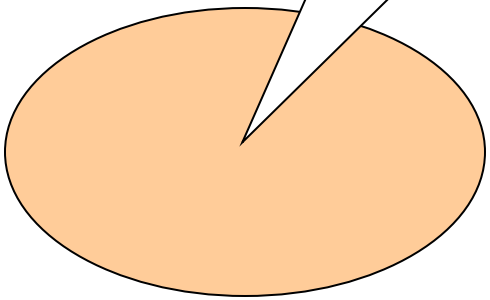
Learning Relational Bayesian Classifiers from RDF Data

Harris Lin, Neeraj Koul, and Vasant Honavar
Artificial Intelligence Research Laboratory
Department of Computer Science
Iowa State University
htlin@iastate.edu

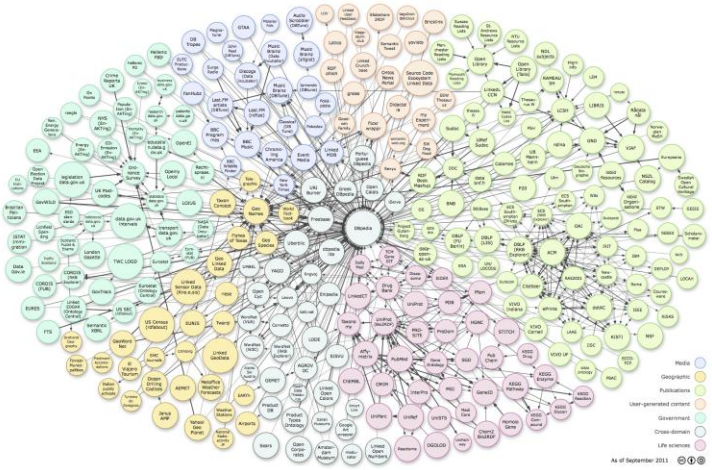
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Probability of patient P getting Alzheimer's disease within next 10 years?



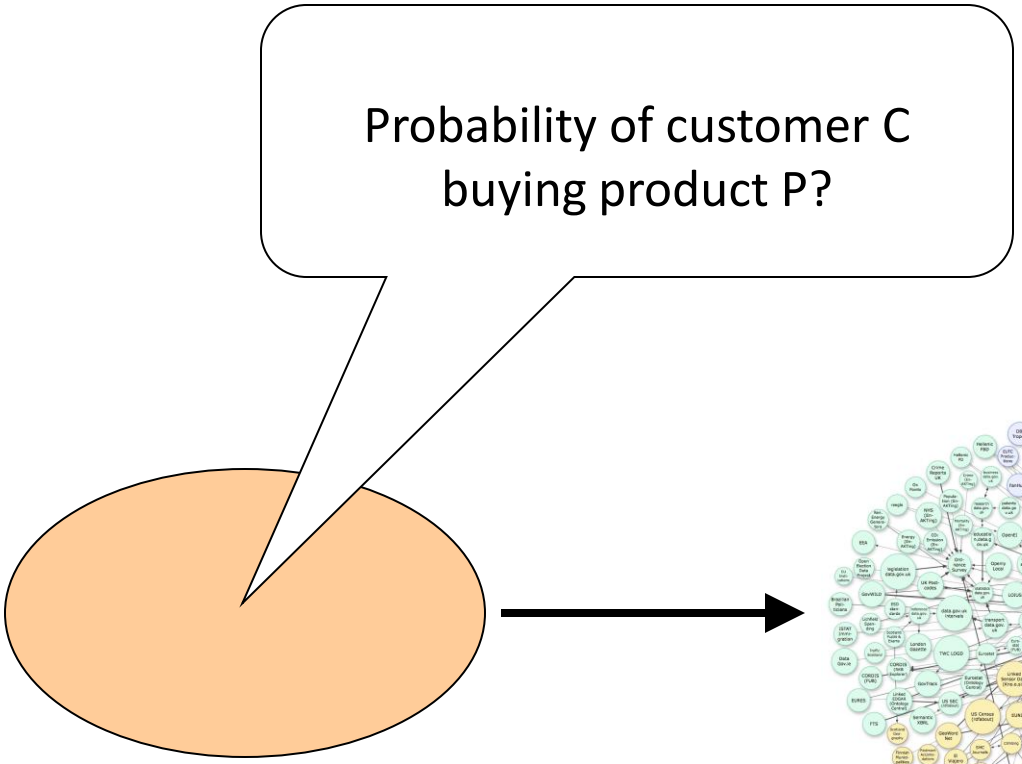
Patient Health Records

Disease Prevention

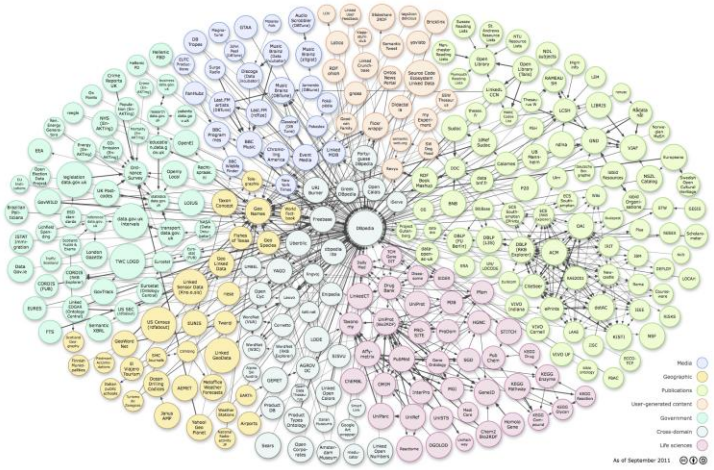


Linked Open Data cloud diagram,
by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

