Unreasonably shallow deep learning

Kyunghyun Cho. NYU & Genentech

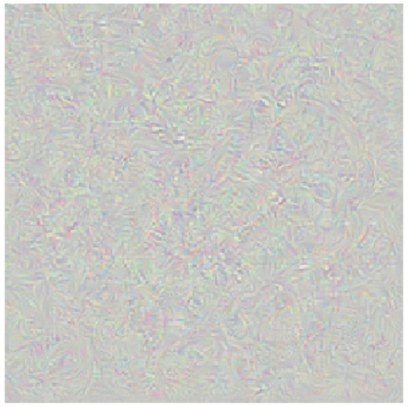
- Non-identifiability in deep learning
  - Taxonomy of solutions in statistical learning
  - Bayesian machine learning
- Spurious correlations arising from confounders
  - De-confounding by counterfactually-augmented data
  - Invariance risk minimization: beyond empirical risk minimization
- Feedback loops in machine learning
  - Data contamination and statistical watermarking
  - Exploration-exploitation dilemma
- What can we do?

(1) Non-identifiability in deep learning

Ambiguity

- Many *seemingly perfect* solutions exist with finite data & feature dependencies

Geirhos et al. (2020)



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 a 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

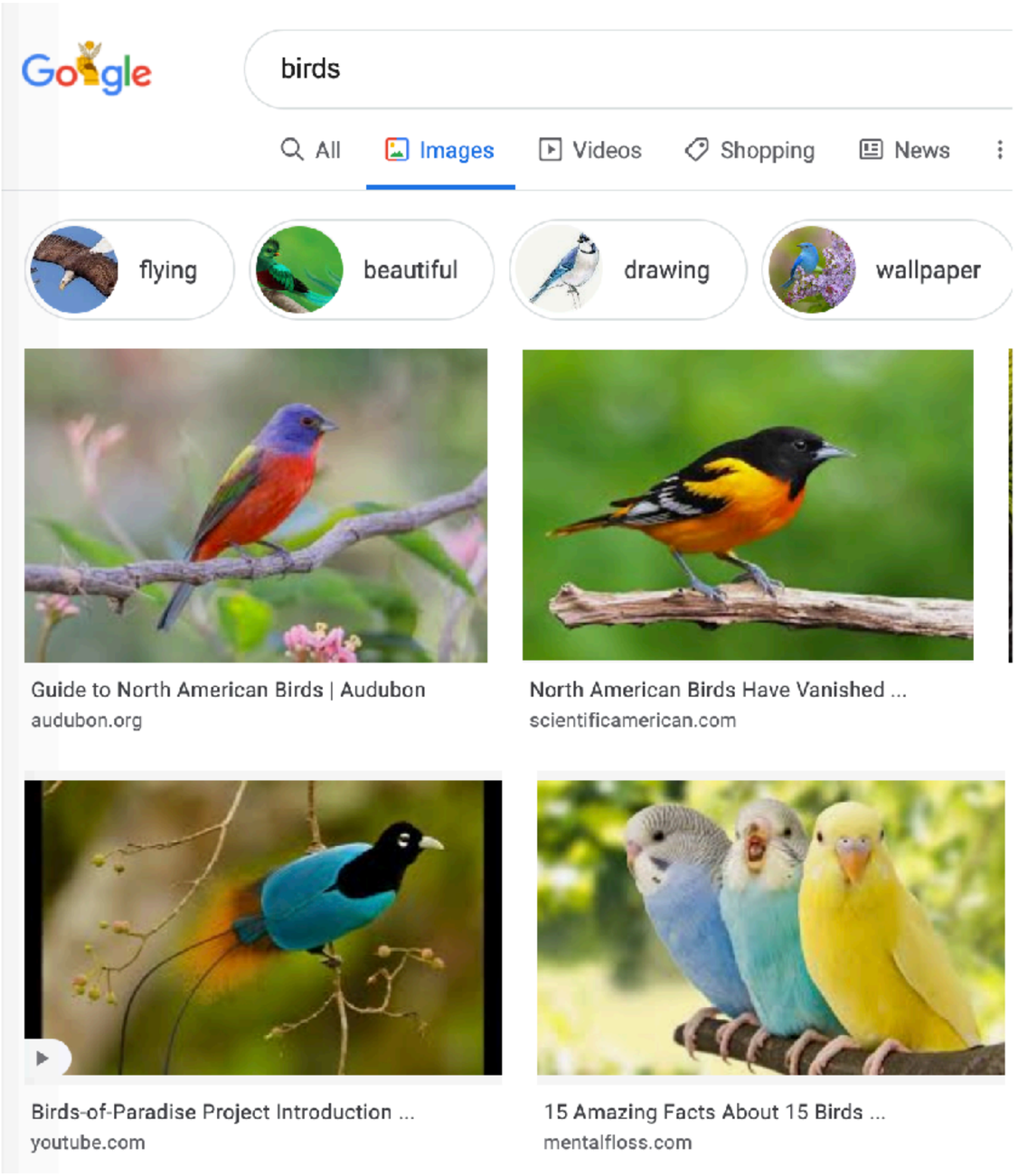
Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Ambiguity

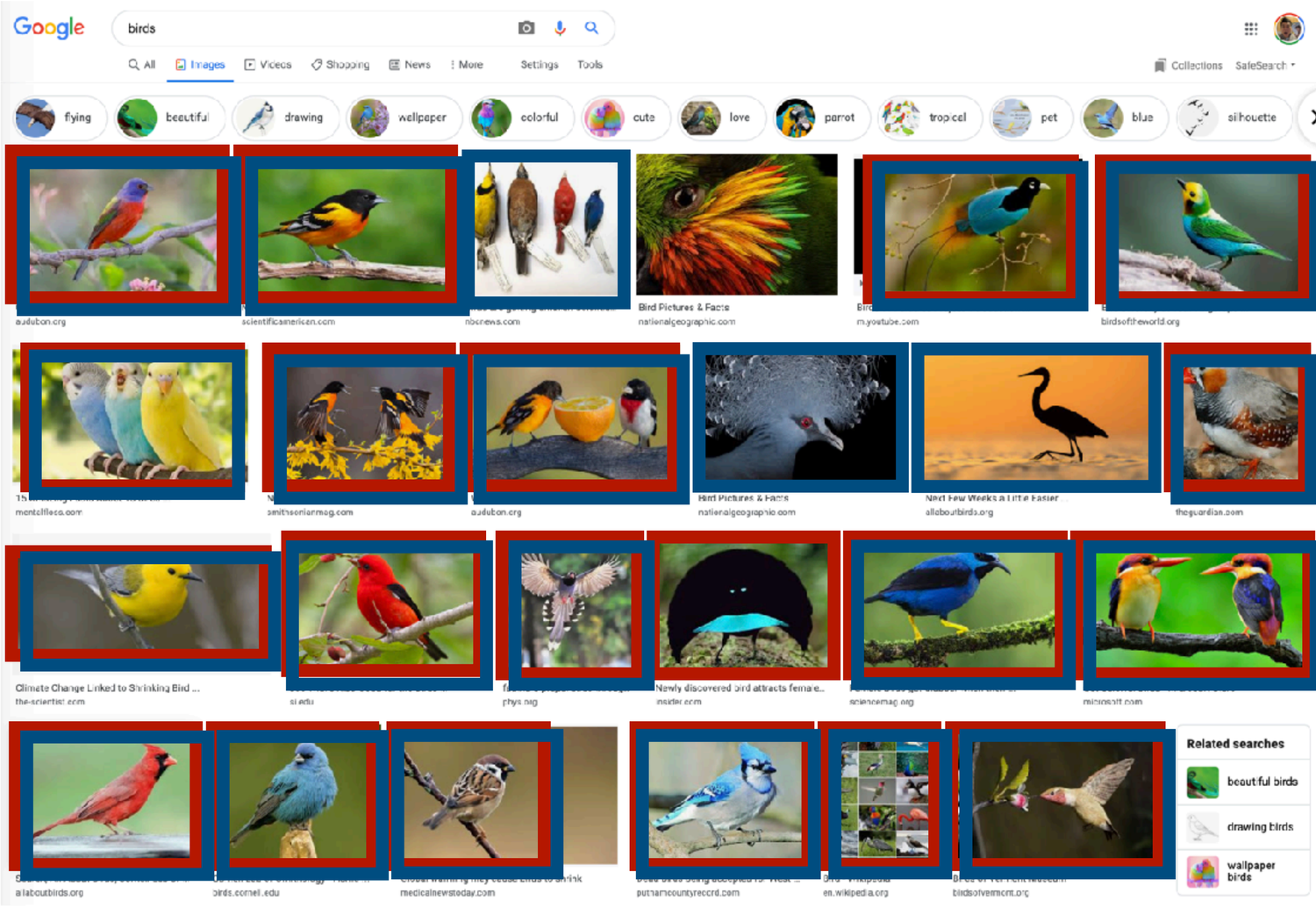
Finite data and feature dependencies

- Given the *four* training *examples*, build a *bird detector*
  - If a bird is in a picture, positive.
  - Otherwise, negative.
- Many *perfect* solutions
  - If there are branches, there is a bird.
  - If there is a beak, there is a bird.
  - …



Not necessarily overfitting either…

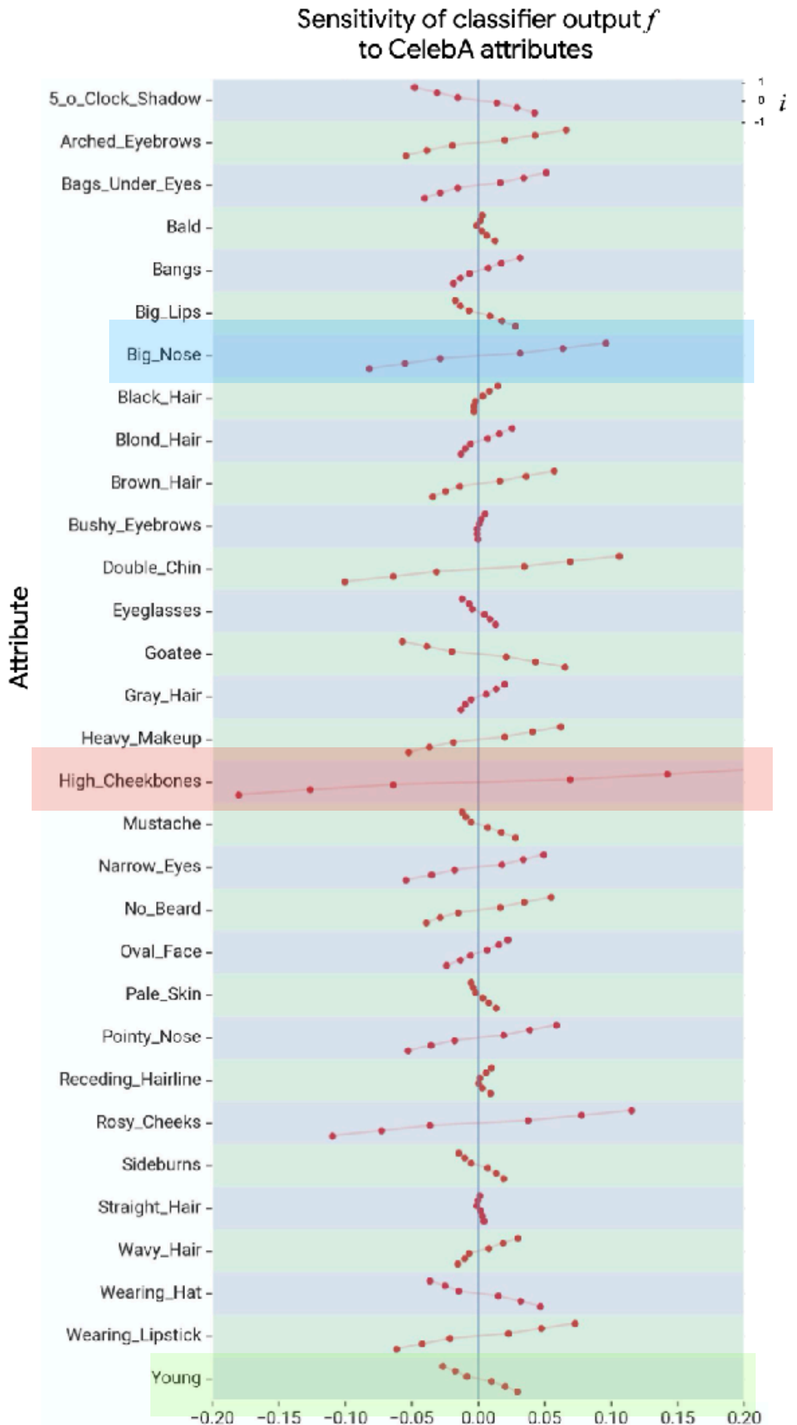
- Both solutions work well
  - If there are branches, there is a bird.
  - If there is a beak, there is a bird.



A few examples in practice

Smile detection

- Denton et al. [2019]
- The impact of various attributes on the prediction by a smile detector
- The smile detector captures various spurious correlations
  - Higher cheekbones make the probability of smiling lower.
  - Bigger noses make the probability of smiling lower.
  - Younger looking people are deemed more likely to smile.



A few examples in practice

Vision API

- DeVries et al. 2019

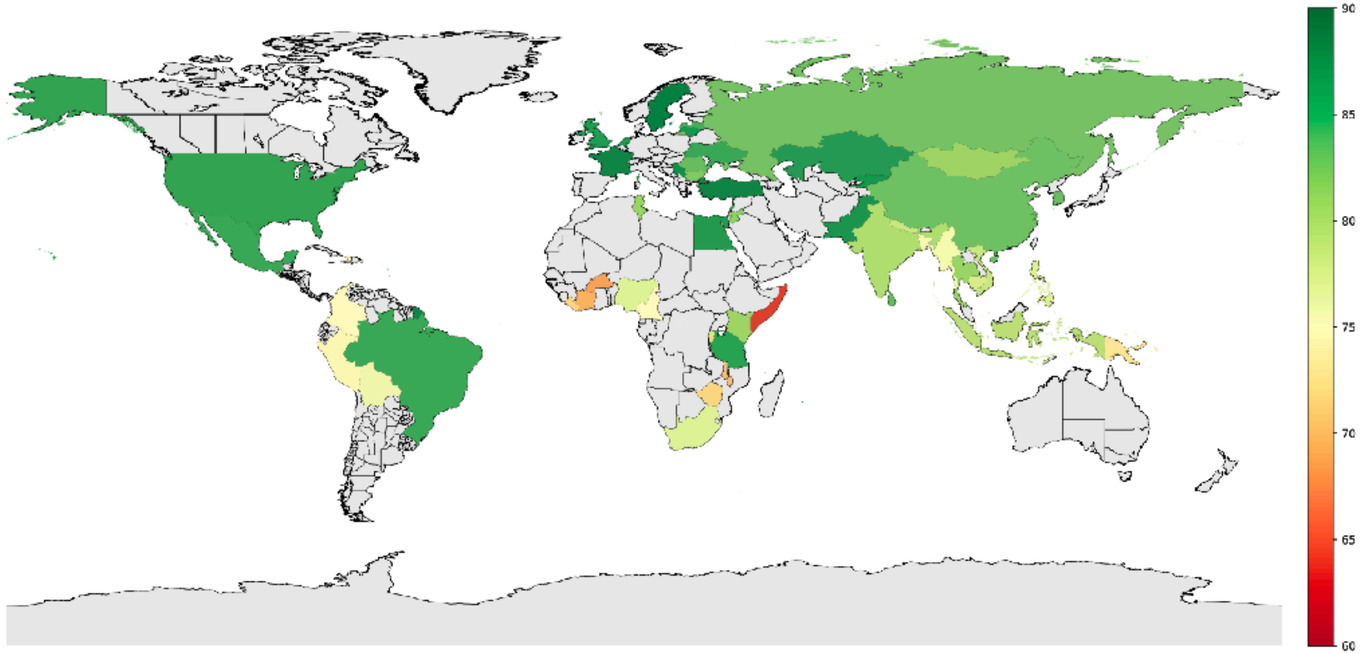


Figure 4: Choropleth map displaying the average accuracy of six object-classification systems per country. The color red indicates an accuracy of ∼60%, yellow an accuracy of ∼75%, and green an accuracy of ∼90%.

