Fourth Brazil-France Workshop

On High Performance Computing and Scientific Data Management Driven by Highly Demanding Applications

Supporting in-silico Science with Data Management

Fabio Porto (fporto@lncc.br)

Collaborations

Esther Pacitti (INRIA – LIRMM)

Patrick Valduriez (INRIA – LIRMM)

Reza Akbarinia (INRIA – LIRMM)

José Antônio F. Macedo (UFC)

Ministério da Ciência, Tecnologia e Inovação

GOVERNO FEDERAL BRASIL PAÍS RICO E PAÍS SEM POBREZA

LIRMM

Outline

- DEXL + HOSCAR
- The Science Cockpit: Managing Science as Data
- Hypothesis as Data
- Group Presentation

FAPERJ-INRIA (Montpellier) Team Associé -2014 - 2016

http://dexl.lncc.br/music

MUSIC

Scientific Data Management in a Multi-Site Cloud

FAPERJ Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

Inria

LIRMM

Laboratório Nacional de Computação Científica

DEXL LAB
EXTREME DATA LAB

@HOSCAR - Gramado

SIMULATION

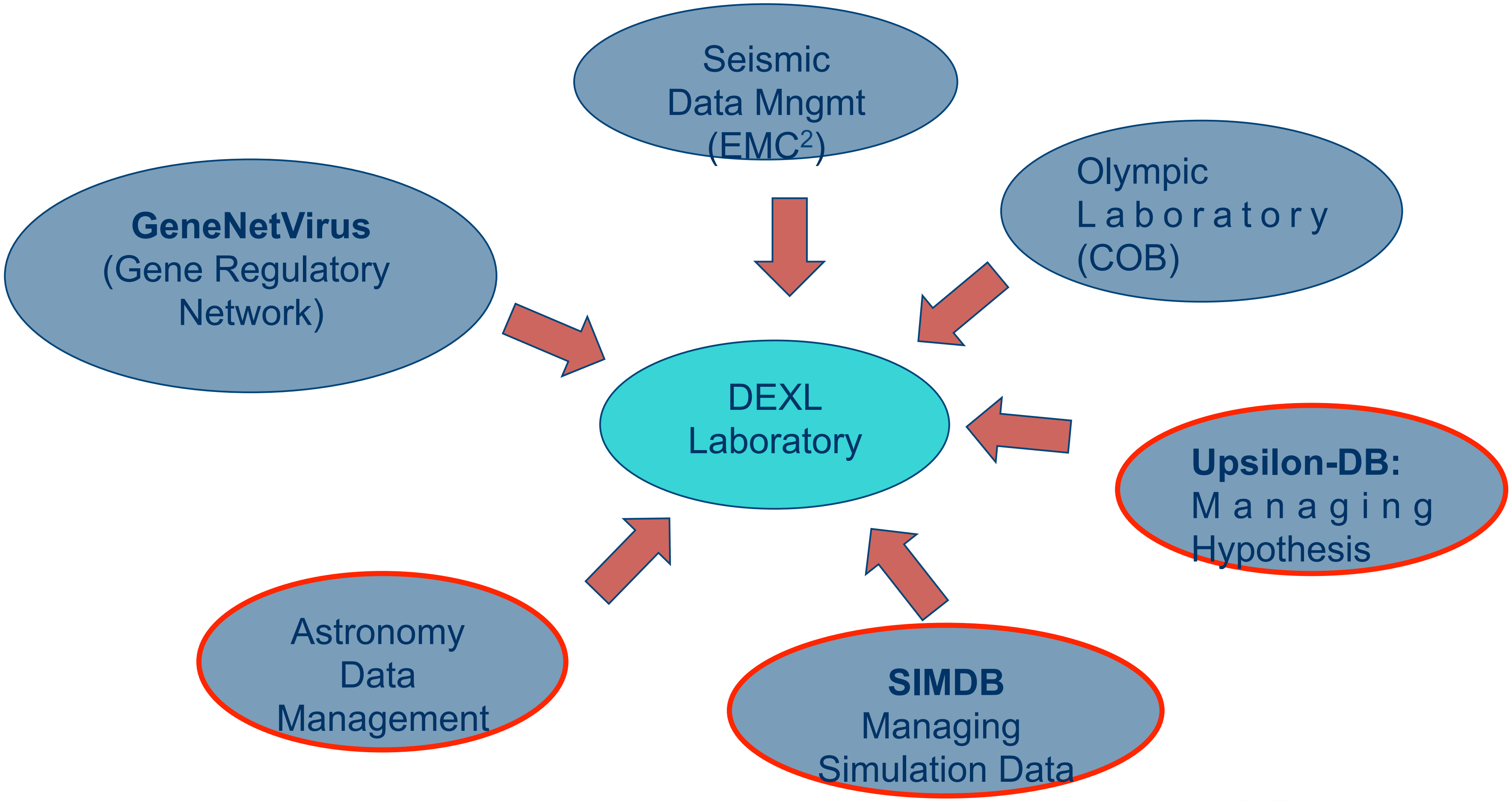
- UPSILON-DB (Bernardo Gonçalves)
  talk today
- SIMDB (Hermano Lustosa)
  collab. with Patrick Valduriez

