University of Zurich[UZH]

GILLES KRATZER

JOINT WORK WITH PROF. DR. REINHARD FURRER

EBPI PHD RETREAT JANUARY 29 - JANUARY 30, 2018

ADVANCES IN ADDITIVE BAYESIAN NETWORK APPLIED TO OBSERVATIONAL SYSTEM EPIDEMIOLOGY DATASETS

- *Classical aim in epidemiology is to investigate relationship between covariate and ONE outcome*
- *Typically based on expert knowledge*

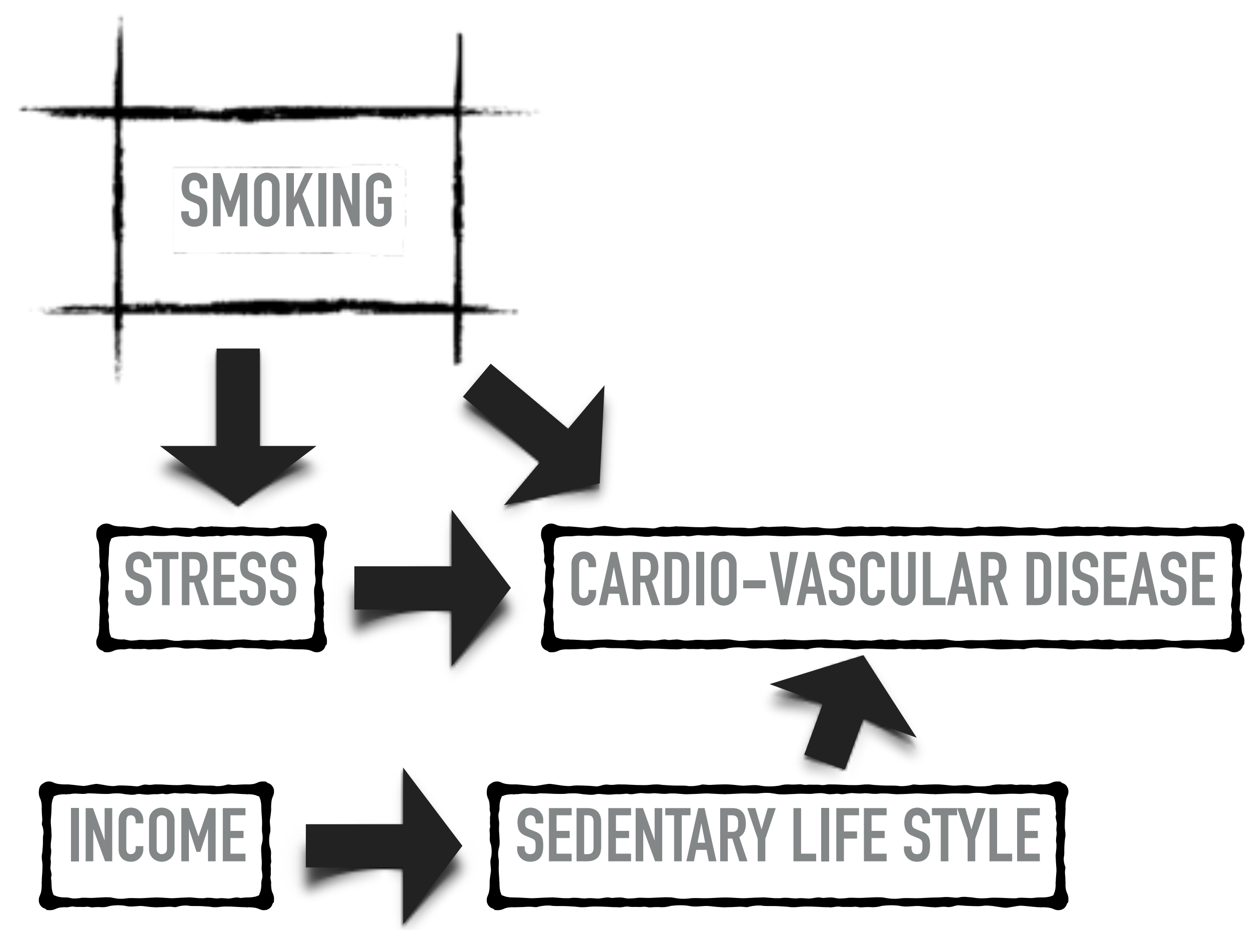
Issues:

- Multi-collinearity
- Dependence
- Confounders
- **Multivariate** versus **Multivariables**

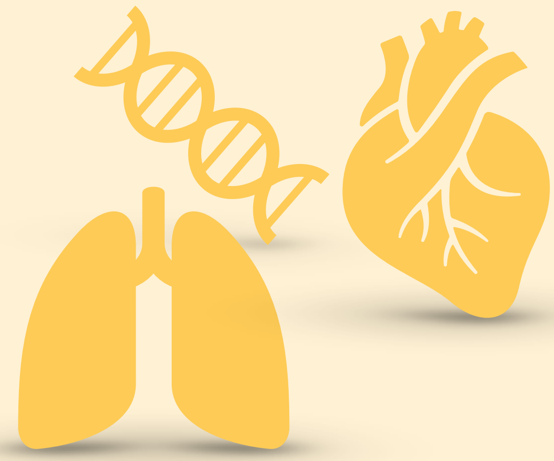
- *Classical aim in epidemiology is to investigate relationship between covariate and ONE outcome*
- *Typically based on expert knowledge*

Issues:

- Multi-collinearity
- Dependence
- Confounders
- **Multivariate** versus **Multivariables**



Enderlein and al. (1996)



DISEASE LEVEL

- Multiple outcomes/Scores
- Target variables for intervention
- Beginning of the coil of discovery