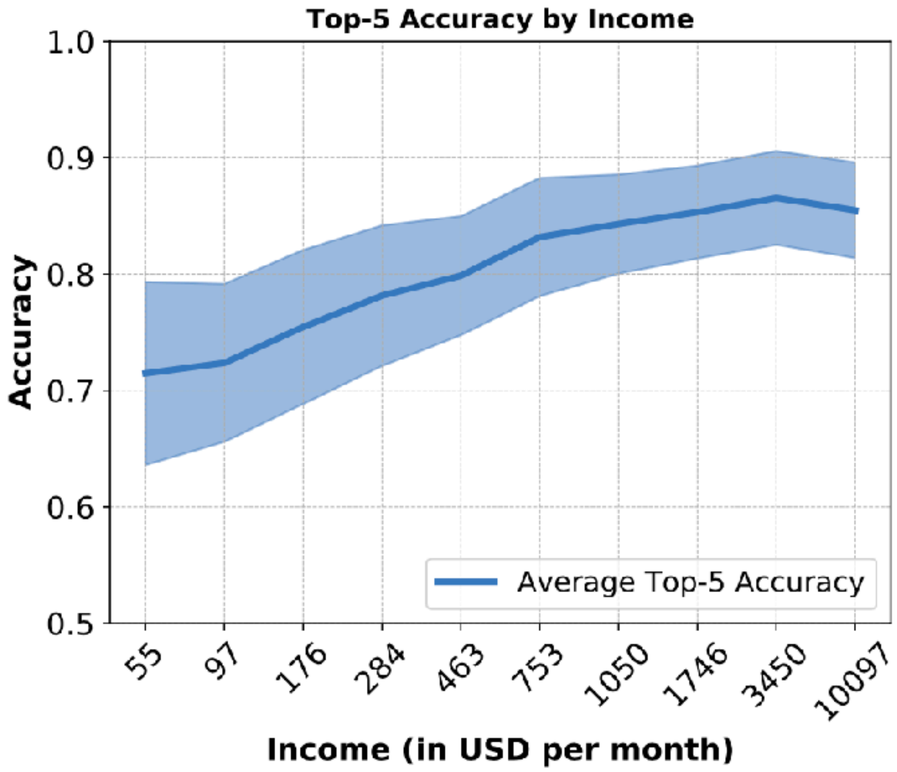


Figure 3: Average accuracy (and standard deviation) of six object-recognition systems as a function of the normalized consumption income of the household in which the image was collected (in US$ per month).

A few examples in practice

Vision API

- DeVries et al. 2019.

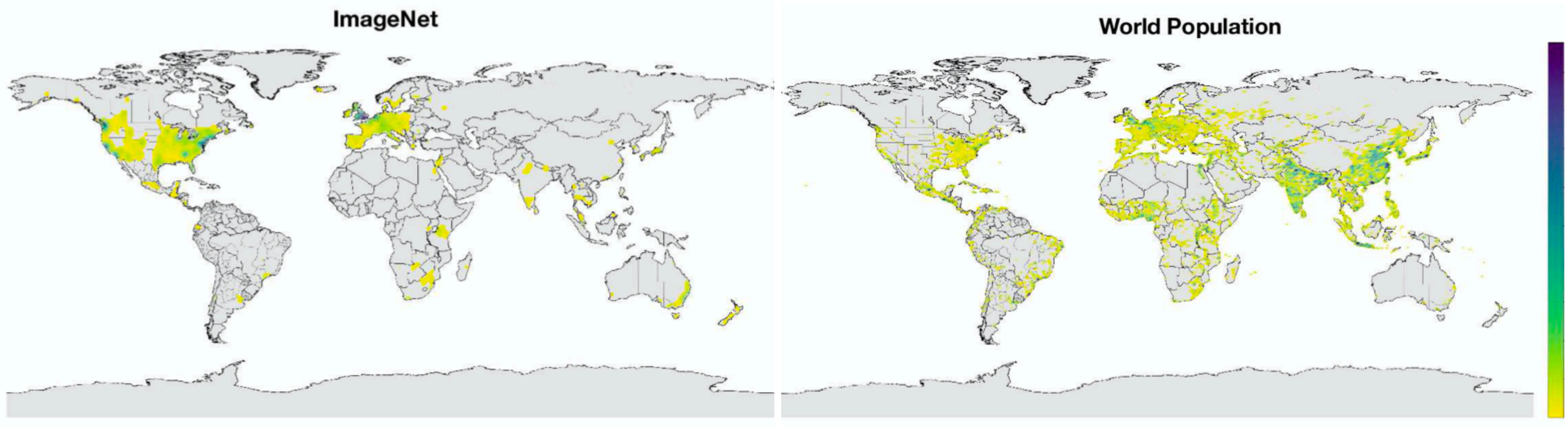
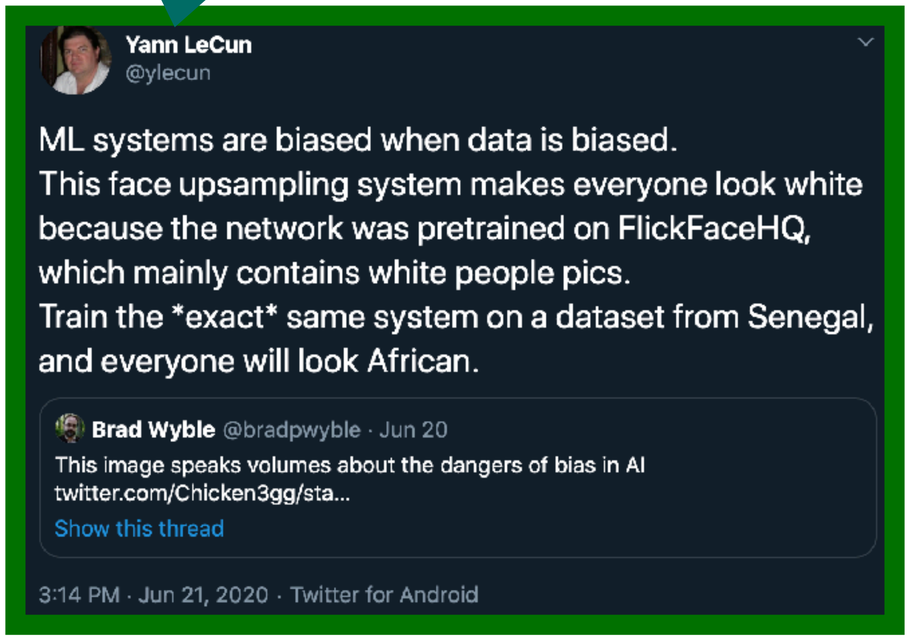


Figure 6: Density maps showing the geographical distribution of images in the ImageNet (top-left), COCO (top-right), and OpenImages (bottom-left) datasets. A world population density map is shown for reference (bottom-right).

It’s not just data but algorithms

- The apparent success of all these solutions create false illusion
  - Too much blame on data curation, too little blame on algorithms
  - Unreasonable trust in algorithms and uneducated claims



또 다른 축은 편향된 데이터 주입에 따른 결과 도출이다. 범죄율 예측이나 대출 한도 문제에 있어, 흑인에게 불리한 결과를 도출해 낸 사례가 있다. 여기서 반드시 짚고 넘어가야 할 점은 기계는 ‘주입된 정보’에 따라 결과를 출력한다는 것이다.

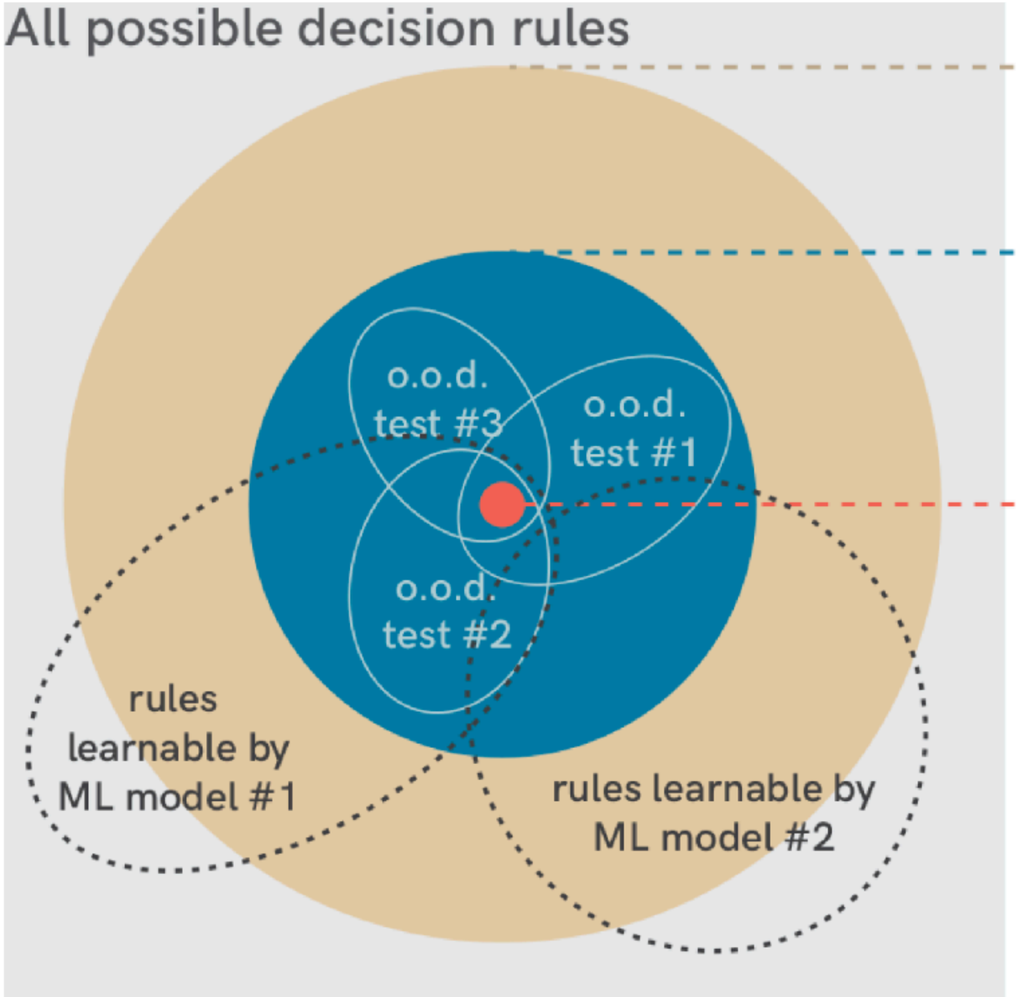
어떤 배경지식이나 선입견도 포함되지 않고, 학습한 대로 결과를 출력할 뿐이다. 따라서 기술을 두고 편견이나 공정성에 대해 판단하는 것은 앞뒤가 맞지 않다. 오히려 주입된 정보의 공정성이나 학습 과정의 문제를 지적해야 한다.

Yann, sorry! this tweet of yours was just a perfect example I was looking for!

Solution space in machine learning

Taxonomy of decision rules

- Training solutions
- Shortcut solutions
- Intended solutions
- Learnable solutions



training solution
performs well on training set

shortcut solution
performs well on training set and i.i.d. test set

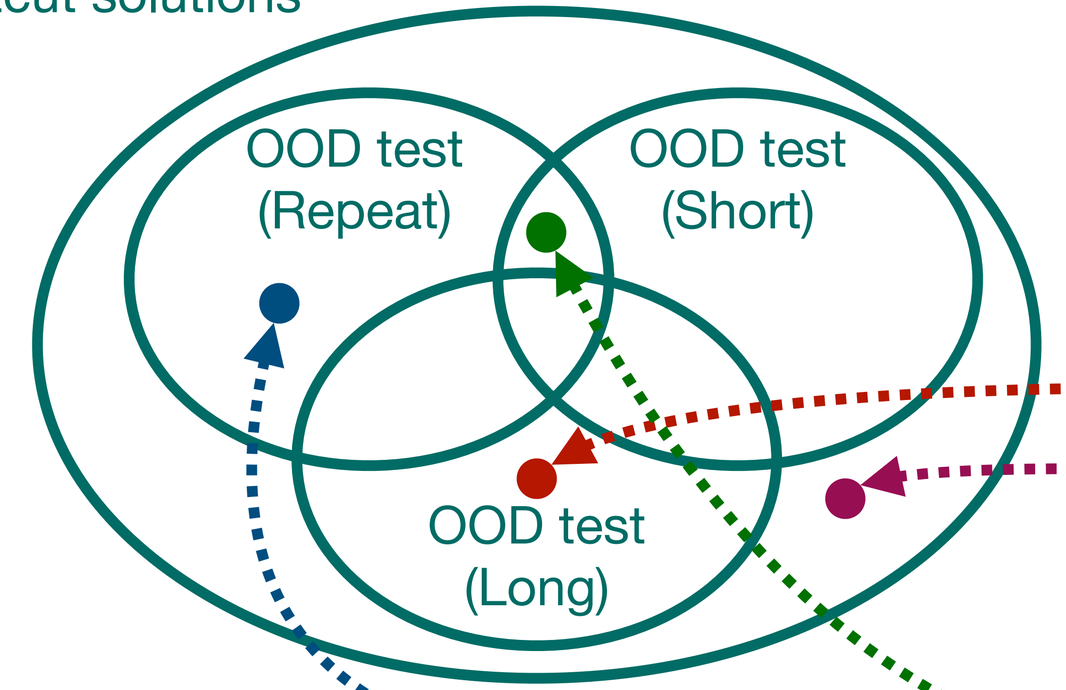
intended solution
performs well on training set, i.i.d. and all relevant o.o.d. test sets

Geirhos et al. (2020)

Decision rules in practice

It actually happens

Shortcut solutions



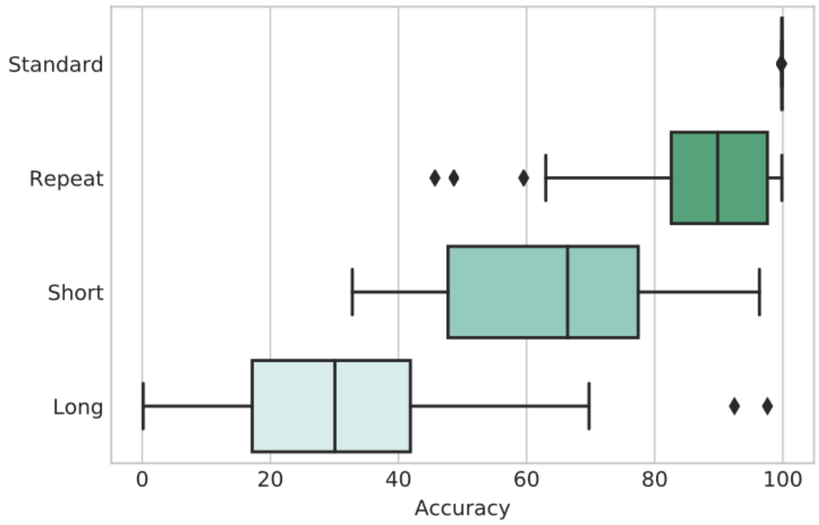


Figure 1: Accuracy % distribution across 50 runs with different random seeds on the four test sets.

Seed	Standard	Repeat	Short	Long
2787	**99.88**	94.65	42.05	23.05
5740	99.86	**45.70**	56.55	**97.60**
10000	99.86	98.55	**32.80**	**0.15**
14932	**99.73**	87.05	42.20	29.75
28897	99.87	**99.85**	47.40	1.40
30468	99.87	86.35	**96.35**	12.90

Table 3: Accuracy % on the test sets for selected runs out of 50 with different random seeds.