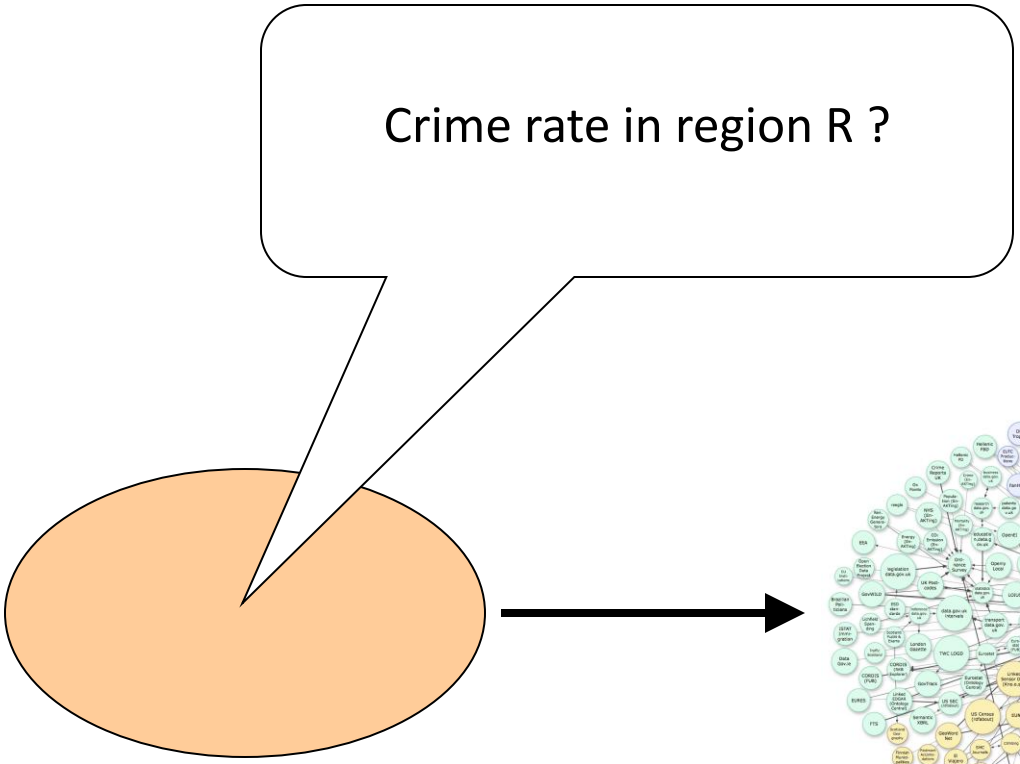
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.



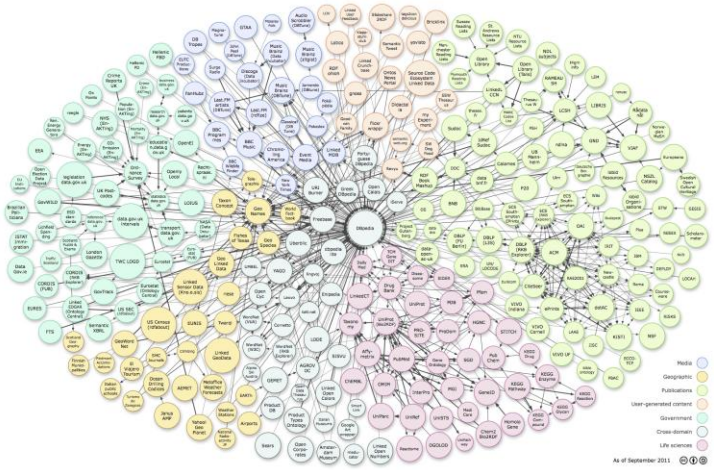
Product Recommendation



Linked Open Data cloud diagram,
by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/



City Planning



Linked Open Data cloud diagram,
by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Overview

We

- Show how to learn relational Bayesian classifiers from RDF data in a setting where the data can be accessed only through statistical queries against a SPARQL endpoint
- Demonstrate the advantages of the statistical query based approach over one that requires direct access to RDF data
- Establish the conditions under which the predictive model can be updated incrementally in response to changes in the underlying data
- Show how to cope with settings where the relevant data attributes to be used for building the predictive model are not known a priori
- Provide open source implementation of the resulting algorithms

