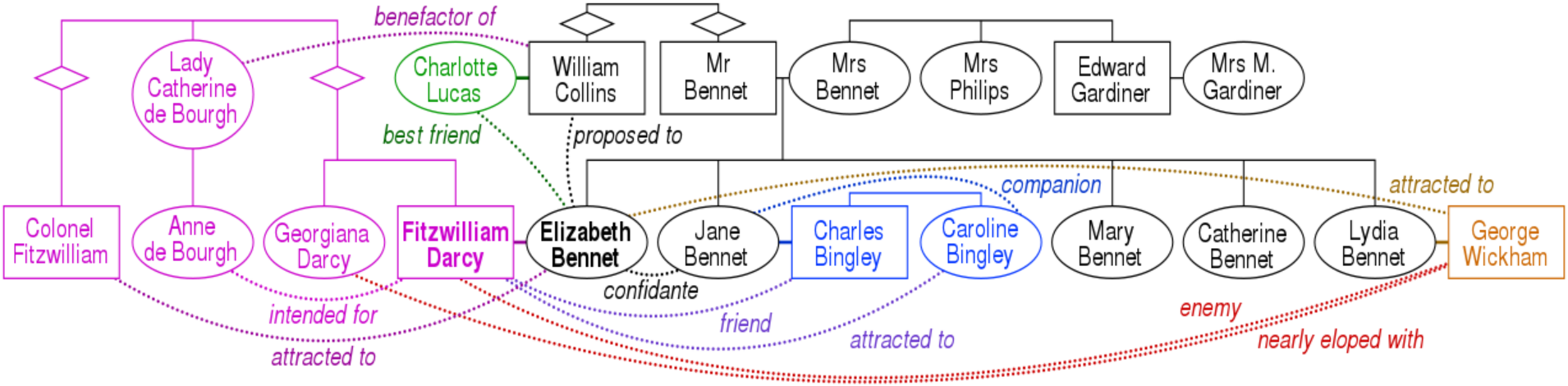
Natural Language Processing

10: Information Extraction – Data Programming

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2022-06-03

Technology
Arts Sciences
TH Köln



parent(Mr. Bennet, Jane Bennet)

Named entity recognition

[tim cook]$_{PER}$ is the ceo of [apple]$_{ORG}$

- Identifying spans of text that correspond to typed entities

Main entities according to ACE (Automatic Content Extraction):

- Person (PER)
- Organization (ORG)
- Geo-political Entity (GPE)
- Location (LOC)
- Facility (FAC)
- Vehicle (VEH)
- Weapon (WEA)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf

Relation extraction

The Big Sleep (1946 film)

From Wikipedia, the free encyclopedia

The Big Sleep is a 1946 American film noir directed by Howard Hawks,[2][3] the first film version of the 1939 novel of the same name by Raymond Chandler. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results".[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.[5][6]

subject	predicate	object
The Big Sleep	directed_by	Howard Hawks
The Big Sleep	stars	Humphrey Bogart
The Big Sleep	stars	Lauren Bacall

Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

Hearst patterns

pattern	sentence
NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP,}* {(or|and)} NP	European countries, especially France, England, and Spain

NP_H is the parent or hyponym

Hearst (1992): Automatic Acquisition of Hyponyms from Large Text Corpora - https://www.aclweb.org/anthology/C92-2082.pdf

Distant supervision

Start with seed relation: chancelor(Angela Merkel, Germany) and search in a large dataset (like the Web or Wikipedia)

- **Angela Merkel** (née Kasner; born 17 July 1954) is a **German** politician serving as the **chancellor** of **Germany** since 2005.
- **Chancellor Angela Merkel**, who hosted the online meeting, pledged further ...
- Biography of **German** politician **Angela Merkel**, who in 2005 became the first female **chancellor** of **Germany**.
- Annalena Baerbock, Greens ... The only woman in the race to succeed **Angela Merkel**, she is the Greens' first ever candidate for **chancellor**, as ...

Wikipedia Infoboxes

The Big Sleep (1946 film)

From Wikipedia, the free encyclopedia

The Big Sleep is a 1946 American film noir directed by Howard Hawks,[2][3] the first film version of the 1939 novel of the same name by Raymond Chandler. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results".[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.[5][6]

