THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Building Resilient Data Pipelines

A/Prof. Gianluca Demartini
Data Science Discipline
School of Information Technology and Electrical Engineering

Research Interests

- **Entity-centric Information Access** (since 2005)
  - Structured/Unstruct data (SIGIR12), Types (ISWC13, WSemJ16)
  - Entity Extraction (WWW14), Prepositions (CIKM14), Entity Cards (SIGIR19)
  - IR Eval (IRJ15, ECIR16 Best Paper, CIKM17, SIGIR18, CIKM19, WWW22, TOIS23)
- **Human-in-the-loop Information Systems** (since 2012)
  - Entity Linking (WWW12, VLDBJ), CrowdQ (CIDR 13)
  - Learnersourcing (LAK21, IEEE TLT), HITL (COMNET15, FnT17)
- **Better Crowdsourcing Platforms** (since 2013)
  - Platforms (WWW15, CSCWJ18), Experiments (CSCW21), Pricing (HCOMP14)
  - Task Allocation (WWW13, WWW16, COR), Workers (CHI15), Attacks (HCOMP18 Best Paper, JAIR), Reward (CSCW20 Hon. Mention)
  - Modus Operandi (UBICOMP17, HT19, WSDM20), Bias (SIGIR18, ECIR20 Best Paper)
  - Time (HCOMP16), Complexity (HCOMP16), Abandonment (WSDM19, TKDE)
- **Better Data** (since 2019)
  - Know. Graphs (ISWC19), Noise (WWW19), Metadata (IPM), SES (WebSci22)
  - Unknown Unknowns (ECAI20, HCOMP21), Behaviors (CIKM20)
  - Data Workers (SIGIR20, TOIS, TKDE, WWW23), Fairness (CIKM22, SIGIR23)
- **Data and AI for Public Good** (since 2020)
  - Conservation (w/ Google); Gender (w/ Wiki); Environment (ECIR21, ADCS21)
  - Fake News (w/ Meta; SIGIR20, CIKM20, IPM); Democracy (ADCS21)

Thanks to:

Australian Government
Australian Research Council

European Commission

Swiss National Science Foundation

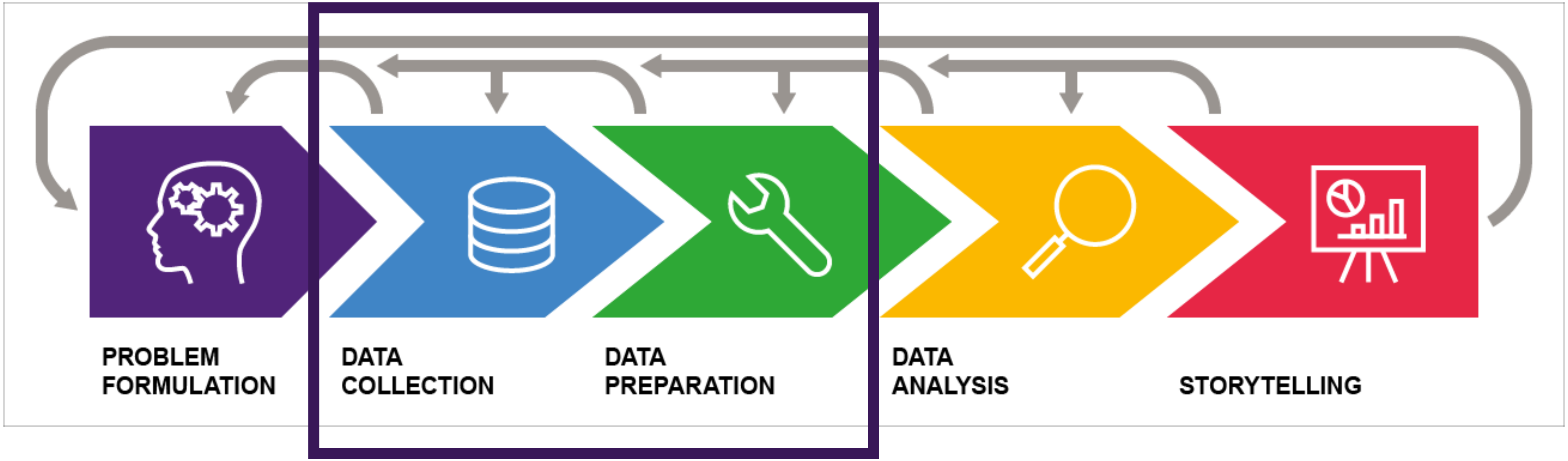
EPSRC
Engineering and Physical Sciences Research Council

Meta

WIKIMEDIA FOUNDATION

Google AI

The Data Science Process



The Impact of Data – What's at stake

Robust data pipelines: Enhance what humans can do with data

The data we are collecting, and the data we are not collecting

Medical doctors prescribing certain diagnostic lab tests (and not others)

The risk of uncollected data

Biased data collection because of expert decision to focus on certain aspects (and not others)

Shazia Sadiq, Amir Aryani, Gianluca Demartini, Wen Hua, Marta Indulska, Andrew Burton Jones, Hassan Khosravi, Diana Benavides Prado, Timos Sellis, Ida Asadi Someh, Rhema Vaithianathan, Sen Wang, and Xiaofang Zhou. Information Resilience: The Nexus of Responsible and Agile Approaches to Information Use. In: The International Journal on Very Large Data Bases (**VLDBJ**), Springer. 2022.

arc training centre for **information resilience**

Outline

Data Collection

- Participation bias: Wikidata editors and knowledge graphs (CSCWJ + ISWC 2019)
- Unknown unknowns (HCOMP 2021)

Data Preparation and Data Quality

- Data curation behaviors (SIGIR 2020 + TOIS, TKDE)
- Behavior embeddings (CIKM 2020)

