

# **PRI Presentation, 21/22 Edition**

---

PRI · Information Processing and Retrieval  
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes  
Dept. Informatics Engineering  
FEUP · U.Porto

# Today's Plan

---

- Information processing and retrieval: context and motivation
- Course presentation: topics, materials, classes, evaluation
- Course projects: groups, themes, rules
- Q&A

# Context and Motivation

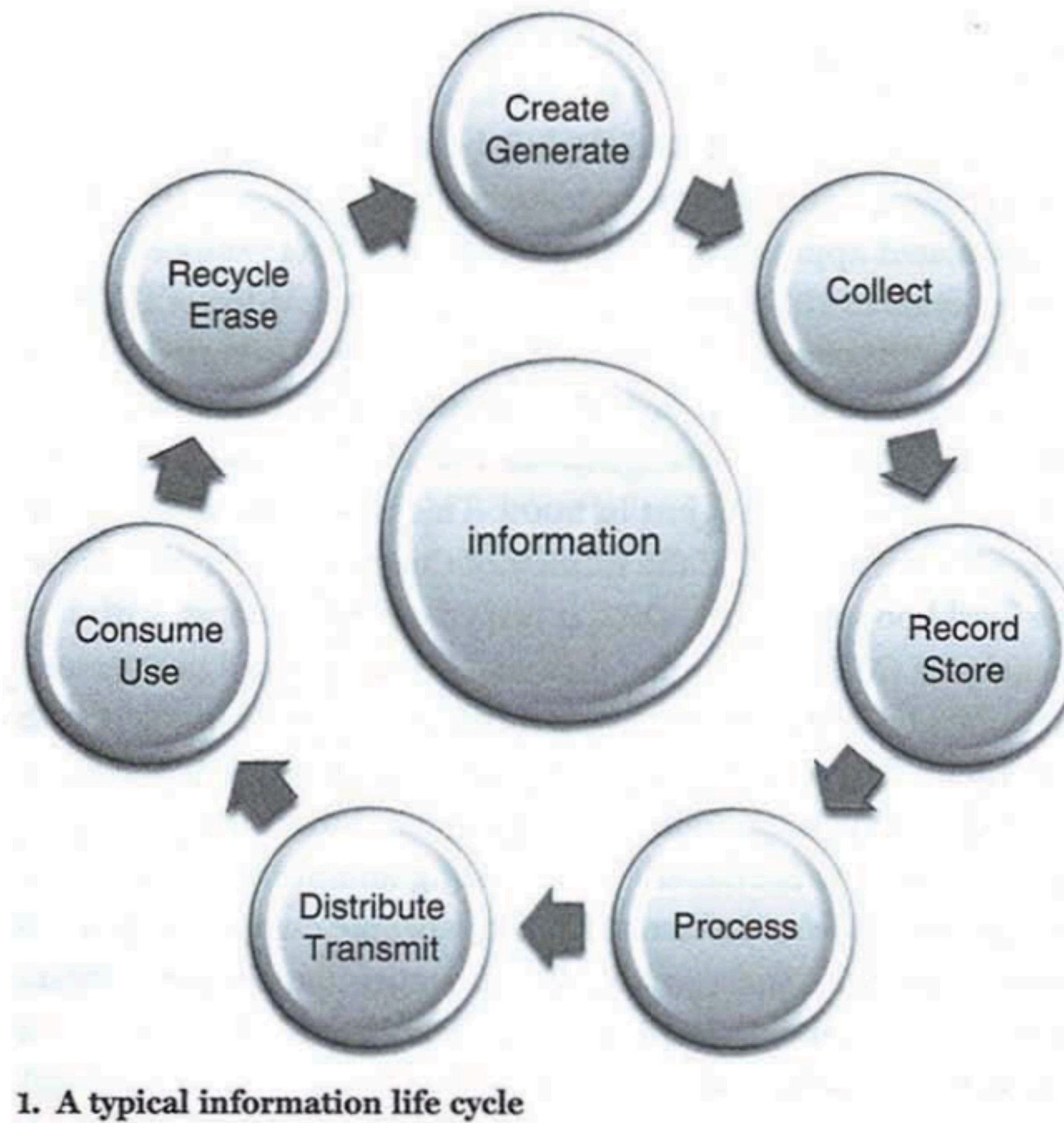
# Information Society

---

- Information communication technologies are ubiquitous in modern societies.
- An ever-growing number of activities depend on the ability to extract value from information.
- Human progress and welfare is largely dependent on an efficient management of the life cycle of information.
- New professional profiles: data engineer, data architect, data analyst, data scientist.
- This course is an introduction to information processing and to information retrieval.

