

Open-domain Question Answering



facebook
research

Danqi Chen

Princeton University

 @danqi_chen

Scott Wen-tau Yih

Facebook AI Research

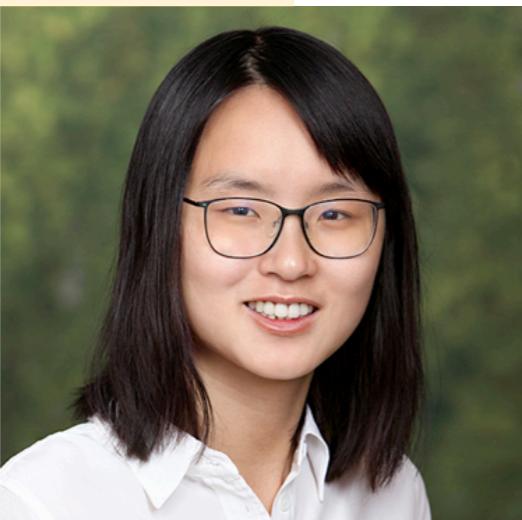
 @scottyih

<https://github.com/danqi/acl2020-openqa-tutorial>

July 5, 2020

Who we are

Danqi



Scott



- Assistant prof at Princeton University since 2019 fall
- Working in QA since 2016
- PhD thesis on neural reading comprehension and question answering
- Author of DrQA, CoQA, SpanBERT/RoBERTa, dense passage retriever (DPR), Stanford Attentive Reader
- Research scientist at Facebook AI Research
- Working in QA since 2013 (or 2001 ☺)
- Taught “Question Answering with Knowledge Base, Web and Beyond” in NAACL & SIGIR 2016
- Author of WikiQA, WebQSP, DPR, retrieval-augmented generation (RAG)

Tutorial slides

Check out our latest slides at:

<https://github.com/danqi/acl2020-openqa-tutorial>

ACL2020 Tutorial: Open-Domain Question Answering

This ACL2020 tutorial will be held on July 5th, 2020, by [Danqi Chen <danqic@cs.princeton.edu>](#) and [Scott Yih <scottyih@fb.com>](#). The video is *NOT* pre-recorded and we will have a 3.5-hour live session at 3-6:30pm PST. You can find more information below and hope to see you there!

Tutorial Slides

NEW: All the slides are available now!

We recommend reading our draft slides before the tutorial. We may have some minor last-minute changes, so please check out the latest version before the live session.

1. [Introduction & problem definition](#)
2. [A history of open-domain \(textual\) QA](#)
3. [Datasets & evaluation](#)
4. [Two-stage retriever-reader approaches](#)
5. [Dense retriever and end-to-end training](#)
6. [Retrieval-free approaches](#)
7. [Open-domain QA using KBs and text](#)
8. [Open problems and future directions](#)

Participation + Q & A

RocketChat: <https://acl2020.rocket.chat/channel/tutorial-8>

The screenshot shows the RocketChat interface. On the left, there's a sidebar with sections for Discussions, Channels, Private Groups, and Direct Messages. The Channels section lists several channels, including #announcements, #general, #helpdesk, #incidents, #live, #professional-conduct-committee, #social-media-posts, and #tutorial-8. A red arrow points from the text "Channel: tutorial-8" at the bottom to the #tutorial-8 channel in the sidebar. The main area shows a channel named "#tutorial-8" with a star icon. Below it, a message from "ACL2020 Admin Bot" is displayed, starting with "Room topic changed to: Open-Domain Question Answering - Danqi Chen and Scott Wen-tau Yih by acl2020-admin-bot". Another message from the same bot follows, providing a detailed overview of the tutorial. At the bottom of the channel view, there's a message from "Ratthachat Chatpatanasiri" and a reply from "Wen-tau Yih".

#tutorial-8
Open-Domain Question Answering - Danqi Chen and Scott Wen-tau Yih

Start of conversation

June 24, 2020

ACL2020 Admin Bot @acl2020-admin-bot 12:11 PM Room topic changed to: Open-Domain Question Answering - Danqi Chen and Scott Wen-tau Yih by acl2020-admin-bot

ACL2020 Admin Bot @acl2020-admin-bot 12:11 PM Room description changed to: This tutorial provides a comprehensive and coherent overview of cutting-edge research in open-domain question answering (QA), the task of answering questions using a large collection of documents of diversified topics. We will start by first giving a brief historical background, discussing the basic setup and core technical challenges of the research problem, and then describe modern datasets with the common evaluation metrics and benchmarks. The focus will then shift to cutting-edge models proposed for open-domain QA, including two-stage retriever-reader approaches, dense retriever and end-to-end training, and retriever-free methods. Finally, we will cover some hybrid approaches using both text and large knowledge bases and conclude the tutorial with important open questions. We hope that the tutorial will not only help the audience to acquire up-to-date knowledge but also provide new perspectives to stimulate the advances of open-domain QA research in the next phase. by acl2020-admin-bot

July 4, 2020

R Ratthachat Chatpatanasiri @Jung 9:36 AM
Hi, I can see only the first page of the slide. Not sure if other people have the same issue?