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

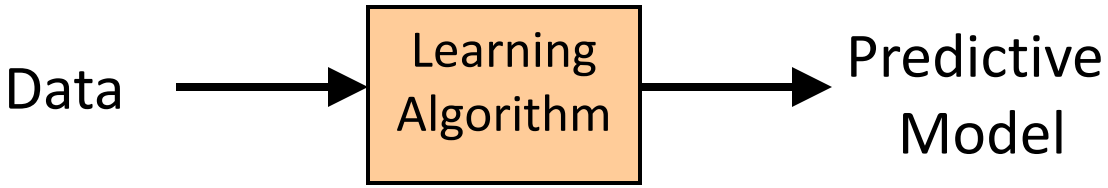
Outline

- Background and Motivation
- Statistical Query Based Approach to Learning Relational Bayesian Classifiers from RDF data
- Experiment 1: Communication Complexity
- Extensions of the basic approach to settings where
  - The RDF data store is updated over time
  - The attributes of interest are not known a priori
  - Experiment 2: Selective Attribute Crawling
- Conclusion and Future Work

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Motivation: Learning Predictive Models from RDF Data

- The growth in RDF data on the web offers unprecedented opportunities for
  - building predictive models e.g., classifiers from RDF data
  - using the resulting models to guide decisions in a broad range of application domains
- Machine learning offers one of the most effective approaches to building predictive models from data



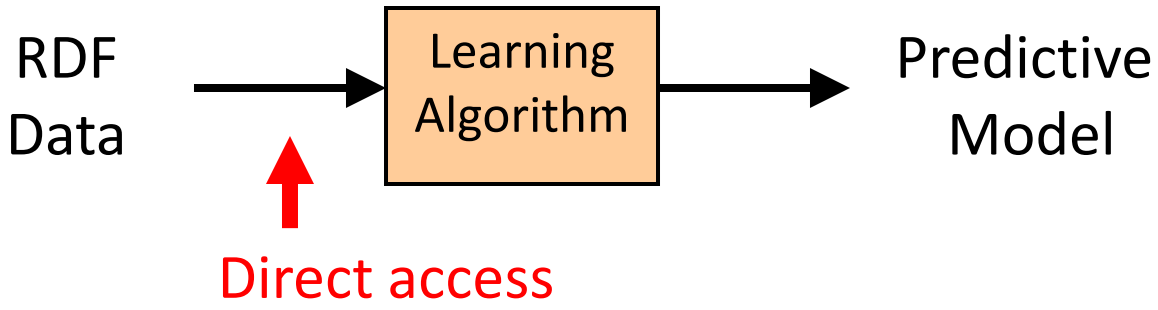
- Statistical relational learning [Getoor, 2007] offers a natural starting point for design of algorithms for learning predictive models from RDF data

Related Work: Learning Predictive Models from RDF Data

- Kiefer et al. [2008] extend SPARQL with additional constructs to support data mining to obtain SPARQL-ML
  - ✓ CREATE MINING MODEL statement for constructing a predictive model from RDF data
  - ✓ PREDICT statement for applying the model to make predictions
- Tresp et al. [2009] use matrix completion methods e.g., Latent Dirichlet Allocation (LDA) to make predictions from a Boolean Matrix representation of RDF data
- Bicer et al. [2011] combine kernel-based support vector machines with relational learning to obtain Relational Kernel machines defined over RDF data

Limitations of Existing Methods for Learning Predictive Models from RDF Data

- Key limitation: Current methods assume that the learning algorithm has direct access to RDF Data



Limitations of Existing Methods for Learning Predictive Models from RDF Data

- In practice:
  - RDF stores may not always provide data dumps (e.g. streaming social network data)
  - Access constraints may prevent access to RDF data (e.g. patient health records, Dr. Pentland's keynote, "Linked Closed Data" [@COLD workshop])
  - Bandwidth constraints may limit the ability to transfer data from a remote RDF store to the local site where the learning algorithm resides (e.g. sensor data, and possibly "Linked Sensor Data" [@SSN workshop])
  - Learning algorithms that assume in-memory access to data cannot handle RDF datasets that are too large to fit in memory
- Needed: Approaches for learning from RDF data without direct access to RDF Data in settings where the data can be accessed only through statistical queries (e.g. against a SPARQL endpoint)

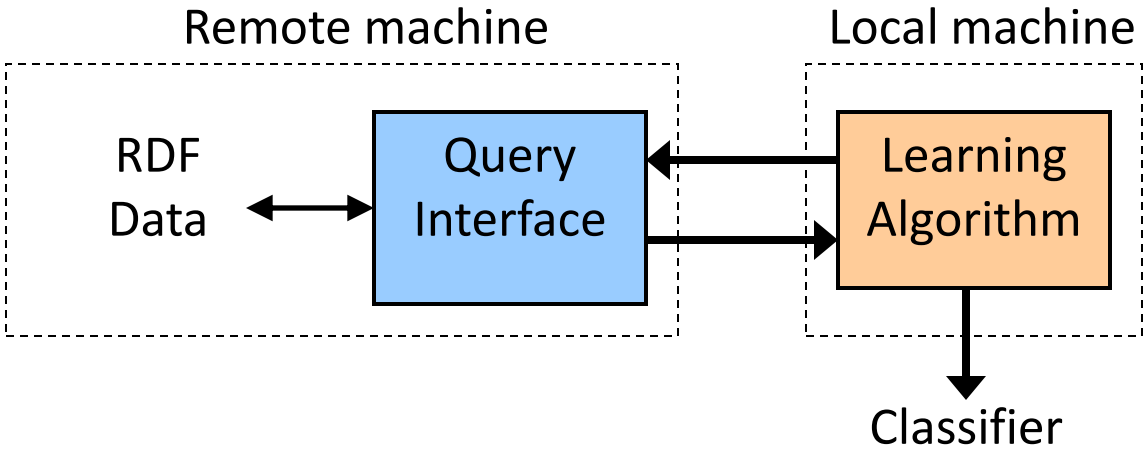
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Outline

