CLOPINET

PASCAL CHALLENGES

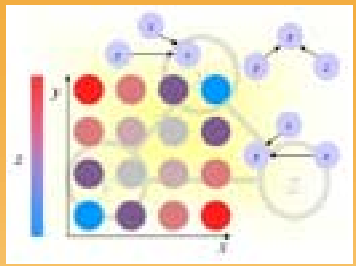
NSF

MPS

Microsoft

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Unipen



Open-access datasets for time series causality discovery validation

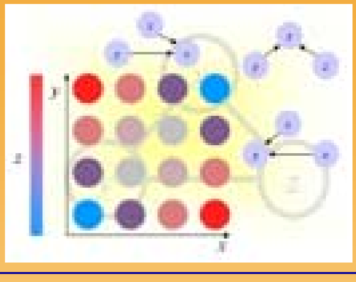
I. Guyon, C. Aliferis, G. Cooper,

A. Elisseff, O. Guyon, J.-P. Pellet,

A. Statnikov, P. Spirtes

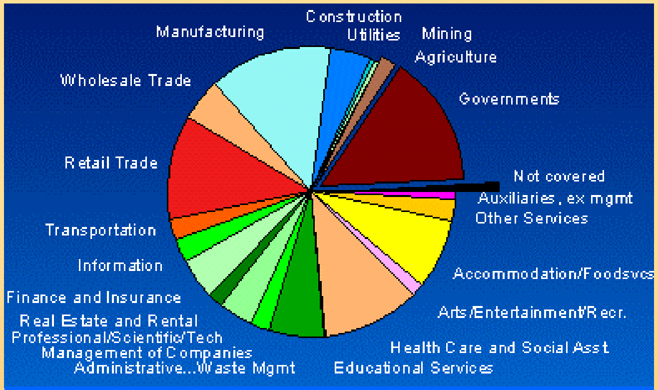
http://clopinet.com/causality/ ***causality@clopinet.com***

The challenges of causality discovery



What affects…

…your health?



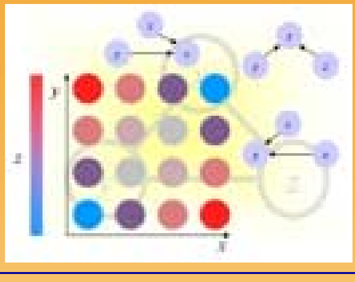
… the economy?

…climate changes?

and…

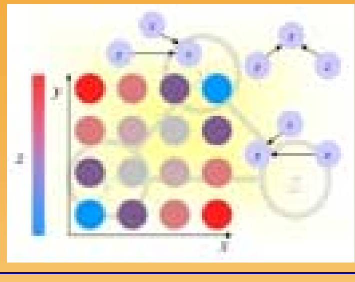
which actions will have beneficial effects?

