Redescription Mining

An Introduction

Esther Galbrun

BOSTON UNIVERSITY

Department of Computer Science, Boston University

December, 2014

2009-2013

UNIVERSITY OF HELSINKI

FACULTY OF SCIENCE

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

2014

BOSTON UNIVERSITY

Redescriptions
An Example with World countries

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

 Canada
 Chile
 China
 France
 United Kingdom
 Mexico
 Mozambique
 Russia
 United States

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

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

 Canada
 Chile
 China
 France
 United Kingdom
 Mexico
 Mozambique
 Russia
 United States

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

— 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

Canada
France
Mozambique
Chile
United Kingdom
Russia
China
Mexico
United States

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

Canada
Chile
China

France
United Kingdom
Mexico

Mozambique
Russia
United States

…finding multiple things
that share common characterizations

Redescriptions
An Example with World countries

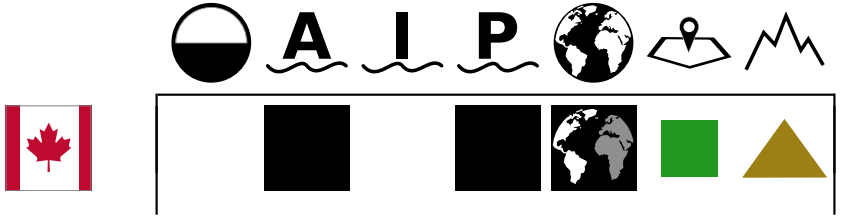
Let's get a bit more specific…

Geographic attributes

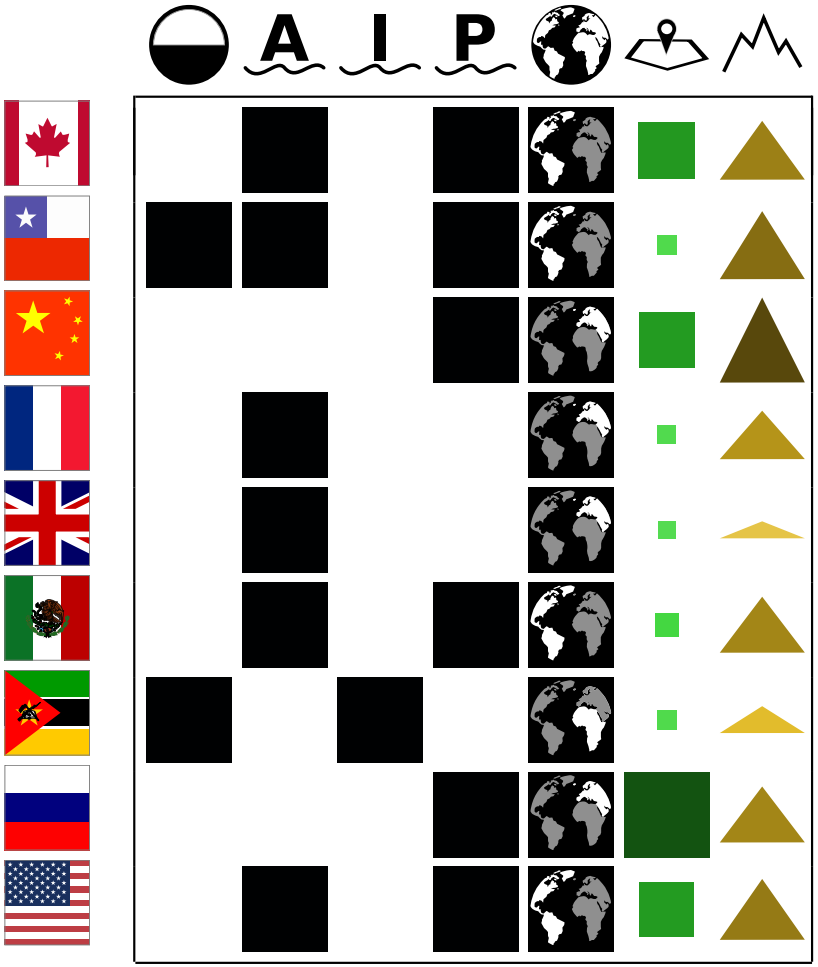
- South Hemisphere
- Border to the Atlantic Ocean
- Border to the Indian Ocean
- Border to the Pacific Ocean
- Continent
- Land area
- Highest elevation

Geographic attributes

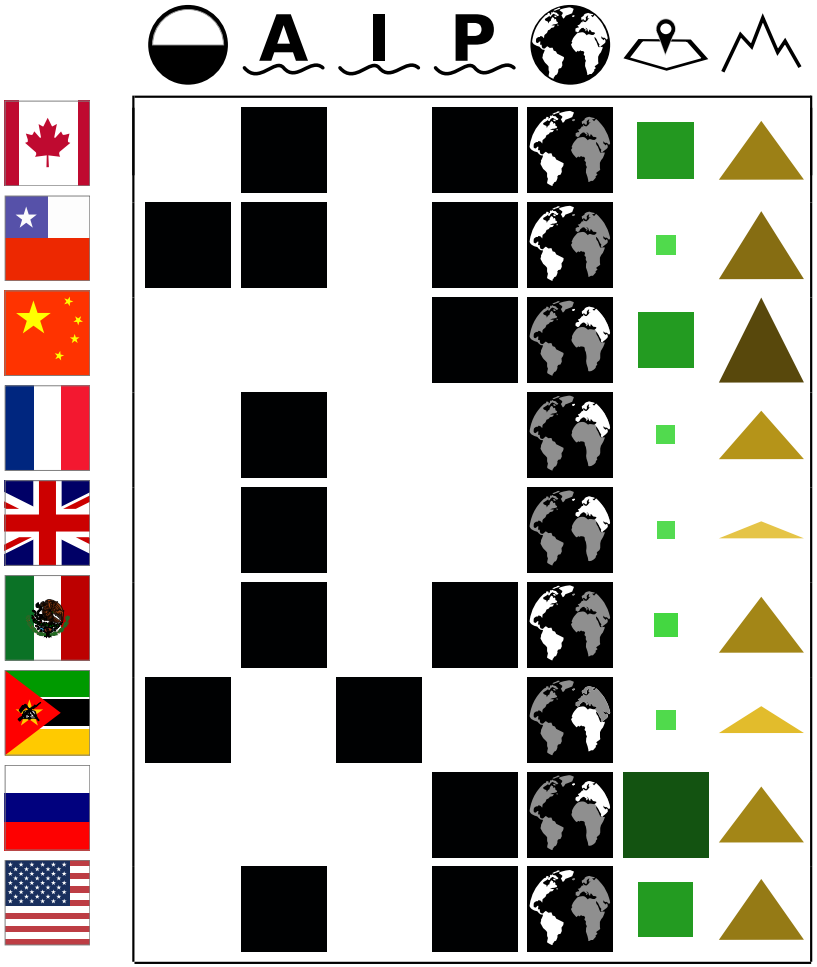
	Canada
South Hemisphere	No
Border to the Atlantic Ocean	Yes
Border to the Indian Ocean	No
Border to the Pacific Ocean	Yes
Continent	Americas
Land area	9 985 M km^2
Highest elevation	5959 m



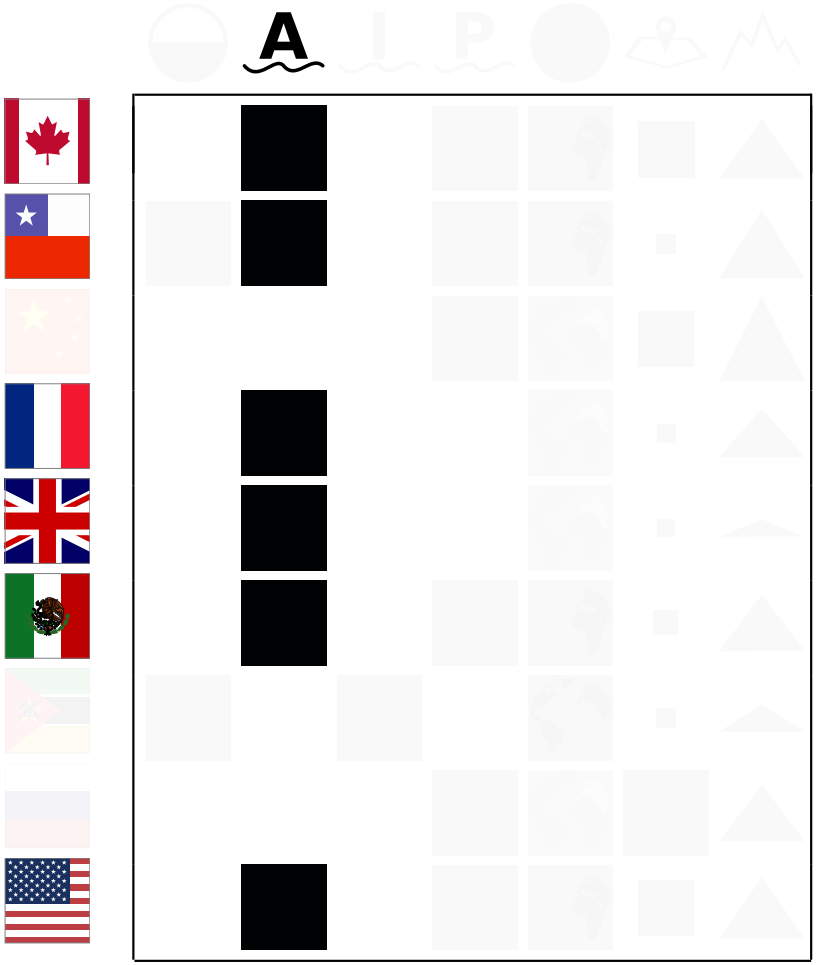
Geographic attributes



Geographic descriptions



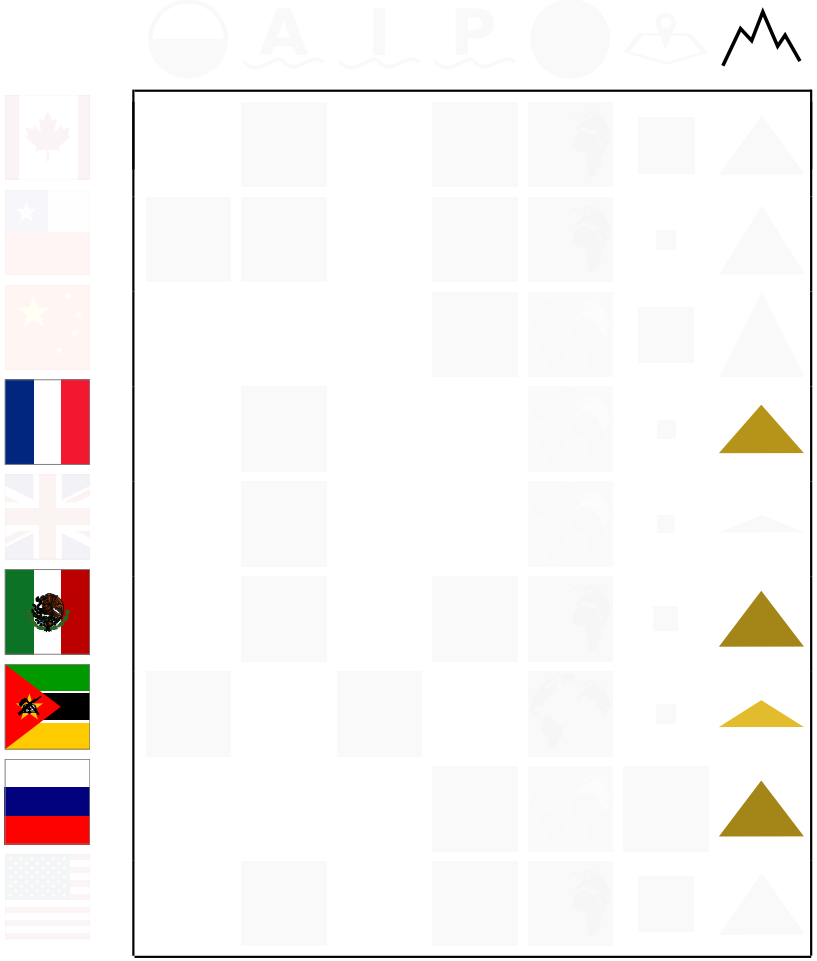
Geographic descriptions



A

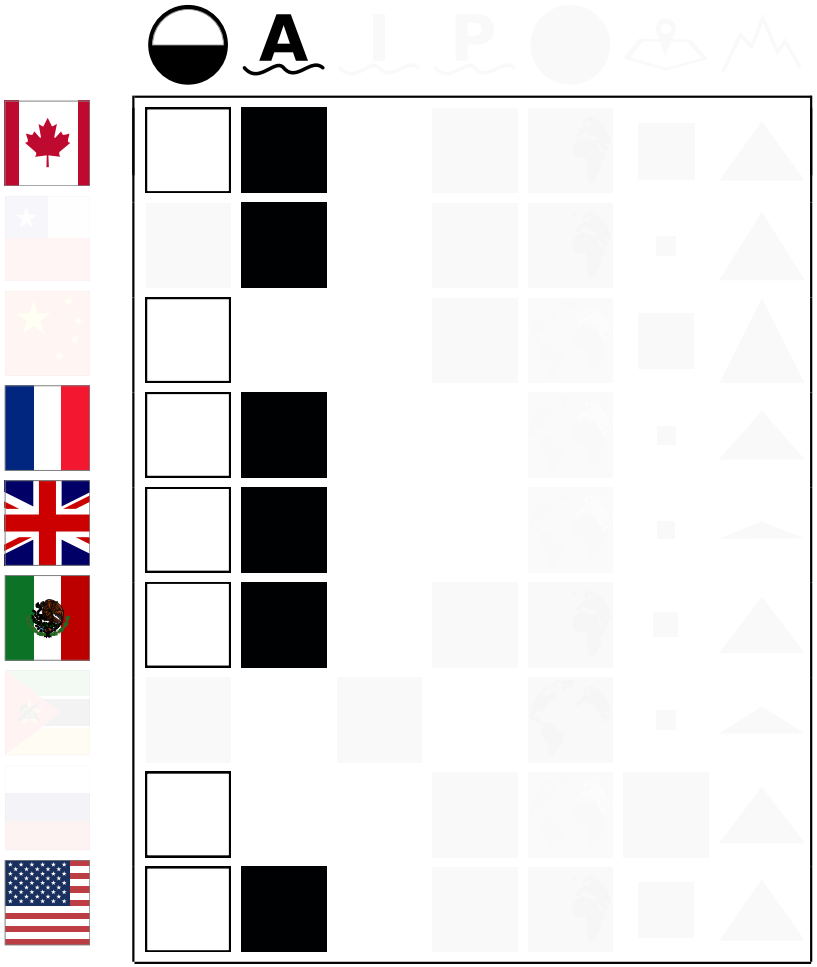
Countries bordering the Atlantic Ocean

Geographic descriptions



Countries with highest elevation between 2400 and 5600 meters

Geographic descriptions



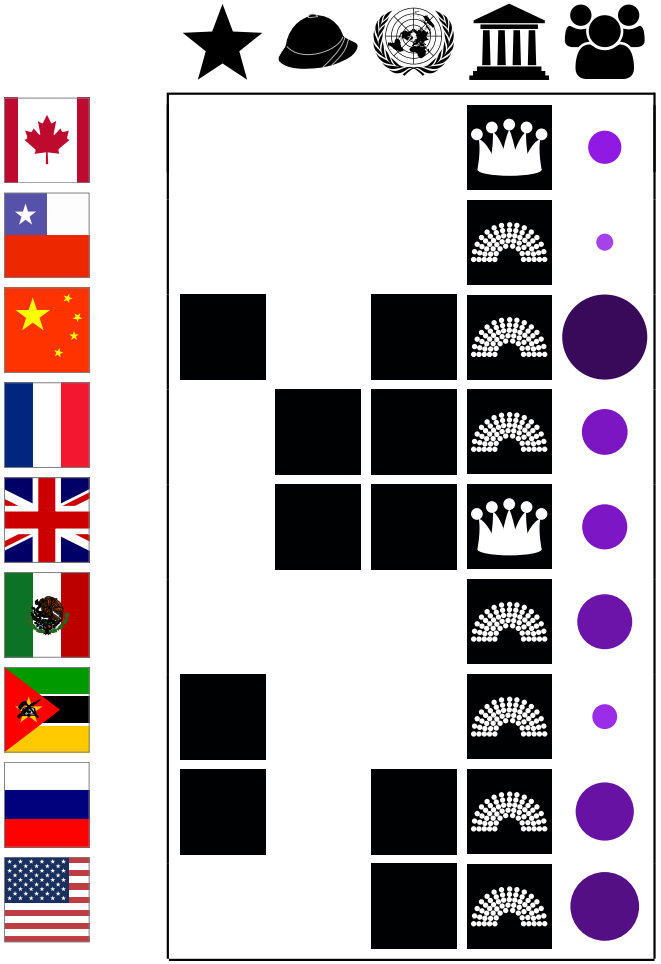
¬ AND A

Countries in the North hemisphere bordering the Atlantic Ocean

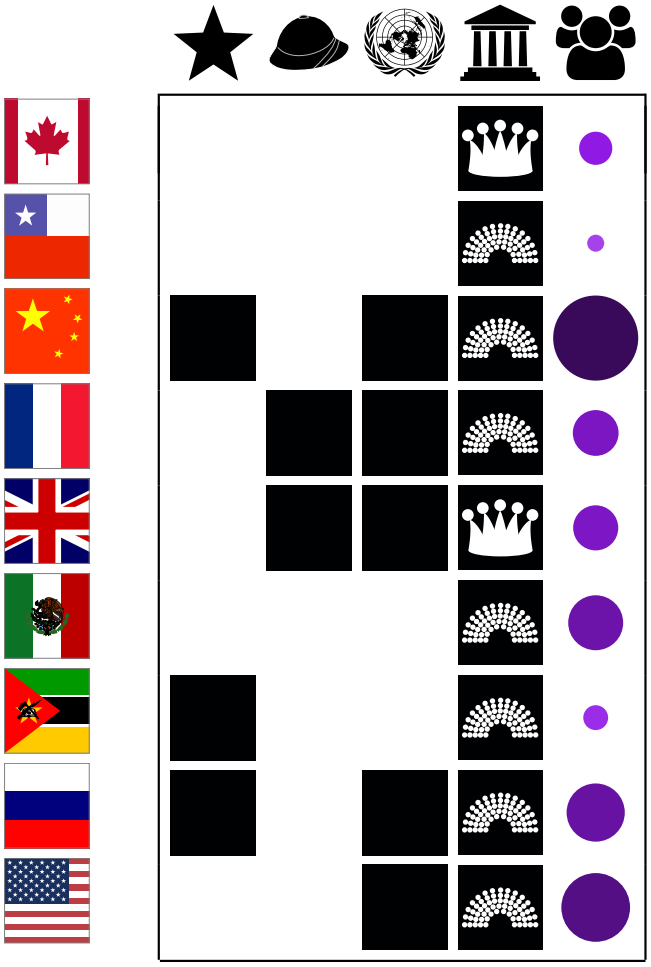
Geopolitical attributes

- History of state communism
- History of colonialism
- Permanent member of the UNSC
- Type of government
- Population

