Reasoning about Sets using Redescription Mining

Mohammed J. Zaki
zaki@cs.rpi.edu

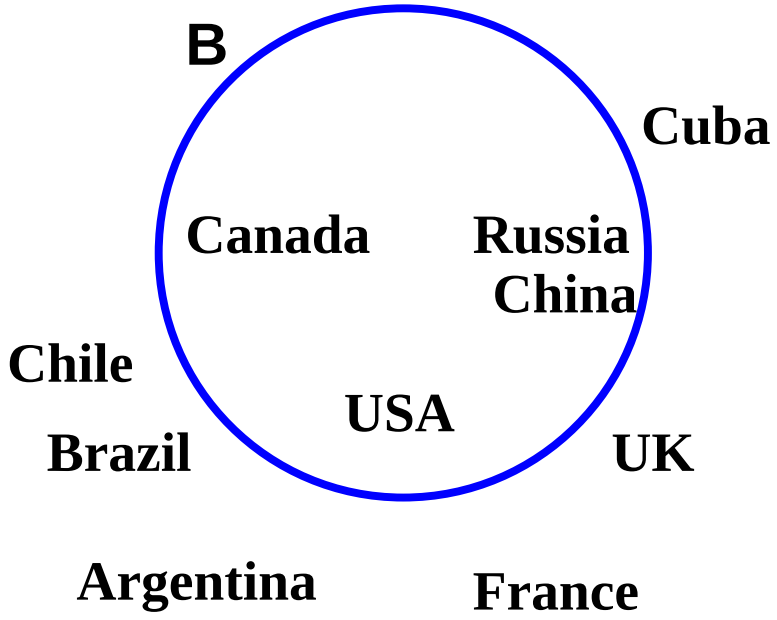
Naren Ramakrishnan
naren@cs.vt.edu

What are redescriptions?

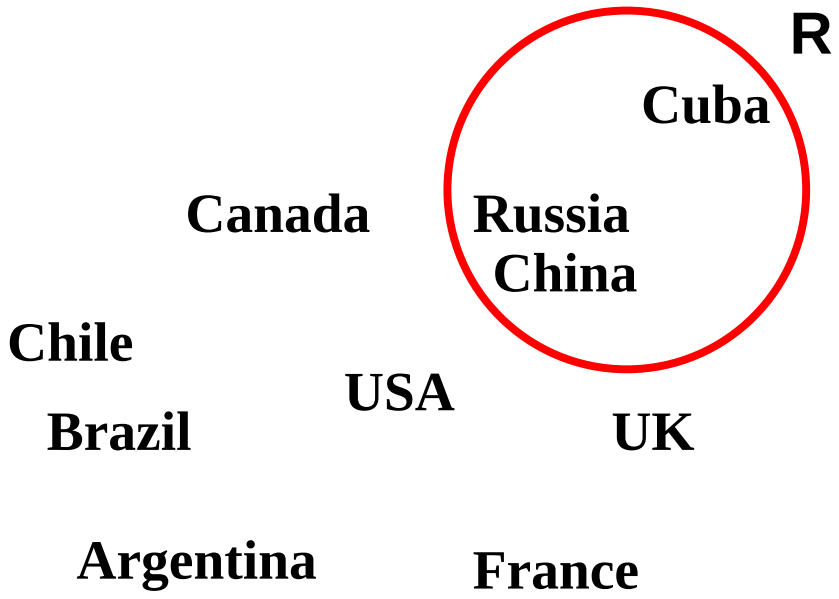
A shift-of-vocabulary, or a different way of communicating a given piece of information.

Input to Redescription Mining

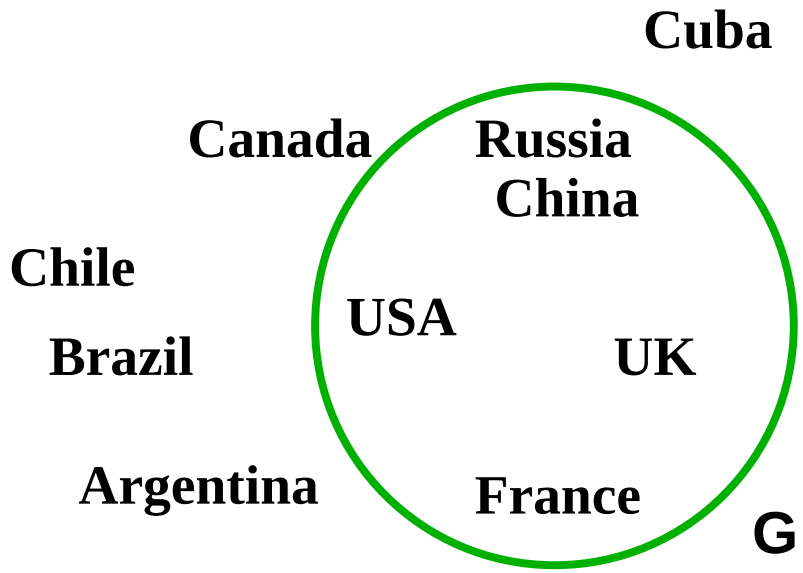
Input to Redescription Mining (contd.)



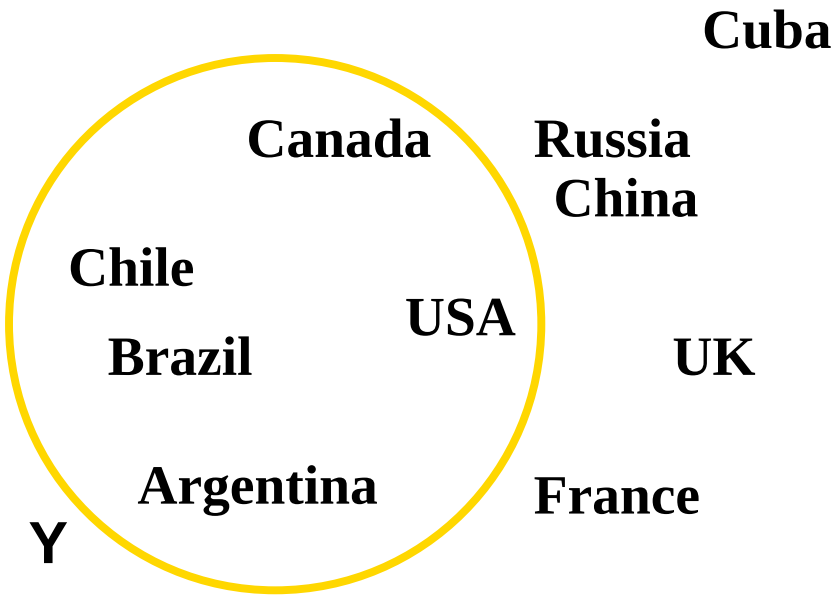
Input to Redescription Mining (contd.)



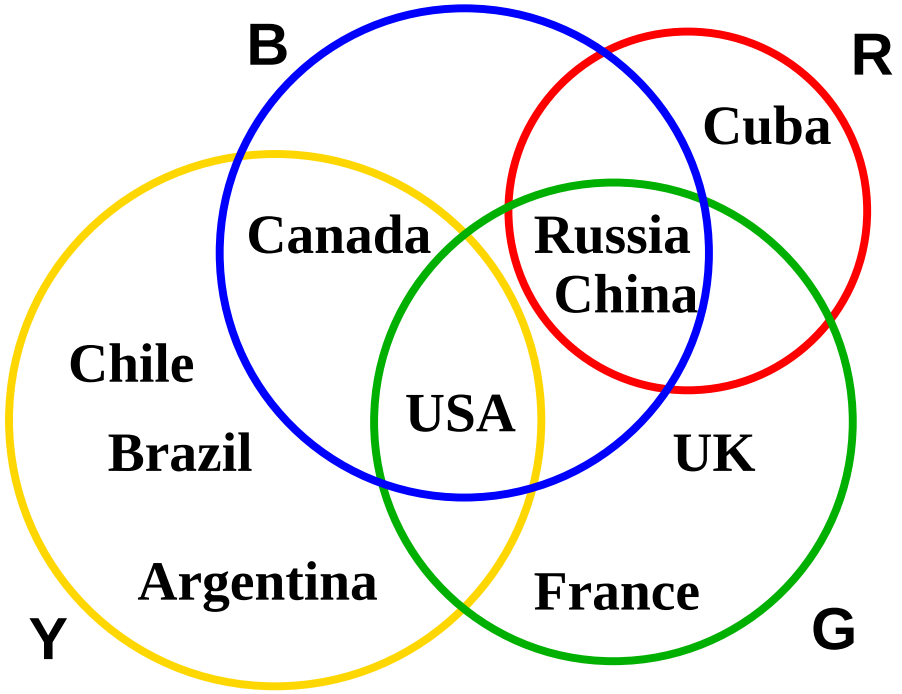
Input to Redescription Mining (contd.)



Input to Redescription Mining (contd.)



Input to Redescription Mining (contd.)



Basic Problem

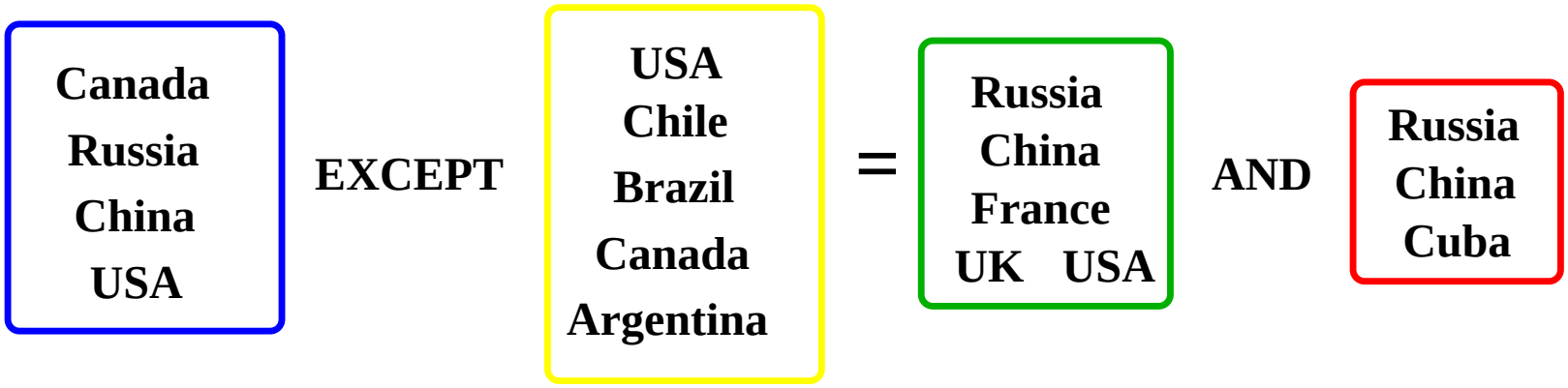
Given

- a set $\mathcal{O}$ of objects (e.g., countries)
- a collection of subsets (*descriptors*) of $\mathcal{O}$

Find

- subsets of $\mathcal{O}$ that can be defined in at least two ways

A Redescription



‘Countries with land area $>$ 3,000,000 sq. miles’ $-$

‘Tourist Destinations in the Americas’

$\Leftrightarrow$

‘Permanent members of U.N. Security Council’ $\cap$

‘Countries with history of communism’

Redescription is sort of like ...