1 reply July 4, 2020 10:32 AM

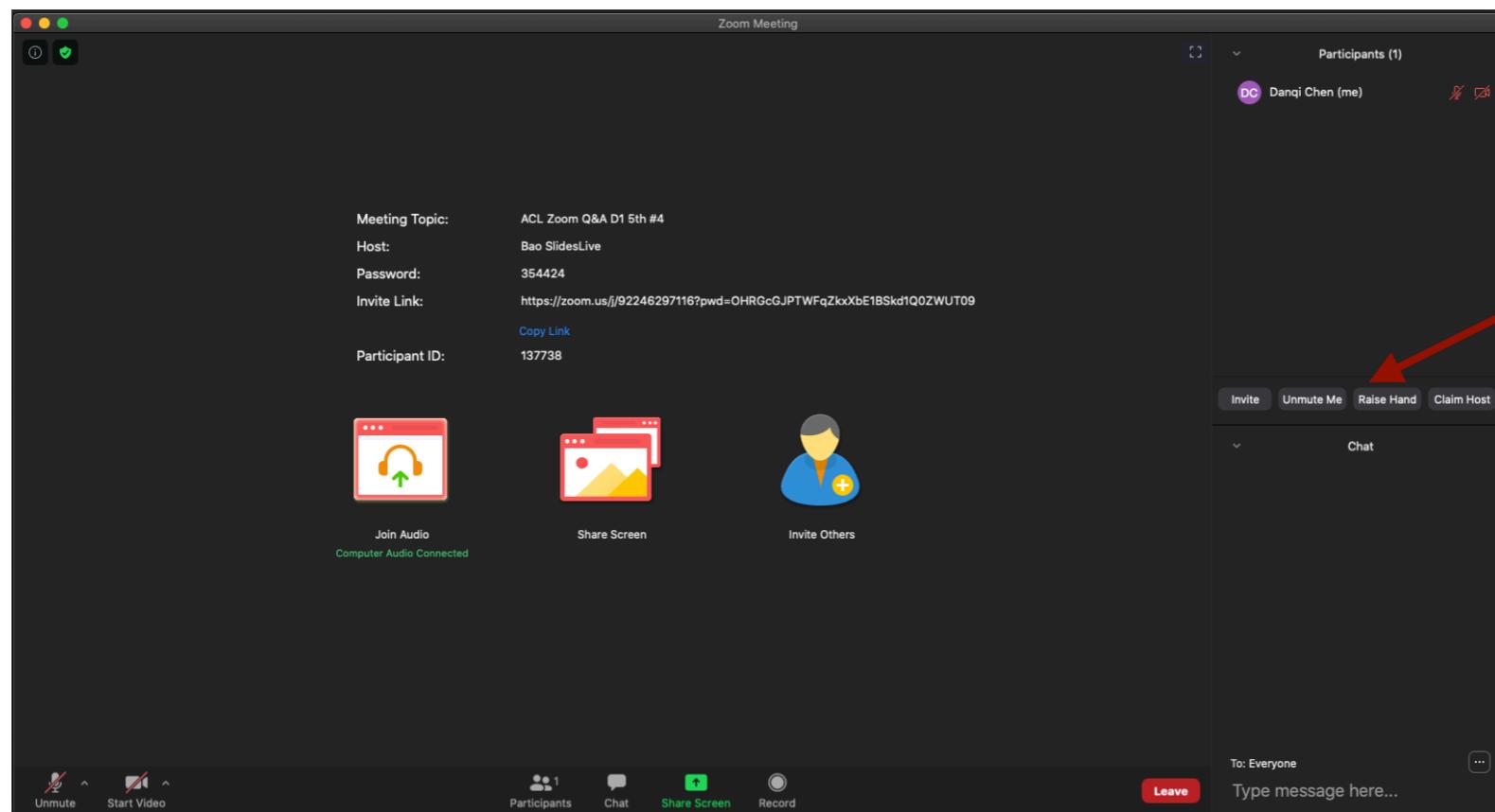
D Hey, I am still communicating with the chairs about fixing the website. In the meantime, all the slides are available on the Github link already: https://gi...

