UNIVERSITÀ
DEGLI STUDI
DI MILANO

LA STATALE

Master Degree in Computer Science

Information Retrieval

Introduction to Information Retrieval

Prof. Alfio Ferrara

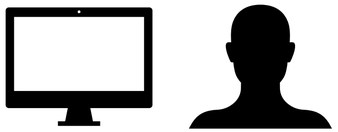
Department of Computer Science, Università degli Studi di Milano
Room 7012 via Celoria 18, 20133 Milano, Italia alfio.ferrara@unimi.it

sed noli modo

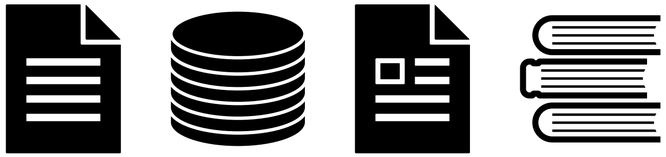
“Like all men of the Library, I have traveled in my youth; I have wandered in search of a book, perhaps the catalogue of catalogues…”

— Jorge Luis Borges

Information retrieval (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).



INFORMATION NEED



DOCUMENT COLLECTIONS
(Aka *corpora*)

Different types of information need

Source: https://trends.google.com/

General info search

1) Election results
2) Coronavirus
3) Kobe Bryant
4) Coronavirus update
5) Coronavirus symptoms
6) Zoom
7) Who is winning the election
8) Naya Rivera
9) Chadwick Boseman
10) PlayStation 5

Entities (e.g., movies)

1) Parasite
2) 1917
3) Black Panther
4) Harley Quinn: Birds of Prey
5) Little Women
6) Just Mercy
7) Bad Boys 3
8) Sonic the Hedgehog
9) Contagion
10) Fantasy Island

How to ...

1) How to make hand sanitizer
2) How to make a face mask with fabric
3) How to make whipped coffee
4) How to make a mask with a bandana
5) How to make a mask without sewing
6) How to make cloud bread
7) How to make Facebook avatar
8) How to make Bitmoji classroom
9) How to make disinfectant wipes
10) how to make a live wallpaper

Others: Specific structured info (e.g., how old is Joe Biden), rankings (e.g., top 10 boardgames 2021), opinion finding (e.g., best restaurant in NYC), similarity (e.g., find an image similar to this one), ...

Information needs change in time

Source: https://trends.google.com/

2012

1) Whitney Houston
2) Hurricane Sandy
3) Election 2012
4) Hunger Games
5) Jeremy Lin
6) Olympics 2012
7) Amanda Todd
8) Gangnam Style
9) Michael Clarke Duncan
10) KONY 2012

2016

1) Powerball
2) Prince
3) Hurricane Matthew
4) Pokémon Go
5) Slither.io
6) Olympics
7) David Bowie
8) Trump
9) Election
10) Hillary Clinton

2020

1) Election results
2) Coronavirus
3) Kobe Bryant
4) Coronavirus update
5) Coronavirus symptoms
6) Zoom
7) Who is winning the election
8) Naya Rivera
9) Chadwick Boseman
10) PlayStation 5

Information needs change in space

Source: https://trends.google.com/

GLOBAL 2020

1) Coronavirus
2) Election results
3) Kobe Bryant
4) Zoom
5) IPL
6) India vs New Zealand
7) Coronavirus update
8) Coronavirus symptoms
9) Joe Biden
10) Google Classroom

ITALY 2020

1) Coronavirus
2) Elezioni USA
3) Classroom
4) Weschool
5) Nuovo Dpcm
6) Diego Armando Maradona
7) Kobe Bryant
8) Meet
9) Contagi
10) Protezione Civile

USA 2020

1) Election results
2) Coronavirus
3) Kobe Bryant
4) Coronavirus update
5) Coronavirus symptoms
6) Zoom
7) Who is winning the election
8) Naya Rivera
9) Chadwick Boseman
10) PlayStation 5

Information needs VS queries

Information needs are not queries. A query is a way we express our information need, but quite often the **real need is implicit in the query**.

Query

“list of exercises in the IR exam”

Need

How to pass the exam the easy way?

