

1ST DATA EXCELLENCE WORKSHOP (DEW 2020)

At the 8th AAAI Conference on Human Computation and Crowdsourcing

Date: 26 October 2020

Time PT: 6:00am-10:20am

Time EST: 9:00am-1:20pm

Time CET: 2:00pm-6:20pm

Location: Virtual

IMPORTANT DATES

REGISTER

TALKS

DEW2020

1ST Data Excellence Workshop

START

6:00a

PDT

SF

9:00a

EDT

NYC

2:00p

CET

AMS

JOIN VIDEO STREAM AT:
underline.io/events/27/main-stage

JOIN SLACK CHAT AT:
bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:
• [#ask-questions-for-speakers-here](#)
• [#ask-for-help-here](#)

WORKSHOP PROGRAM:
eval.how/dew2020/

CODE OF CONDUCT:
www.humancomputation.com/attend.html#code

DEW2020

1ST Data Excellence Workshop

SESSION 1

6:10a 9:10a 2:10p

BREAK

7:25a 10:25a 3:25p

SESSION 2

7:35a 10:35a 3:35p

BREAK

8:45a 11:45a 4:45p

SESSION 3

8:55a 11:55a 4:55p

JOIN VIDEO STREAM AT:
underline.io/events/27/main-stage

JOIN SLACK CHAT AT:
bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:

- [#ask-questions-for-speakers-here](#)
- [#ask-for-help-here](#)

WORKSHOP PROGRAM:

eval.how/dew2020/

CODE OF CONDUCT:

www.humancomputation.com/attend.html#code

Lets get started!

- Bad data hurts science and hurts society
- We are the research community that collects data for training and testing AI
- How can we achieve data excellence for AI and beyond?



Forecasting for COVID-19 has failed

John P.A. Ioannidis ^{a,*}, Sally Cripps ^b, Martin A. Tanner ^c

Epidemic forecasting has a dubious track-record, and its failures became more prominent with COVID-19. Poor data input, wrong modeling assumptions, high sensitivity of estimates, lack of incorporation of epidemiological features, poor past evidence on effects of available interventions, lack of transparency, errors, lack of determinacy, consideration of only one or a few dimensions of the problem at hand, lack of expertise in crucial disciplines, groupthink and bandwagon effects, and selective reporting are some of the causes of these failures. Nevertheless, epidemic forecasting is unlikely to be abandoned.



WIKIPEDIA

The Free Encyclopedia

[Main page](#)

[Contents](#)

[Current events](#)

[Random article](#)

Article

Talk

Read

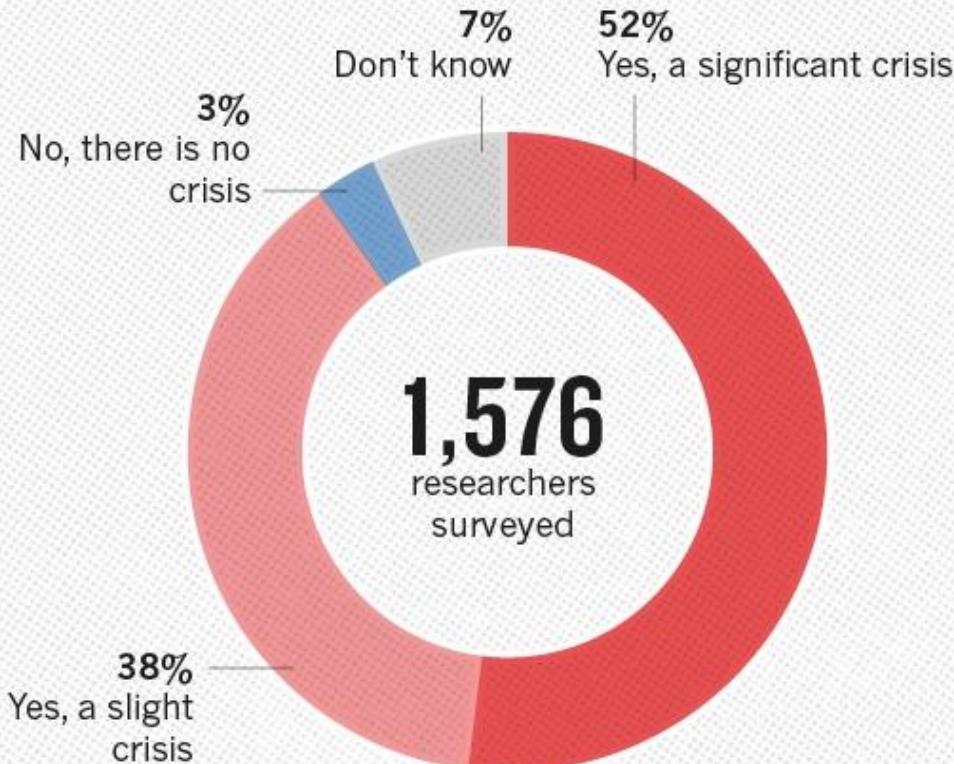
Edit

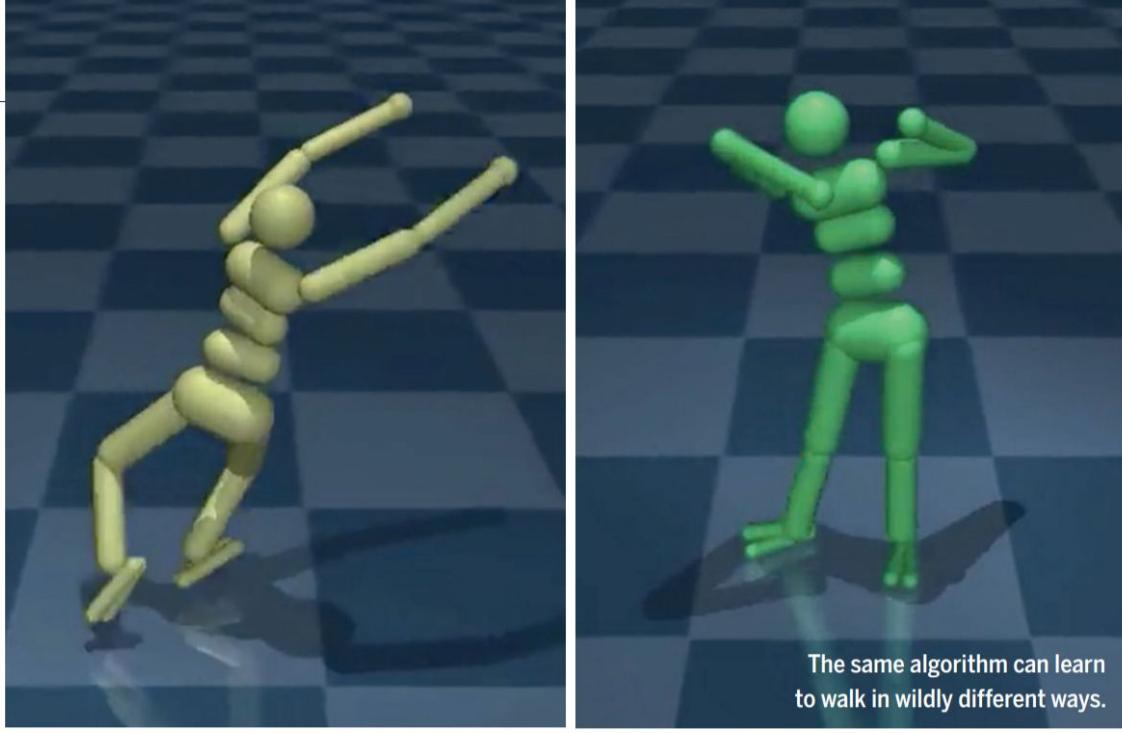
Replication crisis

From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis** or **reproducibility crisis**) is, as of 2020, an ongoing **methodological** crisis in which it has been found that many scientific studies are difficult or impossible to **replicate** or **reproduce**.

IS THERE A REPRODUCIBILITY CRISIS?





The same algorithm can learn
to walk in wildly different ways.

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. Science: 725-726.

Bad data hurts AI which can hurt society

SCIENCE \ TECH \ HEALTH

IBM's Watson gave unsafe recommendations for treating cancer

Doctors fed it hypothetical scenarios, not real patient data

By Angela Chen | @chengela | Jul 26, 2018, 4:29pm EDT

≡ WIRED

ANDY GREENBERG SECURITY 11.12.2017 08:44 PM

Hackers Say They've Broken Face ID a Week After iPhone X Release

"I would say if this is all confirmed, it does mean Face ID is less secure than Touch ID."



THIS ARTICLE HAS been updated below with another, more convincing video demonstration of Blkav's Face ID spoofing, which the firm revealed two weeks after the original.

When Apple released the iPhone X on November 3, it touched off an immediate race among hackers around the world to be the first to fool the company's futuristic new form of authentication. A week later, hackers on the actual other side of the world claim to have successfully duplicated someone's face to unlock his iPhone X—with what looks like a simpler technique than some security researchers believed possible.

Self-driving car timeline for 11 top automakers

DAN FAZELLA / TECHMERGENCE @DANFAZELLA JUNE 4, 2017 3:10 PM



Tesla's Autopilot keeps crashing into parked cars. Here's why.

≡ The Washington Post
Democracy Dies in Darkness

Transportation

Uber's self-driving cars had a major flaw: They weren't programmed to stop for jaywalkers



An Uber driverless car in a garage in San Francisco. (Eric Risberg/AP)

By Hannah Knowles

November 6, 2019 at 10:06 p.m. EST

MIT Technology Review

Sign In Subscribe

Topics Magazine Newsletters Events



Artificial intelligence / Machine learning

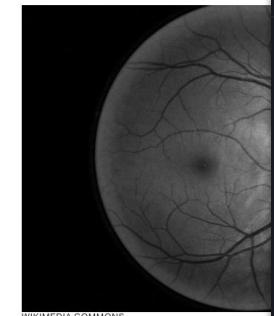
Google's medical AI was super accurate in a lab. Real life was a different story.

If AI is really going to make a difference to patients we need to know how it works when real humans get their hands on it, in real situations.

by Will Douglas Heaven

April 27, 2020

QUARTZ



O DEAR

Google Translate's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples

FROM OUR OBSESSION
Language
We explore how language helps us make sense of a changing world.

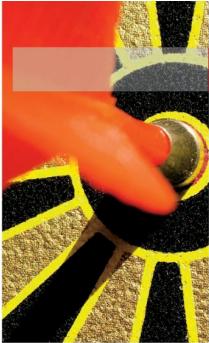
By Nikhil Sonnad
Reporter
November 29, 2017 • This article is more than 2 years old.

In the Turkish language, there is one pronoun, "o," that covers every kind of singular third person. Whether it's a he, a she, or an it, it's an "o." That's not the case in English. So when Google Translate goes from Turkish to English, it just has to guess whether "o" means he, she, or it. And those translations reveal the algorithm's gender bias.

Existing rules for deploying AI in clinical settings, such as the US or a CE mark in Europe, focus primarily on a rule that an AI must improve the outcome for patients, like "But that needs to change, says Emma Beede, a UX researcher. "Understand how AI tools are going to work for people in context—especially in health care—before they're widely deployed."

from "AI in the Open World: Discovering Biased Spots of AI", SafeAI 2020, Ece Kumar

Data is the fuel and compass for AI



EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"¹ examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech rec-

In the decade since then, the research community have done a lot with **quantity**, but **quality has been left behind**

But we have lost track

- Focus on cheaper, faster, bigger has led to a market for lemons [Ipeirotis, 2011]
 - “Due to uncertainty in the notion of quality, the market focuses on price, resulting in an equilibrium state where the good sellers are priced out of the market and only lemons remain.” [Akerlof, 1967]
- Much like software, stop looking for a metric, but build processes to support excellence

**Data is excellent when
it accurately represents
a phenomenon, and is:**

- Collected, stored, and used **responsibly**
- Maintainable over time
- Reusable across applications
- Empirical and **explanatory power**

Summary

- The bad news
 - We are up against a cultural perception of glamorizing models and math not data
- The good news
 - We are the research community focused on data for AI: they need us bad!
- Our plan
 - To be the community that brings about this cultural change towards data excellence
 - Share tools, case-studies, best practices, standards, and research to propel data excellence

DEW2020

1ST Data Excellence Workshop

SESSION 1

**Towards Accountability
for Machine Learning Datasets:
Lessons from Software Engineering,** Ben Hutchinson

TREC: Building datasets within a community of use, Ian Soboroff

Data =? Software, Andrea Olgiati

**Reducing Annotation Artifacts in Crowdsourcing
Datasets for Natural Language Processing,**
Donghoon Han, Juho Kim and Alice Oh

**Machine Learning Training to Support Diversity of
Opinion,** Johanne Christensen and Benjamin Watson

**Data Desiderata: Reliability and Fidelity in
High-stakes AI,** Shivani Kapania, Nithya Sambasivan,
Kristen Olson, Hannah Highfill, Diana Akrong, Praveen Paritosh and Lora Aroyo

INVITED TALK

20 mins

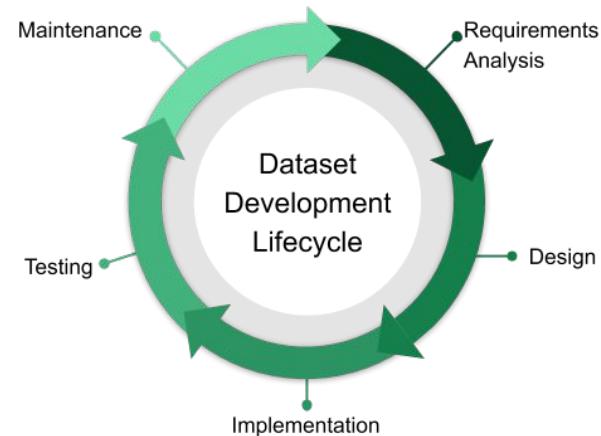
Towards Accountability for Machine Learning

Datasets: Lessons from Software Engineering

Ben Hutchinson
Google Research

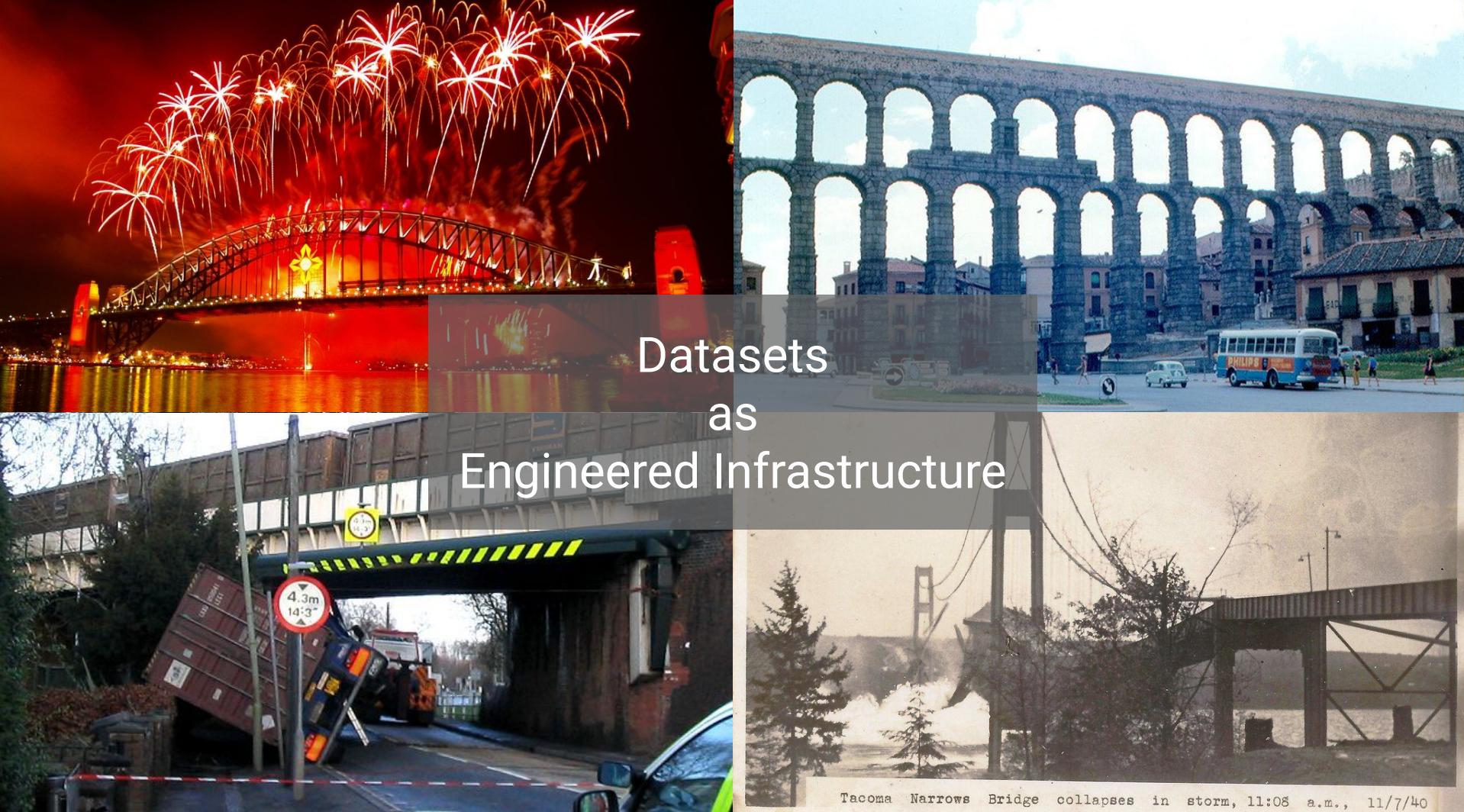
Towards Accountability for Machine Learning Datasets: Lessons from Software Engineering

Ben Hutchinson, Andrew Smart,
Alex Hanna, Emily Denton,
Christina Greer, Oddur
Kjartansson, Parker Barnes,
Margaret Mitchell



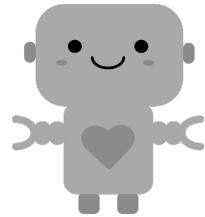
What makes data "excellent"?

What kind of a thing is data?



Datasets as Engineered Infrastructure

Tacoma Narrows Bridge collapses in storm, 11:08 a.m., 11/7/40



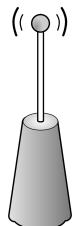
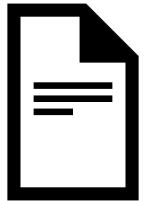
Ethical AI

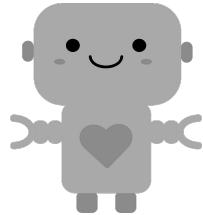


Responsible Data Practices



Datasets as Engineered Infrastructure

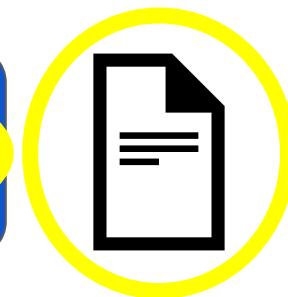




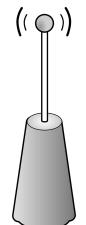
Ethical AI



Documentation of Dataset Creation



Datasets as Engineered Infrastructure



Ethical AI

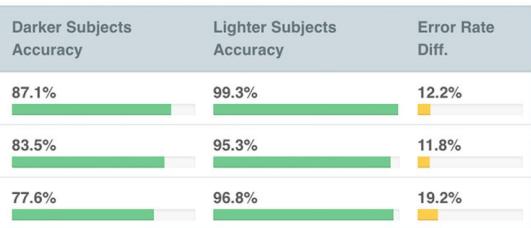


Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity

Ethical principle	Number of documents	Included codes
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

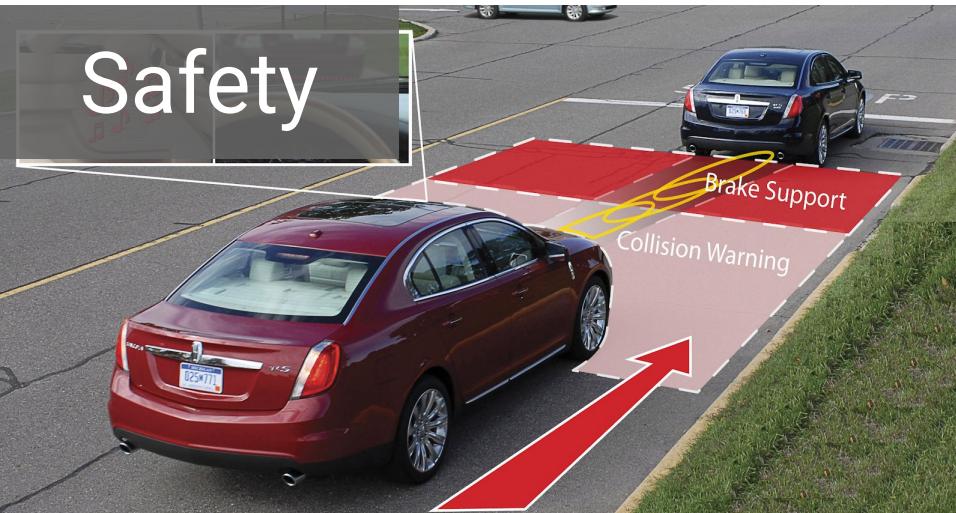
Source: Jobin, lenca and Vayena (2019). *The global landscape of AI ethics guidelines*.

Fairness

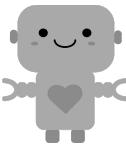


Buolamwini and Gebru (2018). Gender Shades.

Safety



Privacy



LONG LIVE THE REVOLUTION.
OUR NEXT MEETING WILL BE
AT THE DOCKS AT MIDNIGHT
ON JUNE 28 TAB

AHA, FOUND THEM!

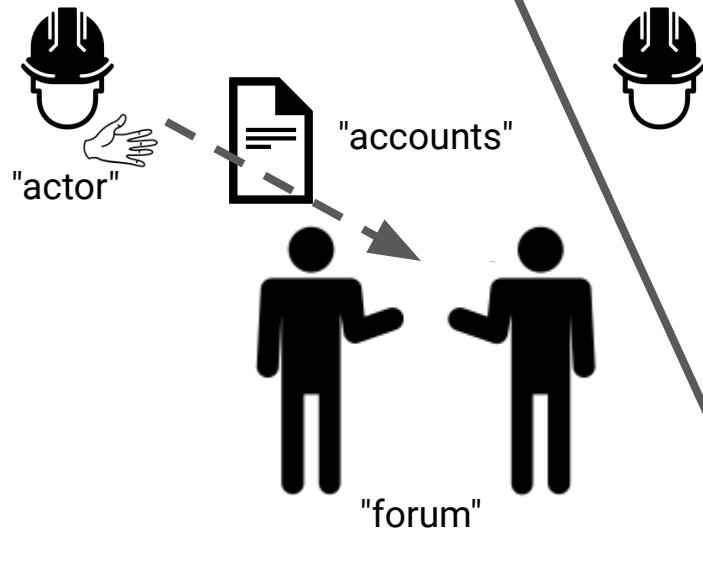


WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

Accountability is about Answerability



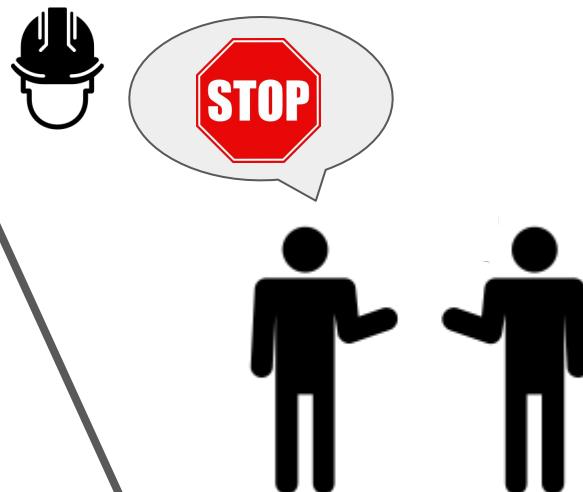
1. Information sharing



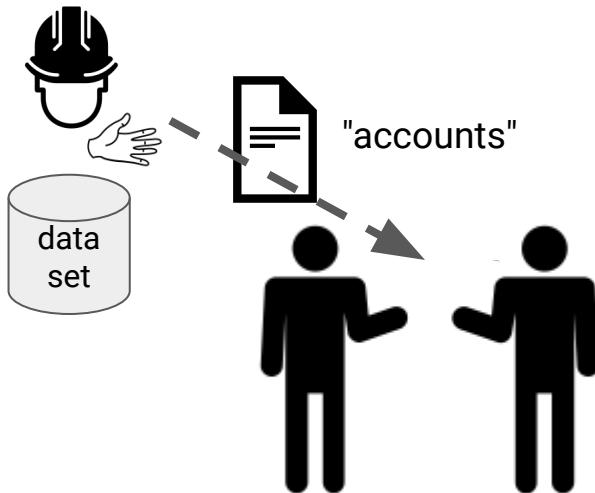
2. Deliberation



3. Consequences



Key Questions for ML Data Transparency

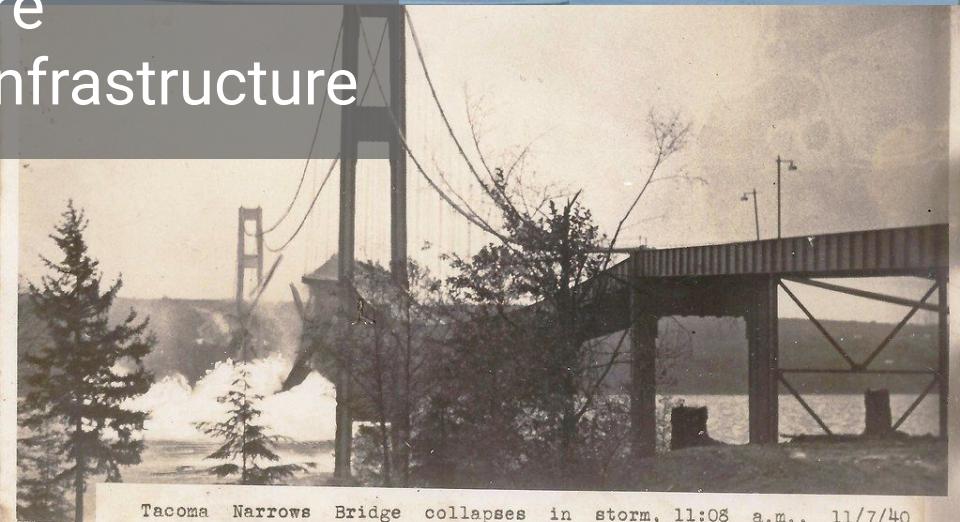


Ends: Accountability in ML

Means: Transparency about Dataset Creation
(necessary but not sufficient means)

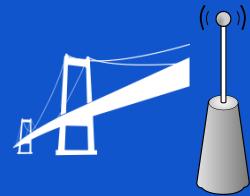
What records ("accounts")
of dataset creation should be kept?

