Web Information Retrieval

Academic year 2020/2021

Instructors

- Luca Becchetti
- Fabrizio Silvestri

A quick tour of (tentative) topics

What is Web IR

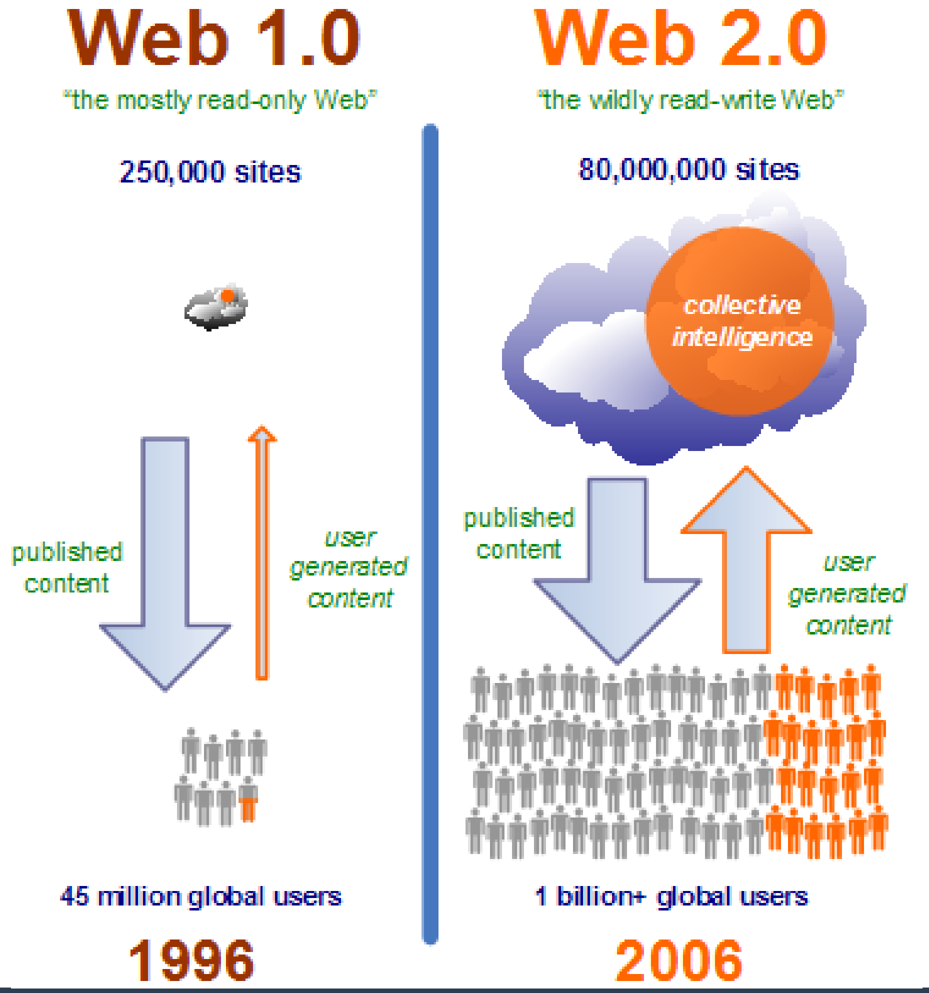
- **Information retrieval when the corpus is the Web**
- **Information Retrieval (IR)**
  - Retrieving unstructured material (usually textual documents) meeting an information need from large collections (usually stored on computers)
- **Live example**

Why is the Web different?

1) Distributed and larger than traditional information resources
2) Linked
3) Evolving
4) Information is semi-structured → view source of HTML pages
5) Multiple-content types (i.e. images, scripts, text etc.) coming in different formats
6) Quality of documents is not homogeneous

