Hypothesis management in support of the e-scientific method

Bernardo Gonçalves
bng@br.ibm.com

Workshop eScience @UFABC – June 22, 2017

IBM: São Paulo's Building in Vila Mariana

The e-Scientific Method

"Originally, there was just experimental science, and then there was theoretical science, with Kepler's Laws, Newton's Laws of Motion, Maxwell's equations, and so on. Then, for many problems, the theoretical models grew too complicated to solve analytically, and people had to start simulating. These simulations have carried us through much of the last half of the last century. At this point, these simulations are generating a whole lot of data, along with a huge increase in data from the experimental sciences." — Jim Gray, 2007

Hypothesis Management

- Scientific research is based on the central idea of a hypothesis, meant to be established or refuted.

Hypothesis Management

- Scientific research is based on the central idea of a hypothesis, meant to be established or refuted.

- Over time and from multiple sources, we may collect evidence that support their (gray-shaded) truth or falsehood.

Hypothesis Management

- Scientific research is based on the central idea of a hypothesis, meant to be established or refuted.

- Over time and from multiple sources, we may collect evidence that support their (gray-shaded) truth or falsehood.

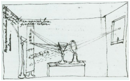
- *Hypothesis management* can be, therefore, closely related to the management of probabilistic data.

Hypothesis Management

- Scientific research is based on the central idea of a hypothesis, meant to be established or refuted.

- Over time and from multiple sources, we may collect evidence that support their (gray-shaded) truth or falsehood.

- *Hypothesis management* can be, therefore, closely related to the management of probabilistic data.



- ...A 'crucial experiment' allegedly establishes the truth of one of a set of competing theories (Routledge Encyc. of Philosophy).

From Hypotheses to Data

Law of free fall

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

(i)

$$
\begin{aligned}
a &= -g \\
v &= -g\,t + v_0 \\
s &= -(g/2)\,t^2 + v_0\,t + s_0
\end{aligned}
$$

(ii)

```
for k = 0:n;
   t = k * dt;
   v = -g*t + v_0;
   s = -(g/2)*t^2 + v_0*t + s_0;
   t_plot(k) = t;
   v_plot(k) = v;
   s_plot(k) = s;
end
```

(iii)

FALL	t	v	s
	0	0	5000
	1	−32	4984
	2	−64	4936
	3	−96	4856
	4	−128	4744
	...	...	...

(iv)

Rival Hypotheses: a Probability Distribution

- Rival hypotheses supposed to explain the same phenomenon.

H_1. Free fall law

$a = -g$

$v = -g\, t \,+\, v_0$

$s = -(g/2)\, t^2 \,+\, v_0\, t + s_0$

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

$Pr := 0.33$

H_2. Stokes' law

$a = 0$

$v = -\sqrt{gD/\, 4.6 \times 10^{-4}}$

$s = -t\sqrt{gD/4.6 \times 10^{-4}} + s_0$

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
	...	...	...

H_3. Velocity-squared law

$a = 0$

$v = -gD^2 /\, 3.29 \times 10^{-6}$

$s = -(gD^2 /\, 3.29 \times 10^{-6})\; t + s_0$

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
	...	...	...	...	...

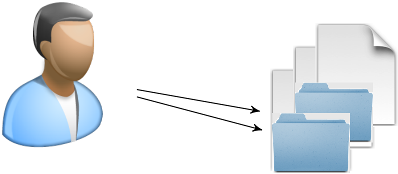
$Pr := 0.33$

Concrete Use Scenario in Computational Science

Example 1.

Bob is a computational scientist who is playing with a number of models and different parameter settings to see which one gives a best fit to his observation samples.

Each run constitutes a specific model instantiation that is associated with a unique file ('big table'). In the end of the day his resulting datasets are spread over many files and folders. □



Beyond Files: a Big Table Database

- User's **default** choice: struggle with the files to find relevant data.

"*There is life beyond files.*" (Jim Gray)

FALL	tid	t	g	v_0	s_0	a	v	s
	1	0	32	0	5000	-32	0	5000
	1	1	32	0	5000	-32	-32	4984
	1	2	32	0	5000	-32	-64	4936
	...	...	...	...	...	...	...	...
	2	0	32.2	0	5000	-32.2	0	5000
	2	1	32.2	0	5000	-32.2	-32.2	4983.9
	2	2	32.2	0	5000	-32.2	-64.4	4935.6
	...	...	...	...	...	...	...	...

