



# Methods for Redescription Mining

*Lectio praecursoria*

Esther Galbrun

Department of Computer Science  
University of Helsinki



# Menetelmiä jälleenkuvausten louhintaan

*Lectio praecursoria*

Esther Galbrun

Tietojenkäsittelytieteen laitos  
Helsingin Yliopisto



# Méthodes pour la fouille de redescriptions

*Lectio praecursoria*

Esther Galbrun

Département d'Informatique  
Université de Helsinki

# Data Mining

*Redescription Mining is a Data Mining task*

# Data Mining



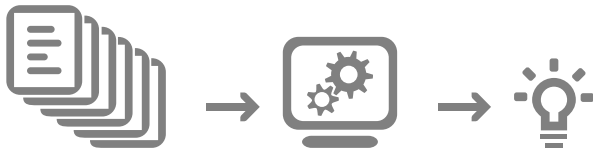
*Data Mining is a subarea of Computer Science*

# Data Mining



*The aim: Turn data into knowledge*

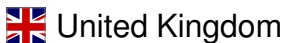
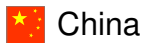
# Data Mining



*The aim: Turn lots of data into knowledge*

# Redescriptions

## An Example with World countries





# Redescriptions

## An Example with World countries

— Countries outside the Americas with land area above 8 billion square kilometers



Canada



France



Mozambique



Chile



United Kingdom



Russia



China



Mexico

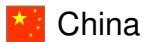


United States

# Redescriptions

## An Example with World countries

— Countries outside the Americas with land area above 8 billion square kilometers



# Redescriptions

## An Example with World countries

— Permanent members of the UN Security Council with a history of state communism



Canada



France



Mozambique



Chile



United Kingdom



Russia



China



Mexico

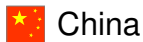


United States

# Redescriptions

## An Example with World countries

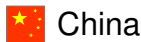
— Permanent members of the UN Security Council with a history of state communism



# Redescriptions

## An Example with World countries

- Countries outside the Americas with land area above 8 billion square kilometers
- Permanent members of the UN Security Council with a history of state communism

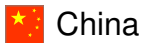


# Redescription Mining

- Countries outside the Americas with land area above 8 billion square kilometers
- Permanent members of the UN Security Council with a history of state communism

*Finding different ways  
to characterize the same things ...*

# Redescription Mining



*...finding multiple things  
that share common characterizations*








# Redescriptions

## An Example with World countries

*Let's get a bit more specific...*



# Geographic attributes

-  South Hemisphere
-  **A** Border to the Atlantic Ocean
-  **I** Border to the Indian Ocean
-  **P** Border to the Pacific Ocean
-  Continent
-  Land area
-  Highest elevation

# Geographic attributes





 South Hemisphere	No
 Border to the Atlantic Ocean	Yes
 Border to the Indian Ocean	No
 Border to the Pacific Ocean	Yes
 Continent	Americas
 Land area	9985M km <sup>2</sup>
 Highest elevation	5959 m

# Geographic attributes



## Canada

 South Hemisphere	No
 Border to the Atlantic Ocean	Yes
 Border to the Indian Ocean	No
 Border to the Pacific Ocean	Yes
 Continent	Americas
 Land area	9985M km <sup>2</sup>
 Highest elevation	5959 m



# Geographic attributes




 South Hemisphere	No
 Border to the Atlantic Ocean	Yes
 Border to the Indian Ocean	No
 Border to the Pacific Ocean	Yes
 Continent	Americas
 Land area	9985M km <sup>2</sup>
 Highest elevation	5959 m



# Geographic attributes

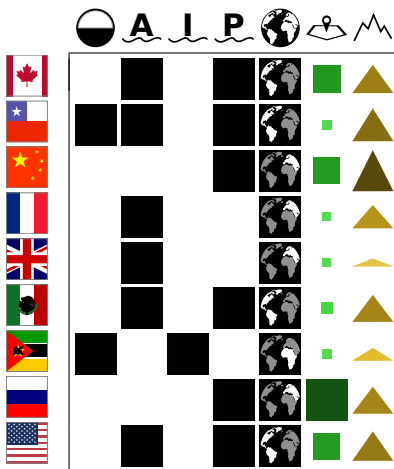


 South Hemisphere	No
 Border to the Atlantic Ocean	Yes
 Border to the Indian Ocean	No
 Border to the Pacific Ocean	Yes
 Continent	Americas
 Land area	9985M km <sup>2</sup>
 Highest elevation	5959 m

