Gestione dell'Informazione

Part A – Full-Text Information Management

Query Languages

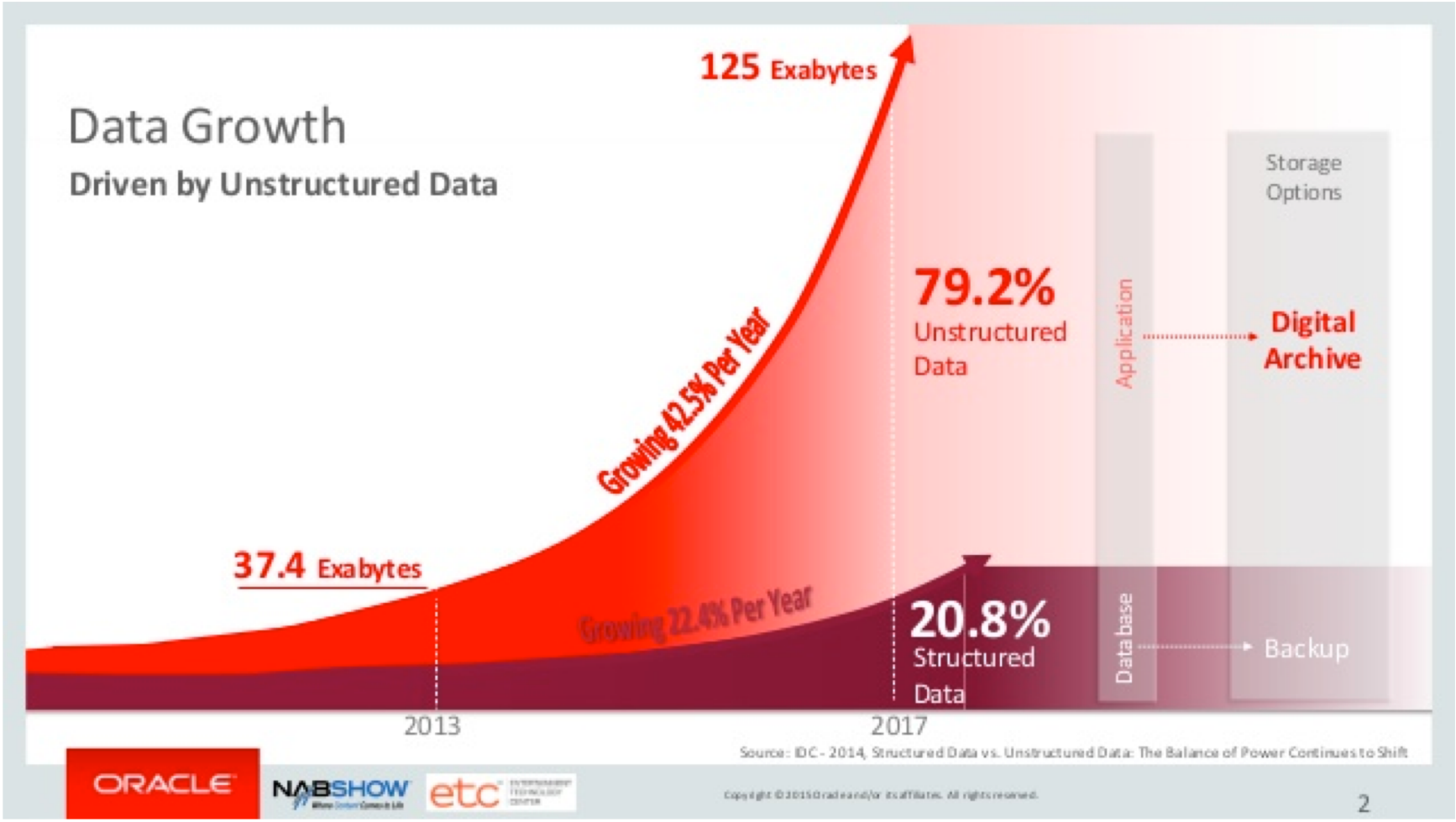
Contents

Text is ubiquitous

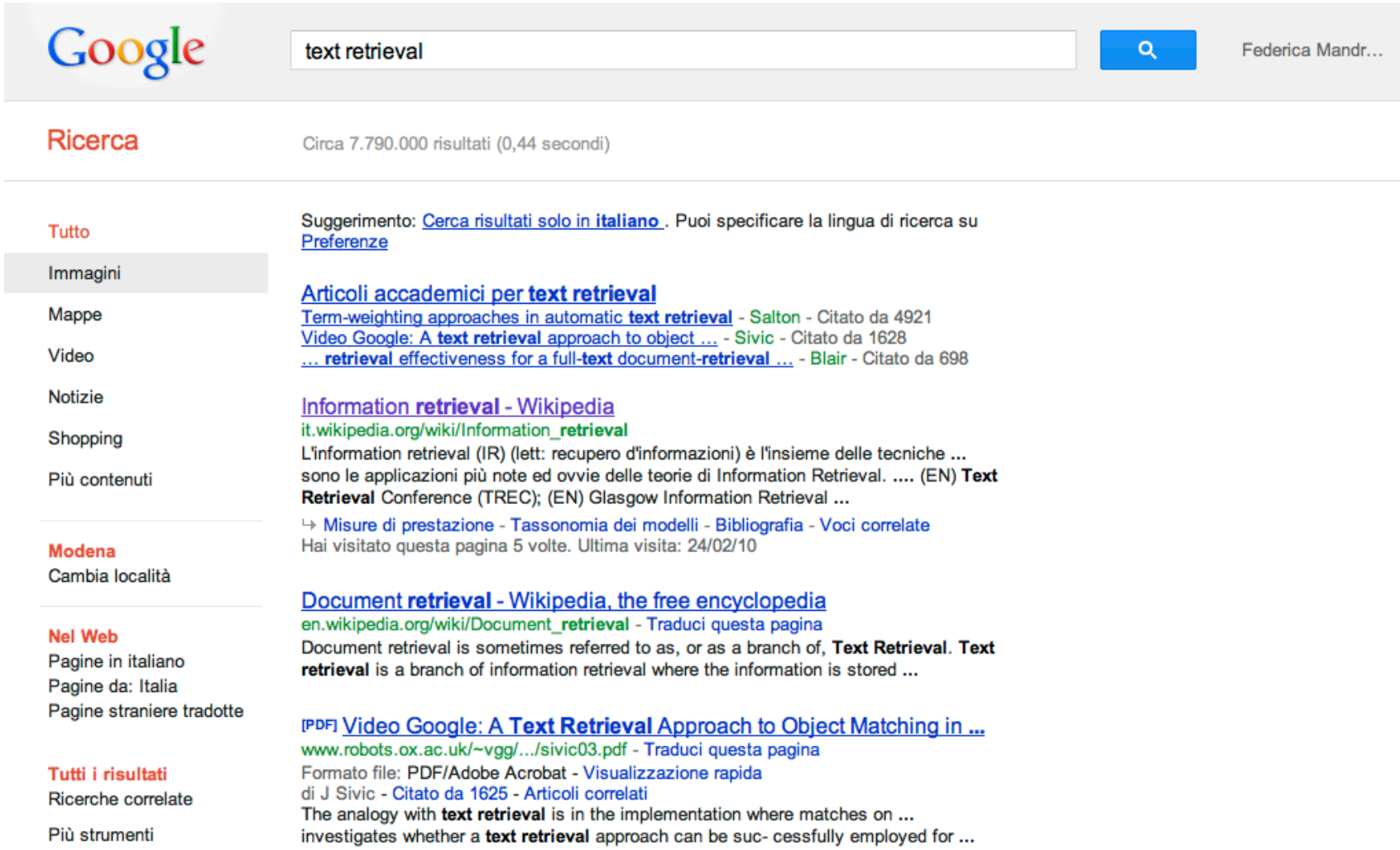
- text is in web pages
- text is in documents
  - Newpapers
  - Articles
  - …..
- text is in data items
  - product names and descriptions
  - names, surnames, addresses
- text is in blogs, forum, social networks
- E-mails, SMSs are text items
- Biological data are often represented as sequence of characters
- …