- Background and Motivation
- Statistical Query Based Approach to Learning Relational Bayesian Classifiers from RDF data
- Experiment 1: Communication Complexity
- Extensions of the basic approach to settings where
  - The RDF data store is updated over time
  - The attributes of interest are not known a priori
  - Experiment 2: Selective Attribute Crawling
- Conclusion and Future Work

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Statistical Query Based Approach to Learning Classifiers from RDF Data



- Uses the statistical query based learning framework of Caragea et al. [2005]
- Decompose the learner into two components
  - Statistical query generation: poses statistical queries to a data source to acquire the information needed by the learner
  - Model construction: uses the answers to statistical queries to update or refine a partial model
- The learner interacts with the RDF data store only through queries posed against a SPARQL access point

Learning Relational Bayesian Classifiers from RDF Data

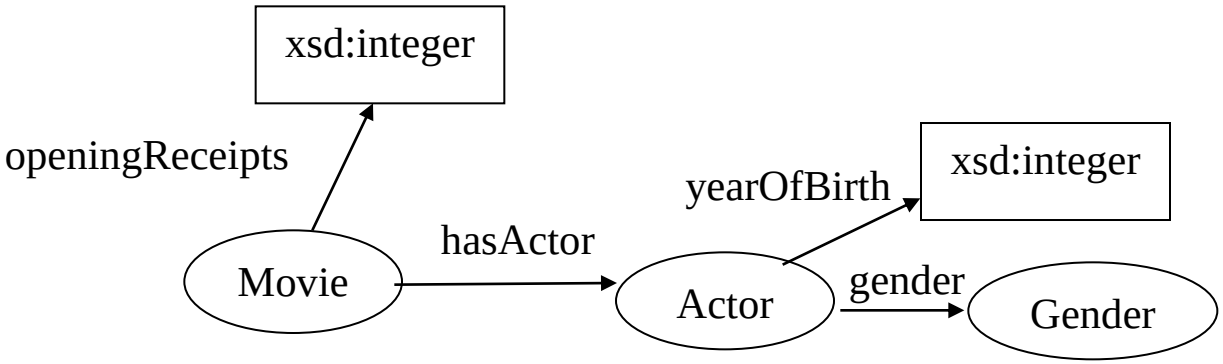
- A Relational Bayesian Classifier (RBC) [Getoor, 2007] is a relational variant of the Simple Bayesian Classifier (SBC)
- SBC
  - Each attribute in a data instance assumes a single value from the domain of the corresponding random variable
  - Assumes that data attributes are conditionally independent of class C
- RBC
  - Each data attribute takes a value that is a multi-set of elements chosen from the domains of the corresponding random variable
  - Assumes that each multi-set valued attribute is independent given the class

$$h_{RBC}(x) = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\, p(c) \prod_i p(\mathcal{L}_i(x) : \ c)$$

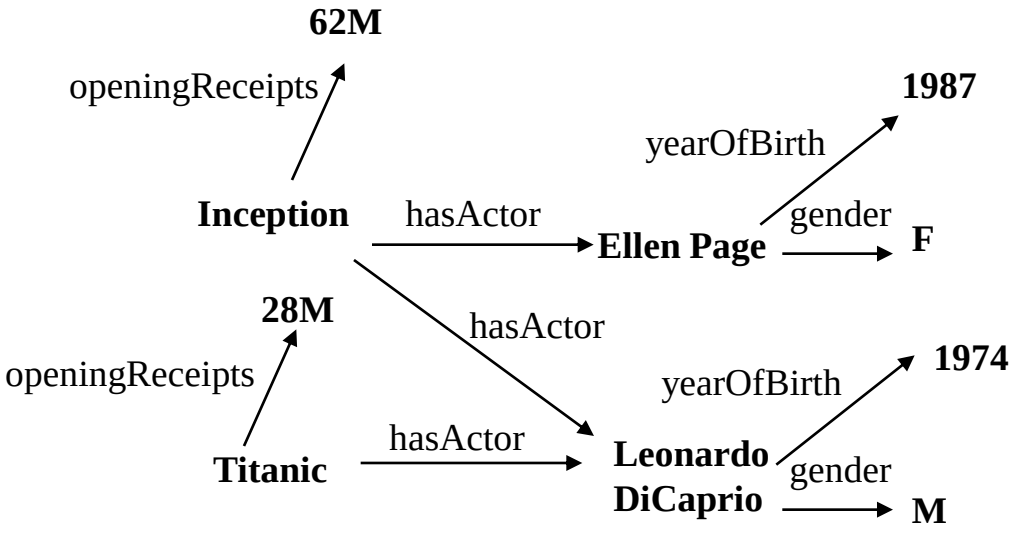
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

