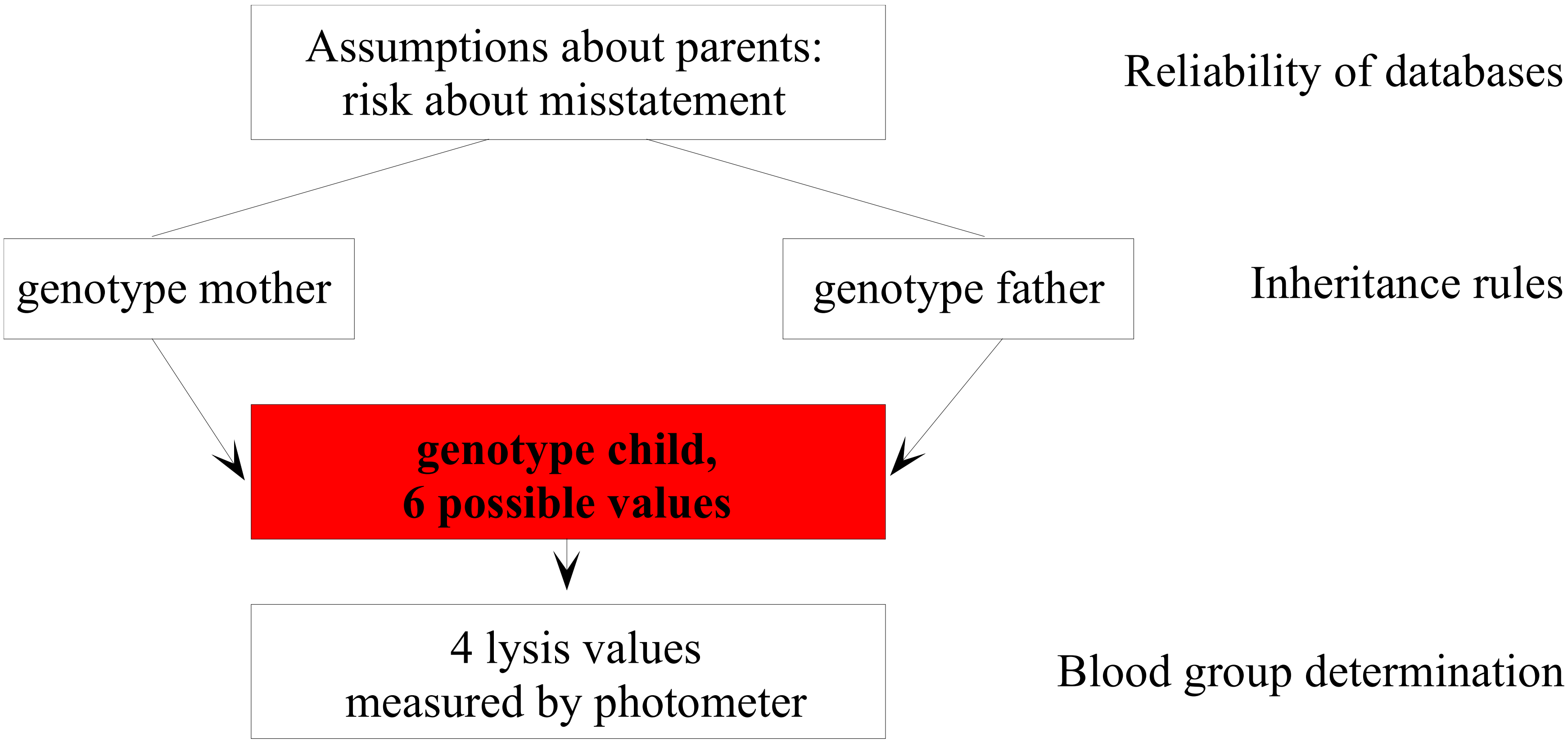
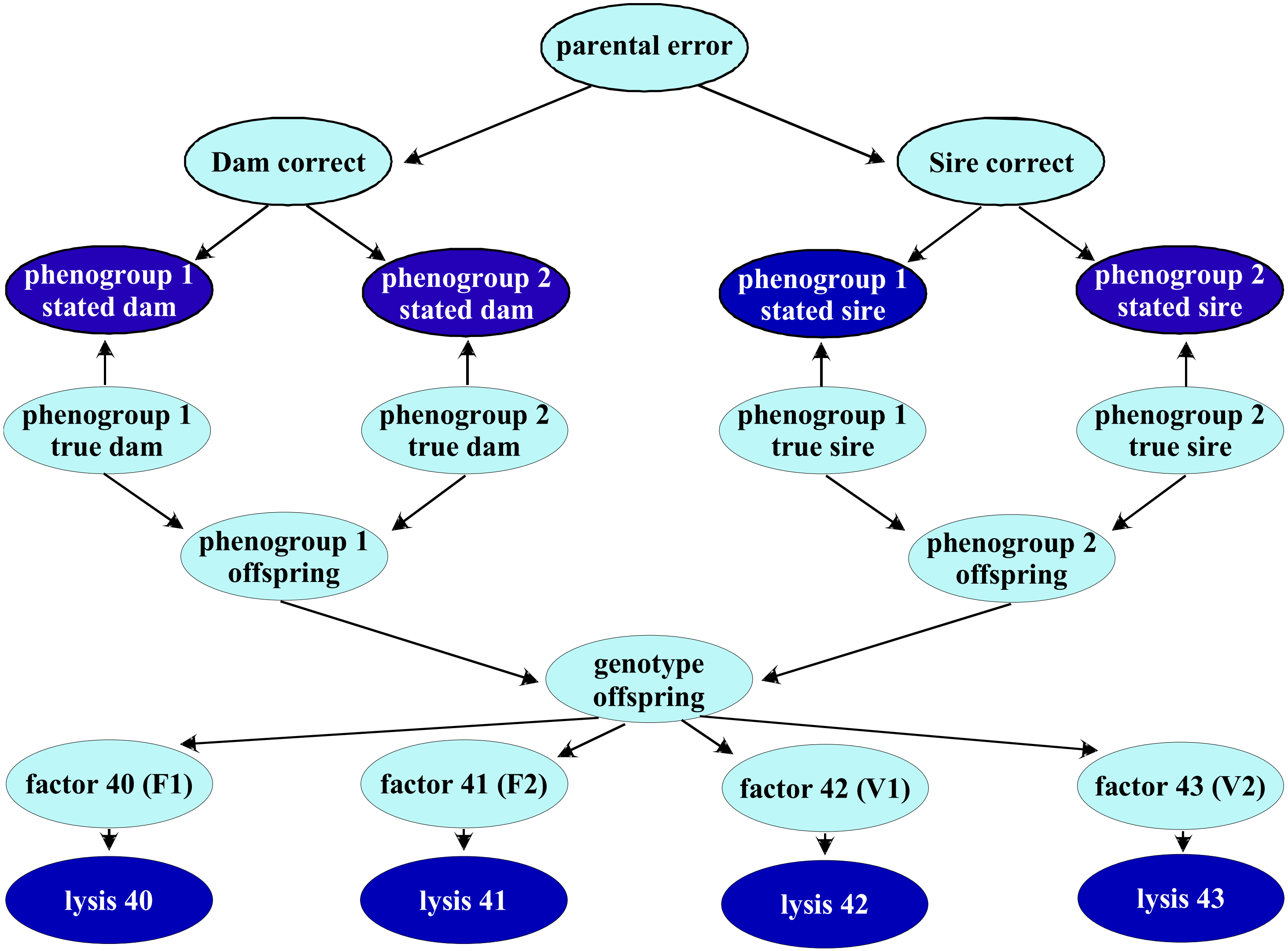
Genotype Determination of Danish Jersey Cattle