Causality and time

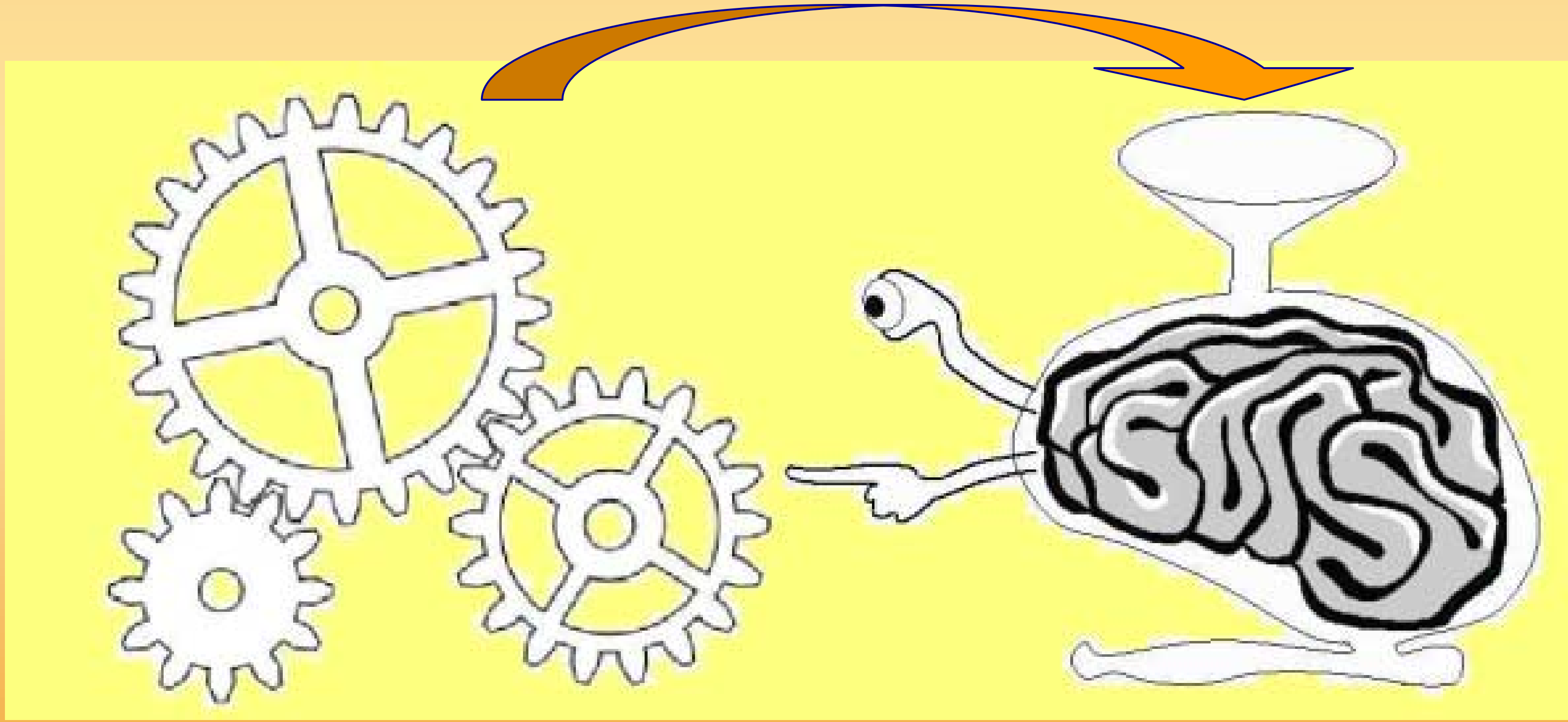


- Everyday notion of causality involves time:
  The causes precede their effects
- Is that always true?
  - Delayed/weak measurements; reverse causation
  - Final cause (objective)
- Time does not resolve:
  - Variability
  - Confounding
  - Sample bias
- Other difficulties:
  - Non i.i.d. samples: redundancy; correlation misleading.
  - Seasonality.
  - Censored data.