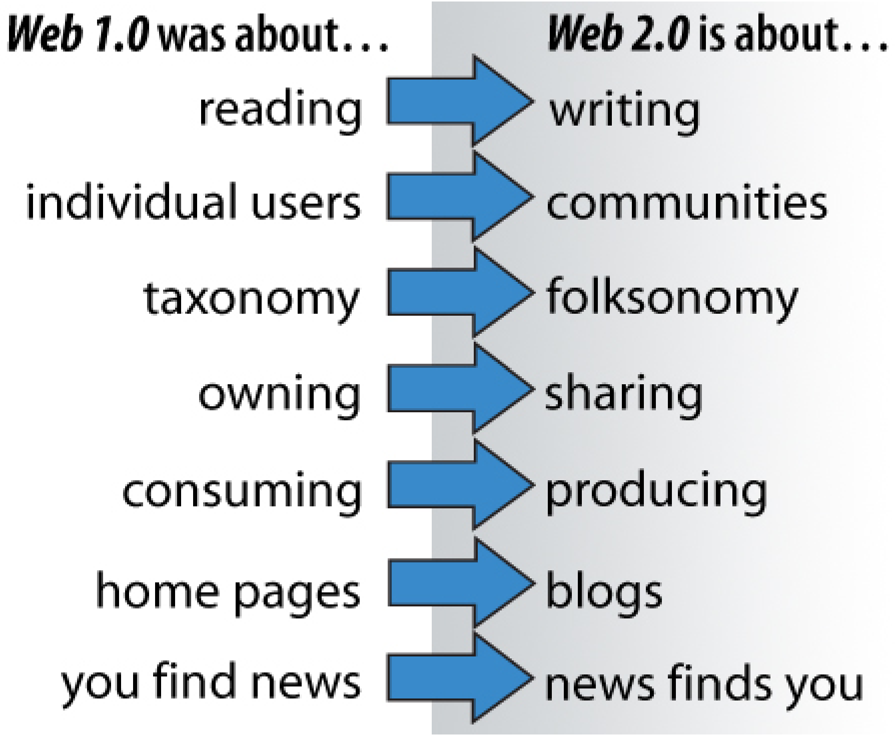
The Web

- As it was
- 10 years later



The Web

- As it was
- As it is



Web 1.0 was about…

Web 2.0 is about…

reading → writing

individual users → communities

taxonomy → folksonomy

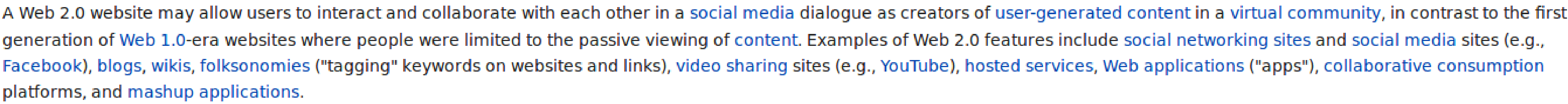
owning → sharing

consuming → producing

home pages → blogs

you find news → news finds you

According to Wikipedia



A Web 2.0 website may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated content in a virtual community, in contrast to the first generation of Web 1.0-era websites where people were limited to the passive viewing of content. Examples of Web 2.0 features include social networking sites and social media sites (e.g., Facebook), blogs, wikis, folksonomies ("tagging" keywords on websites and links), video sharing sites (e.g., YouTube), hosted services, Web applications ("apps"), collaborative consumption platforms, and mashup applications.

Course outline

- Collecting a Web corpus
- Pre-processing and organizing a Web corpus
- (Web) document retrieval (querying and searching the corpus)
- Analyzing documents using NLP
- Web page classification and clustering

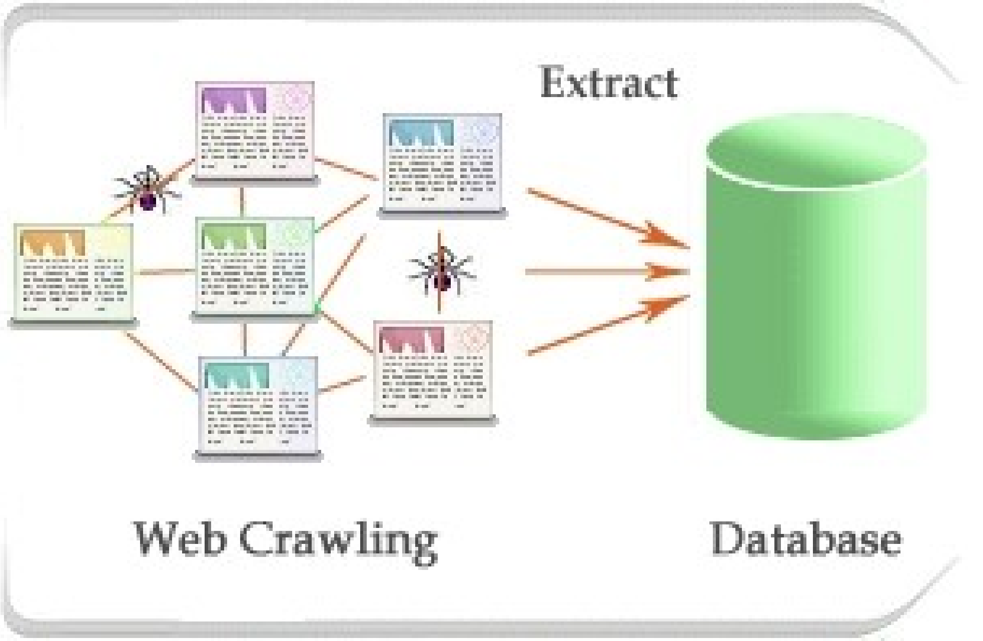
Course outline

- Collecting a Web corpus
- Pre-processing and organizing a Web corpus
- (Web) document retrieval (querying and searching the corpus)
- Using the Web as a platform to provide services

Collecting a Web corpus

Crawling the Web

- **Exploit link structure**
- **Simplified scheme**
  - Start from an initial page
  - Retrieve all linked pages
  - Iterate on new pages
- **Example**
- **At this point you should have**

