Non-Standard-Datenbanken

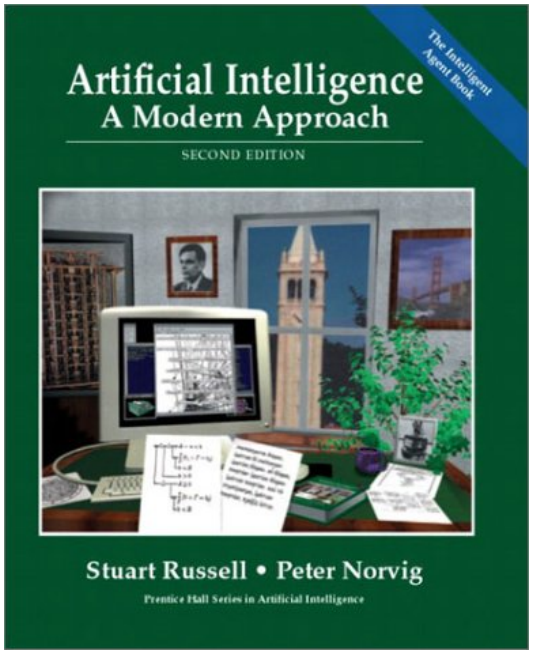
Bayesian Networks, Inference, and Learning

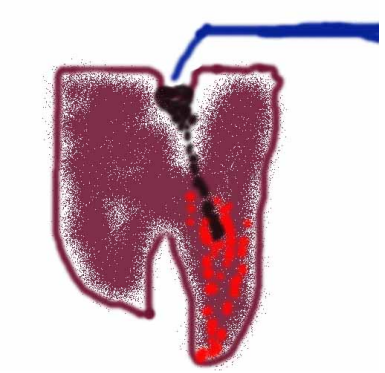
Prof. Dr. Ralf Möller
Universität zu Lübeck
Institut für Informationssysteme

Beispiel

Zahnarzt-Problem mit vier Variablen:

- Toothache (Sind besagte Schmerzen wirklich Zahnschmerzen?)
- Cavity (Es könnte ein Loch sein?)
- Catch (Stahlinstrument erzeugt Testschmerz?)
- Weather (Wetter: sunny,rainy,cloudy,snow)





Nachfolgende Präsentationen enthalten Material aus Kapitel 14 (Sektion 1 and 2)

Prior probability

- Prior or unconditional probabilities of propositions

  e.g., P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72 correspond to belief prior to arrival of any (new) evidence

- Probability distribution

  gives values for all possible assignments:

  P(*Weather*) = <0.72, 0.1, 0.08, 0.1>

  (normalized, i.e., sums to 1 because one condition must be the case)

Full joint probability distribution

- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables

  P(Weather,Cavity) is a 4 × 2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

- Full joint probability distribution: all random variables involved
  - **P**(Toothache, Catch, Cavity, Weather)

- Every query about a domain can be answered by the full joint distribution

Discrete random variables: Notation

- Dom(Weather) = {sunny, rainy, cloudy, snow} and Dom(Weather) disjoint from domain of other random variables:
  - Atomic event Weather=rainy often written as rainy
  - Example: P(rainy), the random variable Weather is implicitly defined by the value rainy
- Boolean variable Cavity
  - Atomic event Cavity=true written as cavity
  - Atomic event Cavity=false written as ¬cavity
  - Examples: P(cavity) or P(¬cavity)

Conditional probability

- Conditional or posterior probabilities
  e.g., P(*cavity* | *toothache*) = 0.8
  or: <0.8>
  i.e., given that *toothache* is all I know

- (Notation for conditional distributions:
  P(*Cavity* | *Toothache*) is a 2-element vector of 2-element vectors

- If we know more, e.g., *cavity* is also given, then we have
  P(*cavity* | *toothache, cavity*) = 1

- New evidence may be irrelevant, allowing simplification, e.g.,
  P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8

- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

- Definition of conditional probability (in terms of uncond. prob.):
  $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$

- Product rule gives an alternative formulation ($\wedge$ is commutative):
  $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

- A general version holds for whole distributions, e.g.,
  $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\, \mathbf{P}(Cavity)$

  View as a set of 4 × 2 equations, not matrix mult.
  (1,1) $P(Weather=sunny \mid Cavity=true)\, P(Cavity=true)$
  (1,2) $P(Weather=sunny \mid Cavity=false)\, P(Cavity=false)$, ….

- Chain rule is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})\end{aligned}$$

Inference by enumeration

- Start with the joint probability distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- For any proposition ϕ, sum the probability where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

- Unconditional or **marginal probability** of toothache
- Process is called marginalization or summing out

Ax_3: $P(A \cup B) = P(A) + P(B)$
for disjoint events $A, B \subseteq \Omega$

Marginalization and conditioning

- Let Y, Z be sequences of random variables s.th. $Y \cup Z$ denotes all random variables describing the world

- Marginalization
  - $\mathbf{P}(\mathbf{Y}) = \Sigma_{z \in Z} \mathbf{P}(\mathbf{Y}, \mathbf{z})$

- Conditioning
  - $\mathbf{P}(\mathbf{Y}) = \Sigma_{z \in Z} \mathbf{P}(\mathbf{Y} \mid \mathbf{z}) P(\mathbf{z})$

Inference by enumeration

- Start with the joint probability distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

- P(cavity $\vee$ *toothache*) = 0.108 + 0.012 + 0.072 + 0.008+ 0.016 + 0.064 = 0.28

(P(*cavity* $\vee$ *toothache*) = P(*cavity*) + P(*toothache*) – P(*cavity* $\wedge$ *toothache*))

Inference by enumeration

- Start with the joint probability distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \quad \text{Product rule}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.08/0.2 = 0.4$$

$$P(cavity \mid toothache) = (0.108+0.012)/0.2 = 0.6$$

Normalization

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Denominator **P**(**z**) (or P(toothache) in the example before) can be viewed as a normalization constant α

P(Cavity | toothache) = α **P**(Cavity,toothache)

= α [**P**(Cavity,toothache,catch) + **P**(Cavity,toothache,¬ catch)]

= α [<0.108,0.016> + <0.012,0.064>]

= α <0.12,0.08> = <0.6,0.4>

General idea: compute distribution on query variable by fixing evidence variables (Toothache) and summing over hidden variables (Catch)

Inference by enumeration, contd.

Typically, we are interested in
the posterior joint distribution of the query variables **Y**
given specific values **e** for the evidence variables **E**
(X are all variables of the modeled world)

Let the hidden variables be **H** = **X** - **Y** – **E** then the required summation of joint entries is done by summing out the hidden variables:

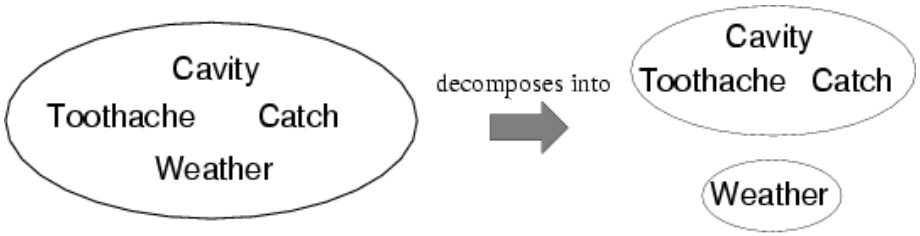
$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables (**X**)

- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity and n denotes the number of random variables
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Independence

- A and B are independent iff
  P(A|B) = **P**(A) or **P**(B|A) = **P**(B) or **P**(A, B) = **P**(A) **P**(B)



P(Toothache, Catch, Cavity, Weather)
= **P**(Toothache, Catch, Cavity) **P**(Weather)

- 32 entries reduced to 12;
- *Absolute* independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- **P**(Toothache, Cavity, Catch) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  (1) **P**(catch | toothache, cavity) = **P**(catch | cavity)

- The same independence holds if I haven't got a cavity:
  (2) **P**(catch | toothache,¬cavity) = **P**(catch | ¬cavity)

- Catch is conditionally independent of Toothache given Cavity:
  P(Catch | Toothache,Cavity) = **P**(Catch | Cavity)

- Equivalent statements:
  P(Toothache | Catch, Cavity) = **P**(Toothache | Cavity)

  P(Toothache, Catch | Cavity) = **P**(Toothache | Cavity) **P**(Catch | Cavity)

Conditional independence contd.

- Write out full joint distribution using chain rule:

  P(Toothache, Catch, Cavity)

  = **P**(Toothache | Catch, Cavity) **P**(Catch, Cavity)

  = **P**(Toothache | Catch, Cavity) **P**(Catch | Cavity) **P**(Cavity)

  conditional independence

  = **P**(Toothache | Cavity) **P**(Catch | Cavity) **P**(Cavity)

  i.e., 2 + 2 + 1 = 5 independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n
- Conditional independence is our most basic and robust form of knowledge about uncertain environments

Naïve Bayes Model

$$\begin{aligned}
&\mathbf{P}(Cavity|toothache \wedge catch) \\
&= \alpha\,\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\
&= \alpha\,\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

Usually, the assumption that effects are independent is wrong,
but works well in practice

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

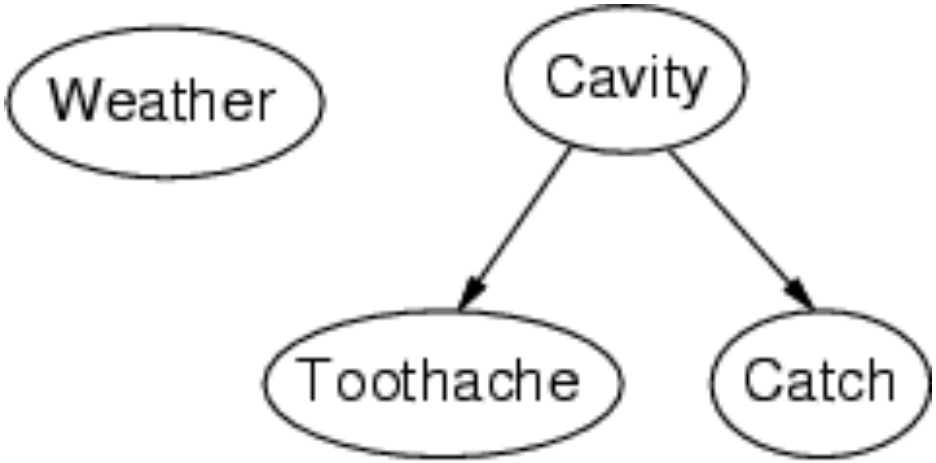
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

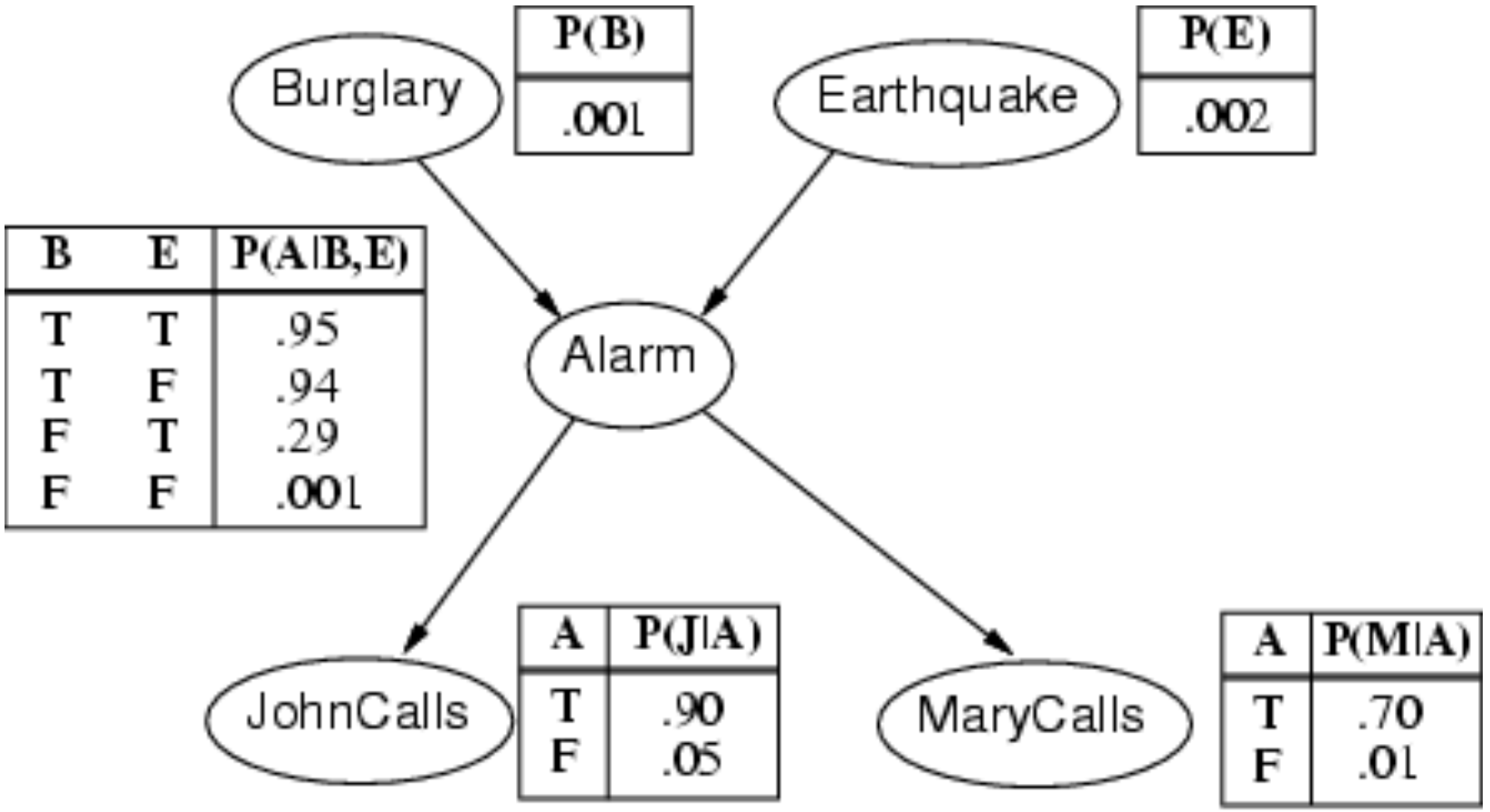
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglary?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

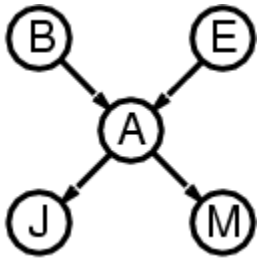
Example contd.



