointness in type taxonomies, geospatial, and temporal reasoning.

Q: “In 1594 he took a job as a tax collector in Andalusia,”

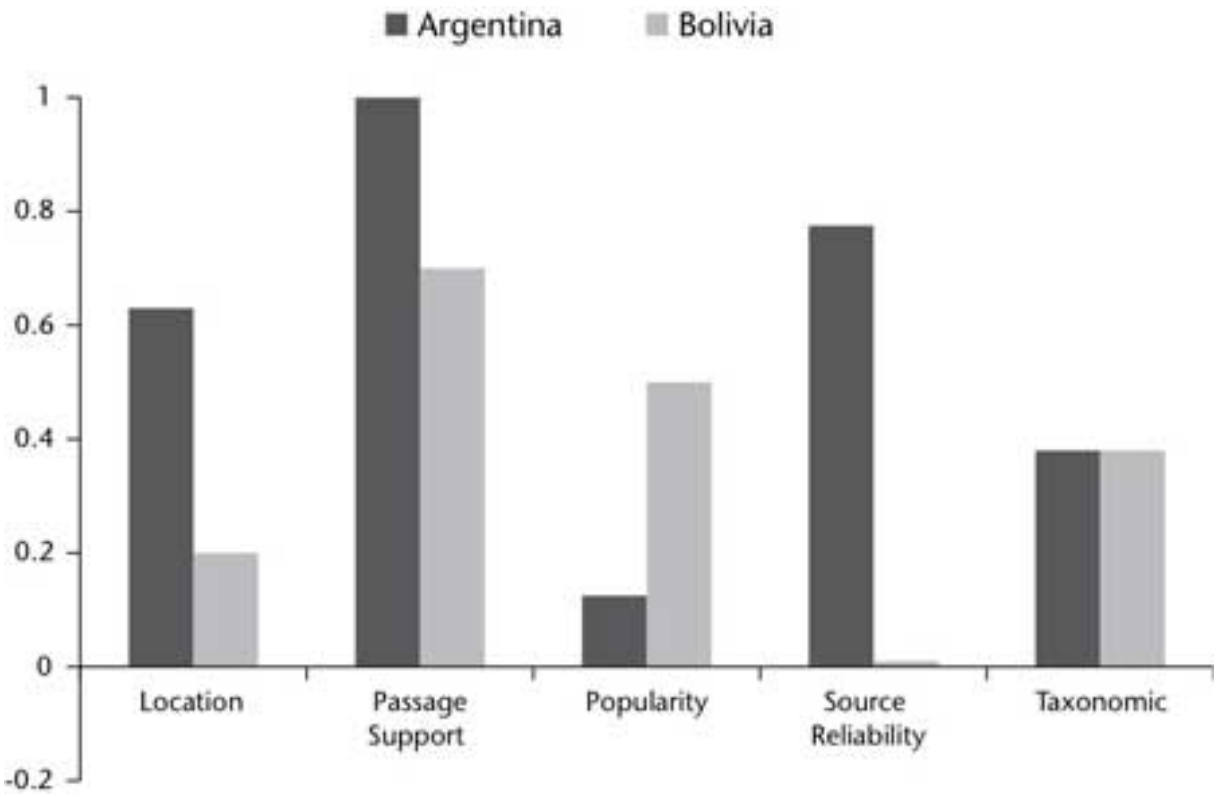
H: “Thoreau” and “Cervantes.”

Temporal reasoning is used to rule out Thoreau as he was not alive in 1594, having been born in 1817, whereas Cervantes, the correct answer, was born in 1547 and died in 1616.

Data Intensive Domains:
IBM WATSON DeepQA

Q: Chile shares its longest land border with this country.

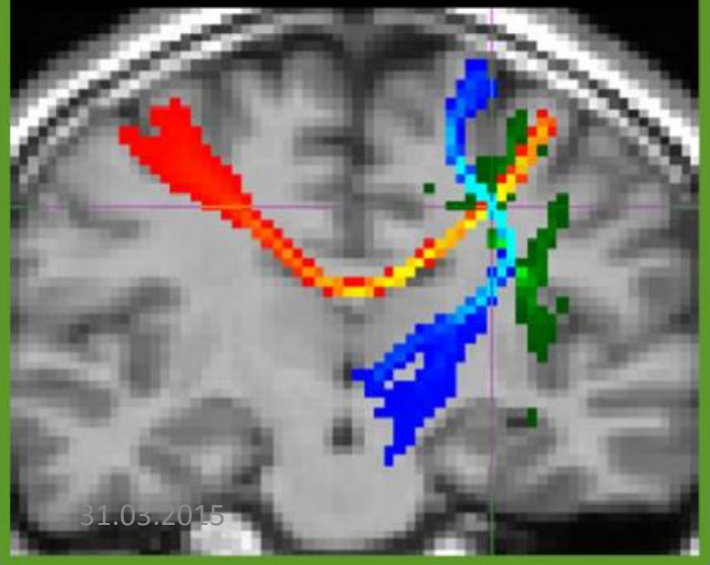
H: Argentina and Bolivia



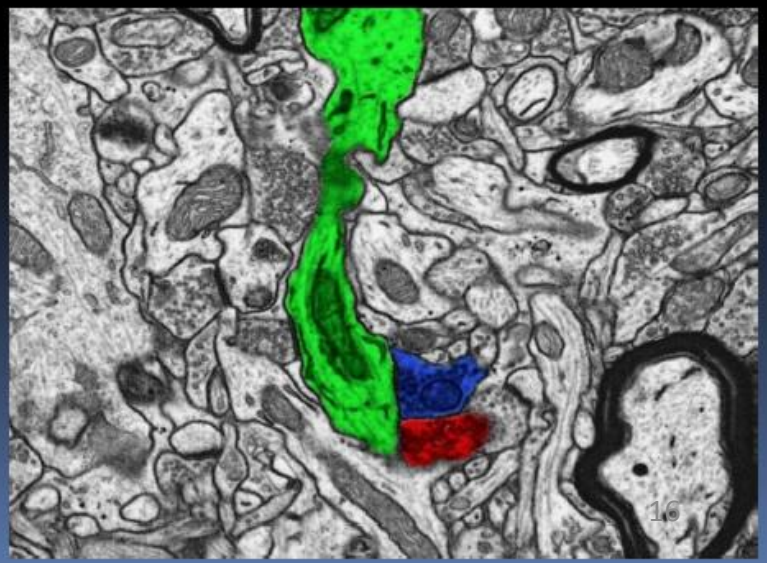
The Human Connectome Project:

- An NIH-funded effort to chart a comprehensive map of neuronal connections and its variability in healthy adults (on the macro-scale)

Macro-connectome
(whole-brain, long-distance)

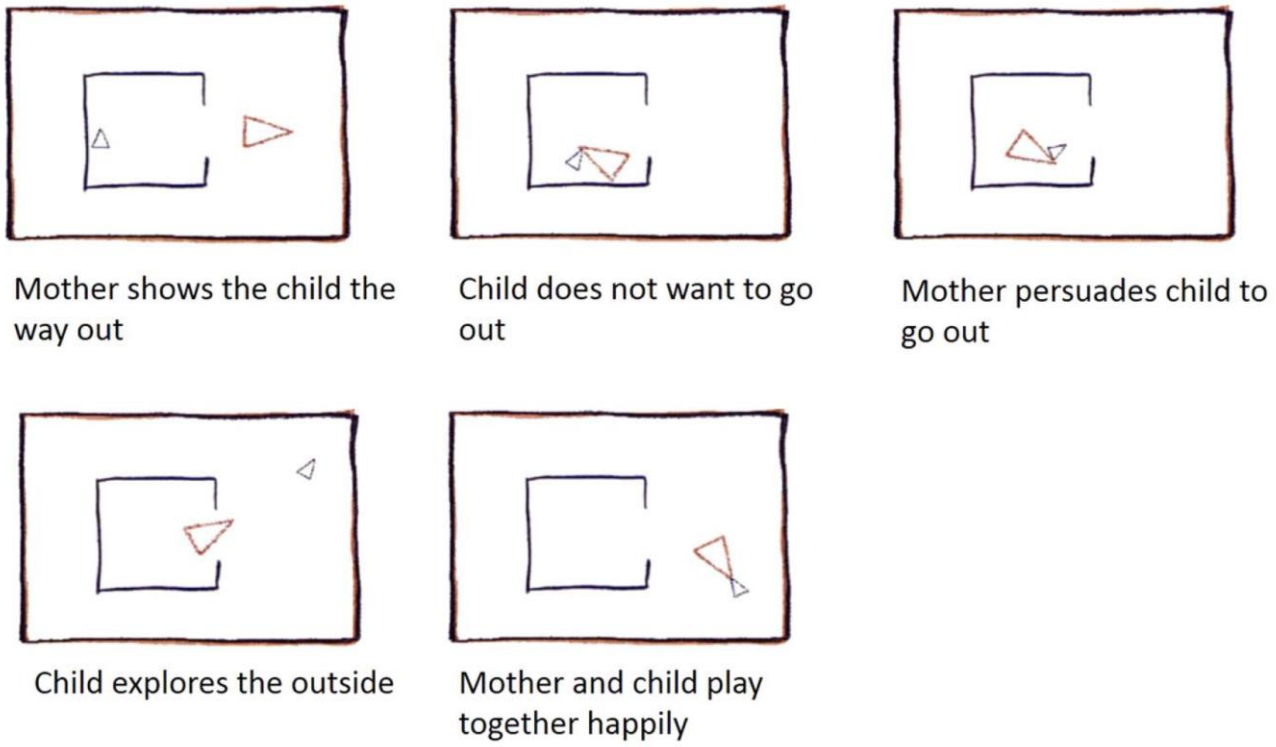


Micro-connectome
(synapses, neurons)