Variety in the document collections

Scale

- Single document scale
- Personal scale (i.e., mailbox)
- Institutional scale (i.e., University library)
- Global scale (i.e., web)

Update frequency

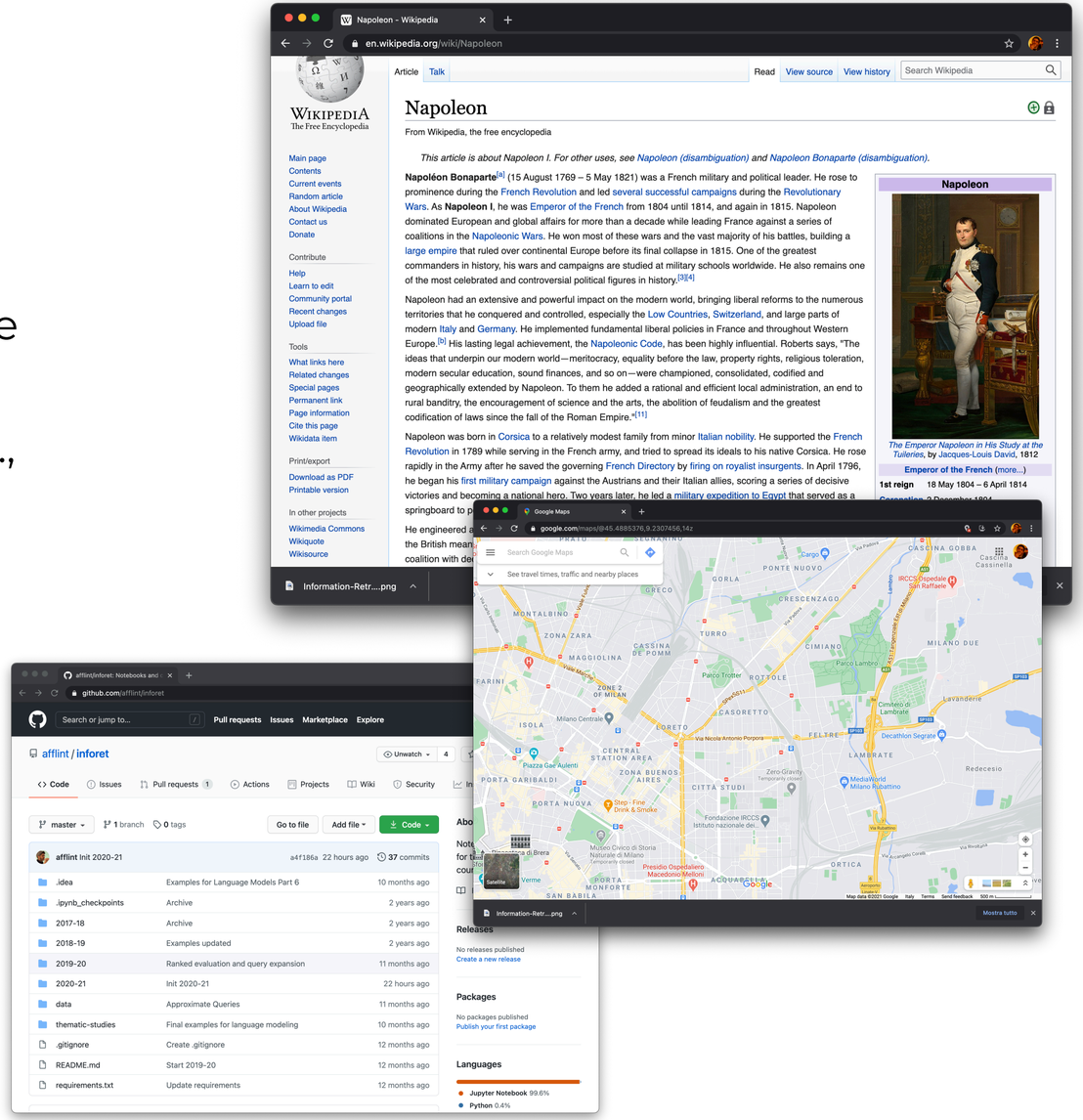
- Stable (i.e., archive)
- Daily update (i.e., newspapers)
- Highly frequent update (i.e., web)
- Continuous update (i.e., sensor data)

Structure

- Structured data (i.e., company database)
- Semi-structured data (i.e., wikipedia)
- Unstructured data (i.e., text document collection)
- Hybrid (i.e., social data)