Compactness of Full Joint Encoding Due to Network Sparseness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1-p$)



- If each variable has no more than k parents, the complete network requires $n \cdot 2^k$ numbers

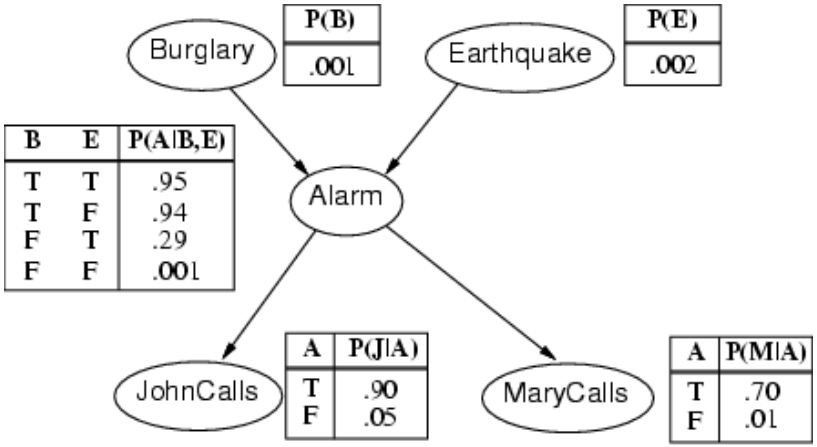
- i.e., grows linearly with n, vs. 2^n for the full joint distribution

- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$\boldsymbol{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \boldsymbol{P}(X_i \mid Parents(X_i))$$



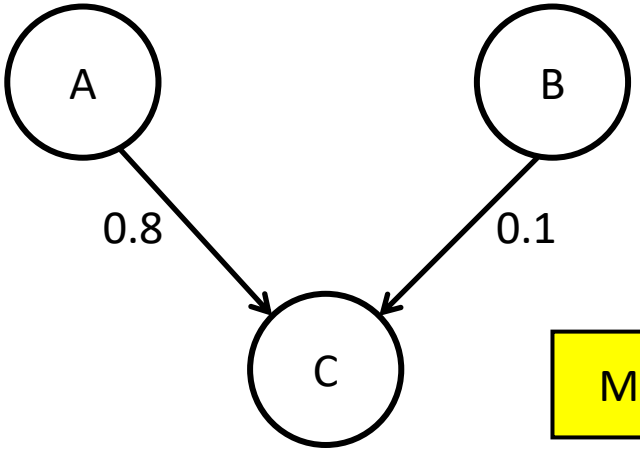
e.g., $\boldsymbol{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$= \boldsymbol{P}(j \mid a)\, \boldsymbol{P}(m \mid a)\, \boldsymbol{P}(a \mid \neg b, \neg e)\, \boldsymbol{P}(\neg b)\, \boldsymbol{P}(\neg e)$

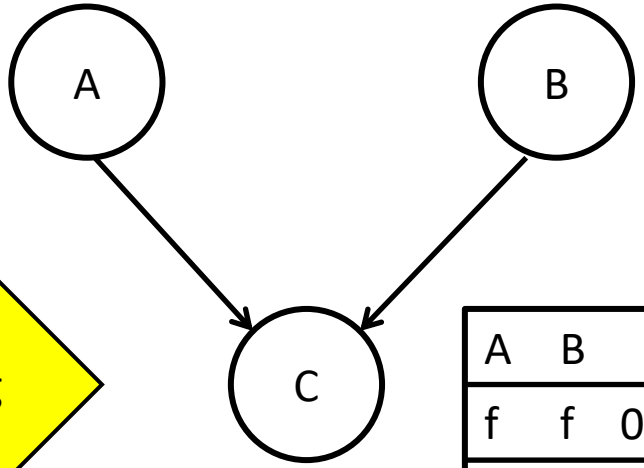
$= 0.90 \times 0.7 \times 0.001 \times 0.999 \times 0.998$

≈ 0.00063

Noisy-OR-Representation



Meaning



A	B	c
f	f	0.0
f	t	0.1
t	f	0.8
t	t	0.1+0.8 - 0.1*0.8

- Substantial reduction of the probabilities to be stored

Inference tasks

Simple queries: compute posterior marginal $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$
e.g., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$

Conjunctive queries: $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

Optimal decisions: decision networks include utility information;
probabilistic inference required for $P(outcome|action, evidence)$

Value of information: which evidence to seek next?

Sensitivity analysis: which probability values are most critical?

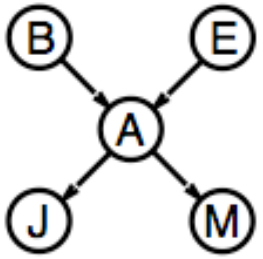
Explanation: what must be true to allow for a certain (desired) conclusion?

Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:

$$\begin{aligned}&\mathbf{P}(B|j,m)\\&= \mathbf{P}(B,j,m)/P(j,m)\\&= \alpha\mathbf{P}(B,j,m)\\&= \alpha \sum_e \sum_a \mathbf{P}(B,e,a,j,m)\end{aligned}$$



Rewrite full joint entries using product of CPT entries:

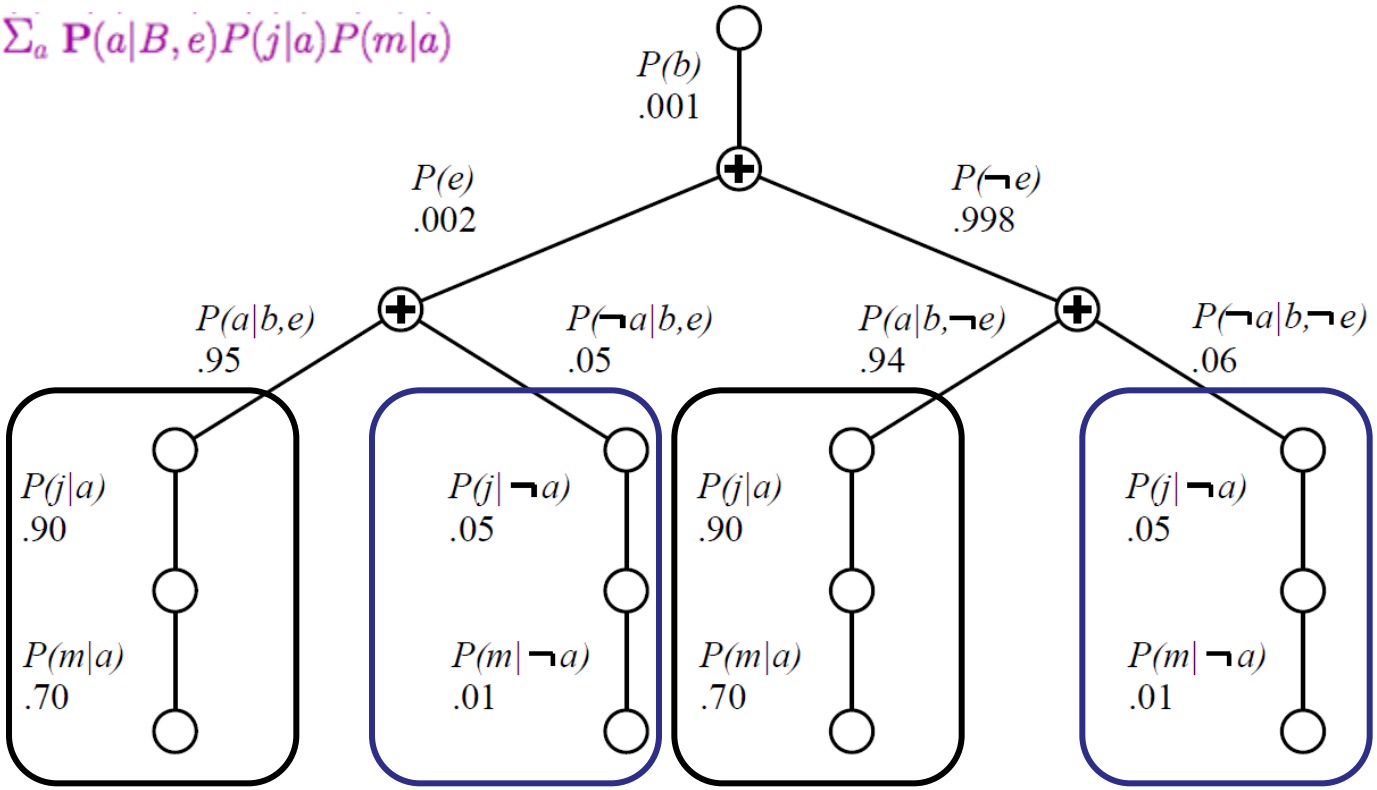
$$\begin{aligned}&\mathbf{P}(B|j,m)\\&= \alpha \sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)\\&= \alpha\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)\end{aligned}$$

Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

(Calculation over tree sum-product evaluation tree with path length n (=number of RVs) and degree d (=maximal number domain elements))

Evaluation Tree

$$\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) P(m|a)$$



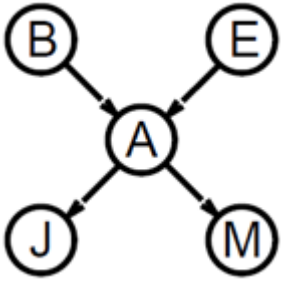
Enumeration is inefficient: repeated computation
e.g., computes $P(j|a)P(m|a)$ for each value of e

Can do better with Variable Elimination (VE)

Basic Objects of VE

$$\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) P(m|a)$$

- Track objects called factors
  (right-to-left, in tree: bottom up)
- Initial factors are local CPTs



$$\underbrace{P(B)}_{f_B(B)} \quad \underbrace{P(J|A)}_{f_J(A,J)} \quad \underbrace{P(A|B,E)}_{f_A(A,B,E)} \quad \underbrace{P(M|A)}_{f_M(A)}$$

- During elimination create new factors
- Form of Dynamic Programming

Basic Operations: Pointwise Product

- Pointwise product of factors $\mathbf{f}_1$ and $\mathbf{f}_2$
  - for example: $\mathbf{f}_1(A,B) * \mathbf{f}_2(B,C) = \mathbf{f}(A,B,C)$
  - in general:

    $\mathbf{f}_1(X_1,\ldots,X_j,Y_1,\ldots,Y_k) * \mathbf{f}_2(Y_1,\ldots,Y_k,Z_1,\ldots,Z_l) =$
    $\mathbf{f}(X_1,\ldots,X_j,Y_1,\ldots,Y_k,Z_1,\ldots,Z_l)$
  - has 2^{j+k+l} entries (if all variables are binary)

Join by pointwise product

A	$f_{JM}(A)$
T	.9 * .7
F	.05 * .01

=

A	$f_J(A)$
T	.9
F	.05

A	$f_M(A)$
T	.7
F	.01

$$\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) P(m|a)$$

A	B	E	$f_{AJM}(A,B,E)$
T	T	T	.95 * .63
T	T	F	.94 * .63
T	F	T	.29 * .63
T	F	F	.001 * .63
F	T	T	.05 * .0005
F	T	F	.06 * .0005
F	F	T	.71 * .0005
F	F	F	.999 * .0005

=

A	B	E	$f_{A\cdots}(A,B,E)$
T	T	T	.95
T	T	F	.94
T	F	T	.29
T	F	F	.001
F	T	T	.05
F	T	F	.06
F	F	T	.71
F	F	F	.999

A	$f_{JM}(A)$
T	.63
F	.0005

Basic Operations: Summing out

- Summing out a variable from a product of factors
  - Move any constant factors outside the summation
  - Add up submatrices in pointwise product of remaining factors

$$\Sigma_x f_1 * \ldots * f_k = f_1 * \ldots * f_i * \Sigma_x f_{i+1} * \ldots * f_k$$
$$= f_1 * \ldots * f_i * f_X$$

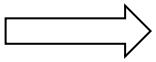
assuming $\mathbf{f_1, \ldots, f_i}$ do not depend on *X*

Summing out

$$\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) P(m|a)$$

A	B	E	$f_{AJM}(A,B,E)$
T	T	T	.95 * .63
T	T	F	.94 * .63
T	F	T	.29 * .63
T	F	F	.001 * .63
F	T	T	.05 * .0005
F	T	F	.06 * .0005
F	F	T	.71 * .0005
F	F	F	.999 * .0005

