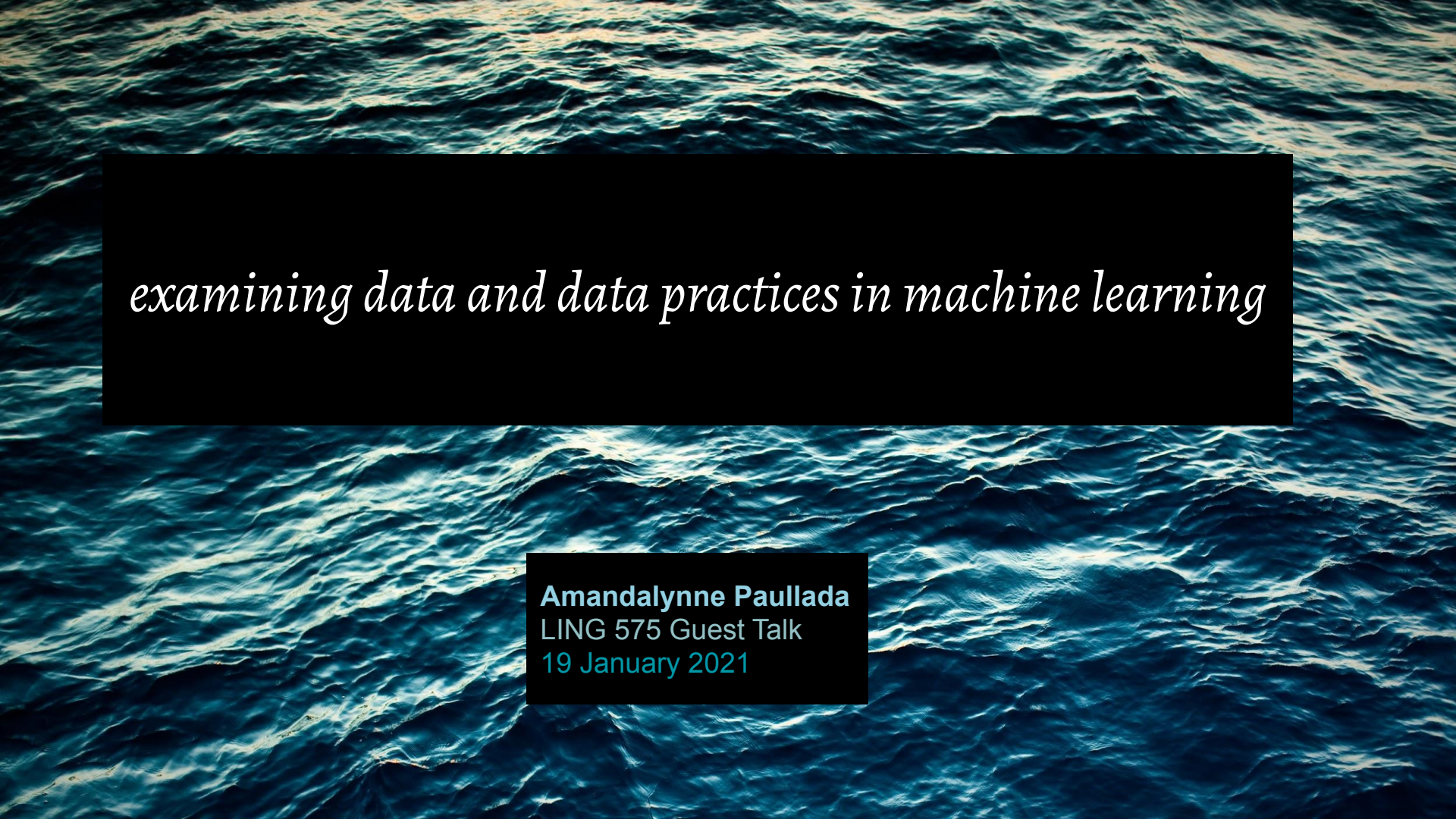
