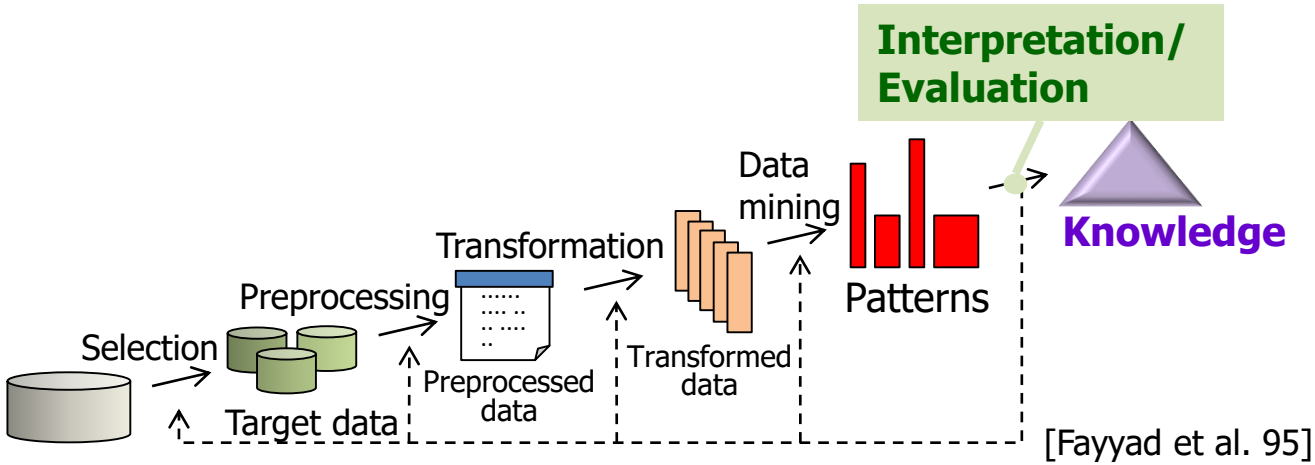
Verbal characterization of probabilistic clusters by minimal discriminative propositions

Yoshitaka Kameya, Satoru Nakamura, Tatsuya Iwasaki and Taisuke Sato
Tokyo Institute of Technology

Motivation

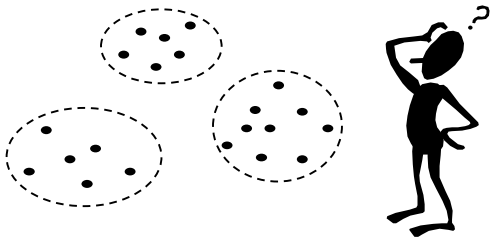
- Interpretation and evaluation are indispensable in practice



[Fayyad et al. 95]

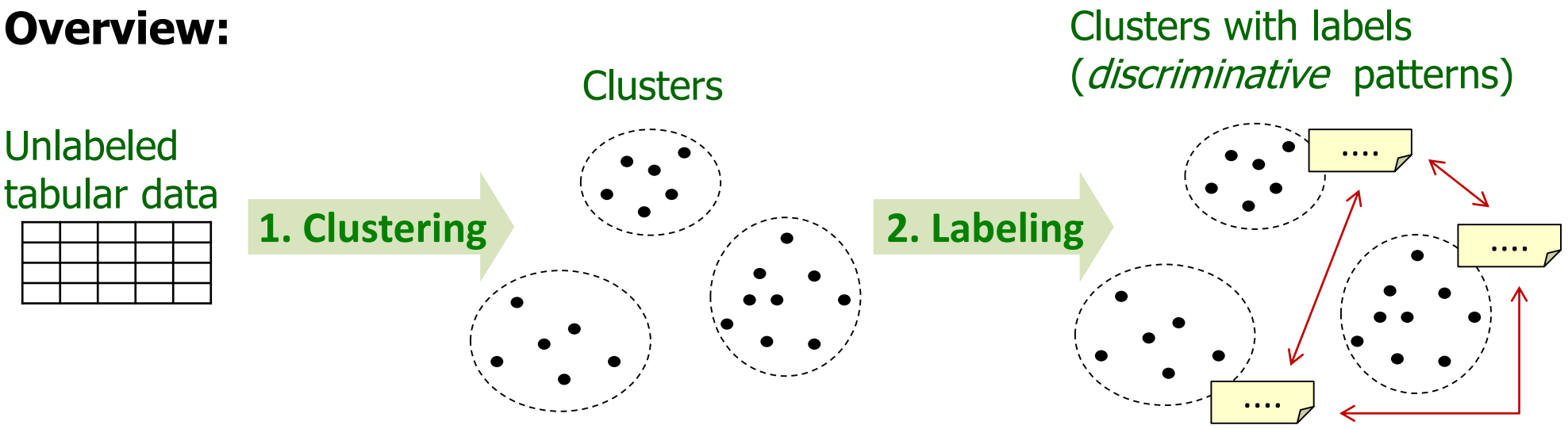
- Data clustering: a data mining technique
- It is sometimes difficult to see
  in what aspect the clusters have been formed

