Very high dimensional causal structure and Markov boundary discovery: key algorithmic developments and the insights gained about the R&D process

Constantin F. Aliferis MD, PhD, FACMI

Professor of Medicine,

Chief Research Informatics Officer,

Director, Institute for Health Informatics,

University of Minnesota

Chief Analytics Officer, M-Health

Talk Motivation

- In 2000 sound and complete computational causal graph algorithms could be used with up to approx. 100 variables with conventional hardware.
- In 2015 analyses with more than 1,000,000 variables (for local graphs) and more than 10,000 variables (for complete graphs) are routine with very modest hardware.

Goals

(a) Summarize the extraordinary progress accomplished in the last 2 decades and where the field is.

(b) R&D process model we used, some insights about the discovery process, and a few empirical principles for developing and validating highly practical algorithms for causal discovery.

Caveats

(a) Emphasize:

local algorithms,

local-to-global,

Markov Boundary,

multiplicity and

experimentation minimization algorithms.

(b) Perspective heavily influenced by the work done in my group since 2000 (and our approach to such R&D).

Assumptions

Audience is familiar with:

- Key principles and applications of machine learning including predictive modeling, feature selection, probabilistic causal graphs/causal discovery

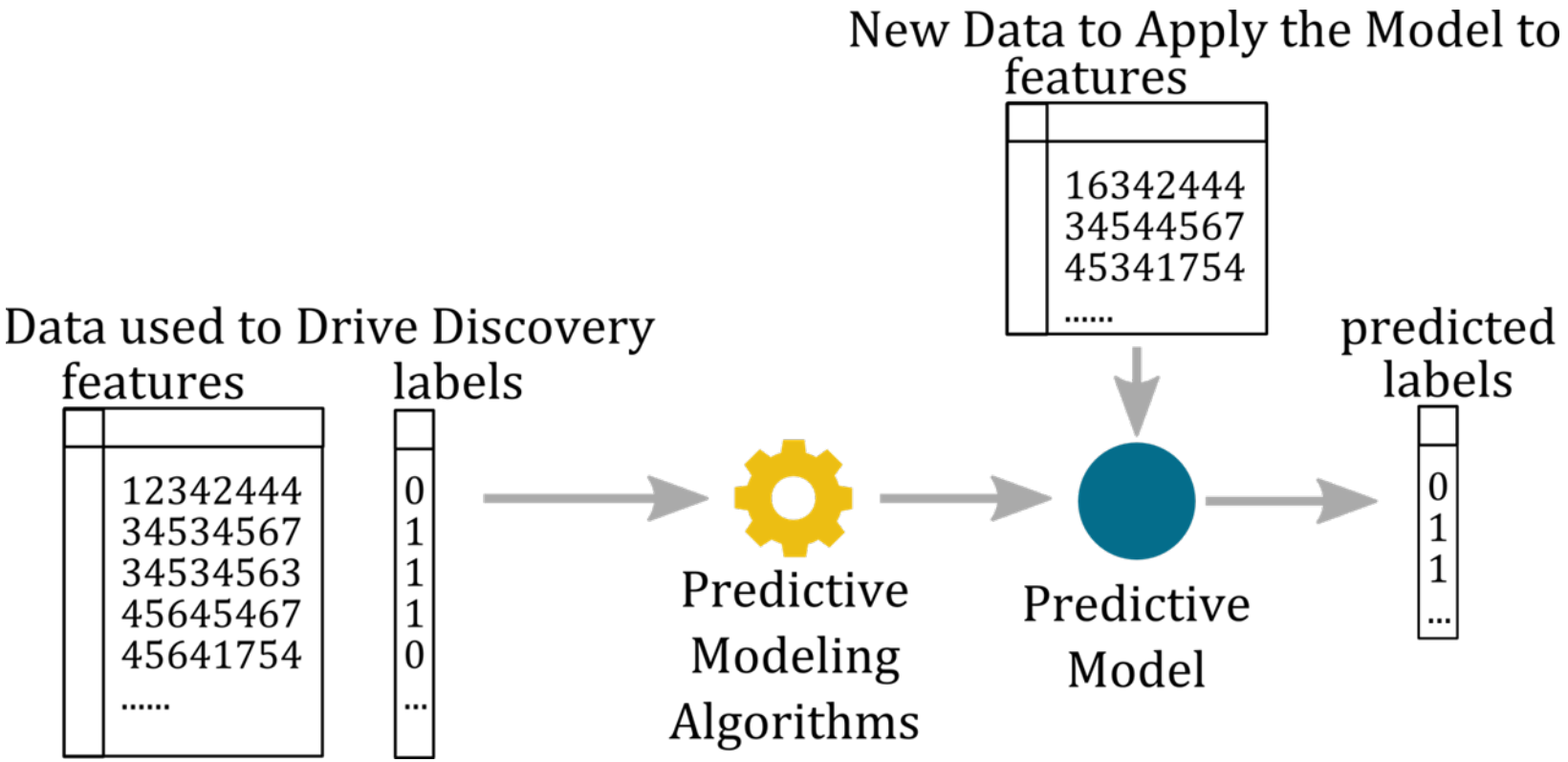
Goal #1: Predictive Modeling

- Forecast the future
- Anticipate events

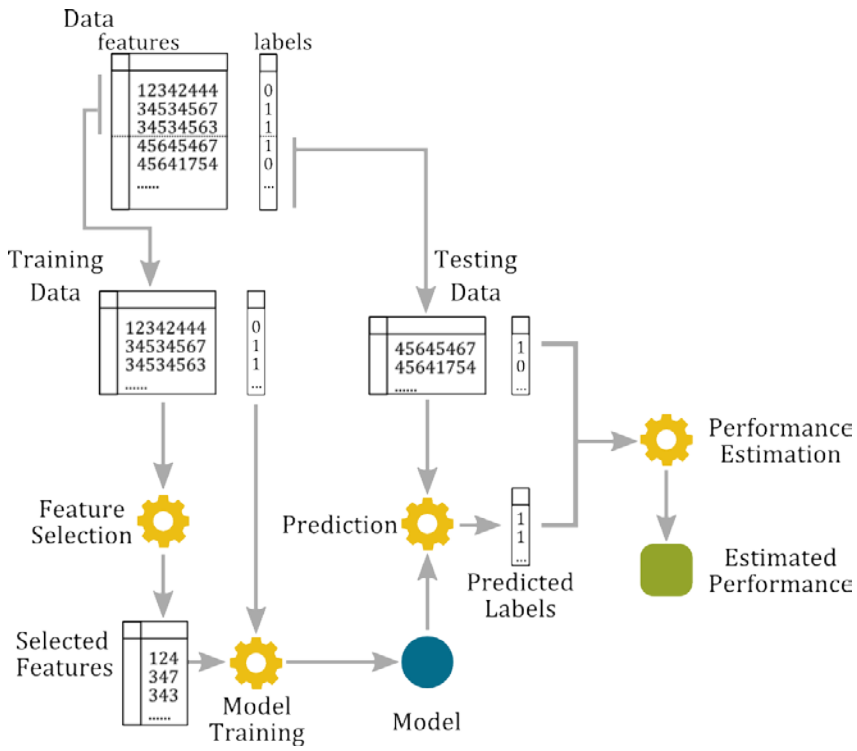
But also:

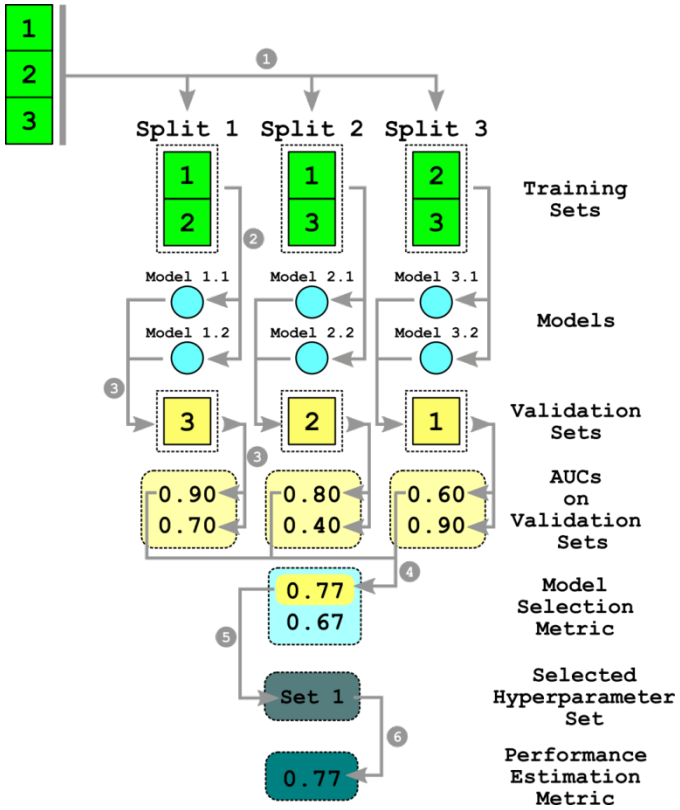
- Recognize patterns
- Assign objects to predefined categories
- Approximate functions (I/O behavior of systems)

Goal #1: Predictive Modeling



Goal #1: Predictive Modeling

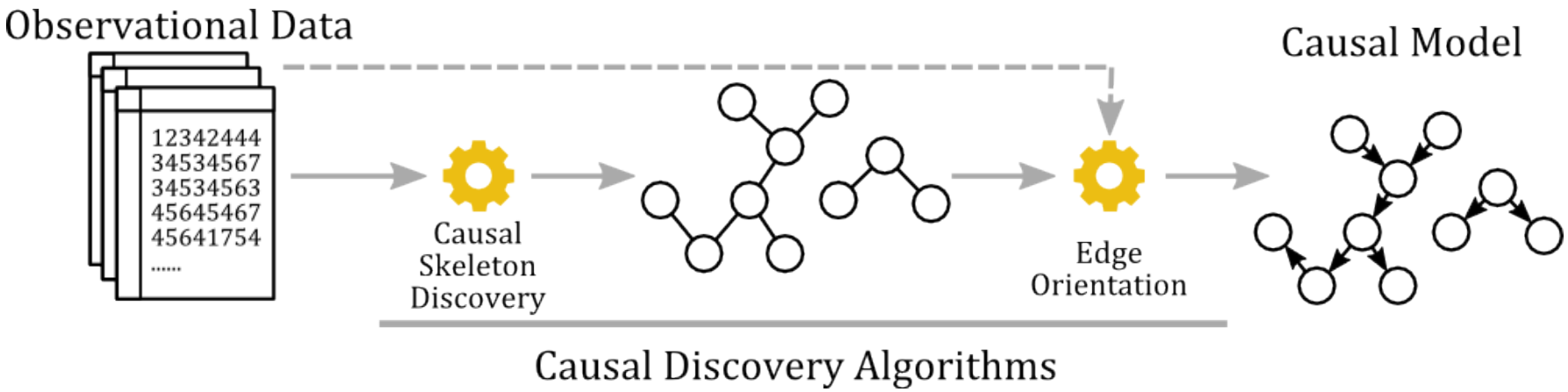




Goal #2: Causal Modeling

- Recognize causes of events
- Recognize complex causal relationships
- Predict events that follow interventions ("manipulations") of a system
- Attribute events to their causes

Goal #2: Causal Modeling

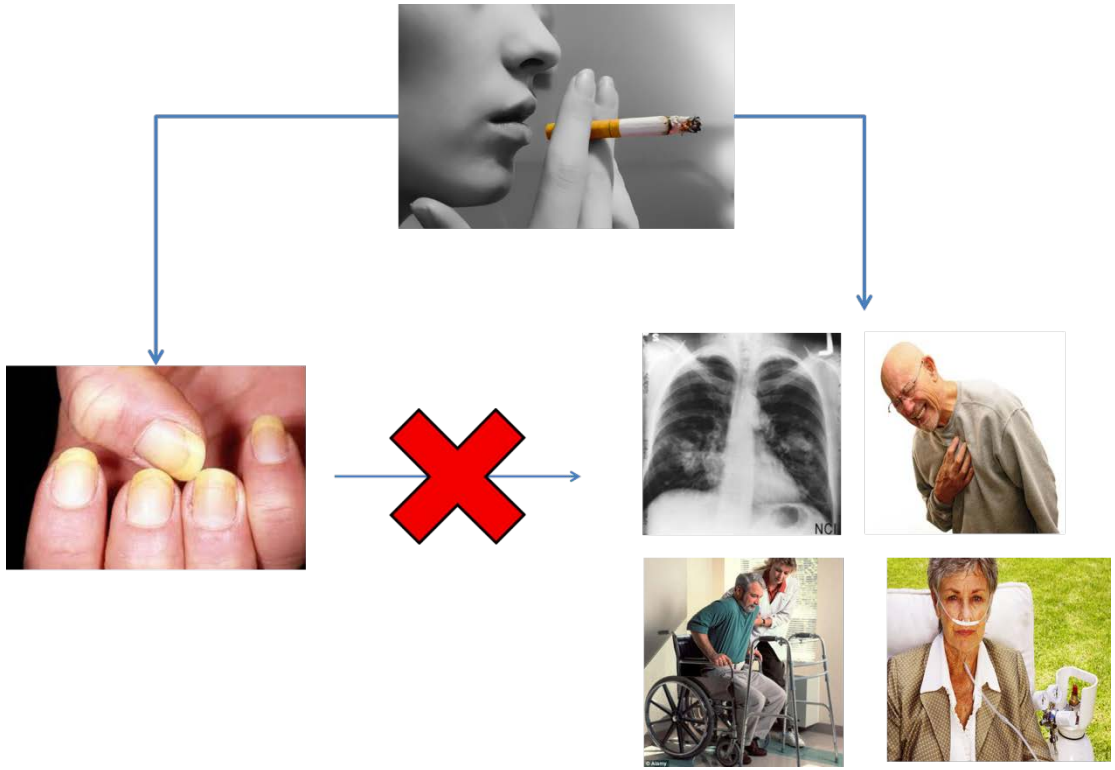


Causality

- Hard to define philosophically
- Good operational way via hypothetical Randomized Experiments

Causality without Experiments

- Dismissive attitude: “Correlation is not causation”



Critique of: “Correlation is not causation” and the strict & blind adherence to an experimental discovery approach

1. **Some correlations are causative and some are not**. Is there a way to systematically differentiate reliably between the two types? It turns out there is.
2. **Is there a way to infer what effects at least certain manipulations would have?** It turns out there is.
3. **REs are neither sound, nor complete**. They admit both false positive, false negative, and true but inflated causal conclusions
4. **REs are typically expensive, slow, low-dimensional and unethical or otherwise infeasible.**

Remainder of talk: take a peek at methods that allow causal discovery without experiments, and combined causal and predictive modeling without experiments.

Generation #1: Simon/Pearl/ Spirtes/Glymour/Scheines/Cooper/Granger

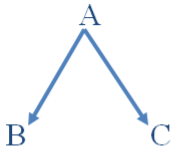
- Learn a causal model if no hidden variables exist

- Key references:

1. *J. Pearl "Causality: Models, Reasoning and Inference". Cambridge University Press, 2000*
2. *P. Spirtes, C. Glymour, R. Scheines "Causation Prediction and Search". MIT Press, 1993, 2000*
3. *C. Glymour, G. Cooper "Computation, Causation and Discovery" AAAI Press 1999*

We need an adequate language for causal discovery. Causal Bayesian Networks simplest and most commonly used one

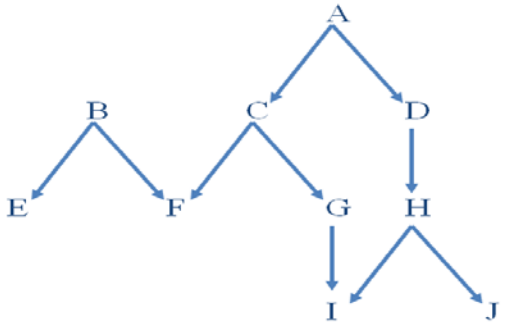
- BN=Graph (Variables (nodes), dependencies (arcs)) + Joint Probability Distribution + Causal Markov Property
- Causal Markov property captures usual semantics of causality



JPD

P(A+, B+, C+)=0.006
P(A+, B+, C-)=0.014
P(A+, B-, C+)=0.054
P(A+, B-, C-)=0.126
P(A-, B+, C+)=0.240
P(A-, B+, C-)=0.160
P(A-, B-, C+)=0.240
P(A-, B-, C-)=0.160

- Markov Property: the probability distribution of any node ***N*** given its parents ***P*** is independent of any subset of the non-descendent nodes ***W*** of ***N***

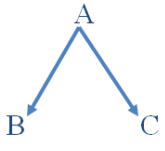


e.g., :

D ⊥ {B,C,E,F,G | A}

F ⊥ {A,D,E,F,G,H,I,J | B, C }

Any JPD can be represented in BN form



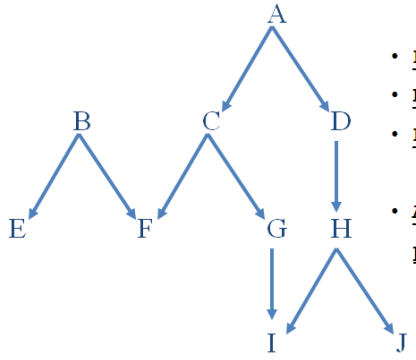
The original JPD:

P(A+, B+, C+)=0.006
P(A+, B+, C-)=0.014
P(A+, B-, C+)=0.054
P(A+, B-, C-)=0.126
P(A-, B+, C+)=0.240
P(A-, B+, C-)=0.160
P(A-, B-, C+)=0.240
P(A-, B-, C-)=0.160

Becomes:

P(A+)=0.8
P(B+ | A+)=0.1
P(B+ | A-)=0.5
P(C+ | A+)=0.3
P(C+ | A-)=0.6

Up to Exponential Saving in Number of Parameters!



- Forward: P(D+,I-| A+)=?
- Backward: P(A+| C+, D+)=?
- Forward & Backward:

  P(D+,C-| I+, E+)=?
- Arbitrary abstraction/Arbitrary predictors/predicted variables

Causal Modeling: PC Algorithm
a prototypical causal discovery algorithm

PC algorithm: Skeleton Discovery

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

B.)

$n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given $\mathbf{S}$ delete edge X - Y from C and record $\mathbf{S}$ in **Sepset**(X,Y) and **Sepset**(Y,X);

until all ordered pairs of adjacent variables X and Y such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n have been tested for d-separation;

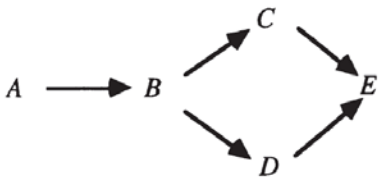
$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(C,X)\backslash\{Y\}$ is of cardinality less than n.

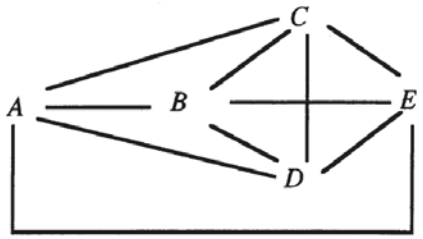
Sprites et al., 1993

Causal Modeling: PC Algorithm

PC algorithm: Skeleton Discovery, Trace



True Graph



Complete Undirected Graph

---

n = 0 No zero order independencies

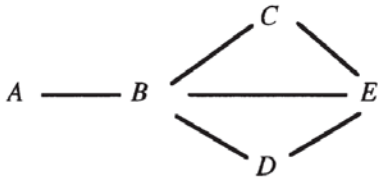
---

n = 1 First order independencies

$A \perp\!\!\!\perp C \mid B$ $A \perp\!\!\!\perp D \mid B$

$A \perp\!\!\!\perp E \mid B$ $C \perp\!\!\!\perp D \mid B$

Resulting Adjacencies

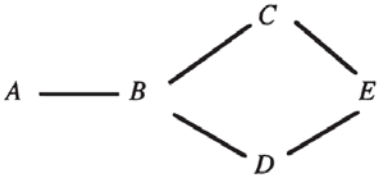


---

n = 2: Second order independencies

$B \perp\!\!\!\perp E \mid \{C,D\}$

Resulting Adjacencies



Causal Modeling: PC Algorithm

PC algorithm: Orientation

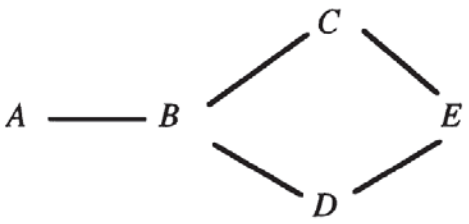
C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as X -> Y <- Z if and only if Y is not in **Sepset**(X,Z).

D. repeat

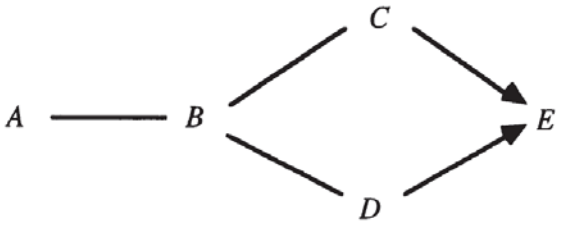
If A -> B, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as B -> C.

If there is a directed path from A to B, and an edge between A and B, then orient A - B as A -> B.

until no more edges can be oriented.



A - B - C; A - B -D;
C - B - D; B - C - E;
B - D - E; C - E - D



Generation #2: Pearl & Spirtes/Glymour/Scheines

- Learn a causal model if hidden variables exist

- 2 major algorithms:

1. ***FCI*** *P. Spirtes et al "Causation Prediction and Search". MIT Press, 1993, 2000*
2. **IC*** *J. Pearl "Causality: Models, Reasoning and Inference". Cambridge University Press, 2000*

Problem #1: Scalability

"In our view, inferring complete causal models […] is essentially impossible in large-scale data mining applications with thousands of variables".

Silverstein, Brin, Motwani, Ullman.

Data Mining and Knowledge Discovery, 2000, pp. 163-192.

Indeed in 2000 one could use sound causal algorithms with up to 100 variables with conventional hardware and slightly more with super computers.

Approaches to Scalability

- Special distributions (e.g., multivariate normal, or Simple Bayes etc.)
- Structural constraints (e.g., connectivity)
- Incomplete learning (output some but not all causal relations)
- Heuristic search
- Focus on skeleton but omit edge orientation
- **Local learning: learn a local causal neighborhood**
- **Related to local learning: local to global**

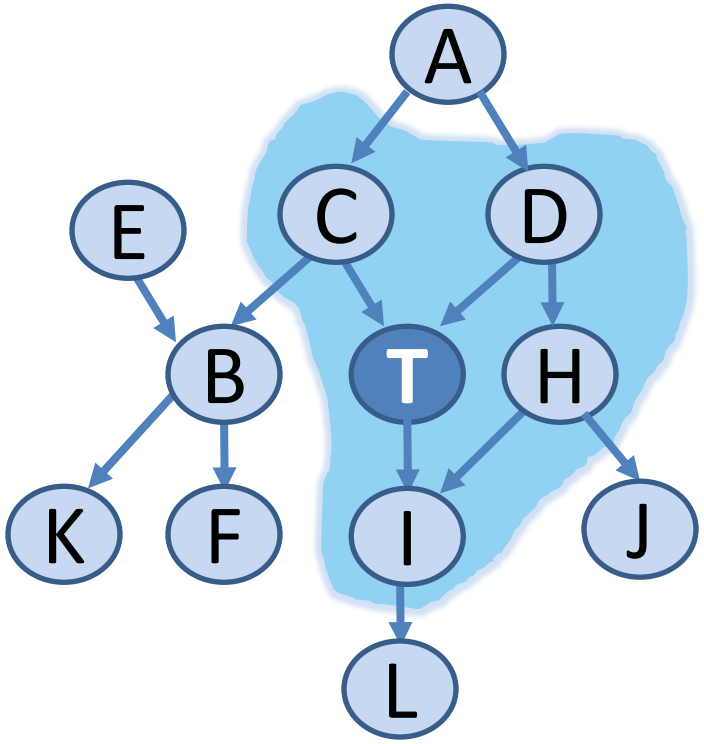
Local causal learning and relationship to Prediction

- Ideally we wish to **blend predictive and causal modeling** because each side has distinct advantages.
- (Obviously) we do not wish to fall in to the trap of confusing predictive with causal knowledge when they do not coincide.
- (Not so obviously) we do not want to use incoherent models for prediction and causal inference.

Approach for Hybrid Predictive + Causal Modeling

The **Markov Boundary** is the set of variables that provides a **principled and mathematically optimal** way to

- reduce variable dimensionality,
- achieve optimal predictivity and –
- discover direct causes and effects for a target/response variable of interest.



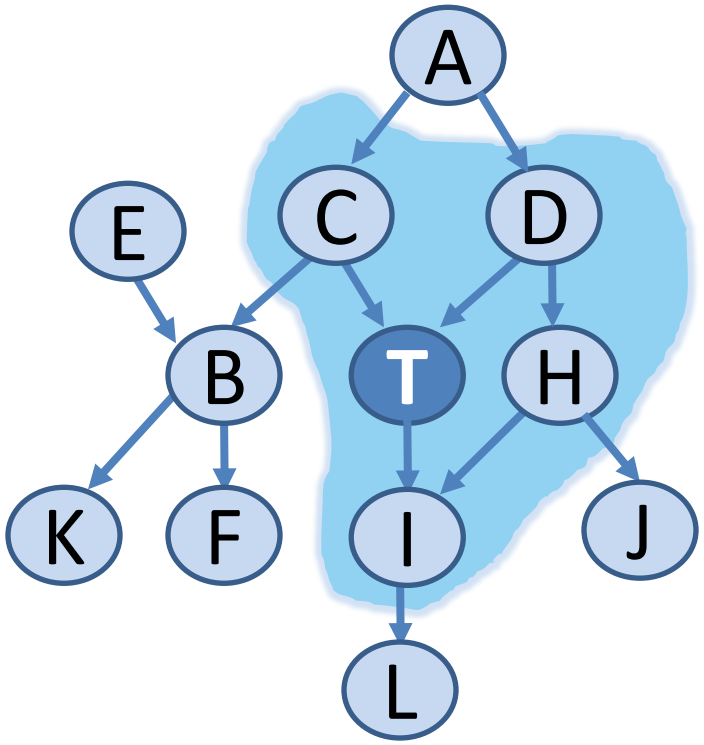
A bit of theory underlying **hybrid causal+predictive** modeling

- There is no single definition of relevancy that covers all combinations of distributions, learners and loss functions (No uniformly optimal filter algorithm exists).
- It is not possible to use wrapper (search and estimate) algorithms for feature selection (No Free Lunch Theorem for feature selection).
- Under broad classes of above, Markov Boundary is optimal predictor set and coincides with Kohavi and John's "Strongly Relevant Features".
- In most distributions, the MB has local causal properties: direct causes + direct effects + direct causes of the direct effects.
- Technicalities in:

"Towards Principled Feature Selection: Relevance, Filters, and Wrappers". I. Tsamardinos and C.F. Aliferis. In Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, USA, January 3-6, 2003.

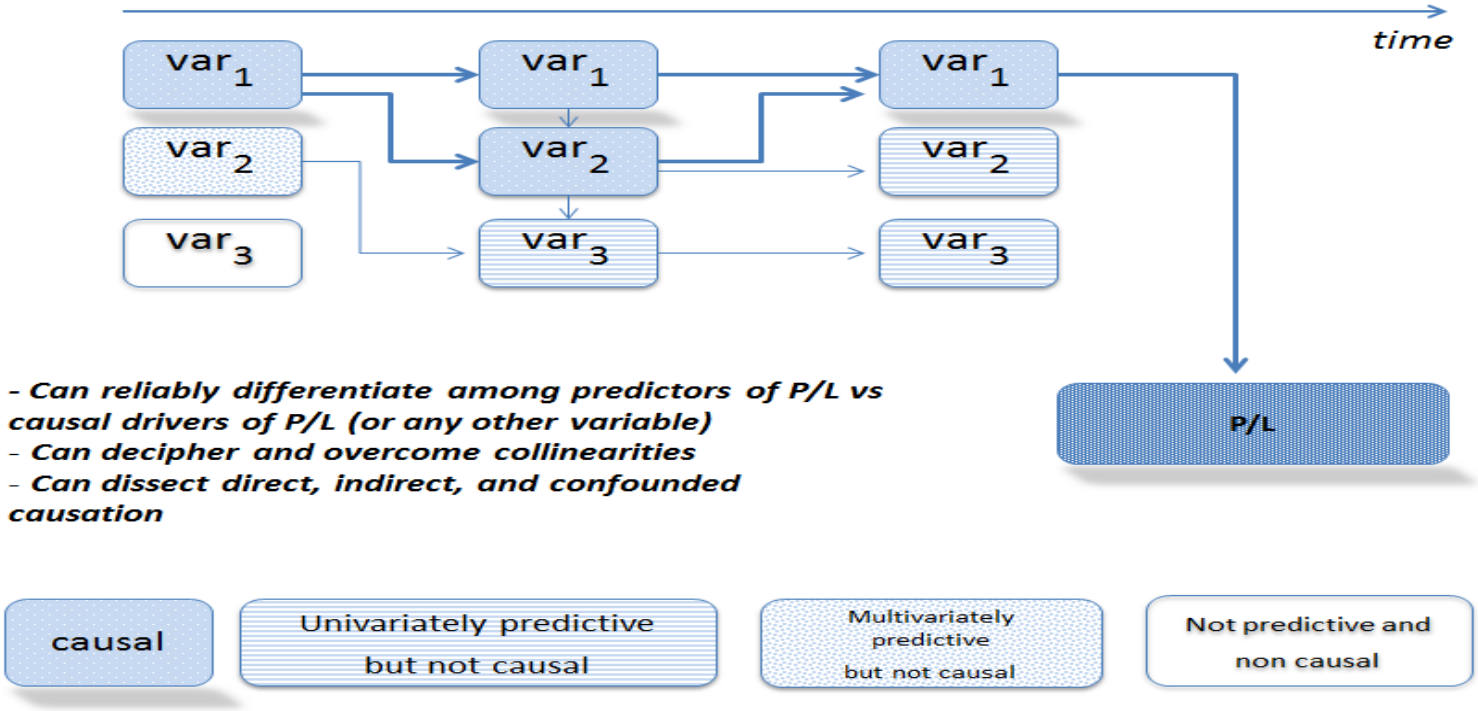
Practical Approach for Hybrid Predictive + Causal Modeling

- If you know the Markov Boundary you can use any standard powerful classifier or regression algorithm to build a predictive model.
- This model will contain all information about the response contained in the full distribution (ie will be optimally predictive)
- Yet by keeping only the MB variable we can safely ignore unnecessary input variables (ie MB is smallest set of optimal predictor variables).



