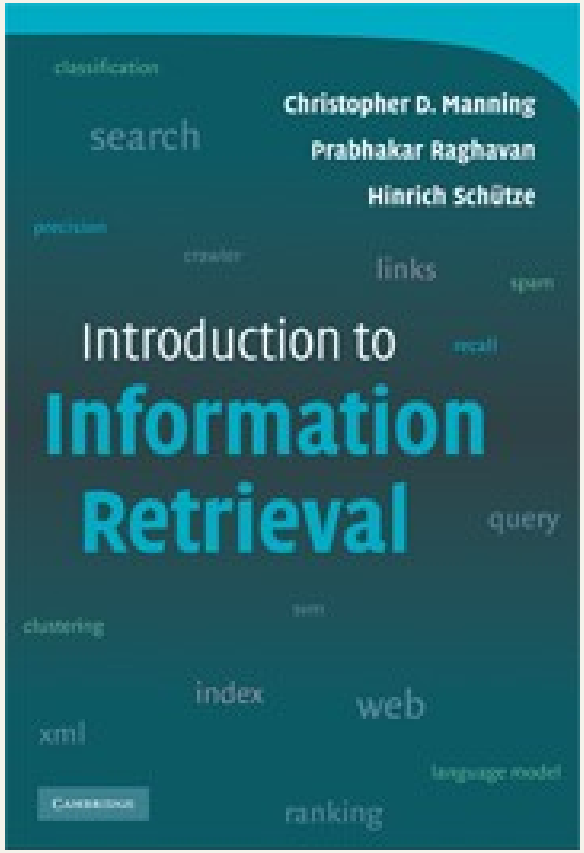
Nowadays IR is much more than building search engines !

IR



Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

Reading Chapter 1

Registrazione Entra

DidaWiki

Cerca

Ultime modifiche Gestore Media Indice

Ti trovi qui: DidaWiki » Corso di Laurea Magistrale in Informatica » Information Retrieval » **Information Retrieval - Academic Year 2019/2020**

magistraleinformatica:ir:ir19:start

Information Retrieval - Academic Year 2019/2020

General Information

- **Teacher** : Paolo Ferragina
- **Course ID**: 289AA
- **CFU:** 6 (first semester)
- **Language:** English
- **Question time:** Monday 15-17, or appointment
- **Official Lecture's Log:** Here it is the registro.
- News about this course will be distributed via a Telegram channel

Indice

- Information Retrieval - Academic Year 2019/2020
- General Information
- Goals
- Schedule of the Lectures
- Exams
- Materials for study
- Lectures
  - Content of the Lectures (LAST YEAR -- just for BACK UP)

Goals

Study, design and analysis of IR systems which are efficient and effective to process, mine, search, cluster and classify documents, coming from textual as well as any unstructured domain. In the lectures, we will:

The exam & 2 midterms: One written test with theory questions + exercises (two rounds, with small penalty)

What is IR today?

Paolo Ferragina

ALTAVISTA
Technology

View Multimedia From Our Vantage Point

Click here for advertising information - reach millions every month!

Search the Web and Display the Results in Standard Form

Submit

Search with Digital's Alta Vista [Advanced Search] [Add URL]

Make Me Laugh...

Create a Site...

Download free demo versions of AltaVista Technology software

ALTAVISTA

[Creative][Search][Humor][Email]

Only text in pages

Search the web using Google!

10 results | Google Search | I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

About Google!

Stanford Search Linux Search

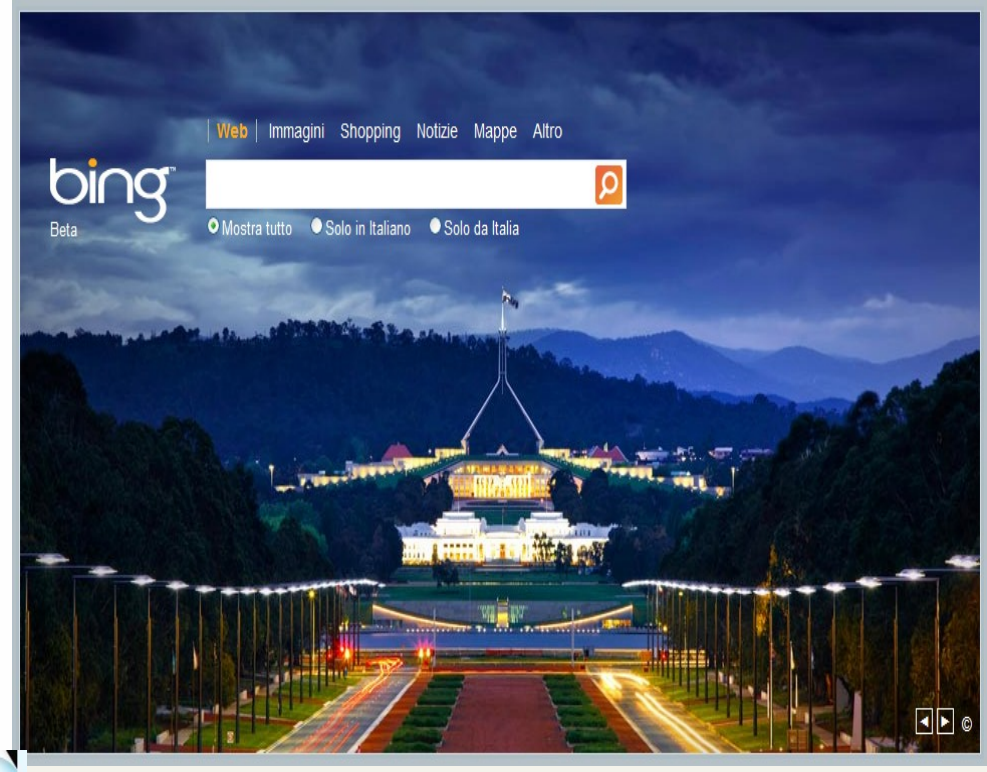
Get Google! updates monthly!

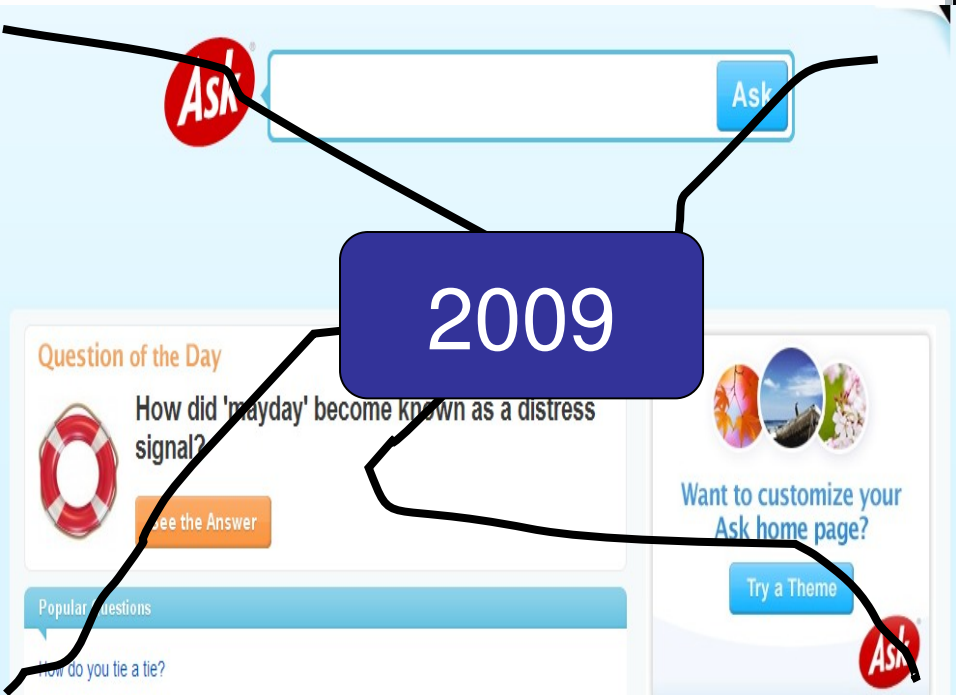
your e-mail | Subscribe

Copyright ©1997-8 Stanford Universit

- Anchor texts
- Hyperlinks


Google
Italia
Web Immagini News Maps Novità! Gruppi Desktop altro »
Ricerca avanzata
Preferenze
Strumenti per le lingue
Cerca con Google
Mi sento fortunato
Cerca: il Web pagine in Italiano pagine provenienti da: Italia
Novità! Usa Gmail e le altre applicazioni di Google Apps per la tua organizzazione.
Pubblicità - Soluzioni Aziendali - Tutto su Google - Google.com in English
Scegli Google come pagina iniziale!
©2007 Google


Web Immagini Shopping Notizie Mappe Altro
bing
Beta
Mostra tutto
Solo in Italiano
Solo da Italia