Experimenting is needed…

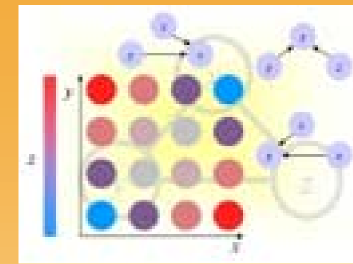


Experimenting is usually needed to determine cause-effect relationships

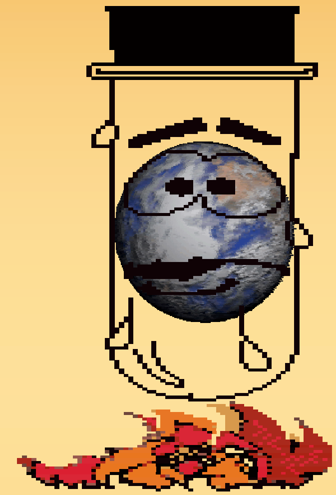


but …

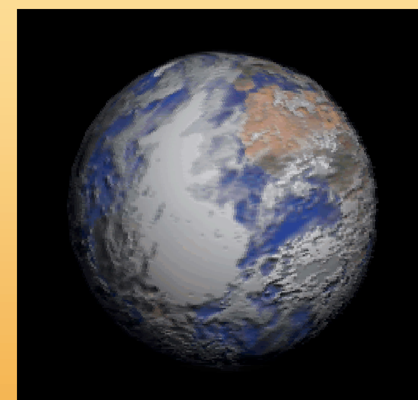
but…



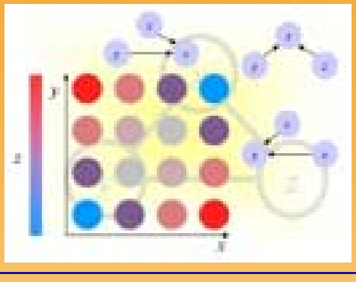
- Experiments are often:
  - Costly
  - Unethical
  - Infeasible



- Non-experimental "observational" data is abundant and costs less.



The Causality Workbench



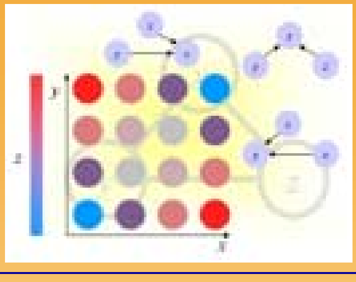
Our goal:

Identify algorithms both

- ***efficient to identify causes***
- ***cost effective***

How ?

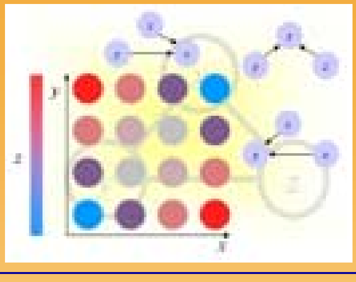
The Causality Workbench



Our challenges:

- Finding adequate data
  - Ground truth of causal relationships
  - Experimental data
  - Large sample size
- Conducting “life” experiments
  - Costly
  - Impractical in a challenge setting

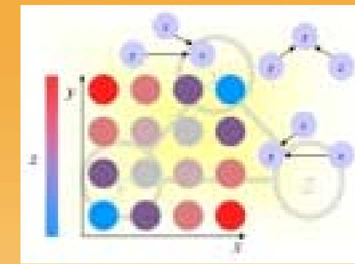
The Causality Workbench



Our methodology:

- Collecting donations or real data
- Acquiring or designing good simulators of real systems
  - – Trained with real data
  - – Used in the field to simulate systems, or
  - – Including real data + artificial “probe” variables
- Defining tasks with well defined objectives

