

Web Information Retrieval

Academic year 2017/2018

Instructors

- **Luca Becchetti**
- **Andrea Vitaletti**



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**

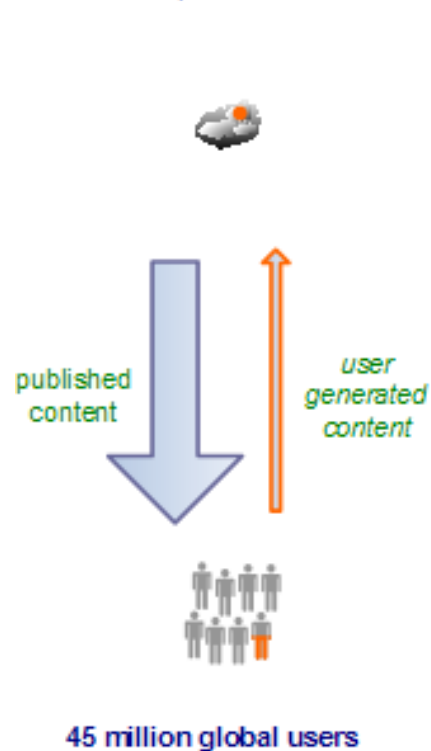


The Web

- **As it was**
Web 1.0
- **As it is**
Web 2.0

"the mostly read-only Web"

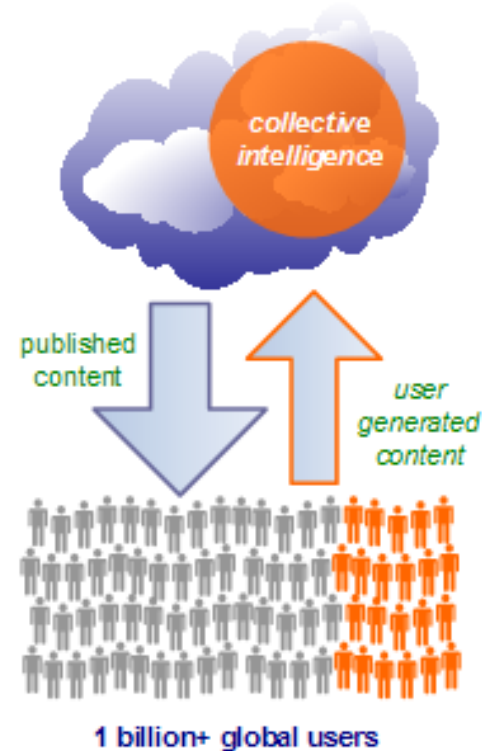
250,000 sites



1996

"the wildly read-write Web"

80,000,000 sites



2006



The Web



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)**
- **Using the Web as a platform to provide services**



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**

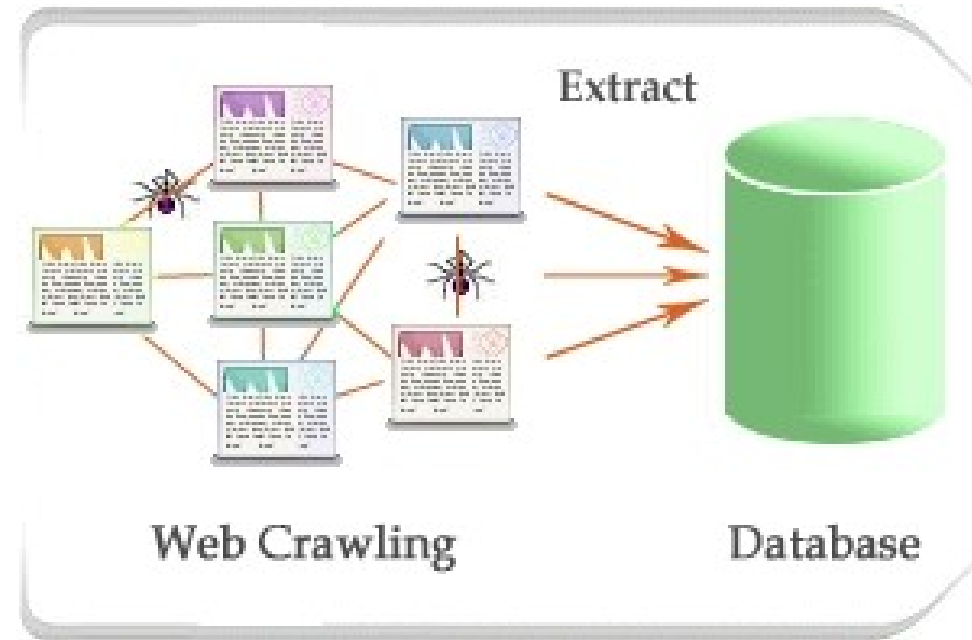


Collecting a Web corpus



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.

