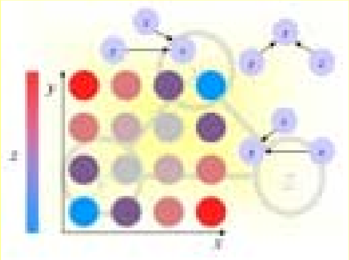
The LOCANET task

(Pot-luck challenge, NIPS 2008)

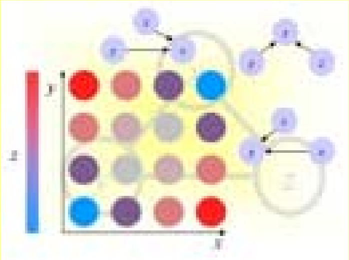
Isabelle Guyon, Clopinet
Alexander Statnikov, Vanderbilt Univ.
Constantin Aliferis, New York University

Acknowledgements



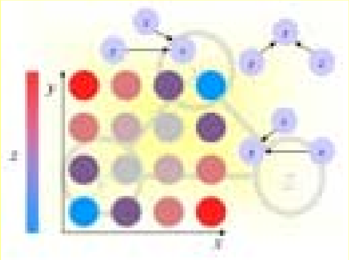
- This work is part of the causality workbench effort of data exchange and benchmark.
- *André Elisseeff and Jean-Philippe Pellet, IBM Zürich, Gregory F. Cooper, Pittsburg University, and Peter Spirtes, Carnegie Mellon*

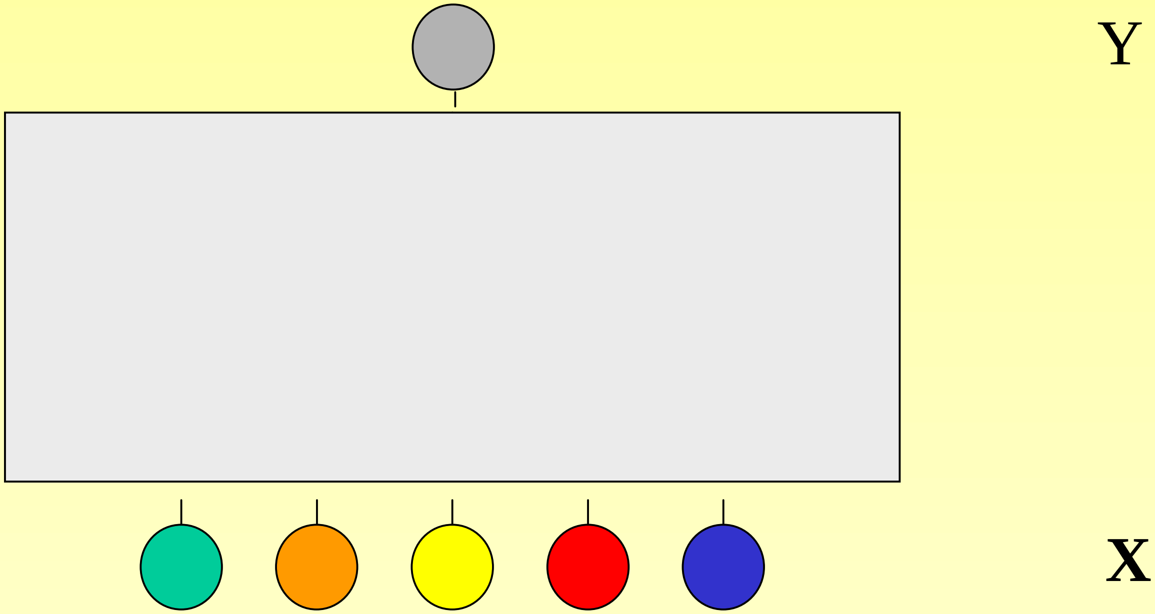
  collaborated on the design of the "causation nd prediction challenge" (WCCI 2008) in which the datasets of the LOCANET task were first used.



Problem description

Feature Selection

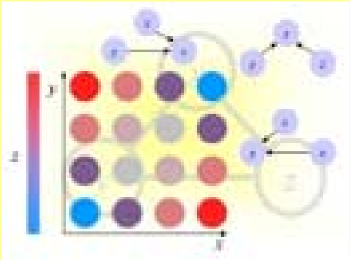


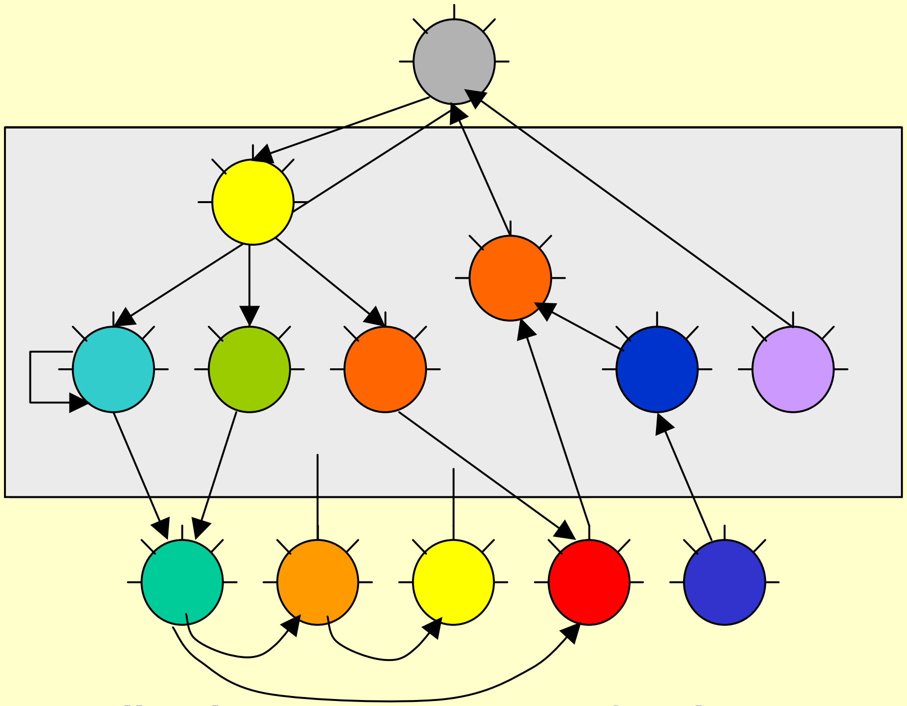


Predict Y from features X_1, X_2, …

Select most predictive features.

Causation



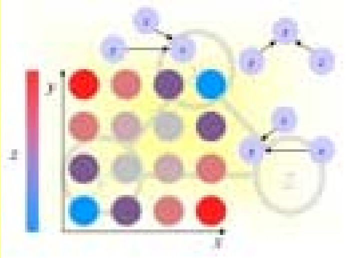


X

Predict the consequences of actions:

Under “manipulations” by an external agent, some features are no longer predictive.

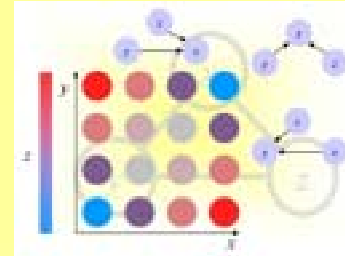
The LOCANET tasks

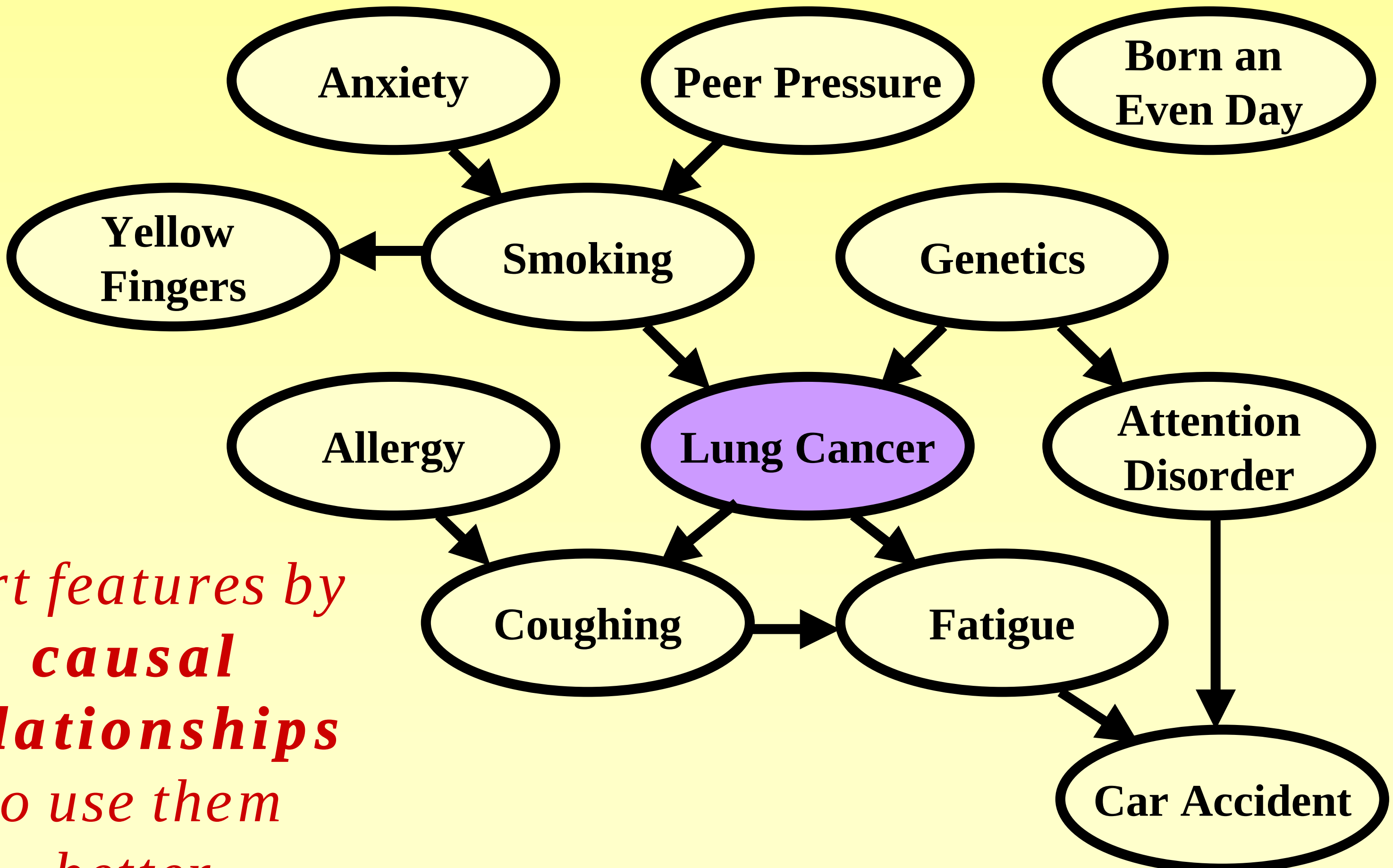


- ***LOCANET** stands for **LOcal CAusal NETwork***
- *Same datasets as Causation and Prediction challenge.*
- *Different goal: find a **depth 3 causal network** around the target (oriented graph structure).*

Challenge	Causation and Prediction	LOCANET (pot-luck)
Task	Predict a target variable in manipulated data	Find the local causal structure around the target
Data	- Un-manipulated training data - Manipulated test data	Only un-manipulated training data

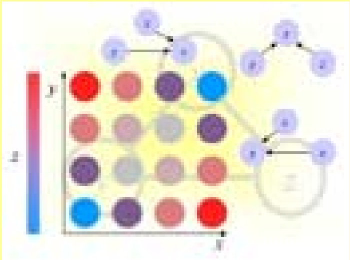
Why LOCANET?

