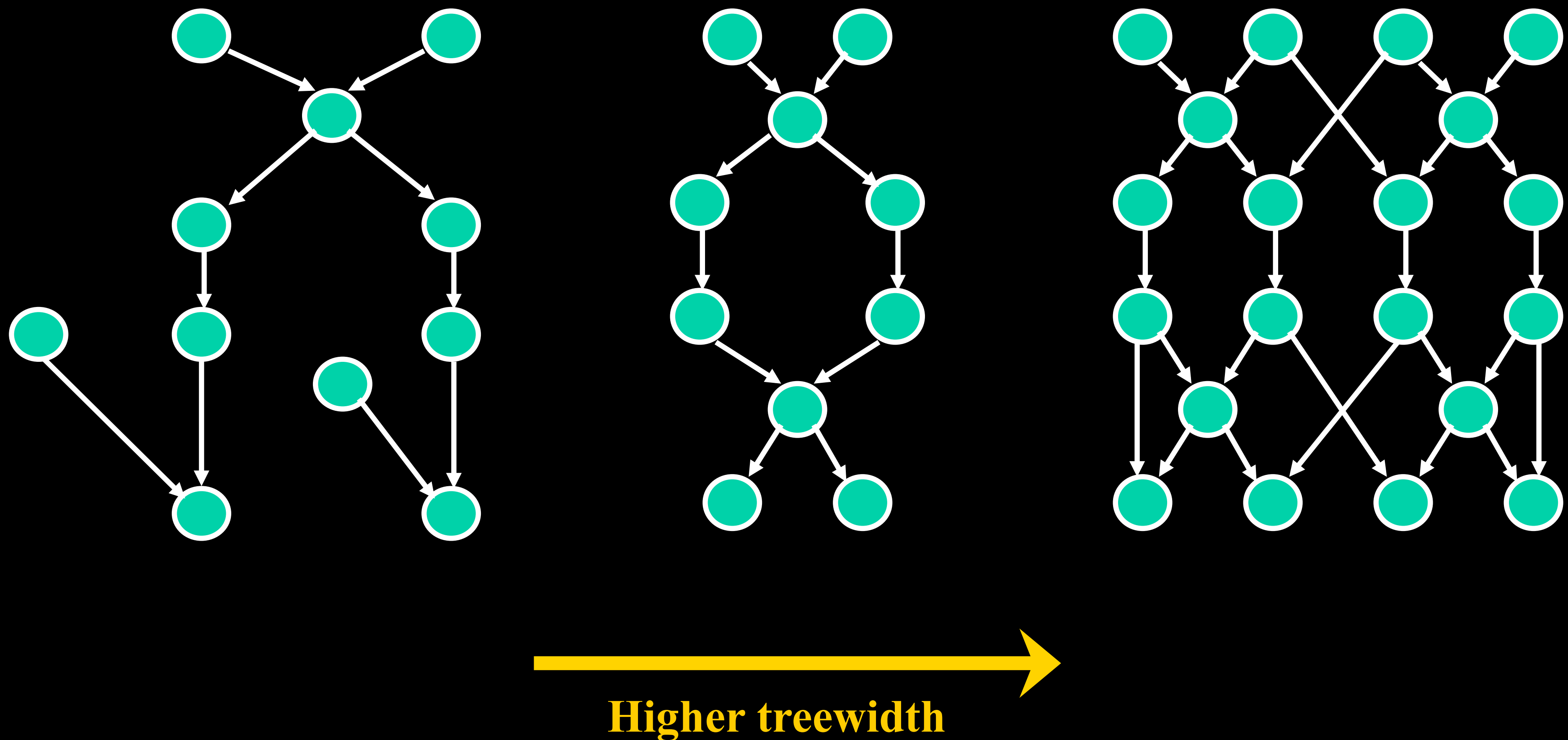
Inference in Bayesian Networks

A. Darwiche

Query Types

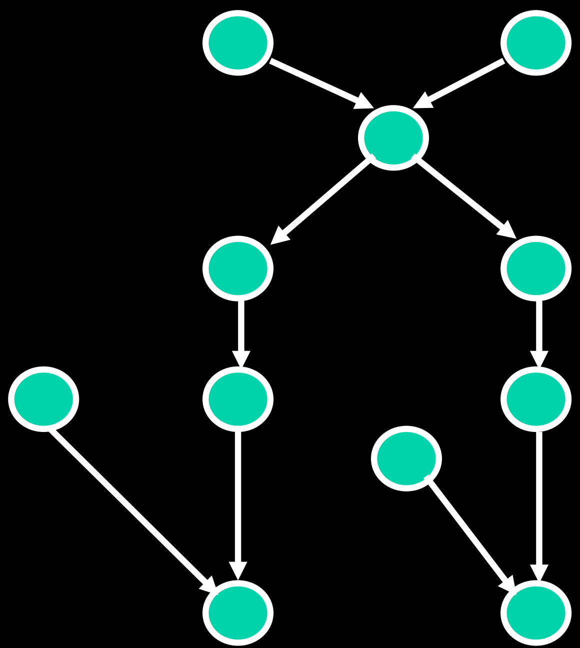
- **Pr:**
  - Evidence: Pr(e)
  - Posterior marginals: Pr(x|e) for every X
- **MPE:** Most probable explanation:
  - Instantiation y such that Pr(y|e) is maximal ($Y = \overline{E}$)
- **MAP:** Maximum a posteriori hypothesis:
  - Intantiation y such that Pr(y|e) is maximal (Y is subset of $\overline{E}$)

TreeWidth
(Measures connectivity of Networks)

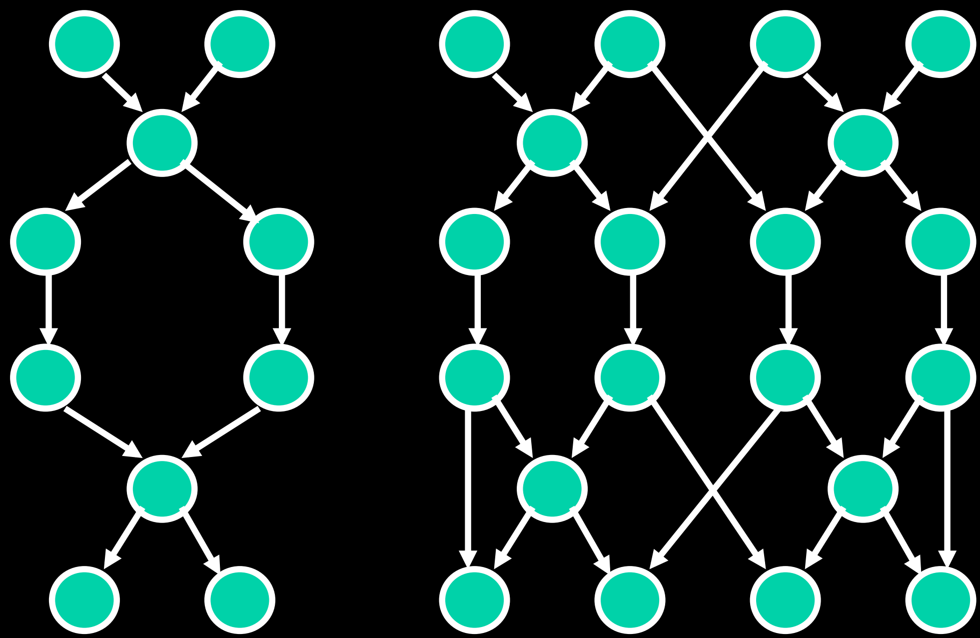


A. Darwiche

TreeWidth
(Measures connectivity of Networks)



Singly-connected network (polytree)



Multiply-connected networks

Treewidth

- The treewidth of a polytree is m, where m is the maximum number of parents per node
- If each node has at most one parent, the polytree is called a tree
- The treewidth of a tree is 1

Treewidth

- Given a Bayesian network **N** with:
  - Number of nodes: n
  - Treewith: w
- We can compute posterior marginals in:
  - $O(n\ 2^w)$ time
  - $O(n\ 2^w)$ space
- It is easy to do inference on polytrees (assuming number of parents is small enough)

Learning

Machine Learning

- Using experiences and observations to improve future performance (actions):
  - What aspect of performance to be improved?
    Irrelevant aspects of the world, how the world evolves, what are desirable/undesirable situations.
  - What feedback is available?
    Supervised, Unsupervised, Reinforcement learning.
  - How to represent the output of a learning process?
    Logical knowledge, Probabilistic knowledge, Neural networks.
- Supervised: Give observations and actions they should lead to
- Unsupervised: Give observations only (find patterns)
- Reinforcement: Give positive/negative feedback on actions

