

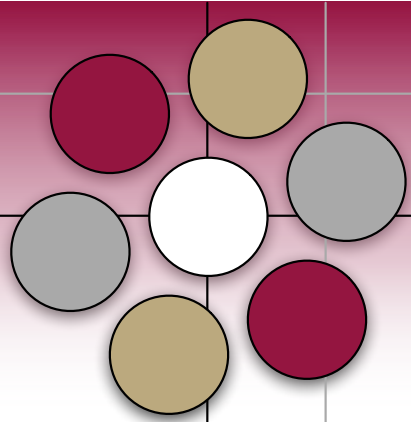
Web Search Basics

Introduction to Information Retrieval

Content adapted from Hinrich Schütze
<http://www.informationretrieval.org>

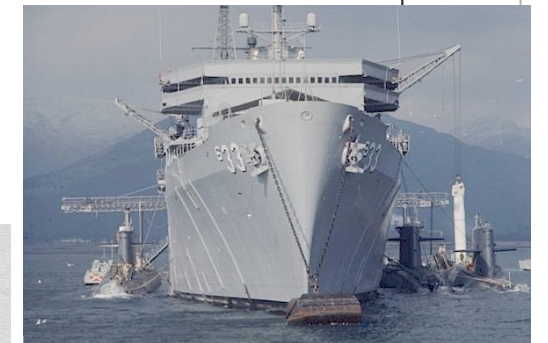
My Family

- I'm married and have four kids and live on the East Side
- I'm pretty open about my story



My Background

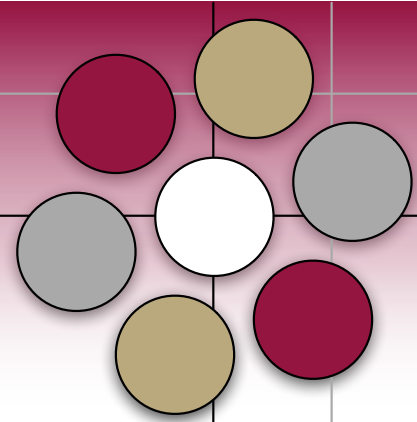
- B.S., Computer Science from Cornell (NY)
- M.Eng, Electrical Engineering
- STRIKE Officer
 - USS CURTIS WILBUR, DDG-54, Japan
- Operations Officer
 - USS SIMON LAKE, AS-33, Italy
- Ph.D in Computer Science and Engineering at UW
- Director of the Laboratory for Ubiquitous Computing and Interaction
- Co-founder of quub, whisper.fm, swayr, waitscout, audia



My Research

- Artificial Intelligence + Collapse Informatics + Ubiquitous Computing
- Example: I made a sensor system for babies in the NICU to detect cerebral palsy





ManTech

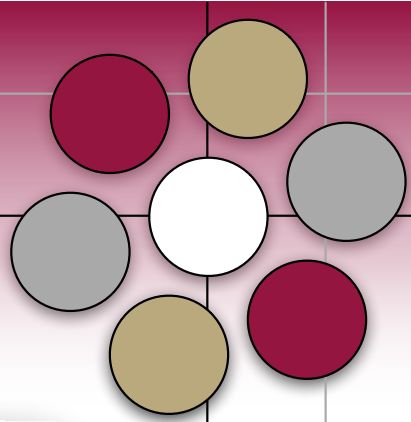
International Corporation®

•

VS



Introduction

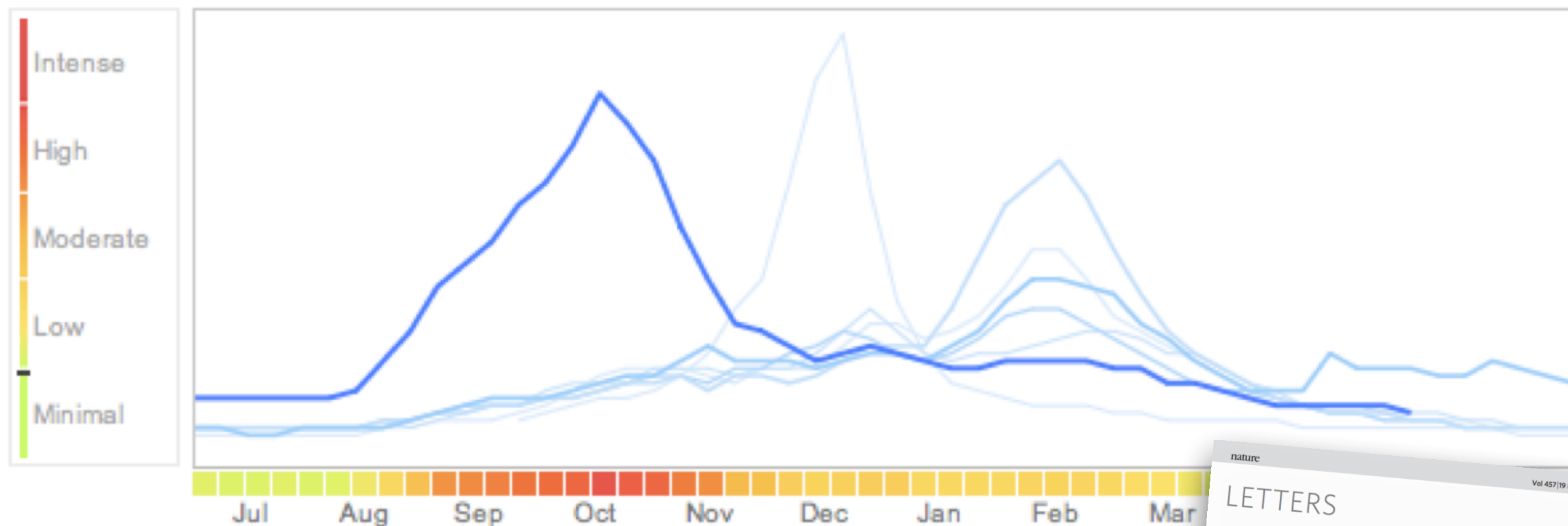


The Google logo is displayed in its multi-colored font (blue, red, yellow, blue, green, red). A small, stylized figure of a person is standing at the end of the word, near the bottom of the final 'e'.

- Google rebranded yesterday

Google Flu Trends

National



LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year¹. In addition to seasonal immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities². Early detection of disease activity, when followed by a rapid health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

Traditional surveillance systems, including those used by the US Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virological and clinical data, including influenza-like illness (ILI) physician visits. The CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1–2-week reporting lag. In an attempt to provide faster detection, innovative surveillance systems have been created to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines³ and over-the-counter drug sales⁴. About 90 million American adults are believed to search online for information about specific diseases or

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \epsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t , α is the multiplicative coefficient, and ϵ is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza Sentinel Provider Surveillance Network (<http://www.cdc.gov/flu/surveillance>) was used to help build our models. For each of the nine age percentage regions of the United States, the CDC reported the average ILI-related on a weekly basis. No data were provided for weeks outside of the annual influenza season, and we excluded such dates from model fitting, although our model was used to generate unvalidated ILI estimates for these weeks.

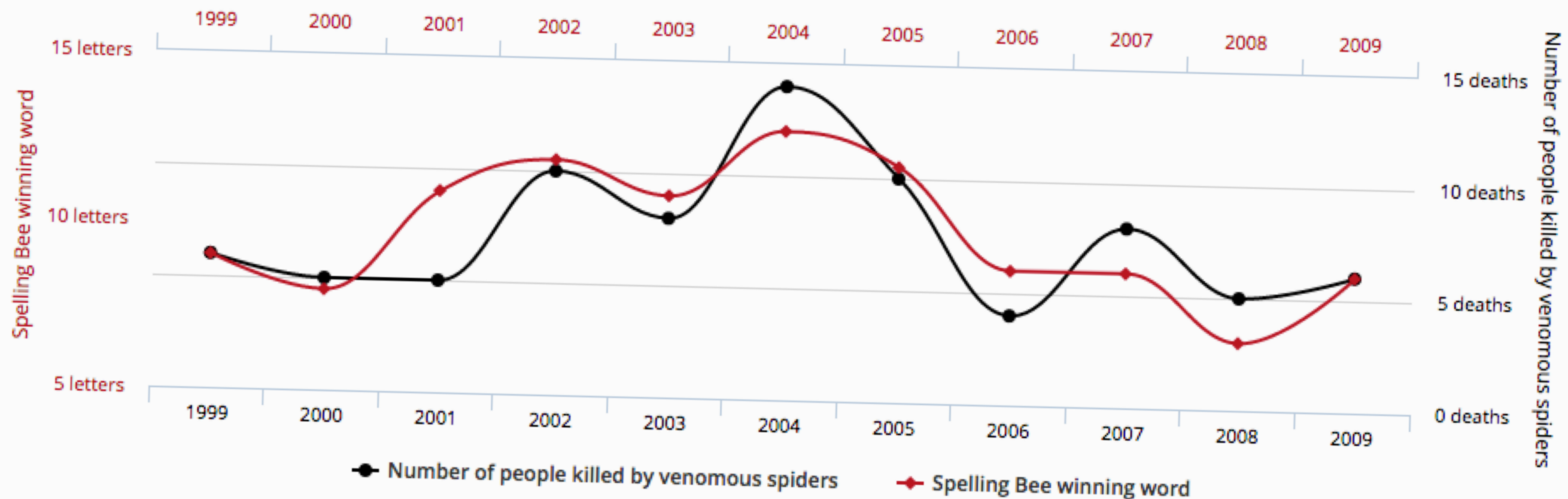
We designed an automated method of selecting ILI-related search queries, requiring no previous knowledge about influenza. We