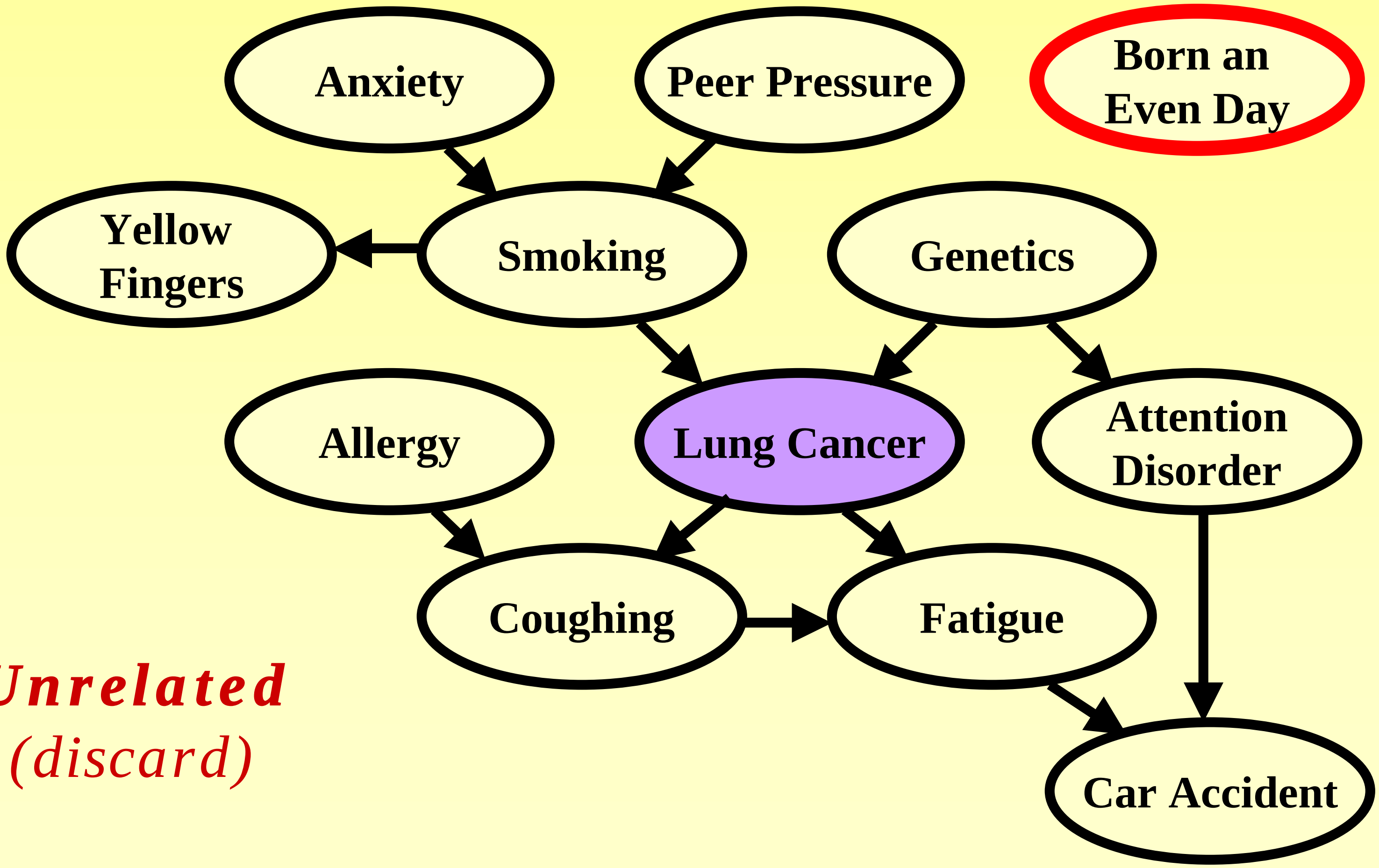




*Sort features by **causal relationships** to use them better*

Why LOCANET?

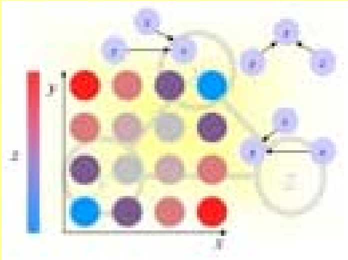


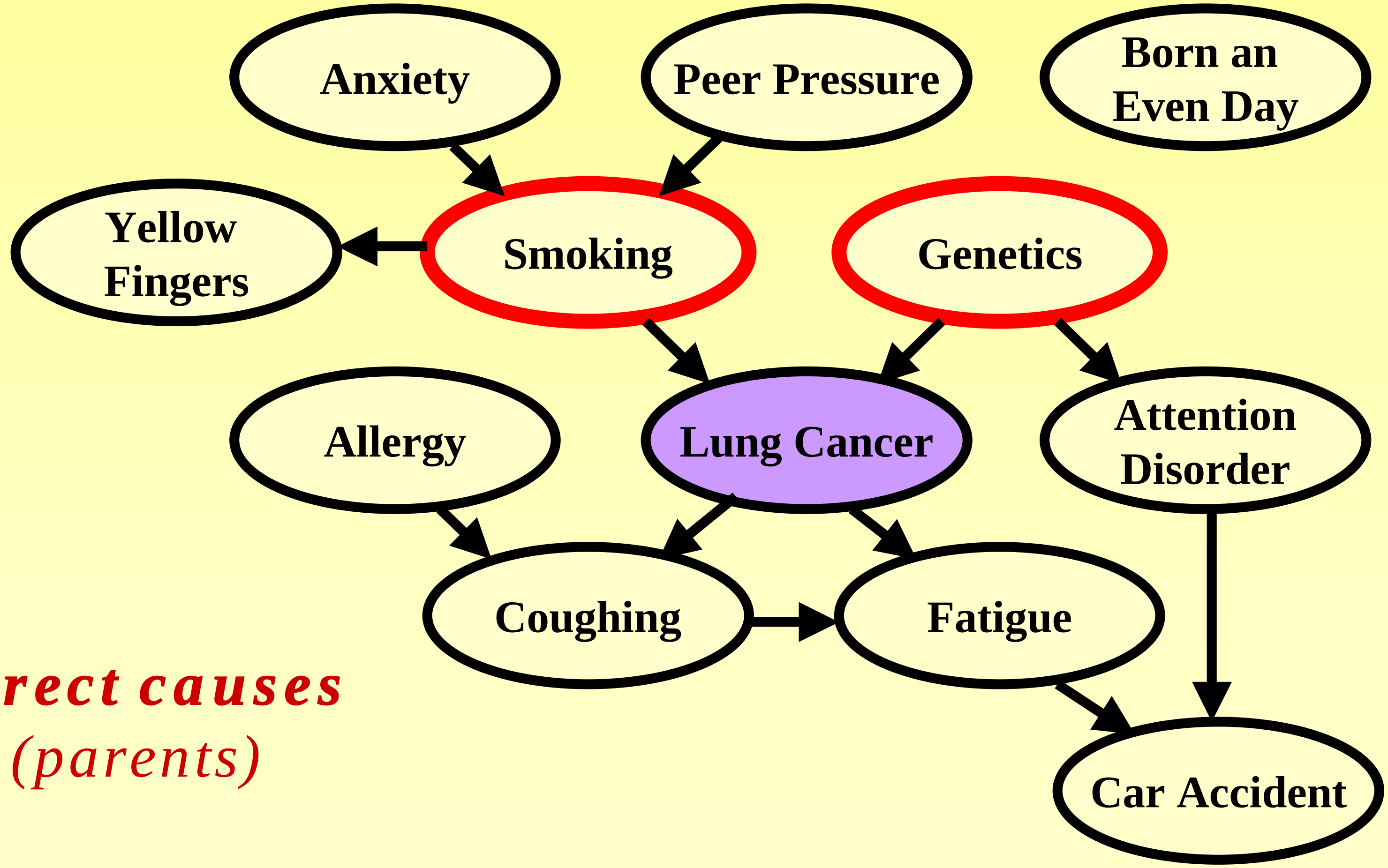


Unrelated

(discard)

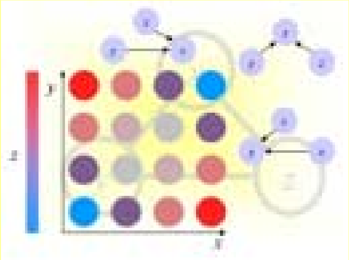
Why LOCANET?

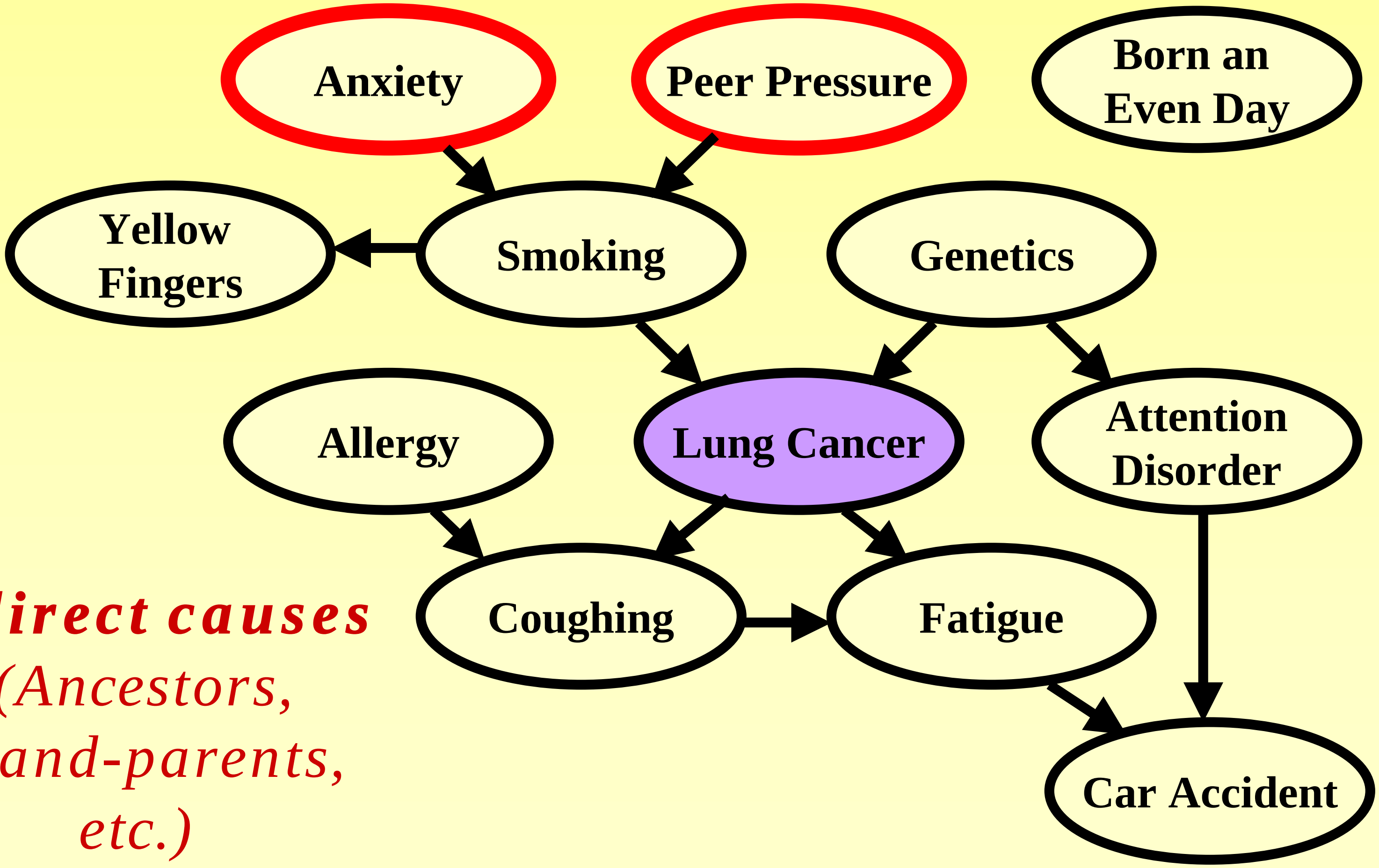




Direct causes
(parents)

Why LOCANET?