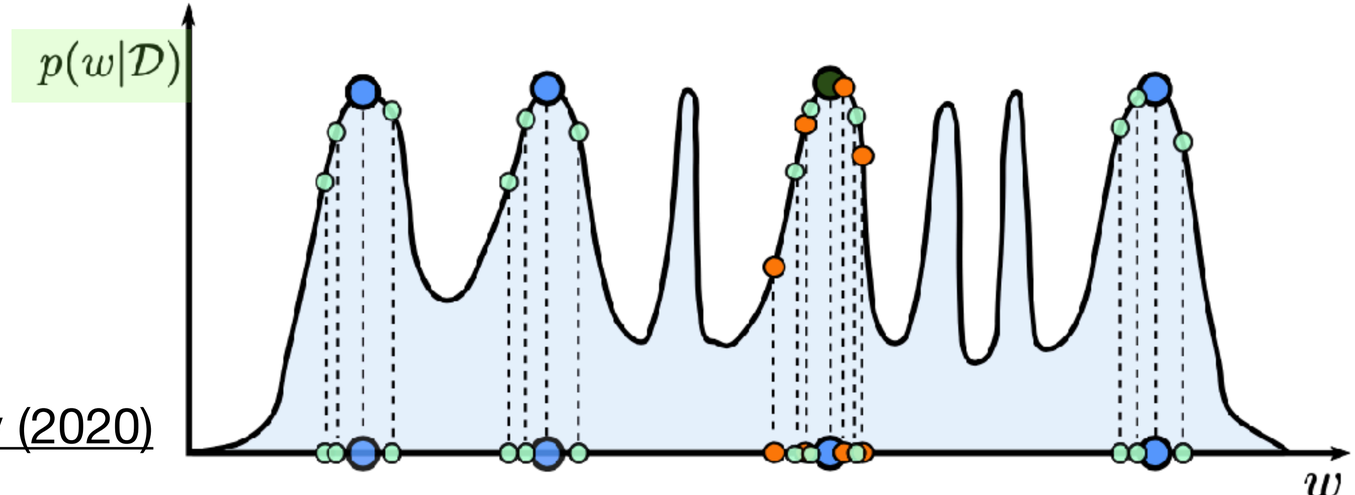
Weber, Shekhar & Balasubramanian (2018)

Bayesian machine learning

- Non-convex optimization as Bayesian posterior inference
  - Likelihood $p(D \mid \theta) \propto \exp\left\{-\sum_{(x,y)\in D} l(x, y; \theta)\right\}$ & prior $p(\theta) \propto \exp\left\{-\eta \|\theta\|^2\right\}$
  - Posterior $p(\theta \mid D) \propto p(D \mid \theta) p(\theta)$
  - Predictive distribution $p(y \mid x, D) = \dfrac{1}{p(x \mid D)} \int p(x, y \mid \theta) p(\theta \mid D) \mathrm{d}\theta$

Bayesian machine learning

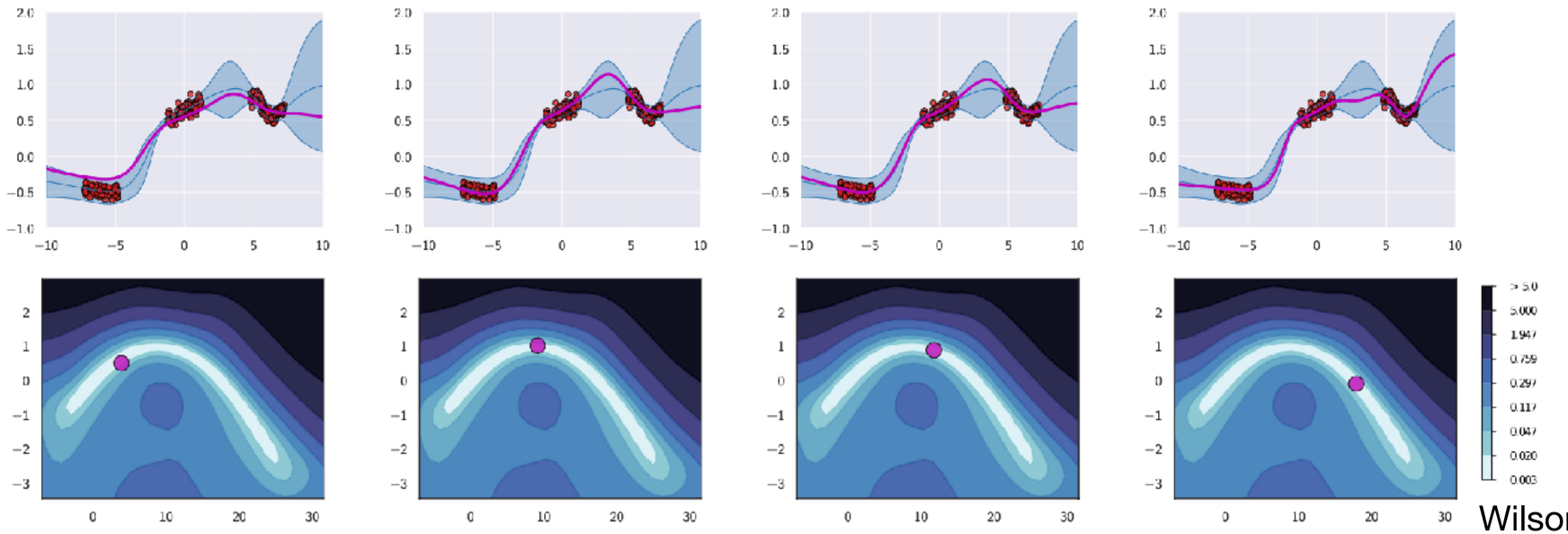
- Non-convex optimization as Bayesian posterior inference
  - Predictive distribution $p(y \mid x, D) = \mathbb{E}_{\theta|D}\left[p(y \mid x, \theta)\right] \approx \frac{1}{M}\sum_{m=1}^{M} p(y \mid x, \theta^m)$
  - The prediction is the average of the predictions from different solutions



Wilson & Izmailov (2020)

Bayesian learning is only a partial remedy

- Each solution corresponds to a different way of solving the problem.
- Averaging them avoids an individual solution's idiosyncrasy to a certain degree.



Wilson & Izmailov (2020)

- But, it's not perfect: observations alone cannot remove spurious correlations.

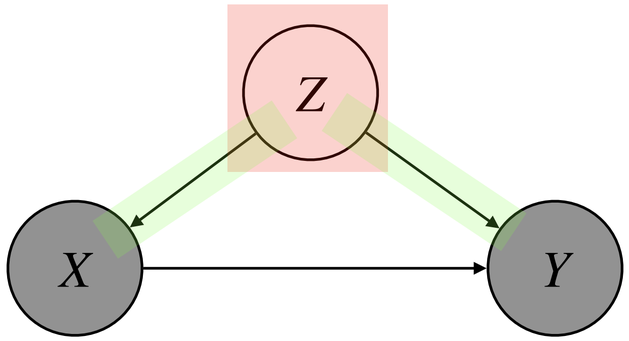
Discussion

Does my AI system work well?

- A **very difficult question** to answer correctly
  - Does my AI system work well on training instances?
  - Does my AI system work well on iid test instances?
    - What does my AI system exploit to work well?
  - Does my AI system work well on out-of-distribution test instances?
    - What does my AI system exploit so that it doesn’t work well?
- Is observation data alone enough?

(2) Causality vs. Correlation

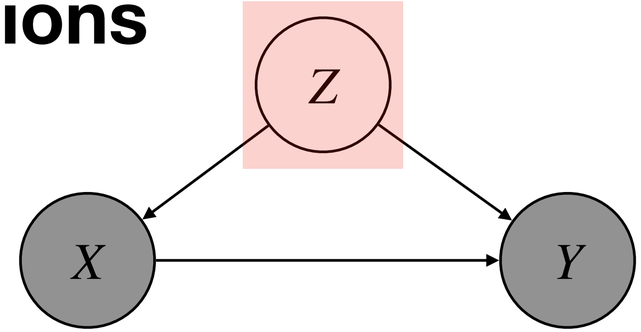
One aspect of undesirable solutions



- Some of the shortcut solutions exploit spurious correlations arising from unobserved confounding factors, creating a hidden path

ML meets causality

Hidden backdoors give rise to spurious correlations

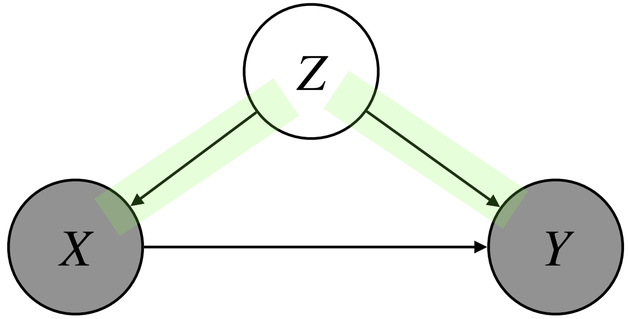


- "Correlation does not imply causality"
- Two possible perspectives
  - The category "Birds" *causes* the display of an actual realization of a "bird"
  - The visual perception of a "bird" *causes* our brain to recognize it as "Birds"
- In either of the cases, there exists a hidden confounder
  - **Forest**: birds are often spotted sitting on tree branches in a forest.
- The hidden confounder creates a hidden path
  - $X \leftarrow Z \rightarrow Y$ in *addition* to the true causal path $X \rightarrow Y$

ML meets causality

Hidden backdoors give rise to spurious correlations

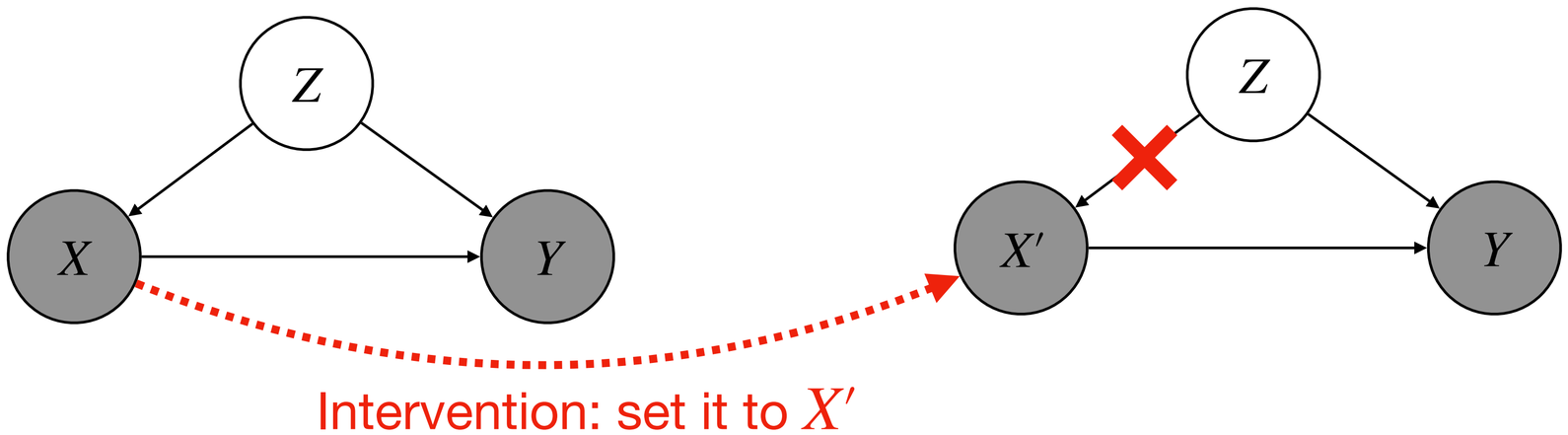
- $X \leftarrow Z \rightarrow Y$ and $X \rightarrow Y$ cannot be distinguished from observing (X, Y) alone
  - Correlation between X and Y arises from either or both of these paths
  - That is, “correlation does not imply causality”
- Back to our bird detector:
  - $X \rightarrow Y$: A bird has a beak and a pair of wings and is covered with feathers.
  - $X \leftarrow Z \rightarrow Y$: A bird often sits on a tree branch in a forest.
- Largely impossible to estimate based on observations alone because
  - $p(Y \mid \text{do}(X)) \neq p(Y \mid X) = \int p(Y \mid X, Z) p(X \mid Z) p(Z) \text{d}Z$



ML meets causality

Human and model in the loop

- But, what are those unobserved confounders?
  - Even if we knew those confounders, how do we observe all of them?
  - Even if we observed them, how do we model $X \mid Z$?
- If it cannot be done, can we **intervene** on X to break $Z \rightarrow X$ in $X \leftarrow Z \rightarrow Y$?



ML meets humans

Human and model in the loop

- Somehow *we* know those backdoor paths
  - *We figured out that our bird detector was in fact a branch detector!*
- Can we (humans) help machines ignore these backdoor paths?

Human and model in the loop

Breaking $X \leftarrow Z$

- *"Given a context (also often called a "premise" in NLI), and a desired target label, we ask the human writer to provide a hypothesis that fools the model into misclassifying the label."* - Nie et al. [2019]

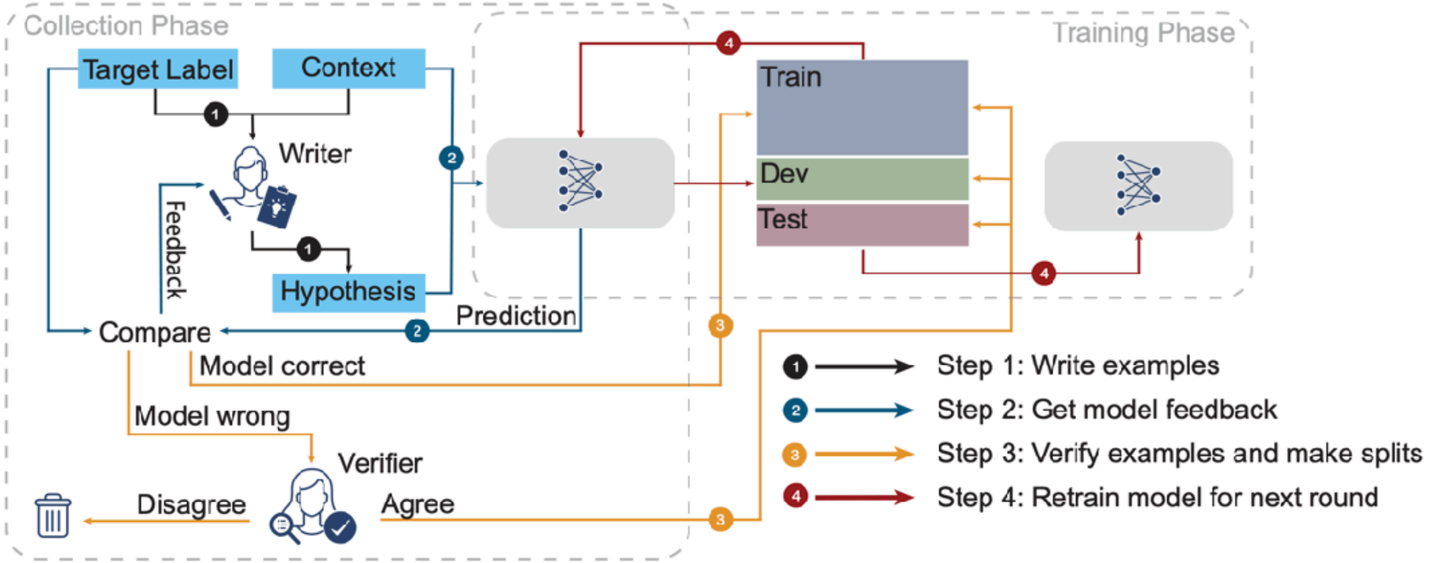


Figure 1: Adversarial NLI data collection via human-and-model-in-the-loop enabled training (HAMLET). The four steps make up one round of data collection. In step 3, model-correct examples are included in the training set; development and test sets are constructed solely from model-wrong verified-correct examples.

Human and model in the loop

Breaking $X \leftarrow Z$

- *"By leveraging humans not only to provide labels but also to intervene upon the data, revising documents to accord with various labels, we can elucidate the difference that makes a difference."* Kaushik et al. [2019]

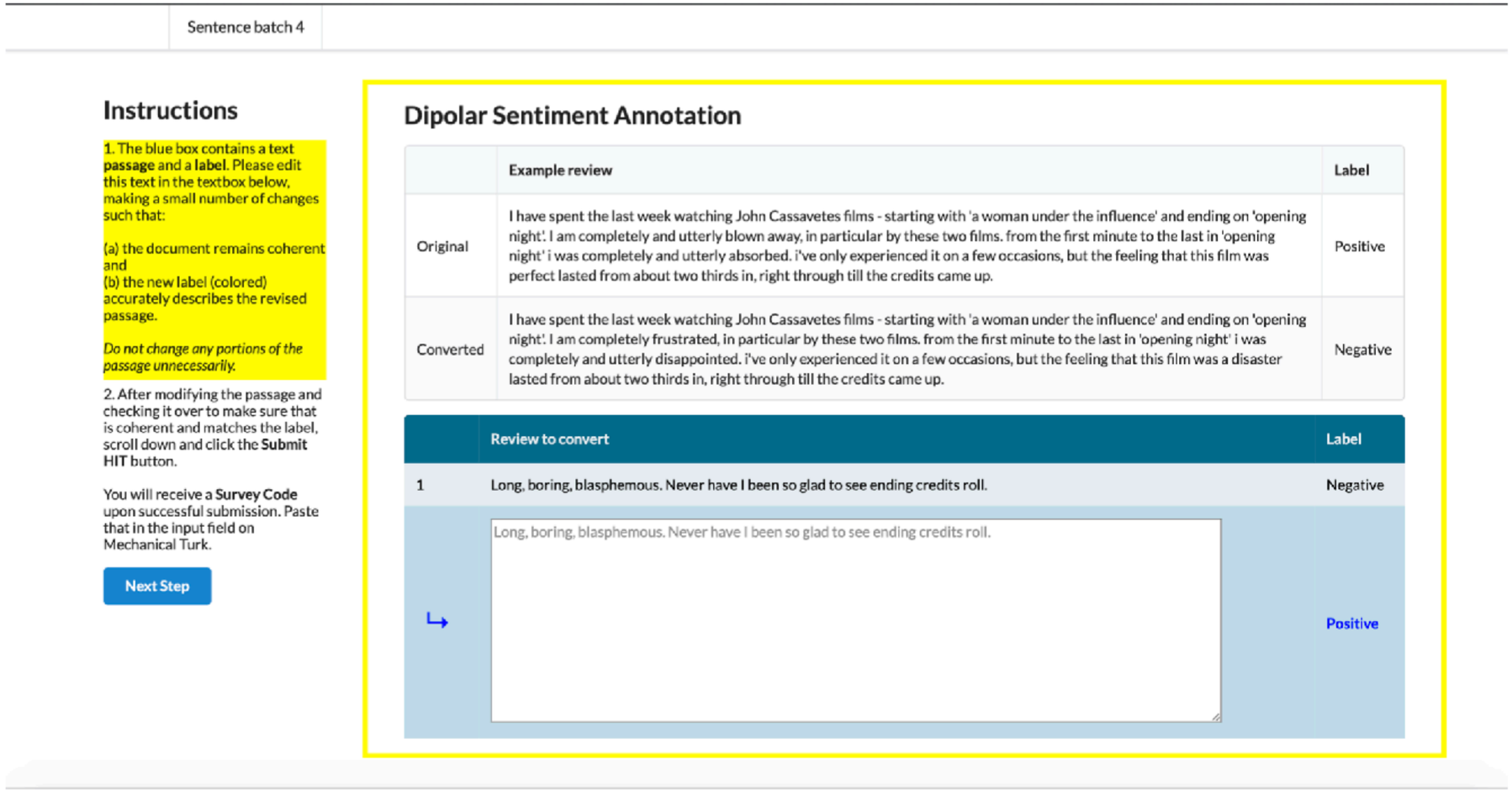


Figure 2: Annotation platform for collecting counterfactually annotated data for sentiment analysis

Human and model in the loop

It helps, but is increasingly more difficult as …

1. The quality of a target model improves [Nie et al., 2019]

Dataset	Genre	Context	Train / Dev / Test	Model error rate		Tries	Time (sec.)
				Unverified	Verified	mean/median per verified ex.	
A1	Wiki	2,080	16,946 / 1,000 / 1,000	29.68%	18.33%	3.4 / 2.0	199.2 / 125.2
A2	Wiki	2,694	45,460 / 1,000 / 1,000	16.59%	8.07%	6.4 / 4.0	355.3 / 189.1
A3	Various	6,002	100,459 / 1,200 / 1,200	17.47%	8.60%	6.4 / 4.0	284.0 / 157.0
	(Wiki subset)	1,000	19,920 / 200 / 200	14.79%	6.92%	7.4 / 5.0	337.3 / 189.6

2. The problem becomes more sophisticated.
   - e.g., Write a news article for which an automatic summarizer would fail (!?)
3. The assumption on underlying causal graphs leads to different experiment designs
   - see, e.g., [Kaushik et al., 2020] for the impact of different causal assumptions
   - Underspecified instruction to humans may result in sub-optimal outcomes [Huang et al., 2020]

Human and model in the loop

“Human” may suffer from the process

THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST

Illustrations by Corey Brickley | Photography by Jessica Chou

- “***The video depicts a man being murdered***. *Someone is stabbing him, dozens of times, while he screams and begs for his life.* ***Chloe’s job is to tell*** *the room* ***whether this post should be removed***.”