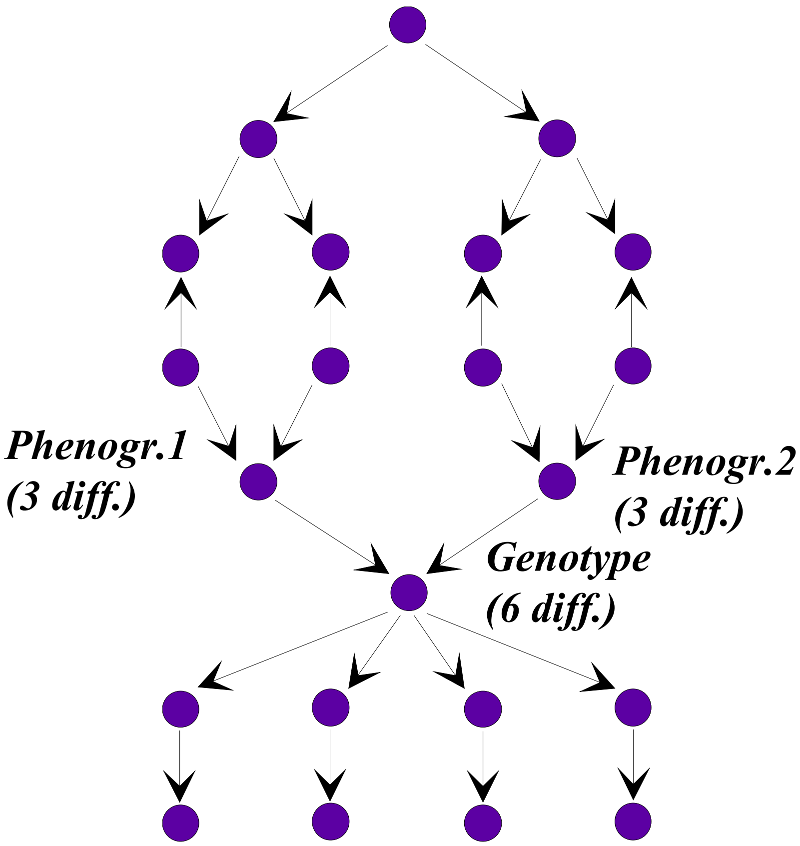


Qualitative Knowledge



Example: Genotype Determination of Jersey Cattle

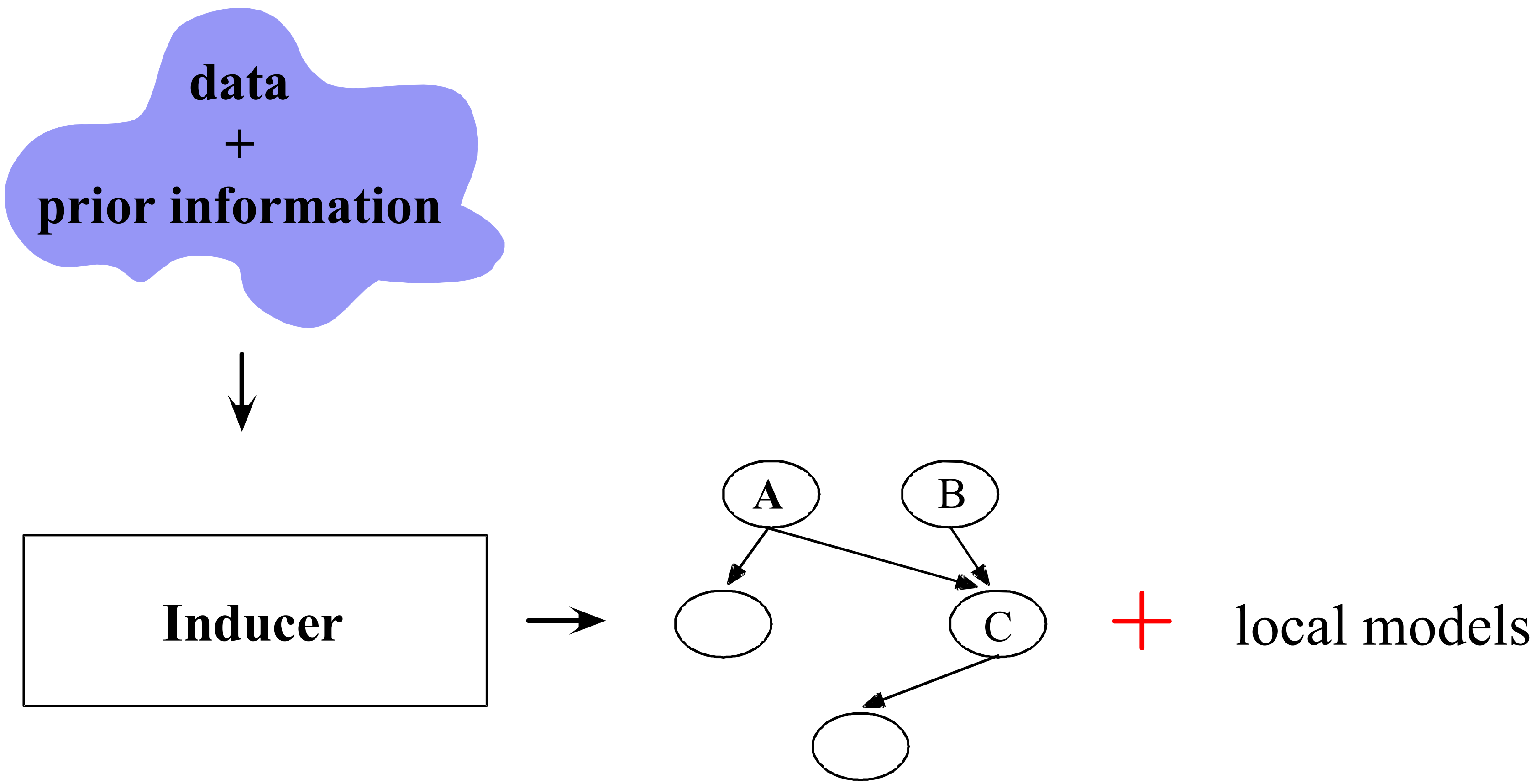
variables: 22, state space $6 \cdot 10^{13}$, parameters: 324