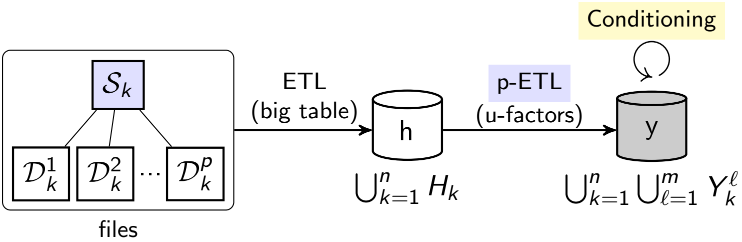
Do Better: a Hypothesis Database

$$
\begin{aligned}
a &= -g \\
v &= -g\,t + v_0 \\
s &= -(g/2)\,t^2 + v_0\,t + s_0
\end{aligned}
$$

FALL	tid	t	g	v_0	s_0	a	v	s
	1	0	32	0	5000	−32	0	5000
	1	1	32	0	5000	−32	−32	4984
	1	2	32	0	5000	−32	−64	4936
	...	...	...	...	...	...	...	...
	2	0	32.2	0	5000	−32.2	0	5000
	2	1	32.2	0	5000	−32.2	−32.2	4983.9
	2	2	32.2	0	5000	−32.2	−64.4	4935.6
	...	...	...	...	...	...	...	...

- Predictive structure: strong **correlations** from the math models (!).
- Project-level **standardization**: pick a favorite MathML editor to report and manage model equations declaratively.

Design-by-Synthesis Pipeline



- Technical challenges:
  1. **Encoding**: math equations → structural eqs. → functional deps.;
  2. **Causal reasoning**: inferring the causal ordering and u-factors;
  3. **Probabilistic DB synthesis**: normalization based on the u-factors;
  4. **Conditioning**: probability distribution update in face of evidence.

Given set $\mathcal{E}$ of equations over set $\mathcal{V}$ of variables... (1)

Law of free fall:

$$\mathcal{H} = \{ \begin{array}{rcl} g &=& 32, \\ v_0 &=& 0, \\ s_0 &=& 5000, \\ a &=& -g, \\ v &=& -g\,t + v_0, \\ s &=& -(g/2)t^2 + v_0\,t + s_0 \;\}. \end{array} \xrightarrow{h\text{-}encode} \Sigma = \{ \begin{array}{rcl} \boxed{\phi} &\to& g, \\ \boxed{\phi} &\to& v_0, \\ \boxed{\phi} &\to& s_0, \\ g\ \boxed{\upsilon} &\to& a, \\ g\ v_0\ t\ \boxed{\upsilon} &\to& v, \\ g\ v_0\ s_0\ t\ \boxed{\upsilon} &\to& s \;\}. \end{array}$$

Given set $\mathcal{E}$ of equations over set $\mathcal{V}$ of variables... (2)

Law of free fall:

$$\mathcal{S} = \{ \begin{aligned} f_1(g) &= 0, \\ f_2(\mathrm{v}_0) &= 0, \\ f_3(s_0) &= 0, \\ f_4(a, g) &= 0, \\ f_5(\mathrm{v}, g, t, \mathrm{v}_0) &= 0, \\ f_6(s, g, t, \mathrm{v}_0, s_0) &= 0 \quad \}. \end{aligned}$$

$$\xrightarrow{h\text{-}encode}$$

$$\Sigma = \{ \begin{aligned} \phi &\to g, \\ \phi &\to v_0, \\ \phi &\to s_0, \\ g\ \upsilon &\to a, \\ g\ v_0\ t\ \upsilon &\to v, \\ g\ v_0\ s_0\ t\ \upsilon &\to s\ \}. \end{aligned}$$