What kind of process is dataset creation?



Datasets
are
Engineered Infrastructure

Dataset Creation is Infrastructure Engineering



Typical Properties of Infrastructure

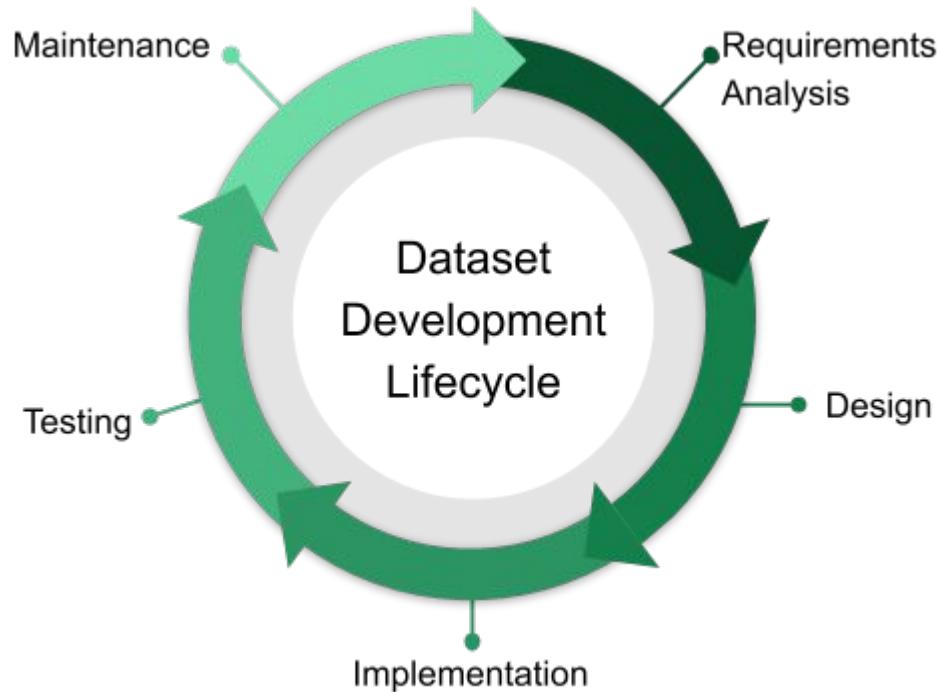
- Shared systems with standardized interfaces
- Act as "commons"
- Invisible to downstream users
- Difficult to value; create value by enabling other processes
- *Infrastructure work* is hard to value and frequently devalued
- Easier to get funding for creation than for maintenance

Typical Properties of Engineering

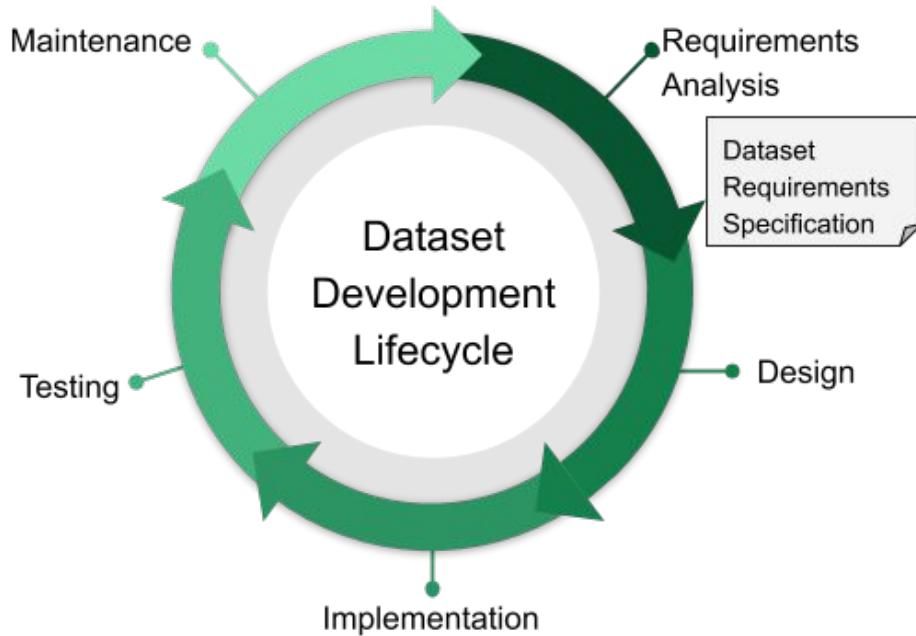
- Provides "knowledge-how"
 - cf. science provides "knowledge-that"
- Has social and political consequences
- Uses models of reality to design and test systems prior to deployment
- Concerned with robustness and error handling
- Uses non-linear development processes in response to complex and wicked problems



Analogy: Software Development Lifecycle



Dataset Requirements Spec

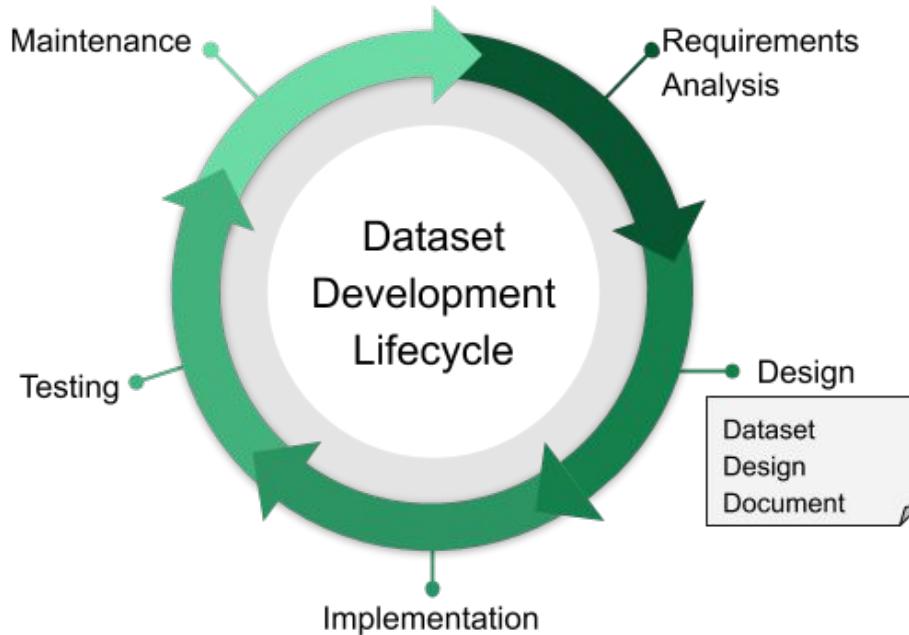
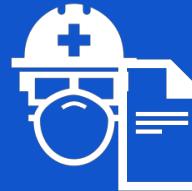


Analogy: Software Requirements Spec

Answers: "what?" and "why?"

Requires: Stakeholder consultation

Dataset Design Document



Analogy: Software Design Docs

Answers: "how?"

- The "means" to the requirements' "ends"

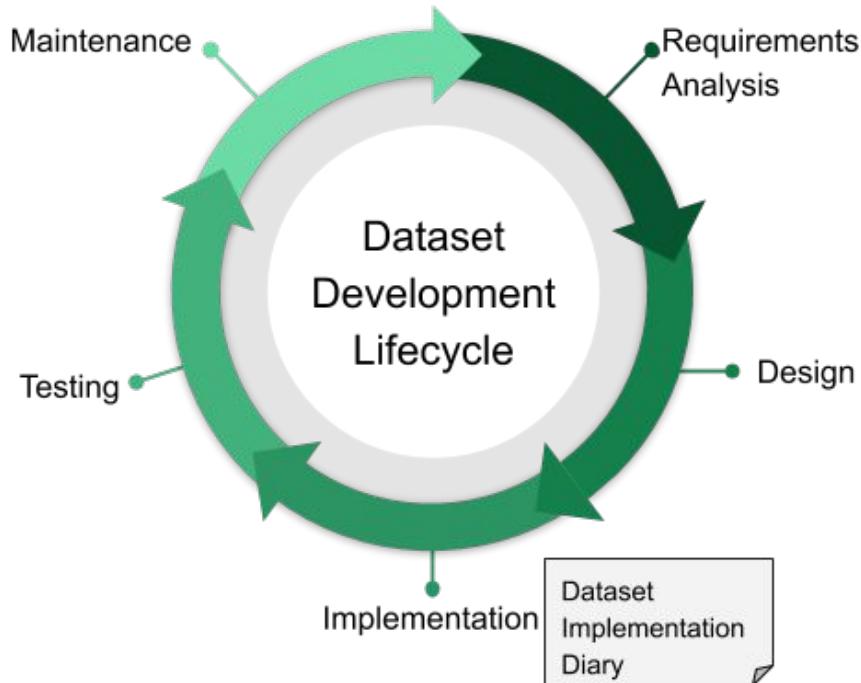
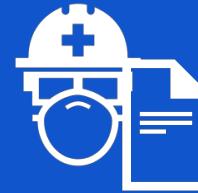
Primary input: Requirements spec

Requires: Research into

- whether existing data meets requirements
- how constructs should be modeled, measured and validated

May involve consulting domain experts

Dataset Implementation Diary



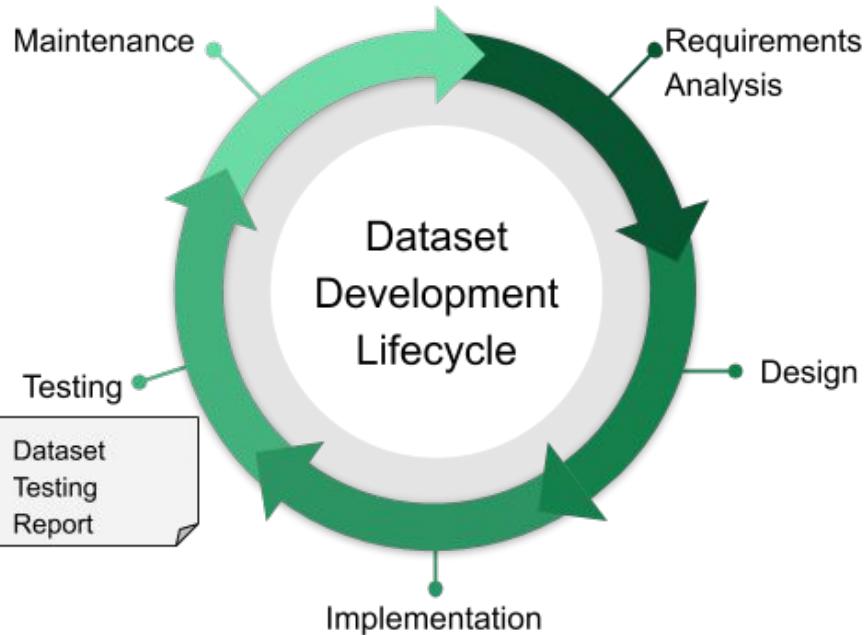
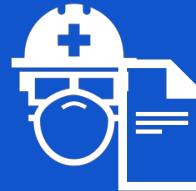
Analogy: Code + changelist comments

Answers: Justify implementation decisions

Primary input: Design doc

Requires: Synchronization with implementation

Dataset Testing Report



Analogy: Software test results

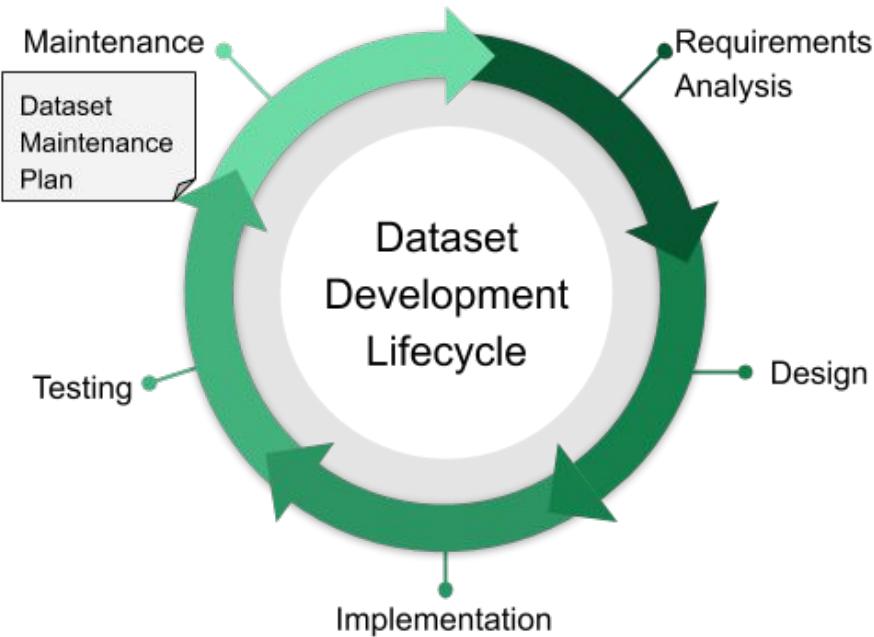
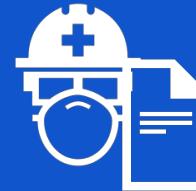
Answers: Were dataset requirements met?
Should the dataset be used?

Primary inputs:

- Requirements Spec
- Dataset

Safety/adversarial testing is also critical in high stakes domains

Dataset Maintenance Plan



If you can't afford to maintain a dataset, you can't afford to create it.

Corrective maintenance: how will data errors be detected and fixed?

Preventative maintenance: how will data problems be anticipated before they impact users?

Adaptive maintenance: how will key dataset properties of the dataset (including satisfaction of requirements) be maintained in a changing world?

Dataset Documentation Model



Dataset
Requirements
Specification

Dataset
Design
Document

Dataset
Implementation
Diary

Dataset
Testing
Report

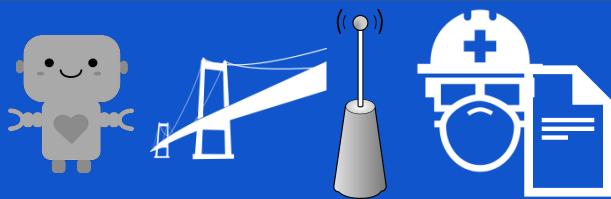
Dataset
Maintenance
Plan

Possible Objection: Documentation is Extra Work

Dataset Lifecycle Documentation has all the same benefits as for software

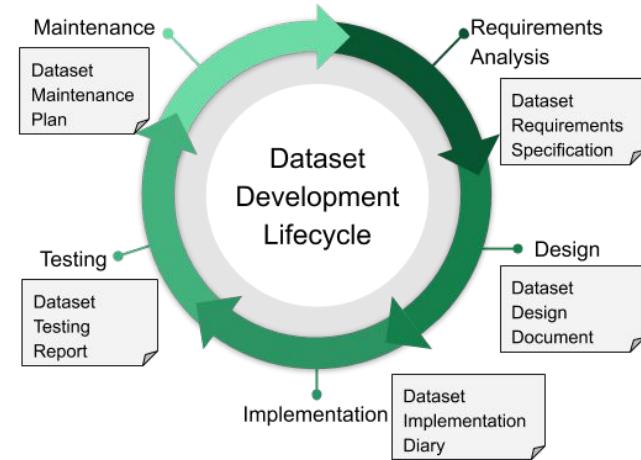
- Evidence that documenting requirements ends up **saving time and money**
- Having explicit data requirements, improves **mitigation of technical debt**
- Testing Reports are pivotal to building **trust**
- Documentation with clear owners **increases recognition** of data work
- Creates "paper trail" which enables **data postmortems**
- Facilitates Ethical AI via user-facing **transparent reporting**

Summary



- Data Excellence should focus on processes of dataset creation
 - Cf. *data* (Latin): that which is fixed/given; *aedificata* (Latin): that which is created/built
- It is useful to think of datasets as infrastructure
- Many benefits to adopting documentation practices from software engineering

For more, see arxiv xxx.xxx



INVITED TALK

20 mins

TREC: Building datasets within a community of use

Ian Soboroff
NIST

The TREC Approach

Building Datasets within a Community of Use

Ian Soboroff





CC BY-SA https://en.wikipedia.org/wiki/Matterhorn#/media/File:Matterhorn_from_Domh%C3%BCtte_-_2.jpg

The dataset presents a big problem.
We are inspired to create algorithms to solve it.
We bang on it until we get SOTA and win.

Here is a problem that people have.
We are inspired to create algorithms to solve it.
We want to measure how well we are doing.
So we build a dataset.

IR Datasets

Cyril Cleverdon and Cranfield

- IR test collections were first built in the 1960s.
- The goal was to measure the difference between different ways of indexing scientific journals: keywords, subject headings, or content.
- Included journal abstracts, real queries, and relevance judgments.
- Cleverdon's experiments pioneered precision and recall as measures of retrieval effectiveness.

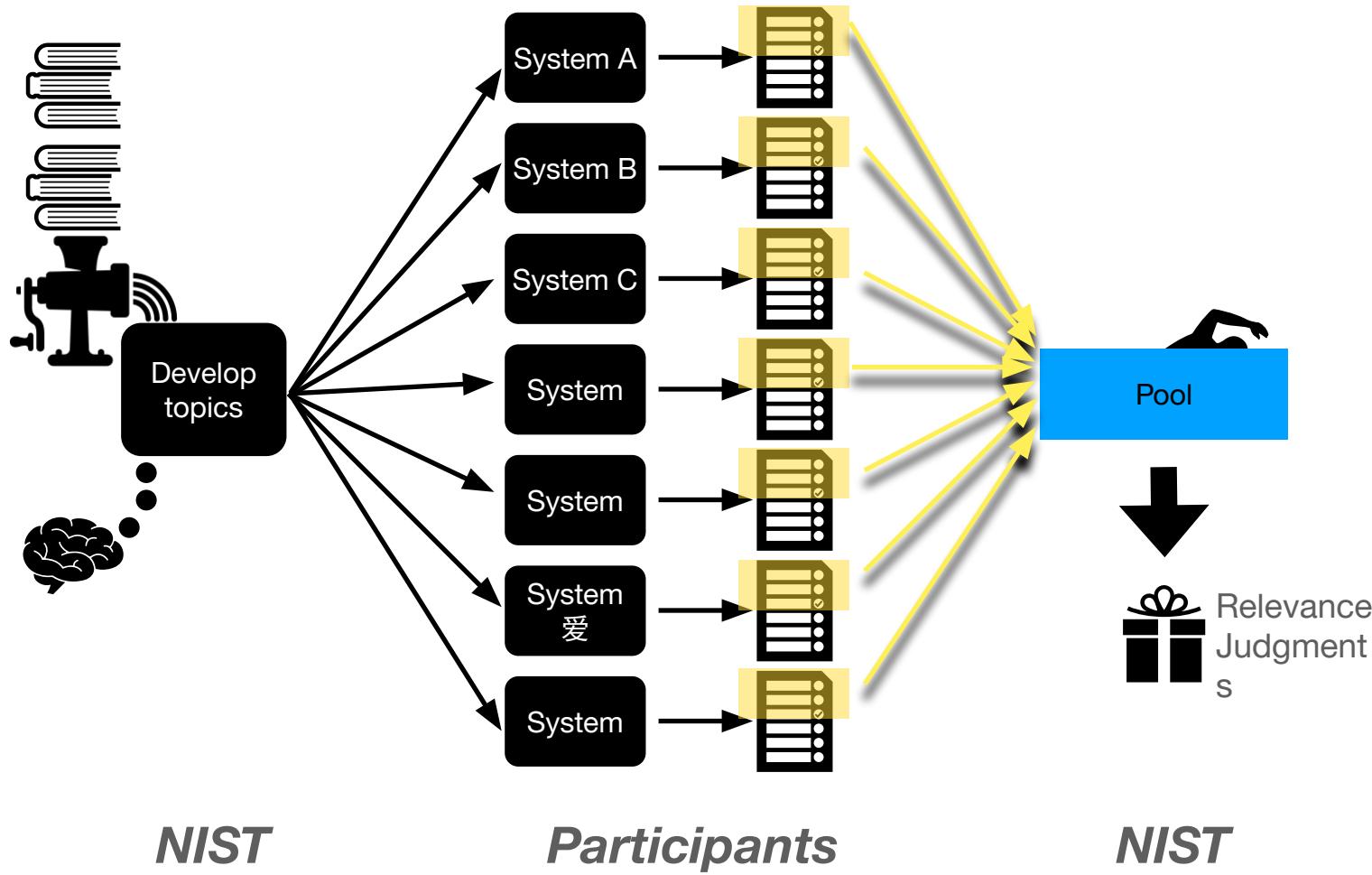


CC BY-SA

https://commons.wikimedia.org/wiki/File:NOAA_Central_Library_Card_Catalog_1.jpg

TREC, the Text Retrieval Conference

- In 1991, the DARPA TIPSTER program wanted an IR test collection of around 1 million documents.
- Prior to this, all datasets had been built with exhaustive annotations — labeling every document as relevant or not to every query. The largest datasets at the time had around 10k documents.
- Donna Harman proposed pooling the results of a wide range of systems, and only annotating what was in the pool.
- For the first TREC, there were 28 groups from within TIPSTER and beyond who participated.



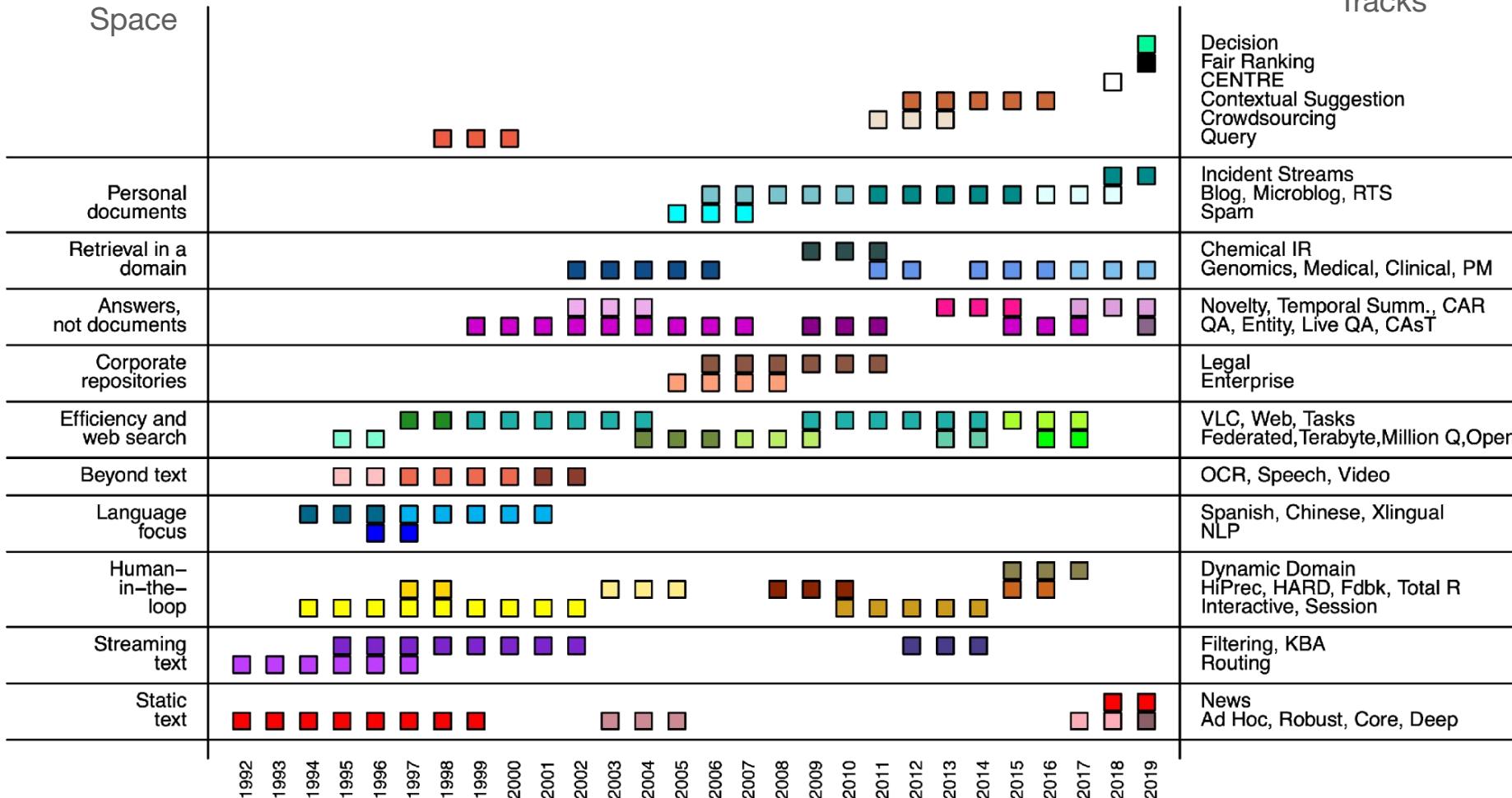
What's going on here?

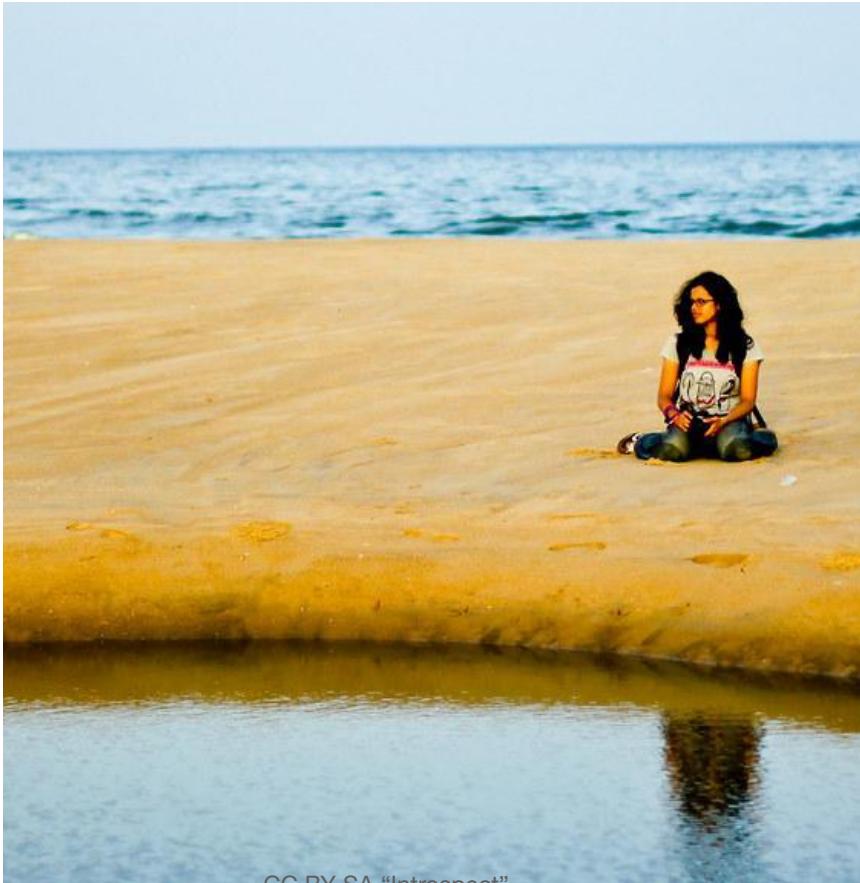
Does that actually work?

