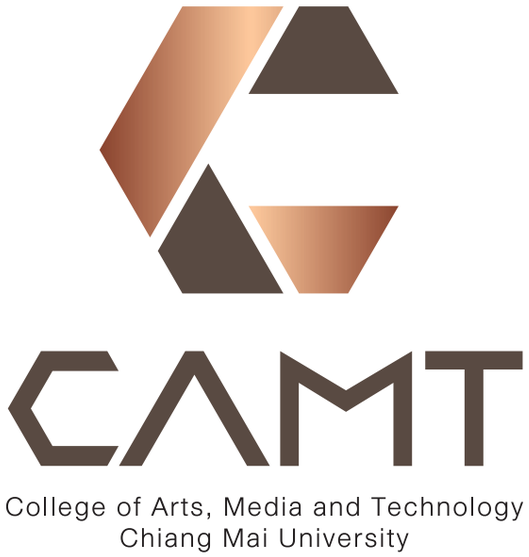
SE 481 Introduction to Information Retrieval (IR for SE)

Module #0 — Homeroom



Passakorn Phannachitta, D.Eng.

passakorn.p@cmu.ac.th

College of Arts, Media and Technology

Chiang Mai University, Chiangmai, Thailand

Prerequisite (SE program)

- SE 201 (953201) — Algorithms Design and Analysis

Know your lecturer

Passakorn Phannachitta, D.Eng. (Aj. Kong)

Office: CAMT 417

Email: passakorn.p@cmu.ac.th

What will be in the course ?

- Fundamental of IR
- Indexing
- IR Models
- Spell correction using IR
- Evaluation
- Web search (Search engine)
- IR and some machine-learning applications
- Add ons

Schedule

- Lectures & Hand ons

November 13	Course introduciton
November 20	Overview
November 27	Index
December 4	Index
December 11 (We have class on this holiday)	IR Models
December 18	IR Models
December 25	Spell correction
Long holiday & Midterm exam weeks	
January 22	Evaluation
January 29	Web search
February 5	Web search
February 12	IR & Machine learning
February 19	IR & Machine learning
February 26	Holiday — No class
March 4	Add ons & Wrap up
Final exam weeks	

Grading - based on criteria

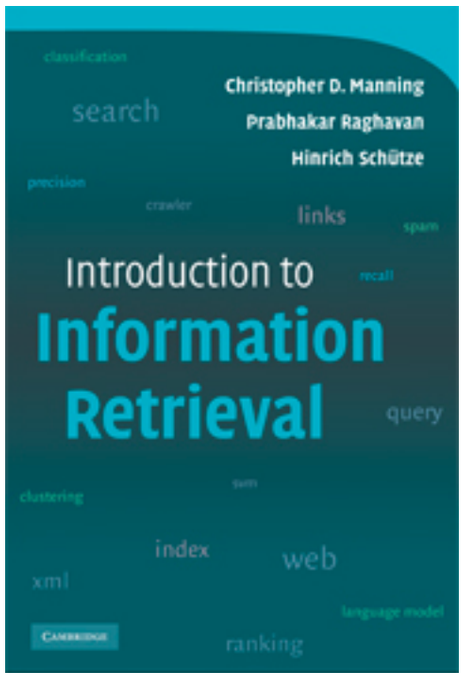
- Hands on 20%
- Many assignments 20%
- Project 20%
- Exam 40%

- **A** is **85** to 100, and **F** is 0 to less than **55**

Grading - based on criteria

A	**[85, 100]**
B+	[80, 85)
B	[75, 80)
C+	[70, 75)
C	[65, 70)
D+	[60, 65)
D	[55, 60)
F	**[0, 55)**

Main reference



https://nlp.stanford.edu/IR-book/

Class communication

- MS TEAM
- SCOTT

Class policy

- Please bring your laptop to every class
- No late submission for everything.

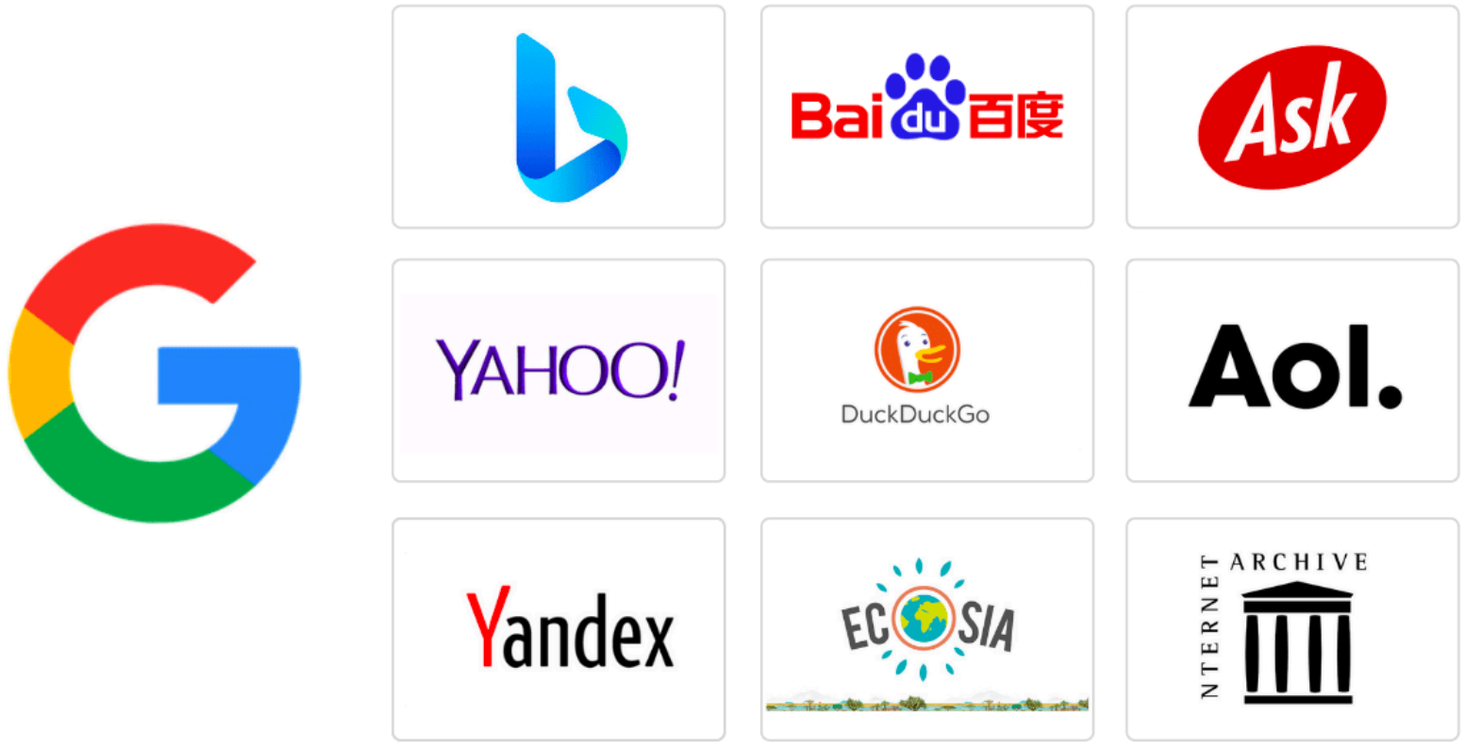
Information Retrieval (IR)

