Microsoft Research

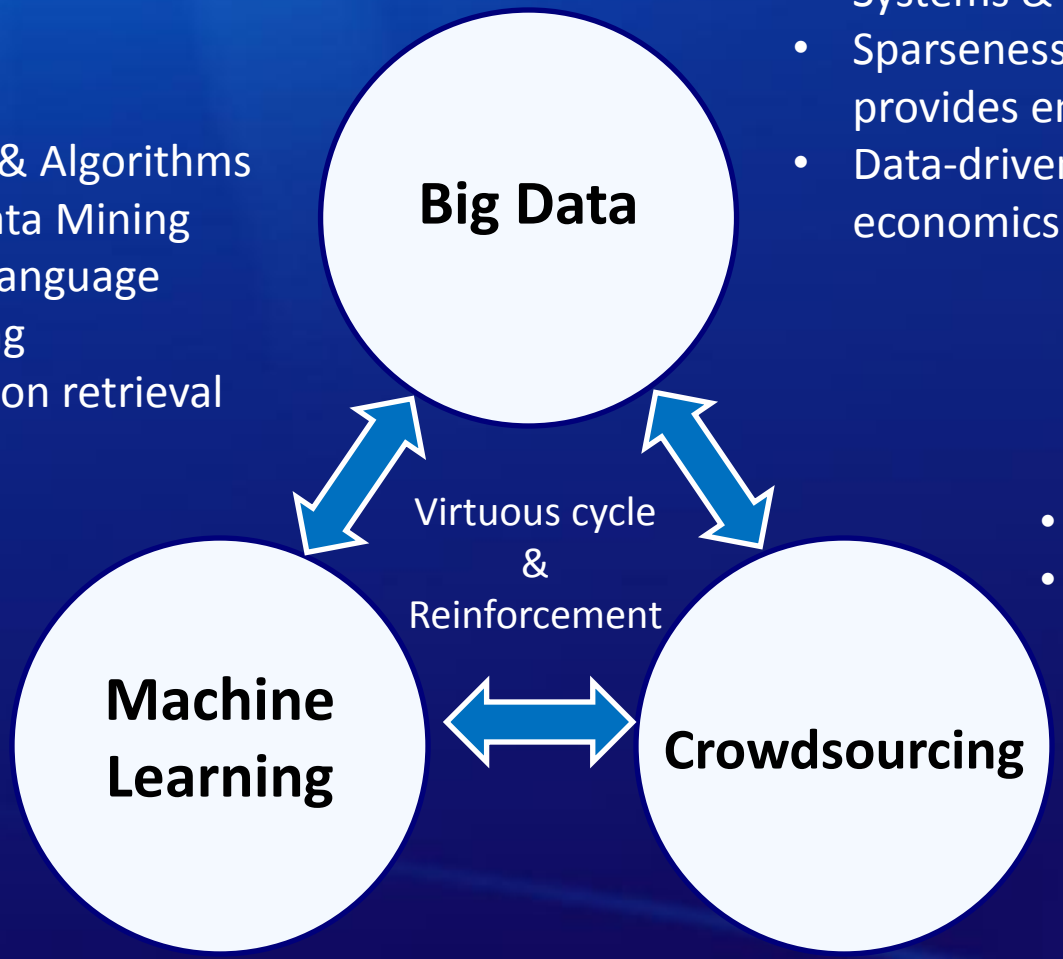
Semantic Search and a New Moore's Law in Knowledge Engineering

Wei-Ying Ma
Microsoft Research Asia

A New Moore's Law in Knowledge Engineering – Driven by Three Major Trends

- Systems & Infrastructure
- Sparseness -> abundance of data provides enough signals for ML
- Data-driven businesses and data economics

- Sciences & Algorithms
- Text & Data Mining
- Natural Language Processing
- Information retrieval
- Analytics



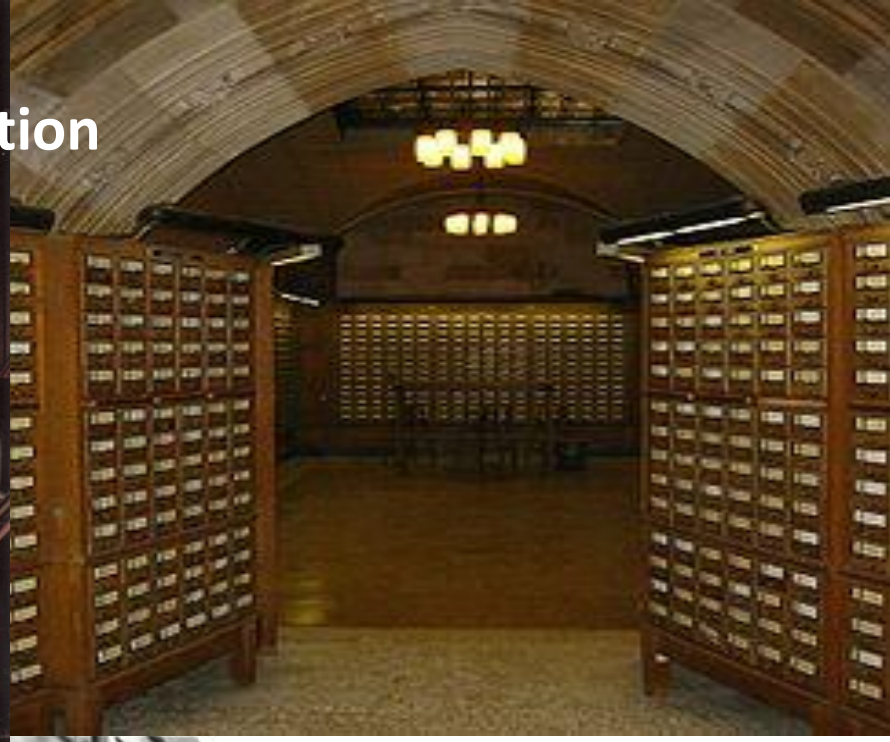
- Human computation
- Human tasking
  - Incentives
  - Network effect

Library: organize the world's information

Library card index

Query → Indexing → Documents

- Query: book title, author name
- Indexing: inverted indices
- Documents: books



The First Generation of Search Engines

- Essentially were invented to replace library card index
  - A “search document” paradigm
  - Advantages: Fast and scalable

