Static datasets aren't enough: Where deployed systems differ from research

Bernease Herman
WHYLABS

May 5, 2021
csv,conf,v6

I hold financial interest in WhyLabs, this work does not represent the views or work conducted at the University of Washington.

My varied experiences have informed these views

Data scientist at WHYLABS

The AI Observability company. We created and maintain **whylogs**, an open-source data logging library that uses statistical profiling for an efficient logging solution that scales and works in real-time.

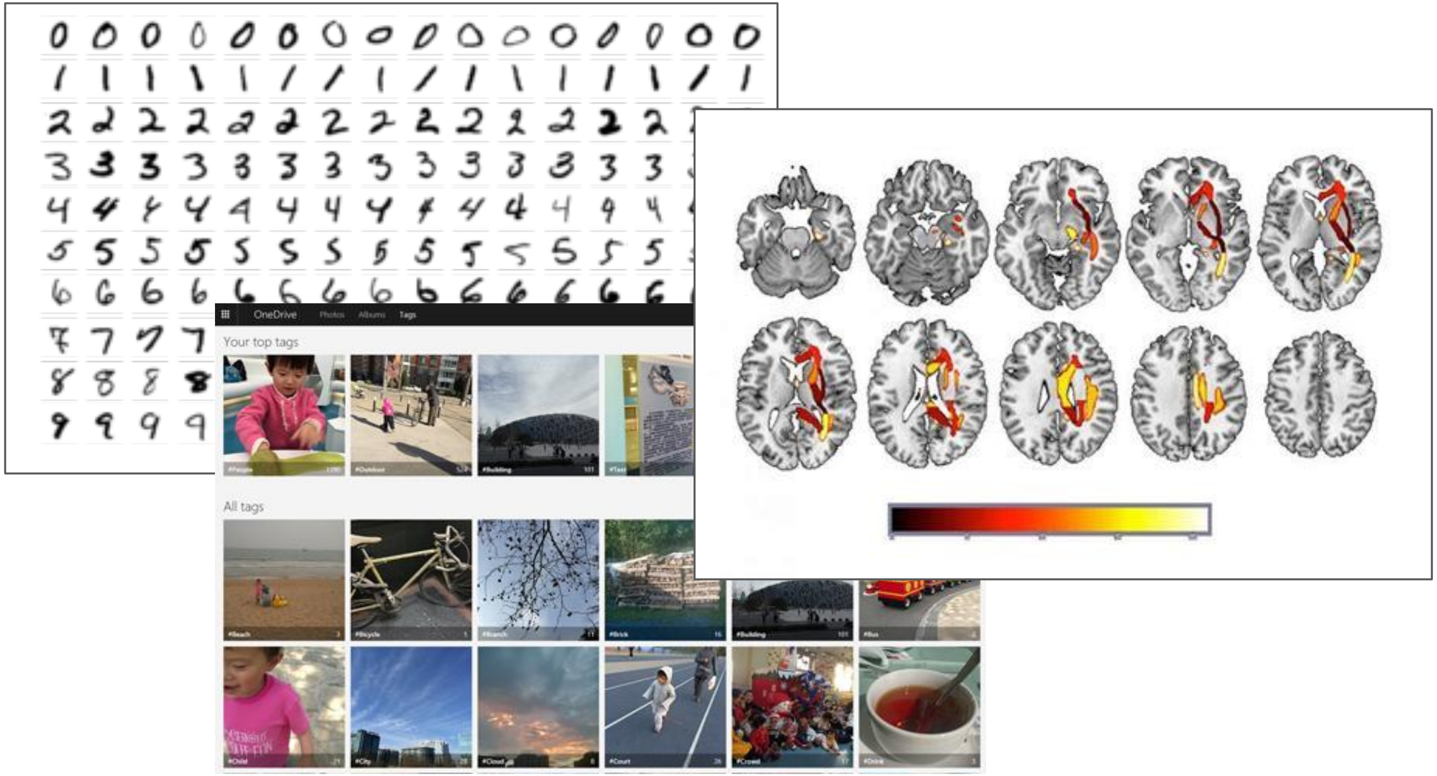
Research scientist at eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

