Lecture 2 - finishing up from last time

The humans strike back,
The humans-in-the-loop

Ranjay Krishna | ranjay@cs.washington.edu

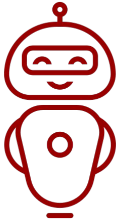
Course logistics

Discussion sections:

- We will discuss two papers
- We will combine both papers together across roles to save time

Project proposals are due Jan 24 at 11:59pm

The humans-in-the-loop: two perspectives



Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Human-Computer Interaction

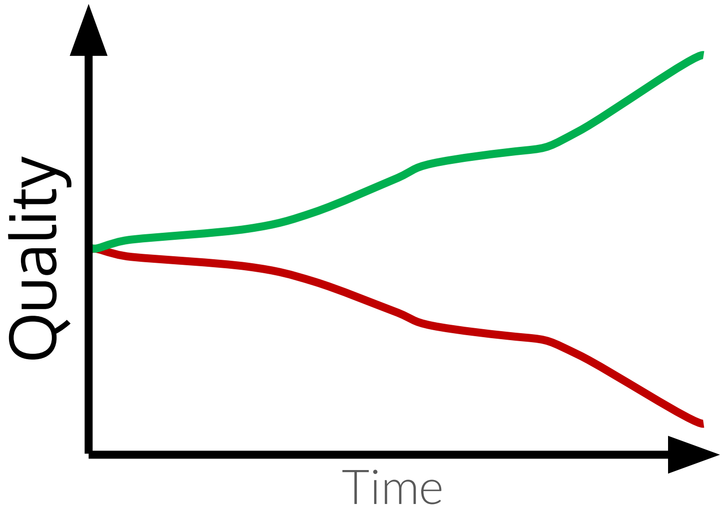
Goal: To support a labor force achieve their financial and career goals

Artifact: automations that structure work

Impacts across long time horizon

Ranjay Krishna | ranjay@cs.washington.edu

Studying long term annotator quality



Quality increases over time:

Quality decreases over time:

[Hata et al. A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality. CSCW 2017]

Speeding up annotation

[Krishna et al. Embracing Error to Enable Rapid Crowdsourcing. CHI 2016]

Job Characteristic Model

Hackman & Oldham, 1980

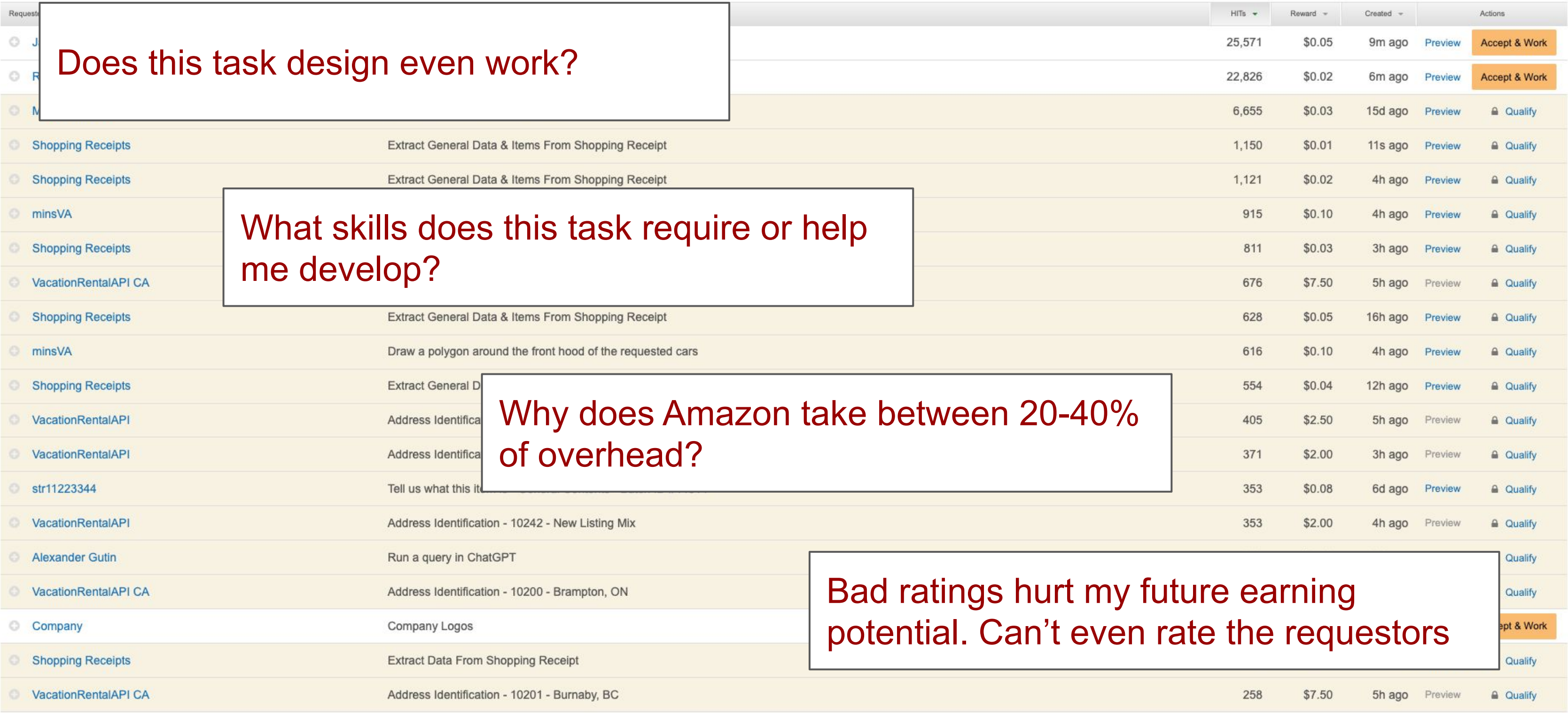
Core Job Characteristics → Critical Psychological States → Outcomes

Core Job Characteristics	Critical Psychological States	Outcomes
Skill variety Skill identity Skill significance	Experience meaningfulness of the work	High internal work motivation
Autonomy	Experience responsibility of the outcomes of the work	High "growth" satisfaction
Feedback	Knowledge of the actual results of the work	High general job satisfaction
		High work effectiveness

Existing platforms do not support these job characteristics

Requester	Title	HITs	Reward	Created	Actions	
James Billings	Market Research Survey	25,571	$0.05	9m ago	Preview	Accept & Work
Research Rewards	Quick Market Research Survey	22,826	$0.02	6m ago	Preview	Accept & Work
Mayanksoniphd	Generate praise, given a persona.	6,655	$0.03	15d ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	1,150	$0.01	11s ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	1,121	$0.02	4h ago	Preview	Qualify
minsVA	Draw a polygon around the tailgate of the requested cars	915	$0.10	4h ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	811	$0.03	3h ago	Preview	Qualify
VacationRentalAPI CA	Address Identification - 10207 - Kelowna, BC	676	$7.50	5h ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	628	$0.05	16h ago	Preview	Qualify
minsVA	Draw a polygon around the front hood of the requested cars	616	$0.10	4h ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	554	$0.04	12h ago	Preview	Qualify
VacationRentalAPI	Address Identification - 10227 - Minneapolis, MN	405	$2.50	5h ago	Preview	Qualify
VacationRentalAPI	Address Identification - 10243 - New Listing Mix	371	$2.00	3h ago	Preview	Qualify
str11223344	Tell us what this item is - General Contents - Batch ID #44814	353	$0.08	6d ago	Preview	Qualify
VacationRentalAPI	Address Identification - 10242 - New Listing Mix	353	$2.00	4h ago	Preview	Qualify
Alexander Gutin	Run a query in ChatGPT	326	$0.02	11d ago	Preview	Qualify
VacationRentalAPI CA	Address Identification - 10200 - Brampton, ON	321	$7.50	5h ago	Preview	Qualify
Company	Company Logos	297	$0.01	17s ago	Preview	Accept & Work
Shopping Receipts	Extract Data From Shopping Receipt	294	$0.01	1m ago	Preview	Qualify
VacationRentalAPI CA	Address Identification - 10201 - Burnaby, BC	258	$7.50	5h ago	Preview	Qualify

Existing platforms do not support these job characteristics



Ranjay Krishna | ranjay@cs.washington.edu

Humans-in-the-loop from an HCI perspective:

Can we develop a platform that supports worker needs?

Daemo: a Self-Governed Crowdsourcing Marketplace

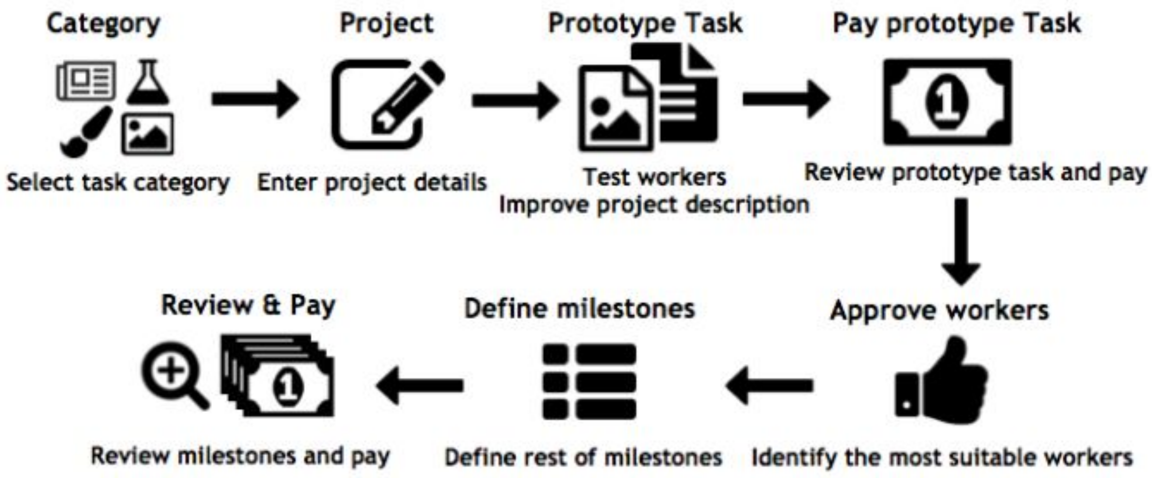
V1 launched with :

Prototype tasks

- Workers improve task design

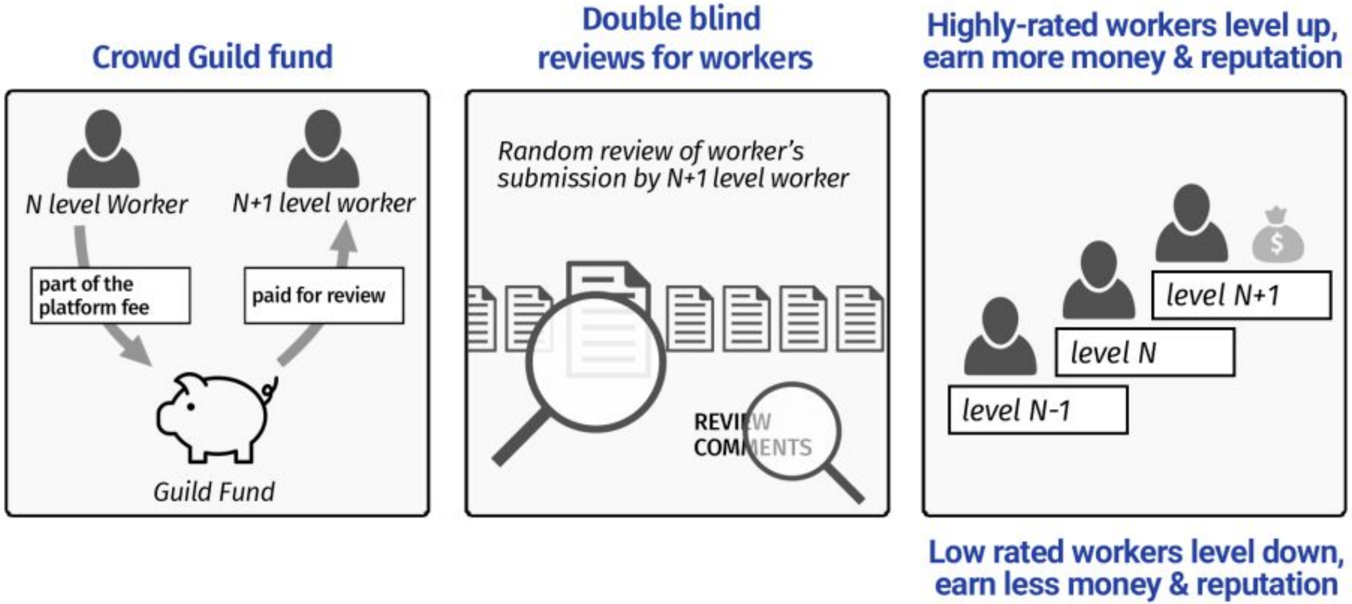
Open governance

- 3 workers
- 3 requesters
- 1 researcher



Gaikwad et al. Daemo: a Self-Governed Crowdsourcing Marketplace. UIST 2017

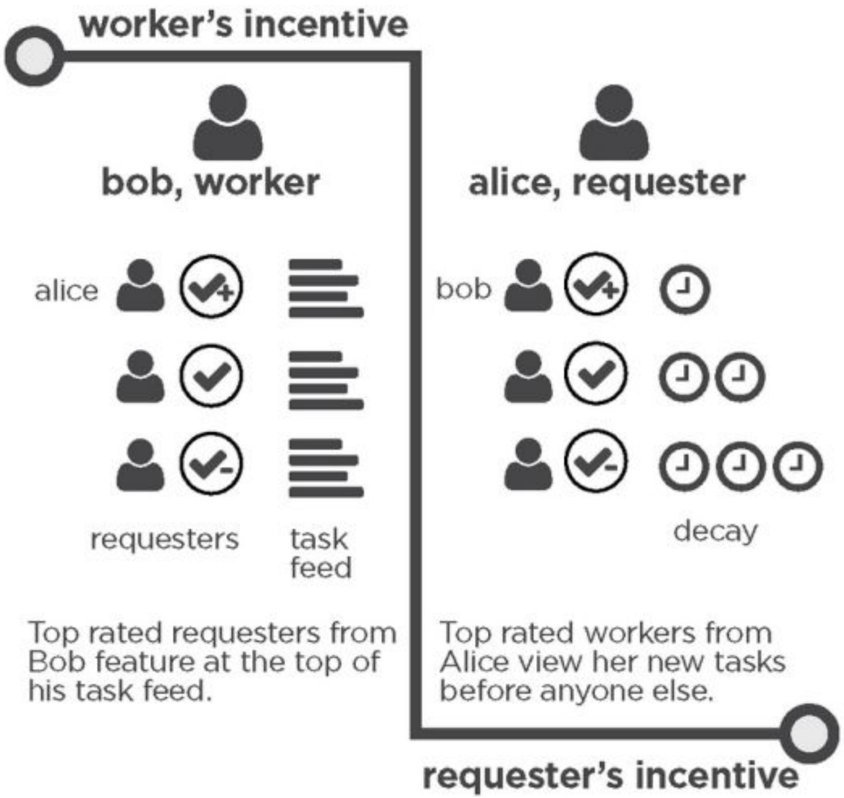
A reputation protocol: workers received feedback



Ranjay Krishna | ranjay@cs.washington.edu

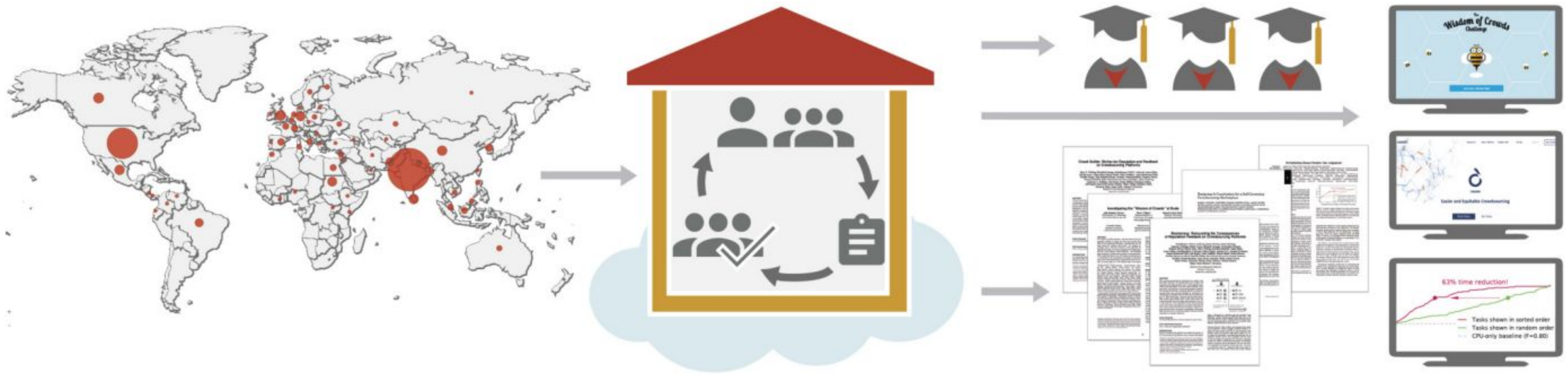
Whiting et al. Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. CSCW 2017

A rating system: To trade off skill variety of identity



Gaikwad et al. Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms. UIST 2016

Building a new decentralized crowdsourcing system with a crowd of researchers

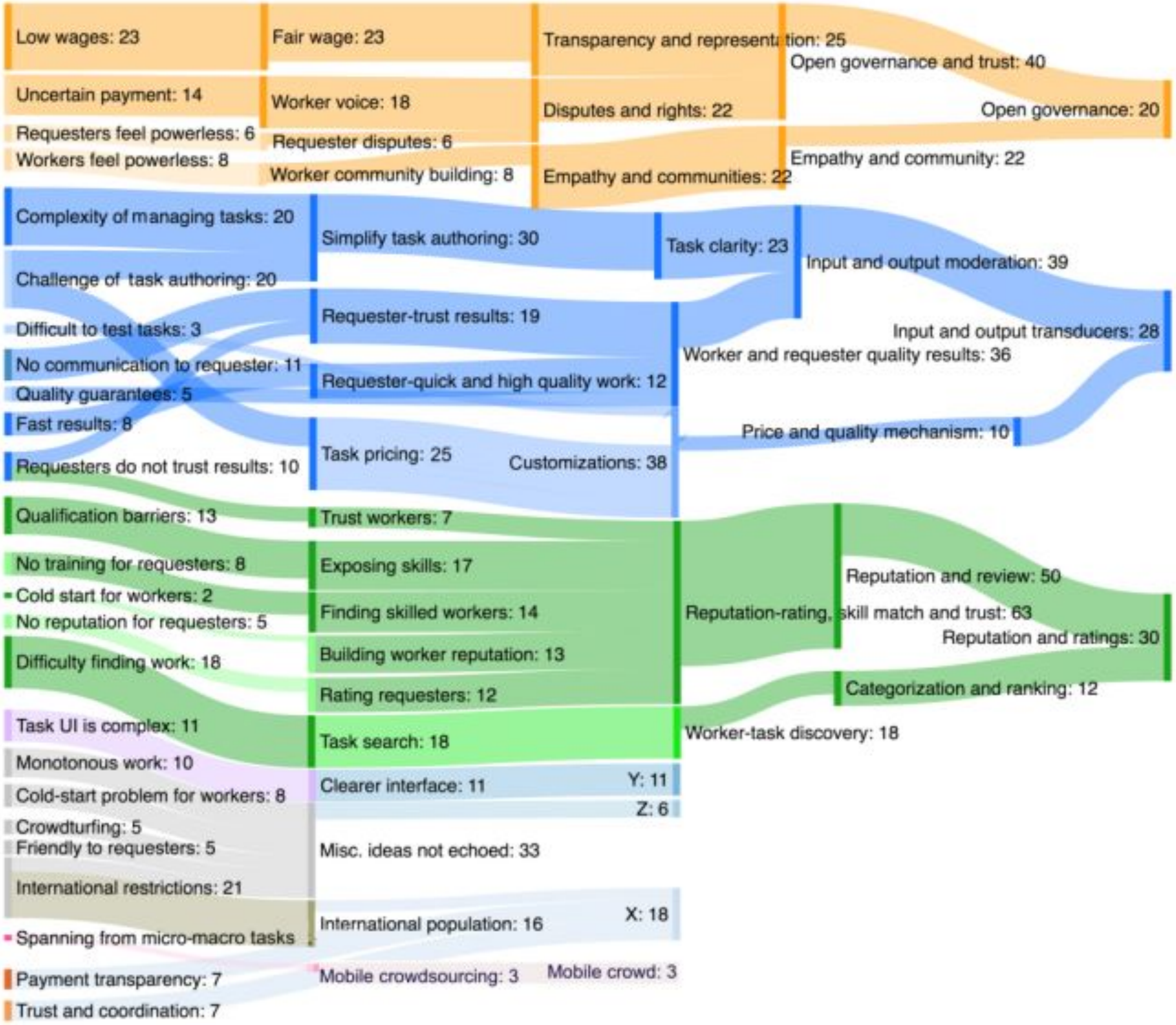


Achieve upward educational mobility while creating research systems and co-authoring papers

Vaish et al. Crowd Research: Open and Scalable University Laboratories. UIST 2017

Ideas

Changes to the platform were ideated on transparently and collectively prioritized

