Web Information Retrieval

Academic year 2017/2018

Instructors

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

A quick tour of (tentative) topics

What is Web IR

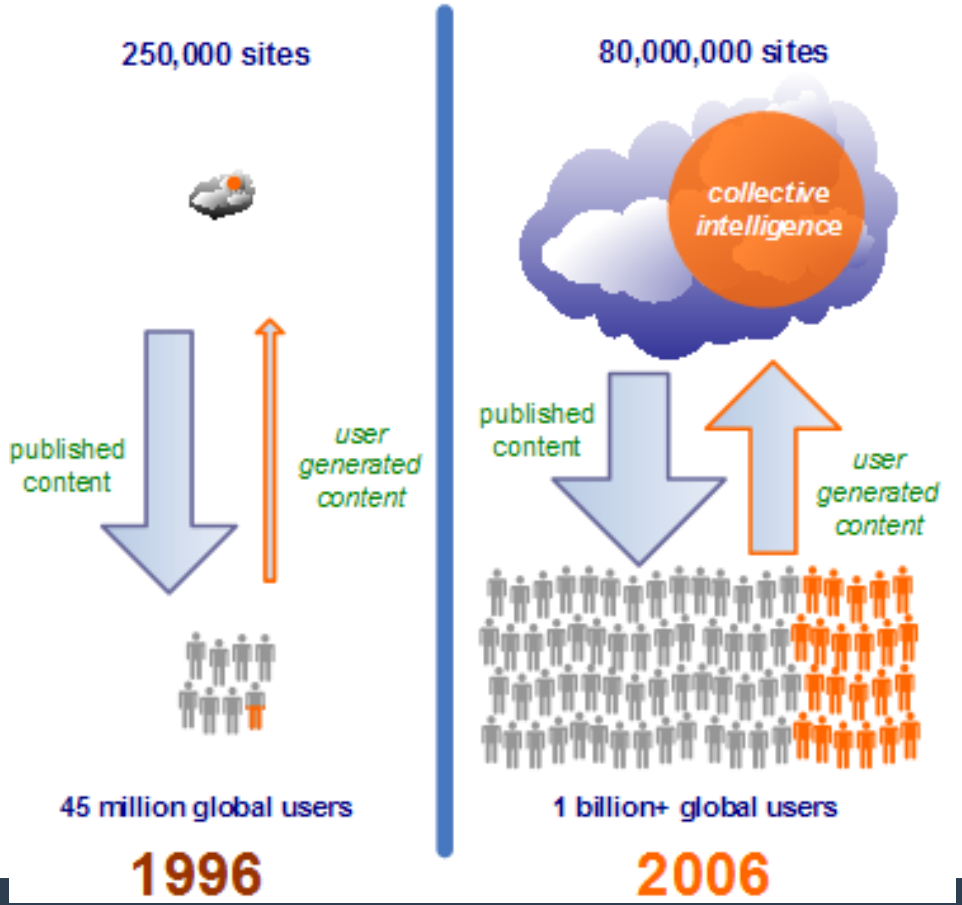
- **Information retrieval when the corpus is the Web**
- **Information Retrieval (IR)**
  - Retrieving unstructured material (usually textual documents) meeting an information need from large collections (usually stored on computers)
- **Live example**