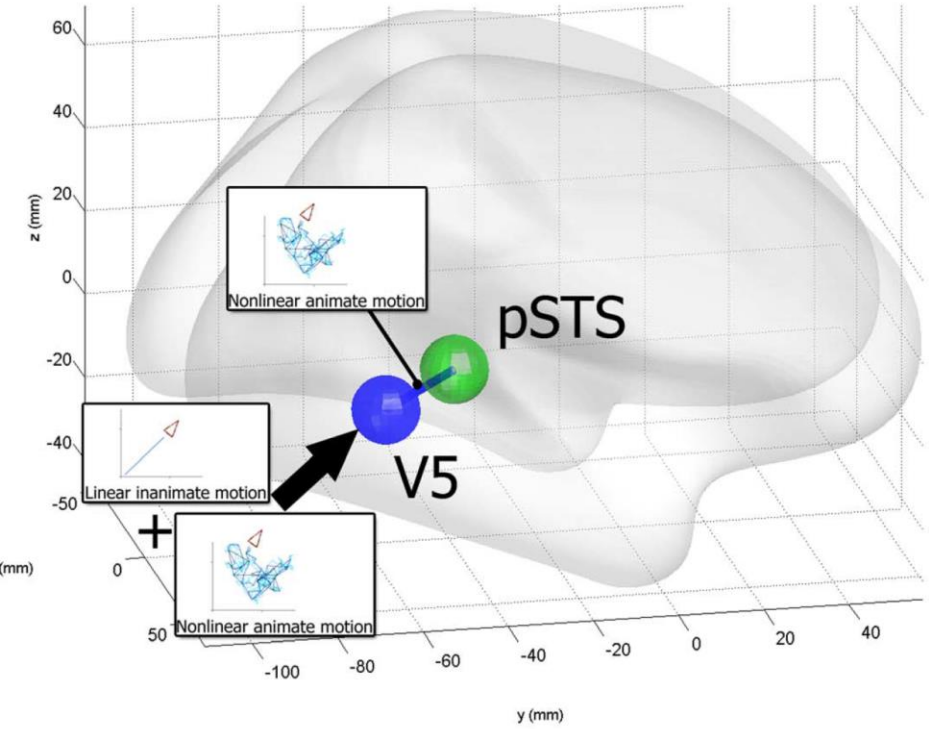
Structural	Unprocessed	70 MB
	Preprocessed	1 GB
Resting State fMRI (each of 2 runs)	Unprocessed	2 GB
	Preprocessed	2.5 GB
	FIX (compact, both runs)	1.3 GB
	FIX_extended	3.5 GB
Task fMRI (avg per Task) (all smoothing levels) (all 7 Tasks) (all smoothing levels)	Unprocessed	490 MB
	Preprocessed	600 MB
	Analyzed	700 MB
	Unprocessed	3.4 GB
	Preprocessed	4.2 GB
	Analyzed	2.6 GB
Diffusion	Unprocessed	2.6 GB
	Preprocessed	790 MB
MEG (10 datasets on 14 subjects)	Unprocessed	106 GB
	Preprocessed	157 GB
Group-Average on U100 and R440	Additionally Processed	200 MB
Group-Average “dense” connectomes (each of 2)	Additionally Processed	33 GB
Total (per Subject, MR only)	Unprocessed	10 GB
	Preprocessed	20 GB
	Both	30 GB
	Both+Analyzed	32 GB
Total (10 Subjects, MR only)	Unprocessed	124 GB
	Preprocessed	199 GB
	Both+Analyzed	348 GB
Total (100 Subjects, MR only)	Unprocessed	1.2 TB
	Preprocessed	2.0 TB
	Both+Analyzed	3.5 TB
Total (All imaging datasets from 507 Subjects)	Unprocessed	6.3 TB
	Preprocessed	10.1 TB
	Both+Analyzed	17.6 TB

Hypotheses in Connectomics



- Animate-Inanimate Motion videos were shown to 500+ humans
- Task-fMRI data was collected and published in Human Connectome Project (now gigabytes of data, soon terabytes and petabytes)

Hypotheses in Connectomics

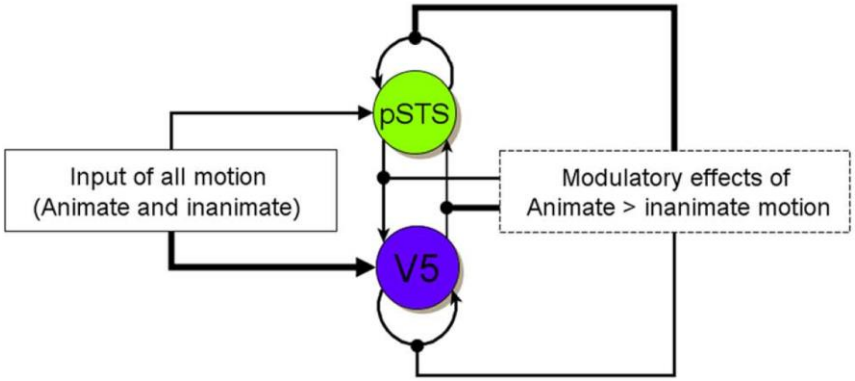


Hypothesis:

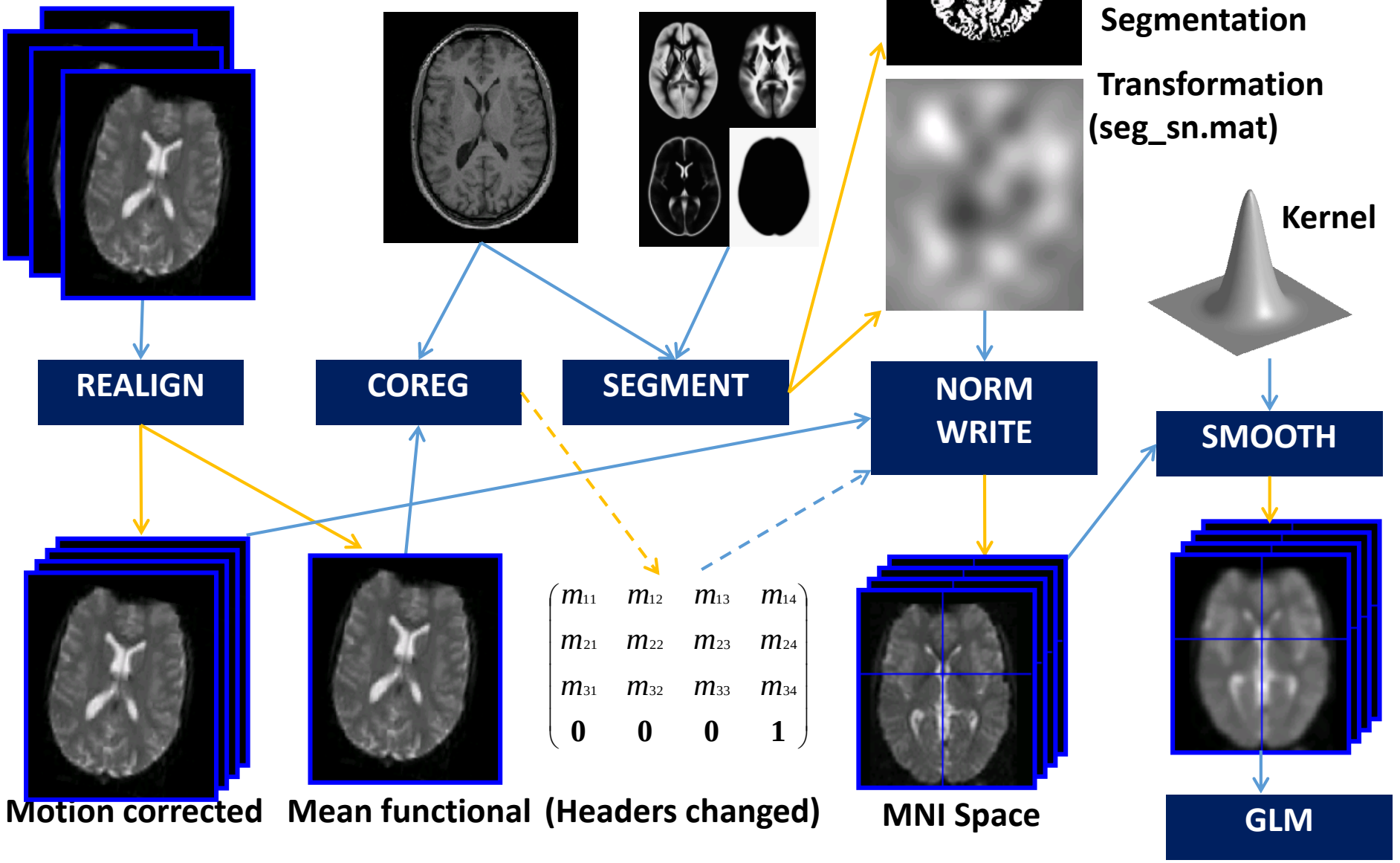
… Predictions about animate motion – relative to inanimate motion – should increase signal passing from lower level sensory area MT+/V5, which is responsive to all motion, to higher-order posterior superior temporal sulcus (pSTS), which is selectively activated by animate motion.

Hypotheses in Connectomics

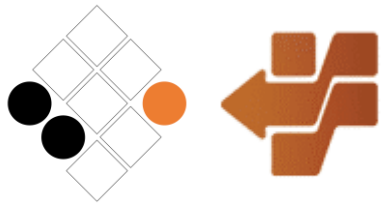
- General Linear Models and Dynamic Causal Modeling were used to test the connectivity
- Matlab code published on the Web to make research reproducible
  - Works well enough on 100s gigabytes of data
  - Unfortunately, not efficient for large scale computations…



Spatial Preprocessing



Why Besancon Galaxy Model is interesting to us?