RBC Example

- RDF Schema:



- Goal: predict whether a movie receives more than $2M in its opening week
  - Target class: Movie
  - Target attribute: (openingReceipts)
  - Attribute 1: (hasActor, yearOfBirth)
  - Attribute 2: (hasActor, gender)



Flattening

Movie	openingReceipts	hasActor, yearOfBirth	hasActor, gender
Inception	62M	{1987,1974}	{F,M}
Titanic	28M	{1974}	{M}

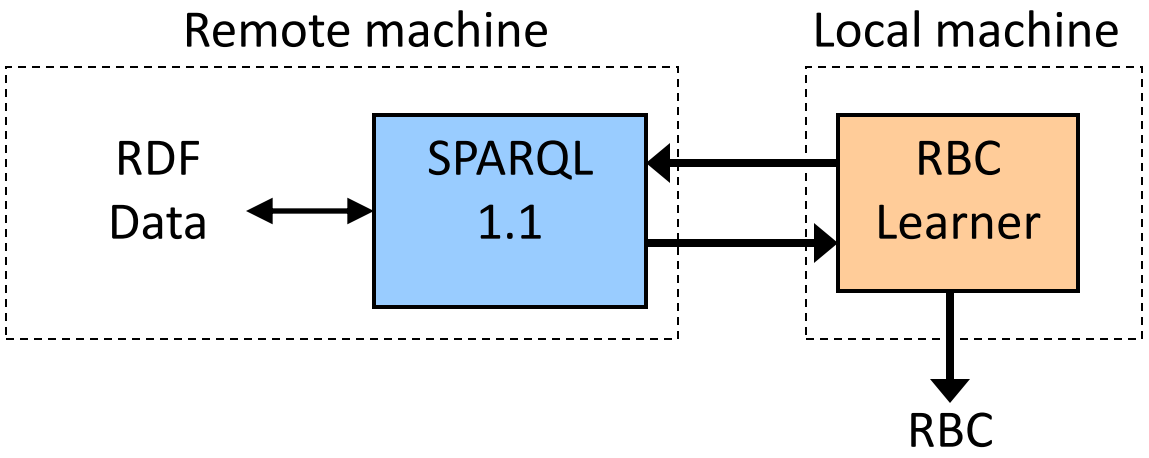
Counts

RBC

Counts via SPARQL Queries

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Learning Relational Bayesian Classifiers from Statistical Queries against RDF data



$$h_{RBC}(x) = \operatorname*{argmax}_{c \in \mathcal{C}} p(c) \prod_i p(\mathcal{L}_i(x) : \ c)$$

- Our implementation relies on the aggregate queries supported by SPARQL 1.1 to express the statistical queries posed by RBC Learner

