Statistical Predicate Invention

Stanley Kok

Dept. of Computer Science and Eng.

University of Washington

Joint work with Pedro Domingos

Overview

- Motivation
- Background
- Multiple Relational Clusterings
- Experiments
- Future Work

Motivation

Statistical Relational Learning

Statistical Learning
- able to handle noisy data

Relational Learning (ILP)
- able to handle non-i.i.d. data

Motivation

Statistical Predicate Invention

Discovery of new concepts, properties, and relations from data

Statistical Relational Learning

Latent Variable Discovery

[Elidan & Friedman, 2005; Elidan et al., 2001; etc.]

Statistical Learning

- able to handle noisy data

Predicate Invention

[Wogulis & Langley, 1989; Muggleton & Buntine, 1988; etc.]

Relational Learning (ILP)

- able to handle non-i.i.d. data

SPI Benefits

- More compact and comprehensible models
- Improve accuracy by representing unobserved aspects of domain
- Model more complex phenomena

State of the Art

- Few approaches combine statistical and relational learning
  - Only cluster objects [Roy et al., 2006; Long et al., 2005; Xu et al., 2005; Neville & Jensen, 2005; Popescul & Ungar 2004; etc.]
  - Only predict single target predicate [Davis et al., 2007; Craven & Slattery, 2001]
- Infinite Relational Model [Kemp et al., 2006; Xu et al., 2006]
  - Clusters objects and relations simultaneously
  - Multiple types of objects
  - Relations can be of any arity
  - #Clusters need not be specified in advance

Multiple Relational Clusterings

- Clusters objects and relations simultaneously
- Multiple types of objects
- Relations can be of any arity
- #Clusters need not be specified in advance
- Learns **multiple** cross-cutting clusterings
- Finite second-order Markov logic
- First step towards general framework for SPI

Overview

- Motivation
- **Background**
- Multiple Relational Clusterings
- Experiments
- Future Work

Markov Logic Networks (MLNs)

- A logical KB is a set of **hard constraints** on the set of possible worlds
- Let's make them **soft constraints**:
  When a world violates a formula,
  it becomes less probable, not impossible
- Give each formula a **weight**
  (Higher weight $\Rightarrow$ Stronger constraint)

$$P(\text{world}) \propto \exp\left(\sum \text{weights of formulas it satisfies}\right)$$

Markov Logic Networks (MLNs)

$$P(\mathbf{x}) = \frac{1}{Z} \exp(\sum_{i=1}^{F} w_i n_i)$$

Vector of
truth assignments to
ground atoms

Partition function.
Sums over all possible
truth assignments to
ground atoms

Weight of i^{th} formula

#true groundings
of i^{th} formula

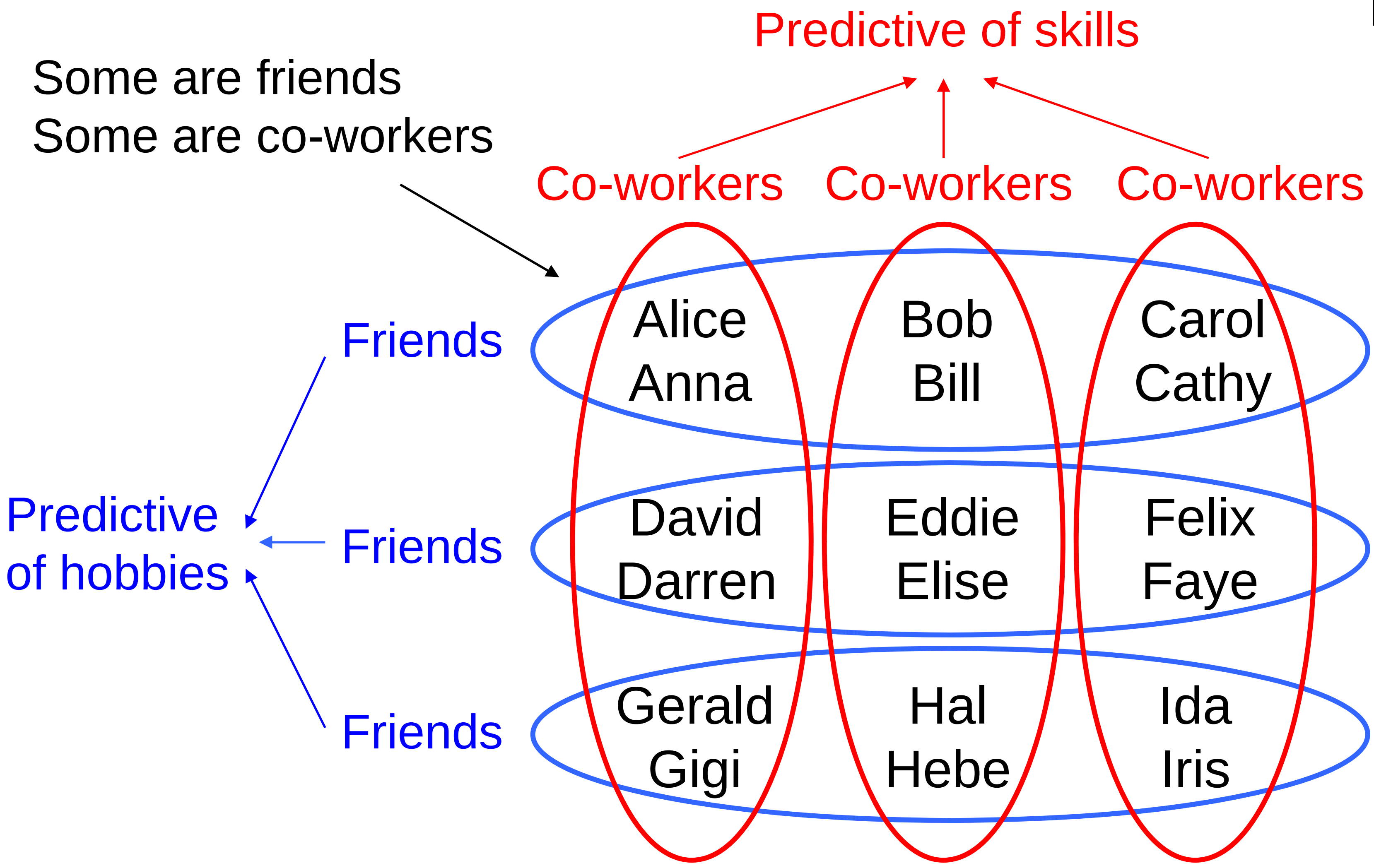
Overview

- Motivation
- Background
- **Multiple Relational Clusterings**
- Experiments
- Future Work

Multiple Relational Clusterings

- Invent unary predicate = Cluster
- Multiple cross-cutting clusterings
- Cluster relations by objects they relate and vice versa
- Cluster objects of same type
- Cluster relations with same arity and argument types

Example of Multiple Clusterings



Second-Order Markov Logic

- Finite, function-free
- Variables range over relations (predicates) and objects (constants)
- Ground atoms with all possible predicate symbols and constant symbols
- Represent some models more compactly than first-order Markov logic
- Specify how predicate symbols are clustered

Symbols

- Cluster: γ_i
- Clustering: Γ
- Atom: $x_i \in \gamma_i$, $r(x_1, \ldots, x_i, \ldots, x_n)$
- Cluster combination: $(\gamma_r, \gamma_1, \ldots, \gamma_i, \ldots, \gamma_n)$

MRC Rules

- Each symbol belongs to at least one cluster

$$\infty \quad \forall x \, \exists \gamma \; x \in \gamma$$

- Symbol cannot belong to >1 cluster in same clustering

$$\infty \quad \forall x, \gamma, \gamma', \Gamma \;\; x \in \gamma \wedge \gamma \in \Gamma \wedge \gamma' \in \Gamma \wedge \gamma \neq \gamma' \Rightarrow x \notin \gamma'$$

- Each atom appears in exactly one combination of clusters

$$\infty \quad \begin{array}{l} \forall r, x_1, \ldots, x_n \; \exists! \gamma_r, \gamma_1, \ldots, \gamma_n \\ \quad r \in \gamma_r \wedge x_1 \in \gamma_1 \wedge \ldots \wedge x_n \in \gamma_n \end{array}$$

MRC Rules

- **Atom prediction rule:** Truth value of atom is determined by cluster combination it belongs to

$$w \quad \begin{array}{l} \forall r, x_1, \ldots, x_n, +\gamma_r, +\gamma_1, \ldots, +\gamma_n \\ \quad r \in \gamma_r \wedge x_1 \in \gamma_1 \wedge \ldots \wedge x_n \in \gamma_n \Rightarrow r(x_1, \ldots, x_n) \end{array}$$

- Exponential prior on number of clusters

$$-\lambda \quad \forall \gamma \; \exists x \; x \in \gamma$$

Learning MRC Model

Learning consists of finding

- Cluster assignment $\{\Gamma\}$: assignment of truth values to all $r \in \gamma_r$ and $x \in \gamma_x$ atoms
- Weights of atom prediction rules

that maximize log-posterior probability

$$\log P(\{\Gamma\}|R) \propto \log P(\{\Gamma\}) + \log P(R|\{\Gamma\})$$

Vector of truth assignments to
all observed ground atoms

Learning MRC Model

$$\log P(\{\Gamma\}|R) \propto \underbrace{\log P(\{\Gamma\})} + \log P(R|\{\Gamma\})$$