- 30+ years of continuous development
- Explicit and implicit hypotheses are presented in the model
- Model is well-known and used by many researchers
- Model ingredients develop with time:
  - Hypotheses combinations and interrelations,
  - New model parameters,
  - New data sources used as new data arrives,
  - Different statistical tools and methods used
  - Possible underlying infrastructure changes
  - Self-consistency
  - Users are allowed to manipulate some parameters (soon)

Why Besancon Galaxy Model is interesting to us?

- Example of interrelated hypotheses:
  - the age distribution, the density laws and the potential are linked with the age-velocity dispersion via the Boltzmann equation. It needs to be consistent.
- Model evolution:
  - The model ingredients are a priori selected according to the previous knowledge (common values in the literature for example). Then if it happens that there is a discrepancy between model predictions and data, we have to identify the parameter(s) that are bad and should be adjusted.

Ingredients	Old model	New default models	
		model A	model B
IMF	Haywood-Robin	Haywood-Robin (*A*)	Kroupa-Haywood v6 (*B*)
SFR	constant	a decreasing $\exp(-0.12\tau)$ Aumer & Binney (2009)	a decreasing $\exp(-0.12\tau)$ Aumer & Binney (2009)
evolutionary tracks	see Table 2	package E2 Table 2	package E2 Table 2
age-metallicity relation	Twarog (1980)	Haywood (2006)	Haywood (2006)
atmosphere models	BaSeL 2.2	BaSeL 3.1	BaSeL 3.1
binarity	no	yes: from Arenou (2011)	yes: from Arenou (2011)
thin disc age	10 Gyr	10 Gyr	10 Gyr
thick disc parameters	x_l = 400 pc, h_z = 800 pc density = 0.0083 $*/pc^3$	x_l = 400 pc, h_z = 800 pc density = 0.0083 $*/pc^3$	x_l = 400 pc, h_z = 800 pc density = 0.0083 $*/pc^3$
extinction model	Drimmel & Spergel (2001)	Drimmel & Spergel (2001) + Marshall et al. (2006)	Drimmel & Spergel (2001) + Marshall et al. (2006)
ISM local density	Robin et al. (2003)	Binney & Tremaine (2008)	Binney & Tremaine (2008)
local stellar mass density	Wielen (1974)	Wielen (1974)	Jahreiß & Wielen (1997)
age-velocity relation	Gómez et al. (1997)	Gómez et al. (1997)	Gómez et al. (1997)
warp	Reylé et al. (2009)	Reylé et al. (2009)	Reylé et al. (2009)
scale length	young disc h_R = 5000.0 pc old disc h_R = 2400.0 pc	young disc h_R = 5000.0 pc old disc h_R = 2400.0 pc	young disc h_R = 5000.0 pc old disc h_R = 2400.0 pc

Research hypothesis

Has description

Star Formation Rate, Ψ(t) represents the total mass of stars born per unit time per unit mass of Galaxy ($f(m) = m^{-\alpha}$)

Competing hypotheses

SFR

subtype

subtype

Constant SFR

Decreasing SFR

subtype

subtype

$$SFR(\tau) = \langle SFR \rangle \frac{(t + t_0)t_n^3}{(t^2 + t_1^2)^2},$$

SFR(τ) ∝ exp(γτ)

It is one of the Aumer and Binney (2009) best fits.

Cited in

instance

instance

<SFR> is the mean SFR, t0 = 5.6 Gyr, t1 = 8.2 Gyr and tn = 9.9 Gyr

SFR(τ) ∝ exp(0.25τ)

SFR(τ) ∝ exp(0.12τ)

...

Cited in

such the input function of the model A from Just and Jahreis (2010)

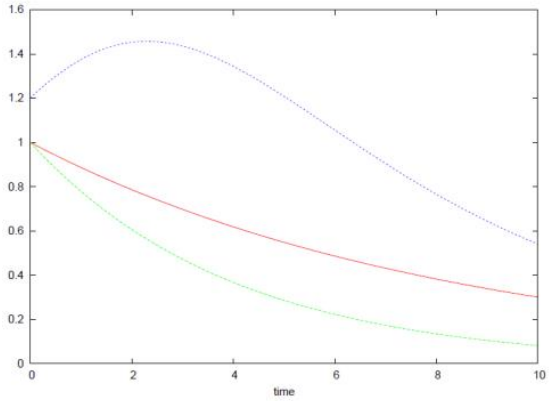


Figure 3.3: Three functions chosen to represent a decreasing SFR as a function of time t. In blue the SFR of the model A of Just and Jahreiß (2010), in red a simple exponential with the γ value 0.12 (proposed by Aumer and Binney (2009)) and in green an exponential with the γ value of 0.25.

Why Besancon Galaxy Model is interesting to us?

However, we want to emphasize that the **γ parameter is correlated with the values of other parameters used in the model and especially with the slopes of IMF and the age of the disc.** This multi-dependency and interplay between different model's ingredients oblige us to always look for the best global fit. In this context one could not give the best solution for a particular variable without correlating it with others

Why BGM is interesting to us?

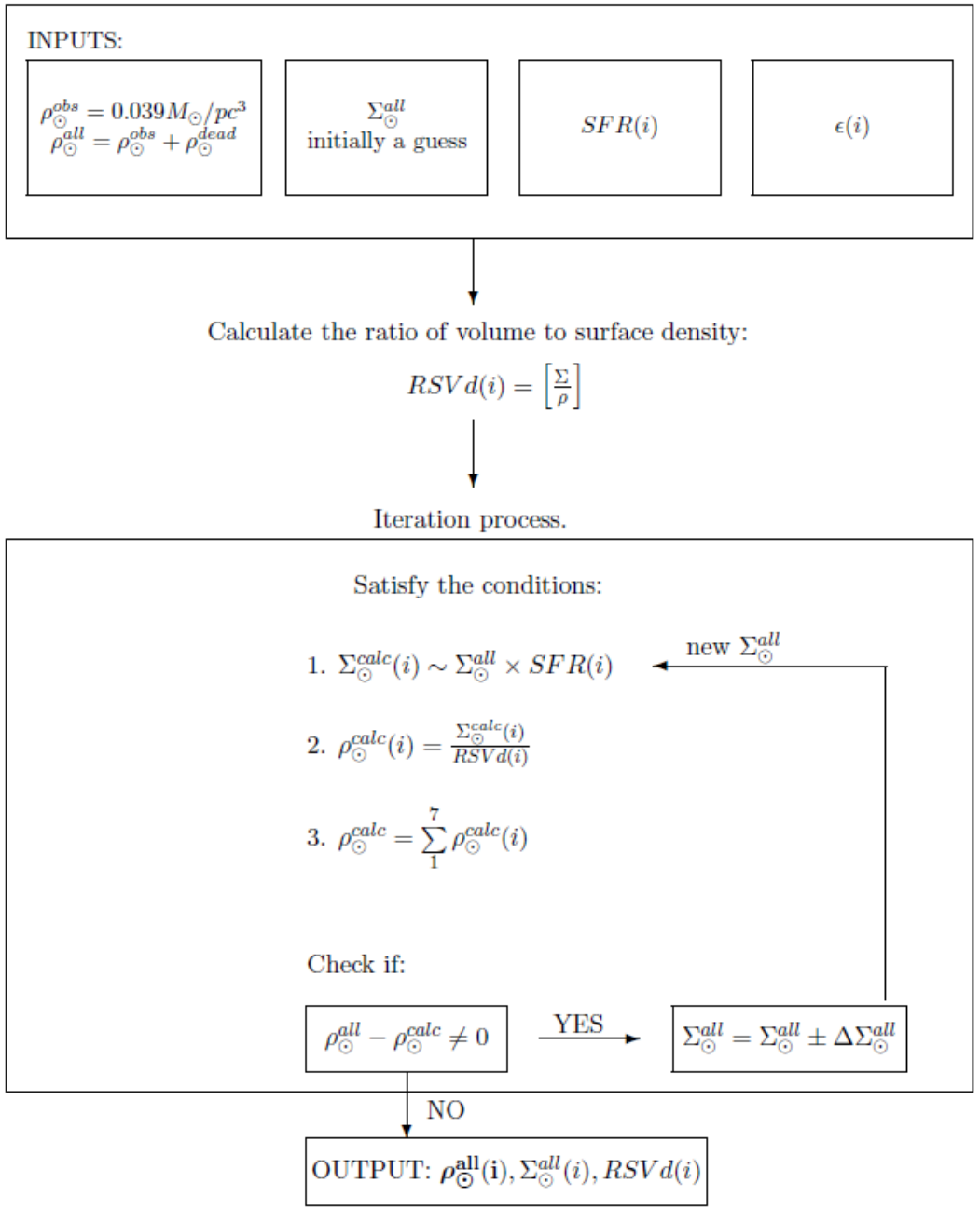
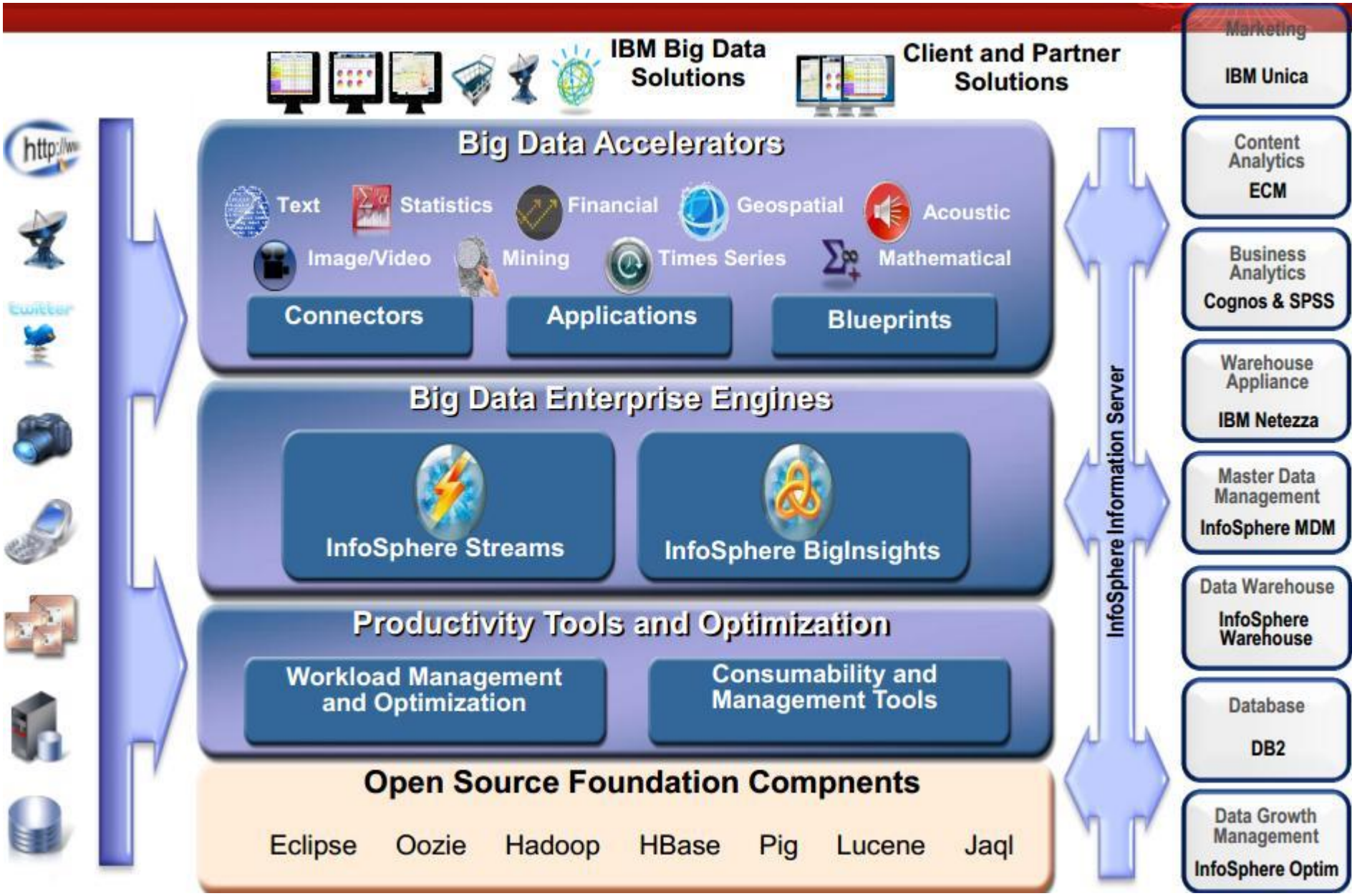
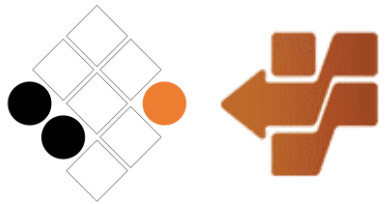