→ **Any help?**

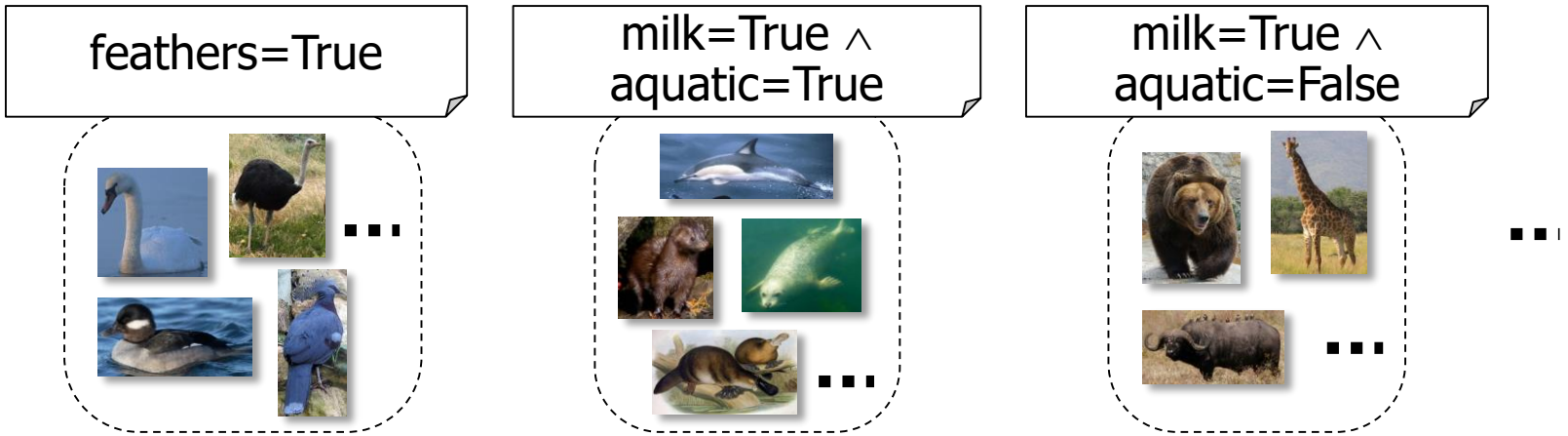


Verbal characterization of clusters

Overview:



Example:



Characteristic labels: Definition

- Characteristic label $\boldsymbol{x}$ of cluster C_k should satisfy:

1. $p(k \mid \boldsymbol{x}) \geq r$ (Relevance condition)
2. $p(\boldsymbol{x}) \geq s_{\text{global}}$ (Global support condition)
3. $p(\boldsymbol{x} \mid k) \geq s_{\text{local}}$ (Local support condition)
4. There is no $\boldsymbol{x}' \subset \boldsymbol{x}$ that satisfies $1 \sim 3$ (Minimality condition)