Three hard rules
+ Exponential prior rule

$$\log P(\{\Gamma\}) = \begin{cases} -\infty & \text{if } \{\Gamma\} \text{ violates hard rule} \\ -\lambda \times \#clusters & \text{otherwise} \end{cases}$$

Learning MRC Model

$$\log P(\{\Gamma\}|R) \propto \log P(\{\Gamma\}) + \underbrace{\log P(R|\{\Gamma\})}$$

Atom prediction rules

Can be computed in closed form

Wt of rule is log-odds of atom in its cluster combination being true $= \log \dfrac{t+\beta}{f+\beta}$

Smoothing parameter

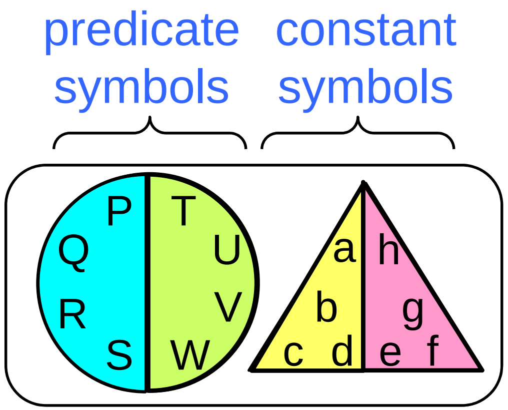
#true & #false atoms in cluster combination

Search Algorithm

- Approximation: Hard assignment of symbols to clusters
- Greedy with restarts
- Top-down divisive refinement algorithm
- Two levels
  - Top-level finds clusterings
  - Bottom-level finds clusters