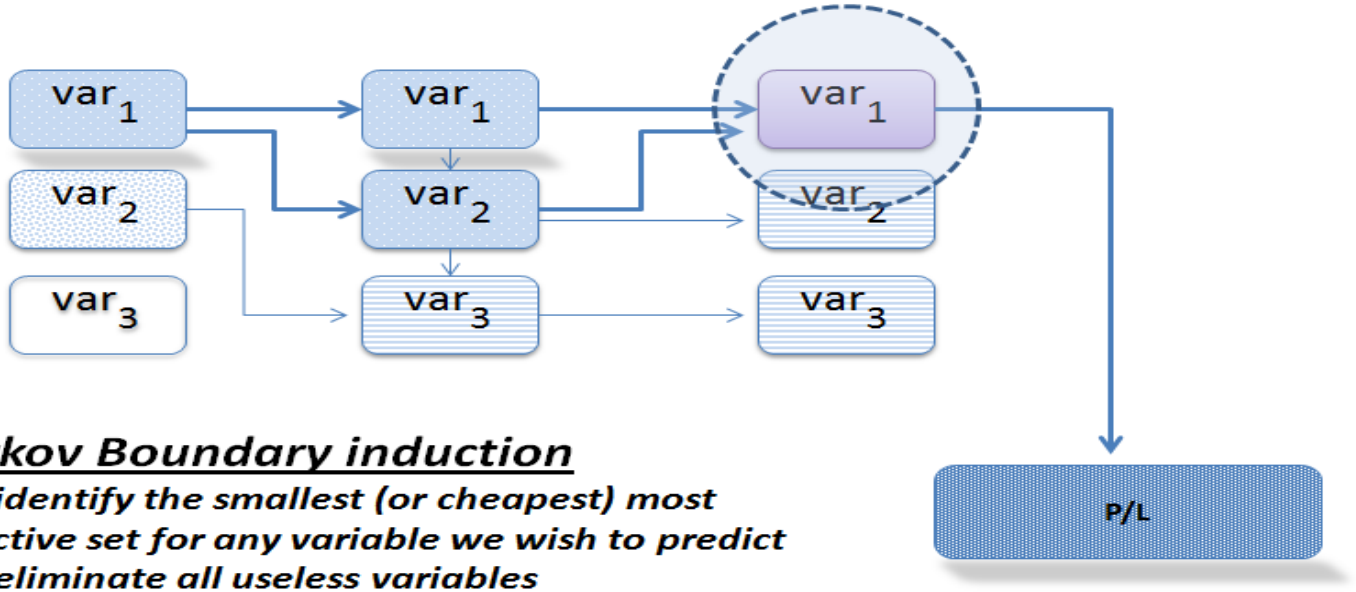
Advantageous Properties of Hybrid Causal-Predictive Analytics **1**

Dissect Predictivity vs Causation



Advantageous Properties of Hybrid Causal-Predictive Analytics **2**

Optimal Predictivity and Parsimony

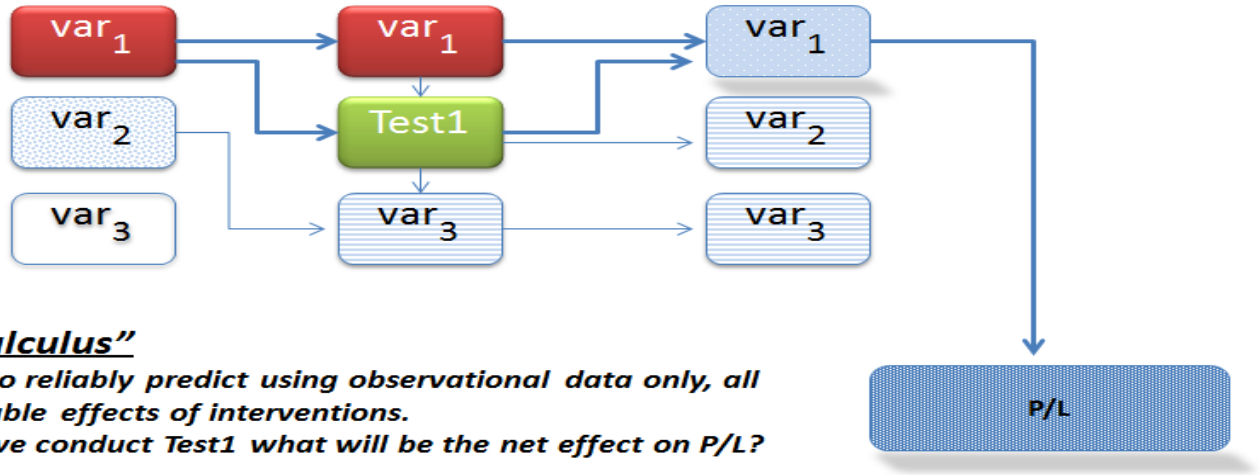


Markov Boundary induction

- can identify the smallest (or cheapest) most predictive set for any variable we wish to predict
- can eliminate all useless variables
- can compress predictive models for model explanation and scalable/convenient use

Advantageous Properties of Hybrid Causal-Predictive Analytics **3**

Estimate Effects of Interventions By Blocking Specific Confounders Revealed by the Causal Graph ("Do Calculus")



"Do calculus"

allows to reliably predict using observational data only, all identifiable effects of interventions.
E.g., if we conduct Test1 what will be the net effect on P/L?

In the example vignette, the Do calculus allows for accurate estimation once we condition on {Var1 (time1), Var1 (time2)} // (this is an application of the so-called "back-door criterion")

Advantageous Properties of Hybrid Causal-Predictive Analytics **4**

- Model multiplicity and optimize models
- Amenable to parallelization, federated analysis, sequential analysis and chunking
- Sound, sample efficient, and scalable in most real life distributions
- Robust to violation of assumptions

Generation #3: Localized MB ("Definitional")

- How do we find the MB?
- One way is to learn a full causal graph, then look at parents, children and spouses.
- NOT practical.
- **Kohler-Sahami**: heuristic, non-scalable.
- **K2MB**: heuristic, non scalable
- Algorithm **Grow-Shrink** (Margaritis and Thrun 2000) returns Markov Boundary only. Sound but sample inefficient and non-scalable.

Generation #4: Scalable Localized MB (Definitional)

- **IAMB** family.
- Return the MB.
- Sound in faithful distributions.
- Sample inefficient (but more efficient than GS)
- Very Scalable (>1,000,000 variables with conventional hardware).
- Robust to hidden variables.
- First paper:

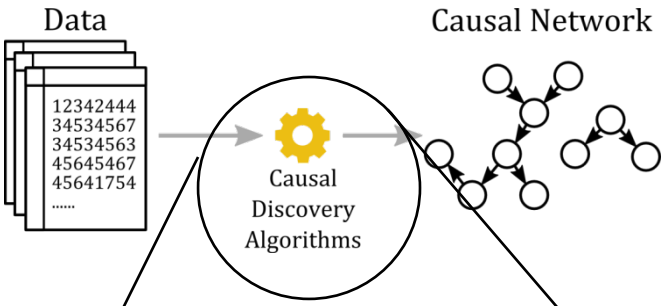
"Algorithms for Large Scale Markov Blanket Discovery". I. Tsamardinos, C.F. Aliferis, A. Statnikov. In Proceedings of the 16th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, St. Augustine, Florida, USA; AAAI Press, pages 376-380, May 12-14, 2003.

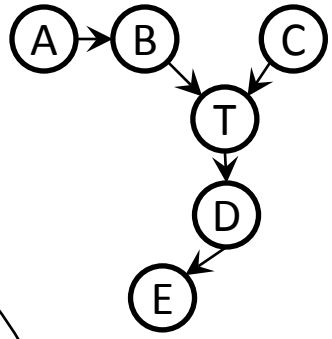
Generation #5: Localized Edges

- Algorithms **MMPC** and **HITON-PC**
- Return the direct causes and direct effects only
- Sound in faithful distributions with no hidden variables locally.
- Sample efficient
- Very Scalable (>1,000,000 variables with conventional hardware).
- Robust to violations of assumptions.
- First papers:

1. *Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations". I. Tsamardinos, C.F. Aliferis, A. Statnikov. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA; ACM Press, pages 673-678, August 24-27, 2003.*
2. *"HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection". C. F. Aliferis, I. Tsamardinos, A. Statnikov. In Proceedings of the 2003 American Medical Informatics Association (AMIA) Annual Symposium, pages 21-25, 2003.*

Causal Modeling: HITON-PC Algorithm (simple version: without symmetry correction or optimizations)





HITON-PC(Data *D*, Target *T*)
"returns parents and children of *T*"
CurrentPC = {}
Repeat
Find variable $V_i \notin CurrentPC$ that maximizes
association(Vi, T) and admit V_i into *CurrentPC*
If there is a variable *X* and a subset *S* of *CurrentPC*
s.t. $\perp(X : T \mid S)$
remove *X* from *CurrentPC;*
do not consider *X* again for admission
Until no more variables are left to consider
Return *CurrentPC*

Trace of HITON-PC

	Vi	CurrentPC	X:T|S	Remove
1	D	{D}	D:T|{}	
2	E	{D,E}	D:T|{E}	
			E:T|{D}	E
3	B	{D,B}	D:T|{B}	
			B:T|{D}	
4	A	{D,B,A}	D:T|{A}	
			D:T|{B}	
			D:T|{B,A}	
			A:T|{D}	
			A:T|{B}	A
			A:T|{D,B}	
			B:T|{D}	
			B:T|{A}	
			B:T|{D,A}	
5	B	{D,C,B}	D:T|{C}	
			D:T|{B}	
			D:T|{B,C}	
			C:T|{D}	
			C:T|{B}	
			C:T|{D,B}	
			B:T|{D}	
			B:T|{C}	
			B:T|{D,C}	

Causal Modeling: Semi-Interleaved HITON-PC a more efficient implementation

Algorithm **Semi-Interleaved HITON-PC** (without "symmetry correction")

Input: dataset $\mathbb{D}$ (a sample from distribution $\mathbb{P}$) for variables $\boldsymbol{V}$, including a response variable T.

Output: a Markov boundary $\boldsymbol{M}$ of T.

Phase I: Forward

1. Initialize $\boldsymbol{M}$ with an empty set
2. Initialize the set of eligible variables $\boldsymbol{E} \leftarrow \boldsymbol{V} \setminus \{T\}$
3. Repeat
4. $\quad Y \leftarrow \text{argmax}_{X \in \boldsymbol{E}}\, \text{Association}(T,\ X)$
5. $\quad \boldsymbol{E} \leftarrow \boldsymbol{E} \setminus \{Y\}$
6. $\quad$ If there is no subset $\boldsymbol{Z} \subseteq \boldsymbol{M}$ such that $T \perp Y \mid \boldsymbol{Z}$ then
7. $\quad\quad \boldsymbol{M} \leftarrow \boldsymbol{M} \cup \{Y\}$
8. Until $\boldsymbol{E}$ is empty

Phase II: Backward

9. For each $X \in \boldsymbol{M}$
10. $\quad$ If there is a subset $\boldsymbol{Z} \subseteq \boldsymbol{M} \setminus \{X\}$ such that $T \perp X \mid \boldsymbol{Z}$ then
11. $\quad\quad \boldsymbol{M} \leftarrow \boldsymbol{M} \setminus \{X\}$
12. End
13. Output $\boldsymbol{M}$

- Efficient, and robust.
- Scalable to very BIG DATA.
- Easily extended for global causal discovery with the LGL framework.
- An instantiation of the GLL framework.

Generation #6: Scalable Region

- Learn causal graph (or Markov network) up to distance *k* from target T by recursive application of local algorithms.

Generation #7: Parallelizing/Chunking/Distributing/ Sequential Scalable MB (Definitional)

- Framework that allows
  - Distributing IAMB-style MB computation among n processors
  - Computing IAMB-style MBs in federated databases
  - Computing IAMB style MBs when data does not fit in a processor by chunking data
  - Computing IAMB style MBs in sequential series of analyses

Aliferis CF, Tsamardinos I. Method, System, and Apparatus for Casual Discovery and Variable Selection for Classification. United States Patent, US 7,117,185 B1, 2006.

Generation #8: Scalable MB ("Compositional")

- Build MB one edge at a time.
- Sound in faithful distributions.
- Sample efficient.
- Robust to violations of some assumptions (e.g. feedback loops)
- Very saleable (>1,000,000 variables with conventional hardware)
- First papers:

1. *Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations". I. Tsamardinos, C.F. Aliferis, A. Statnikov. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA; ACM Press, pages 673-678, August 24-27, 2003.*
2. *"HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection". C. F. Aliferis, I. Tsamardinos, A. Statnikov. In Proceedings of the 2003 American Medical Informatics Association (AMIA) Annual Symposium, pages 21-25, 2003.*

Generation #9: DAQ Local to Global – Full Causal Graph – Algorithm **MMHC**

- Builds local neighborhoods, connects them and then repairs graph with search and score Bayesian approach
- Sound skeleton in faithful distributions.
- Heuristic orientation, best of class overall quality of graph discovery
- Sample efficient.
- Discrete variables only.
- Very scaleable (>10,000 variables with conventional hardware)

- First paper:

"The Max-Min Hill Climbing Bayesian Network Structure Learning Algorithm". I. Tsamardinos, L.E. Brown, C.F. Aliferis. Machine Learning, 65:31-78, 2006.

Generation #10: Generalized Learning Frameworks: **GLL & LGL**

- Generalize the algorithms for local causal edges and compositional MB.
- Generalize the divide and conquer approach of MMHC for full causal graph discovery.
- Generalization in form of **generative algorithms** that can be instantiated in an infinity of ways.
- **Admissibility rules** describe constraints on instantiation that when followed guarantee soundness.
- Specific new instantiations achieve higher scalability, applicability on continuous data and even better quality of reconstruction.

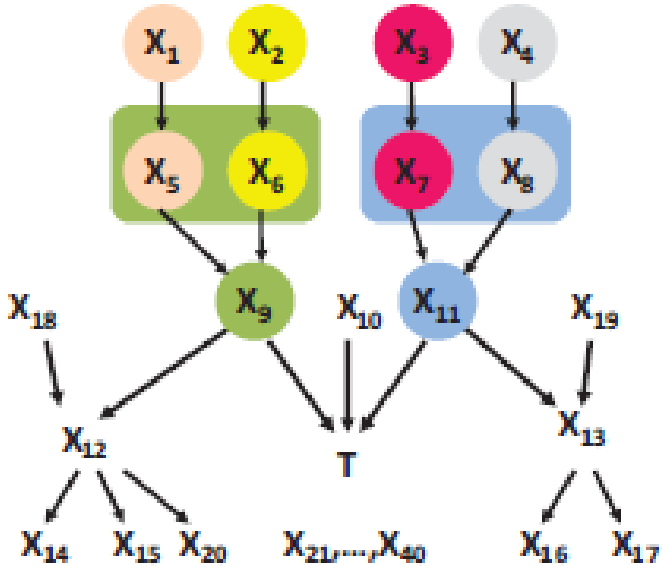
Key papers:

"Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation" C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Journal of Machine Learning Research, 11(Jan):171- 234, 2010.

"Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions" Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos . Journal of Machine Learning Research, 11(Jan):235 - 284, 2010.

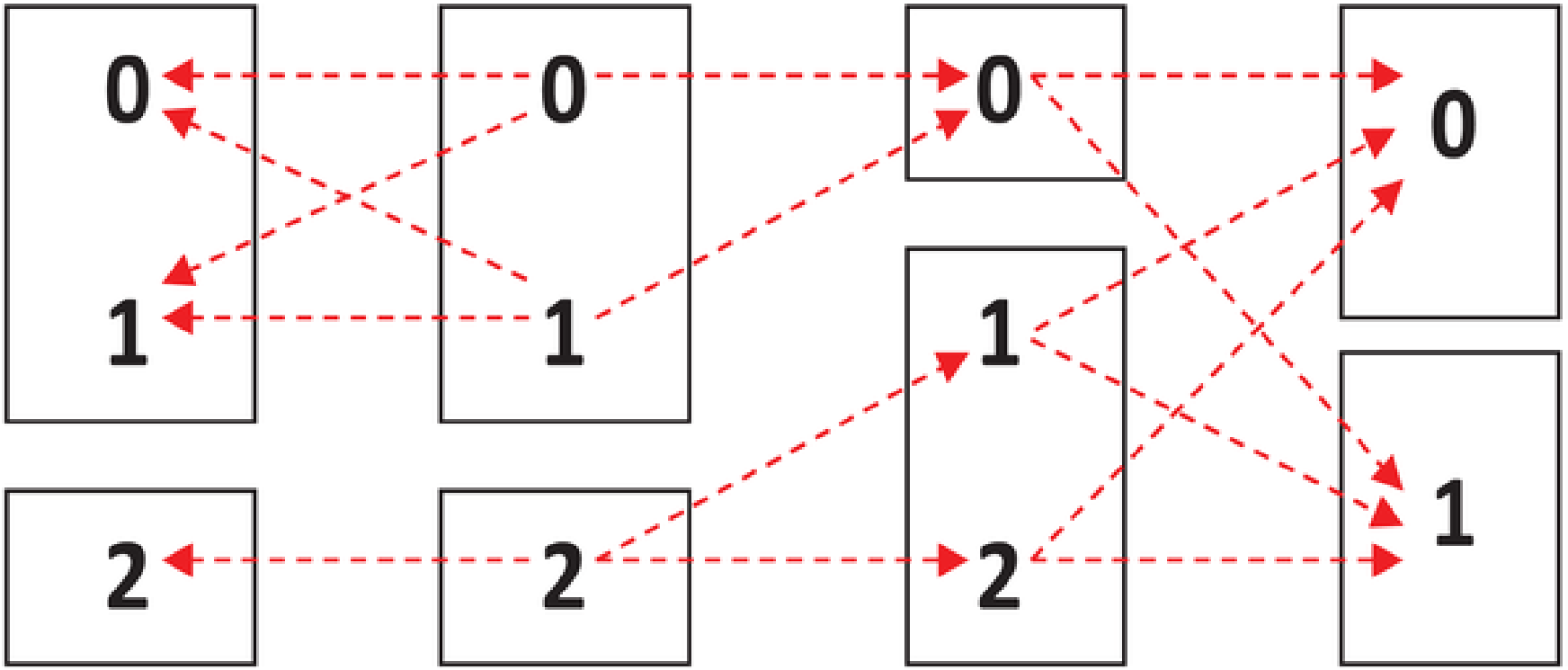
Generation #11: Target Information Equivalency & Modeling Multiplicity

- In some distributions: not one but many MBs.
- No need for determinism!
- Distinct from collinearity.
- Number of MBs can be exponential to number of variables!
- All MBs have optimal predictive information; all are irreducible; some have some have more local causal variables than others; some are more proximal than others; some are larger than others.



Graph of a causal Bayesian network used to trace the TIE∗ algorithm. The network parameterization is provided in Table 8 in Appendix B. The response variable is T. All variables take values {0,1}. Variables that contain equivalent information about T are highlighted with the same color, for example, variables X1 and X5 provide equivalent information about T; variable X9 and each of the four variable sets {X5,X6}, {X1,X2}, {X1,X6}, {X5,X2} provide equivalent information about T.

Figure 1. The figure describes a class of Bayesian networks that share the same pathway structure (with 3 gene variables A, B, C and a phenotypic response variable T) and their joint probability distribution obeys the constraints shown below the structure.



Statnikov A, Aliferis CF (2010) Analysis and Computational Dissection of Molecular Signature Multiplicity. PLoS Comput Biol 6(5): e1000790. doi:10.1371/journal.pcbi.1000790
http://127.0.0.1:8081/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000790

High-level pseudocode of the TIE* algorithm.

Algorithm **TIE*** (high-level pseudocode)

Inputs: (a) dataset with predictive variables (e.g., genes) and a phenotypic response variable,
(b) base Markov boundary induction (gene/variable selection) algorithm.

Output: the set of maximally predictive and non-redundant signatures of the phenotype.

1. Use the base algorithm to learn a Markov boundary **M** of the phenotype from data for all measured variables. Output **M**.
2. Repeat
3. Generate the smallest subset of variables **G** of the so far discovered Markov boundaries of the phenotype such that: (i) **G** was not considered in the previous iteration of the algorithm, and (ii) **G** does not properly include any subset of variables that was generated in the previous iteration of the algorithm when $\mathbf{M}_{new}$ was found not to be a Markov boundary of the phenotype.
4. Use the base algorithm to learn a *candidate* Markov boundary $\mathbf{M}_{new}$ of the phenotype from data for all measured variables but **G**.
5. If the phenotypic predictivity of the signature $\mathbf{M}_{new}$ is at least as good as that of **M** (estimated by holdout validation or other unbiased estimator) according to a statistical significance test or some other criterion, then $\mathbf{M}_{new}$ is indeed a Markov boundary of the phenotype and it is output.
6. Until no subset **G** can be generated in step 3.

Statnikov A, Aliferis CF (2010) Analysis and Computational Dissection of Molecular Signature Multiplicity. PLoS Comput Biol 6(5): e1000790. doi:10.1371/journal.pcbi.1000790
http://127.0.0.1:8081/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000790

Generation #11: Target Information Equivalency & Modeling Multiplicity CONT'D

- TIE* family of algorithms extracts all MBs in a distribution.
- Sample efficient.
- Sound.
- Scalable (>1,000,000 variables with conventional hardware).
- Like GLL and LGL generative framework describes generative algorithm, admissibility criteria and meta properties.
- Papers:

"Analysis and Computational Dissection of Molecular Signature Multiplicity" A. Statnikov, C.F. Aliferis. (Cover Article) PLoS Computational Biology, 2010; 6(5): e1000790.

Algorithms for Discovery of Multiple Markov Boundaries. Alexander Statnikov, Nikita I. Lytkin, Jan Lemeire, Constantin F. Aliferis; JMLR, 14(Feb):499–566, 2013.

Generation #12: Compositional MBs with Hidden Variables (**Algorithm CIMB**)

- IAMB family (definitional MB algortihms) robust to hidden variables but GLL-MB family (compositional algorithms) admit false negatives.
- **CIMB** is a compositional family that avoids false negatives.
- Same sample efficiency, soundness and scalability as GLL-MB.

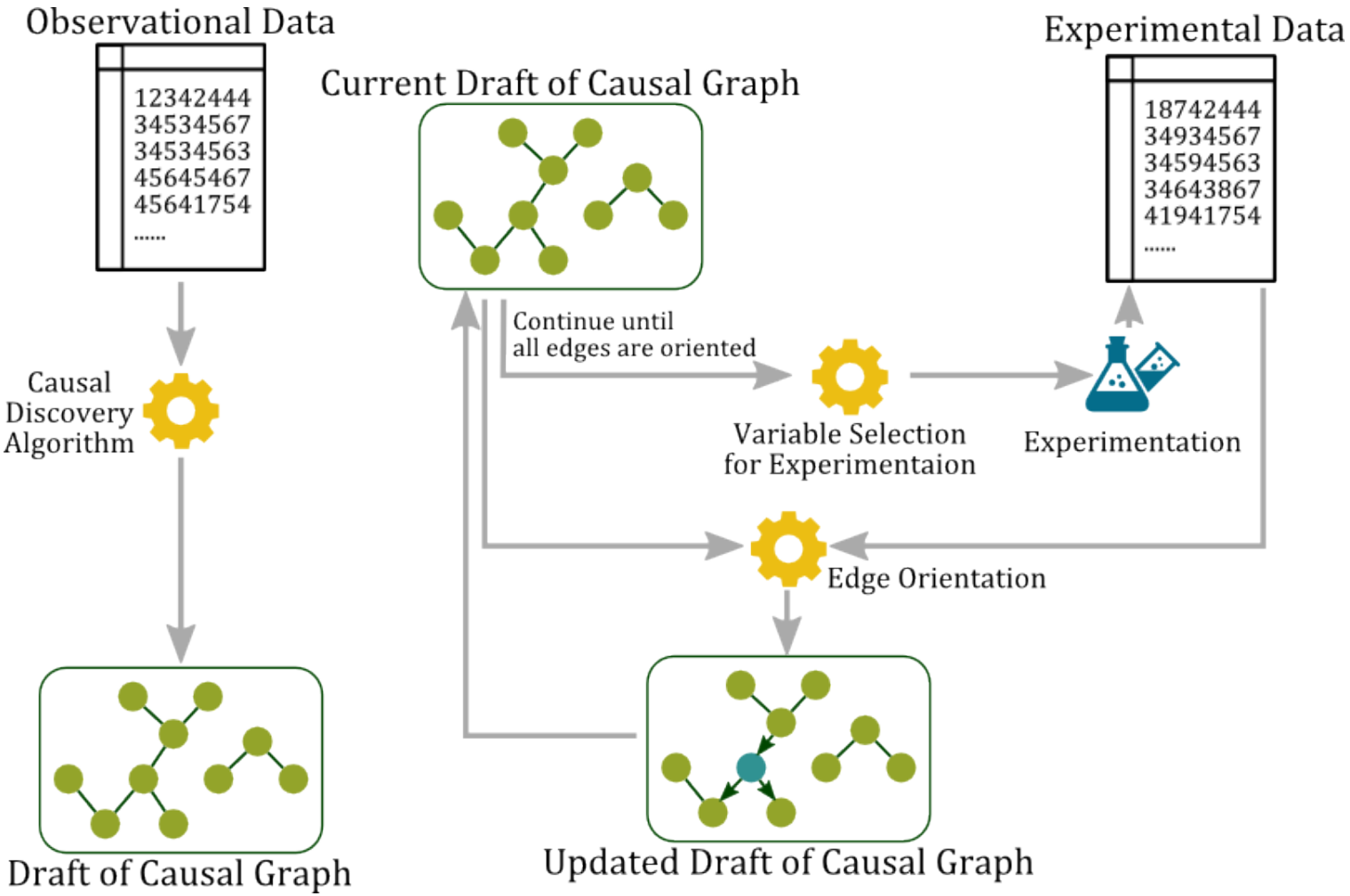
Generation #13: Experimentation Minimizing with Algorithm **ODLP**

- Causal Model-Guided Experimental Minimization and Adaptive Data Collection
- Intends to help experimentalists reduce the number of experiments needed to learn a causal model.
- Especially useful when experimentation is needed to resolve causal ambiguity that is undiscoverable without experimentation.

"New Ultra-Scalable and Experimentally Efficient Methods for Local Causal Pathway Discovery".

Alexander Statnikov, Mikael Henaff, Nikita Lytkin, Efstratios Efstathiadis, Eric R. Peskin, Constantin F. Aliferis (to appear in JMLR)

Simplified view of the Framework:



Causal Model Guided Experimental Minimization and Adaptive Data Collection

The ODLP Algorithm:

Output:

- Local causal pathway (parents and children) of the variable of interest.

Two Phases:

- Identify local causal pathway consistent with the data and information equivalent clusters.
- Adaptively recommend experiments to perform, integrate experimental results to refine and orient the local causal pathway.

Statnikov et al., 2015 (Accepted)

Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP: Pseudo Code:

Algorithm ***ODLP***

- *Input*:
  - Observational data D^O, including a target variable *T*;
  - Experimental protocols/methods to manipulate one variable at a time and generate experimental data D^E that quantifies response of the system to the manipulation.
- *Output*: Local causal pathway of *T*.

1. Apply TIE* or iTIE* to the observational data D^O to identify all local causal pathways of *T* consistent with the data.
2. ***V*** ← Union of all variables that participate in local causal pathways of *T* consistent with the data (*this is a draft of the local causal pathway*).
3. Form equivalence clusters over variables in ***V*** such that each equivalence cluster contains variables that have equivalent information about T (*this can be accomplished directly from the output or the operation of TIE* or iTIE**).

Identify effects of T

4. Manipulate *T* and obtain experimental data D^E.
5. Mark all variables in ***V*** that change in D^E due to manipulation of T as "*effects*".

