Bayes statisztika és Bayes hálók az asztma/allergia kutatásban

Antal Péter
BME
Méréstechnika és Információs Rendszerek Tanszék
Mesterséges Intelligencia Csoport

Tartalom

- Bayes hálók (grafikus modellek)
  - Kronológia
  - Az általános Bayes hálós modell
  - Értelmezési szintek
  - Bayes hálók tanulása
- Bayes statisztika
- Bayes hálók és Bayes statisztika orvosbiológiai kutatásban
  - Genetikai asszociációs kísérletek
  - Fúziós kihívások
- Eddigi alkalmazások asthma/allergia kutatásban

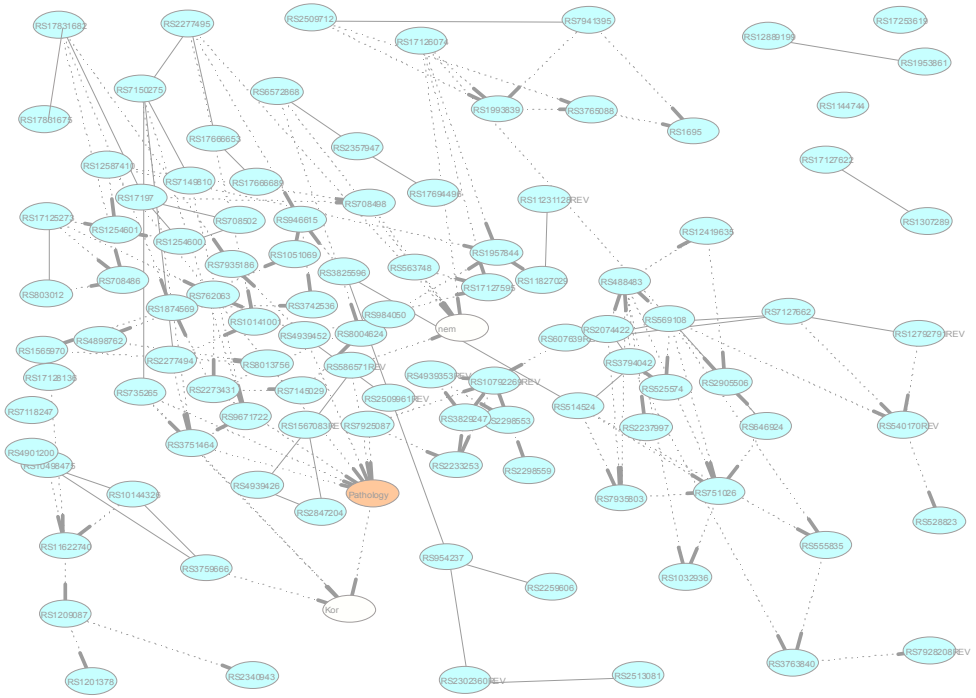
Bayes-*

- **Thomas Bayes (c. 1702 – April 17, 1761),**
  brit matematikus és presbiteriánus lelkész

- Bayes tétel
- Bayes statisztika
- Bayesi döntéselmélet
- Bayes hálók
- Bayes faktor
- Bayes hiba
- Bayes rizikó
- Bayes döntés
- Bayes becslő
- ...

$$p(Model|Data) \propto p(Data|Model)\,p(Model)$$

$$\hat{f} \approx \bar{f} = E_{p(\theta,S|Data)}[\,f(\theta,S)]$$



Történeti áttekintés

- [1921]: S. Wright's diagrams, „Correlation and causation"
- [1970] The first (medical) applications (Dombal).
- [1979] The axiomatization of independencies (Dawid)
- [1982] The decomposition of a distribution (Lauritzen).
- [1988] Representation of independencies (Pearl).
- [1988] Inference methods in polytrees (Pearl).
- [1989] Exact general inference methods (Lauritzen).
- [1990] Closed form for the structure posterior (Cooper, Herskovits)
- [1995] Bayesian inference over BNs (Madigan)
- [1998] Bayesian inference about causation (Heckerman)
- [2000] Bayesian inference about relevance (Friedman)
- [2000] J.Pearl: Causality, Cambridge University Press
- …..

Bayes tétel

Egy algebrai trivialitás

$$p(X \mid Y) = \frac{p(Y \mid X)\,p(X)}{p(Y)} = \frac{p(Y \mid X)\,p(X)}{\sum_X p(Y \mid X)\,p(X)}$$

Egy tudománytörténeti paradigmaváltás

$$p(Model \mid Data) \propto p(Data \mid Model)\,p(Model)$$

És pszichológiai konzekvenciák, megerősítés

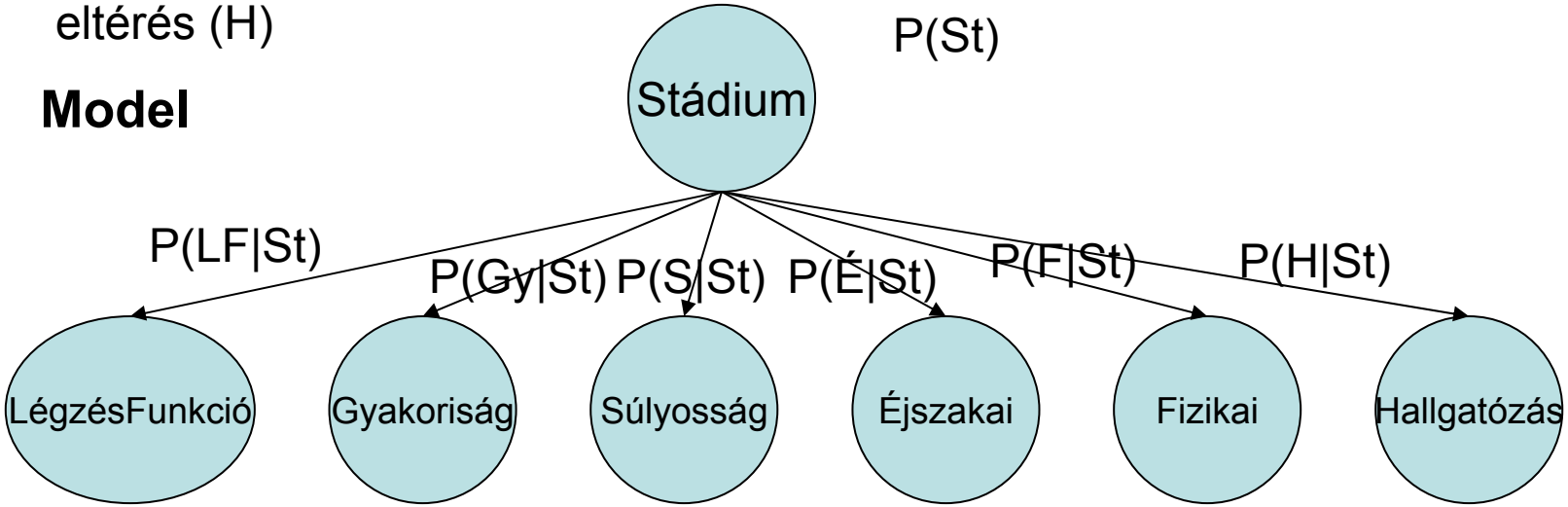
P(betegség|tünet)=???

p(tünet|betegség)=0.79
p(betegség)=0.12

Bayes tételtől a Bayes hálókig: Naív/idióta Bayes háló

Változók:
Asthma-stádium (St), Légzésfunkció(LF), TünetGyakoriság (Gy), TünetSúlyosság (S), ÉjszakaiTünet (É), FizikaiTerhelés (F), Hallgatózási eltérés (H)

