6.036 Introduction to Machine Learning

(meets with 6.862)

Probabilistic modeling and inference

Administrivia

HW4 will be out by the end of the week.

Project 3 Due next Friday 5/5 at 9PM.

Drop Date: **today Thursday 4/27.**

As always:

- Check LMOD/Piazza for announcements.
- To contact staff, use Piazza (6036-staff@lists.csail.mit.edu for exceptions only)

Bayesian networks

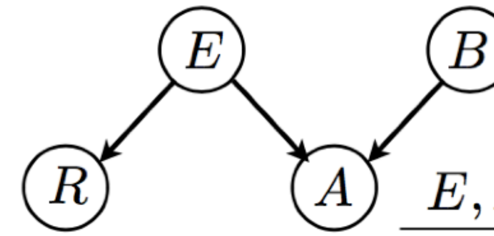
Rich class of generative models, combining **graphs** and **probability**

$E:$

T	F
0.01	0.99

$B:$

T	F
0.01	0.99



$R:$

E	T	F
T	1	0
F	0	1

$A|E,B:$

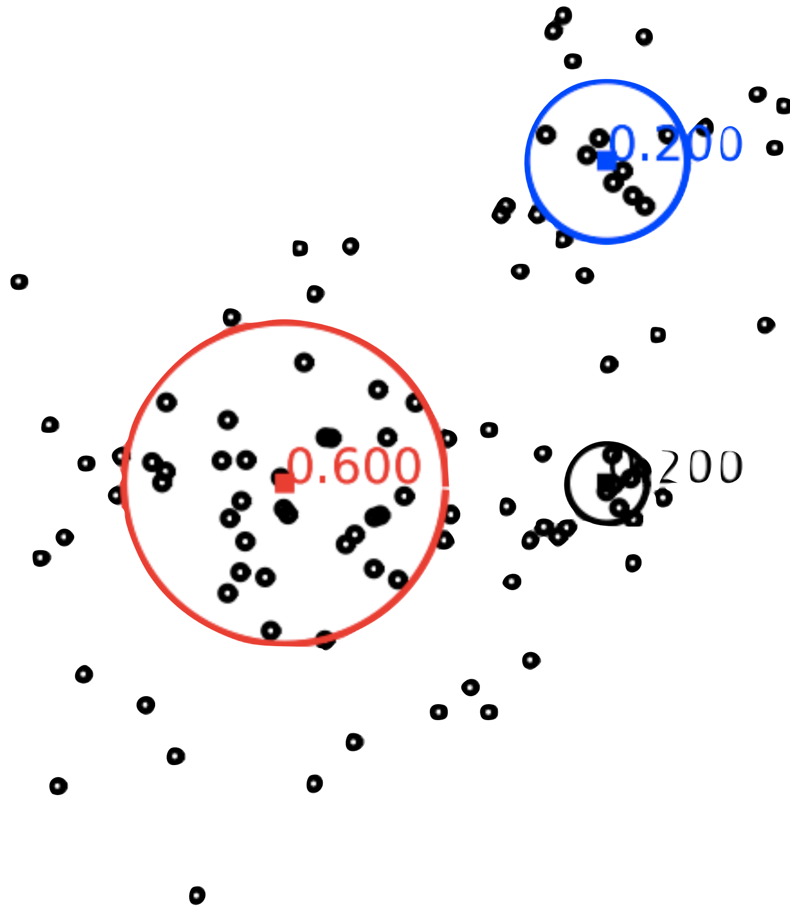
E,B	T	F
T,T	1	0
T,F	1	0
F,T	1	0
F,F	0	1

Key elements:

- A **directed acyclic graph (DAG)** with variables as nodes
- An associated **probability distribution**
- Joint distribution **factors** according to the graph

- Graph makes explicit and summarizes useful properties of the underlying distribution
- Can use the graph structure for efficient inference (compute marginals and conditionals).

Mixtures



- Recall Mixture model
  (e.g., **Gaussian mixture**)

$$y \qquad y \in \{1, \ldots, K\} \qquad P(y) = p_y$$

$$\downarrow$$

$$x \qquad x \in \mathbb{R}^d \qquad P(x|y) = N(x; \mu^{(y)}, \sigma_y^2 I)$$

- Joint density

$$p(x, y) = \sum_{y \in \{1, \ldots, K\}} p_y \, N(x; \mu^{(y)}, \sigma_y^2 I)$$

What is the associated graph?