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

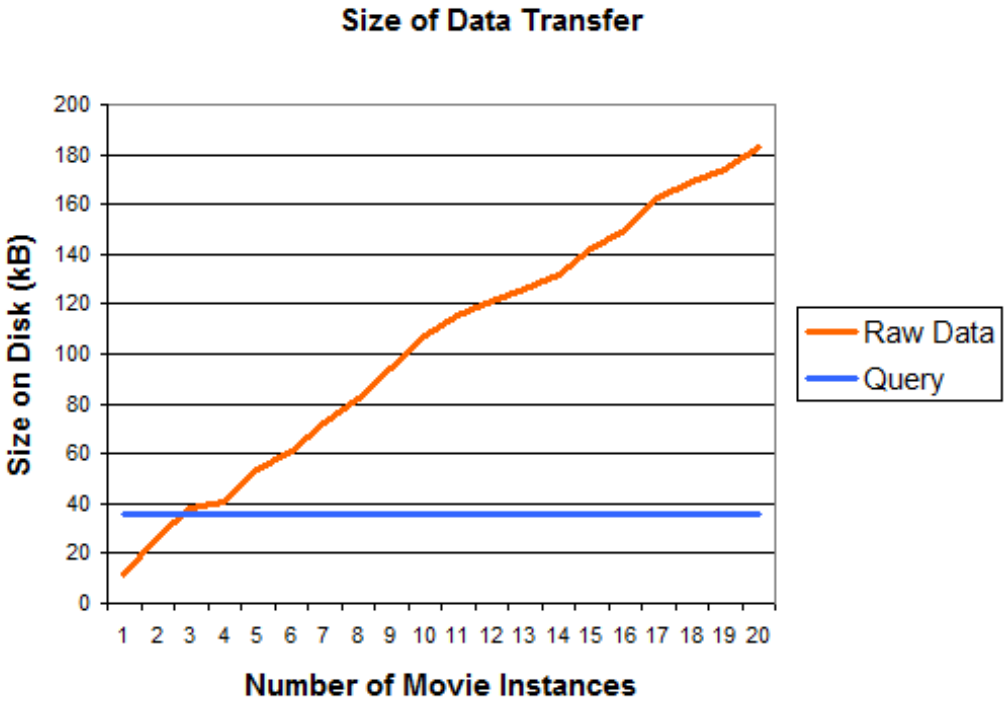
Outline

- Background and Motivation
- Statistical Query Based Approach to Learning Relational Bayesian Classifiers from RDF data
- Experiment 1: Communication Complexity
- Extensions of the basic approach to settings where
  - The RDF data store is updated over time
  - The attributes of interest are not known a priori
  - Experiment 2: Selective Attribute Crawling
- Conclusion and Future Work

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Communication Complexity: Statistical Query Based RBC Learner compared with RBC Learner with Direct Access to Data

- Data prepared from LinkedMDB and Freebase
- SPARQL queries and results logged in plain text format
- Raw data dump in RDF/XML (also compared with compressed dumps)
- The cost of retrieving the statistics needed for learning < < the cost of retrieving the entire dataset



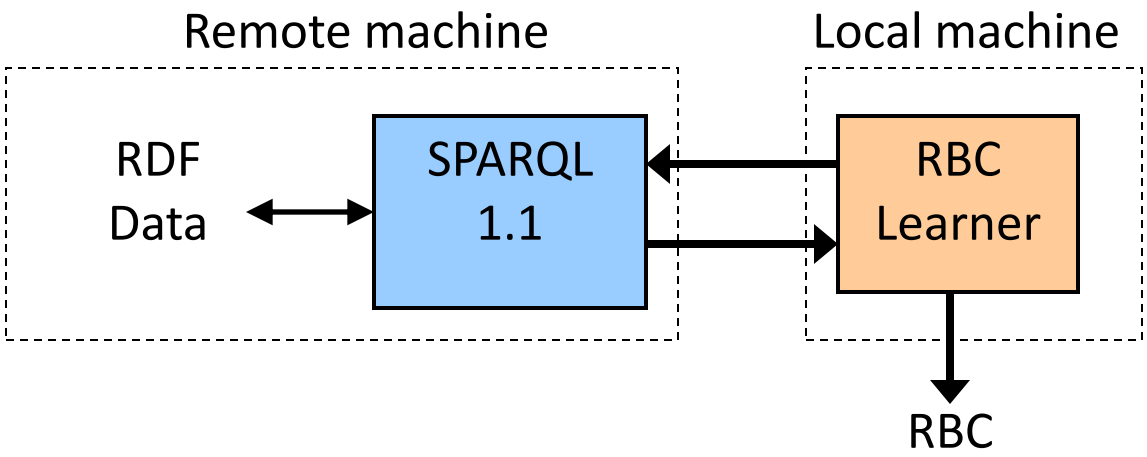
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Outline

- Background and Motivation
- Statistical Query Based Approach to Learning Relational Bayesian Classifiers from RDF data
- Experiment 1: Communication Complexity
- Extensions of the basic approach to settings where
  - The RDF data store is updated over time
  - The attributes of interest are not known a priori
  - Experiment 2: Selective Attribute Crawling
- Conclusion and Future Work

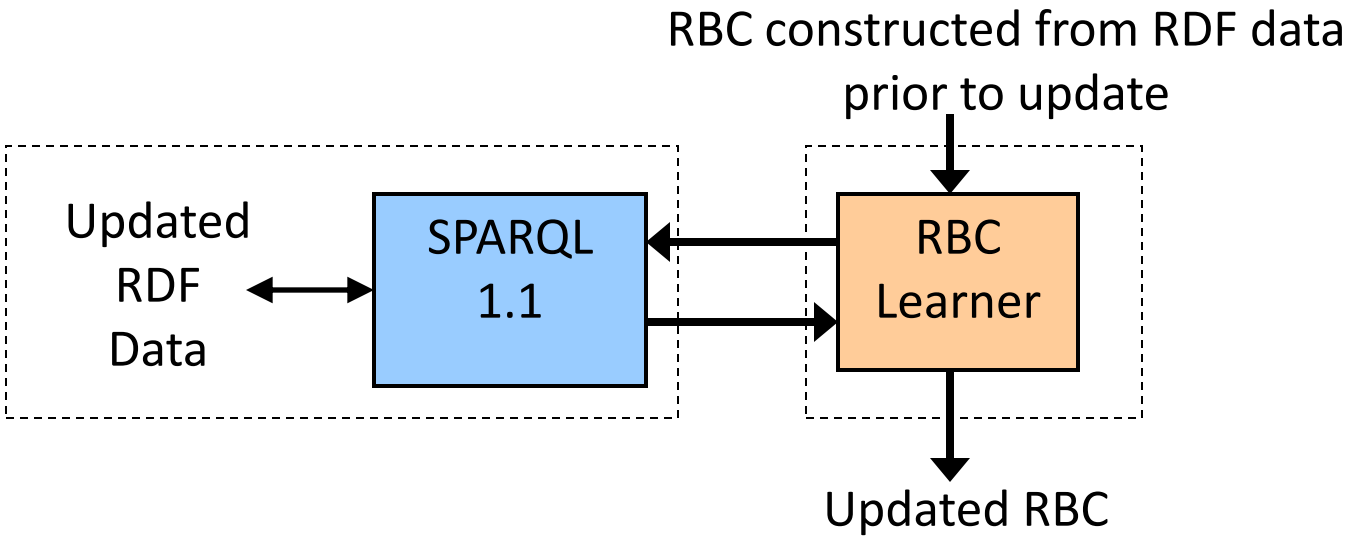
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Learning Relational Bayesian Classifiers from Statistical Queries against RDF data: Further Extensions