# Information Life Cycle

---



# The Unreasonable Effectiveness of Data

---

- In 2009, Google researchers published a paper highlighting the virtues of data in successfully tackling complex language-related problems.
- *The Unreasonable Effectiveness of Data (2009) [\[link\]](#)*  
*Alon Halevy, Perter Norvig, Fernando Pereira*
  - "[I]nvariably, simple models and a lot of data trump more elaborate models based on less data"
  - "simple  $n$ -gram models or linear classifiers based on millions of specific features perform better than elaborate models that try to discover general rules"

# An Era of Information Abundance

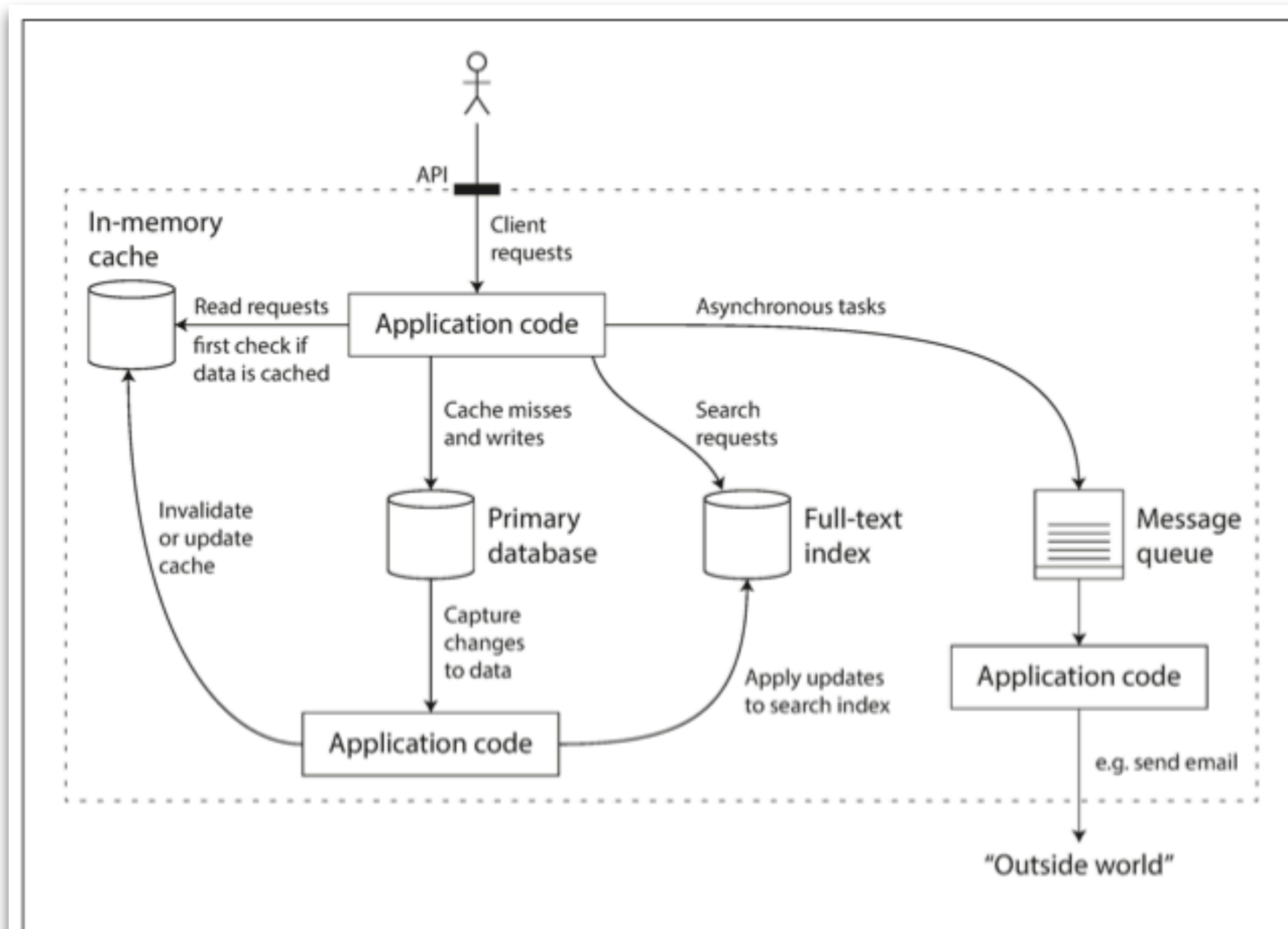
---

- Information consumes attention — leading to attention scarcity
  - *"In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."*  
Herbert A. Simon (1971!)
- Information is complemented by analysis — information growth results in growth in information analysis
  - *"The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."*  
Hal Varian (2009)