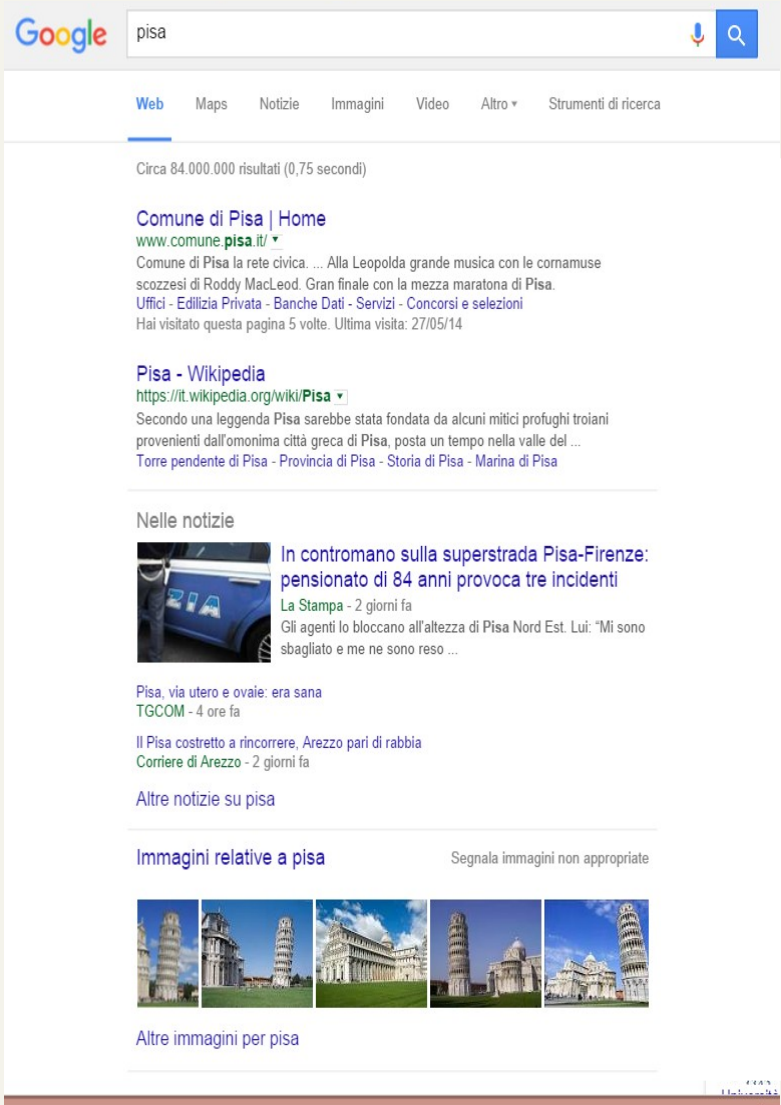

Ask
2009
Question of the Day
How did 'mayday' become known as a distress signal?
See the Answer
Want to customize your Ask home page?
Try a Theme
Popular Questions
How do you tie a tie?


YAHOO!
Search:
Web Search
My Yahoo!
My Mail
Answers
Autos
Finance
Games
Groups
HotJobs
Maps
Mobile Web
Movies
Music
Personals
Real Estate
Shopping
Sports
Tech
Travel
TV
Yellow Pages
Featured
Sports
Video
Hi, Paolo
Sign Out
Mail
Messenger
Radio
Local
Horoscopes
2009-12
In the News
World
Marketplace
All Stars: Most Viewed Cars

Third generation (≈ 2005)

- More sources:
  - News
  - Immages
  - Maps
  - Wikipedia
  - …

Answer
User needs



Google galileo

Web Immagini Maps Video Notizie Altro Strumenti di ricerca

Circa 56.200.000 risultati (0,49 secondi)

Home | Galileo - Giornale di Scienza
www.**galileo**net.it/
News, magazine, recensioni e dossier monografici sui temi della scienza. Il primo magazine Online italiano su scienza e problemi globali. On line dal 1996.

Galileo Galilei - Wikipedia
it.wikipedia.org/wiki/**Galileo_Galilei**
Galileo Galilei (Pisa, 15 febbraio 1564 – Arcetri, 8 gennaio 1642) è stato un fisico, filosofo, astronomo e matematico italiano, considerato il padre della scienza ...
Processo a Galileo Galilei - Categoria:Galileo Galilei - Casa di Galileo Galilei

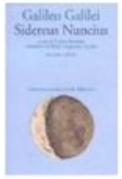
Galileo Galilei - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Galileo_Galilei** Traduci questa pagina
Galileo Galilei often known mononymously as **Galileo**, was an Italian physicist, mathematician, engineer, astronomer, and philosopher who played a major role ...

Aeroporto Galileo Galilei - Sito ufficiale - Aeroporto di Pisa ...
www.pisa-airport.com/
Aeroporto Internazionale **Galileo Galilei**. Include le informazioni, gli orari dei voli, le infrastrutture.
Hai visitato questa pagina molte volte. Ultima visita: 05/09/14

Liceo Classico Galilei Pisa
www.lc**galilei**.pisa.it/
Liceo Classico **Galileo Galilei** - Pisa. Il futuro ha un cuore antico. Home; Registro

Galileo Galilei
Fisico

Libri



Sidereus Nuncius
1610


Dialogo sopra i due massimi s...
1632


Discorsi e dimostraz... matemati...
1638


Il Saggiatore
1623

Lettera a Madama Cristina d...
1636

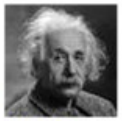
Ricerche correlate
Visualizza altri 15 elementi


Niccolò Copernico


Isaac Newton


Giovanni Keplero

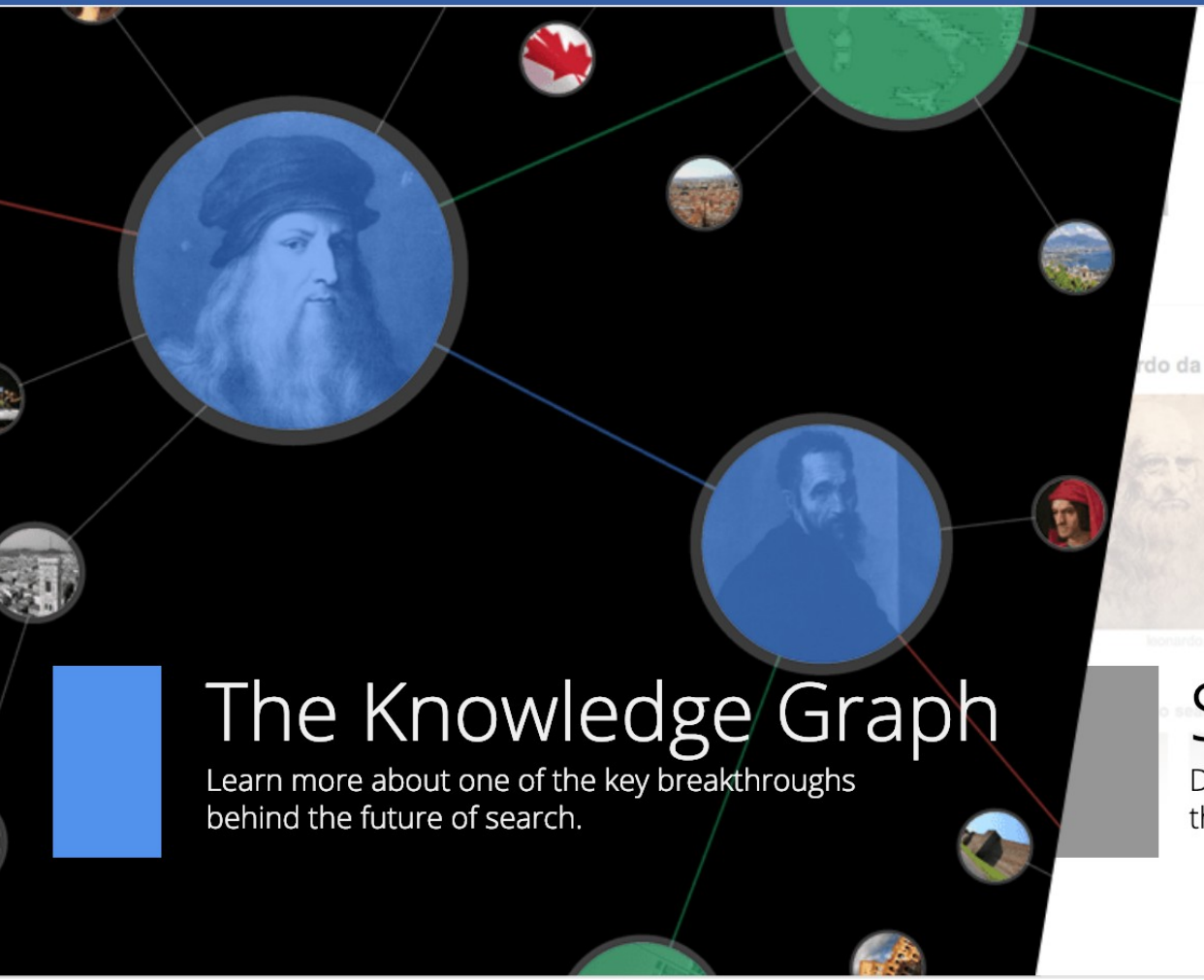

Aristotele


Albert Einstein

Feedback

Google | Inside Search

Home | How Search Works | Tips & Tricks | **Features** | Search Stories | Playground | Blog | Help



May 2012

The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

See it in action

Discover answers to questions you never thought to ask, and explore collections and lists.