The image shows a Google search interface for the query "web information retrieval". The search bar at the top contains the text "web information retrieval" and a magnifying glass icon. Below the search bar, the "All" tab is selected, and the search results are displayed. A green oval highlights the text "About 2,490,000 results (0.39 seconds)", with a line pointing to the word "Efficiency". Below this, a grey box contains a list of scholarly articles for "web information retrieval", including "Web Information retrieval - Ceri - Cited by 32", "Web Information Retrieval - Lewandowski - Cited by 56", and "Introduction to Information retrieval - Manning - Cited by 13742". A blue box highlights the book "Web Information Retrieval | Stefano Ceri | Springer" with its URL and authors. A line points from this box to the word "Relevance". To the right, a white box titled "See results about" lists the book "Web Information Retrieval (Book by Alessandro Bozzon,...)" with its publication date and authors. A line points from this box to the word "Relevance". A small book cover image is also visible next to the book title.

Google

web information retrieval

All Images Videos News Shopping More Settings Tools

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

Web Information Retrieval

Efficiency

Relevance

How Google puts it ...

- <https://www.google.com/search/howsearchworks/>

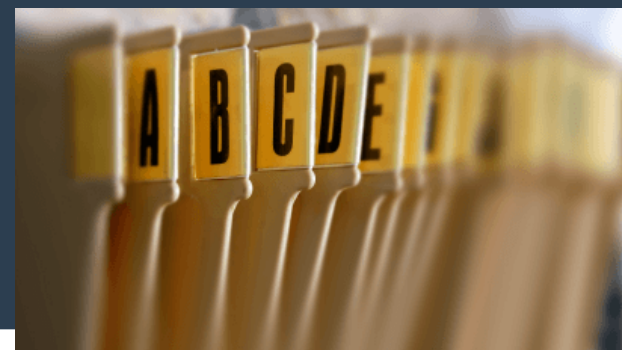


How Google puts it ...