ASTRONOMY

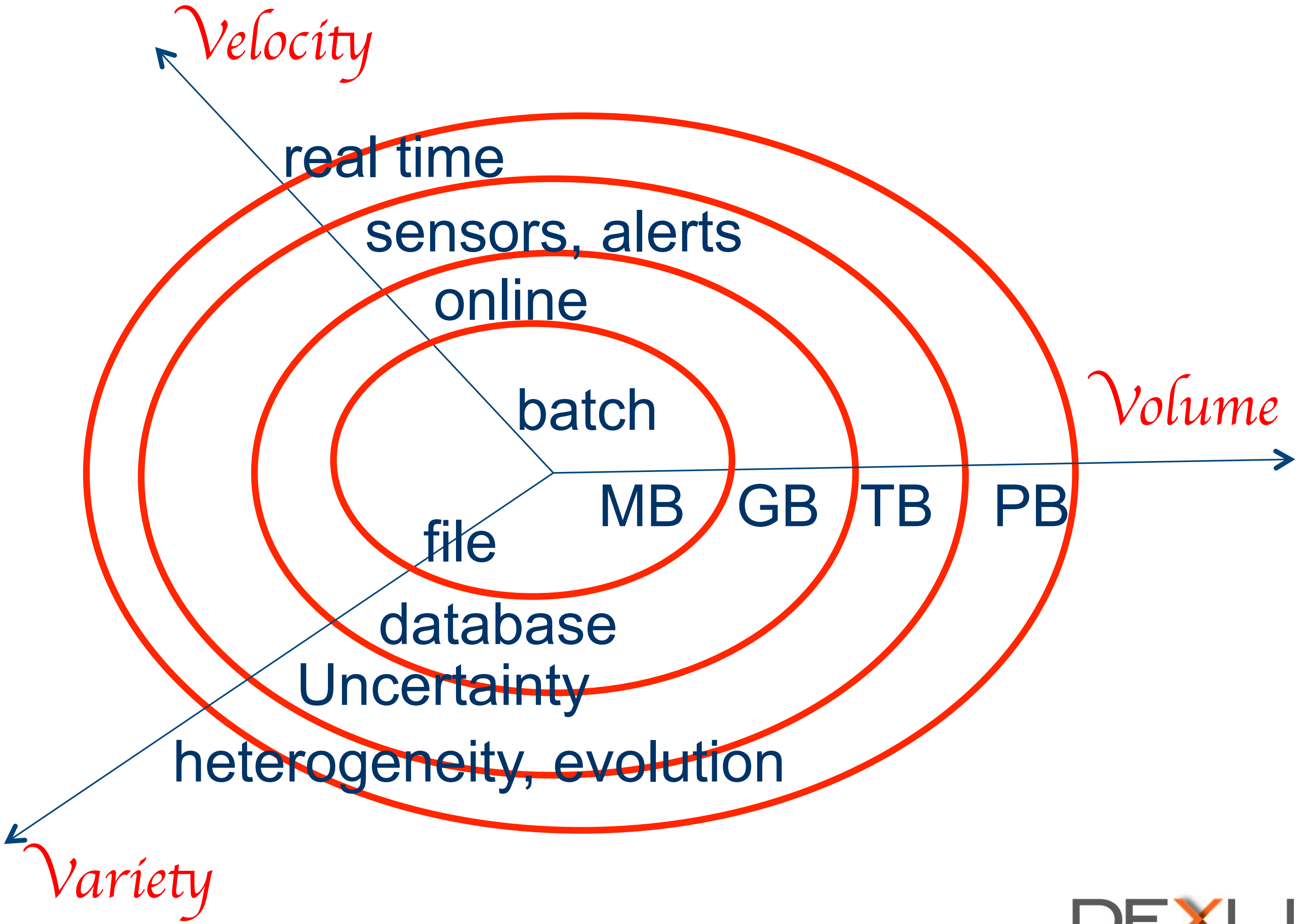
- ENTITY-RESOLUTION (Vinicius Freire)
  collabo. with Reza Akbarinia
- 2D DATA PARTITIONING (Daniel Gaspar)
  IN SUPPORT TO E.R.
  Sandwich at INRIA (2015-2016)
- PATTERN QUERY (Amir Khatibi)

DEXL LAB
EXTREME DATA LAB

DEXL – On going Projects



Big Data – V's Characterization



DEXL LAB
EXTREME DATA LAB

Big-Data (in science) Data Challenges

- Data Representation
  - Different Data Models:
    - Data structure and query languages
    - Graphs, Matrixes, Key-Value,...
- Data Uncertainty
  - Data is uncertain
  - uncertainty quantification on data
- Data Partitioning
  - in sync with data processing
- Data Heterogeneity
  - Data Granularity

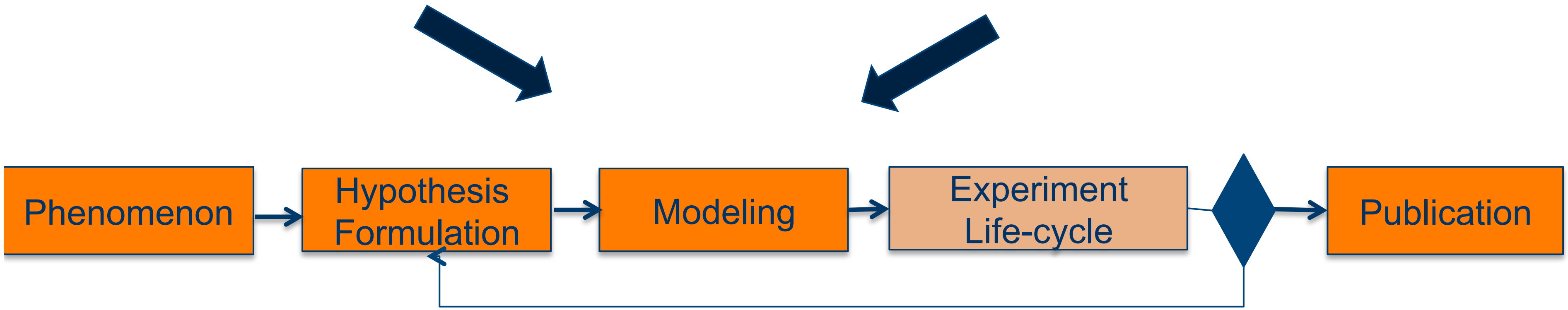
DEXL LAB
EXTREME DATA LAB

Why to use Data Management in science (in HOSCAR) ??

- Big Data does not fit in memory
  - efficient data access in disk
  - various indexing available
  - efficient data transformation algorithms
- High level query languages
  - to uniformly analyse data
  - with "free" automatic algebraic optimization
- Enables Data Sharing
  - manage thousands of datasets
- Data Transparency
  - standardize the communication with different applications
    - visualization
    - analytics
    - reproducibility results / rerun jobs

