A Tutorial on Graphical Mode
and How to Learn Them from D

Christian Borgelt

Intelligent Data Analysis and Graphical Models Res
European Center for Soft Computing
c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres (Astu

christian.borgelt@softcomputing.e
http://www.softcomputing.es/
http://www.borgelt.net/

Overview

- **Graphical Models: Core Ideas and Notions**
- **A Simple Example: How does it work in princi**
- **Conditional Independence Graphs**
  - conditional independence and the graphoid axioms
  - separation in (directed and undirected) graphs
  - decomposition/factorization of distributions
- **Evidence Propagation in Graphical Models**
- **Building Graphical Models**
- **Learning Graphical Models from Data**
  - quantitative (parameter) and qualitative (structure) le
  - evaluation measures and search methods
  - learning by measuring the strength of marginal depen
  - learning by conditional independence tests
- **Summary**

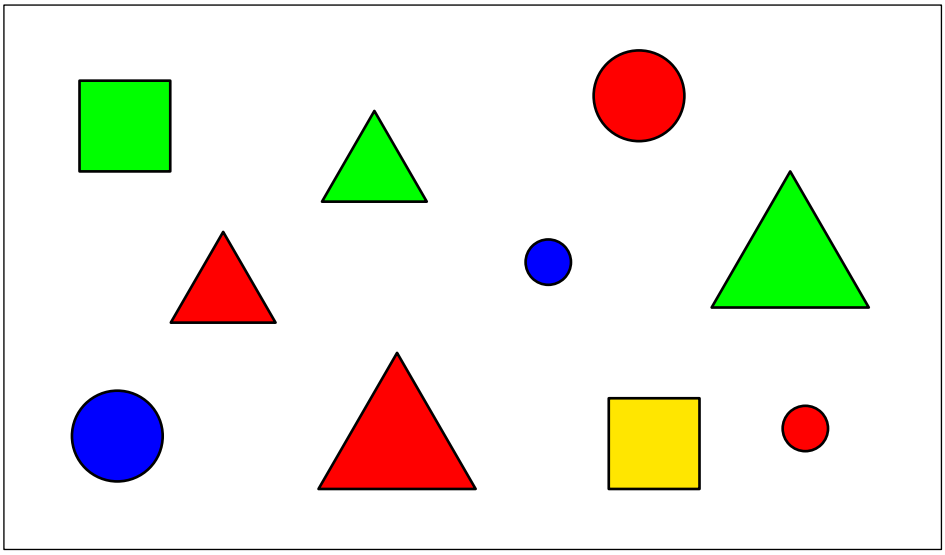
Graphical Models: Core Ideas and M

- **Decomposition:** Under certain conditions a distribution
  distribution) on a multi-dimensional domain, which encod
  knowledge about this domain, can be decomposed into a
  (usually overlapping) distributions on lower-dimensional s
- **Simplified Reasoning:** If such a decomposition is pos
  to know the distributions on the subspaces to draw all infe
  under consideration that can be drawn using the original
- Such a decomposition can nicely be represented as a **gra**
  graph theory), and therefore it is called a **Graphical M**
- The graphical representation
  - encodes **conditional independences** that hold in
  - describes a **factorization** of the probability distribut
  - indicates how **evidence propagation** has to be car

A Simple Example: The Relational Case

A Simple Example

Example Domain

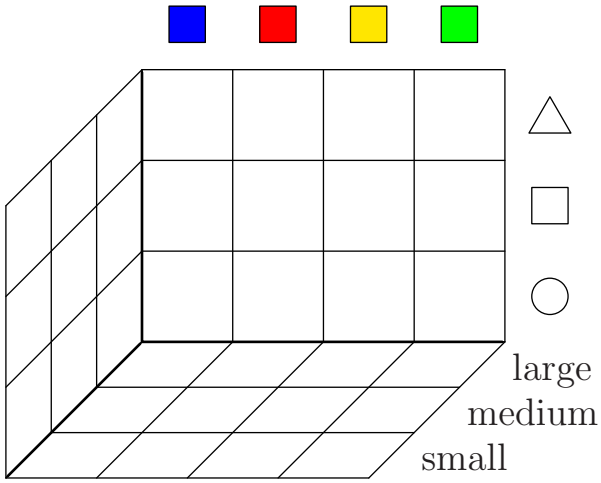


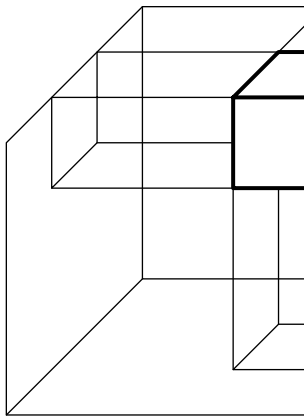
Relati

color

- 10 simple geometrical objects, 3 attributes.
- One object is chosen at random and examined.
- Inferences are drawn about the unobserved attributes.

The Reasoning Space





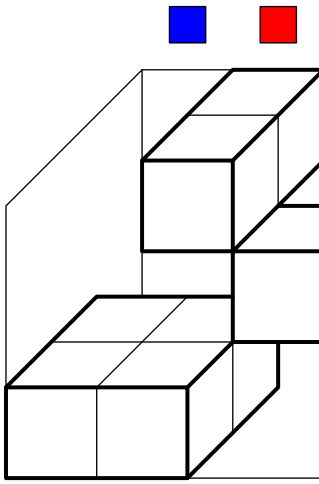
- The reasoning space consists of a finite set Ω of states.
- The states are described by a set of n attributes A_i, $i = 1$ whose domains $\{a_1^{(i)}, \ldots, a_{n_i}^{(i)}\}$ can be seen as sets of prop
- The events in a domain are mutually exclusive and exhau
- The reasoning space is assumed to contain the true, but u

The Relation in the Reasoning S

Relation

color	shape	size
■	○	small
■	○	medium
■	○	small
■	○	medium
■	△	medium
■	△	large
■	□	medium
■	□	medium
■	△	medium
■	△	large

Relation in the Rea

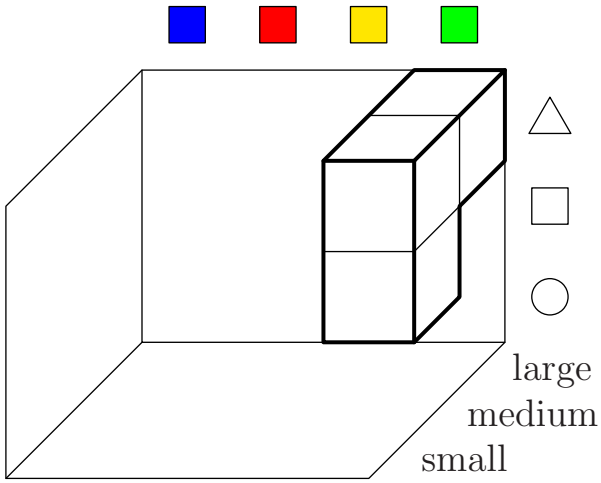


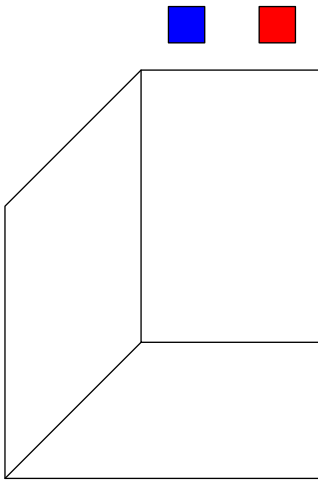
Each cube repres

- The spatial representation helps to understand the decom

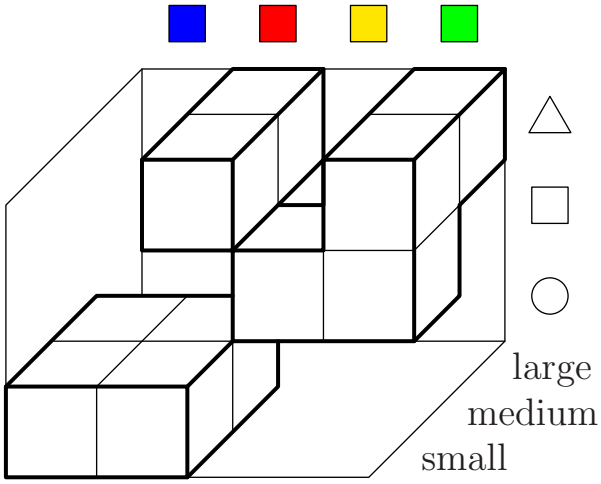
Reasoning

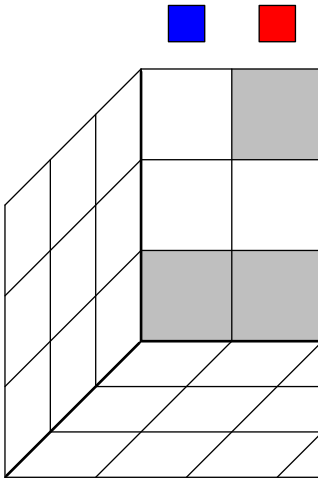
- Let it be known (e.g. from an observation) that the given
  This information considerably reduces the space of possible
- From the prior knowledge it follows that the given object
  - either a triangle or a square and
  - either medium or large.

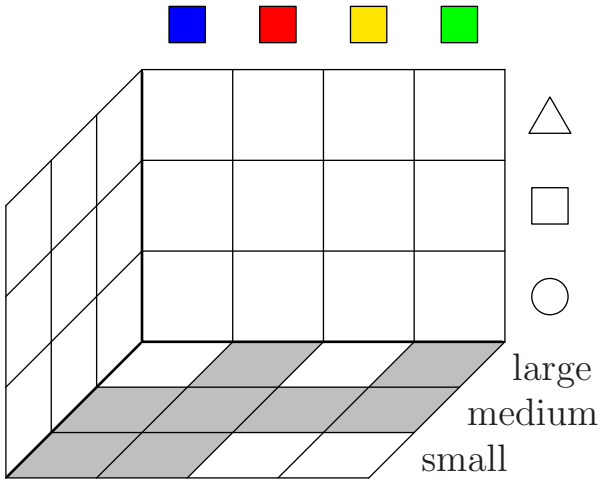


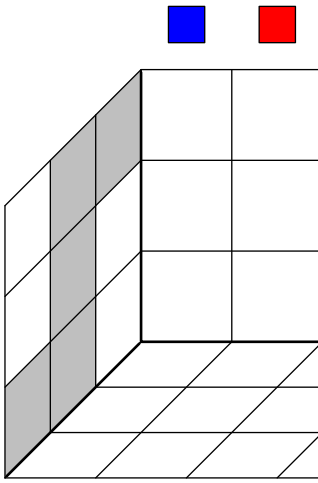


Prior Knowledge and Its Project

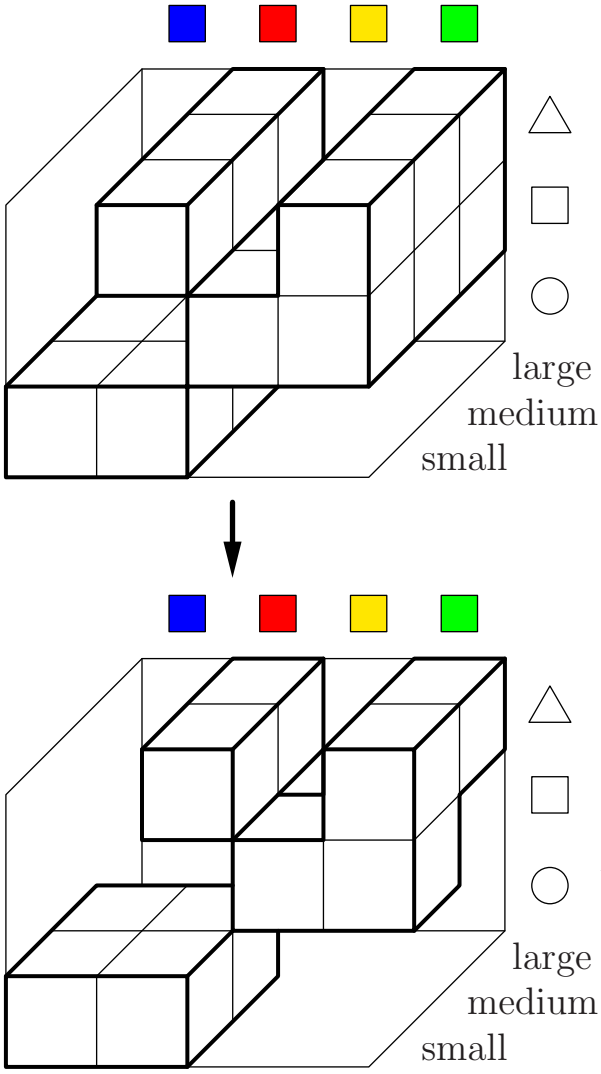








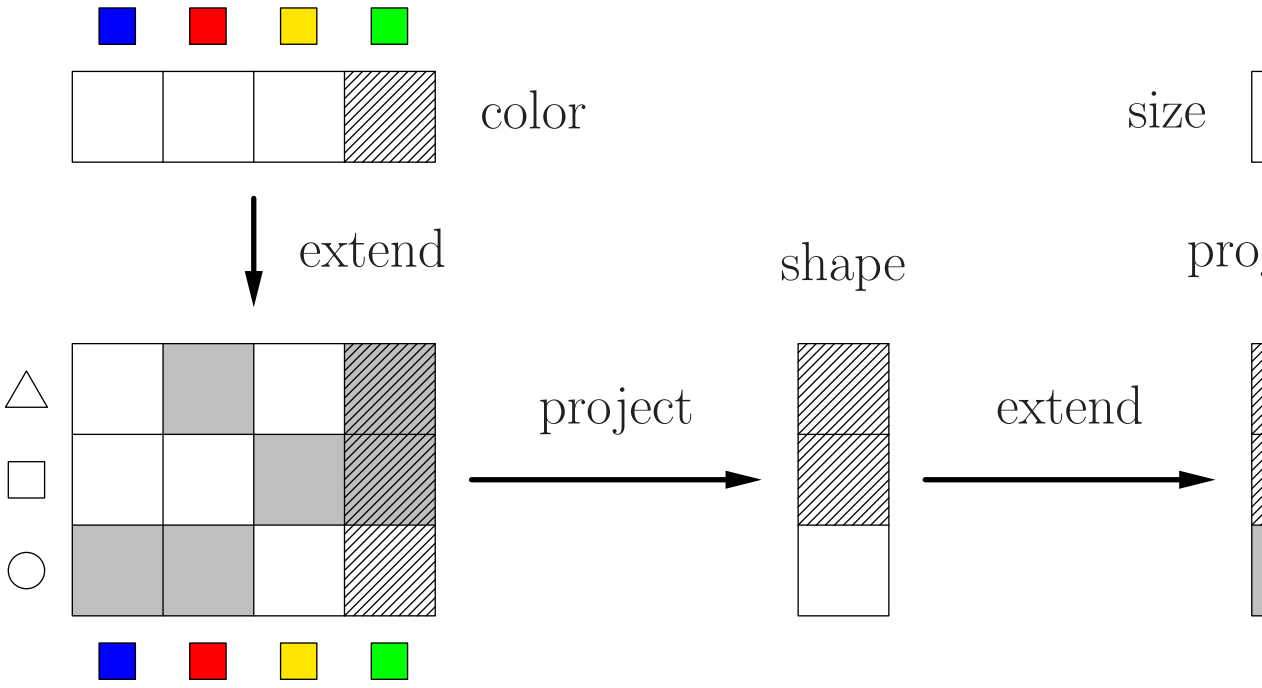
Cylindrical Extensions and Their Int



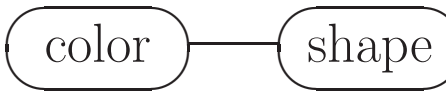
Intersecting the
tensions of the
the subspace spa
and shape and
tion to the subspa
shape and size y
nal three-dimensi

Reasoning with Projections

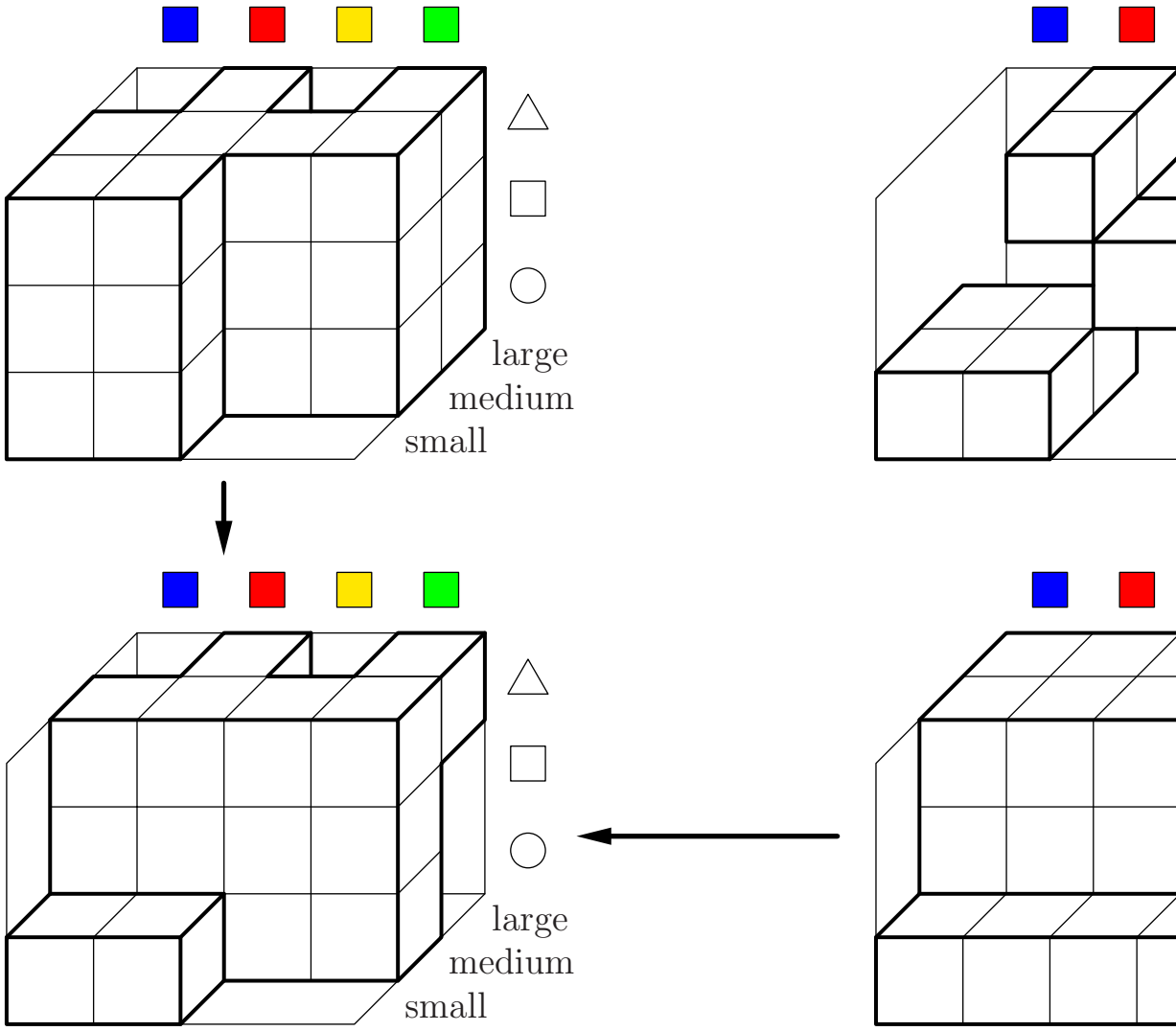
The same result can be obtained using only the projections to
without reconstructing the original three-dimensional relation:



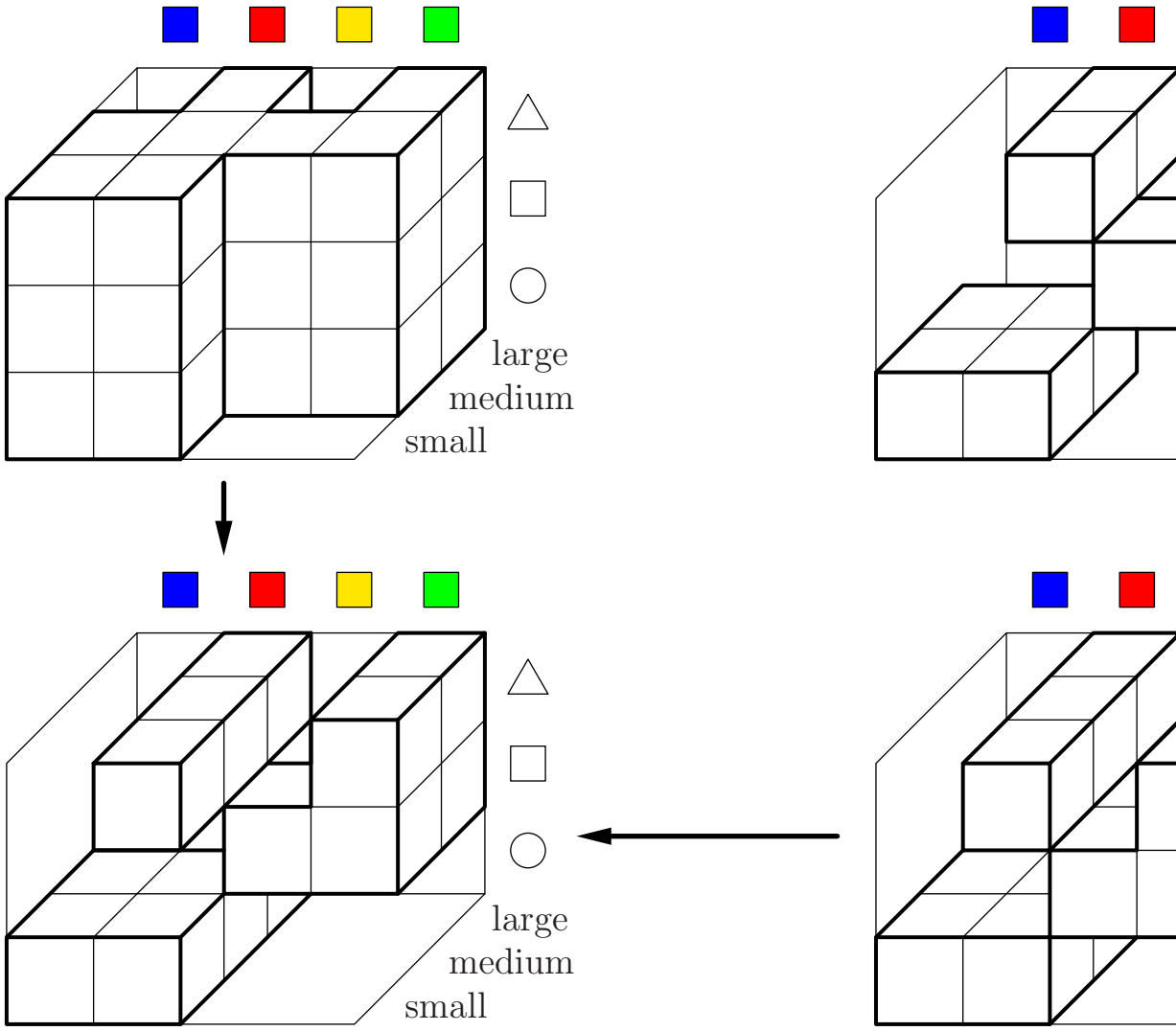
This justifies a graph representation:



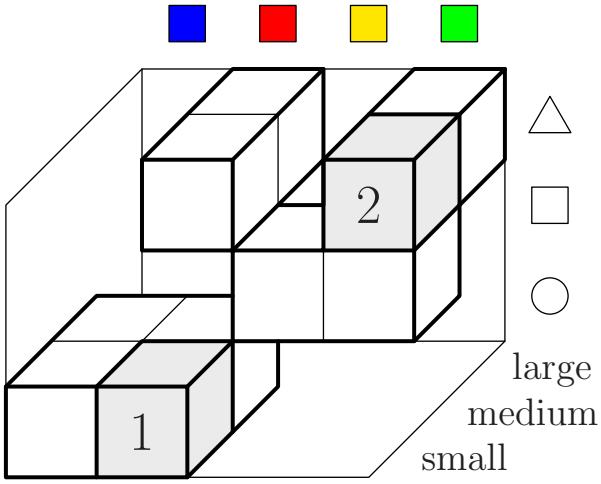
Using other Projections 1

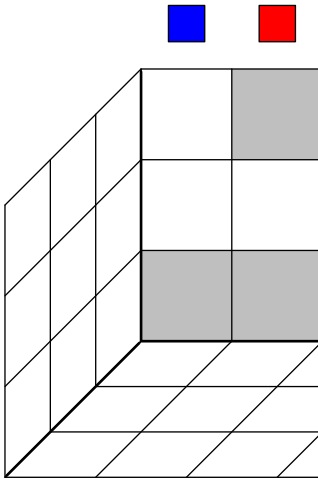


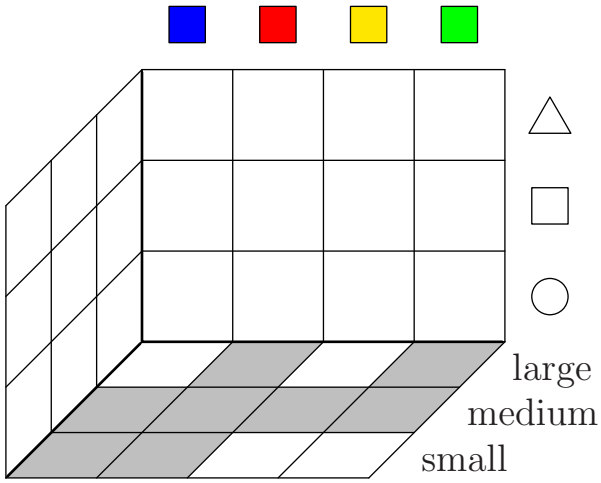
Using other Projections 2

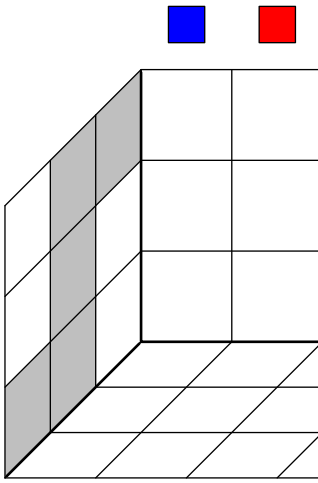


Is Decomposition Always Possib









Relational Graphical Models: Formalization

Possibility-Based Formalizatio

Definition: Let Ω be a (finite) sample space.
A **discrete possibility measure** R on Ω is a function $R : 2^{\Omega}$

1. $R(\emptyset) = 0$ and
2. $\forall E_1, E_2 \subseteq \Omega : R(E_1 \cup E_2) = \max\{R(E_1), R(E_2)\}$.

- Similar to Kolmogorov's axioms of probability theory.
- If an event E can occur (if it is possible), then $R(E) = 1$, otherwise (if E cannot occur/is impossible) then $R(E) =$
- $R(\Omega) = 1$ is not required, because this would exclude the
- From the axioms it follows $R(E_1 \cap E_2) \leq \min\{R(E_1), R$
- Attributes are introduced as random variables (as in prob
- $R(A = a)$ and $P(a)$ are abbreviations of $R(\{\omega \mid A(\omega) =$

Possibility-Based Formalization (con

Definition: Let $U = \{A_1, \ldots, A_n\}$ be a set of attributes
sample space Ω with respective domains $\mathrm{dom}(A_i)$, $i = 1, \ldots$
over U is the restriction of a discrete possibility measure R
all events that can be defined by stating values for all attrib
$r_U = R|_{\mathcal{E}_U}$, where

$$\begin{aligned}
\mathcal{E}_U &= \Big\{ E \in 2^\Omega \;\Big|\; \exists a_1 \in \mathrm{dom}(A_1) : \ldots \exists a_n \in \mathrm{do} \\
&\qquad\qquad E \mathrel{\widehat{=}} \bigwedge_{A_j \in U} A_j = a_j \Big\} \\
&= \Big\{ E \in 2^\Omega \;\Big|\; \exists a_1 \in \mathrm{dom}(A_1) : \ldots \exists a_n \in \mathrm{do} \\
&\qquad\qquad E = \Big\{ \omega \in \Omega \;\Big|\; \bigwedge_{A_j \in U} A_j(\omega) = a
\end{aligned}$$

- Corresponds to the notion of a probability distribution.
- Advantage of this formalization: No index transformation
  for projections, there are just fewer terms in the conjunct

Possibility-Based Formalization (con

Definition: Let $U = \{A_1, \ldots, A_n\}$ be a set of attributes over U. Furthermore, let $\mathcal{M} = \{M_1, \ldots, M_m\} \subseteq 2^U$ be a se not necessarily disjoint) subsets of U satisfying

$$\bigcup_{M \in \mathcal{M}} M = U.$$

r_U is called **decomposable** w.r.t. $\mathcal{M}$ iff

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$r_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \min_{M \in \mathcal{M}} \Big\{ r_M\Big(\bigwedge_{A_i \in M} A_i =$$

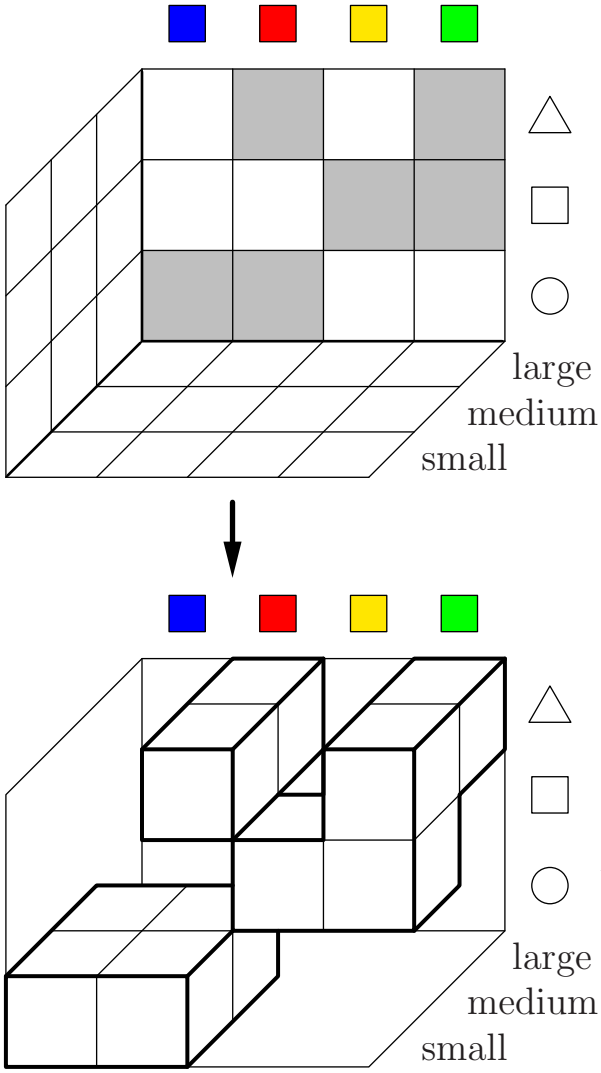
If r_U is decomposable w.r.t. $\mathcal{M}$, the set of relations

$$\mathcal{R}_\mathcal{M} = \{r_{M_1}, \ldots, r_{M_m}\} = \{r_M \mid M \in \mathcal{M}$$

is called the **decomposition** of r_U.

- Equivalent to **join decomposability** in database theor

Relational Decomposition: Simple E



Taking the min
projection to t
spanned by col
and of the proj
subspace spann
and size yields
three-dimension

Conditional Possibility and Indepe

Definition: Let Ω be a (finite) sample space, R a discrete po
Ω, and $E_1, E_2 \subseteq \Omega$ events. Then

$$R(E_1 \mid E_2) = R(E_1 \cap E_2)$$

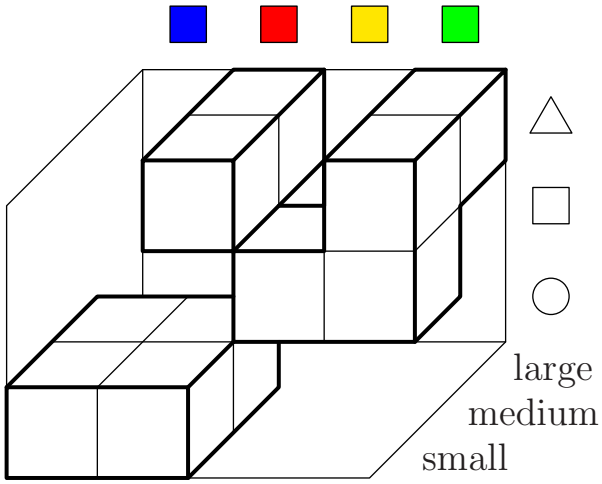
is called the **conditional possibility** of E_1 given E_2.

Definition: Let Ω be a (finite) sample space, R a discrete
on Ω, and A, B, and C attributes with respective domains
and $\mathrm{dom}(C)$. A and B are called **conditionally relationa**
given C, written $A \perp\!\!\!\perp_R B \mid C$, iff

$$\begin{aligned}
&\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) : \\
&\qquad R(A = a, C = c \mid B = b) = \min\{R(A = a \mid B = b), R \\
\Leftrightarrow\quad &R(A = a, C = c,\ B = b) = \min\{R(A = a,\ B = b), R
\end{aligned}$$

- Similar to the corresponding notions of probability theory

Conditional Independence: Simple E

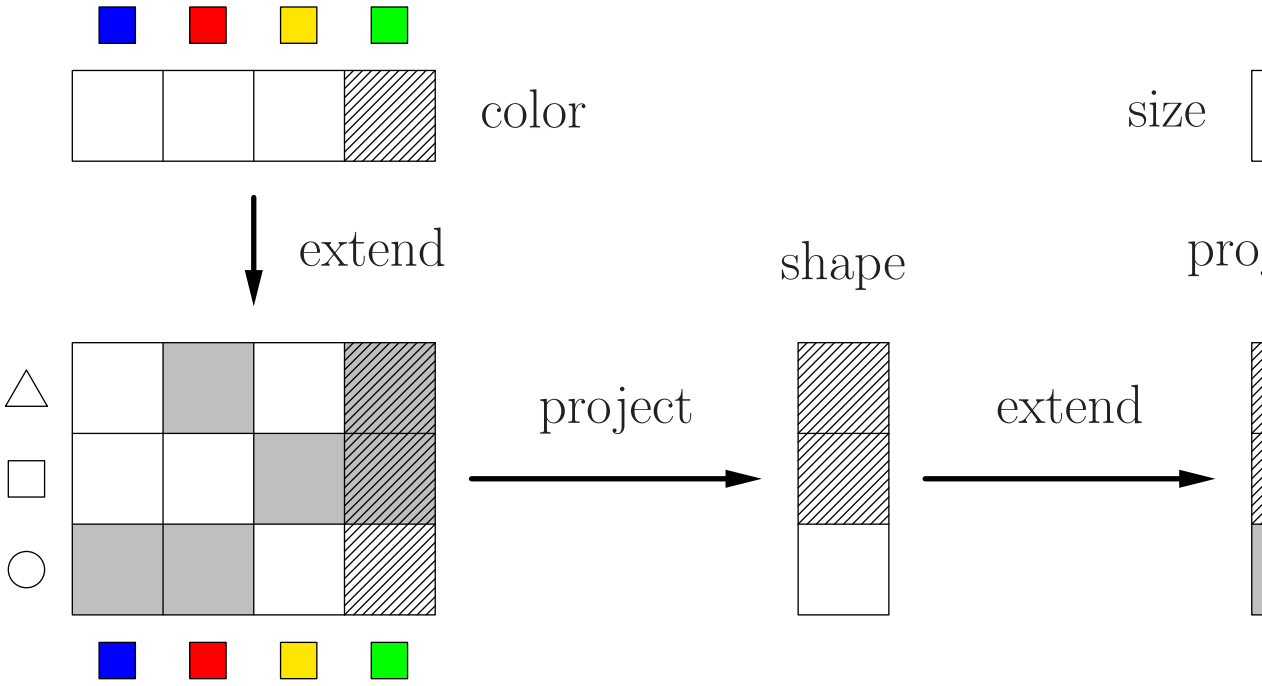


Example relatio
simple geometri
attributes: color

- In this example relation, the color of an object is condi
independent of its size given its shape.
- Intuitively: if we fix the shape, the colors and sizes that a
with this shape can be combined freely.
- Alternative view: once we know the shape, the color does n
information about the size (and vice versa).

Relational Evidence Propagati

Due to the fact that color and size are conditionally independe
the reasoning result can be obtained using only the projection



This reasoning scheme can be formally justified with discrete p

Relational Evidence Propagation, S

$$
\begin{aligned}
& R(B = b \mid A = a_{\text{obs}}) \\
&= R\Big(\bigvee_{a \in \text{dom}(A)} A = a, B = b, \bigvee_{c \in \text{dom}(C)} C = c \,\Big|\, A = a_{\text{obs}} \Big) \\
&\overset{(1)}{=} \max_{a \in \text{dom}(A)} \Big\{ \max_{c \in \text{dom}(C)} \{ R(A = a, B = b, C = c \mid A = a_{\text{obs}}) \\
&\overset{(2)}{=} \max_{a \in \text{dom}(A)} \Big\{ \max_{c \in \text{dom}(C)} \{ \min\{ R(A = a, B = b, C = c), R(A \\
&\overset{(3)}{=} \max_{a \in \text{dom}(A)} \Big\{ \max_{c \in \text{dom}(C)} \{ \min\{ R(A = a, B = b), R(B = b, C \\
&\qquad\qquad R(A = a \mid A = a_{\text{obs}}) \} \} \} \\
&= \max_{a \in \text{dom}(A)} \{ \min\{ R(A = a, B = b), R(A = a \mid A = a_{\text{obs}}), \\
&\qquad\qquad \underbrace{\max_{c \in \text{dom}(C)} \{ R(B = b, C = c) \} \} \}}_{= R(B = b) \geq R(A = a, B = b)} \\
&= \max_{a \in \text{dom}(A)} \{ \min\{ R(A = a, B = b), R(A = a \mid A = a_{\text{obs}}) \}
\end{aligned}
$$

Relational Evidence Propagation, Step 1

(1) holds because of the second axiom a discrete possibility me

(3) holds because of the fact that the relation R_{ABC} can be de
set $\mathcal{M} = \{\{A, B\}, \{B, C\}\}$.

(2) holds, since in the first place

$$R(A = a, B = b, C = c | A = a_{obs}) = R(A = a, B = b, C$$
$$= \begin{cases} R(A = a, B = b \\ 0, \end{cases}$$

and secondly

$$R(A = a \mid A = a_{\text{obs}}) = R(A = a, A = a_{\text{obs}}$$
$$= \begin{cases} R(A = a), & \text{if } a \\ 0, & \text{othe} \end{cases}$$

and therefore, since trivially $R(A = a) \geq R(A = a, B = b$

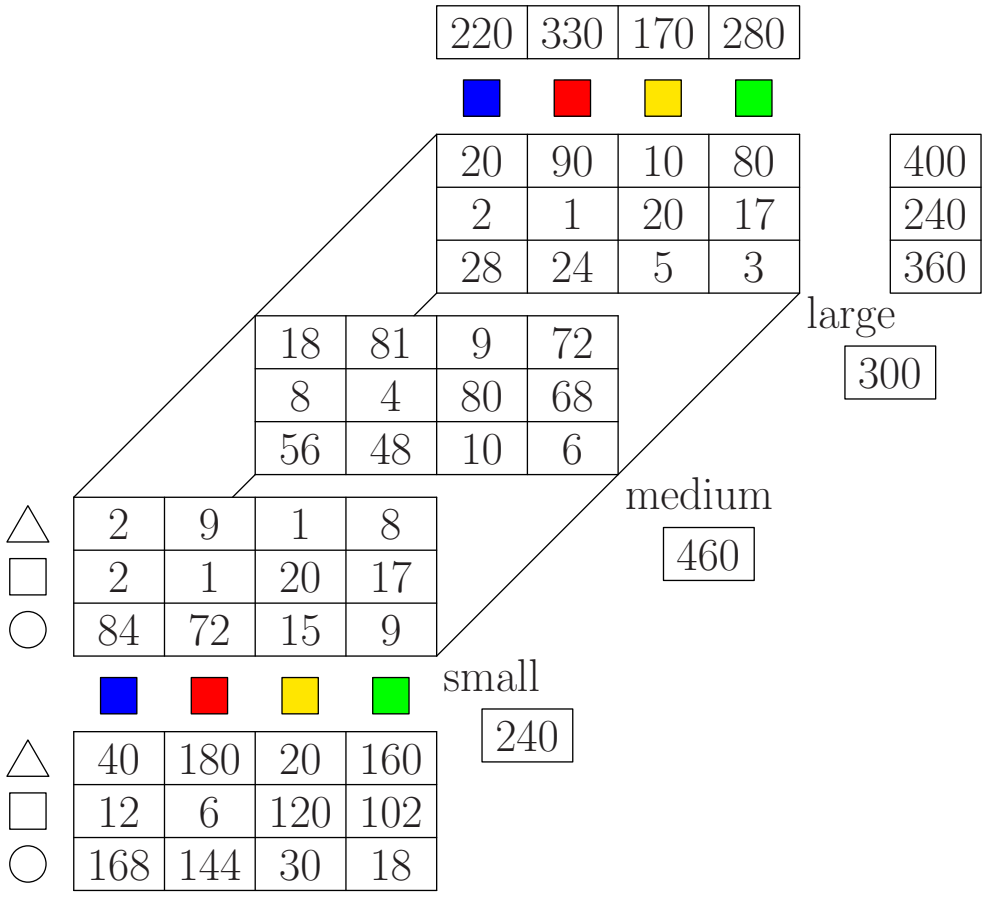
$$R(A = a, B = b, C = c \mid A = a_{obs})$$
$$= \min\{R(A = a, B = b, C = c), R(A = a \mid A$$

Relational Evidence Propagation, S

$$
\begin{aligned}
& R(C = c \mid A = a_{\text{obs}}) \\
&= R\Big(\bigvee_{a \in \text{dom}(A)} A = a, \bigvee_{b \in \text{dom}(B)} B = b, C = c \,\Big|\, A = a_{\text{obs}} \Big) \\
&\overset{(1)}{=} \max_{a \in \text{dom}(A)} \Big\{ \max_{b \in \text{dom}(B)} \{ R(A = a, B = b, C = c \mid A = a_{\text{obs}}) \\
&\overset{(2)}{=} \max_{a \in \text{dom}(A)} \Big\{ \max_{b \in \text{dom}(B)} \{ \min\{ R(A = a, B = b, C = c), R(A \\
&\overset{(3)}{=} \max_{a \in \text{dom}(A)} \Big\{ \max_{b \in \text{dom}(B)} \{ \min\{ R(A = a, B = b), R(B = b, C \\
&\qquad\qquad R(A = a \mid A = a_{\text{obs}}) \} \} \} \\
&= \max_{b \in \text{dom}(B)} \{ \min\{ R(B = b, C = c), \\
&\qquad \underbrace{\max_{a \in \text{dom}(A)} \{ \min\{ R(A = a, B = b), R(A = a \mid }_{= R(B = b \mid A = a_{\text{obs}})} \\
&= \max_{b \in \text{dom}(B)} \{ \min\{ R(B = b, C = c), R(B = b \mid A = a_{\text{obs}}) \}
\end{aligned}
$$