- Epidemic monitoring from search queries

Google Trends

Letters in Winning Word of Scripps National Spelling Bee correlates with Number of people killed by venomous spiders

Correlation: 80.57% ($r=0.8057$)

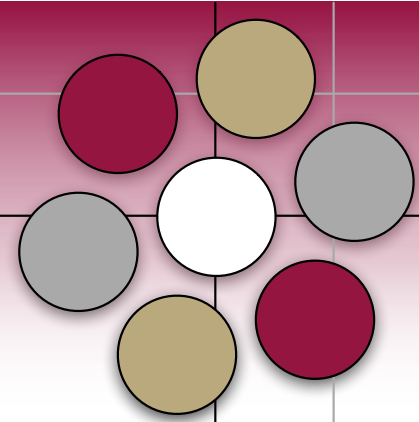


Data sources: National Spelling Bee and Centers for Disease Control & Prevention

tylervigen.com

- Correlation != Causality
- <http://www.tylervigen.com/spurious-correlations>

Google Trends



Compare Search terms ▼

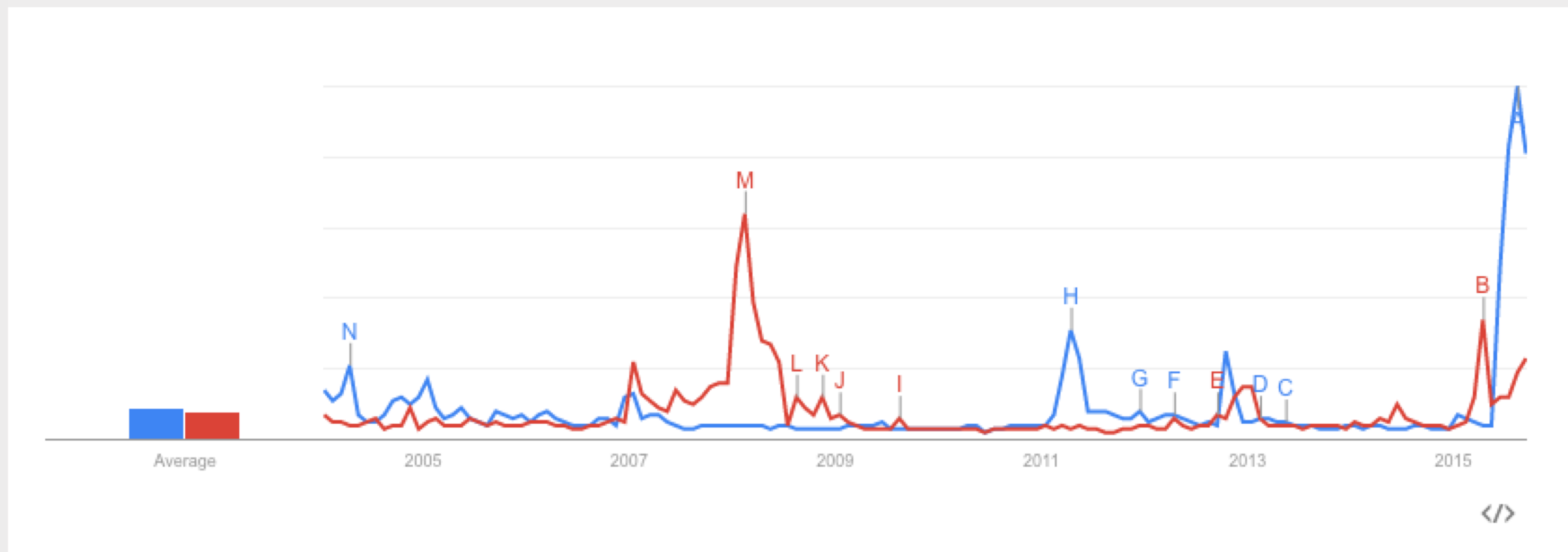
donald trump
Search term

hillary clinton
Search term

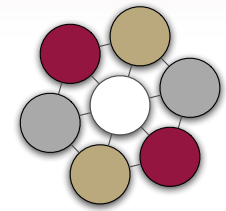
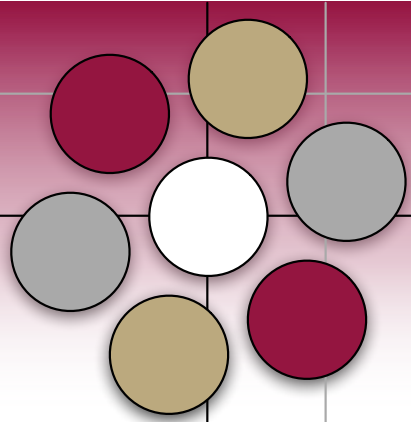
+ Add term

Interest over time ?

☒ News headlines ☐ Forecast ?



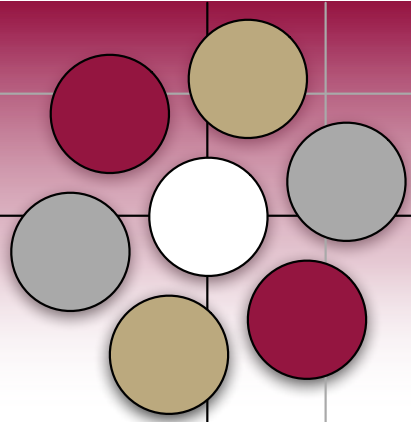
Overview



Introduction

- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
 - Size of the Web
- Web Users
- Spam

Introduction



Where did Big Data come from?

- Classic IR
- Web IR
- Generalized data processing frameworks
- Cheap cloud storage
- Lots of data
 - Apply the frameworks to them

Handwritten notes on a piece of paper, likely a whiteboard or a large sheet of paper, showing data generation rates for Google's self-driving car and the NYSE.

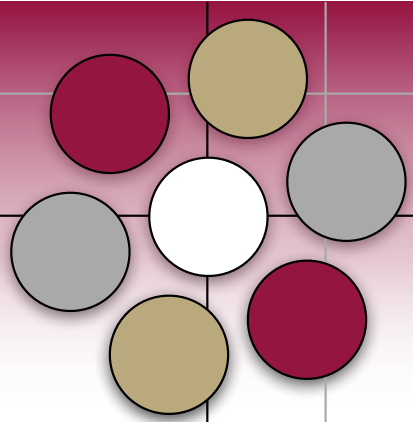
1 G.S.D.C. = 7.03 NYSE
DATA GENERATED

GOOGLE SELF-DRIVING CAR
1GB/sec

NYSE generates
1TB/day

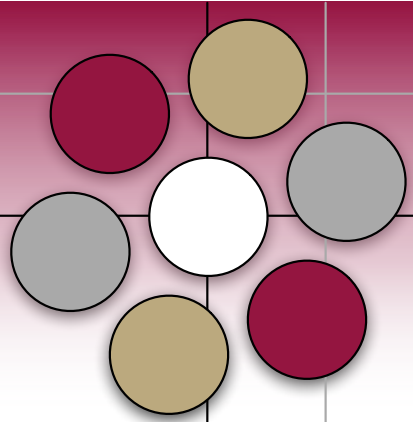
7200 GB
7.03 NYSE

Introduction



- Search behavior has changed a lot over time
- How often do you use search?

