



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



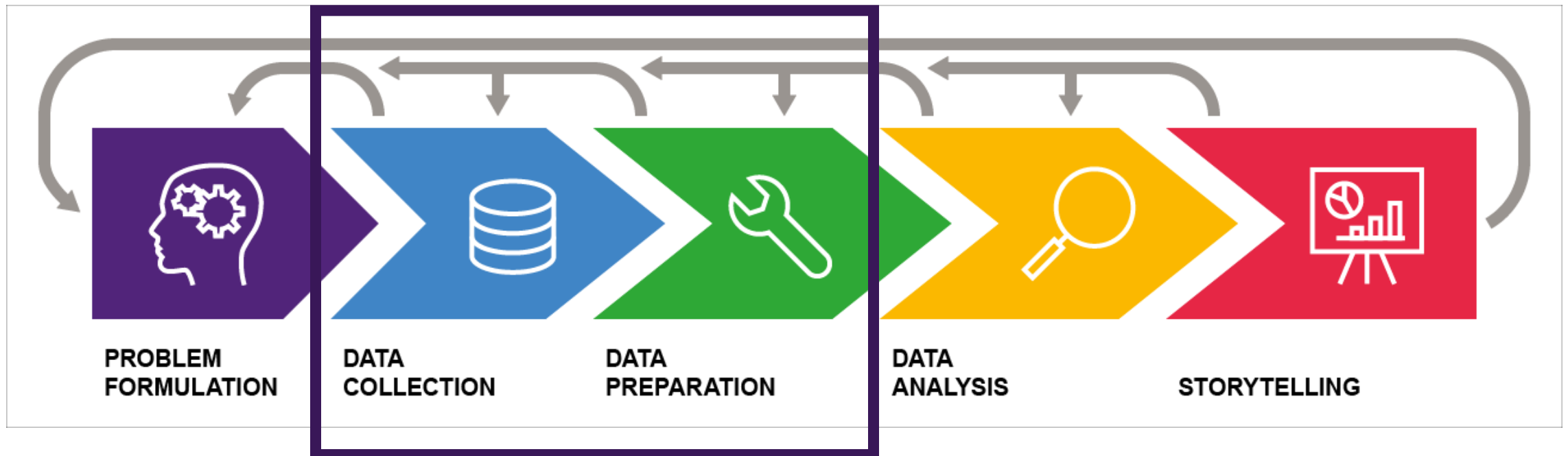
Swiss National
Science Foundation



Engineering and Physical Sciences
Research Council



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.

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)

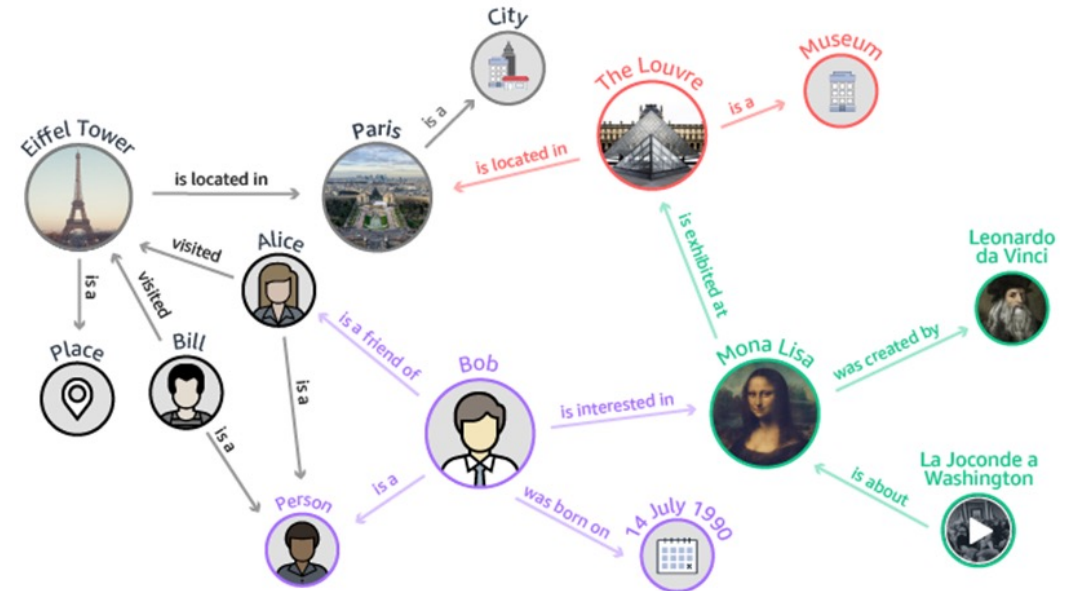
The impact of bias on ML models (CIKM 2022)