Graphical Model

- node
  → random variable
- edges
  → conditional dependencies
- decomposition
  → $P(X_1, \ldots, X_{22}) = \prod_{i=1}^{22} P(X_i \mid \text{parents}(X_i))$
- diagnosis
  → P(· | knowledge)

Learning Graphical Models



Genotype Determination of Danish Jersey Cattle: Database of Cases

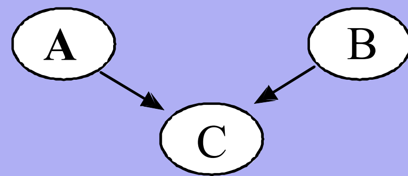
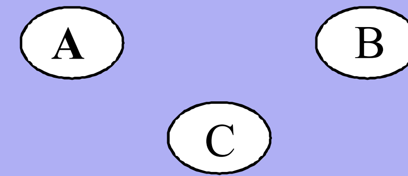
747 cases
22 entries per case

Case 657:

(0, 6, 5, 5,	*f*1, *v*2, *f*1, *v*1,	*f*1*v*1,	***f*1 or *v*2,**	*f*1,	*f*1, *v*2, *f*1, *v*1, , , **...**)
lysis factors from 0 to 7 (photometer	observed phenogroups of parents	genotype offspring	phenogroup offspring 1	phenogroup offspring 2	true phenogroups of parents

- ESPRIT Project DRUMS 2, BR 6156
- Problems:
  - How to reduce complexity problems?
  - How to handle imprecise (fuzzy, vague, ...) data?

The Learning Problem

	known structure	unknown structure
		

complete data

A	B	C
<a_4,	b_3,	c_1>
<a_3,	b_2,	c_4>

known structure:

Statistical Parametric Estimation (closed from eq.):

- statistical parameter fitting,
- ML Estimation,
- Bayesian Inference, ...

unknown structure:

Discrete Optimization over Structures (discrete search):

- likelihood scores,
- MDL

Problem:
search complexity → heuristics

incomplete data

(missing values, hidden variables,...)