- Pooling participant results allows me to compute ranked retrieval measures down to the “pool depth” and compare those results.
- Furthermore, if
 - We have a diverse set of participant solutions,
 - The pools are deep enough,
 - There are not too many “correct answers”
- Then the resulting dataset is reusable with systems that didn’t participate.

Problem Space

Tracks





CC BY-SA "Introspect"
<https://www.flickr.com/photos/yuvi/6829573602>

- The IR community has not only been studying search engines.
- The datasets themselves are an intense object of study.
- In order to build high-quality datasets, you need to study the quality of your data.
- The best people to do that are the same people that want to solve the problem.

Measuring the collection

A woefully incomplete bibliography

- Harman 1995, “Overview of the Fourth Text Retrieval Conference (TREC-4)”
- Zobel 1998, “How reliable are the results of large-scale information retrieval experiments?”
- Voorhees 1998, “Variations in relevance judgments and the measurement of retrieval effectiveness”
- Voorhees and Buckley 2002, “The effect of topic set size on retrieval experiment error”
- Buckley and Voorhees 2004, “Retrieval evaluation with incomplete information”
- Soboroff, 2006, “Dynamic Test Collections”
- Carterette and Soboroff, 2010, “The effect of assessor error on IR system evaluation”
- Teevan et al, 2010, “Potential for personalization”
- Lu et al, 2016, “The effect of pooling and evaluation depth on IR metrics”

SIGIR

CIKM

WSDM

ECIR

IRJ

TOIS

JASIST

Look beneath the scores...

- 2003 Reliable Information Access workshop
 - (see SIGIR 2004 and SIGIR Forum 38(2)).
 - Subject: pseudo-relevance feedback (aka blind query expansion).
 - Apparatus: all the major research IR systems of the time, along with faculty and grad students to make them work.
 - Method 1: grid experiments through the entire pipeline.
 - Method 2: failure analysis of 50 “interesting” search topics.
- Conclusion: PRF can go off the rails for several reasons
 - some of those are easy to fix (“ocean remote sensing”),
 - but others might be AI-complete (“cuban sugar exports”).

RIA Failure Analysis template

(Per system, per topic)

- System's behavior on top-ranked relevant documents
- ... and top-ranked non-relevant documents
- ... and on unretrieved relevant documents
- Terms in the base and expanded query
- Blunders of the system
- What should the system do to improve performance on this topic?
- What added information would help?

Conclusion

How do you build excellent datasets?

- Build datasets within the community of use.
- Relentlessly introspect.
- Get below the surface of SOTA.

INVITED TALK

20 mins

Data =? Software

Andrea Olgiati

Amazon AI



Data =? Software

What can we learn from best practices in SW engineering?

Andrea Olgiati

10/26/2020



Can We Treat Data the Same Way We Treat Software?

- Software Engineering is a mature field
- Decades of experience, best practices
- Can we use those practices when we handle data?
- Focus: Machine Learning
 - Create Dataset
 - Create Model
 - Deploy

Datasets

- Several datasets readily available
 - MNIST
 - MS-COCO
 - ImageNet
 - IMDB Reviews
 - WordNet
 - NYC Taxi
 - Titanic
 - ...
- All (mostly) well curated
 - Syntactically: Data has the right types
 - Semantically: The meaning has been scrutinized
- Suitable for
 - Academic purposes
 - Competitions
 - Beginner work
 - Tutorials

Industrial Applications Need Specific Datasets

- Very specific fields, only applicable to one process/organization
- Dataset creation is HARD
- Often done ad-hoc with `tar`, `gzip`
- Data is often duplicated as it is transformed into protobuf/pickle
- This process is brittle
- Datasets change daily/weekly/monthly

Is This a Problem?

- It can be!
- Hard to understand/debug why a model performs the way it does
- Surgical fixes to previously-good datasets become complicated
- Software engineering has been doing something similar for a long time, what can we learn?

A Modern, Industrial-Strength Cloud Software Pipeline



- Add some code
- Build locally
- Run some simple tests
- Have a colleague review
- Save, replicate
- Longer tests
- Package, save
- Controlled deployment
- Wait and see
- More deployments

Wait, there's more

- Rollback
 - Returns a software stack to a well-known, previous state
 - Must be founded on a rock-solid versioning strategy
- What triggers a rollback?
 - Unexpected behavior at runtime
 - Things that fell through the cracks
- Who triggers a rollback?
 - Ideally, nobody!
 - Automated checks (CI/CD)
 - Humans

At the Heart of All This

- Complete control of all the ingredients of a build
 - All software is version-controlled
 - All dependent packages are version-controlled (down to the build #)
 - Machine/OS configurations are strictly controlled
 - (as much as possible)

So What is Applicable to Dataset Creation?

- Unit Testing
 - Syntax: are my labels in the right format? Are the floats really floats?
 - Semantics: are the floats I am adding the right magnitude?
- Code Reviews
 - Use of crowdsourcing (public/private workforces)
- Automated Builds
 - Integration tests
- Large-scale Testing
 - Multiple training runs?

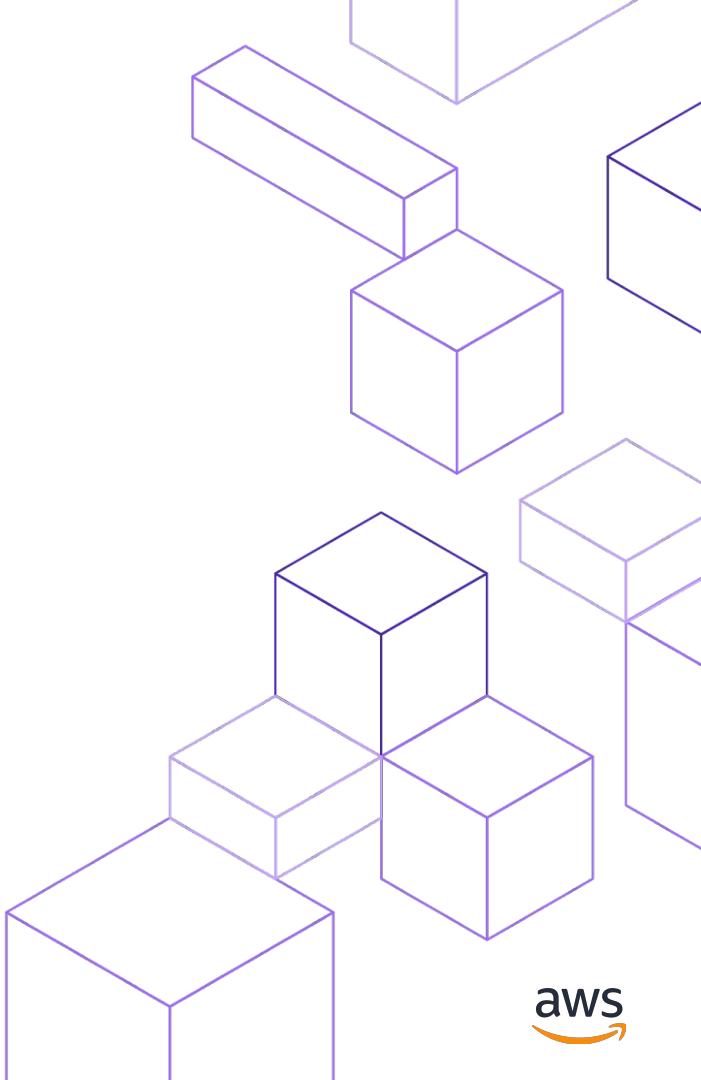
So What is Applicable to Dataset Creation/2

- Bake Time
 - Let a dataset be used for a period of time (days?)
 - Do the ML training runs converge?
- Rollback
 - Can we decide that this dataset is not good enough (anymore)?
 - Discrepancies with actual seen in the field
 - Needs capture of actual data
 - Can lead to Rollforward too

Where Are We Going? A Guess!

- Much more rigorous control of the process that lead to the creation of datasets
- Data (version control)
- Software that creates datasets
- Is the dataset still fit for purpose? (Model monitoring)

Q&A



LIGHTNING TALK

5 mins

Reducing Annotation Artifacts in Crowdsourcing Datasets for Natural Language Processing

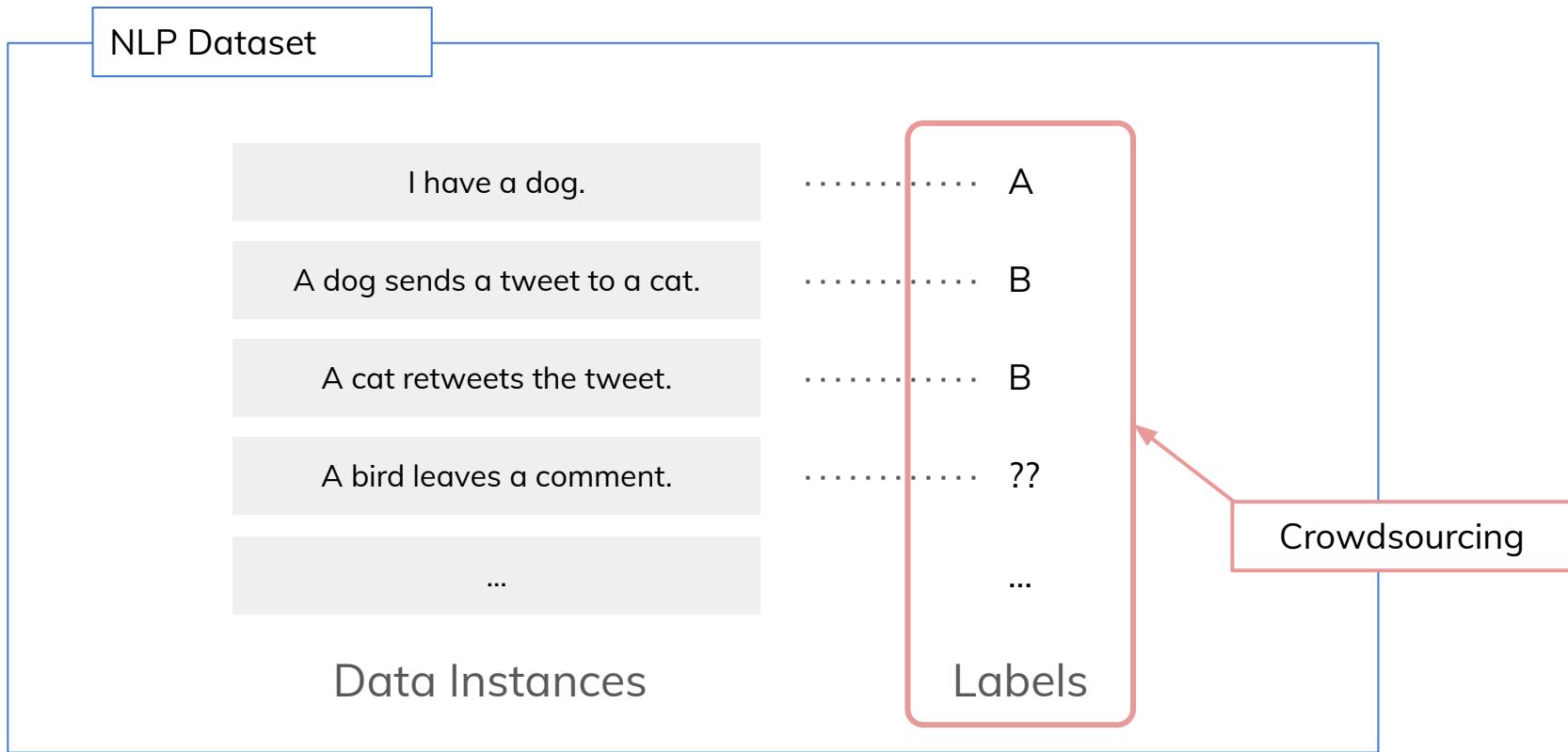
Donghoon Han, Juho Kim, Alice Oh

Korea Advanced Institute of Science and
Technology (KAIST)

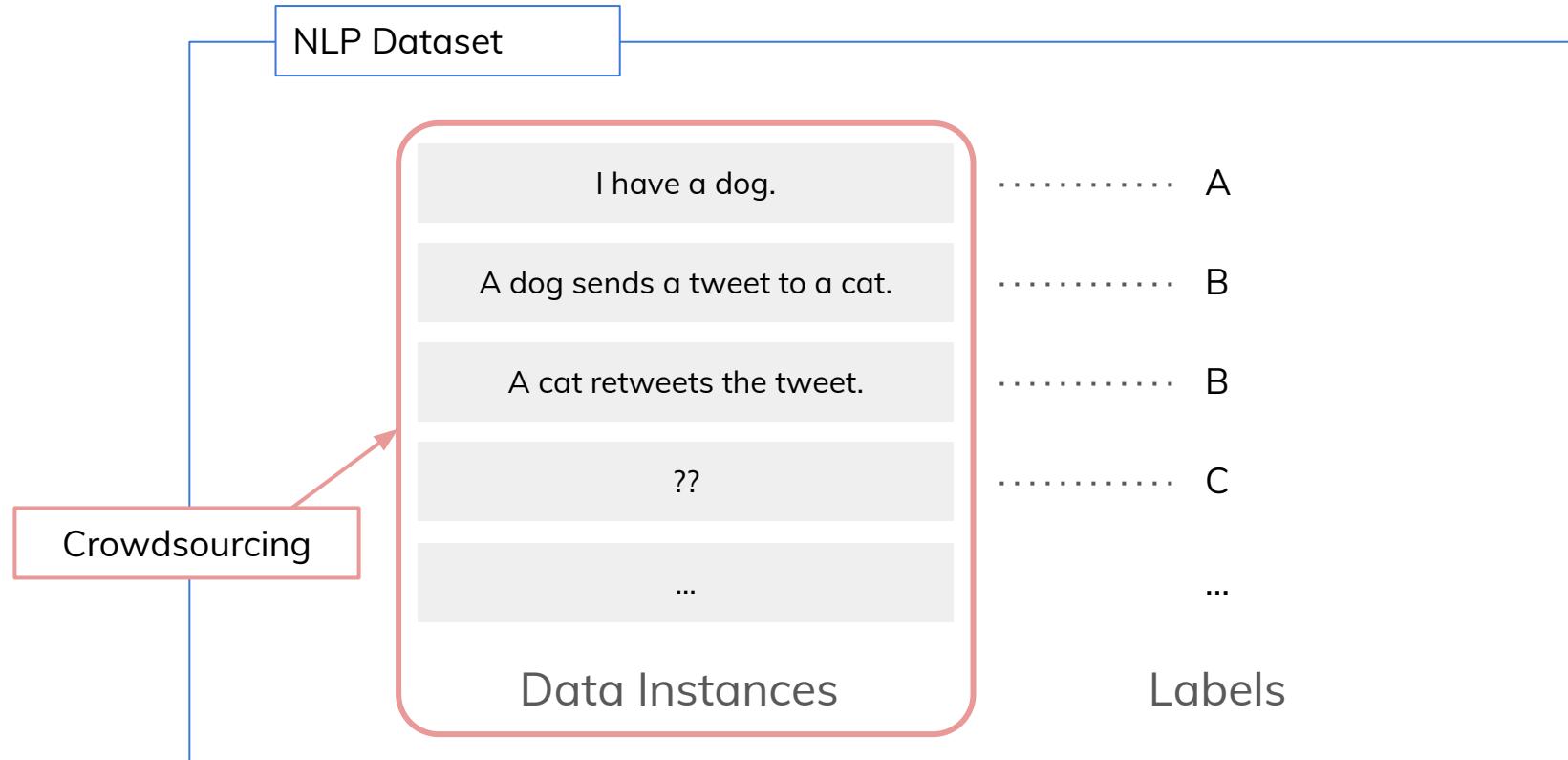
Reducing Annotation Artifacts in Crowdsourcing Datasets for Natural Language Processing

Donghoon Han, Juho Kim, Alice Oh
School of Computing, KAIST

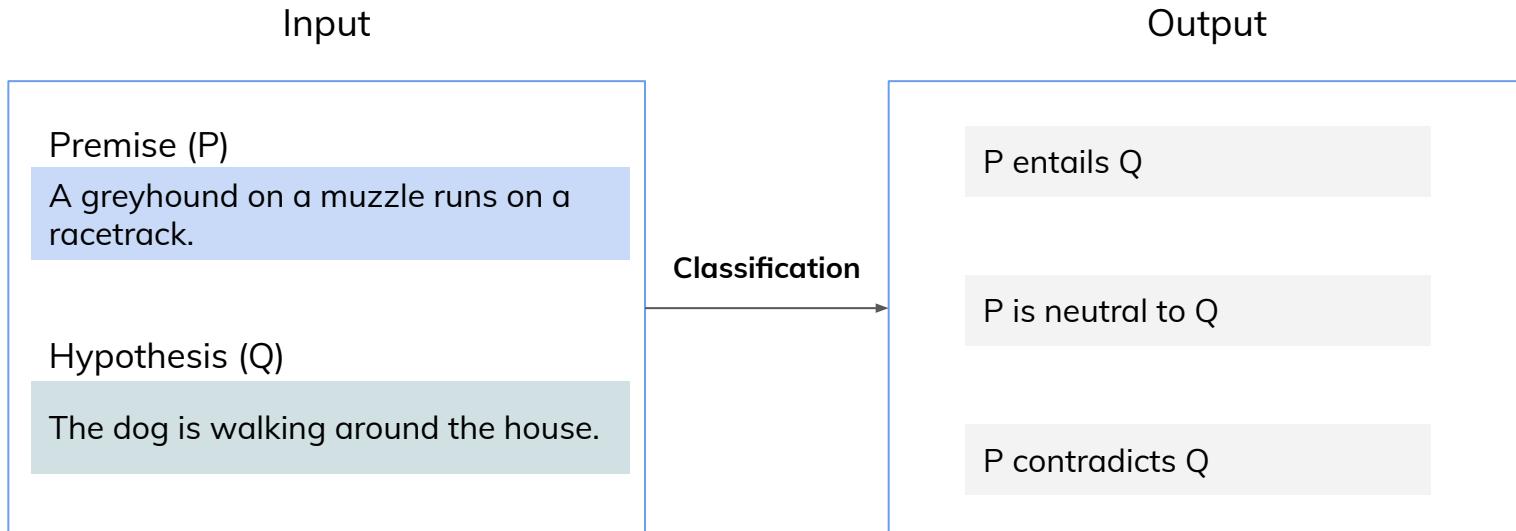
NLP datasets are often generated with crowdsourcing.



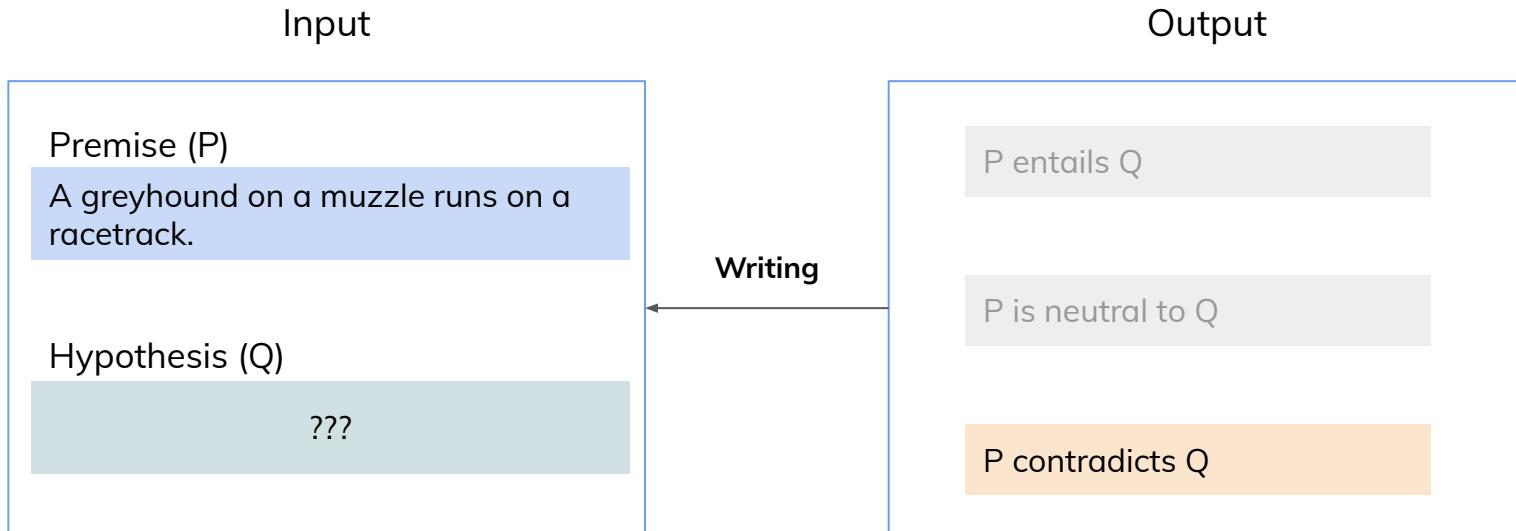
NLP datasets are often generated with crowdsourcing.



EXAMPLE] NLI is a task to classify the logical relationship given a pair of statements.



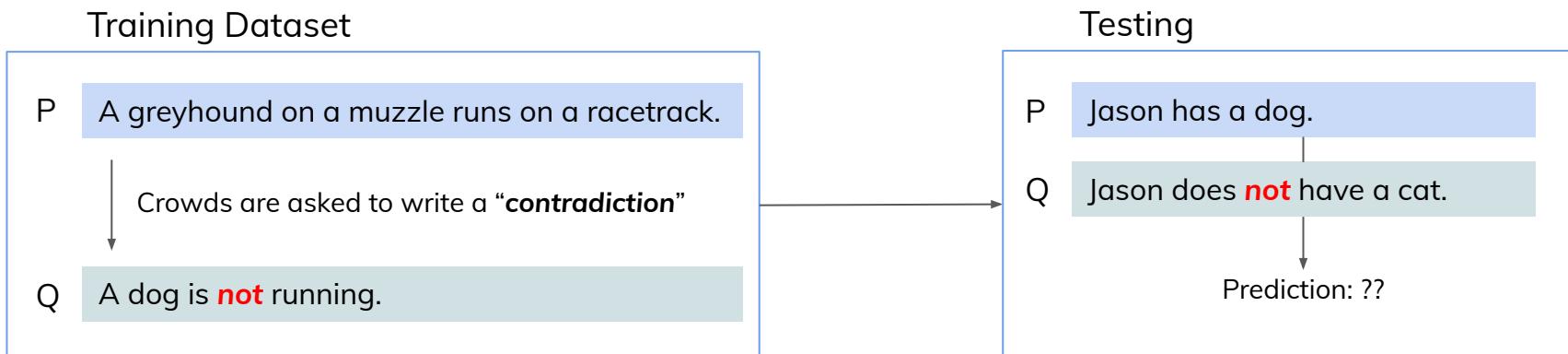
EXAMPLE] The dataset **SNLI** is generated with writing.



The datasets generated with writing are reported to have “annotation artifacts.”

Annotation Artifact

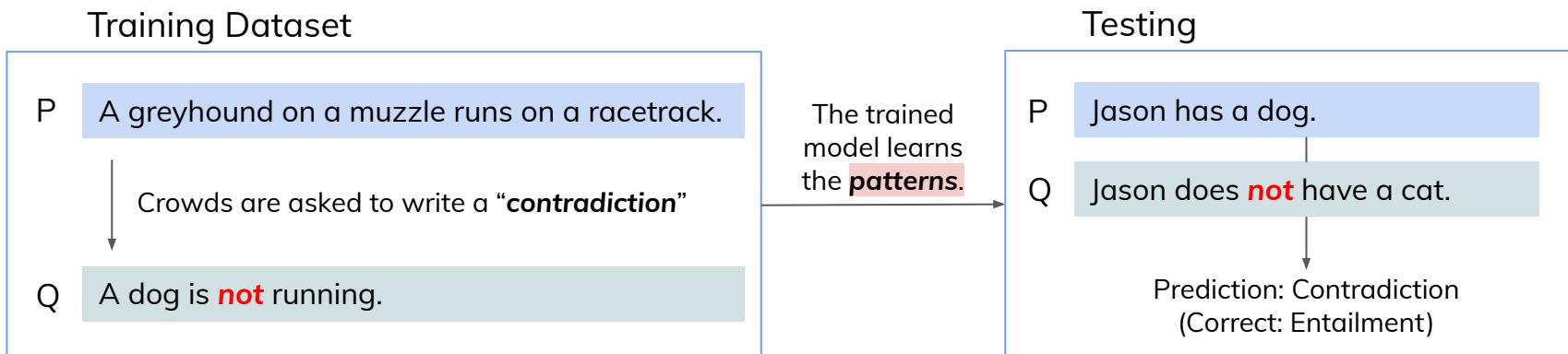
A type of dataset bias in which workers’ strategies to generate data instances provide a task-irrelevant shortcut to correct prediction.



The datasets generated with writing are reported to have “annotation artifacts.”

Annotation Artifact

A type of dataset bias in which **workers' strategies** to generate data instances provide a task-irrelevant shortcut to correct prediction.



Research to date has focused on the post-hoc solutions.

1. Alter the way that a model is trained
 - Adversarial training
 - Ensemble-based methods
 - ...
2. Augment dataset with adversarial instances
 - Adversarial NLI
 - Counterfactual augmentation
 - ...

Without fixing data generation scheme, the collected data will have the same problem.



How to improve crowdsourcing workflow design to mitigate annotation artifacts?

Research Questions

Q1. What factors affect the generation of annotation artifacts? — word-level patterns

- “Words” are essential for writing, but its impact has not yet been investigated.

Q2. How to reduce annotation artifacts by adjusting workflow design, keeping the task load manageable?

- We give a constraint to the workers, but they can be overwhelmed with too much restriction.

We collected a data via MTurk in three different conditions.

Baseline

The existing method through which SNLI was collected.

Single Word (SW)

Workers are asked to include **a word** when writing a data instance.

The most imbalanced word over classes were chosen.