Unstructured vs Structured data growth

1 Exabyte = 1 million of Terabytes



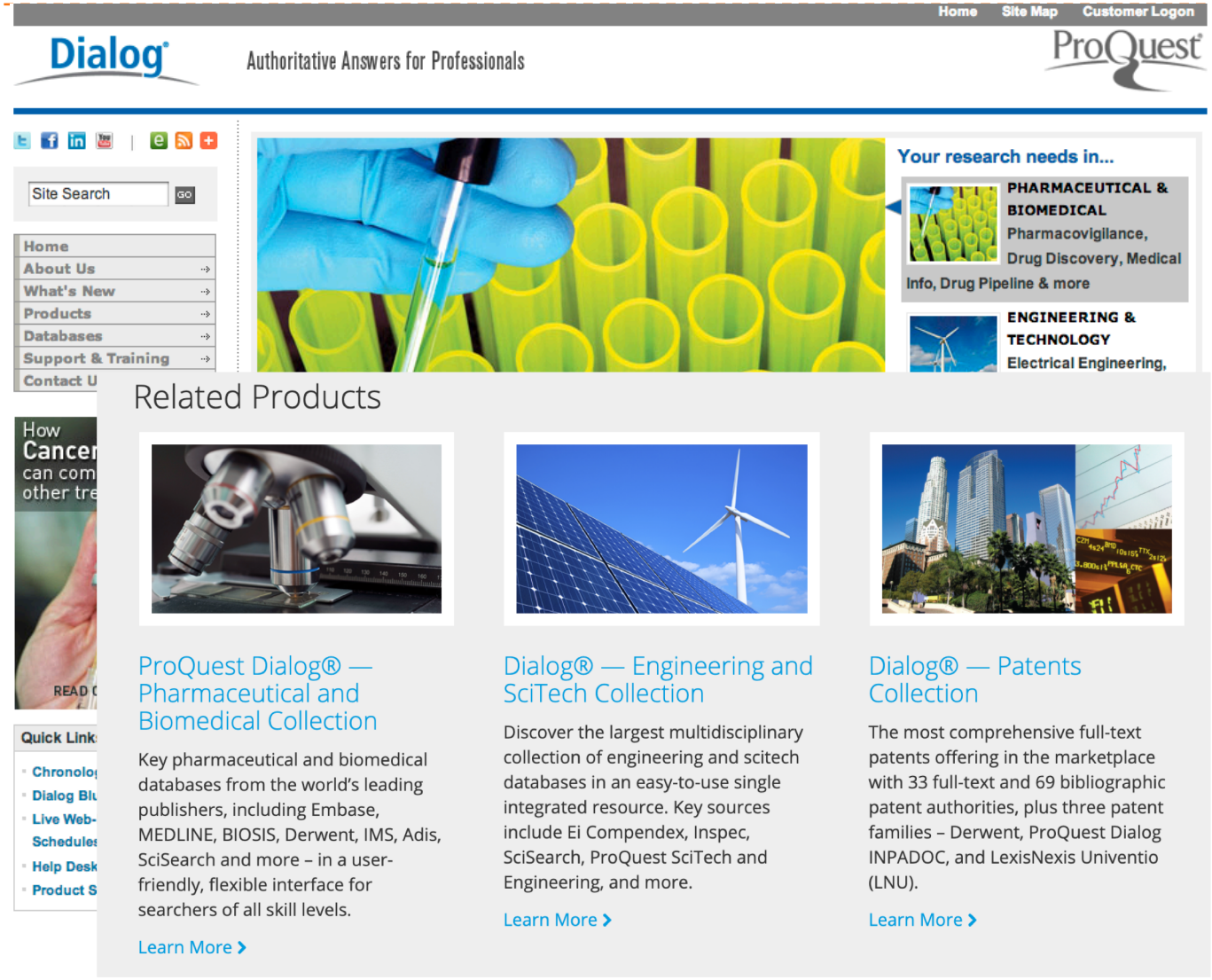
Samples of Applications searching text: Search Engines



Samples of Applications searching text: Digital libraries

DIALOG:

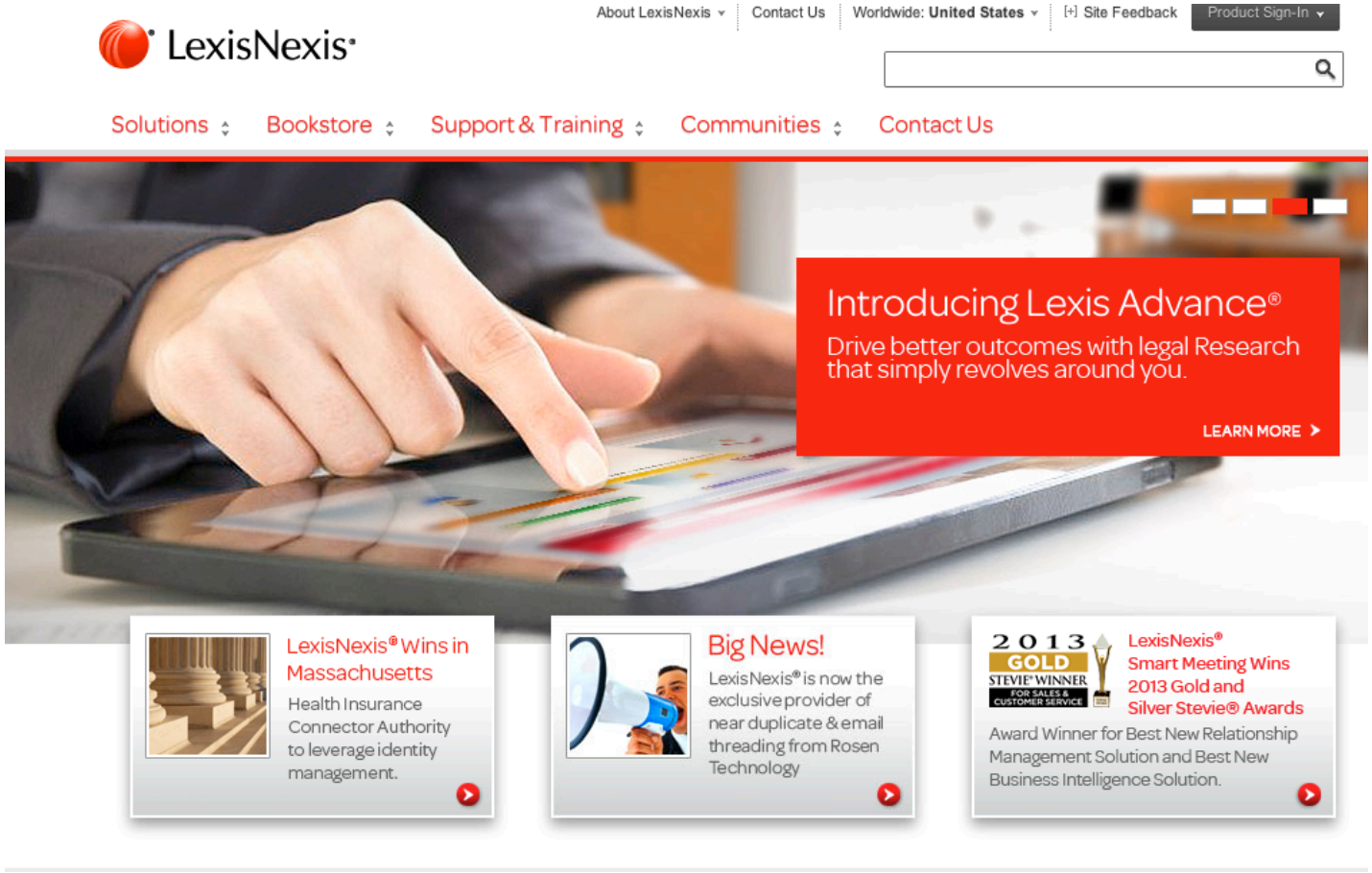
- ✓ 15 terabytes of content
- ✓ 900 databases
- ✓ 700,000 searches handled per month
- ✓ over 17 million document