INTRODUCTION TO INFORMATION RETRIEVAL

Information Retrieval (IR)?

- IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
- IR deals with the representation, storage, organization and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of the information items should be such as to provide the users with easy access to information of their interest.
- IR is the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system

Information Retrieval (IR)?

- IR
  - is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.
- IR is interdisciplinary,
  - based on computer science, mathematics, library science, information science information architecture, cognitive psychology, linguistics, statistics, and physics.
  - are used to reduce what has been called "information overload".
  - Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Examples de search engines

DuckDuckgo

Information Retrieval (IR)?

- These days we frequently think first of web search, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval
  - Digital library
  - Enterprise search

Data Age 2025

The Evolution of Data to Life-Critical

Don't Focus on Big Data; Focus on the Data That's Big

Before 1980

- Data sits almost exclusively in datacenters
- Data and compute centralized
- Business-focused

1980—2000

- Data and compute are distributed
- Datacenters expand role in managing data
- Quick expansion in entertainment

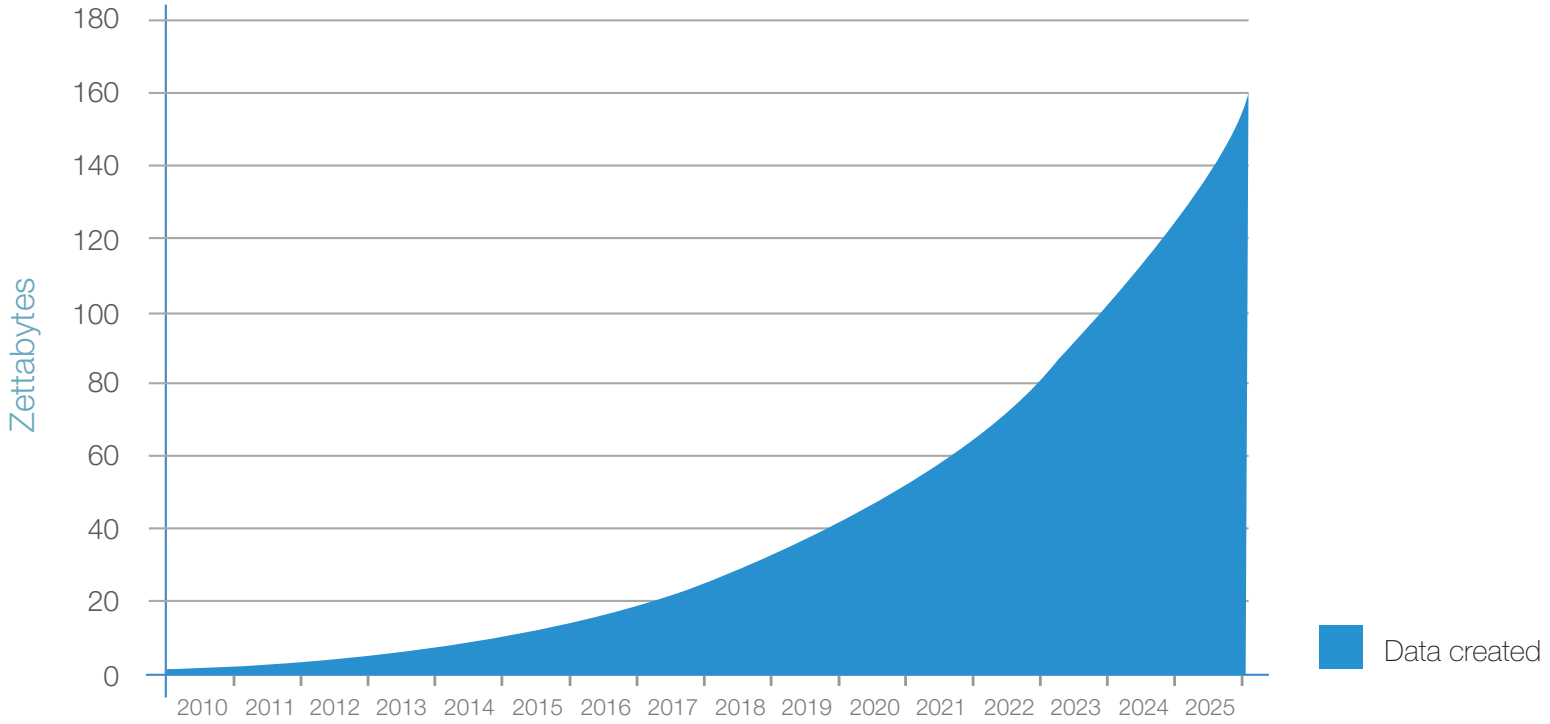
2000 to Today

- Datacenters expand to cloud infrastructures
- Compute continues to be distributed; data begins to contract
- Add social to the mix