POPULATION LEVEL

- Demographic data
- Meta population information
- Cluster

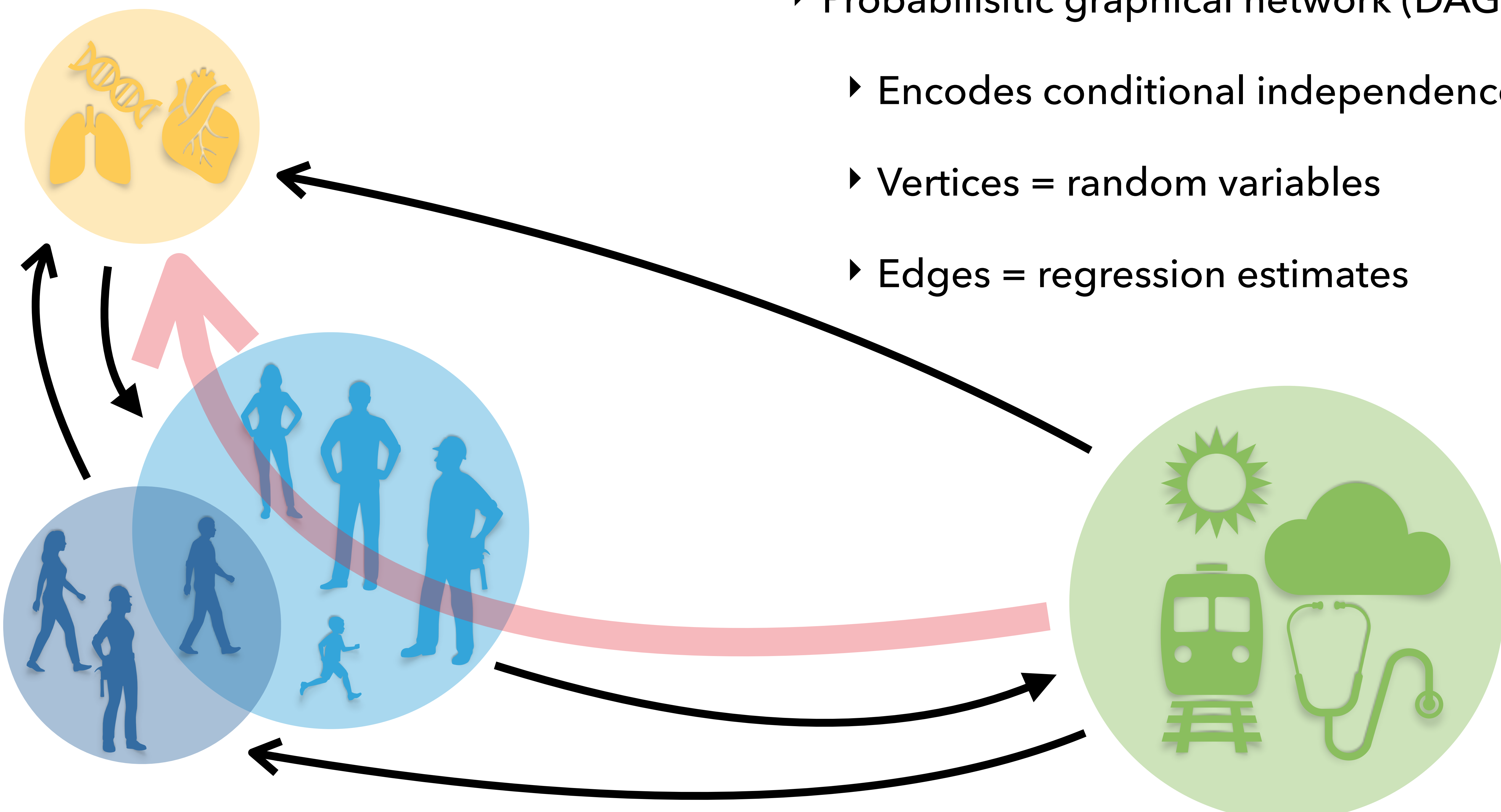
ENVIRONMENT LEVEL

- External factors
- Ecology
- Living condition

Example

- **Metabolic syndrom**
- A clustering of 3/5 medical conditions
- Observational data
- Age, gender, …
- Random effect
- Weather condition
- Socio-economic condition
- Housing

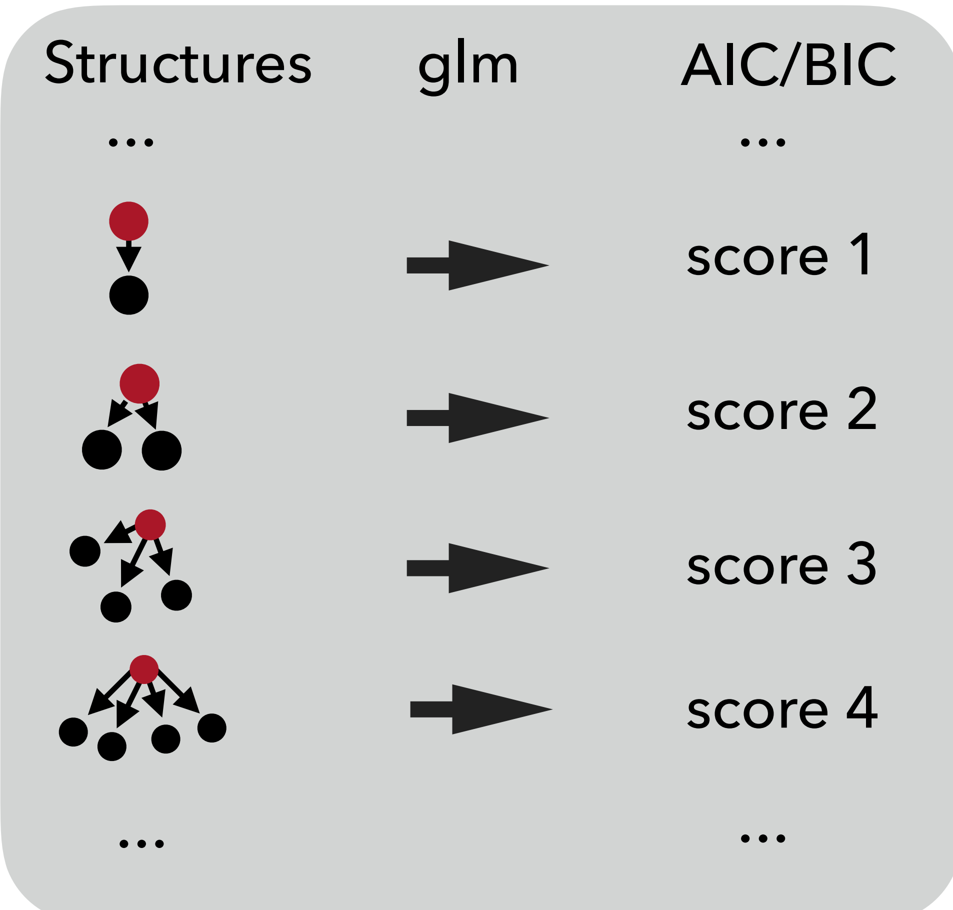
Main purpose of ABN: Sort out **directly** associated versus **indirectly** associated, as they are not primary target for intervention