association rule mining

- generalize from implications to equivalences

conceptual clustering

- find clusters with dual characterizations

constructive induction

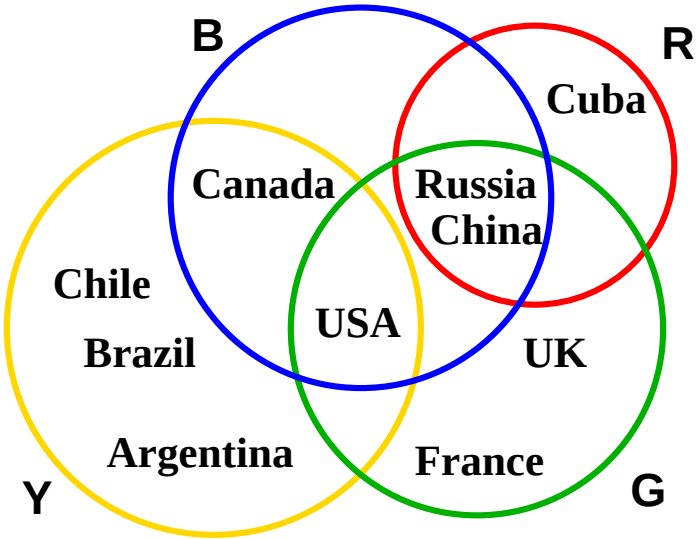
- build features that mutually reinforce each other

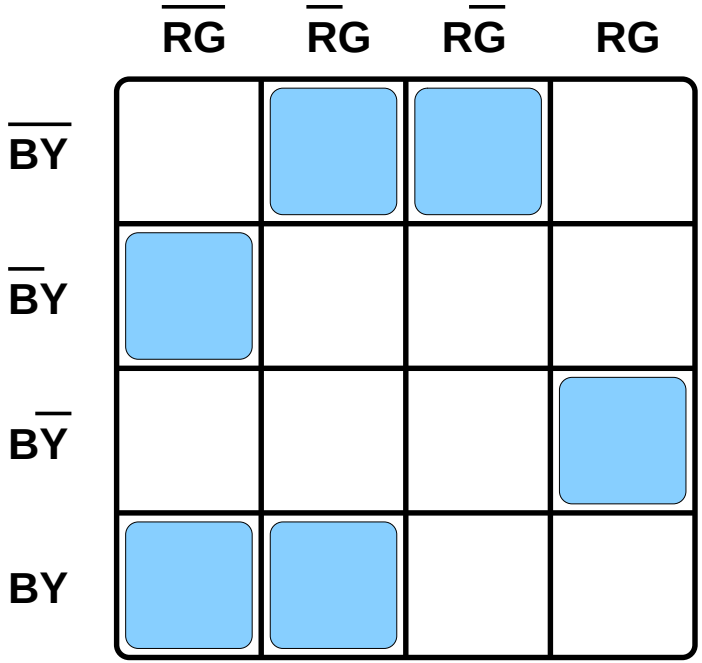
Applications in Bioinformatics

(Gene) subsets galore!

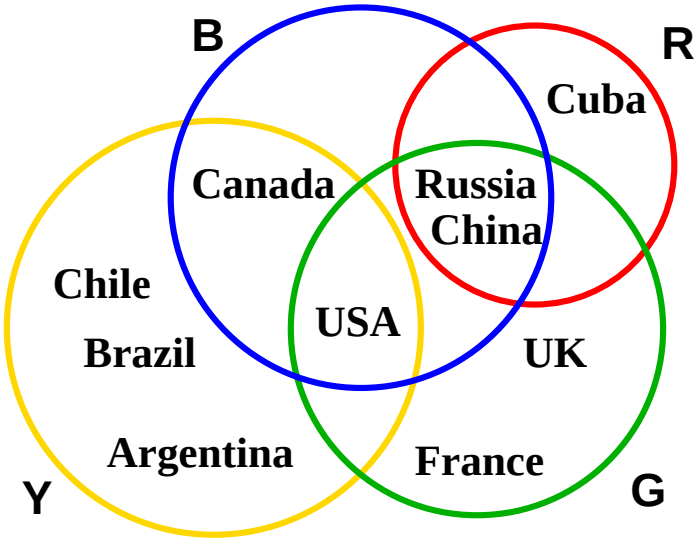
- Genes localized in the mitochondrion
- Genes up-expressed two-fold or more in heat stress
- Genes encoding for proteins forming the immunoglobin complex
- Genes involved in glucose biosynthesis
- Genes handpicked by Prof. Genie for further study
- Genes clustered together by your favorite algorithm
- . . .

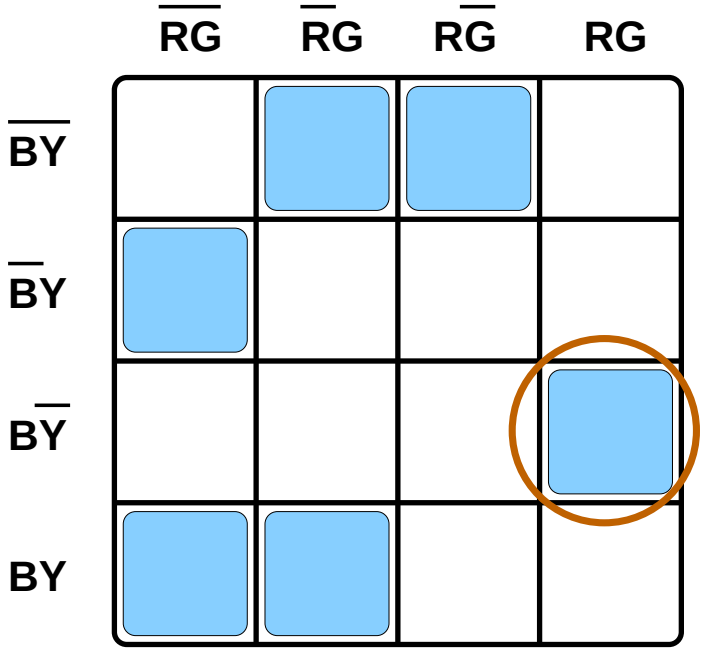
How do redescriptions happen?



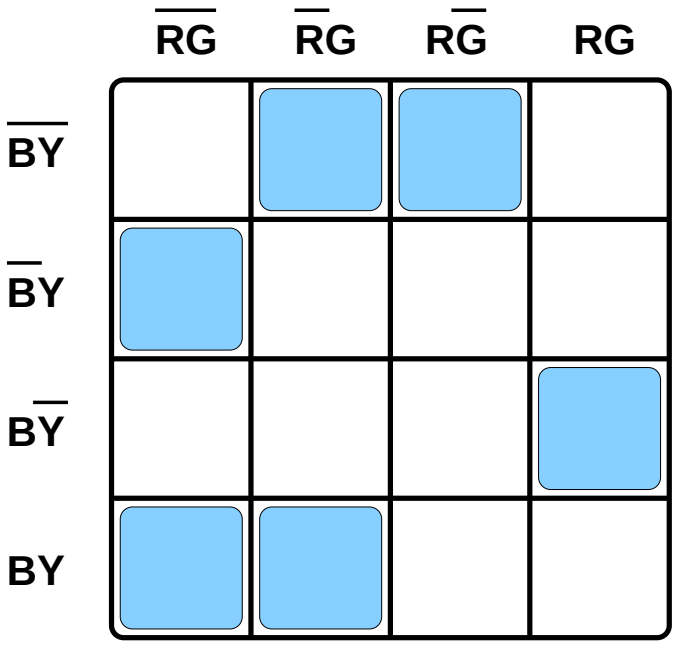


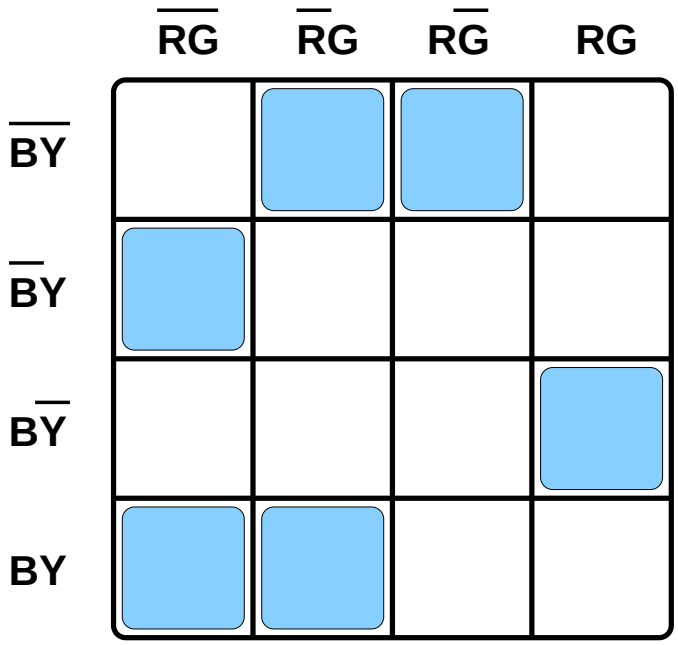
How do redescriptions happen?