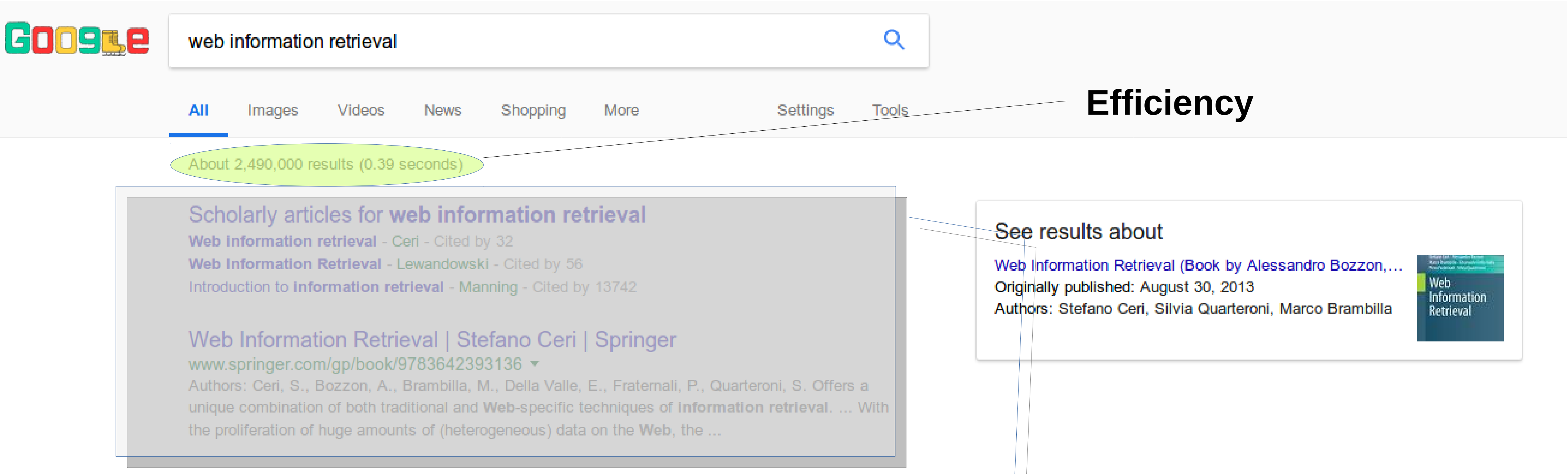


Crawling/Caveats and traps

- **Design/algorithmic challenges**
  - E.g.: Multiple Web crawlers
    - How to ensure we are not crawling the same pages?
  - We are not visiting all pages
    - Bias in data
- **Web applications**
  - e.g., social networking platforms
    - Not all pages accessible
    - Specific APIs/restrictions

Organizing a Web corpus

We have goals in mind – e.g.



web information retrieval

About 2,490,000 results (0.39 seconds)

Scholarly articles for **web information retrieval**

Web information retrieval - Ceri - Cited by 32

Web Information Retrieval - Lewandowski - Cited by 56

Introduction to **information retrieval** - Manning - Cited by 13742

Web Information Retrieval | Stefano Ceri | Springer

www.springer.com/gp/book/9783642393136

Authors: Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., Quarteroni, S. Offers a unique combination of both traditional and **Web**-specific techniques of **information retrieval**. ... With the proliferation of huge amounts of (heterogeneous) data on the **Web**, the ...

See results about

Web Information Retrieval (Book by Alessandro Bozzon,...

Originally published: August 30, 2013

Authors: Stefano Ceri, Silvia Quarteroni, Marco Brambilla

Efficiency

Relevance

How Google puts it …

- **https://www.google.com/search/howsearchworks/**

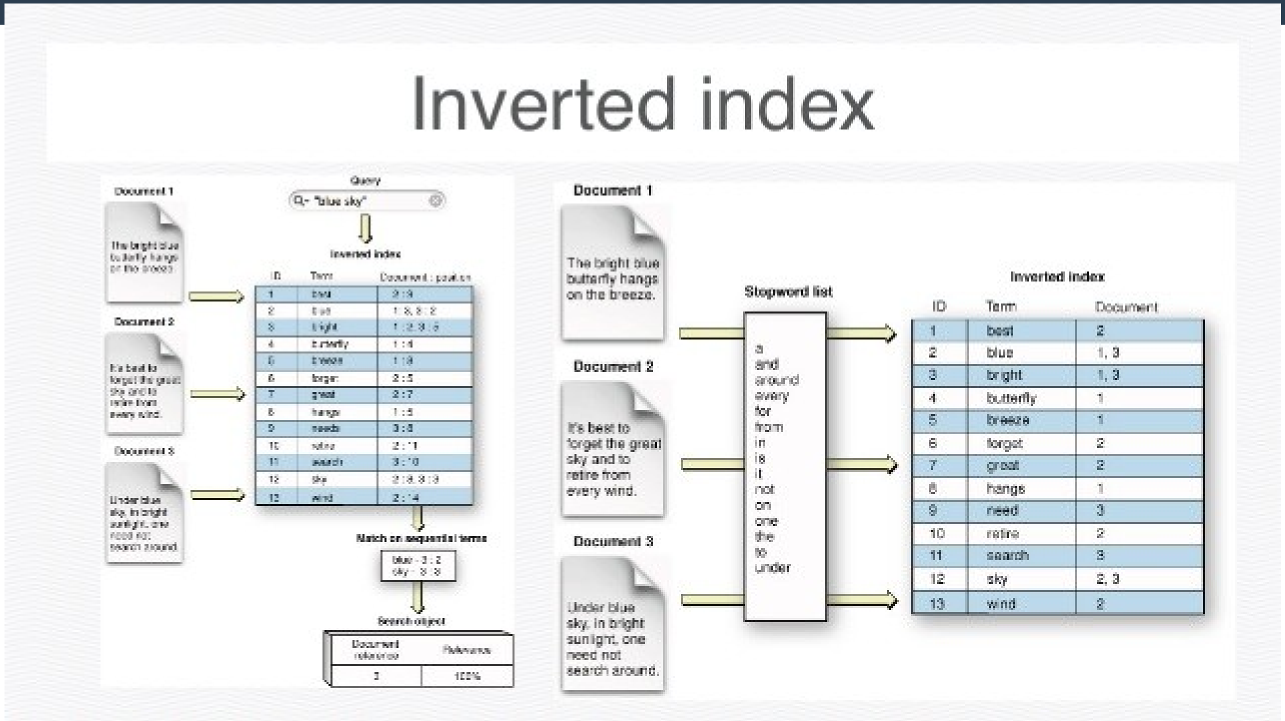
How Google puts it ...

