On Data: Considerations

CS 490A, Fall 2021

Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs490a_f21

Brendan O'Connor & Laure Thompson

College of Information & Computer Sciences
University of Massachusetts Amherst

Administrivia – HW Submissions

- 5 total late days, up to 3 can be used on a **single** assignment
  - Only 3 late days may be applied to the collective HW2 deadlines!
- No late submissions accepted once late days are exhausted
- Extensions may be granted for specific circumstances
  - Religious holidays, unforeseen circumstances (illness, computer / internet issues, personal emergency, …)
  - **Not** job interviews, other coursework
  - When in doubt, as instructors (via email or private Piazza post)

Administrivia – HW Submissions

- Locked out of Gradescope?
  - Create a **private Piazza post** ASAP and **include your submission** as part of the post (via upload/attachment).
- Trouble generating PDFs?
  - Exporting to pdf in Jupyter has **non-Python dependencies**

  A. Install needed dependencies, come to office hours for assistance

  B. Export to html and then generate pdf separately

     1. In Jupyter, export notebook to .html
     2. Open .html file in browser, print to file (as pdf)

Administrivia – HW2

- **Early deliverable**: Peer-annotated data due **Tuesday, October 5th**
  - Submission: 2 CSVs, one for each annotator
- What is the overall annotation task?

X: text?, tweet

y: label, set $\mathcal{L}$

$f(x) \rightarrow y$

Administrivia – HW2

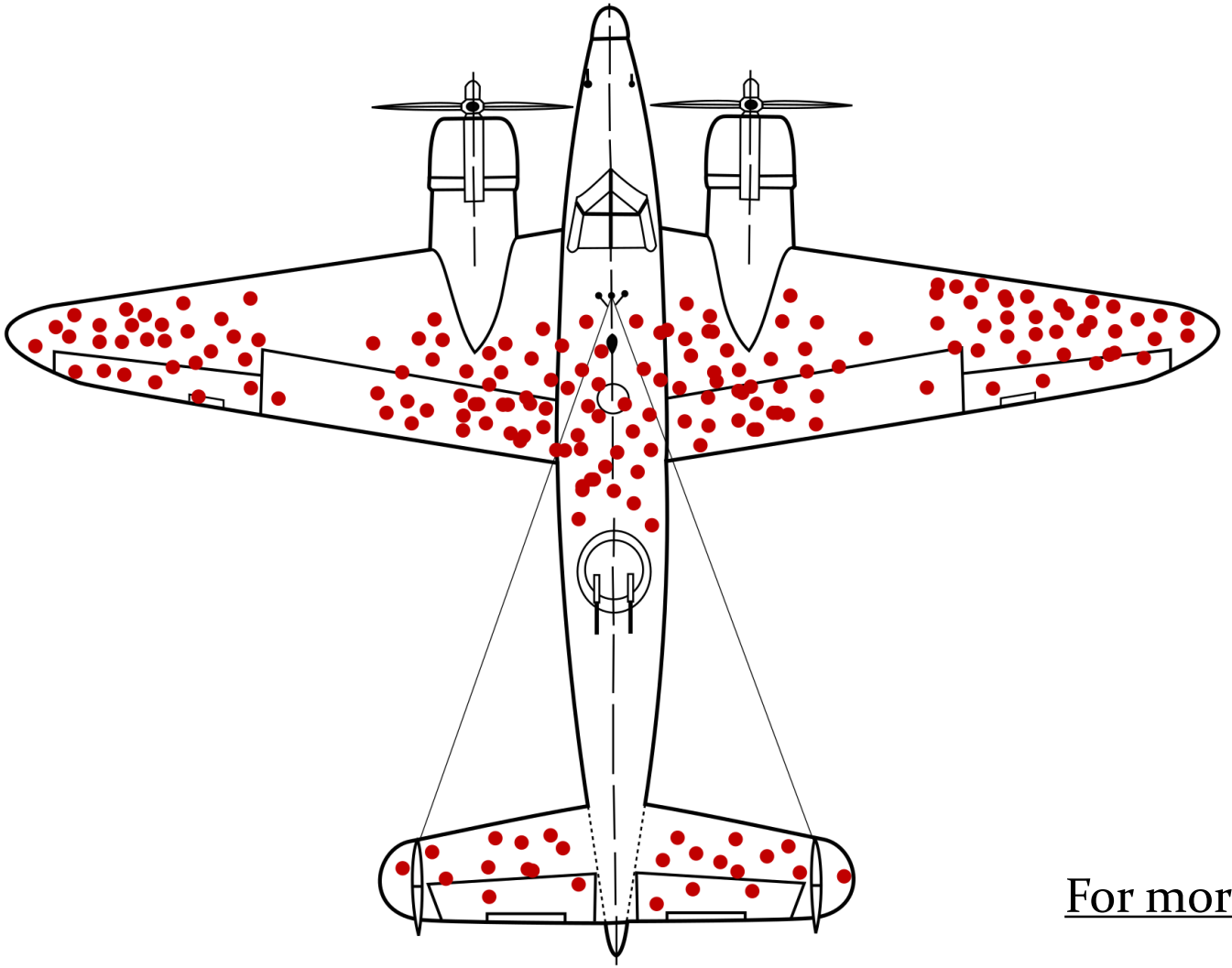
- Each annotator labels your full dataset of 250 tweets
- Choosing label set:
  - No more than 5 categories
  - Avoid rare categories, say less than ~5%
    - 10 tweets → 4%
    - 15 tweets → 6%