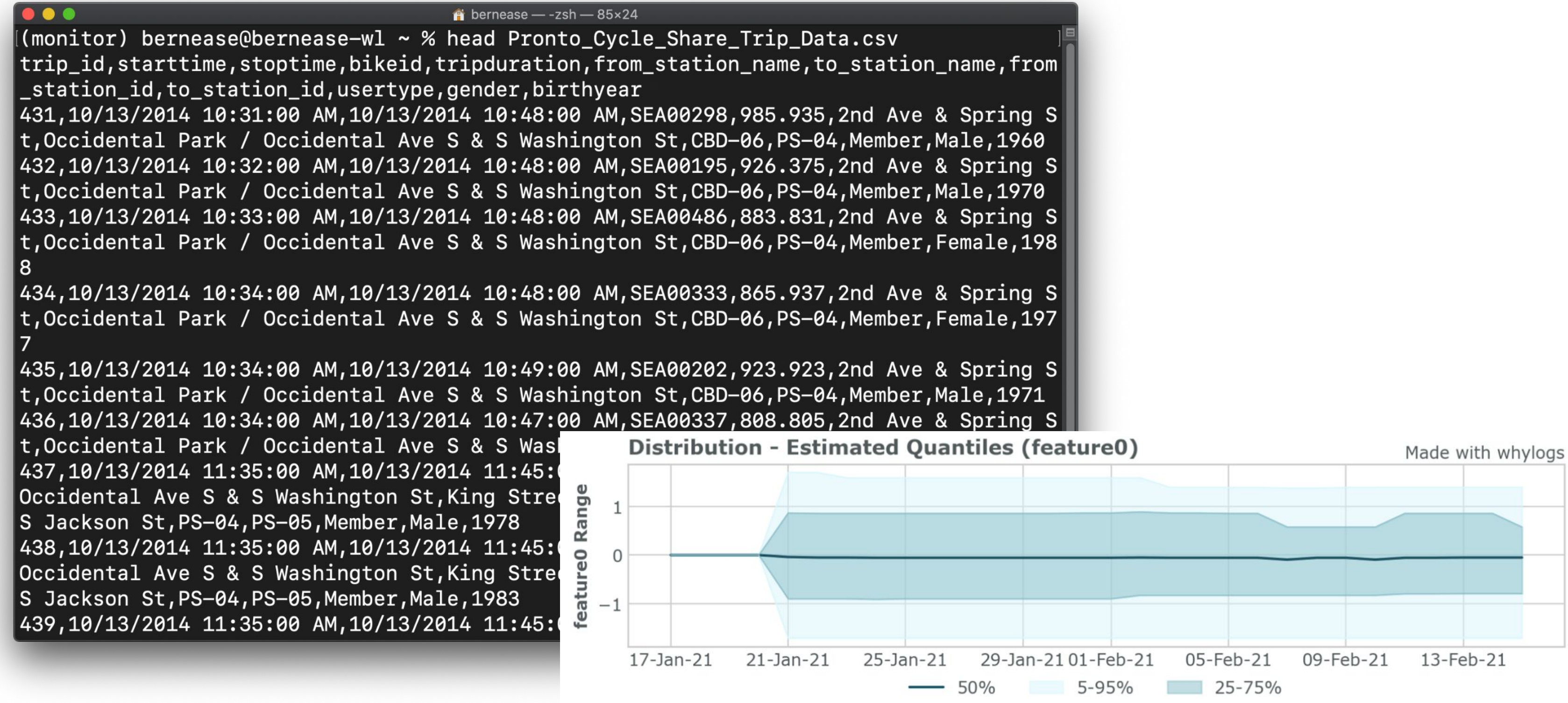
At University of Washington. Run academic hackweeks and summer Data Science for Social Good. Research on evaluation metrics and interpretability.