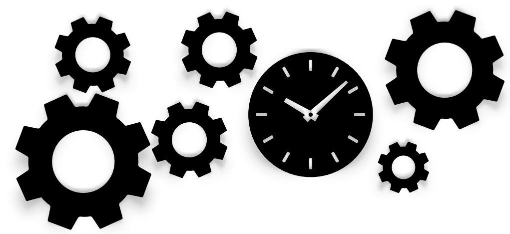
Bayesian Network

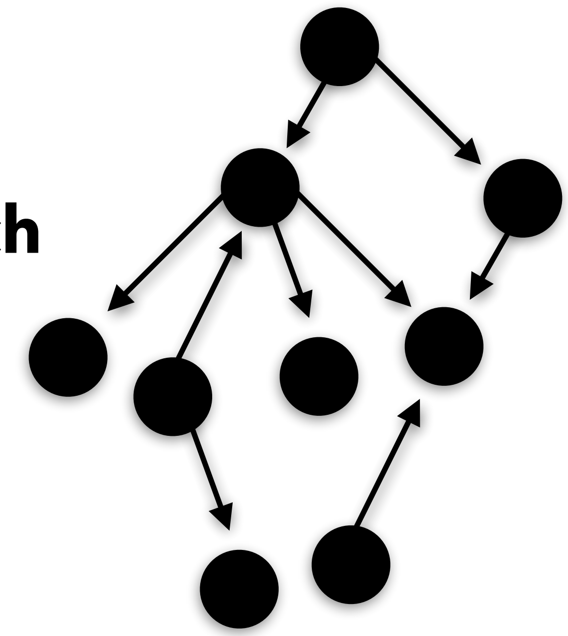
- Probabilisitic graphical network (DAG)
  - Encodes conditional independence
  - Vertices = random variables
  - Edges = regression estimates

