Towards semantics-enabled infrastructure for knowledge acquisition from distributed data

Vasant Honavar and Doina Caragea
Artificial Intelligence Research Laboratory
Department of Computer Science
Bioinformatics and Computational Biology Graduate Program
Center for Computational Intelligence, Learning, & Discovery
Iowa State University
honavar@cs.iastate.edu
www.cs.iastate.edu/~honavar/

In collaboration with Jun Zhang (Ph.D., 2005), Jie Bao (Ph.D., 2007)

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Outline

- Background and motivation
- Learning from data revisited
- Learning predictive models from distributed data
- Learning predictive models from semantically heterogeneous data
- Learning predictive models from partially specified data
- Current Status and Summary of Results

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Representative Application: Gene Annotation

Discovering potential errors in gene annotation using machine learning (Andorf, Dobbs, and Honavar, BMC Bioinformatics, 2007)

- Train on human kinases, and test on mouse kinases – surprisingly poor accuracy!
- Nearly 95 percent of the GO annotations returned by AmiGO for a set of mouse protein kinases are inconsistent with the annotations of their human homologs and are likely, erroneous
- The mouse annotations came from Okazaki et al, Nature, 420, 563-573, 2002
- They were propagated to MGI through the Fantom2 (Functional Annotation of Mouse) Database and from MGI to AmiGO
- 136 rat protein kinase annotations retrieved using AmiGO had functions assigned based on one of the 201 potentially incorrectly annotated mouse proteins
- Postscript: Erroneous mouse annotations were traced to a bug in the annotation script and have since been corrected by MGI

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

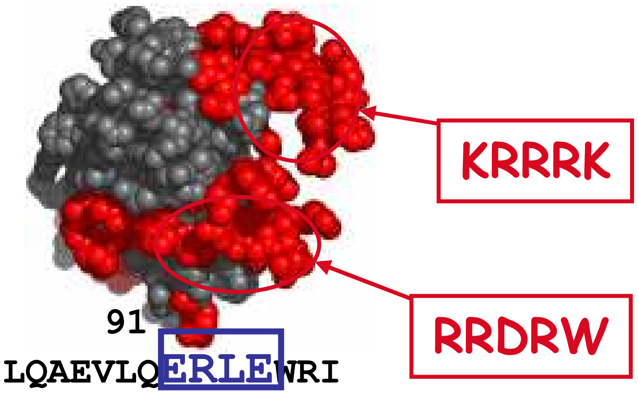
Representative Application - Predicting Protein-RNA Binding Sites

PREDICTED:

Structure +

Protein binding residues +

RNA binding residues



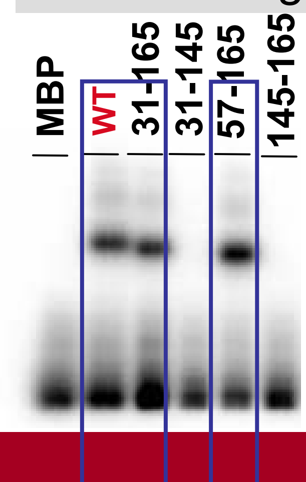

```
  41        51       61        71        81         91
GPLESDQWCRVLRQSLPEEKISSQTCI ARRHLGPGPTQHTPSRRDRWIREQILQAEVLQERLEWRI
++++++++      ++         +++++++++++++++
                           +++++++++++++++

       131        141        151        161
QRGDFSAWGDYQQAQERRWGEQSSPRVLRPGDSKRRRKHL
++++++++++              ++   +++   ++++++
       +              ++++++++++++++++++++
```

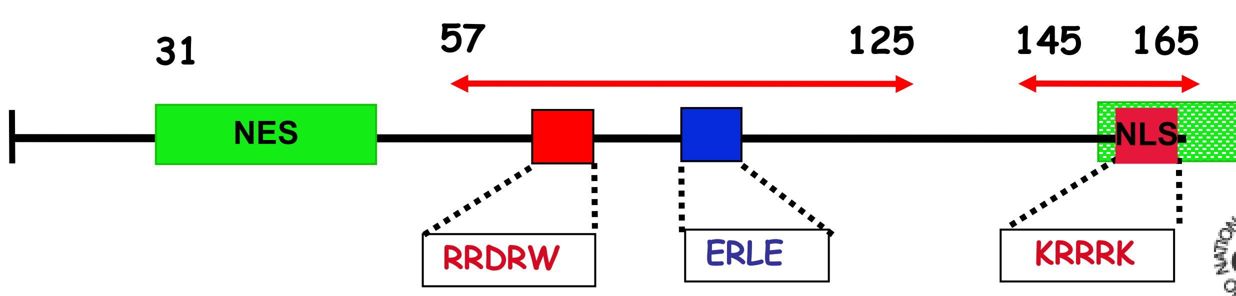
VALIDATED:

Protein binding residues

RNA binding residues



EIAV Rev: Predictions vs Experiments



Terribilini, M., Lee. J-H., Yan, C., Carpenter, S., Jernigan, R., Honavar, V. and Dobbs, D. (2006)

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Background

Data revolution

- Bioinformatics
  - Over 200 data repositories of interest to molecular biologists alone (Discala, 2000)
- Environmental Informatics
- Enterprise Informatics
- Medical Informatics
- Social Informatics ...

Information processing revolution: Algorithms as theories

- Computation: Biology::Calculus:Physics

Connectivity revolution (Internet and the web)

Integration revolution

- Need to understand the elephant as opposed to examining the trunk, the tail, etc.

Needed – infrastructure to support collaborative, integrative analysis of data

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Predictive models from Data

- Supporting collaborative, integrative analysis of data across geographic, organizational, and disciplinary barriers requires coming to terms with:
  - Large, distributed autonomous data sources
    - Memory, bandwidth, and computing limitations
    - Access and privacy constraints
  - Differences in data semantics
    - Same term, different meaning
    - Different terms, same meaning
    - Different domains of values for semantically equivalent attributes
    - Different measurement units, different levels of abstraction
- Can we learn without centralized access to data?
- Can we learn in the presence of semantic gaps between user and data sources?
- How do the results compare with the centralized setting?

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

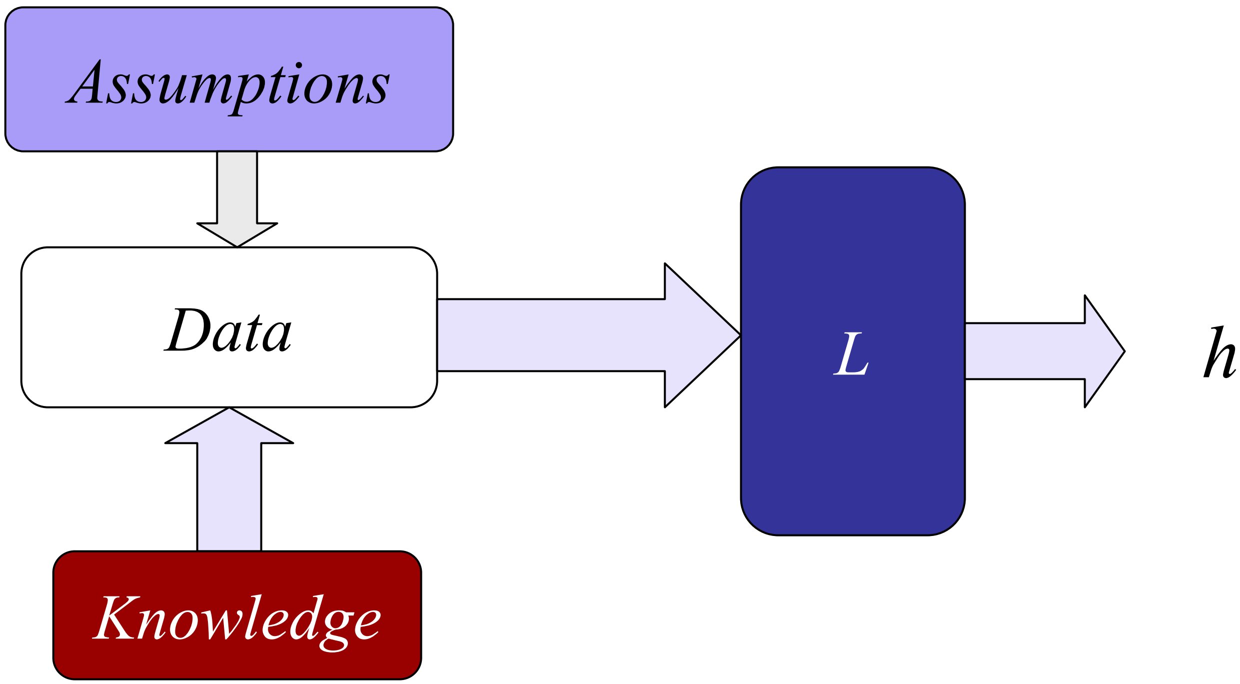
Outline

- Background and motivation
- Learning from data revisited
- Learning predictive models from distributed data
- Learning predictive models from semantically heterogeneous data
- Learning predictive models from partially specified data
- Current Status and Summary of Results

