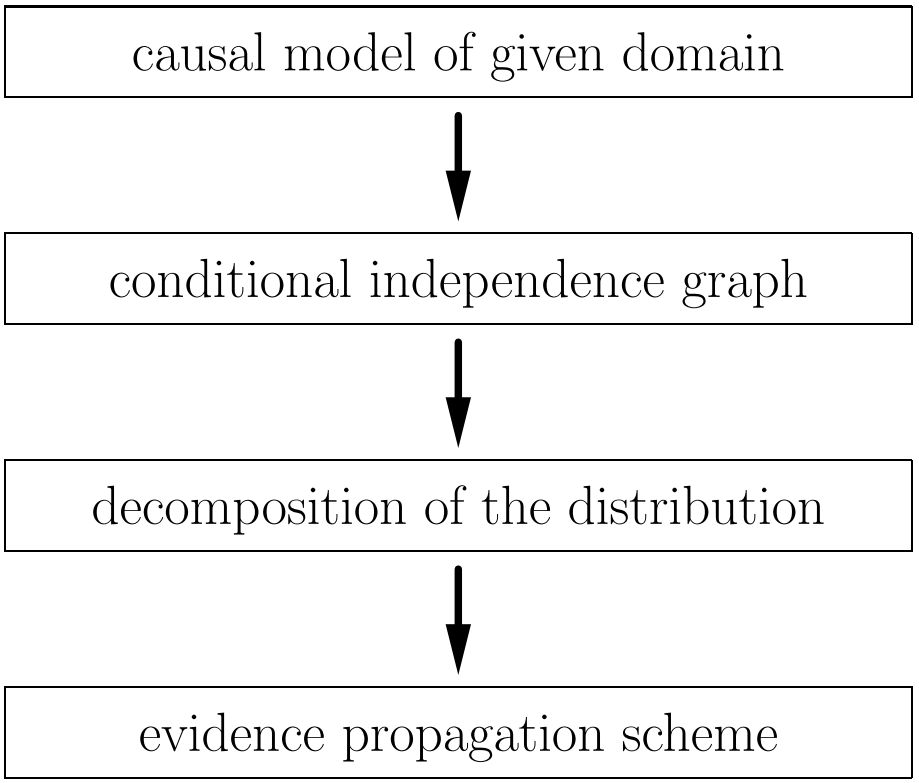
Manual Building of Bayes Networks

Building Graphical Models: Causal Modeling

Manual creation of a reasoning system based on a graphical model:



heuristics!