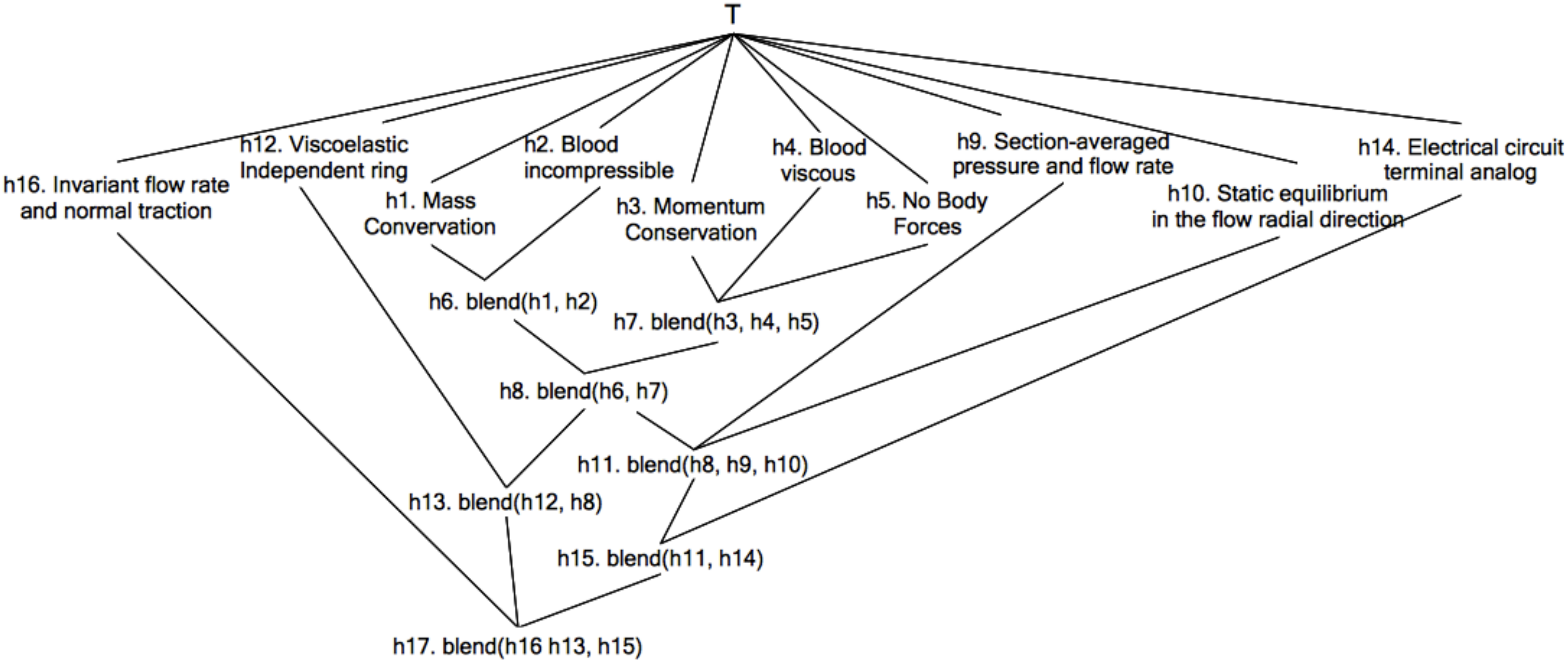
The Vision: Science life-cycle Cockpit

in-silico science life-cycle



DEXL LAB
EXTREME DATA LAB

Research Lattice for the Human Cardio Vascular System



B. Gonçalves, F. Porto, SSDBM 2013

DEXL LAB
EXTREME DATA LAB

MODELLING – HYPOTHESIS-DRIVEN BIG DATA RESEARCH

These PhD: Bernardo Gonçalves

Hypotheses in the Dark Energy Survey Project

- Phenomenon
  - The universe is increasing its expansion acceleration
    - Discovered in 1998 during supernovae investigation
    - Supported by redshift observation of far away supernovae
- Hypotheses
  - A new behaviour, **Dark Energy,** pushes the acceleration
  - The Universe density is not uniform
- Evidences
  - gravitational lenses
  - Galaxy clusters

To make sense of Big Data we need models

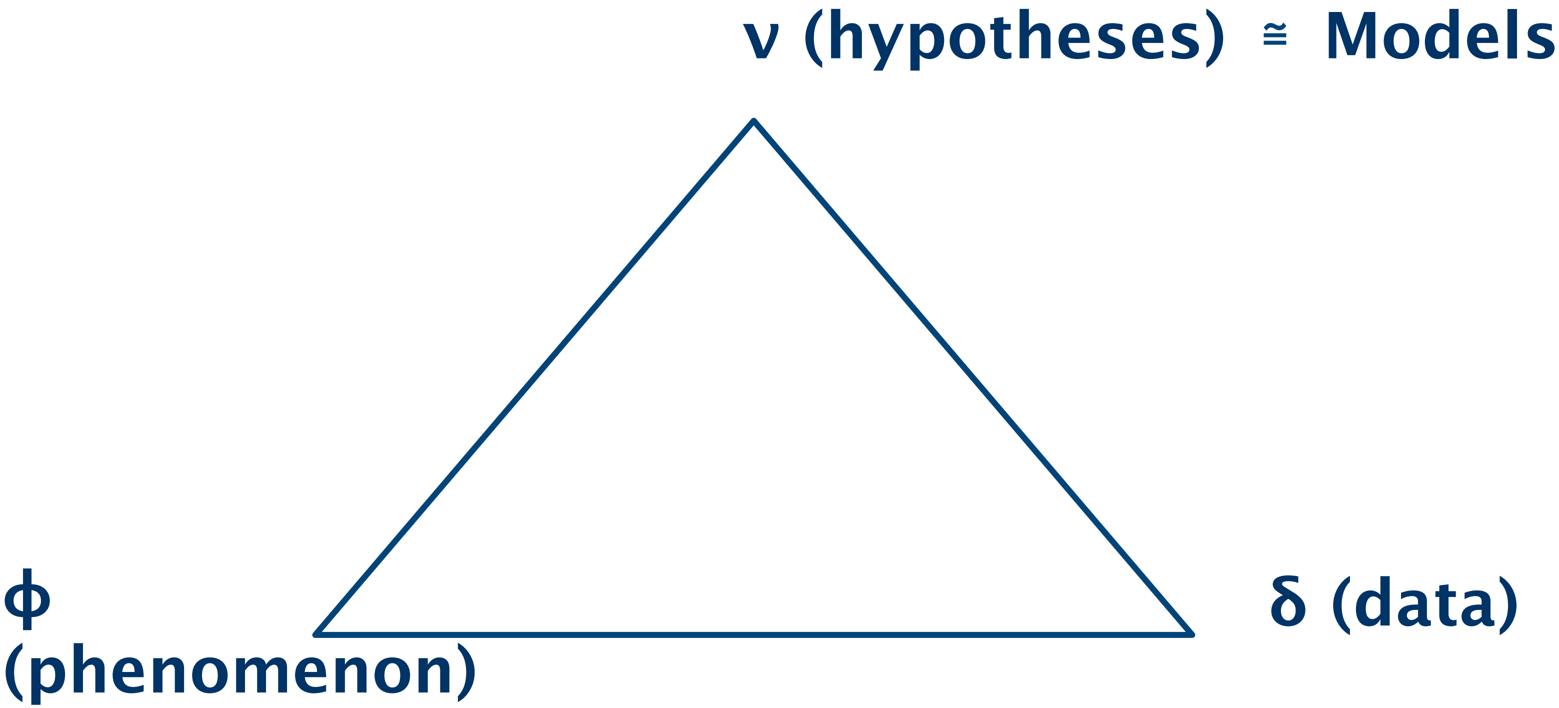
[Peter Haas – Data is Dead without what-if models, PVLDB 2011]

- In new Big Data prediction analysis – identify first principles that guide predictions – deep vs shallow prediction

Hypothesis driven Big Data analysis

- Scientific Hypothesis – a model for scientists' interpretation of a phenomenon;
- Science method – prove falsifiable hypotheses
  - Popper, K. Conjectures and Refutations
- Big Data analytics – hypotheses exploration
  - Can we probe the data to prove hypotheses?
  - How is the hypothesis related to the data?