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

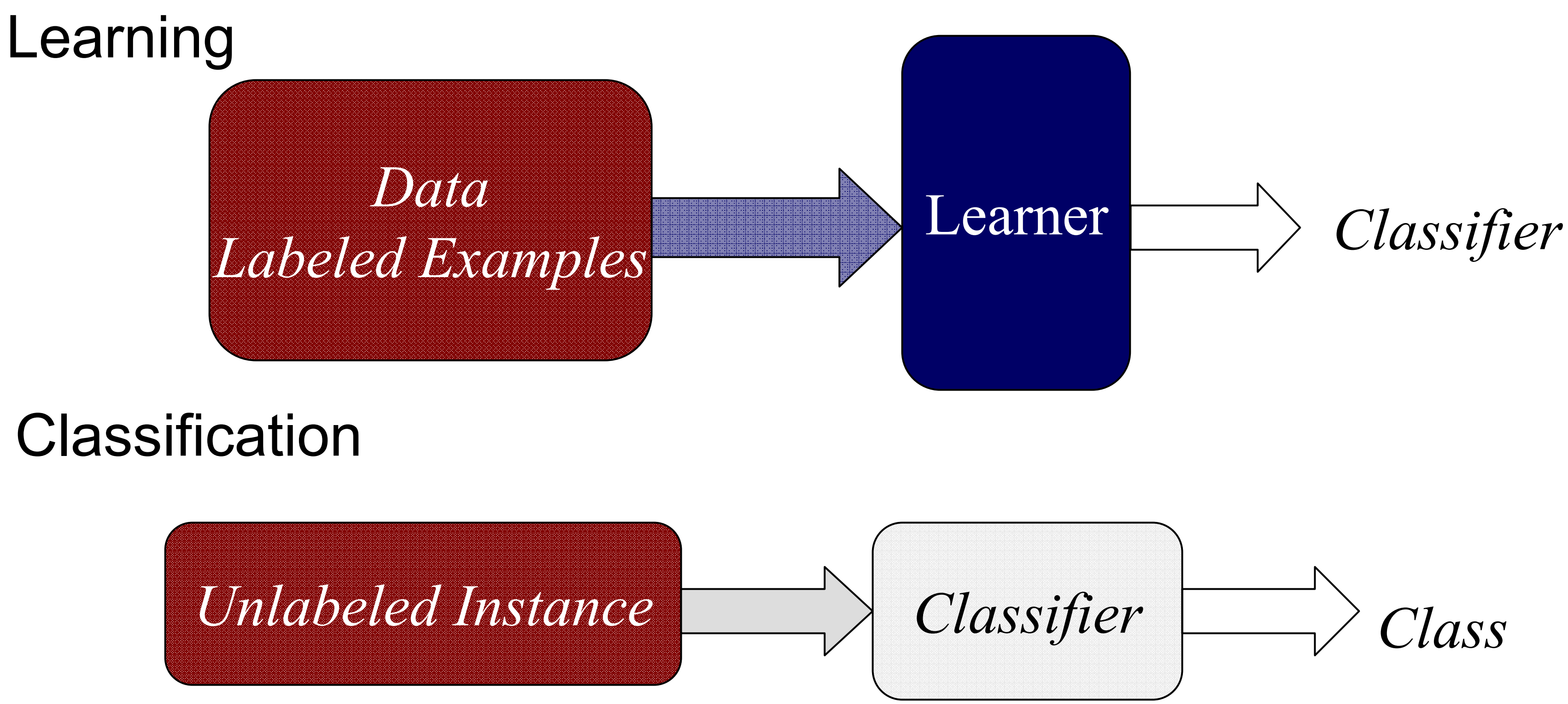
Acquiring knowledge from data

Most machine learning algorithms assume centralized access to a semantically homogeneous data



Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning Classifiers from Data



Standard learning algorithms assume centralized access to data

Can we do without direct access to data?

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

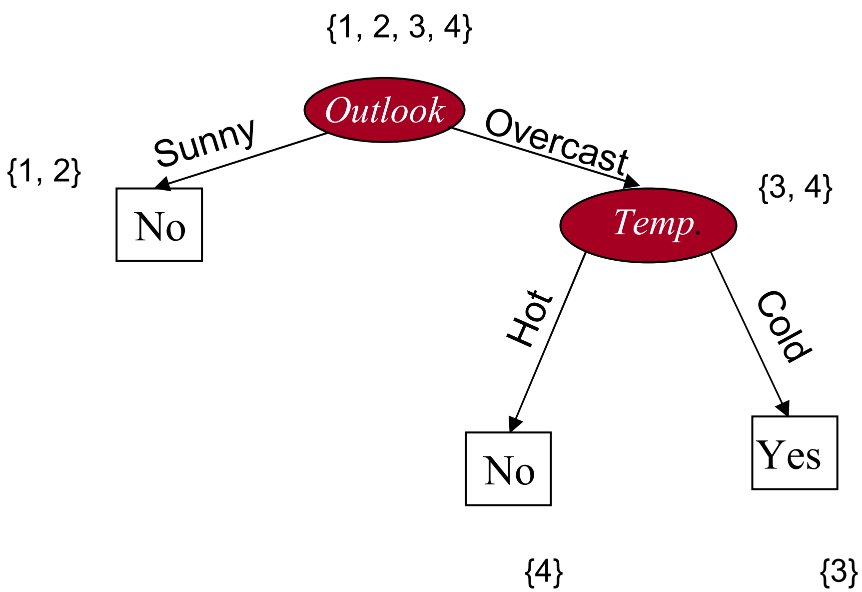
Example: Learning decision tree classifiers

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Overcast	Cold	Normal	Weak	No

Day	Outlook	Temp	Humid.	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No

Day	Outlook	Temp	Humid.	Wind	Play
3	Overcast	Hot	High	Weak	Yes
4	Overcast	Cold	Normal	Strong	No

…
…



Entropy

$$H(D) = -\sum_{i \in Classes} \frac{|D_i|}{|D|} \cdot \log_2 \left(\frac{|D_i|}{|D|} \right)$$

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Example: Learning decision tree classifiers

- Decision tree is constructed by recursively (and greedily) choosing the attribute that provides the greatest estimated information about the class label
- What information do we need to choose a split at each step?
  - Information gain
  - Estimated probability distribution resulting from each candidate split
  - Proportion of instances of each class along each branch of each candidate split

- Key observation: If we have the relevant counts, we have no need for the data!

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Example: Learning decision tree classifiers

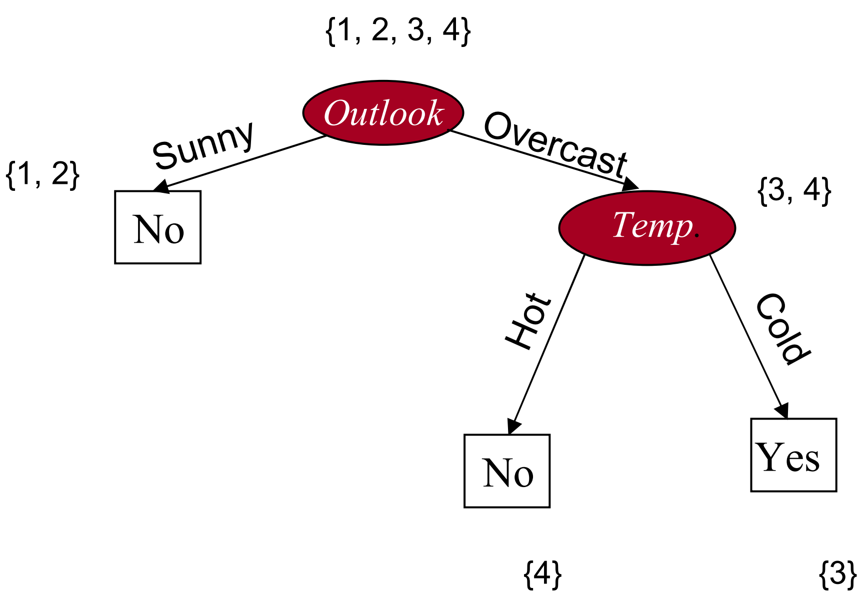
Day	Outlook	Temp.	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Overcast	Cold	Normal	Weak	No

