

Empowering People with Knowledge – the Next Frontier for Web Search

Chin-Yew Lin
Microsoft Research Asia

Important Trends for Web Search

- Organize all information →
Address user's information need
- Semantic search
 - Keyword → Intent
 - Search what you type →
Search what you want
+ Social Search
- The cloud platform and developer ecosystem

Semantic Search

- Semantics?

- *Anything added to but not in the original text is semantics*
- *Anything fancier than indexing original strings*
- *Anything that has "language"*

– CIKM 2010, Gregory Grefenstette, Chief Science Officer, Exalead

- Q1: “5 star hotel in Beijing”
 - “5 star” => RATING
 - “hotel” => ACCOMMODATION
 - “Beijing” => CITY/LOCATION
 - ACCOMMODATION **HAS** RATING
 - ACCOMMODATION **IS_IN** LOCATION
 - **SELECT * FROM Hotel WHERE rating = 5 AND location = ‘Beijing’**
- Q2: “**Safari** plugin download” => Safari: SOFTWARE
- Q3: “**Safari** in South Africa” => Safari: ACTIVITY

Traditional Search

Semantic Search

Search what people TYPE ⇒ Search what people MEAN

Semantic Search

- Semantics?

Intent + Knowledge

- Q1: “5 star hotel in Beijing”
 - “5 star” => RATING
 - “hotel” => ACCOMMODATION
 - “Beijing” => CITY/LOCATION
 - ACCOMMODATION **HAS** RATING
 - ACCOMMODATION **IS_IN** LOCATION
 - **SELECT * FROM Hotel WHERE rating = 5 AND location = ‘Beijing’**
- Q2: “**Safari** plugin download” => Safari: SOFTWARE
- Q3: “**Safari** in South Africa” => Safari: ACTIVITY

Traditional Search

Semantic Search

Search what people TYPE ⇒ Search what people MEAN

Library Card Index



- **Search**
 - Paradigm: Query → Indexing → Documents
 - Query: book title, author name, ...
 - Indexing: inverted indices
 - Documents: books
- **Browsing**
 - Documents are organized into categories

The First Generation of Search Engines

- Essentially were invented to replace library card index
 - Based on information retrieval techniques
- **Search**
 - Paradigm: Query → Indexing → Documents → **Ranking**
 - Query: any words appearing in pages
 - Indexing: inverted indices
 - Documents: pages, images
 - Ranking: classical IR techniques + PageRank
- **Browsing**
 - Pages are organized into categories

Current Search Engines

- **Search (have not changed much)**
 - Paradigm: 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
- **Browsing**
 - Authoritative pages are organized into categories
- **Challenges: information explosion and information overload**
 - Index selection, index quality, and freshness
 - Relevance ranking (10 blue links)

Limits of Current Search Paradigm

- Current Search Paradigm:
Query → Indexing → Documents → Ranking
- Hypotheses
 1. I know my information need
 2. My information need can be expressed by keywords
 3. There are documents containing the information I want
 4. The documents contain the keywords
 5. The keywords are good indicators of the documents
 6. Ranking can put the documents at top n
- Hold well for library book search, but not as well for web search

The Next Frontier for Web Search

Enable people to gain knowledge and creativity from the web by computationally understanding user intent and matching that with published content, services, and people

- **Intent**
 - Computationally understand what the user is trying to accomplish
 - “Knowing” user needs, attitudes, and desires enables us to help the consumer better enrich their lives
- **Knowledge**
 - Computationally distill concepts and entities – such as people, places, products, businesses – and the relationship between them
 - Enable people and businesses to derive insights and knowledge from the web, and take actions
- **Semantic Search + Social Search**
 - Search what you mean not what you type (not only content, but services, and people)

Markets

Languages

...