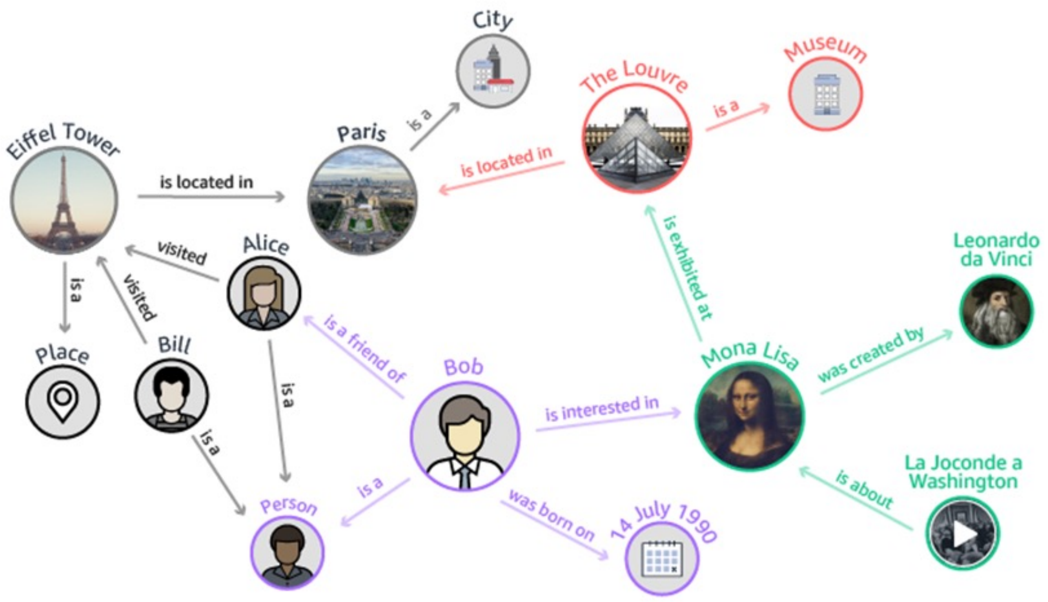
Data Labelling

- Political bias: misinformation annotations (ECIR 2020, SIGIR 2020)
- Cultural bias: socio-economic diversity of annotations (WebSci 2022)
- Bandwagon effect influencing human annotations (IP&M)

The impact of bias on ML models (CIKM 2022)

KGs and data quality

Knowledge graphs store information about entities and their relations



Data quality:

- Missing/wrong entities
- Missing/wrong information about entities
- Missing/wrong relations between entities

Wikidata - A Collaborative Knowledge Graph

KG for humans and machines

Started in 2012

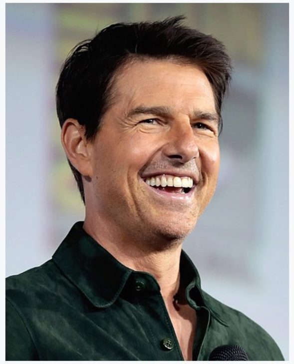
Source of structured open data

Collaboratively edited

Used to power Wikipedia infoboxes

WIKIDATA

Tom Cruise



Cruise at the 2019 San Diego Comic-Con

Born	Thomas Cruise Mapother IV July 3, 1962 (age 59) Syracuse, New York, U.S.
Occupation	Actor · producer
Years active	1981–present
Works	Full list
Spouse(s)	Mimi Rogers (m. 1987; div. 1990) Nicole Kidman (m. 1990; div. 2001) Katie Holmes (m. 2006; div. 2012)
Children	3
Relatives	William Mapother (cousin)
Awards	Full list
Website	tomcruise.com

Signature

Wikidata Editors

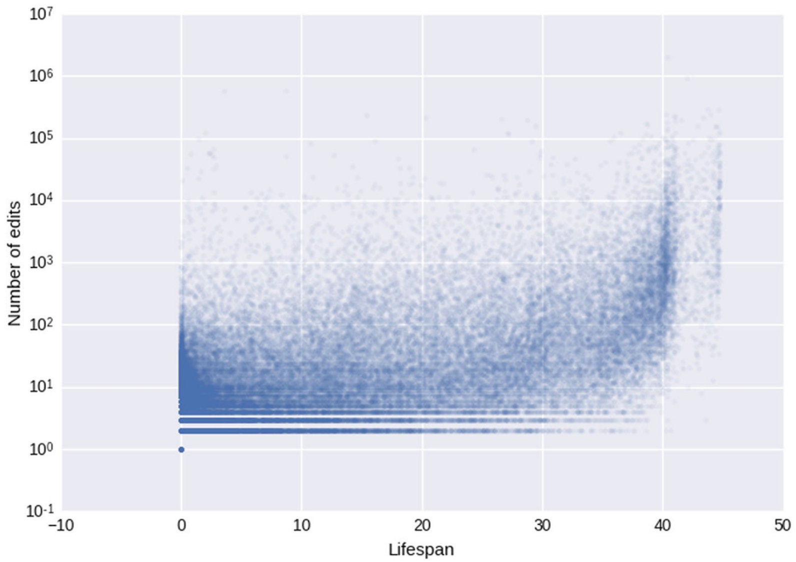
Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel Difallah, Michael Feldman, and Lydia Pintscher. **The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits**. In: Computer Supported Cooperative Work (CSCW) Special Issue on Crowd Dynamics: Conflicts, Contradictions, and Cooperation Issues in Crowdsourcing, Springer, 2018.