Day	Outlook	Temp	Humid.	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No

Day	Outlook	Temp	Humid.	Wind	Play
3	Overcast	Hot	High	Weak	Yes
4	Overcast	Cold	Normal	Stron g	No

…

…

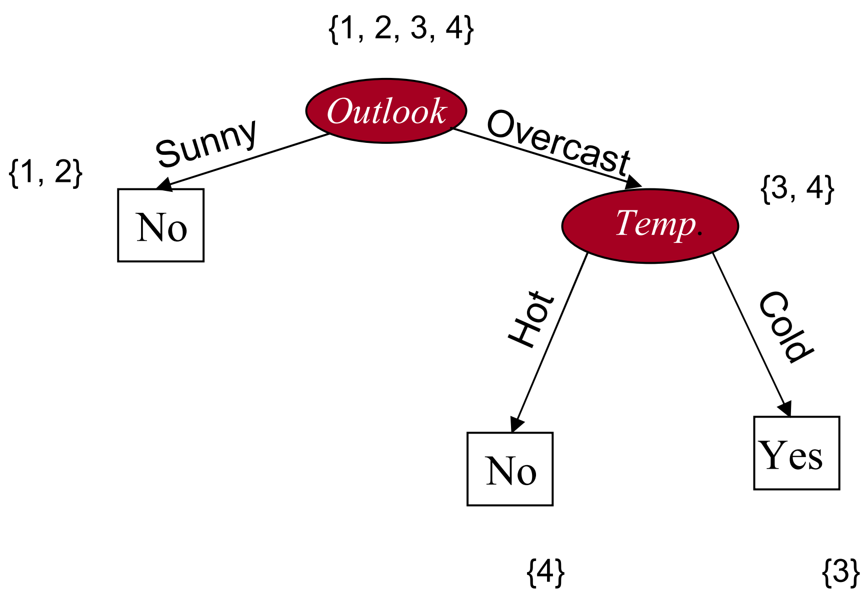


Entropy

$$H(D) = -\sum_{i \in Classes} \frac{|D_i|}{|D|} \cdot \log_2 \left(\frac{|D_i|}{|D|} \right)$$

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Sufficient statistics for refining a partially constructed decision tree



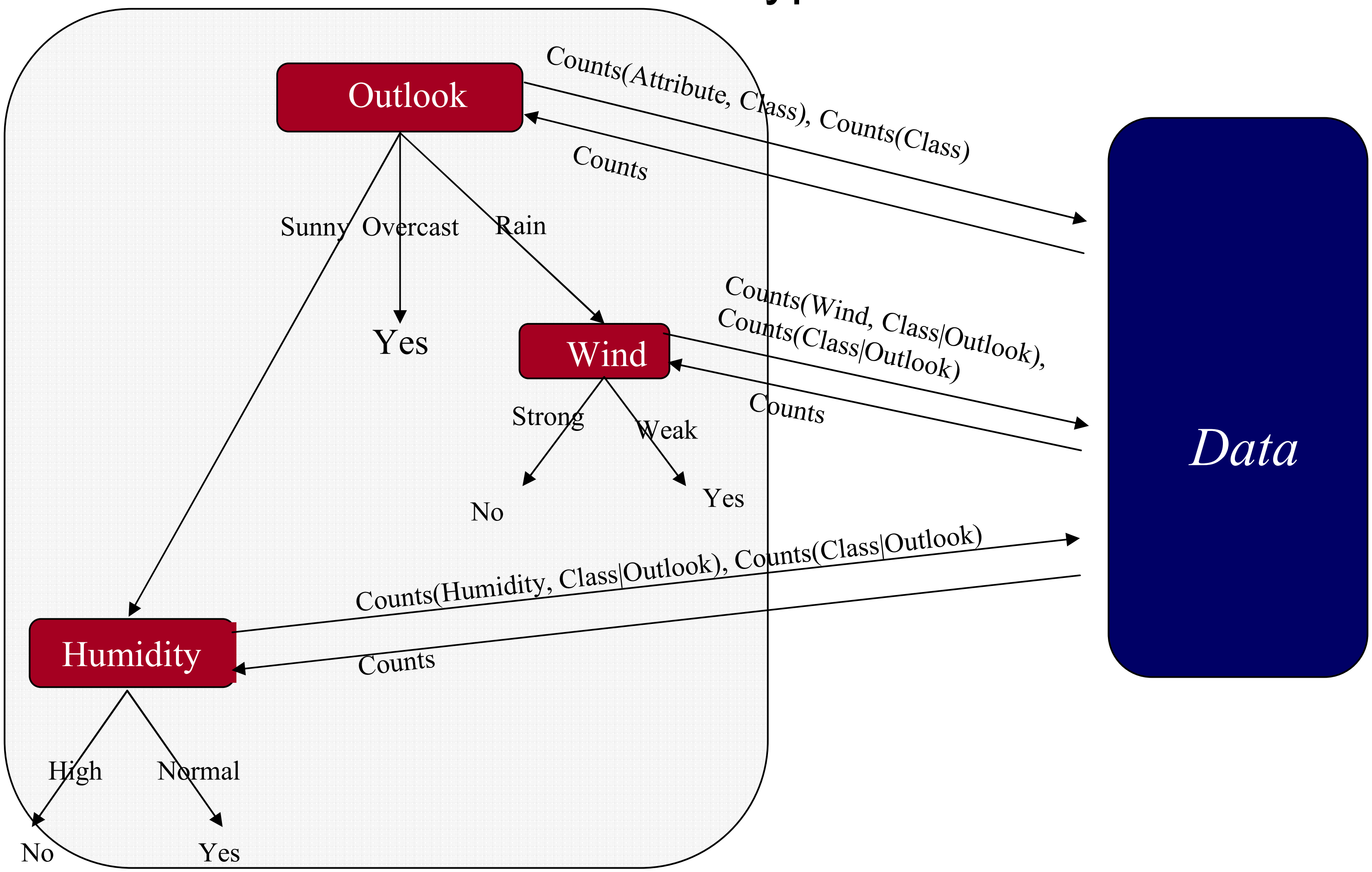
Entropy

$$H(D) = -\sum_{i \in Classes} \frac{|D_i|}{|D|} \cdot \log_2\left(\frac{|D_i|}{|D|}\right)$$

Sufficient statistics for refining a partially constructed decision tree

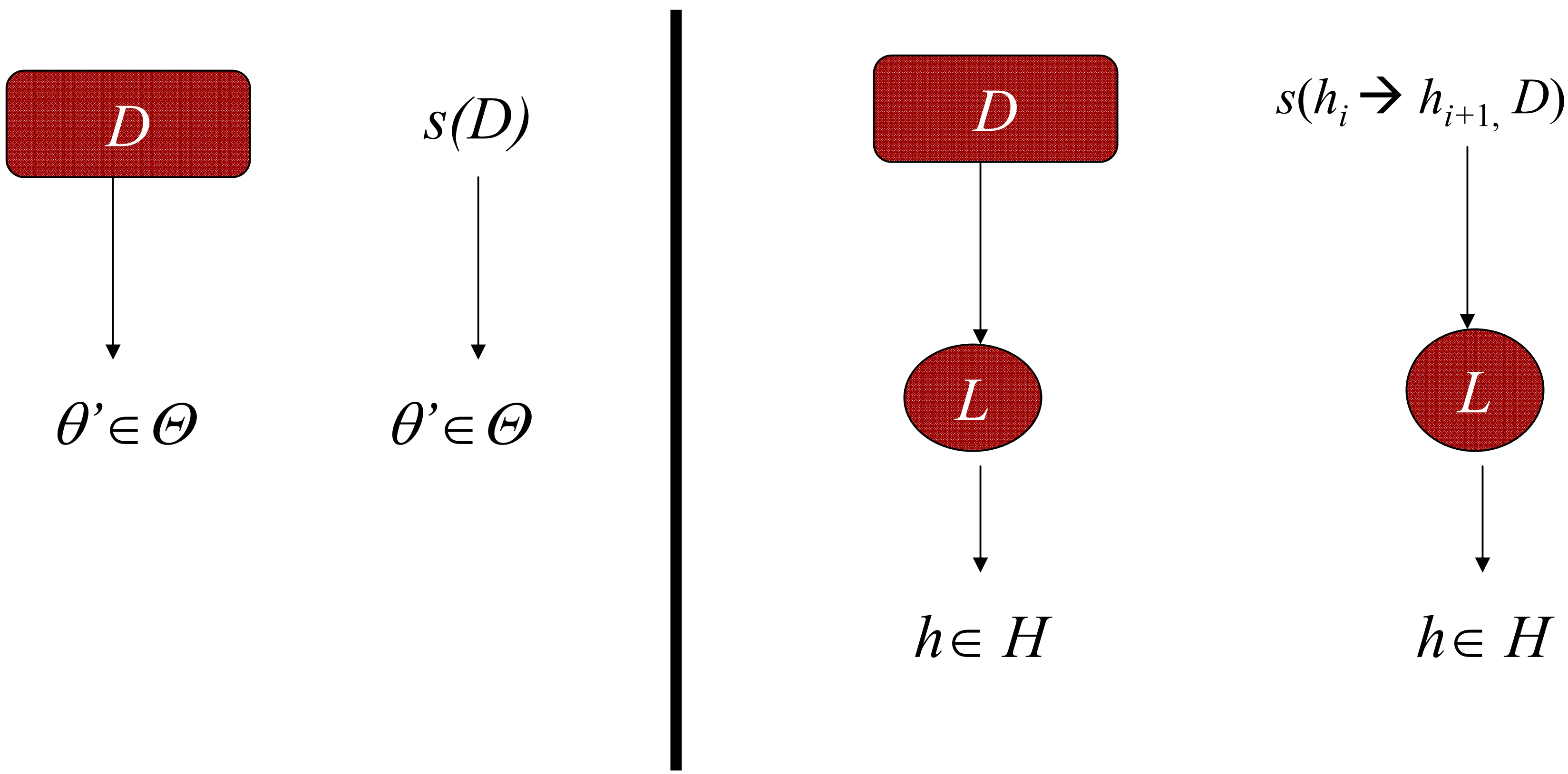
$$count(attribute\ value, class|path)$$
$$count(class|path)$$

