University of Milano-Bicocca
Master's Degree in Data Science
Data Science Lab in Biosciences
Academic Year 2022-2023

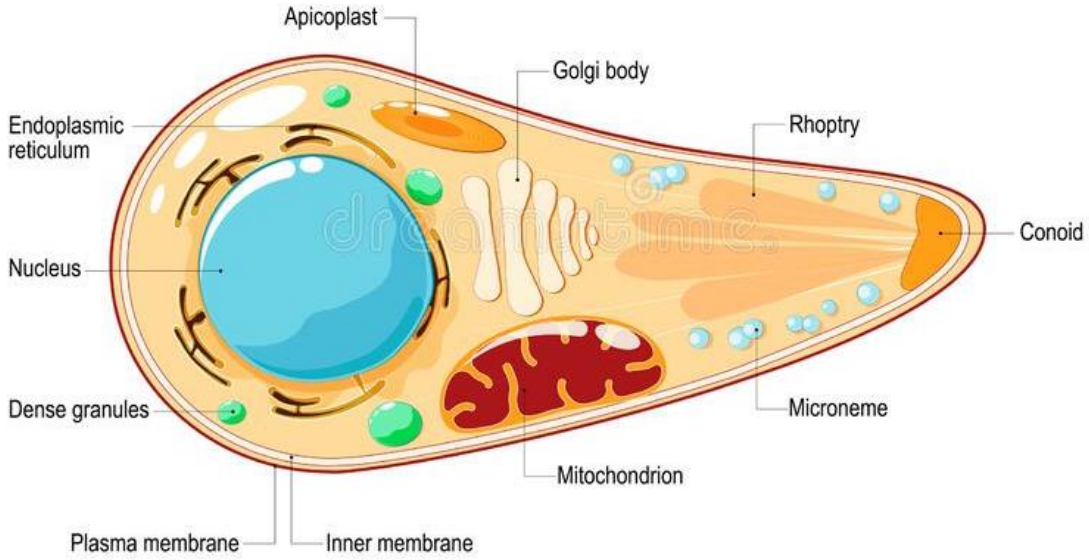
TOXOPLASMA GONDII: NCBI DATABASES (TAXONOMY, NUCLEOTIDE & PubMed) DATA ACQUISITION AND ANALYSIS

Authors:
Giorgio CARBONE matr. n° 811974
Emilio LINGENTHAL matr. n° 889111
Lorenzo LONGO matr. n° 846738