- **https://www.google.com/search/howsearchworks/**
- **In a nutshell**
  - Crawling
  - Indexing
  - Search algorithms

Indexing

- **Organize Web corpus so as to efficiently answer (implicit or explicit) queries**
- **Challenging task**
  - Multiple objectives
  - Multiple trade-offs

Efficient data structures



Inverted index

Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Query
"blue sky"
Inverted index

ID	Term	Document : position
1	best	2 : 3
2	blue	1 : 3, 3 : 2
3	bright	1 : 2, 3 : 5
4	butterfly	1 : 4
5	breeze	1 : 8
6	forget	2 : 5
7	great	2 : 7
8	hangs	1 : 5
9	needs	3 : 8
10	retire	2 : 11
11	search	3 : 10
12	sky	2 : 8, 3 : 3
13	wind	2 : 14

Match on sequential terms
blue - 3 : 2
sky - 3 : 3
Search object

Document reference	Relevance
3	100%

Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Stopword list
a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
Inverted index

ID	Term	Document
1	best	2
2	blue	1, 3
3	bright	1, 3
4	butterfly	1
5	breeze	1
6	forgot	2
7	great	2
8	hangs	1
9	need	3
10	retire	2
11	search	3
12	sky	2, 3
13	wind	2

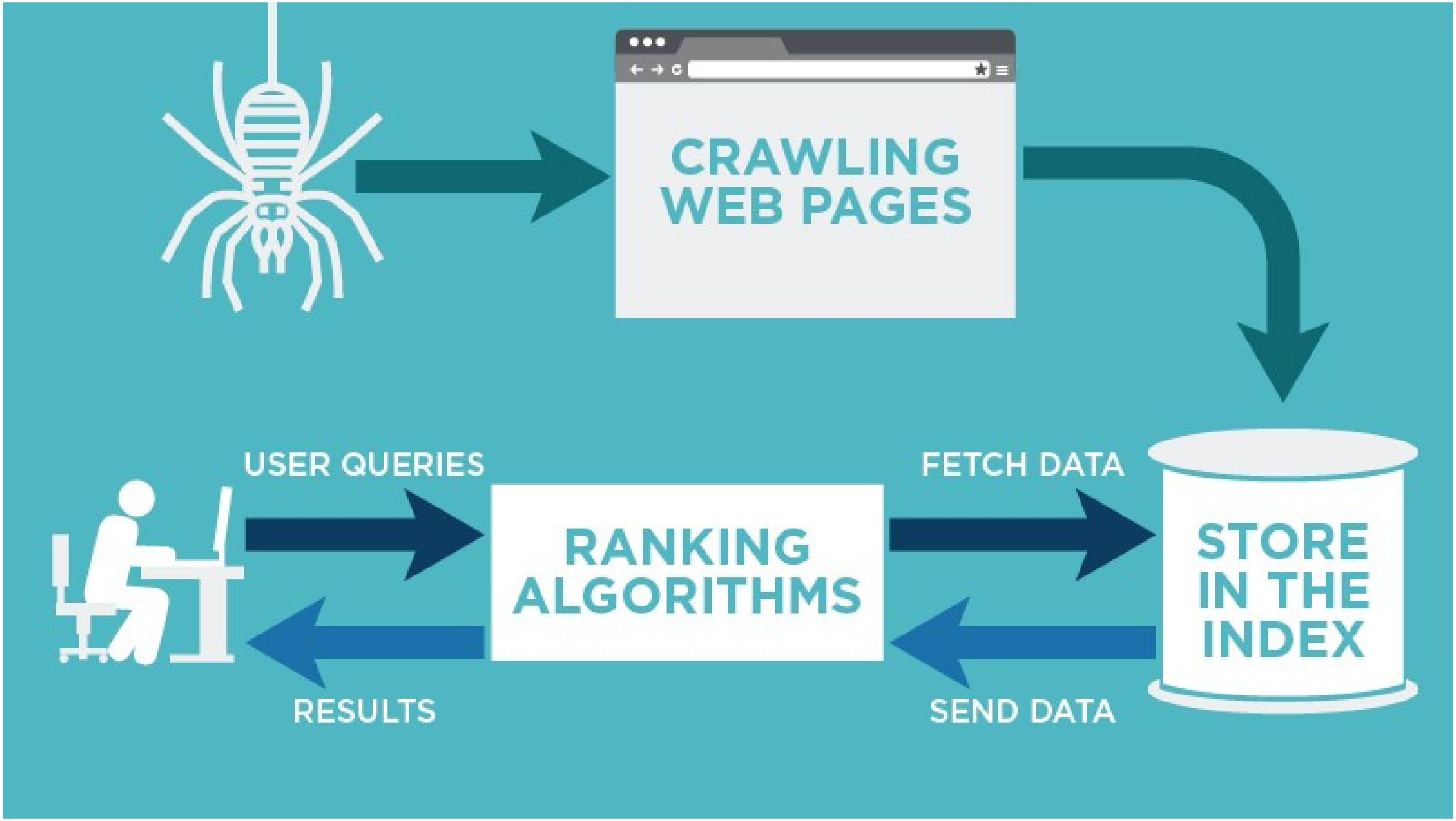
- **Typically an inverted index**
  - Index construction
  - Search using an inverted index
  - Compression, metadata enrichment ...

