Presenting *genderphoto*: An Open-Source Tool for Recovering Missing Gender Labels in Innovation Data

Diletta Abbonato[1,2] and Cecilia Maronero[3]

[1]Department of Cultures, Politics and Society (CPS), University of Turin, Italy
(diletta.abbonato@unito.it)
[2]BETA, University of Strasbourg, France
[3]Bordeaux School of Economics (BxSE), Université de Bordeaux, France
(cecilia.maronero.1@u-bordeaux.fr)

Extended abstract submitted to the NBER Innovation Information Initiative Technical Working Group, 4–5 December 2026.

Software: https://github.com/zabbonat/genderphoto (MIT licence).

Abstract

Research on inventors and scientists often uses gender inferred from first names. Standard dictionaries work well for many names, but they also leave a large number of cases unclassified. These cases are commonly dropped, even though there is little reason to expect them to be random. We present *genderphoto*, an open-source Python package designed for the cases that name-based methods do not resolve. The package combines country-specific name dictionaries, pronouns in public search-result snippets, public images, a face classifier and a locally run vision-language model. It records the source and confidence of each label and keeps uncertain cases unknown. We use the USPTO Artificial Intelligence Patent Dataset to show how the tool works. The source data contain no gender attribution for 153,755 inventor listings, or 8.0% of the total. *genderphoto* raises coverage from 66.4% of names after the name stage to 92.6% after the full pipeline, which is 91.9% of all inventor listings on the 106,030 affected patents. Importantly, the recovered cases are also more likely to be classified as female.

1. The problem and the tool

Almost every statistic on women in invention starts from a first name. A name dictionary assigns a male or female label, and researchers use those labels to measure participation, collaboration, and inventive performance. The WIPO World Gender Name Dictionary (WGND 2.0) is the standard source for patent data (Lax-Martínez et al., 2016, 2021), and the USPTO's reports on women inventors rely on name-based attribution (Miguelez et al., 2019). Benchmarks show that these tools are accurate on the names they classify (Santamaría and Mihaljević, 2018), but they do not always return an answer, and their errors are unevenly distributed across groups (Lockhart et al., 2023). Researchers then drop the observation, keep a low-confidence guess or leave the label missing.

Missing labels are common in large innovation datasets. Romanized East Asian given names often carry little gender information. Other names change meaning across countries: Andrea is usually a man's name in Italy and a woman's name in Germany, Austria, and the United States, so a dictionary can fail even when it reports a high probability. If the missing cases differ from those a dictionary can classify, dropping them changes the measured participation of women, and so may keeping a weak guess.

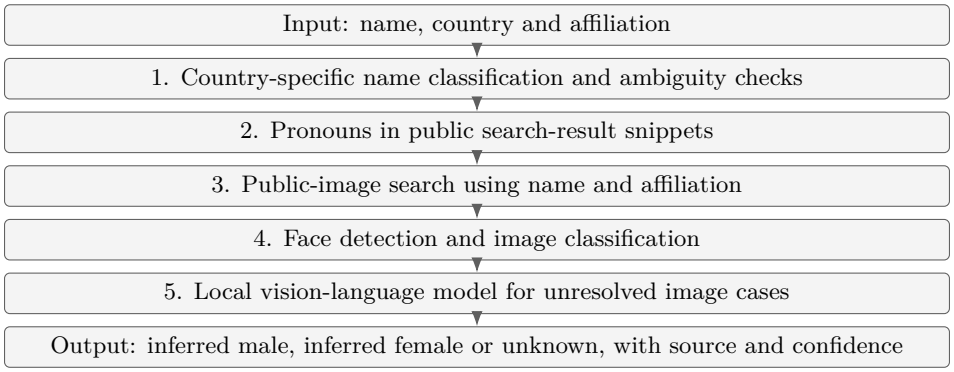
The tool *genderphoto* is designed to address these cases. It combines name dictionaries with public text and public images, records how each label was obtained, and leaves a case unknown when the

evidence is not clear enough. It can be run as a full pipeline or used only for name screening and missing-data diagnostics. We apply it to US AI patents as a large and difficult test case.

2. How *genderphoto* works

The input is a table with a person's name, country, and optionally affiliation, which makes the searches more precise. Each case moves through the stages in Figure 1 and leaves the pipeline as soon as one of them provides enough evidence. If none does, the label remains unknown. Thresholds and decision rules are set by the user rather than fixed in the code.

Figure 1: The *genderphoto* pipeline. The non-classified gender case leaves the pipeline as soon as a stage provides enough evidence.



1. Name classification. The first stage queries WGND 2.0 through the *global-gender-predictor* library, conditional on country, and adds a dictionary of romanized Chinese given names. A second library, *gender-guesser*, provides a cross-check. A name moves forward when the probability is below 0.75, when the two sources disagree, or when it is on a curated list of names with different gender associations across countries. These checks are essential as later stages can examine an unresolved case but cannot repair a confident label accepted too early.