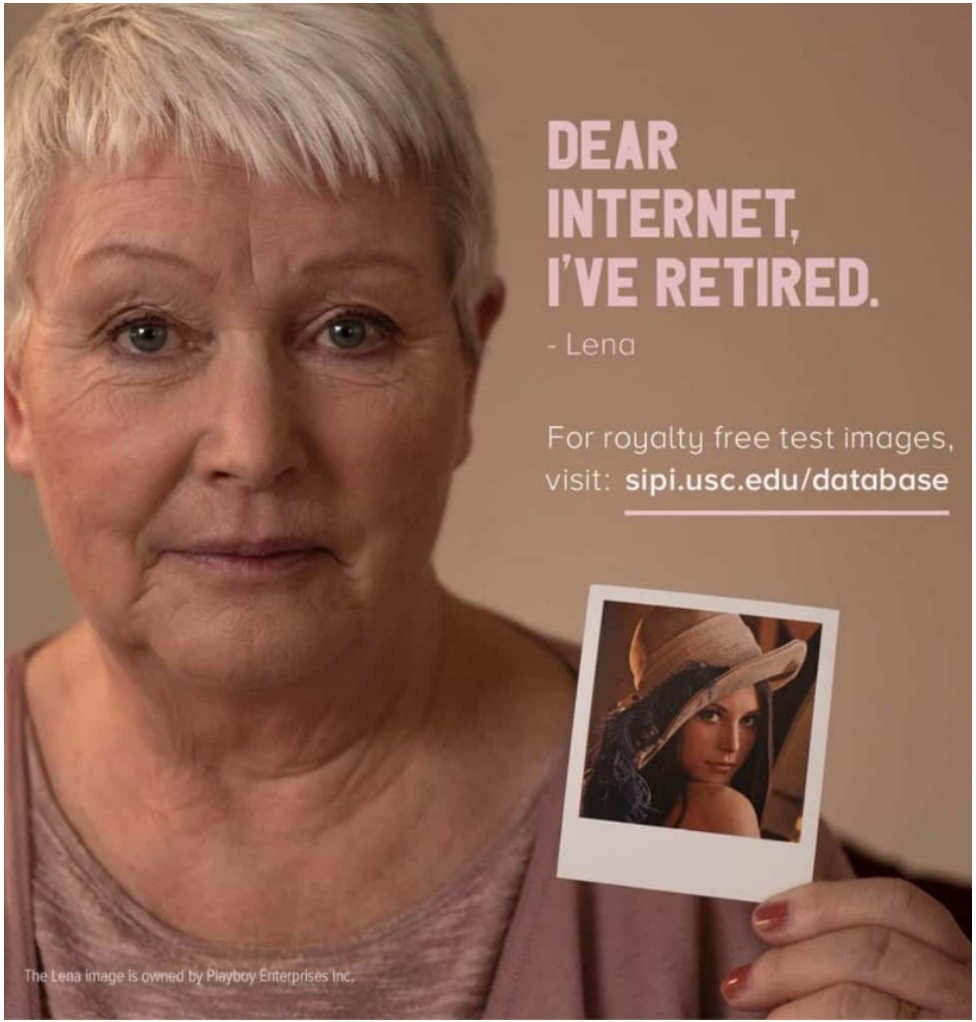