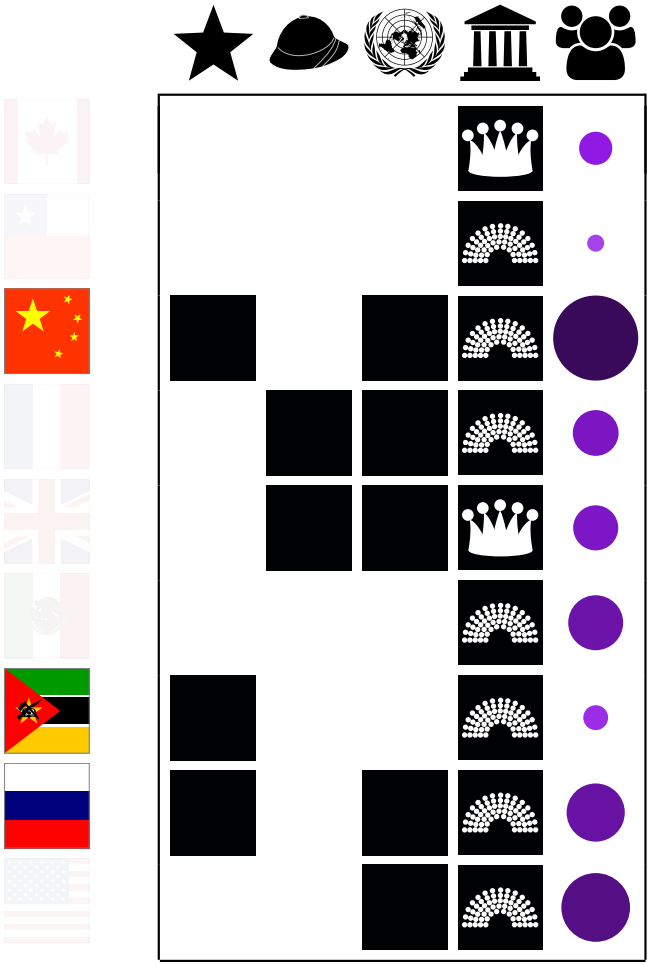
Geopolitical attributes



Geopolitical descriptions



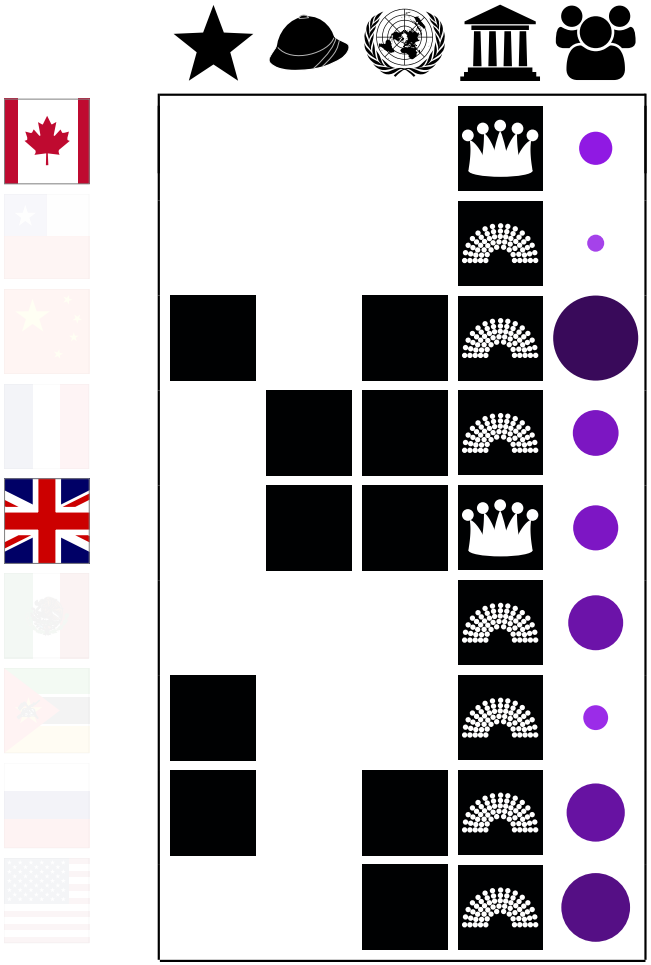
Geopolitical descriptions

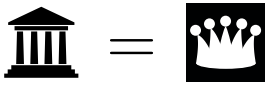


★

Countries with a history of state communism

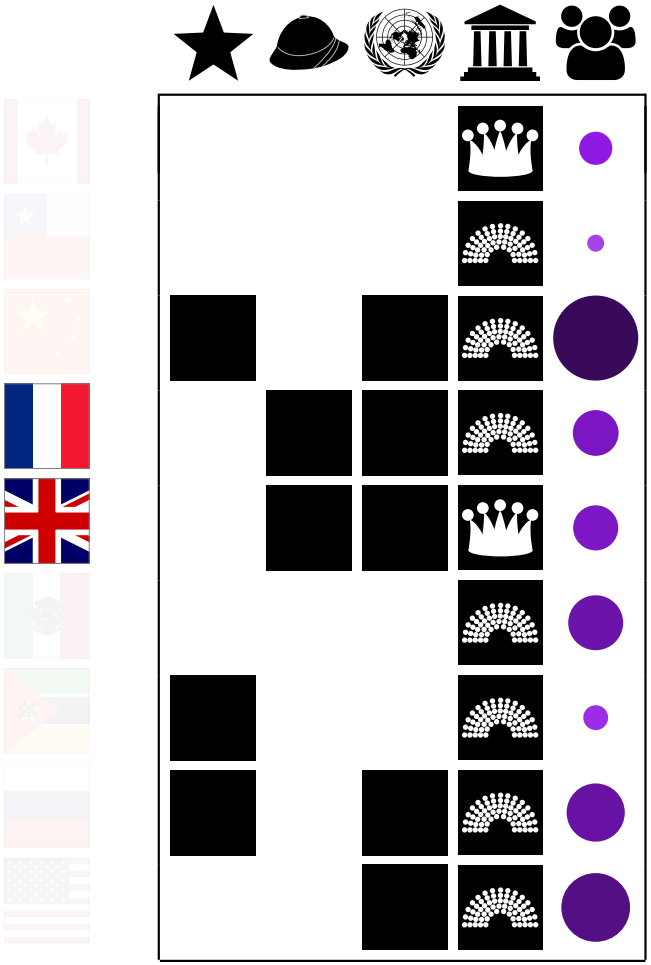
Geopolitical descriptions





Monarchies

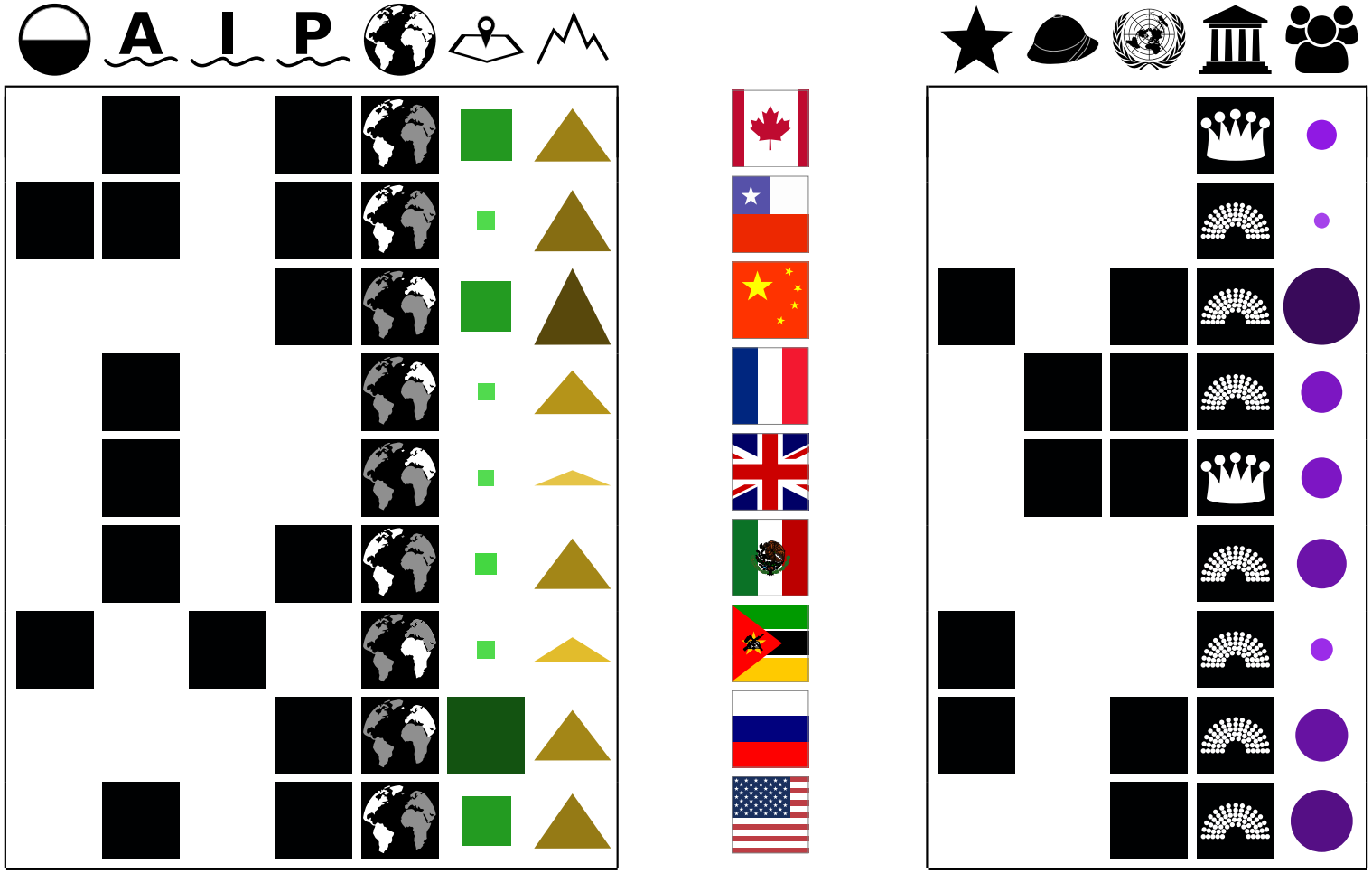
Geopolitical descriptions



AND

Countries with a history of colonialism members of UNSC

Two views on the objects

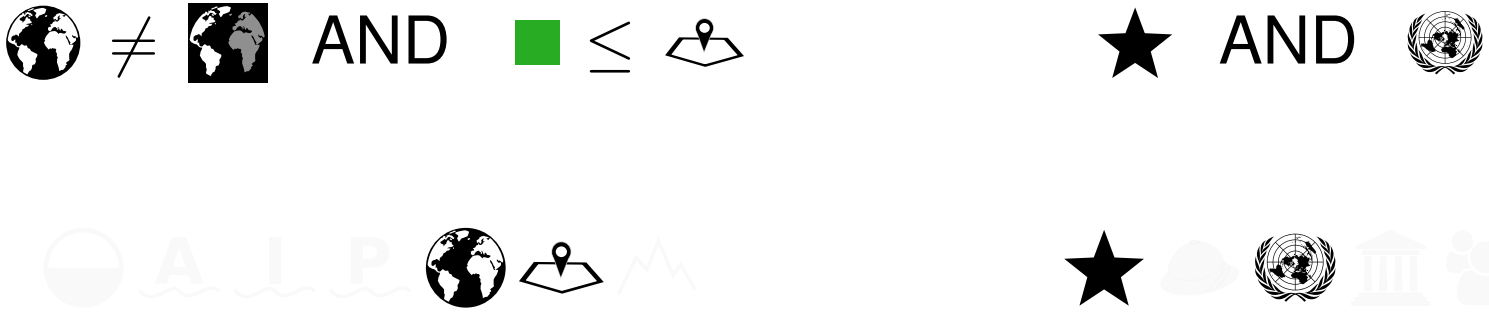


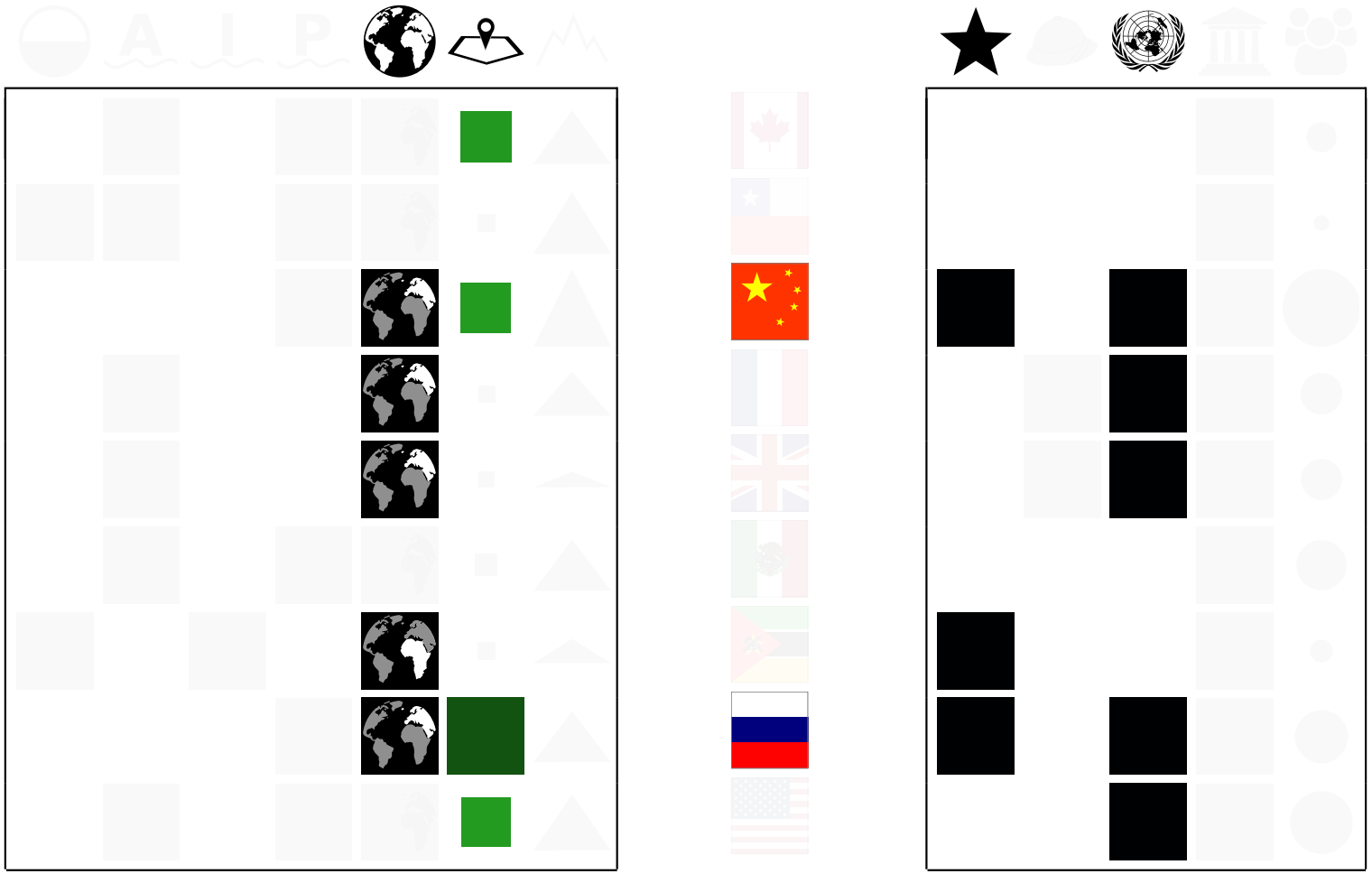
Redescriptions

— 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

Canada
Chile
China
France
United Kingdom
Mexico
Mozambique
Russia
United States

Redescriptions



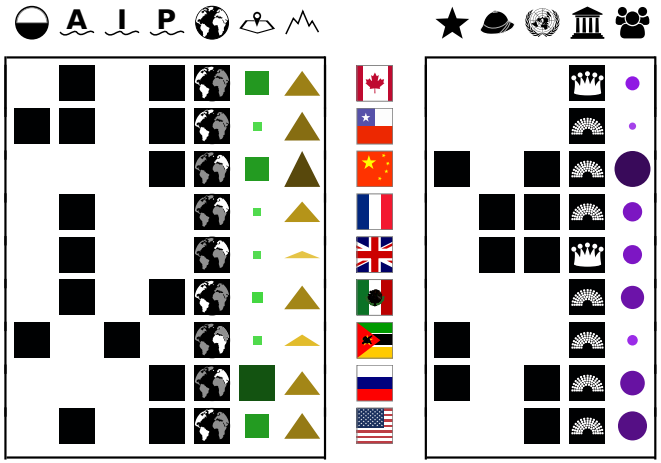


Definitions

Redescription Given two datasets with identity between the rows, a **redescription** is a pair of queries $(q_{\mathbf{L}}, q_{\mathbf{R}})$ over the columns characterizing approximately the same sets of rows.

