So, You Want to Release a Dataset?

Reflections on Benchmark Development, Community Building, and Making Robust Scientific Progress

Bhaskar Mitra

Principal Researcher, Microsoft Research

@UnderdogGeek bmitra@microsoft.com

Neu-IR workshop

nn4ir

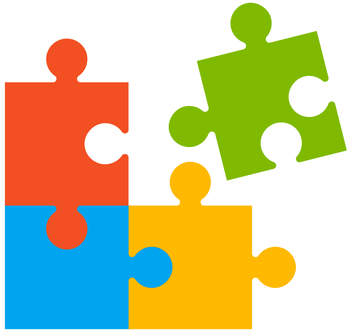
MS MARCO

Text REtrieval Conference (TREC)
...to encourage research in information retrieval from large text collections.

Why?

Releasing a dataset is a means to an end, not an end in itself

Ideally, a dataset release should be part of a larger vision to stimulate new research directions and build community



What is the larger problem you are trying to solve? Who is interested in working on this problem? Who benefits from progress on this problem? What is blocking the community from making progress?

Note: This is the task you care about, not the family of approaches or solutions you are interested in

Can you state the problem in 8 words or less?

MS MARCO

E.g., with MS MARCO our goal was to create a research agenda around “Ranking in the large-data regime”

It is true that the motivation behind MS MARCO was partly to explore deep learning methods for search, but we were keen on any methods (machine learning based or not) that can take advantage of large training datasets

What is the larger problem you are trying to solve? Who is interested in working on this problem? Who benefits from progress on this problem? What is blocking the community from making progress?

Is there an existing community of researchers who would be interested to work on this problem, or do you need to build one?

If the community exists, have you talked to them to understand what's blocking their progress?

If it doesn't exist, why not? Lack of interest or something blocking them from getting started?

What is the larger problem you are trying to solve? **Who is interested in working on this problem?** Who benefits from progress on this problem? What is blocking the community from making progress?

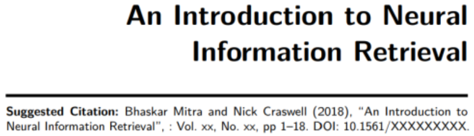
You don't need all of these to release a dataset but investing in some helps ensure you are working to close an actual gap

MS MARCO

Before there was MS MARCO…

Neu-IR workshop

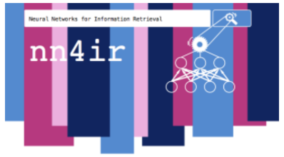
There was the Neu-IR workshop trying to gather a community of interested researchers



An Introduction to Neural Information Retrieval

Suggested Citation: Bhaskar Mitra and Nick Craswell (2018), "An Introduction to Neural Information Retrieval", : Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

There was the monograph trying to establish a common starting point and vocabulary for the community



There were several tutorials popularizing recent developments in the field

What is the larger problem you are trying to solve? Who is interested in working on this problem? Who benefits from progress on this problem? What is blocking the community from making progress?

Datasets and benchmarks create incentive structures that may lead significant section of the community down specific lanes of research

We must critically reflect on where we are leading the community to and where we are choosing not to invest (potential blind spots)

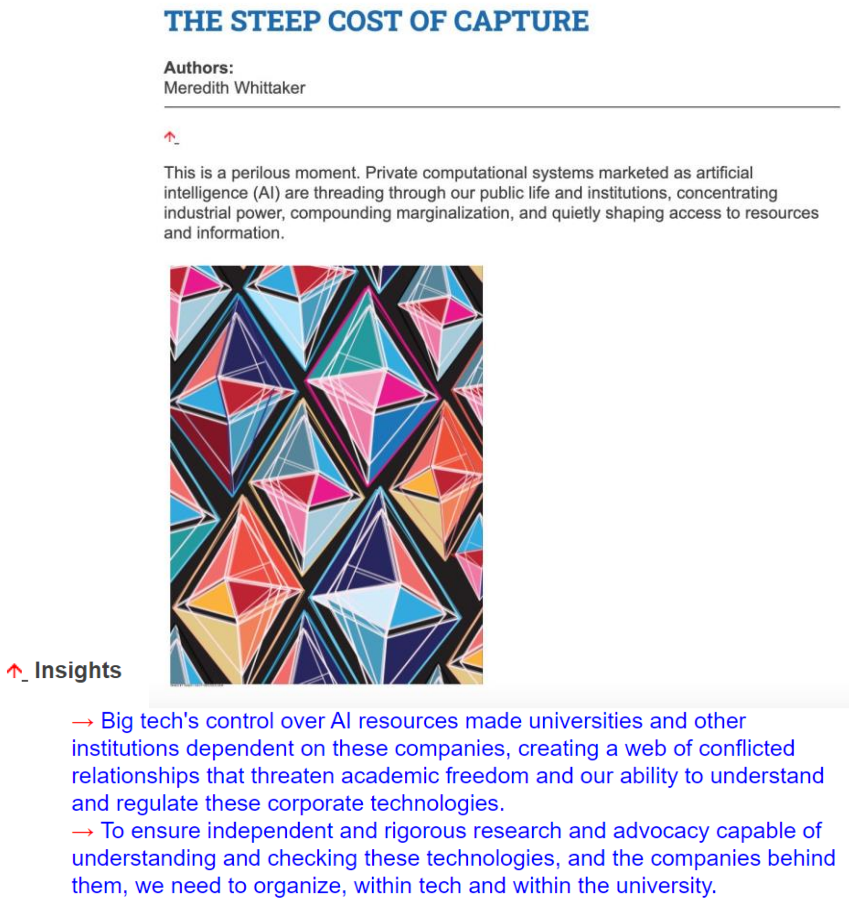
What is the larger problem you are trying to solve? Who is interested in working on this problem? **Who benefits from progress on this problem?** What is blocking the community from making progress?

E.g., for dataset release from industry, consider the different stakeholders:

Business stakeholders. Releasing datasets encourages the academic research community to make progress on problems important to business. But that may translate to disproportionate industry influence in shaping academic research agenda.

Academic stakeholders. Releasing datasets may enable new academic research but academic interests are often broader than business interests and the benchmark design should cater to those broader needs.

Society. Science is not apolitical. We should be proactive in identifying and safeguarding against potential harms and disparate benefits to different subpopulations. We are responsible for any harmful impact from technologies whose development is aided by our datasets and benchmarks. When in doubt exercise caution.



THE STEEP COST OF CAPTURE

Authors:
Meredith Whittaker

This is a perilous moment. Private computational systems marketed as artificial intelligence (AI) are threading through our public life and institutions, concentrating industrial power, compounding marginalization, and quietly shaping access to resources and information.

Insights

→ Big tech's control over AI resources made universities and other institutions dependent on these companies, creating a web of conflicted relationships that threaten academic freedom and our ability to understand and regulate these corporate technologies.

→ To ensure independent and rigorous research and advocacy capable of understanding and checking these technologies, and the companies behind them, we need to organize, within tech and within the university.

What is the larger problem you are trying to solve? Who is interested in working on this problem? **Who benefits from progress on this problem?** What is blocking the community from making progress?

Is it just the lack of datasets?

Consider that there may be other barriers to progress in the field that if left unaddressed may hinder the community from leveraging the new dataset—e.g., Does the community agree on a common problem statement? Do they have the right computational resources and software tools to quickly try different approaches? Does this problem require interdisciplinary expertise?

What is the larger problem you are trying to solve? Who is interested in working on this problem? Who benefits from progress on this problem? **What is blocking the community from making progress?**

Ethical considerations

Privacy. Does the dataset leak personally-identifiable / private information, either by itself or when cross-referenced with other datasets?

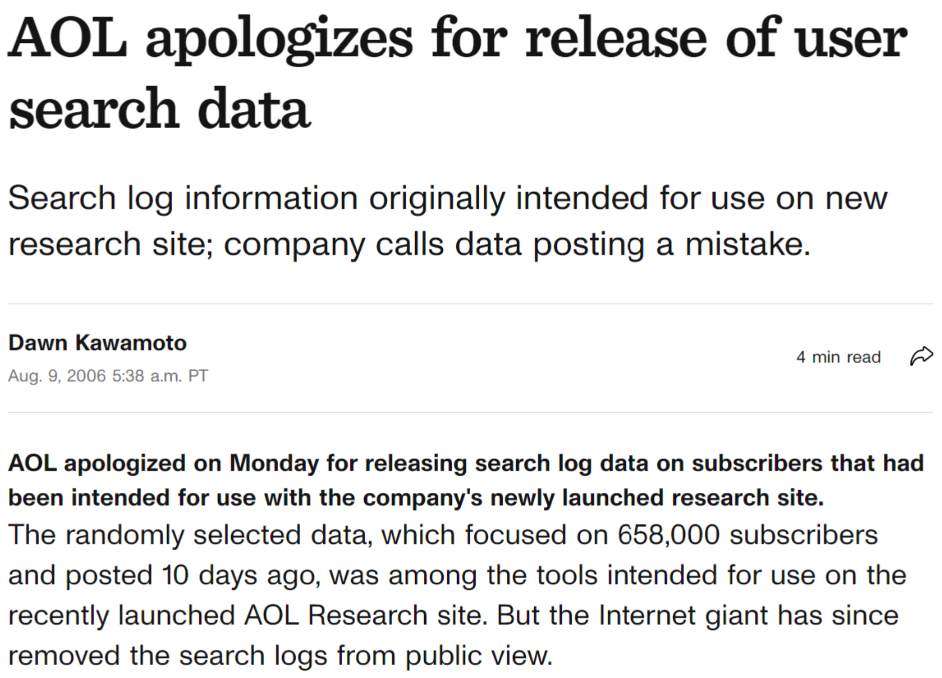
Erasure and under-representation. Does the dataset under-represent certain groups that may result in unfair disparities in quality of service in models trained on this data?

Denigration and stereotyping. Are subjects represented in ways that may be considered denigrating and/or stereotyping? Can these be further amplified by models trained on this dataset?

Politics of classification.
labels that make inhere
gender labels that assu
the right to self-identif

Misuse. Can this datase
originally anticipated d

Recourse. Are there an
be harmed by their inc



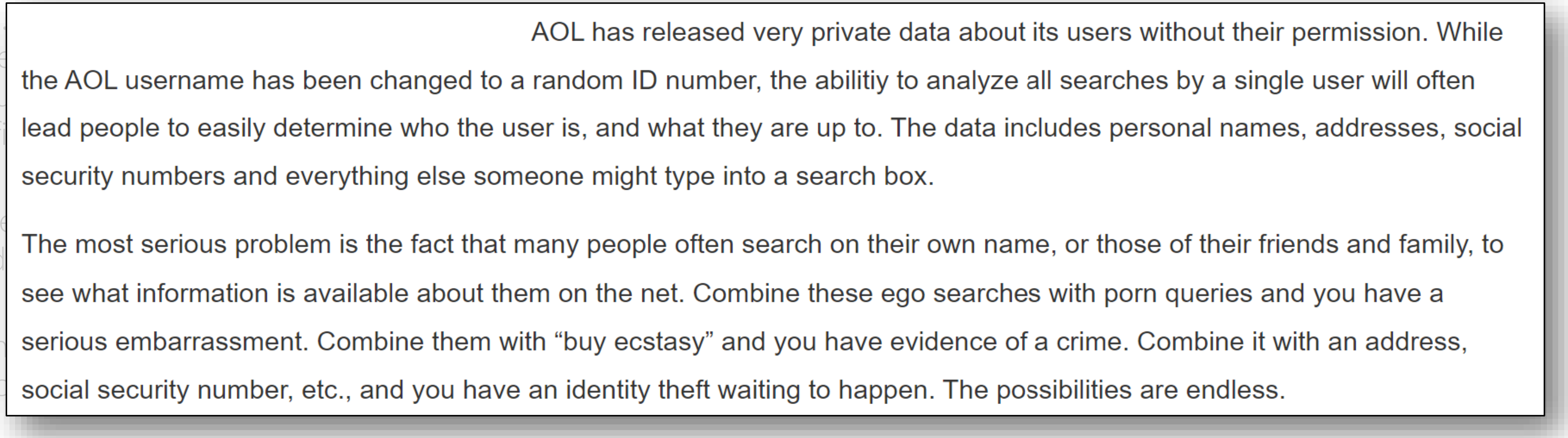
AOL apologizes for release of user search data

Search log information originally intended for use on new research site; company calls data posting a mistake.