Markov models and HMM

Markov Models

$$y_1 \longrightarrow y_2 \longrightarrow \cdots \longrightarrow y_{n-1} \longrightarrow y_n$$

$$P(y_1, \ldots, y_n) \stackrel{def}{=} P(y_1)P(y_2|y_1)P(y_3|y_2)\cdots P(y_n|y_{n-1})$$

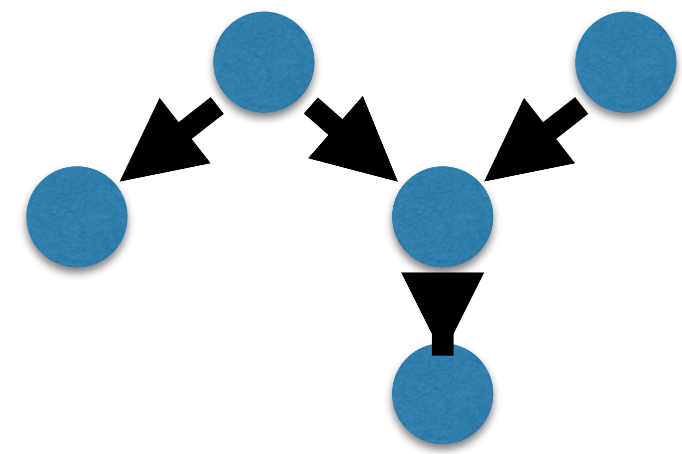
Hidden Markov Models (HMM)

$$\begin{array}{ccccccccc}
y_1 & \longrightarrow & y_2 & \longrightarrow & \cdots & \longrightarrow & y_{n-1} & \longrightarrow & y_n \\
\downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
x_1 & & x_2 & & \cdots & & x_{n-1} & & x_n
\end{array}$$

$$P(y_1, \ldots, y_n, x_1, \ldots, x_n) = P(y_1)P(x_1|y_1)\prod_{i=2}^{n}[P(y_i|y_{i-1})P(x_i|y_i)]$$

Efficient descriptions

BNs can be compactly described.



- If variables are binary, how many parameters do we need for this 5-node network?

- What about for an *n*-step HMM model?

- What if it is time-homogeneous?

$$\begin{array}{ccccccccc}
y_1 & \rightarrow & y_2 & \rightarrow & \cdots & \rightarrow & y_{n-1} & \rightarrow & y_n \\
\downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
x_1 & & x_2 & & \cdots & & x_{n-1} & & x_n
\end{array}$$

(In)dependence properties

- How to understand the dependence/independence properties more formally?

- Recall there's a distinction between
  - **A** and **B** and *independent* (marginal pdf factors)
  - **A** and **B** are *conditionally independent* given **C** (conditional pdf factors)

- Example**: Hair and football**

Hair and football (I)

- **A: Short/Long hair**
- **B: Likes/Dislikes football**

A \ B	Likes football	Dislikes football
Short hair	78	32
Long Hair	32	58

- Not independent! (formally, it does not factor).
- One variable gives *some* information about the other
- Q: Is this true? Is this useful? Does it have explanatory power? It depends for what…

- A better model: add *gender* as a latent variable
  - **C: Male/Female**

Hair and football (II)

- **A** and **B** are *conditionally independent* given **C**
- Full dataset aggregates (mixture of) two subpopulations

Male

A \ B	Likes football	Dislikes football
Short hair	72	18
Long Hair	8	2

+

Female

A \ B	Likes football	Dislikes football
Short hair	6	14
Long Hair	24	56

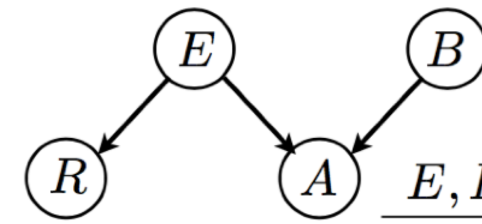
Causality and intervention

E:

	T	F
	0.01	0.99

B:

	T	F
	0.01	0.99



R:

E	T	F
T	1	0
F	0	1

$A|E,B$:

E,B	T	F
T,T	1	0
T,F	1	0
F,T	1	0
F,F	0	1

- What do the arrows really mean?
- Relationships with causality?
- Sometimes (e.g., Burglary/Alarm), is cause/effect
  Sometimes, only correlation, or modeling effect
- Structurally, only a factorization of the joint pdf
  (direction of arrows can be meaningless!)
- How to distinguish between them? *Intervention*

When does independence hold?