Introduction



- Search behavior has changed a lot over time
- How often do you use search?
- How often do you use search engines?

Introduction

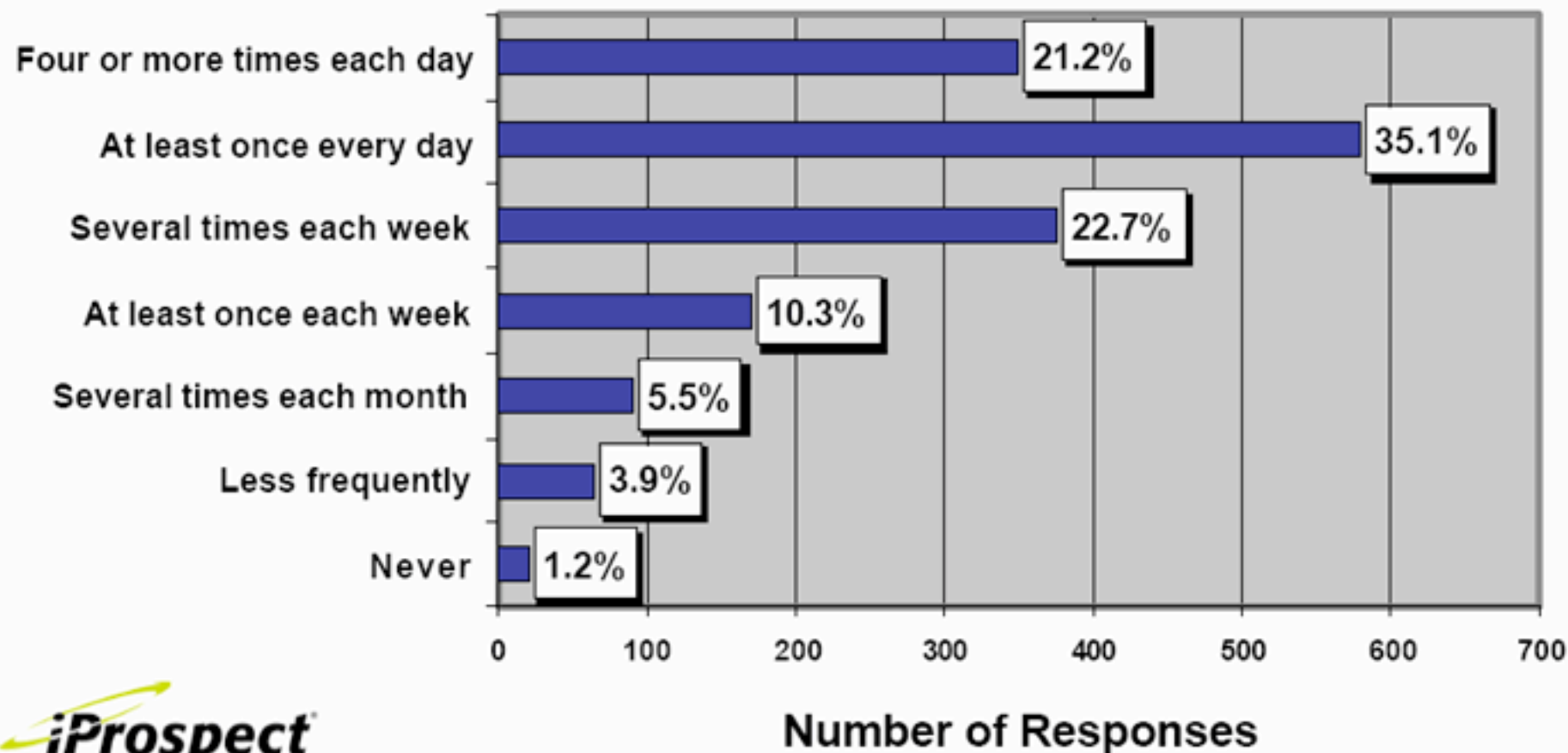
- Search behavior has changed a lot over time

Search use ...

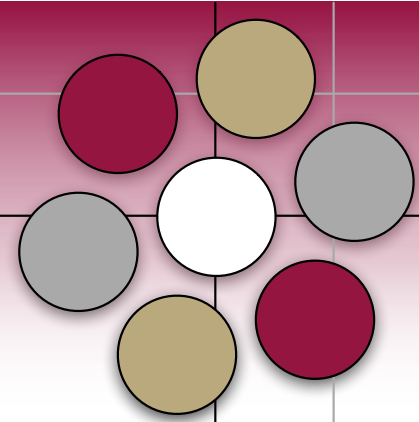
(iProspect Survey, 4/04,

<http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf>)

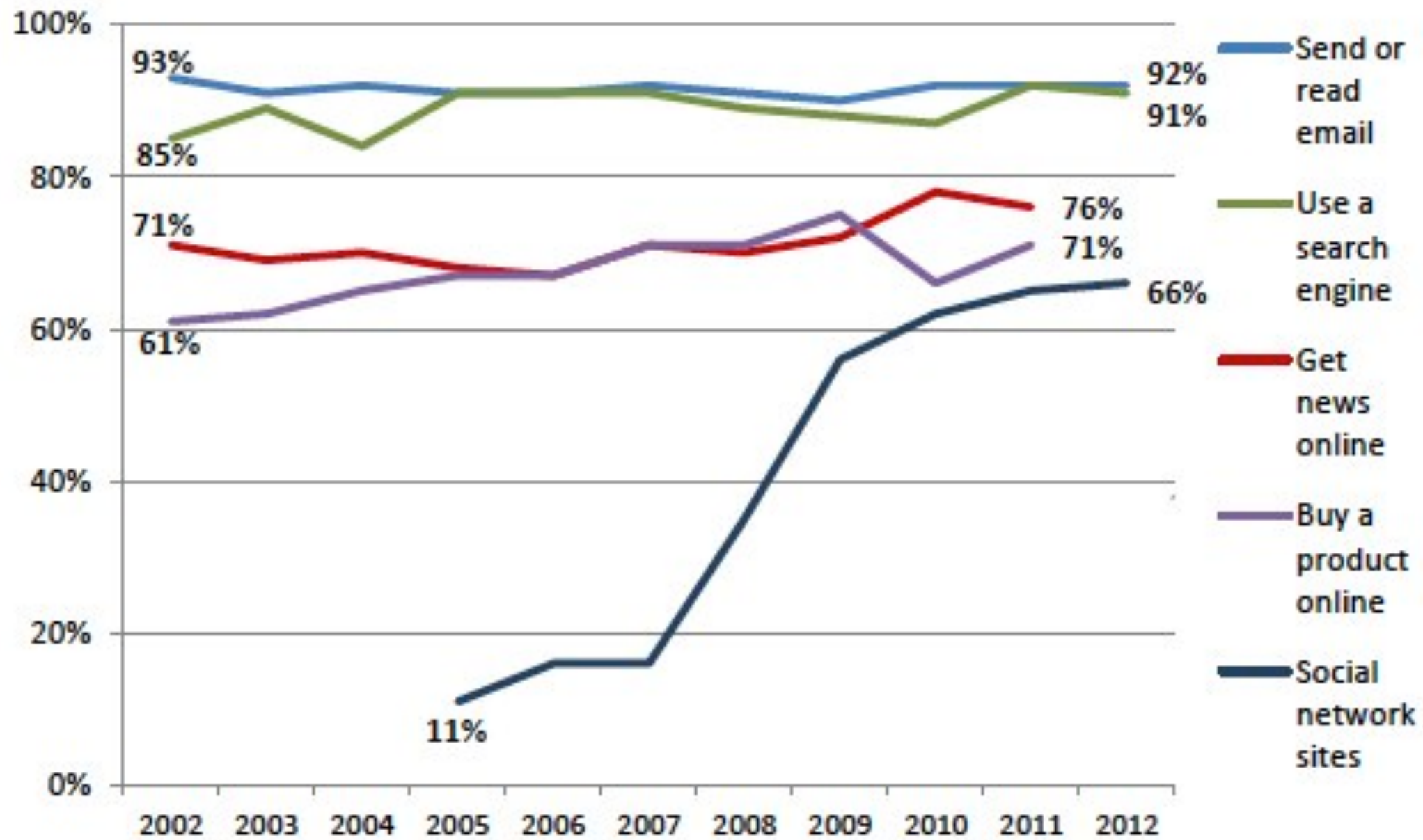
How often do you use search engines on the Internet?