GitHub Repository

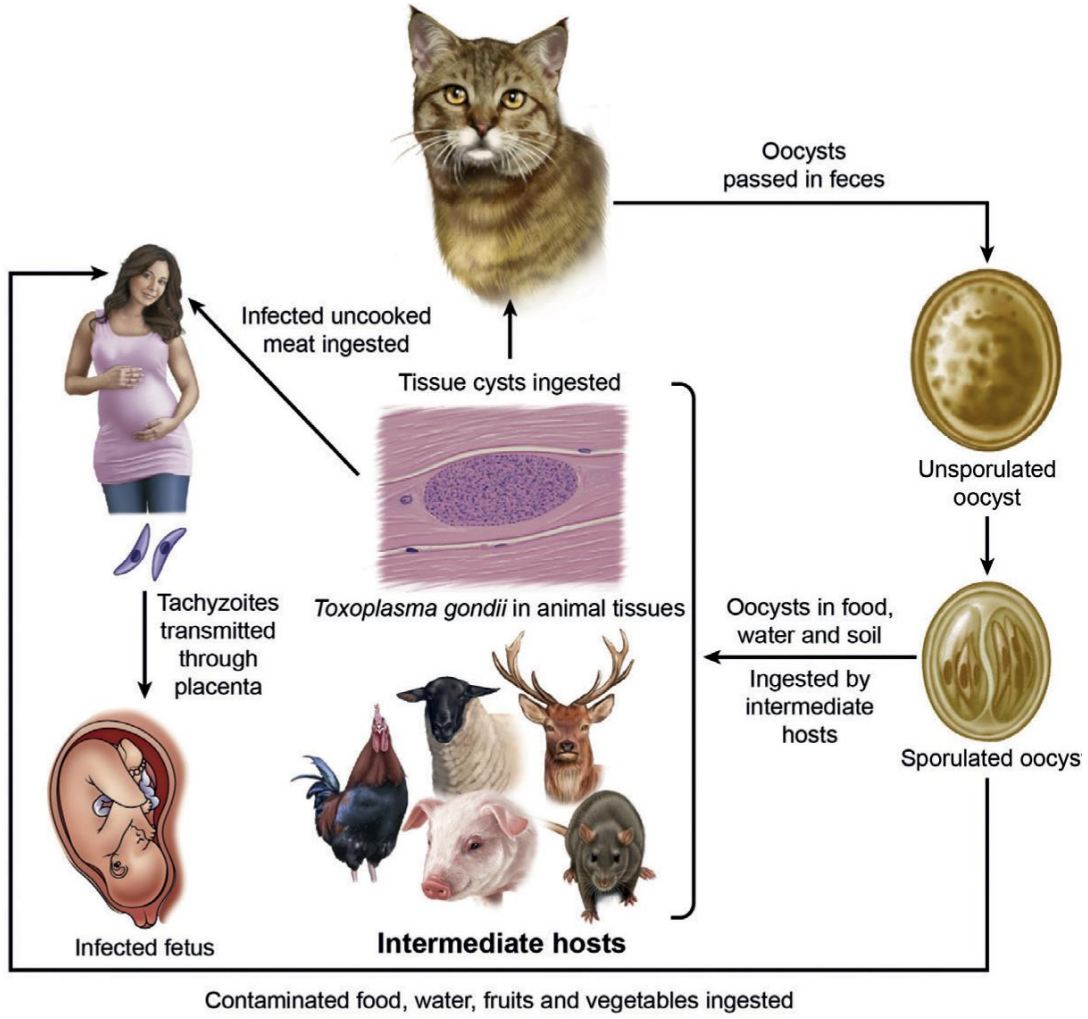
Toxoplasma Gondii

- It is a **parasitic protista** that lives mainly in cats
- It can cause **toxoplasmosis** in **humans**
  - especially in individuals with a **weak immune system**
- Usually asymptomatic, but can create convulsions or poor coordination in severe cases



Toxoplasma Gondii: Life cycle

- **Cats** become **infected** by **eating meat** containing the parasite
- The parasite **reproduces** sexually in the cats in the cells of the **intestine**, forming **oocysts**
- In a **second stage**, the parasite can **infect another host** (e.g. humans)
  - by propagating in the cells through asexual reproduction



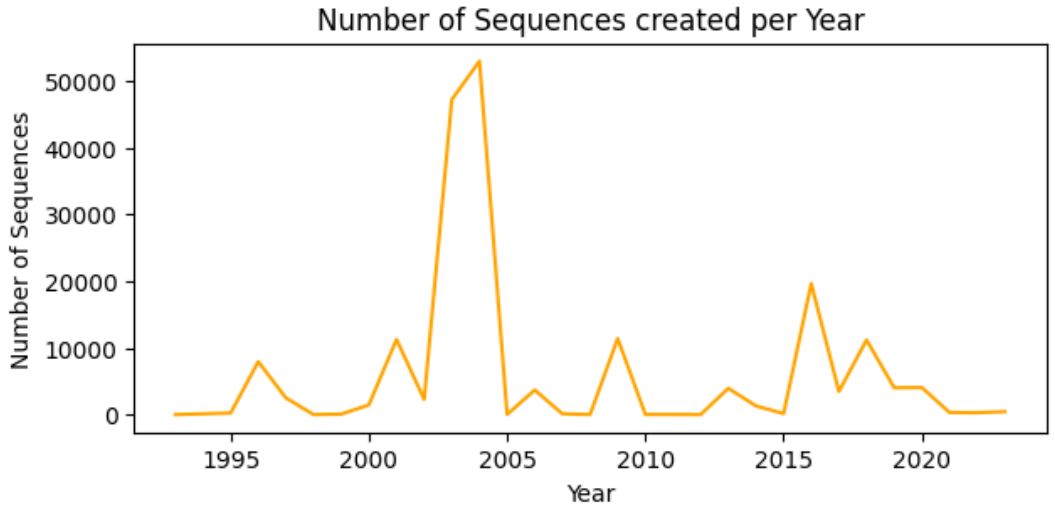
Taxonomy

- The Taxonomy Database
  - Query → `"toxoplasma gondii"[All]`
- ***Apicomplexa*** is a subset of protists capable of attacking host cells
- ***Eucoccidiorida*** is a collection of microscopic unicellular parasites
- ***Sarcocystidae*** refers to ***Apicomplexan*** protists capable of generating disease in humans and animals
- ***Gondii*** is the only known member of the ***Toxoplasma***

Domain	Eukaryota
Kingdom	Protista
Phylum	Apicomplexa
Class	Conoidasida
Order	Eucoccidiorida
Family	Sarcocystidae
Genus	Toxoplasma
Species	gondii