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

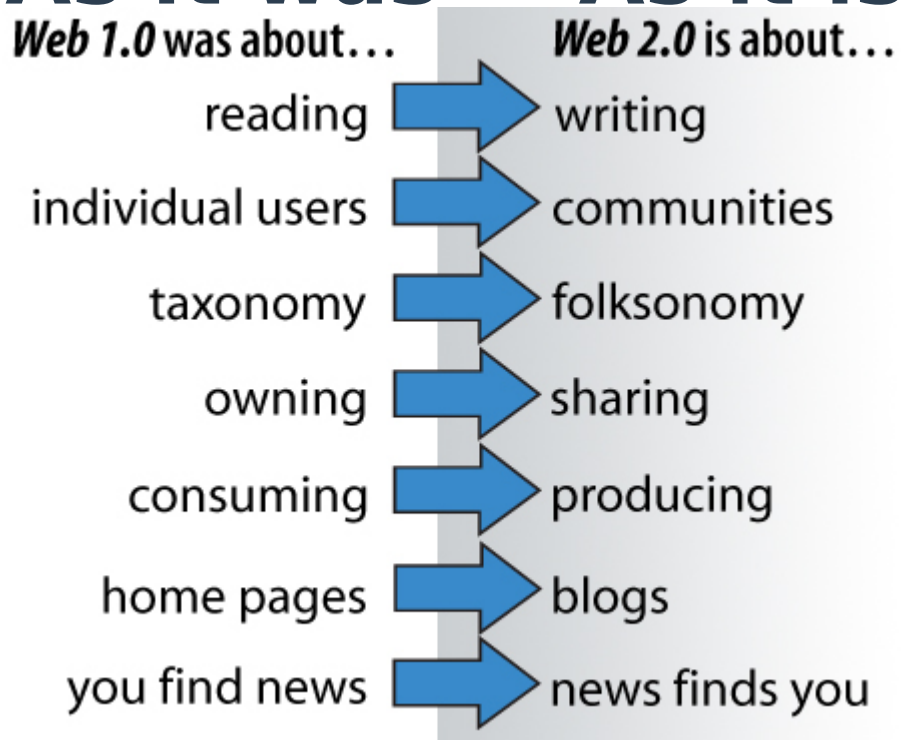
The Web

- As it was
- As it is



The Web

- As it was
- As it is



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**