Figure 2.12: The process of local mass normalization in the BGM.

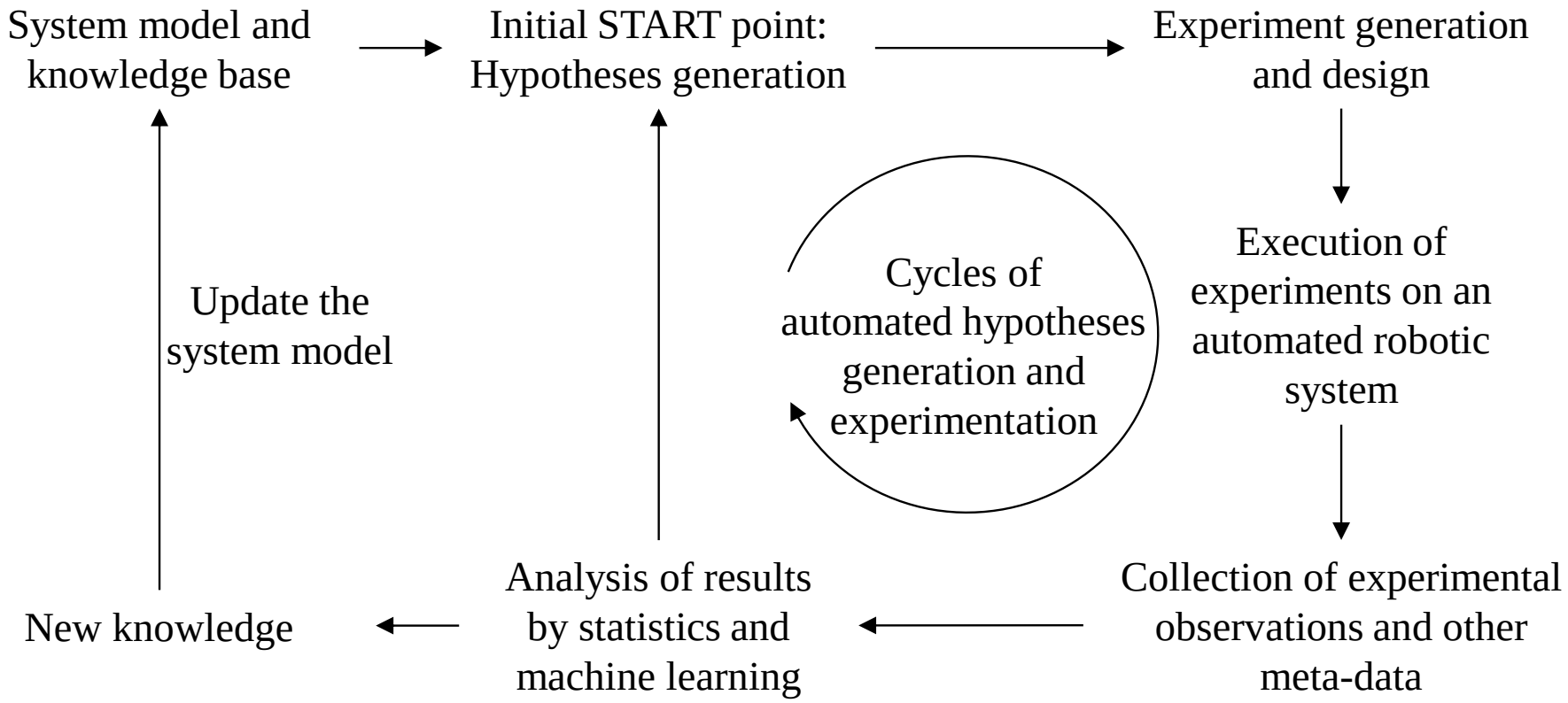

IBM Big Data Solutions
Client and Partner Solutions
Big Data Accelerators
Text
Statistics
Financial
Geospatial
Acoustic
Image/Video
Mining
Times Series
Mathematical
Connectors
Applications
Blueprints
Big Data Enterprise Engines
InfoSphere Streams
InfoSphere BigInsights
Productivity Tools and Optimization
Workload Management and Optimization
Consumability and Management Tools
Open Source Foundation Compnents
Eclipse
Oozie
Hadoop
HBase
Pig
Lucene
Jaql
InfoSphere Information Server
Marketing
IBM Unica
Content Analytics
ECM
Business Analytics
Cognos & SPSS
Warehouse Appliance
IBM Netezza
Master Data Management
InfoSphere MDM
Data Warehouse
InfoSphere Warehouse
Database
DB2
Data Growth Management
InfoSphere Optim

Hypothesis-driven approach generalized

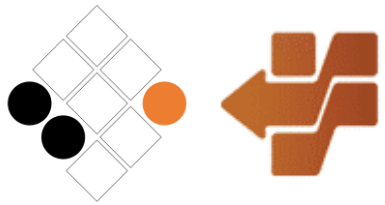


- We are interested in generalizing hypothesis-driven approaches to managing scientific experiment:
  - How to conceptualize application areas?
  - How to encode hypotheses correctly?
  - How to encode the research cycle – from hypothesis development; through experiment and data collection; to interpretation and drawing of conclusions; to communication of results to other scientists; to assimilating, criticizing and synthesizing the communications of colleagues?
  - How to maintain model independence from data resources?