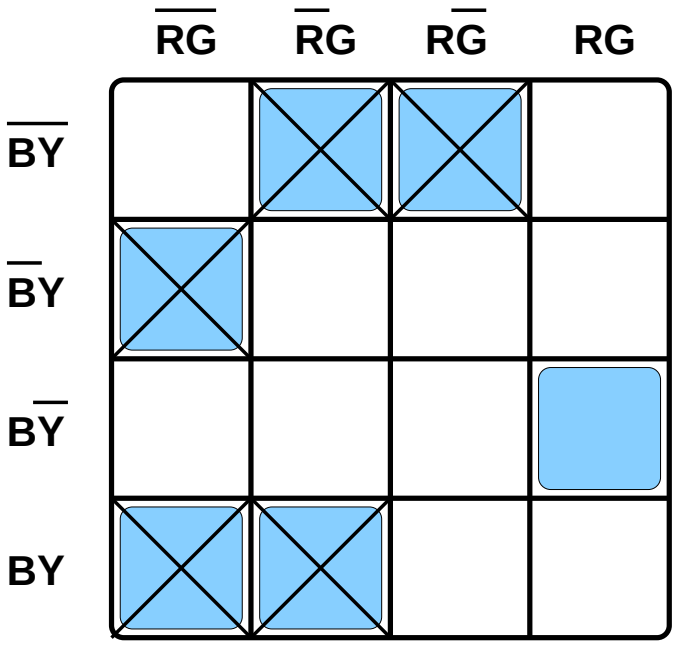


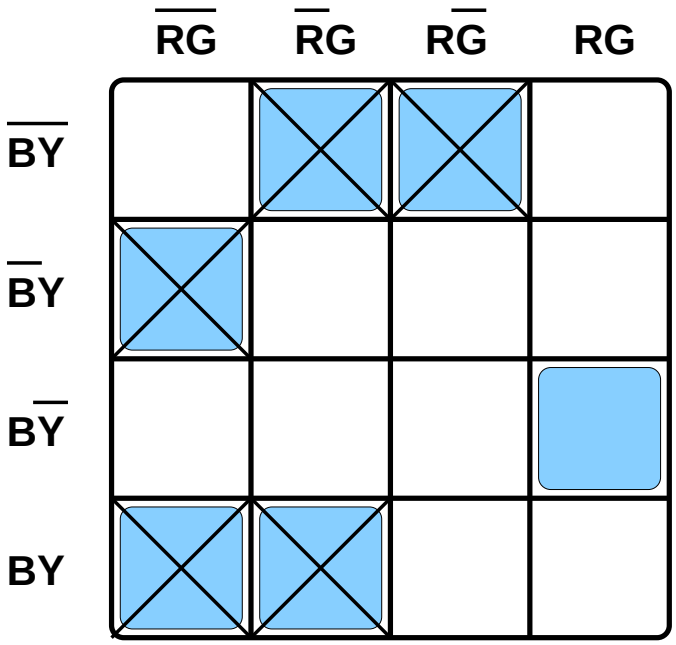
A game on Karnaugh maps



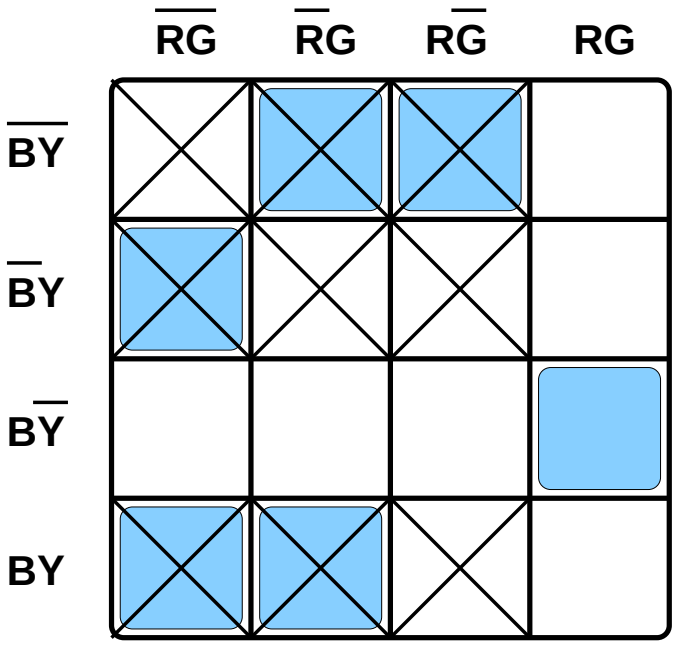


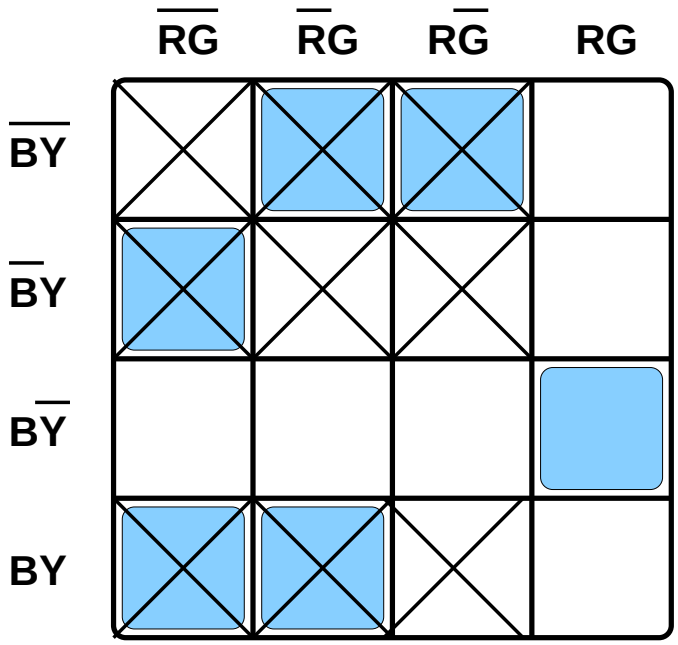
A game on Karnaugh maps



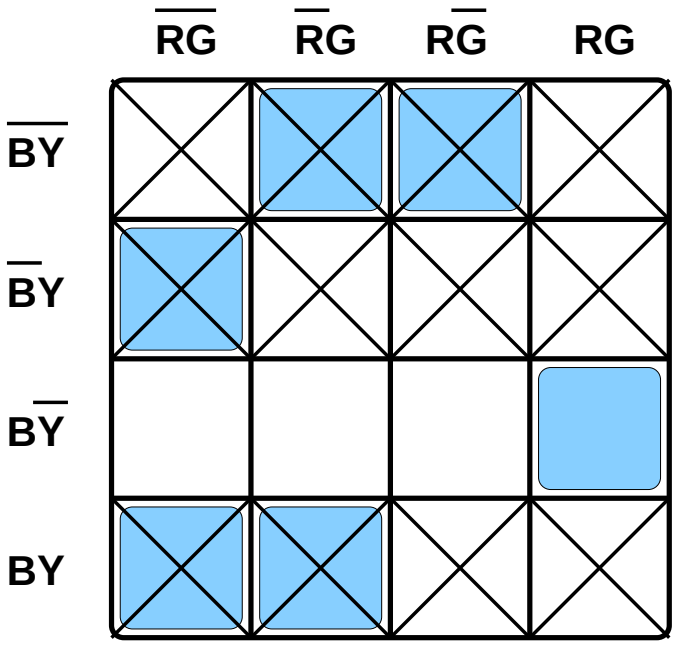


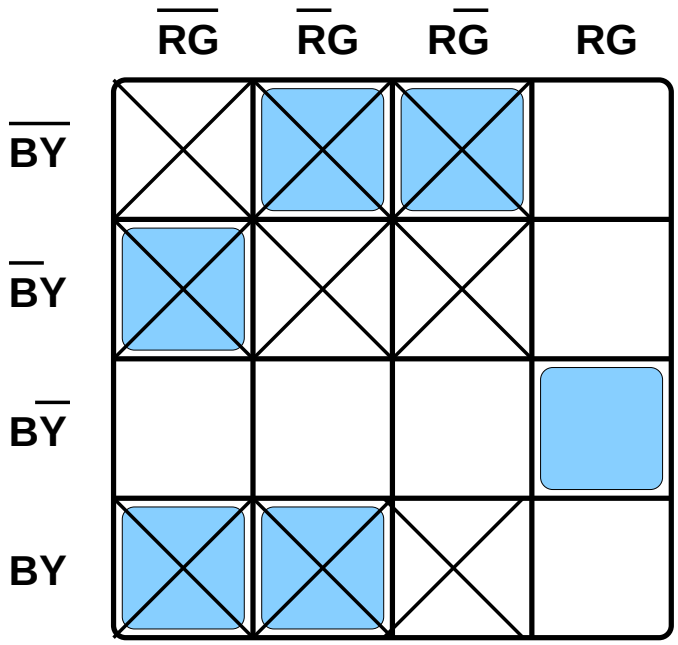
A game on Karnaugh maps



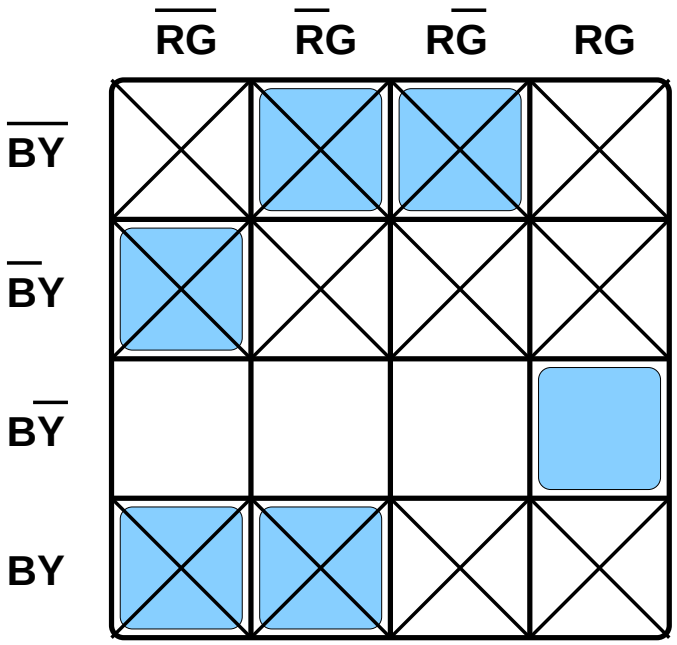


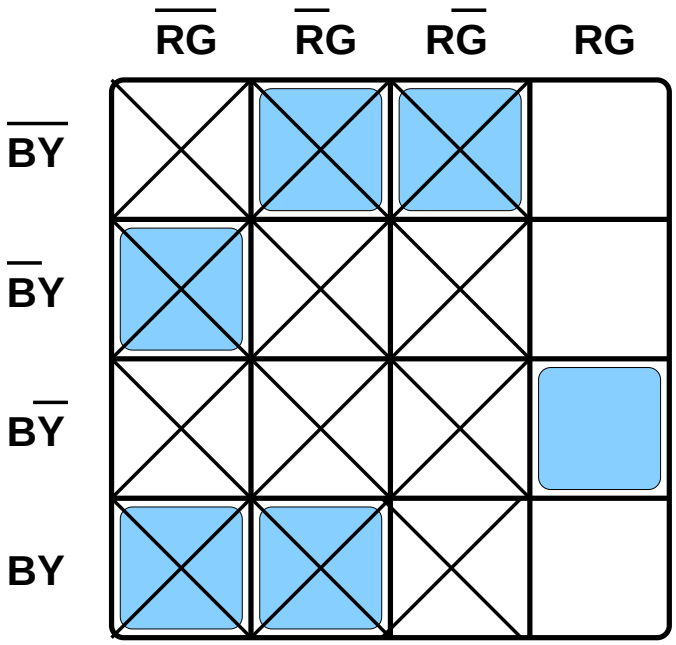
A game on Karnaugh maps



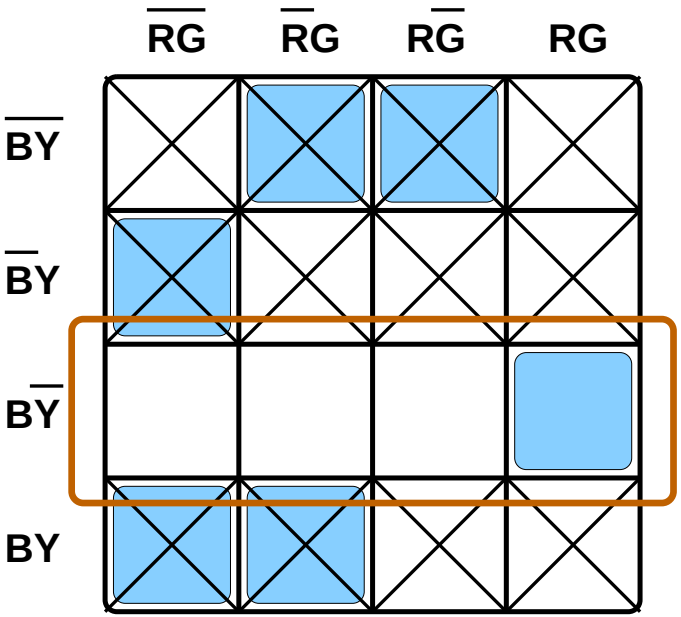


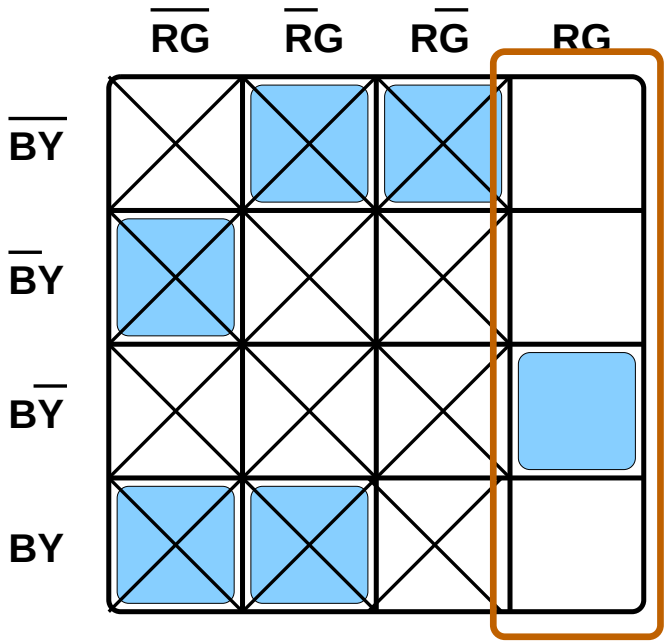
A game on Karnaugh maps



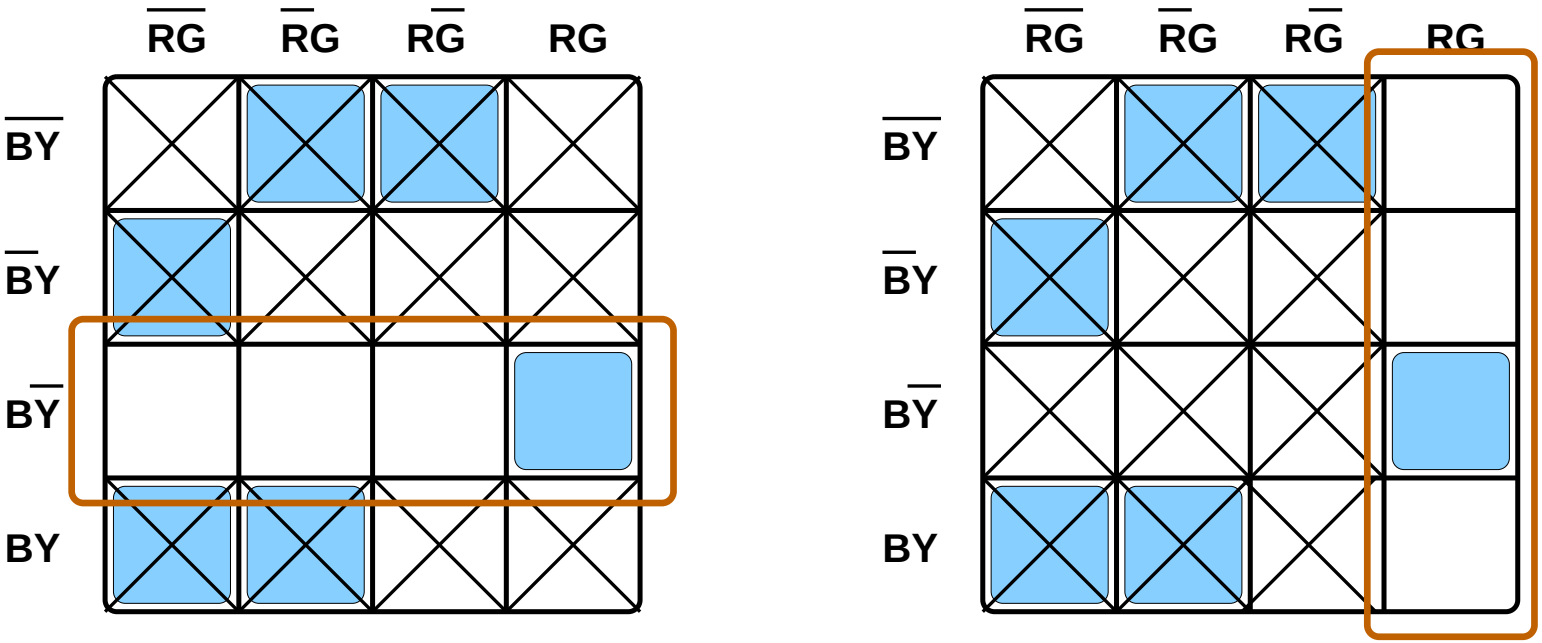


Reading off a redescription



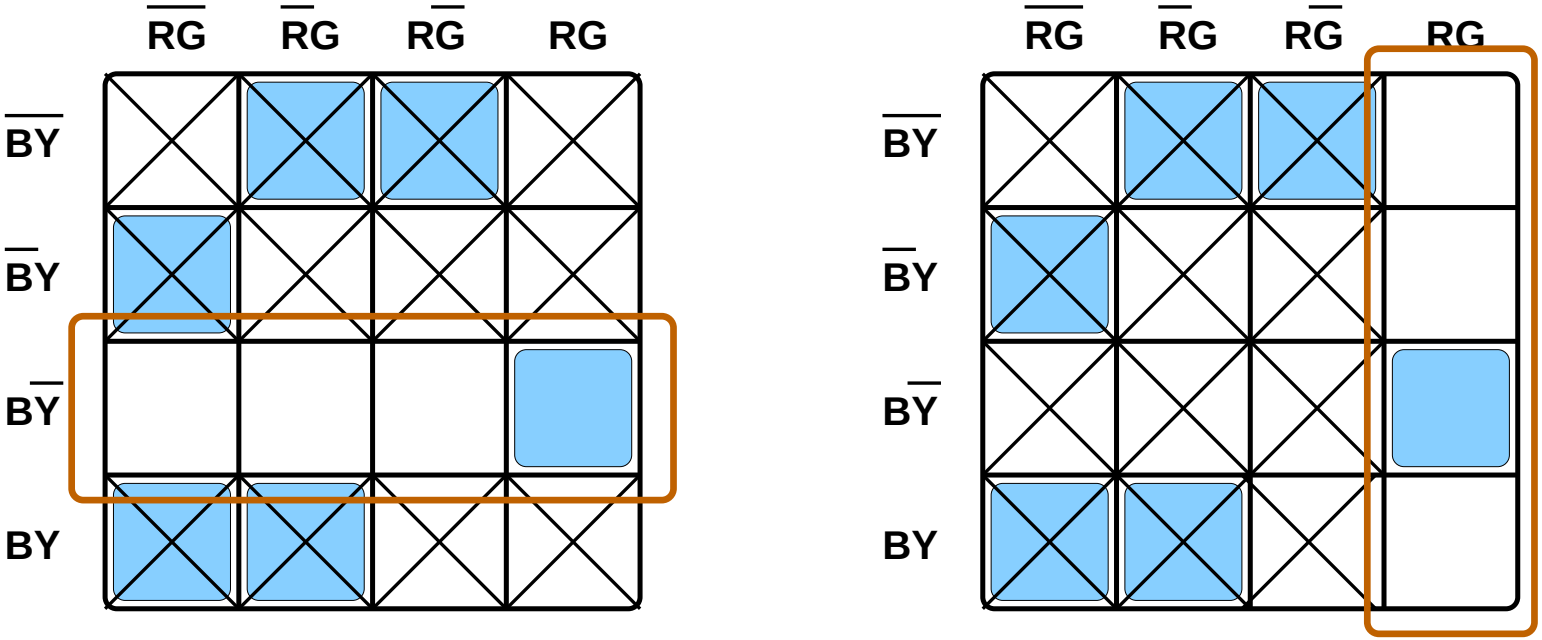


Reading off a redescription