Channel: tutorial-8

Participation + Q & A



Please join us
on Zoom if
possible!

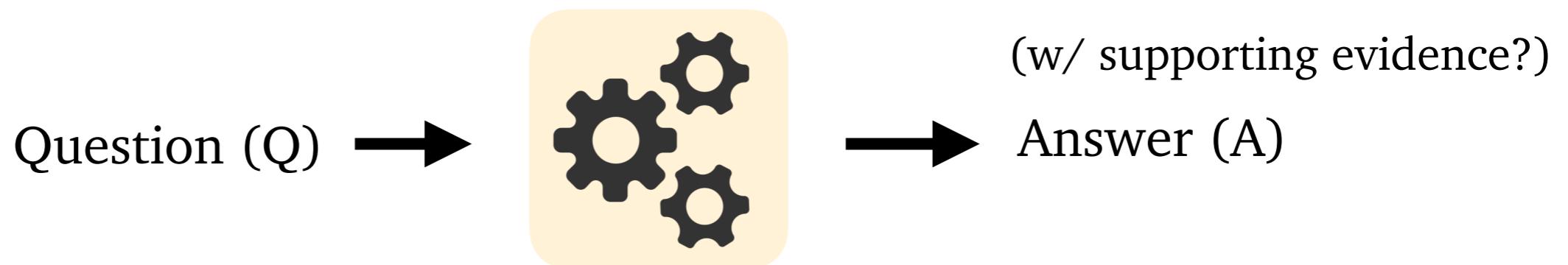


Click "Participants"

Use RocketChat instead of Zoom Chat!

Open-domain QA

- **Question answering** = build computer systems that **automatically** answer questions posed by humans in a **natural language**



- **Open-domain** = deal with questions about nearly anything, usually rely on *general ontologies* and *world knowledge*

Q: Where does the energy in a nuclear explosion come from?

A: high-speed nuclear reaction

Q: Where is Einstein's house?

A: 112 Mercer St, Princeton, NJ

Q: How many papers were accepted by ACL 2020?

A: 779 papers

Open-domain QA

Knowledge Bases



Structured

Tables

Category	Structure	Country	City	Height (metres)	Height (feet)
Mixed use	Burj Khalifa	United Arab Emirates	Dubai	829.8	2,722
Self-supporting tower	Tokyo Skytree	Japan	Tokyo	634	2,080
Mixed use	Shanghai Tower	China	Shanghai	632	2,073
Clock building	Abraj Al Bait Towers	Saudi Arabia	Mecca	601	1,972
Military structure	Large masts of INS Kattabomman	India	Tirunelveli	471	1,545
Mast radiator	Lualualei VLF transmitter	United States	Lualualei, Hawaii	458	1,503
Twin towers	Petronas Twin Towers	Malaysia	Kuala Lumpur	452	1,482
Residential	432 Park Avenue	United States	New York	425.5	1,396
Chimney	Ekibastuz GRES-2 Power Station	Kazakhstan	Ekibastuz	419.7	1,377
Radar	Dimona Radar Facility	Israel	Dimona	400	1,312
Lattice tower	Kiev TV Tower	Ukraine	Kiev	385	1,263
Electricity pylon	Zhoushan Island Overhead Powerline Tie	China	Zhoushan	370	1,214

Semi-structured

Web Documents & Wikipedia



Unstructured

- This tutorial mostly focuses on open domain **textual** QA
- In Part 6, we will discuss hybrid approaches using both KBs and text

Open-domain QA

We mostly focus on **factoid question answering**:

- Require systems to return a *short* and *concise* answer to these questions
- In contrast to other QA problems: community question answering, non-factoid or long-form question answering
- Focus more on *retrieval* and *NLU instead of generation*

How do Jellyfish function without brains or nervous systems?

Why can't humans see in the dark?

How to protect yourself from COVID-19?



QUESTION

Why do you need to bring your temperature down?



ANSWER

Up to a point, having a fever is a good thing when you're fighting an infection as in the case of sepsis (infection in the blood). Many pathogens don't fare well in even a degree or two of average raised temperature, while your body is much more resilient. It's still a pretty serious condition on its own, and sepsis is frequently fatal regardless of the not only the body's attempts to fight it, but with medical intervention.

The problems in general however, start when the fever is too high, or just high for too long. Your body will release something called chaperone molecules that help your proteins fold correctly, but there will still be errors and it's more energetically expensive. This chaperone molecules also have limits, and past a certain point your body fails on a number of levels.