Identify direct and other causes of T

6. Repeat
   a. If there is an equivalence cluster that contains a single unmarked variable *X* and all marked variables in this cluster (if any) are only passengers and/or effects, then mark *X* as a "*direct cause*" and go to step 6.
   b. Select (according to some heuristic function or at random) an unmarked variable *X* from an equivalence cluster.
   c. Manipulate *X* and obtain experimental data D^E.
   d. If *T* does not change in D^E due to manipulation of *X*, mark *X* as a "*passenger*" and mark all other non-effect variables that change in D^E due to manipulation of *X* as "*passengers*"; otherwise mark *X* as a "*cause*".
7. Until there are no equivalence clusters with unmarked variables.
8. For every cause *X*, mark *X* as a "*direct cause*" if there exist no other cause in the same equivalence cluster that changes due to manipulation of *X*; otherwise mark *X* as an "*other cause*".

Identify direct effects of T

9. Repeat
   a. If there is an equivalence cluster that contains a single effect variable *X* which has neither been marked as "other effect" nor as "direct effect" and other effect variables in this cluster (if any) are only other effects, then mark *X* as a "*direct effect*" and to go step 9.
   b. Select (according to some heuristic function or at random) an effect variable *X* that has neither been marked as "*other effect*" nor as "*direct effect*".
   c. Manipulate *X* and obtain experimental data D^E.
   d. Mark all effect variables that change in D^E due to manipulation of *X* and belong to the same equivalence cluster as "other *effects*".
10. Until all effect variables are either marked as "*other effects*" or "*direct effects*".
11. Return the local causal pathway of *T*, i.e. only direct causes and direct effects of *T*.

The ODLP Algorithm:

Output:

- Local causal pathway (parents and children) of the variable of interest.

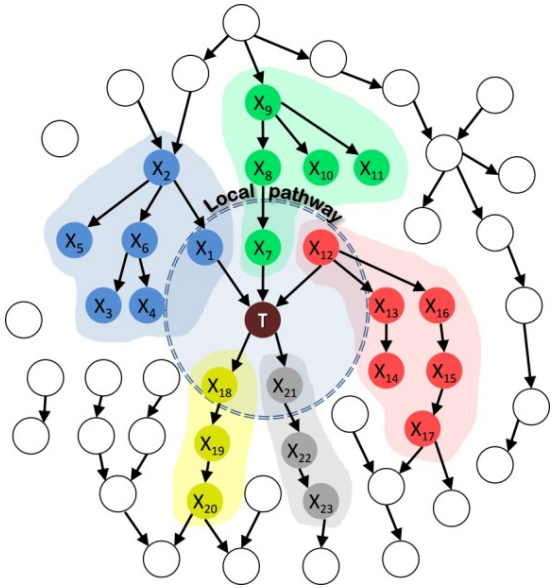
Two Phases:

- Identify local causal pathway consistent with the data and information equivalent clusters.
- Adaptively recommend experiments to perform, integrate experimental results to refine and orient the local causal pathway.

Causal Model Guided Experimental Minimization and Adaptive Data Collection

The ODLP Algorithm Phase I:

- Identify local causal pathway consistent with the data and information equivalent clusters (TIE*, iTIE* algorithms).



Definition of target information equivalency: Two subsets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ from $\boldsymbol{V}$ are target information equivalent with respect to a variable T iff the following conditions hold $T \not\perp \boldsymbol{X}$, $T \not\perp \boldsymbol{Y}$, $T \perp \boldsymbol{X}|\boldsymbol{Y}$, and $T \perp \boldsymbol{X}|\boldsymbol{Y}$ (Lemeire, 2007).

Causal Model Guided Experimental Minimization and Adaptive Data Collection

The ODLP Algorithm Phase I: iTIE*

Algorithm **iTIE***

Input: dataset $\mathbb{D}$ (a sample from distribution $\mathbb{P}$) for variables ***V***, including a target variable *T*.

Output: multiple Markov boundaries of *T* that exist in $\mathbb{P}$.

Phase I: Forward

1. Initialize Θ with an empty set
2. Initialize ***M*** with an empty set
3. Initialize the set of eligible variables ***E*** ← ***V*** \ *T*
4. Repeat
5. $Y \leftarrow \text{argmax}_{X \in \boldsymbol{E}} \text{Association}(T, X)$
6. ***E*** ← ***E*** \ *Y*
7. If there is no subset ***Z*** ⊂ ***M*** such that $T \perp Y \mid \boldsymbol{Z}$ then
8. ***M*** ← ***M*** ∪ *Y*
9. Else if ***Z*** exists and the following relations hold: $T \perp Y$, $T \perp \boldsymbol{Z}$, $T \perp \boldsymbol{Z} \mid Y$
10. Record in Θ that *Y* and ***Z*** contain equivalent information with respect to *T*
11. Until ***E*** is empty

Phase II: Backward

12. For each $X \in \boldsymbol{M}$
13. If there is a subset ***Z*** ⊂ ***M*** \ *X* such that $T \perp X \mid \boldsymbol{Z}$ then
14. ***M*** ← ***M*** \ *X*

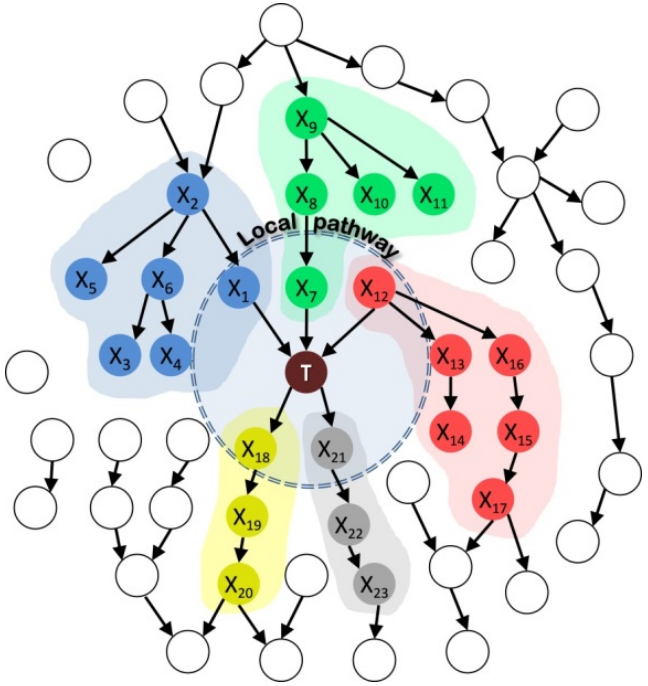
Phase III: Construction of multiple Markov boundaries

15. Compute the Cartesian product of target information equivalency relations for subsets of ***M*** that are stored in Θ to construct multiple Markov boundaries of *T*
16. Output multiple Markov boundaries of *T*

Causal Model Guided Experimental Minimization and Adaptive Data Collection

The ODLP Algorithm Phase II:

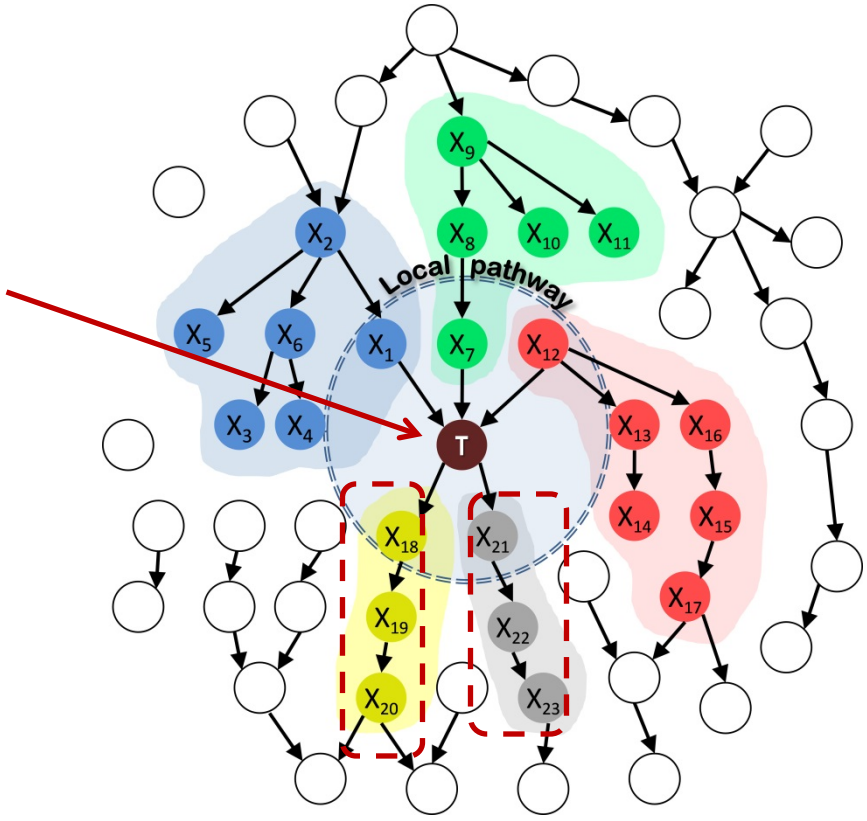
- Adaptively recommend experiments to perform, integrate experimental results to refine and orient the local causal pathway. (i.e. ***Identify Causes, Effects, and Passengers***).



Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP: Identifying effects

- Manipulate *T* and obtain experimental data D^E.
- Mark all variables in ***V*** that change in D^E due to manipulation of T as *effects*.

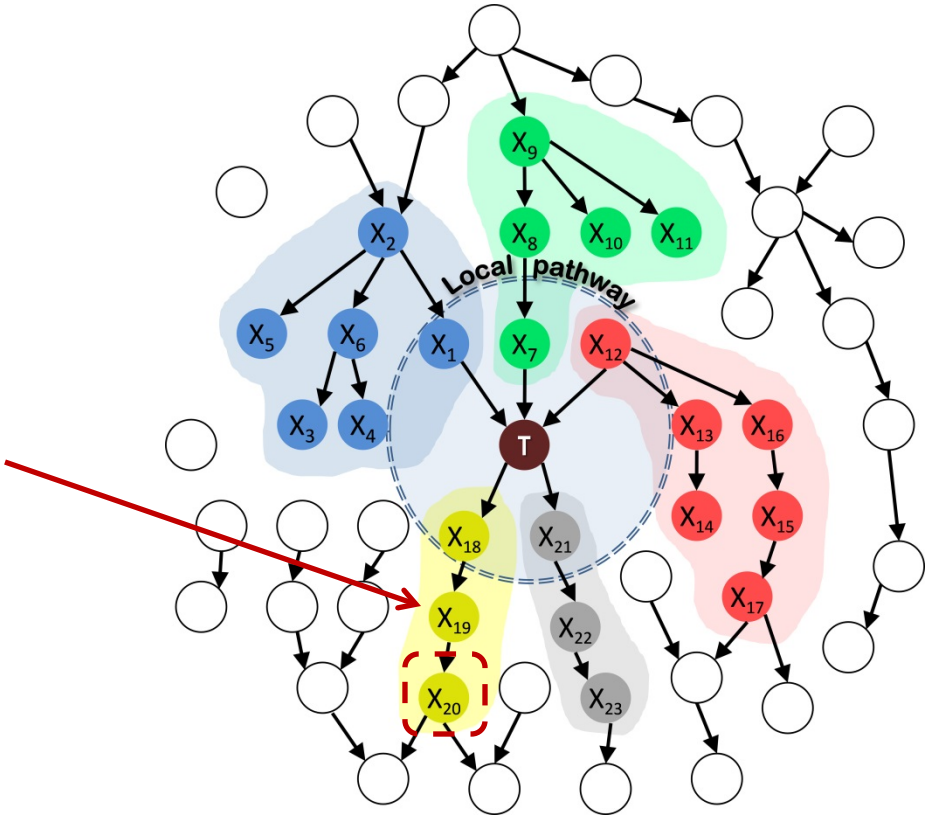


effects

Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP: direct and indirect effects

- Select an effect variable *X* that has neither been marked as *indirect effect* nor as *direct effect*.
- Manipulate *X* and obtain experimental data D^E.
- Mark all effect variables that change in D^E due to manipulation of *X* and belong to the same equivalence cluster as indirect *effects*.
- The last effect variable in an equivalent cluster that is not marked as *indirect effect* is a *direct effect*.

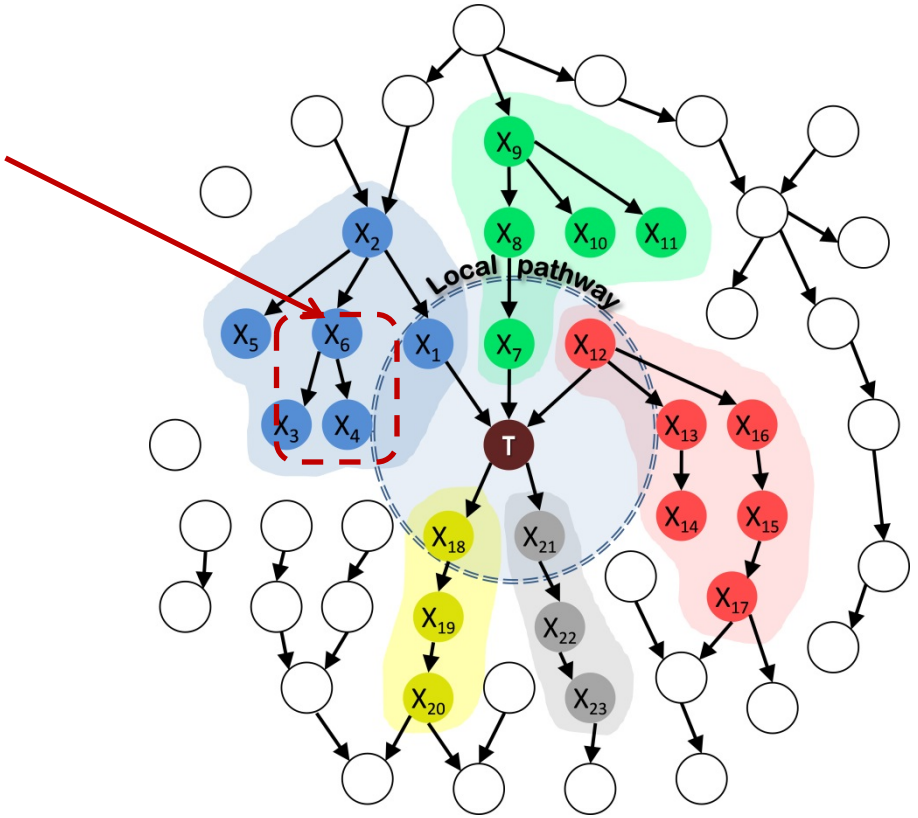


Indirect effect

Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP: Identifying Passengers

- Select an unmarked variable *X* from an equivalence cluster.
- Manipulate *X* and obtain experimental data D^E.
- If *T* does not change in D^E due to manipulation of *X*, mark *X* as a *passenger* and mark all other non-effect variables that change in D^E due to manipulation of *X* as *passengers*; otherwise mark *X* as a *cause*.

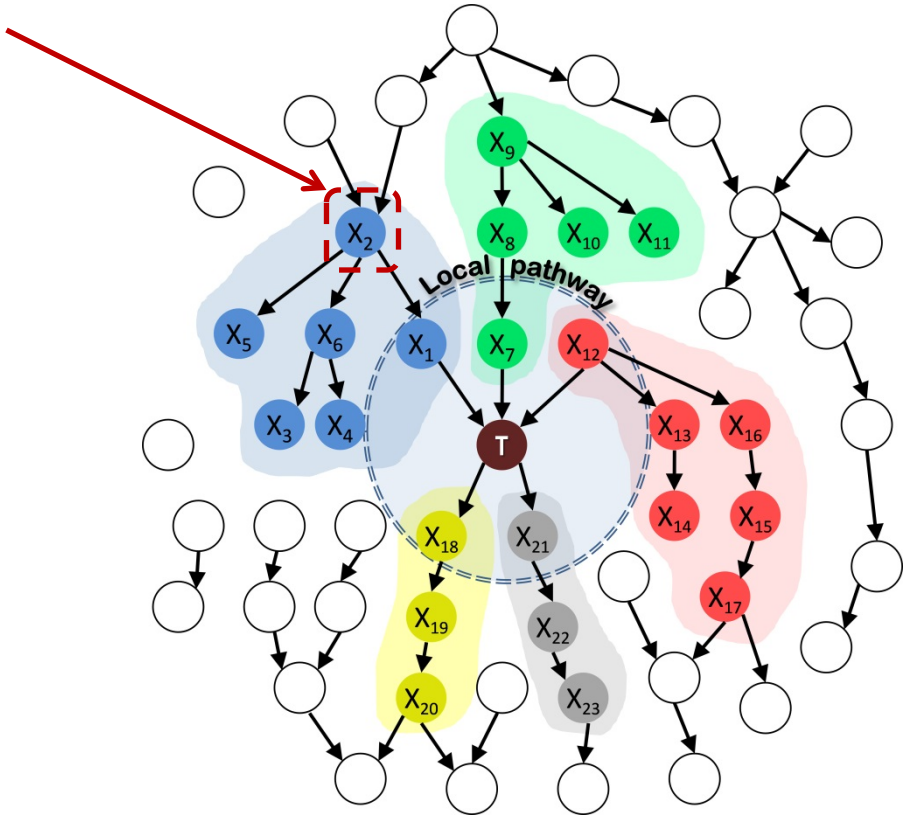


Passengers

Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP: Identifying Causes

- For every cause *X*, mark *X* as a *direct cause* if there exist no other cause in the same equivalence cluster that changes due to manipulation of *X*; otherwise mark *X* as an *Indirect cause*.
- If there is an equivalence cluster that contains a single unmarked variable *X* and all marked variables in this cluster (if any) are only passengers and/or effects, then mark *X* as a *direct cause*.



Generation #14: Generalized Framework for Parallel/ Chunked/ Sequential/Distributed Processing

- As in P/D/S/C framework for definitional MB algorithms but extends to local causal, MB compositional and TIE algortihms

APPLICATION/PROVING GROUND #1

1. Optimal predictivity and maximum feature selection parsimony

First Results: General Distributions

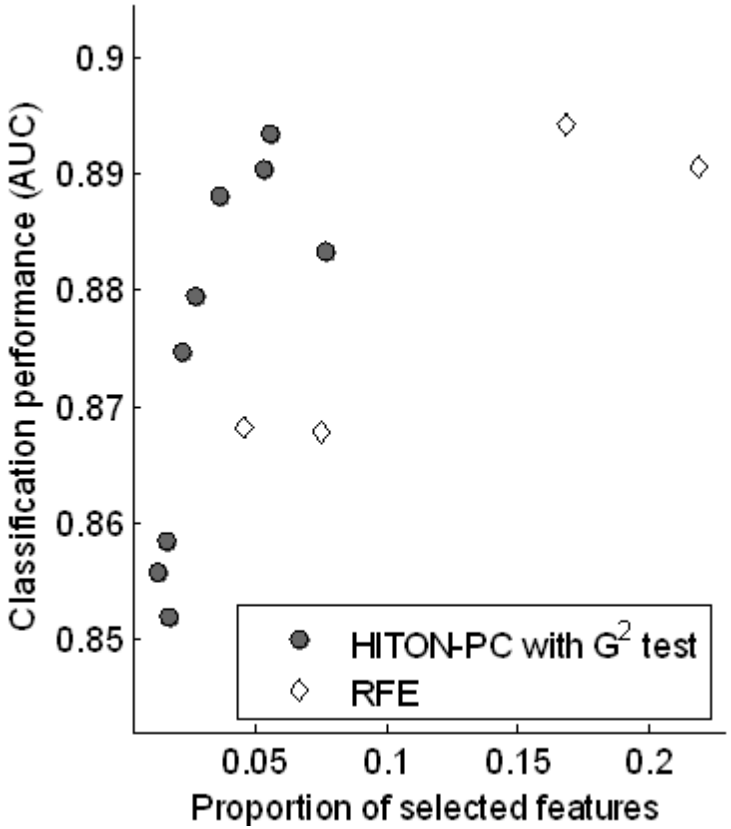
- >100 algorithms
- >40 datasets
- Key references

"Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation" C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. *Journal of Machine Learning Research*, 11(Jan):171-234, 2010.

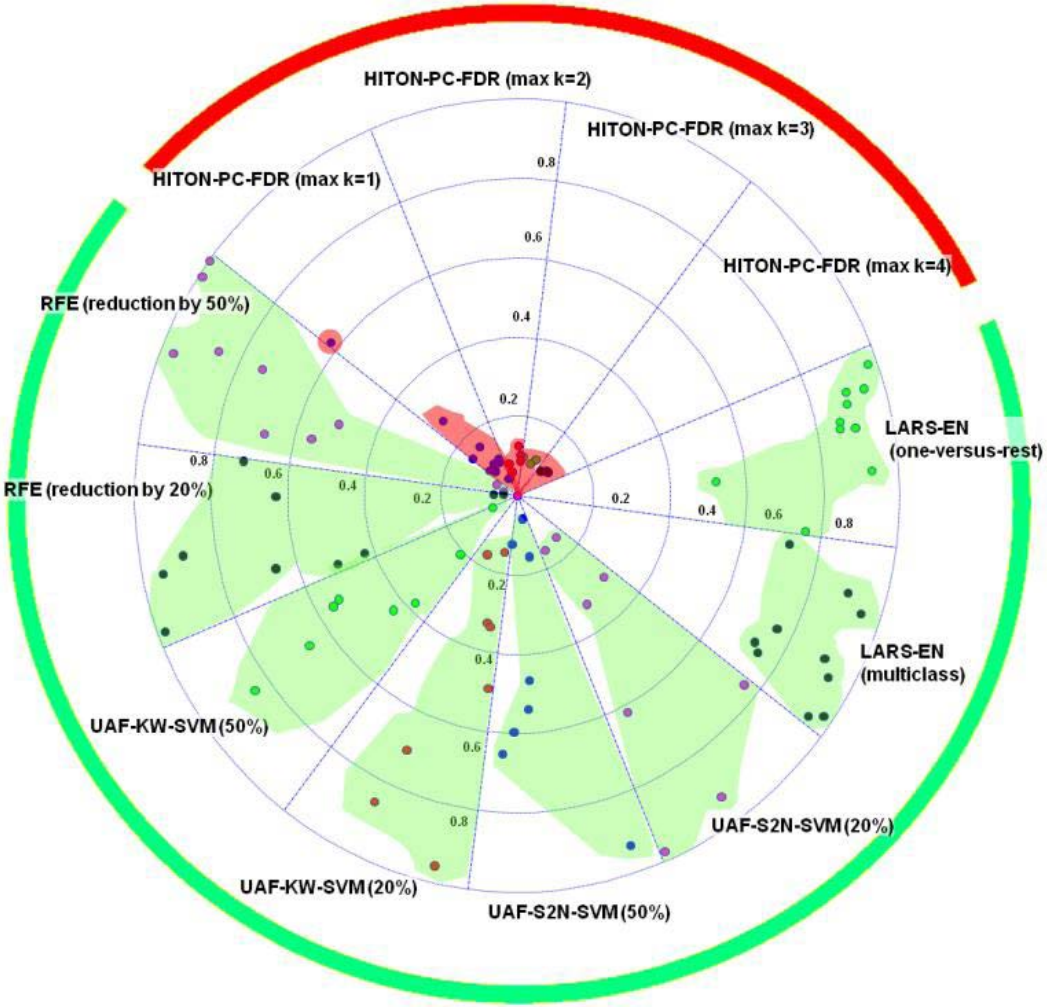
"Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions" Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos . *Journal of Machine Learning Research*, 11(Jan):235 - 284, 2010.

Development of maximally parsimonious and maximally predictive models and predictive variable sets

	Predicitivity		*Reduction*	
Feature selection method	P-value	Nominal winner	P-value	Nominal winner
No feature selection	0.1890	Other	**<0.0001**	HITON-PC
RFE: 4 variants	0.9754	Other	**0.0046**	HITON-PC
	0.8030	Other	**0.0042**	HITON-PC
	0.1312	HITON-PC	0.3634	HITON-PC
	0.1008	HITON-PC	0.6816	Other
UAF-KruskalWallis-SVM: 4 variants	0.2248	Other	**0.0028**	HITON-PC
	0.0098	Other	**0.0004**	HITON-PC
	1.0000	HITON-PC	0.1414	HITON-PC
	0.3232	HITON-PC	0.3998	HITON-PC
UAF-Signal2Noise-SVM: 4 variants	0.0710	Other	**0.0018**	HITON-PC
	0.0752	Other	**0.0030**	HITON-PC
	0.4420	HITON-PC	0.7850	HITON-PC
	0.2820	HITON-PC	0.6604	HITON-PC
UAF-Neal-SVM: 4 variants	0.5046	Other	**<0.0001**	HITON-PC
	0.9782	HITON-PC	**<0.0001**	HITON-PC
	0.6980	HITON-PC	**0.0044**	HITON-PC
	0.3806	HITON-PC	**0.0186**	HITON-PC
Random Forest Variable Selection: 2 variants	0.6064	HITON-PC	0.3252	HITON-PC
	0.5050	HITON-PC	0.1338	Other
LARS-Elastic Net: 2 variants	1.0000	Other	0.1112	HITON-PC
	0.0832	HITON-PC	0.5216	Other
RELIEF: 8 variants	0.2032	Other	**<0.0001**	HITON-PC
	0.9362	Other	**<0.0001**	HITON-PC
	0.4388	Other	**0.0014**	HITON-PC
	0.8432	Other	**0.0010**	HITON-PC
	0.4290	HITON-PC	**0.0108**	HITON-PC
	0.3114	HITON-PC	0.0518	HITON-PC
	0.4424	HITON-PC	0.0706	HITON-PC
	0.2748	HITON-PC	**0.0404**	HITON-PC
L0-norm	**0.0258**	HITON-PC	0.1942	HITON-PC
Forward Stepwise Selection	**0.0028**	HITON-PC	0.2758	Other
Koller-Sahami: 6 variants	0.7506	HITON-PC	**<0.0001**	HITON-PC
	0.6234	HITON-PC	**<0.0001**	HITON-PC
	0.6278	HITON-PC	**<0.0001**	HITON-PC
	<0.0001	HITON-PC	**<0.0001**	Other
	0.1278	HITON-PC	0.3856	HITON-PC
	0.1236	HITON-PC	**<0.0001**	HITON-PC
IAMB: 3 variants	**<0.0001**	HITON-PC	**<0.0001**	Other
	<0.0001	HITON-PC	**<0.0001**	Other
	<0.0001	HITON-PC	0.1202	Other
K2MB	**<0.0001**	HITON-PC	**<0.0001**	Other
BLCD-MB	**<0.0001**	HITON-PC	**<0.0001**	Other
FAST-IAMB	**<0.0001**	HITON-PC	**<0.0001**	Other



Simultaneous identification of causative and predictive determinants of the response variable using induction of Markov Blankets (i.e., partial causal graph induction)



New Results: HT Molecular Data

- 43 dataset-tasks
- GLL algorithm (HITON-PCnonsym instantiation) vs 35 Comparator algorithms including:
  - Univariate association + wrapping – based
  - PCA-based
  - SVM-based (RFE)
  - Random Forest –based
  - Regularized regression – based
  - Various other heuristic

43 dataset-tasks

Name	Data type	Assaying platform	Task	Num. variables	Num. samples
Adam	Proteomics mass-spectrometry	SELDI-TOF-MS	Dx	779	326
Conrads	Proteomics mass-spectrometry	High Resolution QqTOF	Dx	2190	216
Alexandrov	Proteomics mass-spectrometry	MALDI-TOF	Dx	16331	112
Ressom1	Proteomics mass-spectrometry	MALDI-TOF	Dx	214	150
Ressom3	Proteomics mass-spectrometry	MALDI-TOF	Dx	191	123
Ressom5	Proteomics mass-spectrometry	MALDI-TOF	Dx	250	129
Bhattacharjee 2	Microarray gene expression	Affymetrix HG-U95A	Dx	12600	203
Bhattacharjee 3	Microarray gene expression	Affymetrix HG-U95A	Dx	12600	160
Savage	Microarray gene expression	Affymetrix HG-133A and HG-133B	Dx	32403	210
Dave1	Microarray gene expression	Human LymphDx 2.7k GeneChip	Dx	2745	303
Dyrskjot1	Microarray gene expression	MDL Human 3k	Dx	1381	404
Miller1	Microarray gene expression	Affymetrix HG-U133A	Dx	22283	251
Miller2	Microarray gene expression	Affymetrix HG-U133A	Dx	22283	247
Miller3	Microarray gene expression	Affymetrix HG-U133A	Dx	22283	251
Vijver3	Microarray gene expression	Agilent Hu25K	Px	24496	215
Rosenwald4	Microarray gene expression	Lymphochip	Px	7399	227
Rosenwald5	Microarray gene expression	Lymphochip	Px	7399	208
Rosenwald6	Microarray gene expression	Lymphochip	Px	7399	194
Taylor2	Microarray gene expression	Affymetrix Human Exon 1.0 ST Array	Dx	43419	150
Blaser1	Microbiomics	Roche 454 sequencing	Dx	660	66
Blaser2	Microbiomics	Roche 454 sequencing	Dx	660	66
Blaser3	Microbiomics	Roche 454 sequencing	Dx	660	66