2. Search-result text. The tool searches for the name and affiliation and accounts for gendered pronouns, where present, in the result snippets without opening the pages. A label is assigned only when one gender clearly dominates. This stage resolves relatively few cases, but provides a rapid alternative to image retrieval when sufficient textual information is available.

3. Image search. For the remaining cases, the tool retrieves up to three public images, searching first for the name with the affiliation, then with a professional keyword, and finally by the name alone. Including the affiliation in the search helps disambiguate individuals with identical names, particularly for common East Asian names.

4. Image classification. DeepFace classifies the faces detected by RetinaFace (Serengil and Ozpinar, 2021; Deng et al., 2020), and images without a detected face are discarded. A label is accepted when the classifications agree and average confidence is at least 90%.

5. Vision-language model. When the classifications disagree, or confidence is lower, the best image is passed to Qwen2.5-VL (Wang et al., 2024), run locally through Ollama and instructed to use the image rather than the name. Where the model disagrees with the majority of the face classifier, its label is kept. This rule reflects evidence that face-attribute classifiers have higher error rates for women than for men, with particularly large disparities for women of color (Buolamwini and Gebru, 2018), and exhibit uneven performance across racial groups (Kärkkäinen and Joo, 2021).

The output retains the classifications produced by both image-based methods, allowing prospective users to audit disagreements or apply an alternative decision rule without repeating the image search. By default, images are processed on the researcher's machine and deleted after classification. They are not sent to an external model, the package does not create or retain facial embeddings, and the resulting variable is an inferred binary gender classification rather than self-reported gender and does not capture non-binary identities. Image-based attribution has been used before in this community: Cook et al. (2026) combine image- and name-processing algorithms with clerical review to identify the race of more than 100,000 startup founders. *genderphoto* applies the same logic to gender and packages it as a reusable tool: the procedure is fully automated, each label carries the stage that produced it, and the cases that remain unlabelled are reported as bounds.

3. Application to US AI patents

The patent data come from the USPTO Artificial Intelligence Patent Dataset (AIPD), which identifies AI patents from the text of claims and abstracts (Giczy et al., 2022), in its 2023 update (Pairolero et al., 2025). We combine it with disambiguated inventor, assignee, and location data from PatentsView and keep patents granted between 2012 and 2023. The data matching process results in 626,268 patents with 1,916,480 inventor listings. The USPTO gender attribution, based on the WIPO dictionary, provides no gender classification for approximately one inventor in twelve. At the patent level, one patent in six lists at least one inventor without a gender classification, corresponding to 106,030 patents. We define these patents as our target sample and apply *genderphoto* to all inventors listed on them. This sample is not representative of AI patents as a whole: inclusion is determined by the failure of the name-based procedure to classify at least one inventor on the patent.

Table 1 summarizes the composition of the target sample and the classification results at each stage of the pipeline. The target patents carry 429,404 inventor listings, which reduce to 169,406 distinct unclassified names. The pipeline is run once per name since the cost is in retrieving and reading images; the label is then attached back to every listing of that name. The name stage classifies 66.4% of the names. The later stages recover 77.9% of the cases sent forward, so the share of names carrying a label rises to 92.6%, and 91.9% of the inventor listings on the target patents carry one.

Table 1: Application of *genderphoto* to USPTO AI patents

Patents in the target sample	106,030
Total inventor listings on them	429,404
Distinct unclassified inventor names run	169,406
classified at the name stage	112,438 (66.4%)
sent to the later stages	56,968 (33.6%)
recovered by the later stages	44,361 (77.9%)
left unknown	12,607 (22.1%)
Names carrying a label, name stage to full pipeline	66.4% to 92.6%
Inventor listings carrying a label	91.9%
Female share among name-classified inventors	17.2%
Female share among recovered inventors	34.9%

Notes. The target contains patents with at least one inventor left unattributed in the source data. A listing is a patent–inventor pair, so a person on several patents appears several times. The pipeline is run once per distinct name and the label is then attached back to every listing of that name. Grant years are 2012–2023.

The missing cases are not a random part of the data. Women account for 17.2% of the inventors classified at the name stage and 34.9% of those recovered later, and about seven eighths of the difference

lies within countries rather than in the different country composition of the two groups. Where the vision-language model overrides the face classifier, it moves 3,368 of 4,301 cases towards female, but the direction of the result does not depend on it: accepting only labels on which the face classifier is unanimous and confident, the tool recovers 32,235 inventors, of whom 26.7% are women. Keeping the dictionary's low-confidence guess instead would put the female share of the ambiguous names as a whole at 13.5%, against the 34.9% the pipeline finds. These comparisons are not estimates of the gender gap in AI patenting. They show that a technical choice about missing labels changes the descriptive statistics used in that work.