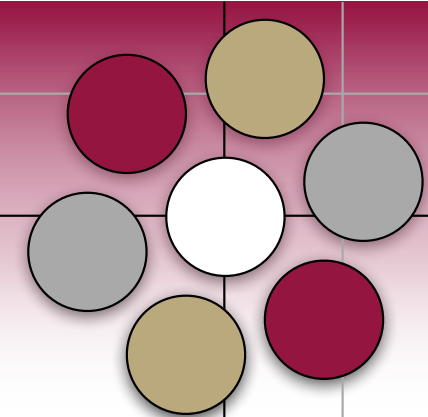
Introduction



% of internet users who do each activity



Introduction

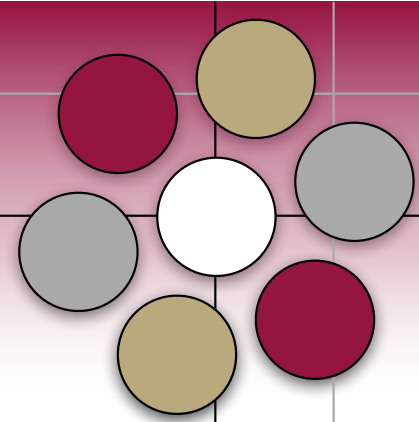


Who uses search?

% of online adults in each group who use search engines

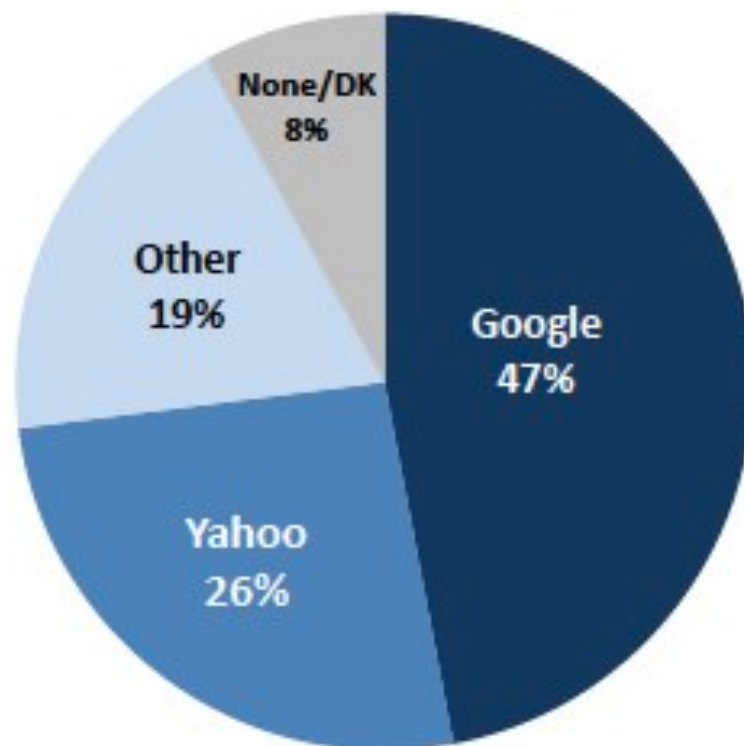
	% of each group who ever use search engines	% of each group who used a search engine yesterday
All online adults	91%	59%
Gender		
Male	90	59
Female	92	60
Race/Ethnicity		
White	93*	63*
African American	89*	44
Hispanic	79	44
Age		
18-29	96	66*
30-49	91	65*
50-64	92	52*
65+	80	38
Education		
Some high school	78	34
High school	88*	45*
Some college	94*	65*
College graduate	95*	74*
Household income		
< \$30,000	84	45
\$30,000 - \$49,999	93*	54*
\$50,000 - \$74,999	97*	66*
\$75,000+	95*	76*

Introduction

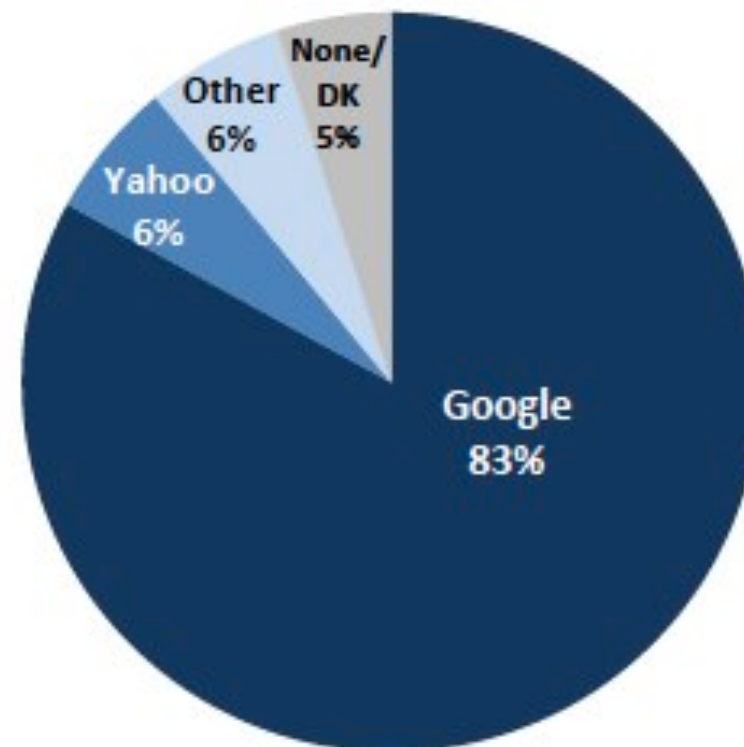


Google is far and away the search engine of choice, preferred by 83% of search users

% of search users who answered the question: Which search engine do you use MOST OFTEN?



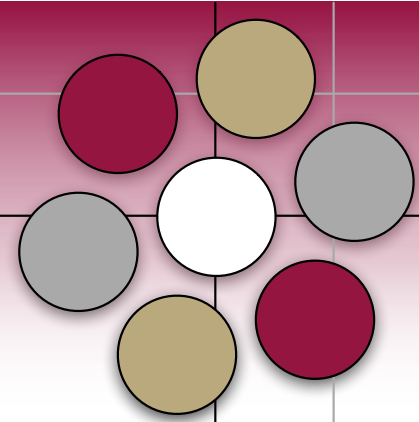
2004



2012

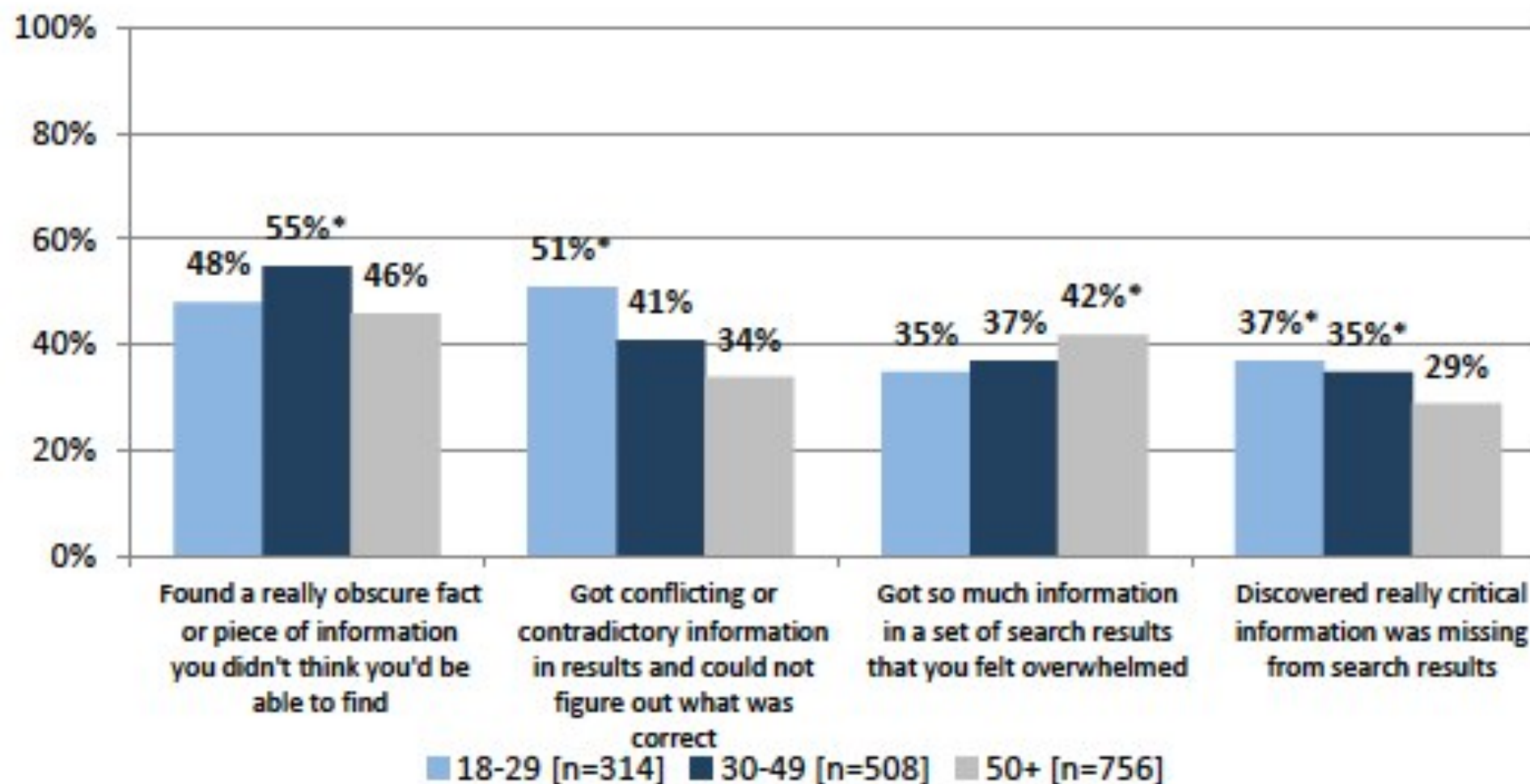
Source: The Pew Research Center's Internet & American Life Project Winter 2012 Tracking Survey, January 20-February 19, 2012. N=2,253 adults, age 18 and older, including 901 cell phone interviews. Interviews conducted in English and Spanish.