Search engines: from keyword matching to question answering



Search needs a shake-up

Search engines: from keyword matching to question answering

Google X |

All News Shopping Videos Images More Settings Tools

About 13,100,000 results (0.41 seconds)

779 Accepted Papers

ACL 2020 Announces Its 779 Accepted Papers | Synced. May 20, 2020

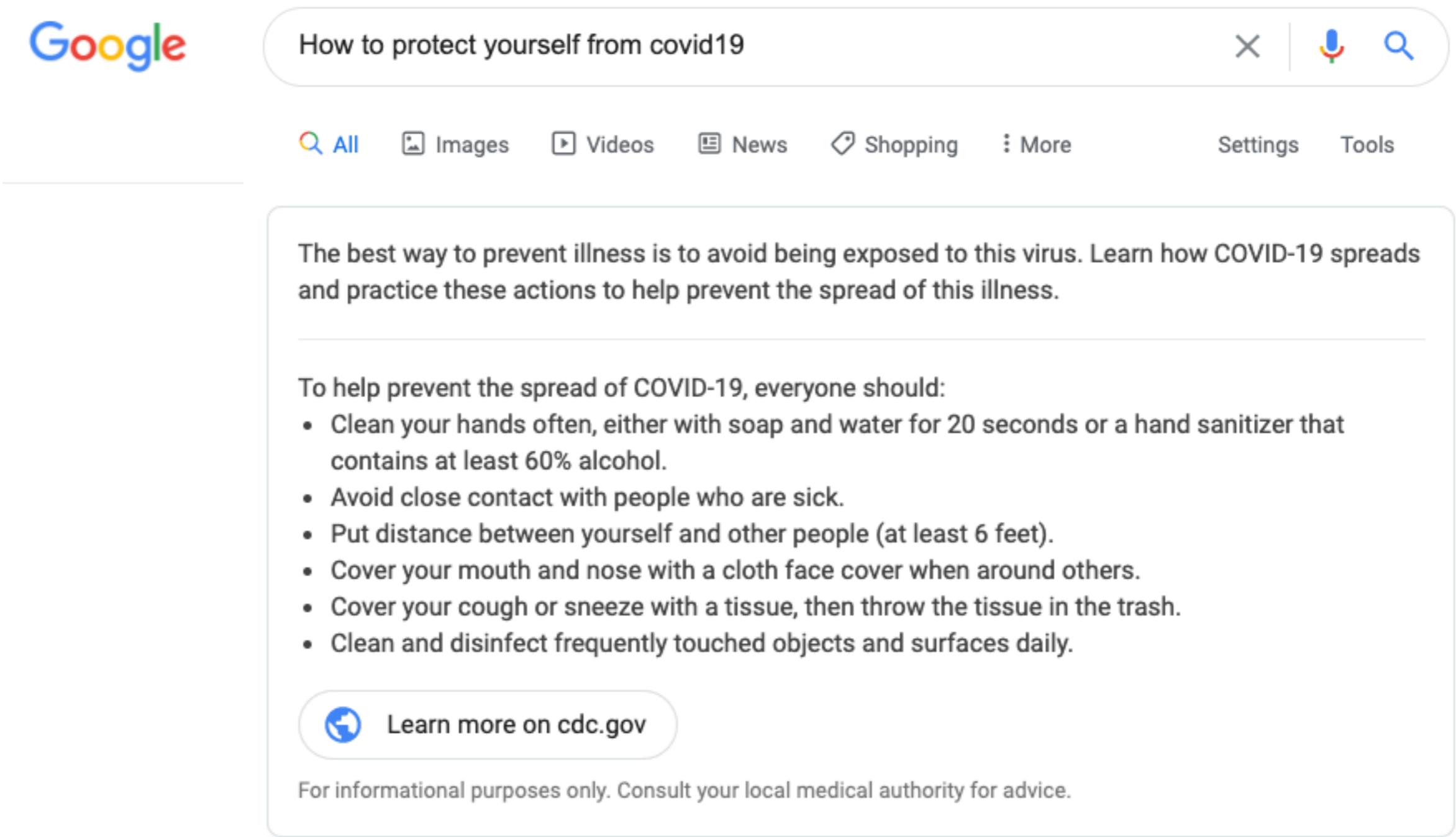
[syncedreview.com › 2020/05/20 › acl-2020-announces... ▾](https://syncedreview.com/2020/05/20/acl-2020-announces-779-accepted-papers/)

[ACL 2020 Announces Its 779 Accepted Papers | Synced](https://syncedreview.com/2020/05/20/acl-2020-announces-779-accepted-papers/)



[About Featured Snippets](#) [Feedback](#)

Search engines: from keyword matching to question answering



The screenshot shows a Google search results page. The search query "How to protect yourself from covid19" is entered in the search bar. Below the search bar are navigation links for All, Images, Videos, News, Shopping, More, Settings, and Tools. The main content area displays a snippet from the CDC website. The snippet starts with a general statement about prevention and then lists specific actions to help prevent the spread of COVID-19. At the bottom of the snippet is a button labeled "Learn more on cdc.gov". A footer note at the bottom of the page states: "For informational purposes only. Consult your local medical authority for advice."