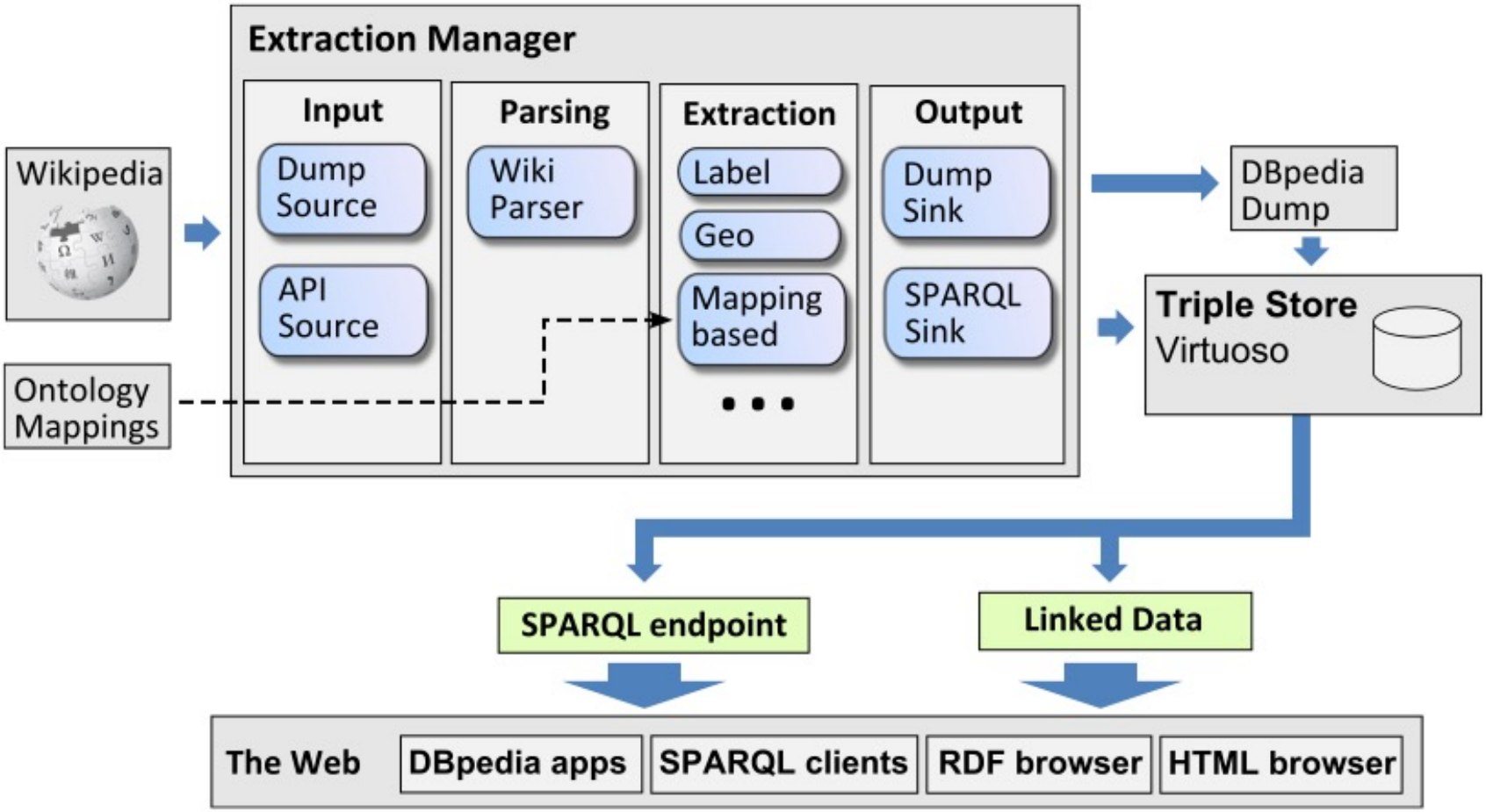
The Big Sleep

Theatrical release lobby card

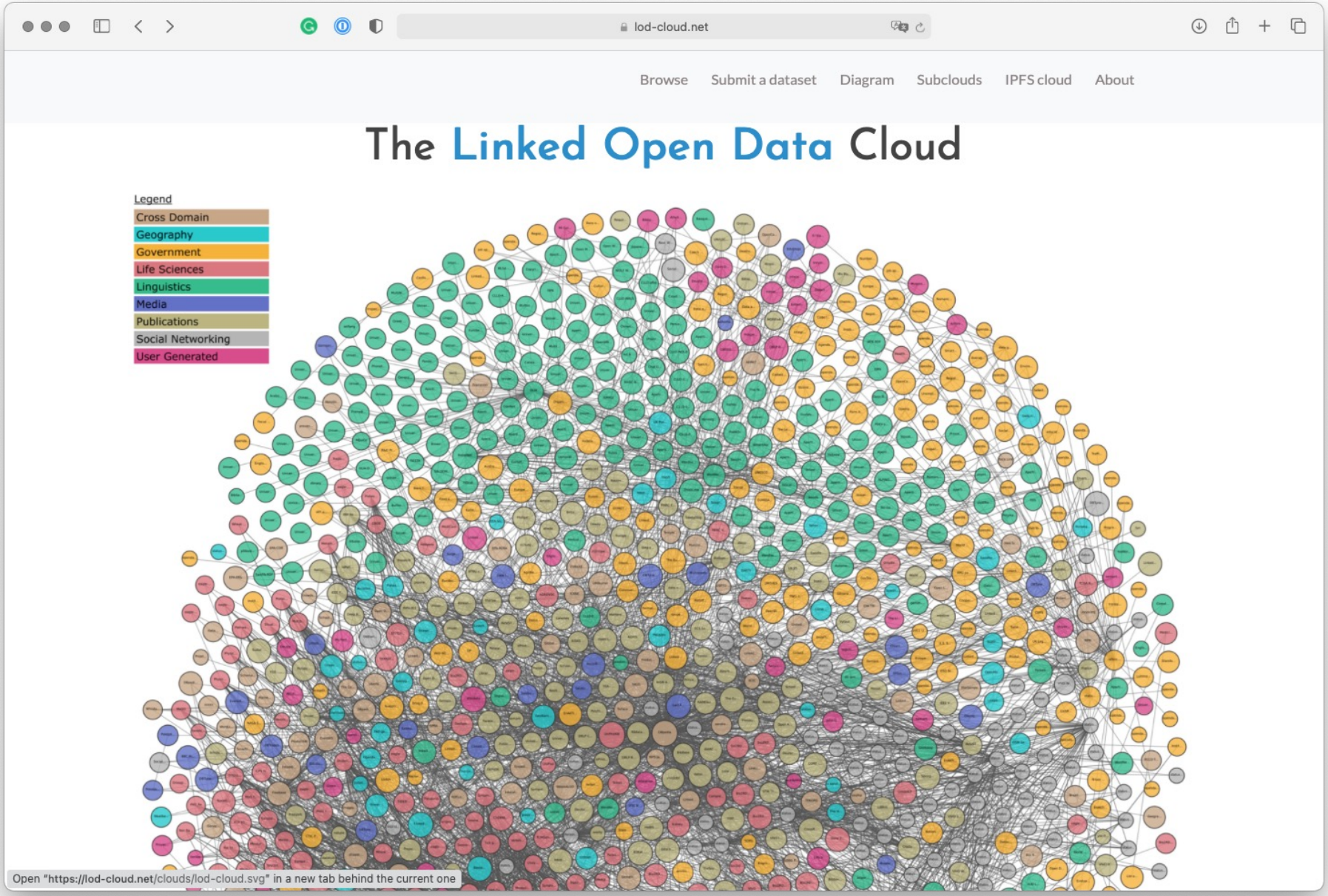
Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	*The Big Sleep* by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall Martha Vickers Dorothy Malone
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby

https://en.wikipedia.org/wiki/The_Big_Sleep_(1946_film)

DBpedia