- **Information retrieval** is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

— Wikipedia

Information Retrieval (IR)

- Mostly we think of an IR system as web search engines



Ref: https://www.reliablesoft.net/wp-content/uploads/2016/12/top-search-engines-oct-2020.png

Information Retrieval (IR)

- While often associated with web search engines, IR systems are prevalent in various other contexts.
  - E-mail search functionality for locating messages and attachments.
  - File search utilities within operating systems for organizing and retrieving documents.
  - And many other domain-specific search tools.

Basic assumption of IR

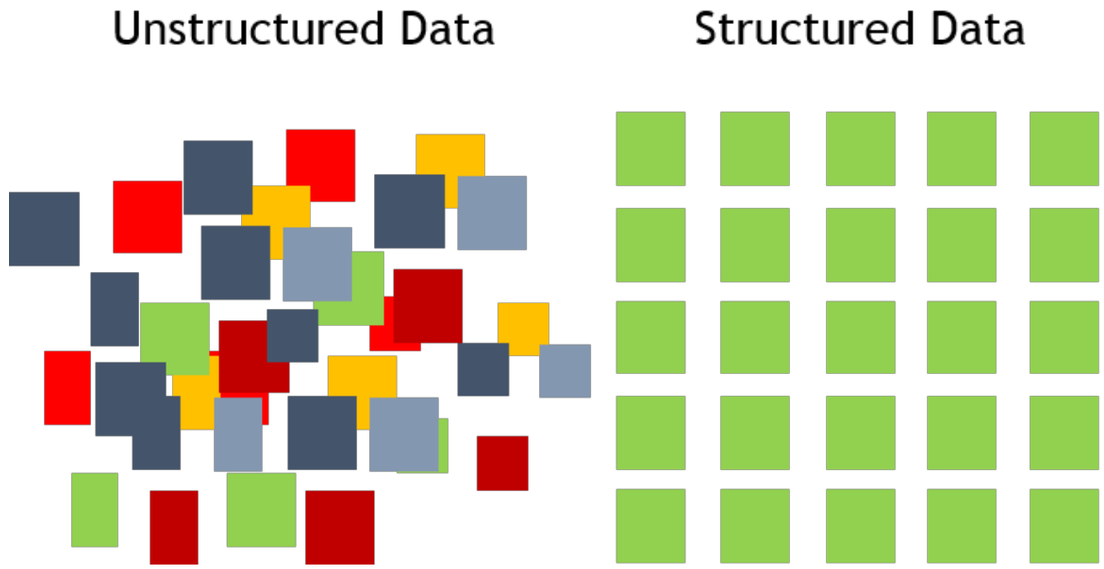
- Assumes that information is encapsulated within a collection of documents.

- An IR system's core function is to fetch documents that **best satisfy** the user's **specific information** requirements, thereby aiding in **task completion**.

- Recognizes that data inherently exists in an **unstructured format**, necessitating sophisticated processing for meaningful retrieval.