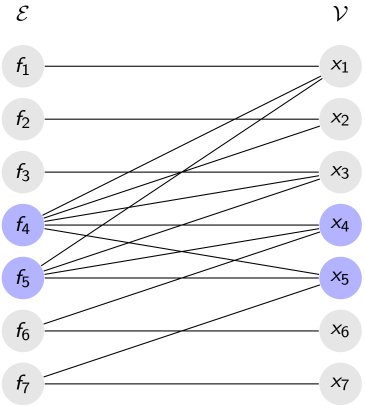
Causal Ordering Algorithm (AI Literature)

1. Identify '**minimal substructures**' at step k;
2. Reduce the matrix by eliminating them;
3. Call step $k+1$ **recursively**.

$$
\begin{array}{c|ccccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\
\hline
f_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
f_2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
f_3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
f_4 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
f_5 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
f_6 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
f_7 & 0 & 0 & 0 & 0 & 1 & 0 & 1
\end{array}
\;\rightarrow\;
\begin{array}{c|ccccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\
\hline
f_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
f_2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
f_3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
f_4 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
f_5 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
f_6 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
f_7 & 0 & 0 & 0 & 0 & 1 & 0 & 1
\end{array}
$$

Equivalent to Finding a Biclique $K_{m,n}$ in a Bipartite Graph

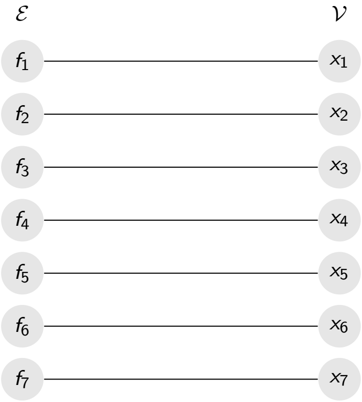
$$
\begin{aligned}
\mathcal{S} = \{ & \\
f_1(x_1) &= 0, \\
f_2(x_2) &= 0, \\
f_3(x_3) &= 0, \\
f_4(x_1, x_2, x_3, x_4, x_5) &= 0, \\
f_5(x_1, x_3, x_4, x_5) &= 0, \\
f_6(x_4, x_6) &= 0, \\
f_7(x_5, x_7) &= 0 \quad \}.
\end{aligned}
$$



Theorem 1.

Let $\mathcal{S}(\mathcal{E}, \mathcal{V})$ be a complete structure. The extraction of its causal ordering by COA($\mathcal{S}$) tries to solve an NP-Hard problem.

Easier: Complete Matching in a Bipartite Graph



- Hopcroft-Karp algorithm to solve it in $O(\sqrt{|\mathcal{E}|}\,|\mathcal{S}|)$;

Proposition 2.

Let $\mathcal{S}(\mathcal{E}, \mathcal{V})$ be a structure. Then a total causal mapping $\varphi\colon \mathcal{E} \to \mathcal{V}$ over $\mathcal{S}$ exists iff $\mathcal{S}$ is complete.

Provably Correct Approach to Hypothesis Encoding

$$C_{\varphi} = \{\, (x_a, x_b) \mid \text{there exists } f \in \mathcal{E} \text{ such that } \boxed{\varphi(f) = x_b} \text{ and } x_a \in \mathit{Vars}(f)\,\}$$

Proposition 1.

Let $\mathcal{S}(\mathcal{E}, \mathcal{V})$ be a structure, and $\varphi_1 \colon \mathcal{E} \to \mathcal{V}$ and $\varphi_2 \colon \mathcal{E} \to \mathcal{V}$ be any two total causal mappings over $\mathcal{S}$. Then $C^{+}_{\varphi_1} = C^{+}_{\varphi_2}$.

Causal Ordering: Sub-Quadratic Complexity on $|\mathcal{S}|$

Corollary 1.
Let $\mathcal{S}(\mathcal{E}, \mathcal{V})$ be a complete structure. Then a total causal mapping $\varphi\colon \mathcal{E} \to \mathcal{V}$ over $\mathcal{S}$ can be found by TCM$(\mathcal{S})$ in time that is bounded by $O(\sqrt{|\mathcal{E}|}\,|\mathcal{S}|)$.

