Ontology Summit 2014 Session 05 Track D: Tackling the Variety Problem in Big Data – I

Ken Baclawski

Anne Thessen

Track D Co-Champions

Session Outline

- Ken Baclawski - Introduction
- Eric Chan (Oracle) - **Enabling OODA Loop with Information Technology**
- Nathan Wilson (Encyclopedia of Life) - **The Semantic Underpinnings of EOL TraitBank**
- Ruth Duerr (National Snow and Ice Data Center) - **Semantics and the SSIII Project**
- Anatoly Levenchuk - Hackathon (Track E) Announcement
- Open Panel Discussion

History of Big Data

- Recognition of the problem in early 1980s: originally called the “information onslaught” [1]
- Different fields experience the problem at different times.
- The transition from information scarcity to information abundance in a field appears to occur abruptly.
- Organizations have difficulty coping with the transition.
- Many organizations are still based on the information scarcity model.

Characteristics of Big Data

- Large amounts of data (volume)
- Rapid pace of data acquisition (velocity)
- Complexity of the data (variety)
    - Complexity of individual data items
    - Both structured and unstructured data
    - Complexity of data from one source
    - Multiple sources

History of Big Data

- Development of new storage and indexing strategies for handling volume and velocity
  - “Map Reduce” was developed in 1994. [2]
- Development of techniques for handling variety
  - Schema mapping
  - Controlled vocabularies
  - Knowledge representations
  - Ontologies and semantic technologies
- Connection between these two?
  - Surprisingly little collaboration and communication.
  - A notable exception is the early work starting in 1992 on representing biological research papers. [3]

The Potential of Big Data

- Could address important social and commercial needs.
- Traditional techniques are inadequate
  - Cost, feasibility, ethical concerns
- New techniques are much better, but
  - Still have ethical issues
  - Privacy has become a prominent issue
  - The variety problem is typically handled with ad hoc techniques

Why is semantics important?

- Misunderstanding the data can result in invalid or misrepresented analyses.
  - A recognized problem well before Big Data.
  - John P. A. Ioannidis claimed that most published scientific research findings are false. [4] [5]
    - Many forms of bias can invalidate results
    - Bias and observer effects are difficult to recognize
    - Standard experimental procedures do not eliminate bias.
- Could Big Data exacerbate this problem?

What can ontologies do to help?

- Background knowledge of the domain
- The structure of the data
- Annotation of data and metadata
- Provenance of the data
    - Transformations
    - Analyses
    - Interpretations
- Data processing workflows
- Privacy concerns
- Hypothesis generation and their workflows

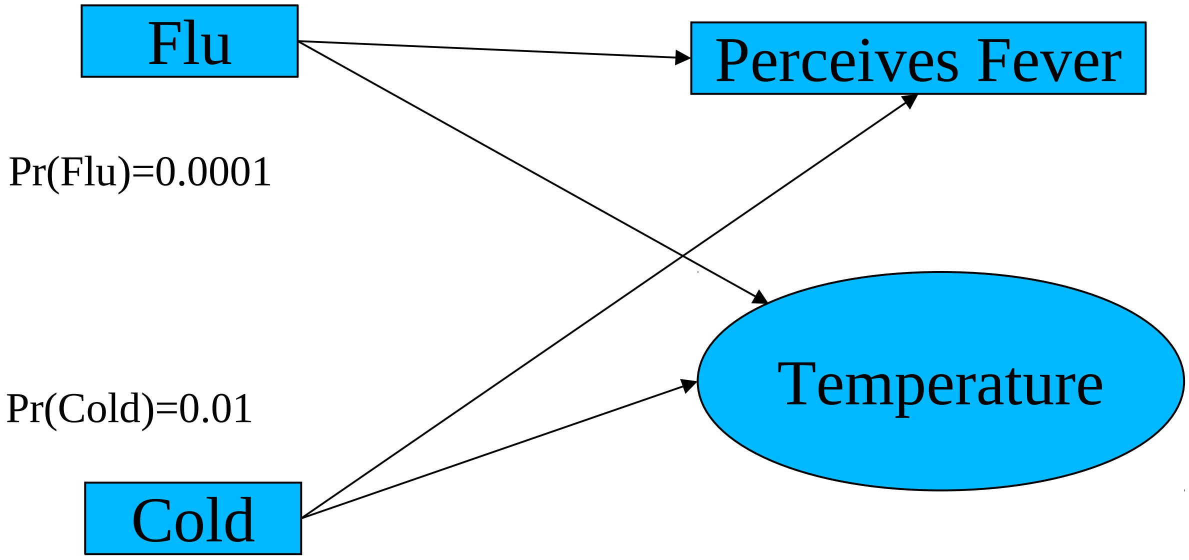
Some Challenges

- Little collaboration between the communities
- Big Data focus on volume and velocity, assuming someone else will handle variety
- Tool incompatibility
- Incompatibility between statistical and logical techniques (hybrid reasoning gap)

Statistical versus Logical Reasoning

- There is no intrinsic incompatibility.
  - Probability has been axiomatized
  - A formal ontology of probability and statistics is possible
  - One can formally represent statistical statements in such an ontology [6]
- Bayesian networks are especially powerful.
  - Can be developed using the same techniques used for developing ontologies [7]