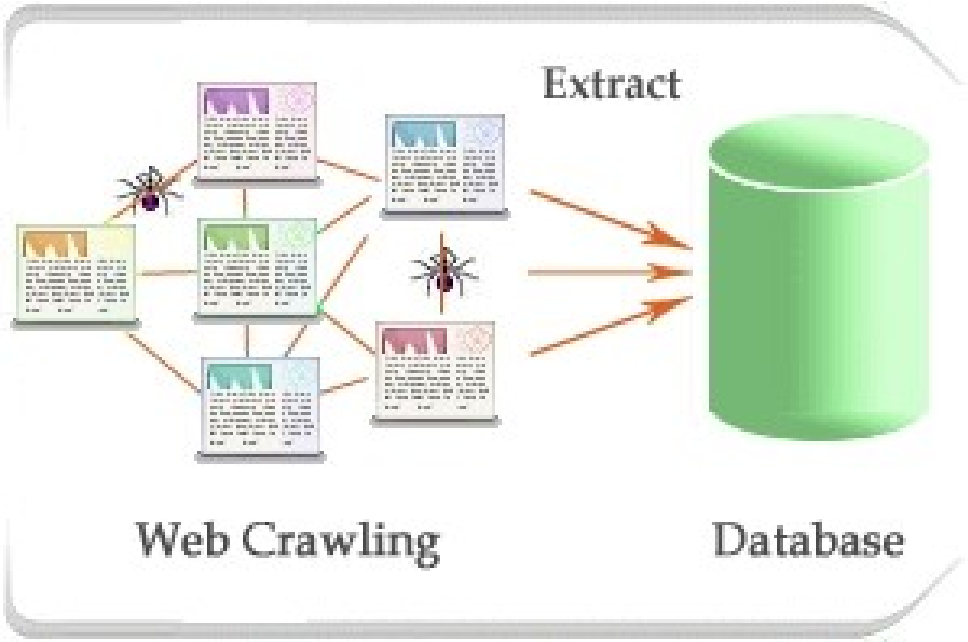
Course outline

- 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**

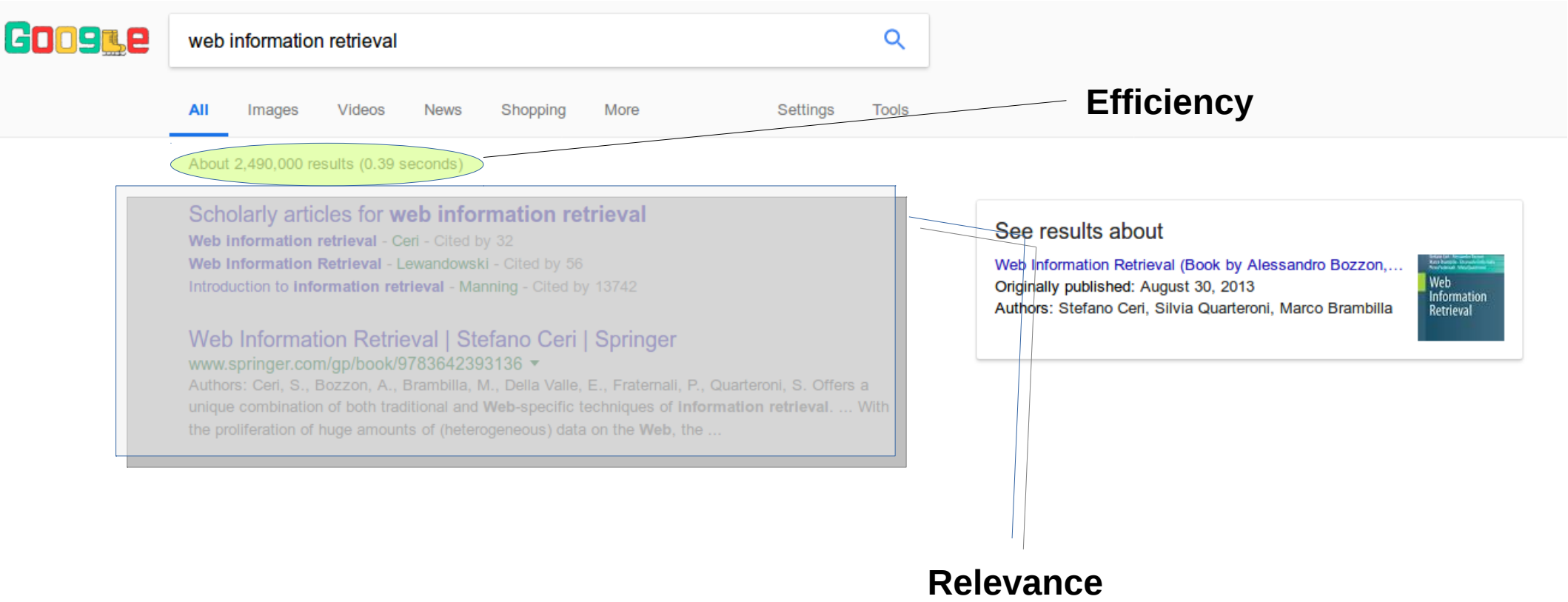


Crawling/Caveats and traps

- **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.