ELSEVIER

Artificial Intelligence

Volume 238, September 2016, Pages 154-165

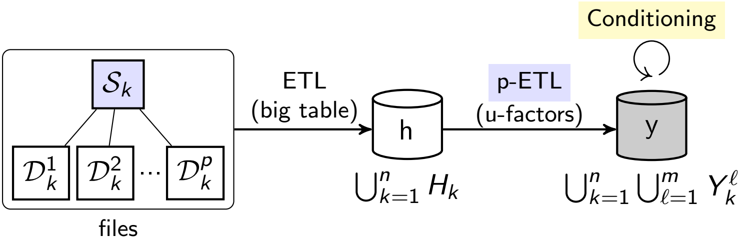
A note on the complexity of the causal ordering problem

Bernardo Gonçalves[a], Fabio Porto[b]

⊞ Show more

https://doi.org/10.1016/j.artint.2016.06.004

Coming Back from Detail



- Technical challenges:
  1. **Encoding**: math equations → structural eqs. → functional deps.;
  2. **Causal reasoning**: inferring the causal ordering and u-factors;
  3. **Probabilistic DB synthesis**: normalization based on the u-factors;
  4. **Conditioning**: probability distribution update in face of evidence.

Bob's Big Table (before our help)

H_3	tid	ϕ	υ	t	x_0	b	p	y_0	d	r	x	y
	1	1	3	0	30	.5	.02	4	.75	.02	30	4
	1	1	3	...	30	.5	.02	4	.75	.02	...	...
	2	1	3	0	30	.5	.018	4	.75	.023	30	4
	2	1	3	...	30	.5	.018	4	.75	.023	...	...
	3	1	3	0	30	.4	.02	4	.8	.02	30	4
	3	1	3	...	30	.4	.02	4	.8	.02	...	...
	4	1	3	0	30	.4	.018	4	.8	.023	30	4
	4	1	3	...	30	.4	.018	4	.8	.023	...	...
	5	1	3	0	30	.397	.02	4	.786	.02	30	4
	5	1	3	...	30	.397	.02	4	.786	.02	...	...
	6	1	3	0	30	.397	.018	4	.786	.023	30	4
	6	1	3	5	30	.397	.018	4	.786	.023	50.1	62.9
	6	1	3	10	30	.397	.018	4	.786	.023	13.8	8.65
	6	1	3	15	30	.397	.018	4	.786	.023	79.3	8.23
	6	1	3	20	30	.397	.018	4	.786	.023	12.6	30.7
	6	1	3	...	30	.397	.018	4	.786	.023	...	...

Synthesized Tables: Ready for Predictive Analysis

Y_0	$V \mapsto D$	ϕ	υ
	$x_1 \mapsto 1$	1	1
	$x_1 \mapsto 2$	1	2
	$x_1 \mapsto 3$	1	3

Y_3^1	$V \mapsto D$	ϕ	x_0
	$x_2 \mapsto 1$	1	30

Y_3^2	$V \mapsto D$	ϕ	b
	$x_3 \mapsto 1$	1	.5
	$x_3 \mapsto 2$	1	.4
	$x_3 \mapsto 3$	1	.397

Y_3^3	$V \mapsto D$	ϕ	p
	$x_4 \mapsto 1$	1	.020
	$x_4 \mapsto 2$	1	.018

Y_3^4	$V_1 \mapsto D_1$	$V_2 \mapsto D_2$	$V_3 \mapsto D_3$	$V_4 \mapsto D_4$	ϕ	υ	t	y	x
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 1$	$x_4 \mapsto 1$	1	3	1900	4	30
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 1$	$x_4 \mapsto 1$	1	3	...	...	...
	...	...	...	...	1	3	...	...	...
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	1	3	1900	4	30
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	1	3	1901	4.12	41.5
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	1	3	1902	5.78	56.7
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	1	3	1903	11.7	72.8
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	1	3	1904	31.1	75.9
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	1	3	..	..	..

Querying Rival Predictions with Probabilities

Y_3^4	$V_1 \mapsto D_1$	$V_2 \mapsto D_2$	$V_3 \mapsto D_3$	$V_4 \mapsto D_4$	ϕ	υ	t	y	x
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 1$	$x_4 \mapsto 1$	2	3	1900	4	30
	...	...	...	...	...	...	...	...	...
	$x_1 \mapsto 3$	$x_2 \mapsto 1$	$x_3 \mapsto 3$	$x_4 \mapsto 2$	2	3	1904	..	75.92
	...	...	...	...	...	...	...	...	...