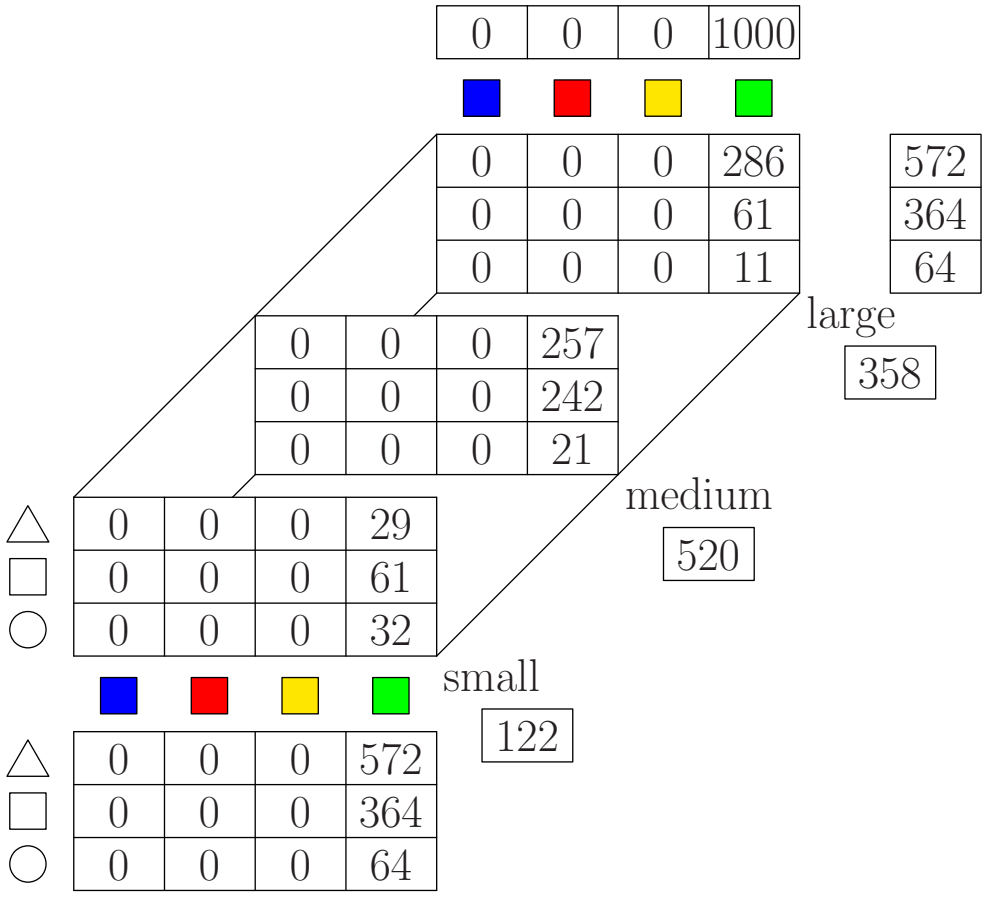
A Simple Example: The Probabilistic Case

A Probability Distribution



The numbers state the probability of the corresponding value
Compared to the example relation, the possible combinations

Reasoning: Computing Conditional Pr



Using the information that the given object is green:
The observed color has a posterior probability of 1.

Probabilistic Decomposition: Simple

- As for relational graphical models, the three-dimensional
  tion can be decomposed into projections to subspaces, n
  distribution on the subspace spanned by color and shap
  distribution on the subspace spanned by shape and size.

- The original probability distribution can be reconstructe
  distributions using the following formulae $\forall i, j, k$:

$$P\Big(a_i^{(\text{color})}, a_j^{(\text{shape})}, a_k^{(\text{size})}\Big) = P\Big(a_i^{(\text{color})}, a_j^{(\text{shape})}\Big) \cdot P$$

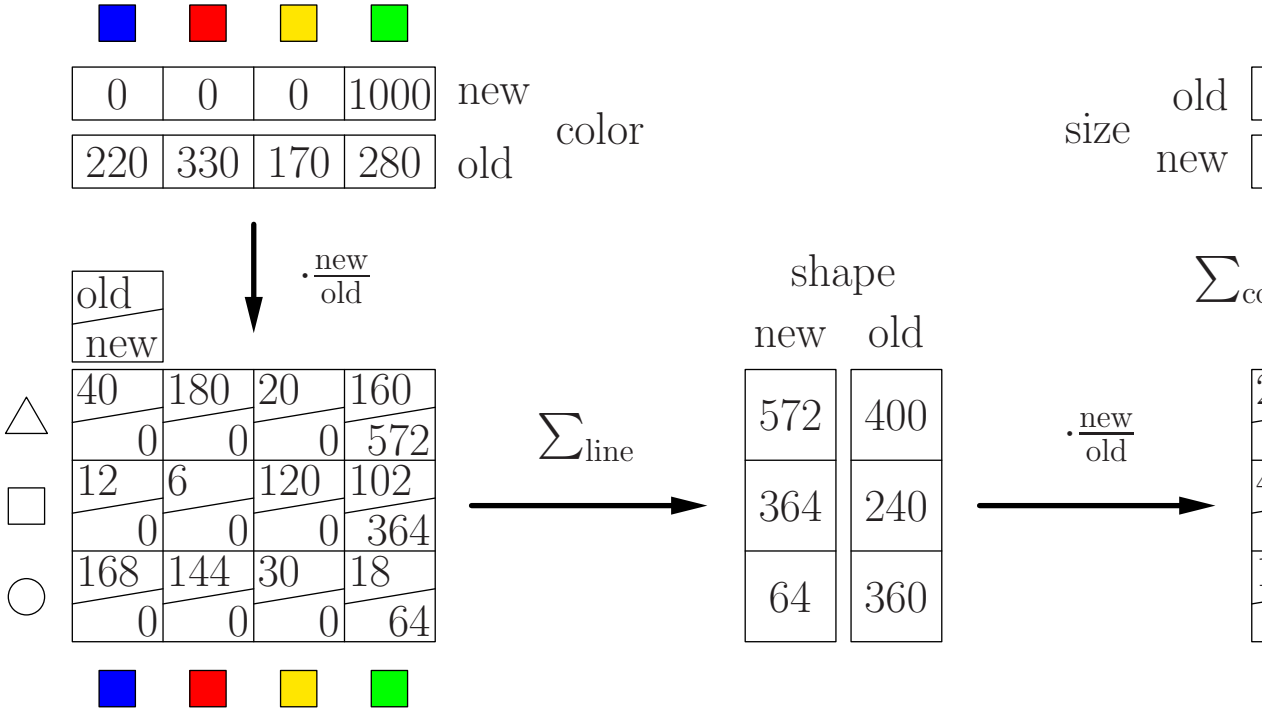
$$= \frac{P\Big(a_i^{(\text{color})}, a_j^{(\text{shape})}\Big) \cdot P}{P\Big(a_j^{(\text{shap}}$$

- These equations express the *conditional independence* of
  size given the attribute *shape*, since they only hold if $\forall i,$

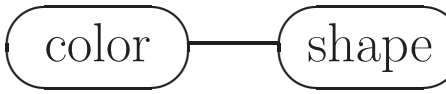
$$P\Big(a_k^{(\text{size})} \,\Big|\, a_j^{(\text{shape})}\Big) = P\Big(a_k^{(\text{size})} \,\Big|\, a_i^{(\text{color})}, a_j^{(}$$

Reasoning with Projections

Again the same result can be obtained using only projections
(marginal probability distributions):



This justifies a graph representation:



Probabilistic Graphical Model Formalization

Probabilistic Decomposition

Definition: Let $U = \{A_1, \ldots, A_n\}$ be a set of attributes a distribution over U. Furthermore, let $\mathcal{M} = \{M_1, \ldots, M_m\}$ nonempty (but not necessarily disjoint) subsets of U satisfying

$$\bigcup_{M \in \mathcal{M}} M = U.$$

p_U is called **decomposable** or **factorizable** w.r.t. $\mathcal{M}$ iff it product of m nonnegative functions $\phi_M : \mathcal{E}_M \to \mathbb{R}_0^+$, $M \in \mathcal{M}$

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$p_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \prod_{M \in \mathcal{M}} \phi_M\Big(\bigwedge_{A_i \in M} A_i =$$

If p_U is decomposable w.r.t. $\mathcal{M}$ the set of functions

$$\Phi_{\mathcal{M}} = \{\phi_{M_1}, \ldots, \phi_{M_m}\} = \{\phi_M \mid M \in \mathcal{M}$$

is called the **decomposition** or the **factorization** of p_U. T are called the **factor potentials** of p_U.

Conditional Independence

Definition: Let Ω be a (finite) sample space, P a probability
A, B, and C attributes with respective domains $\mathrm{dom}(A)$, do
A and B are called **conditionally probabilistically inde**
written $A \perp\!\!\!\perp_P B \mid C$, iff

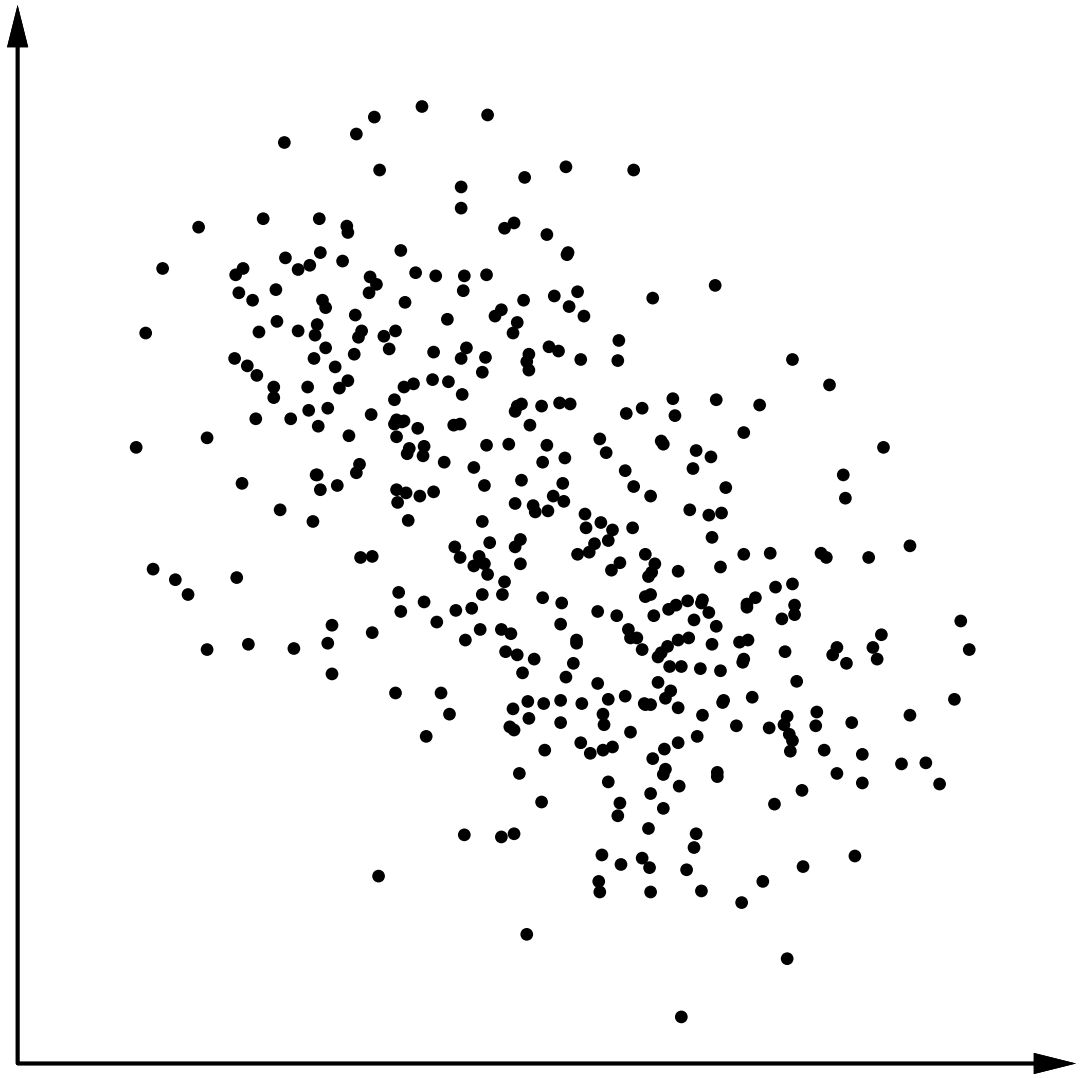
$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) :$$
$$P(A = a, B = b \mid C = c) = P(A = a \mid C = c) \cdot P(B$$

Equivalent formula:

$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) :$$
$$P(A = a \mid B = b, C = c) = P(A = a \mid C$$

- Conditional independences make it possible to consider p
  distribution independent of others.
- Therefore it is plausible that a set of conditional indepen
  decomposition of a joint probability distribution.

Conditional Independence: An Ex



Dependence
smoking an

Each dot re

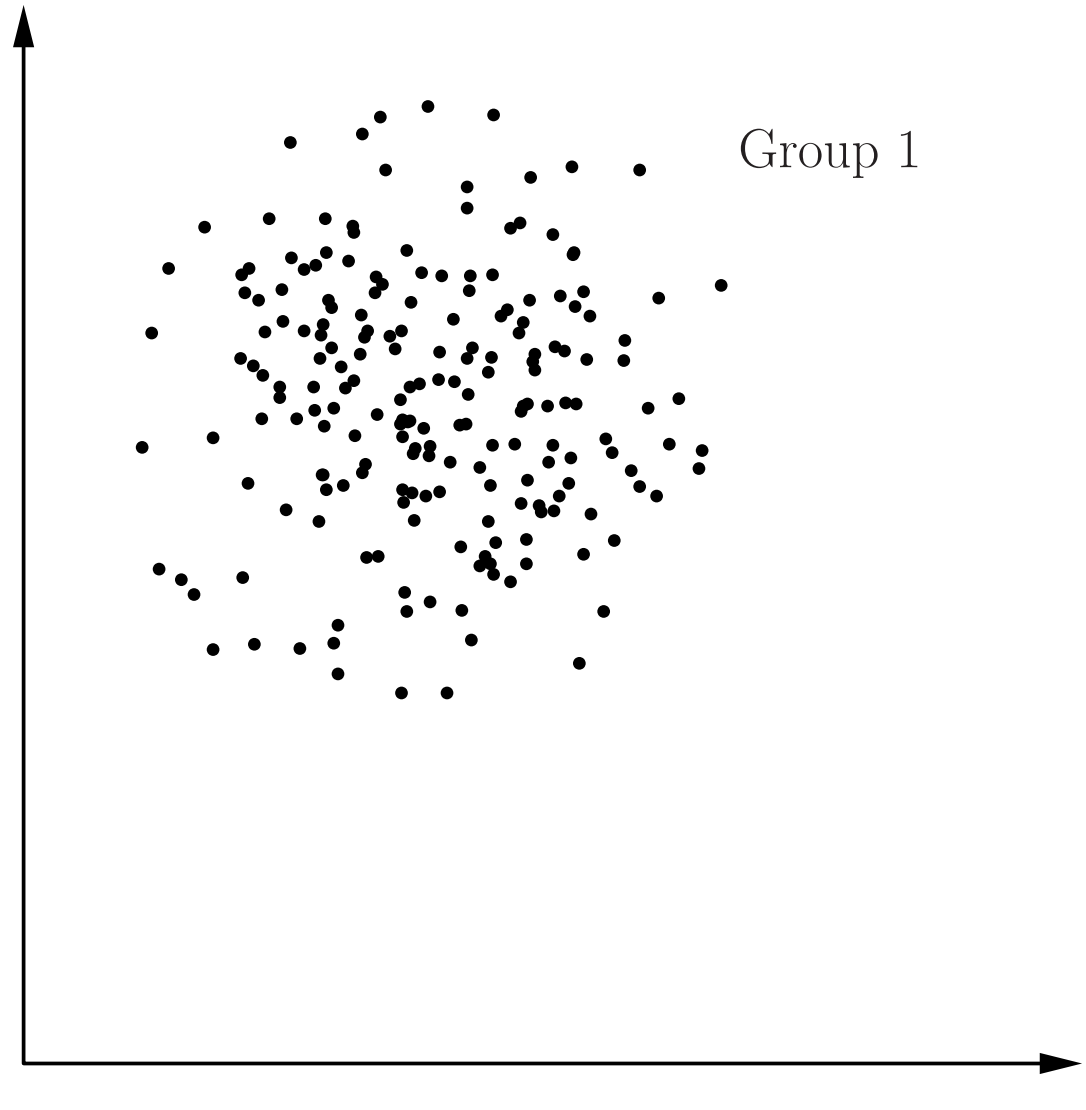
x-axis: ag
y-axis: av
cig

Weak, but

The more ci
the lower th

(Note that
and thus s
revealing a

Conditional Independence: An Ex



Conjectured

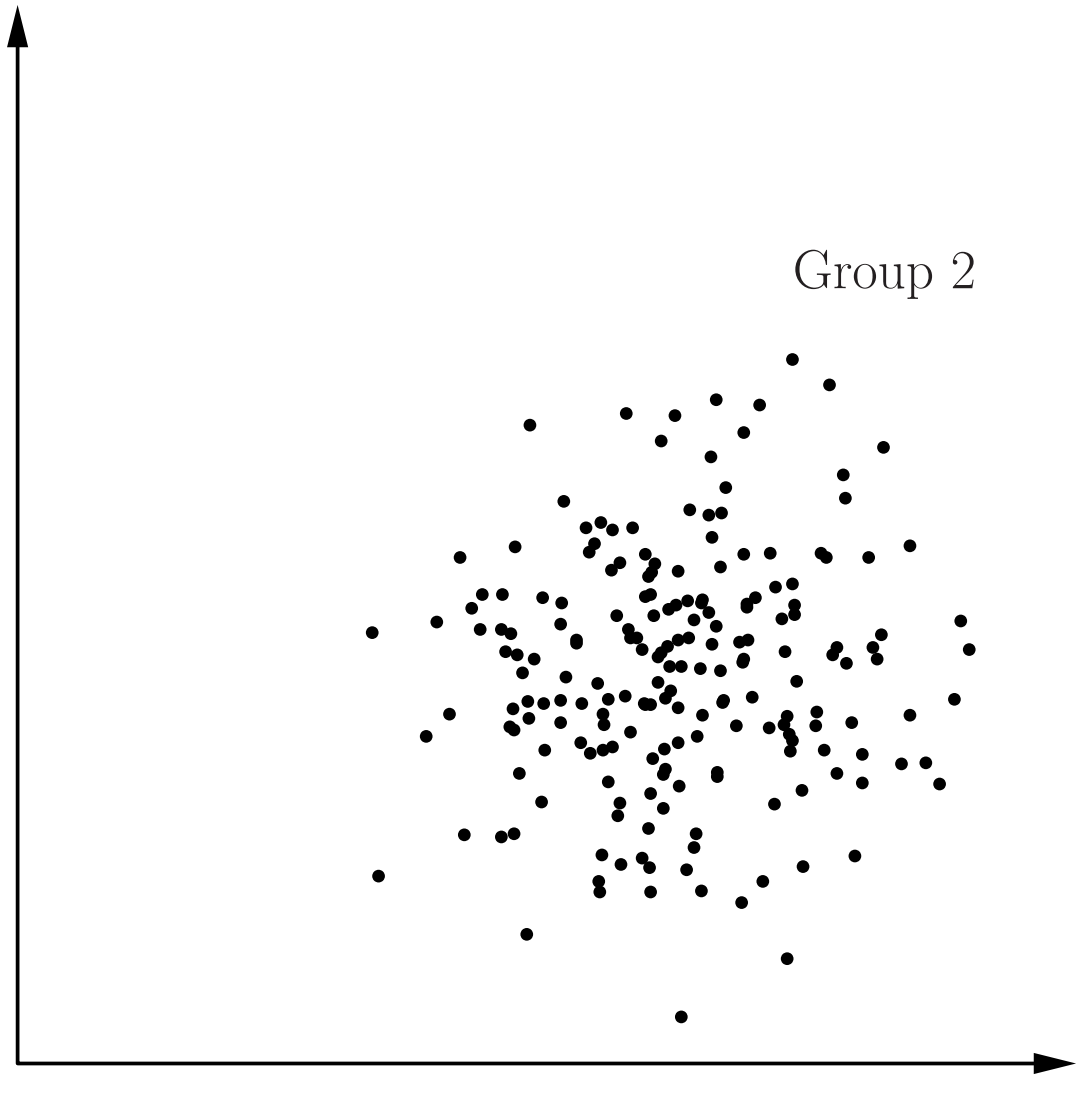
There is a c
namely whe
is exposed t

If this were
splitting the
remove the

Group 1:
exposed to s

(Note that
and theref
as an argu
hazards ca

Conditional Independence: An Ex



Conjectured

There is a c
namely whe
is exposed t

If this were
splitting the
remove the

Group 2:
not exposed

(Note that
and theref
as an argu
hazards ca

Probabilistic Decomposition (conti

Chain Rule of Probability:

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$P\Big(\bigwedge_{i=1}^{n} A_i = a_i\Big) = \prod_{i=1}^{n} P\Big(A_i = a_i \,\Big|\, \bigwedge_{j=1}^{i-1} A_j$$

- The chain rule of probability is valid in general (or at least for strictly positive distributions).