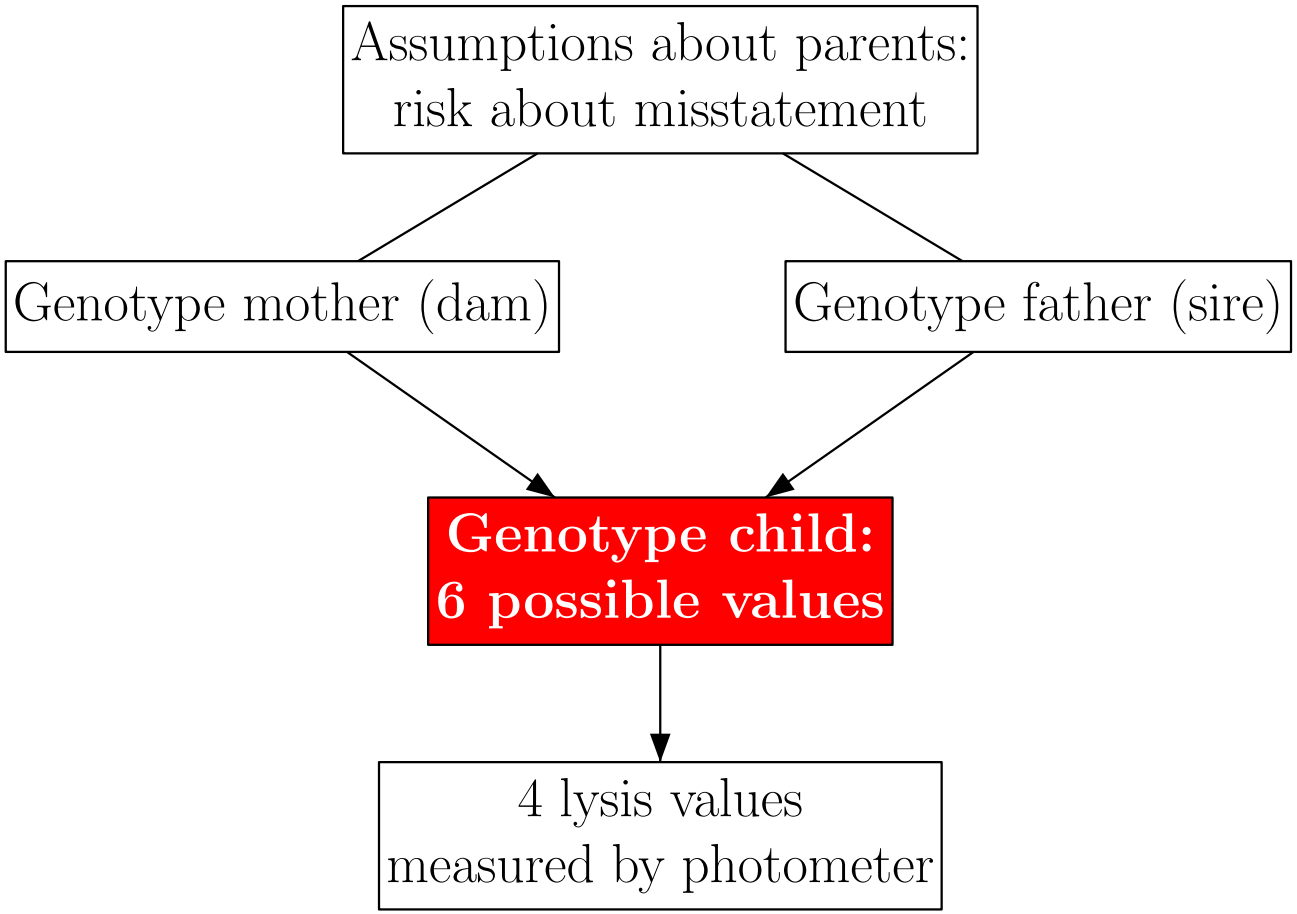
formally provable

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



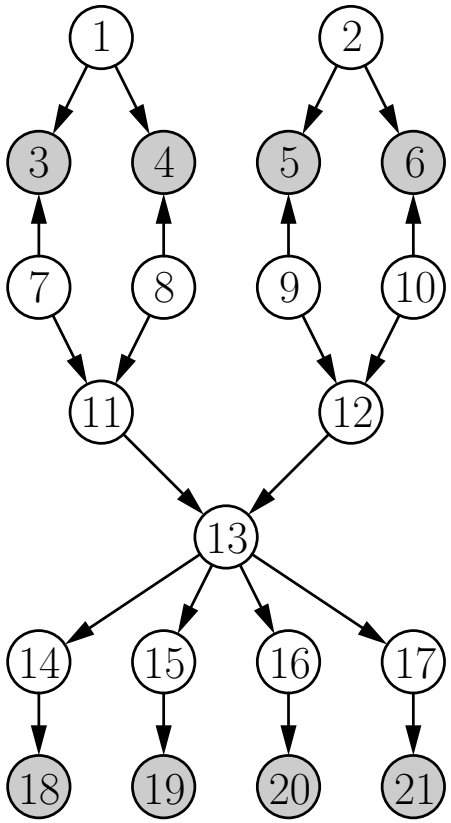
Reliability of databases

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.

