Web Search Basics

Introduction to Information Retrieval

Informatics 141 / CS 121
Donald J. Patterson

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

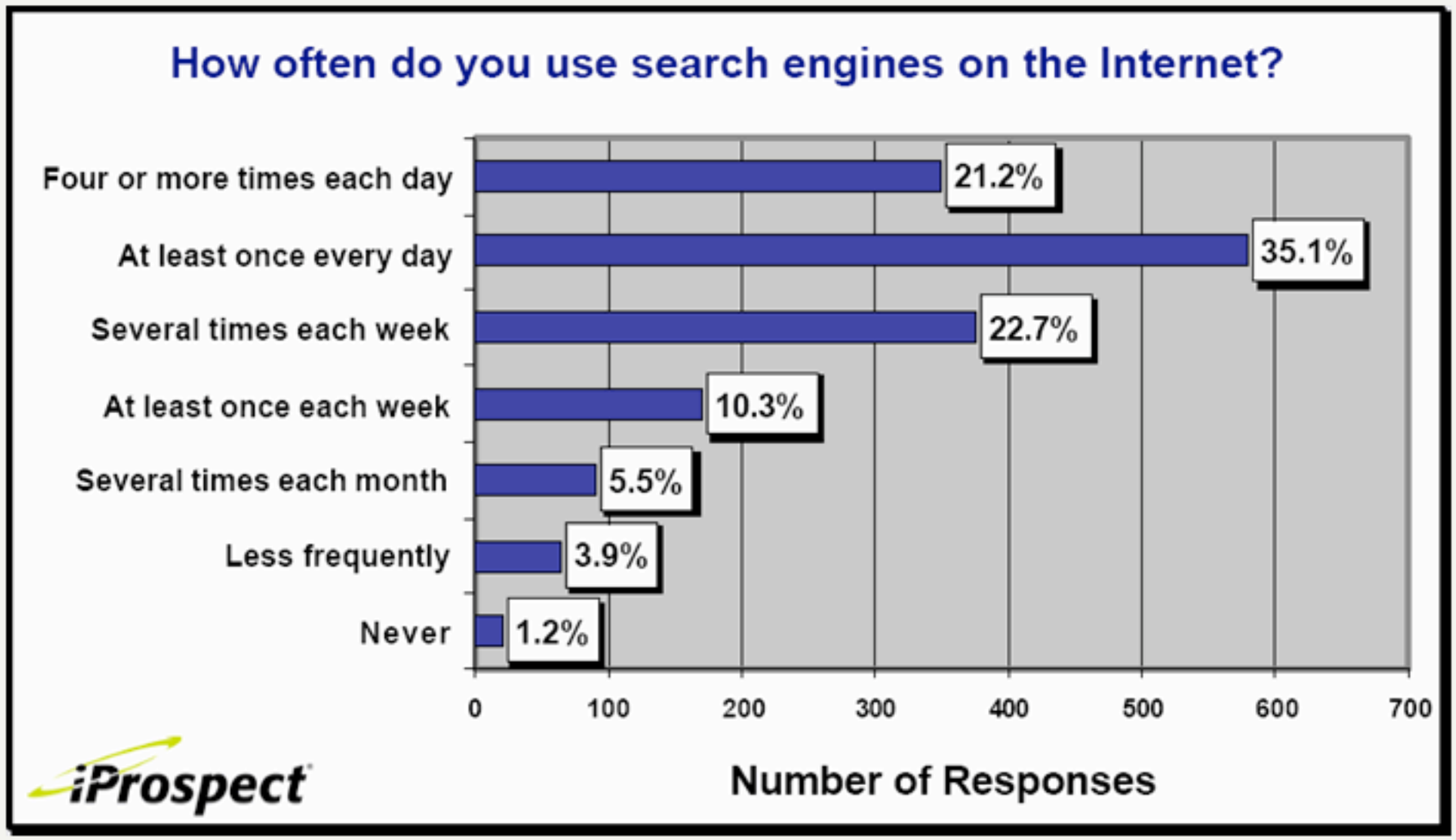
Overview

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

Introduction

Search use …

(iProspect Survey, 4/04,
http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf)



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? (de.licio.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 constitue an every-shrinking minority" [Pew Foundation report, January 2005]

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.