- “***“Autistic people should be sterilized”*** *seems offensive to him, but it* ***stays up*** *as well.* ***Autism is not a “protected characteristic”*** *the way race and gender are, and so it doesn’t violate the policy. (“Men should be sterilized” would be taken down.)*”

- “***The job also changed the way he saw the world***. *After he saw so many videos saying that 9/11 was not a terrorist attack, he came to believe them. Conspiracy videos about the Las Vegas massacre were also very persuasive, he says, and he now believes that multiple shooters were responsible for the attack.*”

Human and model in the loop

Disproportionately problematic to disadvantaged populations

- https://twitter.com/colinmadland/status/1307111816250748933

Human and model in the loop

Disproportionately problematic to disadvantaged populations

- https://twitter.com/AnimaAnandkumar/status/1307470107401170947?s=20

Human and model in the loop

Early deployment and iterative refinement is often not an option

- We often **cannot deploy an imperfect system**, let humans find loopholes and patch them iteratively, because doing so often **negatively affects disadvantaged populations** in the society disproportionately.
- This often leaves (potentially) **irreversible damages**.

Human and model in the loop

What I least suspected…

- "***Israeli police arrested a Palestinian man*** *last week* ***after a Facebook post*** *he made saying "good morning" in Arabic* ***was mistranslated to read "attack them"*** *in Hebrew, local media have reported."*

<Israel Facebook: Mistranslated post creates security alert> BBC 2017.

Human and model in the loop

Early deployment and iterative refinement is often not an option

- <Wrongfully Accused by an Algorithm> by Kashmir Hill on NYTimes (2020)
  - "***Two officers got out and handcuffed Mr. Williams*** *on his front lawn,* ***in front of his wife and two young daughters*** *… The police drove Mr. Williams to a detention center. He had his mug shot, fingerprints and DNA taken, and was held overnight.*"
  - "*The detective turned over … a still image from a surveillance video … his case may be the first known account of* ***an American being wrongfully arrested based on a flawed match from a facial recognition algorithm****.*"
  - "*… Mr. Williams asked if he was free to go. "Unfortunately not," one detective said. …* ***Mr. Williams was kept in custody until that evening****, 30 hours after being arrested,* ***and released on a $1,000 personal bond****.*"

Human and model in the loop

Early deployment and iterative refinement is often not an option

- §2 Automating Eligibility in the Heartland from <Automating Inequality>
  - by Virginia Eubanks
  - "***The Indiana automated eligibility system*** *… achieved stunning welfare roll reductions. …* ***As poverty in Indiana increased****, caseloads dropped. … in 2006, 38 percent of poor families with children were* ***receiving cash benefits from TANF****. By 2014, the number* ***had dropped to 8 percent****.*"
  - "*In 2000, African Americans made up 46.5 percent of the state's TANF rolls, and whites had a very slim majority in the program …* ***At the end of the automation experiment in 2010****, the gap … had widened precipitously. … the TANF rolls were now 54.2 percent white and only 32.1 percent African American. …* ***tested on primarily white communities, Black families still felt its worst effects.***"

Invariance to environments

Causal graphs are often difficult to work with

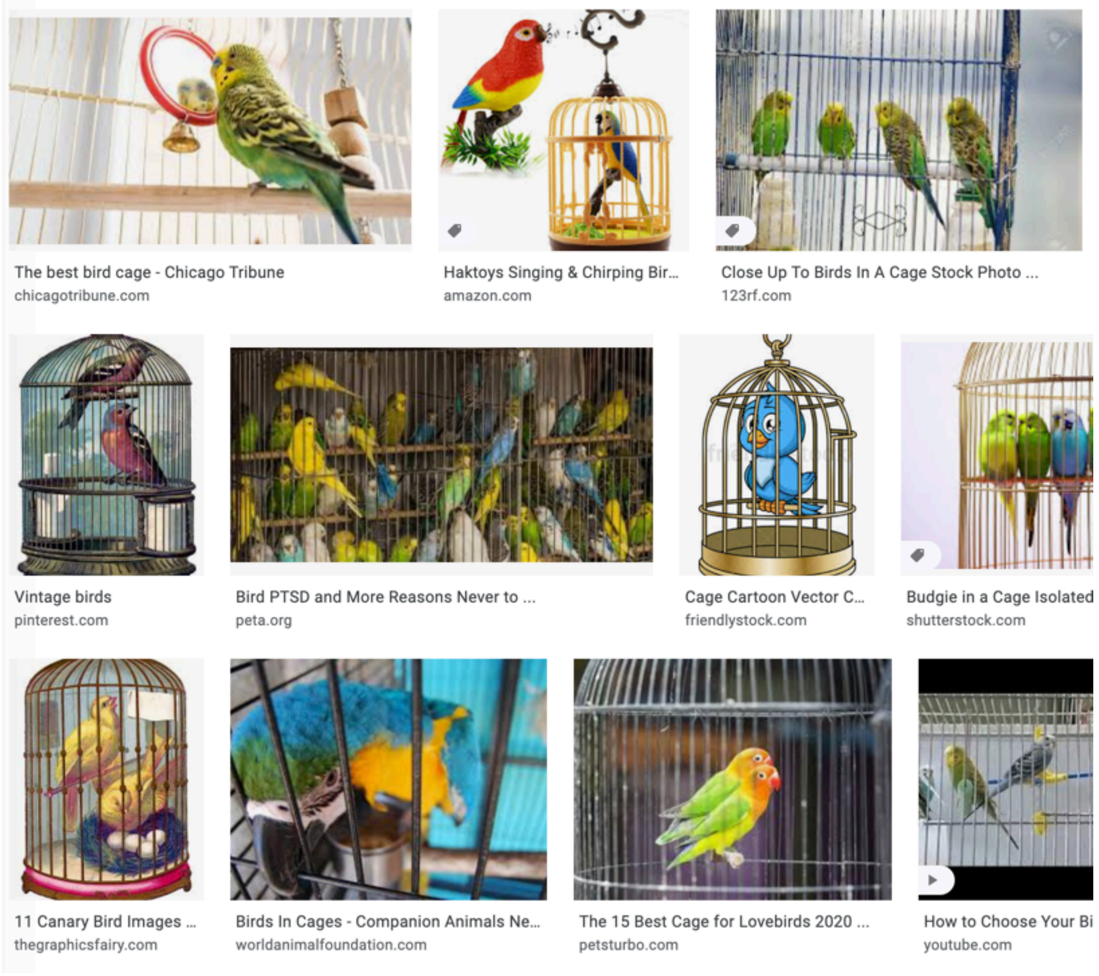
- We must do better before deploying our systems.
  - Can we draw a causal graph and perform causal analysis each time?
- Arjovsky et al. (2019) however ask:
  - "*For instance, consider the task of image classification. Here,* ***the observed variables are hundreds of thousands of correlated pixels****.* ***What is the causal graph governing them?*** *It is reasonable to assume that causation does not happen between pixels, but between the real-world concepts captured by the camera. In these cases,* ***invariant correlations in images are a proxy into the causation at play in the real world****. To find those invariant correlations,* ***we need*** *methods which can* ***disentangle the observed pixels into latent variables closer to the realm of causation***"

Invariance to environments

Intended solutions are often invariant to environmental changes

- Bird classifiers
  - ~~If there are branches, there is a bird.~~
  - **If there is a beak, there is a bird.**
  - …
- Intended solutions should not be sensitive to the perturbations in the environment
- More desirable than "*a meaningful causal graph relating the observed variables,* [which is] *an awkward assumption when dealing with perceptual inputs such as pixels.*"

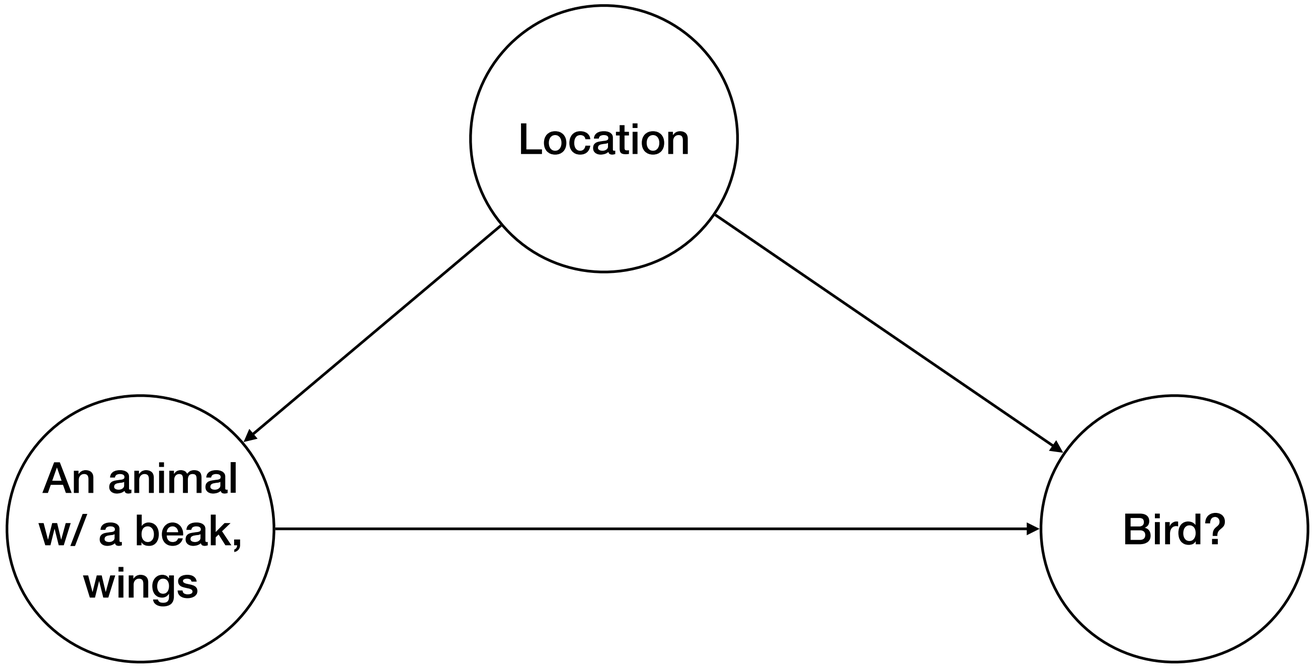
  Arjovsky et al. [2019]



Invariance to environments

Environments share a common causal structure

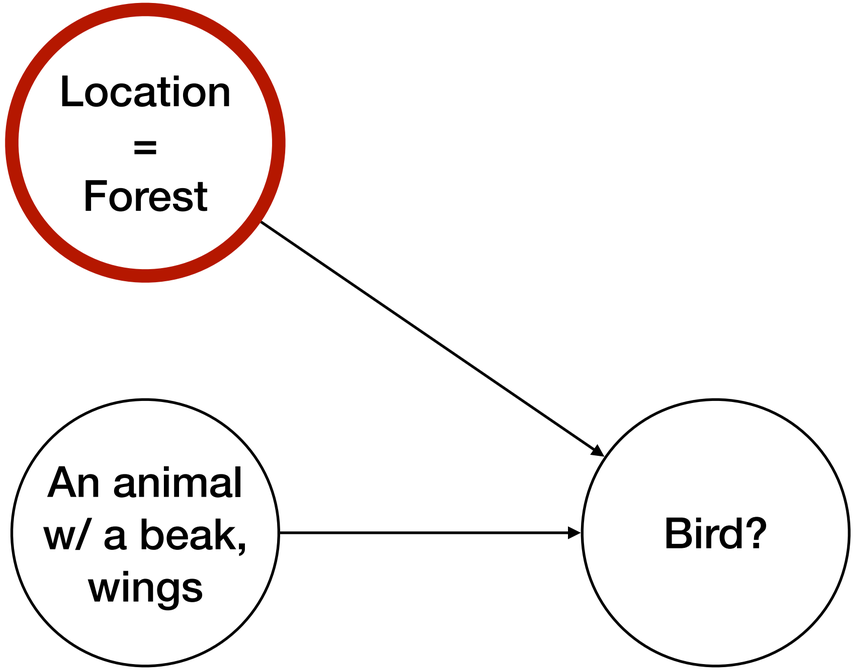
- Invariant prediction principle: "*if we consider all "direct causes" of a target variable of interest, then* ***the conditional distribution of the target given the the direct causes will not change when we interfere experimentally with all other variables*** *in the model except the target itself.*" [Peters et al., 2015]



