An Overview of
Learning Bayes Nets From Data

Chris Meek
Microsoft Research

http://research.microsoft.com/~meek

What's and Why's

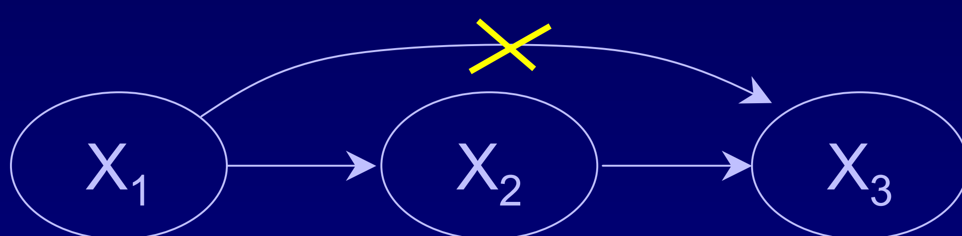
- What is a Bayesian network?
- Why Bayesian networks are useful?
- Why learn a Bayesian network?

What is a Bayesian Network?

also called belief networks, and (directed acyclic) graphical models

- Directed acyclic graph
  - Nodes are variables (discrete or continuous)
  - Arcs indicate dependence between variables.
- Conditional Probabilities (local distributions)

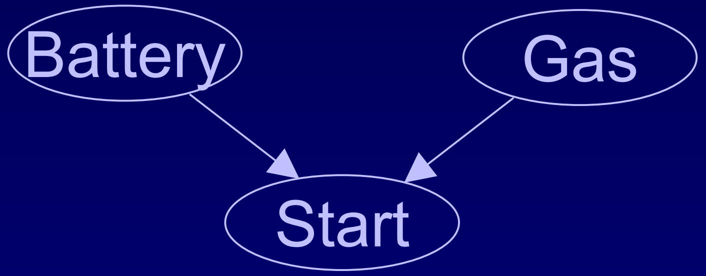
- Missing arcs implies conditional independence
- Independencies + local distributions => modular specification of a joint distribution



$$p(x_1)\; p(x_2 \mid x_1)\; p(x_3 \mid \cancel{x_1}, x_2) = p(x_1, x_2, x_3)$$

Why Bayesian Networks?

- Expressive language
  - Finite mixture models, Factor analysis, HMM, Kalman filter,…
- Intuitive language
  - Can utilize causal knowledge in constructing models
  - Domain experts comfortable building a network
- General purpose "inference" algorithms
  - P(Bad Battery | Has Gas, Won't Start)



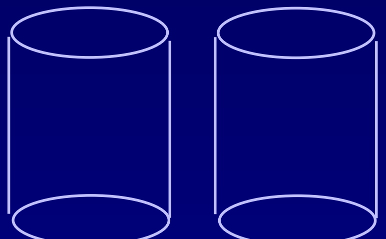
  - Exact: Modular specification leads to large computational efficiencies
  - Approximate: "Loopy" belief propagation

Why Learning?

knowledge-based
(expert systems)

-Answer Wizard, Office 95, 97, & 2000
-Troubleshooters, Windows 98 & 2000

-Causal discovery
-Data visualization
-Concise model of data
-Prediction



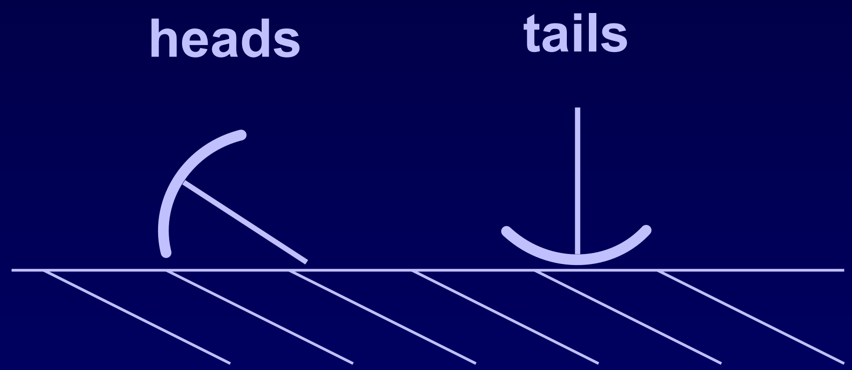
data-based

Overview

- Learning Probabilities (local distributions)
  - Introduction to Bayesian statistics: Learning a probability
  - Learning probabilities in a Bayes net
  - Applications
- Learning Bayes-net structure
  - Bayesian model selection/averaging
  - Applications

Learning Probabilities: Classical Approach

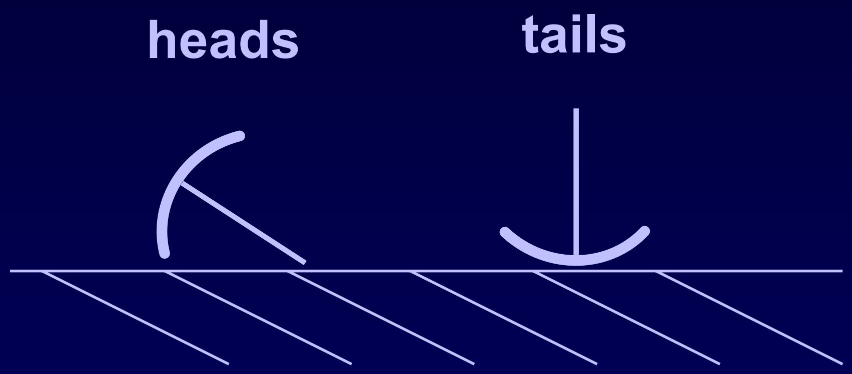
Simple case: Flipping a thumbtack



<u>True probability</u> θ is unknown

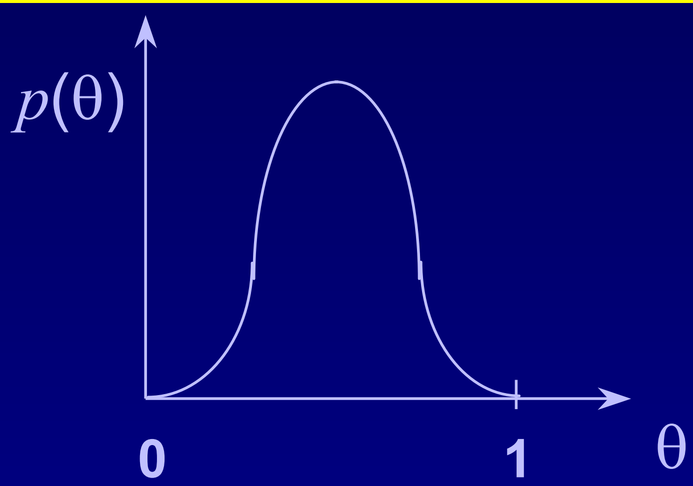
Given iid data, estimate θ using an estimator with good properties: low bias, low variance, consistent (e.g., ML estimate)

Learning Probabilities: Bayesian Approach



True probability θ is unknown

Bayesian probability density for θ



Bayesian Approach: use Bayes' rule to compute a new density for θ given data

prior

likelihood

posterior

$$p(\theta \mid data) = \frac{p(\theta)\, p(\mathrm{d}ata \mid \theta)}{\int p(\theta)\, p(\mathrm{d}ata \mid \theta)\, d\theta}$$

$$\propto p(\theta)\, p(data \mid \theta)$$

The Likelihood

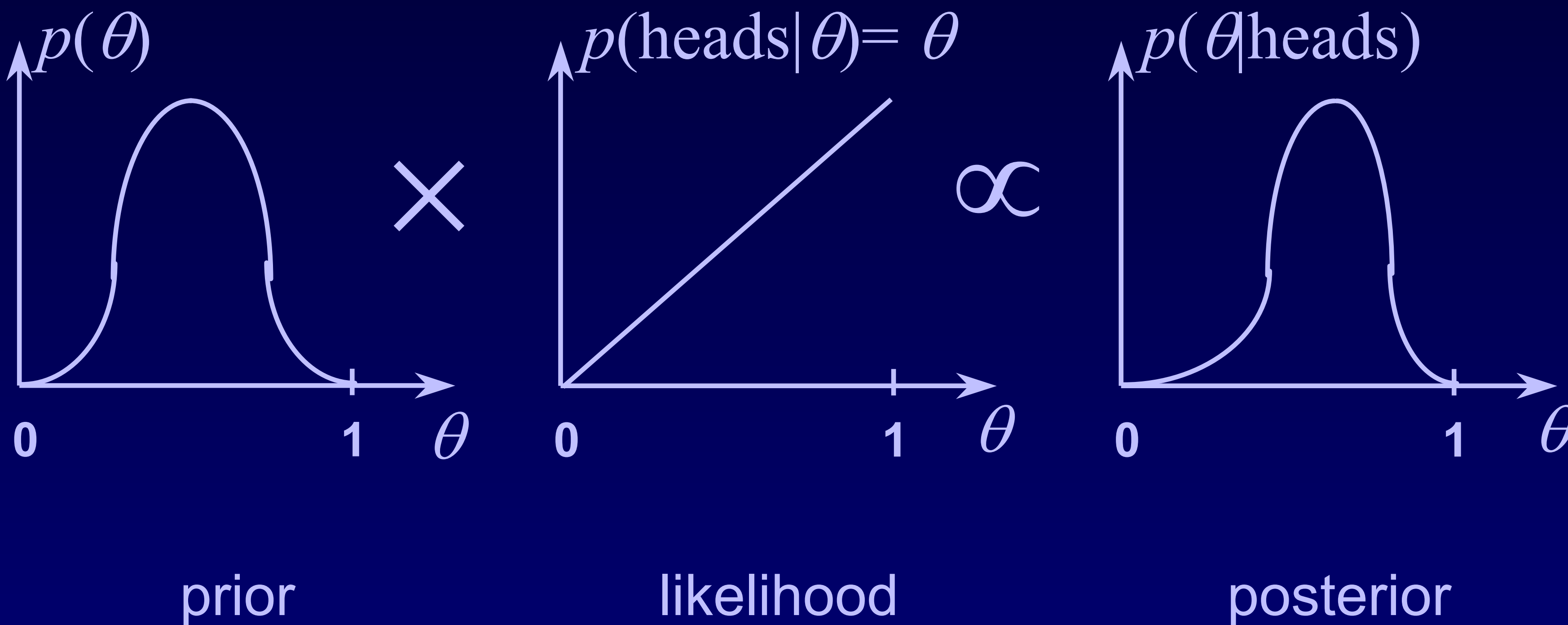
$$p(\text{heads} \mid \theta) = \theta$$

$$p(\text{tails} \mid \theta) = (1-\theta)$$

$$p(hhth \ldots ttth \mid \theta) = \theta^{\#h}(1-\theta)^{\#t}$$

"binomial distribution"