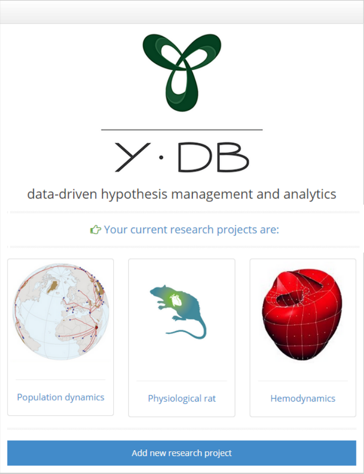
W	$V \mapsto D$	Pr
	$x_1 \mapsto 2$	.33
	$x_1 \mapsto 3$	.33
	$x_1 \mapsto 3$	.33
	$x_2 \mapsto 1$	1
	$x_3 \mapsto 1$	.33
	$x_3 \mapsto 2$	.33
	$x_3 \mapsto 3$	.33
	$x_4 \mapsto 1$	.5
	$x_4 \mapsto 2$	.5

Operation **conf()**;

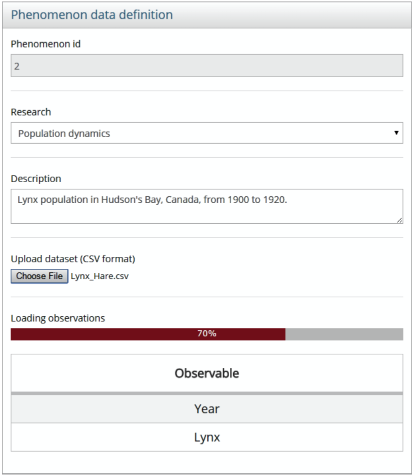
$$\theta = \{\, x_1 \mapsto 3,\ x_2 \mapsto 1,\ x_3 \mapsto 3,\ x_4 \mapsto 2 \,\},$$

$$Pr = .33 * 1 * .33 * .5 \approx .055$$

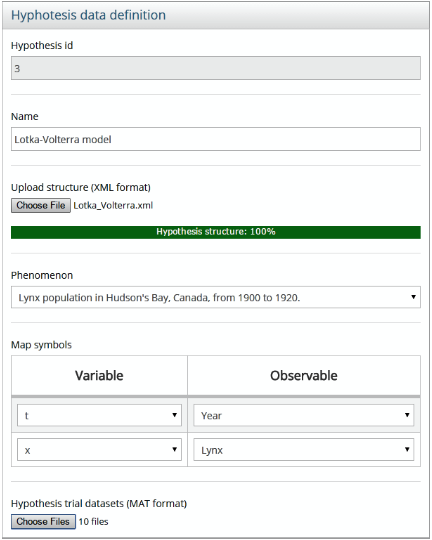
Prototype System (1)



Prototype System (2)



Prototype System (3)



Prototype System (4)

Hypothesis management

Hypothesis

Lotka-Volterra model

Phenomenon

Lynx population in Hudson's Bay, Canada, from 1900 to 1920.

Simulation trial

6

Key 1 | Key 2

« 1 2 3 4 5 6 7 8 9 10 ... 20 »

t	y	x
1904.0	31.1083920070696	75.9196961932191
1904.1	34.1895828779035	74.4878105315043
1904.2	37.4356490187431	72.6675156977604
1904.3	40.8008312965646	70.4705874385644
1904.4	44.2262696820135	67.9230251780413
1904.5	47.6417755935588	65.0649920502872
1904.6	50.9691309633260	61.9493743700121
1904.7	54.1267031673551	58.6389813583626

Prototype System (5)

Hypothesis Analytics

Phenomenon

Lynx population in Hudson's Bay, Canada, from 1900 to 1920.

Observations | Predictions

« 1 2 3 »

	Year	Lynx	Hare
	1900	30	4
	1901	47.2	6.1
	1902	70.2	9.8
	1903	77.4	35.2
	1904	36.3	59.4
	1905	20.6	41.7
	1906	18.1	19
	1907	21.4	13
	1908	22	8.3

Prototype System (6)

Hypothesis Analytics

Phenomenon

Lynx population in Hudson's Bay, Canada, from 1900 to 1920.