Model

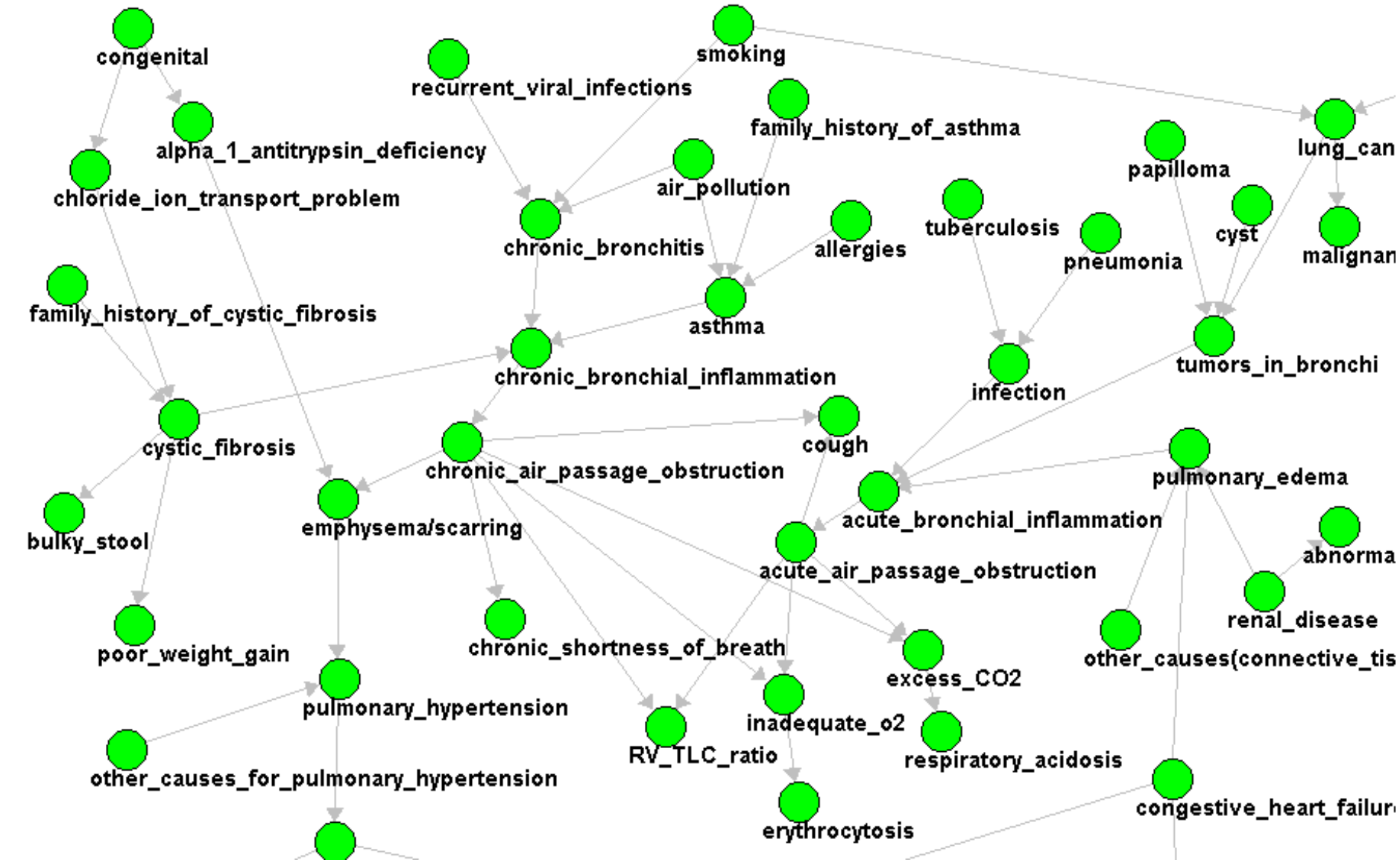


Diagnosztika

$$p(St \mid LF, Gy, S, É, F, H)$$

$$\propto p(St)\, p(LF \mid St)\, p(Gy \mid St)\, p(S \mid St)\, p(É \mid St)\, p(F \mid St)\, p(H \mid St)$$

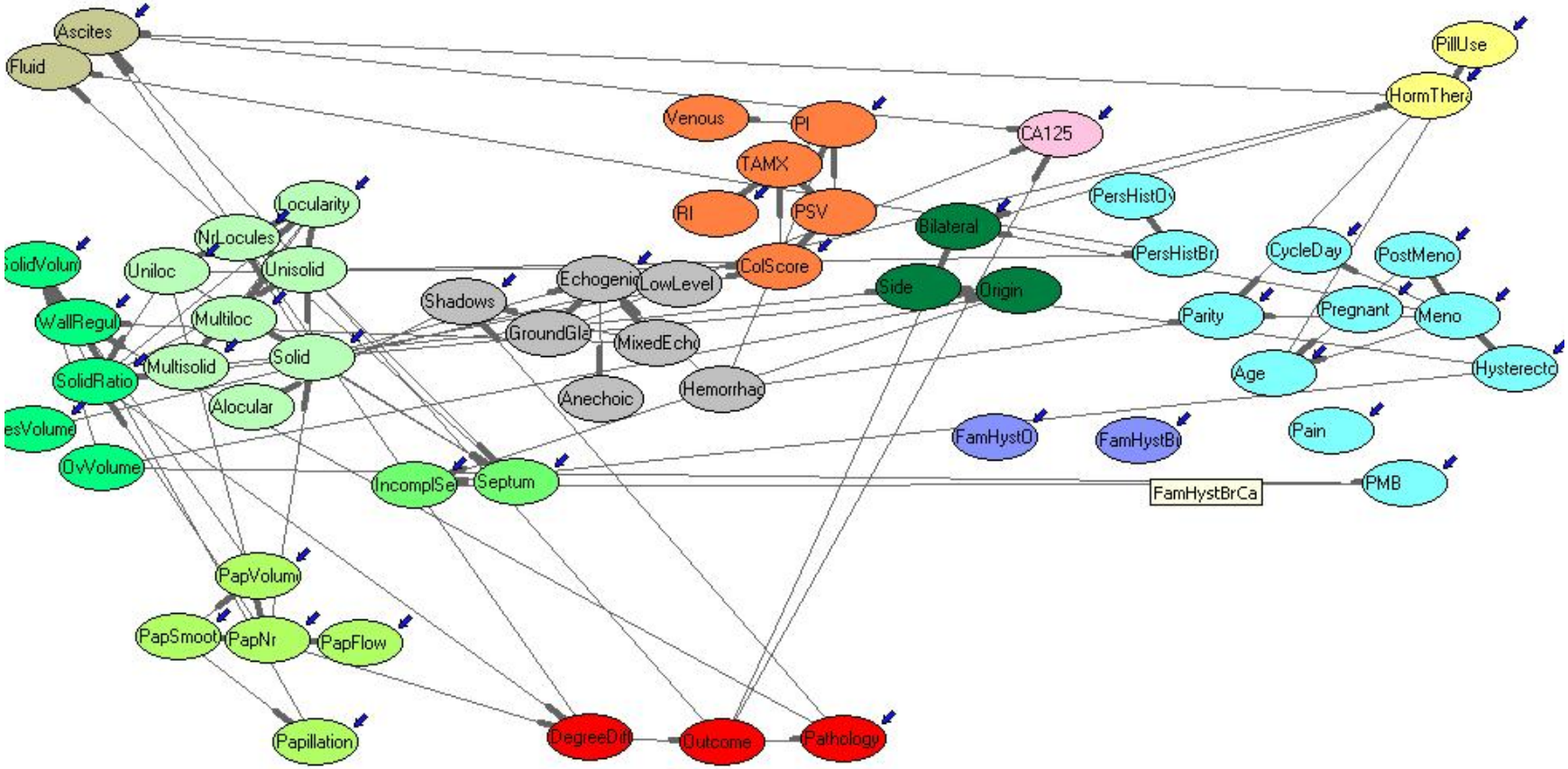
$$= p(St) \prod_i p(Tünet_i \mid St.)$$


congenital
alpha_1_antitrypsin_deficiency
chloride_ion_transport_problem
family_history_of_cystic_fibrosis
cystic_fibrosis
bulky_stool
poor_weight_gain
other_causes_for_pulmonary_hypertension
pulmonary_hypertension
emphysema/scarring
recurrent_viral_infections
smoking
air_pollution
chronic_bronchitis
allergies
asthma
chronic_bronchial_inflammation
chronic_air_passage_obstruction
chronic_shortness_of_breath
RV_TLC_ratio
inadequate_o2
erythrocytosis
cough
family_history_of_asthma
tuberculosis
pneumonia
infection
acute_bronchial_inflammation
acute_air_passage_obstruction
excess_CO2
respiratory_acidosis
papilloma
cyst
tumors_in_bronchi
lung_can
malignan
pulmonary_edema
abnorma
renal_disease
other_causes(connective_tis
congestive_heart_failur

Pathfinder

- Pathfinder system. (Heckerman, Probabilistic Similarity
- Networks, MIT Press, Cambridge MA).
- • Diagnostic system for lymph-node diseases.
- • 60 diseases and 100 symptoms and test-results.
- • 14,000 probabilities.
- • Expert consulted to make net.
- • 8 hours to determine variables.
- • 35 hours for net topology.
- • 40 hours for probability table values.
- • Apparently, the experts found it quite easy to invent the causal links and probabilities.
- Pathfinder is now outperforming the world experts in diagnosis. Being extended to several dozen other medical domains.

És még egy model



Bayes háló értelmezési szintek

- Együttes eloszlás hatékony reprezentálása
  - Hatékony következtetés és tanulás
- Függetlenségi állítások reprezentálása
  - Direkt függések
- Oksági modell reprezentálása
  - „A csomópontok változók,
  - Az élek direkt, okozati relációk”

Eloszlások faktorizációja

- Általános (➔Markov hálók)
  - Potenciálok

$$p(X_1,\ldots,X_n) \propto \prod_{j=1,..,L} \varphi_j(X_{j_1},...,X_{j_{L_j}})$$

- Speciális (➔Bayes hálók)
  - Feltételes eloszlások

$$p(X_1,\ldots,X_n) = \prod_{i=1,..,n} p(X_i \mid X_1,\ldots,X_{i-1})$$

$$= \prod_{i=1,..,n} p(X_i \mid Pa(X_i))$$

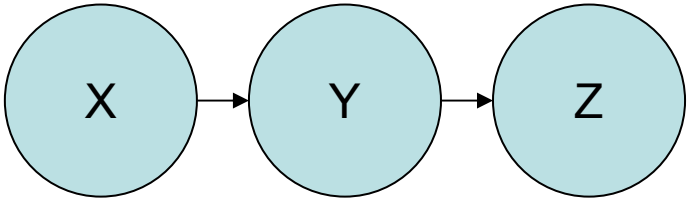
Eloszlások függetlenségeinek gráfos reprezentálása

„A valószínűségszámítás nem a számokról szól”

- Jelölje $I_p(X|Z|Y)$, hogy X,Y függetlenek Z feltétellel p eloszlásban
- Az M_p függetlenségi modell tartalmazza az összes p-beli függetlenségi állítást
- Kérdések
  - Lehet-e minden M_p-t egy gráffal reprezentálni? (nem)
  - Pontosan melyiket lehet? (nem ismert)
  - Mi a legjobb „közelítés”? (problémafüggő)

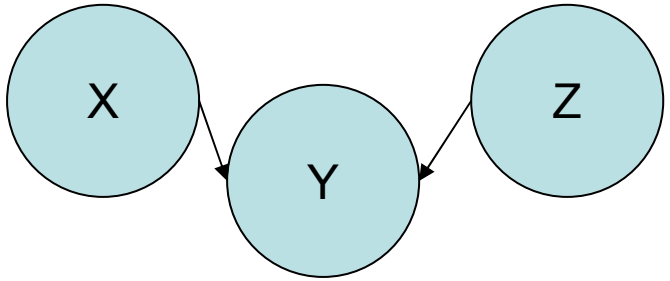
Gráffal nem reprezentálható eloszlások/függetlenségi modellek

- Intranzitív függés (pl. X és Z lehet független)



- Alsóbbrendű függetlenségből nem következik magasabbrendű

  (X,Z)-tól együtt függhet Y, mégha páronként független is



A függetlenségi modell vs gráfok II.