Example: Application of Bayes rule to the observation of a single "heads"



The probability of heads on the next toss

$$
\begin{aligned}
p(h \mid \mathbf{d}) &= \int p(h \mid \theta, \cancel{\mathbf{d}})\; p(\theta \mid \mathbf{d})\, d\theta \\
&= \int \theta \;\; p(\theta \mid \mathbf{d})\, d\theta \\
&= E_{p(\theta \mid \mathbf{d})}(\theta)
\end{aligned}
$$

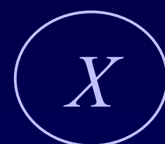
Note: This yields nearly identical answers to ML estimates when one uses a “flat” prior

Overview

- Learning Probabilities
  - Introduction to Bayesian statistics: Learning a probability
  - Learning probabilities in a Bayes net
  - Applications
- Learning Bayes-net structure
  - Bayesian model selection/averaging
  - Applications

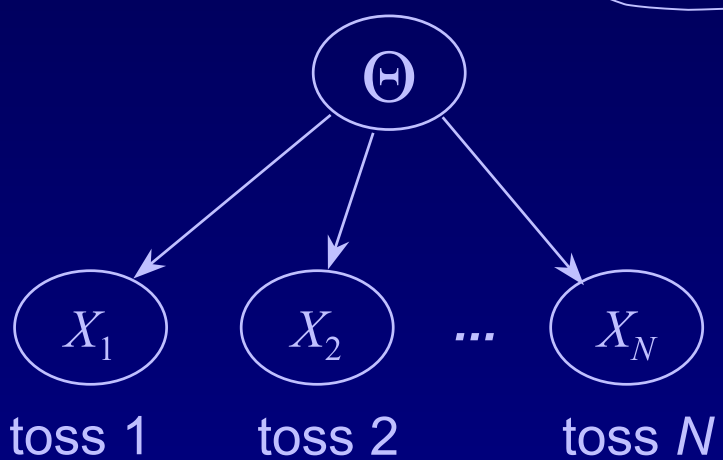
From thumbtacks to Bayes nets

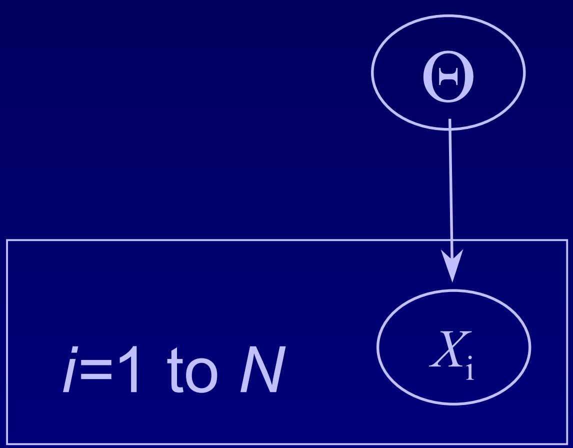
Thumbtack problem can be viewed as learning the probability for a very simple BN:



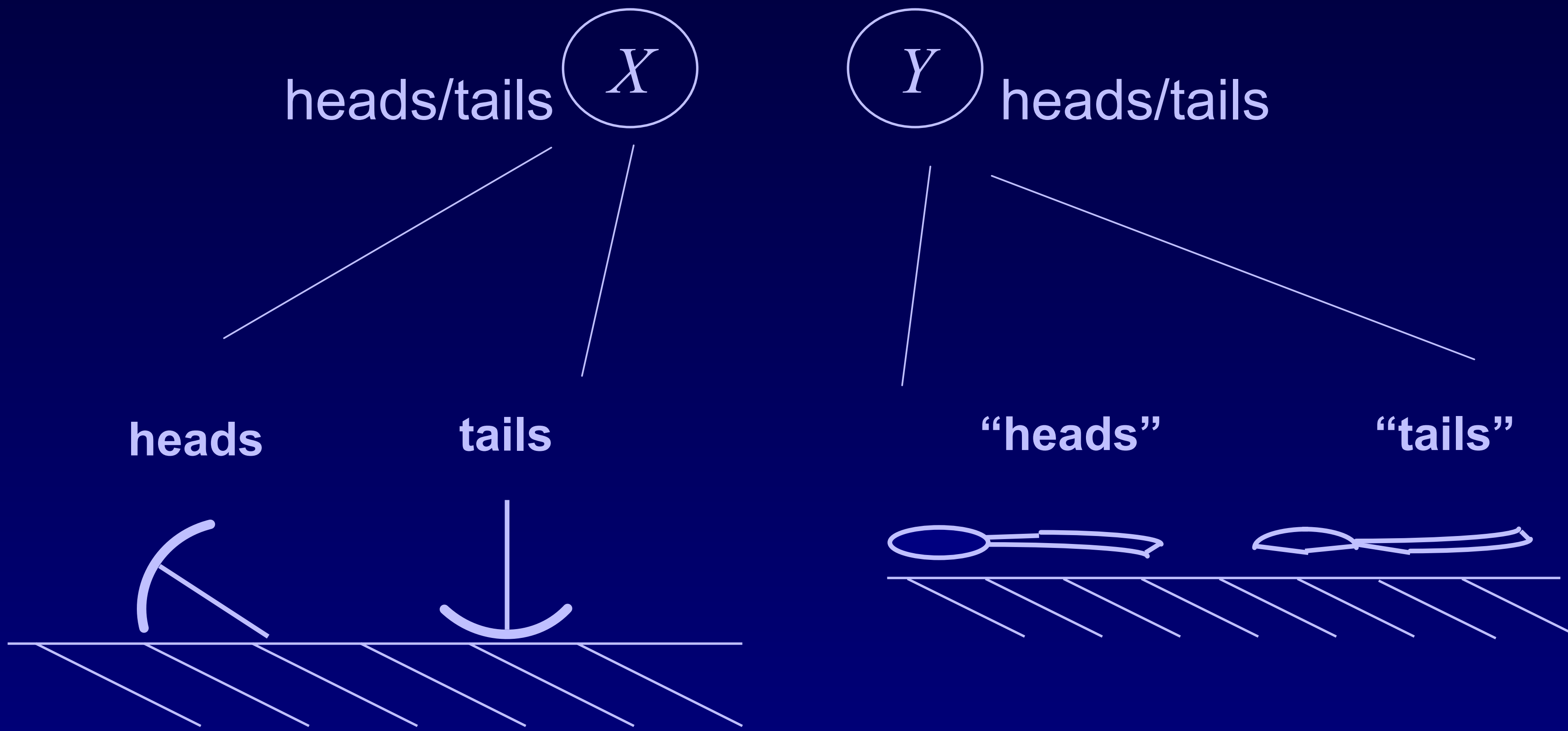
heads/tails

$$P\left(X = heads\right) = f\left(\theta\right)$$

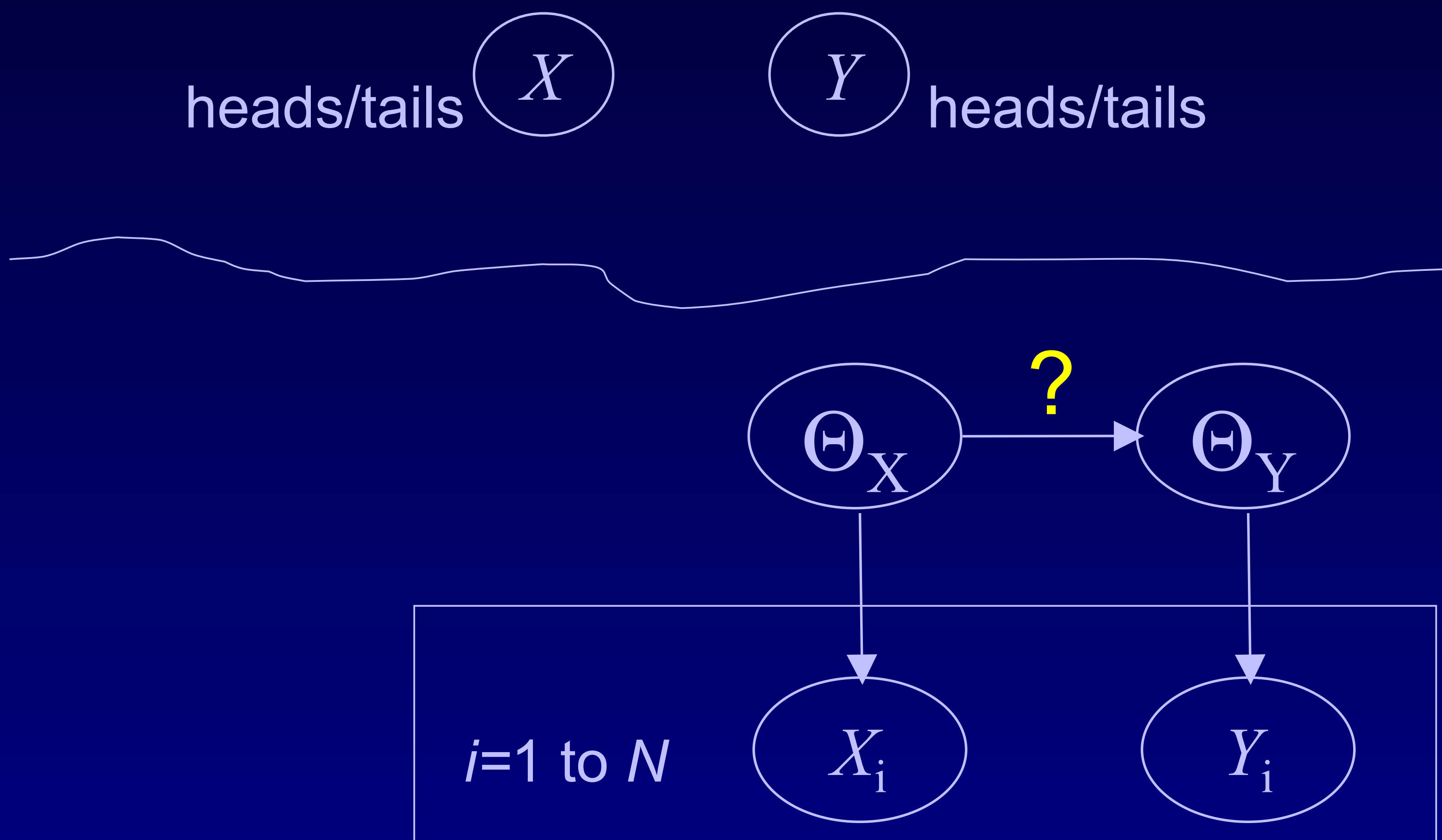




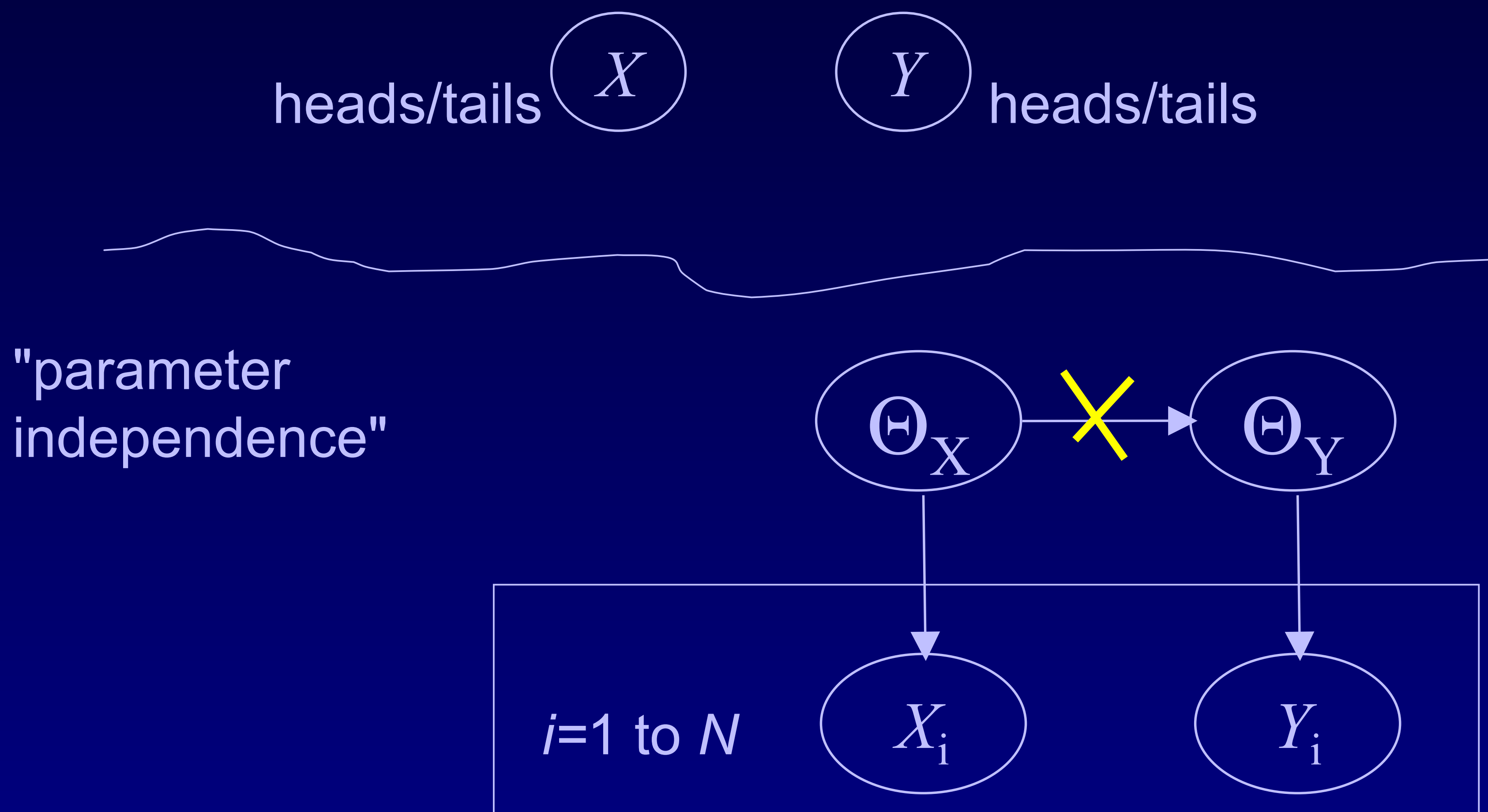
The next simplest Bayes net



The next simplest Bayes net



The next simplest Bayes net

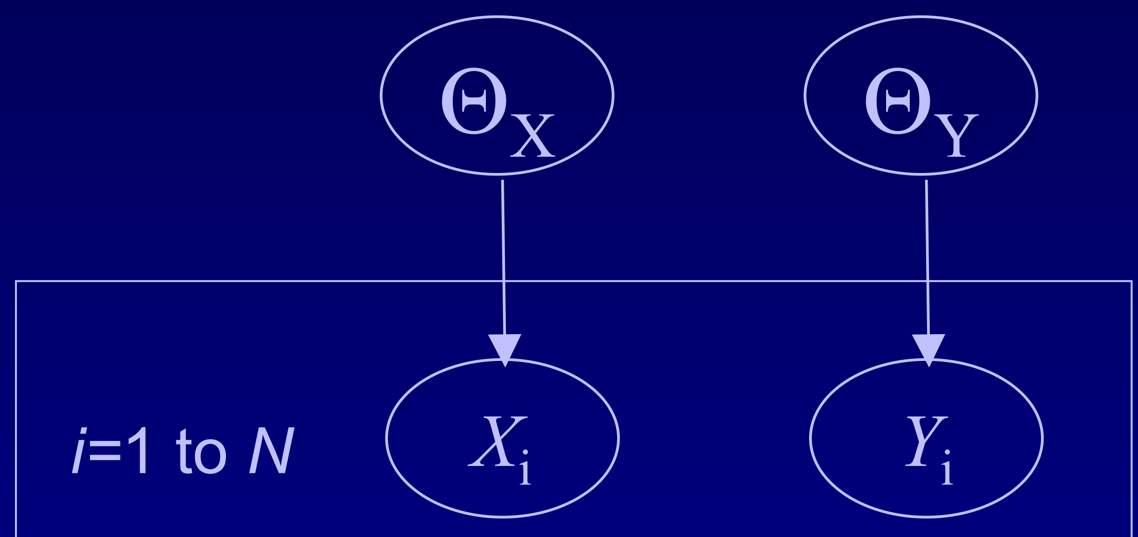


The next simplest Bayes net

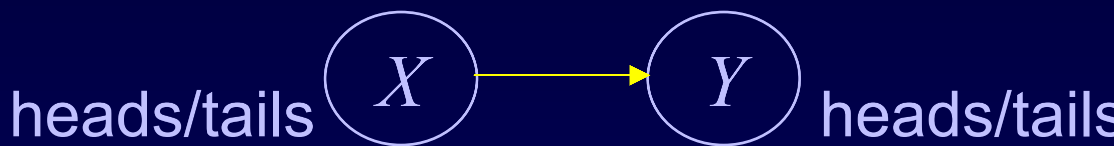
"parameter independence"

$\Downarrow$

two separate thumbtack-like learning problems



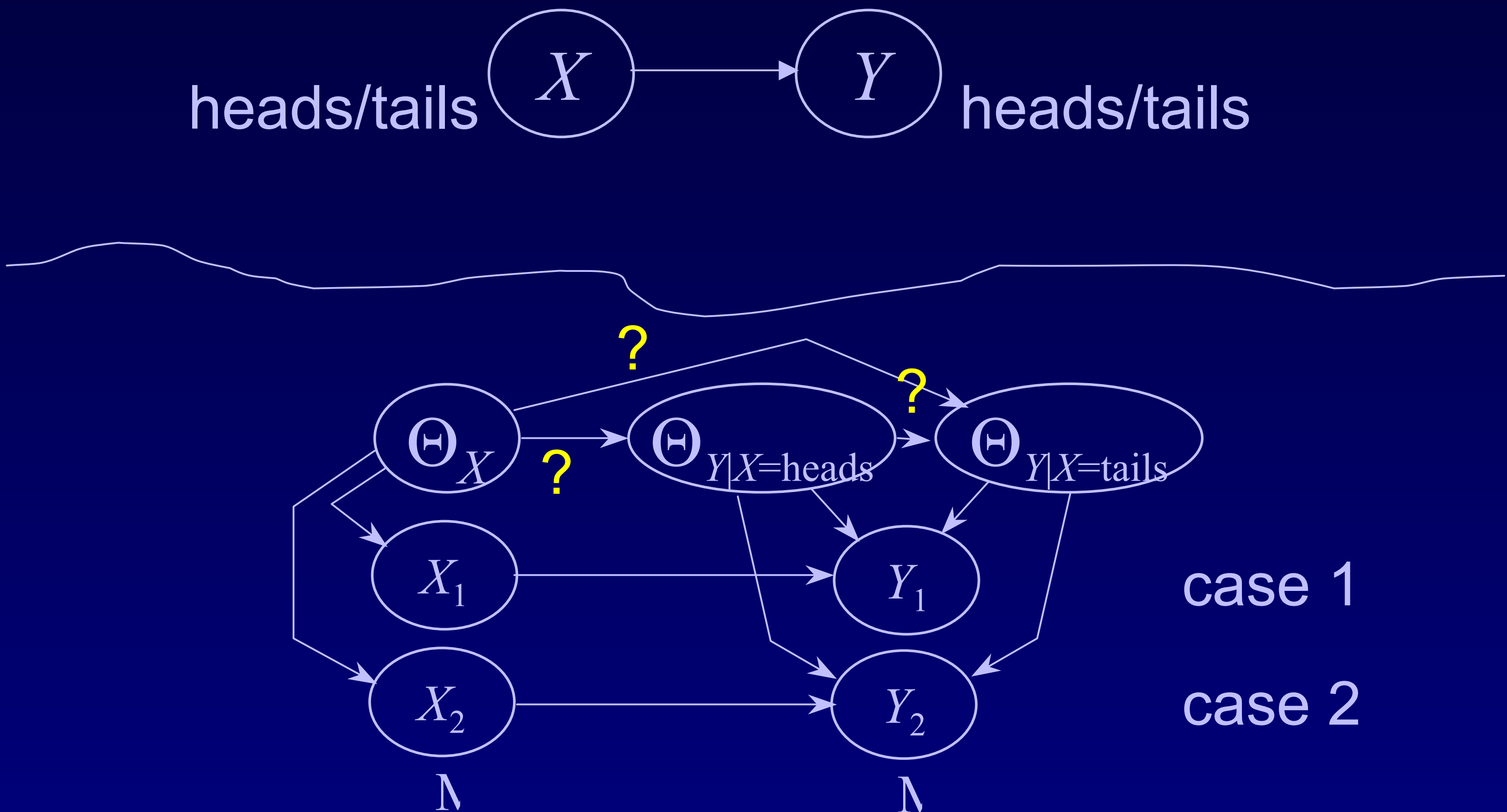
A bit more difficult...