Decision Tree Learning = Answering Count Queries + Hypothesis refinement



Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Sufficient statistics for learning: Analogy with statistical parameter estimation



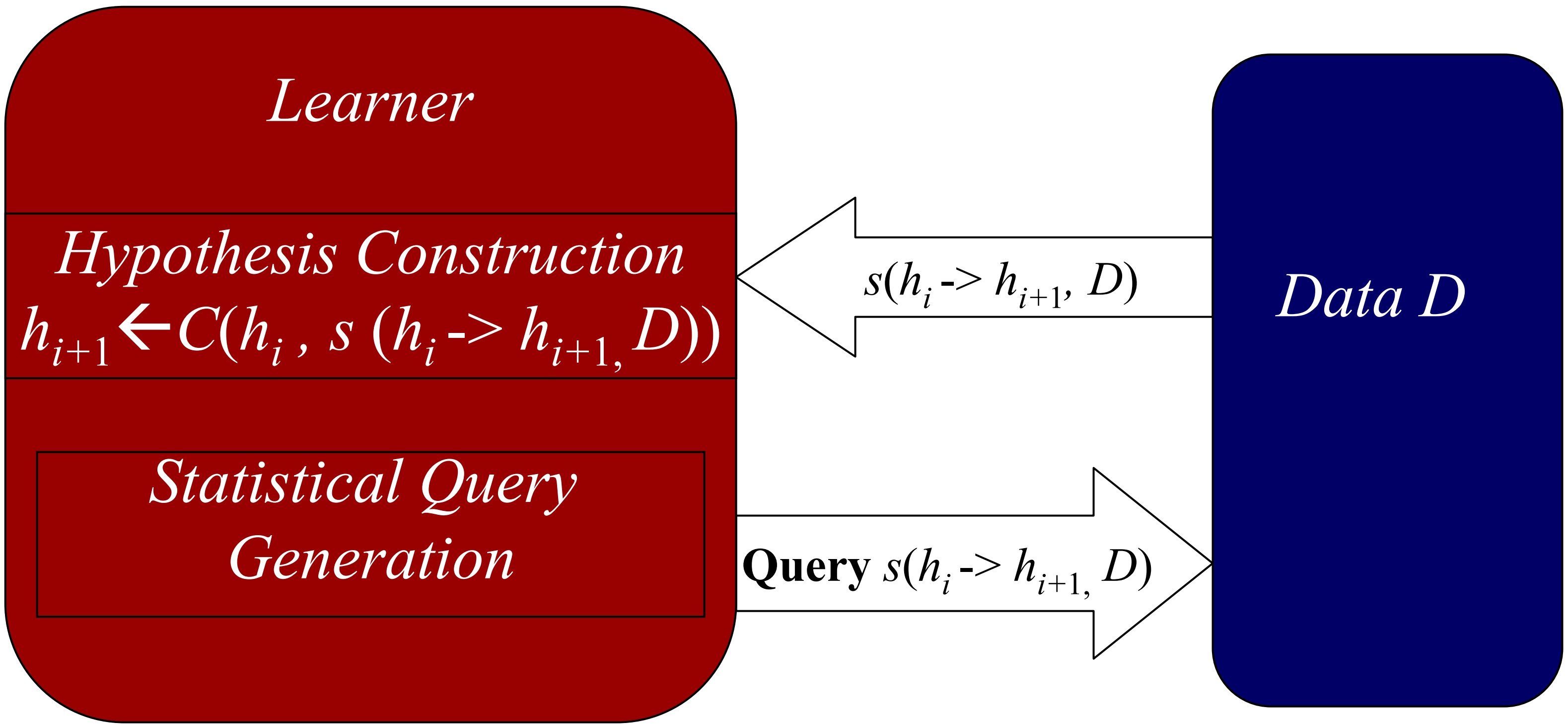
Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Sufficient statistics for learning a hypothesis from data

- It helps to break down the computation of $s_L(D,h)$ into smaller steps
  - queries to data D
  - computation on the results of the queries
- Generalizes the classical sufficient statistics by interleaving computation and queries against data
- Basic operations
  - Refinement
  - Composition

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning from Data Reexamined



Learning = Sufficient statistics Extraction
+ Hypothesis Construction

[Caragea, Silvescu, and Honavar, 2004]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning from Data Reexamined

Designing algorithms for learning from data reduces to

- Identifying of minimal or near minimal sufficient statistics for different classes of learning algorithms
- Designing procedures for obtaining the relevant sufficient statistics or their efficient approximations

Leading to

- Separation of concerns between hypothesis construction (through successive refinement and composition operations) and statistical query answering

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Outline

- Background and motivation
- Learning from data revisited
- Learning predictive models from distributed data
- Learning predictive models from semantically heterogeneous data
- Learning predictive models from partially specified data
- Current Status and Summary of Results

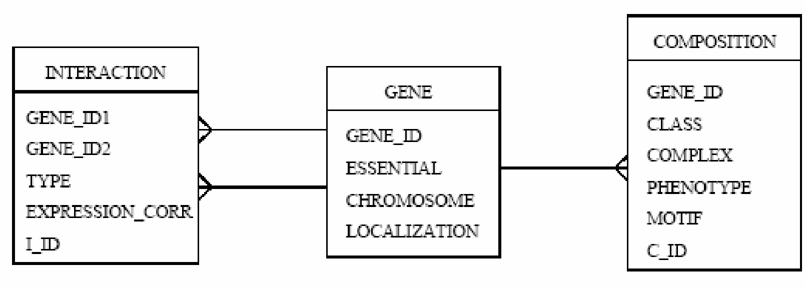
Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning Classifiers from Distributed Data

Learning from distributed data requires learning from dataset fragments without gathering all of the data in a central location