examining data and data practices in machine learning
Amandalynne Paullada
LING 575 Guest Talk
19 January 2021

examining data and data practices in machine learning

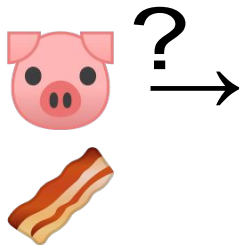
general content warning

Amandalynne Paullada
LING 575 Guest Talk
19 January 2021


DEAR
INTERNET,
I'VE RETIRED.
- Lena
For royalty free test images,
visit: sipi.usc.edu/database
The Lena image is owned by Playboy Enterprises Inc.

As data-driven machine learning algorithms demand increasingly huge collections of data, there is a tendency to avoid attending to the contents and provenance of the data. Vinay Prabhu calls this **the abattoir effect** -- everyone wants to use datasets, but they don't care to know how they were made.

Uncritical use of datasets obscures details like:



- human labor, human subjectivity & positionality
- the manner in which the data was produced & collected
- actual contents of the dataset

These details are useful for assessing whether and how a dataset might be useful for a particular application, and for enabling more systematic error analysis.

Roadmap

- A closer look at dataset contents
  - Case study 1: Image descriptions and social biases
  - Case study 2: Textual entailment and spurious cues
  - Focusing on American English data
- Some qualitative and quantitative approaches to auditing datasets

By no means a comprehensive survey!

Case study 1:

Image descriptions and social biases

Flickr30k Dataset (Young et al., 2014)

- 30,000 images from Flickr, each with 5 crowdsourced descriptions
  - Ostensibly limited to US annotators
- Goal: "visual denotations of linguistic expressions"
- *Not* the original captions from Flickr -- trying to solicit descriptions that contain only information that can be inferred from the image
  - "Our beloved Skippy" → "a black and white dog sitting on a sofa"
- Can be used to train models for tasks such as caption generation

Young et al (2014), *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*

Image descriptions from Flickr30k

Describe the image:

Figure 1: Image 8063007 from the Flickr30K dataset.

Image descriptions from Flickr30k

1. A blond girl and a bald man with his arms crossed a standing inside looking at each other.

2. A **worker** is **being scolded** by her **boss** in a **stern lecture**.

3. A **manager talks to an employee about job performance**.

4. A hot, blond girl **getting criticized by her boss**.

5. Sonic employees **talking about work**.

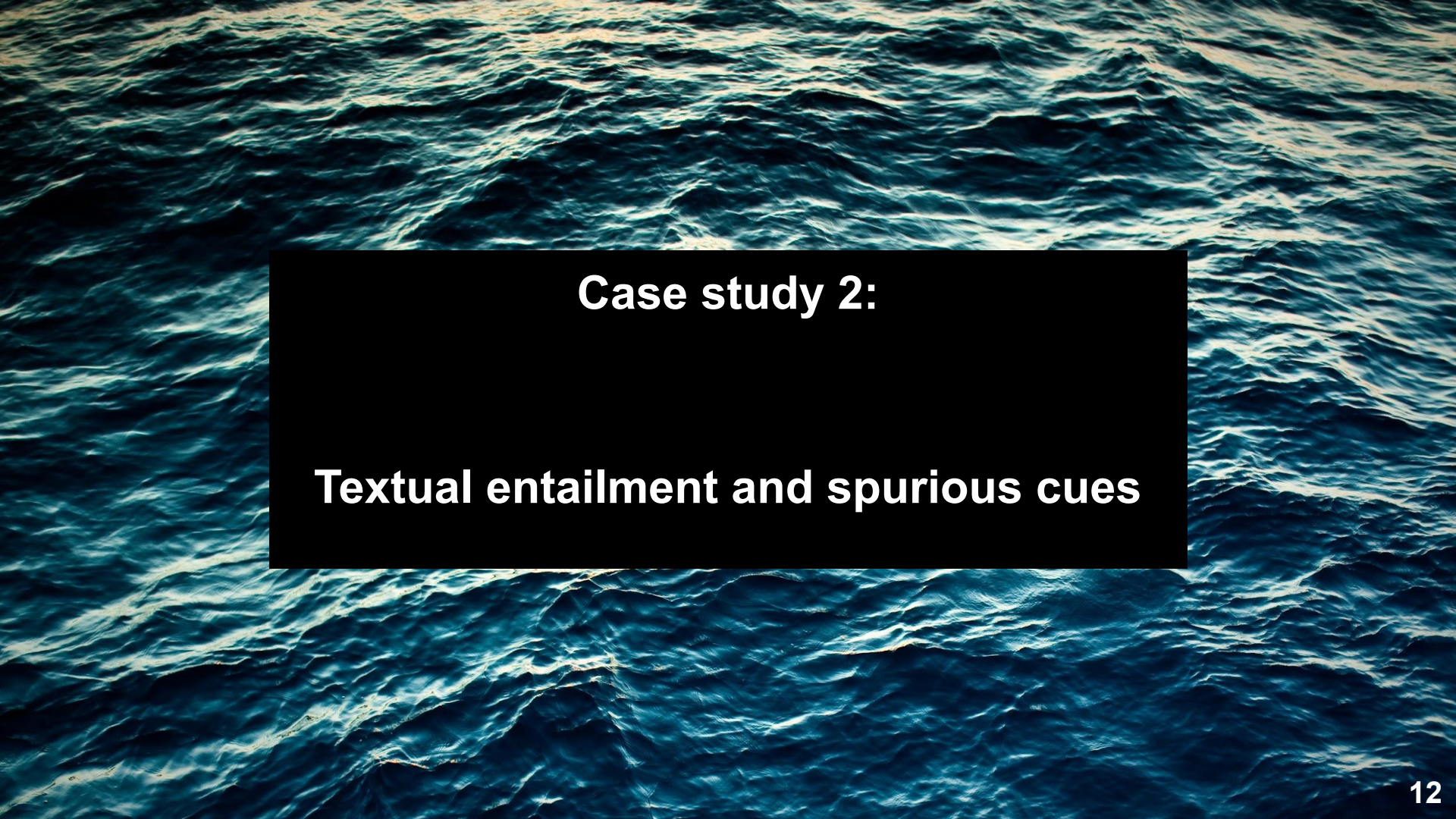
Figure 1: Image 8063007 from the Flickr30K dataset.