How to protect yourself from covid19

All Images Videos News Shopping More Settings Tools

The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.

To help prevent the spread of COVID-19, everyone should:

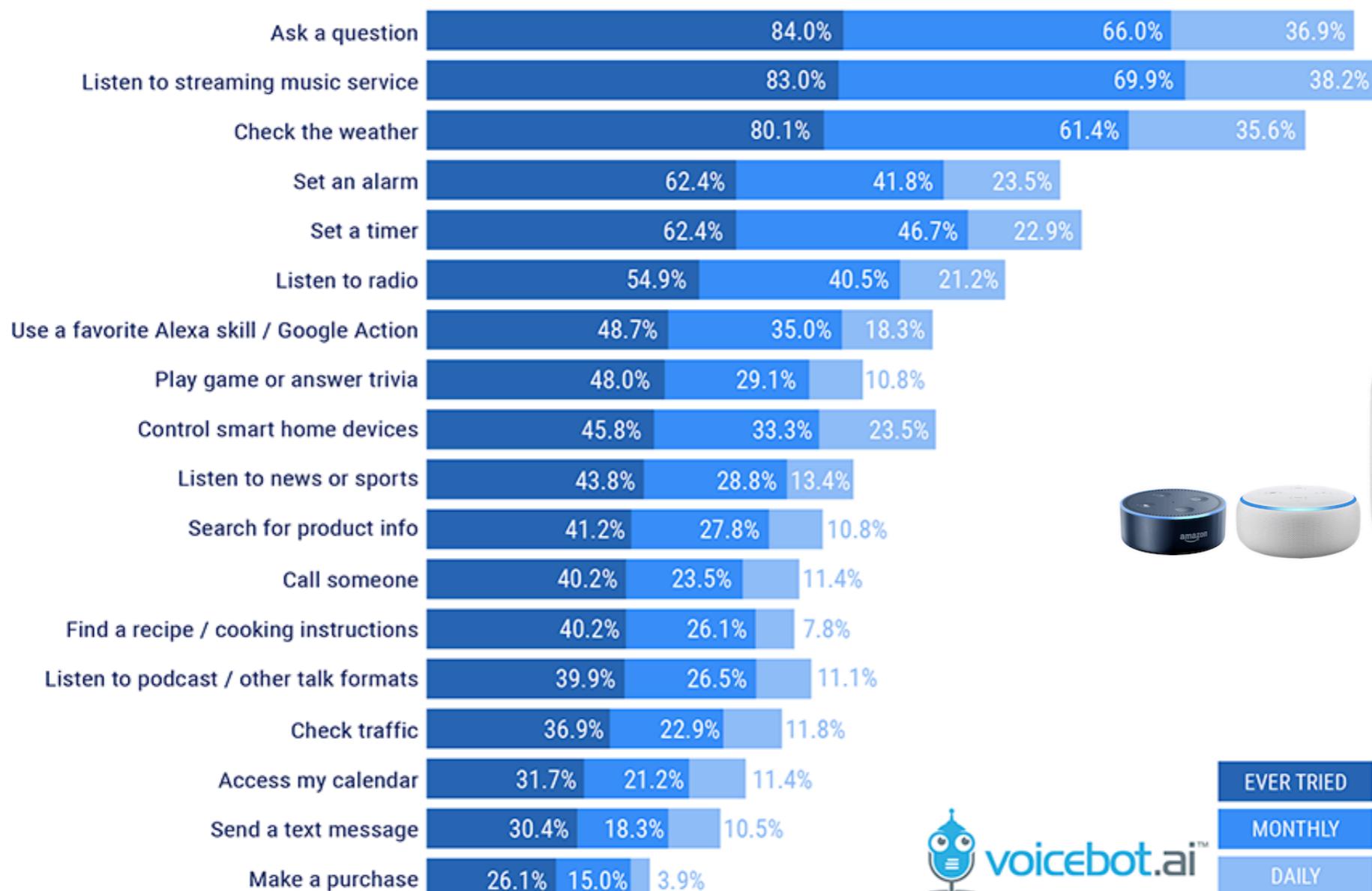
- Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
- Avoid close contact with people who are sick.
- Put distance between yourself and other people (at least 6 feet).
- Cover your mouth and nose with a cloth face cover when around others.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces daily.

