Redescription Mining

10 July 2014

An Example

In last season of Italy’s Serie A, the games in which
the away team won
and the home team didn’t score in the first half
and the away team scored in the first half
were (approximately) the games in which
the home team scored at most once
and the away team was leading after the first half

Another Example

EDUSKUNTA
RIKSDAGEN

In the 2011 parliamentary elections in Finland, the candidates who
were female
or were at most 39 years old
were (approximately) the candidates who
supported gay families right to adopt outside the family

Third Example

The areas in Europe where
the Eurasian elk (*A. a. alces*) lives
are (approximately) the areas where
January’s maximum temperature is between –10℃ and +0.5℃
and June’s maximum temperature is between +12℃ and +25℃
and August’s average precipitation is between 50 and 140 mm

What do these statements have in common?

An Example

In last season of Italy’s Serie A, the games in which
the away team won
and the home team didn’t score in the first half
and the away team scored in the first half
were (approximately) the games in which
the home team scored at most once
and the away team was leading after the first half

An Example

In last season of Italy's Serie A, the games in which

the away team won
and *the home team didn't score in the first half*
and *the away team scored in the first half*

were (approximately) the games in which

the home team scored at most once
and *the away team was leading after the first half*

Another Example

EDUSKUNTA
RIKSDAGEN

In the 2011 parliamentary elections in Finland, the candidates who

were female
or *were at most 39 years old*

were (approximately) the candidates who

supported gay families right to adopt outside the family

Third Example

The areas in Europe where

*the Eurasian elk (*A. a. alces*) lives*

are (approximately) the areas where

January’s maximum temperature is between –10℃ and +0.5℃
and *June’s maximum temperature is between +12℃ and +25℃*
and *August’s average precipitation is between 50 and 140 mm*

What are redescriptions?

Informal Definition

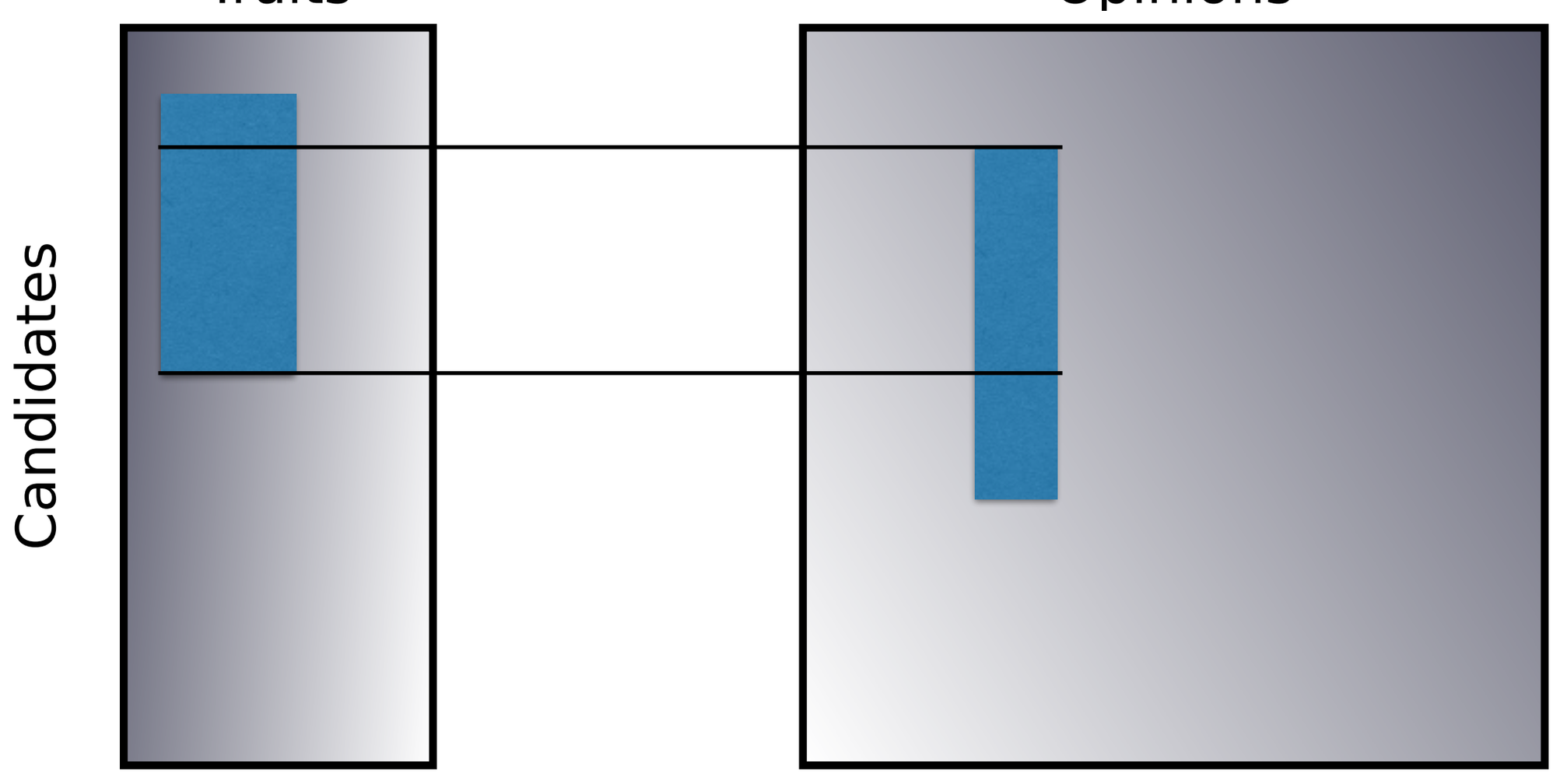
- A **redescription** provides two ways of describing the same set of entities
  - Descriptions are statements over entities' attributes
    - Tells us something about interesting attributes
  - Also the set of entities is interesting