- Extensions of the basic approach to settings where
  - The RDF data store is updated over time
  - The attributes of interest are not known a priori

Updating Relational Bayesian Classifiers in Response to Updates to the Underlying RDF Store



- RBC classifiers are updatable (definition adapted from [Koul, 2010]) if the updates to the underlying RDF data store are clean
- An update is clean if any new tuples that are added as a result of update share objects with only the elements of the multi-sets that make up the attribute values used to construct the RBC before the update (see paper for precise definition)

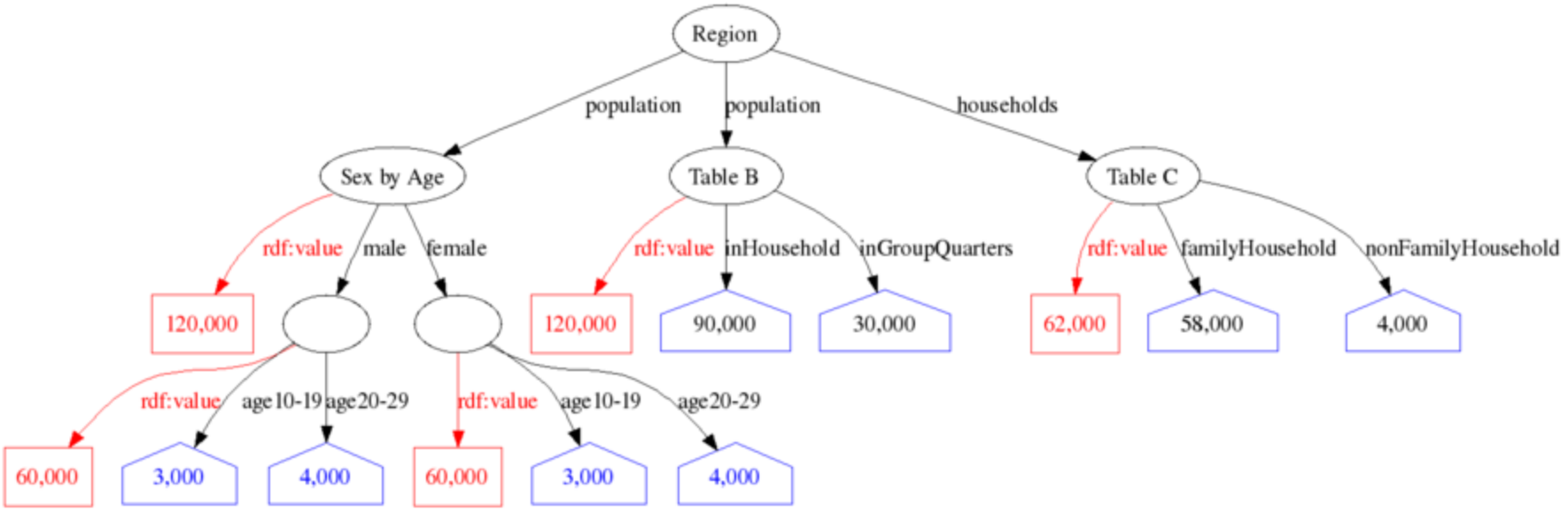
Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Building Relational Bayesian Classifiers when the Attributes of Interest are not known A priori

- In order to pose statistical queries, the learner needs to know the attributes of RDF data (i.e., the RDF schema)
- If the attributes are not known a priori, we selectively crawl for attributes that are predictive of the class label using information gain to score candidate attributes