Prior, software engineer at amazon, research at Morgan Stanley

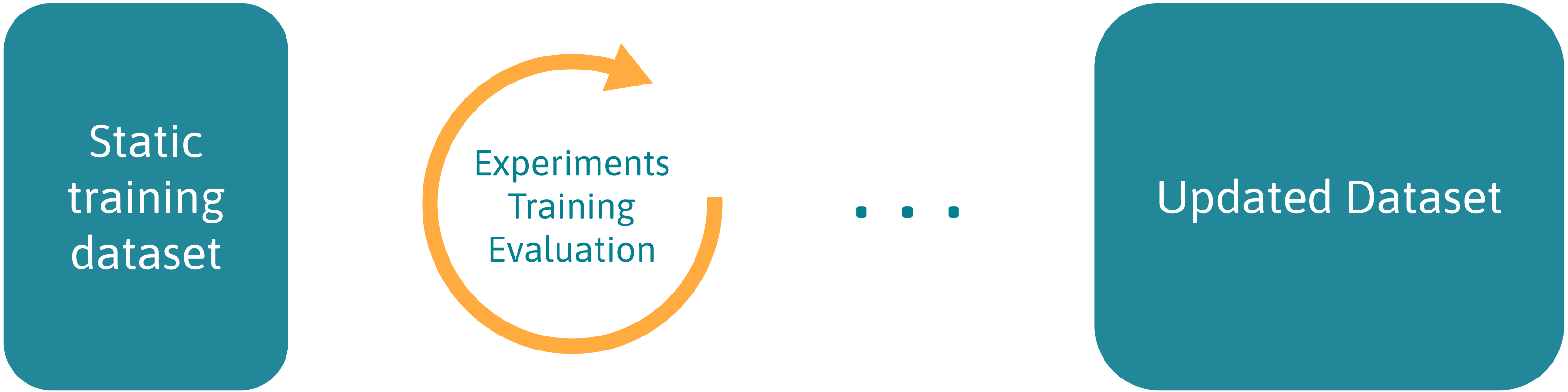
Many ML and data science resources use ***static*** datasets



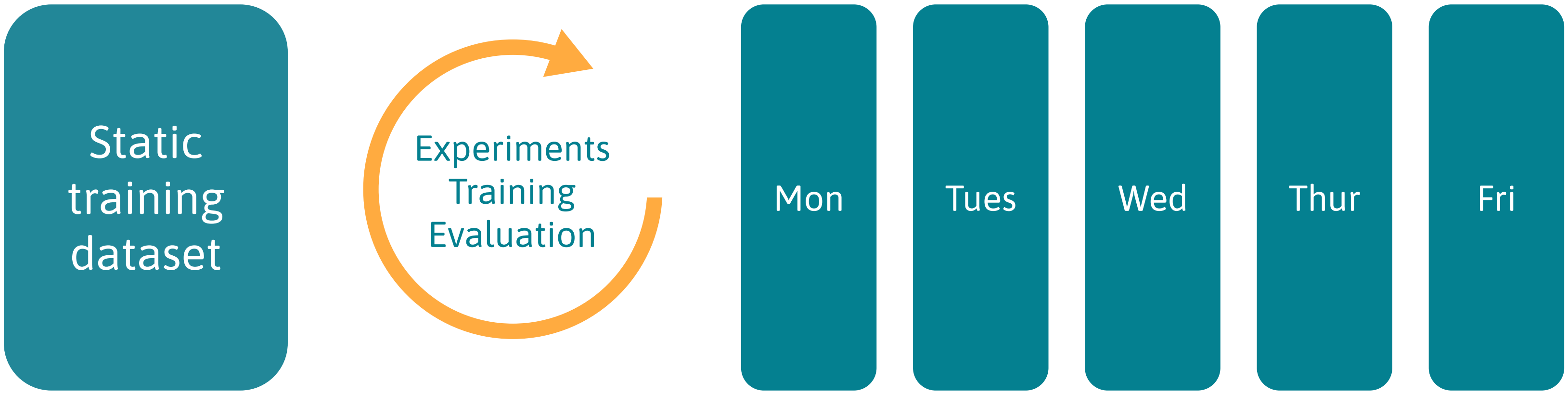
But many realistic datasets and metrics ***change over time***



