A probabilistic logic incorporating posteriors of hierarchic graphical models

András Millinghoffer, Gábor Hullám and Péter Antal

Department of Measurement and Information Systems
Budapest University of Technology and Economics

Motivation and background

Fusion of factual knowledge and complex posteriors

The „Most Probable Sentences” problem

Applications

Summary

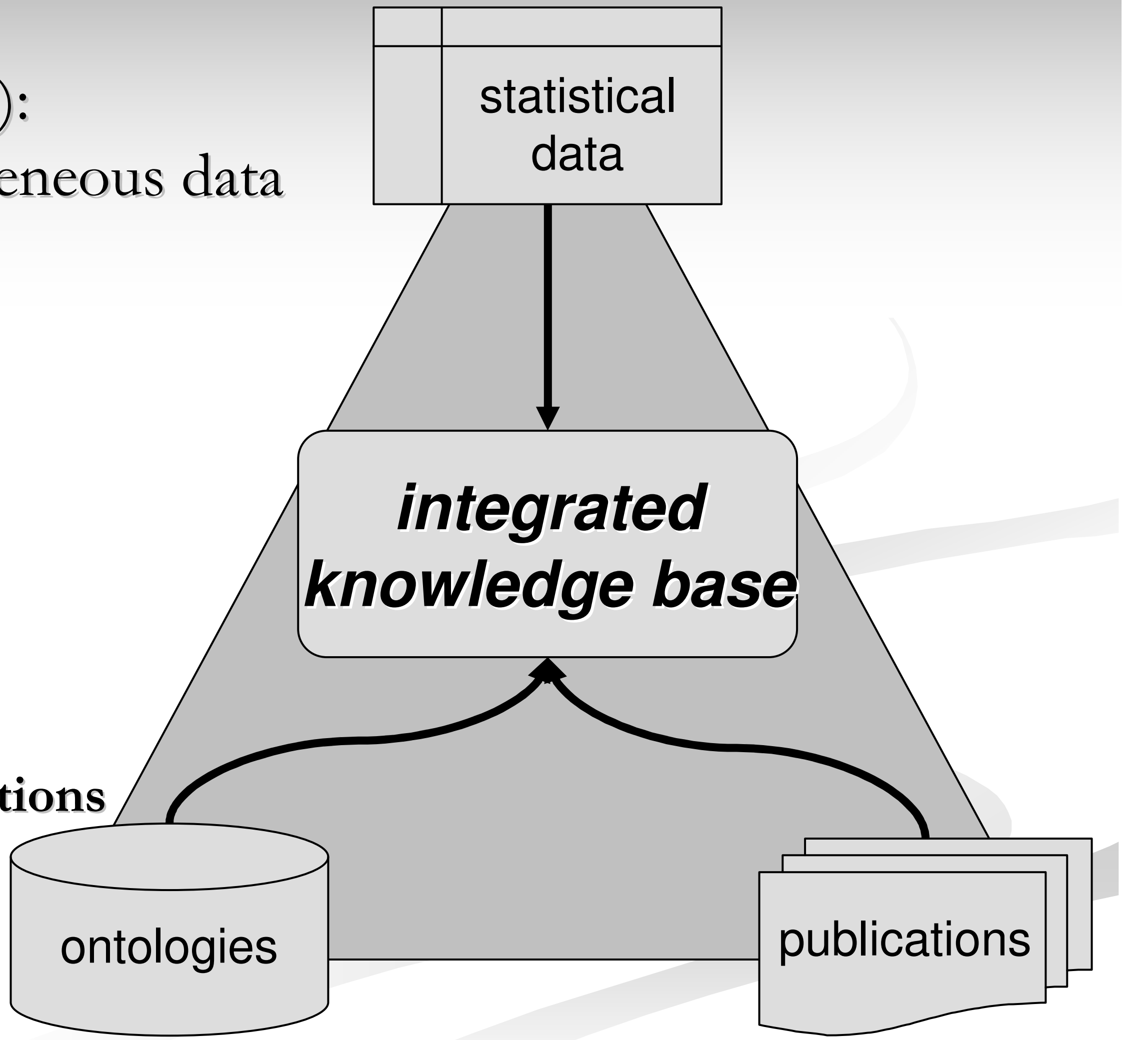
Motivation

Background (e.g. biomedicine):
Rapidly accumulating heterogeneous data

Uncertain (statistical) sources
- **Clinical observations**
- Gene activity measurements
- Expert-defined models

