

# بازیابی اطلاعات

دکتر امین گلزاری اسکویی

[a.golzari@azaruniv.ac.ir](mailto:a.golzari@azaruniv.ac.ir)

[a.golzari@tabrizu.ac.ir](mailto:a.golzari@tabrizu.ac.ir)

<https://github.com/Amin-Golzari-Oskoue>



دانشگاه صنعتی ارومیه

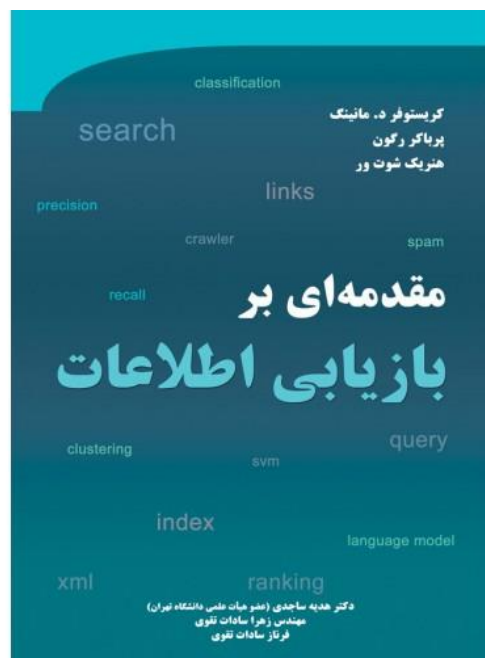
پاییز ۱۴۰۲

## مقدمه ای بر بازیابی اطلاعات

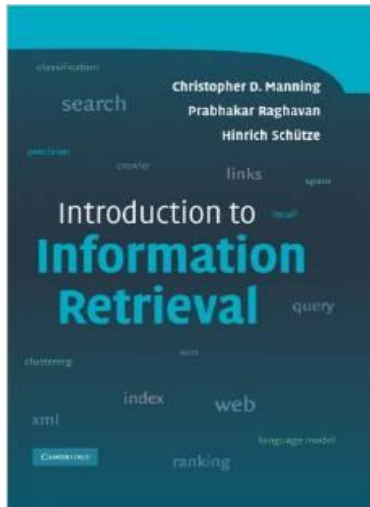
دکتر هدیه ساجدی، فرناز سادات تقوی،

زهرا سادات تقوی

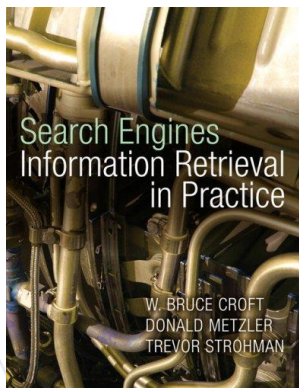
انتشارات نیاز دانش



## Text books

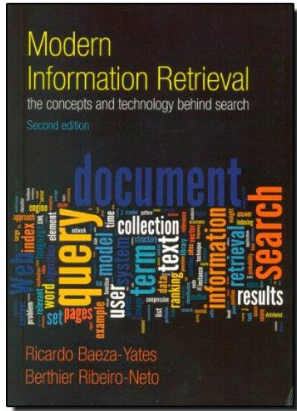


- ◎ *Introduction to Information Retrieval*. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

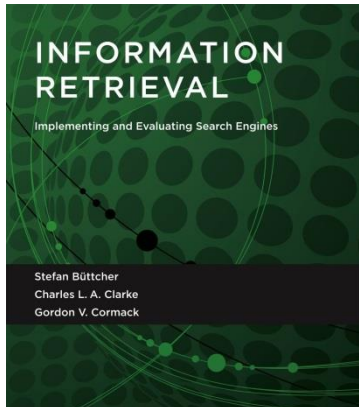


- ◎ *Search Engines: Information Retrieval in Practice*. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

## Text books




- ◎ ***Modern Information Retrieval***. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2011.



- ◎ ***Information Retrieval: Implementing and Evaluating Search Engines***. Stefan Böttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.

## نمرات و آزمون‌ها

نمرات (درصد)	فعالیت
5	حاضر و غایب 
25	تکالیف و پروژه‌ها 
10	ارائه‌ها 
60	امتحان نهایی 

در صورت تقلب در پروژه‌ها، تکالیف و ارائه‌ها (حتی یکبار و در یکی از این موارد) نمره نهایی ۹ ثبت خواهد شد. 

## دستیاران



عرفان فیاطی

دانشجو سال آخر علوم کامپیوتر  
موزه یادگیری ماشین و یادگیری عمیق



سوگل یگانه

دانشجو سال آخر علوم کامپیوتر  
موزه یادگیری ماشین و یادگیری عمیق



یونس فورابلو

دانشجو سال آخر مهندسی کامپیوتر  
موزه یادگیری ماشین، یادگیری عمیق  
و برنامه‌نویسی پایتون



امیرمسین میاتی

دانشجو سال دوم مهندسی کامپیوتر  
موزه یادگیری ماشین، یادگیری عمیق  
و برنامه‌نویسی پایتون



محمّد ظهیری

دانشجو سال سوم مهندسی کامپیوتر  
موزه ی برنامه‌نویسی پایتون



# فصل ۱

## مقدمه

مطالب این فصل

بازیابی اطلاعات چیست؟

تاریخچه بازیابی اطلاعات



## بازیابی اطلاعات چیست؟

The image shows a Google search results page for the query "what is information retrieval". The search bar at the top shows the query and the Google logo. Below the search bar, there are tabs for "Web", "Videos", "Images", "News", "Shopping", and "More". The search results show "About 14,300,000 results (0.43 seconds)". The first result is a definition of information retrieval, which is highlighted with a red box. A red line points from this box to a second red box that highlights a specific definition of information retrieval from the Stanford NLP. Another red line points from the second box to a third red box that highlights a definition of information retrieval from Merriam-Webster.

bing Google what is information retrieval

Web Videos Images News Shopping More Search tools

About 14,300,000 results (0.43 seconds)

in·for·ma·tion re·triev·al

noun COMPUTING

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems...

based on metadata or on full-text (or other content-based) indexing.  
Category: Information retrieval - Relevance - Human-computer information ...

[PDF] Introduction to Information Retrieval - The Stanford NLP  
nlp.stanford.edu/IR-book/pdf/01bool.pdf  
Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Information retrieval  
www.iva.dk/.../inf... The Royal School of Library and Information Science  
Oct 15, 2006 - Information retrieval (IR). The term IR may be considered a research field, but it may also be considered a research tradition (or rather a set of ...

