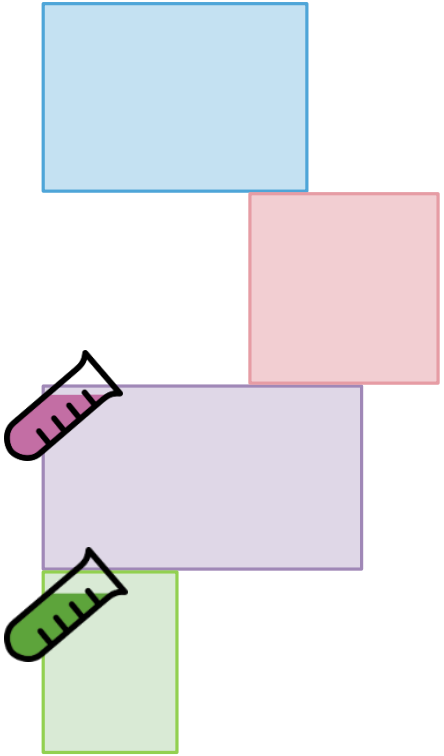
LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY

Ioannis Tsamardinos
Computer Science Department, University of Crete
June 2016

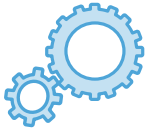
HETEROGENEOUS DATA SETS MEASURING THE SAME SYSTEM UNDER STUDY

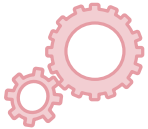
Study \ Variables	Thrombosis 	Contraceptives 	Protein C 	Breast Cancer 	Protein Y 	Protein Z 
1 observational data	Yes	No	10.5	Yes	-	-
	No	Yes	5.3	No	-	-
					-	-
	No	Yes	0.01	No	-	-
2 observational data	-	-	-	Yes	0.03	9.3
	-	-	-			
	-	-	-	No	3.4	22.2
3 experimental data	No	No	0 (Control)	No	3.4	-
	Yes	No	0 (Control)	Yes	2.2	-
					-	-
	Yes	Yes	5.0 (Treat.)	Yes	7.1	-
	No	Yes	5.0 (Treat.)	No	8.9	-
4 experimental data	No	No (Ctrl)	-	-	-	-
	No	No (Ctrl)	-	-	-	-
			-	-	-	-
	Yes	Yes(Treat)	-	-	-	-

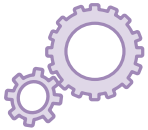
ISOLATED ANALYSIS

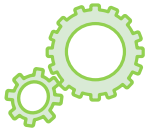


Analyze data









Publish results

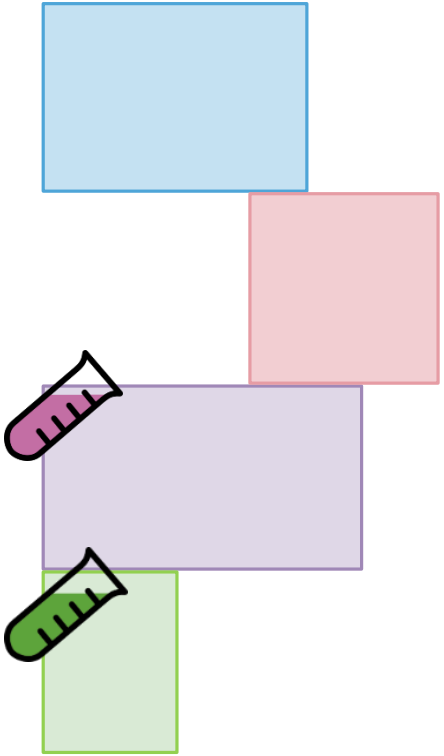
"…The use of contraceptives is correlated with Thrombosis, negatively correlated with Breast Cancer and levels of Protein E …"

"…Protein E is a risk factor for Breast Cancer…"

"…Drugs reducing protein C reduced the probability of Breast Cancer and lowered the levels of Protein E…"

"… In the randomized control trial, women taking contraceptives had 30% more chances of being diagnosed with thrombosis … "

INTEGRATIVE CAUSAL ANALYSIS

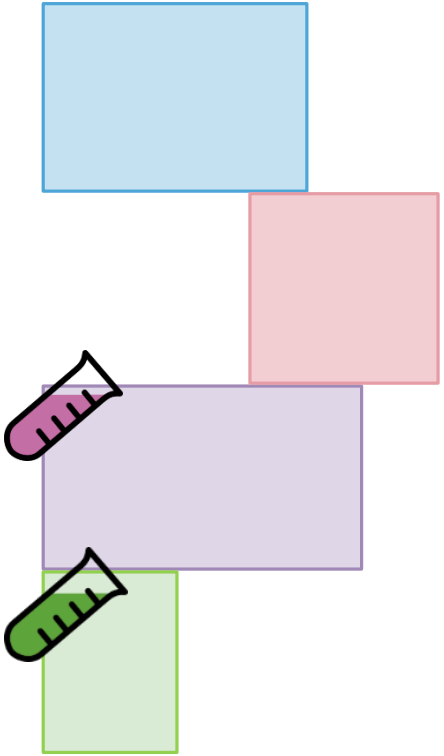


Data can not be pooled together:

Missing variables cannot be treated as missing values.

They come from **different experimental/sampling conditions (different distributions).**

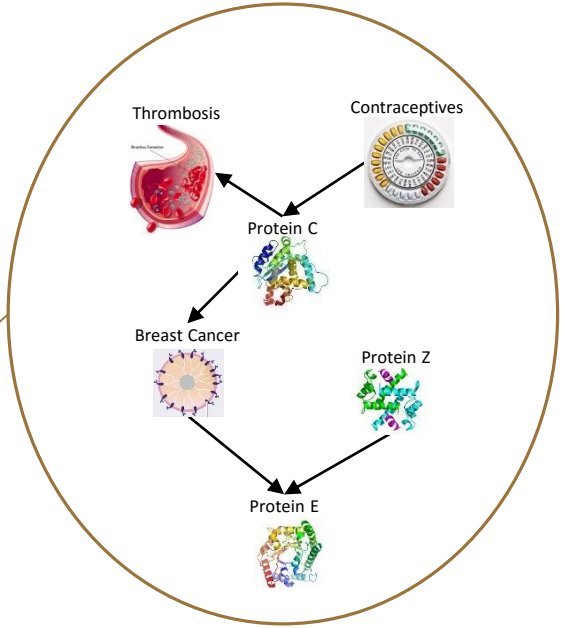
INTEGRATIVE CAUSAL ANALYSIS



Data can not be pooled together:

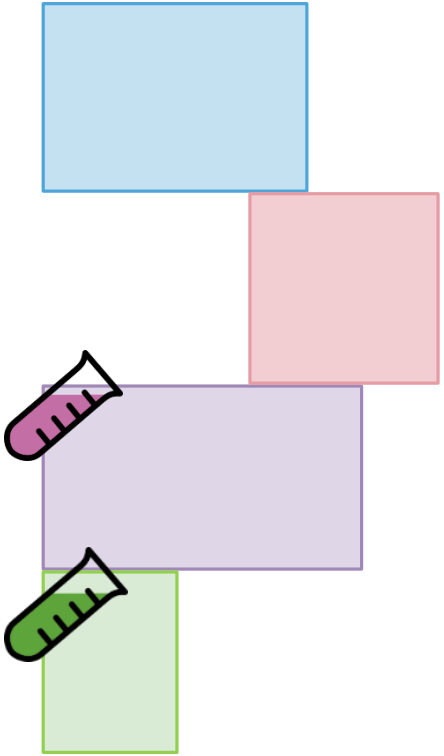
Missing variables cannot be treated as missing values.

They come from **different experimental/sampling conditions (different distributions).**

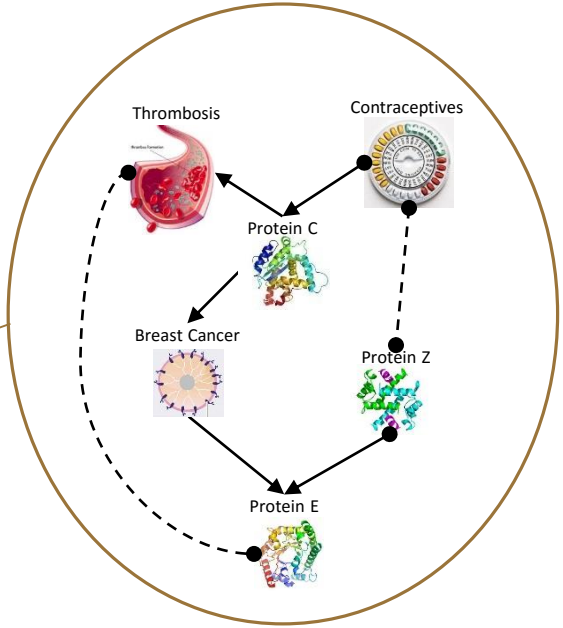


Data come from the **same causal mechanism**.

INTEGRATIVE CAUSAL ANALYSIS

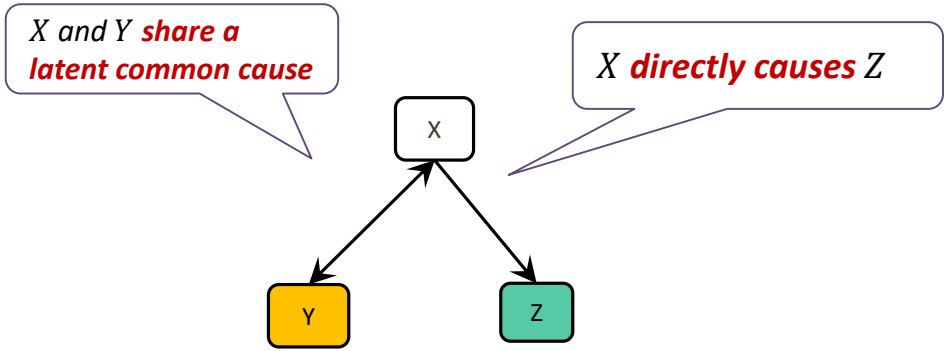


Identify the **causal graphs** that **simultaneously fit** all data.



SEMI MARKOV CAUSAL GRAPHS

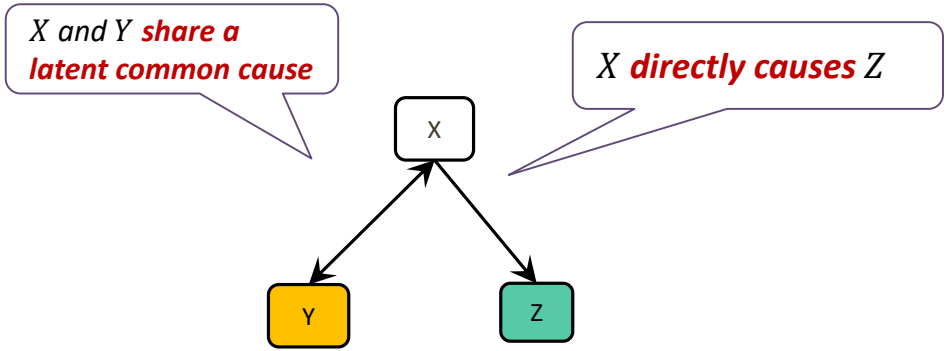
Semi Markov Causal Graph $\mathcal{G}$



- Directed edges represent direct causal relationships.
- Bi-directed edges represent confounding (**latent confounders)**.
- Both types of edges allowed for a single pair of variables.
- No directed cycles (no causal feedback).

SEMI MARKOV CAUSAL GRAPHS

Semi Markov Causal Graph $\mathcal{G}$



- Directed edges represent direct causal relationships.
- Bi-directed edges represent confounding (**latent confounders)**.
- Both types of edges allowed for a single pair of variables.
- No directed cycles (no causal feedback).