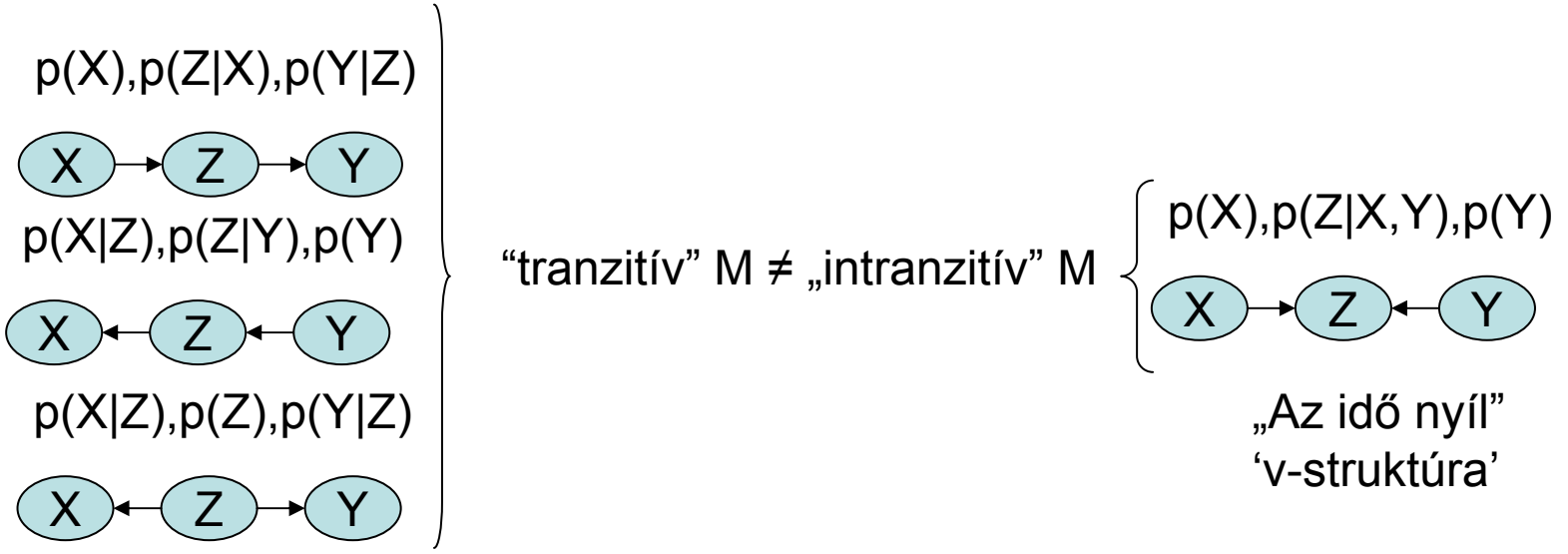
- Jó hírek:
  - Véges karakterizáció egzakt Markov hálós reprezentációra
  - D-elválasztottság helyes és teljes kalkulus
- Sejtések
  - Nincs véges karakterizáció egzakt Bayes hálós reprezentációra
- Rossz hírek
  - Léteznek egzakt módon nem reprezentálható eloszlások (sem Markov, sem Bayes hálókkal)
  - A reprezentáció redundáns ➔oksági értelmezés???

Principles of causality

- strong **association**,
- X precedes **temporally** Y,
- plausible **explanation** without alternative explanations,
- **necessity** (generally: if cause is removed, effect is decreased or actually: y would not have been occurred with that much probability if x had not been present),
- **sufficiency** (generally: if exposure to cause is increased, effect is increased or actually: y would have been occurred with larger probability if x had been present).

Függetlenségi vs. oksági modell

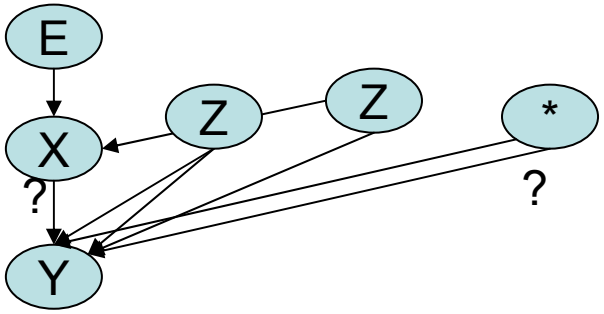
BAJ: egy függetlenségi modellhez több Bayes háló is tartozhat



Oksági Bayes hálók

„Correlation≠causation”, „no cause in, no cause out”

„Csak oksági kapcsolatok” feltevés: több változó esetén bizonyos élek egyértelműek
Több változó esetén bizonyos esetekben még rejtett változók is kizárhatók



KONKLÚZIÓ1: tabula rasa oksági következtetés lehetséges

- K2: oksági következtetések spektruma a tárgyterület független feltevések függvényében
  - Csak oksági relációk
  - Rejtett változók
  - Modell minimalitás
  - Változószám

„Az okozatiság matematizálása”
„a statisztikai idő iránya”

Feladatok Bayes hálóknál

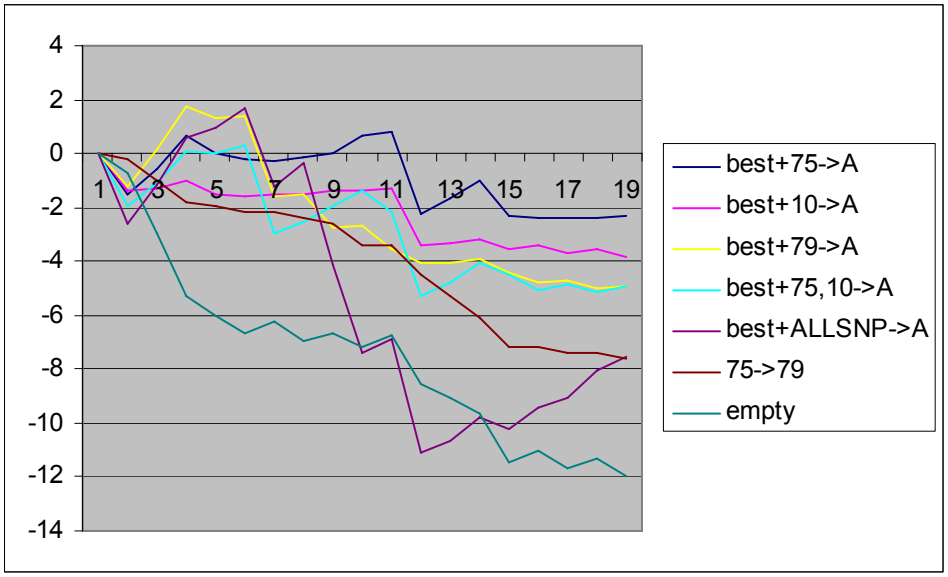
- (Passzív megfigyeléses) következtetés
  - P(Cél-változók|Megfigyelt-változók)
  - Legvalószínűbb értékkonfiguráció
- Beavatkozásos következtetés
  - P(Cél-változók|Megfigyelt-változók, Beállított-változók)
- Kontrafaktuális következtetés
  - P(Cél-változók|Megfigyelt-változók, Átállított-változók)
- Döntési háló, hasznosság és beavatkozási változók
  - További információ értéke
  - Maximális hasznú döntés
- Paraméter tanulás
- Struktúra tanulás
  - ➔BNET 1995-

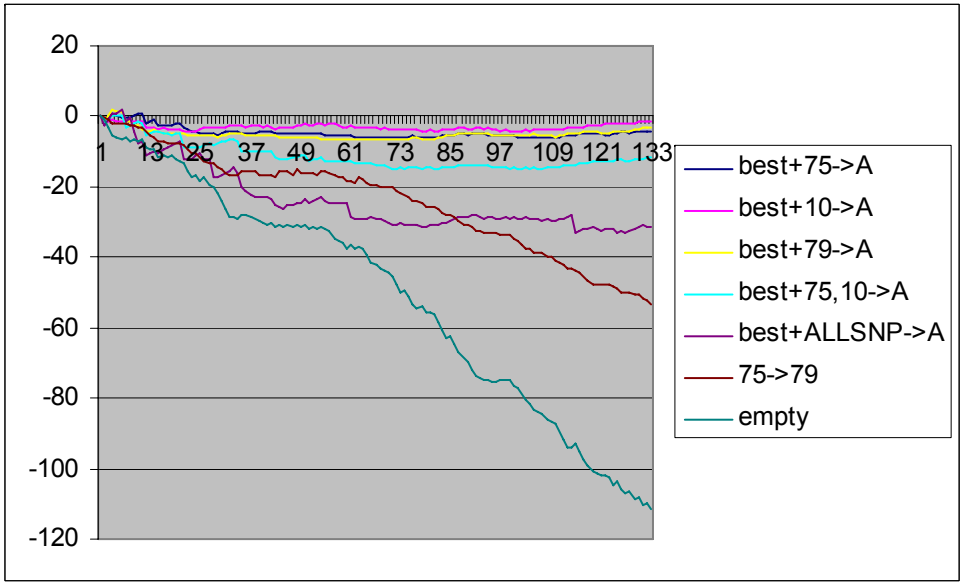
Bayes hálók tanulása

- Hipotézistesztéléssel
  - Függetlenségi tesztek diktálják a hálót
- Bayes statisztika esetén

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)$$

Változók: SNP75, SNP10,SNP79, Asthma („best model”: 79<-75->10)





Bayes háló tulajdonságok

- Ötlet: hipotézisek, mint Bayes háló jegyek
- Ötlet^2: (modellalapú) posterior a hipotézisekre

$$p(ModelTulajdonság \mid Data)$$

$$= \sum_m p(m \mid Data) 1(Tulajdonság - igaz - m - ben)$$

$$= E_{p(m|Data)}[1(Tulajdonság - igaz - m - ben)]$$

- Génregularizációs kísérletben
  - Páronkénti oksági kapcsolatok: független, közös ok, oksági kapcsolat (direkt/indirekt)
- Asszociációs kísérletben
  - Relevancia???

