https://gilleskratzer.netlify.com/

http://www.r-bayesian-networks.org/

University of Zurich[UZH]

GILLES KRATZER, APPLIED STATISTICS GROUP, UZH

CAUSALITY WORKSHOP, UZH 14.12.2018

BAYESIAN NETWORKS MEET OBSERVATIONAL DATA

gilles.kratzer@math.uzh.ch

MOTIVATIONAL EXAMPLE: CREDIT CARD FRAUD DETECTION PREDICTION

Credit Card Fraud Detection Using Bayesian and Neural Networks

Sam Maes Karl Tuyls Bram Vanschoenwinkel
Bernard Manderick
Vrije Universiteit Brussel - Department of Computer Science
Computational Modeling Lab (COMO)
Pleinlaan 2
B-1050 Brussel, Belgium
{sammaes@,ktuyls@,bvschoen@,bernard@arti.}vub.ac.be

Abstract

This paper discusses automated credit card fraud detection by means of machine learning. In an era of digitalization, credit card fraud detection is of great importance to financial institutions. We apply two machine learning techniques suited for reasoning under uncertainty: artificial neural networks and

do the fraud detection. After a process of learning, the program is supposed to be able to correctly classify a transaction it has never seen before as fraudulent or not fraudulent, given some features of that transaction.

The structure of this paper is as follows: first we introduce the reader to the domain of credit card fraud detection. In Sections 3 and 4 we briefly ex-

MOTIVATIONAL EXAMPLE: CREDIT CARD FRAUD DETECTION PREDICTION

Credit Card Fraud Detection Using Bayesian and Neural Networks

Sam Maes Karl Tuyls Bram Vanschoenwinkel

experiment	±10% false pos	±15% false pos
ANN-fig 2(a)	60% true pos	70% true pos
ANN-fig 2(a)	47% true pos	58% true pos
ANN-fig 2(c)	60% true pos	70% true pos
BBN-fig 2(c)	68% truc pos	74% truc pos
BBN-fig 2(g)	68% true pos	74% true pos

Table 1: This table compares the results achieved with ANN and BBN, for a false positive rate of respectively 10% and 15%.

Abstract

This paper discusses
tection by means of
of digitalization, cre
great importance to
two machine learning techniques suited for reasoning under uncertainty: artificial neural networks and

rocess of learning,
e to correctly clas-
n before as fraud-
ie features of that

s follows: first we introduce the reader to the domain of credit card fraud detection. In Sections 3 and 4 we briefly ex-

University of Zurich[UZH]

Contents lists available at SciVerse ScienceDirect

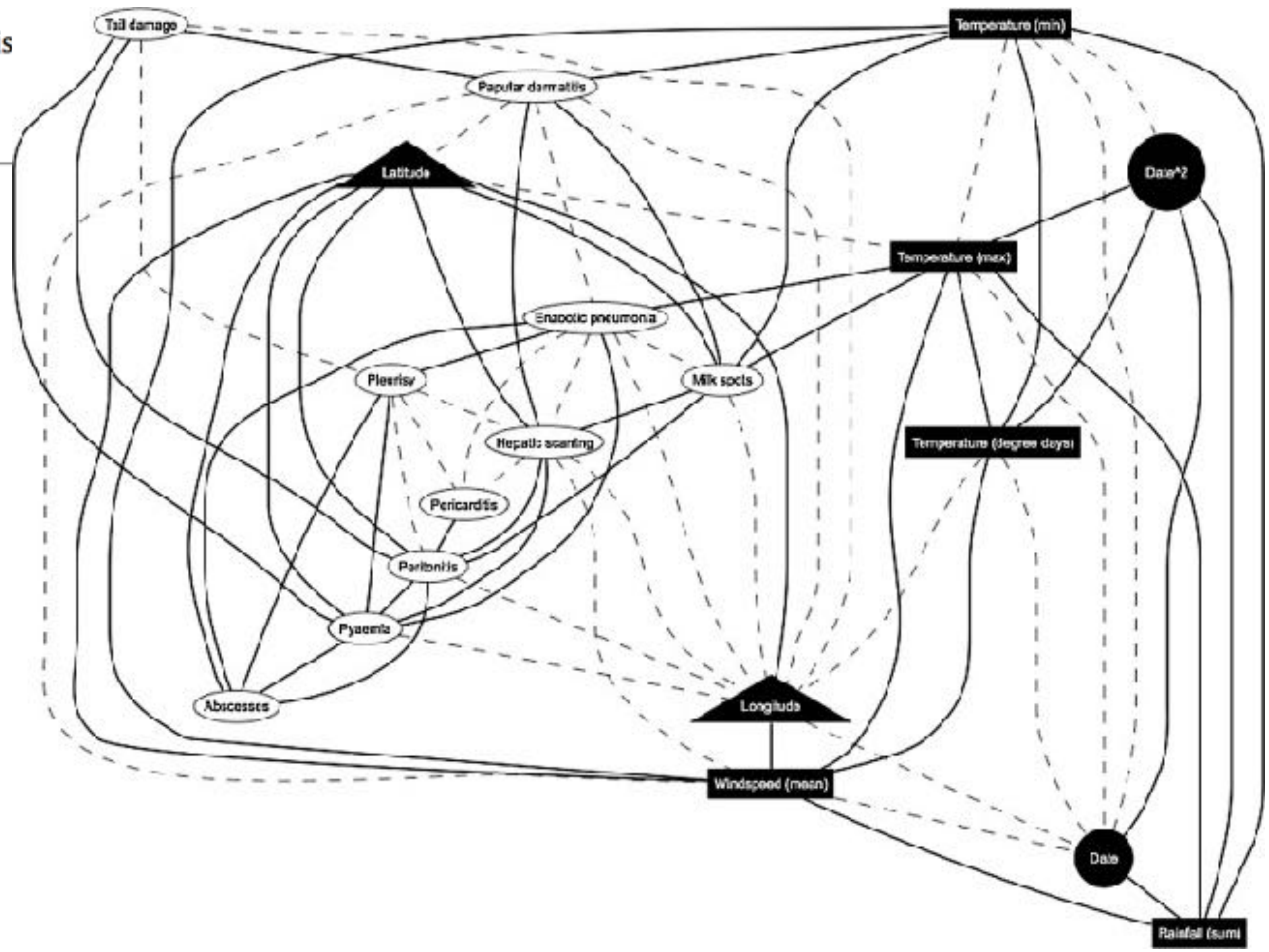
Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Using Bayesian networks to explore the role of weather as a potential determinant of disease in pigs

B.J.J. McCormick[a], M.J. Sanchez-Vazquez[b], F.I. Lewis

[a] Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA
[b] OIE Organisation Mondiale de la Santé Animale, 12, rue de Prony, 75017 Paris, France
[c] Section of Epidemiology, University of Zurich, Zurich, Switzerland



Discovering complex interrelationships between socioeconomic status and health in Europe: A case study applying Bayesian Networks

Javier Alvarez-Galvez [a, b, *]

[a] Loyola University Andalusia, Department of International Studies, Campus de Palmas Altas, Faculty of Political Sciences and Law, Seville 41014, Spain
[b] Complutense University of Madrid, Department of Sociology IV (Research Methodology and Communication Theory), Campus de Somosaguas, Faculty of Political

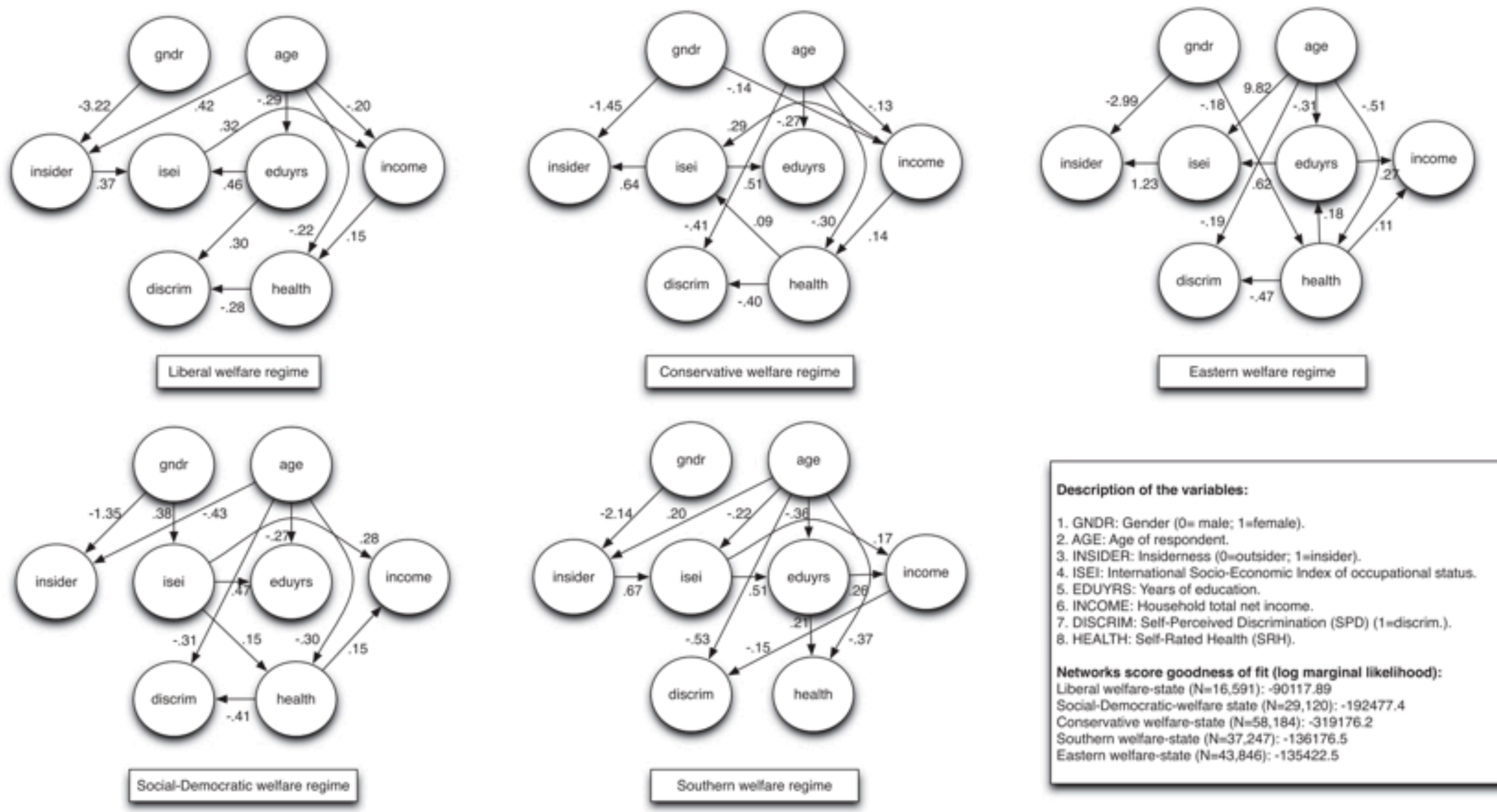


Fig. 1. Bayesian networks describing interrelationships between SES and health in five European welfare states.

BAYESIAN NETWORKS IN THE MACHINE LEARNING WORLD

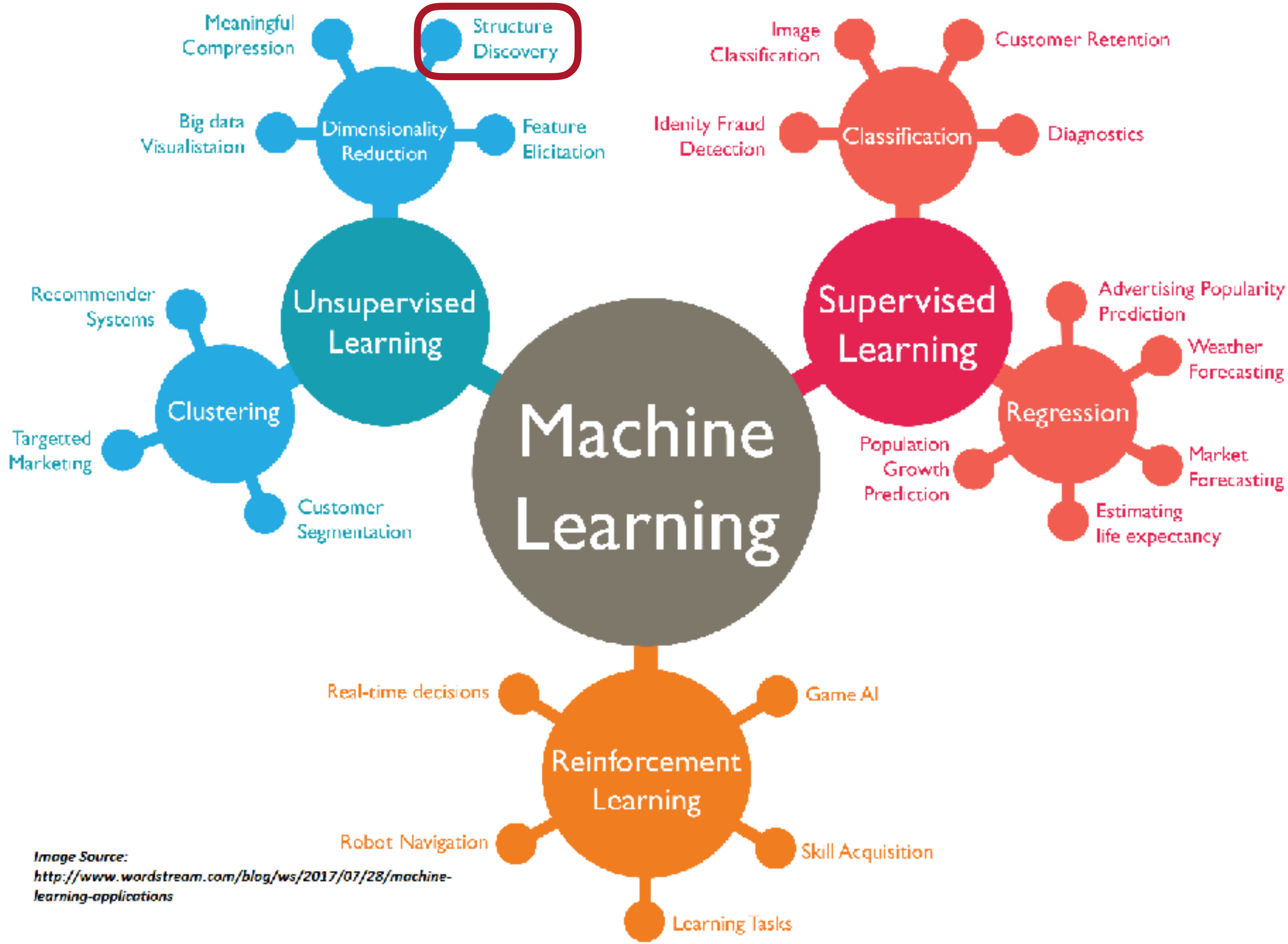


Image Source: http://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Objectif of the talk:

How to learn Bayesian networks from observational data?

Objectif of the talk:

How to ~~learn~~ select Bayesian networks from observational data?

Bayesian Networks are defined by two elements:

Network structure:

Directed Acyclic Graph (DAG): G = (V, A)

in which each node $v_i \in V$ corresponds to a random variable X_i

Probability distribution:

Probability distribution X with parameters Θ, which can be factorised into smaller local probability distributions according to the arcs $a_{ij} \in A$ present in the graph.

A BN encodes the factorisation of the joint distribution

$$P(\mathbf{X}) = \prod_{j=1}^{n} P(X_j \mid \mathbf{Pa}_j, \Theta_j), \text{ where } \mathbf{Pa}_j \text{ is the set of parents of } X_j$$

Objectif of the talk:

~~How to learn Bayesian networks from observational data?~~

Which approaches do exist?

Which assumptions/limitations are involved when learning a Bayesian network form observational dataset?