emphasis & figure from van Miltenburg (2016), *Stereotyping and Bias in the Flickr30K Dataset*

Sources of bias in Flickr 30k (van Miltenburg 2016)

- **linguistic bias**: "a systematic asymmetry in word choice as a function of the social category to which the target belongs" (Beukeboom 2014)
  - The labels "Black baby" and "Asian baby" appear with higher frequency than "White baby" in the dataset, but there are vastly more White babies (n=504) depicted in the data. (vs. 66 Asian, 36 Black babies)*
  - The US-based annotators treated "White" as an unmarked default, revealing a (racist) reporting bias
- **unwarranted inferences**: "speculation about the image"
  - Previous example: how do we know who is the "manager" in the photo, or if the individuals are in a hierarchical professional relationship at all?

* based on manual labels from author of paper


Case study 2:
Textual entailment and spurious cues

Natural Language Inference (Textual Entailment)

Task: Given a premise X and a sentence Y, predict whether their relationship is entailment, neutral, or contradictory.

Premise: "Billy is my favorite grandchild."

1. "I am Billy's grandparent."
2. "Billy visited me for my birthday."
3. "I have no grandchildren."

Entailment: If X is true, Y must also be true.

Neutral: The truth of X and the truth of Y are unrelated.

Contradiction: X and Y cannot both be true.

Natural Language Inference (Textual Entailment)

Task: Given a premise X and a sentence Y, predict whether their relationship is entailment, neutral, or contradictory.

Premise: "Billy is my favorite grandchild."

1. "I am Billy's grandparent." **Entailed**
2. "Billy visited me for my birthday." **Neutral**
3. "I have no grandchildren." **Contradiction**

Can a machine do this?

Entailment: If X is true, Y must also be true.

Neutral: The truth of X and the truth of Y are unrelated.

Contradiction: X and Y cannot both be true.

Stanford Natural Language Inference (Bowman et al 2015)

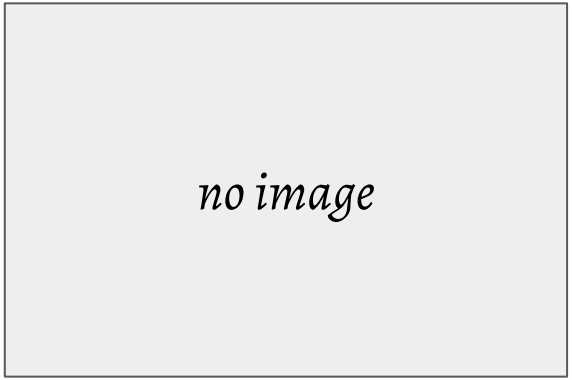
Premises sourced from photo captions in an existing dataset (Flickr 30k). Annotators shown only caption, no photo.

Instructions for human annotators:

"Write one alternate caption that…
...is definitely a true description of the photo.

...might be a true description of the photo.

...is definitely a false description of the photo."



Premise: "A black and white dog with a stick in its mouth is swimming."