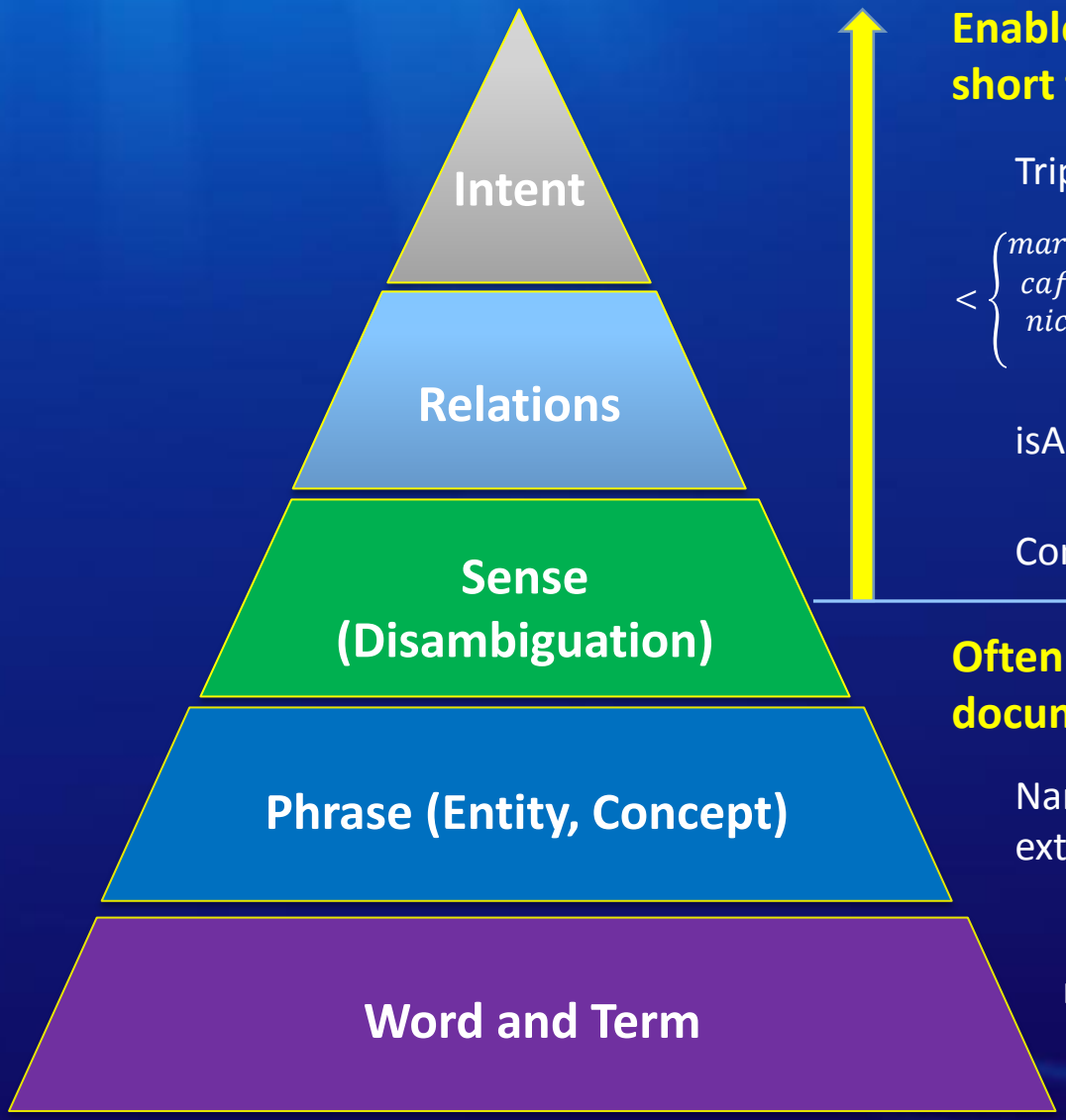
- Search Paradigm:
  Query → Indexing → Documents → Ranking
  - Query: any words appearing in pages
  - Indexing: inverted indices
  - Documents: pages, images
  - Ranking: Classical information retrieval (IR) techniques + PageRank

Search Engines Today

- Search Paradigm (has not changed much)
  Query → Indexing → Documents → Ranking
  - Query: any words appearing in pages
  - Indexing: inverted indices
  - Documents: pages, images, videos, books, answers,…
  - Ranking: More signals (features) are used; machine learning; log mining; human feedbacks, etc.

- Challenges and opportunities
  - Semantic understanding of documents and queries
  - Information explosion & information overload

Different Level of Semantic Matching



Enable micro-level understanding of short text, e.g. queries, ads, sentences

Triple relations: subject + verb + object

$$< \begin{Bmatrix} marijuana \\ caffeine \\ nicotine \\ \dots \end{Bmatrix}, \begin{Bmatrix} result\ in \\ be\ risk\ factor\ for \\ be\ major\ cause\ of \\ \dots \end{Bmatrix}, \begin{Bmatrix} insomnia \\ emphysema \\ breast\ cancer \\ \dots \end{Bmatrix} >$$

isA, isPropertyOf, and general relations

Conflation, Disambiguation

Often for macro-level understanding of document (e.g. topics)

Named entity recognition (NER), Phrase extraction

ngram, Bag of words (BOW), etc.

Organize the World's Information → Directly Fulfill People's Information Need

Document Space

Trillions of Web pages
(and still growing rapidly)

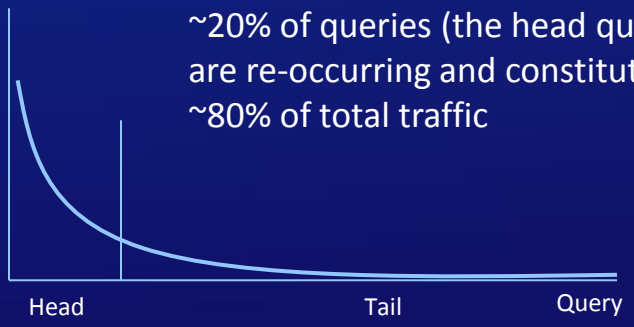
Indexed by search engines

A small (and decreasing) percentage of Web pages in the index could appear in SERP

Query Space

Unique queries per month – much smaller and finite

~20% of queries (the head queries) are re-occurring and constitute ~80% of total traffic

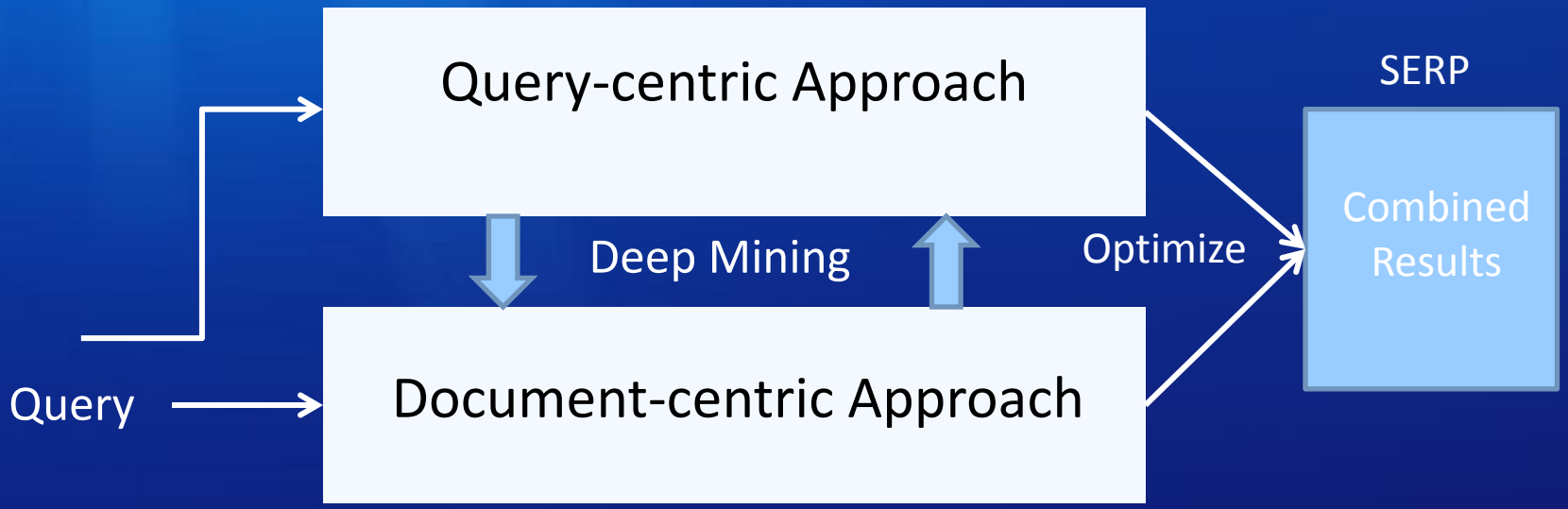


How ***information*** is organized → How people express their ***information needs***

Organize and Accumulate Knowledge around People's Information Need

Document-centric Model	Query-centric Model
Index and keywords are primary, queries are transient	Queries (both on aggregate and a single user's) are a first-class object
Organize and search documents (crawling → indexing → documents → ranking)	Organize and accumulate knowledge around queries — make queries searchable
Index is centralized and massive (requires lightweight, common query processing for scale)	Smaller query "footprint" allows for deeper mining of queries
Document-based results in SERP	Hybrid search: Document- and query-based models running in parallel, combined results in optimized SERP

New Search Paradigm – a Hybrid Model



- Hybrid Search, combine benefits of:
  - Document-centric + Query-centric
  - Cloud (shared – head/body) + Clients (personal – tail/complex)
- Enables:
  - Deeper mining for better results (intent + knowledge)
  - New user experiences (task, app search, dialog, user control)

Two Types of Information Need

Noun-centric	Verb-centric
• **Informational intent** ➢ Often about **entities** (people, places, things) • **Wikipedia** is a good example of knowledge manually crafted by human ➢ Often ranked top in search result if there is a match ➢ High precision but low coverage	• **Transactional intent** ➢ Rent a car ➢ Book a table at restaurant ➢ Buy a product ➢ Check in a flight ➢ Print boarding pass ➢ …
• **Question: can we create 1 trillion of entity/knowledge pages** ➢ High coverage with reasonable precision ➢ Leverage social, interactive mining, and crowdsourcing	• Identify and define important tasks ➢ Leverage developer ecosystem and crowdsourcing

Building an Entity Graph

- Nodes – Entity pages with profile information
- Links – Relationships among entities
- Links to other graphs (e.g. Web graph, other social graph such as Facebook, etc.)