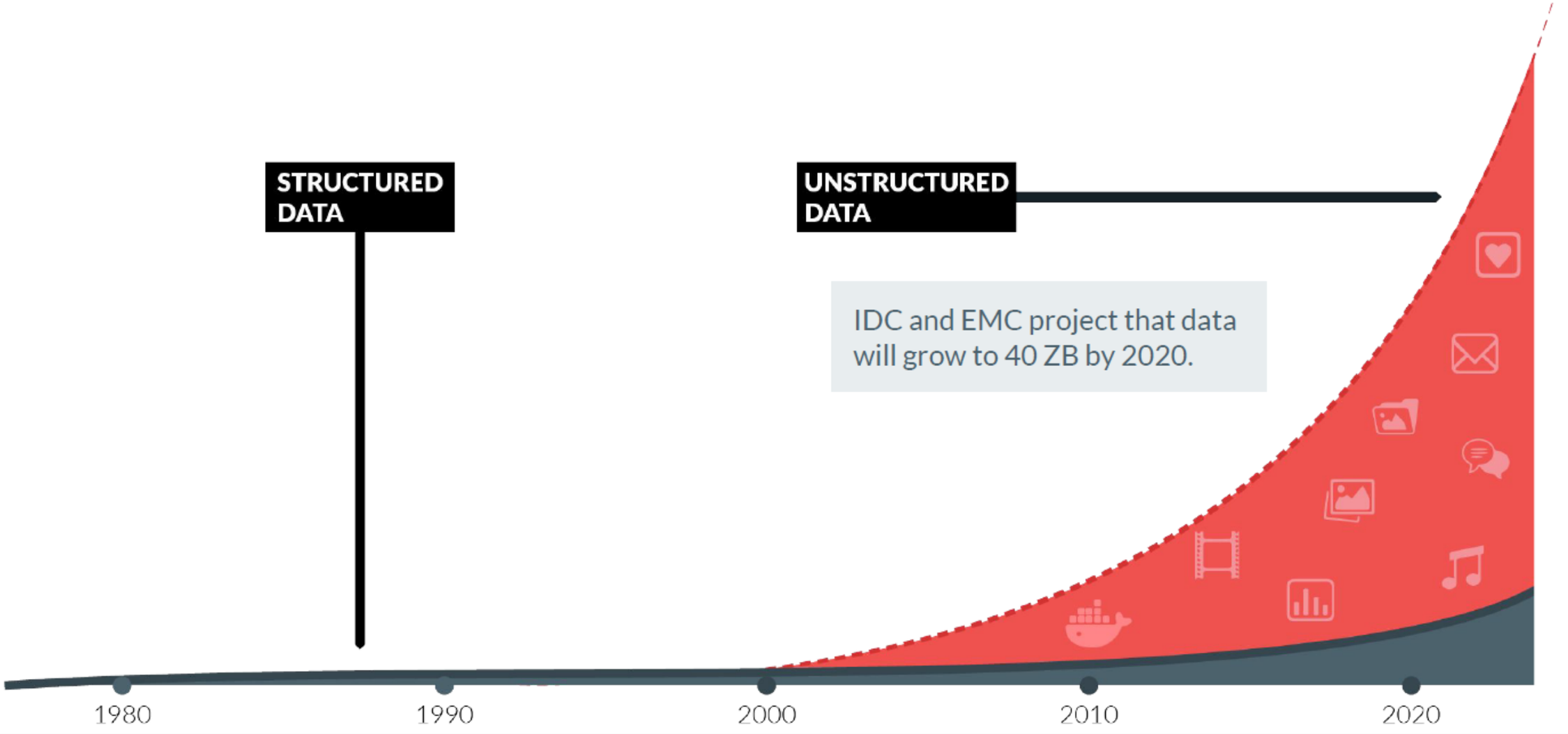
Unstructured data

- Structured data
  - Characterized by a specific format that facilitates efficient search, linkage, and comparison due to its clear type definition and pattern.



- Unstructured data
  - Lacks a predefined data model or schema, which makes it more challenging to organize and analyze without additional processing.

Unstructured data is everywhere



Information Retrieval (IR)

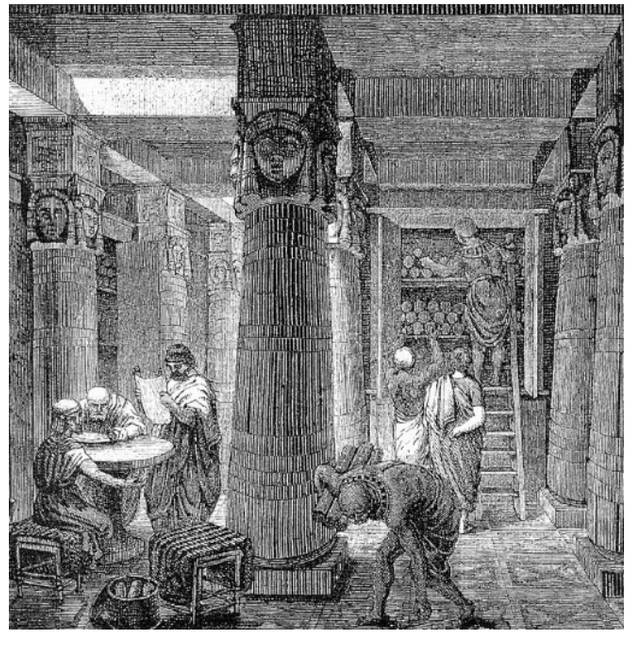
- Google's mission statement

Our mission is to organise the world's information and make it universally accessible and useful.

Ref: https://about.google/

Organizing a large amount of information

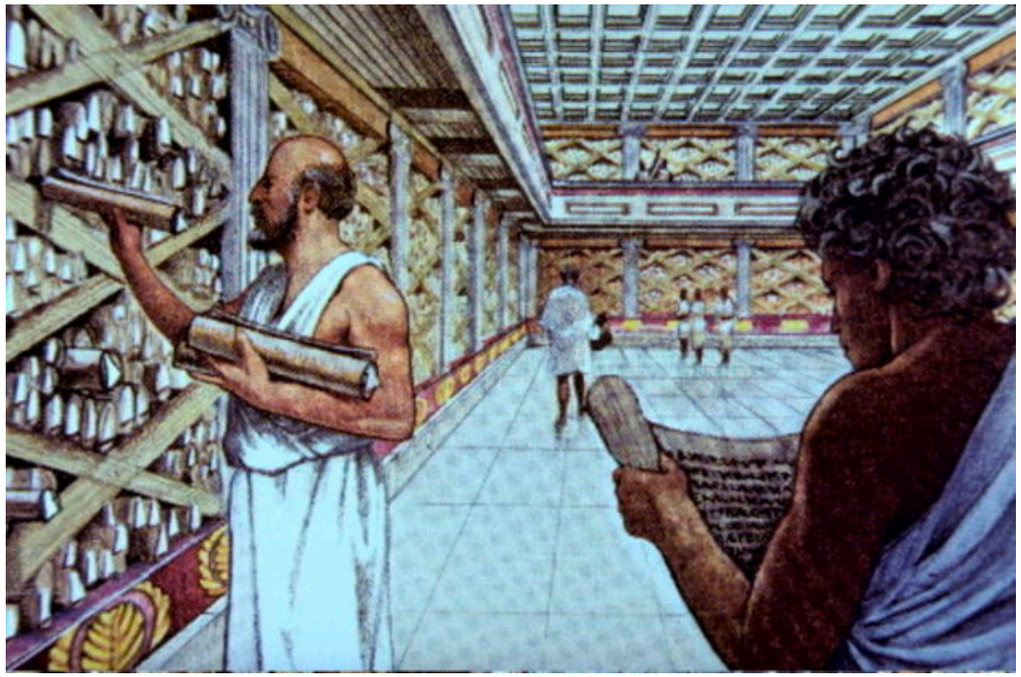
- The very first form might be (physical) libraries
- E.g., the Library of Alexandria — the largest libraries in ancient world
  - Built ~300 BC
  - Collected numerous papyrus scrolls