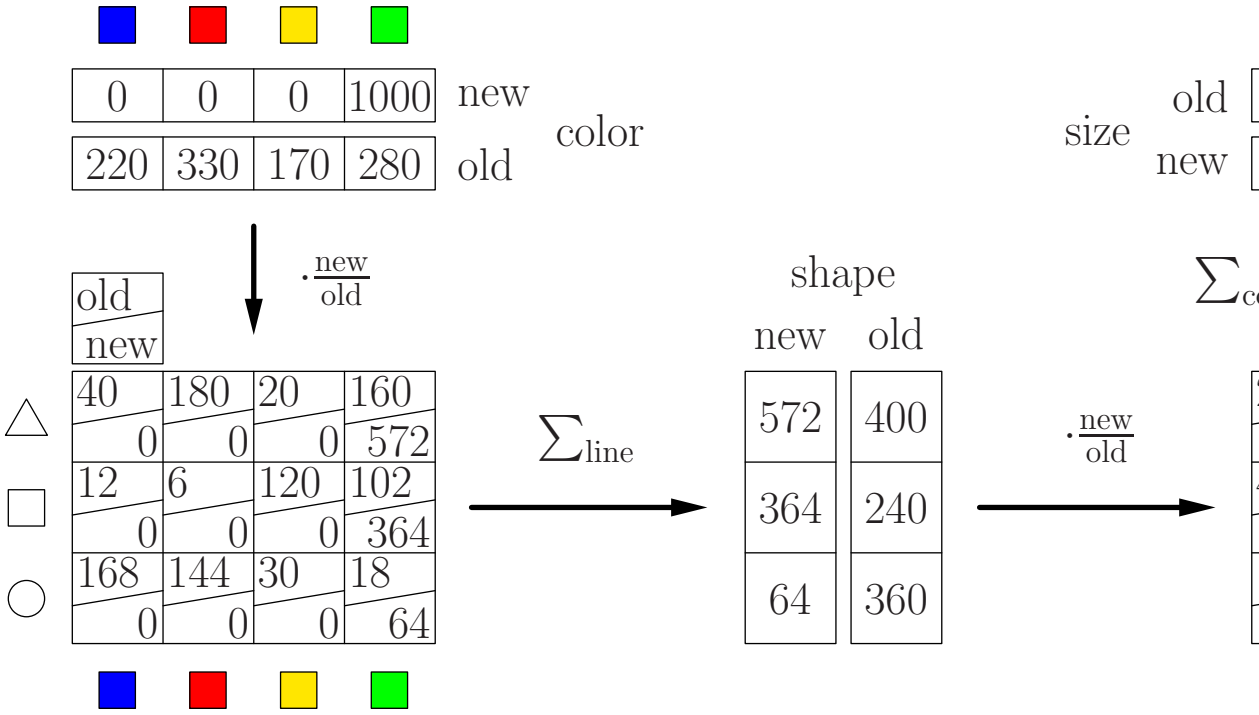
Chain Rule Factorization:

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$P\Big(\bigwedge_{i=1}^{n} A_i = a_i\Big) = \prod_{i=1}^{n} P\Big(A_i = a_i \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A}$$

- Conditional independence statements are used to "cancel"

Reasoning with Projections

Due to the fact that color and size are conditionally independe
the reasoning result can be obtained using only the projection



This reasoning scheme can be formally justified with probabili

Probabilistic Evidence Propagation,

$$
\begin{aligned}
& P(B = b \mid A = a_{\text{obs}}) \\
&= P\Big(\bigvee_{a \in \text{dom}(A)} A = a, B = b, \bigvee_{c \in \text{dom}(C)} C = c \,\Big|\, A = a_{\text{obs}} \Big) \\
&\overset{(1)}{=} \sum_{a \in \text{dom}(A)} \sum_{c \in \text{dom}(C)} P(A = a, B = b, C = c \mid A = a_{\text{obs}}) \\
&\overset{(2)}{=} \sum_{a \in \text{dom}(A)} \sum_{c \in \text{dom}(C)} P(A = a, B = b, C = c) \cdot \frac{P(A = a \mid \\
&\overset{(3)}{=} \sum_{a \in \text{dom}(A)} \sum_{c \in \text{dom}(C)} \frac{P(A = a, B = b) P(B = b, C = c)}{P(B = b)} \cdot \\
&= \sum_{a \in \text{dom}(A)} P(A = a, B = b) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)} \underbrace{\sum_{c \in \text{dom}}} \\
&= \sum_{a \in \text{dom}(A)} P(A = a, B = b) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)}.
\end{aligned}
$$

Probabilistic Evidence Propagation, Step

(1) holds because of Kolmogorov's axioms.

(3) holds because of the fact that the distribution p_{ABC} can b
the set $\mathcal{M} = \{\{A, B\}, \{B, C\}\}$.

(2) holds, since in the first place

$$P(A = a, B = b, C = c \,|\, A = a_{obs}) \;=\; \frac{P(A = a, B = b, C}{P(A =}$$

$$= \begin{cases} \dfrac{P(A = a, B =}{P(A = a_{\mathrm{o}} \\ 0, \end{cases}$$

and secondly

$$P(A = a, A = a_{\mathrm{obs}}) \;=\; \begin{cases} P(A = a), & \text{if } a = \\ 0, & \text{othe} \end{cases}$$

and therefore

$$P(A = a, B = b, C = c \mid A = a_{obs})$$

$$= \; P(A = a, B = b, C = c) \cdot \frac{P(A = a \mid A}{P(A = a}$$

Probabilistic Evidence Propagation,

$$
\begin{aligned}
&P(C = c \mid A = a_{\text{obs}}) \\
&= P\Big(\bigvee_{a \in \text{dom}(A)} A = a, \bigvee_{b \in \text{dom}(B)} B = b, C = c \,\Big|\, A = a_{\text{obs}} \Big) \\
&\overset{(1)}{=} \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} P(A = a, B = b, C = c \mid A = a_{\text{obs}}) \\
&\overset{(2)}{=} \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} P(A = a, B = b, C = c) \cdot \frac{P(A = a \mid} {P(A} \\
&\overset{(3)}{=} \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} \frac{P(A = a, B = b) P(B = b, C = c)}{P(B = b)} \cdot \\
&= \sum_{b \in \text{dom}(B)} \frac{P(B = b, C = c)}{P(B = b)} \underbrace{\sum_{a \in \text{dom}(A)} P(A = a, B = b) \cdot}_{= P(B = b \mid A =} \\
&= \sum_{b \in \text{dom}(B)} P(B = b, C = c) \cdot \frac{P(B = b \mid A = a_{\text{obs}})}{P(B = b)}.
\end{aligned}
$$

Graphical Models:
The General Theory

(Semi-)Graphoid Axioms

Definition: Let V be a set of (mathematical) objects and $(\cdot \perp$
relation of subsets of V. Furthermore, let W, X, Y, and Z be
of V. The four statements

symmetry:	$(X \perp\!\!\!\perp Y \mid Z) \;\Rightarrow\; (Y \perp\!\!\!\perp X \mid Z)$
decomposition:	$(W \cup X \perp\!\!\!\perp Y \mid Z) \;\Rightarrow\; (W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp$
weak union:	$(W \cup X \perp\!\!\!\perp Y \mid Z) \;\Rightarrow\; (X \perp\!\!\!\perp Y \mid Z \cup W)$
contraction:	$(X \perp\!\!\!\perp Y \mid Z \cup W) \wedge (W \perp\!\!\!\perp Y \mid Z) \;\Rightarrow\; (W \cup$

are called the **semi-graphoid axioms**. A three-place relation
isfies the semi-graphoid axioms for all W, X, Y, and Z is called
The above four statements together with

intersection:	$(W \perp\!\!\!\perp Y \mid Z \cup X) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \;\Rightarrow$

are called the **graphoid axioms**. A three-place relation $(\cdot \perp$
the graphoid axioms for all W, X, Y, and Z is called a **graph**

Illustration of the (Semi-)Graphoid

decomposition:

weak union:

contraction:

intersection:

- Similar to the properties of **separation in graphs**.
- **Idea:** Represent conditional independence by separation

Separation in Graphs

Definition: Let $G = (V, E)$ be an undirected graph and
disjoint subsets of nodes. Z **u-separates** X and Y in G, wr
iff all paths from a node in X to a node in Y contain a node
contains a node in Z is called **blocked** (by Z), otherwise it is

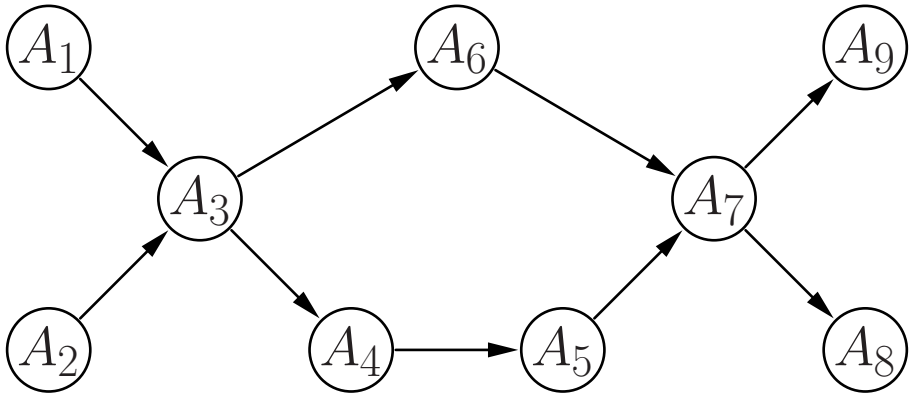
Definition: Let $\vec{G} = (V, \vec{E})$ be a directed acyclic graph and
disjoint subsets of nodes. Z **d-separates** X and Y in $\vec{G}$, wr
iff there is *no* path from a node in X to a node in Y along whi
conditions hold:

1. every node with converging edges either is in Z or has a d
2. every other node is not in Z.

A path satisfying the two conditions above is said to be **activ**
otherwise it is said to be **blocked** (by Z).

Separation in Directed Acyclic Gr

Example Graph:



Valid Separations:

$\langle \{A_1\} \mid \{A_3\} \mid \{A_4\} \rangle$ $\quad$ $\langle \{A_8\} \mid \{A_7\} \mid$

$\langle \{A_3\} \mid \{A_4, A_6\} \mid \{A_7\} \rangle$ $\quad$ $\langle \{A_1\} \mid \emptyset \mid \{A_2$

Invalid Separations:

$\langle \{A_1\} \mid \{A_4\} \mid \{A_2\} \rangle$ $\quad$ $\langle \{A_1\} \mid \{A_6\} \mid$

$\langle \{A_4\} \mid \{A_3, A_7\} \mid \{A_6\} \rangle$ $\quad$ $\langle \{A_1\} \mid \{A_4, A_9$

Conditional (In)Dependence Gra

Definition: Let $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ be a three-place relation rep
conditional independence statements that hold in a given distrib
of attributes. An undirected graph $G = (U, E)$ over U is ca
dependence graph or a **dependence map** w.r.t. δ, iff fo
$X, Y, Z \subseteq U$ of attributes

$$X \perp\!\!\!\perp_\delta Y \mid Z \;\Rightarrow\; \langle X \mid Z \mid Y \rangle_G,$$

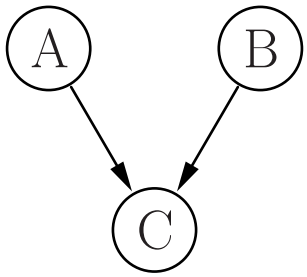
i.e., if G captures by u-separation all (conditional) independe
and thus represents only valid (conditional) dependences. Sim
conditional independence graph or an **independence**
all disjoint subsets $X, Y, Z \subseteq U$ of attributes

$$\langle X \mid Z \mid Y \rangle_G \;\Rightarrow\; X \perp\!\!\!\perp_\delta Y \mid Z,$$

i.e., if G captures by u-separation only (conditional) independ
in δ. G is said to be a **perfect map** of the conditional (in)de
is both a dependence map and an independence map.

Limitations of Graph Representa

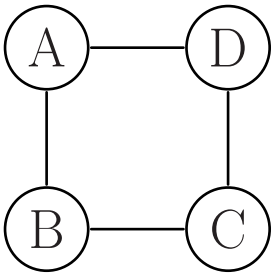
Perfect directed map, no perfect undirected map:



p_{ABC}	$A = a_1$	
	$B = b_1$	$B = b_2$
$C = c_1$	$4/24$	$3/24$
$C = c_2$	$2/24$	$3/24$

$A =$	
$B = b_1$	
$3/24$	
$3/24$	

Perfect undirected map, no perfect directed map:



p_{ABCD}		$A = a_1$	
		$B = b_1$	$B = b_2$
$C = c_1$	$D = d_1$	$1/47$	$1/47$
	$D = d_2$	$1/47$	$1/47$
$C = c_2$	$D = d_1$	$1/47$	$2/47$
	$D = d_2$	$2/47$	$4/47$

Markov Properties of Undirected G

Definition: An undirected graph $G = (U, E)$ over a set U of
have (w.r.t. a distribution δ) the
pairwise Markov property,
iff in δ any pair of attributes which are nonadjacent in the gra
independent given all remaining attributes, i.e., iff

$$\forall A, B \in U, A \neq B: \quad (A, B) \notin E \;\Rightarrow\; A \perp\!\!\!\perp_\delta B \mid U$$

local Markov property,
iff in δ any attribute is conditionally independent of all remain
its neighbors, i.e., iff

$$\forall A \in U: \quad A \perp\!\!\!\perp_\delta U - \text{closure}(A) \mid \text{boundary}$$

global Markov property,
iff in δ any two sets of attributes which are u-separated by a th
independent given the attributes in the third set, i.e., iff

$$\forall X, Y, Z \subseteq U: \quad \langle X \mid Z \mid Y \rangle_G \;\Rightarrow\; X \perp\!\!\!\perp_\delta Y$$

Markov Properties of Directed Acycli

Definition: A directed acyclic graph $\vec{G} = (U, \vec{E})$ over a set U to have (w.r.t. a distribution δ) the
pairwise Markov property,
iff in δ any attribute is conditionally independent of any non-de its parents given all remaining non-descendants, i.e., iff

$$\forall A, B \in U : \; B \in \text{nondescs}(A) - \text{parents}(A) \;\Rightarrow\; A \perp\!\!\!\perp_\delta B \mid \text{r}$$

local Markov property,
iff in δ any attribute is conditionally independent of all remain given its parents, i.e., iff

$$\forall A \in U : \quad A \perp\!\!\!\perp_\delta \text{nondescs}(A) - \text{parents}(A) \mid \text{par}$$

global Markov property,
iff in δ any two sets of attributes which are d-separated by a th independent given the attributes in the third set, i.e., iff

$$\forall X, Y, Z \subseteq U : \quad \langle X \mid Z \mid Y \rangle_{\vec{G}} \;\Rightarrow\; X \perp\!\!\!\perp_\delta Y$$

Equivalence of Markov Propert

Theorem: If a three-place relation $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ representing t
independence statements that hold in a given joint distributio
attributes satisfies the graphoid axioms, then the pairwise, the
Markov property of an undirected graph $G = (U, E)$ over U a

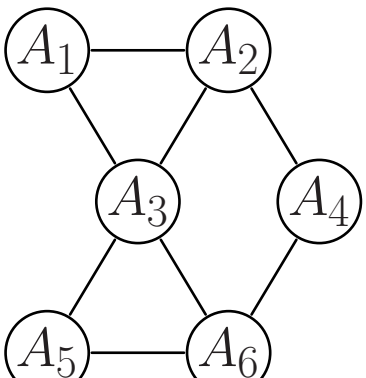
Theorem: If a three-place relation $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ representing t
independence statements that hold in a given joint distributio
attributes satisfies the semi-graphoid axioms, then the local an
property of a directed acyclic graph $\vec{G} = (U, \vec{E})$ over U are eq

If $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ satisfies the graphoid axioms, then the pairwise
global Markov property are equivalent.

Undirected Graphs and Decompos

Definition: A probability distribution p_V over a set V o
decomposable or **factorizable w.r.t. an undirected**
over V iff it can be written as a product of nonnegative functi
cliques of G. That is, let $\mathcal{M}$ be a family of subsets of varia
subgraphs of G induced by the sets $M \in \mathcal{M}$ are the maximal
there exist functions $\phi_M : \mathcal{E}_M \to \mathbb{R}_0^+$, $M \in \mathcal{M}$, $\forall a_1 \in$ do
$\mathrm{dom}(A_n)$:

$$p_V\Big(\bigwedge_{A_i \in V} A_i = a_i\Big) = \prod_{M \in \mathcal{M}} \phi_M\Big(\bigwedge_{A_i \in M} A_i =$$

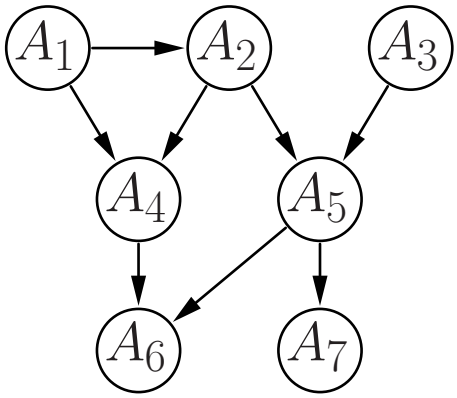


$$\begin{aligned} & p_V(A_1 = a_1, \ldots, A_6 = a_6) \\ = \;& \phi_{A_1A_2A_3}(A_1 = a_1, A_2 = \\ \cdot \;& \phi_{A_3A_5A_6}(A_3 = a_3, A_5 = \\ \cdot \;& \phi_{A_2A_4}(A_2 = a_2, A_4 = a \\ \cdot \;& \phi_{A_4A_6}(A_4 = a_4, A_6 = a \end{aligned}$$

Directed Acyclic Graphs and Decomp

Definition: A probability distribution p_U over a set U of att
composable or **factorizable w.r.t. a directed acyclic**
over U, iff it can be written as a product of the conditional
attributes given their parents in $\vec{G}$, i.e., iff

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$p_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \prod_{A_i \in U} P\Big(A_i = a_i \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}_{\vec{G}}($$



$$\begin{aligned} & P(A_1 = a_1, \ldots, A_7 = a_7) \\ = \;& P(A_1 = a_1) \cdot P(A_2 = a_2 \mid A \\ \cdot \;& P(A_4 = a_4 \mid A_1 = a_1, A_2 = \\ \cdot \;& P(A_5 = a_5 \mid A_2 = a_2, A_3 = \\ \cdot \;& P(A_6 = a_6 \mid A_4 = a_4, A_5 = \\ \cdot \;& P(A_7 = a_7 \mid A_5 = a_5). \end{aligned}$$

Conditional Independence Graphs and De

Core Theorem of Graphical Models:
Let p_V be a strictly positive probability distribution on a set ables. A directed or undirected graph $G = (V, E)$ is a condi graph w.r.t. p_V if and only if p_V is factorizable w.r.t. G.

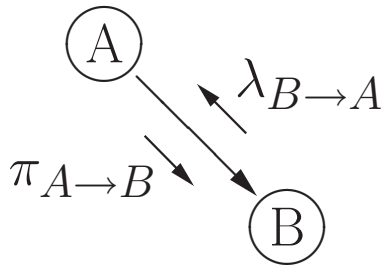
Definition: A **Markov network** is an undirected condi graph of a probability distribution p_V together with the fam tions ϕ_M of the factorization induced by the graph.

Definition: A **Bayesian network** is a directed conditional of a probability distribution p_U together with the family of cond of the factorization induced by the graph.

- Sometimes the conditional independence graph is required
- For correct evidence propagation it is not required that th Evidence propagation may just be less efficient than possi

Probabilistic Graphical Model
Evidence Propagation in Polytr

Evidence Propagation in Polytr



Idea: Node processors com
by message passing: π-messag
from parent to child and λ-m
sent from child to parent.

Derivation of the Propagation Formulae

Computation of Marginal Distribution:

$$P(A_g = a_g) \;=\; \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \mathrm{dom}(A_i)}} P\Big(\bigwedge_{A_j \in U} A_j =$$

Chain Rule Factorization w.r.t. the Polytree:

$$P(A_g = a_g) \;=\; \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{pare}}$$

Evidence Propagation in Polytrees (co

Decomposition w.r.t. Subgraphs:

$$P(A_g = a_g) = \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \mathrm{dom}(A_i)}} \Big(P\Big(A_g = a_g \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}($$
$$\cdot \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}($$
$$\cdot \prod_{A_k \in U_-(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}($$

Attribute sets underlying subgraphs:

$$U_B^A(C) = \{C\} \cup \{D \in U \mid D \underset{\vec{G}'}{\approx} C, \vec{G}' = (U, E - \{($$

$$U_+(A) = \bigcup_{C \in \mathrm{parents}(A)} U_A^C(C), \qquad U_+(A,B) = \bigcup_{C \in \mathrm{parents}($$

$$U_-(A) = \bigcup_{C \in \mathrm{children}(A)} U_C^A(C), \qquad U_-(A,B) = \bigcup_{C \in \mathrm{children}}$$

Evidence Propagation in Polytrees (co

Terms that are independent of a summation variable can b
corresponding sum. This yields a decomposition into two mai

$$
\begin{aligned}
P(A_g = a_g) \;=\; & \Bigg(\sum_{\substack{\forall A_i \in \text{parents}(A_g): \\ a_i \in \text{dom}(A_i)}} P\Big(A_g = a_g \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_g} \\
& \cdot \Big[\sum_{\substack{\forall A_i \in U_+^*(A_g): \\ a_i \in \text{dom}(A_i)}} \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \\
& \cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_g): \\ a_i \in \text{dom}(A_i)}} \prod_{A_k \in U_-(A_g)} P\Big(A_k = a_k \,\Big|\, \\
= \;& \pi(A_g = a_g) \cdot \lambda(A_g = a_g),
\end{aligned}
$$

where $U_+^*(A_g) = U_+(A_g) - \text{parents}(A_g)$.

