Probabilistic Models for Understanding Health Care Data

26-05-2015

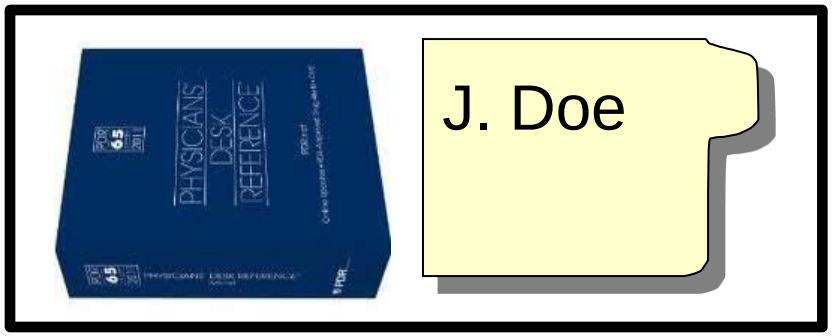
Arjen Hommersom

Open Universiteit
www.ou.nl

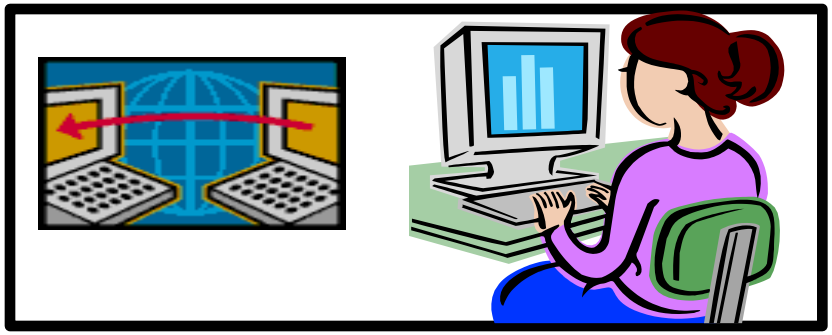
Overview

- Motivation: the health-care domain
- Probabilistic graphical models
- Recent research projects
- Identification of states in probabilistic automata
  - state-based representation of Bayesian networks
  - score-based structure learning
  - treatment of patients with psychotic depression
- Conclusions and plans

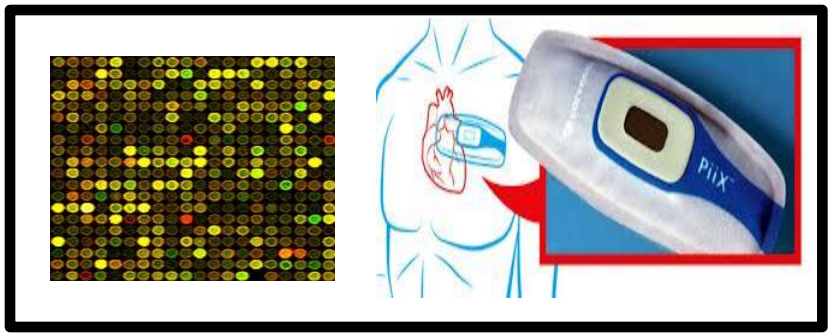
Evolution of health-care



Past



Present



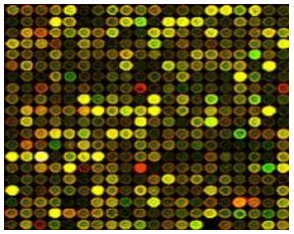
Soon

Open Universiteit
www.ou.nl

Challenge

Complex Data

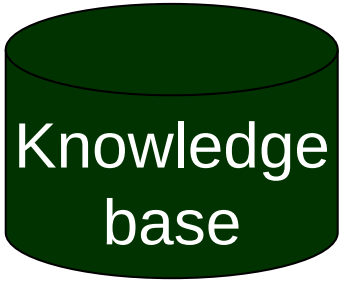
Clinical



Genetics

Lots of knowledge

Papers



Artificial Intelligence

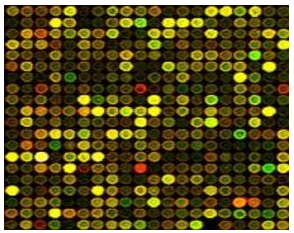
Diagnosis
Treatment
Etc.

How can we deal with all this knowledge and data?

How does AI help?

Complex Data

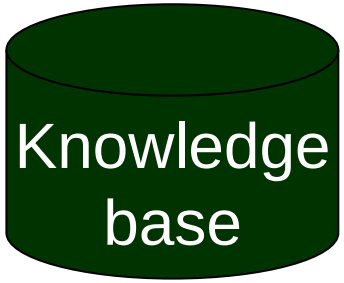
Clinical



Genetics

Lots of knowledge

Papers



Predictive modelling

$\text{MassSize} > 10 \Rightarrow \text{Cancer}$
Etc.

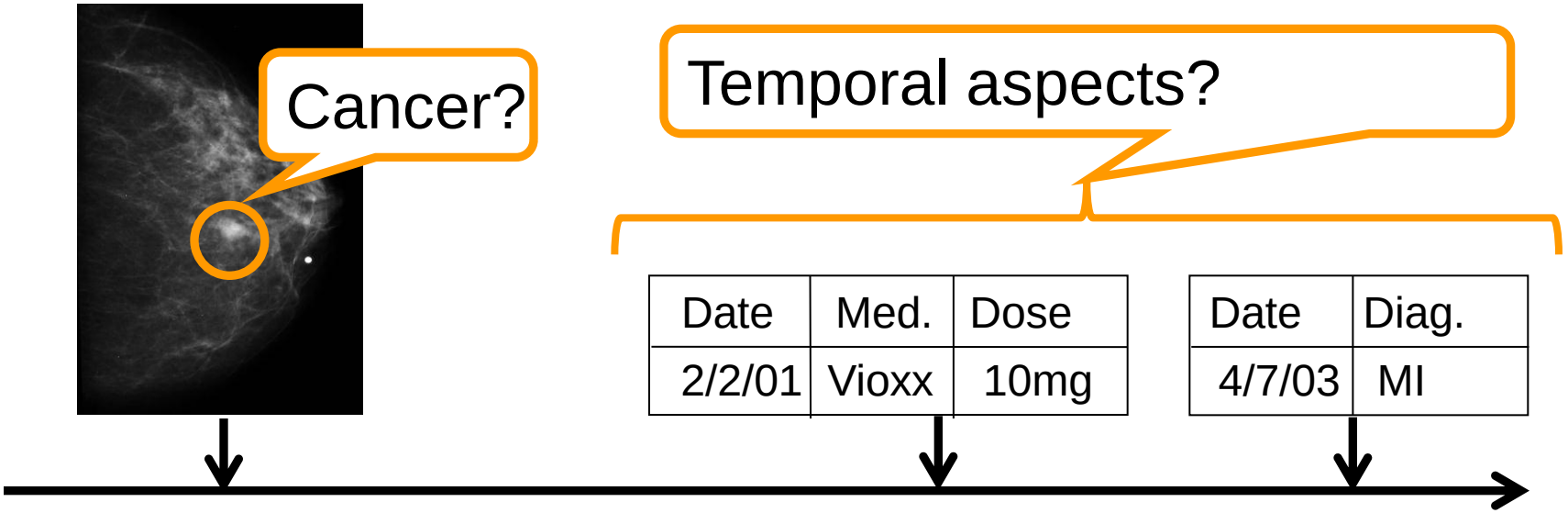
Reasoning about data

Prob(Flu | Fever) = ?