Without search engines the web wouldn't scale

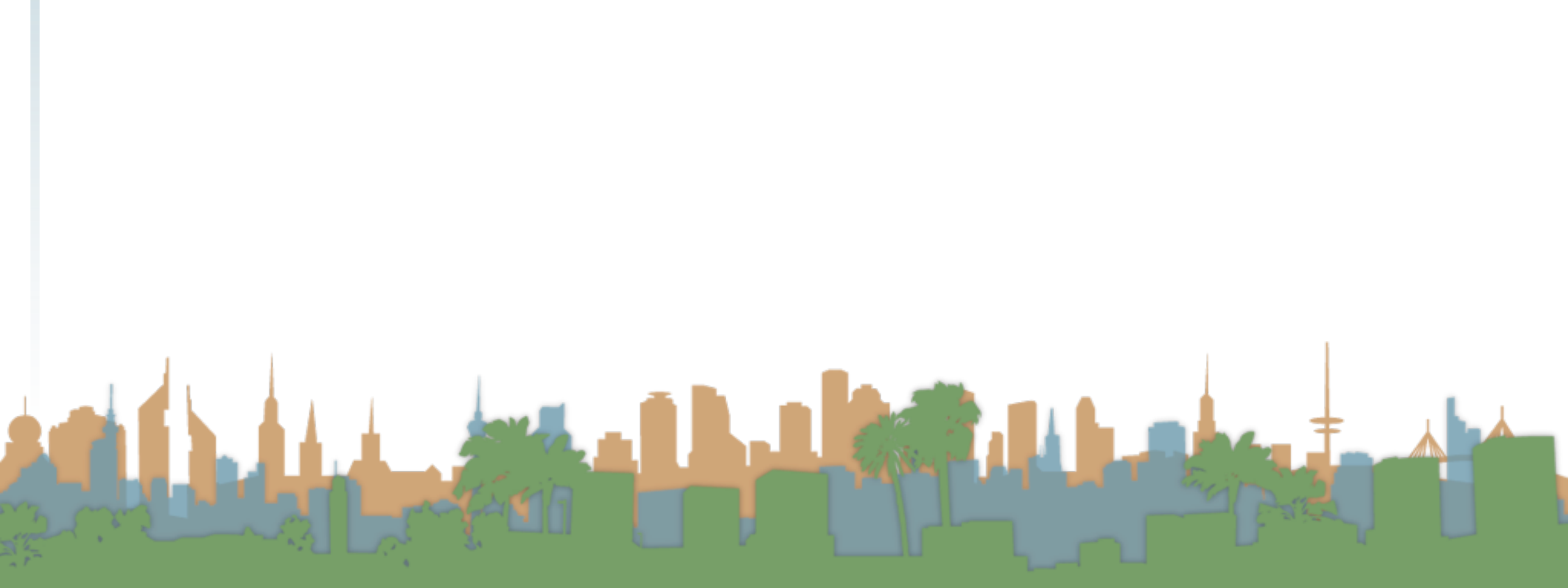
- Search turned out to be the best mechanism for advertising on the web, a $15 billion plus industry.
  - Growing very fast (entire US advertising industry is $250 billion though)
  - Sponsored search marketing is about $10 billion

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

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

www.goto.com/d/search/;$sessionid$AQ4Z14AAAH0R5QFIEF3QRJQ?type=home&tm=18&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 on
   anything to do with Wilmington.
   www.buddyblake.com (Cost to advertiser: $0.38)

2. **Coldwell Banker Sea Coast Realty**
   Wilmington's number one real estate company.
   www.cbseacoast.com (Cost to advertiser: $0.37)

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: $0.35)