OEM's

App's

UX

Ranking & Intent

**Indexing &
Knowledge**

Domains

Tasks

EcoSystem

**1st
page**

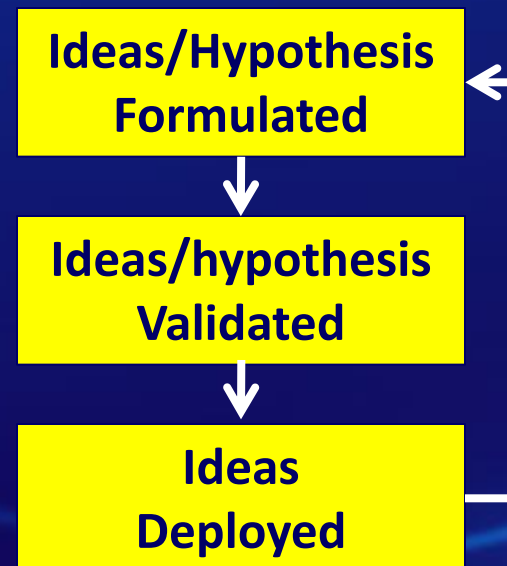
**2nd
page**

Search Platform

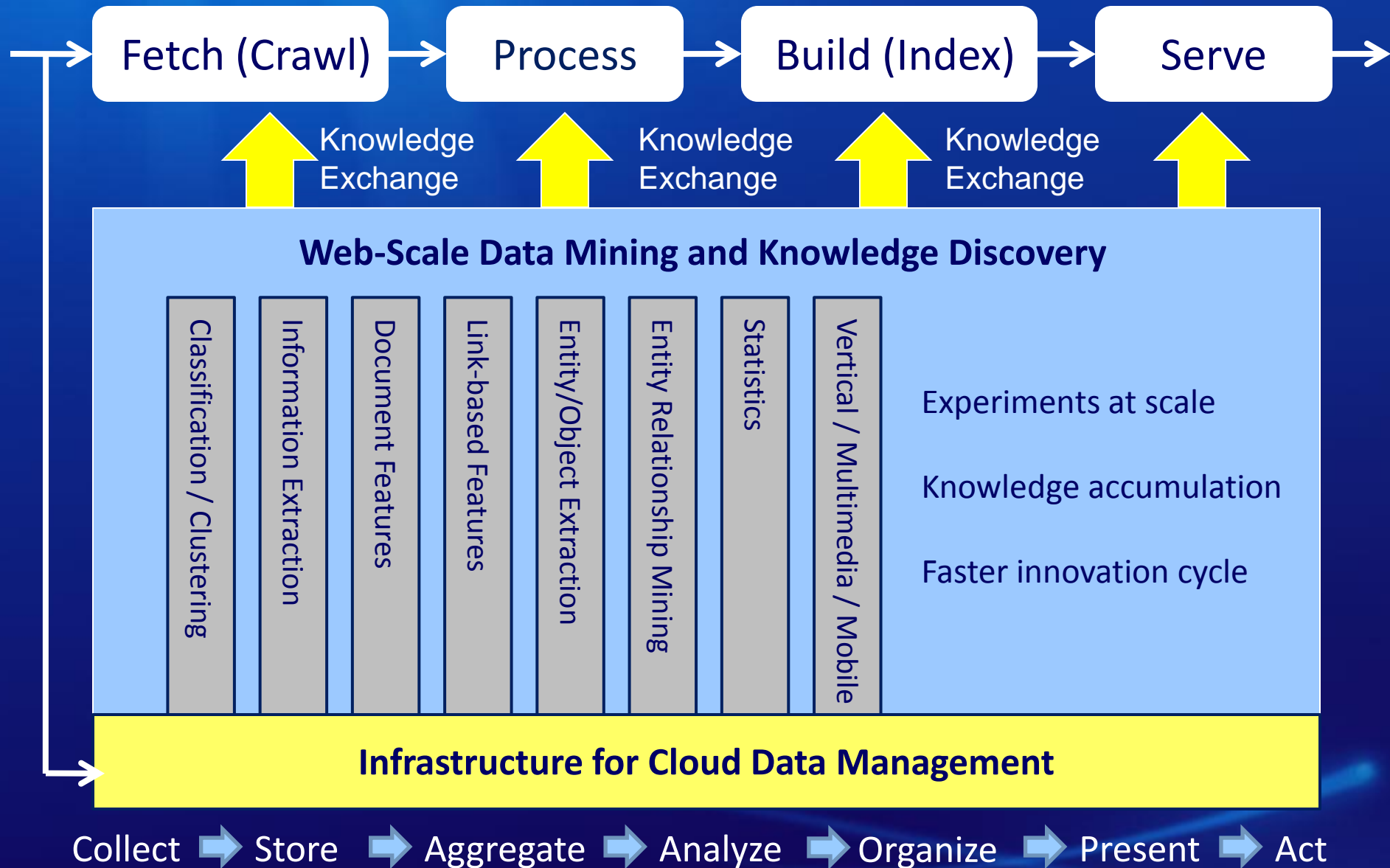
Infrastructure

Infrastructure for Web-scale Data Mining and Knowledge Discovery