Search and score algorithm



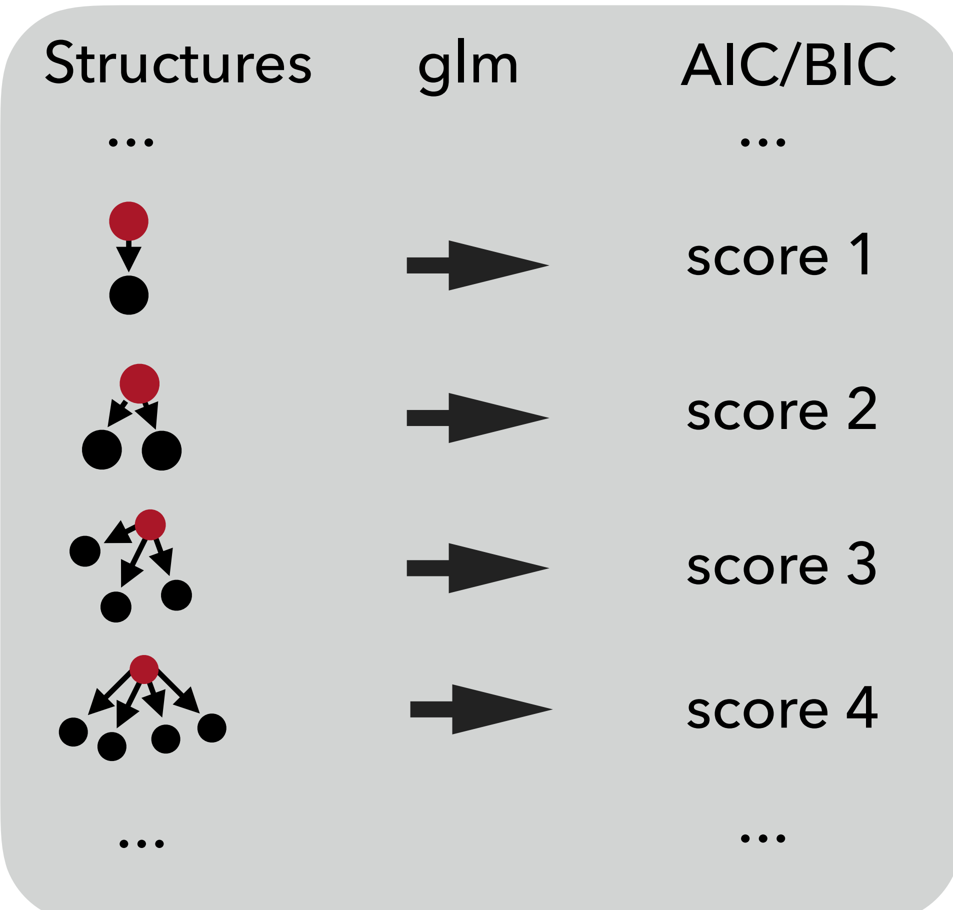
Exact or heuristic search



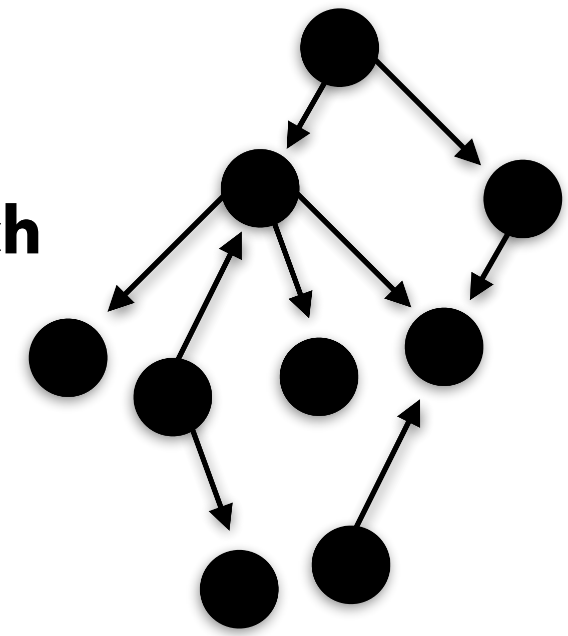


Bayesian network with highest posterior probability

Search and score algorithm



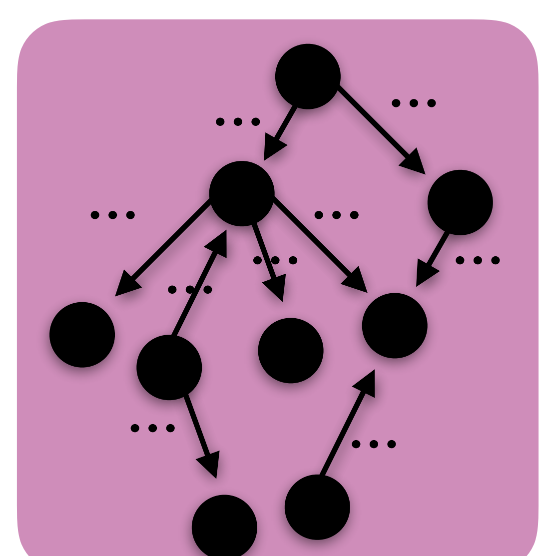
Exact or heuristic search



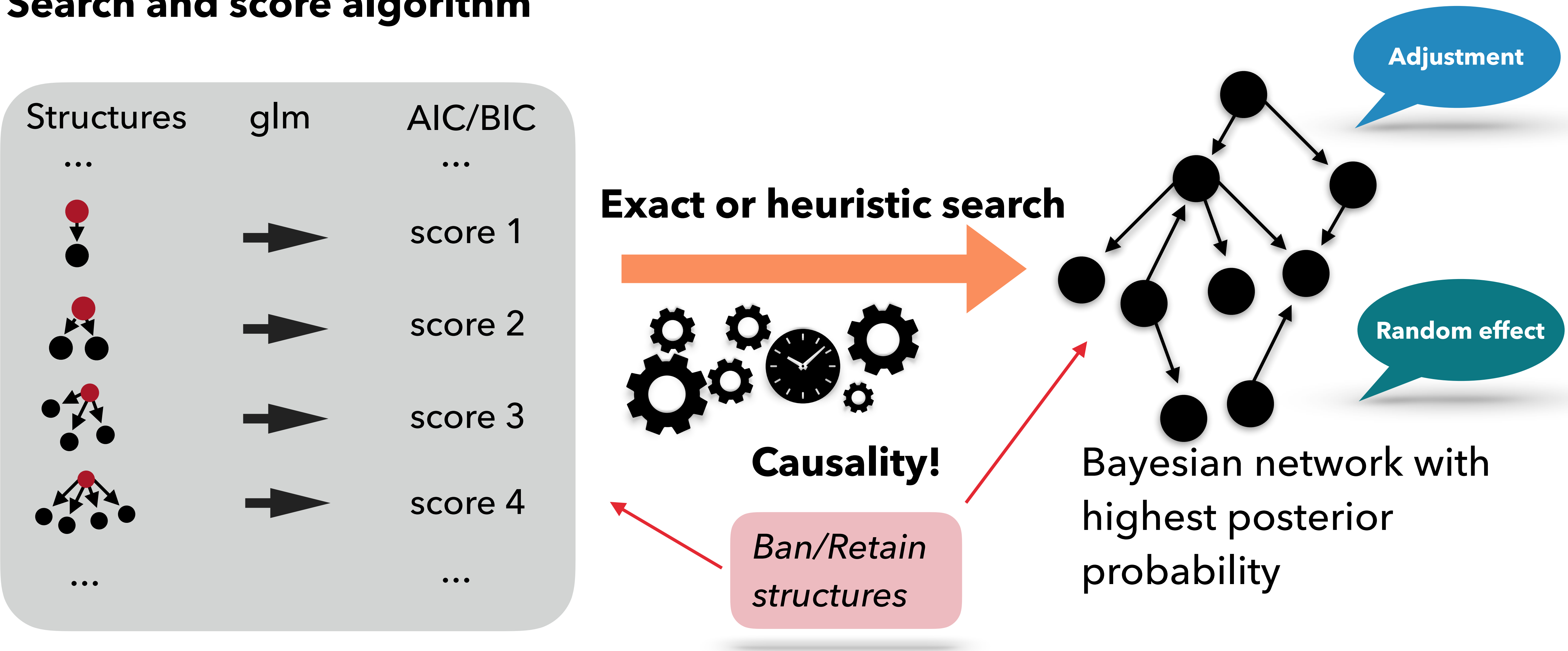
Bayesian network with highest posterior probability