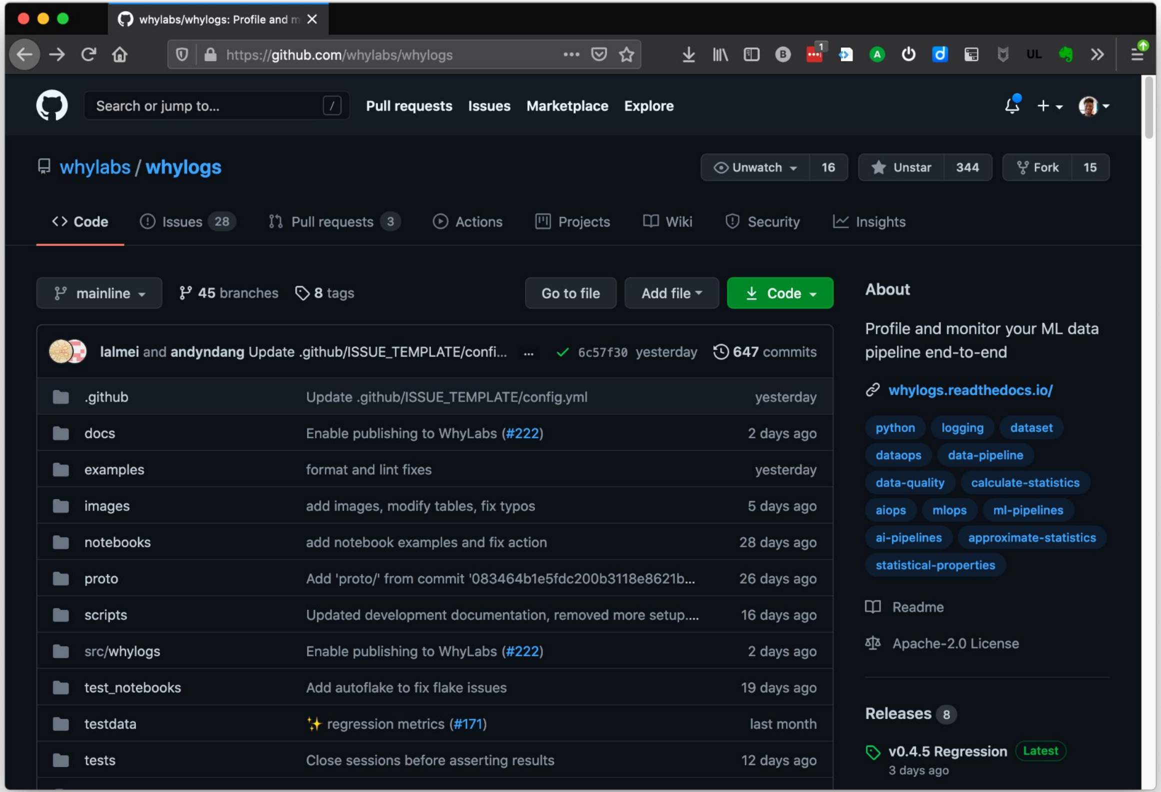
In deployed systems, the static approach leads to ***periodic dataset patches*** and model retrainings