Feature learning

Feature posterior

$$P(F = f) = \sum_{G:F^G = f} P(G)$$

- Goal: approximate the full-scale summation (integral)
- Practical methods: MCMC (Markov Chain Monte Carlo) sampling

A kapcsolt BN jegyek relevanciára

– **Markov Blanket (sub)Graphs (MBGs)**

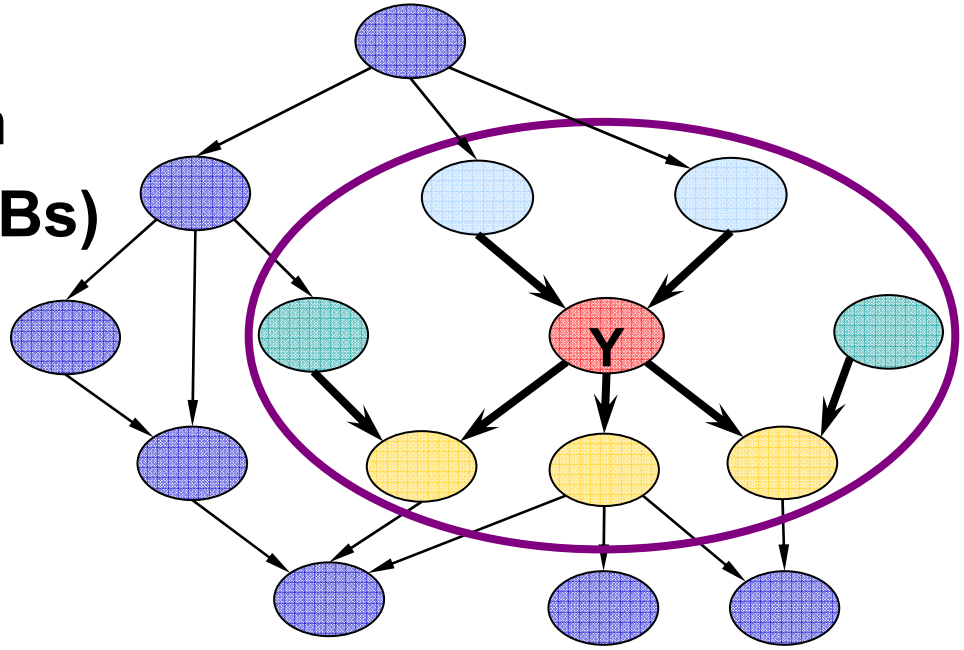
(1) parents of the node

(2) its children

(3) parents of the children

Markov Blanket Sets (MBs)

- the set of nodes which probabilistically isolate the target from the rest of the model



– **Markov Blanket Membership (MBM)**

- pairwise relationship

Kapcsolt, eltérő absztrakciós szintű hipotézisosztályok

- **Az asszociációs tanulmányok kérdései Bayes hálók strukturális jegyeivel formalizálhatók**
  - Páronkénti asszociáció: Markov Blanket Memberhsips (MBM)
  - Többváltozós/multifaktoriális asszociáció: Markov Blanket sets (MB)
  - Multifaktoriális asszociáció interakciókkal: Markov Blanket Subgraphs (MBG, C-RPDAG)
  - Teljes interakciós modell: Részlegesen irányított Bayes háló (PDAG)
  - Teljes oksági modell:Bayes háló (BN)
- **Kapcsolt, absztrakciós szinteket** alkotnak ezek az asszociációs (feltételes) jegyek
  - DAG=>PDAG=>MBG=>C-RPDAG=>MB=>MBM

Adatoktól az orvosbiológiai ígéretekig

- Populációs adatok (családfa)
- Klinikai adatok
- Genetikai profil
  - Egyed
  - Szövet
- Genomikai profil
  - Gén-expressziós profil
  - mikroRNS profil
- Proteomikai profil

Statisztikai adatelemzés

Tudásmérnökség

Rendszermodellek

- Génregularizációs hálók
- Útvonal elemzés
- Modell alapú gyógyszertervezés

- Személyre szabott
  - prevenció
  - szűrés
  - diagnosztika
  - gyógyszer
  - kezelés
- (Nép)betegségek
  - „biológiai” szintű megértése
  - kezelése
  - Pl. malária, TBC, asztma, allergia, csontritkulás, elhízás,...

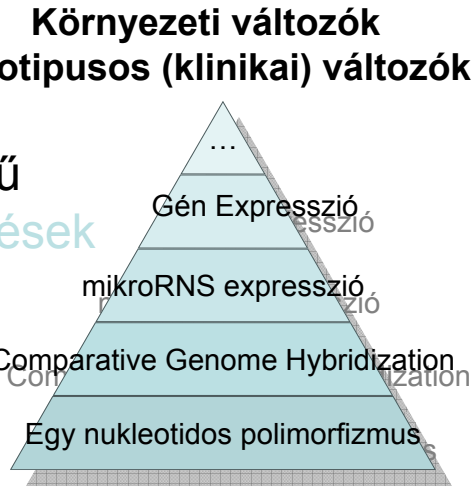
Bioinformatika: tudás és adat fúzió

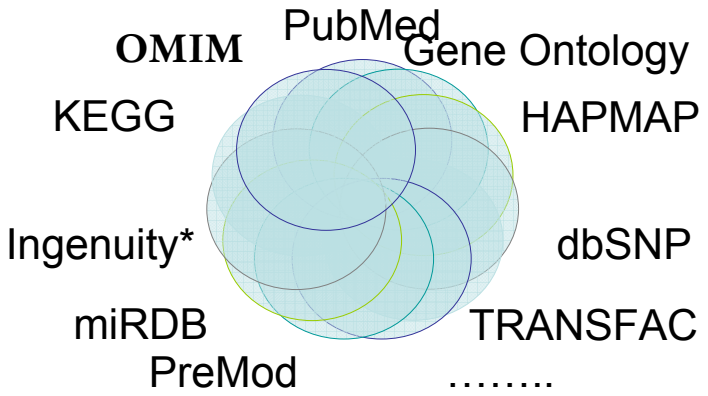
Nagy dimenziójú adatok

Nagy mennyiségű, elektronikus tudás

Nagy áteresztőképességű molekuláris biológiai mérések

- DNS chip
- (mikroRNS chip)
- CGH chip
- SNP chip



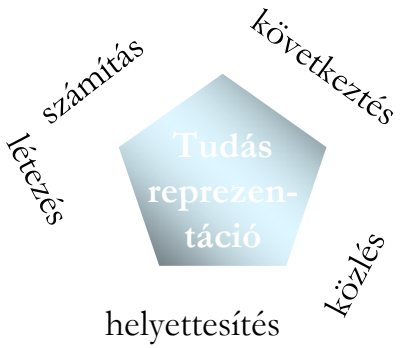


Komplex, hierarchikus, tudásgazdag hipotézisek

- Asszociációk
- Multifaktoriális interakciók
- Okozati relációk
- Alrendszerek, rendszerek

Tudásmérnöki trendek (fúzióhoz)

- Szabadszöveges tudástárak
  - Pl. Pubmed 15 millió absztrakt
  - Cikk állítása formálisan
- Szemi-formális tudásbázisok
- Elektronikus ontológiák, taxonómiák
  - UMLS, MeSH, Gene Ontology, MGED ontológia, SNP ontológia
- Formális tudásbázisok
- **Valószínűségi tudásmodellek/logikák**
- Szemantikus világháló
- Annotált, integrált adatbázisok
- Annotált, integrált web-szolgáltatások



Tudásmérnöki feladatok:
- reprezentáció,
- kinyerés,
- következtetés,
- fenntartás

Enciklopédisták (D.Diderot)
Logikai pozitivizmus/empirizmus (R. Carnap)
"World Brain" (H.G.Wells,1938)

Elektronikus, hatalmas mennyiségű, egységesített, valószínűségi, empirikusan lehorgonyzott, formálisan reprezentált tudás

Indukciós trendek (fúzióhoz)

- Statisztikai, számításelméleti tanulási elméletek
  - Nem-parametrikus konzisztencia eredmények
  - Nem-aszimptotikus konvergencia eredmények
  - ...
- Bayes statisztika
  - Markov Lánc Monte Carlo módszerek (MCMC)
  - Hibrid/Hierarchikus/Kapcsolt/... MCMC-ek
  - ..
- Oksági következtetés
  - Oksági kapcsolatok indukciója passzív megfigyelésekből
  - Oksági kapcsolatok formalizálása (kauzális Bayes hálókkal/ strukturális egyenletekkel)
  - Megfigyelések és beavatkozások tervezhetősége (kiváltása)
  - (Beavatkozások hatásának valószínűségi modellezése
  - Kontrafaktuális következtetések)