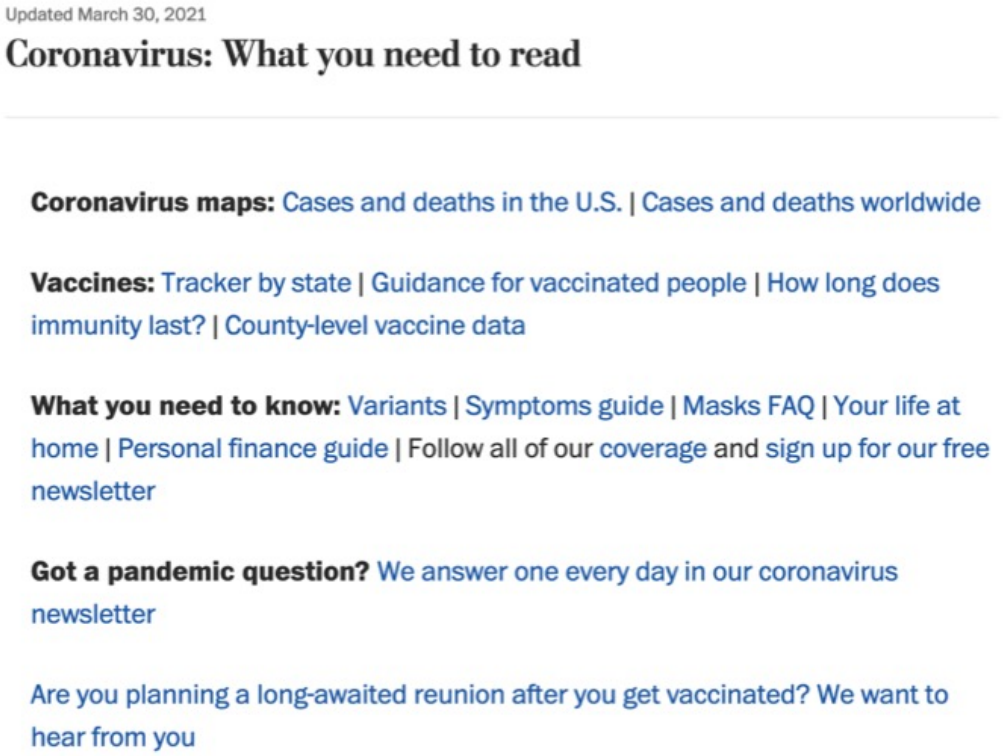
Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia



https://lod-cloud.net

Fresh from the Research Front

Background linkling



Updated March 30, 2021

Coronavirus: What you need to read

Coronavirus maps: Cases and deaths in the U.S. | Cases and deaths worldwide

Vaccines: Tracker by state | Guidance for vaccinated people | How long does immunity last? | County-level vaccine data

