Manual Building of Bayes Networks
Manual creation of a reasoning system based on a graphical model:

- causal model of given domain
- conditional independence graph
- decomposition of the distribution
- evidence propagation scheme

heuristics!
formally provable
formally provable

Problem: strong assumptions about the statistical effects of causal relations. Nevertheless this approach often yields usable graphical models.
Example 1: Genotype Determination of Danish Jersey Cattle

Assumptions about parents:
- risk about misstatement

Genotype mother (dam)  Genotype father (sire)

Genotype child:
- 6 possible values

4 lysis values measured by photometer

Reliability of databases
Inheritance rules
Blood group determination

See paper on our website.
Danish Jersey Cattle Blood Type Determination

21 attributes:

1 – dam correct?
2 – sire correct?
3 – stated dam ph.gr. 1
4 – stated dam ph.gr. 2
5 – stated sire ph.gr. 1
6 – stated sire ph.gr. 2
7 – true dam ph.gr. 1
8 – true dam ph.gr. 2
9 – true sire ph.gr. 1
10 – true sire ph.gr. 2
11 – offspring ph.gr. 1
12 – offspring ph.gr. 2
13 – offspring genotype
14 – factor 40
15 – factor 41
16 – factor 42
17 – factor 43
18 – lysis 40
19 – lysis 41
20 – lysis 42
21 – lysis 43

The grey nodes correspond to observable attributes.

This graph was specified by human domain experts, based on knowledge about (causal) dependences of the variables.
Example 1: Genotype Determination of Danish Jersey Cattle

Full 21-dimensional domain has $2^6 \cdot 3^{10} \cdot 6 \cdot 8^4 = 92,876,046,336$ possible states.

Bayesian network requires only 306 conditional probabilities.

Example of a conditional probability table (attributes 2, 9, and 5):

<table>
<thead>
<tr>
<th>sire correct</th>
<th>true sire</th>
<th>stated sire phenogroup 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>phenogroup 1</td>
<td>F1</td>
</tr>
<tr>
<td>yes</td>
<td>F1</td>
<td>1</td>
</tr>
<tr>
<td>yes</td>
<td>V1</td>
<td>0</td>
</tr>
<tr>
<td>yes</td>
<td>V2</td>
<td>0</td>
</tr>
<tr>
<td>no</td>
<td>F1</td>
<td>0.58</td>
</tr>
<tr>
<td>no</td>
<td>V1</td>
<td>0.58</td>
</tr>
<tr>
<td>no</td>
<td>V2</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The probabilities are acquired from human domain experts or estimated from historical data.
Example 1: Genotype Determination of Danish Jersey Cattle

moral graph
(already triangulated)

join tree
Example 1: Genotype Determination of Danish Jersey Cattle

Marginal distributions before setting evidence:
Conditional distributions given evidence in the input variables:
Example 2: Item Planning at Volkswagen

**Strategy of the VW Group**

<table>
<thead>
<tr>
<th>Marketing strategy</th>
<th>Vehicle specification by clients</th>
<th>Bestsellers defined by manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>Huge number of variants</td>
<td>Small number of variants</td>
</tr>
</tbody>
</table>

**Vehicle specification**

<table>
<thead>
<tr>
<th>Equipment</th>
<th>fastback</th>
<th>2,81, 150 kW</th>
<th>Type Alpha</th>
<th>4</th>
<th>leather</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>car body type</td>
<td>engine</td>
<td>radio</td>
<td>doors</td>
<td>seat cover</td>
<td>...</td>
</tr>
</tbody>
</table>
Example 2: Model “Golf”

Approx. 200 equipment groups

2 to 50 items per group

Therefore more than $2^{200}$ possible vehicle specifications

Choice of valid specifications is constrained by a rule system
(10000 technical rules, plus marketing and production rules)

Example of technical rules:

If Engine = $e_1$ then Transmission = $t_3$

If Engine = $e_4$ and Heating = $h_2$ then Generator $\in \{g_3, g_4, g_5\}$
Problem Representation

**Historical Data**
Sample of produced *vehicle specifications*
(representative choice, context-dependent, e.g. Golf)

**System of Rules**
*Rules* for the validity of item combinations
(specified for a vehicle class and a planning interval)

**Prediction & Planning**
Predicted / assigned *planning data*
(production program, demands, installation rates, capacity restrictions, ...)

Rudolf Kruse, Alexander Dockhorn
Bayesian Networks
**Complexity of the Planning Problem**

**Equipment table**

<table>
<thead>
<tr>
<th></th>
<th>Engine</th>
<th>Transmission</th>
<th>Heating</th>
<th>Generator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$e_1$</td>
<td>$t_3$</td>
<td>$h_1$</td>
<td>$g_1$</td>
</tr>
<tr>
<td>2</td>
<td>$e_2$</td>
<td>$t_4$</td>
<td>$h_3$</td>
<td>$g_5$</td>
</tr>
<tr>
<td></td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>100000</td>
<td>$e_7$</td>
<td>$t_1$</td>
<td>$h_3$</td>
<td>$g_2$</td>
</tr>
</tbody>
</table>

**Installation rates**

<table>
<thead>
<tr>
<th>Engine</th>
<th>Transmission</th>
<th>Heating</th>
<th>Generator</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>$t_1$</td>
<td>$h_1$</td>
<td>$g_1$</td>
<td>0.0000012</td>
</tr>
<tr>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

Result is a 200-dimensional, finite probability space

\[
P(\text{Engine} = e_1, \text{Transmission} = t_3) = ?
\]

\[
P(\text{Heating} = h_1 \mid \text{Generator} = g_3) = ? \quad \text{Problem of complexity!}
\]
Solution: Decomposition into Subspaces

\[ P(E, H, T, A) = P(A | E, H, T) \cdot P(T | E, H) \cdot P(E | H) \cdot P(H) \]

\[ \text{here} \quad P(A | E, H) \cdot P(T | E) \cdot P(E) \cdot P(H) \]

Bayesian Network

Hypergraph Decomposition
Typical Planning Operation: Focusing

**Application:**
- **Compute item demand**
  Calculation of installation rates of equipment combinations
- **Simulation**
  Analyze customer requirements (e.g. of persons having ordered a navigation system for a VW Polo)

**Input:** Equipment combinations

**Operation:** Compute
- the conditional network distribution and
- the probabilities of the specified equipment combinations.
Implementation and Deployment

Project leader: Intelligent System Consulting (Gebhardt)

Client server system
Server on 6–8 machines
Quadcore platform
Terabyte hard drive
Java, Linux, Oracle
WebSphere application server
Software used daily worldwide
20 developers
5000 Bayesian networks are currently used