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



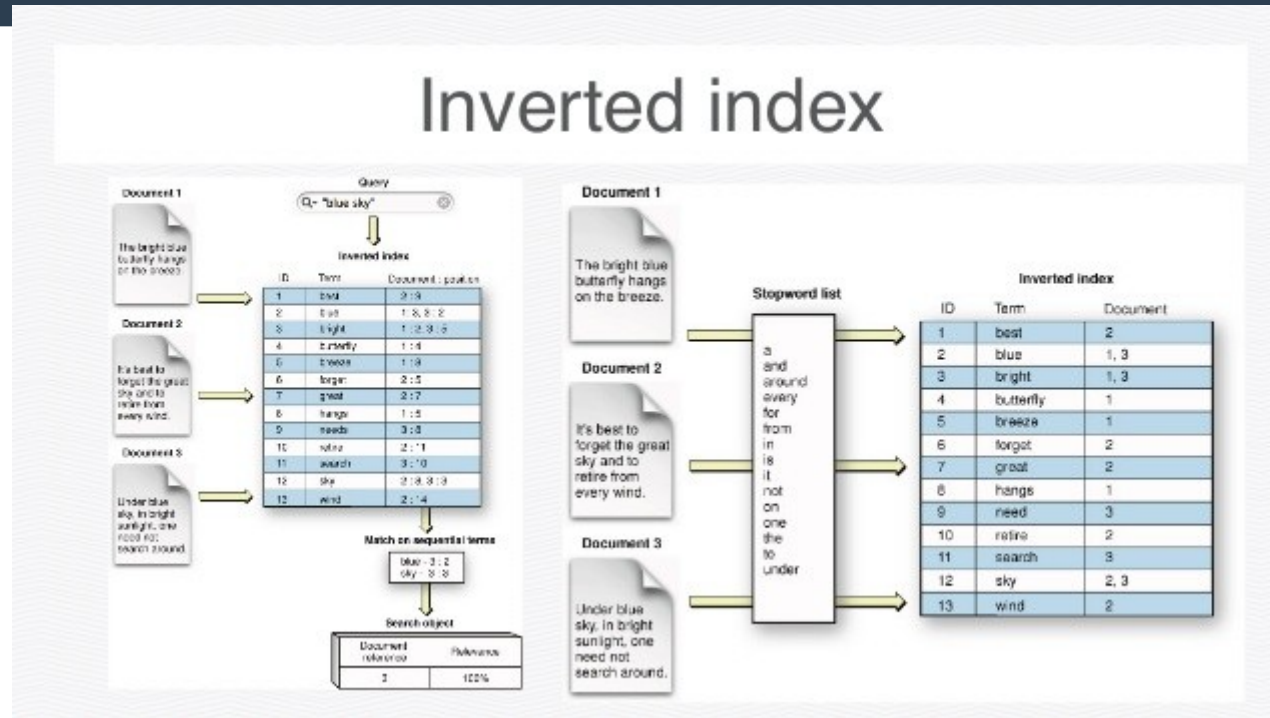
Indexing



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



Efficient data structures



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



Querying the corpus (search)