What you need to know: Variants | Symptoms guide | Masks FAQ | Your life at home | Personal finance guide | Follow all of our coverage and sign up for our free newsletter

Got a pandemic question? We answer one every day in our coronavirus newsletter

Are you planning a long-awaited reunion after you get vaccinated? We want to hear from you

The Washington Post implements manually linked "explainer boxes". This explainer box appeared at the end of most articles about COVID during the pandemic.

http://trec-news.org/guidelines-2021.pdf

Our approach

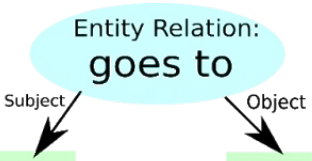
- Large corpus makes costly NLP techniques for all documents infeasible.
  - Use index and TF-IDF to select (approx. 200 documents/topic) interesting articles.
  - Extract entities and relations for selected documents.

- Assumption: Articles are more likely to be relevant if both articles share many common relations.

Thx @Björn

Our approach

- Entities often occur by chance, possibly in different contexts, in both articles:
  - **Nixon** *goes to* **China.**
  - **Nixon** was born in 1913. … In 2010, China became the world's second largest economy.



The shorthand is "Nixon goes to China," meaning a moment in which a leader reverses his past positions to do something that is shocking but beneficial.

Thx @Björn

Our approach

Extract all entities for each of the selected documents (spacy):

- New York, Trump, Labor Department, Nasa, Air Force

Find relations between the entities:

- Nixon visits China, Obama pass healthcare

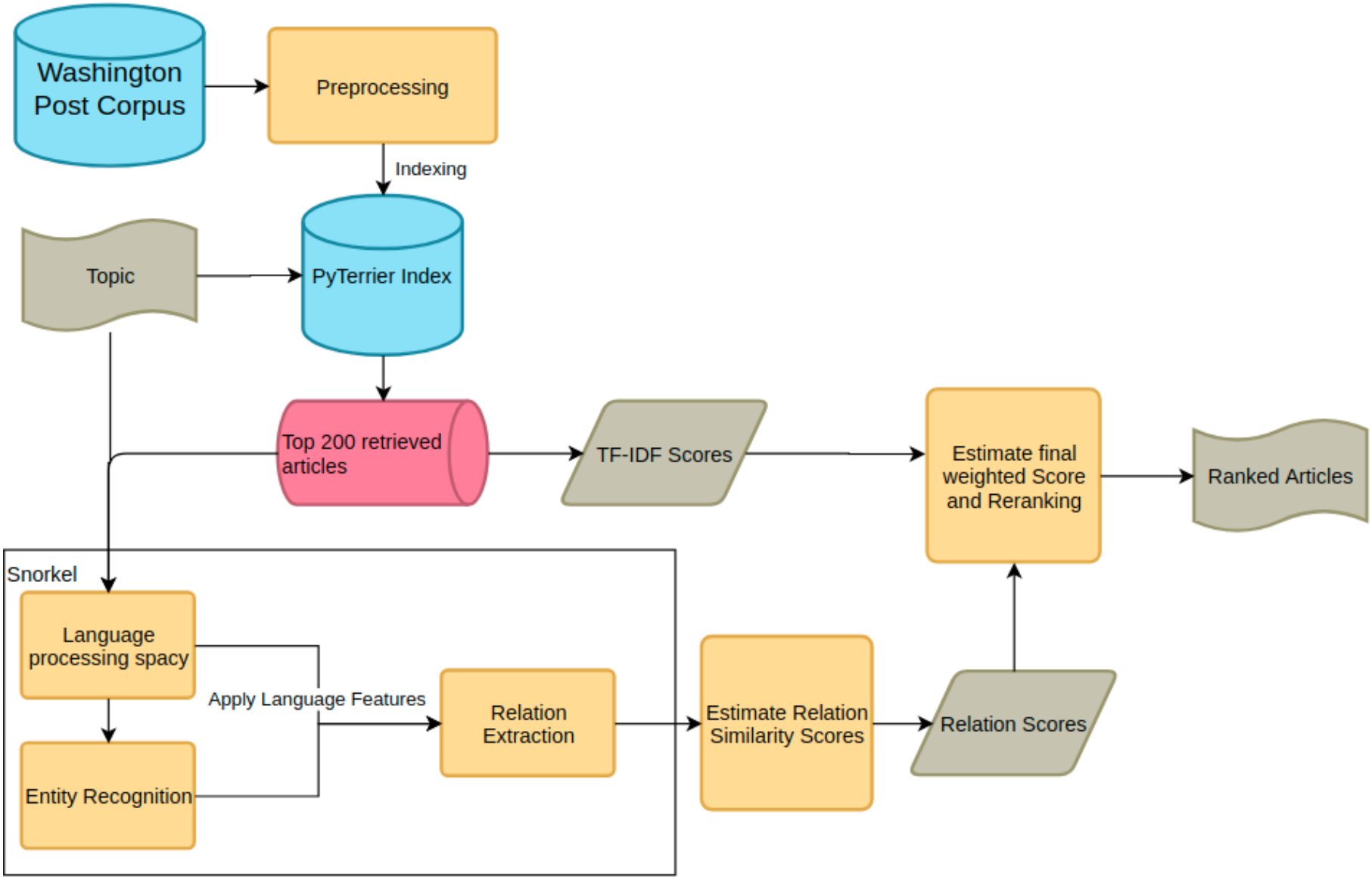
Simple recognition of relations via linguistic features (Snorkel):