The Wikidata edit history (2012-2016)

- 35M (human) edits, 8M items, 140K editors

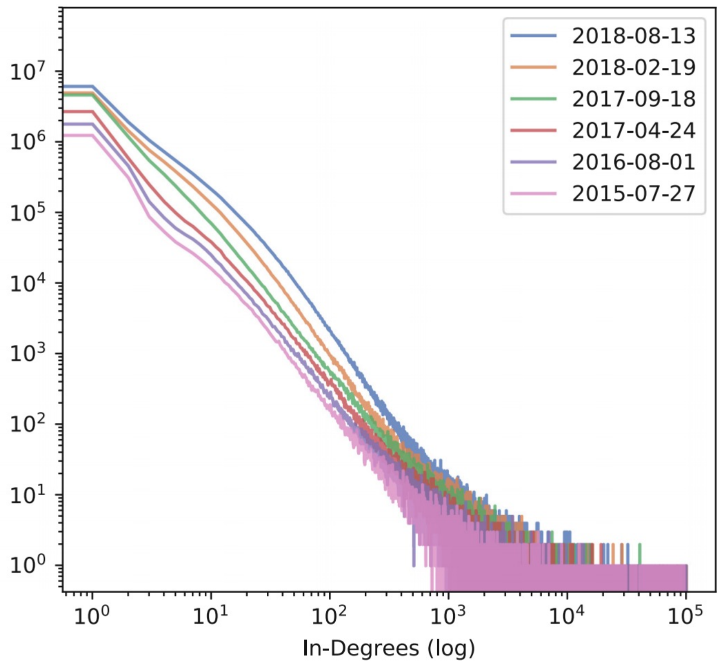
Why do certain editors have a lifetime longer than others?

- **It's a habit**: Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not
- **It's not boring**: Editors with a long lifespan increase the diversity of the type of their edits



Bias: longer lifespan editors contribute more and thus their views and focus dominate the KG data

Knowledge Graph - Completeness



Estimating Class Completeness

- Do we have all the cities of Germany in the KG?

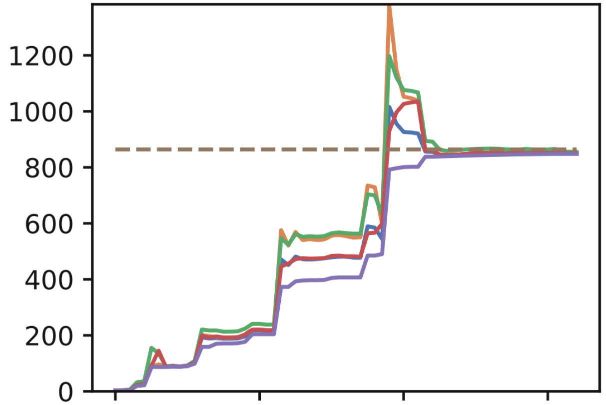
Need to know class cardinality

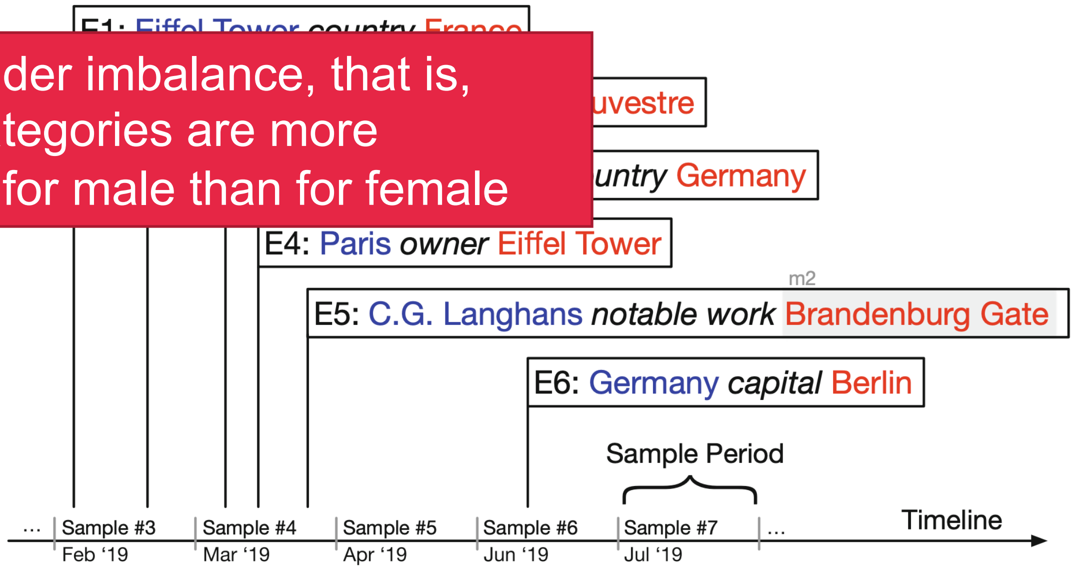
- Easy for US States, difficult for others (need to estimate)

Estimation based on capture/recapture

- Need sampling/mentions over time

(h) Paintings by Vincent van Gogh





Bias: gender imbalance, that is, certain categories are more complete for male than for female

Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs. In: The **International Semantic Web Conference** (ISWC 2019 - Research Track).

Unknown Unknowns

Bias leads to fairness issues!