Ref: https://en.wikipedia.org/wiki/Library_of_Alexandria

Searching for a piece of information

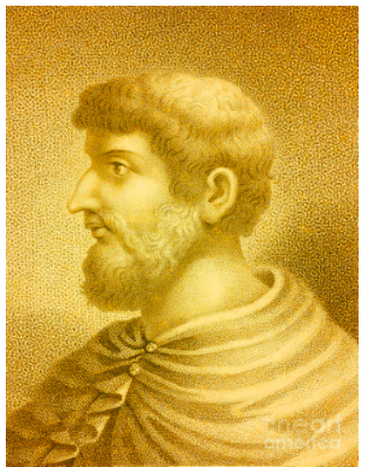
- Suppose there are thousands of scrolls, how can we find the most relevant one for the information we are seeking for?



Ref: https://thesomathread.com/2016/09/08/the-role-and-fate-of-the-library-at-alexandria/

A very first attempt

- **Callimachus**, a scholar at the Library of Alexandria made the tool called **Pinakes.**
- **Pinakes** were considered as the earliest library catalog.
- Works are divided in genres and categories, e.g., law, poetry, history, medicine, and mathematics.
- The **Pinakes** proved indispensable to librarians for centuries, and they became a model for organizing knowledge throughout the Mediterranean



Ref: https://www.geni.com/people/Callimachus/6000000078663146846

A librarian's standard

- Local variations for cataloging and library classification continued throughout the late 1800s,
- when Anthony Panizzi and Melvil Dewey proposed a more standard approach called **Decimal Classification**.

- E.g., 500 — Natural sciences and mathematics
  - 510 — Mathematics
    - 516 — Geometry
      - 516.3 — Analytic geometries