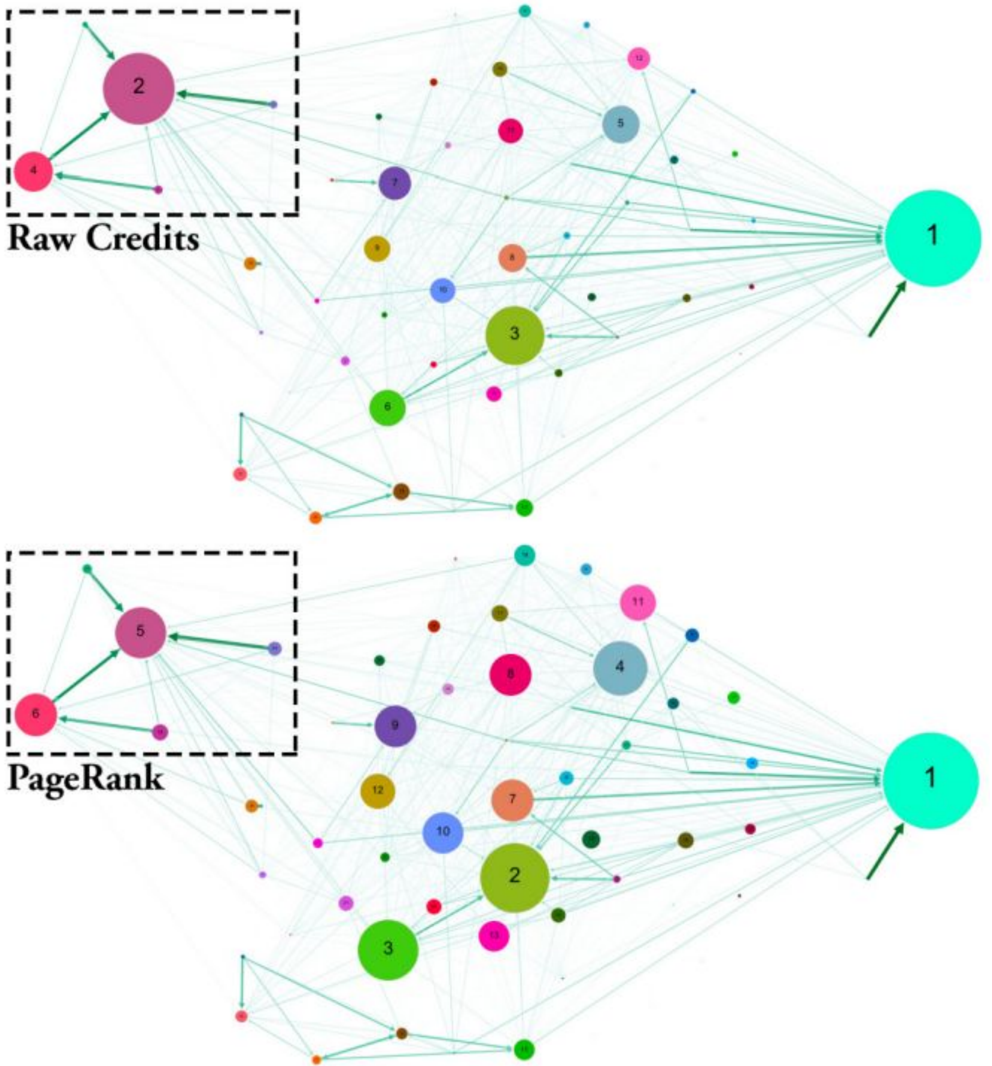


Ranjay Krishna | ranjay@cs.washington.edu

Author order determined using crowdsourced points and page rank

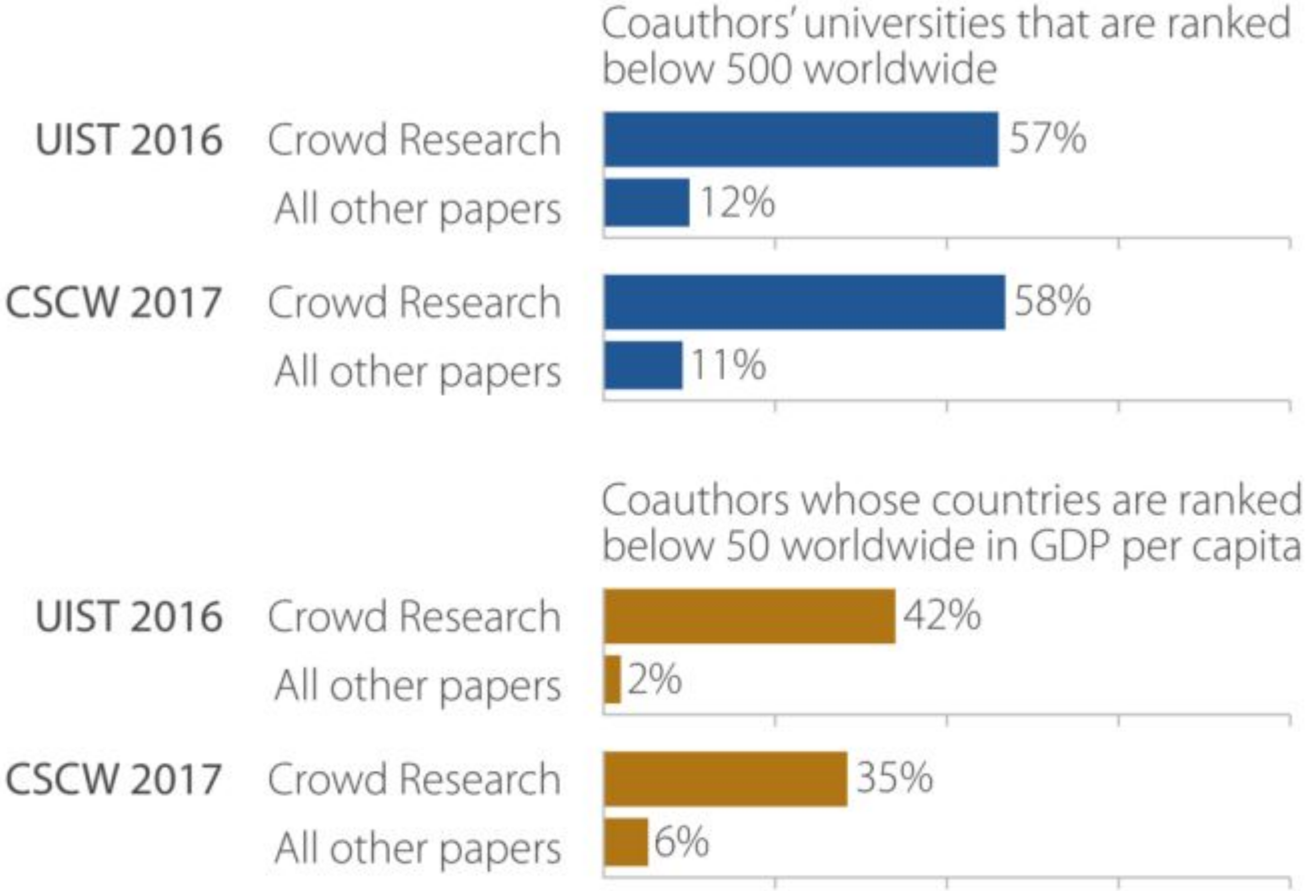
Potential challenges:

- Link ring
- Quid-proquo strategy



Supporting upward mobility

Our authors were more diverse than those from other papers at the same venue



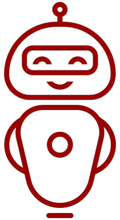
Job Characteristic Model

Hackman & Oldham, 1980

Core Job Characteristics → Critical Psychological States → Outcomes

Core Job Characteristics		Critical Psychological States		Outcomes
Skill variety Skill identity Skill significance	→	Experience meaningfulness of the work	→	High internal work motivation High "growth" satisfaction High general job satisfaction High work effectiveness
Autonomy	→	Experience responsibility of the outcomes of the work		
Feedback	→	Knowledge of the actual results of the work		

The humans-in-the-loop: two perspectives



Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Human-Computer Interaction

Goal: To support a labor force achieve their financial and career goals

Artifact: automations that structure work

Impacts across long time horizon

Ranjay Krishna | ranjay@cs.washington.edu

Lecture 3

Return of the metrics,
The challenges with evaluating models

Ranjay Krishna | ranjay@cs.washington.edu

Today’s questions: We will take an AI perspective today

Is a model good enough for deployment?

Which model is better?

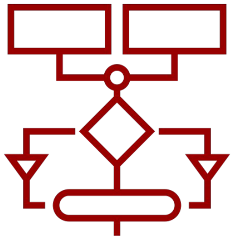
How do we design effective evaluation metrics?

How do we utilize these metrics within an appropriate evaluation protocols?

Main take away from today’s lecture

Machine learning evaluation is a challenging unsolved problem.

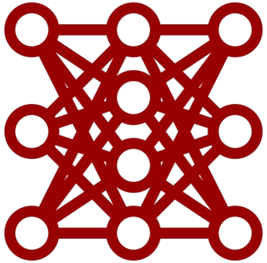
A shift in AI: From algorithms to machine learning



Classical algorithms

Problems: precisely defined algebraically

Example: Graphcut algorithm

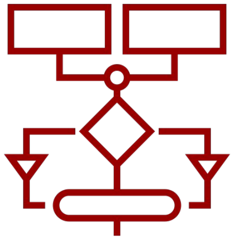


Empirical machine learning

Problems: loosely defined by datasets

Example: ResNet50 trained on ImageNet 1K

A shift in AI: From algorithms to machine learning



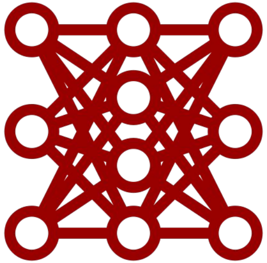
Classical algorithms

Problems: precisely defined algebraically

Example: Graphcut algorithm

Accuracy: measured by correctness

Artifact: provably correct, transparent process



Empirical machine learning

Problems: loosely defined by datasets

Example: ResNet50 trained on ImageNet 1K

Accuracy: measured using test set

Artifact: stochastic black box model

Object Classification: The ImageNet task

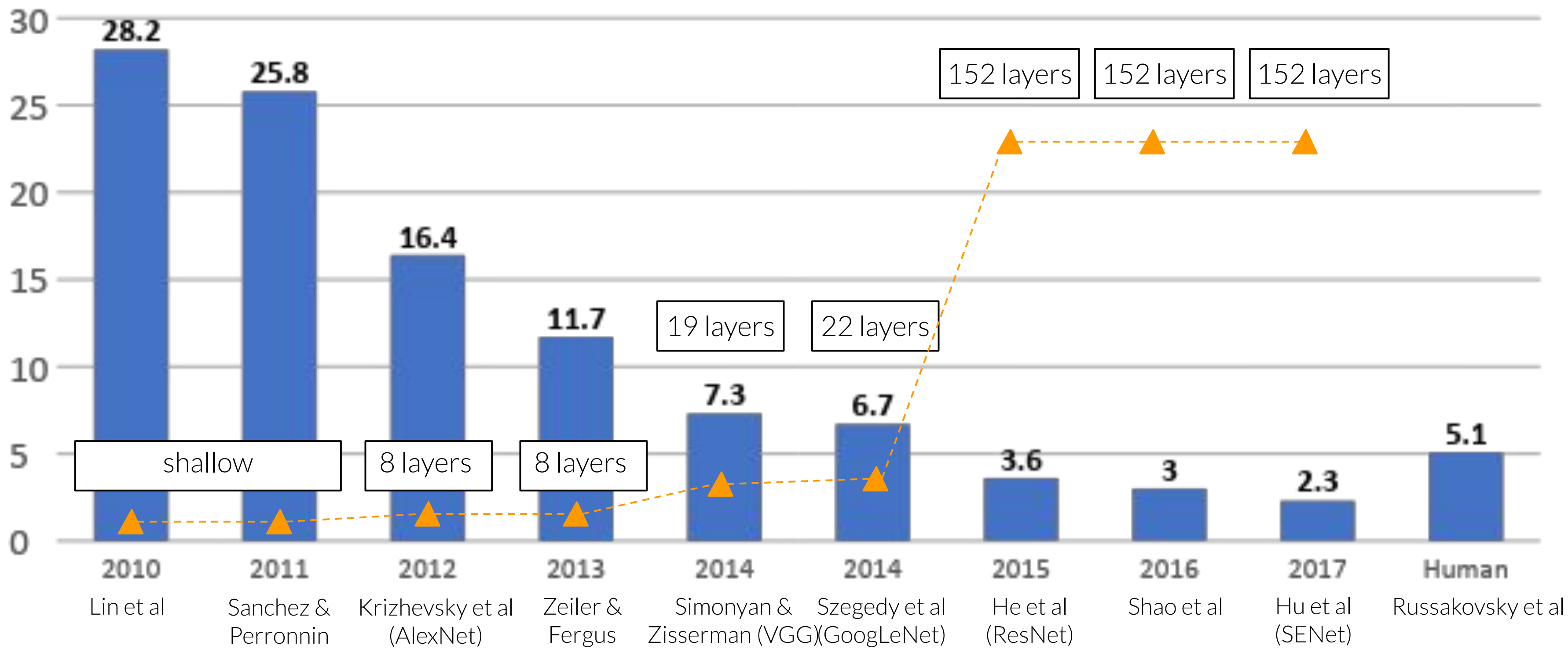
This image by Nikita is licensed under CC-BY 2.0

(assume given a set of possible labels)
{dog, cat, truck, plane, ...}

→ cat

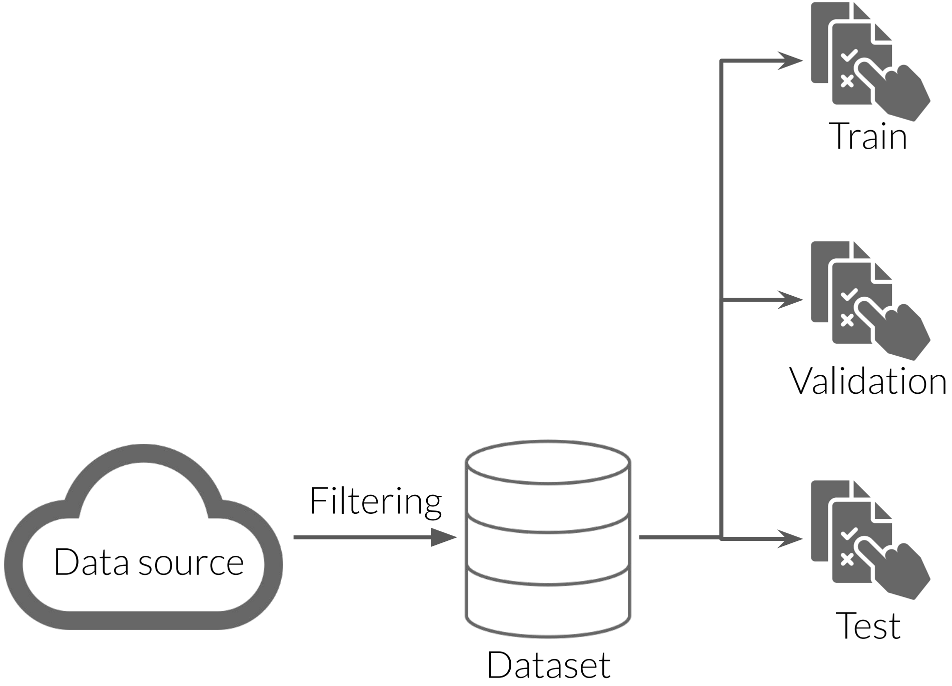
Evaluated using either top-1 or top-5 accuracy

Top-5 accuracy on ImageNet challenge over the years

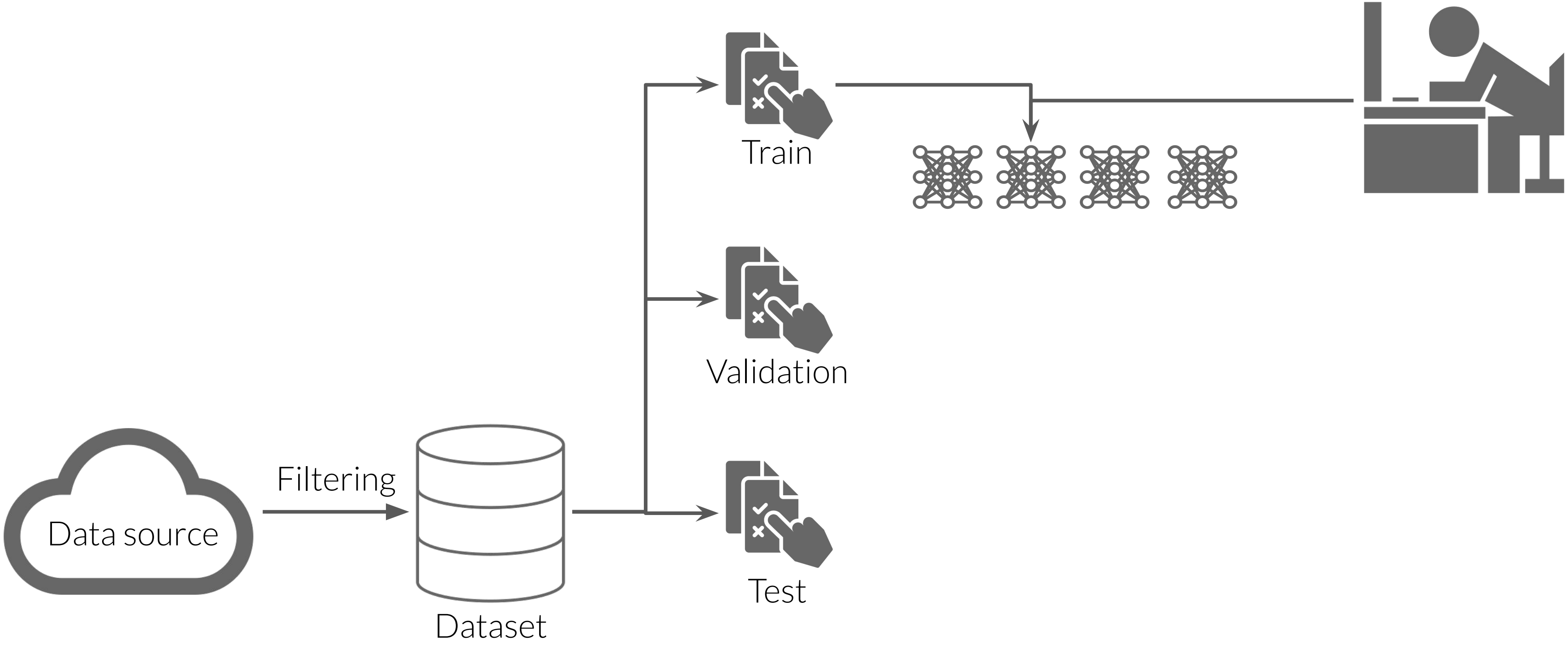


Ranjay Krishna | ranjay@cs.washington.edu

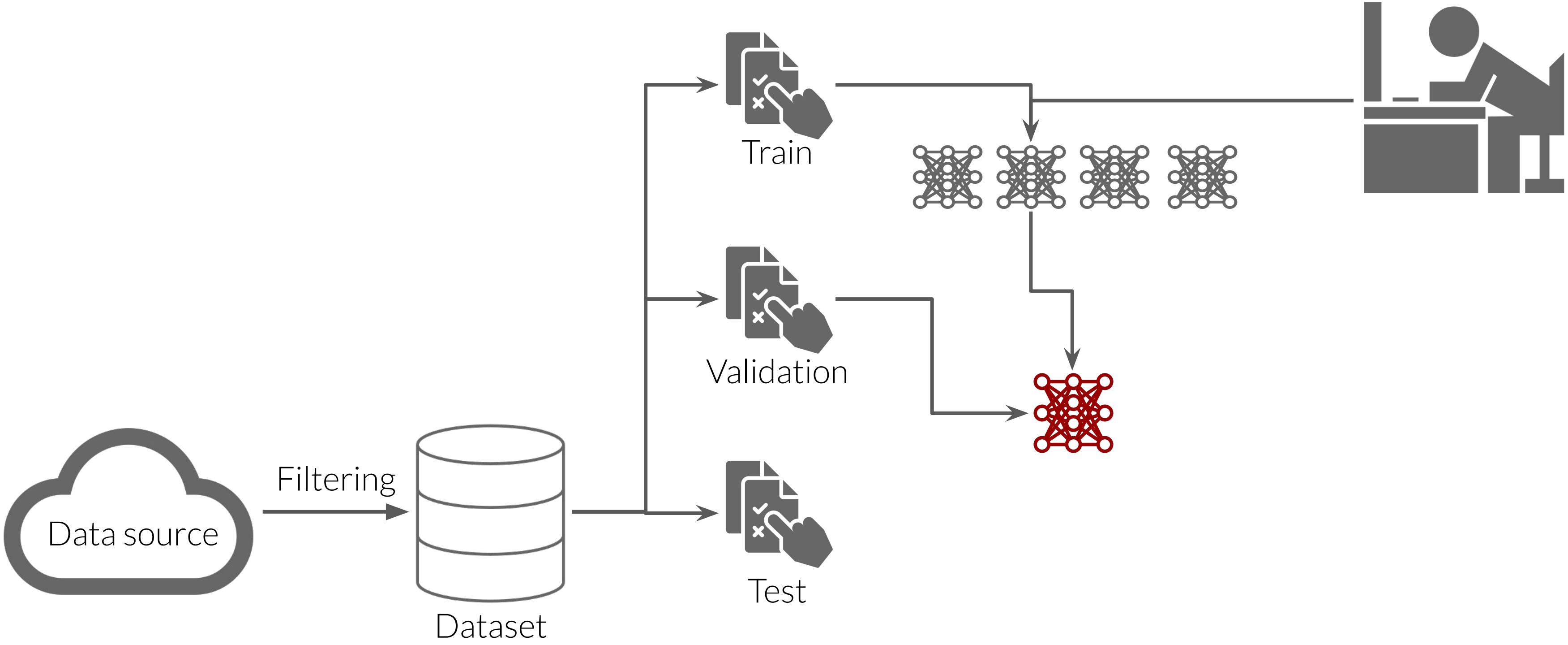
High level evaluation protocol for empirical machine learning



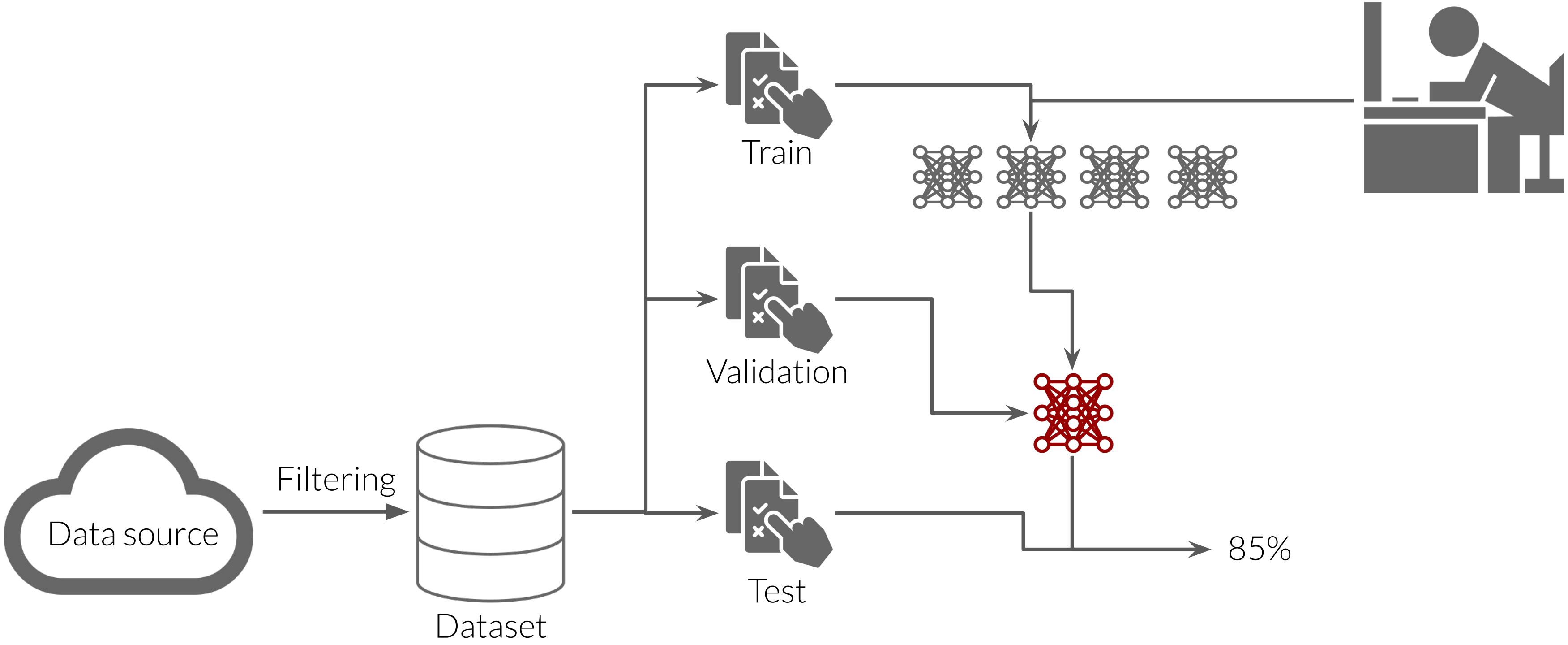
High level evaluation protocol for empirical machine learning