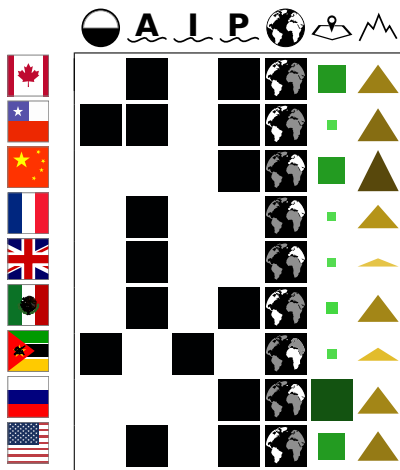


# Geographic attributes

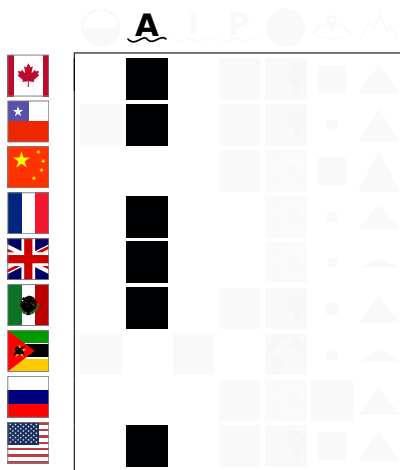
Example



# Geographic descriptions



# Geographic descriptions

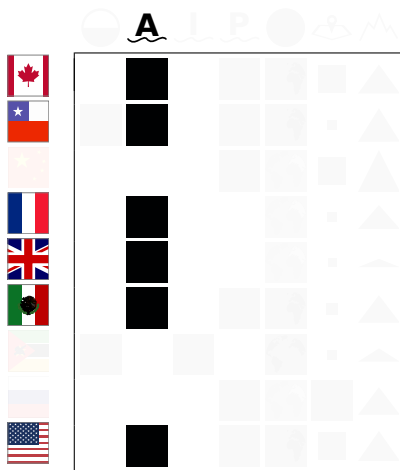


**A**

*Countries bordering  
the Atlantic Ocean*



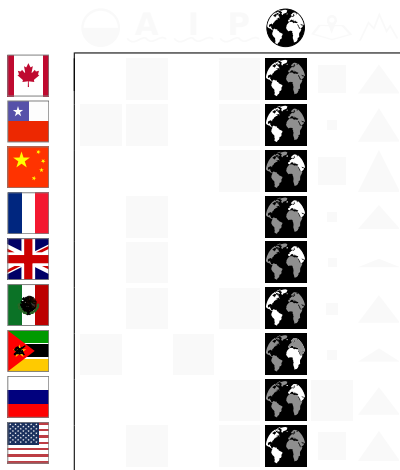
# Geographic descriptions



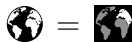
**A**

*Countries bordering  
the Atlantic Ocean*

# Geographic descriptions



A

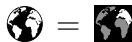


*Countries belonging  
to Eurasia*

# Geographic descriptions

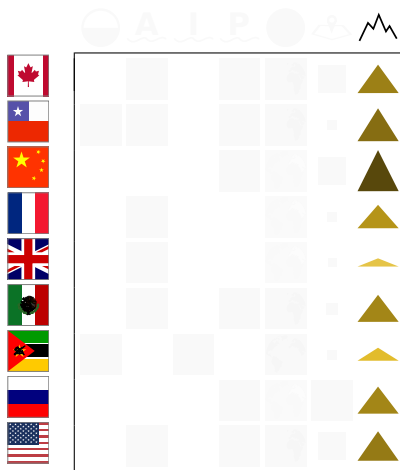


A



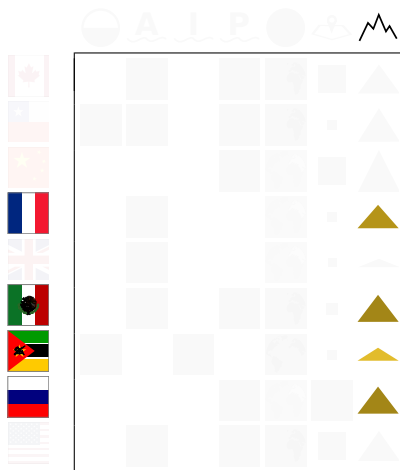
*Countries belonging  
to Eurasia*

# Geographic descriptions



*Countries with highest elevation  
between 2400 and 5600 meters*

# Geographic descriptions

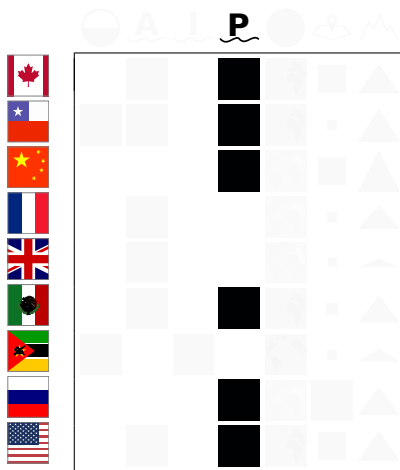


$$\overset{\mathbf{A}}{\sim} \text{Globe} = \text{Globe}$$

$$\text{Yellow Triangle} \leq \text{Mountain Range} \leq \text{Brown Triangle}$$