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



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

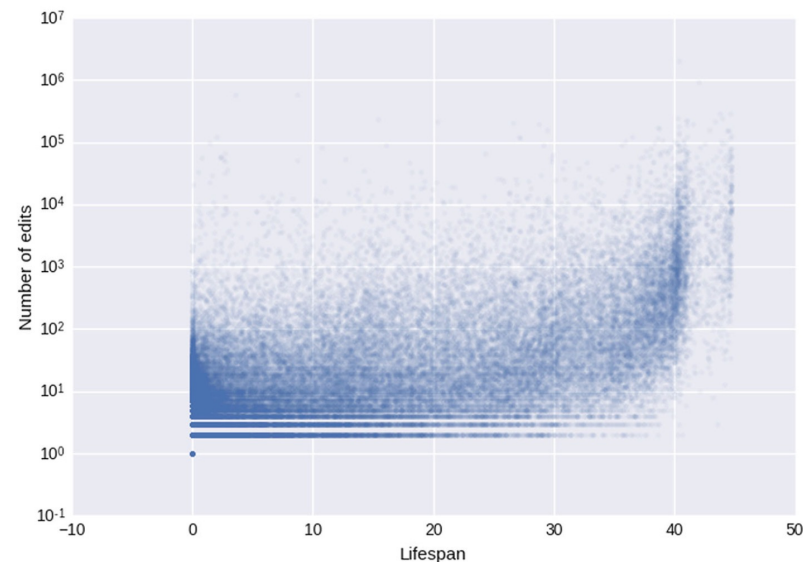
The Wikidata edit history (2012-2016)

- 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

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.



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

Knowledge Graph - Completeness

Estimating Class Completeness

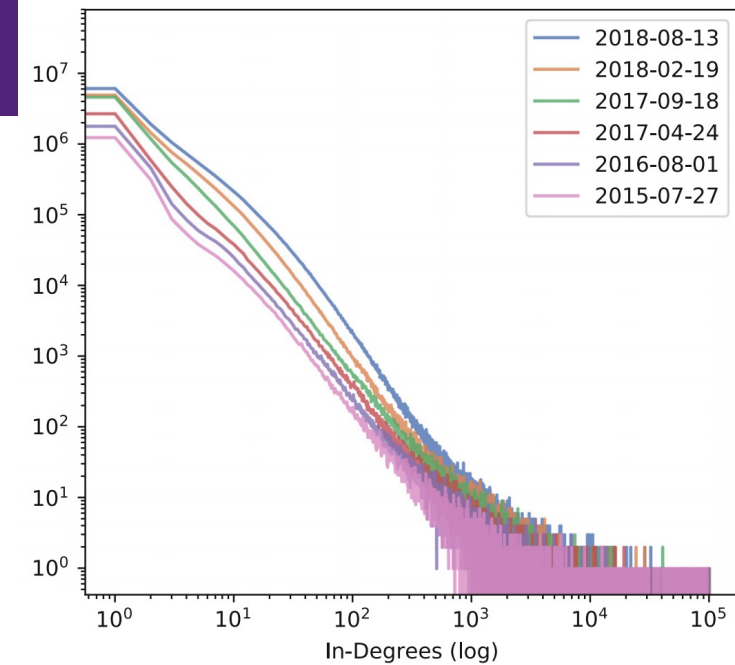
- Do we have all the cities of Germany in the KG?

Need to know class cardinality

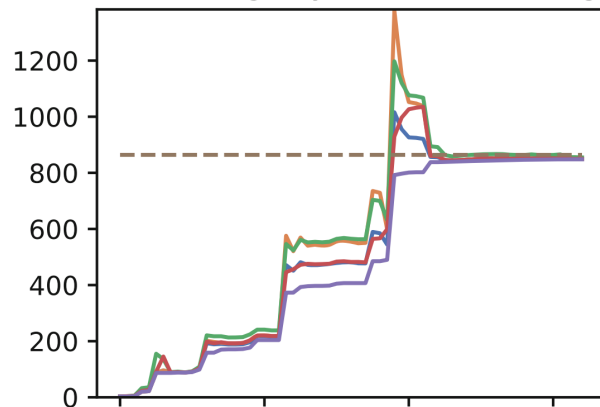
- Easy for US States, difficult for others (need to estimate)