Administrivia - Health

- If you feel unwell, **do not come to class**
- Reminder: Complete the COVID-19 **daily** self-checklist before coming to campus/class
- Flu shots are now available. Book an appointment via the Campus Health Hub

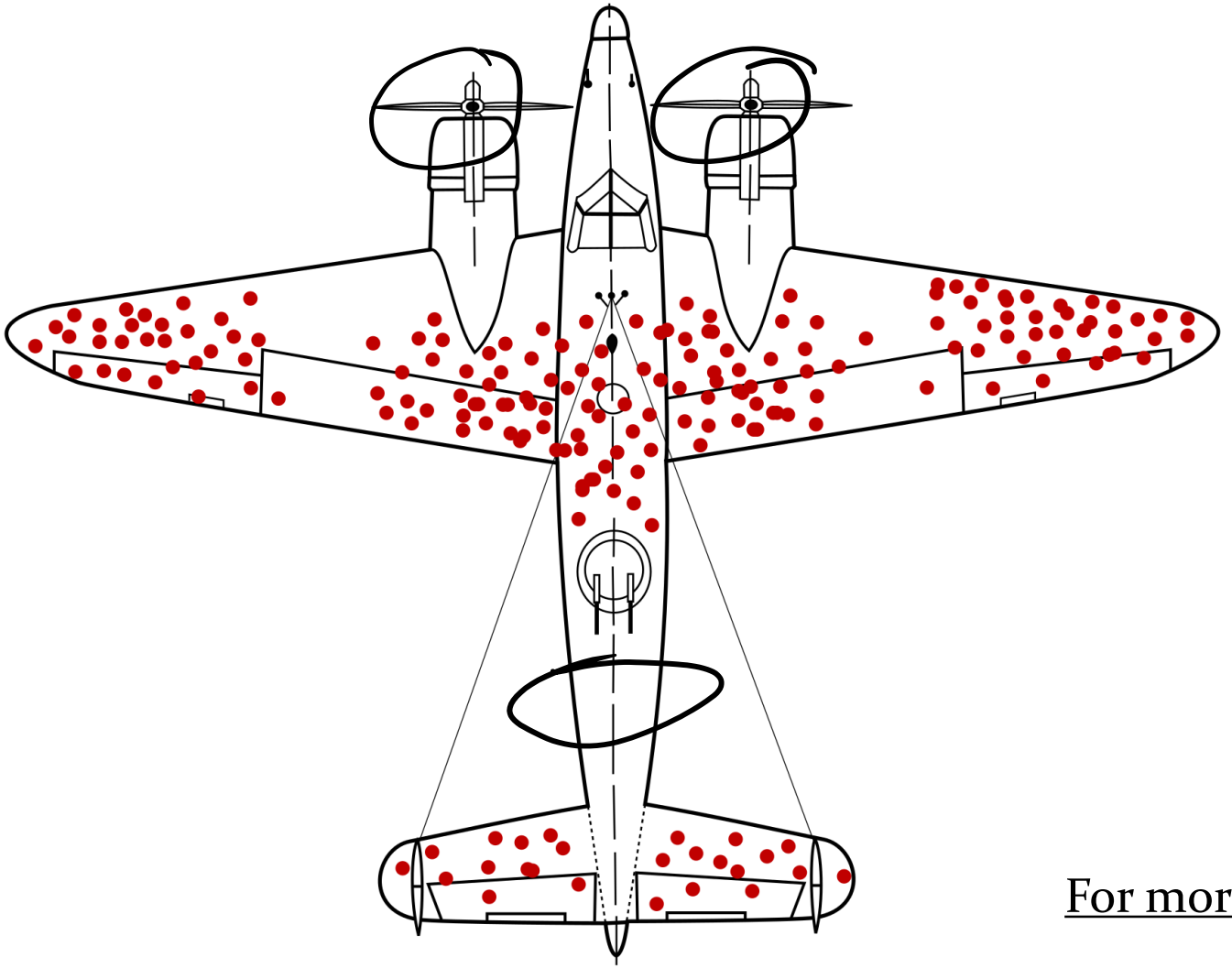
There’s no such thing as raw data.

What does this data tell us?