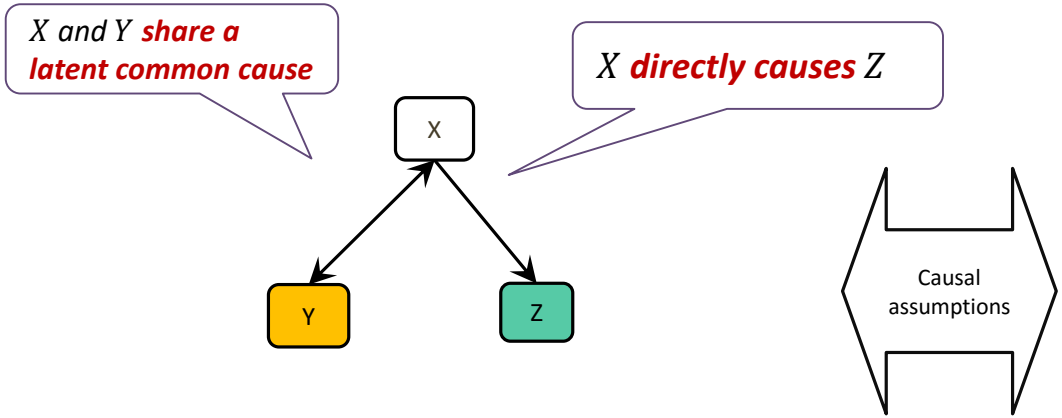
Joint Probability Distribution $\mathcal{P}$

		Z	
X	Y	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

- Joint probability distribution entails conditional (in) dependencies.
- $Ind(X, Y|\boldsymbol{Z}): P(X|Y, \boldsymbol{Z}) = P(X|\boldsymbol{Z})$
- $Dep(X, Y|\boldsymbol{Z}): P(X|Y, \boldsymbol{Z}) \neq P(X|\boldsymbol{Z})$

SEMI MARKOV CAUSAL GRAPHS

Semi Markov Causal Graph $\mathcal{G}$



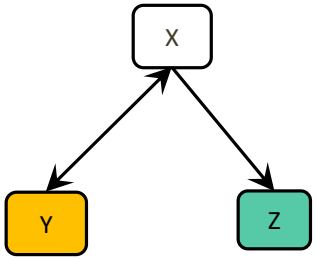
Joint Probability Distribution $\mathcal{P}$

		Z	
X	Y	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

- Directed edges represent direct causal relationships.
- Bi-directed edges represent confounding (**latent confounders)**.
- Both types of edges allowed for a single pair of variables.
- No directed cycles (no causal feedback).

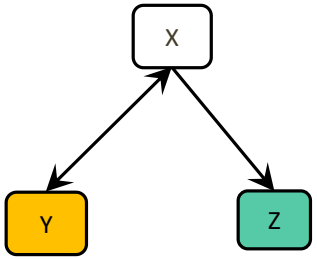
- Joint probability distribution entails conditional (in) dependencies.
- $Ind(X, Y|\boldsymbol{Z}): P(X|Y, \boldsymbol{Z}) = P(X|\boldsymbol{Z})$
- $Dep(X, Y|\boldsymbol{Z}): P(X|Y, \boldsymbol{Z}) \neq P(X|\boldsymbol{Z})$

CAUSAL ASSUMPTIONS



Causal Markov Assumption:
Every variable is independent of its **non-effects** given its **direct causes**.

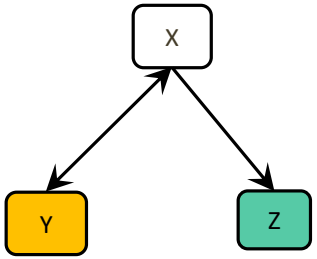
CAUSAL ASSUMPTIONS



$Ind(Y, Z \mid X)$

Causal Markov Assumption:
Every variable is independent of its **non-effects** given its **direct causes**.

CAUSAL ASSUMPTIONS

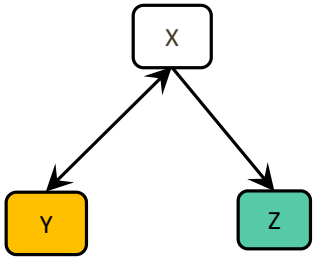


$Ind(Y, Z \mid X)$

Causal Markov Assumption:
Every variable is independent of its **non-effects** given its **direct causes**.

Causal Faithfulness Assumption:
Independences stem **only** from the causal structure, **not the parameterization** of the distribution.

CAUSAL ASSUMPTIONS



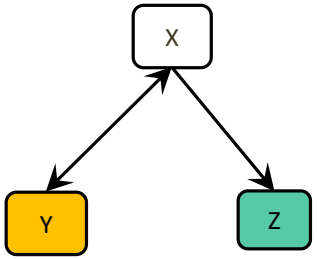
$Ind(Y, Z \mid X)$

$Dep(Y, Z \mid \emptyset)$
$Dep(X, Z \mid \emptyset)$
$Dep(X, Z \mid Y)$
$Dep(Y, X \mid \emptyset)$
$Dep(Y, X \mid Z)$

Causal Markov Assumption:
Every variable is independent of its **non-effects** given its **direct causes.**

Causal Faithfulness Assumption:
Independences stem **only** from the causal structure, **not the parameterization** of the distribution.

CAUSAL ASSUMPTIONS



$Ind(Y, Z \mid X)$

$Dep(Y, Z \mid \emptyset)$
$Dep(X, Z \mid \emptyset)$
$Dep(X, Z \mid Y)$
$Dep(Y, X \mid \emptyset)$
$Dep(Y, X \mid Z)$

Causal Markov Assumption:
Every variable is independent of its **non-effects** given its **direct causes.**

Causal Faithfulness Assumption:
Independences stem **only** from the causal structure, **not the parameterization** of the distribution.

All independencies in the joint probability distribution can be identified in $\mathcal{G}$ using the graphical criterion of **m-separation.**

m-SEPARATION

A path $X_1, \ldots, X_n$ between X_1 and X_n is **m-connecting given $\boldsymbol{V}$** if for every triple $\langle X_{i-1}, X_i, X_{i+1} \rangle$ on the path:

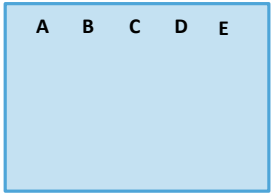
- If $X_{i-1} *\!\!\rightarrow X_i \leftarrow\!\!* X_{i+1}$ (colliding triplet),
  X_i or one of its descendants $\in \boldsymbol{V}$
- Otherwise, $X_i \notin \boldsymbol{V}$

m-connecting path **=>** information flow **=> dependence**

No **m-connecting** path **=>** no information flow **=> independence** (m-separation)

Colliders $X_{i-1} *\!\!\rightarrow X_i \leftarrow\!\!* X_{i+1}$ are **special** and create an asymmetry that will allow us to orient causal direction.

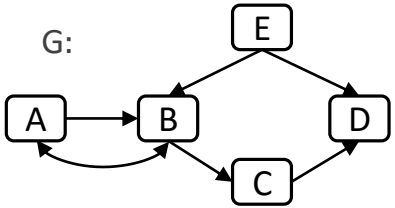
CAUSAL MODELLING



Data set D measuring a set of variables

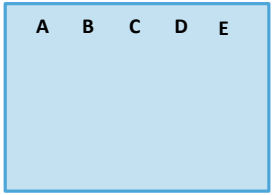
$A, B \mid E, C$	Ind
$A, B \mid \emptyset$	Dep
…	…
$E, C \mid A, B, C$	Dep

Conditional (in)dependencies (expected) in the joint probability distribution

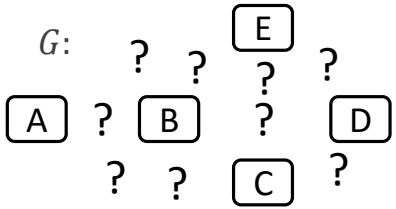


Paths (m-separations/connections) in the causal graph

REVERSE ENGINEERING

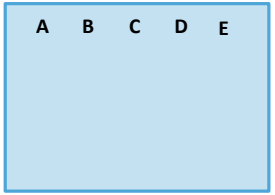


Data set D measuring a set of variables



causal graph?

REVERSE ENGINEERING

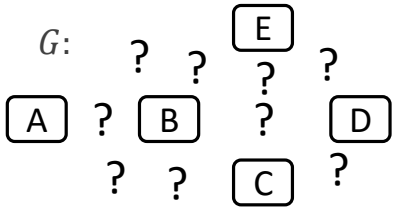


Data set D measuring a set of variables

$A, B \mid E, C$	Ind
$A, B \mid \emptyset$	Dep
…	…
$E, C \mid A, B, C$	Dep

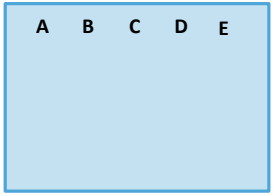
Find the (in)dependencies using statistical tests.

??



causal graph?

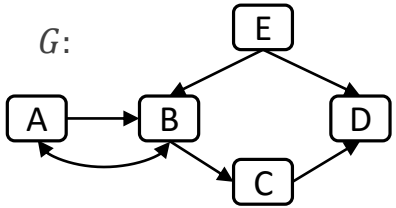
REVERSE ENGINEERING



Data set D measuring a set of variables

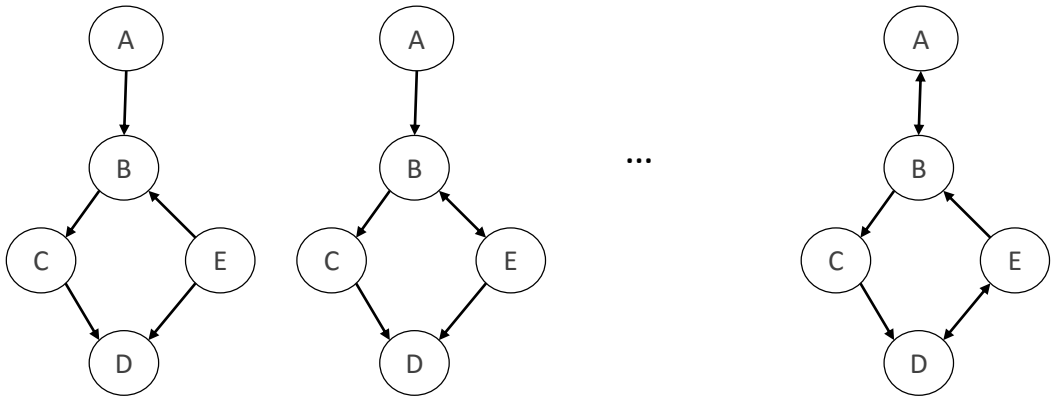
$A, B \mid E, C$	Ind
$A, B \mid \emptyset$	Dep
…	…
$E, C \mid A, B, C$	Dep

Find the (in)dependencies using statistical tests.



Find a graph that satisfies the implied m-connections/separations.

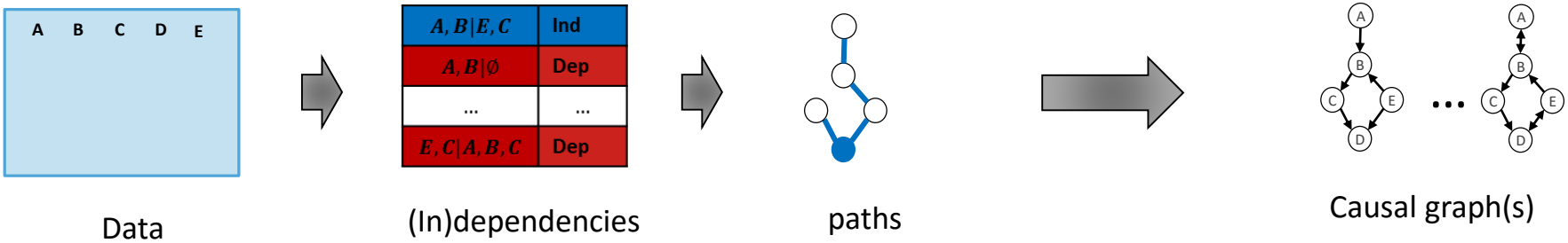
MARKOV EQUIVALENCE



$A, B \mid E, C$	Ind
$A, B \mid \emptyset$	Dep
...	...
$E, C \mid A, B, C$	Dep

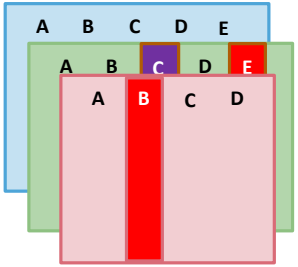
- More than one graphs entail the same set of conditional independencies.
- The graphs have some common features (edges/orientations).
- For some types of causal graphs, Markov equivalence classes share the same skeleton.
    - not semi-Markov causal graphs

CAUSAL DISCOVERY

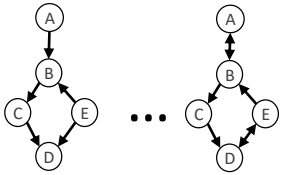


Sound and complete algorithms (e.g., FCI) take as input a data set and output a summary of all the graphs that satisfy all identified conditional independencies.

INTEGRATIVE CAUSAL DISCOVERY

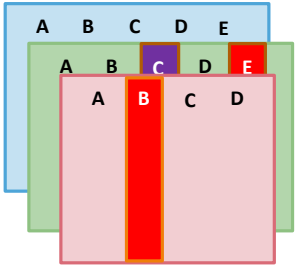


Data sets measuring overlapping variable sets under intervention/selection.

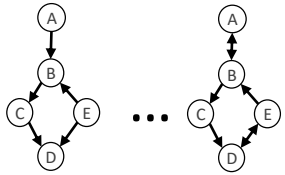


Causal graph(s) that simultaneously fit all data.