Nucleotide: Data Acquisition

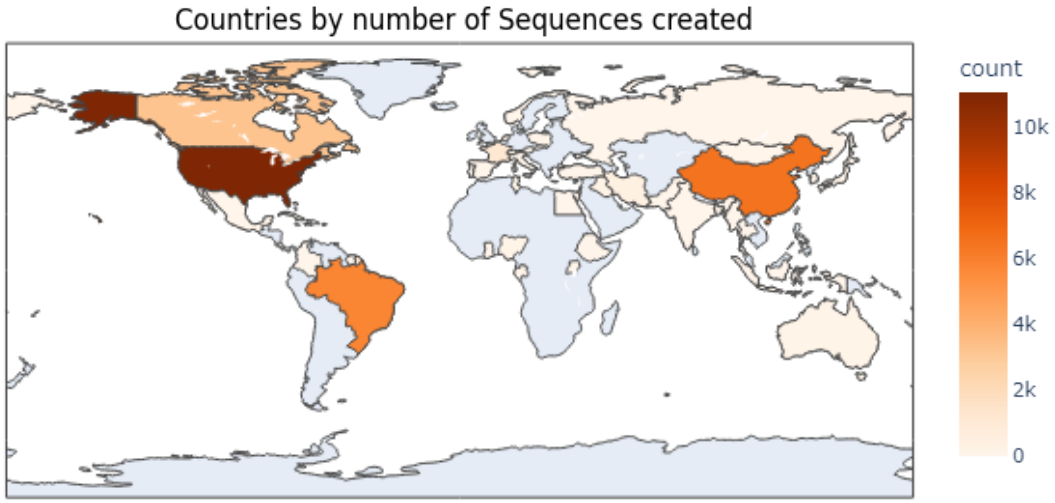
- 189’981 sequences from The Nucleotide Database
  - Query → `"toxoplasma gondii"[Organism]`
  - Using **BioPython** Library
- **Parsing** → 189981 instances and 13 attributes datasets

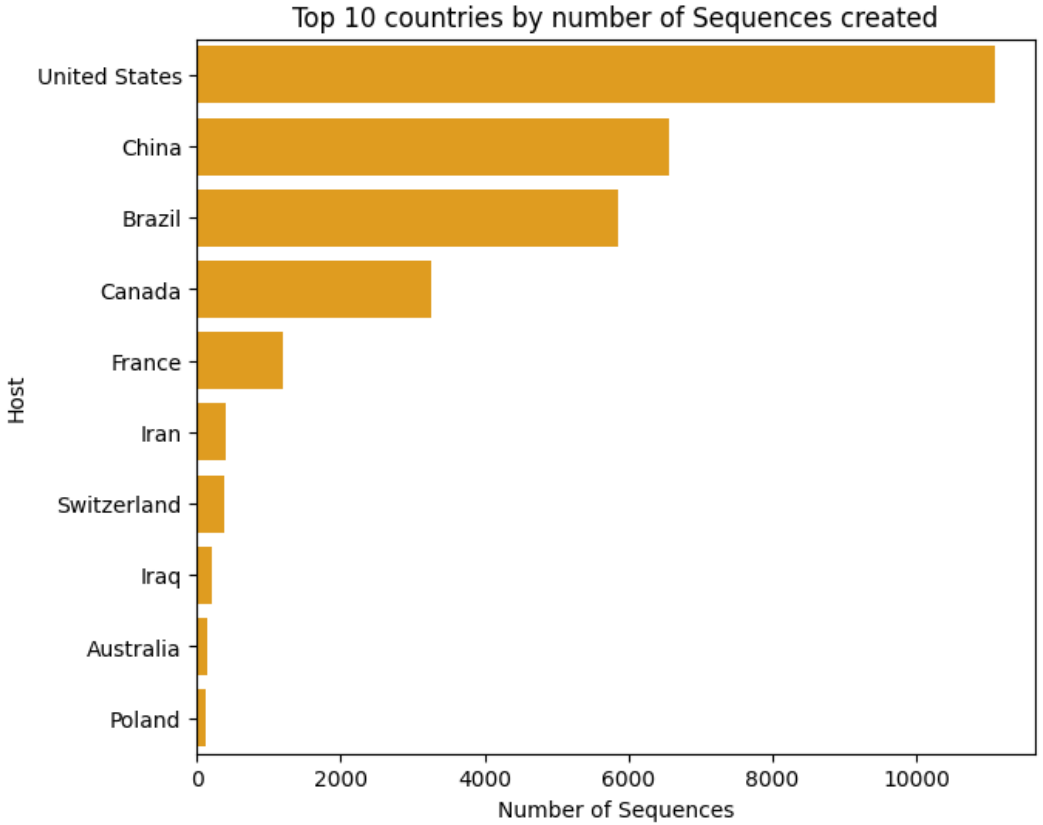


Attribute	Example
GBSeq_locus	OQ402957
GBSeq_length	597
GBSeq_strandedness	double
GBSeq_create-date	26-FEB-2023
GBSeq_update-date	26-FEB-2023
GBSeq_definition	Toxoplasma gondii voucher W22_6329 Apico allel...
Submission Type	Direct Submission
Sequencing Type	##Assembly-Data-START## ; Sequencing Technolog...
mol_type	genomic DNA
isolation_source	brain
host	Castor fiber
country	Switzerland
collection_date	2022