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

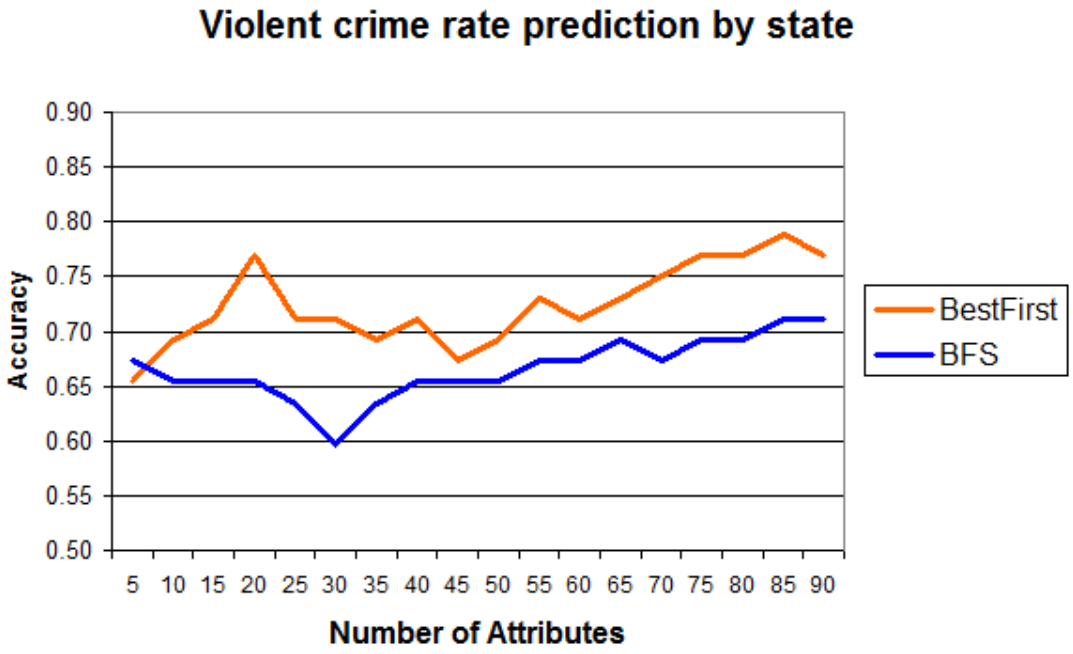
US Census RDF schema (by Tauberer)



J. Tauberer. The 2000 U.S. census: 1 billion RDF triples. http://www.rdfabout.com/demo/census/

Selective Attribute Crawling Experiment: Census

- Combined with Data.gov violent crime rate dataset
- Predict: A state's violent crime rate is over 400 per 100,000 population?
- Cross validation of 52 US states into 13 folds



Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Top 6 Attributes Selected by BestFirst Strategy

- ./families
- ./families/type
- ./households
- ./households/1personHousehold/femaleHouseholder
- ./households/2OrMorePersonHousehold/familyHouseholds/marriedcoupleFamily/noOwnChildrenUnder18Years
- ./households/2OrMorePersonHousehold/familyHouseholds/otherFamily/femaleHouseholder_NoHusbandPresent

Open Source Implementation

- Google Code - http://code.google.com/p/induslearningframework/
- Implemented as part of INDUS [Koul2008a], an open source system that learns predictive models from remote data source using sufficient statistics

Outline

- Background and Motivation
- Statistical Query Based Approach to Learning Relational Bayesian Classifiers from RDF data
- Experiment 1: Communication Complexity
- Extensions of the basic approach to settings where
  - The RDF data store is updated over time
  - The attributes of interest are not known a priori
  - Experiment 2: Selective Attribute Crawling
- Conclusion and Future Work

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Contributions

- Statistical Query Based Approach to Learning Relational Bayesian Classifiers from RDF data
  - ✓ Applicable to settings where data can be accessed only through statistical queries against a SPARQL endpoint
  - ✓ Far more bandwidth efficient than the alternatives that assume direct access to RDF data
  - ✓ Generalizable to other statistical relational learning algorithms
- Extensions to settings where
  - ✓ The RDF data store is updated over time
  - ✓ The attributes of interest are not known a priori
- Open source implementation of the algorithms

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Future Work

- Statistical Query based variants of other relational learning algorithms for building predictive models from RDF data
- Extensions of the approach to learn predictive models from multiple distributed RDF data stores
  (cf. RDF Query – Multiple Sources session @ Maritim)
- Open source implementation of other algorithms
- Real-world applications

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Thank You!

- Questions?

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

SPARQL Aggregate Queries

```
SELECT COUNT(*) WHERE {
  ?x rdf:type <C> .
  ?x <c1> ?c1 .   ...   ?cm-1 <cm> c .
  ?x <p1> ?v1 .   ...   ?vj-1 <pj> a .
}
```

```
SELECT COUNT(*) WHERE {
  { SELECT (agg(?vj) AS ?aggvalue) WHERE {
      ?x rdf:type <T> .
      ?x <c1> ?c1 .   ...   ?cm-1 <cm> c .
      OPTIONAL { ?x <p1> ?v1 .   ...   ?vj-1 <pj> ?vj . }
    } GROUP BY ?x
  } FILTER(?aggvalue >= vl && ?aggvalue <= vh)
}
```

Harris Lin, Neeraj Koul, and Vasant Honavar. ISWC 2011. Research supported in part by NSF grant IIS 0711356.

Updatable Definition

Definition 10 (Updatable Model [19]) *Given datasets D_1 and D_2 such that $D_1 \subseteq D_2$, we say that a primitive query θ is updatable iff we can specify functions f and g such that:*

1. $\theta(D_2) = f(\theta(D_2 - D_1), \theta(D_1))$
2. $\theta(D_1) = g(\theta(D_2), \theta(D_2 - D_1))$

We say that the predictive model constructed using L is updatable iff all primitive queries required over the dataset D to build $L(D)$ are updatable.