Estimation based on capture/recapture

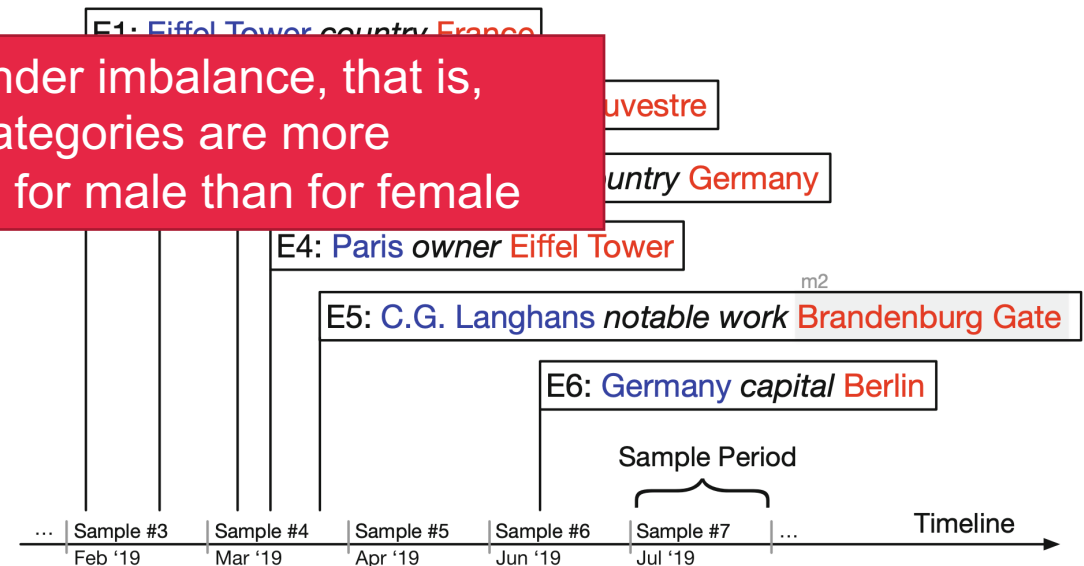
- Need sampling/mentions over time



(h) Paintings by Vincent van Gogh

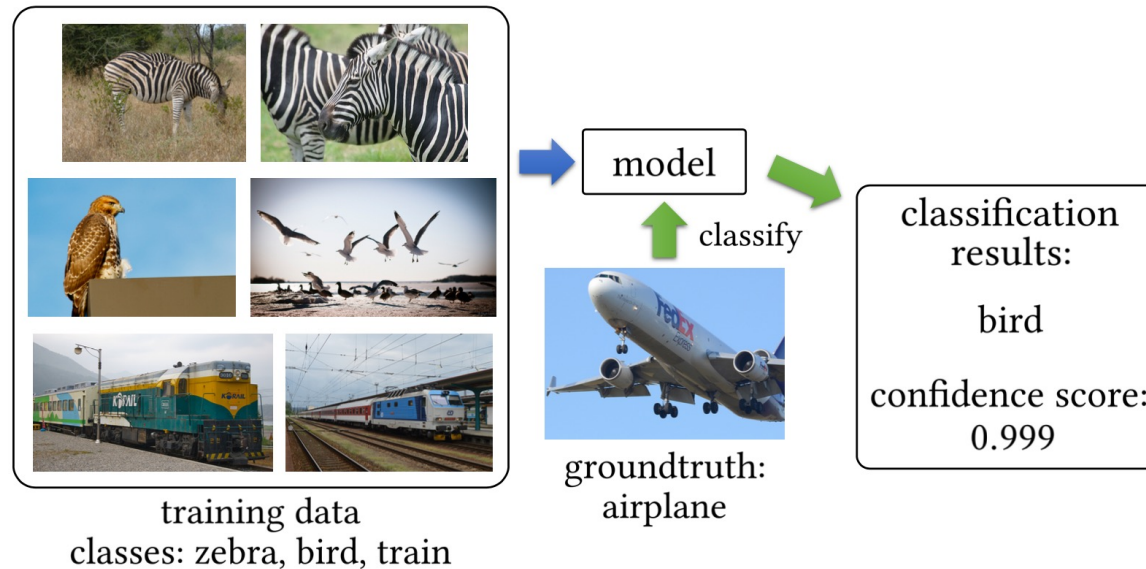


Bias: gender imbalance, that is, certain categories are more complete for male than for female



Unknown Unknowns

Bias leads to fairness issues!



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

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.

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

left panel

necessary libs and load DB

port pandas as pd

import numpy as np

dbName = 'dataCleaningDB/formal_dataset_v1_18'

pd.read_csv(filepath_or_buffer = DB_FileName)

table structure

Show record with specific row number

Show records within a row number range

List all distinct values in a column

Select records with conditions

Count distinct values in a column

Sort records by values in columns

Sort unique values in a column

Select rows with attribute containing specific characters

Select rows with attribute containing specific RegEx

Show records with empty value

Using a value set/list to filter records

Using a value set/list to exclude record

Show data type of a specific cell

List all data types and counts in a column

List all values and their counts in a column

List all values and counts in a column, and sort them by both their counts and values

List all values with more than one occurrence in a column

List all records of which some values in a column appear more than once

middle panel

Jupyter notebook-formal

Last Checkpoint: a minute ago (autosaved)