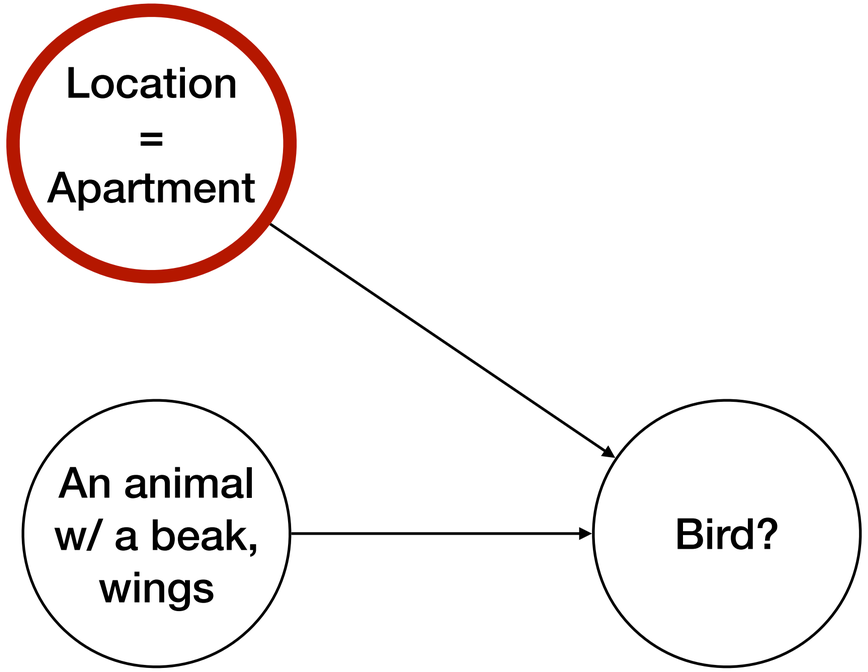
Invariance to environments

An environment = an intervention

- Environment 1: Location = Forest



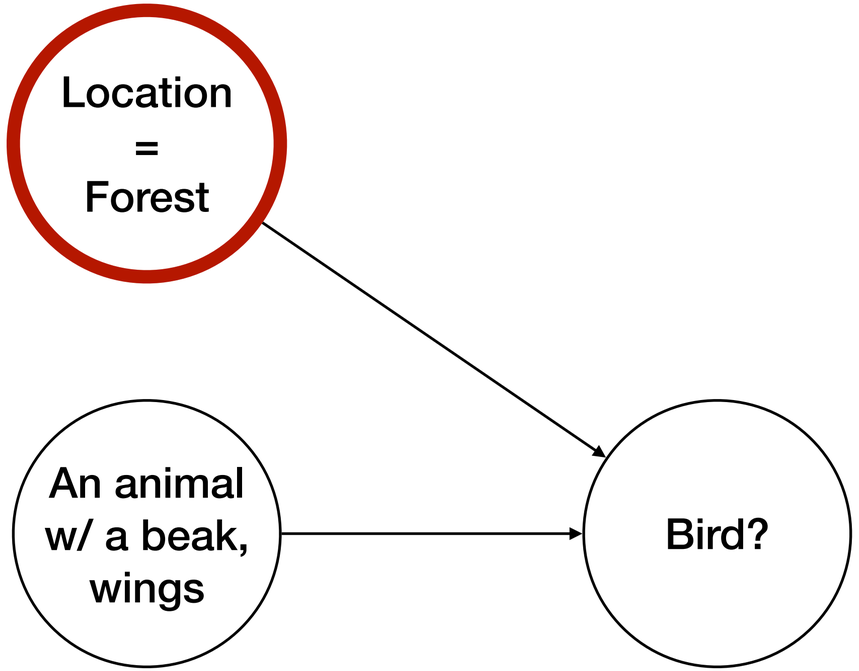
- Environment 2: Location = Apartment

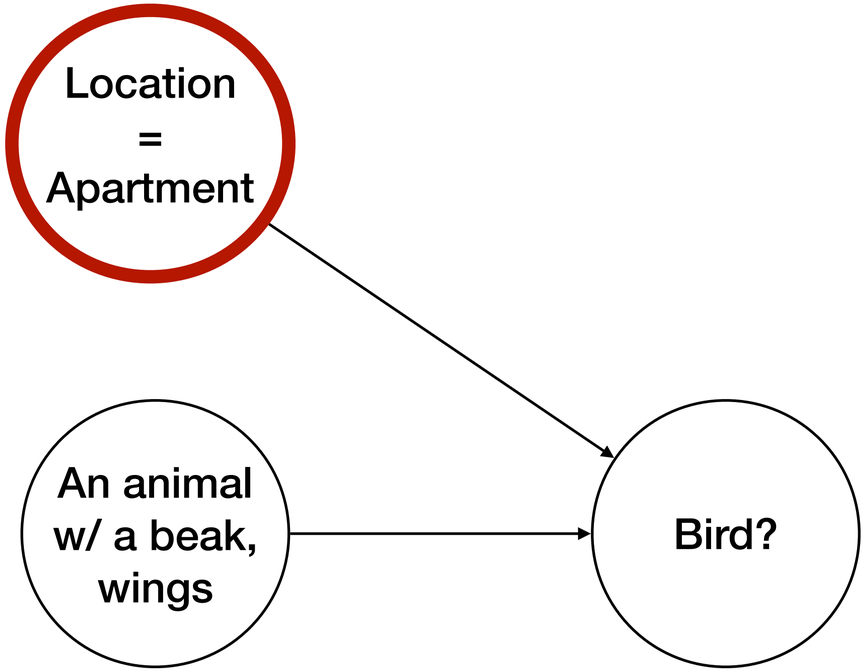


Invariance to environments

Invariance vs. Causal mechanism

- When would a classifier work well on both environments?





Invariant risk minimization

IRM directly maximizes the invariance

- IRM [Arjovsky et al., 2019]

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi)$$

$$\text{subject to} \quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e \in \mathcal{E}_{\mathrm{tr}}$$

  - Φ: a feature extractor & w: a classifier weight
  - $\mathscr{E}_{\mathrm{tr}}$: a set of training environments
  - R^e: a loss function for the environment e

Invariant risk minimization

IRM directly maximizes the invariance

- IRM [Arjovsky et al., 2019]

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi)$$

$$\text{subject to} \quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\text{tr}}$$

  - **Inner-loop**: for each $e \in \mathscr{E}_{\text{tr}}$, find w that works well with Φ.
  - **Outer-loop**: find Φ that works well for all training environments.

Invariant risk minimization

Connection to meta-learning

- MAML [Finn et al., 2018][Luketina et al., 2016]
  - **Inner-loop**: for each $\mathcal{T}_i \sim p(\mathcal{T})$, find θ_i' that is easily reachable from θ.
  - **Outer-loop**: find θ to be a good initial point for all tasks.

Algorithm 1 Model-Agnostic Meta-Learning

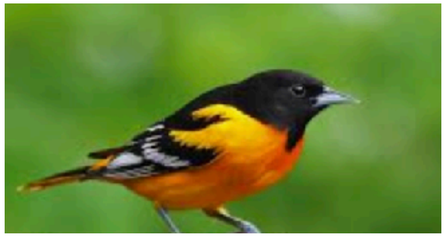
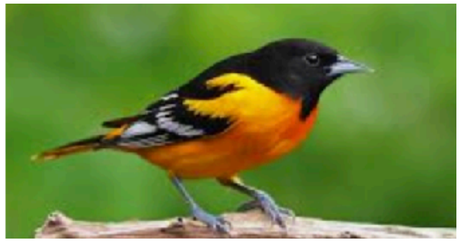
Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters

```
randomly initialize θ
while not done do
    Sample batch of tasks T_i ~ p(T)
    for all T_i do
        Evaluate ∇_θ L_{T_i}(f_θ) with respect to K examples
        Compute adapted parameters with gradient descent: θ'_i = θ − α∇_θ L_{T_i}(f_θ)
    end for
    Update θ ← θ − β∇_θ Σ_{T_i ~ p(T)} L_{T_i}(f_{θ'_i})
end while
```

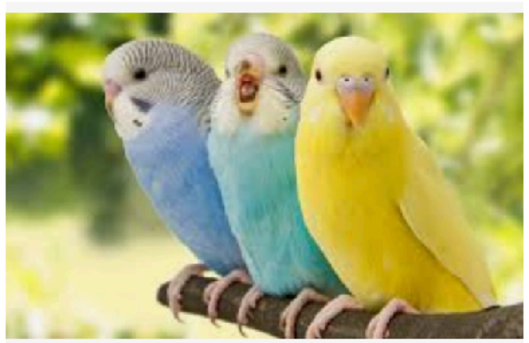
Data augmentation

Poor man's invariance risk minimization

- One of the key components from Krizhevsky et al. (2012)
- In most of birds' pictures, branches are placed at the bottom. Data augmentation by translation can destroy this spurious correlation between branches and birds:

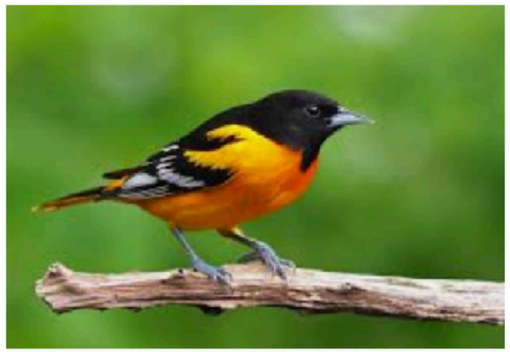



- Limited intervention: intervention on camera configuration



15 Amazing Facts About 15 Birds ...
mentalfloss.com

Guide to North American Birds | Audubon
audubon.org



North American Birds Have Vanished ...
scientificamerican.com

Invariant risk minimization

Still quite limited

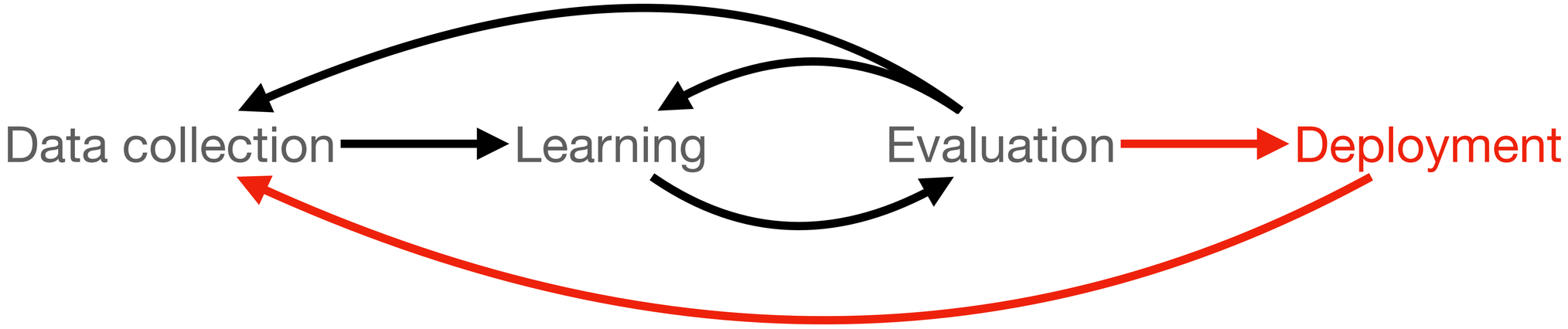
- It is not trivial to build a dataset to have multiple domains
  - How do we define a domain?
  - How do we collect data to cover as many domains as possible?
  - How do we detect source domains of existing instances?
- Generalization to unseen domains is not guaranteed
  - e.g. “The Risk of Invariant Risk Minimization” by Rosenfeld et al. (2021, ICLR)

Discussion

AI cannot be solved by existing algorithms and practices

- The world is complicated with many variables causally affecting each other
- We only observe a few variables and model unobserved variations as noise.
- We cannot measure the *true* relationship between two variables.
- Causal analysis seems to provide some insight
  - Intervention: Human-and-machine-in-the-loop systems
  - Invariant predictor: invariant risk minimization
- Still in its infancy…

(3) A feedback loop in machine learning



1. Data contamination
2. Exploration vs. Exploitation

Google Cloud Vision API

Works beautifully

- Try it yourself!

Faces | **Objects** | Labels | Text | Properties | Safe Search

Image from iOS (1).jpg

Person	92%
Person	88%
Person	88%
Backpack	87%
Person	87%
Person	87%
Person	85%
Person	83%

Landmarks | Objects | **Labels** | Text | Properties | Safe Search

IMG_0527.jpg

Building	99%
Metropolitan Area	99%
Skyscraper	98%
City	97%
Urban Area	97%
Metropolis	97%
Landmark	97%
Tower Block	97%
Architecture	96%
Daytime	96%
Tower	95%
Condominium	93%

Building ImageNet

The standard dataset for computer vision research/development

- *ImageNet was constructed in 2009 through Internet search and crowdsourcing. The research team obtained* ***a vocabulary of categories from WordNet****, an English database that represents each category as a synonym set (or "synset") consisting of synonymous words, for example, dogsled, dog sled, dog sleigh. WordNet was manually constructed by psychologists and linguists at Princeton University starting in 1985, and has been widely used in the research of Natural Language Processing.*

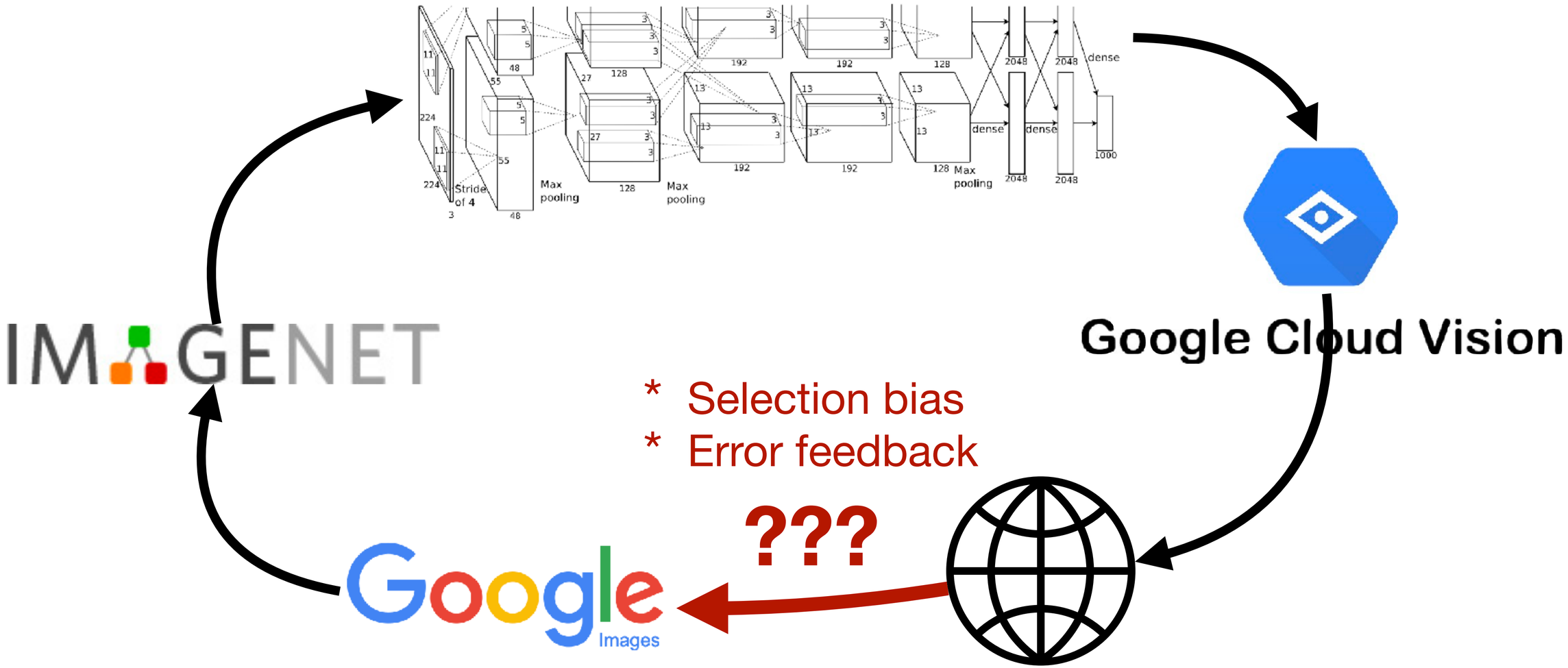
  For each synset, the ImageNet team in 2009 automatically downloaded images by querying search engines*. At that time, search engines did not have accurate algorithms to understand the content of images, so they retrieved images mostly based on their captions or tags. This resulted in many irrelevant images for each category.*