Information Retrieval - Merriam-Webster  
www.merriam-webster.com/.../information%20retrieva... Merriam-Webster  
the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system. ADVERTISEMENT ...



## بازیابی اطلاعات – تعریف

بازیابی اطلاعات یافتن مواد (معمولا اسناد) از یک ماهیت بدون ساختار (معمولا متن) است که یک نیاز اطلاعاتی را از داخل مجموعه های بزرگ (که معمولا در کامپیوتر ذخیره می شوند) برآورده می کند.

فعالیت دستیابی به منابع اطلاعاتی مرتبط با یک نیاز اطلاعاتی از یک مجموعه منابع اطلاعاتی

بازیابی اطلاعات با نمایش، ذخیره سازی، سازماندهی، و دسترسی به آیتم های اطلاعاتی از قبیل سندها، صفحات وب، رکوردهای ساخت یافته و نیمه ساخت یافته و آبجکت های مالتی مدیا سروکار دارد.

نمایش و سازماندهی آیتم های اطلاعاتی باید بگونه ای باشد که تا به کاربران امکان دسترسی آسان به اطلاعات مورد نظرشان را بدهد.

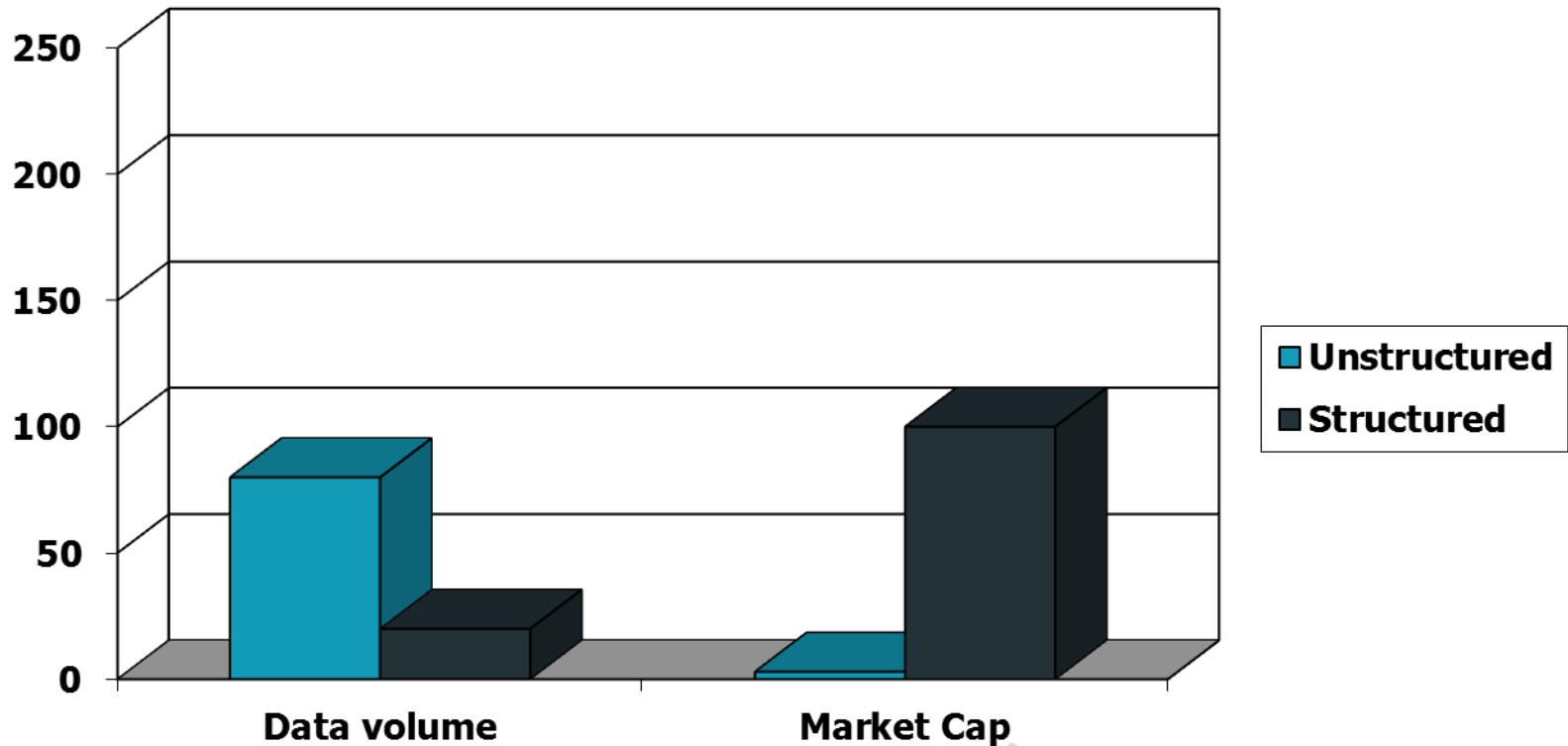
