# Data Quality Dimensions for Fair AI

Camilla Quaresmini and Giuseppe Primiero


UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FILOSOFIA
"PIERO MARTINETTI"

28 October 2021

# Table of Contents

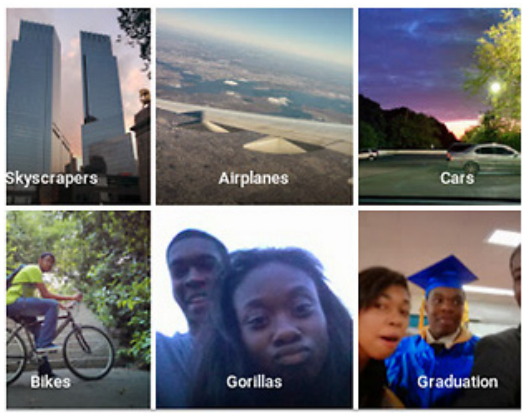
# Classification

## Definition

Classification is a task related to a predictive modeling problem, where a class label is predicted for a given input data, based on a classification model.

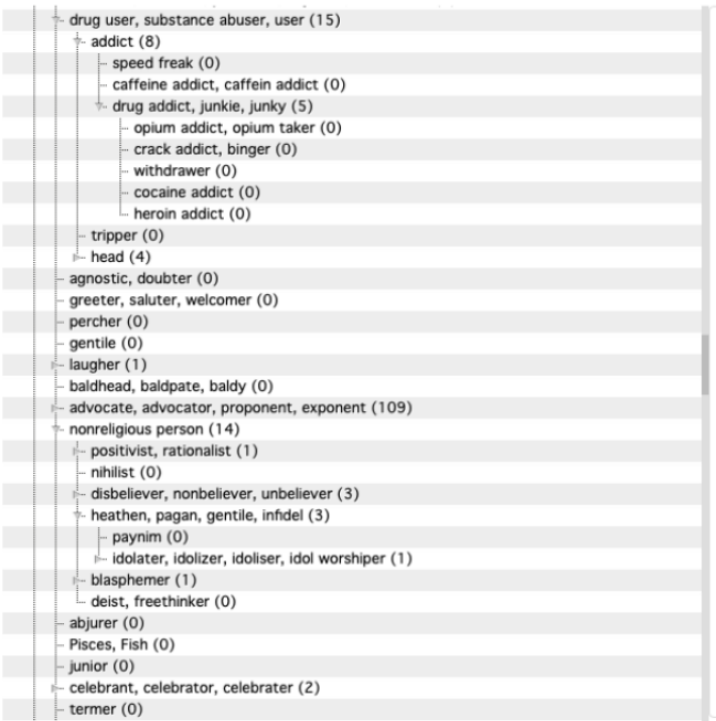
# Classification: Problems

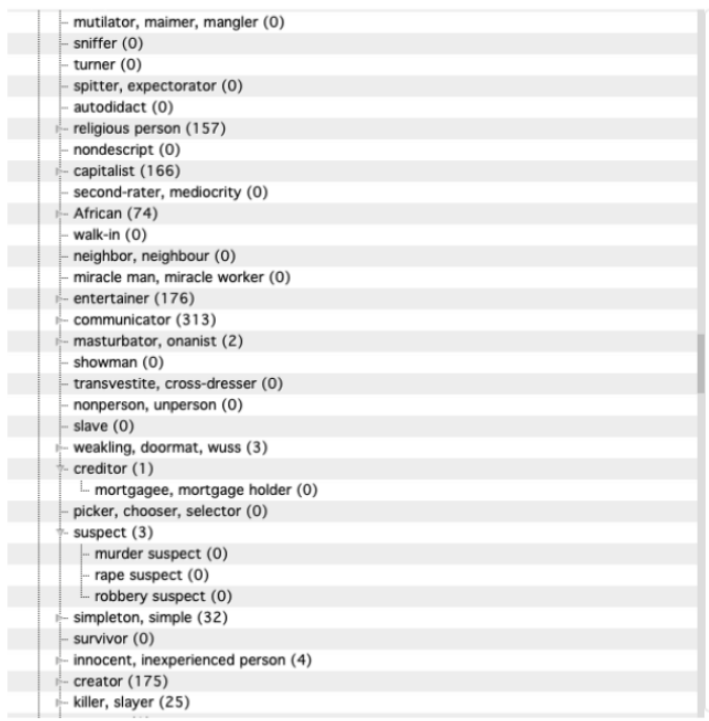
- Input data are datasets.
- Theoretically, datasets are representative of the world. But in the real life they are not.
- This results in systems full of biases.