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

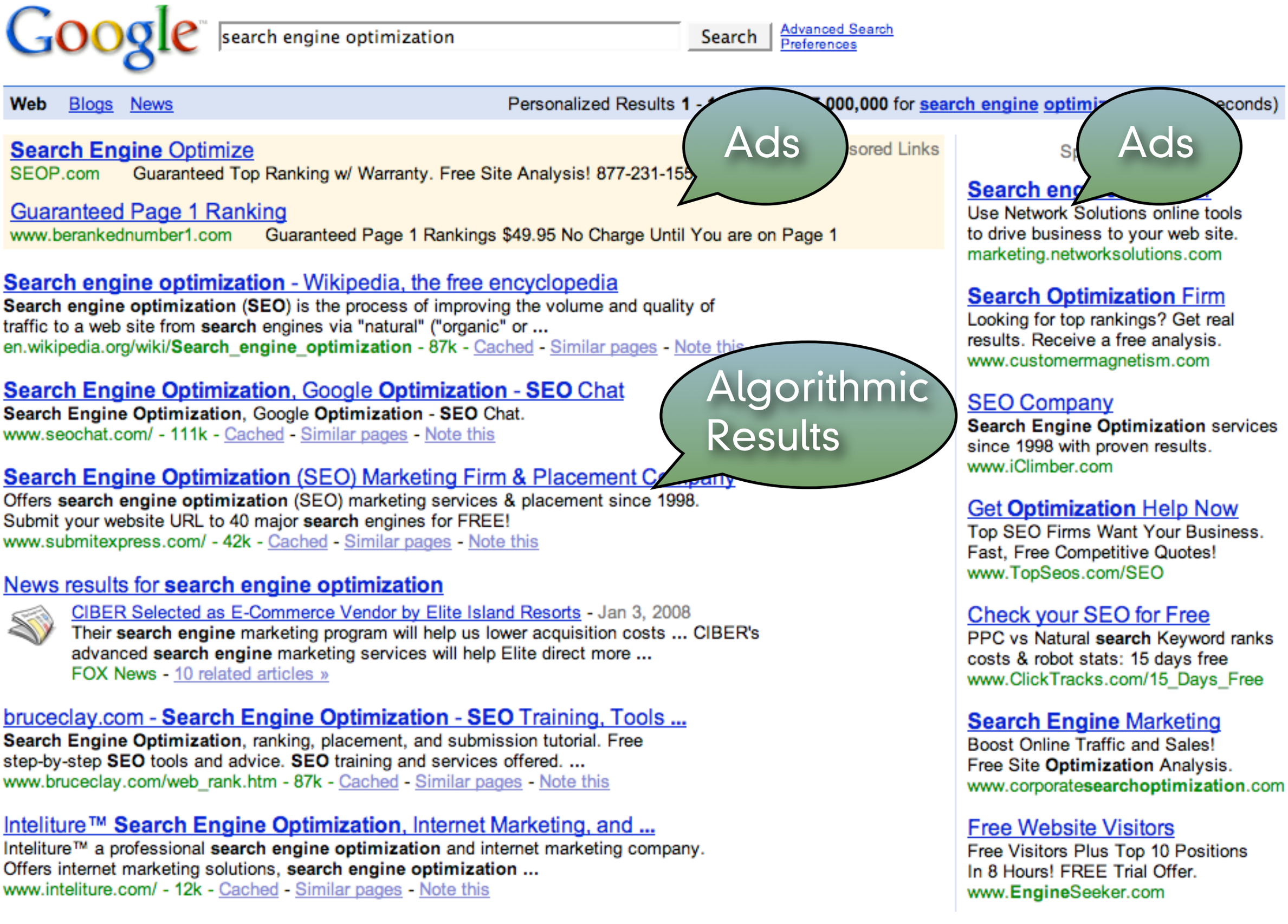
History

- 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

Overview

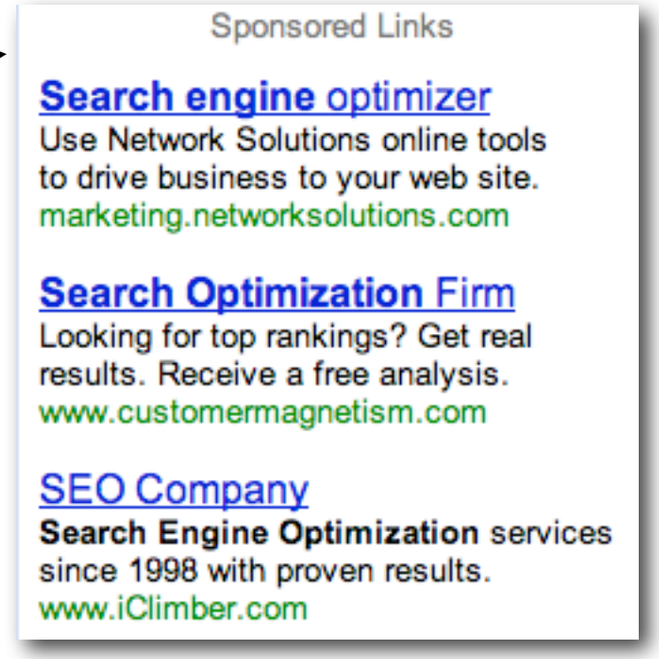
- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
  - Size of the Web
- Web Users
- Spam