Hypothesis-driven closed loop cycle

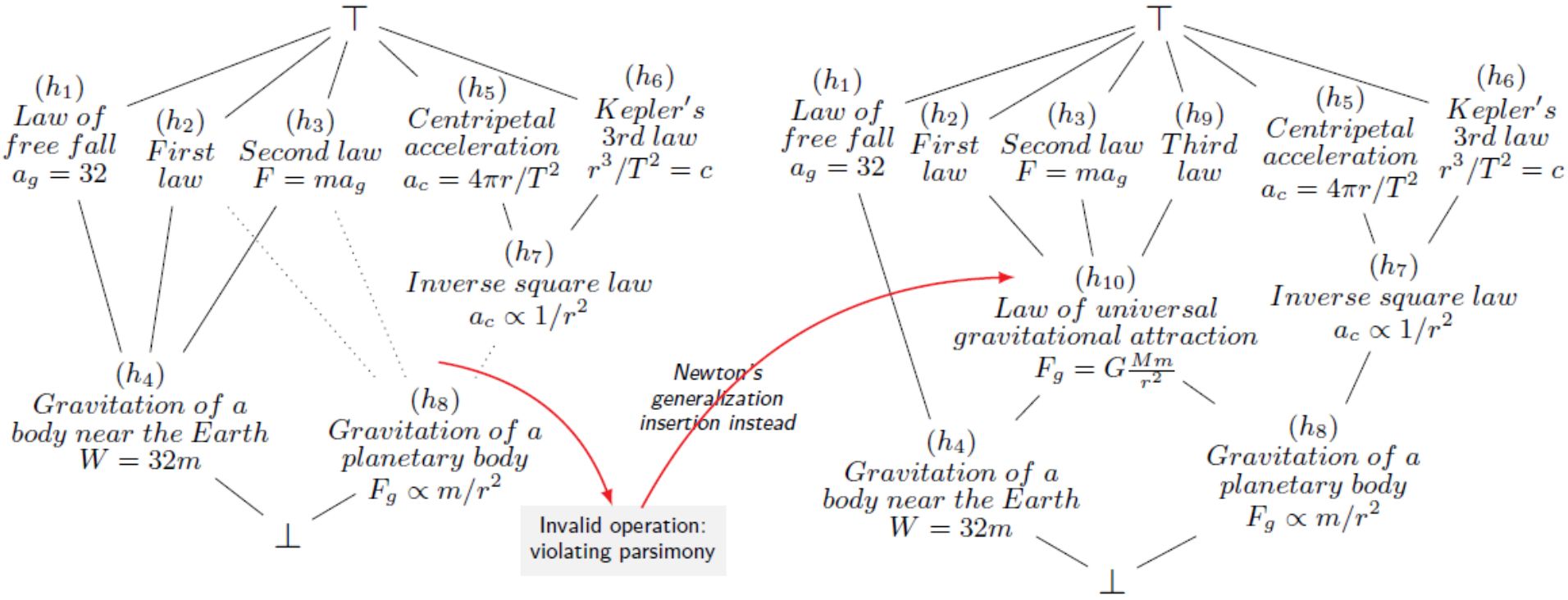


Research Lattices

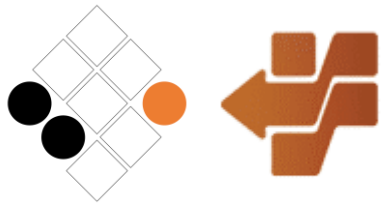


- Dealing with multiple hypotheses at a time
- Possible to find consistent and inconsistent hypotheses
- Add/delete hypotheses
  - consistently keep the partial ordering;
  - automatic placement of hypotheses in the RL
- Querying
  - finding hypotheses based on chosen hypothesis

Research Lattices



Research Lattices: queries

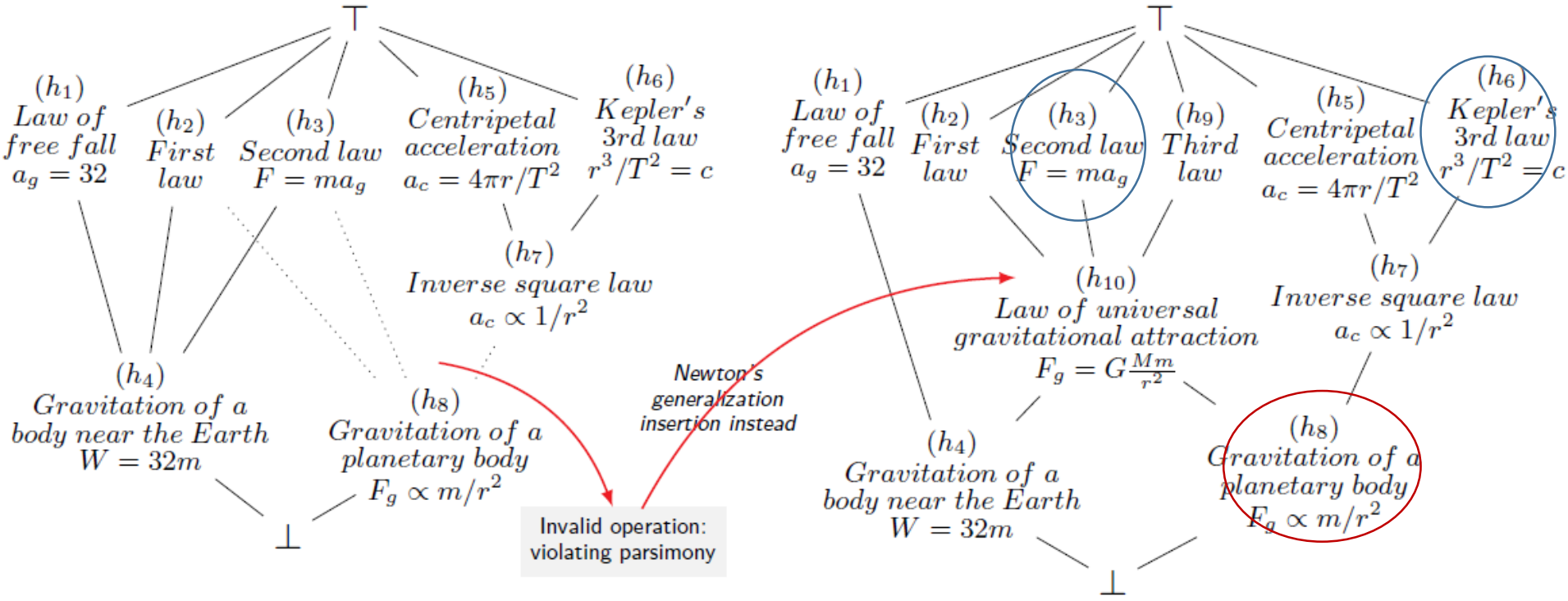


(Q1) Given h_4 and h_8, find their *join* (or strongest weakest hypothesis): $\{\, h_q \in R \mid h_q = h_4 \vee h_8 \,\}$. $[\, h_{10} \,]$.

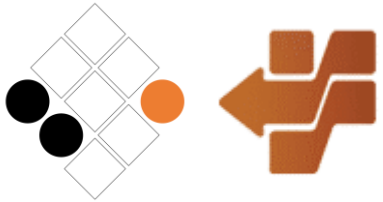
(Q2) Given h_3 and h_6, find their *meet* (or weakest strongest hypothesis): $\{\, h_q \in R \mid h_q = h_3 \wedge h_6 \,\}$. $[\, h_8 \,]$.

(Q3) List all hypotheses that h_{10} is based on or equal to: $\{h \in \uparrow h_{10} \subseteq R\}$. $[\, h_{10}, h_2, h_3, h_9, \top \,]$.

Research Lattices



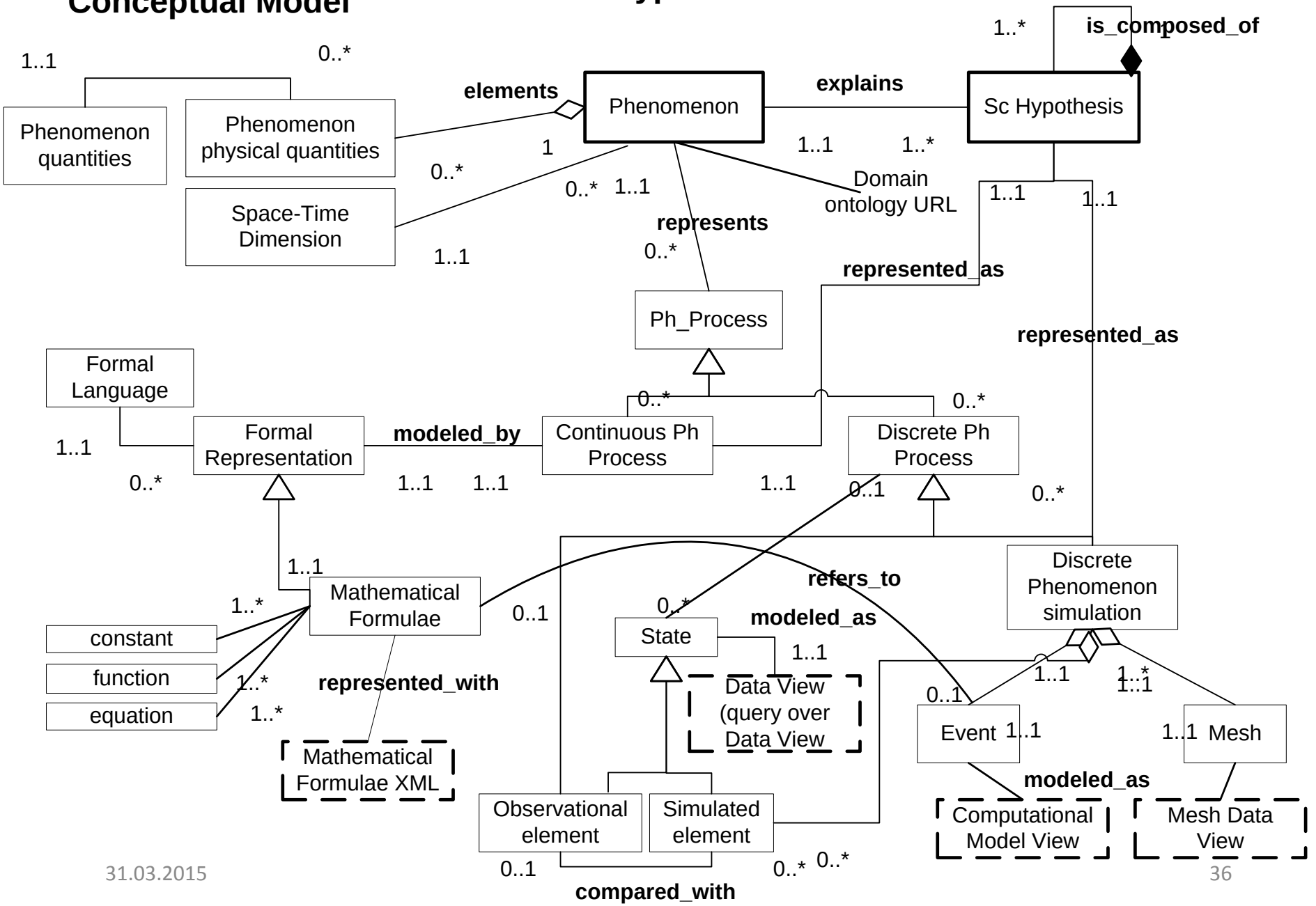
Conceptual Modeling of Data Intensive Domains