Parameter estimation

- compute marginal posterior density
- regression estimate

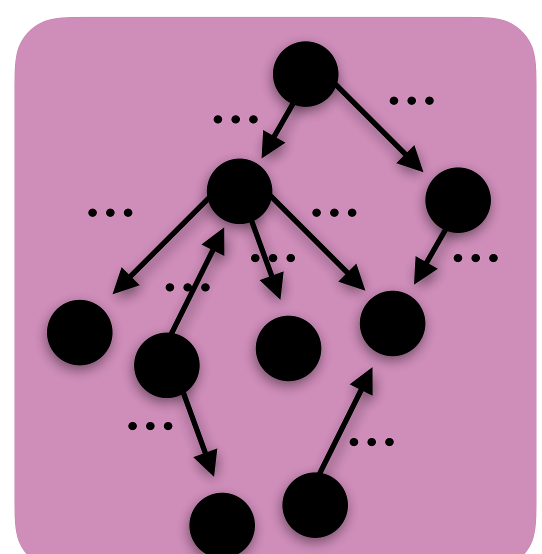


Search and score algorithm



Parameter estimation

- compute marginal posterior density
- regression estimate

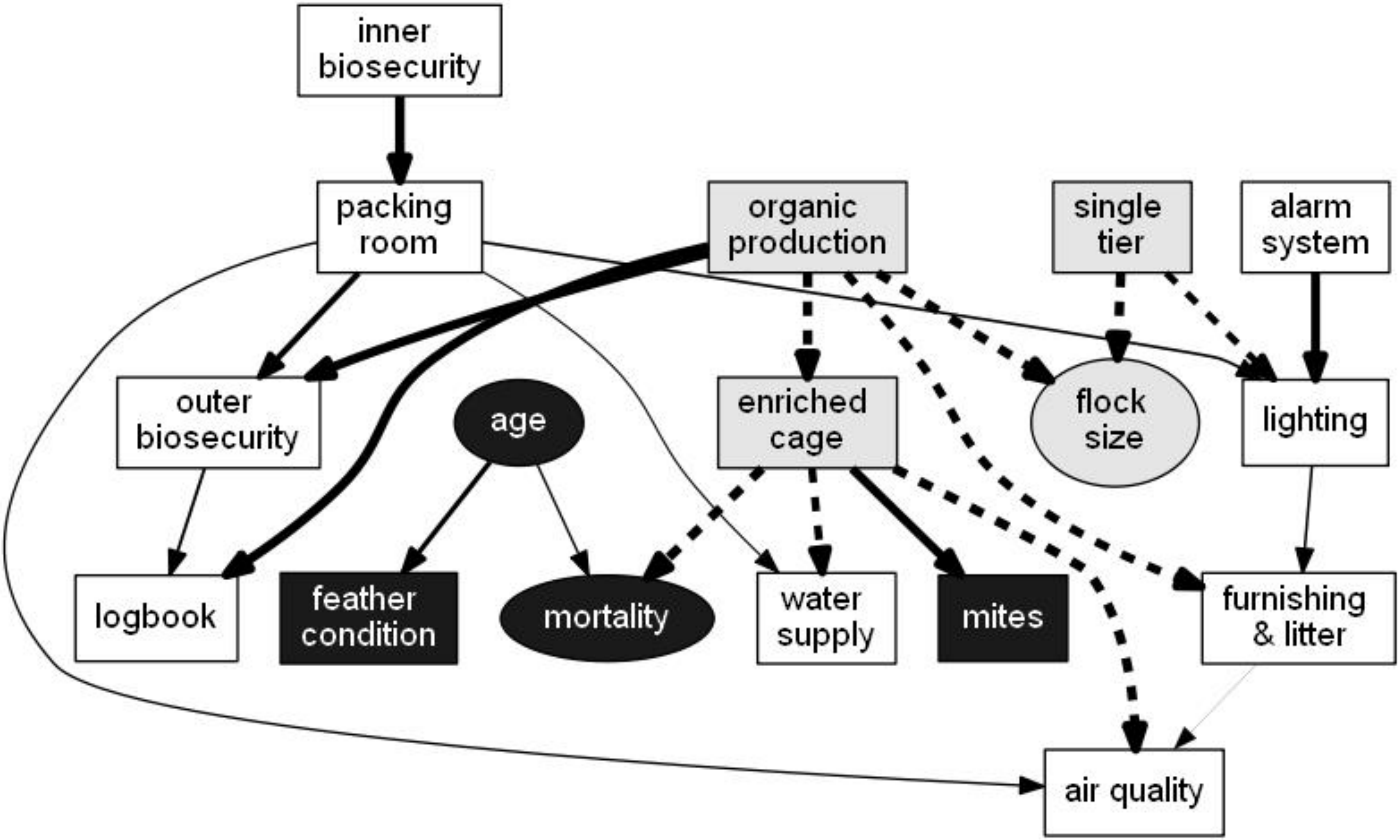


Using R

buildscorecache()

mostprobable()

fitabn()



Arianna Comin et al (2017); Revealing the structure of the associations between housing system, facilities, management and welfare of commercial laying hens using Additive Bayesian networks

- Simple output
- Arc coefficients: easy to interpret
- Statistical guarantees

Current implementation

- Distributed as an R package (CRAN)
- Bayesian regression based on INLA (lm, logit and Poisson) with possibly **random effect**
- Most probable search (**exact search**) and Hill climber (heuristic approach)

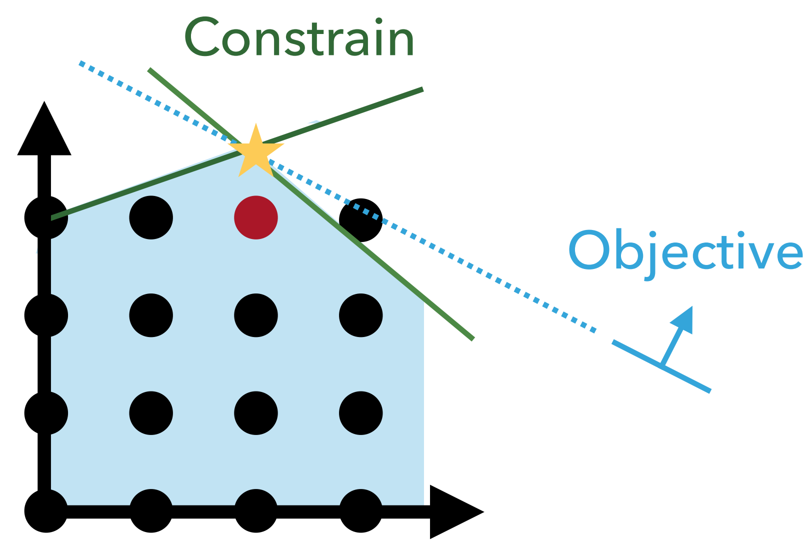
(Very!) Near Future features

- Arc strength based on Mutual Information
    - Significance not p-value based
- GLM implementation (data separation, multinomial variable, adjustment)
    - Multiple scores: AIC, BIC, MDL

FUTUR WORK

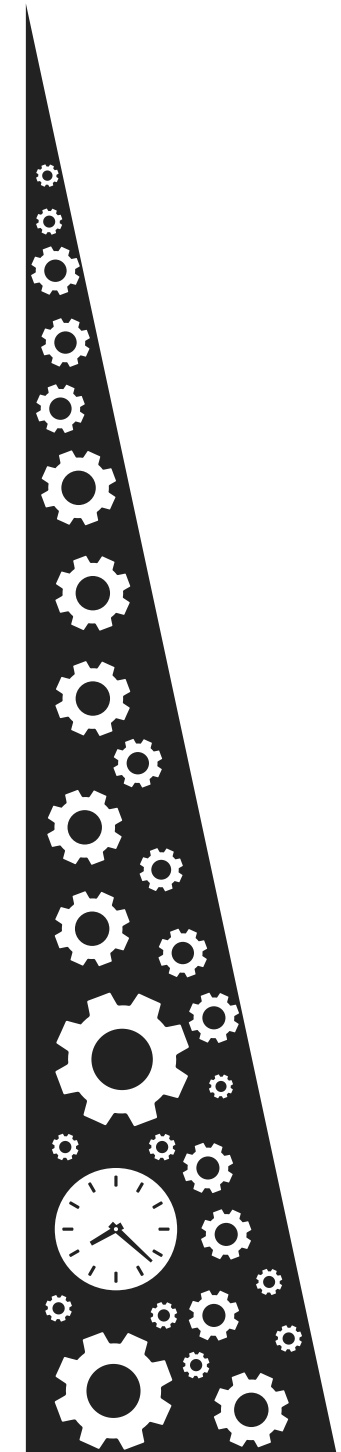
- Search algorithm based on Mutual Information
  - Penalized by χ^2 *(de Campos et al, 2006)*

Increasing order of complexity!