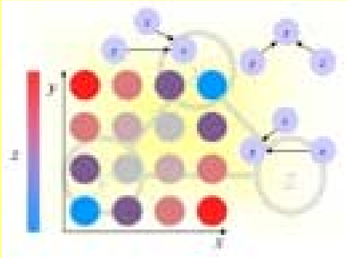


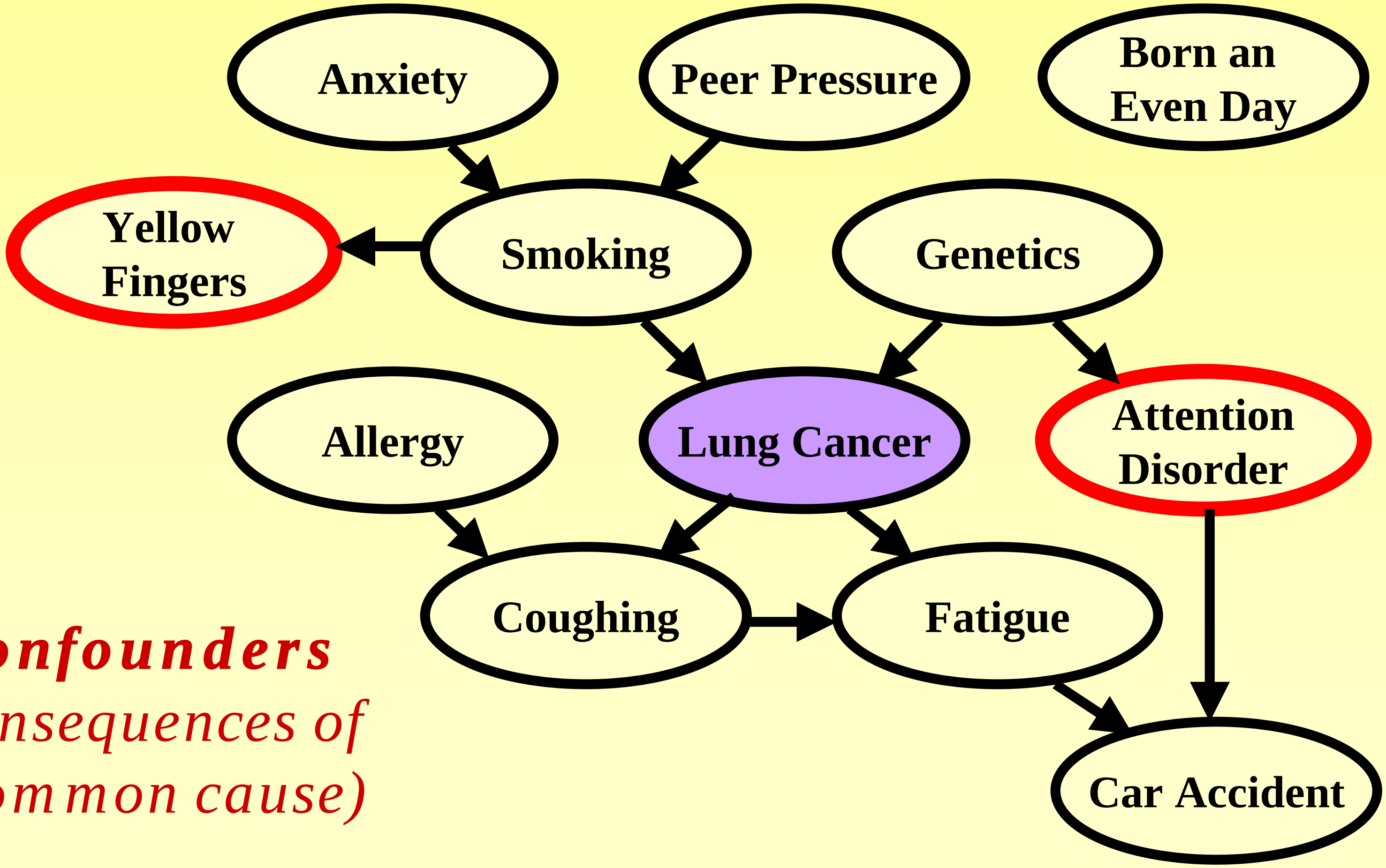


Indirect causes

(Ancestors, Grand-parents, etc.)

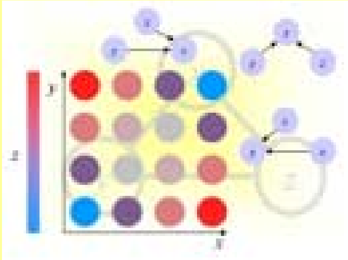
Why LOCANET?

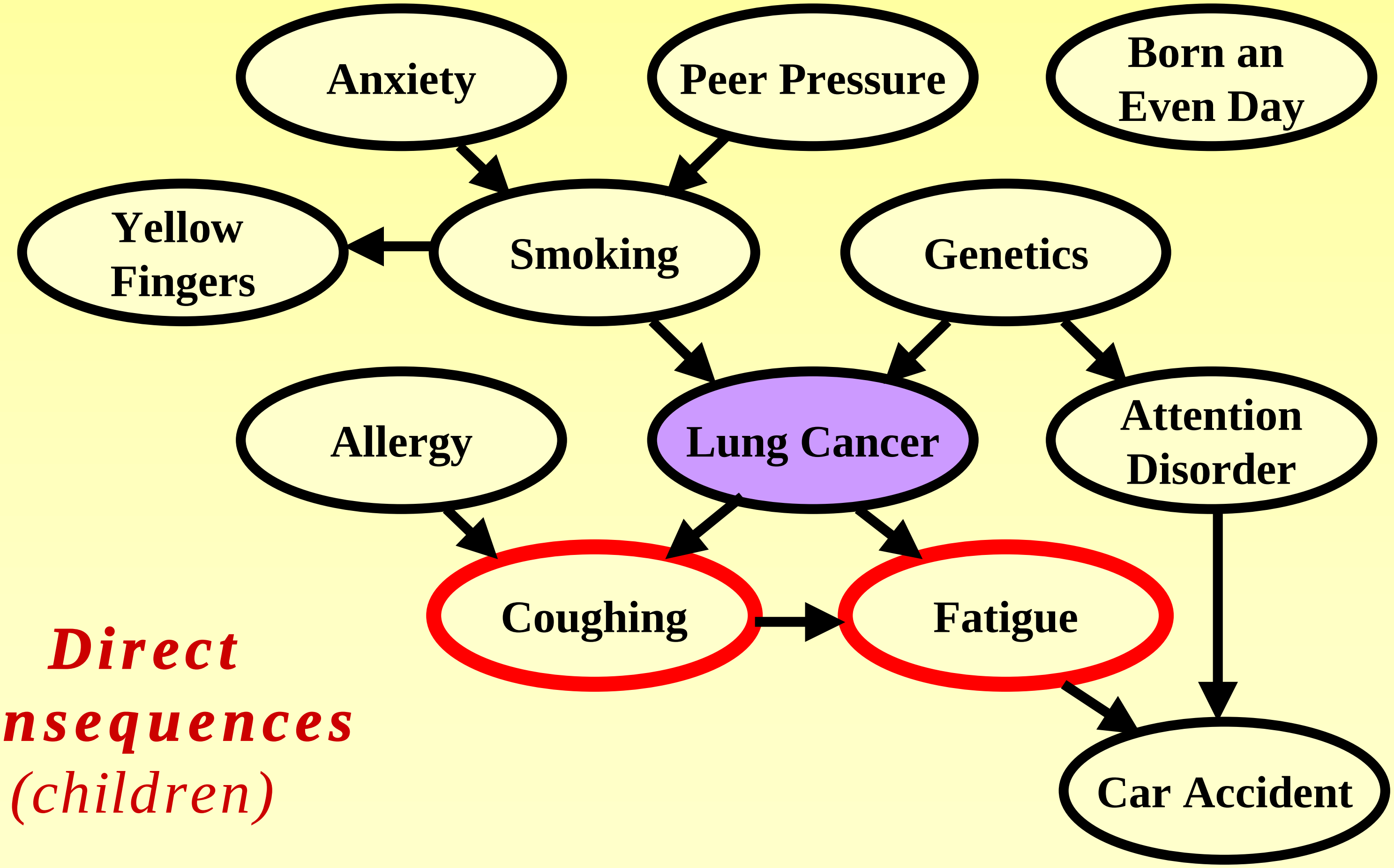




Confounders
(Consequences of a common cause)

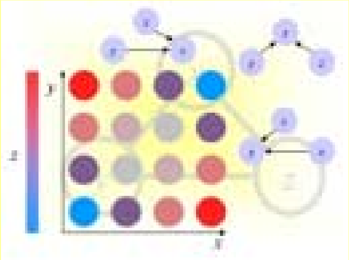
Why LOCANET?

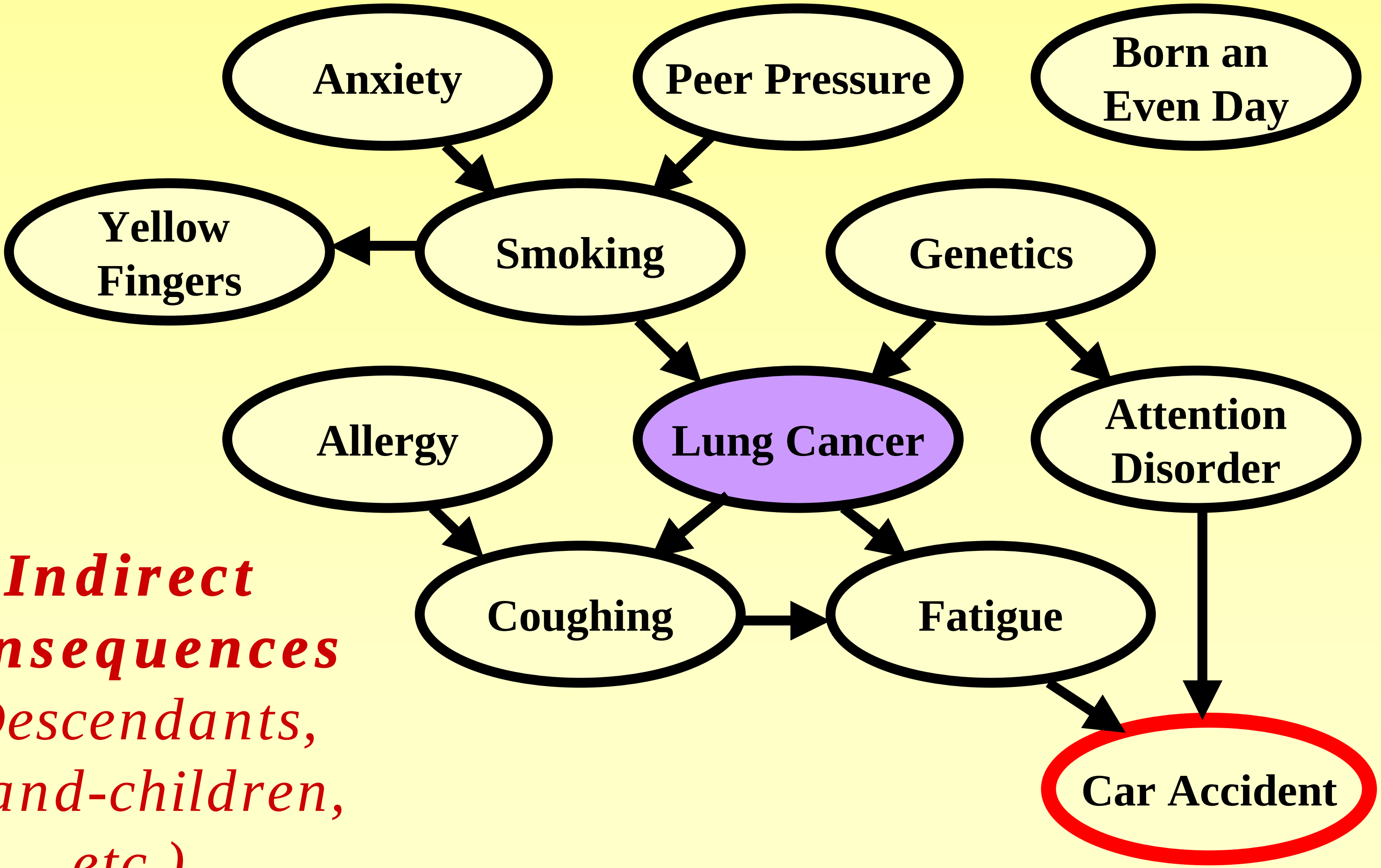




Direct consequences
(children)

Why LOCANET?

