

TopiQAL

Topic Modeling based QA Model to answer topical COVID-19 questions

Hamsa Shwetha Venkataram
Data Scientist



Jet Propulsion Laboratory
California Institute of Technology

We all have questions.

Who am I ?

How big is the Universe ?

How many days in a Martian year ?

Is there a vaccine to COVID-19 ?

**The important
thing is not to stop
questioning.**

Albert Einstein

<https://pbs.twimg.com/media/D1lyBC7X0AAYn4d.jpg>

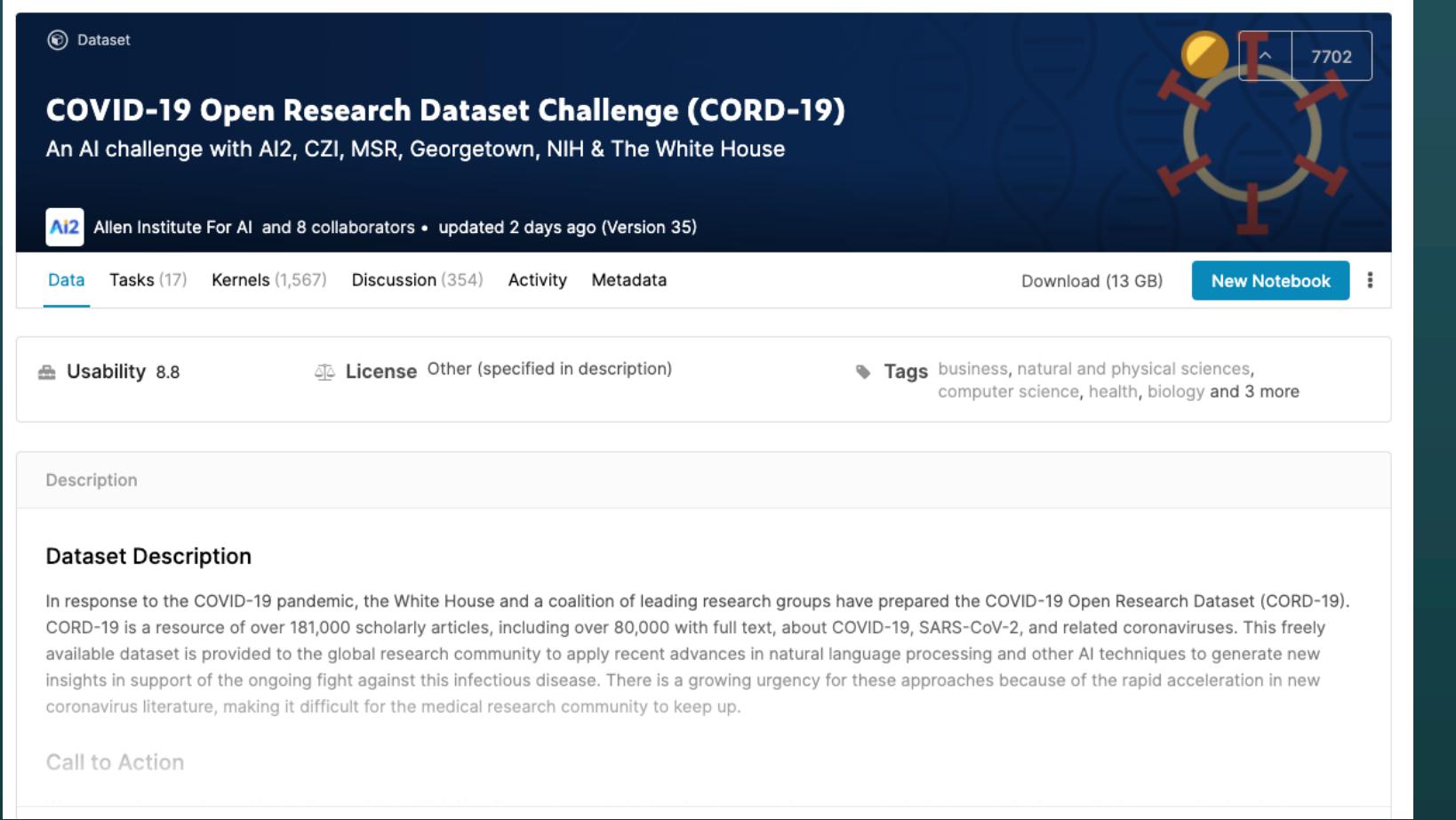
Let's step back a bit ..