Search Algorithm

Inputs: sets of



Greedy search with restarts

Outputs: Clustering of each set of symbols

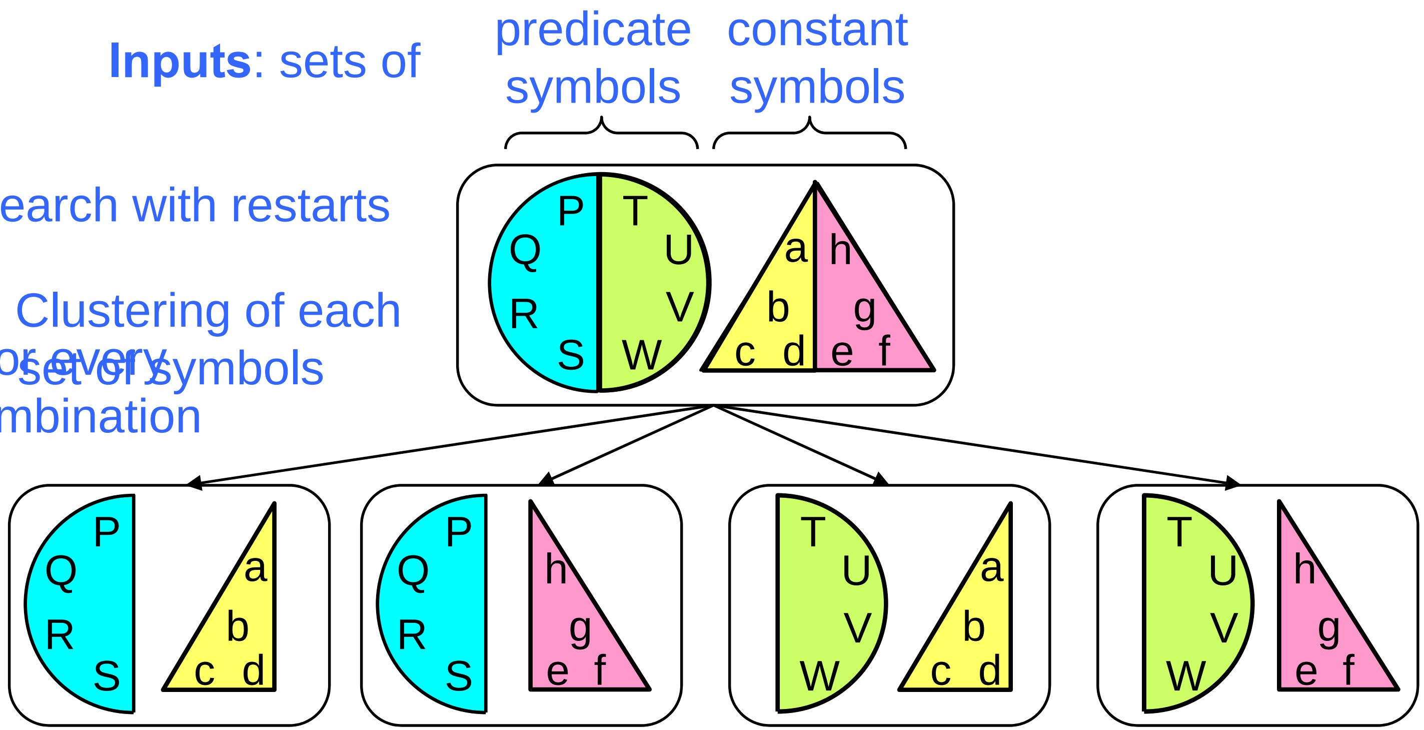
Search Algorithm

Inputs: sets of

Greedy search with restarts

Outputs: Clustering of each set of symbols

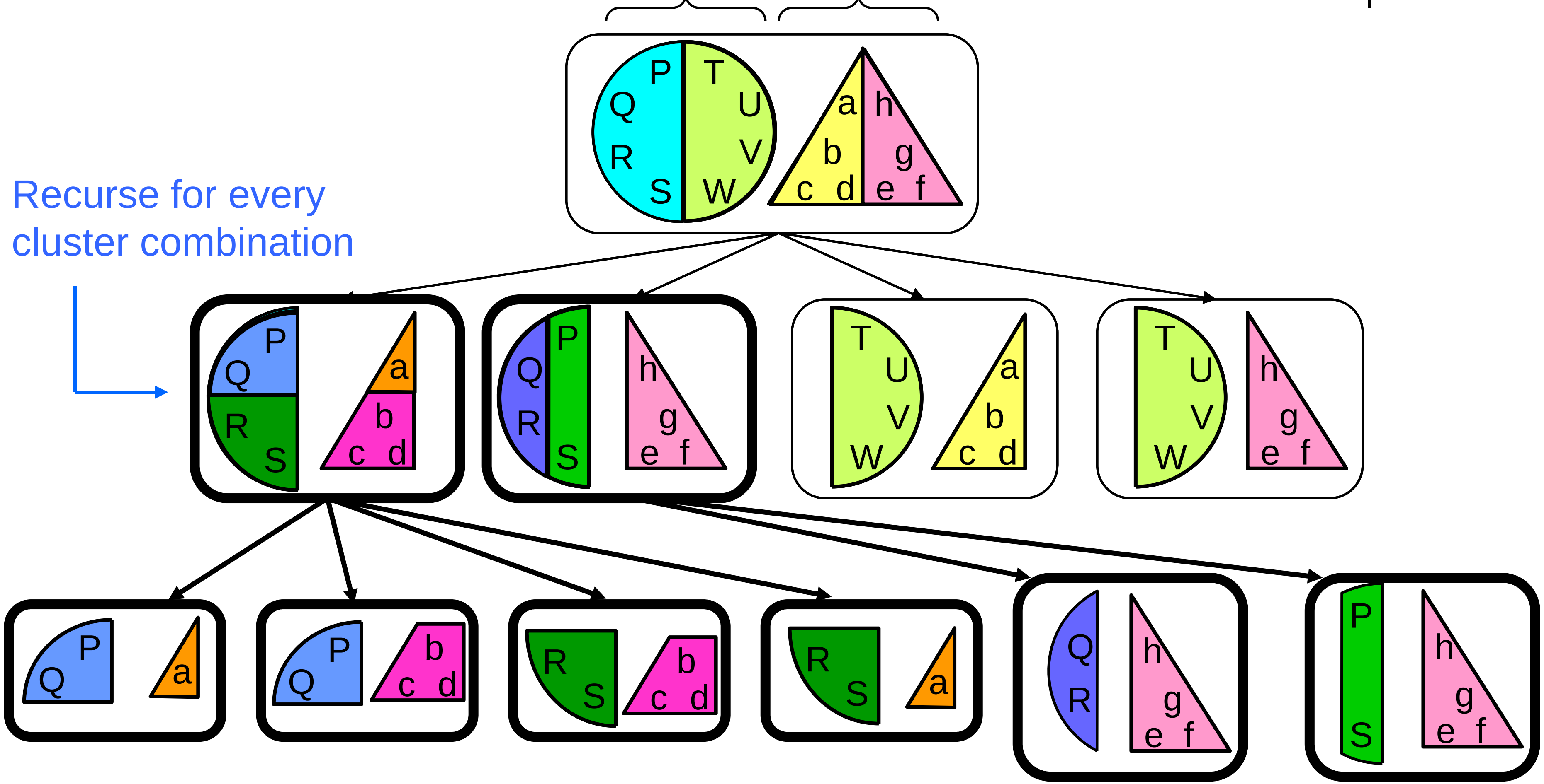
Recurse for every cluster combination



Search Algorithm

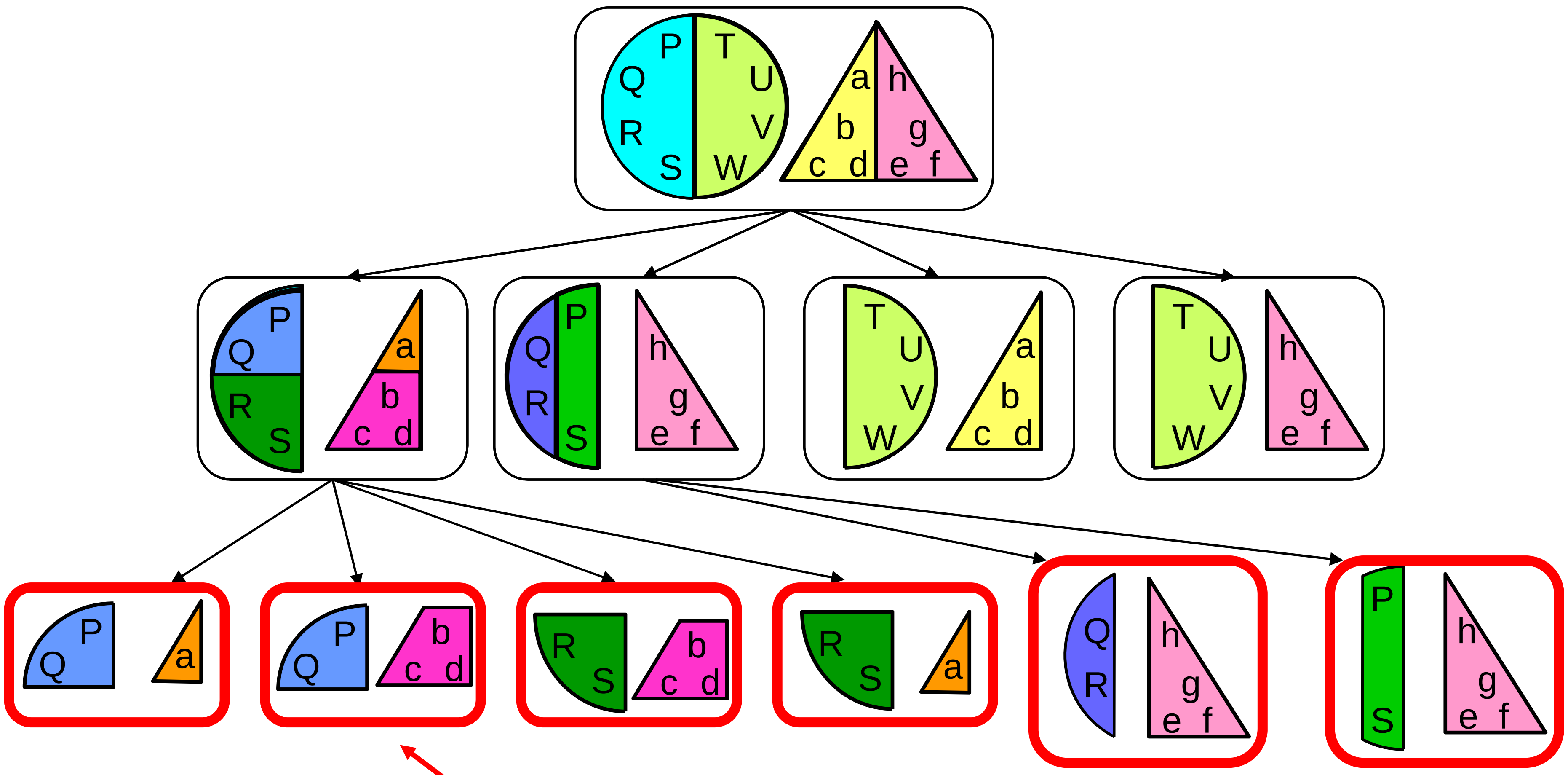
Inputs: sets of predicate symbols constant symbols

Recurse for every cluster combination



Terminate when no refinement improves MAP score

Search Algorithm

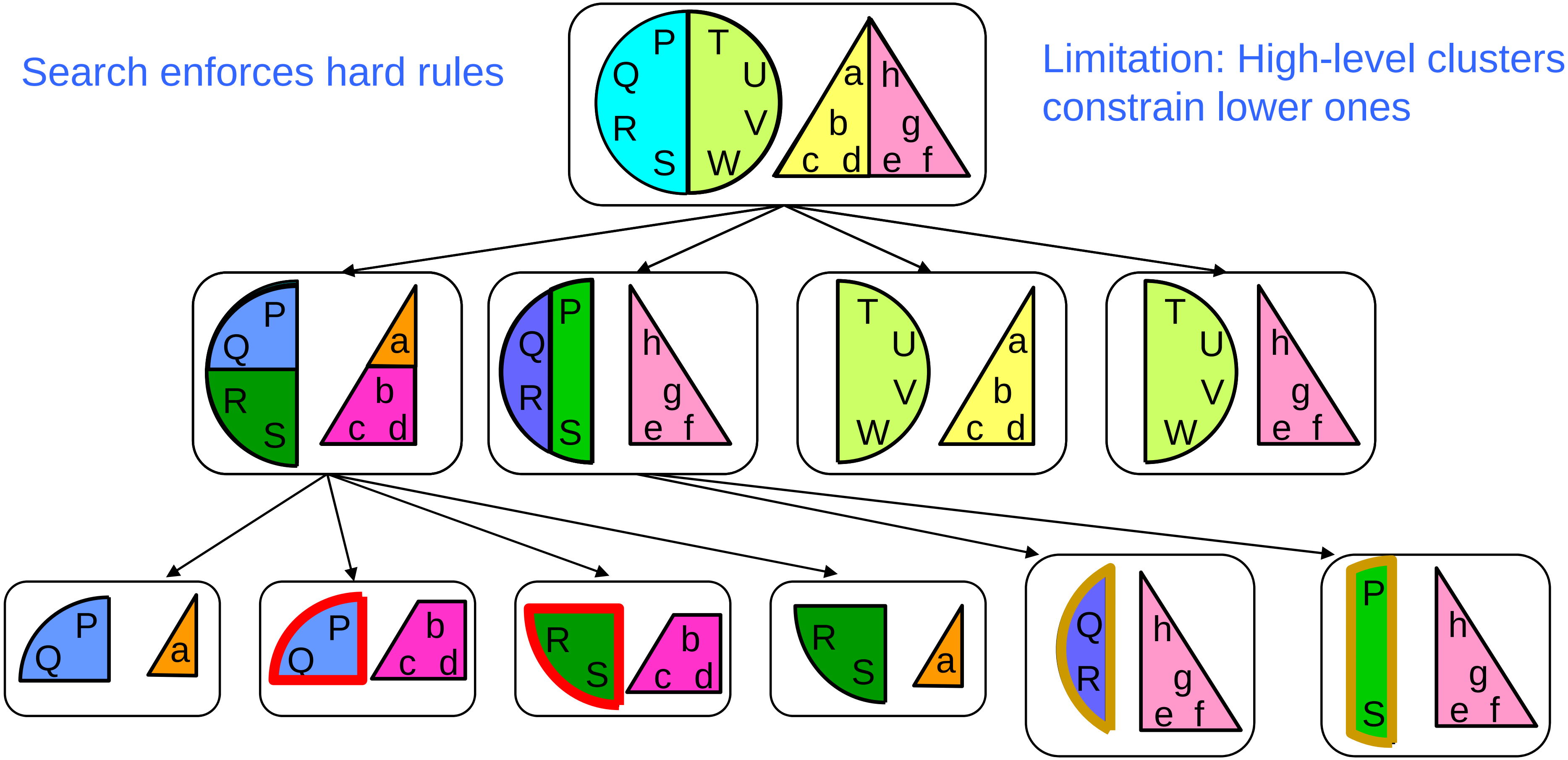


$9r, x \quad r \; 3 \; \gamma_r \; ^3 \; x \; 3 \; \gamma_x \; * \quad r(x)$

Search Algorithm

Search enforces hard rules

Limitation: High-level clusters constrain lower ones



: Multiple clusterings

Overview

- Motivation
- Background
- Multiple Relational Clusterings
- **Experiments**
- Future Work

Datasets

- **Animals**
  - Sets of animals and their features, e.g., *Fast(Leopard)*
  - 50 animals, 85 features
  - **4250** ground atoms; **1562** true ones
- **Unified Medical Language System (UMLS)**
  - Biomedical ontology
  - Binary predicates, e.g., *Treats(Antibiotic,Disease)*
  - 49 relations, 135 concepts
  - **893,025** ground atoms; **6529** true ones