مثال

## علاوه بر جستجوی وب

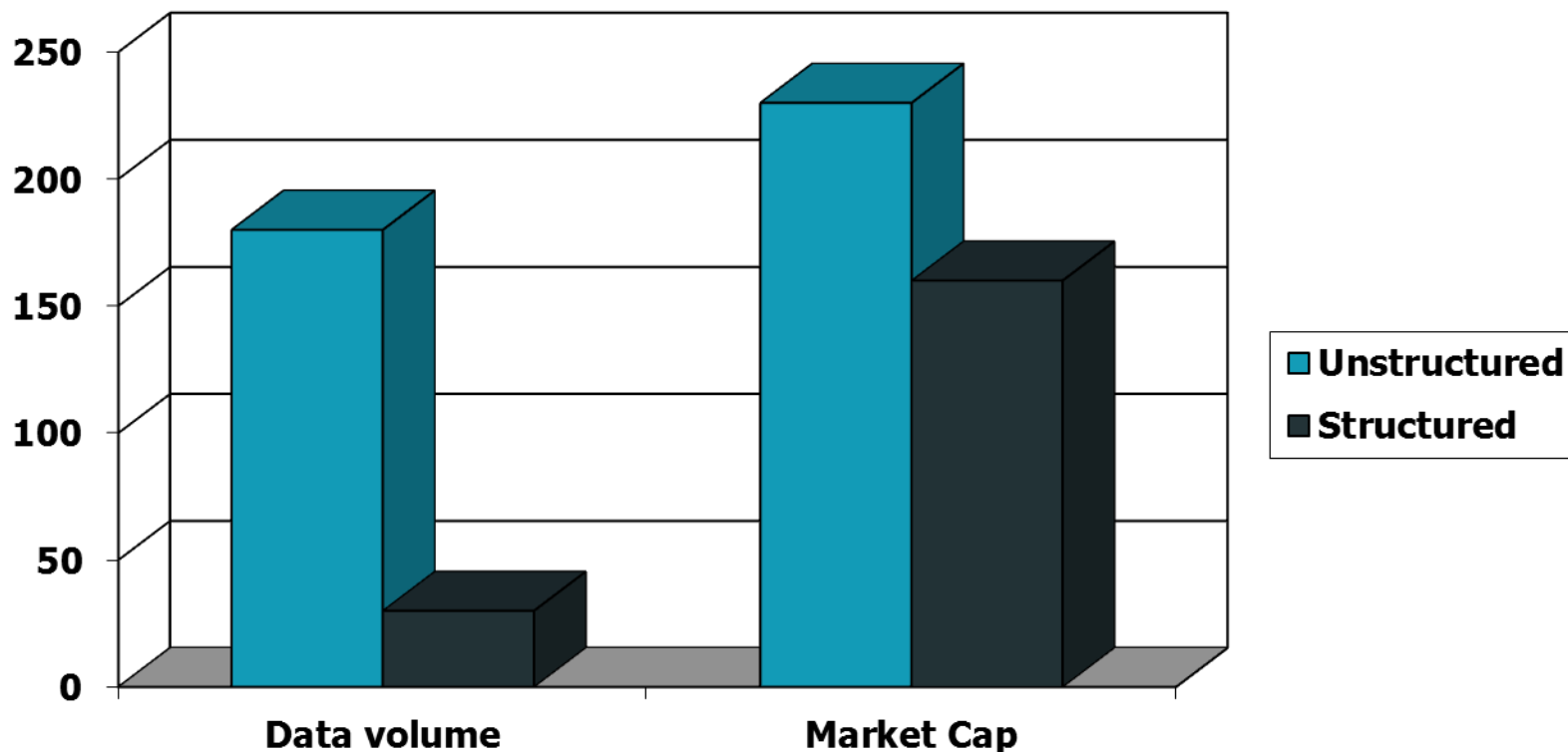
سایر کاربردها:

- E-mail search
- Searching your laptop
- Corporate knowledge bases
- Legal information retrieval

# مقایسه بین داده های ساخت یافته و غیر ساخت یافته در اواسط دهه ۹۰



# مقایسه بین داده های ساخت یافته و غیر ساخت یافته – امروز



## تفاوت بین بازیابی داده و بازیابی اطلاعات

هدف اصلی یک سیستم IR بازیابی اطلاعاتی مرتبط (Relevant) با نیاز کاربر است.  
○ تاکید روی بازیابی اطلاعات است نه داده.

◎ یک سیستم IR باید **محتوای** یک آیتم اطلاعاتی را **تفسیر** کند و آیتم های اطلاعاتی را بر اساس درجه ارتباط (relevance) آن ها با query کاربر رتبه بندی (Rank) کند.

## تفاوت بین بازیابی داده و بازیابی اطلاعات

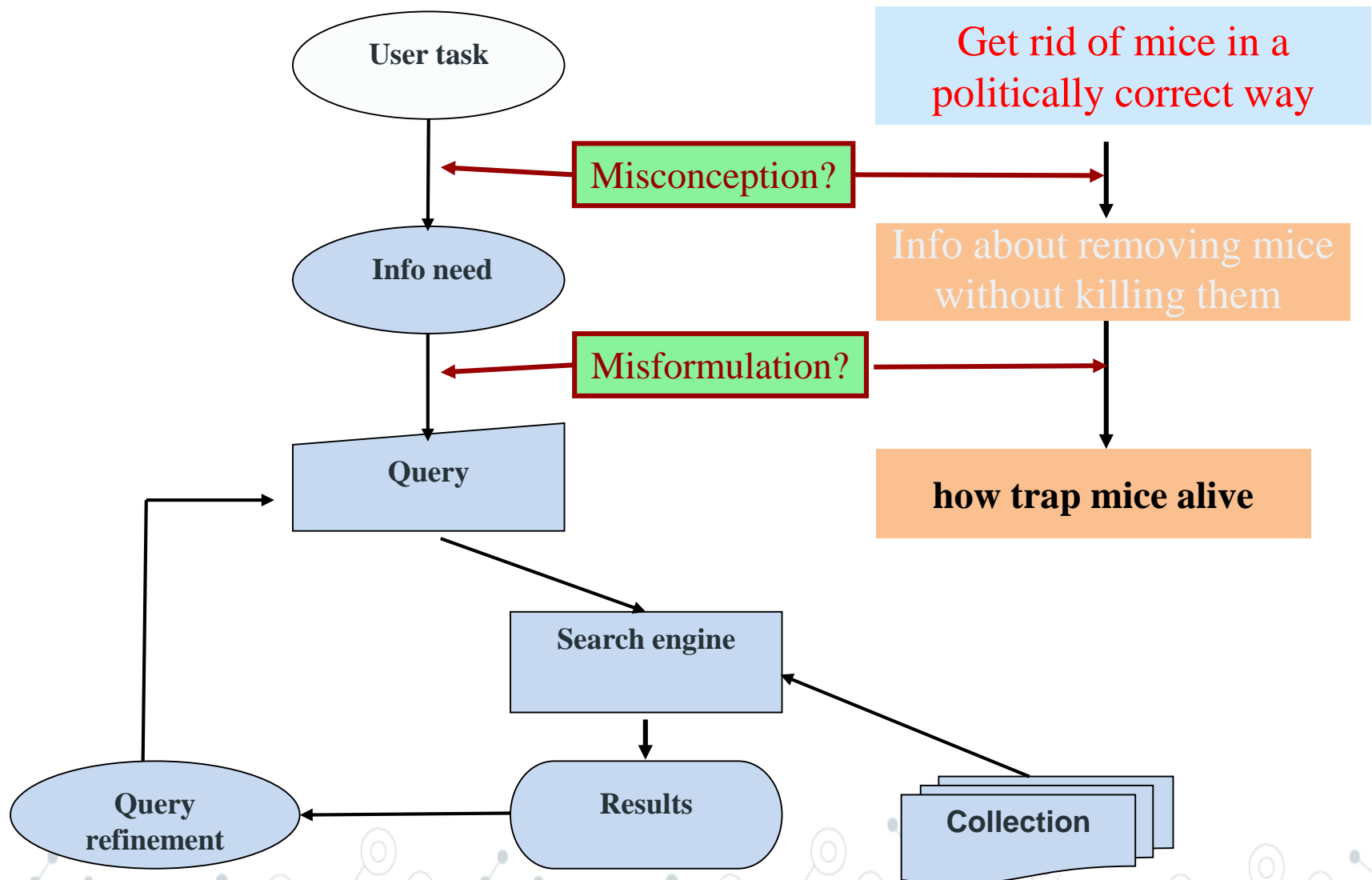
**بازیابی داده:** کدام سند دارای کلمه های استفاده شده در پرس وجوی کاربر است.