The White House and a coalition of leading research groups have prepared the

COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 29,000

scholarly articles, including over 13,000 with full text, about COVID-19, SARS-CoV-2,

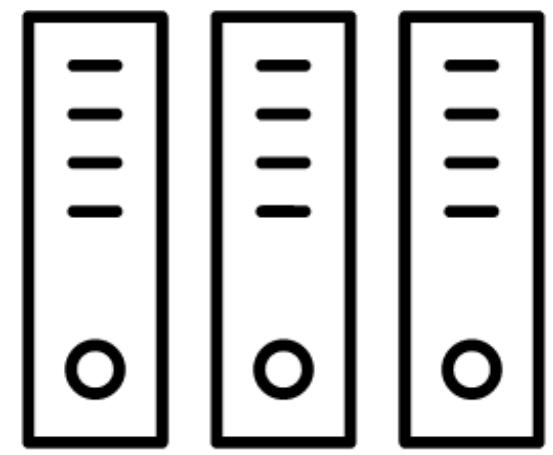
and related coronaviruses.



The screenshot shows the Kaggle dataset page for the COVID-19 Open Research Dataset Challenge (CORD-19). The page has a dark blue header with the title "COVID-19 Open Research Dataset Challenge (CORD-19)" and a subtitle "An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House". Below the header, there is a banner for "Allen Institute For AI and 8 collaborators" with a note "updated 2 days ago (Version 35)". The main navigation bar includes links for "Data", "Tasks (17)", "Kernels (1,567)", "Discussion (354)", "Activity", and "Metadata". On the right side, there are buttons for "Download (13 GB)" and "New Notebook". The page features a circular logo with a DNA helix and a central figure. Below the navigation, there are sections for "Usability 8.8", "License Other (specified in description)", and "Tags business, natural and physical sciences, computer science, health, biology and 3 more". The "Description" section contains a brief overview of the dataset. The "Dataset Description" section provides more detailed information about the dataset's purpose and scope. The "Call to Action" section encourages users to participate in the challenge.