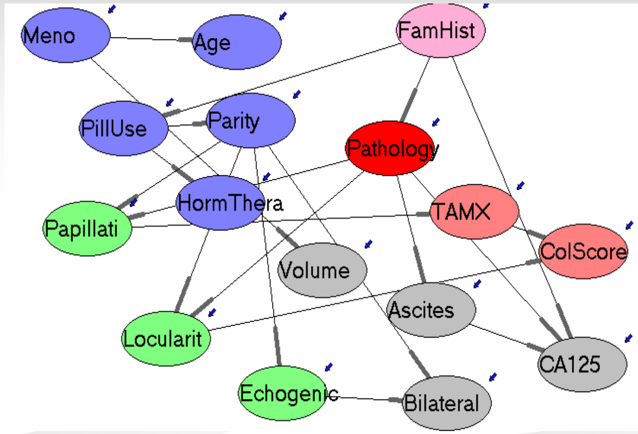
Factual knowledge
- Domain ontologies
- **Natural-language publications**

Goal:
- Fusion



The model class: Bayesian networks

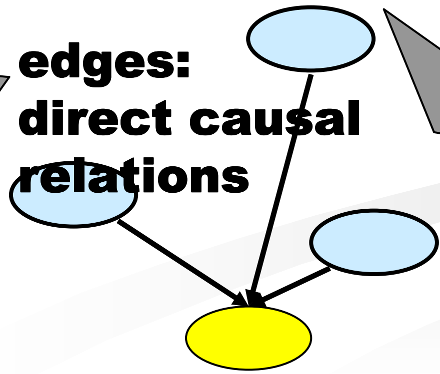
- directed acyclic graph (DAG)
  - nodes – domain entities
  - edges – direct probabilistic relations
- conditional probability models P(X|Pa(X))
- interpretations:



	effective representation of the distribution

$$P(X_1,..X_N)=\prod_{i=1}^{N} P(X_i \mid Pa(X_i))$$

DAG structure: dependency map (d-separation)



edges: direct causal relations

Bayesian statistics and inference

Knowledge representation:

- set of models (feature values)
- distribution over them

Learning (predictive inference):

$$P(G \mid D) = \frac{P(D \mid G) \times P(G)}{P(D)}$$

Parametric inference:

application: feature learning

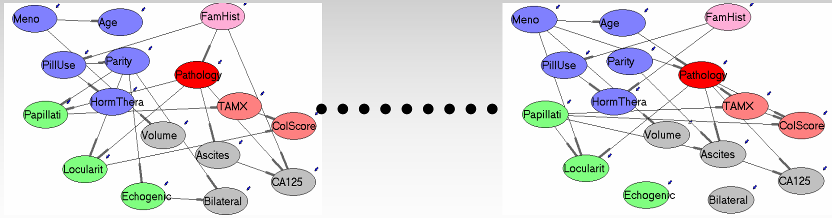
$$P(F = f) = \sum_{G:F^G = f} P(G)$$

- Practical methods: MCMC (Markov Chain Monte Carlo) sampling

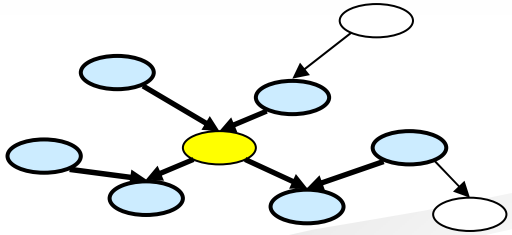
Interesting Bayesian network features

Levels of model features/posteriors:

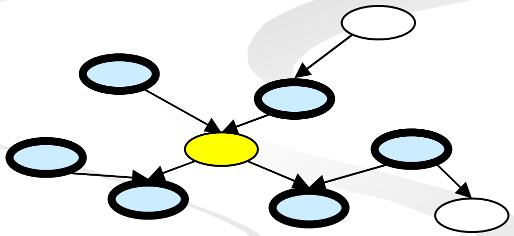
- full structures/DAGs



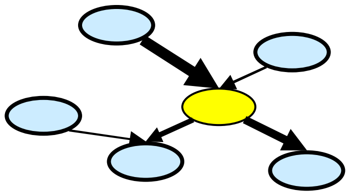
- Markov Blanket Graphs (MBGs)
  - (1) parents of the node,
    (2) its children,
    (3) parents of the children



- Markov Blankets (MBs)
  - the set of nodes which probabilistically isolate the target from the rest of the model



- Markov Blanket Membership (MBM)
- directed edges



Motivations:

- simpler (lower-level) features are easier to learn

Basics of MCMC methods I.