Entity Extraction, Integration, & Summarization

- Different Technologies for Knowledge Acquisition
  - Domain-specific vs. Across-domain (general purpose)
  - Automatic vs. Interactive
  - Breadth vs. Depth vs. Density (i.e. link/relationship)
  - Ability to understand a question or answer a question
- Sources of Knowledge
  - The web crawl (unstructured, semi-structured, structured data)
  - Wikipedia and other human crafted knowledge bases
  - Search query log (user interaction data, etc.)
  - Crowdsourcing (game, social, or Mechanical Turk, etc.)

Entities in Academic Domain

http://academic.research.microsoft.com

Author

Jiawei Han
University of Illinois Urbana Champaign
Publications: 546 | Citations: 14815 | G-Index: 113 | H-Index: 56
Interest: Databases, Data Mining, Artificial Intelligence
275 related publication(s)

Christos Faloutsos
Carnegie Mellon University
Publications: 392 | Citations: 12265 | G-Index: 104 | H-Index: 52
Interest: Databases, Data Mining, Multimedia
122 related publication(s)

Philip S. Yu
University of Illinois Chicago
Publications: 673 | Citations: 9488 | G-Index: 78 | H-Index: 45
Interest: Databases, Data Mining, Distributed and Parallel Computing
200 related publication(s)

Vipin Kumar
University of Minnesota
Publications: 472 | Citations: 9427 | G-Index: 86 | H-Index: 46
Interest: Distributed and Parallel Computing, Data Mining, Artificial Intel
100 related publication(s)

Rakesh Agrawal
Microsoft
Publications: 261 | Citations: 19966 | G-Index: 140 | H-Index: 56
Interest: Databases, Data Mining, Artificial Intelligence
62 related publication(s)

Heikki Mannila
University of Helsinki
Publications: 254 | Citations: 7031 | G-Index: 80 | H-Index: 37
Interest: Data Mining, Databases, Artificial Intelligence
105 related publication(s)

Wei Wang
University of North Carolina Chapel Hill
Publications: 1361 | Citations: 5231 | G-Index: 55 | H-Index: 32
Interest: Databases, Data Mining, Networks and Communications

Conference

Knowledge Discovery and **Data Mining** -
Publications: 2,015 | Citation Count: 38,635 | Year Range: 1994-2010
2015 related publication(s)

IEEE International Conference on **Data Mining** - ICDM
Publications: 1,683 | Citation Count: 8,585 | Year Range: 2000-2009
1683 related publication(s)

Pacific-Asia Conference on Knowledge Discovery and **Data Mining** - PAKDD
Publications: 1,247 | Citation Count: 3,082 | Year Range: 1997-2010
1247 related publication(s)

Principles of **Data Mining** and Knowledge Discovery - PKDD
Publications: 1,034 | Citation Count: 4,692 | Year Range: 1997-2010
1034 related publication(s)

SIAM International Conference on **Data Mining** - SDM
Publications: 685 | Citation Count: 4,372 | Year Range: 2001-2010
685 related publication(s)

Workshop on Knowledge Discovery and **Data Mining** - WKDD
Publications: 524 | Citation Count: 39 | Year Range: 2002-2010
524 related publication(s)

Rough Sets, Fuzzy Sets, **Data Mining**, and Granular-Soft Computing - RSFDGrC
Publications: 485 | Citation Count: 494 | Year Range: 1999-2009
485 related publication(s)

Advanced **Data Mining** and Applications - ADMA
Publications: 446 | Citation Count: 135 | Year Range: 2005-2009
446 related publication(s)

Research Issues on **Data Mining** and Knowledge Discovery - DMKD
Publications: 99 | Citation Count: 2,100 | Year Range: 1996-2004
99 related publication(s)

Int. Conf. on **Data Mining** - DMIN
Publications: 374 | Citation Count: 81 | Year Range: 2005-2009

Journal

Data Mining and Knowledge Discovery -
Publications: 344 | Citation Count: 9,209 | Year Range: 1997-2010
344 related publication(s)

SIGKDD Explorations
Publications: 378 | Citation Count: 4,396 | Year Range: 1999-2009
182 related publication(s)

International Journal of Business Intelligence and **Data Mining** - IJBIDM
Publications: 101 | Citation Count: 49 | Year Range: 2005-2010
101 related publication(s)

Statistical Analysis and **Data Mining**
Publications: 62 | Citation Count: 33 | Year Range: 2008-2010
62 related publication(s)

Expert Systems with Applications - ESWA
Publications: 3,556 | Citation Count: 3,973 | Year Range: 1990-2010
140 related publication(s)

International Journal of **Data Mining** and Bioinformatics - IJDMB
Publications: 63 | Citation Count: 65 | Year Range: 2005-2009
63 related publication(s)

IEEE Transactions on Knowledge and **Data** Engineering - TKDE
Publications: 2,127 | Citation Count: 29,698 | Year Range: 1988-2010
194 related publication(s)

The Computing Research Repository - CORR
Publications: 22,876 | Citation Count: 87,355 | Year Range: 1980-2010
171 related publication(s)

International Journal of **Data** Warehousing and **Mining** - IJDWM
Publications: 81 | Citation Count: 108 | Year Range: 2005-2009
81 related publication(s)

SIGMOD Record
Publications: 2,751 | Citation Count: 83,593 | Year Range: 1969-2009
86 related publication(s)