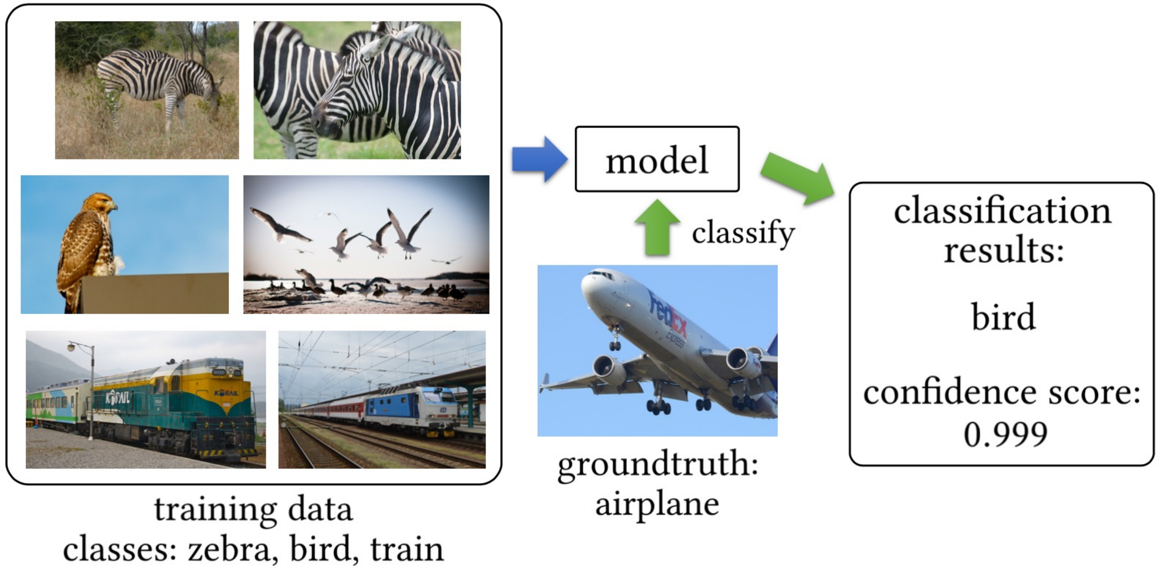


Figure 1: An example of UUs where the model makes a *wrong* classification but with a high confidence score. In this case, the classification model is not able to identify such mistakes automatically.

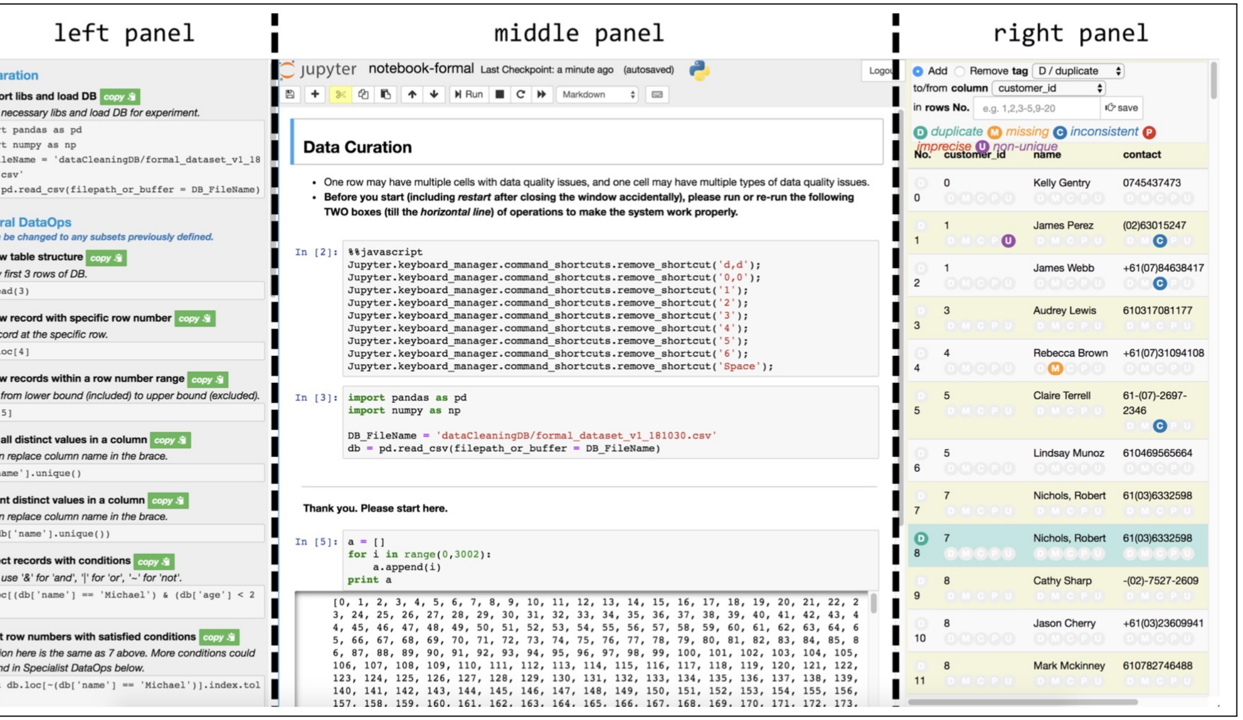
A Human-in-the-loop approach to UU detection:

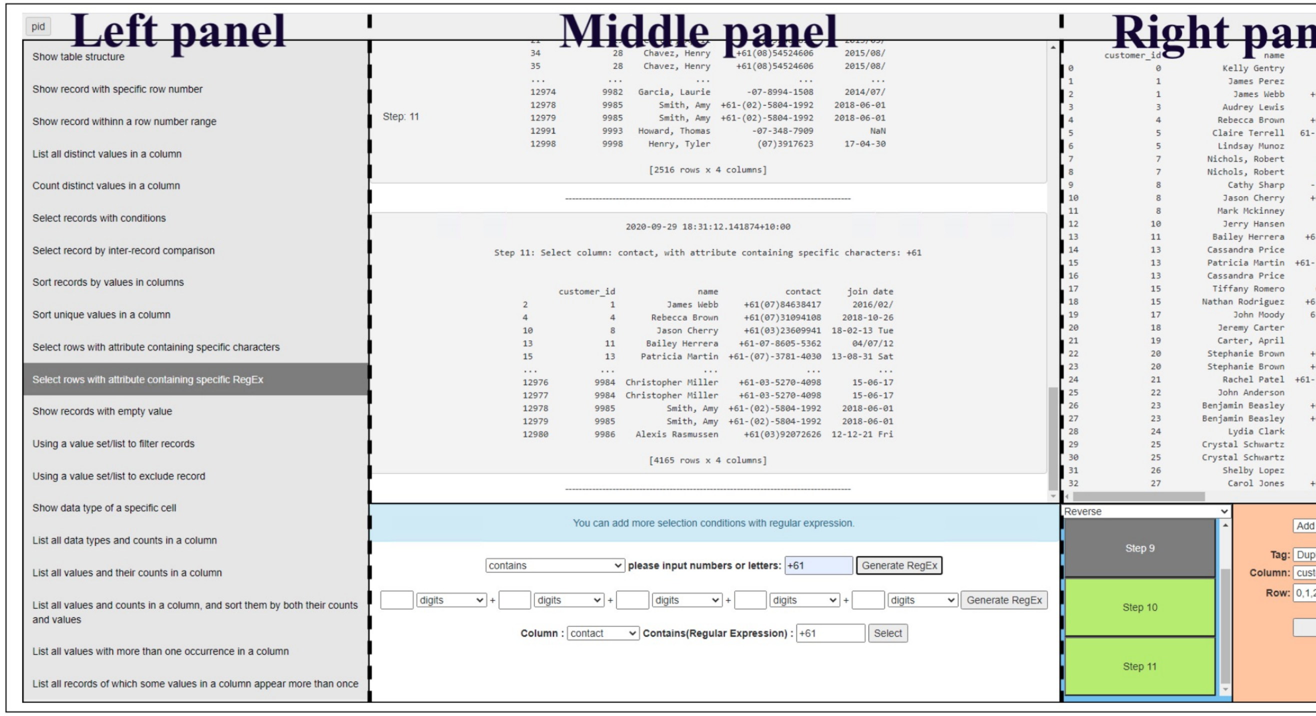
- Extract features from human generated text
- Compare human-generated features with the features learned by the model from training images
- Identify images represented differently in the two spaces (likely to be UUs)
- Collect labels, re-train, and iterate

Lei Han, Xiao Dong, and Gianluca Demartini. Iterative Human-in-the-Loop Discovery of Unknown Unknowns in Image Datasets. In: Proceedings of the 9th AAAI Conference on Human Computation and Crowdsourcing (**HCOMP** 2021). November 2021.

Bias in Data Preparation and Data Quality

Data Curation Behaviors





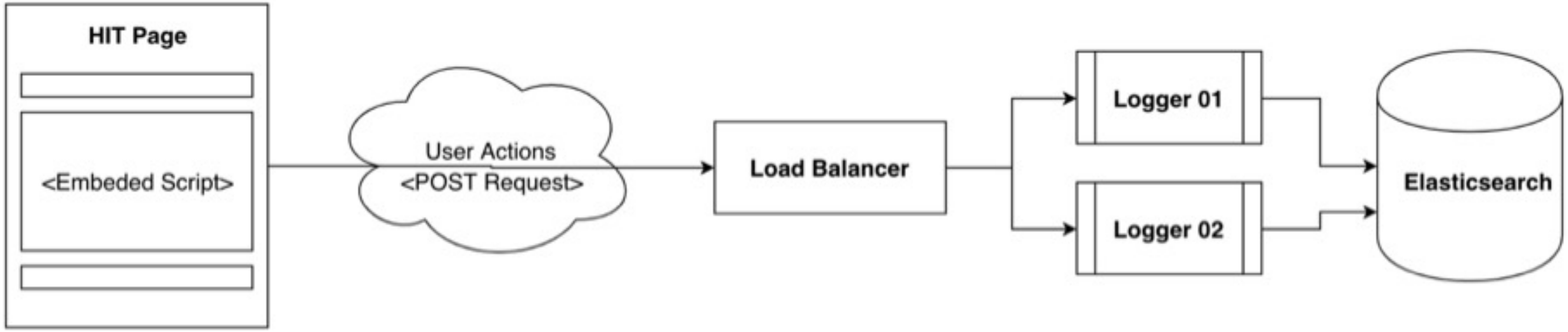
Lei Han, Tianwa Chen, Gianluca Demartini, Marta Indulska, and Shazia Sadiq. A Data-Driven Analysis of Behaviors in Data Curation Processes. In: ACM Transactions on Information Systems (**TOIS**). 2022.

Shaochen Yu, Tianwa Chen, Lei Han, Gianluca Demartini, and Shazia Sadiq. DataOps-4G: On Supporting Generalists in Data Quality Discovery. In: IEEE Transactions on Knowledge and Data Engineering (**TKDE**). 2022.

Logging Behaviors

UQCrowd Logging System

- JS code embedded in the data annotation tasks
- Send msg (for every click, keystroke, scroll, new tab, etc.) to our server



Observe human annotator online behaviors while they complete tasks

https://github.com/d-lab/uqcrowd-log

Behavior embeddings

Model human annotator behavior using embeddings

- Raw actions from logs as sequences of tokens + CBOW
- Vector representations of user behaviors

Compare user behaviors (e.g., high performers / low performers)