43 dataset-tasks CONT'D

Sreekumar	Metabolomics	High-throughput LC-MS and GC-MS	Dx	1061	107
Schulte	miRNA	RT-qPCR	Px	307	69
Leidinger	miRNA	Geniom Biochip miRNA	Dx	864	57
Taylor1	miRNA	Agilent-019118 Human miRNA Microarray 2.0	Dx	373	113
Landi	miRNA	CCDTM miRNA700-V3	Dx	198	290
Guo	miRNA	Tsinghua University mammalian 2K microRNA microarray	Dx	1932	257
Taylor3	aCGH	Agilent-014693 Human Genome CGH Microarray 244A	Dx	231021	218
Stransky	aCGH	UCSF Hum Array 2.0 CGH	Dx	2143	57
Trolet	aCGH	Custom 4K BAC clones array	Px	3649	78
Blaveri	aCGH	UCSF Hum Array 2.0 CGH	Dx	2142	98
Snijders	aCGH	UCSF Hum Array 2.0 CGH	Dx	1934	75
Lindgren1	aCGH	SWEGENE_BAC_32K_Full	Dx	31935	103
Lindgren2	aCGH	SWEGENE_BAC_32K_Full	Px	31935	84
Teschendorff	DNA Methylation	Illumina HumanMethylation27 BeadChip	Dx	27578	540
Christensen1	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1413	109
Christensen2	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1413	176
Christensen3	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1413	215
Holm1	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1452	174
Holm2	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1452	174
Holm3	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1452	148
Holm4	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1452	89
Holm5	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1452	78
Holm6	DNA Methylation	Illumina GoldenGate Methylation Cancer Panel I	Dx	1452	81

Experimental Results : Accuracy + Parsimony

Number of selected features

K=3

Dataset name	Dataset type	HPC_Z alpha=0.05	SVM_RFE1	SVM_RFE2	UAF_KW1	UAF_KW2	UAF_KW_FDR	UAF_SN1	UAF_SN2	UAF_BW1	UAF_BW2	UAF_T1	UAF_T2	UAF_T_FDR	UAF_X21	UAF_X22	UAF_X2_FDR	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2	SPCA1	SPCA2
Average	Proteomics	6.3	6.5	153.4	5.6	432.0	1496.5	6.5	416.8	63.8	400.5	34.4	379.4	1469.9	24.9	311.8	1857.2	21.8	396.9	5.8	45.5	230.8	22.2	119.3	199.5	1170.9	462.4	1641.1
	Microarray	9.9	11.0	1512.0	8.6	3502.5	3007.6	3.8	2864.4	3.1	3421.6	3.1	3421.6	3251.1	531.1	6338.1	5487.3	9.8	63.6	1.8	30.1	5178.1	63.2	1266.2	72.9	5432.7	3389.2	9654.6
	Microbiomics	3.2	1.7	18.7	1.5	42.7	7.4	1.1	74.1	198.0	341.0	1.4	43.3	1.7	3.1	90.9	82.5	3.5	5.7	1.2	28.7	30.7	15.5	25.5	6.0	165.0	32.9	227.9
	Metabolomics	5.4	2.1	48.6	1.0	180.1	0.1	1.2	200.8	1.3	81.7	1.3	81.7	0.0	28.9	197.3	8.8	17.5	27.4	1.2	121.3	2.0	58.2	264.8	2.6	430.7	75.7	349.0
	miRNA	4.3	3.1	127.1	5.3	378.9	381.2	7.5	322.3	8.3	174.1	8.3	174.1	395.0	11.0	142.4	466.0	12.2	28.2	2.7	24.5	68.3	26.5	66.5	130.6	480.6	262.6	514.2
	aCGH	7.2	4.2	4589.4	3.5	20552.9	15804.6	5.9	28666.2	9.9	30654.9	9.9	30654.9	19289.8	117.8	20966.7	28208.9	5.7	32.1	2.0	36.9	3396.4	3317.7	11105.2	153.1	11341.1	1643.9	10362.4
	DNA Methylation	9.1	97.7	2937.4	28.6	3026.5	1076.2	3.5	3124.2	5.3	3073.1	5.2	3073.1	541.5	744.4	1233.6	1597.4	28.7	75.0	2.2	34.9	83.8	517.4	1628.2	1131.8	3038.7	1452.6	3289.2
	Grand	**7.7**	**26.9**	**1840.4**	**10.8**	**4988.0**	**3808.9**	**4.6**	**6081.7**	**26.3**	**6537.2**	**9.2**	**6514.6**	**4300.2**	**342.5**	**5434.5**	**6633.3**	**14.9**	**97.1**	**2.5**	**35.6**	**2083.3**	**657.5**	**2485.9**	**337.9**	**4239.0**	**1652.3**	**5430.9**

Classification performance
(AUC)

K=3

		HPC_Z alpha=0.05	SVM_RFE1	SVM_RFE2	UAF_KW1	UAF_KW2	UAF_KW_FDR	UAF_SN1	UAF_SN2	UAF_BW1	UAF_BW2	UAF_T1	UAF_T2	UAF_T_FDR	UAF_X21	UAF_X22	UAF_X2_FDR	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2	SPCA1	SPCA2
Average	Proteomics	0.964	0.936	0.981	0.925	0.972	0.984	0.943	0.980	0.942	0.975	0.936	0.973	0.979	0.939	0.976	0.986	0.957	0.977	0.922	0.979	0.939	0.960	0.980	0.919	0.978	0.962	0.985
	Microarray	0.819	0.747	0.826	0.799	0.820	0.805	0.799	0.829	0.778	0.829	0.778	0.829	0.801	0.807	0.826	0.825	0.818	0.817	0.781	0.811	0.798	0.800	0.800	0.680	0.813	0.801	0.825
	Microbiomics	0.843	0.699	0.749	0.732	0.780	0.624	0.719	0.755	0.672	0.615	0.767	0.697	0.692	0.708	0.746	0.806	0.827	0.799	0.760	0.758	0.713	0.691	0.690	0.559	0.639	0.570	0.602
	Metabolomics	0.750	0.560	0.628	0.447	0.505	0.460	0.425	0.493	0.401	0.519	0.401	0.519	0.500	0.603	0.672	0.519	0.682	0.623	0.391	0.615	0.519	0.559	0.577	0.397	0.656	0.468	0.544
	miRNA	0.923	0.894	0.942	0.896	0.934	0.949	0.893	0.922	0.900	0.937	0.900	0.937	0.945	0.911	0.916	0.948	0.920	0.933	0.898	0.922	0.843	0.895	0.916	0.833	0.921	0.907	0.935
	aCGH	0.797	0.708	0.794	0.762	0.806	0.713	0.755	0.801	0.729	0.815	0.729	0.815	0.725	0.802	0.829	0.826	0.751	0.771	0.724	0.793	0.735	0.744	0.781	0.666	0.749	0.696	0.792
	DNA Methylation	0.899	0.845	0.910	0.861	0.909	0.924	0.853	0.908	0.854	0.913	0.853	0.913	0.921	0.894	0.921	0.929	0.883	0.904	0.851	0.885	0.806	0.896	0.908	0.828	0.905	0.871	0.918
	Grand	**0.865**	**0.797**	**0.864**	**0.822**	**0.861**	**0.837**	**0.820**	**0.860**	**0.807**	**0.856**	**0.812**	**0.861**	**0.842**	**0.844**	**0.869**	**0.876**	**0.849**	**0.858**	**0.810**	**0.851**	**0.802**	**0.832**	**0.846**	**0.745**	**0.842**	**0.811**	**0.853**

Experimental Results: over all data types Predictivity and Parsimony

	Predictivity		Reduction	
Feature Selection Method	P-value	Nominal winner	P-value	Nominal winner
ALL	0.5	Other	0	HITON-PC
SVM_RFE1	0	HITON-PC	0.3764	HITON-PC
SVM_RFE2	0.4508	HITON-PC	0	HITON-PC
UAF_KW1	0	HITON-PC	0.3793	HITON-PC
UAF_KW2	0.3477	HITON-PC	0	HITON-PC
UAF_KW_FDR	0.032	HITON-PC	0	HITON-PC
UAF_SN1	0	HITON-PC	0.0012	Other
UAF_SN2	0.3273	HITON-PC	0	HITON-PC
UAF_BW1	0	HITON-PC	0.0314	HITON-PC
UAF_BW2	0.2444	HITON-PC	0	HITON-PC
UAF_T1	0	HITON-PC	0.4689	HITON-PC
UAF_T2	0.3651	HITON-PC	0	HITON-PC
UAF_T_FDR	0.0496	HITON-PC	0	HITON-PC
UAF_X21	0.0085	HITON-PC	0	HITON-PC
UAF_X22	0.2633	Other	0	HITON-PC
UAF_X2_FDR	0.0868	Other	0	HITON-PC
mRMR1	0	HITON-PC	0.0011	HITON-PC
mRMR2	0.123	HITON-PC	0	HITON-PC
mRMR3	0	HITON-PC	0.0053	Other
mRMR4	0.0241	HITON-PC	0	HITON-PC
mRMR5	0	HITON-PC	0.0683	HITON-PC
mRMR6	0.1496	HITON-PC	0	HITON-PC
RFVS1	0.0107	HITON-PC	0.0163	HITON-PC
RFVS2	0.1832	HITON-PC	0	HITON-PC
LARS_EN1	0	HITON-PC	0	Other
LARS_EN2	0.0126	HITON-PC	0	HITON-PC
SIMCA	0	HITON-PC	0	HITON-PC
SIMCA_SVM1	0.0015	HITON-PC	0	HITON-PC
SIMCA_SVM2	0.0244	HITON-PC	0	HITON-PC
PCA1	0	HITON-PC	0	HITON-PC
PCA2	0.0163	HITON-PC	0	HITON-PC
SPCA1	0.0003	HITON-PC	0	HITON-PC
SPCA2	0.1763	HITON-PC	0	HITON-PC
TGDR1	0	HITON-PC	0	Other
TGDR2	0.0164	HITON-PC	0.0224	HITON-PC
TGDR3	0.0667	HITON-PC	0	HITON-PC

reference HPC method: HPC_Z, K=3, alpha=0.05

Experimental Results By Data Type: Accuracy + Parsimony

Proteomics

HPC_Z	ALL	SVM_RFE2	UAF_KW_FDR	UAF_SN2	UAF_T_FDR	UAF_X2_FDR	RFVS2	LARS_EN2	SIMCA_SVM2	PCA2	SPCA2
0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.99
23.02	3,325.83	153.35	1,496.48	416.83	1,469.90	1,857.17	396.85	45.45	119.25	1,170.90	1,641.10

Microarray

HPC_Z	ALL	SVM_RFE2	UAF_SN2	UAF_BW2	UAF_T2	UAF_X22	UAF_X2_FDR	SPCA2
0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.82
44.42	16,822.31	1,512.00	2,864.38	3,421.65	3,421.65	6,338.10	5,487.31	9,654.64

Microbiomics

HPC_Z
0.85
2.13

Metabolomics

HPC_Z
0.75
5.40

Experimental Results By Data Type: Accuracy + Parsimony CONT'D

miRNA						
HPC_Z						
0.95						
21.14						
aCGH						
HPC_Z	ALL	UAF_BW2	UAF_T2	UAF_X22	UAF_X2_FDR	mRMR6
0.81	0.83	0.82	0.82	0.83	0.83	0.81
285.17	43,537.00	30,654.93	30,654.93	20,966.66	28,208.86	53.36

DNA-Methylation													
HPC_Z	ALL	SVM_RFE2	UAF_KW2	UAF_KW_FDR	UAF_SN2	UAF_BW2	UAF_T2	UAF_T_FDR	UAF_X22	UAF_X2_FDR	mRMR2	SIMCA_SVM2	SPCA2
0.91	0.92	0.91	0.91	0.92	0.91	0.91	0.91	0.92	0.92	0.93	0.91	0.91	0.92
59.29	4,052.90	2,937.38	3,026.45	1,076.22	3,124.15	3,073.08	3,073.08	541.53	1,233.62	1,597.40	224.08	1,628.20	3,289.16

ALL		
HPC_Z	UAF_X22	UAF_X2_FDR
0.87	0.87	0.88
118.59	5,434.46	6,633.34

Experimental Results
Reproducibility

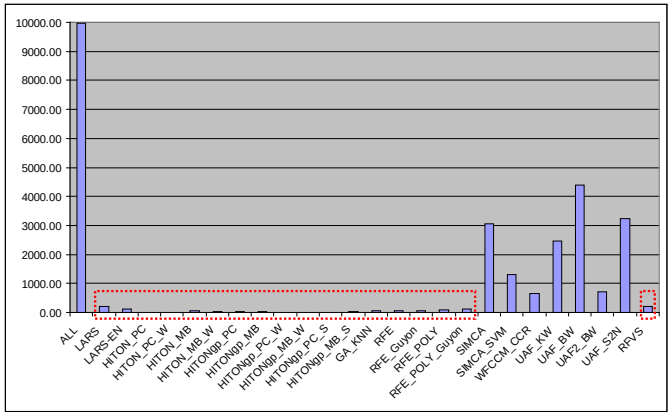
Area under ROC curve absolute nominal difference

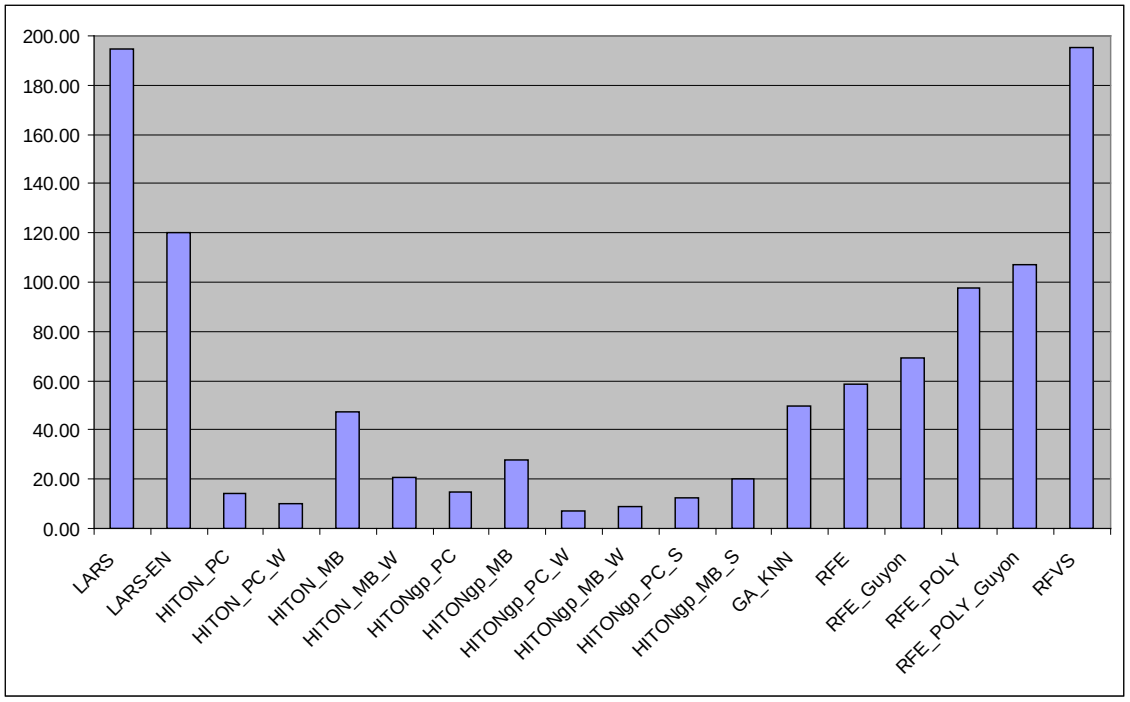
Dataset name	K=3													
	HPC_Z alpha=0.01	HPC_Z alpha=0.05	HPC_Z alpha=0.10	SVM_RFE1	SVM_RFE2	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2
Beer	0.000	**0.001**	0.000	0.000	0.000	0.008	0.004	0.002	0.003	0.002	0.000	0.000	0.019	0.130
Su	0.004	**0.002**	0.002	0.103	0.009	0.040	0.005	0.010	0.038	0.000	0.000	0.000	0.316	0.049
Sotiriou1	0.089	**0.036**	0.002	0.146	0.017	0.099	0.047	0.146	0.061	0.020	0.023	0.041	0.218	0.015
Sotiriou3	0.106	0.023	0.058	0.024	0.010	0.006	0.010	0.144	0.070	0.074	0.133	0.060	0.103	0.000
Freije	0.025	0.053	0.065	0.106	0.106	0.085	0.020	0.004	0.028	0.050	0.107	0.015	0.031	0.013
Ross3	0.156	0.005	0.118	0.149	0.149	0.018	0.121	0.186	0.083	0.068	0.099	0.099	0.141	0.017
Average	0.063	0.020	0.041	0.088	0.049	0.043	0.035	0.082	0.047	0.036	0.060	0.036	0.138	0.037
Median	0.057	0.014	0.030	0.105	0.014	0.029	0.015	0.077	0.050	0.035	0.061	0.028	0.122	0.016
Min	0.000	0.001	0.000	0.000	0.000	0.006	0.004	0.002	0.003	0.000	0.000	0.000	0.019	0.000
Max	0.156	0.053	0.118	0.149	0.149	0.099	0.121	0.186	0.083	0.074	0.133	0.099	0.316	0.130
Coefficient of variation	1.000	1.064	1.175	0.709	1.297	0.945	1.312	1.041	0.629	0.919	0.984	1.088	0.826	1.291

Area under ROC curve statistical difference

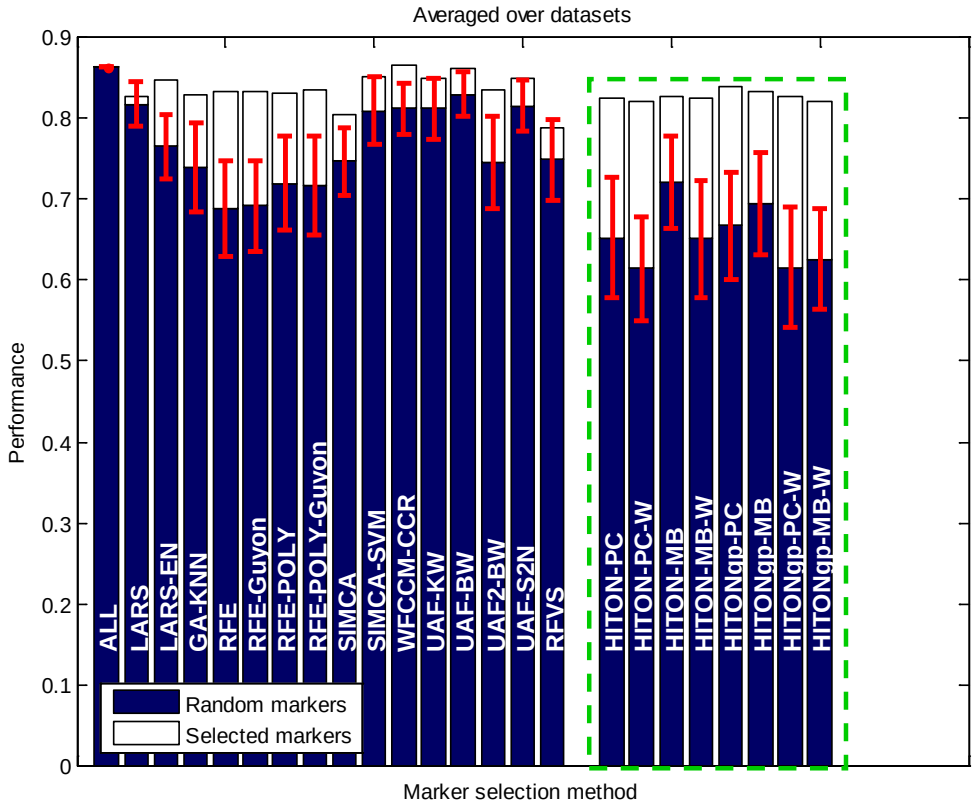
Dataset name	K=3													
	HPC_Z alpha=0.01	HPC_Z alpha=0.05	HPC_Z alpha=0.10	SVM_RFE1	SVM_RFE2	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2
Beer	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	-0.002	0.000	-0.002	0.000	0.000	0.000	0.000
Su	0.000	**0.000**	0.000	0.000	0.000	-0.007	0.000	0.000	-0.029	0.000	0.000	0.000	-0.181	-0.027
Sotiriou1	0.000	**0.000**	0.000	0.000	0.000	-0.019	-0.009	-0.074	-0.074	0.000	-0.022	0.000	0.000	0.000
Sotiriou3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Freije	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ross3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Average	0.000	0.000	0.000	0.000	0.000	-0.004	-0.002	-0.013	-0.017	0.000	-0.004	0.000	-0.030	-0.004

Experimental Results: Parsimony



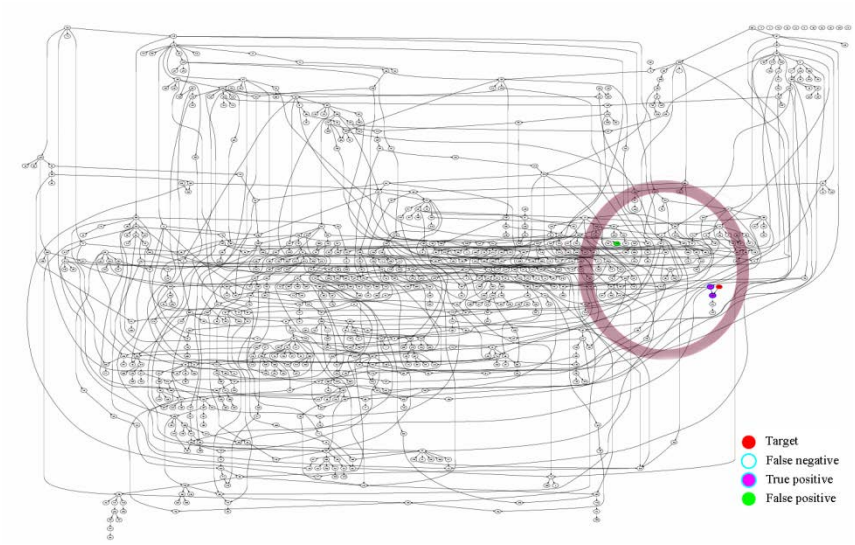


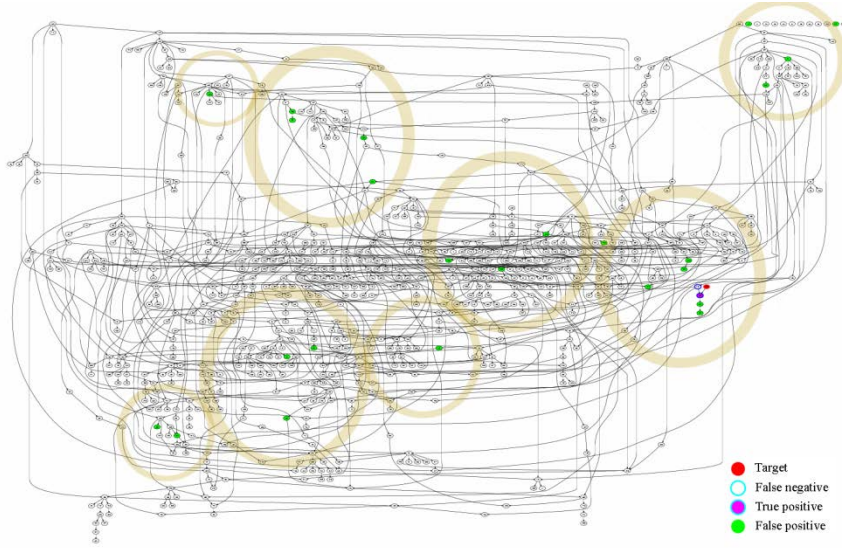
Experimental Results
Classification performance vs random selection



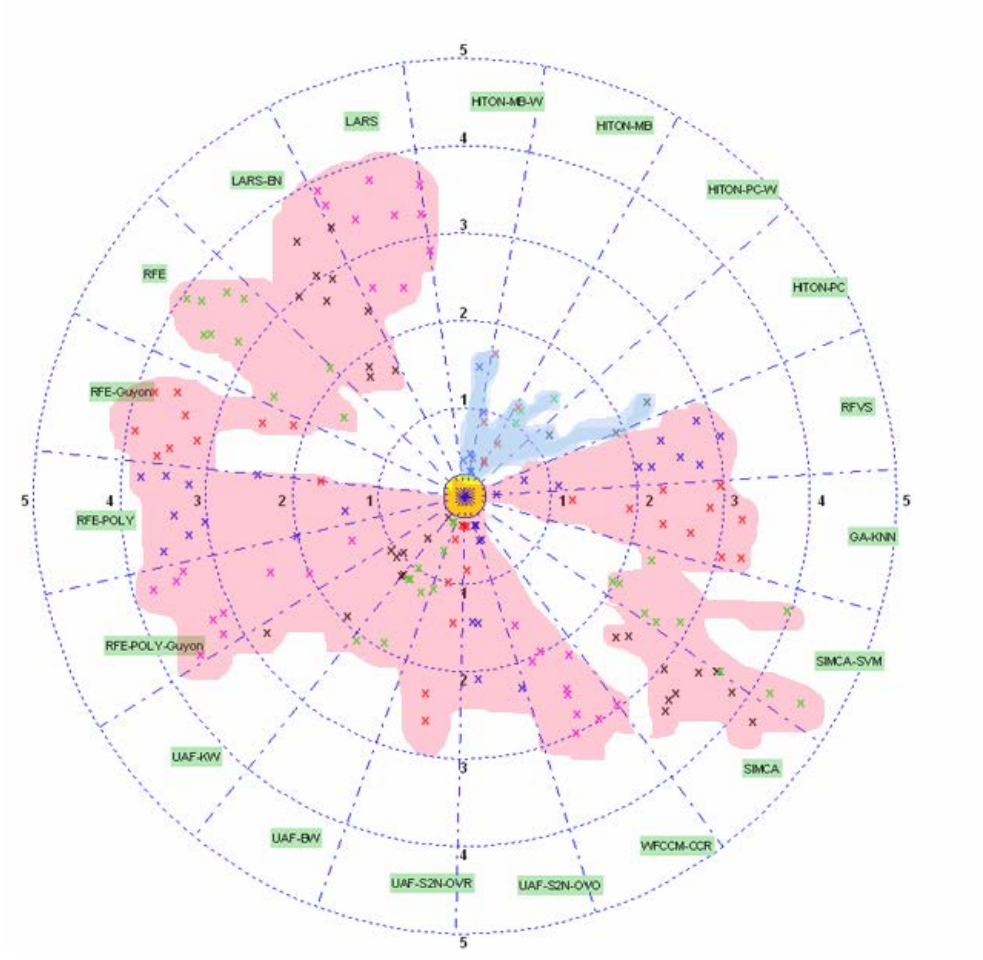
2. Network reverse-engineering methods (Causal Discovery)

Experimental Results
Pathway localization





Experimental Results
Pathway localization



Passengers, Drivers, Irrelevant

REGED with 10,000 irrelevant variables

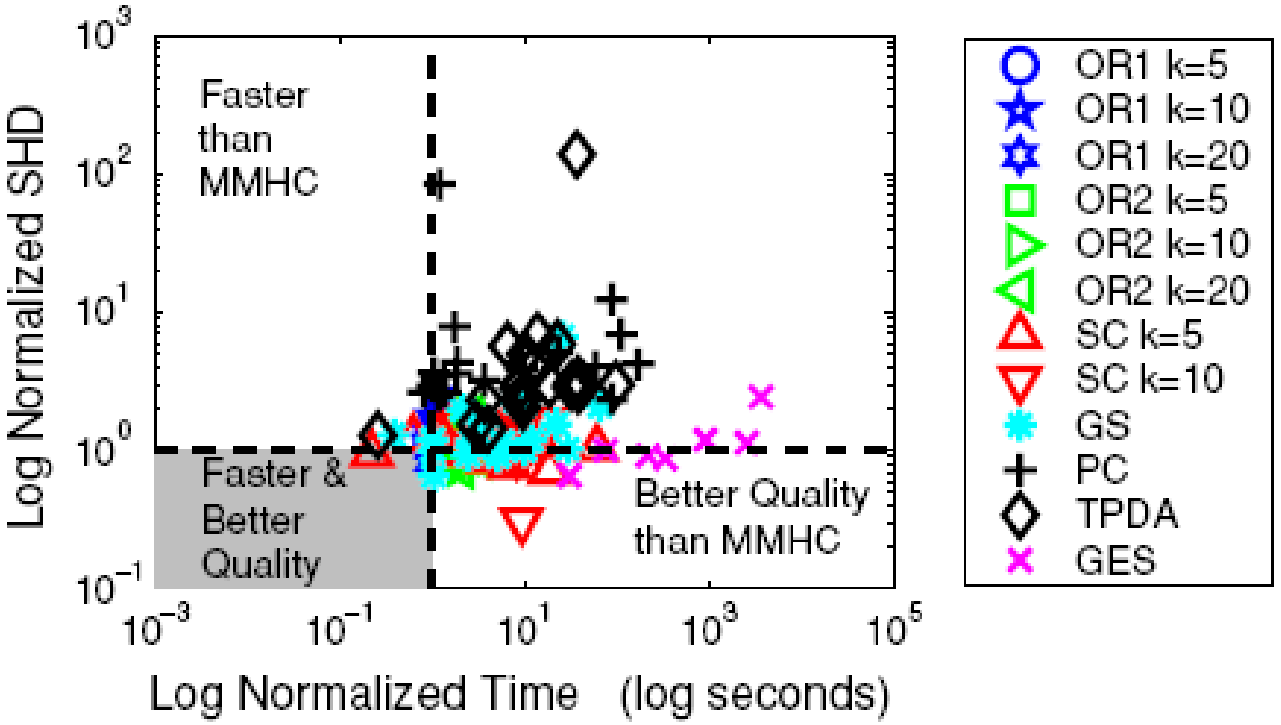
Dataset name	TPC	ALL	K=3 HPC_Z alpha=0.05	SVM_RFE1	SVM_RFE2	UAF_KW_FDR	UAF_T_FDR	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	PCA1	PCA2
AUC	1.000	0.961	1.000	0.990	0.998	0.998	0.998	0.999	1.000	0.967	1.000	0.961	0.971	0.994
Number of selected features	15	10999	15	3	5	633	646	7	18	2	24	10999	687	1375
Undirected Graph Distance	0.000	1.000	0.000	0.000	0.000	0.600	0.601	0.020	0.053	0.000	0.091	1.000	0.645	0.673
False Negative Proportion	0.0%	0.0%	13.3%	80.0%	66.7%	6.7%	6.7%	60.0%	20.0%	86.7%	13.3%	0.0%	53.3%	13.3%
False Positive Proportion	0.0%	100.0%	0.0%	0.0%	0.0%	60.6%	61.1%	0.1%	0.6%	0.0%	0.5%	100.0%	69.1%	76.3%
DC	2	2	2	1	2	2	2	1	2	1	2	2	2	2
IC	0	57	0	0	0	57	56	1	2	0	0	57	56	57
DE	13	13	11	2	3	12	12	5	10	1	11	13	5	11
IE	0	6	0	0	0	6	6	0	3	0	1	6	3	6
Passenger	0	711	0	0	0	533	538	0	1	0	4	711	621	680
IR	0	10210	2	0	0	23	32	0	0	0	6	10210	0	619

First Results: general Distributions, MMHC algorithm

- 7 algorithms (13 total variants)
- Applied to >20 simulated data from known Bayesian networks
- Key reference

"The Max-Min Hill Climbing Bayesian Network Structure Learning Algorithm". I. Tsamardinos, L.E. Brown, C.F. Aliferis. *Machine Learning*, 65:31-78, 2006.