Observations | Predictions

« 1 2 3 4 5 6 7 8 9 10 ... 20 »

upsilon	tid	Year	Lynx	conf
3	2	1904	65.060410460081	0.183505
3	6	1904	75.919696193219	0.179993
3	4	1904	77.459735769215	0.175992
3	1	1904	89.592307430943	0.131452
3	5	1904	88.321831841064	0.127000
3	3	1904	90.083803232660	0.124023
1	1	1904	16.487212706992	0.047211
2	2	1904	77.822475573932	0.017372
2	1	1904	79.812581025093	0.013234
1	2	1904	18.221188003898	0.000220

Takeaways

We have seen:

- The automatic construction of a probabilistic DB (out of math equations and datasets) to support the analysis of crucial exps.
- This is hypothesis management (as data management and analytics) in support of the e-scientific method.

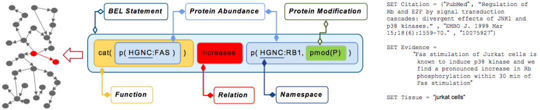
(Papers: *PVLDB'14, IEEE Computing in Science & Eng.'15, Artif. Intell.'16*)

Future Work: in the field of Bioinformatics

- Recommendation of crucial experiments;
  - Find 'rival' (structurally similar) math models from a repository;
  - Example: BioModels (EMBL-EBI), with 1.6K+ models stored.

Future Work: in the field of Bioinformatics

- Recommendation of crucial experiments;
  - Find 'rival' (structurally similar) math models from a repository;
  - Example: BioModels (EMBL-EBI), with 1.6K+ models stored.
- Refutation attempts (sense of Karl Popper);
  - Look for negative claims in the literature.
  - Example: Causal Biological Networks, with 120+ models stored (each has hundreds of hypothetical claims).



Thank you.

Acknowledgements

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

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

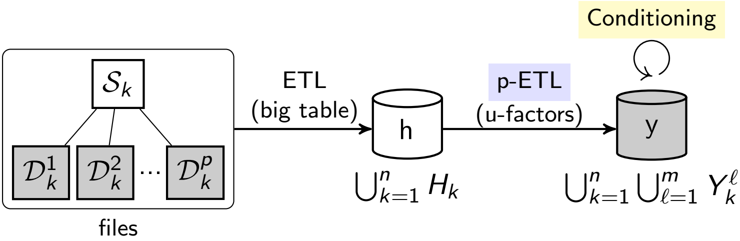
IBM

Questions?

Bernardo Gonçalves

bng@br.ibm.com

Design-by-Synthesis Pipeline



- Technical challenges:
  1. **Encoding**: math equations $\rightarrow$ structural eqs. $\rightarrow$ functional deps.;
  2. **Causal reasoning**: inferring the causal ordering and u-factors;
  3. **Probabilistic DB synthesis**: normalization based on the u-factors;
  4. **Conditioning**: probability distribution update in face of evidence.

The Folding $\Sigma^{\looparrowright}$ of Σ

Acyclic pseudo-transitive reasoning

Algorithm 1 Folding of an fd set.

procedure $\textsc{Folding}(\Sigma\colon \text{fd set})$
Require: Σ given encodes complete structure $\mathcal{S}$
Ensure: Returns fd set $\Sigma^{\looparrowright}$, the folding of Σ
    $\Sigma^{\looparrowright} \leftarrow \varnothing$
    for all $\langle X, A\rangle \in \Sigma$ **do**
        $Z \leftarrow$ AFolding(Σ, A)
        $\Sigma^{\looparrowright} \leftarrow \Sigma^{\looparrowright} \cup \langle Z, A\rangle$
    return $\Sigma^{\looparrowright}$