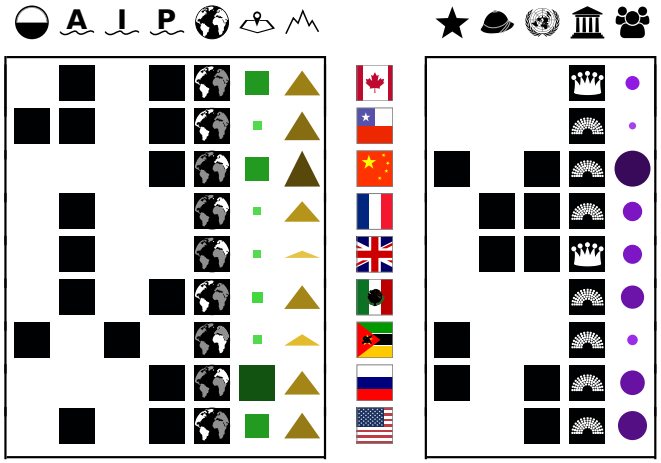
Redescription Mining Given such a pair of datasets and a set of constraints, find the best redescriptions satisfying the constraints.

Definitions



Dataset Two data matrices

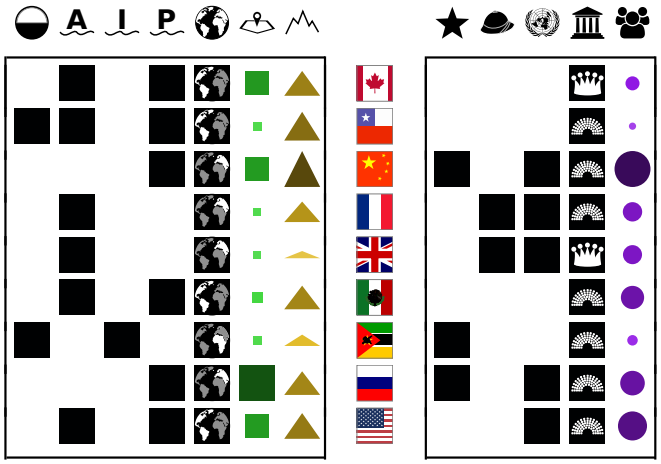
Definitions



Dataset Two data matrices

Queries Logical formulae

Definitions



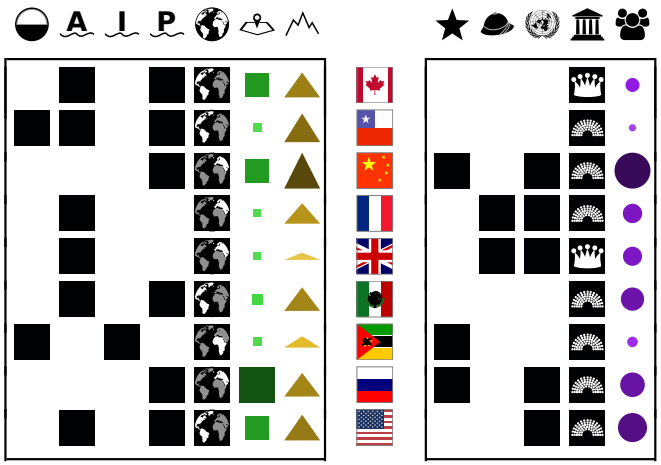
Dataset Two data matrices

Queries Logical formulae

Accuracy Jaccard coefficient

$$\mathrm{J}(q_{\mathbf{L}}, q_{\mathbf{R}}) = \frac{|\mathrm{supp}(q_{\mathbf{L}}) \cap \mathrm{supp}(q_{\mathbf{R}})|}{|\mathrm{supp}(q_{\mathbf{L}}) \cup \mathrm{supp}(q_{\mathbf{R}})|}$$

Definitions



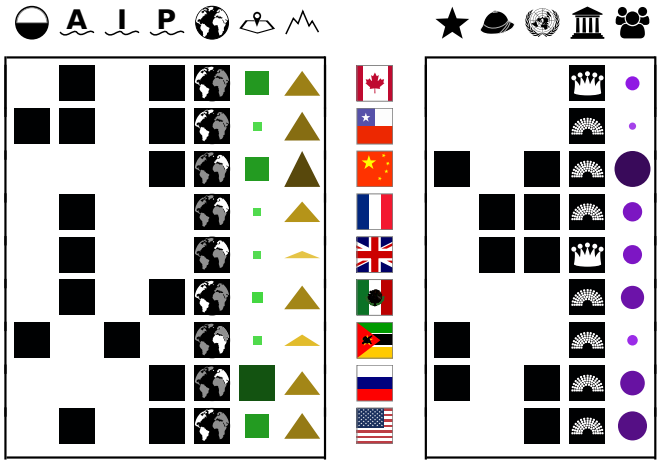
Dataset Two data matrices

Queries Logical formulae

Accuracy Jaccard coefficient

Constraints Support, accuracy, length of the query, p-value, …

Special Cases

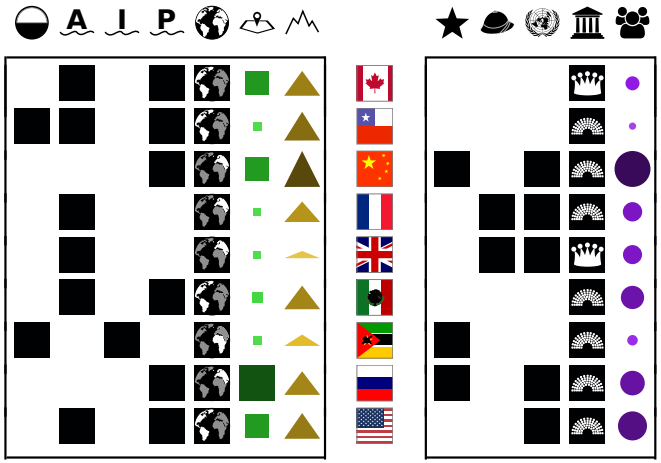


? AND ? $\Rightarrow$? AND ? AND ?

? AND ? $\Leftarrow$? AND ? AND ?

Only conjunctive queries:
bi-directional
association rules

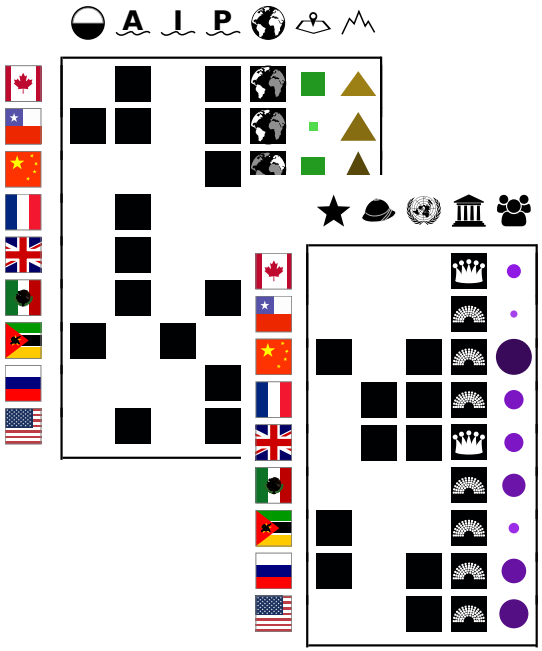
Special Cases



? ? ? ⇒ (AND) OR

One query given: classification task

Special Cases



→  →

(🌍 ≠ 🗺 AND ■ ≤ 🧭 ,
★ AND 🇺🇳)
{ 🇨🇳 , 🇷🇺 }

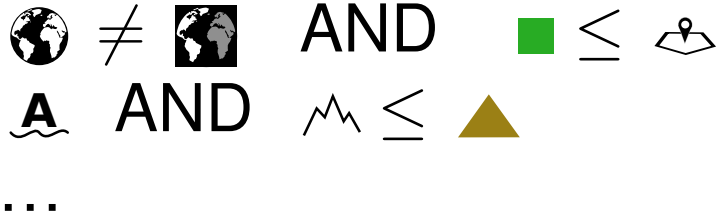
(🅐 AND ⛰ ≤ ▲ ,
🏛 = 👑 OR ● ≤ 👥 ≤ ●)
{ 🇨🇦 , 🇫🇷 , 🇬🇧 , 🇲🇽 , 🇷🇺 }

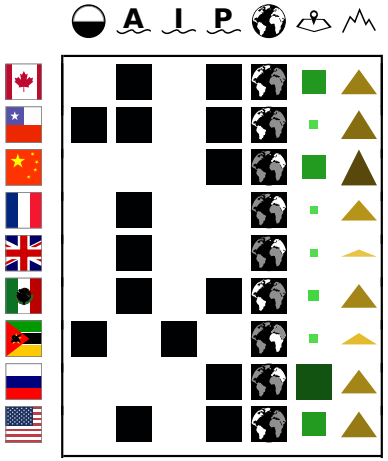
...

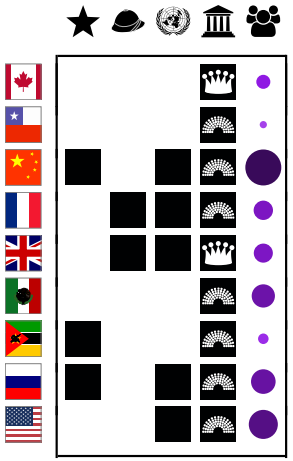
Exploration Strategies

How do we find redescriptions?

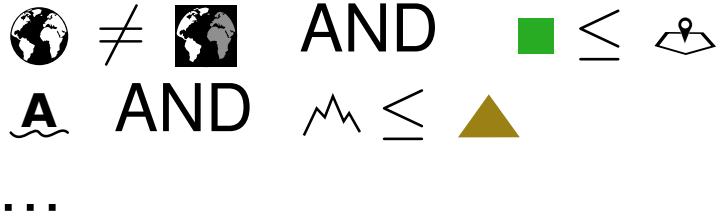
Exploration Strategies

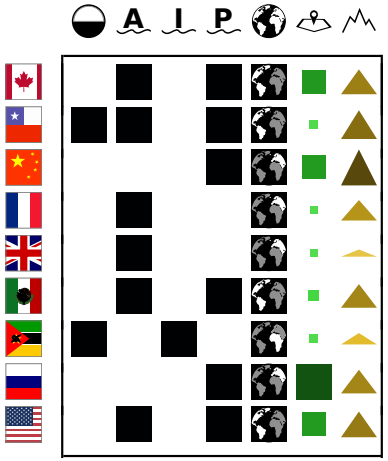


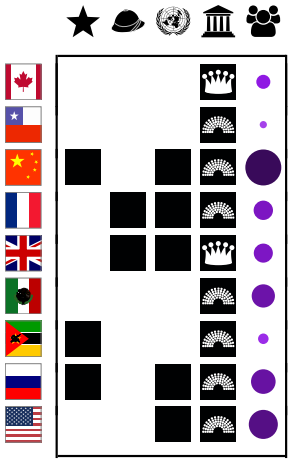




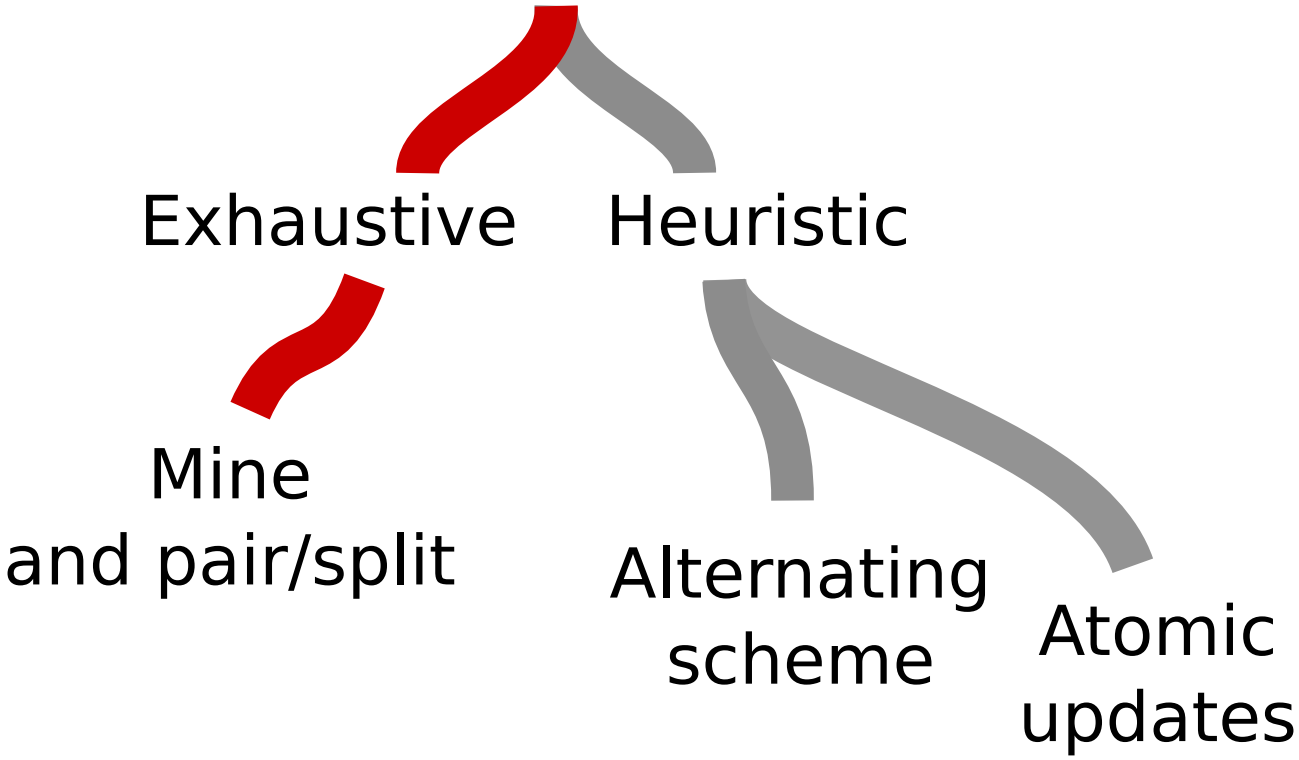
Exploration Strategies



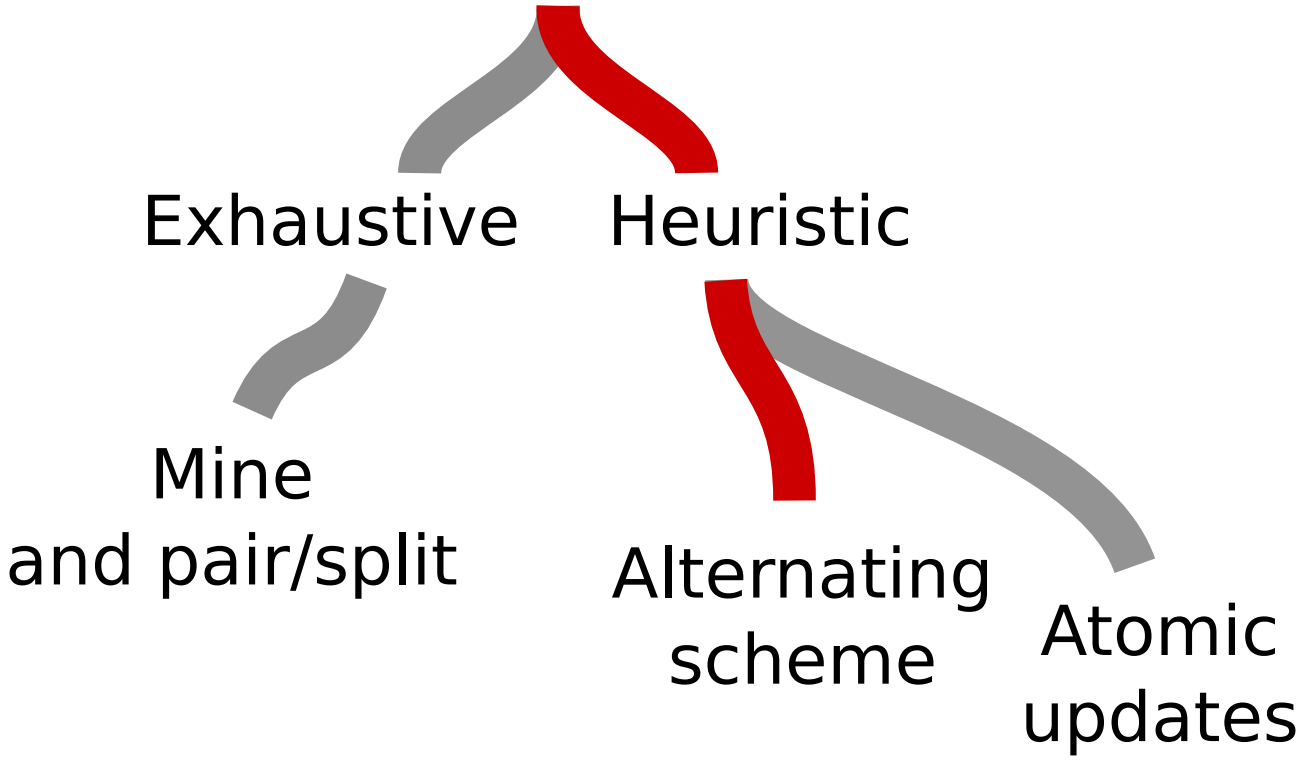




Exploration Strategies

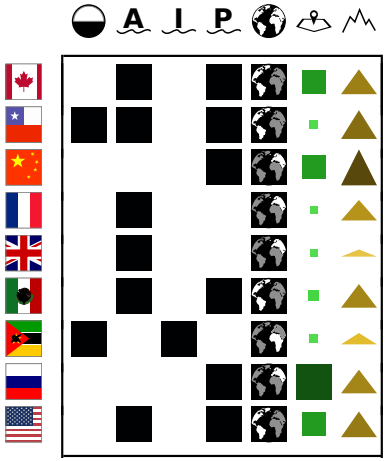


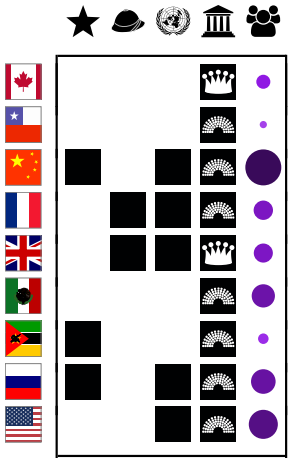
Exploration Strategies



Exploration Strategies

(¬ **A** AND ¬ **I**) OR (**A** AND **P**)