Summing out a



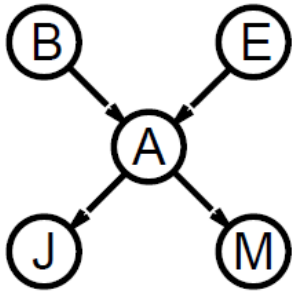
B	E	$f_{\bar{A}JM}(B,E)$
T	T	.95 * .63 + .05 * .0005 = .5985
T	F	.94 * .63 + .06 * .0005 = .5922
F	T	.29 * .63 + .71 * .0005 = .1830
F	F	.001 * .63 + .999 * .0005 = .001129

VE result for Burglary Example

$$
\begin{aligned}
&\mathbf{P}(B|j,m) \\
&\quad = \alpha \underbrace{\mathbf{P}(B)}_{B} \Sigma_e \underbrace{P(e)}_{E} \Sigma_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a f_A(a,b,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A\text{)} \\
&\quad = \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)} \\
&\quad = \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

Further Optimization by Finding Irrelevant Variables

Consider the query $P(JohnCalls|Burglary = true)$



$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(J|a) \sum_m P(m|a)$$

Sum over m is identically 1; M is **irrelevant** to the query

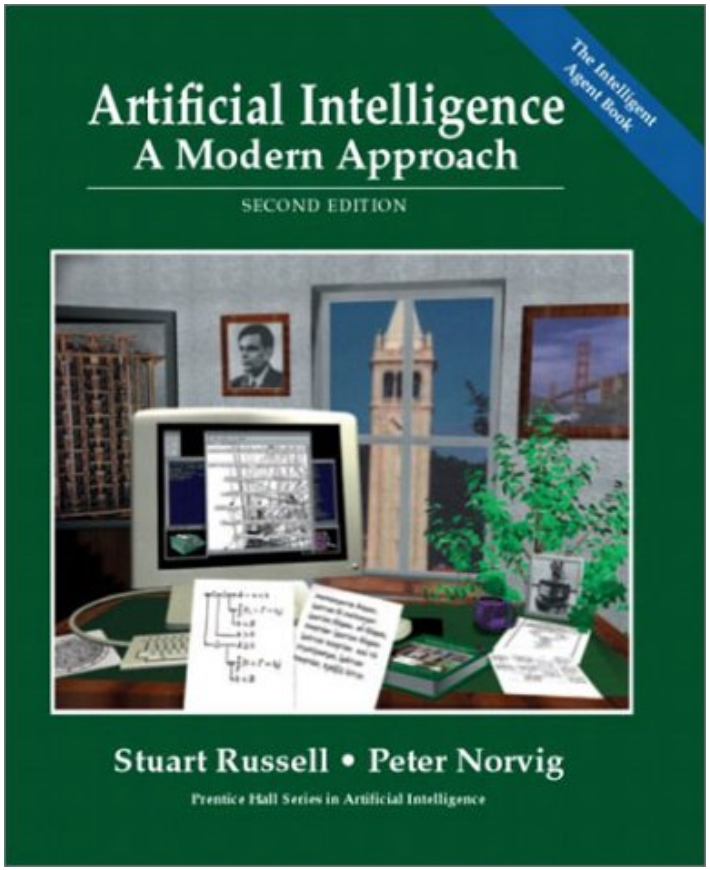
Thm 1: Y is irrelevant unless $Y \in Ancestors(\{X\} \cup \mathbf{E})$

Here, $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$, and
$Ancestors(\{X\} \cup \mathbf{E}) =$ {Alarm, Earthquake, Burglary}
so $MaryCalls$ is irrelevant

Irrelevant variables can be identified by a graph theoretical criterion on the associated moral graph by m-separation
Many other inference algorithms and optimizations known
→ Discussed in detail in course on Bayesian Knowledge Representation

Acknowledgements

- Slides from AIMA book provided by Cristina Conati, UBC



Data Mining Bayesian Networks

- Full Bayesian Learning
- MAP learning
- Maximum Likelihood Learning
- Learning Bayesian Networks
  - Fully observable
  - With hidden (unobservable) variables

Full Bayesian Learning

- In the learning methods we have seen so far, the idea was always to find the best model that could explain some observations
- In contrast, full Bayesian learning sees learning as Bayesian updating of a probability distribution over the hypothesis space, given data
  - H is the hypothesis variable
  - Possible hypotheses (values of H) $h_1 \ldots, h_n$
  - $P(H)$ = prior probability distribution over hypothesis space
- j_{th} observation d_j gives the outcome of random variable D_j
  - training data $\mathbf{d} = d_1, .., d_k$

Example

- Suppose we have 5 types of candy bags
  - 10% are 100% cherry candies (h_{100})
  - 20% are 75% cherry + 25% lime candies (h_{75})
  - 40% are 50% cherry + 50% lime candies (h_{50})
  - 20% are 25% cherry + 75% lime candies (h_{25})
  - 10% are 100% lime candies (h_0)

- Then we observe candies drawn from some bag

- Let's call θ the parameter that defines the fraction of cherry candy in a bag, and h_θ the corresponding hypothesis
- Which of the five kinds of bag has generated my 10 observations? $P(h_\theta \mid \mathbf{d})$.
- What flavour will the next candy be? Prediction $\mathbf{P}(X \mid \mathbf{d})$

Full Bayesian Learning

- Given the data so far, each hypothesis h_i has a posterior probability:
  - $P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i)\, P(h_i)$ (**Bayes theorem**)
  - where $P(\mathbf{d} \mid h_i)$ is called the likelihood of the data under each hypothesis
- Predictions over a new entity X are a weighted average over the prediction of each hypothesis:
  - $\mathbf{P}(X \mid \mathbf{d}) =$
    $= \sum_i \mathbf{P}(X, h_i \mid \mathbf{d})$
    $= \sum_i \mathbf{P}(X \mid h_i, \mathbf{d})\, P(h_i \mid \mathbf{d})$
    $= \sum_i \mathbf{P}(X \mid h_i)\, P(h_i \mid \mathbf{d})$
    $\sim \sum_i \mathbf{P}(X \mid h_i)\, P(\mathbf{d} \mid h_i)\, P(h_i)$
  - The weights are given by the data likelihood and prior of each h

The data does not add anything to a prediction given a hypothesis

+ No need to pick one best-guess hypothesis!

- Need to iterate over all hypotheses

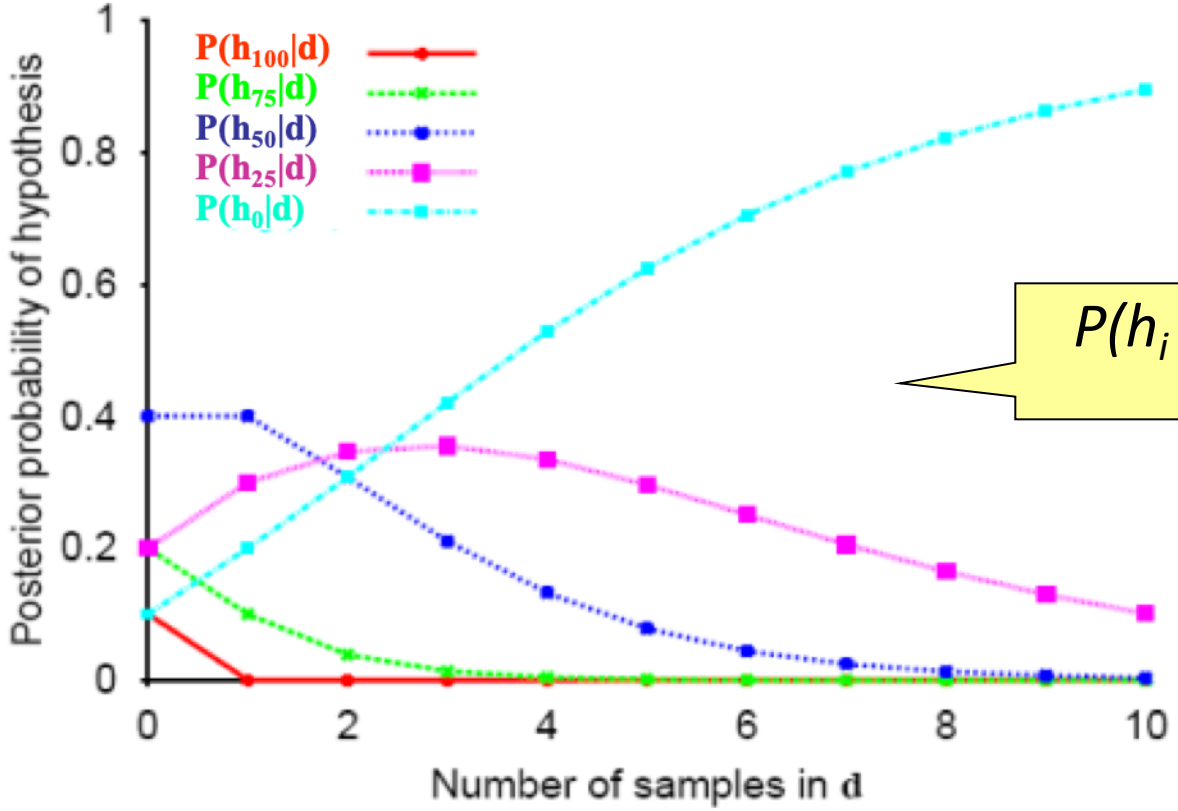
Example

- If we re-wrap each candy and return it to the bag, our 10 observations are independent and identically distributed, i.i.d, so
  - $P(\mathbf{d} \mid h_\theta) = \prod_j P(d_j \mid h_\theta)$ for j=1,..,10
- For a given h_θ , the value of $P(d_j \mid h_\theta)$ is
  - $P(d_j = \text{cherry} \mid h_\theta) = \theta$; $P(d_j = \text{lime} \mid h_\theta) = (1-\theta)$
- And given N observations, of which c are cherry and l = N-c lime

$$P(\boldsymbol{d} \mid h_\theta) = \prod_{j=1}^{c} \theta \prod_{j=1}^{\ell} (1-\theta) = \theta^c (1-\theta)^\ell$$

  - **Binomial distribution**: probability of # of successes in a sequence of N independent trials with binary outcome, each of which yields success with probability θ.
- For instance, after observing 3 lime candies in a row:
  - $P([\text{lime, lime, lime}] \mid h_{50}) = 0.5^3$ because the probability of seeing lime for each observation is 0.5 under this hypotheses

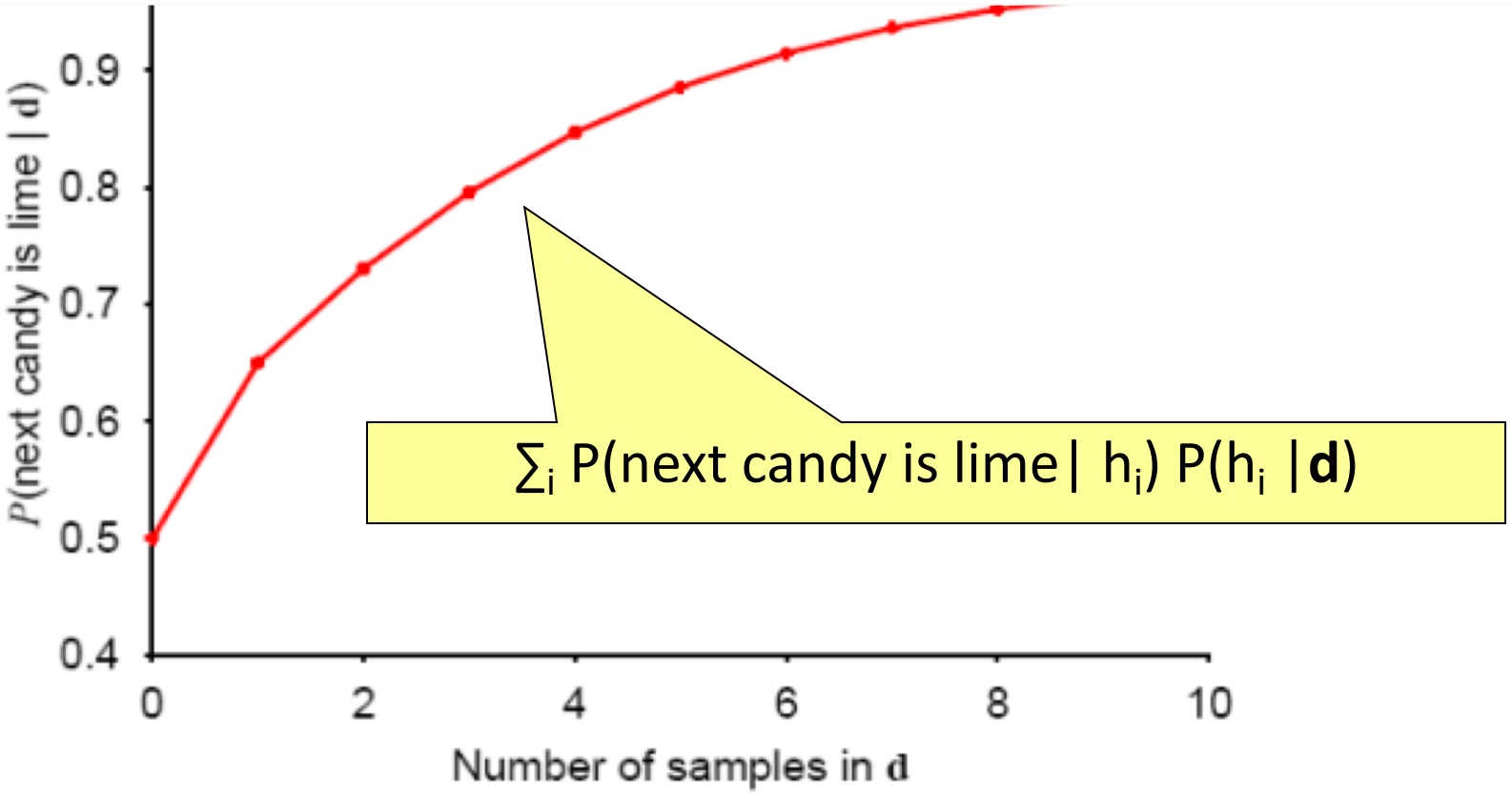
All-limes: Posterior Probability of H



$$P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i)\, P(h_i)$$

- Initially, the h_p with higher priors dominate (h_{50} with prior = 0.4)
- As data comes in, the finally best hypothesis (h_0) starts dominating, as the probability of seeing this data given the other hypotheses gets increasingly smaller
  - After seeing three lime candies in a row, the probability that the bag is the all-lime one starts taking off