# Information Processing and Retrieval Tasks

- Classic workflow and typical tasks.
- Data Ingestion
  - Collect data
  - Describe data
  - Move data
- Data Transformation
  - Data modeling
  - Data migration
  - Pipeline orchestration
- Data Optimization
  - Selection, export, assessment



**Figure 1-1.** One possible architecture for a data system that combines several components.



# Course Presentation

# PRI Team, 21/22 Edition

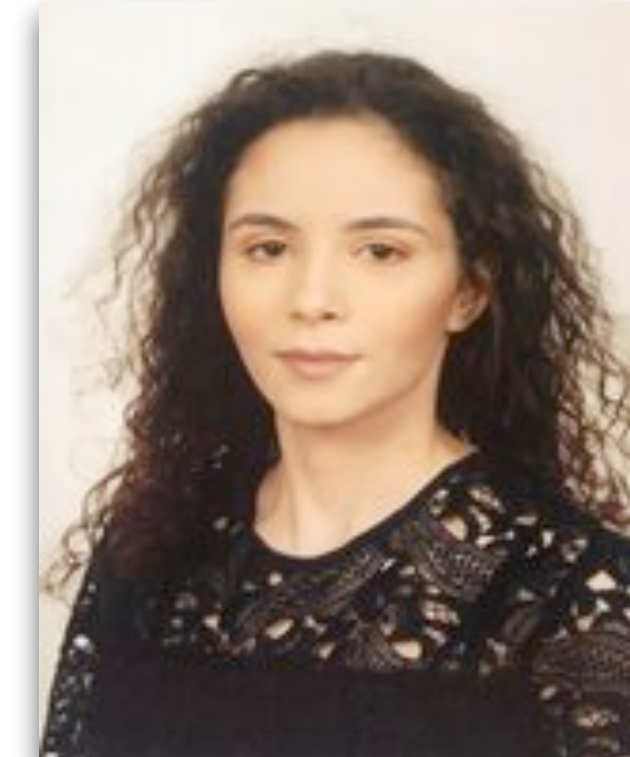
---



Sérgio Nunes  
(regente)



João Damas



Sara Fernandes

# PRI @ M.EIC

---

- This is the first edition of PRI
- M.EIC 1st-year course from the area of Information Systems
- The course evolves from DAPI (Information Description, Storage and Retrieval)
- Significant increase in the number of students, from 1 class (~30 students) to 6 classes (~140 students?)
- Lectures will be on Mondays, 17h00, online.
- Practical classes will be at FEUP, on Tuesdays, Wednesdays and Fridays

# Course Objectives

---

- Prepare students to know, understand, design and develop solutions for information processing and retrieval
  - Make students aware of the challenges associated with building information search systems
  - Familiarize students with the main concepts and techniques associated with information processing and retrieval
  - Enable students to design, implement and evaluate information search systems on document collections

# Learning Outcomes

---

- Identify and describe the main tasks associated with information processing and retrieval
- Describe the architecture and functioning of an information search system
- Describe the tasks associated with the processing phases of a collection (offline) and interrogation processing (online)
- Distinguish the different information retrieval models, identifying their principles, models for document representation, and similarity measures
- Describe and implement different techniques for indexing information
- Describe and implement different techniques for retrieving and ordering results

# Information Processing

---

- Data sources, data provenance and datasets
- Data acquisition and data exploration
- Data pipelines
- Data processing and extraction
- Data characterization

# Information Retrieval

---

- Field of Information Retrieval: history, basic concepts and tools
- Architecture of IR Systems: indexing and retrieval processes
- IR Models: ranking, boolean model, vector space model
- Evaluation in IR: methods and metrics
- Web IR: link analysis, classic algorithms
- Tentative: Entity-Oriented Search, User Interfaces for Search, Applications



# Classes

---

- Lectures, 2h, Mondays at 17h00, online
  - Topic presentation and discussion.
  - Recorded and made available.
- Practical, 2h, in-person
  - Brief presentation of working guides and tutorials
  - Status report with each group
- We plan on having a set of invited talks at the end of the semester (tbd)

# Evaluation

---

- Distributed evaluation with final exam
- Final grade =
  - 60% Group Project +
  - 40% Exam
- Group Project =
  - 20% Data Processing (M1) +
  - 40% Information Retrieval (M2) +
  - 40% Search System (M3)
- The final grade of the project can vary between members of the same group, by plus or minus 3 values, based on the opinion of the teachers and in the self- and hetero- assessment to be carried out internally in each group.
- Minimum grade of 40% (8) required (but not sufficient!) in each of the components.