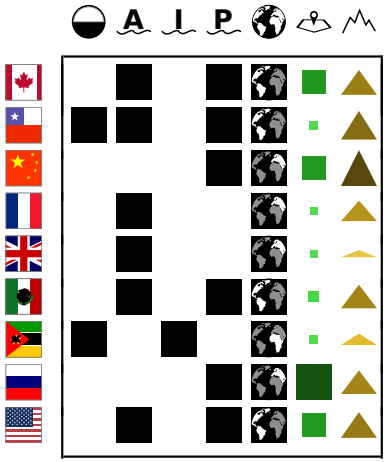


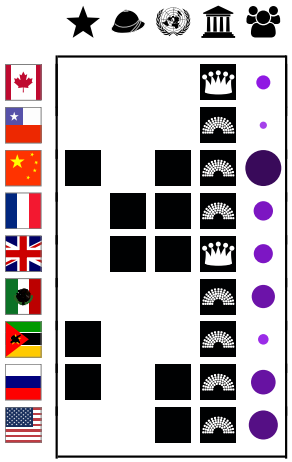


Exploration Strategies

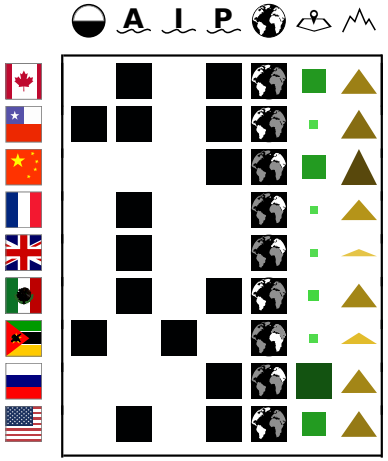
(¬ **A** AND ¬ **I**) OR (**A** AND **P**)

