Machine Learning Systems Design

ML beyond accuracy: Fairness, Security, Governance

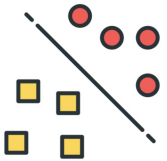
Sara Hooker

Stanford University

CS 329S | Chip Huyen

My research agenda to-date has focused on:

- Going beyond test-set accuracy
- Training models that fulfill multiple desired criteria



Model Compression - compact machine learning models to work in resource constrained environments.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

Model fragility and security - deploy secure models that protect user privacy.

Model Interpretability - reliable explanations for model behavior.

The Role of Model Design: Characterizing Bias and Developing Trustworthy AI Models

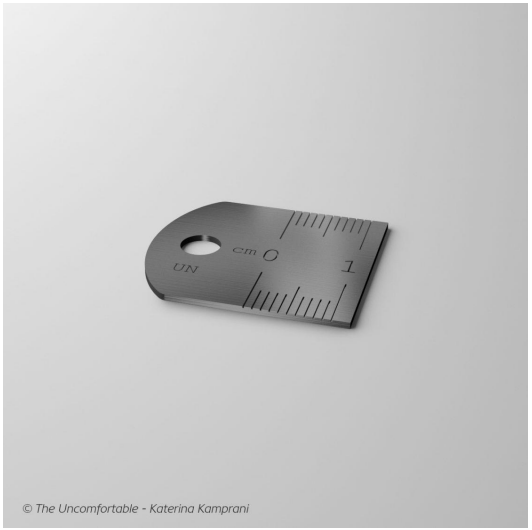
The Imperfect Objective

The role of model design: characterizing bias and developing trustworthy AI models

Understanding trade-off between desiderata.

I'll mention research collaborations with my colleagues:
Nyalleng Moorosi, Gregory Clark, Samy Bengio, Emily Denton, Aaron Courville, Yann Dauphin, Andrea Frome, Chirag Agarwal, Daniel Souza, Dumitru Erhan, Oreva Ahia, Julia Kreutzer.

The imperfect objective.



Google

Katerina Kamprani

What if discomfort is not uniform, but targeted?

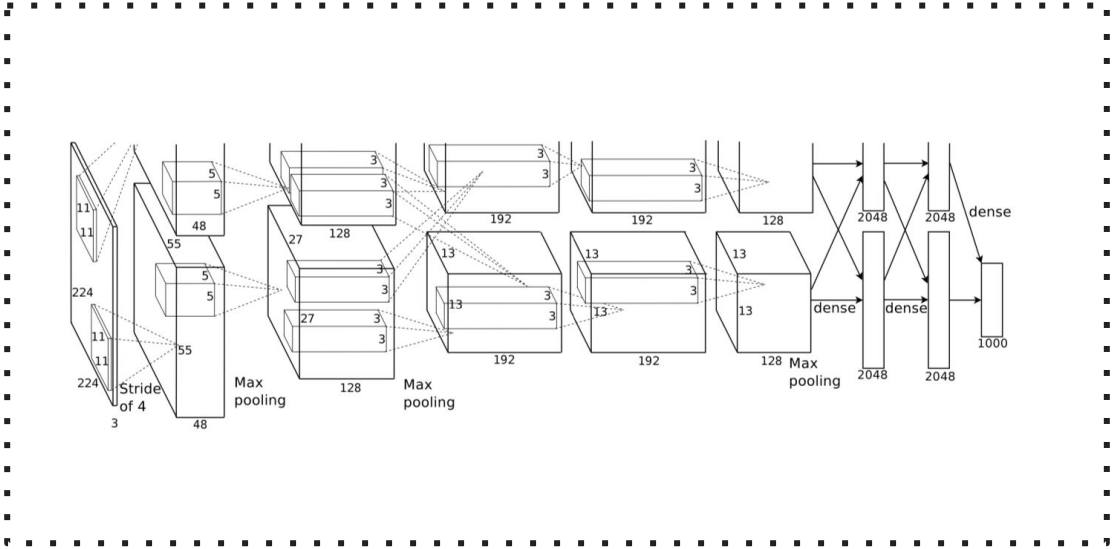
Google

Further reading

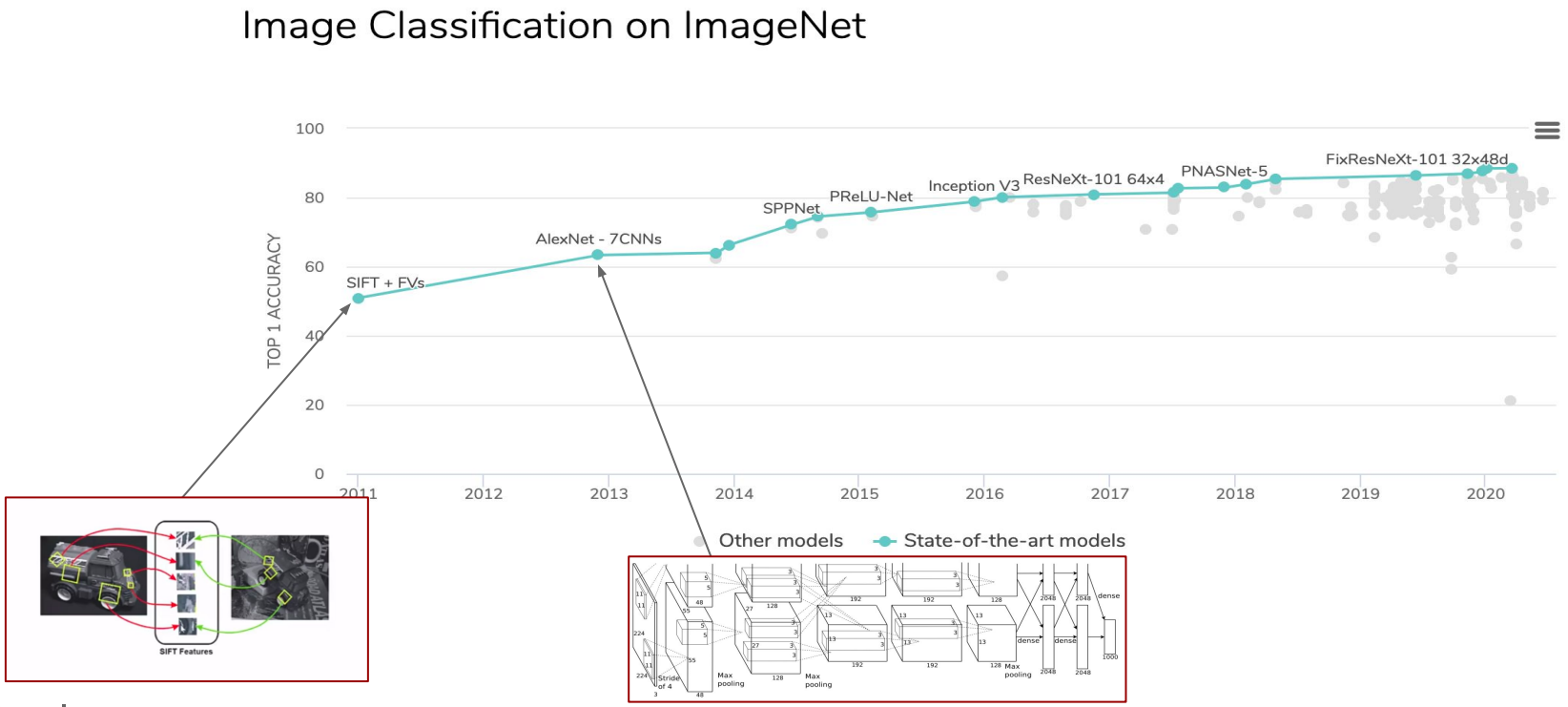
Our goal is to bridge the technology design gap - develop technology that works for everyone.

Our goal is to bridge the technology design gap - develop technology that works for everyone.

Achieving this requires understanding how our modelling choices impact downstream impact.



Over the last decade “performance” of a model treated as synonymous with pursuit of top-1 accuracy.



Google

Top-line metrics do not guarantee that the trained function fulfills other properties we may care about.

$$loss = \sum_{i=1}^{B} \mathcal{L}(y_i, \hat{y}_i)$$

Empirical risk minimization - train a representation to minimize average error.

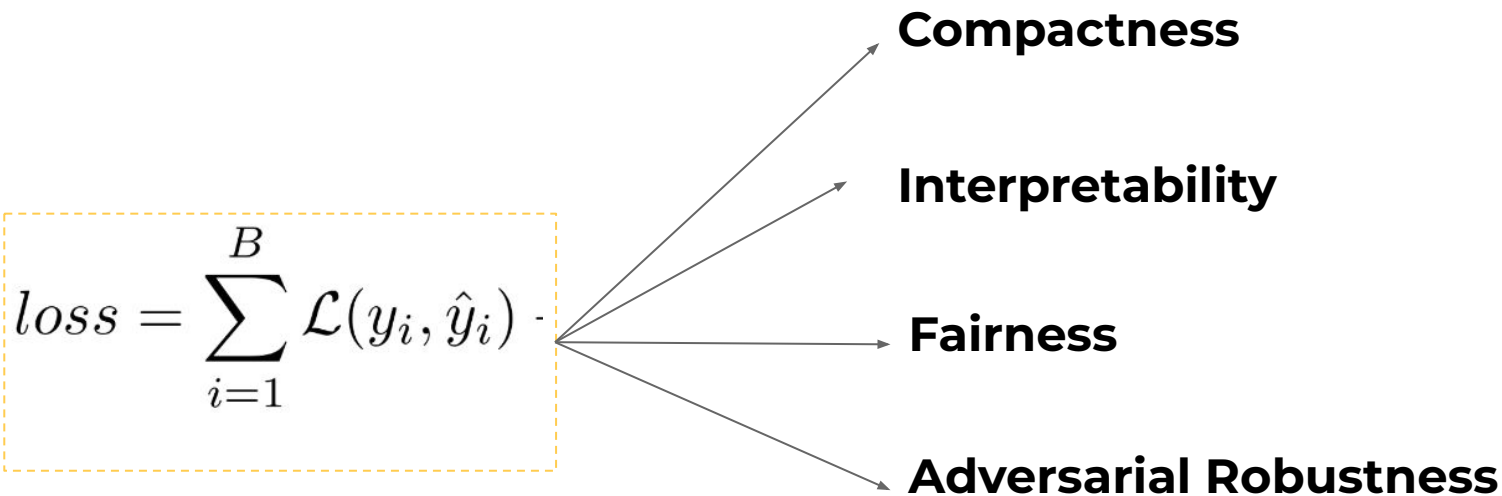
Compactness

Interpretability

Fairness

Robustness

Typical loss functions in machine learning (MSE, Hinge-Loss and CE) impose no preference for functions that are interpretable, fair, robust or guarantee privacy.



Donald Knuth said “computers do exactly what they are told, no more and no less.”

A model can fulfill an objective in many ways, while violating the spirit of said objective.

The Clever Hans Effect 1891 - 1907

Hans the horse:

- arithmetic functions
- identify colours
- Count the crowd

High accuracy without true learning.

Experimental Design -
Can Hans answer a question if the human does not know the answer?

Hans answered correctly by picking up on microscopic clues.

The under-specification of our objective function often leads to undesirable model behavior termed “shortcut learning.”

Cow

Limousine

Berry et al. (paper link)
Hooker et al. 2019 (paper link)

The under-specification of our objective function often leads to undesirable model behavior “shortcut learning.”

Sheep

A herd of sheep grazing on a lush green hillside
Tags: grazing, sheep, mountain, cattle, horse

Dog

Left: A man is holding a dog in his hand
Right: A woman is holding a dog in her hand
Image: @SouperSarah

Blog [link]

The under-specification of our objective function often leads to undesirable model behavior termed “shortcut learning.”

Task is predicting Ending of Story:

Context	Right Ending	Wrong Ending
Sammy’s coffee grinder was broken. He needed something to crush up his coffee beans. He put his coffee beans in a plastic bag. He tried crushing them with a hammer.	It worked for Sammy.	Sammy was not that much into coffee.
Gina misplaced her phone at her grandparents. It wasnt anywhere in the living room. She realized she was in the car before. She grabbed her dads keys and ran outside.	She found her phone in the car.	She didnt want her phone anymore.
Sarah had been dreaming of visiting Europe for years. She had finally saved enough for the trip. She landed in Spain and traveled east across the continent. She didn’t like how different everything was.	Sarah decided that she preferred her home over Europe.	Sarah then decided to move to Europe.

Table 1: Example Story Cloze Test instances from the Spring 2016 release.

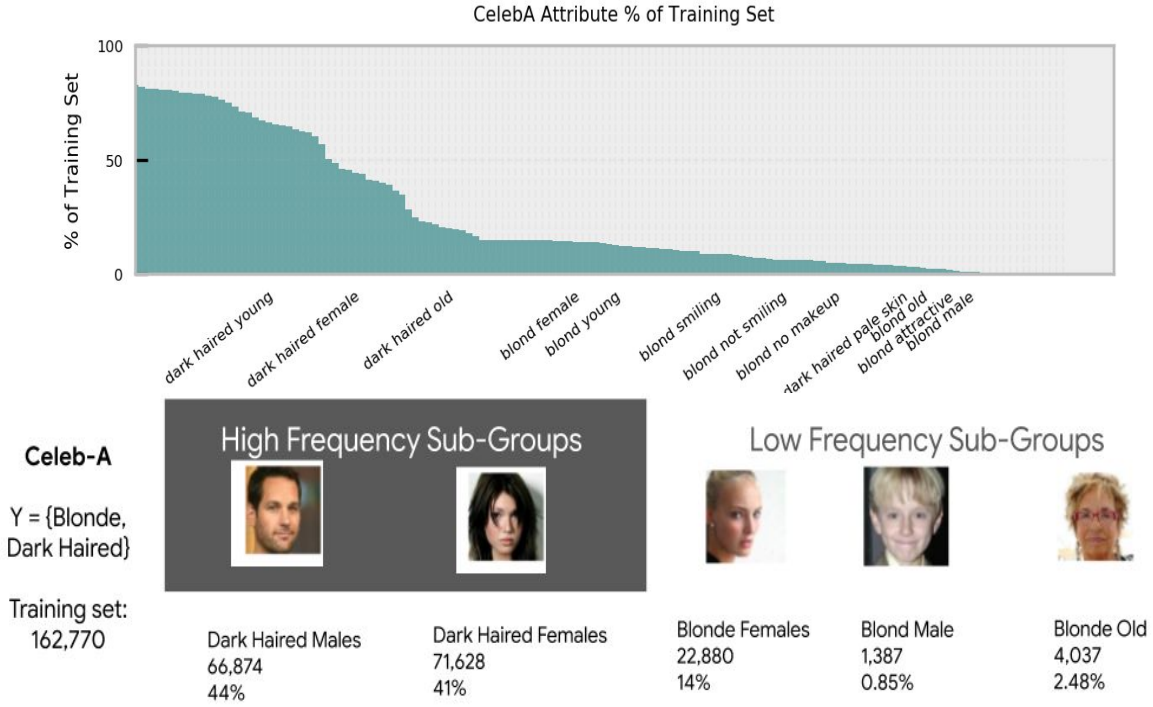
Pay Attention to the Ending -

Work by Cai et al. show show you can discard the story plots and only train to choose the better of two endings reaches 72.5% accuracy.

[[Cai et al. 2017]]

High accuracy without “true” learning.

“Shortcut learning” is due to relative under/overrepresentation of training features.

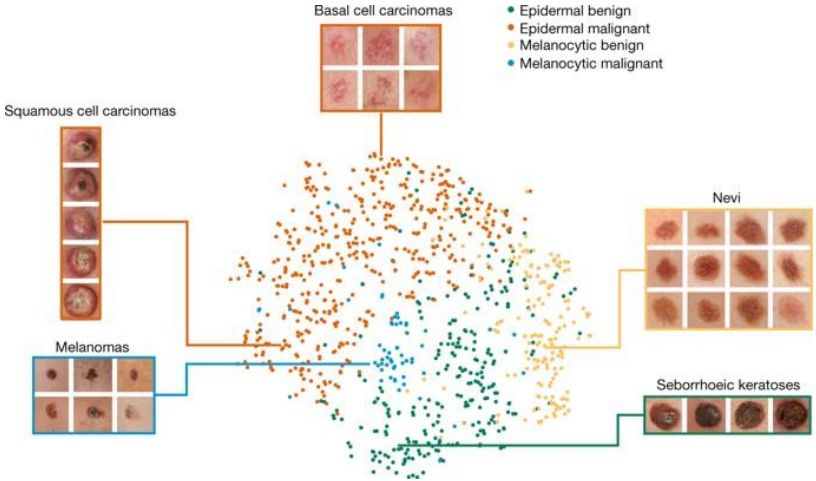


For example, a model may learn to correlate blonde with being a female because there are far fewer blonde males in the training dataset.

This results in higher error rates on the long-tail -- the underrepresented features in the dataset.

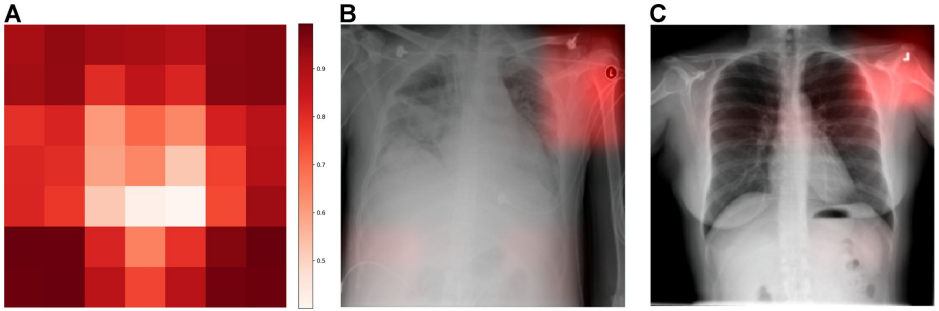
When this happen in sensitive domains, there can be a huge cost to human welfare.

Skin lesions



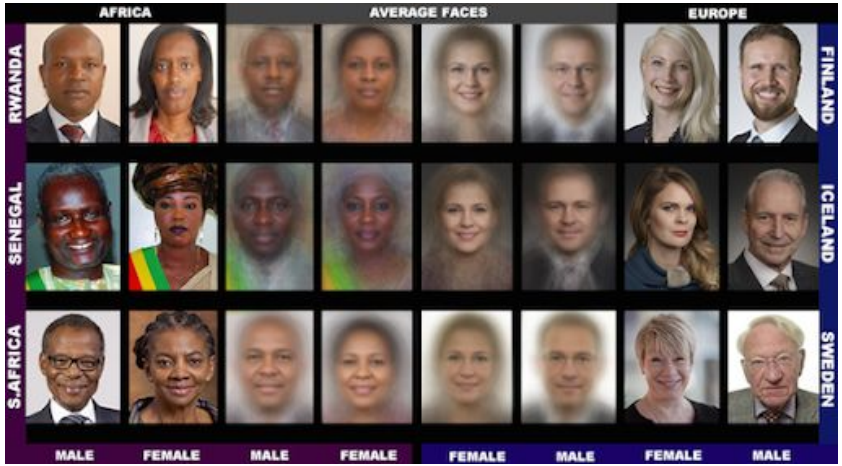
Esteva et al. (link)
Zech et al. 2018 (link)
AlBadaway et al. 2018 (link)

Pneumonia



High accuracy without “true” learning.

How a model treats underrepresented features in the long-tail of the distribution often coincides with notions of fairness.



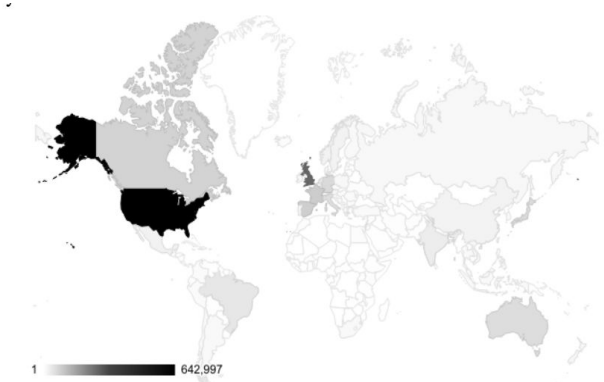


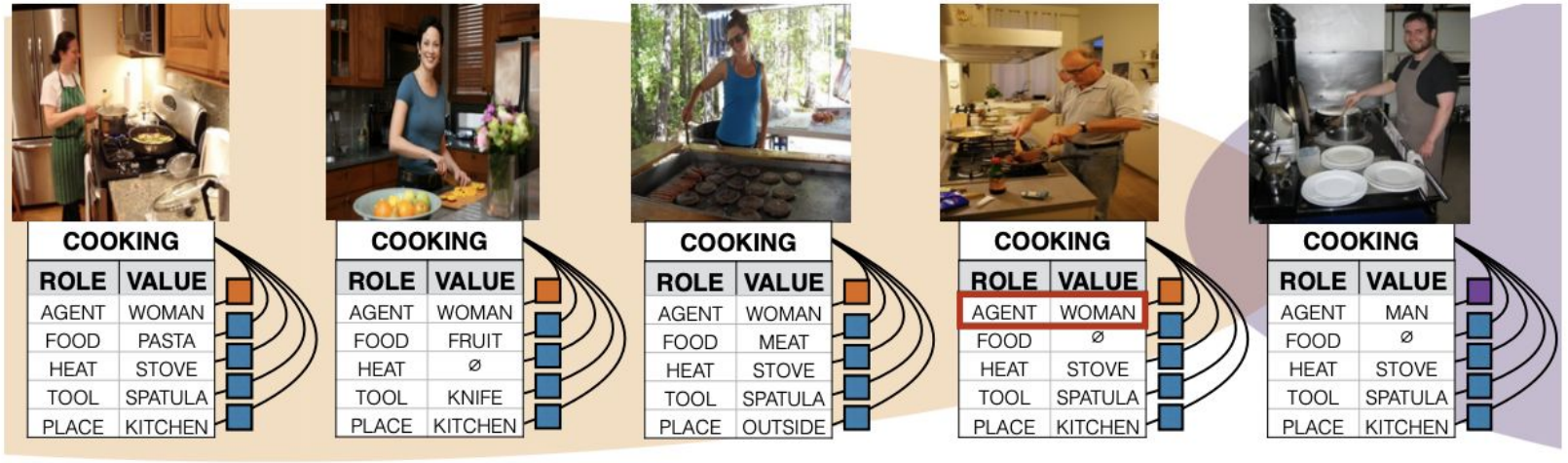
Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe.