But we should be logging and storing our data in a ***dynamic*** way

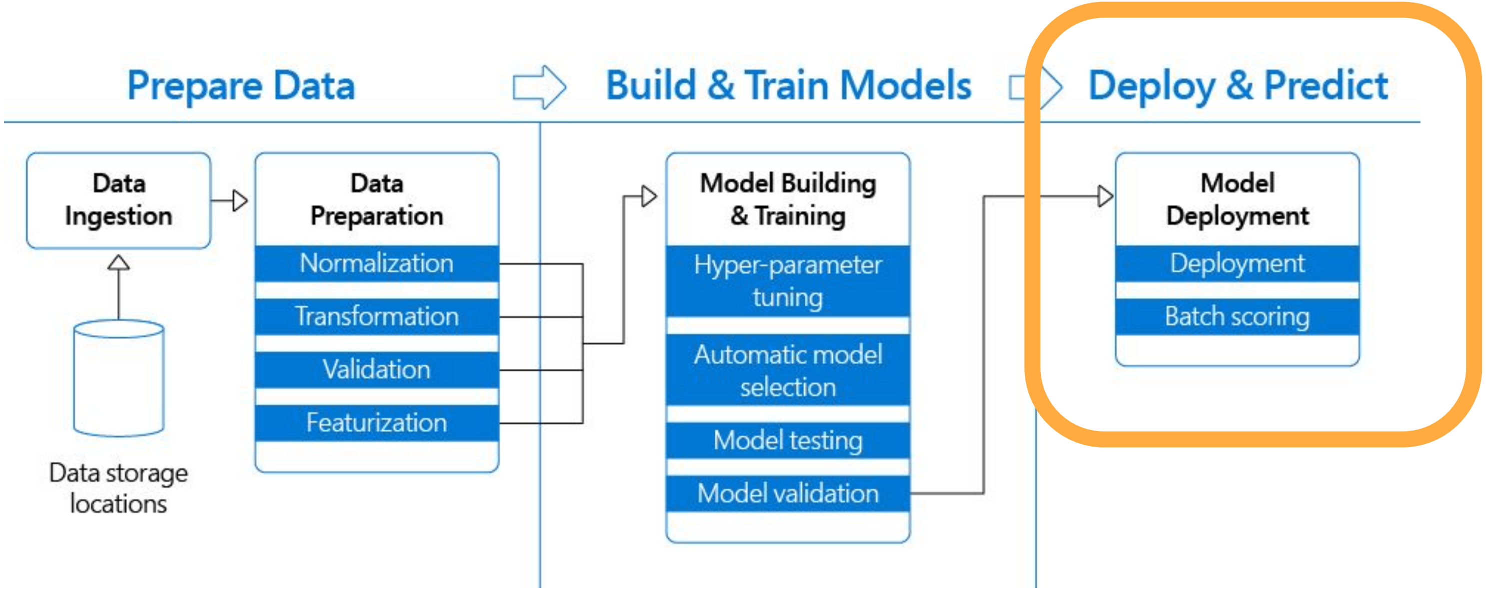


That’s why we’ve open-sourced our library, **whylogs**



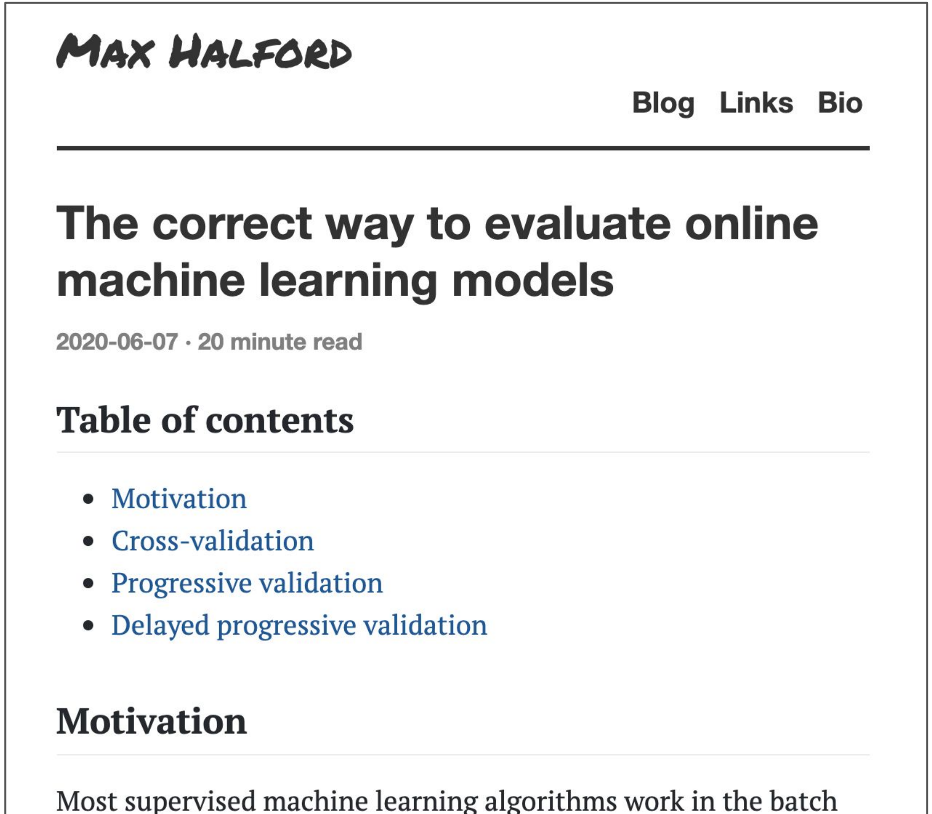
Let's start a conversation about time-batched data and other skills that are missing from data science learning pathways.