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Wikimedia Commons
Wikiquote
Wikisource

Languages
Адыгэбзэ

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).
This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,[1][2][3] he epitomised the Renaissance humanist ideal.

Many historians and scholars regard him as the archetype of the "Renaissance Man", a "Universal Genius" or "Renaissance M... ...al ...nly ...th of his ...ed in Rome, Bologna and Venice, and he spent his last ... ...d to him by Francis I of France.

Leonardo was, and is, renowned primarily as a painter. Among his works, the *Mona* ... most famous and most parodied portrait[6] and *The Last Supper* the most reproduce... ... painting of all time, their fame approached only by Michelangelo's *The Creation of A...* Leonardo's drawing of the *Vitruvian Man* is also regarded as a cultural icon,[7] being

Leonardo da Vinci

The Vitruvian Man

Science

From Wikipedia, the free encyclopedia

This article is about the general term. For other uses, see Science (disambiguation).

Science[nb 1][2]:58[3] is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.[nb 2]

Contemporary science is typically subdivided into the natural sciences, which study the material universe; the social science... which study people and societies; and the formal sciences, such as mathematics. The formal sciences are often excluded ... they do not depend on empirical observations.[4] Disciplines which use science like engineering and medicine may also be considered to be applied sciences.[5]

During the Middle Ages in the Middle East, foundations for the scientific method were laid by Ibn al-Haytham in his *Book of Optics*.[6][7][8][9][10] From classical antiquity through the 19th century, science as a type of knowledge was more closely linke... philosophy than it is now and, in fact, in the Western world, the term "natural philosophy" encompassed fields of study that ... today associated with science, such as astronomy, medicine, and physics.[11][nb 3] While the classification of the material wo... the ancient Indians and Greeks into air, earth, fire and water was more philosophical, medieval Middle Eastern scientists us... practical, experimental observation to classify materials.[12]

In the 17th and 18th centuries, scientists increasingly sought to formulate knowledge in terms of *laws of nature*. Over the c... of the 19th century, the word "science" became increasingly associated with the scientific method itself, as a disciplined way to study the natural world. It was in the 19th century that scientific disciplines such as biology, chemistry, and physics reached their modern shapes. The same time period also included the origin of the terms "scientist" and "scientific community," the founding of scientific institutions, and increasing significance of the interactions with society and other aspects of culture.[13][14]

Invention

From Wikipedia, the free encyclopedia

"Inventor" and "Invented" redirect here. For other uses, see Invention (disambiguation).
For more details on inventions throughout history, see Timeline of historic inventions.
For the CAD design software, see Autodesk Inventor.

An **invention** is a unique or novel device, method, composition or process. The invention proc... engineering and product development process. It may be an improvement upon a machine or ... an object or a result. An invention that achieves a completely unique function or result may be ... novel and not obvious to others skilled in the same field. An inventor may be taking a big step i...

Some inventions can be patented. A patent legally protects the intellectual property rights of the ... claimed invention is actually an invention. The rules and requirements for patenting an inventio... process of obtaining a patent is often expensive.

Another meaning of invention is **cultural invention**, which is an innovative set of useful social ... passed on to others.[1] The Institute for Social Inventions collected many such ideas in magazi... important component of artistic and design creativity. Inventions often extend the boundaries o... capability.

Astronomy

From Wikipedia, the free encyclopedia

This article is about the scientific study of celestial objects. For other uses, see Astronomy

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planets... nebulae) and processes (such as supernovae explosions, gamma ray bursts, and cosmic micr... physics, chemistry, and evolution of such objects and processes, and more generally all phen... atmosphere of Earth. A related but distinct subject, physical cosmology, is concerned with stud...

Astronomy is the oldest of the natural sciences. The early civilizations in recorded history, such ... Egyptians, Nubians, Iranians, Chinese, and Maya performed methodical observations of the n... included disciplines as diverse as astrometry, celestial navigation, observational astronomy an... professional astronomy is nowadays often considered to be synonymous with astrophysics.[2]

During the 20th century, the field of professional astronomy split into observational and theoret... is focused on acquiring data from observations of astronomical objects, which is then analyzed... Theoretical astronomy is oriented toward the development of computer or analytical models to ... phenomena. The two fields complement each other, with theoretical astronomy seeking to exp... observations being used to confirm theoretical results.

Astronomy is one of the few sciences where amateurs can still play an active role, especially in ... transient phenomena. Amateur astronomers have made and contributed to many important as... new comets.

From word to concepts

Leonardo is the scientist that has painted Mona Lisa

Science

Italia

Renaissance

Leonardo da Vinci

Mona Lisa (painting)

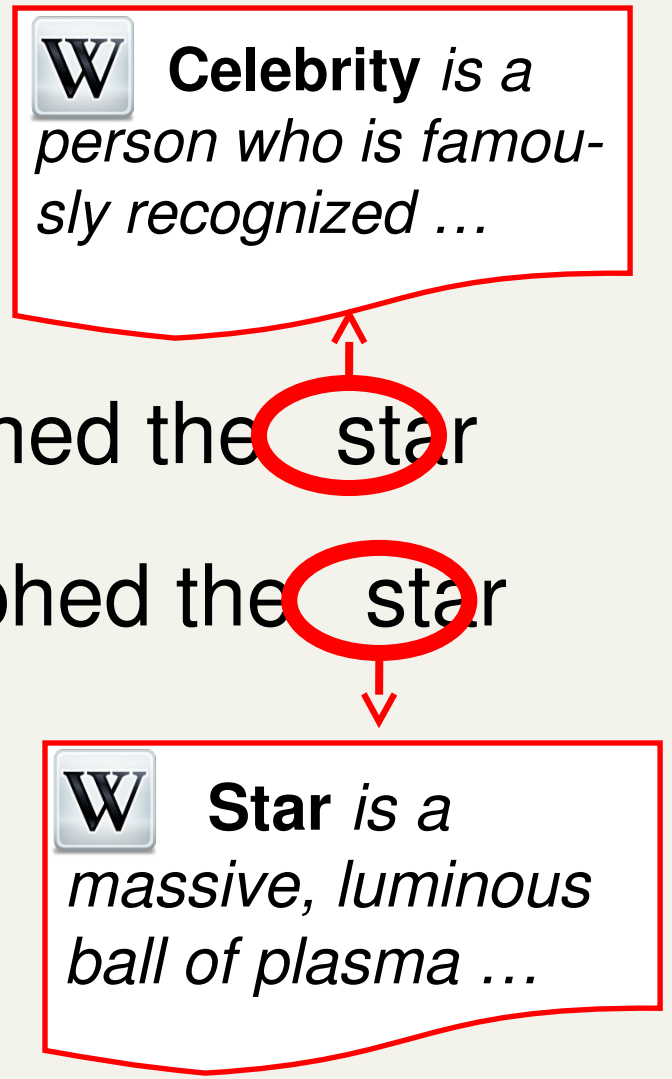
Cartography

Louvre

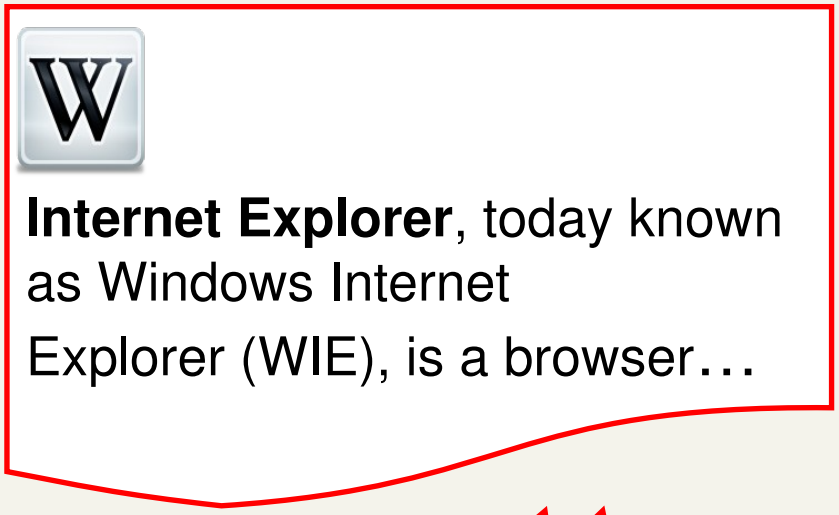
Arts

Florence

Polysemy


Celebrity is a person who is famou-sly recognized …
the paparazzi photographed the star
the astronomer photographed the star
Star is a massive, luminous ball of plasma …