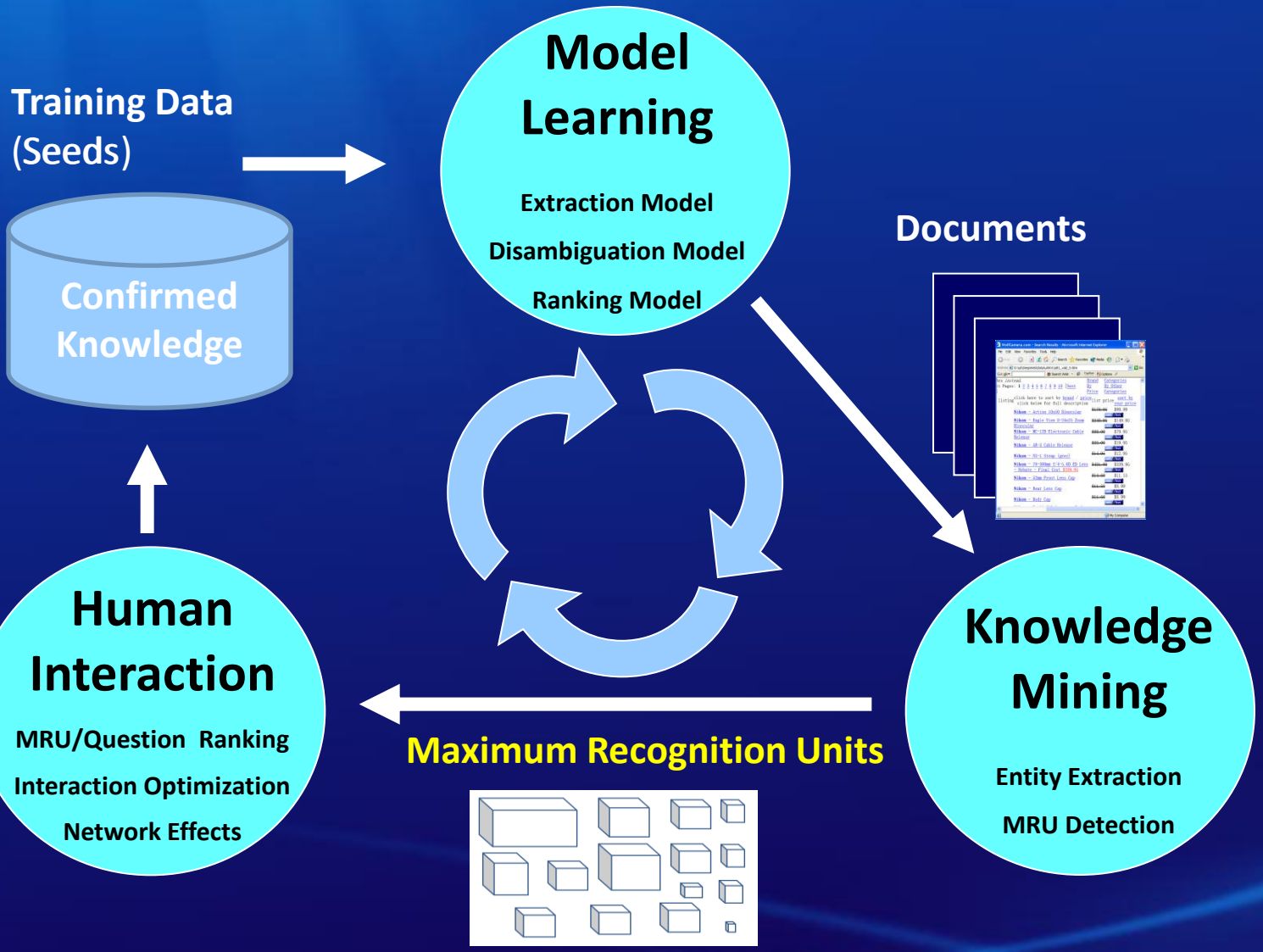
Serious Name Disambiguation Issues in Automated Mining

- Many people with popular names
  - For example, we have 200k+ Lei Zhang in China and 10+ Lei Zhang in MSFT
- Fully automated name disambiguation impossible
  - Simply don't get enough signals
  - Need knowledge from people to disambiguate/connect
- Solution: add people into knowledge mining loop
  - Interactive mining + crowdsourcing

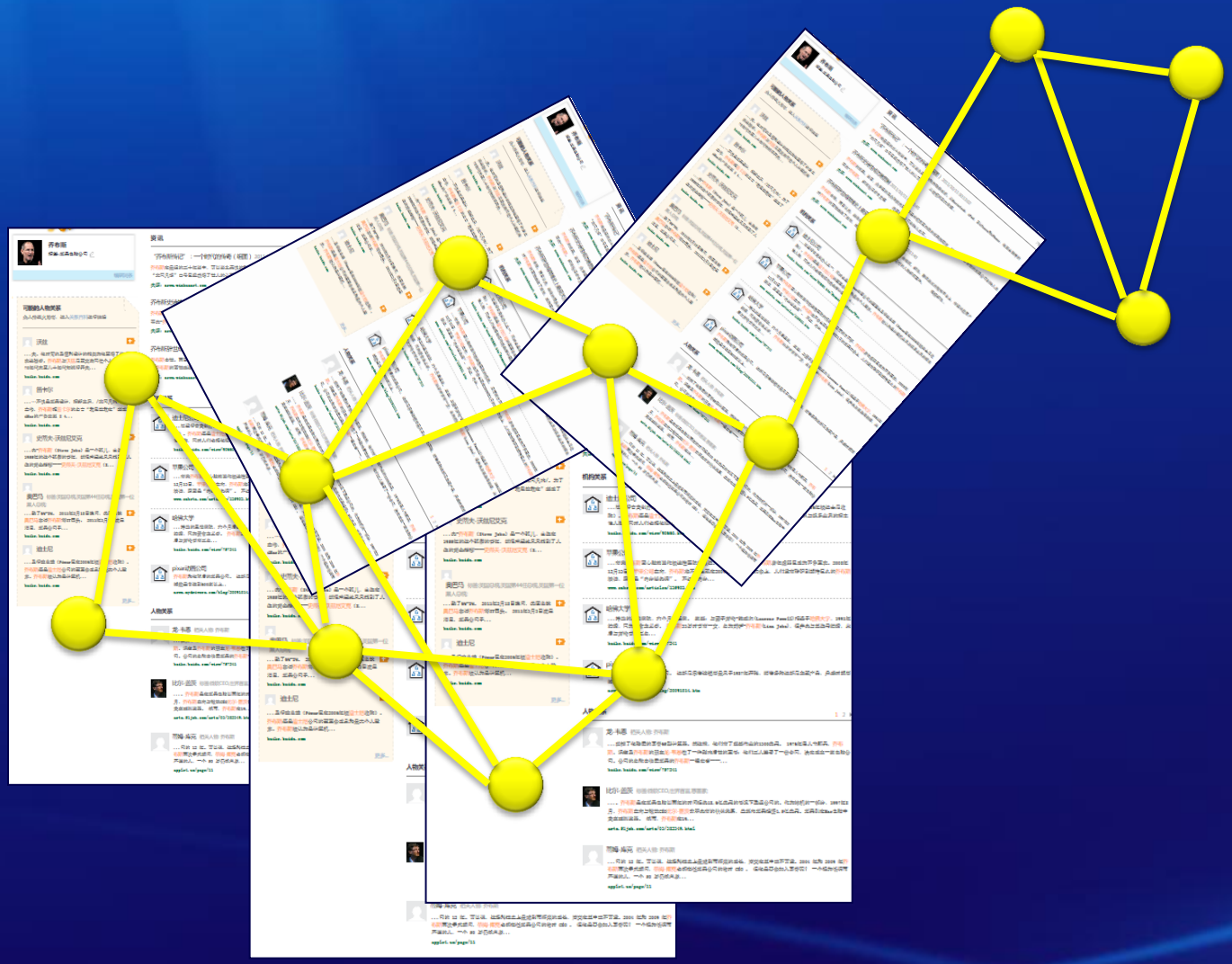
Demo

- An Interactive Wiki Editing Tool for Entity Graph
- Q20 (a 20 Question Game)
  - Use the knowledge mined from the web to create a game
  - Connected with celebrity profile editing
  - Help us collect more knowledge from users
  - http://renlifang.msra.cn/Q20/index.aspx (Chinese version)