Gender shades (link)
Shankar et al. (link)
Zhao, Jieyu et al. (2017)

Geographic bias in how we collect our datasets. Shankar et al. (2017) show models perform far worse on locales undersampled in the training set.

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World (Shankar et al. (link))

Undersampling/oversampling leads to undesirable spurious correlations.
Zhao, Jieyu et al. (2017) show Activity recognition datasets exhibit stereotype-aligned gender biases.



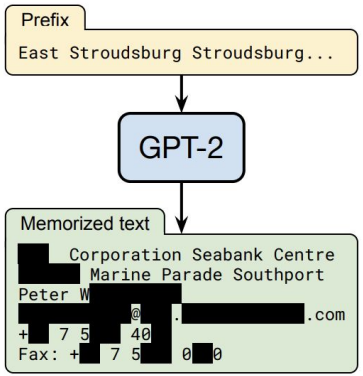
Men also like shopping (and cooking too).

Zhao, Jieyu et al. (2017).

Delegating learning of the function to the model can (and has) led to Clever Hans moments.

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Overfitting to pattern matching



Memorization leads to Leakage of private text
Carlini et al. 2021

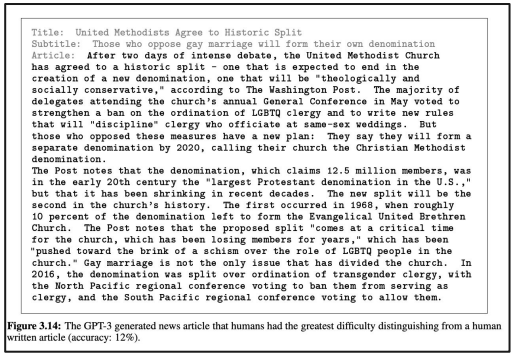

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Can generate factually incorrect statements
Brown et al, 2020

High accuracy without "true" learning.

Open challenges in auditing and mitigating harmful bias.

Fairness

Preferences about how our trained model should behave on subset of sensitive or protected features.

Legally protected features:
Certain attributes are protected by law. For example, in the US it is illegal to discriminate based upon race, color, religion, sex, national origin, disability.

Legal framework will differ by country.

Sensitive features:
Income, eye color, hair, skin color, accent, locale.

These features may not be protected by law, but are often correlated with protected attributes .

Your choice of tool to audit and mitigating algorithmic bias will depend upon whether you know:

- the sensitive features which are adversely impacted
- have comprehensive labels for these features

- Unknown bias
- Incomplete or no labels for sensitive features

- Known concern
- Comprehensive labels

Your choice of tool to audit and mitigate algorithmic bias will depend upon whether you know:

- the sensitive features which are adversely impacted
- have comprehensive labels for these features

1.

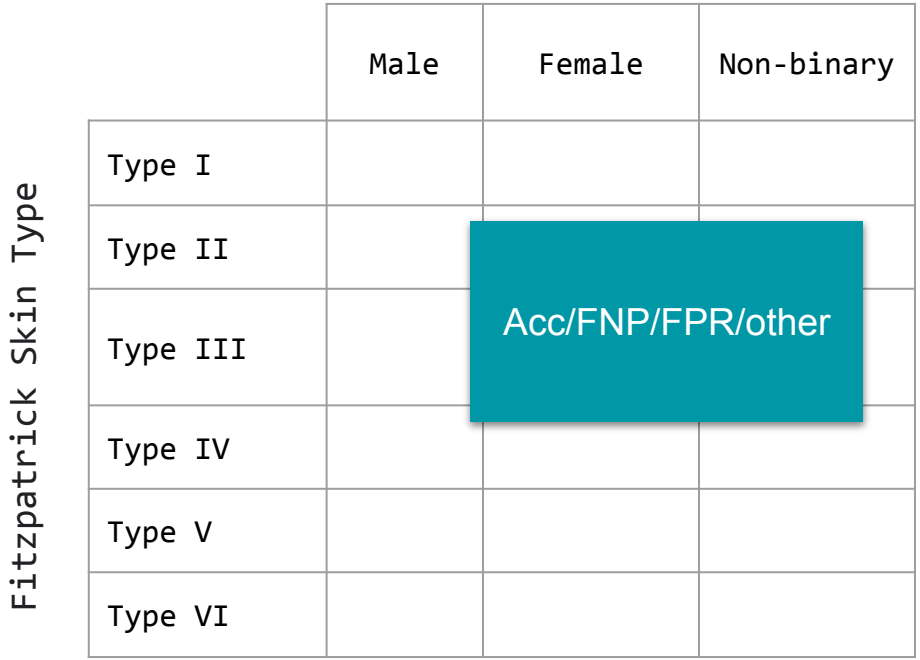
- Unknown bias
- Incomplete or no labels for sensitive features

- Known concern
- Comprehensive labels

1. With known and comprehensive labels - track impact using intersectional metrics

What is it?
Statistically evaluate model performance (e.g. accuracy, error rates) by “subgroup” e.g. skin tone, gender, age

Requires
Good, “balanced” test sets that are representative of the actual use-case(s) for the model in production



Example of intersectional audit

Gender Shades - Evaluated classifiers' performance across genders, skin types, and intersection of gender and skin type

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Paper

When labels are known and complete - opens up range of remedies to mitigate impact

Data-Based

1. Re-balance or re-weight sensitive features to balance training set.

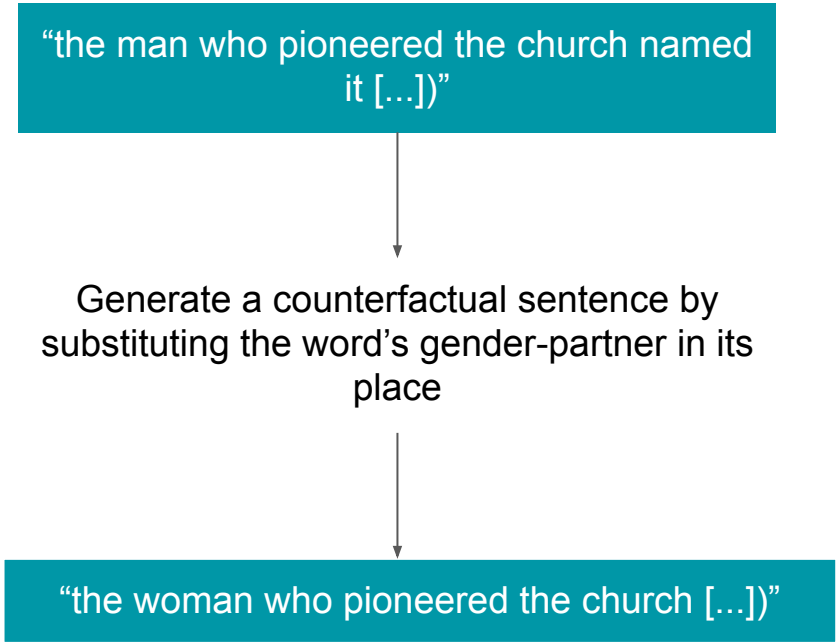
2. Remove problematic feature from training set (**not always feasible**)

3. Tailored Augmentation Strategies

Examples of counterfactual data augmentation remedies:

Counterfactual data augmentation strategies (CDA):

- Can involve duplicating examples and swapping gendered terms in training data can help with debiasing word embeddings, pretrained language models, coreference resolution models.



[[Lu et al. (2018), Maudslay et al. (2019), Webster et al. (2021)]]

When labels are known and complete - range of remedies to mitigate impact.

Data-Based

1. Re-balance or re-weight sensitive features to balance training set.

2. Remove problematic feature from training set (**not always feasible due to proxy variables**)

3. Tailored Augmentation Strategies

Model-Based

1. Min diff - penalizes model for differences in treatment of distributions

2. Rate constraint - guaranteeing recall, data co-occurrence or another rate metric is at least/most [x%].

[[zhao et al. 2017]]

3. Worst case error constraint

What about where we don't have complete labels for the sensitive attribute we care about?

2.

- Unknown bias
  Incomplete or no labels for sensitive features

- Known concern
- Comprehensive labels

Can you spot a metric which doesn't require labels?

How **does** my model perform...

Classification accuracy / precision-recall curve / logarithmic loss / area under the curve / mean squared error / mean absolute error / F1 score / standard deviation / variance / confidence intervals / KL divergence / false positive rate / false negative rate /

How **might** my model perform...

on a sample of test data / on cross-slices of test data / on an individual data point / if a datapoint is perturbed / if model thresholds were different / / across all values of a feature / when compared to a different model

Most of our remedies are centered around the assumption that we have comprehensive labelling available. However, this is a tenuous assumption.

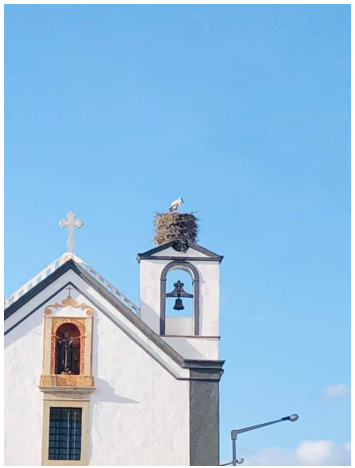
Most of our remedies are centered around the assumption that we have comprehensive labelling available. However, this is a tenuous assumption.

Any thoughts why? What challenges exist around ensuring comprehensive labelling?

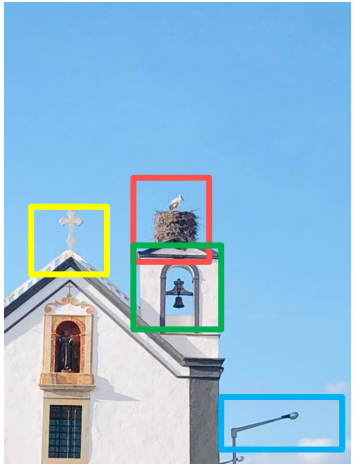
More often than not, we do not have comprehensive labels from human annotators.

1) For high dimensional problems, often time consuming/infeasible to comprehensively label.
2) Labelling sensitive features are insufficient, necessary to label all proxy features.
3) Legal obstacles around collecting certain sensitive features [[Andrus et al. 2021, Veal 2017]].
4) Even when collected, issues with consistency in annotation [[Khan et al. 2021]]
5) Annotation can be biased by the lived experience of the annotators
6) What we deem to be harm is not static. It is shaped by political/geographical/economic/historical considerations.

1. For high dimensional problems, often time consuming/infeasible to comprehensively label.



church



Bird, nest, street lamp, cross, statue, window, window grid.

1.1. For NLP tasks, often require separately curating labels for different languages.

the man who pioneered the church named it [...])”

Requires a curated list of words to substitute. This has to be curated for different languages.

Generate a counterfactual sentence by substituting the word’s gender-partner in its place

“the woman who pioneered the church [...])”

The GEM Benchmark [[Germann et al. 2021]]

2. Labelling protected features is often insufficient, necessary to label all proxy features.





Task: Sleeping or awake?

Consider *species* to be the protected attribute, many other variables may be proxy variables (**indoor/outdoor**).

2. Labelling protected features is often insufficient, necessary to label all proxy features.

In languages like Portuguese, remedies like augmentation by swapping protected attributes require not only identifying protected nouns but also altering any words that are modified by the noun.

2. Labelling protected features is often insufficient, necessary to label all proxy features.

A minha mãe é Professora.

O meu pãe é Professor.

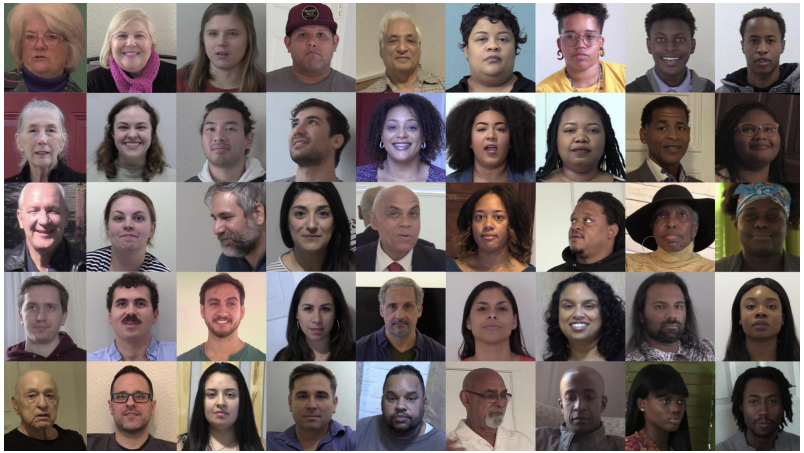
My mom is a teacher.

My dad is a teacher.

Gender swapping is much harder in languages where the adjectives, articles, and pronouns that agree with these nouns also adjust to comply with gender.

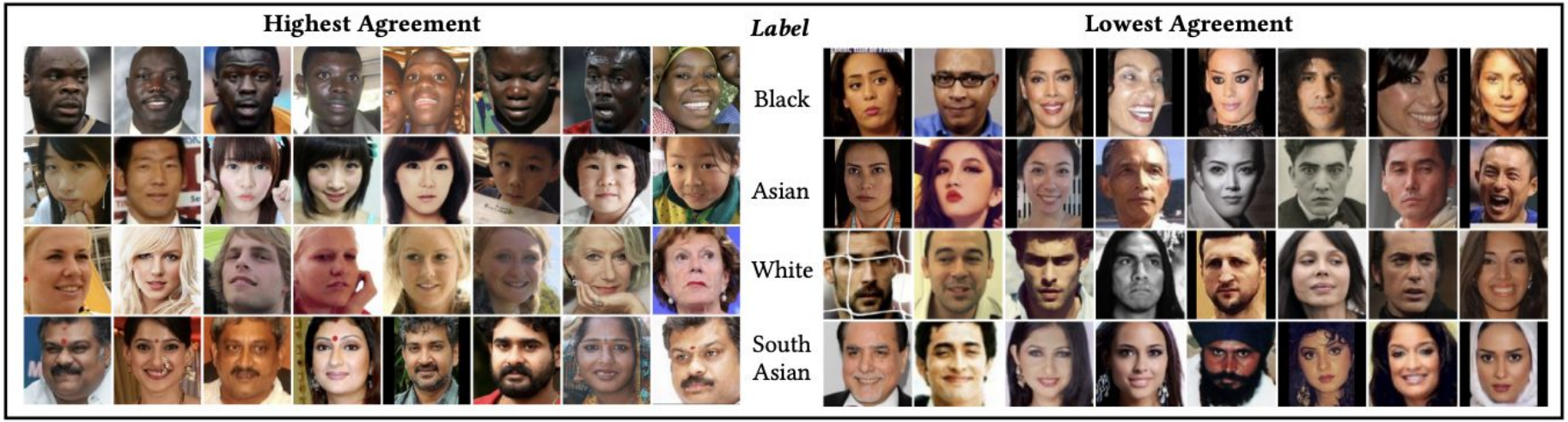
3. Legal obstacles around collecting certain sensitive features.

Recent dataset release by Facebook notable for both compensating and getting consent.



[[Andrus et al. 2021, Veal 2017]].

4. Inconsistency in how sensitive features are labelled.



Images with **high** (left) and **low** (right) levels of agreement on how to annotate race.

[[Khan et al. 2021]]

4.1 Annotation errors are prevalent in most large scale datasets.

Tagged for the profession **model**:

"Hank Sheinkopf is a **model for brilliant communications** in a world where messages are broadcast at astounding rates. He's a player in the PR world of political campaigns, both in domestic and foreign sectors."

Example found by Preethi Seshadri.

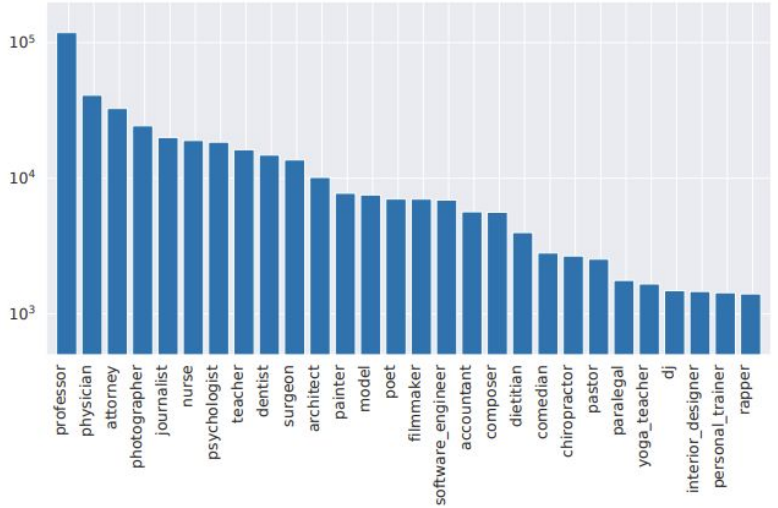


Figure 1: Distribution of the number of biographies for the twenty-eight different occupations, shown on a log scale.

[[Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes SettingDe-Arteaga et al. 2019]]

4.1 The subjectivity of certain labels may result in high annotator disagreement /variance.

What constitutes toxic speech?

[[Akin et al. 2018]]

4.1 The subjectivity of certain labels may result in high annotator disagreement /variance.

What constitutes toxic speech?

Toxicity without swear words. Davidson et al. (2017) phrase the problem that hate speech may not contain hate or swear words at all.

Example: *"she looks like a horse"*

[[Akin et al. 2018]]

5. Annotation can be biased by the lived experience of the annotators

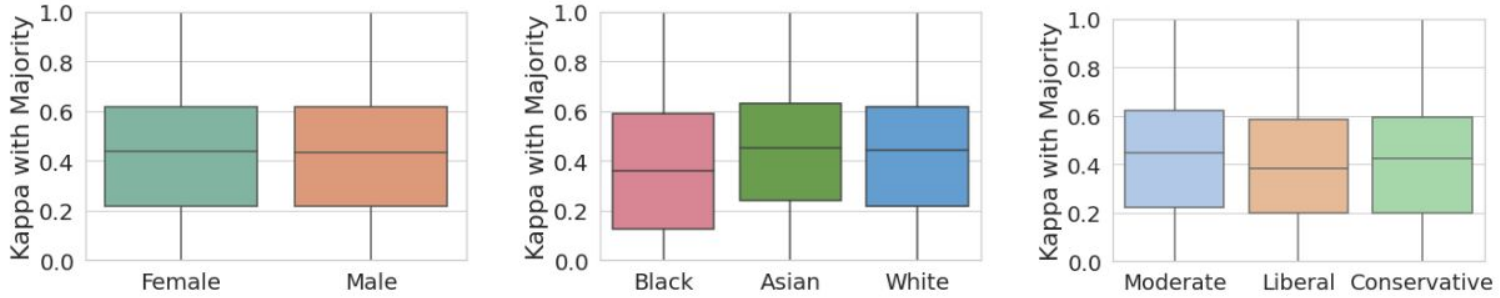


Figure 2: Average and standard deviation of annotator agreement with aggregated labels, calculated for annotators grouped by their socio-demographics under gender, race, and political affiliation.

label aggregation may introduce representational biases of individual and group perspectives.

[[On Releasing Annotator-Level Labels and Information in Datasets

Prabhakaran, Davani et al. 2021]]

6. What we deem to be harm is not static. It is shaped by political/geographical/economic/historical considerations.

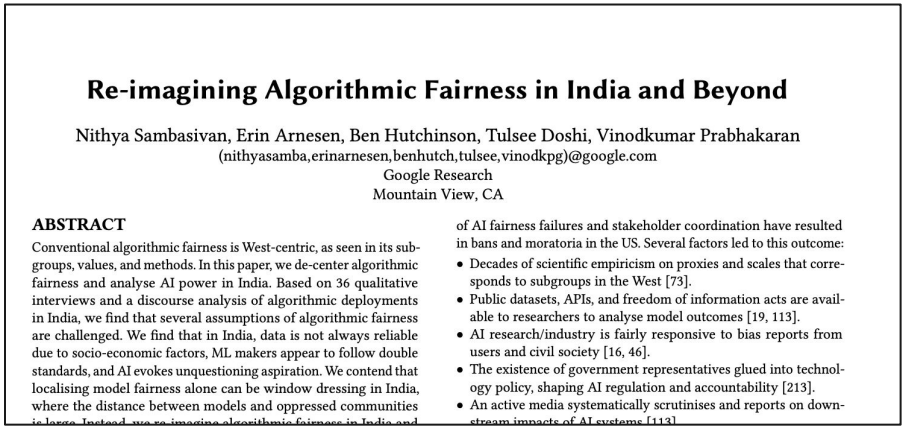
6. What we deem to be harm is not static. It is shaped by political/geographical/economic/historical considerations.

The rules and social contract we ask citizens to abide by has changed over time, and varies by country.

In 1963, the Swiss government passed the first measures to make sure every inhabitant had access to a nuclear shelter.

1999 - Michigan removes law prohibiting citizens from “reproachful or contemptuous language” in print against anyone who declines a duel challenge.