Dawn Kawamoto
Aug. 9, 2006 5:38 a.m. PT

4 min read

AOL apologized on Monday for releasing search log data on subscribers that had been intended for use with the company's newly launched research site.
The randomly selected data, which focused on 658,000 subscribers and posted 10 days ago, was among the tools intended for use on the recently launched AOL Research site. But the Internet giant has since removed the search logs from public view.



AOL has released very private data about its users without their permission. While the AOL username has been changed to a random ID number, the abilitiy to analyze all searches by a single user will often lead people to easily determine who the user is, and what they are up to. The data includes personal names, addresses, social security numbers and everything else someone might type into a search box.

The most serious problem is the fact that many people often search on their own name, or those of their friends and family, to see what information is available about them on the net. Combine these ego searches with porn queries and you have a serious embarrassment. Combine them with “buy ecstasy” and you have evidence of a crime. Combine it with an address, social security number, etc., and you have an identity theft waiting to happen. The possibilities are endless.

Privacy. Does the dataset leak personally-identifiable / private information, either by itself or when cross-referenced with other datasets?

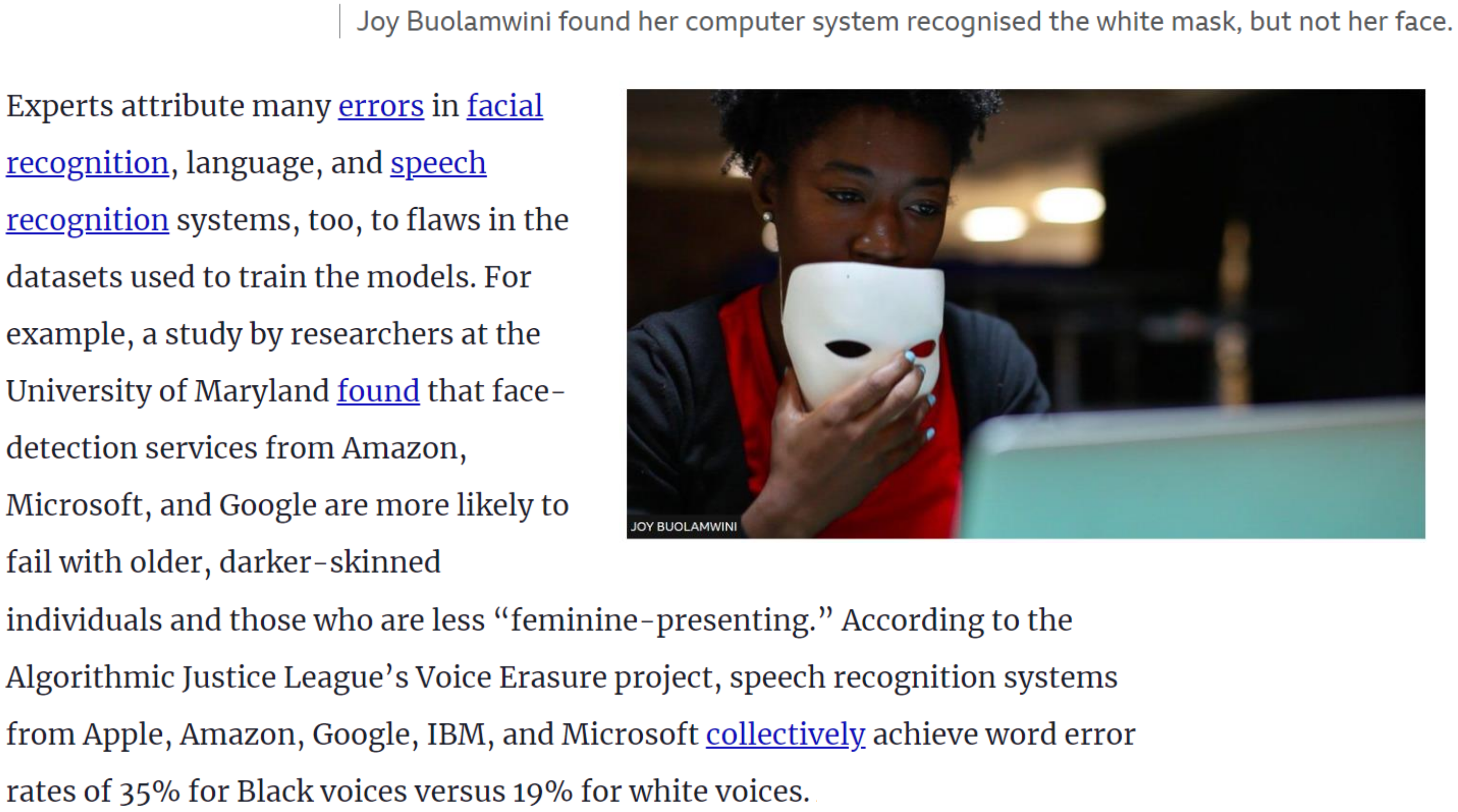
Erasure and under-representation. Does the dataset under-represent certain groups that may result in unfair disparities in quality of service in models trained on this data?

Denigration and stereotyping. Are subjects represented in ways that may be considered denigrating and/or stereotyping? Can these be further amplified by models trained on this dataset?

Politics of classification. Does this dataset contain class labels that make inherently harmful assumptions—e.g., gender labels that assume gender is binary or denies the right to self-identification?

Misuse. Can this dataset be abused for purposes not originally anticipated during its design?

Recourse. Are there any recourse for subjects who may be harmed by their inclusion/exclusion in the dataset?



Joy Buolamwini found her computer system recognised the white mask, but not her face.

JOY BUOLAMWINI

Experts attribute many errors in facial recognition, language, and speech recognition systems, too, to flaws in the datasets used to train the models. For example, a study by researchers at the University of Maryland found that face-detection services from Amazon, Microsoft, and Google are more likely to fail with older, darker-skinned individuals and those who are less "feminine-presenting." According to the Algorithmic Justice League's Voice Erasure project, speech recognition systems from Apple, Amazon, Google, IBM, and Microsoft collectively achieve word error rates of 35% for Black voices versus 19% for white voices.

Privacy. Does the dataset leak personally-identifiable / private information, either by itself or when cross-referenced with other datasets?

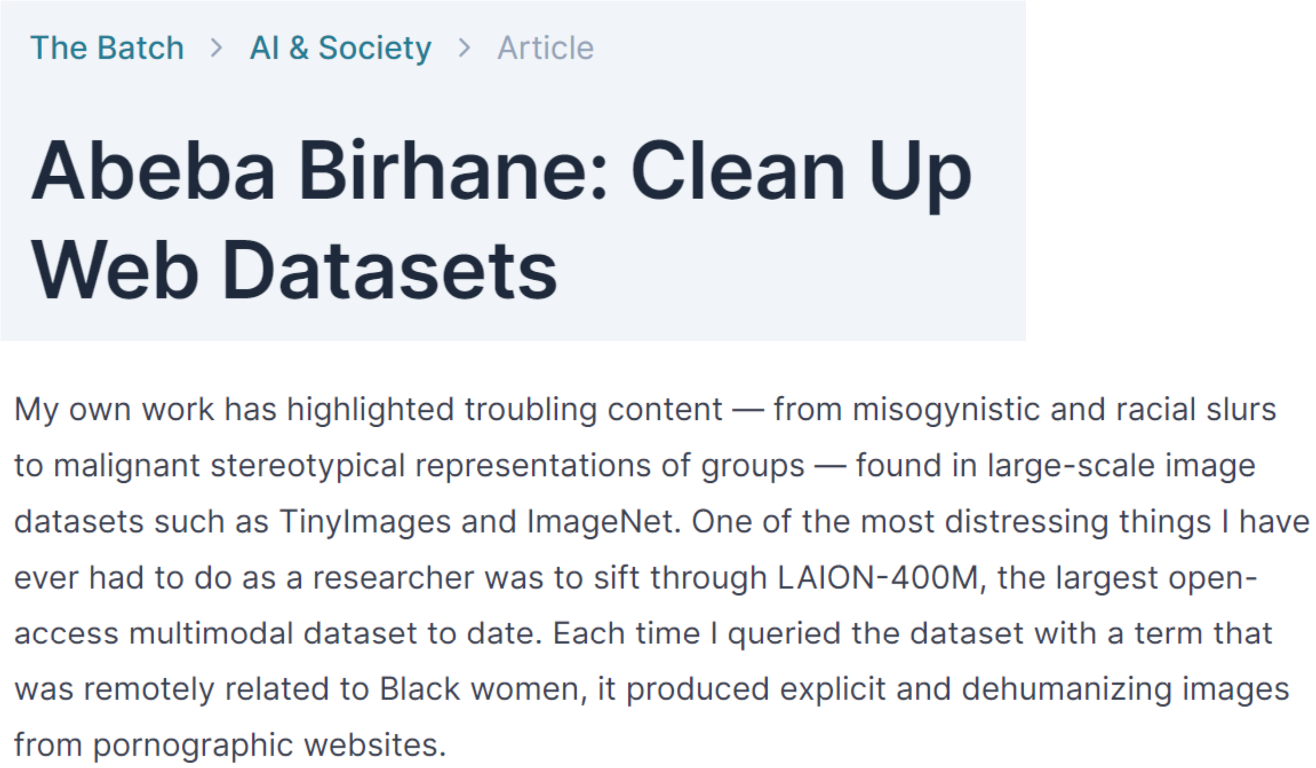
Erasure and under-representation. Does the dataset under-represent certain groups that may result in unfair disparities in quality of service in models trained on this data?

Denigration and stereotyping. Are subjects represented in ways that may be considered denigrating and/or stereotyping? Can these be further amplified by models trained on this dataset?

Politics of classification. Does this dataset contain class labels that make inherently harmful assumptions—e.g., gender labels that assume gender is binary or denies the right to self-identification?

Misuse. Can this dataset be abused for purposes not originally anticipated during its design?

Recourse. Are there any recourse for subjects who may be harmed by their inclusion/exclusion in the dataset?



The Batch > AI & Society > Article

Abeba Birhane: Clean Up Web Datasets

My own work has highlighted troubling content — from misogynistic and racial slurs to malignant stereotypical representations of groups — found in large-scale image datasets such as TinyImages and ImageNet. One of the most distressing things I have ever had to do as a researcher was to sift through LAION-400M, the largest open-access multimodal dataset to date. Each time I queried the dataset with a term that was remotely related to Black women, it produced explicit and dehumanizing images from pornographic websites.

Privacy. Does the dataset leak personally-identifiable / private information, either by itself or when cross-referenced with other datasets?

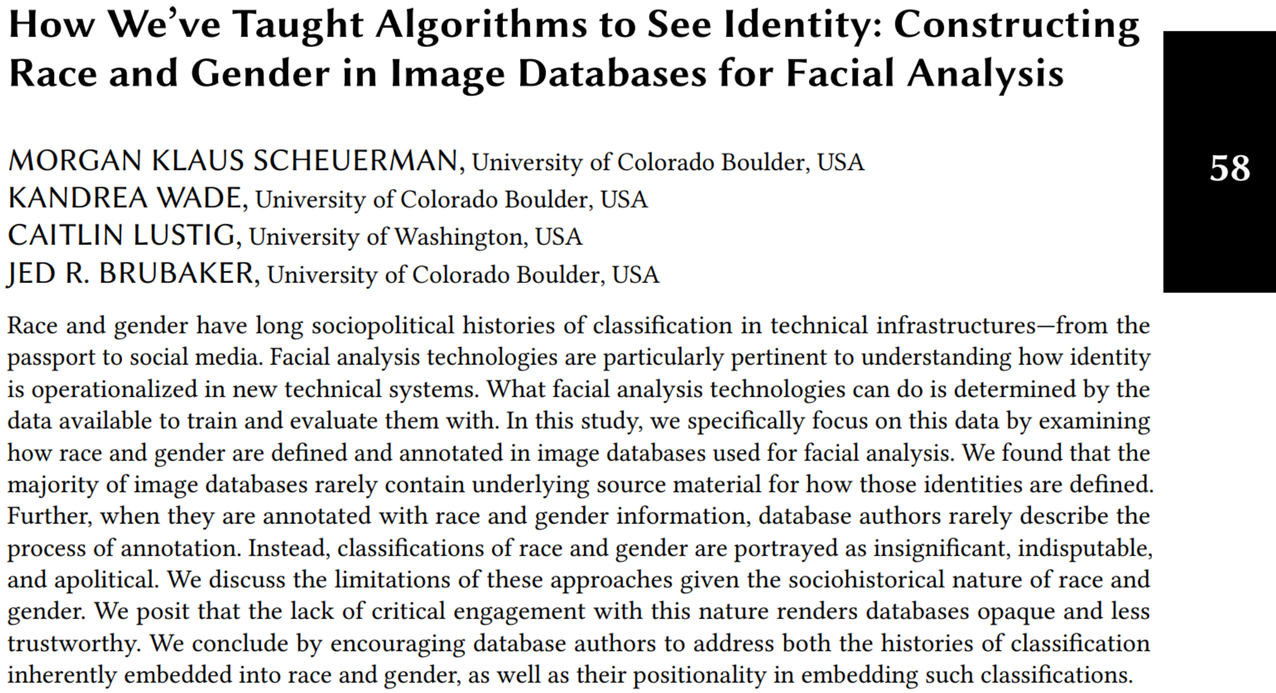
Erasure and under-representation. Does the dataset under-represent certain groups that may result in unfair disparities in quality of service in models trained on this data?