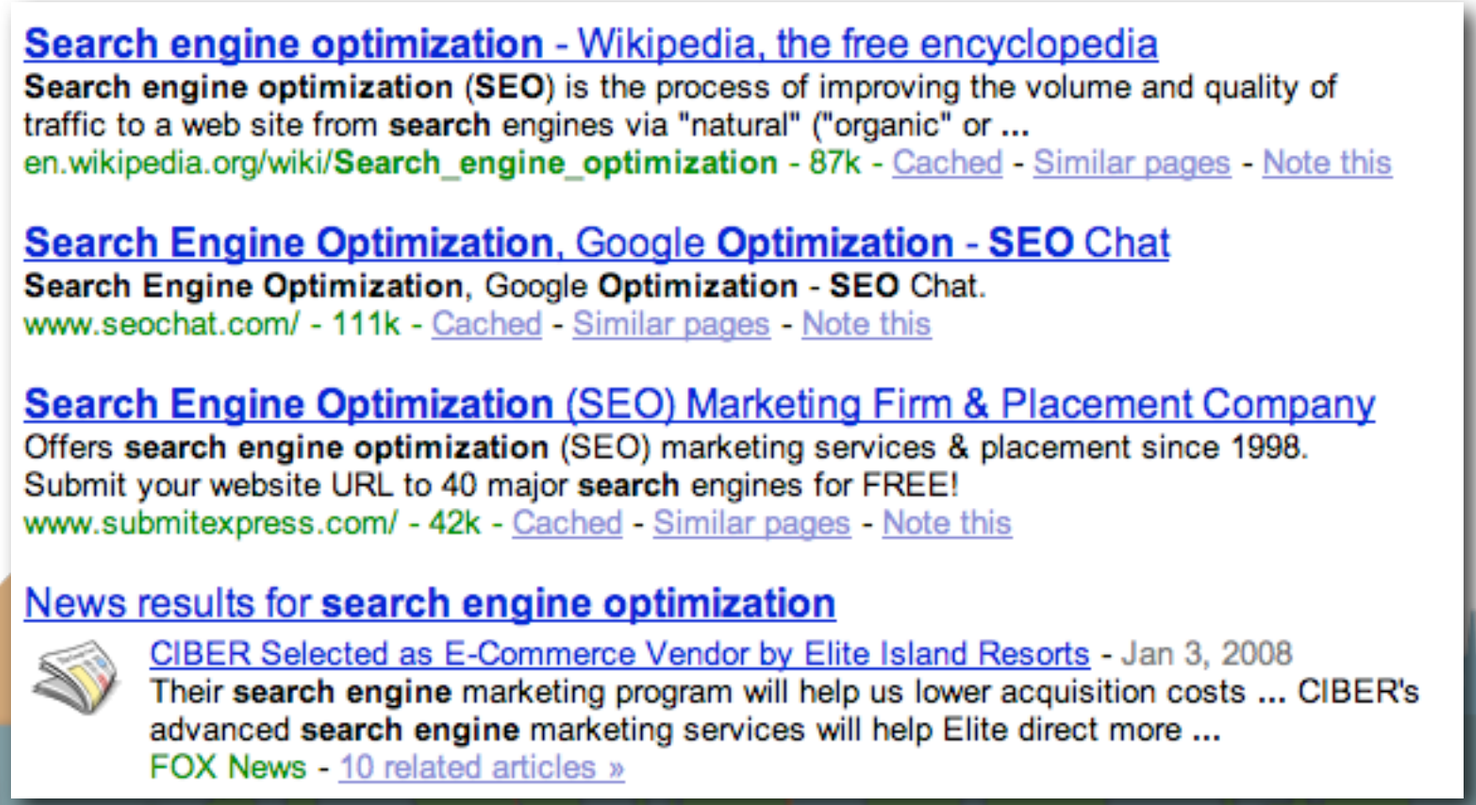
Sponsored Search



Ads vs. Search Results

- Google has maintained that ads (based on vendors bidding for search queries) do not affect vendors ranking in search results





Ranking of ads

- Other search engines (Yahoo!, MSN) have made similar statements on occasion
  - Any of them can change at any time
  - Facebook is currently testing the waters in their "Newsfeeds"
- We will ignore the possibility of paid placement ads being interspersed in search results.

Ranking of ads

- Goto model:
  - Rank according to how much advertiser pays
- Current model:
  - Balance auction price and relevance
    - Irrelevant ads (few click-throughs)
      - Decrease opportunities for relevant ads
      - Harm the user experience
    - Idea: Well-targeted advertising is good for everyone