INTEGRATIVE CAUSAL DISCOVERY



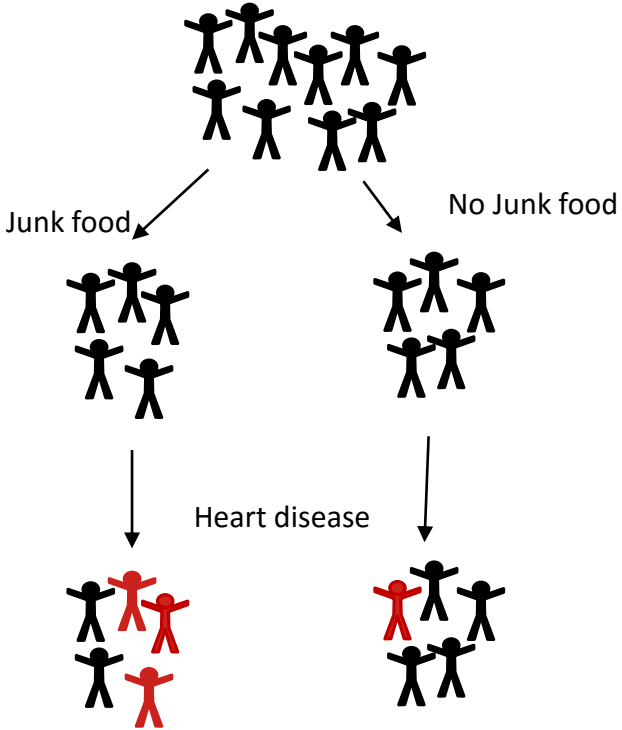
Data sets measuring overlapping variable sets under intervention/selection.



Causal graph(s) that simultaneously fit all data.

- Every data set imposes some constraints.
- Observational data impose m-separation/m-connection constraints on the candidate graph.
- Different variables?
- Experimental data?
- Data sampled under selection?

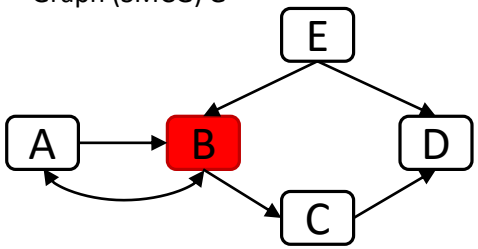
INTERVENTIONS (MANIPULATIONS)



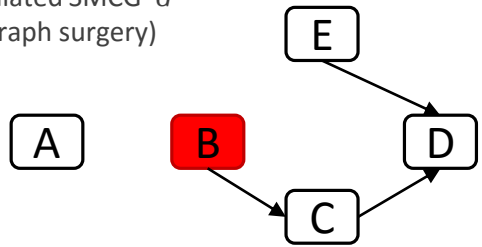
Values of the manipulated variable are **set** solely **by the intervention procedure**
e.g. a randomized variable in a randomized control trial.

INTERVENTIONS

Graph (SMCG) G

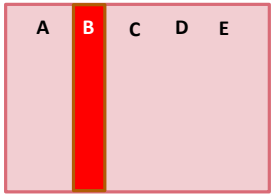


Manipulated SMCG G^B
(after graph surgery)

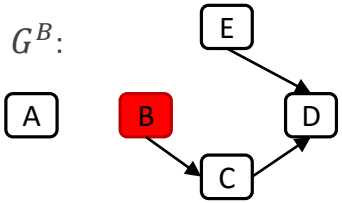


- If you know the causal model, you can model interventions.
- Values of B are **set** solely **by the intervention procedure:** If you know direct causal relations, **remove all edges into the manipulated variable.**
- This procedure is called graph surgery.
  - The resulting graph is called the **manipulated graph** (symb. G^B)

CAUSAL DISCOVERY WITH INTERVENTIONS



$A, B \mid E, C$	Ind
$A, B \mid \emptyset$	Dep
…	…
$E, C \mid A, B, C$	Dep



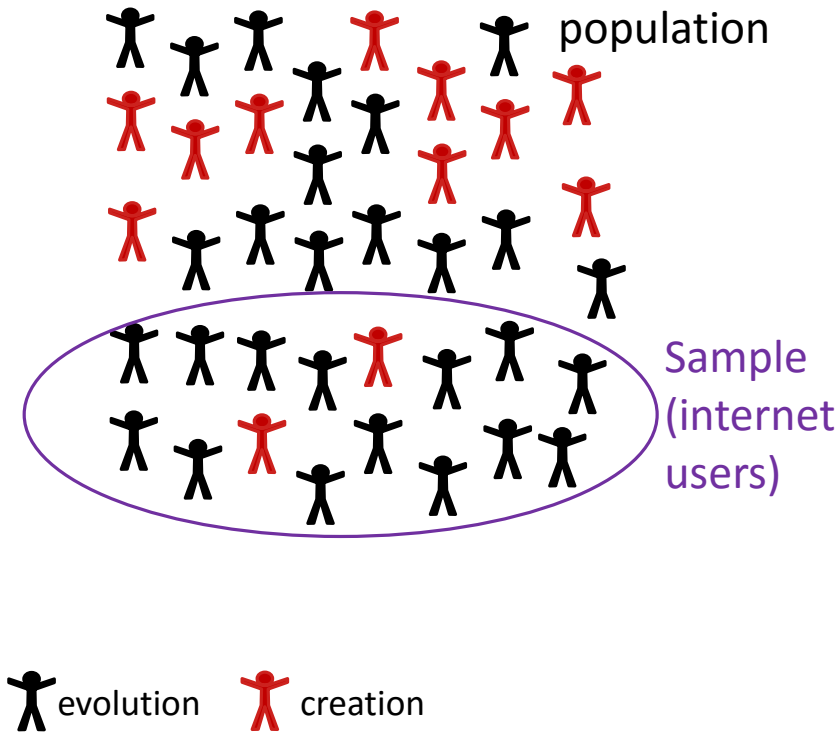
$\nexists$ *m-connecting path from A to D given* $\emptyset$ *in* G^B
$\nexists$ *m-connecting path from A to D given* B *in* G^B
⋮
$\nexists$ *m-connecting path from A to D given* B, C *in* G^B
⋮
$\exists$ *m-connecting path from B to C given* $\emptyset$ *in* G^B

Dataset D_i measuring a subset of variables, some of which are manipulated

Conditional independencies in D_i

Path constraints on the causal graph after **manipulation**

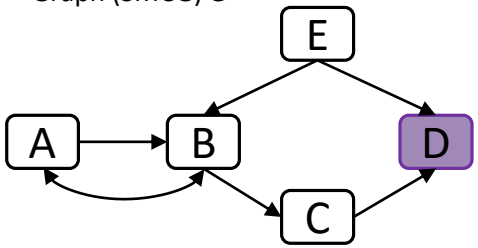
SELECTION BIAS



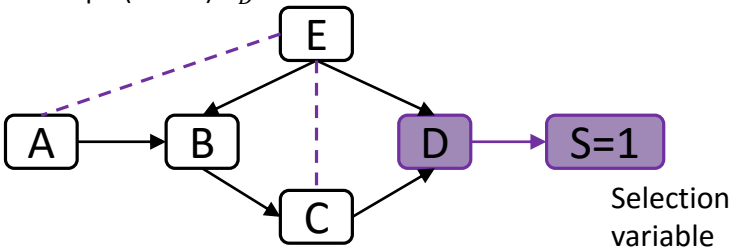
- Samples are selected based on the value of one of your variables.
- e.g. you perform your study in a specific region/on the internet; case-control study for a rare disease.

SELECTION BIAS IN CAUSAL MODELS

Graph (SMCG) G

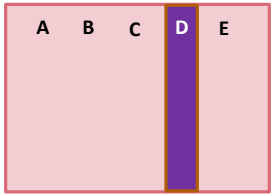


Selected Graph (SMCG) G_D

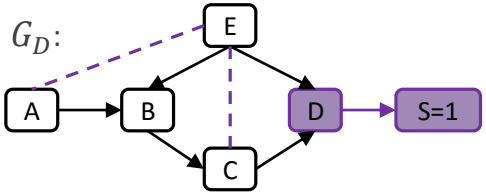


- If you know the causal model, you can model selection bias.
- Samples are selected based on the value of D; The value of D directly affects the probability of being selected.
- S is a child of D, S=1 for all your samples.
- Selected graph, symb. G_D

CAUSAL DISCOVERY WITH SELECTION BIAS



$A, B \mid E, C, S=1$	Ind
$A, B \mid S=1$	Dep
…	…
$E, C \mid A, B, D, S=1$	Dep



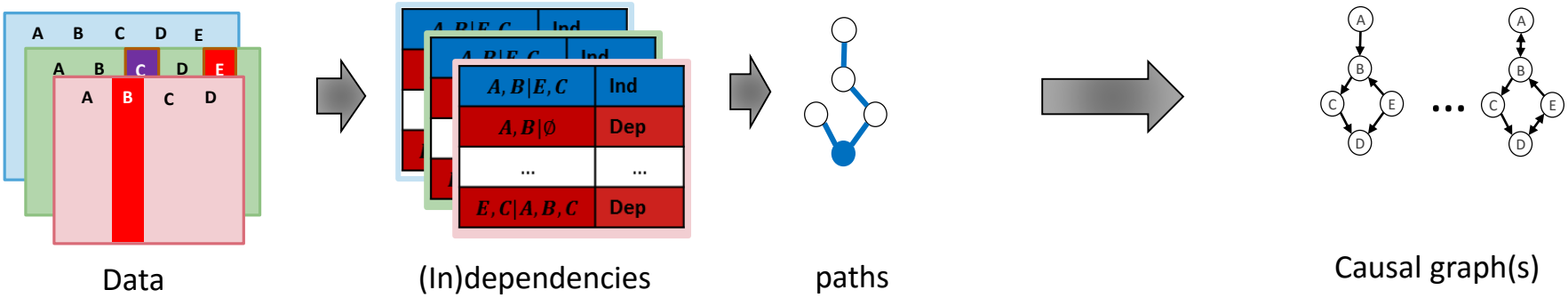
$\nexists$ *m-connecting path from A to D given* $\emptyset$ *in* G_D
$\nexists$ *m-connecting path from A to D given* B *in* G_D
⋮
$\nexists$ *m-connecting path from A to D given* B, C *in* G_D
⋮
$\exists$ *m-connecting path from B to C given* $\emptyset$ *in* G_D

Dataset D_i measuring a subset of variables, some of which are selected upon

Conditional independencies in D_i

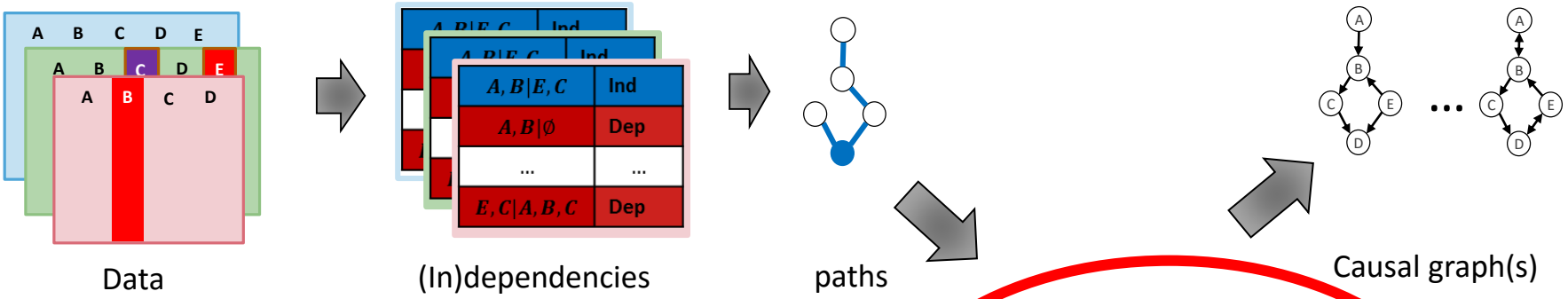
Path constraints on the underlying causal graph after **selection**

INTEGRATIVE CAUSAL DISCOVERY



- Every data set imposes some constraints.
- Observational data impose path constraints on the candidate graph.
- Experimental data impose path constraints on the candidate graph after manipulation.
- Data sampled under selection impose path constraints on the candidate graph after selection.
- Easily handles overlapping variable sets
    - Each study imposes constraints on the observed variables.

LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY



Convert to logic formula!

Variables of the formula correspond to graph features (edges, orientations).

Truth setting assignments encode graphs that satisfy all path constraints after manipulation/selection.