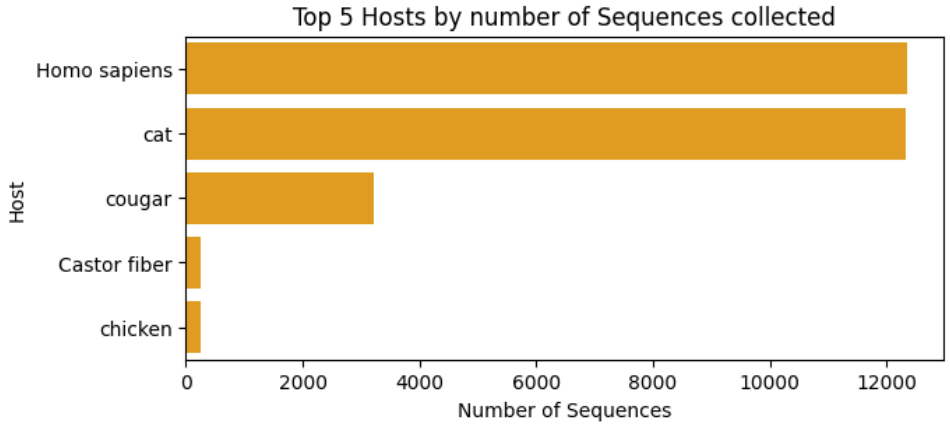
Nucleotide: Countries Analysis

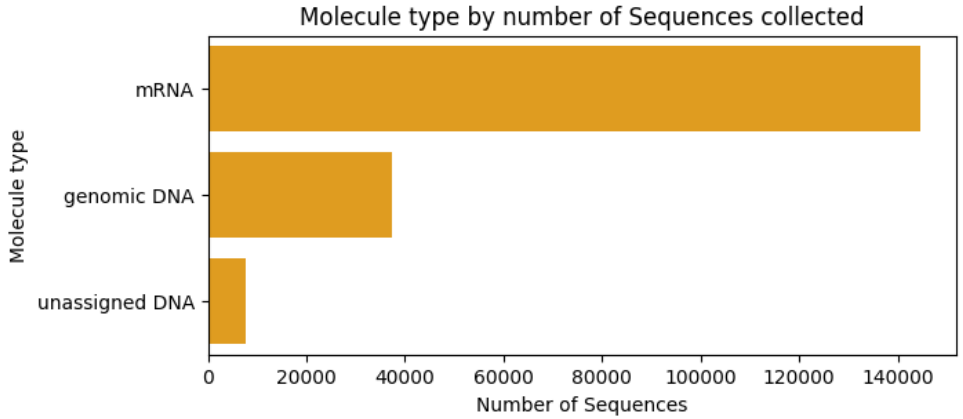
- Most of the sequences are updated from **USA, China, Brazil, Canada and France**