*Countries with highest elevation  
between 2400 and 5600 meters*

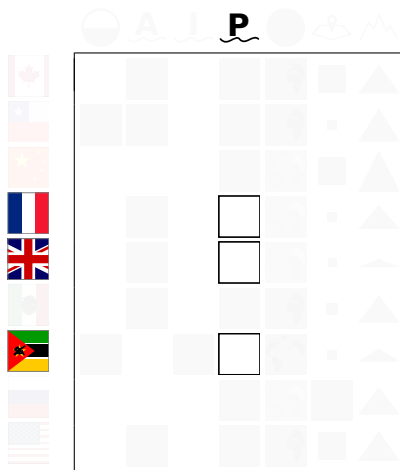
# Geographic descriptions



$\neg$  P

*Countries not bordering  
the Pacific Ocean*

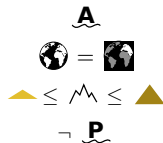
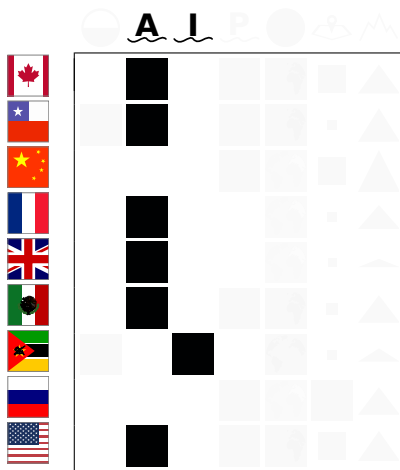
# Geographic descriptions



$\neg$  **P**

*Countries not bordering  
the Pacific Ocean*

# Geographic descriptions

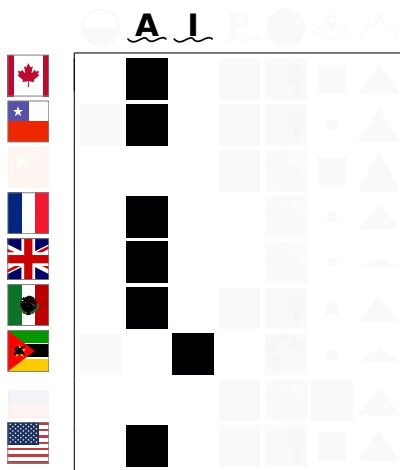


A OR I

*Countries bordering the Atlantic  
or the Indian Oceans*



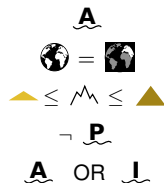
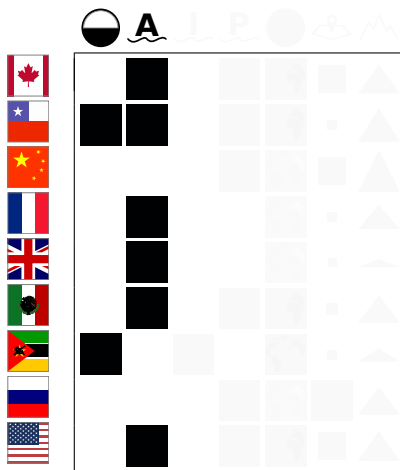
# Geographic descriptions



A OR I

*Countries bordering the Atlantic  
or the Indian Oceans*














# Geographic descriptions

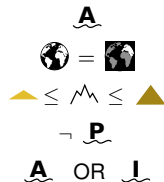


$\neg$   AND  $\overline{A}$

*Countries in the North hemisphere  
bordering the Atlantic Ocean*

# Geographic descriptions

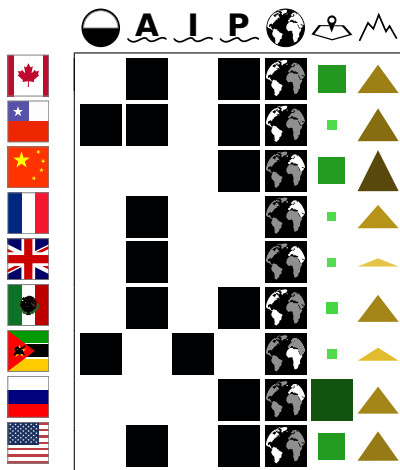
	 <b>A</b>	<b>I</b>	<b>P</b>			
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