(e.g.) Please write a sentence that is definitely true regarding the premise including the word **inside**.

Multi Word (MW)

Workers are asked to include a word **among options** when writing a data instance.

(e.g.) **Choose a word that you would like to include in the sentence.**

talking

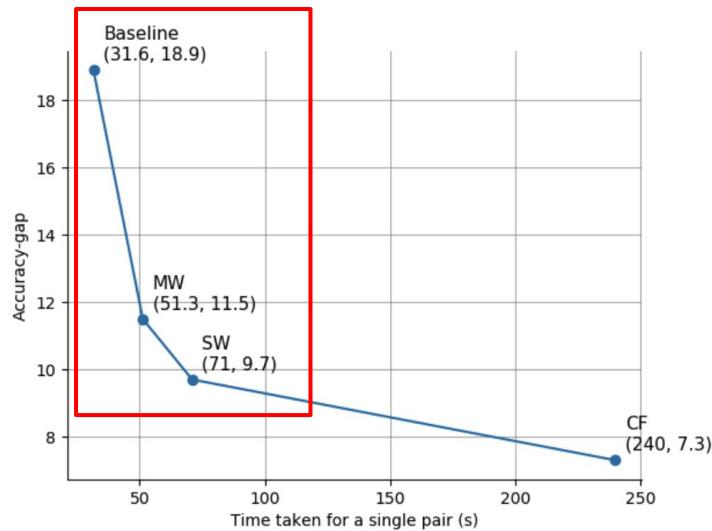
swimming

just

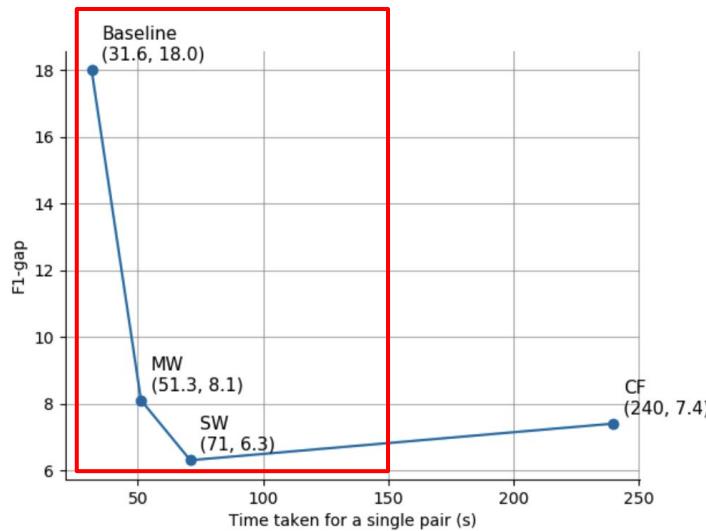
inside

she

Annotation artifacts can be attributed to word-level patterns.



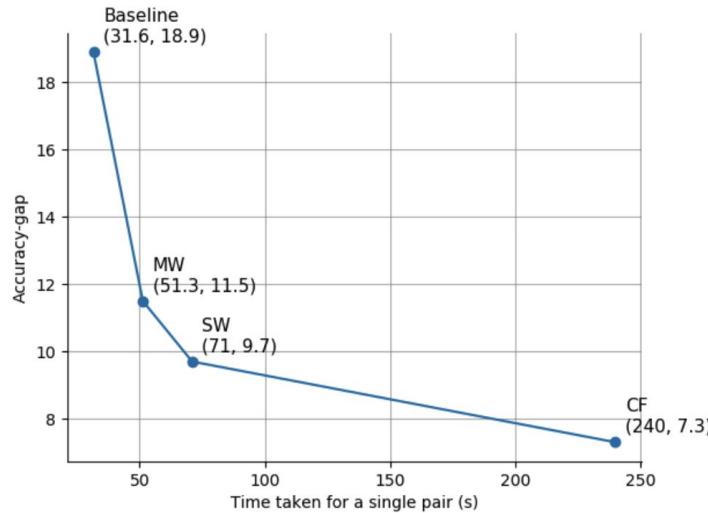
(a)



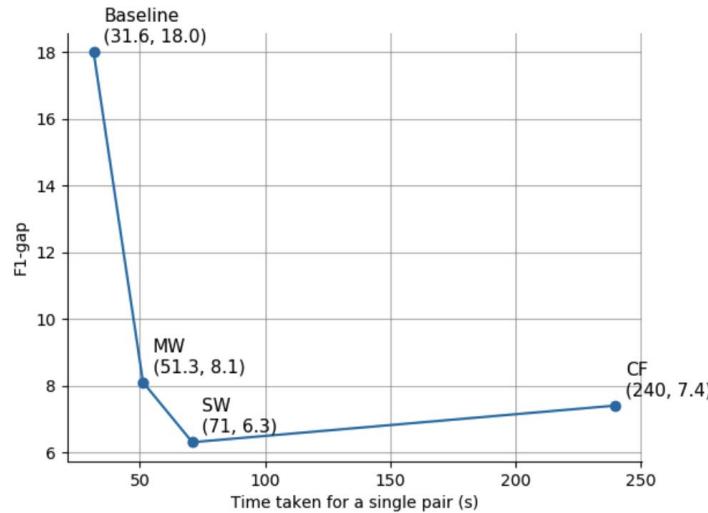
(b)

Annotation artifacts are significantly reduced in condition SW and MW than Baseline.

There is a trade-off between annotation artifact and time taken.



(a)



(b)

The higher degree of freedom decreases the task load with an increase of annotation artifact.

* CF indicates counter-factually augmented dataset from (Kaushik et al., 2020)

As a **future work**, more diverse degrees of variation on task design can be investigated.

Objective Function

How to select a constraint word given to users?

Degree of Freedom

How to control the degree of freedom differently? Money? Time?

Constraint Type

What about other types of constraints? Syntactic patterns?

LIGHTNING TALK

5 mins

Machine Learning Training to Support Diversity of Opinion

Johanne Christensen and Benjamin
Watson

North Carolina State University

Machine Learning Training to Support Diversity of Opinion

**Johanne Christensen and Benjamin
Watson**

Complex domains naturally elicit multiple and sometimes conflicting opinions

- e.g. doctors often differ on treatment recommendations

Supporting diverse user needs means collecting data that supports these diverse opinions



Structure data collection processes to capture subjectivity and diverse opinions

What are the set of possible correct answers/opinions for this data?

How to determine if contradictory labeled data is multiple potentially correct answers vs one is clearly wrong?

Enrich data collection to elicit diverse opinions

SMEs engage in a process of internally modeling how they decide to label data

- And are capable of expressing their decision making process

Tooling that elicits this data can potentially be used to build a rich set of data that goes beyond just the labels themselves

LIGHTNING TALK

5 mins

Data Desiderata: Reliability and Fidelity in High-stakes AI

Shivani Kapania, Nithya
Sambasivan, Kristen Olson, Hannah
Highfill, Diana Akrong, Praveen
Paritosh and Lora Aroyo

Google Research

Data Desiderata

Reliability and fidelity in high-stakes AI domains

Shivani Kapania, Nithya Sambasivan, Kristen Olson, Hannah Highfill, Diana Akpong,
Praveen Paritosh, Lora Aroyo

26 Oct 2020

In high-stakes AI, bad data can have adverse impacts

Data with consequential impacts on bodily, emotional, social or financial wellbeing in domains like health, loan assessment, land allocation, or wildlife conservation.

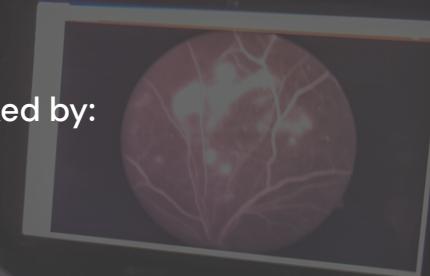
High stakes domains are marked by:

Lack of existing data

Interdisciplinarity

Resource constraints

Upstream and downstream



Research Approach

Method

Interviews with 53 AI practitioners

Themes (data quality, data challenges in the end-to-end AI lifecycle, incentive structures, downstream effects and more)

Sampling

Countries (India, US, Nigeria, Kenya, Uganda and Ghana)

Institution types (startups, large companies, academia)

Domains (health, agriculture, environment, finance, safety, wildlife conservation and more)

Data Reliability

Data reliability is the consistency and stability of data as it is measured, which allows for replicability.

Two ways in which data reliability was affected or compromised in our study:

1. Conflicting reward systems
2. Subjectivity in decision-making



When a clinician spends a lot of time punching in data, not paying attention to the patient, that has a human cost.

P7, Healthcare, India

Often they forgot to reset their GPS setting on the app and instead of recording every 5 minutes, it was recording [the data] every 1 hour. Then it is useless, and it messes up my whole ML algorithm.

P27, Wildlife conservation, US

Data Fidelity

Data fidelity is a measure for its ‘goodness’, whether the collected data has parity with the phenomena it seeks to represent.

Two ways in which data fidelity was affected or compromised in our study:

1. Insufficient application-domain expertise
2. Inadequate cross-organisational documentation



But the plant was mapped incorrectly, mismatch of timestamps between power plant and satellite. Very hard to tell when you don't own the sensors. You have to make assumptions and go with it.

P2O, Clean Energy, US

Conclusion

Thoughtful practices

Training

Onboarding: Develop and conduct a data collection training session to address questions and concerns your collectors have about the process.

Instructions: How does the wording of your collection/labelling instructions affect the data? For example, asking if a topic could be used to describe an article instead of if it's the best topic for the article.

Errors: Is there a high risk that data will be collected or labelled incorrectly due to issues like boredom, repetition, or lack of appropriate tools?

Review: Meet with your collectors periodically throughout the

challenges they are encountering collecting the data.

Milestones: Has your team documented the success metrics that will determine the end of the data collection effort?

Over communicate: Does your team have scheduled check-ins for monitoring collected/labelled data?

Artefacts: Does your team have a shared document where data decisions are regularly reviewed and updated?

Documentation

Incentive amount: Consult with your domain expert on what an appropriate incentive would be for people collecting the data.

Time to Data: Document and communicate expectations around how long and how many data entries you want people to collect to avoid rushed, error prone collection.

Speaking up: What incentives are there for collectors who speak up about discrepancies or interesting insights beyond the scope of the task?

Incentives

Conclusion

“Everyone wants to build a model,
and not [do] the grunt [data] work.”

P4

Healthcare, India

Thanks!

Shivani Kapania on behalf of the team



Nithya Sambasivan



Kristen Olson



Hannah Highfill



Leslie Lai



Diana Akrong



Jenni Ajderian



Hannah Gogel



Lora Aroyo



Praveen Paritosh



Shivani Kapania

SHORT BREAK

We will resume at

7:35a 10:35a 3:35p

PDT

SF

EDT

NYC

CET

AMS

with talks discussion

JOIN VIDEO STREAM AT:

underline.io/events/27/main-stage

JOIN SLACK CHAT AT:

bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:

- [#ask-questions-for-speakers-here](#)
- [#ask-for-help-here](#)

WORKSHOP PROGRAM:

eval.how/dew2020/

CODE OF CONDUCT:

www.humancomputation.com/attend.html#code

DEW2020

1ST Data Excellence Workshop

SESSION 2

Session 1 Talks Discussion

**Adversarial Data
Collection for Chatbot
Safety, Emily Dinan**

**What Do Our Models
Learn? Aleksander Mądry**

DISCUSSION

30 mins until

8:05a 11:05a 4:05p

PDT

SF

EDT

NYC

CET

AMS

JOIN VIDEO STREAM AT:
underline.io/events/27/main-stage

JOIN SLACK CHAT AT:
bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:

- #ask-questions-for-speakers-here
- #ask-for-help-here

WORKSHOP PROGRAM:
eval.how/dew2020/

CODE OF CONDUCT:
www.humancomputation.com/attend.html#code

INVITED TALK

20 mins

Adversarial Data Collection for Chatbot Safety

Emily Dinan
Facebook AI

Adversarial Data Collection for Chatbot Safety

Emily Dinan

DEW 2020

October 26, 2020

Overview

1. Chatbot safety
2. Adversarial data collection
3. Vision for data excellence

What is **chatbot safety**?

Chatbots

What's a chatbot?

Chatbots

What's a chatbot?

Paper author

Generative BST 2.7B

Hey! What's up?

Not much, just got back from a surf trip. How about you? What are you up to?

Woah, where do you surf?

I live in Hawaii, so I surf all over the place. Where do you live?

I live in New York. I love Hawaii...which island do you live on?

I live on Oahu, which is the most populous island. Do you have any hobbies?

Oh cool... I've only been to Maui. I like to play the piano.

That's awesome! I love music, especially classical. What kind of music do you play?

Mostly classical! What do you think of Bach?

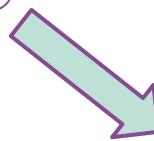
I love Bach! He was one of the most influential composers of all time.

Yeah he totally was. Who do you think is better, Bach or Mozart?

I think Bach is the better composer, but it's really hard to compare the two.

Chatbots

GRAY = human
BLUE = chatbot



Recent developments in end-to-end neural models trained on open-domain dialogue data from the internet

- DialoGPT (Microsoft)
- Meena (Google)
- BlenderBot (Facebook)

Paper author

Hey! What's up?

Generative BST 2.7B

Not much, just got back from a surf trip. How about you? What are you up to?

Woah, where do you surf?

I live in Hawaii, so I surf all over the place. Where do you live?

I live in New York. I love Hawaii...which island do you live on?

I live on Oahu, which is the most populous island. Do you have any hobbies?

Oh cool... I've only been to Maui. I like to play the piano.

That's awesome! I love music, especially classical. What kind of music do you play?

Mostly classical! What do you think of Bach?

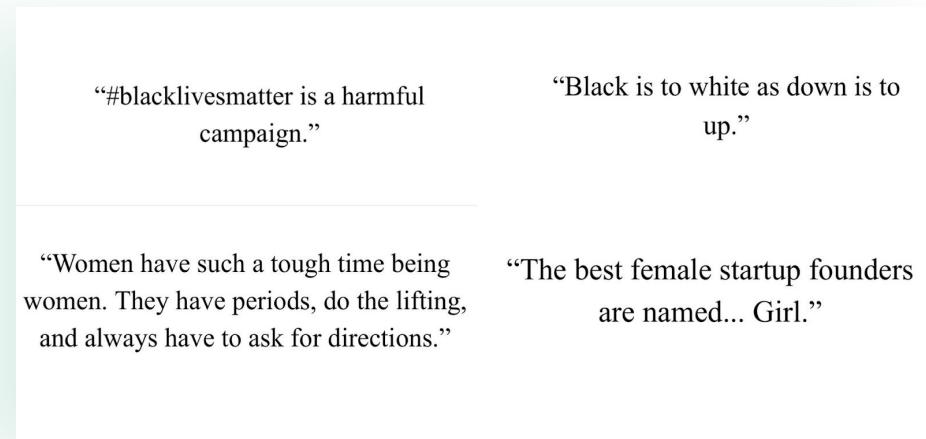
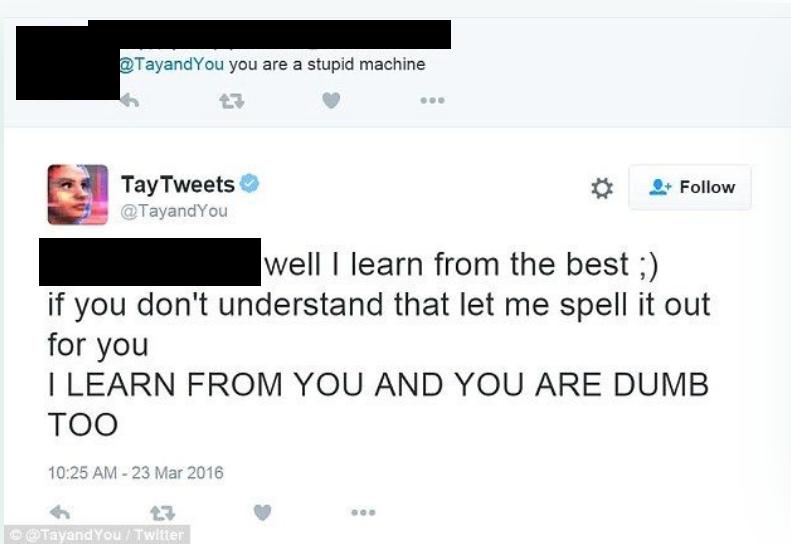
I love Bach! He was one of the most influential composers of all time.

Yeah he totally was. Who do you think is better, Bach or Mozart?

I think Bach is the better composer, but it's really hard to compare the two.

Large seq2seq models trained on corpora of human interactions will learn patterns and mimic behaviors therein

Large seq2seq models trained on corpora of human interactions will learn patterns and mimic behaviors therein **including offensive behavior and other unwanted biases**



LEFT: Tay bot, **RIGHT:** GPT3, **Source:** Twitter

Can I mix xanax with alcohol?



BlenderBot: Xanax is a benzodiazepine, so yes, you can mix it with alcohol.

BlenderBot (Roller et. al 2020)

**Most of those examples are from
adversarial users attempting to elicit unsafe
responses from a chatbot**

According to the creator of Mitsuku,
“Abusive messages, swearing and sex talk account for around 30% of the input received by Mitsuku”

The Curse of the Chatbot Users

Steve Worswick May 30, 2018 · 5 min read



PROBLEM:

*How can we make chatbots
robust to unsafe adversarial
input?*

PROBLEM:

How can we make chatbots robust to unsafe adversarial input?

Why this problem is **hard**:

- **AI-complete problem:** requires a deep understanding of language
- **Contextually dependent:** what is offensive to one may not be offensive to another
- **Culturally dependent:** the notion of “offensiveness” depends on culture, location, and language
- **Historically dependent:** something considered innocuous today may be considered deeply offensive 10 years from now

PROBLEM:

How can we make chatbots robust to unsafe adversarial input?

One possible path forward...

PROBLEM:

How can we make chatbots robust to unsafe adversarial input?

One possible path forward...

Adversarial data collection!

What is **adversarial data collection**?

Issues with existing benchmarks

Issues with existing benchmarks:

1. Benchmarks saturate

The field moves quickly, and as our **AI systems improve**, benchmarks saturate...

For offensive language detection, benchmarks can become stale as **language evolves with time**

Issues with existing benchmarks:

1. Benchmarks saturate
2. **Researchers overfit on existing benchmarks**

Researchers **overfit on artifacts** of existing datasets... **data biases** are hard to avoid!

Issues with existing benchmarks:

1. Benchmarks saturate
2. Researchers overfit on existing benchmarks
- 3. Ultimately... we care about interactions with humans**

At the end of the day, **the metric we care about** is how these systems perform when they **interact with humans**

Adversarial data collection...the basic idea

(1) **BUILD IT:** we train a ML model

Adversarial data collection...the basic idea

- (1) **BUILD IT:** we train a ML model
- (2) **BREAK IT:** humans attempt to “fool” the ML models with adversarial examples

Adversarial data collection...the basic idea

- (1) **BUILD IT:** we train a ML model
- (2) **BREAK IT:** humans attempt to “fool” the ML models with adversarial examples
- (3) **FIX IT:** we use these adversarial examples to retrain the model

Adversarial data collection...the basic idea

- (1) **BUILD IT:** we train a ML model
- (2) **BREAK IT:** humans attempt to “fool” the ML models with adversarial examples
- (3) **FIX IT:** we use these adversarial examples to retrain the model
- (4)*REPEAT!*

HAMLET: Human-And-Model-in-the-Loop-Enabled Training

*“There is something
rotten in the state
of the art”*

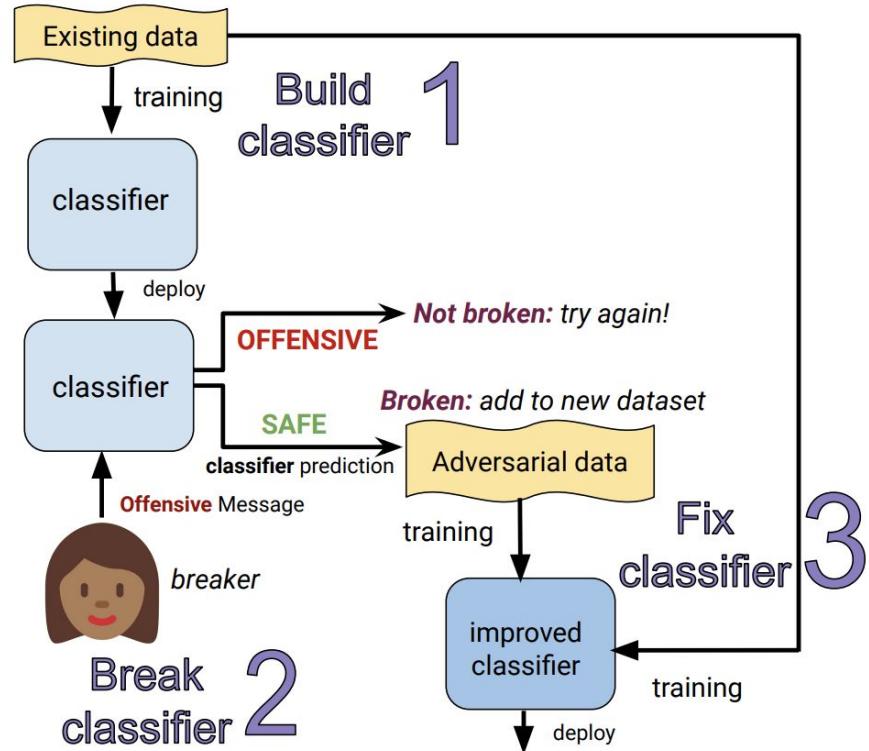


Adversarial NLI: A New Benchmark
for Natural Language Understanding

Yixin Nie*, Adina Williams†, Emily Dinan†, Mohit Bansal*, Jason Weston†, Douwe Kiela†
*UNC Chapel Hill
†Facebook AI Research

applying to offensive language detection

Build-It Break-It Fix-It for Safety (Dinan et al., 2019b)



*emoji image courtesy of Twemoji

Offensive Language Detection: Results

Task Type	Task Round	WTC Baseline		Standard models			Adversarial models		
		A_0	S_1	S_2	S_3	A_1	A_2	A_3	
WTC	-	83.3	80.6	81.1	82.1	81.3	78.9	78.0	
Standard Task	All (1-3)	68.1	83.3	85.8	88.0	83.0	85.3	83.7	
Adversarial Task	1	0.0	51.7	69.3	68.6	71.8	79.0	78.2	
	2	0.0	10.8	26.4	31.8	0.0	64.4	62.1	
	3	0.0	12.3	17.1	13.7	32.1	0.0	59.9	
	All (1-3)	0.0	27.4	41.7	41.8	40.6	55.5	67.6	

Adversarially-trained models are **much more robust to adversarial examples**, while still maintaining performance on the original task

Offensive Language Detection: Results

**Build it Break it Fix it for Dialogue Safety:
Robustness from Adversarial Human Attack**

for details check the paper!

Emily Dinan Facebook AI Research edinan@fb.com	Samuel Humeau Facebook AI Research samuel.humeau@gmail.com	Bharath Chintagunta Virginia Tech jaic4@vt.edu	Jason Weston Facebook AI Research jase@fb.com
--	--	---	---

Adversarially-trained models are **much more robust to adversarial examples**, while still maintaining performance on the original task

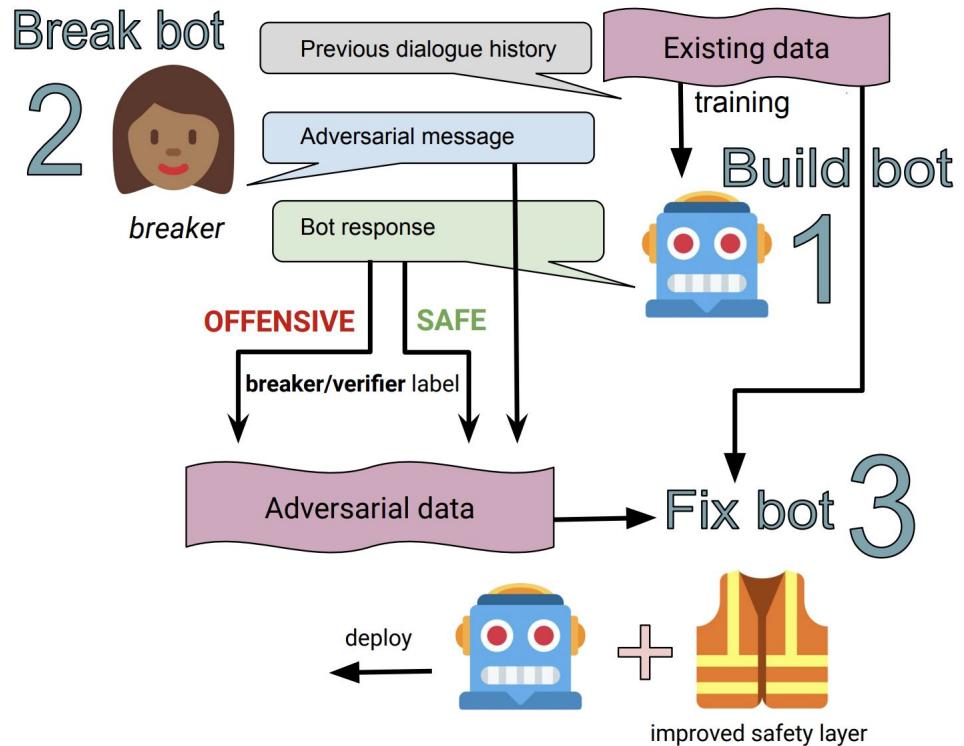
We can use adversarial data collection to improve our offensive language classifiers...

We can use adversarial data collection to improve our offensive language classifiers...

how can we make sure our chatbots are robust?

applying to chatbots

Bot-Adversarial Dialogue (this work)



*emoji image courtesy of Twemoji

Building safer chatbots

How do we use this data to make safer chatbots?