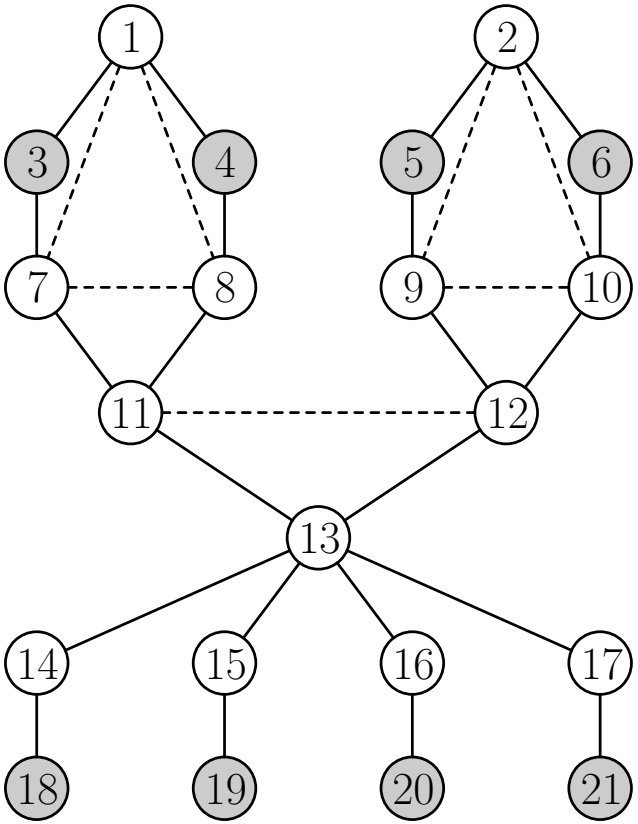
Bayesian network requires only 306 conditional probabilities.

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

sire correct	true sire phenogroup 1	stated sire phenogroup 1 F1	V1	V2
yes	F1	1	0	0
yes	V1	0	1	0
yes	V2	0	0	1
no	F1	0.58	0.10	0.32
no	V1	0.58	0.10	0.32
no	V2	0.58	0.10	0.32

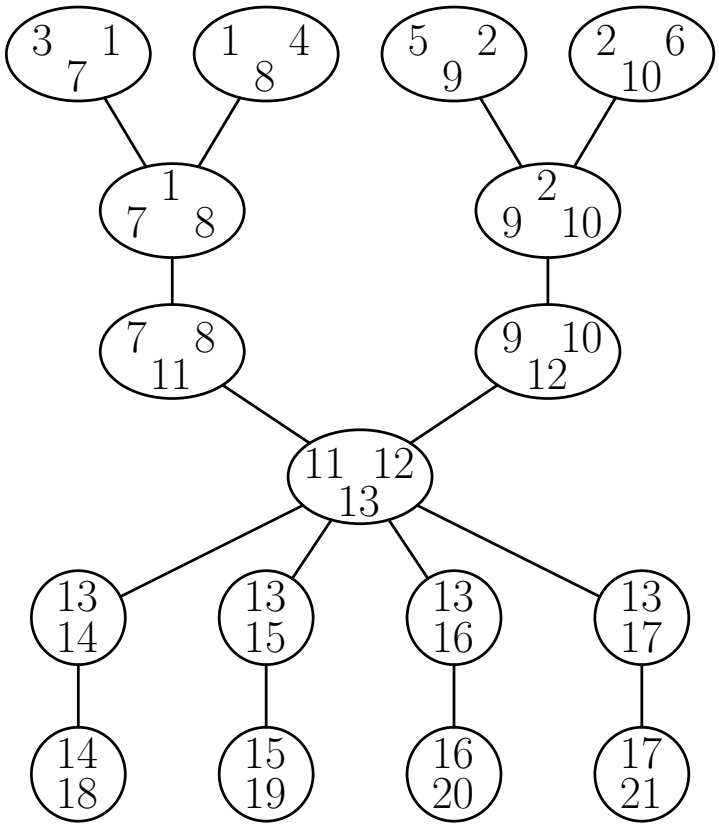
The probabilities are acquired from human domain experts
or estimated from historical data.