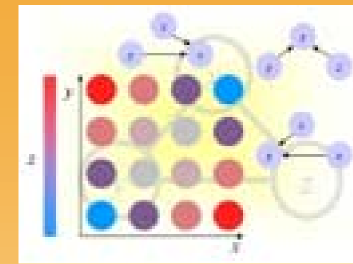
To benchmark algorithms, we built a …

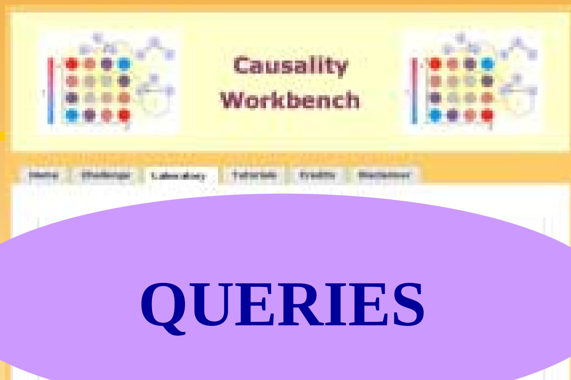


Virtual Lab

http://clopinet.com/causality

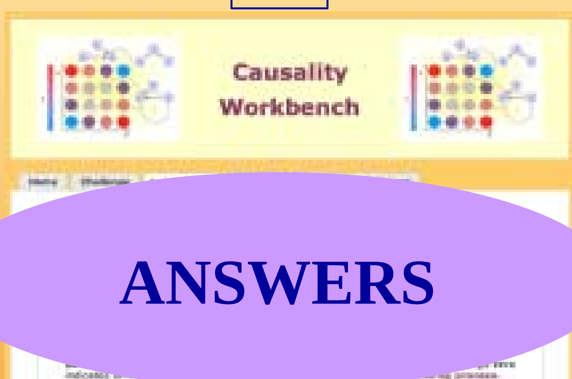
Virtual Lab





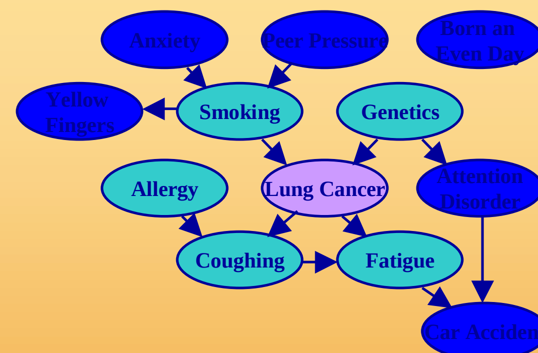
QUERIES

Database

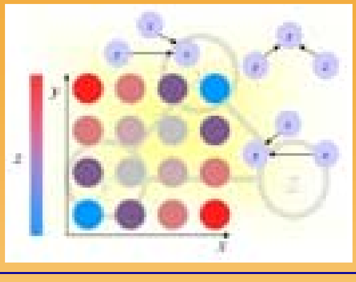


ANSWERS

Models of systems



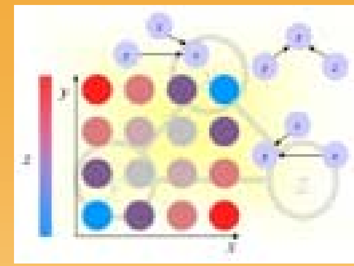
Virtual Lab



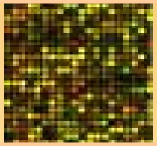
What we can do for you:

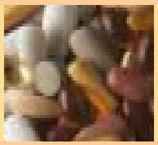
- Let you intervene on the system
  - Perform virtual experiments
- Serve you the data you want
  - For a virtual cash fee
- Include
  - Real data
  - Semi-artificial data
  - Simulated data