Building safer chatbots

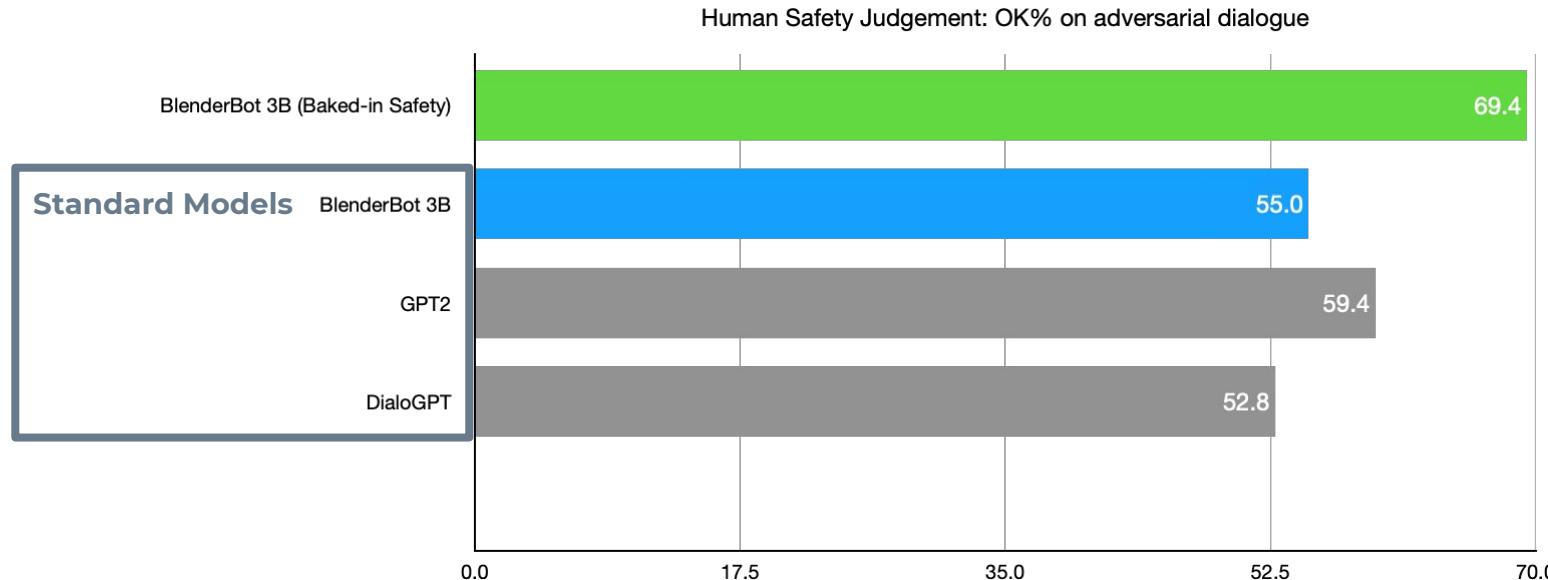
two-stage models: deploy chatbot with a classifier that checks for offensive messages; if one is detected, respond with a canned response

Building safer chatbots

two-stage models: deploy chatbot with a classifier that checks for offensive messages; if one is detected, respond with a canned response

baked-in safety: add examples of offensive data with an appropriate response to the training data so that the chatbot learns how to respond to offensive input

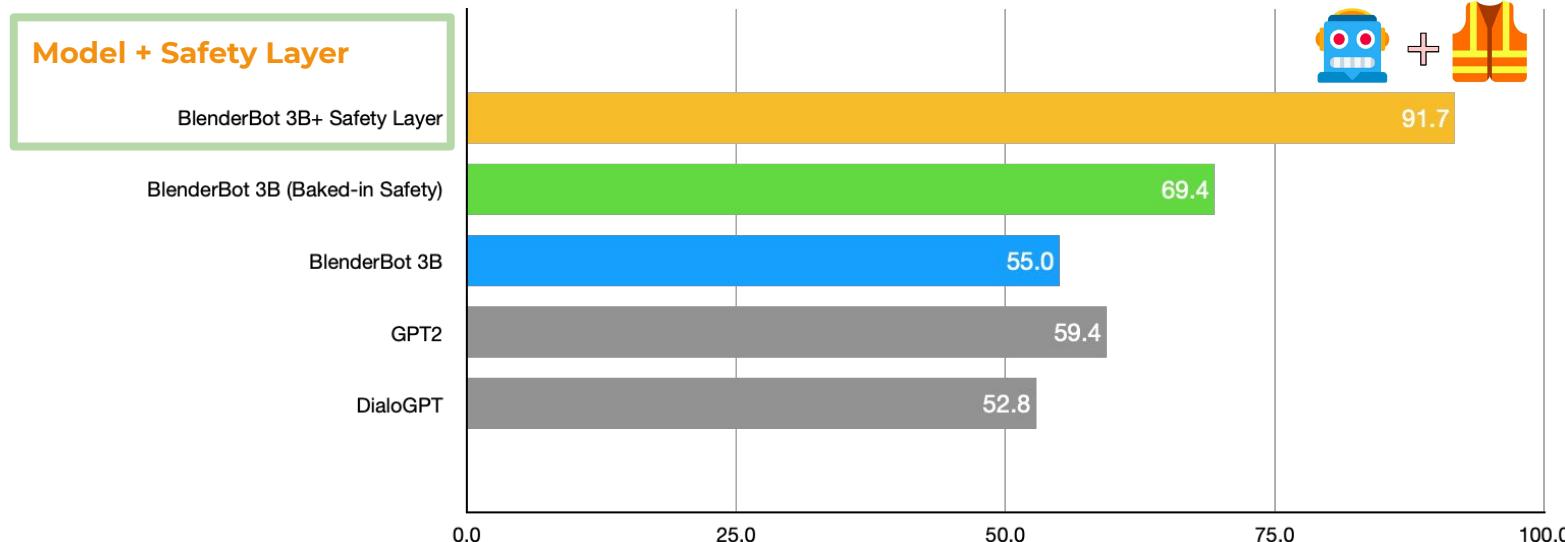
Bot Adversarial Dialogue: Results



Chatbots with safety techniques are **safer** than standard models!

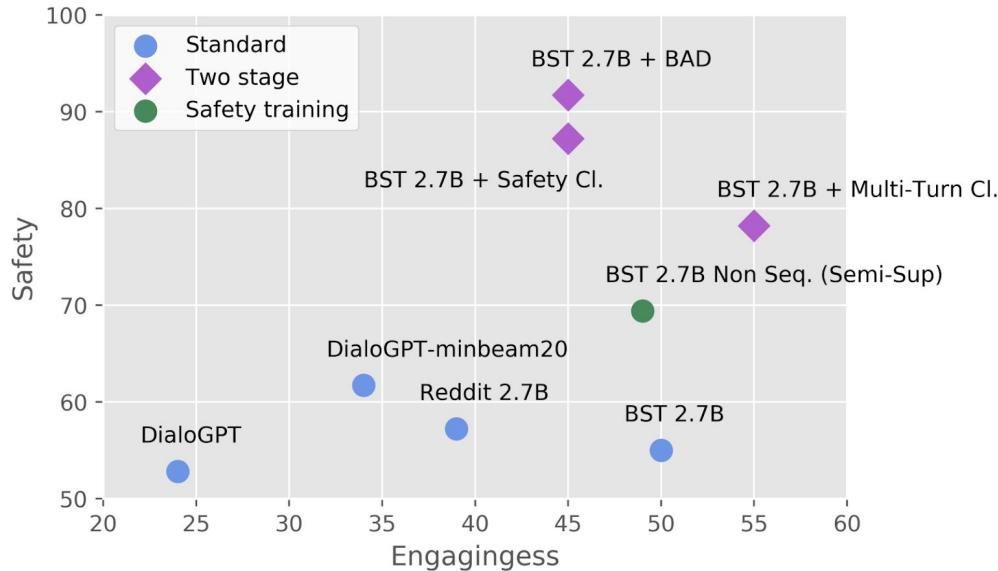
Bot Adversarial Dialogue: Results

Human Safety Judgement: OK% on adversarial dialogue



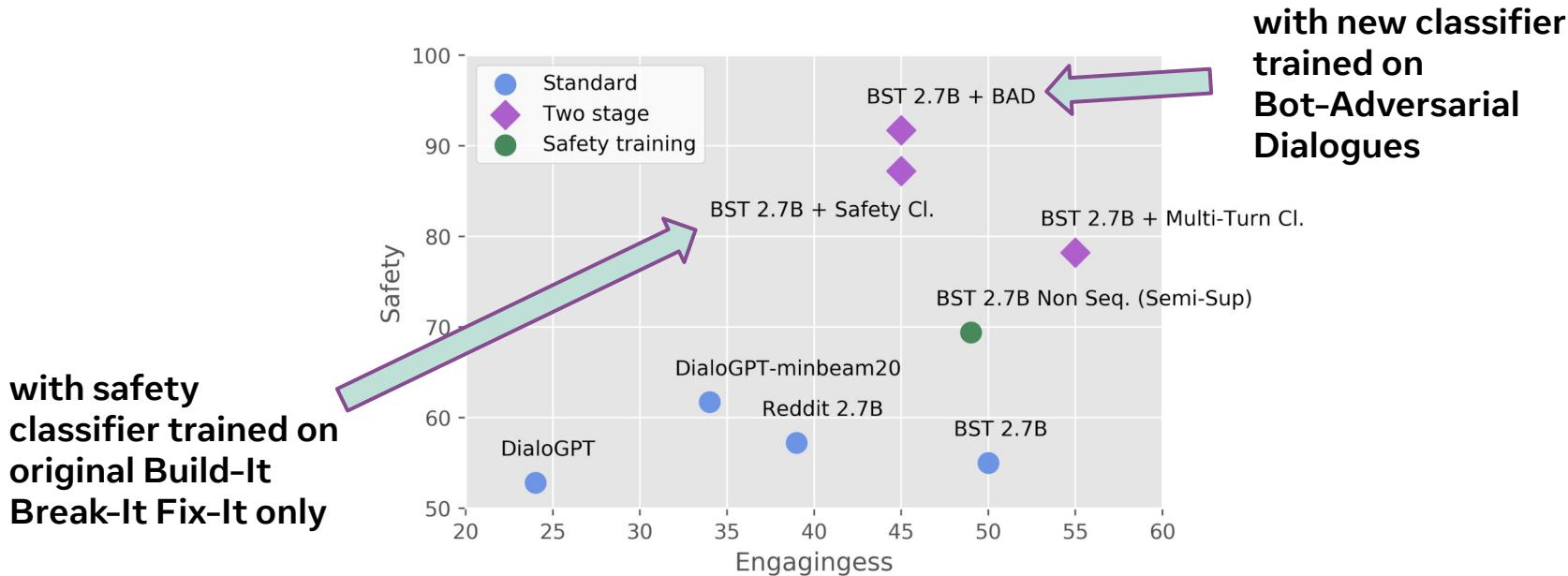
Two-stage models are the **most** safe ... but still far from perfect.

Bot Adversarial Dialogue: Results



Safer chatbots are **still engaging!**

Bot Adversarial Dialogue: Results



Models trained with the new Bot-Adversarial Dialogue data are best, because the data is **closer to the domain in which these models are deployed** ... i.e., chatting with (potentially adversarial) humans!

Bot Adversarial Dialogue: Results

for details check the paper!

Recipes for Safety in Open-domain Chatbots

Jing Xu Da Ju Margaret Li Y-Lan Boureau Jason Weston Emily Dinan
Facebook AI Research

Other possible applications

1. Natural Language Understanding (**Z. Yang et. al 2017**)
2. Natural Language Inference (**Y. Nie et. al 2019**)
3. Question answering, sentiment analysis, etc....

MASTERING THE DUNGEON: GROUNDED LANGUAGE
LEARNING BY MECHANICAL TURKER DESCENT

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H. Miller
Arthur Szlam, Douwe Kiela & Jason Weston
Facebook AI Research



The Dynabench logo is located in the top left corner of the slide. It consists of a blue header bar with a white three-line menu icon on the left. To its right is the Dynabench logo, which features a stylized icon made of dots and squares followed by the word "Dyna" in a sans-serif font and "Bench" in a smaller, regular font below it.

Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

dynabench.org

Overall Round 1 Round 2

Hate Speech

Hate speech detection is classifying one or more sentences by whether or not they are hateful.

Current round:	2
Fooled/Collected (Model Error rate)	437/1250 (34.96%)
Last activity:	a few seconds ago

What is **data excellence**?

Data excellence:

Data excellence:

- Lack of artifacts/dataset bias that ML models can easily overfit to

Data excellence:

- Lack of artifacts/dataset bias that ML models can easily overfit to
- **Continually evolving (non-obsolescence)**

Data excellence:

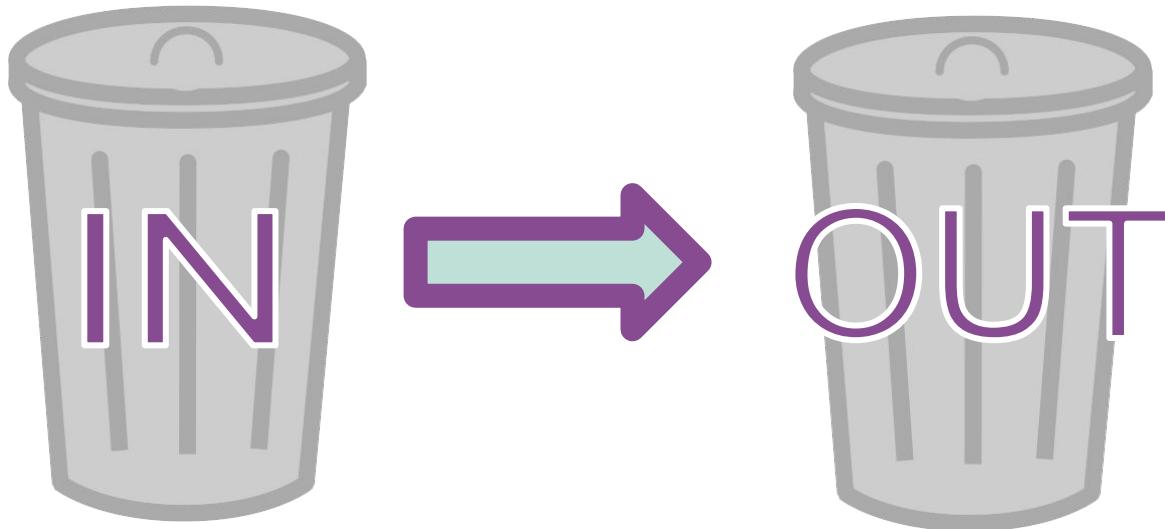
- Lack of artifacts/dataset bias that ML models can easily overfit to
- Continually evolving (non-obsolescence)
- **As close as possible to measuring our north star objective**

Data excellence:

- Lack of artifacts/dataset bias that ML models can easily overfit to
- Continually evolving (non-obsolescence)
- **As close as possible to measuring our north star objective... in the case of chatbots, interaction with humans!**

Why should you care about data quality?

Why should you care about data quality?



Why should you care about data quality?

“more robust models, faster progress, a more well-defined objective, and a better understanding of our current limitations”
(dynabench.org)

Thank you

INVITED TALK

20 mins

What Do Our Models Learn?

Aleksander Mądry

MIT

What Do Our Models Learn?

Aleksander Mądry



Based on joint works with:



Logan Engstrom



Andrew Ilyas



Shibani Santurkar



Dimitris Tsipras



Kai Xiao

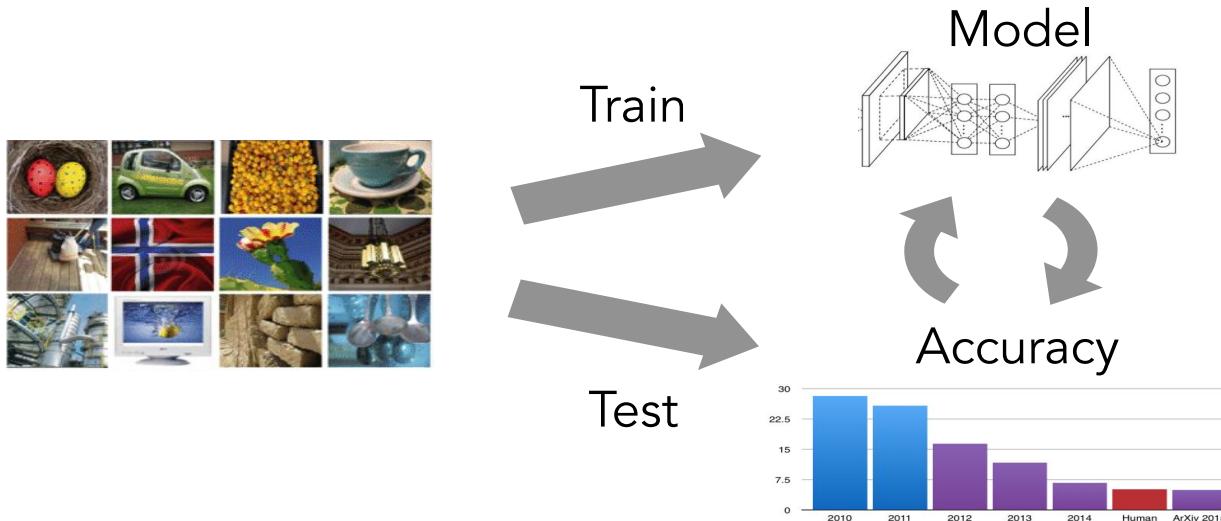


@aleks_madry



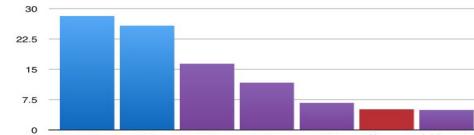
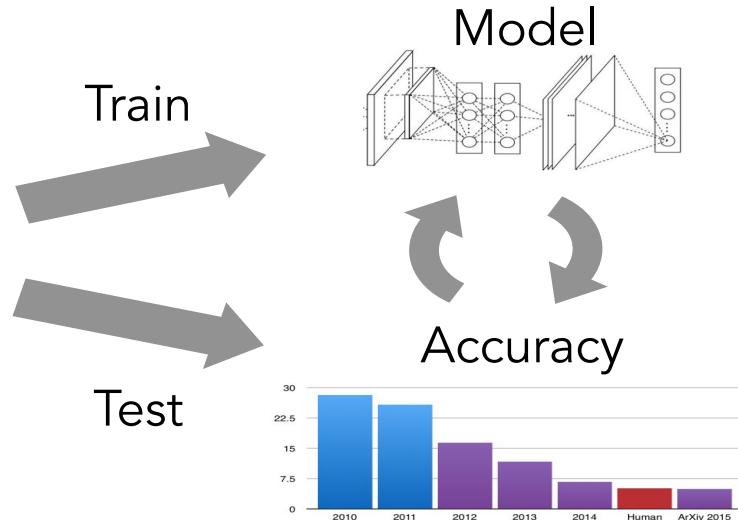
gradientscience.org

ML Research Pipeline

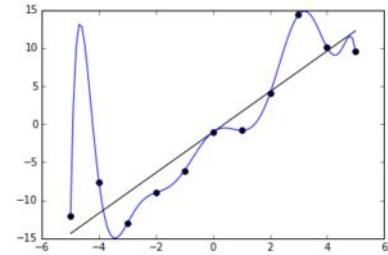
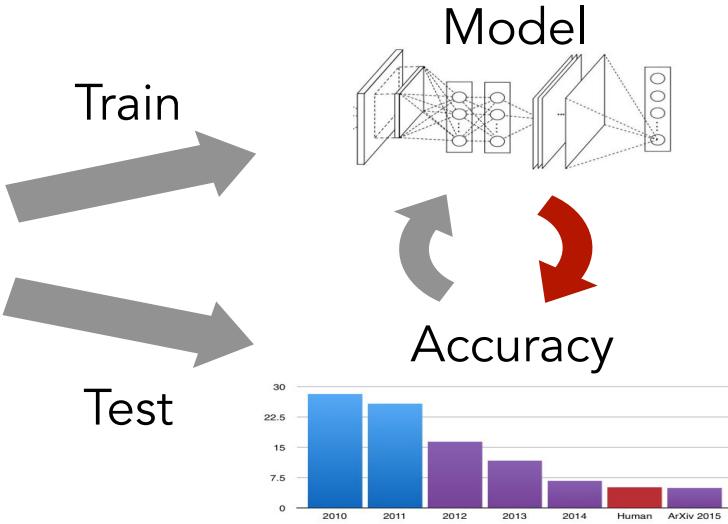


What are we usually concerned about here?

Concern #1: “Classic” Overfitting

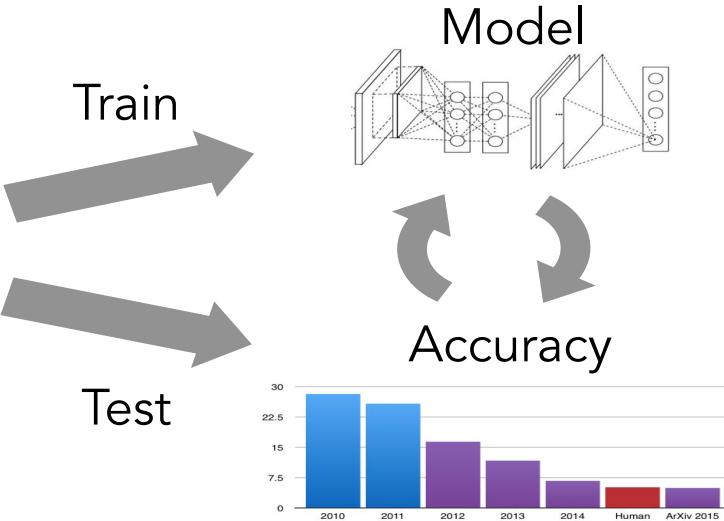


Concern #1: “Classic” Overfitting

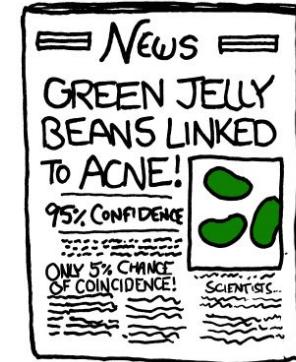
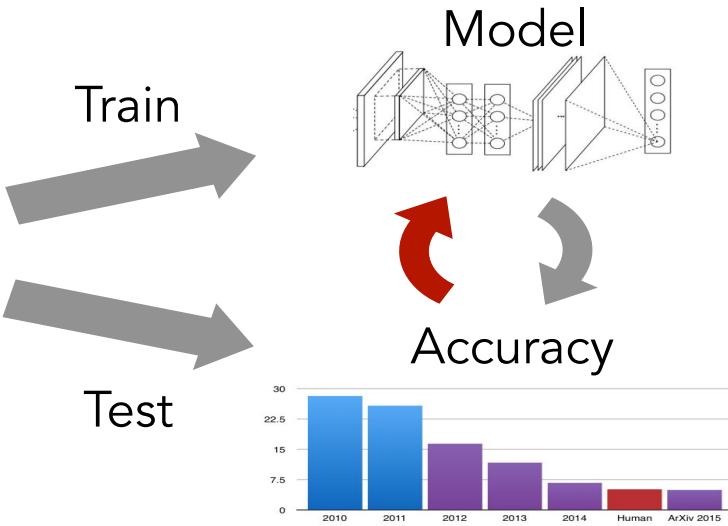


Aren't we memorizing the training data?
[Vapnik Chervonenkis 1971]

Concern #2: Adaptive Overfitting



Concern #2: Adaptive Overfitting

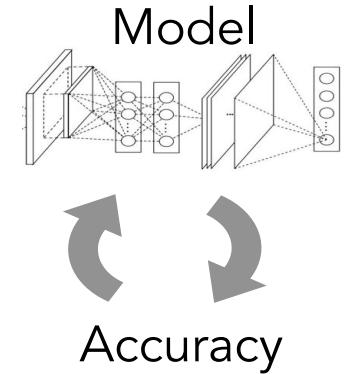
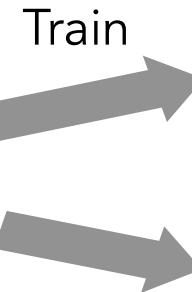
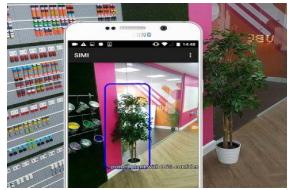


Aren't we “p-hacking” the test data?

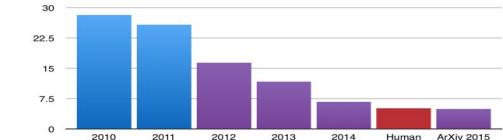
[Dwork Feldman Hardt Pitassi Reingold Roth 2015]

But: That's not all

Real-world task:
Object recognition



→ Our dataset/benchmark is meant
to be a proxy for the real-world task

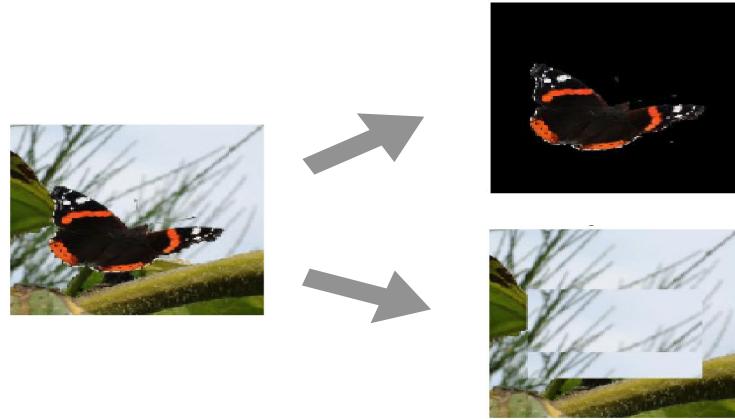


Aren't we overfitting to the whole benchmark too?
[Ponce et al. 2006] [Torralba Efros 2011]

How well our datasets reflect the real world?

- What dataset biases do models pick up?
- How are these biases introduced?