- Do entities appear in the same sentence?
- Is there a subject-object relationship between entities?
- Which verb connects the entities?

Give more weight to rare entities (IDF).

Final score for document calculated from TF-IDF score and weighted amount of shared relations.

Thx @Björn


Washington Post Corpus
Preprocessing
Indexing
Topic
PyTerrier Index
Top 200 retrieved articles
TF-IDF Scores
Estimate final weighted Score and Reranking
Ranked Articles
Snorkel
Language processing spacy
Apply Language Features
Entity Recognition
Relation Extraction
Estimate Relation Similarity Scores
Relation Scores

Thx @Björn

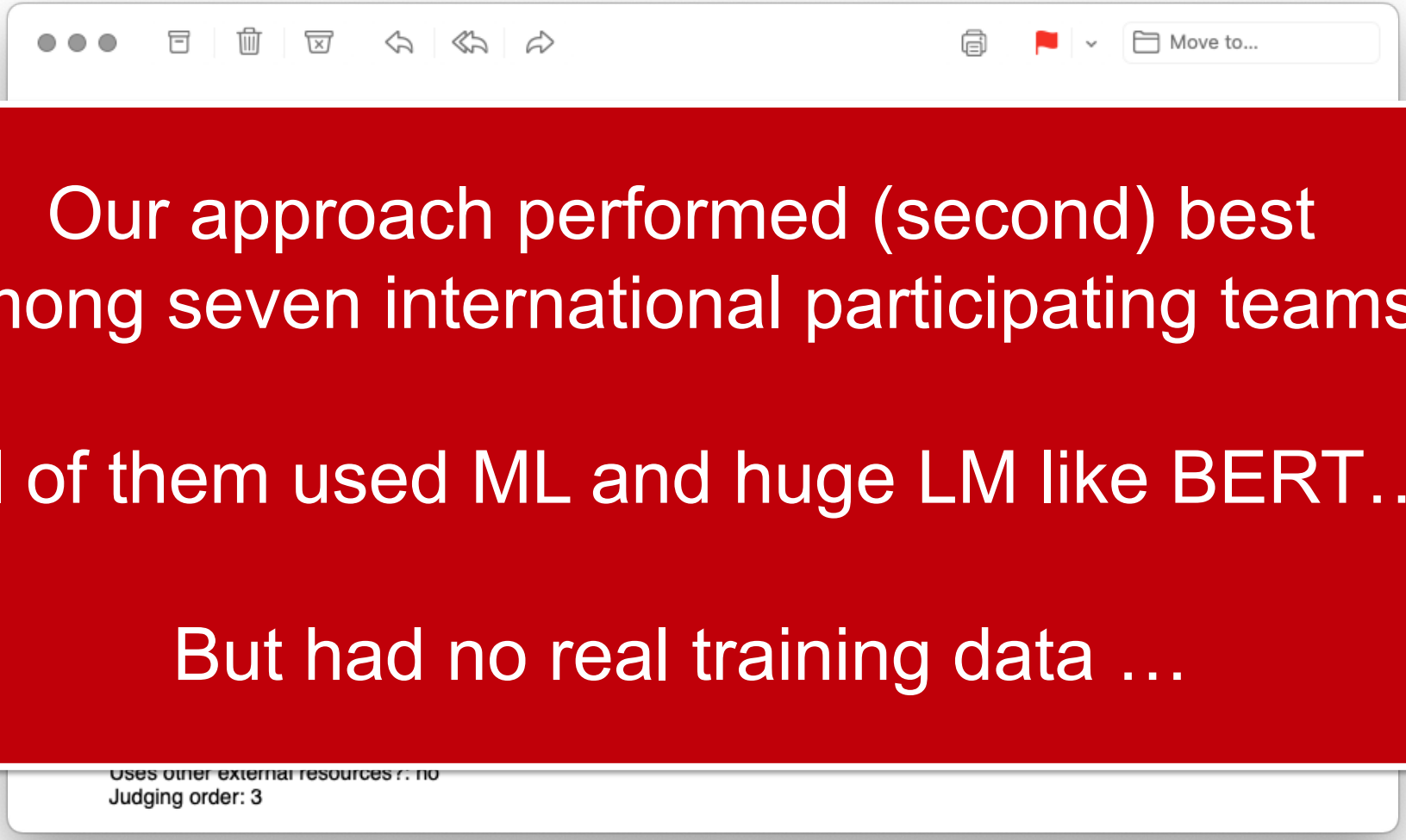
Real-world Example

Retrieved Example:

- **Topic:** Olympics 2020
- **Topic Title:** For 2020 Olympic hopefuls, postponement is another challenge to overcome.
- **Title of retrieved Article:** For finely tuned Olympic athletes, a one-year postponement changes everything
- **Shared Relation Entities:**
  - 'U.S.' - 'Tokyo'
  - 'British' - 'Adam Peaty'
  - 'Helen Maroulis' - 'Olympic'
  - 'American' - 'Emma Coburn'

Thx @Björn

Fun fact



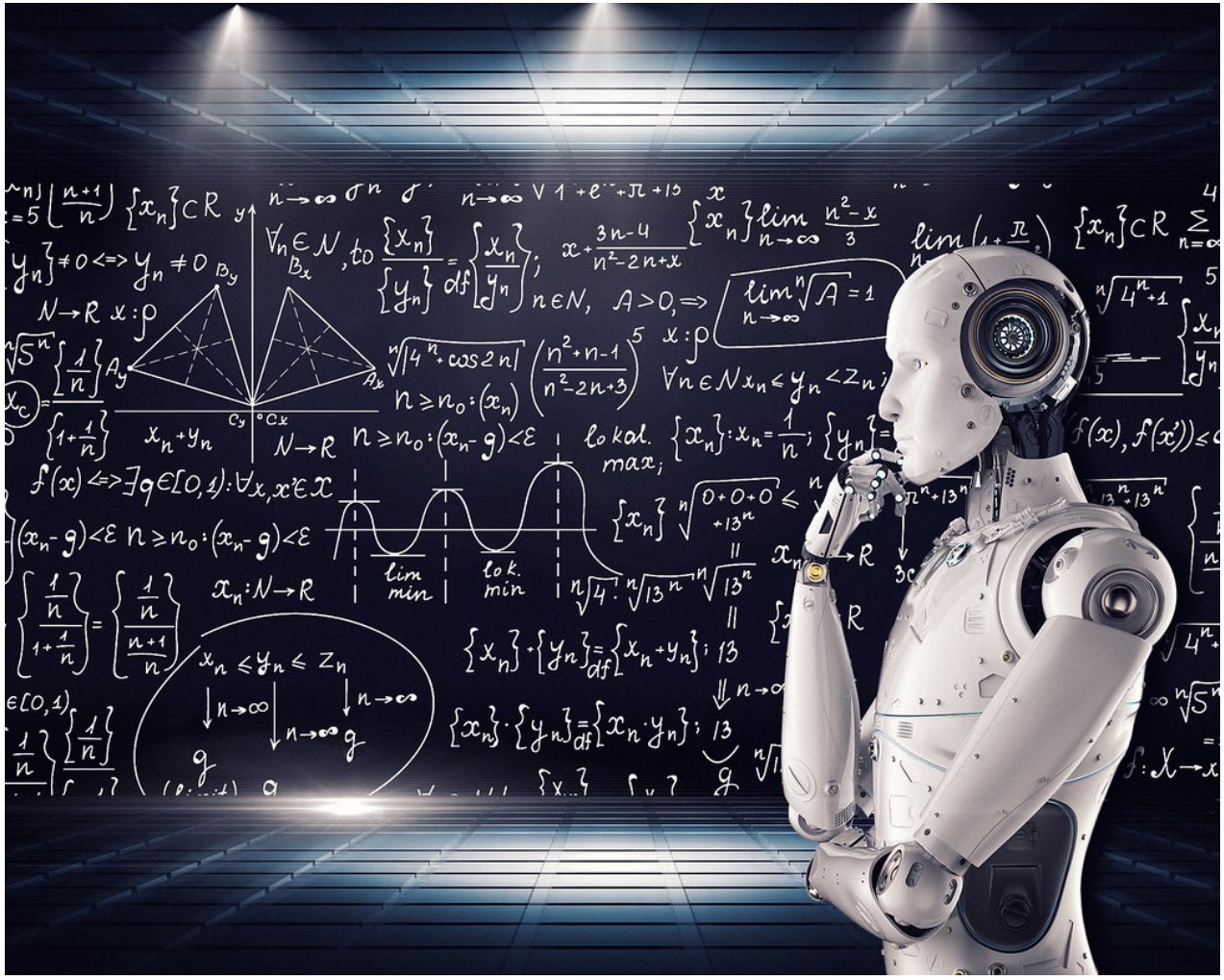
Our approach performed (second) best among seven international participating teams.