Interactive Mining + Crowdsourcing



Network Effect in Entity Graph Editing

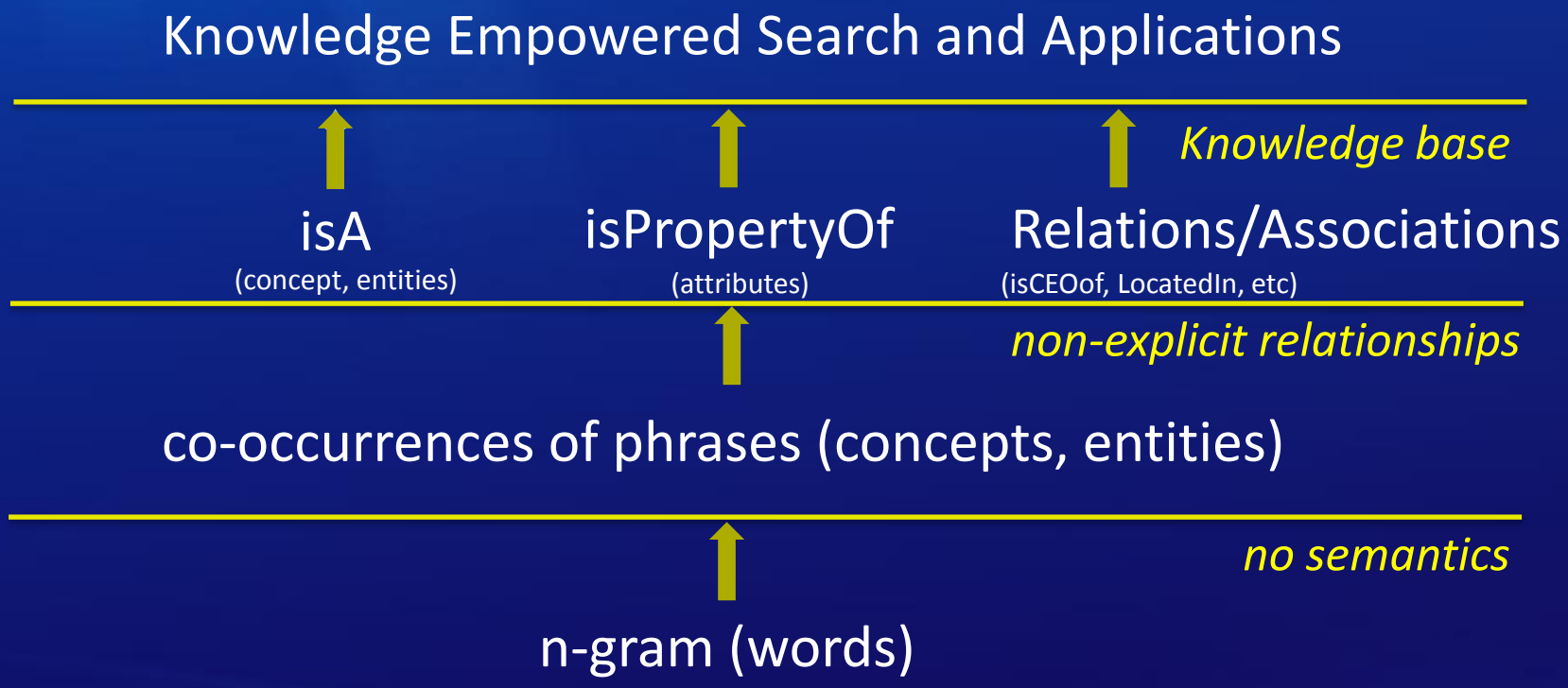


Understanding vs. Answering

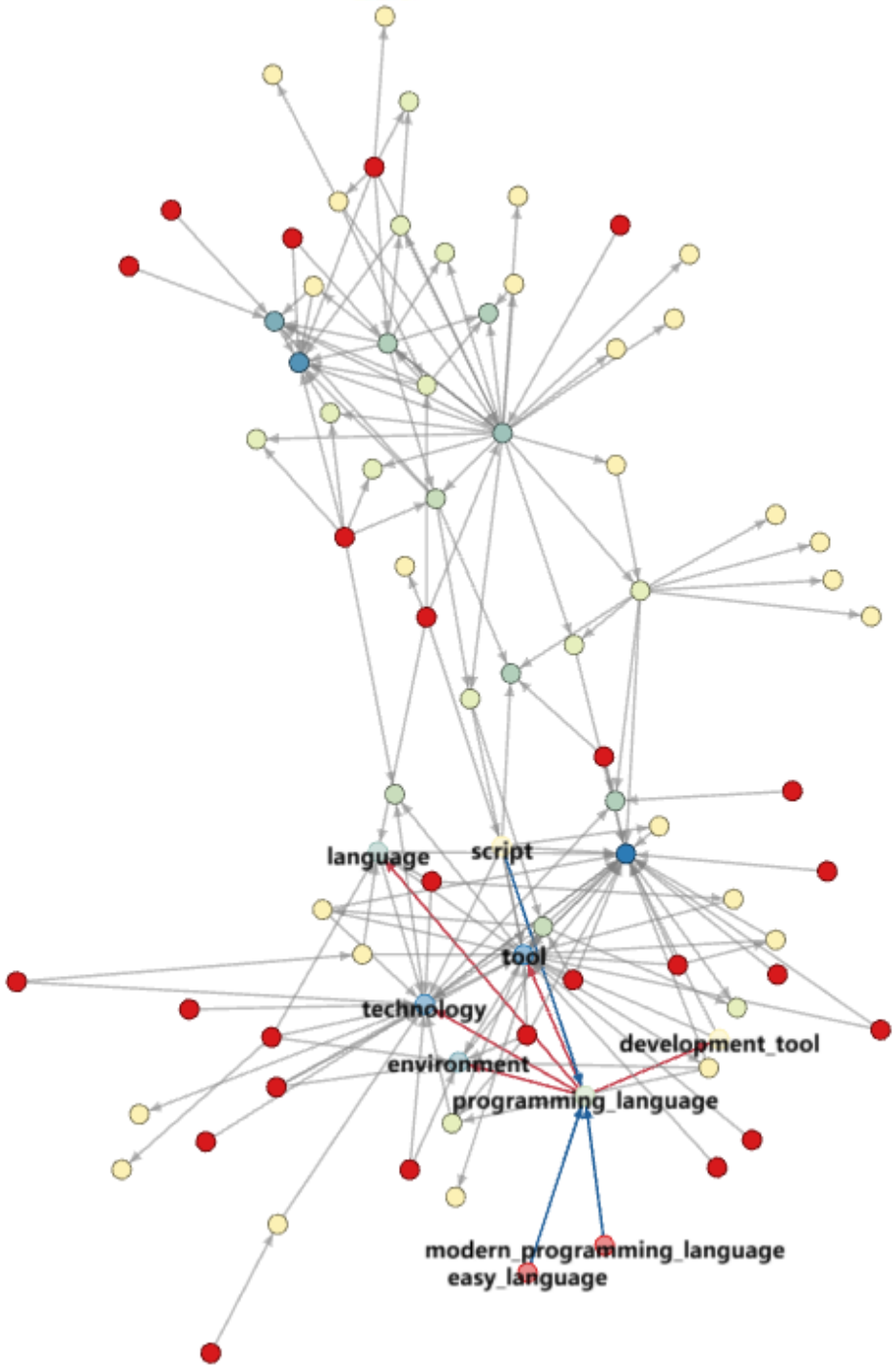
- All computing tasks, from those in a *calculator* to the *Turing Test*, face two tasks
  - Understand a question
  - Answer a question

- A biggest challenges of computing in the 21st century
  - Understand questions in natural language and answer them by knowledge inference
  - Understand user's intent and help him complete the task