The Folding $\Sigma^{\looparrowright}$ of Σ

Acyclic pseudo-transitive reasoning

$$
\begin{aligned}
\Sigma = \{ \quad \phi \quad &\to \quad x_0, \\
\phi \quad &\to \quad b, \\
\phi \quad &\to \quad p, \\
\phi \quad &\to \quad y_0, \\
\phi \quad &\to \quad d, \\
\phi \quad &\to \quad r, \\
x_0\, b\, p\;\; t\, \upsilon\, y \quad &\to \quad x, \\
y_0\, d\, r\;\; t\, \upsilon\, x \quad &\to \quad y \;\; \}.
\end{aligned}
\qquad \xrightarrow{\textit{folding}} \qquad
\begin{aligned}
&\Upsilon(\Sigma)^{\looparrowright} = \{ \\
&x_0\, b\, p\;\; t\, \upsilon\, y\;\; y_0\, d\, r \quad \to \quad x, \\
&y_0\, d\, r\;\; t\, \upsilon\, x\;\; x_0\, b\, p \quad \to \quad y \;\; \}.
\end{aligned}
$$

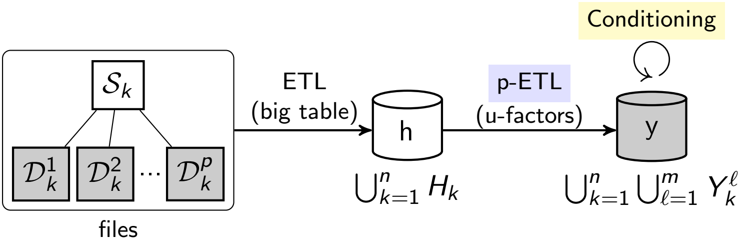
FD set Σ_{89} (a real Physiology Model that "fits the screen")

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

Its Folding $\Sigma_{89}^{\looparrowright}$

$$\Upsilon(\Sigma_{89})^{\looparrowright} = \{ \quad \text{C1a C2a } \phi\ \upsilon \quad \to \quad \text{A\_t\_min Ac D\_t\_min Dc Tc}, \\ \text{C1a DelP } \phi\ \upsilon \quad \to \quad \text{P}, \\ \text{C1a C2a DelP } \phi \text{ t } \upsilon \text{ T} \quad \to \quad \text{A Atarget D Ttarget} \quad \}.$$

Design-by-Synthesis Pipeline



- Technical challenges:
  1. **Encoding**: math equations $\rightarrow$ structural eqs. $\rightarrow$ functional deps.;
  2. **Causal reasoning**: inferring the causal ordering and u-factors;
  3. **Probabilistic DB synthesis**: normalization based on the u-factors;
  4. **Conditioning**: probability distribution update in face of evidence.

Defining the Theoretical U-factor

$$Y_0 \ := \ \pi_{\phi,\upsilon}(\text{repair-key}_{\phi}(H_0)).$$

H_0	ϕ	υ
	1	1
	1	2
	1	3

Y_0	$V \mapsto D$	ϕ	υ
	$x_1 \mapsto 1$	1	1
	$x_1 \mapsto 2$	1	2
	$x_1 \mapsto 3$	1	3

W	$V \mapsto D$	Pr
	$x_1 \mapsto 1$	.33
	$x_1 \mapsto 2$	.33
	$x_1 \mapsto 3$	.33

U-factor Learning

Discovery of contingent functional dependencies

- '**Input**' relation of the Lotka-Volterra hypothesis. Observe:
  - **Multiplicity** of parameter values;
  - **Correlations** between parameter values.

H_3^1	tid	ϕ	x_0	b	p	y_0	d	r
	1	1	30	.5	.020	4	.75	.020
	2	1	30	.5	.018	4	.75	.023
	3	1	30	.4	.020	4	.8	.020
	4	1	30	.4	.018	4	.8	.023
	5	1	30	.397	.020	4	.786	.020
	6	1	30	.397	.018	4	.786	.023

$$\Omega = \{ \begin{array}{lcl} \phi\, x_0 & \to & y_0, \\ \phi\, b & \to & d, \\ \phi\, p & \to & r \end{array} \}.$$

U-factorization

Defining the Empirical U-factors

$$Y_k^i := \pi_{\phi\, A\, G}\,(\text{repair-key}_{\,\phi\,@\,count}\,(\,\gamma_{\phi,\,A,\,G,\,count(*)}(H_k)\,)\,).$$

Y_3^1	$V \mapsto D$	ϕ	x_0	y_0
	$x_2 \mapsto 1$	1	30	4

Y_3^2	$V \mapsto D$	ϕ	b	p
	$x_3 \mapsto 1$	1	.5	.5
	$x_3 \mapsto 2$	1	.4	.8
	$x_3 \mapsto 3$	1	.397	.786