Theoretical limitations:

- BN learning is **ill-posed on two levels**
  - Finite sample (any stats problem is ill-posed)
  - Complete knowledge of observational distribution usually does not determine the underlying causal model

Objectif of the talk:

~~How to learn Bayesian networks from observational data?~~

Which approaches do exist?

Which assumptions/limitations are involved when learning a Bayesian network form observational dataset?

Technical limitations:

- Approximate learning process
- Proxies
- Combinatorial wall!!!
  - Simplification needed

COMBINATORIAL WALL

# Nodes	# DAGs	Inference
1 - 15 Nodes	$< 10^{41}$ DAGs	Exact inference
16 - 25 Nodes	$< 10^{100}$ DAGs	Exact inference possible
26 - 50 Nodes	$< 10^{400}$ DAGs	Approximate inference
51 - 100 Nodes	$< 10^{1700}$ DAGs	Approximate inference
101 - 1000 Nodes	$< 10^{100000}$ DAGs	(very) approximative inference

Typical domain of interest

Approximations:

- limiting number of parents per node
- Decomposable scores/efficient algorithm
- Score equivalence

University of Zurich[UZH]

PLAN

1. From observationnal dataset deduce probabilistic model
    - Usually discrete BN or jointly Gaussian
    - Epidemiological constrain: mixture of distributions
2. From probabilistic model deduce structure

EXPONENTIAL FAMILY

Observational dataset

X1	X2	X3	…
12	23	53	…
32	31	23	…
10	16	45	…
…	…	…	…

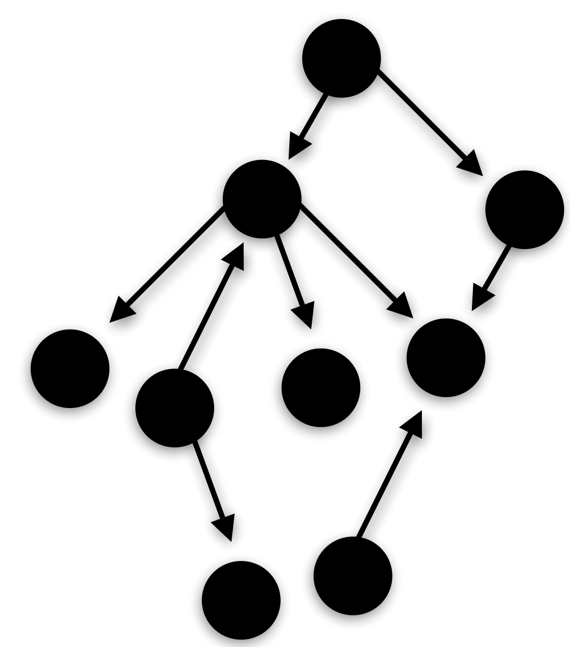
1

Probabilistic model

$$P(X_1, \ldots, X_n) = P(X_i | X_j, \ldots) \ldots$$

2

Network structure



Independance testing

Computing directly

SOME ELEMENTS OF PROBABILITY THEORY

The conditional probability of A given B is: $$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Bayes theorem: $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Let A, B and C non intersecting subsets of nodes in a DAG G

A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

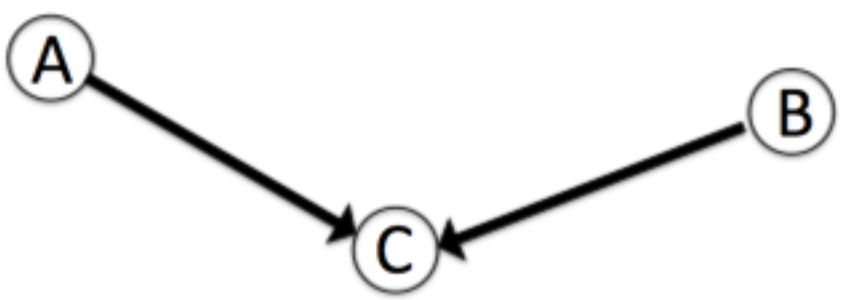
ELEMENT OF GRAPH THEORY

Let A, B and C non intersecting subsets of nodes in a DAG G

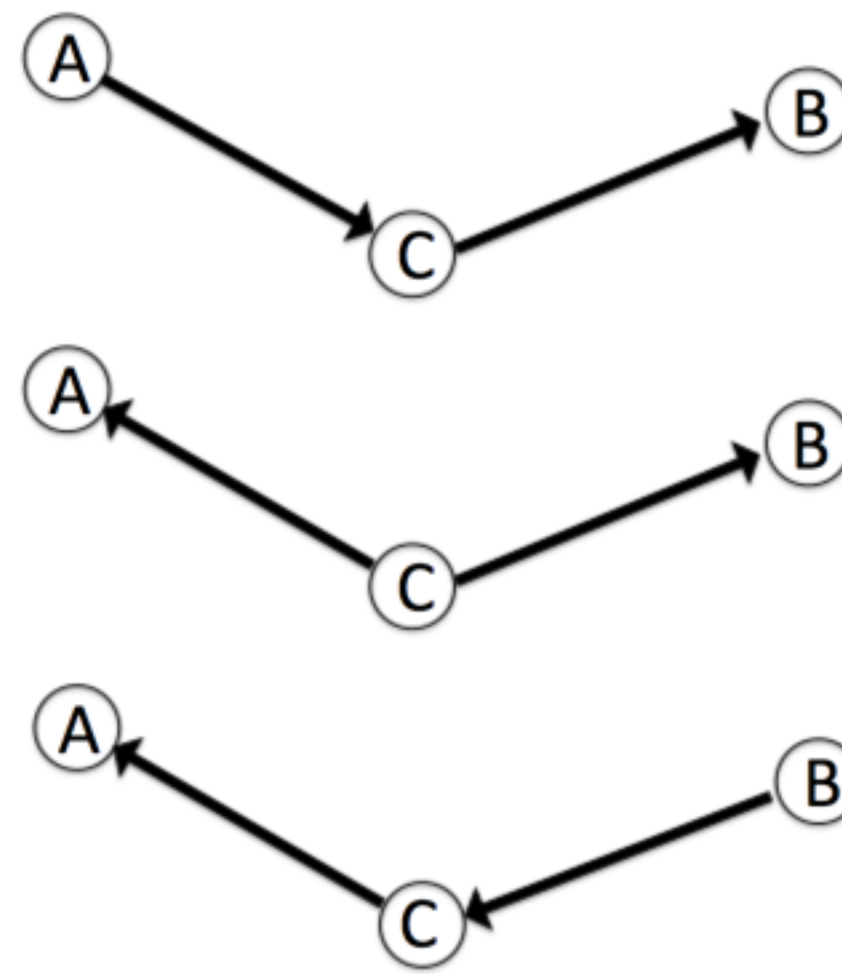
A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

$$A \not\perp\!\!\!\perp_P B|C$$



$$A \perp\!\!\!\perp_P B|C$$

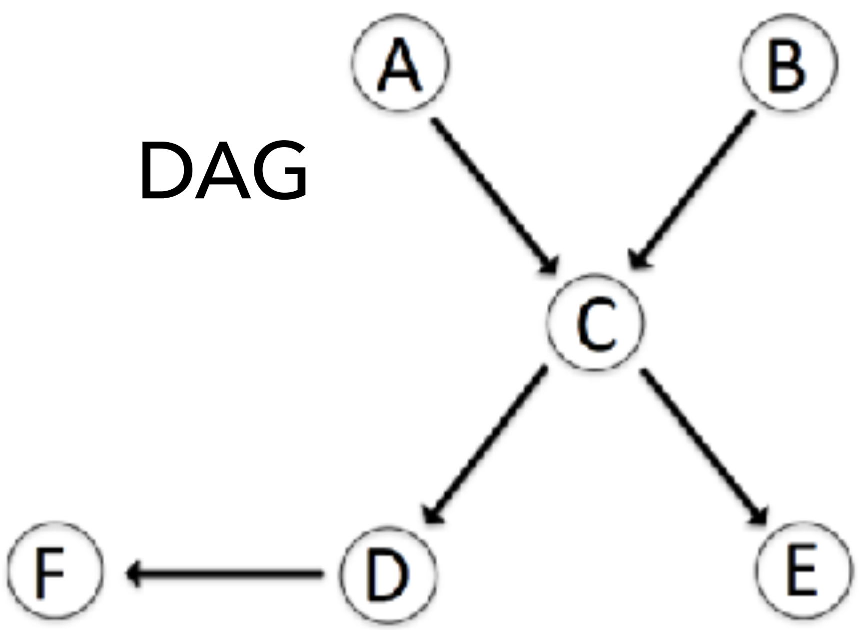


LEARNING BAYESIAN NETWORKS

- In a practical perspective, for observational data, if learning algorithms rely on probabilistic learning algorithm. Then one can learn up to the Markov equivalence class.

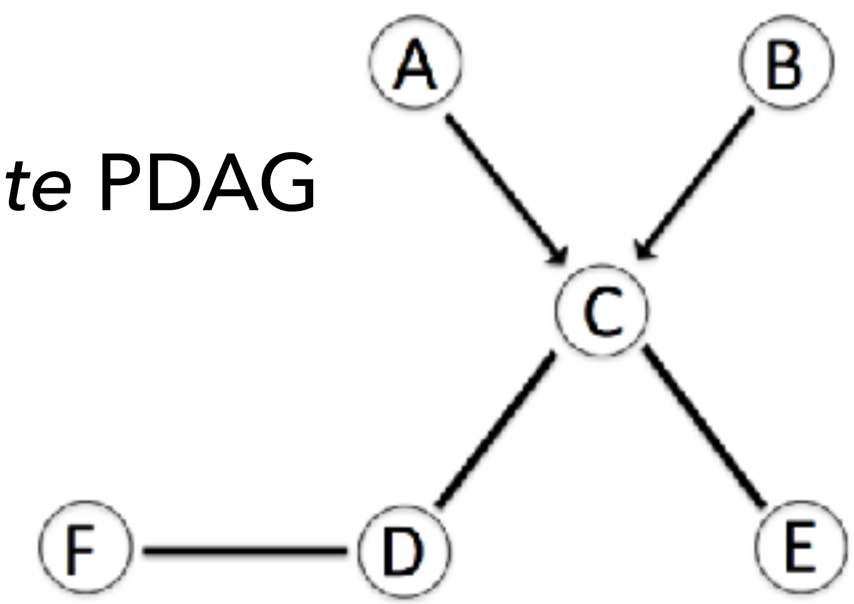
- Markov equivalence class are the set of DAGs that have the same skeleton and v-structure.

DAG



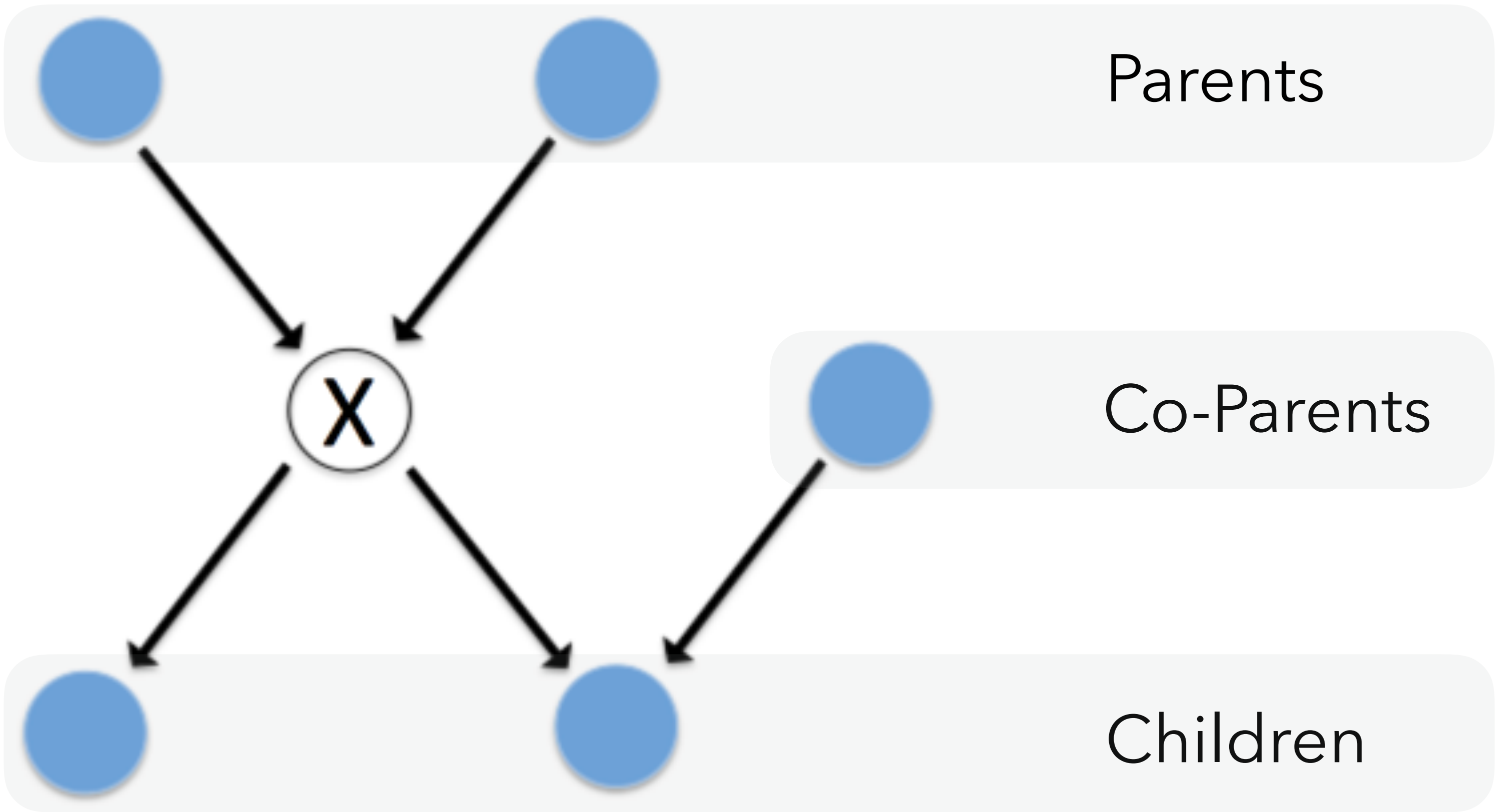
=

complete PDAG



ELEMENT OF GRAPH THEORY: MARKOV BLANKET

The Markov Blanket of a node is the set of parents, co-parents and children.



$$P(X_k \mid X_n, k \neq n) = P(X_k \mid X_{\mathrm{MB}(k)}), \forall k$$

The Markov Blanket of a node is the set of nodes that shields the index node from the rest of the network

Local Markov property:

$$X \perp \text{ Non-Descendants(X) } | Pa(X)$$

$$\mathcal{M} = (\mathcal{S}, \Theta_{\mathcal{M}})$$

Model selection

Structure learning

Parameter estimation

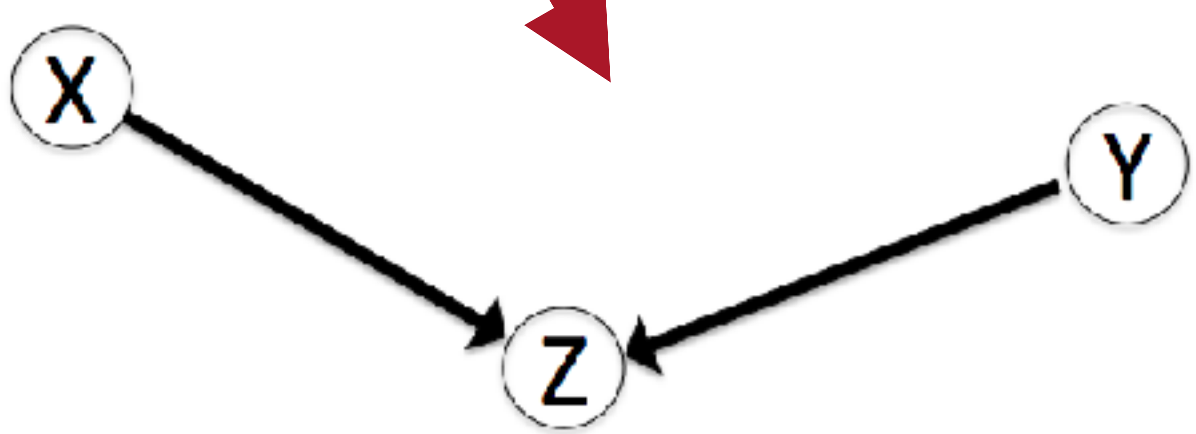
Parameter learning

$$P(\mathcal{M}|\mathcal{D}) = \underbrace{P(\Theta_{\mathcal{M}}, \mathcal{S}|\mathcal{D})}_{\text{model learning}} = \underbrace{P(\Theta_{\mathcal{M}}|\mathcal{S}, \mathcal{D})}_{\text{parameter learning}} \cdot \underbrace{P(\mathcal{S}|\mathcal{D})}_{\text{structure learning}}$$