Prediction Probability



- The probability that the next candy is lime increases with the probability that the bag is an all-lime one

Overview

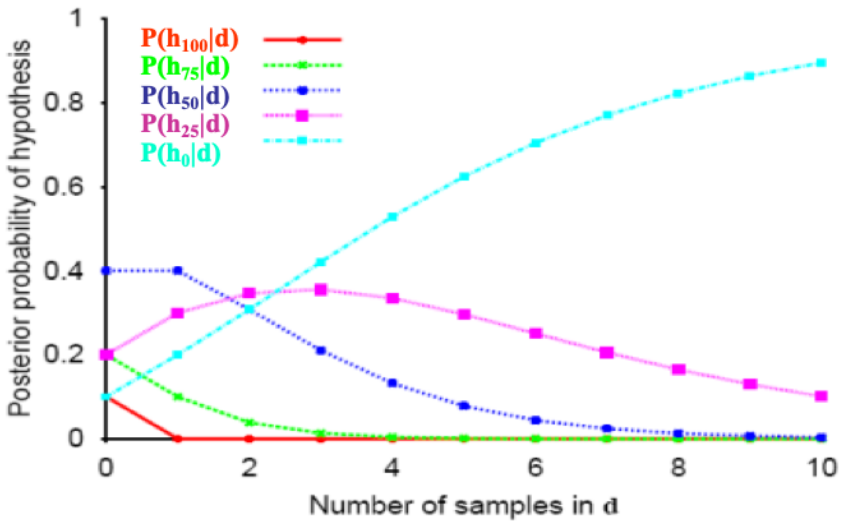
- Full Bayesian Learning
- MAP learning
- Maximum Likelihood Learning
- Learning Bayesian Networks
  - Fully observable
  - With hidden (unobservable) variables

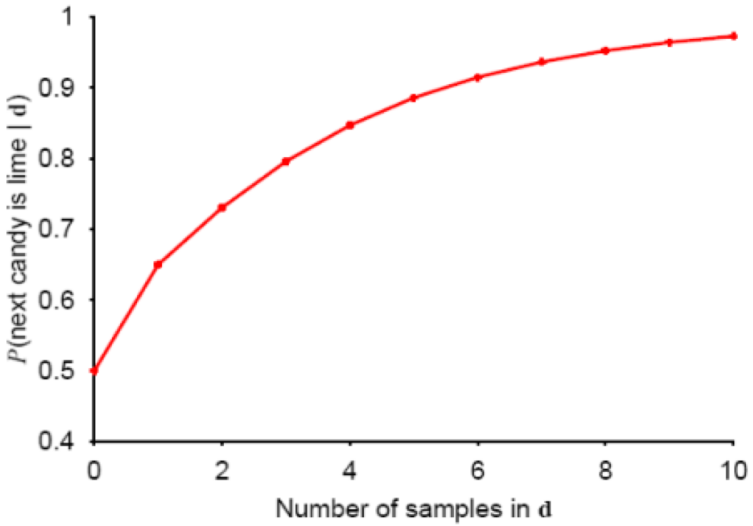
MAP approximation

- Full Bayesian learning seems like a very safe bet, but unfortunately it does not work well in practice
  - Summing over the hypothesis space is often intractable (e.g., 18,446,744,073,709,551,616 Boolean functions of 6 attributes)
- Very common approximation: Maximum a posterior (MAP) learning:
  - Instead of doing prediction by considering all possible hypotheses , as in
    - $\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|\ h_i)\ P(h_i\ |\mathbf{d})$
  - Make predictions based on h_{MAP} that maximises $P(h_i\ |\mathbf{d})$
    - I.e., maximize $P(\mathbf{d}|\ h_i)\ P(h_i)$
    - $\mathbf{P}(X|\mathbf{d}) \approx \mathbf{P}(X|\ h_{MAP}\ )$

MAP approximation

- MAP is a good approximation when $P(X \mid \mathbf{d}) \approx P(X \mid h_{MAP})$
  - In our example, h_{MAP} is the all-lime bag after only 3 candies, predicting that the next candy will be lime with $p = 1$
  - The Bayesian learner gave a prediction of 0.8, safer after seeing only 3 candies





Bias

- As more data arrive, MAP and Bayesian prediction become closer, as MAP's competing hypotheses become less likely
- Often easier to find MAP (optimization problem) than deal with a large summation problem
- *P(H)* plays an important role in both MAP and Full Bayesian Learning (defines learning bias)
- Used to define a tradeoff between model complexity and its ability to fit the data
  - More complex models can explain the data better => higher $P(\mathbf{d} \mid h_i)$ **danger of overfitting**
  - But they are less likely a priory because there are more of them than simpler model => lower $P(h_i)$
  - I.e. common learning bias is to penalize complexity

Overview

- Full Bayesian Learning
- MAP learning
- Maximum Likelihood Learning
- Learning Bayesian Networks
  - Fully observable
  - With hidden (unobservable) variables

Maximum Likelihood (ML)Learning

- Further simplification over full Bayesian and MAP learning
  - Assume uniform priors over the space of hypotheses
  - MAP learning (maximize $P(\mathbf{d} \mid h_i)\, P(h_i)$) reduces to maximize $P(\mathbf{d} \mid h_i)$
- When is ML appropriate?

Howard, R.: Decision analysis: Perspectives on inference, decision, and experimentation. Proceedings of the IEEE 58(5), 632-643, **1970**

Maximum Likelihood (ML) Learning

- Further simplification over Full Bayesian and MAP learning
  - Assume uniform prior over the space of hypotheses
  - MAP learning (maximize $P(\mathbf{d}|\ h_i)\ P(h_i)$) reduces to maximize $P(\mathbf{d}|\ h_i)$
- When is ML appropriate?
  - Used in statistics as the standard (non-bayesian) statistical learning method by those who distrust subjective nature of hypotheses priors
  - When the competing hypotheses are indeed equally likely (e.g. have same complexity)
  - With very large datasets, for which $P(\mathbf{d}|\ h_i)$ tends to overcome the influence of $P(h_i)$

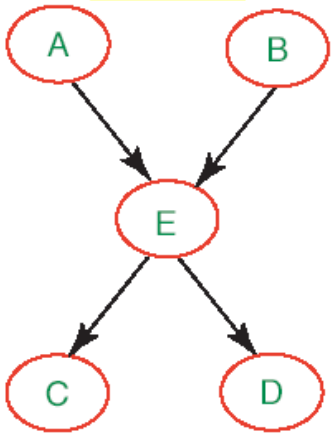
Overview

- Full Bayesian Learning
- MAP learning
- Maximum Likelihood Learning
- Learning Bayesian Networks
  - Fully observable (complete data)
  - With hidden (unobservable) variables

Learning BNets: Complete Data

- We will start by applying ML to the simplest type of BNets learning:
  - known structure
  - Data containing observations for all variables
    - ✓ All variables are observable, no missing data
- The only thing that we need to learn are the network's parameters

Model → Data → Probabilities



A	B	C	D	E
t	f	t	t	f
f	t	t	t	t
t	t	f	t	f
		...		

$P(A)$
$P(B)$
$P(E|A, B)$
$P(C|E)$
$P(D|E)$

Maximum-Likelihood-Parameterschätzung

- Nehme an, die Struktur eines BNs sei bekannt
- Ziel: Schätze BN-Parameter θ
  - Einträge in CPTs, P(X | Parents(X))
- Eine Parametrierung θ ist gut, falls hierdurch die beobachteten Daten wahrscheinlich generiert werden:

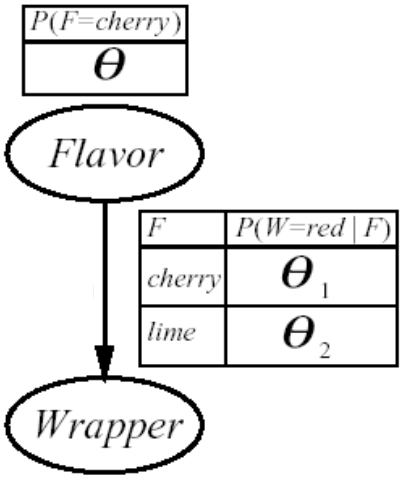
$$P(D \mid \theta) = \prod_m P(x[m] \mid \theta)$$

- Maximum Likelihood Estimation (MLE) Prinzip: Wähle θ^* so, dass $P(D \mid \theta^*)$ maximiert wird

Gleichverteilte, unabhängige Stichprobem (i.i.d. samples)

Anwendungsbeispiel Bonbonfabrik

- Ein Hersteller wählt die Farbe des Bonbonpapiers mit einer bestimmten Wahrscheinlichkeit je nach Geschmack, wobei die entsprechende Verteilung nicht bekannt sei
  - Wenn Geschmack=cherry, wähle rotes Papier mit W'keit θ_1
  - Wenn Geschmack=lime, wähle rotes Papier mit W'keit θ_2
- Das Bayessche Netzwerk enthält drei zu lernende Parameter
  - $\theta\, \theta_1\, \theta_2$

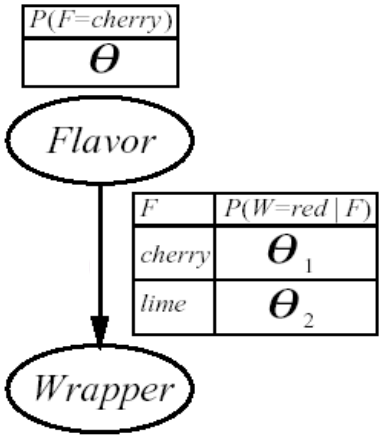


Anwendungsbeispiel Bonbonfabrik

- P(W=green, F = cherry| $h_{\theta\theta_1\theta_2}$) = **(*)**

  = P(W=green|F = cherry, $h_{\theta\theta_1\theta_2}$) P(F = cherry| $h_{\theta\theta_1\theta_2}$)

  = $\theta\ (1-\theta_1)$



- Wir packen N Bonbons aus
  - *c* sind cherry und ℓ sind lime
  - r^c cherry mit rotem Papier, g^c cherry mit grünem Papier
  - r^l lime mit rotem Papier, g^ℓ lime mit grünem Papier
  - Jeder Versuch liefert eine Kombination aus Papier und Geschmack wie bei (*)
- P(**d**| $h_{\theta\theta_1\theta_2}$)

  = $\prod_j P(d_j|\ h_{\theta\theta_1\theta_2})\ = \theta^c\ (1-\theta)^{\ell}\ (\theta_1)^{r^c}\ (1-\theta_1)^{g^c}\ (\theta_2)^{r^\ell}\ (1-\theta_2)^{g^\ell}$

Anwendungsbeispiel Bonbonfabrik

- Maximierung des Logarithmus der Zielfunktion
  - $c \log\theta + \ell \log(1-\theta) + r^c \log\theta_1 + g^c \log(1-\theta_1) + r^l \log\theta_2 + g^\ell \log(1-\theta_2)$
- Bestimmung der Ableitungen bzgl. $\theta, \theta_1, \theta_2$
  - Ausdrücke ohne Term, nach dem abgeleitet wird, verschwinden

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+g_c}$$

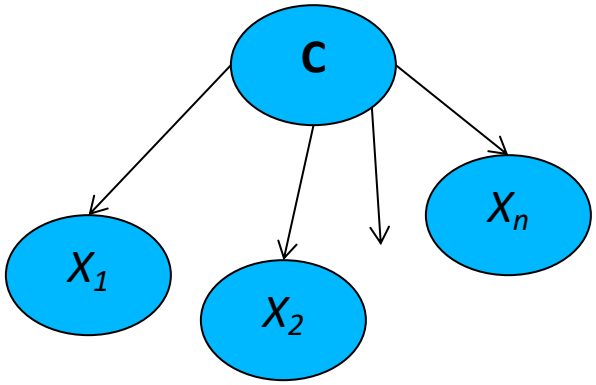
$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$

Maximum-Likelihood-Parameterschätzung

- Schätzung durch Bildung relativer Häufigkeiten
- Dieser Prozess ist auf jedes voll beobachtbare BN anwendbar
- Mit vollständigen Daten und Maximum-Likelihood-Parameterschätzung:
  - Parameterlernen zerfällt in separate Lernprobleme für jeden Parameter (CPT) durch Logarithmierung
  - Jeder Parameter wird durch die relative Häufigkeit eines Knotenwertes bei gegebenen Werten der Elternknoten bestimmt

Beliebte Anwendung: Naives Bayes-Modell

- Naïve Bayes-Modell: Sehr einfaches Bayessches Netzwerk zur Klassifikation
  - *Klassenvariable C* (vorherzusagen) bildet Wurzel
  - Attributvariablen X_i (Beobachtungen) sind Blätter



- Naiv, weil angenommen wird, dass die Attributwerte bedingt unabhängig sind, wenn die Klasse gegeben ist

$$\boldsymbol{P}(C|x_1,x_2,..,x_n) = \frac{\boldsymbol{P}(C,x_1,x_2,..,x_n)}{\boldsymbol{P}(x_1,x_2,..,x_n)} = \alpha \boldsymbol{P}(C)\prod_i \boldsymbol{P}(x_n \mid C)$$

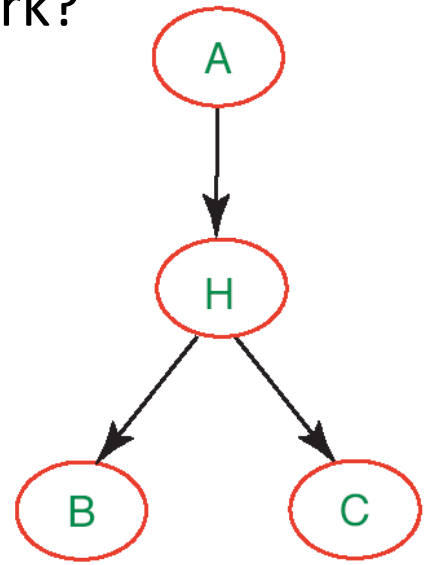
- Deterministische Vorhersagen können durch Wahl der wahrscheinlichsten Klasse erreicht werden
- Skalierung auf realen Daten sehr gut:
  - 2n + 1 Parameter benötigt

Überblick

- Volles Bayessches Lernen (BMA)
- MAP-Lernen
- Maximum-Likelihood-Lernen
  - Vollständig beobachtbar (vollständige Daten)
  - Mit versteckten (unbeobachtbaren) Variablen

Parameterlernen mit versteckten Variablen

- Bisher haben wir angenommen, Daten für jede Variable des BNs stehen zur Verfügung
- Was machen wir, wenn das nicht der Fall ist, d.h. es gibt versteckte (hidden) Variablen im Netzwerk?