Example

[Gender = F] ∨ [Age ≤ 39]

⇔

[Supports Gay Adoption Rights = True]



Some Definitions

- An **attribute** x has **domain** $dom(x)$
  - $dom(x) = \{0,1\}$ (binary), $dom(x) = \{a, b, \ldots, z\}$ (categorical), or $dom(x) \subseteq \mathbb{R}$ (numerical)
- If $X=\{x_1, x_2, \ldots, x_n\}$ is an ordered set of attributes, then $dom(X)$ is the set of all possible attributes' value tuples,

  $dom(X) = \{\langle y_1, y_2, \ldots, y_n\rangle :$

  $y_1 \in dom(x_1), y_2 \in dom(x_2), \ldots, y_n \in dom(x_n)\}$

More Definitions

- An **entity** e that has attributes X is a tuple in $dom(X)$
- **Data set** D_X is a set of entities,

  $D_X = \{e_i \in dom(X) : 1 \leq i \leq n\}$
- If the data set has **missing values**, we add special value *?* to each attribute's domain,

  $dom(x') = dom(x) \cup \{?\}$

Still More Definitions

- A **literal** over attribute x is a function

  l_x: $dom(x) \rightarrow \{\top, \bot\}$

  - E.g. [x], [x = ”Class”], or [x ≥ 10.5]

- A **query** over attribute set X is a Boolean function q_X over the literals of X’s attributes

  - Query q_X evaluates true on entity e, if the Boolean function evaluates true when the literals are evaluated with e’s values

Last Slide of Definitions

- The **support set** of query q_X in data D, $supp_D(q_X)$ is the set of entities in D where q_X evaluates true:

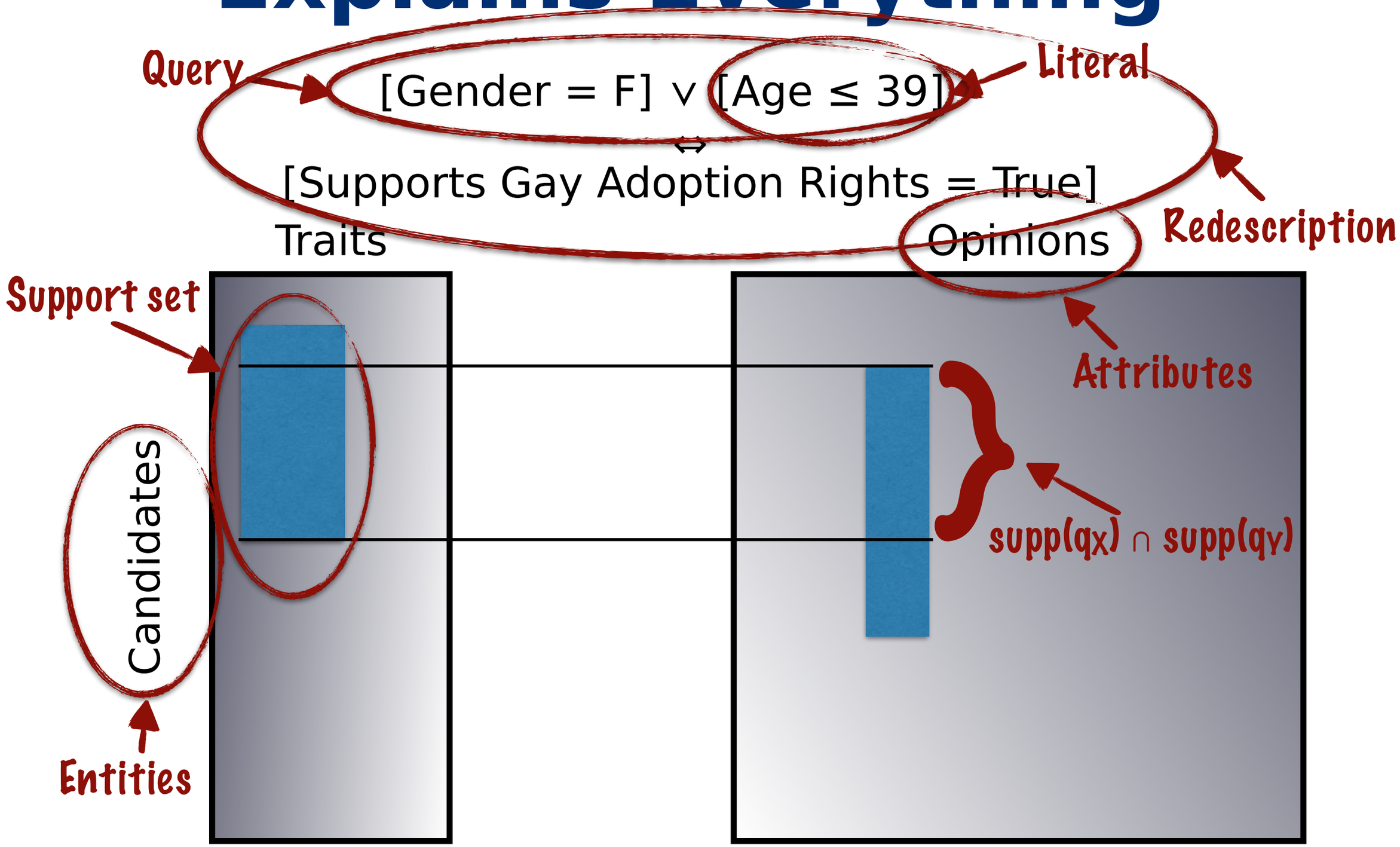
  $supp_D(q_X) = \{e \in D : q_X(e) = \top\}$

- The **support size** of q_X in D is $|supp_D(q_X)|$

… Just Kidding

- Let X and Y be two (non-overlapping) sets of attributes of entities in D and let q_X and q_Y be queries over X and Y
- The pair (q_X, q_Y) is called a **redescription**
- The **Jaccard coefficient** between q_X and q_Y is $$J(q_X, q_Y) = \frac{|supp_D(q_X) \cap supp_D(q_Y)|}{|supp_D(q_X) \cup supp_D(q_Y)|}$$

The One Slide that Explains Everything



Types of Redescriptions

- Types of data (only Boolean, with categorical, with numerical, with missing values)
- Types of queries (monotone conjunctive, monotone, tree-type, linear parsing tree, …)
- Other restrictions (min Jaccard, min support, max support, max number of attributes, *p*-value, …)

Why Redescriptions?

Two Views are Better than One

- Redescriptions help us to understand the data
  - E.g. in Finnish politics, women and young candidates express more liberal opinions
- Redescriptions find very complicated form of correlation
  - E.g. Eurasian Elk and it’s bioclimatic niche

Algorithms

Redescription Mining as Association Rule Mining

- Bi-directional association rules
  - Only binary variables
  - q_X and q_Y restricted to monotone conjunctive queries
- Jaccard coefficient is symmetric confidence
  - $q_X \Rightarrow q_Y$ and $q_Y \Rightarrow q_X$ must both have high confidence

Redescription Mining as Classification

- Query q_Y given, build q_X
  - q_Y defines a binary labeling of data entities (is in the support or not)
- A binary classification task
  - But the classifier must return query-type classification rules