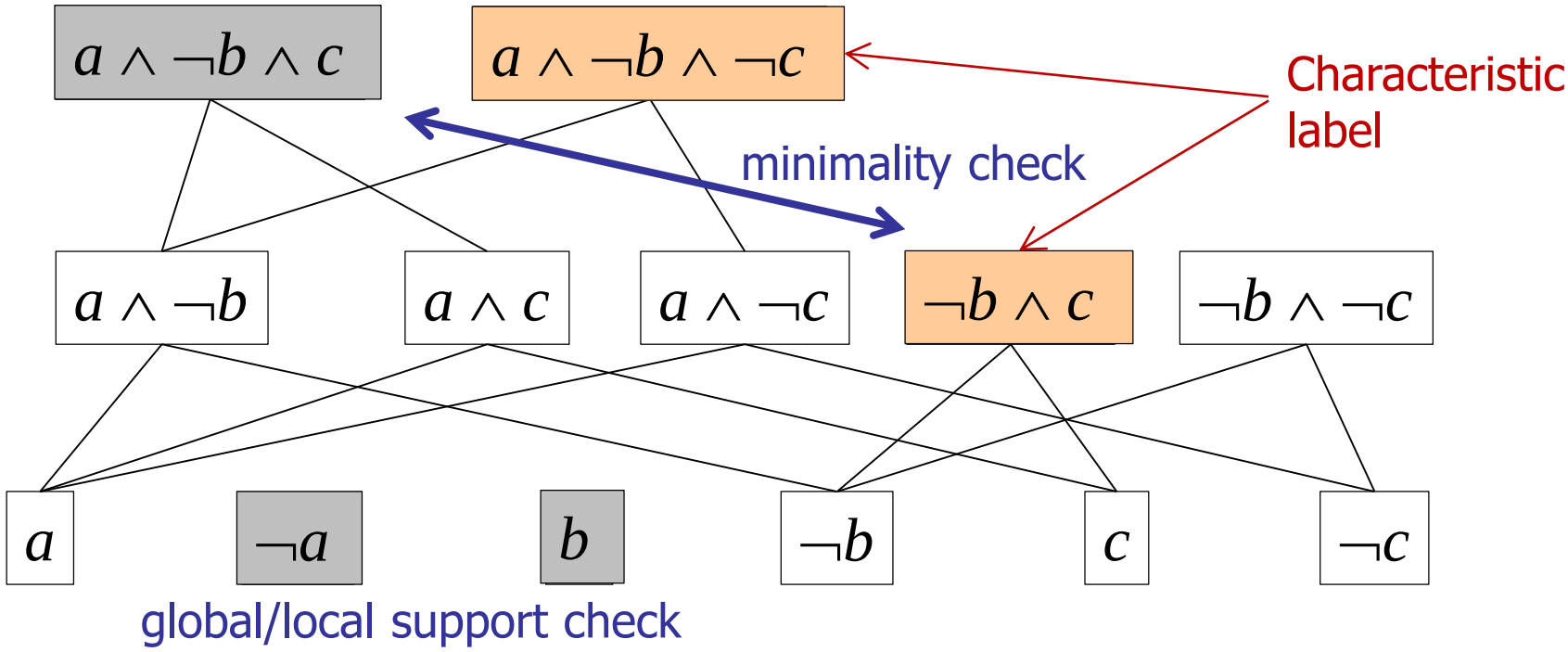
→ Primary filters

$$p(k \mid \boldsymbol{x}) \propto \theta(k) \prod_i \theta(x_i \mid k)$$

- Model-based computation of probabilities:
  - No need to scan the whole dataset D to get the counts
  - Well-abstracted data (when the model fits to D)
  - Problem with missing values automatically eliminated
  - Continuous values easily handled by assuming Gaussians

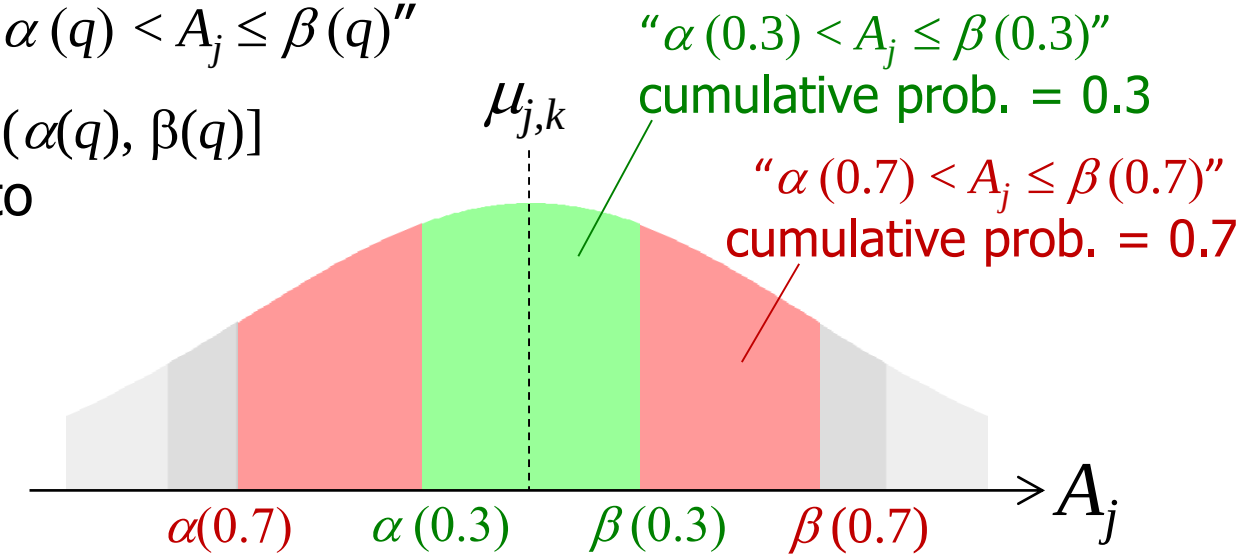
Exhaustive search for characteristic labels