- What does the network structure tell us about conditional independence?
- A formal criterion (*d-separation*, “Bayes-Ball” algorithm) to check whether **A** and **B** are conditionally independent given $\mathbf{\{C_1,\ldots,C_k\}}$
- Full details in recitation, let’s work out a few simple (but important) examples
- Why do we care? When variables are not independent, can gain information (infer) about one from the other

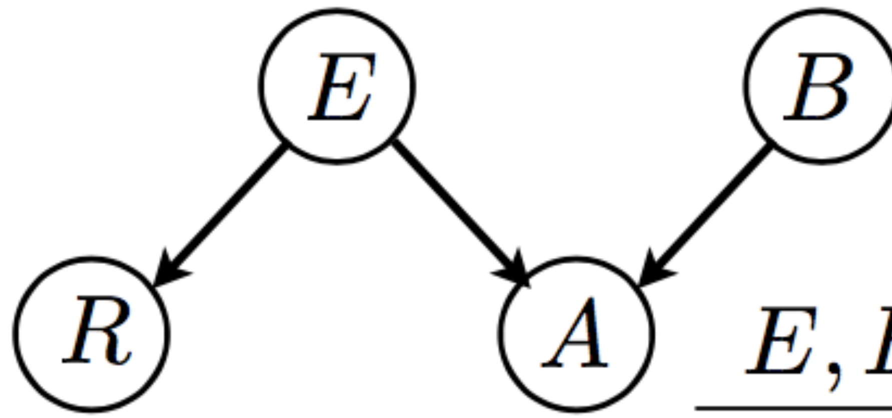
Alarm Example

$E:$

T	F
0.01	0.99

$B:$

T	F
0.01	0.99



$R:$

E	T	F
T	1	0
F	0	1

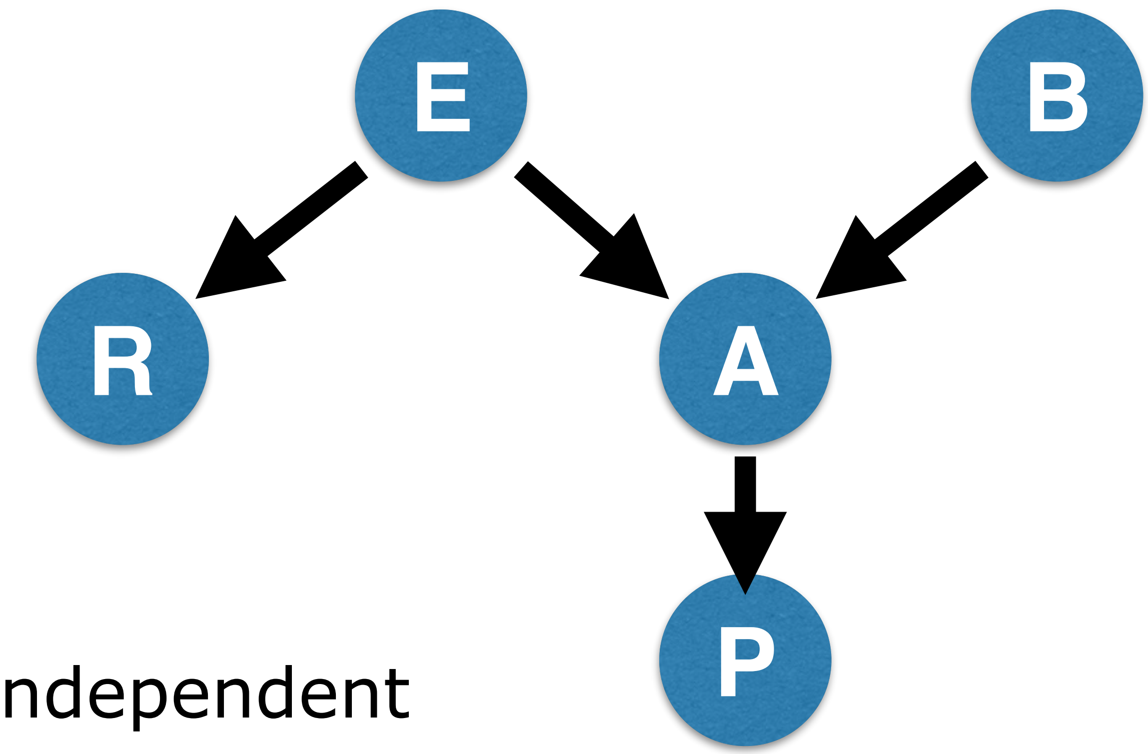
$A|E,B:$

E, B	T	F
T, T	1	0
T, F	1	0
F, T	1	0
F, F	0	1

Binary (T/F) variables:

- **B**: Burglary
- **E**: Earthquake
- **R**: Radio Report
- **A**: Alarm

Marginal Independence



Statements are based on network structure *only* (specific probabilities are irrelevant)

Add event **P** ("police shows up") below **A**.