Logic encoding Φ of path constraints in the causal graph

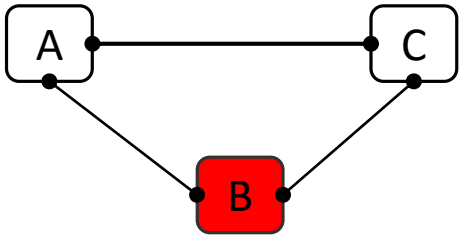
$$[E_{A\to D} \vee [E_{A\to B} \wedge E_{B\to D}] \vee [E_{A\to C} \wedge E_{C\to D}] \vee \cdots]$$

$$\vdots$$

$$[E_{A\to C} \vee [E_{A\to B} \wedge E_{B\to C}] \vee [E_{A\leftrightarrow C} \wedge E_{C\to D}] \vee \ldots]$$

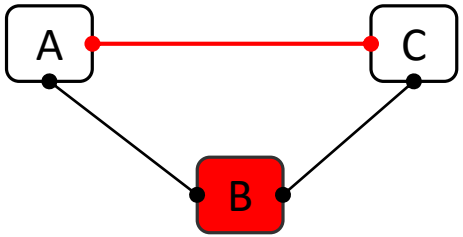
CONVERSION TO LOGIC FORMULA: EXAMPLE (INTERVENTION)

- Suppose you know nothing about the causal structure G of A, B, C.
- In a data set where B is manipulated, $\mathrm{Ind}(\mathrm{A}, \mathrm{C} | \emptyset)$
- In path terms: $\nexists$ m-connecting path between A and C given $\emptyset$ in G^B.



CONVERSION TO LOGIC FORMULA: EXAMPLE (INTERVENTION)

- Suppose you know nothing about the causal structure G of A, B, C.
- In a data set where B is manipulated, $\mathrm{Ind}(\mathrm{A}, \mathrm{C}|\emptyset)$
- In path terms: $\nexists$ m-connecting path between A and C given $\emptyset$ in G^B.

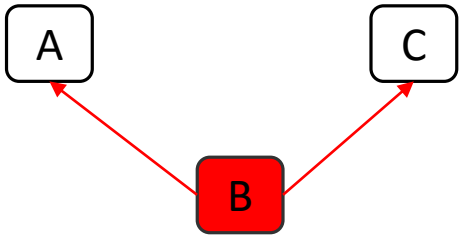


A-C does not exist

$\neg E_{A\to C} \wedge \neg E_{A\leftarrow C} \wedge \neg E_{A\leftrightarrow C}$

CONVERSION TO LOGIC FORMULA: EXAMPLE (INTERVENTION)

- Suppose you know nothing about the causal structure G of A, B, C.
- In a data set where B is manipulated, $\mathrm{Ind}(\mathrm{A}, \mathrm{C}|\emptyset)$
- In path terms: $\nexists$ m-connecting path between A and C given $\emptyset$ in G^B.



A-C does not exist

$\neg E_{A\to C} \wedge \neg E_{A\leftarrow C} \wedge \neg E_{A\leftrightarrow C}$

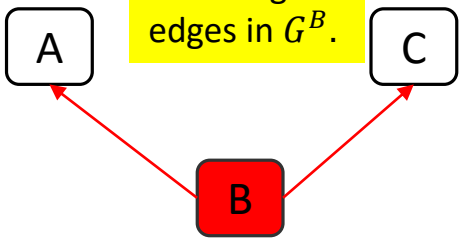
A-B-C is not m-connecting

$\neg(E_{B\to A} \wedge E_{B\to C})$

CONVERSION TO LOGIC FORMULA: EXAMPLE (INTERVENTION)

- Suppose you know nothing about the causal structure G of A, B, C.
- In a data set where B is manipulated, $\mathrm{Ind}(\mathrm{A}, \mathrm{C}|\emptyset)$
- In path terms: $\nexists$ m-connecting path between A and C given $\emptyset$ in G^B.

B has no incoming edges in G^B.



A-C does not exist

$\neg E_{A\to C} \wedge \neg E_{A\leftarrow C} \wedge \neg E_{A\leftrightarrow C}$

A-B-C is not m-connecting

$\neg(E_{B\to A} \wedge E_{B\to C})$

Logic formula:

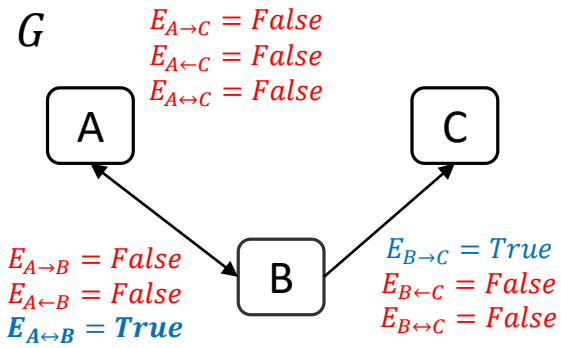
$$(\neg E_{A\to C} \wedge \neg E_{A\leftarrow C} \wedge \neg E_{A\leftrightarrow C}) \wedge \neg(E_{A\leftarrow B} \wedge E_{B\to C})$$

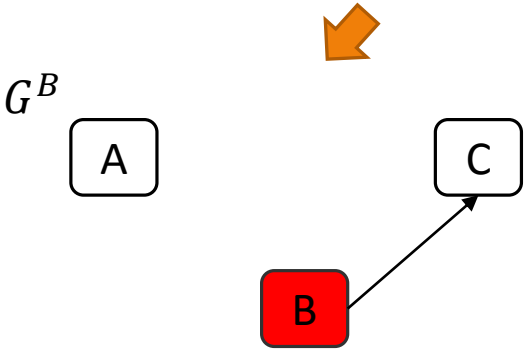
CONVERSION TO LOGIC FORMULA: EXAMPLE

Logic formula: TRUE

$$(\neg E_{A\to C} \wedge \neg E_{A\leftarrow C} \wedge \neg E_{A\leftrightarrow C}) \wedge$$

$$\neg(E_{A\leftarrow B} \wedge E_{B\to C}) \wedge$$





CONVERSION TO FIRST-ORDER LOGIC: INPUT CONSTRAINTS

1) As many (conditional) **dependencies** and **independencies** from multiple datasets as desired, even datasets over different variables

2) Meta-Information about the datasets

- for each variable and dataset, whether **it was used for selection, or not, or unknown**
- for each variable and dataset, whether it was **manipulated** (soft or hard), not, or unknown

3) Structural prior knowledge

- presence/absence of direct edges, paths or dependencies
- root/leaf nodes
- any structural constraint that can be expressed in first-order logic

CONVERSION TO FIRST-ORDER LOGIC: LOGIC VARIABLES AND SEMANTICS

Logic variables represent **features of the graph** and datasets:

edges, directed paths, m-connecting paths, selection targets, intervention targets

Set to **true** if $Dep(X, Y|\, \boldsymbol{Z})$ is determined in dataset D and **false** otherwise

Set to **true** if *X* is known to be used for selection in dataset D

Set to **true** if *X* is a known target of a manipulation in dataset D

$X \rightarrow Y$	X has an arrow into Y
$X \leftrightarrow Y$	X and Y are confounded
$X \dashrightarrow Y$	X is an ancestor of Y
$X \underset{Z,D}{\cdots} Y$	mconn$(X, Y, \mathbf{Z})$ in dataset D
$X \underset{Z,D}{\cdot\cdot>} Y$	mconn$(X, Y, \mathbf{Z})$ in dataset D (path into Y)
$X \underset{Z,D}{\cdot\cdot-} Y$	mconn$(X, Y, \mathbf{Z})$ in dataset D (path out of Y)
X^S_D	X is used for selection in dataset D
X^I_D	X is manipulated (hard) in dataset D

CONVERSION TO FIRST-ORDER LOGIC : INFERENCE RULES ETIO ALGORITHM (KDD 2016)

$$X \dashrightarrow Y \Leftrightarrow X \rightarrow Y \vee (X \dashrightarrow U \wedge U \dashrightarrow Y) \quad (1)$$

Ancestry

$$\neg Y \rightarrow X \Leftarrow X \dashrightarrow Y \quad (2)$$

Acyclicity

$$X \underset{Z,D}{\cdots} Y \Leftrightarrow X \underset{Z,D}{\cdot\cdot>} Y \vee X \underset{Z,D}{\cdot\cdot-} Y \quad (3)$$

m-connections

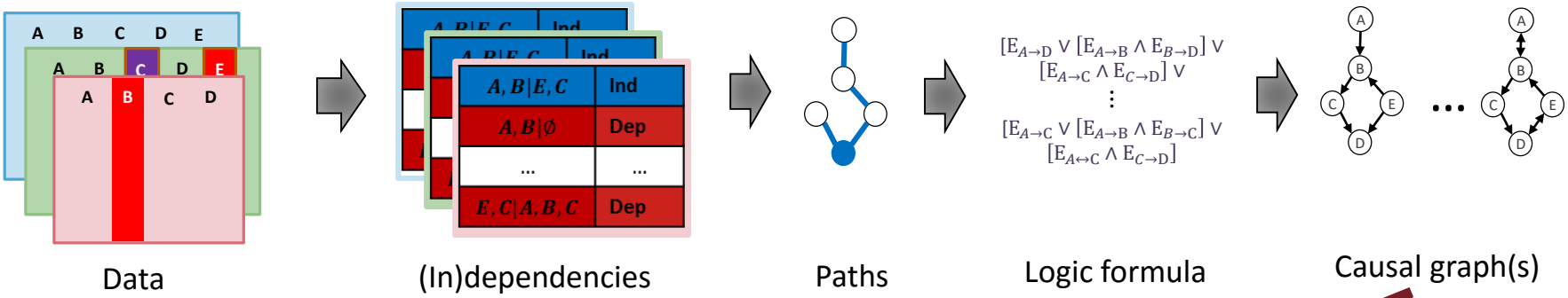
$$\begin{aligned}
X \underset{Z,D}{\cdot\cdot-} U \Leftrightarrow\ & (X = U \wedge X \notin Z) && (4a)\\
\vee\ & (X \underset{Z,D}{\cdot\cdot-} Y \wedge U \rightarrow Y \wedge Y \notin Z \wedge \neg Y_D^I) && (4b)\\
\vee\ & (X \underset{Z,D}{\cdot\cdot>} Y \wedge U \rightarrow Y \wedge Y \in Z \wedge \neg Y_D^I) && (4c)\\
\vee\ & X \underset{Z,D}{\cdot\cdot-} Y \wedge Y_D^S \wedge U_D^S \wedge Y \notin Z && (4d)\\
\vee\ & X \underset{Z,D}{\cdot\cdot>} Y \wedge Y_D^S \wedge U_D^S \wedge Y \notin Z && (4e)
\end{aligned}$$

m-connections out of U

$$\begin{aligned}
X \underset{Z,D}{\cdot\cdot>} U \Leftrightarrow\ & (X \underset{Z,D}{\cdot\cdot-} Y \wedge Y \rightarrow U \wedge Y \notin Z \wedge \neg U_D^I) && (5a)\\
\vee\ & (X \underset{Z,D}{\cdot\cdot>} Y \wedge Y \rightarrow U \wedge Y \notin Z \wedge \neg Y_D^I \wedge \neg U_D^I) && (5b)\\
\vee\ & (X \underset{Z,D}{\cdot\cdot-} Y \wedge Y \leftrightarrow U \wedge Y \notin Z \wedge \neg U_D^I) && (5c)\\
\vee\ & (X \underset{Z,D}{\cdot\cdot>} Y \wedge Y \leftrightarrow U \wedge Y \in Z \wedge \neg Y_D^I \wedge \neg U_D^I) && (5d)
\end{aligned}$$

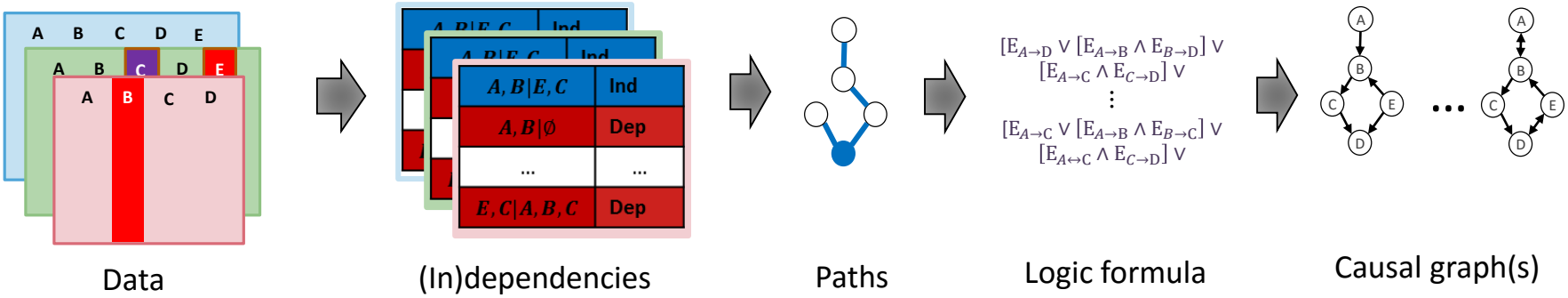
m-connections into U

LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY



Exponential number of
1.Independencies
2.Paths
3. Solutions

LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY

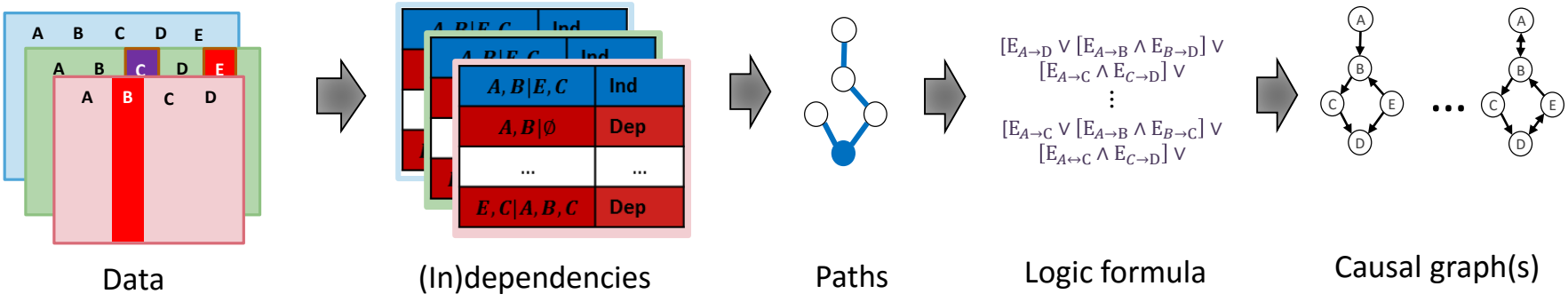


Reduce the number of independencies:

Run FCI and use only the tests performed by FCI.