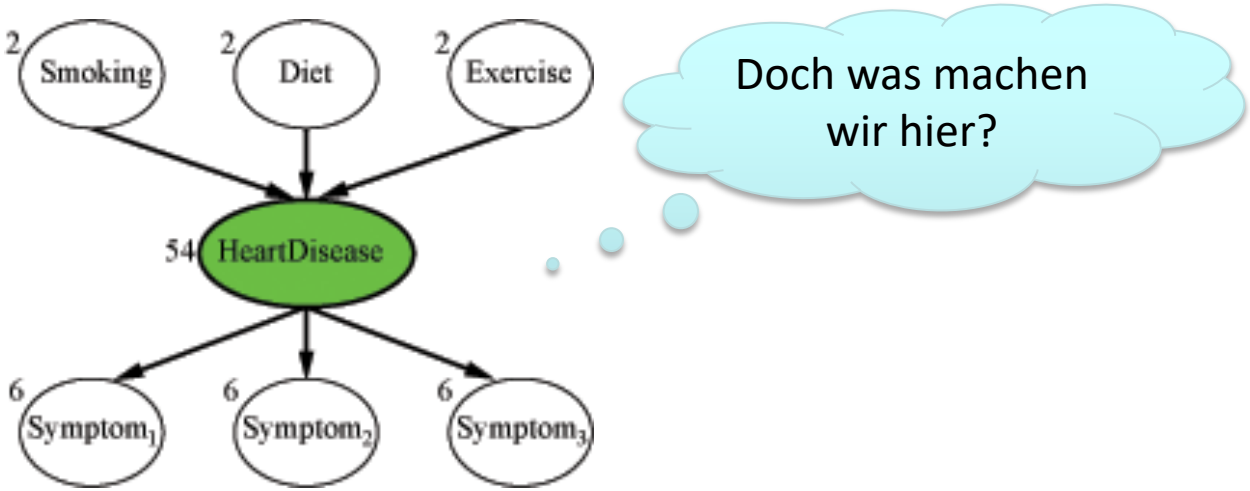


A	B	C
t	f	t
f	t	t
t	t	f
	...	

- Den Ansatz mit relativen Häufigkeiten können wir nicht so einfach übernehmen (keine Zähler für Variable H bekannt)

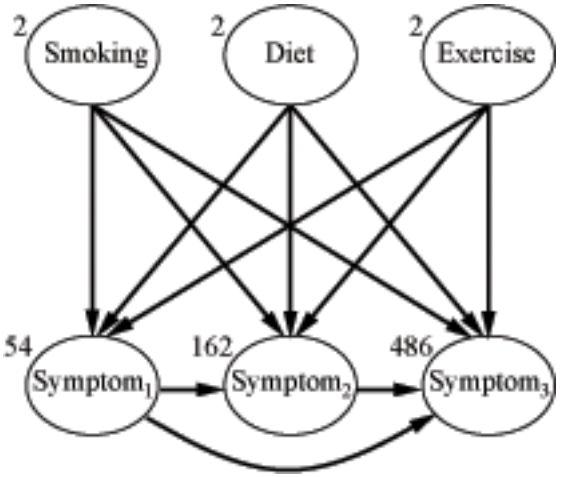
Vermeintliche Lösung

- Vermeide versteckte Variablen
- Könnte klappen bei kleinen Netzwerken



- Jede Variable habe 3 Werte (low, moderate, high)
- Die Zahlen an den Knoten repräsentieren, wie viele Parameter für die CPT des betreffenden Knotens bestimmt werden müssen
- 78 Wahrscheinlichkeitswerte insgesamt

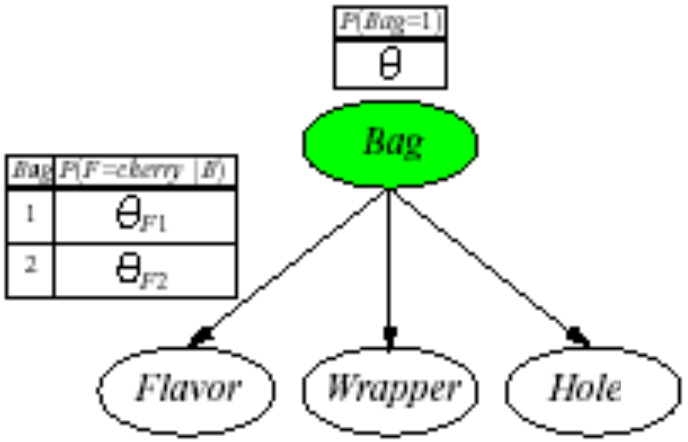
HeartDisease wegzulassen ist gar keine gute Lösung!



- Die Symptome sind nicht länger bedingt unabhängig gegeben die Elternknotenwerte
  - Sehr viel mehr Kanten, sehr viel mehr Wahrscheinlichkeiten zu spezifizieren: 708 insgesamt
  - Wir brauchen sehr viel mehr Daten, um diese ganzen Werte vernünftig zu lernen

Beispiel: Bonbonfabrik wieder einmal

- Zwei Bonbontüten (1 and 2) wurden vermischt
- Bonbons werden durch drei Eigenschaften beschrieben: Flavor und Wrapper wie vorher, plus Hole (ob ein Loch in der Mitte ist)
- Die Merkmale von Bonbon hängen mit bestimmten Wahrscheinlichkeiten von der Tüte ab, aus der sie kommen
- Wir wollen für jedes Bonbon vorhersagen, aus welcher Tüte es kam, je nach vorgefundenen Eigenschaften: Naïve-Bayes-Modell



θ= P(Bag = 1)

θ_{Fj} = P(Flavor = cherry | Bag = j)

θ_{Wj} = P(Wrapper = red | Bag = j)

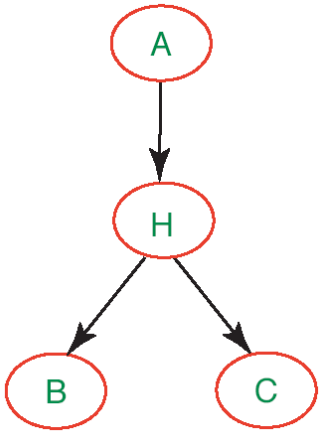
θ_{Hj} = P(Hole = yes | Bag = j)

j =1,2

Expectation-Maximization (EM)

- Wenn wir versteckte Variablen beibehalten und die Netzwerkparameter aus Daten lernen wollen, haben wir eine Form des unüberwachten Lernens (*unsupervised learning*)
  - Die Daten geben nicht über alle Aspekte von Datenpunkten Auskunft
- Expectation-Maximization
  - Genereller Algorithmus zum Lernen von Modellparametern bei unvollständigen Daten
  - Hier für diskrete Datenwerte im BN-Kontext gezeigt

EM: Generelle Idee



A	B	C
t	f	t
f	t	t
t	t	f
	...	

- Falls wir Daten für alle Variablen im BN hätten, könnten wir die Parameter mit ML- (oder MAP-) Ansätzen lernen
  - Relative Häufigkeiten bestimmen, wie vorher besprochen
- Falls wir die Parameter hätten, könnten wir die Posterior-Wahrscheinlichkeiten jedes Ereignisses schätzen:

P(H|A,B,C)

Dempster, A.P., Laird. N.M., Rubin, D.B.: *Maximum-Likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, **1977**

EM: Generelle Idee

- Der Algorithmus startet mit "erfundenen" (zufällig generierten) Informationen, um das Lernproblem zu lösen:
  - Erfundene Netzwerkparameterwerte (virtuell notiert in den CPTs)
- Dann werden die Initialwerte in zwei Schritten verfeinert:
  - Expectation (E): Aktualisiere die Daten (virtuell) mit aus dem aktuellen Modell hergeleiteten Erwartungen
  - Maximization (M): Gegeben die aktualisierten Daten, aktualisieren die Parameter des BNs mitteln Maximum Likelihood (ML) Ansatz
    - ✓ Das ist der gleiche Schritt, wie im voll beobachtbaren Fall

CPT: Conditional Probability Table

EM: Wie funktioniert das mit Naive Bayes-Modelen

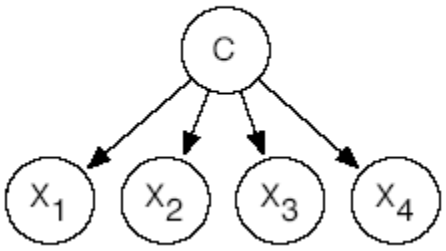
- Betrachtet die folgenden Daten, …
  - N Beispiele mit Booleschen Attributen X_1, X_2, X_3, X_4

Data

X_1	X_2	X_3	X_4
t	f	t	t
f	t	t	f
f	f	t	t
		…	

- … die wir kategorisieren wollen mit möglichen Werten einer Klasse C = {1,2,3}
- Wir verwenden einen naiven Bayes-Klassifikator mit versteckter Variable C

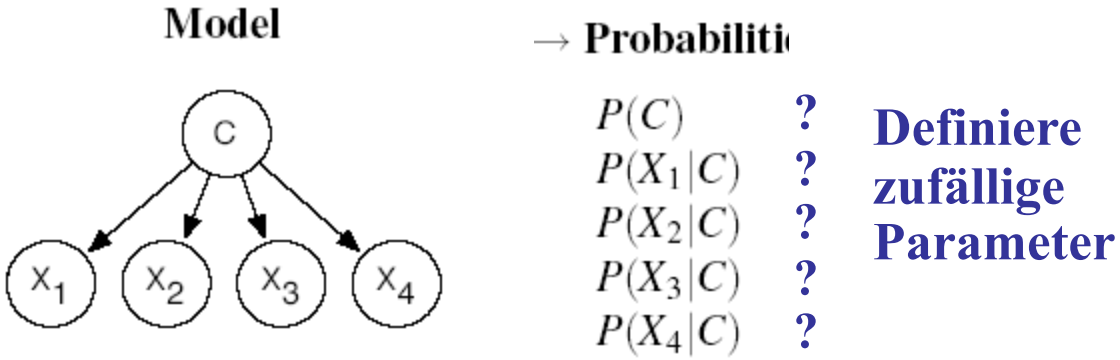
Model → **Probabiliti**



$P(C)$?
$P(X_1|C)$?
$P(X_2|C)$?
$P(X_3|C)$?
$P(X_4|C)$?

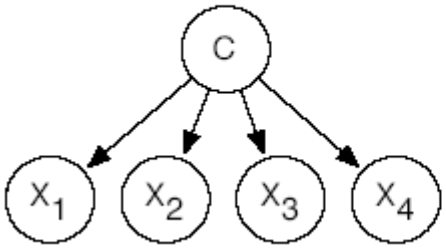
EM: Initialisierung

- Der Algorithmus startet mit "erfundenen" (zufällig generierten) Informationen, um das Lernproblem zu lösen:
  - Erfundene Netzwerkparameterwerte (virtuell notiert in den CPTs)



EM: Expectation-Schritt (Bestimmung der erwarteter Zahlen)

Model



Data

X_1	X_2	X_3	X_4
t	f	t	t
f	t	t	f
f	f	t	t
		...	

→ Probabilities

$P(C)$?
$P(X_1|C)$?
$P(X_2|C)$?
$P(X_3|C)$?
$P(X_4|C)$?