Pattern recognition

$\text{Smoking} \wedge \text{Cancer}$

Solution direction



1. Dealing with uncertainty
2. Grip on the most important relations
3. Understandable models
4. Efficient reasoning

Open Universiteit
www.ou.nl

Uncertainty

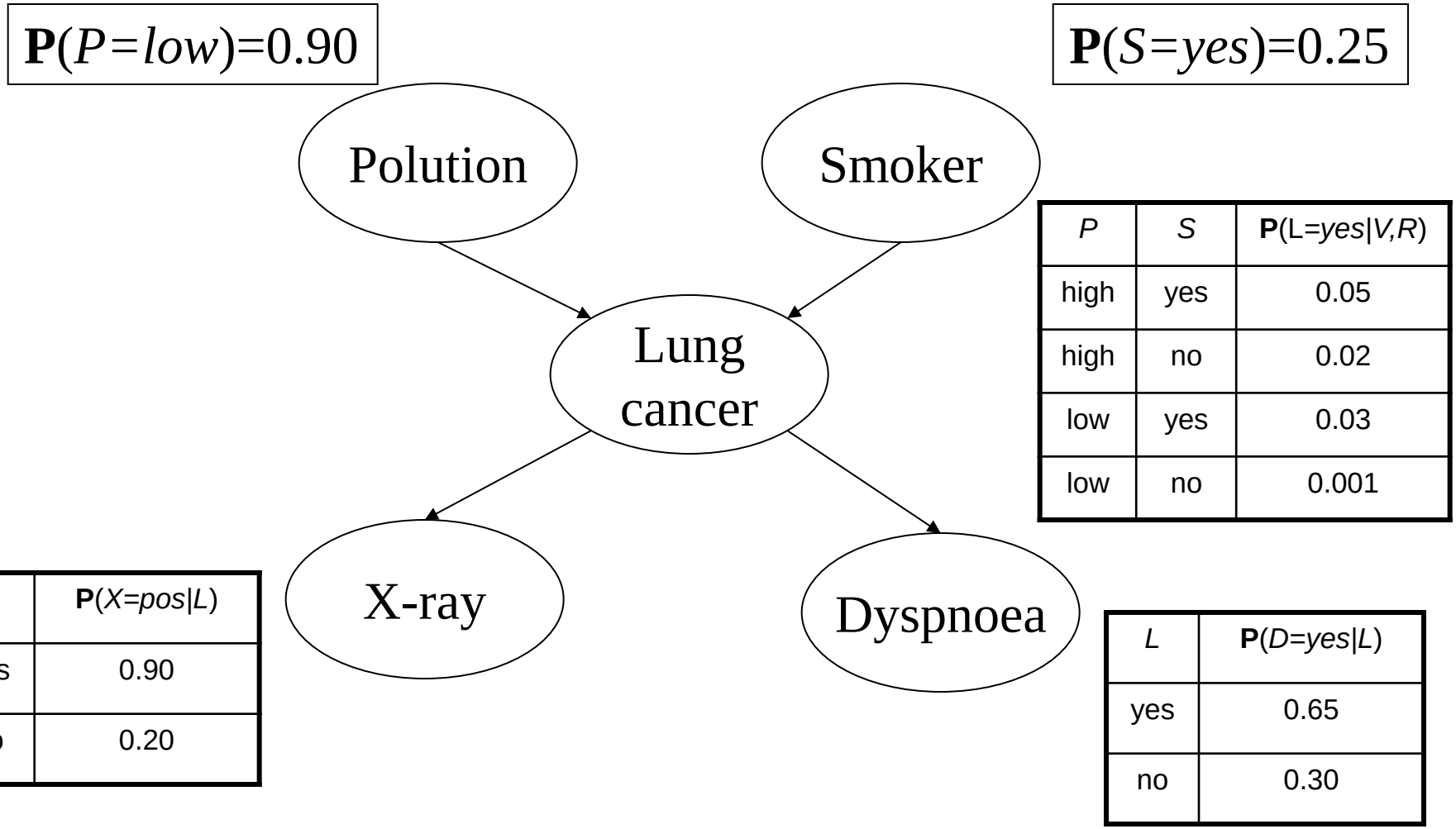
- Let φ, ψ be inconsistent propositional formulas, then:
  1. 0 <= P(φ)
  2. P(true) = 1
  3. P(φ or ψ) = P(φ) + P(ψ) – P(φ and ψ)

- Dutch book argument (agents whose degrees of belief don't satisfy these axioms will be subject to Dutch Book bets where the agent will inevitably lose money)

- Joint distributions over a set of n variables have 2^n parameters

- Key insight in the 80s: exploit independence assumptions (*probabilistic graphical models*)

Introduction Bayesian networks



P	*S*	**P**(L=*yes*|*V*,*R*)
high	yes	0.05
high	no	0.02
low	yes	0.03
low	no	0.001