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

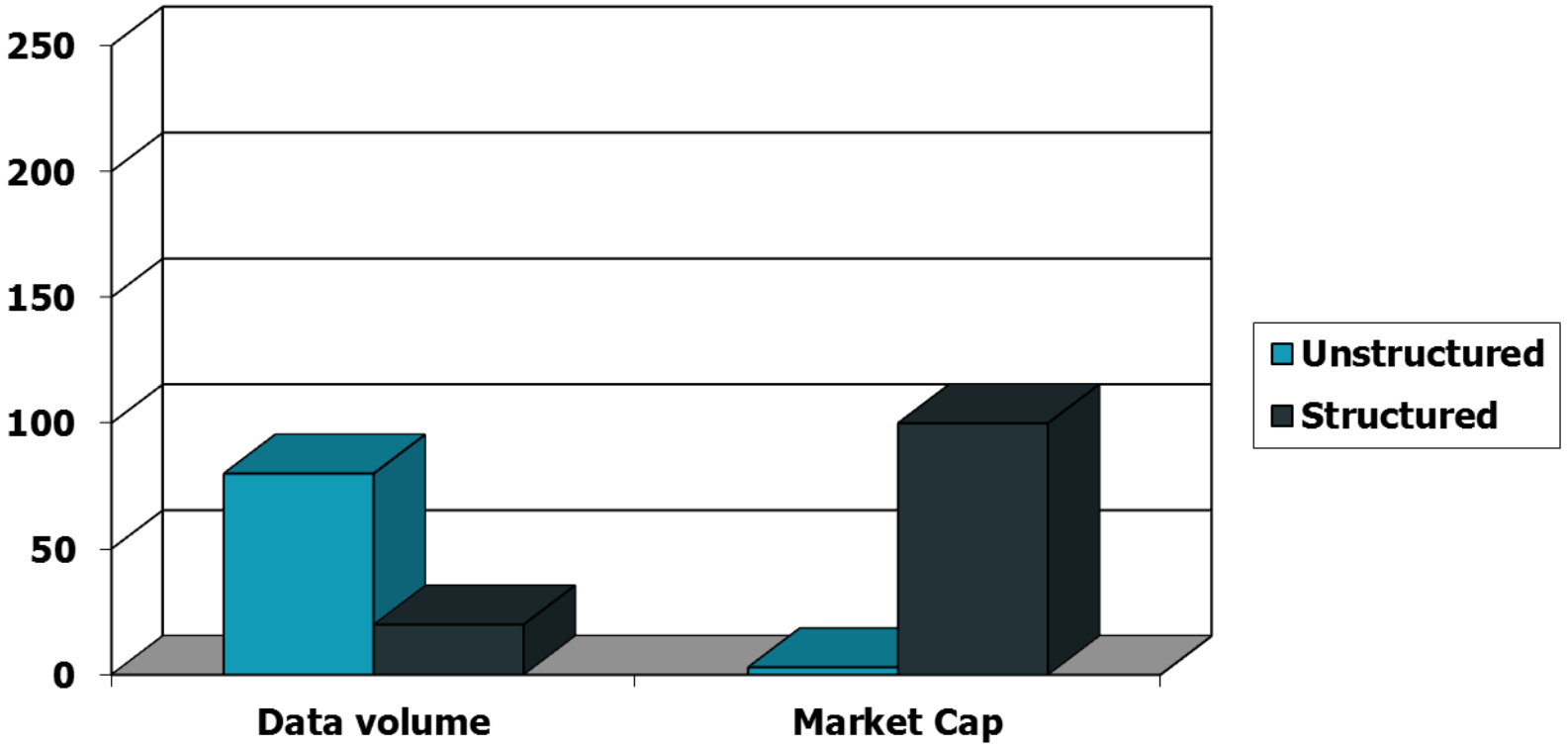
This is the state of our data-driven world today.