Limit max conditioning set size.

LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY

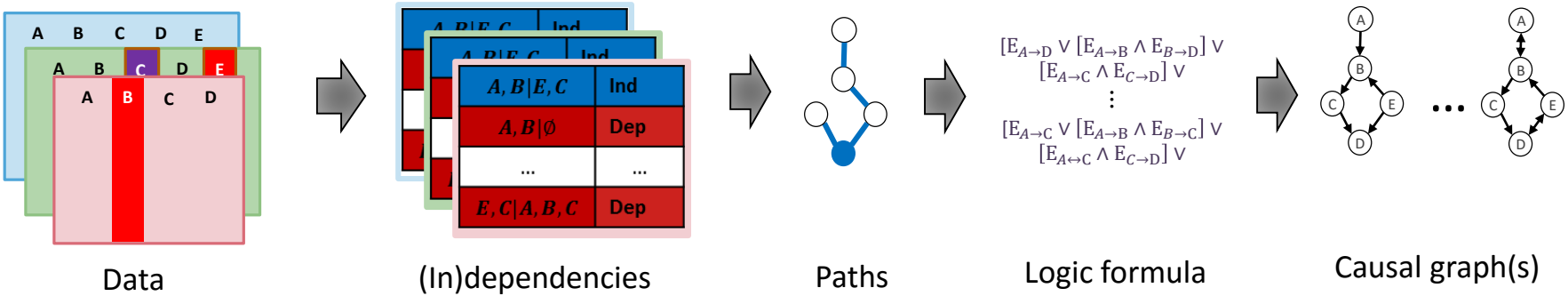


Reduce the number of paths:

Use inducing paths that connect paths on the graph to ∃ of independence (given any set).

Limit the maximum path length.

LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY

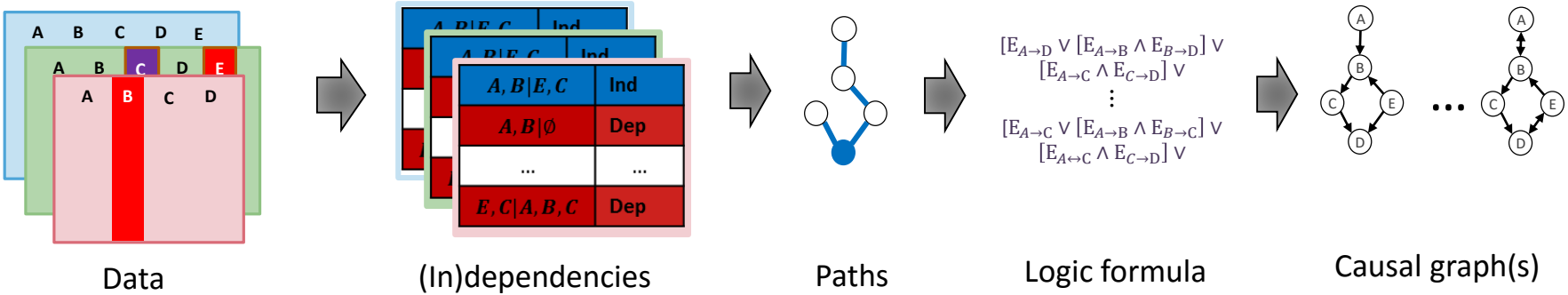


Need a clever way to encode constraints!

e.g. recursively encode paths.

Convert to CNF for most SAT solvers.

LOGIC-BASED INTEGRATIVE CAUSAL DISCOVERY



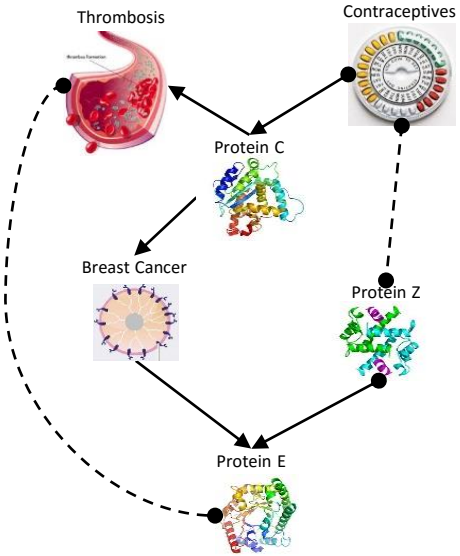
No need to enumerate all solutions!

Query the formula for

- A single causal graph.
- A causal graph with specific features.
- Features that are invariant in all possible causal graphs.

SUMMARIZING PAIRWISE RELATIONS

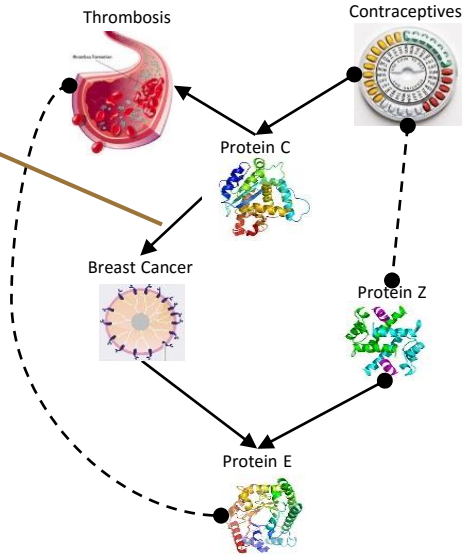
Absent edges:
Absent in **all** solutions



SUMMARIZING PAIRWISE RELATIONS

Absent edges:
Absent in **all** solutions

solid edges:
present in **all** solutions

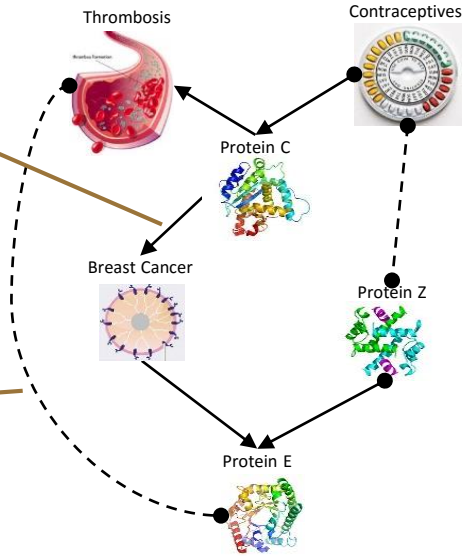


SUMMARIZING PAIRWISE RELATIONS

Absent edges: Absent in **all** solutions

solid edges: present in **all** solutions

dashed edges: present in **some** solutions



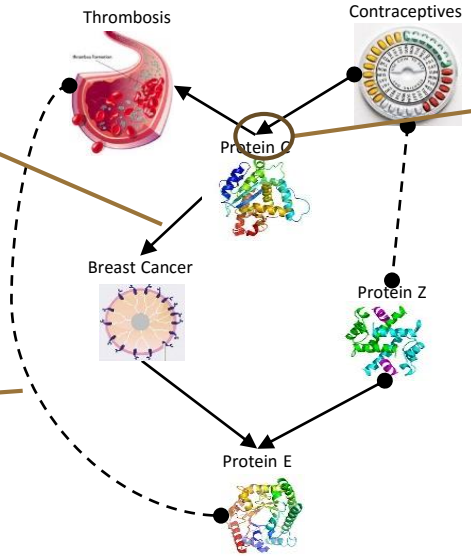
SUMMARIZING PAIRWISE RELATIONS

Absent edges: Absent in **all** solutions

solid edges: present in **all** solutions

solid endpoints: same orientation in **all** solutions

dashed edges: present in **some** solutions

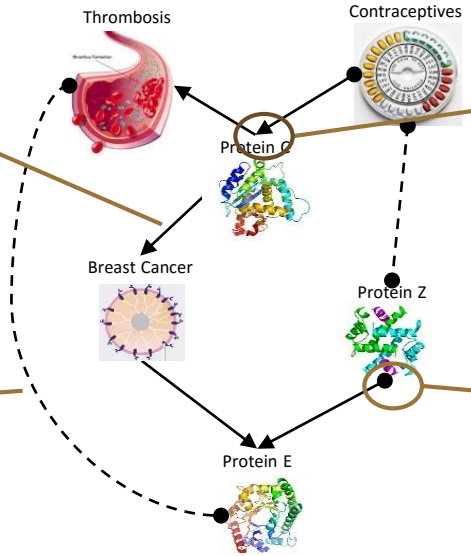


SUMMARIZING PAIRWISE RELATIONS

Absent edges: Absent in **all** solutions

solid edges: present in **all** solutions

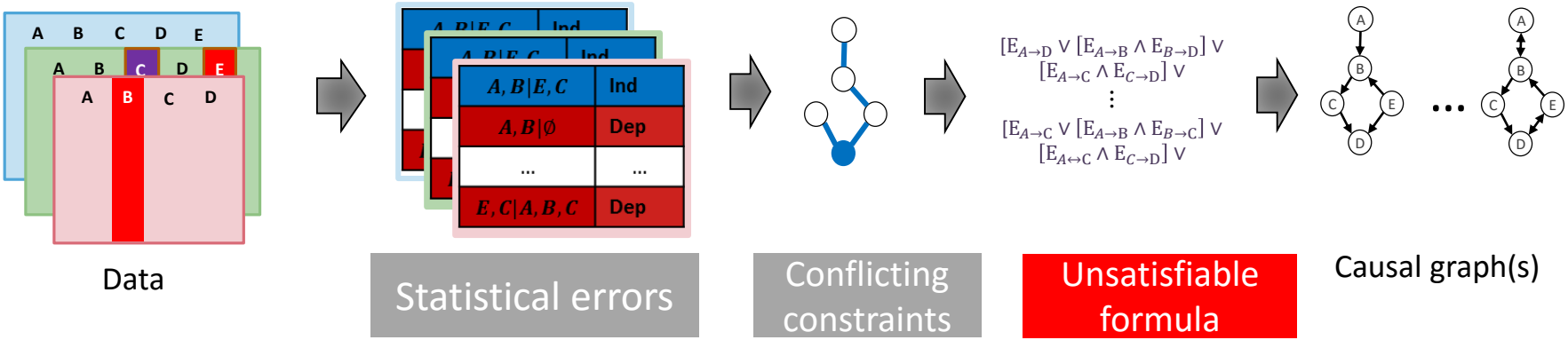
solid endpoints: same orientation in **all** solutions



dashed edges: present in **some** solutions

Circle endpoints: orientation varies in **different** solutions

STATISTICAL ERRORS RESULT IN CONFLICTING INPUTS



Convert p-values to probabilities

Solve a subset of constraints optimizing a function of the probabilities

EXISTING ALGORITHMS

Vary in:

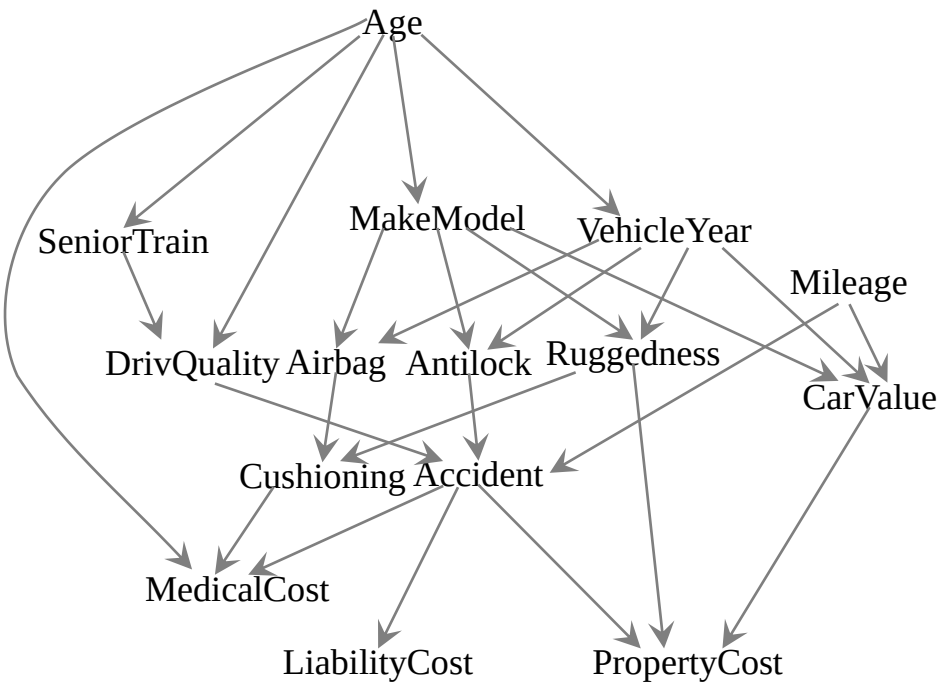
- Type of constraints:
  - different types of paths (m-connecting, inducing, ancestral).
  - translation to logic formula.
- Types of heterogeneity:
  - Soft/hard interventions, selection.
- Preprocessing:
  - Heuristics to limit number of constraints / paths.
- Conflict Resolution
  - Method for calculating probabilities.
  - Conflict resolution strategy (greedy/ max SAT / weighted max SAT).
- CS solver
  - Initially SAT solvers, more recently ASP.
- Scalability
  - Depends on choices above. Be exact/ focus on scalability.
  - Difficult to determine
    - huge variance depending on the problem.