Equivalence of interpretation



Hypotheses as Data - Upsilon-DB

- From the triangular equivalence, we derive that

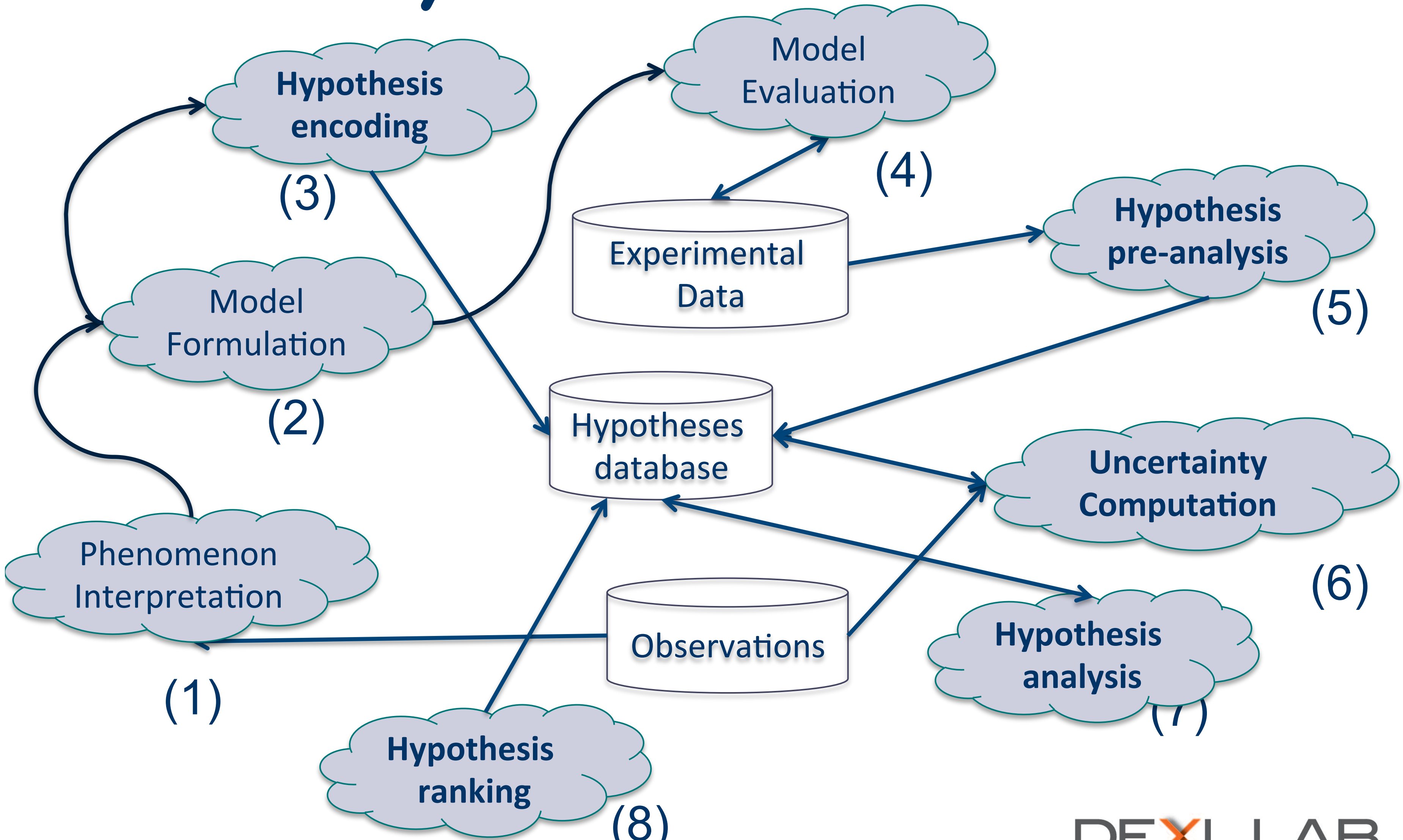
$$Hypothesis \equiv Model \equiv Data$$

- How can we infer data from Model?
  - hypothesis encoding

[Bernardo Gonçalves, Fabio Porto, PVLDB 2014]

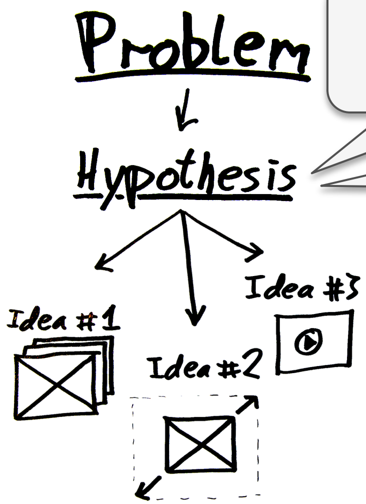
DEXL LAB
EXTREME DATA LAB

in silico science Hypotheses driven life-cycle



DEXL LAB
EXTREME DATA LAB

ϒ-DB: Overall Picture



Is this data?

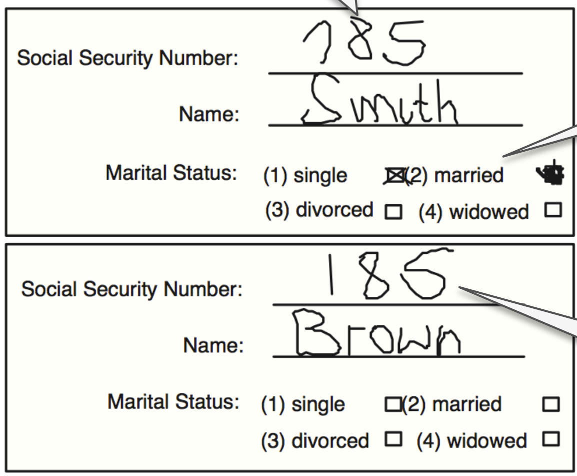
How do we manage it?

Hypotheses as uncertain data

U-Relational MayBMS

[Koch 2009]

185 or 785?


Social Security Number: 785
Name: Smith
Marital Status: (1) single ☒(2) married (3) divorced ☐ (4) widowed ☐

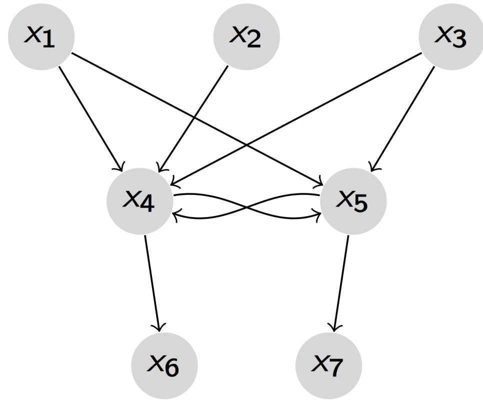
Social Security Number: 185
Name: Brown
Marital Status: (1) single ☐(2) married ☐ (3) divorced ☐ (4) widowed ☐

married?

185 or 186?

Modeling dependencies: Structural equations and causal ordering

FDs



Possible worlds, with different probabilities

DEXL LAB
EXTREME DATA LAB

Hypothesis as Models

Law of free fall

If a body falls from rest, its velocity at any point is proportional to the time it has been falling.