- Algorithmic search implementation
  - Heuristic search/Hybrid search
  - Integer programming (*Cussens, 2012*)

- Bayesian regression work horse
  - Stan implementation
  - Diaconis-Ylvisaker conjugate priors (*Pittavino et al, 2016*)



- Causal belief: Informative prior structure *versus* incomplete synthetic observations

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
- Ranking of 100 variables with 100’000 observations in ~14 minutes! (forward greedy search)

[1] *Walter et al (2009)*

f_i candidate feature to be ranked

C set of variables of importance

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

S set of already selected variables

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

$$\mathrm{score}_i = \mathrm{MI}(f_i; \mathbf{C}) - \beta \sum_{F_s \in \mathbf{S}} \alpha(f_i, f_s, \mathbf{C})\ \mathrm{MI}(f_i; f_s)$$

Estévez and al. (2009)

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

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

Greedy search

Forward - argmax

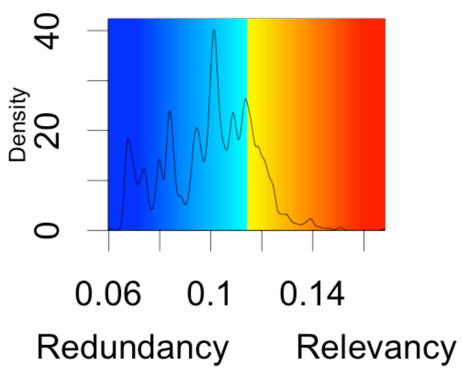
$$\text{score}_i = \underbrace{\mathrm{MI}(f_i;\mathbf{C})}_{\text{Relevance}} - \beta \sum_{F_s \in \mathbf{S}} \underbrace{\alpha(f_i, f_s, \mathbf{C})}_{\text{Normalization}} \underbrace{\mathrm{MI}(f_i; f_s)}_{\text{Redundancy}}$$

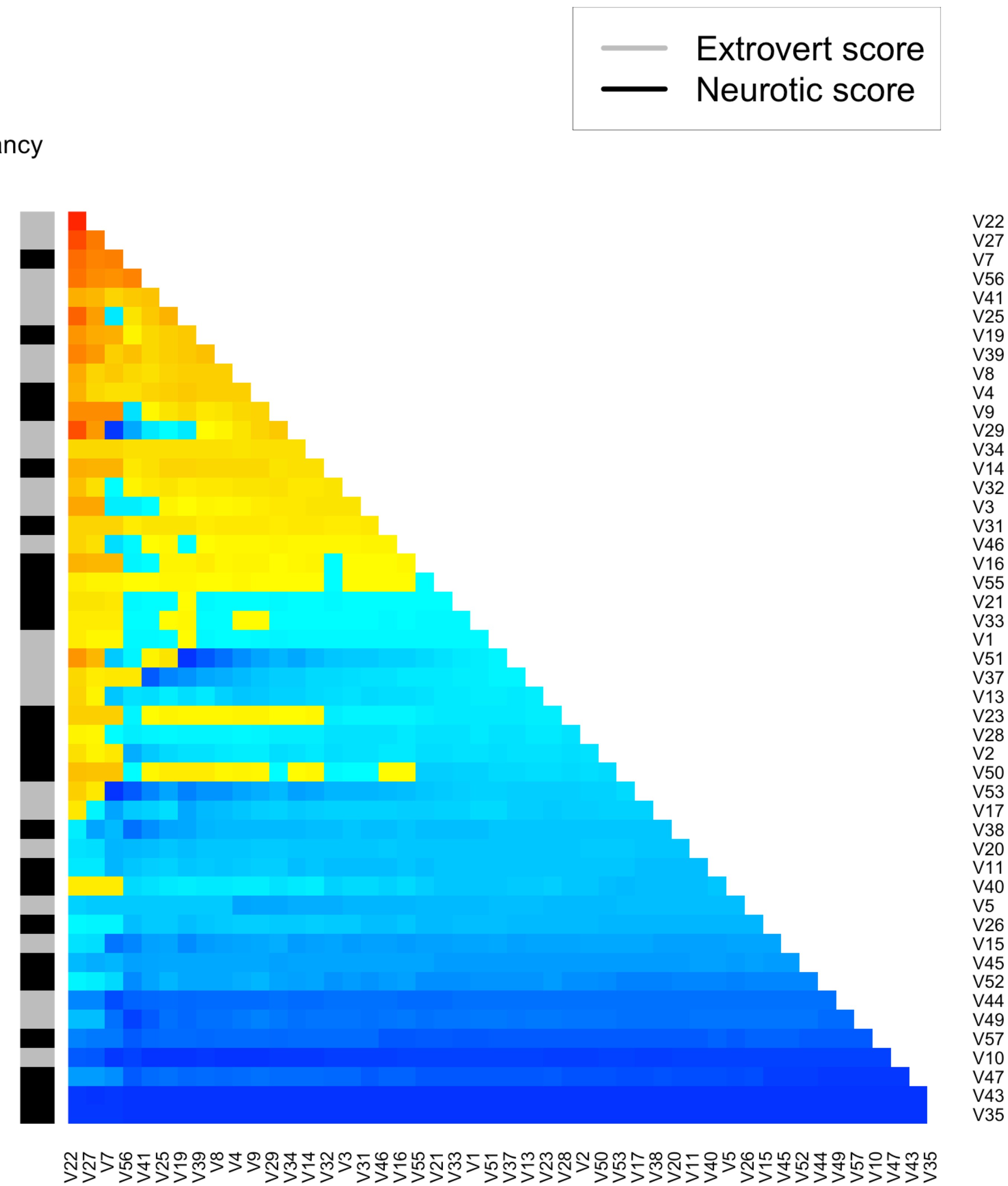
Estévez and al. (2009)