- Was benötigen wir, um die Netzwerkparameter mit dem Maximum-Likelihood-Ansatz zu lernen?
  - Für P(C) = Anzahl(Datenpunkte mit C=i) / Anzahl (alle Datenpunkte) i=1,2,3
  - Für $P(X_h|C)$ = Anzahl(Daten mit $X_h = val_k$ und C=i) / Anzahl(Daten mit C=i)

    für alle Werte val_k von X_h und i=1,2,3

EM: Expectation-Schritt (Bestimmung der erwarteter Zahlen)

- Bislang haben wir nur: N = Anzahl(Datenpunkte)
- Wir approximieren alle weiteren Zähler mit erwarteten Zählerwerten, die aus dem Modell mit den "erfundenen" Parametern abgeleitet werden
- Der erwartete Zähler $\hat{N}(C=i)$ ist die Summe, über alle N Beispieldaten, der Wahrscheinlichkeiten, dass jedes Beispiel in Kategorie i fällt

$$\hat{N}(C=i) = \sum_{j=1}^{N} P(C=i \mid \text{Attribute von Beispiel } e_j)$$

$$= \sum_{j=1}^{N} P(C=i \mid x1_j, x2_j, x3_j, x4_j)$$

EM: Expectation-Schritt (Bestimmung der erwarteter Zahlen)

- Wie erhalten wir die notwendigen Werte aus dem Modell?

$$\hat{N}(C=i) = \sum_{j=1}^{N} P(C=i \mid \text{ Attribute von Beispiel } e_j)$$

$$= \sum_{j=1}^{N} P(C=i \mid x1_j, x2_j, x3_j, x4_j)$$

- Leicht mit naivem Bayesschen Netzwerk

$$P(C=i \mid x1_j, x2_j, x3_j, x4_j) = \frac{P(C=i, x1_j, x2_j, x3_j, x4_j)}{P(x1_j, x2_j, x3_j, x4_j)}$$

$$= \frac{P(x1_j \mid C=i)..,P(x4_j \mid C=i)P(C=i)}{P(x1_j, x2_j, x3_j, x4_j)}$$

Auch direkt aus Netzwerk bestimmbar. Übungsaufgabe

Erfundene Parameter für das Netzwerk

EM: Expectation-Schritt (Bestimmung der erwarteter Zahlen)

- Durch einen ähnlichen Schritt erhalten wir die erwarteten Zähler für Beispiele mit Attribut $X_h = val_k$ und Kategorie i

- Diese werden später benötigt zur Schätzung von $\mathbf{P}(X_h \mid C)$:

$$P(X_h \mid C) = \frac{\text{Exp.Counts(examples with } X_h = val_k \text{ and } C = i)}{\text{Exp.Counts(examples with } C = i)} = \frac{\hat{N}(X_h = val_k, C = i)}{\hat{N}(C = i)}$$

  - für alle Werte val_k von X_h und i=1,2,3

- Zum Beispiel

$$\hat{N}(X_1 = t, C = 1) = \sum_{e_j \text{ with } X_1 = t} P(C = i \mid x1_j = t, x2_j, x3_j, x4_j)$$

Wiederum erhalten wir diese W'keiten aus dem aktuellen Modell

EM: Generelle Idee

- Der Algorithmus startet mit "erfundenen" (zufällig generierten) Informationen, um das Lernproblem zu lösen:
  - Erfundene Netzwerkparameterwerte (virtuell notiert in den CPTs)
- Dann werden die Initialwerte in zwei Schritten verfeinert:
  - Expectation (E): Aktualisiere die Daten (virtuell) mit aus dem aktuellen Modell hergeleiteten Erwartungen
  - Maximization (M): Gegeben die aktualisierten Daten, aktualisieren die Parameter des BNs mitteln Maximum Likelihood (ML) Ansatz
    - ✓ Das ist der gleiche Schritt, wie im voll beobachtbaren Fall

CPT: Conditional Probability Table

Maximization-Schritt: (Verfeinerung der Parameter)

- Nun verfeinern wir die Netzwerkparameter mittels Maximum-Likelihood-Lernen auf den erwarteten Zählern

$$P(C = i) = \frac{\hat{N}(C = i)}{N}$$

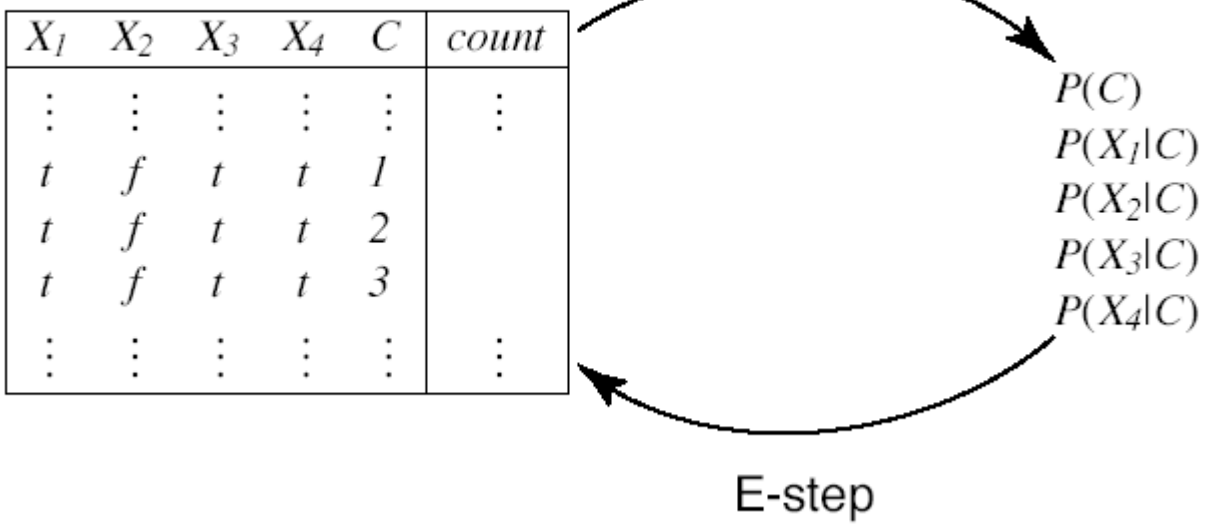
$$P(X_j = val_k \mid C = i) = \frac{\hat{N}(X_j = val_k C = i)}{\hat{N}(C = i)}$$

- für alle Werte val_k von X_j und $i=1,2,3$

EM-Zyklus

- Nun kann der E-Schritt wiederholt werden

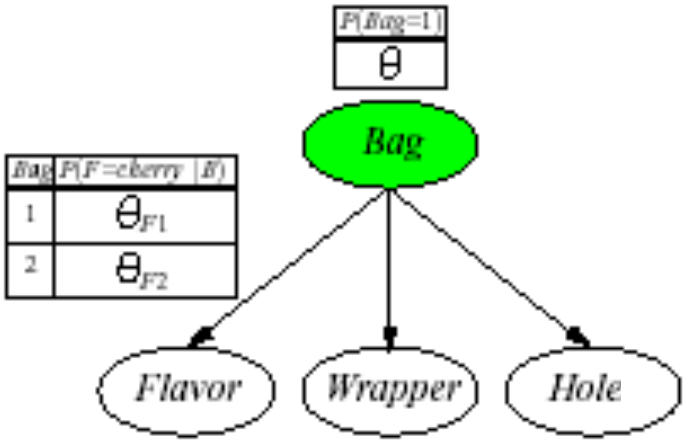
Erwartete Zähler
("Augmented data")



Wahrscheinlichkeiten

Beispiel: Bonbonfabrik wieder einmal

- Zwei Bonbontüten (1 and 2) wurden vermischt
- Bonbons werden durch drei Eigenschaften beschrieben: Flavor und Wrapper wie vorher, plus Hole (ob ein Loch in der Mitte ist)
- Die Merkmale von Bonbon hängen mit bestimmten Wahrscheinlichkeiten von der Tüte ab, aus der sie kommen
- Wir wollen für jedes Bonbon vorhersagen, aus welcher Tüte es kam, je nach vorgefundenen Eigenschaften: Naïve-Bayes-Modell



θ= P(Bag = 1)

θ_{Fj} = P(Flavor = cherry | Bag = j)

θ_{Wj} = P(Wrapper = red | Bag = j)

θ_{Hj} = P(Hole = yes | Bag = j)

j =1,2

Daten

- Nehmen wir an, die wahren Parameter seien:
  - $\theta = 0.5$;
  - $\theta_{F1} = \theta_{W1} = \theta_{H1} = 0.8$;
  - $\theta_{F2} = \theta_{W2} = \theta_{H2} = 0.3$;
- Annahme: Die folgenden Zähler werden "gesampelt" aus **P**(C, F, W, H) (N = 1000): Wir haben einen Datensatz

	W=red		W=green	
	H=1	H=0	H=1	H=0
F=cherry	273	93	104	90
F=lime	79	100	94	167

- Wir wollen nun die wahren Parameter aus diesen Daten mittels EM rekonstruieren

EM: Initialisierung

- Weise Parametern zufällige Initialwerte zu
  - Zu Vereinfachung der Darstellung verwenden wir hier:

$$\theta^{(0)} = 0.6;$$
$$\theta_{F1}^{(0)} = \theta_{W1}^{(0)} = \theta_{H1}^{(0)} = 0.6;$$
$$\theta_{F2}^{(0)} = \theta_{W2}^{(0)} = \theta_{H2}^{(0)} = 0.4$$

- Nun durchlaufen wir einen EM-Zyklus zur Berechnung von $\theta^{(1)}$.

E-Schritt

- Zuerst brauchen wir die erwarteten Zähler für Bonbon von Tüte 1:
  - Addiere die Wahrscheinlichkeiten, dass jedes der N Datenpunkte aus Tüte 1 kommt
  - Seien $flavor_j$, $wrapper_j$, $hole_j$ die Werte der entsprechenden Attribute für den j-ten Datenpunkt

$$\hat{N}(\text{Bag} = 1) = \sum_{j=1}^{N} P(Bag = 1 \mid flavor_j, wrapper_j, hole_j) =$$

$$= \sum_{j=1}^{N} \frac{P(flavor_j, wrapper_j, hole_j | Bag = 1) P(Bag = 1)}{P(flavor_j, wrapper_j, hole_j)}$$

$$= \sum_{j=1}^{N} \frac{P(flavor_j | Bag = 1) P(wrapper_j | Bag = 1) P(hole_j | Bag = 1) P(Bag = 1)}{\sum_i P(flavor_j | Bag = i) P(wrapper_j | Bag = i) P(hole_j | Bag = i) P(Bag = i)}$$

E-step

$$\sum_{j=1}^{N} \frac{P(flavor_j|Bag=1)P(wrapper_j|Bag=1)P(hole_j|Bag=1)P(Bag=1)}{\sum_i P(flavor_j|Bag=i)P(wrapper_j|Bag=i)P(hole_j|Bag=i)P(Bag=i)}$$

- Die Summation kann in die 8 Bonbongruppen aus der Tabelle aufgebrochen werden.
  - Zum Beispiel ergibt die Summe über 273 cherry-Bonbons mit rotem Papier und einem Loch (erster Eintrag in der Tabelle)

	W=red		W=green	
	H=1	H=0	H=1	H=0
F=cherry	273	93	104	90
F=lime	79	100	94	167

$$= 273 \frac{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)}}{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)} + \theta_{F2}^{(0)}\theta_{W2}^{(0)}\theta_{H2}^{(0)}(1-\theta^{(0)})} =$$

$$273 \frac{0.6^4}{0.6^4 + 0.4^4} = 273 \frac{0.1296}{0.1552} = 227.97$$

$$\theta^{(0)} = 0.6;$$

$$\theta_{F1}^{(0)} = \theta_{W1}^{(0)} = \theta_{H1}^{(0)} = 0.6;$$

$$\theta_{F2}^{(0)} = \theta_{W2}^{(0)} = \theta_{H2}^{(0)} = 0.4$$

M-Schritt

- Mit der Berechnung der anderen 7 Bonbongruppe erhalten wir

$$\hat{N}(Bag = 1) = 612.4$$

- Nun führen wir den M-Schritt aus, um θ zu verfeinern. Hierzu nehmen wir den erwarteten Zählerwert der Datenpunkte aus Tüte 1

$$\theta^{(1)} = \frac{\hat{N}(Bag = 1)}{N} = 0.6124$$

Noch ein Parameter…

- Wir machen das gleiche für den Parameter θ_{F1}
- E-Schritt: Bestimme erwarteten Zählerwert von cherry-Bonbons aus Tüte 1

$$\hat{N}(Bag = 1 \wedge Flavor = cherry) = \sum_{j:Flavor_j=cherry} P(Bag = 1 \mid Flavor_j = cherry, wrapper_j, hole_j)$$

  - Bestimmbar aus naivem Bayesschen Modell wie vorher besprochen
  - Als Übung…
- M-Schritt: Verfeinere θ_{F1} durch Bestimmung der relativen Häufigkeiten

$$\theta_{F1}^{(1)} = \frac{\hat{N}(Bag = 1 \wedge Flavor = cherry)}{\hat{N}(Bag = 1)}$$

Lernergebnis

- Nach einem vollen Zyklus über alle Parameter haben wir

$$\theta^{(1)} = 0.6124;$$

$$\theta_{F1}^{(1)} = 0.6684; \quad \theta_{W1}^{(1)} = 0.6483; \quad \theta_{H1}^{(1)} = 0.658;$$

$$\theta_{F2}^{(1)} = 0.3887; \quad \theta_{W2}^{(1)} = 0.3817; \quad \theta_{H2}^{(1)} = 0.3827;$$

- Für jeden Satz von Parametern können wir die Log-Likelihood bestimmen wie auch schon früher

- Man kann zeigen, dass dieser Wert mit jeder EM-Iteration steigt (Konvergenz)

- EM mit Maximum-Likelihood bleibt aber bei lokalen Maxima hängen bleiben (ggf. mehrere Startwerte nehmen, Priors berücksichtigen oder MAP nehmen)