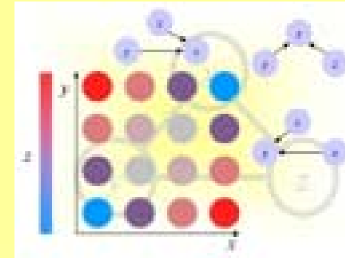


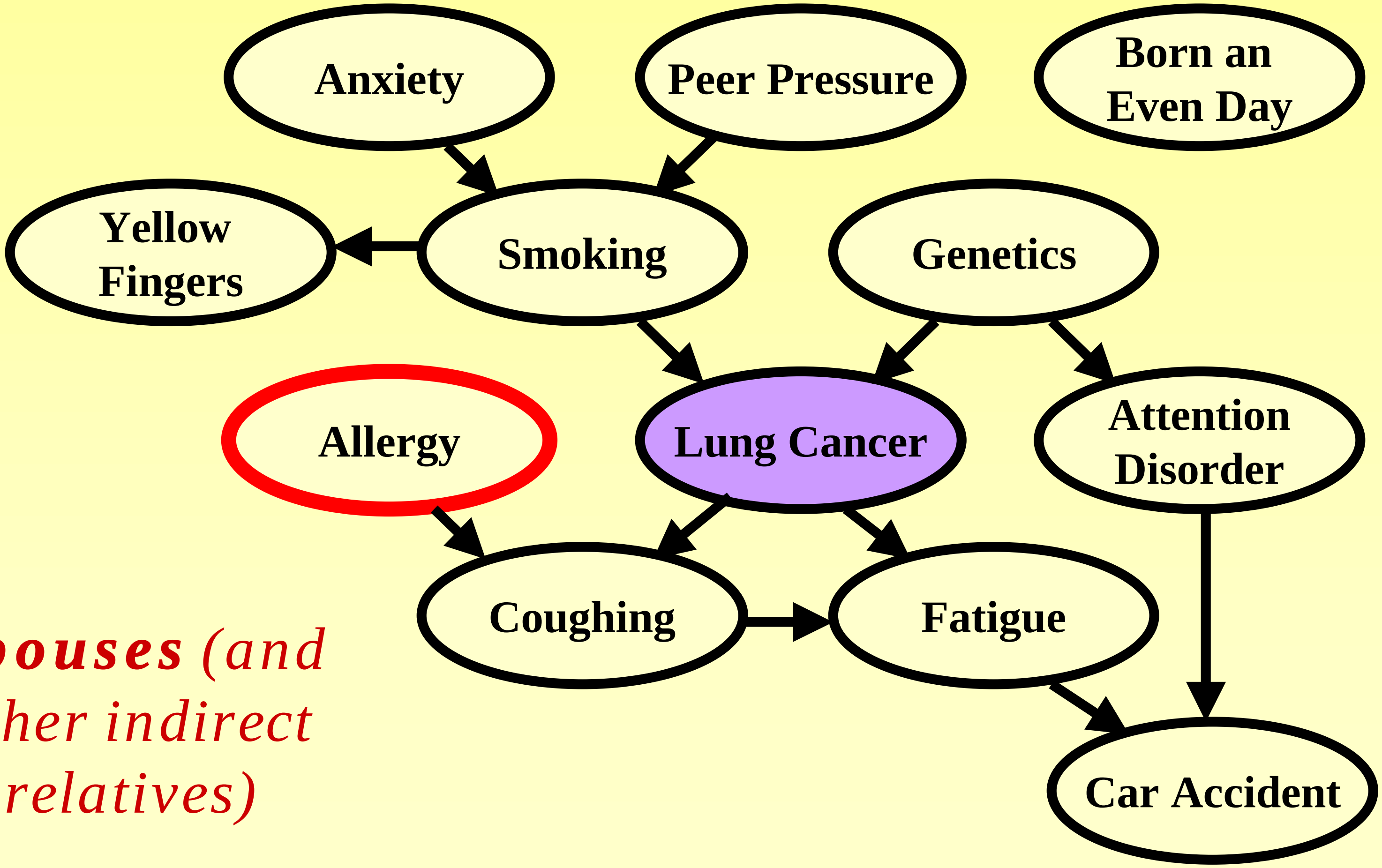


Indirect consequences

(Descendants, Grand-children, etc.)

Why LOCANET?



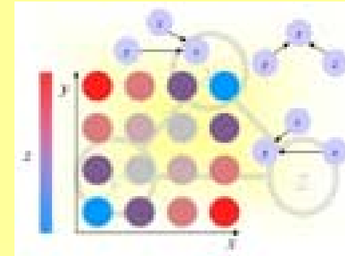


***Spouses** (and other indirect relatives)*

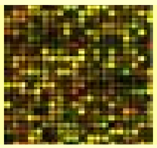
Datasets

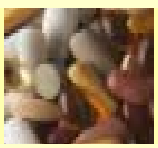
Four challenge datasets

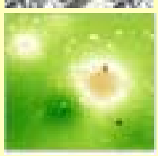
all with binary target variables (classification)



Challenge datasets

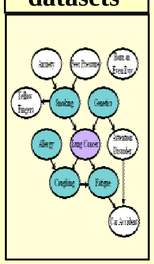






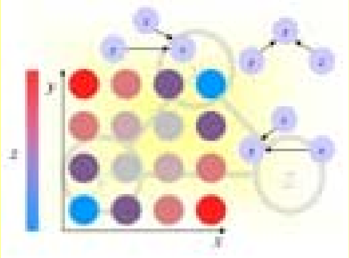
Dataset	**Description**	**Var. type**	**Var. num.**	**Tr. num.**
REGED	Lung cancer (re-simulated)	Numeric	999	500
SIDO	Drug discovery (real w. probes)	Binary	4932	12678
CINA	Marketing (real w. probes)	Mixed	132	16023
MARTI	Lung cancer (re-simulated)	Numeric	1024	500

Toy datasets



LUCAS	Toy medicine data (simulated)	Binary	11	2000
LUCAP	Toy medicine data (simul. w. probes)	Binary	143	2000

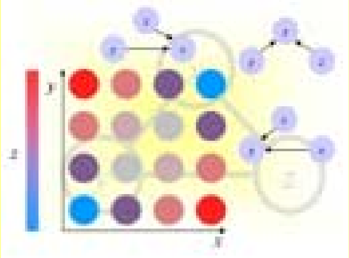
Difficulties

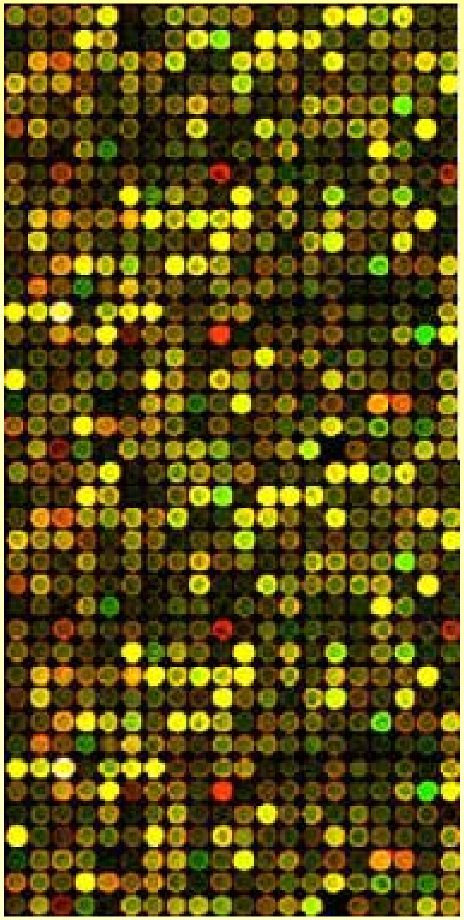


- **Violated assumptions**:
  - Causal sufficiency
  - Markov equivalence
  - Faithfulness
  - Linearity
  - “Gaussianity”
- **Overfitting** (statistical complexity):
  - Finite sample size
- **Algorithm efficiency** (computational complexity):
  - Thousands of variables
  - Tens of thousands of examples

REGED

REsimulated Gene Expression Dataset

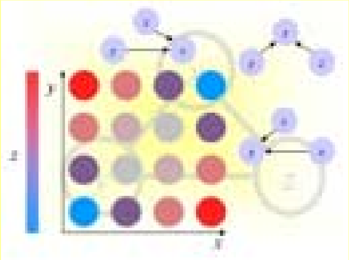


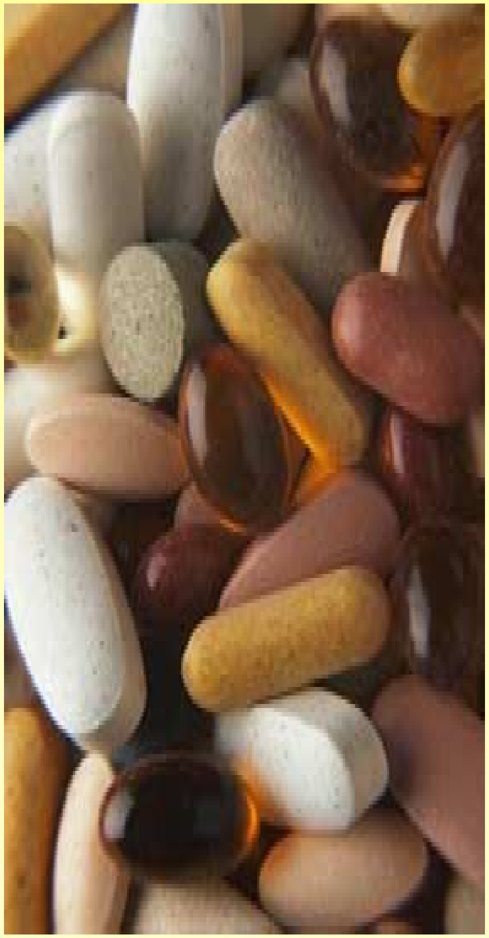


- **GOAL:** Find genes responsible of lung cancer (separate causes from consequences and confounders).
- **DATA TYPE:** “Re-simulated”, *i.e.* generated by a model derived from real human lung-cancer microarray gene expression data.
- **DATA TABLE:** of dim (P, N):
  - N=999 numeric features (gene expression coefficients) and 1 binary target (separating malignant adenocarcinoma samples from control squamous cell samples).
  - P=500 training examples.