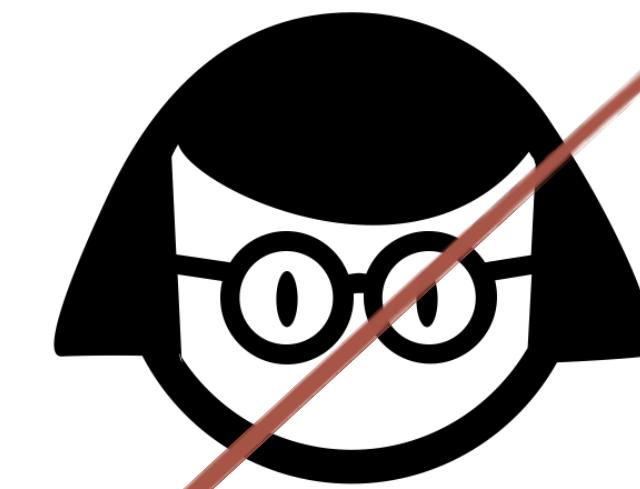
Articles



CORD-19 Dataset



Answers



Subject Matter Experts

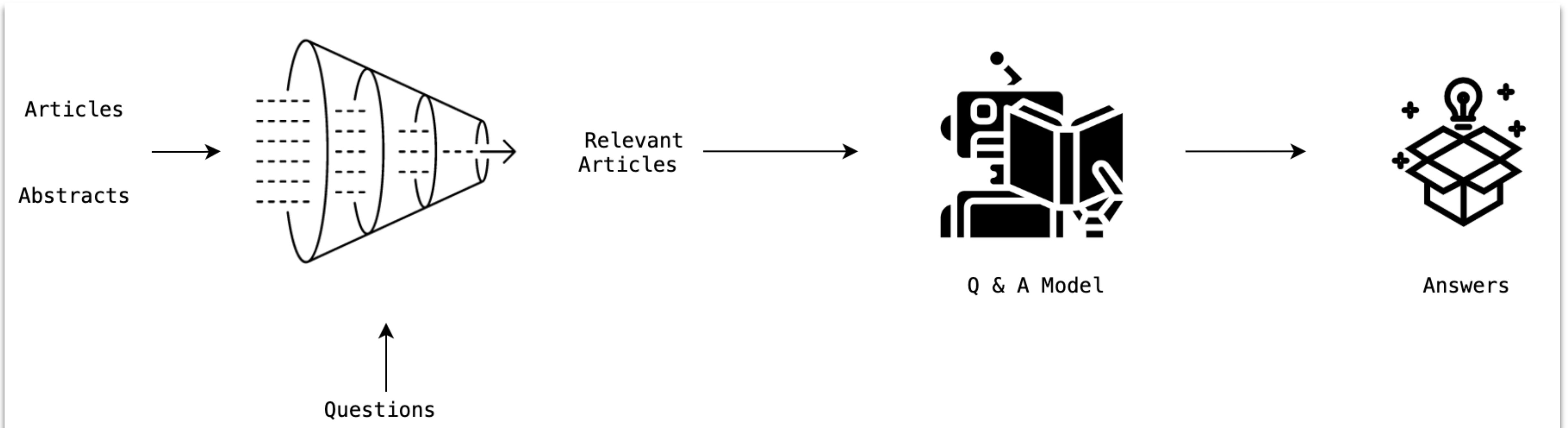


Labels



Tasks - Questions

Why TopiQAL ?