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

EYSENCK PERSONALITY INVENTORY

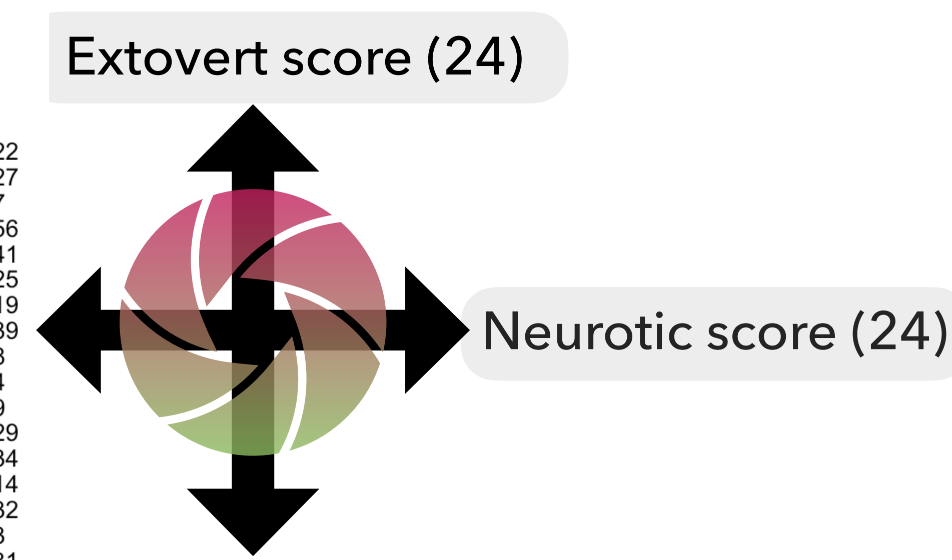
EPI: 3570 observations and 57 variables



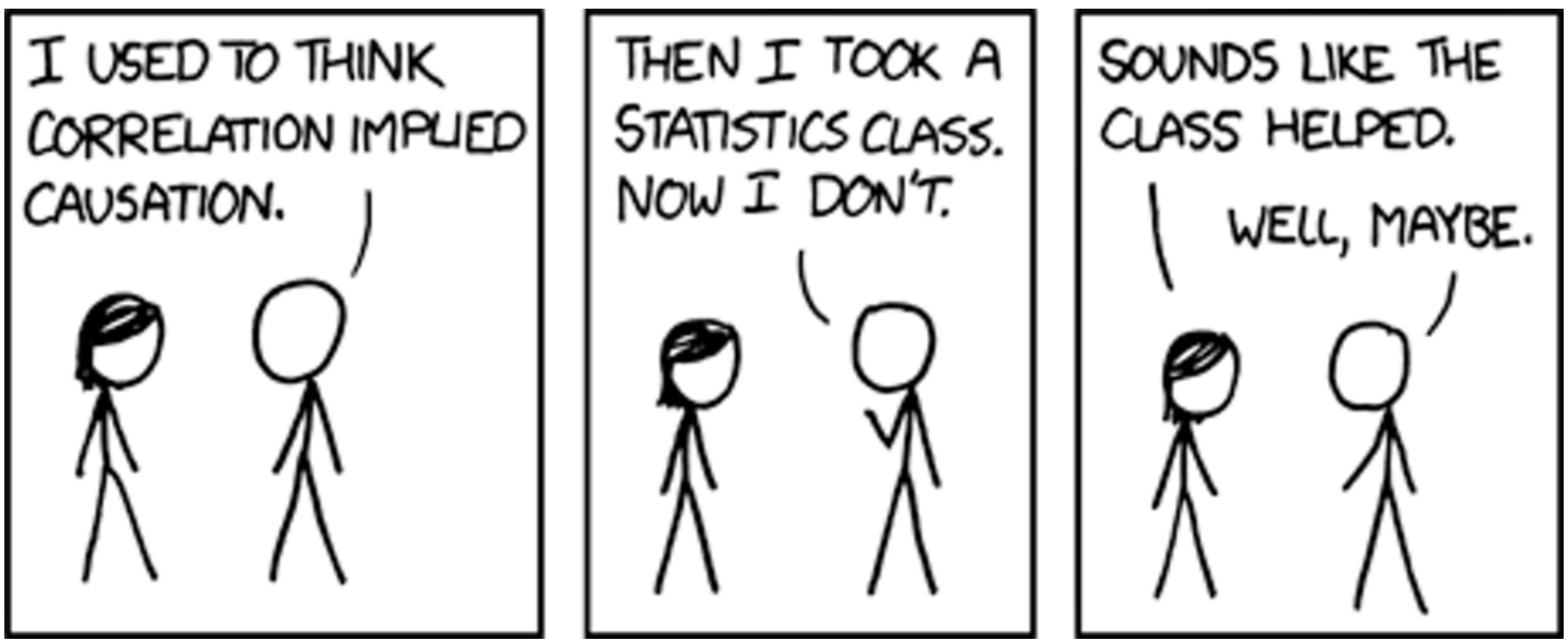


Structure of EPI:

✓ Lie scale (9 responses)

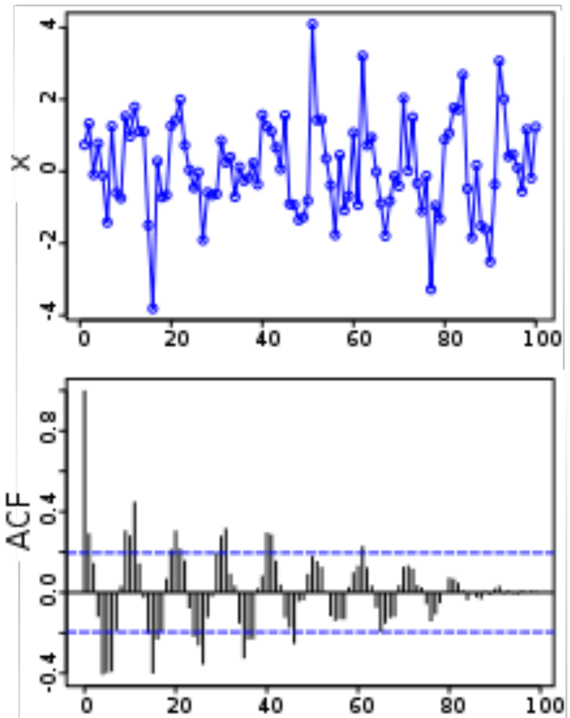


Looking forward for your questions, inputs or remarks …



xkcd.com

Backup slides



Time series regression

- OLS estimates
- Goodness of fit metrics

Variance-Covariance

$$\begin{pmatrix} \sigma^2_{y_1} & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_n} \\ \sigma_{y_1 y_2} & \sigma^2_{y_2} & \cdots & \sigma_{y_2 y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_1 y_n} & \sigma_{y_2 y_n} & \cdots & \sigma^2_{y_n} \end{pmatrix}$$