No more ignoring the deployment stage of the pipeline in data science training and tools



No more doing model evaluation that ***doesn't account for*** the timing and realities of data arrival

Progressive validation & Delayed progressive validation


MAX HALFORD

Blog Links Bio

The correct way to evaluate online machine learning models

2020-06-07 · 20 minute read

Table of contents

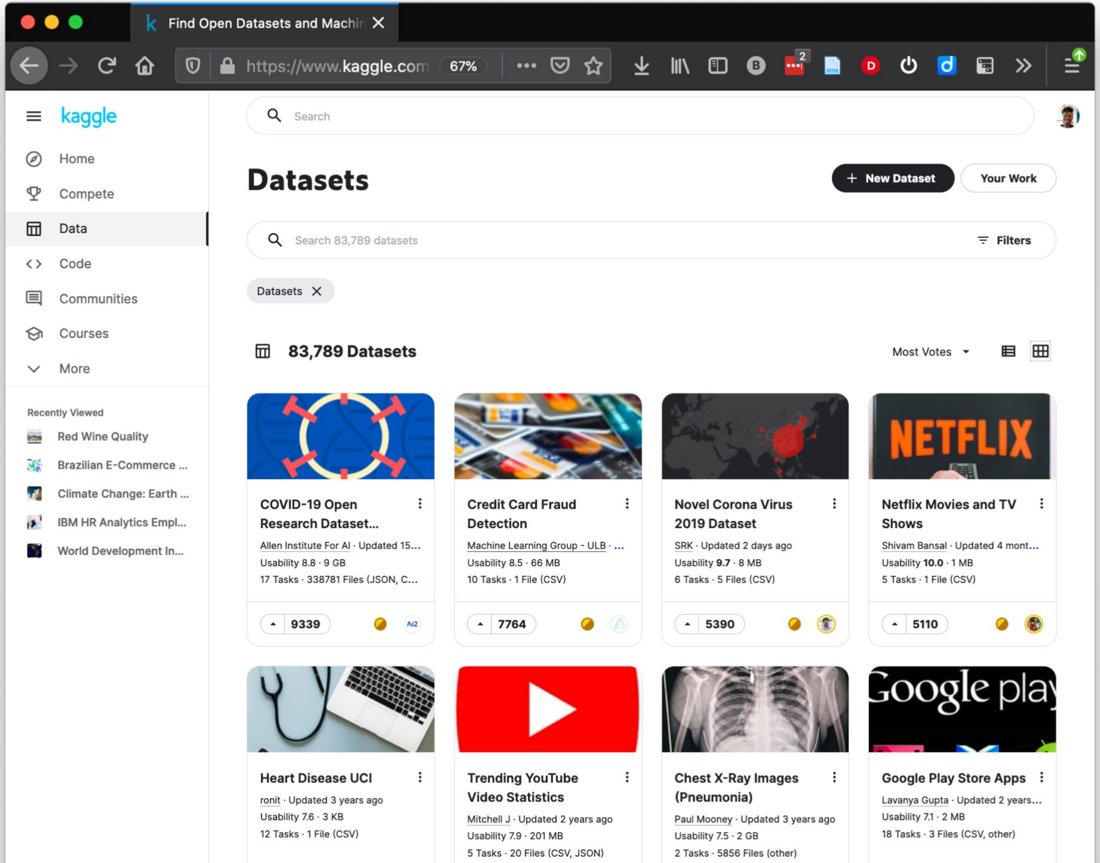
Motivation

Most supervised machine learning algorithms work in the batch

All data scientists need skills on ***time-batched*** data

14 of 40 of the most voted Kaggle datasets include date or time index column.