$\neg$   AND **A**

*Countries in the North hemisphere  
bordering the Atlantic Ocean*

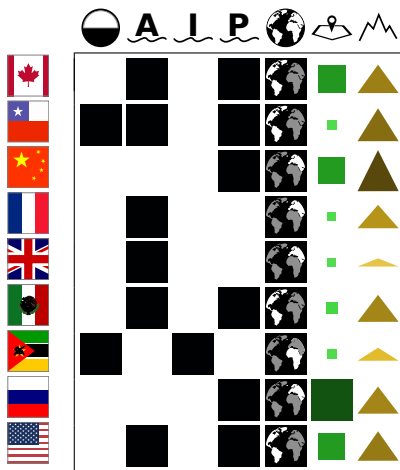
# Geographic descriptions



$$\begin{aligned}
 & \text{A} \\
 & \text{Globe} = \text{Globe} \\
 & \text{Yellow Triangle} \leq \text{Mountain} \leq \text{Yellow Triangle} \\
 & \neg \text{P} \\
 & \text{A} \text{ OR } \text{I} \\
 & \neg \text{Black Circle} \text{ AND } \text{A}
 \end{aligned}$$

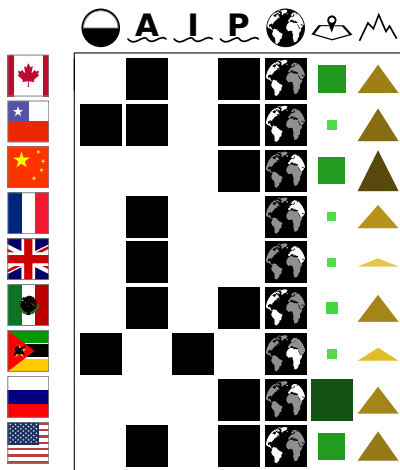
$$\text{Globe} \neq \text{Globe} \text{ AND } \text{Green Square} \leq \text{Location Pin}$$

# Geographic descriptions



$$\begin{aligned}
 & \overline{\mathbf{A}} \\
 & \text{Globe} = \text{Globe} \\
 & \text{Yellow Triangle} \leq \text{Mountain Range} \leq \text{Yellow Triangle} \\
 & \neg \mathbf{P} \\
 & \mathbf{A} \text{ OR } \mathbf{I} \\
 & \neg \text{Half-Black Circle} \text{ AND } \mathbf{A} \\
 & \text{Globe} \neq \text{Globe} \text{ AND } \text{Green Square} \leq \text{Location Pin} \\
 & ( \text{Globe} \neq \text{Globe} \text{ OR } \text{Green Square} \leq \text{Location Pin} \leq \text{Green Square} ) \\
 & \text{AND } \mathbf{A}
 \end{aligned}$$

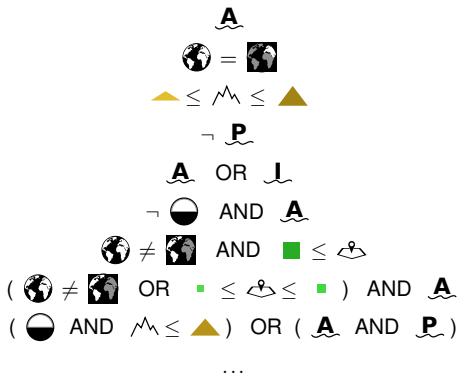
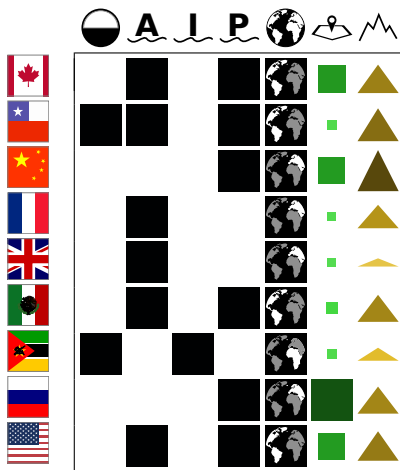
# Geographic descriptions



$$\begin{aligned}
 & \underline{A} \\
 & \text{globe} = \text{globe} \\
 & \text{yellow triangle} \leq \text{mountain range} \leq \text{yellow triangle} \\
 & \neg \underline{P} \\
 & \underline{A} \text{ OR } \underline{I} \\
 & \neg \text{circle with line} \text{ AND } \underline{A} \\
 & \text{globe} \neq \text{globe} \text{ AND } \text{green square} \leq \text{location pin} \\
 & ( \text{globe} \neq \text{globe} \text{ OR } \text{green square} \leq \text{location pin} \leq \text{green square} ) \text{ AND } \underline{A}
 \end{aligned}$$

$$\begin{aligned}
 & ( \text{circle with line} \text{ AND } \text{mountain range} \leq \text{yellow triangle} ) \\
 & \text{OR } ( \underline{A} \text{ AND } \underline{P} )
 \end{aligned}$$