Example 1: Genotype Determination of Danish Jersey Cattle



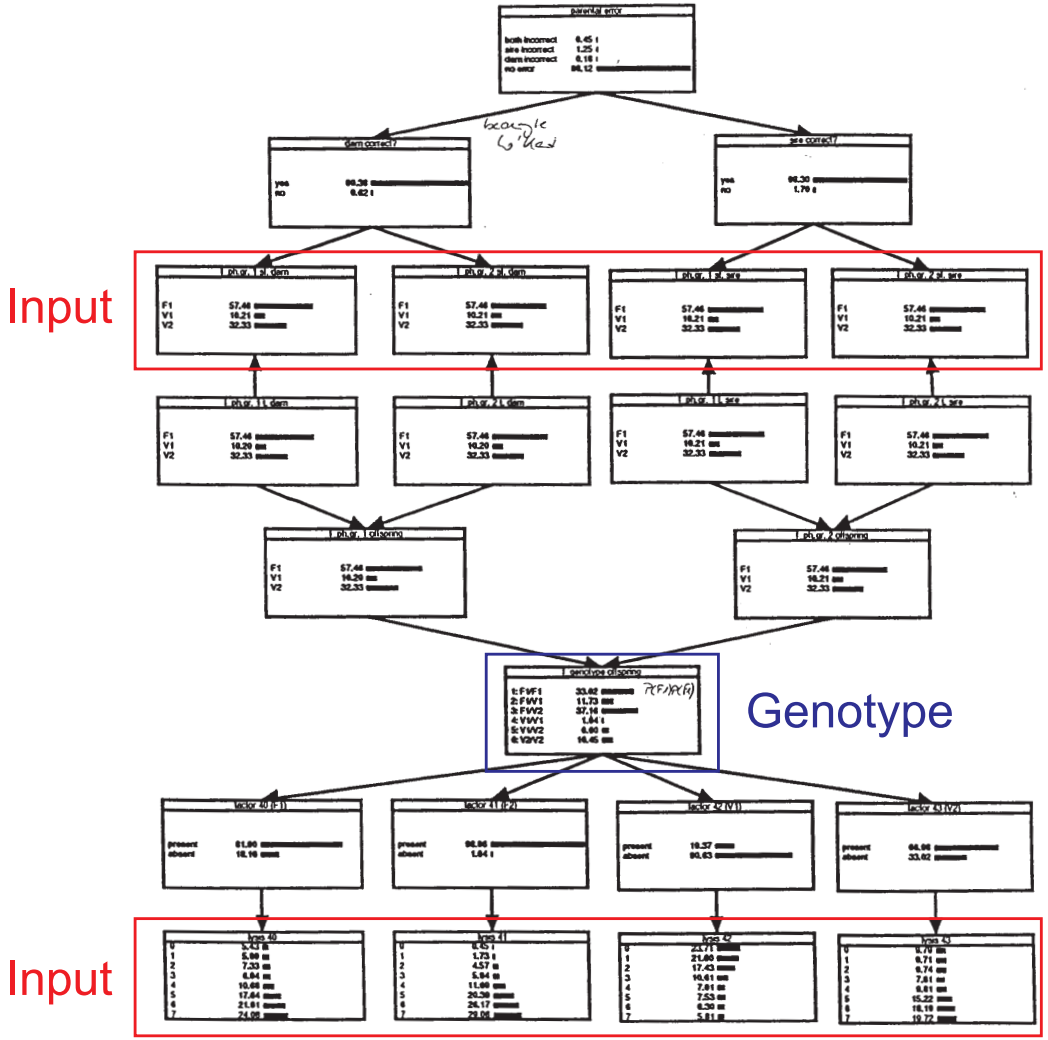
moral graph

(already triangulated)