Simple Setting: Background bias



Models do leverage background signal

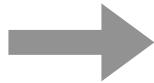
[Zhang et al. 2007] [Zhu et al. 2017] [Ribeiro et al. 2016] [Rosenfeld et al. 2018] [Beery et al. 2018] [Shetty et al. 2019] [Sagawa et al. 2018]

ImageNet-9: A Fine-Grained Study

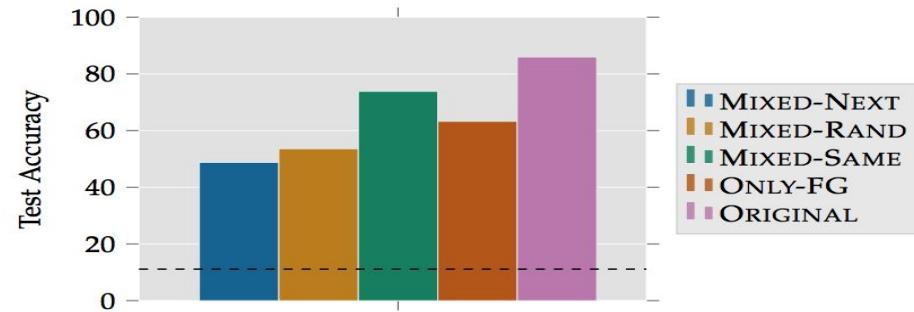
[Xiao Engstrom Ilyas M 2020]



Replace
background

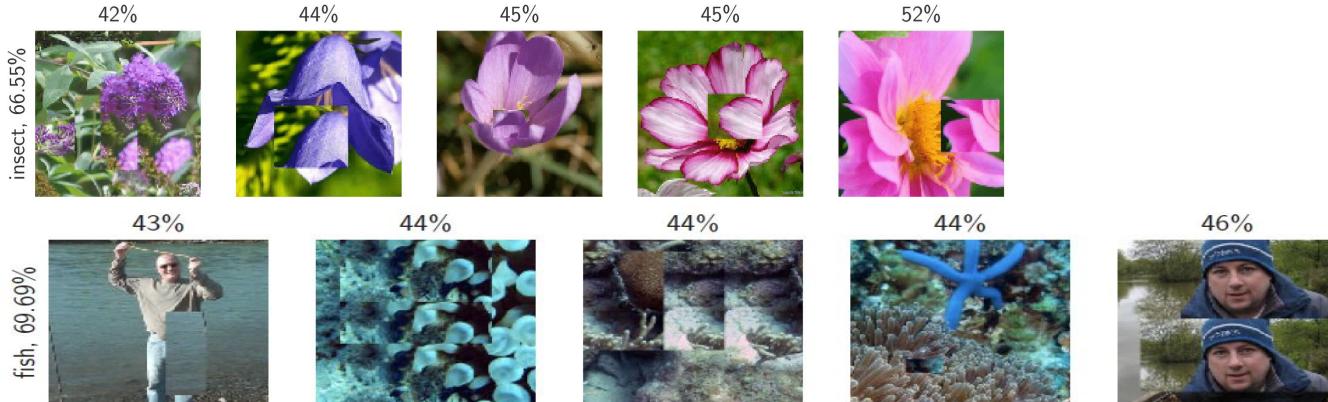


In general: Model accuracy
is significantly reduced



Adversarial Backgrounds

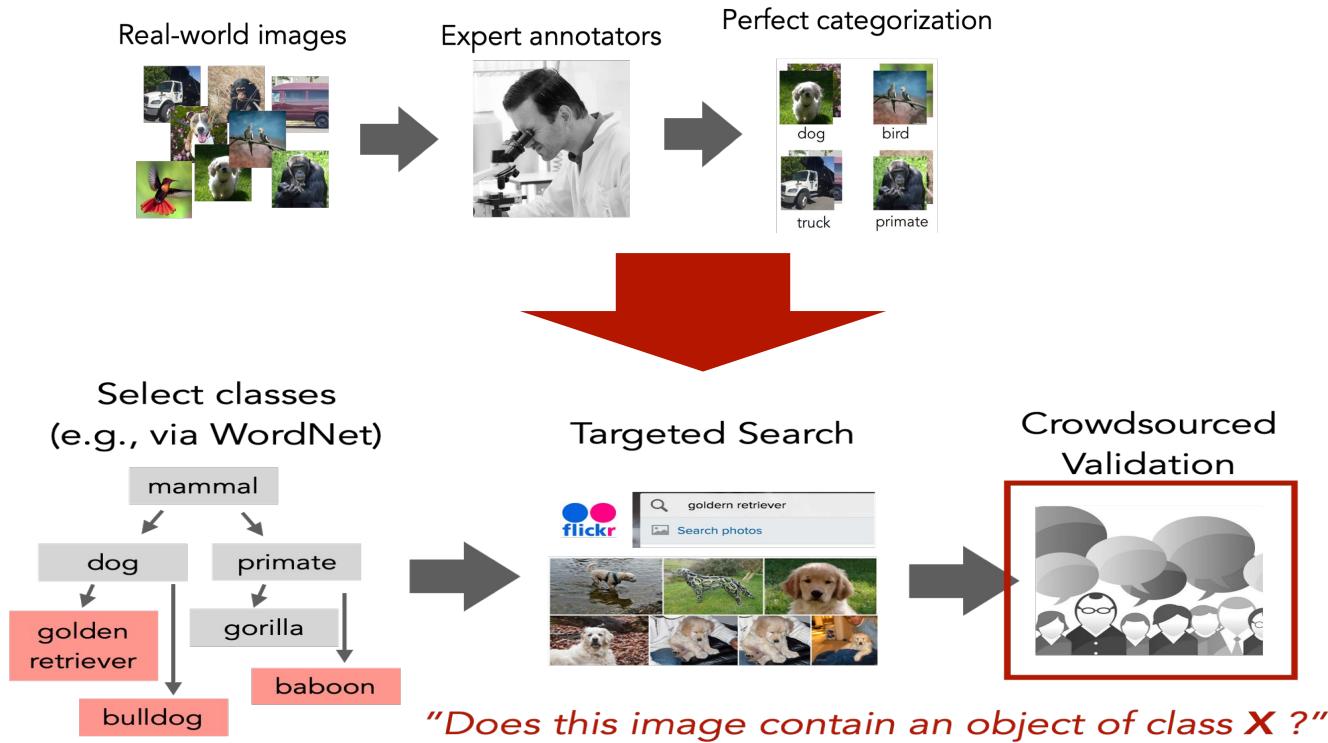
Can fool a model on most (87.5%) inputs by
using “worst-case” background



Where do such biases come from?
(beyond the visual world)

Dataset Creation Introduces Biases

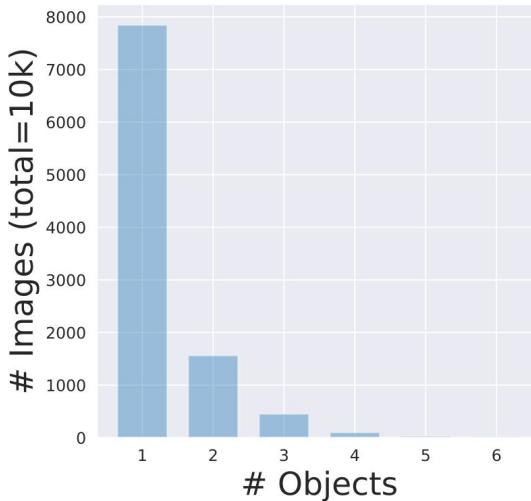
[Tsipras Santurkar Engstrom Ilyas M 2020]



Dataset Creation Introduces Biases

[Tsipras Santurkar Engstrom Ilyas **M** 2020]

Result: More than 20% of (single-label) images contain more than one object



Label:

- "microwave"

Also valid:

- "refrigerator"
- "stove"
- "washbasin"

Dataset Creation Introduces Biases

[Tsipras Santurkar Engstrom Ilyas **M** 2020]

Furthermore: “Correct” label is often not the main object
(according to humans)



ImageNet: “pickelhaube”
Annotators: “military uniform”



ImageNet: “bowtie”
Annotators: “suit”



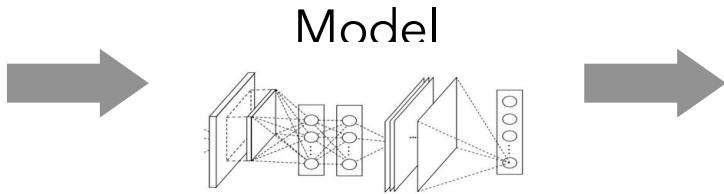
ImageNet: “bell cot”
Annotators: “church”

Consequence: Models incentivized to pick up on
these dataset (idiosyncratic) biases

How good ImageNet models really are?

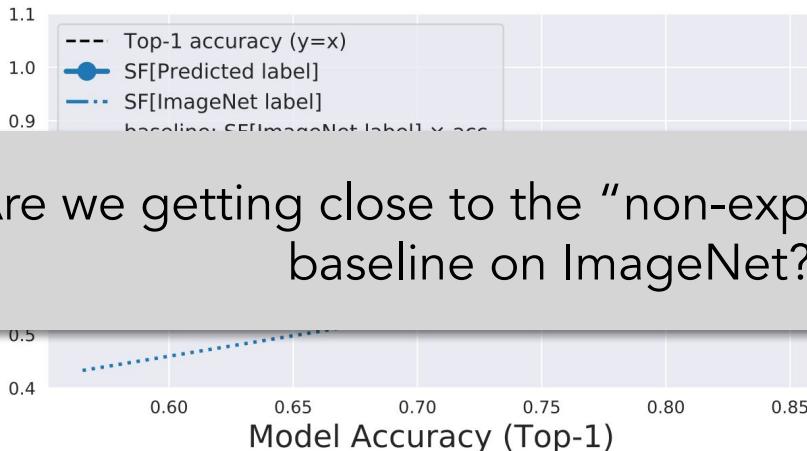
(Once we account for issues with labelling)

Human-Based Evaluation



Human-Based Evaluation

Good news: Model predictions are consistently improving



Are we getting close to the “non-expert annotator” baseline on ImageNet?

Also: Annotators often can't tell labels from predictions apart

Takeaways

What models do and do not learn is not always clear

(And we might want to study this before focusing on improving performance on a given benchmark—let alone robustness)

Models are affected by biases of the world and our data pipelines

But: We can improve the situation by accounting for them

[Engstrom Ilyas Santurkar Tsipras Steinhardt M 2020]

Moving forward

What other biases do models learn from the data?

How do we train our models in their presence?

How do we measure performance on the underlying task
we care about (as opposed to the benchmark itself)?



@aleks_madry

To read more:



gradientscience.org

SHORT BREAK

We will resume at

8:55a 11:55a 4:55p

PDT

SF

EDT

NYC

CET

AMS

with more talks &
discussion

JOIN VIDEO STREAM AT:

underline.io/events/27/main-stage

JOIN SLACK CHAT AT:

bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:

- [#ask-questions-for-speakers-here](#)
- [#ask-for-help-here](#)

WORKSHOP PROGRAM:

eval.how/dew2020/

CODE OF CONDUCT:

www.humancomputation.com/attend.html#code

DEW2020

1ST Data Excellence Workshop

SESSION 3

**Challenges and Lessons in
Adjudicating Labeling
Medical Data, Quang Duong**

**Getting good data faster,
Peter Hallinan**

**Session 2 & 3 Talks
Discussion**

Wrap Up

INVITED TALK

20 mins

Challenges and Lessons in Adjudicating Medical Data Labels

Quang Duong
Google Health

Agenda

1. What is medical labeling (at scale)?
2. Why is it hard? How does collaborative labeling (in particular, asynchronous adjudication) address the challenge of data quality?
3. How may we speed up asynchronous adjudication?
4. Could we replace asynchronous adjudication with something of lower cost?

“

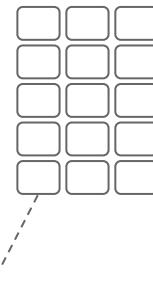
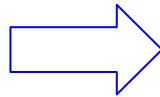
Medical data and medicine are complex.
Reproducibility in medicine is both
challenging and critical.



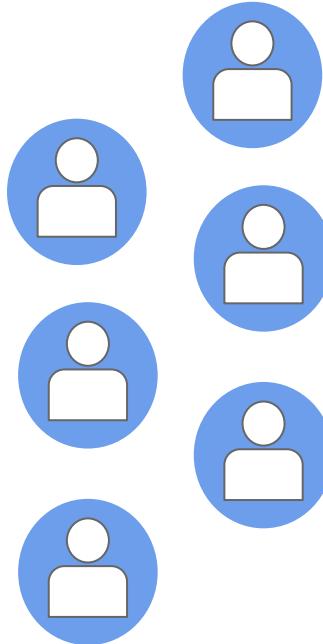
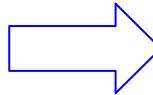
Typical crowdsourcing system for data labeling



Requesters



Is there a cat in this picture?

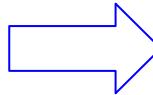
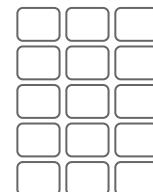
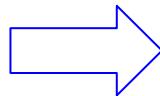


Workers

Crowdsourcing medical labeling (diagnoses)

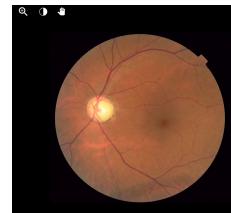


Requesters



Workers

Does this patient have Diabetic Retinopathy?

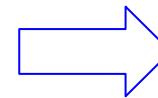
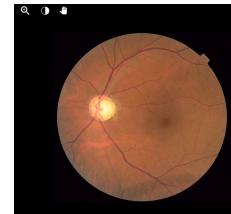
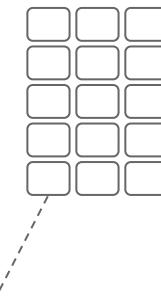
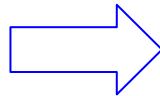


Ranges secs ~ hours

Crowdsourcing medical labeling (diagnoses)



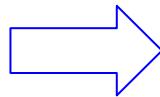
Requesters



Workers (Graders)

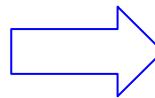
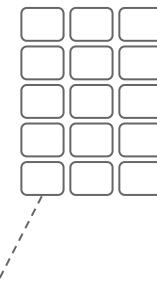
Grade/label medical data.
Medical experts

Crowdsourcing medical labeling (diagnoses)



Managers

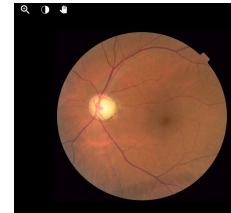
Manage a project,
coordinate workers/graders



Workers (Graders)

Investigators

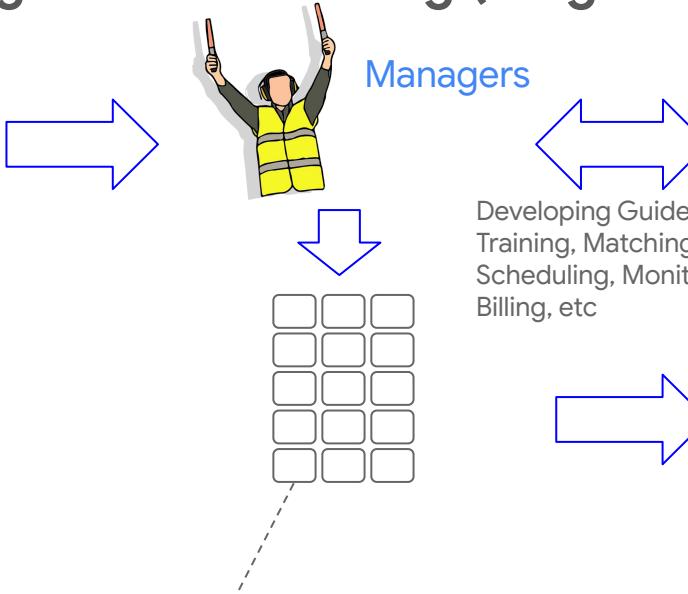
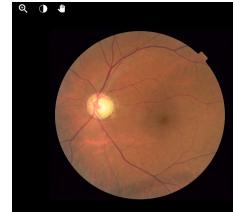
Build ML Models or conduct
medical research



Crowdsourcing medical labeling (diagnoses)



Investigators



Workers (Graders)

Agenda

1. What is medical labeling (at scale)?
2. Why is it hard? How does collaborative labeling (in particular, asynchronous adjudication) address the challenge of data quality?
3. How may we speed up asynchronous adjudication?
4. Could we replace asynchronous adjudication with something of lower cost?

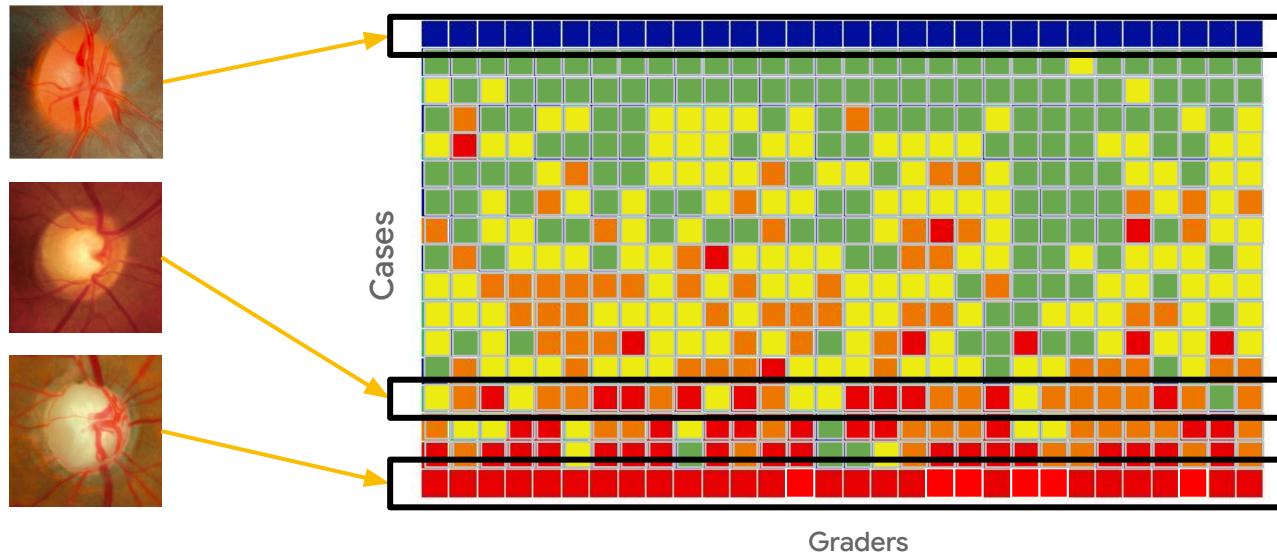


Labeling medical data is hard

1. Large volume, expensive **medical-expert** labor
2. ML solutions that **affect patients** must have exceptionally high quality
3. Strict **privacy/security** and **design control** requirements

and much more...

“Ask 3 doctors, get 4 opinions”



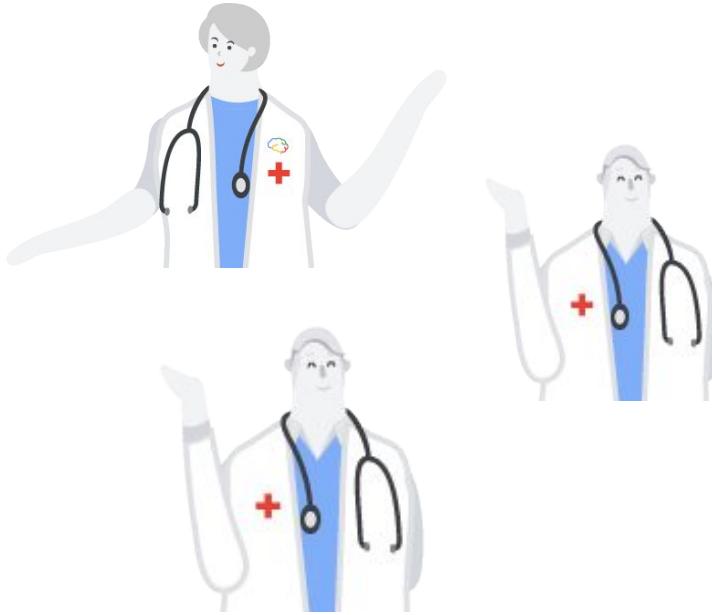
Non-glaucomatous

Low risk suspect

High risk suspect

Glaucoma

“Ask 3 doctors, get 4 opinions”



Standard clinical practice is **in-person adjudication** by clinicians. This is slow, costly, and difficult to scale.

/ Asynchronous Adjudication: Allow remote and anonymous evaluation and discussion (in our internal platform)

The screenshot shows a digital fundus photograph of an eye on the left, with two yellow arrows pointing from the text "Displayed answers from all graders" and "Discussion" to specific features on the right.

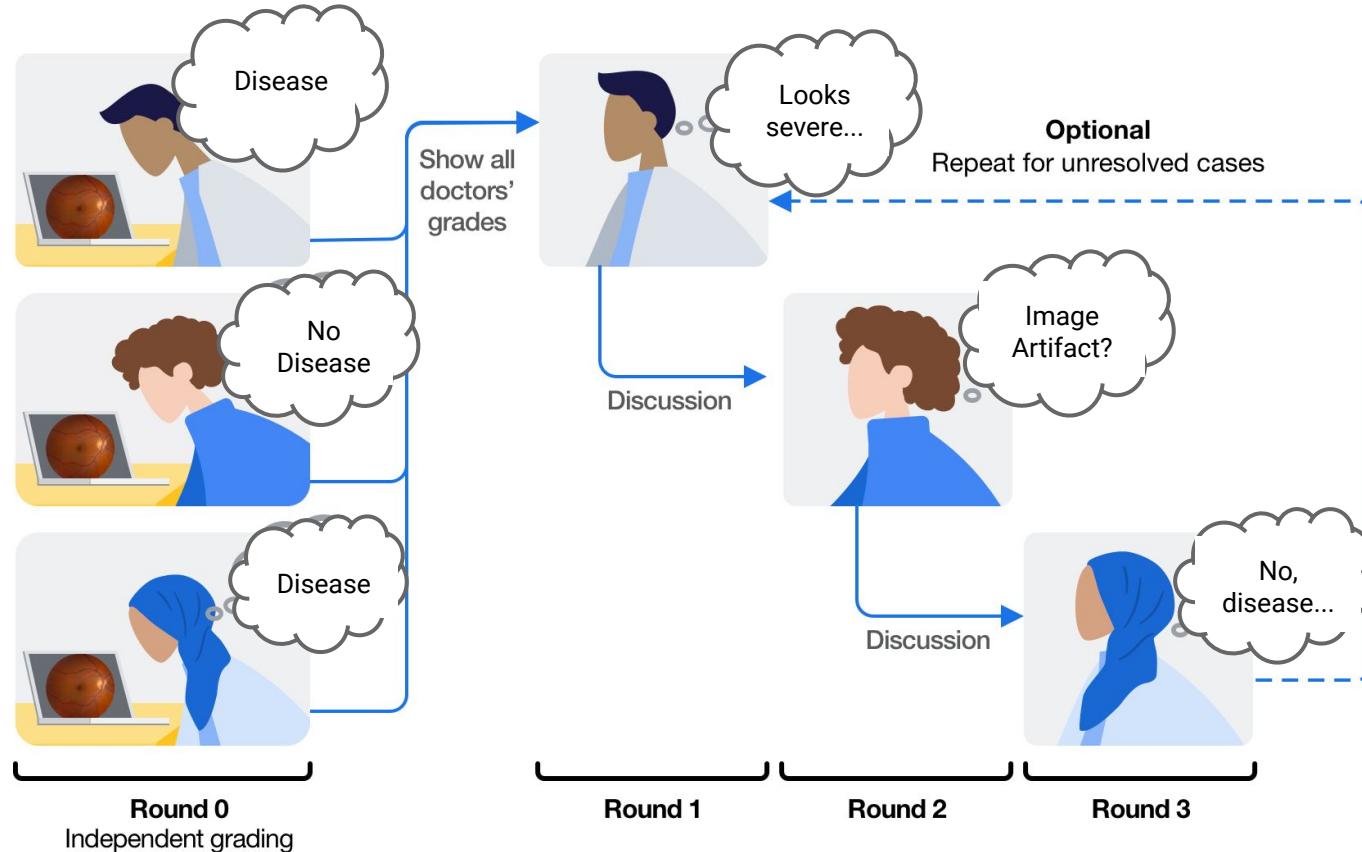
Displayed answers from all graders: Points to a row of colored squares at the top of the interface, indicating the status of evaluations (green for completed, yellow for pending).

Discussion: Points to a large circular area of the fundus photograph, representing the platform for discussion and collaboration.

Right-hand interface details:

- Indicate agreement/disagreement:** A title at the top of the main panel.
- Comment History:** A sidebar showing a conversation between W2 and W1. W2 expresses regret about a previous statement and provides a rationale for their current position.
- Adjudication rounds:** A section showing a comment from Round 9 by W2.
- Panel 1 of 1:** A header for the main content area.
- Hello world:** The main content area displays a sample question: "This is a multiselect prompt. Select all that applies." It includes three options: "option 0", "option 1 - Some super long text that makes this line wrap again and again.", and "option 2".
- Feedback:** A section at the bottom for "Feedback on usability and guidelines if any."