Experimental Results – MMHC
Time-Structural errors



Recent Results: LGL-Bach

- 15 datasets and gold standards
- LGL algorithm (HITON-Back) vs 32 de-novo reverse-engineering methods that work with genome-scale observational data
- Key reference:

"A Comprehensive Assessment of Methods for De-Novo Reverse-Engineering of Genome-Scale Regulatory Networks" Varun Narendra, Nikita I. Lytkin, Constantin F. Aliferis, Alexander Statnikov. *Genomics*, 2010.

Graph:

- Aracne (2)
- Relevance Networks (3)
- SA-CLR (2)
- CLR (4)
- LGL-Bach (6)
- Hierarchical Clustering (1)
- Graphical Lasso (1)
- GeneNet (2)
- Fisher's Z (2)
- qp-graphs (5)

Likelihood of interactions:

- Mutual Information (2)
- SA-CLR (1)
- CLR (2)
- GeneNet (1)
- qp-graphs (5)
- Fisher's Z (1)

Comparator Methods by family

Univariate:

- Relevance Networks (3)
- CLR (4)
- Fisher's Z (2)
- Mutual Information (2)

Random/control:

- Full graph (1)
- Empty graph (1)

Multivariate:

- Aracne (2)
- SA-CLR (2)
- Hierarchical Clustering (1)
- LGL-Bach (6)
- Graphical Lasso (1)
- GeneNet (2)
- qp-graphs (5)

5 simulated datasets and gold-standards

Dataset	Gold-Standard					Gene expression data	
	Description	No. of TFs	No. of genes	No. of edges	Description	No. of arrays	No. of genes
REGED	REGED network	-	1,000	1,148	First 500 instances from REGED dataset	500	1,000
GNW(A)	Yeast regulatory network from GNW 2.0	157	4,441	12,864	25 time series with 21 time points in each generated by GNW 2.0	525	4,441
GNW(B)	1000-gene subnetwork of Yeast regulatory network from GNW 2.0	68	1,000	3,221	25 time series with 21 time points in each generated by GNW 2.0	525	1,000
GNW(C)	E.coli network from GNW 2.0	166	1,502	3,476	25 time series with 21 time points in each generated by GNW 2.0	525	1,502
GNW(D)	1000-gene subnetwork of E.coli regulatory network from GNW 2.0	121	1,000	2,361	25 time series with 21 time points in each generated by GNW 2.0	525	1,000

10 real datasets and gold-standards

Dataset	Gold-Standard				Gene expression data		
	Description	No. of TFs	No. of genes	No. of edges	Description	No. of arrays	No. of genes
ECOLI(A)	TF-gene interactions from RegulonDB 6.4 (strong evidence)	140	1,053	1,982	E.coli gene expression dataset from Many Microbe Microarrays Database	907	4,297
ECOLI(B)	TF-gene interactions from RegulonDB 6.4 (strong and weak evidence)	174	1,465	3,399			
ECOLI(C)	DREAM2 TF-gene network from RegulonDB 6.0	152	1,135	3,070			
ECOLI(D)	DREAM2 TF-gene network from RegulonDB 6.0	152	1,146	3,091	E.coli gene expression dataset from DREAM2	300	3,456
YEAST(A)	TF-gene interactions from the Fraenkel lab, (α = 0.001, C = 0)	116	2,779	6,455	Yeast gene expression dataset from Many Microbe Microarrays Database	530	5,520
YEAST(B)	TF-gene interactions from the Fraenkel lab, (α = 0.001, C = 1)	115	2,295	4,754			
YEAST(C)	TF-gene interactions from the Fraenkel lab, (α = 0.001, C = 2)	115	1,949	3,667			
YEAST(D)	TF-gene interactions from the Fraenkel lab, (α = 0.005, C = 0)	116	3,508	10,915			
YEAST(E)	TF-gene interactions from the Fraenkel lab, (α = 0.005, C = 1)	115	2,872	7,491			
YEAST(F)	TF-gene interactions from the Fraenkel lab, (α = 0.005, C = 2)	115	2,372	5,448			

More on real gold-standards

- Several studies estimated that E. Coli and Yeast gold-standards capture up to 80-90% of all TF-gene relations.
- TF-DNA binding interactions do not always imply functional changes in gene expression.
- Condition-dependent transcription and possible mismatch with gene expression data.
- Small changes in expression cannot be reliably detected by microarrays.
- Cellular aggregation and sampling from mixtures of distributions can hide statistical relations.

Empirical evaluation: causal (mechanism) discovery. Combined PPV/NPV

Method		REGED	GNW(A)	GNW(B)	GNW(C)	GNW(D)	ECOLI(A)	ECOLI(B)	ECOLI(C)	ECOLI(D)	YEAST(A)	YEAST(B)	YEAST(C)	YEAST(D)	YEAST(E)	YEAST(F)
Aracne	$\alpha = 10^{-7}$	0.350	0.796	0.725	0.840	0.864	0.851	0.862	0.826	0.858	0.969	0.970	0.972	0.958	0.962	0.963
	$\alpha = 0.05$	0.826	0.802	0.739	0.841	0.868	0.851	0.862	0.826	0.858	0.969	0.970	0.972	0.958	0.962	0.963
Relevance Networks 1	$\alpha = 10^{-7}$	0.995	0.953	0.888	0.965	0.942	0.985	0.985	0.980	0.975	0.980	0.982	0.983	0.973	0.977	0.980
	$\alpha = 0.05$	0.997	0.981	0.950	0.985	0.979	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.977	0.980
Relevance Networks 2		0.994	0.937	0.903	0.954	0.948	0.984	0.984	0.979	0.968	0.979	0.981	0.983	0.973	0.977	0.979
SA-CLR	$\alpha = 0.05$	0.976	0.944	0.880	0.949	0.933	0.960	0.963	0.956	0.953	0.978	0.980	0.982	0.972	0.976	0.978
	FDR = 0.05	0.718	0.858	0.762	0.873	0.868	0.899	0.908	0.893	0.882	0.970	0.971	0.974	0.962	0.965	0.968
CLR	Normal MI estimator; $\alpha = 0.05$	0.963	0.928	0.850	0.933	0.913	0.951	0.957	0.947	0.947	0.979	0.981	0.982	0.973	0.977	0.978
	Normal MI estimator; FDR = 0.05	0.693	0.846	0.737	0.855	0.849	0.887	0.901	0.879	0.888	0.972	0.972	0.974	0.965	0.969	0.970
	Stouffer MI estimator; $\alpha = 0.05$	0.975	0.934	0.858	0.939	0.920	0.959	0.963	0.955	0.953	0.979	0.981	0.982	0.973	0.977	0.978
	Stouffer MI estimator; FDR = 0.05	0.736	0.858	0.751	0.866	0.859	0.911	0.922	0.907	0.905	0.974	0.975	0.976	0.967	0.971	0.972
LGL-Bach	max-k = 1, w/o symmetry	0.185	0.528	0.665	0.720	0.788	0.552	0.577	0.495	0.611	0.949	0.956	0.950	0.936	0.944	0.935
	max-k = 2, w/o symmetry	0.141	0.571	0.655	0.724	0.565	0.429	0.400	0.356	0.568	0.939	0.941	0.940	0.930	0.942	0.938
	max-k = 3, w/o symmetry	0.127	0.553	0.655	0.734	0.559	0.540	0.521	0.403	0.578	0.928	0.937	0.927	0.921	0.938	0.928
	max-k = 1, with symmetry	0.173	0.528	0.663	0.722	0.790	0.600	0.609	0.508	0.608	0.950	0.957	0.951	0.938	0.945	0.936
	max-k = 2, with symmetry	0.105	0.556	0.655	0.712	0.566	0.509	0.494	0.415	0.557	0.931	0.934	0.923	0.926	0.935	0.921
	max-k = 3, with symmetry	0.087	0.524	0.616	0.522	0.543	0.465	0.439	0.378	0.559	0.941	0.938	0.932	0.927	0.933	0.921
Hierarchical Clustering		0.996	0.944	0.850	0.950	0.914	0.960	0.964	0.956	0.956	0.979	0.981	0.982	0.973	0.976	0.979
Graphical Lasso		0.801	0.393	0.384	0.608	0.686	0.805	0.840	0.786	0.301	0.970	0.973	0.973	0.964	0.969	0.966
GeneNet	$\alpha = 0.05$	0.975	0.974	0.938	0.982	0.972	0.965	0.971	0.961	0.961	0.971	0.972	0.973	0.963	0.967	0.969
	FDR = 0.05	0.805	0.970	0.943	0.977	0.969	0.895	0.912	0.887	0.891	0.960	0.961	0.961	0.951	0.956	0.956
qp-graphs	q = 1	0.996	0.979	0.946	0.984	0.977	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.977	0.980
	q = 2	0.996	0.980	0.949	0.985	0.978	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.978	0.980
	q = 3	0.996	0.981	0.949	0.985	0.979	0.986	0.986	0.981	0.981	0.980	0.984	0.985	0.973	0.978	0.981
	q = 20	0.995	0.981	0.950	0.985	0.979	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.977	0.980
	q = 200	0.996	0.979	0.949	0.983	0.977	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.977	0.980
Fisher	$\alpha = 0.05$	0.996	0.975	0.935	0.980	0.972	0.984	0.985	0.979	0.978	0.980	0.982	0.983	0.973	0.977	0.980
	FDR = 0.05	0.996	0.973	0.932	0.979	0.971	0.984	0.985	0.979	0.978	0.980	0.982	0.984	0.973	0.977	0.980
Full Graph		0.998	0.981	0.952	0.985	0.979	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.977	0.980
Empty Graph		0.998	0.981	0.952	0.985	0.979	0.986	0.986	0.981	0.981	0.980	0.982	0.983	0.973	0.977	0.980

Caveat: LGL-Bach output are most likely to be TFs. LGL-Bach non-returned variables are most likely to not be TFs. However other methods will return more complete sets at the expense of many false negatives.

3. Signature/Marker Multiplicity

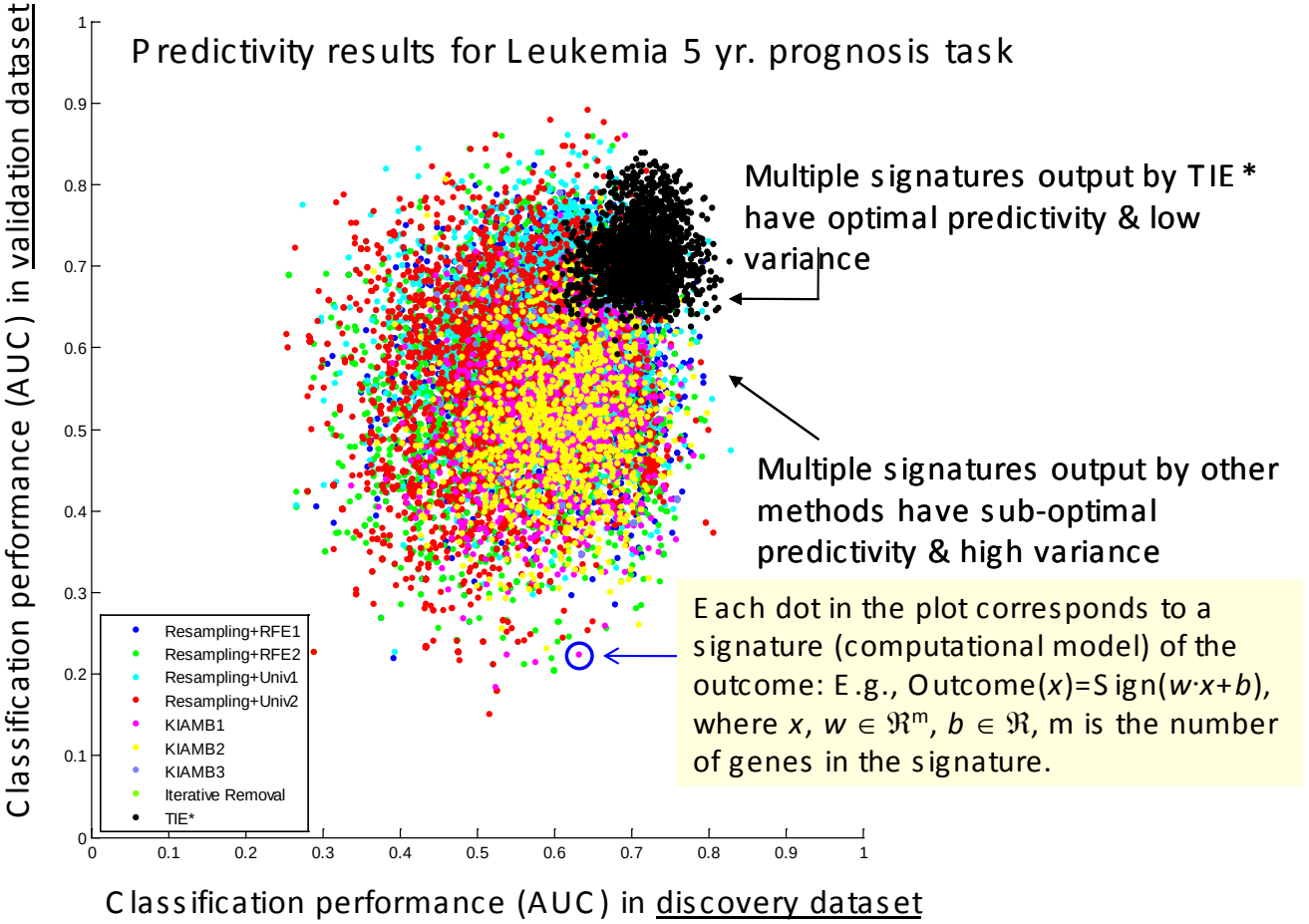
Key reference:

Statnikov A, Aliferis CF. Analysis and Computational Dissection of Molecular Signature Multiplicity. *PLoS Computational Biology* 2010, 6:e1000790.

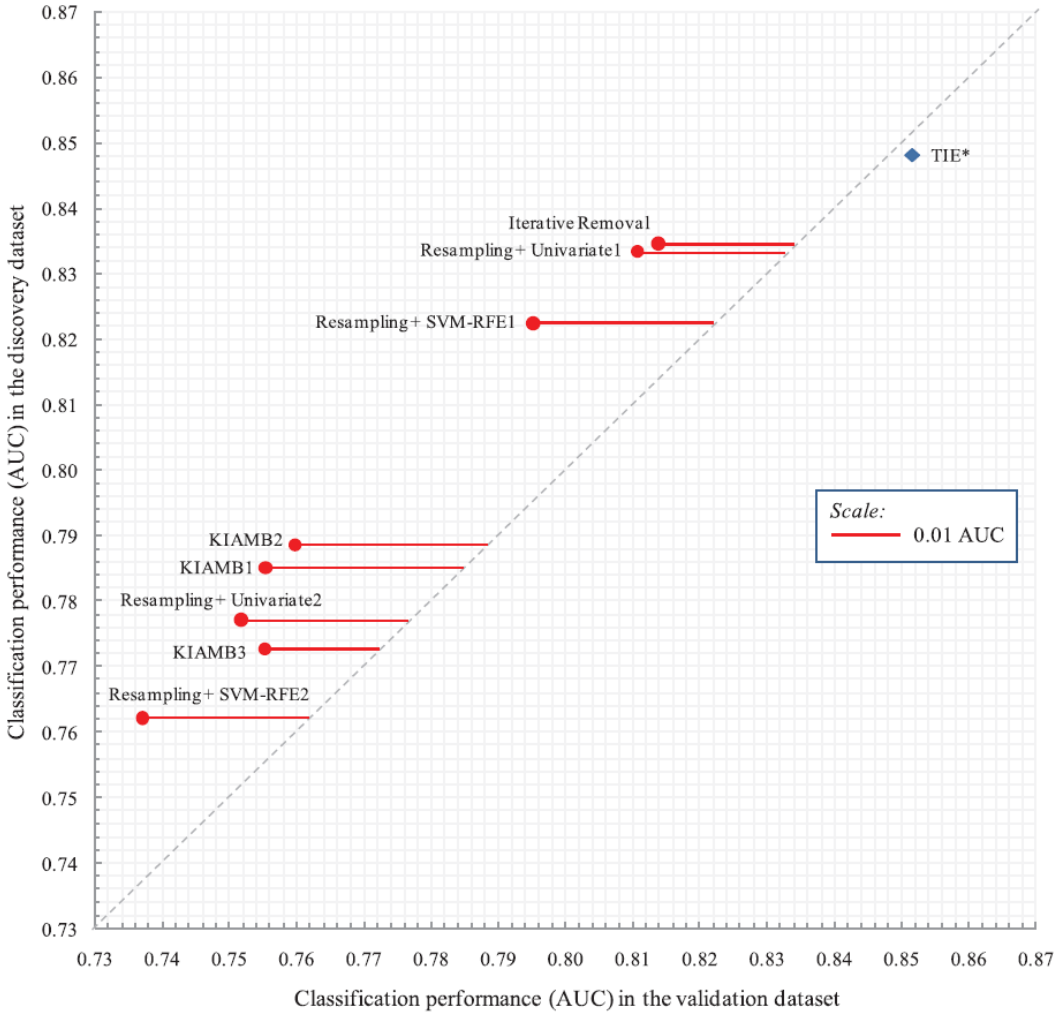
Empirical evaluation: multiplicity

Discovery of not just one of possibly many optimally predictive and maximally compact models but also *all such predictive models that are maximally predictive, and non-redundant*.

TIE* signatures in comparison with other signatures



Empirical evaluation: multiplicity

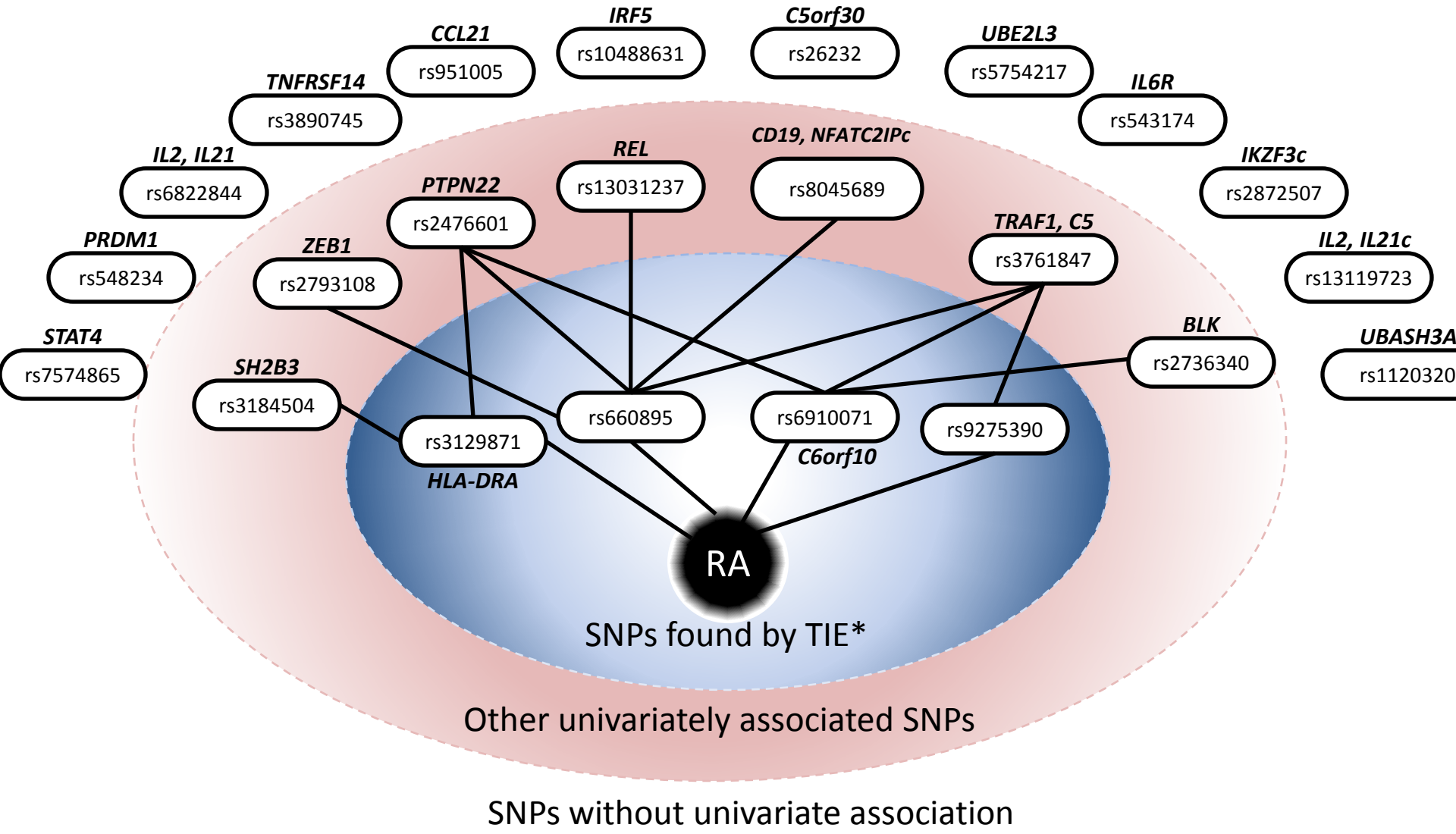


4. Example Recent Applications from NYU

Here are some references with recent GLL/TIE* applications:

- Lytkin NI, McVoy L, Weitkamp JH, Aliferis CF, Statnikov A. Expanding the Understanding of Biases in Development of Clinical-Grade Molecular Signatures: A Case Study in Acute Respiratory Viral Infections. *PLoS ONE*, 2011; 6(6): e20662.
- Alekseyenko AV, Lytkin NI, Ai J, Ding B, Padyukov L, Aliferis CF, Statnikov A. Causal Graph-Based Analysis of Genome-Wide Association Data in Rheumatoid Arthritis. *Biology Direct*, 2011 May; 6(1): 25.
- Narendra V, Lytkin NI, Aliferis CF, Statnikov A. A Comprehensive Assessment of Methods for De-Novo Reverse-Engineering of Genome-Scale Regulatory Networks. *Genomics*, 2011 Jan; 97(1): 7-18.
- Statnikov A, Lytkin NI, McVoy L, Weitkamp JH, Aliferis CF. Using Gene Expression Profiles from Peripheral Blood to Identify Asymptomatic Responses to Acute Respiratory Viral Infections. *BMC Research Notes*, 2010 Oct; 3(1): 264.
- Statnikov A, McVoy L, Lytkin N, Aliferis CF. Improving Development of the Molecular Signature for Diagnosis of Acute Respiratory Viral Infections. *Cell Host & Microbe*, 2010 Feb; 7(2): 100-1.

Application in GWAS



Causal Model Guided Experimental Minimization and Adaptive Data Collection

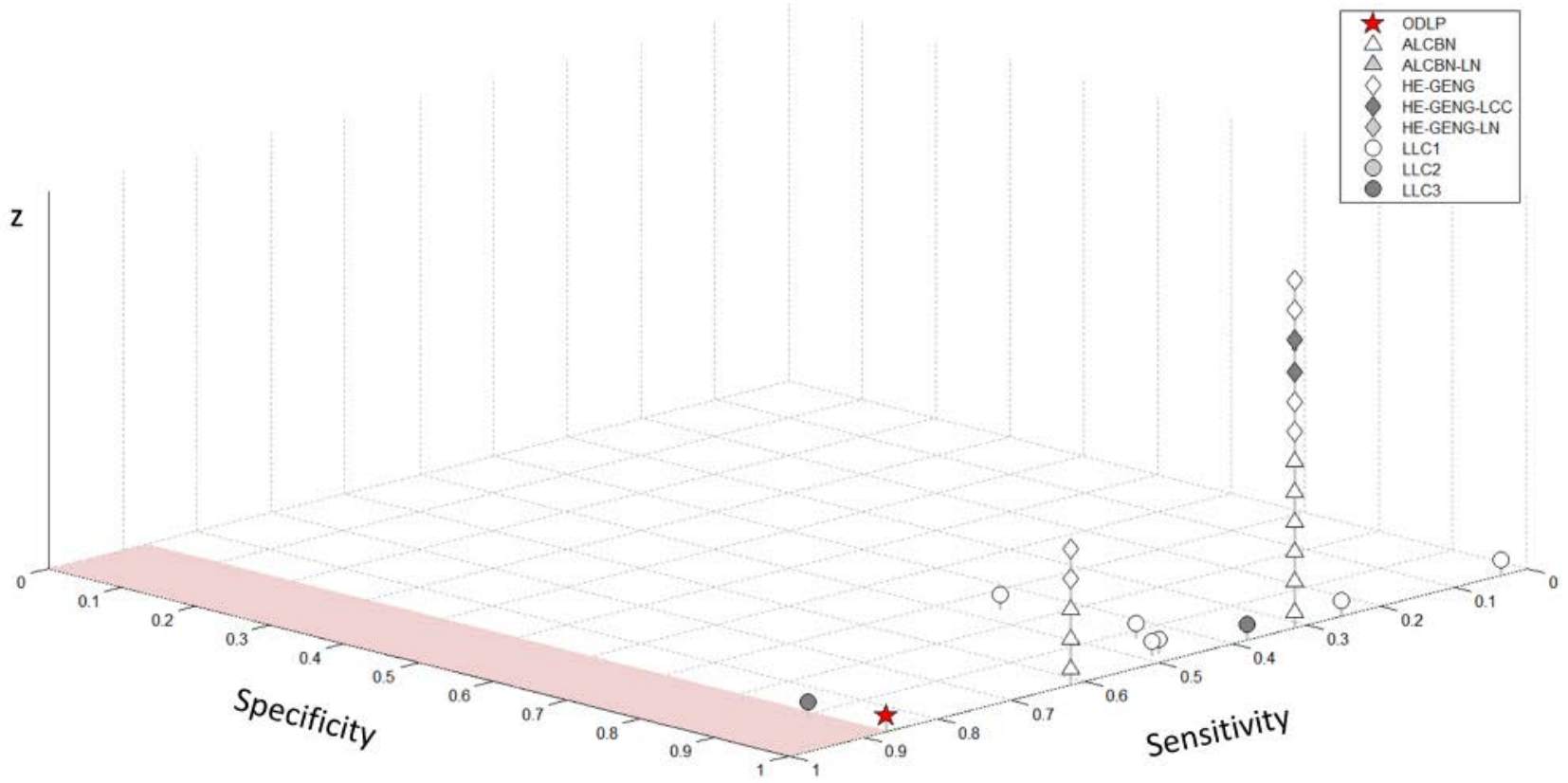
ODLP vs Other Algorithms: Performance on Simulated Data

- Benchmark study
- 58 algorithms/variant from 4 algorithm families.
- 11 networks of different sizes.