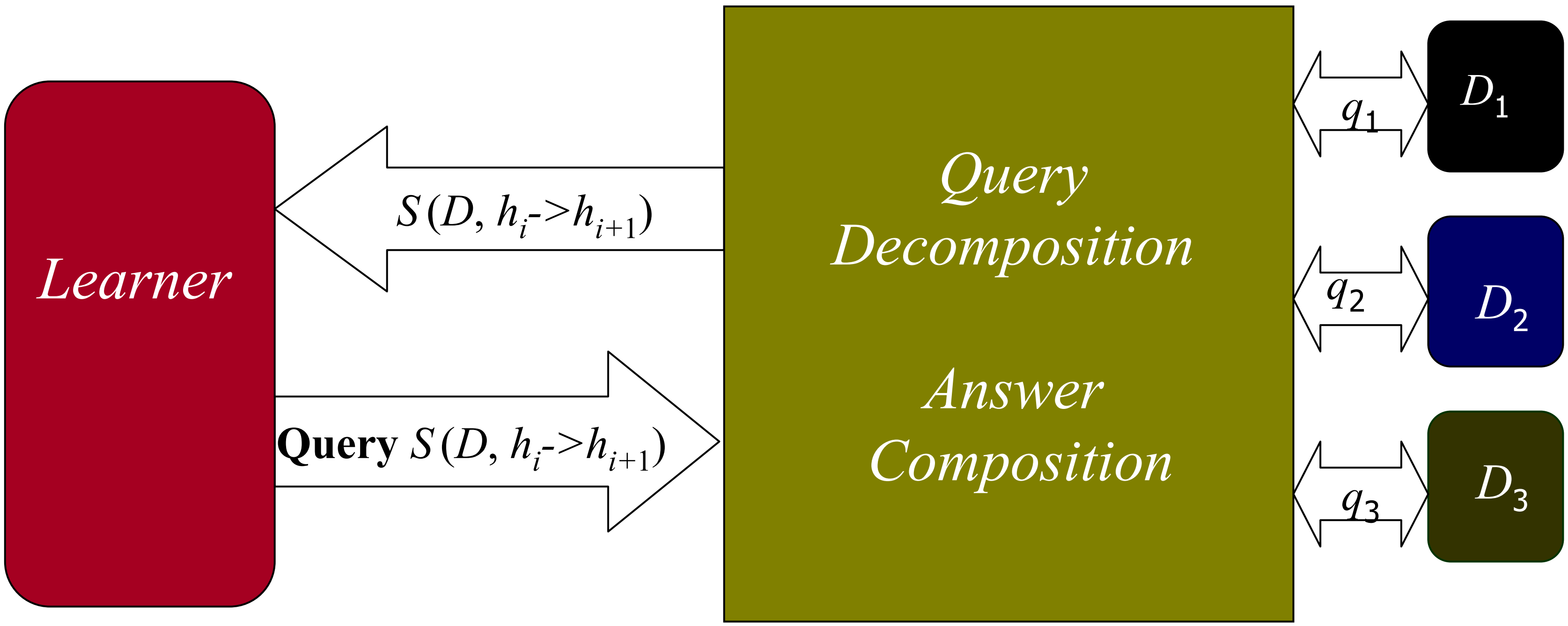
Assuming that the data set is represented in tabular form, data fragmentation can be

- horizontal
- vertical
- or more general (e.g. multi-relational)



Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning from distributed data



Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning from Distributed Data

- Learning classifiers from distributed data reduces to statistical query answering from distributed data
- A sound and complete procedure for answering the desired class of statistical queries from distributed data under
  - Different types of data fragmentation
  - Different constraints on access and query capabilities
  - Different bandwidth and resource constraints

[Caragea, Silvescu, and Honavar, 2004, Caragea et al., 2005]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

How can we evaluate algorithms for learning from distributed data?

Compare with their batch counterparts

- Exactness – guarantee that the learned hypothesis is the same as or equivalent to that obtained by the batch counterpart
- Approximation – guarantee that the learned hypothesis is an approximation (in a quantifiable sense) of the hypothesis obtained in the batch setting
- Communication, memory, and processing requirements

[Caragea, Silvescu, and Honavar., 2003, 2004]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Some Results on Learning from Distributed Data

- Provably exact algorithms for learning decision trees, SVM, Naïve Bayes, Neural Network, and Bayesian network classifiers from distributed data
- Positive and negative results concerning efficiency (bandwith, memory, computation) of learning from distributed data

[Caragea, Silvescu, and Honavar, 2004, Honavar and Caragea, 2008]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Outline

- Background and motivation
- Learning from data revisited
- Learning classifiers from distributed data
- Learning classifiers from semantically heterogeneous data
- Learning Classifier from partially specified data
- Current Status and Summary of Results

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Semantically heterogeneous data

Different schema, different data semantics

D_1

Day	Temperature (C)	Wind Speed (km/h)	Outlook
1	20	16	Cloudy
2	10	34	Sunny
3	17	25	Rainy

D_2

Day	Temp (F)	Wind (mph)	Precipitation
4	3	24	Rain
5	-2	50	Light Rain
6	0	34	No Prec

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Making Data Sources Self Describing

Exposing the schema – structure of data
Specification of the attributes of the data

D_1	Day: day	Temperature: deg C	Wind Speed: kmh	Outlook: outlook
D_2	Day: day	Temp: deg F	Wind: mph	Precipitation: prec

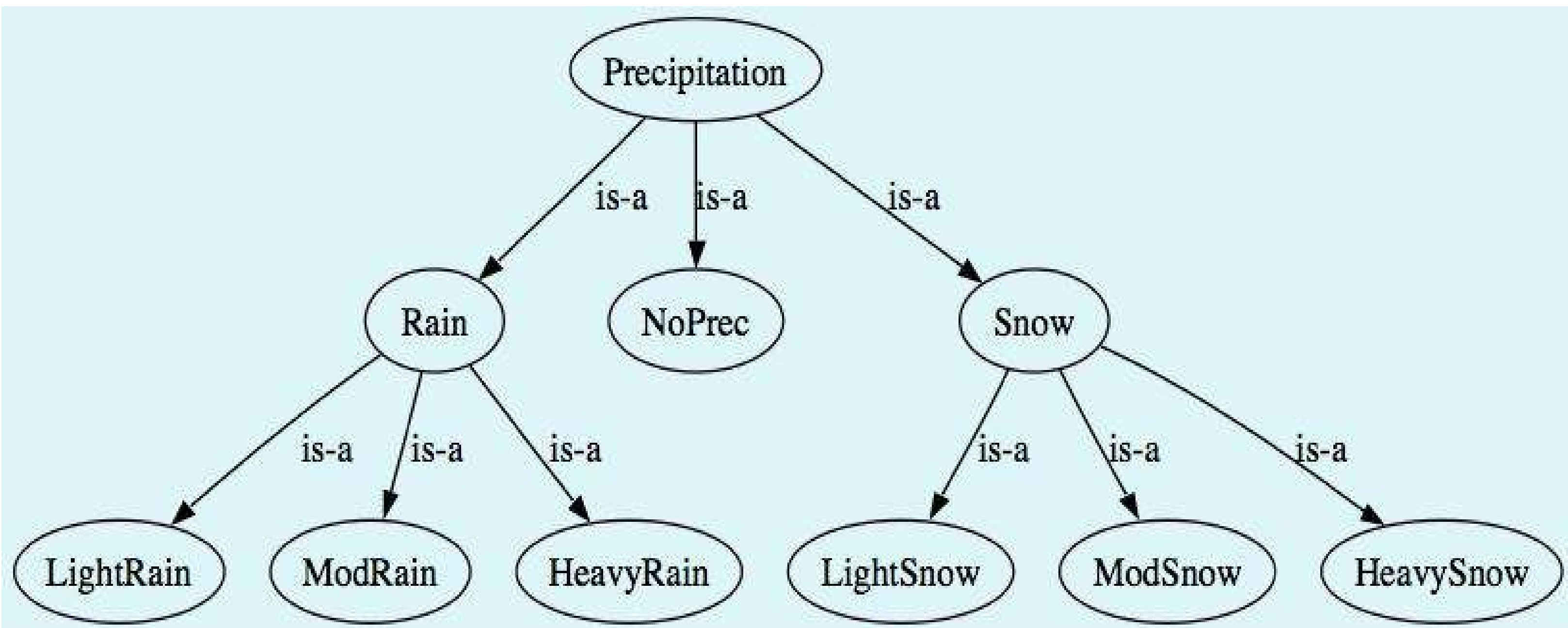
Exposing the ontology

- Schema semantics
- Data semantics

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Ontology Extended Data Sources

- Expose the data semantics
  - Special Case of interest:
    - Values of each attribute organized as an AVH



Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Ontology Extended Data Sources

- Ontology extended data source [Caragea et al, 2005]
- Inspired by ontology-extended relational algebra [Bonatti et al., 2003]
- Querying data sources from a user's point of view is facilitated by specifying mappings
  - From user schema to data source schemas
  - From user AVH to data source AVH
- More systematic characterization of OEDS and mappings within a description logics framework is in progress