A	B	C
<a_4,	?,	c_1>
<a_3,	b_2,	?>

known structure:

Parametric Optimization:

- EM,
- gradient descent, ...

unknown structure:

Combined Methods:

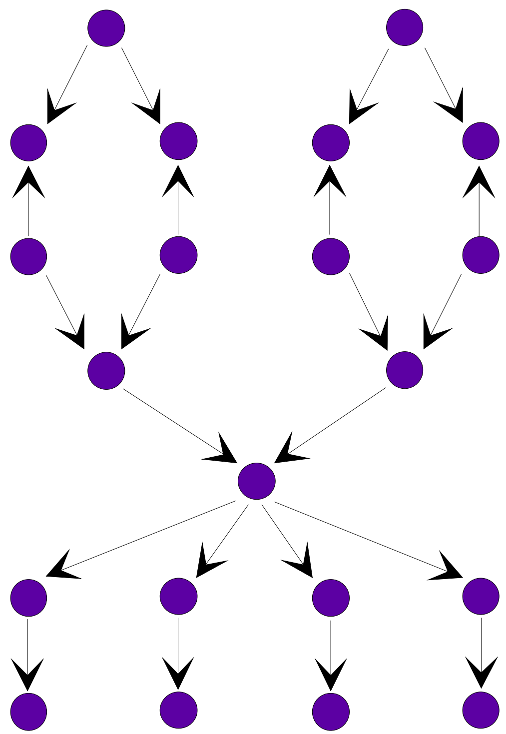
- structured EM
- only few approaches

Problems:

- criterion for fit?
- new variables?
- local maxima?
- fuzzy values?

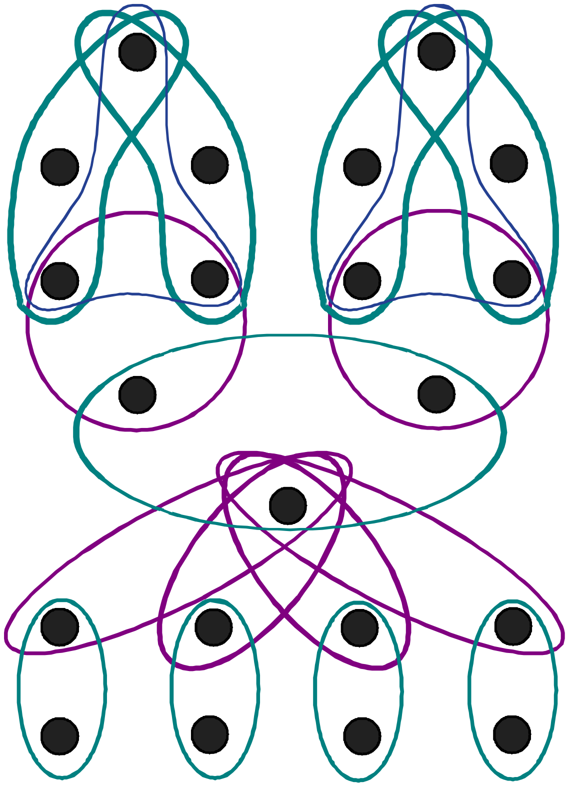
Genotype Determination

Directed dependency network



Rule $\rightarrow$ conditional dependency

Hypergraph representation



Rule $\rightarrow$ constraint

Daimler-Chrysler Research and Technology Ulm, „Data Mining“ Project

Fields of Application

- Improvement of Product Quality by Finding Weaknesses
  - Learn dependency network for vehicle properties and faults
  - Look for unusual conditional fault frequencies
  - Find causes for these unusual frequencies
  - Improve construction of vehicle

- Improvement of Error Diagnosis in Garages
  - Learn dependency network for vehicle properties and faults
  - Record properties of new faulty vehicle
  - Test for the most probable faults

Analysis of Daimler/Chrysler Database

- Database: ~ 18.500 passenger cars
  > 100 attributes per car

- Analysis of dependencies between **special equipment** and **faults**.

- Results used as a starting point for technical experts looking for causes.

General Structure of (most) Learning Algorithms for Graphical Models

- Use a criterion to measure the degree to which a network structure fits the data and the prior knowledge
  (model selection, goodness of hypergraph)

- Use a search algorithm to find a model that receives a high score by the criterion
  (optimal spanning tree, K2: greedy selection of parents, ...)