Statnikov et al., 2015 (Accepted in JMLR)

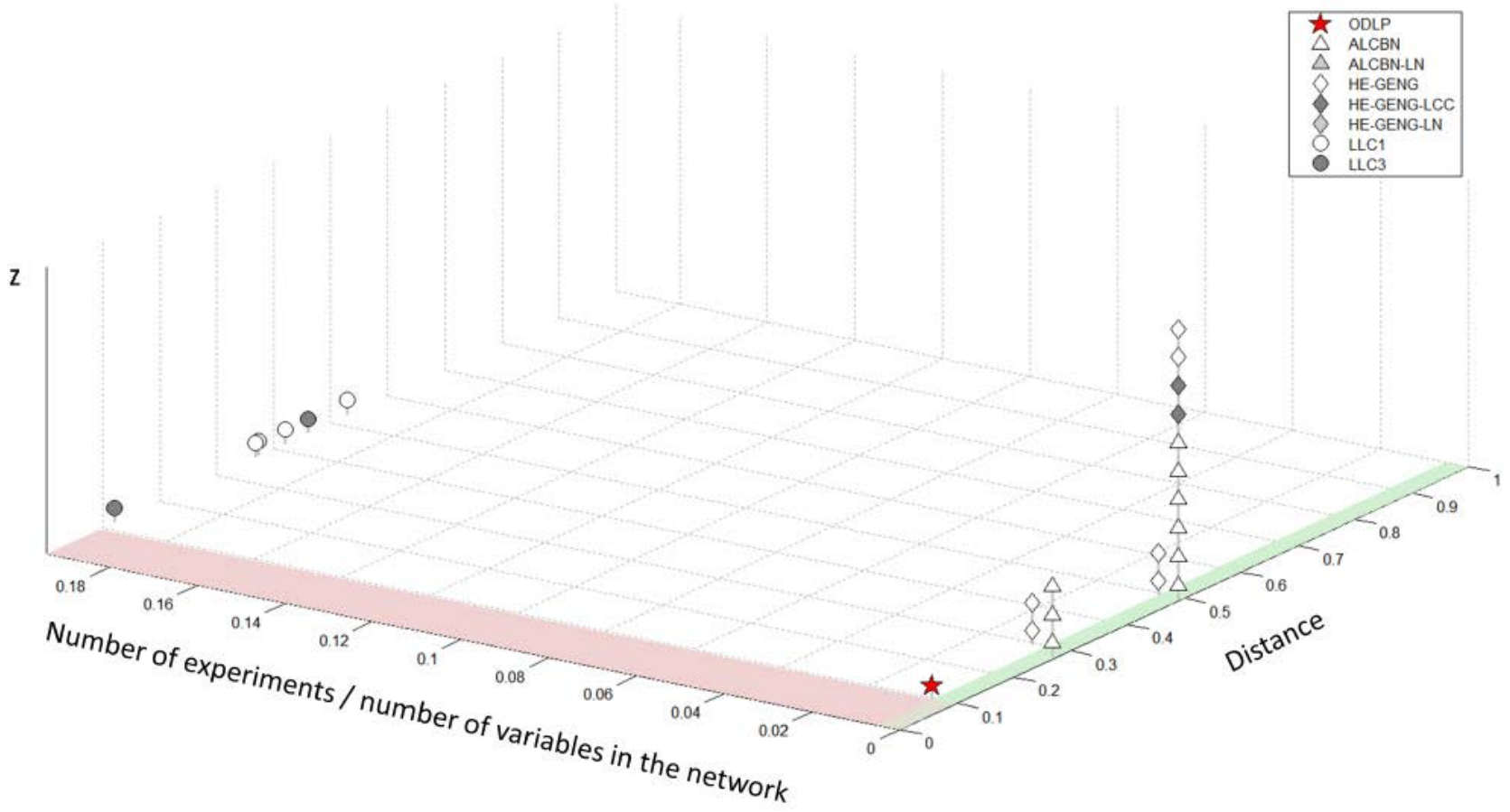
Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP vs Other Algorithms: Network Reconstruction Quality



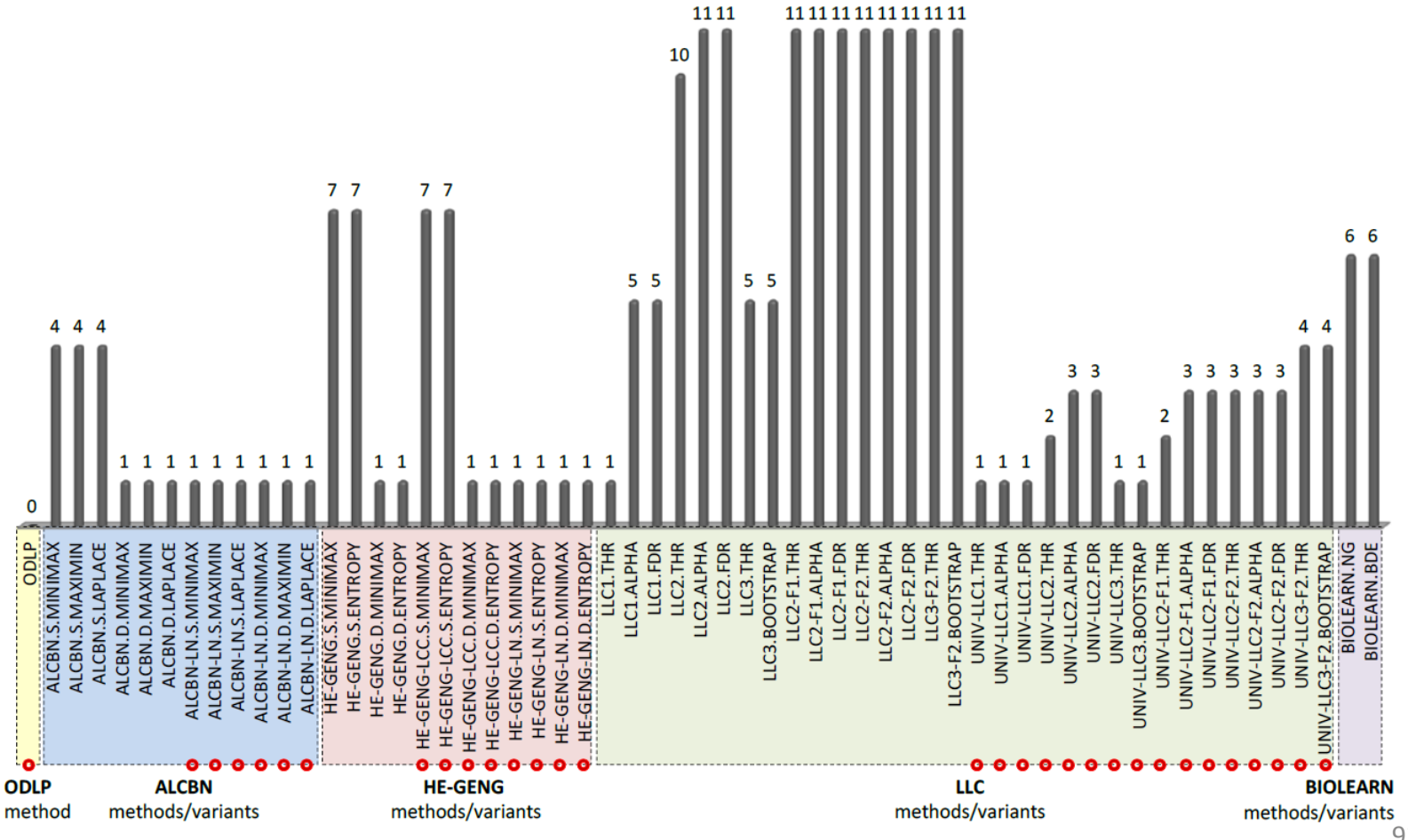
Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP vs Other Algorithms: Reconstruction Quality & Efficiency



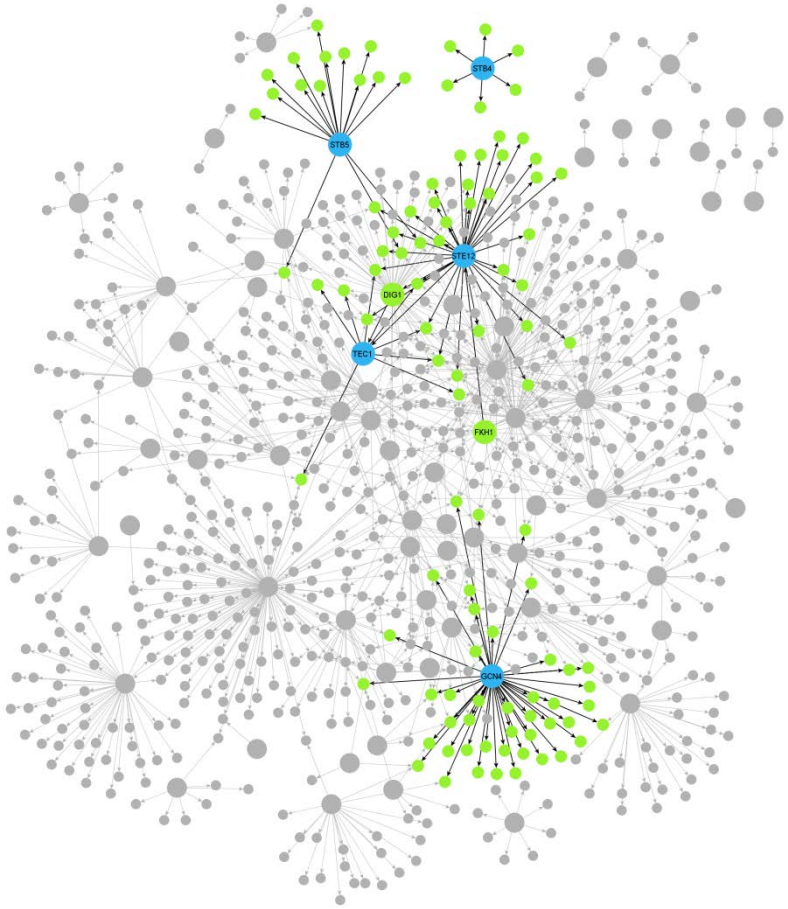
Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP vs Other Algorithms: Scalability



Causal Model Guided Experimental Minimization and Adaptive Data Collection

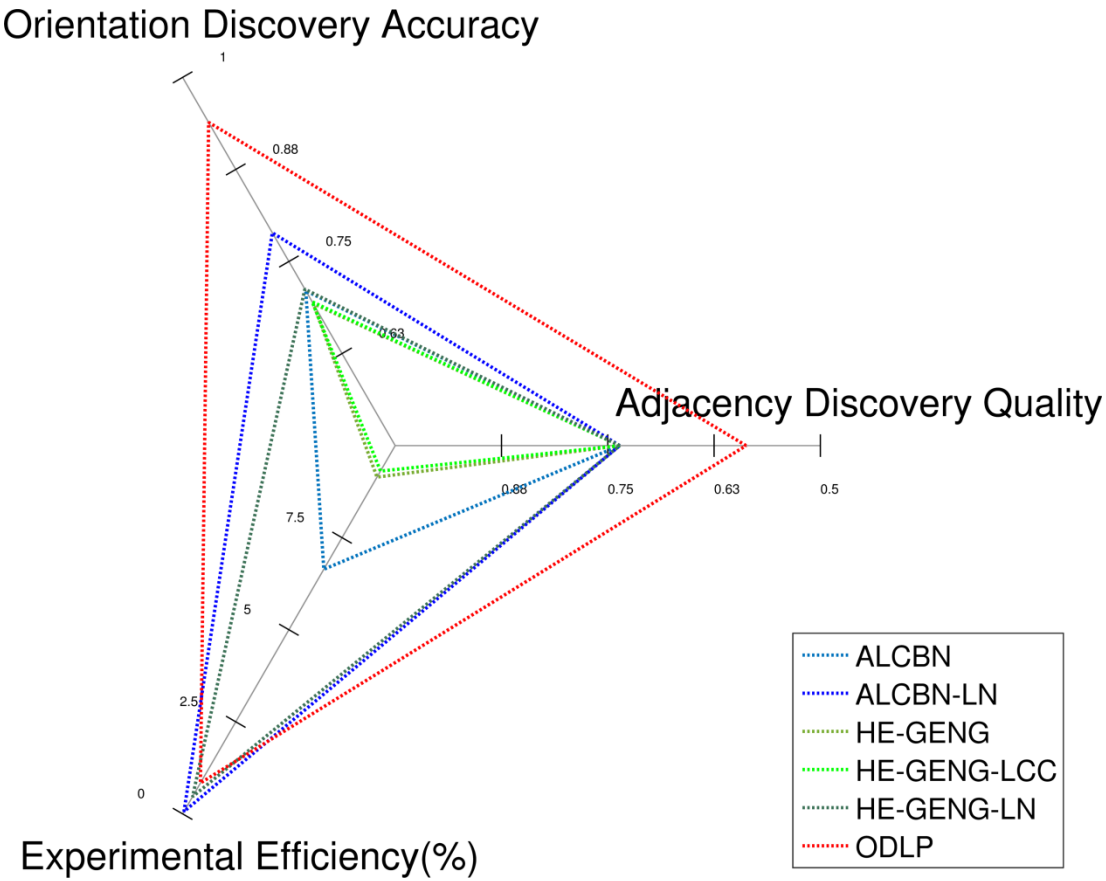
ODLP vs Other Algorithms: Performance on Real Biological Data



Ma et al., 2015 (submitted)

Causal Model Guided Experimental Minimization and Adaptive Data Collection

ODLP vs Other Algorithms: Performance on Real Biological Data



Empirical evaluation: control of false positives

Reduction of false discovery rate with superior sensitivity and specificity than traditional FDR control

Number of false positives (within irrelevant variables) in the parents and children set for features selected by HITON-PC with parameter *max-k*={0,1,2,3,4} on different training sample sizes {100, 200, 500, 1000, 2000, 5000}. The color of each table cell denotes number of false positives with green corresponding to smaller values and red to larger ones.

Lung_Cancer

	Version 1 *(original network)*					***Version 2*** *(original network + irrelevant variables)*					***Version 3*** *(weakened signal + irrelevant variables)*					***Version 4*** *(only irrelevant variables)*				
	max-k parameter																			
Sample size	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
100	0.20	0.00	0.00	0.00	0.00	411.60	1.60	1.50	1.50	1.50	488.80	11.70	8.60	8.60	8.60	411.60	12.70	9.80	9.80	9.80
200	1.50	0.00	0.00	0.00	0.00	488.60	1.20	0.00	0.00	0.00	471.60	14.90	2.90	3.00	3.00	488.60	17.30	5.80	5.50	5.50
500	0.20	0.00	0.00	0.00	0.00	446.00	2.10	0.00	0.00	0.00	424.90	13.30	0.90	1.20	1.40	446.00	28.10	6.40	5.00	4.90
1000	0.50	0.00	0.00	0.00	0.00	422.70	1.60	0.00	0.00	0.00	413.20	12.70	0.20	0.30	0.30	422.70	31.20	6.90	5.30	5.10
2000	0.80	0.00	0.00	0.00	0.00	409.00	1.60	0.00	0.00	0.00	407.90	11.10	0.40	0.00	0.00	409.00	31.80	6.10	4.00	4.00
5000	0.70	0.00	0.00	0.00	0.00	403.10	1.70	0.00	0.00	0.00	397.80	11.80	0.00	0.00	0.00	403.10	30.90	6.20	4.70	4.10

Alarm10

	Version 1 *(original network)*					***Version 2*** *(original network + irrelevant variables)*					***Version 3*** *(weakened signal + irrelevant variables)*					***Version 4*** *(only irrelevant variables)*				
	max-k parameter																			
Sample size	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
100	0.00	0.00	0.00	0.00	0.00	392.10	23.00	22.80	22.80	22.80	408.70	26.20	26.40	26.40	26.40	392.10	23.30	23.40	23.40	23.40
200	0.00	0.00	0.00	0.00	0.00	412.90	5.70	3.80	3.80	3.80	427.80	10.30	6.50	6.50	6.50	412.90	19.30	9.70	9.70	9.70
500	0.00	0.00	0.00	0.00	0.00	411.60	3.90	0.80	0.80	0.80	417.90	14.80	4.40	3.90	3.80	411.60	24.40	6.80	6.60	6.60
1000	0.00	0.00	0.00	0.00	0.00	414.10	2.40	0.90	0.60	0.60	399.90	12.60	3.30	2.80	2.70	414.10	22.70	7.20	6.40	6.30
2000	0.00	0.00	0.00	0.00	0.00	382.00	1.60	0.00	0.00	0.00	380.00	10.10	1.80	1.60	1.50	382.00	25.00	8.80	6.50	5.90
5000	0.00	0.00	0.00	0.00	0.00	381.00	1.40	0.10	0.00	0.00	367.10	7.70	1.00	0.30	0.30	381.00	22.90	6.10	5.00	4.90

Small number of false positives *Large number of false positives*

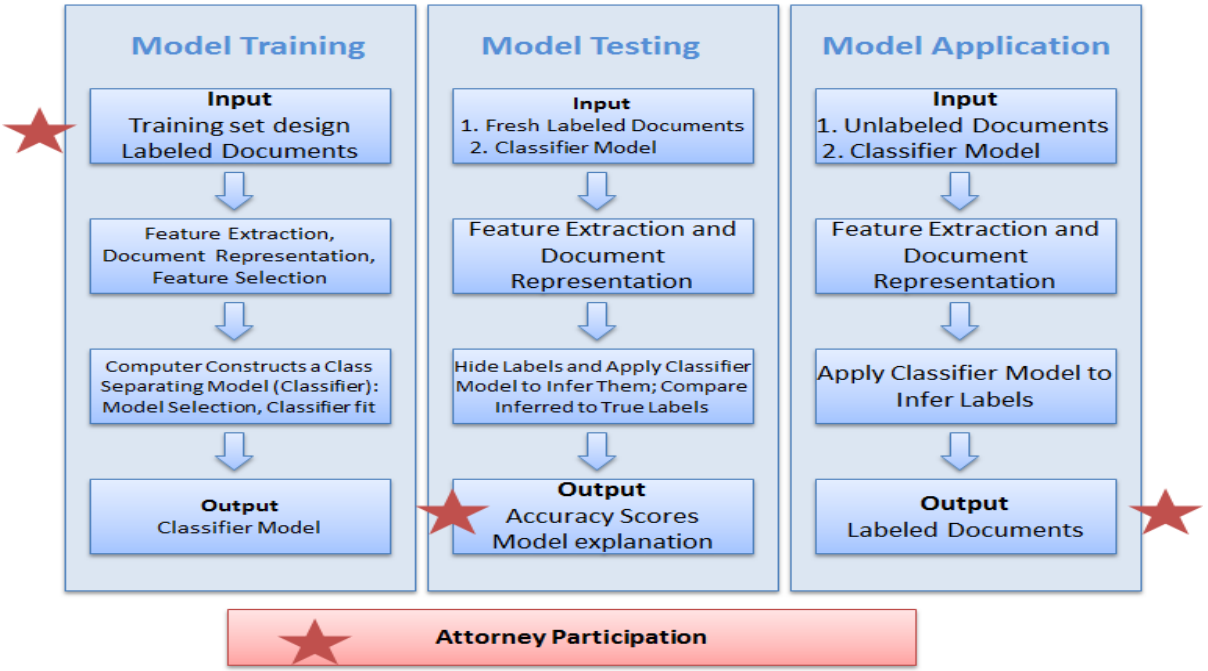
APPLICATION/PROVING GROUND #2: LEGAL PREDICTIVE CODING

Limitations of Human Legal Document Review

- Error-prone
  - Variation in reviewer expertise
  - Intra- and inter-reviewer coding variation
  - Review overconfidence in performance
  - Limitations of adjunctive key word searches

- Expensive
- Time consuming

Predictive Coding: A Great Example of Value of Big Data Analytics



When implemented correctly: **Faster** (often by a factor of 10 or more), **cheaper** (often by a factor of 10 or more), **more accurate** (from about 60-70% accuracy to neighborhood of 95%)

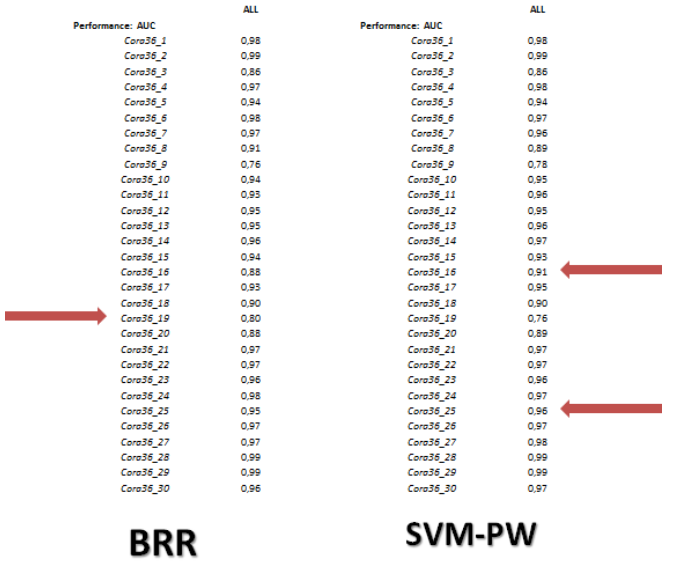
A few Key Findings
I. Not All Methods Are (or Perform) the Same

- Results from largest text categorization benchmark in text categorization ever produced
- >240 dataset-tasks
- 30 classification x 20 feature selection algorithms = 600 main analysis protocols (including commercial engines from Oracle, Google, IBM/SPSS, SAP)
- 4 loss functions
- Nested repeated N-fold cross validation:
  - –ensures rich exploration of different ways to parameterize core models;
  - –ensures avoidance of over fitting/accurate estimation of predictive accuracy
- =>millions of models built & tested, 10,000s of state-of-the-art data analysis setups evaluated

A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization
Aphinyanaphongs, Yindalon; Fu, Lawrence D; Li, Zhiguo; Peskin, Eric R; Efstathiadis, Efstratios; Aliferis, Constantin F; Statnikov, Alexander 2014 OCT;65(10):1964-1987, Journal of the Association for Information Science & Technology id: 1313832, year: 2014, vol: 65, page: 1964

A few Key Findings
I. Not All Methods Are (or Perform) the Same

	AUC	Precision	Recall	F
SVM_LibSVM_Linear_Fixed_C=1	0,95	0,78	0,57	0,64
SVM_LibSVM_Linear_Optimized_C	0,95	0,61	0,39	0,45
SVM_LibSVM_Poly_Optimized_C	0,96	0,65	0,41	0,48
SVM_LibSVM_Weighted_Linear_Fixed_C=1	0,95	0,75	0,61	0,64
SVM_LibSVM_Weighted_Linear_Optimized_C	0,96	0,72	0,62	0,63
SVM_LibSVM_Weighted_Poly_Optimized_C	0,96	0,72	0,61	0,62
SVM_LibLinear_Linear_Fixed_C=1	0,92	0,76	0,55	0,62
SVM_LibLinear_Linear_Optimized_C	0,93	0,77	0,55	0,62
1SVM_LibLinear_Optimized_C	0,93	0,75	0,64	0,68
KRR_Poly_Optimized_C	0,95	0,69	0,24	0,32
Naive_Bayes	0,79	0,61	0,62	0,57
LR_LibLinear_L1_Regularized_Optimized_C	0,93	0,78	0,62	0,68
LR_LibLinear_L2_Regularized_Optimized_C	0,93	0,80	0,52	0,60
BBR	0,96	0,83	0,50	0,59
Google_Prediction_API	0,91	0,59	0,46	0,50



Performance: AUC	ALL
Cora36_1	0,98
Cora36_2	0,99
Cora36_3	0,86
Cora36_4	0,97
Cora36_5	0,94
Cora36_6	0,98
Cora36_7	0,97
Cora36_8	0,91
Cora36_9	0,76
Cora36_10	0,94
Cora36_11	0,93
Cora36_12	0,95
Cora36_13	0,95
Cora36_14	0,96
Cora36_15	0,94
Cora36_16	0,88
Cora36_17	0,93
Cora36_18	0,90
Cora36_19	0,80
Cora36_20	0,88
Cora36_21	0,97
Cora36_22	0,97
Cora36_23	0,96
Cora36_24	0,98
Cora36_25	0,95
Cora36_26	0,97
Cora36_27	0,97
Cora36_28	0,99
Cora36_29	0,99
Cora36_30	0,96

BRR

Performance: AUC	ALL
Cora36_1	0,98
Cora36_2	0,99
Cora36_3	0,86
Cora36_4	0,98
Cora36_5	0,94
Cora36_6	0,97
Cora36_7	0,96
Cora36_8	0,89
Cora36_9	0,78
Cora36_10	0,95
Cora36_11	0,96
Cora36_12	0,95
Cora36_13	0,96
Cora36_14	0,97
Cora36_15	0,93
Cora36_16	0,91
Cora36_17	0,95
Cora36_18	0,90
Cora36_19	0,76
Cora36_20	0,89
Cora36_21	0,97
Cora36_22	0,97
Cora36_23	0,96
Cora36_24	0,97
Cora36_25	0,96
Cora36_26	0,97
Cora36_27	0,98
Cora36_28	0,99
Cora36_29	0,99
Cora36_30	0,97

SVM-PW

A Few Key Findings
I. Not All Methods Are (or Perform) the Same

1. **SVMs, KRR, and BLR are the best performing classifier** algorithms on average
2. There is **no single dominant classification algorithm** over all datasets
3. **Markov Boundary feature selection achieves best data compression** without compromising on accuracy.
4. **Loss functions affect classifier rankings** (or may require tuning).
5. **It is not only the technology but how it is implemented. e.g., Oracle auto classifier.**
6. **Google analytics platform consistently poor performer (better only than Naïve Bayes)**.
7. IBM/SPSS/SAP auto-classifier requires extensive user-provided setup, and is very buggy.
8. **Active Learning often overfits**.
9. **Ensembling** (i.e., combining results from several classifiers) as implemented in Google analytics and IBM/SPSS modeler does not lead to dominant performance.
10. **PLSA methods produce models with highly unstable** classification performance.
11. **TREC competition datasets** and the performance of winners in that competition are **not as informative as a full-scale benchmark**.
12. **Small scale tests should not be trusted since for any algorithm or analysis setup it is easy to find a few datasets where this algorithm seems to outperform other methods.**

A few Key Findings

II. Important Aspects Often Overlooked

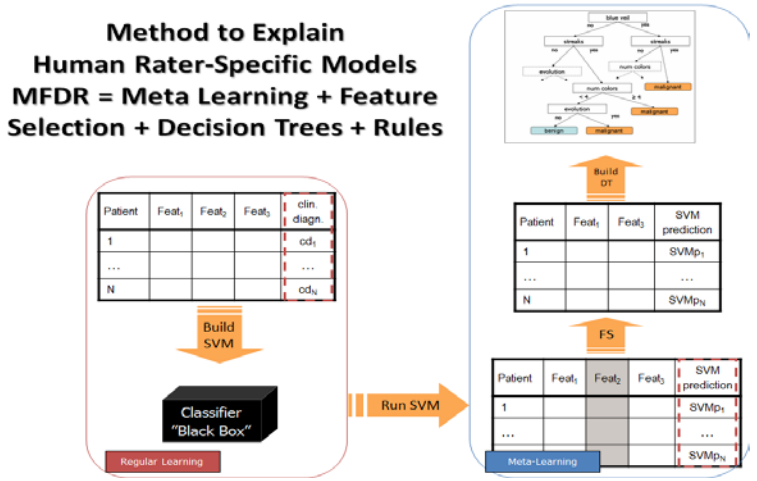
- **Data Design**: how to best (fastest, cheapest) collect data?

Design	Training sample	Manage True HOT, false HOT, Missed HOT	Accuracy & Cost	Preconditions
Human	No sample; applied on all data	Not managed	Very low accuracy, very high cost	Large numbers of human reviewers can be trained for task in short time
Keywords	No sample; applied on all data	Not managed	Very low accuracy, medium-high cost	---
Keywords , then Predictive Coding or human	Various	Not managed	Very low accuracy, medium cost	Same as in component parts
Random Sampling	Random	Accurately managed	High accuracy, low cost	HOT documents not extremely rare
Case Control	Non-random	Not managed	High accuracy, low cost	HOT documents set identified independently
Seeded Iterative	Non-random	Not managed	Medium-high accuracy, medium-low cost	HOT documents set identified independently
HOT-Augmented Random Sampling	Non-random	Not managed	High accuracy, low cost	HOT documents not extremely rare + HOT documents set identified independently

Ideal (if feasible)

Second best if Random Sampling not feasible

- **Defend the results and the process**.

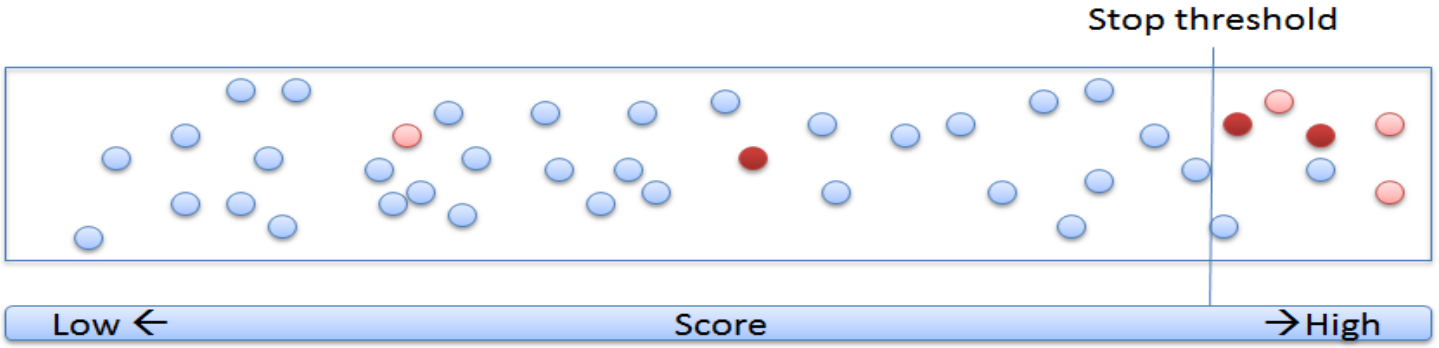


A few Key Findings
II. Important Aspects Often Overlooked

- How to **manage risks for false positives and false negatives** when deciding to stop reviewing documents in the ranked list?