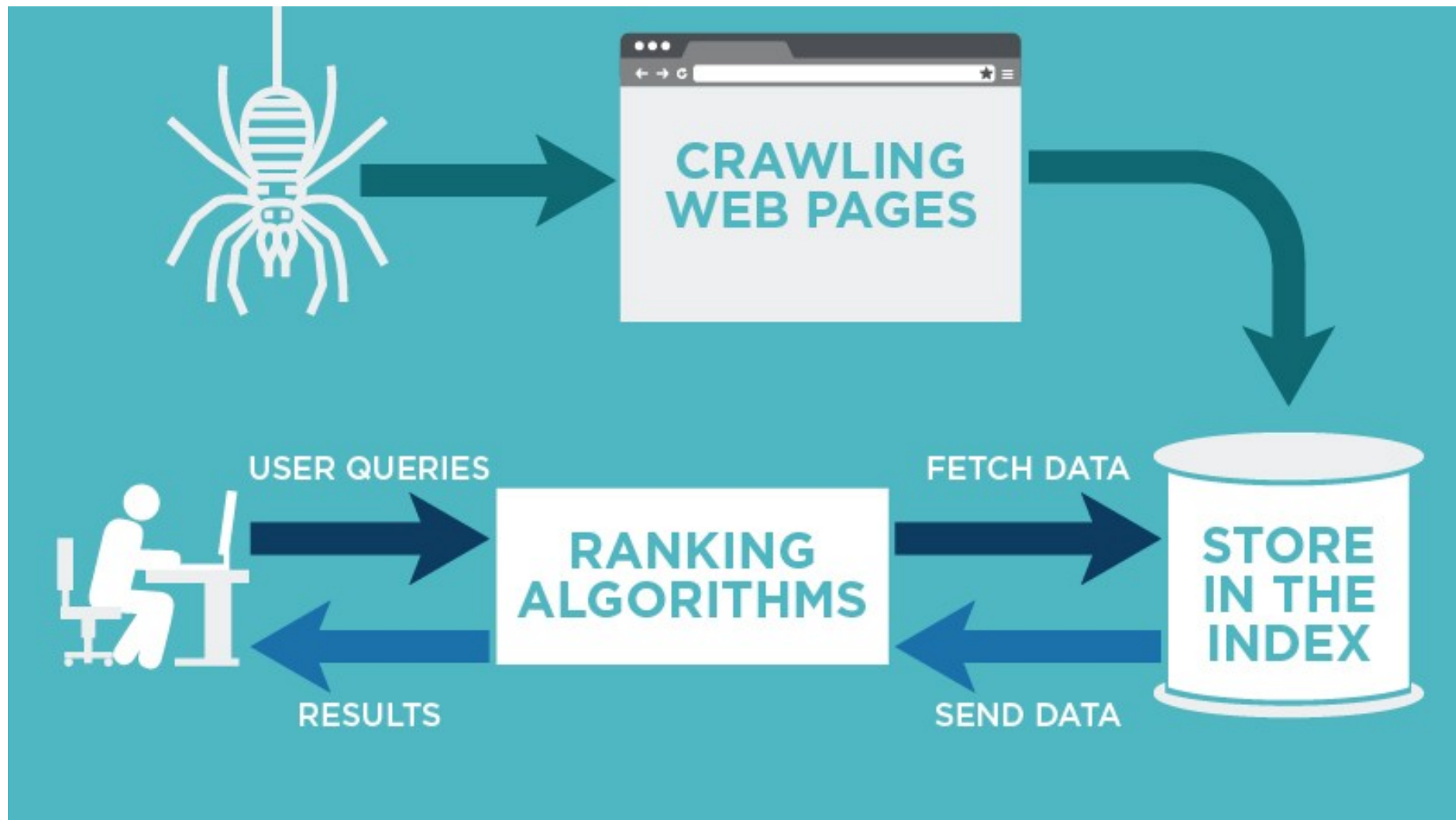


Goals and document scoring

- **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 Web as a platform



Providing services over the Web

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



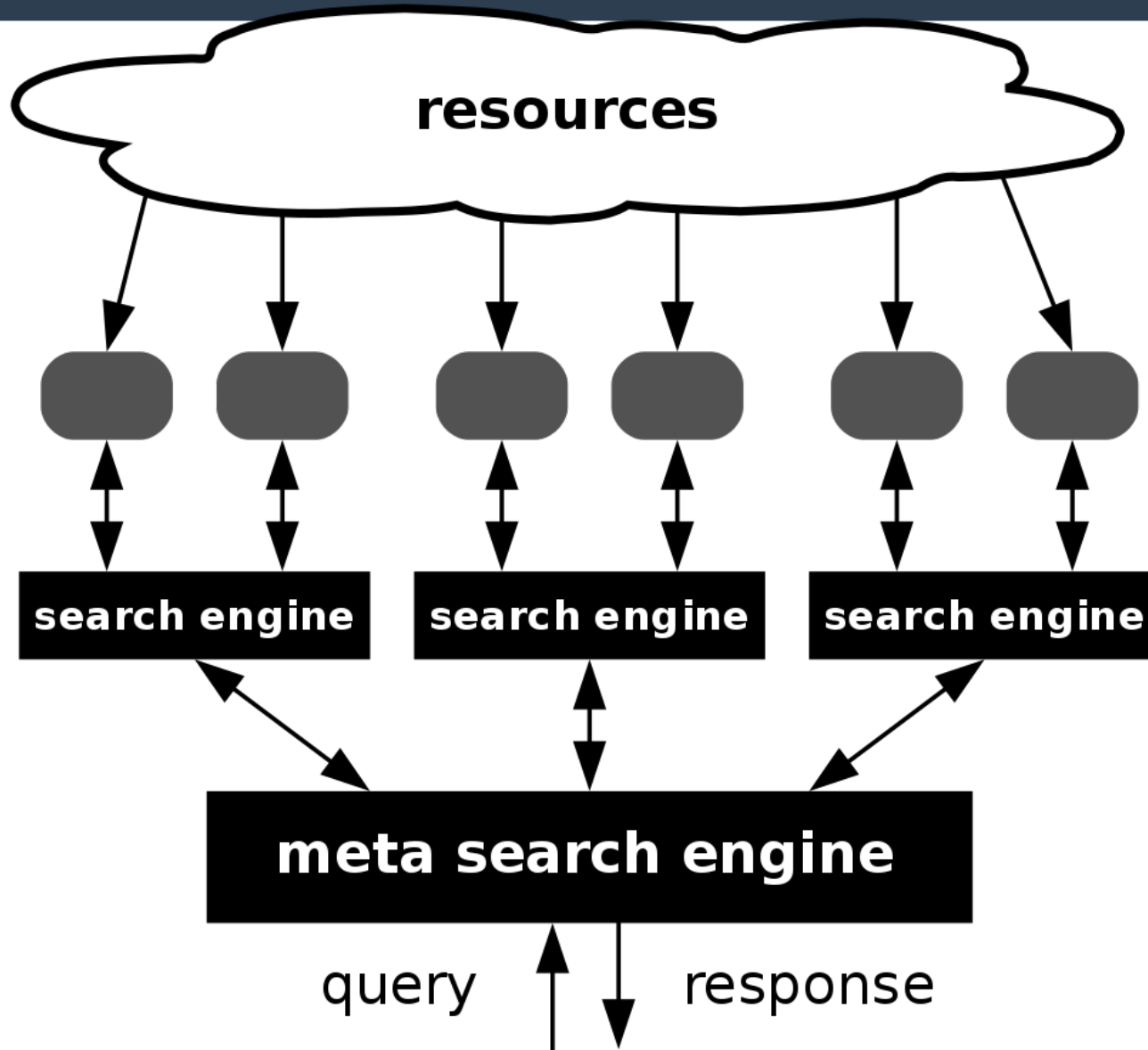
Now (social networking only)



New approaches/challenges



Meta-search



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 Frameworks](#) • [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.