- If **E** is known, **R** and **A** <u>are</u> conditionally independent ("common cause")
- If **A** is known, **B** and **P** <u>are</u> conditionally independent ("chain")
- If **A** is known, **E** and **B** could be dependent (i.e., are not necessarily conditionally independent, "common effect", "explaining away")
- If **P** is known, **E** and **B** could be dependent (i.e., are not necessarily conditionally independent)

Algorithms for inference

- Desired tasks:
  - Update beliefs: compute marginals (what is the probability of event **X**?)
  - Prediction: conditionals (e.g., given these symptoms, which disease is more likely?)
  - Control: which decision will yield best expected outcome?

- Inference on graphical models is hard (for general BNs), since can model arbitrary constraint satisfaction.

- Prototypical algorithmic techniques: *dynamic programming*, *message-passing*, *belief propagation*, …

- Example: Kalman filtering

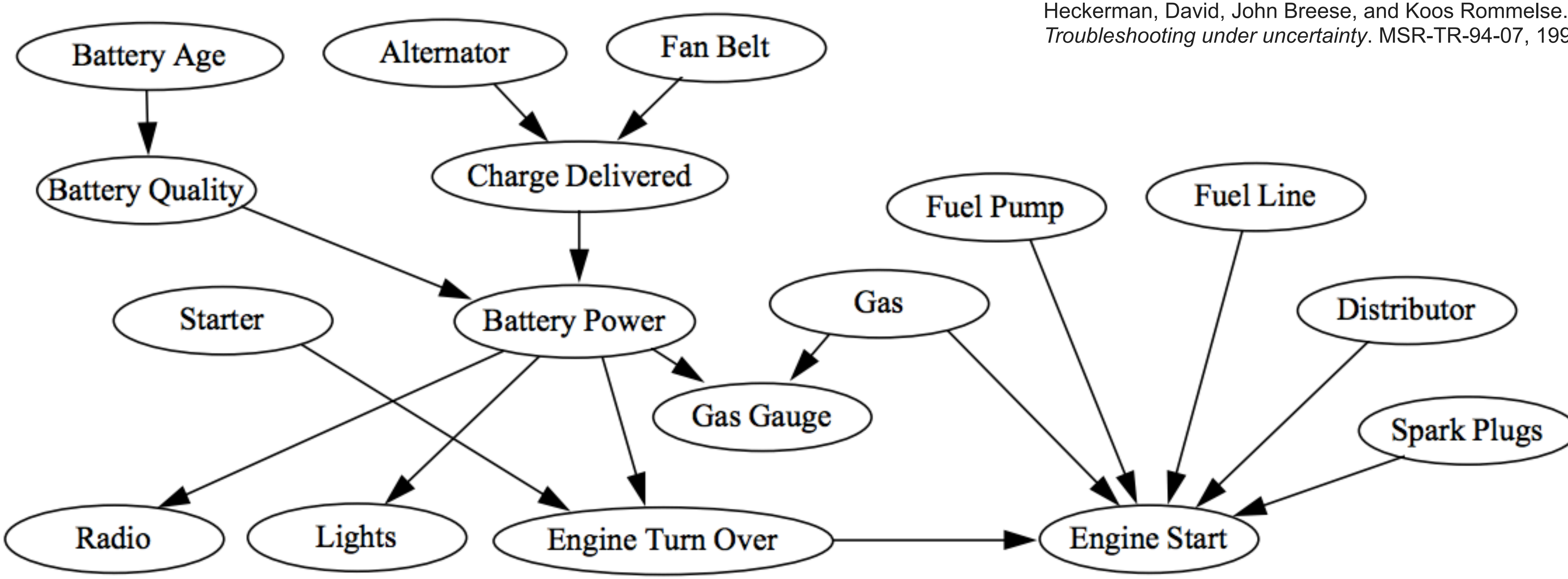
Where do BN models come from?