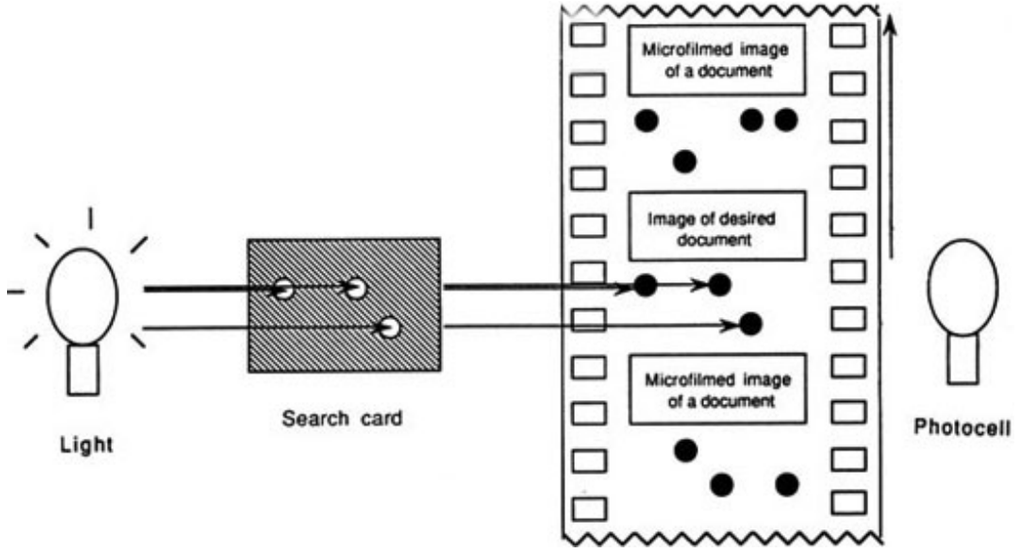
What we have learned

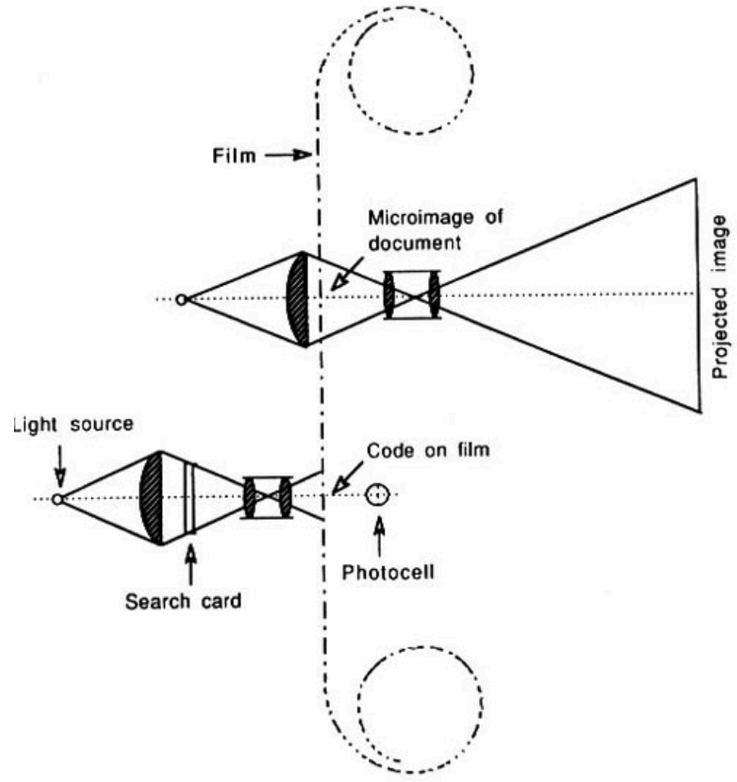
- Categorization can help us narrowing the search scope.
- However, to precisely locate the information in content-level is still far from reality.

- Categorization is made bottom up.
- What about doing the bottom up from the content level?

The very first IR machines

- Emanuel Goldberg's Statistical machine (1931)





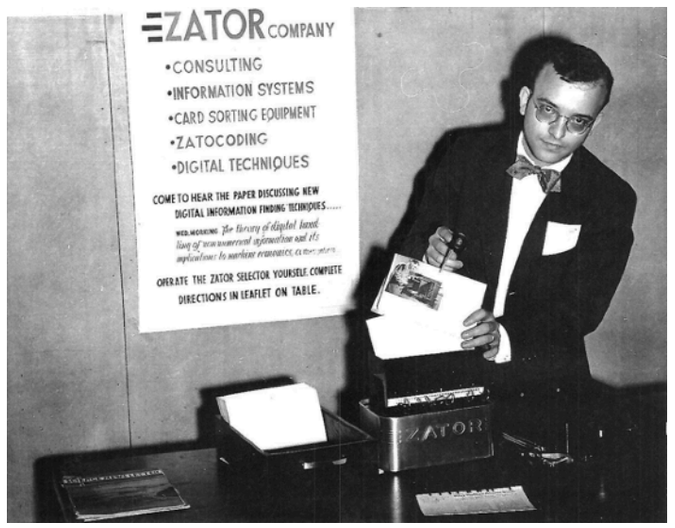
Ref: https://history-computer.com/emanuel-goldberg/

Goldberg’s machine

- Labelling and indexing

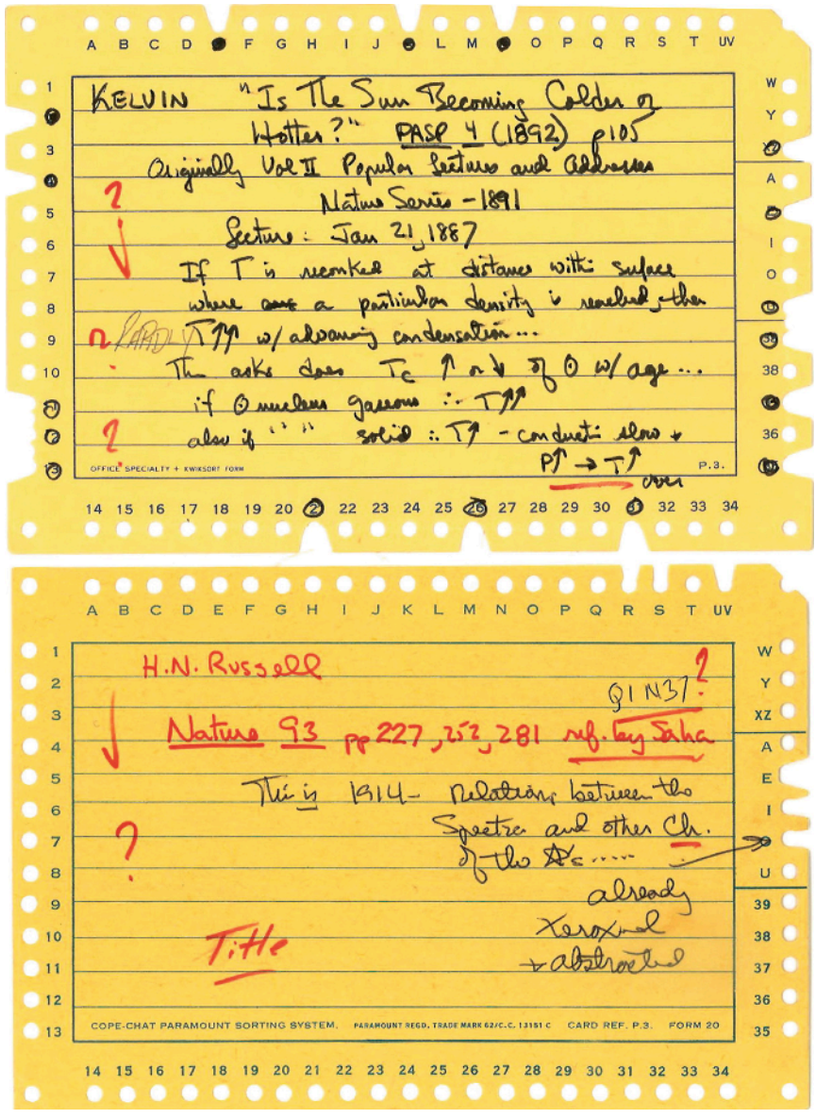
Coining the term “Information Retrieval”

- Calvin Mooers
- At MIT, Mooers developed a mechanical system using superimposed codes of descriptors for information retrieval called **Zatocoding**.
- Mooers founded the Zator Company in 1947 to market this idea, and pursued work in information theory, information retrieval, and artificial intelligence.
- Mooers coined the term "information retrieval" using it first in a conference paper presented in 1950.