Learning in Bayesian Networks

The Learning Problem

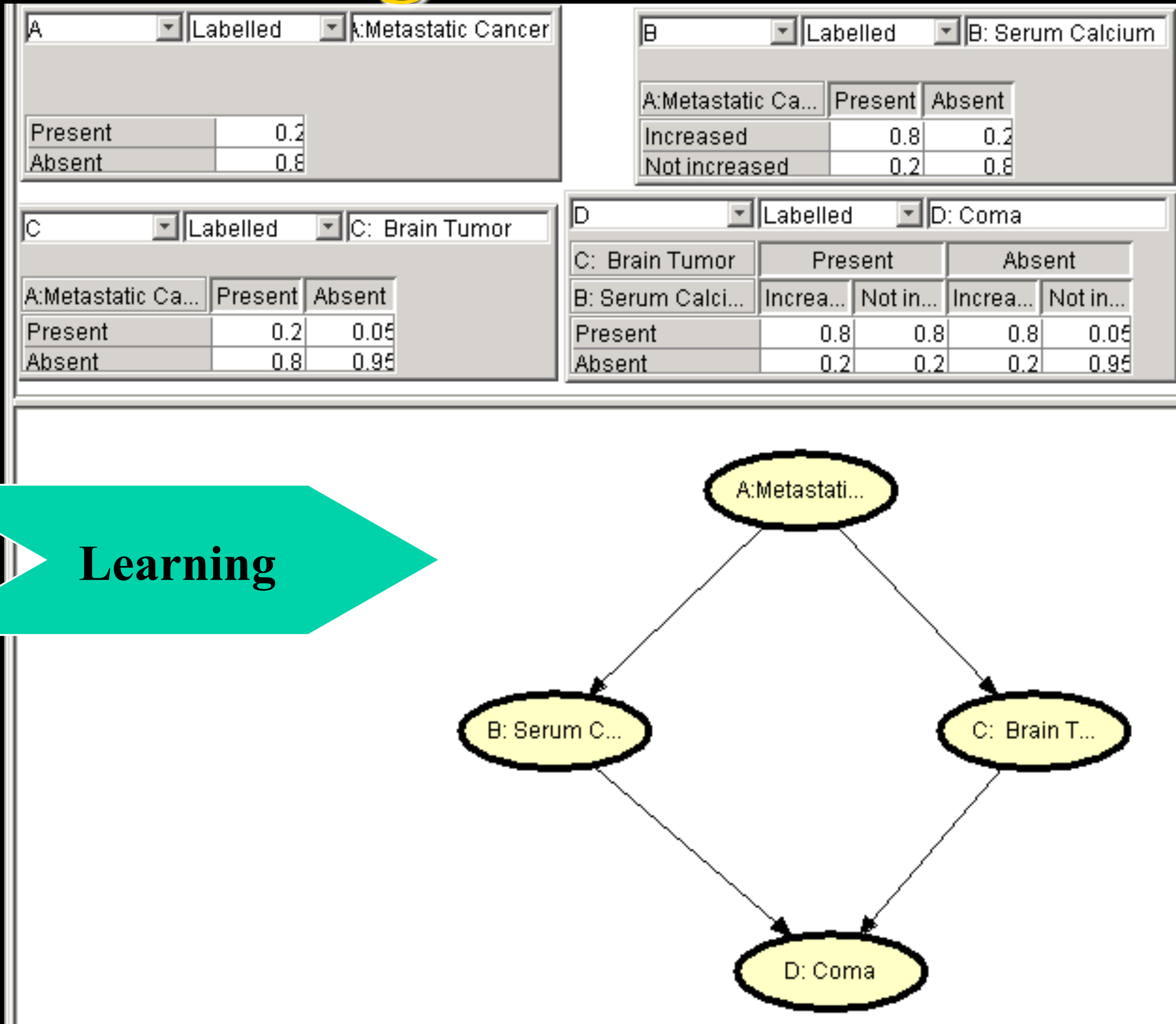
Known Structure Complete Data

Known Structure Incomplete Data

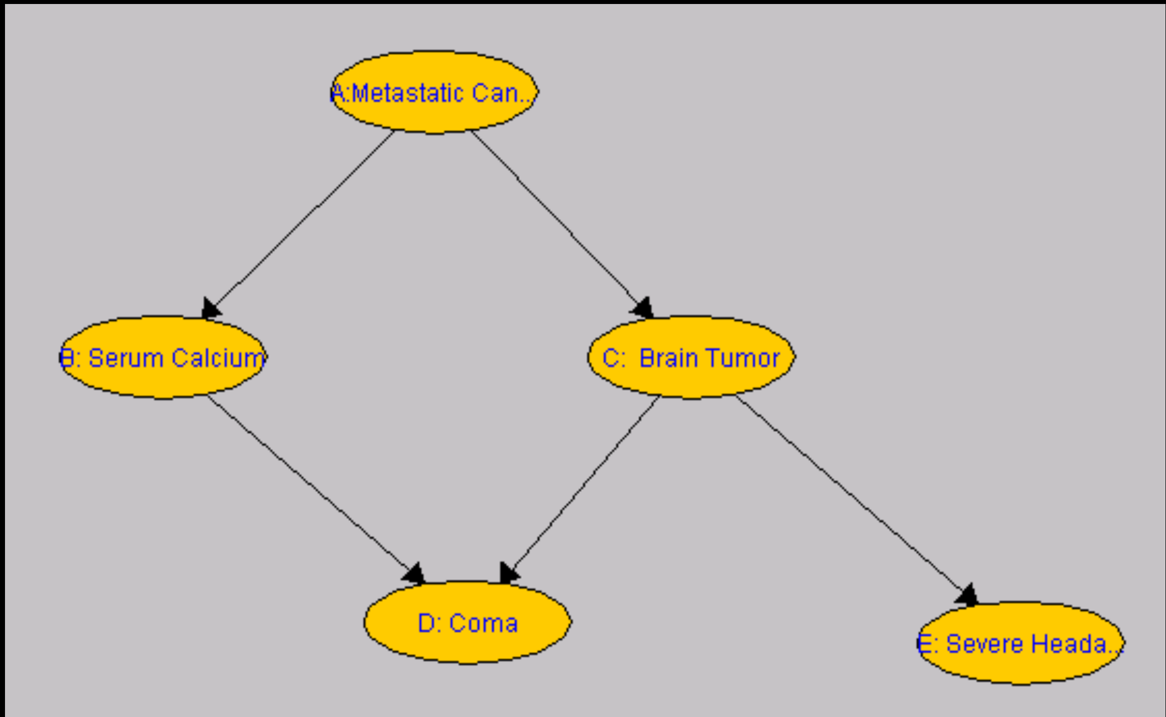
Unknown Structure Complete Data

Unknown Structure Incomplete Data

Learning

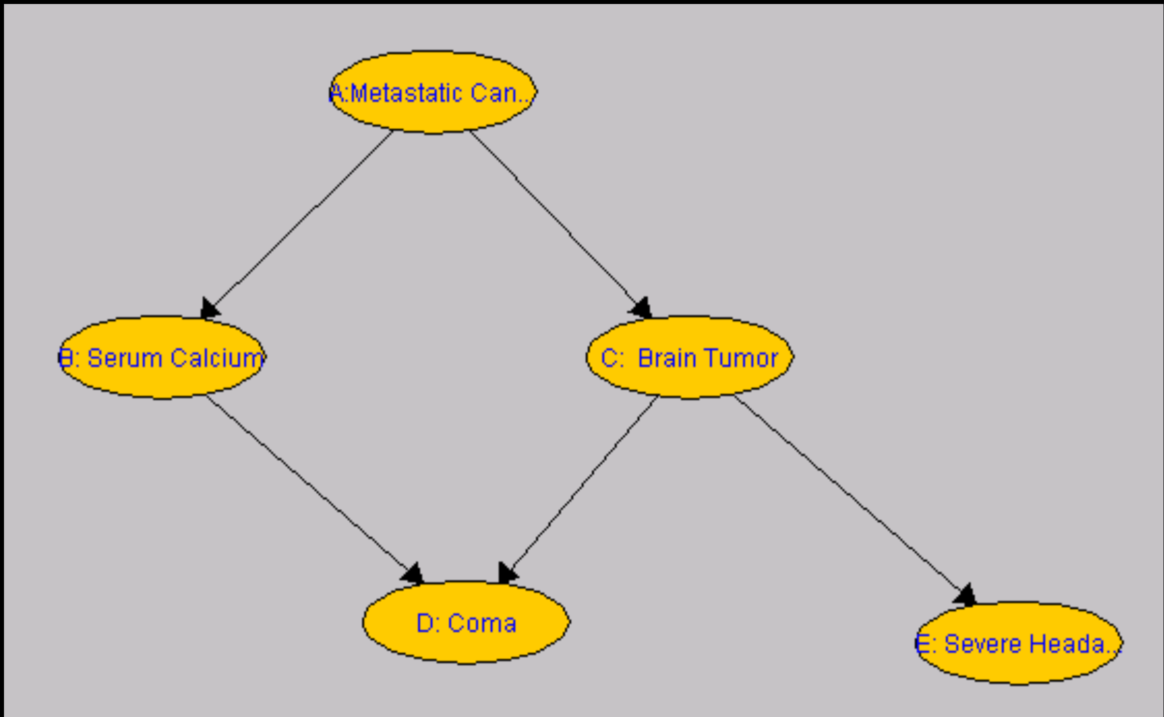


Known Structure Complete Data



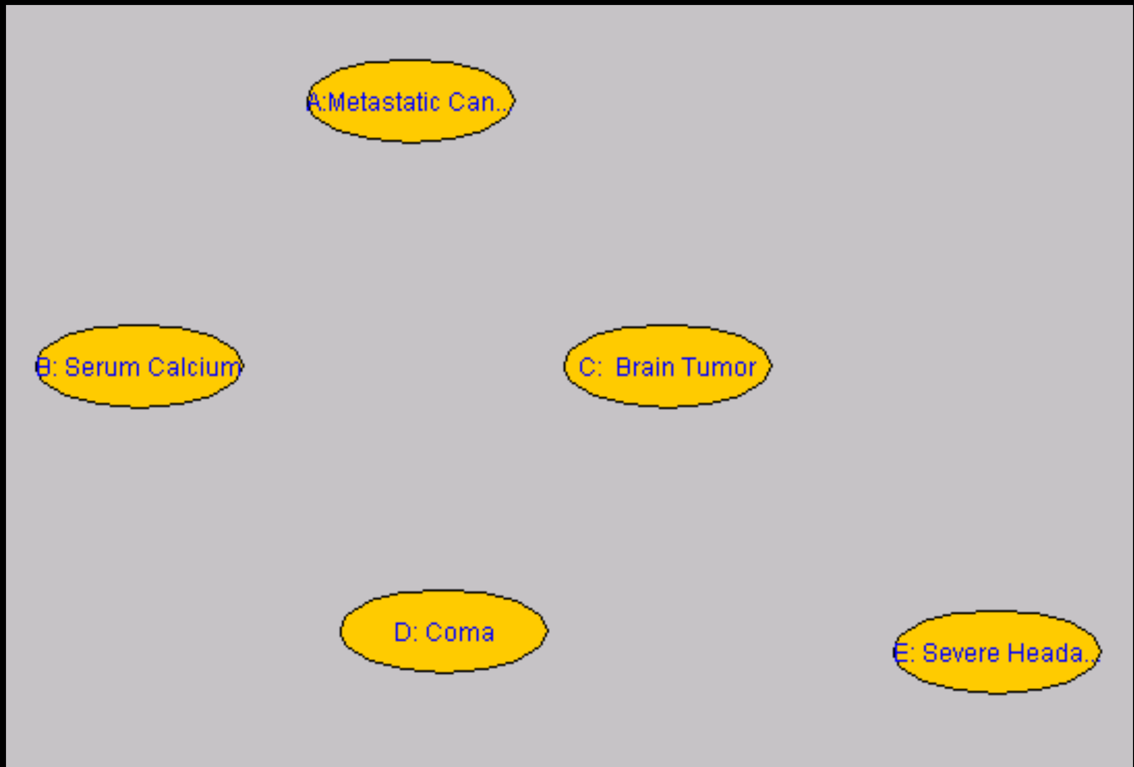

```
A,B,D,C,E
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Absent
Present,Not increased,Absent,Present,Absent
Absent,Increased,Absent,Absent,Absent
```

Known Structure   Incomplete Data



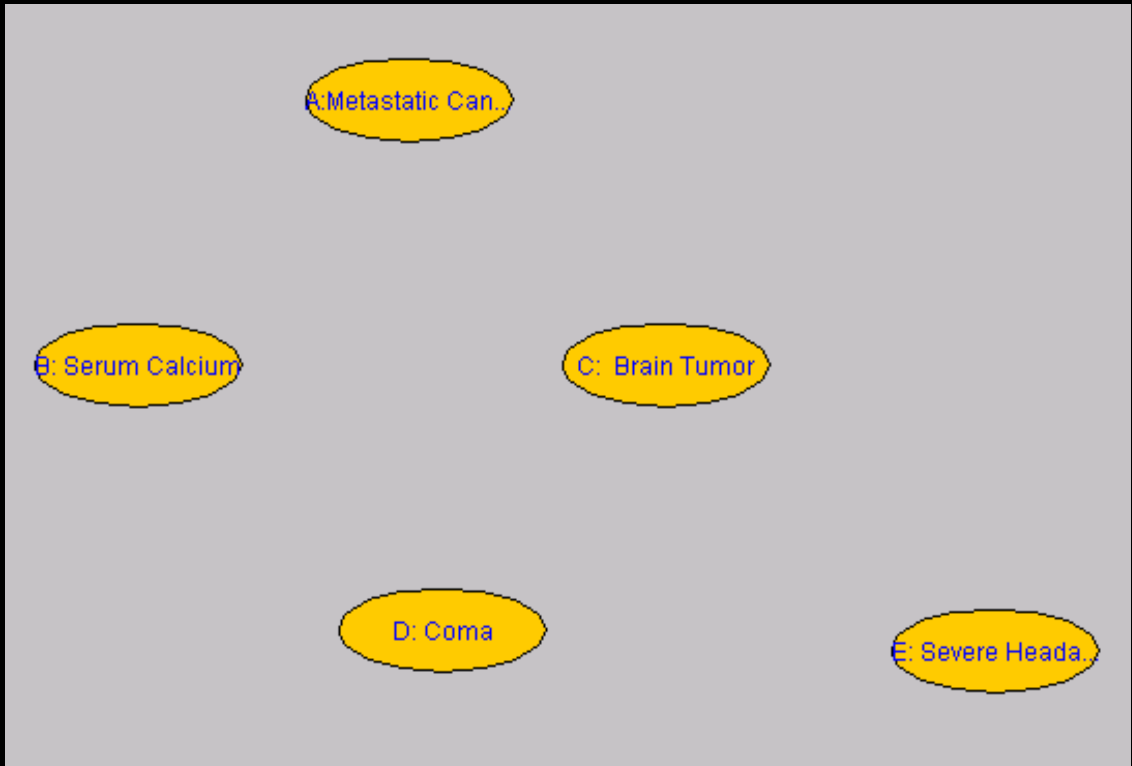

```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not  increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not  increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not  increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not  increased,Absent,Absent,Present