Measuring the Deviation from an Independent Distribution

Probability- and Information-based Measures

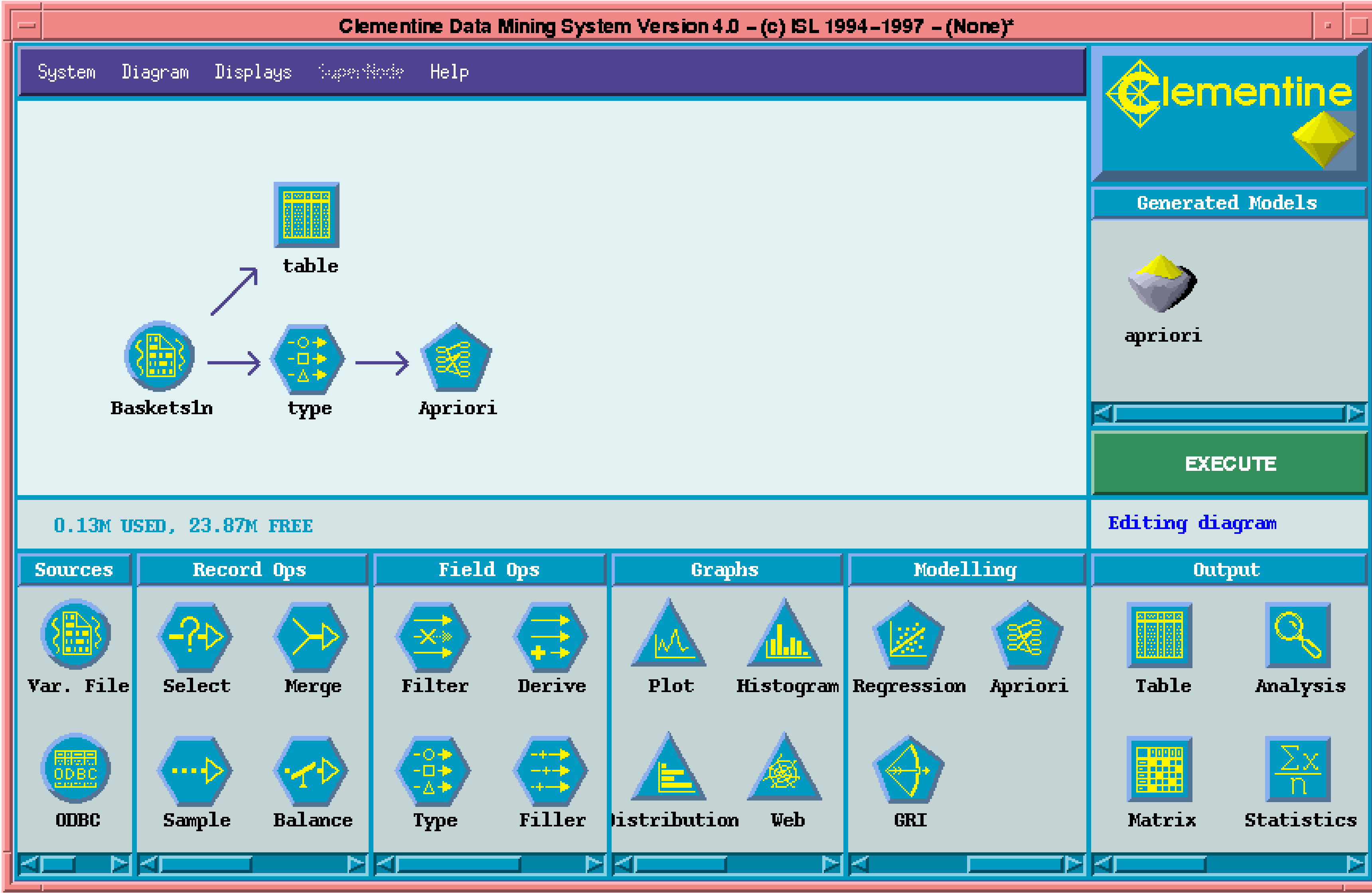
- information gain *
  identical with mutual information
- information gain ratio *
- *g*-function (Cooper and Herskovits)
- minimum description length
- gini index *