Denigration and stereotyping. Are subjects represented in ways that may be considered denigrating and/or stereotyping? Can these be further amplified by models trained on this dataset?

Politics of classification. Does this dataset contain class labels that make inherently harmful assumptions—e.g., inferred gender labels that assume gender is binary and/or denies the right to self-identification?

Misuse. Can this dataset be abused for purposes not originally anticipated during its design?

Recourse. Are there any recourse for subjects who may be harmed by their inclusion/exclusion in the dataset?

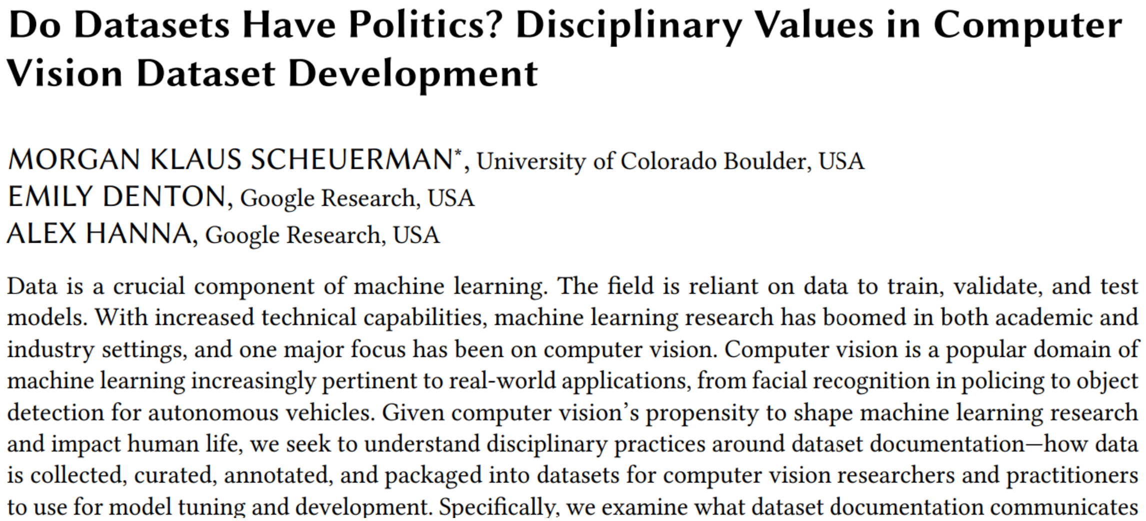


58

How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis

MORGAN KLAUS SCHEUERMAN, University of Colorado Boulder, USA
KANDREA WADE, University of Colorado Boulder, USA
CAITLIN LUSTIG, University of Washington, USA
JED R. BRUBAKER, University of Colorado Boulder, USA

Race and gender have long sociopolitical histories of classification in technical infrastructures—from the passport to social media. Facial analysis technologies are particularly pertinent to understanding how identity is operationalized in new technical systems. What facial analysis technologies can do is determined by the data available to train and evaluate them with. In this study, we specifically focus on this data by examining how race and gender are defined and annotated in image databases used for facial analysis. We found that the majority of image databases rarely contain underlying source material for how those identities are defined. Further, when they are annotated with race and gender information, database authors rarely describe the process of annotation. Instead, classifications of race and gender are portrayed as insignificant, indisputable, and apolitical. We discuss the limitations of these approaches given the sociohistorical nature of race and gender. We posit that the lack of critical engagement with this nature renders databases opaque and less trustworthy. We conclude by encouraging database authors to address both the histories of classification inherently embedded into race and gender, as well as their positionality in embedding such classifications.



Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development

MORGAN KLAUS SCHEUERMAN*, University of Colorado Boulder, USA
EMILY DENTON, Google Research, USA
ALEX HANNA, Google Research, USA

Data is a crucial component of machine learning. The field is reliant on data to train, validate, and test models. With increased technical capabilities, machine learning research has boomed in both academic and industry settings, and one major focus has been on computer vision. Computer vision is a popular domain of machine learning increasingly pertinent to real-world applications, from facial recognition in policing to object detection for autonomous vehicles. Given computer vision's propensity to shape machine learning research and impact human life, we seek to understand disciplinary practices around dataset documentation—how data is collected, curated, annotated, and packaged into datasets for computer vision researchers and practitioners to use for model tuning and development. Specifically, we examine what dataset documentation communicates

Privacy. Does the dataset leak personally-identifiable / private information, either by itself or when cross-referenced with other datasets?

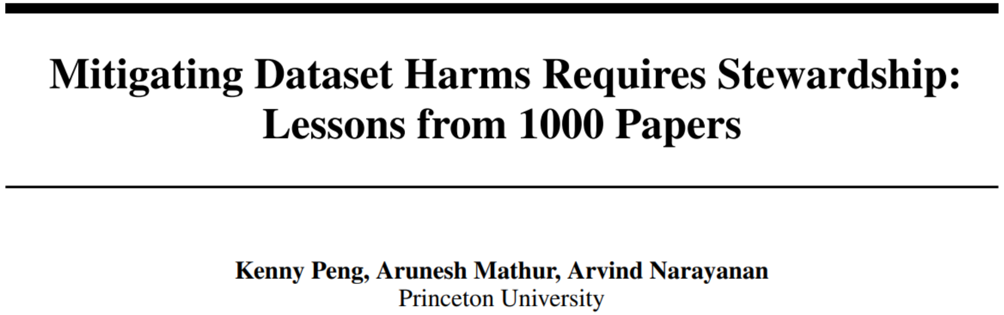
Erasure and under-representation. Does the dataset under-represent certain groups that may result in unfair disparities in quality of service in models trained on this data?

Denigration and stereotyping. Are subjects represented in ways that may be considered denigrating and/or stereotyping? Can these be further amplified by models trained on this dataset?

Politics of classification. Does this dataset contain class labels that make inherently harmful assumptions—e.g., gender labels that assume gender is binary or denies the right to self-identification?

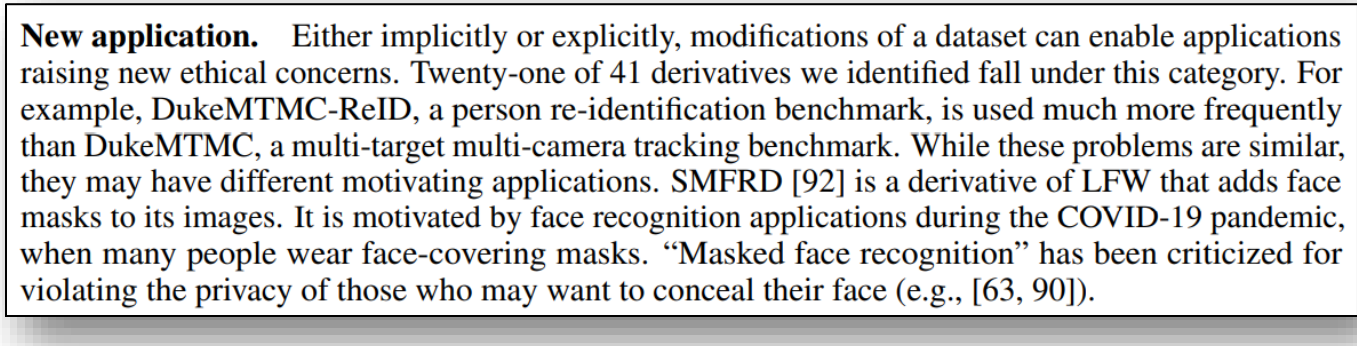
Misuse. Can this dataset be abused for purposes not originally anticipated during its design?

Recourse. Are there any recourse for subjects who may be harmed by their inclusion/exclusion in the dataset?



Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers

Kenny Peng, Arunesh Mathur, Arvind Narayanan
Princeton University



New application. Either implicitly or explicitly, modifications of a dataset can enable applications raising new ethical concerns. Twenty-one of 41 derivatives we identified fall under this category. For example, DukeMTMC-ReID, a person re-identification benchmark, is used much more frequently than DukeMTMC, a multi-target multi-camera tracking benchmark. While these problems are similar, they may have different motivating applications. SMFRD [92] is a derivative of LFW that adds face masks to its images. It is motivated by face recognition applications during the COVID-19 pandemic, when many people wear face-covering masks. "Masked face recognition" has been criticized for violating the privacy of those who may want to conceal their face (e.g., [63, 90]).

Privacy. Does the dataset leak personally-identifiable / private information, either by itself or when cross-referenced with other datasets?

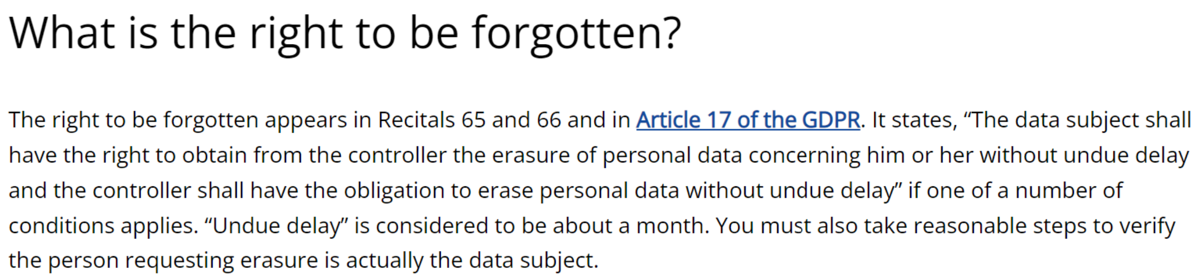
Erasure and under-representation. Does the dataset under-represent certain groups that may result in unfair disparities in quality of service in models trained on this data?

Denigration and stereotyping. Are subjects represented in ways that may be considered denigrating and/or stereotyping? Can these be further amplified by models trained on this dataset?

Politics of classification. Does this dataset contain class labels that make inherently harmful assumptions—e.g., gender labels that assume gender is binary or denies the right to self-identification?

Misuse. Can this dataset be abused for purposes not originally anticipated during its design?