Sinonimy

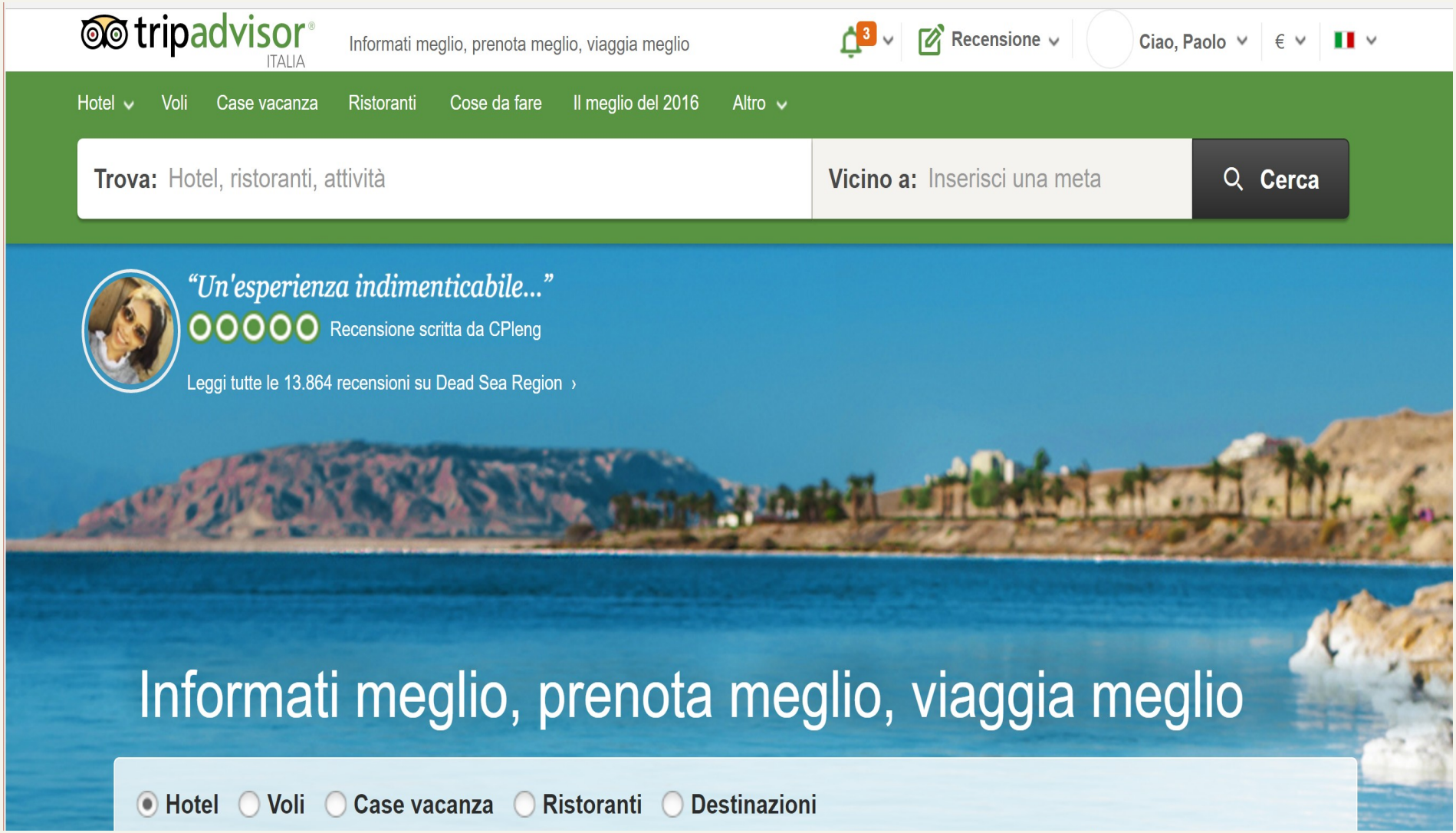


He is using Microsoft's browser

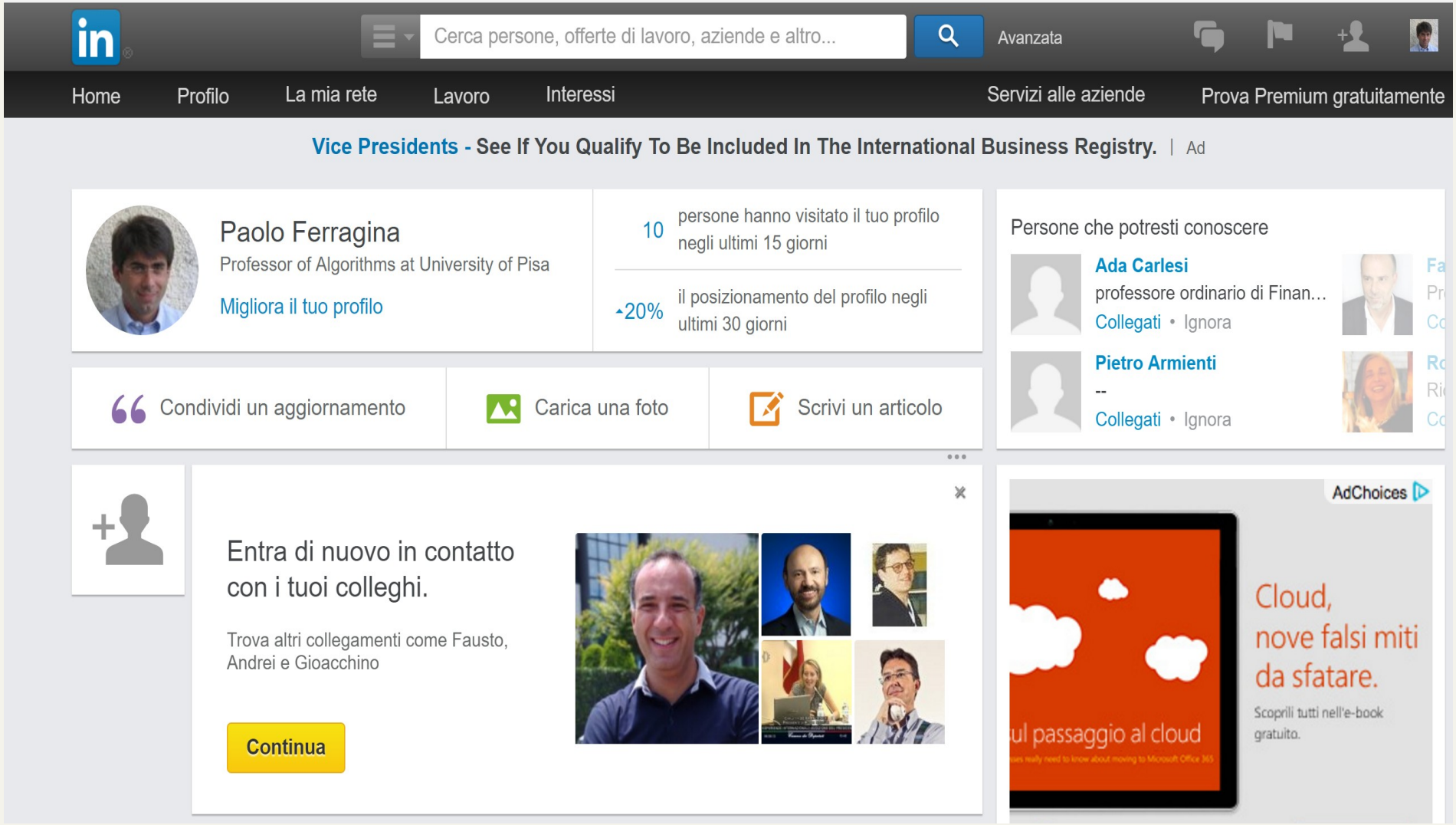
She plays with Internet Explorer

Searching routes

Searching over geo+labels



Searching over labeled graphs



Evolution of Search Engines

- Zero generation -- use metadata added by users

1991-.. Wanderer

- First generation -- use only on-page, web-text data
  - Word frequency and language

1995-1997 AltaVista, Excite, Lycos, etc

- Second generation -- use off-page, web-graph data
  - Link (or connectivity) analysis
  - Anchor-text (How people refer to a page)

1998: Google

- Third generation -- answer "the need behind the query"
  - Focus on "user need", rather than on query
  - Integrate multiple data-sources
  - Click-through data

Google, Yahoo, MSN, ASK,……

Fourth and current generation → Information Supply

Google Search Timeline

Google

II° generation

Did you mean?