  To remove the irrelevant images, they hired ***workers on Amazon Mechanical Turk (MTurk)*** *to* ***verify whether each image correctly depicts the synset****. Due to the sheer scale of the annotation effort (over 50K workers and over 160M candidate images), an automated system was developed to download images from the search engines and assign the images to MTurk workers.*

http://image-net.org/update-sep-17-2019

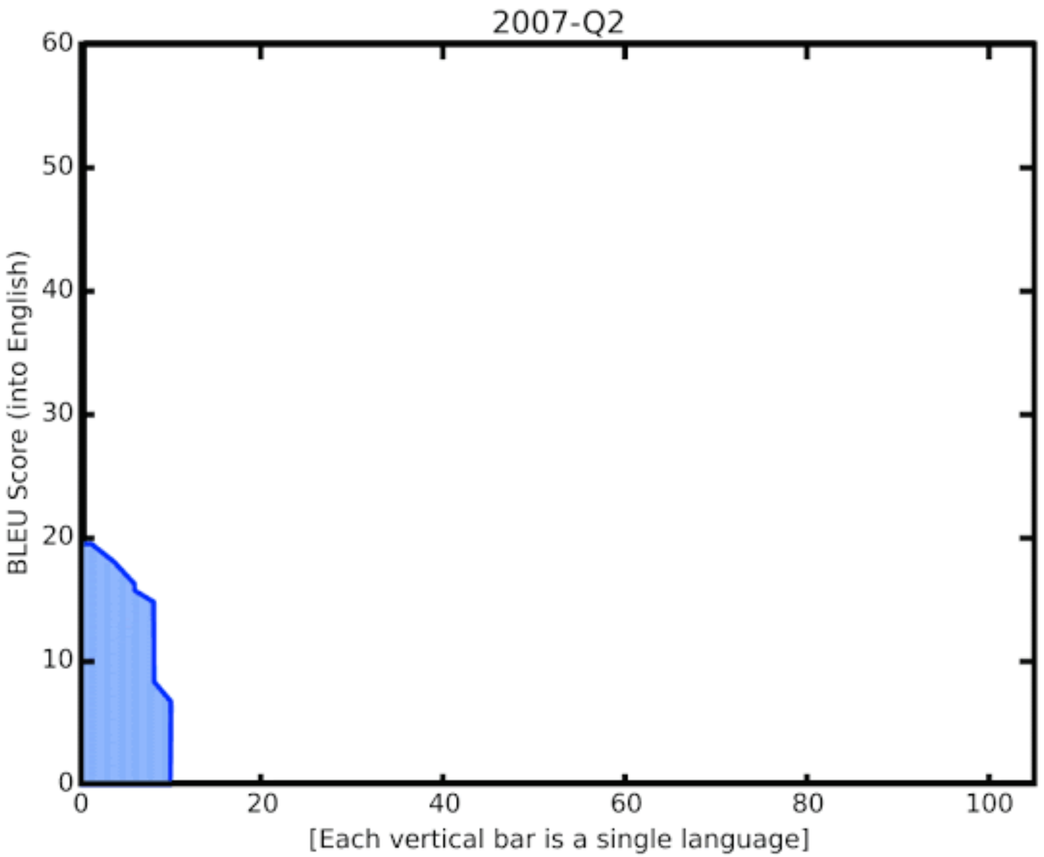
Building ImageNet

Feedback loop in action



Building Google Translate

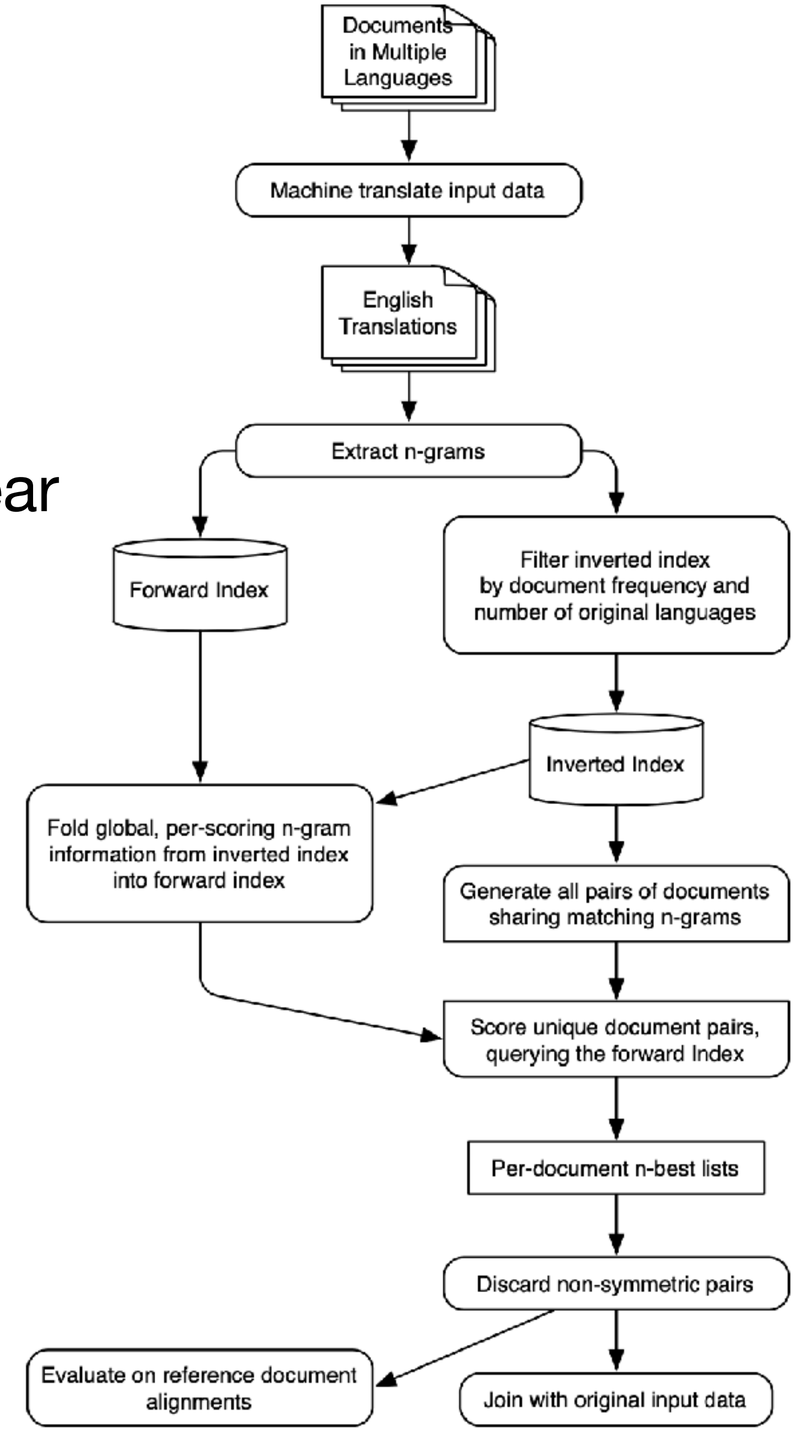
- Machine translation works quite well [Google Translate, 2020]



Building Google Translate

Data collection

- Because it works so well, new problems appear
  1. Where do we get more data from?
     - Web is filled with parallel documents
       - [Uszkoreit et al., 2010]
       - News articles, Wikipedia, e-commerce, etc.



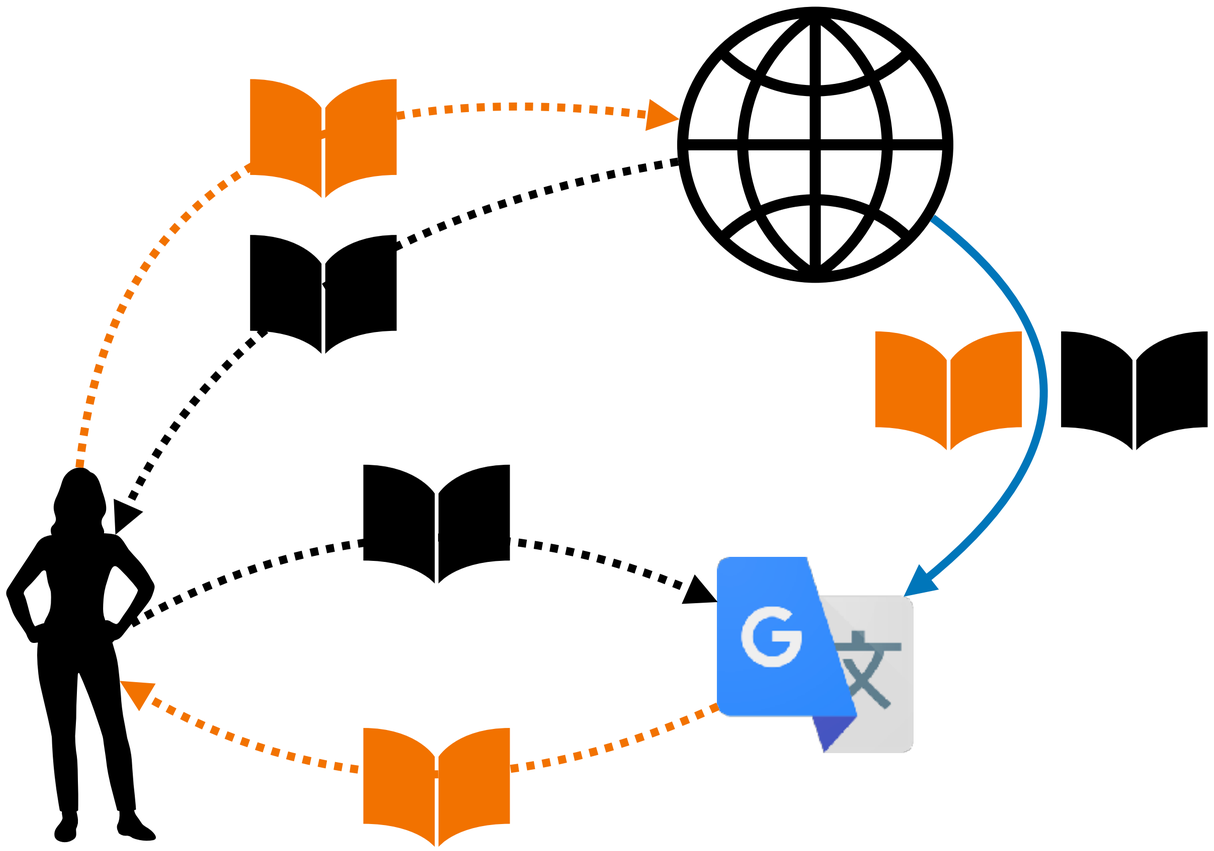
Building Google Translate

Contamination

- Because it works so well, new problems appear
  1. Where do we get more data from?
     - Web is filled with parallel documents
  2. Contamination by machine translated text
     - *"For each document, we inspected the **page content** for source code that indicates the **use of translation modules**/plug-ins … computed the proportion of the content within our corpus that uses these modules. We find that a **significant proportion** of the mined parallel data for some language pairs **is generated** … **Tagalog (50.6%)**, **Hindi (44.5%)** and **Galician (41.9%)** … high enough to **affect machine translations systems** …"* [Venugopal et al., 2011]

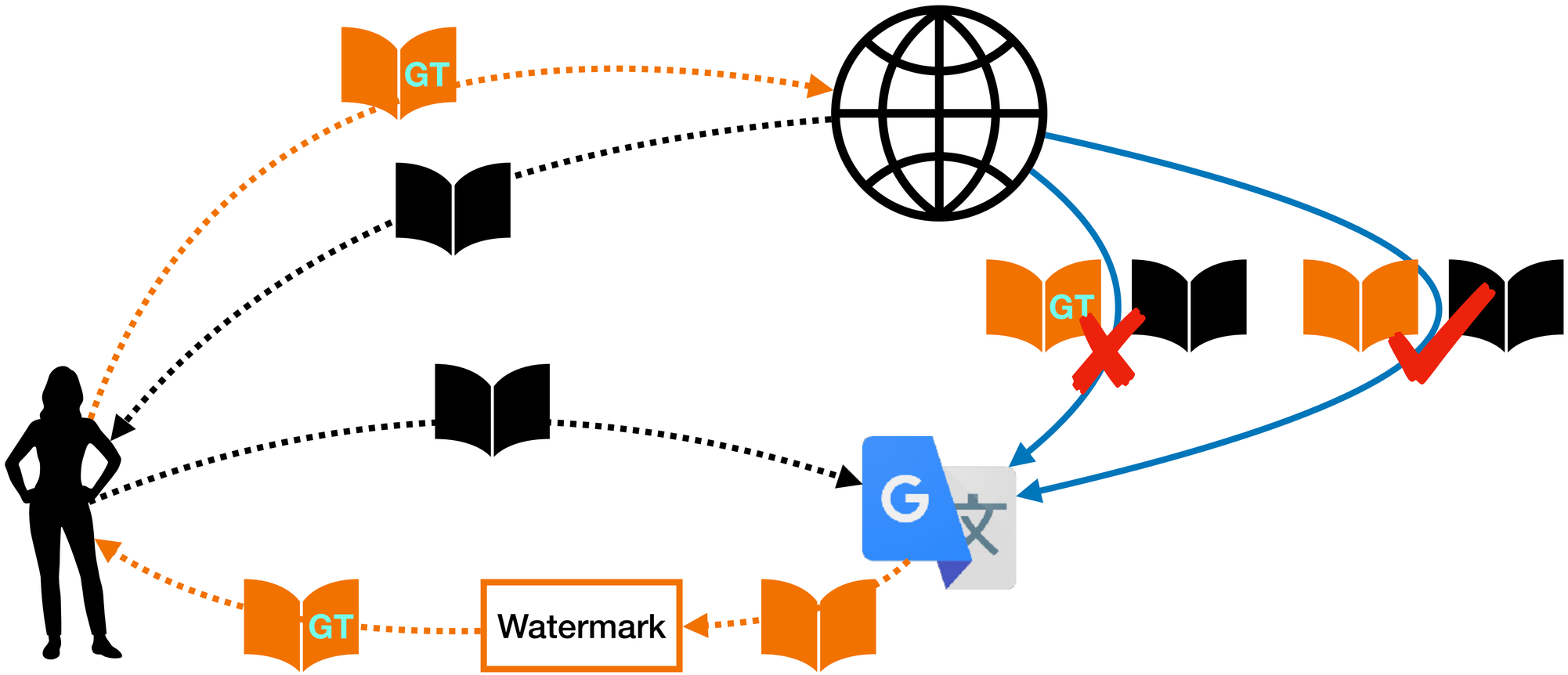
Google Translate

Vicious cycle



Google Translate

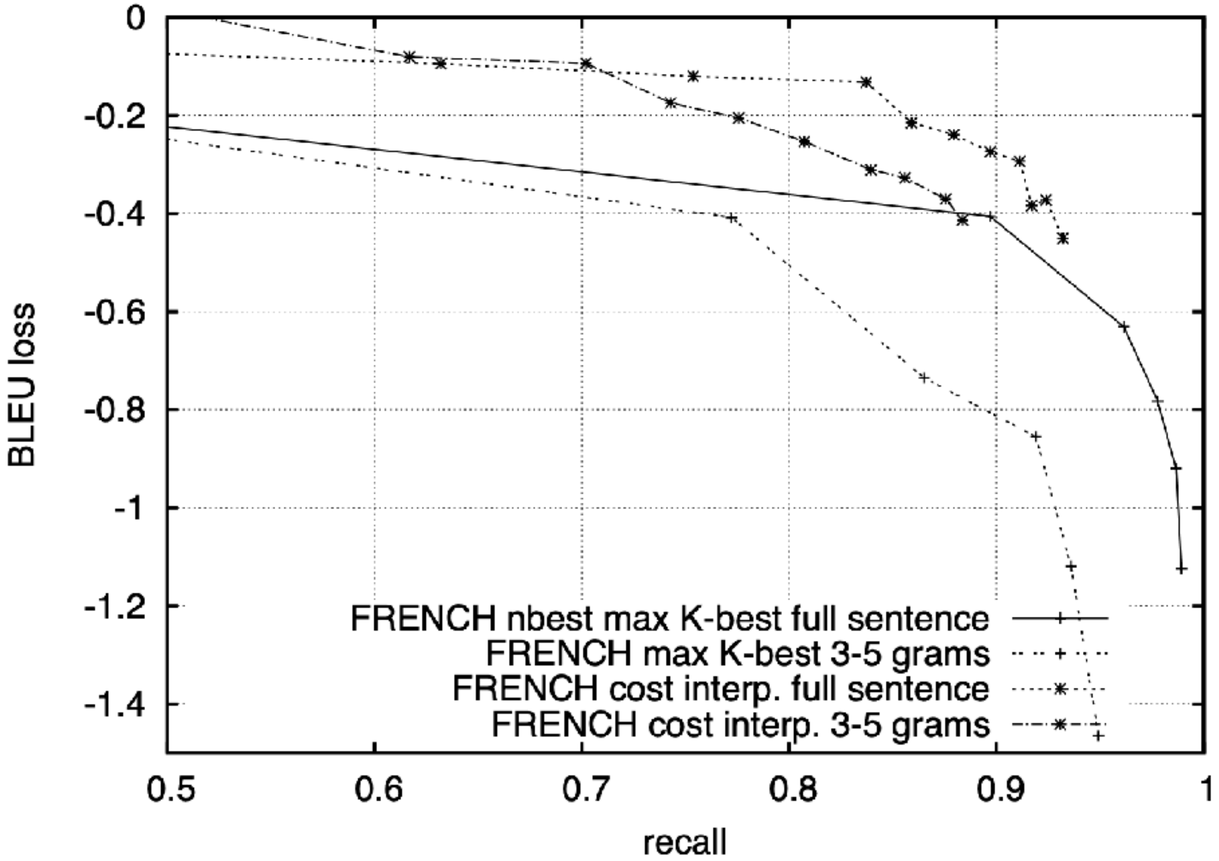
An algorithmic solution: statistical watermarking [Venugopal et al., 2011]