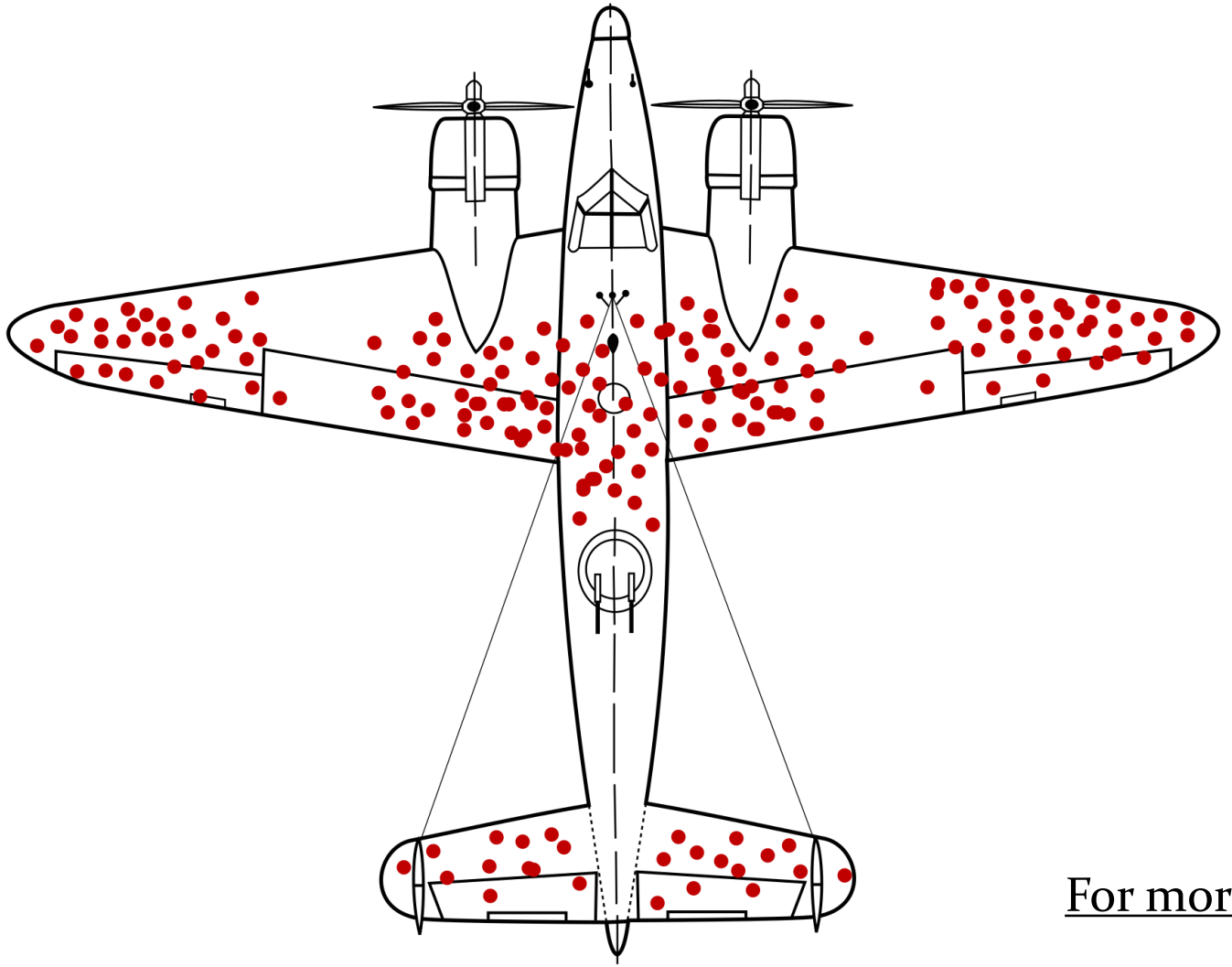
For more: Casselman 2016

What do we need to *know* about the data?