6. What we deem to be harm is not static. It is shaped by political/geographical/economic/historical considerations.



Re-imagining Algorithmic Fairness in India and Beyond

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, Vodkumar Prabhakaran
(nithyasamba,erinarnesen,benhutch,tulsee,vinodkpg)@google.com
Google Research
Mountain View, CA

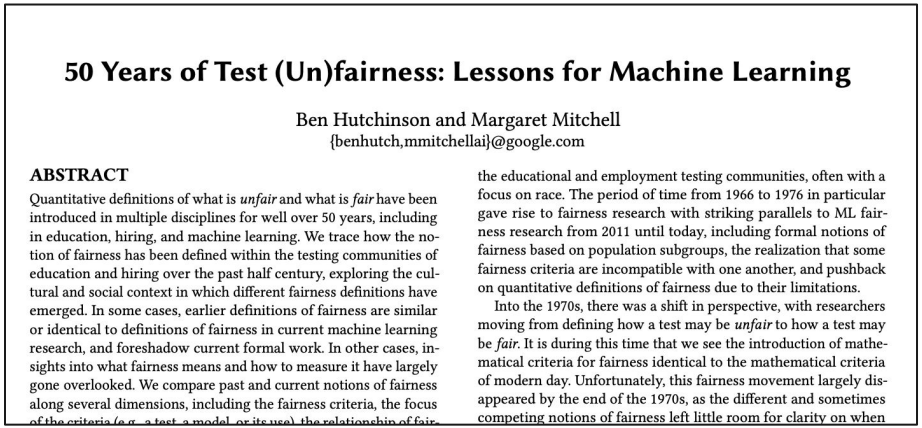
ABSTRACT

Conventional algorithmic fairness is West-centric, as seen in its sub-groups, values, and methods. In this paper, we de-center algorithmic fairness and analyse AI power in India. Based on 36 qualitative interviews and a discourse analysis of algorithmic deployments in India, we find that several assumptions of algorithmic fairness are challenged. We find that in India, data is not always reliable due to socio-economic factors, ML makers appear to follow double standards, and AI evokes unquestioning aspiration. We contend that localising model fairness alone can be window dressing in India, where the distance between models and oppressed communities is large. Instead, we re-imagine algorithmic fairness in India and

of AI fairness failures and stakeholder coordination have resulted in bans and moratoria in the US. Several factors led to this outcome:

- Decades of scientific empiricism on proxies and scales that corresponds to subgroups in the West [73].
- Public datasets, APIs, and freedom of information acts are available to researchers to analyse model outcomes [19, 113].
- AI research/industry is fairly responsive to bias reports from users and civil society [16, 46].
- The existence of government representatives glued into technology policy, shaping AI regulation and accountability [213].
- An active media systematically scrutinises and reports on downstream impacts of AI systems [113]

[[Sambasivan et al. 2021]]



50 Years of Test (Un)fairness: Lessons for Machine Learning

Ben Hutchinson and Margaret Mitchell
{benhutch,mmitchellai}@google.com

ABSTRACT

Quantitative definitions of what is *unfair* and what is *fair* have been introduced in multiple disciplines for well over 50 years, including in education, hiring, and machine learning. We trace how the notion of fairness has been defined within the testing communities of education and hiring over the past half century, exploring the cultural and social context in which different fairness definitions have emerged. In some cases, earlier definitions of fairness are similar or identical to definitions of fairness in current machine learning research, and foreshadow current formal work. In other cases, insights into what fairness means and how to measure it have largely gone overlooked. We compare past and current notions of fairness along several dimensions, including the fairness criteria, the focus of the criteria (e.g., a test, a model, or its use), the relationship of fair-

the educational and employment testing communities, often with a focus on race. The period of time from 1966 to 1976 in particular gave rise to fairness research with striking parallels to ML fairness research from 2011 until today, including formal notions of fairness based on population subgroups, the realization that some fairness criteria are incompatible with one another, and pushback on quantitative definitions of fairness due to their limitations.

Into the 1970s, there was a shift in perspective, with researchers moving from defining how a test may be *unfair* to how a test may be *fair*. It is during this time that we see the introduction of mathematical criteria for fairness identical to the mathematical criteria of modern day. Unfortunately, this fairness movement largely disappeared by the end of the 1970s, as the different and sometimes competing notions of fairness left little room for clarity on when

[[Hutchinson et al. 2018]]

Fairness considerations are not static across time or space.

We just covered several of the key challenges in ensuring comprehensive labelling.

1) For high dimensional problems, often time consuming/infeasible to comprehensively label.
2) Labelling sensitive features are insufficient, necessary to label all proxy features.
3) Legal obstacles around collecting certain sensitive features [[Andrus et al. 2021, Veal 2017]].
4) Even when collected, issues with consistency in annotation [[Khan et al. 2021]]
5) Annotation can be biased by the lived experience of the annotators
6) What we deem to be harm is not static. It is shaped by political/geographical/economic/historical considerations.

If we cannot guarantee we have fully addressed bias in the data pipeline, the overall harm in a system is a product of the interactions between the data and our model design choices.

Recognizing how model design impacts harm opens up new mitigation techniques that are far less burdensome than comprehensive data collection.

- Unknown bias
- Incomplete or no labels for sensitive features

- Known concern
- Comprehensive labels

Leveraging model signal to audit large scale datasets

Global feature importance - Ranks dataset examples by which are most challenging.

Surfaces a tractable subset of the most challenging/least challenging examples for human inspection. Avoids time consuming need to inspect every example.

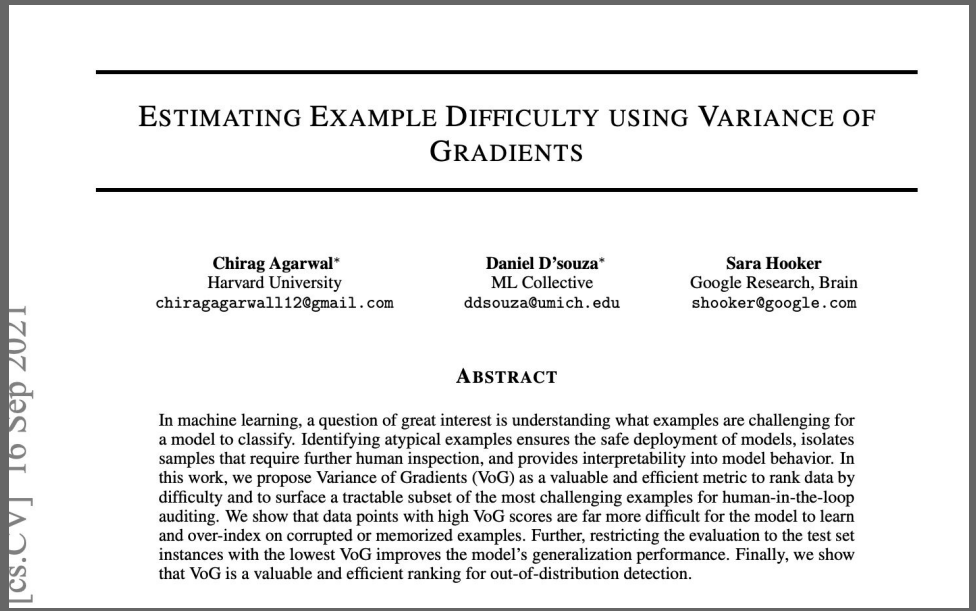
Data Cleaning

Isolating subset for relabelling

Identify issues with fairness

One of the biggest challenges in improving datasets is identifying where to invest limited human annotator time.

CACM



ESTIMATING EXAMPLE DIFFICULTY USING VARIANCE OF GRADIENTS

Chirag Agarwal*
Harvard University
chiragagarwall12@gmail.com

Daniel D'souza*
ML Collective
ddsouza@umich.edu

Sara Hooker
Google Research, Brain
shooker@google.com

ABSTRACT

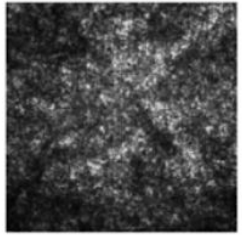
In machine learning, a question of great interest is understanding what examples are challenging for a model to classify. Identifying atypical examples ensures the safe deployment of models, isolates samples that require further human inspection, and provides interpretability into model behavior. In this work, we propose Variance of Gradients (VoG) as a valuable and efficient metric to rank data by difficulty and to surface a tractable subset of the most challenging examples for human-in-the-loop auditing. We show that data points with high VoG scores are far more difficult for the model to learn and over-index on corrupted or memorized examples. Further, restricting the evaluation to the test set instances with the lowest VoG improves the model's generalization performance. Finally, we show that VoG is a valuable and efficient ranking for out-of-distribution detection.

[[Agarwal et al. 2021]]

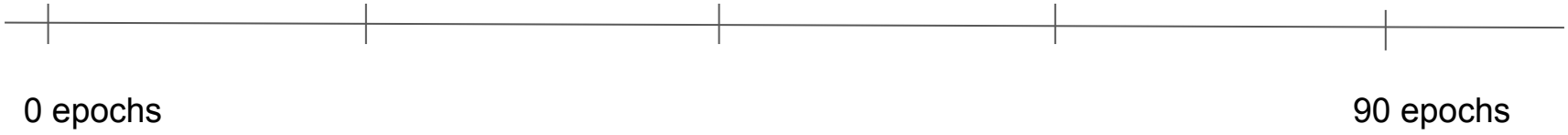
Variance of Gradients (VoG) is an example of a global ranking tool.

$$VOG_i = \frac{1}{N}\sqrt{\left(\frac{1}{N_t}\sum_{t=1}^{N_t}(S_{ti} - \mu_i)^2\right)}$$

gradient



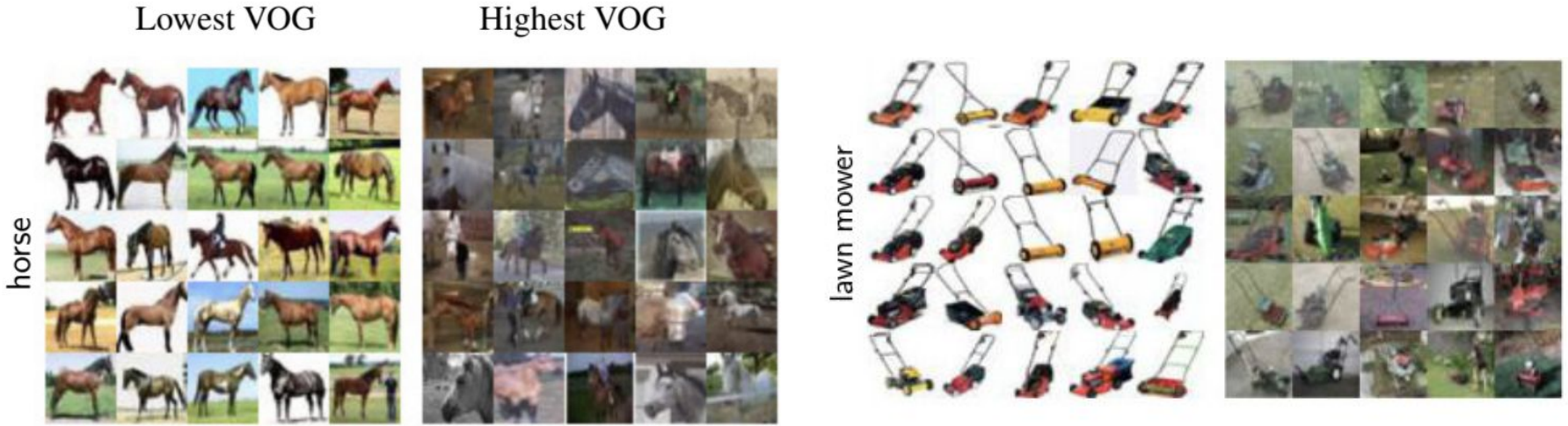
Compute average variance in gradients (VOG) for an image over training.



Estimating Example Difficulty using Variance of Gradients, Agarwal, D,Souza and Hooker, 2020

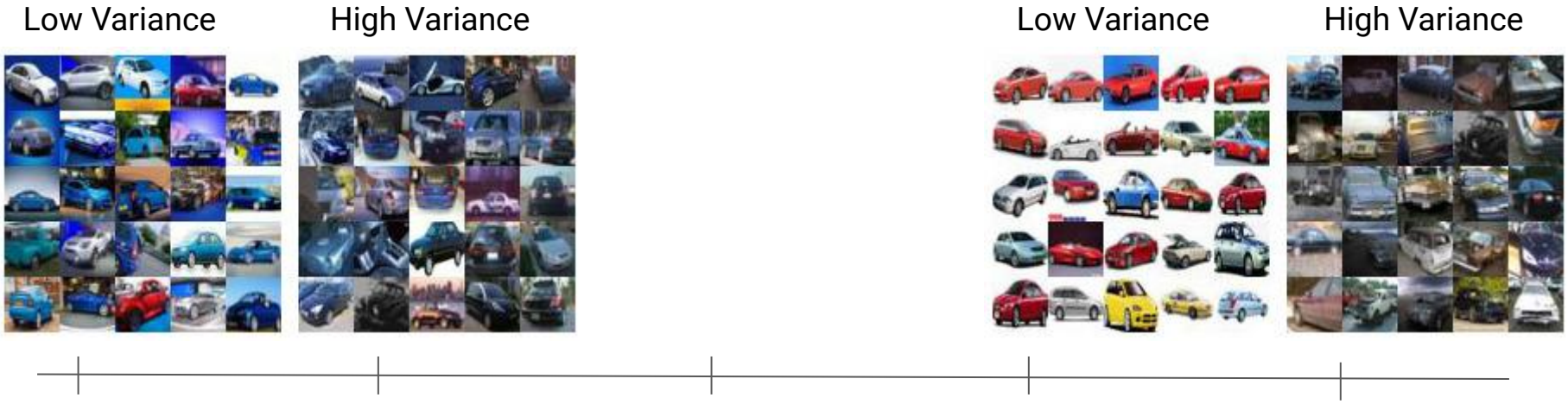
VoG computes a relative ranking of each class.

What examples does the model find challenging or easy to learn?



Estimating Example Difficulty using Variance of Gradients, Agarwal, Souza and Hooker, 2020

Easy examples are learnt early in training, harder examples require memorization later in training.



Estimating Example Difficulty using Variance of Gradients, Agarwal, Souza and Hooker, 2020

A Tale Of Two Long Tails

Daniel D'souza [1 2] **Zach Nussbaum** [1] **Chirag Agarwal** [3] **Sara Hooker** [4]

Abstract

As machine learning models are increasingly employed to assist human decision-makers, it becomes critical to communicate the uncertainty associated with these model predictions. However, the majority of work on uncertainty has focused on traditional probabilistic or ranking approaches – where the model assigns low probabilities or scores to uncertain examples. While this captures what examples are challenging for the model, it does *not* capture the underlying source of the uncertainty. In this work, we seek to identify examples the model is uncertain about *and* characterize the source of said uncertainty. We explore the benefits of designing a targeted intervention – targeted data augmentation of the examples where the model is uncertain over the course of training. We investigate whether the rate of learning in the presence of additional

(a) Reducible Error (b) Irreducible Error

Figure 1. Examples of different predictive uncertainties. **Left:** An instance of the `horse` class representing error reducible using more data examples. **Right:** A `horse` image mislabelled as a `donkey`, representing irreducible error as the model cannot learn this class distribution even with more examples because of the corrupted label.

[[D'souza et al. 2021]]

Convergence rates differ between different types of examples. We can leverage and amplify these differences to distinguish between atypical and noisy examples.

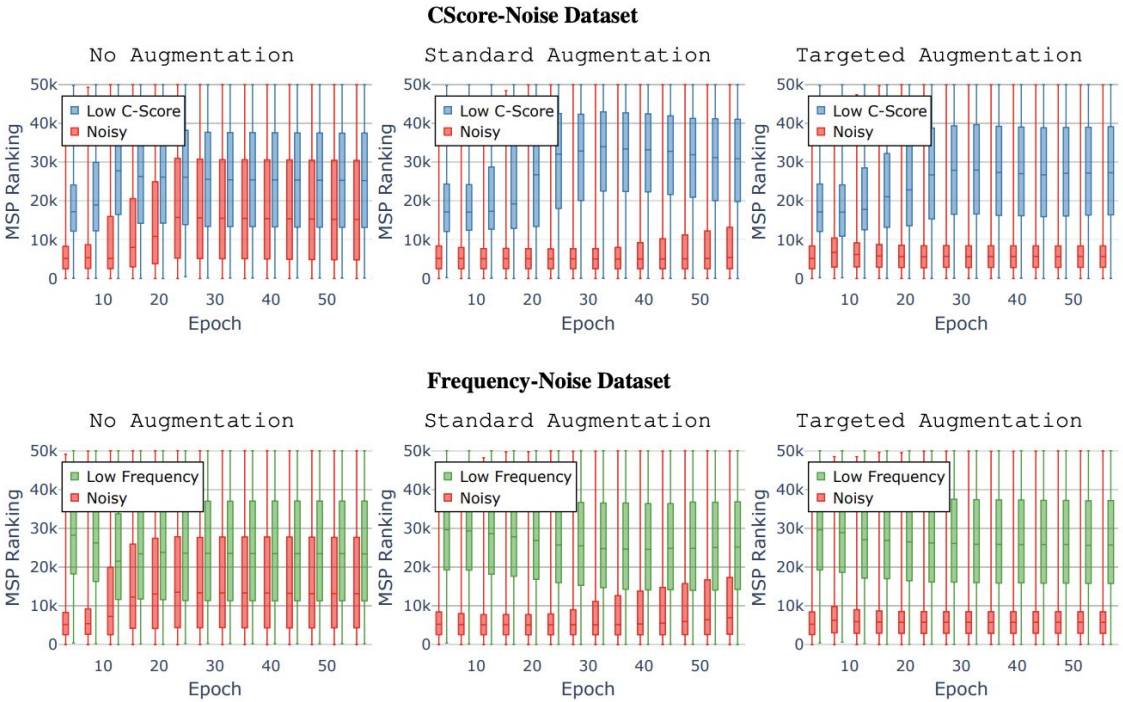
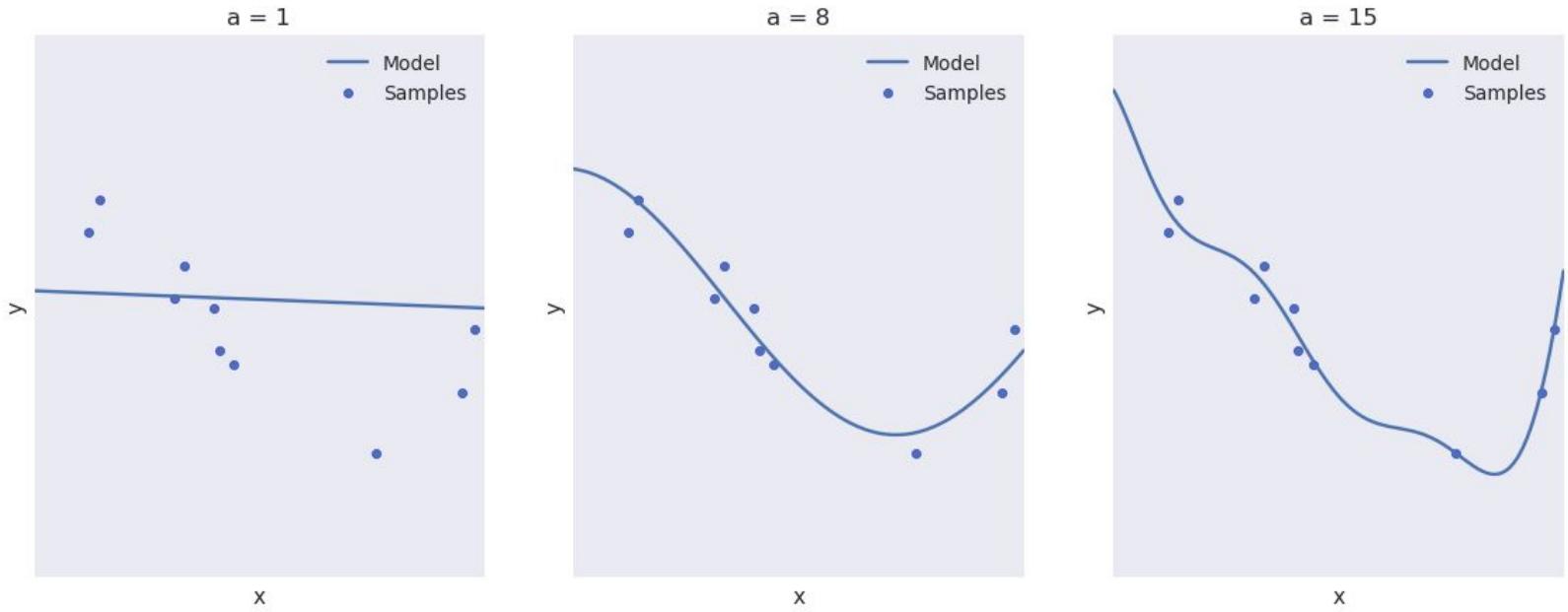


Figure 2. MSP ranking for atypical and noisy subsets in LongTail Cifar-10 dataset across training for different augmentation variants

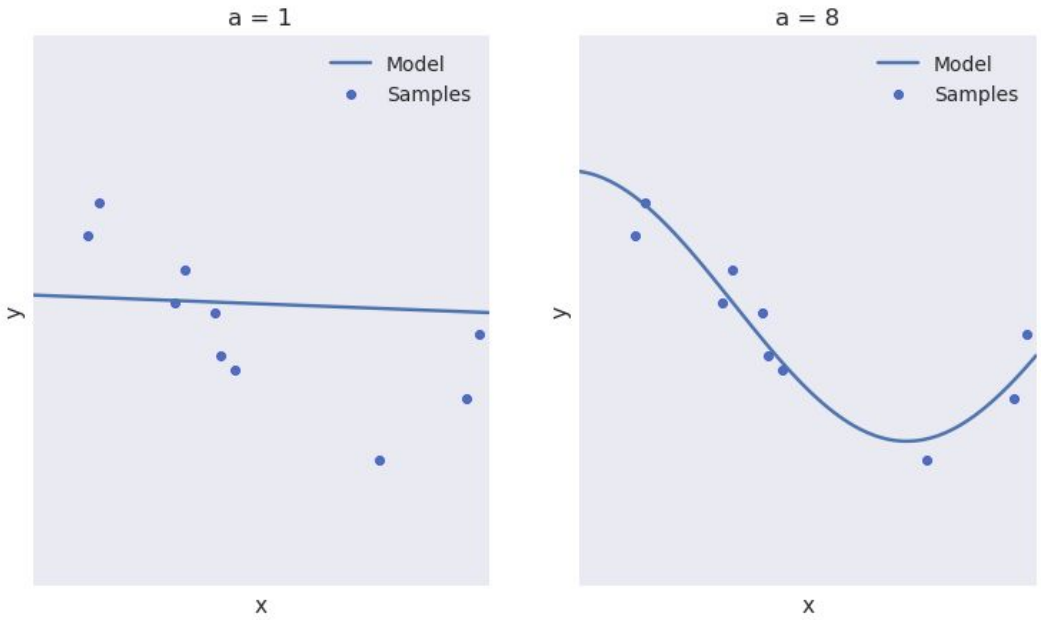
[[A tale of two long tails, D'souza et al. 2021]]

How do system level and algorithm design choices impact model behavior?