# Main Bibliography

---

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.  
Introduction to Information Retrieval. Cambridge University Press, 2008  
<https://nlp.stanford.edu/IR-book/information-retrieval-book.html>
- W. Bruce Croft, Donald Metzler, Trevor Strohman.  
Search Engines: Information Retrieval in Practice. Pearson Education, 2015  
<https://ciir.cs.umass.edu/irbook/>
- Kristin Balog. Entity-Oriented Search. Springer, 2018  
<http://link.springer.com/978-3-319-93935-3>
- Martin Kleppmann. Designing Data-Intensive Applications. O'Reilly, 2017
- Jeroen Janssens. Data Science at the Command Line, 2nd Edition, O'Reilly, 2021  
<https://www.datascienceatthecommandline.com/2e/>

Group Project

# Group Project

---

- Design and implementation of an information processing and retrieval system
- The project is developed in groups of 3 students and starts with the selection of a topic and the relevant data sources
- The project is organized in deliveries and partial presentations, which correspond to the project development phases
  - Milestone 1: Information Processing (week of Nov 8th)
  - Milestone 2: Information Retrieval (week of Dec 13th)
  - Milestone 3: Search System (week of Jan 24th)

# Milestones

---

- Each project delivery (milestone) has a corresponding presentation and discussion
- Electronic submissions of the project deliverables are accepted up to 18:00 on the day before the in-class presentation
- Reports are written as short scientific papers, using a two-column format (4 pages max in each delivery). Each report is a self-contained work-in-progress and is based on the previous deliveries
- In the weeks assigned to project presentations, the practical class will be organized in a workshop format, with project presentations and discussions according to an established schedule
- The final project evaluation corresponds to a weighted average of the milestones evaluations.

# M1: Information Processing

---

- The first milestone is achieved with the preparation and characterization of the datasets selected for the project
- Work on these tasks depends on the nature, volume, organization and accessibility of the selected datasets. As a result of this milestone, a well-documented and reproducible pipeline of data processing is expected
  - search repositories for datasets
  - select convenient data subsets
  - assess the authority of the data source and data quality
  - perform exploratory data analysis
  - prepare and document a data processing pipeline
  - characterize the datasets, identifying and describing some of their properties
  - identify the conceptual model for the data domain
  - identify follow-up information needs in the data domain



## M2: Information Retrieval

---

- The second milestone is achieved with the implementation and use of an information retrieval tool on the project datasets and its exploration with free-text queries
- This task makes use of state-of-the-art retrieval tools and involves the view of the datasets as collections of documents, the identification of a document model for indexing, and the design of queries to be executed on the indexed information
  - choose the information retrieval tool (Solr, Elasticsearch, ...)
  - analyze the documents and identify their indexable components
  - use the selected tool to build the indexes
  - use the selected tool to configure and execute the queries
  - demonstrate the indexing and retrieval processes
  - manually evaluate the returned results
  - evaluate the results obtained for the defined information needs

# M3: Search System

---

- The third milestone is achieved with the development of the final version of the search system
- This version is an improvement over the previous milestone, making use of features and techniques with the goal of improving the quality of the search results
- For this milestone, each group is expected to explore innovative approaches and ideas, and will heavily depend on the context and data of each group
- Additionally, an extended evaluation of the results and a comparison with the previous version of the search system is also expected
- Examples of topics to explore include: incorporate new information retrieval algorithms; expand the information available for each document by adding and linking new datasets; work on user interfaces by developing a frontend for the search system


# Project Themes

---

- Project topics are free, but cannot be repeated in the same class
- Need to be approved by the end of the **first practical class (note)**
- Data source(s) must be of unstructured nature and rich in textual data
- Consider your personal interests and motivations
- Many possibilities: education, sports, law, government, media ...