Ceruzzi P.E. (2019) Calvin Mooers, Zatocoding, and Early Research on Information Retrieval. In: Haigh T. (eds) Exploring the Early Digital. History of Computing. Springer

Zatocoding



Ceruzzi P.E. (2019) Calvin Mooers, Zatocoding, and Early Research on Information Retrieval. In: Haigh T. (eds) Exploring the Early Digital. History of Computing. Springer

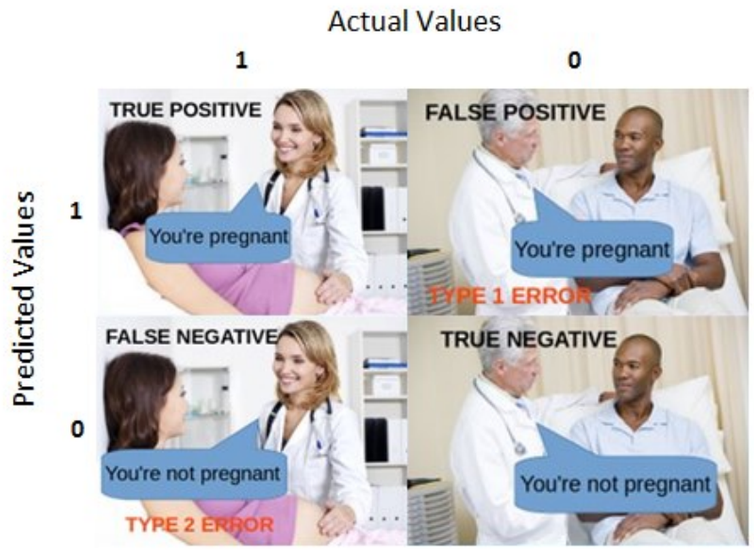
Uniterms

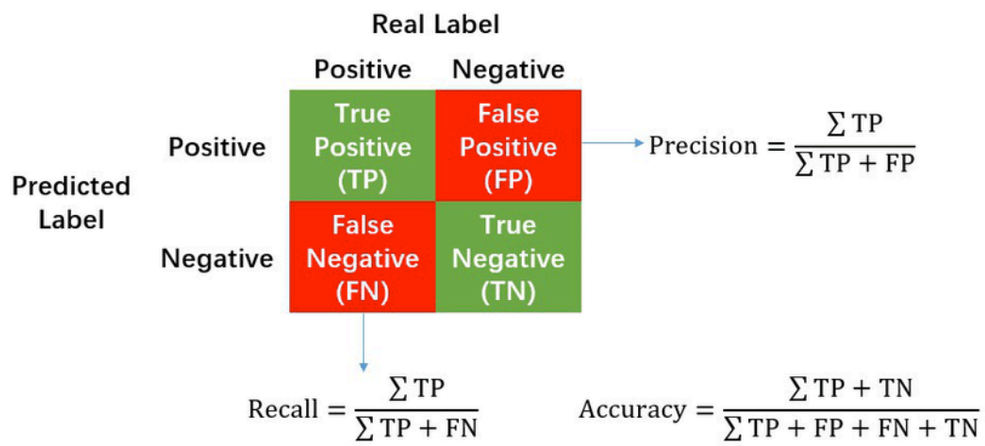
- Mortimer Taube (1952)
- 100 most important leaders in Library and Information Science of the 20th century
- Taube invented **Coordinate Indexing**, which uses **uniterms** in the context of cataloging
- Index card
- E.g., a document on “barefoot running sneakers” might be filed under "barefoot" but perhaps not "sneakers" which would be found on too many documents.