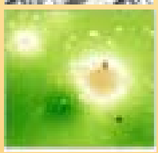
Causation and Prediction challenge



Challenge datasets

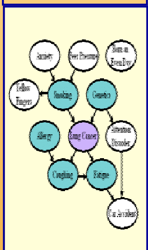






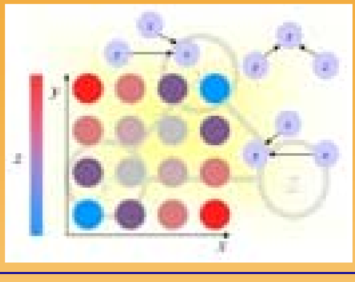
Dataset (click for info)	Description	Test data	Variables (num)	Target	Training examples	Test examples	Download (text format)	Download (Matlab format)
REGED0	Genomics re-simulated data	Not manipulated	Numeric (999)	Binary	500	20000	31 MB	25 MB
REGED1	Genomics re-simulated data	Manipulated (see list of manipulated variables)	Numeric (999)	Binary	500	20000	31 MB	25 MB
REGED2	Genomics re-simulated data	Manipulated	Numeric (999)	Binary	500	20000	31 MB	25 MB
SIDO0	Pharmacology real data w. probes	Not manipulated	Binary (4932)	Binary	12678	10000	12 MB	14 MB
SIDO1	Pharmacology real data w. probes	Manipulated	Binary (4932)	Binary	12678	10000	12 MB	14 MB
SIDO2	Pharmacology real data w. probes	Manipulated	Binary (4932)	Binary	12678	10000	12 MB	14 MB
CINA0	Census real data w. probes	Not manipulated	Mixed (132)	Binary	16033	10000	1 MB	1 MB
CINA1	Census real data w. probes	Manipulated	Mixed (132)	Binary	16033	10000	1 MB	1 MB
CINA2	Census real data w. probes	Manipulated	Mixed (132)	Binary	16033	10000	1 MB	1 MB
MARTI0	Genomics re-simulated data w. noise	Not manipulated	Numeric (1024)	Binary	500	20000	47 MB	35 MB
MARTI1	Genomics re-simulated data w. noise	Manipulated	Numeric (1024)	Binary	500	20000	47 MB	35 MB
MARTI2	Genomics re-simulated data w. noise	Manipulated	Numeric (1024)	Binary	500	20000	47 MB	35 MB

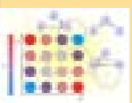
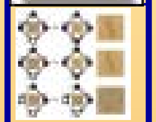
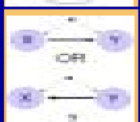
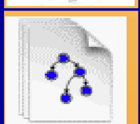
Toy datasets



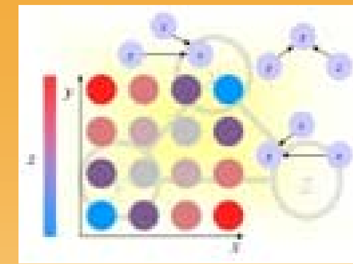
Dataset (click for info)	Description	Test data	Variables (num)	Target	Training examples	Test examples	Download (text format)	Download (Matlab format)
LUCAS0	Toy medicine data	Not manipulated	Binary (11)	Binary	2000	10000	33 KB	22 KB
LUCAS1	Toy medicine data	Manipulated	Binary (11)	Binary	2000	10000	33 KB	23 KB
LUCAS2	Toy medicine data	Manipulated	Binary (11)	Binary	2000	10000	31 KB	32 KB
LUCAP0	Toy medicine data w. probes	Not manipulated	Binary (143)	Binary	2000	10000	341 KB	262 KB
LUCAP1	Toy medicine data w. probes	Manipulated	Binary (143)	Binary	2000	10000	342 KB	654 KB
LUCAP2	Toy medicine data w. probes	Manipulated	Binary (143)	Binary	2000	10000	342 KB	263 KB

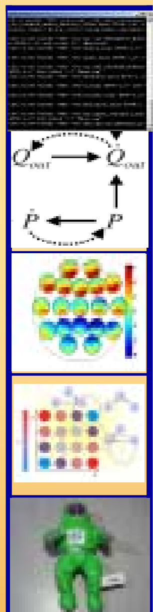
Pot-Luck challenge



	Task	Views	Type	Time dep.
	CYTO	609	real, self eval	
	LOCANET	1372	real, artif	
	PROMO	862	artif, self eval	✓
	SIGNET	918	artif	✓
	TIED	551	artif	
	CauseEffectPairs	580	real	
	Stemmatology	372	real, self eval	

Other donated datasets

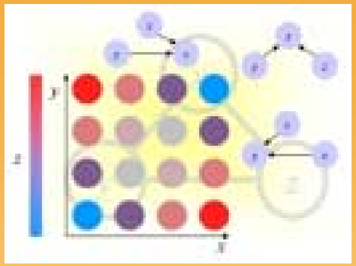


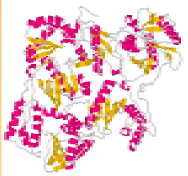


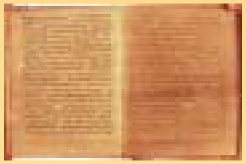
Task	Views	Type	Time dep.
WebLogs	272	real, self eval	
MIDS	232	artif	✓
NOISE	247	real, artif	✓
SECOM	297	real	✓
SEFTI	280	real	✓