SIDO

SImple Drug Operation mechanisms

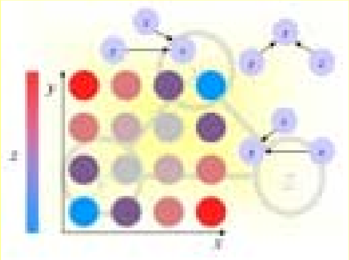




- **GOAL:** Pharmacology problem: uncover mechanisms of action of molecules (separate causes from confounders). This would help chemists in the design of new compounds, retaining activity, but having other desirable properties (less toxic, easier to administer).
- **DATA TYPE:** Real plus artificial probes.
- **DATA TABLE:** of dim (P, N):
  - N=4932 binary features (QSAR molecular descriptors generated programmatically and artificial probes) and 1 binary target (molecular activity against HIV virus).
  - P=12678 training examples.

CINA

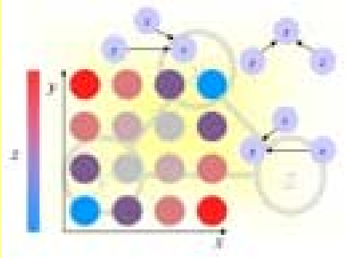
Census is Not Adult

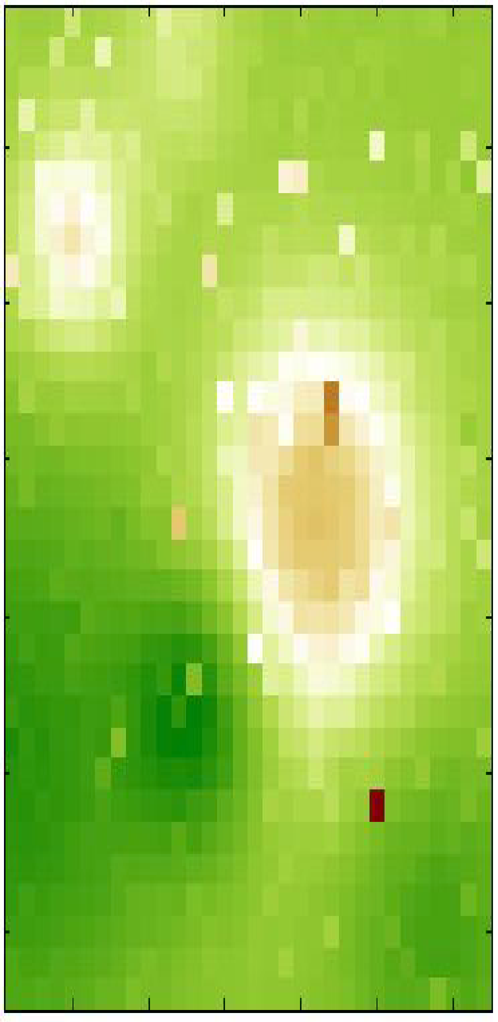


- **GOAL:** Uncover the socio-economic factors (age, workclass, education, marital status, occupation, native country, etc.) affecting high income (separate causes from consequences and confounders).
- **DATA TYPE:** Real plus artificial probes.
- **DATA TABLE:** of dim (P, N):
  - N=132 mixed categorical coded as binary, binary and numeric features (socio-economic factors and artificial probes) and 1 binary target whether the income exceeds 50K USD).
  - P=16023 training examples.

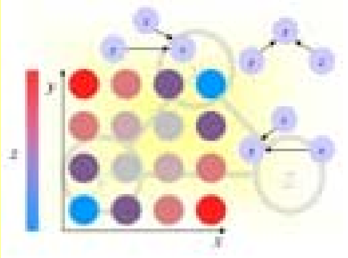
MARTI

Measurement ARTIfact



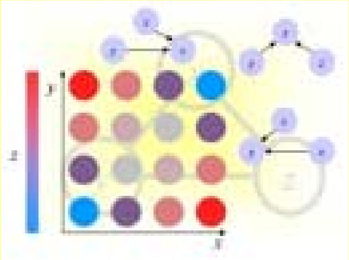


- **GOAL:** Find genes responsible of lung cancer (separate causes from consequences and confounders).
- **DATA TYPE:** Same as REGED (Re-simulated, generated by a model derived from real human lung-cancer microarray gene expression data) but with on top a noise model (correlated noise).
- **DATA TABLE:** of dim (P, N):
  - N=1024 numeric features (gene expression coefficients) and 1 binary target (malignant samples *vs.* control).
  - P=500 training examples.



Evaluation method

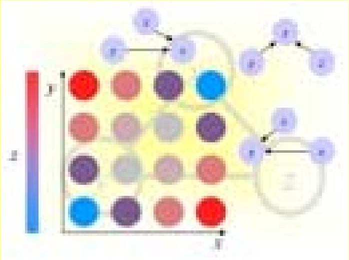
Result format



- Each feature is numbered according to its position in the data table (the target is 0).
- Provide a text file, each line containing a feature followed by a list of parents (up to 3 connections away from the target).
- Example: Guyon_LUCAS_feat.localgraph

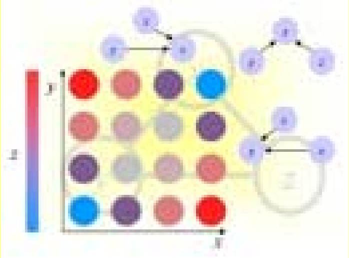
  0: 1 5
  1: 3 4
  2: 1
  6: 5
  8: 6 9
  9: 0 11
  11: 0 10

Relationship to target



- We consider only **local directed acyclic graphs**. We encode the relationship as a string of **up (u)** and **down (d)** arrows, from the target.
  - **Depth 1 relatives:** parents (u) and children (d).
  - **Depth 2 relatives:** spouses (du), grand-children (dd), siblings (ud), grand-parents (uu).
  - **Depth 3 relatives:** great-grand-parents (uuu), uncles/aunts (uud), nices/nephews (udd), parents of siblings (udu), spouses of children (ddu), parents in law (duu), children of spouses (dud), great-grand-children (ddd).
- If there are 2 paths, we prefer the shortest.
- If there are 2 same length paths, both are OK.

Score: average edit distance



- To compare the proposed local network to the true network, a **confusion matrix** C_{ij} is computed, recording the number of relatives confused for another type of relative, among the 14 types of relatives in depth 3 networks.

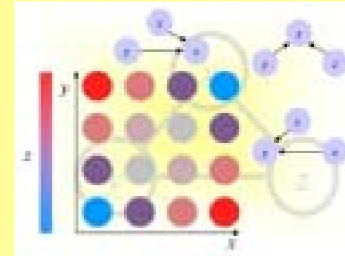
- A **cost matrix** A_{ij}, is applied to account for the distance between relatives (computed with an **edit distance** as the number of substitutions, insertions, or deletions to go from one string to the other).

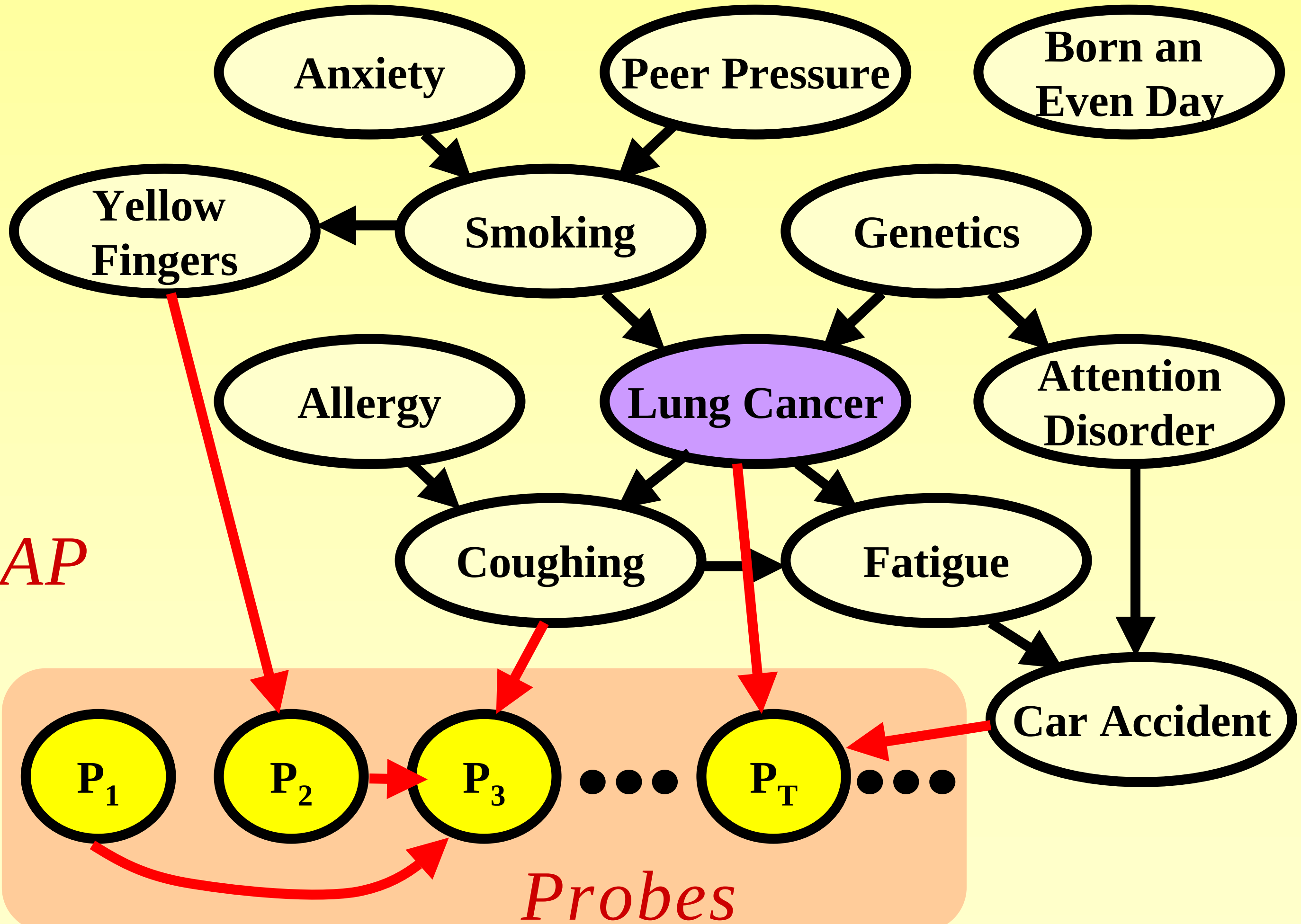
- The score of the solution is then computed as:

$$S = \mathrm{sum}_{ij}\ A_{ij}\ C_{ij}$$

Real data with probes

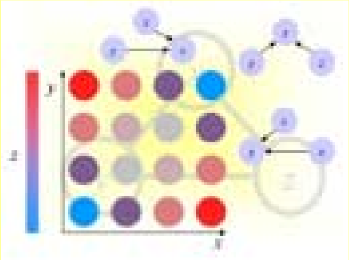
Using artificial "probes"





LUCAP

Evaluation using "probes"

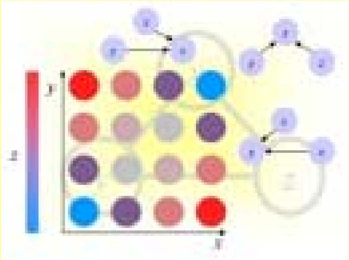


- We compute the score:

$$S = \text{sum}_{ij}\ A_{ij}\ C_{ij}$$

by summing only over probes.

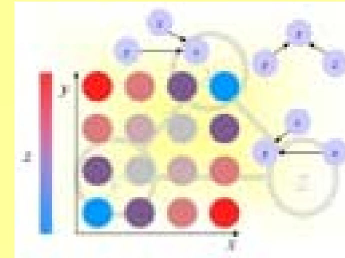
- We verify manually the plausibility of relationships between real variables.