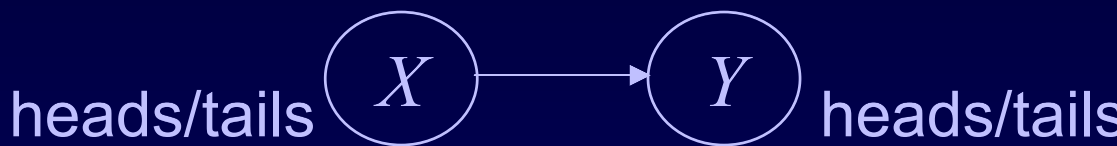
Three probabilities to learn:

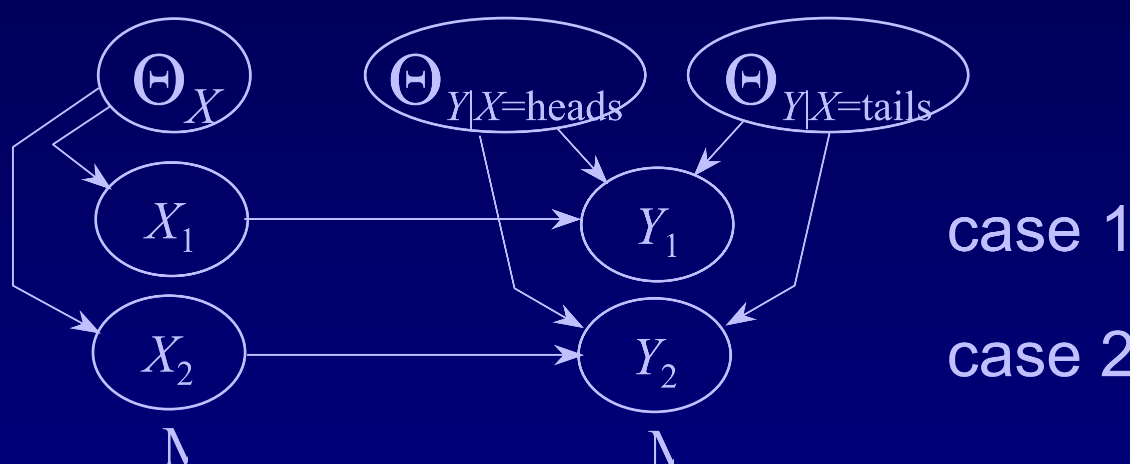
- $\theta_{X=\text{heads}}$
- $\theta_{Y=\text{heads}|X=\text{heads}}$
- $\theta_{Y=\text{heads}|X=\text{tails}}$

A bit more difficult...

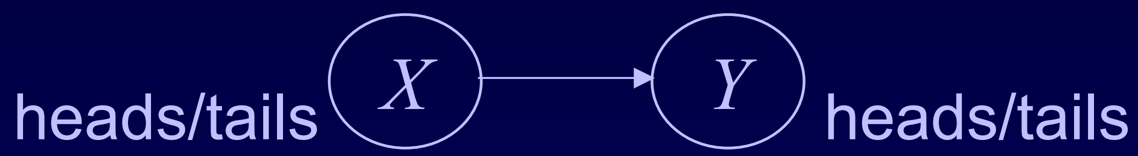


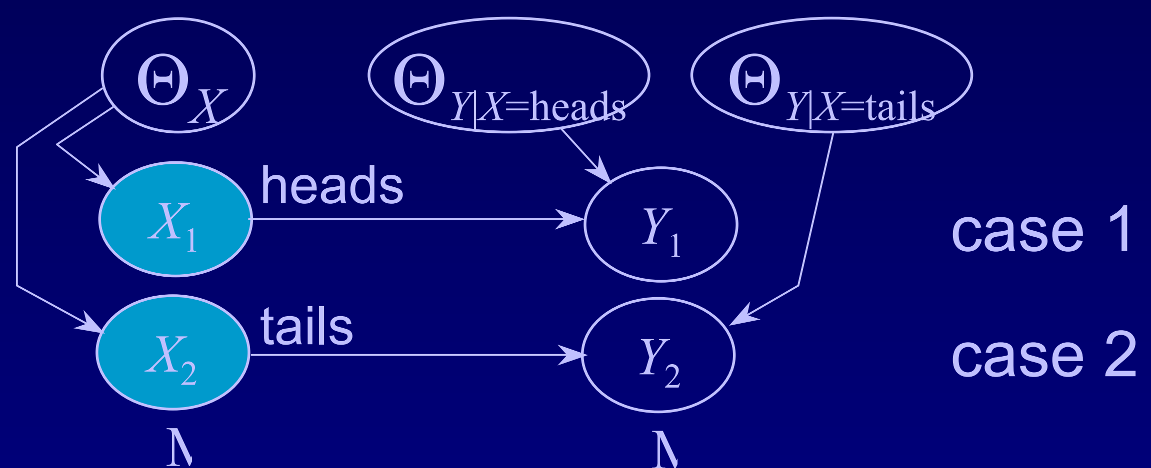
A bit more difficult...





A bit more difficult...





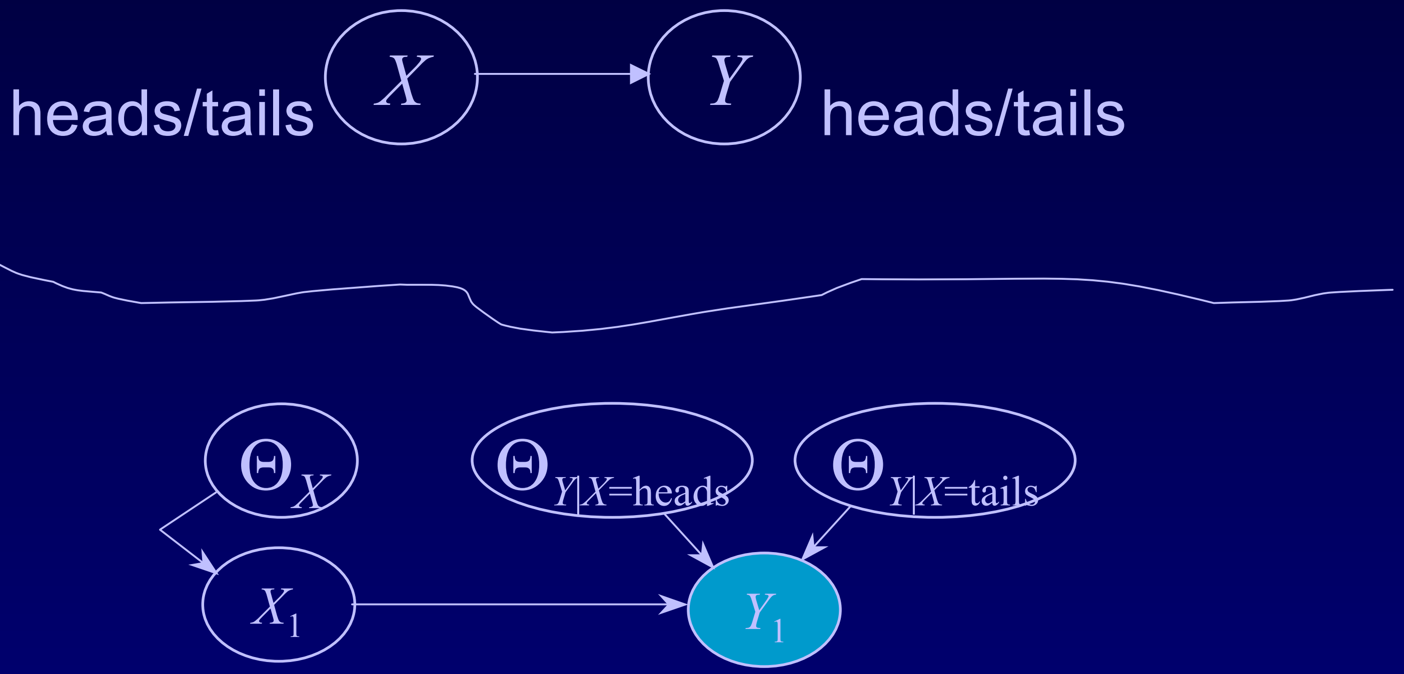
3 separate thumbtack-like problems

In general…

Learning probabilities in a BN is straightforward if

- Likelihoods from the exponential family (multinomial, poisson, gamma, ...)
- Parameter independence
- Conjugate priors
- Complete data

Incomplete data makes parameters dependent



Incomplete data

- Incomplete data makes parameters dependent

Parameter Learning for incomplete data

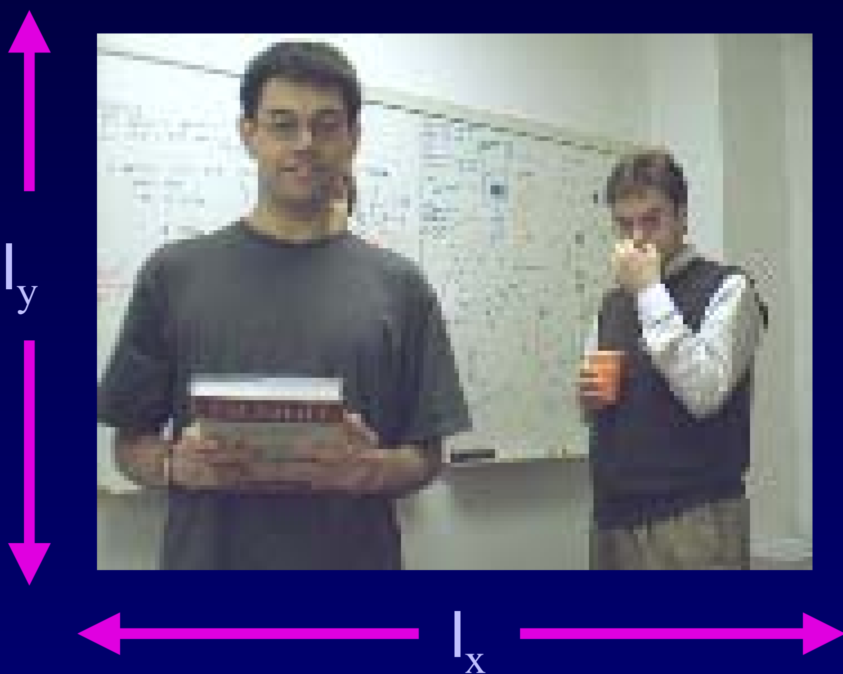
- Monte-Carlo integration
  - Investigate properties of the posterior and perform prediction
- Large-sample Approx. (Laplace/Gaussian approx.)
  - Expectation-maximization (EM) algorithm and inference to compute mean and variance.
- Variational methods