Absent,Not  increased,N/A,Absent,Absent
Absent,Not  increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not  increased,Absent,Absent,Absent
Absent,Not  increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not  increased,Absent,Absent,Present
Absent,Not  increased,Absent,Absent,Present
Absent,Not  increased,Absent,Absent,Absent
Absent,Not  increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not  increased,Absent,Absent,Present
Absent,Not  increased,Absent,Absent,Present
Absent,N/A,Absent,Absent,Present
N/A,Not  increased,Present,N/A,Present
Absent,N/A,Present,Absent,Present
Absent,Not  increased,Absent,Absent,Present
```

Unknown Structure   Complete Data



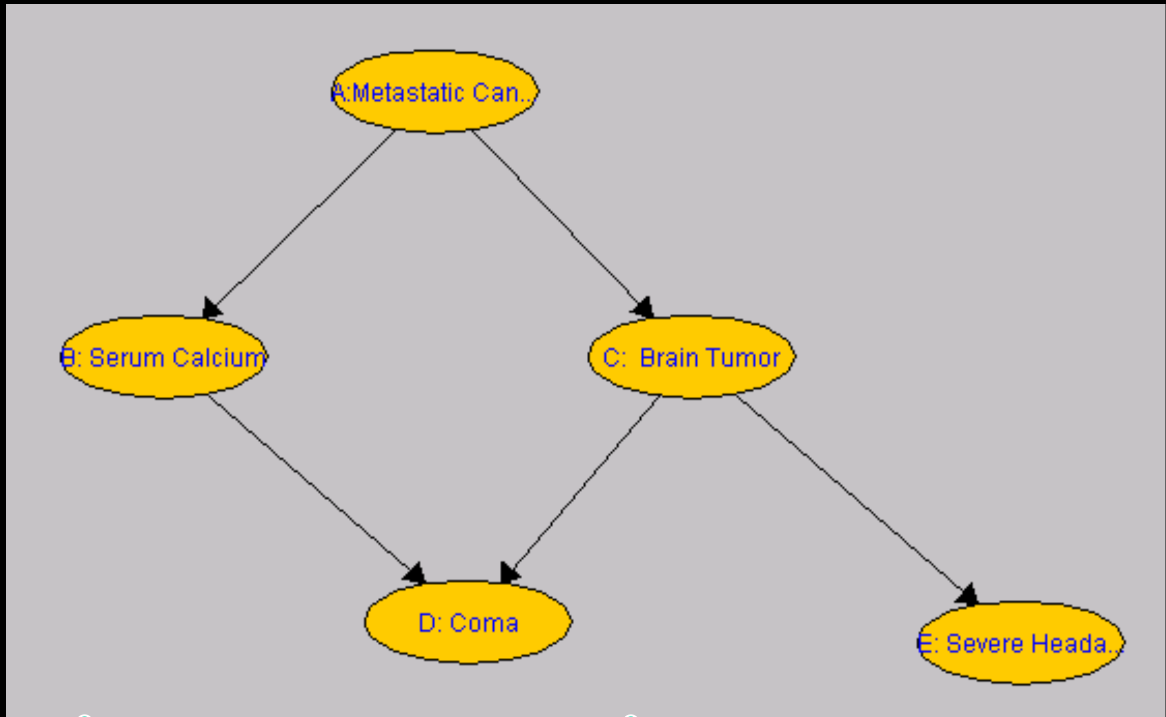

```
A,B,D,C,E
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

Unknown Structure Incomplete Data




