Decision-Driven Models with Probabilistic Soft Logic

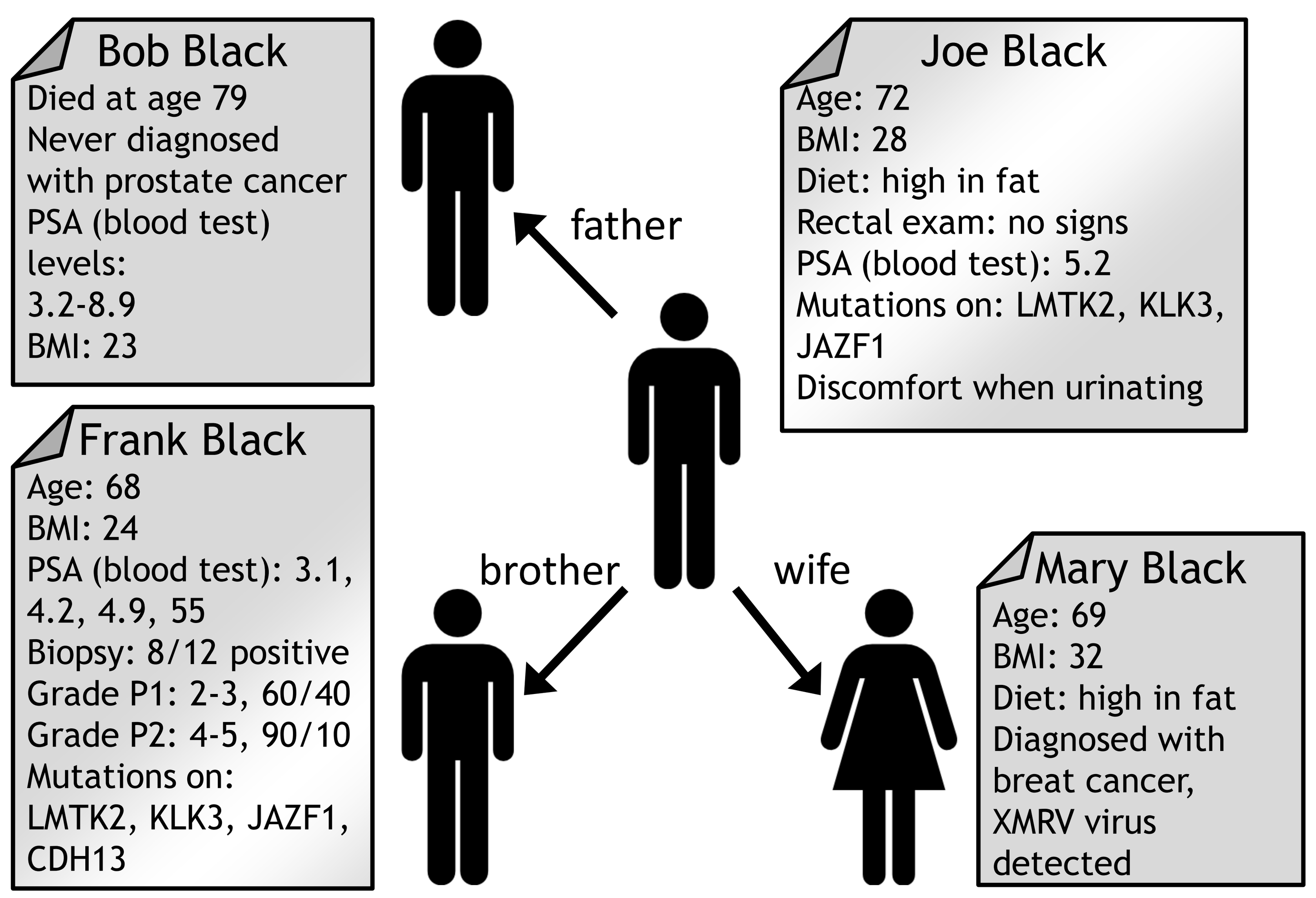
Stephen H. Bach, Matthias Broecheler, Stanley Kok, Lise Getoor

Introduction

- Personalized medical decisions require integrating ever-increasing amounts of uncertain information, making it more difficult to compute the marginal probability distributions of interest to decision makers.
- We propose a new modeling approach: *decision-driven modeling*, which reasons probabilistically about marginals.
- We show how decision-driven models can be constructed easily and represented compactly using *probabilistic soft logic*, a recently introduced framework for statistical relational learning.