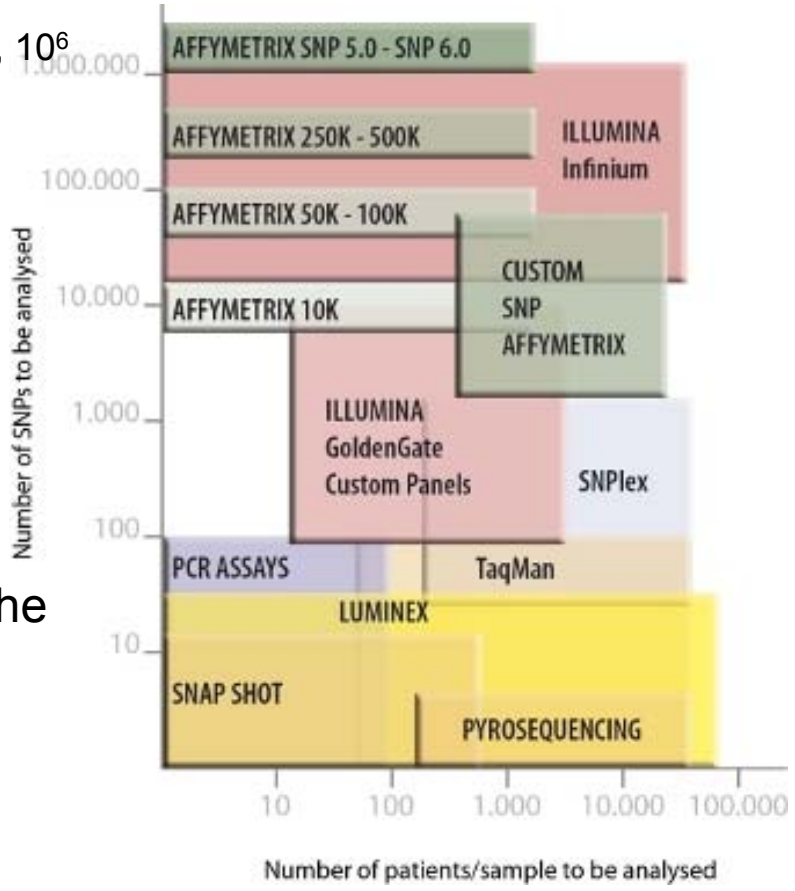
W.Ockham
D.Hume
K.R.Popper

T.Bayes
P.Laplace

D.Hume
H.Reichenbach

Genetic Association Studies (GAS): a statistical view

- Genome-wide association studies (GWASs)
  - Sample size: ~1000
  - |Response/outcome (target) variables|: ~10
  - Predictor/explanatory:
    - |SNPs| (nominal/numeric, unphased genotype): ~10^5, 10^6
    - |Environmental variables|: ~10
  - Cost: 1000x1000 Euro
- Partial genome screening studies (PGSSs)
  - Sample size: ~1000
  - |Response/outcome (target) variables|: ~10
  - Predictor/explanatory:
    - |SNPs| (nominal/numeric, unphased genotype): ~10^2
    - |Environmental variables|: ~10
  - Cost: 10000 Euro
- The goal is the exploration of…
  - the relevance of explanatory variables w.r.t. the target variables
  - the interdependence of explanatory variables



The genomic background of asthma

SU/DGCI: András Falus, Csaba Szalai, Ildikó Ungvári

Sample size: 1100

SNPs: 46 (150)

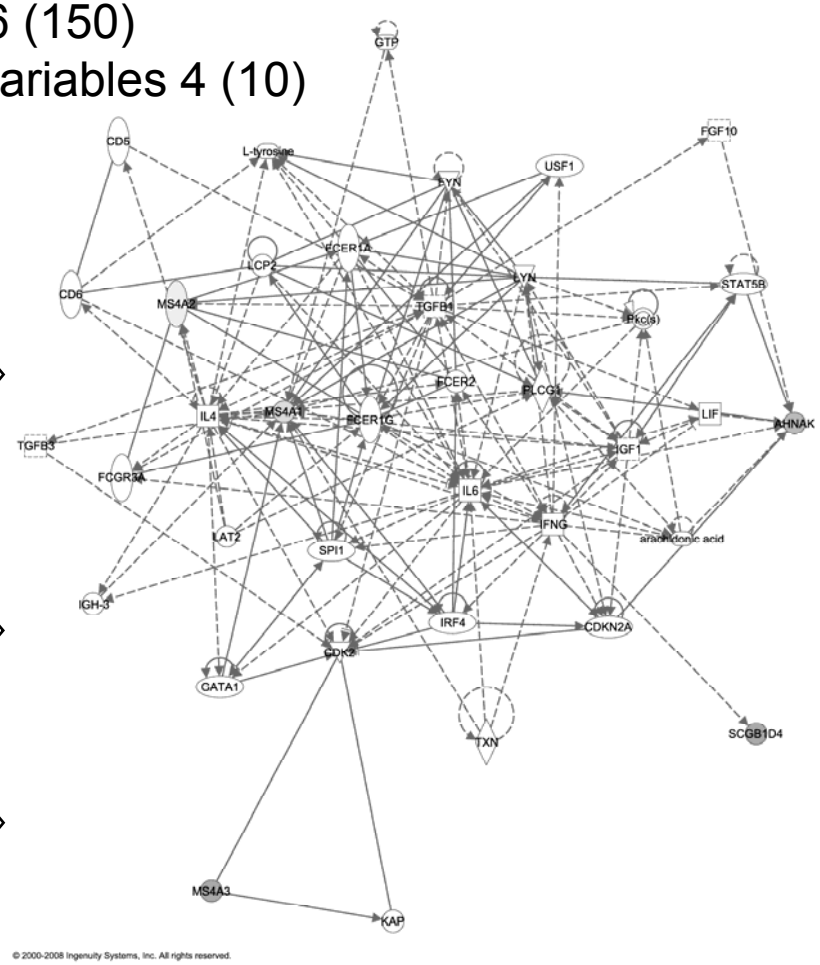
Clinical variables 4 (10)