Evidence Propagation in Polytrees (co

$$\sum_{\substack{\forall A_i \in U_+^*(A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a$$

$$= \prod_{A_p \in \mathrm{parents}(A_g)} \Bigg(\sum_{\substack{\forall A_i \in \mathrm{parents}(A_p): \\ a_i \in \mathrm{dom}(A_i)}} P\Big(A_p = a_p \,\Big|\, \bigwedge_{A_j \in \mathrm{paren}$$

$$\cdot \Bigg[\sum_{\substack{\forall A_i \in U_+^*(A_p): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_+(A_p)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parent}}$$

$$\cdot \Bigg[\sum_{\substack{\forall A_i \in U_-(A_p, A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_p, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{paren}$$

$$= \prod_{A_p \in \mathrm{parents}(A_g)} \pi(A_p = a_p)$$

$$\cdot \Bigg[\sum_{\substack{\forall A_i \in U_-(A_p, A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_p, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{paren}$$

Evidence Propagation in Polytrees (co

$$\sum_{\substack{\forall A_i \in U_+^*(A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a$$

$$= \prod_{A_p \in \mathrm{parents}(A_g)} \pi(A_p = a_p)$$

$$\cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p, A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_p, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{paren}}$$

$$= \prod_{A_p \in \mathrm{parents}(A_g)} \pi_{A_p \to A_g}(A_p = a_p)$$

$$\pi(A_g = a_g) = \sum_{\substack{\forall A_i \in \mathrm{parents}(A_g): \\ a_i \in \mathrm{dom}(A_i)}} P(A_g = a_g \,|\, \bigwedge_{A_j \in \mathrm{parents}}$$

$$\cdot \prod_{A_p \in \mathrm{parents}(A_g)} \pi_{A_p \to A_g}$$

Evidence Propagation in Polytrees (co

$$\lambda(A_g = a_g) = \sum_{\substack{\forall A_i \in U_-(A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_g)} P(A_k = a_k \mid \bigwedge_{A_j \in \mathrm{pa}}$$

$$= \prod_{A_c \in \mathrm{children}(A_g)} \sum_{a_c \in \mathrm{dom}(A_c)}$$

$$\Big(\sum_{\substack{\forall A_i \in \mathrm{parents}(A_c) - \{A_g\}: \\ a_i \in \mathrm{dom}(A_i)}} P(A_c = a_c \mid \bigwedge_{A_j \in \mathrm{parents}(A_c)} A_j$$

$$\cdot \Big[\sum_{\substack{\forall A_i \in U_+^*(A_c, A_g): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_+(A_c, A_g)} P(A_k = a_k \mid \bigwedge_{A_j \in \mathrm{paren}}$$

$$\cdot \underbrace{\Big[\sum_{\substack{\forall A_i \in U_-(A_c): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_c)} P(A_k = a_k \mid \bigwedge_{A_j \in \mathrm{paren}}}_{= \lambda(A_c = a_c)}$$

$$= \prod_{A_c \in \mathrm{children}(A_g)} \lambda_{A_c \to A_g}(A_g = a_g)$$

Propagation Formulae without Evi

$$\pi_{A_p \to A_c}(A_p = a_p)$$
$$= \pi(A_p = a_p) \cdot \Bigg[\sum_{\substack{\forall A_i \in U_-(A_p, A_c): \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_p, A_c)} P\Big(A_k = a_k \,\Big|$$

$$= \frac{P(A_p = a_p)}{\lambda_{A_c \to A_p}(A_p = a_p)}$$

$$\lambda_{A_c \to A_p}(A_p = a_p)$$
$$= \sum_{a_c \in \mathrm{dom}(A_c)} \lambda(A_c = a_c) \sum_{\substack{\forall A_i \in \mathrm{parents}(A_c) - \{A_p\}: \\ a_i \in \mathrm{dom}(A_k)}} P\Big(A_c = a_c \,\Big|$$
$$\cdot \prod_{A_k \in \mathrm{parents}(A_c) - \{A_p\}} \pi_{A_k \to A_p}(A_k$$

Evidence Propagation in Polytrees (co

Evidence: The attributes in a set X_{obs} are observed.

$$P\Big(A_g = a_g \,\Big|\, \bigwedge_{A_k \in X_{\text{obs}}} A_k = a_k^{(\text{obs})}\Big)$$

$$= \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \text{dom}(A_i)}} P\Big(\bigwedge_{A_j \in U} A_j = a_j \,\Big|\, \bigwedge_{A_k \in X_{\text{obs}}} A_k = a_k^{(\text{obs})}$$

$$= \alpha \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \text{dom}(A_i)}} P\Big(\bigwedge_{A_j \in U} A_j = a_j \Big) \prod_{A_k \in X_{\text{obs}}} P\Big(A_k = a$$

where $\alpha = \dfrac{1}{P\left(\bigwedge_{A_k \in X_{\text{obs}}} A_k = a_k^{(\text{obs})}\right)}$

Propagation Formulae with Evid

$$\pi_{A_p \to A_c}(A_p = a_p)$$

$$= P\left(A_p = a_p \,\middle|\, A_p = a_p^{(\text{obs})}\right) \cdot \pi(A_p = a_p)$$

$$\cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p, A_c): \\ a_i \in \text{dom}(A_i)}} \prod_{A_k \in U_-(A_p, A_c)} P\Big(A_k = a_k \,\Big|\, \Big/ \quad A_j \in \text{pare}$$

$$= \begin{cases} \beta, & \text{if } a_p = a_p^{(\text{obs})}, \\ 0, & \text{otherwise,} \end{cases}$$

- The value of β is not explicitly determined. Usually a va the correct value is implicitly determined later by norm probability distribution for A_g.

Propagation Formulae with Evid

$$\lambda_{A_c \to A_p}(A_p = a_p)$$

$$= \sum_{a_c \in \mathrm{dom}(A_c)} P\left(A_c = a_c \,\middle|\, A_c = a_c^{(\mathrm{obs})}\right) \cdot \lambda(A_c = a_c)$$

$$\cdot \sum_{\substack{\forall A_i \in \mathrm{parents}(A_c) - \{A_p\}: \\ a_i \in \mathrm{dom}(A_k)}} P\Big(A_c = a_c \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}($$

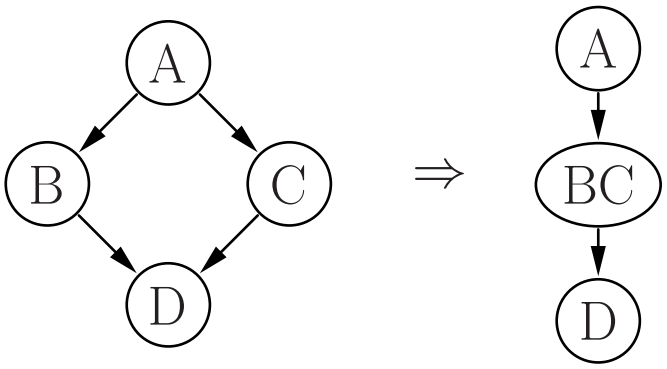
$$\cdot \prod_{A_k \in \mathrm{parents}(A_c) - \{A_p\}} \pi_{A_k}$$

Probabilistic Graphical Model
Evidence Propagation in Multiply Connec

Propagation in Multiply Connected N

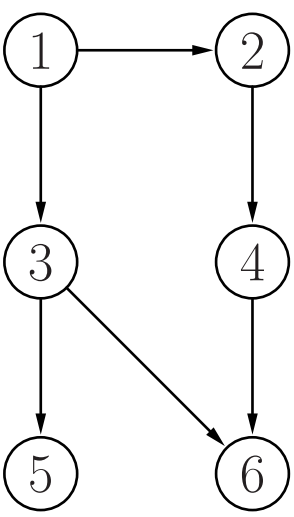
- Multiply connected networks pose a problem:
  - There are several ways on which information can trave (node) to another.
  - As a consequence, the same evidence may be used t probability distribution of an attribute.
  - Since probabilistic update is not idempotent, multiple i evidence usually invalidates the result.
- General idea to solve this problem:

  Transform network into a singly connected stru

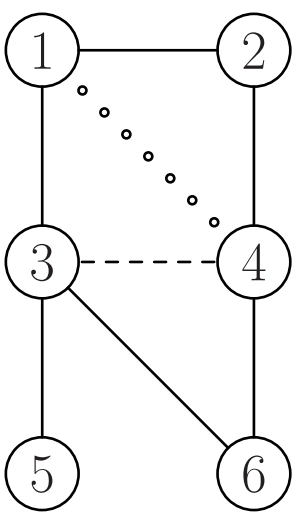


Merging attribu
polytree algorit
multiply connec

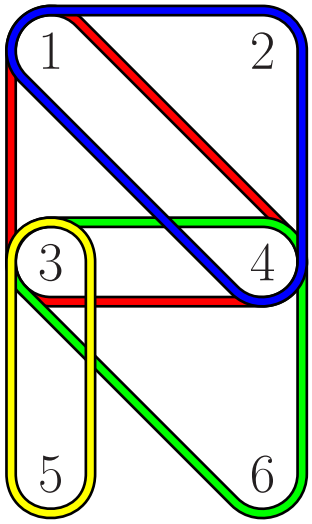
Triangulation and Join Tree Constr



original graph



triangulated moral graph



maximal cliques

- A singly connected structure is obtained by triangulating forming a tree of maximal cliques, the so-called **join tree**
- For evidence propagation a join tree is enhanced by so-ca the edges, which are intersection of the connected nodes -

Graph Triangulation

Algorithm: (graph triangulation)

Input: An undirected graph $G = (V, E)$.
Output: A triangulated undirected graph $G' = (V, E')$ with

1. Compute an ordering of the nodes of the graph using *ma* *search*, i.e., number the nodes from 1 to $n = |V|$, in incr assigning the next number to the node having the large numbered neighbors (breaking ties arbitrarily).

2. From $i = n$ to $i = 1$ recursively fill in edges between any no of the node numbered i having lower ranks than i (including the node numbered i in previous steps). If no edges are add graph is chordal; otherwise the new graph is chordal.

Join Tree Construction

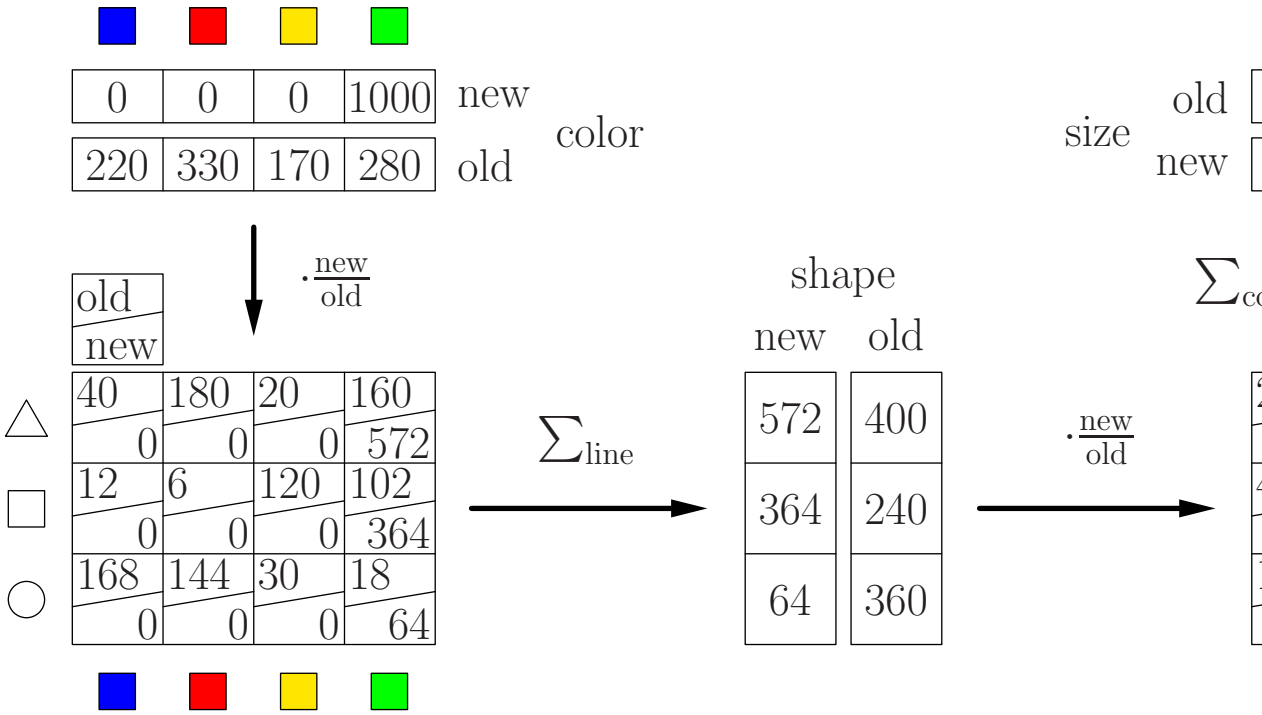
Algorithm: (join tree construction)

Input: A triangulated undirected graph $G = (V, E)$.
Output: A join tree $G' = (V', E')$ for G.

1. Determine a numbering of the nodes of G using maximum
2. Assign to each clique the maximum of the ranks of its node
3. Sort the cliques in ascending order w.r.t. the numbers assig
4. Traverse the cliques in ascending order and for each clique cliques $C_1, \ldots, C_{i-1}$ preceding it the clique with which it ha of nodes in common (breaking ties arbitrarily).

Reasoning in Join/Junction Tre

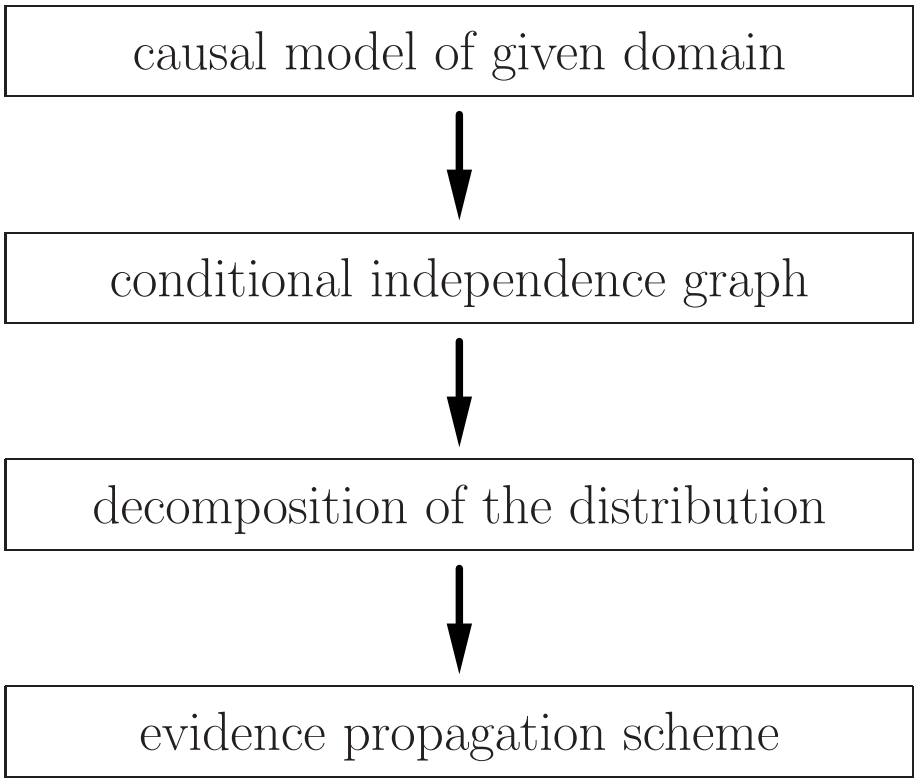
- Reasoning in join trees follows the same lines as shown in
- Multiple pieces of evidence from different branches may b a distribution before continuing by summing/marginalizin



Graphical Models:
Manual Model Building

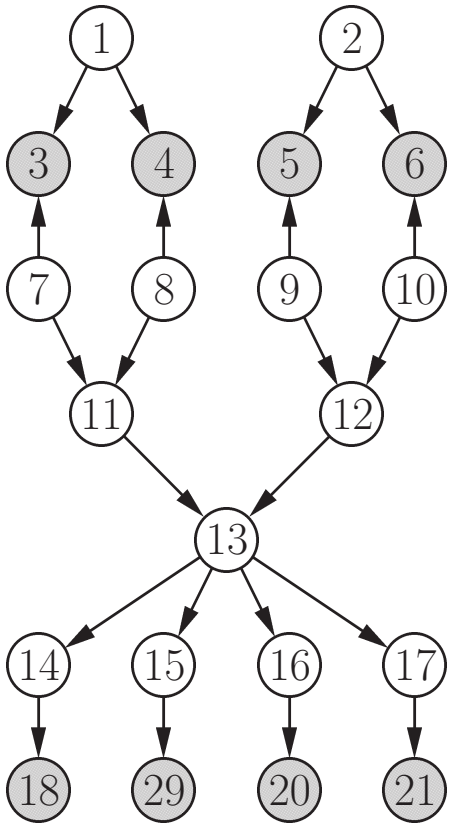
Building Graphical Models: Causal M

Manual creation of a reasoning system based on a



- Problem: strong assumptions about the statistical effects

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
12
13
14
15
16
17
18
19
20
21

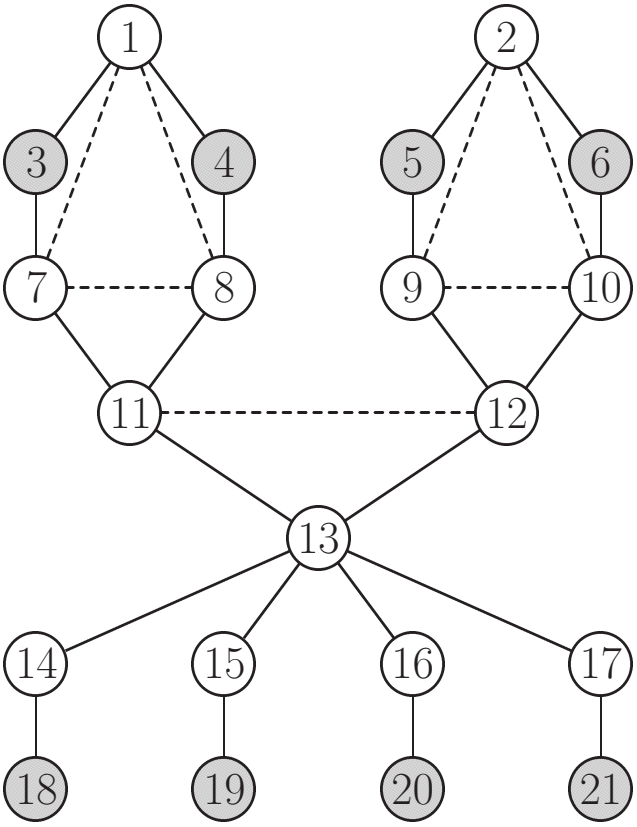
The grey nodes correspond to obs

Danish Jersey Cattle Blood Type Dete

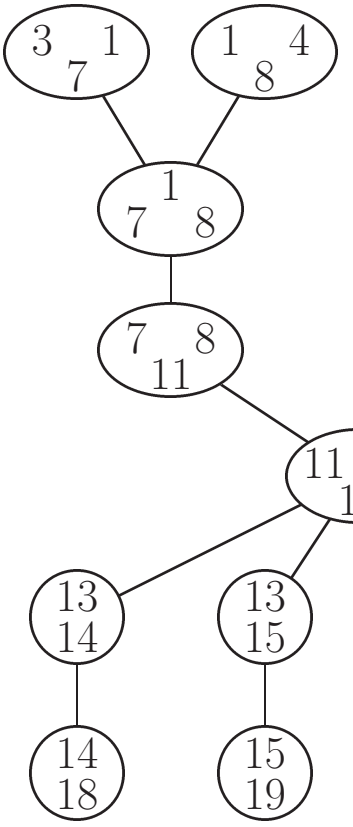
- Full 21-dimensional domain has $2^6 \cdot 3^{10} \cdot 6 \cdot 8^4 = 92\ 876\ 046$
- Bayesian network requires only 306 conditional probabilit
- Example of a conditional probability table (attributes 2, 9

sire correct	true sire phenogroup 1	stated sire phenogrou F1	V1	V2
yes	F1	1	0	0
yes	V1	0	1	0
yes	V2	0	0	1
no	F1	0.58	0.10	0.3
no	V1	0.58	0.10	0.3
no	V2	0.58	0.10	0.3

Danish Jersey Cattle Blood Type Dete



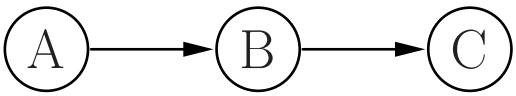
moral graph



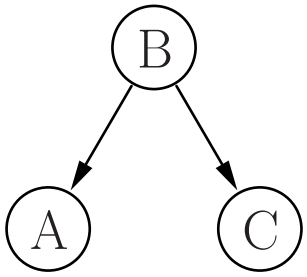
join

Graphical Models and Causali

Graphical Models and Causali



causal chain



common cause

Example:

A – accelerator pedal
B – fuel supply
C – engine speed

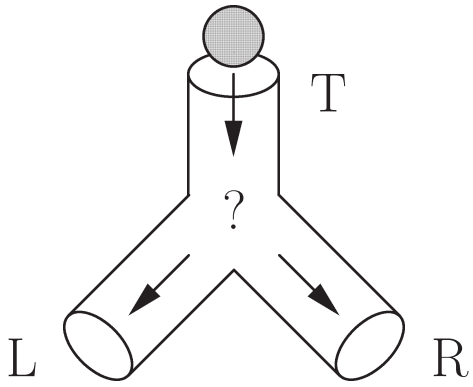
$A \not\perp\!\!\!\perp C \mid \emptyset$
$A \perp\!\!\!\perp C \mid B$

Example:

A – ice cream sales
B – temperature
C – bathing accidents

$A \not\perp\!\!\!\perp C \mid \emptyset$
$A \perp\!\!\!\perp C \mid B$

Common Cause Assumption (Causal Marko



t	r	$\overline{r}$	Σ
l	0	$1/2$	$1/2$
$\overline{l}$	$1/2$	0	$1/2$
Σ	$1/2$	$1/2$	

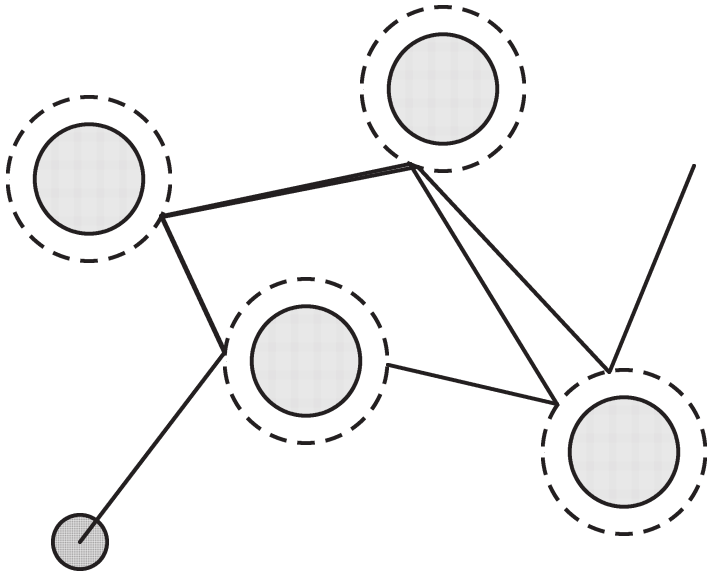
Y-shaped tube arrangement int
dropped (T). Since the ball can
at the left outlet (L) *or* the rig
corresponding variables are dep

Counter argument: The cause is
scribed. If the exact shape, posi
of the ball and the tubes are k
can be determined and the vari
dependent.