$$(B\overline{Y}\,\overline{RG} \;\vee\; B\overline{Y}\,\overline{R}G \;\vee\; B\overline{Y}R\overline{G} \;\vee\; B\overline{Y}RG)$$

Reading off a redescription

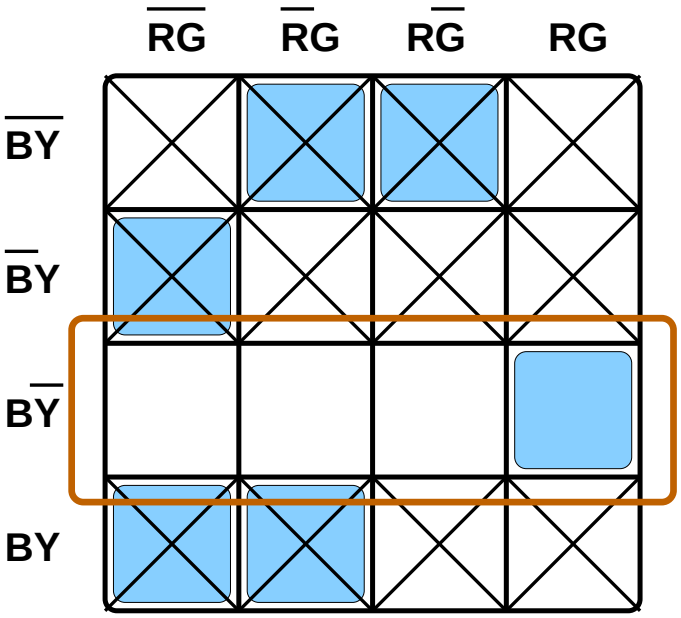


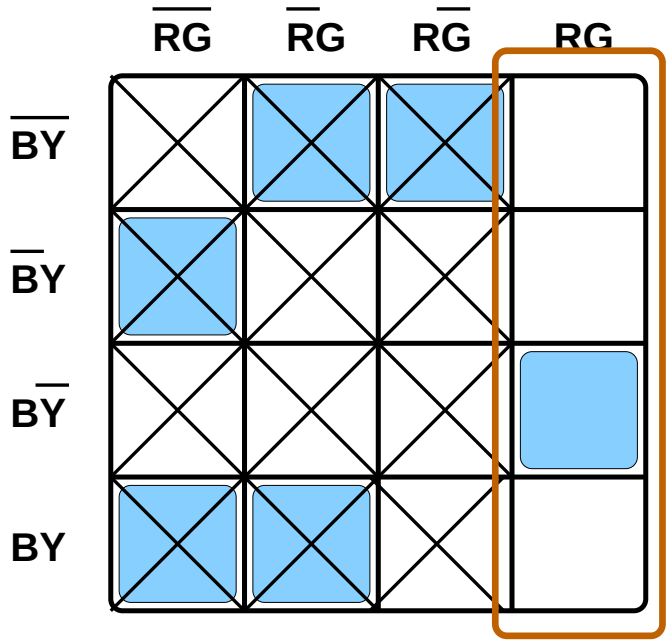
$$(B\overline{Y}\,\overline{RG} \;\vee\; B\overline{Y}\,\overline{R}G \;\vee\; B\overline{Y}R\overline{G} \;\vee\; B\overline{Y}RG)$$

$$\Leftrightarrow$$

$$(\overline{B}\overline{Y}\,\overline{RG} \;\vee\; \overline{B}\overline{Y}RG \;\vee\; B\overline{Y}RG \;\vee\; BYRG)$$

Reading off a redescription



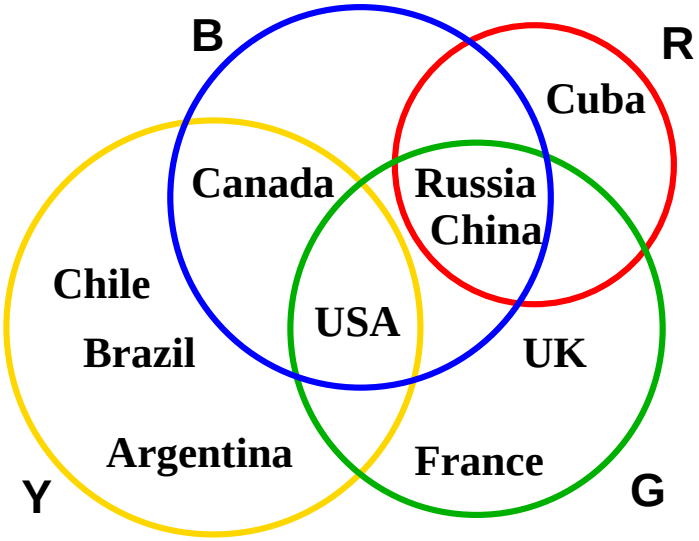


$$(B\overline{Y})$$

$$\Leftrightarrow$$

$$(RG)$$

Redescriptions help reason about sets



Q: How can B be made equal to R?