High level evaluation protocol for empirical machine learning



High level evaluation protocol for empirical machine learning



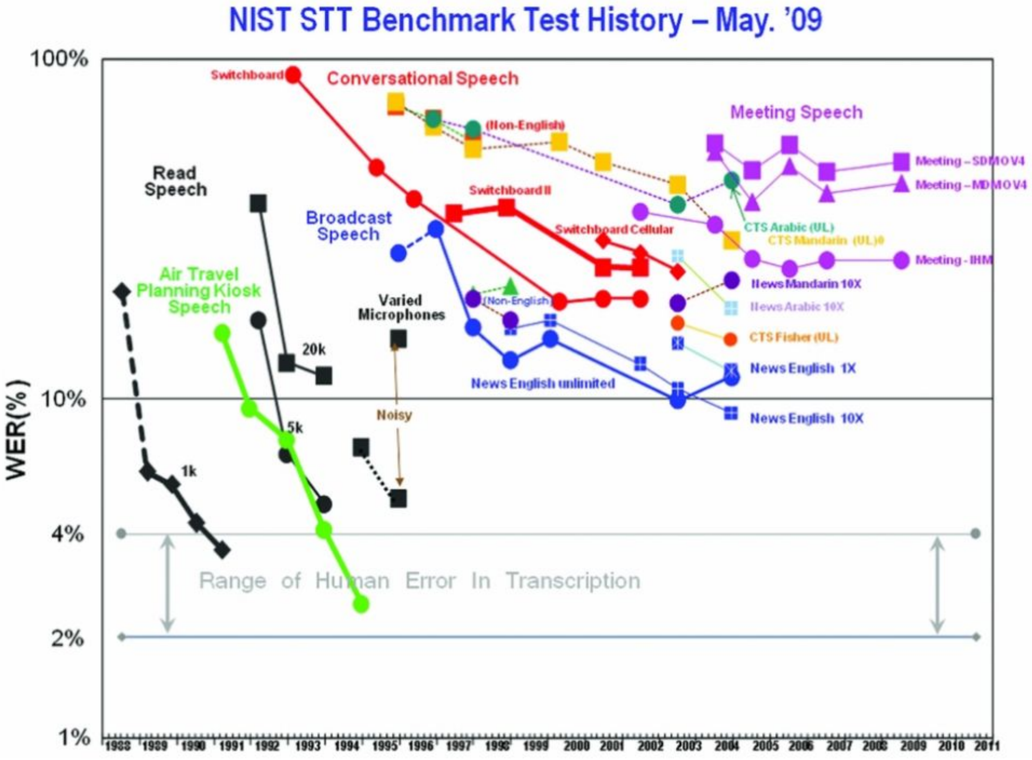
Use of benchmark test datasets and common metrics

- Dates back to 1980s.
- Funded by DAPRA and led by IBM
- Goal: solve general diction problem

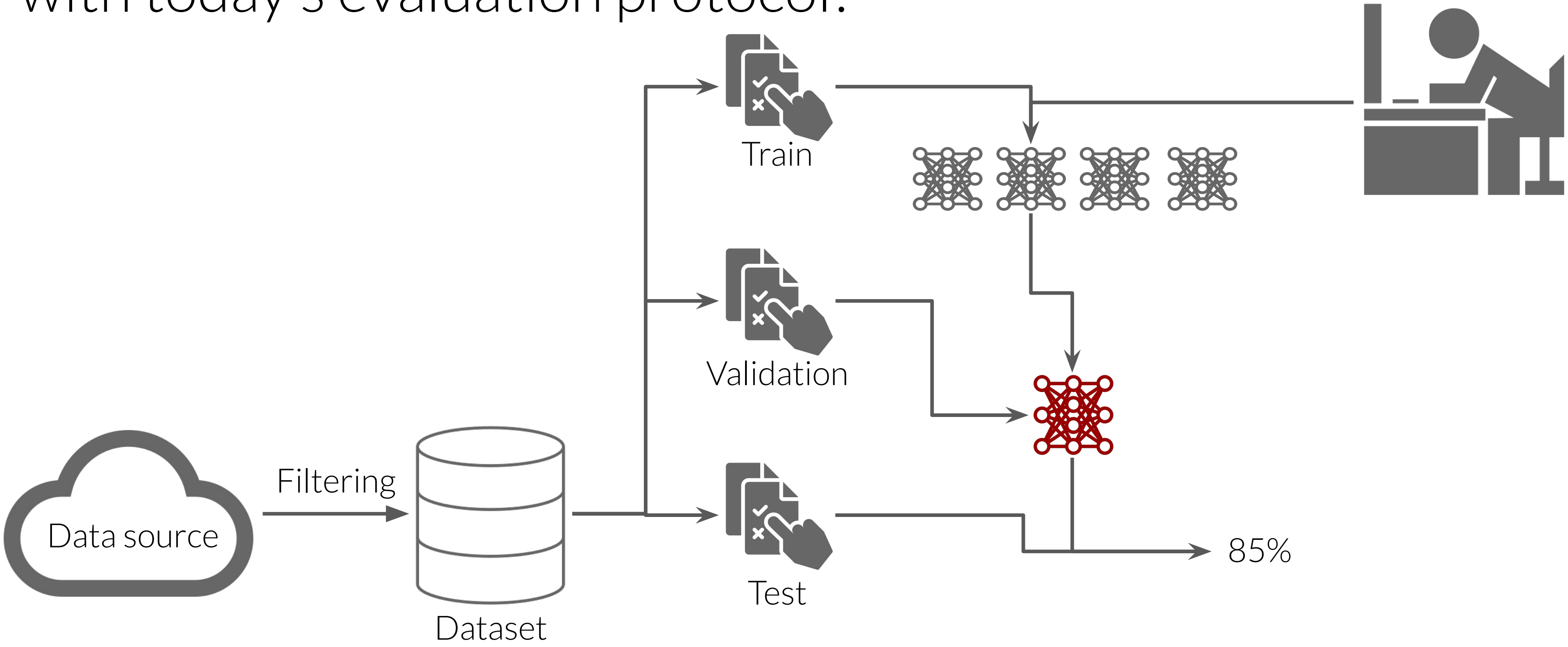
- Metric: Word error rate (WER)
- Artifact: a shared set of datasets, evaluation protocols, common metric, etc.

UCI machine learning collection of datasets

Started in 1987 by David Aha and fellow graduate students at UC Irvine.

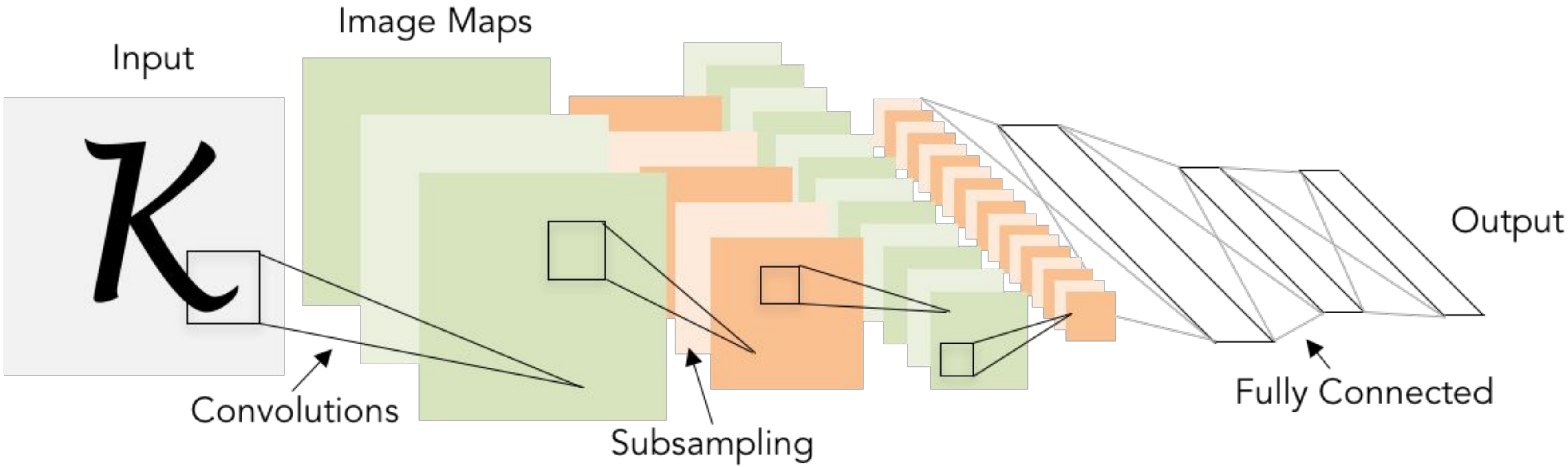


Class activity: So if things are working, can you think of issues with today's evaluation protocol?

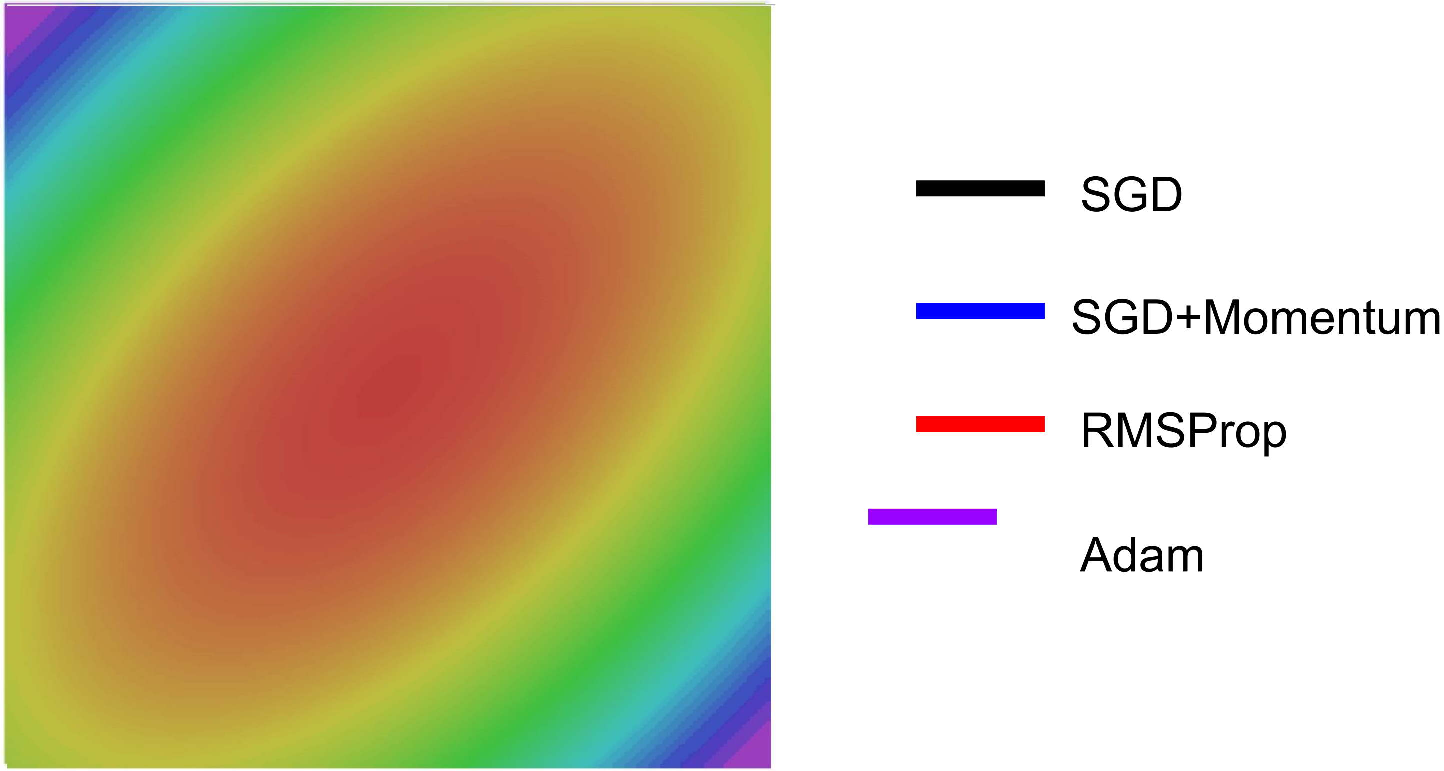


#1: The replication crisis

Take a basic convolution neural network to solve object classification for instance



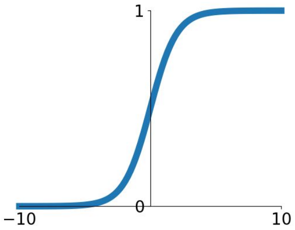
Optimization options



So many choices of activation Functions

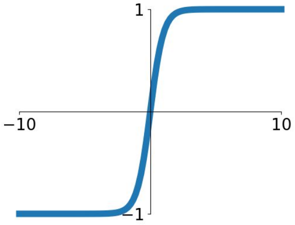
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



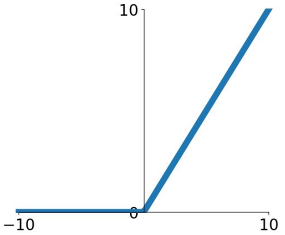
tanh

$$\tanh(x)$$



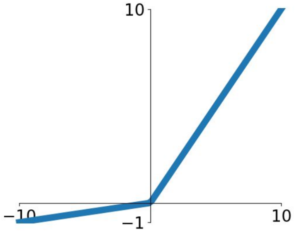
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

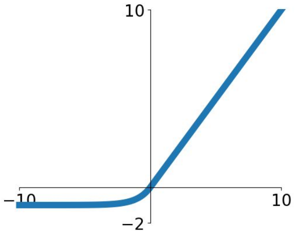


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

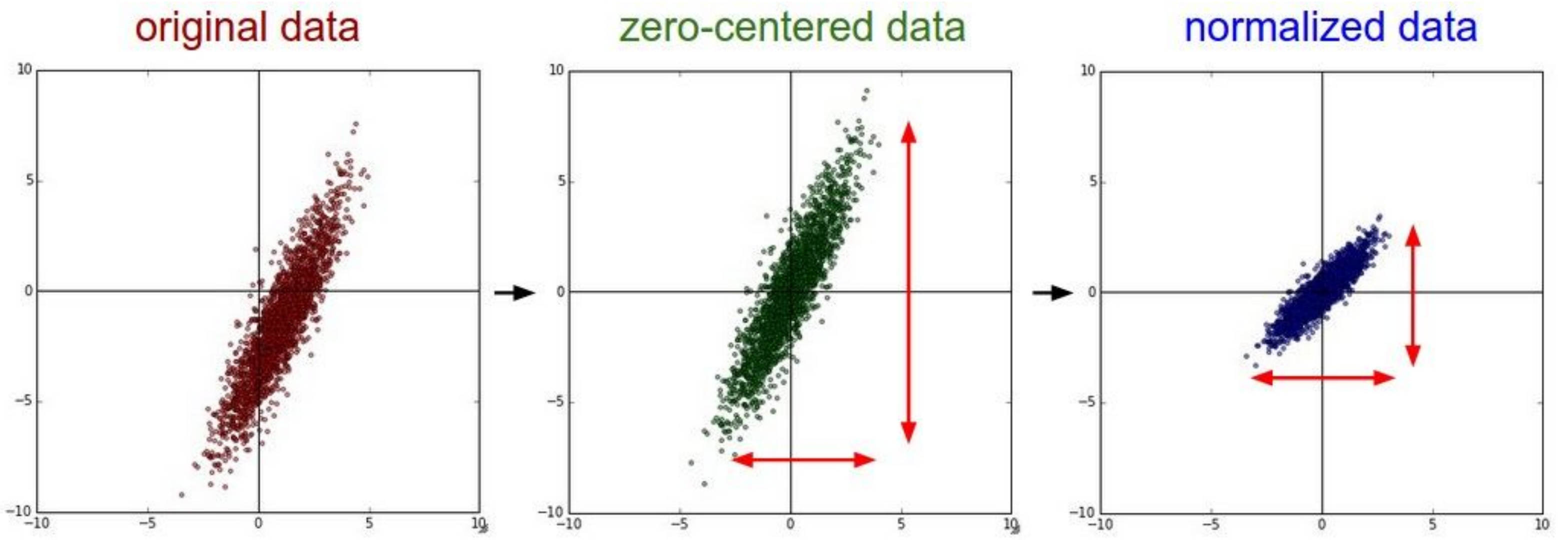
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Ranjay Krishna | ranjay@cs.washington.edu

Data preprocessing

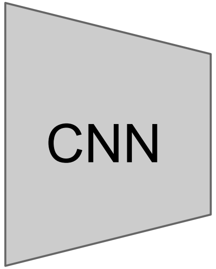


Regularization options: e.g. Mixup

Training: Train on random blends of images
Testing: Use original images

Examples:
Dropout
Batch Normalization
Data Augmentation
DropConnect
Fractional Max Pooling
Stochastic Depth
Cutout / Random Crop
Mixup

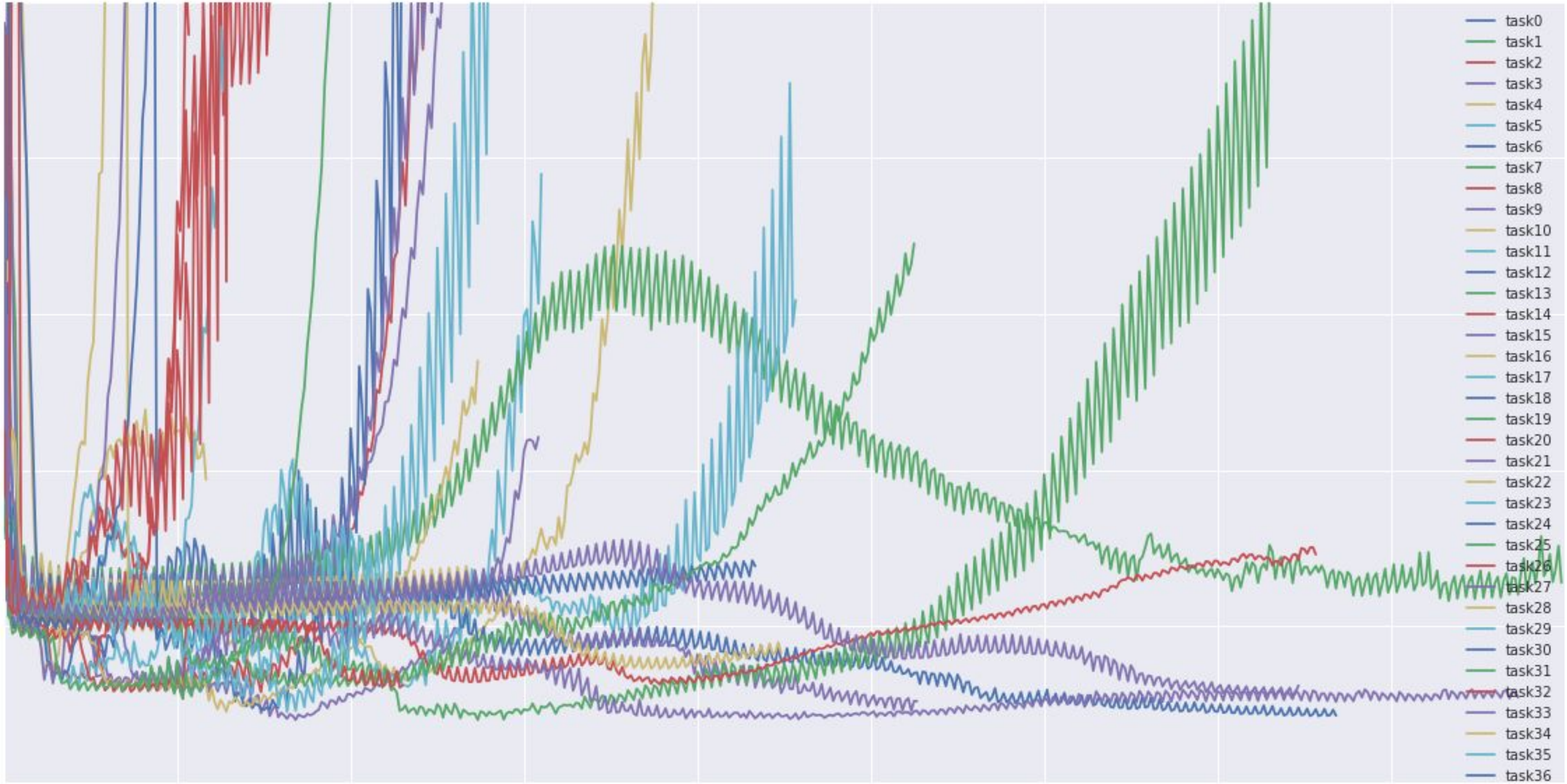
Randomly blend the pixels of pairs of training images, e.g. 40% cat, 60% dog



Target label:
cat: 0.4
dog: 0.6

Zhang et al, “*mixup*: Beyond Empirical Risk Minimization”, ICLR 2018

Loss curves are often used instead of real metrics to make decisions

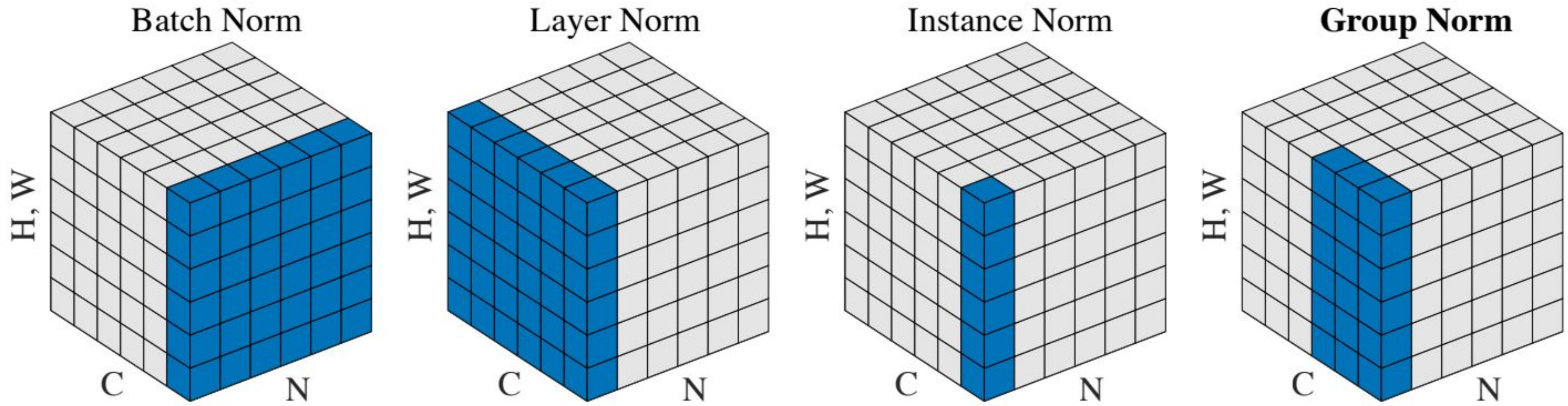


Hardware + Software options

PyTorch

TensorFlow

Normalization layer options



Wu and He, “Group Normalization”, ECCV 2018

#1: The replication crisis

All these details are lost in appendixes or during experiments.