Constraint based algorithms

$$P_{X \perp\!\!\!\perp Y | Z} < \alpha$$

$$X \perp\!\!\!\perp_{\mathcal{S}} Y | Z = X \perp Y | Z$$



Search-and-score algorithms

Maximum a posteriori score

$$G^* = \underset{G}{\operatorname{argmax}} \; f(\mathcal{D}, G, n, \dots)$$

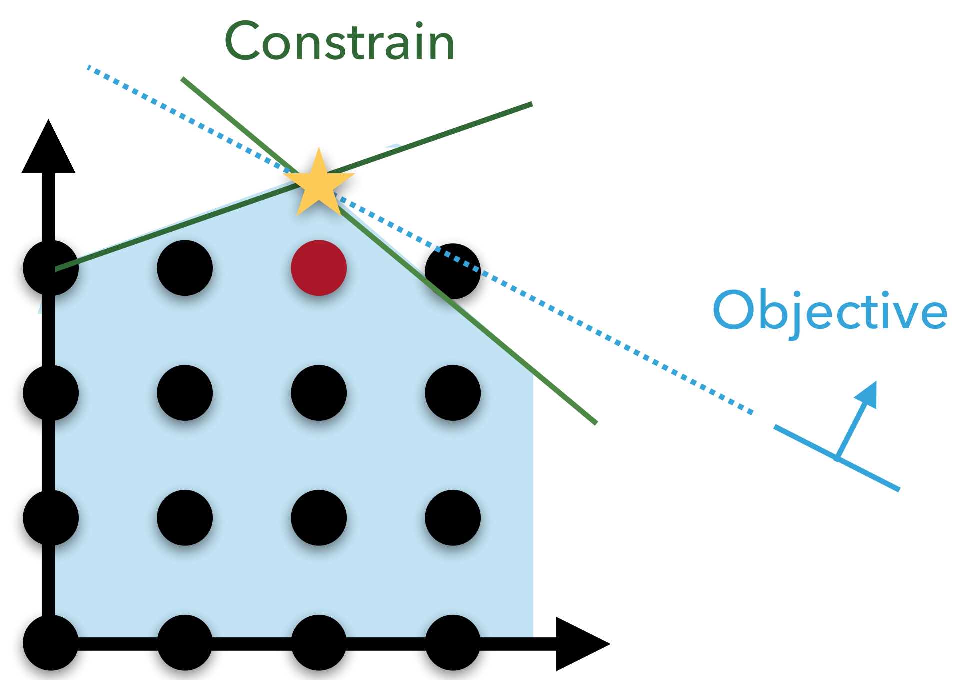
Example of scoring functions:

- Bayesian or ML scores
  - Bayesian Posterior
  - Bayesian-Dirichlet (BDeu,BDs,BDe)
  - Bayesian Information Criterion (BIC)

University of Zurich[UZH]

Score-and-search algorithms

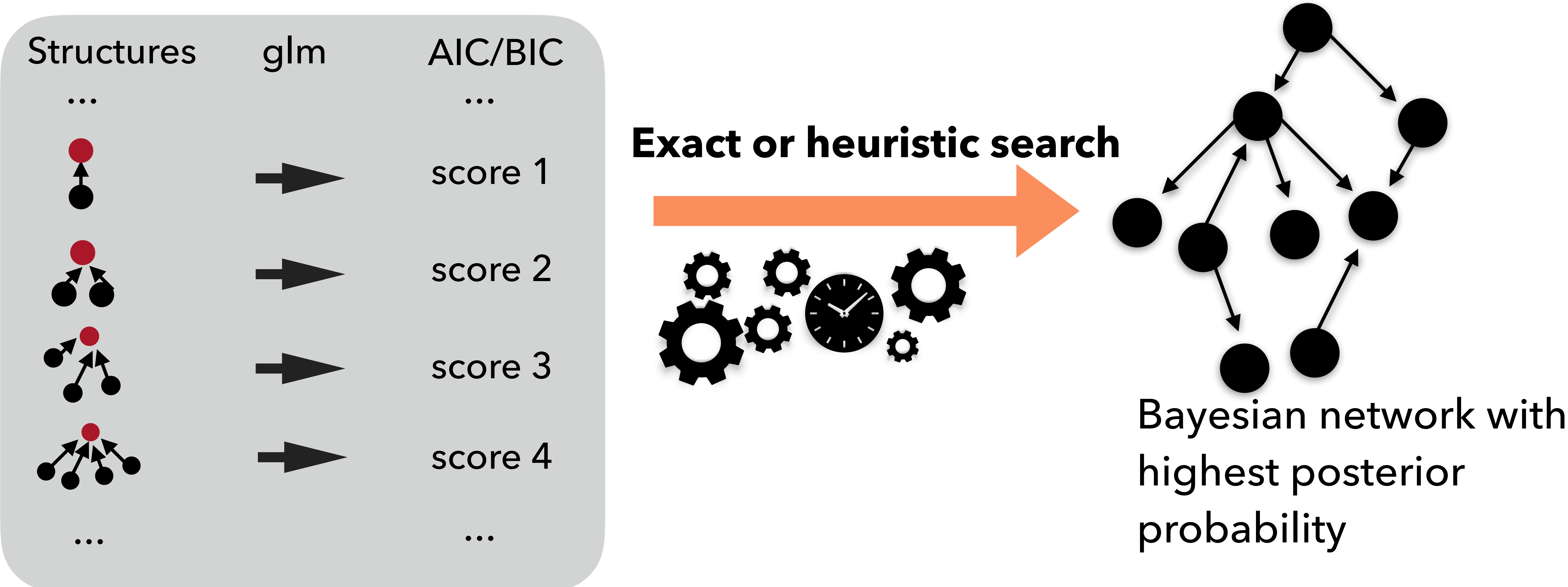
- *Heuristic approaches / Greedy search*
  - Hill-climbing (with possibly random restarts/stochastics …)
  - Tabu search (Glover, 1986)
  - Simulated annealing (Kirkpatrick et al, 1983)
  - Plus an entire zoo of methods …
- *Exact search*
  - Exact node ordering (Koivisto et al. , 2004)
  - Learning with cutting planes (Cussens, 2012)



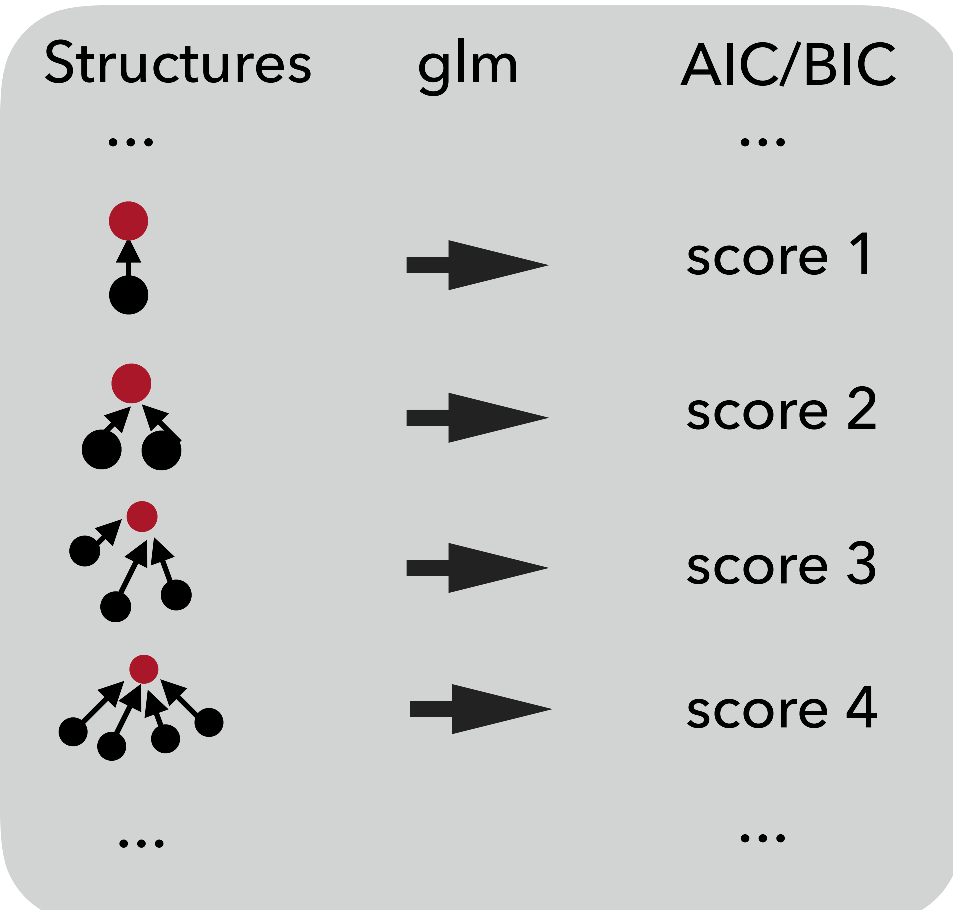
Scores

- Decomposability!
- Discrete BNs:
  - Bayesian-Dirichlet: BDeu (Heckerman et al. ,1995)
- Score equivalence for additive regression framework:
  - Bayesian based scores: not always score equivalent due to the prior!
  - Information theoretic scores: BIC asymptotically score equivalent

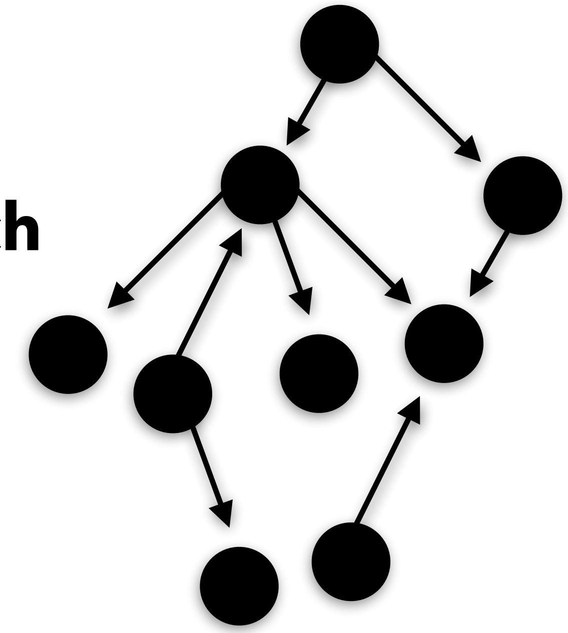
Search and score algorithm



Search and score algorithm



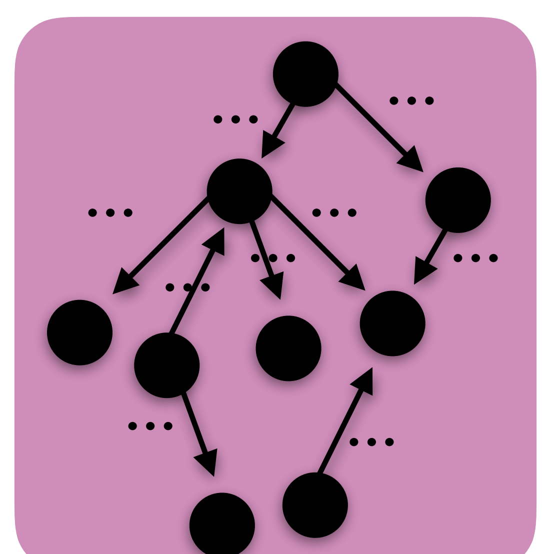
Exact or heuristic search



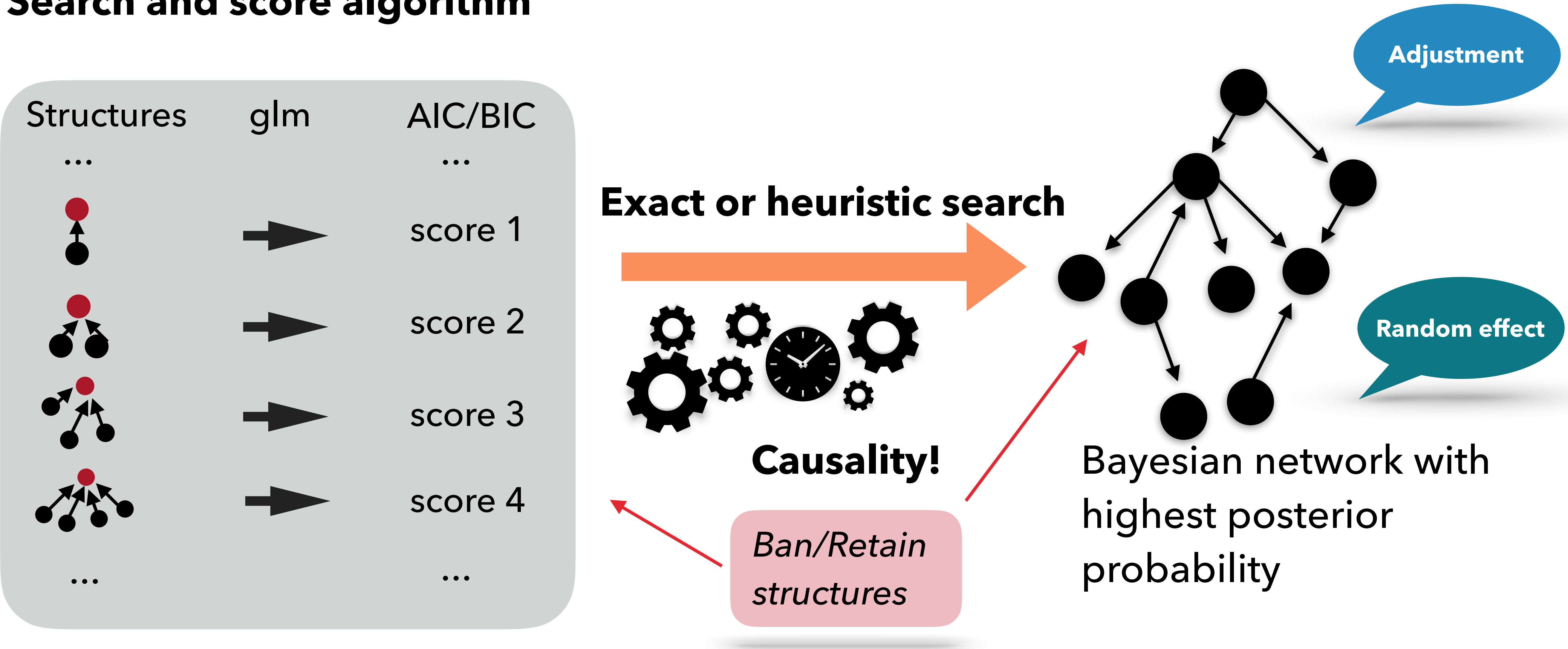
Bayesian network with highest posterior probability

Parameter estimation

- compute marginal posterior density
- regression estimate

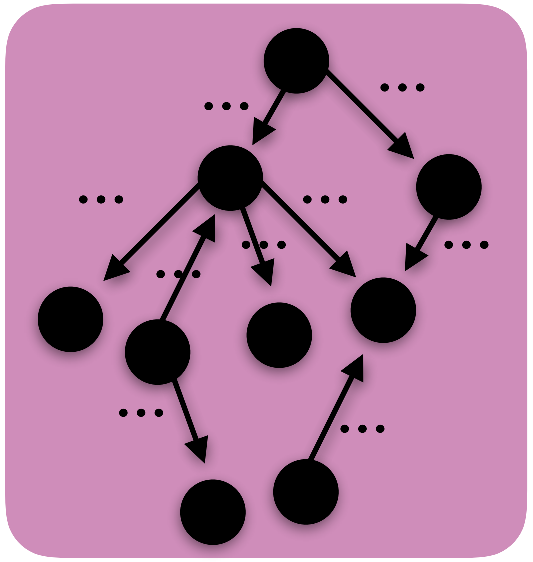


Search and score algorithm



Parameter estimation

- compute marginal posterior density
- regression estimate



Using R

buildscorecache()

mostprobable()

fitabn()

CAUSAL THINKING VERSUS ACAUSAL THINKING

- Strong assumptions … but common in statistics, no?
- *“It seems that if conditional independence judgements are byproducts of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal Bayesian networks.”* (Pearl, 2009)
- The do-calculus
  - Interventions
  - In epidemiology: Randomised Controlled Trial
- So … BN is a nice framework to treat causal and acausal thinking

R CODE: SOFTWARE IMPLEMENTATION

Popular R packages (available on CRAN)

bnlearn

- Learning via constraint-based and score-based algorithms (many!)

pcalg

- Robust estimation of CPDAG via the PC-Algorithm

deal

- Learning BNs with mixed (discrete and continuous) variables

catnet

- Discrete BNs using likelihood-based criteria

abn

- Learning BNs with mixed (discrete, continuous, Poisson) variables
- Score based methods: Bayesian and frequentist estimation
- Exact and heuristic search
- Link strength

Disclaimer: I am author and maintainer of the abn R package. I will use it for the example part.