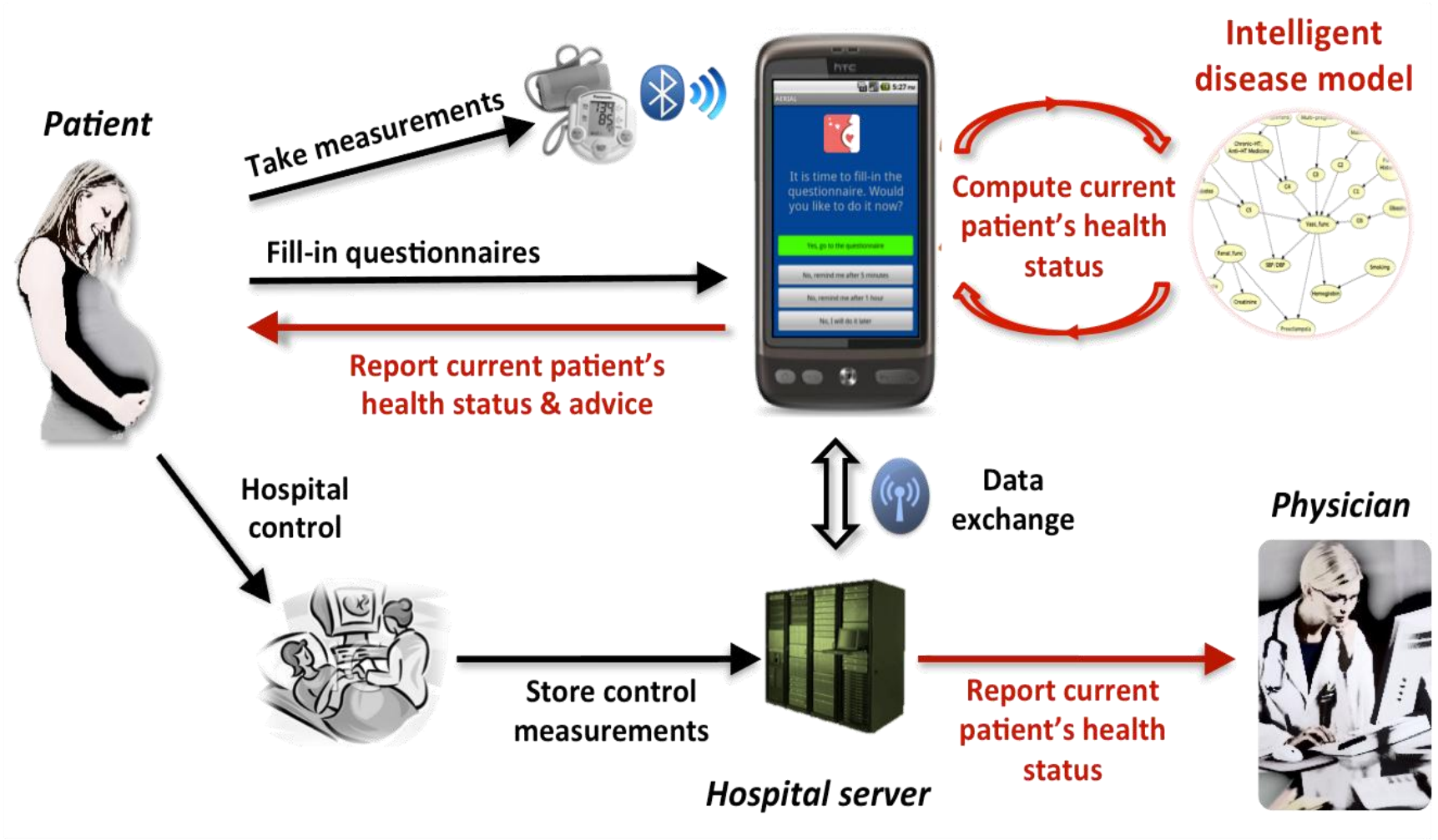
L	**P**(*X*=*pos*|*L*)
yes	0.90
no	0.20

L	**P**(*D*=*yes*|*L*)
yes	0.65
no	0.30

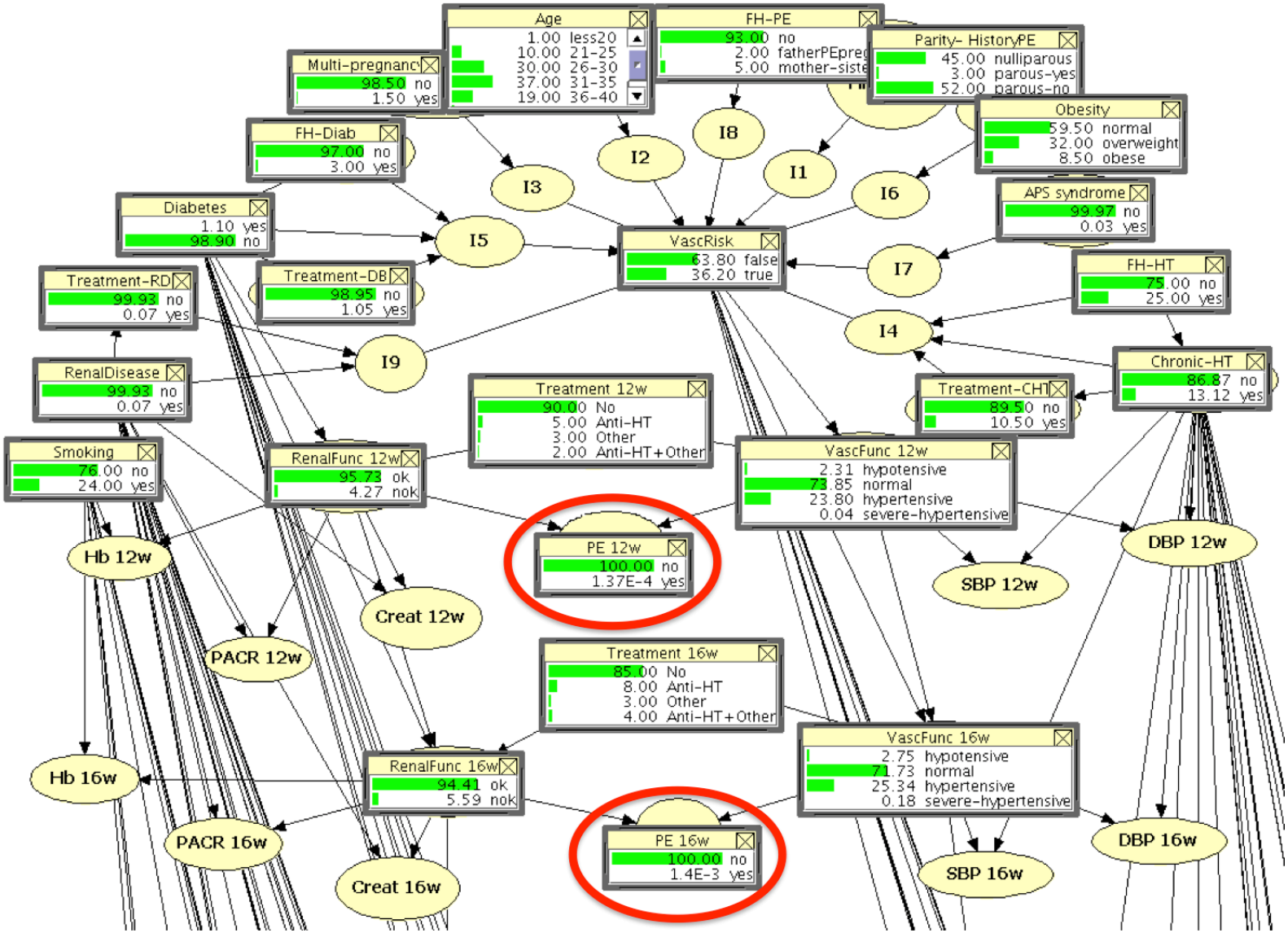
Factorisation:

P(P,S,L,X,D) = **P**(X|L) **P**(D|L) **P**(L|P,S) **P**(P) **P**(S)

e-Health: supporting self-management



Pre-eclampsia network



Continuous-time Models

Move from discrete-time to continuous-time

Models a distribution $P(X_i, X_j, \ldots, X_k)$ for any set of time points {i,j,…,k}

Some interests:

- Building continuous-time models

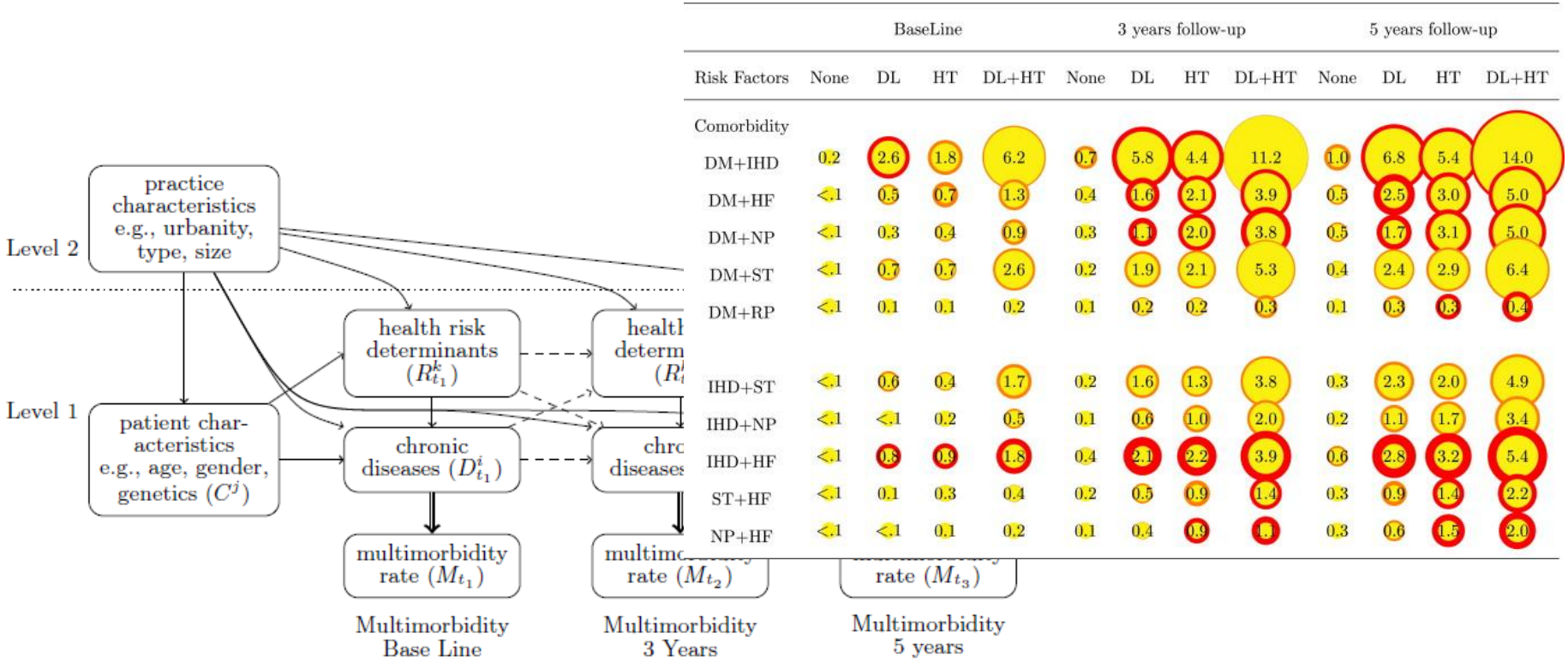
  Maarten van der Heijden, Arjen Hommersom. *Causal Independence Models for Continuous Time Bayesian Networks*. The Seventh European Workshop on Probabilistic Graphical Models, 2014

- Combining different time granularities

  Manxia Liu, Arjen Hommersom, Maarten van der Heijden, Peter Lucas *Hybrid-Time Bayesian Networks*. ECSQARU, 2015.