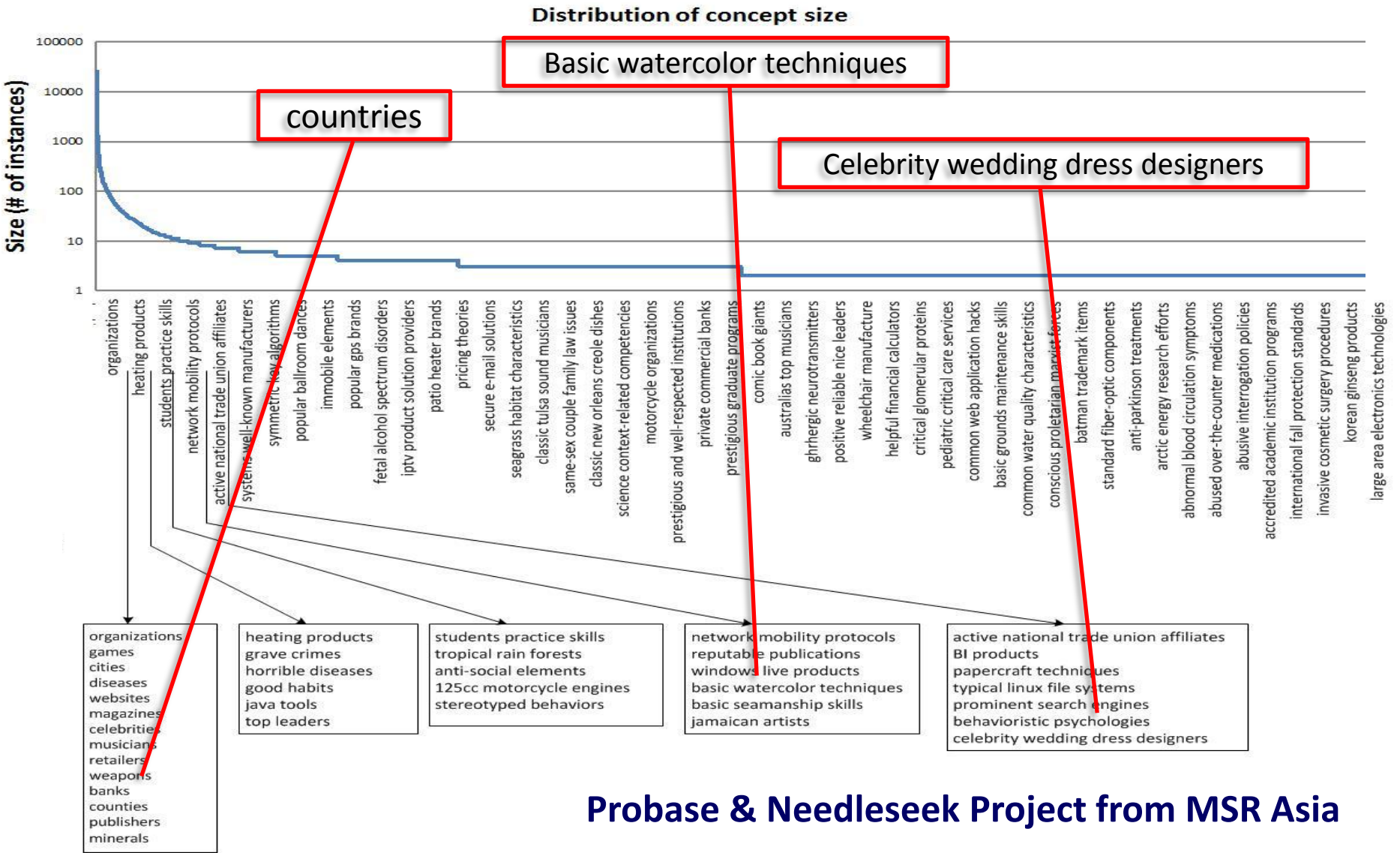
Building a Knowledge Graph



A semantic network for “python”

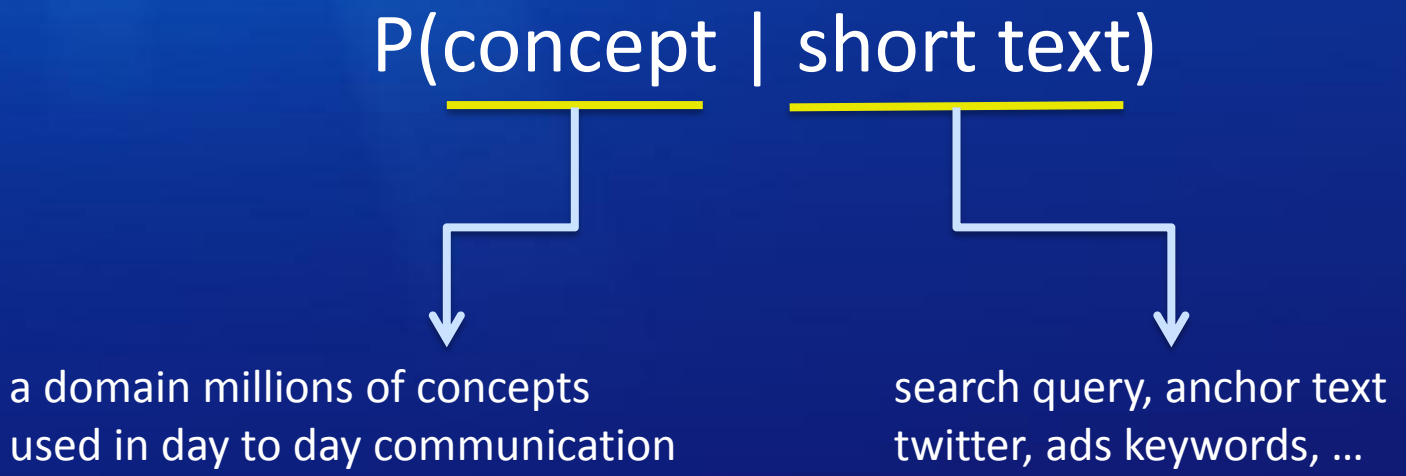


Distribution of Concept Size in Knowledge Base



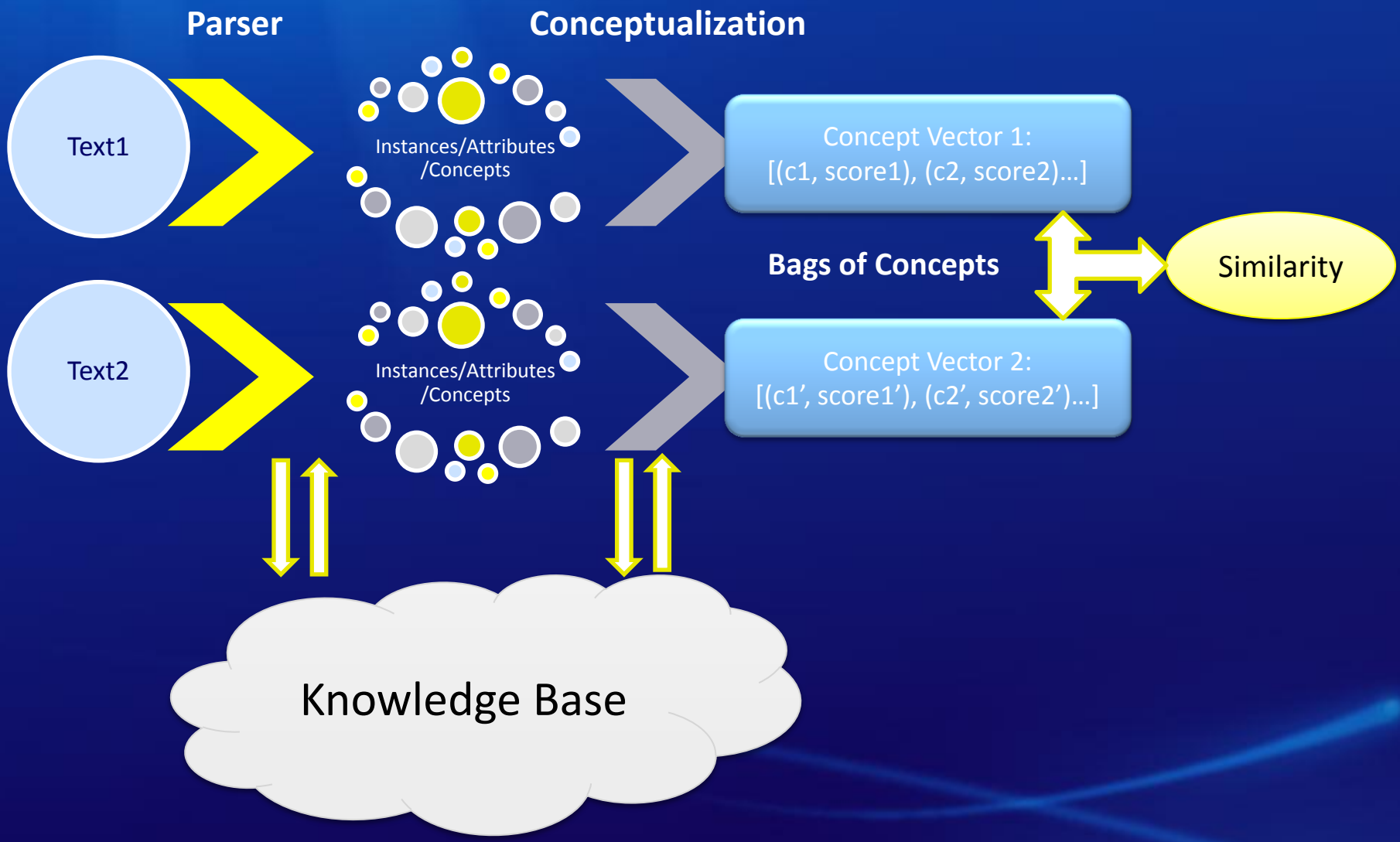
Probase & Needleseek Project from MSR Asia