If a classifier is built well, is calibrated, validated, and we know the prevalence of HOT in the population something very important happens:

- We can calculate that if we accept the first *k* documents then
  - we will have found a specific number of HOT documents
  - we will have a specific number of false positives
  - we will have missed a specific number of HOT documents
- Thus, we can decide where to stop and thus manage our effort against missing an acceptably small number of HOT documents.



Predictive Coding for Discovery Example Case Studies

- F*** (M***) lawsuit.

  *Identification of HOT cases incriminating investment firm as negligent in due diligence for M** firm investments.*

- D** C** vs. M** L**.

  *The analysis identified documents that indicated whether M** was aware of the state of the auction rate securities (ARS) market and whether M** misrepresented its understanding of the risk and liquidity of the market.*

  Notably, achieved 0.99 AUC in HOT document classification.

- J*** vs. N***.

  Undisclosed task. Client only provided labeled documents

- B*** S***.

  Multiple PC categories for litigation preparedness.

- A*** E*** vs Affiliates.

  *Class action lawsuit for fee discrimination. A*** wishing to produce evidence that they did not purposely manipulate their charges to businesses). Notably we created custom data structures and database to enable PC with the A*** CRM software.*

Positive and negative examples

From:
Sent:Thu 7/29/2010
To:
Cc:
Bcc:
Subject:Newedge - Large Trader Reporting

Gentlemen,

I received a call from Josh Stahl of Newedge. He and two of his colleagues, Kevin Zuber and Mike Dempsey, had questions regarding NFX Rule F-8 (documenting the OTC trade that's part of a SwapDrop) and the technical/connectivity requirements for reporting Large Trader Positions to NFA.

I was able to help them understand Rule F-8, but I wasn't as knowledgeable on the technical mechanics of the Large Trader Position reporting process. So this is notice that I gave them your names as initial contact persons at your respective organizations.

Given the general nature of their questions our organizations may want to consider adding both of these topics to an FAQ for new and prospective members.

Thanks,

[redacted]

+

From:
Sent:Fri 9/11/2009
To:
Cc:
Bcc:
Subject:RE: jeffries and co.

Thanks for this and I will reach out to Jason as you suggest

From:
Sent: Friday, September 11, 2009 2:18 PM
To: Winter, Steven; Lewis, Clifford M; Welch, Denise
Cc:
Subject: jeffries and co.

Hello

Our good friends at Jeffries would like to directly | discuss with you their desire/need for an FCM in cleared IRS.

Please feel free to reach out to| r (copied here below h| now running the desk at Jeffries- and like many of the well capitalized BDs, Jeffries are looking to expand their reach back into their old stomping grounds

No more fertile soil | than thru a clearing member in IRS.

More of these types of names to follow and please let us know if there is someone else in your team we need to| have copied on emails for new clients?

Best

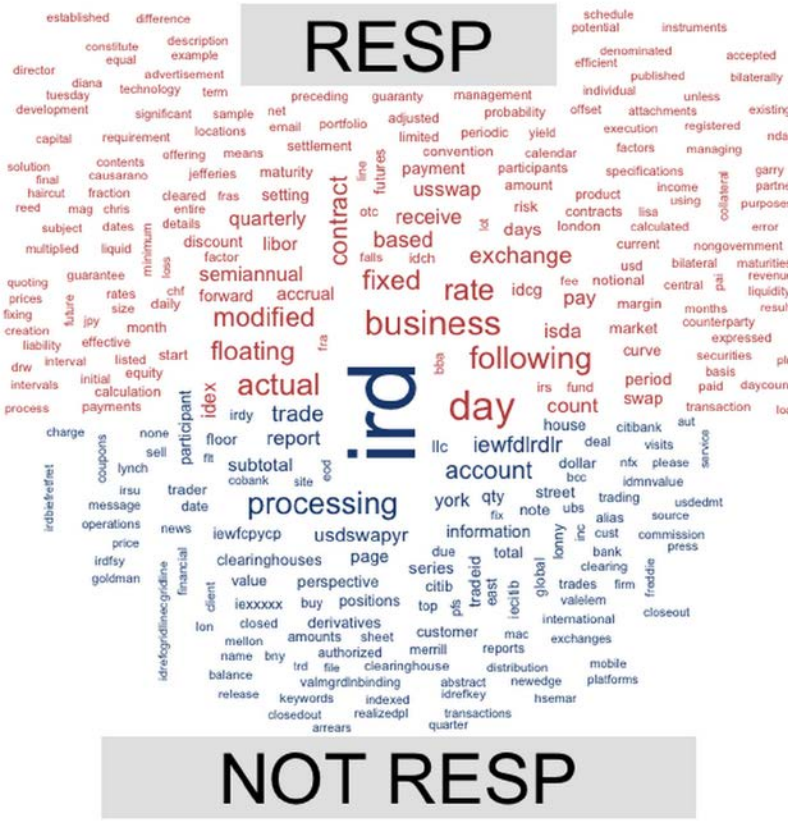
[redacted]

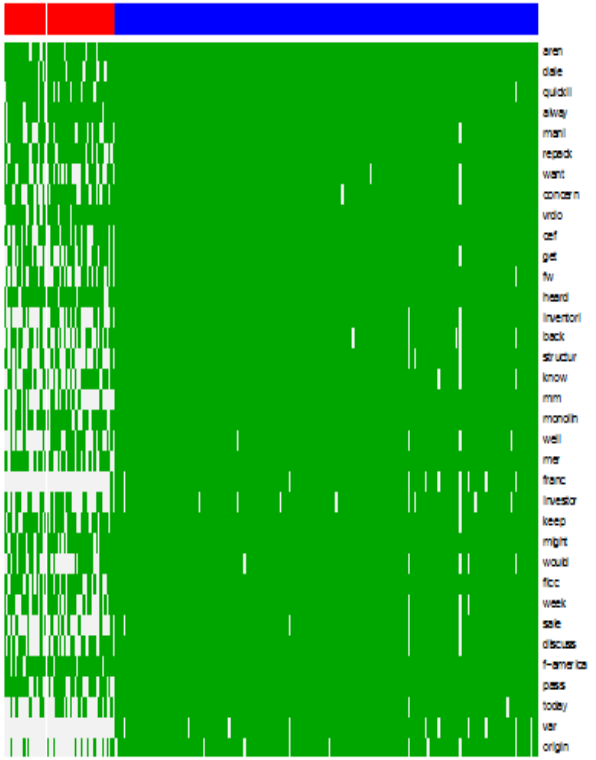
-

Using feature lists for model explanation

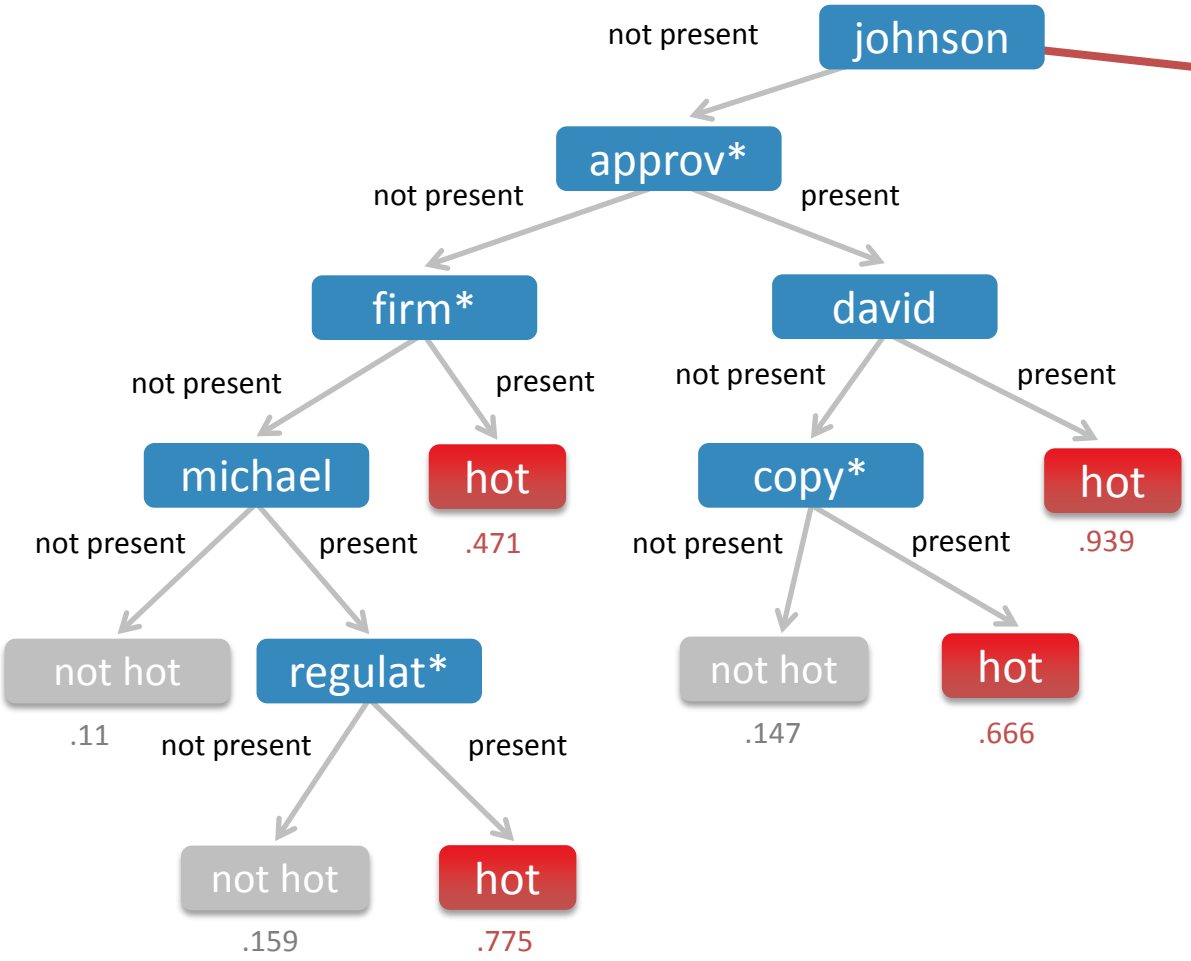
Feature	AUC	Frequency of selection during cross-validation
idcg	0.66	1
current	0.62	1
forward	0.616	1
need	0.612	1
accept	0.609	1
float	0.599	1
jefferi	0.563	1
drw	0.548	1
report	0.373	1
use	0.641	0.98
re	0.617	0.98
portfolio	0.597	0.98
discount	0.568	0.98
bilater	0.555	0.98
affirm	0.545	0.98
fix	0.62	0.94
construct	0.532	0.94
pay	0.578	0.92
par	0.547	0.92
interest	0.631	0.9
counterparti	0.587	0.9
aris	0.571	0.9
factor	0.569	0.9
spread	0.554	0.9
o	0.631	0.88
rate	0.626	0.88
basi	0.598	0.88
exposur	0.561	0.88
pai	0.554	0.88
tighter	0.54	0.88
contract	0.629	0.86
start	0.606	0.86
real	0.547	0.86
limit	0.59	0.84
interv	0.574	0.84
abil	0.554	0.84

Explaining coding using word clouds & heat maps





Using decision trees for model explanation



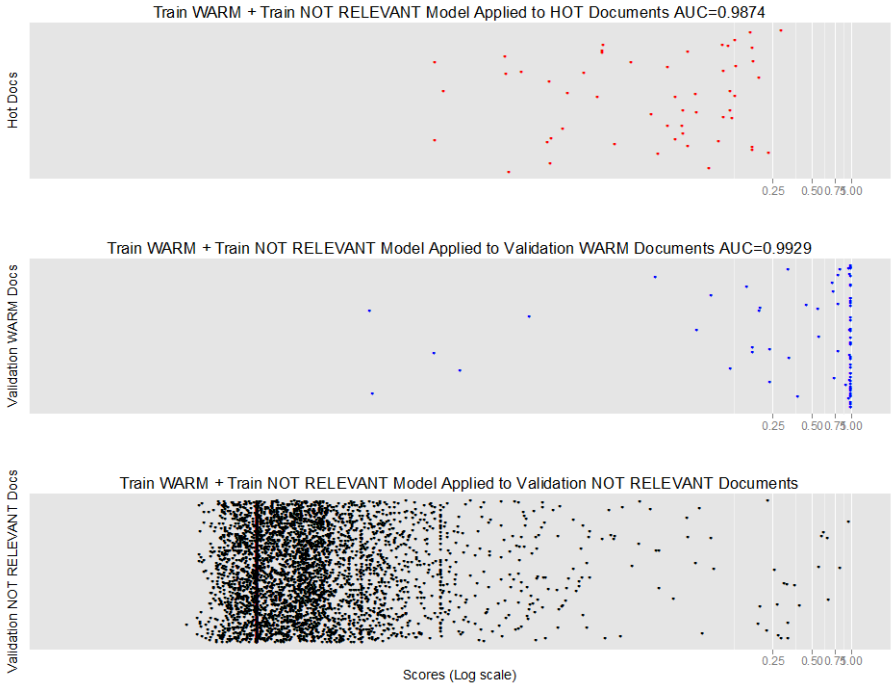
	johnson	Imag*
Document	Yes	Yes

If a document contains “johnson” and “imag*”, then there is a high likelihood of it being a hot document (.888).

Managing misclassification risks when using the model results

Threshold	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	# of Predicted Positives in the Application Corpus	# of Predicted Negatives in the Application Corpus
0.01	0.984	0.222	0.110	0.997	81411	15763
0.02	0.914	0.550	0.162	0.987	46093	51081
0.03	0.856	0.647	0.188	0.980	29263	67911
0.04	0.813	0.712	0.211	0.977	19565	77609
0.05	0.771	0.752	0.229	0.973	13998	83176
0.06	0.733	0.787	0.247	0.970	10486	86688
0.07	0.703	0.813	0.264	0.967	8442	88732
0.08	0.677	0.838	0.285	0.966	7014	90160
0.09	0.642	0.856	0.299	0.963	6165	91009
0.1	0.617	0.870	0.310	0.961	5402	91772
0.11	0.589	0.882	0.323	0.958	4819	92355
0.12	0.564	0.893	0.334	0.956	4282	92892
0.13	0.548	0.903	0.352	0.955	3863	93311
0.14	0.536	0.911	0.368	0.955	3516	93658
0.15	0.518	0.917	0.375	0.953	3262	93912
0.16	0.501	0.922	0.383	0.952	3077	94097
0.17	0.495	0.928	0.396	0.951	2852	94322
0.18	0.482	0.931	0.400	0.950	2676	94498
0.19	0.468	0.934	0.406	0.949	2572	94602
0.2	0.449	0.937	0.407	0.948	2451	94723
0.21	0.442	0.940	0.416	0.947	2353	94821
0.22	0.432	0.944	0.427	0.947	2263	94911
0.23	0.428	0.947	0.439	0.946	1976	95198
0.24	0.420	0.950	0.450	0.946	1913	95261
0.25	0.413	0.953	0.458	0.945	1774	95400
0.26	0.400	0.955	0.460	0.944	1727	95447
0.27	0.394	0.958	0.468	0.944	1525	95649
0.28	0.387	0.960	0.476	0.943	1413	95761
0.29	0.380	0.962	0.487	0.943	1346	95828
0.3	0.375	0.965	0.502	0.943	1328	95846
0.31	0.367	0.967	0.516	0.942	1287	95887
0.32	0.360	0.968	0.520	0.942	1250	95924
0.33	0.353	0.970	0.532	0.941	1035	96139
0.34	0.349	0.972	0.542	0.941	970	96204
0.35	0.346	0.973	0.554	0.941	946	96228
0.36	0.341	0.974	0.561	0.940	910	96264

Examining consistency of experts' labeling by cross-application of models

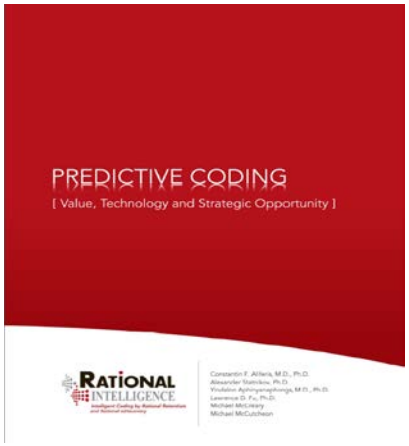


Conclusions

- PC can be used as an efficiency booster or as a transformative technology.
- It can address a variety of client needs including cost reduction, production speed accelerator, profit margin improvement, market share increase, and product de-risking.
- The technology can also be used for fraud detection, insurance risk modeling, and numerous other applications in legal and other domains.

Key References

CF. Aliferis et al. **Predictive Coding: Value, Technology and Strategic Opportunity**, Rational Intelligence 2013.



A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization
Aphinyanaphongs, Yindalon; Fu, Lawrence D; Li, Zhiguo; Peskin, Eric R; Efstathiadis, Efstratios; Aliferis, Constantin F; Statnikov, Alexander
2014 OCT;65(10):1964-1987, Journal of the Association for Information Science & Technology id: 1313832, year: 2014, vol: 65, page: 1964

APPLICATION/PROVING GROUND #3: HEALTHCARE OPERATIONAL MODELING

Value Generation Map

Quality

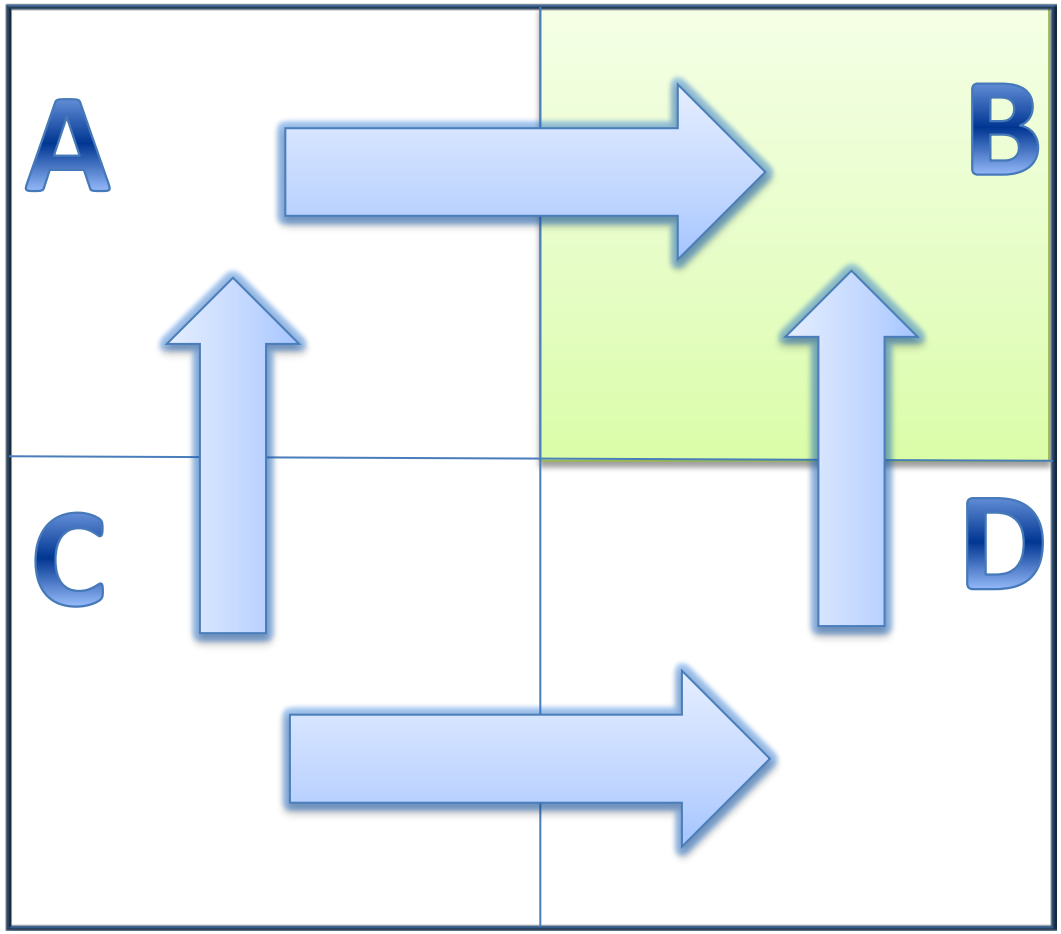
,

Safety,

Risk

Managem

ent



Profitability:

Market Share, Cost containment

Insights about the R&D process

Insights about the R&D process

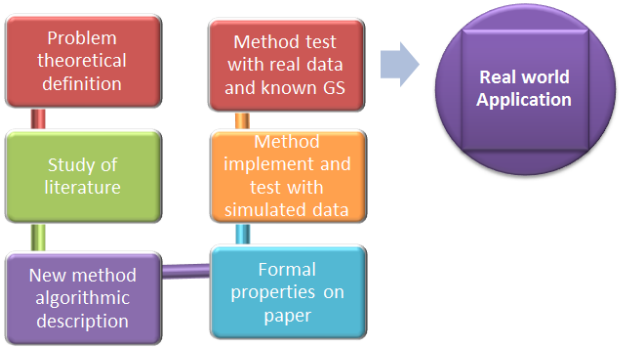
1. Building upon a firm theoretical foundation

Insights about the R&D process

Evidence-based algorithm development

In Medicine there is a hierarchy of Evidence

Similarly for Analytics



Insights about the R&D process

2. Keeping it real: is the new method motivated by a real problem without a solution? Or by a real weakness in pre-existing methods?

How to tell?

Benchmarking

- Thorough
- Realistic
- Unbiased

Insights about the R&D process

More on benchmarking: does the new method/comparator methods really work? When?

a. Extensive testing (datasets, sample sizes, noise, mv etc)
b. Try to systematically make the algorithm “break”
c. Respect authors’ setups/protocols
d. Show all parameterizations
e. Overall robustness
f. Even very “naïve” algorithms will often have their sweet spot

Insights about the R&D process

3. Keeping it real: does new method/comparators fit real life workflows?

a. Sometimes it will help rather than hinder.

E.g., - directionality vs edge discovery;

- allowing acceptable error

b. Other times, it makes things harder:

E.g., Manipulations' specificity

Insights about the R&D process

4. Because it may look like it will not (or should not) work it does not mean it won't! Examples:
    a. The problem of multiple hypothesis testing
    b. PC skeleton phase vs MMHC skeleton phase
    c. Learning with epistasis
    d. The power of edge detection
    e. LCN approximating MB
    f. Connectivity/shielding effects
    g. Real life sparseness etc. etc.

Insights about the R&D process

5. We may assume that finding the right parameter value will be easy/not overfit; this is not always the case.

6. Combining techniques even from entirely different families occasionally works wonders. E.g.:

a. CIT based skeleton with Bayesian orientation and repair.
b. Fitting all sorts of classifiers on MB variable sets
c. Plugging all kinds of CIT inside CIT-based algorithms

Insights about the R&D process

7. Pay attention to legitimate problems of preexisting work. E.g. SPC vs MMHC

8. Go deep into the details of prior work. E.g., Aracne experiments, K-S, GS, univariate associations, etc.

Insights about the R&D process

9. Reuse as much as possible and create an interlocking system of modules as much as possible. ➔ More useful, coherent, robust

10. Progressively fix limitations in successive generations of algorithms ➔ DAQ the R&D

…But know what constitutes a minimal advance vs a an important advance (incremental or not). My advice: do not bother too much with minor steps.

Insights about the R&D process

11. There is great value in establishing general properties (not just algorithmic ones). E.g. GLL says something about a very large number of possible algorithms and discourages frivolous modifications while it points to potentially serious opportunities for improvements.

12. Play to your strengths and respect your weaknesses. E.g.: my working with CIT framework instead of Bayesian.

13. Create a team science environment that all ideas (from the group and outside) can be challenged from within the group and outside. Practice “creative disbelief”. Prevent groupthink.

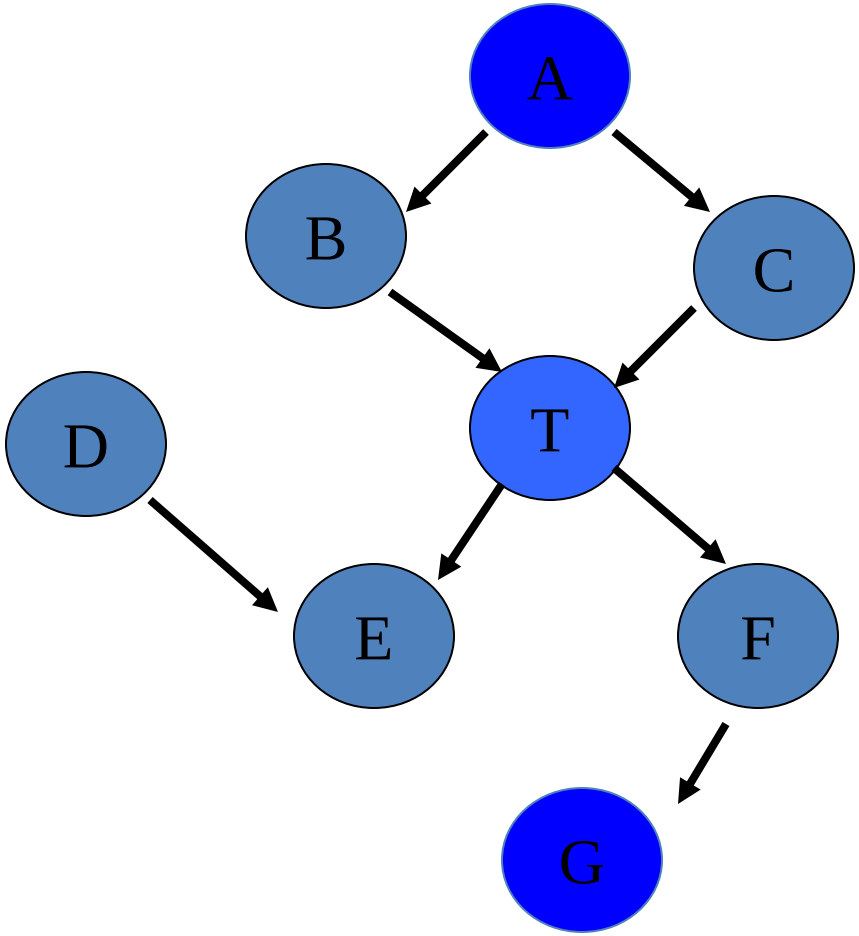
Discussion

A Pictorial presentation of HITON-MB (barring speed-up optimizations)

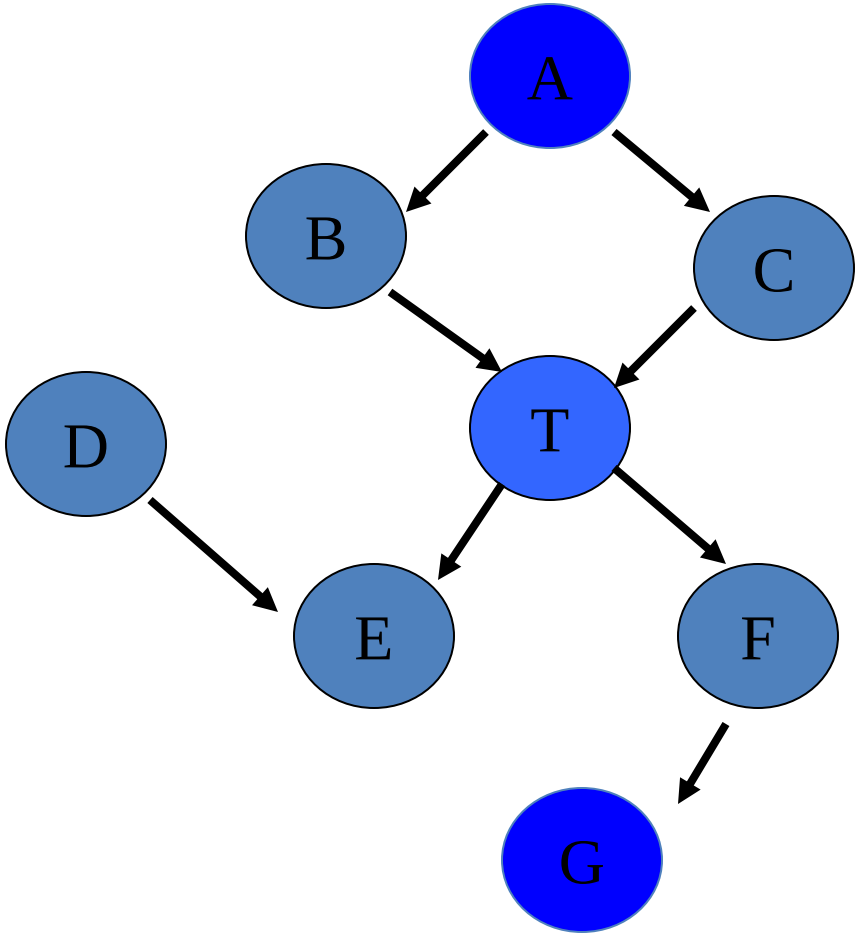
Example Trace of HITON:

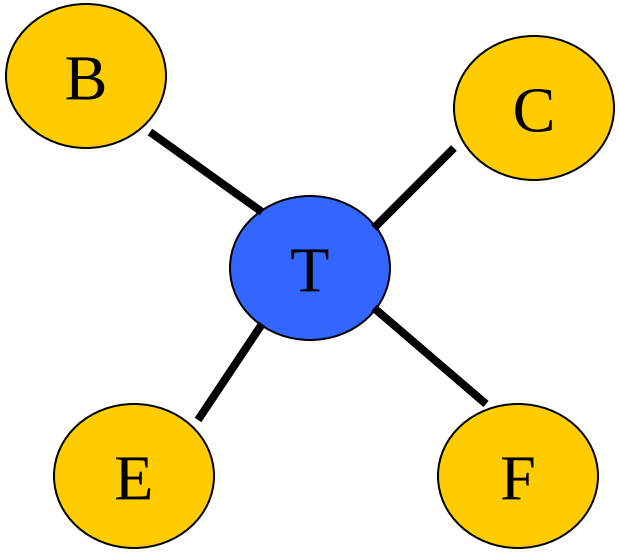
True structure depicted; members of the Markov Blanket of T are cyan

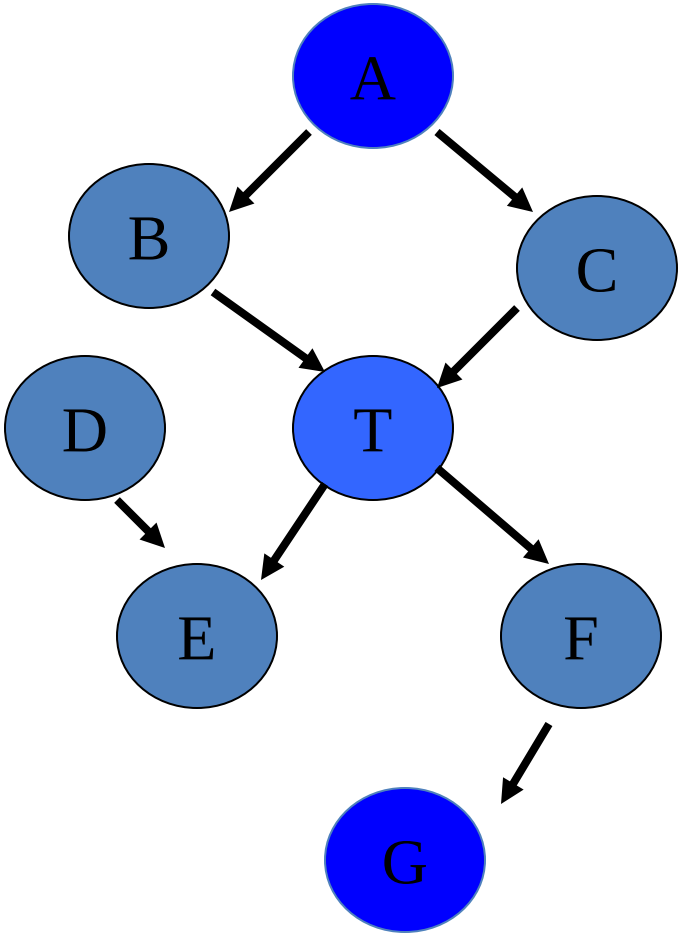
We have access to training data but not the true structure



1. Identify variables with direct edges to the target T







Tentative PC:

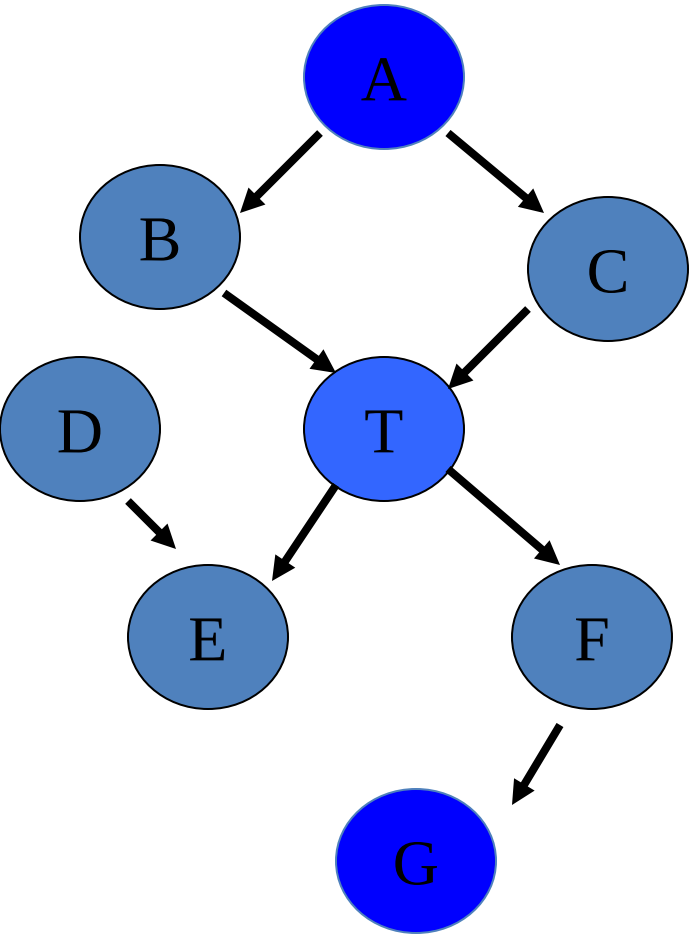
Iteration 1 A

Iteration 2 A B

Iteration 3 A B C

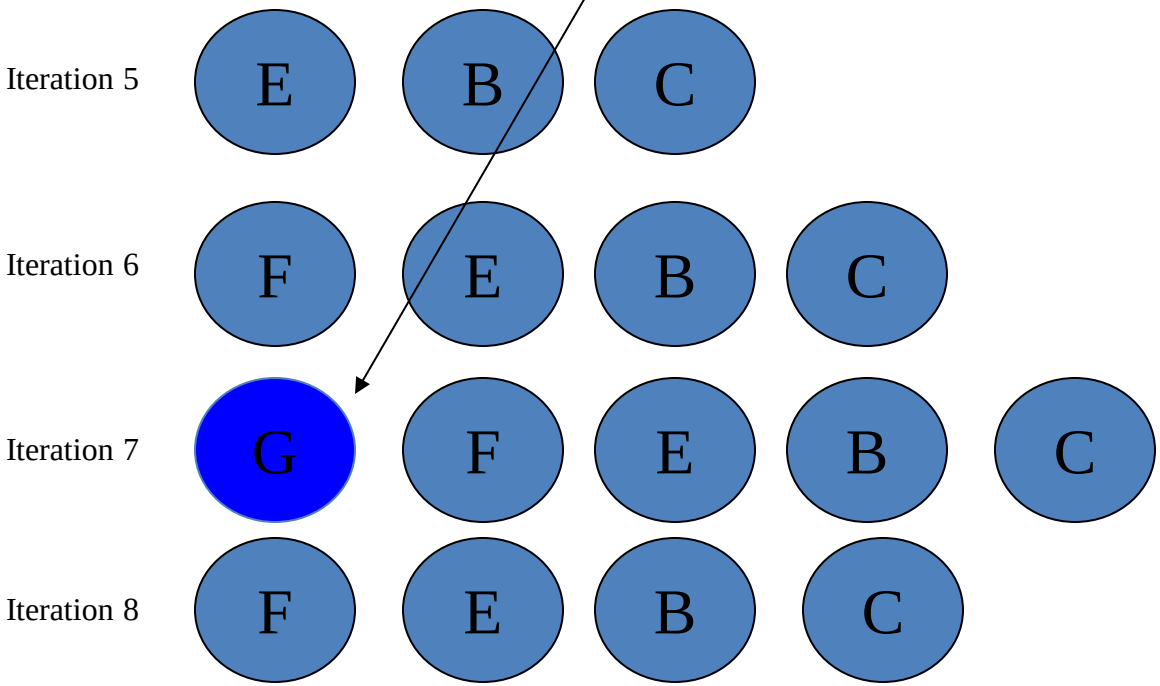
Iteration 4 B C

A is removed because
$\perp(A : T \mid B, C)$

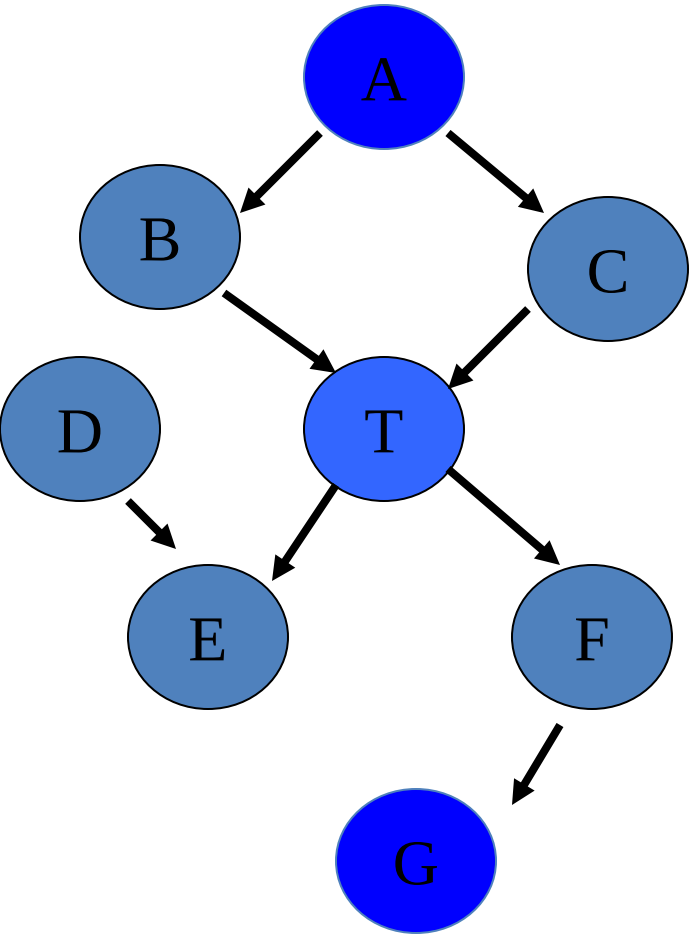


Tentative PC (continued):

G is removed because
$\perp(G : T \mid F)$



Algorithm terminates because there are not other variables left to consider.



Symmetry:

When running the previous procedure for B returns: A, T.

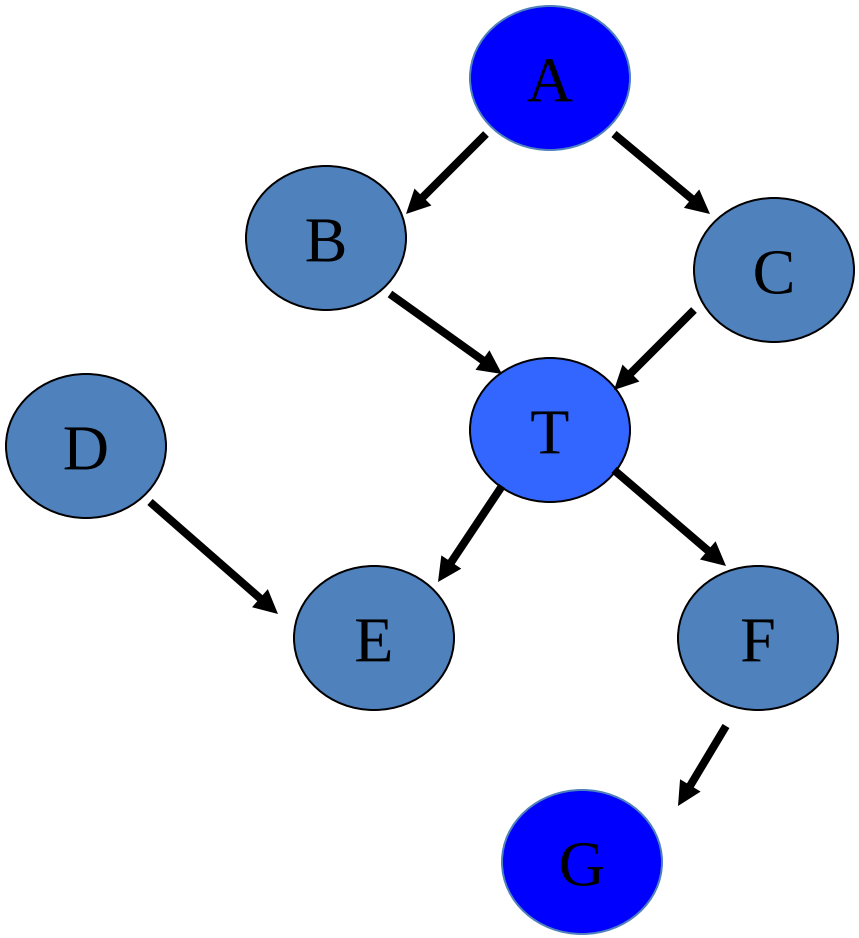
When running the previous procedure for C returns: A, T

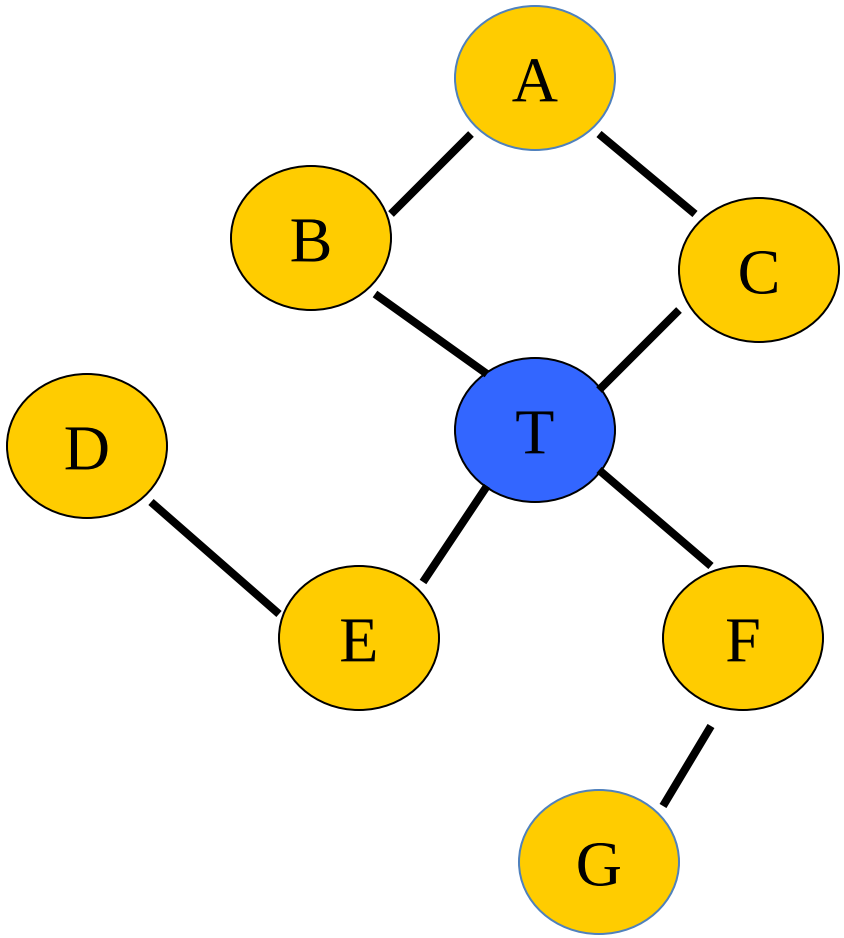
When running the previous procedure for E returns: D, T.

When running the previous procedure for F returns: G, T.

Hence all B,C,E,F satisfy symmetry and are retained.

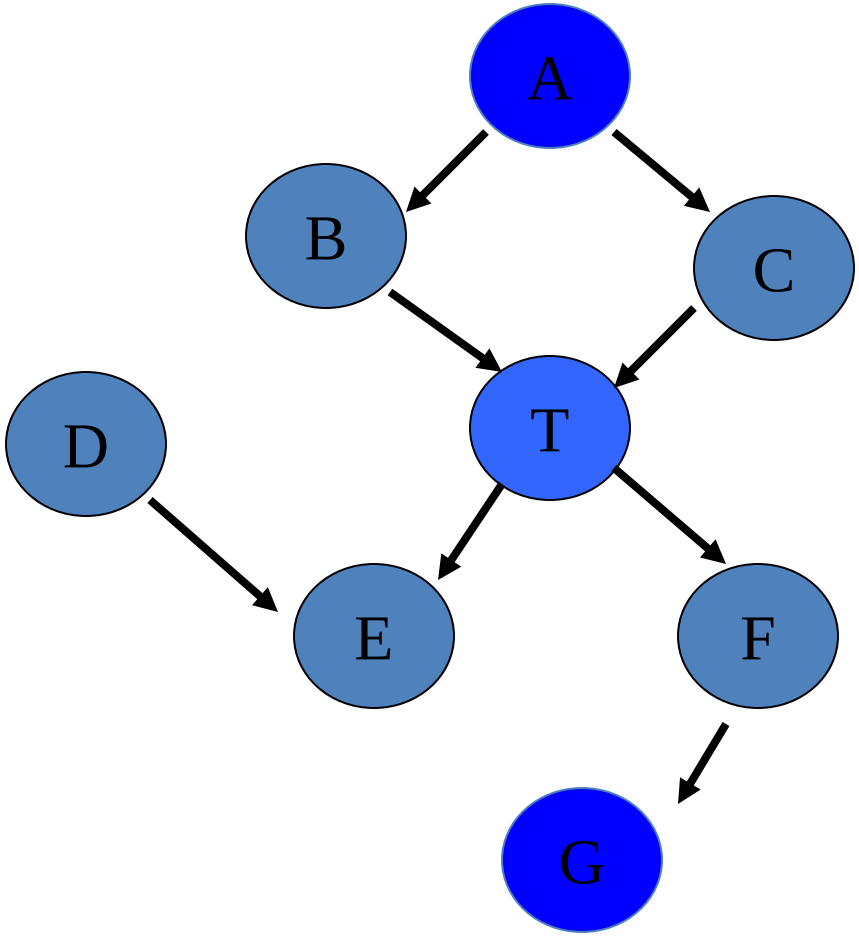
2. **Repeat previous for all members of PC and take the union of the resulting variables to be U.**

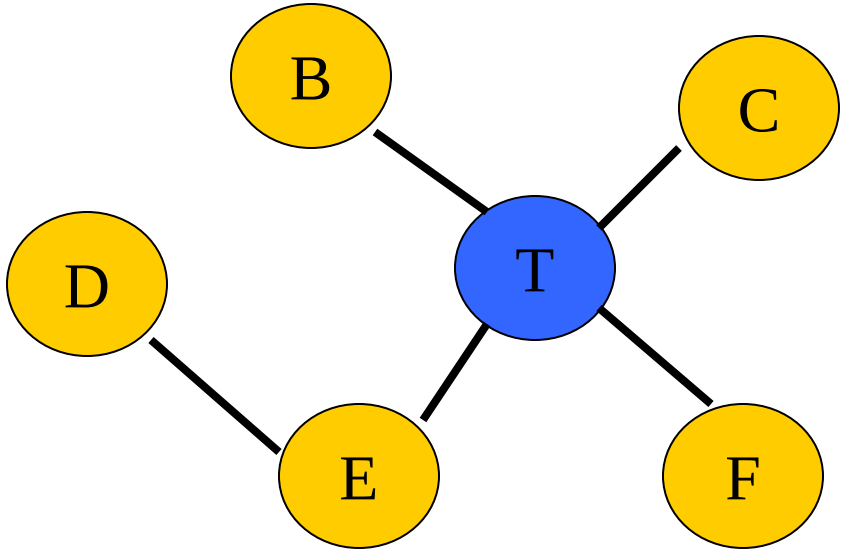




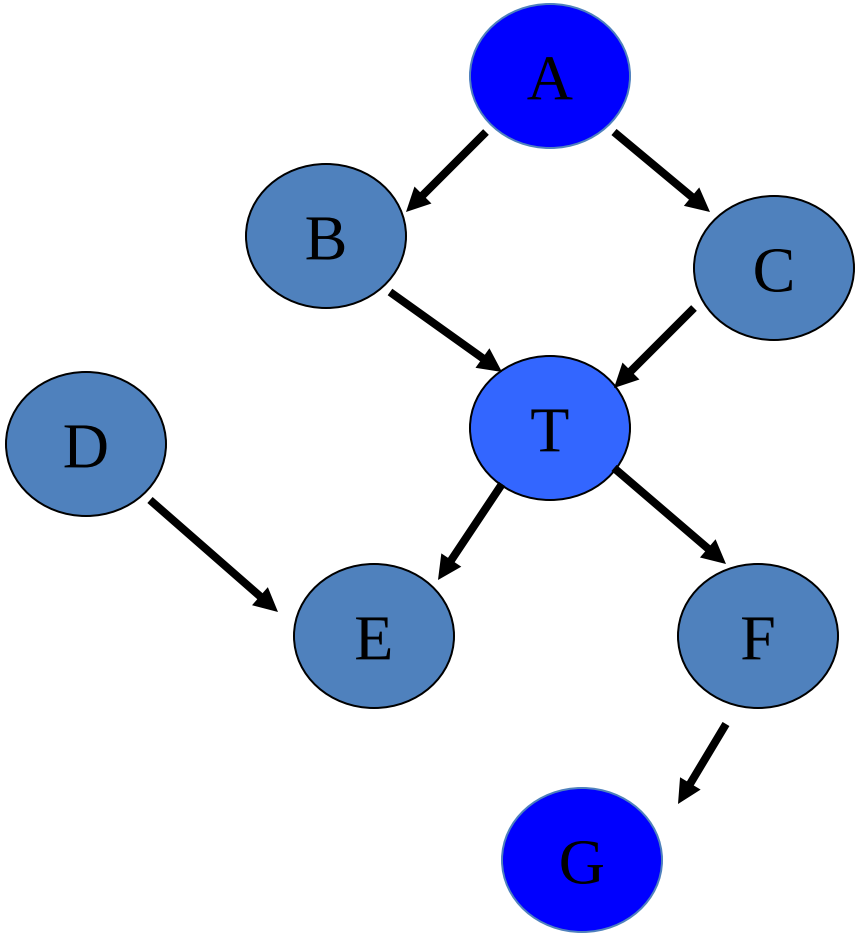
3. Throw away non-members of the Markov Blanket.

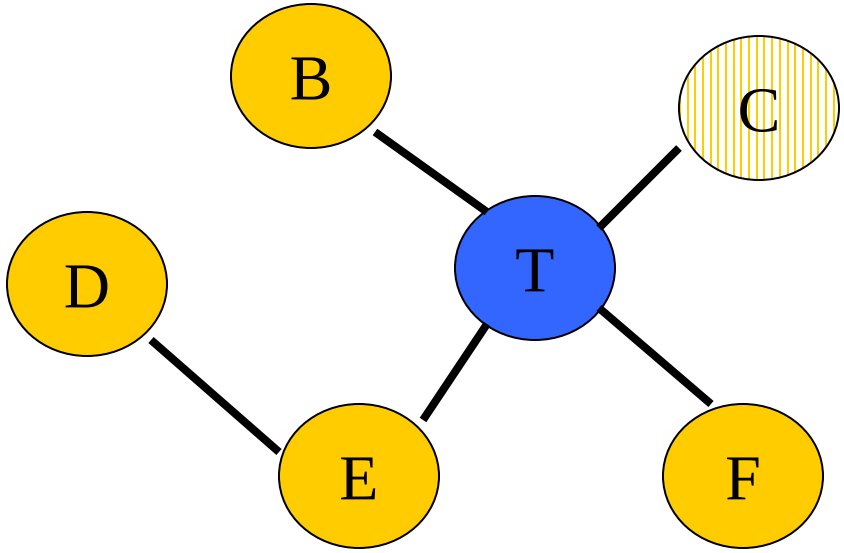
A member *X* of PCPC that is not in PC is a member of the Markov Blanket if there is some member of PC *Y*, such that *X* becomes conditionally dependent with T conditioned on any subset of the remaining variables and *Y* .





4. **If we desire to use the Markov Blanket for classification, eliminate any unnecessary variables by using a wrapping approach and cross-validation.**





Generalized Learning Frameworks (GLL, LGL)

GLL-PC: Generalized Local Learning -Parents and Children

1. Start with empty set *S* of candidates for the true PC set.
2. Inclusion heuristic function: prioritizes variables for inclusion in S and throws away non-eligible variables
3. Elimination strategy: removes variables from inside candidate set *S*
4. Interleaving strategy: combines #2, and #3 until an exit termination criterion met
5. Symmetry requirement: Eliminate from *S after #4* every variable V such that when steps #1-4 are run again with V as the target, T is not in the candidate set after step #4.
6. Report the candidate set *S*

Steps #2,3,4 can be instantiated in infinite ways.

There are rules that determine the admissible instantiations (which are themselves infinite)

GLL-PC: Admissibility rules

1. Start with empty set of candidates.
2. Inclusion heuristic function: Rank variables for priority for inclusion in the candidate set and include the highest-ranked variable(s) according to ANY heuristic ranking function that respects the following requirement:

   > *All variables that have a direct edge to/from the response variable, are eligible for inclusion in the candidate set and each one is assigned a non-zero value by the ranking function. Variables with zero values are discarded and never considered again.*

   Variables may be re-ranked after each update of the candidate set, or the original ranking may be used throughout the algorithm's operation.
3. Elimination strategy: If any variable (inside or outside the candidate set) becomes independent of the response variable given any subset of the candidate set, then discard that variable and never consider it again (whether it is inside or outside the candidate set). Part of the strategy is *prioritizing* the independence tests.
4. Interleaving strategy: Iterate inclusion and elimination ANY way you like provided that you stop iterating when no variable outside the candidate set is eligible for inclusion and when no variable in the candidate set can be removed.
5. Once iterating has stopped, filter the candidate set using symmetry criterion.
6. Output candidate set.

Respecting the admissibility rules of GLL-PC

- Obtain correct local causal neighborhood (direct causes and direct effects) under the following sufficient conditions:
  - Faithful distributions,
  - Correct statistical decisions about independence (affected by choice of test, power-size analysis, and sample size)
  - Local causal sufficiency (i.e., no confounders among direct causes/effects and the target).

HITON-PC as instance of GLL-PC

1. Start with empty set of candidates.
2. Inclusion heuristic function: Rank variables for priority for inclusion in the candidate set by univariate association. Discard variables with zero univariate association. Put in the candidate set the first variable.
3. Elimination strategy: If any variable inside the candidate set becomes independent of the response variable given any subset of the candidate set, then remove that variable from the candidate set and never consider it again.
4. Interleaving strategy: perform elimination every time the candidate PC set receives a new member.
5. Once iterating has stopped, filter the candidate set using symmetry criterion.
6. Output candidate set.

This we call: interleaved HITON-PC with symmetry correction and is a correct algorithm.

MMPC as instance of GLL-PC

1. Start with empty set of candidates.
2. Inclusion heuristic function: Rank each variable for priority for inclusion in the candidate set using the maximum of the minimum associations of the variable and the target (minimizing over all conditioning subsets of current candidate members of PC). Discard variables with zero max-min association with target. Put in the candidate set the first variable.
3. Elimination strategy: If any variable inside the candidate set becomes independent of the response variable given any subset of the candidate set, then remove that variable from the candidate set and never consider it again.
4. Interleaving strategy: Perform elimination only once (when the tentative PC cannot grow any more).
5. Once iterating has stopped, filter the candidate set using symmetry criterion.
6. Output candidate set.

This we call: MMPC with symmetry correction and is a correct algorithm.

GLL-MB: Generalized Local Learning –Markov Blanket

1. Start with empty set *M* of candidates for the true MB set.
2. Find the PC(T) using GLL-PC.
3. Find the PC(X) for every member of PC(T). Create the union U=Union (PC(Xi)).
4. Eliminate non-spouses from U using the SGS criterion.
5. Eliminate non-predictive members of U using a wrapper approach.

Steps #2,5 can be instantiated in infinite ways.

Admissibility requirements: use an admissible GLL-PC and a sufficiently powerful wrapper.

Respecting the admissibility rules of GLL-MB

- Obtain correct minimal Markov Blanket (variable set that renders all other variables independent of T given the MB) under the following sufficient conditions :
  - Faithful distributions,
  - Correct statistical decisions about independence (affected by choice of test, power-size analysis, and sample size).

HITON-MB as instance of GLL-MB

1. Start with empty set *M* of candidates for the true MB set.
2. Find the PC(T) using HITON-PC with symmetry correction (or without).
3. Find the PC(X) for every member of PC(T). Create the union U=Union (PC(Xi)).
4. Eliminate non-spouses from U using the SGS criterion.
5. Eliminate non-predictive members of U using a backward elimination wrapper and the desired classifier and loss function.

This we call: interleaved HITON-MB with (or without) symmetry correction and is a correct algorithm.

LGL: Locally-constrained Global Learning

1. Find PC(X) for all variables X in data using an admissible instantiation of GLL-PC.
2. Piece together the undirected skeleton.
3. Use any desired arc orientation scheme to orient edges.

#1,3 can be instantiated in infinite ways. If an admissible GLL-PC is used in #1, and admissible orientation scheme in #3, then the total algorithm is admissible.

Respecting the admissibility rules of LGL

- Obtain correct causal graph under the following sufficient conditions :
  - Faithful distributions,
  - Correct statistical decisions about independence (affected by choice of test, power-size analysis, and sample size); alternatively correct statistical decisions about graph structure scoring.
  - Causal sufficiency (i.e., no confounders between any pair of variables).

MMHC: instance of LGL

1. Find PC(X) for all variables X in data using MMPC.
2. Piece together the undirected skeleton.
3. Use greedy TABU search and BDeu to orient edges.

MMHC is admissible with respect to the skeleton but inadmissible with respect to orientation.