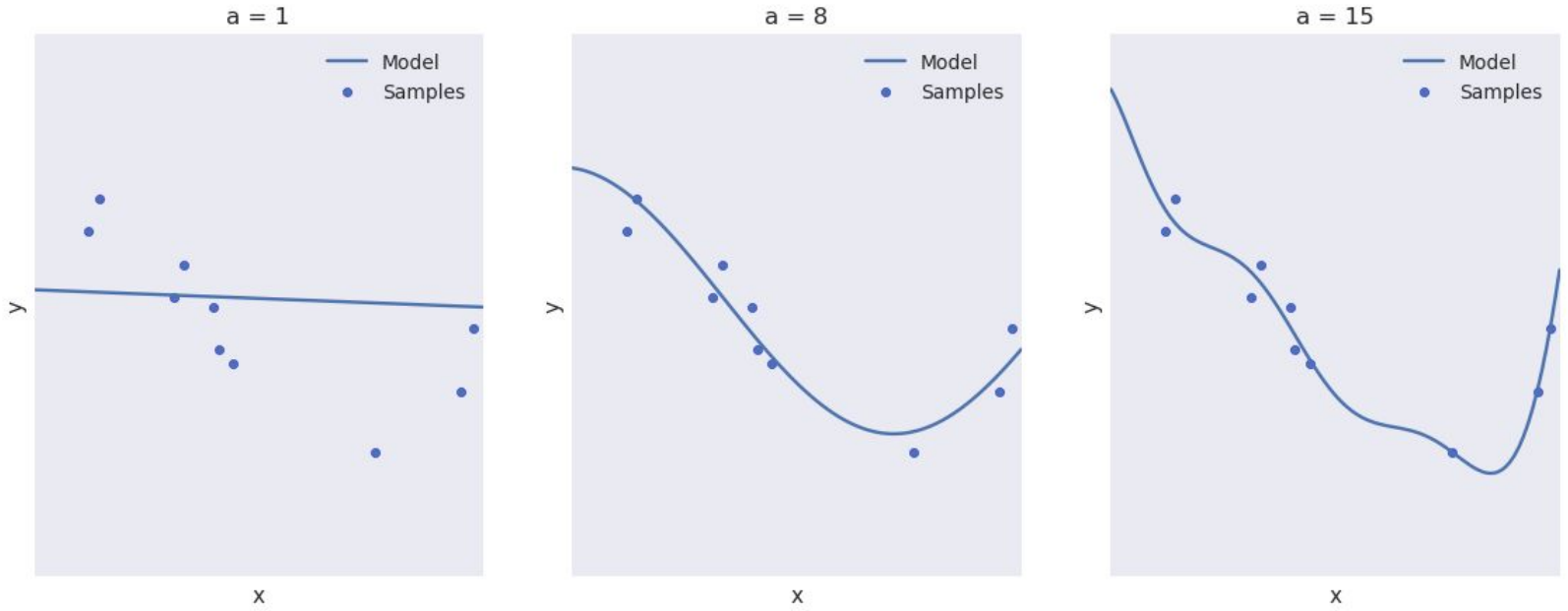
In introduction to machine learning, you are often presented with a linear polynomial function like this.



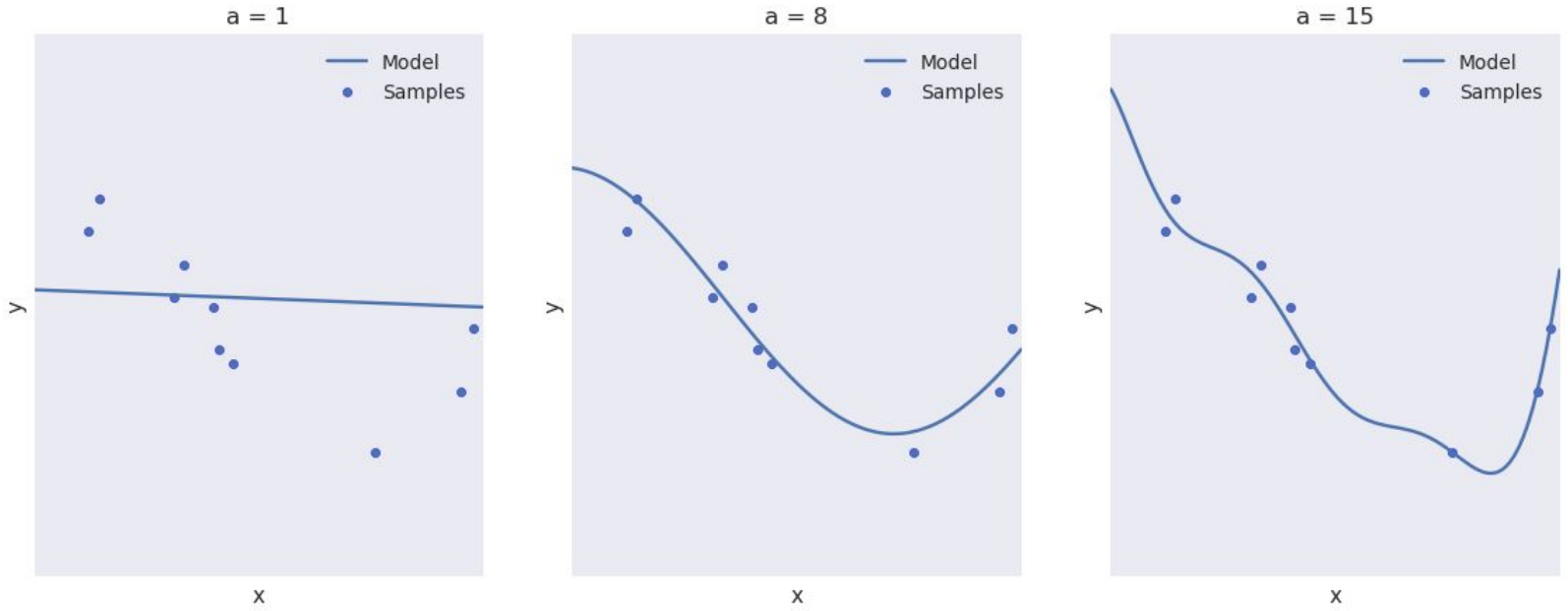
As you increase the degree of the polynomial, you can see how it impacts how the model learns the distribution.



This is one of the first lessons students learn -- the choice of model function matters.

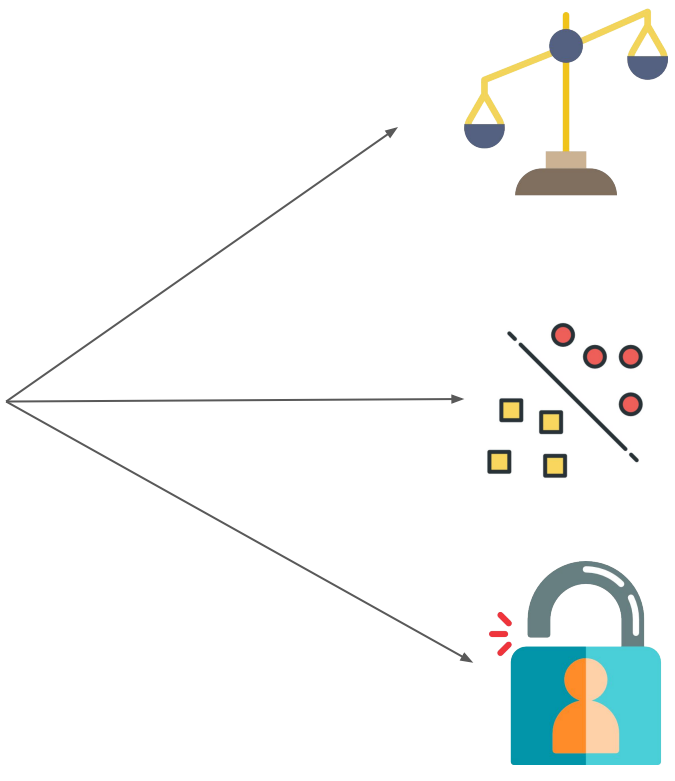


Our modelling choices -- architecture, loss function, optimizer all express a preference for final model behavior.



Often, ML literature makes the unrealistic assumption that optimizing for one property holds all others static.

How we often talk about different properties in the literature.



Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

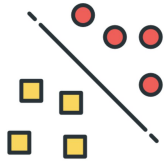
Model Compression - compact machine learning models to work in resource constrained environments.

Model fragility and security - deploy secure models that protect user privacy.

However, our design choices involve trade-offs between objectives.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

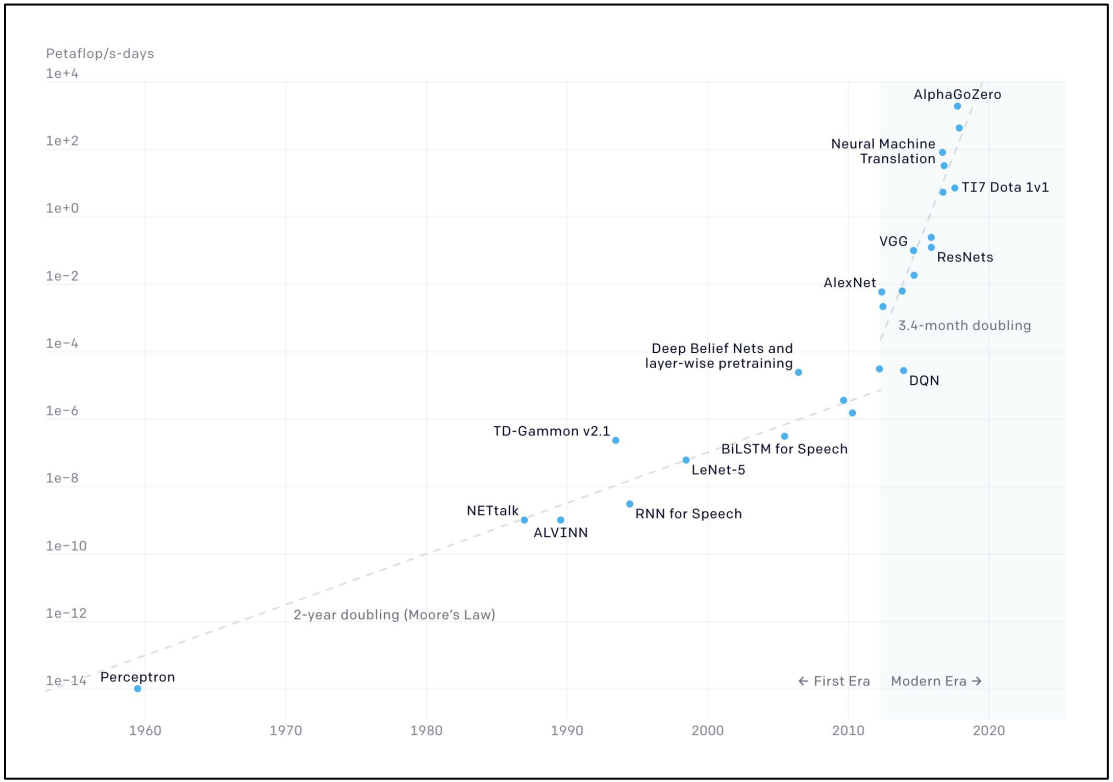
Model fragility and security - deploy secure models that protect user privacy.



Model Compression - compact machine learning models to work in resource constrained environments.

Case Study: How does model compression trade-off against other properties we care about such as robustness and fairness?

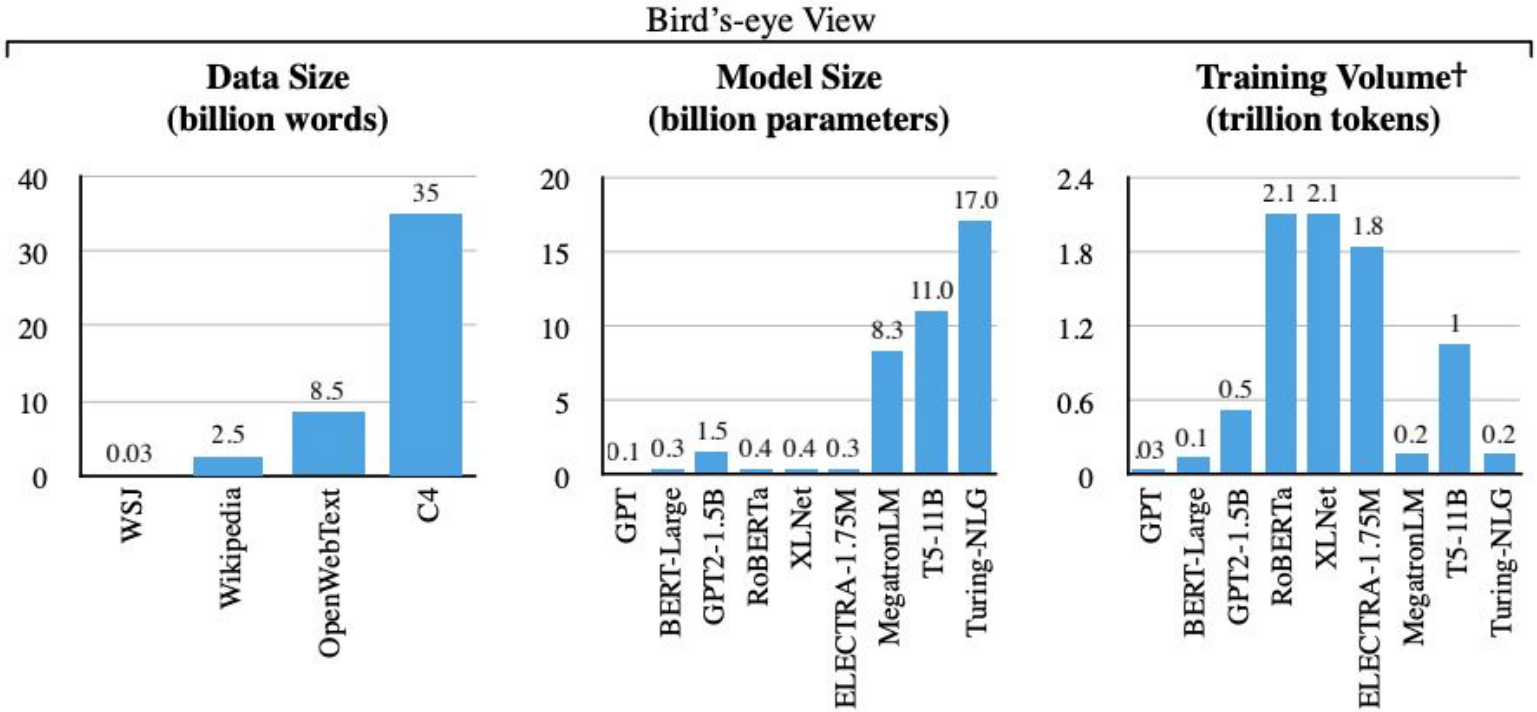
A “bigger is better” race in the number of model parameters has gripped the field of machine learning.



Google

Canziani et al., 2016, Open AI 2019

This characterizes both vision and NLP tasks.



Link here [Sharir et al. 2020]

An argument in favor of this approach:

- Different regimes of capacity appear to allow for different generalization properties.
- It is very simple formula (throw more parameters at the model)

[Nakkiran et al. 2019, Petroni et al., Brown et al.,] [Adam et al.]

A key limitation of this approach:

Relationship between weights and generalization properties is not well understood.

The intriguing relationship between capacity and generalization.

Why do we need so many weights in the first place?

1) Diminishing returns to adding more weights.
2) Many redundancies between weights
3) We can remove most weights after training.

Diminishing returns to adding parameters. Millions of parameters are needed to **eek** out additional gains.

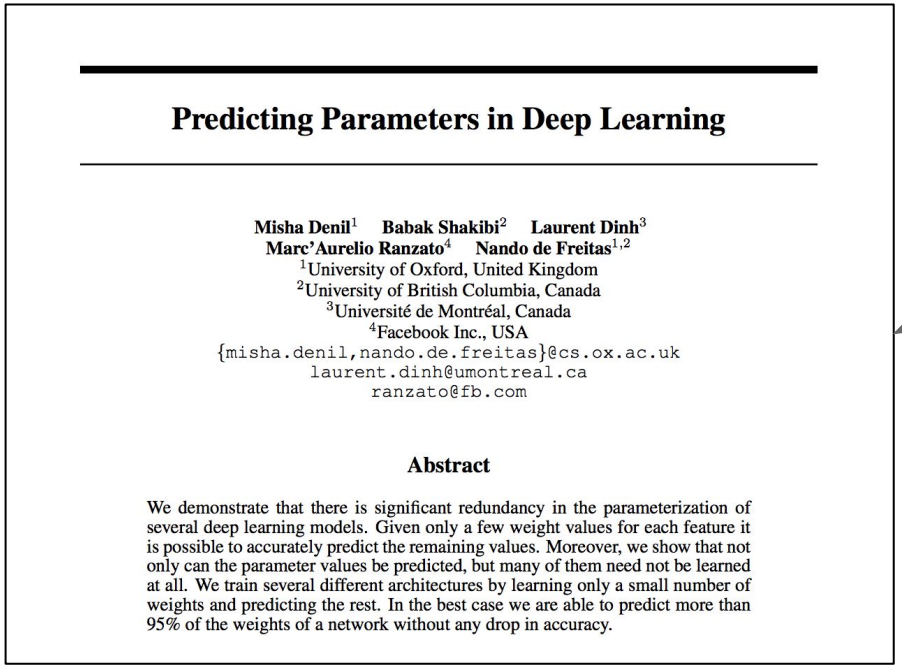
Model	Parameters[a]	Features	Image Size	ImageNet Top-1 Accurac	
				Paper	Public Checkpoint[b]
Inception v1[c] [69]	5.6M	1024	224	73.2	69.8
BN-Inception[d] [34]	10.2M	1024	224	74.8	74.0
Inception v3 [70]	21.8M	2048	299	78.8	78.0
Inception v4 [68]	41.1M	1536	299	80.0	80.2
Inception-ResNet v2 [68]	54.3M	1536	299	80.1	80.4
ResNet-50 v1[e] [29, 26, 25]	23.5M	2048	224	76.4	75.2
ResNet-101 v1 [29, 26, 25]	42.5M	2048	224	77.9	76.4
ResNet-152 v1 [29, 26, 25]	58.1M	2048	224	N/A	76.8
DenseNet-121 [31]	7.0M	1024	224	75.0	74.8
DenseNet-169 [31]	12.5M	1024	224	76.2	76.2
DenseNet-201 [31]	18.1M	1024	224	77.4	77.3
MobileNet v1 [30]	3.2M	1024	224	70.6	70.7
MobileNet v2 [61]	2.2M	1280	224	72.0	71.8
MobileNet v2 (1.4) [61]	4.3M	1792	224	74.7	75.0
NASNet-A Mobile [84]	4.2M	1056	224	74.0	74.0
NASNet-A Large [84]	84.7M	4032	331	82.7	82.7

Almost double the amount of weights for a gain in 2% points.

Google

Table: Kornblith et al., 2018 [Kaplan + 2020]

Redundancies Between Weights



Predicting Parameters in Deep Learning

Misha Denil[1] **Babak Shakibi**[2] **Laurent Dinh**[3]
Marc'Aurelio Ranzato[4] **Nando de Freitas**[1,2]
[1]University of Oxford, United Kingdom
[2]University of British Columbia, Canada
[3]Université de Montréal, Canada
[4]Facebook Inc., USA
{misha.denil,nando.de.freitas}@cs.ox.ac.uk
laurent.dinh@umontreal.ca
ranzato@fb.com

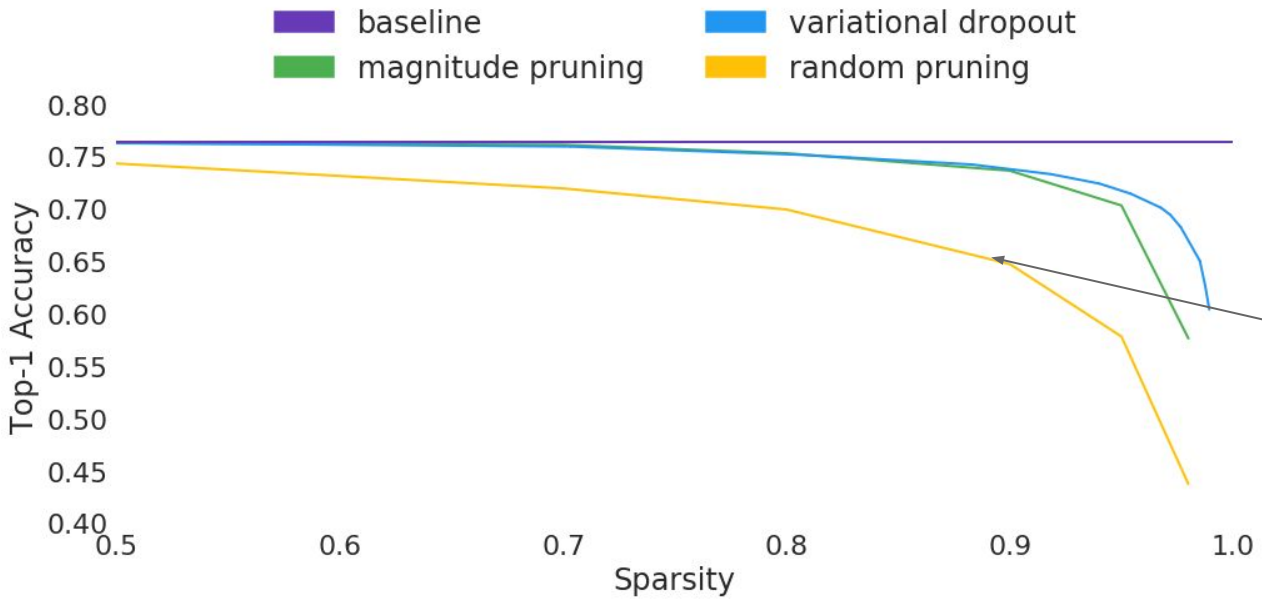
Abstract

We demonstrate that there is significant redundancy in the parameterization of several deep learning models. Given only a few weight values for each feature it is possible to accurately predict the remaining values. Moreover, we show that not only can the parameter values be predicted, but many of them need not be learned at all. We train several different architectures by learning only a small number of weights and predicting the rest. In the best case we are able to predict more than 95% of the weights of a network without any drop in accuracy.

Denil et al. find that a small set of weights can be used to predict 95% of weights in the network.