page views delivered per month



Samples of Applications searching text: Digital libraries

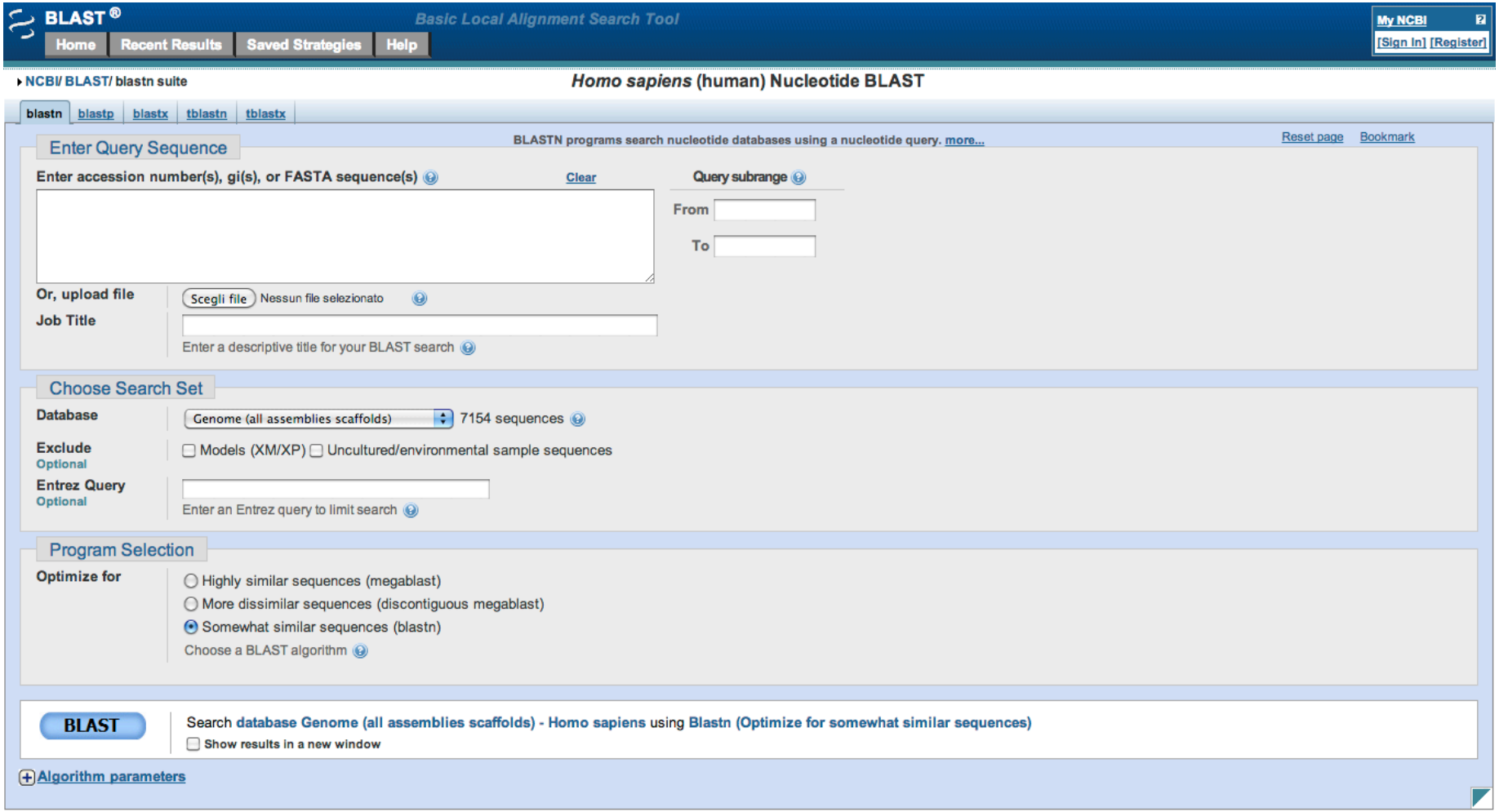
LEXIS-NEXIS:

- ✓ 3.3 billion of data records

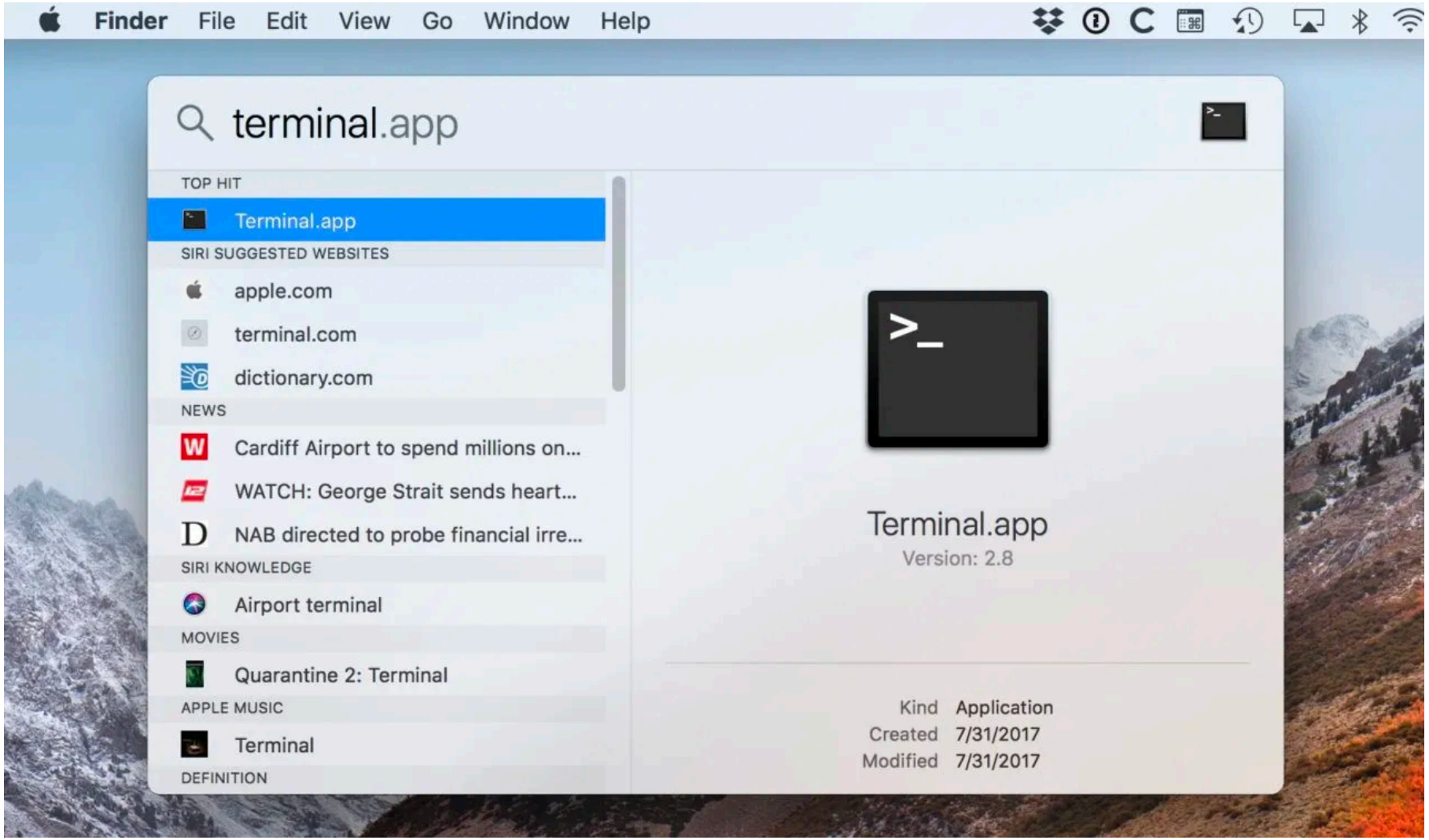


Samples of Applications searching text: search form in web sites and apps

Samples of Applications searching text: Biological databank applications



Samples of Applications searching text: File system find



Information Retrieval: some definitions

- "Dealing with the representation, storage, organization of information items for providing the user with easy access"

  Modern Information Retrieval

- "The term Information Retrieval refers to a search that may cover any form of information: structured data, text, video, image, sound, musical scores, DNA sequences, etc."

  Information Retrieval – Algorithms and Heuristics

Basic assumptions of Information Retrieval

- **Collection**: Rather fixed set of information items
- **Goal**: Retrieve items with information that is relevant to the user's information need and helps the user complete a task

History of IR

- 1951: Calvin N. Mooers coins the term `Information Retrieval'

"IR is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. (. ..). IR embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, and machines that are employed to carry out the operation"

- 1960': **SMART** (System for the Mechanical Analysis and Retrieval of Text) developed at Cornell University by **Gerard Salton** and his group

IBM 7094-II (shown left)

- ✓ Speed: 0.35 MIPS (Millions of Instructions Per Second)
- ✓ Memory: 32 kilobytes

History of IR

- 1972: Roger Summit et al. introduce **DIALOG** as commercial online service
- Late 1980's: First PC systems incorporate retrieval
- Early 1990's: Cheap disks lead to the information storage revolution
- 1992: **Westlaw** is the first large-scale information service using probabilistic retrieval
- Mid 1990's: Multimedia databases
- 1994: The internet and web explosion
- 1995: IR techniques are incorporated in all kinds of information management applications
- …

Information versus Data Retrieval (1/2)

Data retrieval (DR)

- *Dealing* with data having well defined structure and semantics
- *Retrieving* all objects satisfying clearly defined conditions in a regular expression
- *Providing* a solution to the user of DB systems

Information retrieval (IR)

- *Dealing* with natural language text not well structured and semantically ambiguous

⇒ **Content based retrieval**

- *Retrieving* information in large collection of text items
- *Retrieving* objects inaccurate and small errors unnoticed

⇒ **Approximate search, ranking**

Information versus Data Retrieval (2/2)