Goal:

- approximating the full-scale summation/integral with an average over DAGs

- DAG-MCMC algorithm:
  - random walk in the space of DAGs
  - evaluating the feature for the visited models
  - approximating the distribution with the sample / calculating average

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

Basics of MCMC methods II.

- Ordering-based MCMC algorithm
  - Random walk in the space of variable orderings

Motivation and background

Fusion of factual knowledge and complex posteriors

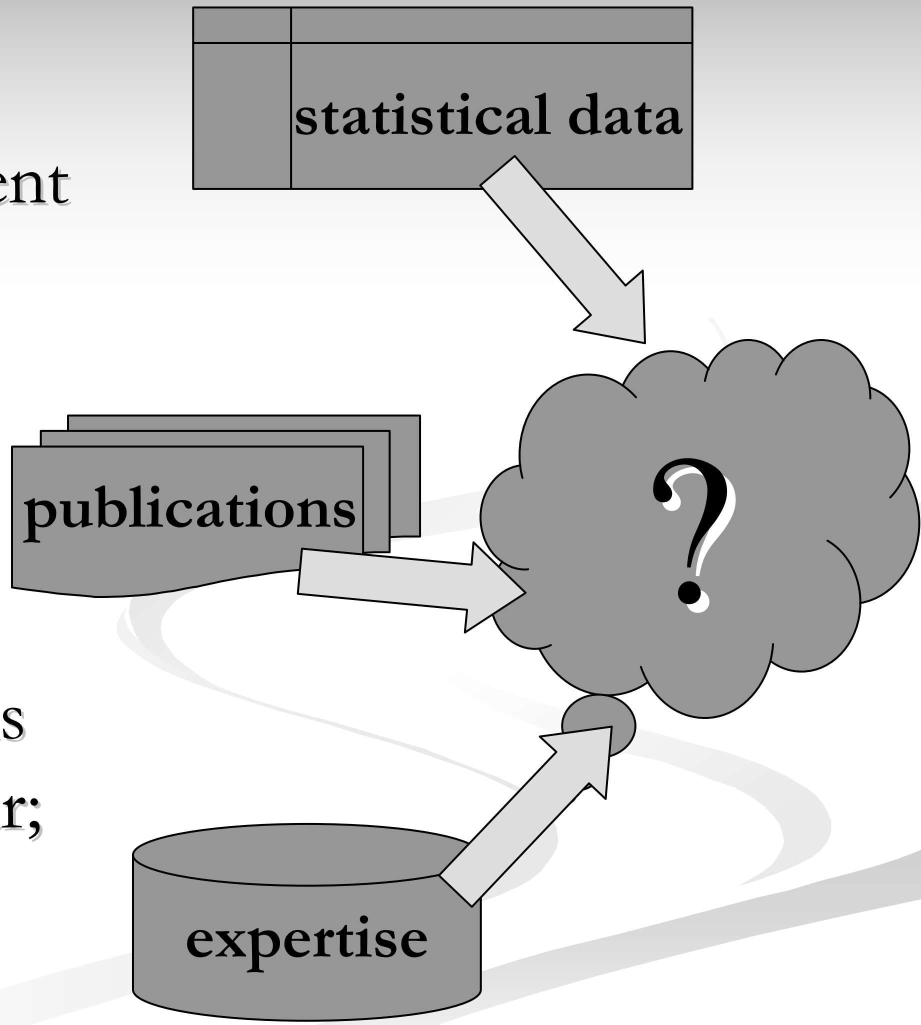
The „Most Probable Sentences" problem

Applications

Summary

Fusion of factual knowledge and complex posteriors

- uncertain (statistical) inference
  - clinical/gene measurement data
  - statistics of publications
- factual knowledge
  - ontologies
  - meta-data of publications (authors, publication year; concept occurrence)