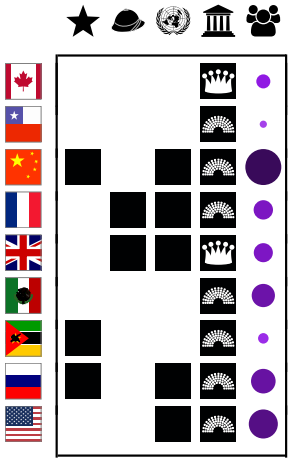
★ OR ¬



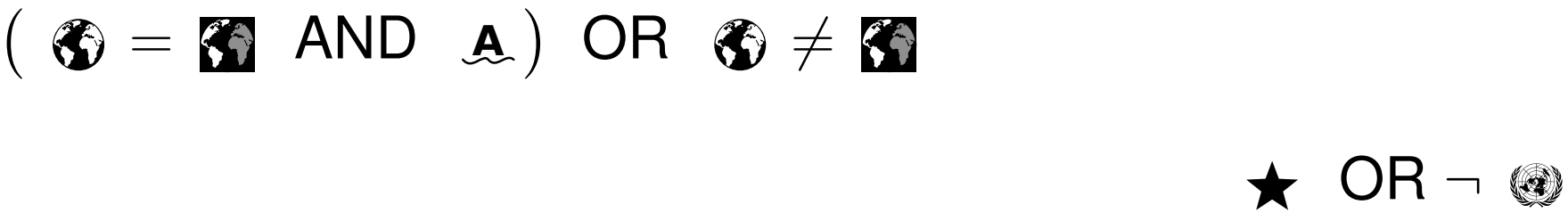


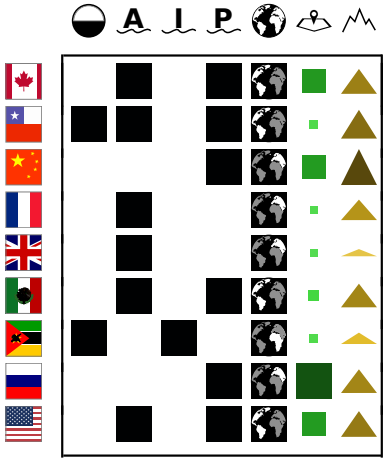
Exploration Strategies

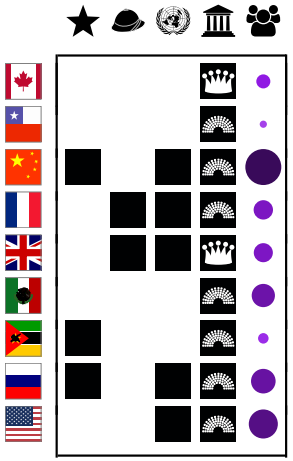




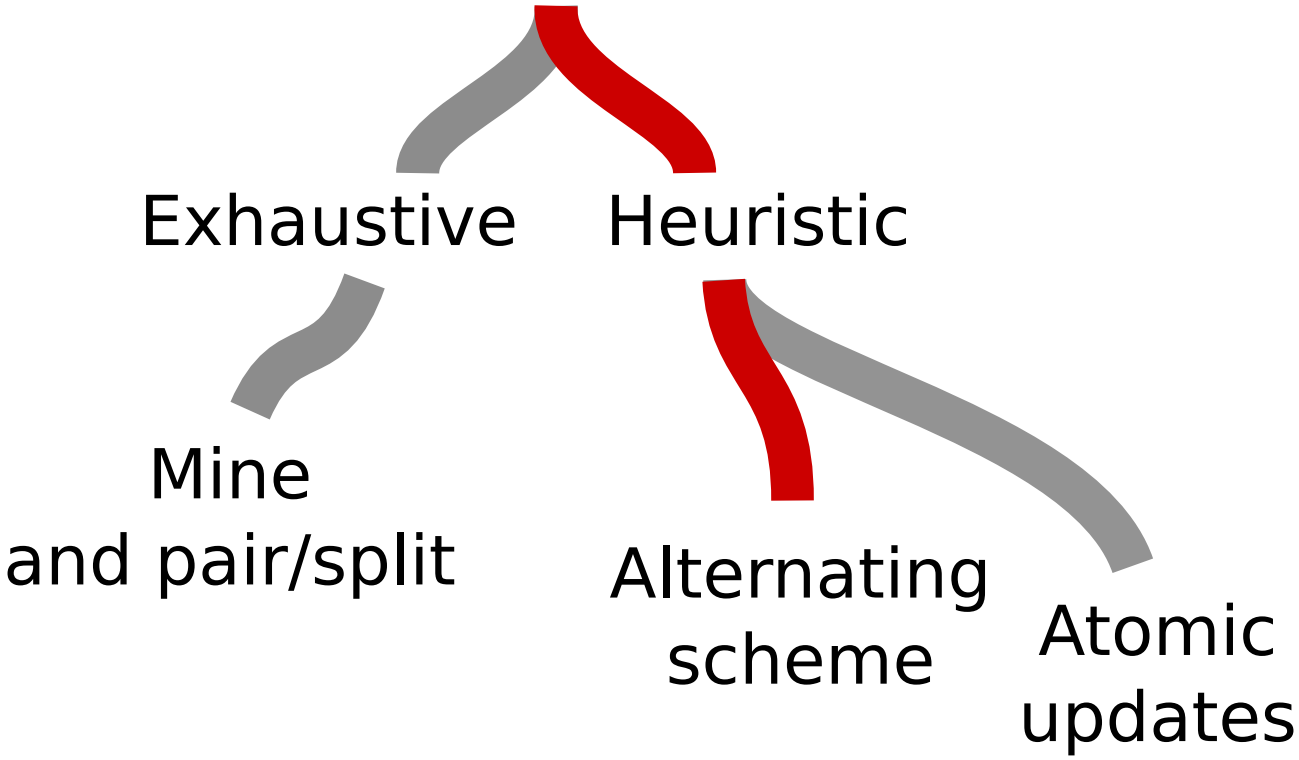
Exploration Strategies



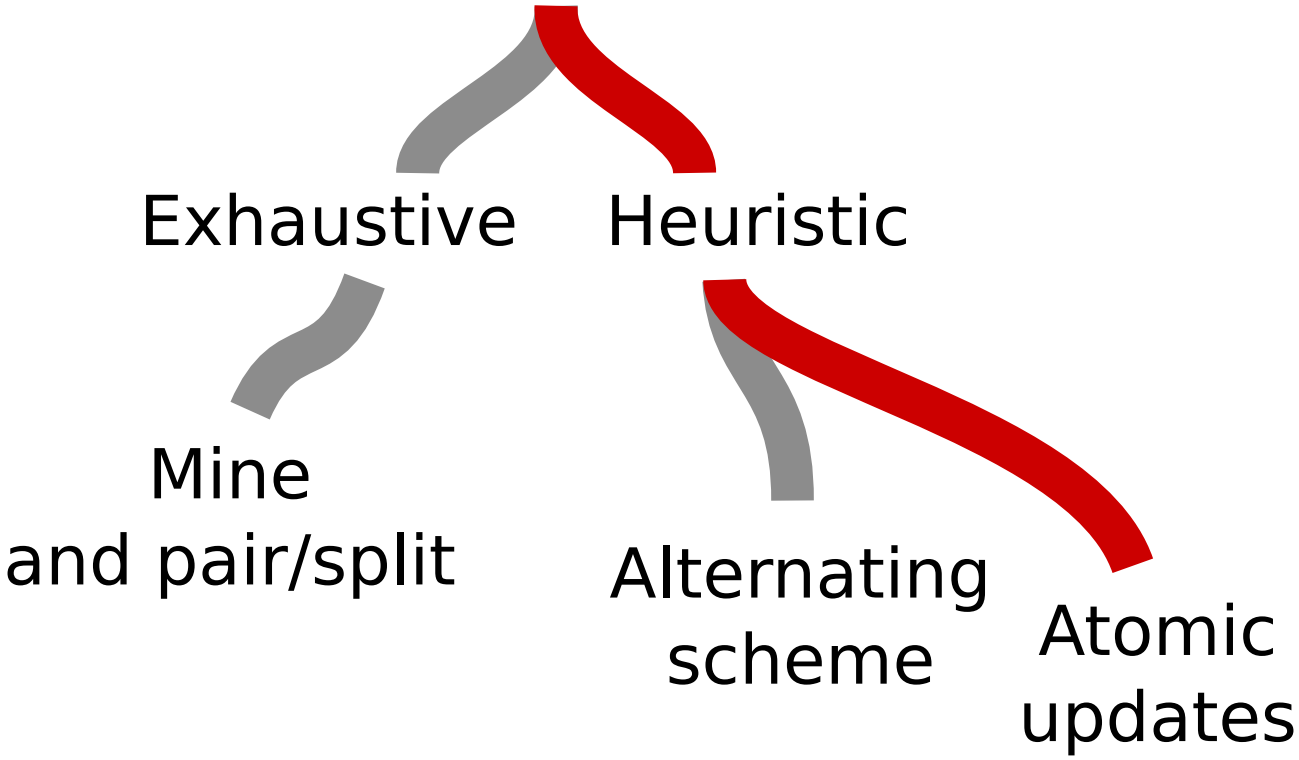




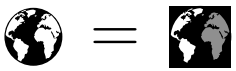
Exploration Strategies

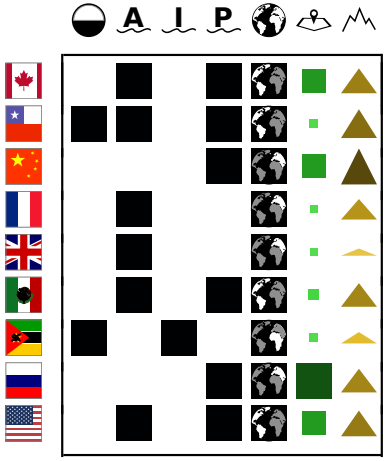


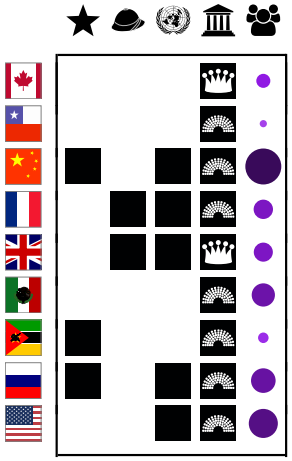
Exploration Strategies



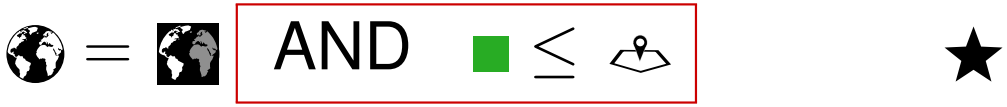
Exploration Strategies

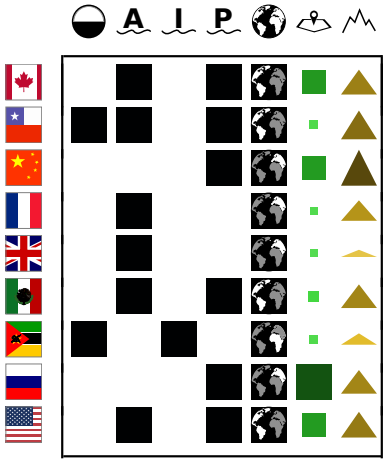


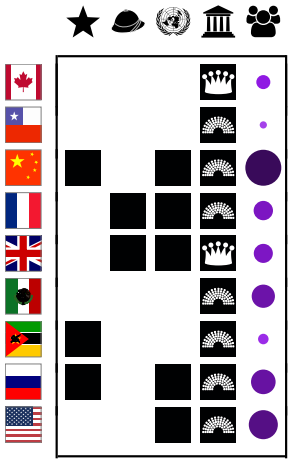




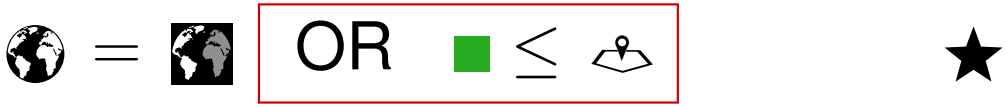
Exploration Strategies

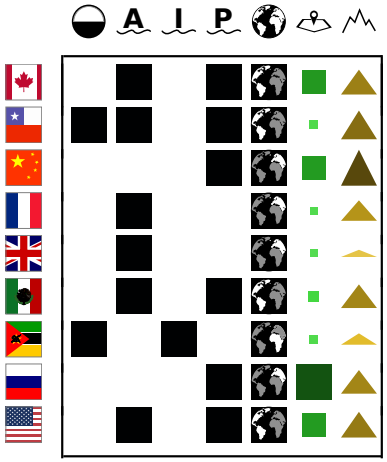


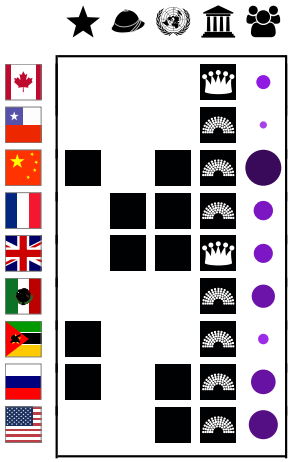




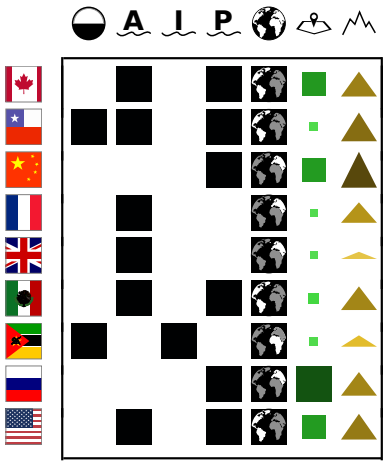
Exploration Strategies

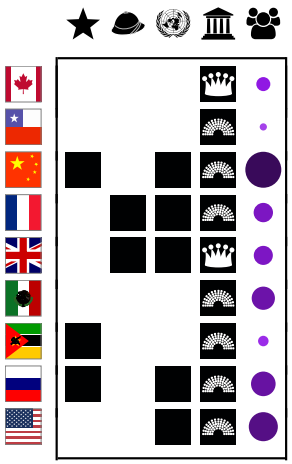




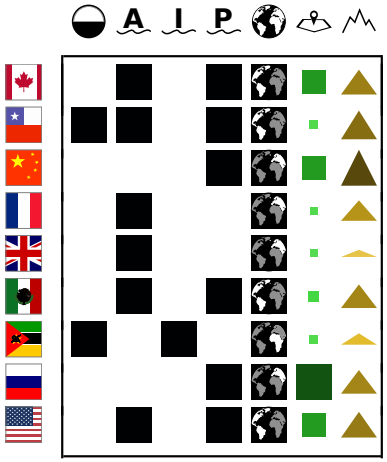


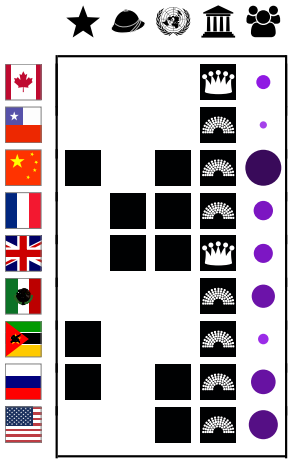
Exploration Strategies



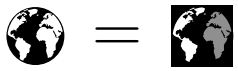


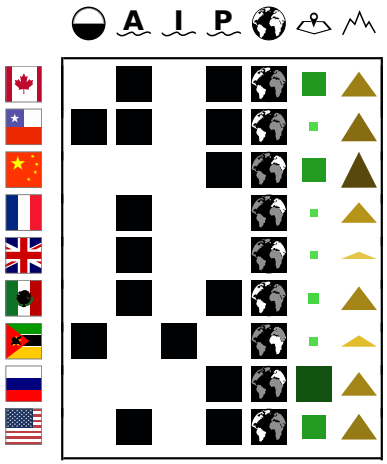
Exploration Strategies

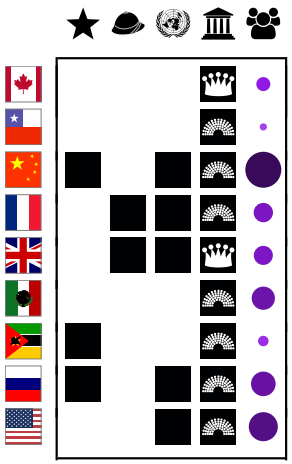




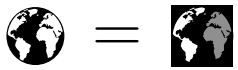
Exploration Strategies

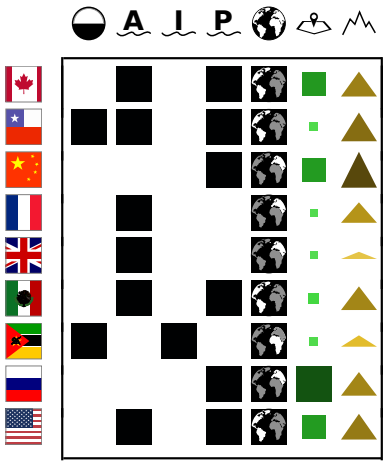


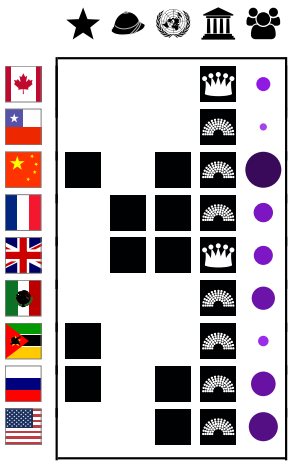




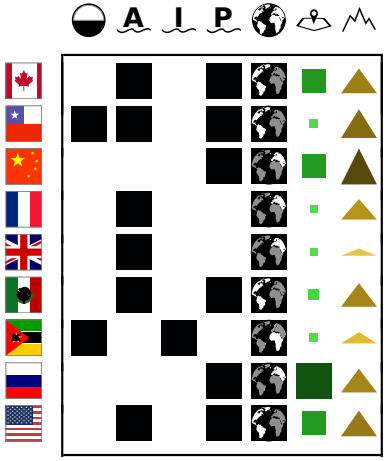
Exploration Strategies

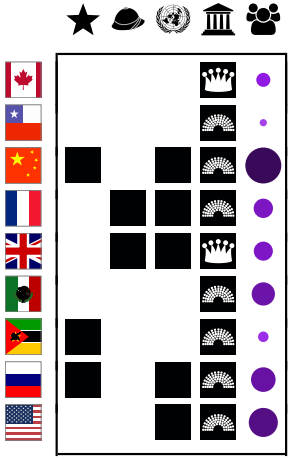




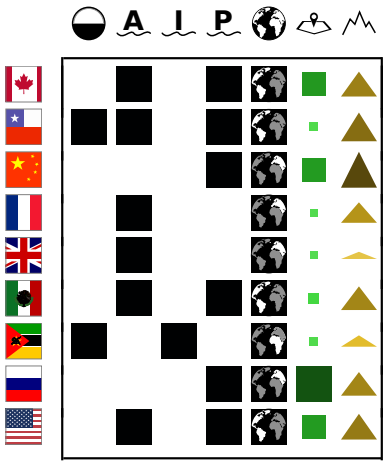


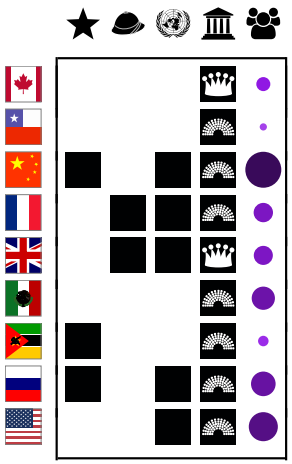
Exploration Strategies



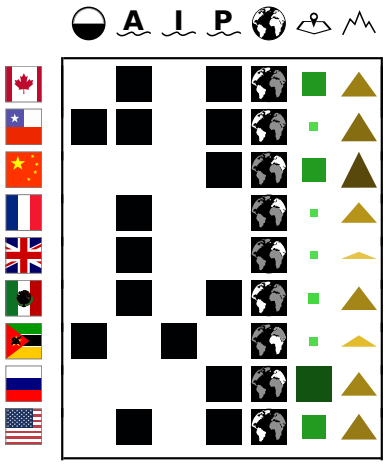


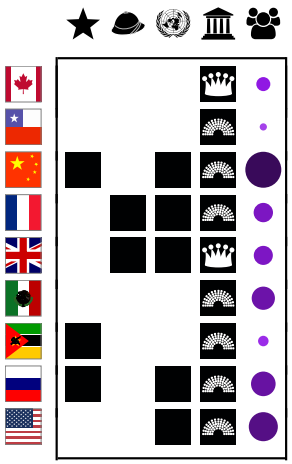
Exploration Strategies



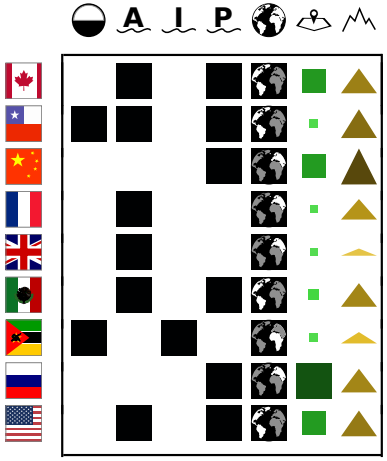


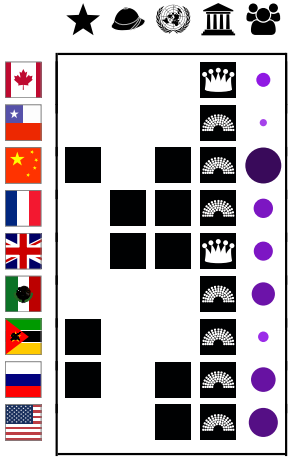
Exploration Strategies



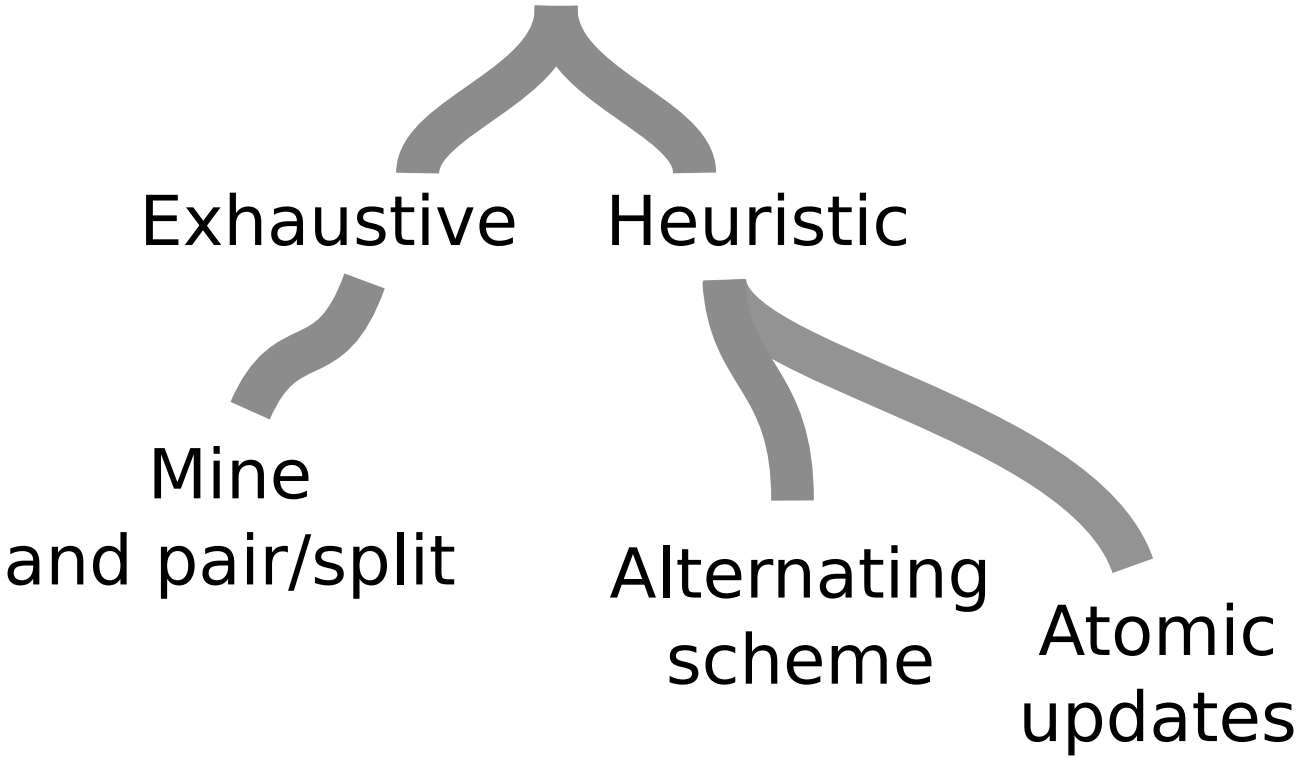


Exploration Strategies





Exploration Strategies

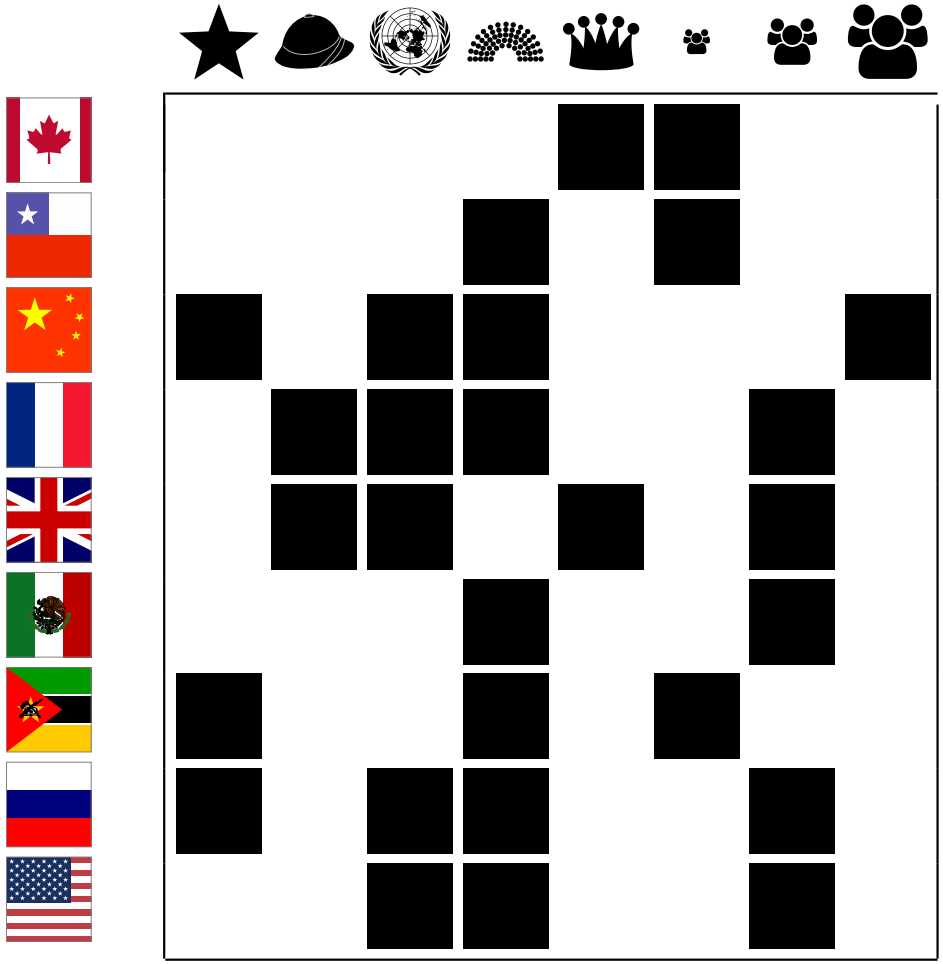


Related work

Redescription mining for Boolean data

Related work

Geopolitical Boolean attributes



Related work

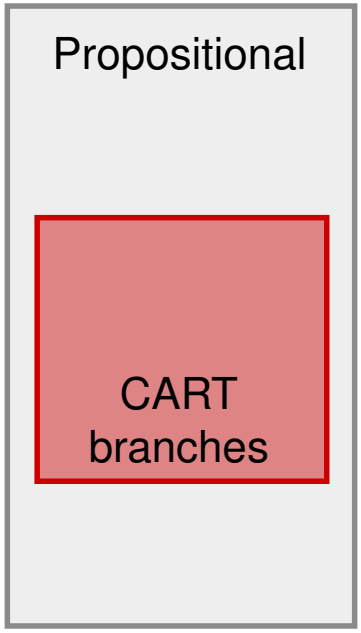
Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions.

N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm.
In *KDD*, 2004.

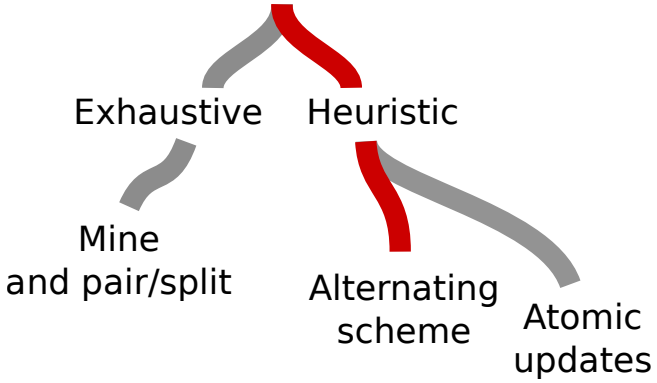
Redescription Mining: Algorithms and Applications in Bioinformatics.

D. Kumar.
PhD Thesis, Virginia Tech, 2007.

Query Languages



Exploration Strategies

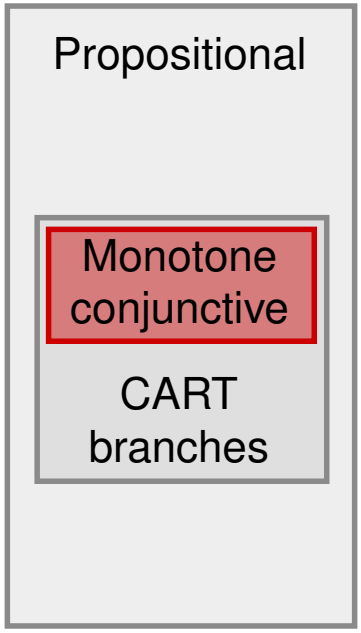


Related work

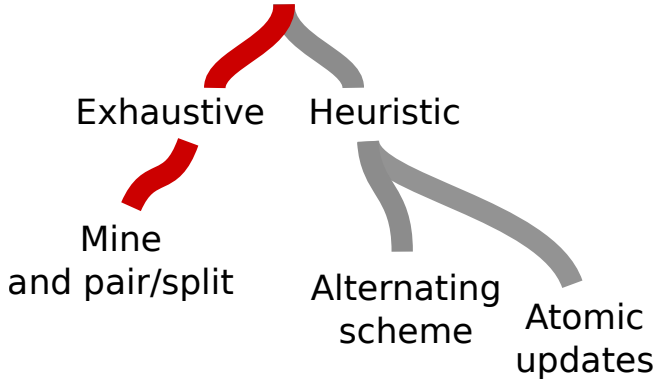
Redescription Mining: Structure Theory and Algorithms.

L. Parida and N. Ramakrishnan.
In *AAAI*, 2005.

Query Languages



Exploration Strategies

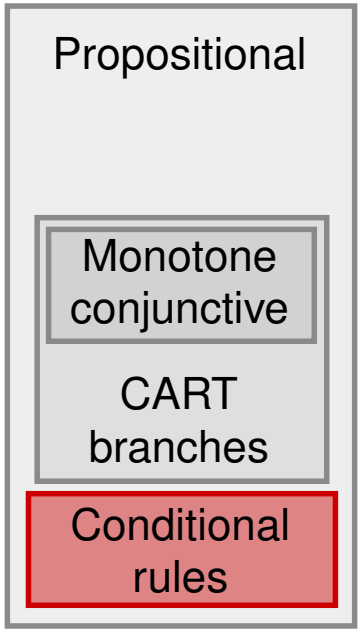


Related work

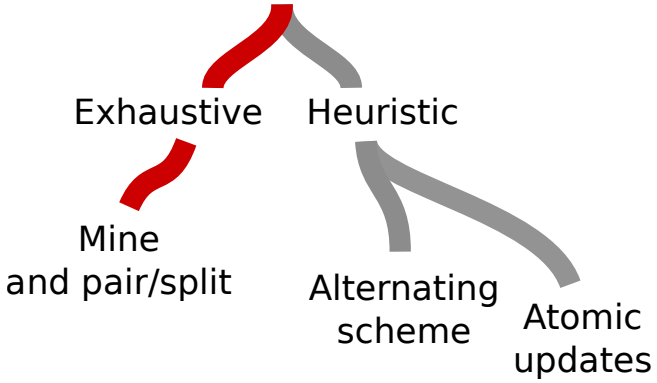
Reasoning About Sets Using Redescription Mining.

M. J. Zaki and N. Ramakrishnan.
In *KDD*, 2005.

Query Languages



Exploration Strategies

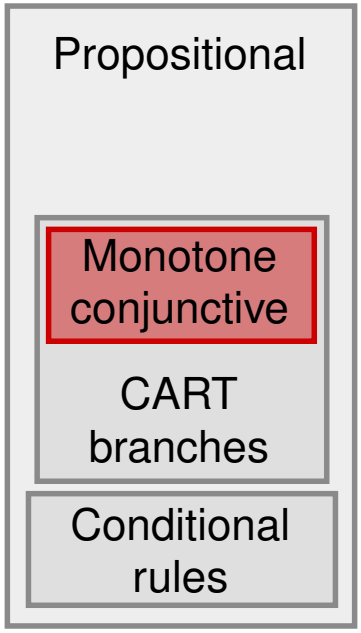


Related work

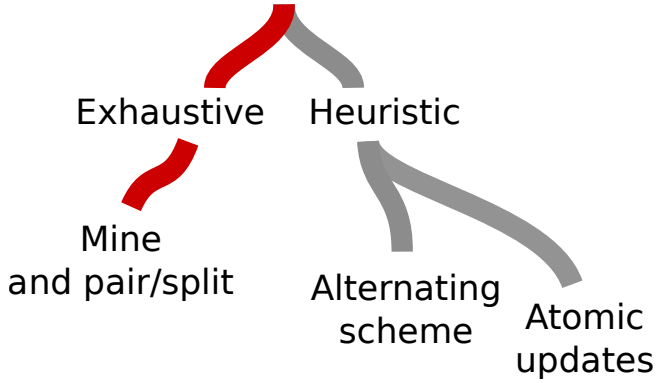
Finding Subgroups Having Several Descriptions: Algorithms for Redescription Mining.