JavaScript | Codecademy
www.codecademy.com/learn/tracks/javascript •
Learn the Fundamentals of JavaScript, the pre

JavaScript | MDN
<https://developer.mozilla.org/en-US/docs/Web/JavaScript>
Aug 18, 2014 • The JavaScript standard is ECMA-262. As of 2012, all modern browsers fully support ECMA-262 5.1. Older browsers support at least ...

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 Frameworks](#) • [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.reinart.net/ •
JSBooks is a showcase of the best free books about Javascript. Find here the best publications about your favorite programming language without spending ...
You recently searched for books.

Amazon.com: JavaScript - Programming: Books
www.amazon.com/ • [Programming](#) • [Amazon.com](#) •
Results 1 - 17 of 3085 • Click to shopping for JavaScript - Programming from a great selection of Books, Books

JavaScript - O'Reilly Media
oreil.ly/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/learn/tracks/javascript • Codecademy •
Learn the Fundamentals of JavaScript, the programming language of the Web.

JavaScript | MDN
<https://developer.mozilla.org/en-US/docs/Web/JavaScript> • Mozilla Developer Network •
Aug 18, 2014 • The JavaScript standard is ECMA-262. As of 2012, all modern browsers fully support ECMA-262 5.1. Older browsers support at least ...

Your profile impacts the outcome

Recommendations

The screenshot displays the Facebook homepage layout. On the left sidebar, there is a search bar, a list of applications (Photos, Groups, Events, Marketplace, Movies, FunWall, Compare People), and a promotional banner for 'We need product testers' featuring a Union Jack flag. The top navigation bar includes links for Profile, edit, Friends, and Inbox, along with home, account, privacy, and logout options. The central 'News Feed' contains several updates: Colette Chapman and Gary Douglas becoming friends, Jonpaul James commenting on Bethan White's photo, Paul Spreadbury and Gary Douglas attending Creamfields 2008, Philip Norton joining a group, Simon Owen sending a comment, Paul Spreadbury attending Creamfields 2008, Liz Ravely adding an application, Jonathan Southcott becoming a fan of Anna Leddra Chapman, and Markus Swede adding new photos. The right sidebar features 'Status Updates' with posts from Darren, Colette Chapman, Leanne Albiston, and Tom Smalley; 'Birthdays' showing no upcoming birthdays; 'People You May Know' with suggestions for Ben Boyer, Simon Draper, and James Cooper; and 'Invite Your Friends' and 'Find Your Friends' sections.

facebook Profile edit Friends ▾ Inbox ▾ home account privacy logout

Search

Applications edit

- Photos
- Groups
- Events
- Marketplace
- Movies
- FunWall
- Compare People

7 more

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?

More Ads | Advertise

News Feed Preferences

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 80 people. Click here to see the comment that Simon Owen sent.

Paul Spreadbury is attending Creamfields 2008.

Liz Ravely added the Send Muppets application.

Jonathan Southcott became a fan of Anna Leddra Chapman.

Anna Leddra Chapman

Musician

59 fans · Become a Fan

See more Pages

Markus Swede added new photos.

Malla · 15 photos

Location: Malla

Status Updates see all

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

13 hours ago · edit

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

Tom Smalley is an accomplished scriptwriter and lecturer of infinite filmic wisdom and doesn't miss his bastarding, woman-enslaving penis one bit. 13h ago

Birthdays see all

No upcoming birthdays.