University of Zurich[UZH]

System epidemiology

- Typically the set of possible variables is formidable
  - The classical approach for variable selection is based on prior scientific knowledge (29%)[1]
  - Change of estimate (18%)[1]
  - Stepwise model selection (16%)[1]

No prior model?

Not one outcome experiment?

varrank Variable ranking for better time allocation

- Variable ranking based on a set of variable of importance
- Model free. Based on information theory metrics
- Mixture of variables (continuous and discrete). Discretisation through rule/clustering

https://CRAN.R-project.org/package=varrank

[1] *Walter et al (2009)*

f_i candidate feature to be ranked

C set of variables of importance

S set of already selected variables

$$\mathrm{H}(X) = \sum_{n=1}^{N} P(x_n) \log P(x_n)$$

Average amount of information of one RV

$$\mathrm{MI}(X;Y) = \sum_{n=1}^{N} \sum_{m=1}^{M} P(x_n; y_m) \log \frac{P(x_n; y_m)}{P(x_n)P(y_m)}$$

Mutual dependence between two RV

Difference (mid) or quotient (miq)

Greedy search

Forward - argmax

$$\mathrm{score}_i = \underbrace{\mathrm{MI}(f_i; \mathbf{C})}_{\text{Relevance}} - \beta \sum_{\mathbf{S}} \underbrace{\alpha(f_i, f_s, \mathbf{C})}_{\text{Normalization}} \underbrace{\mathrm{MI}(f_i; f_s)}_{\text{Redundancy}}$$

Backward - argmax

Discretization

Estévez and al. (2009)

$$\beta = 1/|\mathbf{S}| \text{ and } \alpha(f_i, f_s, \mathbf{C}) = \frac{1}{\min(\mathrm{H}(f_i), \mathrm{H}(f_s))}$$

R CODE: EXAMPLE ASIA = SYNTHETIC DATASET

Proposed by Lauritzen et al.,1988 and provided by Scutari, 2009

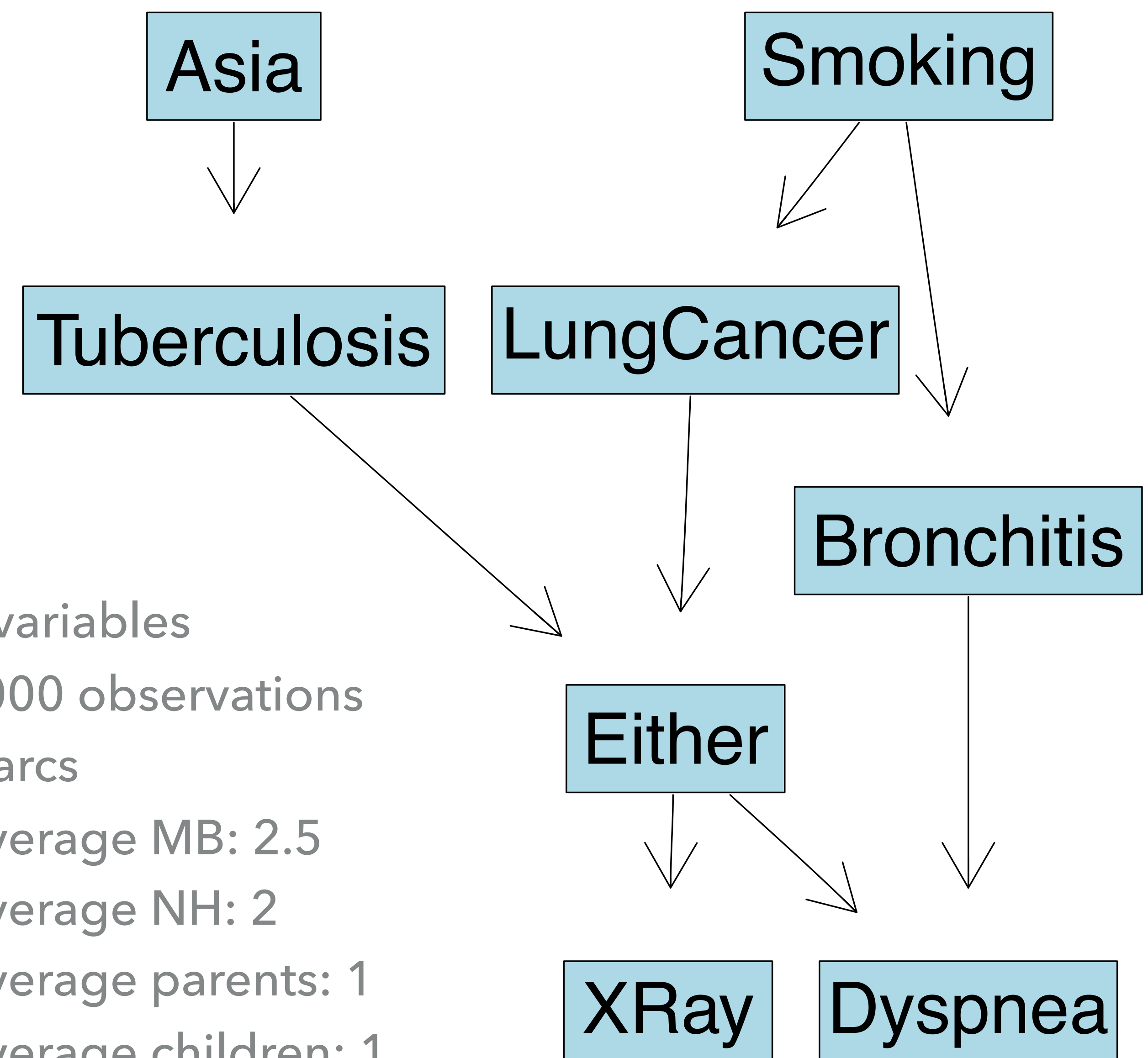
"Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea."

https://CRAN.R-project.org/package=abn

Proposed by Lauritzen et al.,1988 and provided by Scutari, 2009

"Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea."

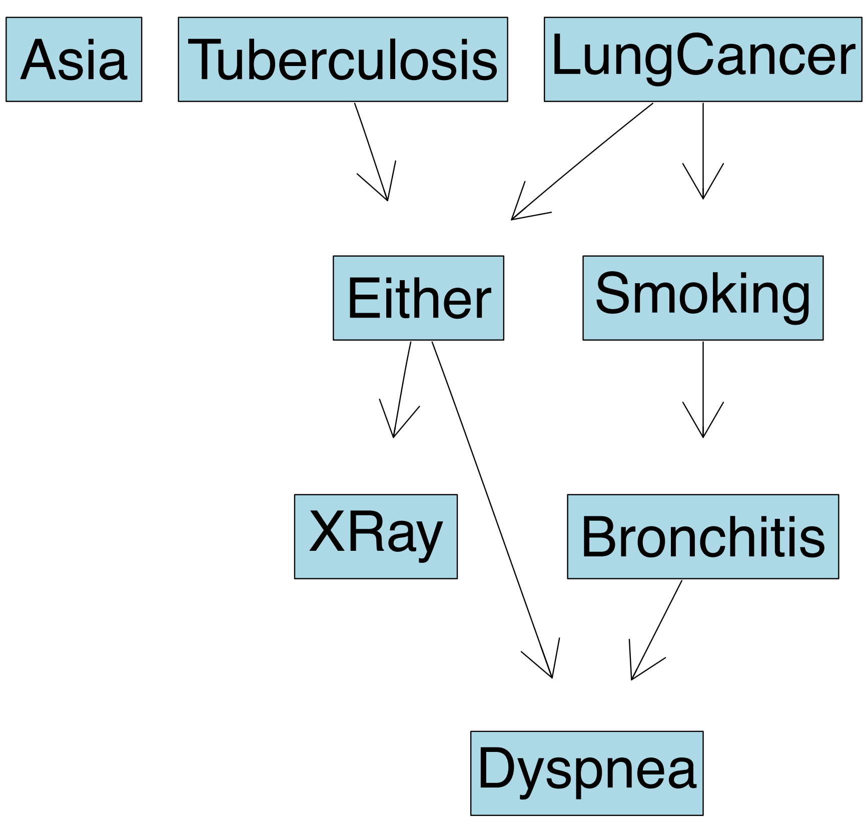
```
##defining distributions
dist = list(Asia = "binomial",
            Smoking = "binomial",
            Tuberculosis = "binomial",
            LungCancer = "binomial",
            Bronchitis = "binomial",
            Either = "binomial",
            XRay = "binomial",
            Dyspnea = "binomial")

#plot BN
plotabn(dag.m = ~Asia|Tuberculosis +
            Tuberculosis|Either +
            Either|XRay:Dyspnea +
            Smoking|Bronchitis:LungCancer +
            LungCancer|Either +
            Bronchitis|Dyspnea,
          data.dists = dist,
          edgedir = "cp",
          fontsize.node = 30,
          edge.arrowwise = 3)
```



8 variables
5000 observations
8 arcs
Average MB: 2.5
Average NH: 2
Average parents: 1
Average children: 1

ASIA: SCORE BASED ALGORITHM

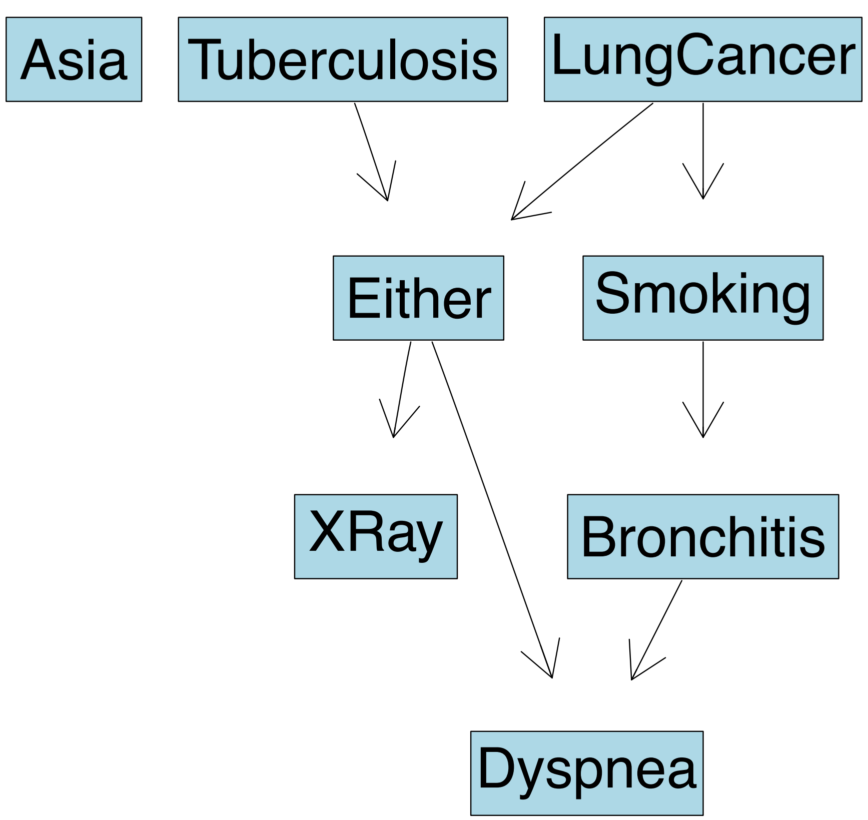
```r
##============================
##score based algorithm
##============================

#loglikelihood score
bsc.compute <- buildscorecache(data.df = asia,
                               data.dists = dist,
                               max.parents = 2)

dag <- mostprobable(score.cache = bsc.compute)
plotabn(dag.m = dag,data.dists = dist, fontsize.node = 30, edge.arrowwise = 3)
```



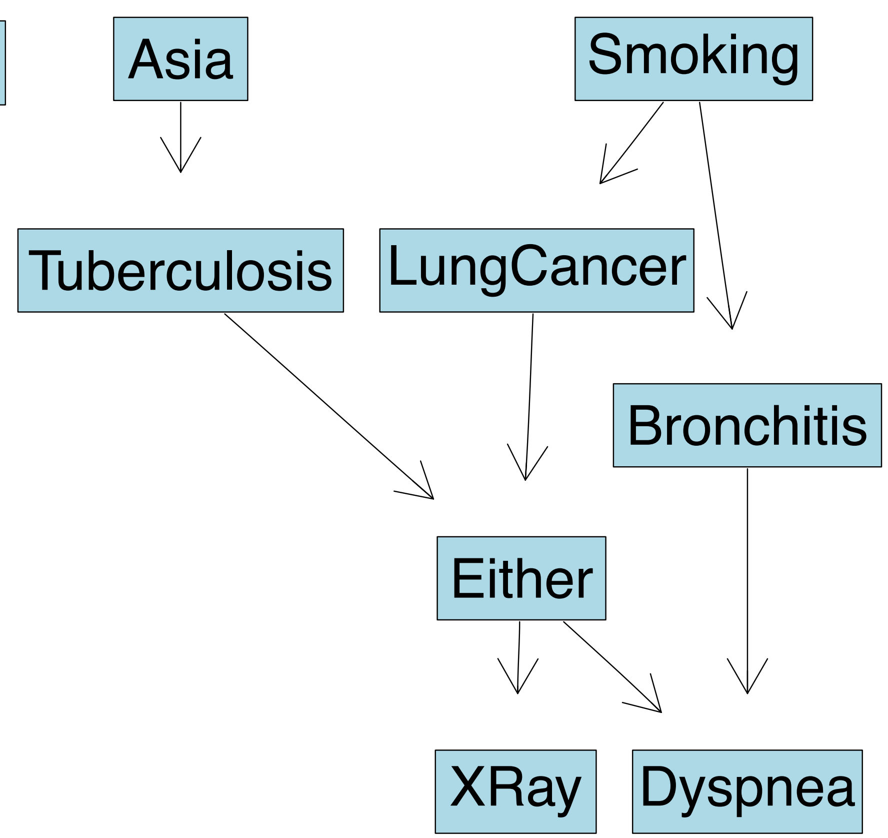
ASIA: SCORE BASED ALGORITHM

```
##=============================
##score based algorithm
##=============================

#loglikelihood score
bsc.compute <- buildscorecache(data.df = asia,
                                 data.dists = dist,
                                 max.parents = 2)

dag <- mostprobable(score.cache = bsc.compute)
plotabn(dag.m = dag,data.dists = dist, fontsize.node = 30, edge.arrowwise = 3)
```