- شرط های تحریف شده به طور مشخص.
- یک خطا یا اشتباه در query باعث خطا در سیستم می شود.
- ساختار و سمانتیک خوش تحریفی دارد.

## بازیابی اطلاعات: ◎

- ساختار مشخصی ندارد و از لحاظ سمانتیکی ابهام وجود دارد
- خطای کوچک قابل صرف نظر است

## مدل جستجوی کلاسیک



## چرا بازیابی اطلاعات مورد نیاز است

**سربار اطلاعاتی:** "اشاره به سختی درک یک موضوع و تصمیم گیری یک فرد که به دلیل وجود اطلاعات خیلی زیاد رخ می‌دهد.

- *"It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information."* - wiki





## چرا بازیابی اطلاعات مورد نیاز است

### سربار اطلاعاتی

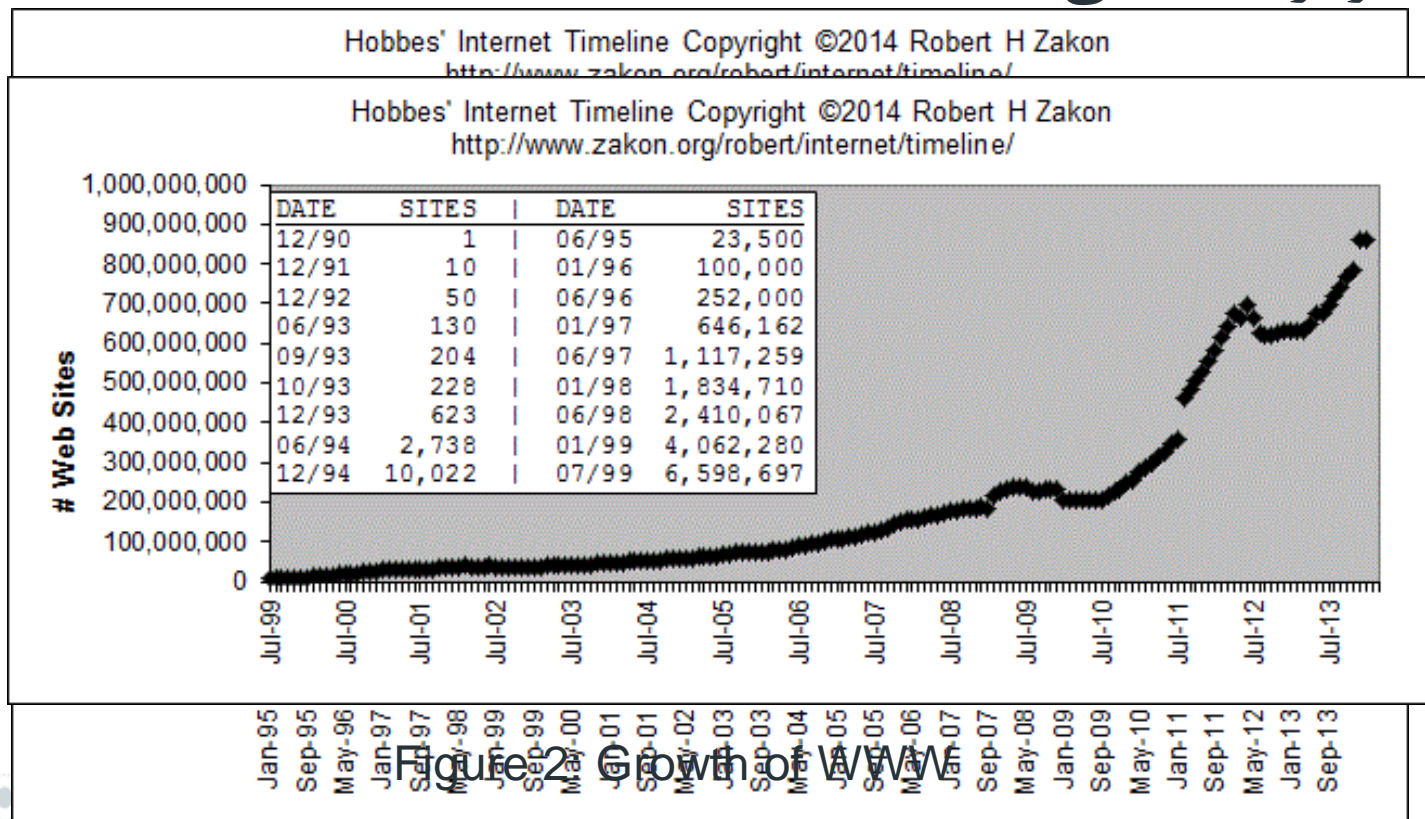


Figure 1: Growth of Internet

## چرا بازیابی اطلاعات مورد نیاز است

پردازش داده های غیرساخت یافته  
داده ساخت یافته: داده های ذخیره شده در یک جدول پایگاه داده.  
مجموع داده های غیرساخت یافته خیلی زیاد است.