Conceptualization of Short Text



Mapping a short text to a concept distribution

Semantic Matching of Two Short Texts



wedding band vs. band for wedding

Keyword 1:	wedding band
Keyword 2:	band for wedding

Step3: Similarity
Semantic Score: 0.009627336278342 58

wedding band:

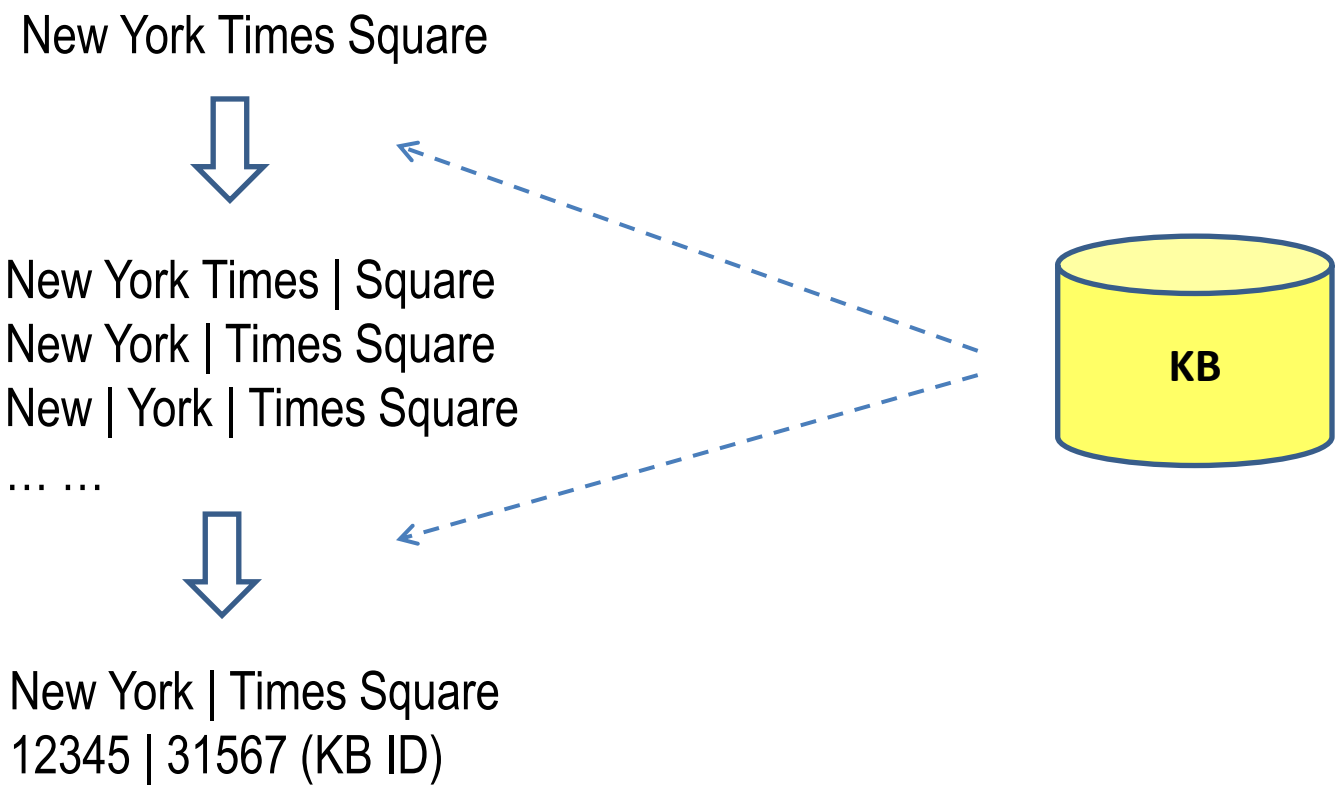
Cluster 1: Modifier: **False**	wedding band	plain ring	0.34
		single plain ring	0.04
		elevated camphorated oleandras	0.02
		status indicator	0.02
		sentimental item	0.02
		design instance	0.02
		consumer product	0.02
		briefest description	0.02
		tone jewelry	0.02
		hand jewelry	0.02

band for wedding:

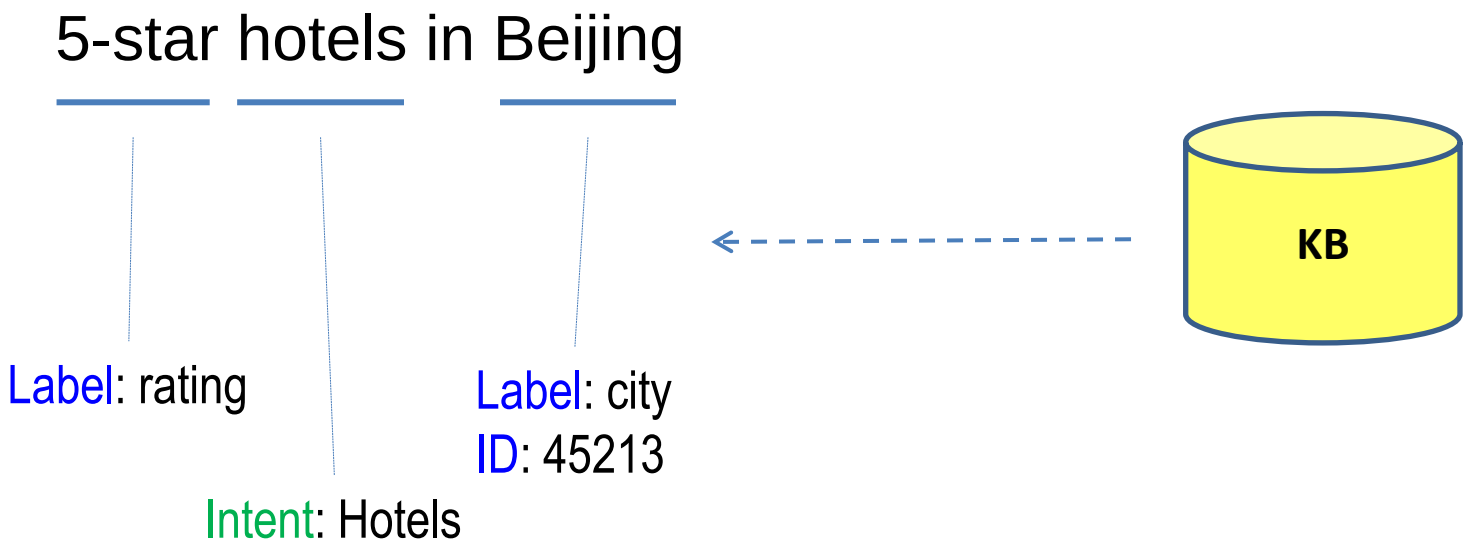
Cluster 1: Modifier: **False**	band for wedding	wedding arrangement service	1
Cluster 2: Modifier: **False**	wedding	celebration	0.0376035691523263
		ceremony	0.0331421287444232
		private event	0.0325047801147228
		private function	0.0286806883365201
		formal occasion	0.0261312938177183
		family event	0.0229445506692161
		formal event	0.0226258763543658
		large event	0.0184831102613129
		private party	0.0130656469088591
		big event	0.0127469725940089

little semantic overlap

Query Segmentation



Query Labeling & Intent Detection



QPE
Query Parsing and Exploration

safari in south africa

Query Segmentation & Labeling

safari		in		south africa	
activity	0.676			country	1.000
interest	0.502			destination	0.710
category	0.424			nation	0.675