CSAT+ [Triantafillou, et al., AISTATS 2010]
LOCI [Claassen and Heskes, UAI 2011]
SAT-Based Causal Discovery [Hyttinen, et al., UAI 2013]
Constraint-Based CD [Hyttinen, et al., UAI 2014]
COmbINE [Triantafillou and Tsamardinos, JMLR 2015]
ETIO [Borboudakis and Tsamardinos, KDD 2016]
ACI [S. Magliacane, T. Claassen, J.M. Mooij, *arXiv*]

MORE

- Using conversion to logic for causal discovery from time-course data
  - Causal Discovery from Subsampled Time Series Data by Constraint Optimization, [Hyttinen, Plis, Järvisalo, Eberhardt and Danks, arXiv, 2016]
- Using conversion to logic for identifying chain graphs.
  - Learning Optimal Chain Graphs with Answer Set Programming
  - [Sonntag, Järvisalo, Penã, Hyttinen, UAI 2015]
- Using conversion to logic to identify semi-Markov causal graphs.
  - [Penã, UAI 2016]
- Using conversion to logic to estimate causal effects for an unknown graph
  - [Hyttinen, Eberhardt and Järvisalo, UAI 2015]
- Massive proof-of-concept proof the techniques work for real data and can become quantitative
  - [Tsamardinos, et al. JMLR 2012]
- <u>More details, examples, references in recent UAI 2016 Tutorial Triantafillou & Tsamardinos</u>

USE CASE: THE INSURANCE DATASET
REAL CAUSAL GRAPH

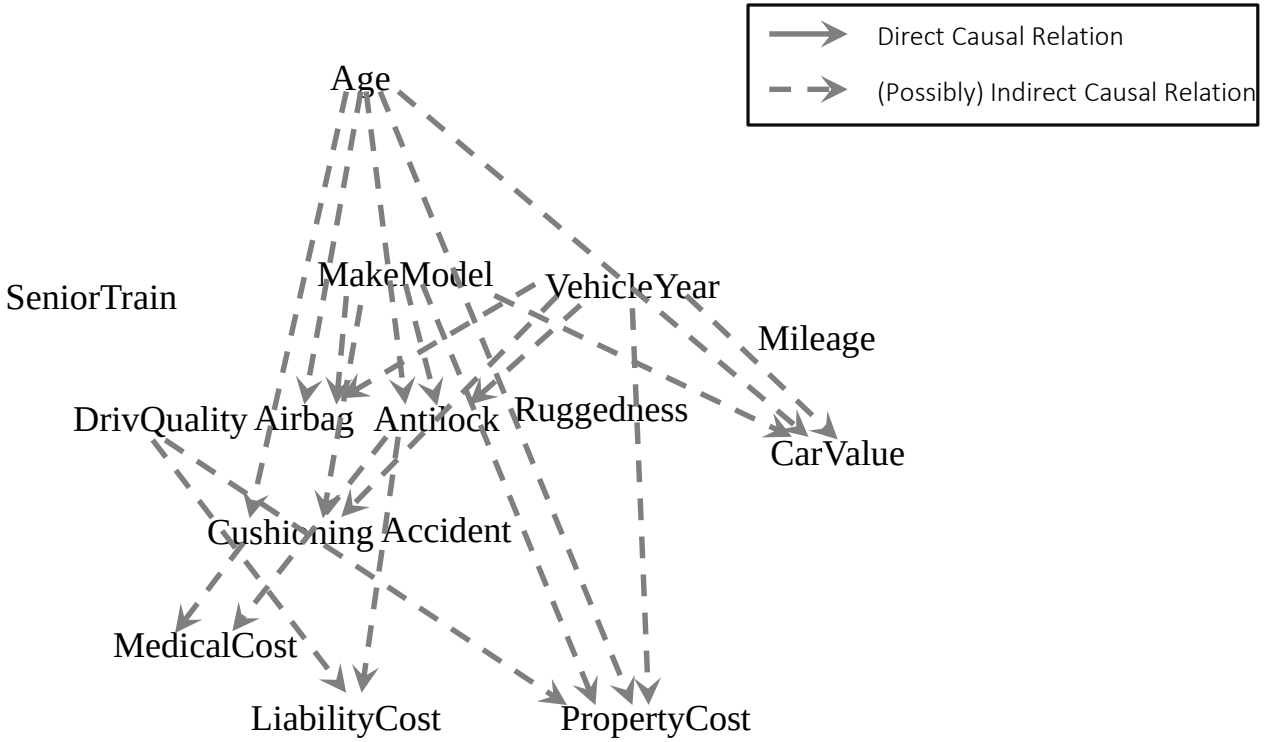


APPLICATION ON REAL PROBLEMS

Insurance Data Datasets	Age	SeniorTrain	MakeModel	VehicleYear	DrivQuality	Airbag	Antilock	Ruggedness	Mileage	CarValue	Cushioning	Accident	MedicalCost	LiabilityCost	PropertyCost
Observational							X			X					
Selected based on Antilock								X			X	X	X		
Soft Intervention on Cushioning							X			X					
Prior Knowledge							Anything $\not\rightarrow$ Age Costs $\not\rightarrow$ Anything Age $\dashrightarrow$ Costs								

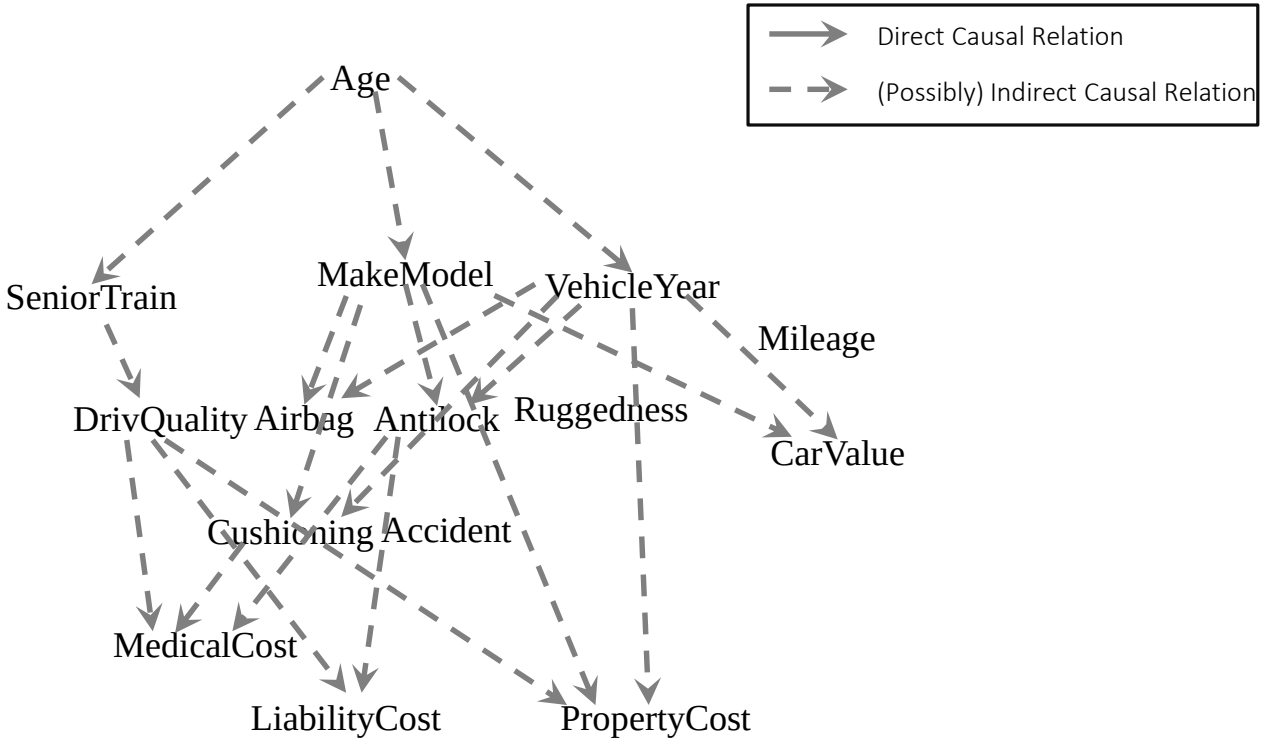
USE CASE: THE INSURANCE DATASET PROVED ANCESTRY RELATIONS

Datasets
Observational



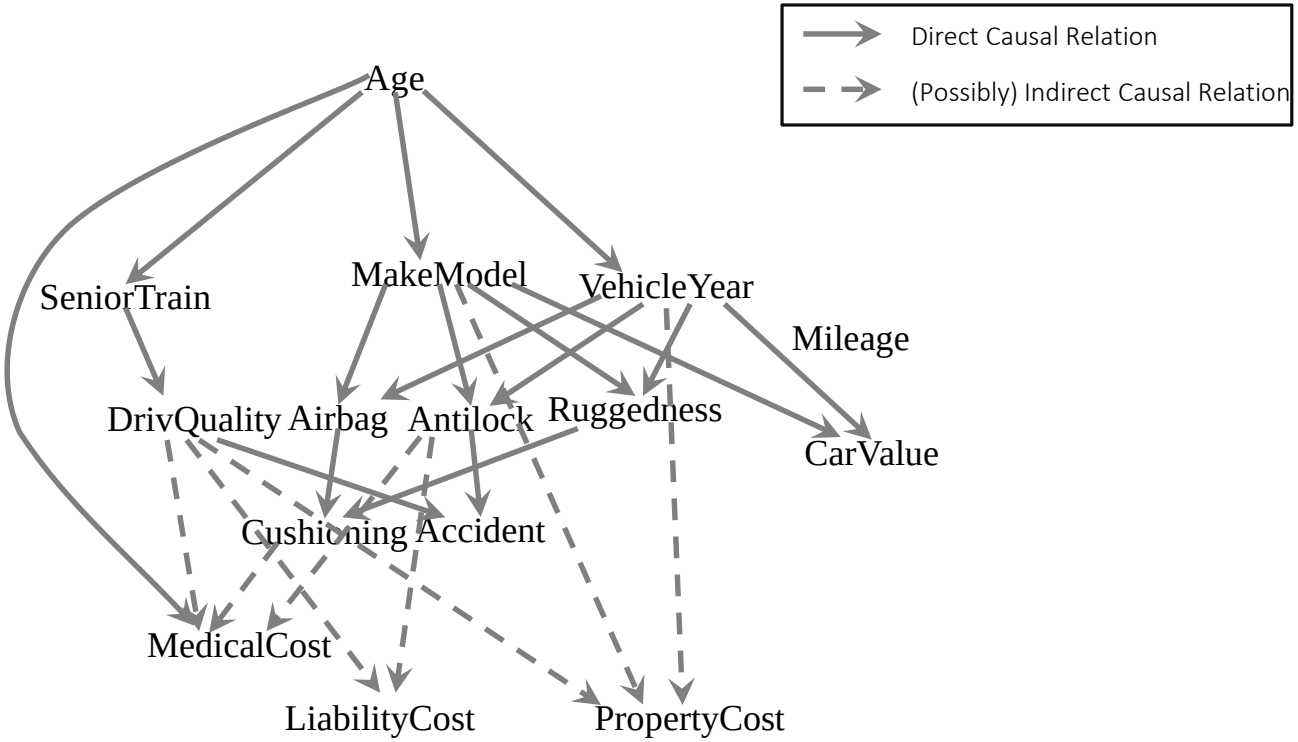
USE CASE: THE INSURANCE DATASET PROVED ANCESTRY RELATIONS (TRANSITIVE REDUCTION)