متن، تصویر، صوت، ویدئو  
"85 درصد اطلاعات شرکت ها به صورت داده های غیرساخت یافته است"

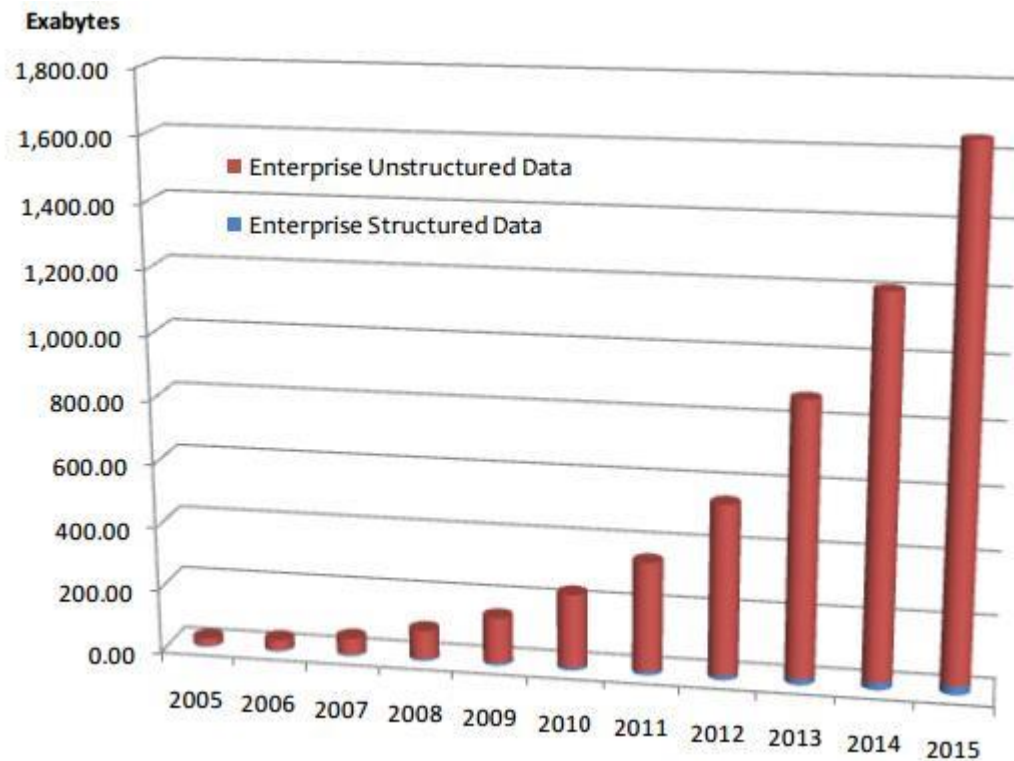
© "85 percent of all business information exists as unstructured data" -  
Merrill Lynch

داده های غیرساخت یافته دارای سمانتیک ناشناخته ای هستند.

Table 1: People in CS Department

ID	Name	Job
1	Jack	Professor
3	David	Stuff
5	Tony	IT support

## چرا بازیابی اطلاعات مورد نیاز است



Total Enterprise Data Growth 2005-2015, IDC 2012

## چرا بازیابی اطلاعات مورد نیاز است

بازیابی اطلاعات یک ابزار اساسی برای مواجهه با سربار اطلاعاتی است.



You are  
here!

## رشد تحقیقاتی IR از گذشته تا به حال

- ◎ Early days (late 1950s to 1960s): foundation of the field
  - Luhn's work on automatic indexing
  - Cleverdon's Cranfield evaluation methodology and index experiments
  - Salton's early work on SMART system and experiments
- ◎ 1970s-1980s: a large number of retrieval models
  - Vector space model
  - Probabilistic models
- ◎ 1990s: further development of retrieval models and new tasks
  - Language models
  - TREC evaluation
  - Web search
- ◎ 2000s-present: more applications, especially Web search and interactions with other fields
  - Learning to rank
  - Scalability (e.g., MapReduce)
  - Real-time search



## تاریخچه بازیابی اطلاعات

### شتاب دهنده

- Academia: Text Retrieval Conference (TREC) in 1992

● هدف آن حمایت از تحقیقات در حوزه بازیابی اطلاعات با فراهم آوردن زیرساخت لازم جهت ارزیابی متدولوژیهای بازیابی متن در مقیاس بسیار بزرگ.

● تقریباً یک سوم پیشرفت ها در حوزه موتورهای جستجوی وب از 1999 تا 2009 به TREC منتسب است.

● هم اکنون هم به عنوان یکی از test-bed های اصلی در زمینه IR است.

### شتاب دهنده

موتور جستجوی وب:

بهبود موتورهای جستجوی وب : توسط شرکتها

WWW باعث رشد انفجارگونه محتوا شد و بسیاری از تکنیک های IR ابداع شد.

### اولین موتور جستجوی وب

First web search engine: “*Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that periodically mirrored these pages and rewrote them into a standard format.*” Sept 2, 1993

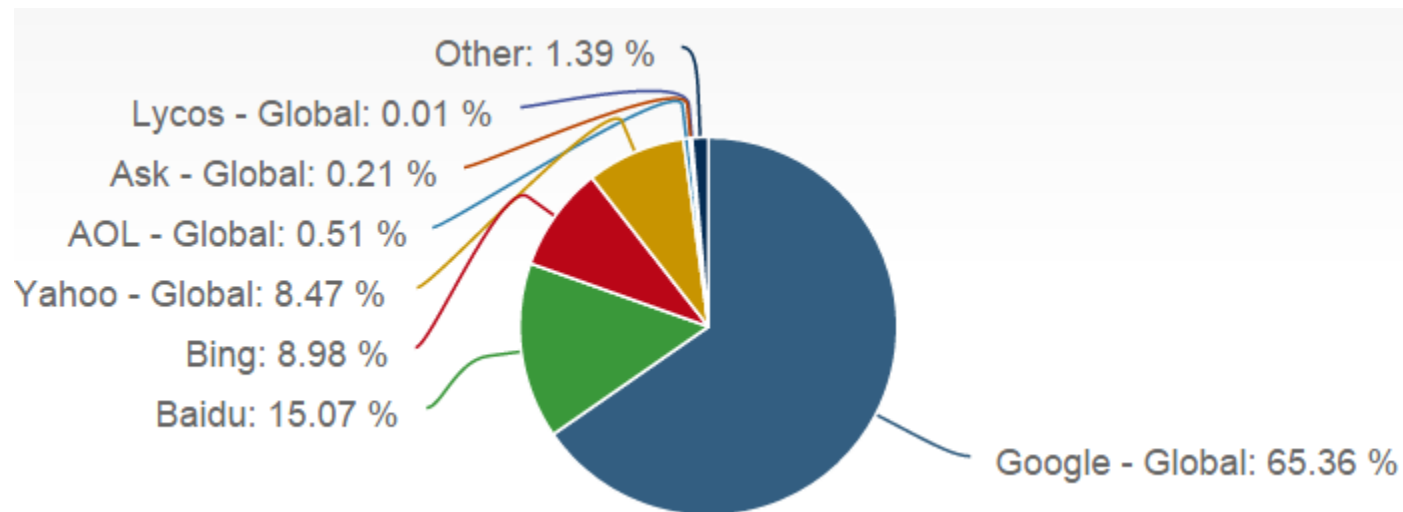
Lycos (started at CMU)