- Currently we take a breadth-first style (Apriori-like)
- Boolean attributes: A, B, C
- "A=True" → a
- "A=False" → $\neg a$



Handling continuous attributes

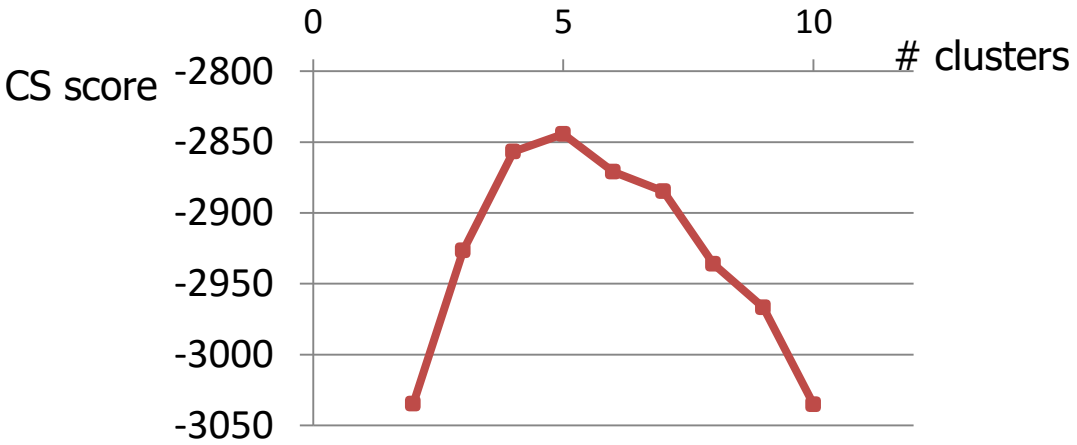
- Both discrete and continuous attributes are handled *consistently* in terms of $p(k \mid \boldsymbol{x})$
  - Gaussian distributions are assumed for continuous attributes
  - Set of cumulative probabilities $Q = \{0.1, 0.2, ..., 0.9\}$ is given
- We introduce propositions "$\alpha(q) < A_j \leq \beta(q)$" ($q \in Q$)
- $q > q'$ ➔ "$\alpha(q) < A_j \leq \beta(q)$" is more general than "$\alpha(q') < A_j \leq \beta(q')$"
- Choosing an appropriate "$\alpha(q) < A_j \leq \beta(q)$"
  - ➔ Adjusting the range $(\alpha(q), \beta(q)]$ automatically so as to maximize $p(k \mid \boldsymbol{x})$



Experiments: Flags dataset

- We used 15 discrete and 8 continuous (integer) *visual* attributes
- No human-annotated classes are given
  - #clusters is estimated as 5 using the Cheeseman-Stutz score (used in AutoClass)

But our in-depth method told us that 6 is better!

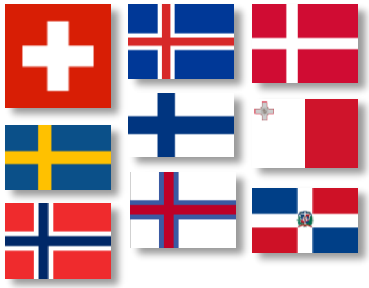


Results:

Label $\boldsymbol{x}$ for C_1	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
#saltires=1	1.000	0.900
topleft=white $\wedge$ #quarters=1	0.817	0.622
stripes=0,1,2 $\wedge$ #quarters=1	0.828	0.540
botright=blue $\wedge$ #quarters=1	0.819	0.505
:	:	:

Label x for C_2	$p(k \mid x)$	$p(x \mid k)$
#bars=1,2,3,4	0.782	0.800

Label x for C_3	$p(k \mid x)$	$p(x \mid k)$
#circles=1,2 $\wedge$ #quarters=0	0.781	0.540
#circles=1,2 $\wedge$ #crosses=0	0.781	0.540
black=T $\wedge$ #circles=1,2	0.766	0.225
blue=F $\wedge$ #circles=1	0.764	0.200
:	:	:



Label x for C_4	$p(k \mid x)$	$p(x \mid k)$
#crosses=1 $\wedge$ #quarters=0	0.829	0.810
#crosses=1 $\wedge$ #saltires=0	0.810	0.810
#crosses=1 $\wedge$ #sunstars=0	0.753	0.720
#circles=0 $\wedge$ #crosses=1	0.768	0.640
green=F $\wedge$ #crosses=1	0.757	0.500
#colors=2,3 $\wedge$ #crosses=1	0.759	0.490

Label $\boldsymbol{x}$ for C_5	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
#bars=0	0.803	0.900
#circles=0	0.752	0.900
#crosses=0 $\wedge$ #quarters=0	0.801	0.810
#crosses=0 $\wedge$ #saltires=0	0.795	0.810
#saltires=0 $\wedge$ #quarters=0	0.790	0.810
:	:	:

Label $\boldsymbol{x}$ for C_3	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
#saltires=0 $\wedge$ #quarters=1	0.960	0.360
topleft=blue $\wedge$ #quarters=1	0.875	0.320

Experiments: Zoo dataset

- 101 examples, 17 attributes (considered as all discrete)
- 7 original (human-annotated) classes
- In clustering, #clusters K is given as 7
- In labeling, $r = 0.9$, $s_{\text{local}} = 1 / (|D| / K) = K / |D|$, $s_{\text{global}} = 1 / |D|$ (ignorable)

Original classes	Obtained clusters						
	C_1	C_2	C_3	C_4	C_5	C_6	C_7
mammals	35	6	0	0	0	0	0
birds	0	0	20	0	0	0	0
fishes	0	0	0	13	0	0	0
amphibians	0	0	0	0	4	0	0
reptiles	0	0	0	0	5	0	0
insects	0	0	0	0	0	8	0
others	0	0	0	0	1	2	7

→ Split into two clusters (C_1 and C_2)

→ Merged into one cluster (C_5)

Experiments: Zoo dataset (Cluster #1 of 7)

- $C_1 \approx$ Terrestrial mammals

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_1	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
milk=T $\wedge$ aquatic=F	1.000	1.000
hair=T $\wedge$ aquatic=F $\wedge$ backbone=T	1.000	1.000
eggs=F $\wedge$ aquatic=F	0.972	1.000
hair=T $\wedge$ backbone=T $\wedge$ fins=F	0.963	1.000
milk=T $\wedge$ fins=F	0.946	1.000
aquatic=F $\wedge$ toothed=T $\wedge$ venomous=F	0.936	1.000
feathers=F $\wedge$ aquatic=F $\wedge$ backbone=T $\wedge$ venomous=F	0.928	1.000
hair=T $\wedge$ aquatic=F $\wedge$ venomous=F	0.916	1.000
hair=T $\wedge$ eggs=F	0.913	1.000
hair=T $\wedge$ toothed=T	0.913	1.000
aquatic=F $\wedge$ toothed=T $\wedge$ backbone=T	0.907	1.000
:	:	:

Experiments: Zoo dataset (Cluster #2 of 7)

- $C_2 \approx$ Aquatic mammals

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_2	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
milk=T $\wedge$ aquatic=T	1.000	1.000
feathers=F $\wedge$ aquatic=T $\wedge$ breathes=T $\wedge$ catsize=T	0.930	1.000
airborne=F $\wedge$ aquatic=T $\wedge$ predator=T $\wedge$ breathes=T $\wedge$ catsize=T	0.920	1.000
airborne=F $\wedge$ aquatic=T $\wedge$ breathes=T $\wedge$ venomous=F $\wedge$ catsize=T	0.905	1.000
eggs=F $\wedge$ aquatic=T $\wedge$ catsize=T	0.980	0.833
aquatic=T $\wedge$ toothed=T $\wedge$ breathes=T $\wedge$ catsize=T	0.933	0.833
eggs=F $\wedge$ aquatic=T $\wedge$ predator=T $\wedge$ venomous=F	0.912	0.833
eggs=F $\wedge$ aquatic=T $\wedge$ backbone=T $\wedge$ venomous=F	0.903	0.833
eggs=F $\wedge$ aquatic=T $\wedge$ breathes=T $\wedge$ venomous=F	0.903	0.833
breathes=T $\wedge$ fins=T	1.000	0.667
milk=T $\wedge$ fins=T	1.000	0.667
:	:	:

Experiments: Zoo dataset (Cluster #3 of 7)

- $C_3 \approx$ Birds

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_3	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
feathers=T	1.000	1.000
milk=F $\wedge$ legs=2	1.000	1.000
eggs=T $\wedge$ legs=2	0.992	1.000
toothed=F $\wedge$ legs=2	0.992	1.000
hair=F $\wedge$ legs=2	0.984	1.000
eggs=T $\wedge$ toothed=F $\wedge$ tail=T	0.941	1.000
milk=F $\wedge$ toothed=F $\wedge$ tail=T	0.934	1.000
eggs=T $\wedge$ toothed=F $\wedge$ backbone=T	0.925	1.000
toothed=F $\wedge$ venomous=F $\wedge$ tail=T	0.923	1.000
toothed=F $\wedge$ fins=F $\wedge$ tail=T	0.922	1.000
hair=F $\wedge$ toothed=F $\wedge$ tail=T	0.922	1.000
:	:	:

Experiments: Zoo dataset (Cluster #4 of 7)

- $C_4 \approx$ Fishes

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_4	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
breathes=F $\wedge$ fins=T	1.000	1.000
milk=F $\wedge$ fins=T	1.000	1.000
eggs=T $\wedge$ fins=T	0.951	1.000
breathes=F $\wedge$ tail=T	0.949	1.000
toothed=T $\wedge$ breathes=F	0.942	1.000
backbone=T $\wedge$ breathes=F	0.935	1.000
milk=F $\wedge$ aquatic=T $\wedge$ legs=0 $\wedge$ tail=T	0.925	1.000
milk=F $\wedge$ aquatic=T $\wedge$ toothed=T $\wedge$ legs=0	0.915	1.000
hair=F $\wedge$ eggs=T $\wedge$ aquatic=T $\wedge$ backbone=T $\wedge$ legs=0	0.912	1.000
eggs=T $\wedge$ aquatic=T $\wedge$ legs=0 $\wedge$ tail=T	0.911	1.000
hair=F $\wedge$ eggs=T $\wedge$ toothed=T $\wedge$ backbone=T $\wedge$ legs=0 $\wedge$ tail=T	0.907	1.000
:	:	:

Experiments: Zoo dataset (Cluster #5 of 7)

- $C_5 \approx$ Amphibians + Reptiles

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_5	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
feathers=F $\wedge$ milk=F $\wedge$ backbone=T $\wedge$ fins=F	1.000	0.899
hair=F $\wedge$ feathers=F $\wedge$ backbone=T $\wedge$ fins=F	0.931	0.899
feathers=F $\wedge$ milk=F $\wedge$ backbone=T $\wedge$ breathes=T	1.000	0.809
milk=F $\wedge$ toothed=T $\wedge$ fins=F	1.000	0.799
hair=F $\wedge$ toothed=T $\wedge$ fins=F	0.935	0.799
hair=F $\wedge$ feathers=F $\wedge$ backbone=T $\wedge$ breathes=T $\wedge$ catsize=F	1.000	0.728
milk=F $\wedge$ toothed=T $\wedge$ breathes=T	1.000	0.719
feathers=F $\wedge$ milk=F $\wedge$ fins=F $\wedge$ tail=T	1.000	0.699
feathers=F $\wedge$ eggs=T $\wedge$ backbone=T $\wedge$ fins=F	0.956	0.720
feathers=F $\wedge$ milk=F $\wedge$ airborne=F $\wedge$ predator=T $\wedge$ breathes=T	0.947	0.720
hair=F $\wedge$ feathers=F $\wedge$ milk=F $\wedge$ predator=T $\wedge$ breathes=T	0.923	0.720
:	:	:

Experiments: Zoo dataset (Cluster #6 of 7)