Add Remove tag Duplicate

to/from column customer_id

in rows No. e.g. 1,2,3-5,9-20

Data Curation

- One row may have multiple cells with data quality issues, and one cell may have multiple types of data quality issues.
- Before you start (including restart after closing the window accidentally), please run or re-run the following TWO boxes (fill the horizontal line) of operations to make the system work properly.

```

In [2]: %%javascript
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('d,d');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('0,0');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('1');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('2');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('3');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('4');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('5');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('6');
Jupyter.keyboard_manager.command_shortcuts.remove_shortcut('Space');

In [3]: import pandas as pd
import numpy as np

DB_FileName = 'dataCleaningDB/formal_dataset_v1_181030.csv'
db = pd.read_csv(filepath_or_buffer = DB_FileName)

Thank you. Please start here.

In [5]: a = []
for i in range(0,3002):
    a.append(i)
print a

```

right panel

customer_id	name	contact
0	Kelly Gentry	0745437473
1	James Perez	(02)63015247
1	James Webb	+61(07)84638417
3	Audrey Lewis	610317081177
4	Rebecca Brown	+61(07)31094108
5	Claire Terrell	61-(07)-2697-2346
6	Lindsay Munoz	610469565664
7	Nichols, Robert	61(03)6332598
7	Nichols, Robert	61(03)6332598
8	Cathy Sharp	-(02)-7527-2609
8	Jason Cherry	+61(03)23609941
8	Mark McKinney	610782746488

Left panel

Show table structure

Show record with specific row number

Show record within a row number range

List all distinct values in a column

Count distinct values in a column

Select records with conditions

Select record by inter-record comparison

Sort records by values in columns

Sort unique values in a column

Select rows with attribute containing specific characters

Select rows with attribute containing specific RegEx

Show records with empty value

Using a value set/list to filter records

Using a value set/list to exclude record

Show data type of a specific cell

List all data types and counts in a column

List all values and their counts in a column

List all values and counts in a column, and sort them by both their counts and values

List all values with more than one occurrence in a column

List all records of which some values in a column appear more than once

Middle panel

Step 11

2020-09-29 18:31:12.141874+10:00

Step 11: Select column: contact, with attribute containing specific characters: +61

customer_id	name	contact	join date
2	1	James Webb	+61(07)84638417 2016/02/
4	4	Rebecca Brown	+61(07)31094108 2018-10-26
10	8	Jason Cherry	+61(03)23609941 18-02-13 Tue
13	11	Bailey Herrera	+61-07-8665-5362 04/07/12
15	13	Patricia Martin	+61-(07)-3761-4030 13-06-31 Sat
...
12976	9984	Christopher Miller	+61-03-5270-4098 15-06-17
12977	9984	Christopher Miller	+61-03-5270-4098 15-06-17
12978	9985	Smith, Amy	+61-(02)-5804-1992 2018-06-01
12979	9985	Smith, Amy	+61-(02)-5804-1992 2018-06-01
12998	9986	Alexis Rasmussen	+61(03)92072626 12-12-21 Fri

[4165 rows x 4 columns]

You can add more selection conditions with regular expression.

contains please input numbers or letters: +61 Generate RegEx

[digits] + [digits] + [digits] + [digits] + [digits] Generate RegEx

Column: contact Contains(Regular Expression): +61 Select

Right panel

customer_id	name
0	Kelly Gentry
1	James Perez
1	James Webb
2	Audrey Lewis
3	Rebecca Brown
4	Claire Terrell
5	Lindsay Munoz
6	Nichols, Robert
7	Nichols, Robert
8	Cathy Sharp
8	Jason Cherry
8	Mark McKinney

Reverse

Step 9

Step 10

Step 11

Tag: Dup

Column: Cust

Row: 0.12

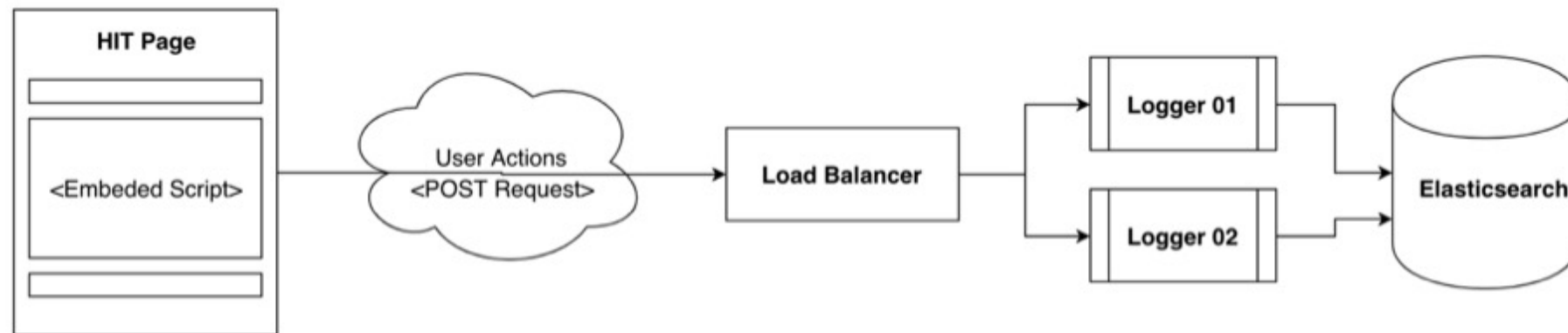
Lei Han, Tianwa Chen, Gianluca Demartini, Marta Indulska, and Shazia Sadiq. A Data-Driven Analysis of Behaviors in Data Curation Processes. In: ACM Transactions on Information Systems (**TOIS**). 2022.