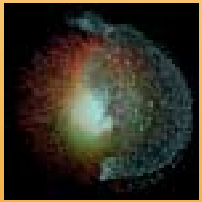
http://clopinet.com/causality

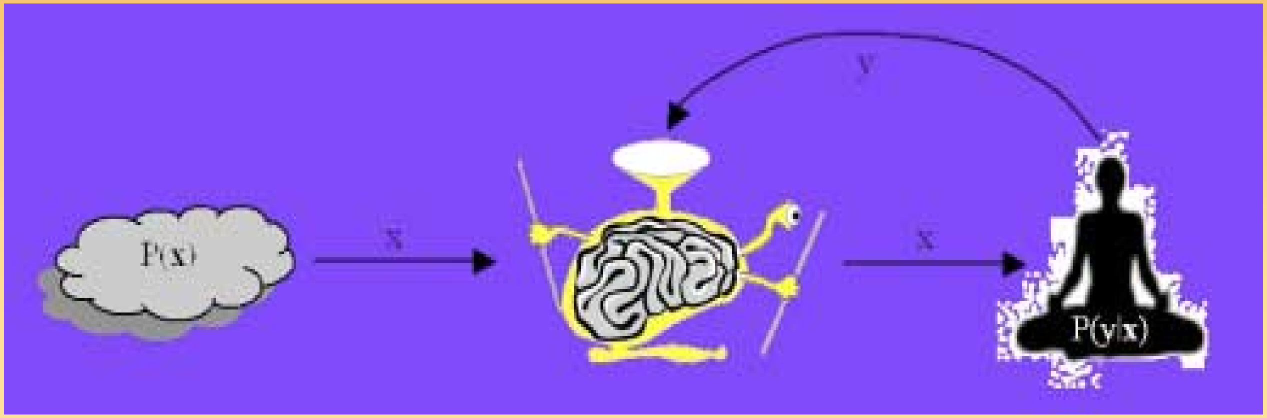
Active Learning Challenge







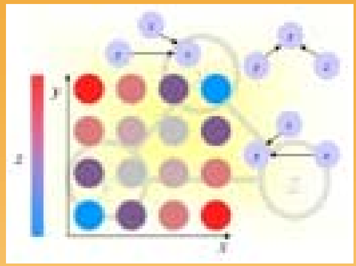




Dataset	Domain	Feat. Type	Feat. num.	Sparsity %	Missing %	Label	Train num.	Test num.	Positive labels %	Seed	Data (zip)	Data (Matlab)
HIVA	Chemo-informatics	binary	1617	90.88	0	binary	21339	21339	3.52	1	5.9 MB	9.3 MB
IBN_SINA	Handwriting recognition	mixed	92	80.67	0	binary	10361	10361	6.53	4	346 KB	537 KB
NOVA	Text processing	binary	16969	99.67	0	binary	9733	9733	28.45	11	2.3 MB	2.3 MB
ORANGE	Marketing	mixed	230	9.57	65.46	binary	25000	25000	7.34	54	6.8 MB	6.4 MB
SYLVA	Ecology	mixed	216	77.88	0	binary	72626	72626	6.15	4	14.5 MB	20.2 MB
ZEBRA	Embryology	continuous	154	0.04	0.004	binary	30744	30744	4.58	23	28.6 MB	53.2 MB

http://clopinet.com/al

Next: Causality and Time Series



With your help:

- Get more datasets
  - of practical and scientific interest
- Get good simulators of real systems
  - paired with the real datasets
- Define tasks and objectives
  - and practical challenge protocols

Organize the next challenge!