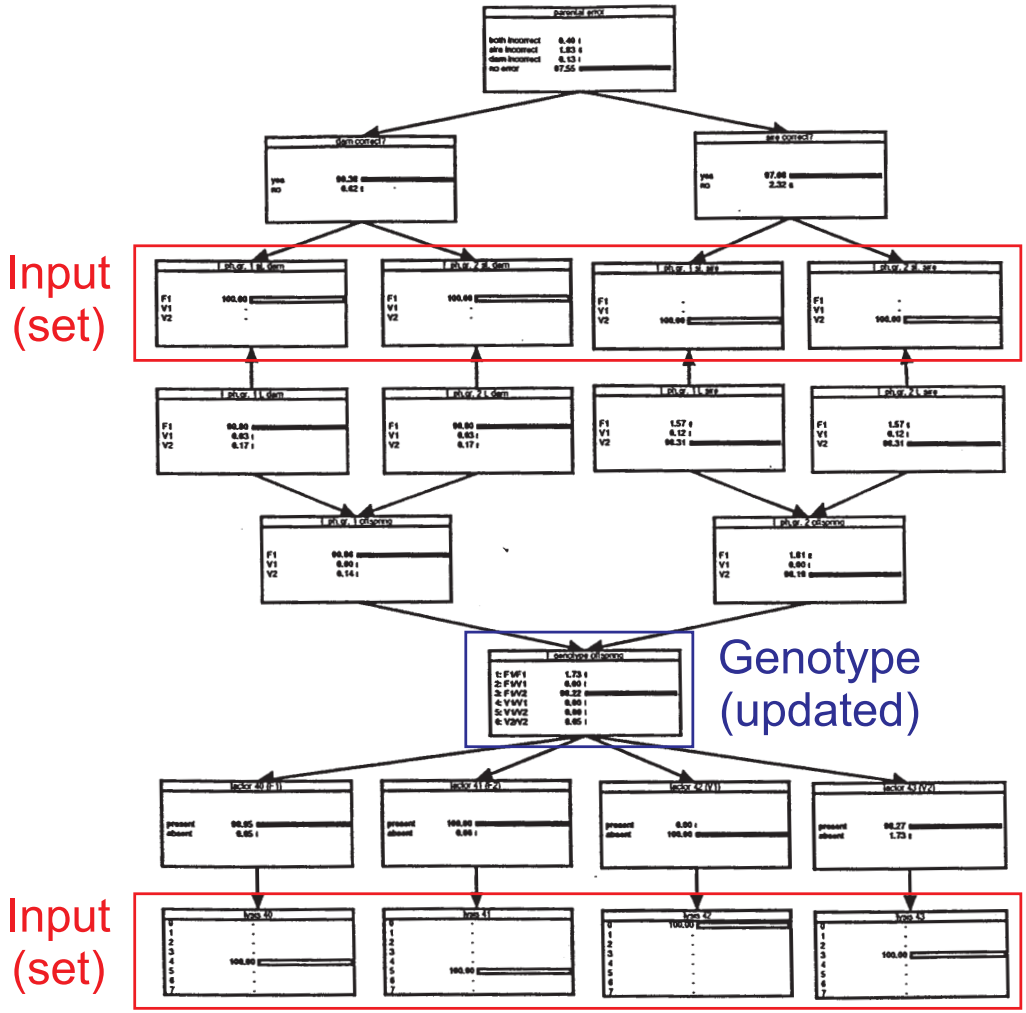


join tree

Marginal distributions before setting evidence:



Conditional distributions given evidence in the input variables:



Example 2: Item Planning at Volkswagen

Strategy of the VW Group

Marketing strategy	Vehicle specification by clients	Bestsellers defined by manufacturer
Complexity	Huge number of variants	Small number of variants

Vehicle specification

Equipment	fastback	2,8 l, 150 kW	Type Alpha	4	leather	...
Group	car body type	engine	radio	doors	seat cover	...