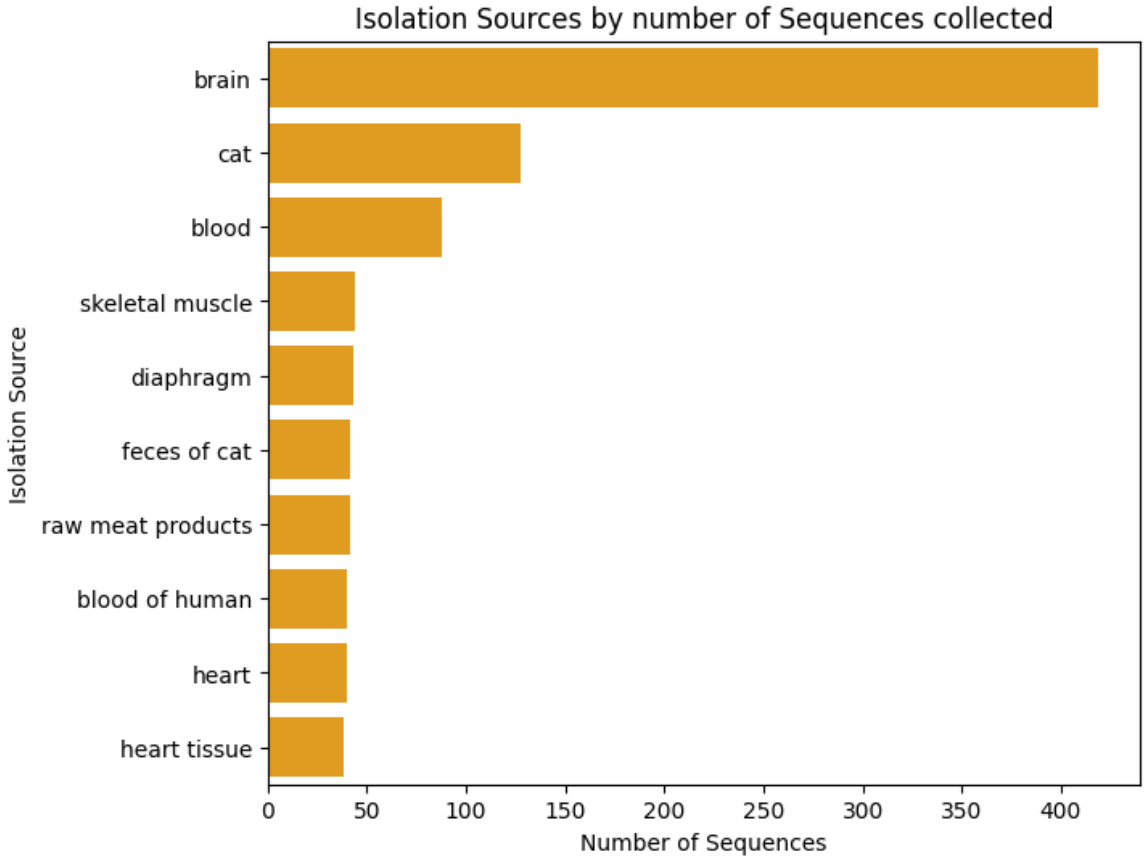




Nucleotide: Sources Analysis







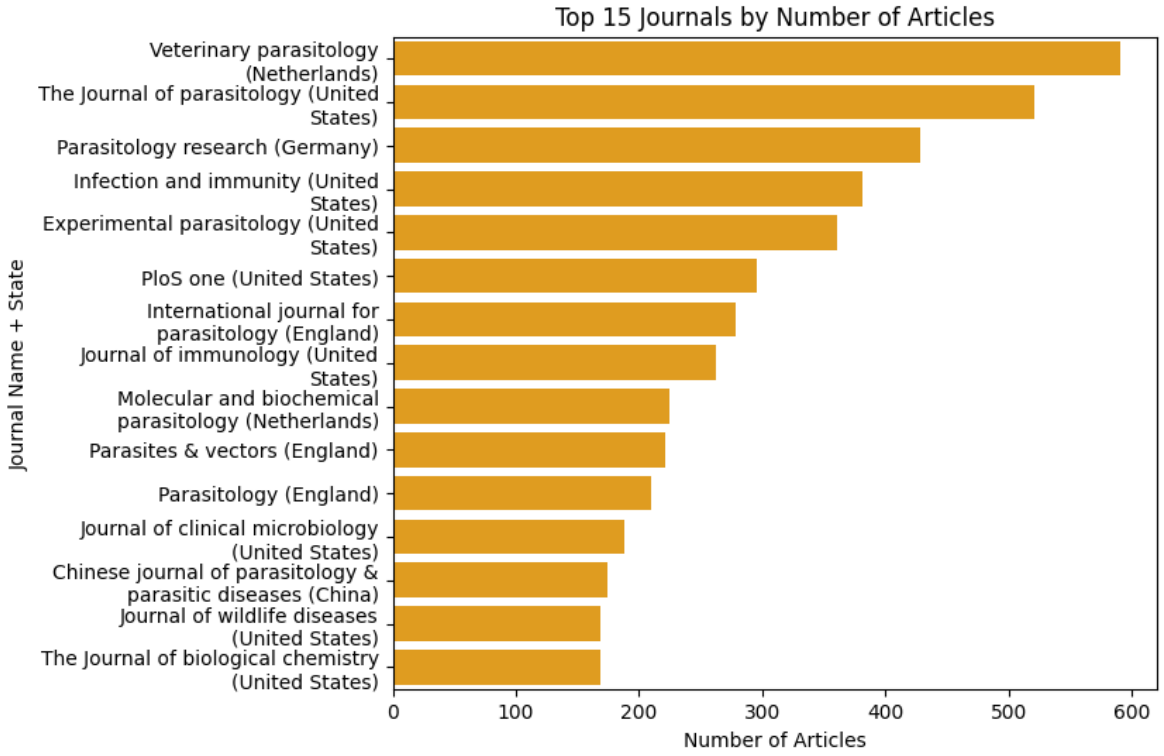
PubMed: Data Acquisition

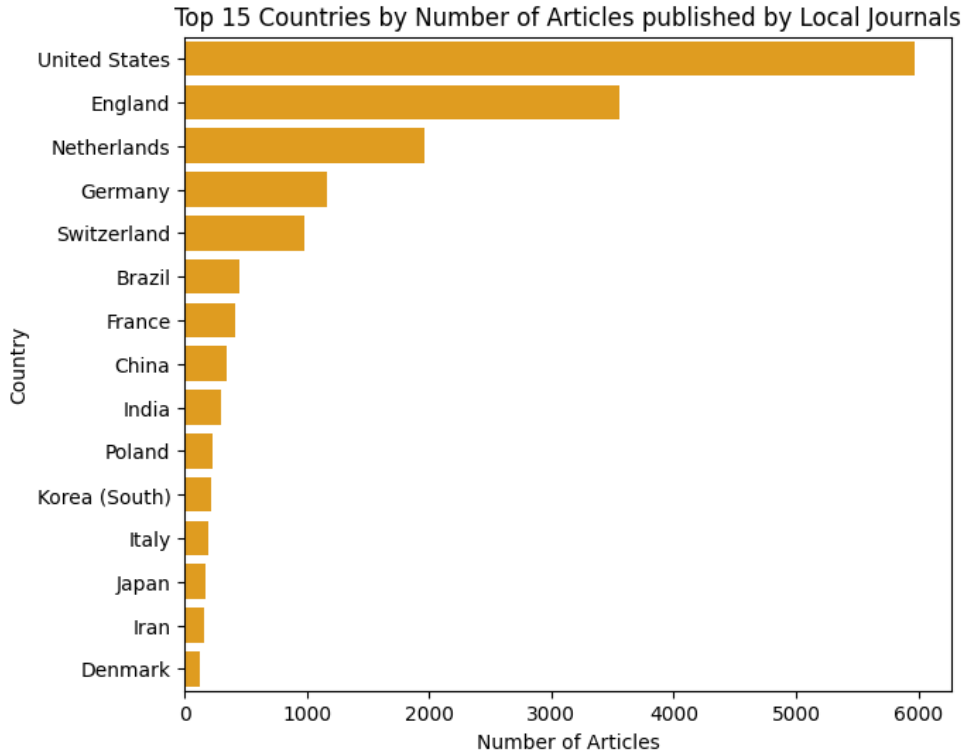
- **17651 publications** containing the string '**toxoplasma gondii**' in the corresponding **title or abstract**.
  - Query → `"toxoplasma gondii"[Title/Abstract]`
  - Using **BioPython** Library
- **Parsing** → 17642 instances and 15 attributes datasets
- Text **Pre-processing** on **Abstracts**
  1. Text **Cleaning**
     - Removal → HTML Tags and Entities, URLs, numbers, control chars, special chars, puntuactions
  2. Text **Normalization**
     - lowering, contractions expansion, stop words removal, **lemmatization**
  3. **Tokenization** (uni-grams)

Attribute	**Example**