# Classification: Problems

- Boulamwini and Gebru (2) demonstrate that training and evaluation data for such algorithms are often biased, producing distorted results.
- According to a recent MIT's study by Northcutt (4), the most well-known AI datasets are full of labeling errors.
- `https://www.image-net.org`



# Bias Mitigation Tools

- Such problems are beginning to be addressed by the ML community.
- A number of tools are available to evaluate the probability that a labelling error has occurred, and repair it.

# Cleanlab

- Cleanlab is an open-source Phyton package for finding and learning with label errors in datasets.
- It is powered by Confident Learning (CL).

# Cleanlab: Description

## Assumption

*Data are categorical, they fit one label.*

## Task

*Given a mapping $y* \rightarrow \tilde{y}$ between variables, where $y*$ is the correct label and $\tilde{y}$ the wrong one, what is the probability*

$$p(\tilde{y} = i | y* = j)$$

*such that label $i$ is wrong, given that label $j$ is correct?*

## Definition (Confident Joint)

$$C_{\tilde{y},y*}[i][j] := |\hat{X}_{\tilde{y}=i,y*=j}|$$

# Case Study

- We consider the FERET database.
- We focus on the particularly difficult (technically: noisy) attribute `Gender`, labels available `male;female`.
- We provide a conceptual and technical analysis with respect to available bias mitigation tools.

# Case Study

- We question the underlying assumption on data dimensions:

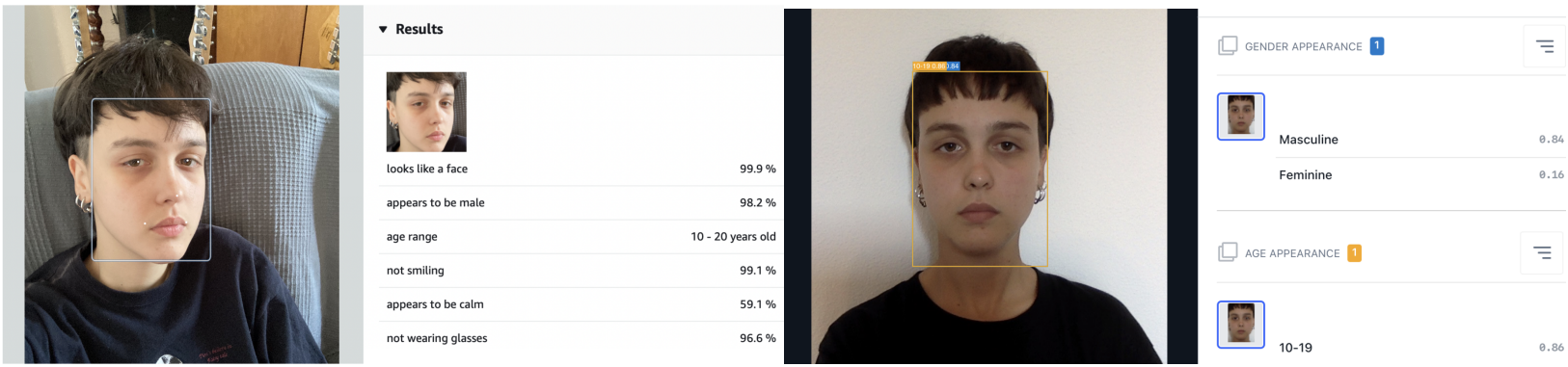
  1. completeness on labeling;
  2. accuracy as a measure of dataset correctness.