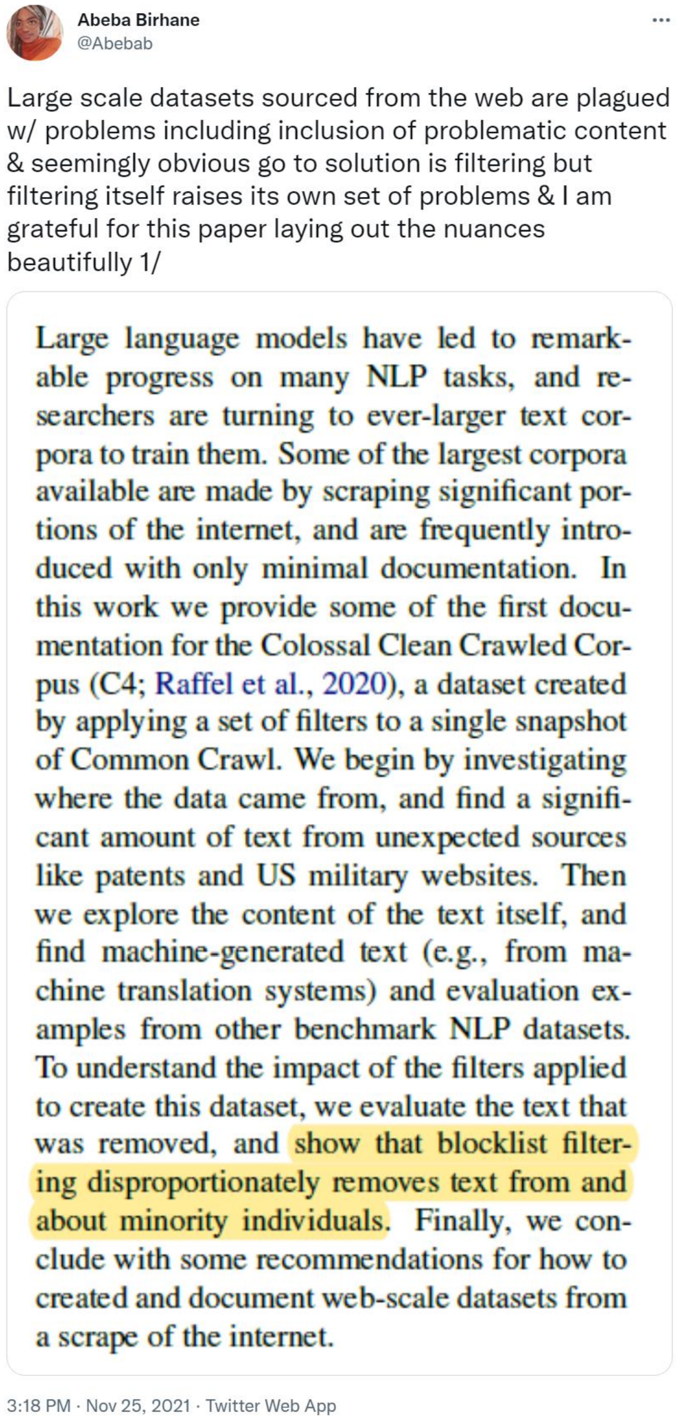
Recourse. Are there any recourse for subjects who may be harmed by their inclusion/exclusion in the dataset?



What is the right to be forgotten?

The right to be forgotten appears in Recitals 65 and 66 and in Article 17 of the GDPR. It states, “The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay” if one of a number of conditions applies. “Undue delay” is considered to be about a month. You must also take reasonable steps to verify the person requesting erasure is actually the data subject.

Mitigation strategies may range from technical solutions (e.g., enforcing k-anonymity for privacy) to legal protections (e.g., disallowing use of the data in commercial applications via data license) but be cautious about simple solutions that themselves may cause secondary harms



Abeba Birhane
@Abebab

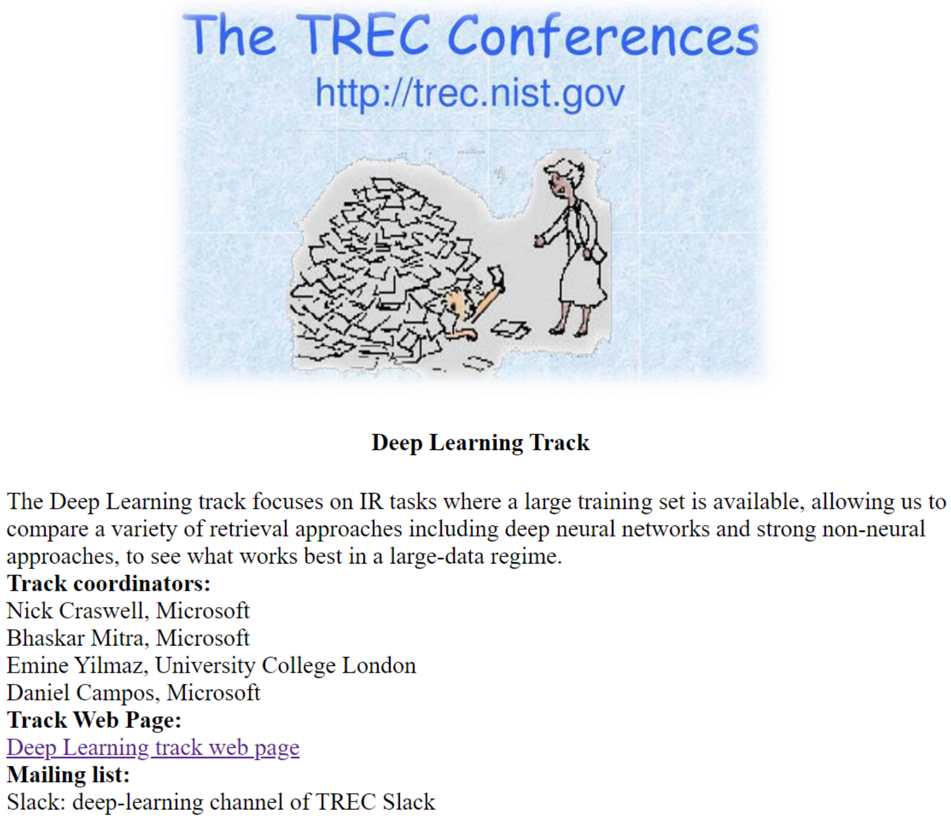
Large scale datasets sourced from the web are plagued w/ problems including inclusion of problematic content & seemingly obvious go to solution is filtering but filtering itself raises its own set of problems & I am grateful for this paper laying out the nuances beautifully 1/

Large language models have led to remarkable progress on many NLP tasks, and researchers are turning to ever-larger text corpora to train them. Some of the largest corpora available are made by scraping significant portions of the internet, and are frequently introduced with only minimal documentation. In this work we provide some of the first documentation for the Colossal Clean Crawled Corpus (C4; Raffel et al., 2020), a dataset created by applying a set of filters to a single snapshot of Common Crawl. We begin by investigating where the data came from, and find a significant amount of text from unexpected sources like patents and US military websites. Then we explore the content of the text itself, and find machine-generated text (e.g., from machine translation systems) and evaluation examples from other benchmark NLP datasets. To understand the impact of the filters applied to create this dataset, we evaluate the text that was removed, and show that blocklist filtering disproportionately removes text from and about minority individuals. Finally, we conclude with some recommendations for how to created and document web-scale datasets from a scrape of the internet.

3:18 PM · Nov 25, 2021 · Twitter Web App

Scientific rigor

MS MARCO



The TREC Conferences

http://trec.nist.gov

Deep Learning Track

The Deep Learning track focuses on IR tasks where a large training set is available, allowing us to compare a variety of retrieval approaches including deep neural networks and strong non-neural approaches, to see what works best in a large-data regime.
Track coordinators:
Nick Craswell, Microsoft
Bhaskar Mitra, Microsoft
Emine Yilmaz, University College London
Daniel Campos, Microsoft
Track Web Page:
Deep Learning track web page
Mailing list:
Slack: deep-learning channel of TREC Slack

Three evaluation protocols:

MS MARCO leaderboard. Participants have access to the test queries but not to the corresponding ground truth labels. Participants can submit runs around the year. The same set of sparse relevance labels are employed for all run evaluation.

TREC Deep Learning Track. The track runs annually and provides the participants with a new test query sets each year. All participants submit their runs by the August deadline. Results are pooled across submitted runs and judged by NIST assessors after the deadline and used for run evaluation.

Offline evaluation with old TREC datasets. Each year the TREC track releases the test labels as reusable benchmarks for the community. Benchmark users are expected to follow appropriate protocols for their own experiments but left to their discretion.

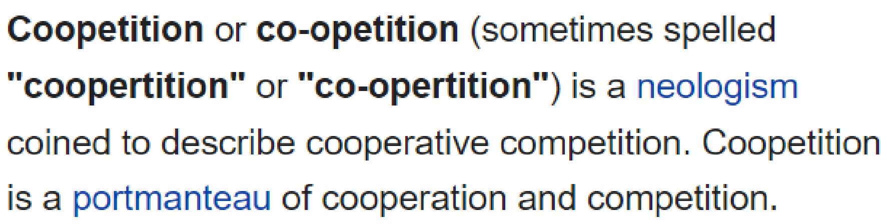
Internal validity. Would improvements on the current dataset hold on a different sample from the same dataset for the same task?

MS MARCO leaderboard allows multiple submissions which over time can make the evaluation less reliable due to multiple testing

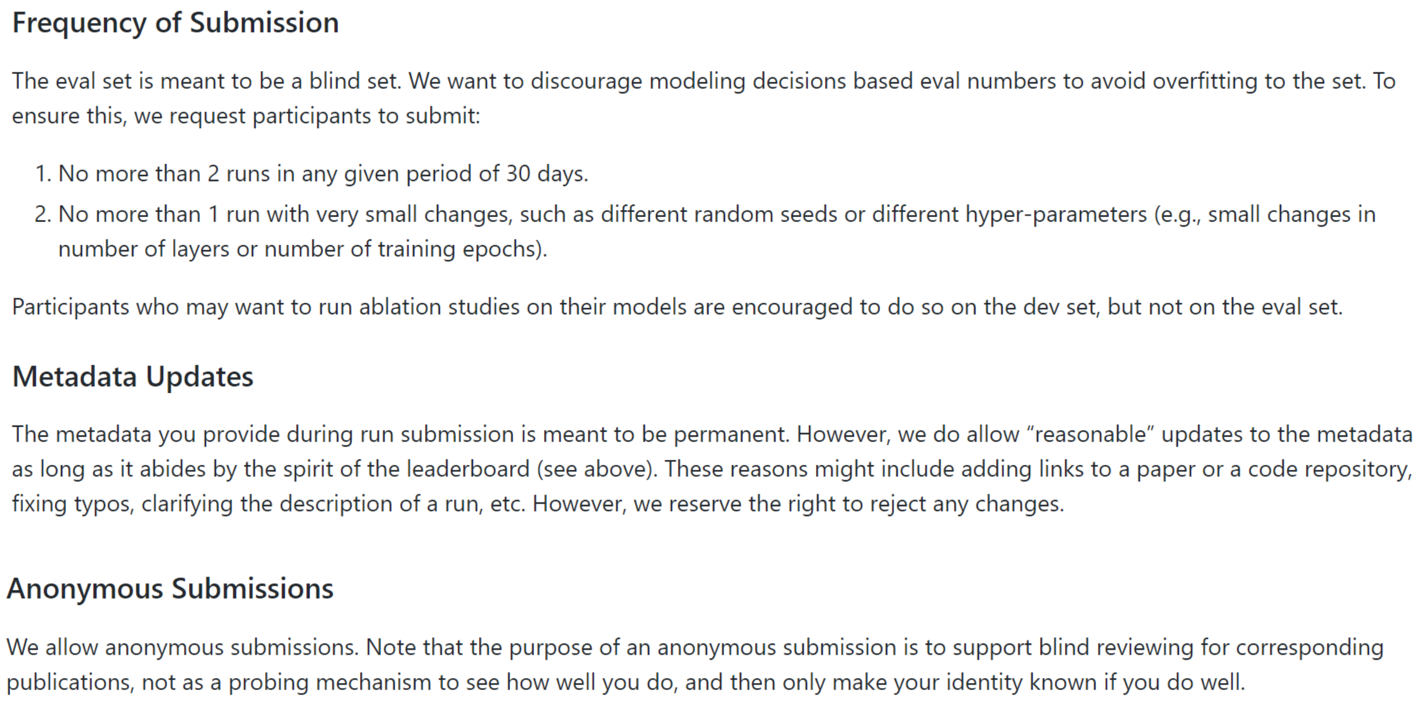
Best practice for avoiding multiple testing → participate at TREC (single-shot submission + pooled judgments)

Least robust (but most flexible): Reuse TREC test set from previous year for offline evaluation—but useful for publication if we follow strict experiment protocols

To improve internal validity of the leaderboard-based evaluation we enforce some strict policies:



Coopetition or **co-opetition** (sometimes spelled **"coopertition"** or **"co-opertition"**) is a neologism coined to describe cooperative competition. Coopetition is a portmanteau of cooperation and competition.



Frequency of Submission

The eval set is meant to be a blind set. We want to discourage modeling decisions based eval numbers to avoid overfitting to the set. To ensure this, we request participants to submit:

1. No more than 2 runs in any given period of 30 days.
2. No more than 1 run with very small changes, such as different random seeds or different hyper-parameters (e.g., small changes in number of layers or number of training epochs).

Participants who may want to run ablation studies on their models are encouraged to do so on the dev set, but not on the eval set.

Metadata Updates

The metadata you provide during run submission is meant to be permanent. However, we do allow “reasonable” updates to the metadata as long as it abides by the spirit of the leaderboard (see above). These reasons might include adding links to a paper or a code repository, fixing typos, clarifying the description of a run, etc. However, we reserve the right to reject any changes.

Anonymous Submissions

We allow anonymous submissions. Note that the purpose of an anonymous submission is to support blind reviewing for corresponding publications, not as a probing mechanism to see how well you do, and then only make your identity known if you do well.