Anecdote: sometimes we can't reproduce our own results because of other processes interfering.

#1: The replication crisis: not just a machine learning challenge

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

News & Comment > News > 2019 > May > Article

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Science Contents News Careers Journals

SHARE

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

\+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

Article | Figures & Data | Info & Metrics | eLetters | PDF

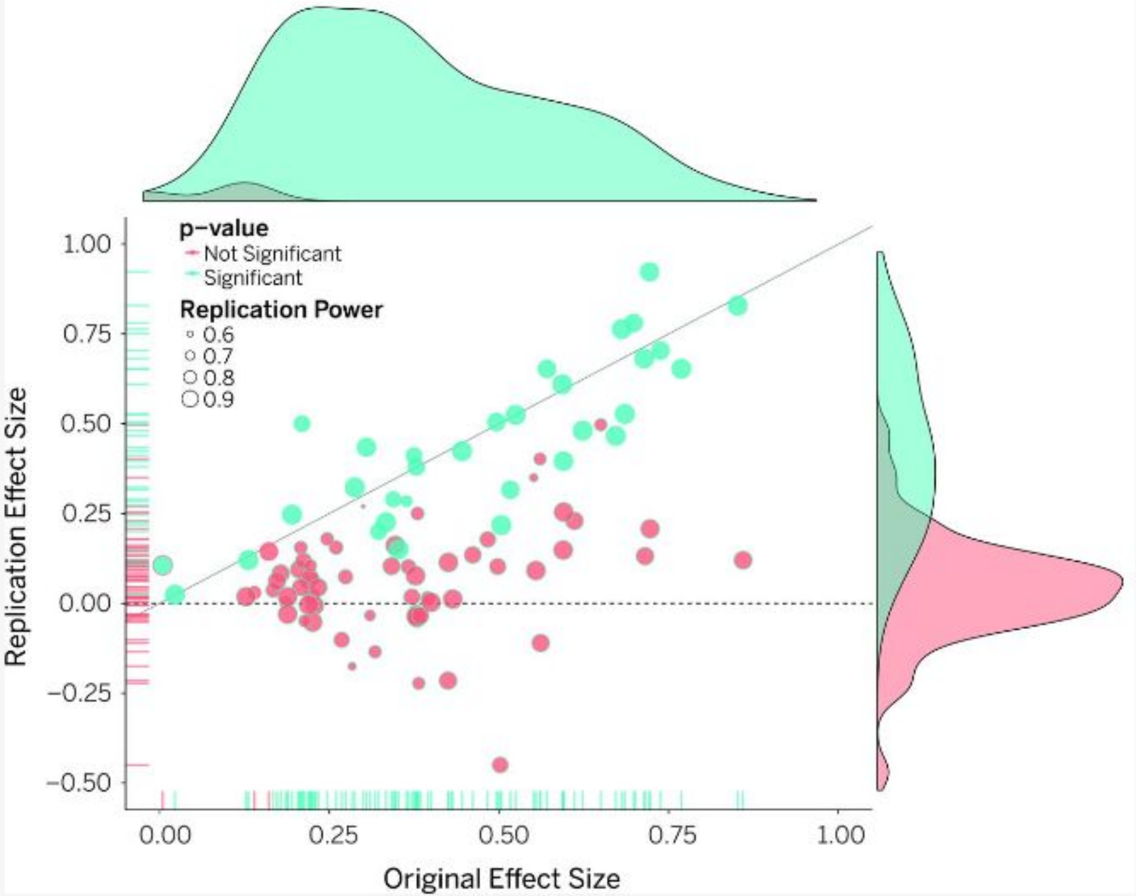
SCIENCE

Psychology's Replication Crisis Is Running Out of Excuses

Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.

ED YONG NOVEMBER 19, 2018

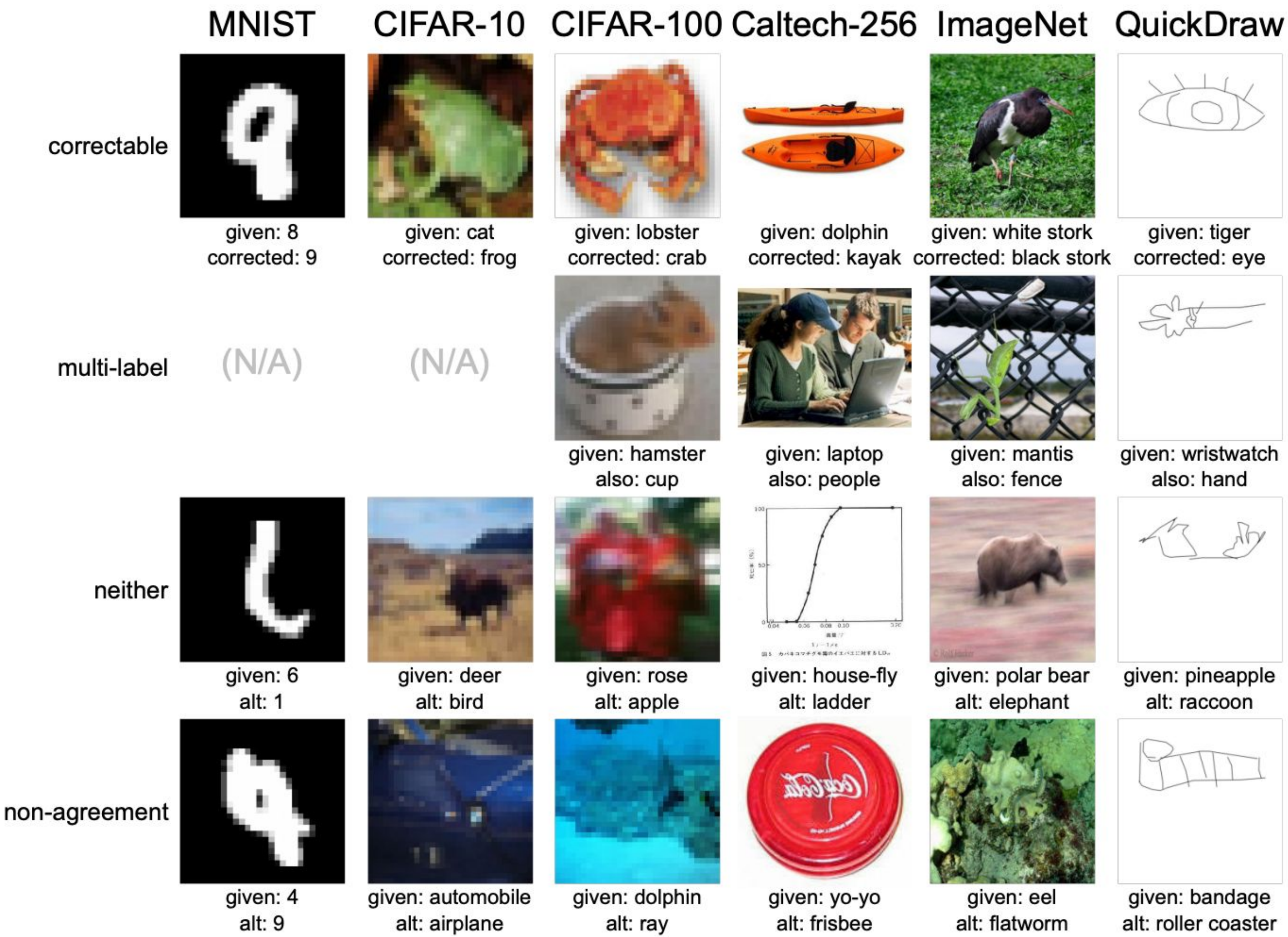
Bad news



Original study effect size versus replication effect size (correlation coefficients).
Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015

#2: Labeling errors



Northcutt et al. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. . NeurIPS 2021

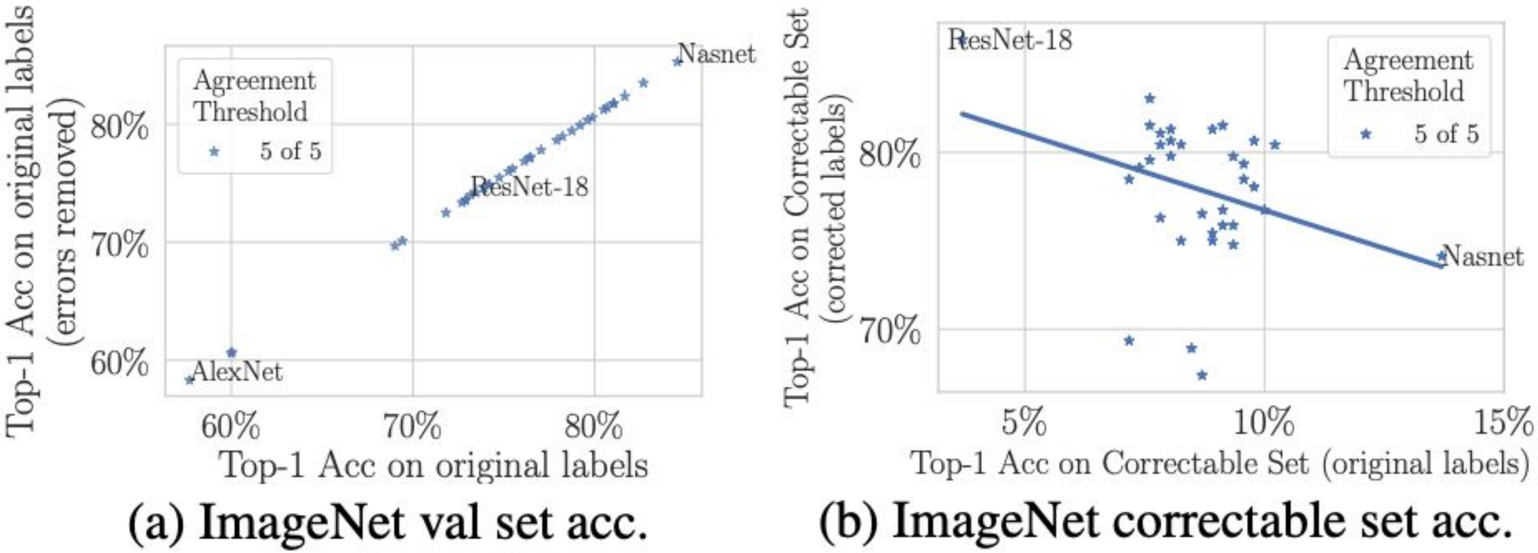
#2: Labeling errors: % errors in test sets

Dataset	Modality	Size	Model	Test Set Errors CL guessed	MTurk checked	validated	estimated	% error
MNIST	image	10,000	2-conv CNN	100	100 (100%)	15	-	0.15
CIFAR-10	image	10,000	VGG	275	275 (100%)	54	-	0.54
CIFAR-100	image	10,000	VGG	2,235	2,235 (100%)	585	-	5.85
Caltech-256†	image	29,780	Wide ResNet-50-2	2,360	2,360 (100%)	458	-	1.54
ImageNet*	image	50,000	ResNet-50	5,440	5,440 (100%)	2,916	-	5.83
QuickDraw†	image	50,426,266	VGG	6,825,383	2,500 (0.04%)	1870	5,105,386	10.12
20news	text	7,532	TFIDF + SGD	93	93 (100%)	82	-	1.09
IMDB	text	25,000	FastText	1,310	1,310 (100%)	725	-	2.90
Amazon Reviews†	text	9,996,437	FastText	533,249	1,000 (0.2%)	732	390,338	3.90
AudioSet	audio	20,371	VGG	307	307 (100%)	275	-	1.35

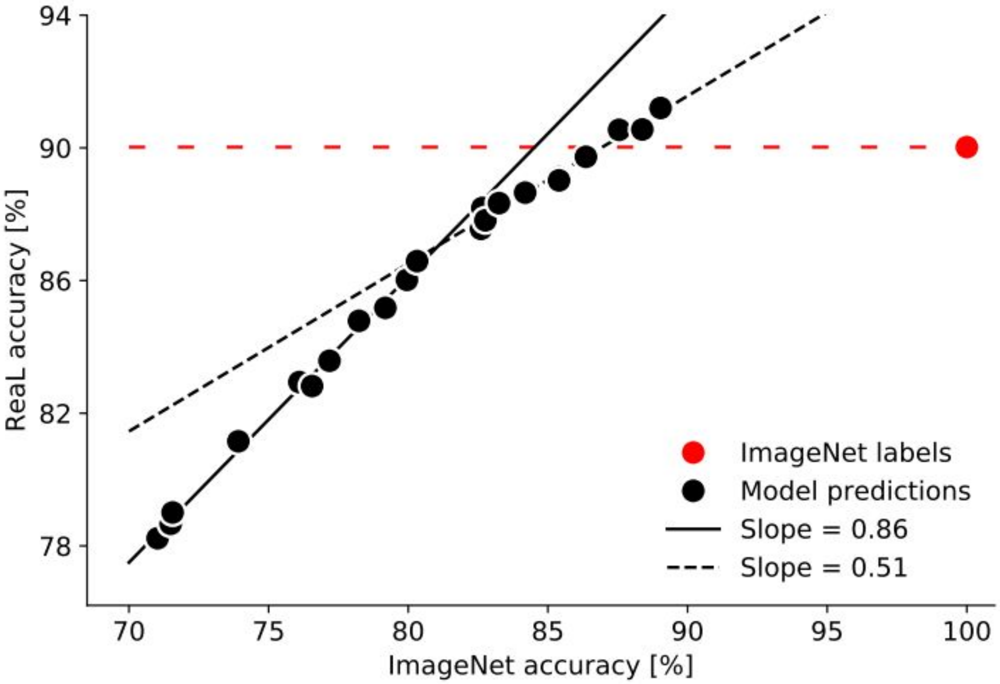
*Because the ImageNet test set labels are not publicly available, the ILSVRC 2012 validation set is used.

Northcutt et al. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. . NeurIPS 2021

#2: Labeling errors: Errors make larger models overfit



#2: Labeling errors: Relabeled ImageNet test set

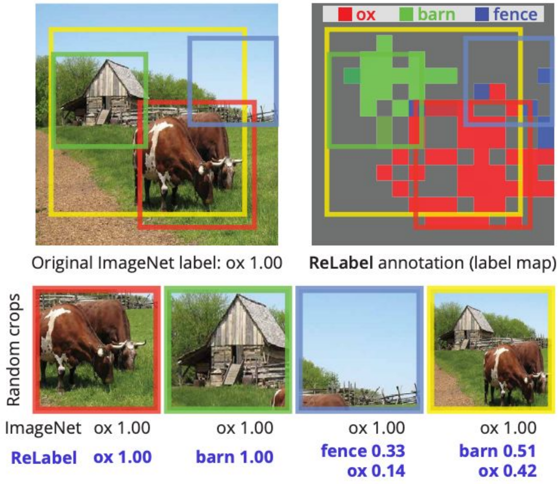


Gains reported using fixed labels is smaller than those with original ImageNet labels is

Beyer et al. Are we done with ImageNet? 2020

#2: Labeling errors: Of course the training set also has errors

Gains reported using fixed labels is smaller than those with original ImageNet labels is



Variants	ImageNet top-1 (%)
ReLabel (localized mutli-labels)	78.9
Localized single labels	78.4 (-0.5)
Global multi-labels	78.5 (-0.4)
Global single labels	77.5 (-1.4)
Original ImageNet labels	77.5 (-1.4)

Yun et al. Re-labeling ImageNet: from Single to Multi-Labels, from Global to Localized Labels. CVPR 2021

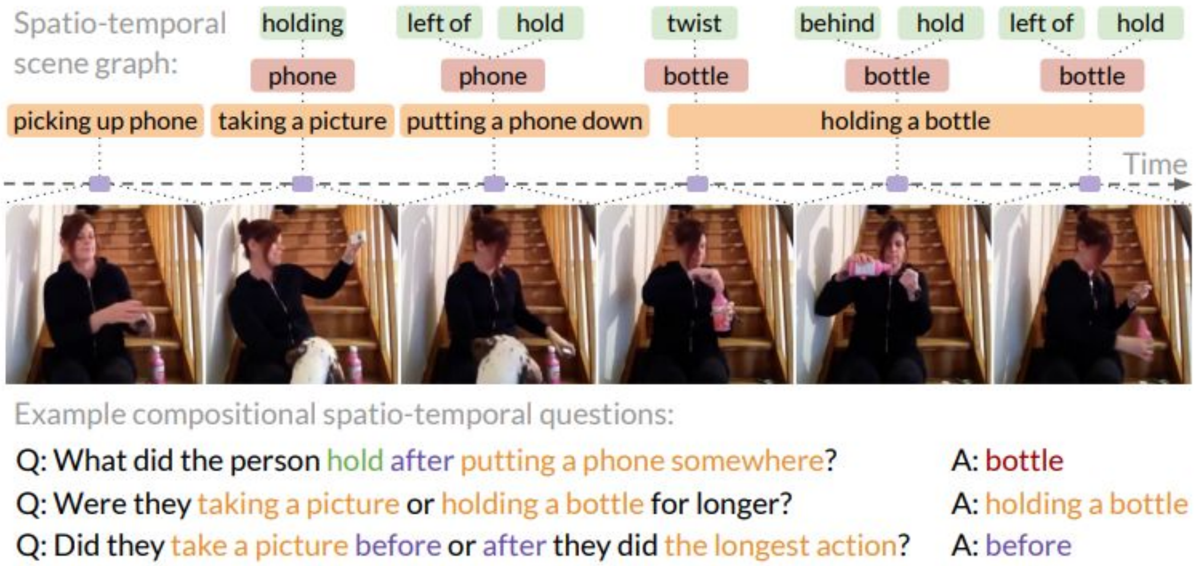
#3: Generalization errors: Test sets represent a small slice of the real world.

Models may have seen:

- people,
- phones,
- bottles,
- people holding bottles

Can they generalize to:

- People holding phones?

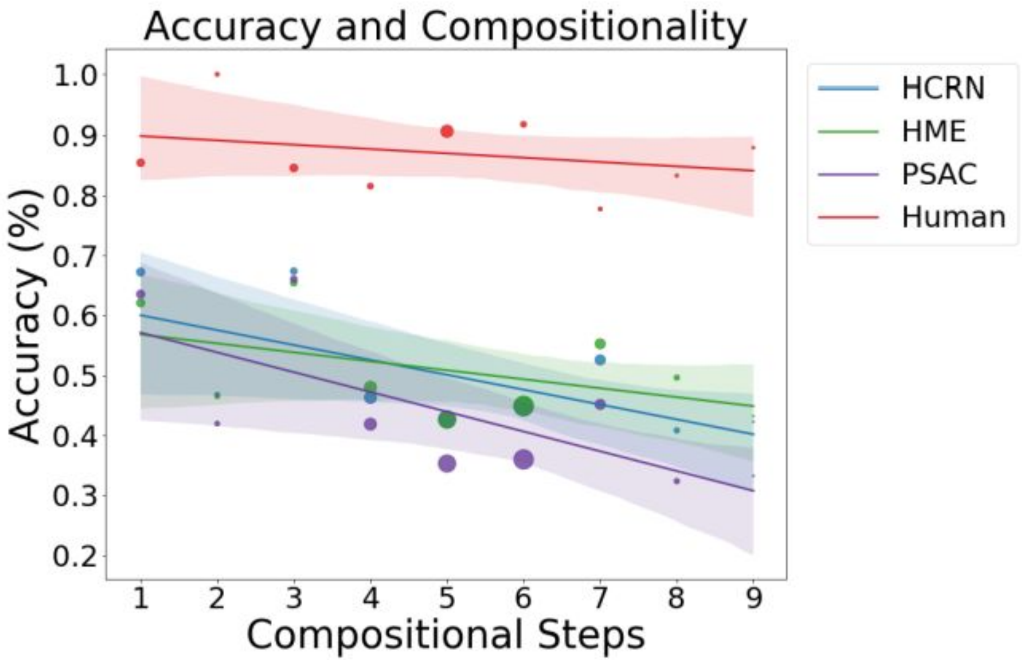


Ranjay Krishna | ranjay@cs.washington.edu

Grunde-McLaughlin et al. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning CVPR 2021

#3: Generalization errors: Systematic generalization in video understanding decreases as composition steps increase

- Human performance: 86%
- Best model performance: 48%



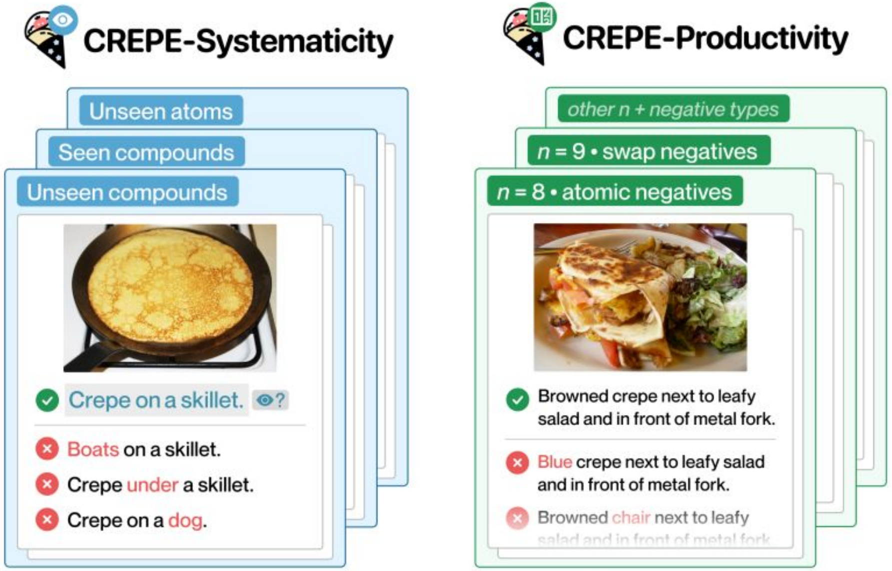
Grunde-McLaughlin et al. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning CVPR 2021

#3: Generalization errors: Maybe videos are too hard… what about images?