Google [[Denil et al., 2014]]

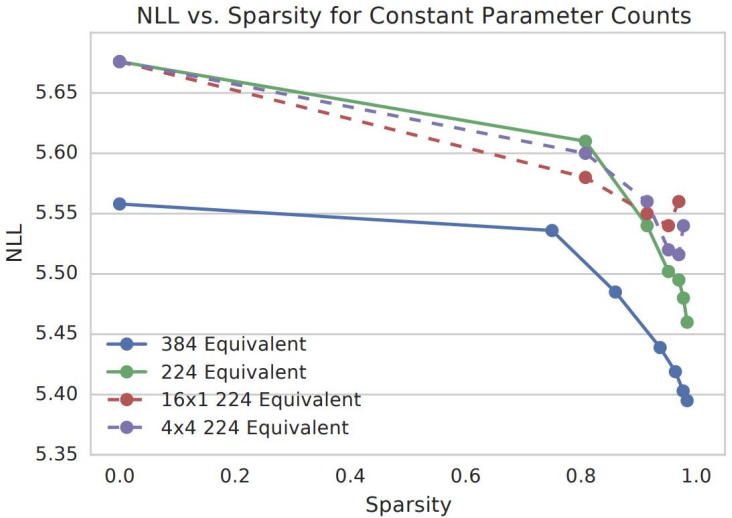
Most weights can be removed after training is finished **(while only losing a few % in test-set accuracy!)**



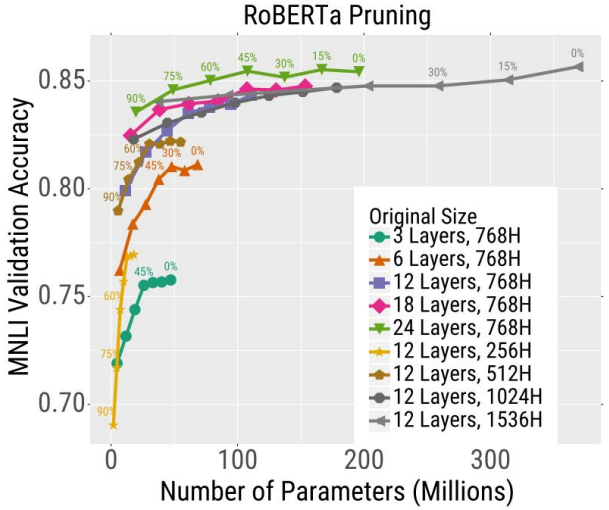
With 90% of the weights removed, a ResNet-50 only loses ~3% of performance (for certain pruning methods).

[[The State of Sparsity in Deep Neural Networks, 2019, Gale, Elsen, Hooker]]

Sparse models **easily** outcompete dense models with same parameter count.



[Efficient Neural Audio Synthesis](), Kalchbrenner et al., 2018



[Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers](), Li et al., 2020

Understanding how capacity impacts generalization is an increasingly urgent question:

How do generalization properties change as models get bigger and bigger?

- Fairness, robustness, privacy.

Increasingly, we are also making design choices at test-time that alter generalization properties - pruning, quantization, fine-tuning.

Why is this interesting? **Theoretical** reasons:

We are in an era of ever bigger models.

Yet, there is a limit to how much we can scale. Understanding the role of capacity can guide us to more efficient solutions.

If most weights are redundant, why do we need them in the first place?

Can these insights guide us to better training protocols?

Why is this interesting? **Practical** reasons:

Most of the world uses ML in a resource constrained environment.

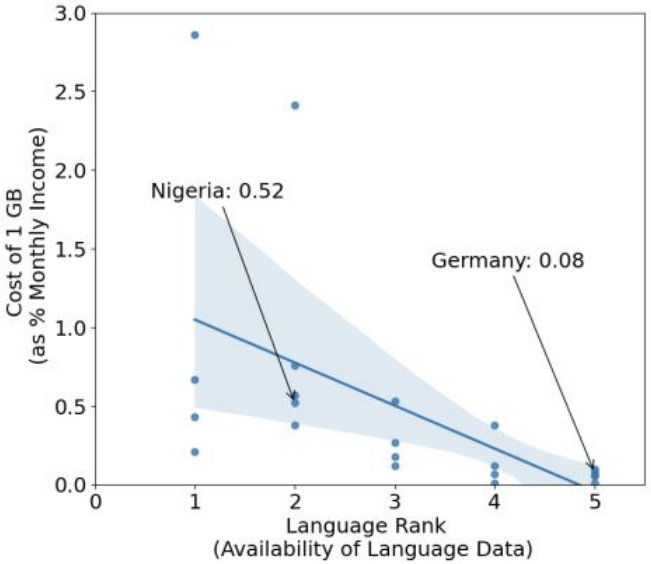


Figure 1: Cost of mobile data by country per language rank according to the taxonomy by Joshi et al. (2020).

Google The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation, link

If we care about access to technology, we need to revisit our model design assumptions:

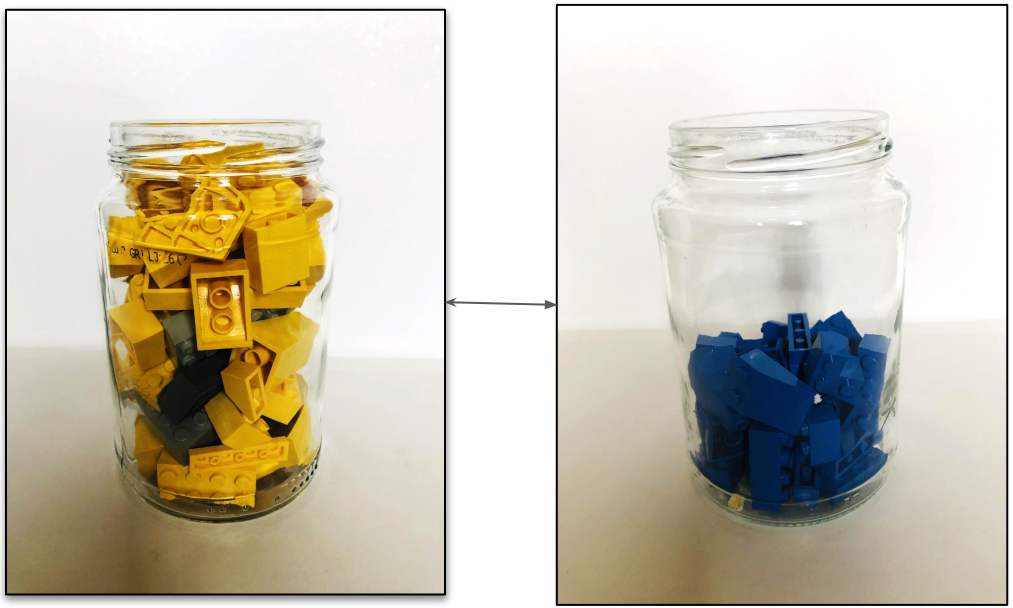
As you increase size of networks:

- More memory to store
- Higher latency for each forward pass in training + inference time

ML at the edge:

- Many different devices, hardware constraints
- Many different resource constraints - memory, compute
- Power, connectivity varies

How can networks with radically different structures and number of parameters have comparable performance?



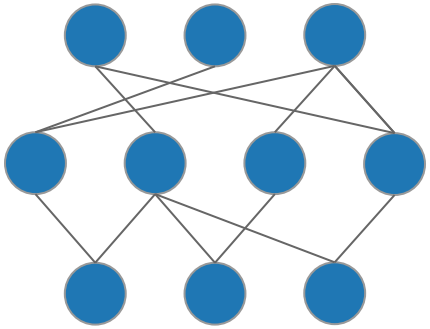
0% pruning
76.70%

50% pruning
76.20%

One possibility is that top-line metrics are not a precise enough measure to capture how capacity impacts the generalization properties of the model.

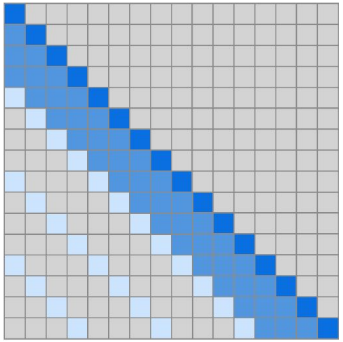
To gain intuition into differences in generalization behavior, we go beyond topline metrics.

Sparsity in Deep Learning



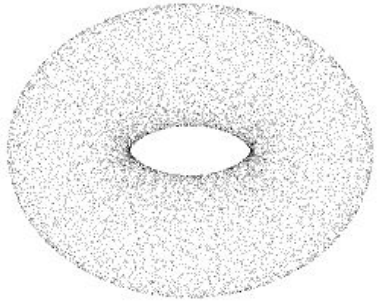
Weight Sparsity

Sources: Pruning, sparse training
Example: 3-5x FLOP advantage for the same accuracy in CNNs [1][2]



Activation Sparsity

Sources: ReLU sparsity, sparse attention
Example: Asymptotic improvement in attention computational complexity. From $O(N^2)$ to $O(N*sqrt(N))$ to $O(N)$ [3][4]

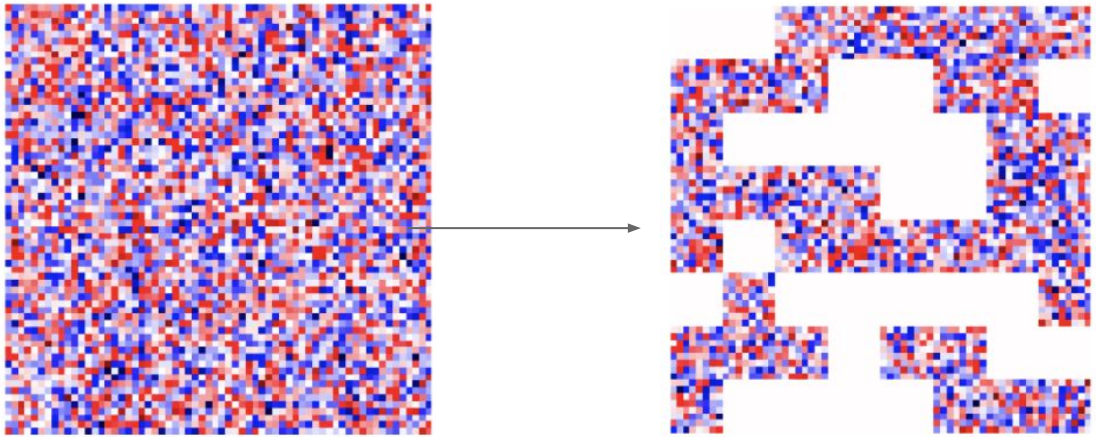


Data Sparsity

Sources: Point clouds, graphs, etc.
Example: 3D object detection with targeted computation [5]

Image sources:
https://bit.ly/3a05oyu

Sparsify networks by “removing” unimportant activations/weights (setting weights/neurons to zero).



Initial weight matrix

After activations have been removed.

Google

Image source: OpenAI

Instead of starting sparse -- most state of art sparsity methods introduce sparsity gradually over the course of training.

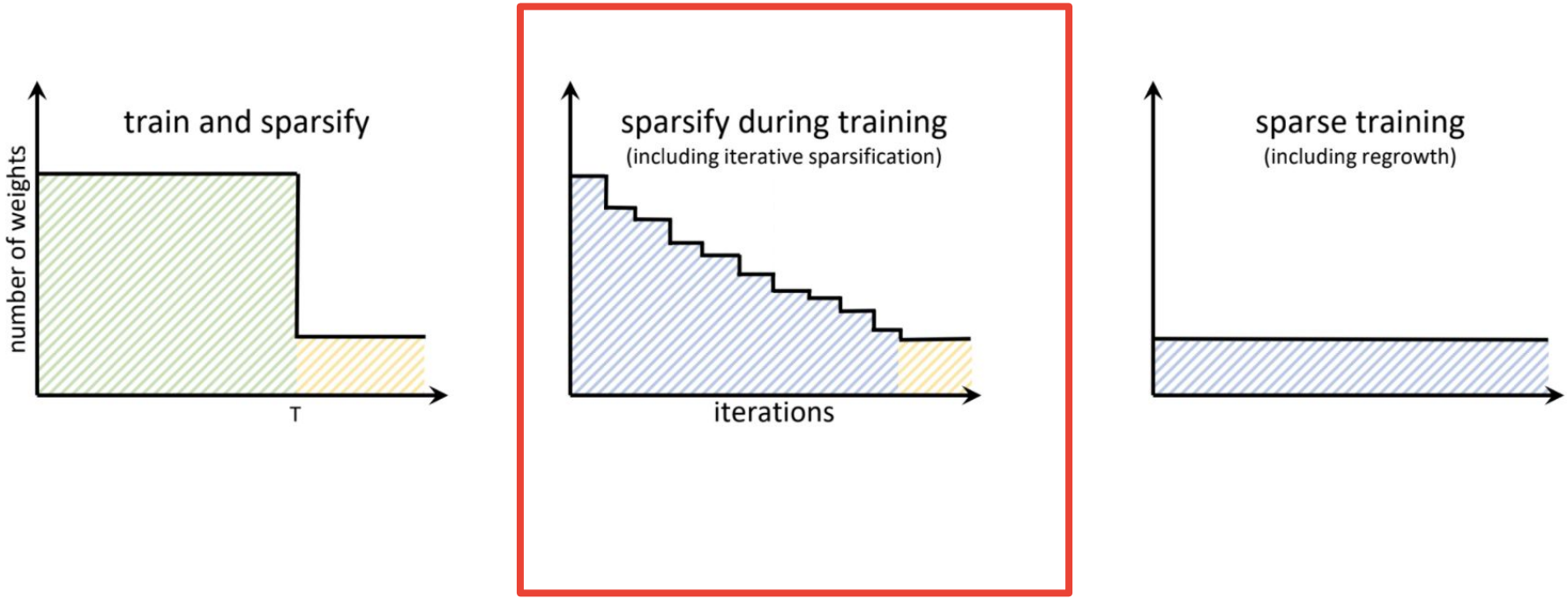
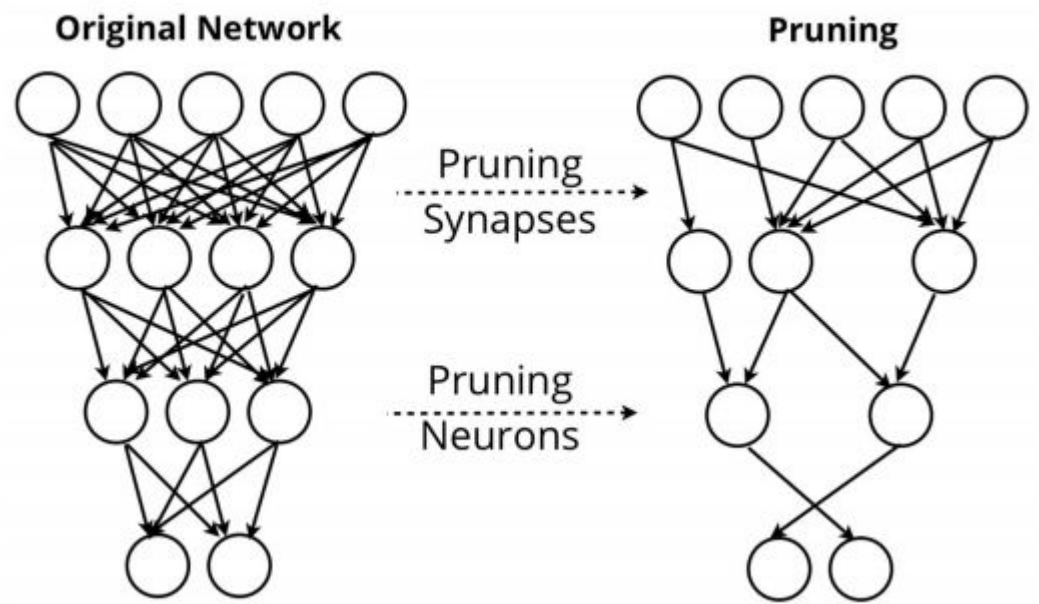


Image from Torsten Hoefler tutorial

Sparsity of 90% means that by the end of training the model only has 10% of all weights remaining. Apply mask of 0 to remaining weights.



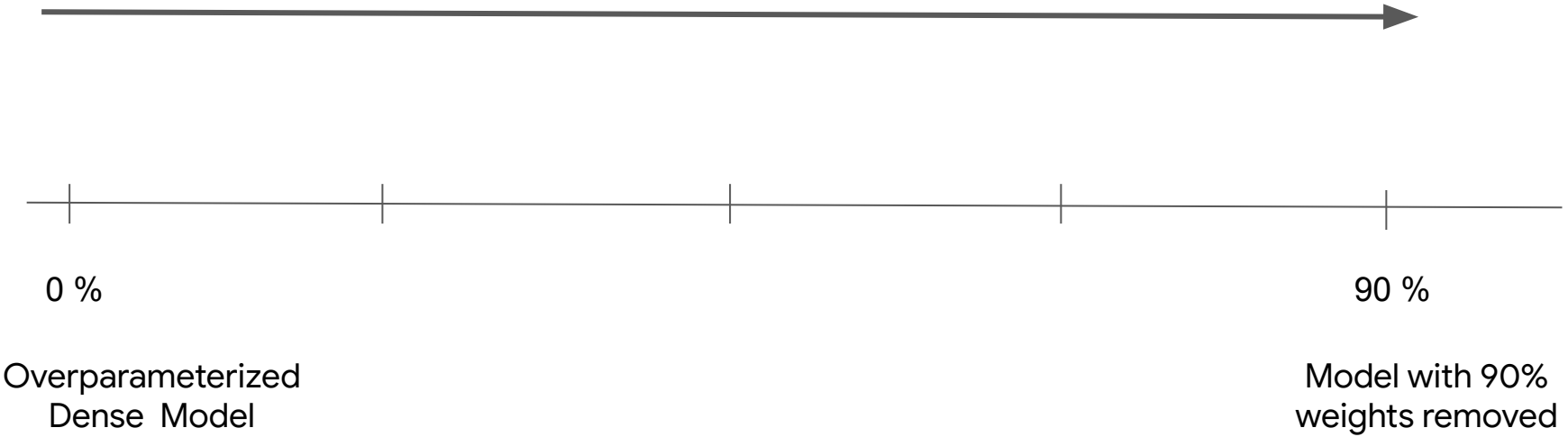
Initial weight matrix

After activations have been removed.

Image source

Valuable experimental set-up - we can precisely vary the level of final sparsity.

Train populations of models with minimal differences in test-set accuracy to different end sparsities [0%, 30%, 50%, 70%, 90%, **95%, 99%**].



Here, we ask - How does model behavior diverge as we vary the level of compression?

1. Robustness to certain types of distribution shift.

2. Measure divergence in class level and exemplar classification performance.

Key results upfront: top level metrics hide critical differences in generalization between compressed and compressed populations of models.

1.

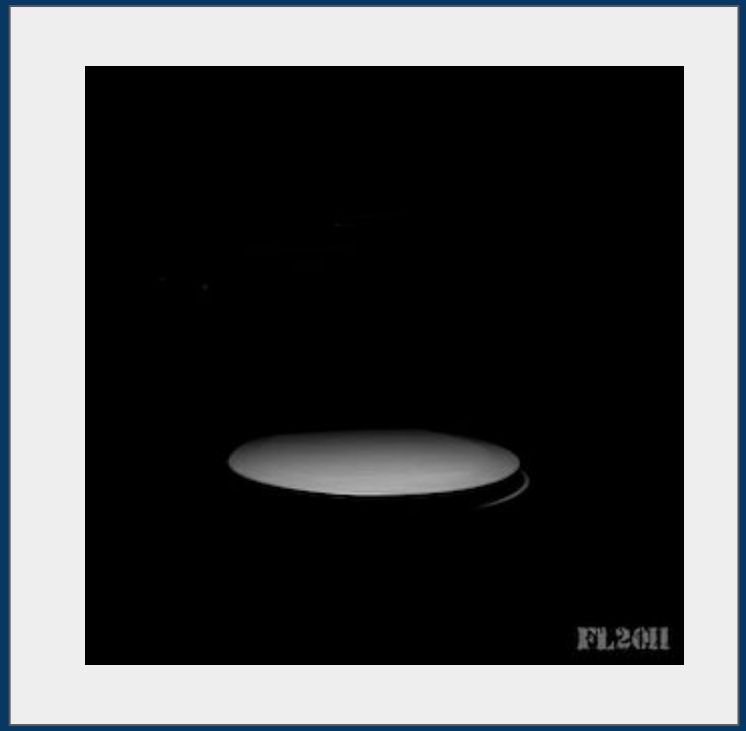
Varying sparsity disproportionately and systematically impact a small subset of classes and exemplars.

→

Why is a narrow part of the data distribution far more sensitive to varying capacity?

Pruning Identified Exemplars (PIEs)

are images where predictive behavior diverges between a population of independently trained compressed and non-compressed models.



ImageNet test-set. True label?

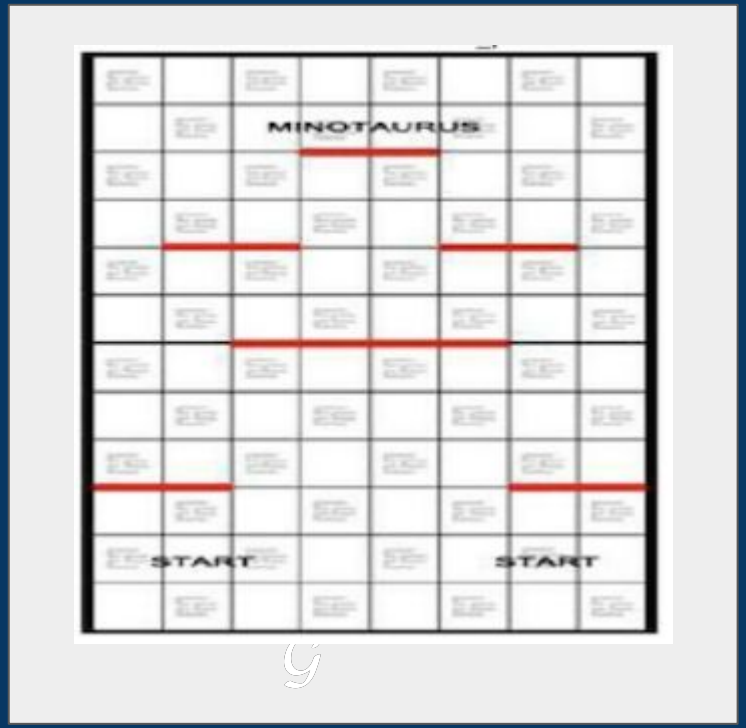

toilet seat
Non-PIE
PIE

ImageNet test-set. True label?

espresso

Non-PIE

PIE



ImageNet test-set. True label?

maze

Non-PIE

PIE

ImageNet test-set. True label?

wool

Non-PIE

PIE

ImageNet test-set.
True label?