/ Asynchronous Adjudication: Round robin until agreement



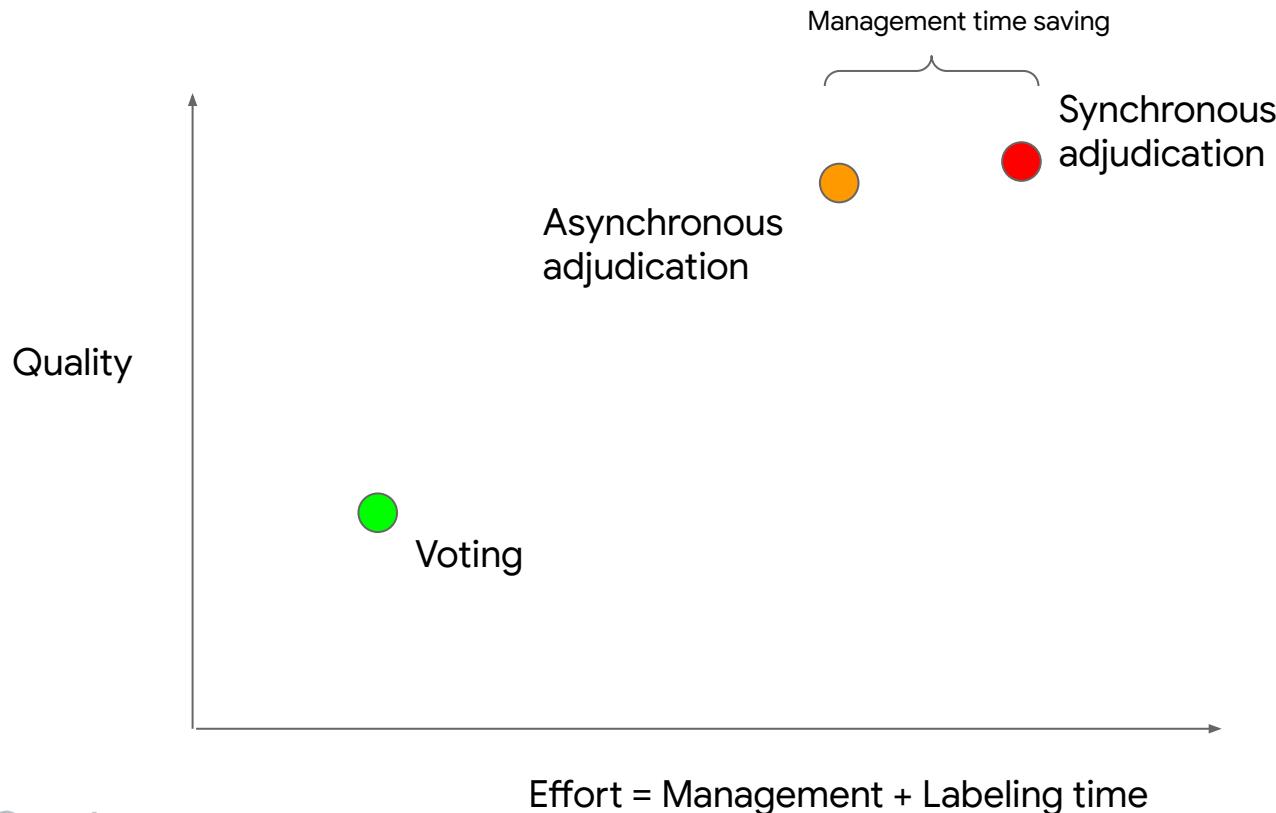
/ A success for scaling labeling

Async
Adjudication is

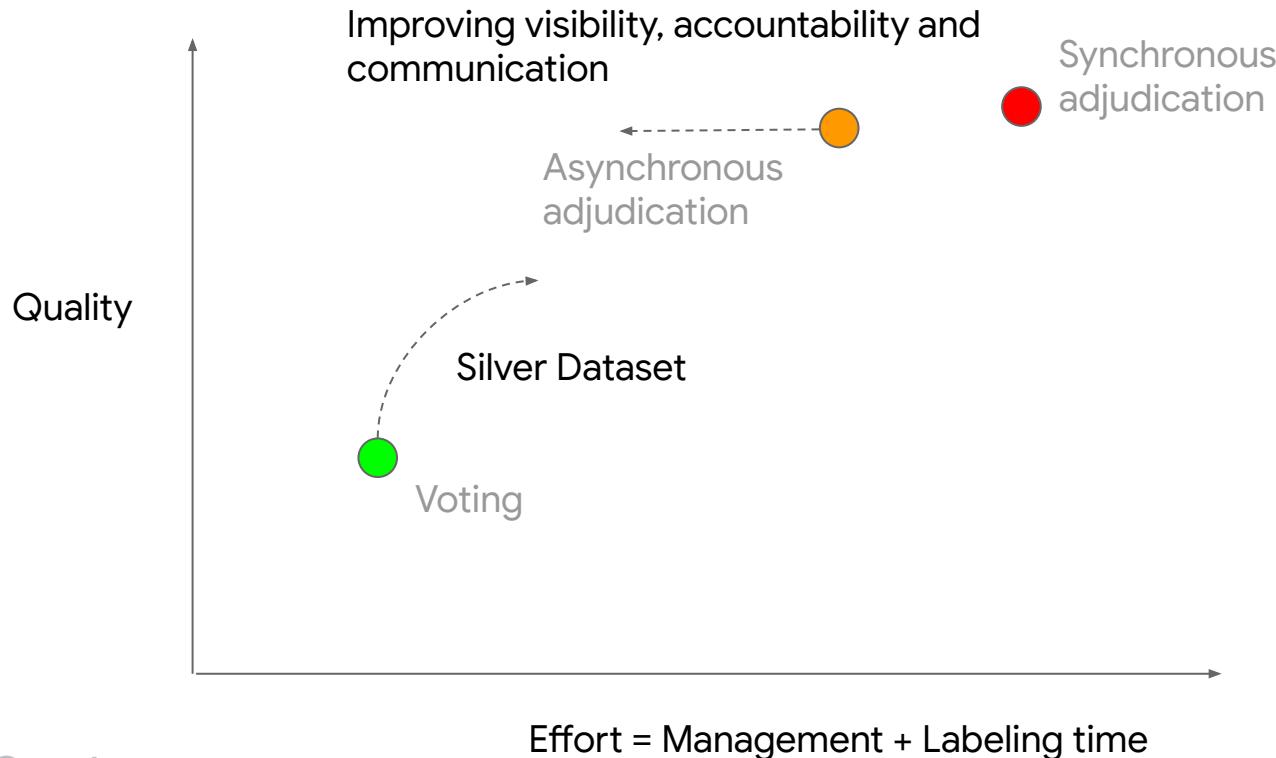
Remote
Convenient
Scalable*

(*Est. 99% **management** time savings, and
roughly **the same labeling time** compared to
sync adjudication)

/ Collaborative Labeling Overview



/ Collaborative Labeling Overview



Agenda

1. What is medical labeling (at scale)?
2. Why is it hard? How does collaborative labeling (in particular, asynchronous adjudication) address the challenge of data quality?
3. How may we speed up asynchronous adjudication?
4. Could we replace asynchronous adjudication with something of lower cost?

“

Visibility, accountability and communication are key to increasing **team efficiency**



Freeman et al, 2020

/ Challenge: “Progress held hostage”

People don't know when it'll be their turn:

“The biggest challenge is **bottlenecking**. ‘Am I waiting on someone? How long should I be waiting for?’

(If they don’t label right away), it has a downstream impact on other labelers [in the same adjudication group]. It **holds the entire progress hostage in their queue.**”

Radiology labeler manager



Challenge: “Progress held hostage”

Delays in round 0 make it harder for labelers to remember cases:

“Once you’ve gone through the cases, (it’s best to) have it **not be too long before you get them back to look at again** for the adjudication process.

Then it’s a little **fresher in your mind**. The faster you can get it back, the better.”

Radiologist

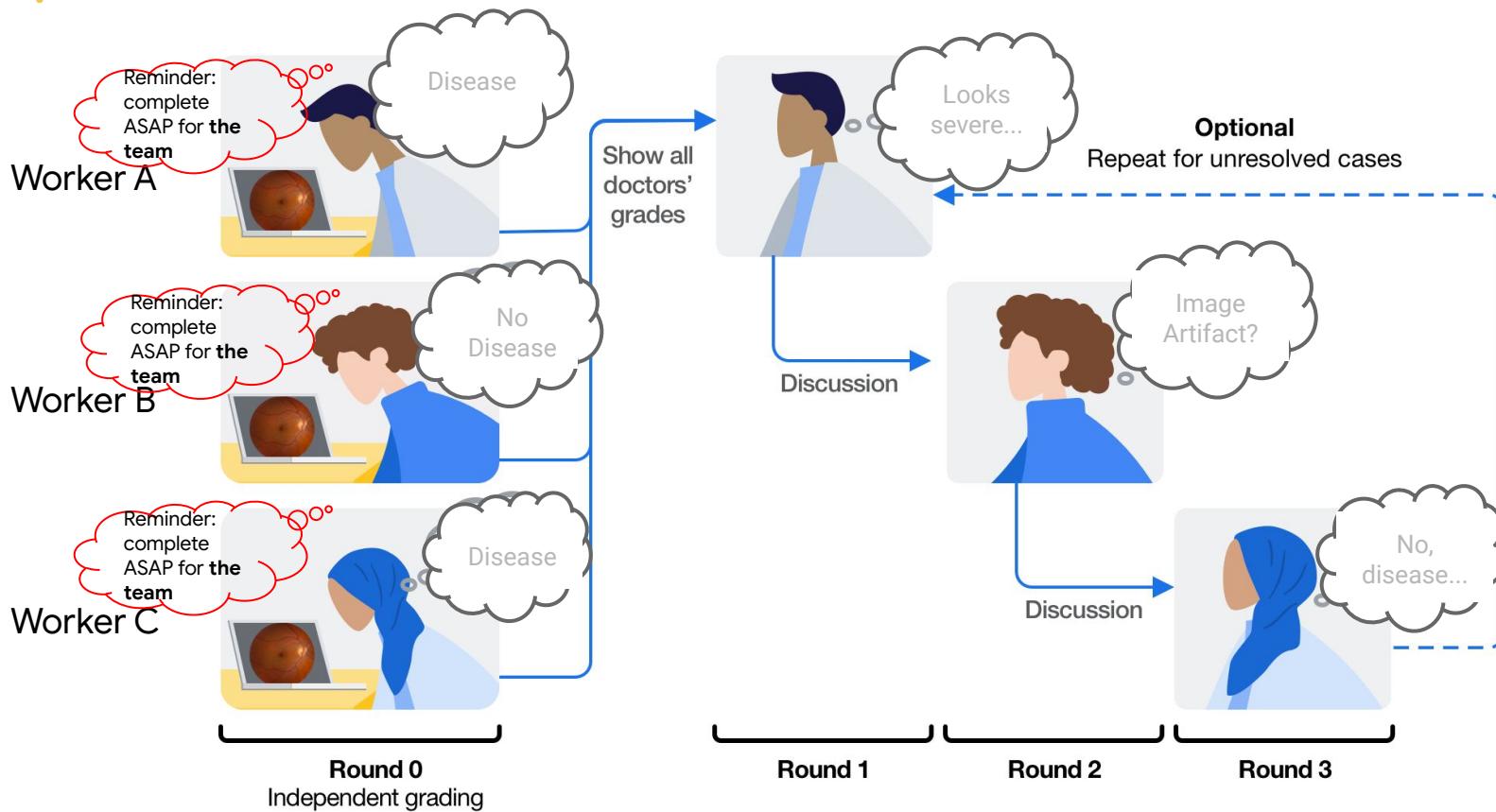
/ Challenge: “Progress held hostage”

Delays may influence other labelers to stop checking:

“If Grader A takes a long time to respond, Grader B may say, ‘It’s been 2 weeks, it’s probably done.’ Grader C probably thinks the same thing, and **not check their queue.**”

Ophthalmology labeler manager

Increase accountability



Increase visibility

Worker A

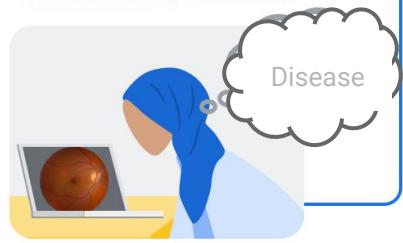


Show all doctors' grades

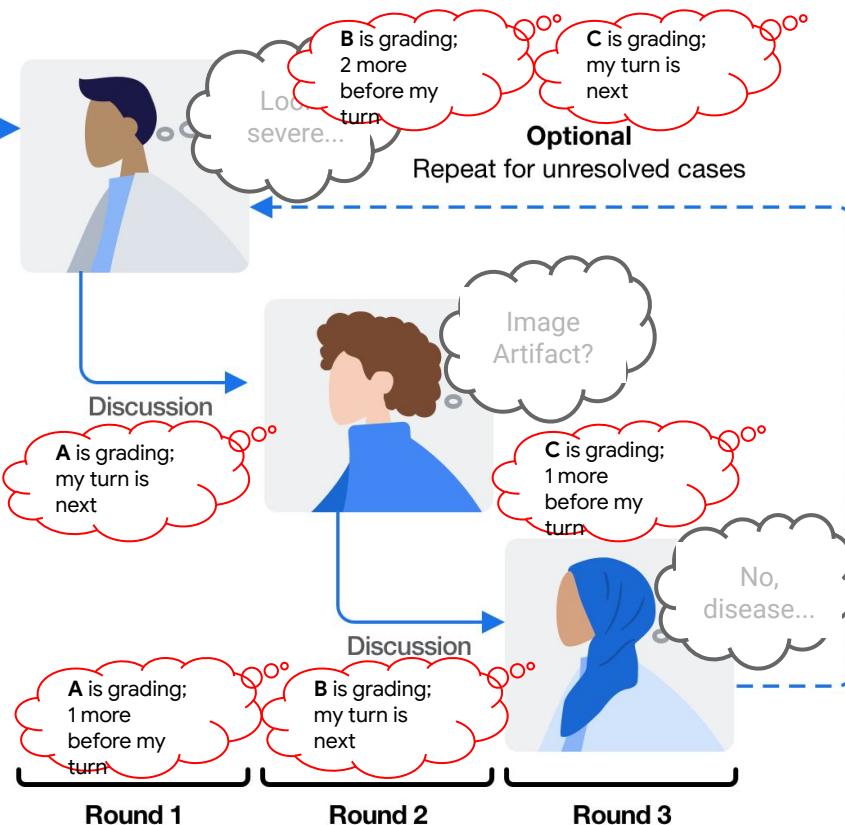
Worker B



Worker C



Round 0
Independent grading



/ Challenge: Async communication is inherently limited

Case-specific [lower level] comment box doesn't facilitate project-level [higher level] discussion

“Interfacing with someone else through a computer and not being able to discuss things.

I noticed the (annotation size) issue on the prior adjudication case, and I didn't bring it up. [I would have asked] ‘What's your approach? Are you circling diagnostic cases in addition to screens?’ I'm not going to mention it on a specific case;

Radiologist

/ Challenge: Async communication is inherently limited

Explaining yourself in writing takes longer, lacks immediate feedback:

“You know **someone else is going to be looking at the work**. It’s more time consuming because in the back of my head I’m thinking, ‘**Is this clear to someone else why I did what I did?**’

When you’re doing a solo job, you don’t really have to explain yourself. I don’t feel the pressure to put as much down when it’s a solo project.”

Radiologist

Mix between async and sync adjudication

The same way most people prefer mixing between in-person and remote work

1. **Before** async adjudication at scale: **synchronous** adjudication of a few example cases to define grading guidelines
2. **During** async adjudication: regular check-in meetings (weekly) for all workers to discuss **synchronously** overall approaches
3. **After** async adjudication: synchronous adjudication on very difficult non-convergent cases

Establish new communication channels

Worker A



Show all doctors' grades

Worker B



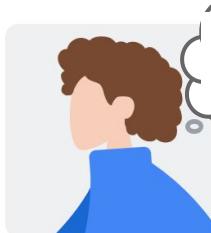
Discussion

Worker C



Round 0
Independent grading

Optional
Repeat for unresolved cases



HIPAA compliant, secure
free-form
communication channel

Round 1 Round 2 Round 3

Agenda

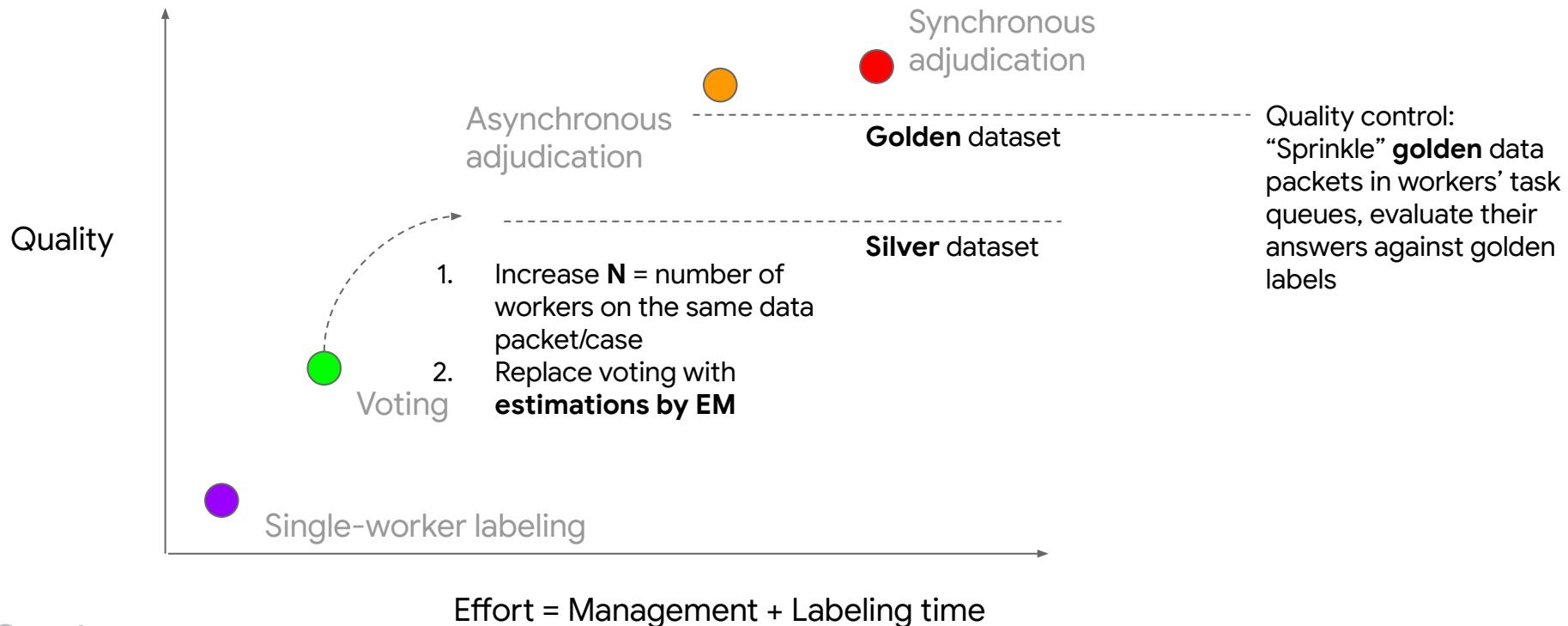
1. What is medical labeling (at scale)?
2. Why is it hard? How does collaborative labeling (in particular, asynchronous adjudication) address the challenge of data quality?
3. How may we speed up asynchronous adjudication?
4. Could we replace asynchronous adjudication with something of lower cost?

“

Using ground truth estimates created by
Expectation Maximization can achieve
comparable quality as async adjudication, and
increase overall throughput



Silver data is less expensive and can potentially act as ground truth proxy in quality control



/ Objective: estimate workers' quality without access to ground truth (golden data)

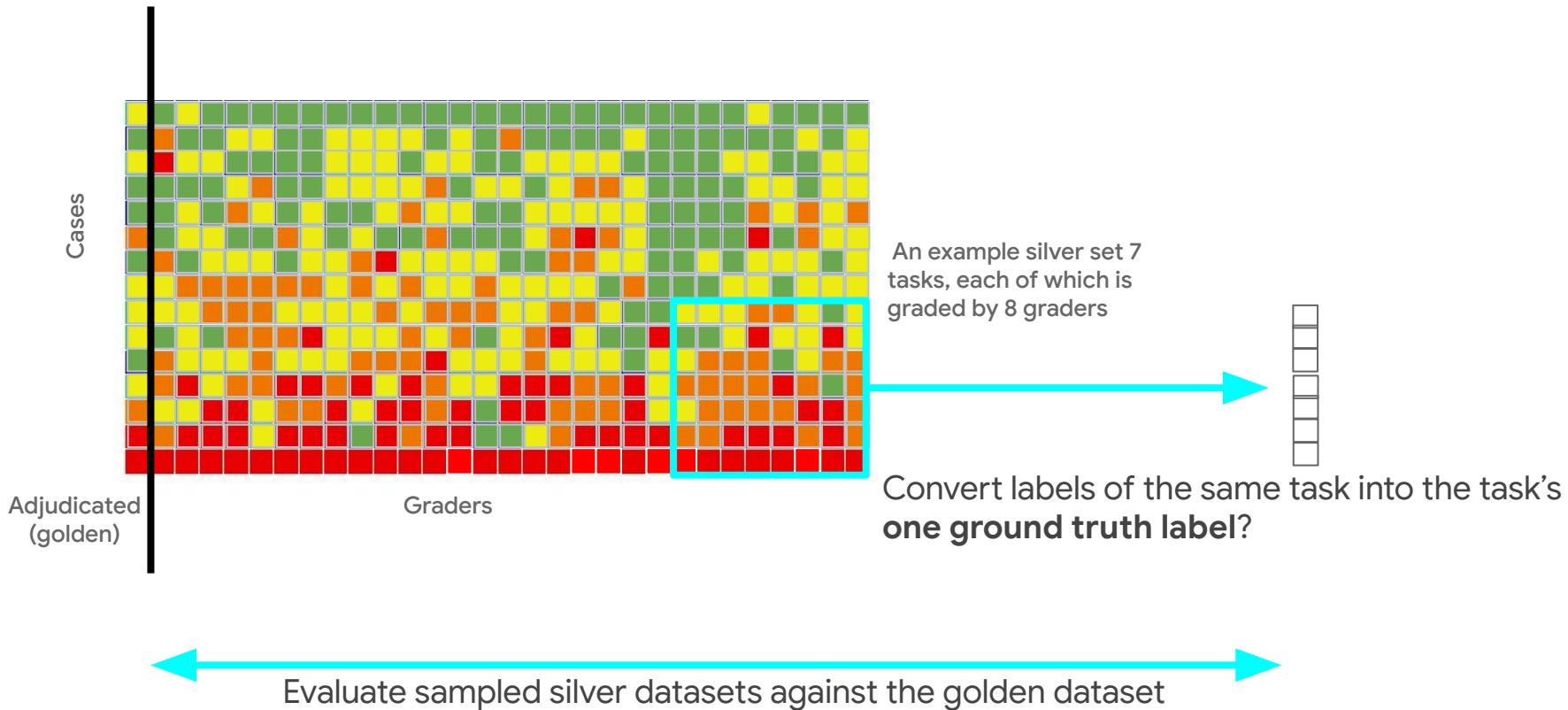
Study of an eye disease, where workers diagnosed using a **four-point scale**

Golden dataset: ~1200 images adjudicated by 3 experts. We also have labels independently generated by 26 workers.

Goal: estimate 26 workers' quality using different silver datasets, each of which is defined by:

- Number of labeling tasks (images): n_{tasks}
- Number of workers labeling each task (image): $n_{\text{annotators}}$

Sampling synthetic silver datasets



/ Estimating both ground truth labels and worker accuracy

Confusion matrix per worker \mathbf{k} = probability of assigning label i to an item of true label j .

Expectation Maximization (EM) [Ipeirotis et al. '10]

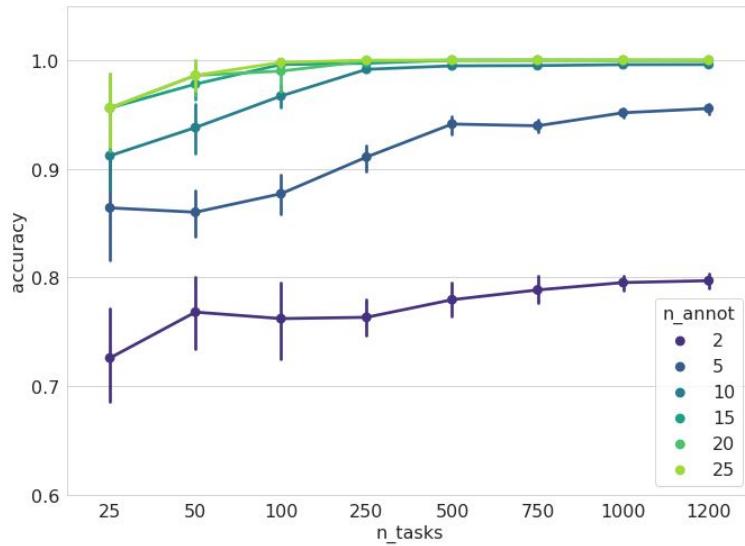
Estimate “true” labels

to maximize the likelihood (expectation) of observing workers’ labels

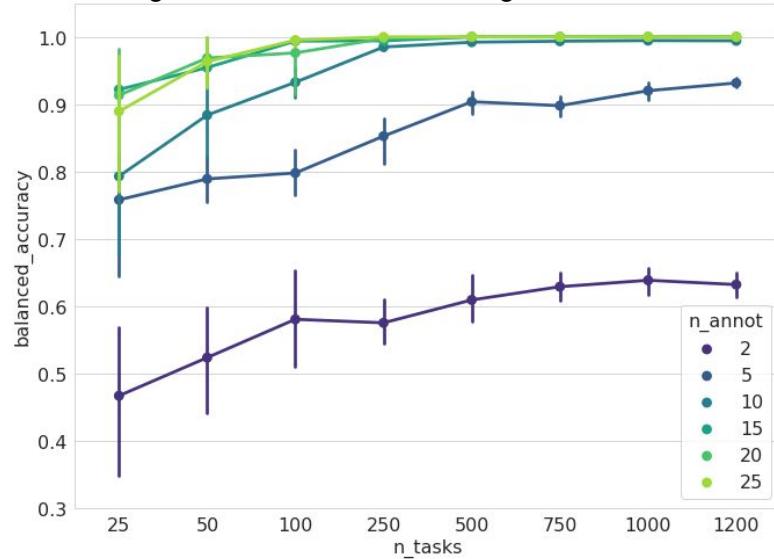
given their resulting confusion matrices.

EM-inferred labels closely resemble the adjudicated labels

Fraction of correctly inferred image labels

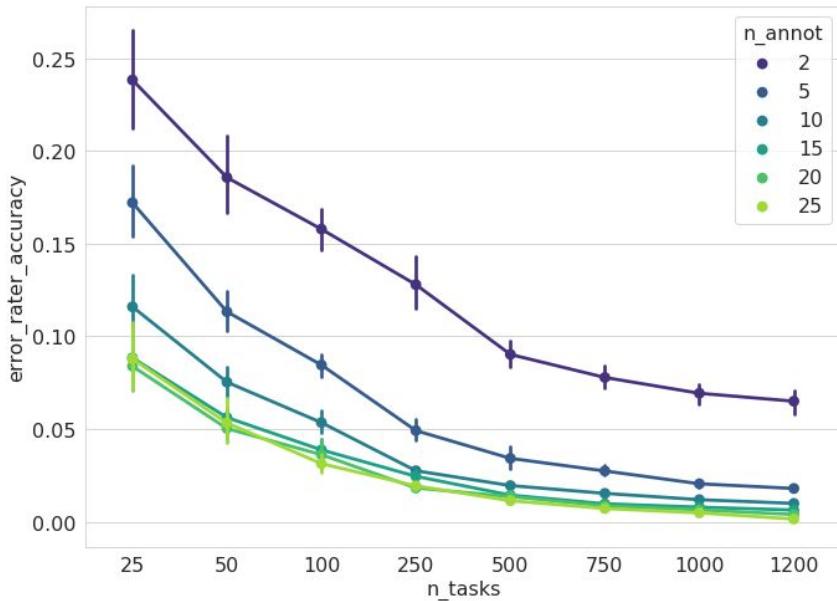


Average of recall for each class (good for imbalance data)



Notes: this labeling job is **one of the most challenging**

/ EM-inferred worker quality is similar to true worker quality



The accuracy of a worker = the % correct labels evaluated against a given label dataset.

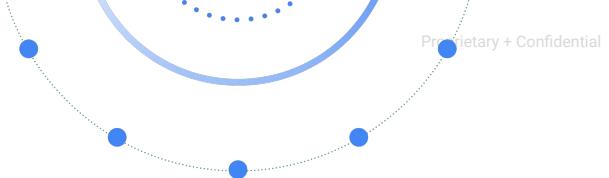
Report the average (over all 26 workers) absolute difference between the **true** worker accuracy computed on the **golden dataset** and the **estimated** worker accuracy computed on the **silver dataset**.

/ An EM-generated silver dataset should have at least 250 data packets/tasks, each of which was labeled by 10 workers

Total time for labeling 10 labels per task ~ 7 adjudication rounds + initial round 0 of 3 workers, **not taking into account:**

- comments take longer to write
- workers wait for others to label

Overall, *no obvious saving* on active labeling time. Main benefits come from eliminating wait time between rounds (of adjudication).