Counter counter argument: Qua
states that location and momen
cannot both at the same time b
arbitrary precision.

Sensitive Dependence on the Initial C

- *Sensitive dependence on the initial conditions* means of the initial conditions (e.g. a change of the initial posi particle) causes a deviation that grows *exponentially* wit
- Many physical systems show, for arbitrary initial condit pendence on the initial conditions. Due to this quantun sometimes have macroscopic consequences.



Example: Billi
(or generally con

Initial imprecisio

after four collisio

Learning Graphical Models from

Learning Graphical Models from

Given: A database of sample cases from a domain of intere

Desired: A (good) graphical model of the domain of interest.

- **Quantitative or Parameter Learning**
  - The structure of the conditional independence graph i
  - Conditional or marginal distributions have to be est statistical methods. (*parameter estimation*)

- **Qualitative or Structural Learning**
  - The structure of the conditional independence graph i
  - A good graph has to be selected from the set of all po (*model selection*)
  - Tradeoff between model complexity and model accura

Danish Jersey Cattle Blood Type Dete

A fraction of the database of sample cases:

```
y y f1 v2 f1 v2 f1 v2 f1 v2 v2 v2 v2v2 n y
y y f1 v2 ** ** f1 v2 ** ** ** ** f1v2 y y
y y f1 v2 f1 f1 f1 v2 f1 f1 f1 f1 f1f1 y y
y y f1 v2 f1 f1 f1 v2 f1 f1 f1 f1 f1f1 y y
y y f1 v2 f1 v1 f1 v2 f1 v1 v2 f1 f1v2 y y
y y f1 f1 ** ** f1 f1 ** ** f1 f1 f1f1 y y
y y f1 v1 ** ** f1 v1 ** ** v1 v2 v1v2 n y
y y f1 v2 f1 v1 f1 v2 f1 v1 f1 v1 f1v1 y y
                ⋮                       ⋮
```

- 21 attributes
- 500 real world sample cases
- A lot of missing values (indicated by **)

Learning Graphical Models from
Learning the Parameters

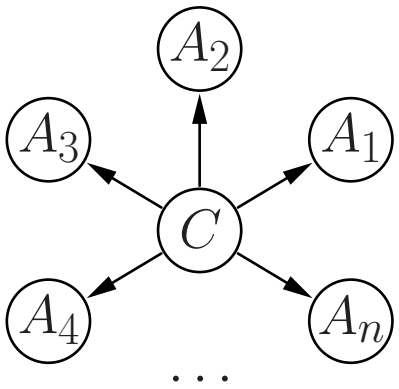
Learning the Parameters of a Graphic

Given: A database of sample cases from a domain of intere
The graph underlying a graphical model for the dor

Desired: Good values for the numeric parameters of the mod

Example: Naive Bayes Classifiers

- A naive Bayes classifier is a Bayesian network with a star-
- The class attribute is the only unconditioned attribute.
- All other attributes are conditioned on the class only.



The structure of a naive Bayes
once the attributes have been se
remaining task is to estimate th
the needed probability distribut

Probabilistic Classification

- A classifier is an algorithm that assigns a class from a pre
  or object, based on the values of descriptive attributes.
- An optimal classifier maximizes the probability of a corre
  - Let C be a class attribute with $\mathrm{dom}(C) = \{c_1, \ldots, c_n$
    which occur with probabilities p_i, $1 \leq i \leq n_C$.
  - Let q_i be the probability with which a classifier assign
    ($q_i \in \{0, 1\}$ for a deterministic classifier)
  - The probability of a correct assignment is

$$P(\text{correct assignment}) = \sum_{i=1}^{n_C} p_i q_i.$$

  - Therefore the best choice for the q_i is

$$q_i = \begin{cases} 1, & \text{if } p_i = \max_{k=1}^{n_C} p_k, \\ 0, & \text{otherwise.} \end{cases}$$

Probabilistic Classification (contin

- Consequence: An optimal classifier should assign the **mos**
- This argument does not change if we take descriptive attr
  - Let $U = \{A_1, \ldots, A_m\}$ be a set of descriptive attribu with domains $\mathrm{dom}(A_k)$, $1 \leq k \leq m$.
  - Let $A_1 = a_1, \ldots, A_m = a_m$ be an instantiation of tributes.
  - An optimal classifier should assign the class c_i for whi

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$

$$\max_{j=1}^{n_C} P(C = c_j \mid A_1 = a_1, \ldots, A_m$$

- **Problem:** We cannot store a class (or the class probabil possible instantiation $A_1 = a_1, \ldots, A_m = a_m$ of the descr (The table size grows exponentially with the number of a

- Therefore: **Simplifying assumptions are necessary**

Bayes' Rule and Bayes' Classifi

- Bayes' rule is a formula that can be used to "invert" cond
  Let X and Y be events, $P(X) > 0$. Then

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}.$$

- Bayes' rule follows directly from the definition of conditio

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \qquad \text{and} \qquad P(X \mid Y)$$

- Bayes' classifiers: Compute the class probabilities as

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$

$$\frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(}{P(A_1 = a_1, \ldots, A_m = a_m)}$$

- Looks unreasonable at first sight: Even more probabilities

Naive Bayes Classifiers

Naive Assumption:

The descriptive attributes are conditionally independent gi

Bayes' Rule:

$$P(C = c_i \mid \vec{a}) = \frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C =}{P(A_1 = a_1, \ldots, A_m = a_m)}$$

Chain Rule of Probability:

$$P(C = c_i \mid \vec{a}) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid A_1 = a_1, \ldots,$$

$c_i)$

Conditional Independence Assumption:

$$P(C = c_i \mid \vec{a}) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid C = c_i)$$

Naive Bayes Classifiers (continu

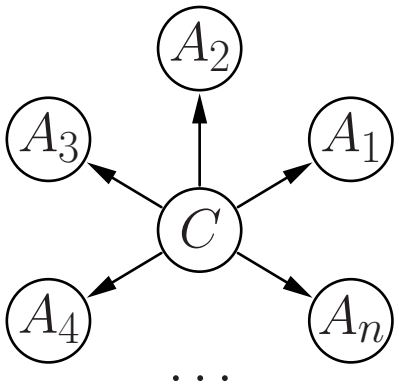
Consequence: Manageable amount of data to store.
Store distributions $P(C = c_i)$ and $\forall 1 \leq j \leq m : P(A_j = a_j \mid$

Classification: Compute for all classes c_i

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) \cdot p_0 \quad = \quad P(C = c_i) \cdot \prod_{j=1}^{n}$$

and predict the class c_i for which this value is largest.

Relation to Bayesian Networks:



Decomposition formula:

$$P(C = c_i, A_1 = a_1, \ldots, A_n$$

$$= \quad P(C = c_i) \cdot \prod_{j=1}^{n} P(A_j$$

Naive Bayes Classifiers: Parameter Es

Estimation of Probabilities:

- **Nominal/Categorical Attributes:**

$$\hat{P}(A_j = a_j \mid C = c_i) = \frac{\#(A_j = a_j, C = c_i}{\#(C = c_i) + n_A}$$

  $\#(\varphi)$ is the number of example cases that satisfy the cond
  n_{A_j} is the number of values of the attribute A_j.

- γ is called **Laplace correction**.

  $\gamma = 0$: Maximum likelihood estimation.

  Common choices: $\gamma = 1$ or $\gamma = \frac{1}{2}$.

- Laplace correction helps to avoid problems with attribute that do not occur with some class in the given data.

  It also introduces a bias towards a uniform distribution.

Naive Bayes Classifiers: Parameter Es

Estimation of Probabilities:

- **Metric/Numeric Attributes:** Assume a normal dist

$$P(A_j = a_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_j(c_i)} \exp\left(-\frac{(a_j -}{2}\right.$$

- Estimate of mean value

$$\hat{\mu}_j(c_i) = \frac{1}{\#(C = c_i)} \sum_{k=1}^{\#(C=c_i)} a_j(k)$$

- Estimate of variance

$$\hat{\sigma}_j^2(c_i) = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} \Big(a_j(k) - \hat{\mu}_j(c_i)\Big)^2$$

$\xi = \#(C = c_i)$: Maximum likelihood estimation
$\xi = \#(C = c_i) - 1$: Unbiased estimation

Naive Bayes Classifiers: Simple Exa

No	Sex	Age	Blood pr.	Drug
1	male	20	normal	A
2	female	73	normal	B
3	female	37	high	A
4	male	33	low	B
5	female	48	high	A
6	male	29	normal	A
7	female	52	normal	B
8	male	42	low	B
9	male	61	normal	B
10	female	30	normal	A
11	female	26	low	B
12	male	54	high	A

$P(\text{Drug})$

$P(\text{Sex} \mid \text{Drug})$
male
female

$P(\text{Age} \mid \text{Drug})$
μ
σ^2

$P(\text{Blood Pr.} \mid \text{D}$
low
normal
high

A simple database and estimated (conditional) probabilit

Naive Bayes Classifiers: Simple Exa

$$
\begin{aligned}
&P(\text{Drug A} \mid \text{male, 61, normal}) \\
&\quad = \quad c_1 \cdot P(\text{Drug A}) \cdot P(\text{male} \mid \text{Drug A}) \cdot P(61 \mid \text{Drug A}) \cdot P \\
&\quad \approx \quad c_1 \cdot 0.5 \cdot 0.5 \cdot 0.004787 \cdot 0.5 \quad = \quad c_1 \cdot 5.984 \cdot 10^{-4} \quad = \quad 0
\end{aligned}
$$

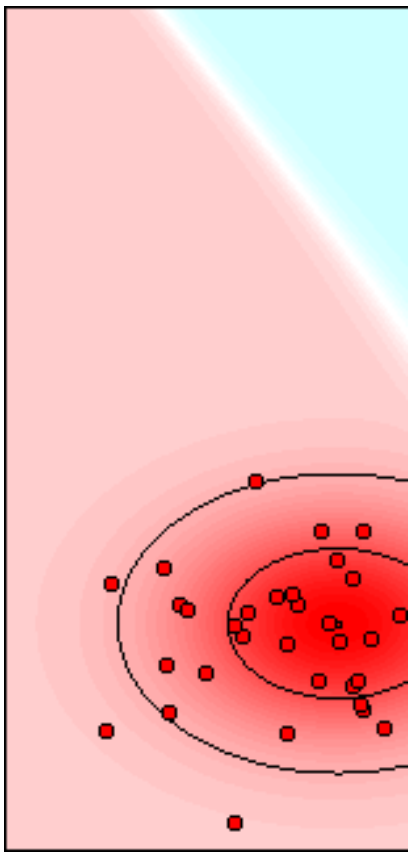
$$
\begin{aligned}
&P(\text{Drug A} \mid \text{male, 61, normal}) \\
&\quad = \quad c_1 \cdot P(\text{Drug B}) \cdot P(\text{male} \mid \text{Drug B}) \cdot P(61 \mid \text{Drug B}) \cdot P(\\
&\quad \approx \quad c_1 \cdot 0.5 \cdot 0.5 \cdot 0.017120 \cdot 0.5 \quad = \quad c_1 \cdot 2.140 \cdot 10^{-3} \quad = \quad 0
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{Drug A} \mid \text{female, 30, normal}) \\
&\quad = \quad c_2 \cdot P(\text{Drug A}) \cdot P(\text{female} \mid \text{Drug A}) \cdot P(30 \mid \text{Drug A}) \cdot \\
&\quad \approx \quad c_2 \cdot 0.5 \cdot 0.5 \cdot 0.027703 \cdot 0.5 \quad = \quad c_2 \cdot 3.471 \cdot 10^{-3} \quad = \quad 0
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{Drug A} \mid \text{female, 30, normal}) \\
&\quad = \quad c_2 \cdot P(\text{Drug B}) \cdot P(\text{female} \mid \text{Drug B}) \cdot P(30 \mid \text{Drug B}) \cdot \\
&\quad \approx \quad c_2 \cdot 0.5 \cdot 0.5 \cdot 0.013567 \cdot 0.5 \quad = \quad c_2 \cdot 1.696 \cdot 10^{-3} \quad = \quad 0
\end{aligned}
$$

Naive Bayes Classifiers: Simple Exa

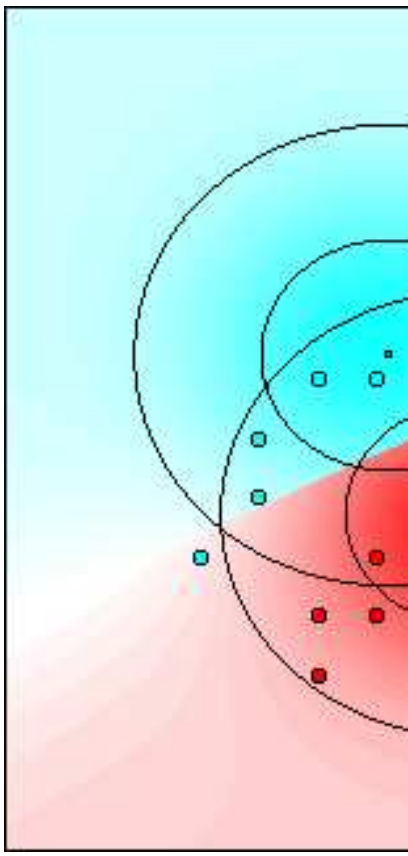
- 100 data points, 2 classes
- Small squares: mean values
- Inner ellipses:
  one standard deviation
- Outer ellipses:
  two standard deviations
- Classes overlap:
  classification is not perfect



Naive Bay

Naive Bayes Classifiers: Simple Exa

- 20 data points, 2 classes
- Small squares: mean values
- Inner ellipses:
  one standard deviation
- Outer ellipses:
  two standard deviations
- Attributes are not conditionally
  independent given the class.



Naive Bay

Naive Bayes Classifiers: Iris Da

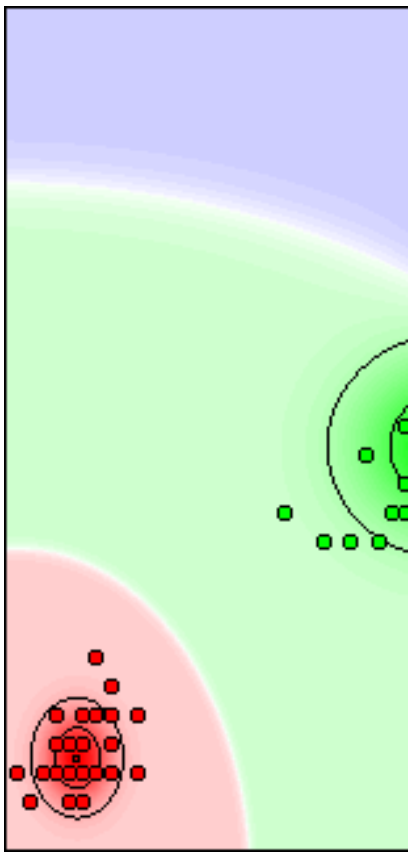
- 150 data points, 3 classes

  Iris setosa (red)
  Iris versicolor (green)
  Iris virginica (blue)

- Shown: 2 out of 4 attributes

  sepal length
  sepal width
  petal length (horizontal)
  petal width (vertical)

- 6 misclassifications
  on the training data
  (with all 4 attributes)



Naive Bay

Learning Graphical Models from
Learning the Structure

Learning the Structure of Graphical Mod

- **Test whether a distribution is decomposable w.r**

  This is the most direct approach. It is not bound to a grap
  but can also be carried out w.r.t. other representations of
  to be used to compute the (candidate) decomposition of th

- **Find a suitable graph by measuring the strength**

  This is a heuristic, but often highly successful approach
  the frequently valid assumption that in a conditional ind
  attribute is more strongly dependent on adjacent attribute
  that are not directly connected to them.

- **Find an independence map by conditional inde**

  This approach exploits the theorems that connect condi
  graphs and graphs that represent decompositions. It has t
  single conditional independence test, if it fails, can exclu
  graphs. However, wrong test results can thus have severe

Evaluation Measures and Search M

- All learning algorithms for graphical models consist of an **evaluation measure** or **scoring function**, e.g.
  - Hartley information gain (relational networks)
  - Shannon information gain, χ^2, K2 metric (probabilisti

  and a (heuristic) **search method**, e.g.
  - conditional independence search
  - greedy search (spanning tree or K2 algorithm)
  - guided random search (simulated annealing, genetic a

- An exhaustive search over all graphs is too expensive:
  - $2^{\binom{n}{2}}$ possible undirected graphs for n attributes.
  - $f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i)$ possible directe

Learning the Structure of a Graphica
Testing for Decomposability

Comparing Relations

- In order to evaluate a graph structure, we need a measur actual relation to the relatio represented by the graph.
- For arbitrary R, E_1, and E_2 it is

$$R(E_1 \cap E_2) \leq \min\{R(E_1), R(E_2)\}.$$

- This relation entails that it is always

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$r_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) \leq \min_{M \in \mathcal{M}} \Big\{ r_M\Big(\bigwedge_{A_i \in M} A$$

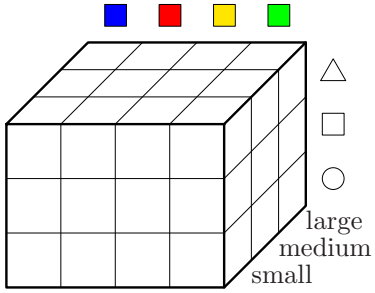
- Therefore: Measure the quality of a family $\mathcal{M}$ of subset c

$$\sum_{a_1 \in \mathrm{dom}(A_1)} \cdots \sum_{a_n \in \mathrm{dom}(A_n)} \Big(\min_{M \in \mathcal{M}} \Big\{ r_M\Big(\bigwedge_{A_i \in M} A_i = a_i\Big)\Big\} -$$

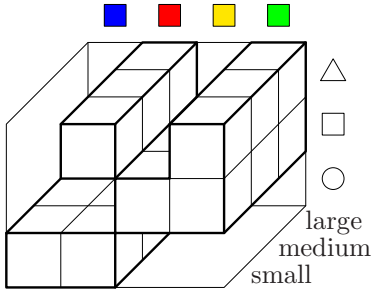
  Intuitively: Count the number of additional tuples.

Direct Test for Decomposability: Re

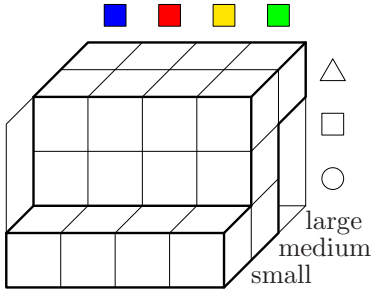
1.