Learn more on cdc.gov

For informational purposes only. Consult your local medical authority for advice.

People ask lots of questions on digital personal assistants

Smart Speaker Use Case Frequency - January 2019



Why give this tutorial today?

NAACL 2001

Open-Domain Textual Question Answering

[Sanda Harabagiu](#) and [Dan Moldovan](#)

Department of Computer Science and Engineering, Southern Methodist University

TREC QA 1999–
2001 competitions
and participant
systems

EACL 2003

Question Answering Techniques for the World Wide Web

Jimmy Lin and Boris Katz
MIT Artificial Intelligence Laboratory

WWW-based QA

NAACL 2012

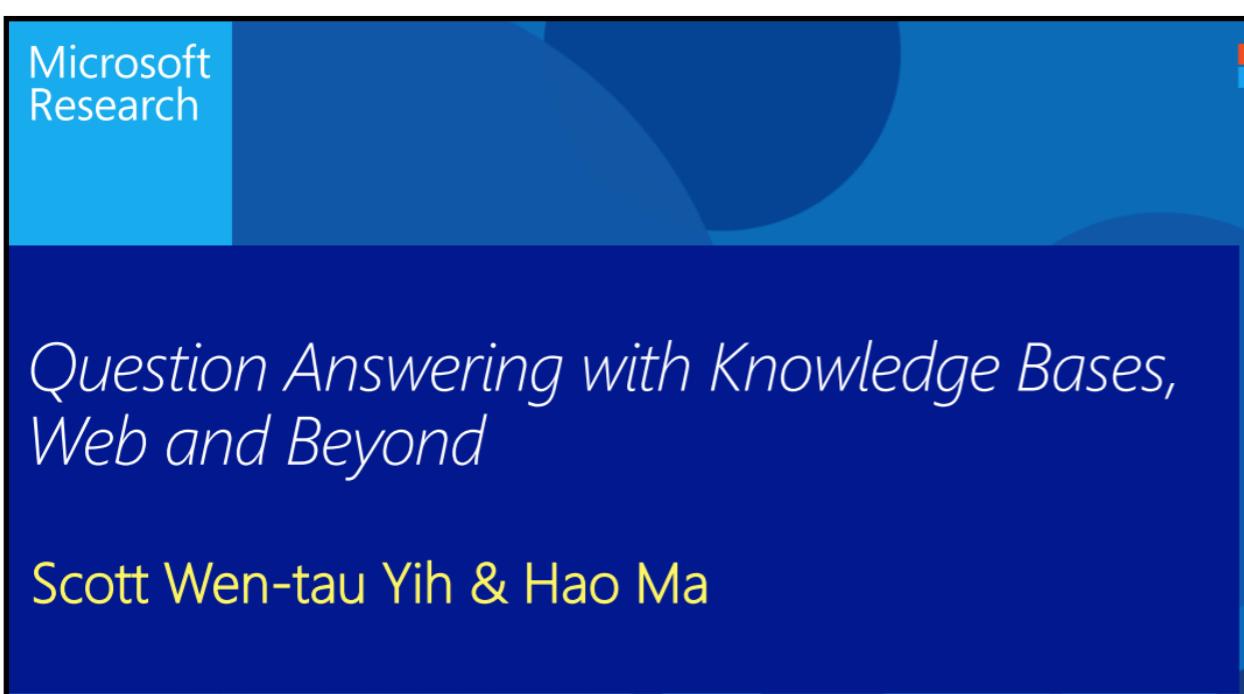
Natural Language Processing in Watson

Alfio M. Gliozzo, Aditya Kalyanpur, James Fan

IBM Watson
system on
Jeopardy questions

Why give this tutorial today?

NAACL 2016



EMNLP 2018

Standardized Tests as benchmarks for Artificial Intelligence?