# Geographic descriptions



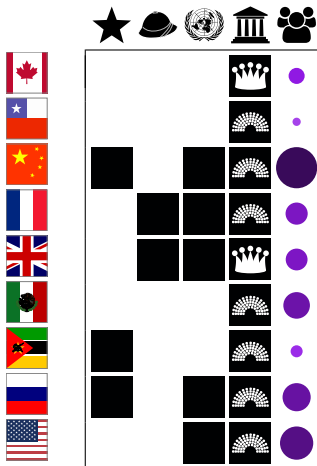
# Geopolitical attributes

- ★ History of state communism
- 🌐 History of colonialism
- 🇺🇳 Permanent member of the UNSC
- 🏛️ Type of government
- 👥 Population

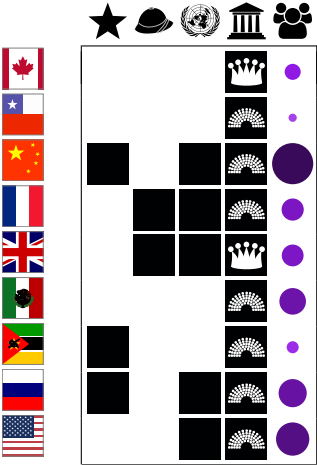


# Geopolitical attributes

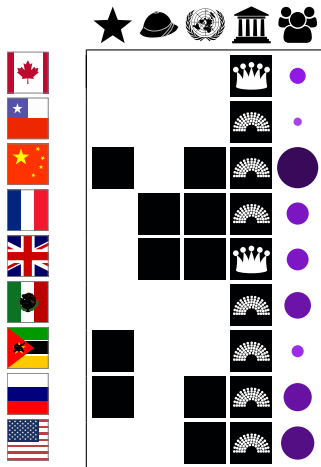
Example



# Geopolitical descriptions

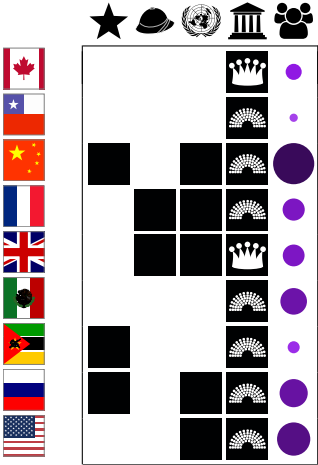


# Geopolitical descriptions



*Countries with a history of state communism*

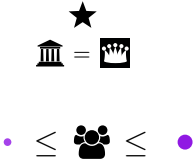
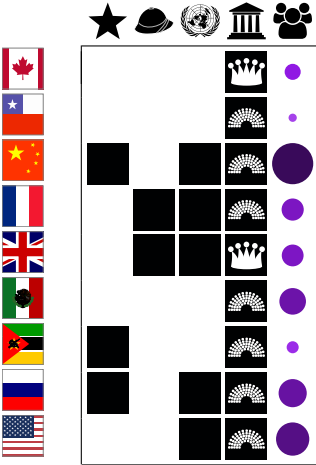
# Geopolitical descriptions



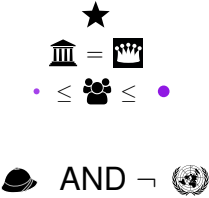
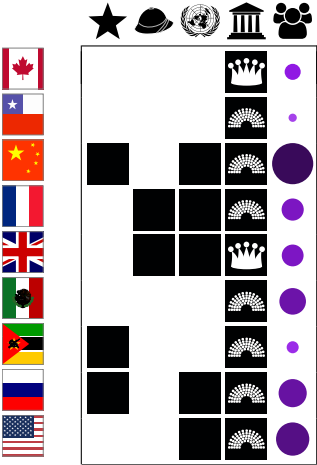
★

 =   
*Monarchies*

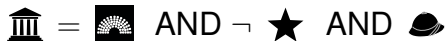
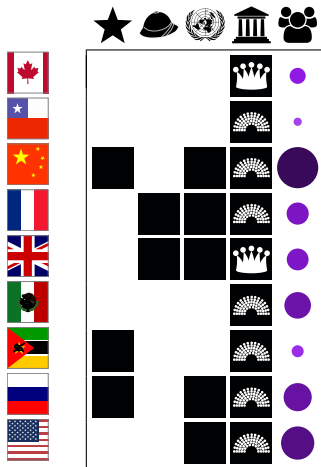
# Geopolitical descriptions