Clinical data

Genetic (SNP) data

Literature

Expertise



Pairwise association

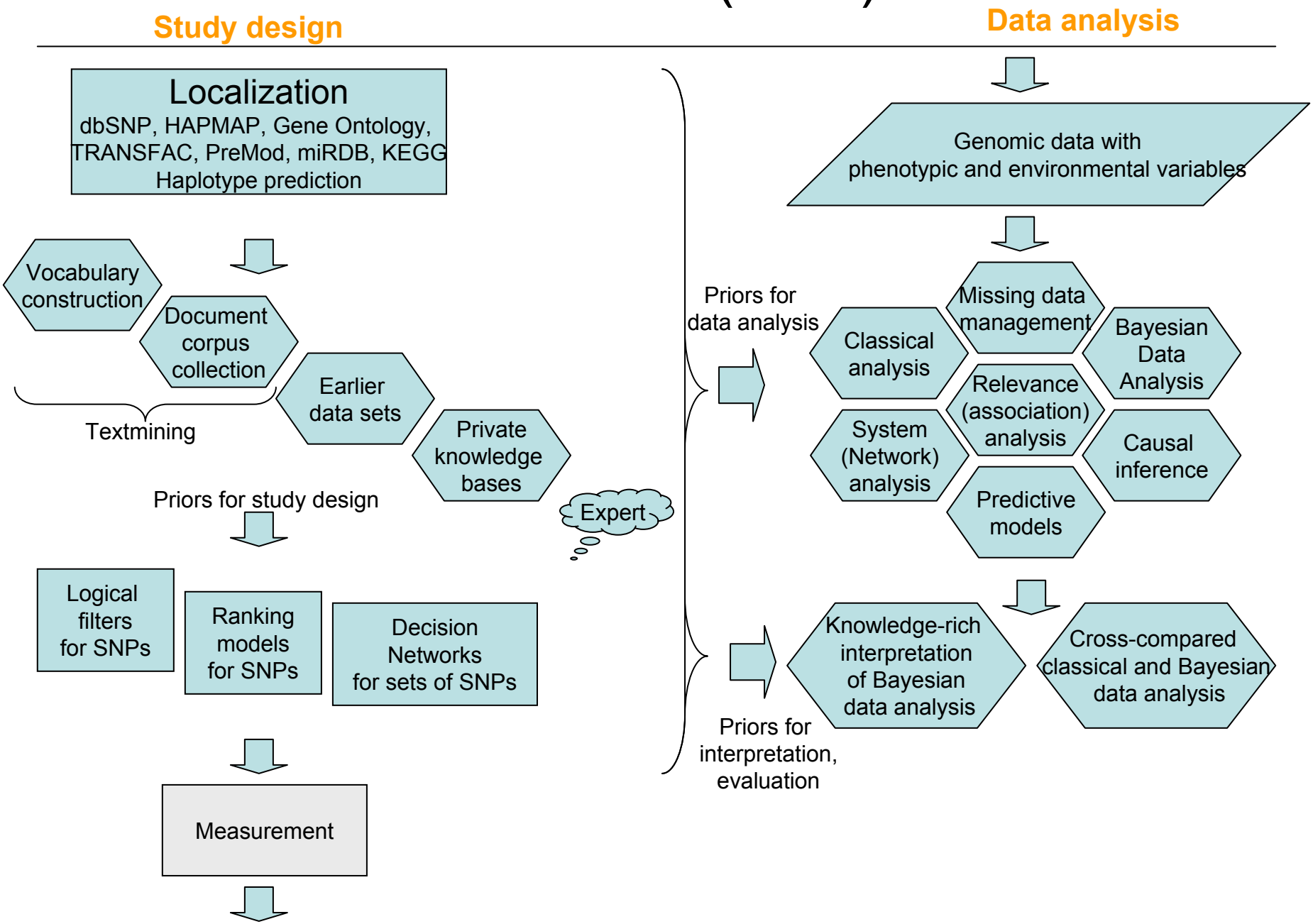
Multivariable analysis

Multivariable analysis with interactions

Complete dependency models

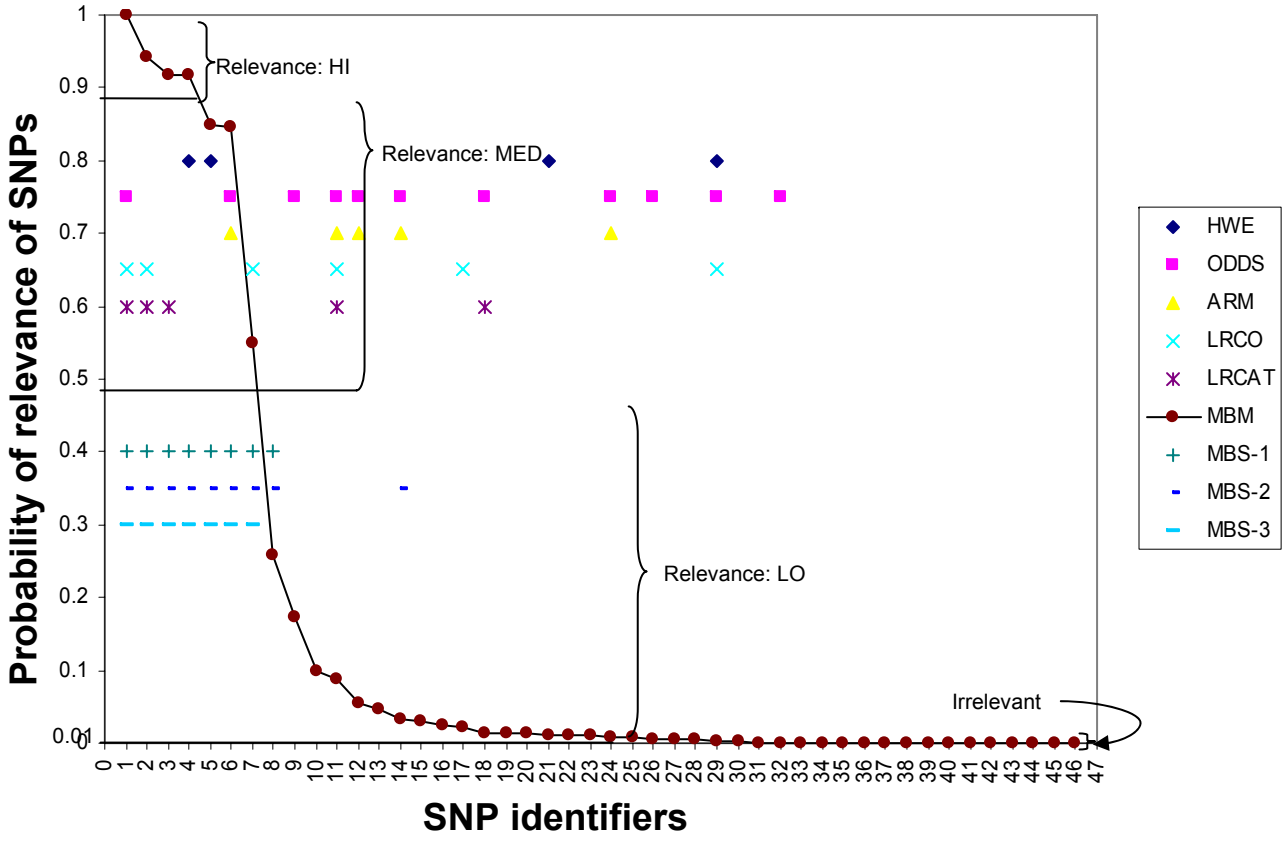
Complete causal models:

Overview (SNP)



An overview of results

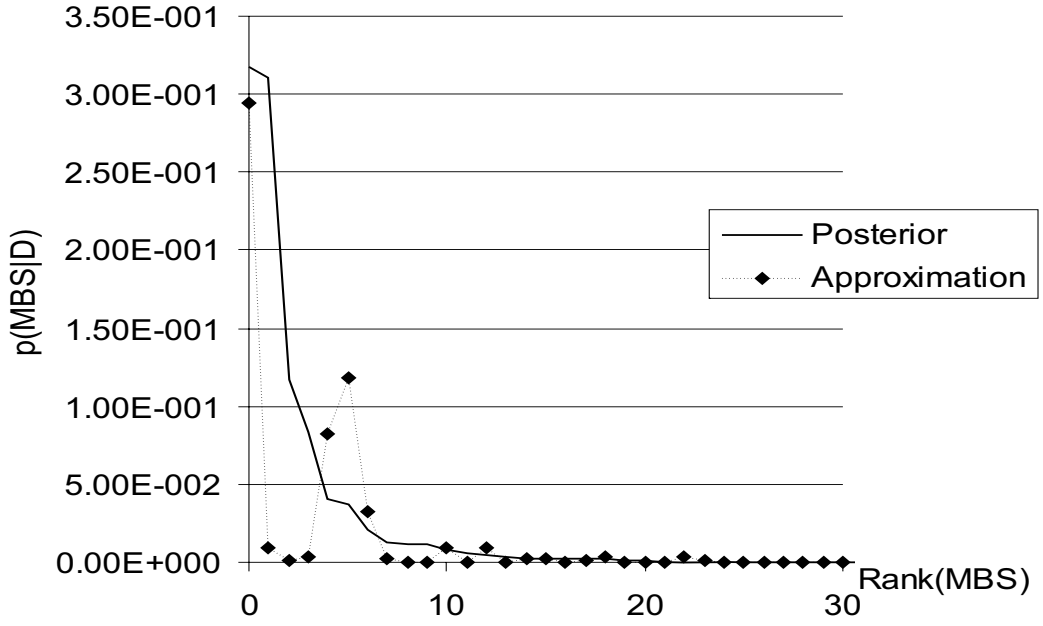
HWE – Hardy-Weinberg equilibrium test, *ODDS* – odds ratio, *ARM* – Cochran-Armitage trend test, *LRCO* – logistic regression (continuous case), *LRCAT* – logistic regression (categorical case), *MBM* – Bayesian pairwise relevance, *MBS-1–9* relevant sets by Bayesian analysis. (only MBM values are numeric, others are arbitrary values for visualization)



Uncertainty/sample complexity in multivariate analysis?

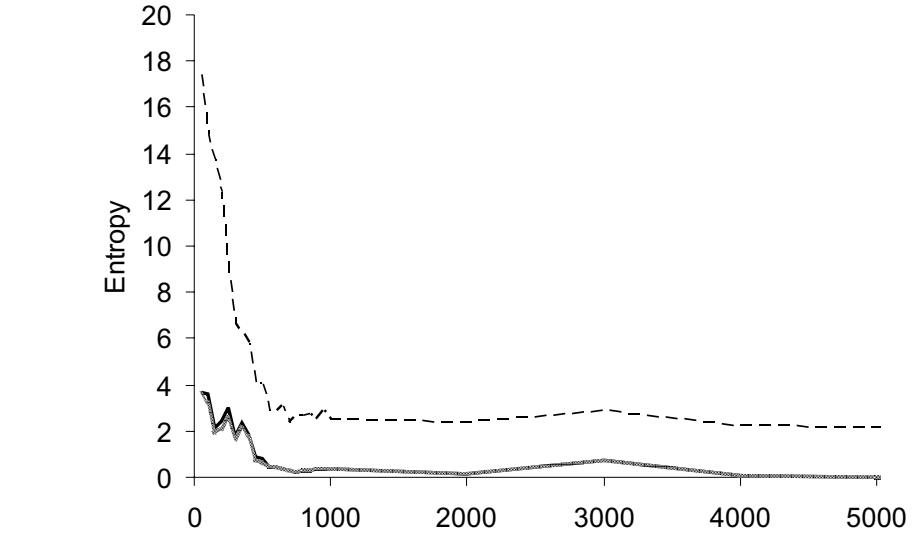
Lessons from the applications of the BMLA

- High-level of uncertainty in multivariate analysis
- There are stable sub-parts (e.g., subset, subgraphs)
- Results for target variables and for certain SNPs could be aggregated

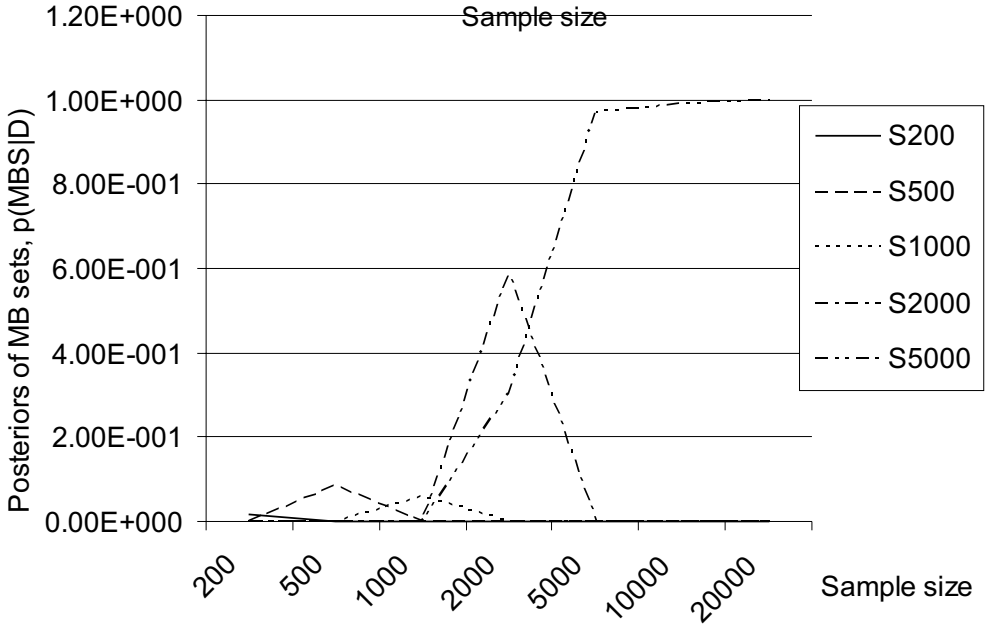


The peakness of the posteriors of the most probable MB sets and their MBM-based approximations.
(46 variables, 1000 samples)

“Sample complexity” for FSS



The summed entropies of the p(MBM(Y,X_i,G)|D) posteriors, and the entropy of the MBS and MBG posteriors for the artificial model (with 7+1 variables).