Annual Size of the Global Datasphere

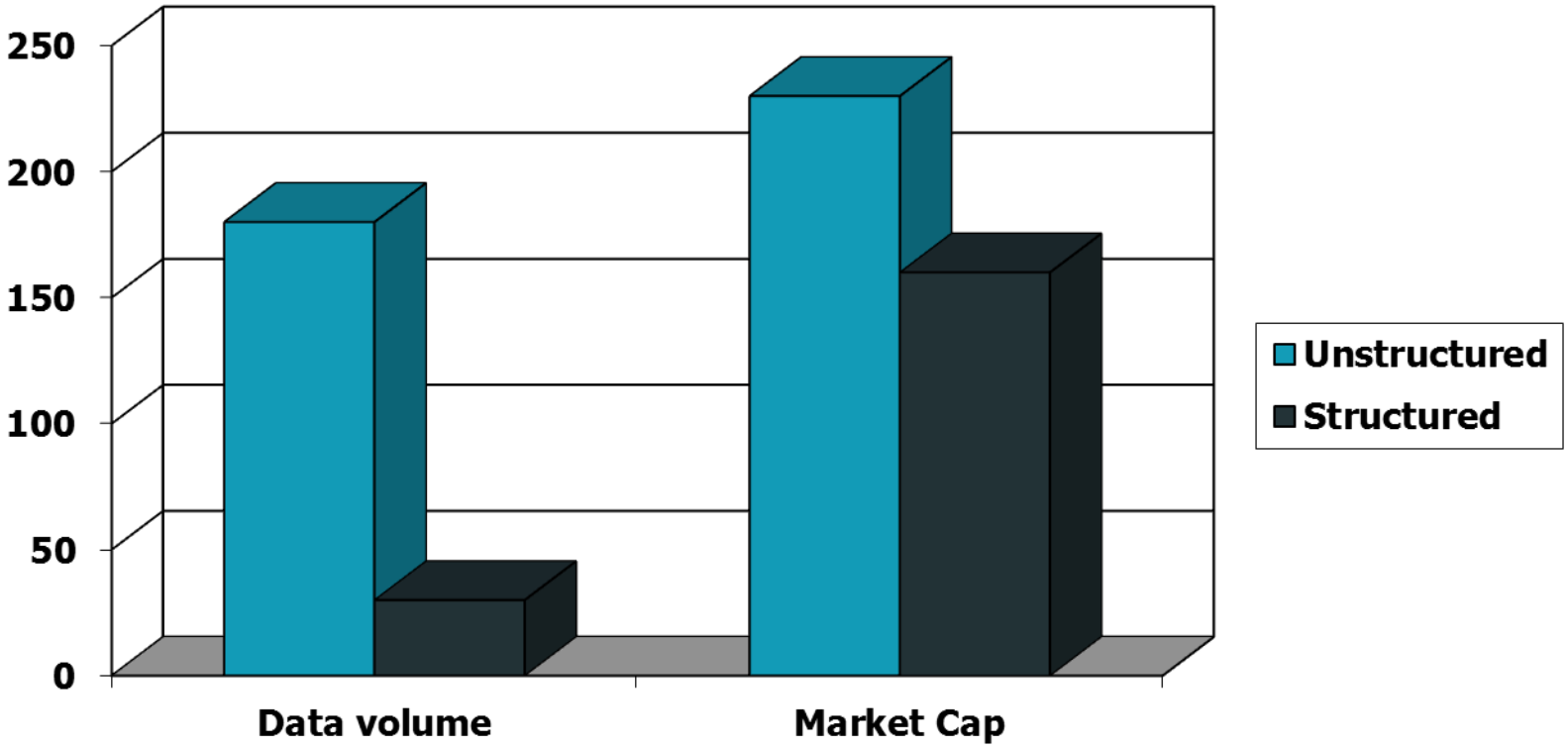


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

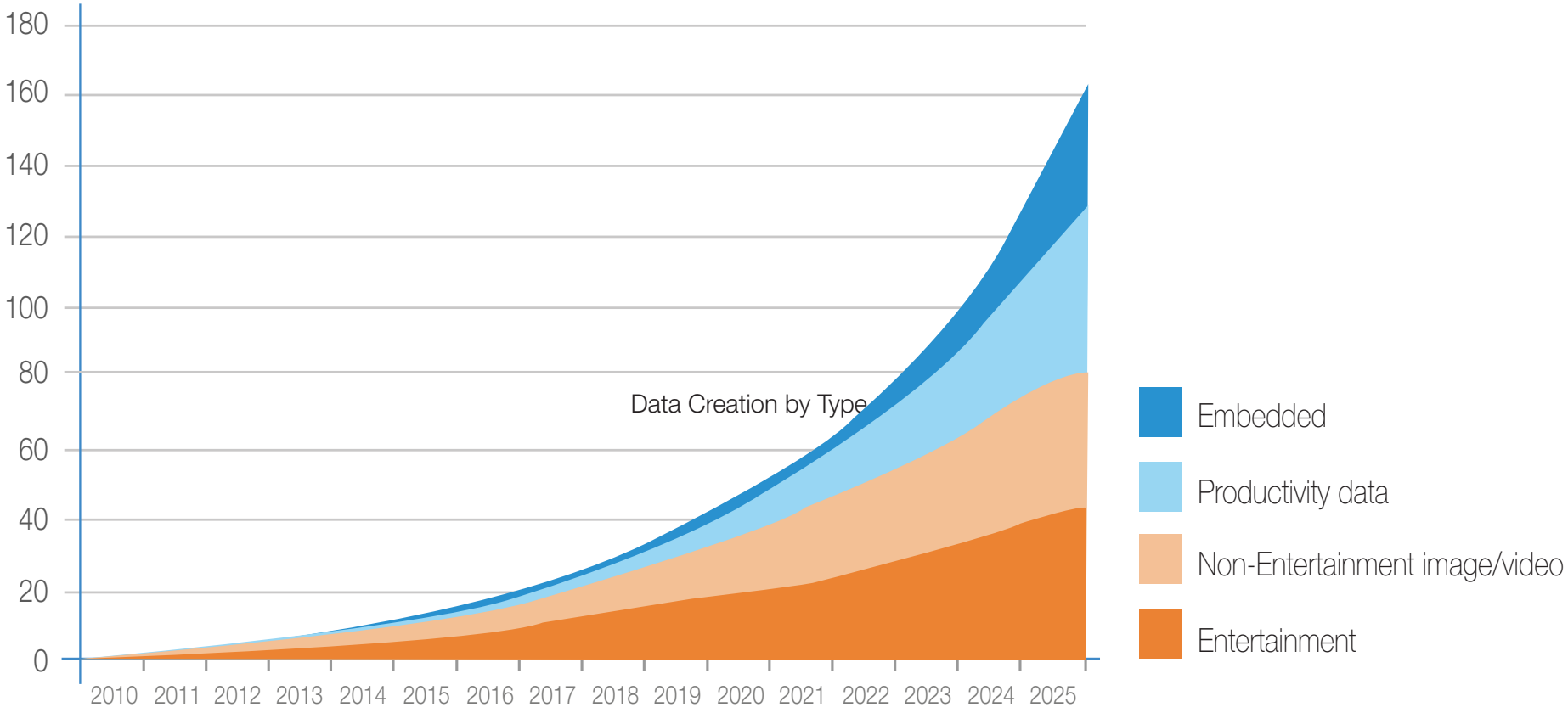
Unstructured (text) vs. structured (database) data in the mid-nineties