A. Gallo, P. Miettinen, and H. Mannila.
In *SDM*, 2008.

Query Languages



Exploration Strategies

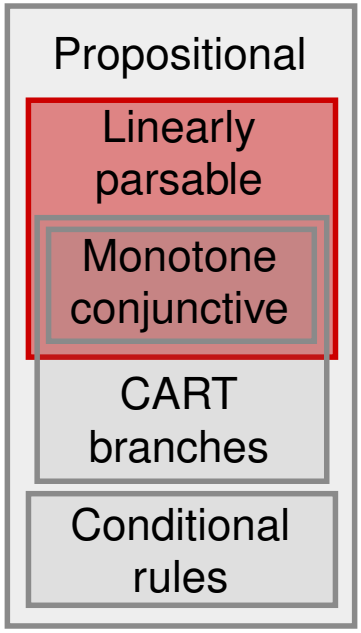


Related work

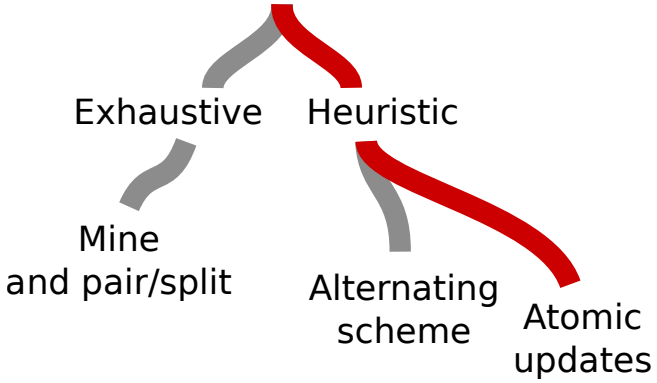
Finding Subgroups Having Several Descriptions: Algorithms for Redescription Mining.

A. Gallo, P. Miettinen, and H. Mannila.
In *SDM*, 2008.

Query Languages



Exploration Strategies



Own work

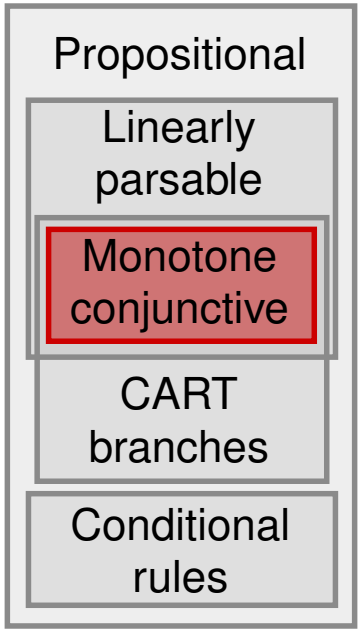
Selecting a good set of redescriptions

Own work

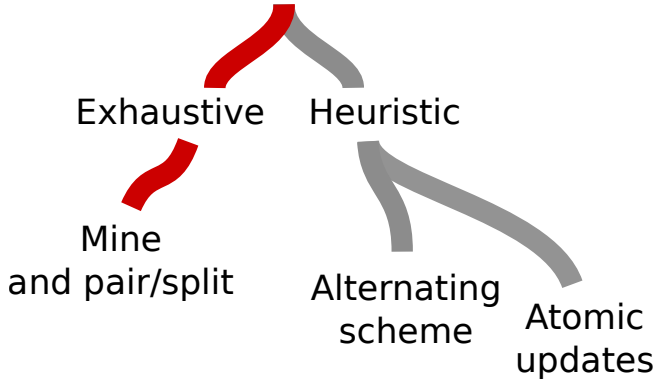
"MDL for Redescription Mining"

with Matthijs van Leeuwen,
Under review.

Query Languages



Exploration Strategies

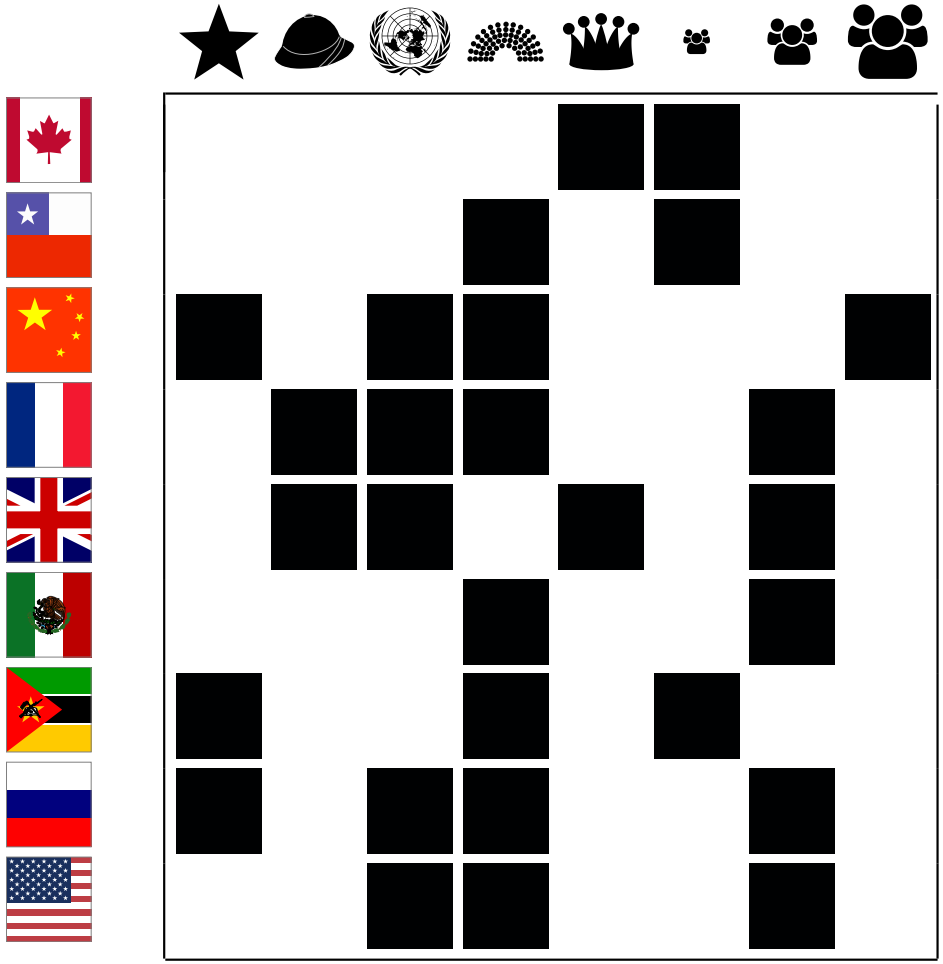


Own work

Extending redescription mining to non-Boolean data

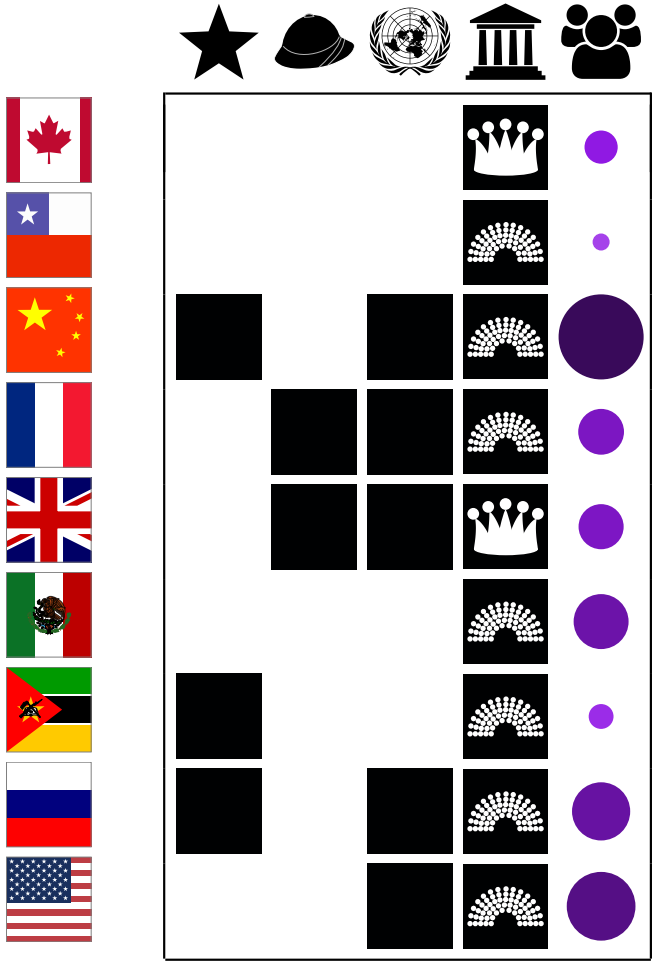
Own work

Geopolitical Boolean attributes



Own work

Geopolitical attributes

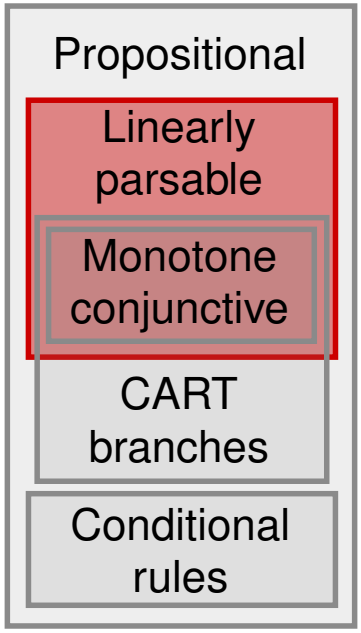


Related work

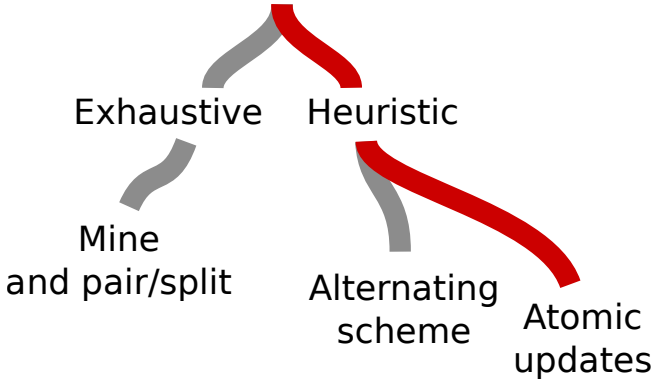
Finding Subgroups Having Several Descriptions: Algorithms for Redescription Mining.

A. Gallo, P. Miettinen, and H. Mannila.
In *SDM*, 2008.

Query Languages



Exploration Strategies

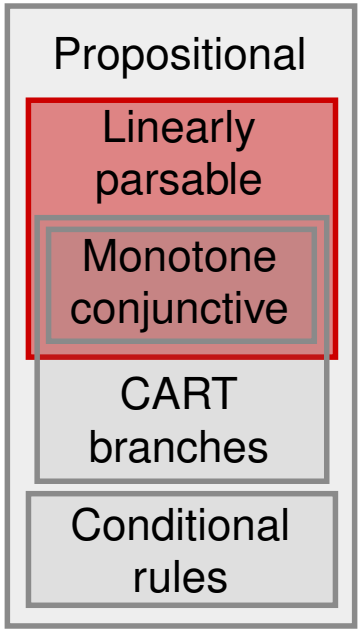


Own work

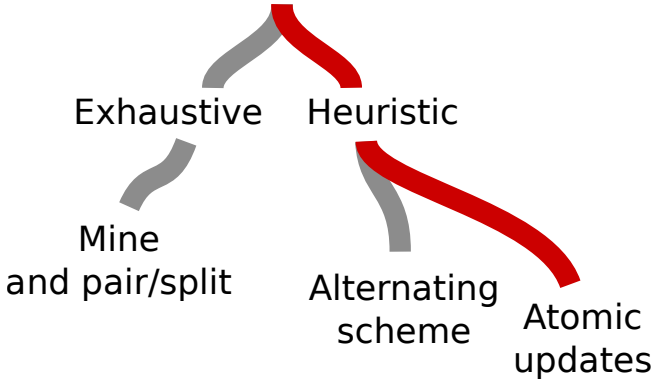
From Black and White to Full Color: Extending Redescription Mining Outside the Boolean World