Ans: Subtract Y from B; intersect G with R, yielding $B\overline{Y} \Leftrightarrow RG$.

Some Definitions

Given a collection of objects $\mathcal{O}$ and descriptors $\mathcal{D}$:

- A redescription $X \iff Y$ ($X, Y \subseteq \mathcal{D}$) holds when
  - $X \cap Y = \emptyset$ and
  - X and Y induce the same set of objects.

Some Definitions

Given a collection of objects $\mathcal{O}$ and descriptors $\mathcal{D}$:

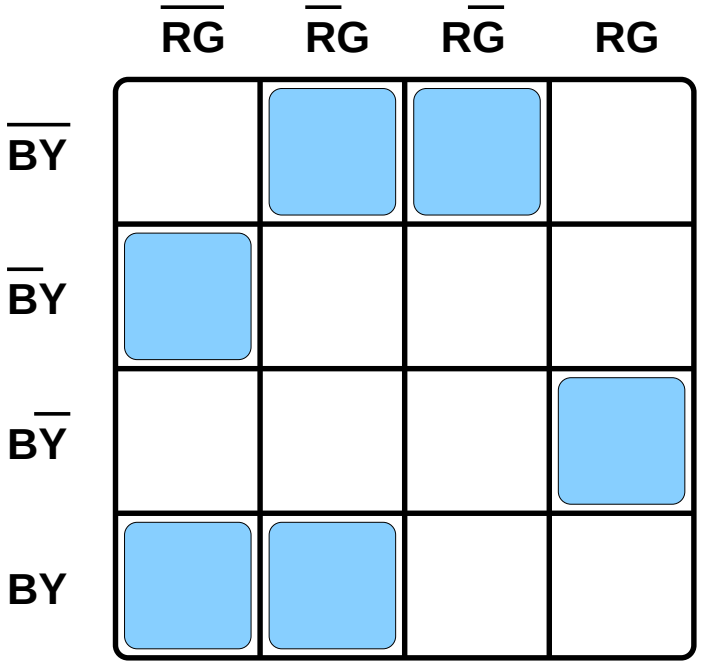
- A redescription $X \iff Y$ ($X, Y \subseteq \mathcal{D}$) holds when
  - $X \cap Y = \emptyset$ and
  - X and Y induce the same set of objects.
- A conditional redescription $X \iff Y|Z$ ($Z \subseteq \mathcal{D}$) holds when
  - $X \cap Y = X \cap Z = Y \cap Z = \emptyset$ and
  - $X \cap Z$ and $Y \cap Z$ induce the same set of objects.

Some Definitions

Given a collection of objects $\mathcal{O}$ and descriptors $\mathcal{D}$:

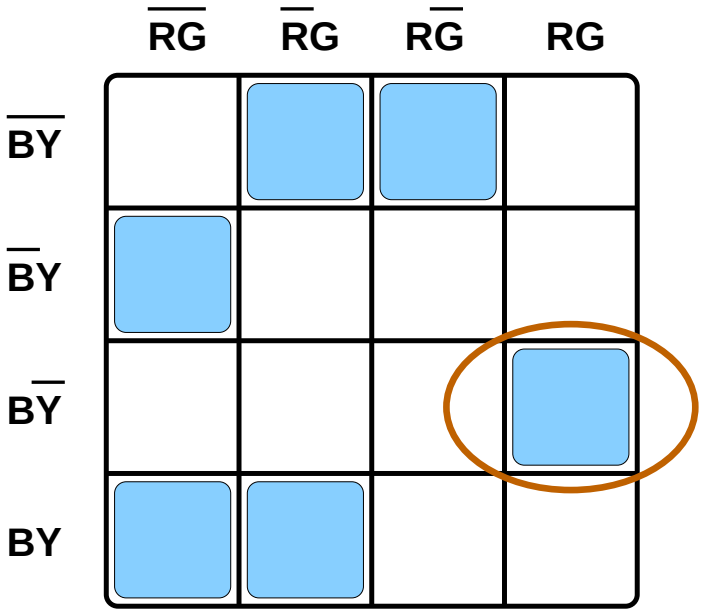
- A redescription $X \iff Y$ ($X, Y \subseteq \mathcal{D}$) holds when
  - $X \cap Y = \emptyset$ and
  - X and Y induce the same set of objects.
- A conditional redescription $X \iff Y|Z$ ($Z \subseteq \mathcal{D}$) holds when
  - $X \cap Y = X \cap Z = Y \cap Z = \emptyset$ and
  - $X \cap Z$ and $Y \cap Z$ induce the same set of objects.
- A redescription $X \iff Y$ is a non-redundant redescription iff there does not exist another redescription $X' \iff Y'$ for the same set of objects, such that $X' \subseteq X$ and $Y' \subseteq Y$

Connections to Association Rule Mining



Objects	=	Transactions
Descriptors	=	Items

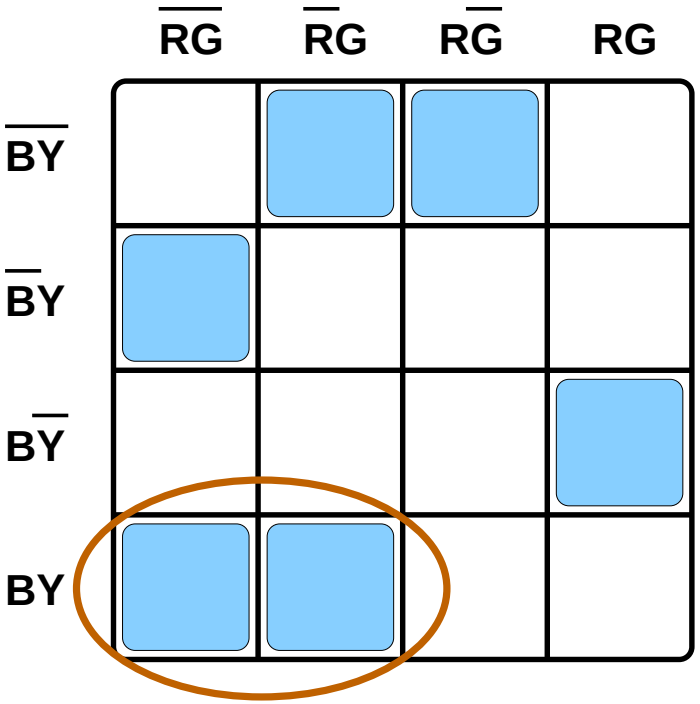
Connections to Association Rule Mining



Objects	=	Transactions
Descriptors	=	Items

Colored cell = closed itemset (e.g., $B\overline{Y}RG$)

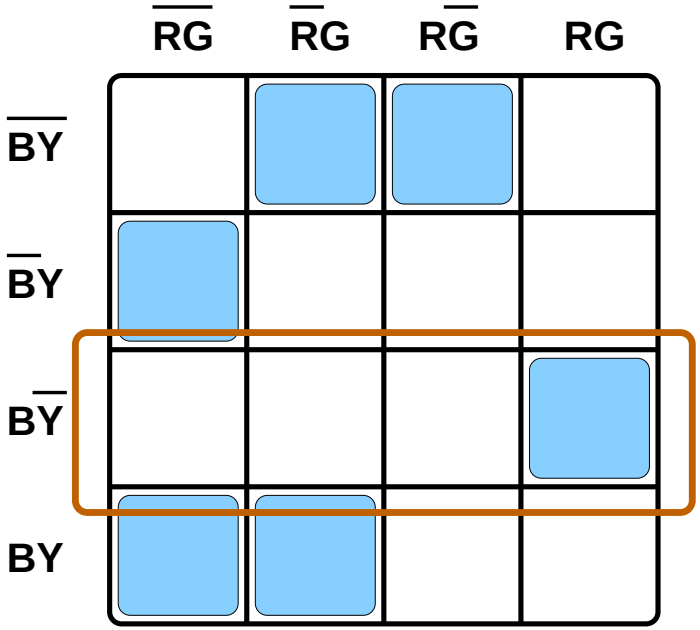
Connections to Association Rule Mining



Objects	=	Transactions
Descriptors	=	Items

Reducible cluster of colored cells = closed itemset (e.g., $BY\overline{R}$)

Connections to Association Rule Mining



Objects	=	Transactions
Descriptors	=	Items

Reducible cluster of mixed cells = non-closed itemset (e.g., $B\overline{Y}$)