Unstructured (text) vs. structured (database) data today



Data Creation by Type

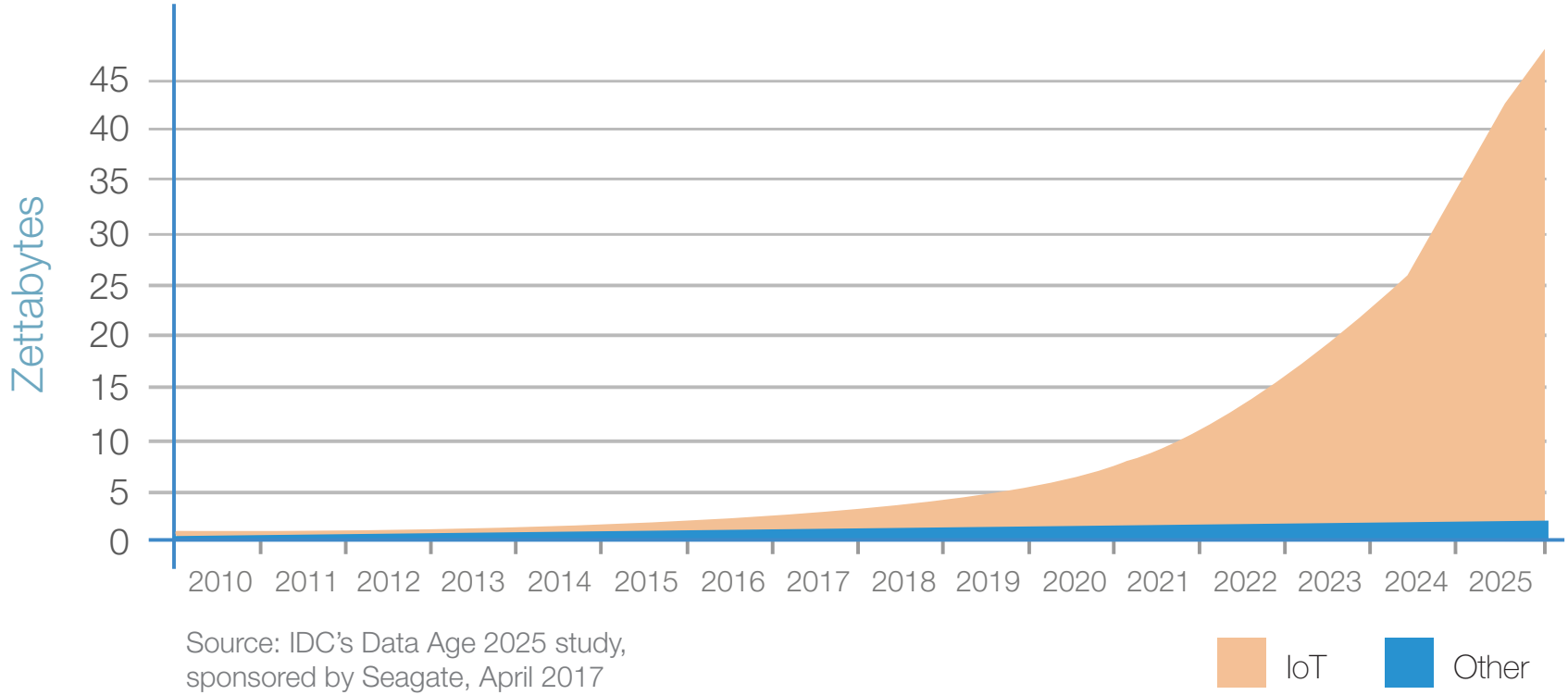


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

Embedded data

Security cameras
Smart meters
Chip cards
RFID readers
Fueling stations
Building automation
Smart infrastructure
Machine tools
Automobiles, boats, planes, busses, and trains
Vending machines
Digital signage
Casinos
Wearables
Medical implants
Toys

IoT Drives Real-Time Data



Source: IDC's Data Age 2025 study,