Bayesian Network Specification



Pr(Flu)=0.0001

Pr(Cold)=0.01

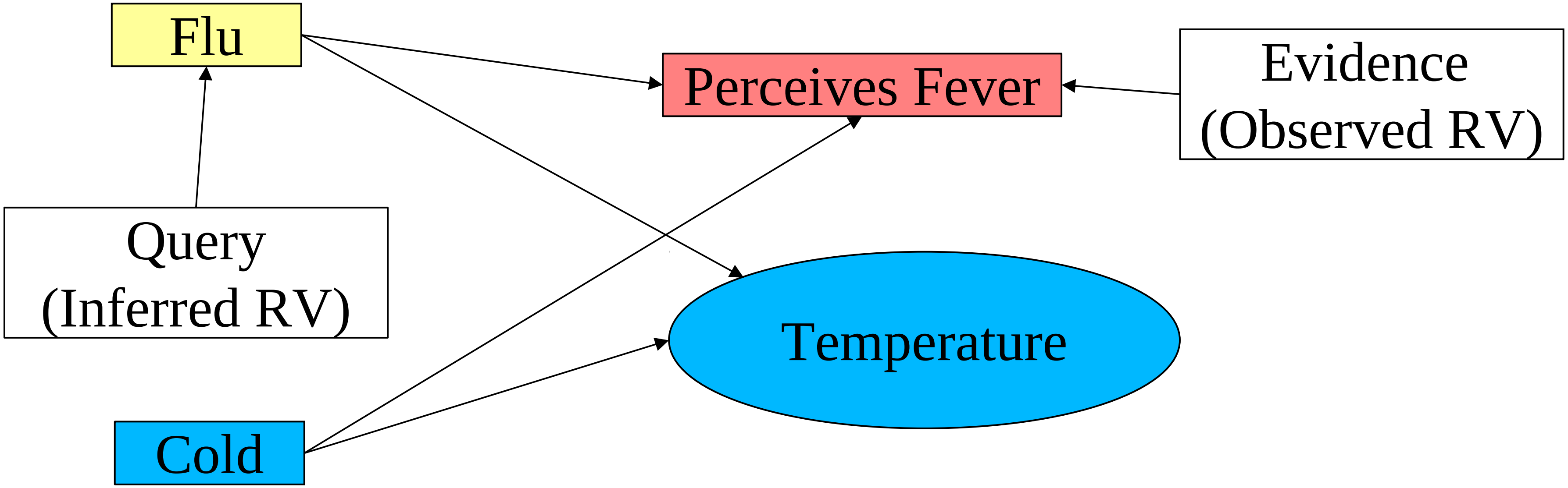
	Perceives Fever	
	TRUE	FALSE
Flu = F & Cold = F	0.01	0.99
Flu = F & Cold = T	0.1	0.9
Flu = T & Cold = F	0.9	0.1
Flu = T & Cold = T	0.95	0.05

	Temperature	
	Mean	Std Dev
Flu = F & Cold = F	37	0.5
Flu = F & Cold = T	37.5	1
Flu = T & Cold = F	39	2
Flu = T & Cold = T	39.5	2.5

Discrete Random Variable

Continuous Random Variable

Bayesian Network Inference



- Inference is performed by observing some random variables (evidence) and inferring some others (query).
- The evidence can be a value or a probability distribution.
- The answer to the query is the marginal distribution.

Speakers

- Eric Chan (Oracle)
  - Tool development addressing many aspects of the variety problem including provenance, metadata, and hypothesis generation and workflow (OODA loop)
- Nathan Wilson (Encyclopedia of Life)
  - Experiences with using semantic technology in the tree of life, a notoriously messy ontology

Speakers

- Ruth Duerr (National Snow and Ice Data Center)
  - Experiences with using ontologies to address very heterogeneous sources of data for snow and ice, including satellite data, model output, point observations, social science data (interviews with Alaskan community members), and many other kinds of data.

Track D Wiki Pages

Synthesis Page

http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2014_Tackling_Variety_In_BigData_Synthesis

Community Input Page

http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2014_Tackling_Variety_In_BigData_CommunityInput

Bibliography

1. D. Kerr, K. Braithwaite, N. Metropolis, D. Sharp, G.-C. Rota (Ed). Science, Computers and the Information Onslaught. Academic Press. (1984)
2. K. Baclawski and J.E. Smith. High-performance, distributed information retrieval. Northeastern University, College of Computer Science. (1994)
3. K. Baclawski. Data/knowledge bases for biological papers and techniques. Northeastern University, College of Computer Science. NSF grant. (1992)
4. Ioannidis JPA, Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124 (2005)
5. Critical Data Conference: Secondary Use of Big Data from Critical Care http://criticaldata.mit.edu/ (2014)
6. K. Baclawski and T. Niu. Ontologies for Bioinformatics. (2005)
7. K. Baclawski. Bayesian network development. In International Workshop on Software Methodologies, Tools and Techniques, pages 18-48. Keynote address. (September, 2004)