Epidemiology of multimorbidity

- 2/3rd of patients older than 65 years have at least two chronic conditions
  - problem of multimorbidity
- Complexity increases exponentially with # of diseases
  - Traditional statistical tools cannot deal with this problem!



	BaseLine				3 years follow-up				5 years follow-up			
Risk Factors	None	DL	HT	DL+HT	None	DL	HT	DL+HT	None	DL	HT	DL+HT
Comorbidity												
DM+IHD	0.2	2.6	1.8	6.2	0.7	5.8	4.4	11.2	1.0	6.8	5.4	14.0
DM+HF	<.1	0.5	0.7	1.3	0.4	1.6	2.1	3.9	0.5	2.5	3.0	5.0
DM+NP	<.1	0.3	0.4	0.9	0.3	1.1	2.0	3.8	0.5	1.7	3.1	5.0
DM+ST	<.1	0.7	0.7	2.6	0.2	1.9	2.1	5.3	0.4	2.4	2.9	6.4
DM+RP	<.1	0.1	0.1	0.2	0.1	0.2	0.2	0.3	0.1	0.3	0.3	0.4
IHD+ST	<.1	0.6	0.4	1.7	0.2	1.6	1.3	3.8	0.3	2.3	2.0	4.9
IHD+NP	<.1	<.1	0.2	0.5	0.1	0.6	1.0	2.0	0.2	1.1	1.7	3.4
IHD+HF	<.1	0.8	0.9	1.8	0.4	2.1	2.2	3.9	0.6	2.8	3.2	5.4
ST+HF	<.1	0.1	0.3	0.4	0.2	0.5	0.9	1.4	0.3	0.9	1.4	2.2
NP+HF	<.1	<.1	0.1	0.2	0.1	0.4	0.9	1.1	0.3	0.6	1.5	2.0

Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity
M Lappenschaar, A Hommersom, PJF Lucas, J Lagro, S Visscher.
Journal of clinical epidemiology (2014).

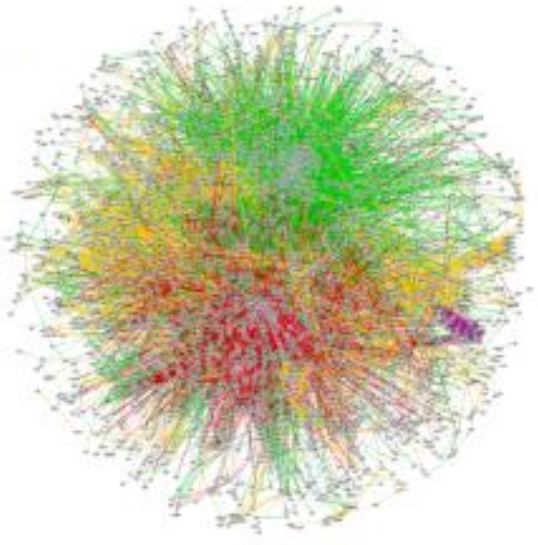
Probabilistic Logic Programming

- Programming language + random variables
- Reason about distribution over executions (*As going from hardware circuits to programming languages)*
- ProbLog: Probabilistic logic programming/datalog
- Example: Gene/protein interaction networks Edges (interactions) have probability “Does there exist a path connecting two proteins?”

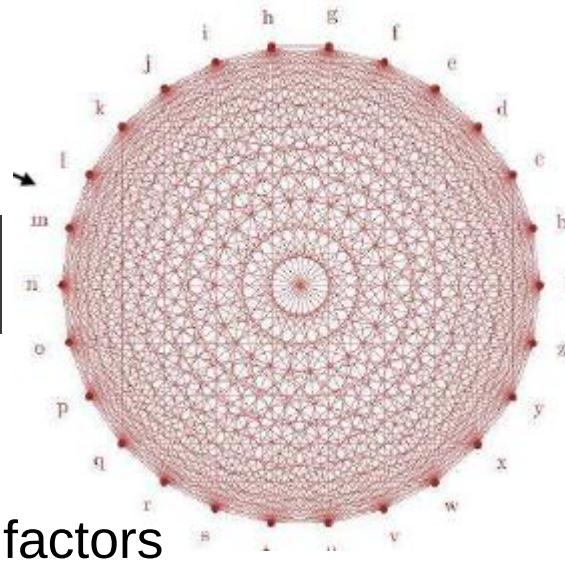
```
path(X,Y) :- edge(X,Y).
path(X,Y) :- edge(X,Z), path(Z,Y).
```

- Cannot be expressed in first-order logic

- **Need a full-fledged programming language!**



Why logic?



- Probabilistic model

$$\text{FacultyPage}(x) \wedge \text{Linked}(x,y) \Rightarrow \text{CoursePage}(y)$$

- As a probabilistic graphical model:
  - 26 pages; 728 variables; 676 factors
  - 1000 pages; 1,002,000 variables; 1,000,000 factors

- Highly intractable?
  - Using probabilstic syllogisms and first-order resolution
  - **Lifted inference in milliseconds!**

- Medical Bayesian networks exhibit large amounts of symmetries that can be exploited
  - Large diagnostic networks ranging between 135 and 1041 variables) may be reduced between **75-85%** (*Is Medical Reasoning Relational? ILP* Conference, Nancy, 2014)

Continuous values in probabilistic logic

In many practical medical application, we also have continuous variables

```
Gluc_if_DM ~ N(7.5,3.8)
Gluc_if_notDM ~ N(5.79, 0.98)

hba1c(1.4 + 0.92 * Gluc_if_DM + N(0, 3.3)) <- dm
hba1c(0.6 + 0.9 * Gluc_if_notDM + N(0, 0.3)) <- not(dm)

e <- hba1c(H), H > 7.2
```

Compute hard bounds on probabilities in this general context:

$$0.416 < P(dm \mid e) < 0.554$$

Constraints can be made arbitrarily small

S. Michels, A.J. Hommersom, P.J.F. Lucas, M. Velikova. *A New Probabilistic Constraint Logic Programming Language Based on a Generalised Distribution Semantics*. Accepted for AI Journal, 2015.

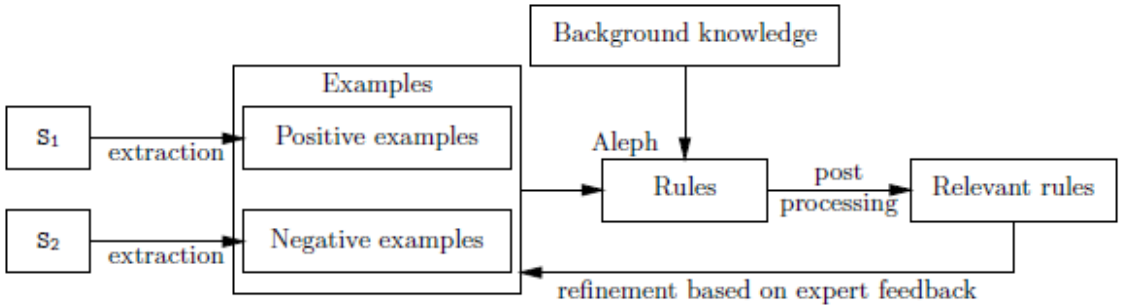
Learning logical rules from data

- PALGA: **63M** pathology excerpts from the Netherlands
- Goal: discovering novel disease associations

Example:

$\texttt{diagnosis}(P, \texttt{auto-immune disease}, T_1)$
$\wedge\ \texttt{topography}(P, \text{liver}, T_2)$
$\wedge\ \texttt{morphology}(P, \texttt{fibrosis}, T_3) \Rightarrow \texttt{cholangitis}(P, T)$

where $T_1, T_2, T_3 < T$



Tim Op De Beeck, Arjen Hommersom, Jan Van Maarten van der Heijden, Jesse Davis, Peter Lucas, Lucy Overbeek, and Iris Nagtegaal. *Mining Hierarchical Pathology Data Using Inductive Logic Programming.* Artificial Intelligence in Medicine (AIME) Conference, 2015.