Example 2: Model "Golf"

Approx. 200 equipment groups

2 to 50 items per group

Therefore more than 2^{200} possible vehicle specifications

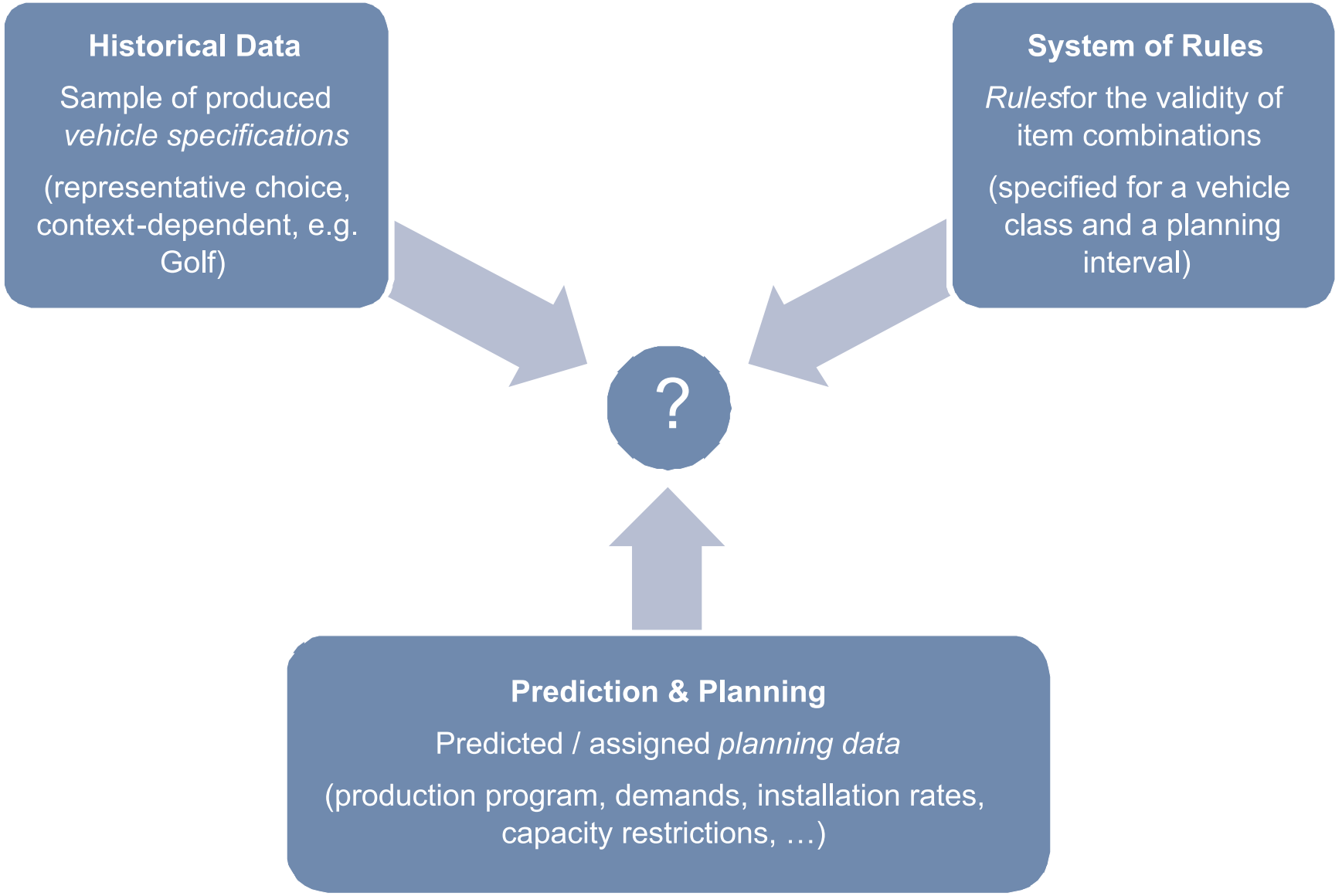
Choice of valid specifications is constrained by a rule system
(10000 technical rules, plus marketing and production rules)