People You May Know see all

Ben Boyer

Add to Friends

Simon Draper

Add to Friends

James Cooper

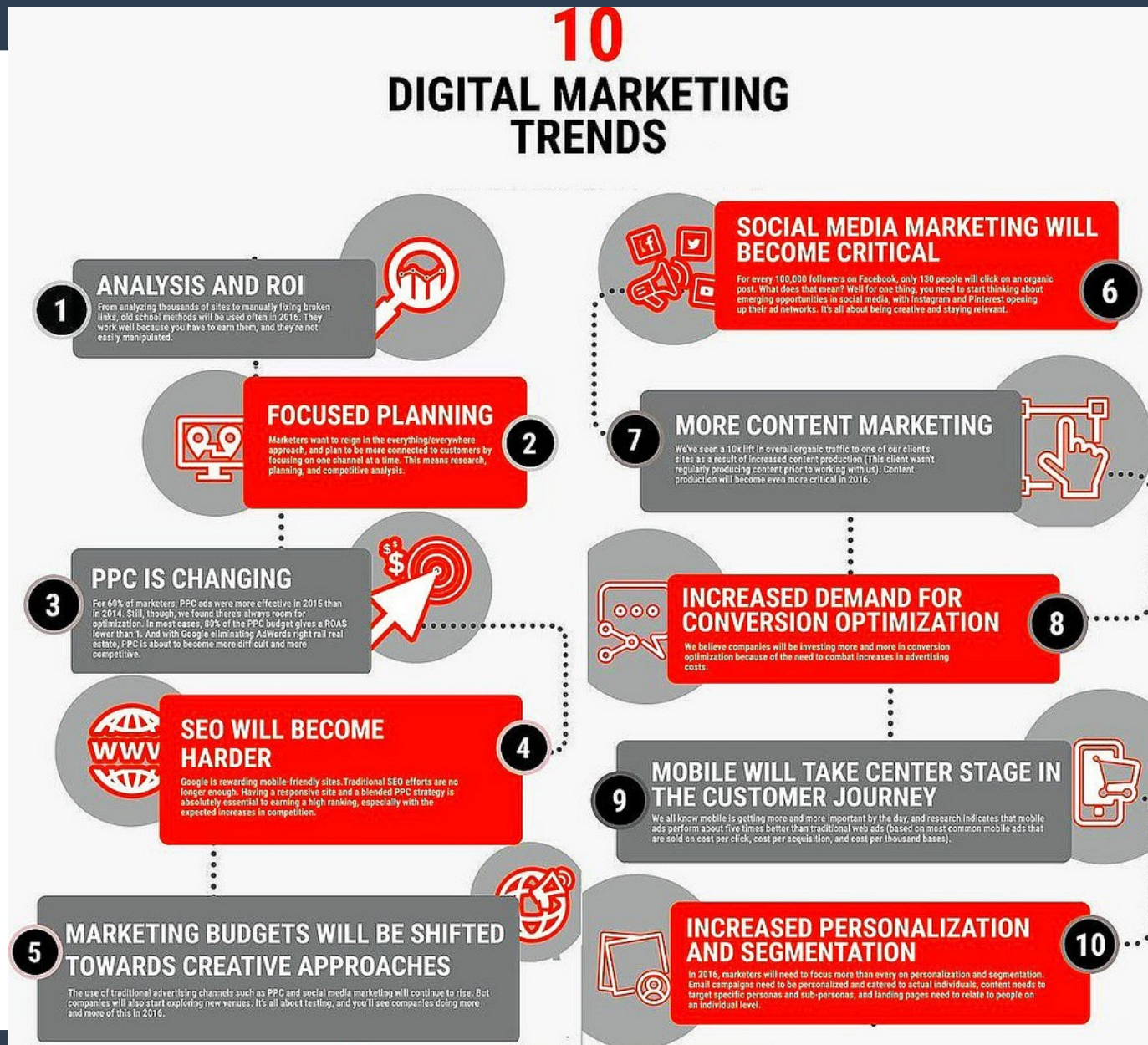
Add to Friends

Invite Your Friends

Invite your friends to join Facebook.

Find Your Friends close

...and much more



Practical info



General info

- **Where**
 - Room A3, via Ariosto 25
- **When**
 - Mondays, 8am – 11am
- **Luca Becchetti**
 - becchetti@diag.uniroma1.it
- **Andrea Vitaletti**
 - vitaletti@diag.uniroma1.it
- **General info, announcements etc.**
 - <https://piazza.com/uniroma1.it/spring2018/wir/home>
 - Please enroll!!
- **Syllabus and material**
 - <https://github.com/andreavitaletti/WIR/wiki>



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**

- Assignments/projects/homeworks: 50%
 - Details to be decided
- Written exam: 50%



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.
- Scientific papers
- On-line material, tutorials etc.