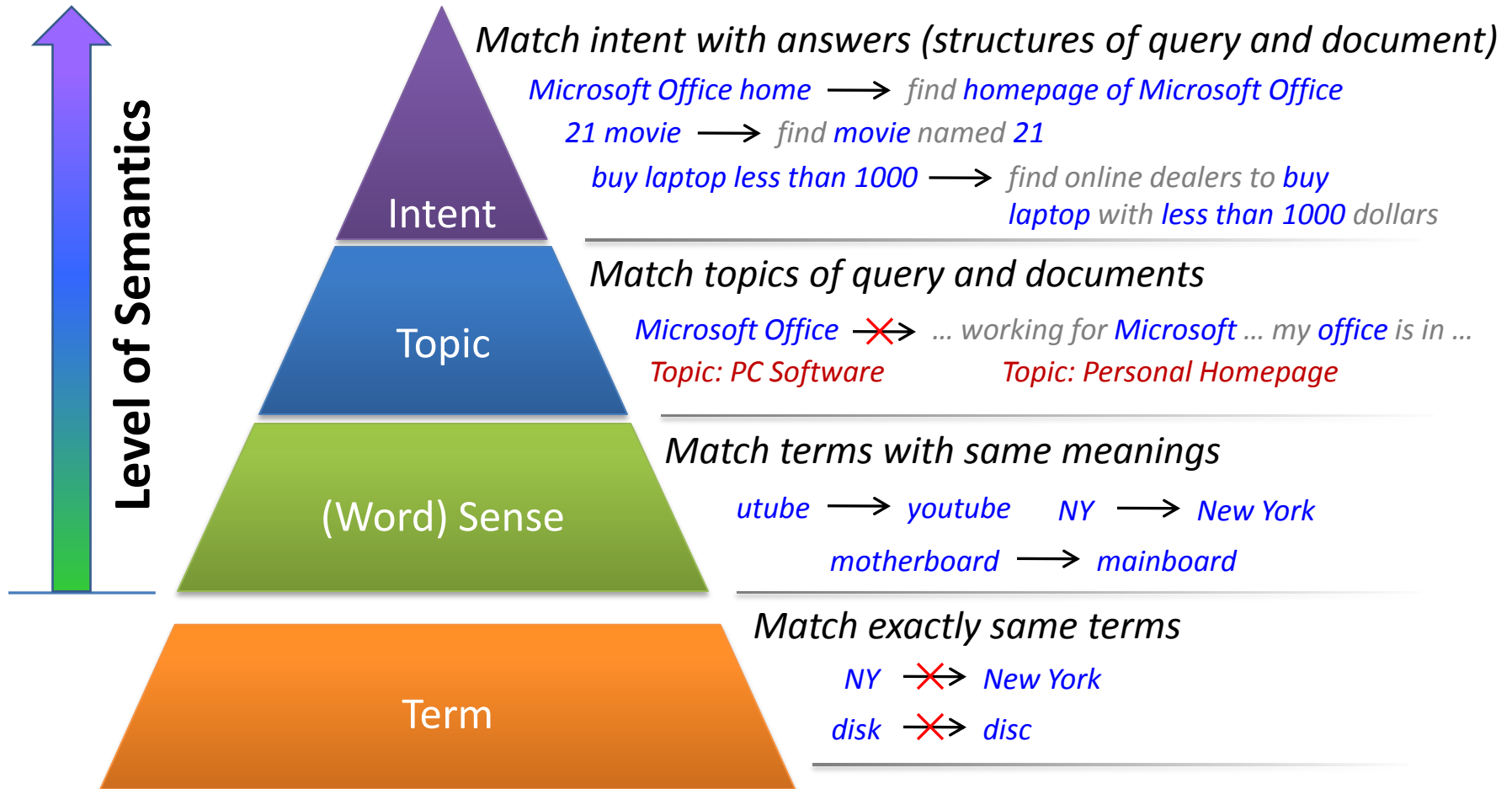
- **Deep understanding of data**
 - Data -> Information -> Knowledge & Intelligence
 - Queries -> Intent
 - Users -> Audience Intelligence -> Personalized & Targeted
- **Experiments at scale**
 - Offline experiments
 - Online experiments
 - Fast cycle of innovation