Lernergebnis

- Nach einem vollen Zyklus über alle Parameter haben wir

$$\theta^{(1)} = 0.6124;$$

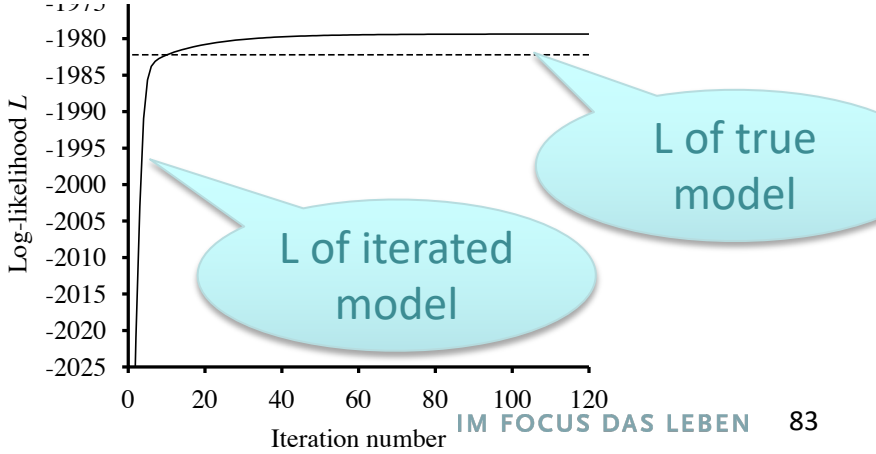
$$\theta_{F1}^{(1)} = 0.6684; \quad \theta_{W1}^{(1)} = 0.6483; \quad \theta_{H1}^{(1)} = 0.658;$$

$$\theta_{F2}^{(1)} = 0.3887; \quad \theta_{W2}^{(1)} = 0.3817; \quad \theta_{H2}^{(1)} = 0.3827;$$

- Für jeden Satz von Parametern können wir die Log-Likelihood bestimmen wie auch schon früher

$$P(\mathbf{d} \mid h_{\theta^{(i)}\theta_{F1}^{(i)}\theta_{W1}^{(i)}\theta_{H1}^{(i)}\ \theta_{F2}^{(i)}\ \theta_{W2}^{(i)}\theta_{H2}^{(i)}}) = \prod_{j=1}^{1000} P(d_j \mid h_{\theta^{(i)}\theta_{F1}^{(i)}\theta_{W1}^{(i)}\theta_{H1}^{(i)}\ \theta_{F2}^{(i)}\ \theta_{W2}^{(i)}\theta_{H2}^{(i)}})$$

- Wir können zeigen, das die Log-Likelihood $L = \log P(d, h_\theta)$ in jeder Iteration ansteigt, so dass die initiale Likelihood schon nach drei Iterationen überflügelt wird



EM: Diskussion

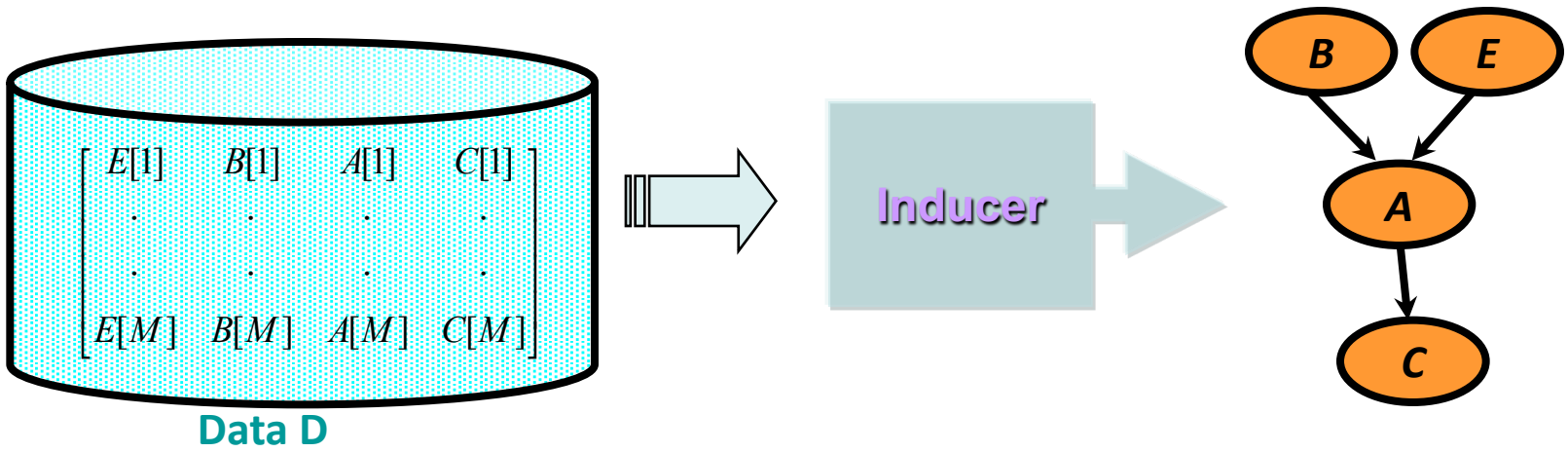
- Für komplexere BNs ist der Algorithmus der gleiche
  - Im allgemeinen müssen wir die CPT-Einträge für jede Variable X_i gegeben die Elternknotenwerte Pa_i berechnen
  - $\theta_{ijk} = P(X_i = x_{ij} \mid Pa_{i=} pa_{ik})$

$$\theta_{ijk} = \frac{\hat{N}(X_i = x_{ij}; Pa_i = pa_{ik})}{\hat{N}(Pa_i = pa_{ik})}$$

- Die erwarteten Zählerwerte werden durch Summierung über die Beispiele berechnet, nachdem all notwendigen $P(X_i = x_{ij}, Pa_{i=} pa_{ik})$ mittels BN-Algorithmen bestimmt sind
- Das Verfahren kann in eine kombinatorische Explosion laufen (ggf. Approximationstechniken angewendet)

Learning Bayesian network structures

- Given training set $\boldsymbol{D} = \{\boldsymbol{x}[1],\ldots,\boldsymbol{x}[\boldsymbol{M}]\}$
- Find model that best matches ***D***
  - model selection
  - parameter estimation



Some of the following slides from an AI course „Graphical models“ by Burgard/De Raedt/Kersting/Nebel

Model selection

Goal: Select the best network structure, given the data

Input:

- Training data
- Scoring function

Output:

- A network that maximizes the score

Structure selection: Scoring

- Bayesian: prior over parameters and structure
  - get balance between model complexity and fit to data as a byproduct

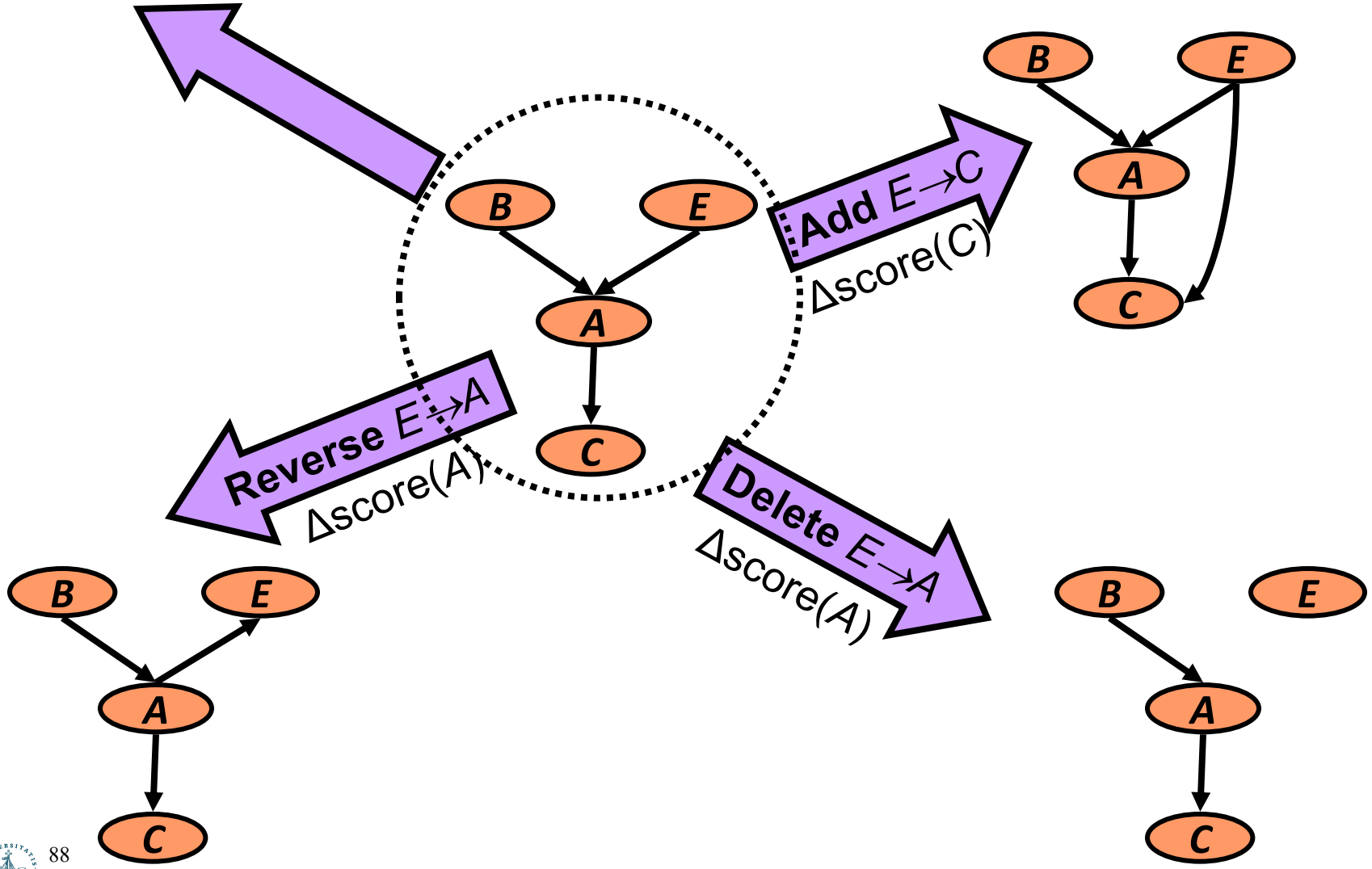
Can we learn G's params from D? | Does G explain D with ML? | Prior w.r.t. MDL

- $\text{Score}_D(G) = \log P(G|D) = \alpha \log [P(D|G)\, P(G)]$
- Marginal likelihood just comes from our parameter estimates
- Prior on structure can be any measure we want; typically a function of the network complexity (MDL principle)

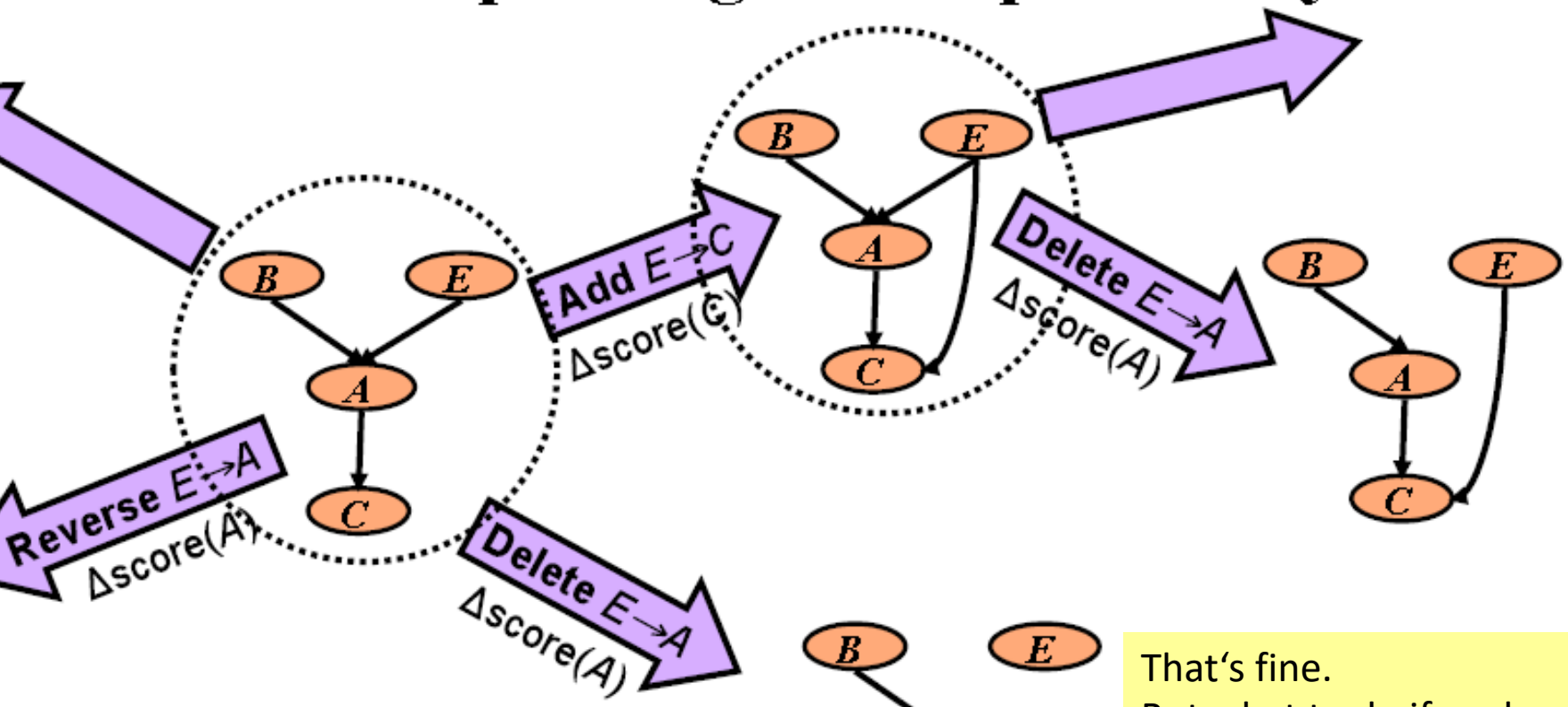
Same key property: Decomposability

$$\text{Score(structure)} = \Sigma_i \text{ Score(substructure of } X_i)$$

Heuristic search



Exploiting decomposability



To recompute scores,
only need to re-score families
that changed in the last move