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

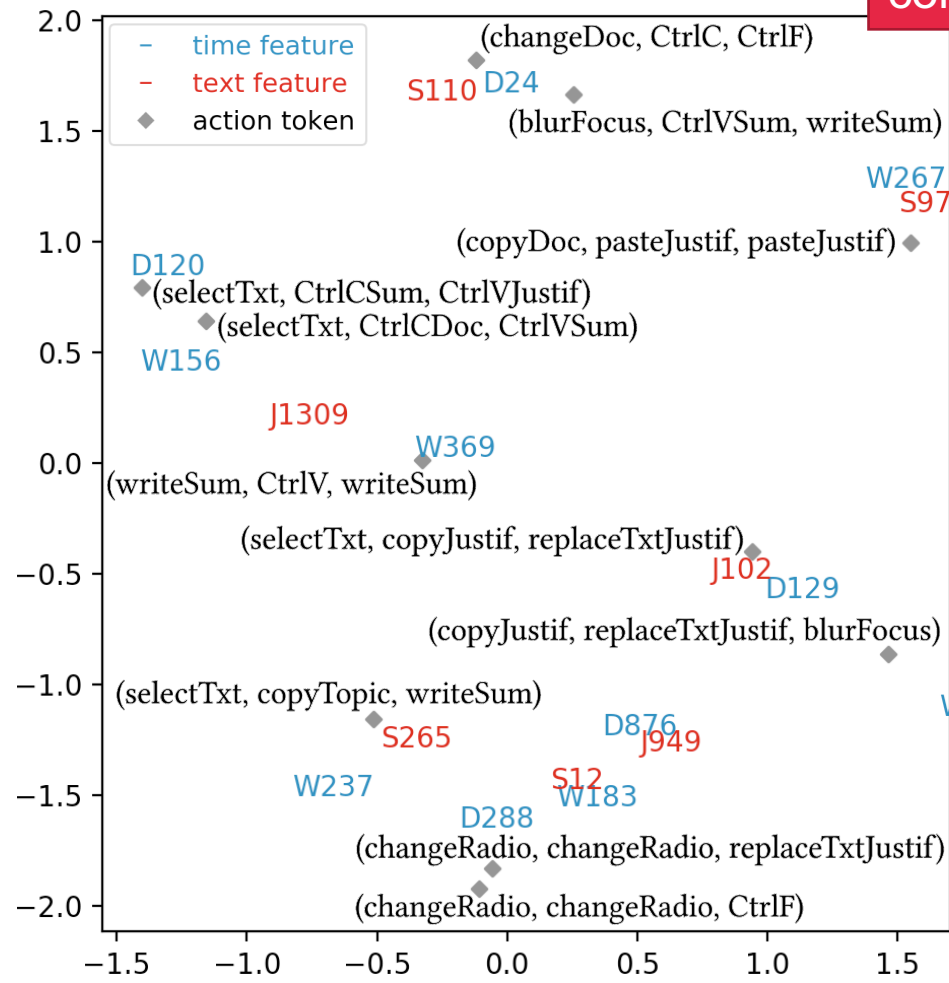
Compare user behaviors (e.g., high performers / low performers)