Hypothesis

$$a(t) = -g$$
$$v(t) = -gt + v_o$$
$$s(t) = -g/2\ t^2 + v_o t + s_o$$

Scientific Model

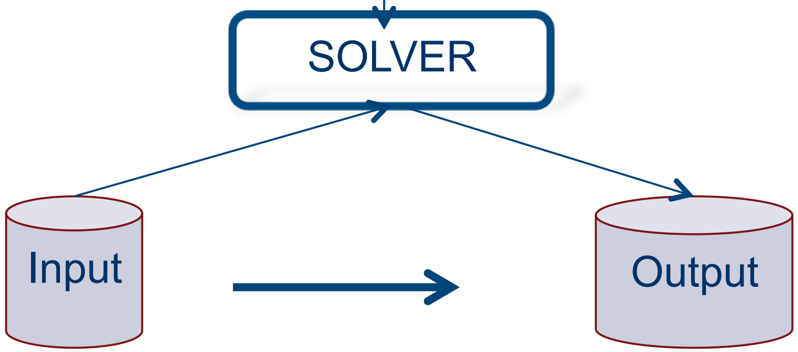
```
for k = 0:n;
   t = k * dt;
   v = -g*t + v0;
   s = -(g/2)*t^2 + v0*t + s0;
   t_plot(k) = t;
   v_plot(k) = v;
   s_plot(k) = s;
end
```

Computational Model

DEXL LAB
EXTREME DATA LAB

Hypothesis - From Models to Data

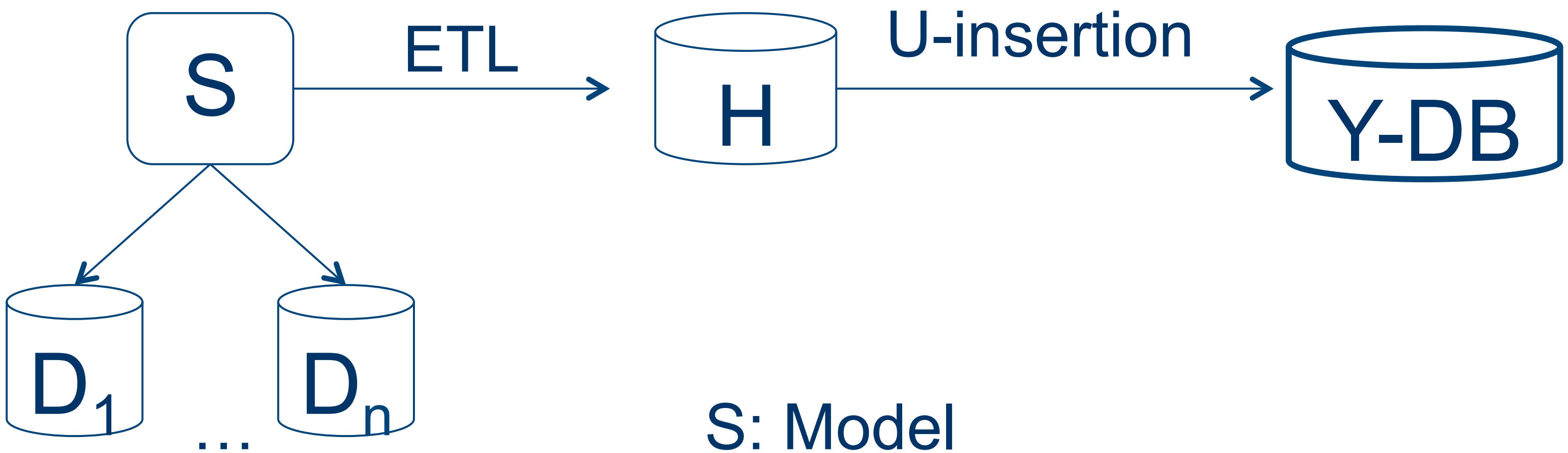
```
for k = 0:n;
   t = k * dt;
   v = -g*t + v0;
   s = -(g/2)*t^2 + v0*t + s0;
   t_plot(k) = t;
   v_plot(k) = v;
   s_plot(k) = s;
end
```



Experimental Phase: **OLTP**

Peter Haas, Model-Data Ecosystem , PODS 2014

Hypothesis encoding: From OLTP to OLAP in hypothesis



S: Model
D: output simulation data
H: Relational DB
Y: U-relational DB

DEXL LAB
EXTREME DATA LAB

Hypothesis as Data

```
for k = 0:n;
  t = k * dt;
  v = -g*t + v0;
  s = -(g/2)*t^2 + v0*t + s0;
  t_plot(k) = t;
  v_plot(k) = v;
  s_plot(k) = s;
end
```

Law of free fall

If a body falls from rest, its velocity at any point is proportional to the time it has been falling.

$$a(t) = -g$$
$$v(t) = -gt + v_o$$
$$s(t) = -g/2\, t^2 + v_o t + s_o$$

Free_Fall	t	v	s
	0	0	5000
	1	-32	4984
	2	-64	4936
	3	-96	4856
	4	-128	4744

DEXL LAB
EXTREME DATA LAB

Hypothesis as Data – DB Synthesis

- Models
  - formalize hypotheses
  - equations establish a functional dependency between dimensions and parameters and predicting variables
    - eg: ***g,t,vo -> v***
  - Derive a DB schema from DFs extracted from equations

DEXL LAB
EXTREME DATA LAB

Hypothesis as Data

- In the Free Fall example:

  - $\Sigma_1 = \{\Phi \text{ -> } g, v_o, s_o$
    $g, \nu \text{ -> } a$
    $g, v_o, t, \nu \text{ -> } v$
    $g, v_o, s_o, t, \nu \text{ -> } s\}$

$v(t) = -gt + v_o$

- Observe that Φ and ν are epistemological variables referring to the phenomenon and the hypothesis, respectively;

DEXL LAB
EXTREME DATA LAB

Hypothesis as Data - schema

- **Φ -> g, v_o, s_o**
  - defines the model parameters
  - It is expected to be violated reproducing the uncertainty in the model input;
  - Such uncertainty contributes to the quality of the hypothesis
- From Σ_1, the schema for predicting *a* under hypothesis h1 would be:
  - h1 (**$\underline{\Phi, v}$, a**)
- From Σ_1, the input parameters are defined as: *key violation
  - h1_input(**$\underline{\Phi}$, g, v_0, s_0**)

DEXL LAB
EXTREME DATA LAB

Hypothesis as *[Un]certain* Data

Uncertainty: 50%

Uncertainty: 33%

INPUT_H1	Φ	g	v_0	s_0
	1	32	0	5000
	1	32	10	5000
	1	32	20	5000
	1	32.2	0	5000
	1	32.2	10	5000
	1	32.2	20	5000

Look for subset of attributes with the same uncertainty

C-Relation after ETL

EXPLANATION	ϕ	v	Conf
	1	1	0.6
	1	2	0.2
	1	3	0.2

H1_INPUT	tid	ϕ	g	v_0	s_0
	1	1	32	0	5000
	2	1	32	10	5000
	3	1	32	20	5000
	4	1	32.2	0	5000
	5	1	32.2	10	5000
	6	1	32.2	20	5000