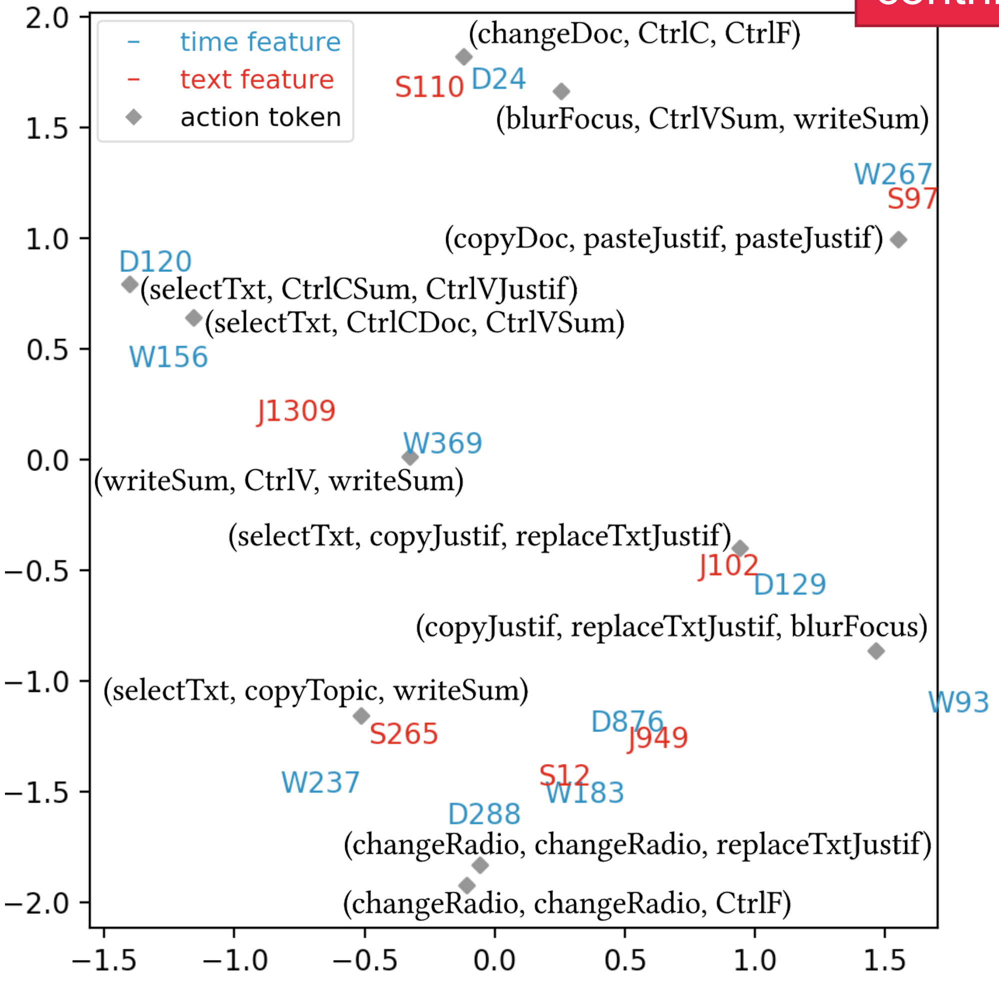
Changes over time

Different time granularities

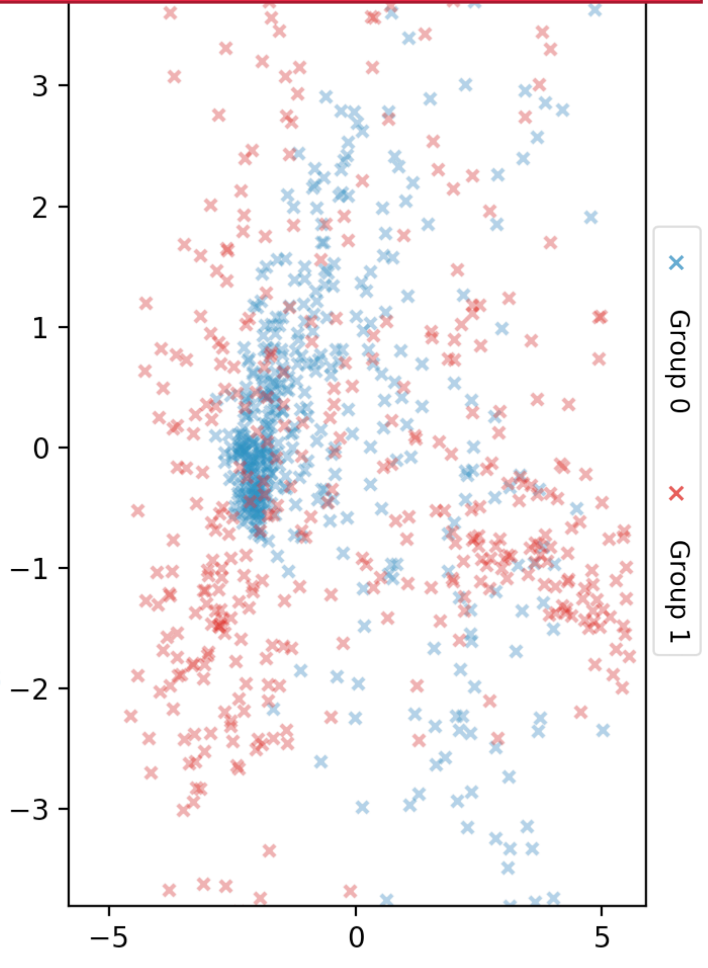
Order	**Single Action**	***n*-gram Token** ($n = 2$)
1	`Ctrl+C`	(`Ctrl+C`, `Ctrl+V`)
2	`Ctrl+V`	(`Ctrl+V`, type characters)
3	type characters	(type characters, delete characters)
4	delete characters	(delete characters, click ‘next’)
5	click ‘next’	—

Lei Han, Alessandro Checco, Djellel E. Difallah, Gianluca Demartini, and Shazia Sadiq. Modelling User Behavior Dynamics with Embeddings. In: 29th ACM International Conference on Information and Knowledge Management (**CIKM** 2020).