Querying the corpus (search)

Goals and document scoring

- **Return documents that are relevant to the query “web information retrieval”**
- **How to define and measure relevance**
  - Textual analysis
    - Use meta-data when available
  - Link analysis (Web structure)
  - Pages can be “more” or “less” relevant → ranking
- **Relevance vs authority**

The final picture



The recent past ...

Providing services over the Web

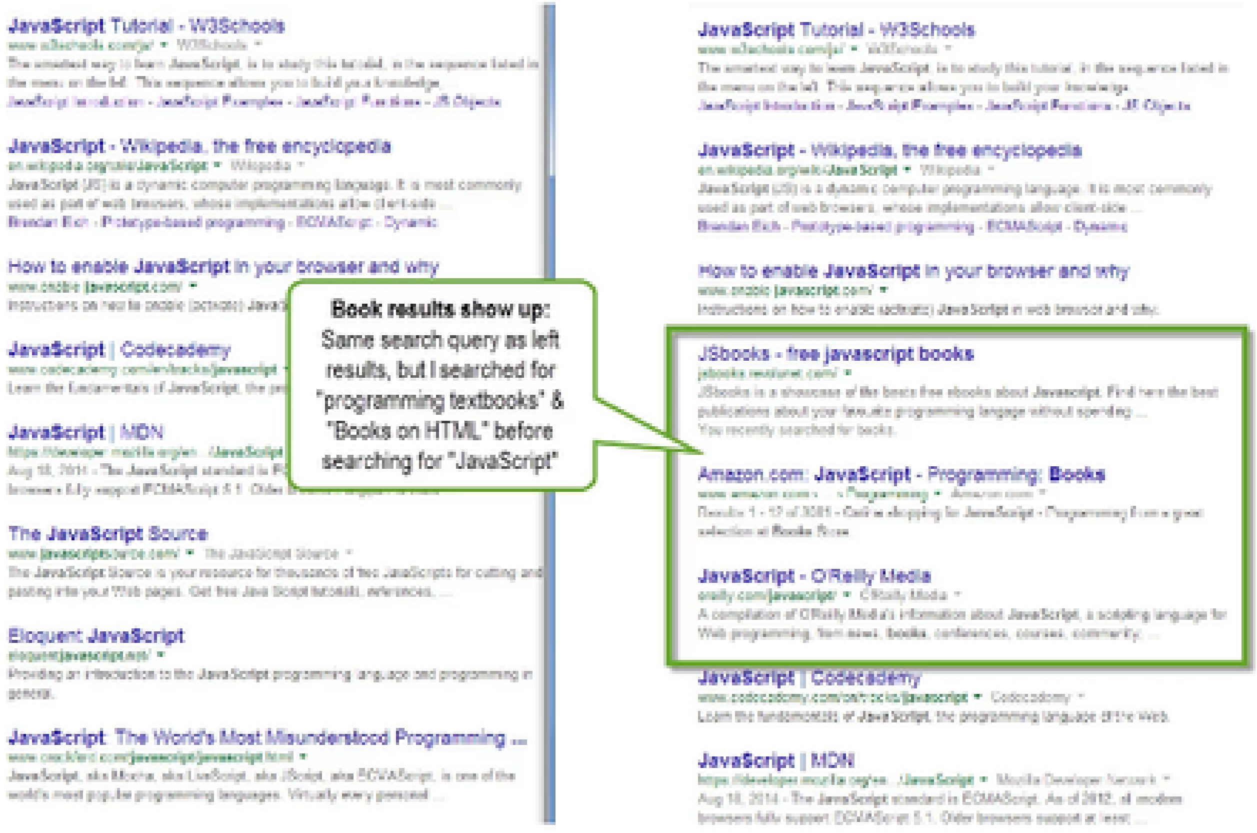
- **Search engines**
  - Web applications providing search
- **More have emerged over the recent past …**

Now (social networking only)

CouchSurfing project
gossipreport.com
gamerDNA
gaia online
Meetcha
Naymz
WISER EARTH
fubar
Flixster
FledgeWing
currensee
film affinity
experience project
cafemom
bigtent
PlentyOfFish
classmates.com
AMIE ST
bebo
mixi
Ning
Helium
care2
itimes
orkut
facebook
hi5
socialgrow
YouTube
twitter
connexion
badoo
NETLOG
XING
babycenter
cyworld
last.fm
brijj.com
LinkedIn
flickr
AVATARS UNITED
disaboom.com
deviantART
ecademy
Cake FINANCIAL
yammer
DailyStrength
gather
Groupsite
goodreads
DOOSTANG
BIGADDA
EntrepreneurConnect
epernicus
ASMALLWORLD
©2010 SocialGrow

New approaches/challenges

Personalization



JavaScript Tutorial - W3Schools
www.w3schools.com/js/ ▾ W3Schools ▾
The smartest way to learn JavaScript, is to study this tutorial, in the sequence listed in the menu on the left. This sequence allows you to build your knowledge ...
JavaScript Introduction - JavaScript Examples - JavaScript Functions - JS Objects

JavaScript - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/JavaScript ▾ Wikipedia ▾
JavaScript (JS) is a dynamic computer programming language. It is most commonly used as part of web browsers, whose implementations allow client-side ...
Brendan Eich - Prototype-based programming - ECMAScript - Dynamic

How to enable JavaScript in your browser and why
www.enable-javascript.com/ ▾
Instructions on how to enable (activate) Java

JavaScript | Codecademy
www.codecademy.com/en/tracks/javascript ▾
Learn the fundamentals of JavaScript, the pr

JavaScript | MDN
https://developer.mozilla.org/en/.../JavaScript
Aug 18, 2014 - The JavaScript standard is EC
browsers fully support ECMAScript 5.1. Older b

The JavaScript Source
www.javascriptsource.com/ ▾ The JavaScript Source ▾
The JavaScript Source is your resource for thousands of free JavaScripts for cutting and pasting into your Web pages. Get free JavaScript tutorials, references, ...