Oct 31, 2018

Mrinmaya Sachan¹ Minjoon Seo² Hannaneh Hajishirzi² Eric P. Xing¹

¹Carnegie Mellon University
{mrinmays,epxing}@cs.cmu.edu

²University of Washington
{minjoon,hannaneh}@cs.washington.edu

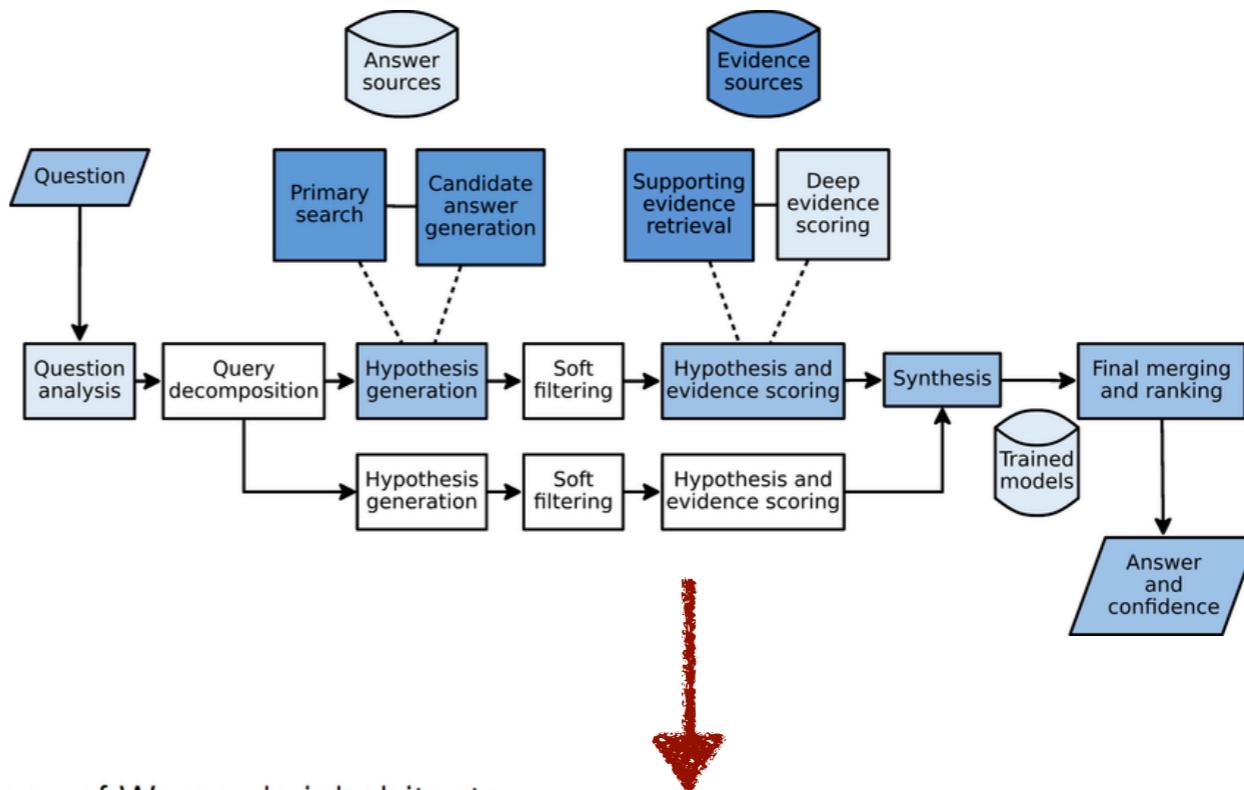
QA with KBs and tables

QA as standardized tests:
reading comprehension and
closed domain problems (science,
geometry, algebra word etc)

This tutorial: open-domain question answering over a large collection of unstructured documents; mostly the new generation and paradigm of the NLP technologies (2017-2020)

Why give this tutorial today?

Classical QA pipeline

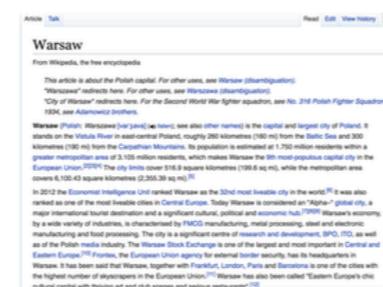


Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Two-stage Retriever-reader approaches