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



Complexity of the Planning Problem

Equipment table

	Engine	Transmission	Heating	Generator	$\cdots$
1	e_1	t_3	h_1	g_1	$\cdots$
2	e_2	t_4	h_3	g_5	$\cdots$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
100000	e_7	t_1	h_3	g_2	$\cdots$

Installation rates

Engine	Transmission	Heating	Generator	$\cdots$	Rate
e_1	t_1	h_1	g_1	$\cdots$	0.0000012
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

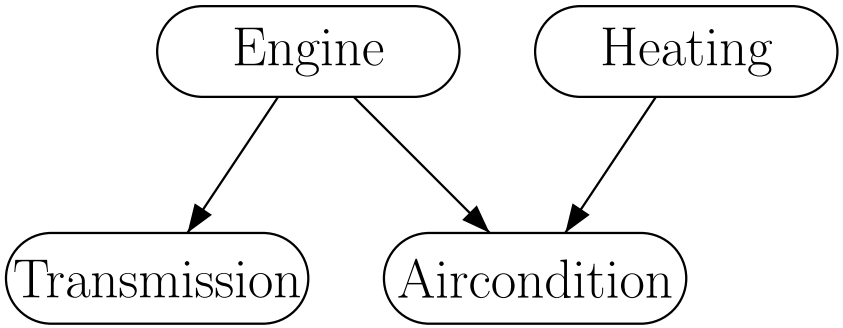
Result is a 200-dimensional, finite probability space

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

$$P(\text{Heating} = h_1 \mid \text{Generator} = g_3) = ?$$

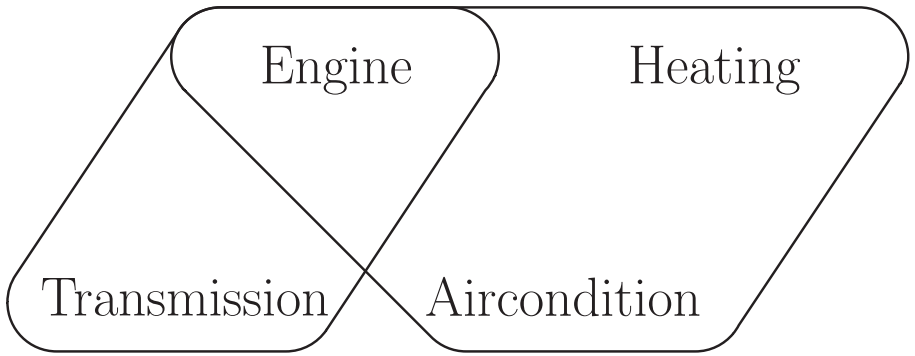
Problem of complexity!