For more: Casselman 2016

Case of selection bias & survivorship bias



For more: Casselman 2016

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

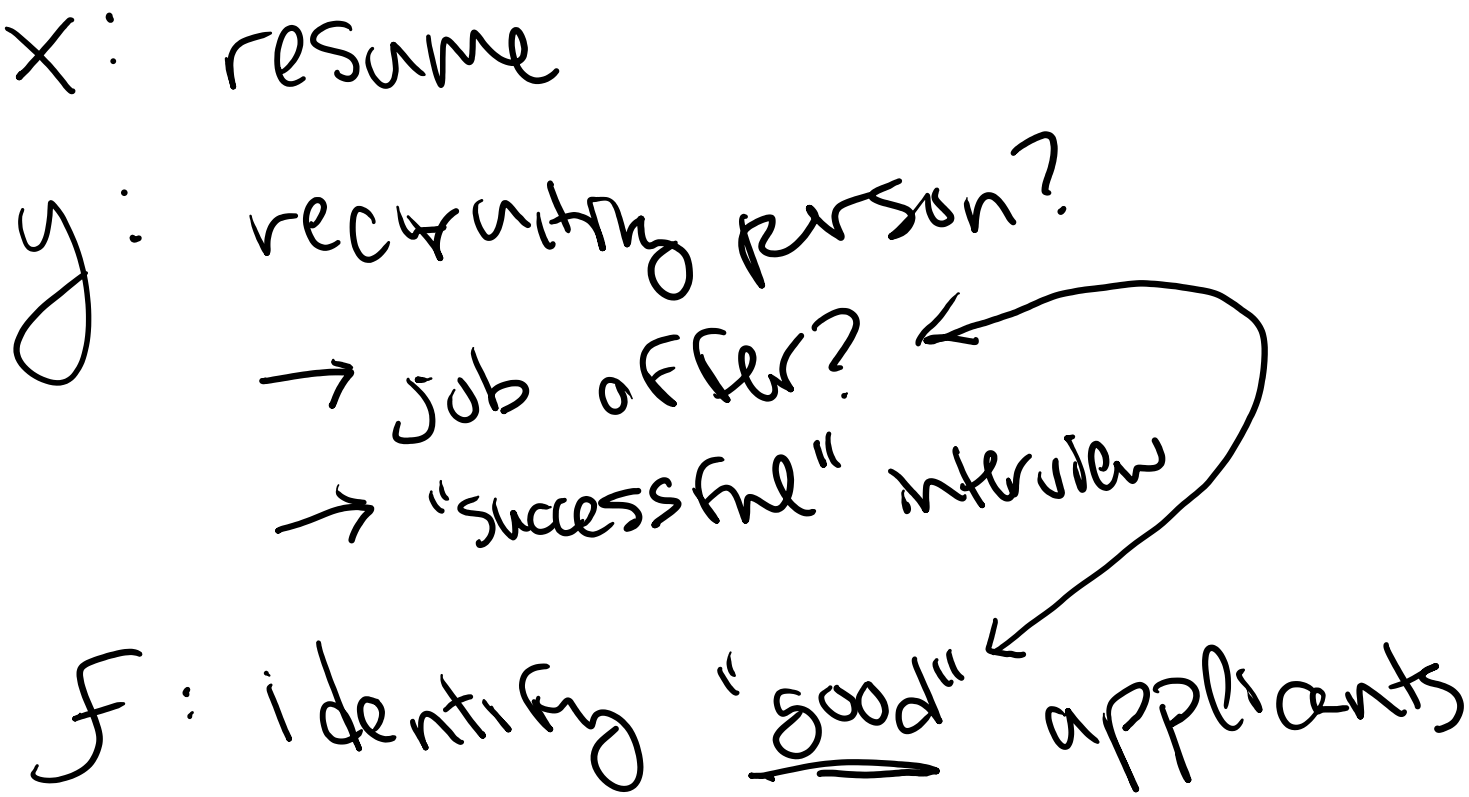
SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Models were trained on 20-years worth of applicant resumes

Penalized resumes:

- containing `"women's"`
- containing `"women's chess captain"`
- of graduates from certain all-women's colleges

Q: What's the problem and task?



The Unknown Perils of Mining Wikipedia

Dr. Benjamin Wilson

If a machine is to learn about humans from Wikipedia, it must experience the corpus as a human sees it and ignore the overwhelming mass of robot-generated pages that no human ever reads. We provide a cleaned corpus (also a Wikipedia recommendation API derived from it).

Problem with “unpopular” pages:

- Irregular coverage of topics
- Receive little to no traffic
- Many are auto-generated

Q: What's the problem and task?

Train general language model

English

Assumption: Language Quality

Exercise / In-Class Activity

Facebook sent flawed data to misinformation researchers.

The data included the interactions of only about half of Facebook's U.S. users, not all of them as the company had said.

nytimes.com

Only included ~**50%** of U.S. users, ***not*** all U.S. users as promised

Only included U.S. users "who engaged with political pages enough to make their political leanings clear"

Q: Is data always objective?

No!

Data is a set of observations. These can be flawed. They can be incomplete, inconsistent, inaccurate, noisy, biased, selective, too general, opinionated, and more.

Issues are not limited to our interpretation and analysis of data but can stem from the data itself.

Q: Is more data always better?

No! More data, more problems.

C4: Colossal Clean Crawled Corpus

- Common Crawl is an archive (TBs) of “web extracted text”
- C4 is a curated version of the Common Crawl