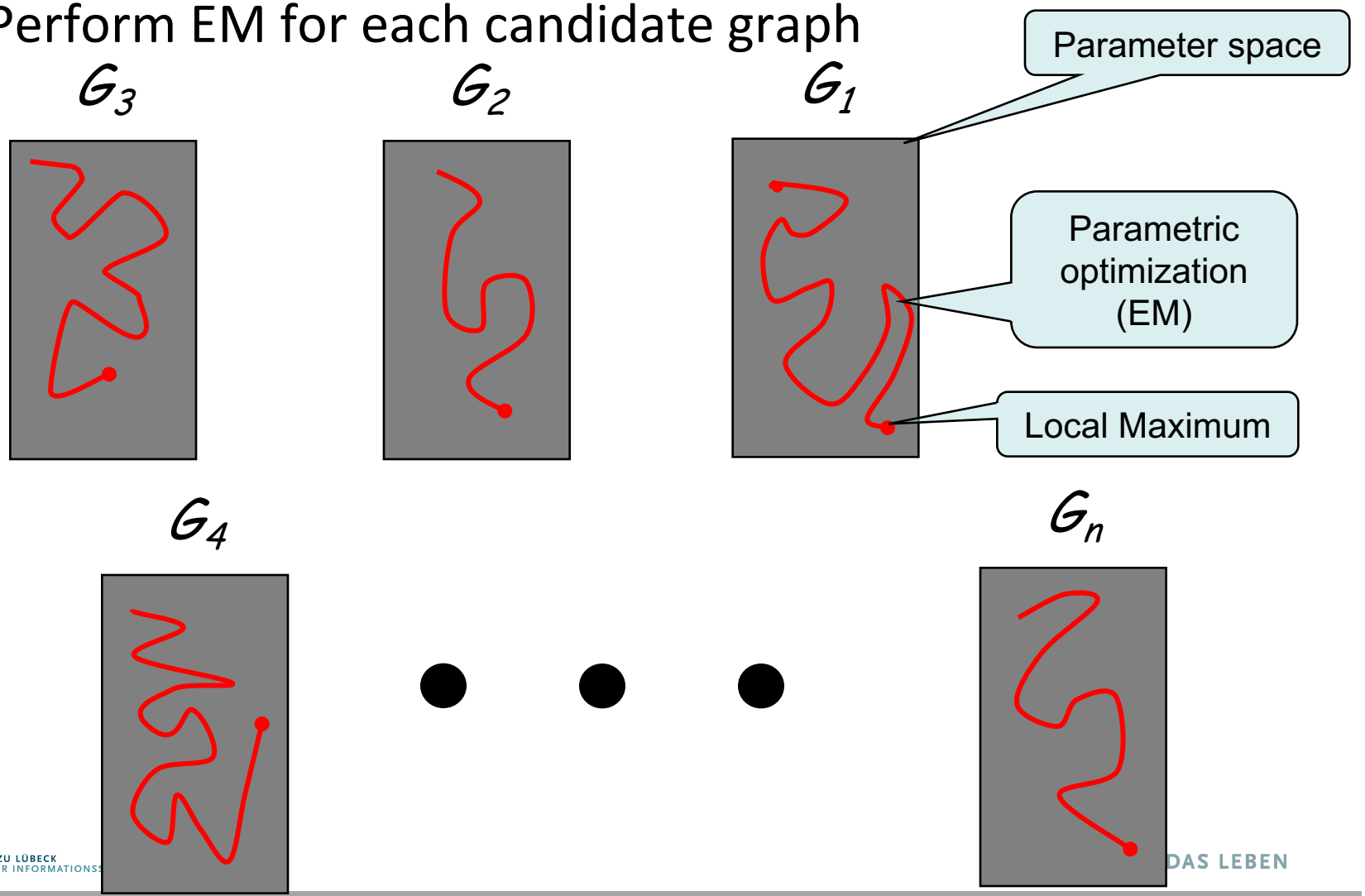
That's fine.
But what to do if we have non-complete data??

Cannot use decomposability

=> Rerun parameter estimation

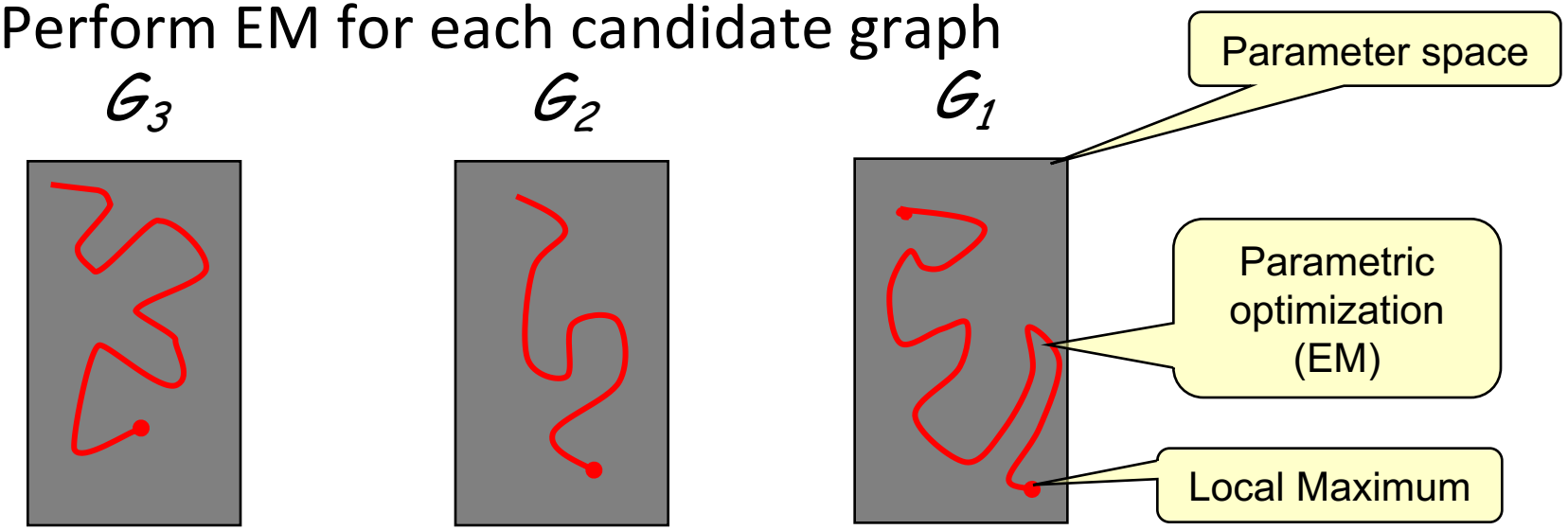
Local Search in Practice

- Perform EM for each candidate graph



Local Search in Practice

- Perform EM for each candidate graph



- Computationally expensive:
  - Parameter optimization via EM — non-trivial
  - Need to perform EM for all candidate structures
  - Spend time even on poor candidates

$\Rightarrow$ In practice, considers only a few candidates

Structural EM [Friedman et al. 98]

Recall, in complete data we had

- Decomposition $\Rightarrow$ efficient search

Idea:

- Instead of optimizing the real score…
- Find decomposable alternative score
- Such that maximizing new score

  $\Rightarrow$ improvement in real score

Structural EM

Idea:

- Use current model to help evaluate new structures

Outline:

- Perform search in (Structure, Parameters) space
- At each iteration, use current model for finding either:
  - Better scoring parameters: "parametric" EM step

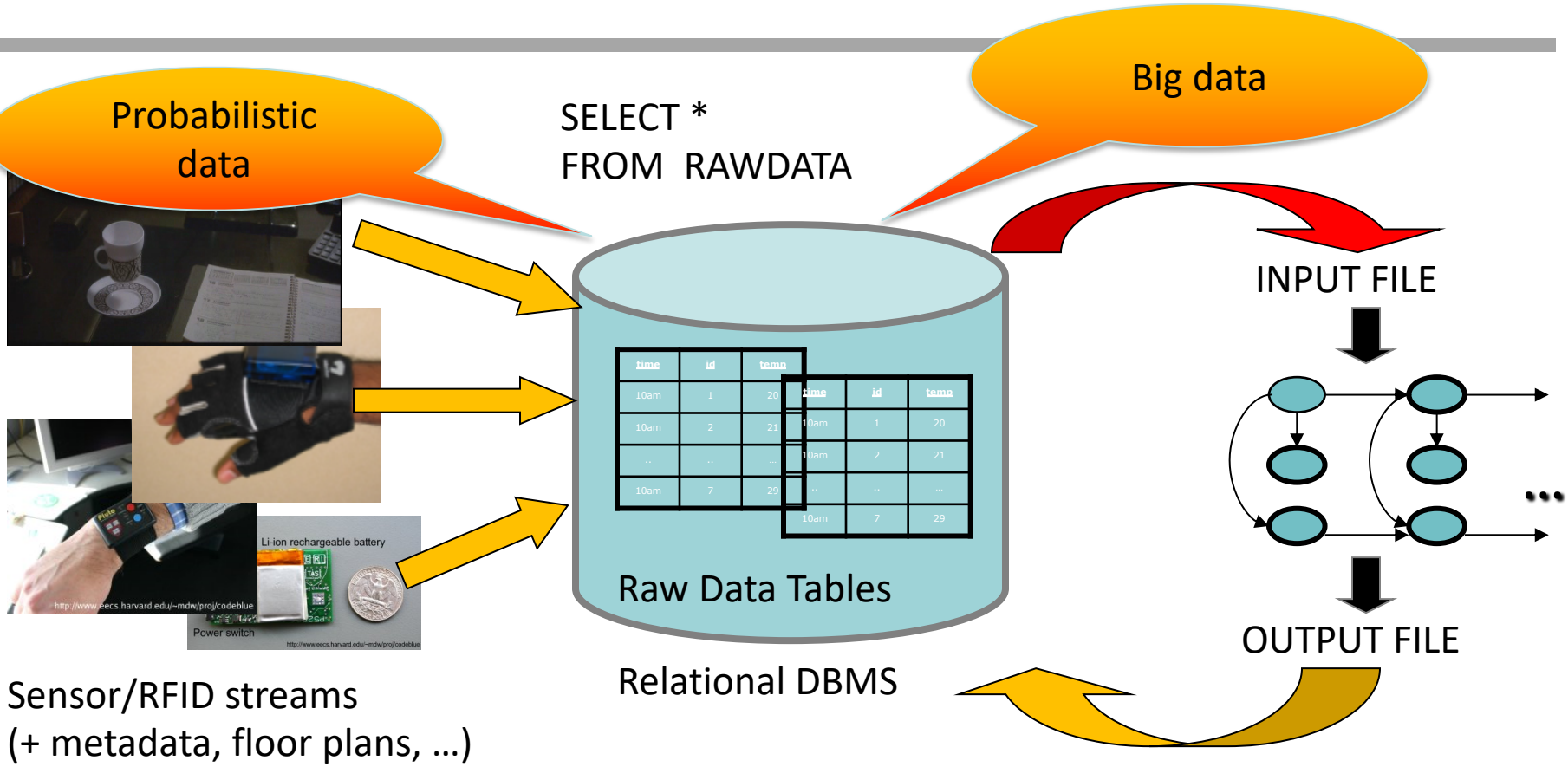
  or
  - Better scoring structure: "structural" EM step

Nir Friedman. The Bayesian structural EM algorithm. In Proceedings of the 14th conference on Uncertainty in artificial intelligence (UAI'98), 129-138. **1998**

Variations on a theme

- **Known structure, fully observable**: only need to do parameter estimation
- **Known structure, hidden variables:** use expectation maximization (EM) to estimate parameters
- **Unknown structure, fully observable:** do heuristic search through structure space, then parameter estimation
- **Unknown structure, hidden variables:** structural EM

Big Datasets Destroy Simple Abstractions



Managing Large, Uncertain Data
Repositories with Probabilistic Graphical Models
Daisy Zhe Wang+, Eirinaios Michelakis+, Minos Garofalakis*+, Joseph M. Hellerstein+ University of California Berkeley+, Yahoo! Research*
25th August **2008**, VLDB

HeisenData -- Towards a Next-Generation Uncertain Data Management System
Project Funding: FP7 Marie-Curie International Reintegration Grants (FP7-PEOPLE-2009-RG, Reference No. 249217)
Project Duration: **1/3/2010 - 28/2/2014**
Fellow & Scientist-in-Charge: Prof. Minos Garofalakis

Daisy Zhe Wang, Eirinaios Michelakis, Minos Garofalakis, and Joseph M. Hellerstein. "**BAYESSTORE**: Managing Large, Uncertain Data Repositories with Probabilistic Graphical Models", Proceedings of VLDB'2008 (PVLDB, Vol. 1), **2008**.

Non-Standard-Datenbanken

Bayessche Netze, Infererenz, Lernen, Relationale Erweiterung