Overview

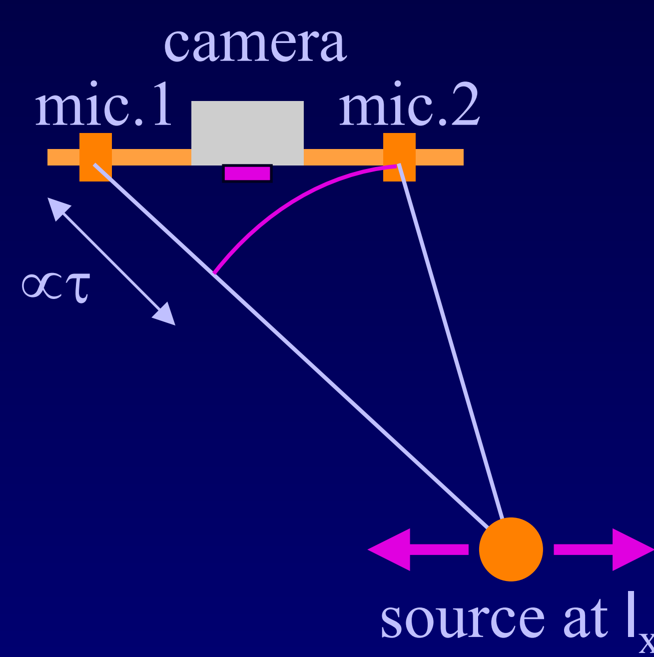
- Learning Probabilities
  - Introduction to Bayesian statistics: Learning a probability
  - Learning probabilities in a Bayes net
  - Applications
- Learning Bayes-net structure
  - Bayesian model selection/averaging
  - Applications