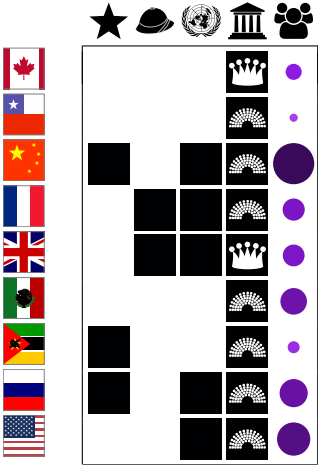
# Geopolitical descriptions



# Geopolitical descriptions

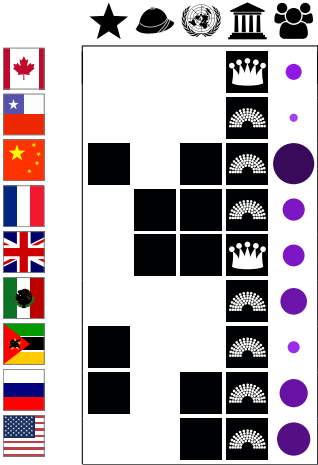


# Geopolitical descriptions

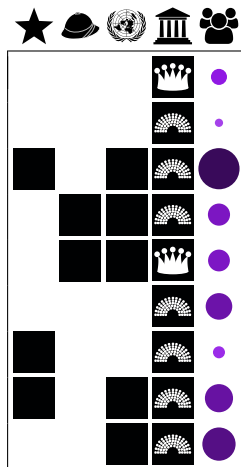




# Geopolitical descriptions

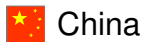


## Two views on the objects

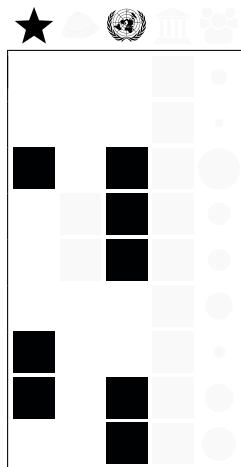
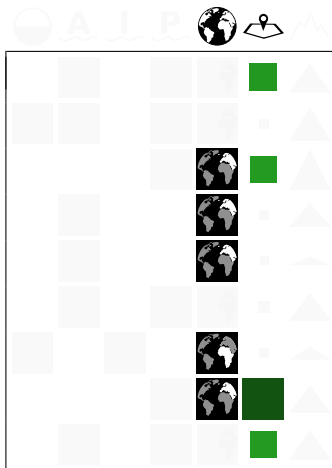


# Redescriptions

- Countries outside the Americas with land area above 8 billion square kilometers
- Permanent members of the UN Security Council with a history of state communism



# Redescriptions

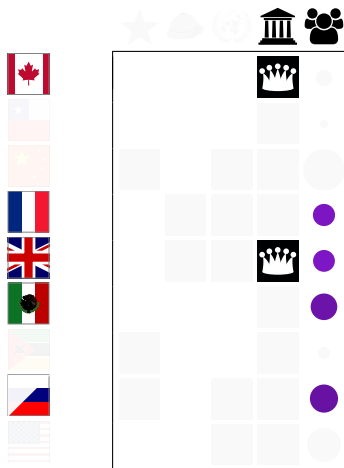
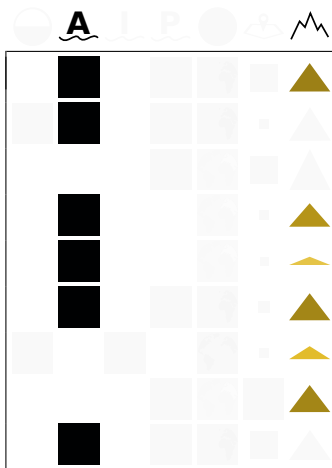


# Redescriptions

Example

**A** AND  $\leq$  

 =  OR   $\leq$    $\leq$  



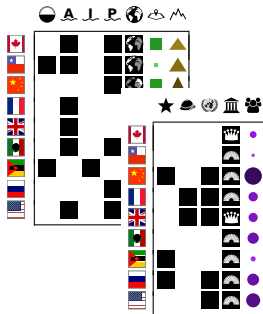
# Redescription Mining

*Finding different ways  
to characterize the same things,*

*and vice-versa,*

*finding multiple things  
that share common characterizations.*

# Redescription Mining



$$\left( \text{globe} \neq \text{globe} \text{ AND } \text{green square} \leq \text{person}, \right. \\
 \left. \text{star} \text{ AND } \text{globe} \right) \\
 \{ \text{China}, \text{Russia} \}$$

$$\left( \text{A} \text{ AND } \text{mountain} \leq \text{triangle}, \right. \\
 \left. \text{building} = \text{crown} \text{ OR } \text{purple circle} \leq \text{paw print} \leq \text{purple circle} \right) \\
 \{ \text{Canada}, \text{France}, \text{UK}, \text{Germany}, \text{Czech Republic} \}$$

...