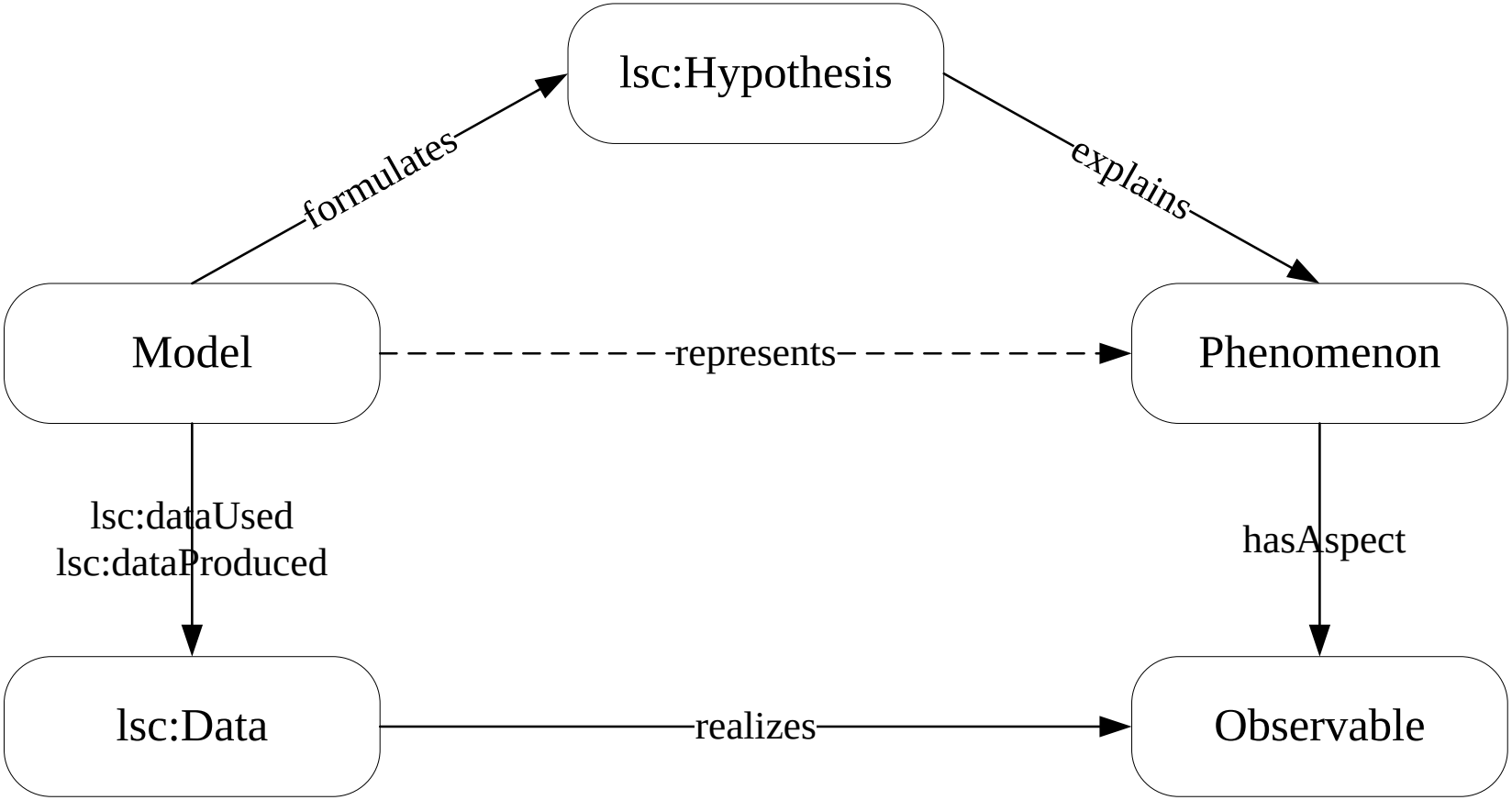
- A *conceptual model* in the field of computer science
  - is also known as a domain model
  - is explicitly chosen to be independent of design or implementation concerns
  - the aim of a conceptual model is to express the meaning of terms and concepts used by domain experts to define the problem
- The proposed approach is aimed to support
  - specifications reusability in various applications
    - over different sets of widely diverse data
  - knowledge semantic integration capability
  - accumulation of reproducible data analysis and problem solving methods and experience in various application domains
  - programming and composition of complex analytical pipelines in an understandable form [Challenges and Opportunities with Big Data (2012)]
    - applying appropriate high-level languages
    - expressing the analytics intended for inferring knowledge from data