Earlier works – first-order probabilistic logic

J. Y. Halpern. 1990. An analysis of first-order logics of probability.

- distribution over possible worlds
- distribution over possible objects

Practical approaches:

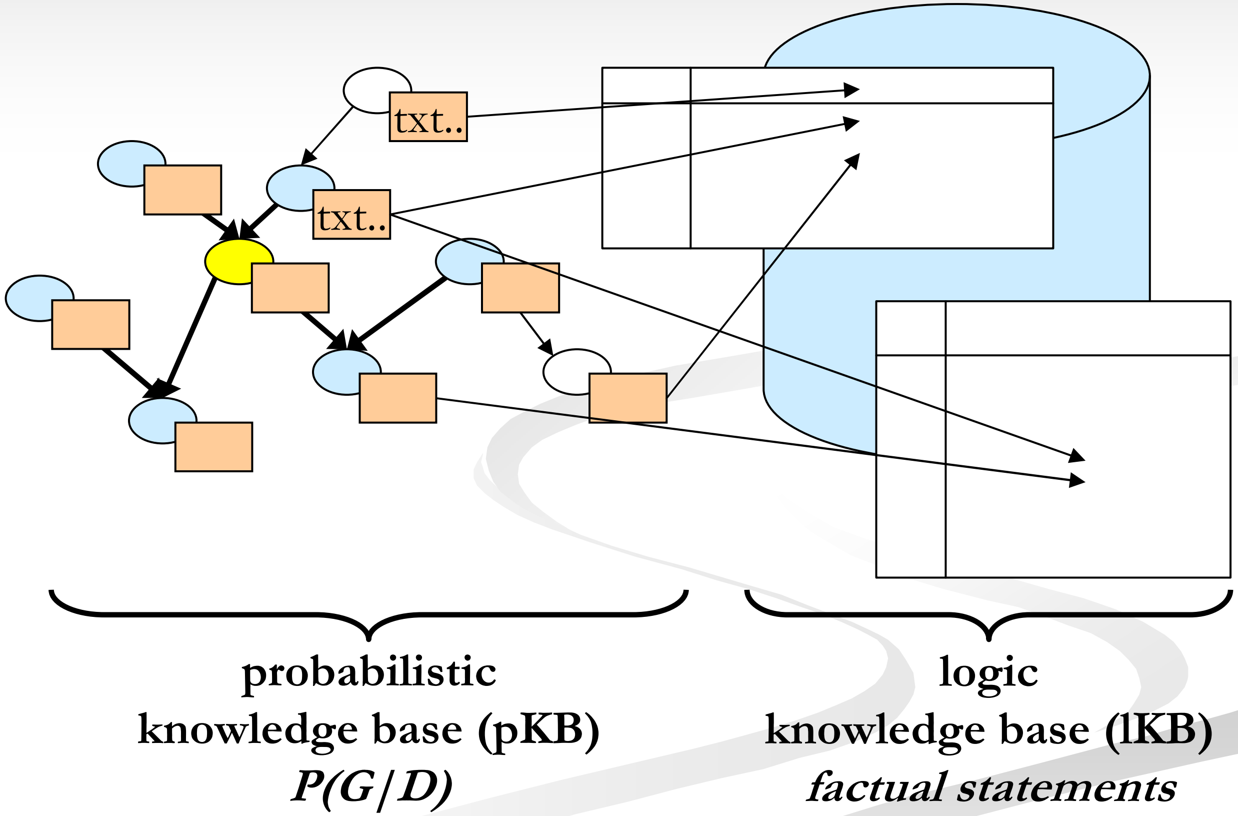
- Knowledge-based model construction
  M. P. Wellman, J. S. Breese and R. P. Goldman. 1992. From knowledge bases to decision models.
- Relational Bayesian networks
  M. Jaeger. 1997. Relational Bayesian networks.
- Bayesian logic programs
  K. Kersting, and L. de Raedt. 2000. Bayesian logic programs.
- Stochastic logic programs
  S. H. Muggleton. 2001. Stochastic logic programs.
- Bayesian logic
  B. Milch, B. Marthi, and S. Russell. 2004. Blog: Relational modeling with unknown objects.
- Markov logic
  P. Domingos and M. Richardson. 2006. Markov logic networks.