Google Translate

Watermarking may hurt quality

- Venugopal et al. (2011) show the trade-off between recall & quality



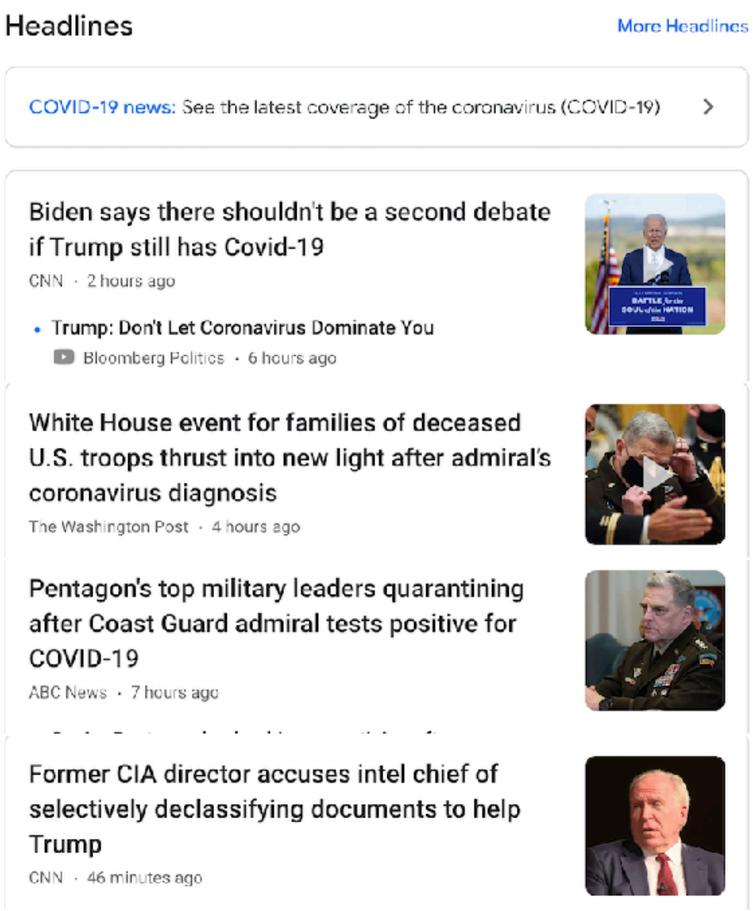
Google Translate

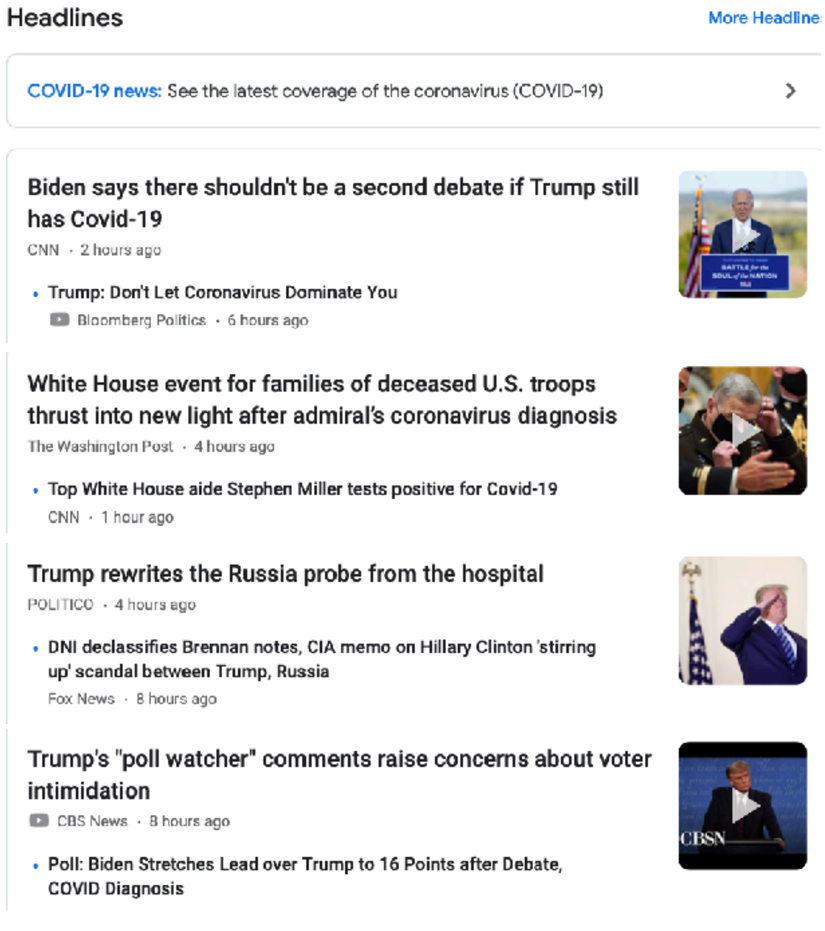
Watermarking, manipulation & privacy

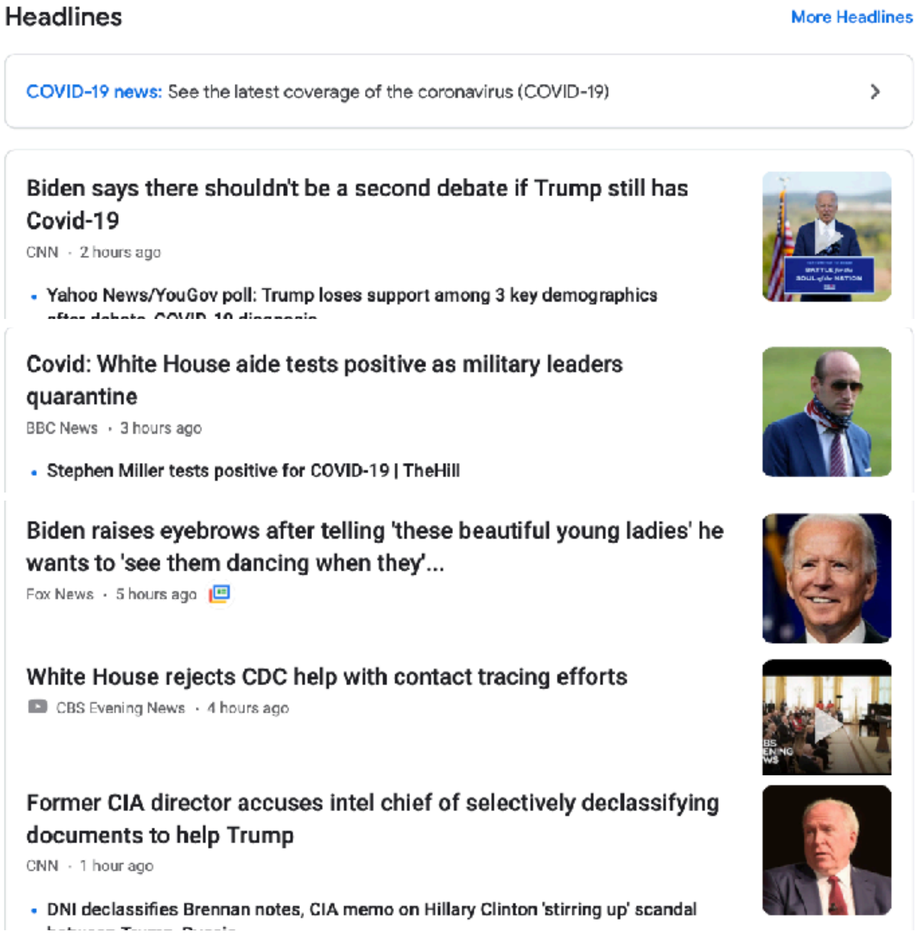
- With ubiquitous watermarking, one can trace who’s written what, and who’s read what, without any explicit tracking software (think of browser cookies)
- How much tracking should we allow?

Personalized feed

Do you read what I read?

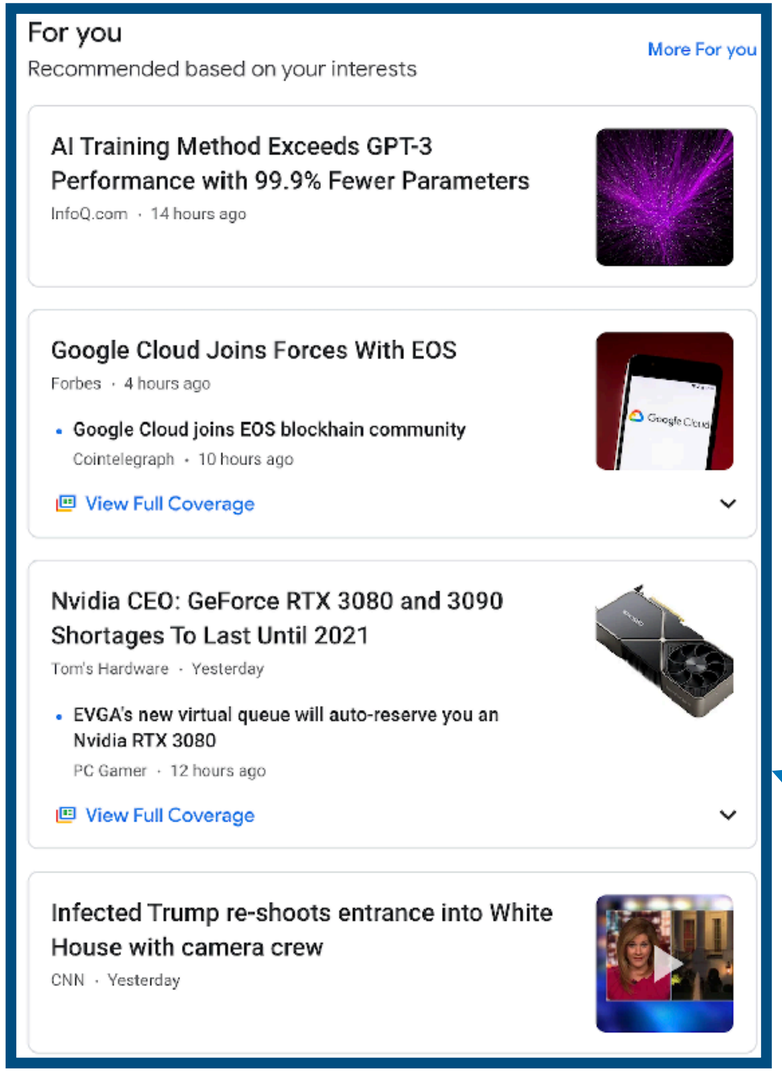




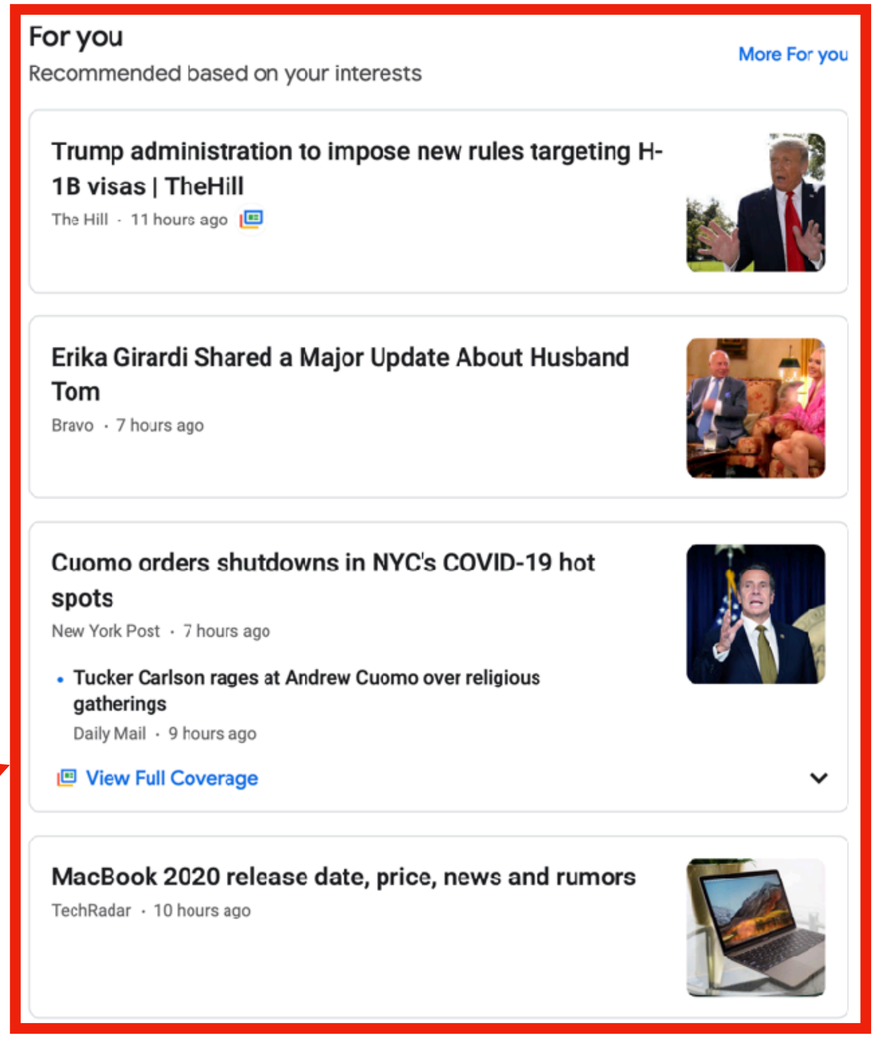


Personalized feed

How personalized is Google News?



Which one of these “For you” lists was for me?



Personalized feed

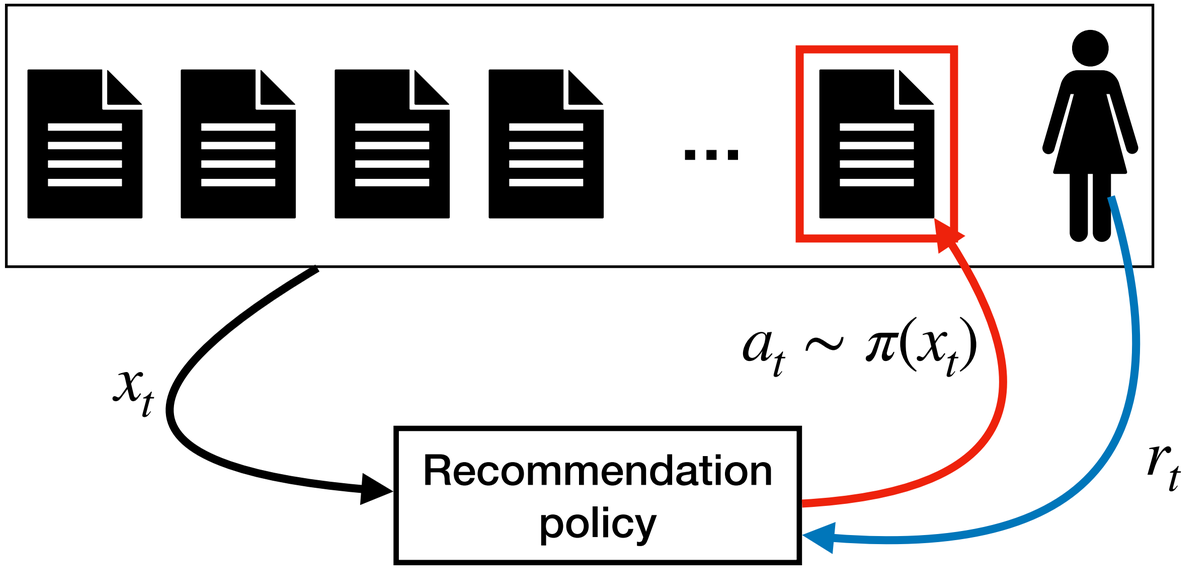
Exploration vs. Exploitation

- Supervised learning of click through rate doesn’t work
  - Because the classifier would only work well on *already recommended* articles.
  - If I had only been recommended and read AI-related articles, how would a classifier know whether i would want to read about the new EO on H1B?
- A good recommendation system must *explore* alternatives.
  - It is likely *detrimental* in a *short* term.
  - It could be *beneficial* in a *long* term.