Possibilistic Measures

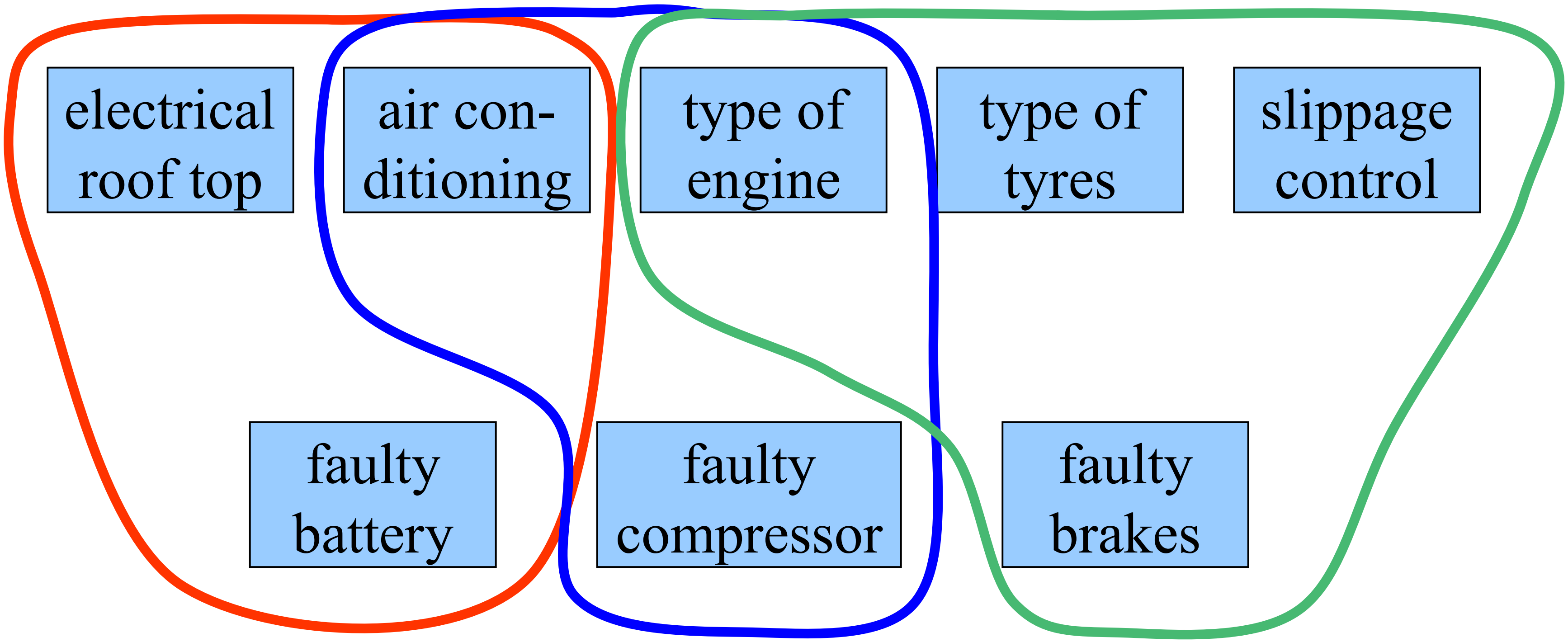
- expected nonspecificity
- specificity gain
- specificity gain ratio

(Measures marked with * originated from decision tree learning)

Data Mining Tool Clementine



Analysis of Daimler/Chrysler Database



Fictituous example:
There are significantly more **faulty batteries**, if both **air conditioning** **and** **electrical roof top** are built into the car.

Example Subnet

Influence of special equipment on battery faults:

(fictitious) frequency of		**air conditioning**	
battery faults		**with**	**without**
electrical sliding roof	**with**	8%	3%
	without	3%	2%

- significant deviation from independent distribution
- hints to possible causes and improvements
- here: larger battery may be required, if an air conditioning system *and* an electrical sliding roof are built in

(The dependencies and frequencies of this example are fictious, true numbers are confidential.)

Data Mining Tool “Information Miner