6.1 TB → 745 GB

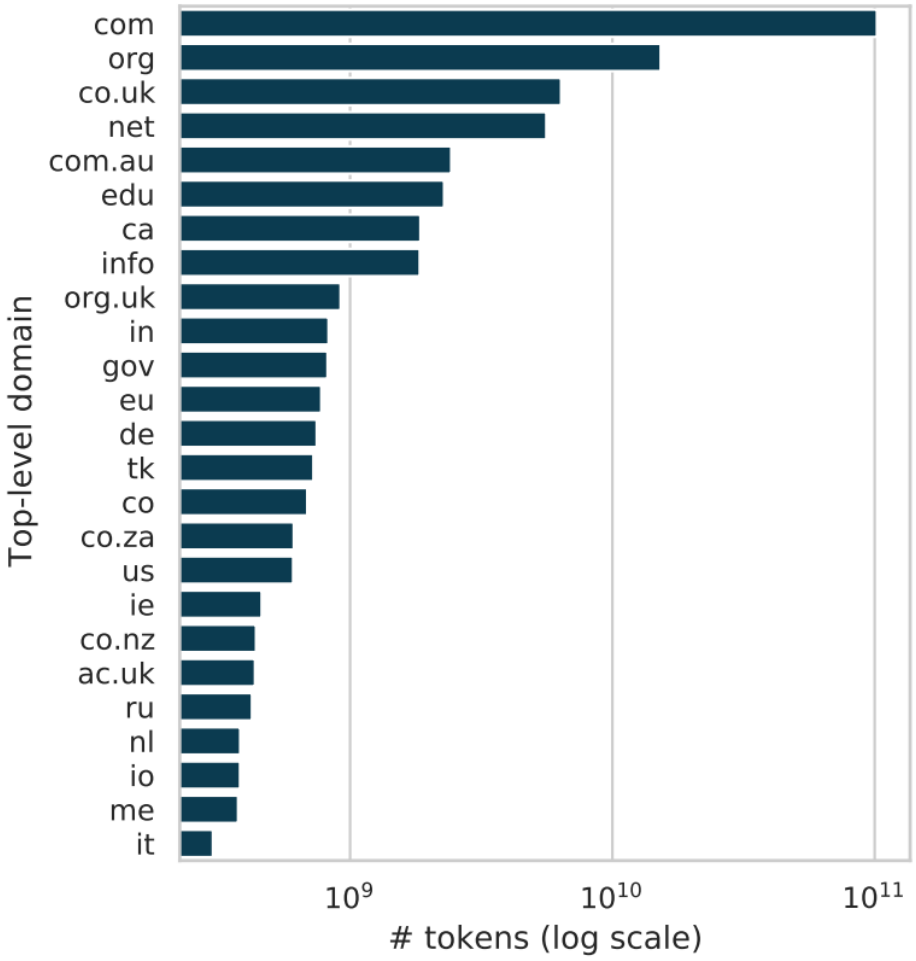
Raffel et al. 2020

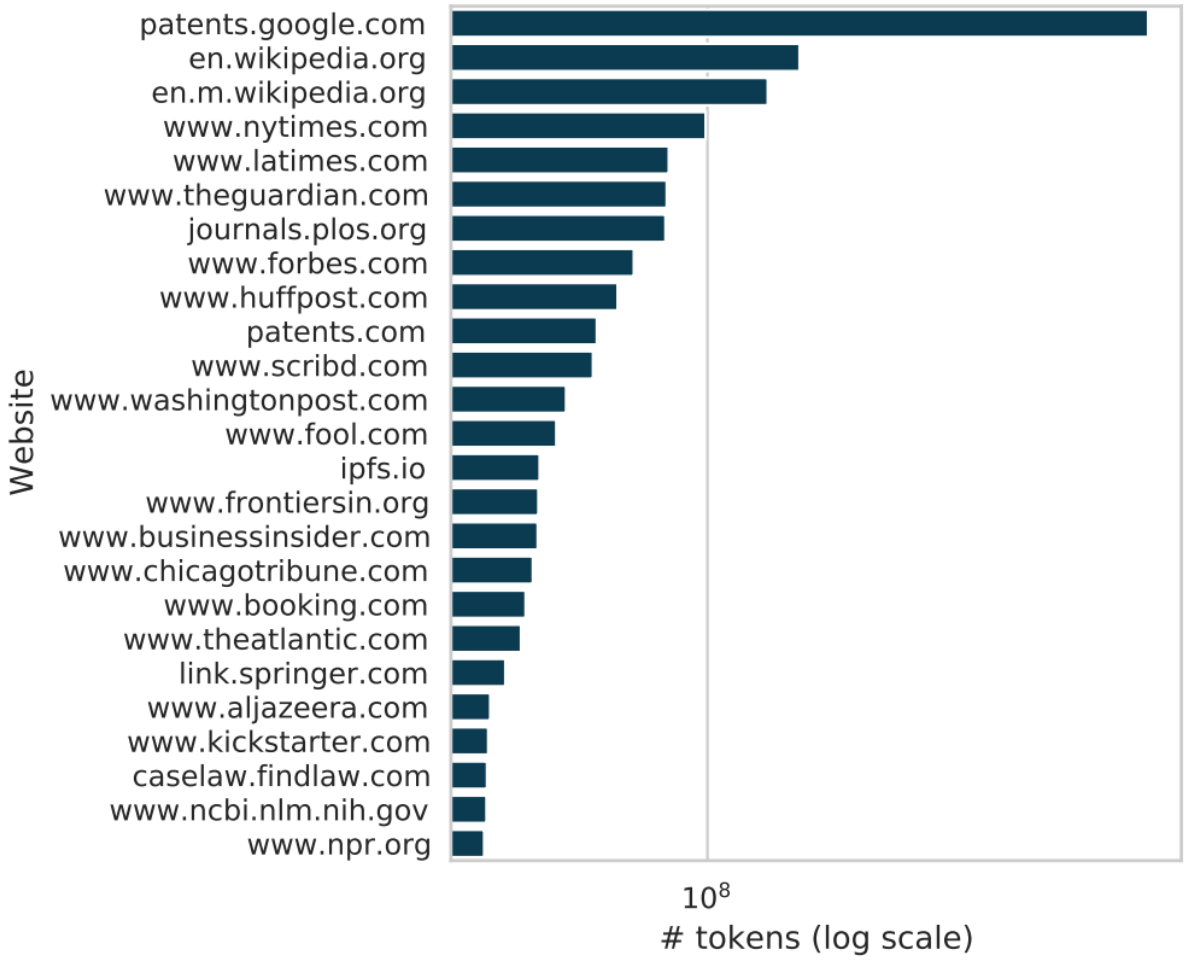
Curation decisions

- Only keep lines ending with a terminal punctuation mark
- Remove pages with < 5 sentences, lines with < 3 words
- Remove pages containing terms in the "List of Dirty, Naughty, Obscene or Otherwise Bad Words"
- Remove pages containing "lorem ipsum"
- Remove pages containing "{"
- Remove all but one of any-three sentence span that occurs multiple times within the data

Raffel et al. 2020

…but what's in it?





Dodge et al. 2021

Issues: Machine Translated Content

Count	Country or WIPO Code	Country or Office Name	Language
70489	US	USA	English
4583	EP	European Patent Office	English, French, or German
4554	JP	Japan	Japanese
2283	CN	China	Chinese (Simplified)
2154	WO	World Intellectual Property Organization	Various
1554	KR	Republic of Korea	Korean
1417	CA	Canada	English
982	AU	Australia	English
747	GB	United Kingdom	English
338	DE	Germany	German
332	TW	Taiwan	Traditional Chinese
271	FR	France	French
138	MX	Mexico	Spanish
118	SE	Sweden	Swedish
711	Other	Various	Various

Dodge et al. 2021

Issues: Whose English is included?

	% in C4	% removed by banned words list
African American English	0.07%	**42%**
Hispanic-Aligned English	0.09%	32%
White-Aligned English	**97.8%**	6.2%