```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,N/A,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

Known Structure




```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,N/A,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Absent,Absent,Present
N/A,Not increased,Present,N/A,Present
Absent,N/A,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

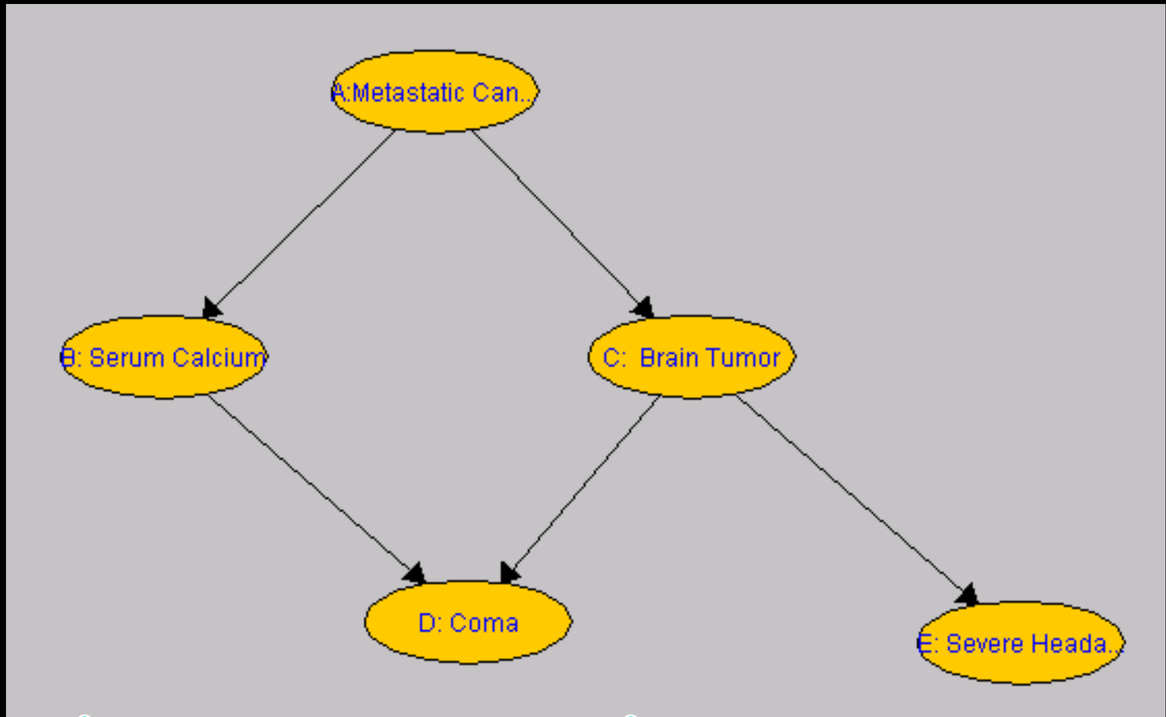
CPTs A

Method A

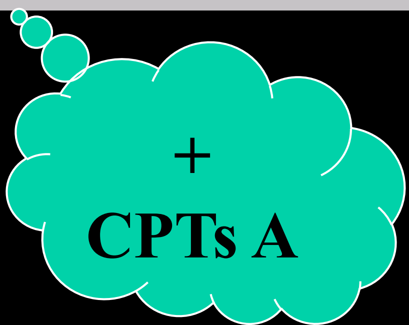
CPTs B

Method B

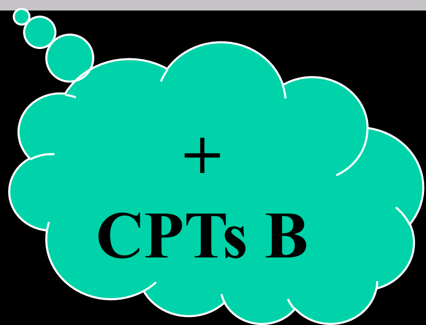
Known Structure




```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,N/A,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Absent,Absent,Present
N/A,Not increased,Present,N/A,Present
Absent,N/A,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```



$= \mathrm{Pr}_A$



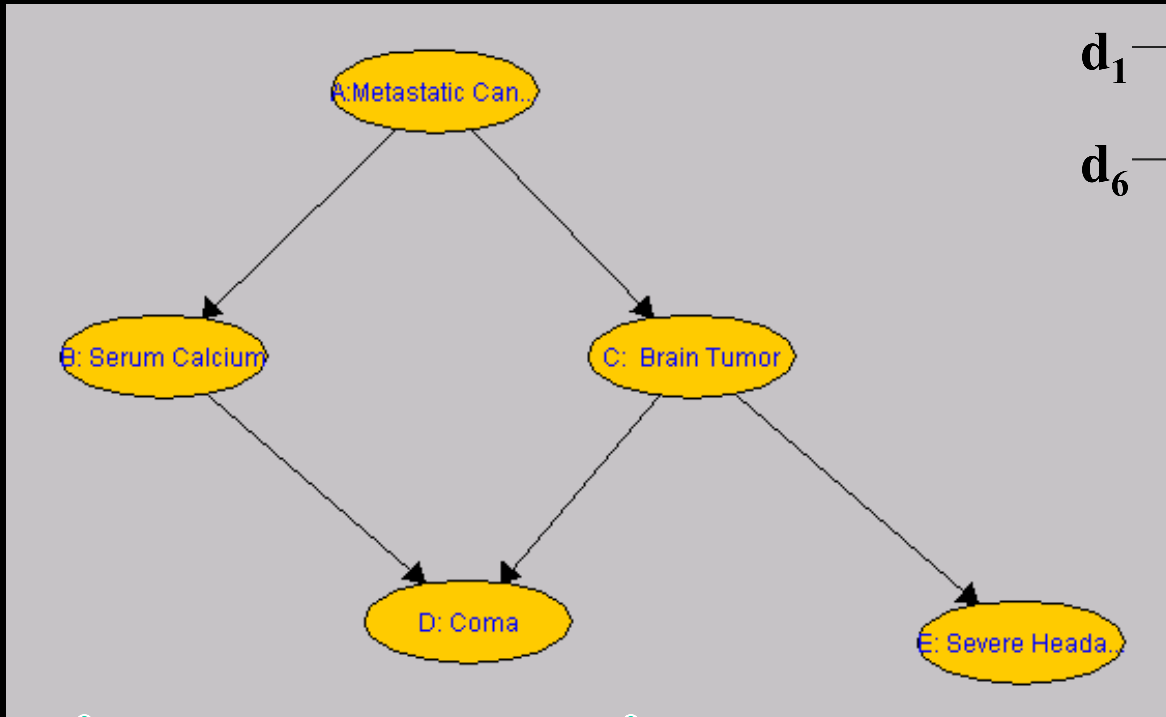
$= \mathrm{Pr}_B$

Which probability distribution should we choose?