(As of May 4, 2021)



All data scientists need skills on ***time-batched*** data

Evolving Academia/Industry Relations in Computing Research

Shwetak Patel (Univ. Washington), Jennifer Rexford (Princeton Univ.),
Benjamin Zorn (Microsoft), Greg Morrisett (Cornell Univ.)
Industry Working Group, Computing Community Consortium (CCC)
June 2019

Executive Summary

In 2015, the CCC co-sponsored an industry round table that produced the document “The Future of Computing Research: Industry-Academic Collaborations.”[1] Since then, several important trends in computing research have emerged, and this document considers how those trends impact the interaction between academia and industry in computing fields. We reach the following conclusions:

- In certain computing disciplines, such as currently artificial intelligence, we observe **significant increases in the level of interaction between professors and companies**, which take the form of extended joint appointments.
- Increasingly, **companies are highly motivated to engage** both professors and graduate students working in specific technical areas because companies view computing research and technical talent as a core aspect of their business success.
- There is also the further potential for principles and values from the academy (e.g., ethics, human-centered approaches, etc.) informing products and R&D roadmaps in new

earch Interactions Between University and Industry
in Computer Science in the
United States and United Kingdom

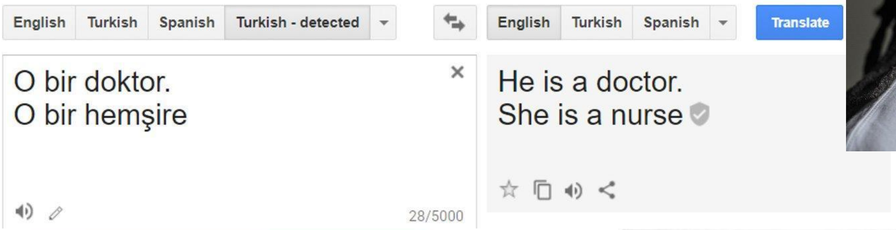
Report Number: 95-8-1

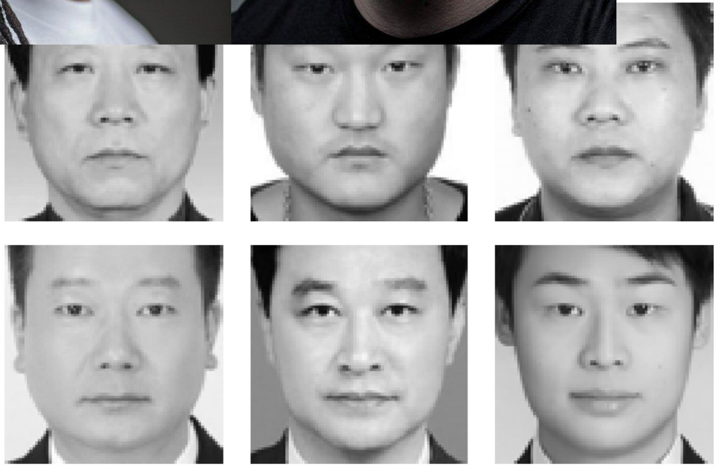
Thomas Haigh

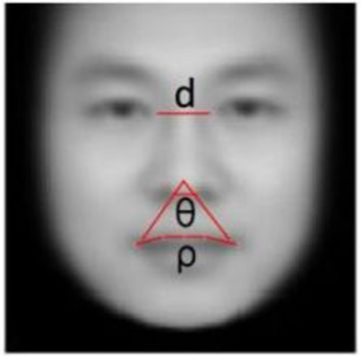
School of Arts and Sciences
Department of History and Sociology of Science
Suite 500, 3440 Market Street
Philadelphia, PA 19104-8400
USA.
c/o ckeirns@sas.upenn.edu

(formerly of)

The *impact* that industrial ML and data science has ***on the world***







Logging your data with a few lines of code

Run the profiling

Note that the result has three entries for three days

[22]:
```
%%time
profile_df = sessionWithModel.aggProfiles().cache()
profile_df.count()
```

```
CPU times: user 1.35 ms, sys: 944 µs, total: 2.3 ms
Wall time: 1.56 s
```