Structure-learning HBNMMs

or: Identifying States in Probabilistic Automata

Arjen Hommersom - joint work with Marcos Bueno, Peter Lucas, Sicco Verwer, Martijn Lappenschaar, and Joost Janzing

Open Universiteit
www.ou.nl

Motivation

- Probabilistic automata: suitable for identifying probabilistic processes given sequences of events (or sequences of actions/words/etc.)
  - certain probabilistic automata (PDFA) are polynomially trainable
  - PNFA are identifiable in the limit with probability 1
- Key problem: identify number of states and transitions between them
- States itself are black boxes
- CAREFUL project: identify states as well

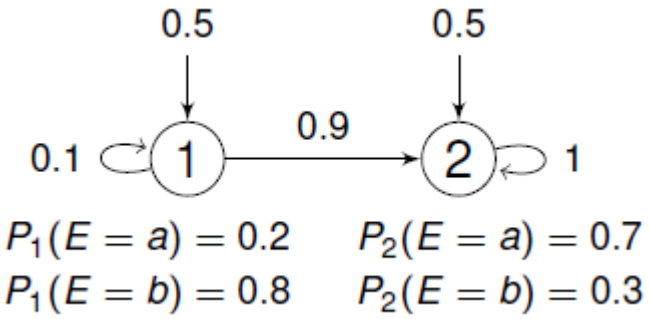
Open Universiteit
www.ou.nl

Outline

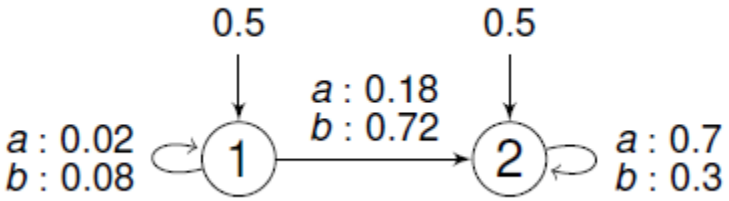
1. State-based representation of Bayesian networks: HBNMM
2. Score-based structure learning
3. Application: treatment of patients with psychotic depression

Probabilistic automata and HMMs

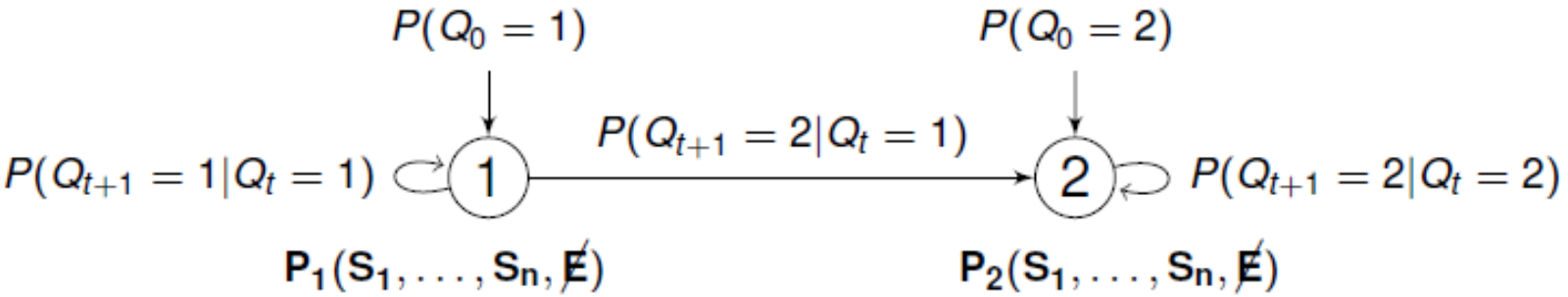
Hidden Markov models = PNFAs without final probabilities
For example, the HMM:



can be translated to the PA (and back):



HBNMM



- Represent P_i (S_1, …, S_n) by a Bayesian network B_i
- Problem: how to learn both transitions and the structure of these B_i?
- Learning structures within HMMs ≈ learning states in PAs

Learning Problem

Given a fixed set of states Q, where $|Q| = n$, let

- T be the the transition probabilities $P(Q_0)$ and $P(Q_{t+1}|Q_t)$
- $B = \{B_i \mid 1 \leq i \leq n\}$ be a set Bayesian networks associated to each state
- $M = (T, B)$ the HMM-BN model with K parameters (details omitted in this talk)
- D a dataset, complete for $S_1, .., S_n$ but varying length of sequences

We aim to find the model with the best score:

$$S(M) = \log P(D \mid M) - Pen(K)$$

where $P(D \mid M) = L(M)$ is called the likelihood and Pen is some penalty function

→ algorithms that learn good Bayesian networks exist

Learning Challenges

- Problem 1 (hidden variables): variables Q_t are unobserved → score will not decompose, which makes exact methods intractable
  - Model selection EM algorithm (Friedman) for learning structure in the presence of missing data

- Problem 2 (dynamics): sequences may be long and data is not available for each time t
  - Learning can be decomposed per state
  - Structure learning *only* involves observed variables

Algorithm

Assuming the penalty can be decomposed (for most scores it can):

$S(M) = \log L(M) - Pen(K)$
$= \log L(T) + \sum_i (\log L(B_i) - Pen(K_i)) - const$
$= \log L(T) + \sum_i S(B_i) - const$

which leads to the following procedure:

Algorithm 2 Expectation-Maximisation for HBNMM

Input: D, a dataset with time series $\mathbf{X}^{(0:T)}$.
Output: An HBNMM that maximises the expected score for D.

1: Choose a model M randomly
2: **while** stopping criterion is not reached **do**
3: **E step for** Compute $P(S^{(t)} \mid D_k)$
4: **for** each state $s \in S$ **do**
5: **M step for** state s. Learn a new model $\phi(s)$ that maximises the expected score
6: set $M_{\downarrow}$ to distributions according to $\phi(s)$
7: **M step** for M_0 and $M_{\rightarrow}$. Estimate new M_0 and $M_{\rightarrow}$ which maximise the expected likelihood
8: $M := (M_0, M_{\rightarrow}, M_{\downarrow})$