Results

Result matrix

(probes only)



	LUCAS	LUCAP	REGED	SIDO	CINA	MARTI
Brown			0.27	3.46	2.23	0.36
De-Prado-Cumplido					3.27	
Dindar					1.70	
Engin				3.48		
Kirkagaclioglu					2.16	
Mwebaze	0.91	1.80	0.22	3.46	2.32	
Oguz					1.75	
Olsen			0.52		3.31	0.21
Tillman			0.34		1.74	
Wang			0.50	3.31	2.17	0.93
Reference A	0.09	1.09	0.01	0.64	0.64	0.02
Reference B	2.36	1.87	0.16	1.92	1.89	0.16
Reference C	2.09	1.43	3.08		1.67	3.01
Reference D	3.56	3.33	0.22	3.67	3.64	0.21

Reference A: Truth graph with 20% of the edges flipped at random.

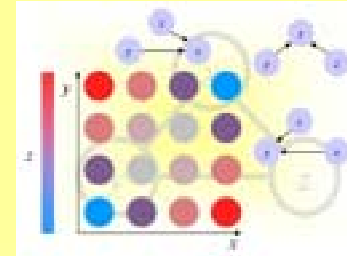
Reference B: Truth graph with connections symmetrized.

Reference C: Variables in the truth graph, fully connected.

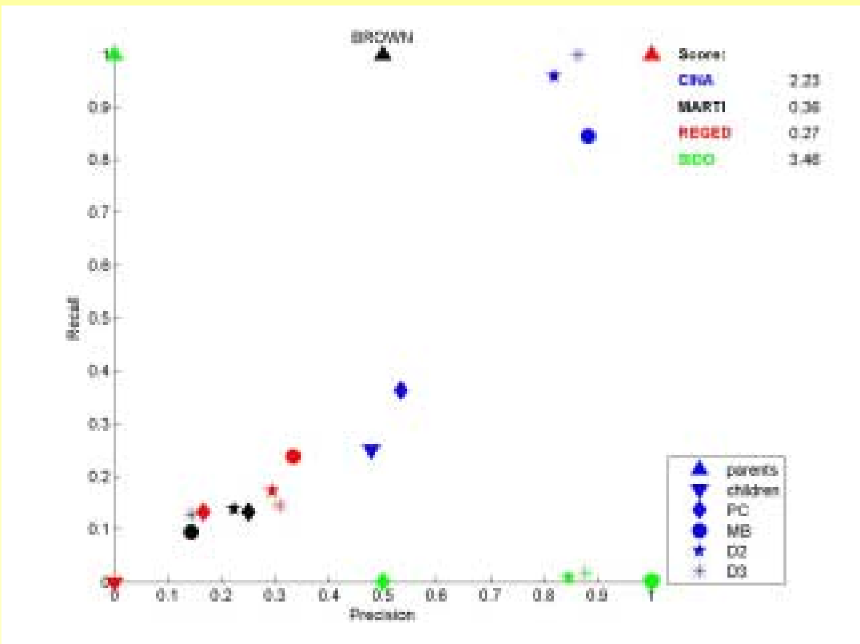
Reference D: Variables in the truth graph are all disconnected.

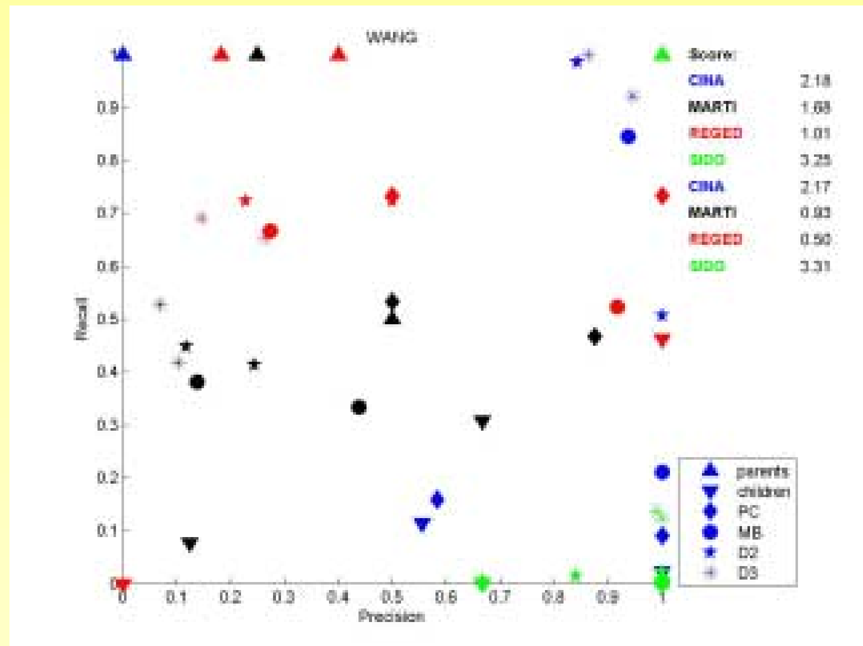
Precision & recall by entrant

(probes only)



http://www.causality.inf.ethz.ch/data/LOCANET.html



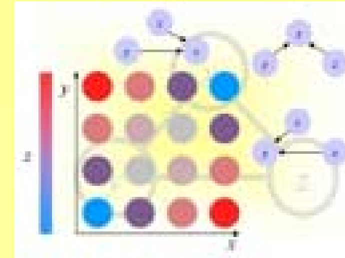


Precision: num_good_found / num_found

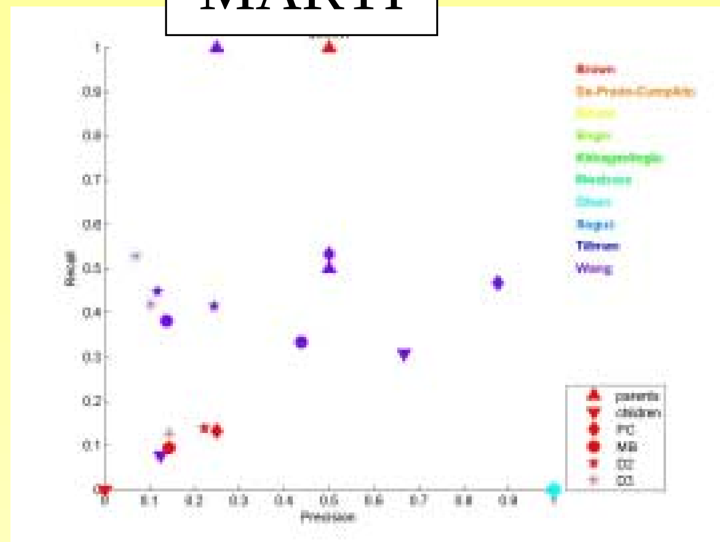
Recall: num_good_found / num_good

Precision & recall by dataset

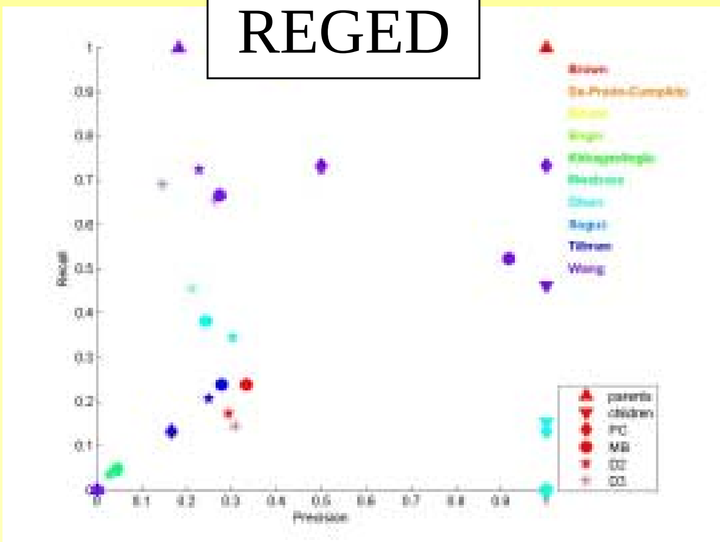
(probes only)



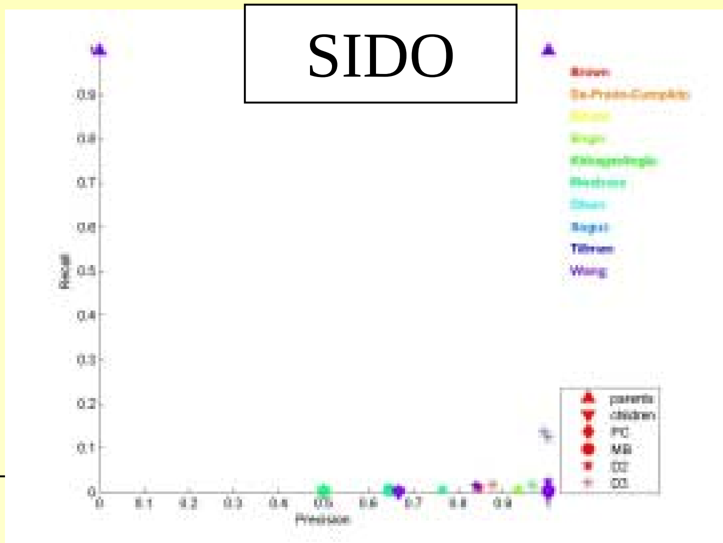
MARTI



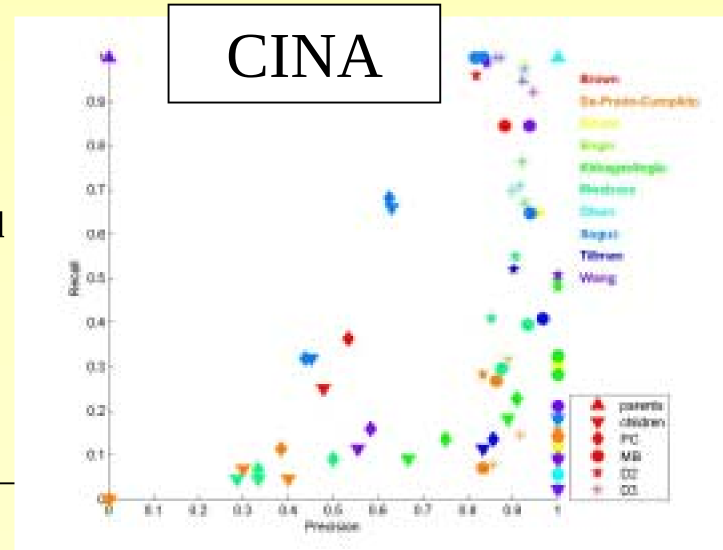
REGED



SIDO



CINA



Precision:

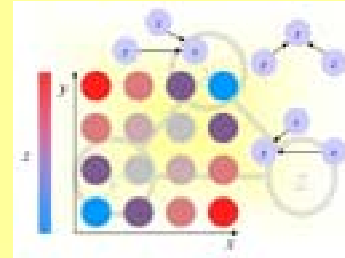
ngood_found / nfound

Recall:

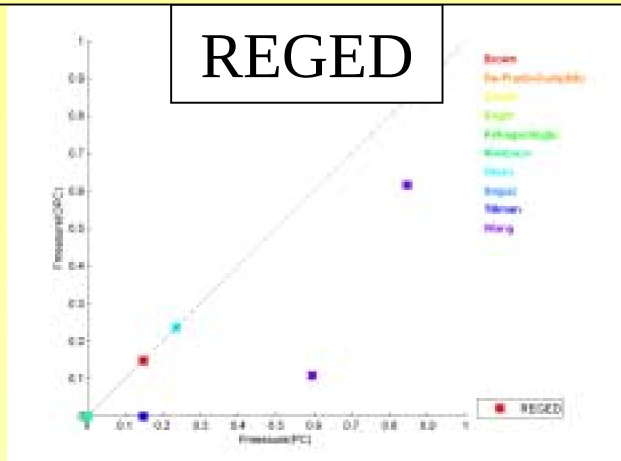
ngood_found / ngood

Fmeasure=2PR/(P+R)

(probes only)