Datasets

- **Kinship**
  - Kinship relations between members of an Australian tribe: *Kinship(Person,Person)*
  - 26 kinship terms, 104 persons
  - **281,216** ground atoms; **10,686** true ones
- **Nations**
  - Set of relations among nations, *e.g.,ExportsTo(USA,Canada)*
  - Set of nation features, *e.g., Monarchy(UK)*
  - 14 nations, 56 relations, 111 features
  - **12,530** ground atoms; **2565** true ones

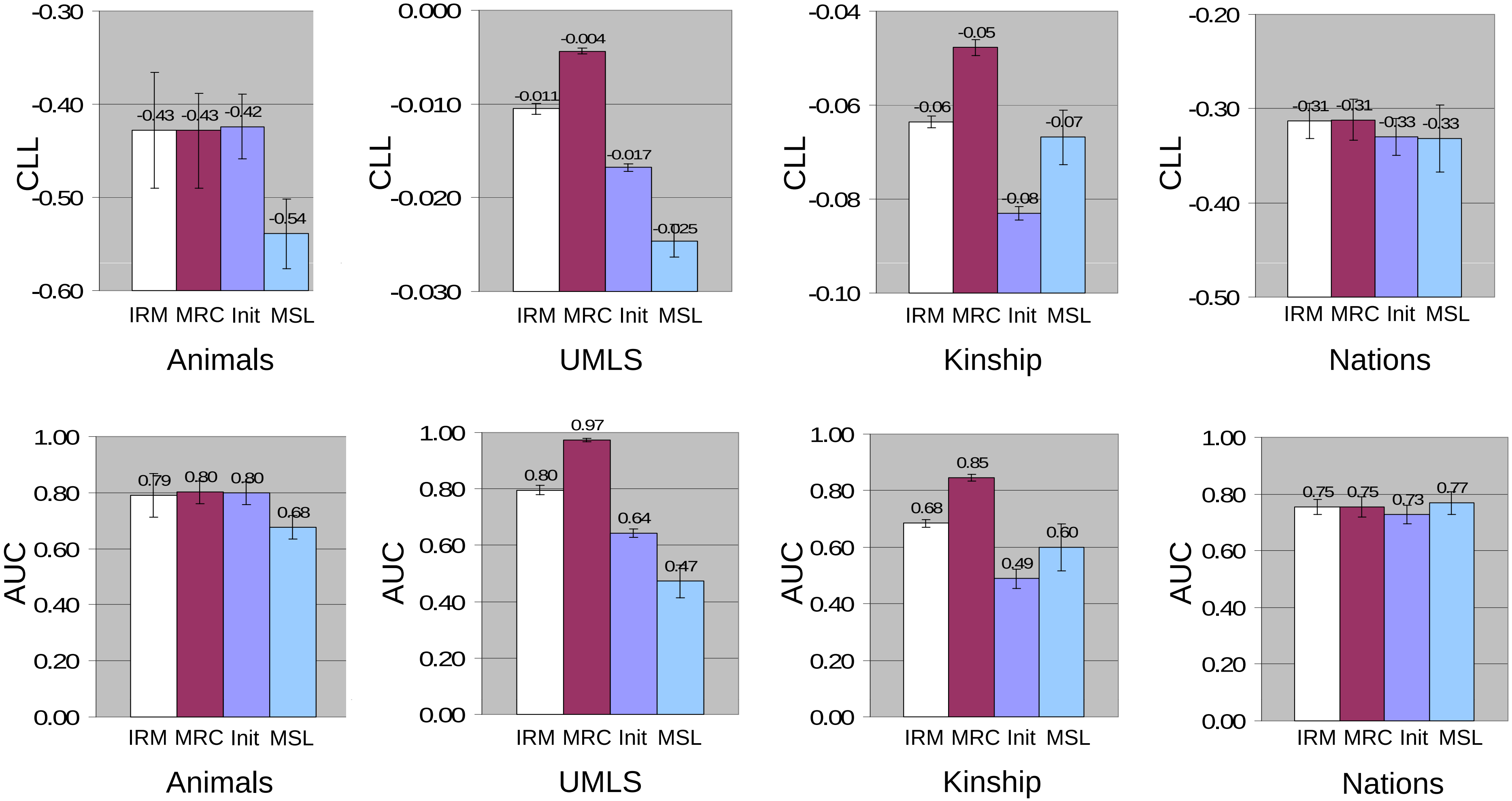
Methodology

- Randomly divided ground atoms into ten folds
- 10-fold cross validation
- Evaluation measures
  - Average conditional log-likelihood of test ground atoms (CLL)
  - Area under precision-recall curve of test ground atoms (AUC)

Methodology

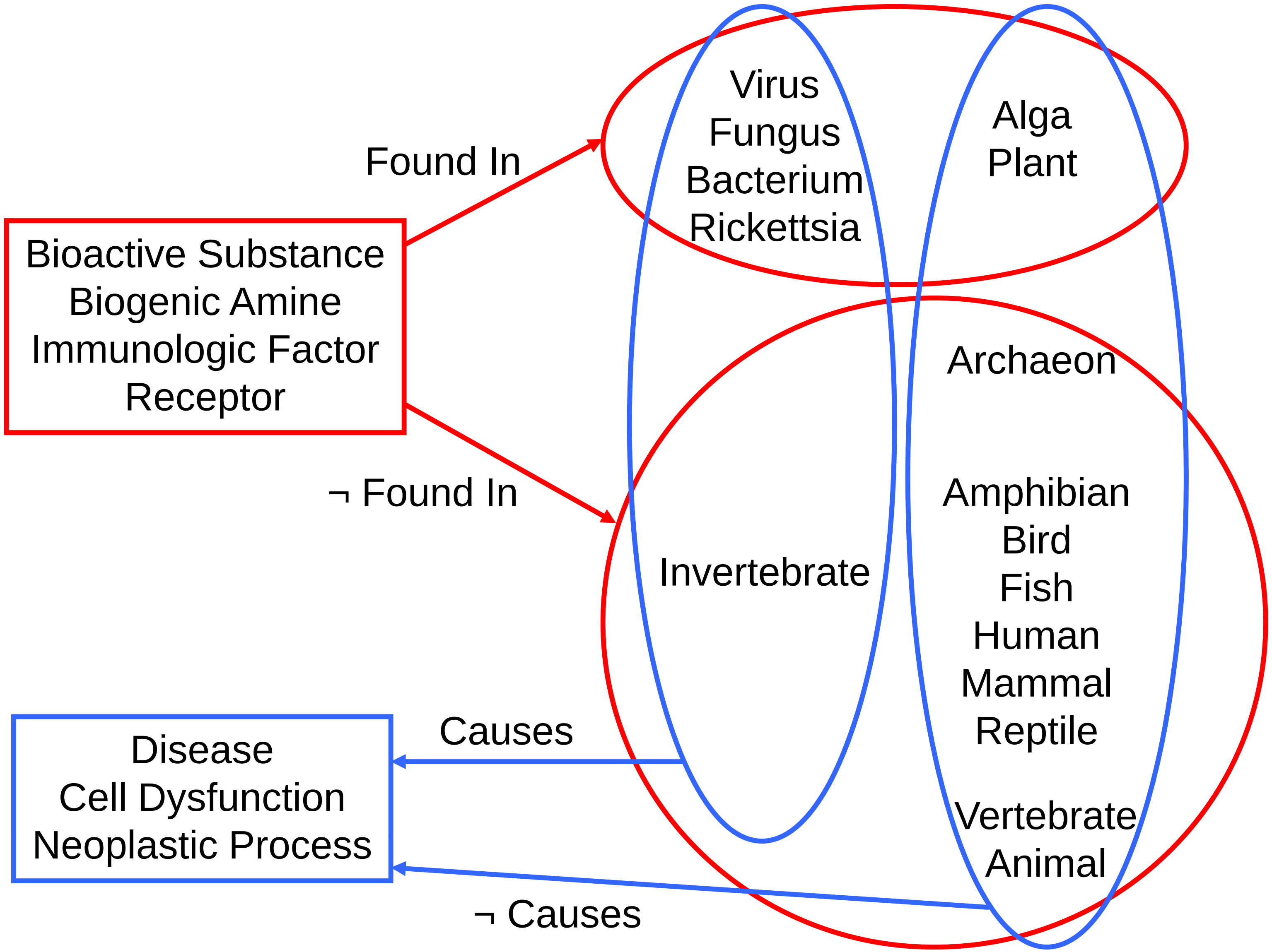
- Compared with IRM [Kemp et al., 2006] and MLN structure learning (MSL) [Kok & Domingos, 2005]
- Used default IRM parameters; run for 10 hrs
- MRC parameters λ and β both set to 1 (no tuning)
- MRC run for 10 hrs for first level of clustering
- MRC subsequent levels permitted 100 steps (3-10 mins)
- MSL run for 24 hours; parameter settings in online appendix