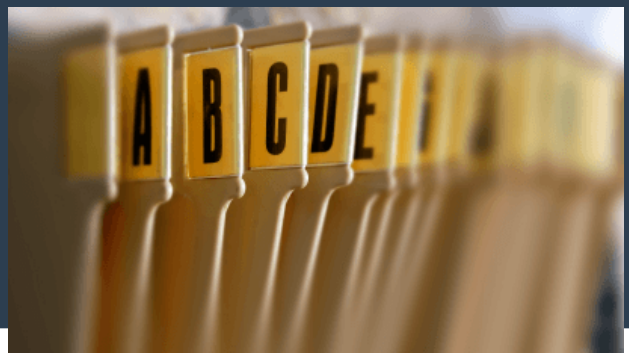
How Google puts it …

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

How Google puts it …

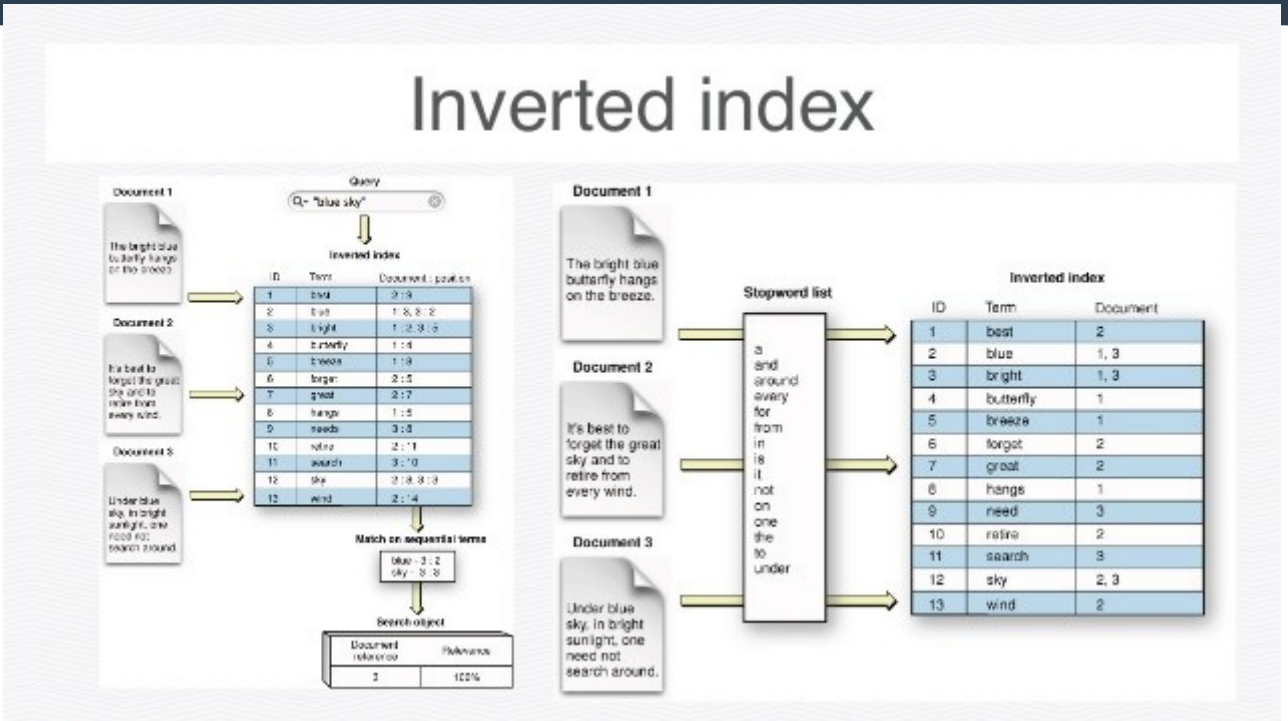
- **https://www.google.com/search/howsearchworks/**
- **In a nutshell**
  - Crawling
  - Indexing