CREPE: a benchmark to test for compositional generalization of CLIP and other image-text models

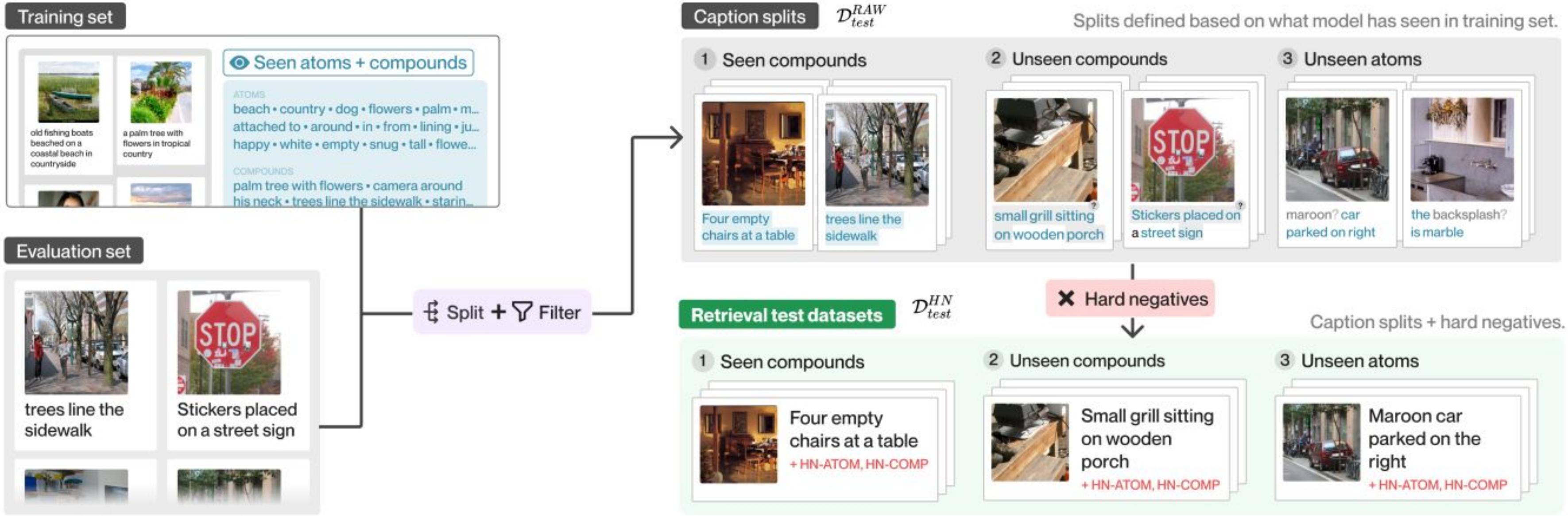
Can models at the very least generalize to new compositions of seens concepts?

Ma et al. CREPE: Can Vision-Language Foundation Models Reason Compositionally? ArXiv 2023



Ranjay Krishna | ranjay@cs.washington.edu

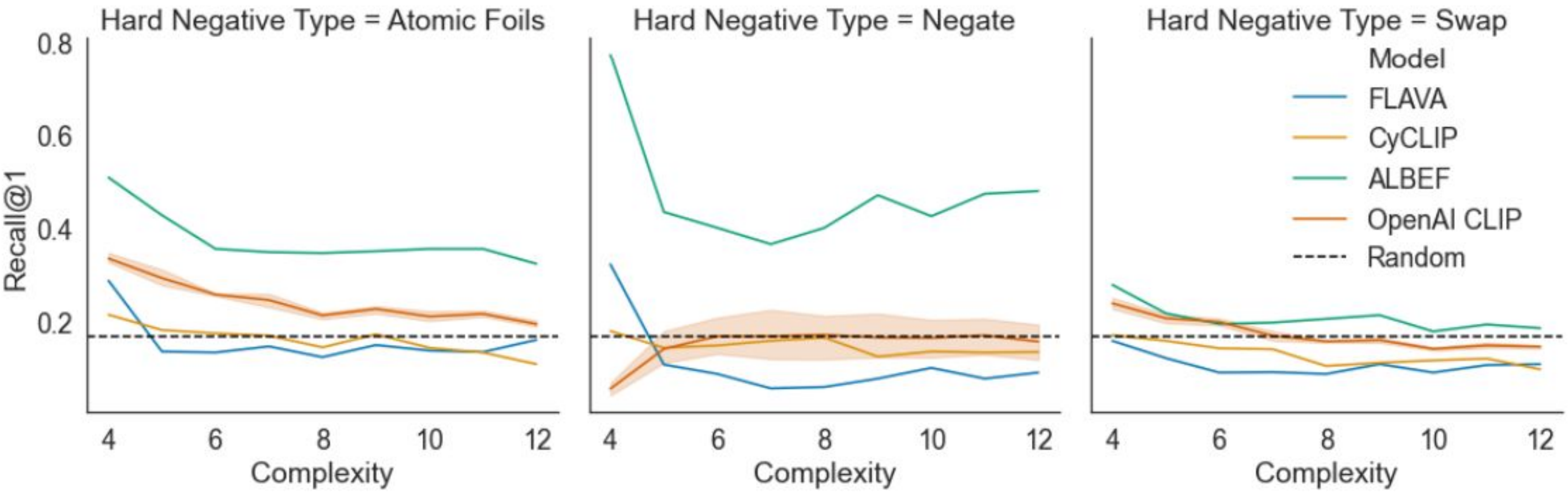
#3: Generalization errors: compositional generalization



Ranjay Krishna | ranjay@cs.washington.edu

Ma et al. CREPE: Can Vision-Language Foundation Models Reason Compositionally? ArXiv 2023

#3: Generalization errors: today's models can't represent composition in language or vision



Ma et al. CREPE: Can Vision-Language Foundation Models Reason Compositionally? ArXiv 2023

#3: Generalization errors: Increasing model size or increasing dataset size doesn't improve compositional generalization

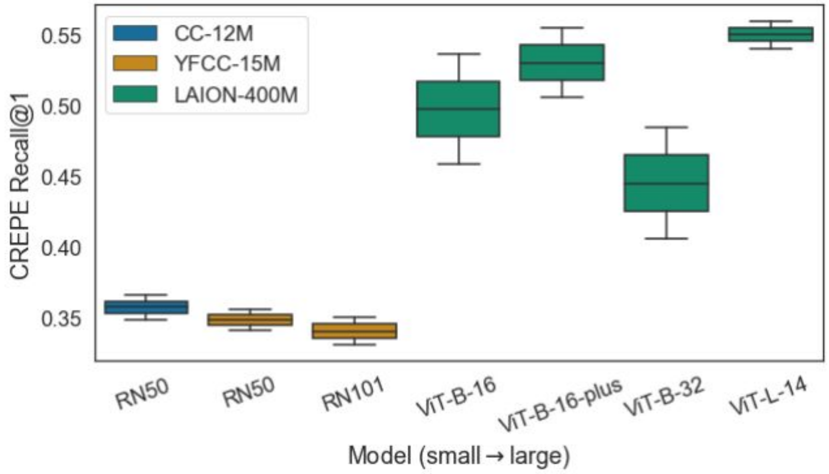


Figure 6. *Systematicity Analysis.* We plot the retrieval Recall@1 of all models pretrained on the three datasets and observe no particular correlation with model size within datasets.

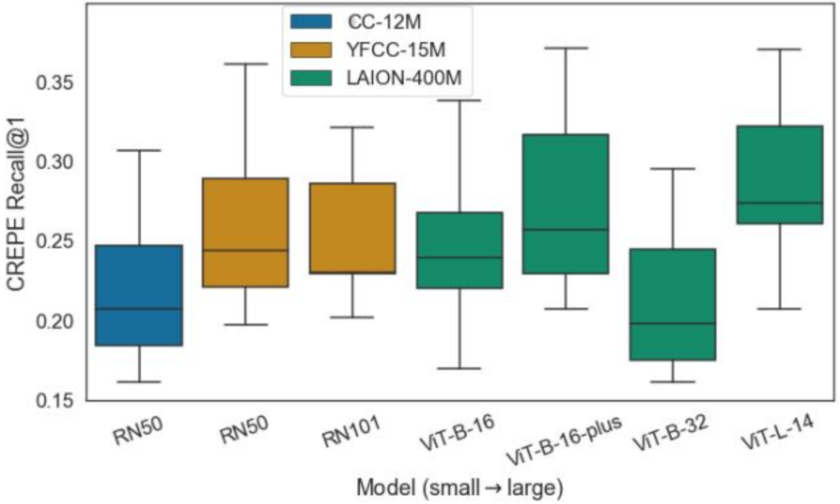
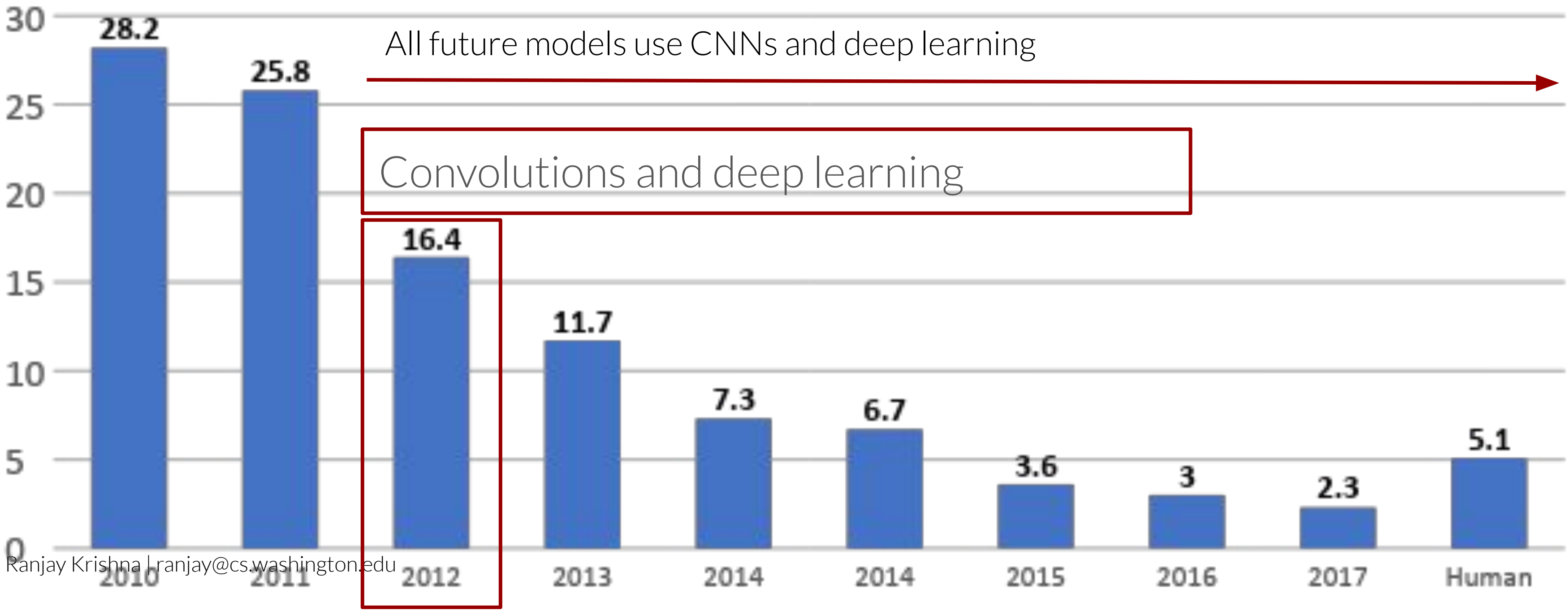


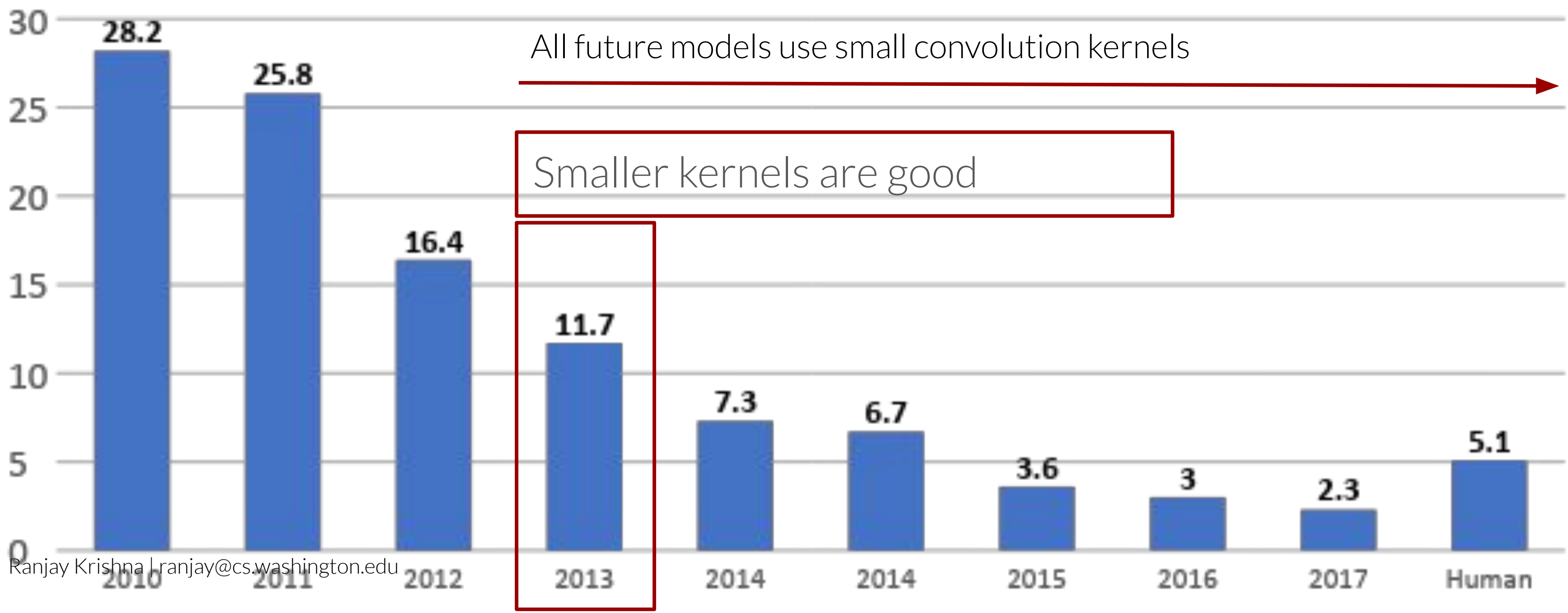
Figure 7. *Productivity Analysis.* We plot the retrieval Recall@1 of all models trained on all three datasets. We observe that there is no consistent correlation with model size within datasets.

Ma et al. CREPE: Can Vision-Language Foundation Models Reason Compositionally? ArXiv 2023

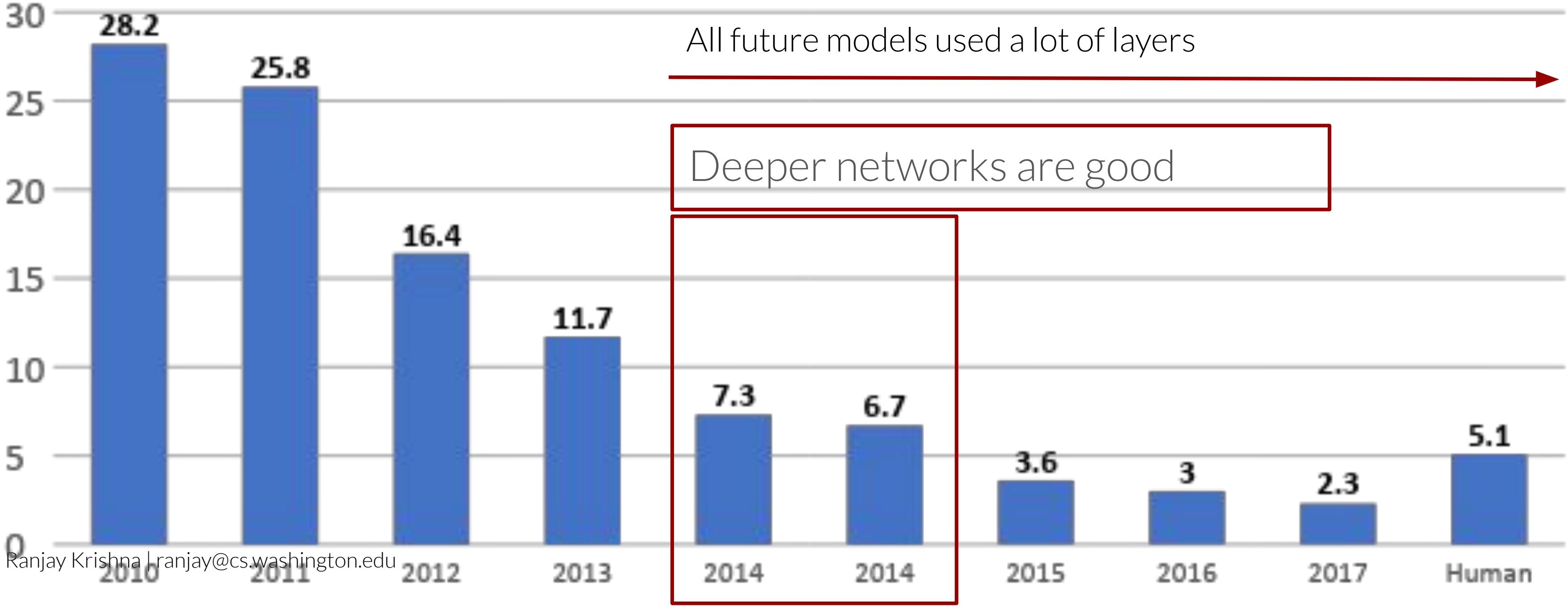
#4: Static test sets: Reusing test sets every year?



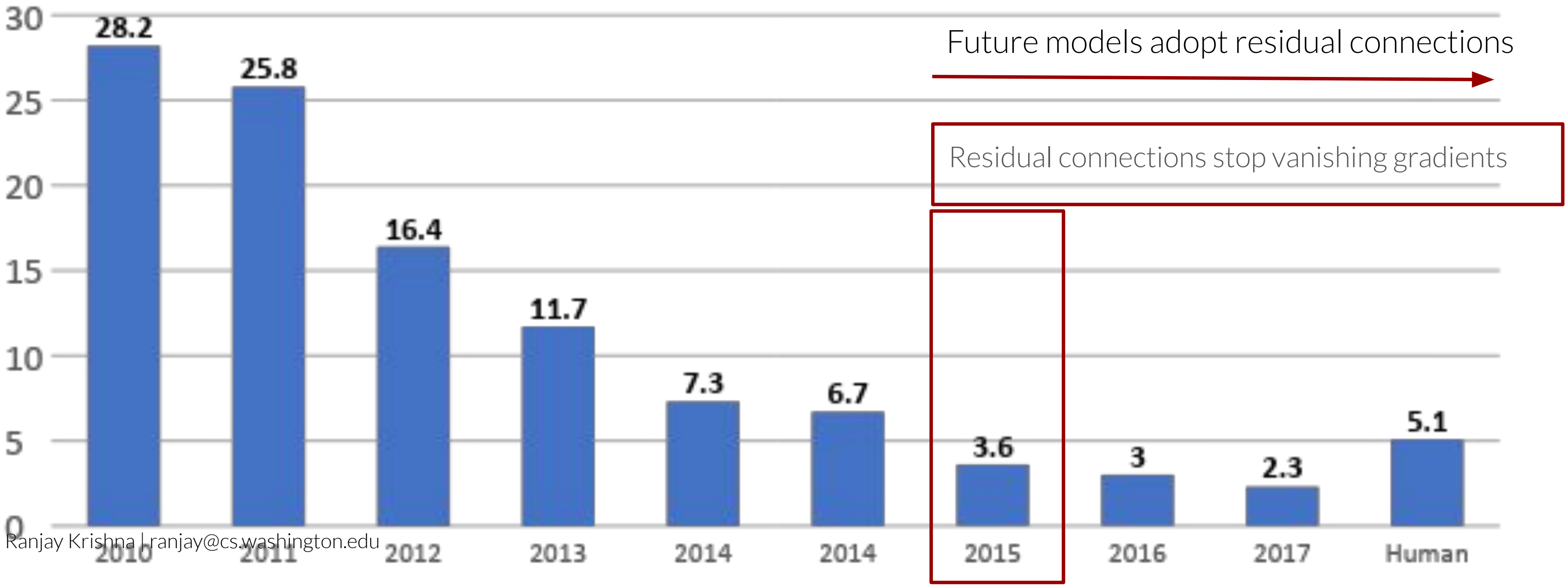
#4: Static test sets: Reusing test sets every year?



#4: Static test sets: Reusing test sets every year?



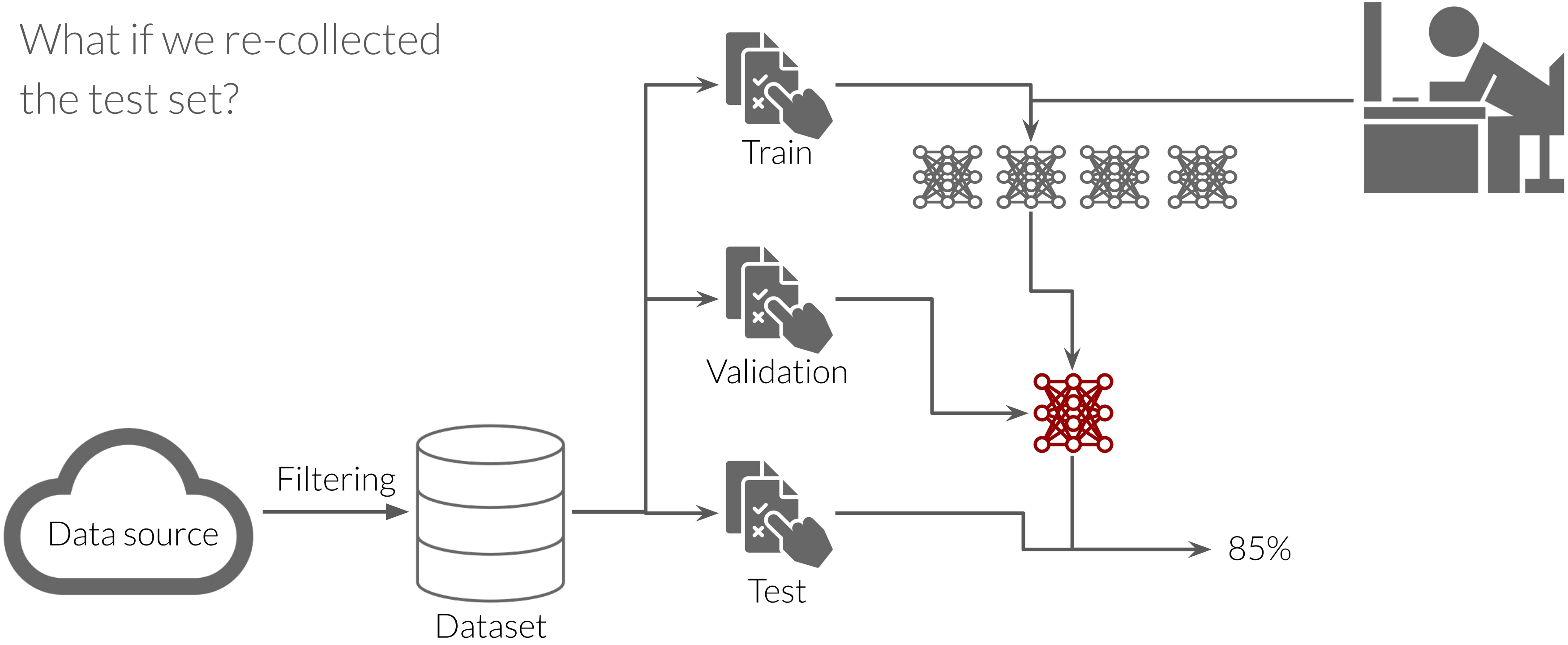
#4: Static test sets: Reusing test sets every year?