- We aim at suggesting improvements on errors identification in the classification of two difficult examples:

  1. non-binary individuals = datapoints which have none of the available labels as the correct one;
  2. transgender individuals = datapoints which have a label that changes under different conditions.

# First Case: Incomplete Label Set

- Consider a datapoint represented by a non-binary individual, whose actual label is missing.




(a) *Amazon Rekognition.* (b) *Clarifai.*

Figure: Example of incomplete label set.

# Scheuerman's Study, (5)

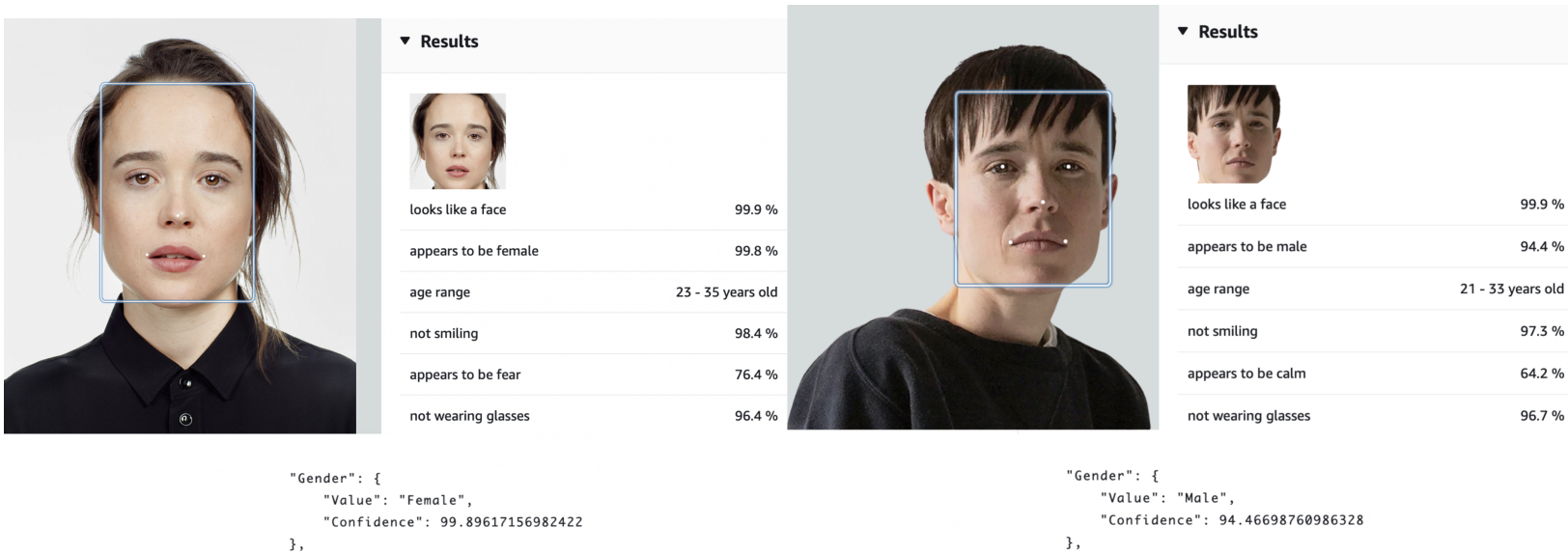
- A possible strategy is to implement a larger partition of labels adding separate categories.
- Technically, consider the completeness of the labels at a higher level of abstraction.
- 7 different genders.
- They consider four famous systems:

| TPR Performance Per Gender Hashtag | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Hashtag*** | ***Amazon*** | | | ***Clarifai*** | | | ***IBM*** | | | ***Microsoft*** | | | ***All*** |
| | *T* | *F* | *TPR* | *T* | *F* | *TPR* | *T* | *F* | *TPR* | *T* | *F* | *TPR* | *Avg* |
| *#woman* | 348 | 2 | 99.4% | 333 | 17 | 95.1% | 345 | 5 | 98.6% | 100 | 0 | 100.0% | 98.3% |
| *#man* | 334 | 16 | 95.4% | 344 | 6 | 98.3% | 341 | 9 | 97.4% | 348 | 2 | 99.4% | 97.6% |
| *#transwoman* | 317 | 33 | 90.6% | 271 | 79 | 77.4% | 330 | 20 | 94.3% | 305 | 45 | 87.1% | 87.3% |
| *#transman* | 216 | 134 | 61.7% | 266 | 84 | 76.0% | 250 | 100 | 71.4% | 255 | 95 | 72.8% | 70.5% |
| *#agender, #genderqueer, #nonbinary* | – | – | – | – | – | – | – | – | – | – | – | – | – |

- Available services offer a static classification non representative of individuals who are not cisnormative.

# Second Case: Inconsistent Datapoint

- A datapoint representing a transgender individual.
- In this case any extension of the label set is in fact misleading.




(a) *Ellen Page in 2015.* (b) *Elliot Page in 2021.*

Figure: Example of inconsistent datapoint in Amazon Rekognition.

# Philosophical Stand

- The assumption that completeness and accuracy are the only relevant data quality dimensions for classification tasks is wrong.
- The literature on data and information quality (1; 3) provides a number of dimensions which can help improving classification and bias mitigation.
- In particular, we believe an important starting point is represented by adding the dimension of timeliness.

# Our Strategy

Add a time frame $\mathcal{T} = t_1, \ldots, t_n$ whose length depends on the dataset and the classification task over the pairing of data points to labels, to measure a probability of a label-change over time.

## Proposition

*Given the label set is complete at time $i$, it can be incomplete at time $i + n$*

## Proposition

*Given the label $i$ is correct for data point $d$ at time $i$, it can be incorrect at time $i + n$*

# Implementing the strategy for Cleanlab (I)

## Assumption

*The probability value of a given label $i$ being wrong, given a label $j$ is correct (their distance) may change over time.*

## Task

*Given a mapping $y* \rightarrow \tilde{y}$ between variables, where $y*$ is the correct label and $\tilde{y}$ the wrong one, what is the probability*

$$p_{\mathcal{T}}[(\tilde{y} = i)_{t_n} | (y* = j)_{t_{n-m}}]$$

*such that label $i$ is wrong at time $t_n$, given that label $j$ was correct at time $t_{n-m}$?*

## Definition (Confident Joint)

$$C_{\tilde{y},y*}[i,j,\mathcal{T}] := |\hat{X}_{\tilde{y}=i_{t_n},y*=j_{t_{n-m}}}|, \forall m < n$$

# Implementing the strategy for Cleanlab (II)

## Assumption

*(Some) Data are not categorical for some labels, they fit one label at any given time, but possibly different ones at different times.*

## Task

*Given a mapping $y* \rightarrow \tilde{y}$ between variables, where $y*$ is the correct label and $\tilde{y}$ the wrong one, what is the probability*

$$p_{\mathcal{T}}[(\tilde{y} = i)_{t_n} | (y* = i)_{t_{n-m}}]$$

*such that label $i$ is wrong at time $t_n$, given that label $i$ was correct at time $t_{n-m}$?*

## Definition (Confident Joint)

$$C_{\tilde{y},y*}[i, \mathcal{T}] := |\hat{X}_{\tilde{y}=i_{t_n},y*=i_{t_{n-m}}}|, \forall m < n$$

# Conclusions

- The present study demonstrates that data quality is not just maximizing accuracy.
- A way to make current bias mitigation tools fairer is to implement more data quality metrics.
- Laying the ground for an extension of associated assessment tools.

# References


[1] Batini Carlo, Scannapieco Monica, "Data Quality: Concepts, Methodologies and Techniques", 2006.

[2] Buolamwini Joy, Gebru Timnit, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", 2018.

[3] Floridi Luciano, Illari Phyllis, "The Philosophy of Information Quality", 2006.

[4] Northcutt Curtis, Athalye Anish, Mueller Jonas, "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks", 2021.

[5] Scheuerman Morgan Klaus, Paul Jacob, Brubaker Jed, "How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services", 2019.