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Mappings between schema

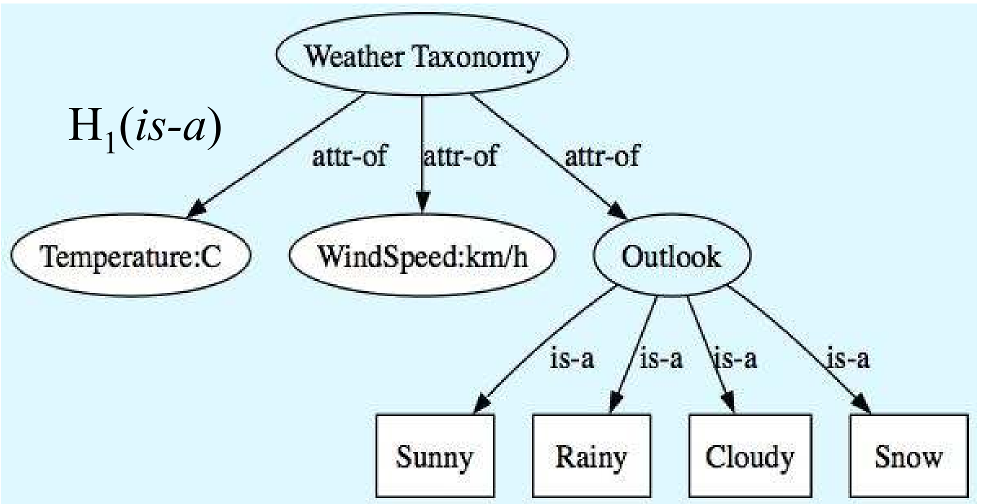
D_1	Day: day	Temperature: deg C	Wind Speed: kmh	Outlook: outlook
D_2	Day: day	Temp: deg F	Wind: mph	Precipitation: prec
D_U	Day: day	Temp: deg F	Wind: kmh	Outlook: outlook

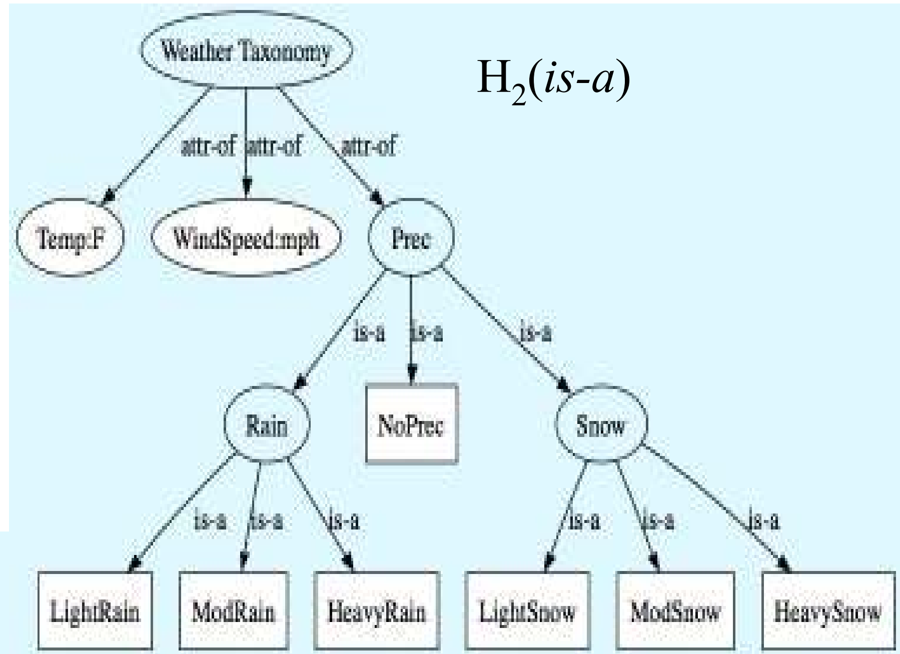
Day : $D_1 \equiv$ Day: D_U
Day : $D_2 \equiv$ Day: D_U

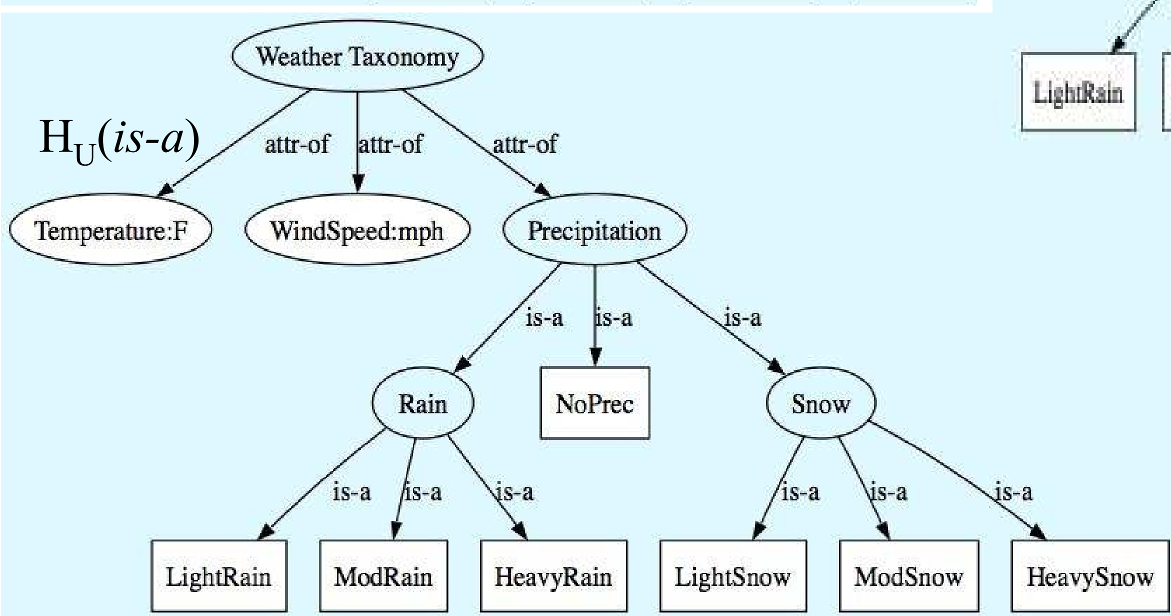
Temperature: $D_1 \equiv$ Temp : D_U
Temp: $D_2 \equiv$ Temp : D_U

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Semantic Correspondence between Ontologies



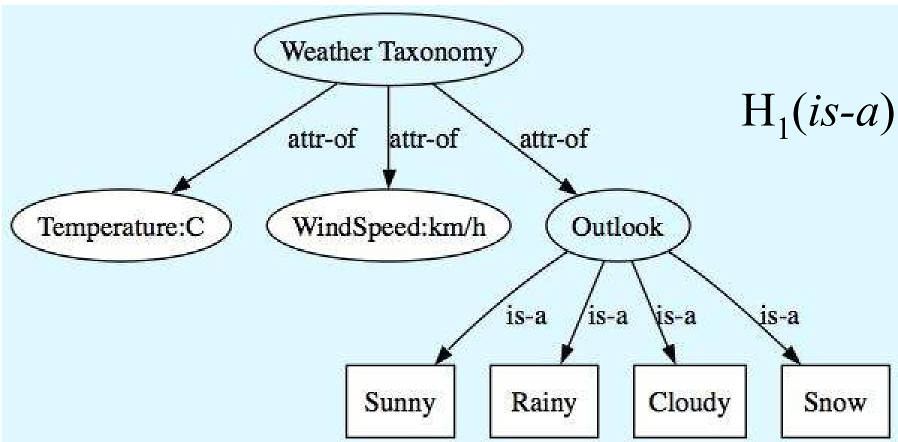


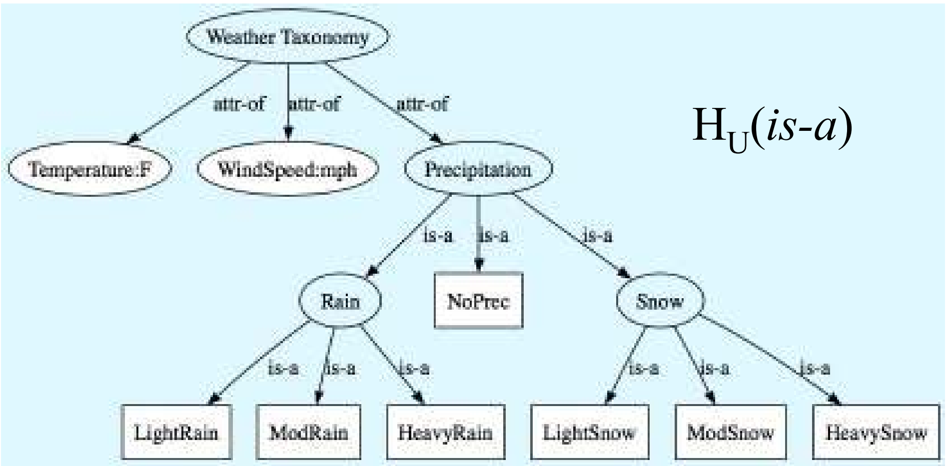


The white nodes represent the values used to describe data

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Data sources from a user's perspective