  - Search algorithms

Indexing



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

Efficient data structures



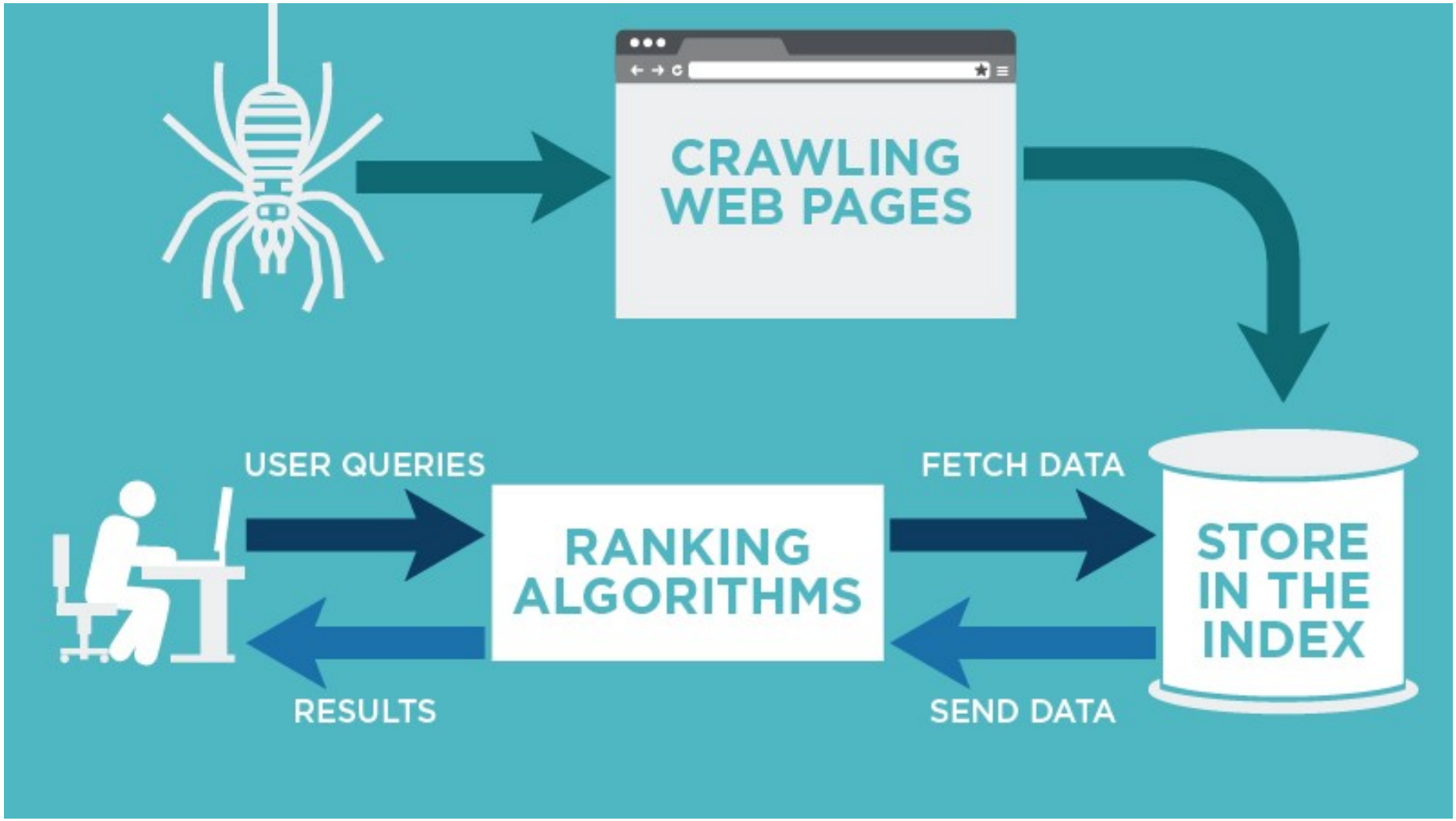
- **Typically an inverted index**
  - Index construction
  - Search using an inverted index
  - Compression, metadata enrichment ...