Introduction



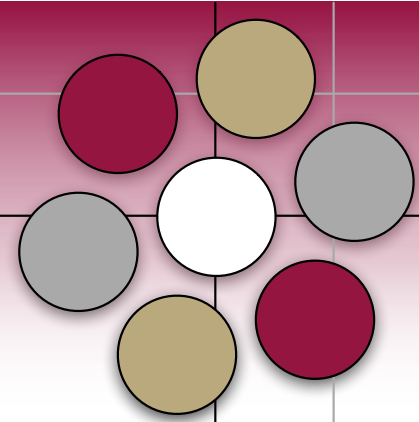
Some search users' experiences vary by age

% of each group who have experienced each of the following...



Source: The Pew Research Center's Internet & American Life Project Winter 2012 Tracking Survey, January 20-February 19, 2012. N=2,253 adults, age 18 and older, including 901 cell phone interviews. Interviews conducted in English and Spanish. An asterisk (*) indicates a significant difference at the 95% confidence level.

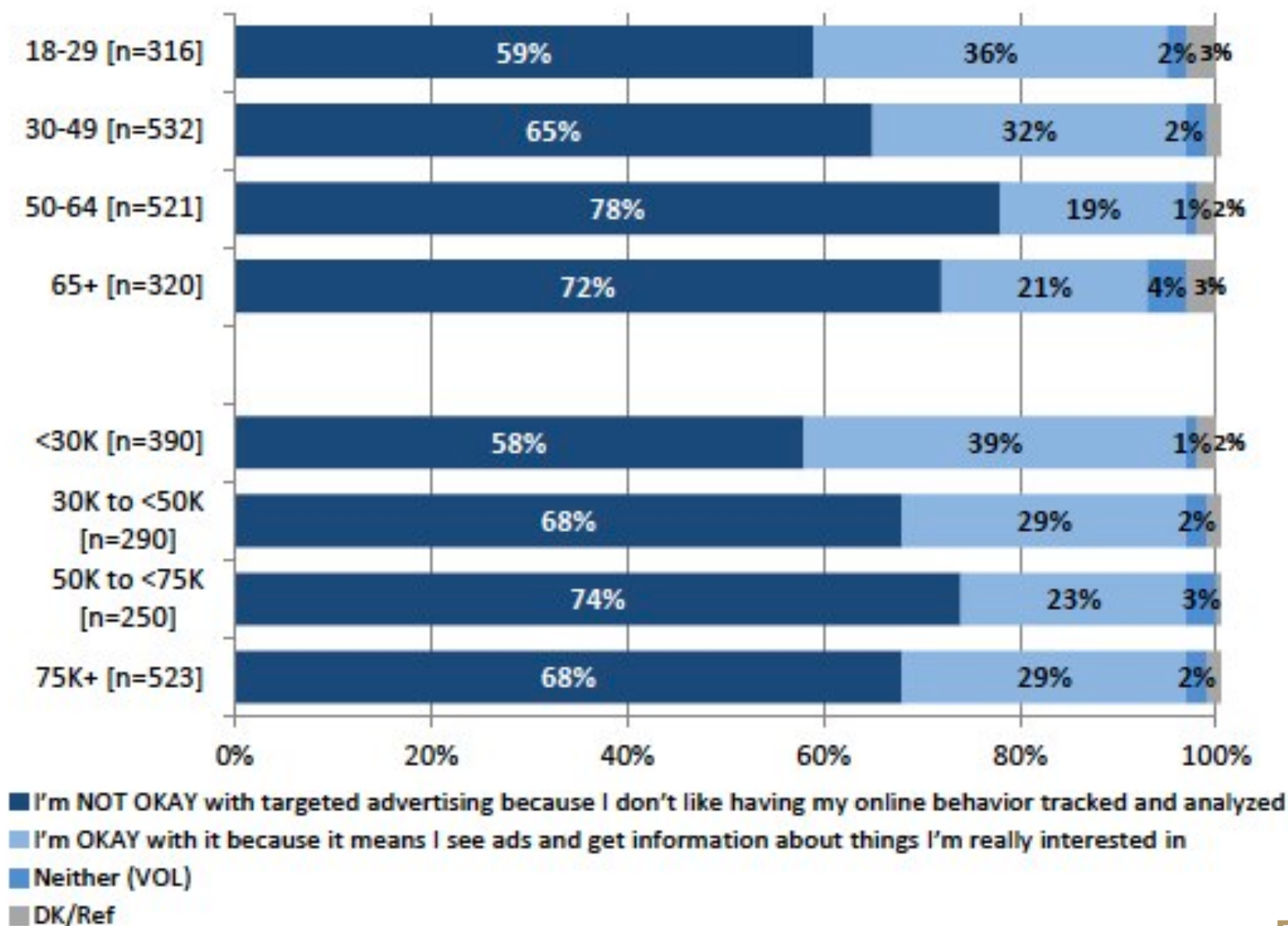
Introduction



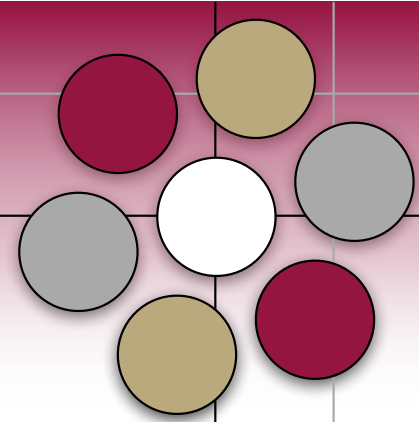
Views of targeted advertising vary by age and income

Which of the following statements comes closest to how you, personally, feel about TARGETED ADVERTISING being used online – even if neither is exactly right?

Asked of adult internet users [n=1,729]



Introduction



Search Education - Google

www.google.com/insidesearch/searcheducation/index.html

Google

Search Education

Home Lesson Plans Live Trainings

Help your students become better searchers

Web search can be a remarkable tool for students, and a bit of instruction in how to search for academic sources will help your students become critical thinkers and independent learners.

With the materials on this site, you can help your students become skilled searchers- whether they're just starting out with search, or ready for more advanced training.

A photograph of students sitting at tables in a library, working on laptops. A large white play button is overlaid on the center of the image.

A green square icon featuring a white outline of an open book with the letters 'Ab' on the right page.

Lesson Plans & Activities

Download lesson plans to develop your students' search literacy skills.

[Browse lesson plans](#)

An orange square icon featuring a white outline of a star with a magnifying glass over it.

Power Searching

Improve your search skills and learn advanced tips with online lessons and activities.