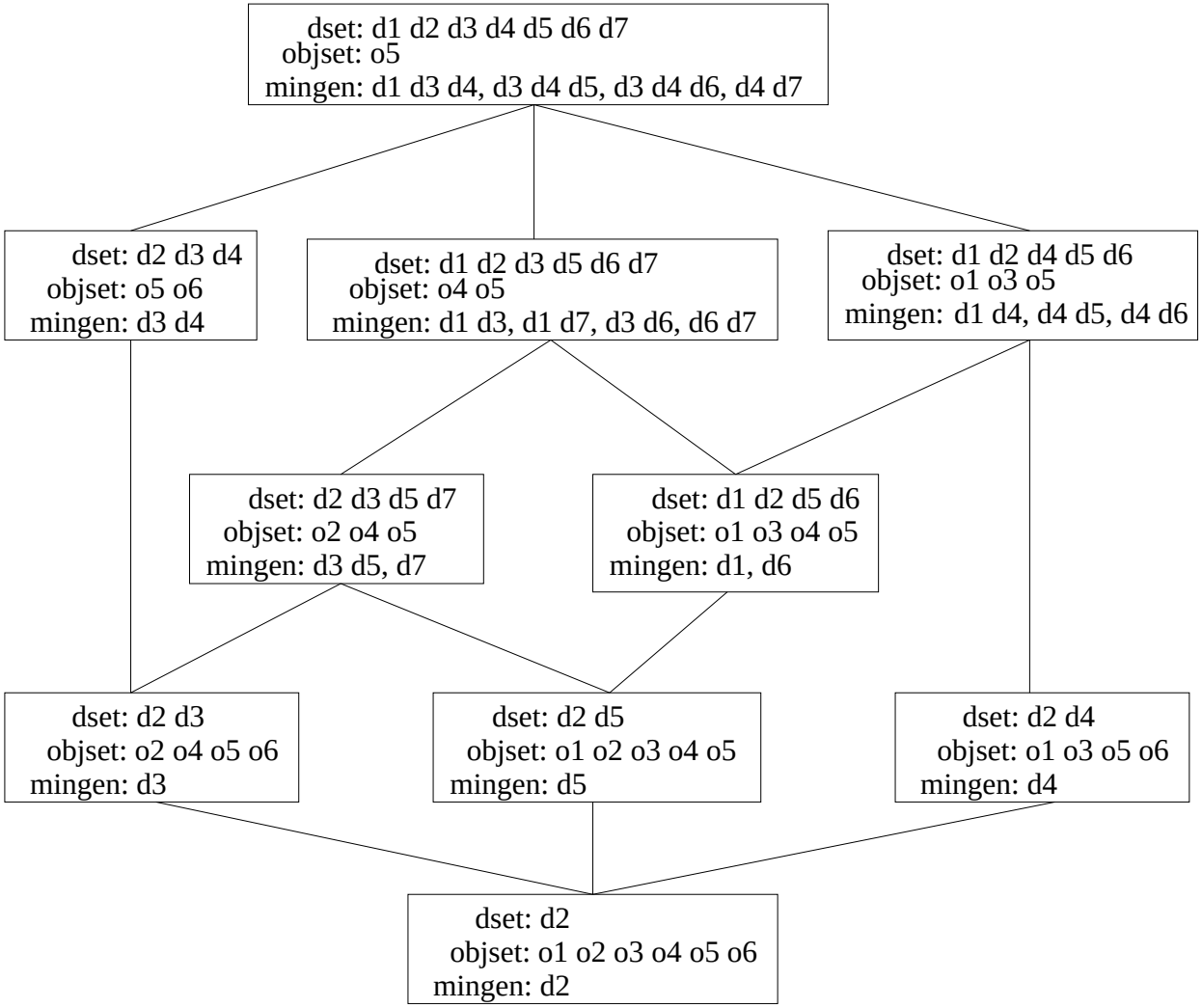
Adapting association mining algorithms

Mining redescriptions reduces to:

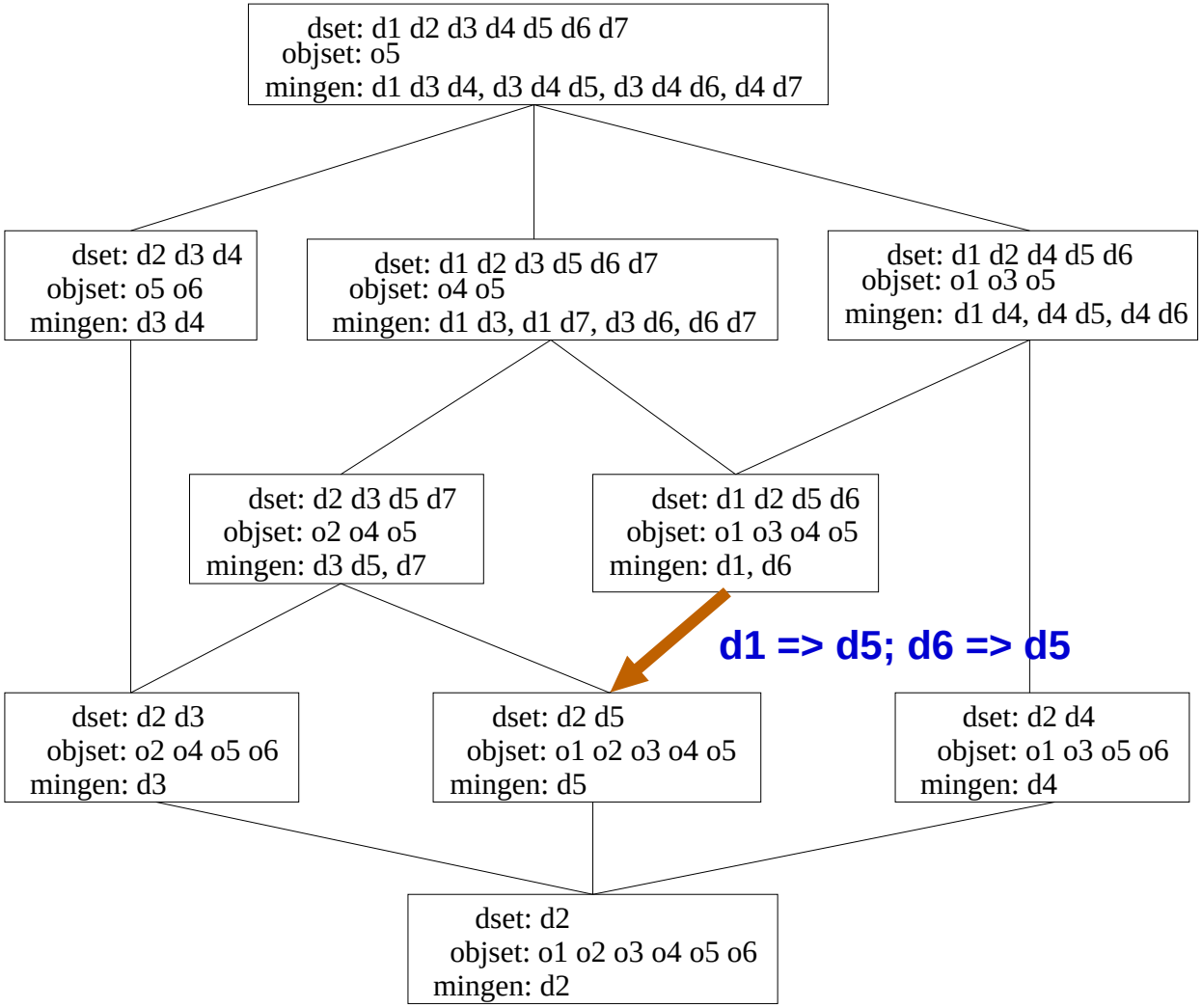
- mining closed itemsets (descriptor sets)
- obtain submatrices reducible to these closed sets (generators)

Object	Descriptors
o_1	$d_1 d_2 d_4 d_5 d_6$
o_2	$d_2 d_3 d_5 d_7$
o_3	$d_1 d_2 d_4 d_5 d_6$
o_4	$d_1 d_2 d_3 d_5 d_6 d_7$
o_5	$d_1 d_2 d_3 d_4 d_5 d_6 d_7$
o_6	$d_2 d_3 d_4$