QPE
Query Parsing and Exploration

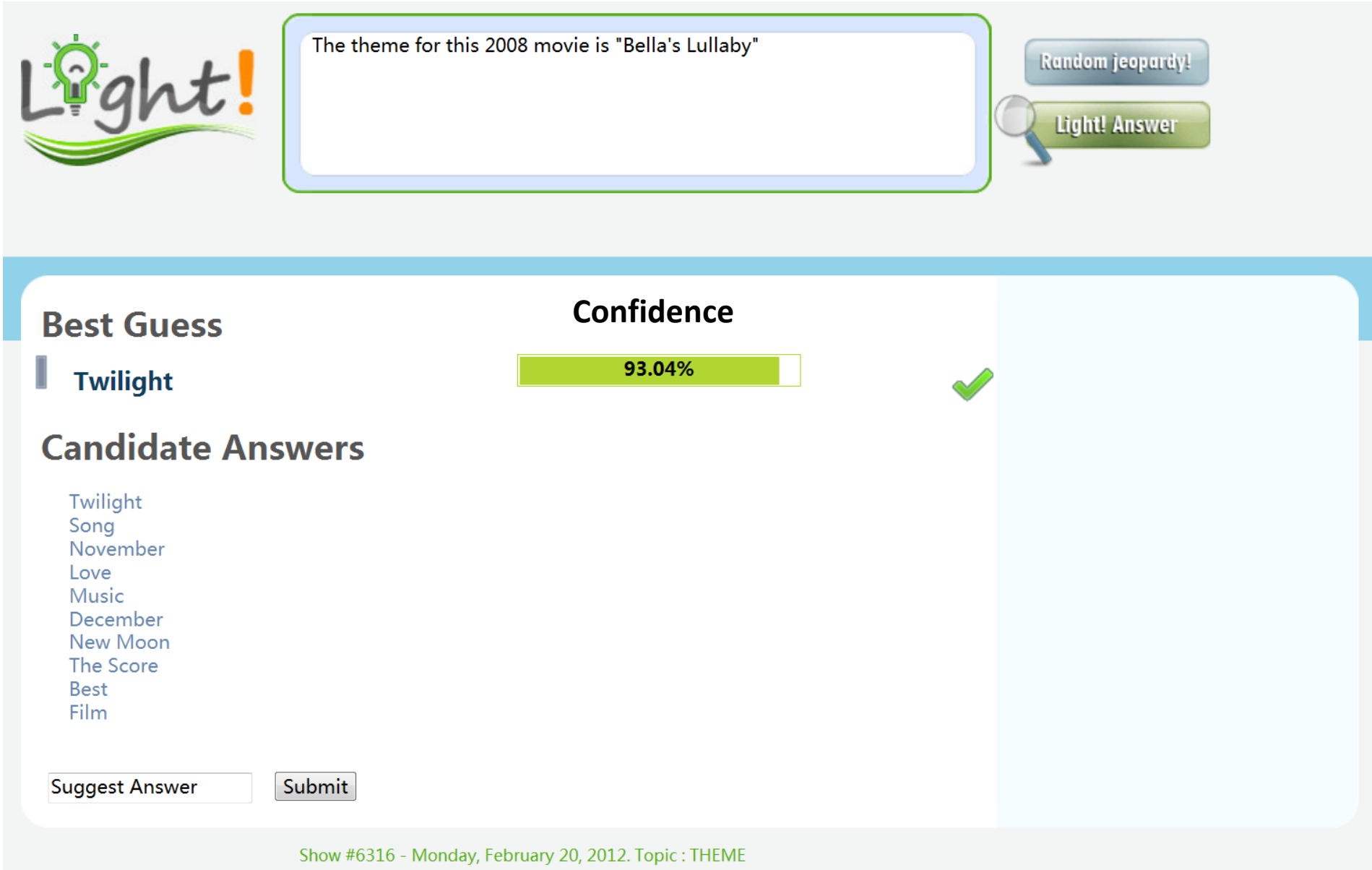
safari plugin download

Query Segmentation & Labeling

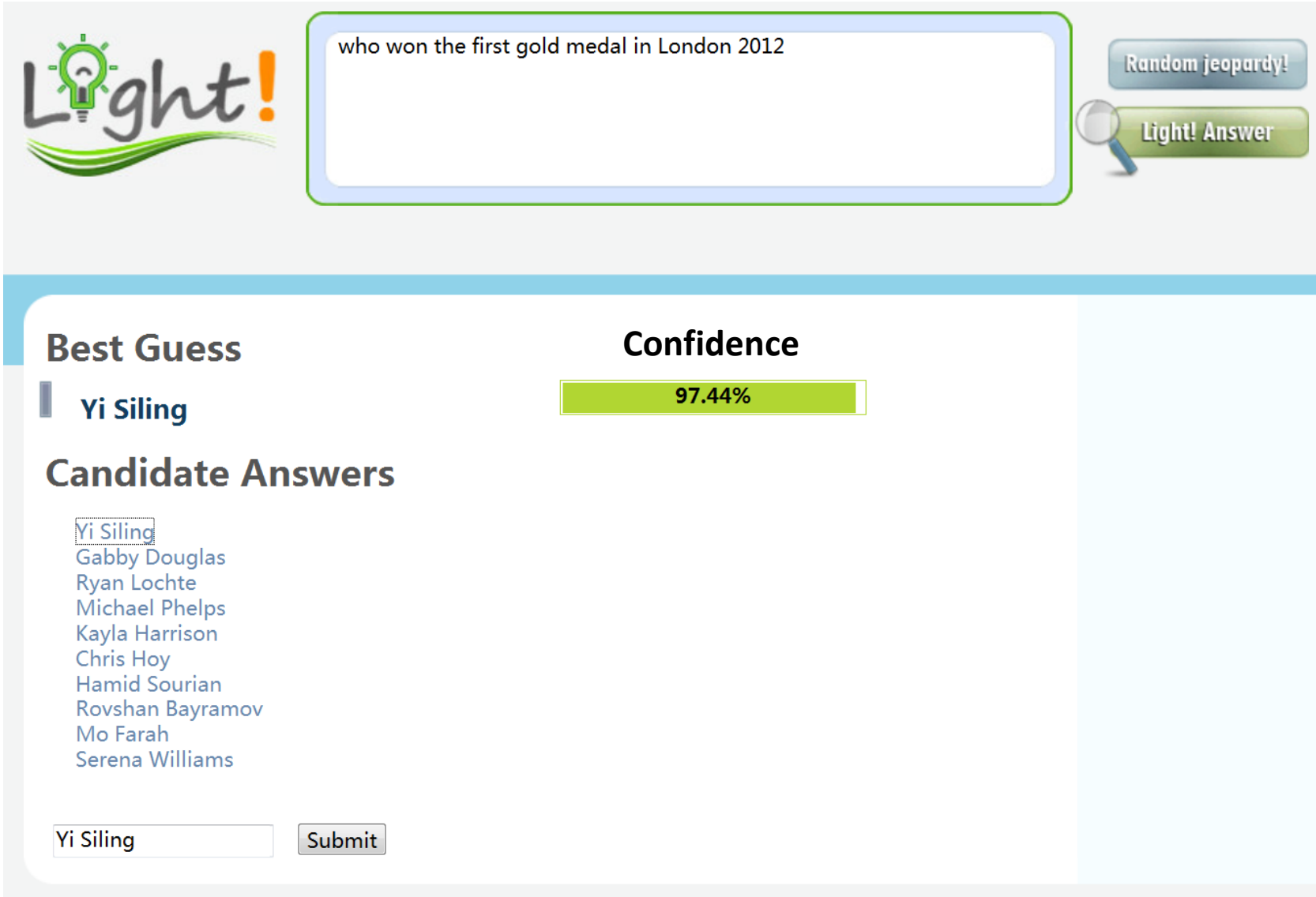
safari		plugin		download	
application	0.550	platform	0.460	service	0.636
program	0.523	product	0.459	feature	0.496
browser	0.446	system	0.384	product	0.427

QA:

The theme for this 2008 movie is "Bella's Lullaby"



QA:
Who won the first gold medal in London 2012



Summary

- Take search to a new level – semantic & knowledge-based search
- Take advantage of new trends and create a virtuous cycle – a new “Moore’s law” in knowledge engineering

Microsoft Research

Thank You!