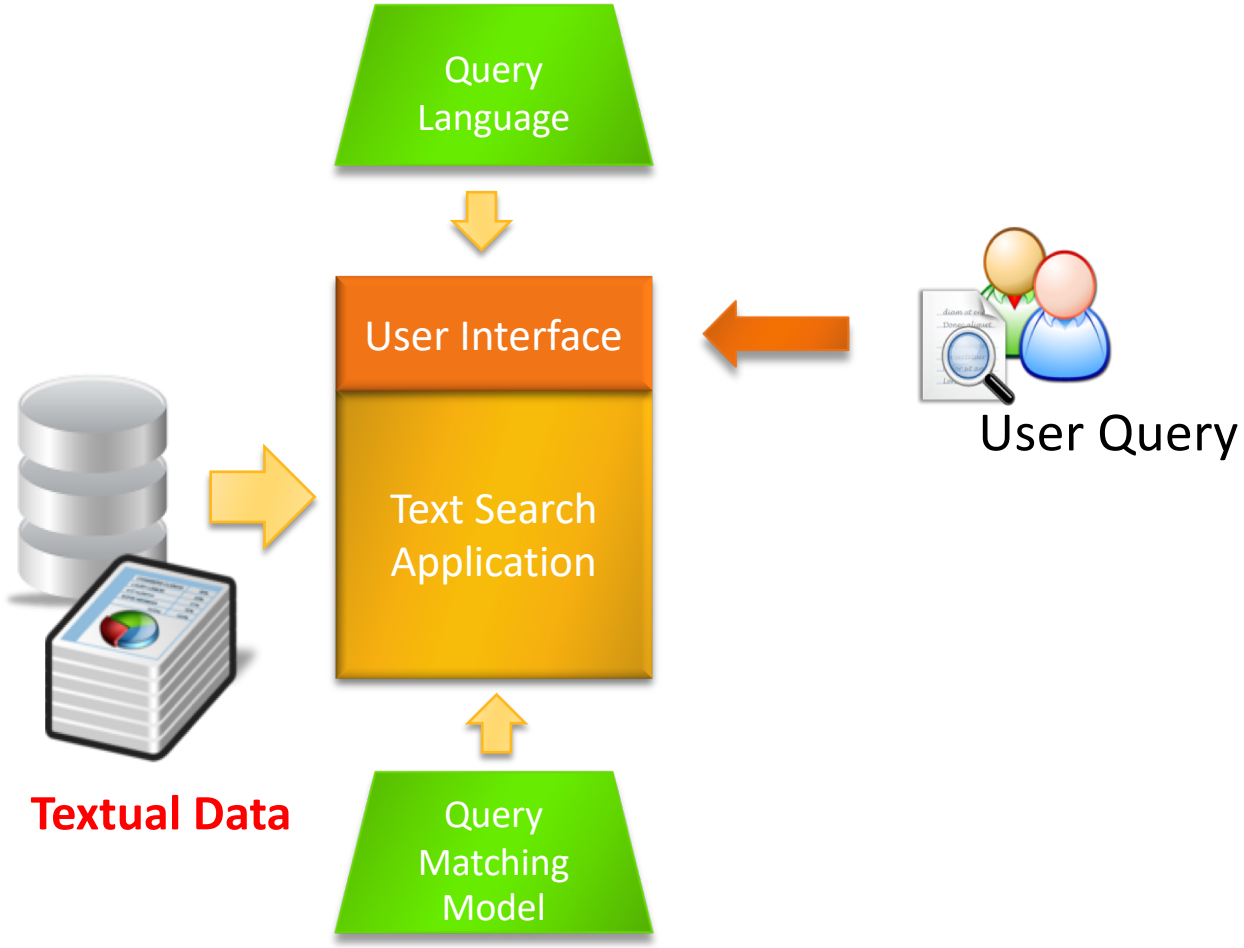
Interpretation of text content for IR

- *Extracting* syntactic and semantic information from data item
- *Using* the extracted information to match the User Information Need (**UIN**)
- Difficulties: knowing how to extract this information and to use it to decide *relevance*

Primary goal of IR

- Retrieve all data items *relevant* to a user query while retrieving as few *non-relevant* data items as possible

Text Search Application overview



Contents

Query languages

- The user
  - Search expert (e.g., librarian) vs. non-expert
  - Background of the user (knowledge of the topic)
  - In-depth searching vs. `just-wanna-get-an-idea' searching
- **UIN** (User Information Need)
  - Natural language declaration of the informational need of a user
    - "Find all the web pages (documents) containing information on college tennis teams which are maintained by an university in the USA."
- Query languages
  - Keyword-based querying
  - Pattern matching
  - Structured querying
  - …
- Any **application searching for text** has to define its own **query language**. This choice is largely dependent on
  - User skills
  - Managed text
- Any **user** has to translate his/her information need into a query in the supported language

Keyword-based querying

Word

- The most elementary query used in full-text applications
- a sequence of letters surrounded by separators

- **Multiple word query**
  Ex) information retrieval
- Search words in a given context
- **Phrase query**
  - Retrieve documents with a specific phrase (ordered list of contiguous words)
    Ex) “enhance retrieval”
  - May allow intervening stop words and/or stemming
    Ex) “enhance retrieval” matches “enhance the retrieval”
- **Proximity query**
  - A phrase query with a maximum allowed distance (character or word) between words in the query
    Ex) If distance = 4 → ‘…*enhance* the power of *retrieval*…’
  - The same order of the words may or may not be required by an application
- **Google Web Search guide**
  https://support.google.com/websearch/answer/134479?hl=it&ref_topic=3081620

Keyword-based querying

- **Boolean query**
  - Basic queries
    - Single-word query, multiple-word query, patterns
  - Query composed of
    - basic queries
    - Boolean operators
      - AND, OR, NOT
- A **natural language query** is simply an enumeration of words and context queries

  "Find all the web pages (documents) containing information on college tennis teams which are maintained by an university in the USA."