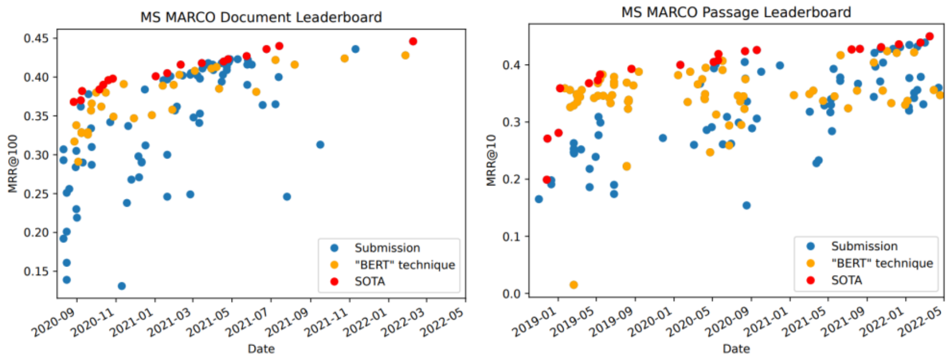


Figure 2: Overview of the MS MARCO document ranking leaderboard (left) and passage ranking leaderboard (right). Each point represents a submission, plotted with its submission date and effectiveness: orange points denote model descriptions that contain the string “BERT” and red points capture improvements in the “state of the art” over time.

But how stable are the leaderboard rankings?

Under bootstrap analysis we find the leaderboard rankings are fairly stable! 😊 👍

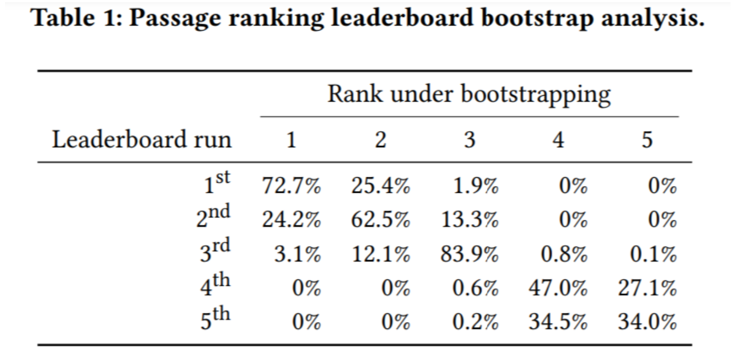


Table 1: Passage ranking leaderboard bootstrap analysis.

	Rank under bootstrapping				
Leaderboard run	1	2	3	4	5
1st	72.7%	25.4%	1.9%	0%	0%
2nd	24.2%	62.5%	13.3%	0%	0%
3rd	3.1%	12.1%	83.9%	0.8%	0.1%
4th	0%	0%	0.6%	47.0%	27.1%
5th	0%	0%	0.2%	34.5%	34.0%

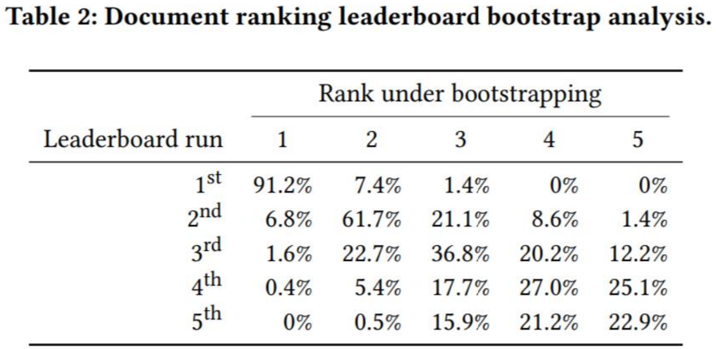


Table 2: Document ranking leaderboard bootstrap analysis.

	Rank under bootstrapping				
Leaderboard run	1	2	3	4	5
1st	91.2%	7.4%	1.4%	0%	0%
2nd	6.8%	61.7%	21.1%	8.6%	1.4%
3rd	1.6%	22.7%	36.8%	20.2%	12.2%
4th	0.4%	5.4%	17.7%	27.0%	25.1%
5th	0%	0.5%	15.9%	21.2%	22.9%

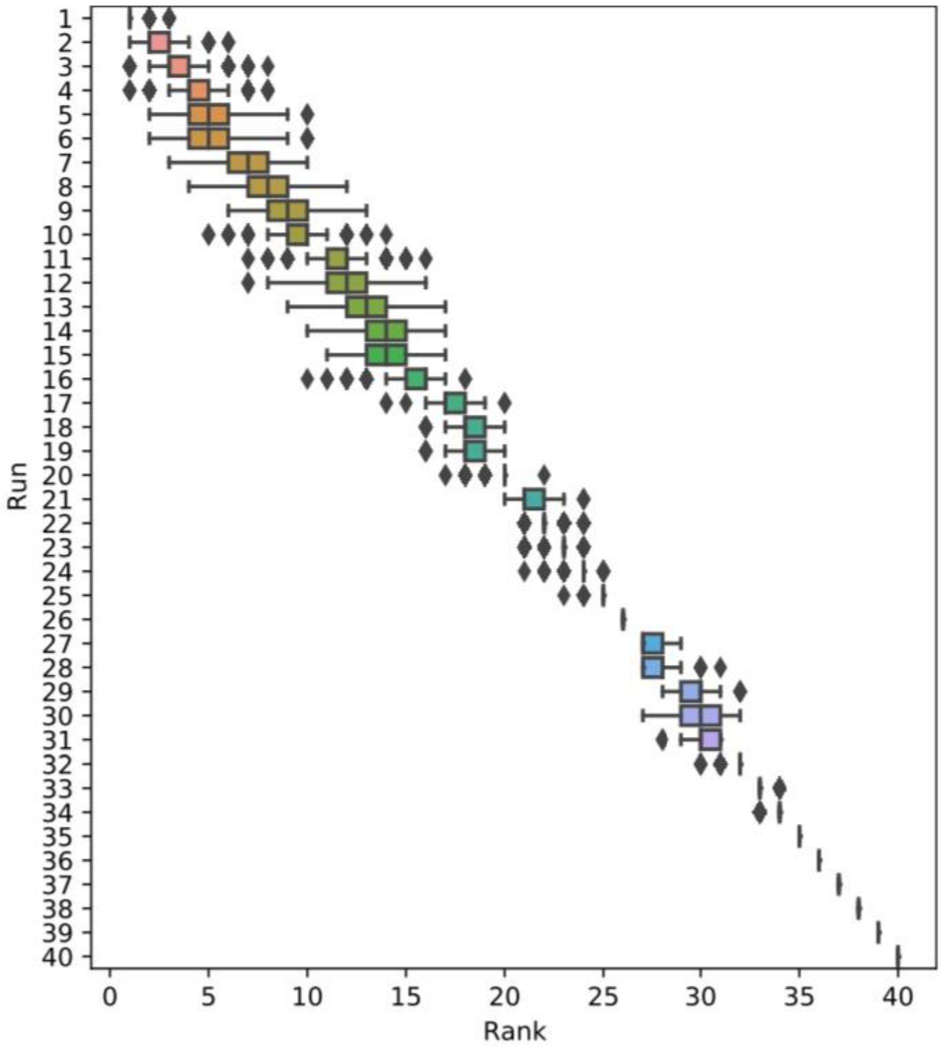


Figure 3: Full results of document leaderboard bootstrap. Runs 1–5 show the same results as Table 2.

Private leaderboard

We included 45 TREC 2020 queries in the document ranking eval set

The top leaderboard run has a more “spread out” rank on the TREC queries and is overtaken by the best TREC 2020

This may be due to distribution difference between the two test sets or the smaller size of the TREC set

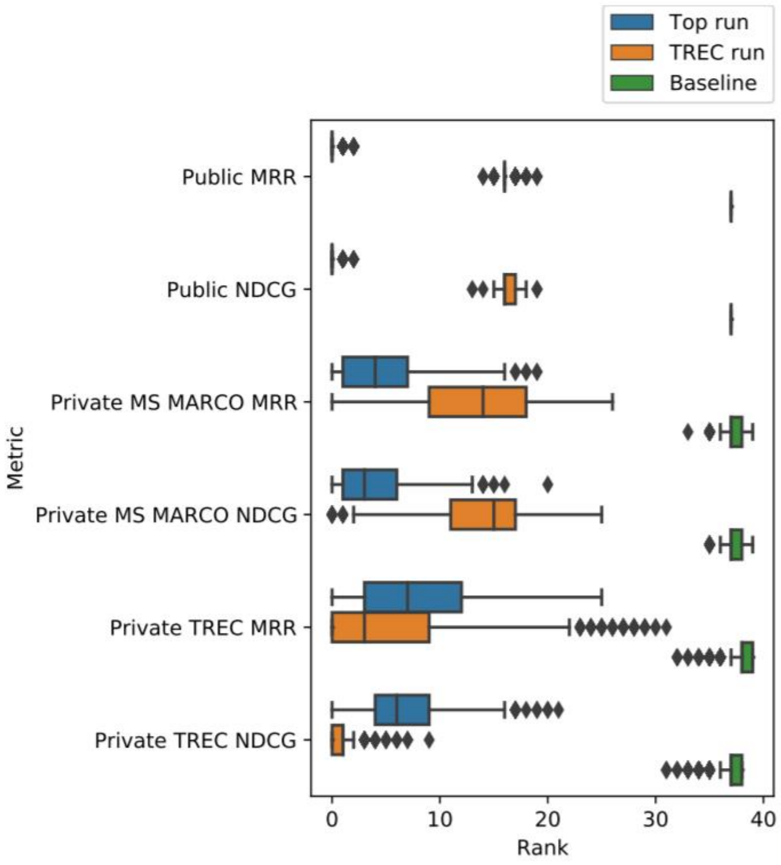
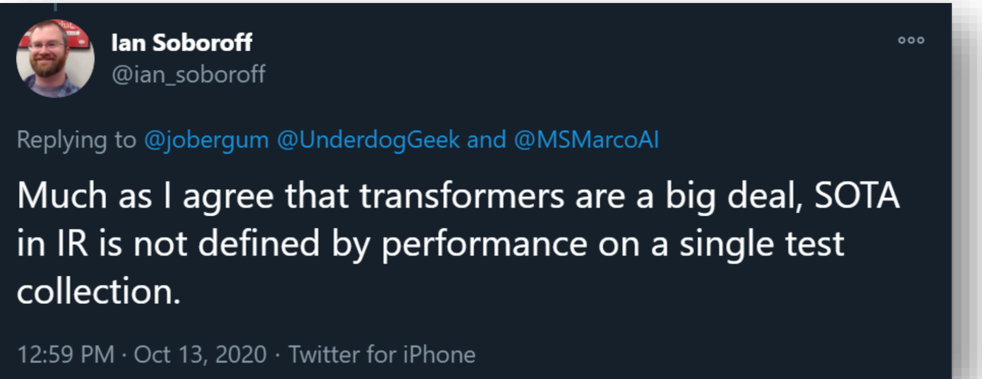


Figure 4: Rank positions of three leaderboard runs under bootstrapping. Metrics are MRR and NDCG@10. The query-sets are the 5,793 Public leaderboard queries and the 45 Private leaderboard queries from TREC-2020. The Private queries can be evaluated with sparse MS MARCO labels or comprehensive TREC labels.

External validity. Would improvements on the current dataset hold on a different dataset with different distribution for the same task or on a different (but closely related) task?



Ian Soboroff
@ian_soboroff

Replying to @jobergum @UnderdogGeek and @MSMarcoAI

Much as I agree that transformers are a big deal, SOTA in IR is not defined by performance on a single test collection.