H1_OUTPUT[a]	tid	ϕ	v	a
	1	1	1	−32
	2	1	1	−32
	3	1	1	−32
	4	1	1	−32.2
	5	1	1	−32.2
	6	1	1	−32.2

Uncertainty Introduction

- Υ_DB is a probabilistic database
  [D. Suciu et al, Probabilistic Databases, 2011]
  - a Y-relation includes certain and conditional columns;
  - a conditional column is a pair (V_i , D_i), where V_i is a random variable and D_i is one of its possible values;
  - ex:
    - Create table Y_g as select U_phi, U_g
      from (**repair key** phi in (select phi, g , **count(*)** as Fr
      from INPUT_H1 **group by** phi, g **weight by Fr**) as U

DEXL LAB
EXTREME DATA LAB

Uncertain in *g* $\Rightarrow Input\_H1(g)$

INPUT_H1	Φ	g
	1	32
	1	32
	1	32
	1	32.2
	1	32.2
	1	32.2

$$3 \div 6 = 0.5$$

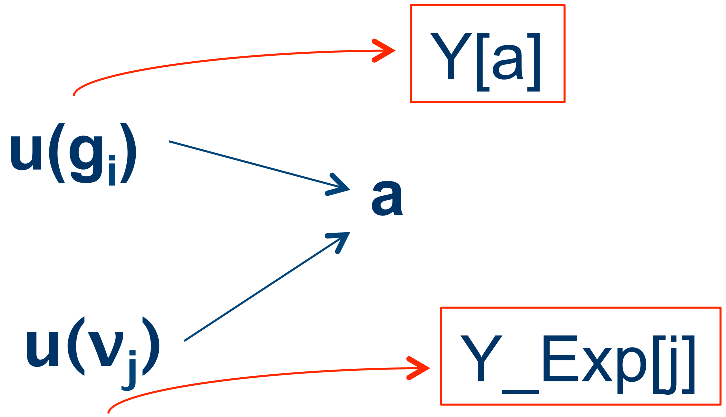
- Create table Y_g as select U_phi, U_g
  from (repair key phi in (select phi, g , count(*) as Fr
  from INPUT_H1 group by phi, g weighted by Fr) as U

Y_g	Φ	V-> D	g
	1	$x_1 \longrightarrow 1$	32
	1	$x_1 \longrightarrow 2$	32.2

DEXL LAB
EXTREME DATA LAB

Uncertainty propagation

- Σ_1 defines a graph of uncertainty propagation:
  - parameters – uncertainty on their data
  - predicting variables – parameter, model
  - Ex:



DEXL LAB
EXTREME DATA LAB

Synthesizing Prediction as a query

- as *g,v* ⟶ *a* in Σ_1, we can predict *a* as a query on uncertain relations Y_g and Y_R
  - create table **Y1_a as select** H.phi, H.upsilon, H.a from *H1_OUTPUT_a* **as H**, *Y_R* as **R**, *Y1_g* as **G**, (select min(tid) as tid, phi, g from H1_INPUT group by phi, g) as **U**

    where H.tid=U.tid and G.phi=U.phi and G.g=U.g

    and H.phi=R.phi and H.upsilon=R.upsilon

DEXL LAB EXTREME DATA LAB

Predicted Y-DB relation Y[a]

Y[a]	Φ	g	a	u
	1	32	32	0.30
	1	32.2	32.2	0.30

Y-DB enables data oriented uncertainty quantification of predicting variables

DEXL LAB
EXTREME DATA LAB

Upsilon-DB for Free-Fall

W	$V \mapsto D$	Pr
	$x_1 \mapsto 1$	.6
	$x_1 \mapsto 2$	.2
	$x_1 \mapsto 3$	.2
	$x_2 \mapsto 1$	.5
	$x_2 \mapsto 2$	.5

Y[Exp]	$V \mapsto D$	ϕ	υ
	$x_1 \mapsto 1$	1	1
	$x_1 \mapsto 2$	1	2
	$x_1 \mapsto 3$	1	3

Y1[g]	$V \mapsto D$	ϕ	g
	$x_2 \mapsto 1$	1	32
	$x_2 \mapsto 2$	1	32.2

Y1[a]	$V_1 \mapsto D_1$	$V_2 \mapsto D_2$	ϕ	υ	a
	$x_1 \mapsto 1$	$x_2 \mapsto 1$	1	1	-32
	$x_1 \mapsto 1$	$x_2 \mapsto 2$	1	1	-32.2

Y[a]	$V_1 \mapsto D_1$	$V_2 \mapsto D_2$	ϕ	υ	a
	$x_1 \mapsto 1$	$x_2 \mapsto 1$	1	1	-32
	$x_1 \mapsto 1$	$x_2 \mapsto 2$	1	1	-32.2
	$x_1 \mapsto 2$	–	1	2	0
	$x_1 \mapsto 3$	–	1	3	0

DEXL LAB
EXTREME DATA LAB

Competing hypotheses

FALL	ϕ	υ	t	v	s
	1	1	0	0	5000.00
	1	1	1	−32	4984.00
	1	1	2	−64	4936.00
	...	...	...	...	...

$Pr := 0.33$

FALL	ϕ	υ	t	v	s
	1	2	0	−607.9E-3	5000.00
	1	2	1	−607.9E-3	4997.32
	1	2	2	−607.9E-3	4994.64
	...	...	...	...	...

$Pr := 0.33$

FALL	ϕ	t	s
	1	0	5000
	1	1	4979.3
	1	2	4932.6
	...	...	...

FALL	ϕ	υ	t	v	s
	1	3	0	−4.17	5000.00
	1	3	1	−4.17	4981.63
	1	3	2	−4.17	4963.26
	...	...	...	...	...

$Pr := 0.33$

Predictive analytics

- What is the value of v in time=100 for hypothesis=1 and how confident we are about that value?
- What is the average velocity among hypotheses {1,2} between time 100 and 150?

Final Remarks

- Υ-DB is an innovative approach for Big Data management;
  - Reflects Hypothesis as data principle
  - Is formal and guards equivalence between data and models
  - Models uncertainty in the model and in the data
  - must be extended
    - to cope with observation validation (Bayesian Model)
    - to support multidimensional representation
  - read our paper at VLDB 2014 ☺

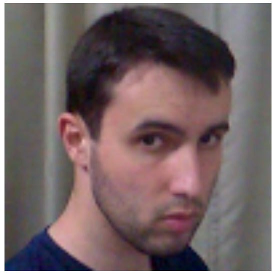
DEXL LAB
EXTREME DATA LAB

DEXL Team @HOSCAR

Dr Ramon G. Costa
UFLA
ramongomescosta@gmail.com

PhD Std. Bernardo Gonçalves
(bgonc@lncc.br)

MSc Std Amir Khatibi
(amir.khatibi.m@gmail.com)



MSc std Hermano Lustosa
(hllustosa@gmail.com)

PhD std Daniel Gaspar
(gaspar@lncc.br)

PhD Std
Vinicius F. Pires
(UFC) (vpires@lncc.br)

DEXL LAB
EXTREME DATA LAB

Obrigado !☺

http://dexl.lncc.br

Physiome Project Hypotheses

http://www.physiome.org/jsim/docs/MML Intro.html.