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

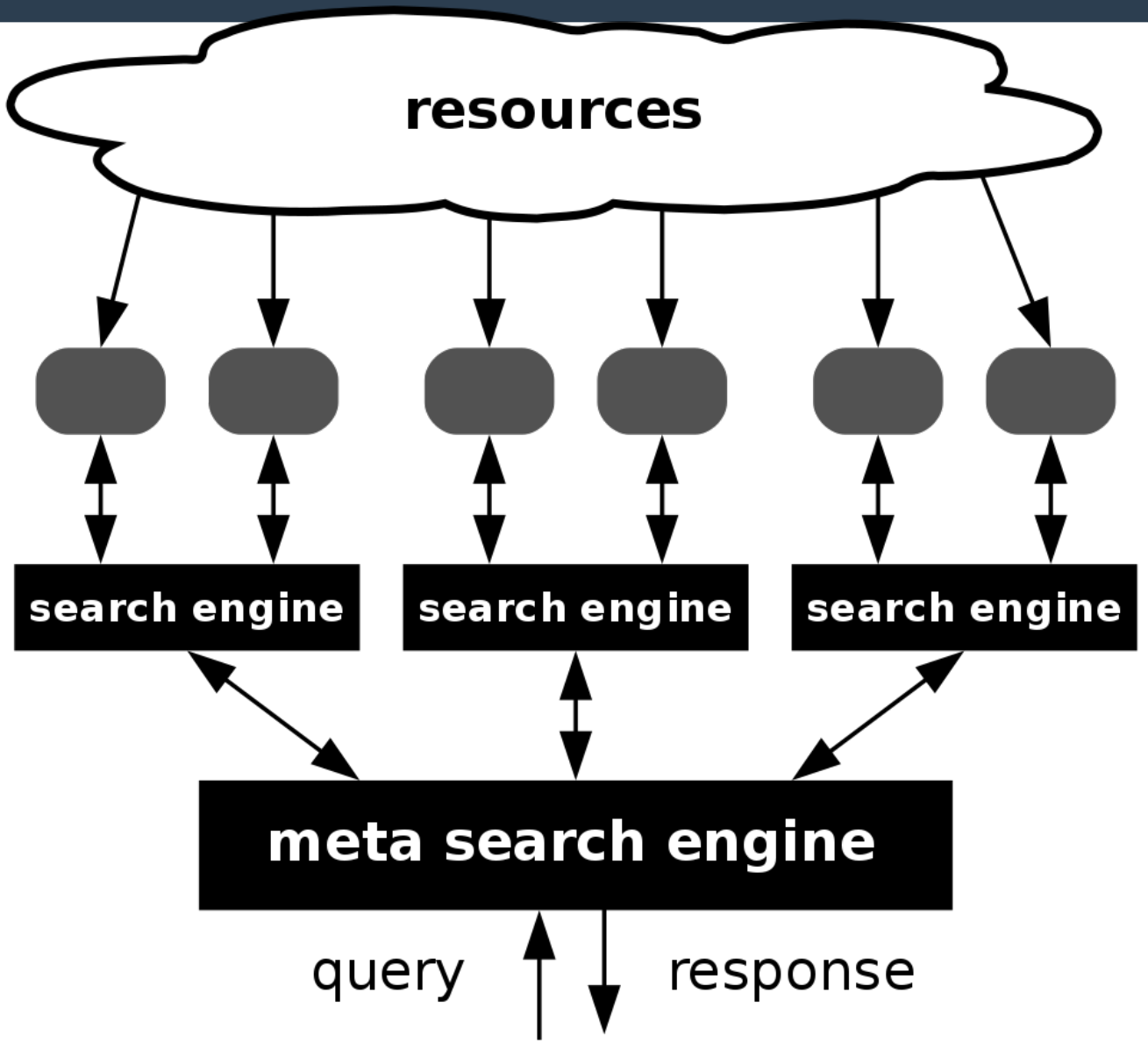
Providing services over the Web

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

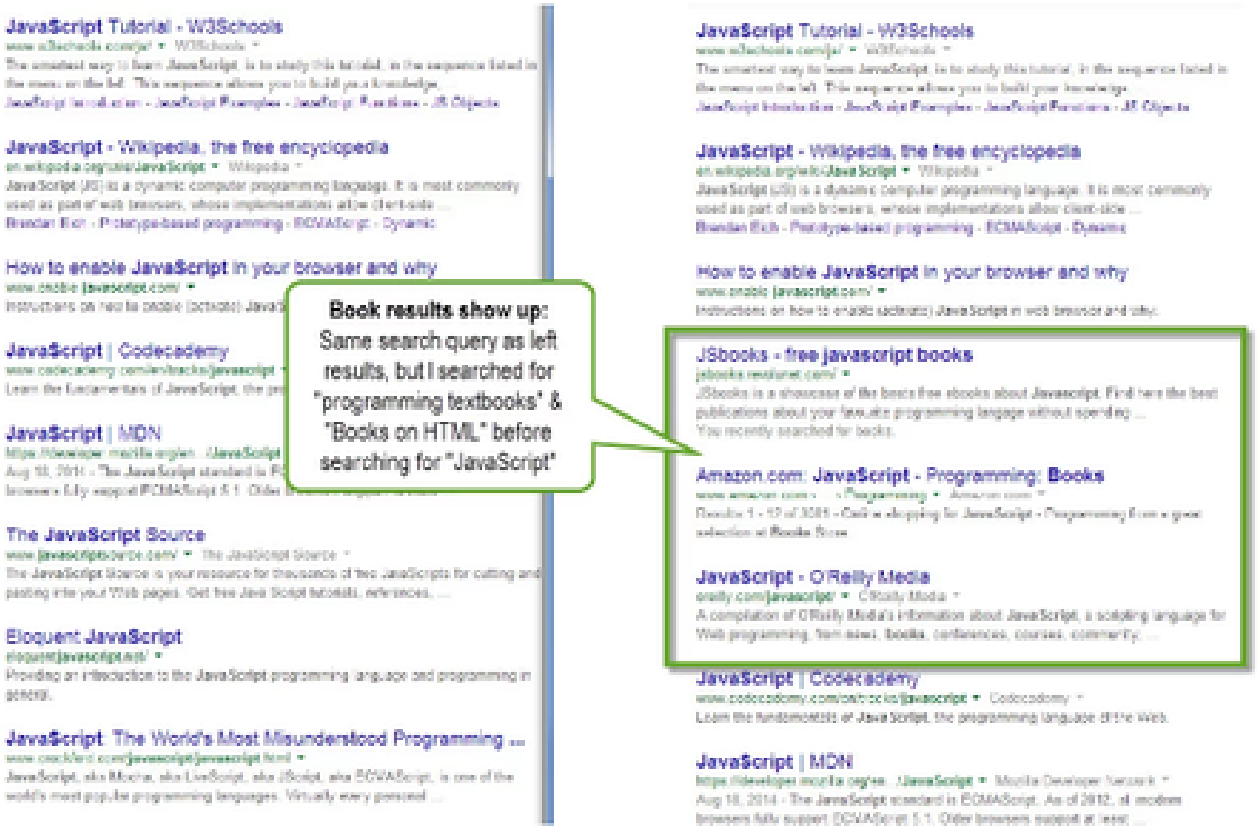
Now (social networking only)