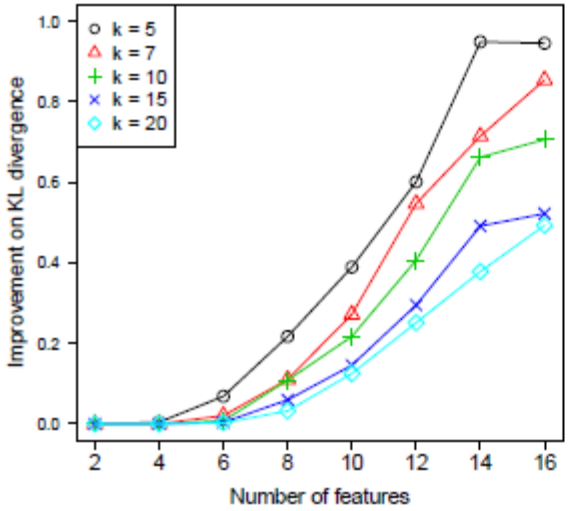
Complexity of learning

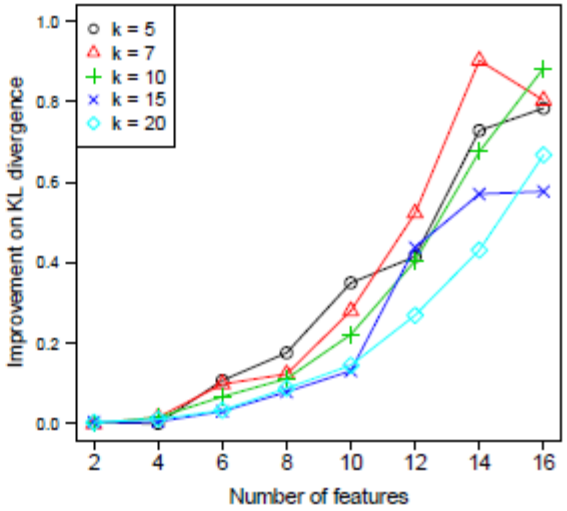
- Mixture of structure learning and the Baum-Welch algorithm for finding unknown parameters of an HMM

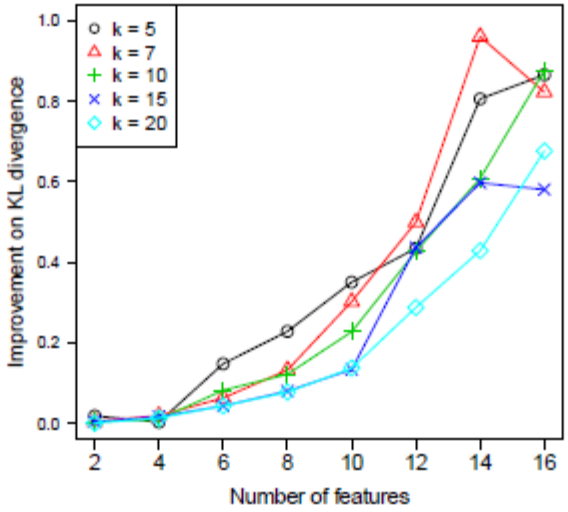
- Computing the E-step relatively easy: quadratic in number of states, linear in data size

- M-step: linear in states, NP-hard learning problem
  - Optimizing expected score not harder than optimizing the score; we just have a weighted likelihood
  - Very feasible for states with limited number of variables

Experiments with artificial data







Comparison with regular HMM and conditional Chow-Liu structures (Kirshner, UAI'2004)

Treatment of psychotic depression

- Data of 122 patients obtained by a randomized controlled trial
- At start of treatment, all patients were diagnosed with DSM-IV-TR psychotic major depression
- Three types of treatments evaluated: venlafaxine, imipramine (antidepressants) or venlafaxine+quetiapine (antidepressant + antipsychotic)
- Previous research focused on Hamilton score
- Primary finding: venlafaxine+quetiapine is more effective than venlafaxine alone

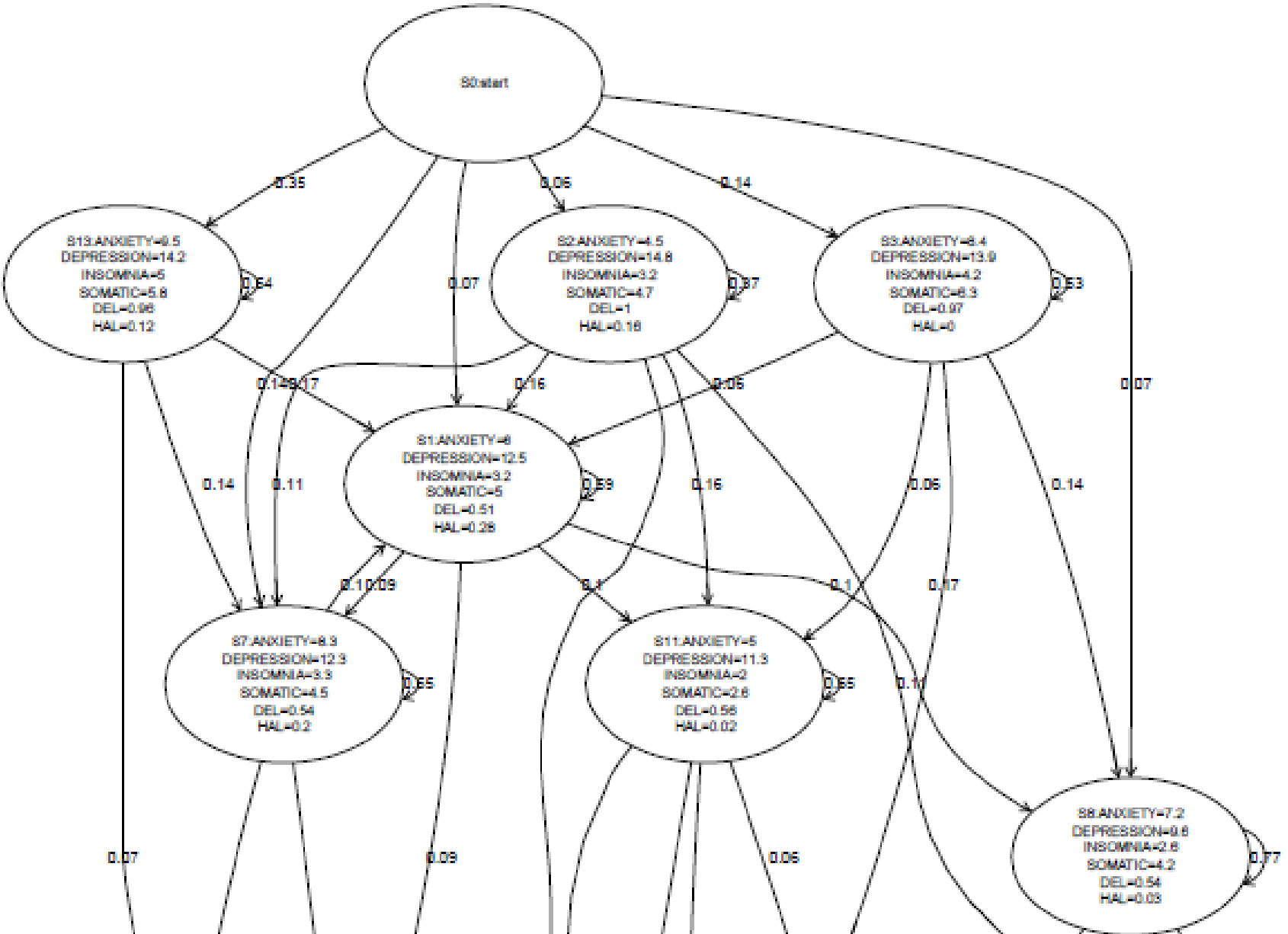
Psychotic depression data

- Collected for 8 weeks (20 patients dropped out earlier)
- Symptoms recorded each week
- 17 items rating the severity of the depression:
  - mood
  - feelings of guilt
  - suicide thoughts
  - insomnia
  - agitation
  - etc.
- Sum of these 17 items is called the Hamilton score (lower = better)
- Two psychotic symptoms (hallucinations, delusions)