- $C_6 \approx$ Insects

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_6	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
toothed=F $\wedge$ breathes=T $\wedge$ tail=F	0.934	1.000
feathers=F $\wedge$ aquatic=F $\wedge$ toothed=F $\wedge$ breathes=T	0.917	1.000
backbone=F $\wedge$ breathes=T	0.917	1.000
aquatic=F $\wedge$ backbone=F $\wedge$ tail=F $\wedge$ catsize=F	0.910	1.000
eggs=T $\wedge$ aquatic=F $\wedge$ toothed=F $\wedge$ tail=F $\wedge$ catsize=F	0.903	1.000
eggs=T $\wedge$ aquatic=F $\wedge$ breathes=T $\wedge$ tail=F	0.902	1.000
aquatic=F $\wedge$ toothed=F $\wedge$ backbone=F	0.901	1.000
predator=F $\wedge$ toothed=F $\wedge$ tail=F	0.987	0.900
predator=F $\wedge$ backbone=F	0.978	0.900
feathers=F $\wedge$ predator=F $\wedge$ toothed=F	0.957	0.900
eggs=T $\wedge$ predator=F $\wedge$ tail=F	0.949	0.900
:	:	:

Experiments: Zoo dataset (Cluster #7 of 7)

- $C_7 \approx$ Other creatures

Labels are ordered by the harmonic mean of $p(k \mid \boldsymbol{x})$ and $p(\boldsymbol{x} \mid k)$

Characteristic label $\boldsymbol{x}$ for C_7	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
backbone=F $\wedge$ breathes=F	0.986	1.000
toothed=F $\wedge$ breathes=F	0.972	1.000
breathes=F $\wedge$ tail=F	0.959	1.000
airborne=F $\wedge$ predator=T $\wedge$ toothed=F $\wedge$ backbone=F	0.925	1.000
hair=F $\wedge$ eggs=T $\wedge$ airborne=F $\wedge$ predator=T $\wedge$ toothed=F $\wedge$ tail=F	0.917	1.000
airborne=F $\wedge$ predator=T $\wedge$ backbone=F $\wedge$ tail=F	0.916	1.000
eggs=T $\wedge$ predator=T $\wedge$ breathes=F $\wedge$ fins=F	0.916	1.000
hair=F $\wedge$ milk=F $\wedge$ airborne=F $\wedge$ predator=T $\wedge$ toothed=F $\wedge$ tail=F	0.906	1.000
hair=F $\wedge$ eggs=T $\wedge$ predator=T $\wedge$ backbone=F $\wedge$ tail=F $\wedge$ domestic=F	0.905	1.000
hair=F $\wedge$ airborne=F $\wedge$ predator=T $\wedge$ toothed=F $\wedge$ fins=F $\wedge$ tail=F	0.904	1.000
hair=F $\wedge$ airborne=F $\wedge$ predator=T $\wedge$ toothed=F $\wedge$ tail=F $\wedge$ domestic=F	0.902	1.000
:	:	:

Experiments: Zoo dataset (in-depth analysis)

Original classes	Obtained clusters						
	C_1	C_2	C_3	C_4	C_5	C_6	C_7
mammals	35	6	0	0	0	0	0
birds	0	0	20	0	0	0	0
:	:	:	:	:	:	:	:

- C_1 & C_2: wrongly split clusters
- These clusters are still meaningful according to characteristic labels

→ The proposed method gives
a new way of in-depth analysis for the obtained clusters
(in the past, we only used *numeric* matching scores)

Characteristic label $\boldsymbol{x}$ for C_1	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
milk=T $\wedge$ aquatic=F	1.000	1.000
hair=T $\wedge$ aquatic=F $\wedge$ backbone=T	1.000	1.000
:	:	:

Characteristic label $\boldsymbol{x}$ for C_2	$p(k \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid k)$
milk=T $\wedge$ aquatic=T	1.000	1.000
feathers=F $\wedge$ aquatic=T $\wedge$ breathes=T $\wedge$ catsize=T	0.930	1.000
:	:	:

Future work

- **Logic-based learning**
  - Our labeling method: an instance of inductive logic programming (ILP) with a simple refinement operator
  - Background knowledge for improving the comprehensibility
  - Apriori with a simple taxonomy [Srikant and Agrawal 96] can be imported

- **Probabilistic modeling**
  - Evidence-based sensitivity analysis of a Bayesian network [Jensen 96]
  - The relevance score $p(k \mid \boldsymbol{x})$ is generalized into $p(q \mid \boldsymbol{e})$