Bias: different people use different strategies that makes them contribute more/less/differently



Crowdsourcing Task



WikiData: 0 less active; 1 more active

Datasets: https://github.com/tomhanlei/20cikm-behavior

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Bias in Data Labelling

Misinformation annotation and video tagging

Crowdsourcing Truthfulness Judgements

~600 MTurk US workers

To assess truthfulness of

- US political statements (Politifact)
- non-US political statements (ABC)

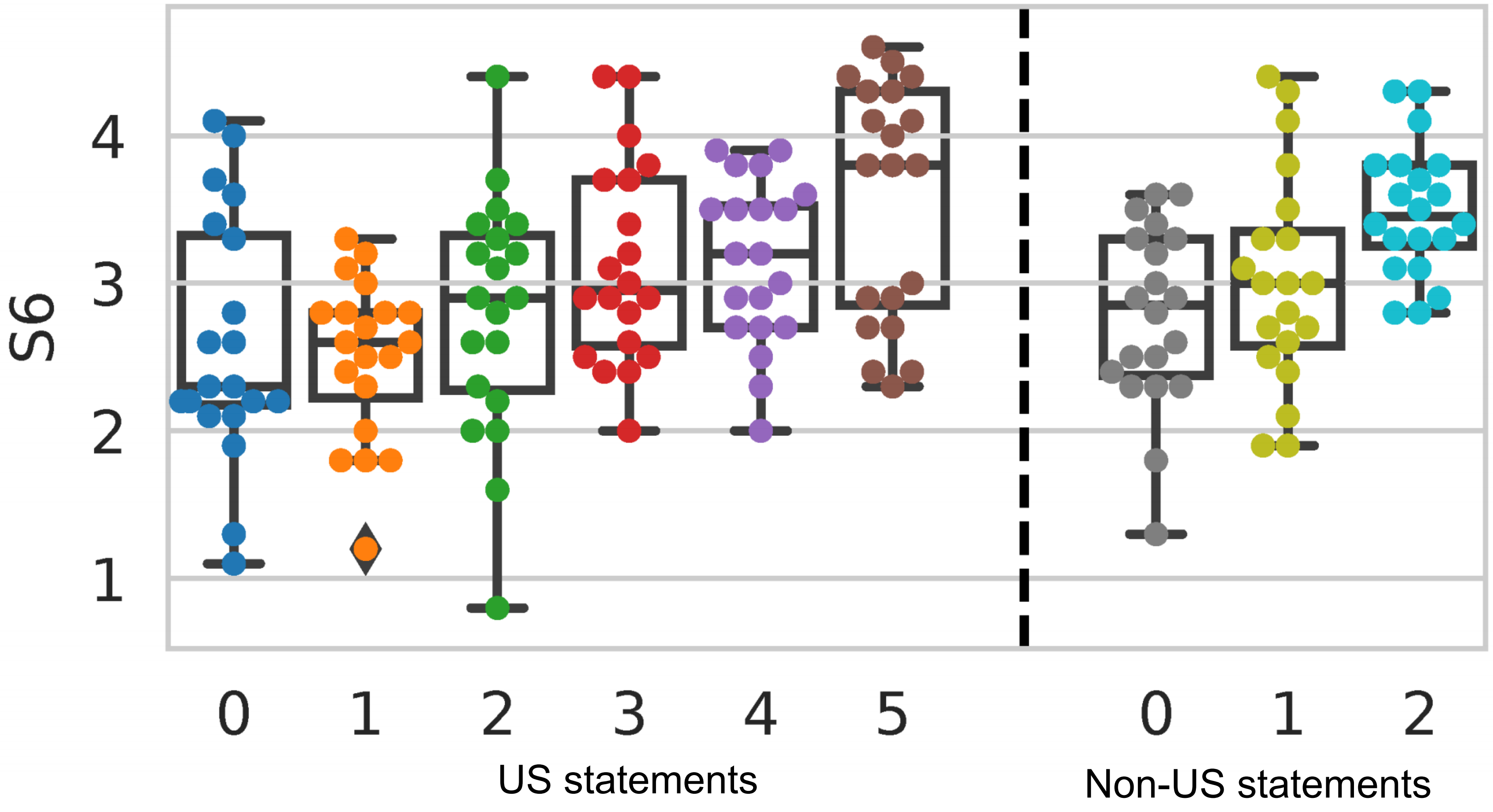
3 scales (3, 6, and 100 levels)

Table 1: Example of statements in the PolitiFact and ABC datasets.