12:59 PM · Oct 13, 2020 · Twitter for iPhone

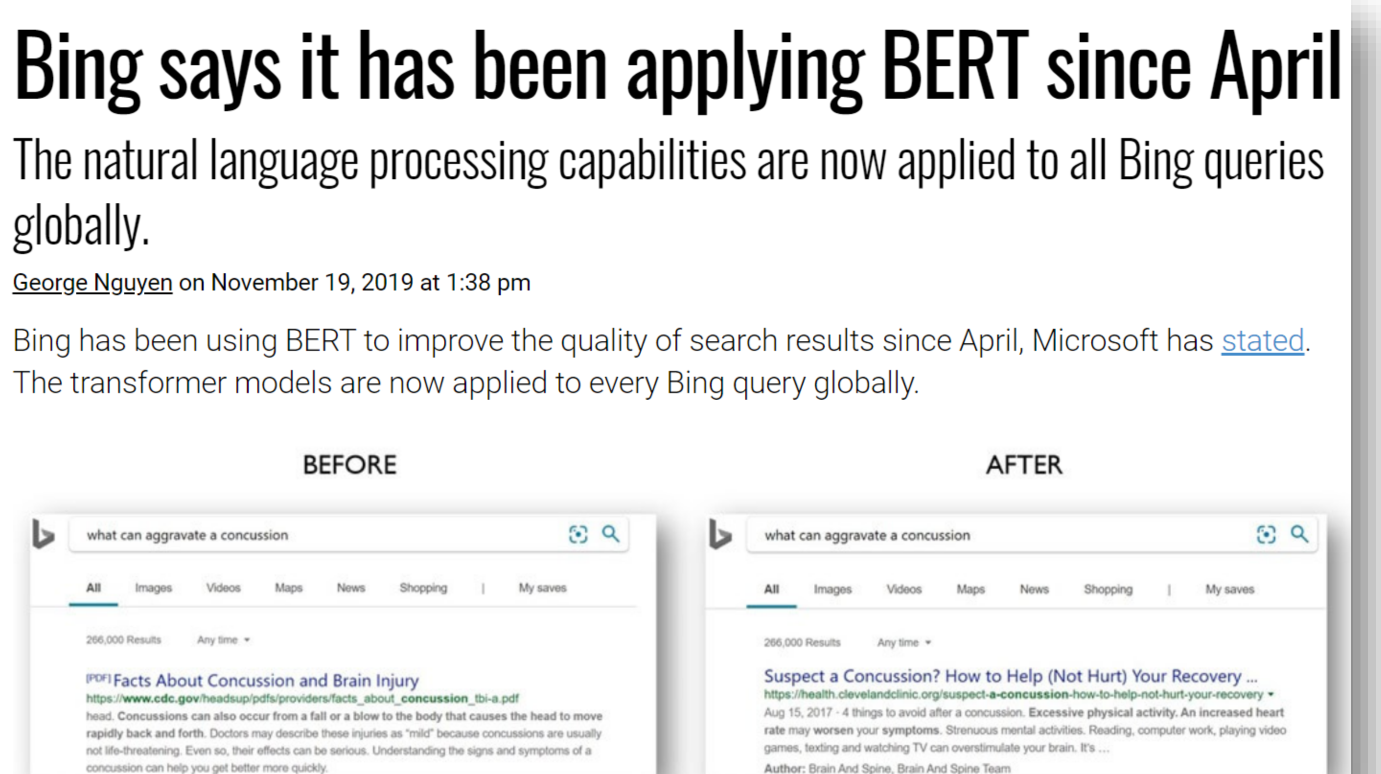
If MS MARCO's training data is only useful for achieving good results on MS MARCO's test set, then it's less useful for the IR community

Important: transfer learning from MS MARCO to other benchmarks

- TREC DL is transfer learning (MS MARCO sparse binary labels → NIST's 5-point labels)
- Promising results: MS MARCO → Robust04, TREC-COVID, TREC-CAsT

Industry impact

BERT-scale deep ranking models in production search systems



Bing says it has been applying BERT since April

The natural language processing capabilities are now applied to all Bing queries globally.

George Nguyen on November 19, 2019 at 1:38 pm

Bing has been using BERT to improve the quality of search results since April, Microsoft has stated. The transformer models are now applied to every Bing query globally.

BEFORE

what can aggravate a concussion

266,000 Results Any time

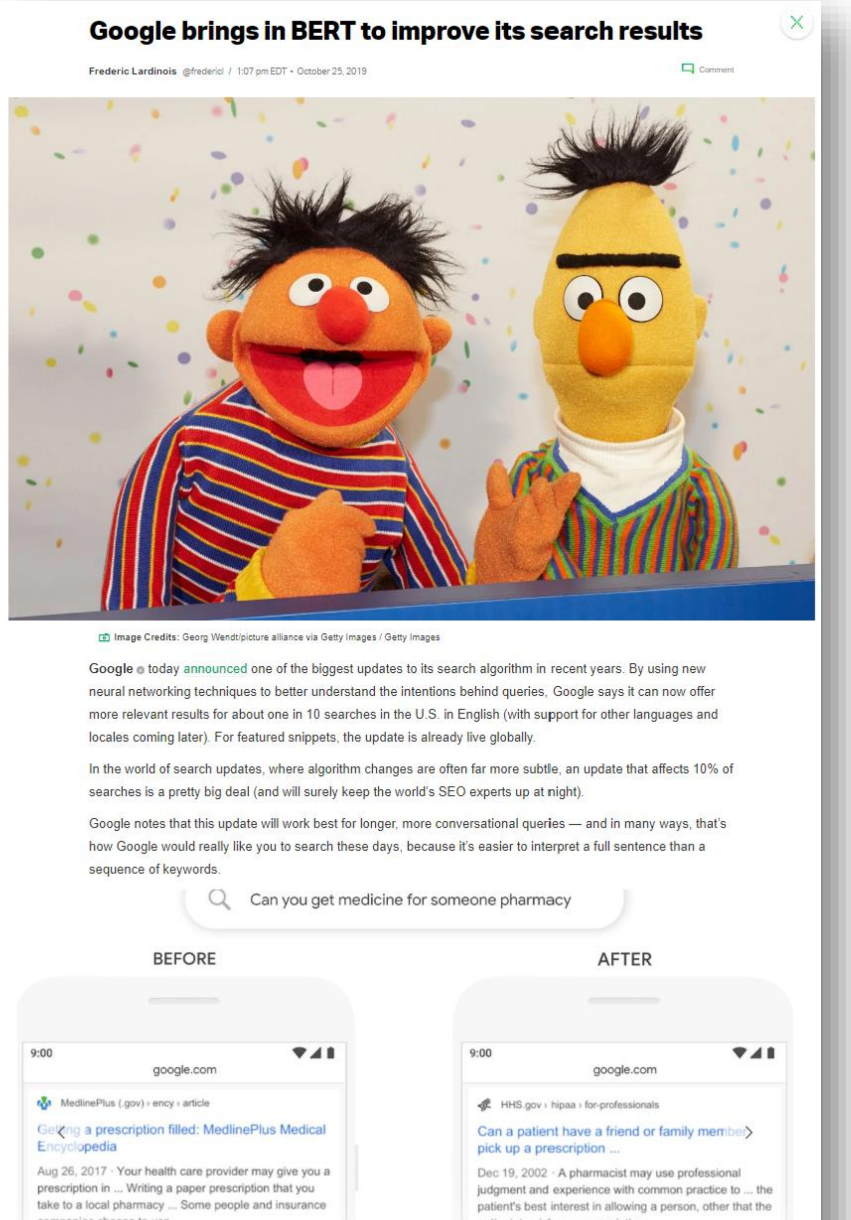
[PDF] Facts About Concussion and Brain Injury
https://www.cdc.gov/headsup/pdfs/providers/facts_about_concussion_tbi-a.pdf
head. Concussions can also occur from a fall or a blow to the body that causes the head to move rapidly back and forth. Doctors may describe these injuries as "mild" because concussions are usually not life-threatening. Even so, their effects can be serious. Understanding the signs and symptoms of a concussion can help you get better more quickly.

AFTER

what can aggravate a concussion

266,000 Results Any time

Suspect a Concussion? How to Help (Not Hurt) Your Recovery ...
https://health.clevelandclinic.org/suspect-a-concussion-how-to-help-not-hurt-your-recovery
Aug 15, 2017 · 4 things to avoid after a concussion. Excessive physical activity. An increased heart rate may worsen your symptoms. Strenuous mental activities. Reading, computer work, playing video games, texting and watching TV can overstimulate your brain. It's ...
Author: Brain And Spine, Brain And Spine Team



Google brings in BERT to improve its search results

Frederic Lardinois @fredericl / 1:07 pm EDT • October 25, 2019

Image Credits: Georg Wendt/picture alliance via Getty Images / Getty Images

Google today announced one of the biggest updates to its search algorithm in recent years. By using new neural networking techniques to better understand the intentions behind queries, Google says it can now offer more relevant results for about one in 10 searches in the U.S. in English (with support for other languages and locales coming later). For featured snippets, the update is already live globally.

In the world of search updates, where algorithm changes are often far more subtle, an update that affects 10% of searches is a pretty big deal (and will surely keep the world's SEO experts up at night).

Google notes that this update will work best for longer, more conversational queries — and in many ways, that's how Google would really like you to search these days, because it's easier to interpret a full sentence than a sequence of keywords.

Can you get medicine for someone pharmacy

BEFORE

9:00
google.com

MedlinePlus (.gov) › ency › article
Getting a prescription filled: MedlinePlus Medical Encyclopedia
Aug 26, 2017 · Your health care provider may give you a prescription in ... Writing a paper prescription that you take to a local pharmacy ... Some people and insurance

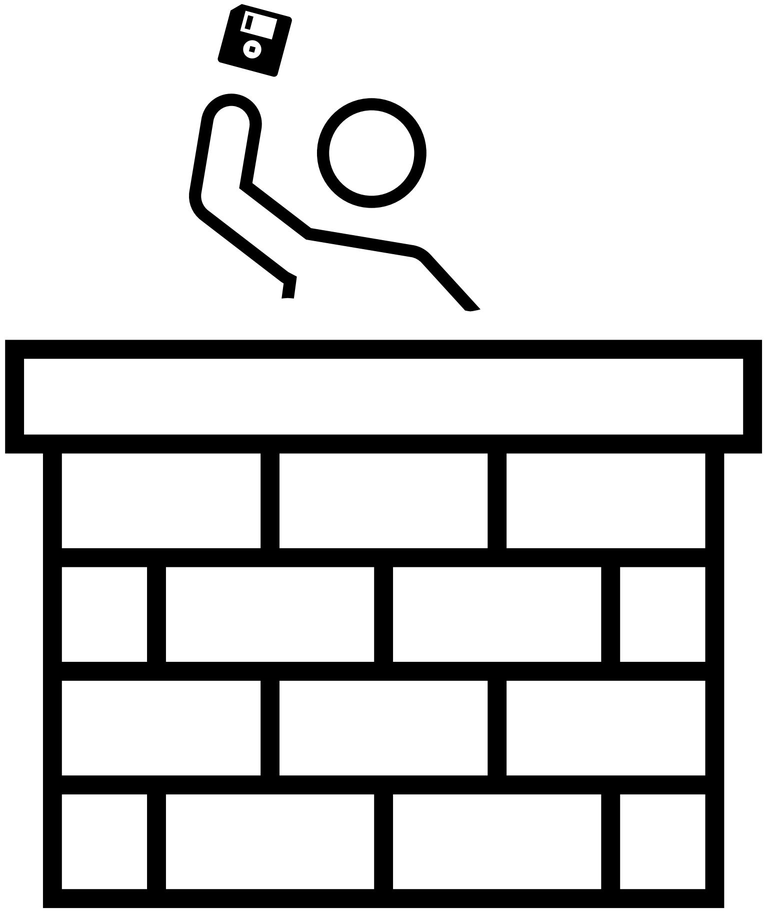
AFTER

9:00
google.com

HHS.gov › hipaa › for-professionals
Can a patient have a friend or family member pick up a prescription ...
Dec 19, 2002 · A pharmacist may use professional judgment and experience with common practice to ... the patient's best interest in allowing a person, other that the