# Project Examples

ant.fe.up.pt/search?q=informação



informação

Todos

Notícias

Cadeiras

Estudantes

Cursos

Pessoal

Salas

Departamentos

Ferramentas de Pesquisa

Qualquer unidade

Qualquer estado

Qualquer curso

Qualquer departamento

Feedback

Licenciatura em Ciência da Informação

Curso

https://sigarra.up.pt/flup/pt/cur\_geral.cur\_view?pv\_ano\_lectivo=2020&...

Faculdade de Engenharia da Universidade do Porto (FEUP) (mais 1)

Áreas Científicas Predominantes: **Ciência da Informação**

Diretores: [Maria Elisa Ramos Morais Cerveira](#)

Mestrado em Ciência da Informação

Curso

https://sigarra.up.pt/flup/pt/cur\_geral.cur\_view?pv\_ano\_lectivo=2020&...

Faculdade de Engenharia da Universidade do Porto (FEUP) (mais 1)

Áreas Científicas Predominantes: **Ciência da Informação**

Diretores: [António Manuel Lucas Soares](#), [Carla Alexandra Teixeira Lopes](#)

Ética da Informação

Cadeira

https://sigarra.up.pt/flup/pt/ucurr\_geral.ficha\_uc\_view?pv\_ocorrencia\_i...

Faculdade de Letras da Universidade do Porto (FLUP)

Cursos Responsáveis: **Licenciatura em Ciência da Informação (CINF)**

Docentes: [Armando Manuel Barreiros Malheiro da Silva](#)

Preservação da Informação

Cadeira

https://sigarra.up.pt/flup/pt/ucurr\_geral.ficha\_uc\_view?pv\_ocorrencia\_i...

Faculdade de Letras da Universidade do Porto (FLUP)

Cursos Responsáveis: **Licenciatura em Ciência da Informação (CINF)**

Docentes: [Maria Manuela Gomes de Azevedo Pinto](#)

Serviços de Informação Empresarial

Cadeira


https://sigarra.up.pt/flup/pt/ucurr\_geral.ficha\_uc\_view?pv\_ocorrencia\_i...

Faculdade de Letras da Universidade do Porto (FLUP)

Cursos Responsáveis: **Licenciatura em Ciência da Informação (CINF)**

Docentes: [Olívia Manuela Marques Pestana](#)

ant.fe.up.pt/search?q=rita



rita

Todos

Estudantes

Notícias

Pessoal

Cadeiras

Salas

Cursos

Ferramentas de Pesquisa


Qualquer unidade

Qualquer estado

Qualquer curso

Qualquer departamento

Feedback



Rita Cerqueira

Estudante


https://sigarra.up.pt/ffup/pt/vld\_entidades\_geral.entidade\_pagina?pct\_...

Faculdade de Farmácia da Universidade do Porto (FFUP)

Curso: **Mestrado Integrado em Ciências Farmacêuticas**

Código: 201103725


Cursos



Curso: **Mestrado Integrado em Ciências Farmacêuticas**

Tipo de Inscrição: Normal

Ano Letivo: 2011/2012



Rita Durães


Estudante

https://sigarra.up.pt/fpceup/pt/vld\_entidades\_geral.entidade\_pagina?p...

Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto (FPCEUP)

Curso: **Luto: Intervenção Psicológica em Diferentes Contextos**

Código: 201812306



Rita Kharchafi


Estudante

https://sigarra.up.pt/faup/pt/vld\_entidades\_geral.entidade\_pagina?pct...

Faculdade de Arquitectura da Universidade do Porto (FAUP)