Learned



Truth

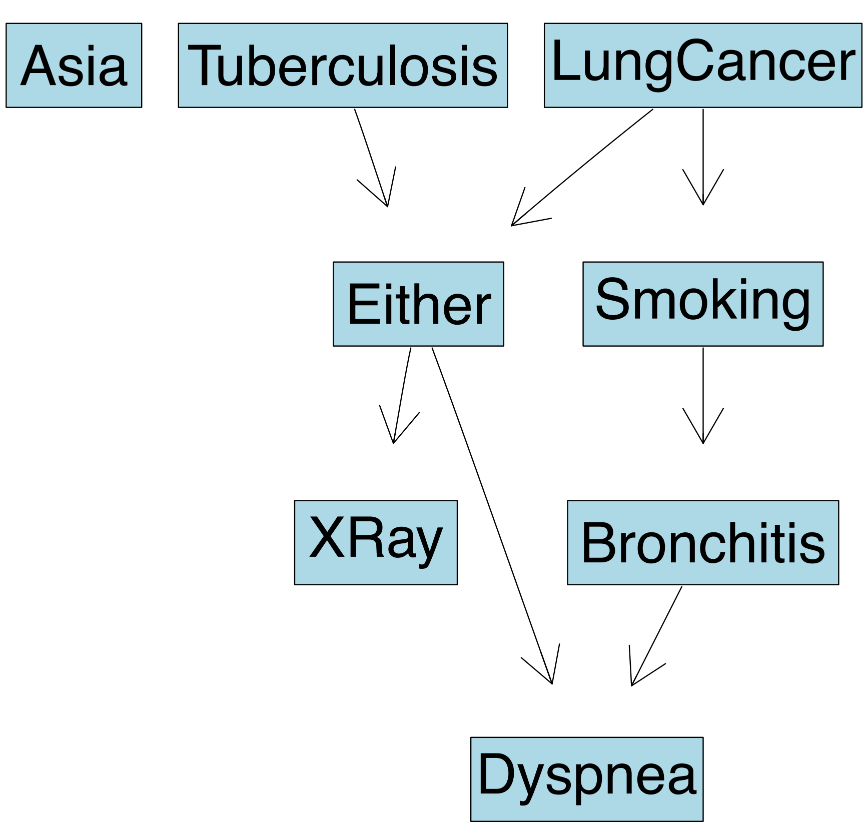


ASIA: SCORE BASED ALGORITHM

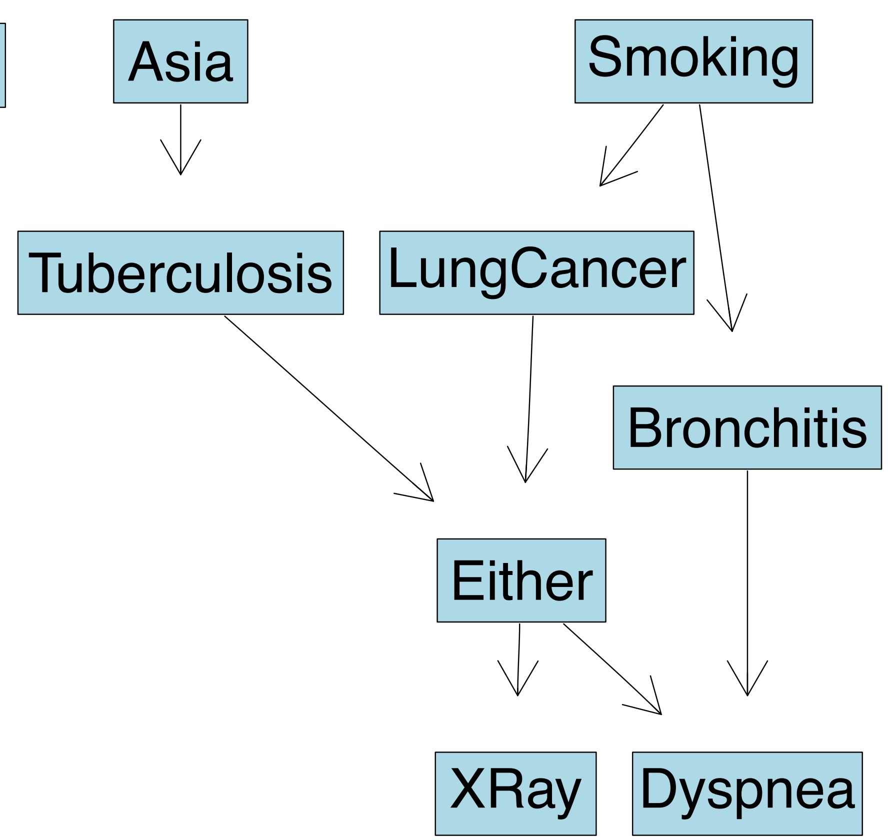
```
##===========================
##score based algorithm
##===========================

#loglikelihood score
bsc.compute <- buildscorecache(data.df = asia,
                               data.dists = dist,
                               max.parents = 2)

dag <- mostprobable(score.cache = bsc.compute)
plotabn(dag.m = dag,data.dists = dist, fontsize.node = 30, edge.arro
```

```
> compareDag(ref = t(dag.adj),
+            test = dag)
$TPR
[1] 0.75

$FPR
[1] 0.01785714

$Accuracy
[1] 0.953125

$FDR
[1] 0.2857143

$`G-measure`
[1] 0.8017837

$`F1-score`
[1] 44.8

$PPV
[1] 0.8571429

$FOR
[1] 0.2857143

$`Hamming-distance`
[1] 3
```

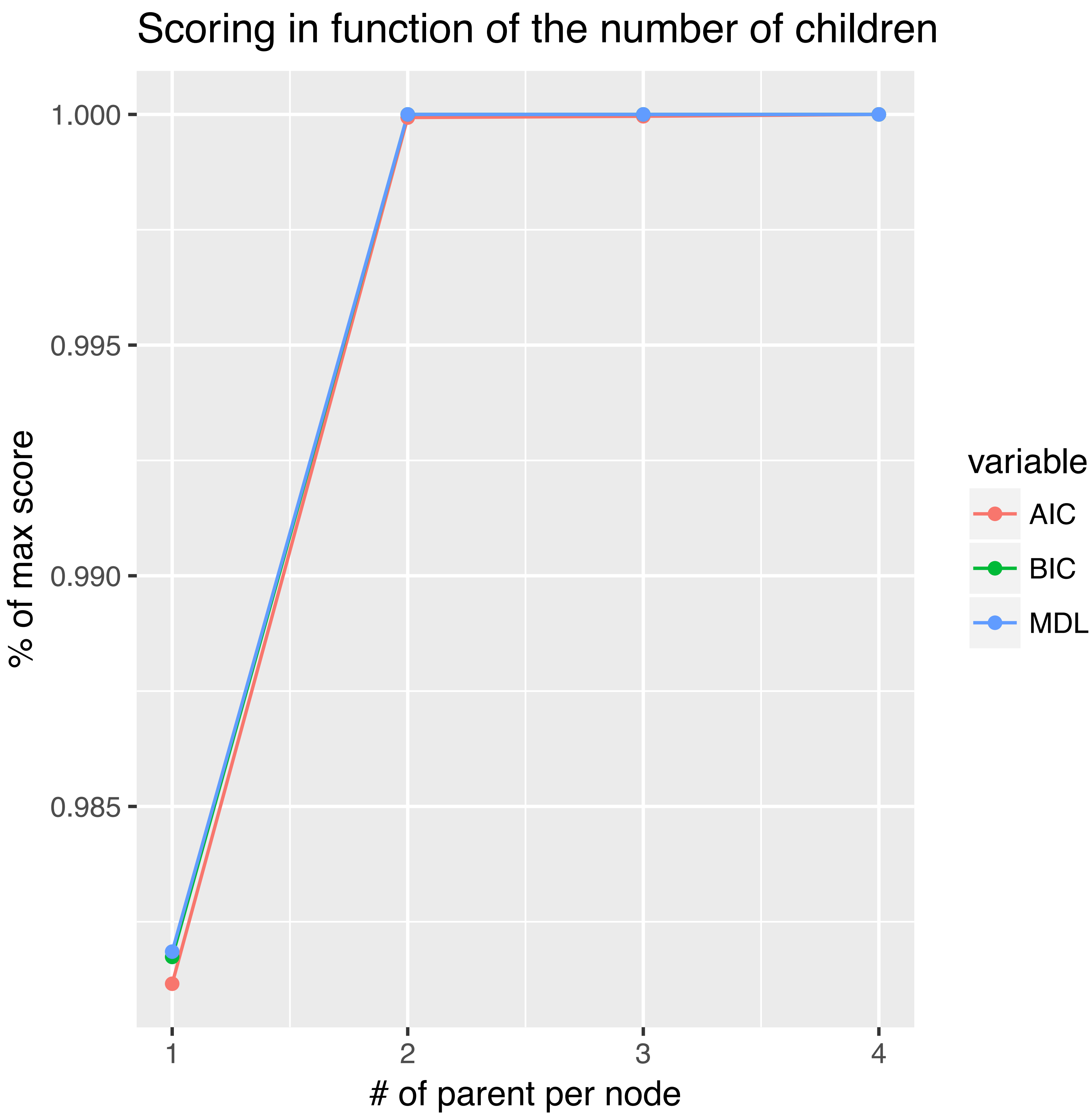
Learned



Truth

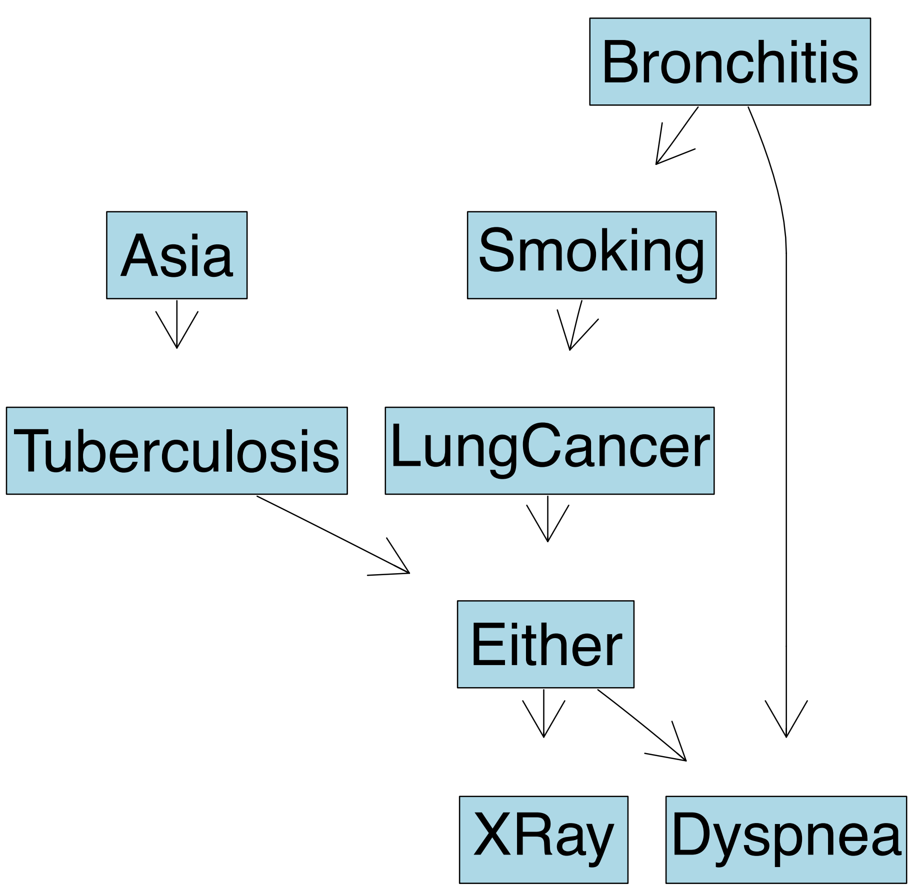


ASIA: HOW MANY PARENT ARE NEEDED?



ASIA: EXTERNAL KNOWLEDGE

```
##===========================
##external knowledge
##===========================

##recent visit to Asia increases risk of tuberculosis
bsc.compute <- buildscorecache.mle(data.df = asia,
                                  data.dists = dist,
                                  max.parents = 2,
                                  dag.retained = ~Tuberculosis|Asia)

dag <- mostprobable(score.cache = bsc.compute,score = "bic")
plotabn(dag.m = dag,data.dists = dist, fontsize.node = 30, edge.arrowwise = 3)
```



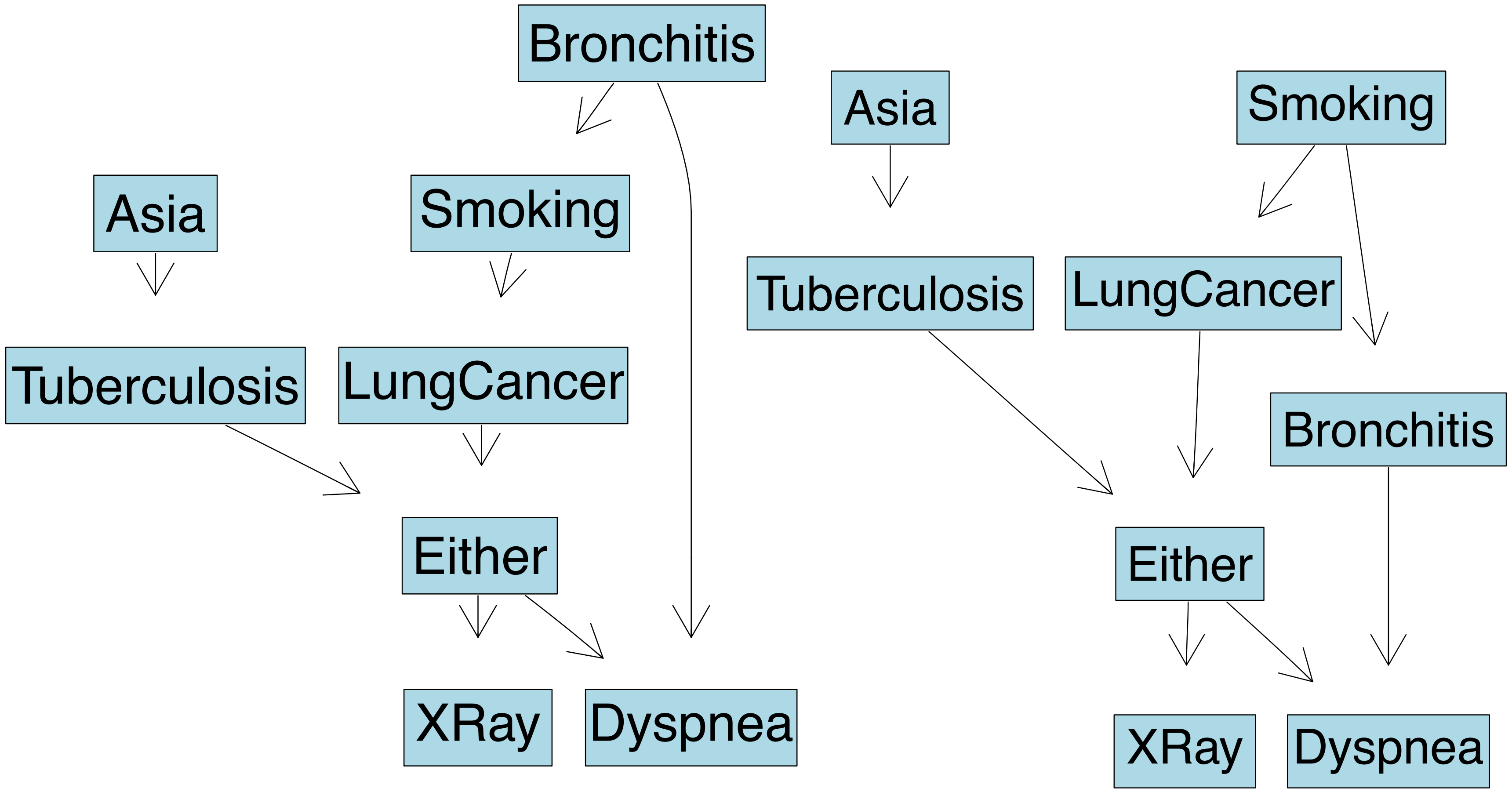
ASIA: EXTERNAL KNOWLEDGE

```
##==========================
##external knowledge
##==========================

##recent visit to Asia increases risk of tuberculosis
bsc.compute <- buildscorecache.mle(data.df = asia,
                                   data.dists = dist,
                                   max.parents = 2,
                                   dag.retained = ~Tuberculosis|Asia)

dag <- mostprobable(score.cache = bsc.compute,score = "bic")
plotabn(dag.m = dag,data.dists = dist, fontsize.node = 30, edge.arro
```

```
> compareDag(ref = t(dag.adj),
+            test = (dag))
$TPR
[1] 0.875

$FPR
[1] 0.01785714

$Accuracy
[1] 0.96875

$FDR
[1] 0.125

$`G-measure`
[1] 0.875

$`F1-score`
[1] 56

$PPV
[1] 0.875

$FOR
[1] 0.125

$`Hamming-distance`
[1] 2
```