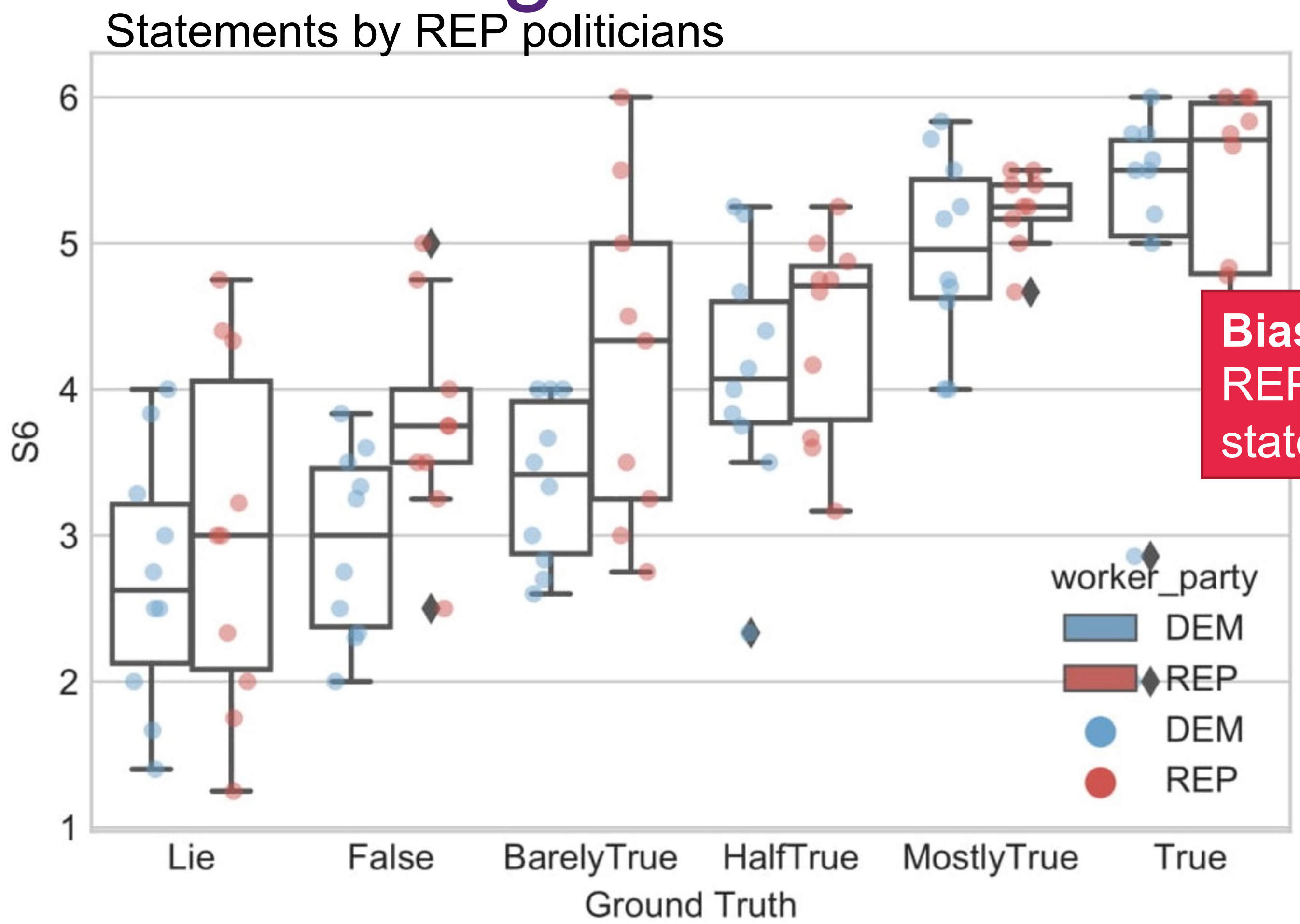
	Statement	Speaker, Year
PolitiFact Label: mostly-true	"Florida ranks first in the nation for access to free prekindergarten."	Rick Scott, 2014
ABC Label: in-between	"Scrapping the carbon tax means every household will be $550 a year better off."	Tony Abbott, 2014

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro and Gianluca Demartini. Can The Crowd Identify Misinformation Objectively? The Effects of Judgments Scale and Assessor's Bias. In: The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR** 2020)

Crowd Performance VS Expert Ground Truth



Fake News labelling - Political bias



Bias: Non-expert people who vote REP are more likely to believe to statements by REP politicians.

David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In: The 42nd European Conference on Information Retrieval (**ECIR** 2020).

Video of people washing hands across different socio-economic statuses

- 4 regions: Africa, Asia, Europe, the Americas; 4 different income level for each region (4*4*7=112)
- Average video duration：13.7 seconds (SD = 9.14 seconds)

Bias in the annotation of SES-diverse content

- **Less accurate** in guessing families' income levels for **African videos**.
- Videos depicting **low-income** households were more likely to receive **negative** annotations
- Videos with **higher-income** families received more **positive** annotations.
- **Negative** annotations were more prevalent for videos shot in **Africa** than in **Asia**.
- Video from **higher income** groups **more appropriate** for inclusion in search results and public service announcements

Shaoyang Fan, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Shazia Sadiq, and Gianluca Demartini. Socio-Economic Diversity in Human Annotations. In: The 14th **ACM Web Science Conference** 2022, Barcelona, Spain. June 2022.

Influencing human annotators

Controlling Bias

- Presenting metadata in tasks can significantly improve the efficiency of annotations.
- Human metadata is a popular resource to assess relevance. **Strong bandwagon effect**
- The role of metadata is subject to its quality.

Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. On the role of Human and Machine Metadata in Relevance Judgment Tasks. In: **Information Processing & Management** (IP&M), Elsevier. 2022.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

What happens when we train ML models with biased labels?

Demo at: https://recant.cyens.org.cy/

Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. **How Does the Crowd Impact the Model? A tool for raising awareness of social bias in crowdsourced training data**. In: The 31st ACM International Conference on Information and Knowledge Management (CIKM 2022) - Demo track. Atlanta, Georgia, USA, October 2022.

1. Input image:

Click here to change the image

Current image: CFD-BF-003-003-N

2. Classification task:

Select a classification task.

Gender | Race | Attractiveness | Trustworthiness

The models try to predict the depicted person's Trustworthiness.

3. Results:

Click to show Results.

Execute

Nine different models were trained on the same images for each task, with different (sub)sets of crowd-worker annotations. The same input image (above) was passed through each of the nine models, resulting in the following outputs (possible outputs: Low, Medium, High):

Model	Model Description	Classification Decision
CFD Annotators	Model trained on the norming data provided with the CFD.	High
All Annotators	Model trained using all the annotations for all images.	Medium
Random	Model that simulates the case where annotators generate labels without considering the image content.	Medium
Men	Model trained using all the annotations provided by male crowdworkers.	Low
Women	Model trained using all the annotations provided by female crowdworkers.	Medium
Black	Model trained using all the annotations provided by Black crowdworkers.	Medium
Asian	Model trained using all the annotations provided by Asian crowdworkers.	Low
White	Model trained using all the annotations provided by White crowdworkers.	Medium
Latino	Model trained using all the annotations provided by Latino crowdworkers.	High

gianlucademartini.net
demartini@acm.org
@eglu81

Lessons learned and what to do

- Bias is present in human-generated data is propagated in data pipelines
- Bias comes from human annotators as much as system design choices

- We need to track and profile data bias across the data pipeline
- Select and diversify the sources of the labels (i.e., human annotators)
- **Bias management** instead of bias removal
  - Demartini et al. “Data Bias Management”, in *Communications of the ACM* https://arxiv.org/abs/2305.09686