- Overview
  Nicos Angelopoulos and James Cussens. 2006. Bayesian learning of Bayesian networks with informative priors.

Probabilistic Annotated Bayesian Network Knowledge Bases (PABN-KBs)

Model elements:

- Bayesian networks
  - uncertain part
  - probabilistic relations
- Textual/xml annotations
  - basic description of entities
  - mapping model elements to „outer" knowledge sources
  - Factual knowledge bases logical relations among objects



The FOPL language

Language elements:

- Bayesian networks – possible worlds
- distribution over models
- factual knowledge sources

- predicates:
  - inherited from the factual part
  - dependence (structural) relations of model elements (nodes)
- logic operators
  - $\wedge, \vee, \neg, \exists, \forall$
- semantics (probability of a sentence):

$$p(\alpha \mid \mathcal{K}) = E_{p(M \mid \mathcal{K})}[\alpha^M] = \sum_{G:M(G) \in \mathcal{M}(K^l)} \alpha^{M(G)} p(G \mid \mathcal{K})$$

A BN oriented FOPL language

Predicates about structural relations of nodes

- directed edge
- directed path
- MBM – Markov blanket membership

- parental set
- MB – Markov blanket set
- MBG – Markov blanket graph

A BN oriented FOPL KB

KB elements:

- set of Bayesian networks
- prior distribution over them

- annotations: nodes → ontology entities

  concepts in articles

- publication repository
- ontologies like GO

FOPL query examples

Basics:

„What is the Markov blanket of variable 'X'?"

Involving annotations:

„What is the probability that the Markov blanket of 'X' will contain variables from a certain class?"

Involving the logic knowledge base:

„What is the probability that every concept in the Markov blanket of 'X' appears in one publication?"

Motivation and background

Fusion of heterogeneous data

The „Most Probable Sentences” problem

Applications

Summary

Feature subset selection and its generalizations I.

Feature subset selection (FSS)

- A relevant subset of features
- Two main approaches:
    1. Wrappers (score function)
    2. Filters (conditional distribution of target variables)

Feature subset selection and its generalizations II.

Feature Subgraph Selection (FSG)

- Identification of the relevant subgraph :
  - Relevant subset of features
  - Dependency between them

Most Probable Sentence (MPS)

- Not enough data to select a feature with a dominant posterior
- Multiple selection : K best features

The „Most Probable Sentences" problem

Given: a set of sentences of interest – target set

Task: find the N highest-scoring (those of highest probability) ones

E.g.: *„find the N most probable MBM sets of variable X"*

Search-and-estimate schemes

Exhaustive enumeration of DAGs:

- for each possible DAG: evaluate target sentences
- calculate the probability of each sentence on-the-fly (sum the probability of the models in which the sentence is true)