[Start now](#)

An orange square icon featuring a white outline of a magnifying glass with a question mark inside.

A Google a Day Challenges

Put your students' search skills to the test with these trivia challenges.

[Browse challenges](#)

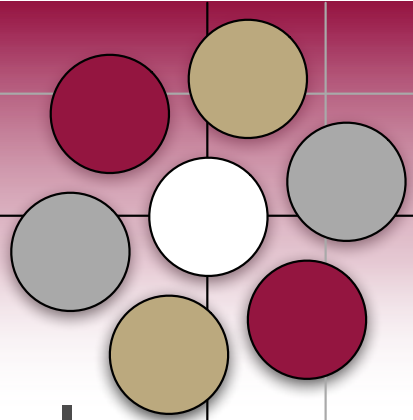
A blue square icon featuring a white outline of a television set with a play button and signal waves.

Live Trainings

Join us for live search trainings or watch past trainings from search experts here at Google.

[Start training](#)

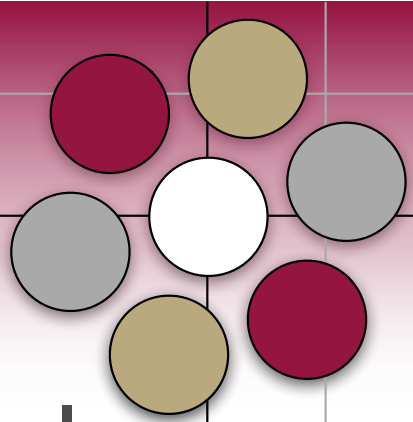
Introduction



Without search engines the web wouldn't scale

- No incentive in creating content unless it can be found.
- Taxonomies, bookmarks can't keep up
- Or can they? (del.icio.us)
- The web is both a technology artifact and a social environment
- “The Web has become the ‘new normal’ in the America way of life; those who don't go online constitute an every-shrinking minority” [Pew Foundation report, January 2005]

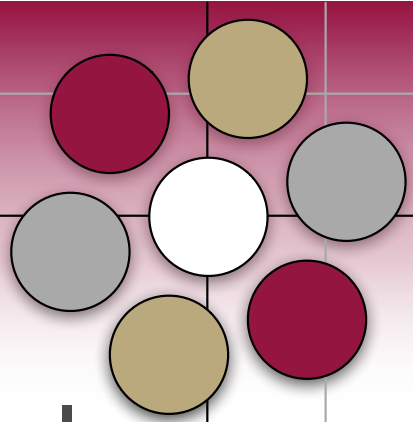
Introduction



Without search engines the web wouldn't scale

- Search engines make aggregation of interest possible:
 - Create incentives for very niche players
 - Economical - specialized stores, providers, etc.
 - Social - narrow interests, specialized communities
- The acceptance of search interaction makes “unlimited selection” stores possible
 - Amazon, Netflix, etc.

Introduction

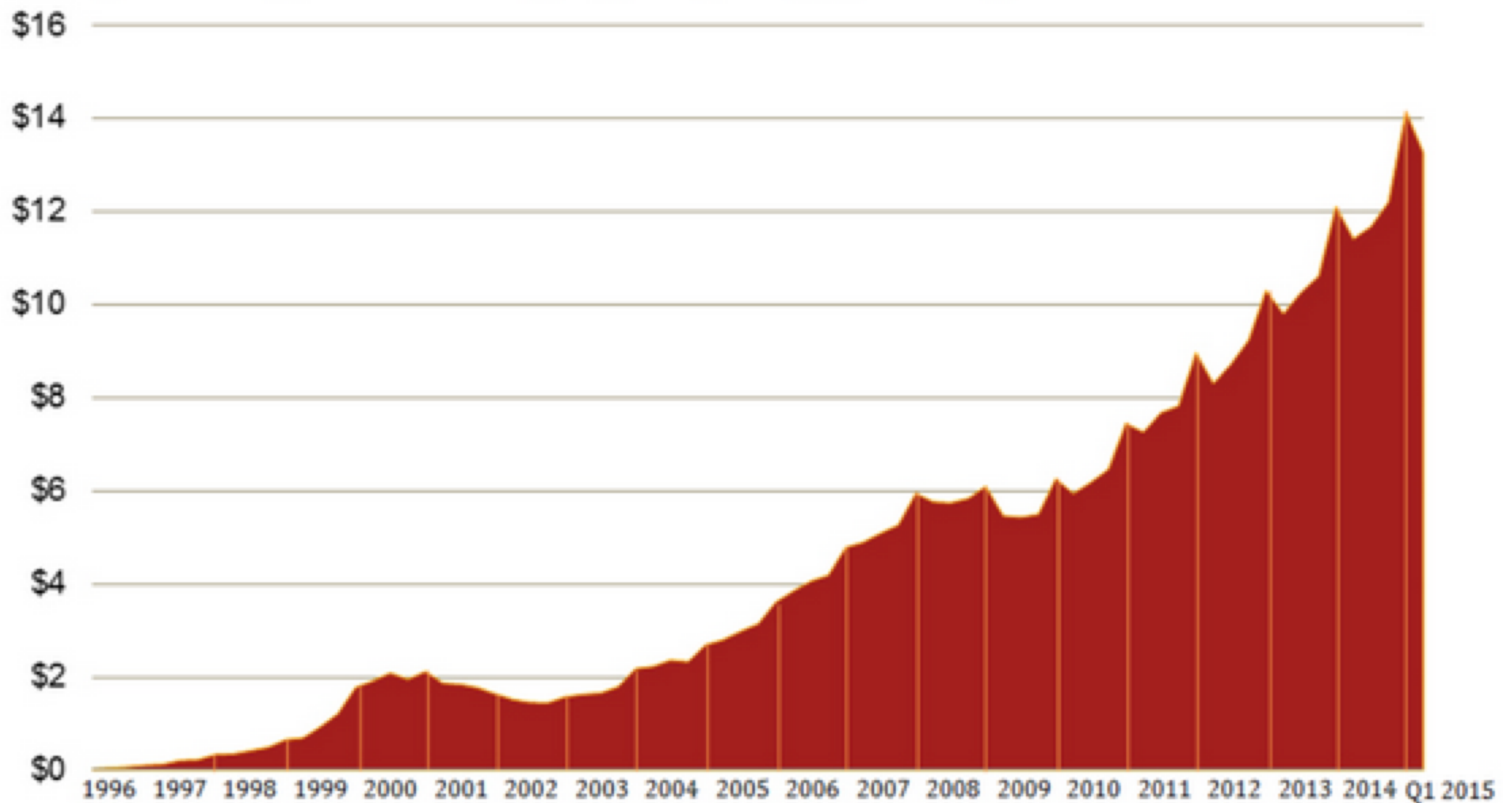


Without search engines the web wouldn't scale

- Search turned out to be the best mechanism for advertising on the web,.
- Growing very fast (entire US advertising industry is \$250 billion though)
- \$15 billion plus industry in 2009
- \$36 billion in 2012

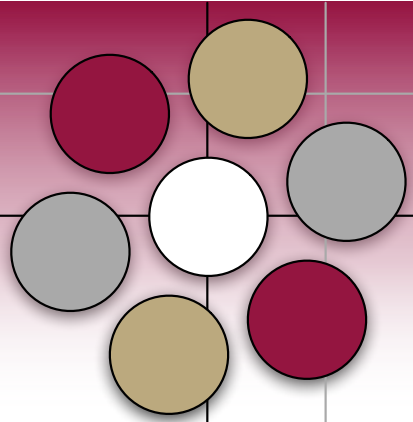
Introduction

Quarterly revenue growth trends Q1 1996 – Q1 2015 (\$ billions)