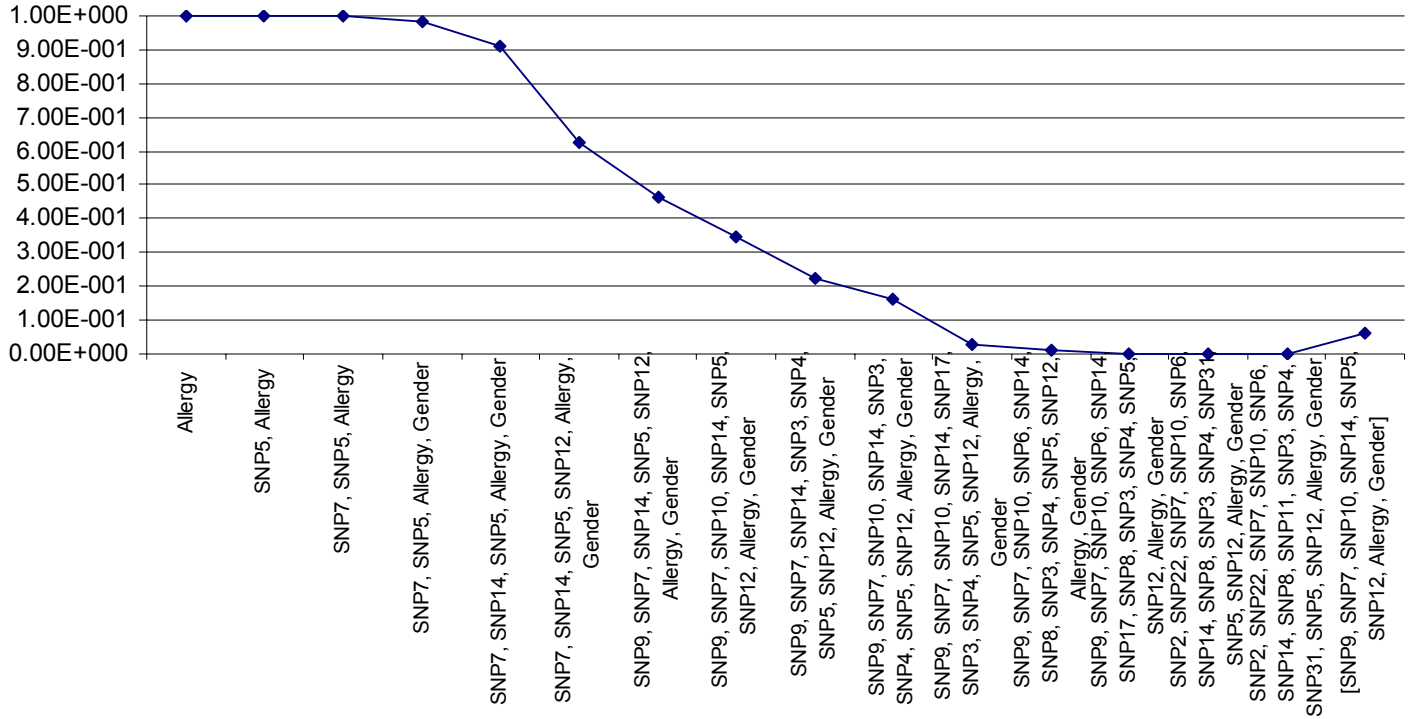
The sequential posteriors of relevant MB sets for the model with 50 variables.

Scalable multivariate analysis

Problem: the "polynomial"gap between simple and complex features
(e.g., MBM (n^2) and MBS (2^n))

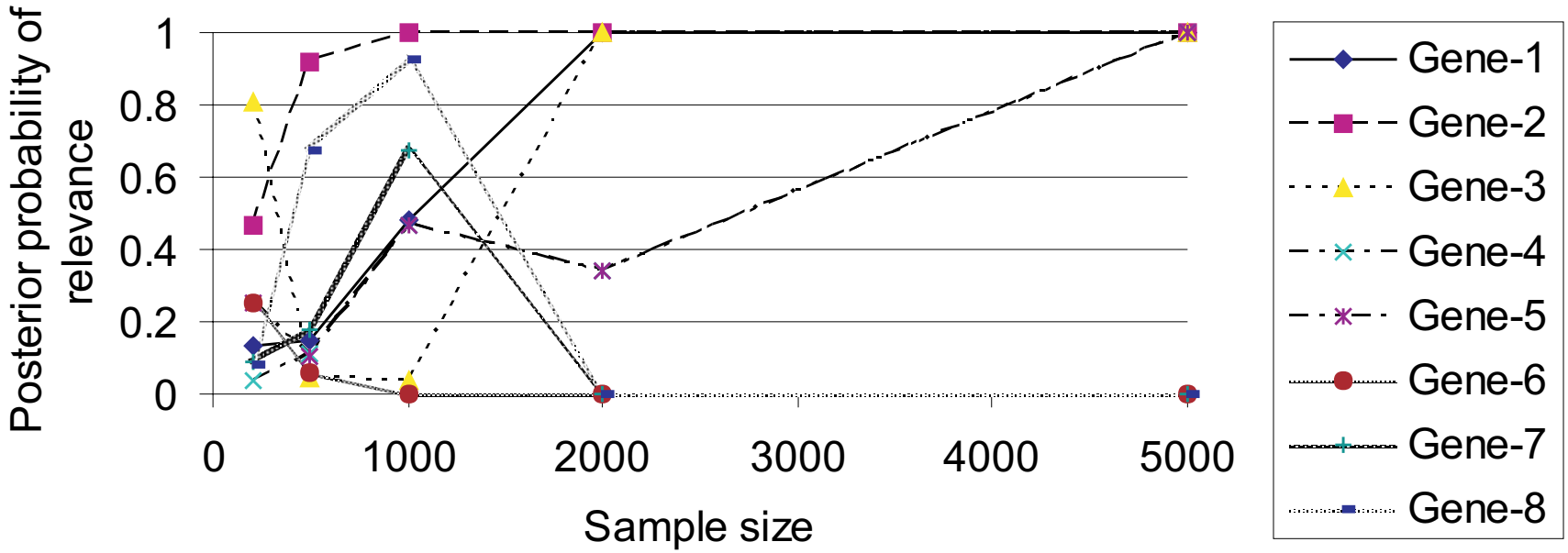
Idea: If all X_i in set S with size k are members of a Markov Boundary set, then S is called a k-ary Markov Boundary subset ($O(n^k)$).

$$p(G : X \subseteq MBS(G) \mid D_N)$$



The maximal posteriors for k-relevance/k-MBS for Asthma.

SNP-to-gene aggregation



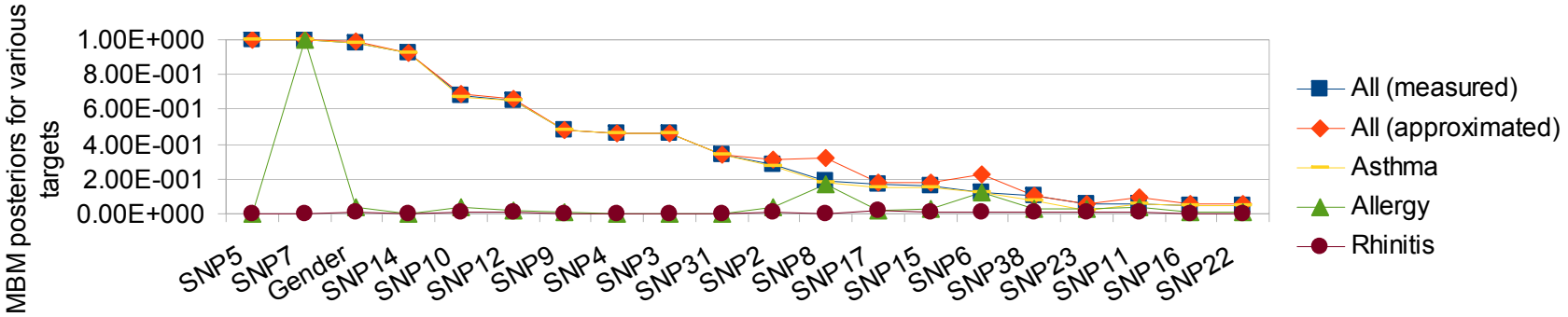
The sequential posteriors that a given gene contains a SNP relevant for asthma

Abstraction levels: SNP, haplo-block, gene,..., pathway

Note that it is different from aggregated multi-variables.