PHENOMENON	ϕ	Description
	1	Steady-sate effects on vessel diameter in response to change in intraluminal pressure.
	2	Dynamics of vessel diameter in response to pulsatile intraluminal pressure.

EXPLANATION	ϕ	υ	Conf
	1	113	1
	1	186	1
	2	60	1
	2	89	1

HYPOTHESIS	υ	Name	Pub	Data	Description
	60	Myogenic_Compliant_Vessel	N	N	This model simulates the flow through a passive and actively responding vessel driven by a sinusoidal pressure input.
	89	Myo_Dyn_Resp_wFit	N	Y	This model describes the dynamic response of a vessel after a step increase in intraluminal pressure.
	113	Vessel_Mechanics	N	Y	This model describes how a microvessel responds to changes in intraluminal pressure in the steady state. This change in vessel diameter to pressure is known as the myogenic response.
	186	Regulatory_Vessel	Y	Y	This model describes the steady state regulatory vessel response to changes in pressure across and shear stress on the vessel wall.

Primitive and Derived FDs - Y=89

$$
\begin{aligned}
\Sigma_{89} = \{ \quad \phi \quad &\rightarrow \quad \text{C1a C1p C2a C2p C3a Cglobal Cmyo} \\
&\quad \text{Dp100 Pc t\_delta t\_max t\_min taua taud}, \\
\phi\ \upsilon\ \text{t} \quad &\rightarrow \quad \text{DelP}, \\
\text{C1a C1p C2a C2p C3a Cglobal Cmyo Dp100 Pc } \upsilon \quad &\rightarrow \quad \text{Dc}, \\
\text{Dc Pc } \upsilon \quad &\rightarrow \quad \text{Tc}, \\
\text{Cglobal Cmyo Dc Pc } \upsilon \quad &\rightarrow \quad \text{Ac}, \\
\text{Dc } \upsilon \quad &\rightarrow \quad \text{D\_\_t\_min}, \\
\text{Ac } \upsilon \quad &\rightarrow \quad \text{A\_\_t\_min}, \\
\text{DelP Pc } \upsilon \quad &\rightarrow \quad \text{P}, \\
\text{D P } \upsilon \quad &\rightarrow \quad \text{T}, \\
\text{A C1a C1p C2a C2p C3a Dp100 P T } \upsilon \quad &\rightarrow \quad \text{Ttarget}, \\
\text{Cglobal Cmyo D P } \upsilon \quad &\rightarrow \quad \text{Atarget}, \\
\text{D\_\_t\_min Dc T Tc Ttarget t taud } \upsilon \quad &\rightarrow \quad \text{D}, \\
\text{A\_\_t\_min Atarget t taua } \upsilon \quad &\rightarrow \quad \text{A } \}.
\end{aligned}
$$

$$
\begin{aligned}
\Sigma'_{89} = \{ \quad \phi \quad &\rightarrow \quad \text{C1a C1p C2a C2p C3a Cglobal Cmyo} \\
&\quad \text{Dp100 Pc t\_delta t\_max t\_min taua taud}, \\
\phi\ \upsilon \quad &\rightarrow \quad \text{A\_\_t\_min Ac D\_\_t\_min Dc Tc}, \\
\phi\ \upsilon\ \text{t} \quad &\rightarrow \quad \text{A Atarget D DelP P T Ttarget } \}.
\end{aligned}
$$

DEXL LAB
EXTREME DATA LAB

Synthesized relations from Σ

H89_KEY1	ϕ	tid	C1a	C1p	C2a	C2p	C3a	Cglobal	Cmyo	Dp100	Pc	t_delta	t_max	t_min	taua	taud
	2	1	2.306	1.043	0.91	8.293	0.374	15.97	38.59	156.4	60	0.1	500	0	60	1
	2	2	1.965	4.924	0.91	18.530	0.374	15.121	35.687	156.992	50	0.01	500	0	90.998	9.034

H89_KEY2	ϕ	υ	tid	A__t_min	Ac	D__t_min	Dc	Tc
	2	89	1	0.271	0.271	97.057	97.057	0.388
	2	89	1	0.217	0.217	116.328	116.328	0.388

H89_KEY3	ϕ	υ	tid	t	A	Atarget	D	DelP	P	T	Ttarget
	2	89	1	0	0.271002840767991	0.27100284077	97.0568250529	0	60	0.388195373624	0.38819537272
	2	89	1	0.1	0.271002840768531	0.27100284140	97.0568250735	0	60	0.388195373707	0.38819537296
	2	89	1	0.2	0.271002840770027	0.27100284192	97.0568250906	0	60	0.388195373775	0.38819537315
	...	...	...	...	...	...	...	...	...	...	...
	2	89	2	0	0.216740664095983	0.216740664096	116.3278134698	0	50	0.387727490126	0.387727492846
	2	89	2	0.01	0.216740664095982	0.216740664078	116.3278134689	0	50	0.387727490123	0.387727492837
	2	89	2	0.02	0.216740664095979	0.216740664060	116.3278134680	0	50	0.387727490120	0.387727492828
	...	...	...	...	...	...	...	...	...	...	...