PubMed ID	31689351
Title	Biochemical and structural ..
Keyword List	['GTPase ', ' enzyme kinetics ' ..
Journal Name	The Biochemical journal
Journal ID	Biochem J
Abstract	Guanylate-binding proteins (GBPs) constitute a ..
Preprocessed Abstract	['toxoplasma', 'gondii', 'protozoan', 'parasite', ...
Article Date	NULL
Completed Date	2020/07/13
Revised Date	2020/07/13
Publication Year	2019
Language	eng
Author List	['Legewie, L ', ' Loschwitz, J ' ..
Publication Type	['Journal Article ', '' Research Support, Non-U.S. Gov't'']

PubMed Publications: Top Journals

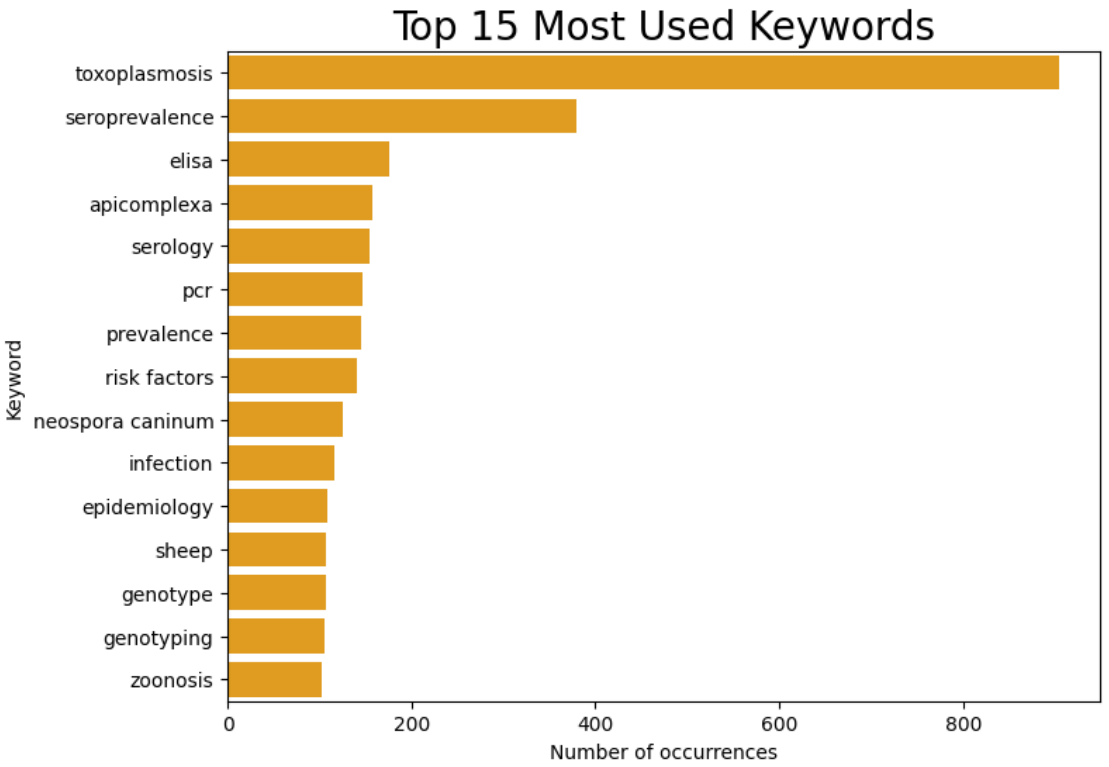
- Articles published in **2173 distinct journals**
- The journal with the **most publications** is **Veterinary parasitology**  with **1128 articles**