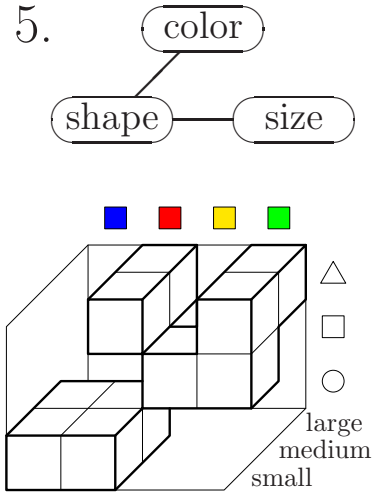
2.



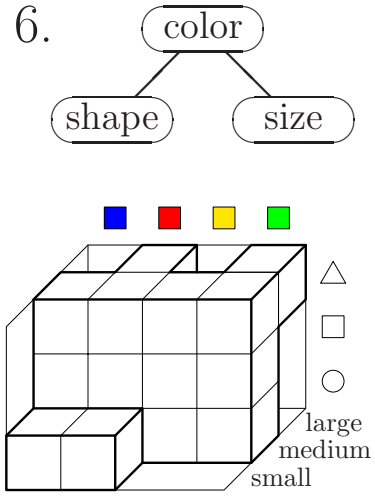
3.



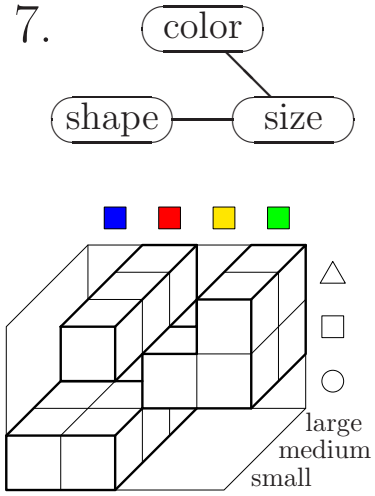
5.



6.



7.



Comparing Probability Distribut

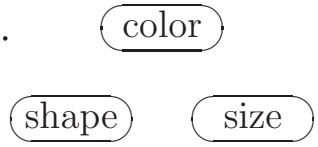
Definition: Let p_1 and p_2 be two strictly positive probability
same set $\mathcal{E}$ of events. Then

$$I_{\mathrm{KLdiv}}(p_1, p_2) = \sum_{E \in \mathcal{E}} p_1(E) \log_2 \frac{p_1(E)}{p_2(E)}$$

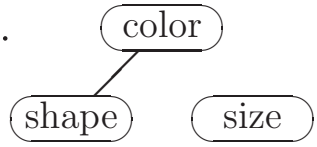
is called the **Kullback-Leibler information divergence**

- The Kullback-Leibler information divergence is non-negat
- It is zero if and only if $p_1 \equiv p_2$.
- Therefore it is plausible that this measure can be used to
  the approximation of a given multi-dimensional distribut
  bution p_2 that is represented by a given graph:
  The smaller the value of this measure, the better the appr

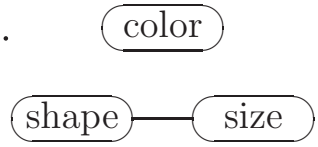
Direct Test for Decomposability: Pro

1. 

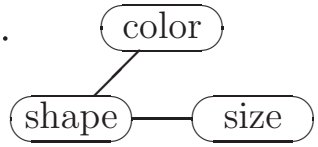
0.640
−5041

2. 

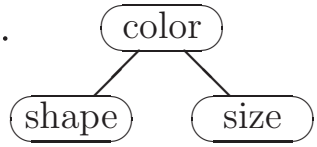
0.211
−4612

3. 

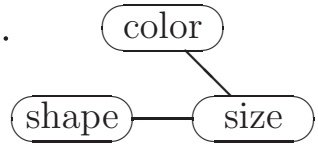
0.429
−4830

5. 

0
−4401

6. 

0.161
−4563

7. 

0.379
−4780

Upper numbers: The Kullback-Leibler information diverge
distribution and its approximation.

Lower numbers: The binary logarithms of the probability of a
(log-likelihood of data).

Learning the Structure of a Graphica
Strength of Marginal Dependen

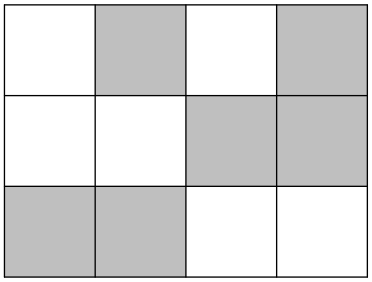
Strength of Marginal Dependences: R

- Learning a relational network consists in finding those sub intersection of the cylindrical extensions of the projection approximates best the set of possible world states, i.e. conta tuples as possible.
- Since computing explicitly the intersection of the cylindri projections and comparing it to the original relation is evaluation functions are used, for instance:

subspace	color × shape	shape × size
possible combinations	12	9
occurring combinations	6	5
relative number	50%	56%

- The relational network can be obtained by interpreting t as edge weights and constructing the minimum weight spa

Strength of Marginal Dependences: R



Hartley information needed to det

coordinates:	$\log_2 4 + \log_2 3 =$
coordinate pair:	$\log_2 6$
gain:	$\log_2 12 - \log_2 6$

Definition: Let A and B be two attributes and R a discrete
with $\exists a \in \mathrm{dom}(A) : \exists b \in \mathrm{dom}(B) : R(A = a, B = b) = 1$. Th

$$
\begin{aligned}
I_{\mathrm{gain}}^{(\mathrm{Hartley})}(A, B) &= \log_2 \Big(\textstyle\sum_{a \in \mathrm{dom}(A)} R(A = a) \Big) + \log_2 \Big(\textstyle\sum_b \\
&- \log_2 \Big(\textstyle\sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} R(A = a, \\
&= \log_2 \frac{\Big(\sum_{a \in \mathrm{dom}(A)} R(A = a) \Big) \cdot \Big(\sum_{b \in \mathrm{dom}(}}{\sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} R(A =}
\end{aligned}
$$

is called the **Hartley information gain** of A and B w.r.t.

Strength of Marginal Dependences: Simp

- **Intuitive interpretation of Hartley information**
  The binary logarithm measures the number of questions
  ing value with a scheme like a binary search. Thus Hartl
  measures the reduction in the number of necessary questi

- Results for the simple example:

$$I_{\text{gain}}^{(\text{Hartley})}(\text{color}, \text{shape}) = 1.00 \text{ bit}$$
$$I_{\text{gain}}^{(\text{Hartley})}(\text{shape}, \text{size}) \approx 0.86 \text{ bit}$$
$$I_{\text{gain}}^{(\text{Hartley})}(\text{color}, \text{size}) \approx 0.58 \text{ bit}$$

- Applying the Kruskal algorithm yields as a learning result

  As we know, this graph describes indeed a decomposition

Strength of Marginal Dependences: Pr

Mutual Information / Cross Entropy / Informatio

Based on Shannon Entropy $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ (Shanno

$$I_{\text{gain}}(A, B) = \overbrace{H(A)}^{} - \overbrace{H(A}^{}$$

$$= \overbrace{-\sum_{i=1}^{n_A} p_{i.} \log_2 p_{i.}}^{} - \overbrace{\sum_{j=1}^{n_B} p_{.j} \left(-\sum_{i=1}^{n_A} p}^{}$$

$H(A)$	Entropy of the distribution on attribute A
$H(A\|B)$	*Expected entropy* of the distribution on A if the value of attribute B becomes know
$H(A) - H(A\|B)$	Expected reduction in entropy or *inform*

Interpretation of Shannon Entro

- Let $S = \{s_1, \ldots, s_n\}$ be a finite set of alternatives having p $P(s_i)$, $i = 1, \ldots, n$, satisfying $\sum_{i=1}^{n} P(s_i) = 1$.
- **Shannon Entropy:**

$$H(S) = -\sum_{i=1}^{n} P(s_i) \log_2 P(s_i)$$

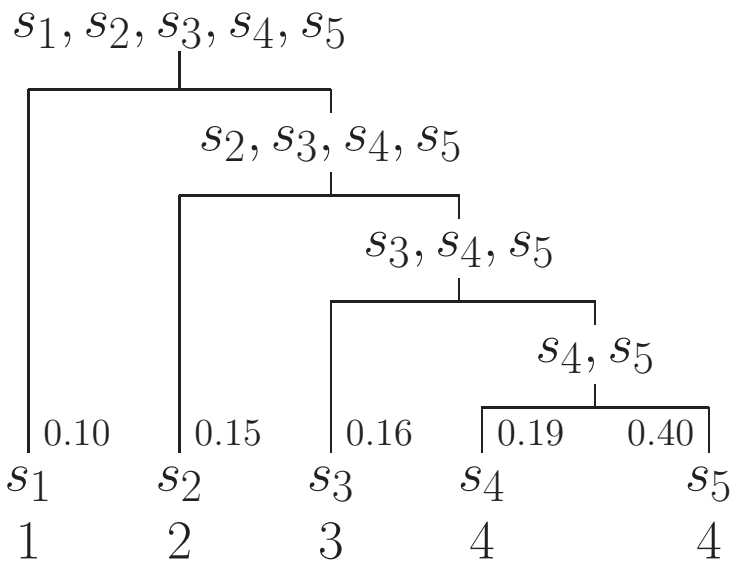
- Intuitively: **Expected number of yes/no questions asked in order to determine the obtaining alter**
  - Suppose there is an oracle, which knows the obtaining but responds only if the question can be answered wit
  - A better question scheme than asking for one alterna can easily be found: Divide the set into two subsets of
  - Ask for containment in an arbitrarily chosen subset.
  - Apply this scheme recursively → number of questions b

Question/Coding Schemes

$P(s_1) = 0.10, \quad P(s_2) = 0.15, \quad P(s_3) = 0.16, \quad P(s_4) = 0.$

Shannon entropy: $-\sum_i P(s_i) \log_2 P(s_i) = 2.15$ b

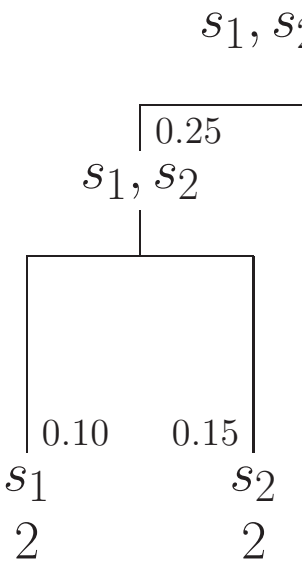
Linear Traversal



Code length: 3.24 bit/symbol

Code efficiency: 0.664

Equal S



Code length:

Code efficienc

Question/Coding Schemes

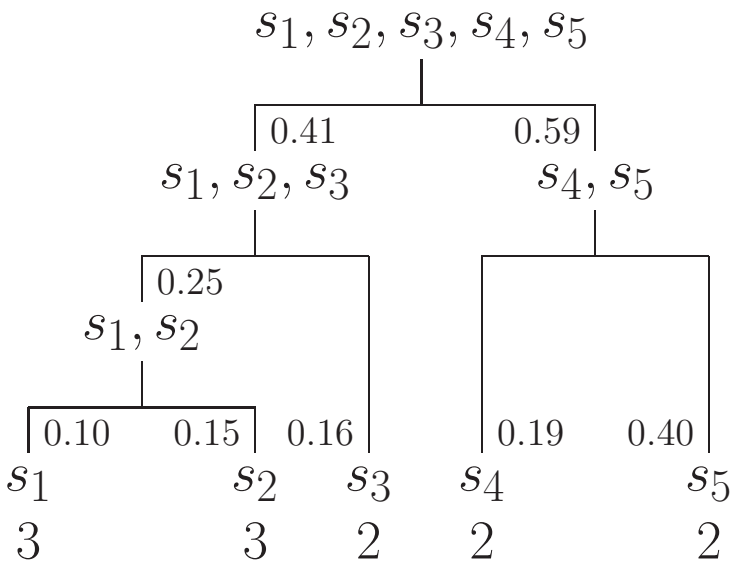
- Splitting into subsets of about equal size can lead to a bad
  alternatives into subsets → high expected number of ques
- Good question schemes take the probability of the alterna
- **Shannon-Fano Coding** (1948)
  - Build the question/coding scheme top-down.
  - Sort the alternatives w.r.t. their probabilities.
  - Split the set so that the subsets have about equal *pro*
    (splits must respect the probability order of the altern
- **Huffman Coding** (1952)
  - Build the question/coding scheme bottom-up.
  - Start with one element sets.
  - Always combine those two sets that have the smallest

Question/Coding Schemes

$P(s_1) = 0.10, \quad P(s_2) = 0.15, \quad P(s_3) = 0.16, \quad P(s_4) = 0.$

Shannon entropy: $-\sum_i P(s_i) \log_2 P(s_i) = 2.15$ b

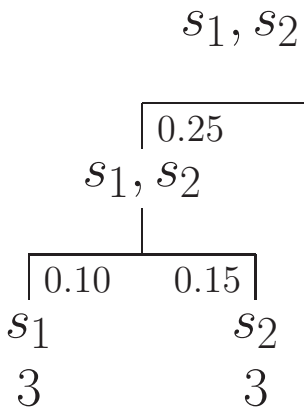
Shannon–Fano Coding (1948)



Code length: 2.25 bit/symbol

Code efficiency: 0.955

Huffman



Code length:

Code efficienc

Question/Coding Schemes

- It can be shown that Huffman coding is optimal if we ha
  obtaining alternative in a single instance.
  (No question/coding scheme has a smaller expected numb
- Only if the obtaining alternative has to be determined in
  pendent) situations, this scheme can be improved upon.
- Idea: Process the sequence not instance by instance, but
  or more consecutive instances and ask directly for the ob
  of alternatives.
- Although this enlarges the question/coding scheme, the
  questions per identification is reduced (because each interro
  obtaining alternative for several situations).
- However, the expected number of questions per identificat
  arbitrarily small. Shannon showed that there is a lower
  Shannon entropy.

Interpretation of Shannon Entro

$$P(s_1) = \tfrac{1}{2}, \quad P(s_2) = \tfrac{1}{4}, \quad P(s_3) = \tfrac{1}{8}, \quad P(s_4) = \tfrac{1}{16},$$

Shannon entropy: $-\sum_i P(s_i) \log_2 P(s_i) = 1.875$ b

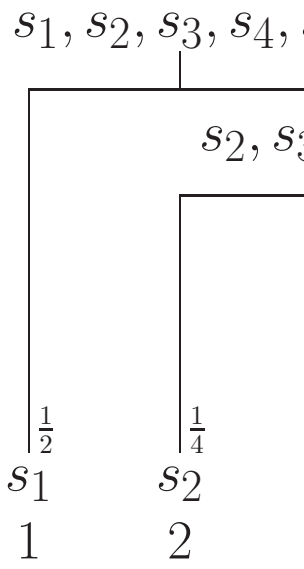
If the probability distribution allows for a perfect Huffman code (code efficiency 1), the Shannon entropy can easily be interpreted as follows:

$$-\sum_i P(s_i) \log_2 P(s_i)$$

$$= \sum_i \underbrace{P(s_i)}_{\substack{\text{occurrence} \\ \text{probability}}} \cdot \underbrace{\log_2 \frac{1}{P(s_i)}}_{\substack{\text{path length} \\ \text{in tree}}} .$$

In other words, it is the expected number of needed yes/no questions.

Perfect Qu



Code length:

Code efficien

Information Gain: Simple Exam

projection to subspace

△	40	180	20	160
□	12	6	120	102
○	168	144	30	18

	s	m	l
△	20	180	200
□	40	160	40
○	180	120	60

large	50	115	35	100
medium	82	133	99	146
small	88	82	36	34

product of marginals

△	88	132	68	112
□	53	79	41	67
○	79	119	61	101

	s	m	l
△	96	184	120
□	58	110	72
○	86	166	108

large	66	99	51	84
medium	101	152	78	129
small	53	79	41	67

Strength of Marginal Dependences: Simp

- Results for the simple example:

$$
\begin{aligned}
I_{\text{gain}}(\text{color}, \text{shape}) &= 0.429 \text{ bit} \\
I_{\text{gain}}(\text{shape}, \text{size}) &= 0.211 \text{ bit} \\
I_{\text{gain}}(\text{color}, \text{size}) &= 0.050 \text{ bit}
\end{aligned}
$$

- Applying the Kruskal algorithm yields as a learning result

- It can be shown that this approach always yields the bes tree w.r.t. Kullback-Leibler information divergence (Chow

- In an extended form this also holds for certain classes of g (for example, tree-augmented naive Bayes classifiers).

- For more complex graphs, the best graph need not be fou (there are counterexamples, see below).

Strength of Marginal Dependences: Gener

- **Optimum Weight Spanning Tree Construction**
  - Compute an evaluation measure on all possible edges (two-dimensional subspaces).
  - Use the Kruskal algorithm to determine an optimum w

- **Greedy Parent Selection** (for directed graphs)
  - Define a topological order of the attributes (to restrict
  - Compute an evaluation measure on all single attribute
  - For each preceding attribute (w.r.t. the topological or add it as a candidate parent to the hyperedge and compute the evaluation measure again.
  - Greedily select a parent according to the evaluation m
  - Repeat the previous two steps until no improvement r

Another Probabilistic Evaluation Measur

- Idea: Compute the probability of a graph given the data

$$P(\vec{G} \mid D) = \frac{1}{P(D)} \int_{\Theta} P(D \mid \vec{G}, \Theta) f(\Theta \mid \vec{G}) F$$

  $\vec{G}$ directed acyclic graph underlying the graphical mode

  Θ probability parameters of the graphical model

  D database to learn from

- In order to compare two graphs, it is sufficient to compute

$$\frac{P(\vec{G}_1 \mid D)}{P(\vec{G}_2 \mid D)} = \frac{P(\vec{G}_1, D)}{P(\vec{G}_2, D)} = \frac{\int_{\Theta_1} P(D \mid \vec{G}_1, \Theta_1) f(\Theta_1 \mid}{\int_{\Theta_2} P(D \mid \vec{G}_2, \Theta_2) f(\Theta_2 \mid}$$

  In this way one can avoid computing the probability $P(D$
  Assuming equal probability of all graphs simplifies further

Another Probabilistic Evaluation Measur

- Assumptions about data and parameter independence yie

$$P(\vec{G}, D) = \gamma \prod_{k=1}^{r} \prod_{j=1}^{m_k} \int_{\theta_{ijk}} \cdots \int \left(\prod_{i=1}^{n_k} \theta_{ijk}^{N_{ijk}} \right) f(\theta_{1jk}, \ldots, \theta_{n_k jk}$$

r	number of attributes describing the domain under
n_k	number of values of the k-th attribute A_k, i.e., n_k
m_k	number of instantiations of the parents of the k-th i.e., $m_k = \prod_{A_j \in \text{parents}(A_k)} n_j = \prod_{A_j \in \text{parents}(A_k)}$ |
θ_{ijk}	probability that the k-th attribute takes its i-th va its parents in $\vec{G}$ take their j-th instantiation
N_{ijk}	number of sample cases in which the k-th attribut and its parents in $\vec{G}$ have their j-th instantiation
γ	normalization factor

Another Probabilistic Evaluation Measur

- Choose $f(\theta_{1jk}, \ldots, \theta_{n_k jk}) = \text{const.}$ [Cooper and Hersk
- Then the solution can be obtained via Dirichlet's integral

$$K_2(\vec{G}, D) = \gamma \prod_{k=1}^{r} \prod_{j=1}^{m_k} \frac{(n_k - 1)!}{(N_{\cdot jk} + n_k - 1)!} \prod_{i=1}^{n_k}$$

- Since this formula is a product over the attributes, each attribute can be handled more or less separately.
- Core ideas of the K2 algorithm:
  - Fix a topological order of the attributes. (Reduces the search space and ensures that the graph
  - Select the parents of each attribute greedily based on the K2 metric (or rather its corresponding fa

A Generalization of the K2 Met

- Choose a maximum likelihood estimation of the probabili

$$f(\theta_{1jk}, \ldots, \theta_{n_k jk}) = \prod_{i=1}^{n_k} \delta \left(\theta_{ijk} - \frac{N_{ijk}}{N_{.jk}} \right)$$

$$\Rightarrow \qquad g_\infty(\vec{G}, D) = \gamma \prod_{k=1}^{r} \prod_{j=1}^{m_k} \prod_{i=1}^{n_k} \left(\frac{N_{ijk}}{N_{.jk}} \right)^{N_{ijk}} \qquad \text{(eq in}$$

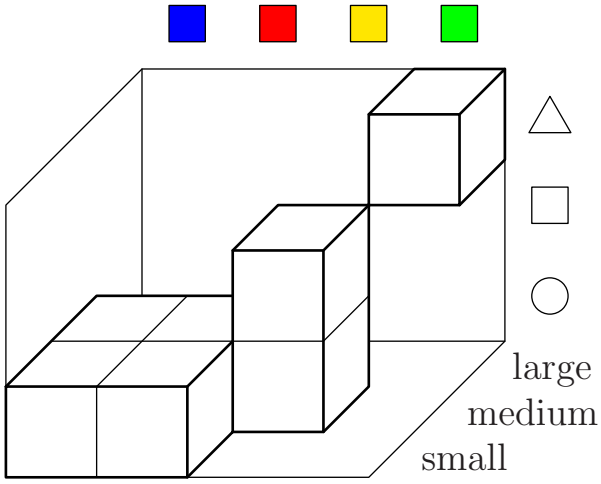
- Choose the likelihood function scaled to maximum 1 and ra