Example: Audio-video fusion
Beal, Attias, & Jojic 2002

Video scenario



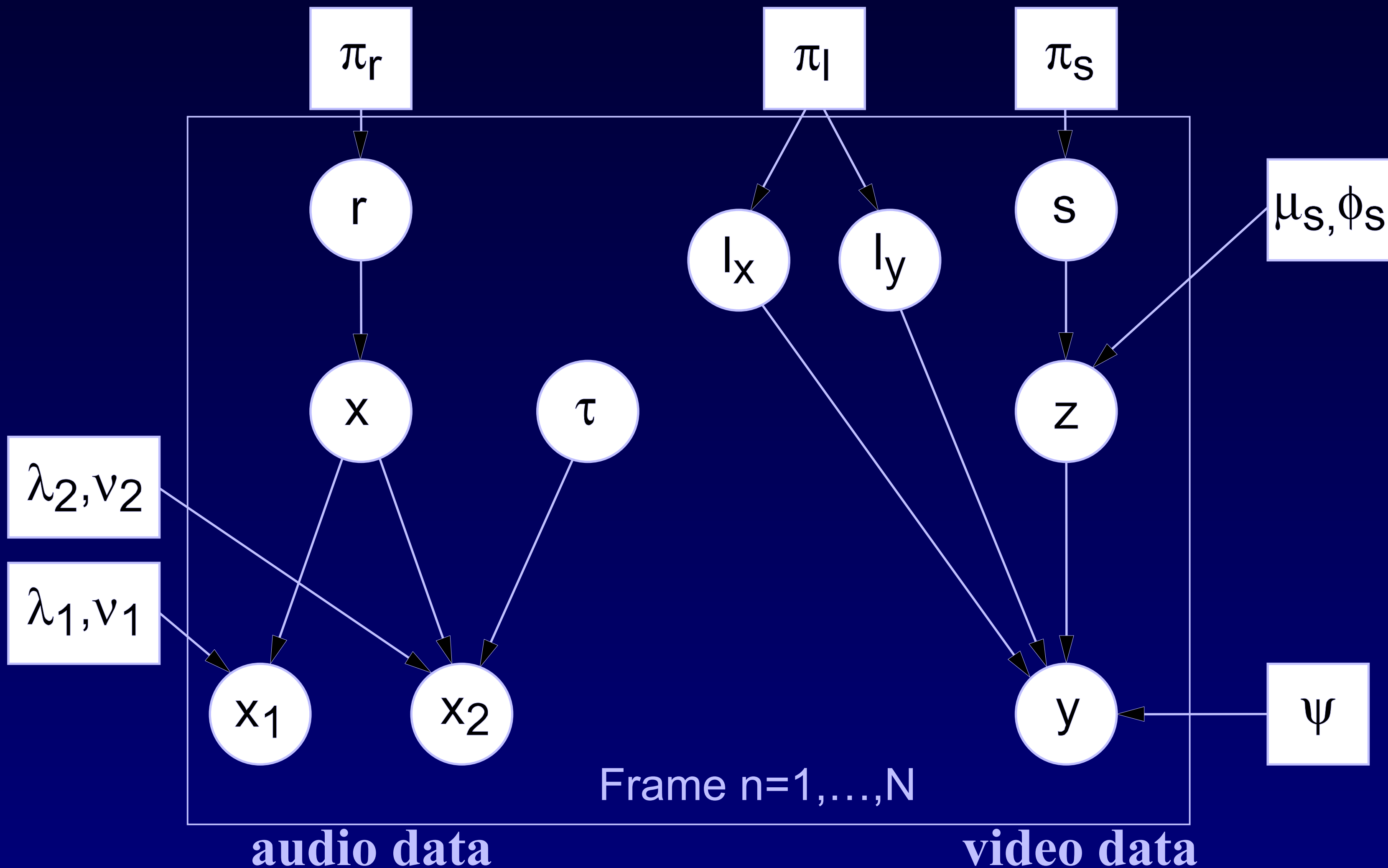
Audio scenario



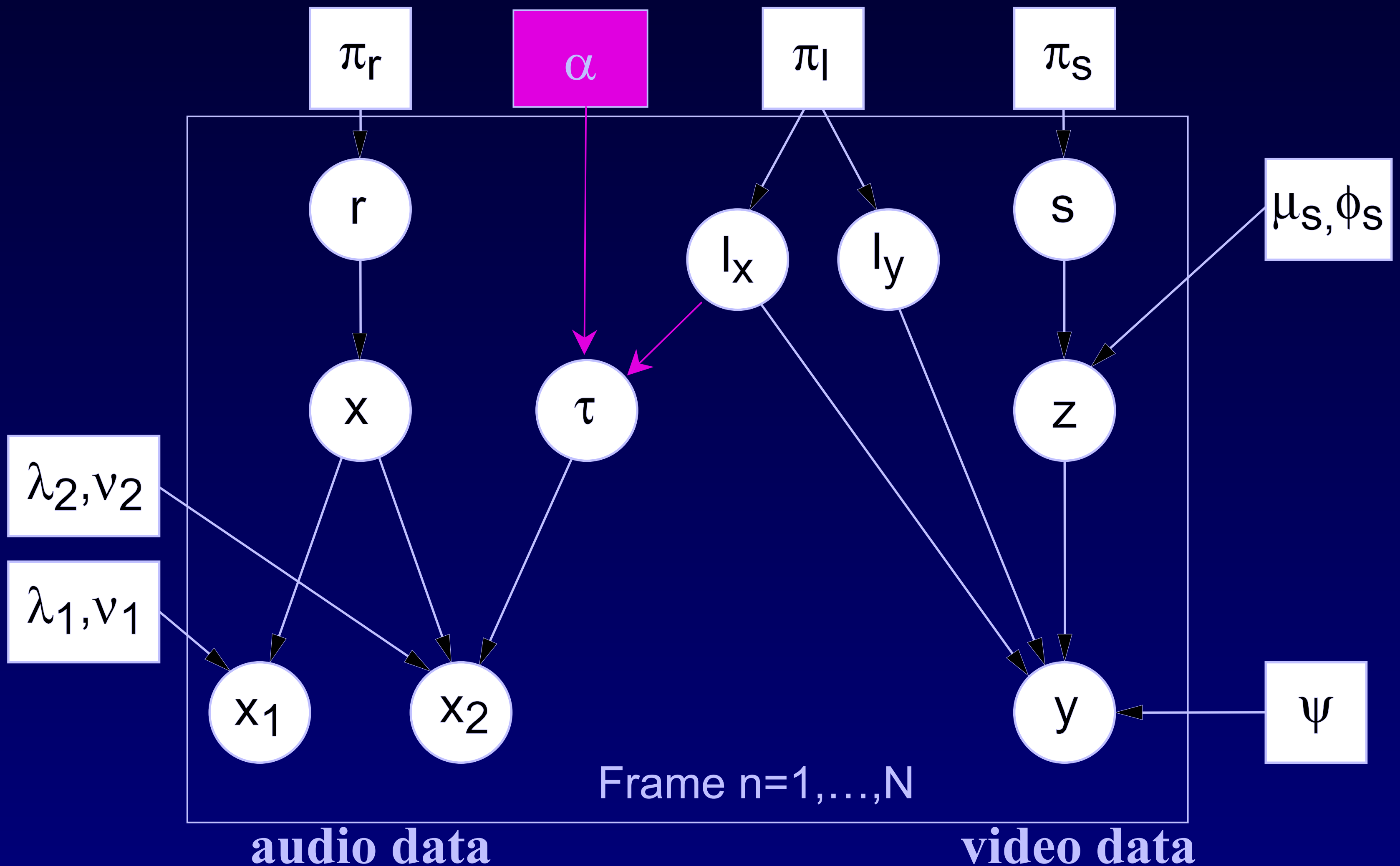
Goal: detect and track speaker

Slide courtesy Beal, Attias and Jojic

Separate audio-video models

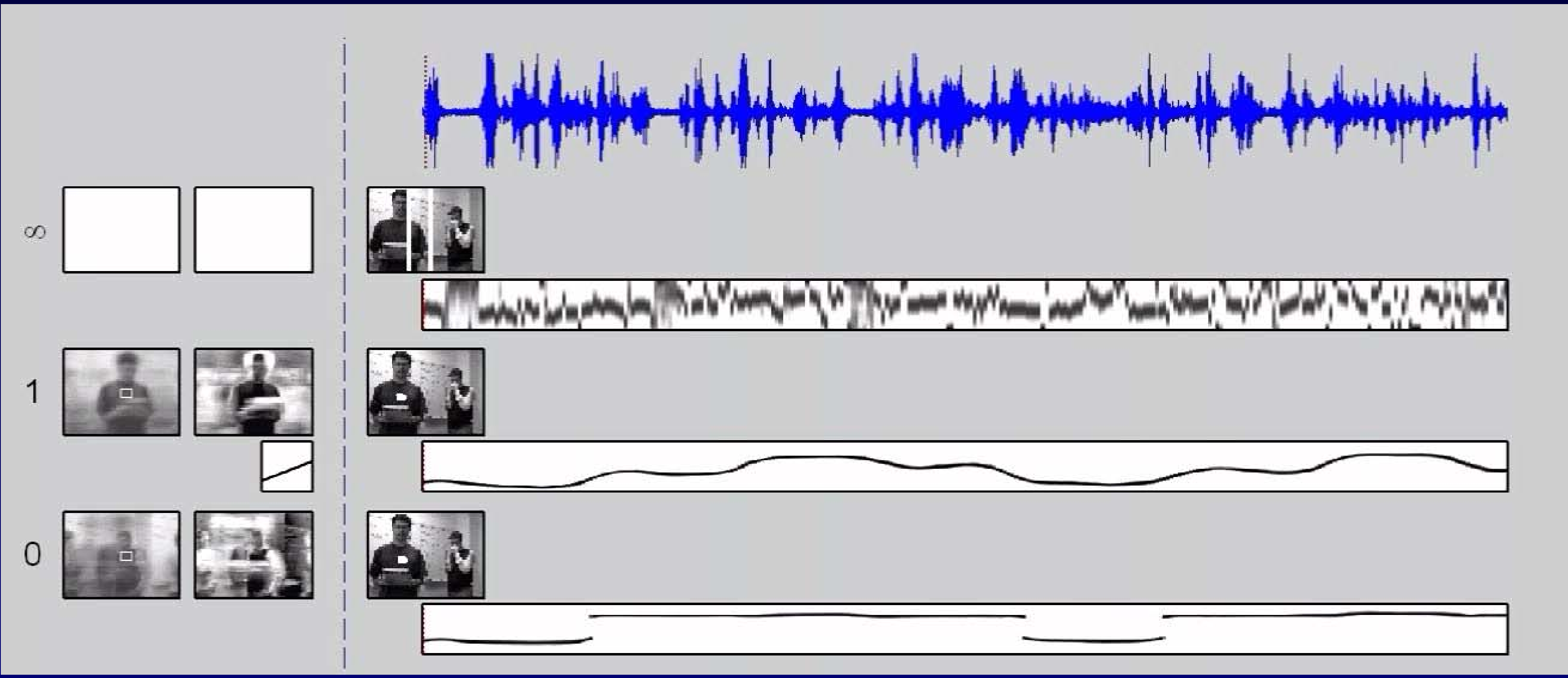


Combined model



Slide courtesy Beal, Attias and Jojic

Tracking Demo



Slide courtesy Beal, Attias and Jojic

Overview

- Learning Probabilities
  - Introduction to Bayesian statistics: Learning a probability
  - Learning probabilities in a Bayes net
  - Applications
- Learning Bayes-net structure
  - Bayesian model selection/averaging
  - Applications

Two Types of Methods for Learning BNs

- Constraint based
  - Finds a Bayesian network structure whose implied independence constraints "match" those found in the data.
- Scoring methods (Bayesian, MDL, MML)
  - Find the Bayesian network structure that can represent distributions that "match" the data (i.e. could have generated the data).

Learning Bayes-net structure

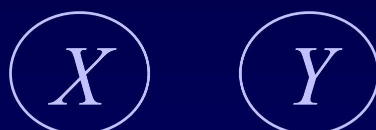
Given data, which model is correct?

model 1: $X \qquad Y$

model 2: $X \rightarrow Y$

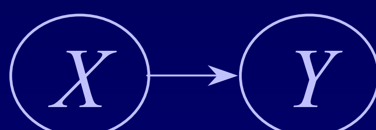
Bayesian approach

Given data, which model is correct? more likely?

model 1: 

$p(m_1) = 0.7$ $\qquad$ $p(m_1 \mid \mathbf{d}) = 0.1$

Data $\mathbf{d}$ $\Longrightarrow$

model 2: 

$p(m_2) = 0.3$ $\qquad$ $p(m_2 \mid \mathbf{d}) = 0.9$

Bayesian approach: Model Averaging

Given data, which model is correct? more likely?

model 1: $X \quad Y$ $\quad p(m_1) = 0.7$ $\qquad p(m_1 \mid \mathbf{d}) = 0.1$

Data $\mathbf{d}$ $\Longrightarrow$

model 2: $X \rightarrow Y$ $\quad p(m_2) = 0.3$ $\qquad p(m_2 \mid \mathbf{d}) = 0.9$

average predictions

Bayesian approach: Model Selection

Given data, which model is correct? more likely?

model 1: $X \quad Y$ $\quad p(m_1) = 0.7 \quad\quad p(m_1 \mid \mathbf{d}) = 0.1$

Data $\mathbf{d}$ $\Longrightarrow$

model 2: $X \rightarrow Y$ $\quad p(m_2) = 0.3 \quad\quad p(m_2 \mid \mathbf{d}) = 0.9$

Keep the best model:
- Explanation
- Understanding
- Tractability

To score a model, use Bayes rule

Given data $\mathbf{d}$:

model score

$$p(m \mid \mathbf{d}) \propto p(m)\,\underline{p(\mathbf{d} \mid m)}$$

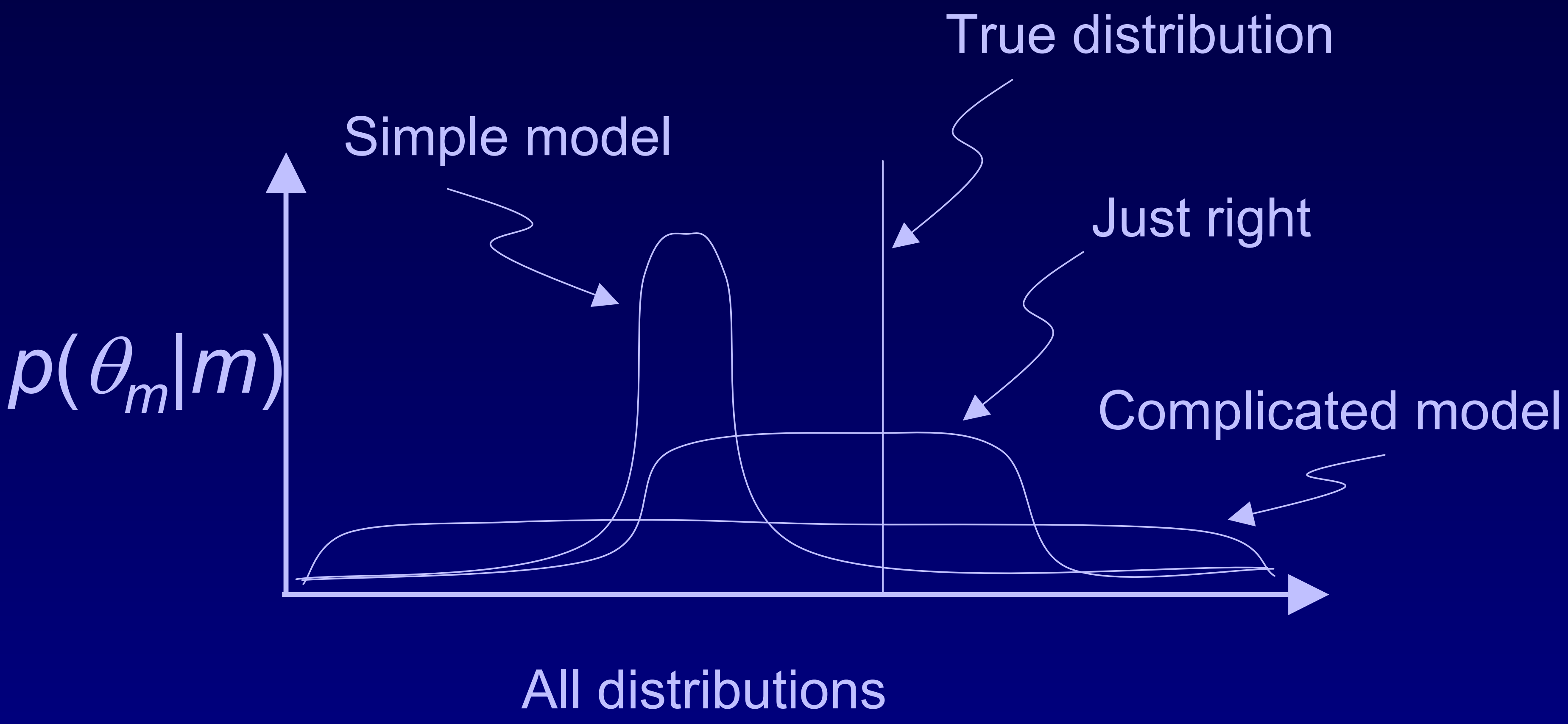
"marginal likelihood"

likelihood

$$p(\mathbf{d} \mid m) = \int p(\mathbf{d} \mid \theta, m)\, p(\theta \mid m)\, d\theta$$

The Bayesian approach and Occam's Razor

$$p(\mathbf{d} \mid m) = \int p(\mathbf{d} \mid \theta_m, m)\, p(\theta_m \mid m)\, d\theta_m$$



Computation of Marginal Likelihood

Efficient closed form if

- Likelihoods from the exponential family (binomial, poisson, gamma, ...)
- Parameter independence
- Conjugate priors
- No missing data, including <u>no hidden variables</u>

Else use approximations

- Monte-Carlo integration
- Large-sample approximations
- Variational methods

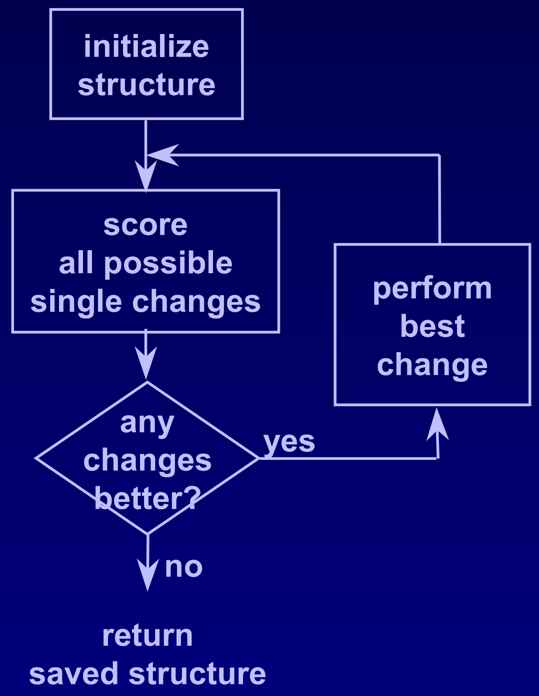
Practical considerations

The number of possible BN structures is super exponential in the number of variables.

- How do we find the best graph(s)?

Model search

- Finding the BN structure with the highest score among those structures with at most *k* parents is NP hard for *k*>1 (Chickering, 1995)
- Heuristic methods
  - Greedy
  - Greedy with restarts
  - MCMC methods



Learning the correct model

- True graph G and P is the generative distribution
- Markov Assumption: P satisfies the independencies implied by G
- Faithfulness Assumption: P satisfies only the independencies implied by G

Theorem: Under Markov and Faithfulness, with enough data generated from P one can recover G (up to equivalence). Even with the greedy method!