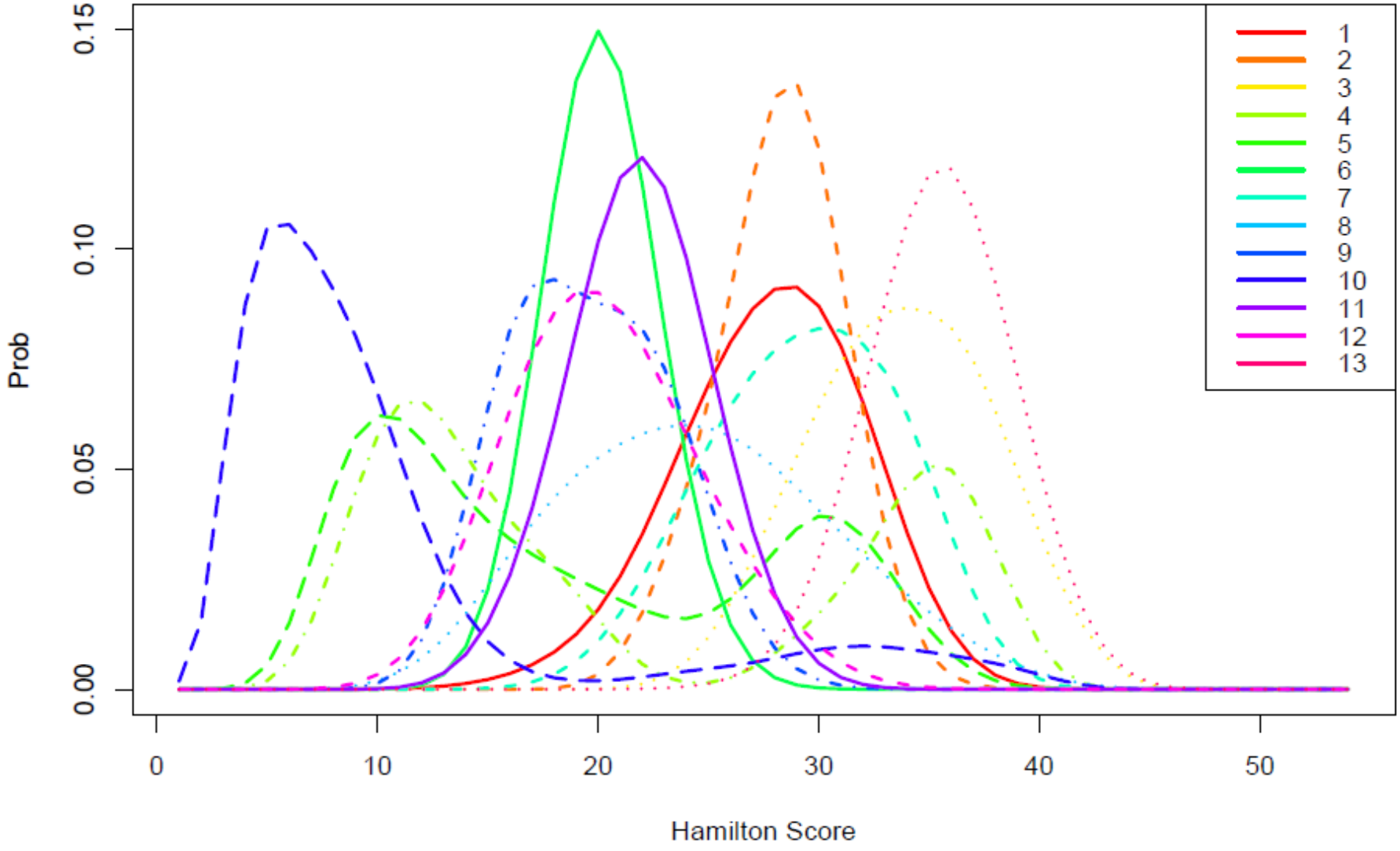
Intended contributions

- In general: obtain more insight compared to regression models
- Identify different patient groups that somehow behave differently (responders – non-responders)
- Identify most important factors that determine recovery
- Explain differences in outcomes between treatments
- Improve fitting of models (linear vs non-linear)

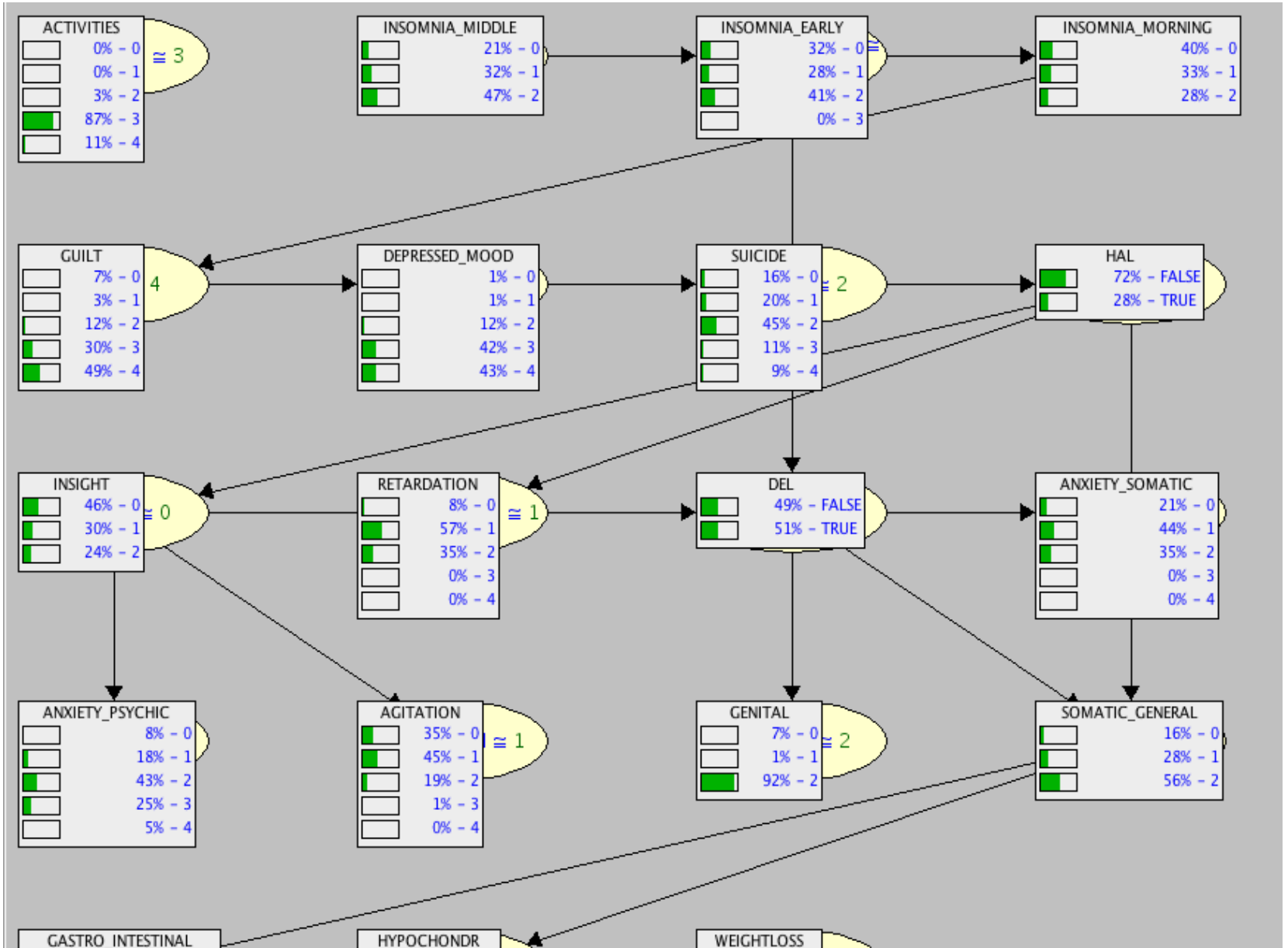
Part of the model (13 states)