Joint analysis for multiple targets

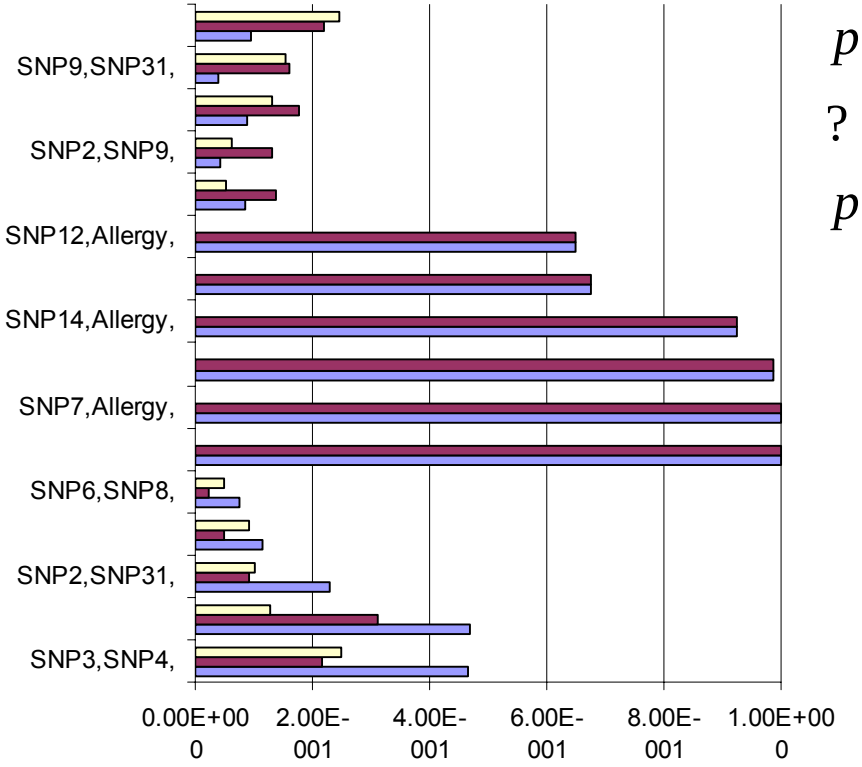
Targets (Outcome): Asthma, Allergy, Rhinitis



The MBM posteriors that a SNP is relevant for asthma, allergy and rhinitis, both separately and jointly.

Note that it cannot be done off-line (contrary to k-MBS/k-MBG and feature aggregation).

Redundancy and interaction



$$p(G : X_i, X_j \subseteq MBS(G) \mid D_N)$$

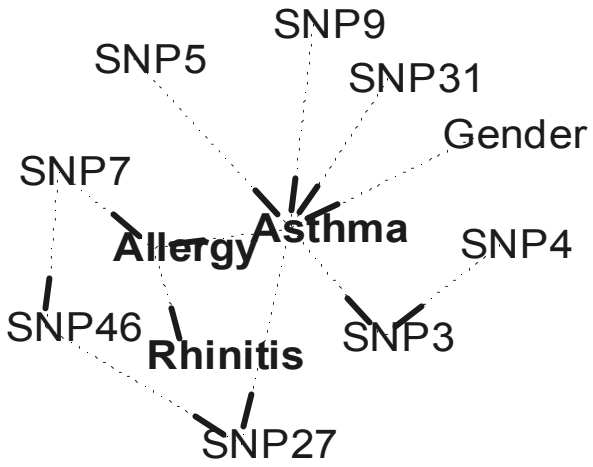
?

$$p(G : X_i \in BS(G) \mid D_N)\, p(G : X_j \in MBS(G) \mid D_N)$$

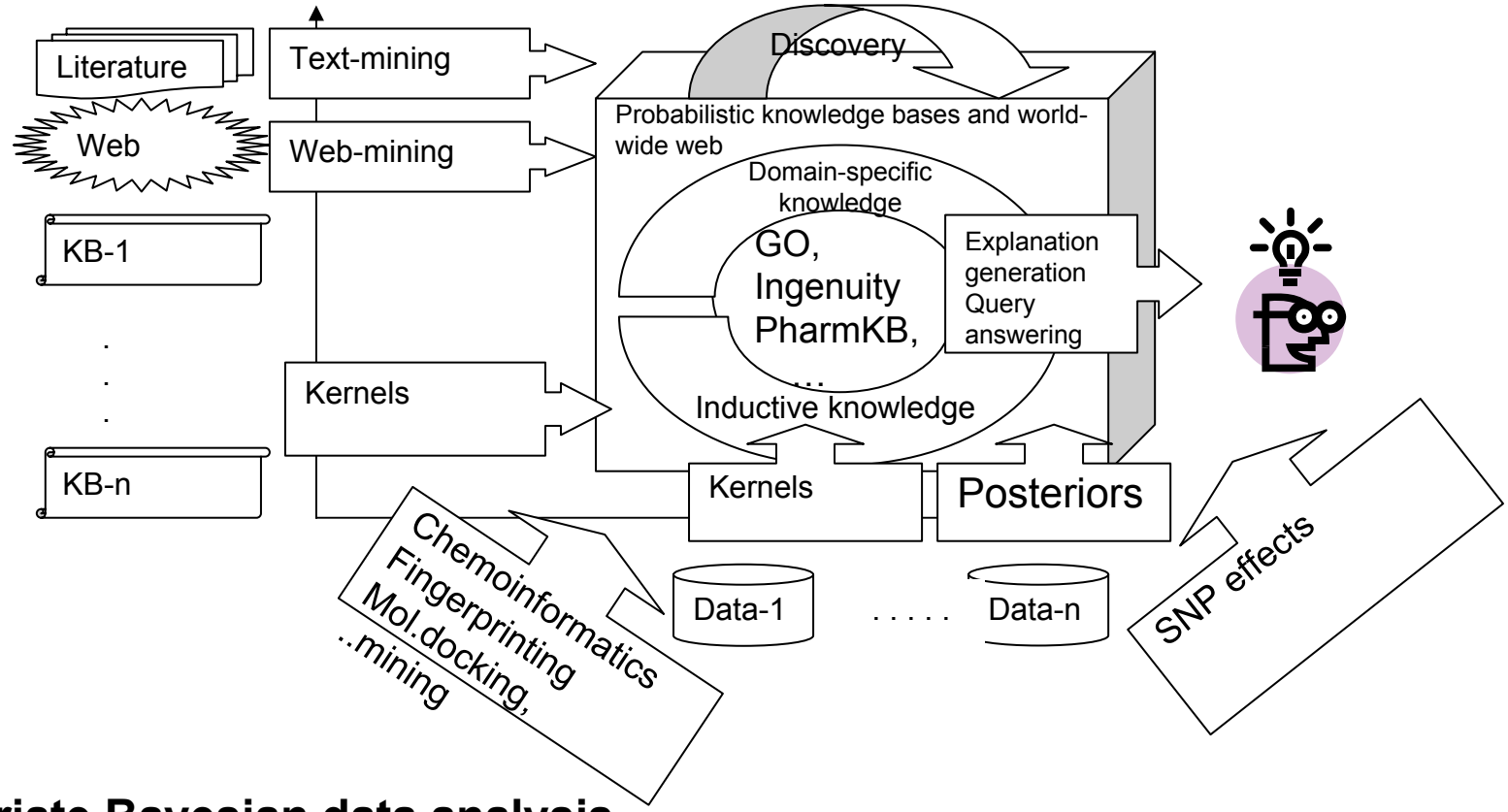
The quantification of interaction and redundancy by the decomposability of the posterior.
Single measure (instead of many Bayes factors).

Conditional (and contextual) relevance

The maximimum a posterior Markov Blanket subgraph



Fusion using probabilistic logic



- **multivariate Bayesian data analysis**,
- the use of more powerful **logic for representing knowledge**,
- **uncertainty management** in knowledge representation, induction, and inference,

syntactic semantics: $P(\phi \mid KB) = \sum_{\text{pKB}} P(\phi \text{ is provable} \mid pKB, lKB, kKB)\, p(pKB)$ $P(\phi \mid KB) = \sum_{\text{pKB}} P(\phi \text{ is provable} \mid pKB, lKB)\, p(pKB)$

Diákok, kollégák

- Végzett diákok
  - **Ercsényi** Dániel: Bayes hálók strukturális tulajdonságainak vizsgálata MCMC módszerekkel (2005)
  - **Gézsi** András: Monte Carlo módszerek Bayes hálók komplex strukturális jegyeinek következtetésére (2006)
  - **Barna** Gergely: Bayes-hálók dekomponált tanulása a priori tudás felhasználásával (2005)
  - **Szabadka** Zoltán (Grolmusz Vince): Fehérje-kismolekula kölcsönhatások vizsgálata molekuláris dokkolással (2006)
  - **Iván** Gábor(Grolmusz Vince): Protein-ligand komplexek kötőhelyeinek vizsgálata adatbányászati algoritmusokkal (2006)
  - **Orbán** György**:** Kontextuális függések szerepe Bayes hálók tanulásában (2006)
  - **Rohla** Tibor: Monte Carlo módszerek Bayes hálók egyszerű strukturális jegyeinek következtetésére (2007)
  - **Palotai** Robin (Csermely Péter): Modulkeresés, klaszterezés fehérje-fehérje hálózatokban
  - **Temesi** Gergely (Falus András): Információ menedzsment és intelligens adatelemzés az SNP analízis támogatására
  - **Gergely** Balázs: Statisztikus szövegbányászat
- Jelenlegi diplomázó, orvosbio. másoddiplomázó, és doktorandusz diákok
  - **Millinghoffer** András: A first-order probabilistic logic for fusing data and prior knowledge
  - **Hullám** Gábor: Valószínűségi tudásmodellek tanulása
  - **Hajós** Gergely: Döntéstámogató keretrendszer SNP asszociációs kísérletek tervezéséhez
  - **Kecskeméthy** Péter (Oxford): Bayesi relevancia elemzés feltételes modellekkel
  - **Huszár** Ferenc: nem-parametrikus Bayesi modellek a kognitív pszichológiában
  - **Erdélyi Boróka**: Bayesi logisztikus regresszió
  - **Szabados** Péter: orvosbiológiai tudásbázisok
  - **Arany Ádám**: Bayesi térszerkezet predikció

Köszönöm a figyelmet!