Learning Bayes Nets From Data

data

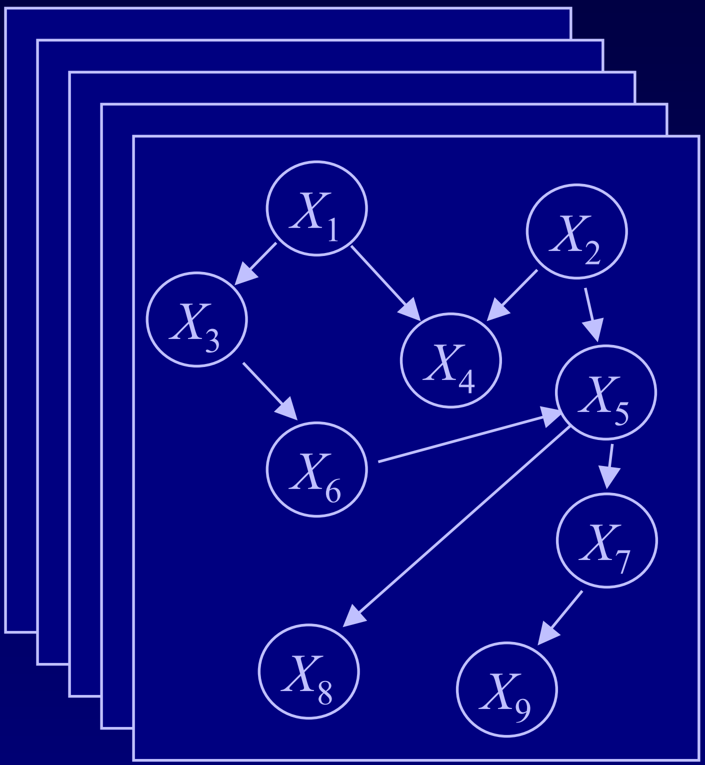
X_1	X_2	X_3	
true	1	Red	
false	5	Blue	
false	3	Green	...
true	2	Red	
	⋮		⋱

+

prior/expert information

Bayes-net learner

Bayes net(s)



Overview

- Learning Probabilities
  - Introduction to Bayesian statistics: Learning a probability
  - Learning probabilities in a Bayes net
  - Applications
- Learning Bayes-net structure
  - Bayesian model selection/averaging
  - Applications

Preference Prediction (a.k.a. Collaborative Filtering)

- Example: Predict what products a user will likely purchase given items in their shopping basket
- Basic idea: use other people's preferences to help predict a new user's preferences.

- Numerous applications
  - Tell people about books or web-pages of interest
  - Movies
  - TV shows

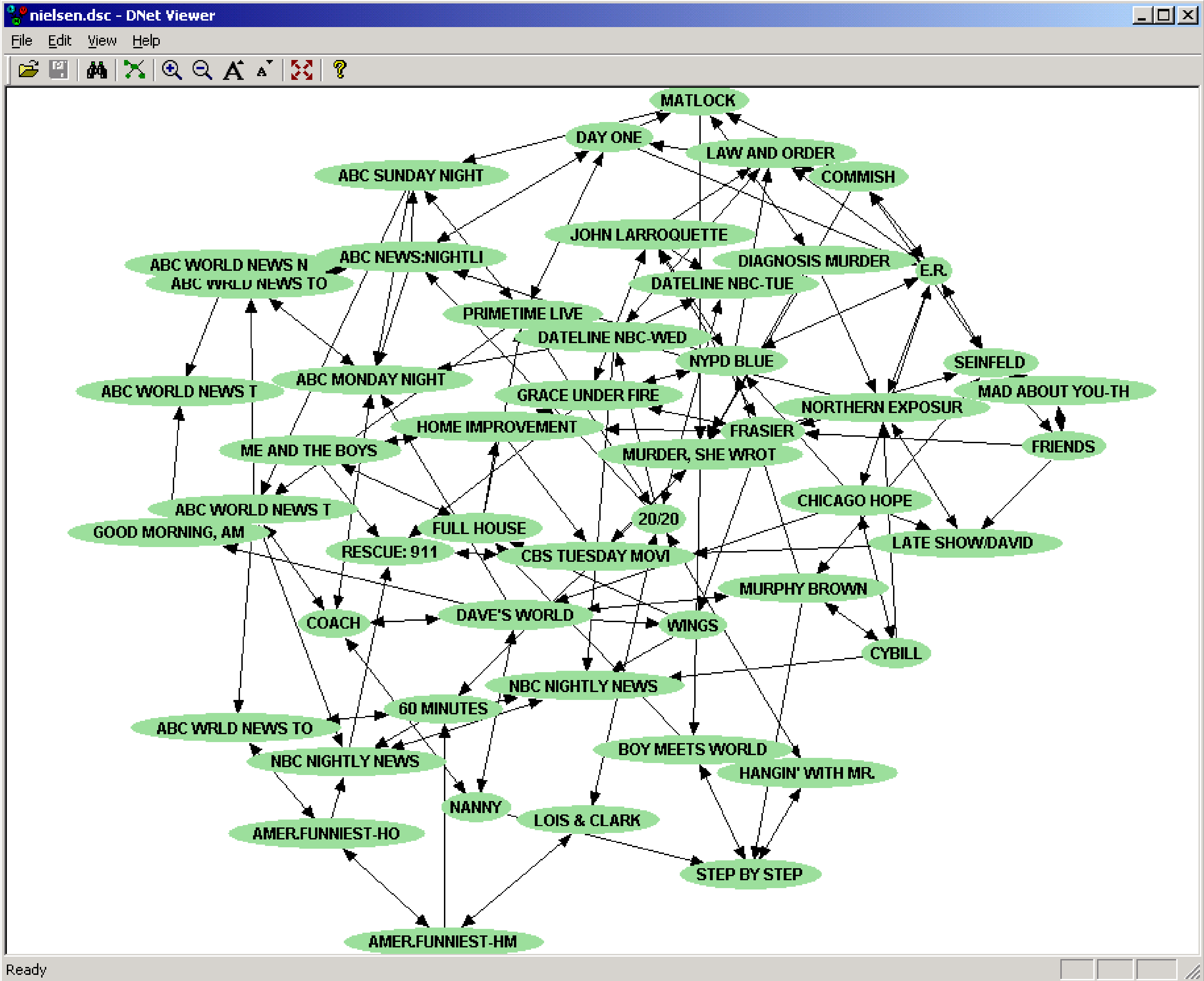
Example: TV viewing

Nielsen data: 2/6/95-2/19/95

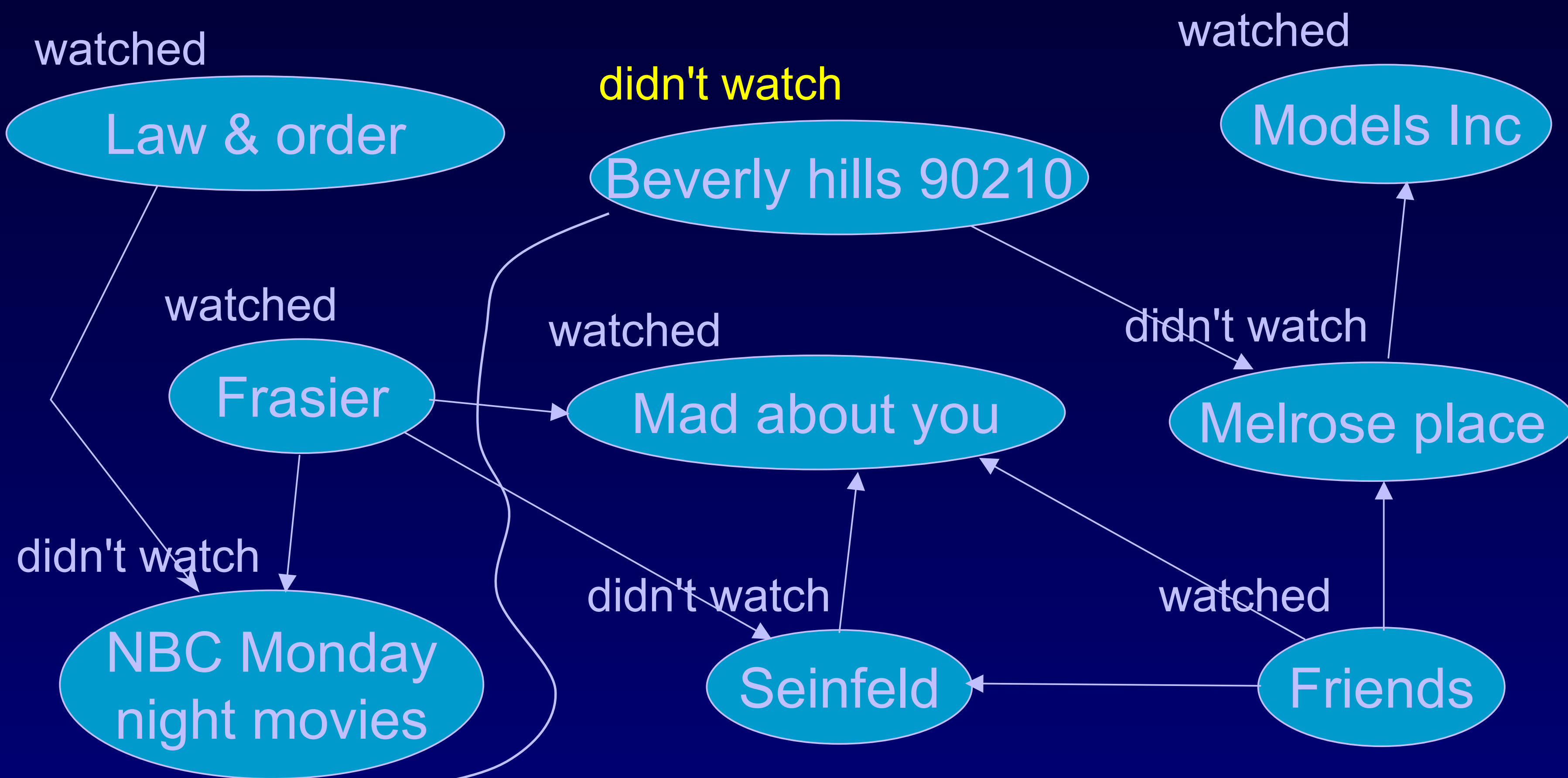
	Show1	Show2	Show3	
viewer 1	y	n	n	
viewer 2	n	y	y	...
viewer 3	n	n	n	
		etc.		