New approaches/challenges

Meta-search


resources
search engine
search engine
search engine
meta search engine
query
response

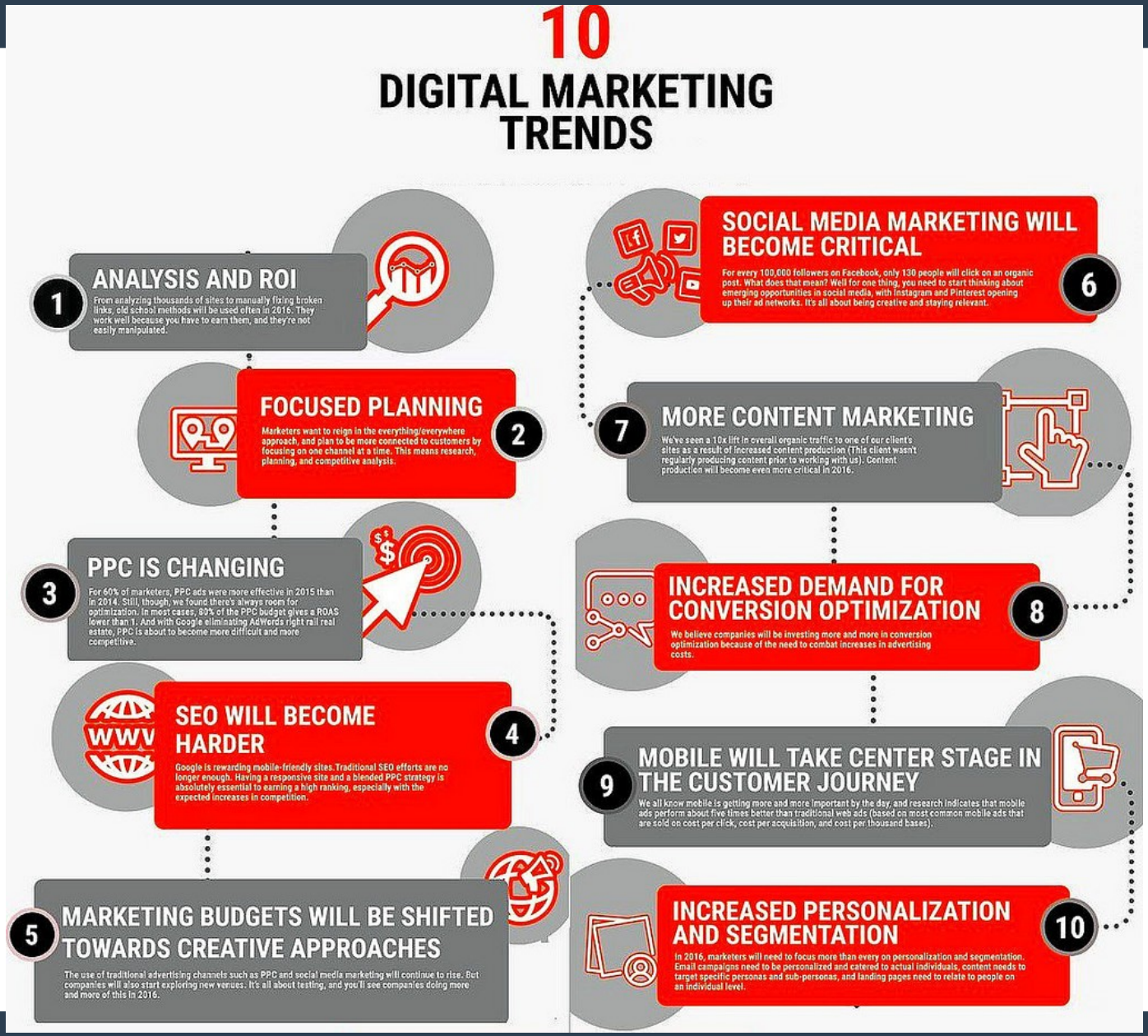
Personalization



Your profile impacts the outcome

Recommendations

...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.