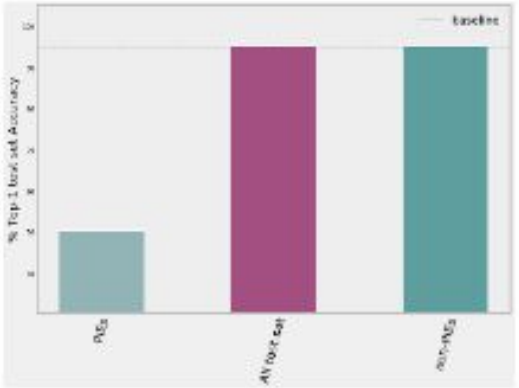
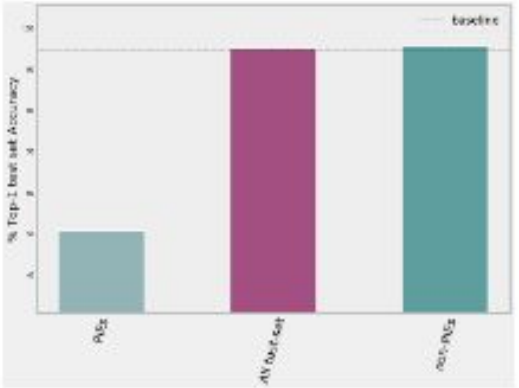
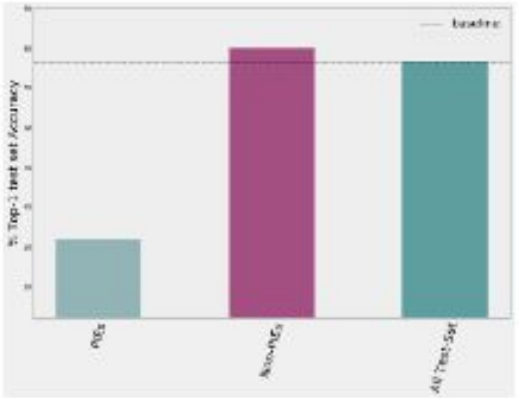
matchstick

Non-PIE

PIE

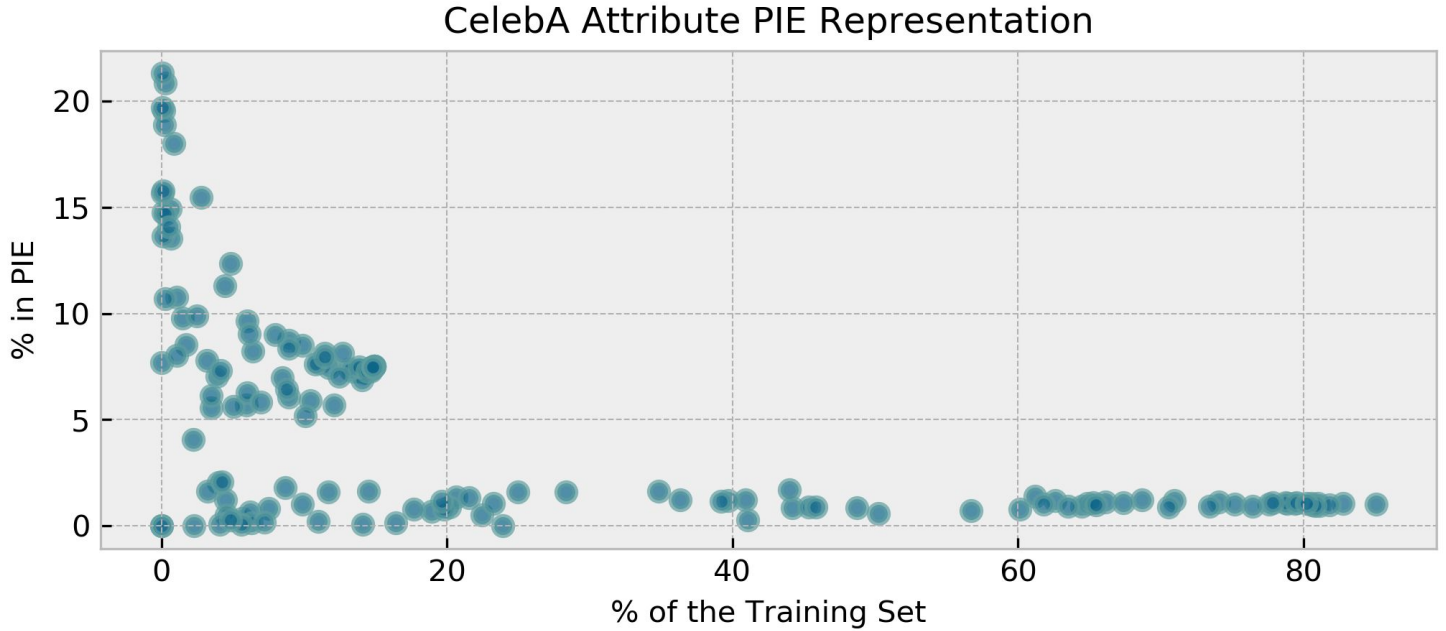
PIEs are also more challenging for algorithms to classify.

Top-1 Accuracy on PIE, All Test-Set, Non-PIE

CelebA	CIFAR-10	ImageNet
		

- Restricting inference to PIEs drastically degrades model performance.
- For ImageNet, removing PIEs from test-set improves top-1 accuracy beyond baseline.

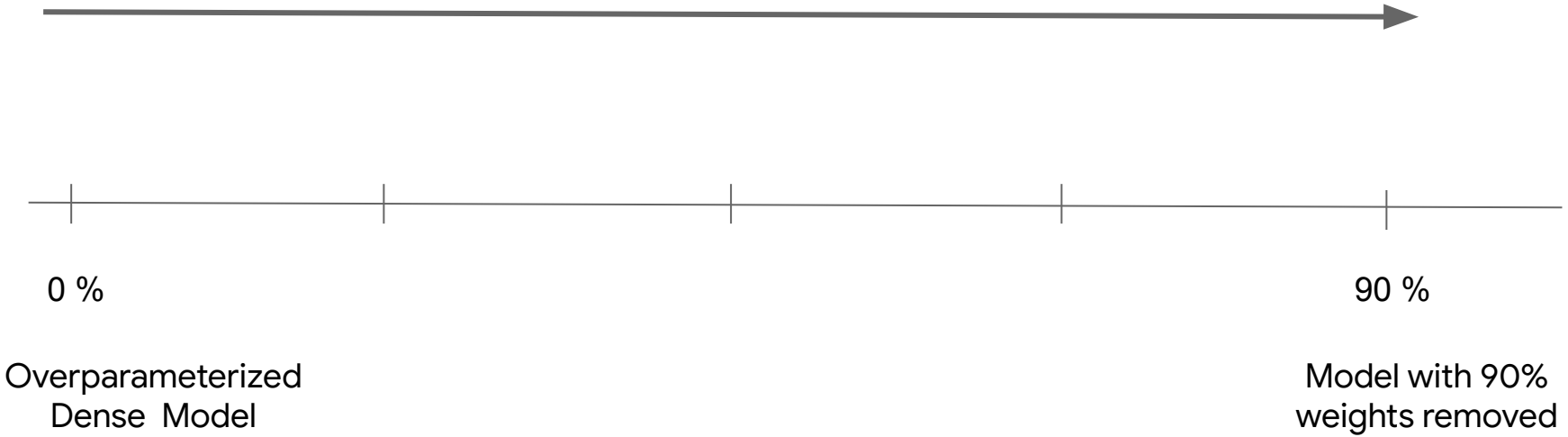
PIEs over-index on the long-tail of underrepresented attributes.



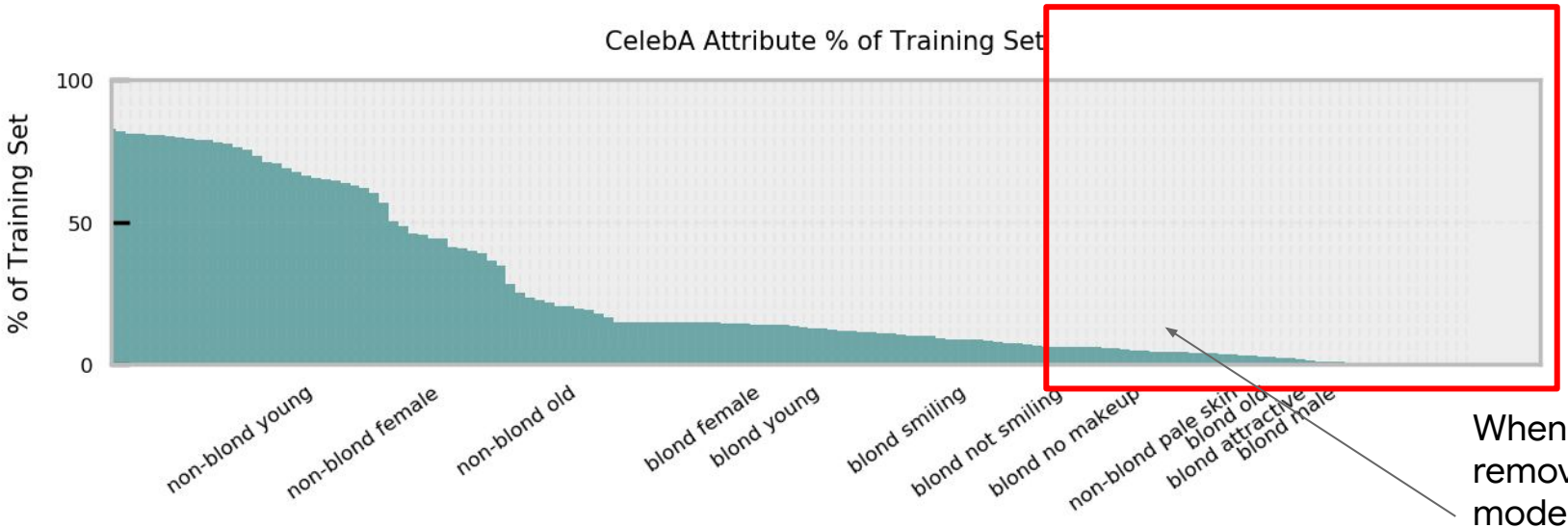
Attribute Proportion of CelebA Training Data vs. relative representation in PIE

We lose the long-tail when we remove the majority of all training weights.

Put differently, we are using the majority of our weights to encode a useful representation for a small fraction of our training distribution.



Low-frequency events - The majority of weights **(90% of all weights)** are used to memorize very rare examples in the dataset.



When we remove weights models lose performance on rare examples.

[[Hooker et al. 2019, Hooker, Moorosi et al, 2020]]

Noisy Data Points

- Data is improperly structured which corrupts information
  - Mislabelled
  - Severely corrupted
  - Multi-object

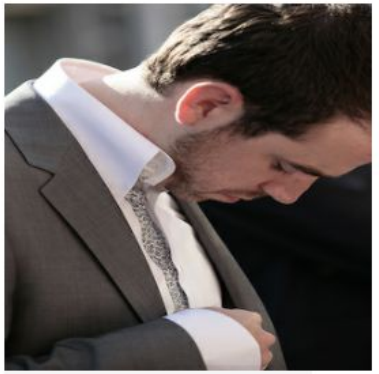
Misuse of parameters to represent these data points.

“Bad memorization”

Noisy PIEs: Incorrectly structured ImageNet data for single-image classification.

True Label: parallel bars
Non-Pruned: parallel bars
Pruned: horizontal bars

True Label: corn
Non-Pruned: corn
Pruned: ear (of corn)



True Label: groom
Non-Pruned: groom
Pruned: suit

Noisy PIEs: Corrupted or incorrectly labeled data.

True Label:
restaurant
Non-Pruned:
meat loaf
Pruned: guacamole

True Label:
envelope
Non-Pruned:
dumbbell
Pruned: maraca

True Label:
tub
Non-Pruned:
cauldron
Pruned: wok

Noisy Data Points

- Data is improperly structured which corrupts information
  - Mislabelled
  - Severely corrupted
  - Multi-object

Misuse of parameters to represent these data points.

"Bad memorization"

Atypical Data Points or Challenging Exemplars

- Underrepresented vantage points (the long-tail of the dataset)
- Image classification entails fine grained task

Valuable use of parameters to represent these data points.

"Good memorization"

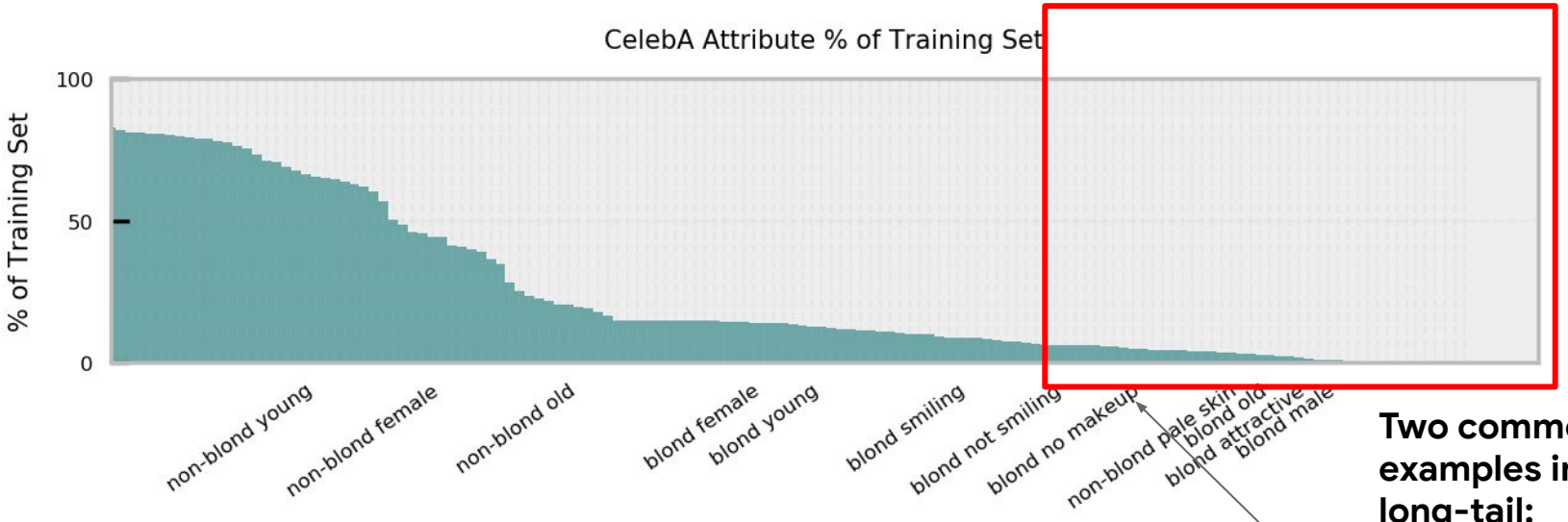
Atypical PIEs: Unusual vantage points of the class category.

True Label: toilet seat
Non-Pruned: toilet seat
Pruned: folding chair

True Label: bathtub
Non-Pruned: bathtub
Pruned: cucumber

True Label: plastic bag
Non-Pruned: gown
Pruned: plastic bag

Whether pruning aids or impedes performance depends upon how relevant learning rare artefacts are for the task.



Two common types of examples in the long-tail:

- **Noisy**
- **Atypical**

Google

[[Ahia et al. 2021, Hooker et al. 2019, Hooker, Moorosi et al. 2020]]

This is very related to whether memorization aids or hurts generalization.

In-distribution considerations:

Noisy Data Points

- Data is improperly structured which corrupts information.

Misuse of parameters to represent these data points.

"Bad memorization"

Atypical Data Points or Challenging Exemplars

- Underrepresented vantage points (the long-tail of the dataset)

Valuable use of parameters to represent these data points.

"Good memorization"

This is very related to whether memorization aids or hurts generalization.

Out-of-distribution considerations:

Test data is very different from my training distribution (typical of low data regimes)

Memorization of rare artefacts in the training data is unlikely to help with generalization to the test data.

"Bad memorization"

Test dataset is very similar to my training distribution (more common with VERY big data regimes)

Memorization of rare artefacts in the training data is likely to help with generalization to the test data.

"Good memorization"

In-distribution considerations - amplification of error on rare/underrepresented attributes.

CHARACTERISING BIAS IN COMPRESSED MODELS

Sara Hooker *
Google Research
shooker@google.com

Nyalleng Moorosi *
Google Research
nyalleng@google.com

Gregory Clark
Google
gregoryclark@google.com

Samy Bengio
Google Research
bengio@google.com

Emily Denton
Google Research
dentone@google.com

ABSTRACT

The popularity and widespread use of pruning and quantization is driven by the severe resource constraints of deploying deep neural networks to environments with strict latency, memory and energy requirements. These techniques achieve high levels of compression with negligible impact on top-line metrics (top-1 and top-5 accuracy). However, overall accuracy hides disproportionately high errors on a small subset of examples; we call this subset Compression Identified Exemplars (*CIE*). We further establish that for *CIE* examples, compression amplifies existing algorithmic bias. Pruning disproportionately impacts performance on underrepresented features, which often coincides with considerations of fairness. Given that *CIE* is a relatively small subset but a great contributor of error in the model, we propose its use as a human-in-the-loop auditing tool to surface a tractable subset of the dataset for further inspection or annotation by a domain expert. We provide qualitative and quantitative support that *CIE* surfaces the most challenging examples in the data distribution for human-in-the-loop auditing.

How does compression impact performance on the long-tail in-distribution?

In-distribution considerations:

Noisy Data Points

- Data is improperly structured which corrupts information.

Misuse of parameters to represent these data points.

"Bad memorization"

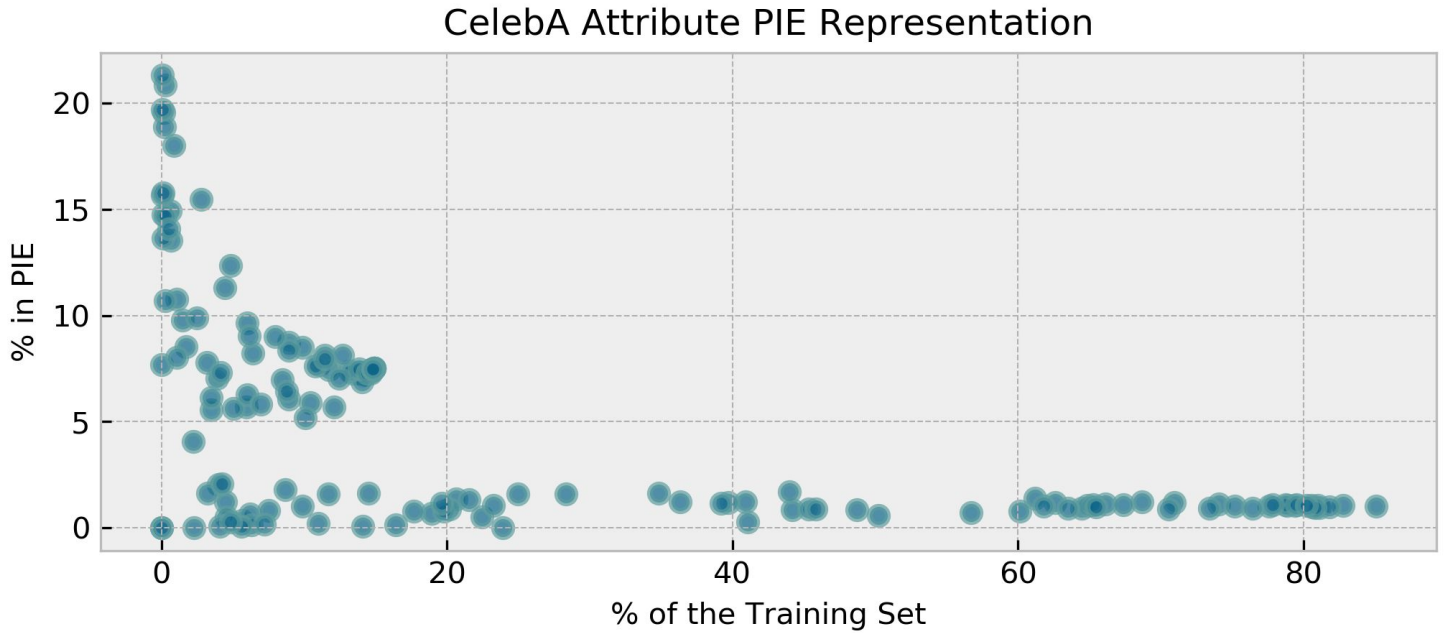
Atypical Data Points or Challenging Exemplars

- Underrepresented vantage points (the long-tail of the dataset)

Valuable use of parameters to represent these data points.

"Good memorization"

Compression amplifies algorithmic harm when the protected feature is in the long-tail of the distribution.

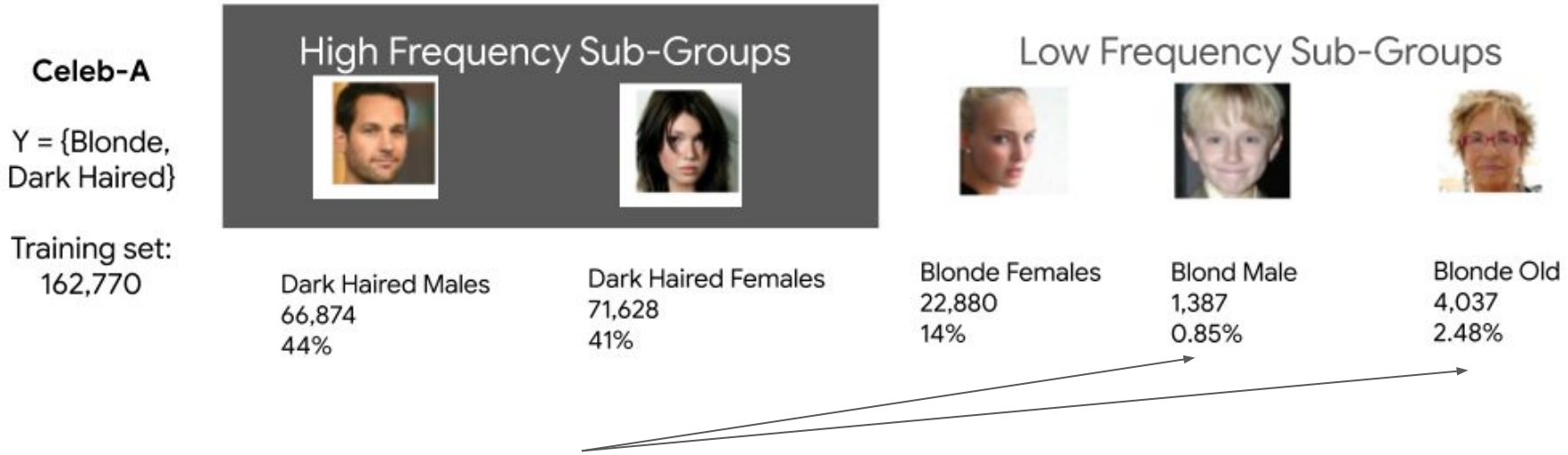


Attribute Proportion of CelebA Training Data vs. relative representation in PIE

[[Hooker et al. 2019, Hooker, Moorosi et al, 2020]]

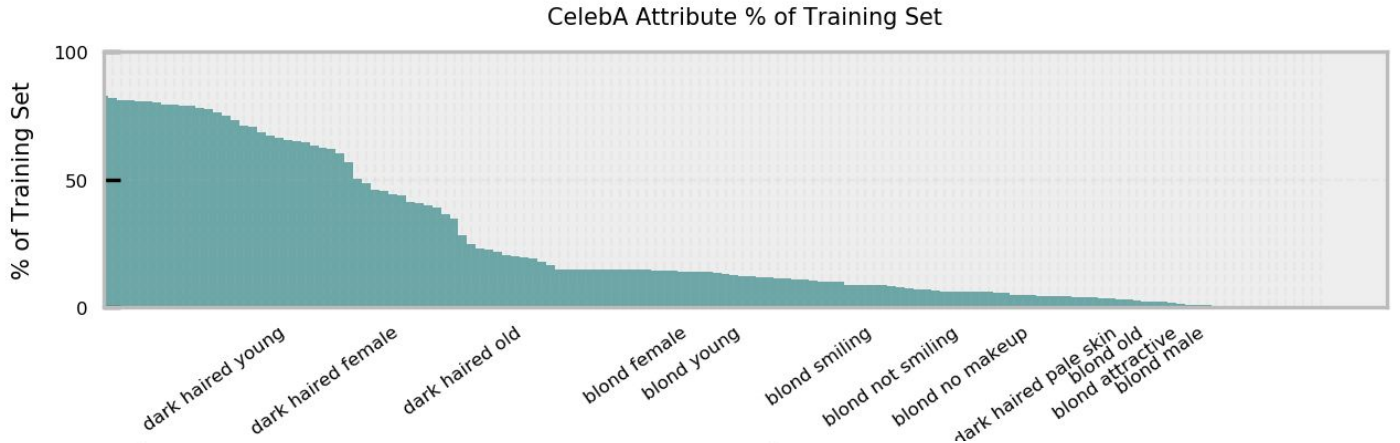
Measuring Impact of Compression on Algorithmic Bias

Celeb-A Spurious correlation between gender, age and hair color {Blond, Non-Blond}



Far fewer examples of ‘Blond Male’ (0.85%) and ‘Blond Old’ (2.43%) of training set.