Rainy : H_1 = Rain : H_U

Snow : H_1 = Snow : H_U [Caragea, Pathak, and Honavar; 2004]

NoPrec : H_U < Outlook : H_1

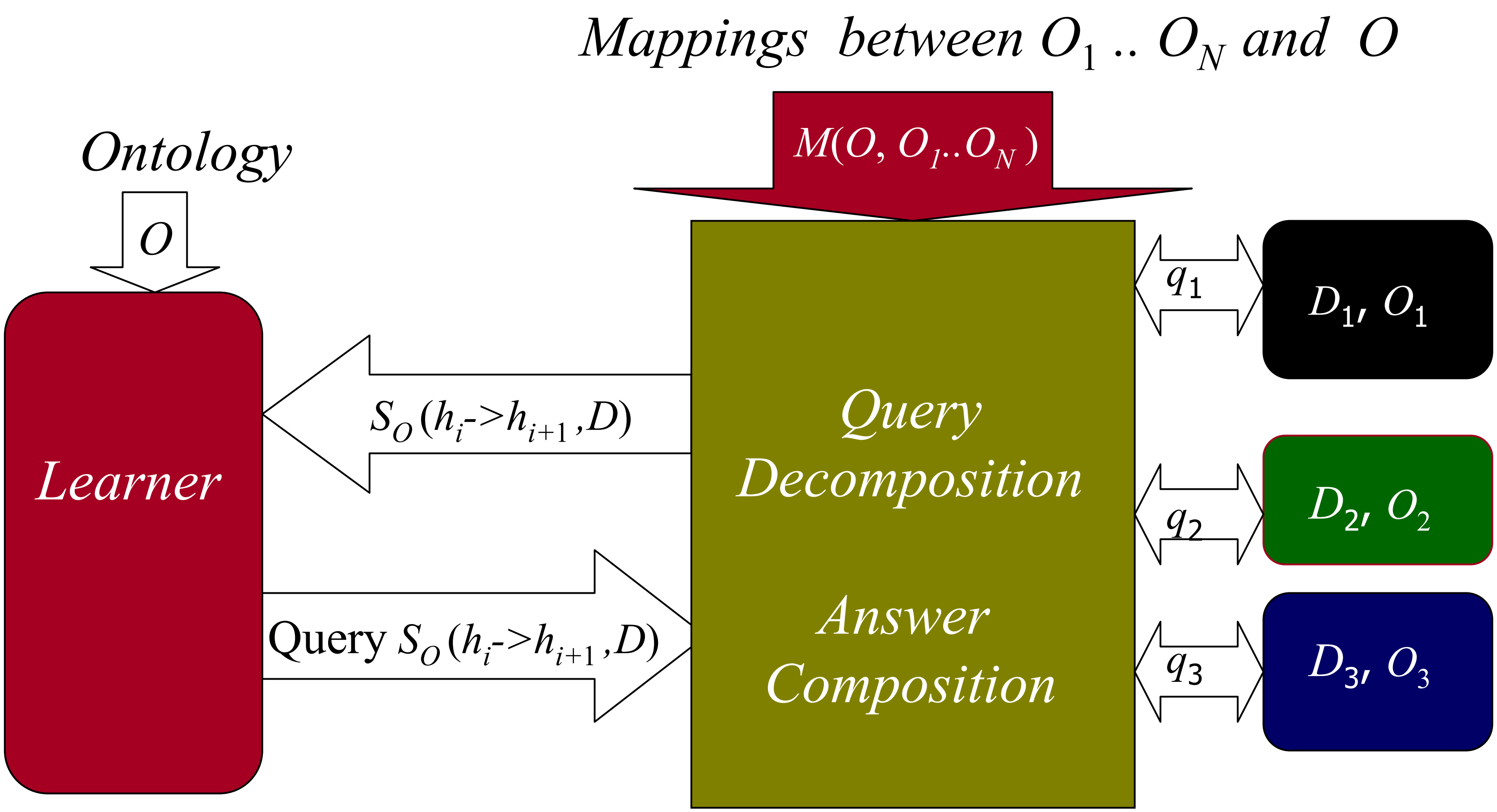
{Sunny, Cloudy} : H_1 = NoPrec : H_U

Conversion functions are used to map units

(e.g. degrees F to degrees C)

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

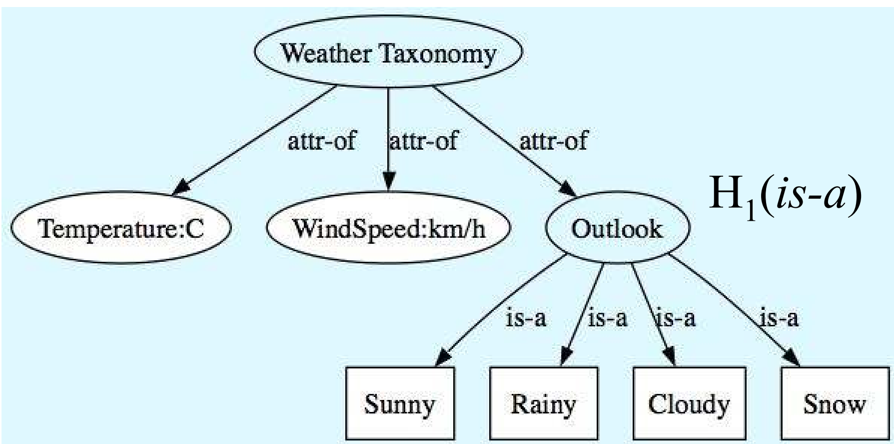
Learning from Semantically Heterogeneous Data

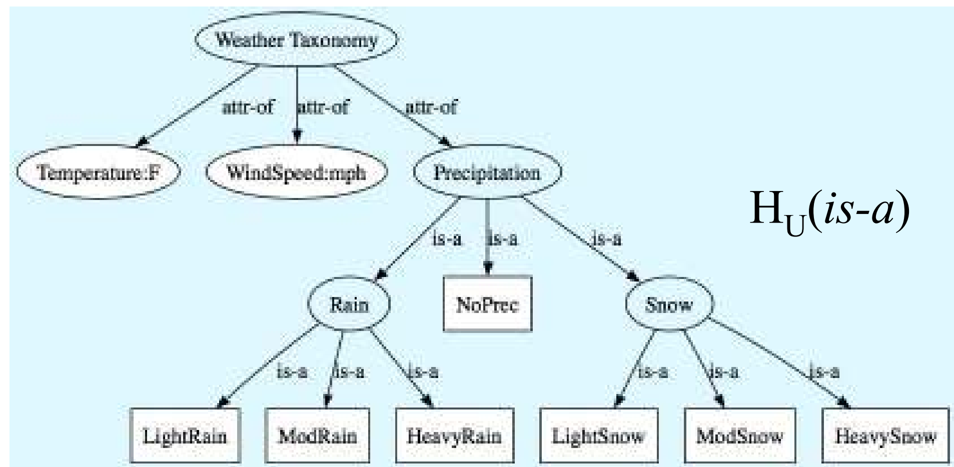


Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Semantic gaps lead to Partially Specified Data

- Different data sources may describe data at different levels of abstraction
- If the description of data is more abstract than what the user expects, additional statistical assumptions become necessary





Snow is under-specified in H_1 relative to user ontology – H_U
Making D_1 partially specified from the user perspective

[Zhang and Honavar, 2003; 2004, 2005]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