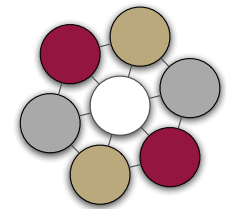


Source: IAB/PwC 2015

Overview

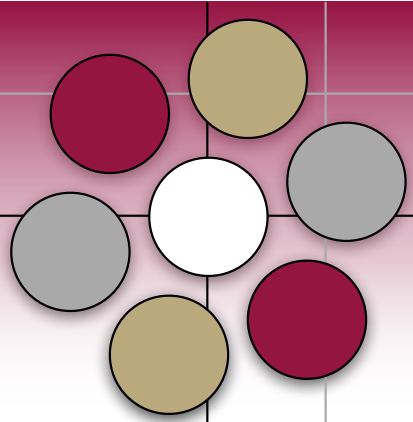


- Introduction



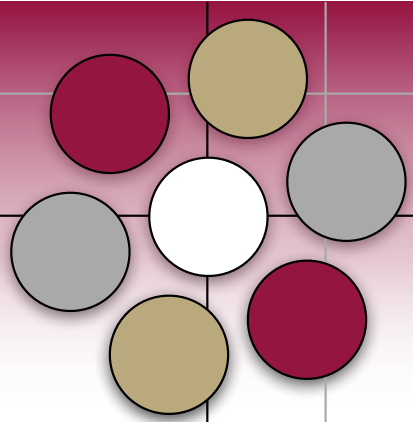
Classic Information Retrieval

- Web IR
- Sponsored Search
- Web Search Basics
 - Size of the Web
- Web Users
- Spam



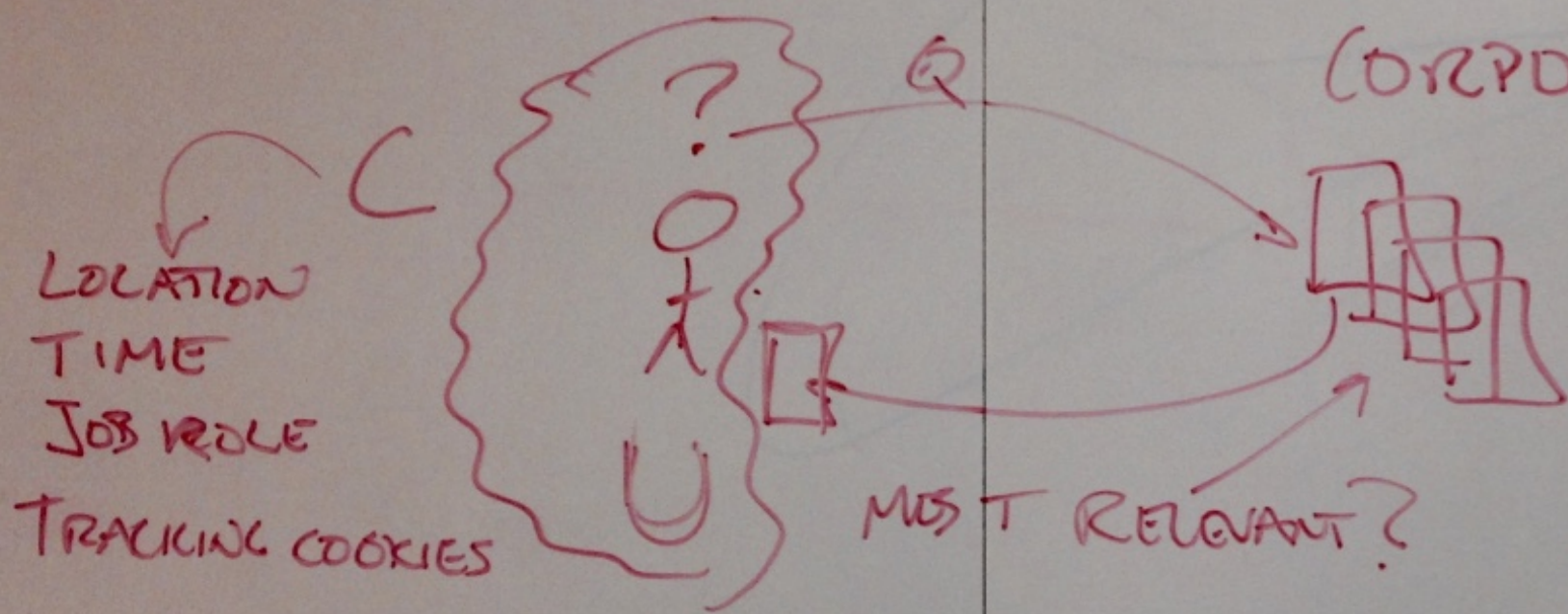
Classic IR assumptions

- Corpus: Fixed document collection
- Goal: Retrieve information content relevant to information need



Classic IR Goal

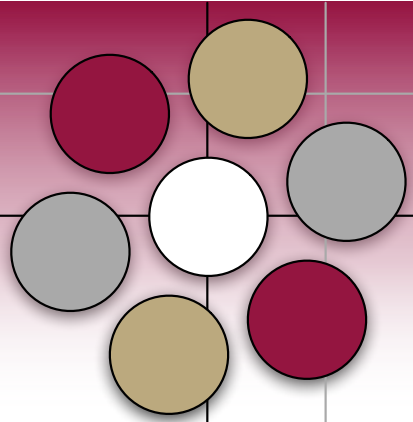
- Classic “Relevance”
 - For each query, Q , and stored document, D , in a corpus there exists a relevance score: $R(Q,D)$
 - $R(Q,D)$ is averaged over users, U , and contexts, C
 - Maximize $R(Q,D)$ instead of $R(Q,D,U,C)$
 - Context is ignored
 - Individuals are ignored
 - Corpus is static



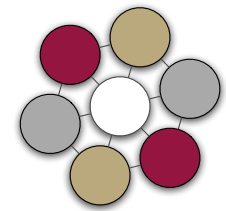
$$R(Q, D) = R$$

$$R(Q, D, U, C) = ?$$

Overview

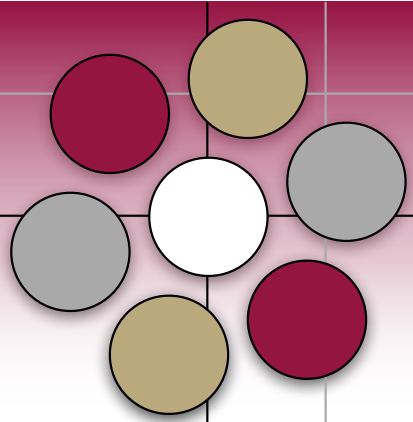


- Introduction
- Classic Information Retrieval



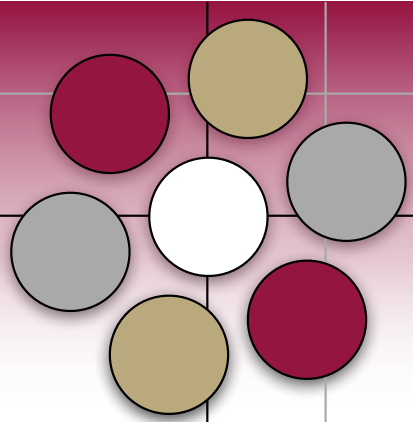
Web IR

- Sponsored Search
- Web Search Basics
 - Size of the Web
- Web Users
- Spam



Web IR: Differences from traditional IR

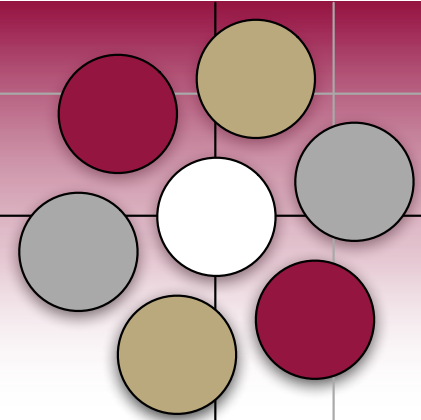
- On the web, search and ads are intricately connected
- The web is huge
- The web is a rapidly changing collection.
- There is spam on the web
 - Adversarial IR
 - Huge difference from traditional IR
- One interface for hugely divergent needs
 - Queries, Maps, Stocks, Weather, Calculations



History

- Early keyword-based engines
 - (1995-1997) Altavista, Excite, Infoseek, Inktomi
- Paid placement ranking
 - Goto.com -> Overture.com -> Yahoo!
 - Results based on auction for keyword placement

Web Information Retrieval

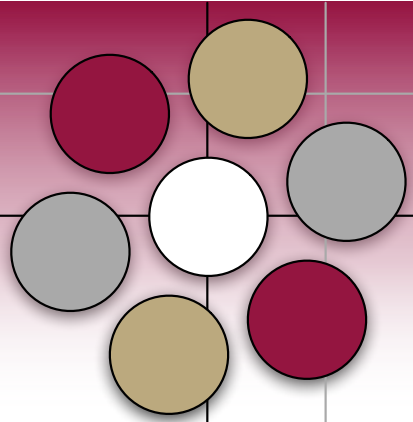


www.goto.com/d/search/?sessionid=1AQ4214AAA095QFIEF3QPUQ?type=home&tm=1&Keywords=Wilmington+

Wilmington real estate.