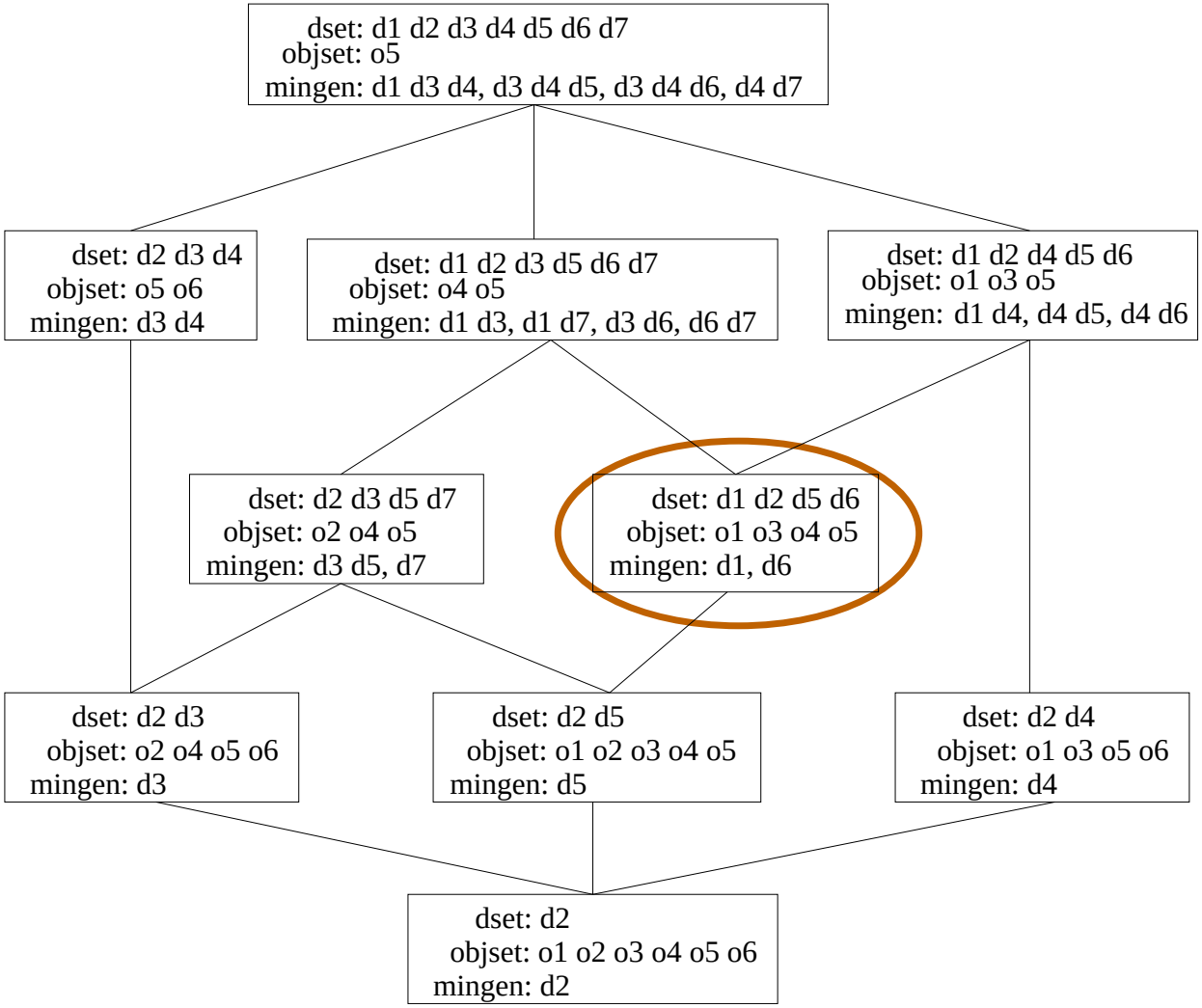
Lattice of Closed Sets



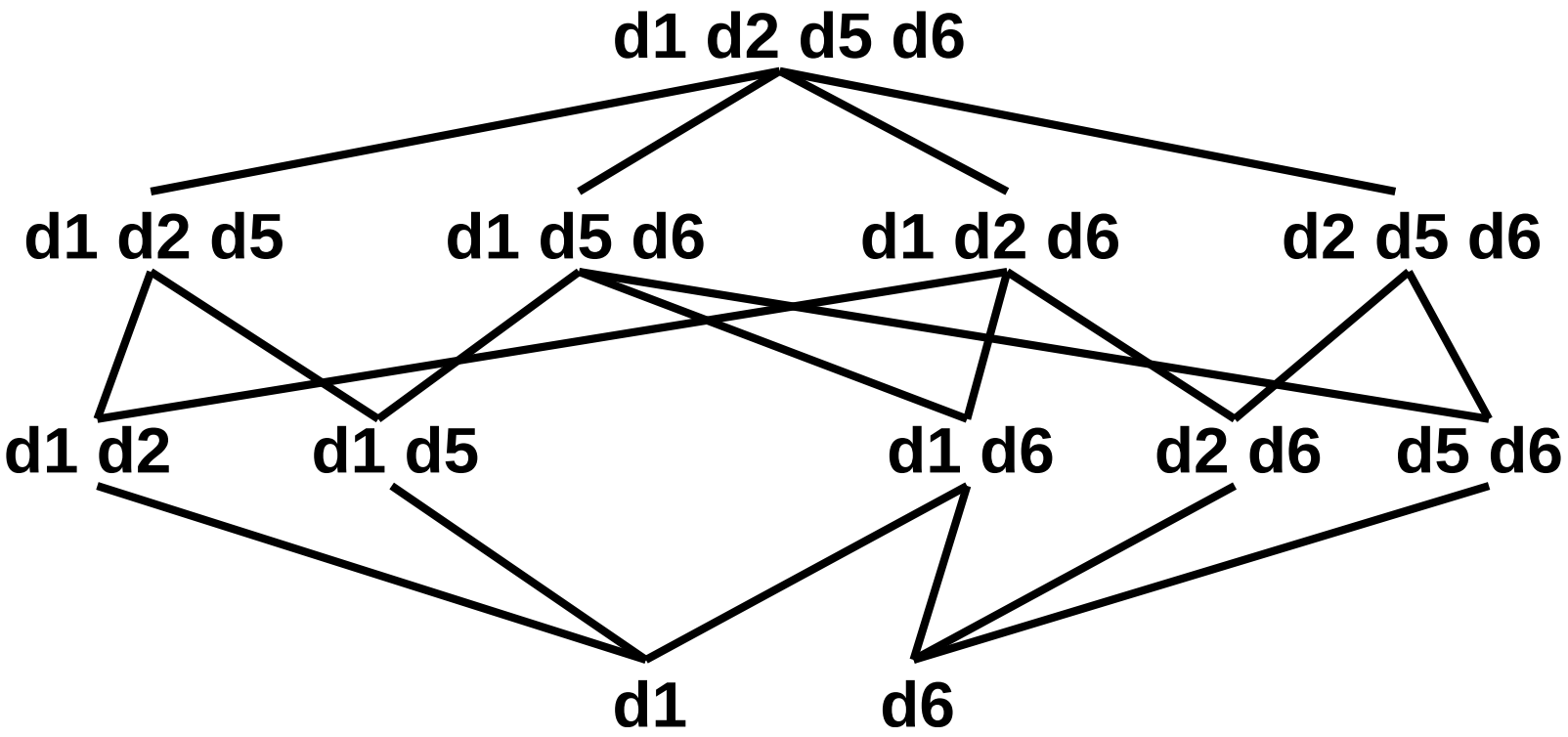
Lattice of Closed Sets



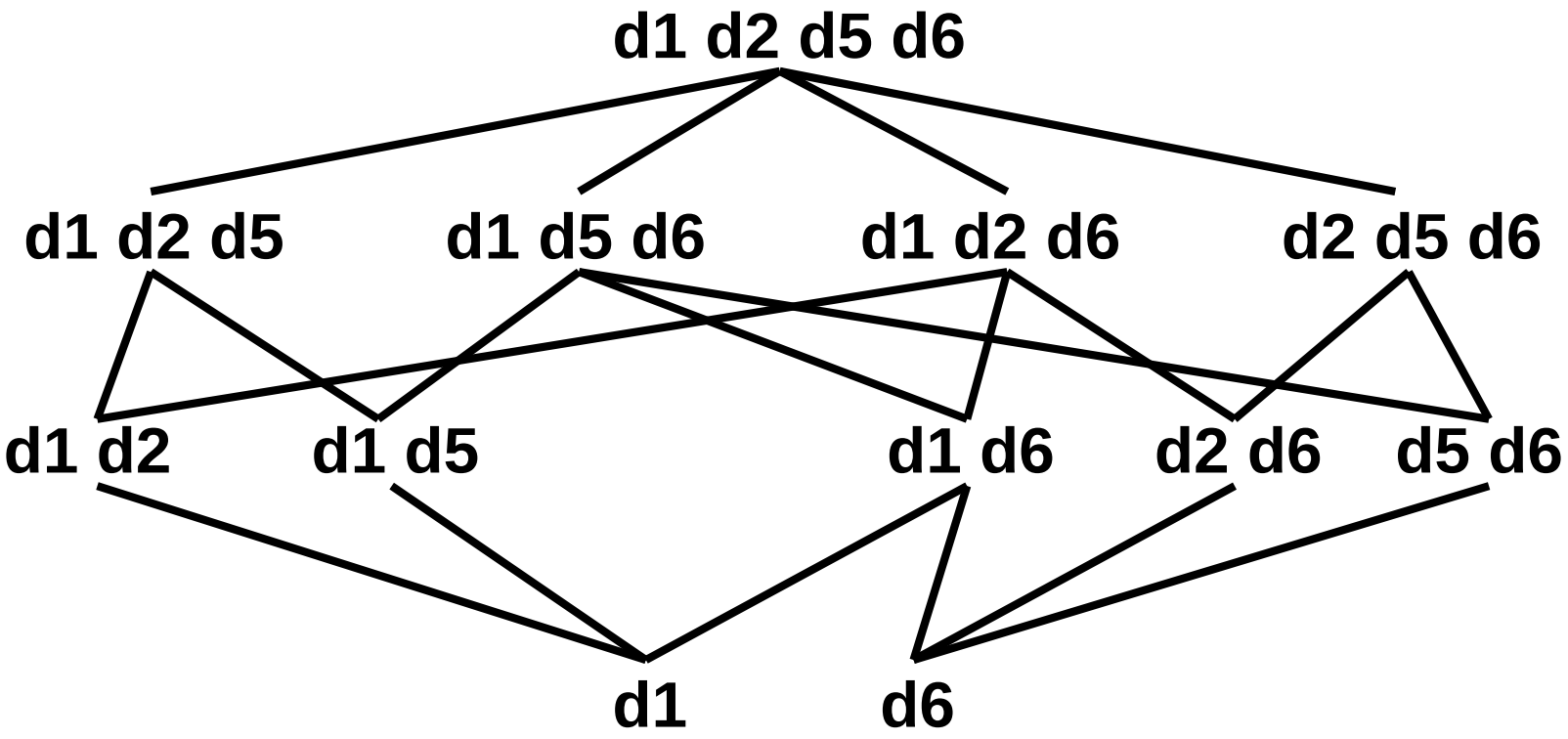
Lattice of Closed Sets



Up Closed and Personal

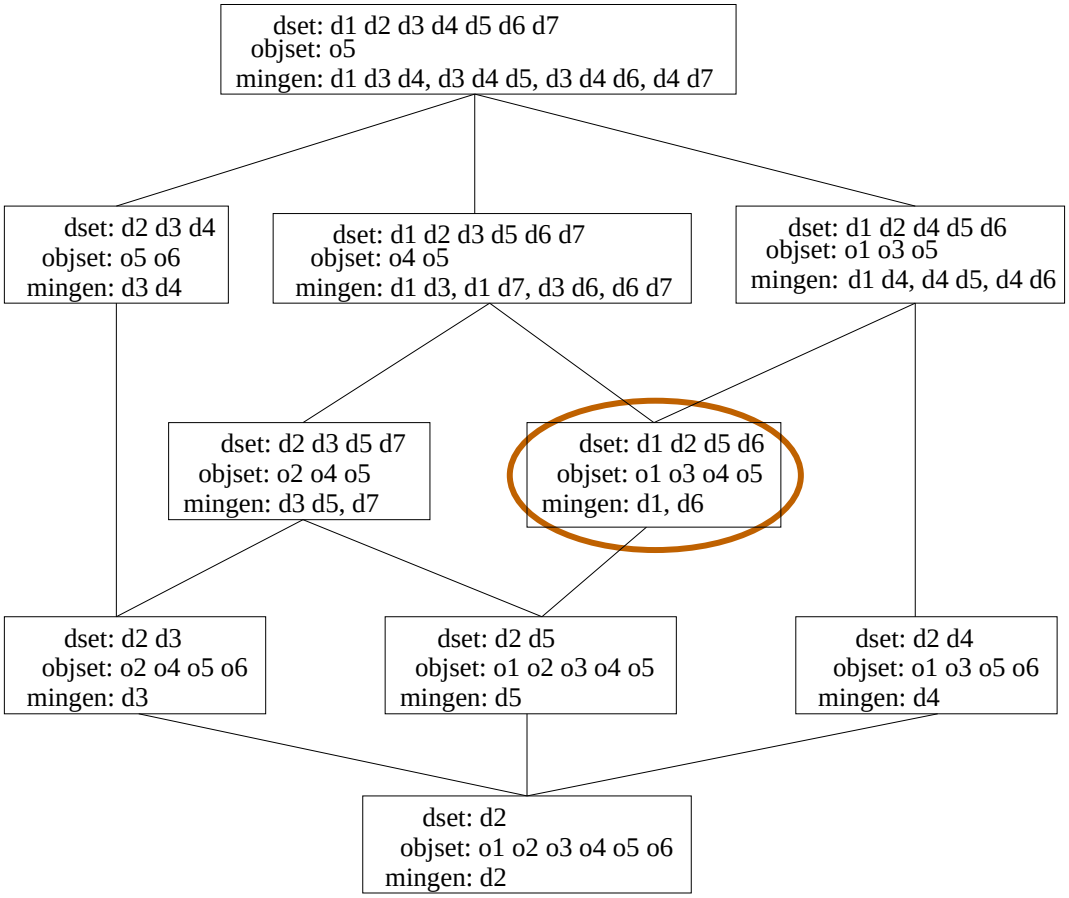


Up Closed and Personal

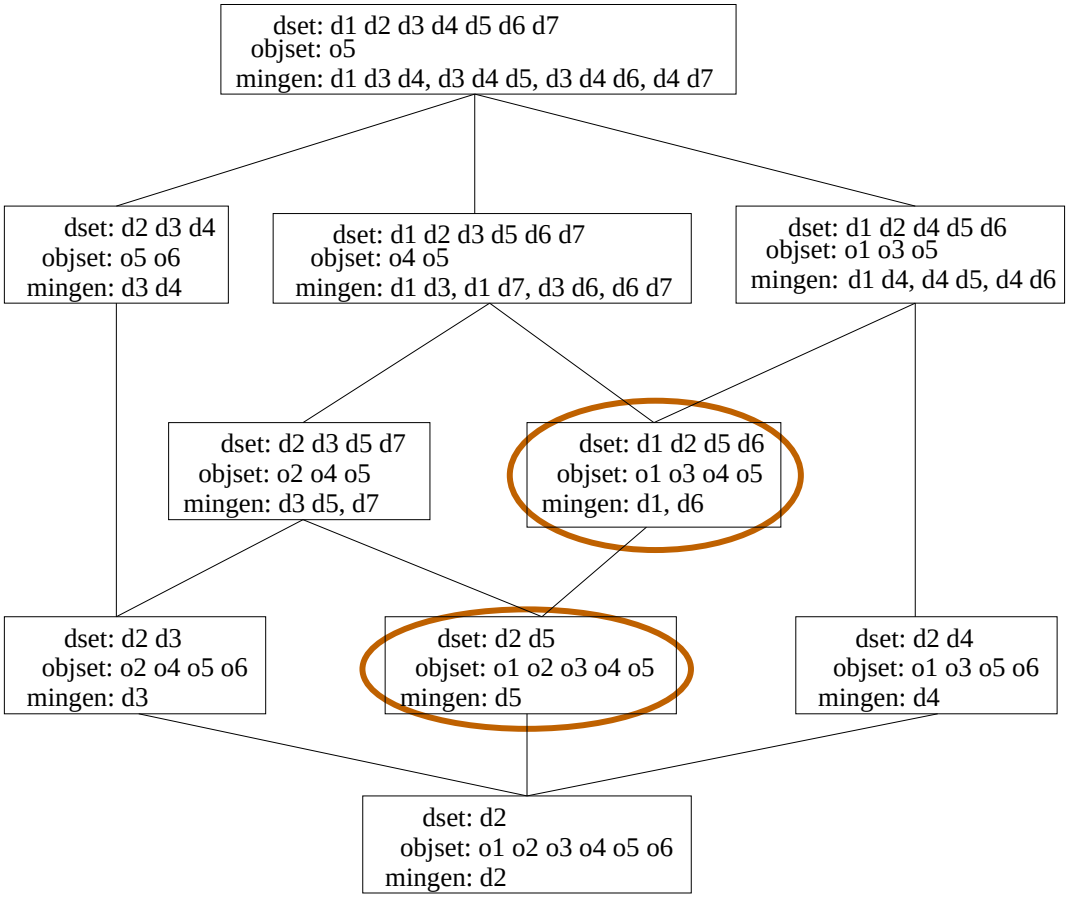


d1 <=> d6

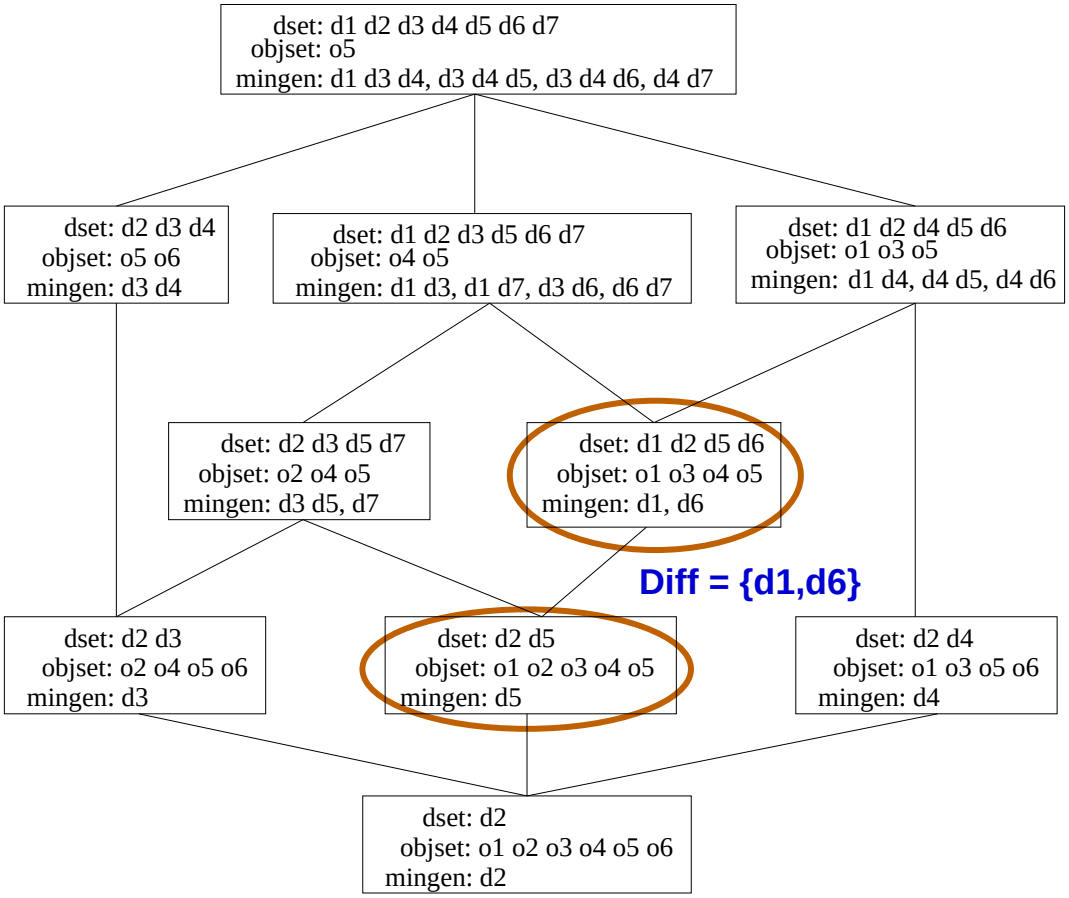
Finding Minimal Generators



Finding Minimal Generators



Finding Minimal Generators



Not so fast...

Datasets are 50% dense!

- cannot rely on pruning to help handle large datasets

Solution approach

- CHARM-L: Mining with constraints
  - Only expand lattice around objects/descriptors of interest

Exploring Gene Sets in Bioinformatics

Vocabularies

- GO functional categories (BIO, CEL, and MOL)
- Expression range buckets in specific microarray experiments
- Gene clusters

Interactive Exploration w/ CHARM-L

What is the relationship between ...

- $d183$ (ORFs ≥ 5 expressed in 15 minutes of heat shock)
- $d184$ (ORFs ≥ 5 expressed in 20 minutes of heat shock)

Answer:

- $d183 - d388 - d460 - d515 \Leftrightarrow d184 - d309$
  - $d388$: (GO MOL mannose transporter)
  - $d460$: (GO CEL external protective structure)
  - $d515$: (GO BIO fructose metabolism)
  - $d309$: (GO MOL molecular function unknown)

Another example

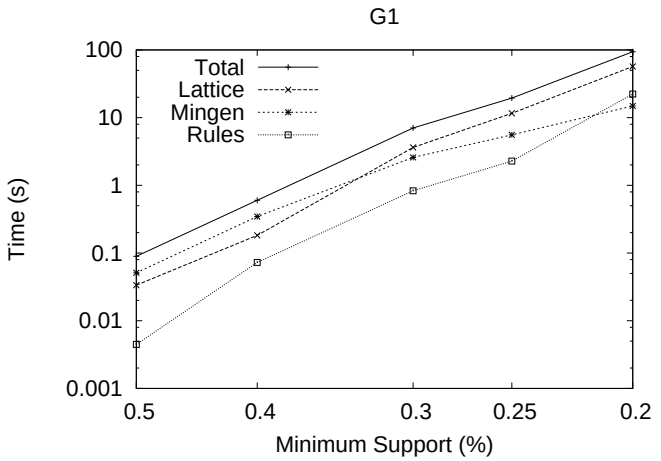
What is the relationship between ...

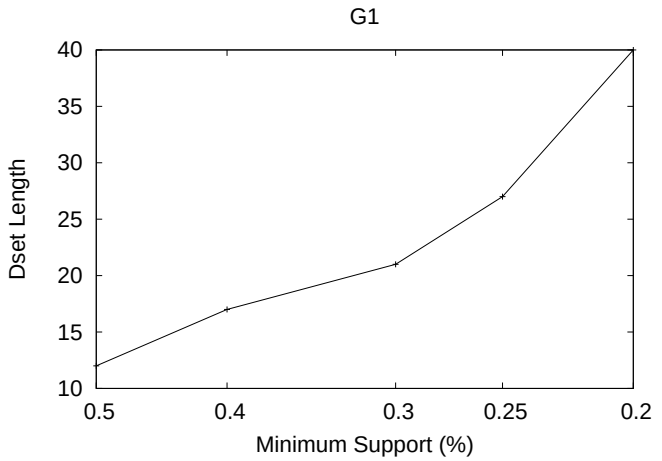
- $d141$ (ORFs ≥ 2 expressed in 10 minutes of heat shock)
- $d184$ (ORFs ≥ 5 expressed in 20 minutes of heat shock)

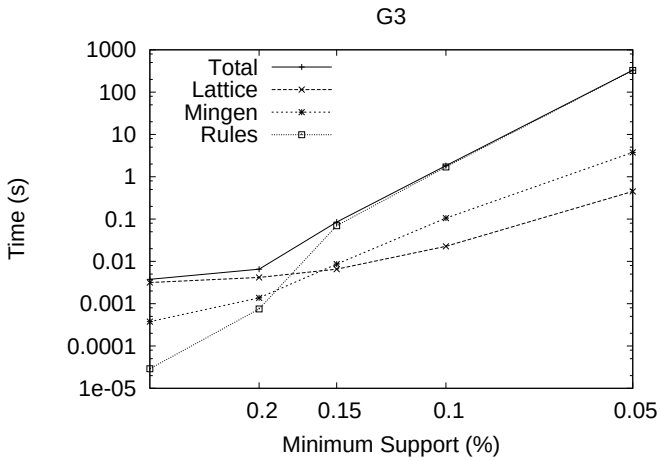