Motivating Example

- Joe Black is a patient with test results indicating a chance of prostate cancer. Should his doctor conduct an invasive biopsy?
- We have multiple sources of personalized information with which to assess the probability that Joe has cancer and the probability that it is aggressive.



Decision-Driven Models

- Intuitively, a decision-driven model (DDM) is a probability distribution over a set of random variables, each of which represents a probability distribution.
- To define a propositional DDM, we must first define the propositions for which we wish to infer marginals, such as A = “Joe has prostate cancer,” B = “Joe’s prostate cancer is aggressive,” and C= “Frank has prostate cancer.”
- Then we define a propositional DDM as follows:

Let $A = \{A_1, \ldots, A_n\}$ be a set of propositions. A decision-driven model for A is defined by the sample space $\Omega_{\mathbb{P}} = [0,1]^n$, the set of marginal distribution denotations $\boldsymbol{M} = \{\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_n}\}$, a probability density function $f(\boldsymbol{x} = < x_1, \ldots, x_n >): \Omega_{\mathbb{P}} \to \mathbb{R}_0^+$ such that $\int_{\boldsymbol{x} \in [0,1]^n} f(\boldsymbol{x})\, d\boldsymbol{x} = \mathbf{1}$, and a mapping $g: \Omega_{\mathbb{P}} \to (\boldsymbol{M} \to [0,1])$.

An Example Decision-Driven Model

- A natural choice to represent a density function for a propositional DDM is a constrained continuous Markov random field (CCMRF).
- A CCMRF has the following form:

$$f(\boldsymbol{x}) = \frac{1}{Z(\Lambda)} \exp[-\sum_{i=1}^{m} \lambda_i \phi_i(\boldsymbol{x})] \;; \quad Z(\Lambda) = \int_{\boldsymbol{x} \in [0,1]^n} \exp[-\sum_{i=1}^{m} \lambda_i \phi_i(x)]\, d\boldsymbol{x}$$

along with a set of constraints on $\boldsymbol{x}$.

- Each ϕ_i is a non-negative compatibility function measuring how compatible dimensions of $\boldsymbol{x}$ are. The value 0 is perfect compatibility.
- Example: Statistical data show that 2 out of 5 men whose brothers have prostate cancer develop prostate cancer as well. We can represent this as the compatibility function

$$\phi_1(< \mathbb{P}_A = x_1, \mathbb{P}_B = x_2, \mathbb{P}_C = x_3 >) = \max(0, x_3 * 0.4 - x_1)$$

Probabilistic Soft Logic

- Probabilistic soft logic (PSL) is a language for compactly representing CCMRFs.
- PSL reasons probabilistically about *atoms*, which are first-order predicates grounded with arguments.
- Each atom corresponds to a continuous random variable in a CCMRF.
- Templates for compatibility functions and constraints are written as rules in first-order logic.
- Example: The compatibility function ϕ_1 can be expressed as the ground rule

```
hasCancer(frank,prostate)*0.4 ⇒ hasCancer(joe,prostate)
```

- We can compactly express this knowledge for all brothers with the first-order rule

```
hasCancer(P,prostate)∧ brother(P,Q)*0.4 ⇒hasCancer(Q,prostate)
```

Inference

- The MAP state of a DDM is the set of most likely marginal distributions.
- Inference in many DDMs, such as those constructed with PSL, is efficient, since finding the MAP state of the random variables can be formulated as a numerical optimization problem.
- Interpreting the MAP state is straightforward, since users can see how each inferred marginal affected the others.
- Marginal inference in such DDMs can be viewed as computing a confidence measure in the inferred marginals.

This material is based upon work supported by the National Science Foundation under Grant No. 0937094. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

University of Maryland, LINQS Machine Learning Research Group http://www.cs.umd.edu/linqs