Personalized feed

Exploration vs. Exploitation: contextual bandits [Li, Chu, Langford & Schapire, 2010]

- At each round $t = 1, \ldots, T$
  1. A context is revealed
     - $x_t \sim D_t$
  2. A policy chooses an action
     - $a_t \sim \pi(x_t) \in \mathscr{A}$
  3. The policy receives a reward
     - $r_t \sim R(r \mid x_t, a_t)$

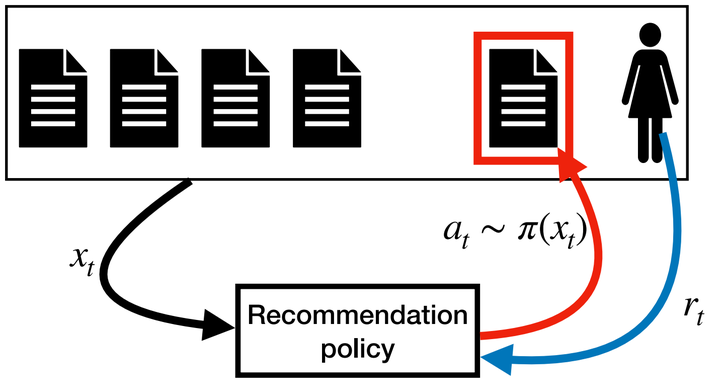


Goal: maximize $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{x \sim D_t} \left[\mathbb{E}_{a \sim \pi(x)} [r_t \mid x, a] \right]$

Personalized feed

Exploration vs. Exploitation: dilemma

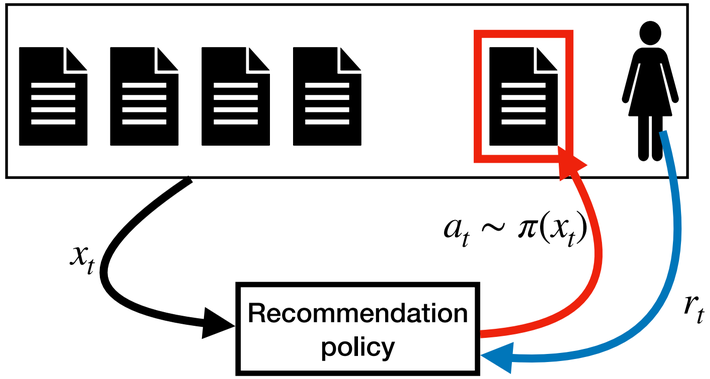
Goal: maximize $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x\sim D_t}\left[\mathbb{E}_{a\sim\pi(x)}[r_t\,|\,x,a]\right]$



- ϵ-greedy exploration
  - With probability ϵ, choose an article uniformly at random.
  - With probability $1-\epsilon$, choose $\arg\max_{a\in\mathcal{A}_t} f(a, x_t)$
- UCB (upper-confidence bound) exploration
  - Choose $\arg\max_{a\in\mathcal{A}_t} \mu(a, x_t) + \alpha\sigma(a, x_t)$

Personalized feed

Contextual bandits often lag behind …

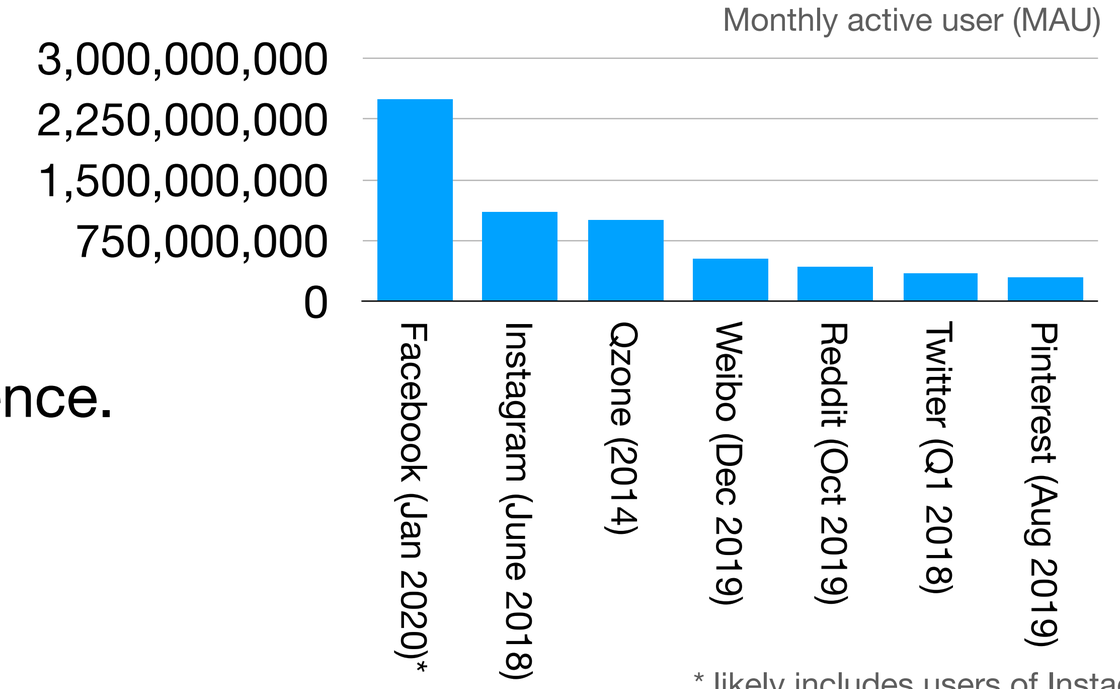


- Contextual bandits for online recommendation has a nice theoretical property
  - The regret is bounded by $O(\sqrt{KdT})$ under strong conditions
    - K: # of candidate articles, d: context dimensionality, T: # of interactions
- Such a property is not nice enough in practice, because
  - There are just **too many candidate articles** to choose from.
  - **Users' preferences and the world change rapidly**.
- Exploration cannot catch up with rapid drift of users' preferences and articles' topics.

Personalized feed

Massive-scale A/B testing

- Exploration often cannot catch up with rapid drift of users' preferences and articles' topics *if* exploration is done *serially* and *individually*.
- What if exploration is done **in *parallel* at a *massive scale***?
  - Test the impact of a small change on millions of users simultaneously.
  - A/B testing (randomized control trial) at a massive scale enables us to measure causal effects effectively with high confidence.



* likely includes users of Instagram

Personalized feed

Case study on emotional contagion in Facebook

- Kramer, Guillory & Hancock [2014] conducted A/B testing on Facebook to test
  1. **whether "*emotions .. are contagious*** *in the absence of direct interaction between experiencer and target.*"
  2. "***whether posts with emotional content are more engaging***."
- They "*manipulated the extent to which people were exposed to emotional expressions in their News Feed.*"
  - "***N = 689,003***" "*for 1 wk*"
  - They **dropped "*each emotional post*" with a probability** "*between a 10% and 90% chance*" from "*News Feed*".
  - "*Posts were determined to be positive or negative if they contained at least one positive or negative word …*"
- They observed
  - "***emotional contagion***": "*the emotions expressed by friends, via online social networks, influence our own moods*"
  - "***withdrawal effect***": "*People .. exposed to fewer emotional posts … were less expressive .. on the following days*"

Personalized feed

Personalization, exploration and ethical concerns

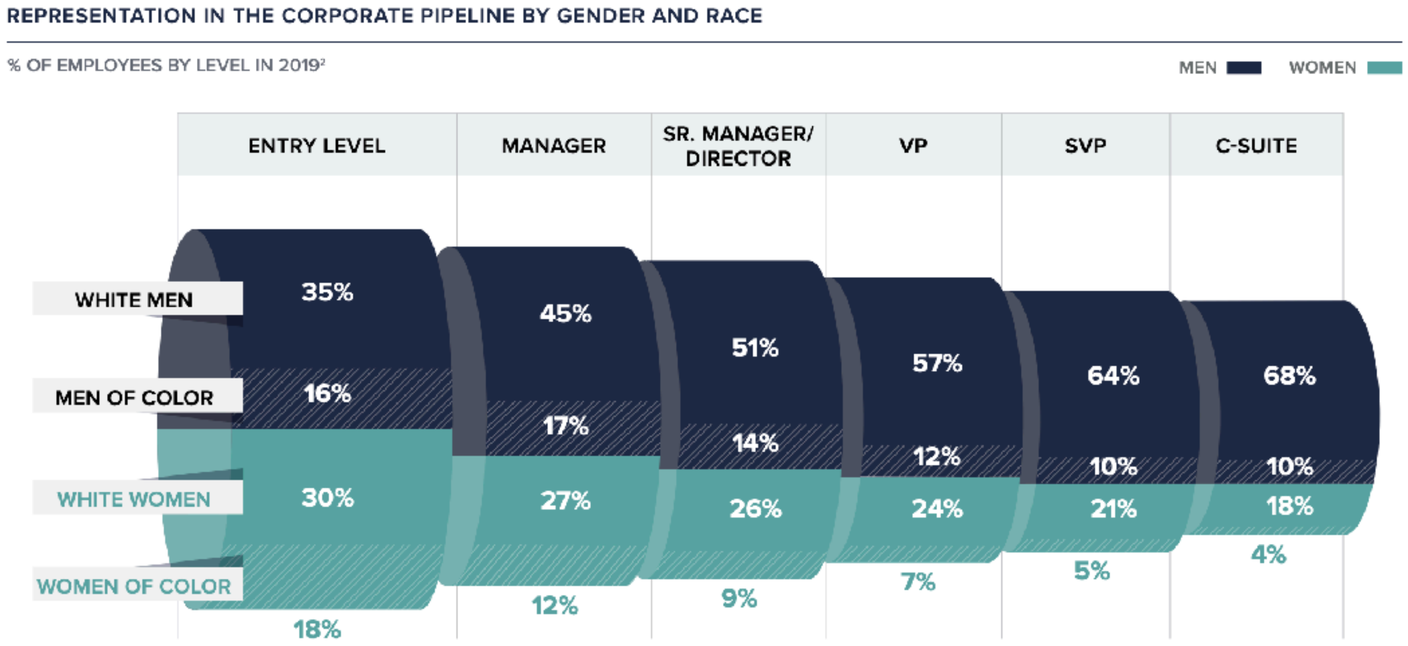
- They observed "*withdrawal effect*": "*People .. exposed to fewer emotional posts … were less expressive .. on the following days*"
- **Hypothetically**, if you were a decision maker at FB, what would you've done?
- Should such exploration be allowed? If so, to which degree?

Checkmate: Hiring

When you cannot watermark, explore nor experiment…

- *"Amazon's system … penalized … the word "women's," as in "women's chess club captain." … downgraded graduates of two all-women's colleges." "… no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory"*

<Amazon scraps secret AI recruiting tool that showed bias against women>
by Jeffrey Dastin



<Women in the Workplace 2019> McKinsey & Co.

Discussion

AI in deployment likely introduces a vicious cycle

- Society 🔄 AI
  - AI systems are not static: they evolve as the surrounding environment evolves.
  - The surrounding environment is not static: it is affected by the AI systems.
- How do we ensure this cycle remains virtuous?

Concluding words

Know your data

Datasets for Datasets [Gebru et al., 2018]

- Pick one dataset you recently used. Let's answer the Q's below.
  1. Who created the dataset? How can they be contacted?
  2. Is the dataset self-contained, or does it link to or otherwise rely on external resources?
  3. Are there recommended data splits? please describe these splits and the rationale.
  4. Who was involved in the data collection process and how were they compensated?
  5. Over what timeframe was the data collected?
  6. Is the software used to preprocess/clean/label the instances available?
  7. Are there tasks for which the dataset should not be used?

Know your model

Model cards for model reporting

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

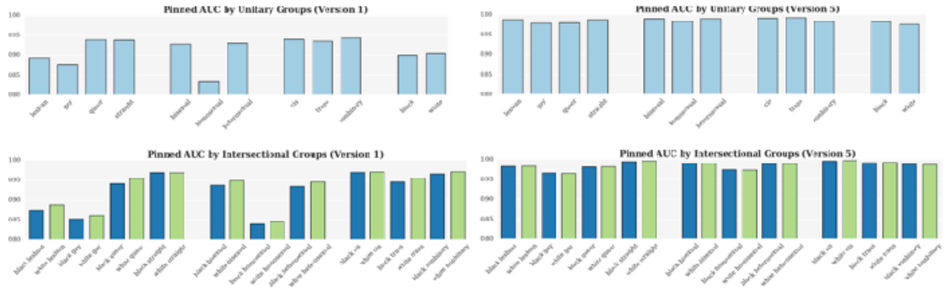
Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

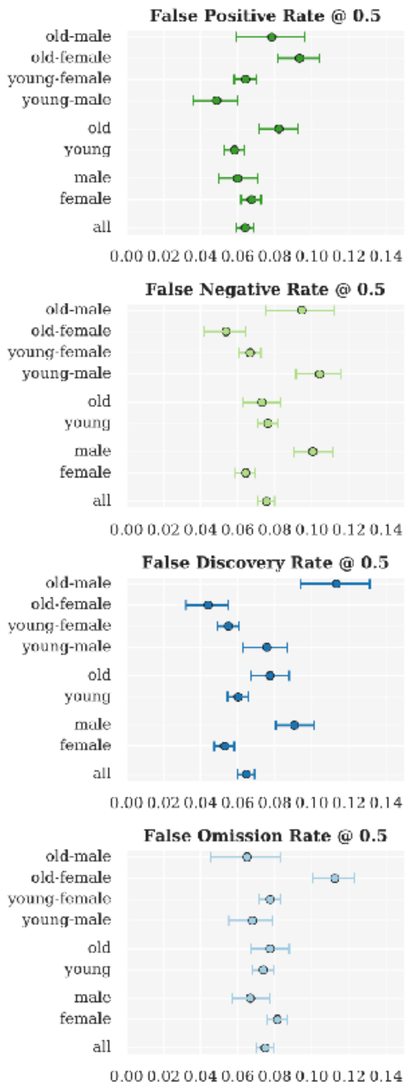
Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses



Mitchell et al. [2018]

Imagine consequences

Value mis-alignment

- You've done everything right. What can go wrong?
- e.g., Growth & engagement
  - More engaging users→more social interaction→happier users→better society
  - Thereby, maximize the growth of user base & engagement level of users.

Total Number of Monthly Active Instagram Users:

1 billion + (source)

Last updated: 1/25/20

The number of Photos & Videos uploaded per day:

100 million+

Last updated: 1/25/20

Bringing you closer to the people and things you love

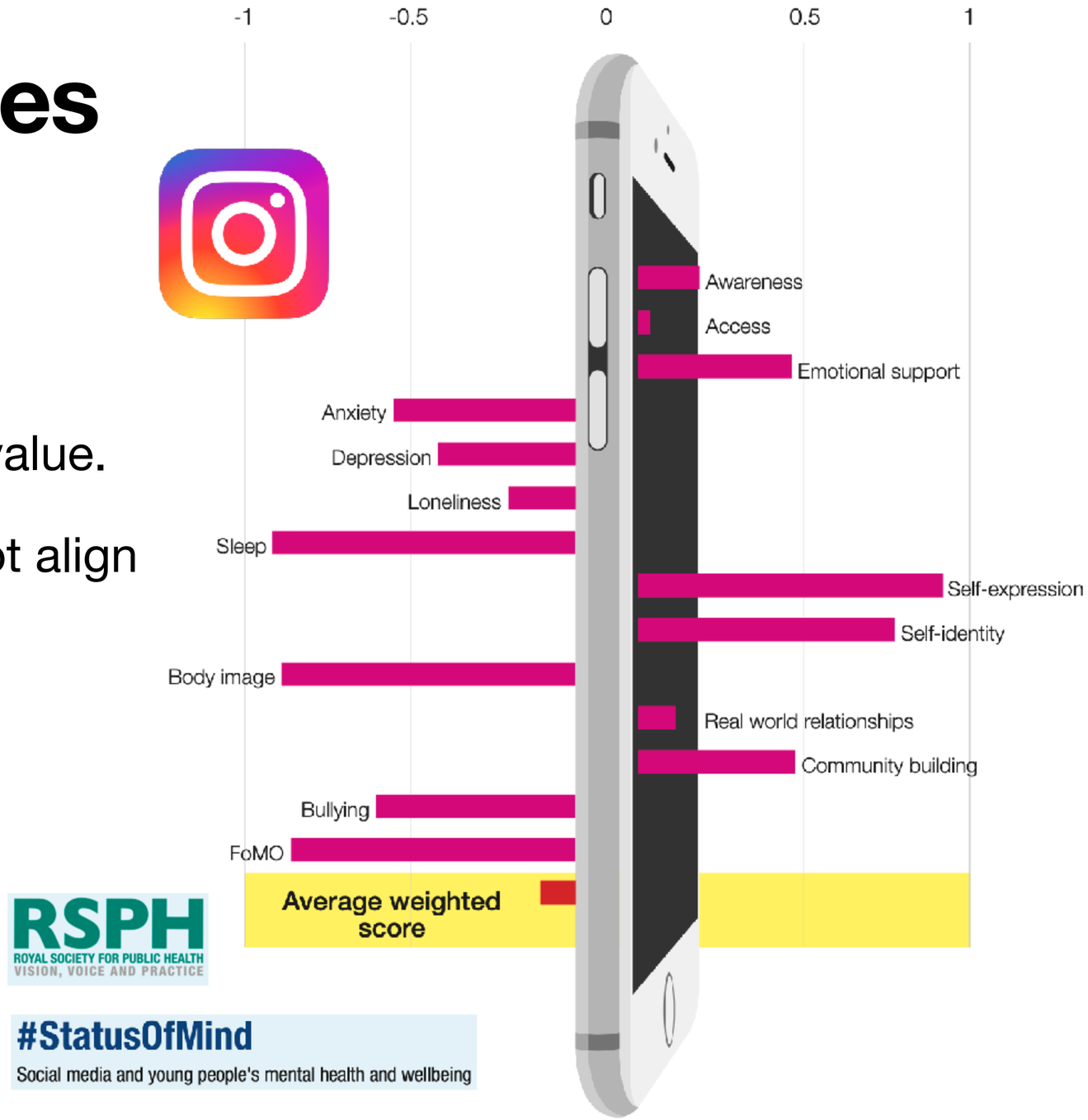
Express yourself in new ways with the latest Instagram features.

Connect with more people, build influence, and create compelling content that's distinctly yours.

Imagine consequences

Value mis-alignment

- True value is never known
  - We build an AI system with a proxy value.
  - A perfect AI system's value would not align perfectly with the true value.
- Imagine both positive & negative consequences



Five recommendations

<Ethical Machine Learning in Health Care> by Chen et al. (2020)

1. Problems should be tackled by **diverse teams** and using frameworks that increase the probability that **equity** will be achieved

2. **Data collection** should be framed as an important **front-of-mind concern** in the ML modelling pipeline

3. **Outcome choice** should reflect the task at hand and should preferably be **unbiased**.

4. Reflection on **the goals of the model** is **essential** during development, and should be articulated **in a pre-analysis plan**.

5. **Audits** should be designed to **identify specific harms**, and **paired with methods and procedures**.