Results

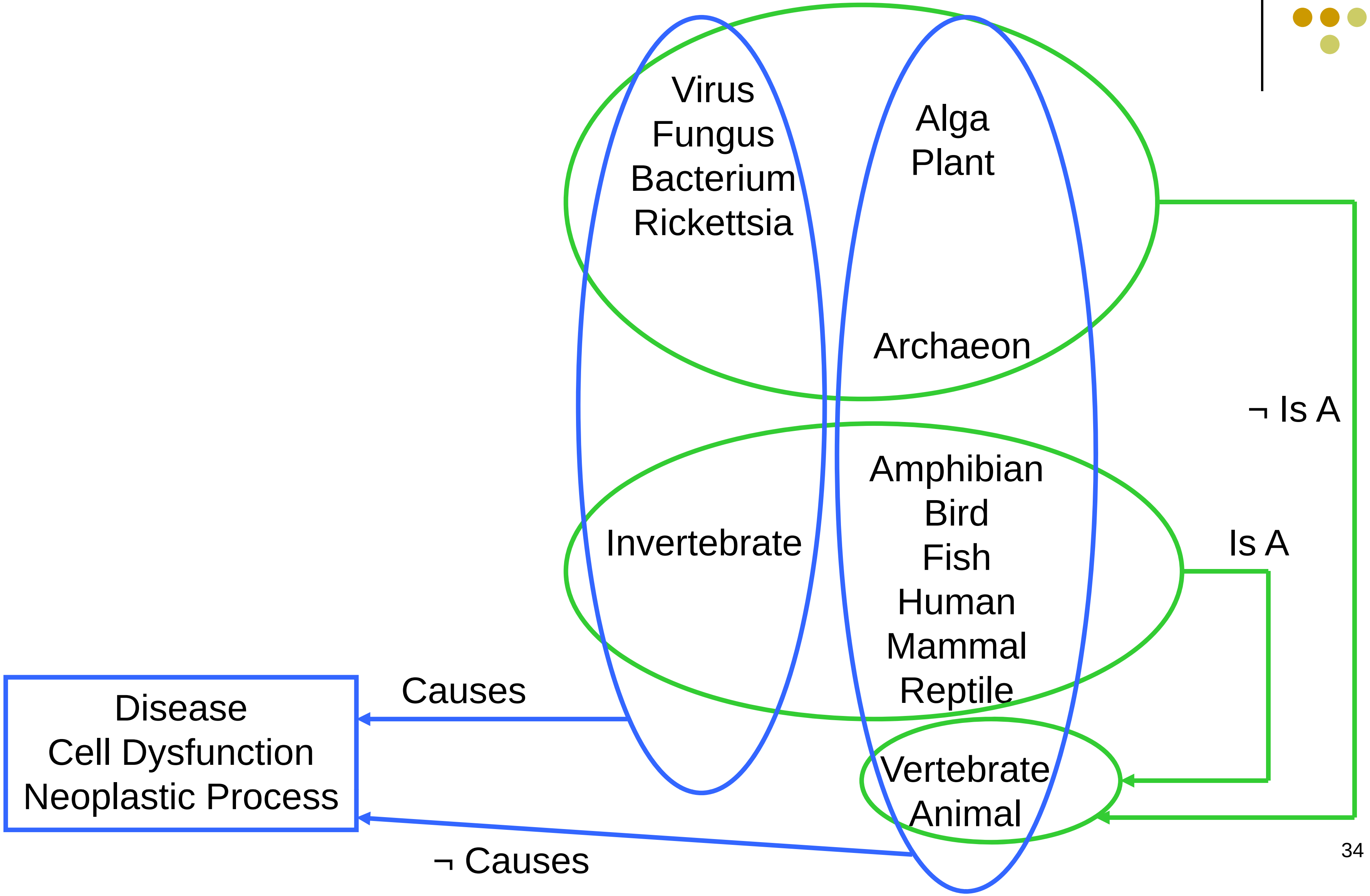


CLL
-0.30
-0.40
-0.50
-0.60
-0.43
-0.43
-0.42
-0.54
IRM MRC Init MSL
Animals
0.000
-0.010
-0.020
-0.030
-0.011
-0.004
-0.017
-0.025
UMLS
-0.04
-0.06
-0.08
-0.10
-0.06
-0.05
-0.08
-0.07
Kinship
-0.20
-0.30
-0.40
-0.50
-0.31
-0.31
-0.33
-0.33
Nations
AUC
1.00
0.80
0.60
0.40
0.20
0.00
0.79
0.80
0.80
0.68
Animals
0.80
0.97
0.64
0.47
UMLS
0.68
0.85
0.49
0.60
Kinship
0.75
0.75
0.73
0.77
Nations

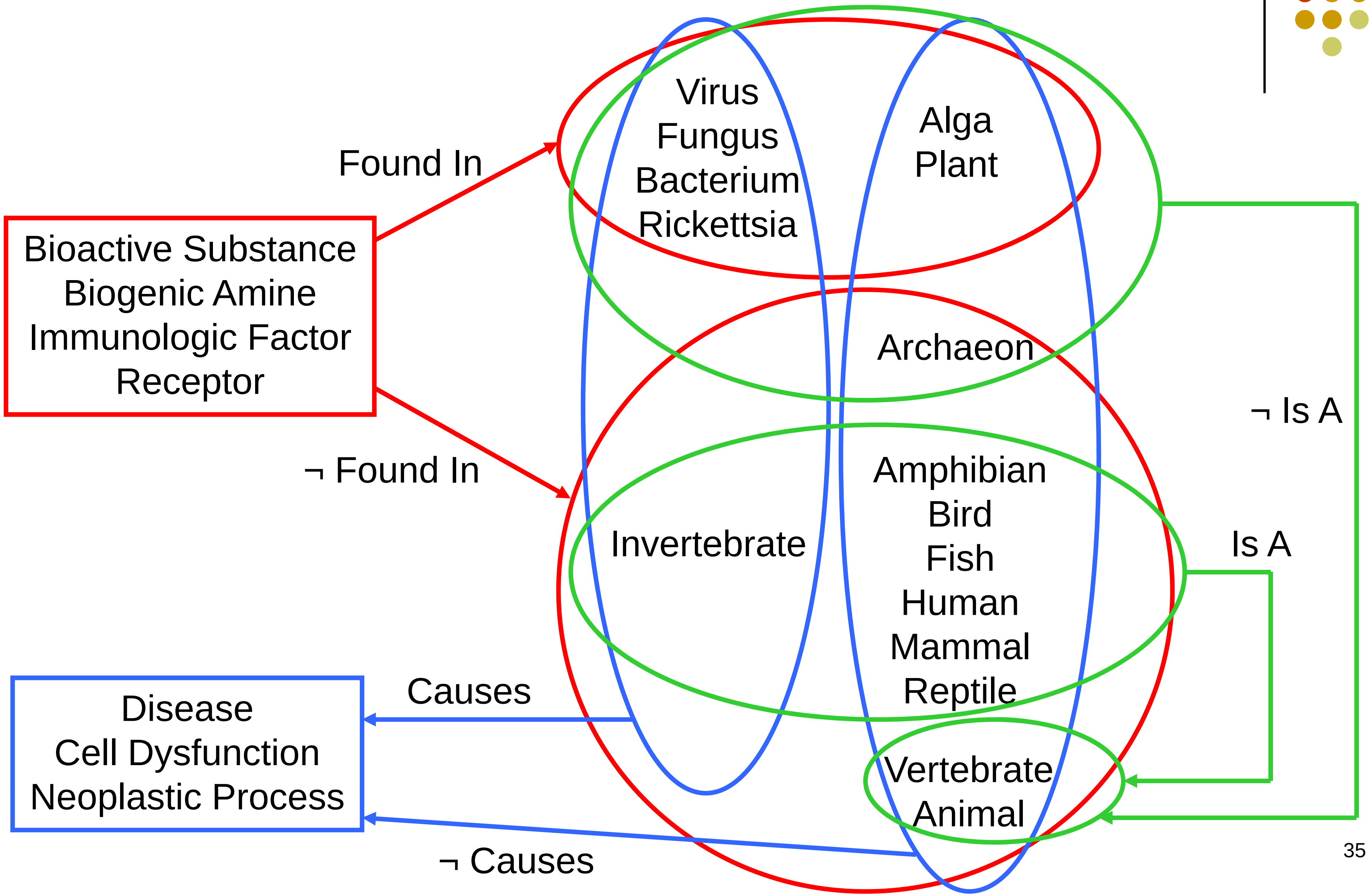
Multiple Clusterings Learned



Multiple Clusterings Learned



Multiple Clusterings Learned



Overview

- Motivation
- Background
- Multiple Relational Clusterings
- Experiments
- **Future Work**

Future Work

- Experiment on larger datasets,
  - e.g., ontology induction from web text
- Use clusters learned as primitives in structure learning
- Learn a hierarchy of multiple clusterings and performing shrinkage
- Cluster predicates with different arities and argument types
- Speculation: all relational structure learning can be accomplished with SPI alone

Conclusion

- Statistical Predicate Invention: key problem for statistical relational learning
- Multiple Relational Clusterings
  - First step towards general framework for SPI
  - Based on finite second-order Markov logic
  - Creates multiple relational clusterings of the symbols in data
  - Empirical comparison with MLN structure learning and IRM shows promise

SPI Benefits

- Compact and comprehensible model
  - Invented predicate efficiently captures dependencies among observed predicates
    - Fewer parameters; lower risk of overfitting
  - Less memory to represent model; potentially speed up inference
- Improve accuracy by representing unobserved aspects of domain
- Invented predicates can be used to learn new formulas
  - Larger search steps; learn more complex models
  - Extend search space by aggregating observed ones

Cluster = Invented Unary Predicate

Statistical Predicate Invention

Latent Variable Discovery
[Elidan & Friedman, 2005; Elidan et al.,2001; etc.]