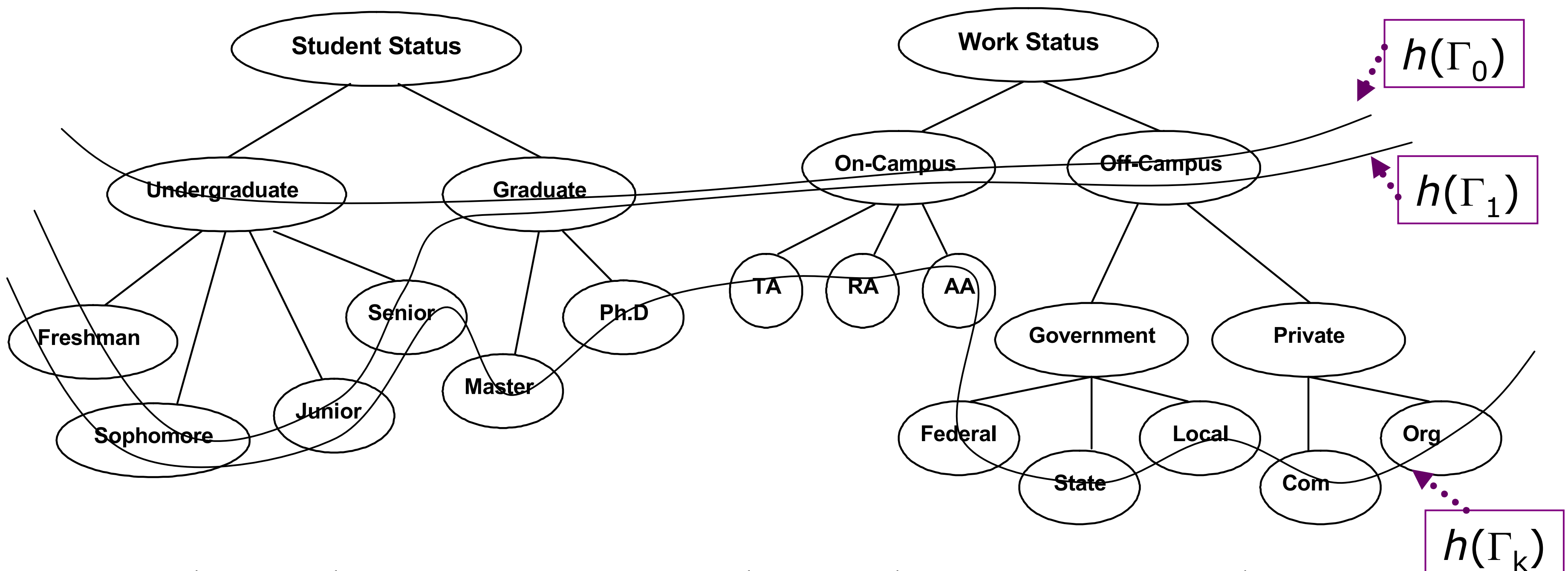
Outline

- Background and motivation
- Learning from data revisited
- Learning predictive models from distributed data
- Learning predictive models from semantically heterogeneous data
- Learning predictive models from partially specified data
- Current Status and Summary of Results

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning Classifiers from Attribute Value Taxonomies (AVT) and Partially Specified Data

Given a taxonomy over values of each attribute, and data specified in terms of values at different levels of abstraction, learn a concise and accurate hypothesis



[Zhang and Honavar, 2003; 2004; Zhang et al., 2006; Caragea et al., 2006]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Learning Classifiers from (AVT) and Partially Specified Data

Cuts through AVT induce a partial order over

- instance representations
- Classifiers

AVT-DTL and AVT-NBL

- Show how to learn classifiers from partially specified data
- Estimate sufficient statistics from partially specified data under specific statistical assumptions
- Use CMDL score to trade off classifier complexity against accuracy

[Zhang and Honavar, 2003; 2004; 2005]

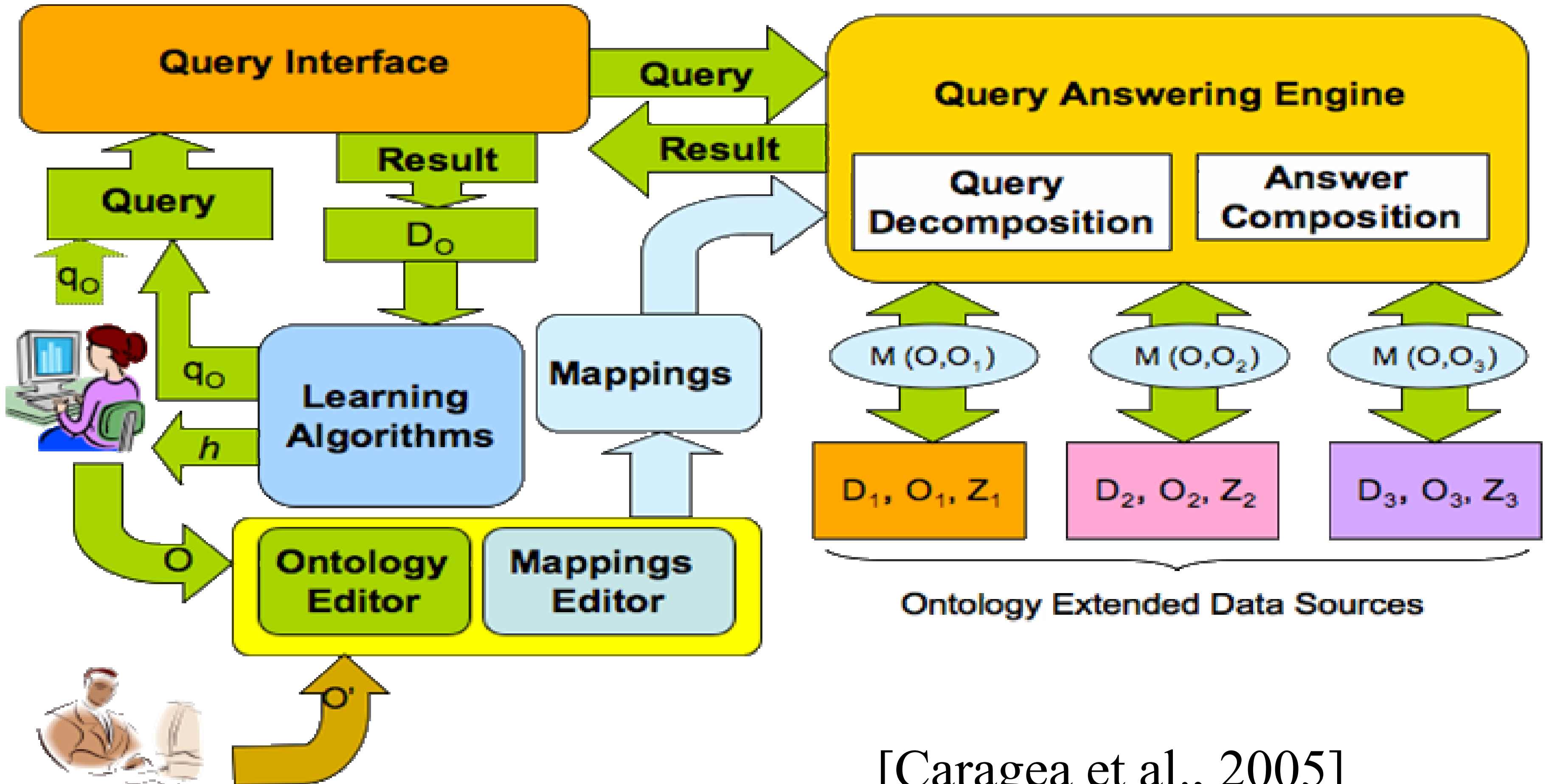
Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Outline

- Background and motivation
- Learning from data revisited
- Learning predictive models from distributed data
- Learning predictive models from semantically heterogeneous data
- Learning predictive models from partially specified data
- Current Status and Summary of Results

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Implementation: INDUS System



[Caragea et al., 2005]

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Summary

- Algorithms learning classifiers from distributed data with provable performance guarantees relative to their centralized or batch counterparts
- Tools for making data sources self-describing
- Tools for specifying semantic correspondences between data sources
- Tools for answering statistical queries from semantically heterogeneous data
- Tools for collaborative construction of ontologies and mappings, distributed reasoning..

Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Current Directions

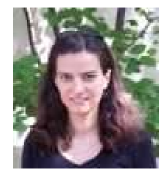
- Further development of the open source tools for collaborative construction of predictive models from data
- Resource bounded approximations of statistical queries under different access constraints and statistical assumptions
- Algorithms for learning predictive models from semantically disparate alternately structured data
- Further investigation of OEDS – Description logics, RDF..
- Relation to modular ontologies and knowledge importing
- Distributed reasoning, privacy-preserving reasoning…
- Applications in bioinformatics, medical informatics, materials informatics, social informatics

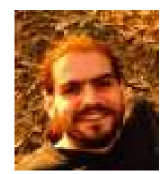
Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)

Acknowledgements

- Students
  - Doina Caragea, Ph.D., 2004
  - Jun Zhang, Ph.D., 2005
  - Jie Bao, Ph.D., 2007
  - Cornelia Caragea, Ph.D., in progress
  - Oksana Yakhnenko, Ph.D., in progress
- Collaborators
  - Giora Slutzki
  - George Voutsadakis
- National Science Foundation



Research supported in part by grants from the National Science Foundation (IIS 0219699, 0711356)