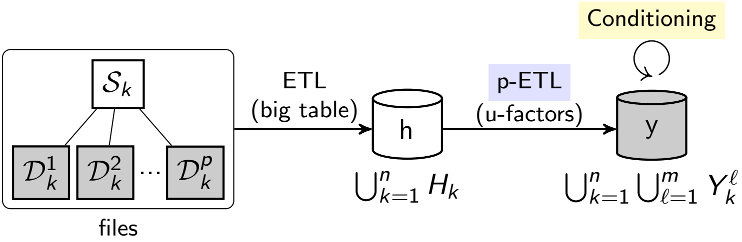
W	$V \mapsto D$	Pr
	...	...
	$x_2 \mapsto 1$	1
	$x_3 \mapsto 1$	.33
	$x_3 \mapsto 2$	.33
	$x_3 \mapsto 3$	.33
	$x_4 \mapsto 1$	.5
	$x_4 \mapsto 2$	.5

Y_3^3	$V \mapsto D$	ϕ	p	r
	$x_4 \mapsto 1$	1	.020	.020
	$x_4 \mapsto 2$	1	.018	.023

Design-Theoretic Properties

- Desirable properties for probability update are ensured;
  - **Claim-centered** decomposition.
    - Theorem 6: **BCNF** w.r.t. causal dependencies.
  - **Correctness** of uncertainty decomposition.
    - Theorem 7: **Lossless join** w.r.t. causal dependencies.

Design-by-Synthesis Pipeline



- Technical challenges:
  1. **Encoding**: math equations $\rightarrow$ structural eqs. $\rightarrow$ functional deps.;
  2. **Causal reasoning**: inferring the causal ordering and u-factors;
  3. **Probabilistic DB synthesis**: normalization based on the u-factors;
  4. **Conditioning**: probability distribution update in face of evidence.

Systematic Application of Bayesian Inference

1. User selection of the **observation** sample;
2. System selection of the competing **prediction** samples;
3. Bayesian **inference** ;
4. Probability distribution **update** ;

Bayes' Rule

- Normal density **likelihood** function:

$$f(y \,|\, \mu_k) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(y-\mu_k)^2} \tag{1}$$

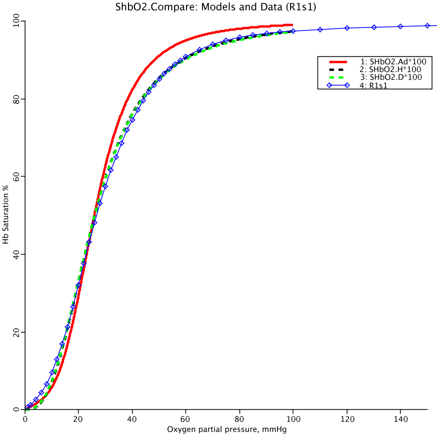
- **Bayes** ' rule:

$$p(\mu_k \,|\, y_1, \ldots, y_n) \;=\; \frac{\prod_{j=1}^{n} f(y_j \,|\, \mu_{kj})\; p(\mu_k)}{\sum_{i=1}^{m} \prod_{j=1}^{n} f(y_j \,|\, \mu_{ij})\; p(\mu_i)} \tag{2}$$

where $y_1, \ldots, y_n$ is the observation sample, μ_k is prediction k and σ is the standard deviation parameter.

Probability Update in Face of Evidence

STUDY	ϕ	υ	pO2	SHbO2	Prior	Posterior
	1	32	100	9.72764121981342E-1	.333	.335441
	1	28	100	9.74346796798538E-1	.333	.335398
	1	31	100	9.90781330988763E-1	.333	.329161



Viewpoint: Why Hypothesis Management?


nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 513 > Issue 7516 > Comment > Article

NATURE | COMMENT

Neuroscience: Where is the brain in the Human Brain Project?

Yves Frégnac & Gilles Laurent

03 September 2014

"Numerical simulations and 'big data' are essential in modern science, but they do not alone yield understanding. Building a massive database to feed simulations without corrective loops between hypotheses and experimental tests seems, at best, a waste of time and money." **Nature** 513, Sept 2014.