Theoretical solution used for testing

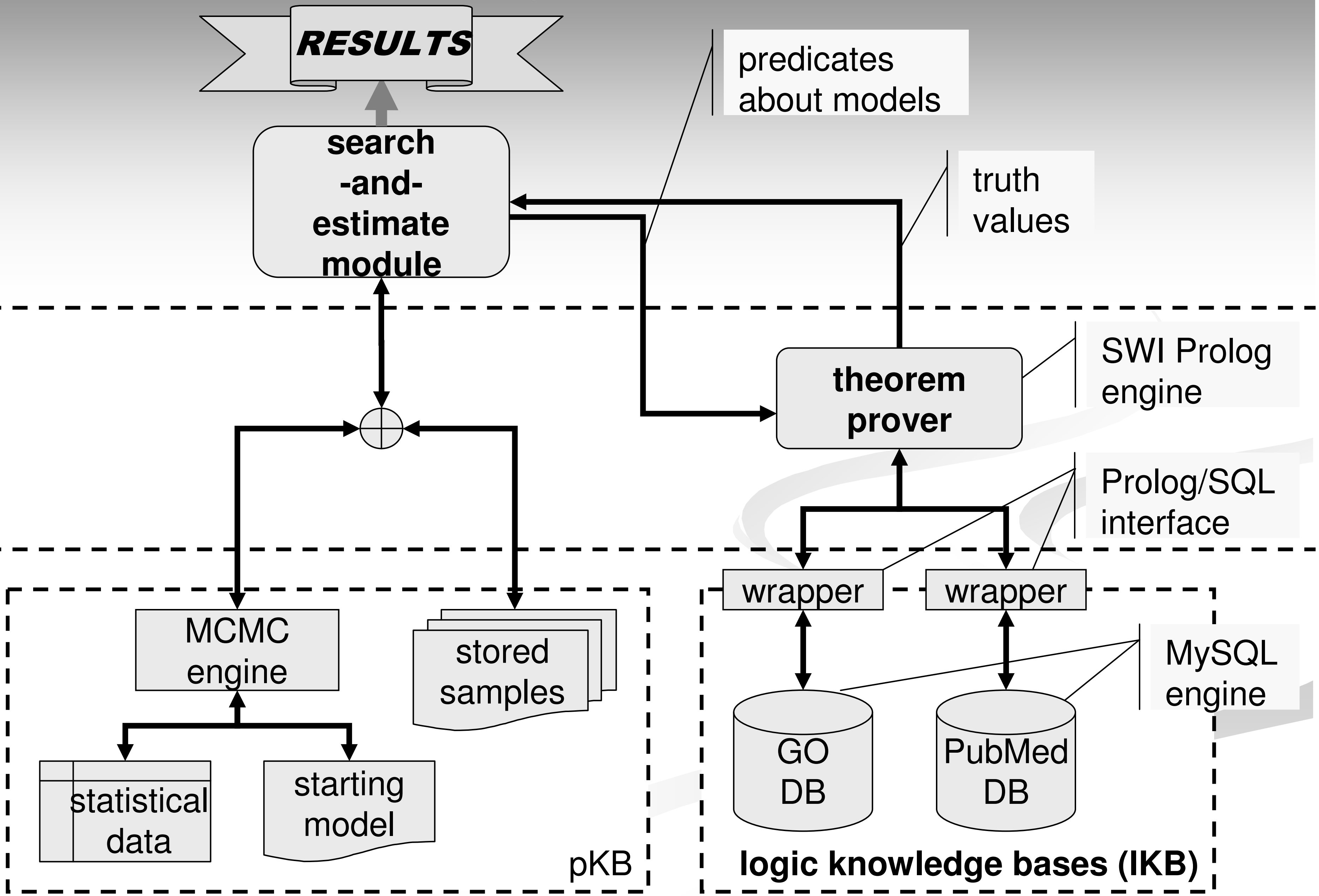
MCMC sampling

- MCMC random walk over DAGs
- for each visited DAG:
  find true sentences / groundings
- update their probabilities:

$$P(S) = \frac{\#(G : KB^G \mapsto S)}{\#(G)}$$

```
listMPS = [];
while( ! MCMC.hasConverged() ){
      model = MCMC.nextModel();
      listNewS = PLEngine.evaluatePredicate(query, model);
      listMPS.insert(listNewS);
}
listMPS.orderBy(prob);
listMPS.truncate(N);
return listMPS;
```

The implemented framework



Motivation and background

Fusion of heterogeneous data

The „Most Probable Sentences” problem

Applications

Summary

Application domain – I. Rheumatoid arthritis

Statistical data:

- clinical observations
  - age, …
  - gender
  - received cures
- gene measurements: single nucleotide polymorphisms (SNP)

Logic knowledge:

- SNP database

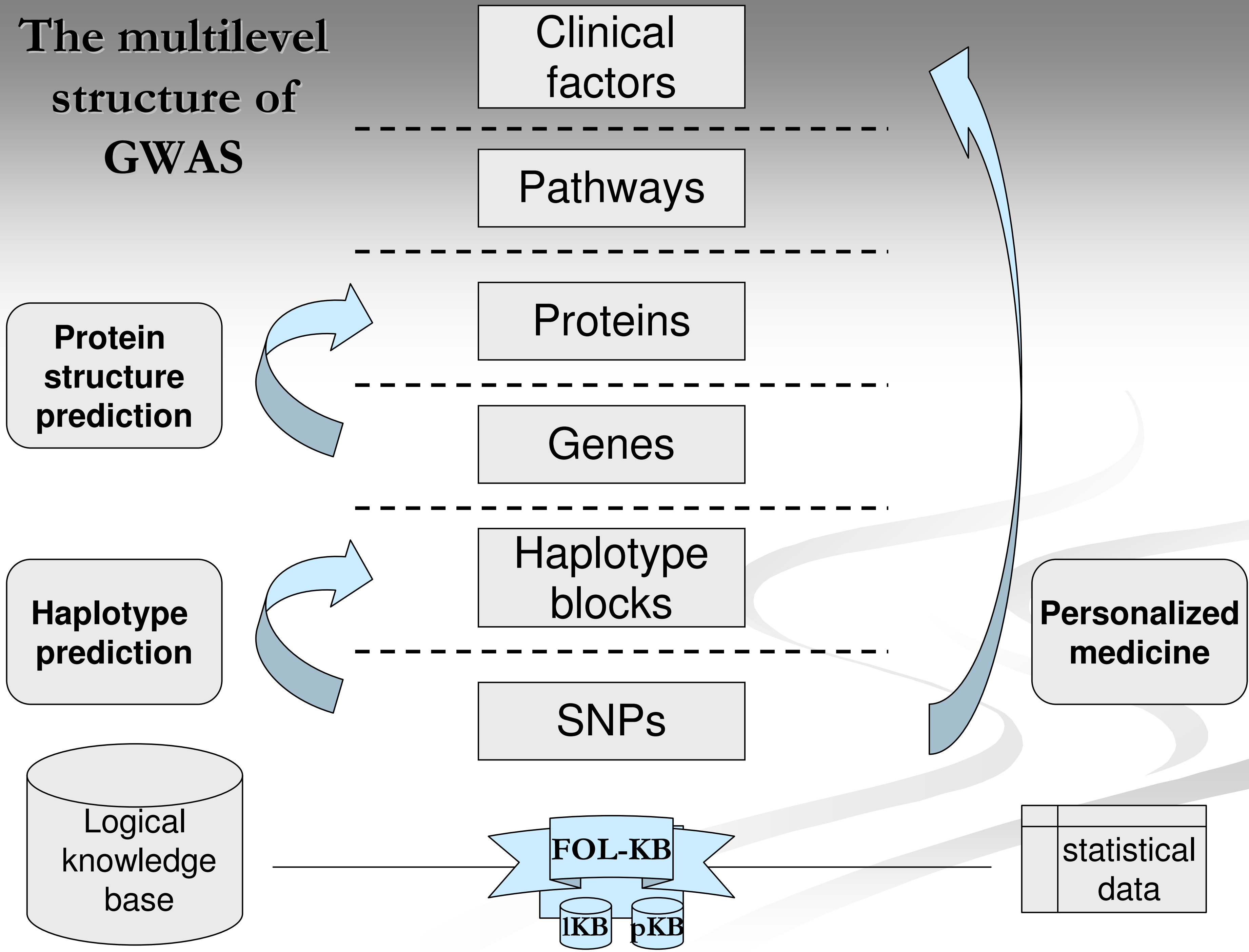
Application domain – II. Genome Wide Association Studies

- moderate number of clinical variables (in the range of 50)
- hundreds of genotypic SNP variables for each patient
- thousands of gene expression measurements

E.g.: Asthma

  - Complex disease mechanism
  - Half of the patients do not respond well to current treatments
  - Unknown pathways in the asthmatic process

The multilevel structure of GWAS



FOPL query examples

„What is the probability that a given SNP 'X' influences certain encoding genes Y_1,Y_2 *that have an effect on certain symptoms of asthma* S_1,S_2,S_3*?"*

„What is the probability that a SNP directly influences the structure of a certain protein, which modifies the „pathway" (the process of the disease), which in turn results in a change of the phenotype (some clinical variable)?"

Motivation and background

Fusion of heterogeneous data

The „Most Probable Sentences" problem

Applications

Summary

Summary

Goal:

Fusion of expertise, data and factual/textual domain knowledge within a first-order logic

Implemented:

Bayesian fusion of a complex posterior over BNs (causal models) and domain literature

Future work

- Extending the model representation
- Hierarchic Bayesian networks
- Describing priors by graph-grammars