Computing the Uncertainty in D

Y[Exp]	$V \mapsto D$	ϕ	υ
	$x_1 \mapsto 1$	2	60
	$x_1 \mapsto 2$	2	89

Y89[tid]	$V \mapsto D$	ϕ	υ	tid
	$x_2 \mapsto 1$	2	89	1
	$x_2 \mapsto 2$	2	89	2

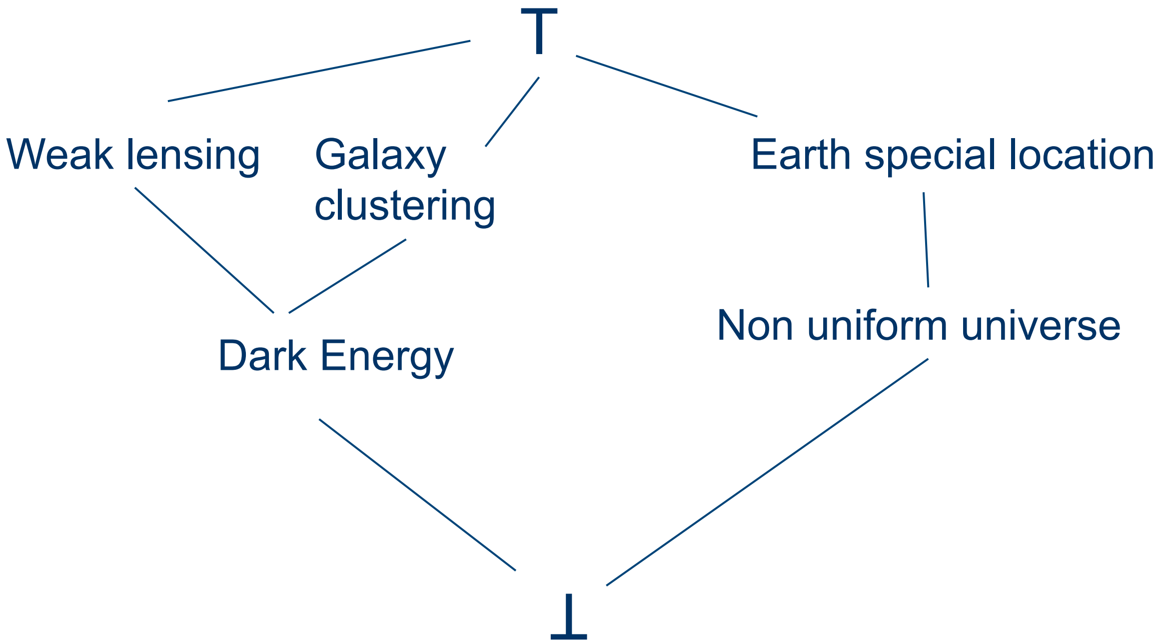
W	$V \mapsto D$	Prior	Post.
	$x_1 \mapsto 1$	.50	0
	$x_1 \mapsto 2$	.50	1
	$x_2 \mapsto 1$	.50	.304
	$x_2 \mapsto 2$	.50	.696

Y[D]	ϕ	υ	tid	t	D	Prior	Posterior
	2	60	1	14.8	194.996792066637	.50	.000
	2	89	1	14.8	97.0568250956827	.25	.304
	2	89	2	14.8	116.327813203282	.25	.696
	2	60	1	30.5	195.684170988267	.50	.000
	2	89	1	30.5	97.0568250767574	.25	.304
	2	89	2	30.5	116.327813337087	.25	.696
	2	60	1	43.7	195.283917335101	.50	.000
	2	89	1	43.7	97.056825073539	.25	.304
	2	89	2	43.7	116.327813382024	.25	.696

Managing a Research

- Different Hypotheses maybe raised;
- Ranking Hypotheses
  - Hypothesis information Capacity

Research Lattices – structure hypotheses of a phenomenon



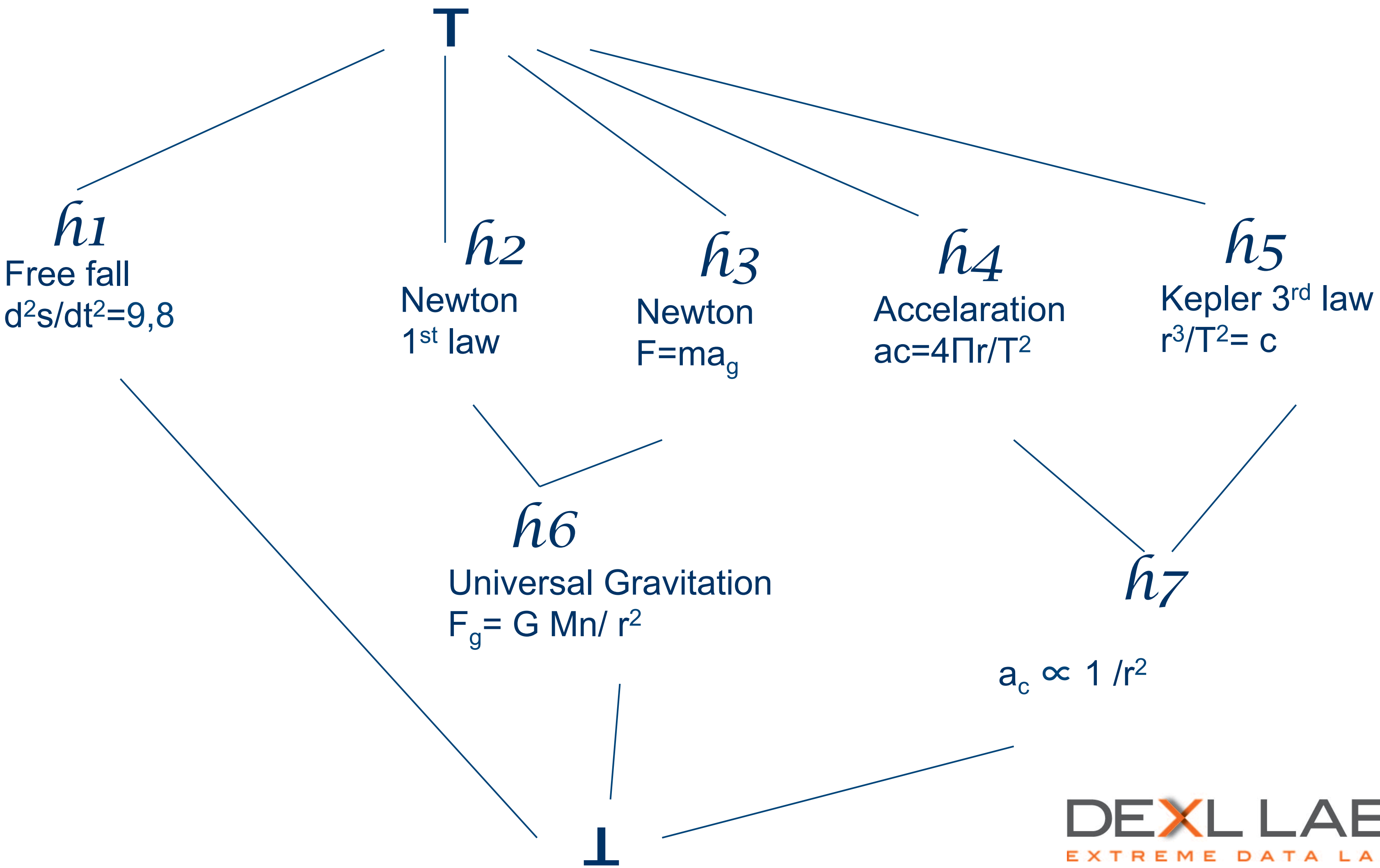
[B. Gonçalves, F. Porto, Research Lattices, AMW 2013]

DEXL LAB
EXTREME DATA LAB

Research Lattices

- Each Node is a hypotheses
- Given two hypotheses h1 and h2, in a R.L., if h1 $\geq$ h2 then h1 ***shows greater predictive*** capacity than h2;
  - capacity, similar to view capacity [Ullman
- *Top* corresponds to all knowledge of a domain;
- *Bottom* is the empty representation of lack of knowledge;

Research Lattice: Acceleration



Research lattice Operations

- Add/delete hypotheses
  - consistently keep the partial ordering;
  - automatic placement of hypotheses in the RL
- Querying
  - finding hypotheses based on "Free Fall" hypothesis
  - find competing hypotheses wrt "Dark Energy" Hypothesis

Sum up

- Research Lattice enables a formal yet bound representation of a research domain
- Different Hypotheses order according to their predictive capacity.

DEXL LAB
EXTREME DATA LAB