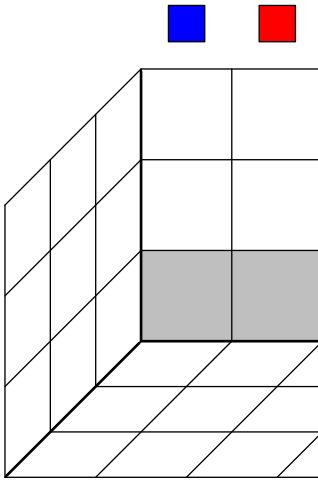
$$f_\alpha(\theta_{1jk}, \ldots, \theta_{n_k jk}) = \beta \cdot \prod_{i=1}^{n_k} \theta_{ijk}^{\alpha N_{ijk}}$$

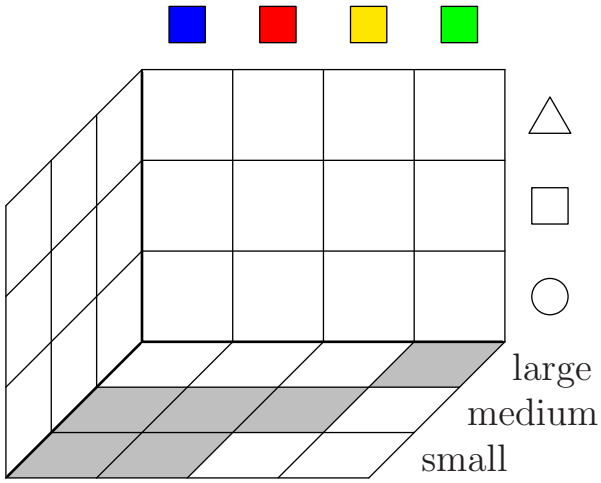
$$\Rightarrow \qquad g_\alpha(\vec{G}, D) = \gamma \prod_{k=1}^{r} \prod_{j=1}^{m_k} \frac{\Gamma(\alpha N_{.jk} + n_k)}{\Gamma((\alpha+1) N_{.jk} + n_k)} \cdot \prod_{i=1}^{n_k}$$

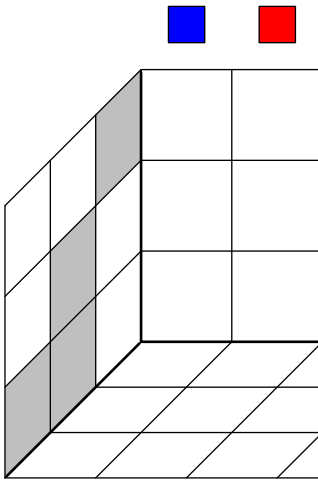
- The parameter α can be interpreted as a sensitivity para mines the strength of the tendency to select parent attrib

Strength of Marginal Dependences: D

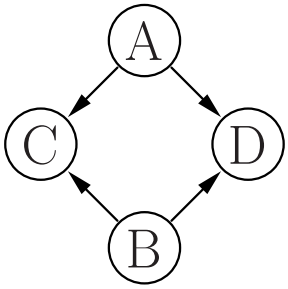








Strength of Marginal Dependences: D



p_A	a_1	a_2
	0.5	0.5

p_B	b_1	b_2
	0.5	0.5

$p_{C\|AB}$	a_1b_1
c_1	0.9
c_2	0.1

$p_{D\|AB}$	a_1b_1
d_1	0.9
d_2	0.1

p_{AD}	a_1	a_2
d_1	0.3	0.2
d_2	0.2	0.3

p_{BD}	b_1	b_2
d_1	0.3	0.2
d_2	0.2	0.3

- Greedy parent selection can lead to suboptimal results if there is more than one path connecting two attributes.
- Here: the edge $C \rightarrow D$ is selected first.

Learning the Structure of a Graphica
Conditional Independence Tes

Structure Learning with Conditional Indep

General Idea: Exploit the theorems that connect condit
graphs and graphs that represent decompositions.

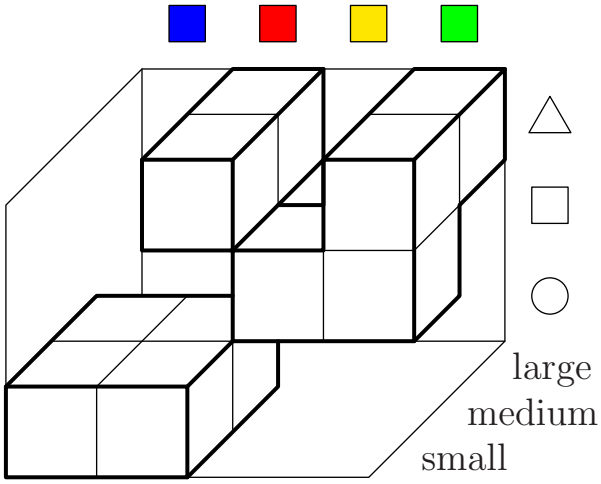
In other words: we want a graph describing a decomposition,
but we search for a conditional independence

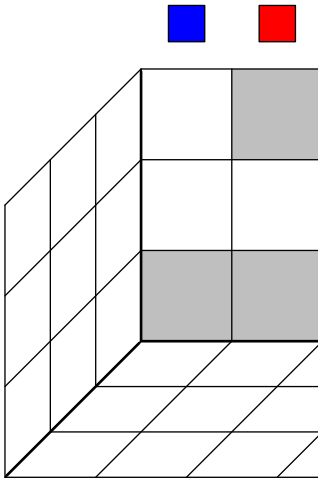
This approach has the advantage that a single conditional ind
if it fails, can exclude several candidate graphs.

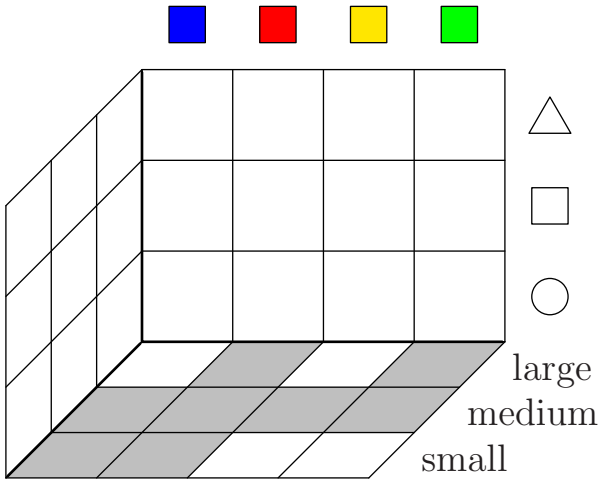
Assumptions:

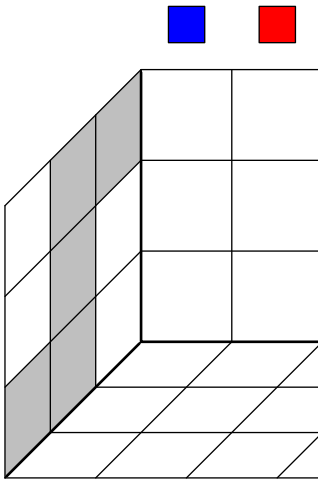
- *Faithfulness:* The domain under consideration can be a
with a graphical model (more precisely: there exists a per
- *Reliability of Tests:* The result of all conditional indepen
with the actual situation in the underlying distribution.
- Other assumptions that are specific to individual algorith

Conditional Independence Tests: Re









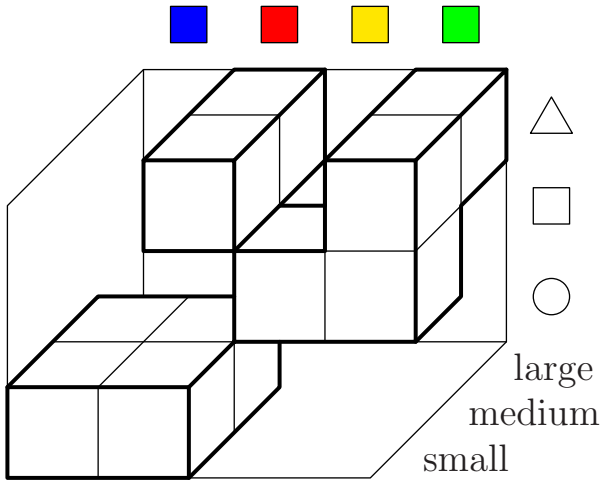
Conditional Independence Tests: Re

- The Hartley information gain can be used directly to tes **marginal independence**.

attributes	relative number of possible value combinations	Hartley inform
color, shape	$\frac{6}{3 \cdot 4} = \frac{1}{2} = 50\%$	$\log_2 3 + \log_2 4$
color, size	$\frac{8}{3 \cdot 4} = \frac{2}{3} \approx 67\%$	$\log_2 3 + \log_2 4$
shape, size	$\frac{5}{3 \cdot 3} = \frac{5}{9} \approx 56\%$	$\log_2 3 + \log_2 3$

- In order to test for (approximate) **conditional indepe**
  - Compute the Hartley information gain for each poss the conditioning attributes.
  - Aggregate the result over all possible instantiations, fo averaging them.

Conditional Independence Tests: Simpl



color	Hartley inf
■	$\log_2 1 + \log$
■	$\log_2 2 + \log$
■	$\log_2 1 + \log$
■	$\log_2 2 + \log$
	average:

shape	Hartley information gain
△	$\log_2 2 + \log_2 2 - \log_2 4 = 0$
□	$\log_2 2 + \log_2 1 - \log_2 2 = 0$
○	$\log_2 2 + \log_2 2 - \log_2 4 = 0$
	average: $= 0$

size	Hartley inf
large	$\log_2 2 + \log$
medium	$\log_2 4 + \log$
small	$\log_2 2 + \log$
	average:

Conditional Independence Tests: Simpl

- The Shannon information gain can be used directly to tes **marginal independence**.
- Conditional independence tests may be carried out by summ gain for all instantiations of the conditioning variables:

$$I_{\text{gain}}(A, B \mid C)$$
$$= \sum_{c \in \text{dom}(C)} P(c) \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} P(a, b \mid c) \log_2$$

  where $P(c)$ is an abbreviation of $P(C = c)$ etc.

- Since $I_{\text{gain}}(\text{color}, \text{size} \mid \text{shape}) = 0$ indicates the only condi we get the following learning result:

Conditional Independence Tests: Genera

Algorithm: (conditional independence graph construction)

1. For each pair of attributes A and B, search for a set S_{AB} that $A \perp\!\!\!\perp B \mid S_{AB}$ holds in $\widehat{P}$, i.e., A and B are independe on S_{AB}. If there is no such S_{AB}, connect the attributes by
2. For each pair of non-adjacent variables A and B with a co (i.e., C is adjacent to A as well as to B), check whether C
   - If it is, continue.
   - If it is not, add arrow heads pointing to C, i.e., $A \rightarrow C$
3. Recursively direct all undirected edges according to the rul
   - If for two adjacent variables A and B there is a strictly d to B not including $A \rightarrow B$, then direct the edge towards
   - If there are three variables A, B, and C with A and B n and $A \rightarrow C$, then direct the edge $C \rightarrow B$.

Conditional Independence Tests: Simpl

Suppose that the following conditional independence statemen

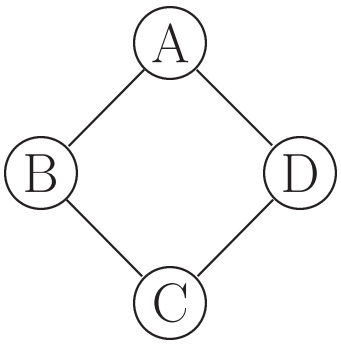
$$A \perp\!\!\!\perp_{\hat{P}} B \mid \emptyset \qquad B \perp\!\!\!\perp_{\hat{P}} A \mid \emptyset$$
$$A \perp\!\!\!\perp_{\hat{P}} D \mid C \qquad D \perp\!\!\!\perp_{\hat{P}} A \mid C$$
$$B \perp\!\!\!\perp_{\hat{P}} D \mid C \qquad D \perp\!\!\!\perp_{\hat{P}} B \mid C$$

All other possible conditional independence statements that the attributes A, B, C, and D (with single attributes on the l

- **Step 1:** Since there is no set rendering A and C, B a independent, the edges $A - C$, $B - C$, and $C - D$ are i
- **Step 2:** Since C is a common neighbor of A and B and w but $A \not\perp\!\!\!\perp_{\hat{P}} B \mid C$, the first two edges must be directed A
- **Step 3:** Since A and D are not adjacent, $C - D$ and $A -$ must be directed $C \rightarrow D$.
  (Otherwise step 2 would have already fixed the orientatio

Conditional Independence Tests: Dr

- The conditional independence graph construction algorith
  there is a **perfect map**. If there is no perfect map, the re



p_{ABCD}		$A = a_1$	
		$B = b_1$	$B = b_2$
$C = c_1$	$D = d_1$	$1/47$	$1/47$
	$D = d_2$	$1/47$	$1/47$
$C = c_2$	$D = d_1$	$1/47$	$2/47$
	$D = d_2$	$2/47$	$4/47$

- **Independence tests of high order**, i.e., with a large n
  may be necessary.
- There are approaches to mitigate these drawbacks.
  (For example, the order is restricted and all tests of highe
  to fail, if all tests of lower order failed.)

The Cheng–Bell–Liu Algorith

- **Drafting:** Build a so-called Chow–Liu tree as an initial
  - Evaluate all attribute pairs (candidate edges) with inf
  - Discard edges with evaluation below independence thr
  - Build optimum (maximum) weight spanning tree.
- **Thickening:** Add necessary edges.
  - Traverse remaining candidate edges in the order of de
  - Test for conditional independence in order to determi whether an edge is needed in the graphical model.
  - Use local Markov property to select a condition set: a conditionally independent of all non-descendants given
  - Since the graph is undirected in this step, the set of adjacent nodes is reduced iteratively and gr in order to remove possible children.

The Cheng–Bell–Liu Algorithm (con

- **Thinning:** Remove superfluous edges.
  - In the thickening phase a conditional independence te
    because the graph was still too sparse.
  - Traverse all edges that have been added to the curren
    and test for conditional independence.
  - Remove unnecessary edges.
    (two phases/approaches: heuristic test/strict test)
- **Orienting:** Direct the edges of the graphical model.
  - Identify the v-structures (converging directed edges).
    (Markov equivalence: same skeleton and same set of v
  - Traverse all pairs of attributes with common neighbo
    common neighbors are in the (maximally) reduced set
  - Direct remaining edges by extending chains and avoid

Learning Undirected Graphical Model

- **Drafting:** Build a Chow–Liu tree as an initial graphical
  - Evaluate all attribute pairs (candidate edges) with spe
  - Discard edges with evaluation below independence thr
  - Build optimum (maximum) weight spanning tree.
- **Thickening:** Add necessary edges.
  - Traverse remaining candidate edges in the order of dec
  - Test for conditional independence in order to determin whether an edge is needed in the graphical model.
  - Use local Markov property to select a condition set: a conditionally independent of any non-neighbor given i
  - Since the graphical model to be learned is undirected, *no (iterative) reduction of the condition set is need* (decisive difference to Cheng–Bell–Liu Algorithm).

Learning Undirected Graphical Model

- **Moralizing:** Take care of possible v-structures.
  - If one assumes a perfect undirected map, this step is u
    However, v-structures are too common and cannot be
    without loss in an undirected graphical model.
  - Possible v-structures can be taken care of by connecti
  - Traverse all edges with an evaluation below the indepe
    that have a common neighbor in the graph.
  - Add edge if conditional independence given the neighb
- **Thinning:** Remove superfluous edges.
  - In the thickening phase a conditional independence te
    because the graph was still too sparse.
  - Traverse all edges that have been added to the curren
    and test for conditional independence.

Learning the Structure of a Graphica
Experiments and Application

Danish Jersey Cattle Blood Type Dete

network	edges	params.	train	test
indep.	0	59	−19921.2	−20087.2
orig.	22	219	−11391.0	−11506.1

Optimum Weight Spanning Tree Construction

measure	edges	params.	train	test
I_{gain}	20.0	285.9	−12122.6	−12339.6
χ^2	20.0	282.9	−12122.6	−12336.2

Greedy Parent Selection w.r.t. a Topological Orde

measure	edges	add.	miss.	params.	train	t
I_{gain}	35.0	17.1	4.1	1342.2	−11229.3	−1181
χ^2	35.0	17.3	4.3	1300.8	−11234.9	−1180
K2	23.3	1.4	0.1	229.9	−11385.4	−1151
$L_{\text{red}}^{(\text{rel})}$	22.5	0.6	0.1	219.9	−11389.5	−1150

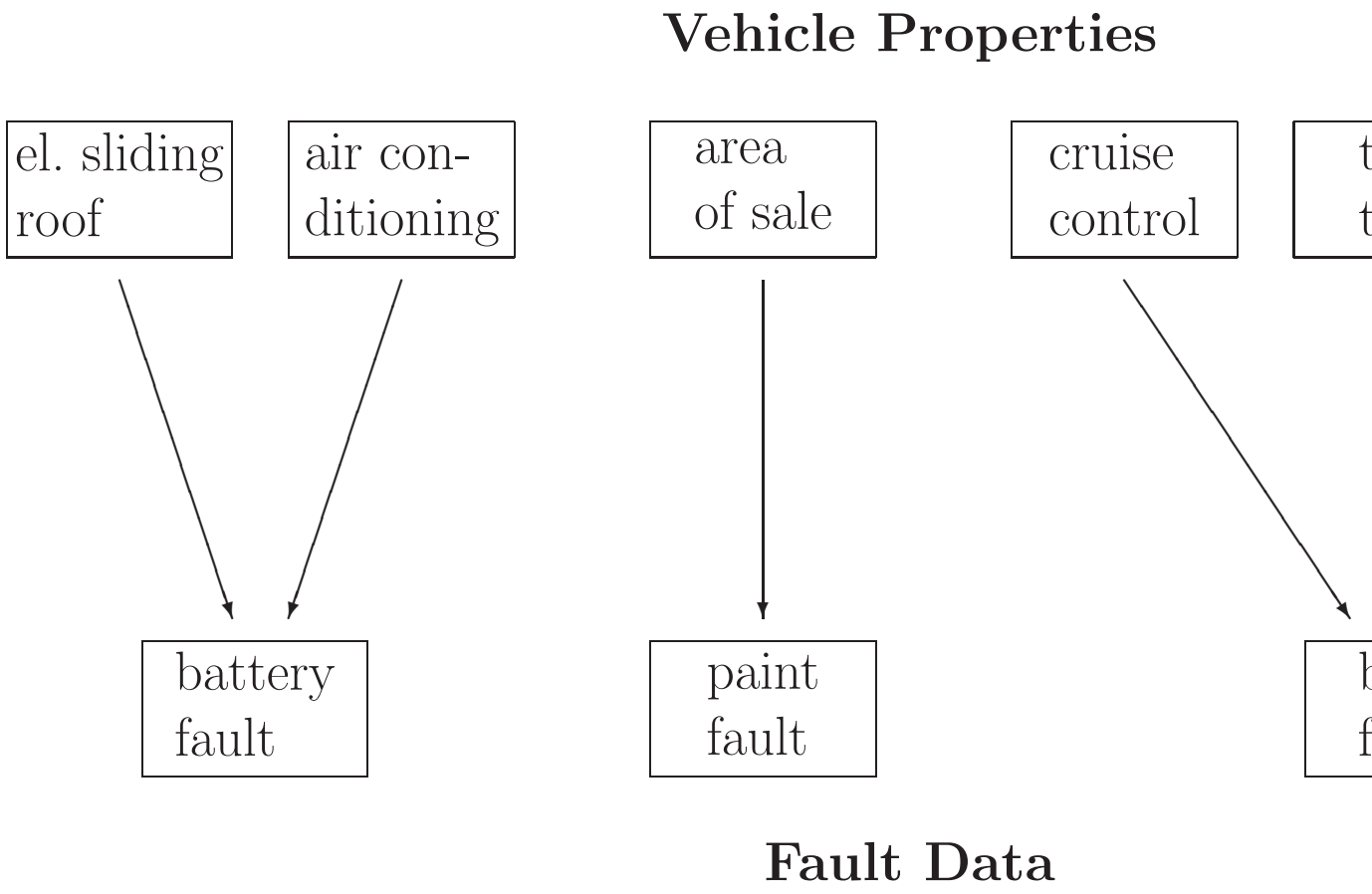
Fields of Application (DaimlerChrys

- **Improvement of Product Quality by Finding W**
  - Learn decision trees or inference 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 decision trees or inference network for vehicle properties and faults.
  - Record properties of new faulty vehicle.
  - Test for the most probable faults.

A Simple Approach to Fault Ana

- Check subnets consisting of an attribute and its parent at
- Select subnets with highest deviation from independent d



Example Subnet

Influence of special equipment on battery faults:

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

- Significant deviation from independent distribution.
- Hints to possible causes and improvements.
- Here: Larger battery may be required, if an air conditioni *and* an electrical sliding roof are built in.

(The dependencies and frequencies of this example are fictitious, true num

Summary

- **Decomposition:** Under certain conditions a distribution
distribution) on a multi-dimensional domain, which encod
knowledge about this domain, can be decomposed into a
(overlapping) distributions on lower-dimensional subspace

- **Simplified Reasoning:** If such a decomposition is pos
to know the distributions on the subspaces to draw all infe
under consideration that can be drawn using the original

- **Graphical Model:** The decomposition is represented
sense of graph theory). The edges of the graph indicate th
evidence has to be propagated. Efficient and correct ev
algorithms can be derived, which exploit the graph struct

- **Learning from Data:** There are several highly succe
learn graphical models from data, although all of them are
Exact learning methods are usually too costly.