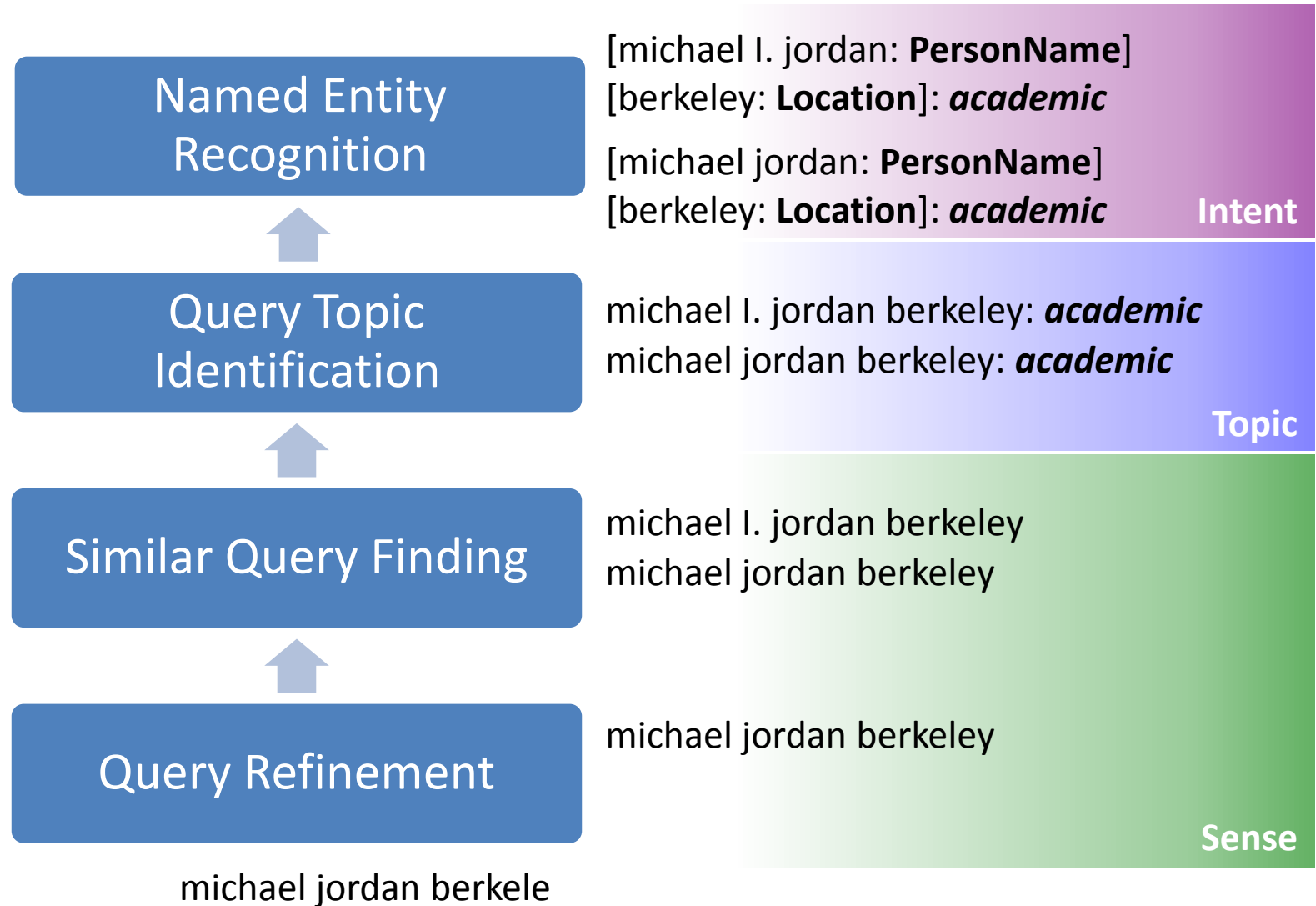
Search Infrastructure + Data Mining



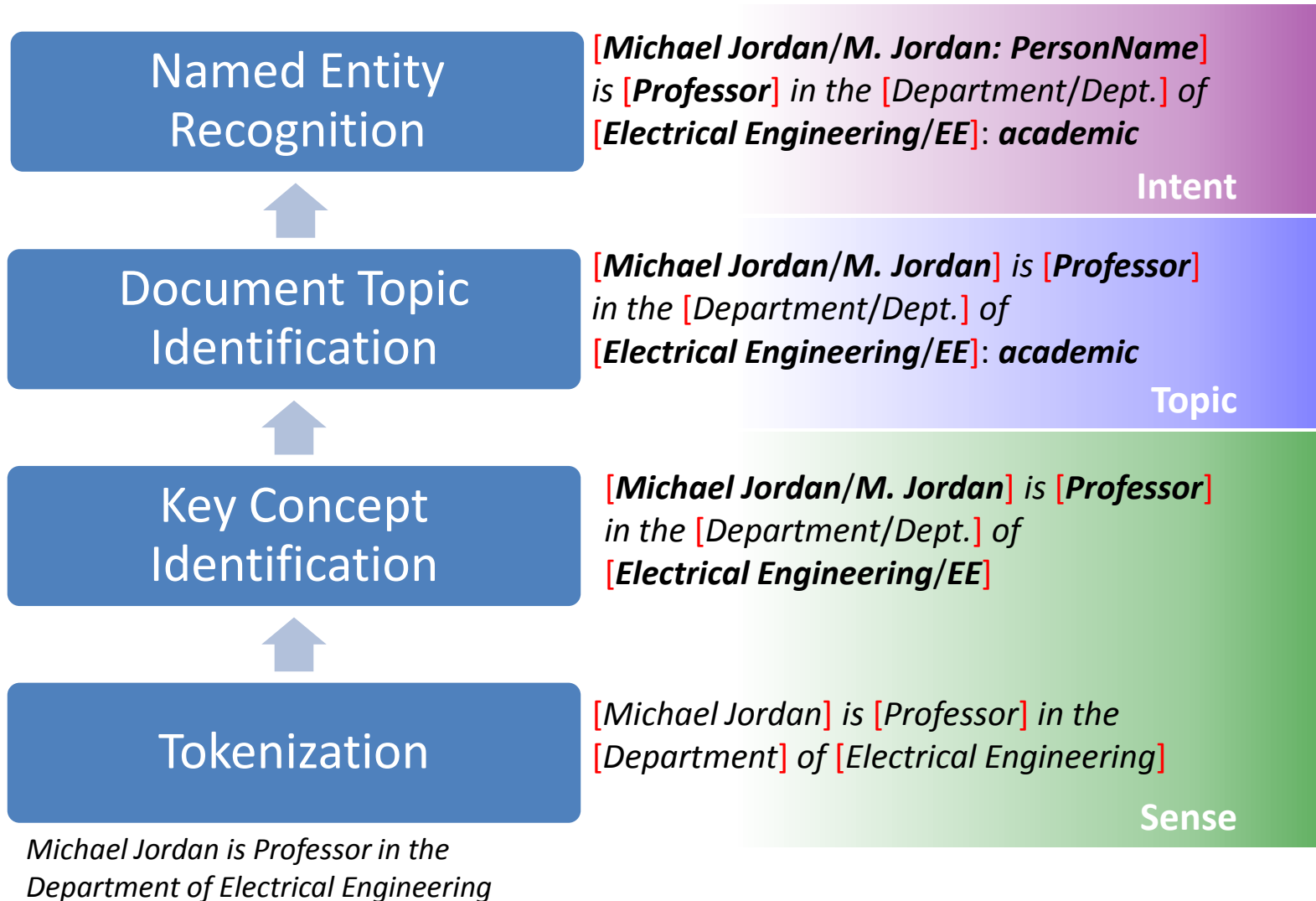
Different Levels of Semantic Matching



Query Processing (Online)



Document Processing (Offline)



Semantic Matching

[Michael I. Jordan's Home Page](#)
Models of visuomotor and other learning (Univ. of California, **Berkeley**, USA)
[www.cs.berkeley.edu/~jordan](#) · [Cached page](#) · [Mark as spam](#)

[Michael Jordan | EECS at UC Berkeley](#)
Michael Jordan Professor Research Areas Artificial Intelligence (AI) Biosystems & Computational Biology (BIO) Control, Intelligent Systems, and Robotics (CIR)
[www.eecs.berkeley.edu/Faculty/Homeworks/jordan.html](#) · [Cached page](#) · [Mark as spam](#)