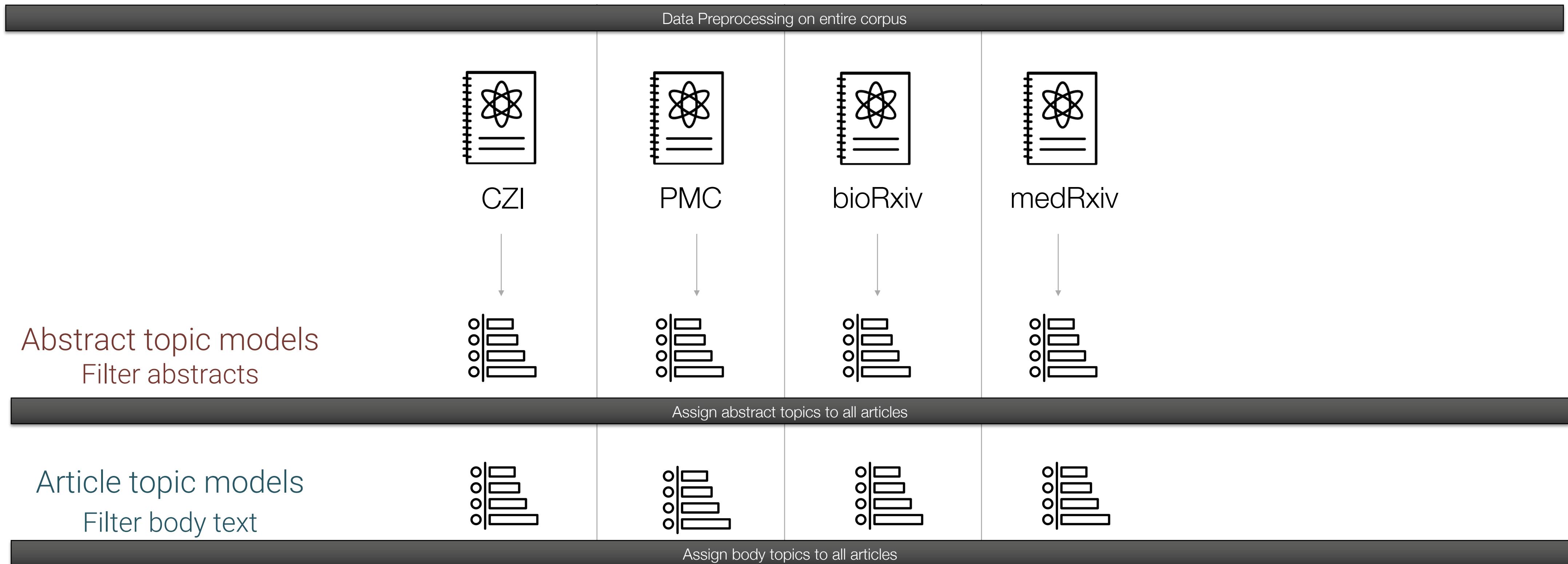


Sums up all that we do!

Topic Models

- Discover hidden thematic structure
- Latent Dirichlet Allocation
- Interpretability of models
- Hierarchical modeling
 - Journals
 - Abstracts
 - Articles
- Setting priors for word-topic combination
 - GuidedLDA
 - Gensim (eta)
- Evaluation
- Visualization
 - [pyLDAVis](#)