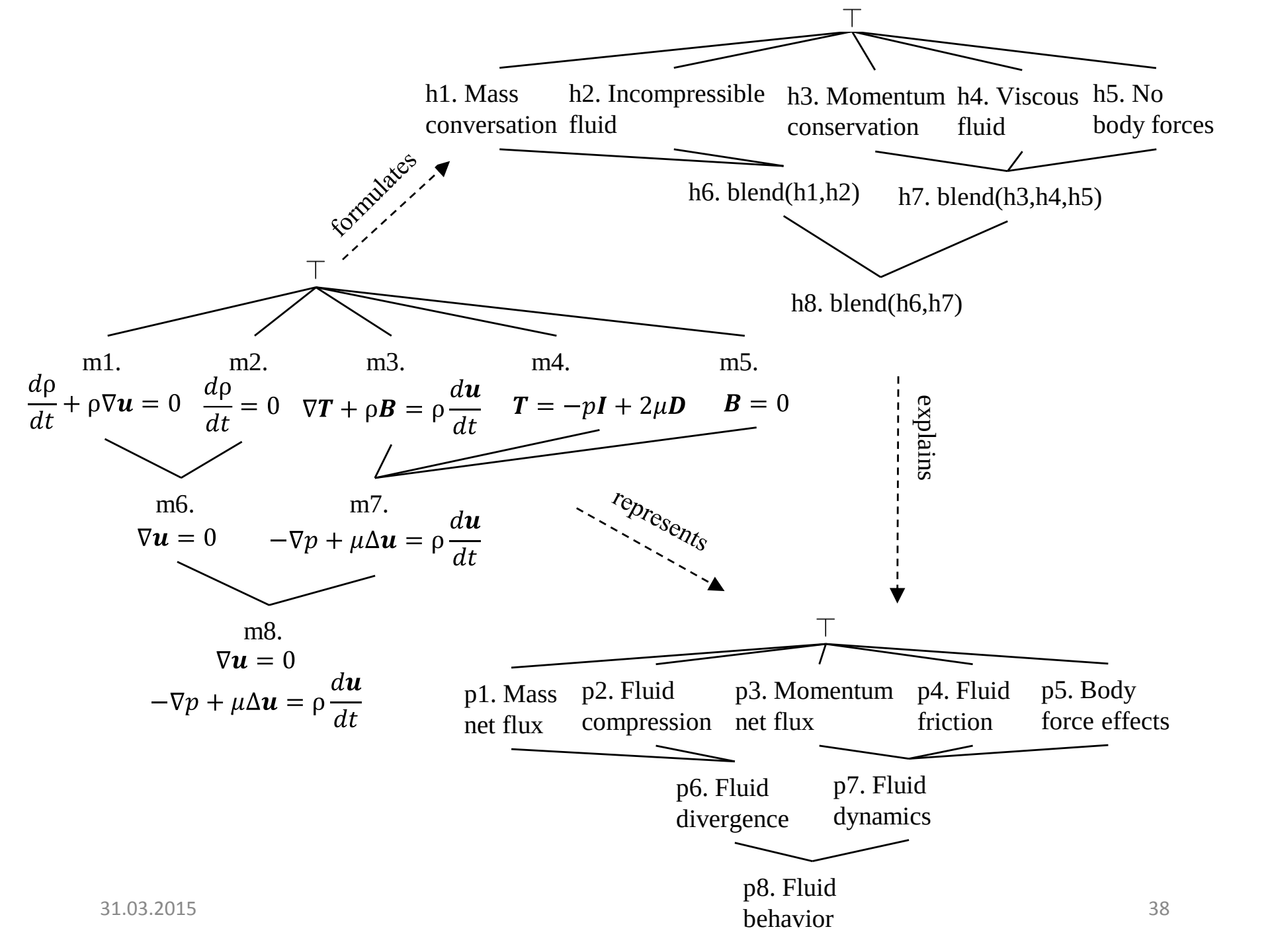
Sc Hypothesis Conceptual Model

Scientific Hypothesis Model

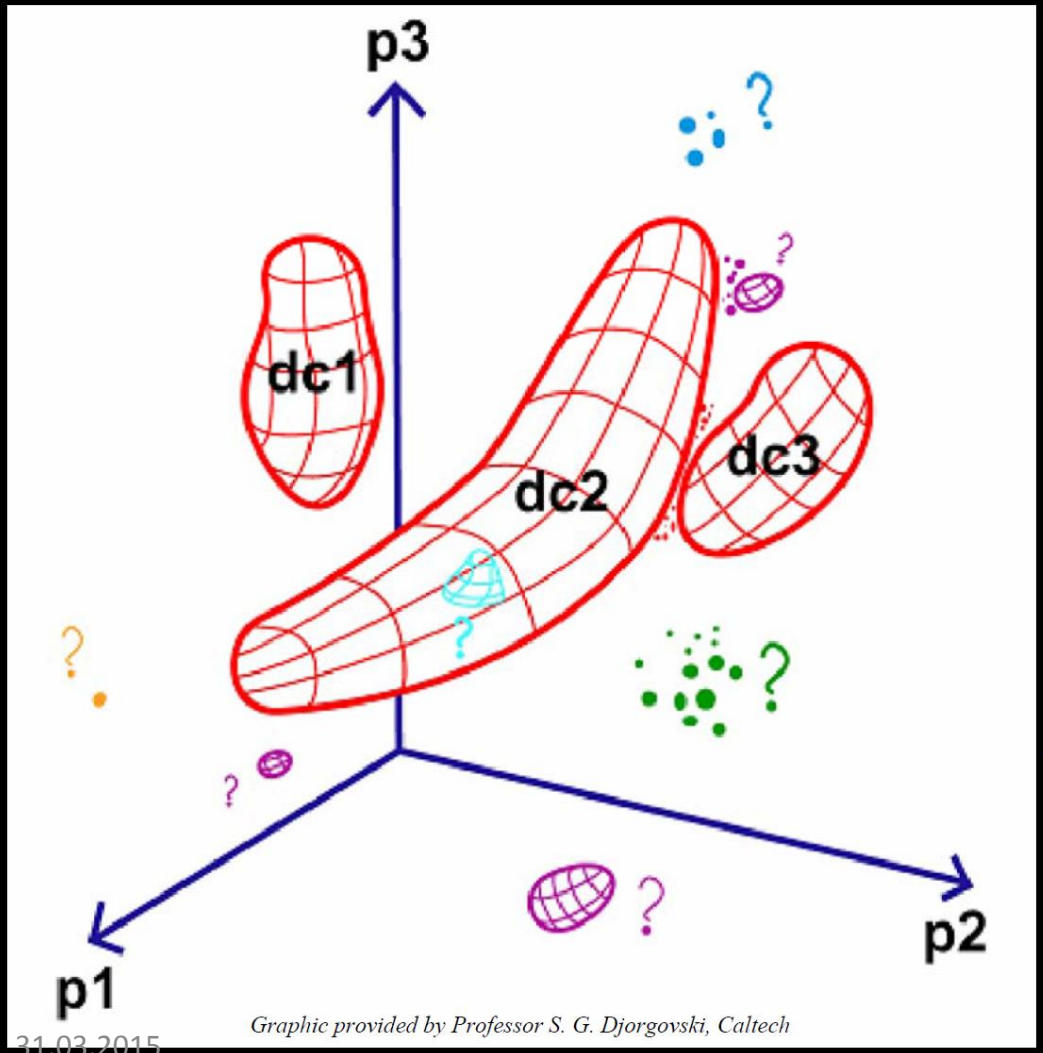


Research Triangle




⊤
h1. Mass conversation
h2. Incompressible fluid
h3. Momentum conservation
h4. Viscous fluid
h5. No body forces
h6. blend(h1,h2)
h7. blend(h3,h4,h5)
h8. blend(h6,h7)
formulates
⊤
m1.
$\frac{d\rho}{dt} + \rho\nabla\boldsymbol{u} = 0$
m2.
$\frac{d\rho}{dt} = 0$
m3.
$\nabla\boldsymbol{T} + \rho\boldsymbol{B} = \rho\frac{d\boldsymbol{u}}{dt}$
m4.
$\boldsymbol{T} = -p\boldsymbol{I} + 2\mu\boldsymbol{D}$
m5.
$\boldsymbol{B} = 0$
explains
m6.
$\nabla\boldsymbol{u} = 0$
m7.
$-\nabla p + \mu\Delta\boldsymbol{u} = \rho\frac{d\boldsymbol{u}}{dt}$
represents
m8.
$\nabla\boldsymbol{u} = 0$
$-\nabla p + \mu\Delta\boldsymbol{u} = \rho\frac{d\boldsymbol{u}}{dt}$
⊤
p1. Mass net flux
p2. Fluid compression
p3. Momentum net flux
p4. Fluid friction
p5. Body force effects
p6. Fluid divergence
p7. Fluid dynamics
p8. Fluid behavior

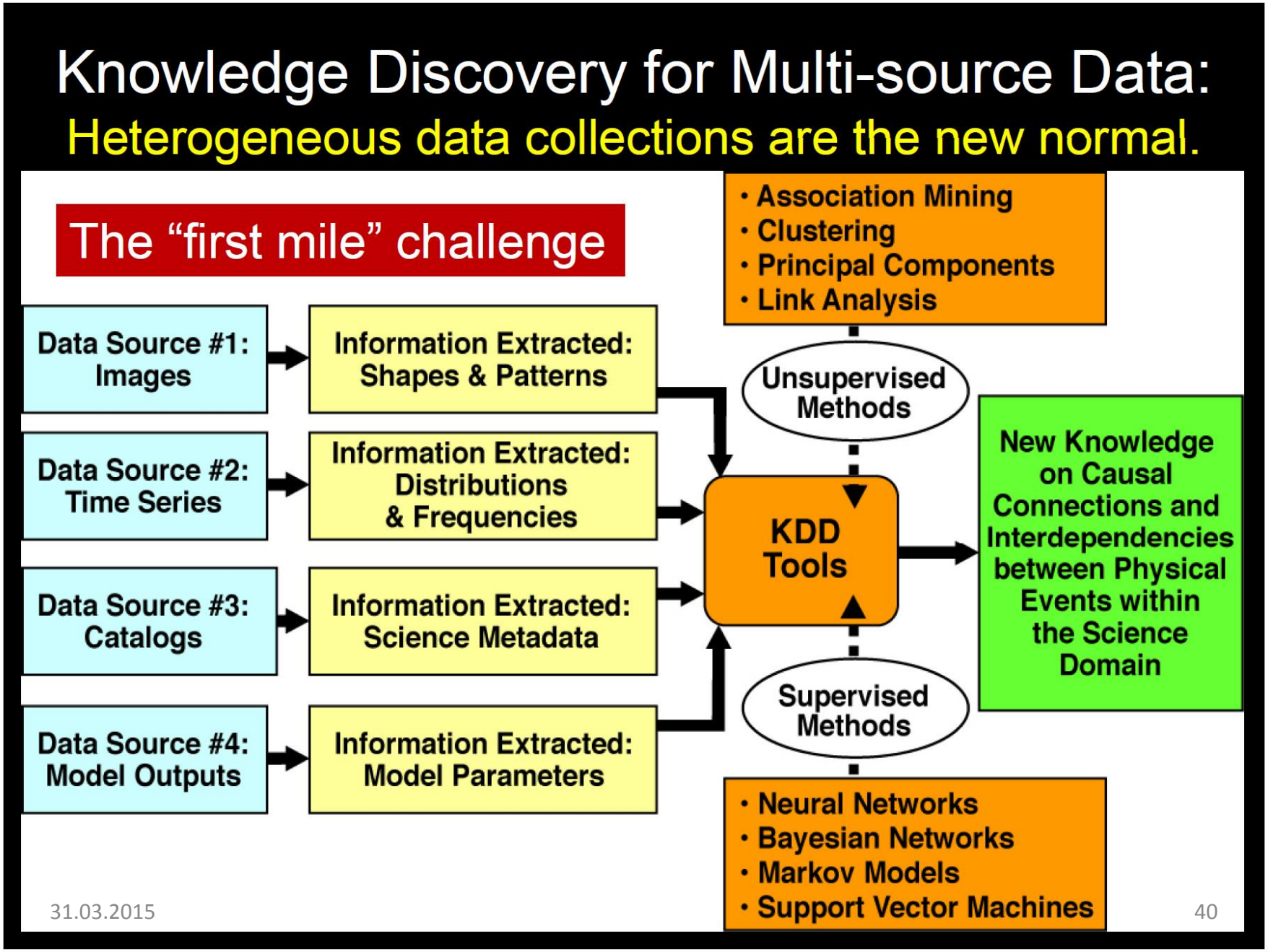
This graph says it all …
3 Steps to Discovery – Data Mining your Big Data



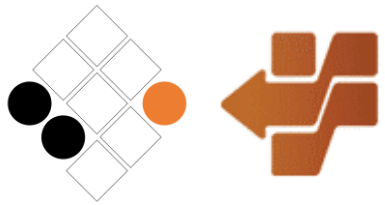
Graphic provided by Professor S. G. Djorgovski, Caltech

- Unsupervised Learning : Cluster Analysis – partition the data items into clusters, without bias, ignoring any initially assigned categories = Class Discovery !
- Supervised Learning : Classification – for each new data item, assign it to a known class (*i.e.*, a known category or cluster) = Predictive Power Discovery !
- Semi-supervised Learning : Outlier/Novelty Detection – identify data items that are outside the bounds of the known classes of behavior = Surprise Discovery !