Synonyms

Products

News

Registered google.com

Images

Adwords

97 98 99 00 01 02 03

III° generation

Autocomplete

Universal Search

Google Instant

News

Stock Quotes

Local

Sports Scores

Flights

Weather

Video

Patents

Movie Times

Google Search App

Search Tools

Voice Search

04 05 06 07 08 09 10

IV° generation

Knowledge Graph

Voice Search on Desktop

Search by Image

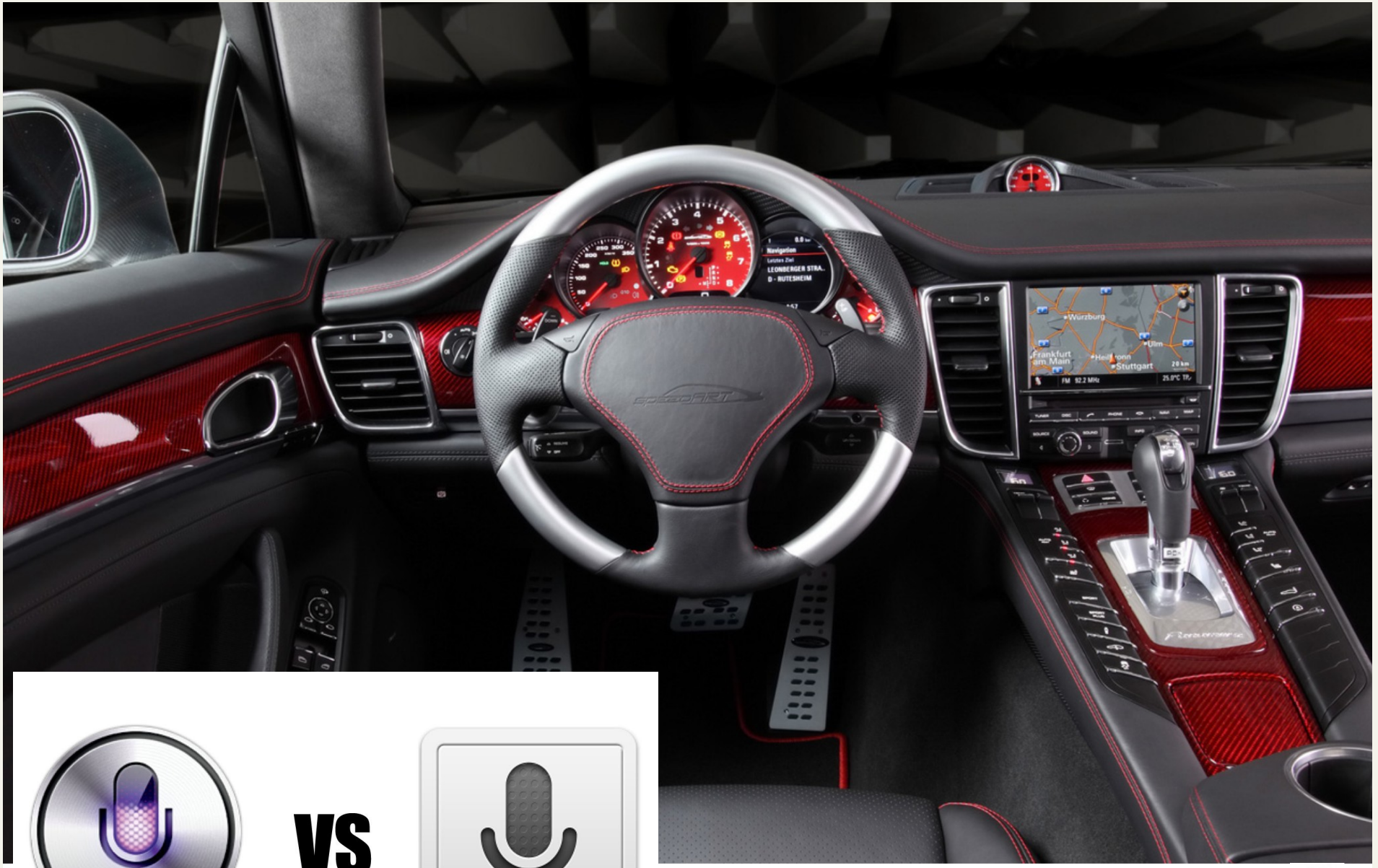
Intelligence Features

Google Now

11 12 13

Paradigm shift

We have now «devices 2.0» that have their ID, Communication capacity, computing and storage, and currently interaction ability.



VS

SIRI

GOOGLE NOW

Three main types of data…

- Opportunistic
  - Credit card transactions
  - Tel calls, bills, web clicks, …
- Purposely sensed
  - pollution, temperature, wind, …
  - movement, accelleration,…
  - Health sensing,…
- User generated
  - Photo, tweet, post, email,…
  - Query-log on search engines

CHANGING THE WAY FOOTBALL WORKS.

The Football platform for the people who live for the beautiful game.
A world of data and video available online.

CLUB | SCOUTS | AGENTS | COACHES | PLAYERS | REFEREES | JOURNALISTS

START NOW

15 DAYS FREE TRIAL | NO CREDIT CARD REQUIRED

Our clients include:

WISER

Wikipedia Expertise Ranking

Search for expertise...

Search by Expertise **Search by Name** **Search by Department**

How it works

WISER is a *fast* and *accurate* **Semantic Search Engine** for *expert finding*, currently working on the faculty of the University of Pisa.

- You can issue an *expert finding* query by inserting some *expertise area* in the *search bar* and pressing *"Search by Expertise"*, this will return the authors from *Faculty of UniPI* that WISER considers experts of the queried research area.
- You can also issue an *expert profiling* query by inserting the name of a researcher/department and then pressing *"Search by Name"* or *"Search by Department"*, this will return the research topics in which the queried researcher/department are considered experts by WISER.

WISER offers a sophisticated GUI that allows to visualize and post-process the results of a query in several ways, eventually useful to dig into the *expertise* of *researchers* and *departments* of the *University of Pisa*.

Basics

Paolo Ferragina

Information Retrieval

Information Retrieval (IR) is finding material (usually documents) of **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers).

IR vs. databases: Unstructured vs Structured data

Structured data tends to refer to “tables”

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,
Salary < 60000 AND Manager = Smith.

Semi-structured data: XML/JSON

- In fact almost no data is “unstructured”
  - E.g., this slide has distinctly identified **zones** such as the *Title* and *Bullets*
- Facilitates “semi-structured” search such as
  - *Title* contains data AND *Bullets* contain search
- Issues:
  - how do you process “about”?
  - how do you rank results?

Unstructured data

Typically refers to **free text**, and allows

- Keyword queries including operators
- More sophisticated “concept” queries e.g.,
  - find all web pages **dealing with** *drug abuse*

Classic model for searching text documents

Boolean queries: Exact match

- The **Boolean retrieval model** is being able to ask a query that is a Boolean expression:
  - Boolean Queries are queries using *AND, OR* and *NOT* to join query terms
    - Views each document as a set of words
    - Is precise: document matches condition or not.
  - Perhaps the **simplest model** to build an IR system on
- Many search systems still use it:
  - Email, library catalog, Mac OS X Spotlight

Implementing the Boolean model

	Antony an	Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony		1	0	0	0	1
Brutus		1	0	1	0	0
Cae		1	0	1	1	1
		1	0	0	0	0
	1	0	0	0	0	0
	1	0	1	1	1	1
	1	0	1	1	1	0

Matrix could be very big

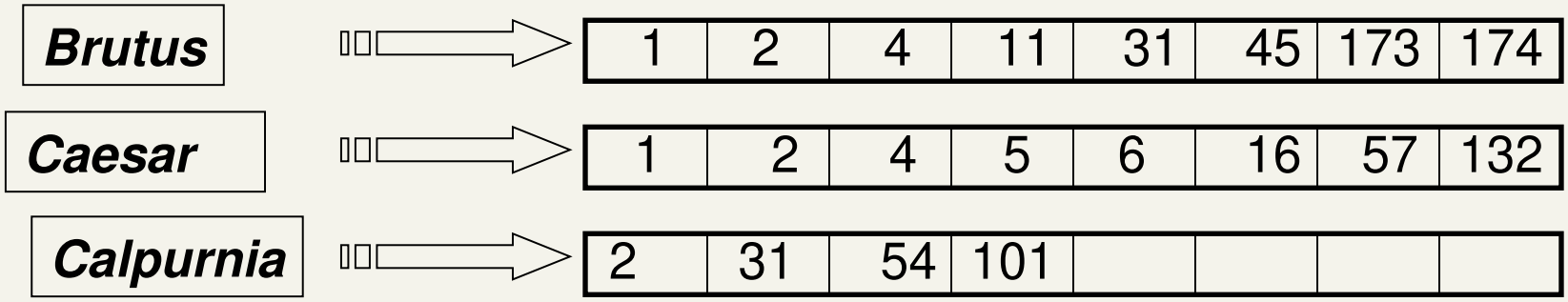
1 if play contains word,
0 otherwise

Brutus *AND* ***Caesar***
BUT NOT ***Calpurnia***

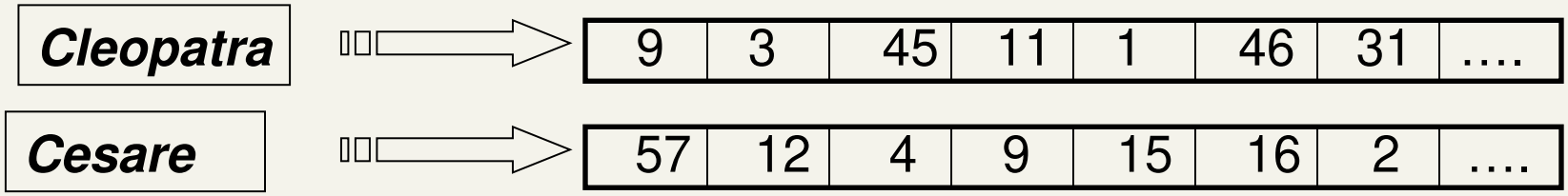
Inverted index

- For each term *t*, we must store a list of all documents that contain *t*.
  - Identify each by **docID**, a document serial number

- **What about inserting a new docID ?**



AND query



If n,m are the lengths of the lists, how many comparisons ?