# Applications Domains

- Bioclimatic niches

*Scandinavia and Baltia, which are characterized by their specific cold climate, are the habitat of the European Elk*



# Applications Domains

- Bioclimatic niches
- Political candidates profiles

*Candidates to the 2011 Finnish parliamentary election below age sixty accord little importance to the question of pension indices*

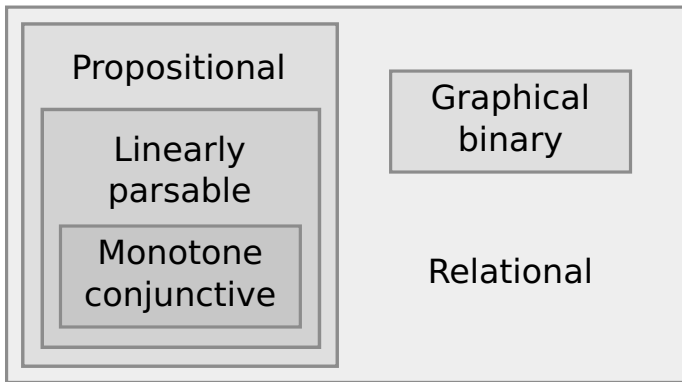
# Applications Domains

- Bioclimatic niches
- Political candidates profiles
- Computer science bibliography

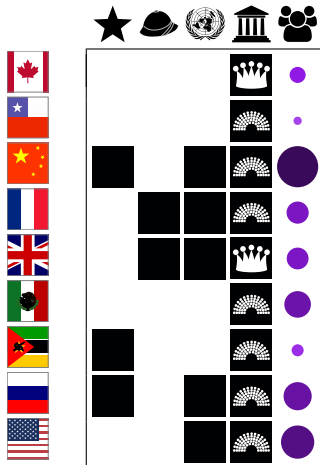
*Researchers with multiple publications in SoCG and CCCG conferences often collaborate with Profs M. Overmars or E. D. Demaine*

# Methods for Redescription Mining

## Query Languages (Chapter 3)



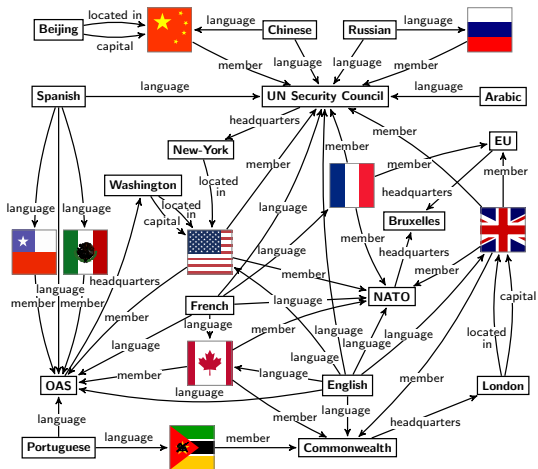
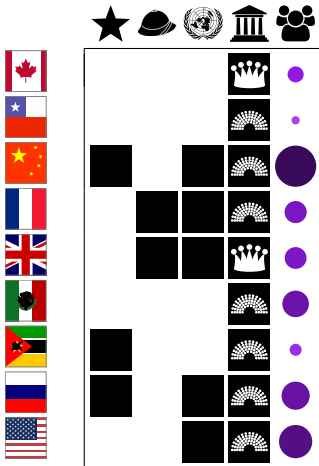
# Geopolitical attributes