sponsored by Seagate, April 2017

- Information is every where:
  - Average Number of Tweets Sent Per Day: 500 million
    - 2 billions queries per day on twitter
  - Every minute 510,000 posted comments FaceBook
  - 45 billions (Google), 25 billions (Bing)
  - 672 Exabytes - 672,000,000,000 Gigabytes (GB) of accessible data.

The issues of IR

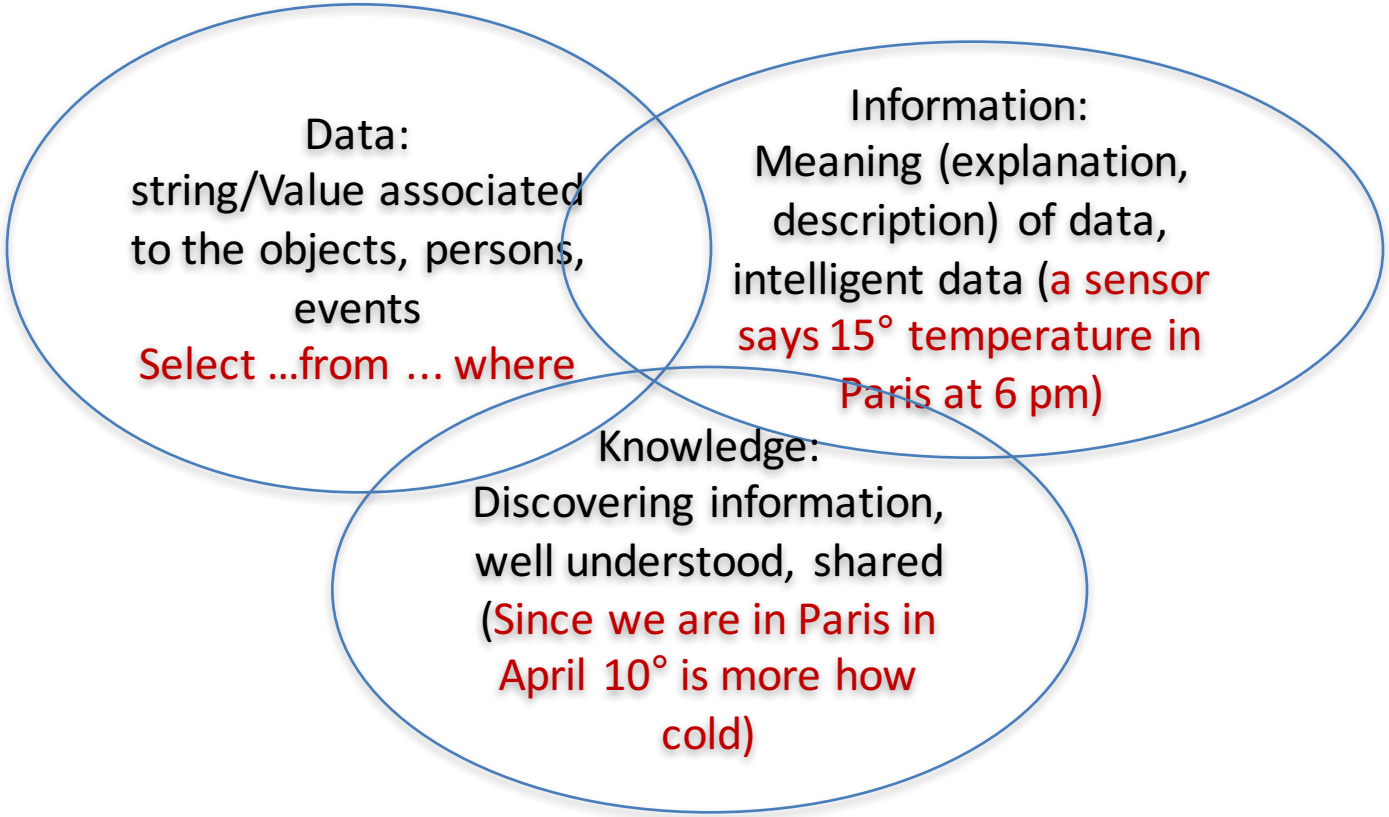
- Is not the availability of information
- But :
  - its selection, its identification : get the useful and right information in a good time