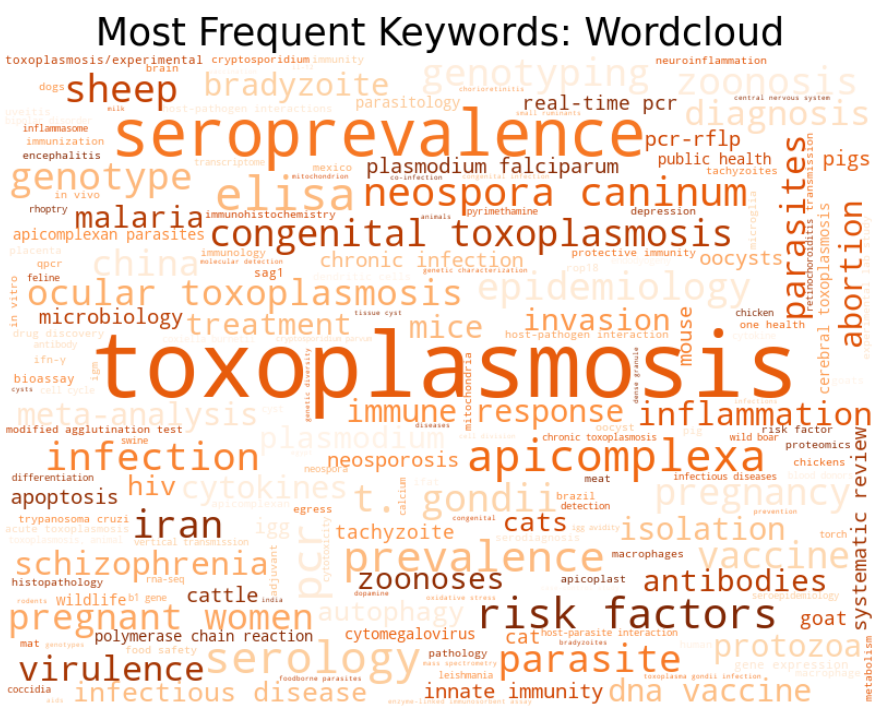




PubMed Publications: Keywords

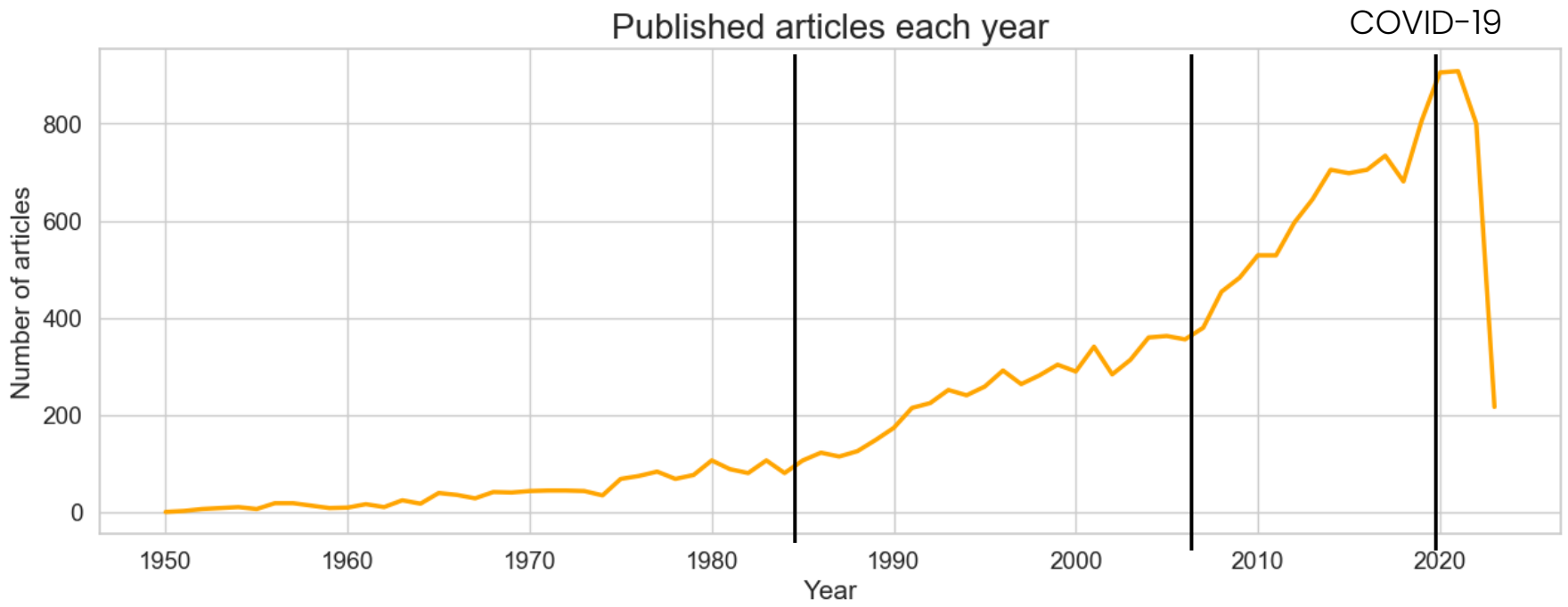
- **“Toxoplasmosis”** is the **most frequent keyword** with 904 occurrences.