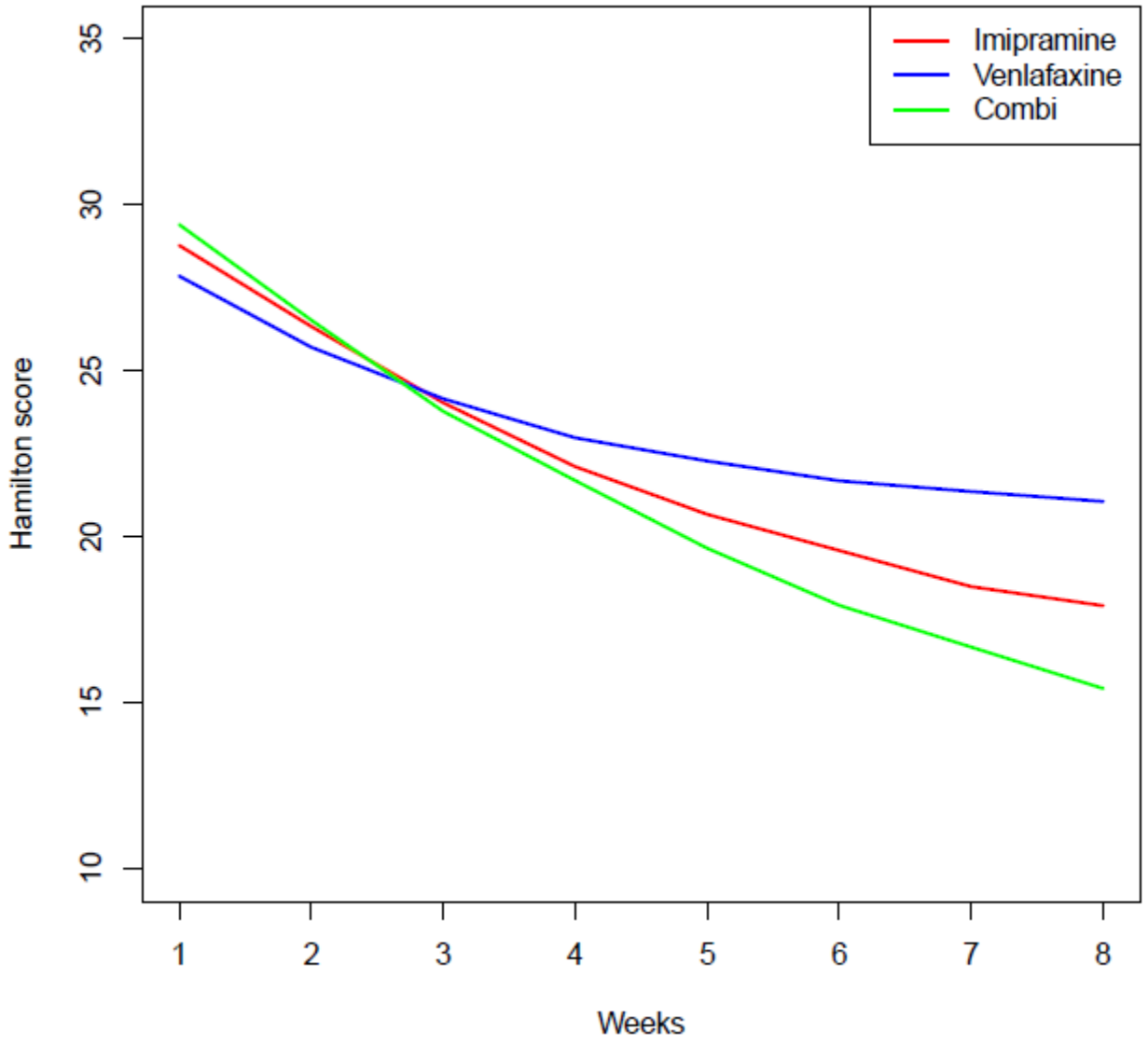
Hamilton score per state



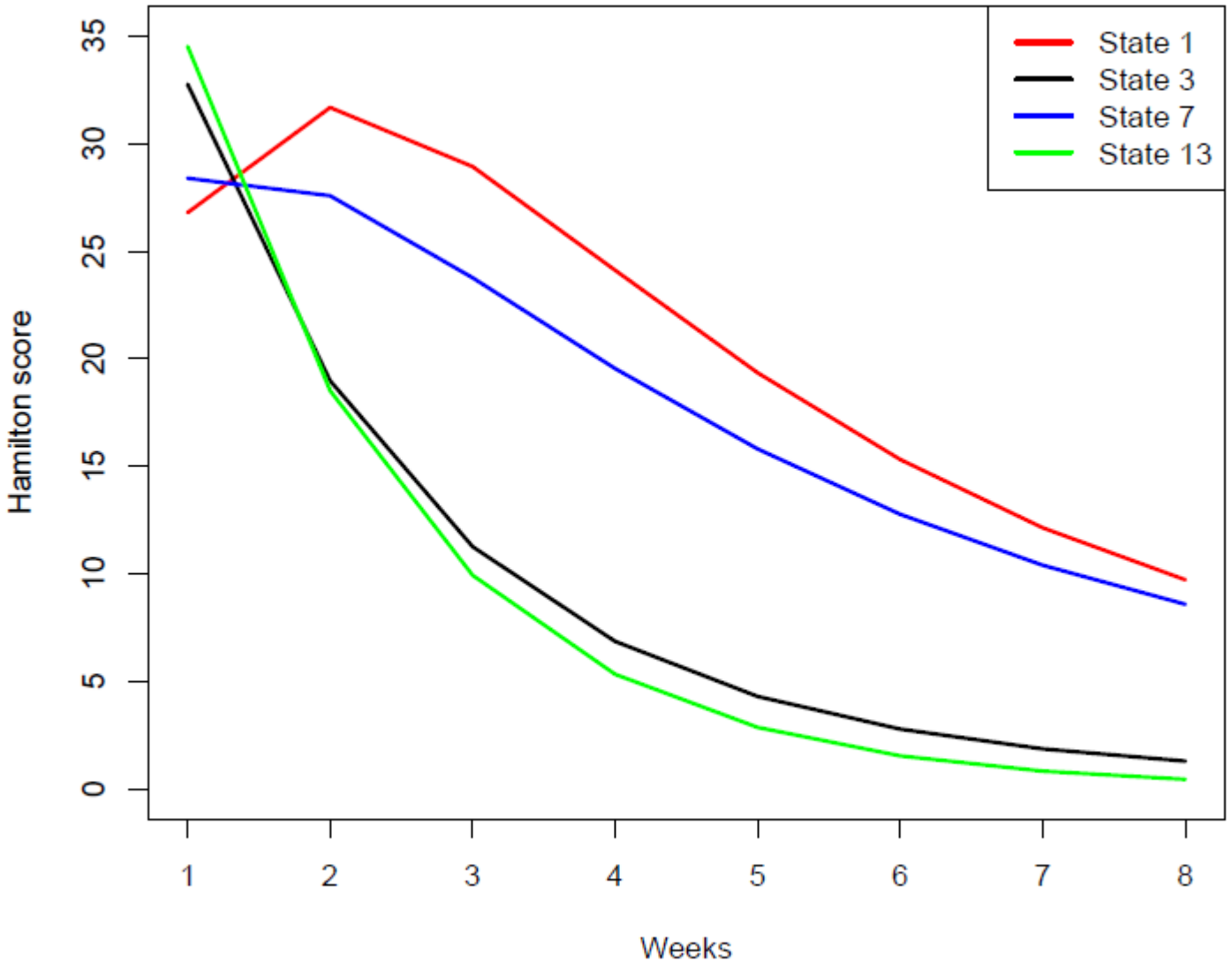
Example state (S1)



Comparison between treatments



Outcomes per state



Conclusions

- Significant challenges in analysing (medical) data
  - complexity, uncertainty
- Introduction of a Bayesian-network based probabilistic automaton
- Application to treatment psychotic depression
- Research directions from OU point of view:
  - Smart technologies in health care services
  - Currently involved in BISS-SIC: first trial project for developing smart interaction centers
    - Improving services with AI techniques
    - Development of intelligent services

Open Universiteit
www.ou.nl