All of them used ML and huge LM like BERT…

But had no real training data …

https://trec.nist.gov/pubs/trec30/appendices/newsBackground.html

NLP tasks are often ML tasks



The Data Programming Paradigm

In many applications, we would like to use machine learning, but we face the following challenges:

- (i) *hand-labeled* training data is not available, and is prohibitively expensive to obtain in sufficient quantities (domain experts)
- (ii) *related external knowledge bases* are either unavailable or insufficiently specific, precluding a traditional distant supervision;
- (iii) *application specifications* are in flux, changing the model we ultimately wish to learn.

In data programming, rather than manually labeling each example, users instead describe the *processes by which* these points could be labeled by providing a set of heuristic rules called *labeling functions*.

https://arxiv.org/abs/1605.07723

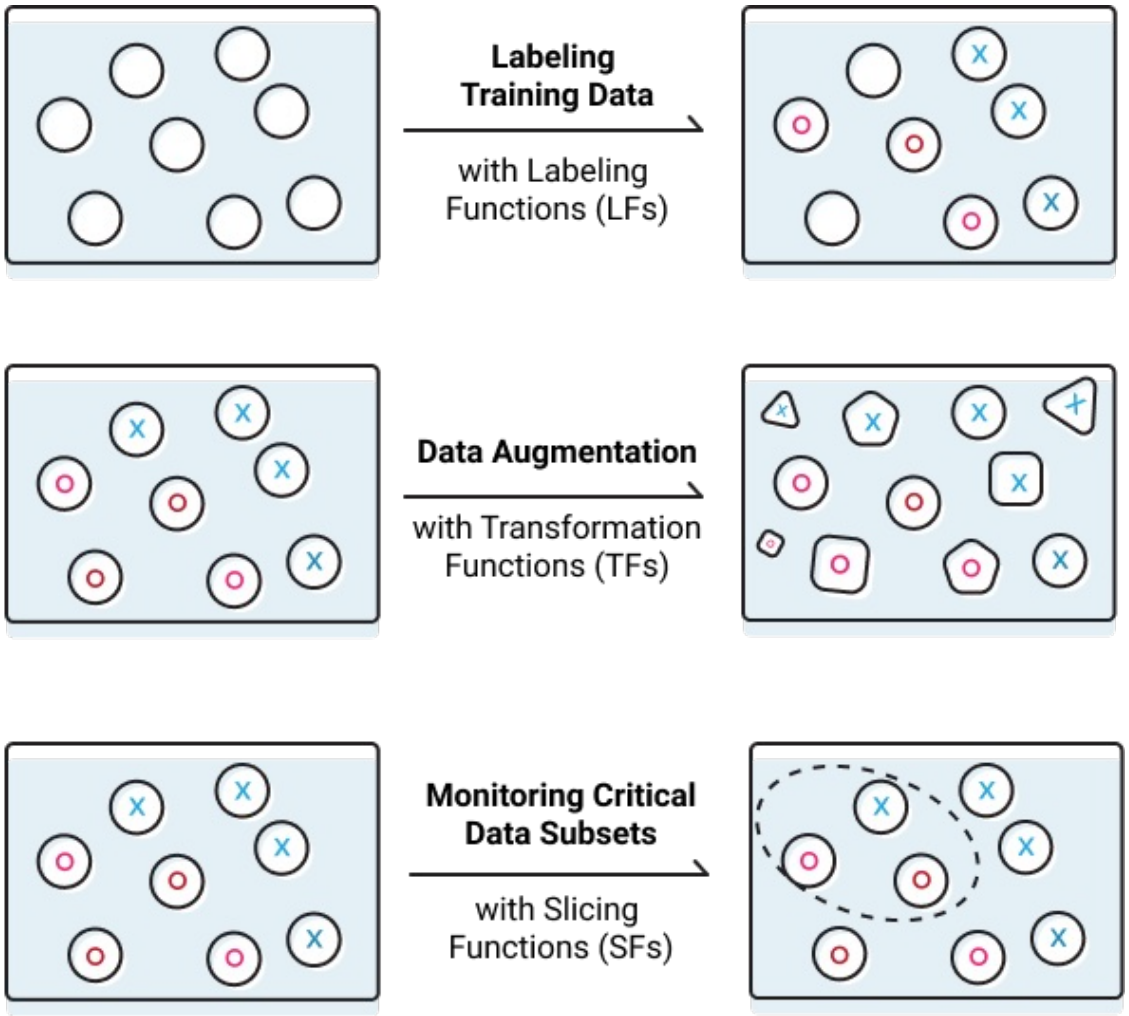
The Data Programming Paradigm

Labeling function

- need not have perfect accuracy or recall;
- rather, it represents a pattern that the user wishes to impart to their model and that is easier to encode as a labeling function than as a set of hand-labeled examples.
- labeling functions can be based on external knowledge bases, libraries or ontologies, can express heuristic patterns, or some hybrid of these types;
- labeling functions are more general than manual annotations, as a manual annotation can always be directly encoded by a labeling function;
- labeling functions can overlap, conflict, and even have dependencies which users can provide as part of the data programming specification