<u>Common criterion:</u> Choose distribution that maximizes probability of data

Known Structure

Data D



d_1

d_6

```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,N/A,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Absent,Absent,Present
N/A,Not increased,Present,N/A,Present
Absent,N/A,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

\+ CPTs A

$= Pr_A$

\+ CPTs B

$= Pr_B$

Probability of data given Pr_A

$$Pr_A(D) = Pr_A(d_1) \ldots Pr_A(d_m)$$

Probability of data given Pr_B

$$Pr_B(D) = Pr_B(d_1) \ldots Pr_B(d_m)$$

Maximizing Probability of Data

- Complete Data: Unique set of CPTs which maximize probability of data

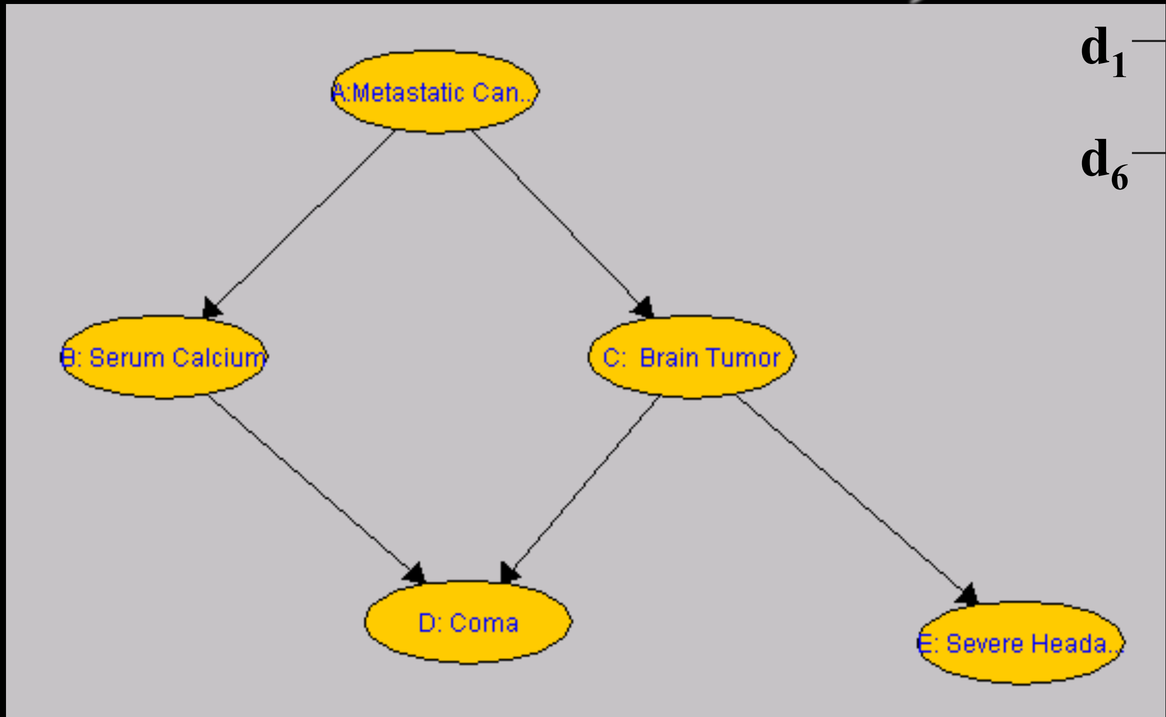
- Incomplete Data: No Unique set of CPTs which maximize probability of data

Maximizing Probability of Data

- Complete Data: <u>Unique</u> set of CPTs which maximize probability of data
- Incomplete Data: <u>No Unique</u> set of CPTs which maximize probability of data

Known Structure, Complete Data

Data D



d$_1$

d$_6$

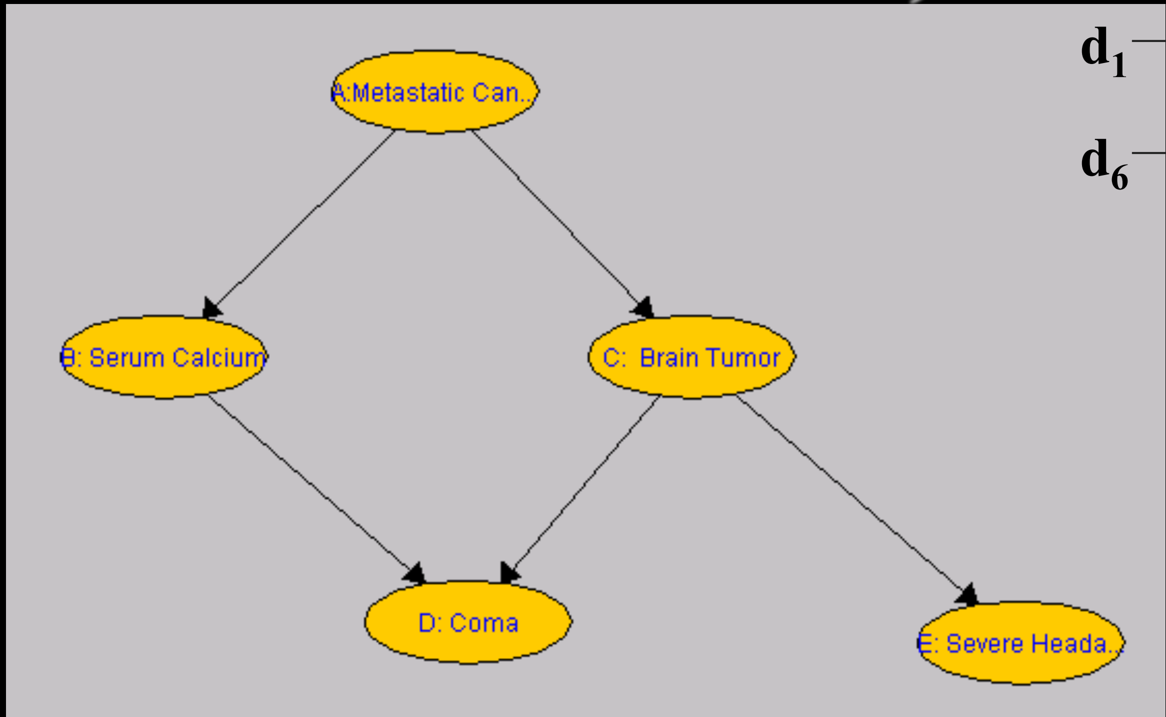
```
A,B,D,C,E
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Absent
Present,Not increased,Absent,Present,Absent
Absent,Increased,Absent,Absent,Absent
```

Estimated parameter:

$$\theta_{d|bc} = \frac{count(dbc, D)}{count(bc, D)} = \frac{\text{Number of data points } d_i \text{ with d b c}}{\text{Number of data points } d_i \text{ with b c}}$$

Data D

Known Structure, Complete Data



d_1, d_6

```
A,B,D,C,E
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Absent
Present,Not increased,Absent,Present,Absent
Absent,Increased,Absent,Absent,Absent
```

Estimated parameter:

$$\theta_{d|bc} = \frac{count(dbc, D)}{count(bc, D)} = \frac{\sum_{j=1}^{m} I(dbc; d_j)}{\sum_{j=1}^{m} I(bc; d_j)}$$

Complexity

- Network with:
  - Nodes: n
  - Parameters: k
  - Data points: m

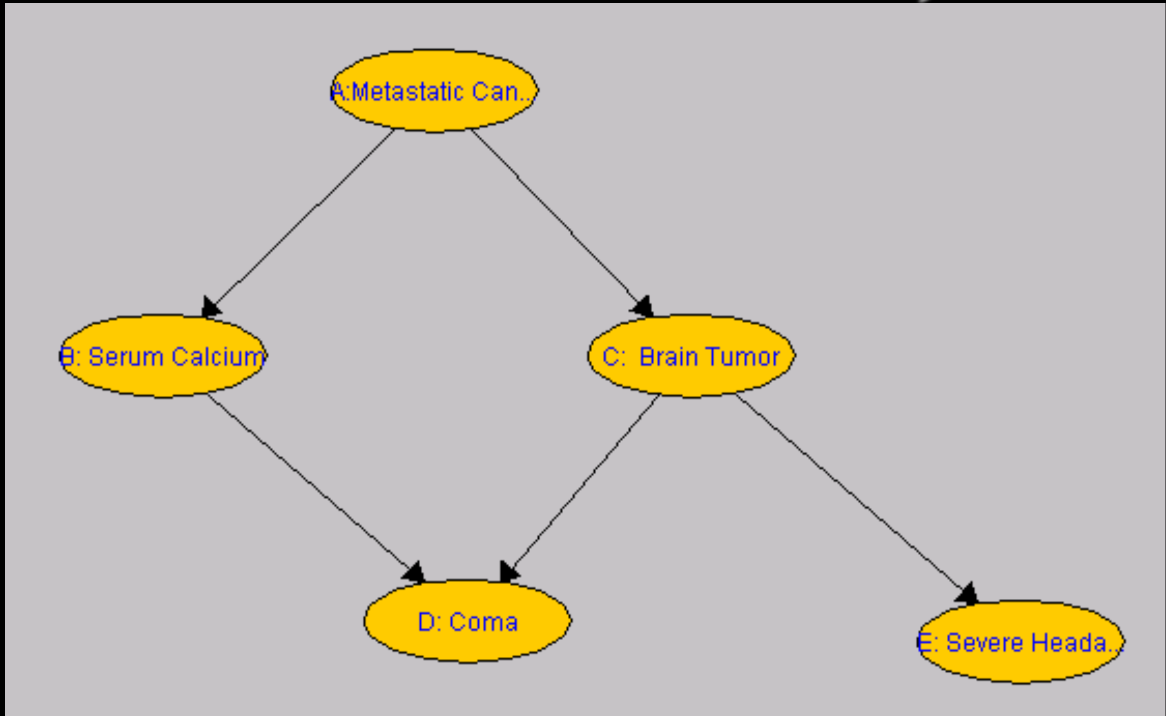
- Time complexity: O(m n k)
  (straightforward implementation)

- Space complexity: O(m n)
  space for data

Complexity

- Network with:
  - Nodes: n
  - Parameters: k
  - Data points: m
- Time complexity: O(m n k) (straightforward implementation)
- Space complexity: O(m n) space for data

Known Structure, Incomplete Data

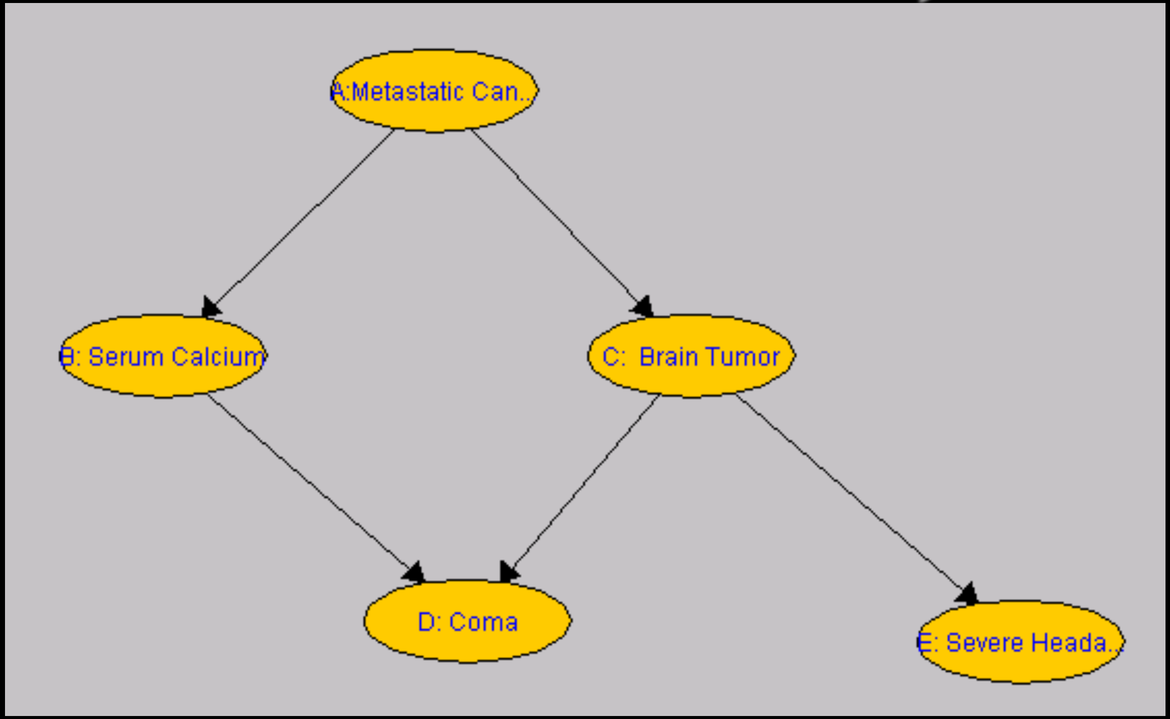