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

Query Languages



Exploration Strategies

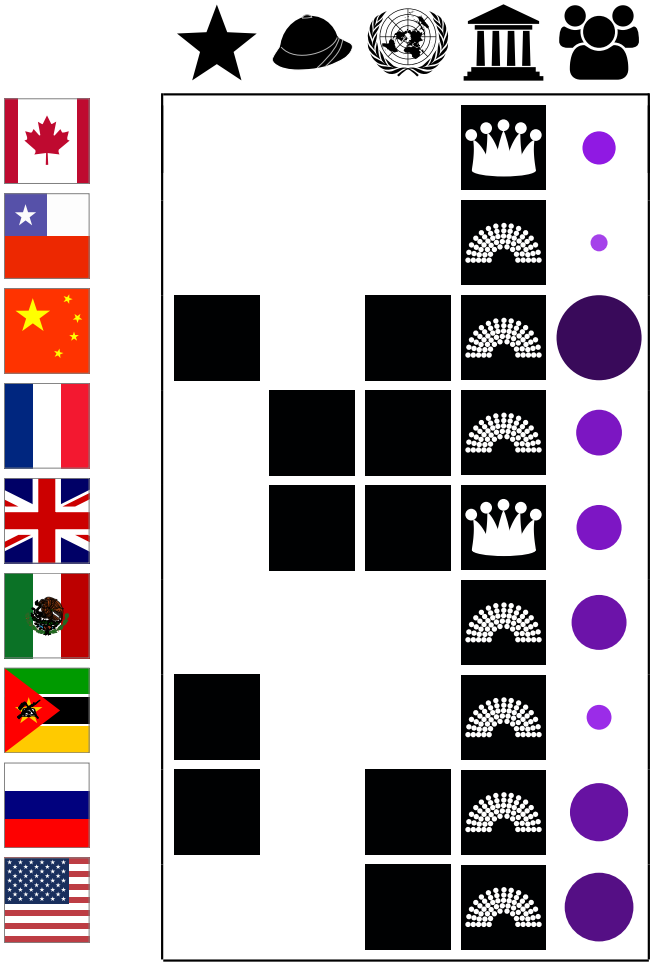


Own work

Extending redescription mining to relational data

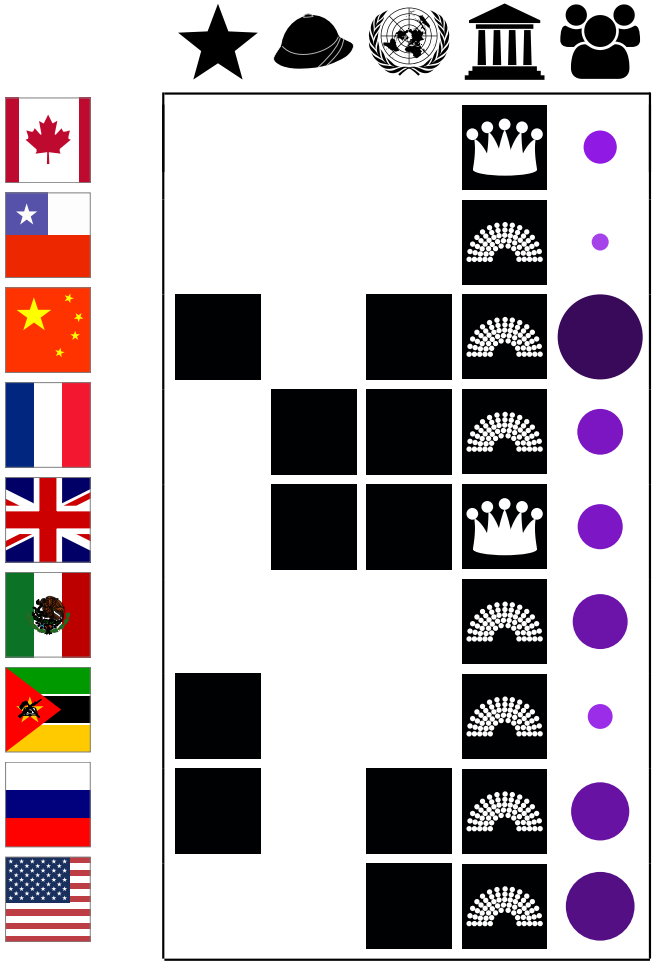
Own work

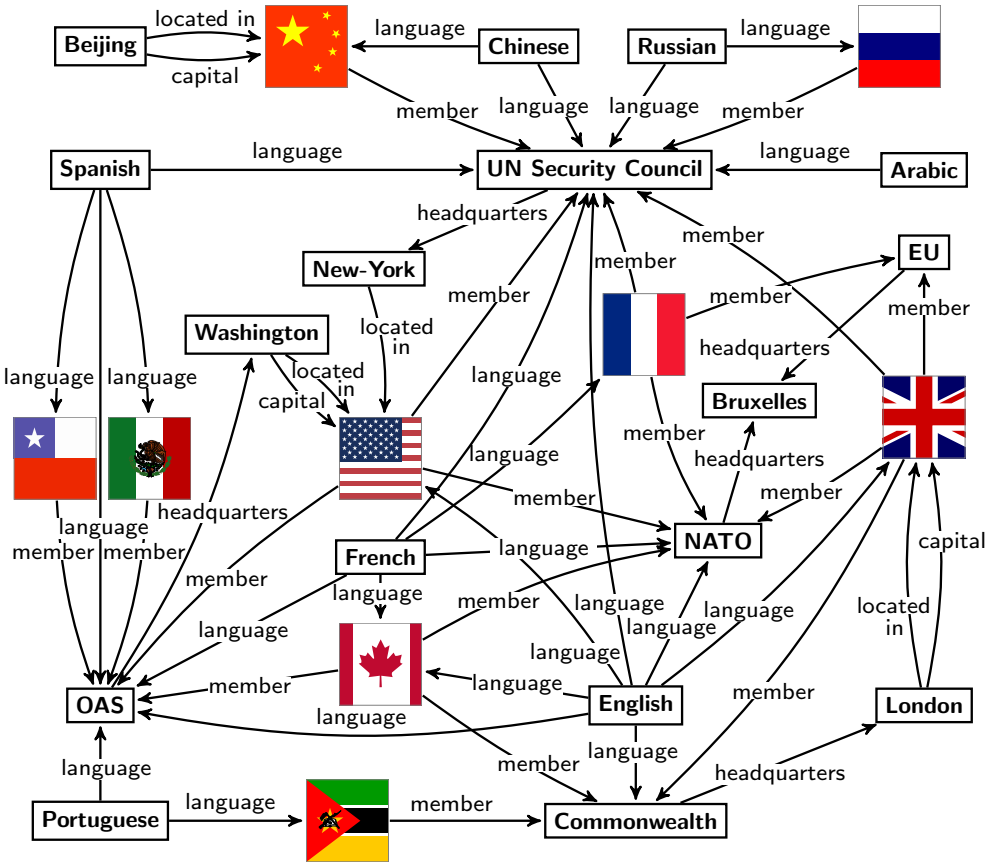
Geopolitical attributes



Own work

Geopolitical attributes and relations



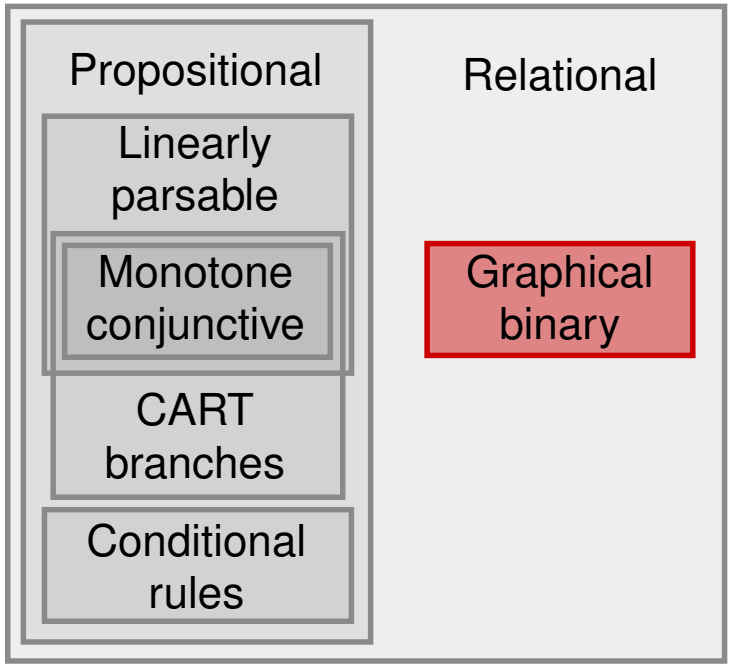


Own work

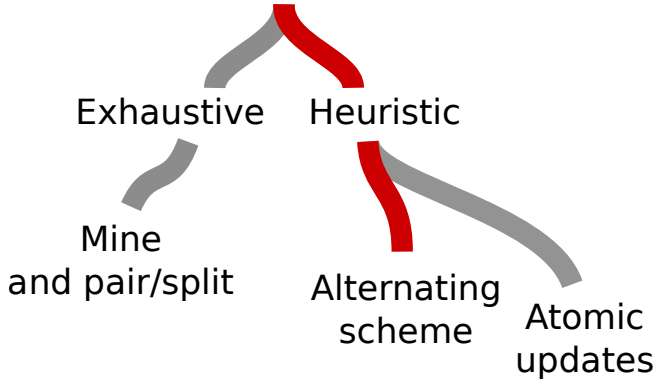
Finding Relational Redescriptions

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

Query Languages



Exploration Strategies



Example Redescriptions

- Computer science bibliography

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

Example Redescriptions

- Computer science bibliography
- Political candidates profiles

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

Example Redescriptions

- Computer science bibliography
- Political candidates profiles
- Bioclimatic niches

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

Bioclimatic Niche Finding

Dataset: Spatial land areas of Europe (2575 entities)

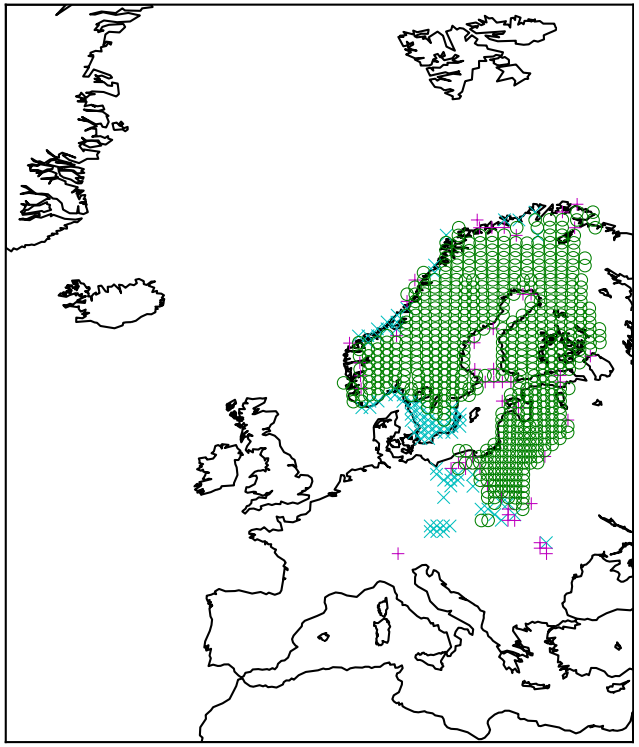
- Presence/absence of mammals (194 species)
- Climatic data (48 temperature and rainfall variables)

Question: Find a query over climatic variables that describes the area inhabited by (a group of) mammal species (and vice versa)

Bioclimatic Niche Finding

European Elk

$$([-9.80 \leq t^{\max}_{\text{Feb}} \leq 0.40] \wedge [12.20 \leq t^{\max}_{\text{Jul}} \leq 24.60] \wedge$$
$$[56.852 \leq p^{\text{avg}}_{\text{Aug}} \leq 136.46]) \vee [183.27 \leq p^{\text{avg}}_{\text{Sep}} \leq 238.78]$$

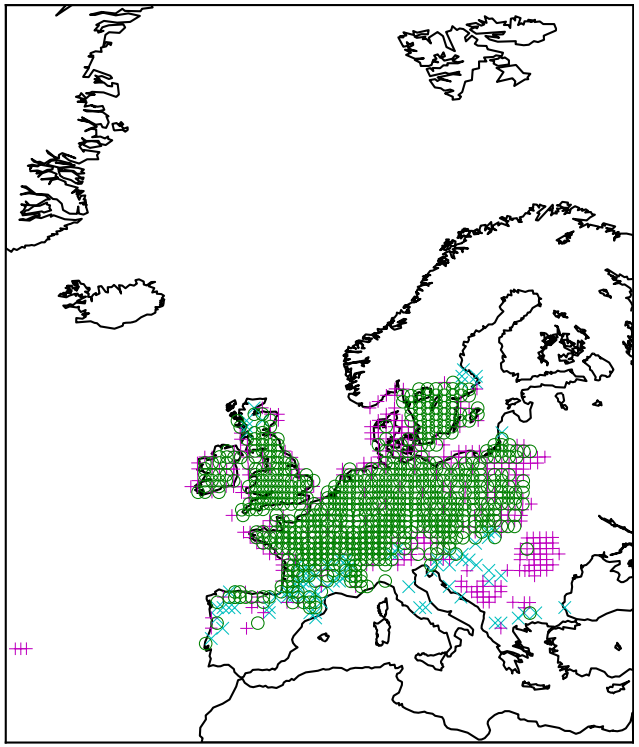


$\text{J} = 0.814 \quad \text{supp} = 582$

Bioclimatic Niche Finding

Wood Mouse $\wedge$ Natterer's Bat $\wedge$ Eurasian Pygmy Shrew

$$([3.20 \leq t^{\max}_{\text{Mar}} \leq 14.50] \wedge [17.30 \leq t^{\max}_{\text{Aug}} \leq 25.20] \wedge$$
$$[14.90 \leq t^{\max}_{\text{Sep}} \leq 22.80]) \vee [19.60 \leq t^{\text{avg}}_{\text{Jul}} \leq 19.956]$$



$\text{J} = 0.623 \quad \text{supp} = 681$

Example Redescriptions

- Computer science bibliography
- Political candidates profiles
- Bioclimatic niches

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

Example Redescriptions

- Computer science bibliography
- Political candidates profiles
- Bioclimatic niches
- Ethnology

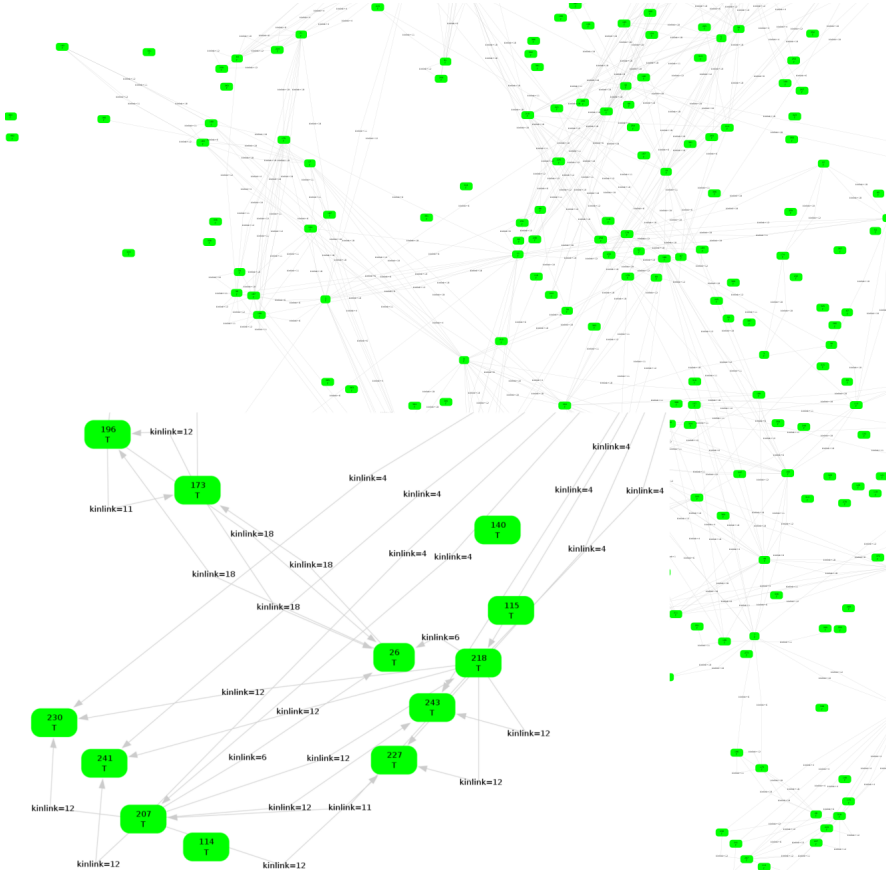
Among Alyawarra, “Aleriya” refers to the son of a male speaker or to the child of the speaker’s brother.

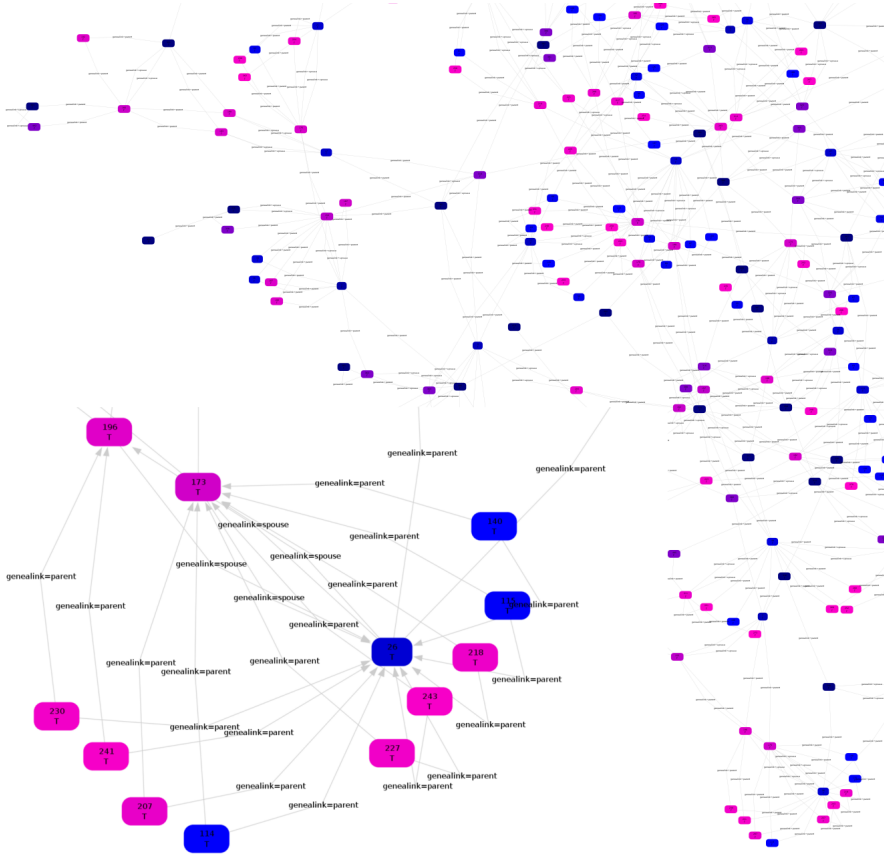
Elicit Kinship Terminology

Dataset: Ethnographic information about Australian *Alyawarra* tribe

- Kinship terminology
- Genealogic, age and sex informations

Elicit Kinship Terminology





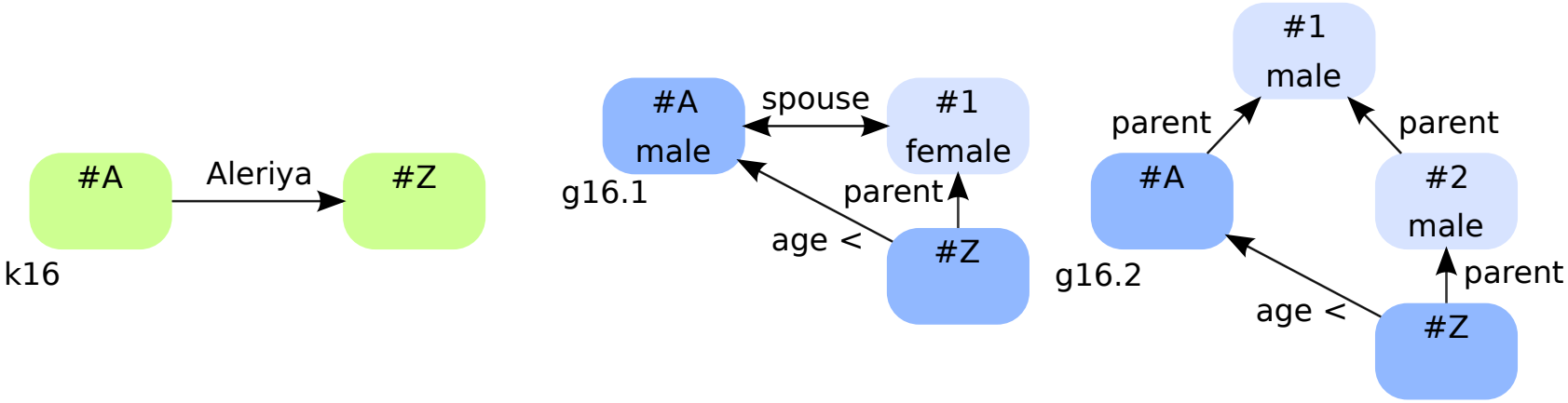
Elicit Kinship Terminology

Dataset: Ethnographic information about Australian *Alyawarra* tribe

- Kinship terminology
- Genealogic, age and sex informations

Question: Elicit the meaning of kinship terms

Elicit Kinship Terminology



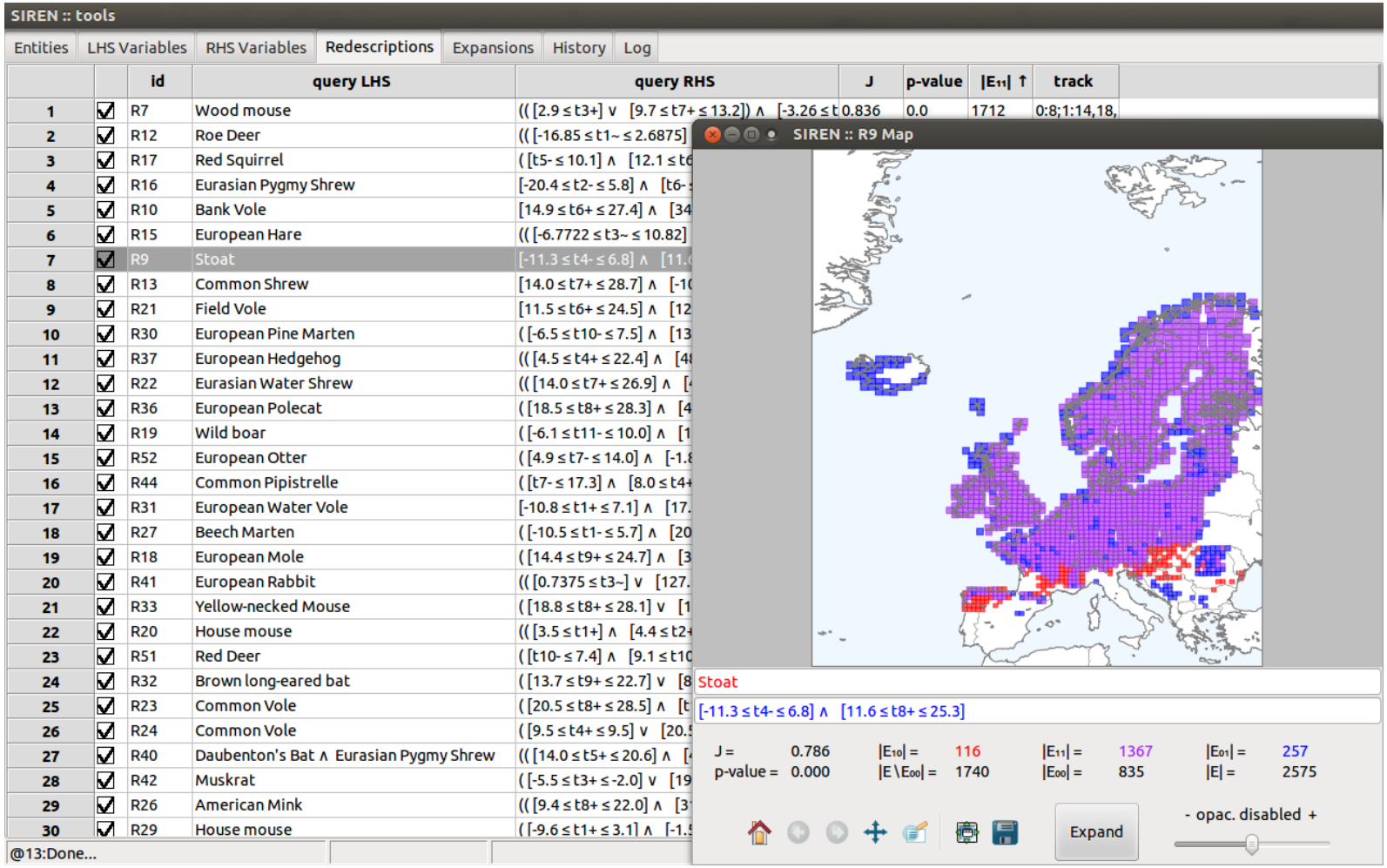
Aleriya *is used to refer to one's father or one's brother's child*

The SIREN interface

Visualizing and interactively mining redescriptions

with Pauli Miettinen,
In *Instant Interactive Data Mining Workshop at ECML/PKDD*, 2012.
Demo at KDD 2012 and SIGMOD 2014.