https://arxiv.org/abs/1605.07723

Introducing: Snorkel



https://www.snorkel.org/get-started/

Demo

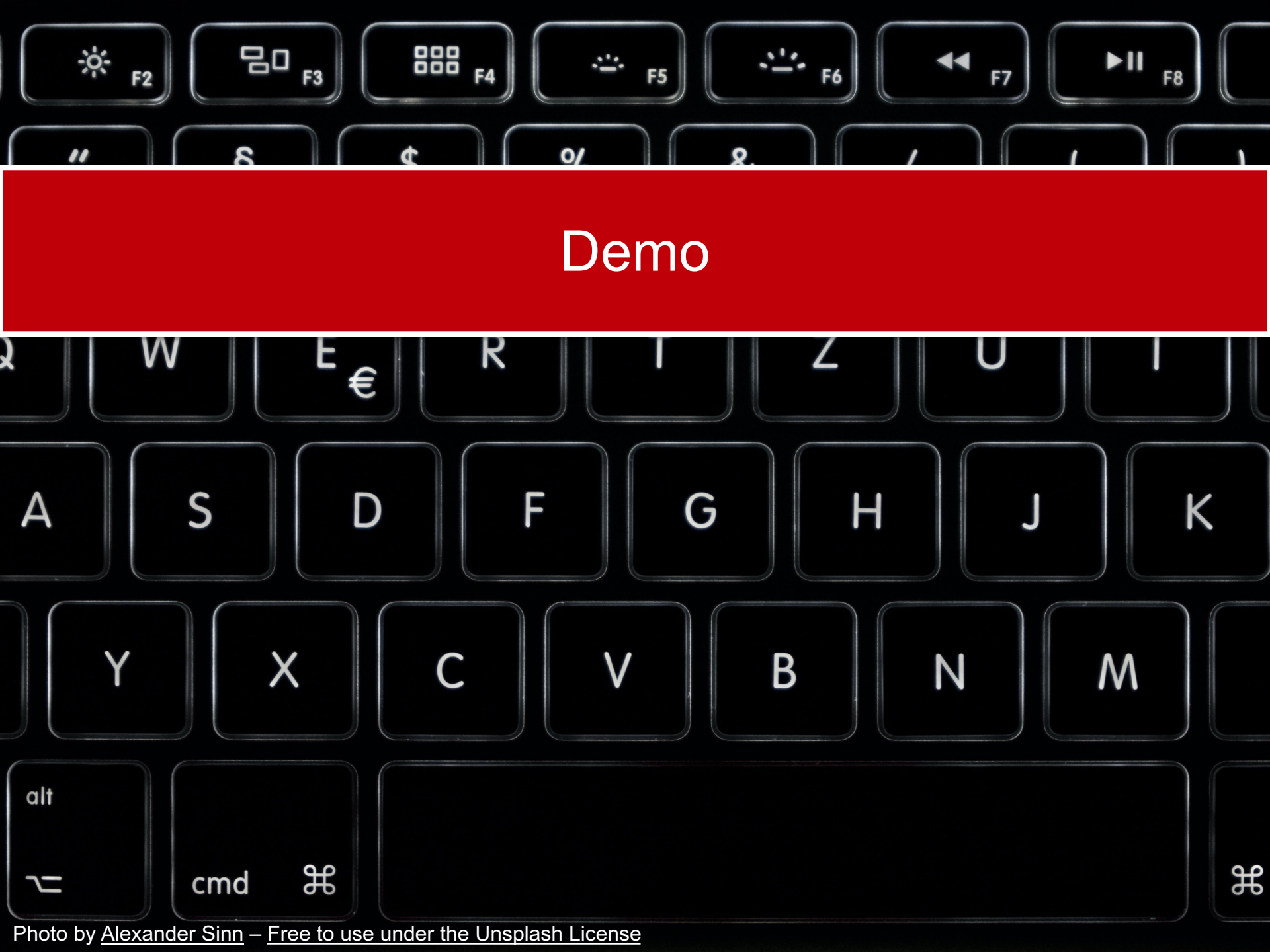


Photo by Alexander Sinn – Free to use under the Unsplash License