Datasets
Observational
Prior Knowledge

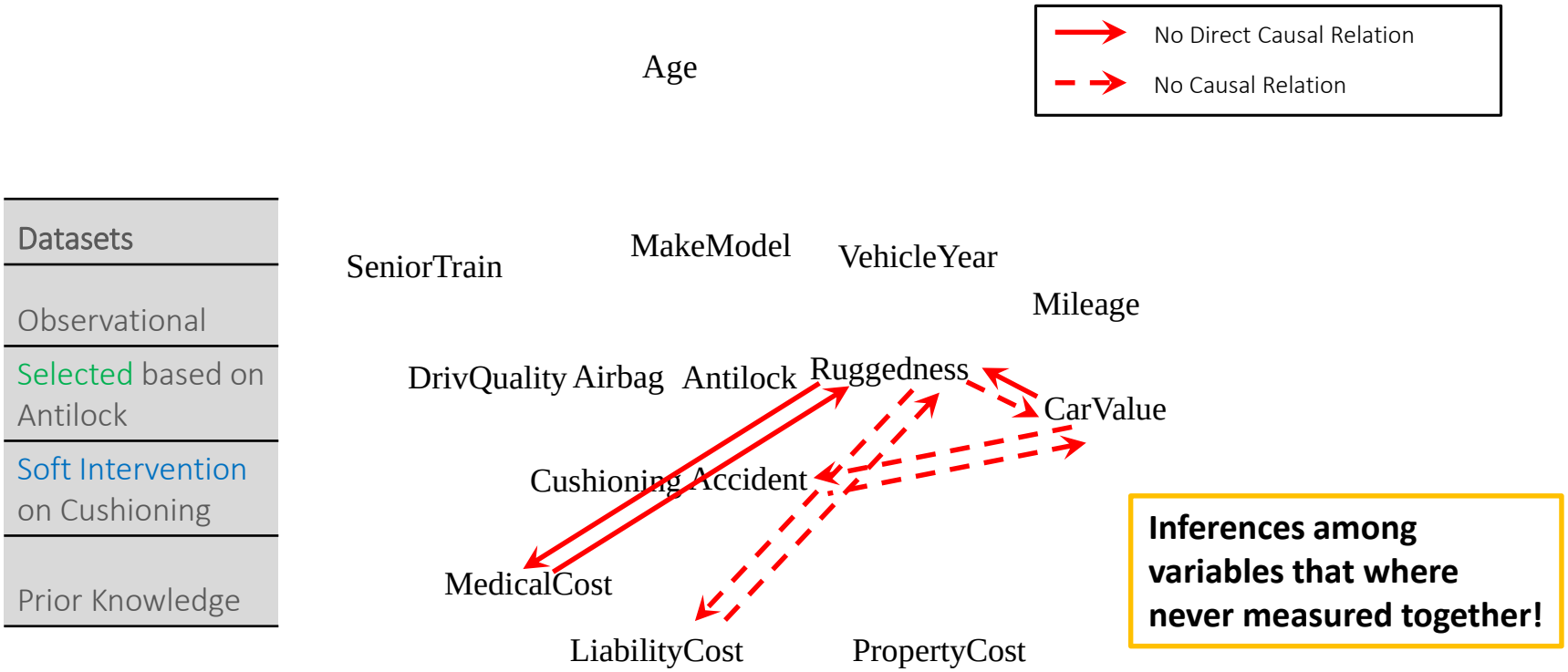


USE CASE: THE INSURANCE DATASET PROVED ANCESTRIES AND DIRECT CAUSAL RELATIONS

Datasets
Observational
Selected based on Antilock
Soft Intervention on Cushioning
Prior Knowledge



USE CASE: THE INSURANCE DATASET
NON-TRIVIAL INFERENCES



Inferences among variables that where never measured together!

KEY-POINTS

Integrative logic-based causal discovery.

Different data distributions, same causal mechanism: use causal modeling to connect.

Can handle datasets of different variable sets, different experimental conditions, prior causal knowledge.

Identify the set of causal graphs that simultaneously fit all datasets and reason with this set.

Convert problem to SAT or ASP; exploit 40 years of SAT-solving technology.

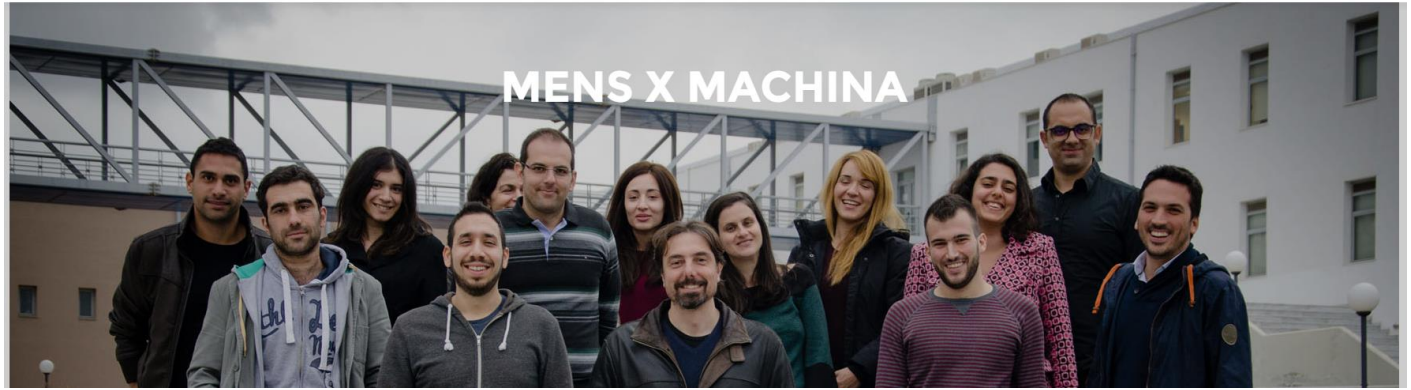
Query-based approach to avoid explosion of possible solutions!

Vision of automatically analyzing a large portion of available datasets in a domain.

ACKNOWLEDGEMENTS

Mens x machina group, University of Crete.
Jan Lemeire, Frederick Eberhardt, Antti Hyttinen, Joris Mooij

Funded by ERC Consolidator Grant CAUSALPATH
Host: University of Crete



SNEAK PREVIEWS TO MXM RESEARCH

LOGIC-BASED CAUSAL DISCOVERY

- Scalability, robustness
- Relax assumptions such as Faithfulness
- Making quantitative predictions
- Extend for temporal data
- Add Verma constraints
- **Application to a real-life insurance problem**

FEATURE SELECTION – FASTER, BETTER, MULTIPLE SOLUTIONS, BIG DATA

Forward-Backward Selection (FBS)

- **very slow**, especially for data with many variables
- returns **single solution**

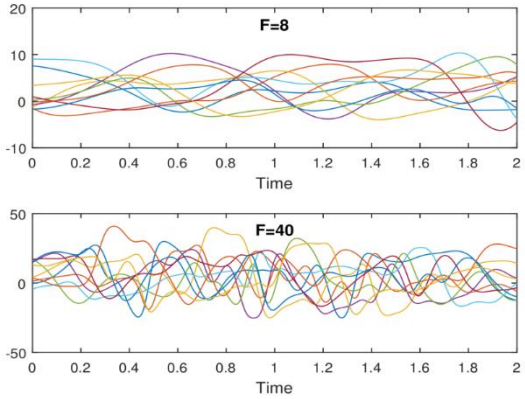
We extended FBS

- Improving **computational performance** by 1-3 orders of magnitude
- Reducing **number of selected variables**, selecting up to 5 times fewer variables
- With comparable or better **predictive performance**
- With the ability to return **multiple**, statistically equivalent **solutions**

Extended single solution FBS also for **Big Data**

- Further improving computational performance, able to **run on millions of samples and variables**
- **Vastly outperforming state-of-the-art** feature selection methods on Big Data
- Almost linear speedup with available cores
- Super-linear scalability with sample size

LEARNING ORDINARY DIFFERENTIAL EQUATION MODELS



Trajectories of Lorenz96 climate model. It is a chaotic system of Ordinary Differential Equations given by:

$$\dot{x}_n = x_{n-1}x_{n+1} - x_{n-1}x_{n-2} - x_n + F,$$
$$n = 1, \dots, N$$

Algorithms for learning the *structure* and the parameters of a Dynamical System from *time-course* measurements

(1) *Eliminate* the time dimension by transforming the original problem to an *atemporal* one.

(2) *Solve* the transformed problem using the *Sparse Signal Identification* theory.

R PACKAGE MXM: DESCRIPTION

Main focus of the package:

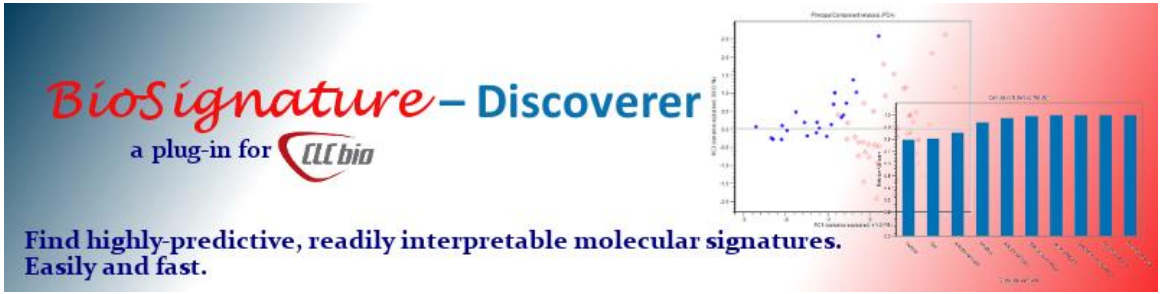
- Variable Selection
- (Causal) Bayesian Networks

Available variable selection methods span prototypical algorithms (forward, backward regression) and advanced ones (SES, MMPC)

- A plethora of different data types can be addressed: continuous, ordinal, categorical, survival, proportions, longitudinal, clustered.

Algorithmic and implementation optimization (e.g., several function are implemented directly in C++)

AYTOMATED MACHINE LEARNING



- Commercial CLC-Bio (a QIAGEN company) plugin for high-throughput data analysis.
- **Automatically identifies multiple** signatures.
- Can handle various data types.
  - Including binary, multi-class, continuous, and **time-to-event** outcomes.
- **Computationally efficient**, fine-tuned implementation.
  - Easily handles even tens of thousands of molecular quantities.
- **High quality results,** using state-of-the art techniques.
- **Interpretable output**, helping the user understand the results.
- **Soon available as a cloud service**

CASE-STUDY: CLASSIFICATION ANALYSIS IN BREAST TUMORS

LaBreche *et al. BMC Medical Genomics* 2011, **4**:61
http://www.biomedcentral.com/1755-8794/4/61

BMC Medical Genomics

RESEARCH ARTICLE Open Access

Integrating Factor Analysis and a Transgenic Mouse Model to Reveal a Peripheral Blood Predictor of Breast Tumors

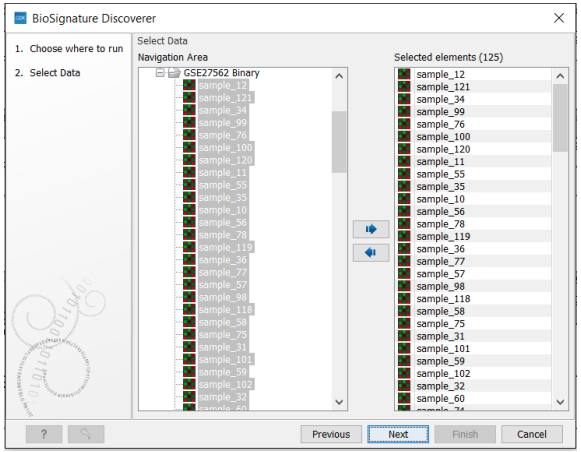
Heather G LaBreche[1,2*], Joseph R Nevins[1,2] and Erich Huang[1,3,4]