Changes over time

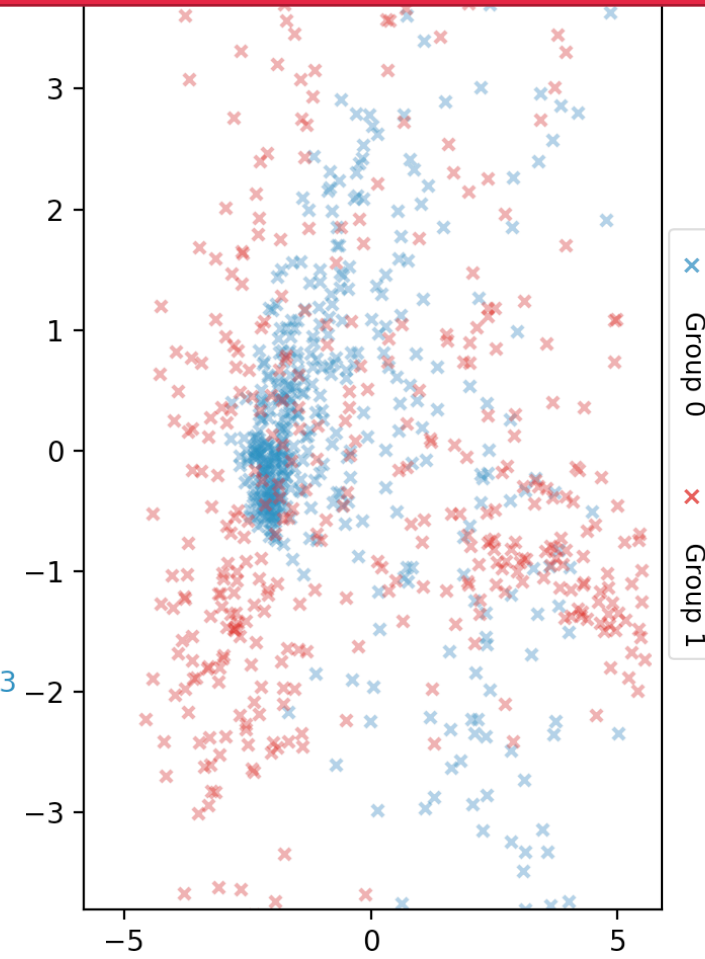
Different time granularities

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'	—

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>



Bias in Data Labelling

Misinformation annotation and video tagging

Crowdsourcing Truthfulness Judgements

~600 MTurk US workers

To assess truthfulness of

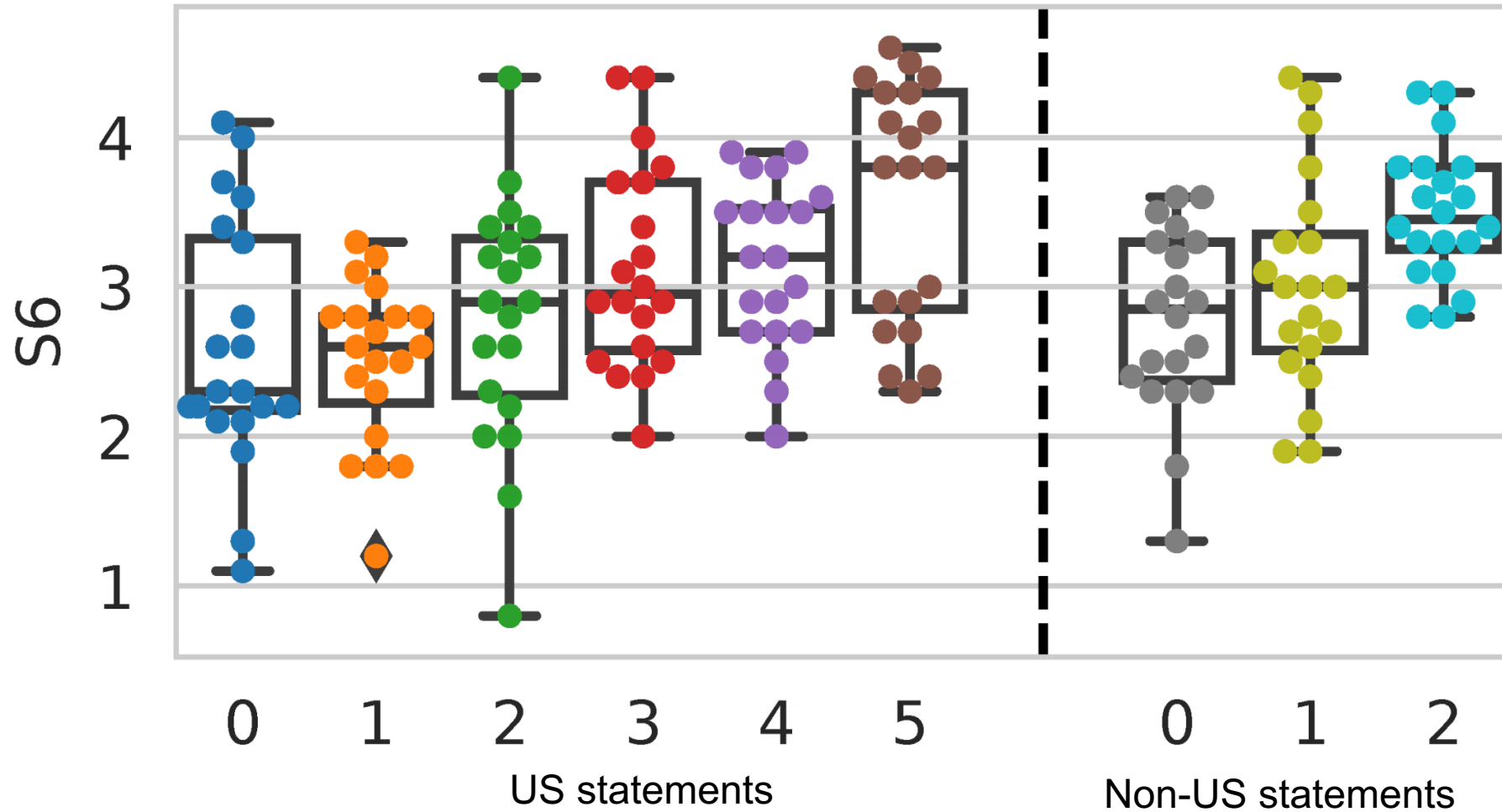
- US political statements (PolitiFact)
- non-US political statements (ABC)