Ref: https://en.wikipedia.org/wiki/Uniterm

Evaluation in IR

- Cyril Cleverdon (1960s)
- Precision and Recall

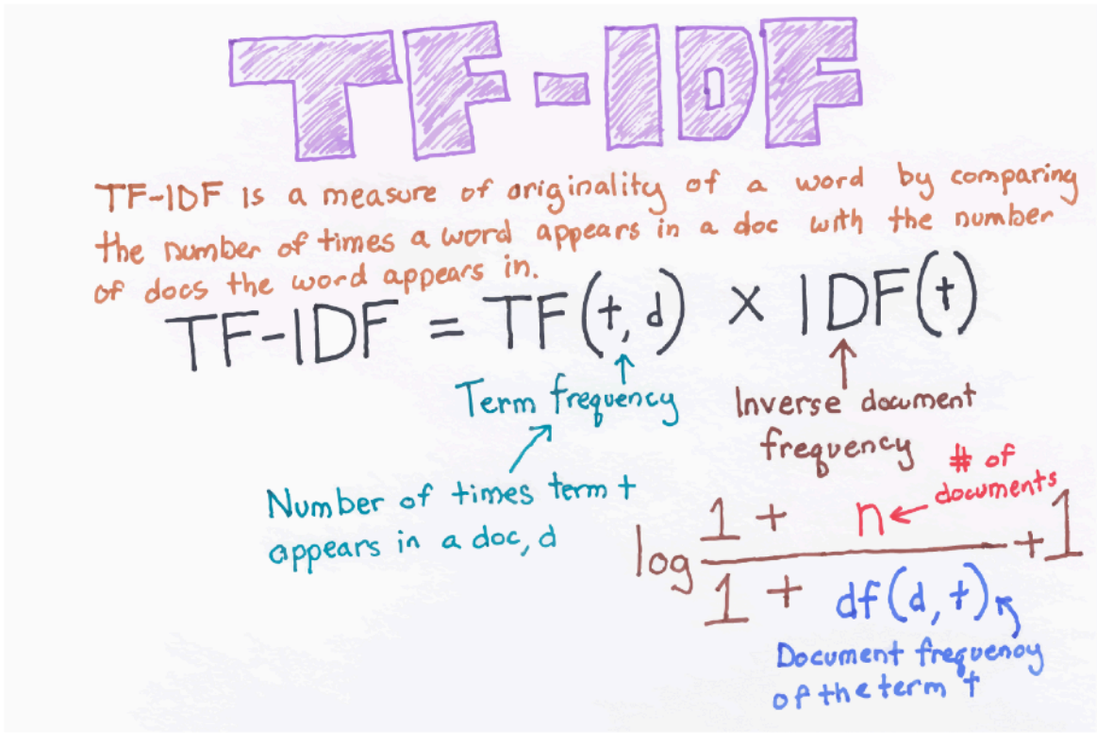




Ref: https://www.bualabs.com/archives/1968/
what-is-confusion-matrix-what-is-metrics-accuracy-precision-recall-f1-score-difference-metrics-ep-1/

IR and ranking

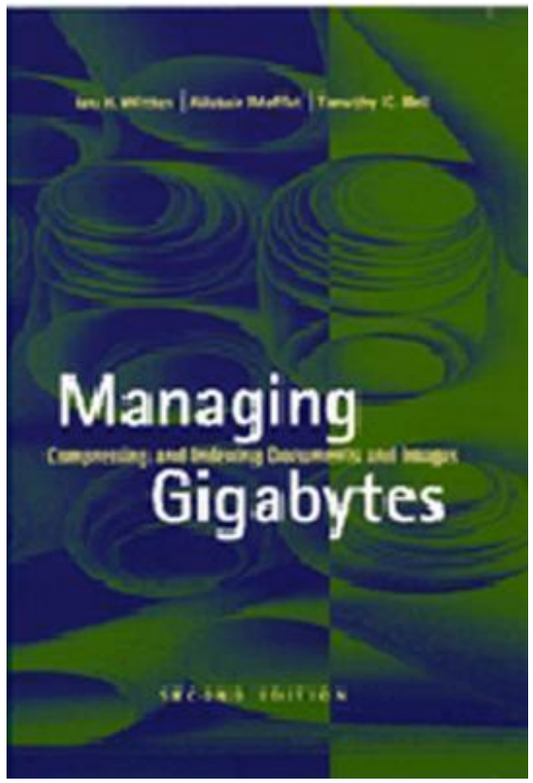
- Hans Peter Luhn (1957) —> term frequencies (tf)
- Karen Sparck-Jones (1972) —> inverse document frequency (idf)
- Gerard Salton (1975) —> tf x idf



Ref: https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

IR and compression

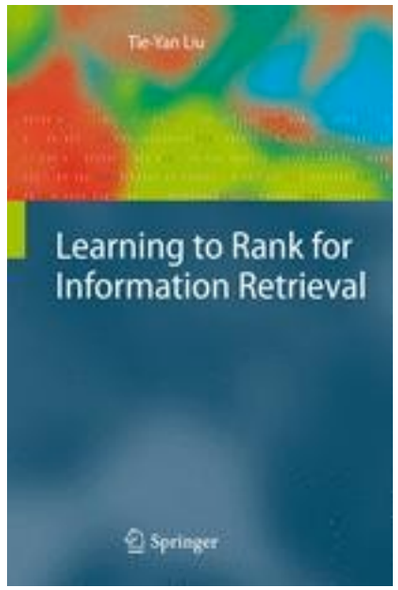
- Ian Witten, Alistair Moffat, and Timothy Bell (1994)



Ref: Witten, I. H., Witten, I. H., Moffat, A., Bell, T. C., Bell, T. C., & Bell, T. C. (1994). Managing gigabytes: compressing and indexing documents and images. Morgan Kaufmann.

Ranking with models

- Stephen Robertson (1994) —> BM25
- Bruce Croft (1998) —> Language model
- Sergey Brin and Larry Page (1998) —> PageRank
- Most recent technology —> Learning to Rank

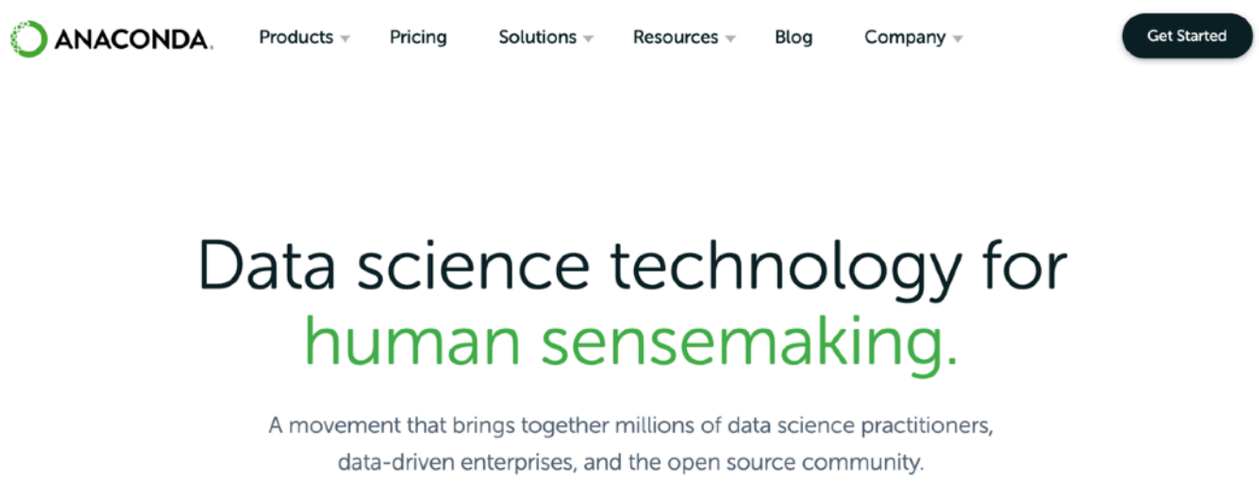


Ref: Liu, T. Y. (2011). Learning to rank for information retrieval.