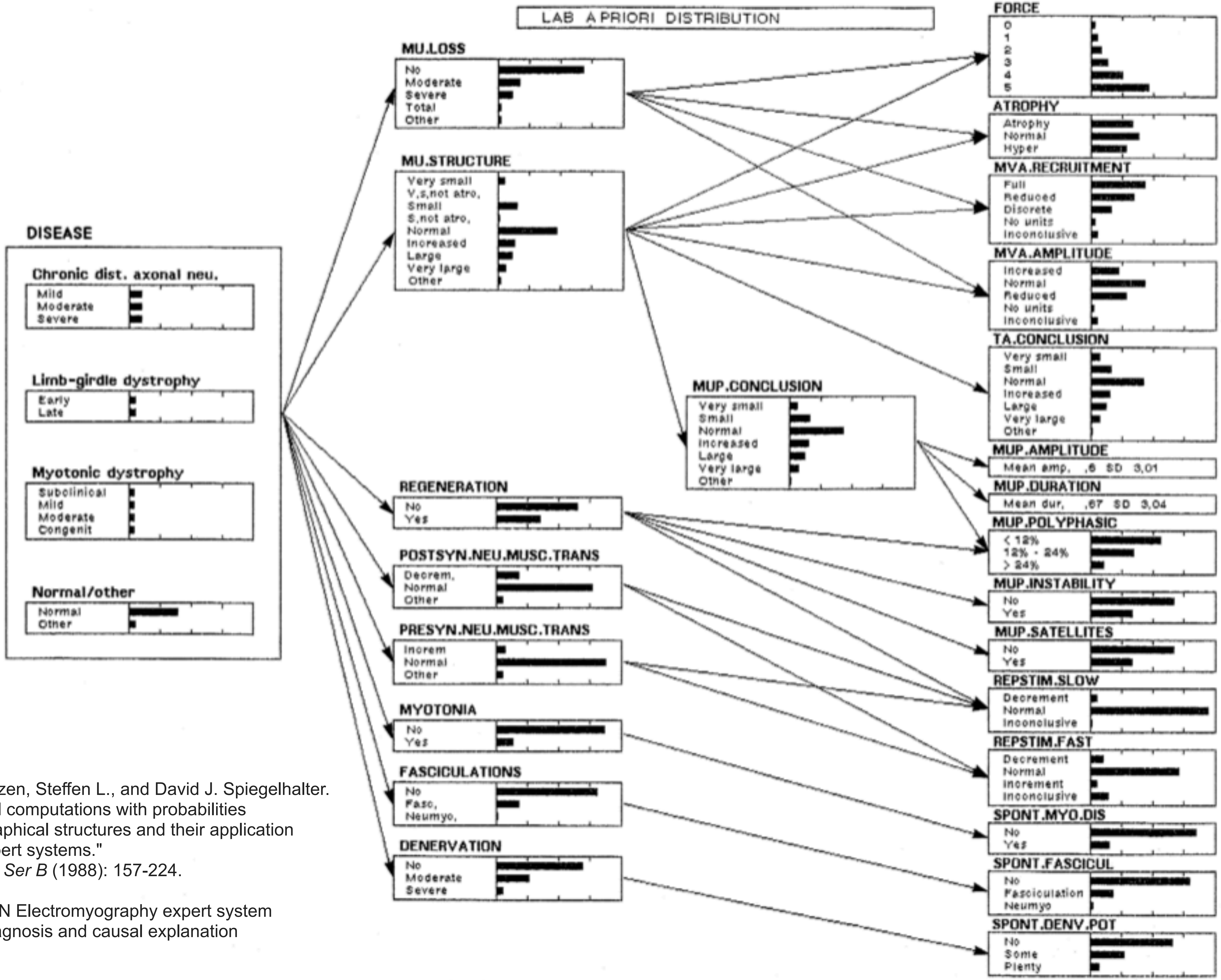
- First-principles modeling, expert knowledge
  - Domain knowledge required (may be expensive)
  - Hard to revise/maintain/update

- *Learned* from data
  - Different versions:
    Known graph, but unknown parameters? Unknown graph?
  - May require extremely large data sets
  - Can be computationally expensive (combinatorial explosion)
  - Explanatory/interpretational power?

Expert knowledge

Heckerman, David, John Breese, and Koos Rommelse. *Troubleshooting under uncertainty*. MSR-TR-94-07, 1994.





Lauritzen, Steffen L., and David J. Spiegelhalter.
"Local computations with probabilities
on graphical structures and their application
to expert systems."
JRSS Ser B (1988): 157-224.