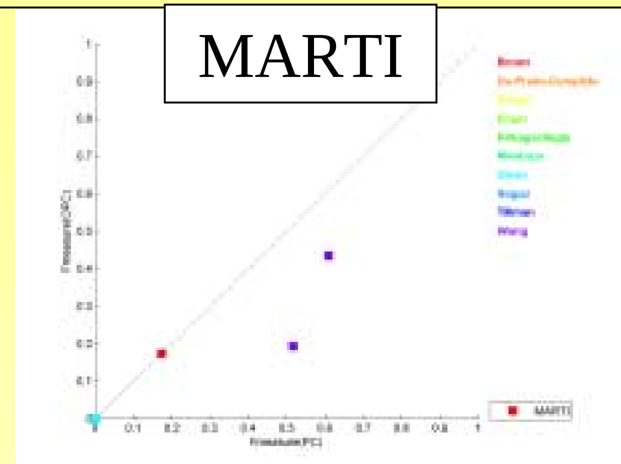


Oriented PC Fmeasure



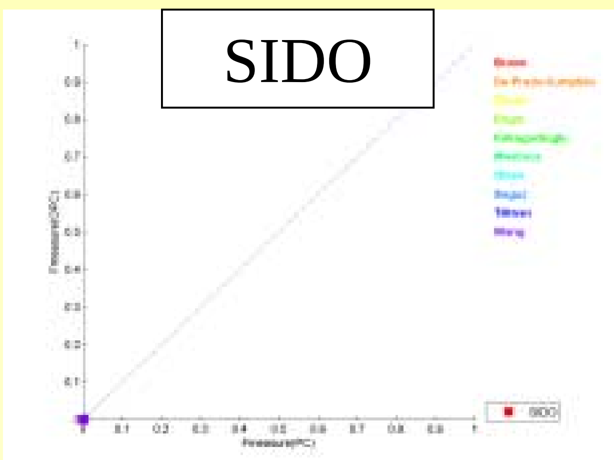
PC Fmeasure

Oriented PC Fmeasure



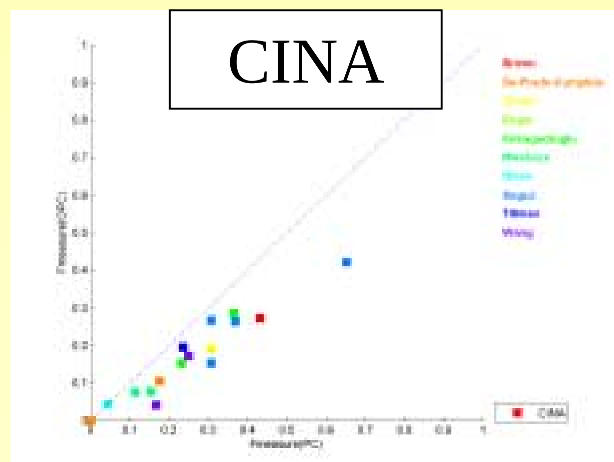
PC Fmeasure

Oriented PC Fmeasure



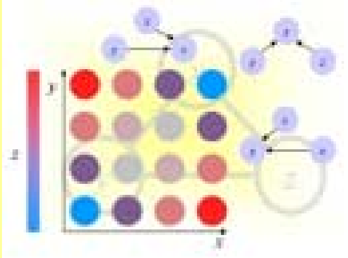
PC Fmeasure

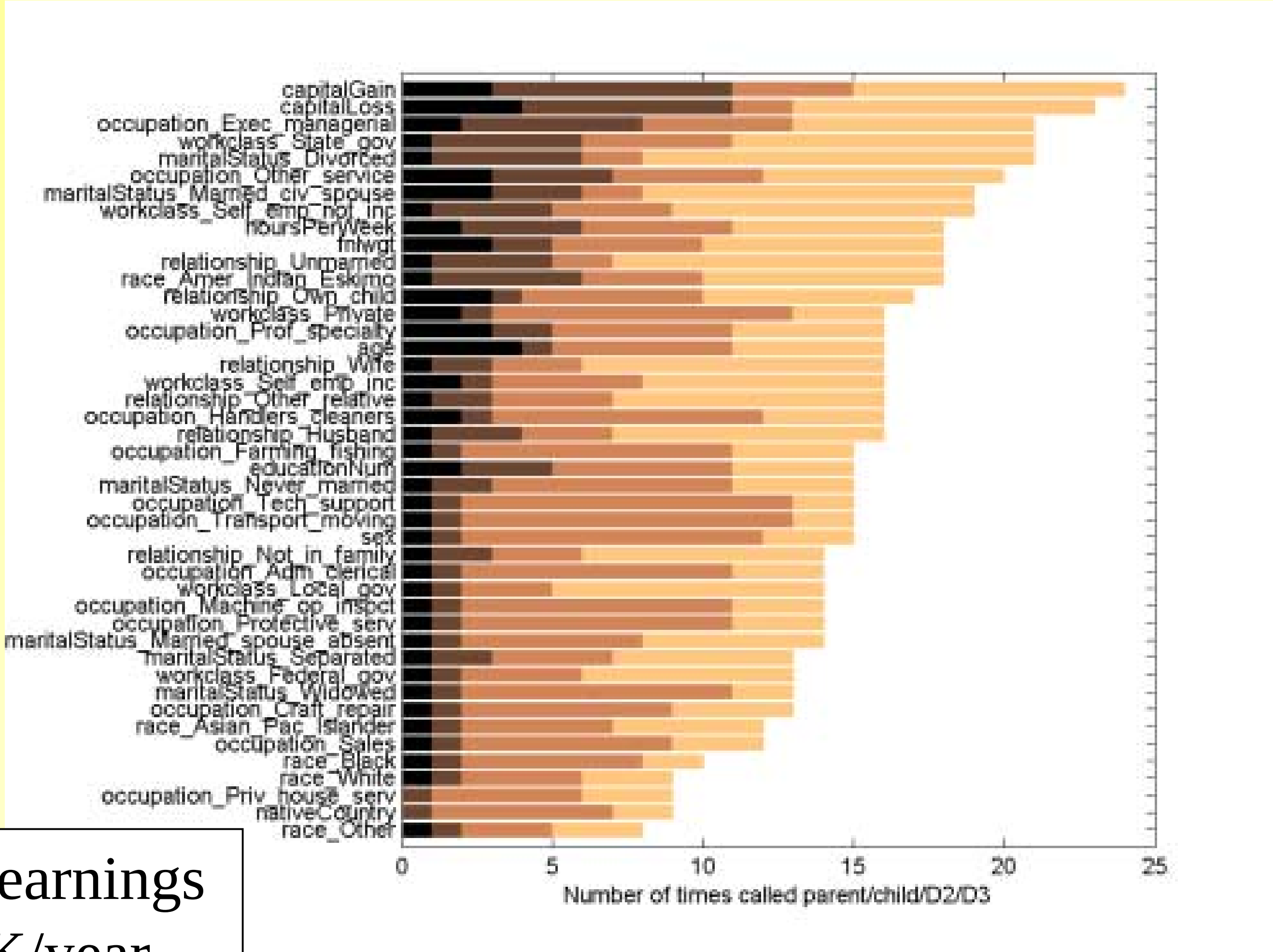
Oriented PC Fmeasure



PC Fmeasure

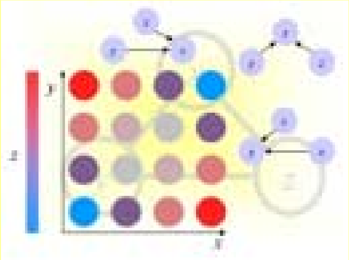
Real features on CINA





Target=earnings ≥$50K/year

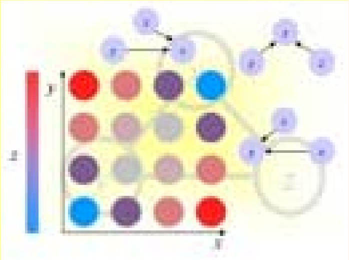
Does this make sense?



age C4 E1 corr= 0.24
occupation_Prof_specialty C3 E2 corr= 0.17
fnlwgt C3 E2 corr=-0.01
maritalStatus_Married_civ_spouse C3 E3 corr= 0.44
educationNum C2 E3 corr= 0.34
occupation_Other_service C3 E4 corr=-0.16
hoursPerWeek C2 E4 corr= 0.23
relationship_Unmarried C1 E4 corr=-0.14
workclass_Self_emp_not_inc C1 E4 corr= 0.02
capitalLoss C4 E7 corr= 0.14
race_Amer_Indian_Eskimo C1 E5 corr=-0.03
maritalStatus_Divorced C1 E5 corr=-0.13
workclass_State_gov C1 E5 corr= 0.01
occupation_Exec_managerial C2 E6 corr= 0.22 <-- ?? why an effect
capitalGain C3 E8 corr= 0.22

Most variables are cited more often as effects than as causes.

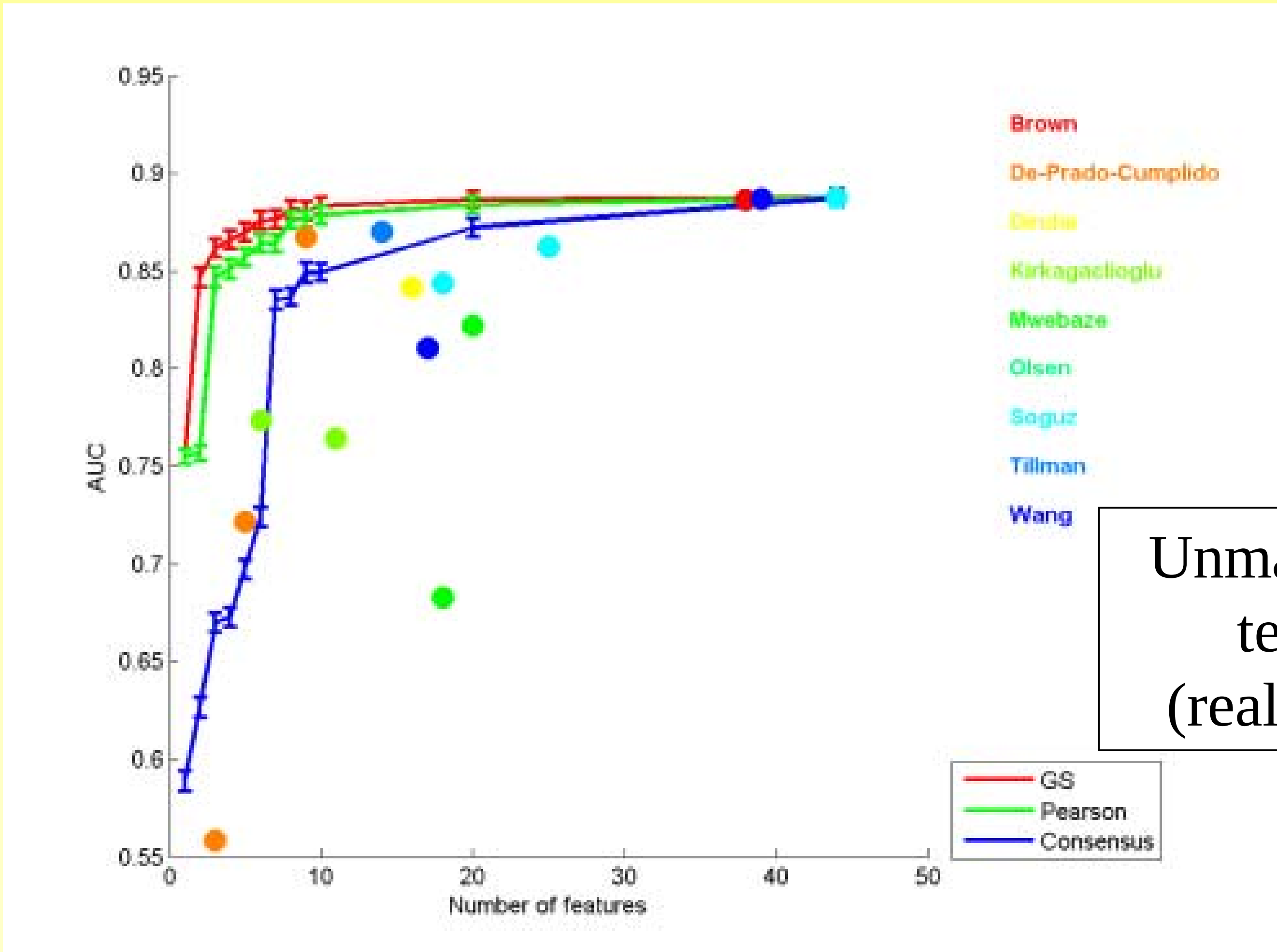
Most correlated features



maritalStatus_Married_civ_spouse
relationship_Husband
educationNum
maritalStatus_Never_married
age
hoursPerWeek
relationship_Own_child
capitalGain
sex
occupation_Exec_managerial
relationship_Not_in_family
occupation_Prof_specialty
occupation_Other_service
capitalLoss
relationship_Unmarried

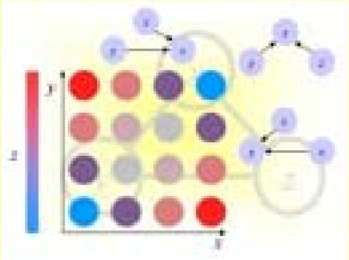
In red: found in the first ½ of the consensus ranking of the challenge.
In orange: tie with the feature exactly at the middle.