Eloquent JavaScript
eloquentjavascript.net/ ▾
Providing an introduction to the JavaScript programming language and programming in general.

JavaScript: The World's Most Misunderstood Programming ...
www.crockford.com/javascript/javascript.html ▾
JavaScript, aka Mocha, aka LiveScript, aka JScript, aka ECMAScript, is one of the world's most popular programming languages. Virtually every personal ...

Book results show up:
Same search query as left results, but I searched for "programming textbooks" & "Books on HTML" before searching for "JavaScript"

JavaScript Tutorial - W3Schools
www.w3schools.com/js/ ▾ W3Schools ▾
The smartest way to learn JavaScript, is to study this tutorial, in the sequence listed in the menu on the left. This sequence allows you to build your knowledge ...
JavaScript Introduction - JavaScript Examples - JavaScript Functions - JS Objects

JavaScript - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/JavaScript ▾ Wikipedia ▾
JavaScript (JS) is a dynamic computer programming language. It is most commonly used as part of web browsers, whose implementations allow client-side ...
Brendan Eich - Prototype-based programming - ECMAScript - Dynamic

How to enable JavaScript in your browser and why
www.enable-javascript.com/ ▾
Instructions on how to enable (activate) JavaScript in web browser and why.

JSbooks - free javascript books
jsbooks.revolunet.com/ ▾
JSbooks is a showcase of the best free ebooks about Javascript. Find here the best publications about your favourite programming langage without spending ...
You recently searched for books.

Amazon.com: JavaScript - Programming: Books
www.amazon.com › ... › Programming ▾ Amazon.com ▾
Results 1 - 12 of 3081 - Online shopping for JavaScript - Programming from a great selection at Books Store.

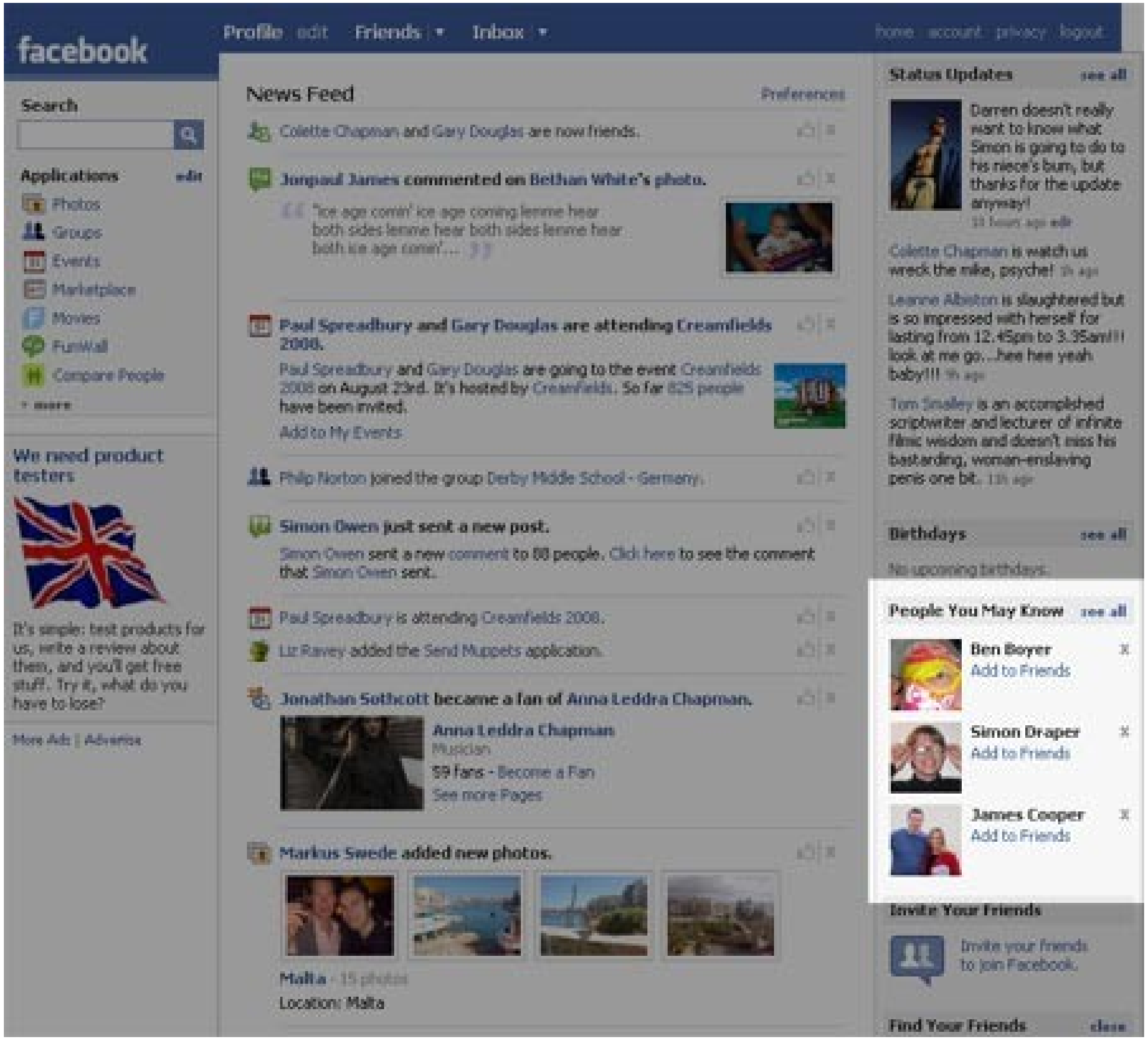
JavaScript - O'Reilly Media
oreilly.com/javascript/ ▾ O'Reilly Media ▾
A compilation of O'Reilly Media's information about JavaScript, a scripting language for Web programming, from news, books, conferences, courses, community, ...