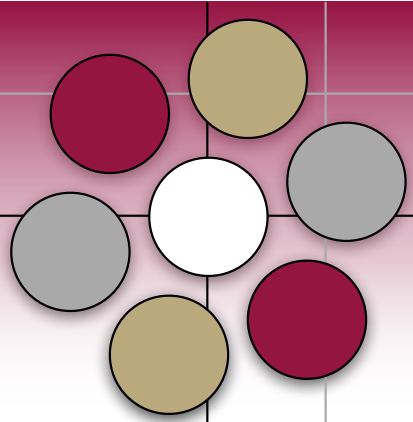
Access 75% of all users now!
Premium Listings reach 75% of all
Internet users. [Sign up](#) for Premium
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)
Wilmington's information and real estate guide. This is your one
anything to do with Wilmington.
[www.buddyblake.com](#) (Cost to advertiser: [10.38](#))
2. [Coldwell Banker Sea Coast Realty](#)
Wilmington's number one real estate company.
[www.cbseacoast.com](#) (Cost to advertiser: [10.37](#))
3. [Wilmington, NC Real Estate Becky Bullard](#)
Everything you need to know about buying or selling a home c
on my Web site!
[www.iwwc.net](#) (Cost to advertiser: [10.35](#))



History

- (1998+) Link-based ranking pioneered by Google
 - Links added the idea of “authoritativeness” to “relevance”
 - Blew away all early engines save Inktomi
 - Great user experience looking for a business model
 - Meanwhile Goto/Overture’s annual revenues were nearing \$1 billion



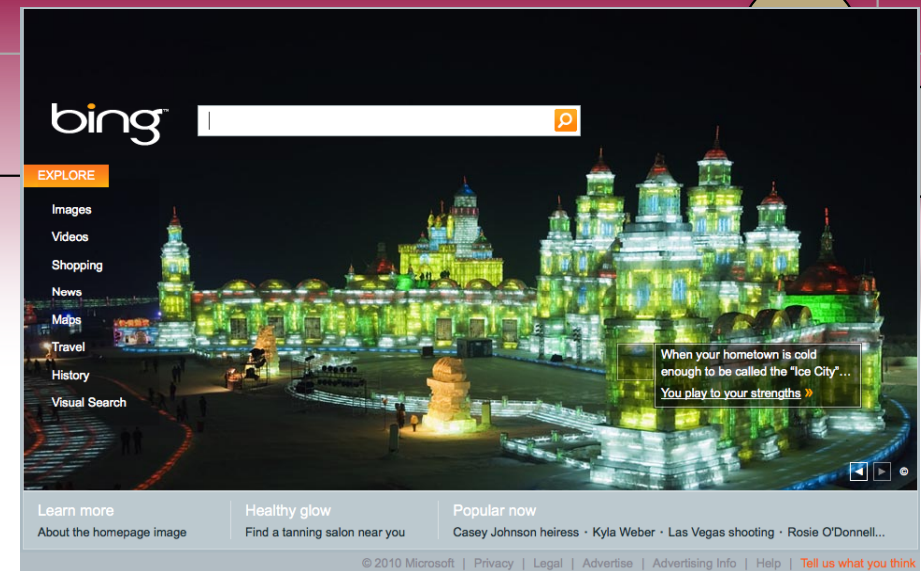
History

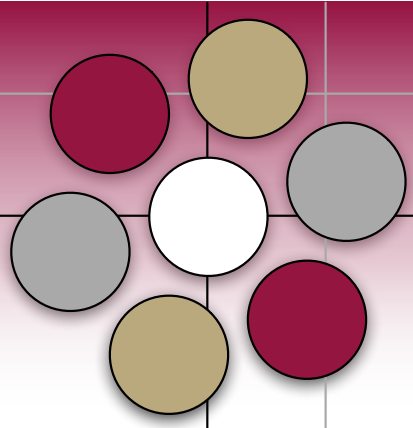
- Result
 - Google:
 - Added paid placement ads on the side
 - Differentiated from search results
 - Yahoo! built a similar architecture
 - Buys Overture for paid placement
 - Buys Inktomi for search

Web Information Retrieval

History

- 2004
 - Microsoft begins in house development of a search engine called Live
- May 28, 2009
 - Microsoft rebrands Live Search to Bing!
 - Search Engine wars intensify
 - New innovations appears at every turn
 - Technology becomes much more closely guarded



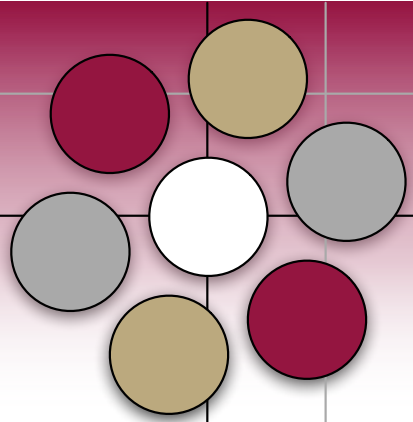


History

- Internationally
 - Chinese search engine Baidu “owns” Chinese search
 - Launched around 2000, specializes in Chinese content



Web Information Retrieval

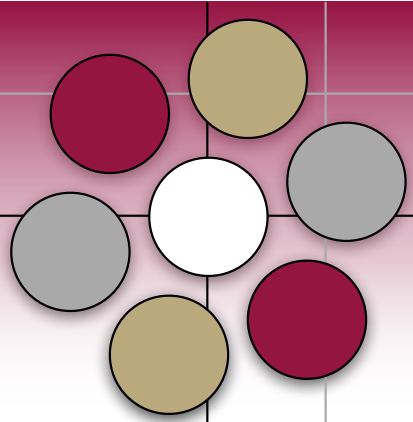


Today (8/29/2015)

- | | | |
|---------------------|-------------------------|-------------------------|
| 1. Google.com | 11. Live.com | 21. Vk.com |
| 2. Facebook.com | 12. Taobao.com | 22. Hao123.com |
| 3. Youtube.com | 13. Sina.com.cn | 23. T.co |
| 4. <u>baidu.com</u> | 14. Linkedin.com | 24. Bing.com |
| 5. Yahoo.com | 15. Yahoo.co.jp | 25. Google.de |
| 6. Amazon.com | 16. Weibo.com | 26. Instagram.com |
| 7. Wikipedia.org | 17. Ebay.com | 27. Msn.com |
| 8. Qq.com | 18. Google.co.jp | 28. Aliexpress.com |
| 9. Twitter.com | 19. Yandex.ru | 29. Microsoft.com |
| 10. Google.co.in | 20. <u>blogspot.com</u> | 30. <u>google.co.uk</u> |

- Top sites globally according to alexa.com

Web Information Retrieval

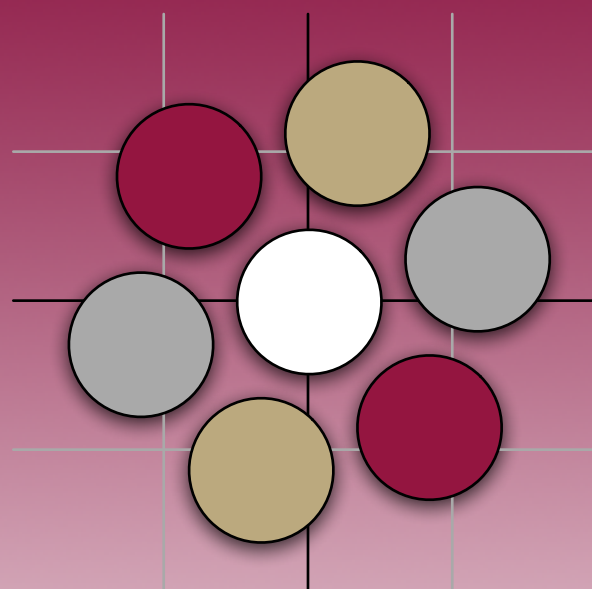


Before (1/7/2010)

1. google.com
2. facebook.com
3. youtube.com
4. yahoo.com
5. live.com
6. wikipedia.org
7. blogger.com
8. baidu.com
9. msn.com
10. yahoo.co.jp

13. twitter.com
15. google.cn
22. bing.com
29. google.co.jp
56. ask.com
64. cnn.com

- Top sites globally according to alexa.com



WESTMONT COMPUTER SCIENCE