Simplest case: latent variable
= cluster
= unary predicate

Predicate Invention
[Wogulis & Langley, 1989;
Muggleton & Buntine, 1988; etc.]

Learning MRC Model

$$\log P(\{\Gamma\}|R) \propto \log P(\{\Gamma\}) + \underbrace{\log P(R|\{\Gamma\})}$$

atom predication rule
+ wt of rule is log-odds of atom
in its cluster combination being true

$$\sum_{k \in K} (t_k + \beta) \log \frac{t_k + \beta}{t_k + f_k + 2\beta} + (f_k + \beta) \log \frac{f_k + \beta}{t_k + f_k + 2\beta}$$

set of cluster combinations

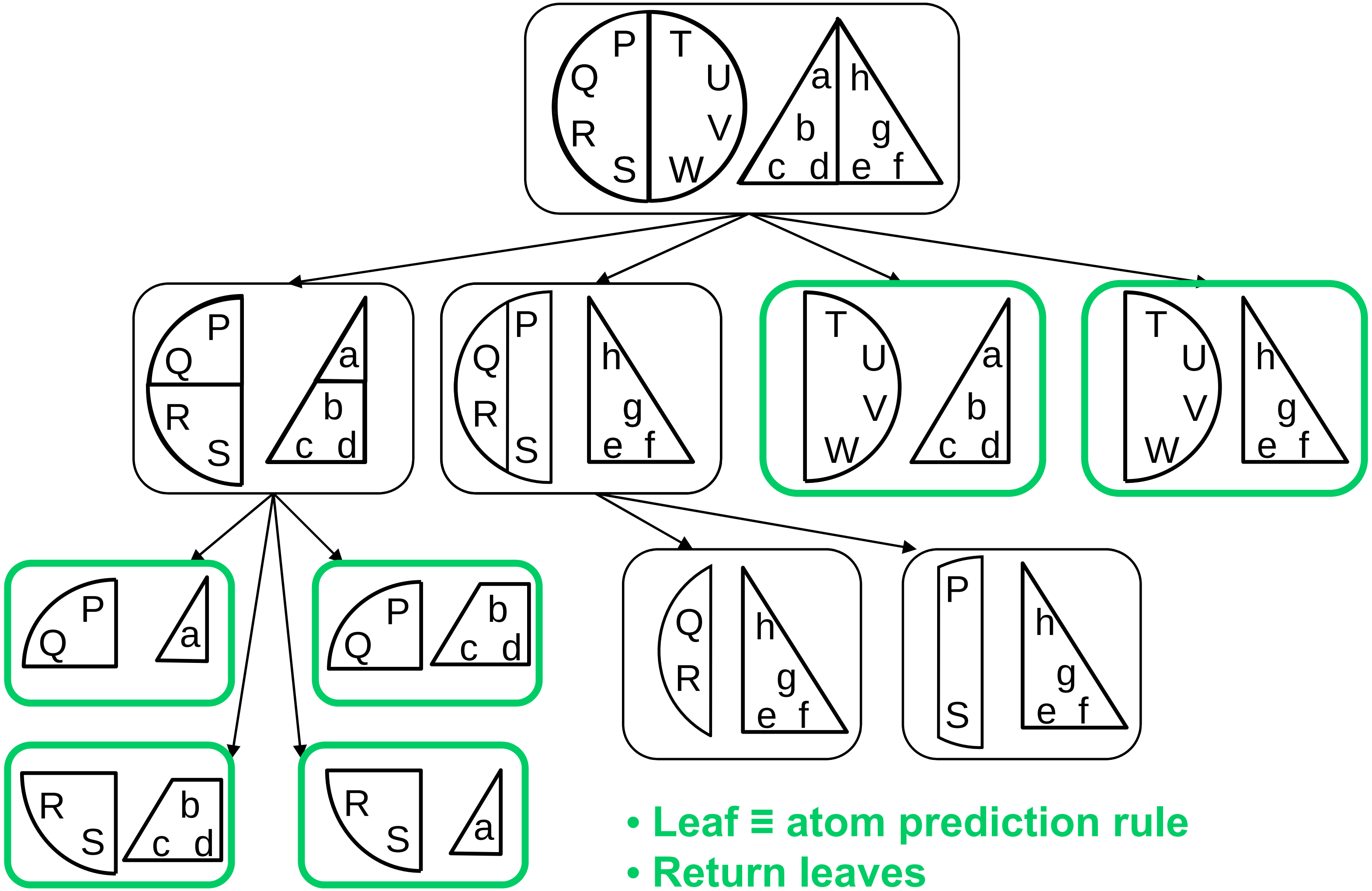
#true & #false atoms in comb.

smoothing parameter

Unknown Atoms

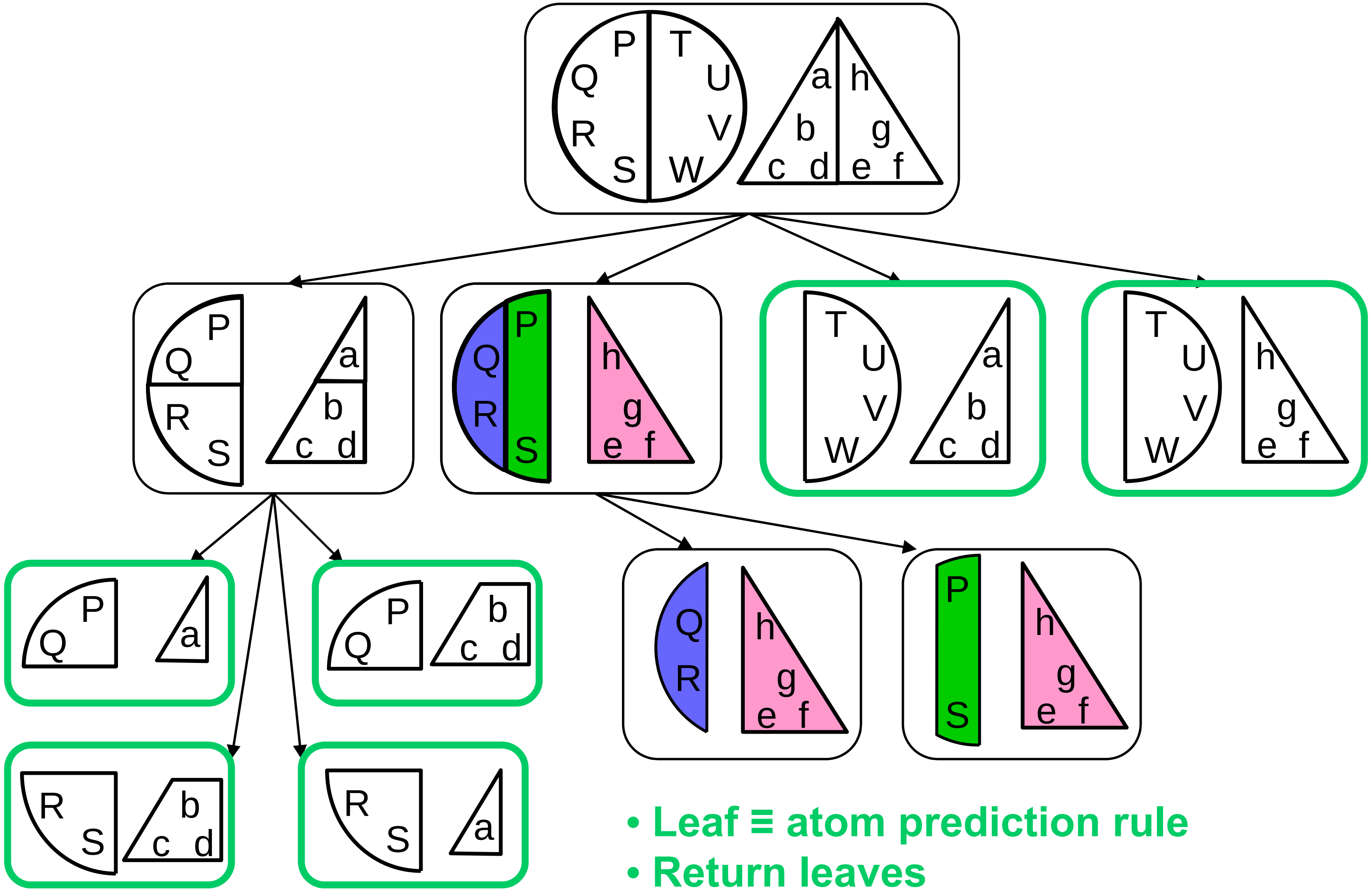
- Atoms with unknown truth values do not affect model
  - Graph-separated from all other atoms by $\{\Gamma\}$
  - Prob(*unknown atom=true*) = $\frac{t+\beta}{t+f+2\beta}$