The SIREN interface



The SIREN interface

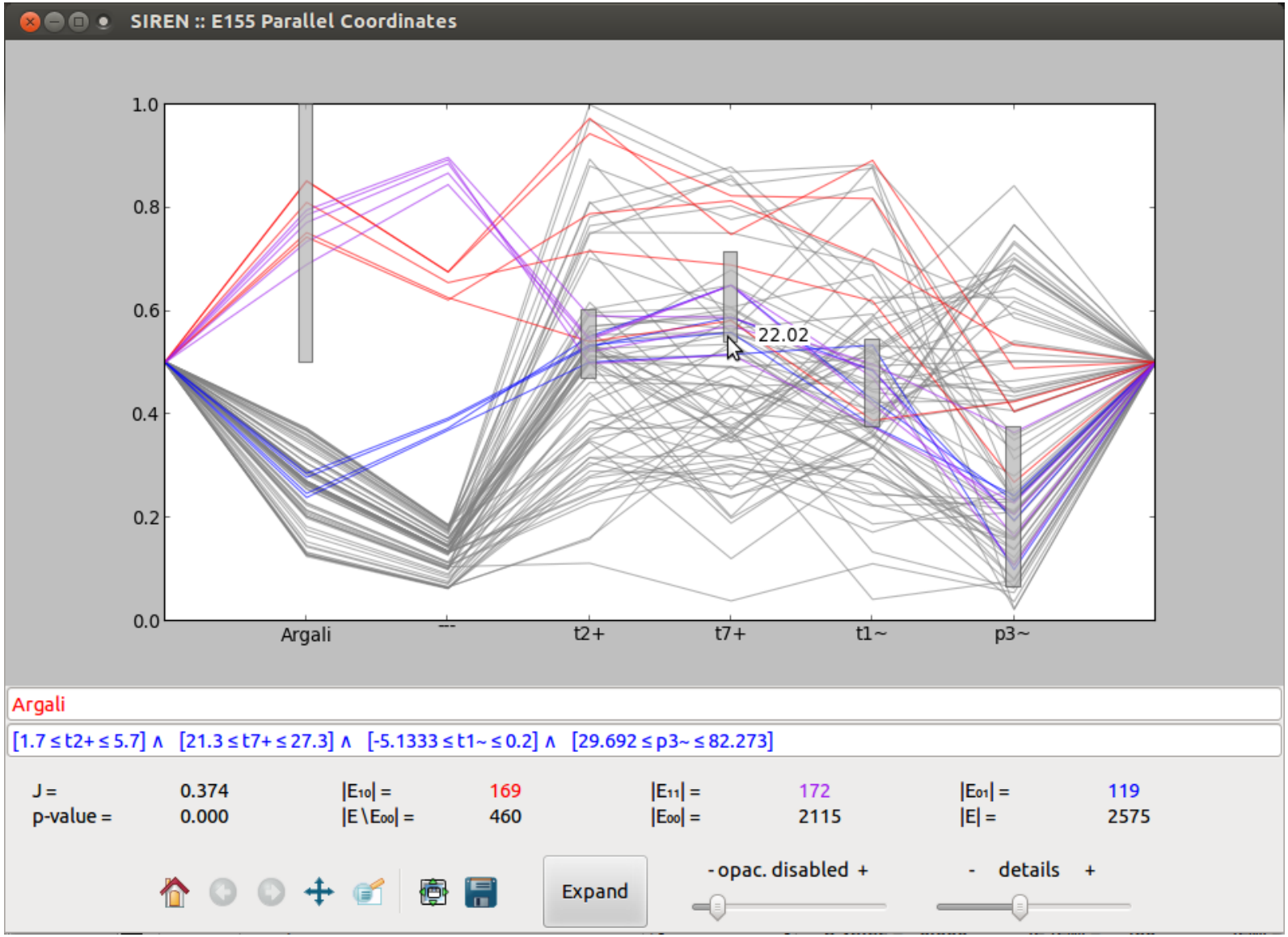
SIREN :: tools

Entities | LHS Variables | RHS Variables | Redescriptions

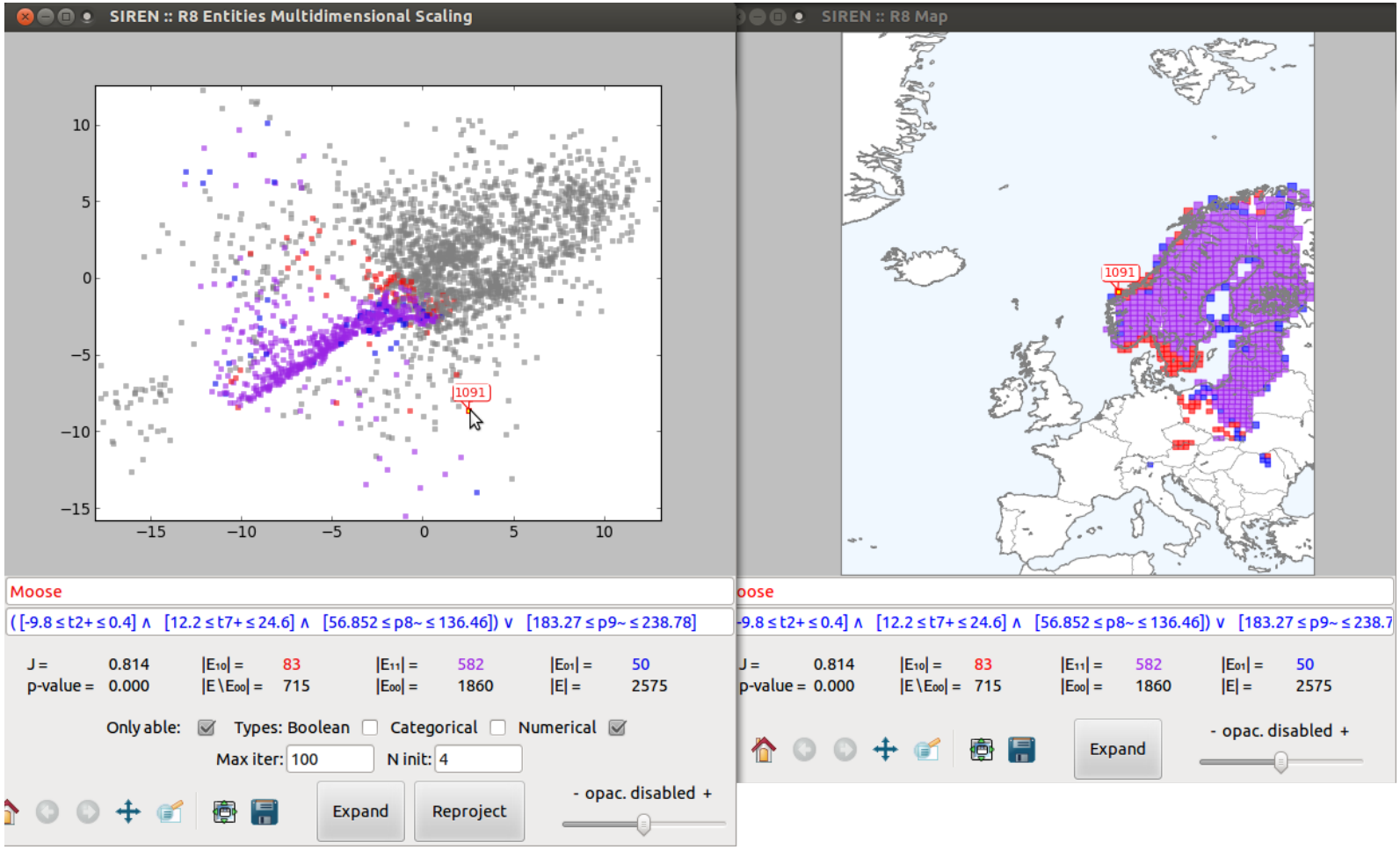
		id	LHS:Wisent ↑	LHS:Wild goat	LHS:European Hedgehog
1	☑	1464	True	False	False
2	☑	1494	True	False	False
3	☑	1891	True	False	False
4	☑	2000	True	False	False
5	☑	2005	True	False	False
6	☑	2016	True	False	False
7	☑	2032	True	False	False
8	☐	2033	True	False	False
9	☐	2037	True	False	False
10	☐	2049	True	False	False
11	☐	2050	True	False	False
12	☐	2264	True	False	False
13	☐	2270	True	False	False
14	☑	2311	True	False	False
15	☑	2357	True	False	False
16	☑	2361	True	False	False
17	☑	2362	True	False	False
18	☑	0	False	False	True
19	☑	1	False	False	True
20	☑	2	False	False	True
21	☑	3	False	False	True
22	☑	4	False	False	False
23	☑	5	False	False	True
24	☑	6	False	False	True
25	☑	7	False	False	True

Loading done

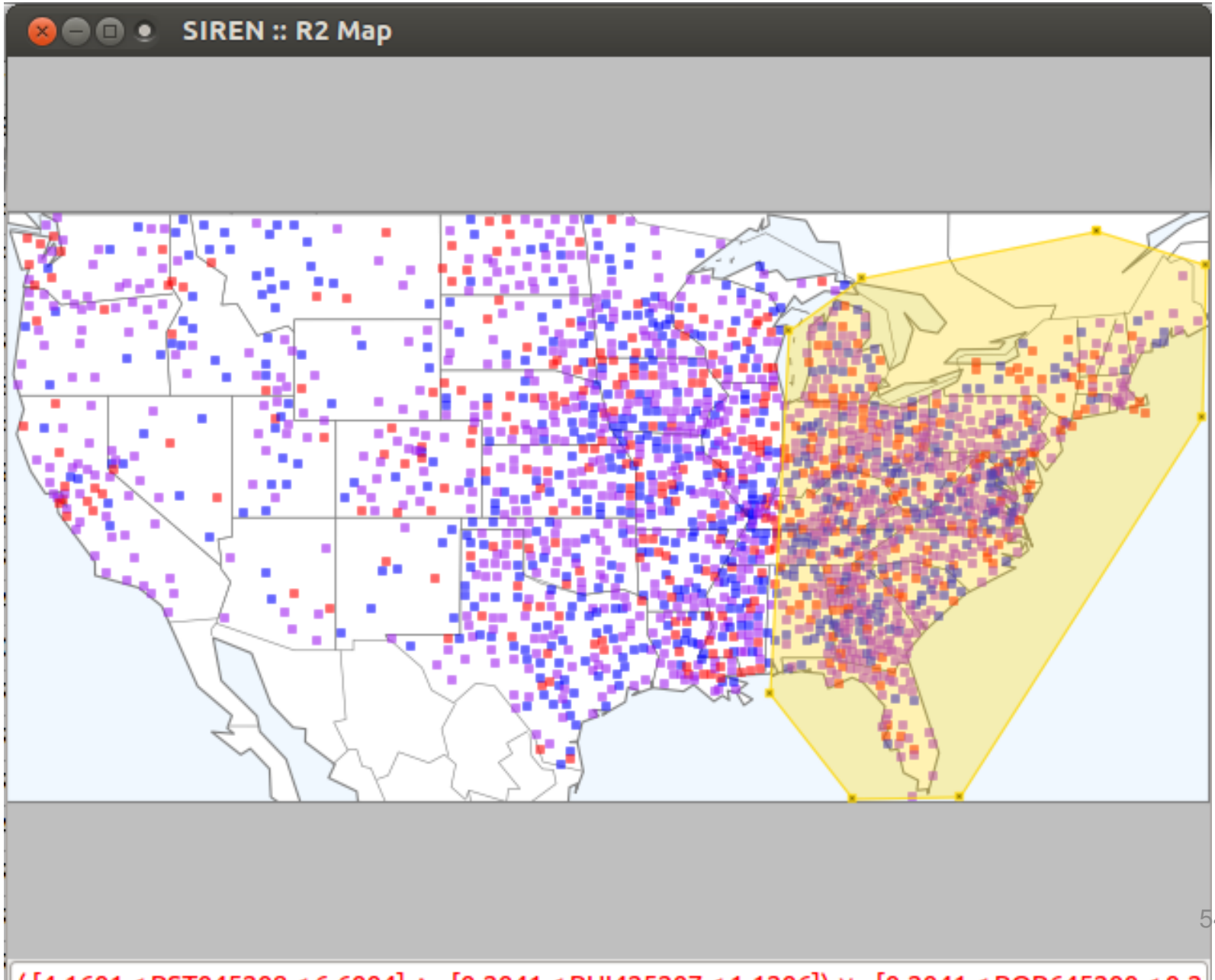
The SIREN interface



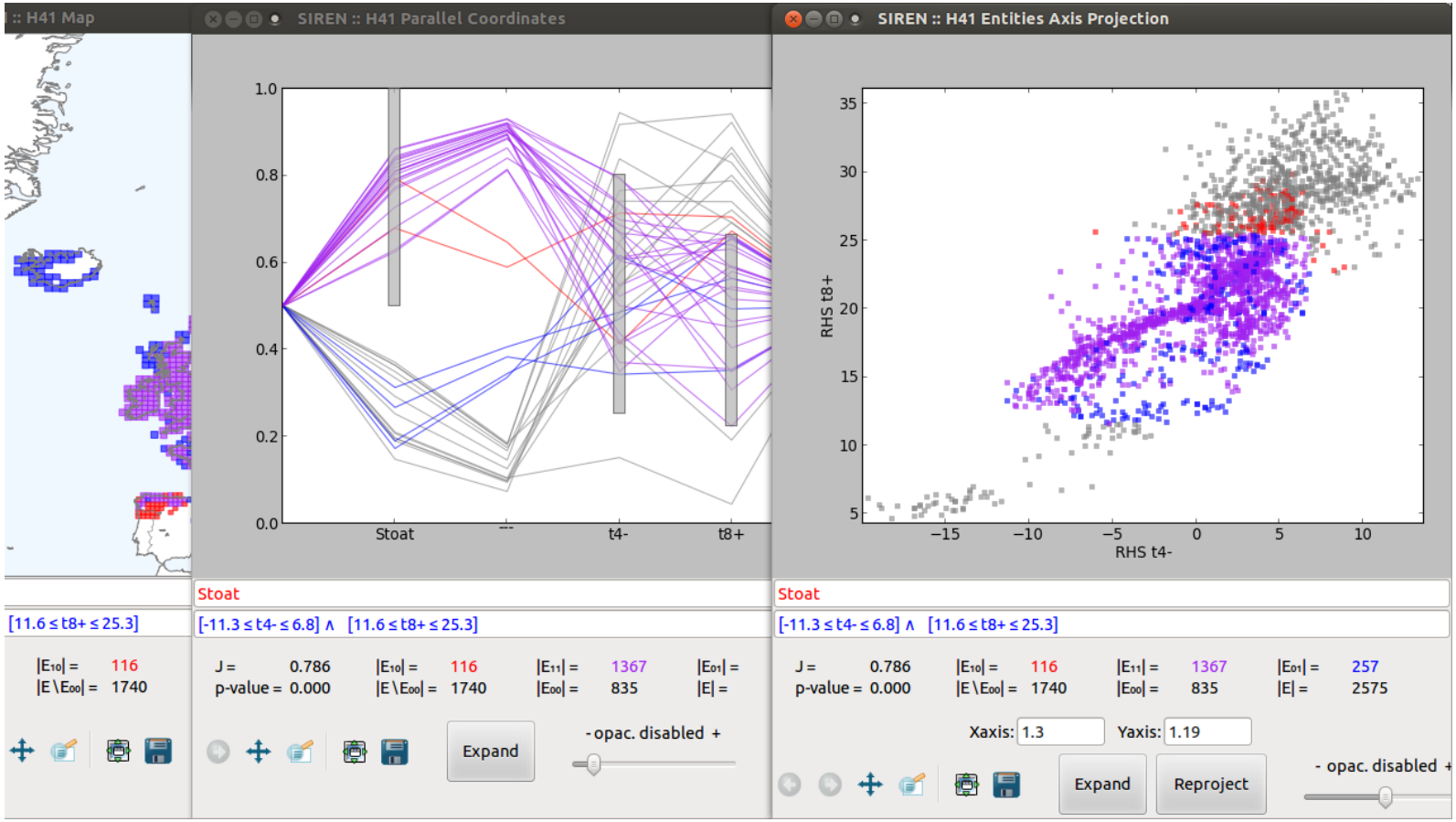
The SIREN interface



The SIREN interface



The SIREN interface



Conclusion

Redescription Mining is a versatile and
powerful data-mining tool, applicable in various domains.

For more details:

- galbrun@cs.helsinki.fi
- http://www.cs.helsinki.fi/u/galbrun/

Conclusion

Redescription Mining is a versatile and
powerful data-mining tool, applicable in various domains.

For more details:

- `galbrun@cs.helsinki.fi`
- `http://www.cs.helsinki.fi/u/galbrun/`

Thank you!