$n * m$

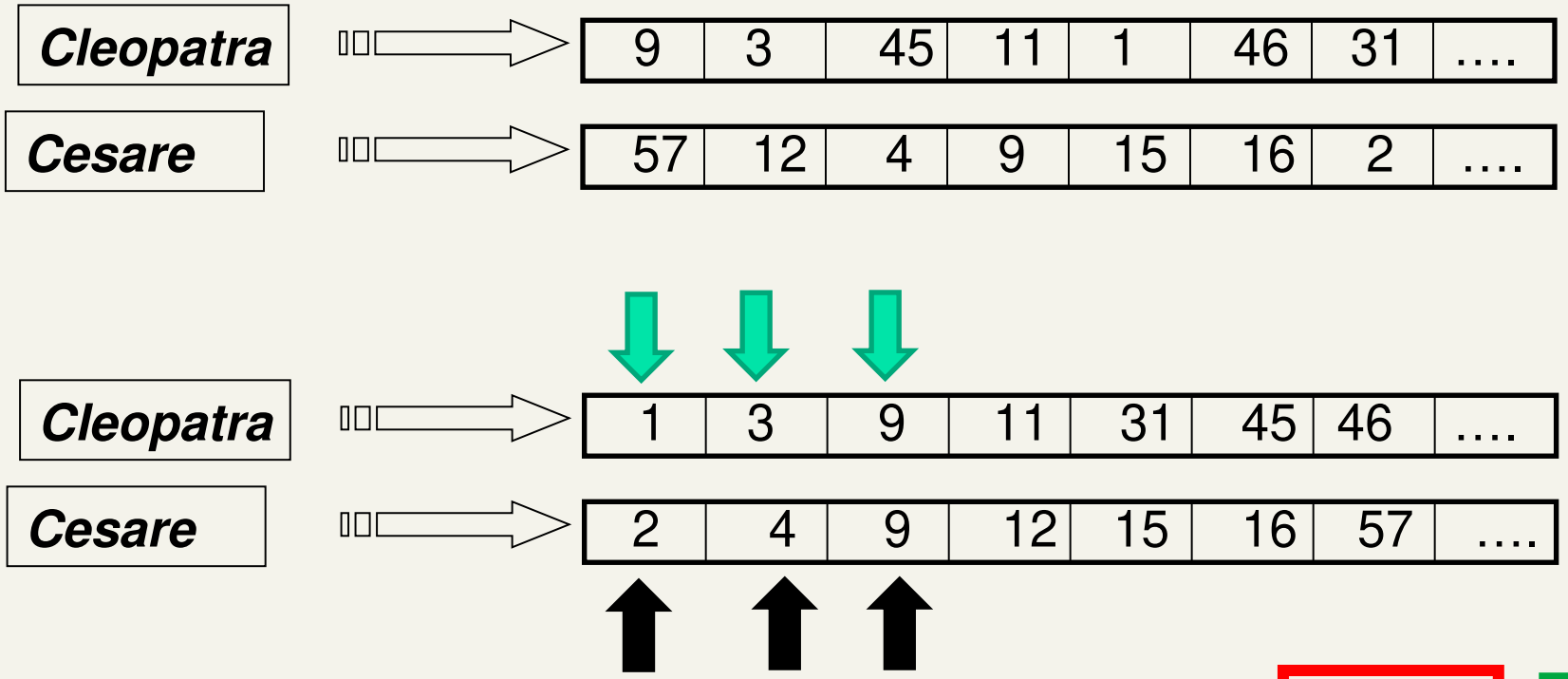
This is not an «engineering problem»,

You need efficient algorithms!

$\approx 10^{12}$ cmp

$\approx 10^{3}$ sec

AND query



How many comparisons ?

How much space ?

Which are the top-10 results ?

n + m

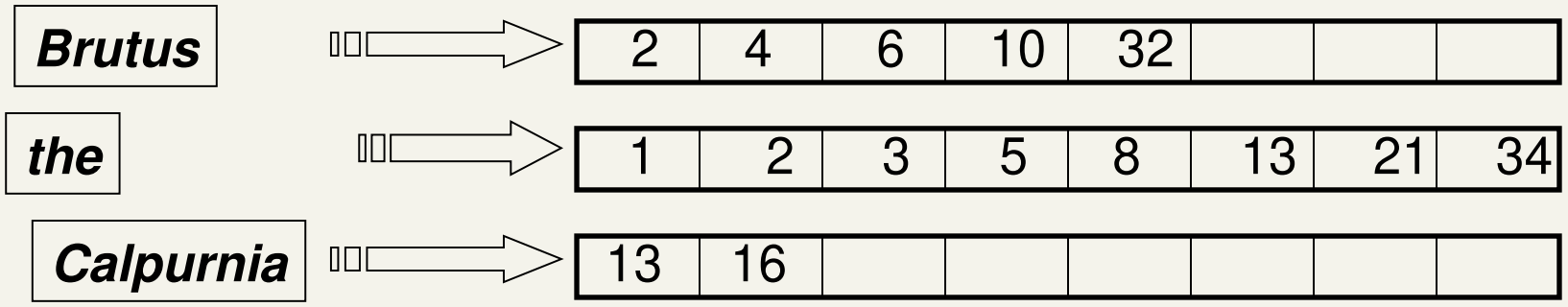
$\approx 10^6$

≈1 msec

Intersecting two postings lists

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

The Inverted index



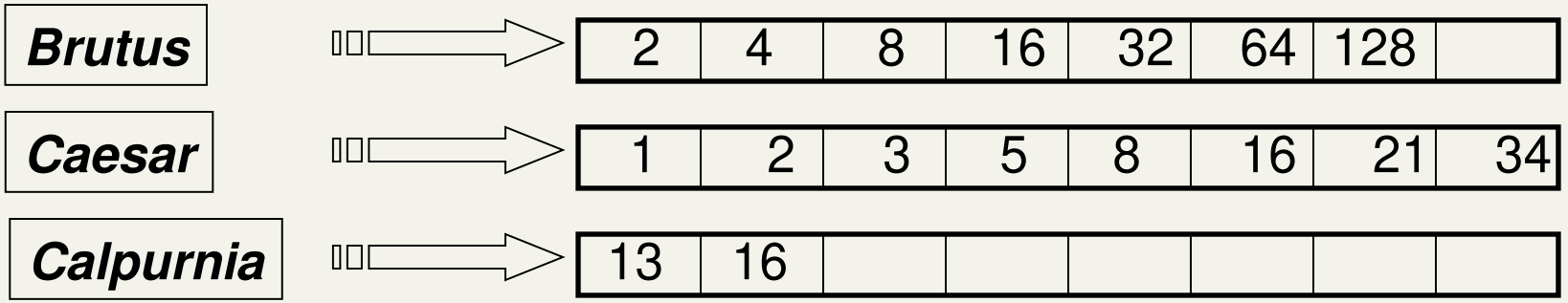
Two advantages:

- **Speed**: query requires just a scan
- **Space**: store smaller integers (gap coding)

Compressed, they occupy 1÷3% original text

Query optimization

- What is the best order for query processing?
  - Consider a query that is an *AND* of *n* terms.
  - For each of the *n* terms, get its postings, then *AND* them together.

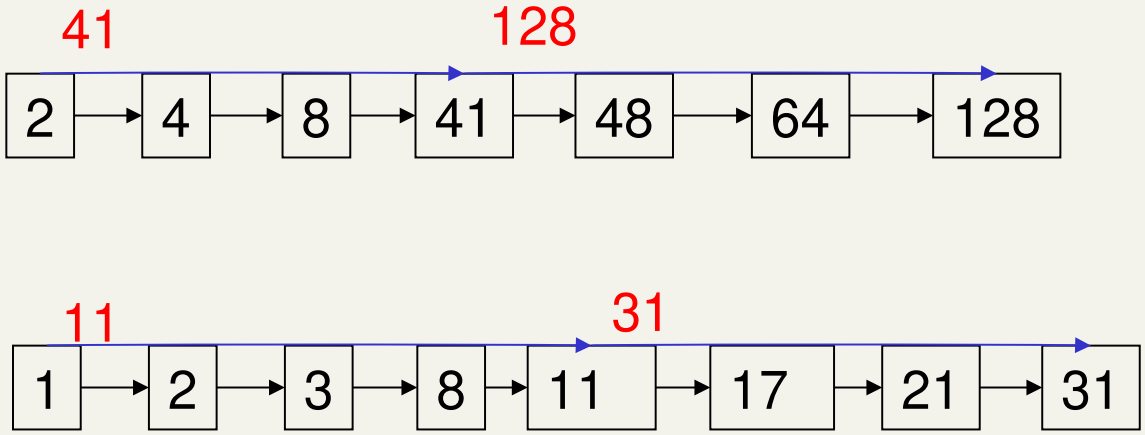


Query: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

Query optimization

- Can we improve scanning-based intersection?
  - Skips (yet scan-based but with shortcuts)