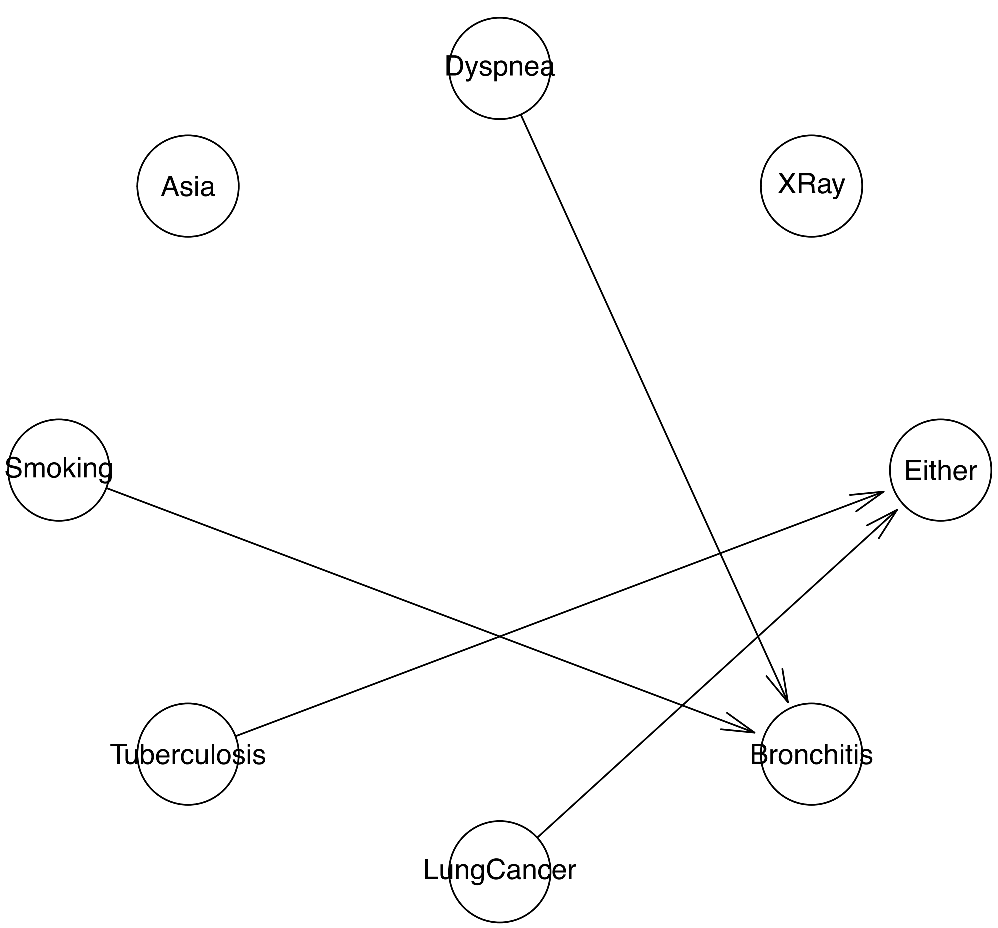
Learned

Truth



ASIA: CONSTRAINT-BASED LEARNING

```
##============================
## constraint-based algorithm
##============================

bn.gs <- gs(asia)
plot(bn.gs)

bn.iamb <- iamb(asia)
plot(bn.iamb)
```

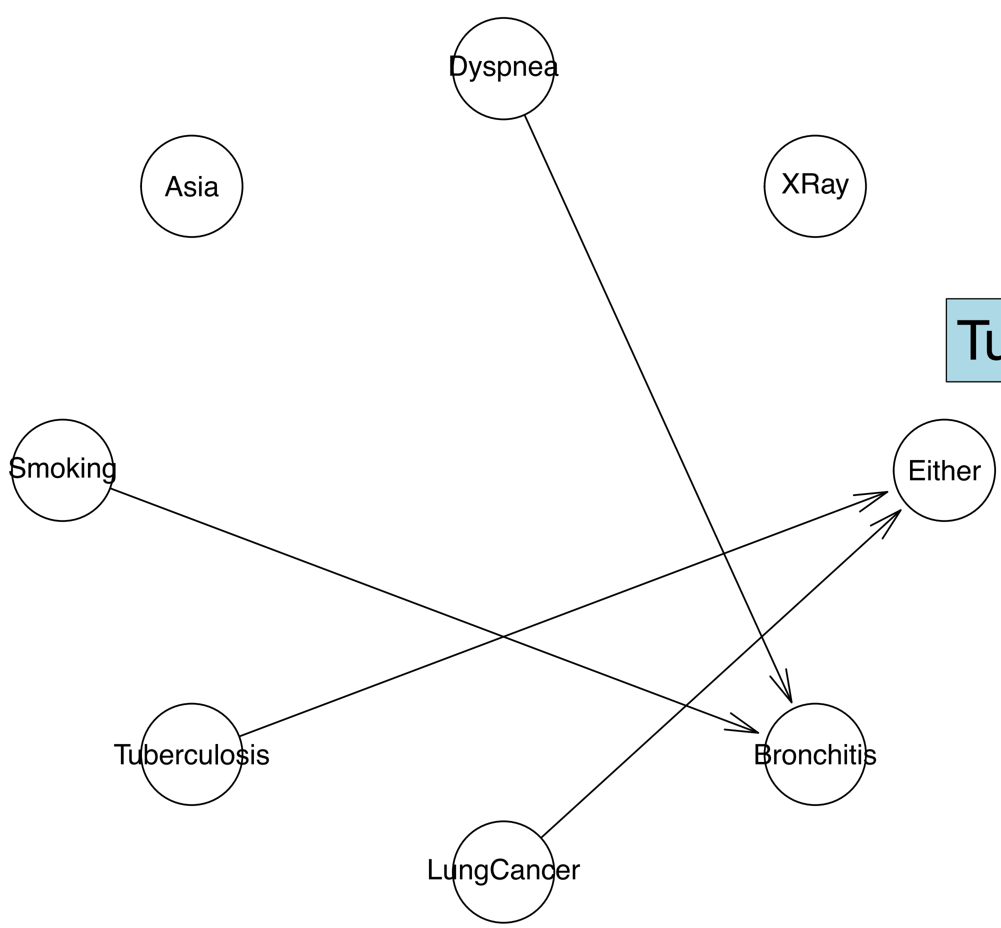


ASIA: CONSTRAINT-BASED LEARNING

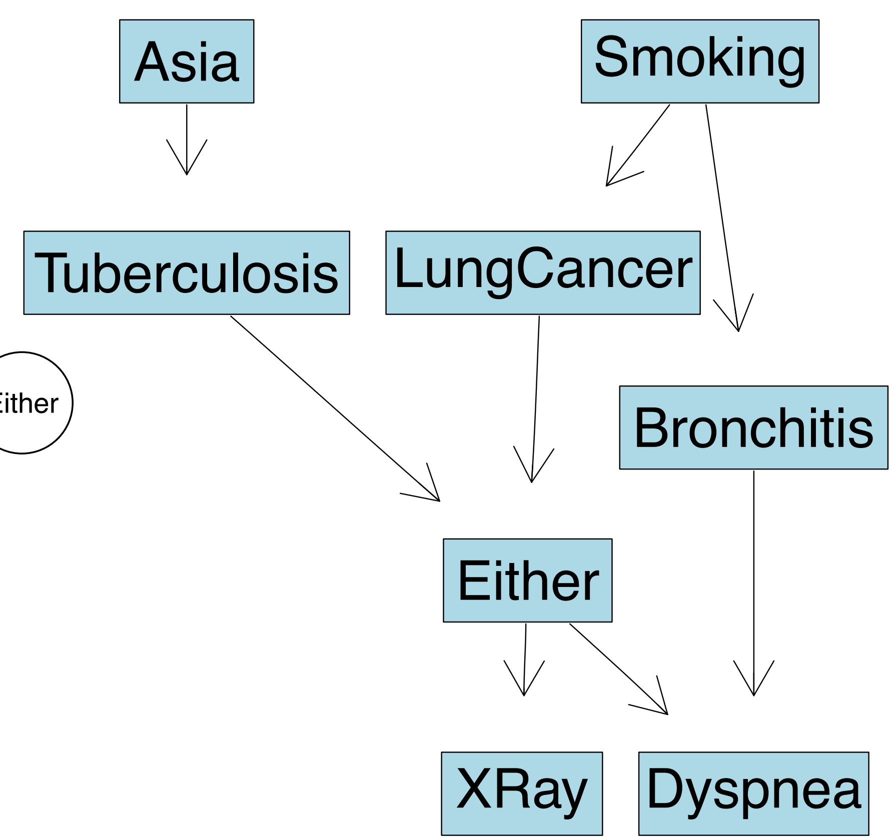
```
##==========================
## constraint-based algorithm
##==========================

bn.gs <- gs(asia)
plot(bn.gs)

bn.iamb <- iamb(asia)
plot(bn.iamb)
```

```
> compareDag(ref = t(dag),
+            test = amat(bn.gs))
$TPR
[1] 0.4285714

$FPR
[1] 0.01754386

$Accuracy
[1] 0.921875

$FDR
[1] 1

$`G-measure`
[1] 0.5669467

$`F1-score`
[1] 15.27273

$PPV
[1] 0.75

$FOR
[1] 1

$`Hamming-distance`
[1] 5
```

Learned

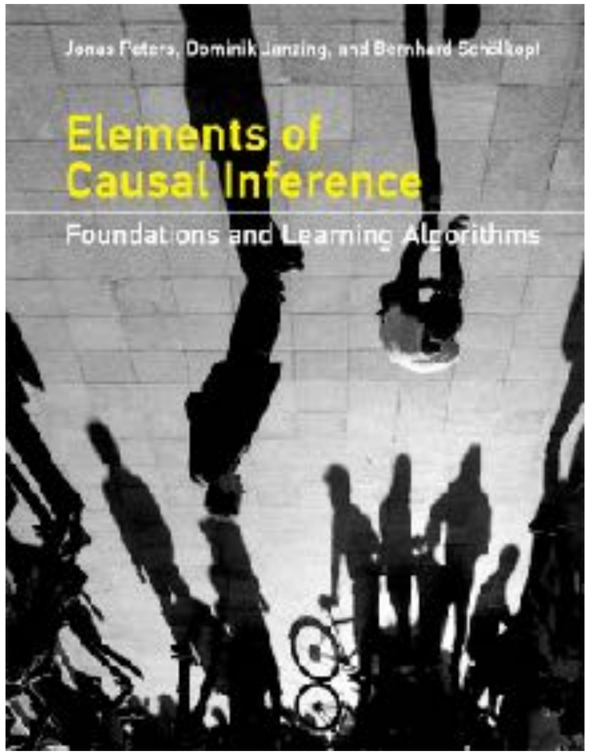


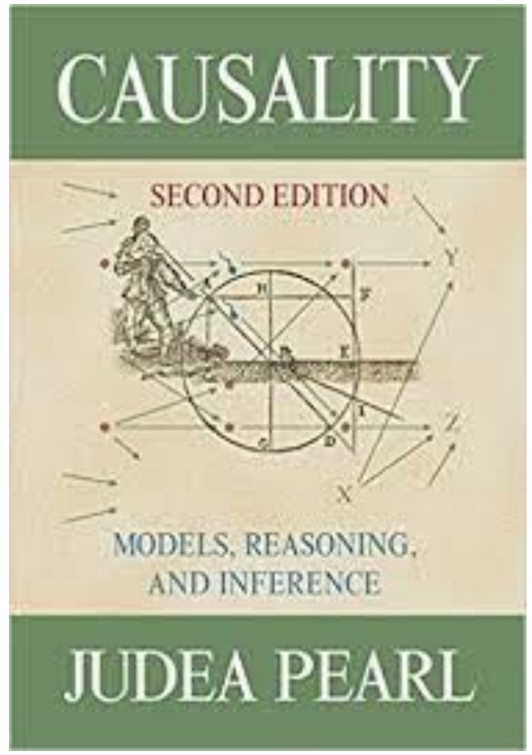
Truth

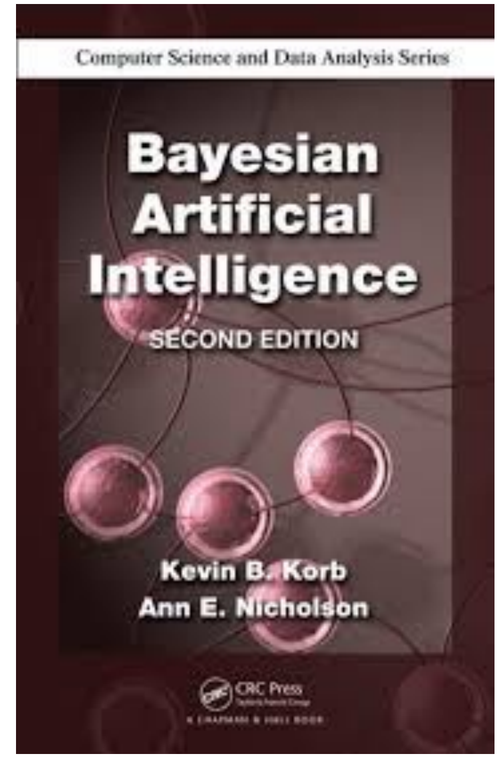


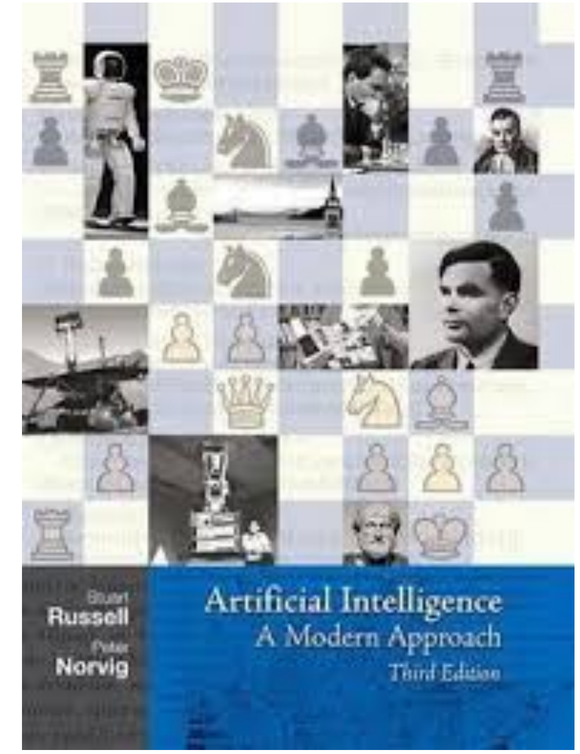
University of Zurich[UZH]