The issues (cont)

- Search information has a cost
  - We spend (35%-average)) of out time to search information
  - to manage it 17% of our time
  - The 1000 biggest enterprises (US) lost $2.5 billion/year because they did not get the right information
- The need to develop automated efficient allowing to: Collect, Organize, search, Select

Data vs Information vs knoweldge

- Confusion:



Steps of IR (Tasks)

- Adhoc search
  - I am searching information (web pages) on a topic
    - I submit a query -> list of results
    - Query «information search » SRI -> return a list of documents processing « information search
- Many types de IR adhoc
  - Adhoc search (spécifique tasks)
  - Specific Domaine (medical, legal, physics, ...)
  - Opinion retrieval (sentiment analysis)
  - Event retrieval Recherche d'événements
  - Person retrieval (expert)

Steps of IR

- Classification
  - grouping the informatios (documents) regarding many criterias
- Question-responses (Query answering)
  - Search responses to the queries
  - example
    - « who is Nobel ? »
    - « what is the width of Mississipi river? »

Steps of IR

- Filtering/recommendation
  - Recommandation
  - Alert systems
  - Selective Dissemination of information
  - Push
  - Profiling

Steps of IR

- document summarization
- Aggregated search
  - Aggregating search engines : querying the results of multiple engines (meta-engines)
  - Aggregating the results : querying multiple sources (vertical search)
  - Aggregating the content: providing a result from multiple contents

Basic assumptions of Information Retrieval

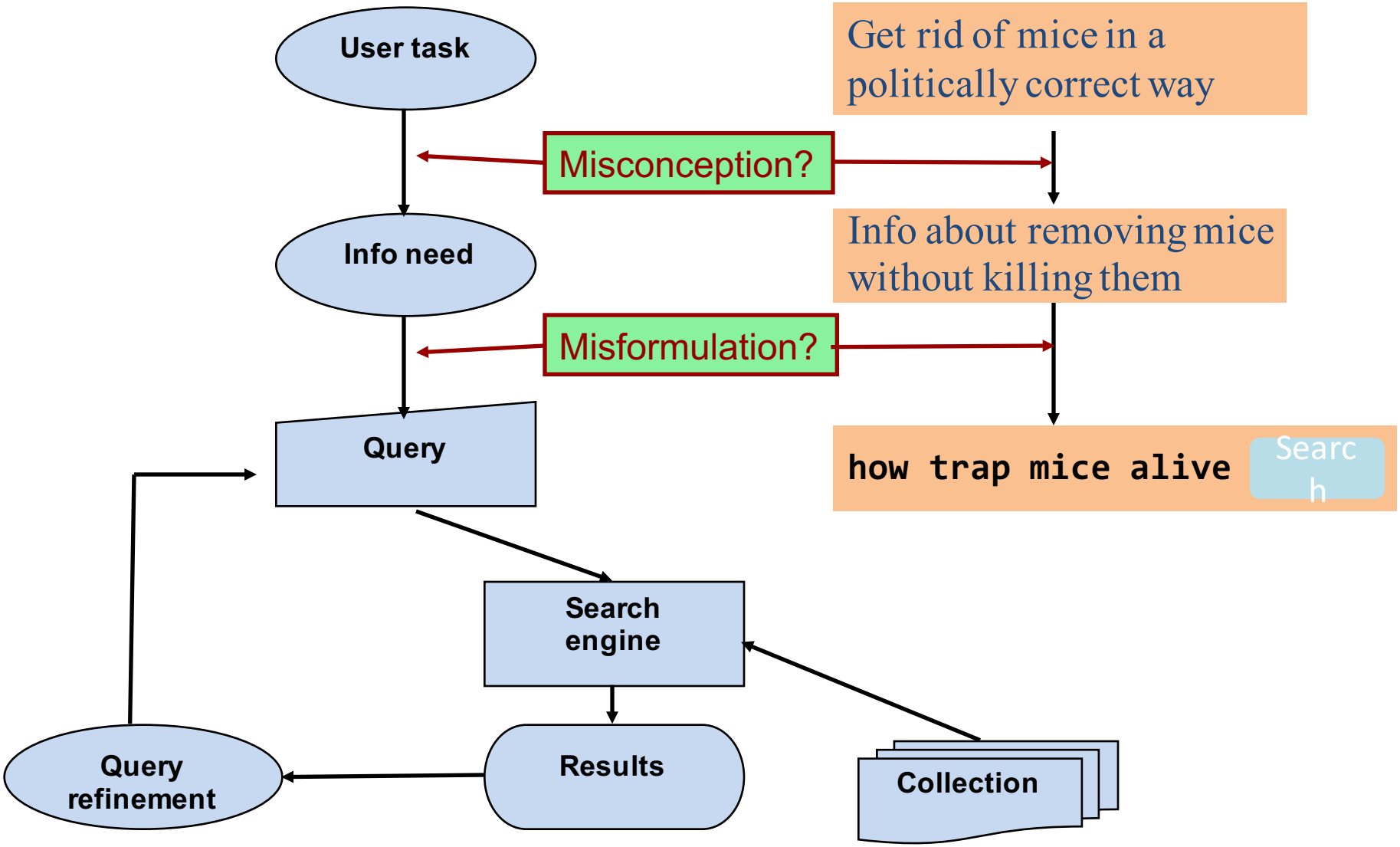
- Collection: A set of documents
  - Assume it is a static collection for the moment