Dodge et al. 2021

Cat Rank

The grass is always greener

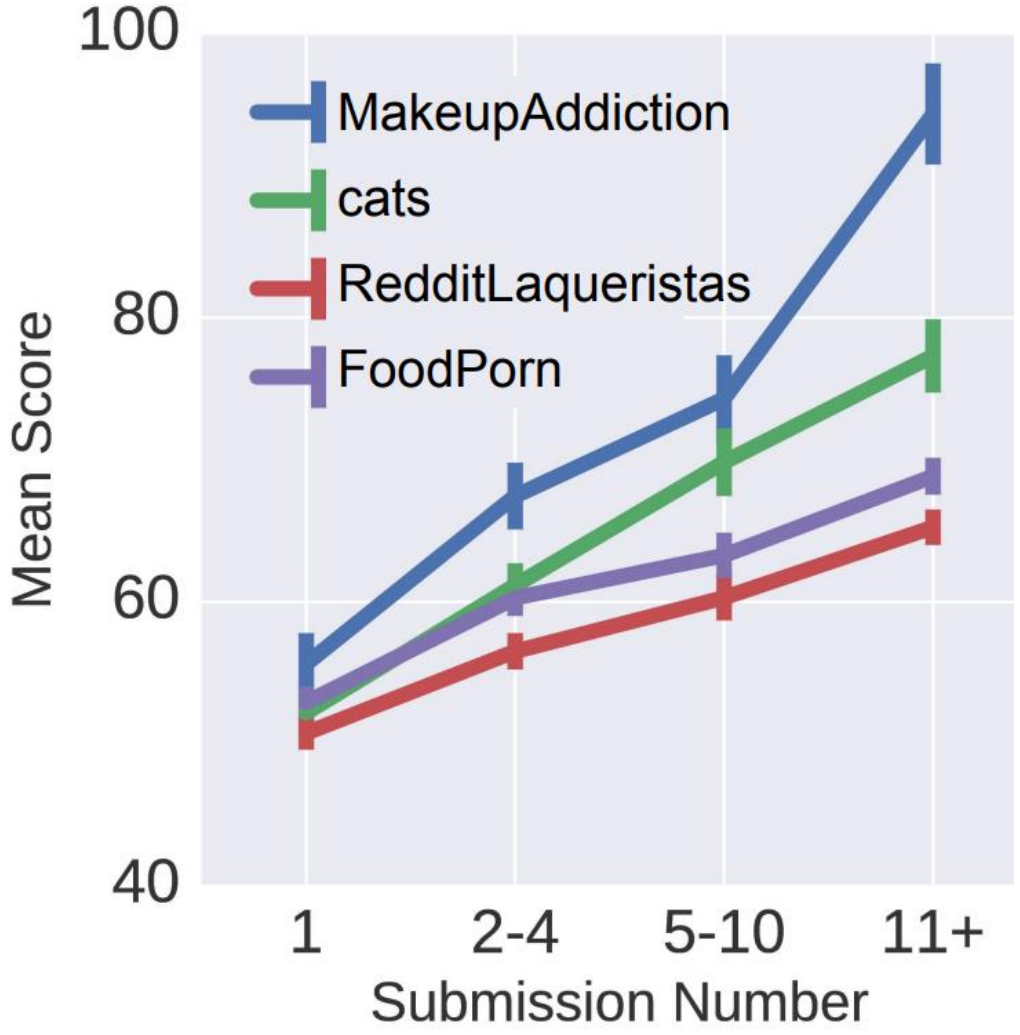
This is why you get two cats

Hessel et al. 2017

Cat Rank

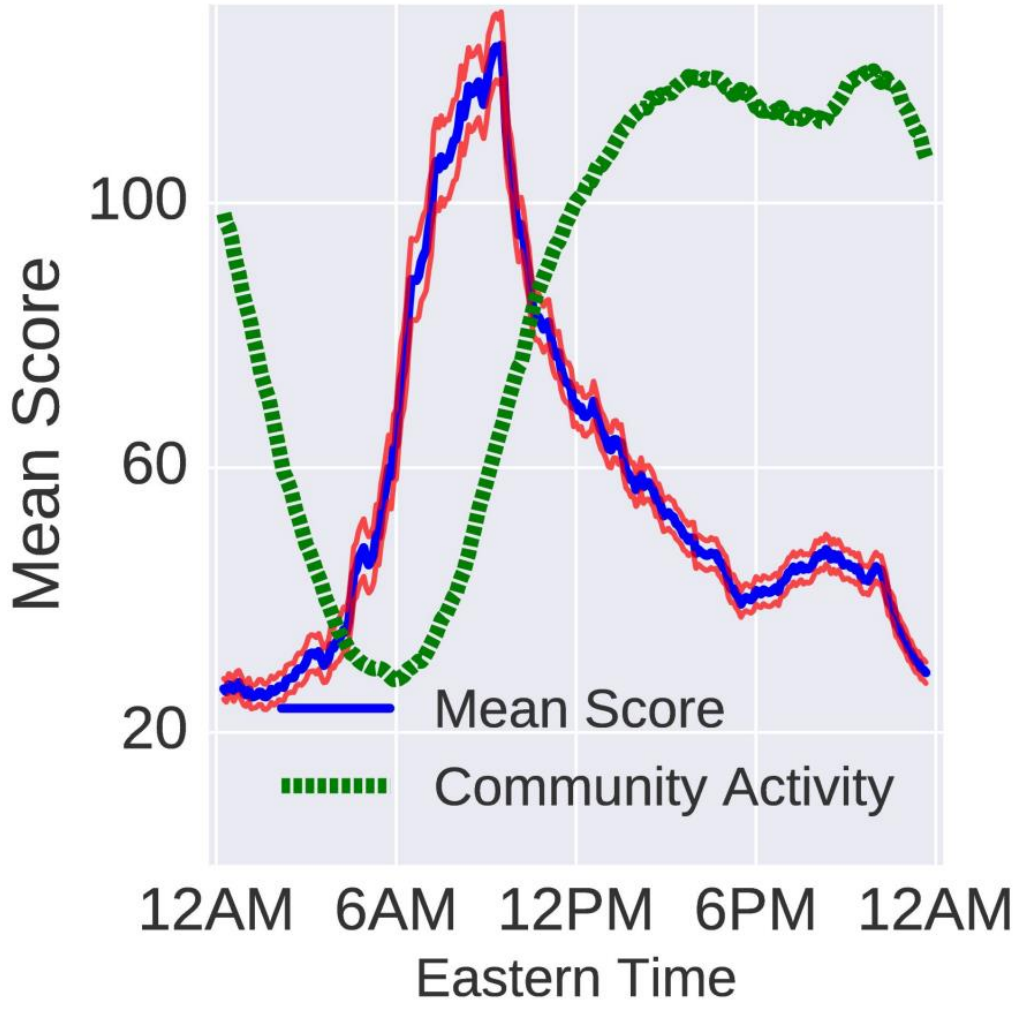
Hessel et al. 2017

Identity matters



Hessel et al. 2017