#4: A static dataset: Are models overfitting to the test set?

#4: A static dataset: Let's collect a new test set

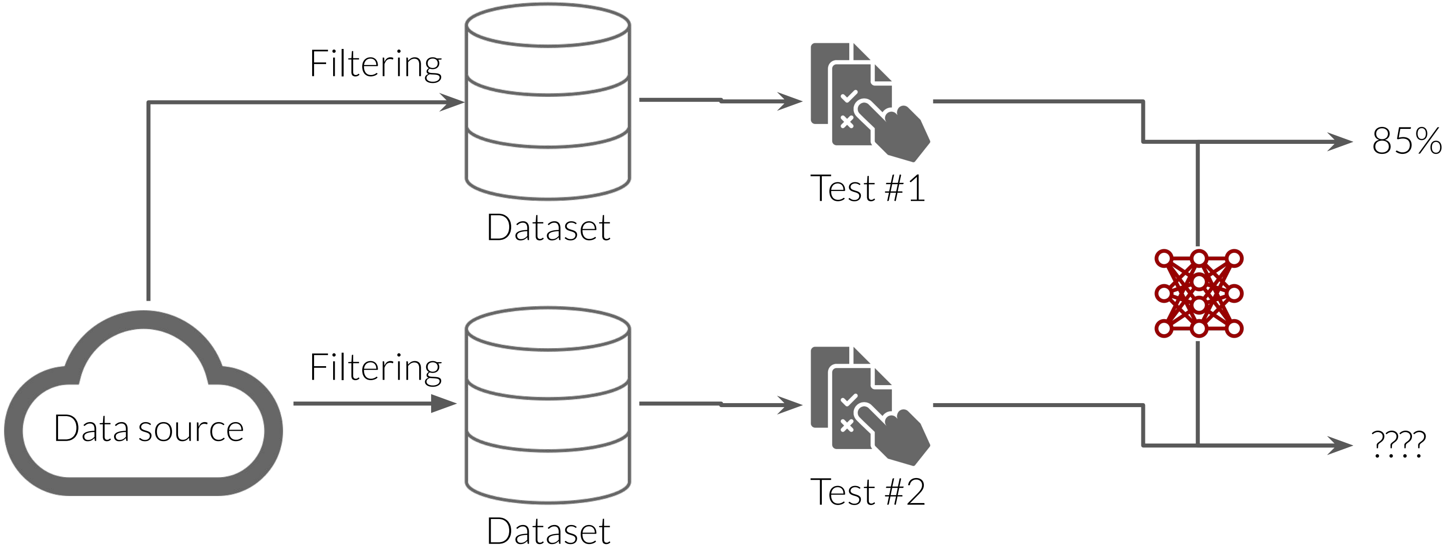
What if we re-collected the test set?



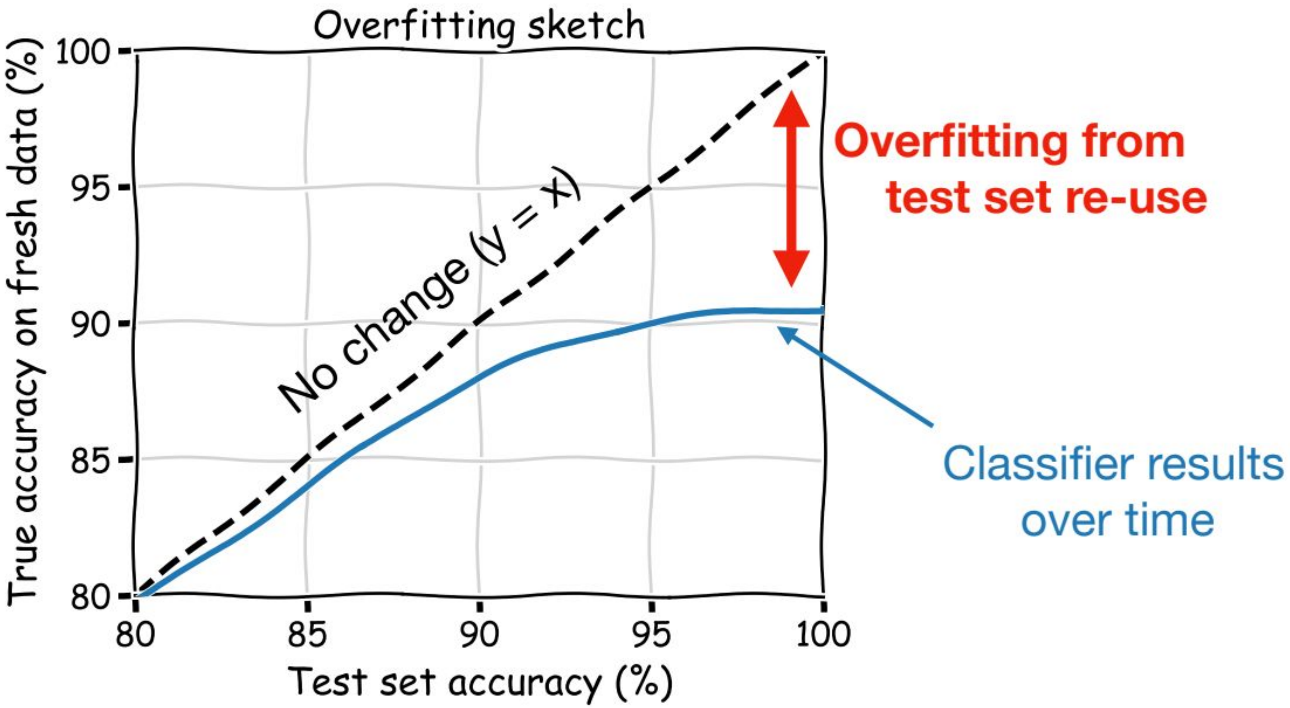
Ranjay Krishna | ranjay@cs.washington.edu

#4: A static dataset: Checking for overfitting

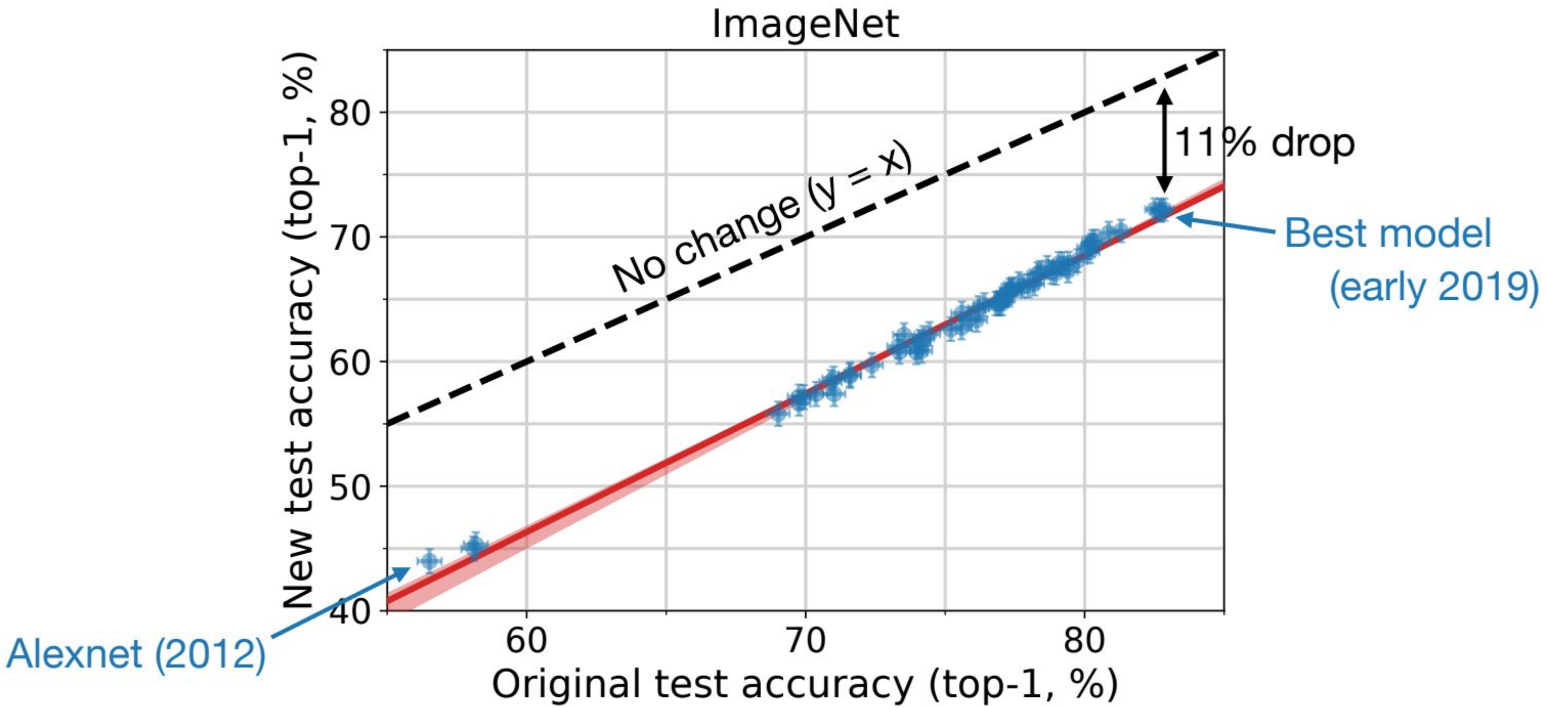
What if we re-collected
the test set?



#4: A static dataset: If models are overfitting to test set



#4: A static dataset: Surprisingly no overfitting



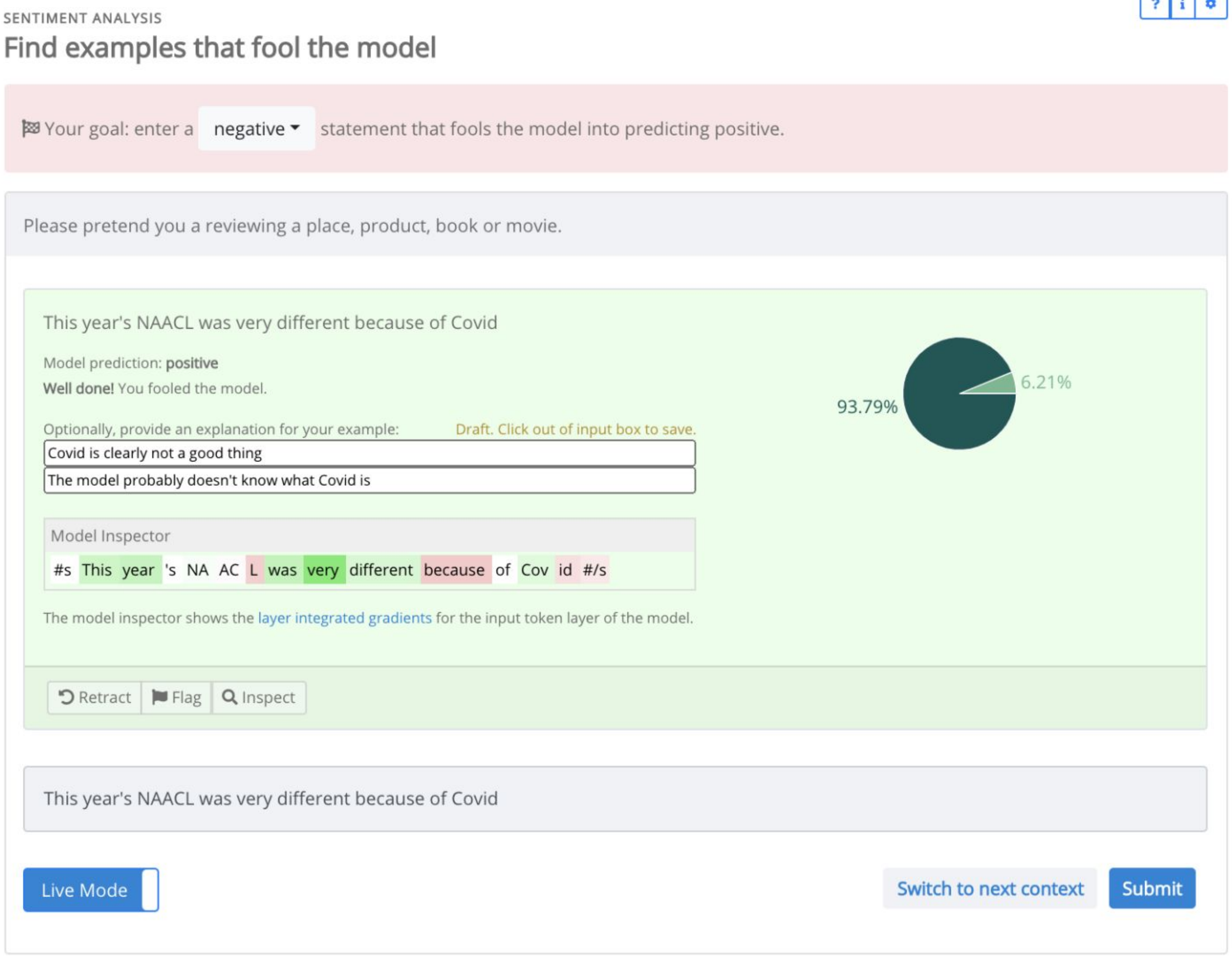
Recht et al. Do imagenet classifiers generalize to imagenet? ICML 2019

#4: A static dataset: creating dynamic benchmarks

Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

Ranjay Krishna | ranjay@cs.washington.edu

Kiela et al. Dynabench: Rethinking Benchmarking in NLP. NAACL 2021



Kiela et al. Dynabench: Rethinking Benchmarking in NLP. NAACL 2021

#4: A static dataset: adversarial training only helps improves performance on adversarial test sets

Adversarially collected training data did not improve model performance

So far, dynamic adversarial testing hasn't resulted in new insights

Kaushik et al. On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study. ArXiv 2021

#4: A static dataset: New guidelines for developing test sets

1. Good performance on the benchmark should imply robust in-domain performance on the task.
   ↪ *We need more work on dataset design and data collection methods.*

2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to remove erroneous examples and to properly handle ambiguous ones.*

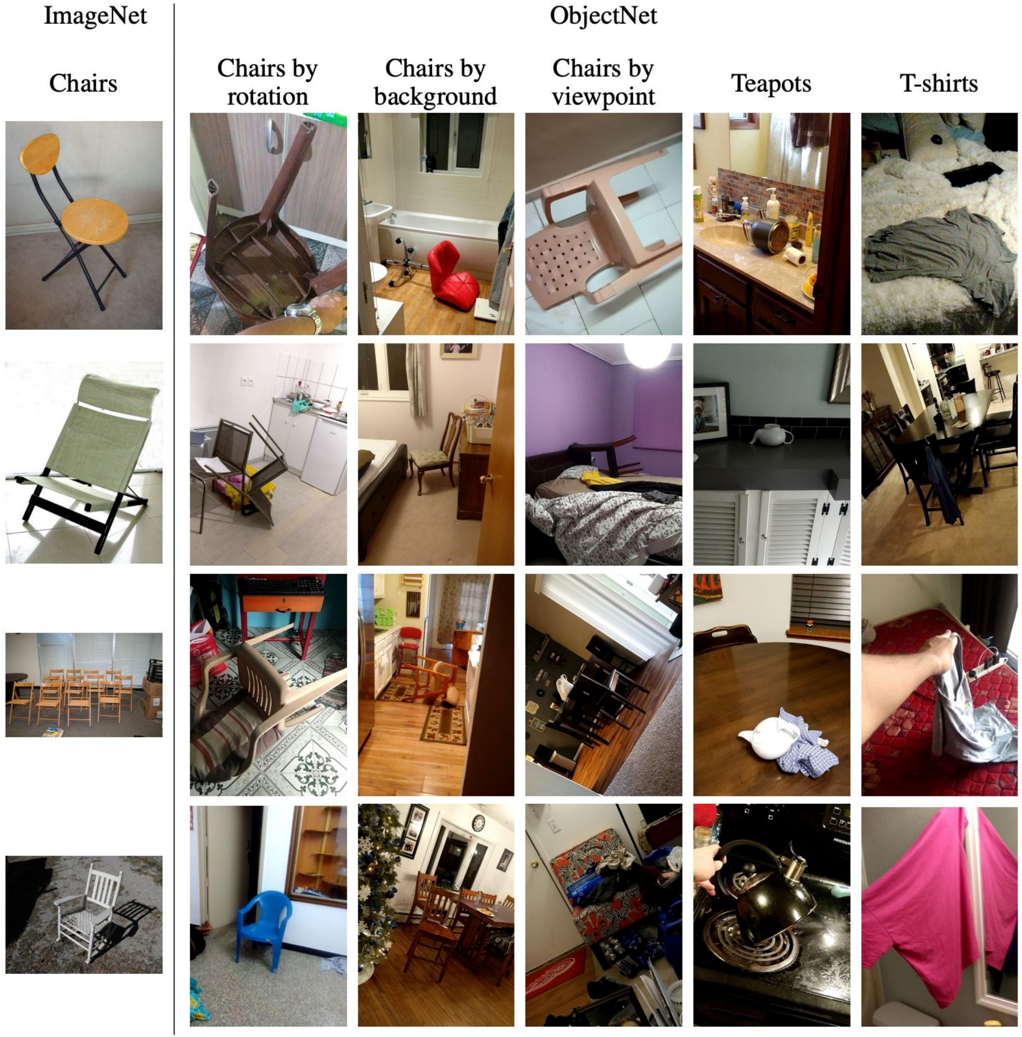
3. Benchmarks should offer adequate statistical power.
   ↪ *Benchmark datasets need to be much harder and/or much larger.*

4. Benchmarks should reveal plausibly harmful social biases in systems, and should not incentivize the creation of biased systems.
   ↪ *We need to better encourage the development and use of auxiliary bias evaluation metrics.*

Bowman et al. What will it take to fix benchmarking in NLP? ArXiv 2021

#5: Distribution shifts

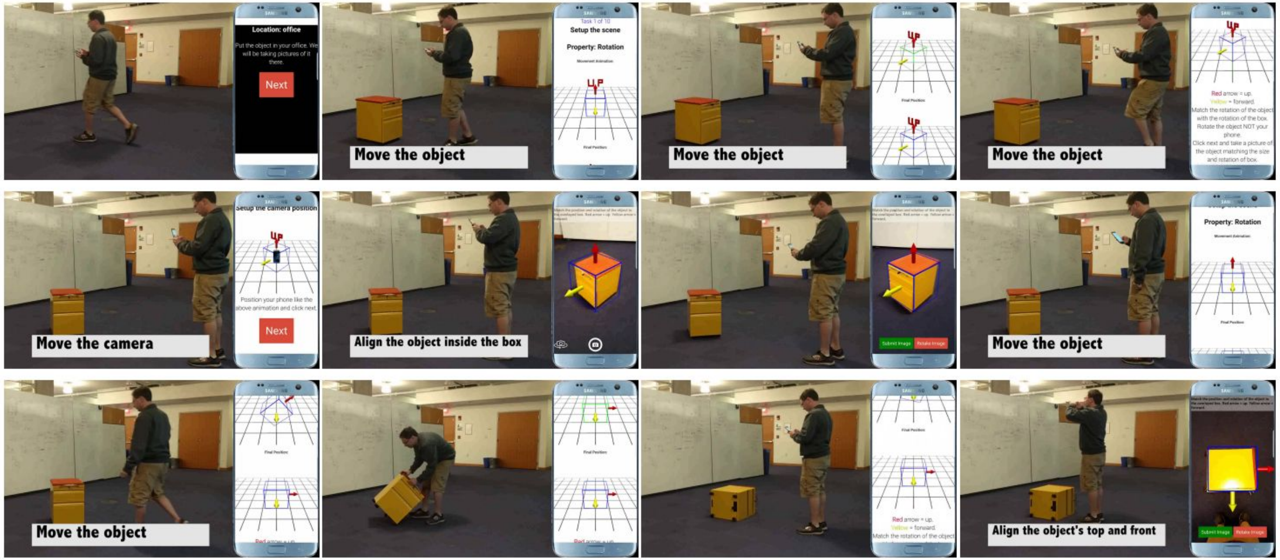
Differences between images in dataset versus images in the real world



Barbu et al. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. NeurIPS 2019

#5: Distribution shifts:

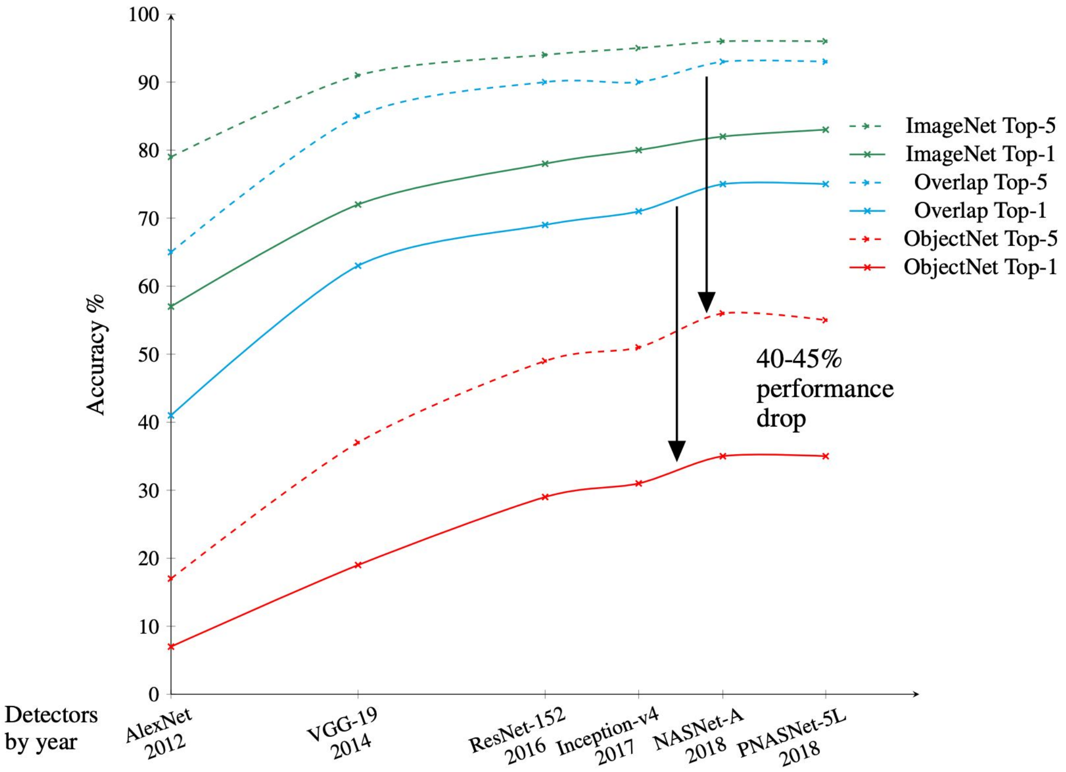
Differences between images in dataset versus images in the real world