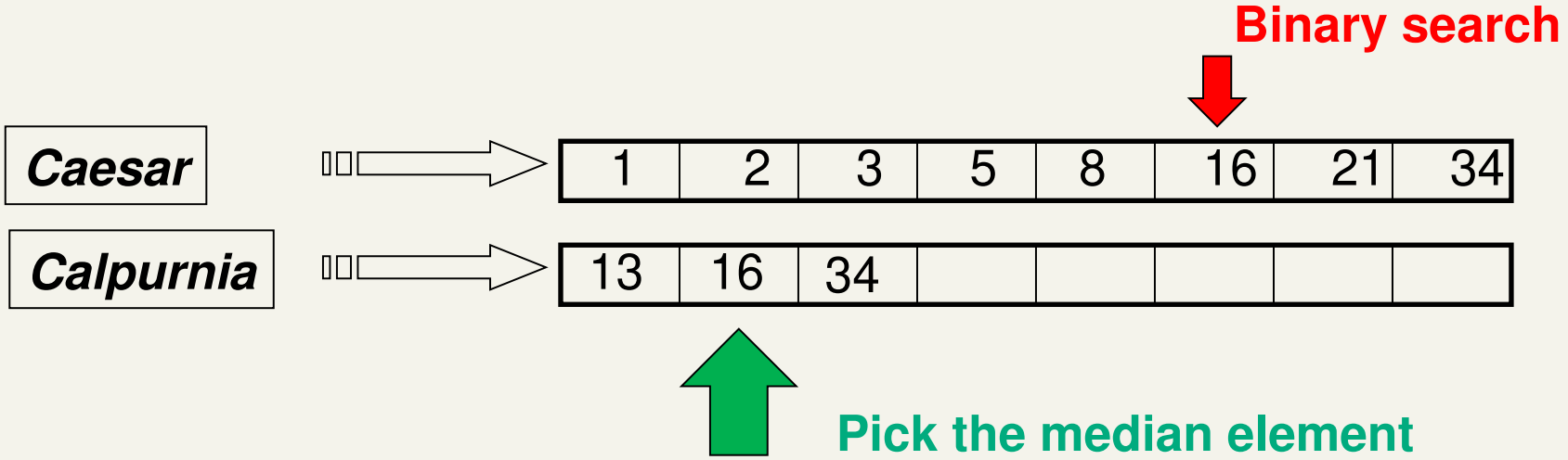
Augment postings with skip pointers (at indexing time)



- Where do we place them ?
- Which is the space/time trade-off ?

Query optimization

- Can we improve scanning-based intersection?
  - Skips (yet scan-based but with shortcuts)
  - Recursive merge (splitting by pivots)



If median is always out: O(log m * log n)

Worst is median always in the middle: T(n,m) = O(log n) + 2 T(n/2,m/2)

Which list you bisect at every recursive step ?

Boolean queries: More general merges

- Exercise: Adapt the merge for :

  Brutus *AND NOT* ***Caesar***

  Brutus *OR NOT* ***Caesar***

Can we still run the merge in time O(*n* + *m*)?

IR is much more…

- What about phrases?
  - ***"Stanford University"***
- Proximity: Find ***Gates*** *NEAR* ***Microsoft***.
  - Need index to capture term positions in docs.
- Zones in documents: Find documents with (*author =* ***Ullman***) *AND* (text contains ***automata***).
- Search for ***Maradona*** and find also "***el pibe de oro***"

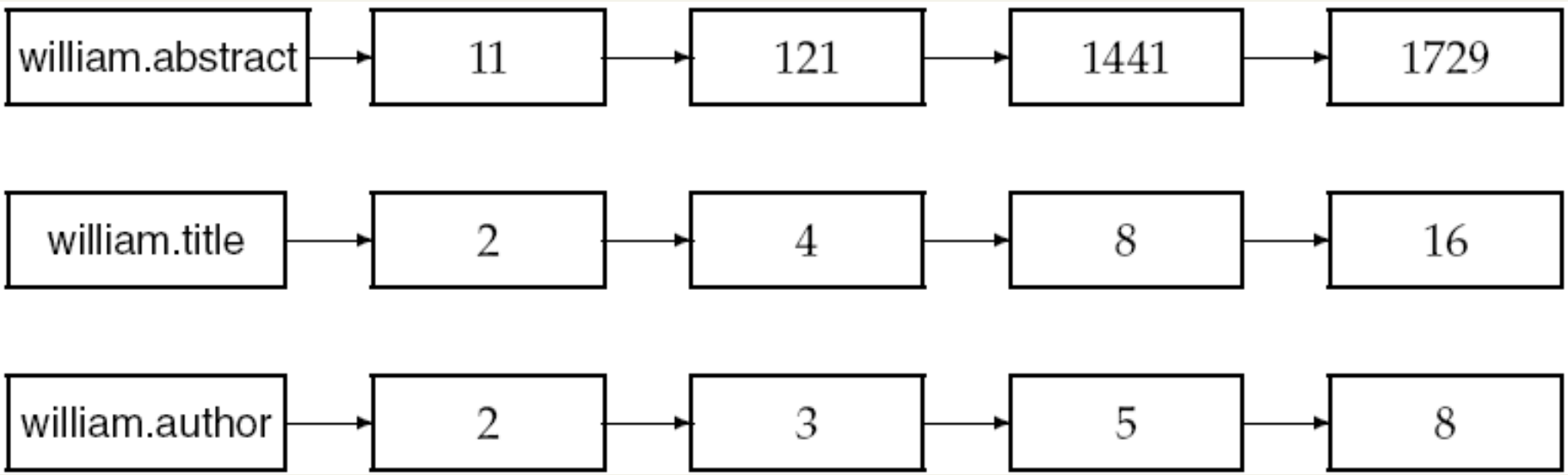
Zone indexes

- A zone is a region of the doc that can contain an arbitrary amount of text e.g.,
  - Title
  - Abstract
  - References …

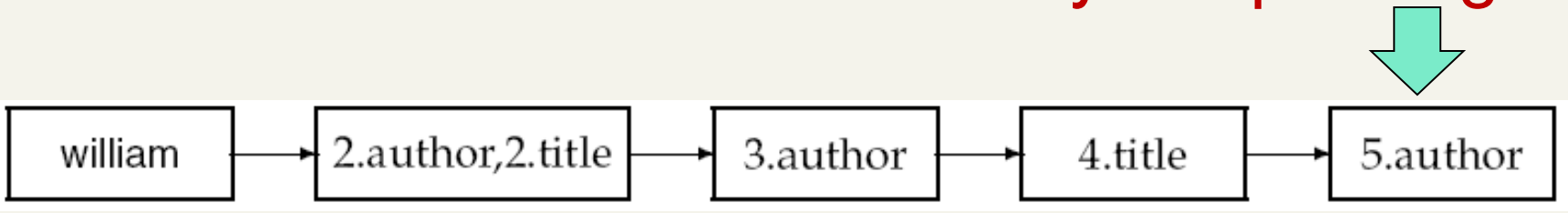
- Build inverted indexes on **fields AND zones** to permit querying

- E.g., "find docs with *merchant* in the title zone and matching the query *gentle rain*"

Example zone indexes



Encode zones in dictionary vs. postings.



Ranking search results

- Boolean queries give inclusion or exclusion of docs.

- But
  - often results are too many and we need to rank results
  - Classification, clustering, summarization, text mining, etc…

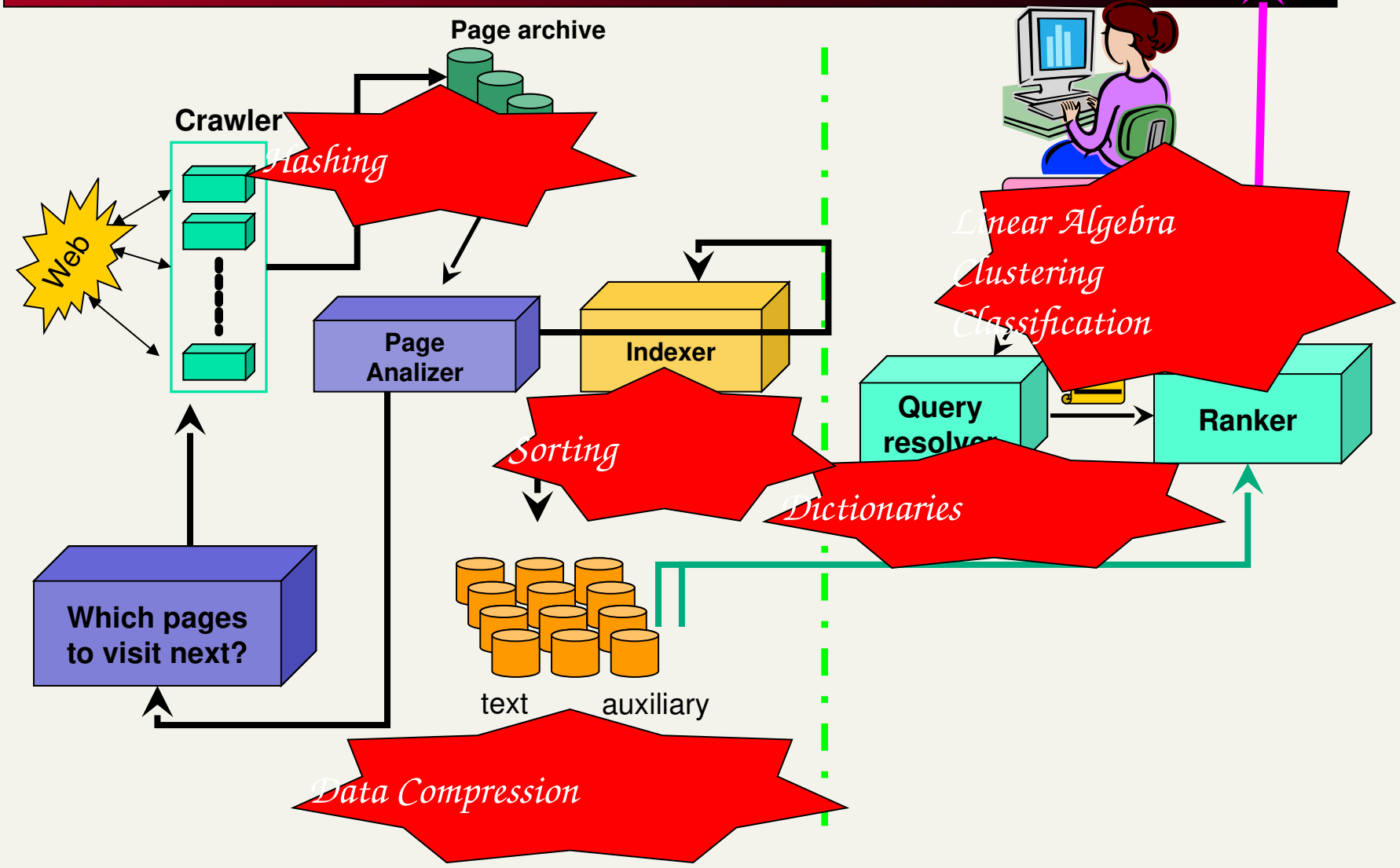
  A lot of AI and Machine Learning on several kinds of features extracted from pages content and the Web for results selection and ranking