Search Algorithm



- Leaf ≡ atom prediction rule
- Return leaves

Search Algorithm

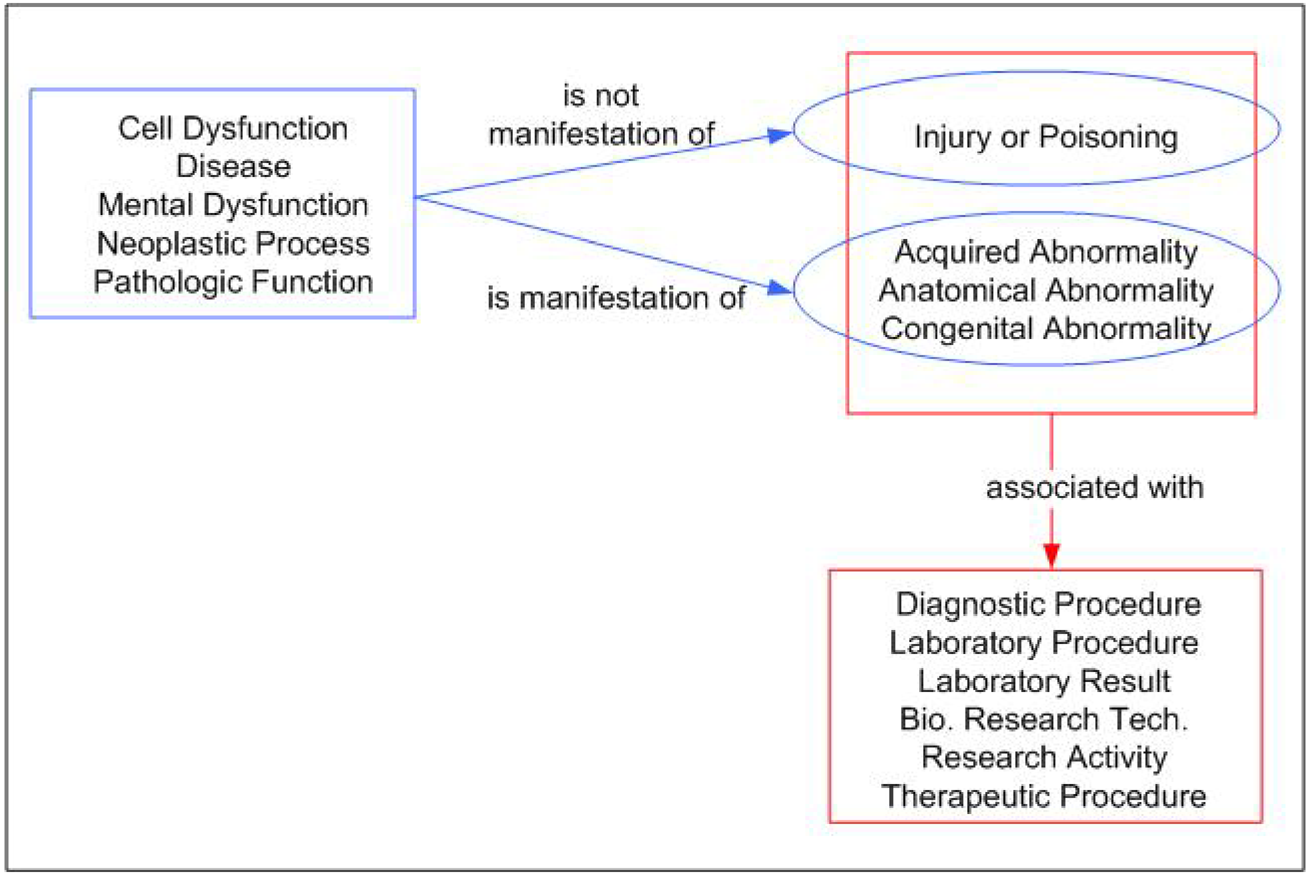


- **Leaf ≡ atom prediction rule**
- **Return leaves**

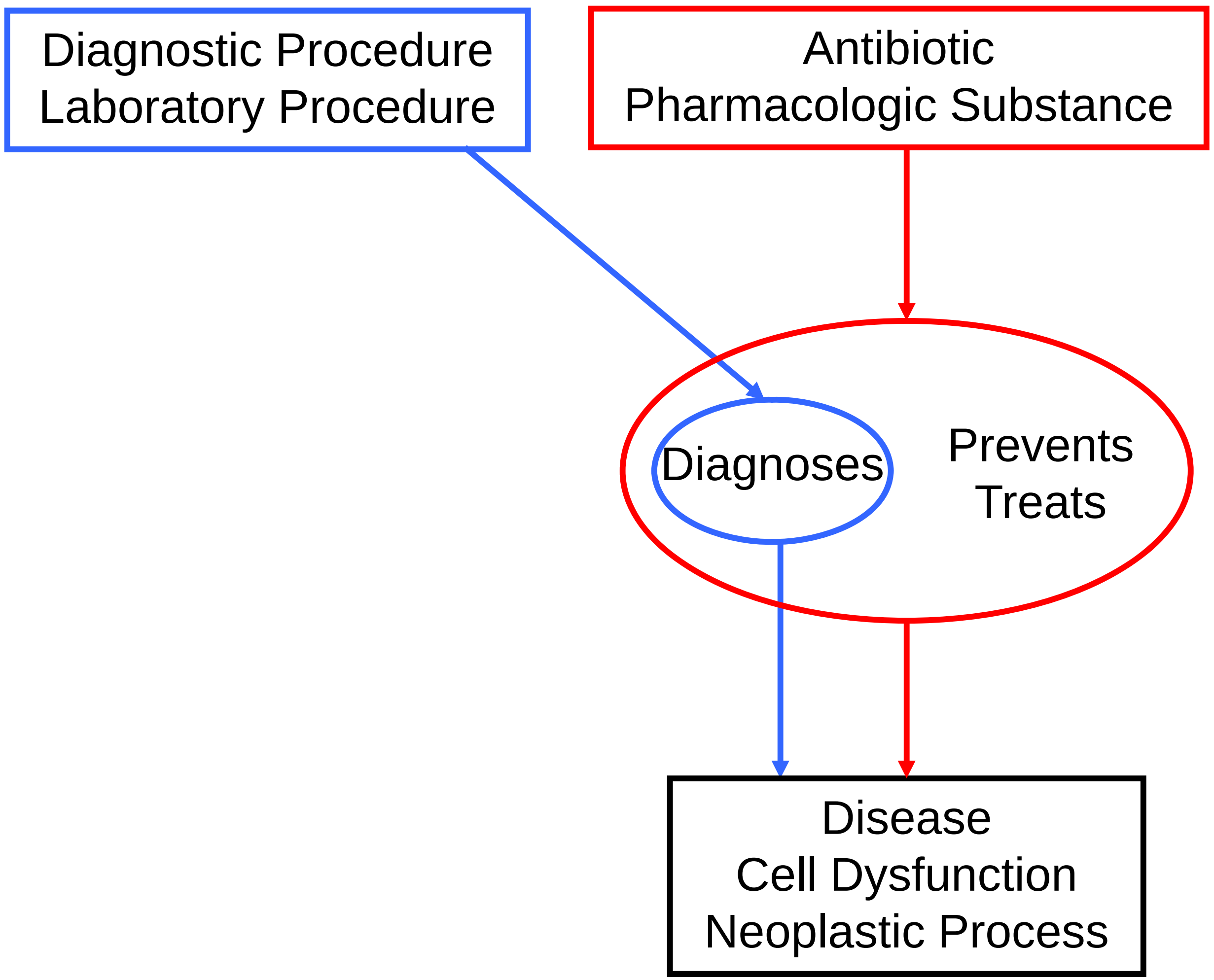
Results

- 3-5 levels of cluster refinement
- Average number of clusters
  - Animals: 202
  - UMLS: 405
  - Kinship: 1044
  - Nations: 586
- Average number of atom predication rules
  - Animals: 305
  - UMLS: 1935
  - Kinship: 3568
  - Nations: 12,169