Stanford Natural Language Inference (Bowman et al 2015)

"Write one alternate caption that…
...is definitely a true description of the photo."

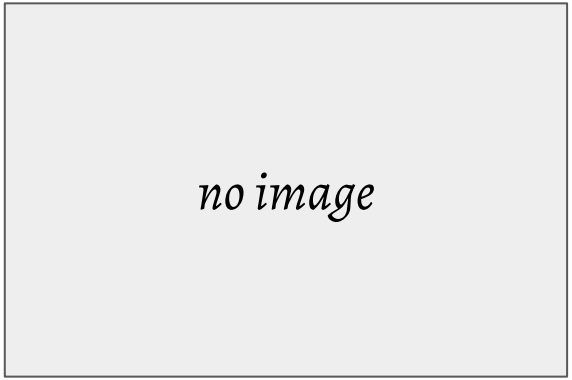
A dog with an object in its mouth is in the water.

"...might be a true description of the photo."

A Dalmatian with a stick in its mouth is swimming slowly.

"...is definitely a false description of the photo."

A cat with a stick in its mouth is in a bathtub.



Premise: "A black and white dog with a stick in its mouth is swimming."

Stanford Natural Language Inference (Bowman et al 2015)

"Write one alternate caption that…
...is definitely a true des iption of the pho

A dog with an object in its m

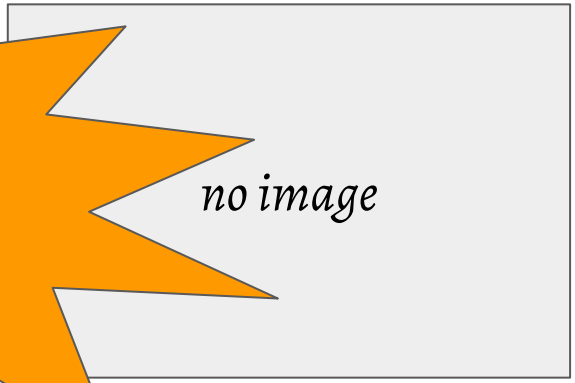
"...might be a true descrip

A Dalmatian with a st
swimming slowly.

"...is definitely a false description of the ph ."

A cat with a stick in its mouth is in a bathtub.

>90% Test Set Accuracy!!
we have achieved sentience



Pre ise: "A black and white dog with a stick in its mouth is swimming."

Breaking NLI models

Poliak et al (2018) and Gururangan et al (2018) found that:

- a model trained to predict entailment relationships *without* ever using the premise sentences performs pretty well!
- presence of the word "cat" was correlated with predictions of "contradiction" (among other so-called artifacts)

The Flickr30k dataset contains lots of 'dog' photos. An 'availability heuristic' for human annotators -- dog vs. cat dichotomy?

The annotators weren't doing anything wrong, of course!

For a summary of issues with NLI datasets and a survey of work done to mitigate these issues, such as adversarial data augmentation, look to Schlegel et al (2020), *Beyond Leaderboards: A survey of methods for revealing weaknesses in Natural Language Inference data and models*

Another interesting perspective: Pavlick & Kwiatkowski (2019), *Inherent Disagreements in Human Textual Inferences* suggest that inter-annotator disagreement on entailment validity is a feature, not a bug


Approaches to understanding datasets

Frameworks for documenting & interrogating data

Data Statements for Natural Language Processing (Bender & Friedman 2018):

- What language varieties are included? Who was involved in the dataset production? ...

Bringing the People Back In: Contesting Benchmark Machine Learning Datasets (Denton et al 2020):

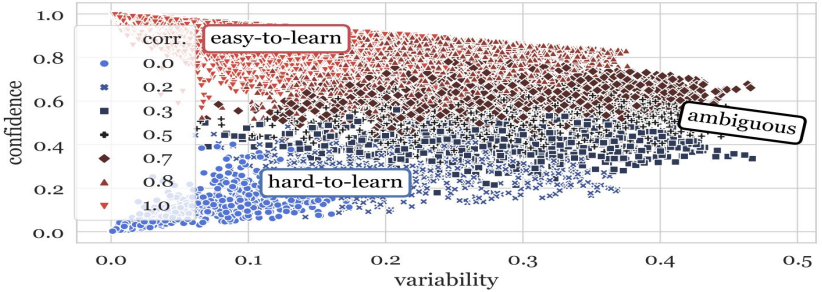
- Why was the dataset created, and in what context? ...