در سال 1994 به عنوان یک موتور جستجوی تجاری توسعه داده شد

© **Booming of search engine industry:** *Magellan, Excite, Infoseek, Inktomi, Northern Light, AltaVista, Yahoo!, Google, and Bing*

## بازیگران اصلی حوزه جستجوی وب

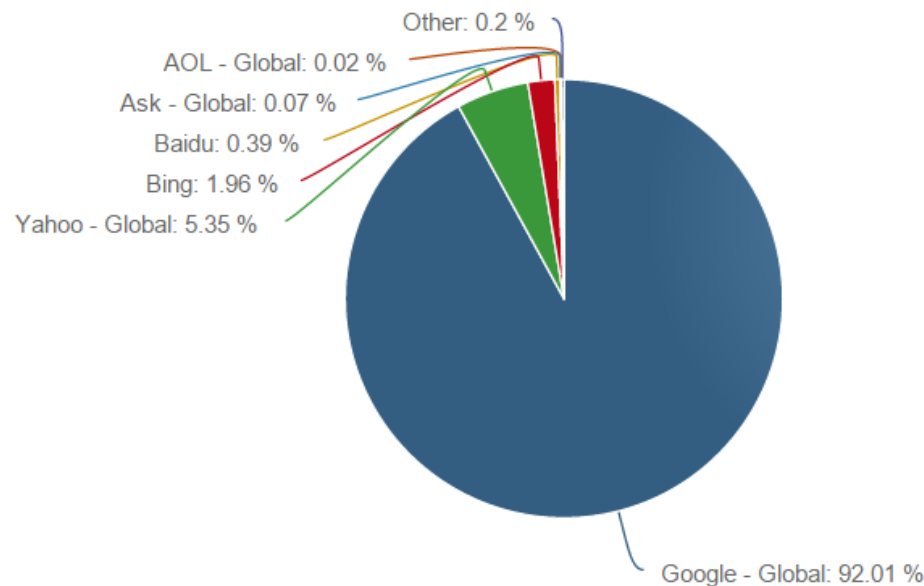
بازار موتور جستجو گوگل - بخش دستکاپ



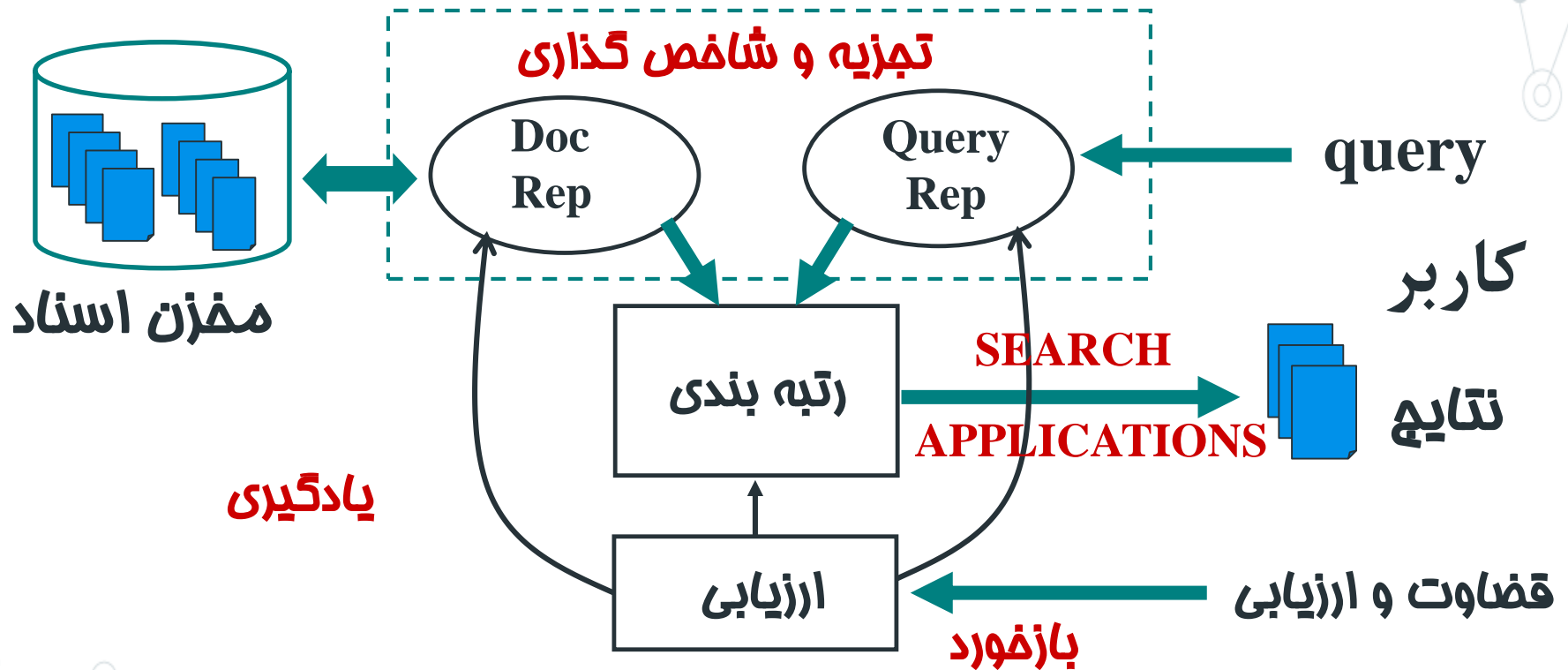


## بازیگران اصلی حوزه جستجوی وب

بازار موتور جستجو گوگل – بخش دستگاه های موبایل



## چگونه بازیابی اطلاعات انجام بدهیم؟



## موتور های جستجوی کنونی

Google

uva

تقاضای درک پرس وجو



Web

Maps

Images

News

Shopping

More ▾

Search tools



About 103,000,000 results (0.65 seconds)