CARTwheels

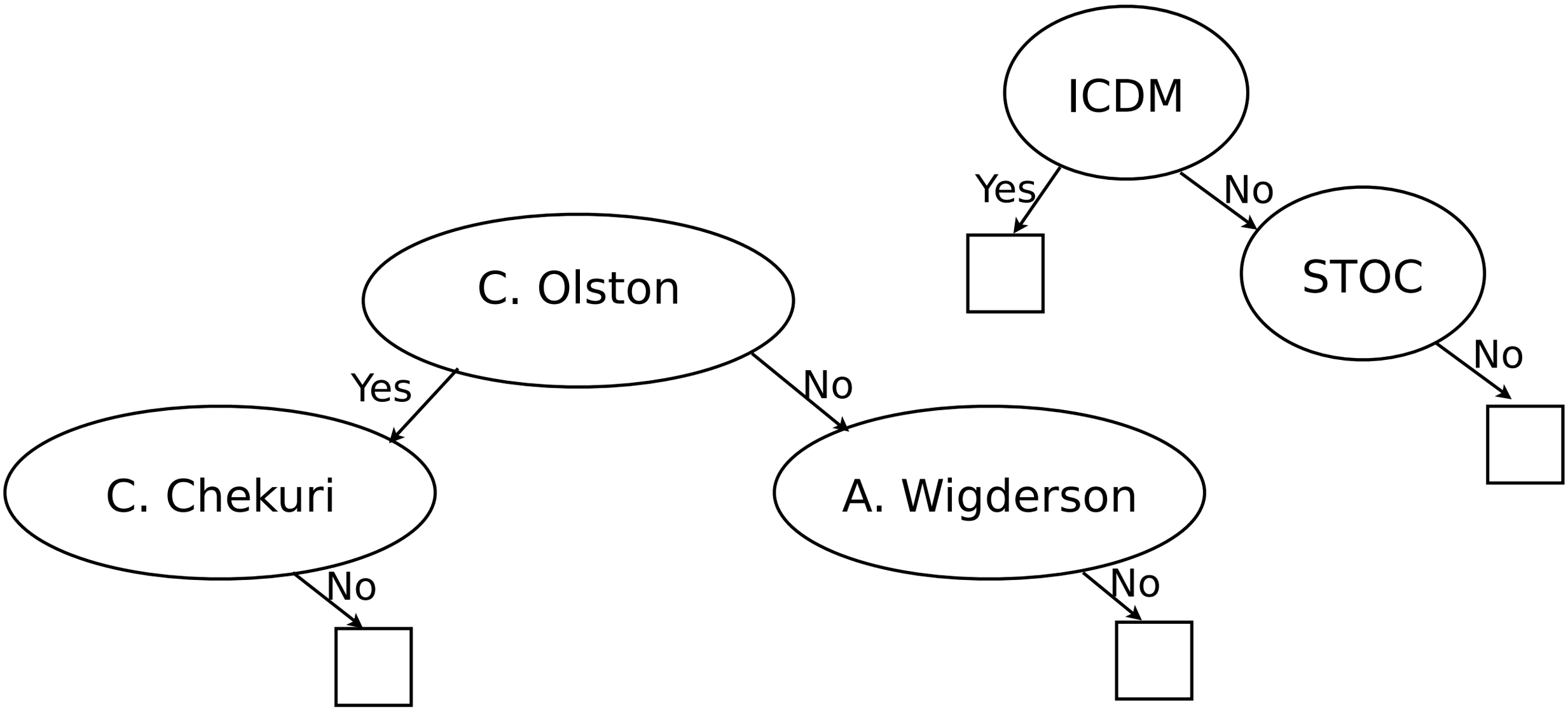
- Classification approach
  - Classification and regression trees (CARTs)
  - Fix one tree and grow the other to match; alternate
  - Leaves are matched and paths are the descriptions

Ramakrishnan, N., Kumar, D., Mishra, B., Potts, M., & Helm, R. F. (2004). Turning CARTwheels: an alternating algorithm for mining redescriptions (pp. 266–275). In KDD ’04.

CARTwheels Example

(ICDM) ∨ (¬ICDM ∧ ¬STOC) ⇔

(C. Olston ∧ ¬C. Chekuri) ∨ (¬C. Olston ∧ ¬A. Wigderson)



ReReMi

- First find a set of good singleton query pairs
  - (q_X, q_Y) where q_X and q_Y both contain just one literal
- Try to extend q_X and q_Y with one new literal
  - $q_X \wedge l$, $q_X \vee l$, $q_X \wedge \neg l$, $q_X \vee \neg l$
  - Use **beam search** for extensions
    - Keep the top-k extensions

Galbrun, E. & Miettinen, P., 2012. From black and white to full color: Extending redescription mining outside the Boolean world. Statistical Analysis and Data Mining, 5(4), pp.284–303.

On the Type of Descriptions

- CARTwheels finds tree-shape queries
  - (A and (B and C) or (not B)) or (not A and…)
  - The published algorithm only works with binary data, but extensions should be doable
- ReReMi linearly-parsable queries
  - ”(A or B) and C”, but not ”A and (B or C)”
- ReReMi can handle real-valued and categorical data
  - And can control the vocabulary of the queries

Suggested Reading

- Kumar, D., 2007. Redescription Mining: Algorithms and Applications in Bioinformatics. PhD thesis, Virginia Tech.
- Galbrun, E., 2013. Methods for Redescription Mining. PhD thesis, University of Helsinki.
- http://www.cs.helsinki.fi/u/galbrun/redescriptors/siren/sigmod/