Hierarchical Topic Modeling



Each topic is seeded with words from 10 Kaggle Tasks

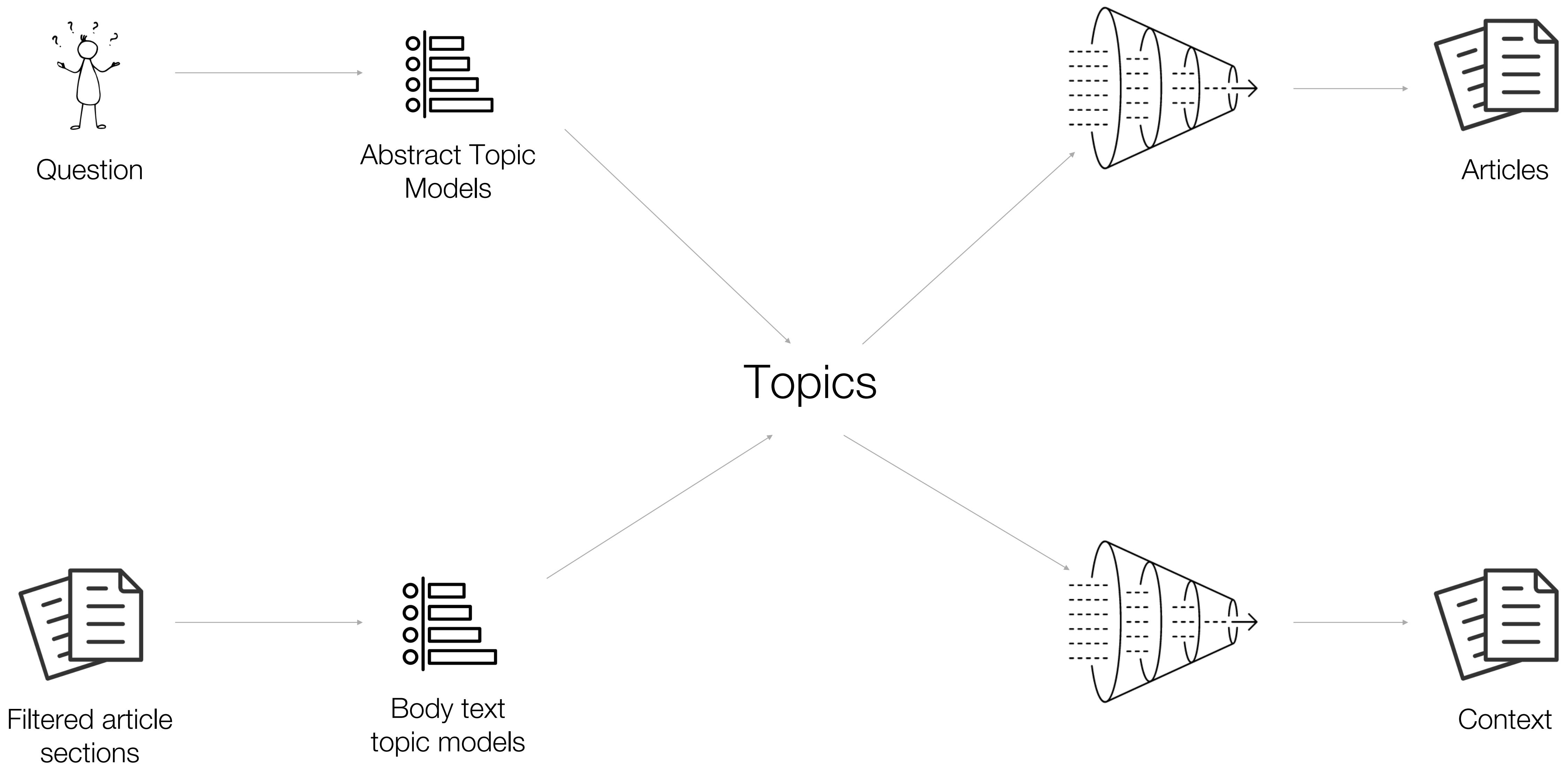
(Bio)BERT- SQuaD



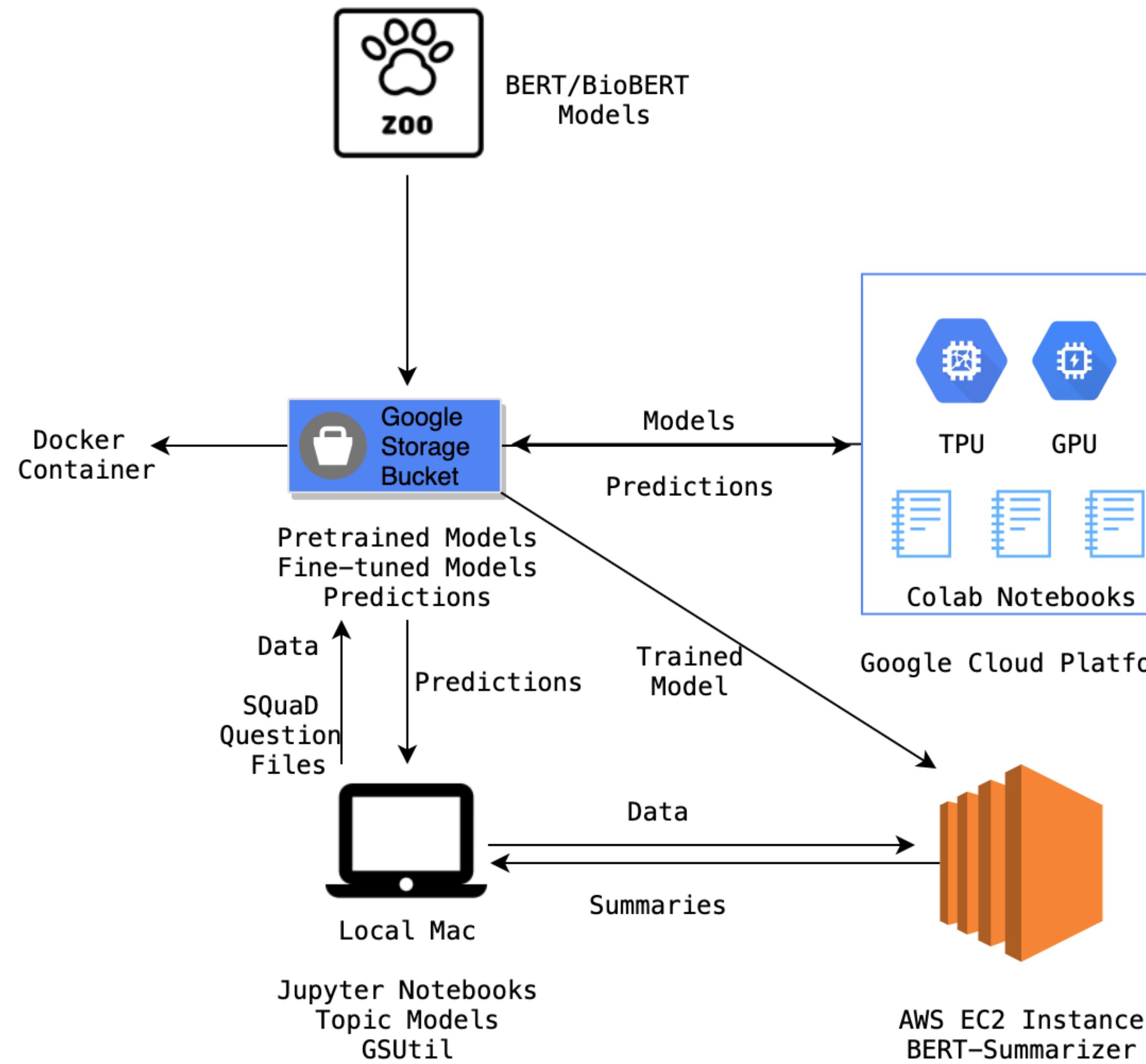
BERT

- What is BERT ?
- How can BERT answer questions ?
 - SQuAD Dataset
- BioBERT
- Fine-tuning and Inference on TPUs
- Google Colab notebooks and [gsutil](#)
- Evaluation

Inference Cycle



Architecture & BioBERT Answers



Question: Methods evaluating potential complication of Antibody-Dependent Enhancement (ADE) in vaccine recipients.

Answer: Nsp3b and E-channel

Question: Exploration of use of best animal models and their predictive value for a human vaccine.