JavaScript | Codecademy
www.codecademy.com/en/tracks/javascript ▾ Codecademy ▾
Learn the fundamentals of JavaScript, the programming language of the Web.

JavaScript | MDN
https://developer.mozilla.org/en/.../JavaScript ▾ Mozilla Developer Network ▾
Aug 18, 2014 - The JavaScript standard is ECMAScript. As of 2012, all modern browsers fully support ECMAScript 5.1. Older browsers support at least ...

Your profile impacts the outcome

Recommendations



News Feed

Colette Chapman and Gary Douglas are now friends.

Jonpaul James commented on Bethan White's photo.

"Ice age comin' ice age coming lemme hear both sides lemme hear both sides lemme hear both ice age comin'..."

Paul Spreadbury and Gary Douglas are attending Creamfields 2008.

Paul Spreadbury and Gary Douglas are going to the event Creamfields 2008 on August 23rd. It's hosted by Creamfields. So far 825 people have been invited.

Add to My Events

Philip Norton joined the group Derby Middle School - Germany.

Simon Owen just sent a new post.

Simon Owen sent a new comment to 88 people. Click here to see the comment that Simon Owen sent.

Paul Spreadbury is attending Creamfields 2008.

Liz Ravey added the Send Muppets application.

Jonathan Sothcott became a fan of Anna Leddra Chapman.

Anna Leddra Chapman
Musician
59 fans - Become a Fan
See more Pages

Markus Swede added new photos.

Malta - 15 photos
Location: Malta

We need product testers

It's simple: test products for us, write a review about them, and you'll get free stuff. Try it, what do you have to lose?

Status Updates

Darren doesn't really want to know what Simon is going to do to his niece's bum, but thanks for the update anyway! 11 hours ago edit

Colette Chapman is watch us wreck the mike, psyche! 1h ago

Leanne Albiston is slaughtered but is so impressed with herself for lasting from 12.45pm to 3.35am!!! look at me go....hee hee yeah baby!!! 9h ago

Tom Smalley is an accomplished scriptwriter and lecturer of infinite filmic wisdom and doesn't miss his bastarding, woman-enslaving penis one bit. 11h ago

Birthdays

No upcoming birthdays.