MUNIN Electromyography expert system
for diagnosis and causal explanation

Learning Bayesian Networks

$$
\begin{array}{cccccccccccc}
 & x_1 & x_2 & & & & \cdots & & & & & x_d \\
 & 2 & 2 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 2 \\
 & 1 & 2 & 2 & 2 & 2 & 2 & 3 & 1 & 1 & 2 & 2 \\
D = & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 2 & 2 & 3 & 1 & 1 & 2 & 3 & 3 & 2 & 3 & 2 \\
 & 1 & 2 & 2 & 2 & 2 & 2 & 3 & 1 & 1 & 2 & 2 \\
 & 1 & 1 & 3 & 1 & 1 & 1 & 3 & 1 & 1 & 1 & 3 \\
 & & & & & & \cdots & & & & &
\end{array}
$$

complete data

- **Parameter estimation:** find the maximum likelihood (or Bayesian) estimates of parameters for a graph *G*
- **Model selection:** appropriately score each *G* based on its degree of fit to the data
- **Structure search:** find the highest scoring structure G

Learning Bayesian Networks

$$D = \begin{array}{ccccccccccc} 2 & 2 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 2 & 2 & 2 & 3 & 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 3 & 1 & 1 & 2 & 3 & 3 & 2 & 3 & 2 \\ 1 & 2 & 2 & 2 & 2 & 2 & 3 & 1 & 1 & 2 & 2 \\ 1 & 1 & 3 & 1 & 1 & 1 & 3 & 1 & 1 & 1 & 3 \\ & & & & & \cdots & & & & & \end{array}$$

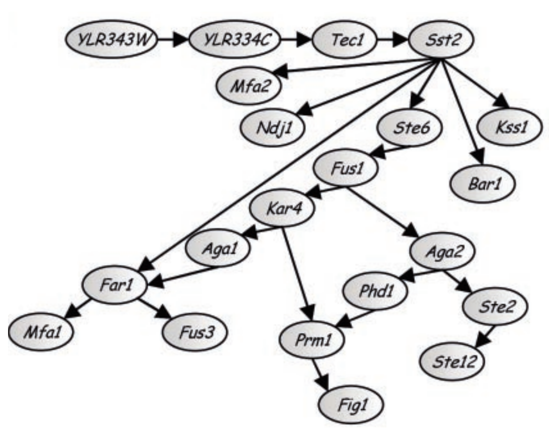
complete data

$P(x_i|x_{pa_i}, \hat{\theta})$ conditional probability estimates for each variable

score$(i|pa_i; D)$ parent selection scores for each variable

$$\text{score}(G; D) = \sum_{i=1}^{d} \text{score}(i|pa_i, D)$$

decomposable scoring function for graphs



$$\arg\max_G \ \text{score}(G; D)$$

highest scoring **acyclic** graph

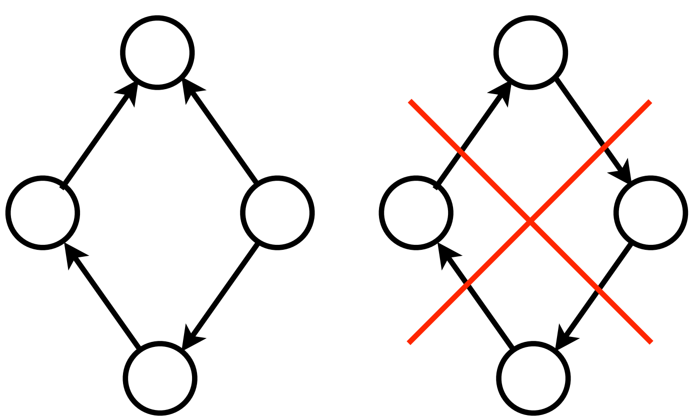
Challenges

- Add a complexity penalty (e.g., Bayesian Information Criterion, BIC) for the number of parameters in the conditional tables

$$\text{score}(i|pa_i; D) = l(i|pa_i; \hat{\theta}) - \frac{(r_i \prod_{j \in pa_i} r_j)}{2} \log(n)$$

BIC score log-likelihood # of parameters/2 log(# of data points)

- Even though graphs are scored locally (score decomposes as sum of scores for each conditional table), selection of structure has global constraint (graph must be acyclic)



- Finding highest-score graph is computationally difficult (some special cases — e.g., trees — may be efficient)

Summary - Modeling and inference

- Graphical models are compact representations of probabilistic descriptions
- Bayesian Networks (directed graphical models) is a rich class, includes mixture models and HMMs
- Dependence/independence properties are often subtle, but can be understood algorithmically
- Relations with causality
- In practice, models are learned from data (scoring functions, etc.), or can developed from first-principles