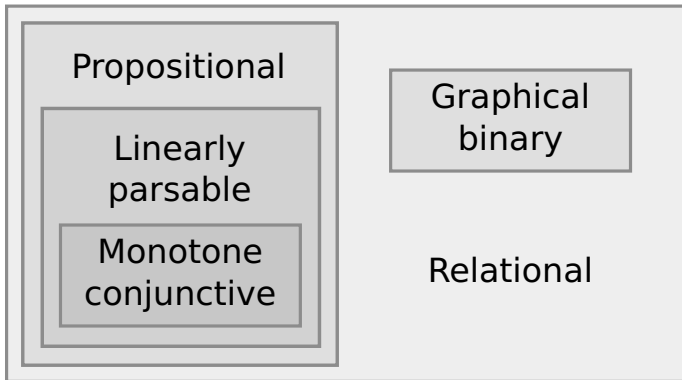
# Geopolitical attributes

# Geopolitical relations

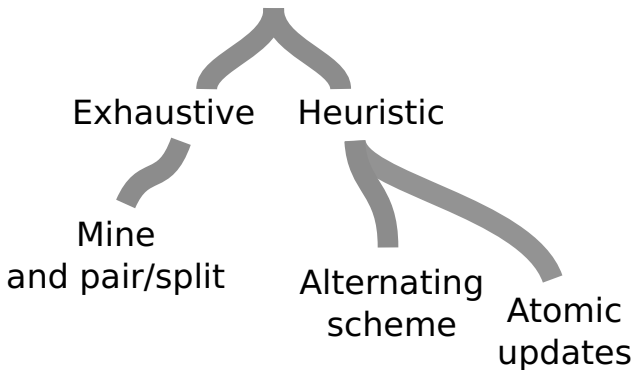
Example



# Query Languages (Chapter 3)



# Exploration strategies (Chapter 4)





# Pattern Selection (Chapter 5)

- Measuring quality
  - Queries and support
  - Individually or as a set
- Enforcing quality constraints



Parameter-full

Parameter-free

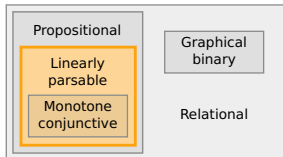
# Article I

## *From Black and White to Full Color: Extending Redescription Mining Outside the Boolean World*

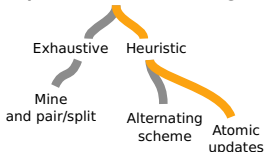
with Pauli Miettinen,  
In *Statistical Analysis and Data Mining*, 2012.

Handling real-values natively with the ReReMi algorithm.

## Query Languages



## Exploration Strategies



## Pattern Selection



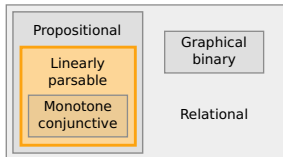
# Article II

## *A Case of Visual and Interactive Data Analysis: Geospatial Redescription Mining*

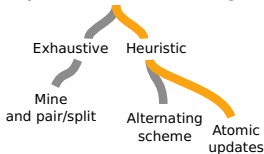
with Pauli Miettinen,  
In *Instant Interactive Data Mining Workshop*  
at *ECML/PKDD*, 2012.

The Siren interactive interface for ReReMi.

## Query Languages



## Exploration Strategies



## Pattern Selection



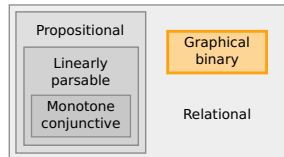
# Article III

## *Finding Relational Redescriptions*

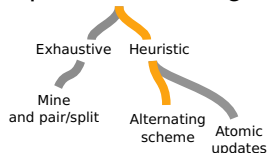
with Angelika Kimmig,  
In *Machine Learning*, 2013.

Extending redescription mining to the relational case.

## Query Languages



## Exploration Strategies



## Pattern Selection



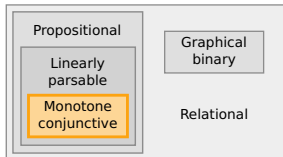
# Article IV

## *Compression-based Association Discovery in Two-View Data*

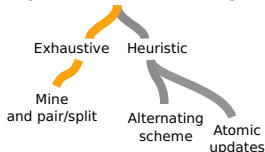
with Matthijs van Leeuwen,  
Submitted for review.

Evaluating the quality of sets of redescriptions.

## Query Languages



## Exploration Strategies



## Pattern Selection

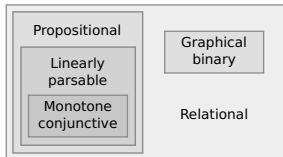
Parameter-full

Parameter-free

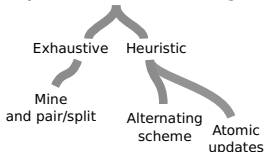
# Methods for Redescription Mining

- I From Black and White to Full Color: Extending Redescription Mining Outside the Boolean World
- II A Case of Visual and Interactive Data Analysis: Geospatial Redescription Mining
- III Finding Relational Redescriptions
- IV Compression-based Association Discovery in Two-View Data

## Query Languages



## Exploration Strategies



## Pattern Selection

Parameter-full

Parameter-free