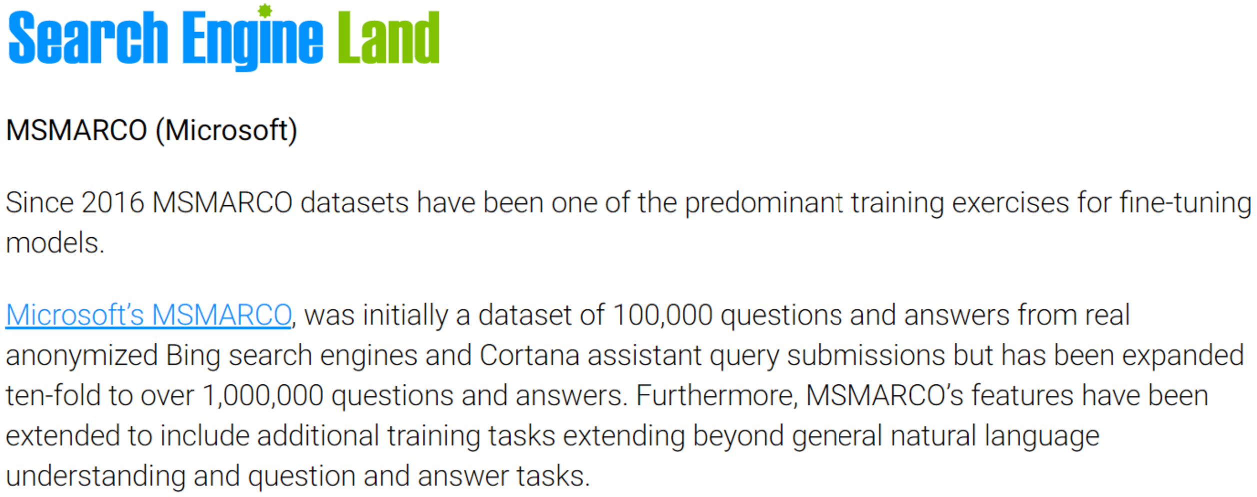
Summary

What benchmark development is **NOT** about: Throwing some data over the wall



What benchmark development is **NOT** about: Throwing some data over the wall

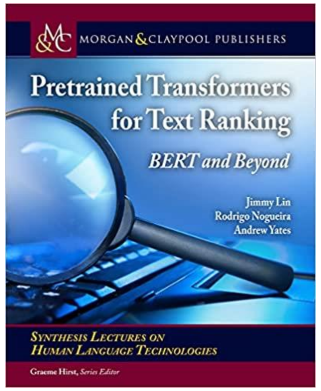
Developing benchmarks can be highly impactful and rewarding research when done thoughtfully—with due focus on community building, scientific rigor, and ethical considerations—and maintained responsibly over time


Search Engine Land

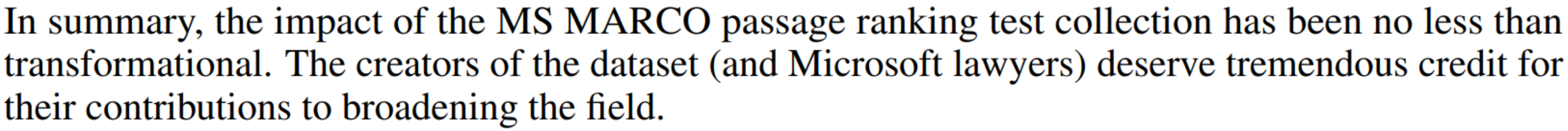
MSMARCO (Microsoft)

Since 2016 MSMARCO datasets have been one of the predominant training exercises for fine-tuning models.

[Microsoft's MSMARCO](), was initially a dataset of 100,000 questions and answers from real anonymized Bing search engines and Cortana assistant query submissions but has been expanded ten-fold to over 1,000,000 questions and answers. Furthermore, MSMARCO's features have been extended to include additional training tasks extending beyond general natural language understanding and question and answer tasks.



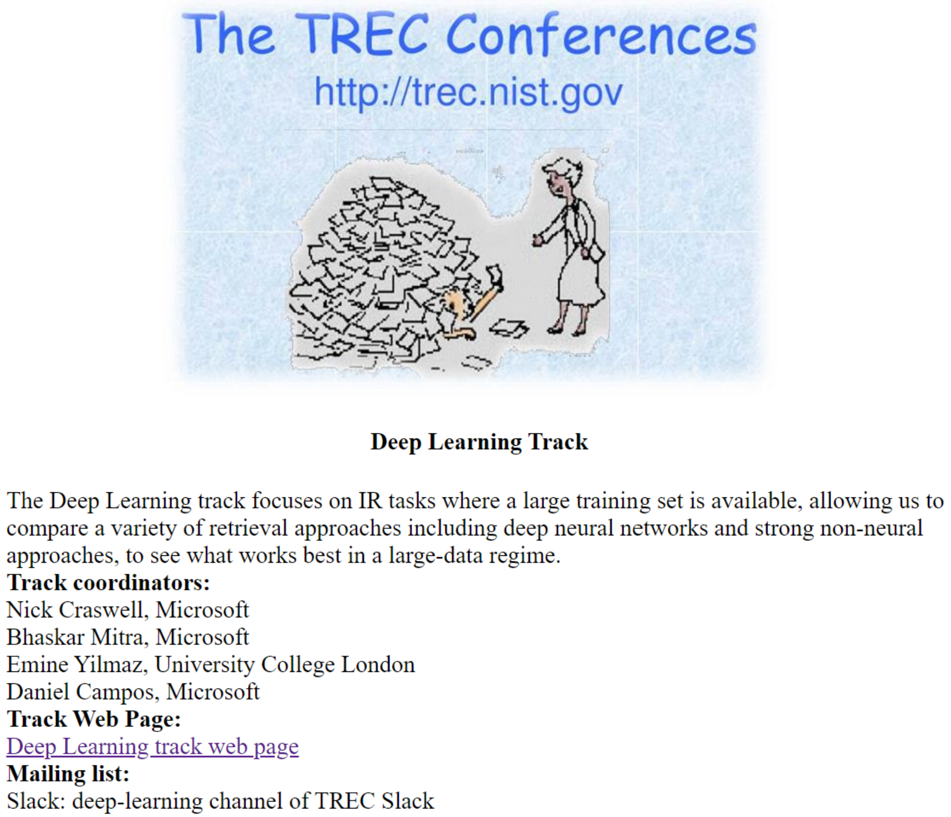
Morgan & Claypool Publishers
Pretrained Transformers for Text Ranking
BERT and Beyond
Jimmy Lin
Rodrigo Nogueira
Andrew Yates
Synthesis Lectures on Human Language Technologies
Graeme Hirst, Series Editor


In summary, the impact of the MS MARCO passage ranking test collection has been no less than transformational. The creators of the dataset (and Microsoft lawyers) deserve tremendous credit for their contributions to broadening the field.

Resources

Reusable research artifacts: Data

https://microsoft.github.io/msmarco/TREC-Deep-Learning



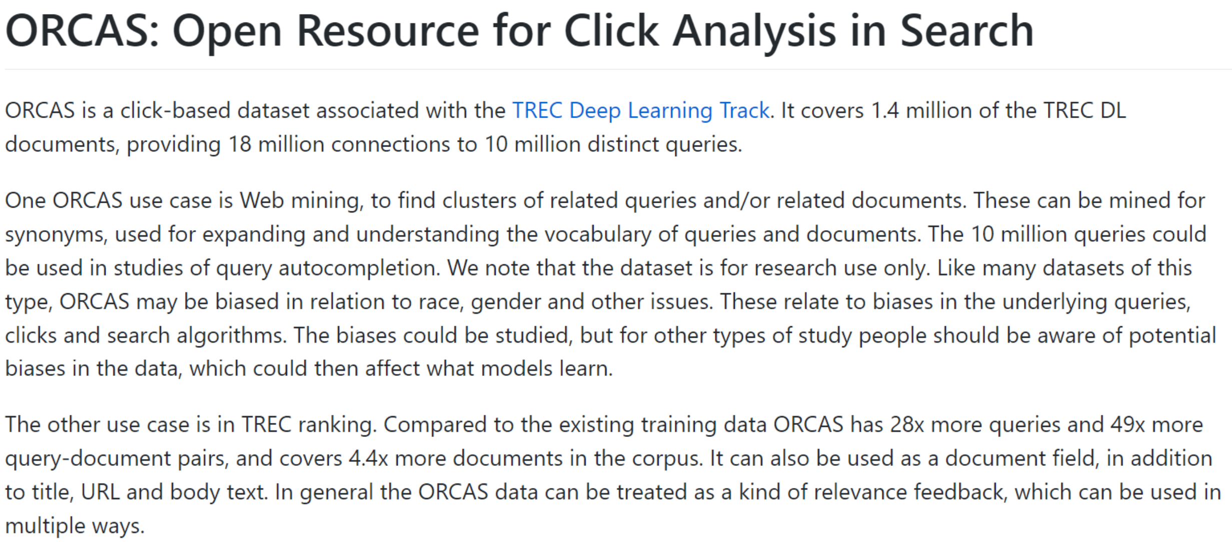
Deep Learning Track

The Deep Learning track focuses on IR tasks where a large training set is available, allowing us to compare a variety of retrieval approaches including deep neural networks and strong non-neural approaches, to see what works best in a large-data regime.
Track coordinators:
Nick Craswell, Microsoft
Bhaskar Mitra, Microsoft
Emine Yilmaz, University College London
Daniel Campos, Microsoft
Track Web Page:
Deep Learning track web page
Mailing list:
Slack: deep-learning channel of TREC Slack

MS MARCO

http://msmarco.org/

https://microsoft.github.io/msmarco/ORCAS



ORCAS: Open Resource for Click Analysis in Search

ORCAS is a click-based dataset associated with the TREC Deep Learning Track. It covers 1.4 million of the TREC DL documents, providing 18 million connections to 10 million distinct queries.

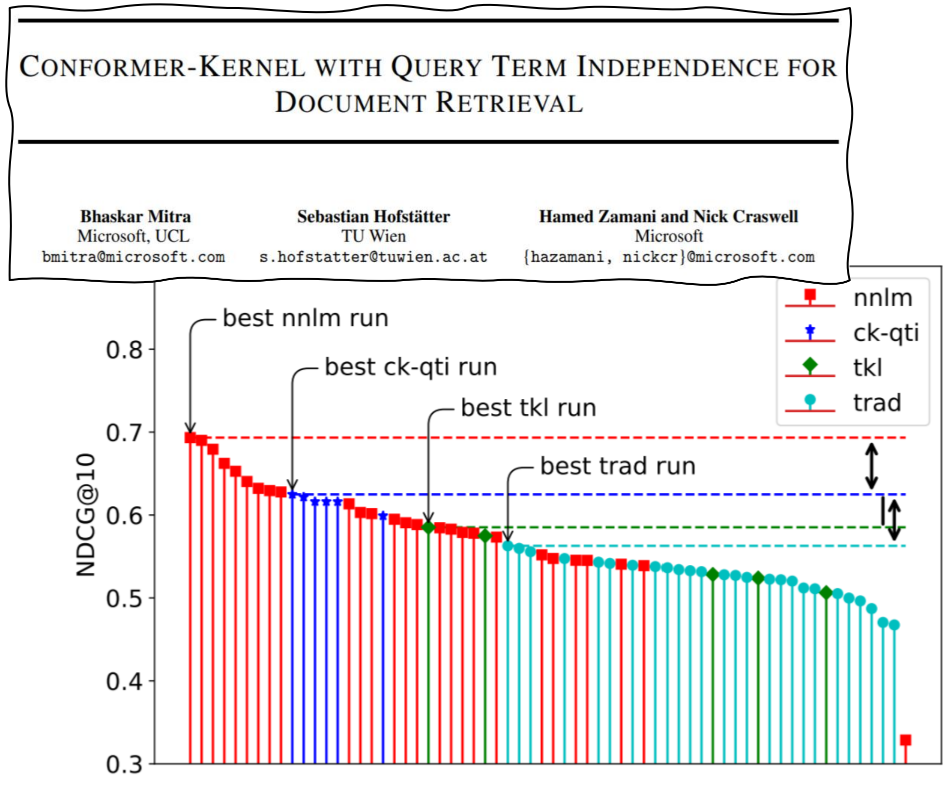
One ORCAS use case is Web mining, to find clusters of related queries and/or related documents. These can be mined for synonyms, used for expanding and understanding the vocabulary of queries and documents. The 10 million queries could be used in studies of query autocompletion. We note that the dataset is for research use only. Like many datasets of this type, ORCAS may be biased in relation to race, gender and other issues. These relate to biases in the underlying queries, clicks and search algorithms. The biases could be studied, but for other types of study people should be aware of potential biases in the data, which could then affect what models learn.

The other use case is in TREC ranking. Compared to the existing training data ORCAS has 28x more queries and 49x more query-document pairs, and covers 4.4x more documents in the corpus. It can also be used as a document field, in addition to title, URL and body text. In general the ORCAS data can be treated as a kind of relevance feedback, which can be used in multiple ways.