```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,N/A,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Absent,Absent,Present
N/A,Not increased,Present,N/A,Present
Absent,N/A,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

Pr_0 corresponds to the initial Bayesian network (random CPTs)

$$\theta_{d|bc} = \frac{\sum_{j=1}^{m} Pr_i(dbc \mid d_j)}{\sum_{j=1}^{m} Pr_i(bc \mid d_j)}$$

Known Structure, Incomplete Data




```
A,B,D,C,E
Present,Increased,N/A,Absent,Absent
Present,Increased,Present,Present,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Increased,Present,Absent,Absent
Absent,N/A,Absent,N/A,Present
Absent,Not increased,N/A,Absent,Absent
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,N/A
Present,Increased,N/A,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,N/A,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Present,Present
Absent,N/A,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Present,Absent,Absent
Present,Increased,Present,Absent,N/A
Present,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,N/A,Absent,Absent,Present
N/A,Not increased,Present,N/A,Present
Absent,N/A,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

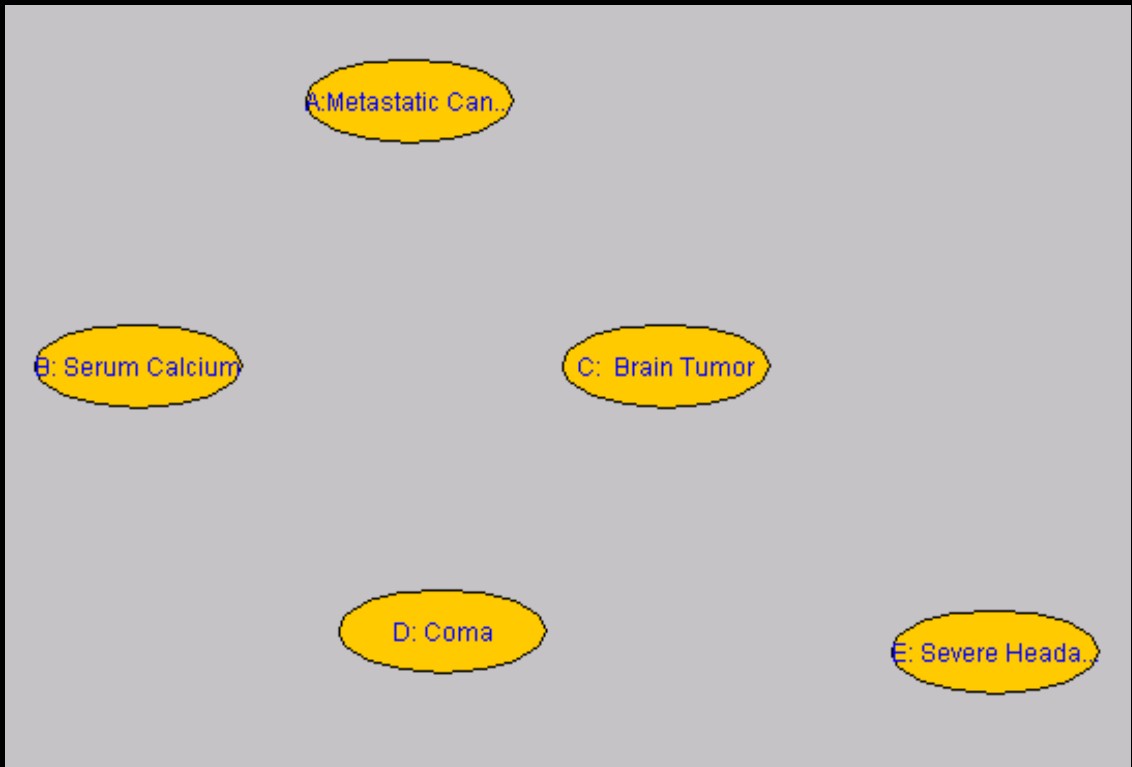
EM Algorithm (Expectation-Maximization):

- Initial CPTs to random values
- Repeat until convergence:
  - Estimate parameters using current CPTs (E-step)
  - Update CPTs using estimates (M-step)

EM Algorithm

- Probability of data cannot get smaller after an iteration
- Algorithm is not guaranteed to return the network which absolutely maximizes probability of data
- It is guaranteed to return a local maxima: Random re-starts
- Algorithm is stopped when
  - change in likelihood gets very small, or
  - Change in parameters gets very small

Unknown Structure   Complete Data