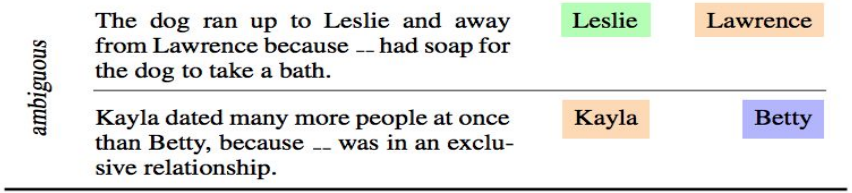
Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (Olteanu et al 2020):

- At what points in the pipeline can bias manifest? What kinds of bias? ...

Dataset Cartography (Swayamdipta et al 2020)

- A model-based tool to characterize and diagnose datasets based on training dynamics
- Produce a 'map' of data to point to easy, hard, and ambiguous instances in the dataset





Figures from Swayamdipta et al (2020), *Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics*

What about pretrained language models?

Embeddings from pretrained language models are trained on massive amounts of data; it would be intractable to examine every document, and the structure of the task doesn't lend very itself well to mapping.

We can make qualitative judgments about sources of training data. For example, GPT-2 (Radford et al 2019) was trained on text linked to from Reddit posts with "at least 3 karma." (OpenAI changed this strategy for GPT-3 (Brown et al 2020))

"Toxicity Prompts" (Gehman et al 2020) are a way to reveal ideological biases in a language model, and highlight the need for more critical choices of pretraining data.

Know Thy Corpus! (Sharoff 2020): techniques such as topic models on the corpus

Artistic interventions

This Person Does Exist by @_pitscher

This Person Exists by Vincent Woo

These pages show you randomly selected images of the real people whose faces were used to train StyleGAN.

*Neither of these people exist.

Everest Pipkin, On *Lacework*

"Moments in Time" dataset (Monfort et al 2019):

- 1 million 3-second videos obtained from various sources by searching keywords based on 339 verbs
- Annotators given binary question about video content

Lacework: an art project based on Moments in Time.

Pipkin looked at 60,000 videos selected for verbs related to 'touch'

Everest Pipkin, On *Lacework*

"When I first started watching the dataset I assumed that the team of researchers who had put it together at MIT had seen the bulk of it, but I'm now convinced that assumption was wrong. This is because so much of the archive is so, so hard to watch.

[There are only two options for an annotator interacting with a video]: include, discard. No ability to report a video, reclassify it or clarify its inclusion, no middle ground."

Everest Pipkin, On *Lacework*

"When you ask:

Does a dog fight match ‘barking’?

Does a sexual assault match ‘kissing’?

Does a police murder match ‘arresting’?

Sometimes the answer is Yes."

Takeaways

"Datasets are the results of their means of collection."

- Mimi Ọnụọha, *The Point of Collection*

- Interesting, surprising, and sometimes disturbing insights arise when we probe our data. There is value in inspecting the data, and there are useful techniques for helping you decide where to begin looking.
- The more familiarity you have with your data, the more informed your insights will be about how to improve data quality and usefulness.
- Making datasets is hard! Iterate, document, think critically about sources and task setup.

Takeaways

For users of datasets

- Interrogate your data using qualitative and quantitative practices in tandem.
- Seek out as much documentation as you can piece together on the conditions of dataset creation.
- Use tools and your own judgment to examine the data and diagnose potential issues.

For producers of datasets

- Document rigorously, using frameworks such as Data Statements.
- Consider task structure: what affordances are made available to human labelers (where relevant).
- Consider data sources.
- Run pilots to iterate on data collection.

More tools and resources

Responsible AI Practices from Google AI

Data Collection + Evaluation worksheet from Google PAIR

AI Fairness 360 from IBM

thanks

get in touch: paullada@uw.edu

Lena looking at research using her image