We find sparsity disproportionately impacts underrepresented features.



Celeb-A

Y = {Blonde, Dark Haired}

Training set: 162,770

High Frequency Sub-Groups

Dark Haired Males
66,874
44%

Dark Haired Females
71,628
41%

Low Frequency Sub-Groups

Blonde Females
22,880
14%

Blond Male
1,387
0.85%

Blonde Old
4,037
2.48%

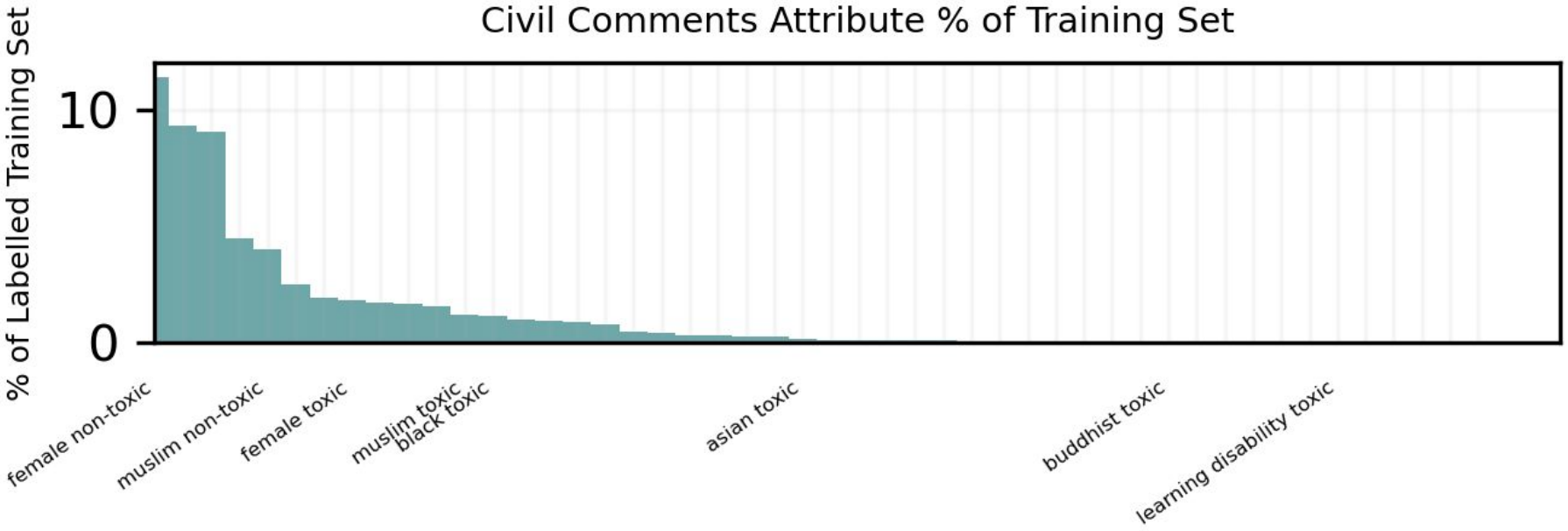
Pruning amplifies algorithmic bias when the underrepresented feature is sensitive (age/gender)

Model	Metric	Aggregate	Unitary				Intersectional			
			M	F	Y	O	MY	MO	FY	FO
Baseline (0% pruning)	Error	5.30%	2.37%	7.15%	5.17%	5.73%	2.28%	2.50%	5.17%	5.73%
	FPR	2.73%	0.93%	4.12%	2.59%	3.18%	0.81%	1.12%	2.59%	3.18%
	FNR	22.03%	62.65%	19.09%	21.35%	24.47%	60.45%	66.87%	21.35%	24.47%
Normalized Difference Between **1)** Compressed and **2)** Non-Compressed Baseline										
Compressed (95% pruning)	Error	24.63%	24.49%	24.67%	20.64%	35.84%	7.96%	49.12%	20.64%	35.84%
	FPR	12.72%	49.54%	6.32%	3.35%	36.02%	5.37%	101.88%	3.35%	36.02%
	FNR	34.22%	8.41%	40.30%	33.83%	35.39%	9.21%	6.98%	33.83%	35.39%

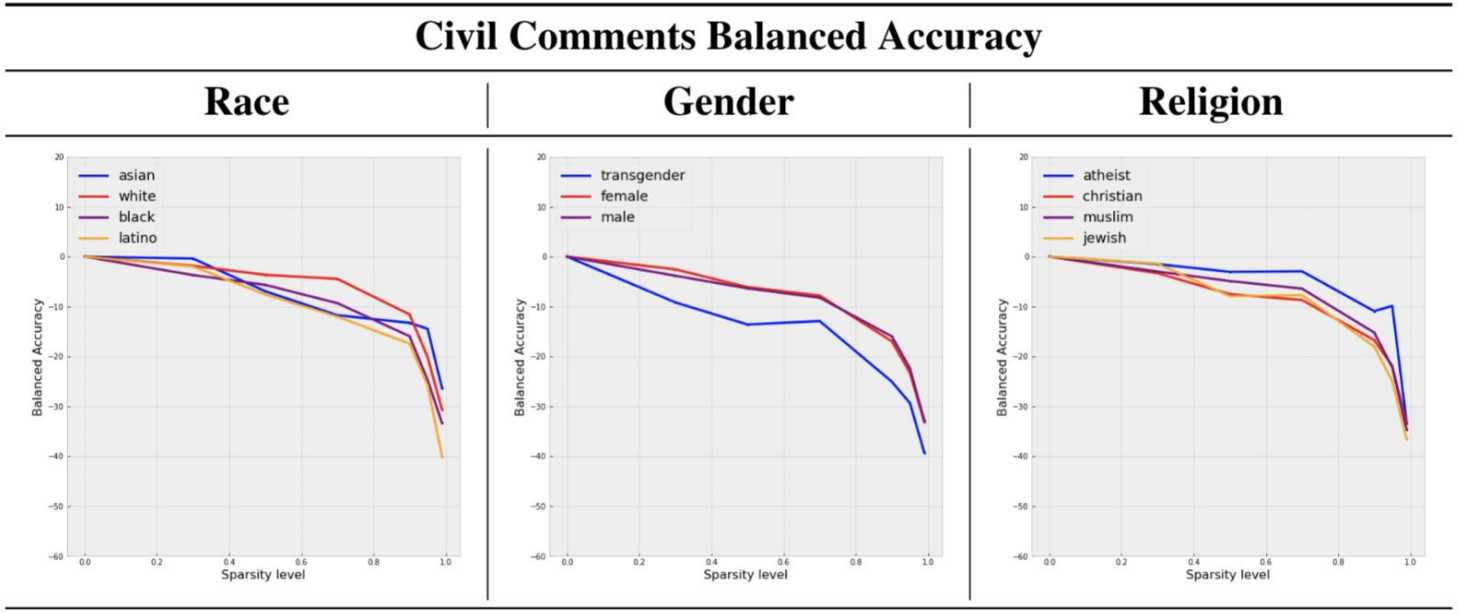
Table 3: Performance metrics disaggregated across `Male` (M), `not Male` (F), `Young` (Y), and `not Young` (O) sub-groups. For all error rates reported, we average performance over 10 models. **Top Row**: Baseline error rates, **Bottom Row:** Relative change in error rate between baseline models and models pruned to 95% sparsity,

[[Hooker et al. 2019, Hooker, Moorosi et al, 2020]]

Civil Comments Task of detecting toxic comments. Target label toxic is only present for ~8% of training set.



Sparsity sharply degrades model ability to detect toxic comments. Most impacted sub-groups are least represented in training set.



[[Hooker et al. 2019, Hooker, Moorosi et al, 2020]]

The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation

Orevaoghene Ahia
Masakhane NLP
oreva.ahia@gmail.com

Julia Kreutzer
Google Research
Masakhane NLP
jkreutzer@google.com

Sara Hooker
Google Research, Brain
shooker@google.com

Abstract

A "bigger is better" explosion in the number of parameters in deep neural networks has made it increasingly challenging to make state-of-the-art networks accessible in compute-restricted environments. Compression techniques have taken on renewed importance as a way to bridge the gap. However, evaluation of the trade-offs incurred by popular compression techniques has been centered on high-resource datasets. In this work, we instead consider the impact of compression in a data-limited regime. We introduce the term *low-resource double bind* to refer to the co-occurrence of data limitations and compute resource constraints. This is a common setting for NLP for low-resource languages, yet the trade-offs in

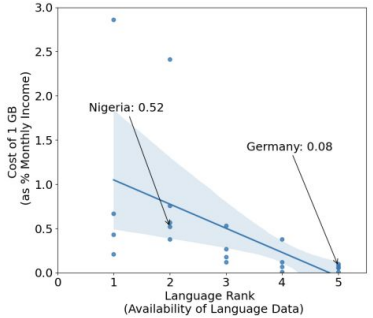


Figure 1: Cost of mobile data by country per language rank according to the taxonomy by Joshi et al. (2020).

Google

The Low Resource Double Bind

How does compression impact performance on OOD data points?

Out-of-distribution considerations:

Test data is very different from my training distribution (typical of low data regimes)

Memorization of rare artefacts in the training data is unlikely to help with generalization to the test data.

"Bad memorization"

Test dataset is very similar to my training distribution (more common with VERY big data regimes)

Memorization of rare artefacts in the training data is likely to help with generalization to the test data.

"Good memorization"

The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation

Low resource double-bind

Limited Data Regime

↕

Compute resource constraints

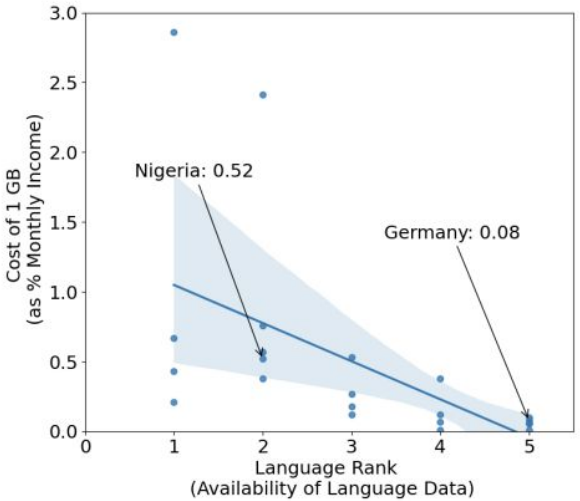
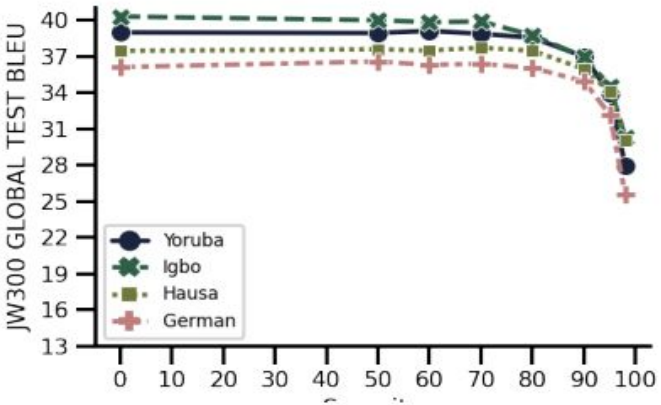
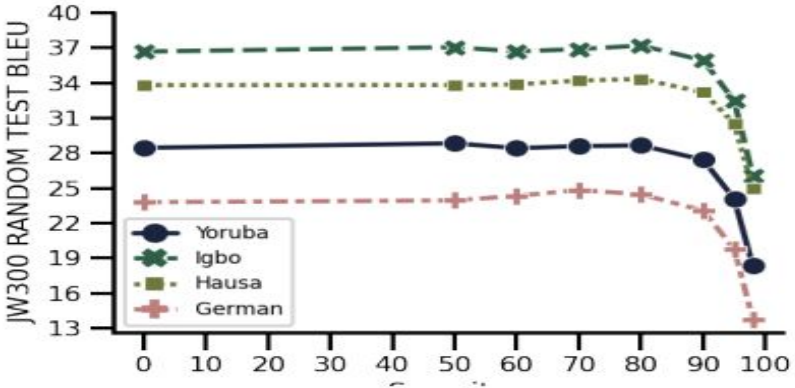


Figure 1: Cost of mobile data by country per language rank according to the taxonomy by Joshi et al. (2020).

Key results upfront: In a low data regime, sparsity disproportionately impacts performance on the long-tail.



Prototypical test-set



Random test set

The Low Resource Double Bind

Surprisingly, we also find that in this setting, high levels of sparsity consistently **improves** generalization to out-of-distribution datasets.

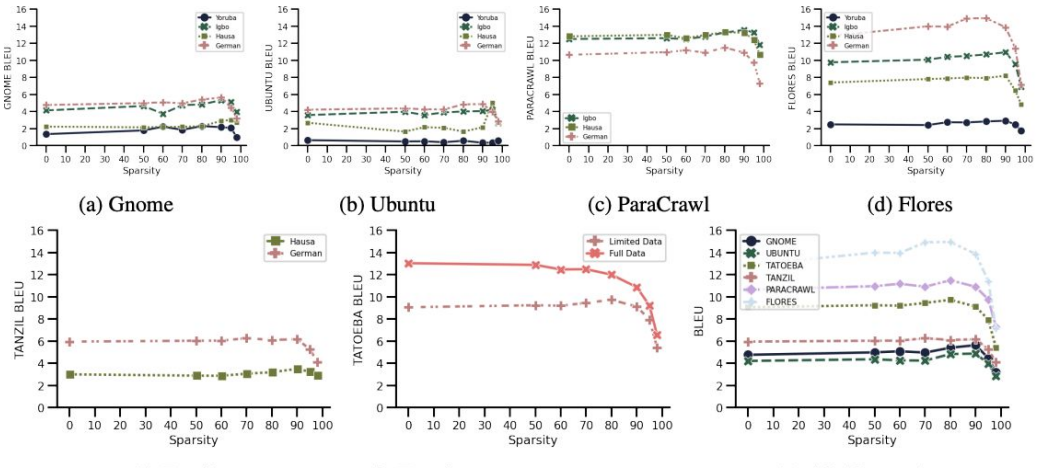
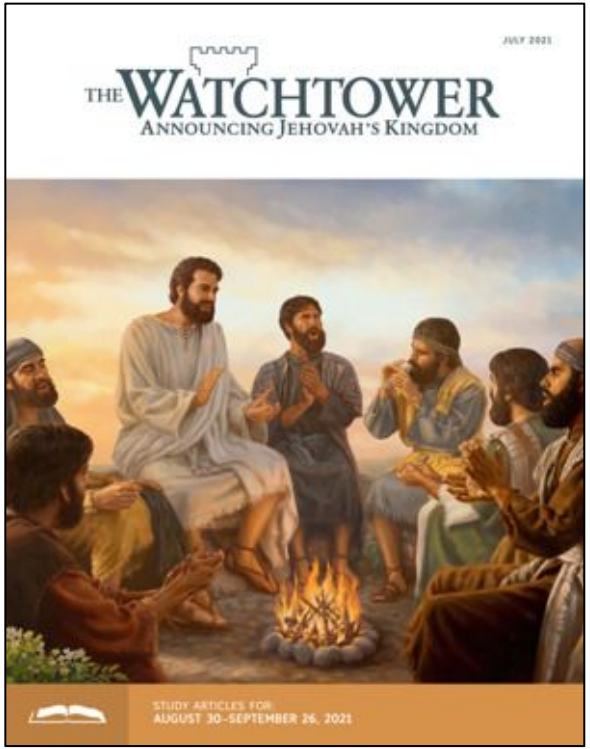


Figure 4: Robustness to distribution shift at different levels of sparsity for models trained in a data-limited regime (`Limited`). For Tatoeba with only German we compare performance for a model trained on `Full` is added for comparison.

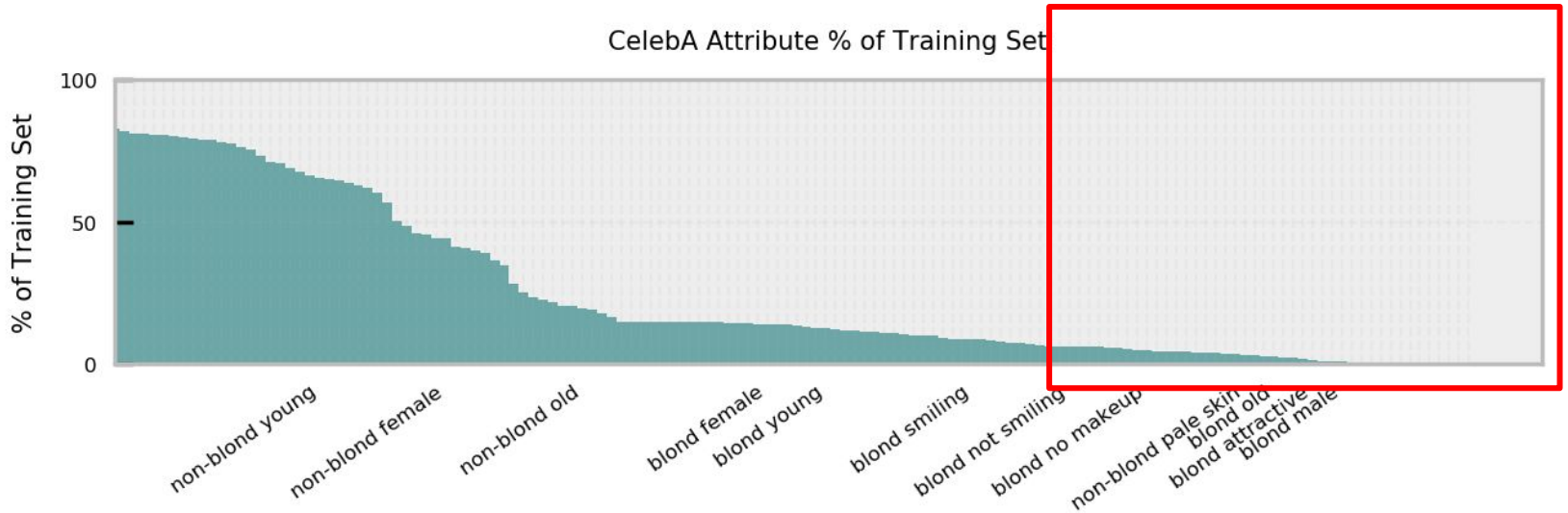
The Low Resource Double Bind

Relates to a wider question - when do we want to curb or aid memorization of rare features?

JW300 is very specialized religious corpus. Rare artefacts **even rarer** in other settings we wish to generalize to.

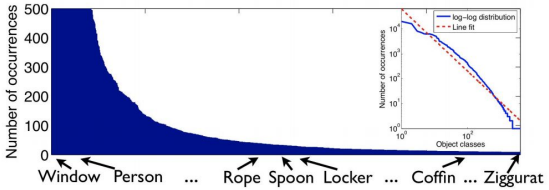


What all these settings have in common is that memorization is currently very expensive.

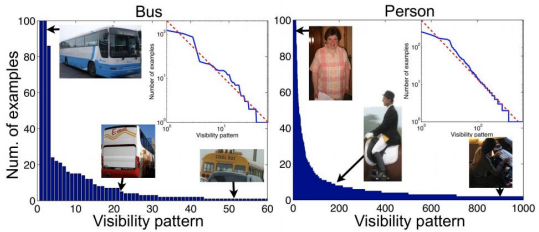


The majority of weights **(90% of all weights)** are used to memorize very rare examples in the dataset.

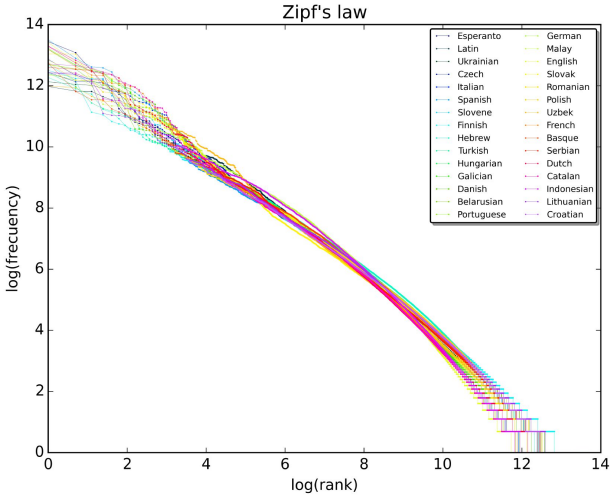
This has far ranging implications. Most natural image, NLP and audio datasets follow a Zipf distribution. If we want to model the world, we need to design train models that can efficiently navigate low-frequency events.