تقاضا برای کارایی

**The University of Virginia**

[www.virginia.edu/](http://www.virginia.edu/) University of Virginia ▾

The University of Virginia in Charlottesville, VA was founded in 1819 by Thomas Jefferson. The

4.9 ★★★★★

**University**

[en.wikipedia.org](https://en.wikipedia.org/wiki/University_of_Virginia)

The University

research univer

**University**

[colleges.usne](http://colleges.usnews.edu/rankings/national-college-rankings/2014/2014-2015)

Is University of

University of

**VIRGINIAS**

[www.virginia.edu](http://www.virginia.edu/)

The University

Network. The n

تقاضا برای صحت

Google



Google Search

I'm Feeling Lucky

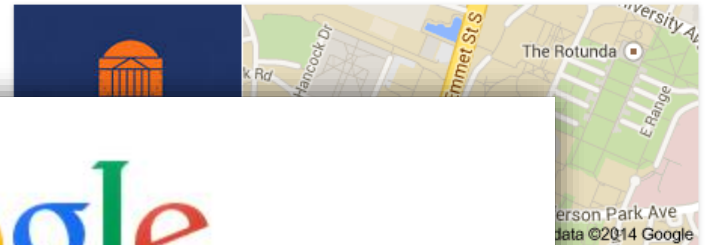
Images for university of virginia

Report images



More images for university of virginia

تقاضا برای راحتی



Directions

Charlottesville,  
U.S. President

**Mascot:** University of Virginia Cavalier

**Founder:** Thomas Jefferson

**Founded:** 1819, Charlottesville, VA

**Colors:** Blue, Orange

**Recent posts**

#UVA's Center for Politics and Politico have teamed up to offer interactive election ratings. #politics #elections #voting 1 hour ago

تقاضا برای تنوع در نتایج

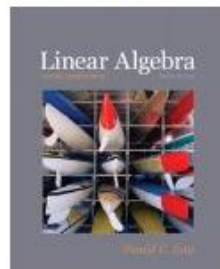
## بازیابی اطلاعات فقط جستجوی وب نیست

جستجوی وب فقط یک حوزه از بازیابی اطلاعات است

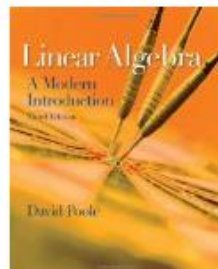
سایر حوزه ها

سیستم های توصیه گر Recommendation systems

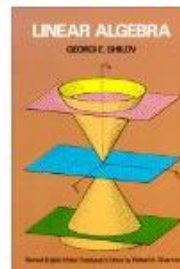
### Recommended Based on Your Browsing History



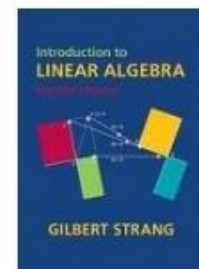
Linear Algebra and Its Applications...  
➤ David C. Lay  
Hardcover  
★★★★☆ (84)  
\$183.33 **\$141.16**



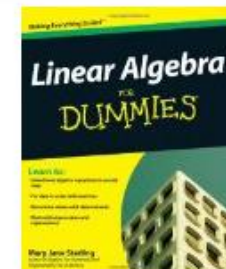
Linear Algebra: A Modern Introduction  
➤ David Poole  
Hardcover  
★★★★☆ (41)  
\$316.95 **\$289.88**



Linear Algebra  
➤ G. E. Shilov  
Paperback  
★★★★☆ (34)  
\$48.95 **\$12.65**



Introduction to Linear Algebra...  
➤ Gilbert Strang  
Hardcover  
★★★★☆ (57)  
\$87.50 **\$83.13**



Linear Algebra For Dummies  
➤ Mary Jane Sterling  
Paperback  
★★★★☆ (29)  
\$49.99 **\$16.23**

## بازیابی اطلاعات فقط جستجوی وب نیست

سیستم های پاسخگویی به سوال (Question answering)



how to calculate derivative of gamma function ☆ ≡

📄 📺 📋 🔗 ≡ Examples 🎲 Random

Derivative: 📄 Step-by-step solution

$$\frac{d}{dx}(\Gamma(x)) = \Gamma(x) \psi^{(0)}(x)$$

$\Gamma(x)$  is the gamma function  
 $\psi^{(n)}(x)$  is the  $n^{\text{th}}$  derivative of the digamma function