Barbu et al. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. NeurIPS 2019

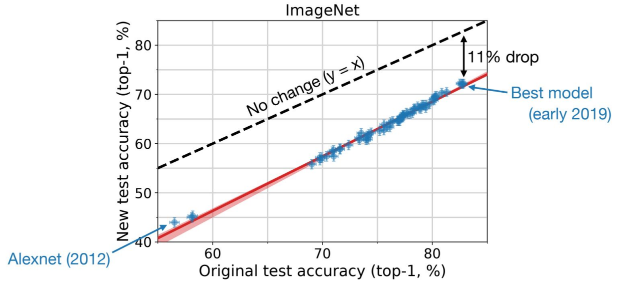
#5: Distribution shifts

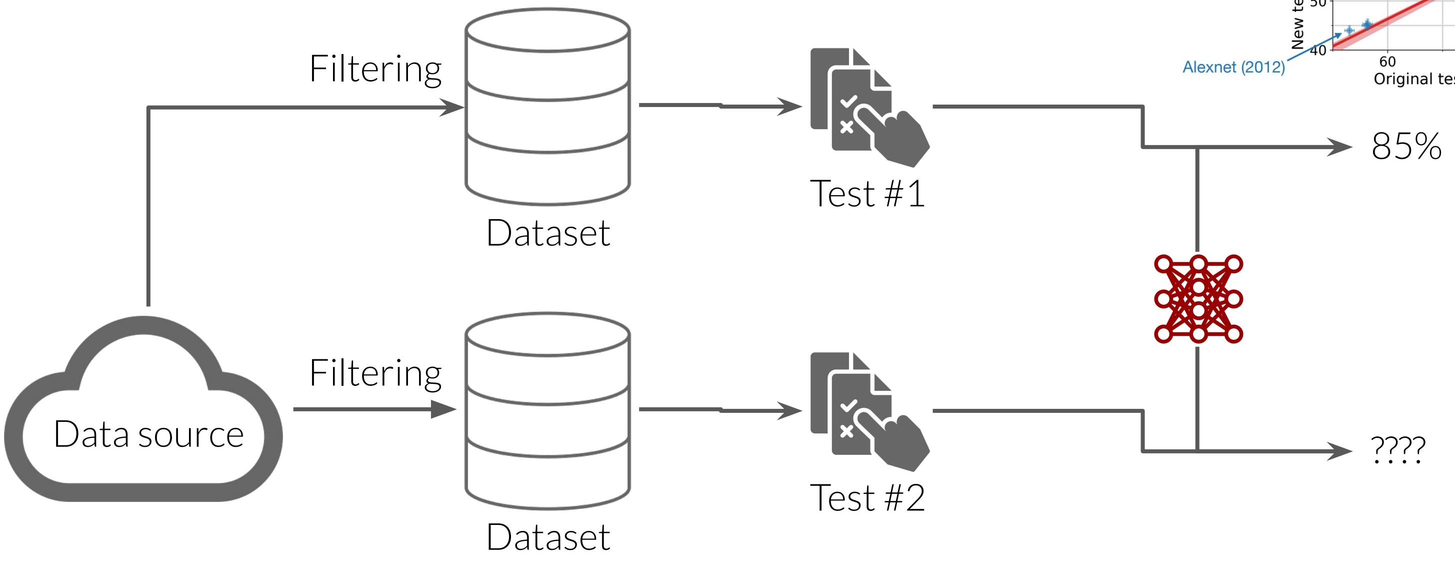
Differences between images in dataset versus images in the real world



Barbu et al. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. NeurIPS 2019

#5: Distribution shifts: in data collection can explain this



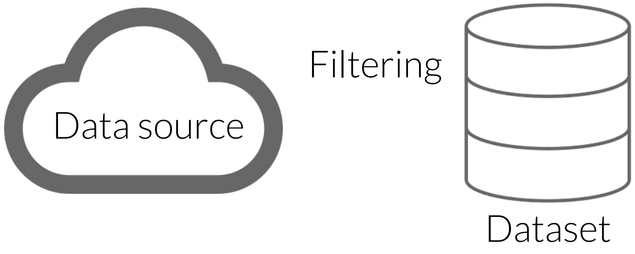


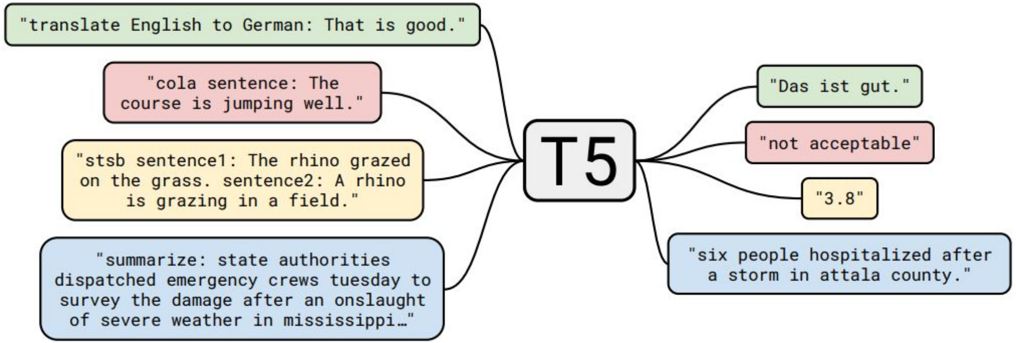
#6: Marginalization: Filtering

T5 trained on Colossal Clean Crawled Corpus

400 words from the [List of filtered words]

- E.g. swastika, white power - implications?
- E.g. twink - implications?

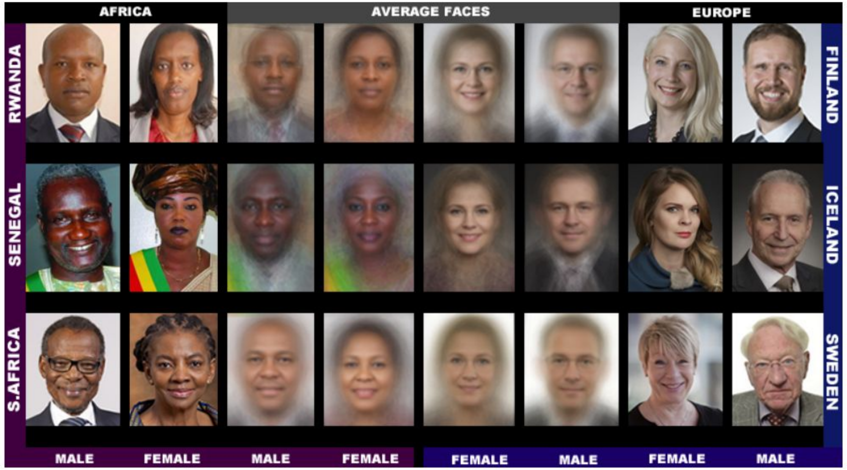




Raffel et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. IJML 2020
Dodge et al. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. ArXiv 2021

#7: Bias in data source

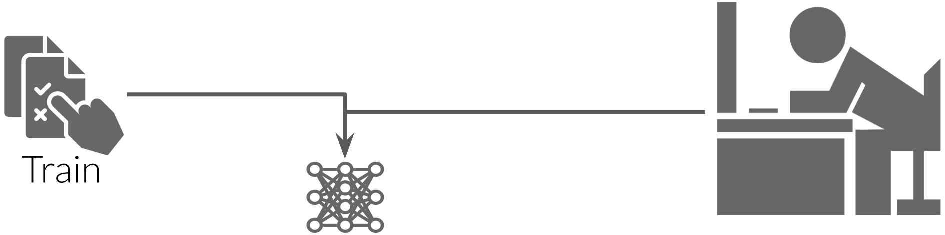
- Then: What was not curated caused bias
- Today: More media coverage = more training data instances



Data source

Buolamwini et al. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAccT 2018
Bender et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT 2021

#8: Environmental and financial costs



Energy for a flight from NY to SF:

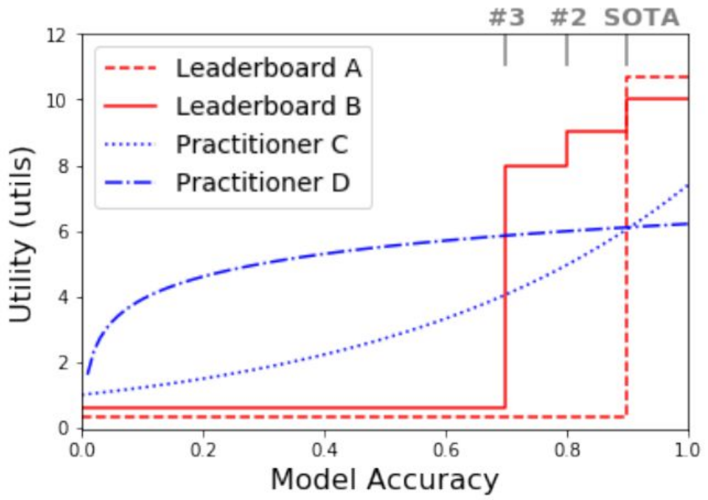
Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Strubell et al. Energy and Policy Considerations for Deep Learning in NLP. ACL 2019

#9: Leaderboard with one metric is not enough

Utility of a new AI model:

- is NOT smooth w.r.t. Accuracy for a leaderboard
- Any improvement along any dimension is good for a practitioner



Ethayarajh et al. Utility is in the Eye of the User: A Critique of NLP Leaderboards. EMNLP 2020

#10: Open ended tasks: Generative models are very hard to evaluate

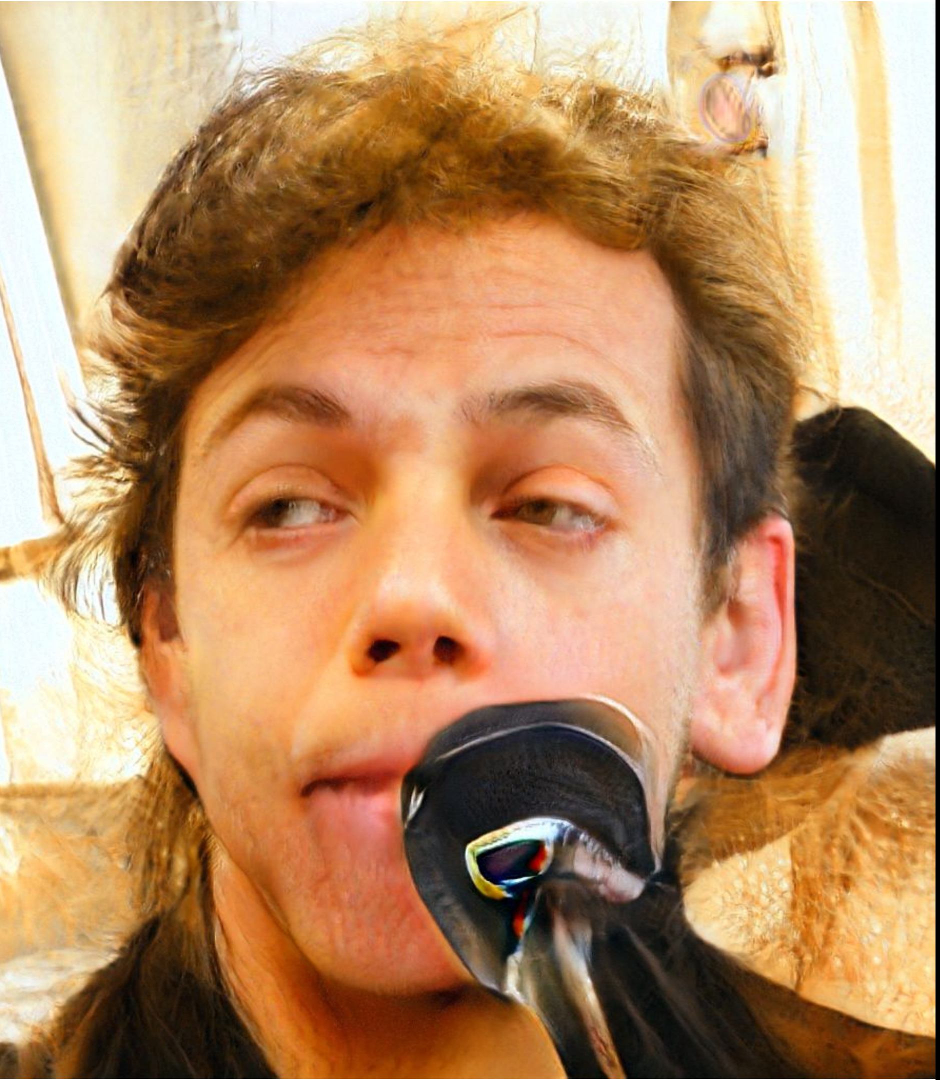
Research question:

How do you evaluate the output of an image generation model?

Zhou et al. HYPE: A Benchmark For Human eYe Perceptual Evaluation of Generative Models. NeurIPS 2019









It used to be easy to measure progress

Ian Goodfellow @goodfellow_ian

Goodfellow, I. J., et al. "Generative Adversarial Networks." (2014).
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." (2015).
Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." (2016).
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." (2017).
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." (2019).

It’s much harder now

Ian Goodfellow @goodfellow_ian

Goodfellow, I. J., et al. "Generative Adversarial Networks." (2014).
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." (2015).
Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." (2016).
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." (2017).
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." (2019).

We don't even have corresponding pairs

2014

2015

2016

2017

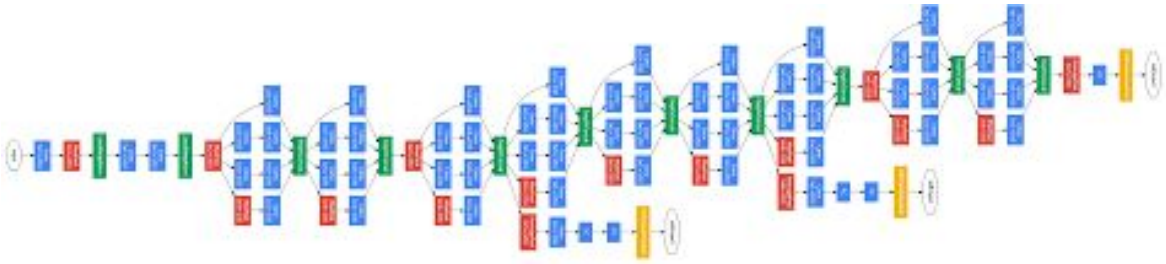
2018

Ian Goodfellow @goodfellow_ian

Goodfellow, I. J., et al. "Generative Adversarial Networks." (2014).
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." (2015).
Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." (2016).
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." (2017).
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." (2019).

How are models evaluated today?

Inception score, FID.

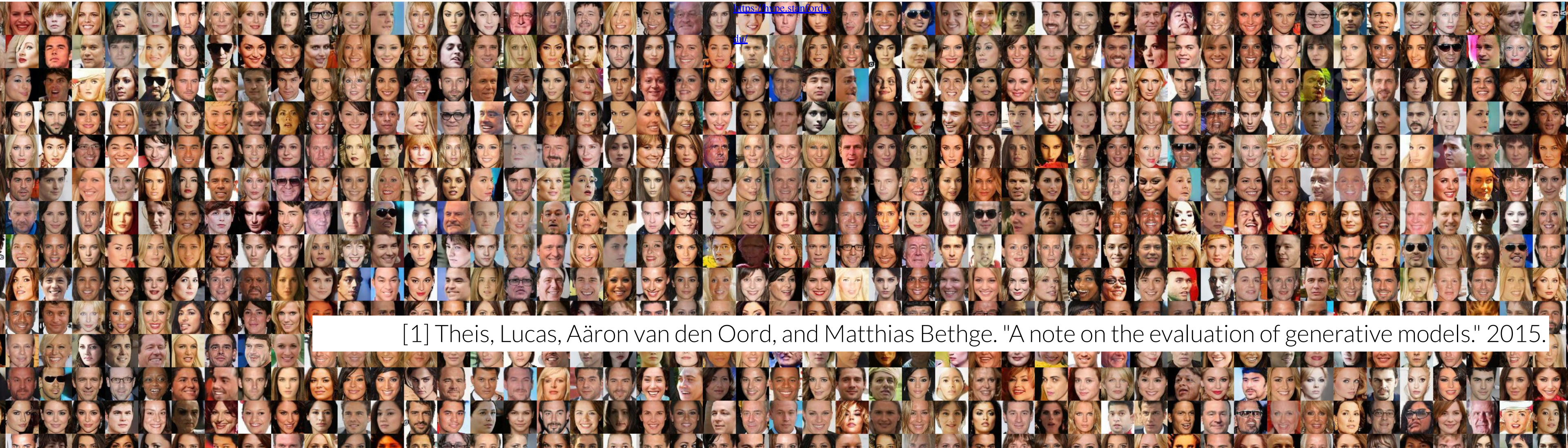


- Trained on imagenet
- **Inception score** is maximized when entropy of predicted output is low
  - Meaning if Inception says with high certainty that it's a "person", the score will be higher
- **FID** calculates distributions from activations of an Inception-v3 layer
- What is the problem with this approach?

Why not use automated metrics?

Why not use automated metrics?

Density estimation has even been shown to be misleading [1].



https://hype.stanford.edu/

[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.

Why not use automated metrics?

Density estimation has even been shown to be misleading [1].

Automated evaluation metrics on sampled outputs (Inception Score [2], FID [3], Precision [4], etc.) rely on ImageNet embeddings.

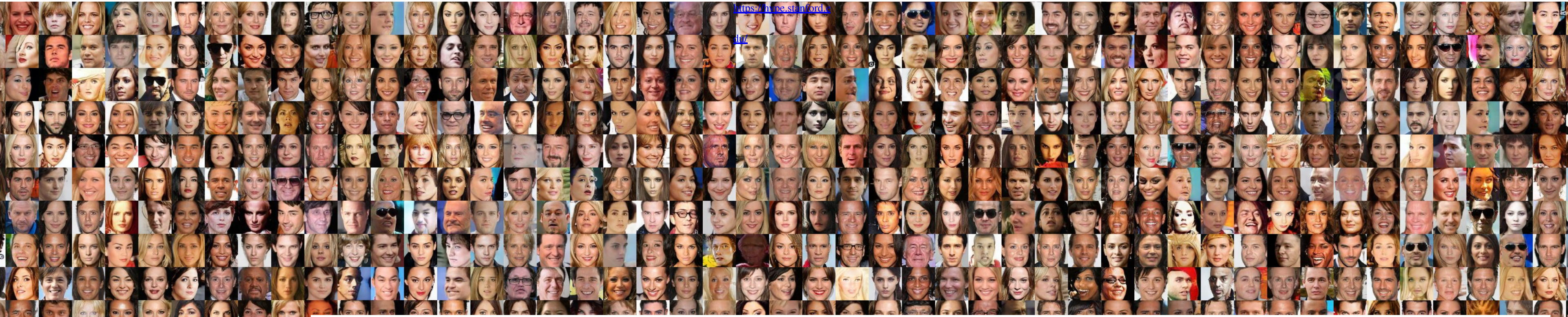
[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.
[2] Salimans, Tim, et al. "Improved techniques for training GANs." 2016.
[3] Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." 2017.
[4] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." 2018.

APA

Why not use automated metrics? Or human metrics?

Density estimation has even been shown to be misleading [1].

Automated evaluation metrics on sampled outputs (Inception Score [2], FID [3], Precision [4], etc.) rely on ImageNet embeddings.



[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.
[2] Salimans, Tim, et al. "Improved techniques for training GANs." 2016.
[3] Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." 2017.
[4] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." 2018.

Why not use automated metrics? Or human metrics?

Density estimation has even been shown to be misleading [1].

Automated evaluation metrics on sampled outputs (Inception Score [2], FID [3], Precision [4], etc.) rely on ImageNet embeddings.

Human evaluation metric are ad-hoc — unreliable and costly.

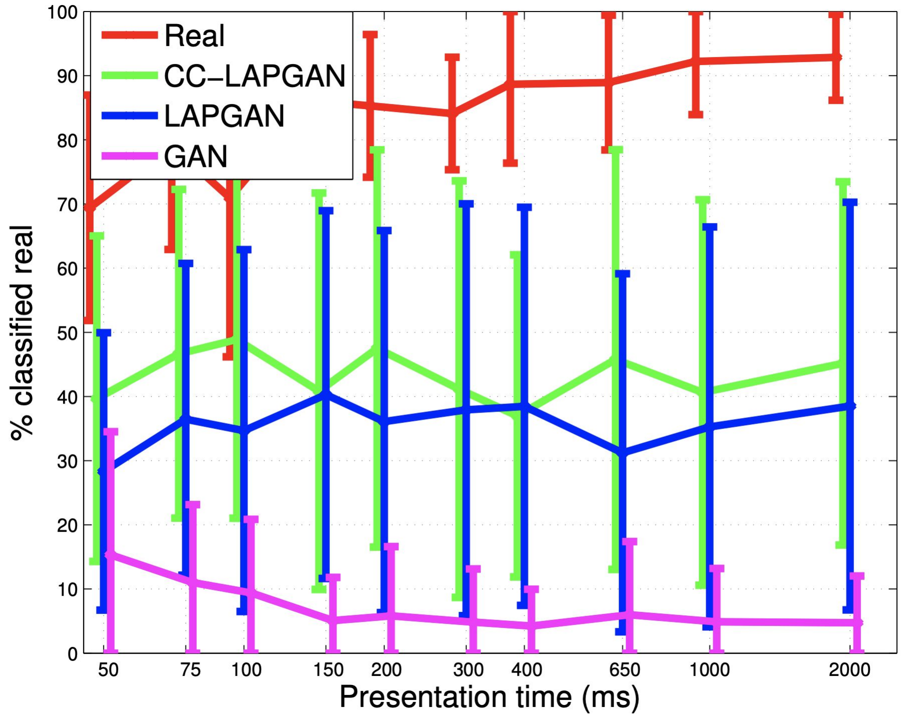
https://hype.stanford.edu/

[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.
[2] Salimans, Tim, et al. "Improved techniques for training GANs." 2016.
[3] Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." 2017.
[4] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." 2018.

APA

Why not use human evaluation?