- Goal: Retrieve documents with information that is relevant to the user's information need and helps the user complete a task

Assumptions

- Need = query
  - The need confused with the query user (a set of keywords)
- Document and query
  - Represented by terms (simple words, groups of words, ...) -> bag of words
- Pertinence
  - represents by the similarity of words bewteen the query and documents

The classic search model



Basic concepts of search model

The method of IR

- Interpret the text rather than understanding it

- Exploit the statistic properties (word counting) of the text rather than lingusitic properties

Basic concepts of search model

d1:
So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

d2:
I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Traitement
=
Indexation

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
did	1	1	
enact	1	1	
hath	1	1	
I	1	2	
i'	1	1	
it	1	1	
julius	1	1	
killed	1	2	
let	1	1	
me	1	1	
noble	1	1	
so	1	1	
the	2	2	
told	1	1	
you	1	1	
was	2	2	
with	1	1	

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

Inverse index

Basic concepts of search model

- Factors used in the most of models
  - Frequency of the terms in a document (tf), = (Number of times term t appears in a document) / (Total number of terms in the document). measures how frequently a term occurs in a document.
  - Frequency in the collection (idf) (**Inverse Document Frequency**)= log_e(Total number of documents / Number of documents with term t in it). ((measures how important a term is.)
  - Its position in the text (p)
  - Size of the document (dl)

$$Score(D) = fonction(tf, idf, dl)$$

- Many theoritical models to formalize the fomula
- It can be learned (machine learning used by most of search engines) https://github.com/gearmonkey/tfidf-python

- Consider a document containing 100 words wherein the word *cat* appears 3 times. The term frequency (i.e., tf) for *cat* is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word *cat* appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

IR Models

- Ensemble theory
  - Boolea Model (more than 1950)
- Algebra
  - Vector space model
  - Spreading activation model
  - LSI (Latent Semantic Indexing)
- Probability
  - Probabilsitic model
  - Inference network model
  - Language model
  - DFR (Divergence from randomness model

- The search engine is different the one used in a directory.
- Software robots (called crawlers, spiders, bots) search the web using the links between pages and documents and indexing automatically the found documents allowing a keywords search
- The indexes of search engines index billions of web pages. (page ranking,

- Each search engine has its own algorithms but the mode processing is the same
- The main elements of a page considered by the algorithms are:
  - The structured elements (URL, name of the domain,..)
  - Title page (balise title)
  - the content of the text
  - Different elements highlighting different html balise
  - The popularity of the page and the site (external links)
  - Internal meshing (network)…..

How good are the retrieved docs?

- *Precision* : Fraction of retrieved docs that are relevant to the user's information need
- *Recall* : Fraction of relevant docs in collection that are retrieved