Most predictive feature sets



Unmanipulated test data (real features)

Forward feature selection

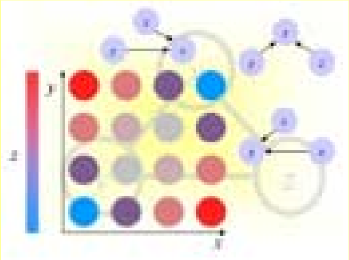


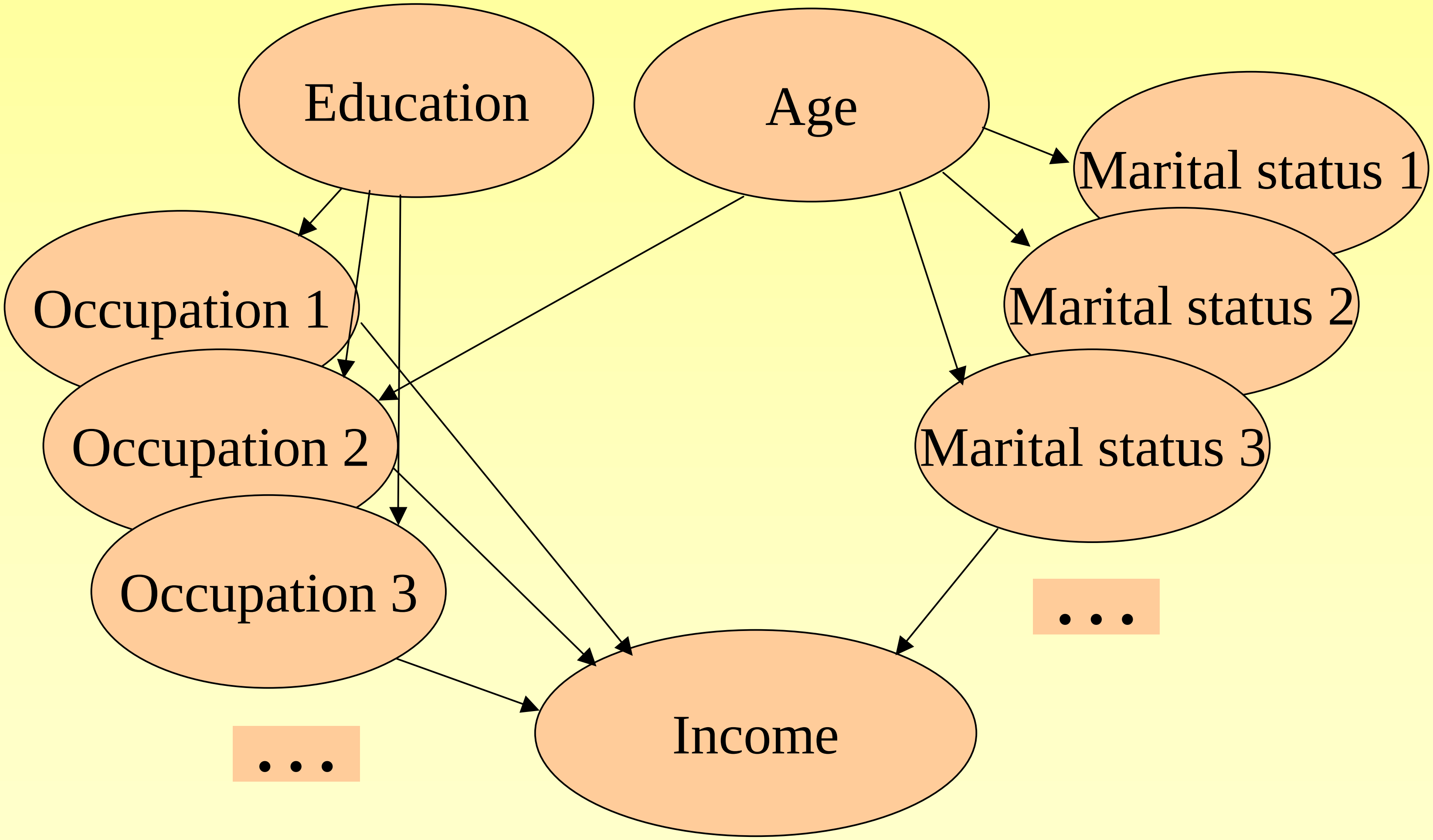
- Gram-Schmidt orthogonalization yields more predictive compact feature subsets than the empirical Markov blanket.
- Top GS features:

maritalStatus_Married_civ_spouse
educationNum
capitalGain
occupation_Exec_managerial
capitalLoss
age

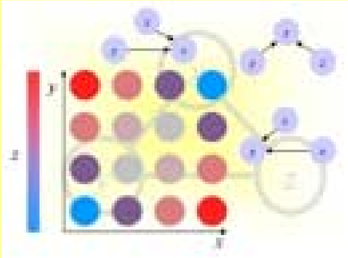
Not found in close neighborhood of the target

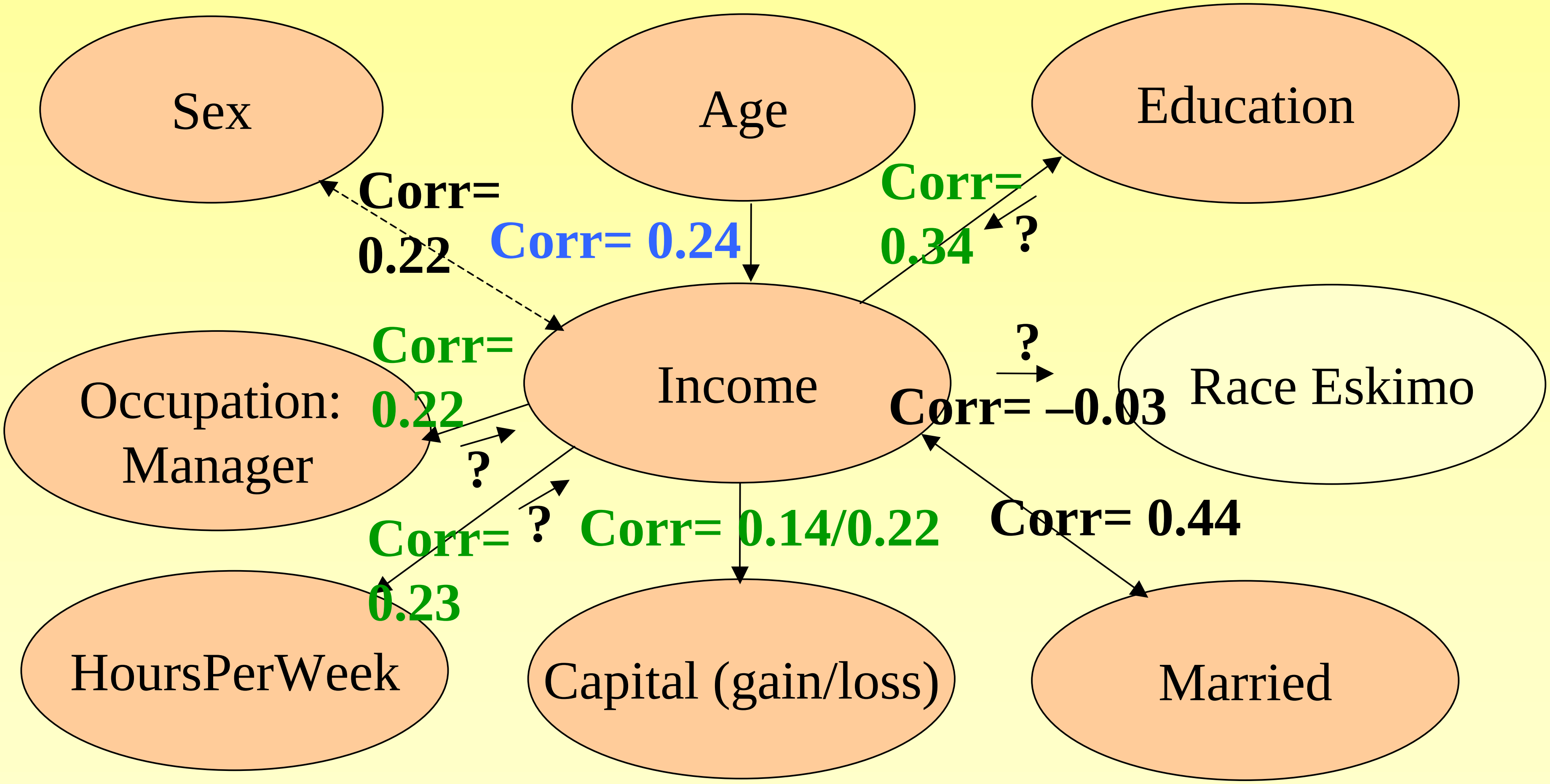
Explanation?



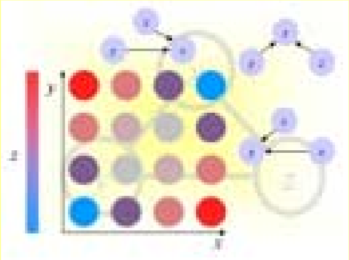


Some findings





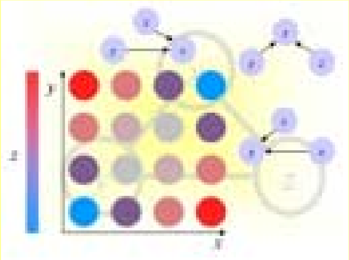
Methods employed



- **Structure learning (independence tests):**
  - **- Brown & Tsamardinos**
  - **- Zhou, Wang, Yin & Geng**
- **Mix of score-based and structure methods:**
  - **- de-Prado-Cumplido &Antonio Artes-Rodrigues**
  - **- Tillman & Ramsey**
- **Mix feature selection and structure methods:**
  - **- Olsen, Meyer & Bontempi**
- **Ensemble of method:**
  - **- Mwebaze and Quinn**

Structure learning gave most promising results
(highest precision, but poor recall)

Conclusion



- Dimensionality kills causal discovery (SIDO).
- Precision generally better than recall.
- Orientation inconsistent and not always plausible in real features across entries.
- Difficult to define a single good quantitative assessment metric.
- CINA offers opportunities to try more algorithms (without probes, without coding).