3 scales (3, 6, and 100 levels)

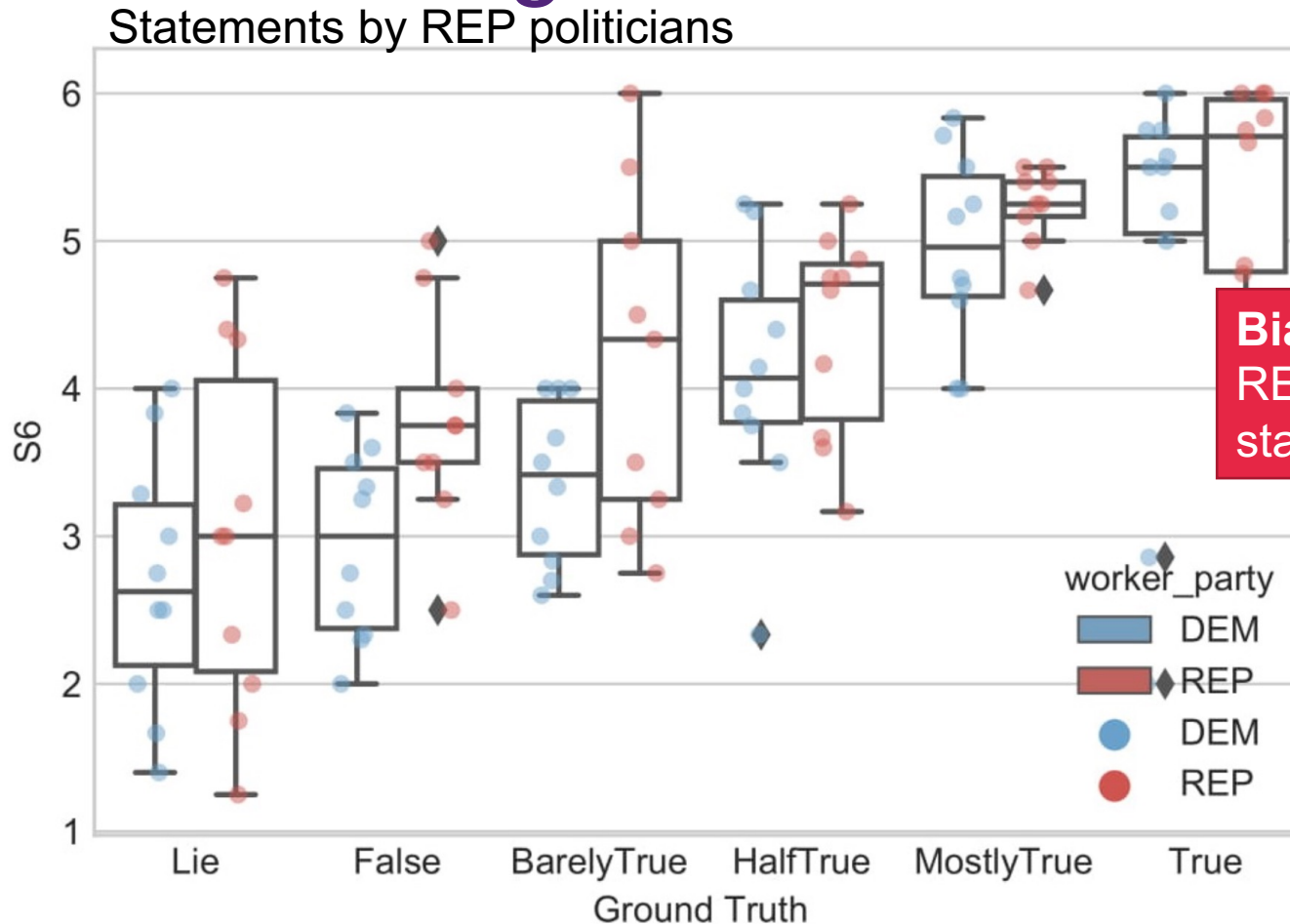
Table 1: Example of statements in the PolitiFact and ABC datasets.