Solution: Decomposition into Subspaces



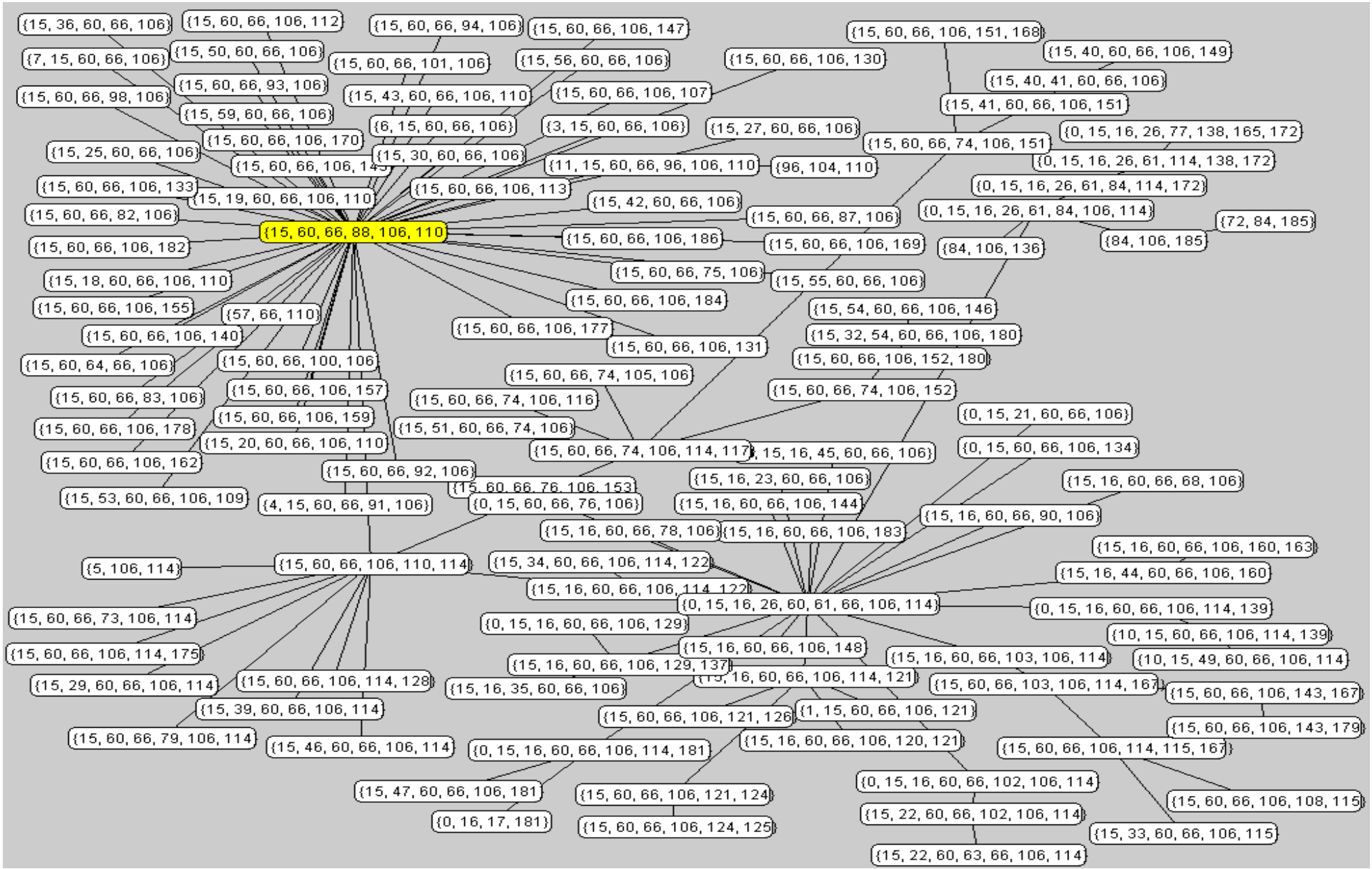
Bayesian Network

$$
\begin{aligned}
P(E, H, T, A) &= P(A \mid E, H, T) \cdot P(T \mid E, H) \cdot P(E \mid H) \cdot P(H) \\
&\overset{\text{here}}{=} P(A \mid E, H) \cdot P(T \mid E) \cdot P(E) \cdot P(H)
\end{aligned}
$$



Hypergraph Decomposition

Clique Tree of the VW Bora



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 maschines

Quadcore platform

Terabyte hard drive

Java, Linux, Oracle

WebSphere application server

Software used daily worldwide

20 developers

5000 Bayesian networks are currently used