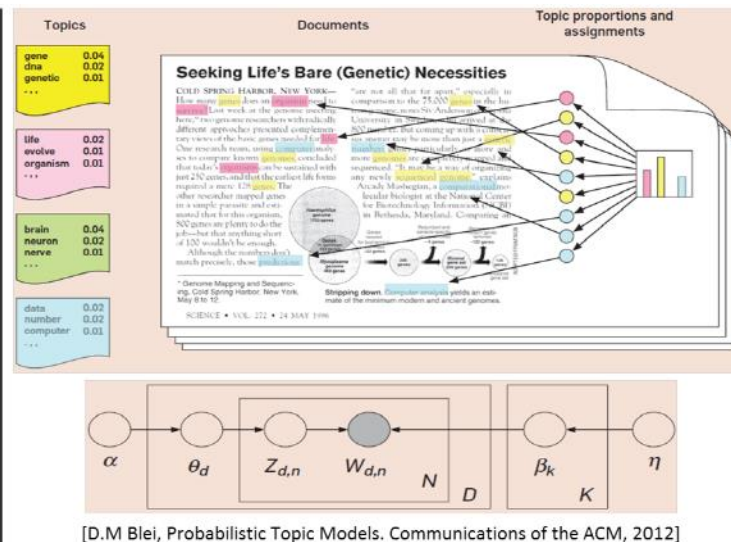
## بازیابی اطلاعات فقط جستجوی وب نیست

بازیابی اطلاعات همچنین شامل متن کاوی است

Text mining

مدلسازی موضوعی

تحلیل احساس



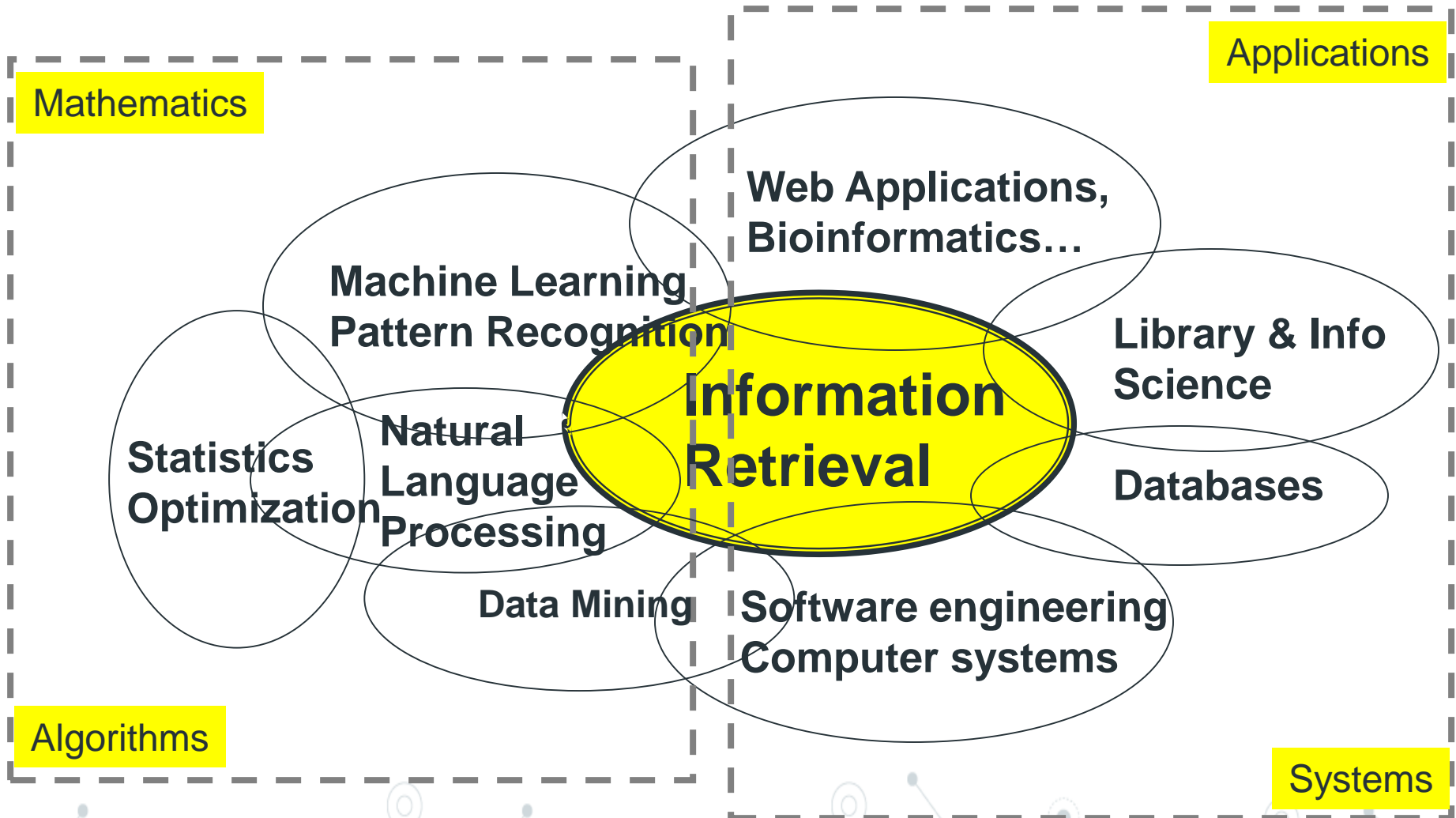
[D.M Blei, Probabilistic Topic Models. Communications of the ACM, 2012]

## بازیابی اطلاعات فقط جستجوی وب نیست

جستجوی اطلاعات سازمان + web search + desktop search:

The screenshot displays the 'PEOPLE/WEB SEARCH / UNIVERSITY OF VIRGINIA' interface. At the top, there are navigation links for 'Discover UVA', 'Schools', 'Athletics', 'Arts', 'Medical Center', 'Research', 'Libraries', 'Public Service', 'New Media', and 'Search'. Below this, a search bar contains the text 'hongning wang' and a 'Search' button. To the right of the search bar, there are tabs for 'People', 'U.Va. Web', and 'Library'. Below the search bar, there are links for 'Search the people directory' and 'Find people named "hongning%20wang"'. A list of search results is shown, including 'All results', 'Current News (2006-present)', 'News Archive (1998-2006)', 'Athletics/Intramurals', 'Health System', 'Human Resources', and 'Computing'. The first result is for 'University of Virginia' and 'Xi Wang - Thursday, May 1, 2014 Synthesizing REST Web Services to Advance a Useful Science of Non-Functional System Properties and Tradeoffs; Chi ...'. The second result is for 'Wang, Hongning' and 'Assistant Professor Name, Hongning Wang. Computing ID, hw5x. Office Phone, (434) 982-2225. Email, [edit] (registered). [add alias]. Title • Classification, Assistant Professor • ...'. The third result is for 'CS New Hires' and 'Jun 2, 2014 ... Hongning Wang, Yue Lu and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. The 16th ...'. The fourth result is for 'Computer Science Colloquia' and 'Hongning Wang, University of Illinois, Champaign-Urbana. Tuesday, March 18, 2014, 3:30 PM, Rice Hall, Rm. 130 (Light refreshments after the seminar Rice ...'.

## حوزه های مرتبط با بازیابی اطلاعات





## IR v.s. DBs

### سیستم های دیتابیس:

○ داده ساخت یافته

○ سمانتیک هر آبجکتی به خوبی تعریف شده است

○ Structured query languages (e.g., SQL)

○ بازیابی دقیق (exact retrieval)

○ تاکید بر کارایی

### بازیابی اطلاعات:

○ داده غیرساخت یافته

○ سمانتیک آبجکت ها قضاوتی است

○ کوئری های ساده متشکل از کلمات

○ بازیابی مبتنی بر مرتبط بودن

○ تاکید بر اثربخشی ولی کارایی هم مدنظر است



# تشکر

## سوال؟

[a.golzari@azaruniv.ac.ir](mailto:a.golzari@azaruniv.ac.ir)

[a.golzari@tabrizu.ac.ir](mailto:a.golzari@tabrizu.ac.ir)

<https://github.com/Amin-Golzari-Oskouei>