[22]:
```
30
```

[23]:
```
profiles = profile_df.toPandas()
profiles
```

[23]:

	date	why_profile
0	2021-02-06 16:00:00	[128, 129, 2, 10, 85, 8, 1, 16, 2, 26, 13, 109...
1	2021-02-07 16:00:00	[140, 133, 2, 10, 85, 8, 1, 16, 2, 26, 13, 109...
2	2021-01-28 16:00:00	[139, 232, 1, 10, 85, 8, 1, 16, 2, 26, 13, 109...
3	2021-02-13 16:00:00	[221, 132, 2, 10, 85, 8, 1, 16, 2, 26, 13, 109...
4	2021-01-19 16:00:00	[160, 231, 1, 10, 85, 8, 1, 16, 2, 26, 13, 109...
5	2021-02-12 16:00:00	[145, 138, 2, 10, 85, 8, 1, 16, 2, 26, 13, 109...

Explore trends in a few lines of code

[25]:
```
from whylogs.viz import ProfileVisualizer

viz = ProfileVisualizer()
```

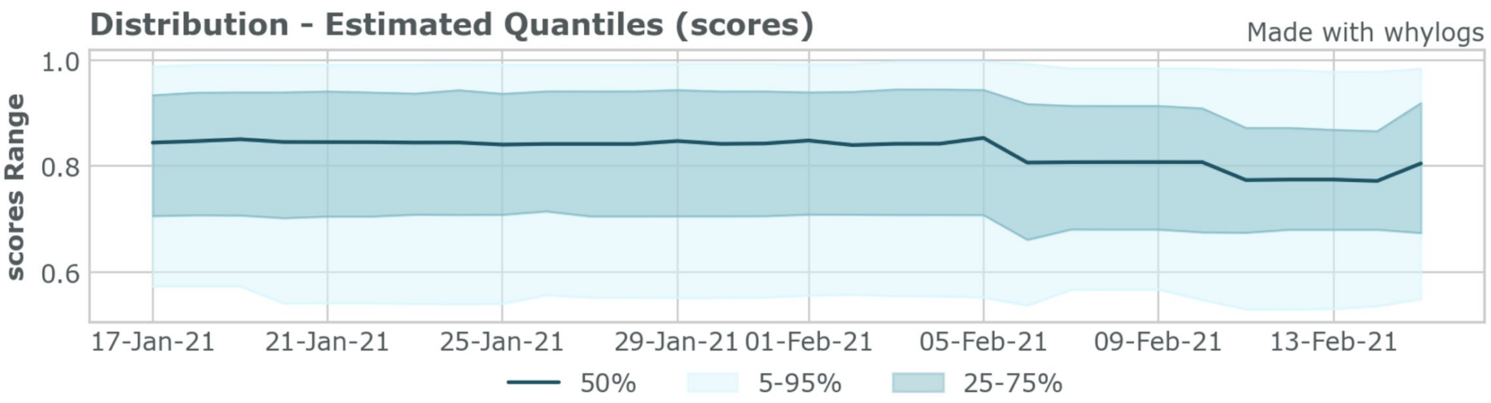
[26]:
```
from whylogs import DatasetProfile
```

[27]:
```
profile_bytes = list(profiles['why_profile'])
dataset_profiles = []
for p in profile_bytes:
    dataset_profiles += DatasetProfile.parse_delimited(p)
```

[28]:
```
viz.set_profiles(dataset_profiles)
```

[29]:
```
viz.plot_distribution("scores")
```

[29]:



Approximate statistics that are storage, computation, and data analysis friendly

Dataset	Size	No. of Entries	No. of Features	Est. Memory Consumption	Output Size (uncompressed)
Lending Club	1.6GB	2.2M	151	**14MB**	7.4MB
NYC Tickets	1.9GB	10.8M	43	**14MB**	2.3MB
Pain pills in the USA	75GB	178M	42	**15MB**	2MB

bit.ly/whylogs

bernease@whylabs.ai

@bernease