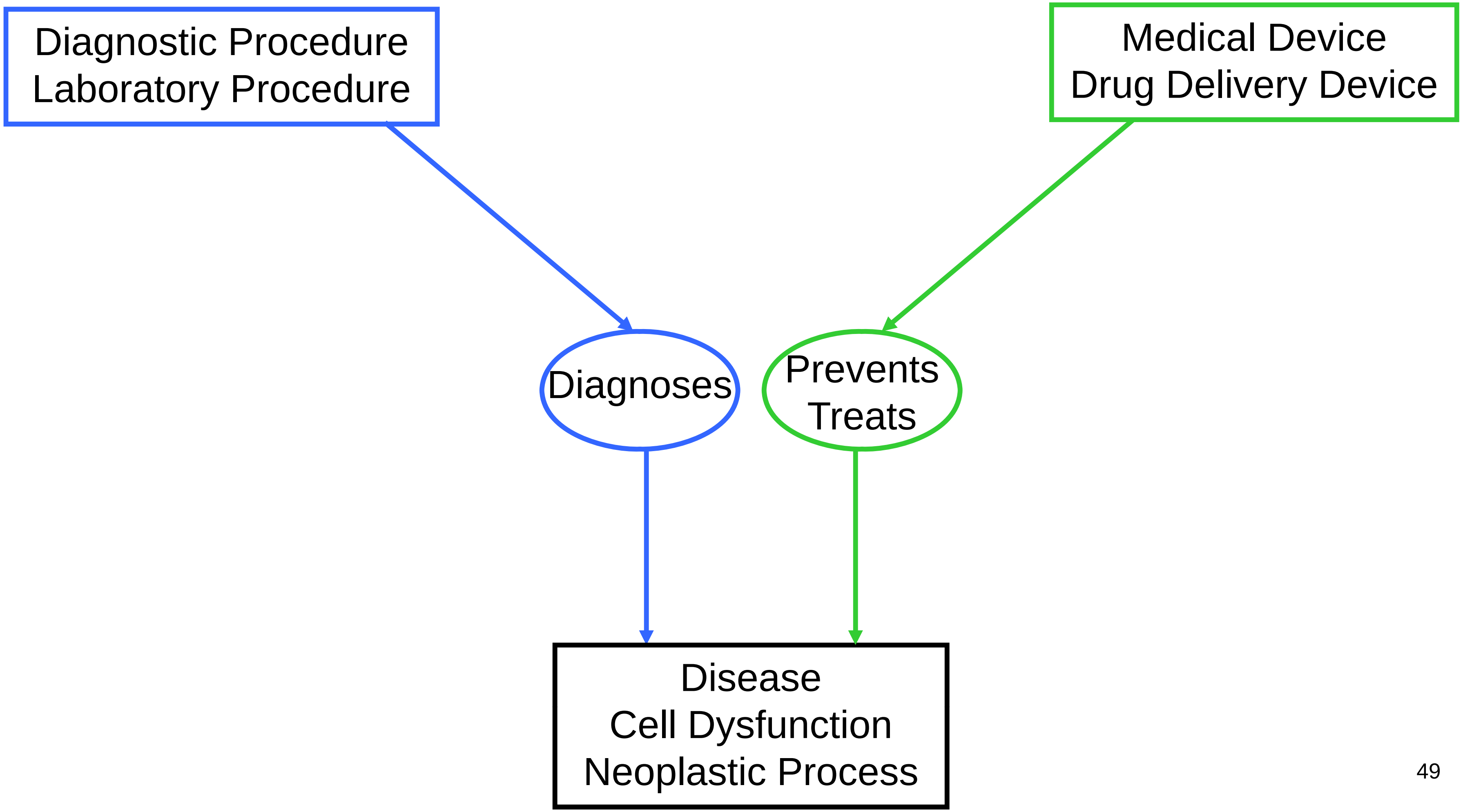
Multiple Clusterings Learned



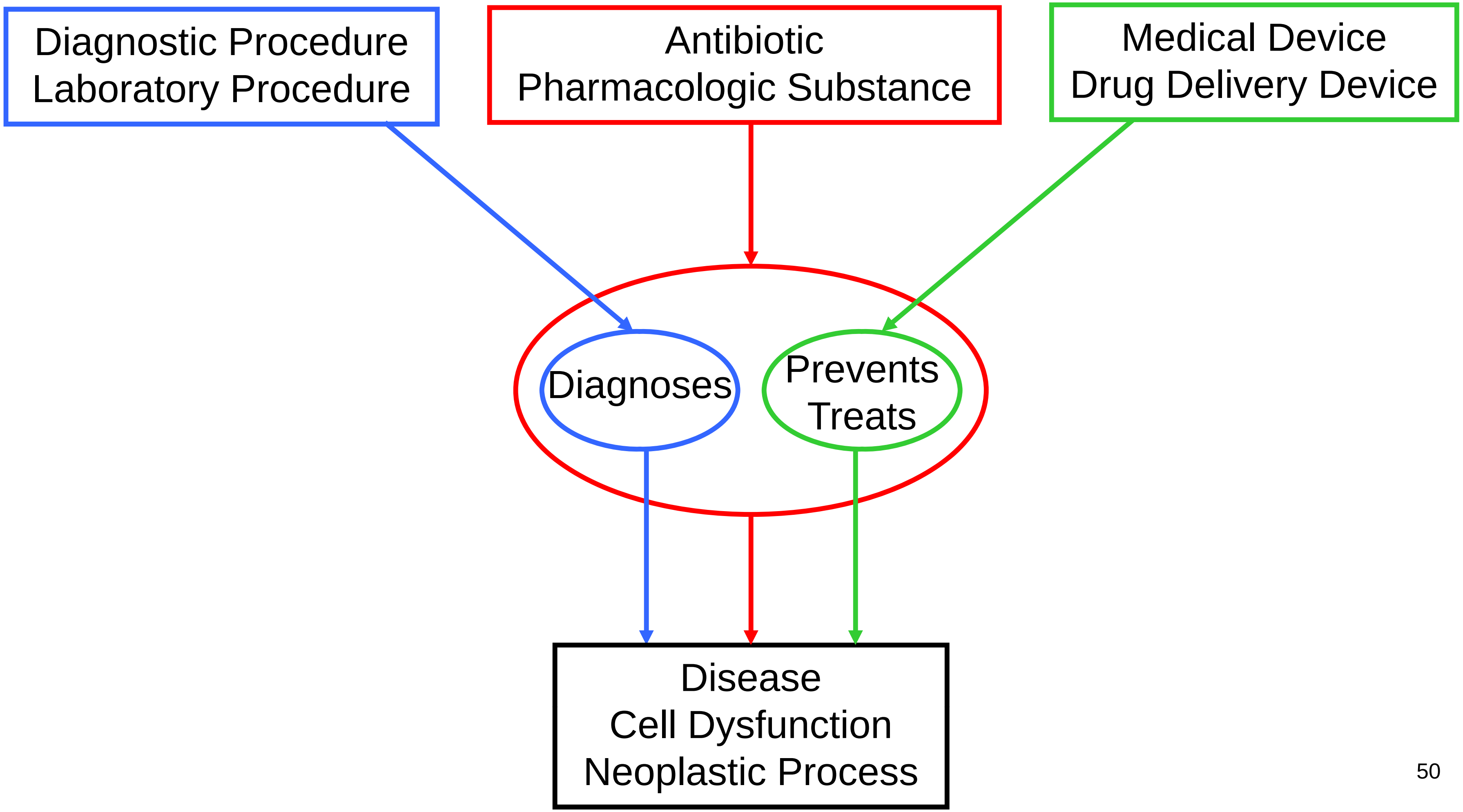
Multiple Clusterings Learned



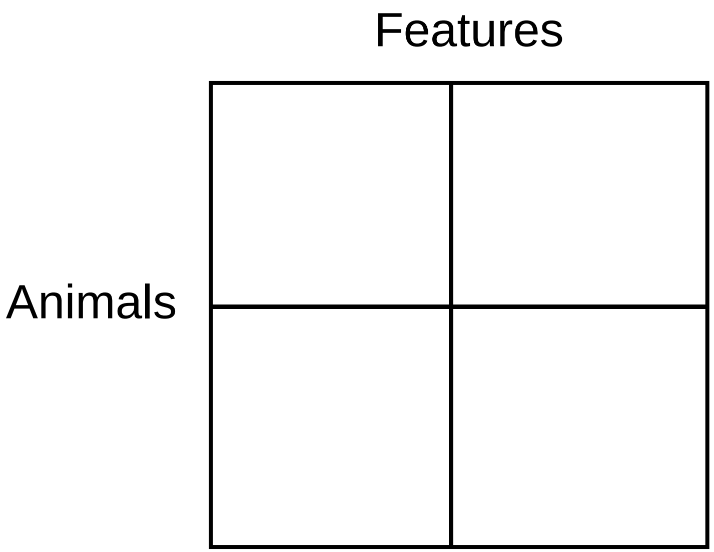
Multiple Clusterings Learned

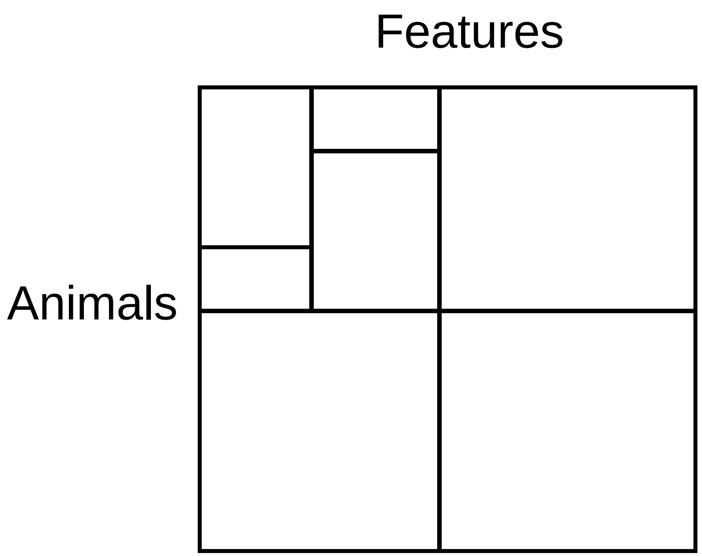


Multiple Clusterings Learned



More Flexible Schema Induction





<u>IRM (one clustering)</u>

<u>MRC (multiple clusterings)</u>