Reusable research artifacts: Code

Relatively cheap to reproduce neural baseline that outperformed all trad + nn runs and two-thirds of all nnlm runs at TREC 2020 Deep Learning Track

https://github.com/bmitra-msft/TREC-Deep-Learning-Quick-Start



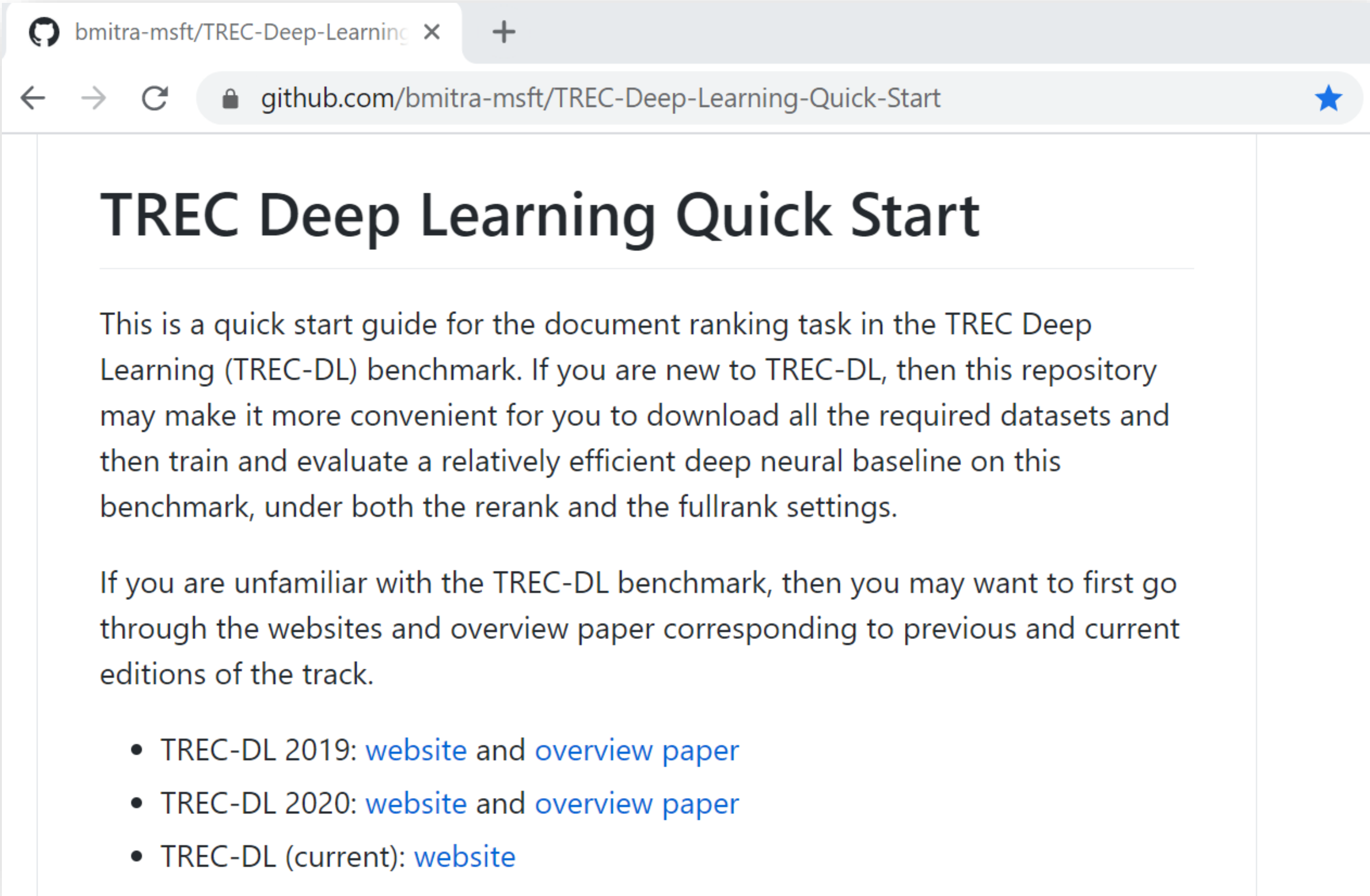
Conformer-Kernel with Query Term Independence for Document Retrieval

Bhaskar Mitra
Microsoft, UCL
bmitra@microsoft.com

Sebastian Hofstätter
TU Wien
s.hofstatter@tuwien.ac.at

Hamed Zamani and Nick Craswell
Microsoft
{hazamani, nickcr}@microsoft.com

(a) NDCG@10



bmitra-msft/TREC-Deep-Learning ×
+
github.com/bmitra-msft/TREC-Deep-Learning-Quick-Start

TREC Deep Learning Quick Start

This is a quick start guide for the document ranking task in the TREC Deep Learning (TREC-DL) benchmark. If you are new to TREC-DL, then this repository may make it more convenient for you to download all the required datasets and then train and evaluate a relatively efficient deep neural baseline on this benchmark, under both the rerank and the fullrank settings.

If you are unfamiliar with the TREC-DL benchmark, then you may want to first go through the websites and overview paper corresponding to previous and current editions of the track.

- TREC-DL 2019: website and overview paper
- TREC-DL 2020: website and overview paper
- TREC-DL (current): website

Learning resources

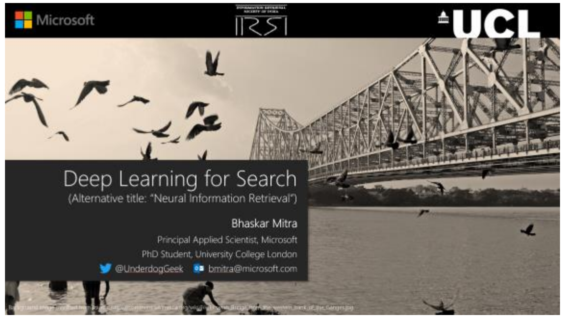
AFIRM: ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search

Learning to Rank with Neural Networks

Instructors
Bhaskar Mitra, Microsoft & University College London, Canada
Nick Craswell, Microsoft, USA
Emine Yilmaz, University College London
Daniel Campos, Microsoft, USA

January 2020

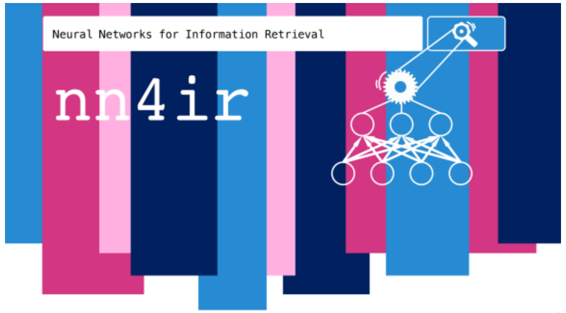
(slides, video)



Deep Learning for Search
(Alternative title: "Neural Information Retrieval")

Bhaskar Mitra
Principal Applied Scientist, Microsoft
PhD Student, University College London
@UnderdogGeek bmitra@microsoft.com

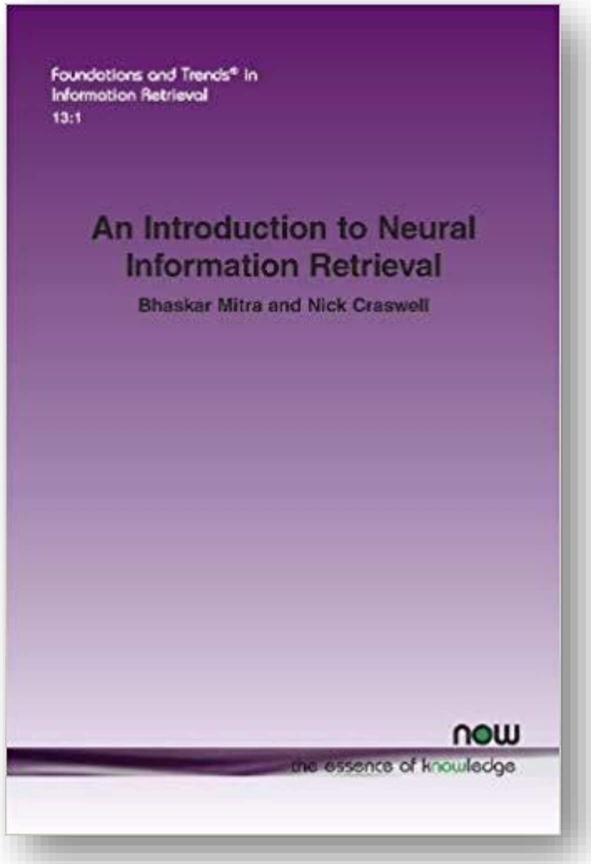
(slides)



Neural Networks for Information Retrieval

nn4ir

(website)



Foundations and Trends® in Information Retrieval
13:1

An Introduction to Neural Information Retrieval

Bhaskar Mitra and Nick Craswell

http://bit.ly/fntir-neural

Shout out to all my collaborators and every member of the neural IR community for this massive joint venture!

Monograph

- Mitra and Craswell. An Introduction to Neural Information Retrieval. FnTIR (2018).

Resources

- Bajaj, Campos, Craswell, Deng, Gao, Liu, Majumder, McNamara, Mitra, and others. MS MARCO: A Human Generated Machine Reading COmprehension Dataset. ArXiv (2016).
- Craswell, Campos, Mitra, Yilmaz, and Billerbeck. ORCAS: 20 Million Clicked Query-Document Pairs for Analyzing Search. CIKM (2020).
- Craswell, Mitra, Yilmaz, Campos, Voorhees, and Soboroff. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. SIGIR (2021).
- Arabzadeh, Mitra, and Bagheri. MS MARCO Chameleons: Challenging the MS MARCO Leaderboard with Extremely Obstinate Queries. CIKM (2021).
- Lin, Campos, Craswell, Mitra, and Yilmaz. Fostering Coopetition While Plugging Leaks: The Design and Implementation of the MS MARCO Leaderboards. SIGIR (2022).

TREC Reports

- Craswell, Mitra, Yilmaz, Campos, and Voorhees. Overview of the TREC 2019 Deep Learning Track. TREC (2020).
- Craswell, Mitra, Yilmaz, and Campos. Overview of the TREC 2020 Deep Learning Track. TREC (2021).
- Craswell, Mitra, Yilmaz, Campos, and Lin. Overview of the TREC 2021 Deep Learning Track. TREC (2022).

Analysis papers

- Yilmaz, Craswell, Mitra, and Campos. On the Reliability of Test Collections for Evaluating Systems of Different Types. SIGIR (2020).
- Craswell, Mitra, Yilmaz, Campos, and Lin. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. SIGIR (2021).
- Lin, Campos, Craswell, Mitra, and Yilmaz. Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard. SIGIR (2021).

Workshops

- Craswell, Croft, Guo, Mitra, and de Rijke. Report on the SIGIR 2016 Workshop on Neural Information Retrieval (Neu-IR). SIGIR Forum (2016).
- Craswell, Croft, de Rijke, Guo, and Mitra. Report on the Second SIGIR Workshop on Neural Information Retrieval (Neu-IR'17). SIGIR Forum (2017).

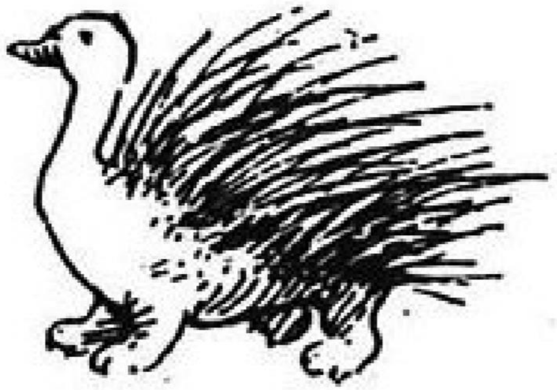
Tutorials

- Mitra and Craswell. Neural text embeddings for information retrieval. WSDM (2017).
- Kenter, Borisov, Van Gysel, Dehghani, de Rijke, and Mitra. Neural networks for information retrieval. SIGIR (2017).
- Kenter, Borisov, Dehghani, de Rijke, and Mitra. Neural networks for information retrieval. WSDM (2018).
- Kenter, Borisov, Van Gysel, Dehghani, de Rijke, and Mitra. Neural networks for information retrieval. ECIR (2018).

Journal special issue

- Craswell, Croft, de Rijke, Guo, and Mitra. Neural information retrieval: introduction to the special issue. IRJ (2017).

Thank you!



হাঁস ছিল, সজারু, (ব্যাকরণ মানি না),
হয়ে গেল 'হাঁসজারু' কেমনে তা জানি না ।

Was a duck, porcupine (to grammar I bow not)
Became Duckupine, but how I know not.

— *Sukumar Ray, Khichuri*

(Translation by Prasenjit Gupta)

@UnderdogGeek bmitra@microsoft.com