125 gene expression profiles of patients

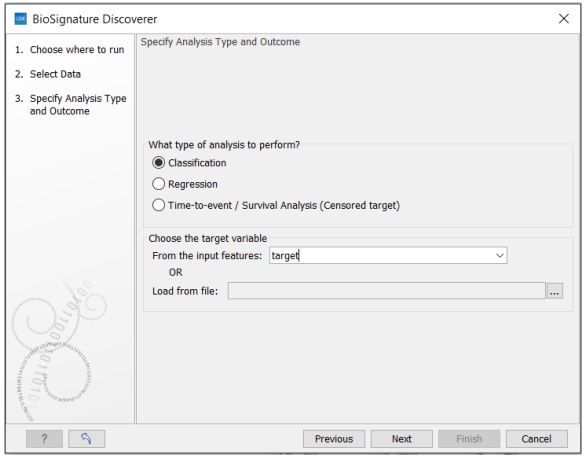
- 31 normal, 94 breast tumor (37 benign, 57 malignant)

54,675 gene expression probesets

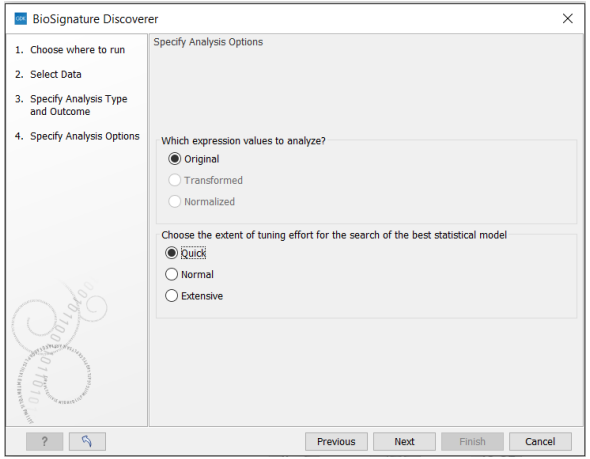
Introduced in LaBreche et al., BMC medical genomics (2011)



1 Selecting the data



2 Choose the type of analysis



3 Tuning-effort level

ANALYSIS RESULTS

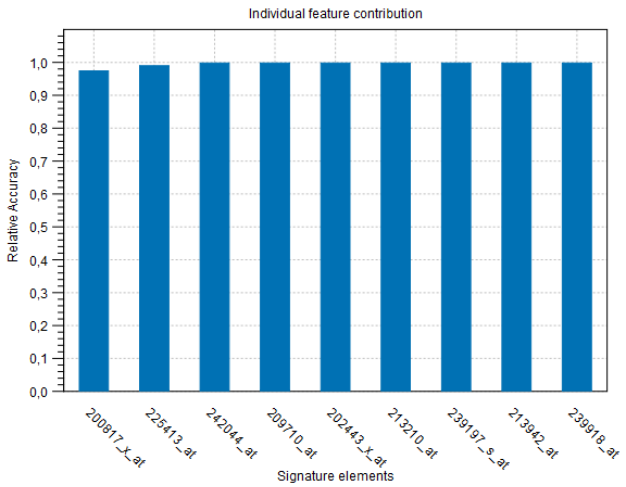
4 Performance Metrics

Metric	Average	95% Confidence Interval
Accuracy	0,977	[0,946, 1,000]
Area Under the ROC Curve	0,985	[0,951, 1,000]
Precision for class 1.0	0,983	[0,900, 1,000]
Precision for class 2.0	0,970	[0,930, 1,000]
Recall for class 1.0	0,917	[0,817, 1,000]
Recall for class 2.0	1,000	[0,940, 1,000]
Sensitivity for class 1.0	0,917	[0,817, 1,000]
Sensitivity for class 2.0	1,000	[0,940, 1,000]
Specificity for class 1.0	1,000	[0,940, 1,000]
Specificity for class 2.0	0,917	[0,817, 1,000]

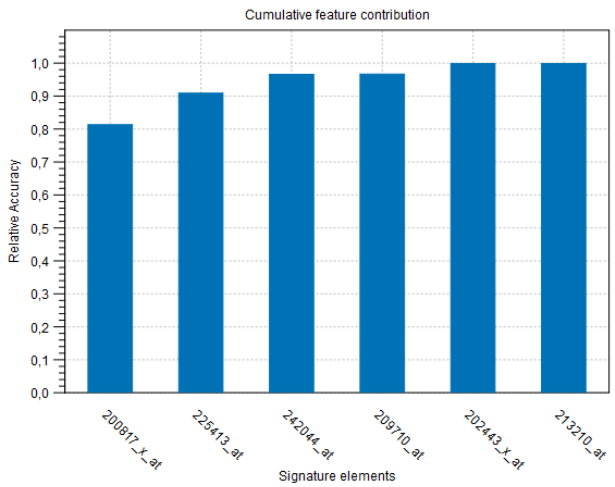
[1]LaBreche et al. achieve 0.97 Area Under the ROC Curve

[2]Analysis took less than 3 minutes on a laptop

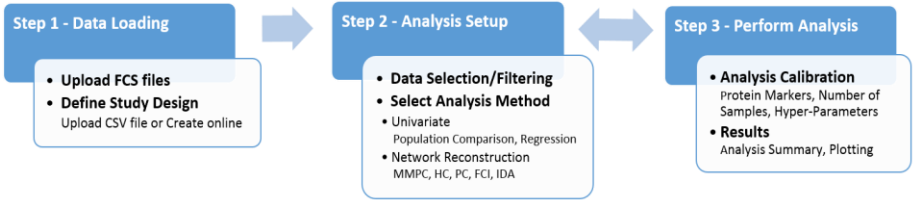
5 Individual feature contribution



6 Cumulative feature contribution



SINGLE-CELL NETWORK RECONSTRUCTION SYSTEM (SCENERY)



Architecture

Web-based, open architecture

Wizard design pattern: Step-based User Interface

Modularity: Easy to incorporate new analysis methods

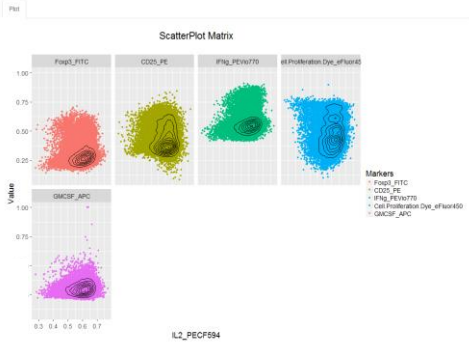
Functionalities

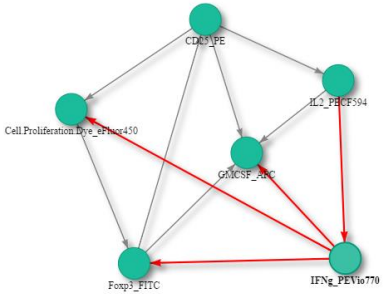
Visualization: Histograms, Scatter/Density/Violin plots, Network graphs

Univariate Analysis : Population Comparison, Regression

Network Reconstruction Analysis

- (Conditional) Association Networks (COR, MMPC)
- Probabilistic Causal Networks (PC, FCI, IDA)
- Bayesian Networks (HC)
- *Currently available methods*: MMPC, PC, HC, FCI, IDA, COR



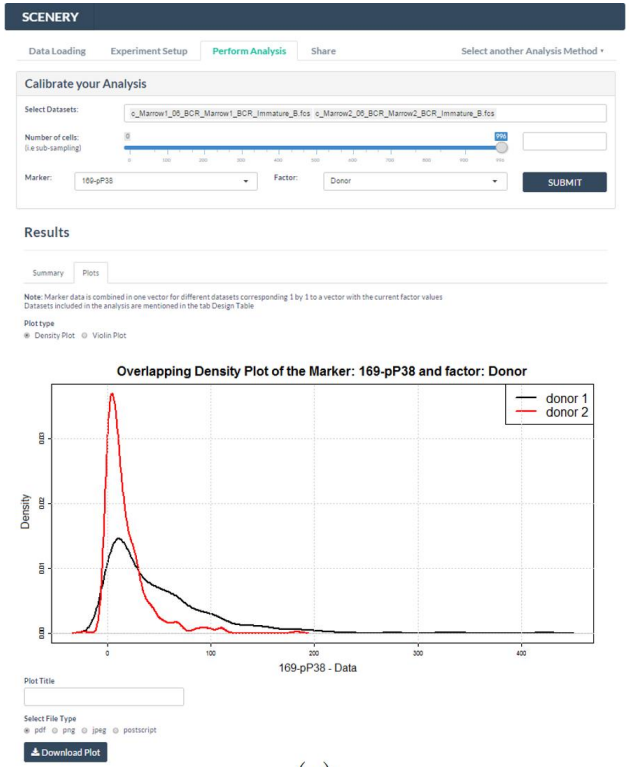


USE CASE

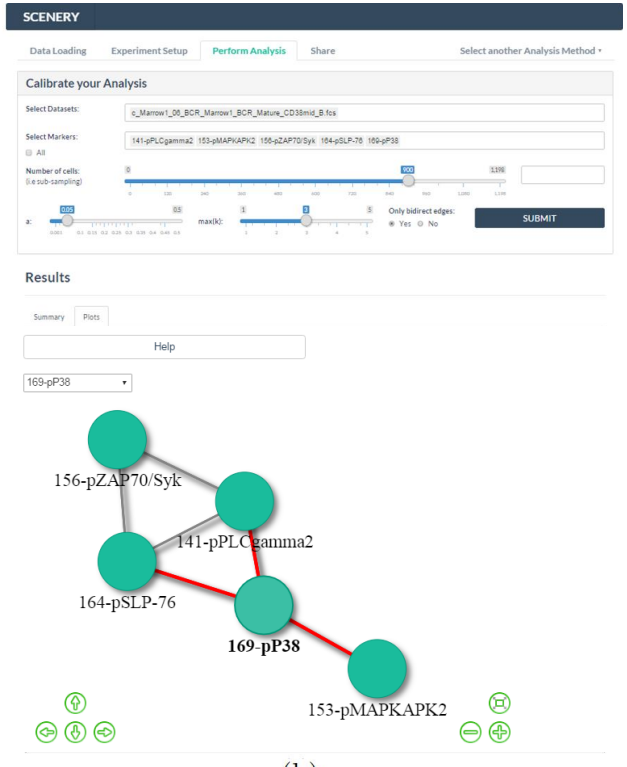
Data: Bendall et al., *Science*, 2011

(a) Overlapping density plots for the marker p38 on 2 donors.

(b) Reconstructed network (MMPC) on selected protein markers: SYK, BLNK, PLC2, p38 and MAPKAPK2.



(a)



(b)

G. Athineou, G. Papoutsoglou, S. Triantafullou, I. Basdekis, V. Lagani, I. Tsamardinos (2016): *SCENERY: a Web-Based Application for Network Reconstruction and Visualization of Cytometry Data*, PACBB 2016.

REFERENCES

1. D. Margaritis and F. Bromberg, **Efficient Markov Network Discovery Using Particle Filters**, *Computational Intelligence (2009).*
2. S. Triantafillou, I Tsamardinos and IG Tollis**, Learning Causal Structure from Overlapping Variable Sets**, *AISTATS 2010*.
3. G. Borboudakis, S. Triantafillou, V. Lagani, I. Tsamardinos, **A Constraint-based Approach to Incorporating Prior Knowledge in Causal Models**, *ESANN 2011.*
4. T. Claassen and T. Heskes, **A Logical Characterization of Constraint-based Causal Discovery**. *UAI 2011.*
5. T. Claassen and T. Heskes, **A Bayesian Approach to Constraint-based Causal Inference**, *UAI 2012*.
6. I. Tsamardinos, S. Triantafillou, V. Lagani, **Towards Integrative Causal Analysis of Heterogeneous Data Sets and Studies,** *Journal of Machine Learning Research (2012).*
7. A. Hyttinen, P. O. Hoyer, F. Eberhardt, and M. Järvisalo, **Discovering Cyclic Causal Models with Latent Variables: A General SAT-Based Procedure ,** *UAI 2013.*

REFERENCES(2)

8. A. Hyttinen, F. Eberhardt, and M. Järvisalo, **Constraint-based Causal Discovery: Conflict Resolution with Answer Set Programming,** *UAI 2014.*
9. S. Triantafillou, I. Tsamardinos, A. Roumpelaki, **Learning Neighborhoods of High Confidence in Constraint-Based Causal Discovery**, *PGM 2014.*
10. S. Triantafillou and I. Tsamardinos*,* **Constraint-based Causal Discovery from Multiple Interventions over Overlapping Variable Sets**, *Journal of Machine Learning Research (2015).*
11. D. Sonntag, M. Järvisalo, Jose M. Pena, A. Hyttinen, **Learning Optimal Chain Graphs with Answer Set Programming,** *UAI 2015.*
12. A. Hyttinen, S. Plis, M. Järvisalo, F. Eberhardt, and D. Danks, **Causal Discovery from Subsampled Time Series Data by Constraint Optimization,** *submitted***.**
13. G. Borboudakis and I. Tsamardinos. **Towards Robust and Versatile Causal Discovery for Business Applications**. *KDD 2016*.
14. S. Magliacane, T. Claassen, J.M. Mooij, **Ancestral Causal Inference**, *arXiv:1606.07035*