Pattern and Pattern Matching

Pattern Pieces of text that have some syntactic properties

- ✓ E.g. comput*

Pattern matching

- allows queries that (approximately) *match text* rather than word tokens
- **Matched pattern**
  - *Text segment* satisfying the pattern specifications
- Such queries can be used in the composition mechanism
  - as basic queries
  - to form phrases and proximity queries
- Each application allows the specification of some kind of patterns from very simple (e.g. words) to rather complex (e.g. regular expressions)

Types of patterns

- **Prefixes:** string forming the beginning of a text word
  Ex) comput* -> computer, computation
- **Suffixes:** string forming the termination of a text word
  Ex) *ters -> computers, painters
- **Substrings:** string appearing within a text word
  Ex) *tal* -> talk, metallic
  any flow -> …many flowers…
- **Ranges:** a pair of strings matching any word (alphabetically) between
  Ex) held and hold -> hoax, hissing
  - Google range query - number…number: $250…$500 laptop

Structural querying

- Assumes documents have structure that can be exploited in search
- Allow queries for text appearing in specific fields:
  - “nuclear fusion” appearing in a chapter title
- Several proposals extend SQL to allow full-text retrieval
  - Database vendor such as Oracle, DB2, etc.
- Google books
- Pubmed: http://www.ncbi.nlm.nih.gov/pubmed
- Gene: http://www.ncbi.nlm.nih.gov/gene
- …

Concept-based querying

- Adoption of a controlled vocabulary
- The user selects one or more terms from the vocabulary to be searched

PROS

- Assist users with proper query formulation
- Provide classified hierarchies that allow the broadening and narrowing of the current query request
- Particularly important for specific domain (e.g. medicine)
- The MESH ontology in PUBMED

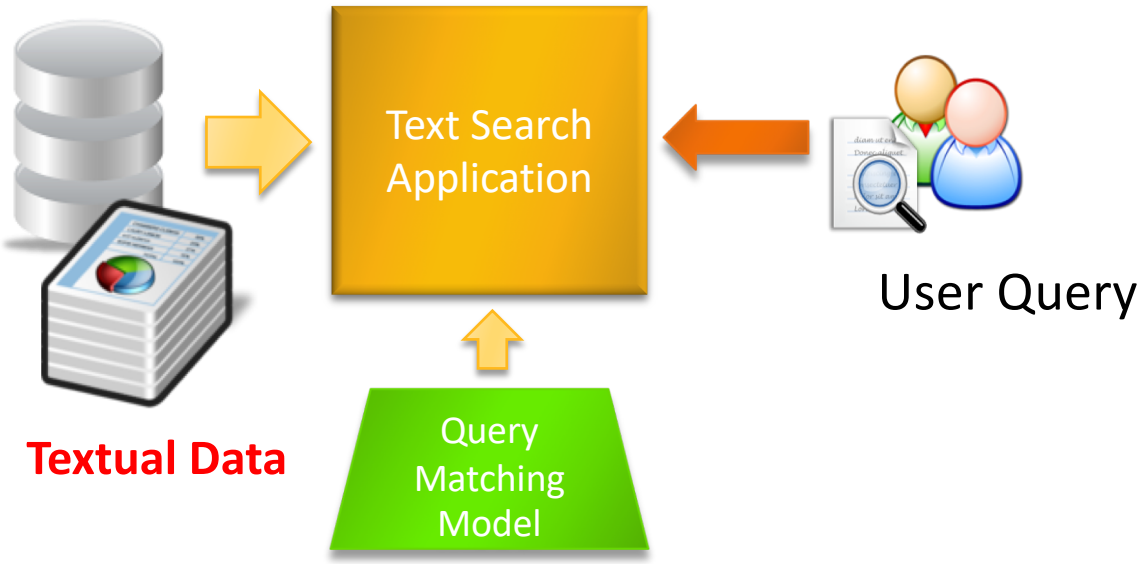
More on user-interfaces

- **Main objective**: to help users in specifying the query that best suits their information need
  - Free text fields are not always the best choice
  - E.g. Vertical portals
- Different strategies
  - User profile
  - Word lists
  - Tag Cloud

    Ex) http://tagcrowd.com/
  - Dictionary visualization
    - E.g. www.visuwords.com
      - Based on Wordnet
      - Source code available

Query processing

- Two approaches
1. **Online query processing**
- When text is
  - Volatile
    - E.g. publish/subscribe systems
  - Short
- Sometimes, it represents the only option
  - Client-side processing
  - Agent
    - E.g. crawler
2. **Store first-query later**
- when text is
  - Big



Query Matching Model

- Given
  - A query ***Q***
  - A text item ***T***
- The query matching model "says"
  - Whether the text ***T*** is relevant for the query ***Q***
  - (How much it is relevant)
    - Ranking

Store first-query later technology