People You May Know

Ben Boyer
Add to Friends

Simon Draper
Add to Friends

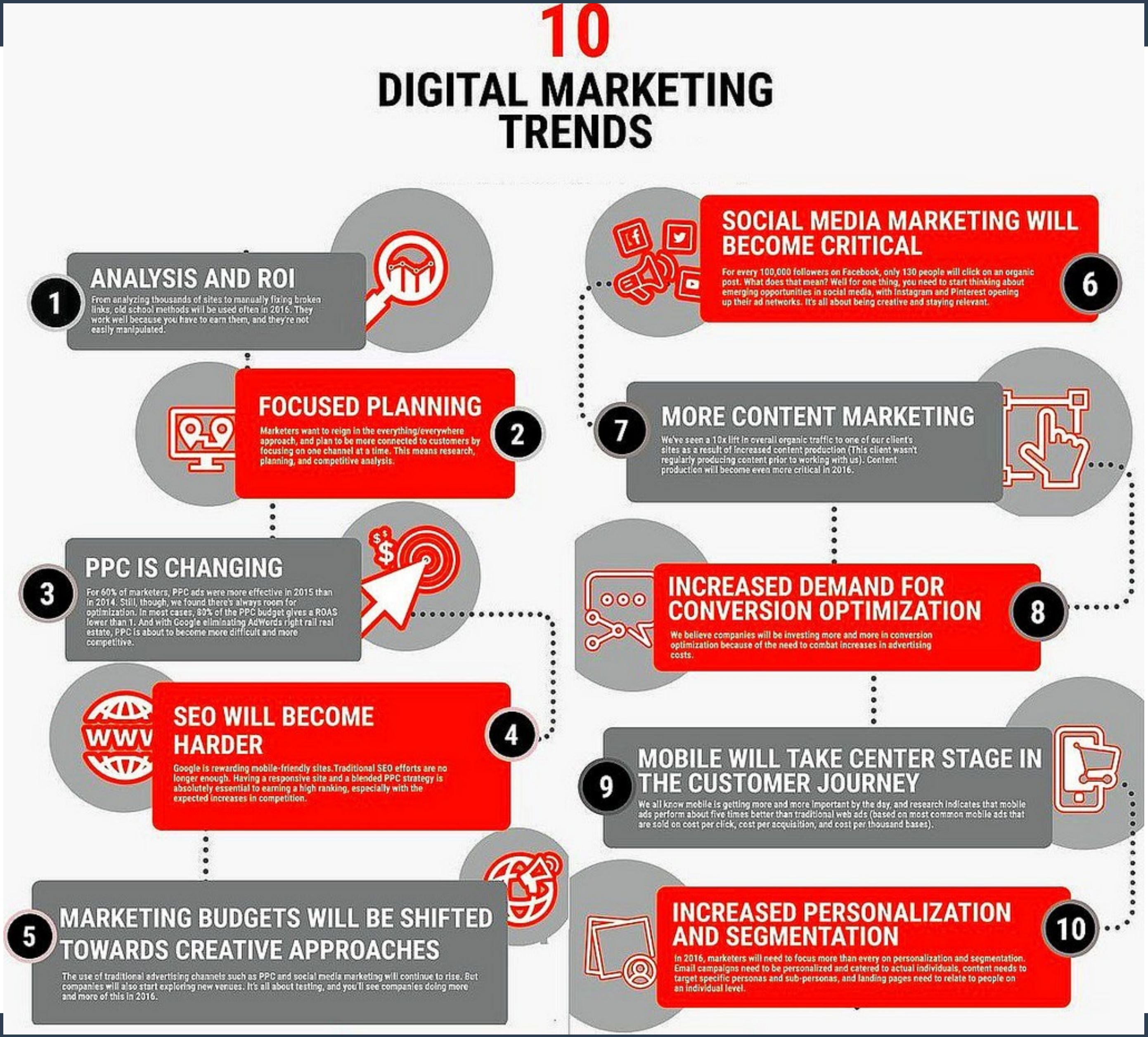
James Cooper
Add to Friends

Invite Your Friends

Invite your friends to join Facebook.

Find Your Friends

…and much more



10 DIGITAL MARKETING TRENDS

1 **ANALYSIS AND ROI**

From analyzing thousands of sites to manually fixing broken links, old school methods will be used often in 2016. They work well because you have to earn them, and they're not easily manipulated.

2 **FOCUSED PLANNING**

Marketers want to reign in the everything/everywhere approach, and plan to be more connected to customers by focusing on one channel at a time. This means research, planning, and competitive analysis.

3 **PPC IS CHANGING**

For 60% of marketers, PPC ads were more effective in 2015 than in 2014. Still, though, we found there's always room for optimization. In most cases, 80% of the PPC budget gives a ROAS lower than 1. And with Google eliminating AdWords right rail real estate, PPC is about to become more difficult and more competitive.

4 **SEO WILL BECOME HARDER**

Google is rewarding mobile-friendly sites.Traditional SEO efforts are no longer enough. Having a responsive site and a blended PPC strategy is absolutely essential to earning a high ranking, especially with the expected increases in competition.

5 **MARKETING BUDGETS WILL BE SHIFTED TOWARDS CREATIVE APPROACHES**

The use of traditional advertising channels such as PPC and social media marketing will continue to rise. But companies will also start exploring new venues. It's all about testing, and you'll see companies doing more and more of this in 2016.

6 **SOCIAL MEDIA MARKETING WILL BECOME CRITICAL**

For every 100,000 followers on Facebook, only 130 people will click on an organic post. What does that mean? Well for one thing, you need to start thinking about emerging opportunities in social media, with Instagram and Pinterest opening up their ad networks. It's all about being creative and staying relevant.

7 **MORE CONTENT MARKETING**

We've seen a 10x lift in overall organic traffic to one of our client's sites as a result of increased content production (This client wasn't regularly producing content prior to working with us). Content production will become even more critical in 2016.

8 **INCREASED DEMAND FOR CONVERSION OPTIMIZATION**

We believe companies will be investing more and more in conversion optimization because of the need to combat increases in advertising costs.

9 **MOBILE WILL TAKE CENTER STAGE IN THE CUSTOMER JOURNEY**

We all know mobile is getting more and more important by the day, and research indicates that mobile ads perform about five times better than traditional web ads (based on most common mobile ads that are sold on cost per click, cost per acquisition, and cost per thousand bases).

10 **INCREASED PERSONALIZATION AND SEGMENTATION**

In 2016, marketers will need to focus more than every on personalization and segmentation. Email campaigns need to be personalized and catered to actual individuals, content needs to target specific personas and sub-personas, and landing pages need to relate to people on an individual level.

Practical info

General info

- **Where**
  - Room A3, via Ariosto 25
- **When**
  - Mondays, 2pm – 5pm
  - Thursdays, 2pm – 4pm
- **Luca Becchetti**
  - becchetti@diag.uniroma1.it
- **Fabrizio Silvestri**
  - fsilvestri@diag.uniroma1.it
- **General info, announcements etc.**
  - https://classroom.google.com/u/0/c/MjcxOTA0NTgyNjEy
    - Please enroll!!

Organization

- **Lectures**
  - New topics
  - Discussions
  - Homeworks
- **Hands on (hopefully)**
  - We try to solve problems together
    - Emphasis on together
    - Bring your laptop if you have one
  - First year – we'll do our best

- **Exam**
  - Possibly: written exam + assignments
  - Details to be decided

More info

- **Prerequisites**
  - Undergraduate in CS or equivalent

- **Useful things**
  - A laptop
  - Curiosity and independence
  - Presence and participation

- **References**
  - Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1. No. 1. Cambridge: Cambridge university press, 2008
    - Thanks to the authors, pdf of chapters is available for free at the book’s Web site
  - Scientific papers
  - On-line material, tutorials etc.