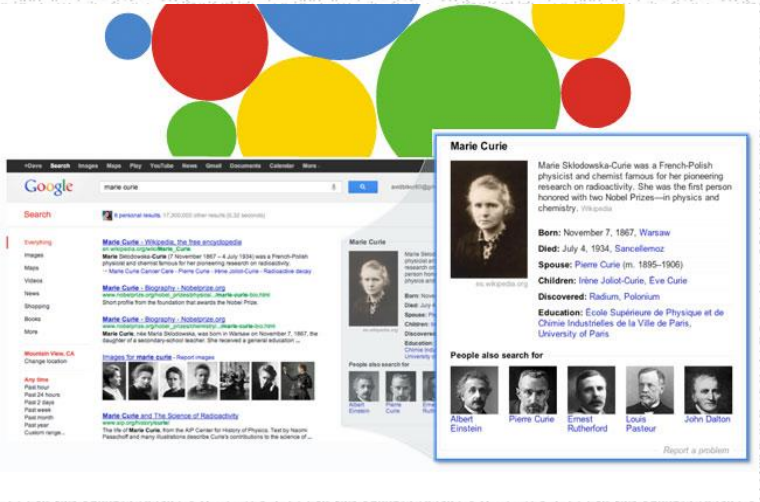
BIG DATA SYSTEMS FOR KNOWLEDGE BASE CONSTRUCTION FROM TEXT IMAGES AND CROWDS

Daisy Zhe Wang
CISE, University of Florida
Data Science Research Lab, Database Research Group
06/19/2013

Sponsors:

Knowledge Bases From Big Data

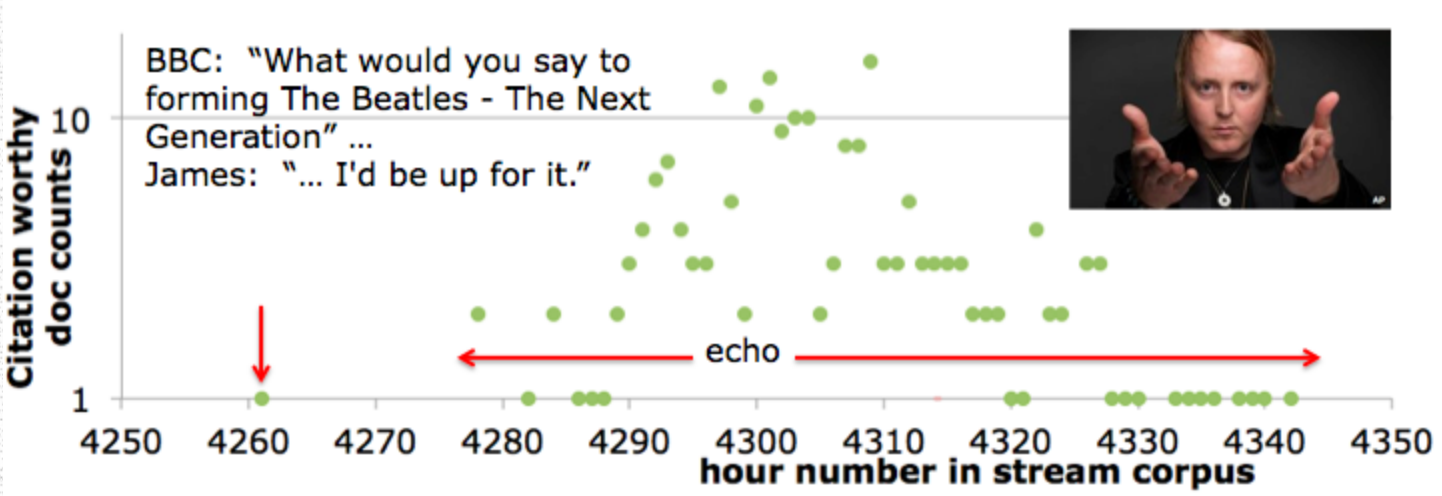
- A *knowledge base* is a collection of entity, facts, relationships that conforms with a certain data model.
- A knowledge base helps machine understand humans, languages, and the world.
- Examples 1: Google Knowledge Graph [Text, Images, Crowd]



DSR @ UF
Data Science Research

Knowledge Bases From Big Data

- Example 2: TREC Knowledge Base Acceleration [News, Glog, Tweets]



- KB Applications:
  - Improve Search Engine/Wikipedia
  - Support Conversation/Q&A Systems
  - Provide Context to Localized Sensing (e.g., Tweet, Image)
  - Domain-specific Knowledge Bases (e.g., EMR analysis)

DSR @ UF
Data Science Research

Life of A Knowledge Base (KB)

- A *knowledge base system* is a special kind of database management system to for knowledge base management.