(a) The number of examples by object class in SUN dataset



(b) Distributions of the visibility patterns for bus and person



Zhu et al., plot of word frequency in Wikipedia

Other model design choices which can amplify or curb harm.

Privacy trade-off with fairness - differential privacy disproportionately impacts underrepresented attributes.

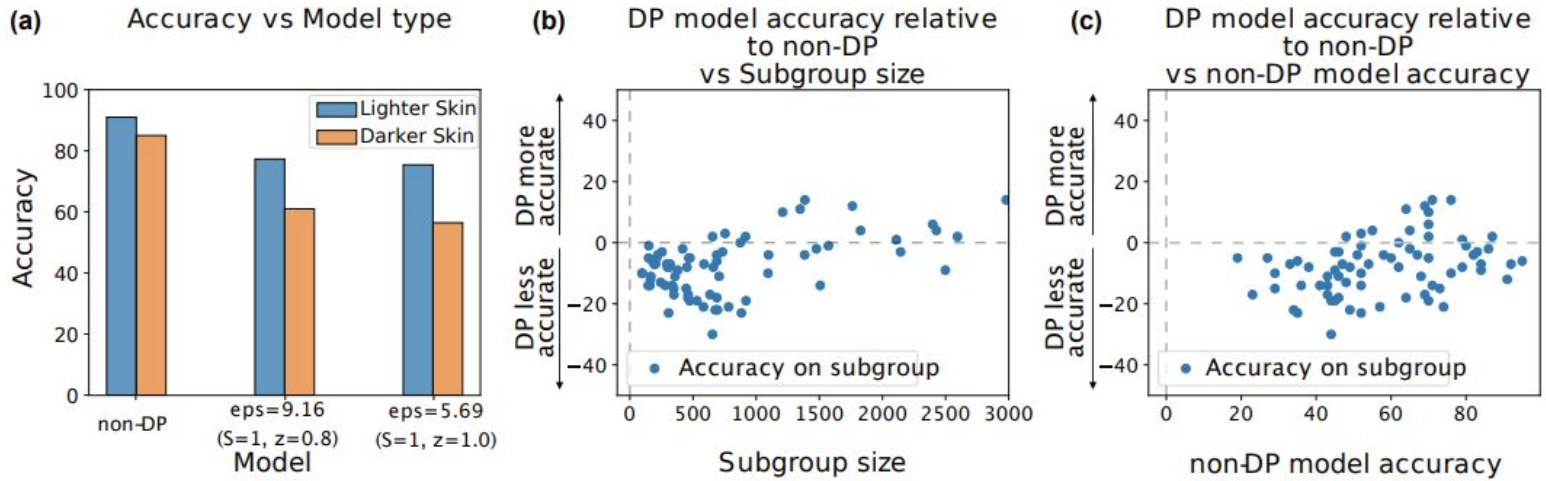


Figure 1: Gender and age classification on facial images.

Bagdasaryan et al.

Stopping training early disproportionately impacts performance on less common and more challenging features.

Recent research suggests there are distinct stages to training.

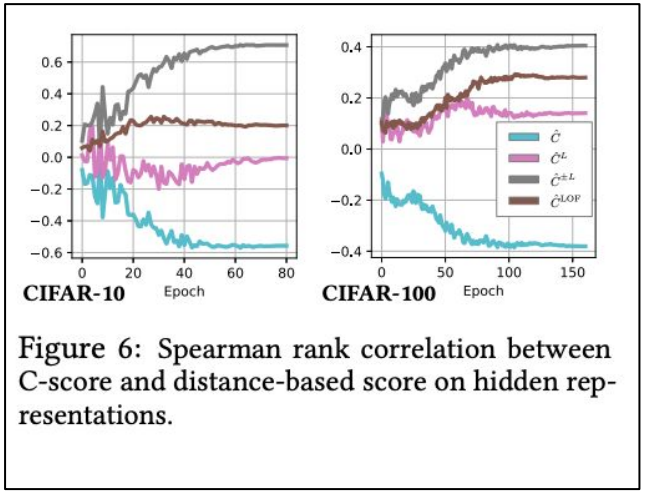
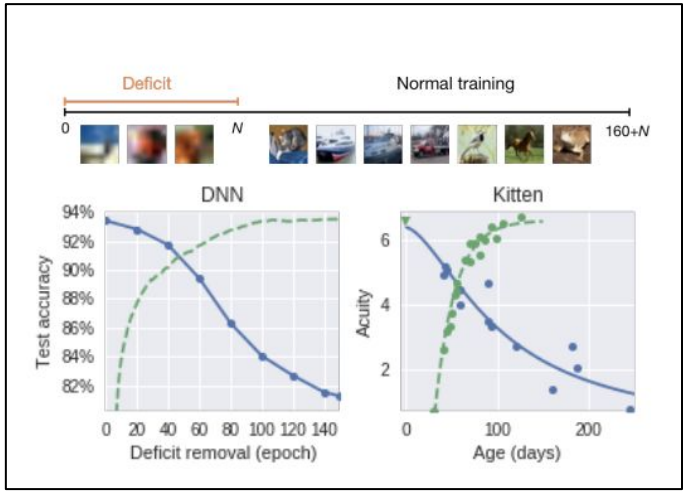


Figure 6: Spearman rank correlation between C-score and distance-based score on hidden representations.



Actionable:

- Understand what features emerge when.
- Allows us to fix incorrect labels/annotation error
- Identify biases

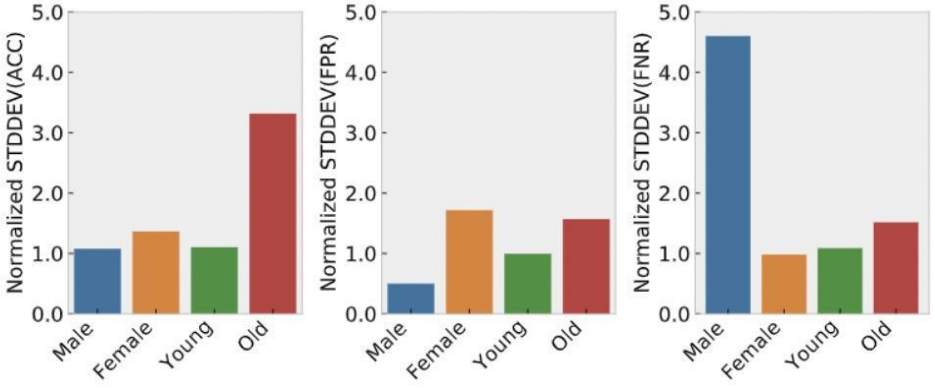
Characterizing Structural Regularities of Labeled Data in Overparameterized Models, 2020 (link)

Critical Learning Periods in Deep Neural Networks, 2017 (link)

Tooling also can impact the generalization properties of your algorithm.

The non-determinism introduced by tooling disproportionately impacts underrepresented attributes.

- Underrepresented attributes disproportionately impacted by the introduction of stochasticity
- High variance is an AI-safety issue on sub-group with sensitive attributes like race, gender, and age.



	Male	Female	Young	Old
Positive Data Points	**1387 (0.8%)**	22880 (14.1%)	20230 (12.4%)	**4037 (2.5%)**
Negative Data Points	66874 (41.1%)	71629 (44.0%)	106558 (65.5%)	31945 (19.6%)

Data points distribution in CelebA dataset

Randomness In Neural Network Training: Characterizing The Impact of Tooling [Donglin Zhuang, Xingyao Zhang, Shuaiwen Song, Sara Hooker,]

Alan Blackwell said in 1997 that in computer science

“many sub-goals can be deferred to the degree that they become what is known amongst professional programmers as S.E.P - somebody else’s problem”

The belief that algorithmic bias is only a dataset problem invites diffusion of responsibility and misses important opportunities to curb harm.

Lord Kelvin reflected, ‘‘If you cannot measure it, you cannot improve it.’’

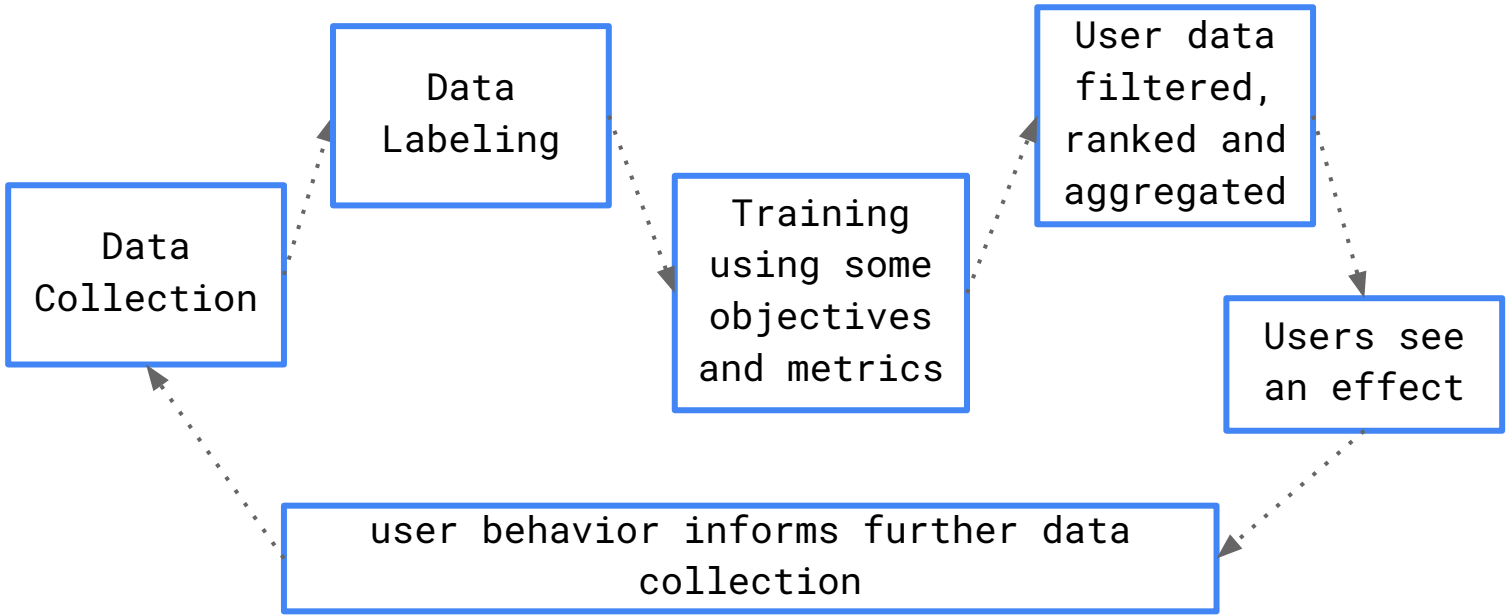
Acknowledging that model design matters has the benefit of spurring more research focus on how it matters and will inevitably surface new insights into how we can design models to minimize harm.

Diffenderfer et al. 2021

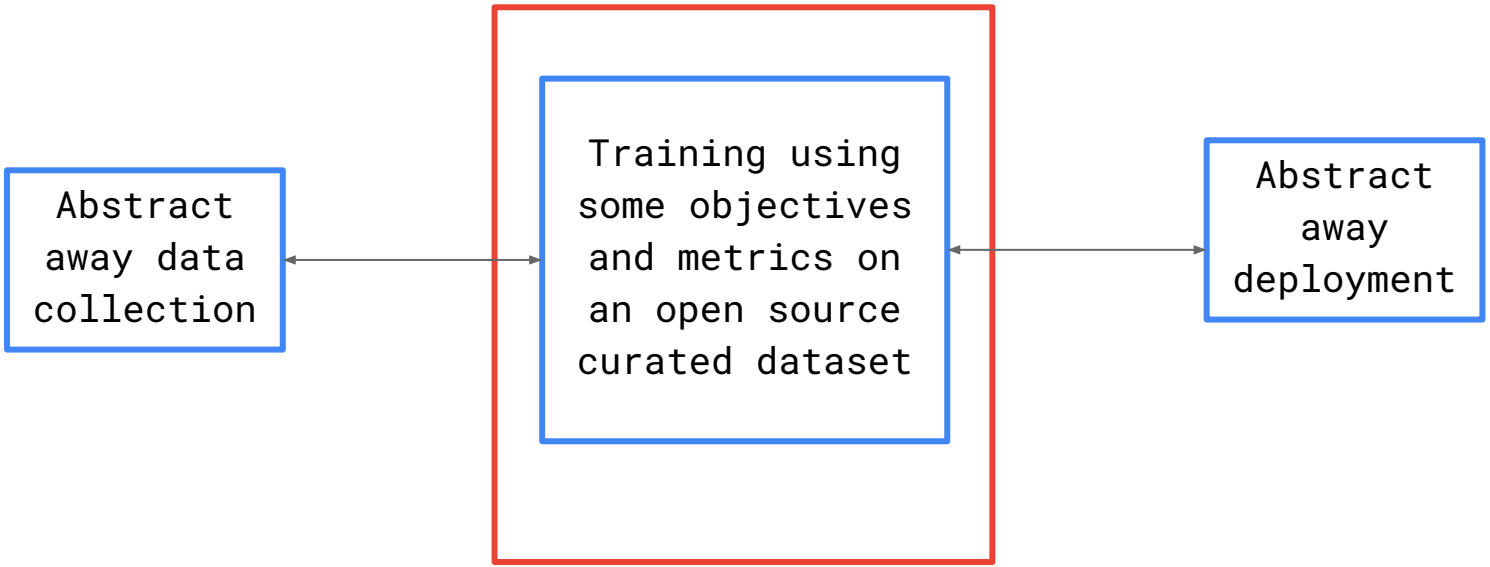
Joseph, Siddiqui et al. 2020

The way forward

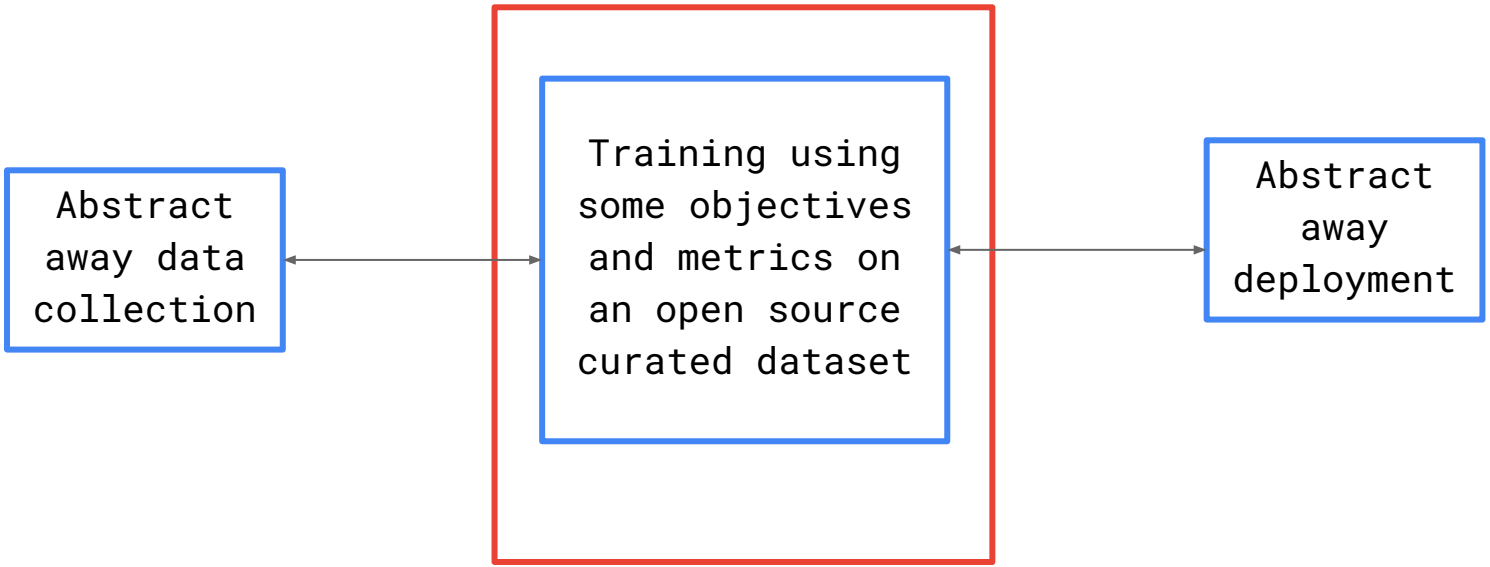
Deploying an algorithm involves many different steps.



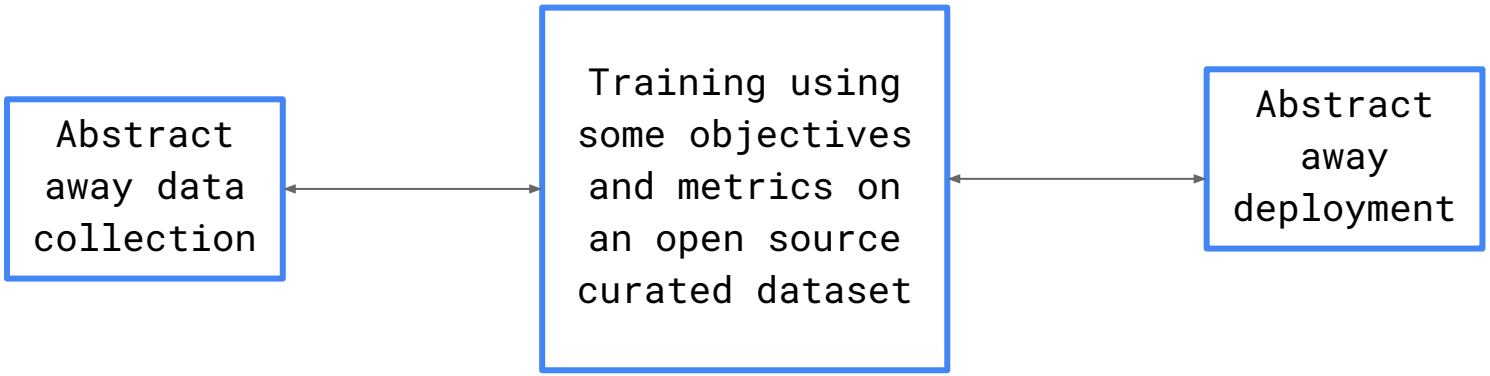
However, the machine learning research community has disproportionately published around one step.



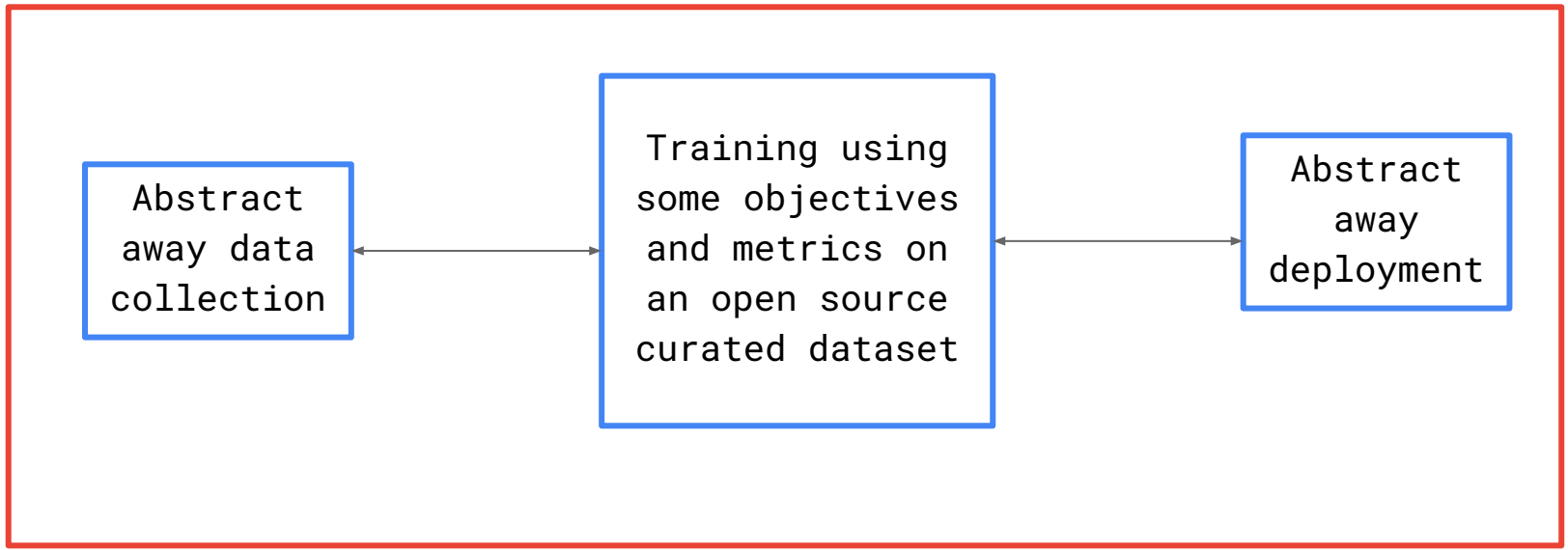
The surprisingly widely held belief that models are impartial displaces responsibility for bias to the those responsible for the data pipeline.



If bias is not fully addressed in the data pipeline, harm is a product of both data and design choices. Model design choices **can and do** amplify harm.



Understanding the interactions between model and dataset can open up new mitigation strategies for designing models that are better specified.



Closing Thoughts (and Q&A)

Questions?

Moving beyond "algorithmic bias is a data problem." *Sara Hooker [[link]]*

Estimating Example Difficulty using Variance of Gradients Chirag Agarwal*, Sara Hooker* [[link]]

The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation *Orevaoghene Ahia, Julia Kreutzer, Sara Hooker [[link]]*

What do compressed deep neural networks forget?, Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, Andrea Frome [[link]]

Characterizing Bias in Compressed Models Sara Hooker*, Nyalleng Moorosi*, Gregory Clark, Samy Bengio, Emily Denton [[link]]

Final takeaways:

Beyond test-set accuracy - It is not always possible to measure the trade-offs between criteria using test-set accuracy alone.

The myth of the compact, private, interpretable, fair model - Desiderata are not independent of each other. Training beyond test set accuracy requires trade-offs in our model preferences.

Understanding the interactions between model and dataset can open up new mitigation strategies.

.

Email: shooker@google.com

Google