Curso: **Educação contínua**

Código: 200804347



Rita Királyfy

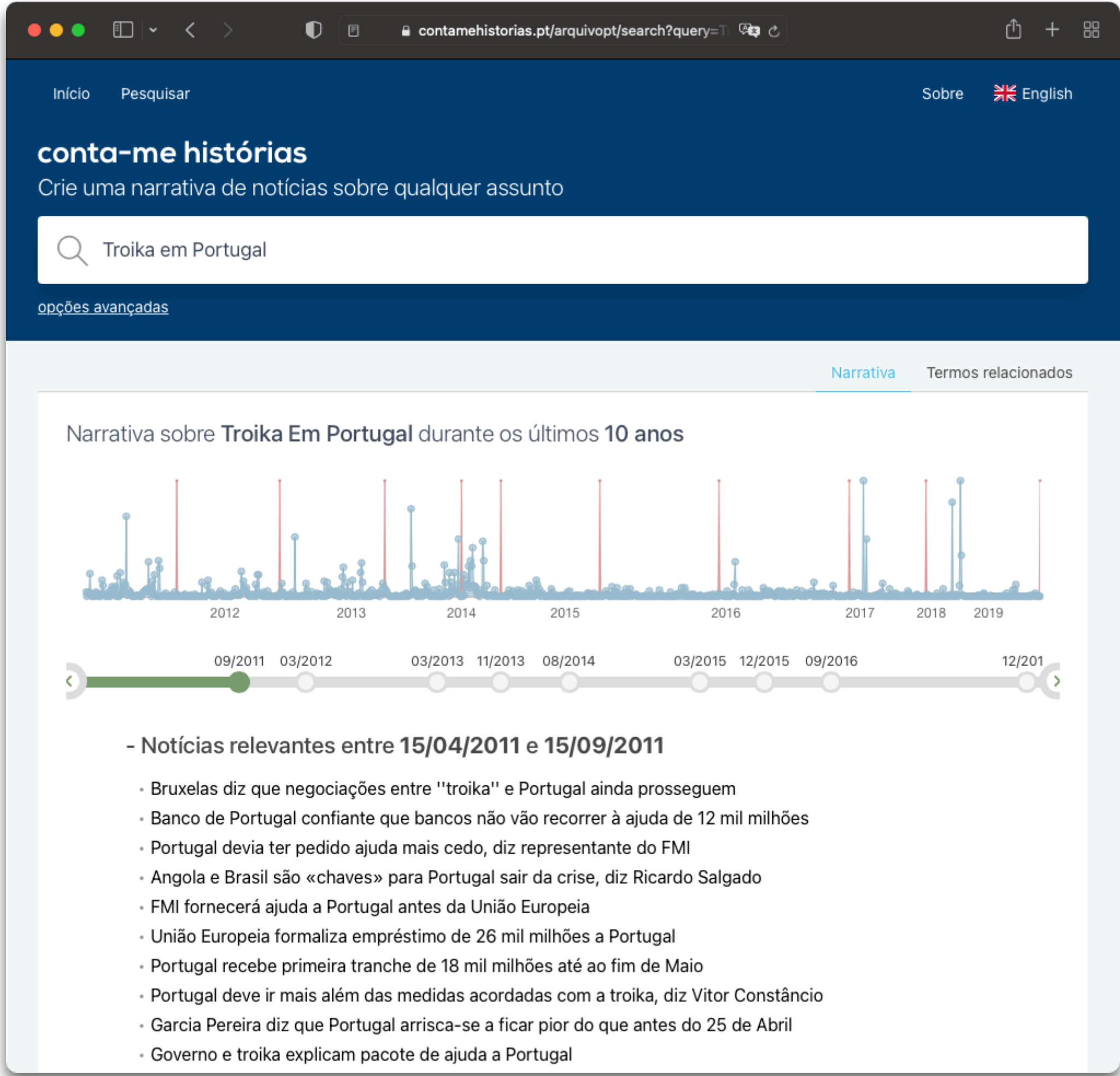
Estudante

https://sigarra.up.pt/fpceup/pt/vld\_entidades\_geral.entidade\_pagina?p...

Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto (FPCEUP)

27

# Conta-me Histórias



# Data Sources

---

- A diversity of data sources exist.
- Keep in mind when choosing sources that:
  - Direct access, i.e. not dependent on third party actions
  - Volume, ideally thousands
  - Rich in textual data, i.e. long texts, not just labels or titles
- <https://web.fe.up.pt/~ssn/wiki/teach/pri/202122/datasets>



# Materials

---

- The course's web page is the starting point:
  - <https://web.fe.up.pt/~ssn/wiki/teach/pri>
  - For each lecture and lab class an information page is available
- Moodle is used for:
  - Group registration
  - Announcements and discussion
  - Submission of milestone materials (report and presentation)
- Slack:
  - Last minute warnings and in-group communication

# PRI Tutorials

---

- You have access to a collection of tutorials to guide to introduce concepts and tools, and guide your explorations during the semester.
- <https://git.fe.up.pt/pri/tutorials>
- Start this week with the first tutorial — Command Line Practice
- *A work in progress. Your feedback is welcomed.*

# Next steps

---

- Answer 'PRI Survey' (if you haven't done so)
- Read the project rules
- Prepare for the first practical class:
  - organize groups before class (3 students) — register in Moodle (you can change later)
  - explore data sources and contexts to identify datasets
- Explore the first PRI tutorial — Command Line
- First delivery in four weeks (15th November week) - Milestone 1: Data Processing.

Questions or comments?