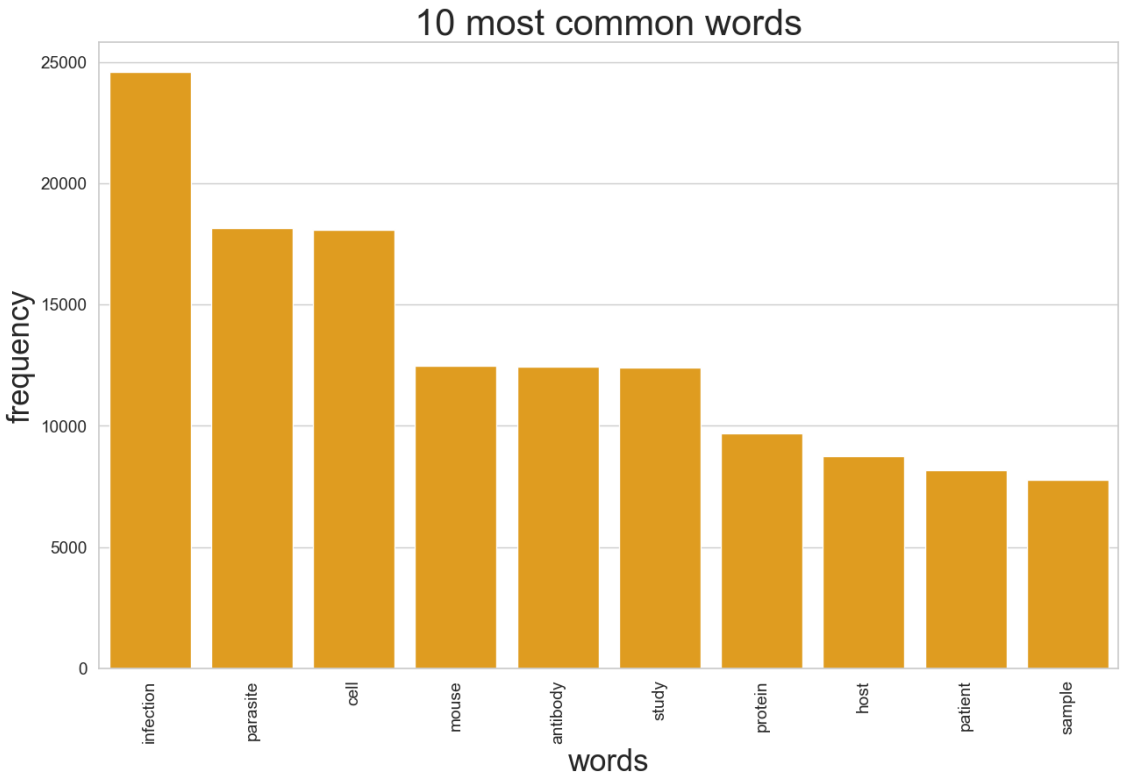
PubMed Publications: publication year

- We can observe an **increasing trend** throughout the years. There are two 2 slope spikes, one in the mid 80's and one in the 2000's



PubMed Publications: Abstracts

- **“Infection”** is the **most frequent** word in all the abstracts with just short than 25k occurrences.



PubMed Publications: Topic modelling on abstracts

- **LDA**-based topic modelling
- hyperparameter tuning to maximize coherence
- 60 topics were created, here's 3 of them
  - Topic 1 → patients
  - Topic 2 → clinical trials
  - Topic 3 → host & transmission

Topic	topic 1	topic 2	topic 3
0	patient	cell	cat
1	hiv	mouse	host
2	cell	infected	assay
3	result	protein	sample
4	study	antibody	infected
5	positive	study	parasite
6	case	parasite	patient
7	parasite	strain	woman
8	antibody	serum	prevalence
9	anti	result	activity