[Publications](#)
Jordan. In M.-H. Chen, D. Dey, P. Mueller, D. Sun, and K. Ye (Eds.), *Frontiers of ...*
Technical Report 661, Department of Statistics, University of California, **Berkeley**, 2004.
[www.cs.berkeley.edu/~jordan/publications.html](#) · [Cached page](#) · [Mark as spam](#)

Results

Query
Representation

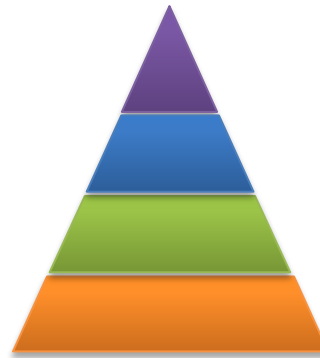
Semantic
Matching

Document
Representation

[michael i. jordan: **PersonName**]
[berkeley: **Location**]: *academic*

[michael jordan: **PersonName**]
[berkeley: **Location**]: *academic*

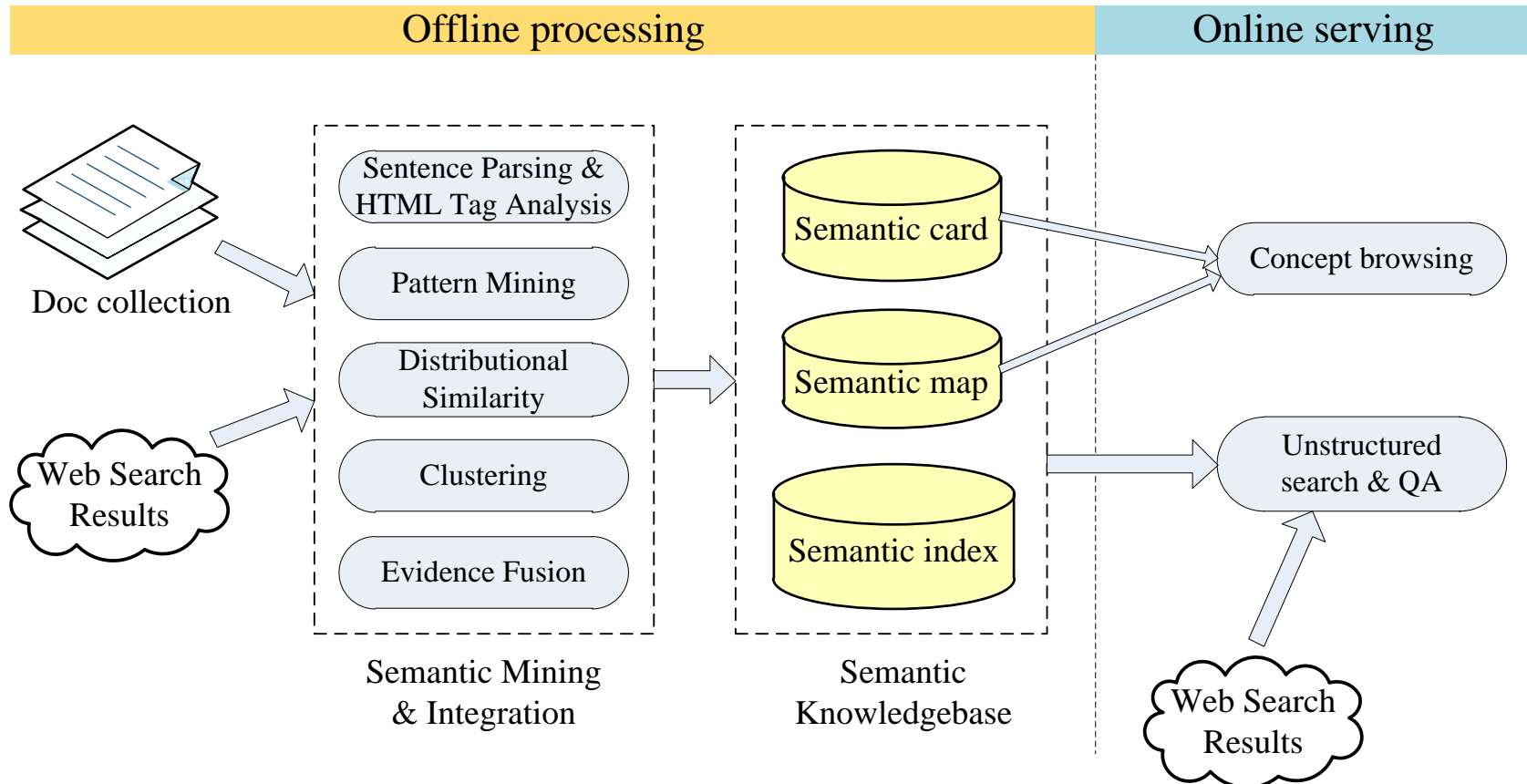
[**Michael Jordan/M. Jordan: PersonName**]
is [**Professor**] in the [**Department/Dept.**] of
[**Electrical Engineering/EE**]: *academic*