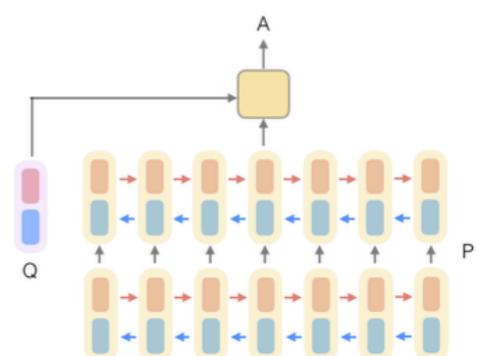


Document Retriever



Document Reader

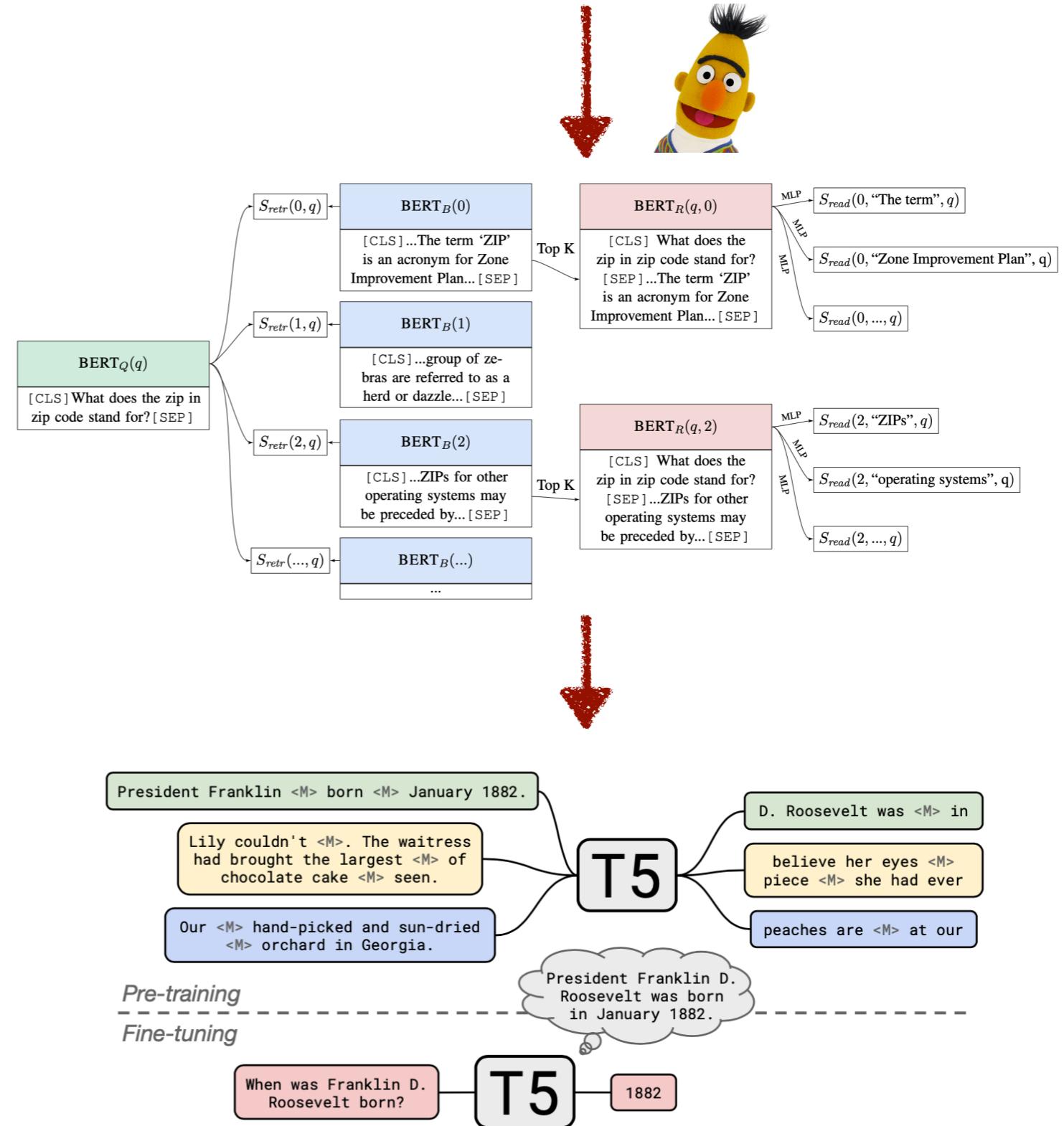
833,500



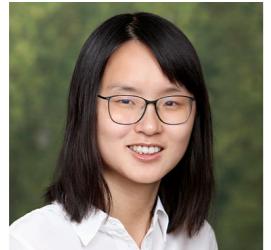
Why give this tutorial today?

End-to-end learning

Retrieval-free
models



Outline



- Part 1. Introduction *<- We are here!*



- Part 2. A history of open-domain QA
- Part 3. Datasets & evaluation
- Part 4. Two-stage retriever-reader approaches
 - ⌚ 30min coffee break
- Part 5. Dense retriever and end-to-end training
- Part 6. Retrieval-free approaches
- Part 7. Open-domain QA using KBs and text
- Part 8. Open problems and future directions