Answer:

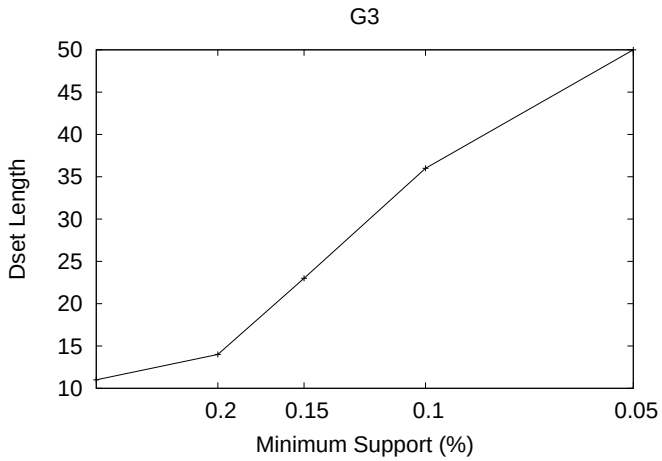
- $d141 - d515 - d608 \Leftrightarrow d184|d183$
  - $d515$: (GO BIO fructose metabolism)
  - $d608$: (ORFS ≥ 4 expressed in histone depletion)
  - $d183$: (ORFs ≥ 5 expressed in 15 minutes of heat shock)

Performance Results









Recap

Redescriptions help reason about set collections

- Conjunctive forms handle set intersections and negations
- Empowers biologist to create and work with vocabularies

Algorithmic innovations

- Lattice mining, finding minimal generators, constraint propagation
- Established connections to boolean formula manipulation

Future Work

Story telling

- Find a sequence of redescriptions connecting disjoint sets X and Y

Schema matching

- $X \subseteq O_1$, $Y \subseteq O_2$, O_1 and O_2 are related by relation R

Generalized boolean expressions

- Mine redescriptions in more expressive forms

Acknowledgements

Collaborators

- Deept Kumar (Virginia Tech)
- Laxmi Parida (IBM TJ Watson)

Funding

- NSF CAREER IIS-0092978, DOE Career DE-FG02-02ER25538, NSF grants EIA-0103708 and EMT-0432098 (Zaki)
- NSF grants IBN-0219332 and EIA-0103660 (Ramakrishnan)

Questions?

For related work, see:

N. Ramakrishnan *et al.*, Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions, in *Proceedings of KDD'04*, pages 266–275, 2004.

- L. Parida and N. Ramakrishnan, Redescription Mining: Structure Theory and Algorithms, in *Proceedings of AAAI'05*, pages 837-844, July 2005.

Contact:

Naren Ramakrishnan
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061
`naren@cs.vt.edu`
`http://www.cs.vt.edu/~naren`