```
A,B,D,C,E
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Absent,Not increased,Absent,Absent,Present
Absent,Not increased,Absent,Absent,Present
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Absent
Present,Increased,Present,Absent,Present
Absent,Not increased,Absent,Absent,Present
```

Minimum Description Length (MDL)

- Includes two components:
  - Maximize probability of data, Pr(D)
  - Minimize dimension of model: number of independent parameters

Heuristic Search

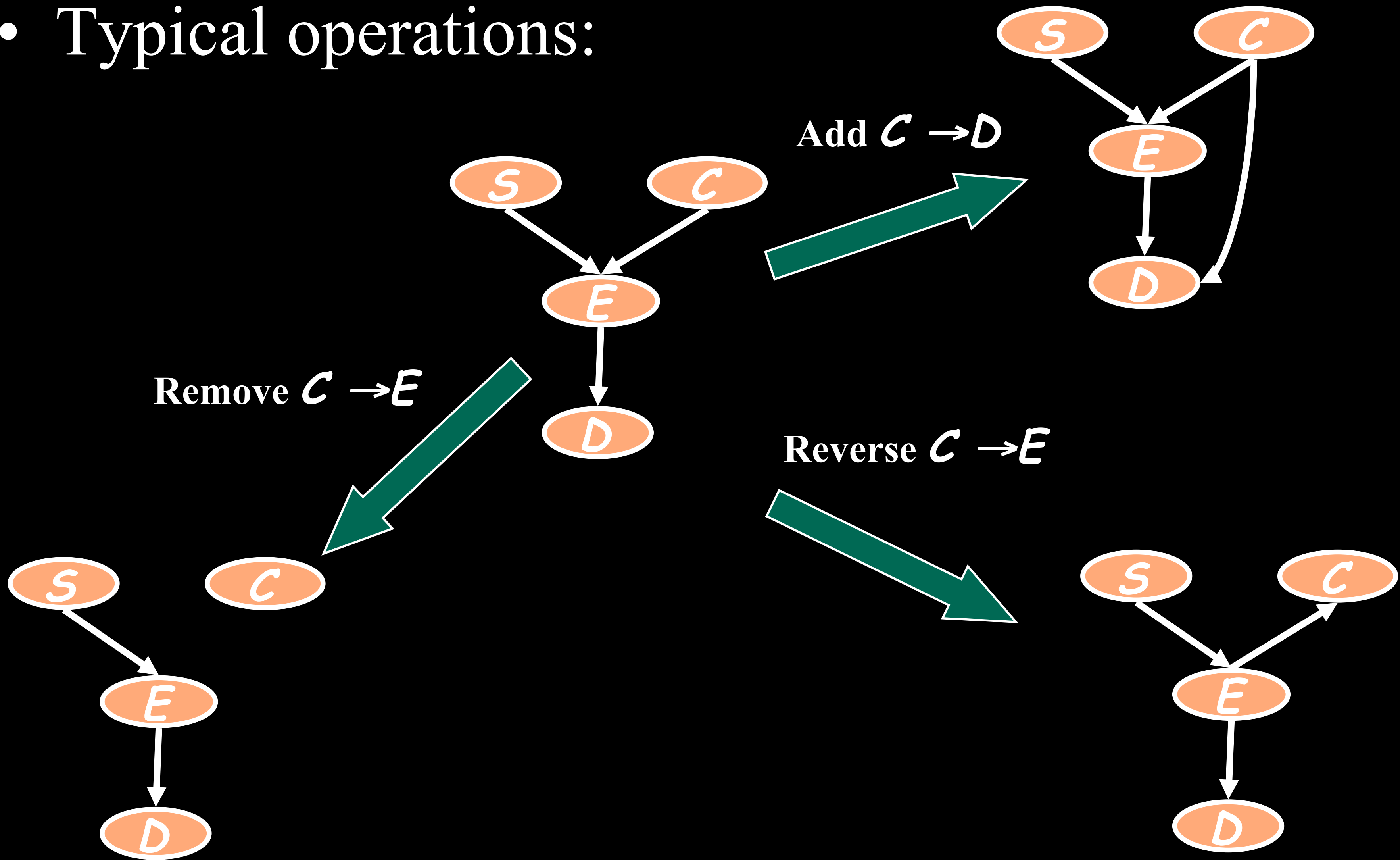
- We address the problem by using heuristic search
- Define a search space:
  - nodes are possible structures
  - edges denote adjacency of structures
- Traverse this space looking for high-scoring structures

Search techniques:

- Greedy hill-climbing
- Best first search
- Simulated Annealing
- ...

Heuristic Search (cont.)

- Typical operations:



Greedy Hill-Climbing

Simplest heuristic local search

- Start with a given network
  - empty network
  - best tree
  - a random network
- At each iteration
  - Evaluate all possible changes
  - Apply change that leads to best improvement in score
  - Reiterate
- Stop when no modification improves score

- Each step requires evaluating approximately n new changes

Greedy Hill-Climbing (cont.)

- Greedy Hill-Climbing can get struck in:
  - **Local Maxima:**
    - All one-edge changes reduce the score
  - **Plateaus:**
    - Some one-edge changes leave the score unchanged

- Both are occur in the search space

Greedy Hill-Climbing (cont.)

To avoid these problems, we can use:

- **TABU-search**
  - Keep list of K most recently visited structures
  - Apply best move that does not lead to a structure in the list
  - This escapes plateaus and local maxima and with “basin” smaller than K structures

- **Random Restarts**
  - Once stuck, apply some fixed number of random edge changes and restart search
  - This can escape from the basin of one maxima to another

Other Local Search Heuristics

- **Stochastic First-Ascent Hill-Climbing**
  - Evaluate possible changes at random
  - Apply the first one that leads "uphill"
  - Stop when a fix amount of "unsuccessful" attempts to change the current candidate
- **Simulated Annealing**
  - Similar idea, but also apply "downhill" changes with a probability that is proportional to the change in score
  - Use a temperature to control amount of random downhill steps
  - Slowly "cool" temperature to reach a regime where performing strict uphill moves

Some Applications

- Biostatistics -- Medical Research Council (Bugs)
- Data Analysis -- NASA (AutoClass)
- Collaborative filtering -- Microsoft (MSBN)
- Fraud detection -- ATT
- Classification -- SRI (TAN-BLT)
- Speech recognition -- UC Berkeley

Collaborative Filtering

- Collaborative Filtering (CF) finds items of interest to a user based on the preferences of other similar users.