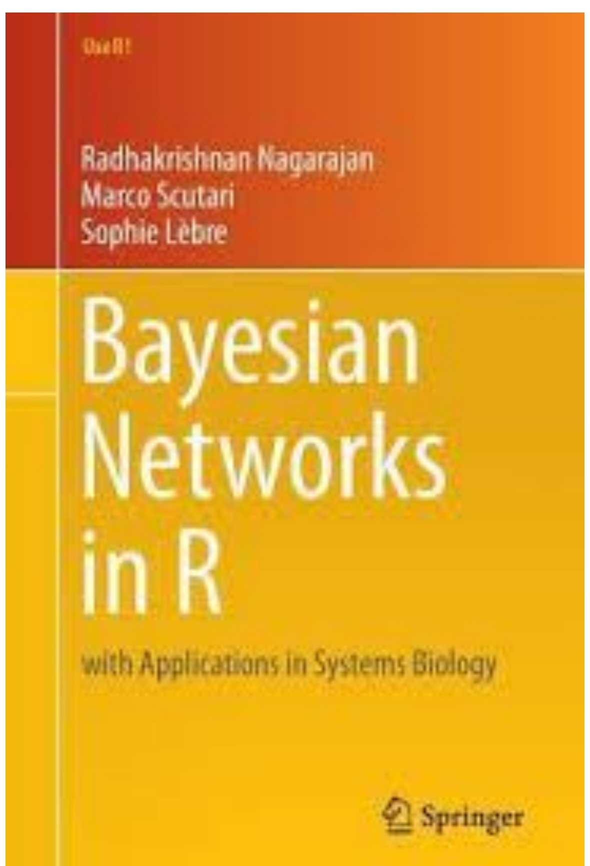
SELECTED BIBLIOGRAPHY

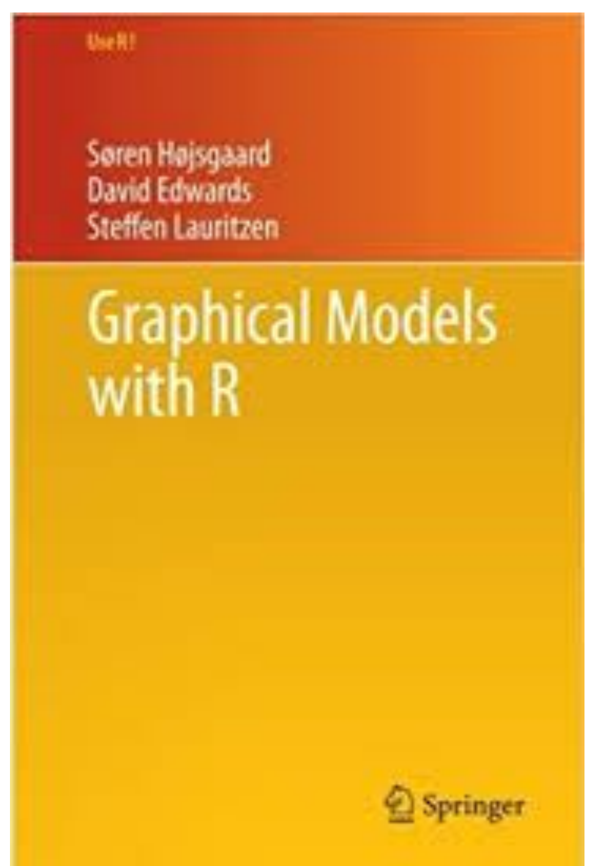


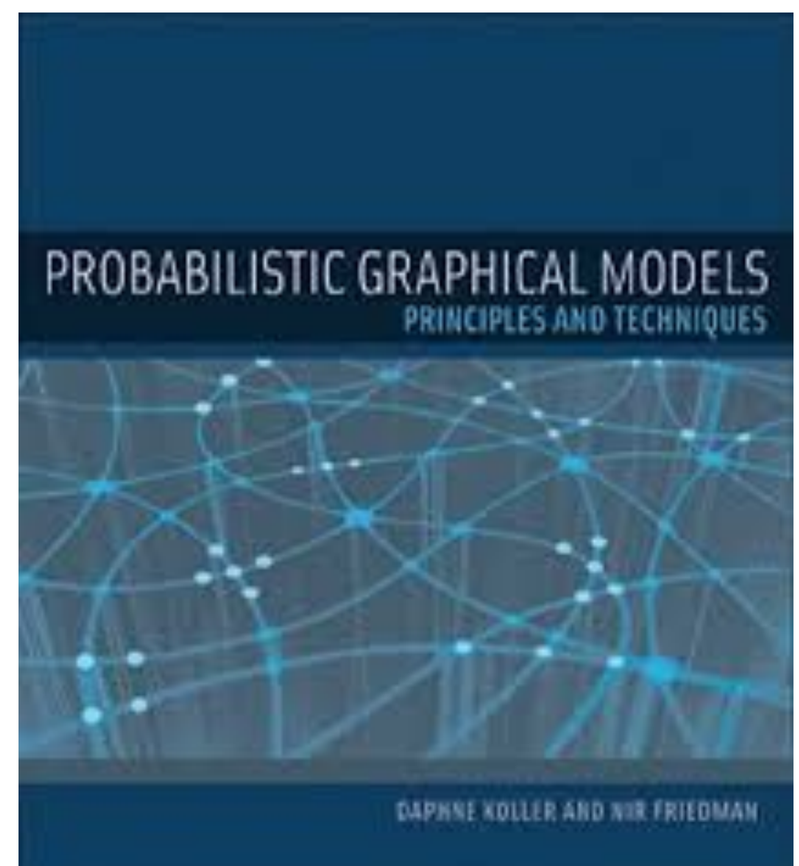




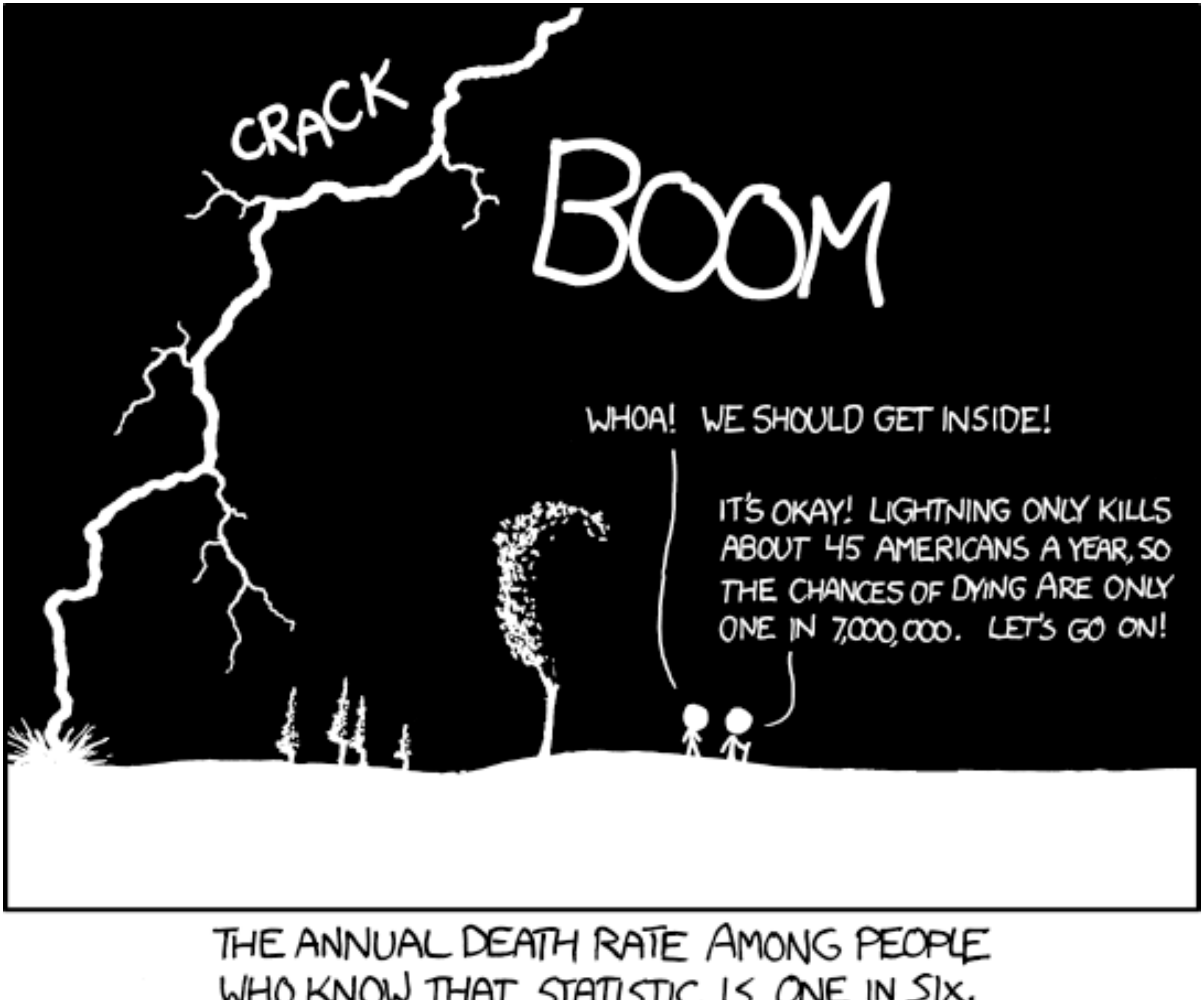








Thank you for your attention



xkcd.com

Backup slides

A path from A to B is blocked if it contains a node s.t. either

- the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or
- the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are C.

If all paths from A to B are blocked, A is said to be d-separated from B by C.

Theorem (Verma & Pearl, 1988): A is d-separated from B by C if, and only if, the joint distribution over all variables in the graph satisfies:

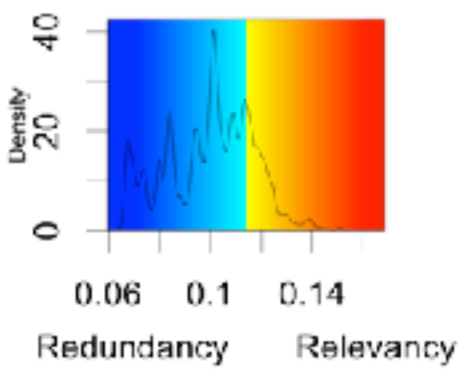
$$\boxed{A \perp\!\!\!\perp_G B | C}$$

Link between statistical statement (conditionally independent) and a graph propriety (d-separation)

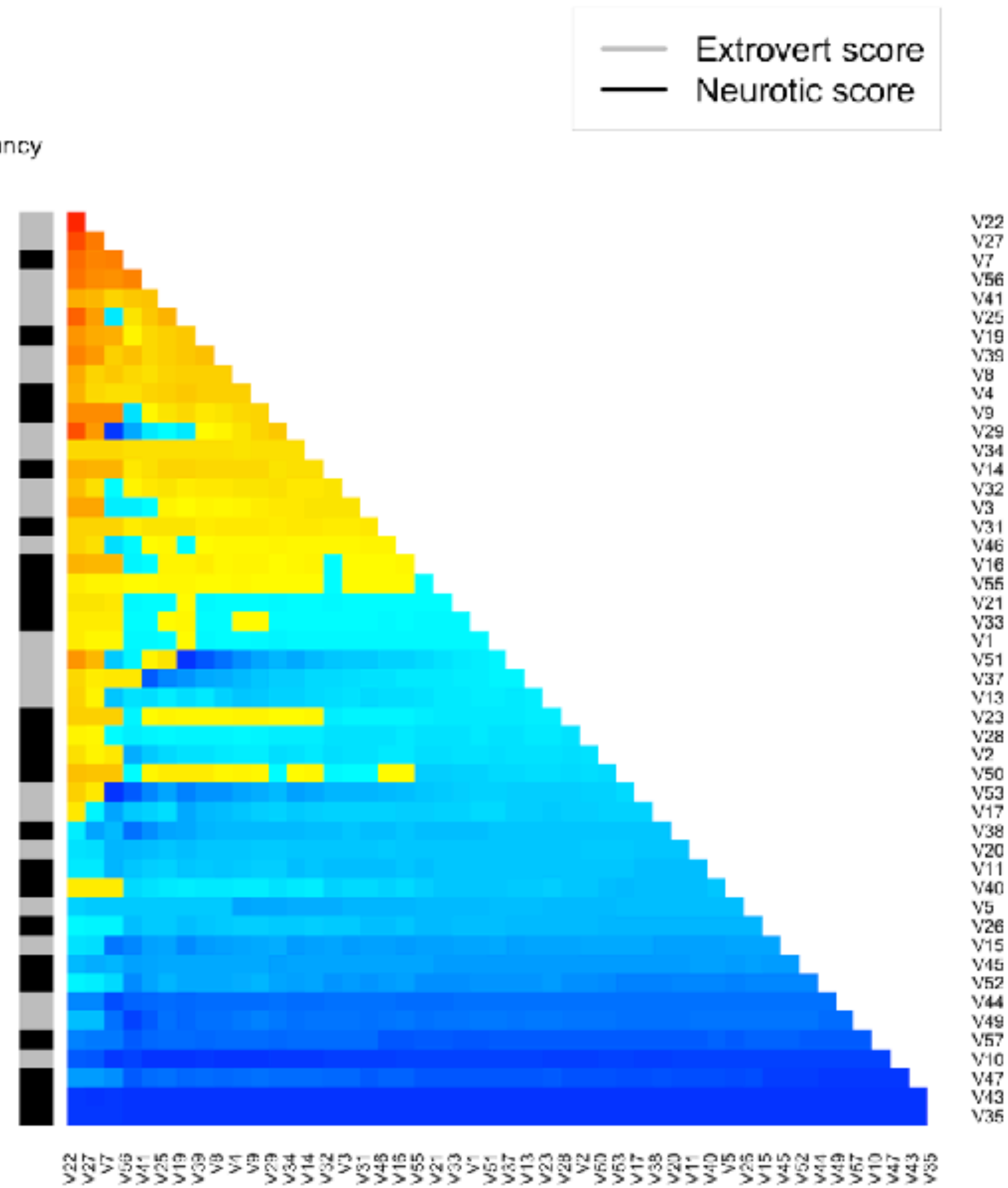
ASIA: HOW MANY PARENT ARE NEEDED?

```
res.mlik <- NULL
res.aic <- NULL
res.bic <- NULL
res.mdl <- NULL

for(i in 1:4){
  mycache.computed.mle <- buildscorecache.mle(data.df = asia,
                                              data.dists = dist,
                                              max.parents = i,
                                              dry.run = FALSE,
                                              maxit = 1000,
                                              tol = 1e-11)

  dag <- mostprobable(score.cache = mycache.computed.mle,score = "aic")
  res.aic <- rbind(res.aic,fitabn.mle(dag.m = dag,data.df = mycache.computed.mle$data.df,data.dists = dist)$aic)
  dag <- mostprobable(score.cache = mycache.computed.mle,score = "bic")
  res.bic <- rbind(res.bic,fitabn.mle(dag.m = dag,data.df = mycache.computed.mle$data.df,data.dists = dist)$bic)
  dag<-mostprobable(score.cache = mycache.computed.mle,score = "mdl")
  res.mdl <- rbind(res.mdl,fitabn.mle(dag.m = dag,data.df = mycache.computed.mle$data.df,data.dists = dist)$mdl)
}

library(ggplot2)
library(reshape)
scoring <- data.frame(AIC = max(-res.aic)/-res.aic, BIC = max(-res.bic)/-res.bic, MDL = max(-res.mdl)/-res.mdl, 1:4)

scoring.long <- melt(scoring, id.vars="X1.4")

ggplot(data = scoring.long, aes(x=X1.4, y=(value), group=variable, color=variable)) +
  geom_line() +
  geom_point() +
  ggtitle("Scoring in function of the number of children", subtitle = NULL) +
  xlab("# of parent per node") +
  ylab("% of max score") +
  scale_x_continuous(breaks=c(1,2,3,4,5,6,7))
```