The package also computes worst-case bounds on female participation (Manski, 2003). Across the 429,404 inventor listings on the target patents, the name stage alone identifies the female share only to within $[10.2, 48.4]$, an interval 38.2% points wide, because it leaves 38% of the listings without a label. After the full pipeline the interval is $[20.4, 28.4]$, just over eight points wide. Researchers who do not want image-based labels can still use the package to report missingness and bounds.

4. Validation

The repository contains a diagnostic set of 100 researchers whose gender is reported on an institutional page, chosen before the final run to include difficult cases: Italian names outside Italy, romanized East Asian names, French and English androgynous names, and male and female controls. The pipeline labels 92 of the 100 correctly, including every case settled at the name stage. It labels 34 cases female against 39 true women: its errors lose women rather than invent them, so a perfect instrument would widen the difference in Section 3.

This set tests the pipeline but is not a random sample of the population to which the tool is applied, and we do not treat 92% as a population accuracy. We are completing a stratified random audit of 538 inventors drawn from the recovered USPTO cases, coded by hand, blind to the label that the pipeline returned. We will report error and non-recovery rates by gender, country group and stage, use the audit to assess disagreements between the face classifier and the vision-language model, and release the coding sheet and scoring script.

The main remaining source of error is person matching: image search may return someone else with the same name, especially when affiliation information is weak. People without public text or a public image cannot be recovered. Search results change over time and the vision-language model is not fully deterministic, so a new run reproduces aggregate patterns rather than every individual label. For this reason each output row reports the stage, confidence and available provenance, not only a final label.

5. Outputs, uses, and public release

The package returns the input data with added columns for the dictionary label and probability, ambiguity flags, the labels from the later stages, confidence, the number of images examined, the stage that produced the result, and the consolidated label. Long jobs write regular checkpoints and can be resumed. Processing takes a few seconds per unresolved person on a desktop machine with a consumer graphics card. The pipeline is not specific to patents: it can be applied to any list of people with a name and a country, such as PatentsView and EPO inventor files, Scopus or OpenAlex author records, grant data and company registers. A researcher can run the full pipeline, stop after the name stage or use only the diagnostics and bounds.

genderphoto is available on GitHub under an MIT licence. The final release will add pinned dependencies and full documentation, an example notebook with sample input files, the 100-case diagnostic set with its expected output, the 538-case audit data with its coding sheet and scoring script, the code that produces the USPTO results reported here, and a versioned archive with a permanent DOI. For

the USPTO application, we will release aggregate tables: coverage, recovery rates and female shares, each with the corresponding bounds, by country of residence, grant year and technology field. We will also release the inventor-level labels, keyed to patent number and inventor name, so that they can be joined to PatentsView and used directly.

At the I3 Technical Working Group Meeting, we would especially welcome feedback on the validation design, including sources of self-reported gender that could serve as an external benchmark, the provenance that should accompany each inferred label. We would also welcome collaboration on validation in other innovation datasets.

References

Buolamwini, J., Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 77–91.

Cook, L.D., Marx, M., Yimfor, E., 2026. Funding black high-growth startups. The Journal of Finance 81, 1619–1660.

Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S., 2020. RetinaFace: Single-shot multi-level face localisation in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 5203–5212.

Giczy, A.V., Pairolero, N.A., Toole, A.A., 2022. Identifying artificial intelligence (ai) invention: A novel ai patent dataset. The Journal of Technology Transfer 47, 476–505.

Kärkkäinen, K., Joo, J., 2021. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558.

Lax-Martínez, G., Raffo, J., Saito, K., 2016. Identifying the gender of PCT inventors. WIPO Economic Research Working Paper .

Lax-Martínez, G., Raffo, J., Saito, K., 2021. The WIPO worldwide gender-name dictionary 2.0. WIPO Economic Research Working Paper .

Lockhart, J.W., King, M.M., Munsch, C., 2023. Name-based demographic inference and the unequal distribution of misrecognition. Nature Human Behaviour 7, 1084–1095.

Manski, C.F., 2003. Partial identification of probability distributions. Springer.

Miguelez, E., Toole, A., Myers, A., Breschi, S., Ferruci, E., Lissoni, F., Sterzi, V., Tarasconi, G., 2019. Progress and Potential: A profile of women inventors on US patents. Technical Report.

Pairolero, N.A., Giczy, A.V., Torres, G., Islam Erana, T., Finlayson, M.A., Toole, A.A., 2025. The artificial intelligence patent dataset (aipd) 2023 update. The Journal of Technology Transfer 50, 2587–2610.

Santamaría, L., Mihaljević, H., 2018. Comparison and benchmark of name-to-gender inference services. PeerJ Computer Science 4, e156.

Serengil, S.I., Ozpinar, A., 2021. Hyperextended lightface: A facial attribute analysis framework, in: 2021 international conference on engineering and emerging technologies (ICEET), IEEE. pp. 1–4.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al., 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 .

Acknowledgments

This research received no external funding. The authors have no financial relationships or other potential conflicts of interest to disclose.