  - More precise definitions and measurements to follow later

Google algorithm zoo

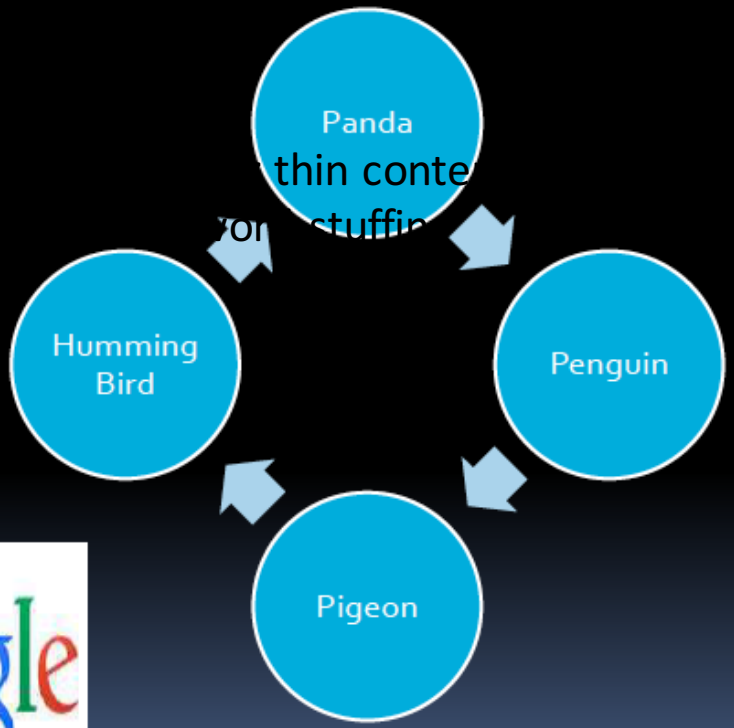
- 4 Different Algorithms
  - a Panda
  - Penguin
  - Pigeon
  - Humming Bird

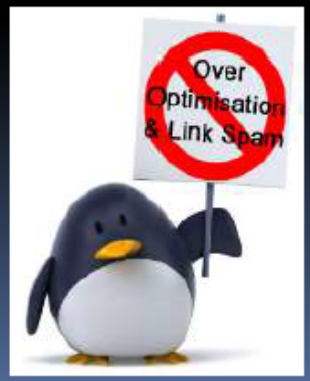
Google Algorithms

- Google's algorithm does the work for you by searching out Web pages that contain the keywords you used to search, then assigning a rank to each page based several factors, including how many times the keywords appear on the page. Higher ranked pages appear further up in Google's search engine results page (SERP), meaning that the best links relating to your search query are theoretically the first ones Google lists.
- From time to time Google algorithm works in order to provide better search results and relevant content for internet users.
- Google algorithm prevents cheating on the website.

Google Algorithms

Duplicate.





rd stuffing

nda Algorithm

lgorithm is used to reduce rankings for
quality sites
which are low - value add for users ,copy content from other
tes
es that are just not very useful

e same time, it will provide better rankings for
quality sites, sites with original content and information

alizes a website that has nothing but only pop-ups and ads.

a updates are focused on penalizing the ranking of sites with a poor
xperience (like the one mentioned above) and improves the ranking
es with a positive user experience.

does Google do this... because you want to find a website that
des an answer to what you are looking for, not a website with a
n of popup and ads.

t

Penguin Algorithm

It is commonly known that the more links back to your website the better SEARCH ENGINE RANKING you will have.
The problem is getting links back to your website is not an easy thing to do. When things are difficult, people cheat. Penguin is designed to penalize those who cheat buy purchasing links from 'link farms' or 'link exchanges' back to their website.

So what is a 'link farm' or 'link exchange'? Websites that are designed to create links back to other websites for a fee in order to achieve a higher SEARCH ENGINE RANKING .

If you are not involved in a 'paid link' scheme or intentionally trying to manipulate search results via links, then Penguin is nothing to worry about.

Spammy or irrelevant links; links with over-optimized anchor text

Pigeon Algorithm

This algorithm ties deeper into their web search capabilities, including the hundreds of ranking signals they use in web search along with search features such as Knowledge Graph, spelling correction, synonyms and more.

This algorithm improves their distance and location ranking parameters.

It provides more relevant and accurate local search results.

Poor on- and off-page SEO

keyword stuffing; low-quality content

Hummingbird Algorithm

Hummingbird's strength is the ability to quickly analyze longer, more complex questions and provide the best answer to the searcher with the fewest possible clicks.

"Hummingbird" algorithm is a more human way to interact with users and provide a more direct answer unlike its previous versions Panda and Penguin.

helps Google better interpret search queries and provide results that match searcher intent

- Conferences
  - ACM SIGIR: Special interest group on Informayion Retrieval
  - CIKM: Conference on Information and Knowdedge Management
  - VLDB: Very large DataBase
  - SIGMOD
  - WWW
- Journals
  - ACM TOIS
  - JIR
  - VLDB

Introduction to

Information Retrieval

Structured vs. Unstructured Data

IR vs. databases: Structured vs unstructured data

- Structured data tends to refer to information in "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.

Unstructured data

- Typically refers to free text
- 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

Semi-structured data

- In fact almost no data is “unstructured”
- E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*
    - … to say nothing of linguistic structure
- Facilitates “semi-structured” search such as
  - *Title* contains data AND *Bullets* contain search
- Or even
  - *Title* is about Object Oriented Programming AND *Author* something like stro*rup
  - where * is the wild-card operator