- *KB extraction*: knowledge extraction using statistical models in NLP/ML literature
- *KB expansion*: knowledge inference using deductive rules and knowledge with uncertainty
- *KB evolution*: knowledge update given new evidence and new data sources
- *KB integration*: knowledge integration from multiple sources (e.g., human, data sets, models)

DSR @ UF
Data Science Research

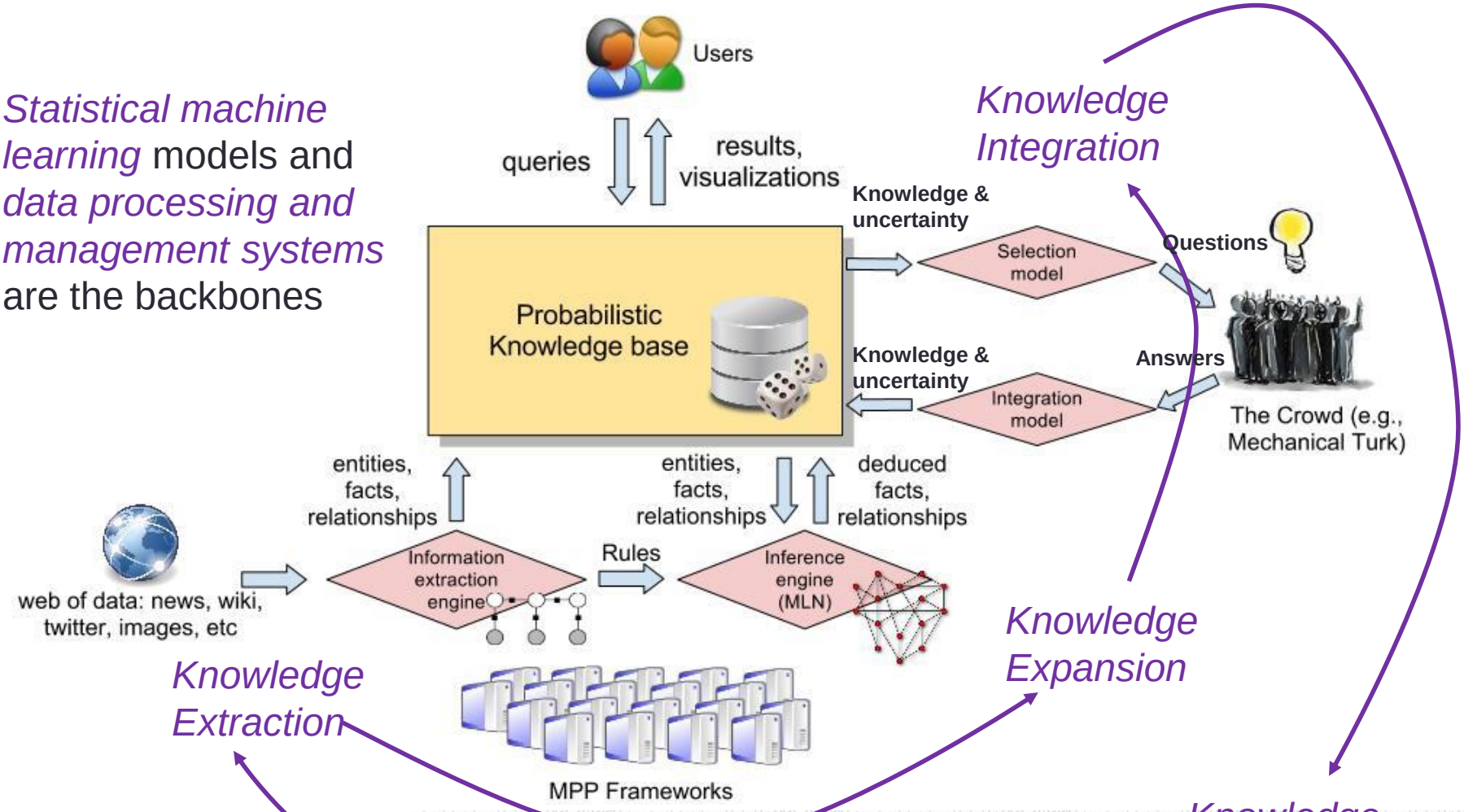
Uncertainty Management

- Where does Uncertainty come from?
  - Inherent in the state-of-the-art NLP/SML results
  - Incorrect, conflicting data sources
- Derived facts and query results
- Uncertainty vs. NULL
- Uncertainty vs. MAP/majority voting

- *Probability Theory* should lay the foundation for uncertainty management → *Probabilistic Knowledge Base System*

Probabilistic Knowledge Base System

Statistical machine learning models and *data processing and management systems* are the backbones



Projects & Big Data Challenges

- Probabilistic Database for Knowledge Extraction
  BayesStore (2006-2011) [VLDB08, VLDB10, ICDE10, SIGMOD11]
- Scalable NLP over MPP frameworks
  MADLib [VLDB12, CIKM12, SIGMOD-DanaC13, ICMLA12]
- Crowd Assisted Machine Learning *CAMeL*
  - Human/Machine Knowledge Integrating
- Knowledge Base for Image Retrieval and Extraction
- Query-Driven Large-scale Co-reference *Archer*
  - Ad-hoc/continuous queries over static/streaming dataset
- Scalable Inference over Probabilistic KB *ProbKB*
  - Scalability, Efficient Updates
- Visualizing Knowledge Bases *VizSearch*

DSR @ UF
Data Science Research

Inter-Disciplinary Collaborations

- Document Retrieval and Predictive coding in E-discovery

- Knowledge Extraction and Outcome Prediction using Medical Notes

- EDEN: Constructing Knowledge Base for Ecological Data Extraction and Exploration

DSR @ UF
Data Science Research