tsabn as a time series extension of abn

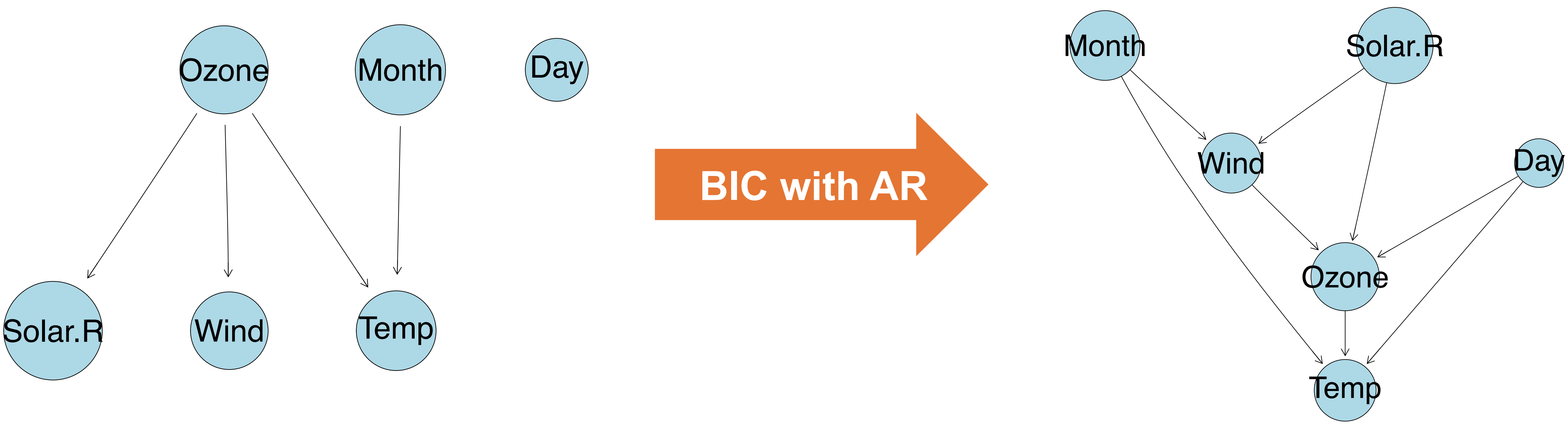
- Extending ABN to correlated errors
- Several implemented scores: AIC, BIC, MDL
- Errors Autocorrelation: ARMA procedure with Autoregressive modelling
  - Kalman filter

Future work

- Implementation of Granger causality score for BN learning

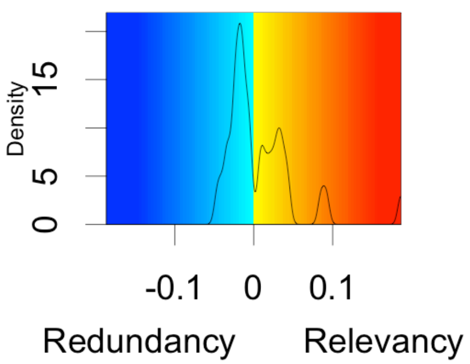
Daily readings of the air quality values from May to September 1973

111 observations on 6 variables



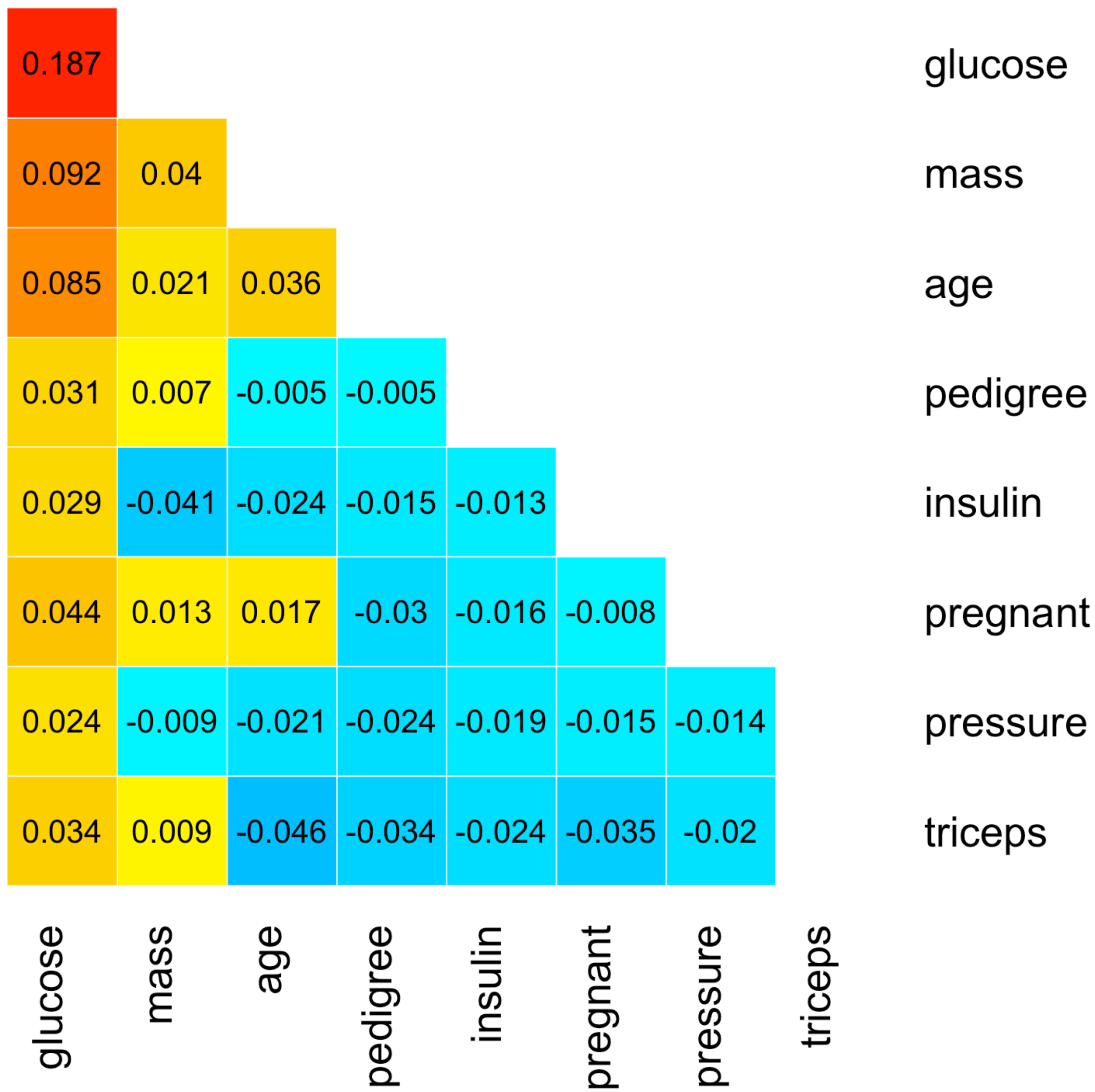
Future work:

Hourly readings of the PM2.5 and 6 other chemical compounds data of US embassy in Beijing with meteorological data from Beijing Capital International Airport from 2013 to 2017



Pima Indians Diabetes Database

768 observations on 9 variables



That is all folks!