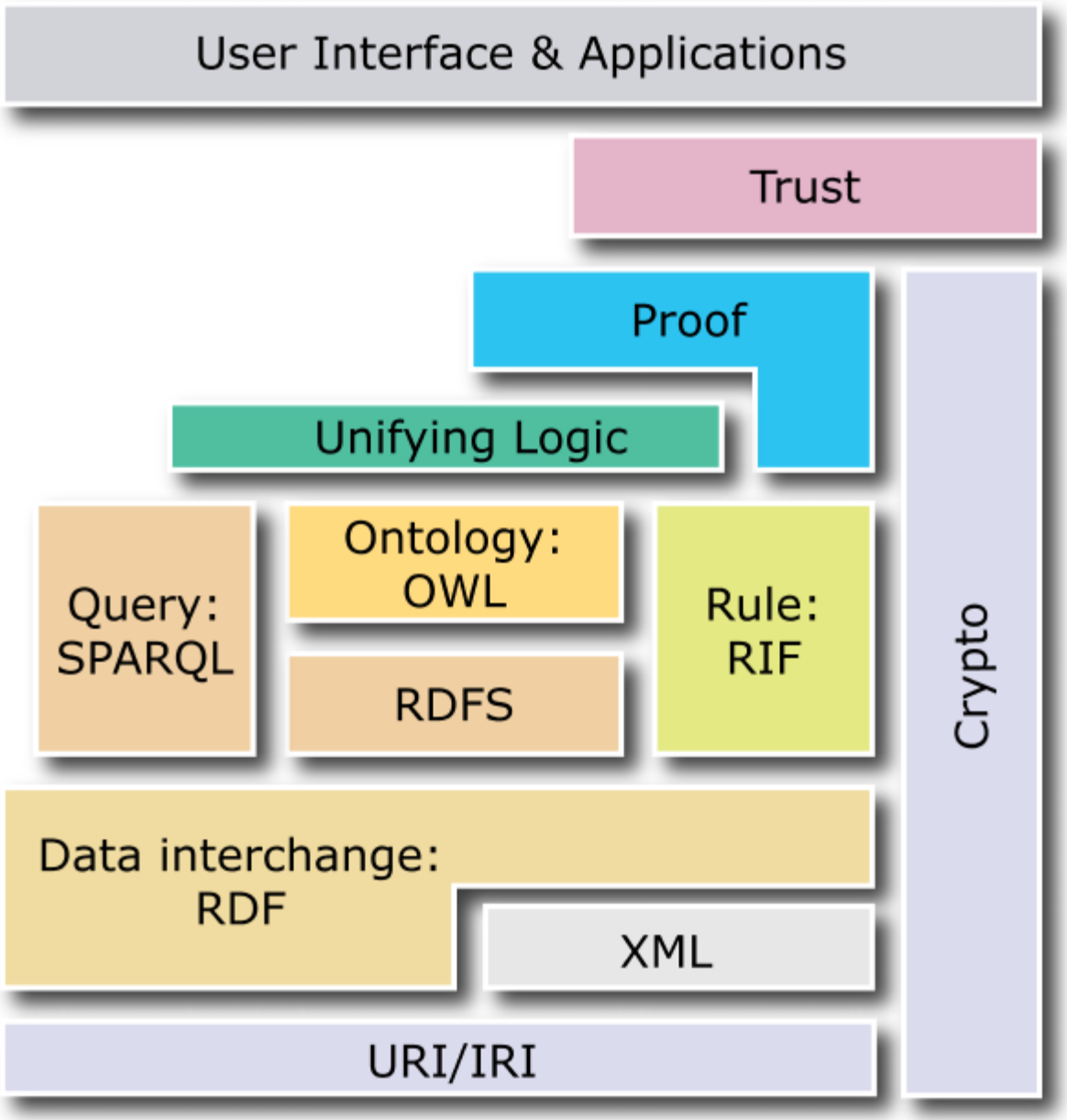
Kirk Borne. Big Data Science for Astronomy & Space, 2013



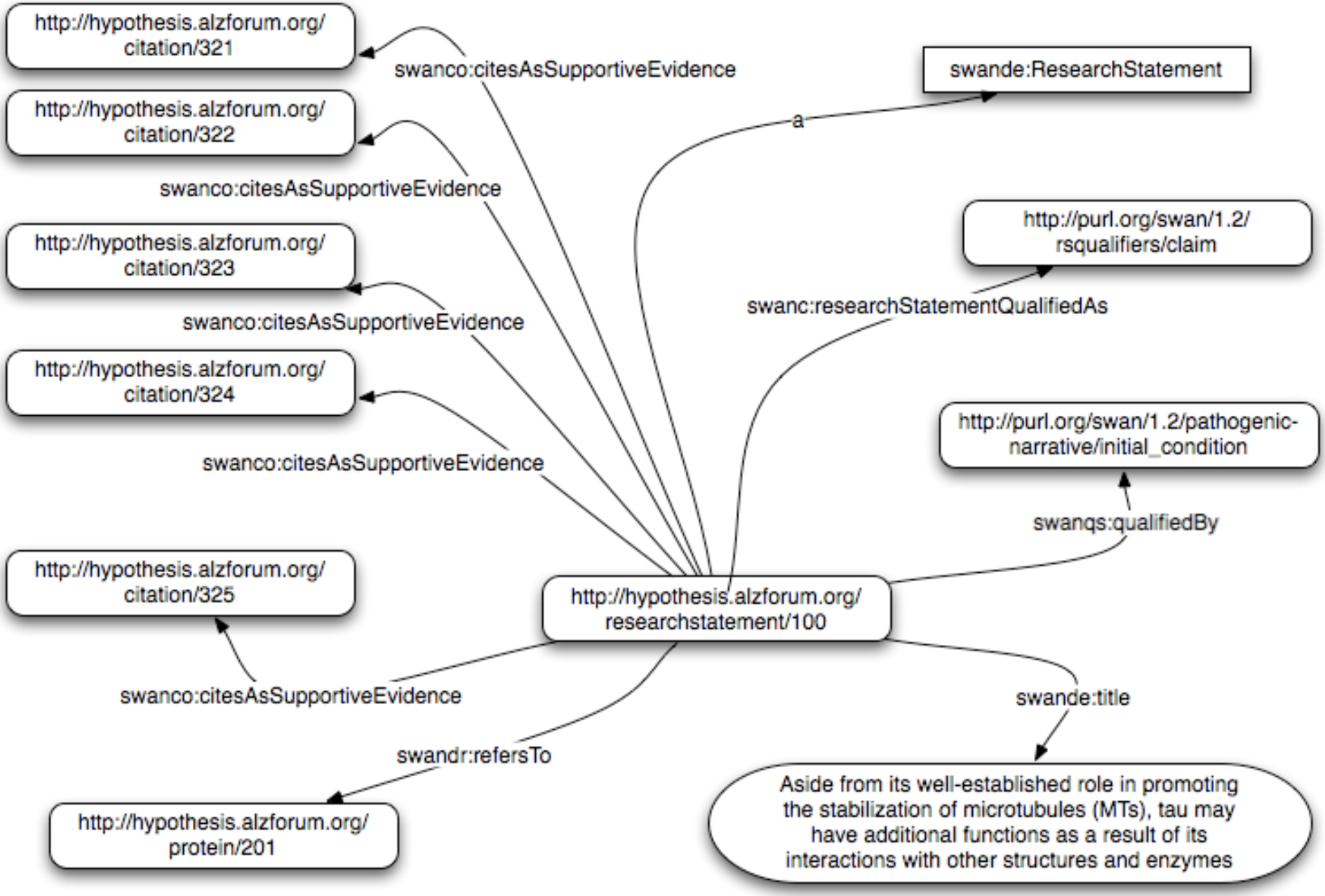
Hypothesis-driven approach generalized: goals



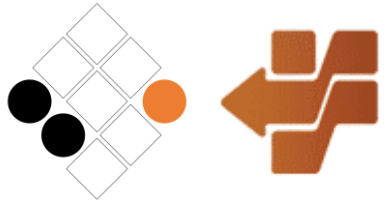
- Identify general requirements for various DID on the organization of hypothesis-driven experiments:
  - hypothesis testing and evolving,
  - model structuring
  - providing for their data independence
- Define methods and tools for their support
- Apply the resulting infrastructure for different applications



Semantic Web Stack


http://hypothesis.alzforum.org/citation/321
swanco:citesAsSupportiveEvidence
swande:ResearchStatement
http://hypothesis.alzforum.org/citation/322
a
swanco:citesAsSupportiveEvidence
http://purl.org/swan/1.2/rsqualifiers/claim
http://hypothesis.alzforum.org/citation/323
swanc:researchStatementQualifiedAs
swanco:citesAsSupportiveEvidence
http://hypothesis.alzforum.org/citation/324
http://purl.org/swan/1.2/pathogenic-narrative/initial_condition
swanco:citesAsSupportiveEvidence
swanqs:qualifiedBy
http://hypothesis.alzforum.org/citation/325
http://hypothesis.alzforum.org/researchstatement/100
swanco:citesAsSupportiveEvidence
swande:title
swandr:refersTo
Aside from its well-established role in promoting the stabilization of microtubules (MTs), tau may have additional functions as a result of its interactions with other structures and enzymes
http://hypothesis.alzforum.org/protein/201

Research Directions



- Use research lattices with Besancon Galaxy Model
- Investigate hypothesis encoding in the model
- Investigate model independence from data
- Collect data about the model evolution and changes:
  - Why new parameters were introduces?
  - Where the data is inconsistent with the model

XVII International Conference

Data Analytics and Management in Data Intensive Domains

DAMDID/RCDL'2015, October 13–16, 2015, Obninsk, Russia

- Interdisciplinary forum of researchers and practitioners from various domains of science
- Approaches to data analyses and management being developed in specific data intensive domains of X-informatics social sciences, as well as in various branches of informatics, industry, new technologies, finance and business are expected to contribute to the conference content

http://damdid2015.iate.obninsk.ru/en/overview_short.html

Thank You! Questions?