EYSENCK PERSONALITY INVENTORY



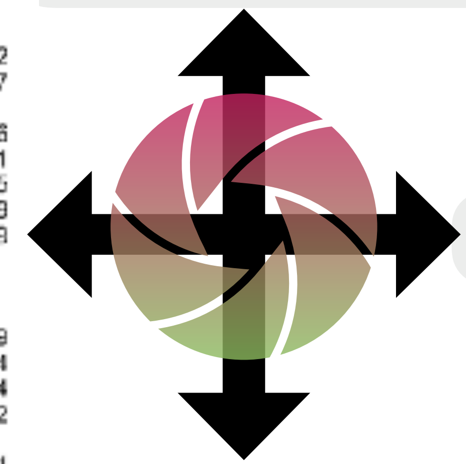
EPI: 3570 observations and 57 variables



Structure of EPI:

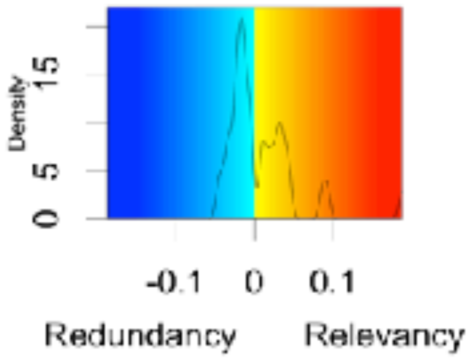
✓ Lie scale (9 responses)

Extovert score (24)



Neurotic score (24)

University of Zurich[UZH]



Pima Indians Diabetes Database

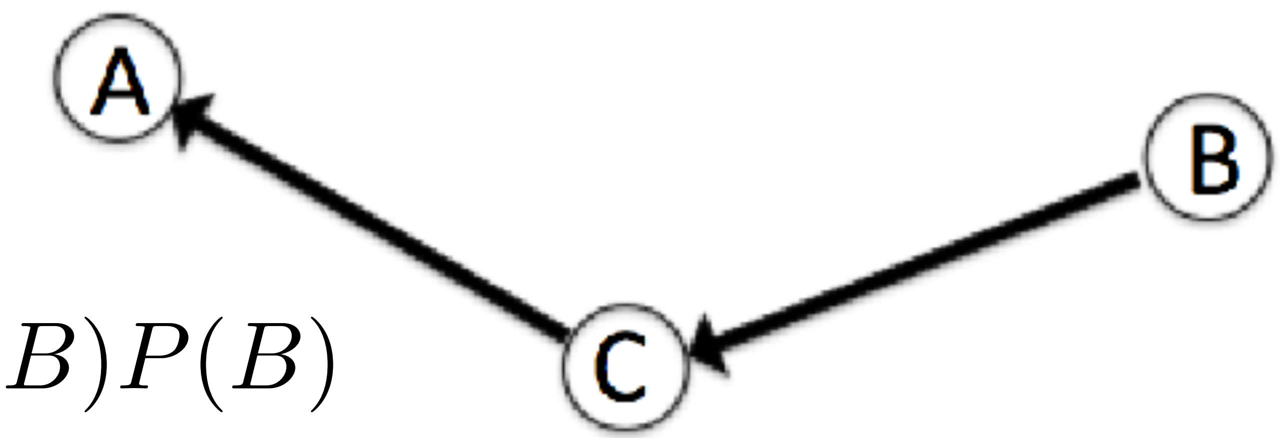
768 observations on 9 variables

	glucose	mass	age	pedigree	insulin	pregnant	pressure	triceps
glucose	0.187							
mass	0.092	0.04						
age	0.085	0.021	0.036					
pedigree	0.031	0.007	-0.005	-0.005				
insulin	0.029	-0.041	-0.024	-0.015	-0.013			
pregnant	0.044	0.013	0.017	-0.03	-0.016	-0.008		
pressure	0.024	-0.009	-0.021	-0.024	-0.019	-0.015	-0.014	
triceps	0.034	0.009	-0.046	-0.034	-0.024	-0.035	-0.02	

Let A, B and C non intersecting subsets of nodes in a DAG G

A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$



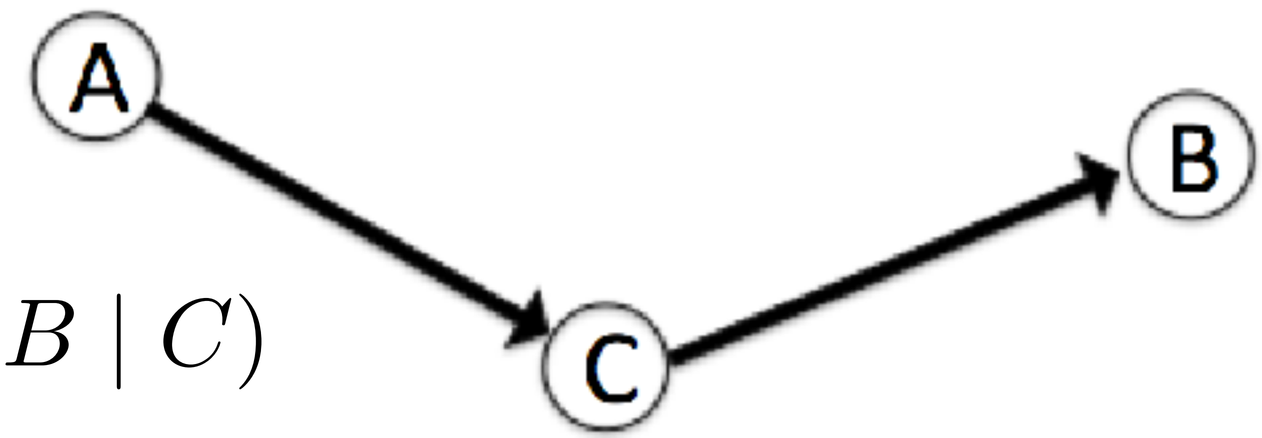
$$P(A, B, C) = P(A \mid C)P(C \mid B)P(B)$$

$$\begin{aligned}
P(A, B \mid C) &= \frac{P(A \mid C)P(C \mid B)P(B)}{P(C)} \\
&= \frac{P(A \mid C)P(B, C)}{P(C)} \\
&= P(A \mid C)P(B \mid C)
\end{aligned}$$

Let A, B and C non intersecting subsets of nodes in a DAG G

A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$



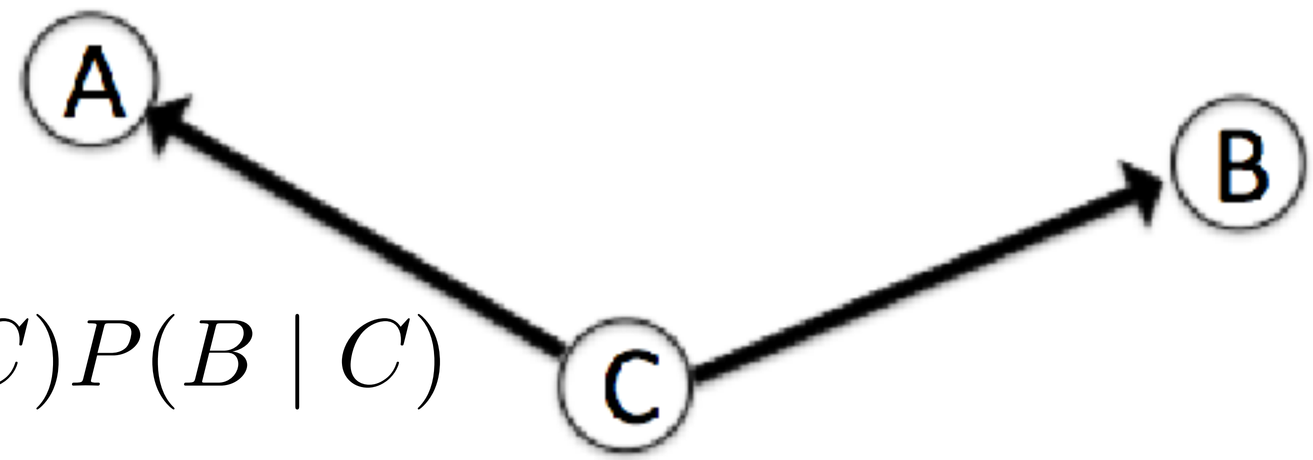
$$P(A, B, C) = P(A)P(C \mid A)P(B \mid C)$$

$$\begin{aligned} P(A, B \mid C) &= \frac{P(A)P(C \mid A)P(B \mid C)}{P(C)} \\ &= \frac{P(A, C)P(B \mid C)}{P(C)} \\ &= P(A \mid C)P(B \mid C) \end{aligned}$$

Let A, B and C non intersecting subsets of nodes in a DAG G

A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$



$$P(A, B, C) = P(C)P(A \mid C)P(B \mid C)$$

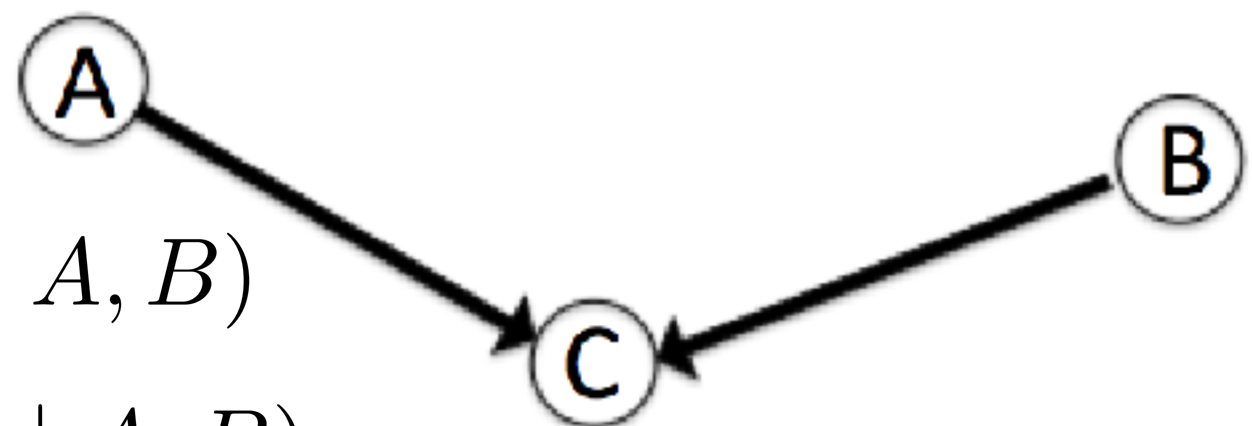
$$P(A, B \mid C) = \frac{P(C)P(A \mid C)P(B \mid C)}{P(C)}$$

$$= P(A \mid C)P(B \mid C)$$

Let A, B and C non intersecting subsets of nodes in a DAG G

A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$



$$P(A, B, C) = P(A)P(B)P(C \mid A, B)$$

$$P(A, B \mid C) = \frac{P(A)P(B)P(C \mid A, B)}{P(C)}$$

$$= \frac{P(A)P(B)P(A, B, C)}{P(A)P(B)P(C)}$$

$$= P(A, B \mid C)$$

$A \not\perp\!\!\!\perp_P B|C$

Constraint-based algorithms

- *Inductive Causation* (IC): (Verma and Pearl, 1991)
  - Provides a framework for learning the structure of Bayesian networks using conditional independence tests in three steps
  - A major problem of the IC algorithm is that the first two steps cannot be applied to any real-world problem due to computational complexity …
- *PC*: first practical application of the IC algorithm (Spirtes et al., 2001)
  - backward selection procedure from the saturated graph
- *Grow-Shrink* (GS) (Margaritis, 2003)
  - Simple forward selection MB detection approach
- *Incremental Association* (IAMB): (Tsamardinos et al., 2003)
  - two-phase selection scheme based on a forward selection followed by a backward selection of the MB

- Constraint-based methods require a Markov and faithfulness assumption
- Conditional independencies in the distribution exactly equal the ones encoded in the DAG via d-separation

$$A \perp\!\!\!\perp_G B | C \underset{\text{Faithful}}{\overset{\text{Markov}}{\rightleftarrows}} A \perp\!\!\!\perp_P B | C$$

- Causal sufficiency: no unmeasured common causes

In a pratical perspective:

- Testing mixture of data?
- Testing assumptions?

ASIA: KNOWN NETWORK

```
fitabn(dag.m = ~Asia|Tuberculosis+
        Tuberculosis|Either +
        Either|XRay:Dyspnea +
        Smoking|Bronchitis:LungCancer +
        LungCancer|Either +
        Bronchitis|Dyspnea,data.df = asia,data.dists = dist)$modes
```

```
$Asia
 Asia|(Intercept) Asia|Tuberculosis
        -4.811200          1.765763

$Smoking
Smoking|(Intercept)  Smoking|LungCancer  Smoking|Bronchitis
          -1.027065            2.356988            1.807460

$Tuberculosis
Tuberculosis|(Intercept)       Tuberculosis|Either
               -12.22120                  10.21823

$LungCancer
LungCancer|(Intercept)        LungCancer|Either
             -12.07565                 14.18547

$Bronchitis
Bronchitis|(Intercept)       Bronchitis|Dyspnea
             -1.388644                 3.200393

$Either
Either|(Intercept)          Either|XRay       Either|Dyspnea
         -8.656348             8.259773             1.538789

$XRay
XRay|(Intercept)
       -2.052496

$Dyspnea
Dyspnea|(Intercept)
         -0.1201444
```

```
fitabn.mle(dag.m = dag.adj,data.df = asia,data.dists = dist)$coef
```

```
$Asia
     Asia|intercept Tuberculosis
[1,]      -4.811371     1.766849

$Smoking
     Smoking|intercept LungCancer Bronchitis
[1,]         -1.027075   2.357079   1.807472

$Tuberculosis
     Tuberculosis|intercept   Either
[1,]              -8.517393 6.516139

$LungCancer
     LungCancer|intercept   Either
[1,]            -8.517393 10.62598

$Bronchitis
     Bronchitis|intercept  Dyspnea
[1,]            -1.388655 3.200415

$Either
     Either|intercept     XRay  Dyspnea
[1,]        -8.665128 8.268402 1.539146

$XRay
     XRay|intercept
[1,]        -2.0525

$Dyspnea
     Dyspnea|intercept
[1,]        -0.1201443
```

ASIA: KNOWN NETWORK


```
fitabn(dag.m = ~Asia|Tuberculosis+
         Tuberculosis|Either +
         Either|XRay:Dyspnea +
         Smoking|Bronchitis:LungCancer +
         LungCancer|Either +
         Bronchitis|Dyspnea,data.df = asia,data.dists = dist)$modes

$Asia
 Asia|(Intercept) Asia|Tuberculosis
        -4.811200          1.765763

$Smoking
Smoking|(Intercept)  Smoking|LungCancer  Smoking|Bronchitis
          -1.027065            2.356988            1.807460

$Tuberculosis
Tuberculosis|(Intercept)       Tuberculosis|Either
               -12.22120                  10.21823

$LungCancer
LungCancer|(Intercept)       LungCancer|Either
             -12.07565                14.18547

$Bronchitis
Bronchitis|(Intercept)       Bronchitis|Dyspnea
             -1.388644                 3.200393

$Either
Either|(Intercept)          Either|XRay      Either|Dyspnea
         -8.656348             8.259773            1.538789

$XRay
XRay|(Intercept)
       -2.052496

$Dyspnea
Dyspnea|(Intercept)
         -0.1201444
```

```
fitabn.mle(dag.m = dag.adj,data.df = asia,data.dists = dist)$coef

$Asia
     Asia|intercept Tuberculosis
[1,]      -4.811371     1.766849

$Smoking
     Smoking|intercept LungCancer Bronchitis
[1,]         -1.027075   2.357079   1.807472

$Tuberculosis
     Tuberculosis|intercept   Either
[1,]              -8.517393 6.516139

$LungCancer
     LungCancer|intercept   Either
[1,]            -8.517393 10.62598

$Bronchitis
     Bronchitis|intercept  Dyspnea
[1,]            -1.388655 3.200415

$Either
     Either|intercept     XRay  Dyspnea
[1,]        -8.665128 8.268402 1.539146

$XRay
     XRay|intercept
[1,]        -2.0525

$Dyspnea
     Dyspnea|intercept
[1,]        -0.1201443
```