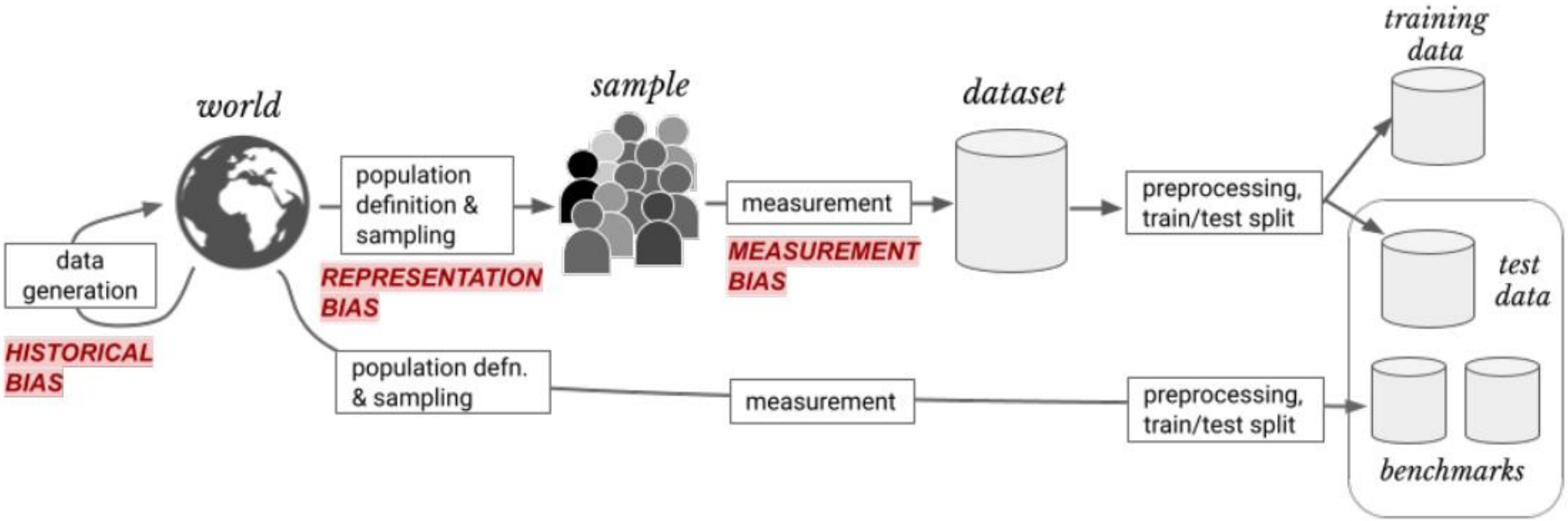
Timing matters



Hessel et al. 2017

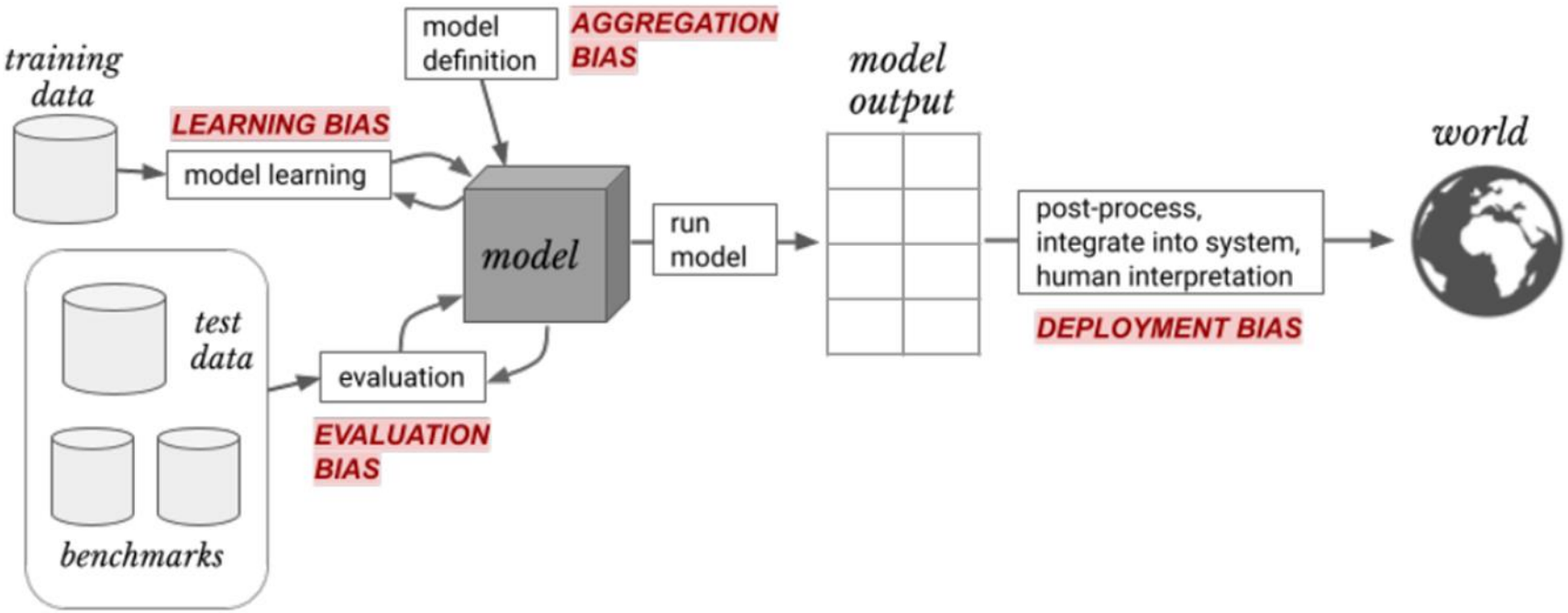
More on bias

Where issues can arise



Source: Suresh & Guttag (2020)

Not just a data issue!



Source: Suresh & Guttag (2020)