  - Assumes that human behavior is predictable

Where is it used?

- E-commerce
  - Recommend products based on previous purchases or click-stream behavior
  - Ex: Amazon.com

amazon.com

- Information sites
  - Rate items based on previous user ratings
  - Ex: MovieLens, Jester

movielens
helping you find the right movies

John	5	-	3	2
Sam	-	4	1	5
Cindy	3	-	5	-

Bob	5	1	-	-

CF

Bob	5	1	3.5	1.7

Memory-based Algorithms

- Use the entire database of user ratings to make predictions.
  - Find users with similar voting histories to the active user.
  - Use these users' votes to predict ratings for products not voted on by the active user.

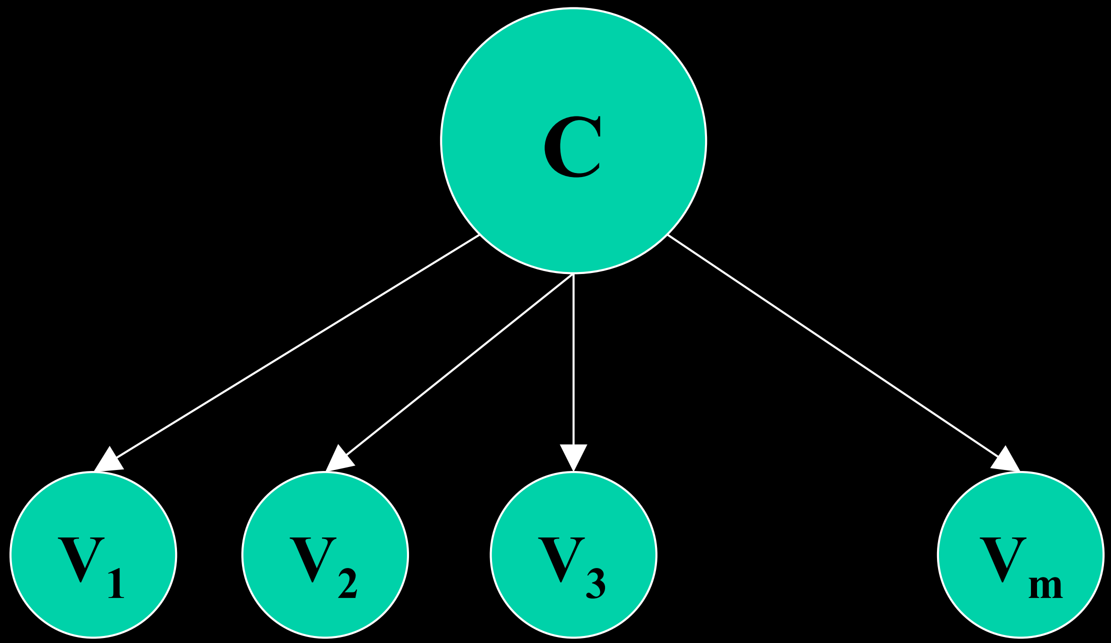
Model-based Algorithms

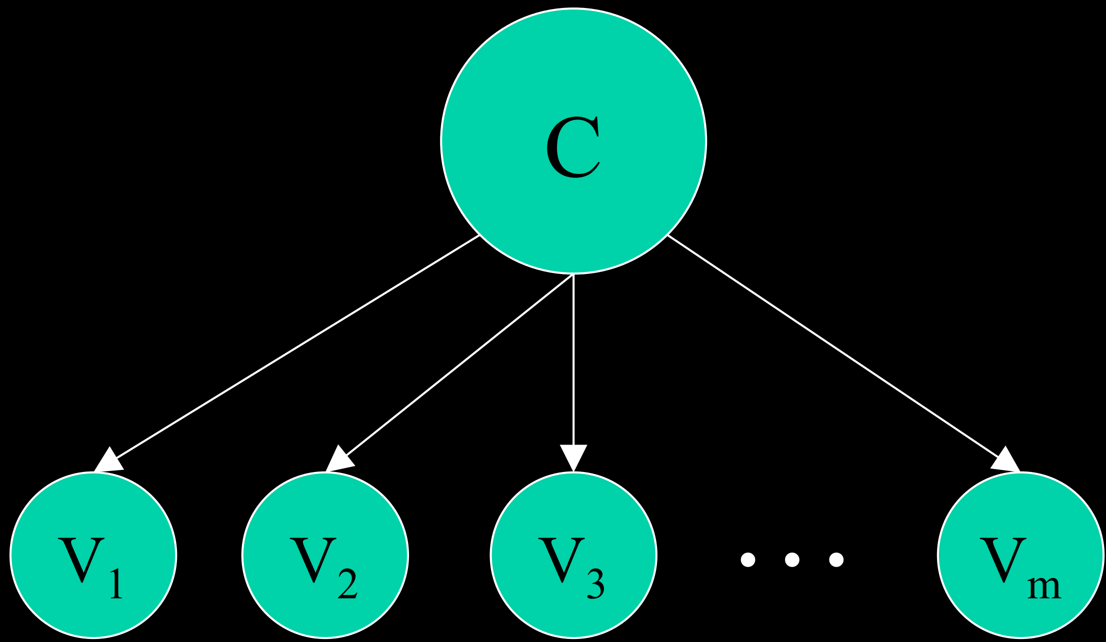
- Construct a model from the vote database.
- Use the model to predict the active user’s ratings.

Bayesian Clustering

- Use a Naïve Bayes network to model the vote database.
- m vote variables: one for each title.
  - Represent discrete vote values.
- 1 “cluster” variable
  - Represents user personalities

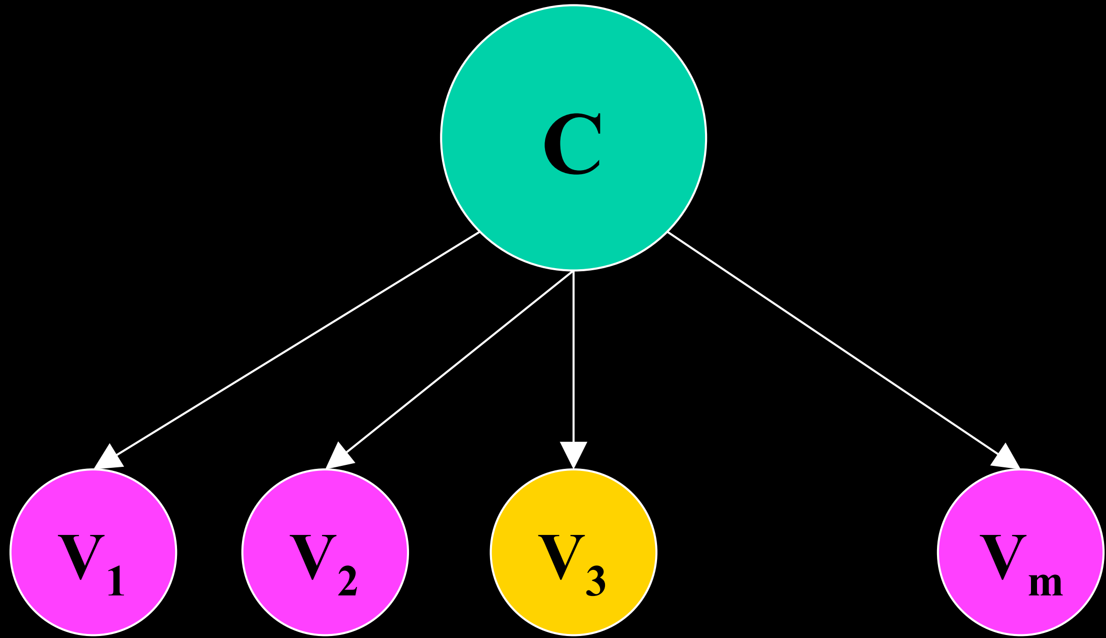
Naïve Bayes





C	$\Pr(c)$
c_1	.5
c_2	.1
c_3	.35
c_4	.05

v_k	C	$\Pr(v_k \mid c)$
1	c_1	.3
2	c_1	.25
$\vdots$	$\vdots$	$\vdots$
5	c_4	.6



- Inference
  - Evidence: known votes v_k for titles $k \in I$
  - Query: title j for which we need to predict vote
- Expected value of vote:

$$p_j = \sum_{h=1}^{w} h \Pr(v_j = h \mid v_k : k \in I)$$

Learning

- Simplified Expectation Maximization (EM) Algorithm with partial data

$$\theta_c \approx \Pr(c) \qquad \theta_{v_k|c} \approx \Pr(v_k \mid c)$$

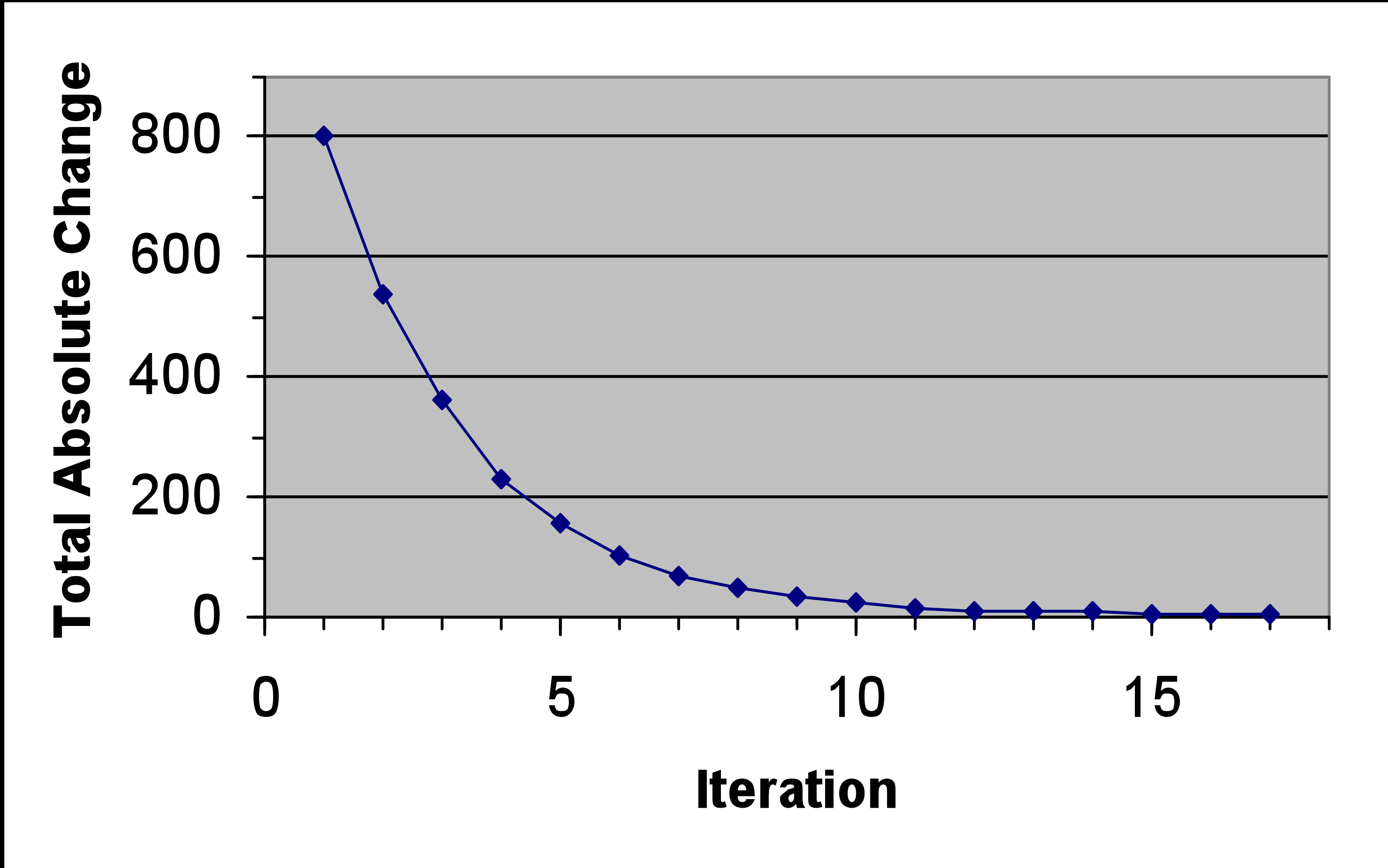
- Initialize CPTs with random values subject to the following constraints:

$$\sum_C \theta_c = 1 \qquad \sum_{v_k} \theta_{v_k|c} = 1$$

Datasets

- MovieLens
  - 943 users; 1682 titles; 100,000 votes (1..5); explicit voting
- MS Web – website visits
  - 610 users; 294 titles; 8,275 votes (0,1) : null votes => 0 : 179,340 votes; implicit voting

- Learning curve for MovieLens Dataset



Protocols

- User database is divided into: 80% training set and 20% test set.
  - One-by-one select a user from the test set to be the active user.
  - Predict some of their votes based on remaining votes

- All-But-One

I_a	e	e	Q	e	e	e	e	e	e	e	e	e	e	e

- Given-{Two, Five, Ten}

I_a	Q	Q	Q	Q	e	Q	Q	Q	Q	Q	Q	e	Q	Q

I_a	Q	e	Q	e	Q	Q	e	Q	Q	Q	e	Q	Q	e

I_a	e	Q	Q	e	Q	e	e	e	e	Q	e	e	e	e

Evaluation Metric

- Average Absolute Deviation

- Ranked Scoring

Results

- Experiments were run 5 times and averaged
- Movielens

Algorithm	Given-Two	Given-Five	Given-Ten	All-But-One
Correlation	1.019	.916	.865	.806
VecSim	.948	.878	.843	.799
BC(9)	.771	.765	.763	.753

- **MS Web**

Algorithm	Given-Two	Given-Five	Given-Ten	All-But-One
Correlation	0.105	0.0911	0.0844	0.0673
VecSim	0.101	0.0885	0.0818	0.0675
BC(9)	0.0652	0.0652	0.0649	0.0507