	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

Crowd Performance VS Expert Ground Truth

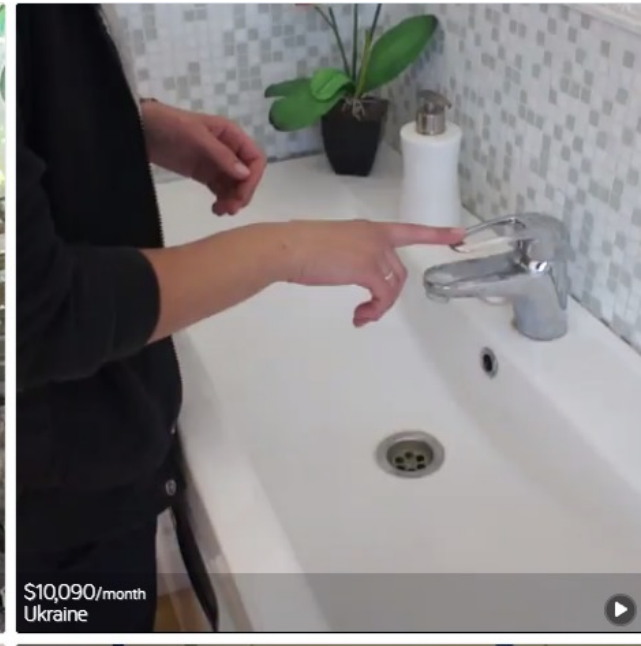


Fake News labelling - Political bias



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

Video of people washing hands across different socio-economic statuses



- 4 regions: Africa, Asia, Europe, the Americas; 4 different income level for each region ($4 \times 4 \times 7 = 112$)
- 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

Influencing human annotators

Controlling Bias

Relevance Assessment

Instructions

Query:
cosmic events
[Show more information](#)

Document 6 of 10

Lo, the Star, which they had seen in the east, went before them, till it came and stood over where the young child was. - St. Matthew, chapter 2, verse 9. Comet? Supernova? Conjunction of planets? Myth? Or miracle? Those, broadly speaking, are the five favourite explanations for the Star of Bethlehem - an apparition whose identity has attracted much speculation from astronomers, amateur and professional, over the years. Of course the atheist who regards the Bible as superstitious nonsense can dismiss the Star as a figment of the imagination of whoever wrote Matthew's gospel. At the other extreme, the fundamentalist who regards the Bible as literal truth can claim the Star simply as a miracle; if so, there is no point in trying to relate a divine sign to known astronomical events. However, there are many Christians who would be prepared to believe that God made use of natural events to give a sign of his Son's coming. And there are many non-Christians who accept that there could be historical - though not supernatural - explanations for some stories in the Bible. So, what natural phenomena could account for the Star of Bethlehem? Anyone who has seen the planet Venus shining bright in the western sky on a crisp winter evening may feel that no further explanation is needed. But Venus could not really be the answer. Nor could any other regular feature of our night sky. Surely only a really remarkable portent would have drawn the wise men or Magi several hundred miles west to Palestine from their presumed homes in Persia. If we are looking for a special astronomical event, we need to know its date as accurately as possible. Our current system of numbering years, supposedly starting from Christ's birth, was not fixed until the 6th century AD, when the historical record was far from clear. Modern Bible scholars believe that the date of the nativity was somewhere between 8 BC when the Emperor Augustus ordered his great tax census and 4 BC when King Herod died. They have no clues about the

Metadata

Tick it if you find it useful!	Attribute	Value
<input type="checkbox"/>	Human previous judgement	Not Relevant: 3 Marginally Relevant: 2 Relevant: 4 Highly Relevant: 1
<input type="checkbox"/>	Average time for human making a judgement decision	45.75 seconds
<input type="checkbox"/>	TF-IDF by machine over relevant documents	star: 84.10 bc: 68.85 bethlehem: 42.93 bible: 41.18 matthew: 38.93
<input type="checkbox"/>	Machine judged relevance	50.39% (95 out of 126)
<input type="checkbox"/>	Query words highlight	highlight
<input type="checkbox"/>	Title	Technology: Mystery of the Star of Bethlehem: 'T' links other words of wisdom on food, drink, nutrition and the legend you're likely to be taking part to this weekend
<input type="checkbox"/>	Length	749 words
<input type="checkbox"/>	Author	By CLIVE COOKSON
<input type="checkbox"/>	TF-IDF (over all documents)	bible: 84.55 matthew: 58.97 comet: 53.50 star: 52.52 bc: 52.16
<input type="checkbox"/>	Word frequency in document (Top 5)	star: 10 comet: 6 bible: 5 event: 5
<input type="checkbox"/>	Date of creation	23-Dec-94

Not Relevant
 Marginally Relevant
 Relevant
 Highly Relevant

Relevance judgement's justification:

[Submit and Go To Next Document](#)

- 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.

What happens when we train ML models with biased labels?

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

Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pinar 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

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>