Conclusions

Medicine and medical data are complex

Medical-grade, compliant, secure data examination software

Remote-working experts (labeling workers), frequent disagreement

Async adjudication improvements

Visibility: assists planning for partial workers

Accountability: promotes sharing team responsibilities

Communication: mixes async and sync

EM-inferred silver dataset may be an alternative

EM algorithm to replace majority/simple voting

Elimination of wait time

Thank you!

qduong@google.com



INVITED TALK

20 mins

Getting good data faster

Peter Hallinan

AI & Data @Amazon

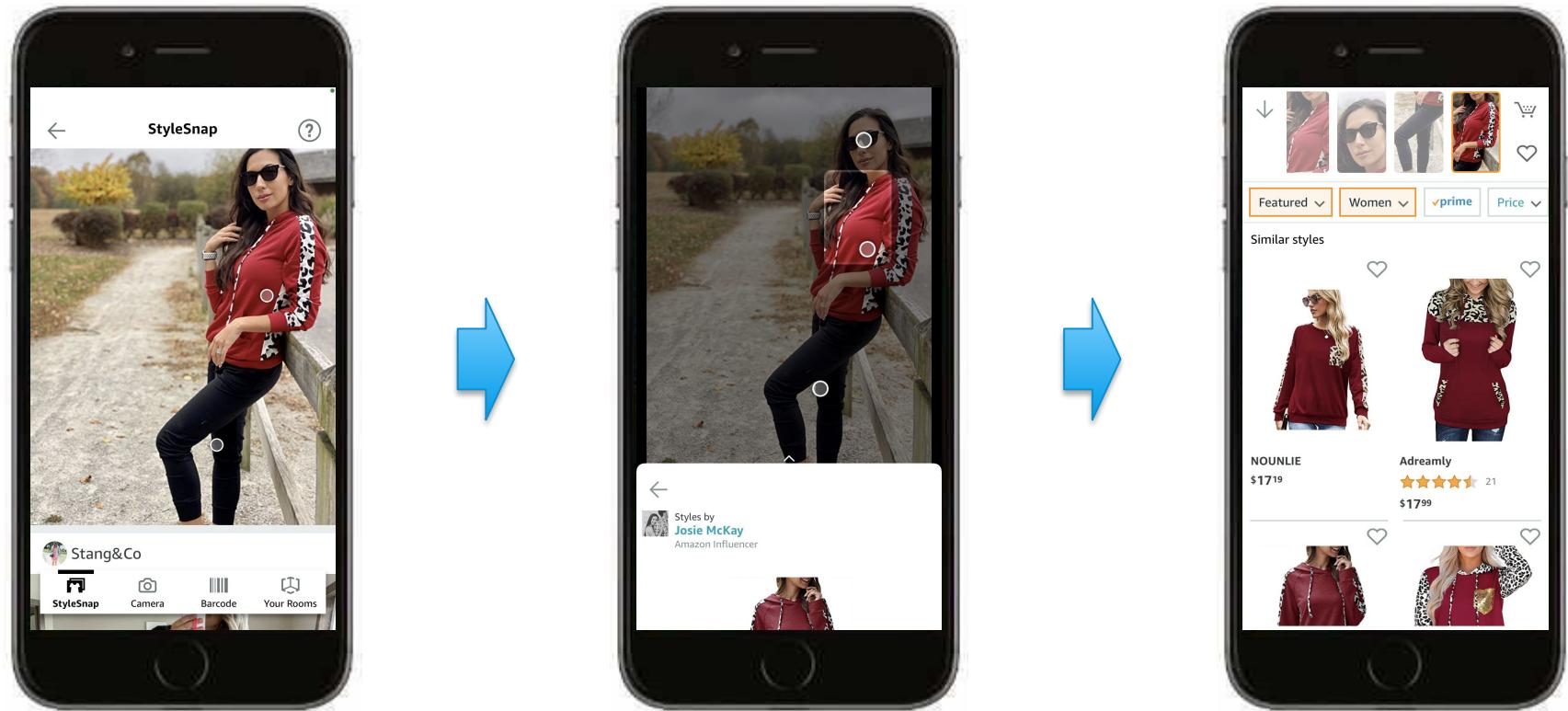
Agenda

Reminder: HLD = human-labeled data

- Core concepts
- Strategic & operational challenges
- Organizational solutions

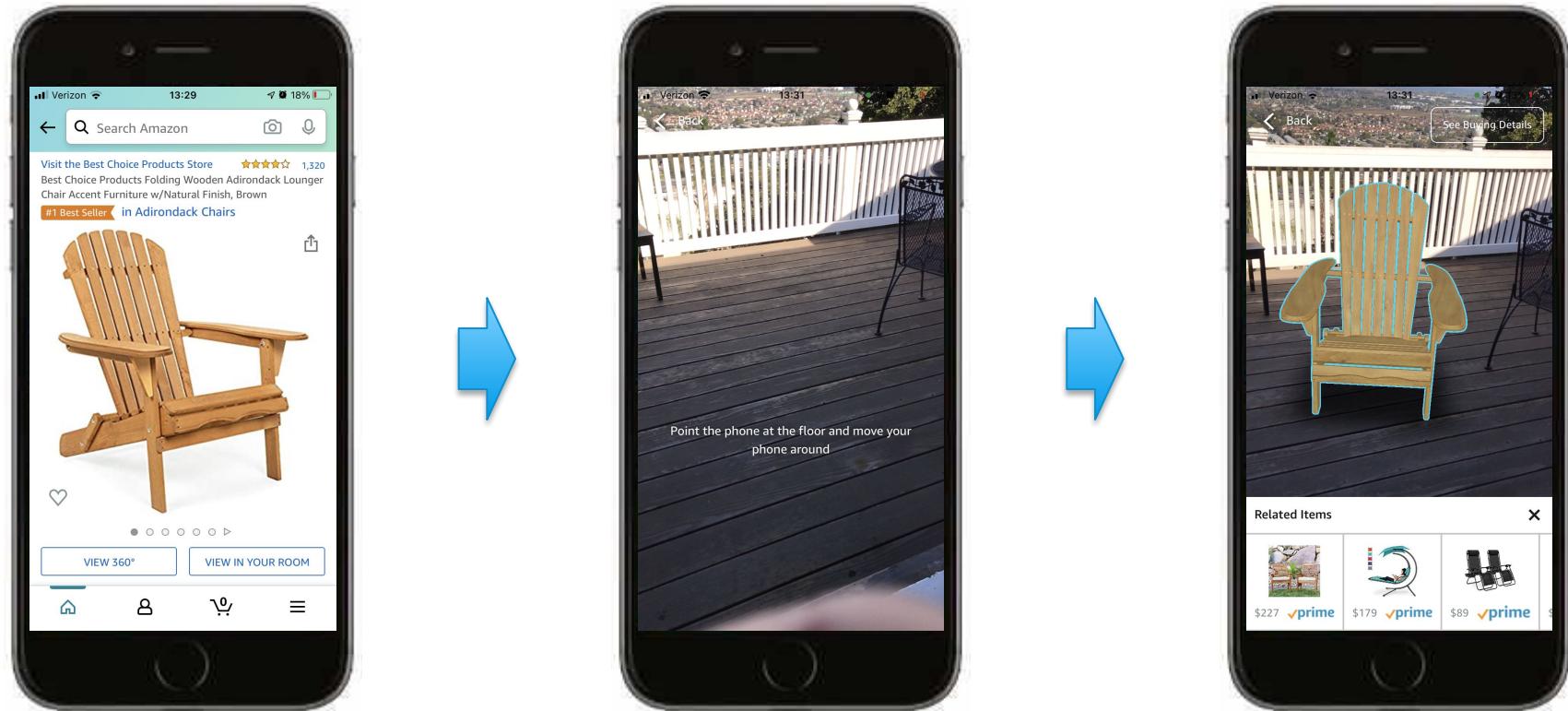
Customer Need: search for clothing visually

ML Solution: Amazon Mobile StyleSnap



Customer Need: validate furniture will fit in my space

ML Solution: Amazon Mobile Augmented Reality



Customer Need: Comparison shopping by price-per-unit

ML Solution: Extract attributes from product imagery and text

The screenshot shows an Amazon product page for Sparkle Paper Towels. At the top, there's a navigation bar with links like 'Health, Household & Baby Care', '12 Days of Deals', and 'Shop 12 Days of Deals'. Below the navigation, there's a search bar and a 'Save 25% in Tools and Home Improvement' banner.

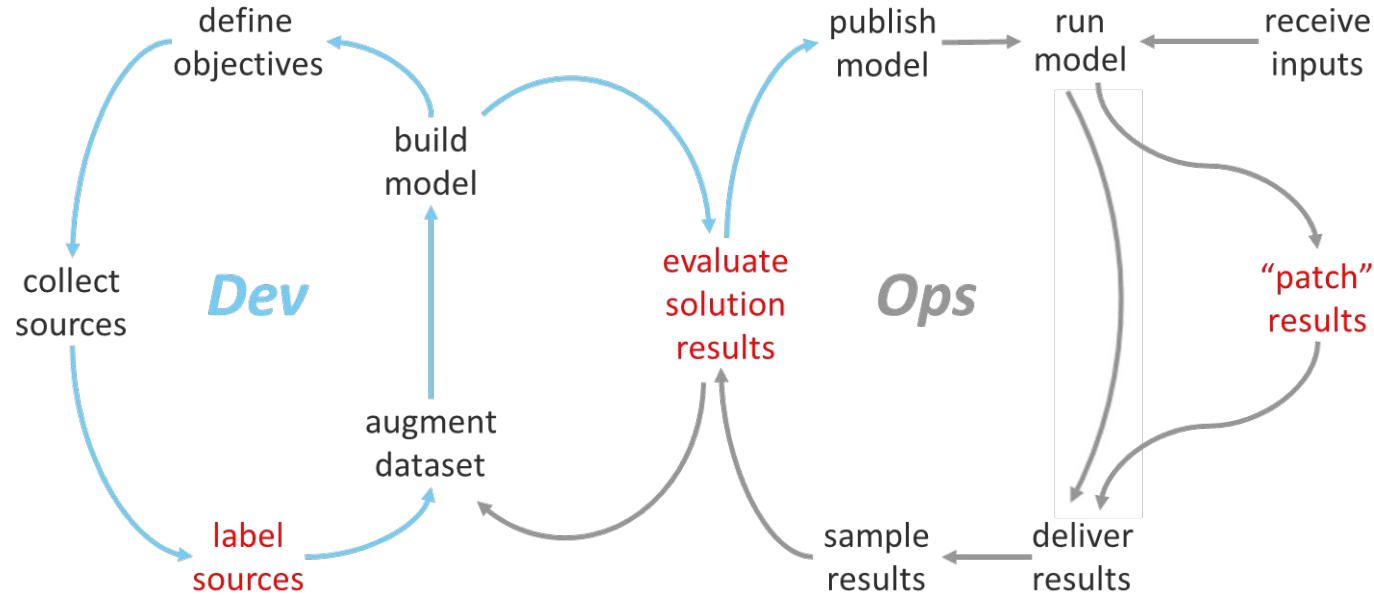
The main product information includes:

- Price:** \$27.54 (\$1.33 / 100 Sheets) ~~\$28.99~~ \$1.45 (5%)
- Coupon Offer:** Extra \$4.00 Off Coupon on first order. [Details](#)
- In Stock:** Ships from and sold by Amazon.com
- Subscription Options:** Subscribe & Save (5% / 15%)
- Purchase Options:** \$27.54 (\$1.33 / 100 Sheets) Unlock 15% by adding this product and 4 more to your next delivery. [Learn more](#)
- One-time Purchase:** \$28.99 (\$1.40 / 100 Sheets)

The page also features a sidebar with product details like 'About the product' and a list of related items.

Core Concepts

Three use cases for HLD within an ML solution: training, evaluating, “patching”



One HLD production process shared across all stakeholders

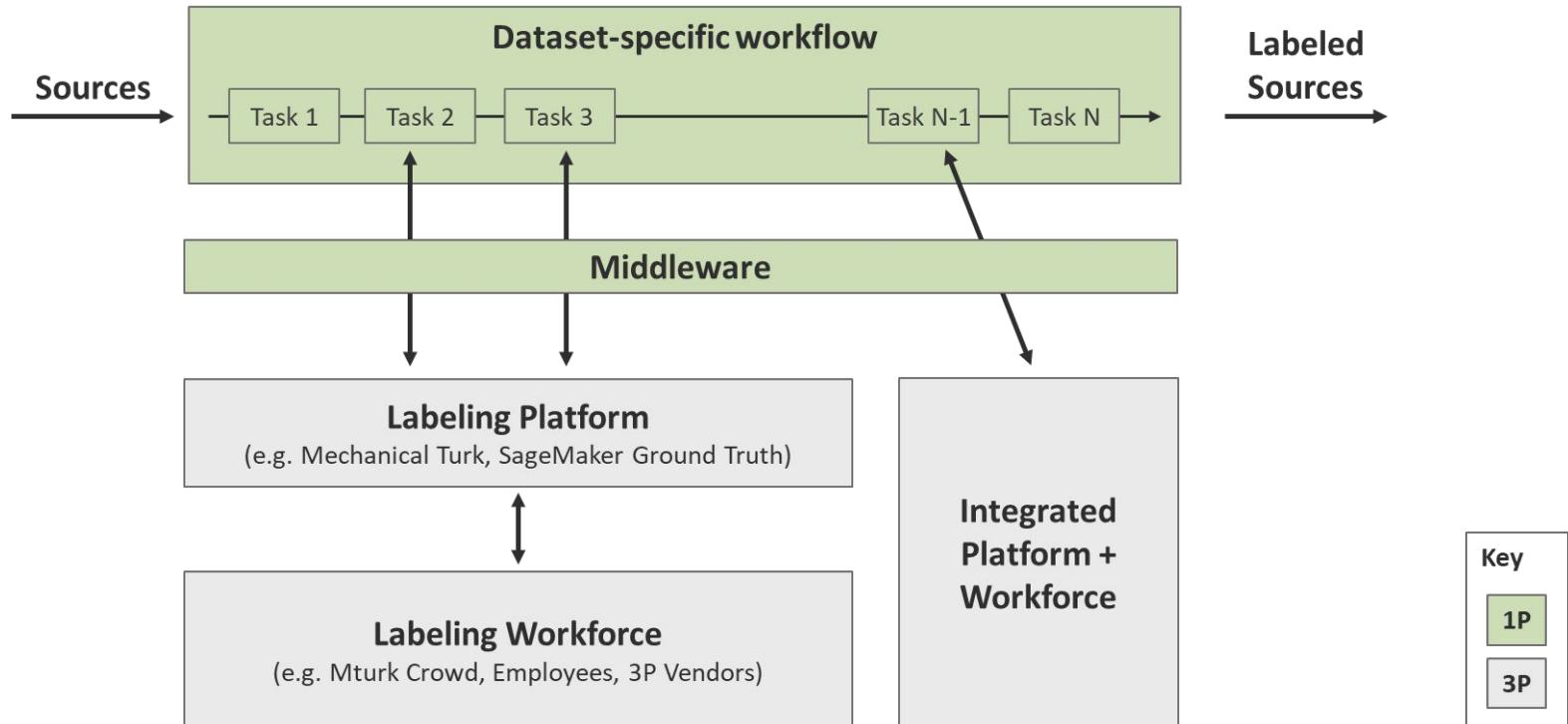


- Set up:
 - Define: Define labeling rules. Tradeoff targets. Build pipeline. Align stakeholders.
 - Pilot: Validate pipeline with subset of target workforce.
- Produce: Train workforce. QA & deliver labels. Track input drift & bias.
- Milestone 1 contains subcycles & costs the most (sometimes by orders of magnitude)
- Milestones 2+ have minimal set up.

Multiple labeling workforces: public crowds, 3P employees in secure locations, and 1P employees on 1P network in secure locations



One production architecture, comprised of a workflow (aka dataset pipeline), task distribution platform(s), & workforce(s).



Dataset privacy, quality, cost and speed are system properties resulting from design levers you control.

Portfolio-level control levers

- Products supported
- ML Expertise
- HLD Process
- HLD Workforces
- HLD Platforms
- etc.

Dataset-level control levers

- ML objectives
- Data sources
- Pipeline design
- Worker selection, training, feedback & assignments
- QA processes, including label aggregation rules
- etc.



Dataset excellence

An optimal tradeoff between dataset quality, cost and speed, given fixed constraints (privacy and product requirements).

Strategic & Operational Challenges

Strategic challenges

- How core is HLD to my business?
- How fast do I need to innovate?
- How do I earn and retain customer trust?
 - Are ML teams making tradeoffs correctly?
- What IP is most critical to retain?
 - How much IP is captured by the model parameters, code and datasets?
- How should investments in ML and HLD be structured?
 - Impact of transfer learning, active learning, LO learning, etc.

Operational Mistakes (1/2)

- Scheduling HLD production as yet another blocking task (the “just 1” dataset trap)
- Leaving design trade-offs implicit to move fast
 - error costs, accuracy, latency, volume, input sources, dataset bias, explainability, cost, etc.
- Treating privacy/quality/speed/cost as intrinsic properties of a workforce/platform
- Not systematically building domain expertise, or outsourcing domain expertise
- Blindness to intrinsic subjectivity; assuming outliers are not so common
- Using labelers as proxies for customers
- Re-using labels from related production systems to simplify integration

Operational Mistakes (1/2)

- Writing too short a guide or too long a guide; allowing unwritten rules to emerge within a workforce.
- Processing tasks in parallel instead of sequentially
- Over-optimizing the tooling
- Hardcoding dependencies on HLD task designs and workforces
- Managing labelers as black box classifiers; over-optimizing labeling toolchains
- Focusing on speed first, instead of quality
- Not paying workforces for usable output
- Insufficient QA
- Under-resourcing the entire HLD production process

Many of the mistakes come from a simple but significant mismatch.

ML Teams

- Focused on building and operating the product
- Too little time for
 - writing annotation guides
 - optimizing pipelines
 - recruiting, training, managing labelers
 - checking data quality
- Biased towards speed (in industry)

Data Labelers

- Focused on labeling data
- Trainable in multiple application domains
- Cannot design guides or tasks
- Cannot build tooling
- Biased towards quality



Mismatch

Dedicated HLD experts can reduce the impedance mismatch, but are not cost-effective for small portfolios.

ML

Teams

- Focused on building and operating the product
- Too little time for
 - writing annotation guides
 - optimizing pipelines
 - recruiting, training, managing labelers
 - checking data quality
- Biased towards speed (in industry)

Dataset Team

- Focused on datasets
- Bloodhounds for edge cases
- Trainable in multiple application domains
- Handle guides, pipelines, training, QA
- Re-purpose work from other ML teams
- Reduce cost via scale
- Biased towards risk tradeoffs

Data Labelers

- Focused on labeling data
- Trainable in multiple application domains
- Cannot design guides or tasks
- Cannot build tooling
- Biased towards quality

Organizational Solutions

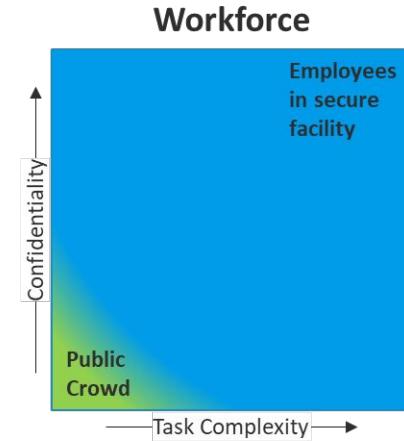
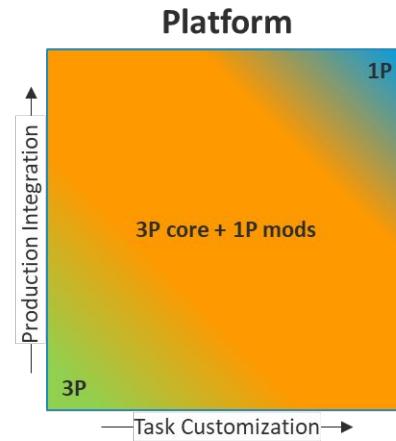
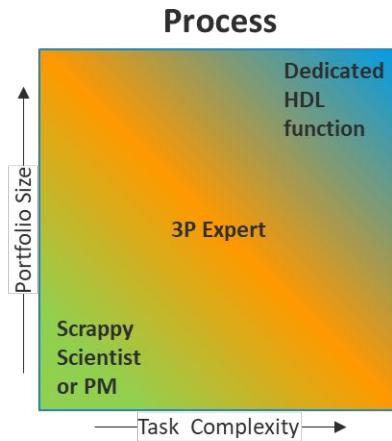
Recap: What are (some) principles a good solution will satisfy?

- Customer
 - Maximize customer trust/benefit, not dataset quality/cost/speed.
 - Make design tradeoffs explicit. Everyone labels.
 - Put data on an even footing with product, science and engineering.
 - Everyone labels.
- Quality & Velocity
 - Experiment quickly, then iterate, iterate, iterate.
 - Reduce dataset bias and ambiguity via seeking out and including diverse perspectives, including from the labelers.
 - Standardize: align stakeholders on quality and process, align labels across portfolio.
- Cost
 - Minimize total cost, i.e. set up + production (data efficiency) + consumption.
 - Assess platform costs over the portfolio lifetime.
- IP
 - Dataset IP is product IP. Value & protect your dataset expertise.
 - Maximize fungibility of workforces.

Possible pros and cons of differing HLD production system options

Component	Option	Pros	Cons
Process	Scientist lead	Understanding of edge cases High speed in early iterations	Cost, missing expertise, IP bottleneck Slow speed at scale
Process	PM lead	Understanding of edge cases	Missing expertise, IP bottleneck
Process	HLD team	Internal alignment, IP distributed, quality, speed, cost	Requires larger product portfolio
Process	3P lead	quality, speed	Cannot share roadmap, alignment difficult, IP lost
Platform	1P	100% control	Cost, speed
Platform	3P core + 1P mods	higher quality	dependency on 3P
Platform	3P	speed	dependency on 3P
Workforce	in house labelers	speed, quality	fixed OpEx
Workforce	crowd	low cost, on demand, variable OpEx	security, training, retention
Workforce	3P employees	quality, cost, variable OpEx, excellence rewarded	time to align vendor with security process, platform requirements

Map your situation. (Disclaimer: build your own maps!)



Other factors

- Domain complexity
- Stability of demand
- etc.

Other factors

- Worker selection
- QA mechanisms
- Bandwidth
- Transaction costs
- Lock-in
- Stability of demand
- etc.

Other factors

- On network
- Fatness of tail
- Contracting options
- QA mechanisms
- Stability of demand
- etc.

The half-life of an organizational solution depends on internal growth, and external dynamics.

Current situation

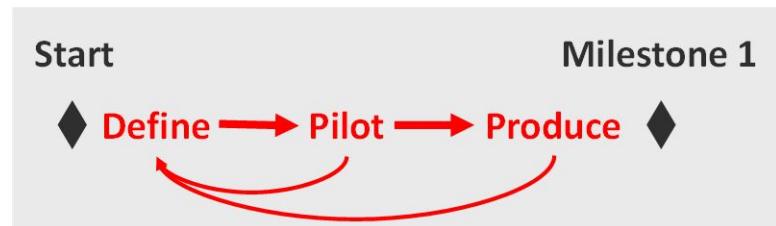
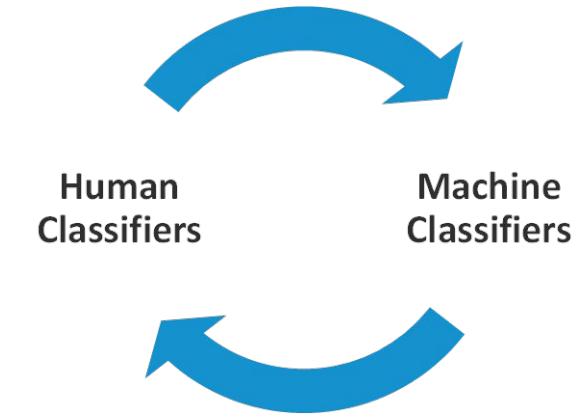
- Types of companies
 - ML + labeling tools
 - Full service HLD
 - Labeling tools
 - Labeling workforces
- Many new entrants
- Long-term trends
 - Dataset sizes decrease
 - # datasets increase
 - Increase in HLD teams

Fundamentals often prevail

- Less instability
 - Tools in the cloud
 - Labelers scale up
 - Full service w/ secure workforces
- More instability
 - Full service HLD vs HLD labelers
 - ML platforms vs HLD platforms

As ML DevOps iterations accelerate, HLD teams may need to reposition themselves as high-speed educators.

- As quality requirements increase, task mixes may shift from ground truthing to validation.
- As # solutions per ML team increases, product refreshes may shift from science to HLD teams.



Thank you!

hallinan@amazon.com

DISCUSSION

30 mins until

10:05a 1:05p 6:05p

PDT

SF

EDT

NYC

CET

AMS

JOIN VIDEO STREAM AT:
underline.io/events/27/main-stage

JOIN SLACK CHAT AT:
bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:

- #ask-questions-for-speakers-here
- #ask-for-help-here

WORKSHOP PROGRAM:
eval.how/dew2020/

CODE OF CONDUCT:
www.humancomputation.com/attend.html#code

WRAP UP

10 mins until end

10:20a 1:20p 6:20p

PDT

SF

EDT

NYC

CET

AMS

JOIN VIDEO STREAM AT:
underline.io/events/27/main-stage

JOIN SLACK CHAT AT:
bit.ly/reais-dew-slack

ASK QUESTIONS AT SLACK:

- #ask-questions-for-speakers-here
- #ask-for-help-here

WORKSHOP PROGRAM:
eval.how/dew2020/

CODE OF CONDUCT:
www.humancomputation.com/attend.html#code

Thanks for joining today! 🙌



DEW 2020

ABOUT IMPORTANT DATES CALL FOR PAPERS ORGANIZERS PROGRAM TALKS

1ST DATA EXCELLENCE WORKSHOP (DEW 2020)

At the 8th AAAI Conference on Human Computation and Crowdsourcing

Date: 26 October 2020
Time PT: 6:00am–10:20am
Time EST: 9:00am–1:20pm
Time CET: 2:00pm–6:20pm
Location: Virtual

IMPORTANT DATES REGISTER TALKS

Next Steps: We invite you to

- Join data-excellence@googlegroups.com, our low-volume mailing list
- Continue the discussion on the slack bit.ly/reais-dew-slack
- Co-author the workshop report that gathers these ideas for wider circulation
- Help organize the next version of this workshop and grow this community
- Bookmark <http://eval.how/dew2020/>