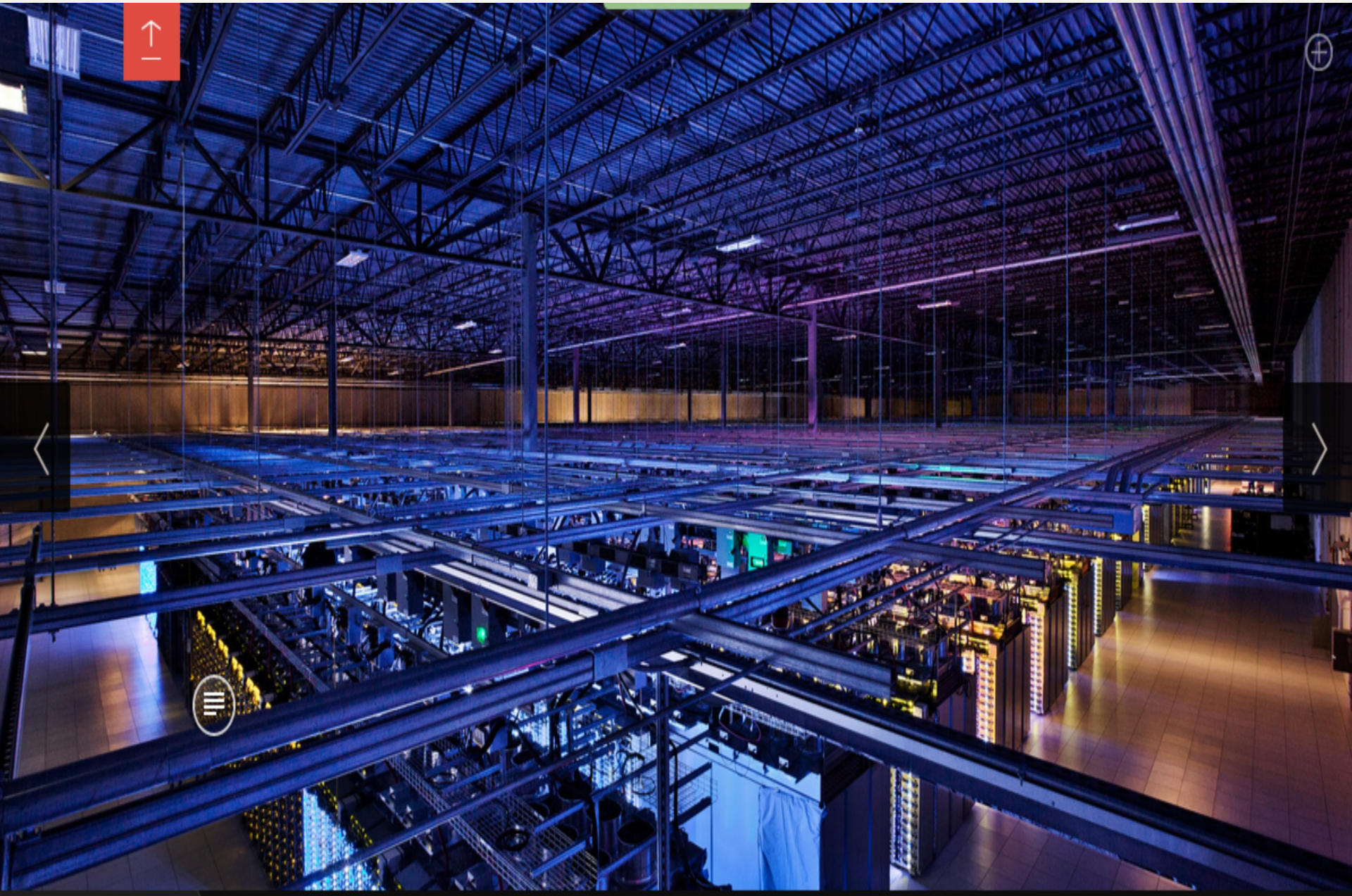
Web IR and its challenges

- Unusual and diverse
  - Documents
  - Users
  - Queries
  - Information needs

- Exploit ideas from social networks
  - link analysis, click-streams, knowledge graphs,...

Our topics, on an example




Council Bluffs, Iowa
Il nostro data center di Council Bluffs si estende per 10.700 metri quadrati. Sfruttiamo in modo ottimale ogni
Condividi
+1 0m

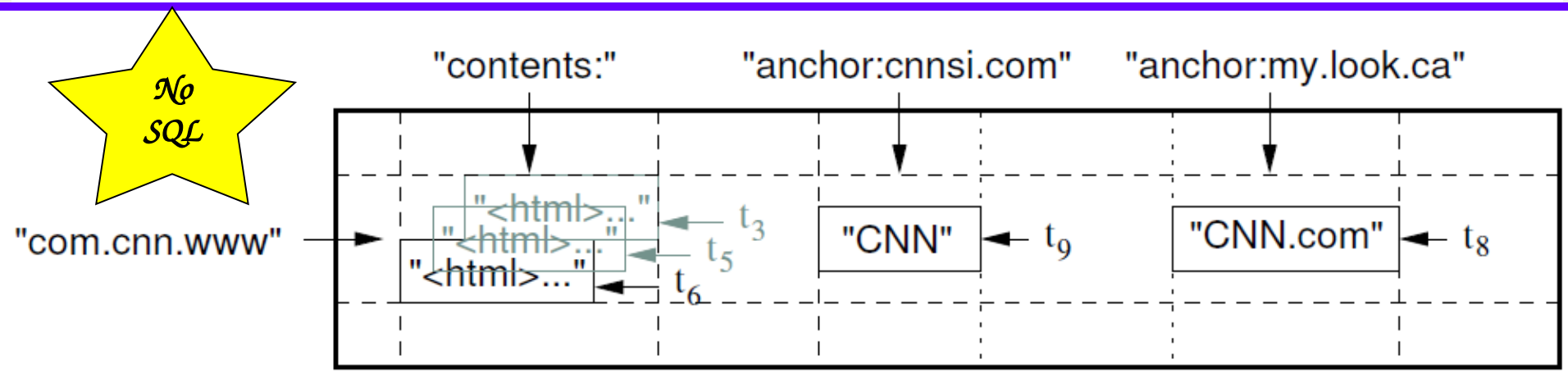
Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber

{fay,jeff,sanjay,wilsonh,kerr,m3b,tushar,fikes,gruber}@google.com

Google, Inc.

[Procs OSDI 2006]



- *Hbase, in Java, Apache license, runs on Hadoop*
- *HyperTable, in C++, GNU license, runs on Hadoop or GlusterFS*
- *Cassandra, in Java, Apache license 2, runs on Amazon's Dynamo*

“Smart” algorithms

2007

Economist.com **BUSINESS**

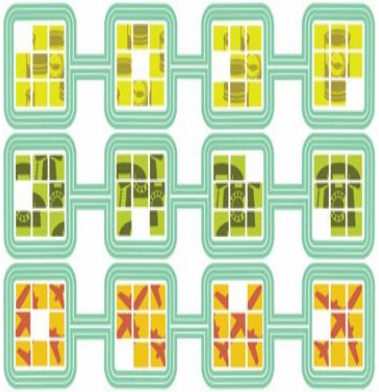
Algorithms

Business by numbers

Sep 13th 2007
From The Economist print edition

Consumers and companies increasingly depend on a hidden mathematical world

Illustration by Gillian Blease



ALGORITHMS sound scary, of interest only to dome-headed mathematicians. In fact they have become the instruction manuals for a host of routine consumer transactions. Browse for a book on Amazon.com and algorithms generate recommendations for other titles to buy. Buy a copy and they help a logistics firm to decide on the best delivery route. Ring to check your order's progress and more algorithms spring into action to determine the quickest connection to and through a call-centre. From analysing credit-card transactions to deciding how to stack supermarket shelves, algorithms now underpin a large amount of everyday life.

“This is rocket science but you don't have to be a rocket scientist to use it”