Answer: no target has been predicted

Question: Capabilities to discover a therapeutic (not vaccine) for the disease, and clinical effectiveness studies to discover therapeutics, to include antiviral agents.

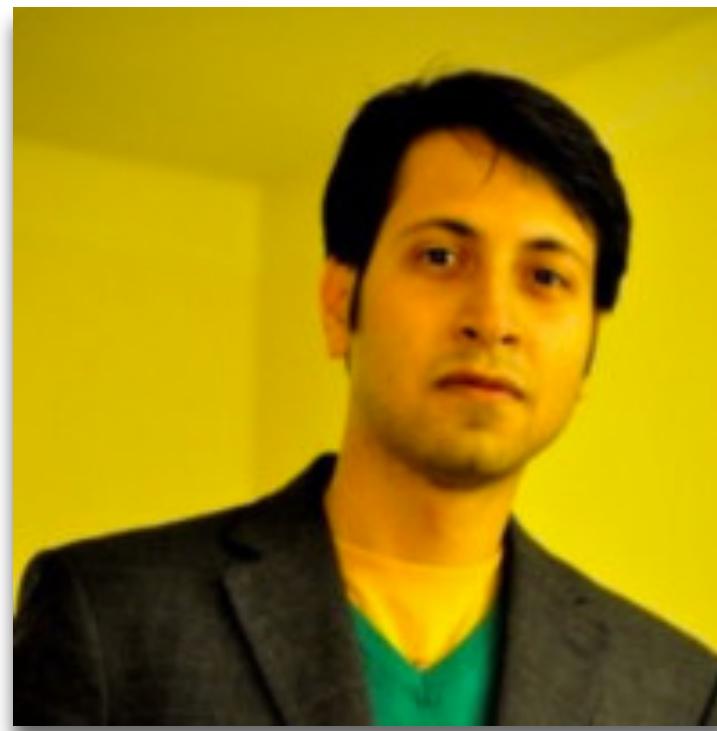
Answer: scientists have come up with three strategies for developing new drugs

Question: Exploration of use of best animal models and their predictive value for a human vaccine.

Team



Hamsa Shwetha



Asitang Mishra



Annie Didier



Anastasia Mensikova



Dr. Chris Mattmann

Thank You!

References

- Icons from Noun Project
- <https://github.com/vi3k6i5/GuidedLDA>
- <https://github.com/nasa-jpl-cord-19/topiQAL>