First hand on

- Install Anaconda or any equivalent in your computer

  https://www.anaconda.com/



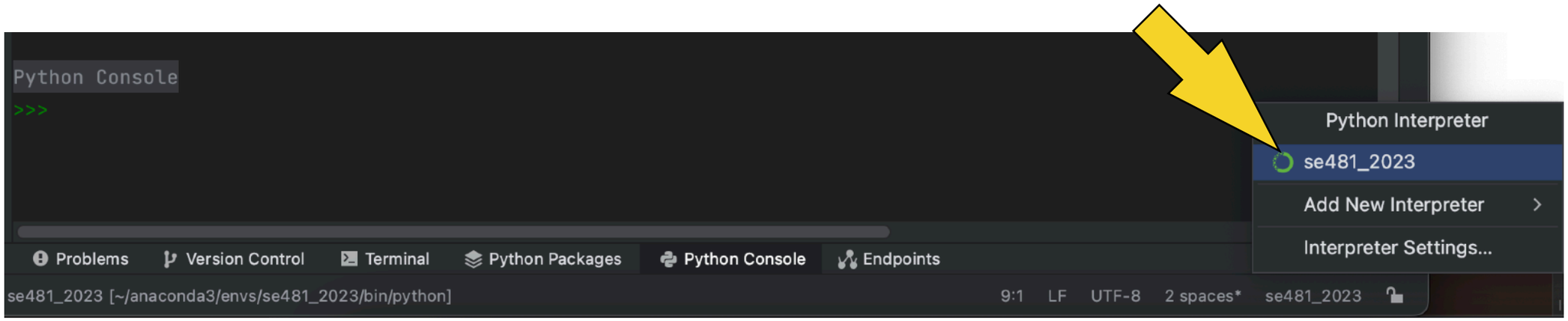
- And an IDE. (the lecturer prefers PyCharm)

  https://www.jetbrains.com/pycharm/

- Show that your IDE is connected to Anaconda

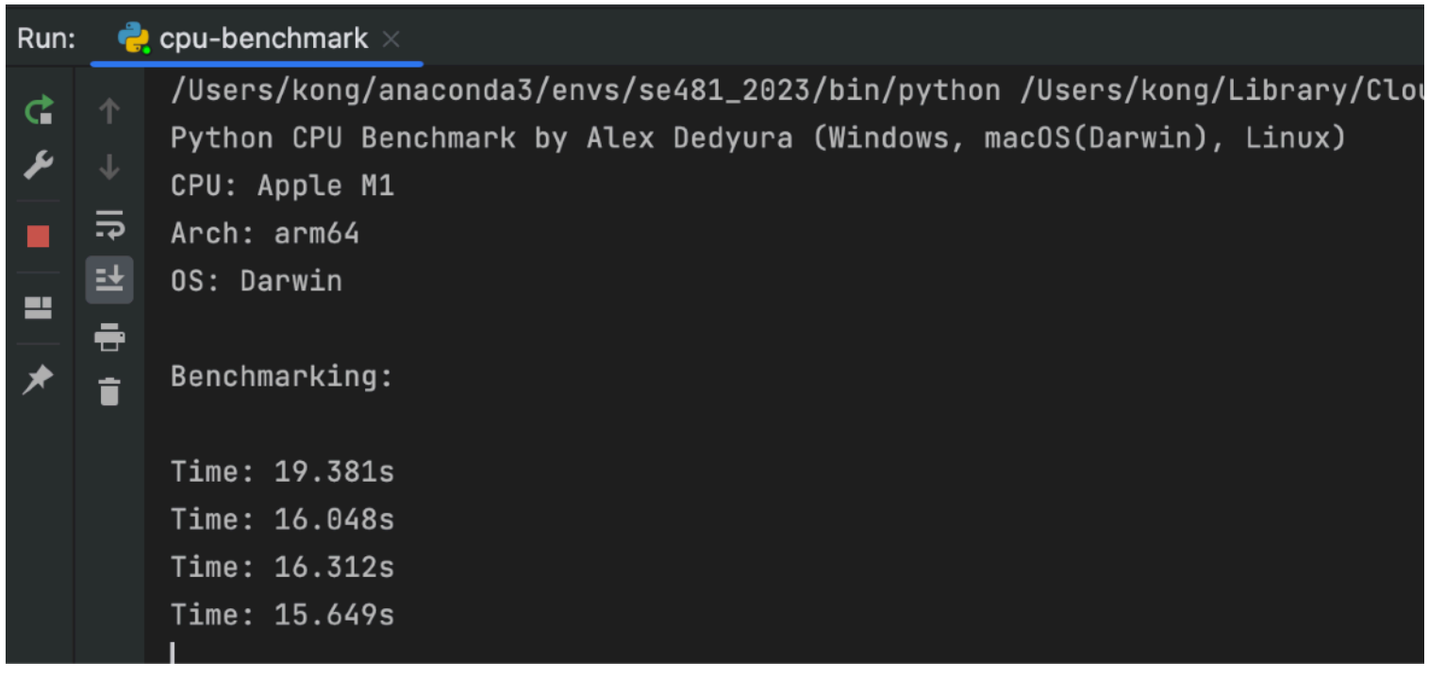
First hand on

- Show that your IDE is connected with your python interpreter, e.g.,



First hand on

- Show that you are able to configure https://github.com/alexdedyura/cpu-benchmark?ref=pythonrepo.com,
  and run it



Time for questions