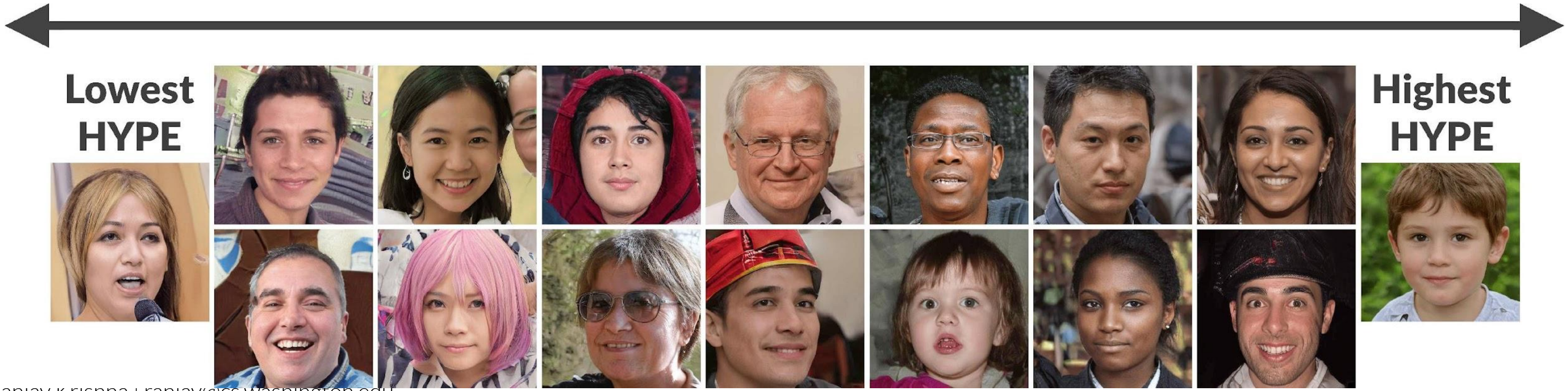
1. Ad-hoc, each executed in idiosyncrasy without proof of reliability or grounding to theory.
2. High variance in their estimates.
3. Lack clear separability between models.
4. Expensive and time-consuming



HYPE measures this progress using human evaluation that is consistent, efficient, and grounded in theory

HYPE is designed to address these problems:

1. **Grounded** method inspired by psychophysics methods in perceptual psychology.
2. **Reliable** and consistent estimator.
3. Statistically **separable** to enable a comparative ranking.
4. Cost and time **efficient**.



Psychophysics method: adaptive staircase procedure

- Staircase methods can determine human perceptual thresholds efficiently and reliably (Cornsweet, 1962).

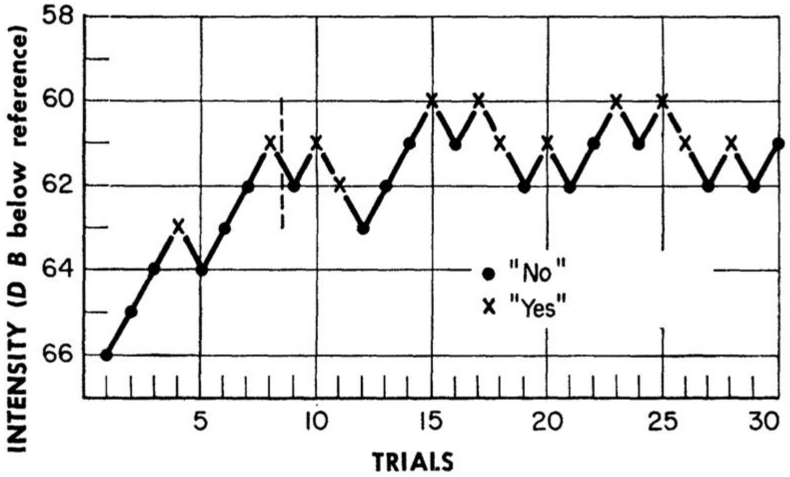
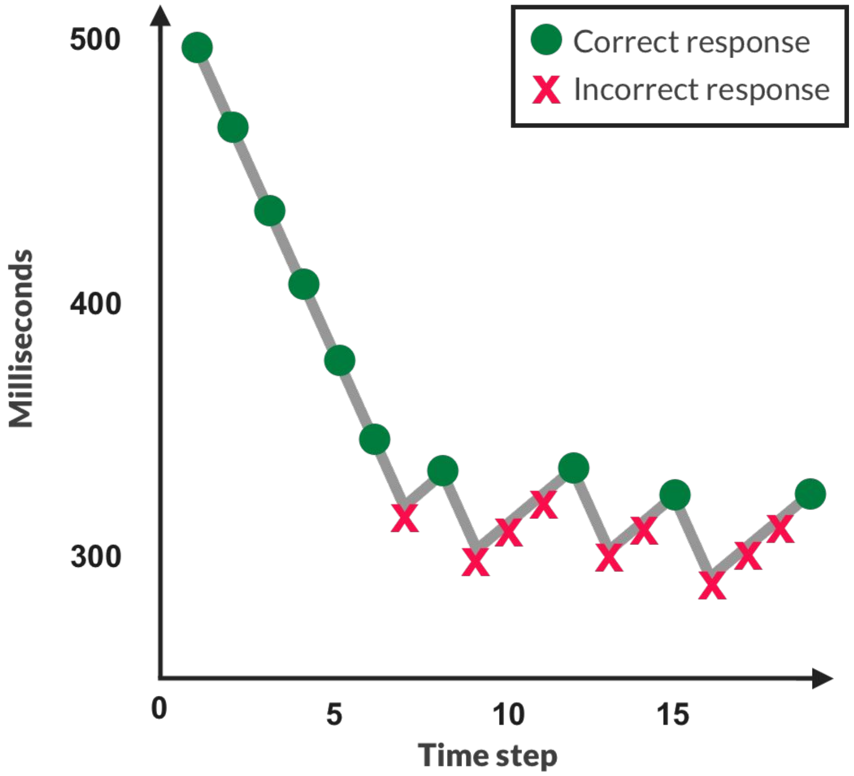


FIG. 1. DATA FROM THE DETERMINATION OF A TYPICAL AUDITORY THRESHOLD BY THE STAIRCASE-METHOD

HYPE: adaptive staircase procedure

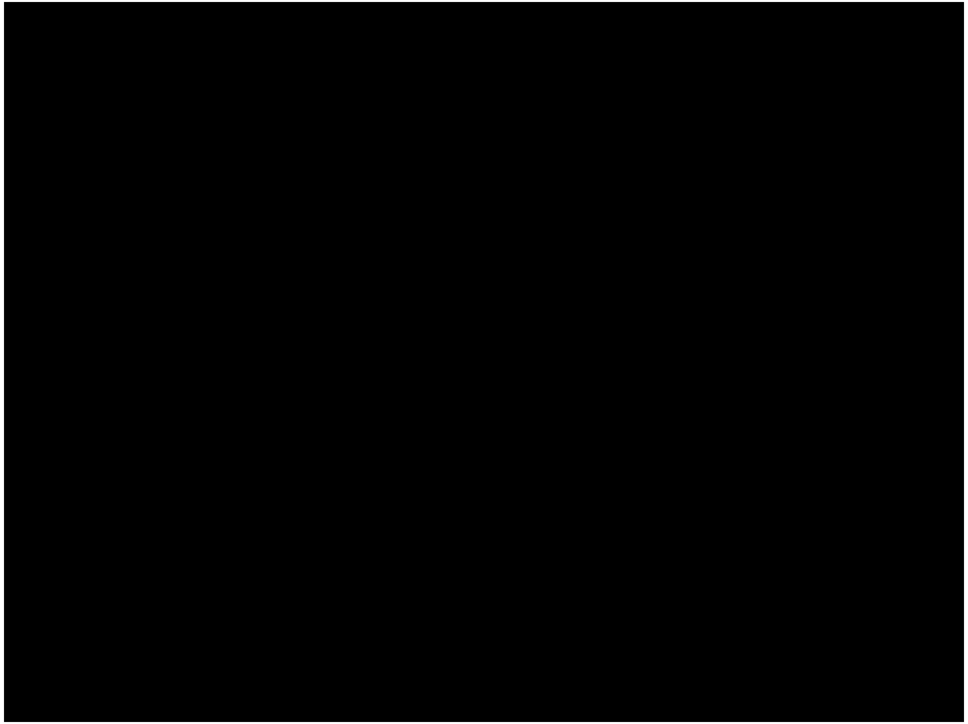


Time: 375ms

real

or

fake

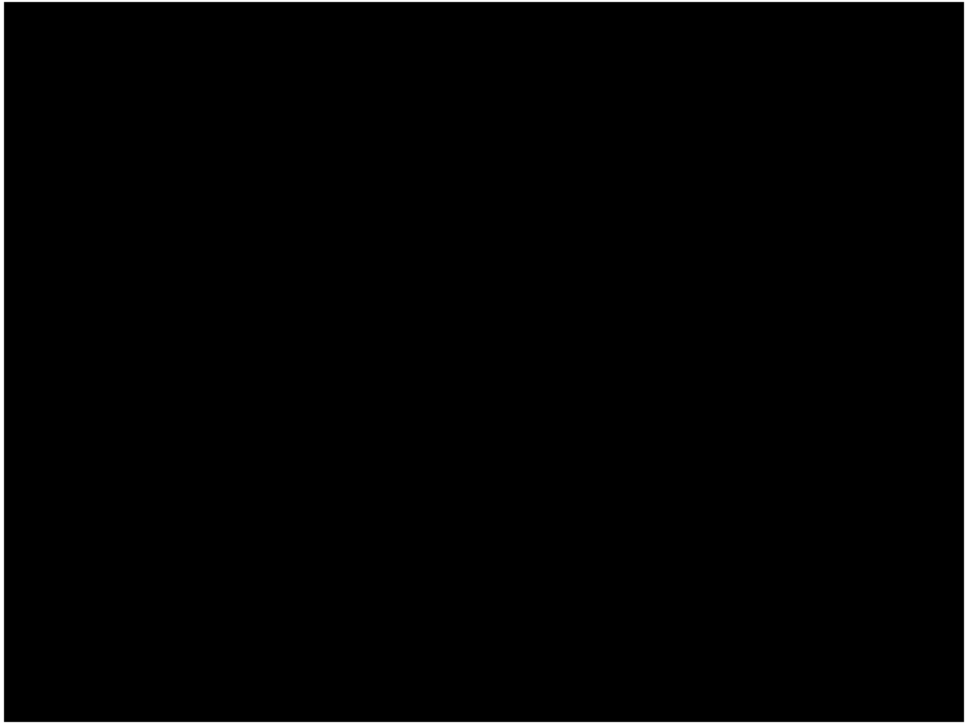


Time: 500ms

real

or

fake

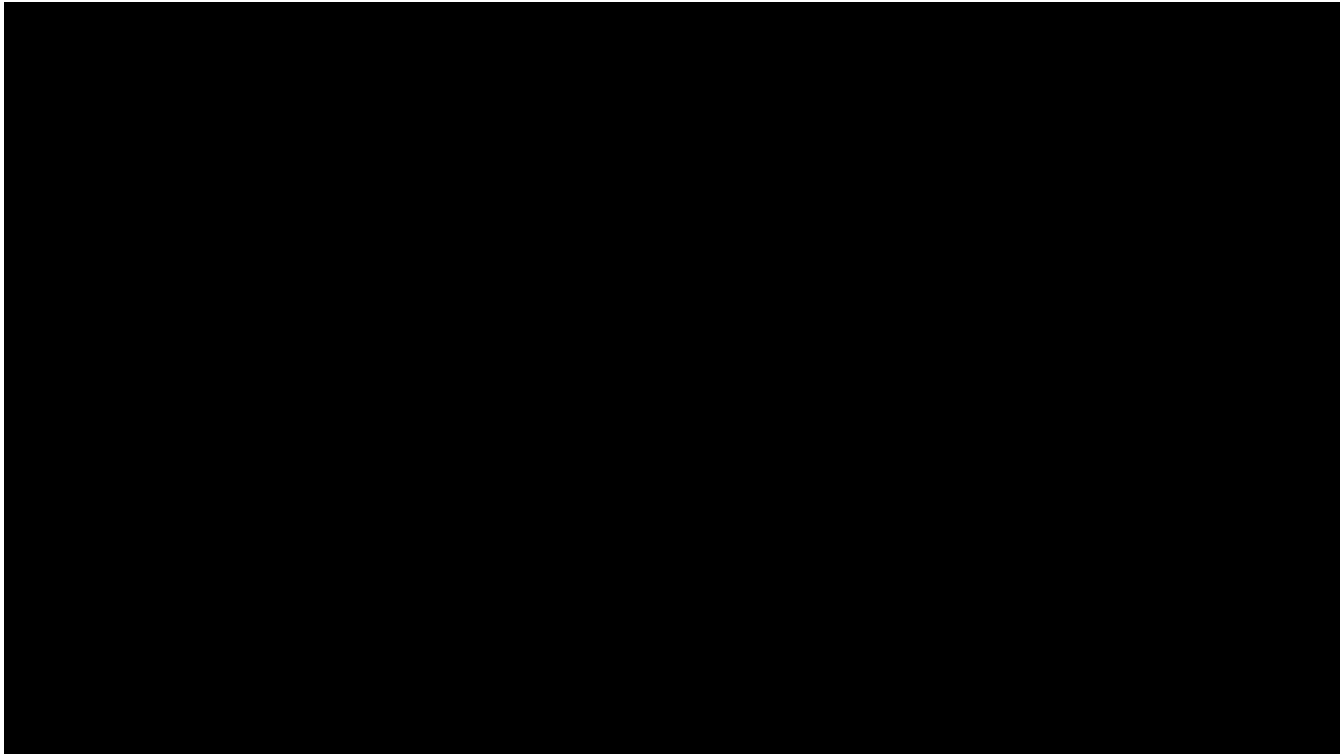


Time: 250ms

real

or

fake



Ranjay Krishna | ranjay@cs.washington.edu

Time: 125ms

real

or

fake

Ranjay Krishna | ranjay@cs.washington.edu

Creating a reliable score

To ensure reliability, we need to:

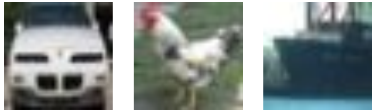
1. Hire and train/filter a sufficient number of evaluators.
2. Sample sufficient outputs.
3. Aggregate.

Experiments

Datasets

- CelebA

- CIFAR-10

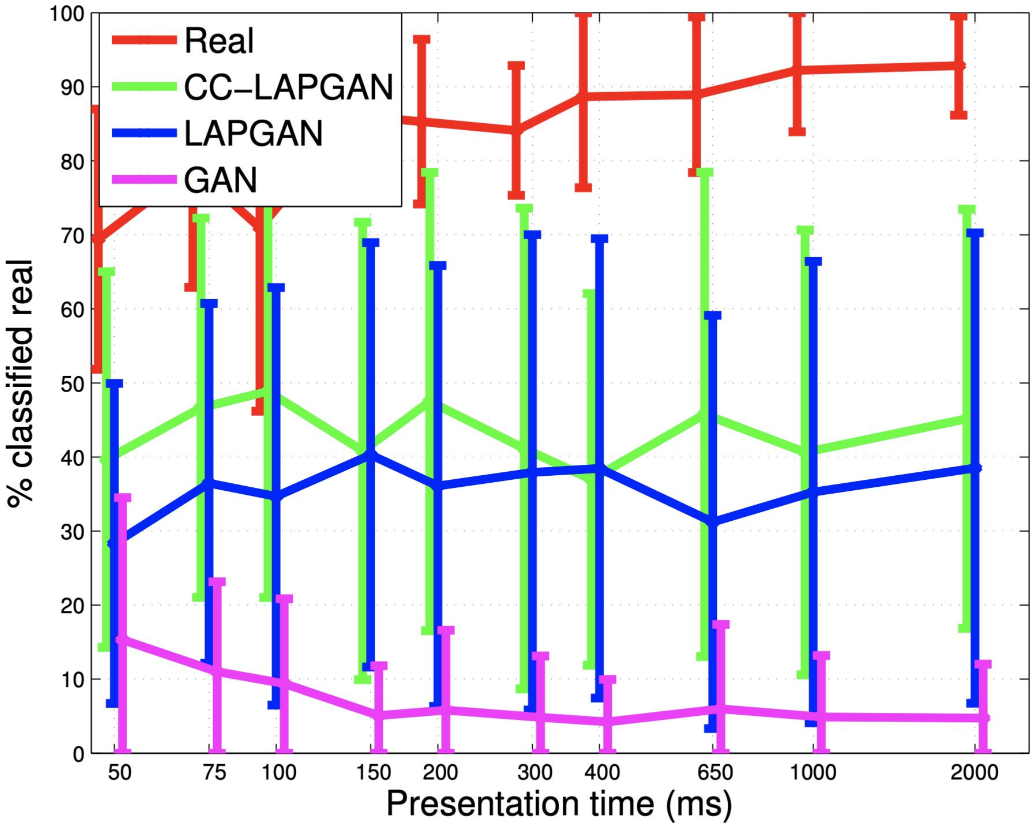


- ImageNet-5

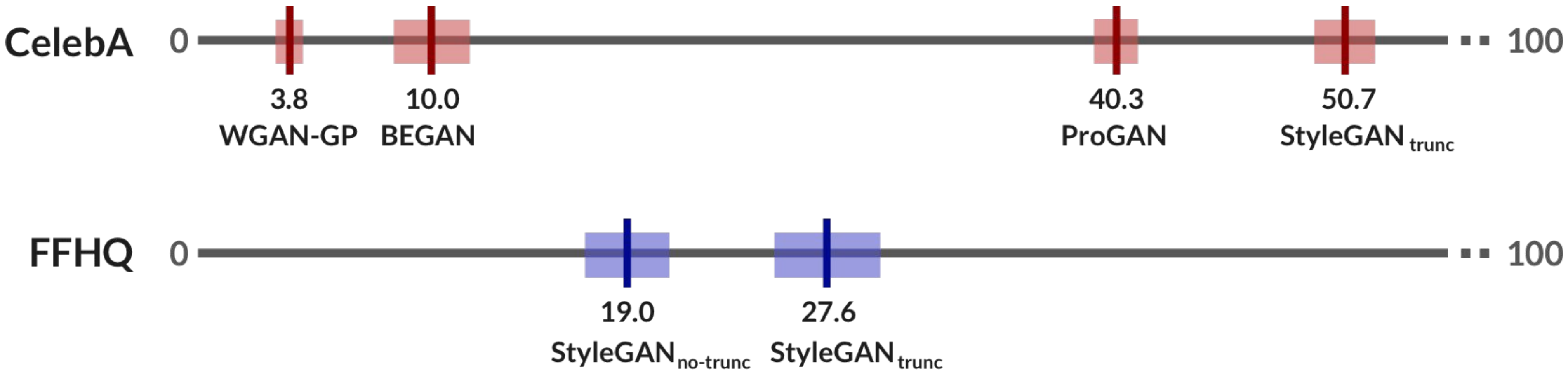
- FFHQ

Results

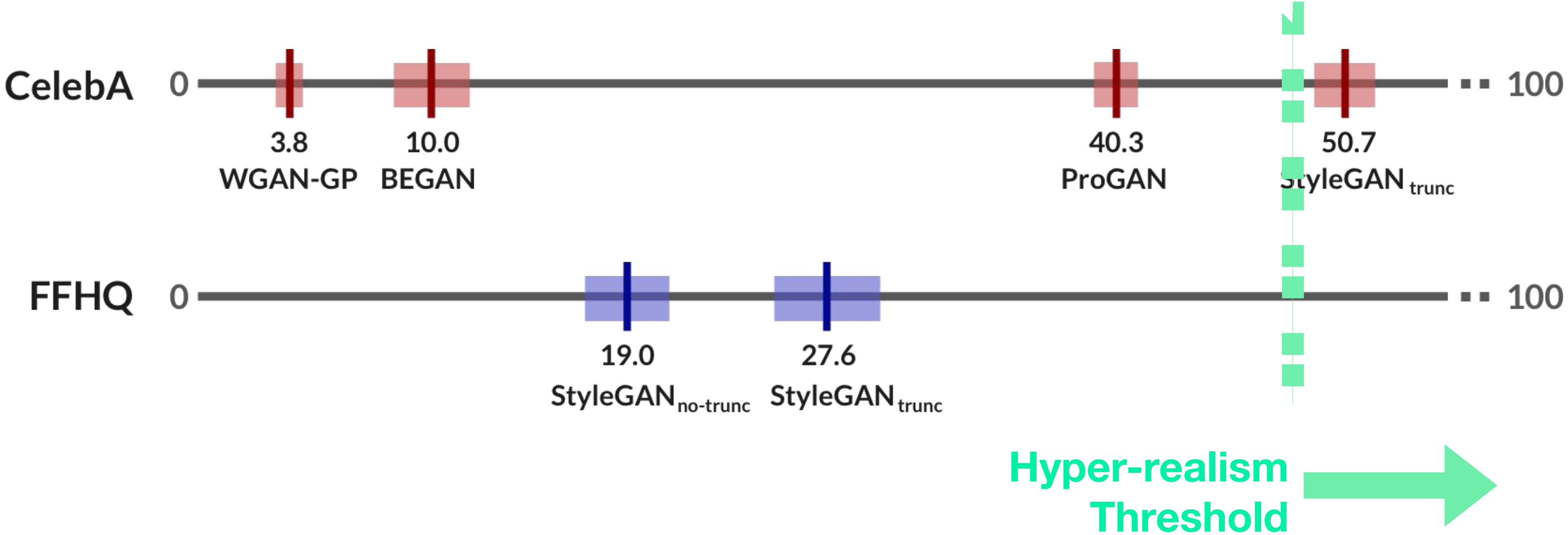
Are HYPE's results statistically separable?



Are HYPE's results statistically separable?



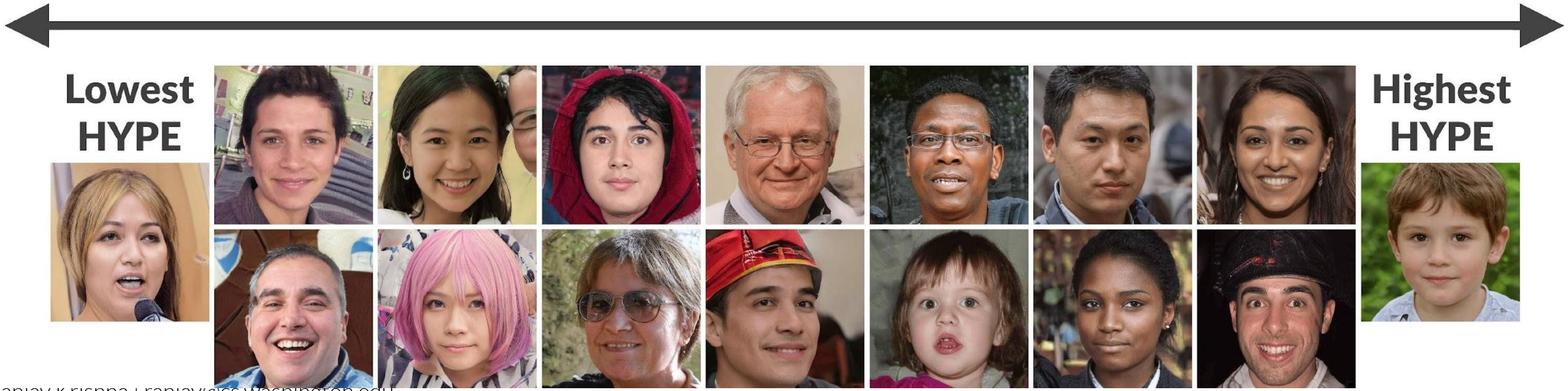
Are HYPE's results statistically separable?



Ranjay Krishna | ranjay@cs.washington.edu

HYPE achieves:

1. **Grounded** method inspired by psychophysics methods in perceptual psychology.
2. **Reliable** and consistent estimator.
3. Statistically **separable** to enable a comparative ranking.
4. Cost and time **efficient**.



Ranjay Krishna | ranjay@cs.washington.edu

Next time:
evaluations with real users from an AI+HCI perspective