Matching can be conducted at different levels

Open-domain Semantic K Mining

- Mine semantic knowledge from web-scale data sources
- Answer & serve queries based on the mined semantic knowledge



Sempute NeedleSeek Knowledge-Base

- Source data
 - **Clue500**: 500 million English pages (3TB compressed)
- Knowledge-base Scale (Dec. 31st, 2010)
 - Instances: **30 million**
 - Links: **1 billion**
 - Categories & Subcategories: **10 million**
 - Attributes: **114,000** (from Clue050)
 - Size on disk: **100 GB**
- Demo
 - <http://needleseek.msra.cn> v2.2 Beta (external)

Sempute NeedleSeek Semantic Card

[Tools](#) [Datasets](#)



Sempute
NeedleSeek

Semantic Card Semantic Map Answer Bing Results

> **beijing** (city)

Basic information:

- city
- capital
- center
- city in china
- chinese city
- city of china



Attributes:

[Country]:	China
[Time zone]:	China Standard Time
[Postal code]:	100000 - 102629
[Area code]:	10
[GDP # Per capita]:	CNY 57,431
[Government # Mayor]:	Guo Jinlong
[Elevation]:	43.5 m
[Settled]:	c. 473 BC
[Mandarin # Postal Map]:	Peking
[Population # Municipality]:	17,430,000

Key sentences:

1. **Beijing** is the capital of New China and previously the capital for nine dynasties in Chinese history.
2. It is believed that **Beijing** was the largest city in the world from 1425 to 1650 and from 1710 to 1825.
3. **Beijing**, Jing for short, is the nation 's political, economic, cultural and educational center as well as China 's most important center for international trade and communications.
4. **Beijing** was the capital city of the Liao, Jin, Yuan, Ming and Qing Dynasties of China.
5. **Beijing** is the capital of the People 's Republic of China.

Related concepts:

[beijing](#) | [shanghai](#) | [guangzhou](#) | [nanjing](#) | [hangzhou](#) | [\[more\]](#)

2 Cards

city

province

Sempute NeedleSeek Semantic Map

[Tools](#) [Datasets](#)



Sempute
NeedleSeek

[Semantic Card](#) [Semantic Map](#) [Answer](#) [Bing Results](#)

Category: **City**

Items

Subcategories

Attributes



1. **beijing**

6. wuhan

11. chongqing

16. changchun



2. shanghai

7. shenyang

12. dalian

17. kunming



3. guangzhou

8. chengdu

13. zhengzhou

18. nanchang



4. nanjing

9. changsha

14. qingdao

19. fuzhou



5. hangzhou

10. jinan

15. tianjin

20. haikou

[TOP]

Category: **Province**

Items

Subcategories

Attributes



1. **beijing**

6. shandong

11. zhejiang

16. gansu



2. sichuan

7. hubei

12. anhui

17. shanxi



3. hunan

8. henan

13. jiangxi

18. liaoning



4. guangdong

9. hebei

14. guizhou

19. guangxi



5. jiangsu

10. fujian

15. yunnan

20. heilongjiang

[TOP]