~200 shows, ~3000 viewers

Goal: For each viewer, recommend shows they haven't watched that they are likely to watch

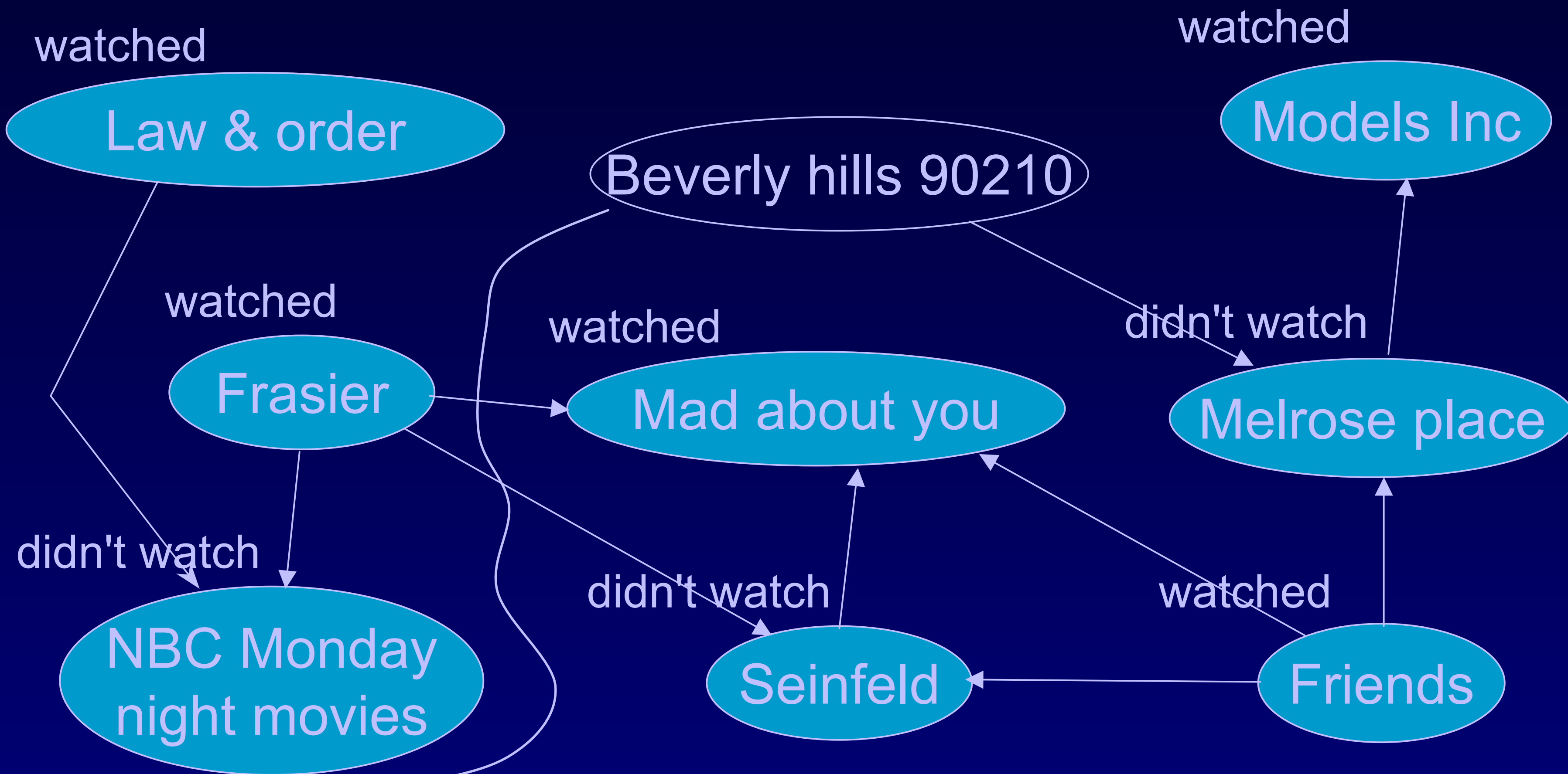

nielsen.dsc - DNet Viewer
File
Edit
View
Help
MATLOCK
DAY ONE
LAW AND ORDER
ABC SUNDAY NIGHT
COMMISH
JOHN LARROQUETTE
ABC NEWS:NIGHTLI
ABC WORLD NEWS N
ABC WRLD NEWS TO
DIAGNOSIS MURDER
E.R.
DATELINE NBC-TUE
PRIMETIME LIVE
DATELINE NBC-WED
NYPD BLUE
SEINFELD
ABC MONDAY NIGHT
ABC WORLD NEWS T
GRACE UNDER FIRE
MAD ABOUT YOU-TH
NORTHERN EXPOSUR
HOME IMPROVEMENT
FRASIER
ME AND THE BOYS
MURDER, SHE WROT
FRIENDS
CHICAGO HOPE
ABC WORLD NEWS T
20/20
GOOD MORNING, AM
FULL HOUSE
LATE SHOW/DAVID
RESCUE: 911
CBS TUESDAY MOVI
MURPHY BROWN
COACH
DAVE'S WORLD
WINGS
CYBILL
NBC NIGHTLY NEWS
60 MINUTES
ABC WRLD NEWS TO
BOY MEETS WORLD
NBC NIGHTLY NEWS
HANGIN' WITH MR.
NANNY
LOIS & CLARK
AMER.FUNNIEST-HO
STEP BY STEP
AMER.FUNNIEST-HM
Ready

Making predictions



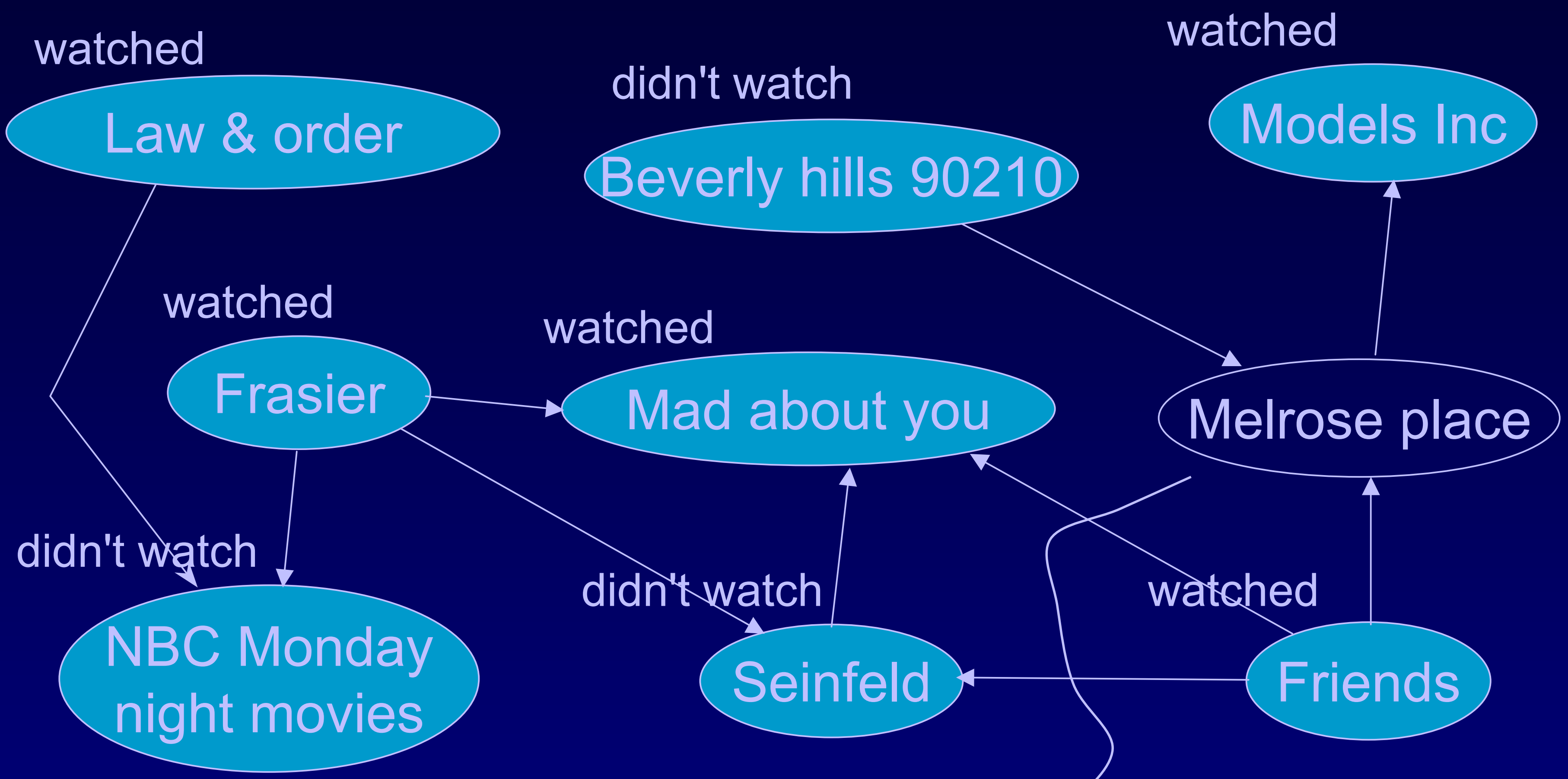
infer: p (watched 90210 | everything else we know about the user)

Making predictions



infer: p (watched 90210 | everything else we know about the user)

Making predictions



infer p (watched Melrose place | everything else we know about the user)

Recommendation list

- p=.67 Seinfeld
- p=.51 NBC Monday night movies
- p=.17 Beverly hills 90210
- p=.06 Melrose place

M

Software Packages

- BUGS: http://www.mrc-bsu.cam.ac.uk/bugs
  parameter learning, hierarchical models, MCMC
- Hugin: http://www.hugin.dk
  Inference and model construction
- xBaies: http://www.city.ac.uk/~rgc
  chain graphs, discrete only
- Bayesian Knowledge Discoverer: http://kmi.open.ac.uk/projects/bkd
  commercial
- MIM: http://inet.uni-c.dk/~edwards/miminfo.html
- BAYDA: http://www.cs.Helsinki.FI/research/cosco
  classification
- BN Power Constructor: BN PowerConstructor
- Microsoft Research: WinMine
  http://research.microsoft.com/~dmax/WinMine/Tooldoc.htm

For more information…

__Tutorials:__

K. Murphy (2001)
http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html

W. Buntine. Operations for learning with graphical models. Journal of Artificial Intelligence Research, 2, 159-225 (1994).

D. Heckerman (1999). A tutorial on learning with Bayesian networks. In Learning in Graphical Models (Ed. M. Jordan). MIT Press.

__Books:__

R. Cowell, A. P. Dawid, S. Lauritzen, and D. Spiegelhalter. Probabilistic Networks and Expert Systems. Springer-Verlag. 1999.

M. I. Jordan (ed, 1988). Learning in Graphical Models. MIT Press.

S. Lauritzen (1996). Graphical Models. Claredon Press.

J. Pearl (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

P. Spirtes, C. Glymour, and R. Scheines (2001). Causation, Prediction, and Search, Second Edition. MIT Press.