Datatypes

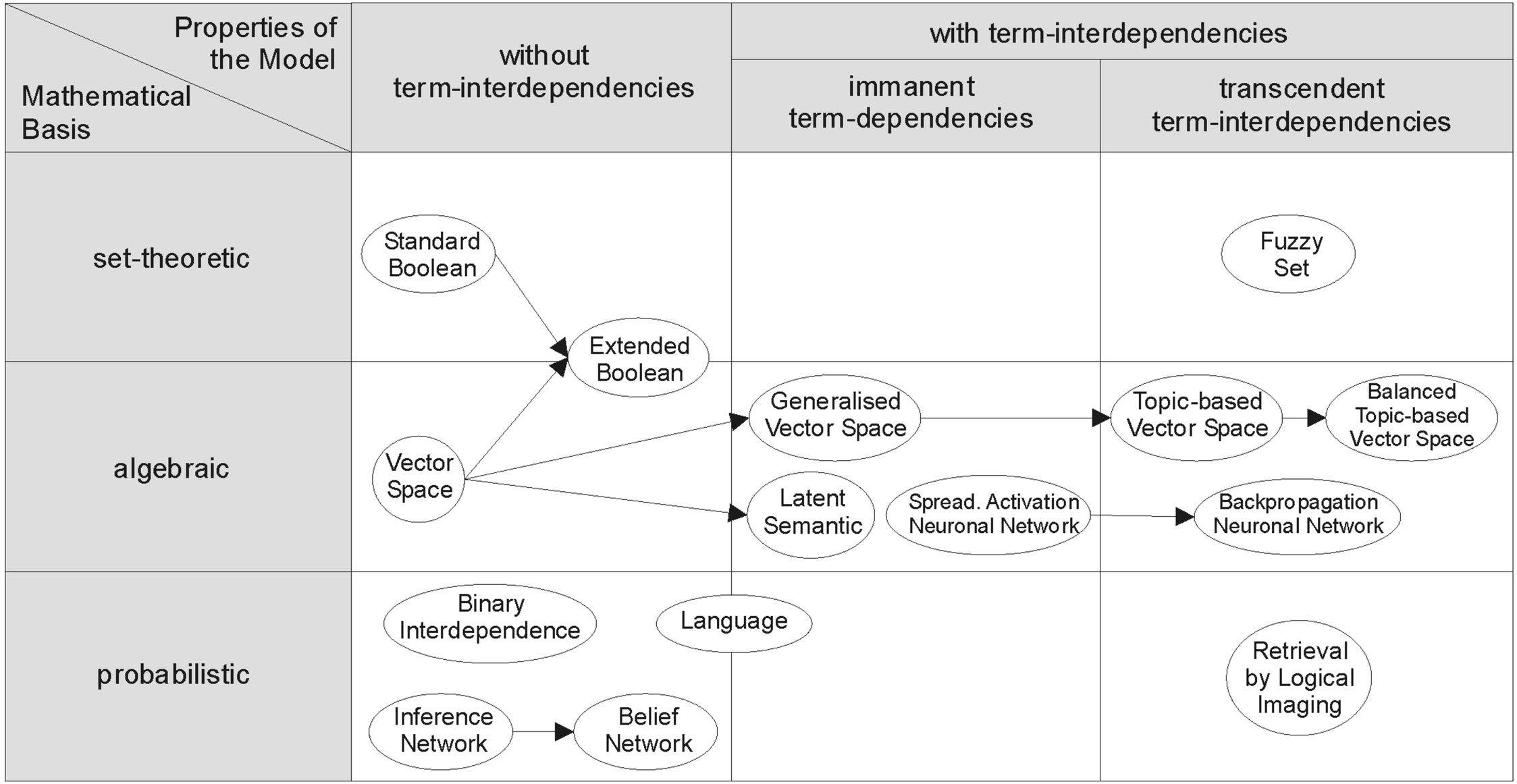
- Text
- Images
- Video
- Geographic data
- ...
- Hybrid



Finding the most suitable task to answer the information need

INFORMATION NEED	TASK	TECHNIQUES
Find documents	Retrieval	NLP
Find topics	Classification	Linear algebra
Find entities	Topic detection	Probabilistic models
Data analytics	Named Entity Recognition	Machine Learning
Spelling correction	Named Entity Resolution	Word Embedding
…	Text generation	Language models
	…	…

Main approaches to information retrieval



BRIDGE THE GAP