Category: **City**

Items

Subcategories

Attributes



1. **beijing**

6. berlin

11. seoul

16. sydney



2. paris

7. tokyo

12. rome

17. taipei



3. london

8. istanbul

13. frankfurt

18. prague



4. moscow

9. madrid

14. bangkok

19. budapest



5. amsterdam

10. brussels

15. new york

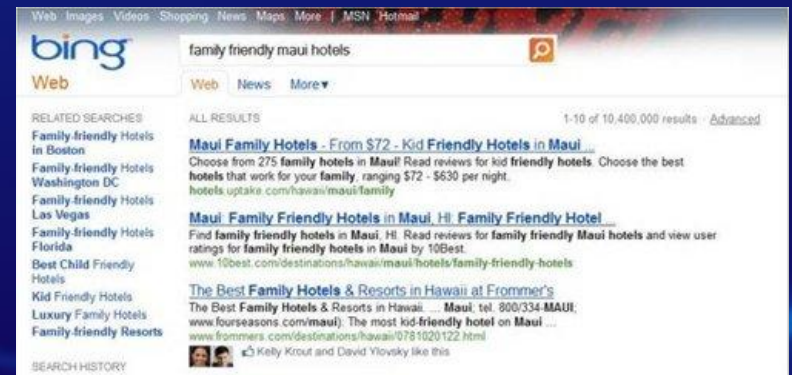
20. hong kong

[TOP]

- Images are generated by the [Glyphy](#) project.

Bing Social Search

- Friend Effect
 - 90% of people seek advice from family and friends as part of the their decision making process
 - 80% of people will delay making a decision until they can get a friend's stamp of approval
- Bing Social Experience
 - Trusted Friends
 - Collective IQ
 - Enabling Conversations





有人的地方就有关系

Web-scale entity extraction and summarization

- Mining implicit social information from web pages
- Mining relationships among people, organizations and locations
- Mining open domain semantic knowledge
- Renlifang (人立方) Search: <http://renlifang.msra.cn>
- EntityCube: <http://entitycube.research.microsoft.com/>
- Travel Guide: <http://travel.msra.cn/>
- Sempute NeedleSeek: <http://needleseek.msra.cn>

Knowledge about People



Knowledge about Places



Travel Guide 旅游指南

海滨度假





China

舟山市 (浙江省)

特色: 群岛 码头 海鲜 鸭蛋 岛屿 海岛 轮渡 沙滩 海水 渔港

查看更多

相关游记

舟山游记 - 普陀区六横...	2007/08/13 ▶
回归-我的05普陀山游记...	2008/07/15 ▶
08年11月普陀山四日扶...	2008/11/20 ▶
美丽舟山清爽游_舟山	2008/07/13 ▶
关于“Charming”...	2005/06/12 ▶

查看更多

相关旅游线路

美丽舟山清爽游_舟山	2008/07/13 ▶
------------	--------------

查看更多

相关问答

美丽舟山清爽游_舟山	2008/07/13 ▶
------------	--------------

导游图

进入目的地页面
收藏

去必应搜索:
机票、火车票、酒店

查看到: 街道 | 城市 | 区域

600 km

AVTEQ © 2009 Zenrin
Image courtesy of NASA

Knowledge about Everything

[Tools](#) [Datasets](#)



[Semantic Card](#) [Semantic Map](#) [Answer](#) [Bing Results](#)

> **seattle** ([city](#))

One Card

Basic information:

- city
- place
- area
- u.s. city
- metropolitan area
- american city



Attributes:

[Country]:	United States
[State]:	Washington
[Elevation]:	0-520 ft
[Time zone # Summer]:	PDT
[Area # Water]:	58.67 sq mi
[Population # Density]:	7,085/sq mi
[Area # Land]:	83.87 sq mi
[County]:	King
[GNIS feature ID]:	1512650
[FIPS code]:	53-63000

Key sentences:

1. Home to great coffee and Microsoft, **Seattle** is the largest city in the Pacific Northwest.
2. **Seattle** was the first city in the US to play a Beatles song on the radio.
3. **Seattle** is the largest city in the Pacific Northwest region of the United States.
4. Fortunately, **Seattle** was a very beautiful city and there was a lot to look at.
5. But outside the downtown core, **Seattle** is a city of neighbourhoods, each with its own, unique personality.

Related concepts:

[seattle](#) | [chicago](#) | [los angeles](#) | [san francisco](#) | [boston](#) | [\[more\]](#)

Summary

- Organize all information →
Address user's information need
- Semantic search **Intent + Knowledge**
 - Keyword → Intent
 - Search what you type →
Search what you want**+ Social Search**
- The cloud platform and developer ecosystem

Q&A