POLITECNICO
MILANO 1863

Gender Discrimination in Data Analysis:
a Socio-Technical Approach

Supervisor: Prof. Viola Schiaffonati   M.Sc. Thesis by:
Co-supervisor: Prof. Letizia Tanca   Riccardo Corona
                             Prof. Pierre Senellart   927975
                             Prof. Karine Gentelet   07/10/2021

Data analysis

Set of processes for inspecting, cleaning, transforming, and modeling data with the aim of discovering useful information, informing conclusions, and supporting decision making.

Gender discrimination

Specific (sub)category of social problems, here expressed in the form of the so-called '**gender gap**', definable as:

A difference between the way men and women are treated in society, or between what men and women do and achieve. (Cambridge Advanced Learner's Dictionary, 2013)

Problem

Data and datasets, on which a lot of actions of our daily routine are based, can be **unfair**. Unfair, or better to say, **biased** data, may influence, directly or indirectly, our perception of reality, and lead us to make decisions that, although seemingly fair and just, contain in turn bias, and discriminate against individuals or groups of individuals.

Example scenarios

- COMPAS tool used in the U.S. to predict recidivism risk biased against Black people. *(Angwin et al., 2016)*
- Amazon software to screen candidates for employment biased against women. *(Dastin, 2018)*

Riccardo Corona   **POLITECNICO** MILANO 1863

Dual perspective on the problem: qualitative and conceptual approach typical of sociology combined with technical and pragmatic approach of engineering and computer science.

- Systematic literature review.
- Experimental case studies through Python and Jupyter Notebook to verify not only whether or not men and women are paid fairly, but also if there are other collateral issues related to gender discrimination in the data and what tools can possibly catch them.
- Technical results discussed in the light of the sociological background.

We try to understand to what extent the problem is to be found in the *instrumentation and technical choices* and to what extent it is instead rooted in *society*.

Sociological Perspective

Global Gender Gap Index

Cumulative measure for ranking countries which benchmarks national gender gaps on economic, education, health, and political criteria.
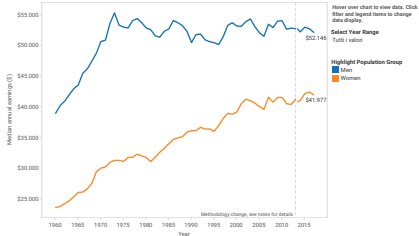
U.S. ranking: 49/144, overall score: 0.718 (1 = gender parity)
$\rightarrow$ men more participatory in labor force; women tend to earn less than men, be employed part-time, or not to be paid for their work, and are underrepresented in managerial and higher-paying jobs.

Some possible reasons:

- Statistical discrimination *(Tilcsik, 2021)*
- Institutional environment *(Beggs, 1995)*
- Unequal bargaining power *(Folbre, 2021)*

Riccardo Corona   **POLITECNICO** MILANO 1863

Median annual earnings by sex
March 1960-2017

Hover over chart to view data. Click filter and legend items to change data display.

Select Year Range
Tutti i valori

Highlight Population Group
Men
Women

$52,146
$41,977
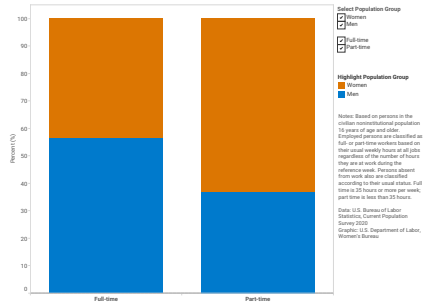
Methodology change, see notes for details

Notes: Earnings are based on median annual earnings of full-time, year-round workers, 15 years old and over beginning in March; 1980, and age 14 years old over as of March of the following year for previous years. Before 1989 earnings are for civilian workers only.
The comparability of historical data has been affected at various times by methodological and other changes in the Current Population Survey. The 2014 CPS ASEC included redesigned questions for income and health insurance coverage for a subsample of the 98,000 addresses using a probability split panel design. Approximately 68,000 addresses were eligible to receive a set of income questions similar to those used in the 2013 CPS ASEC and the remaining 30,000 addresses were eligible to receive the redesigned income questions, resulting in two estimates for 2013. Estimates based on the portion of the sample that received the redesigned income questions are the most appropriate for comparing estimates from ASEC 2014 with ASEC 2015 and beyond.
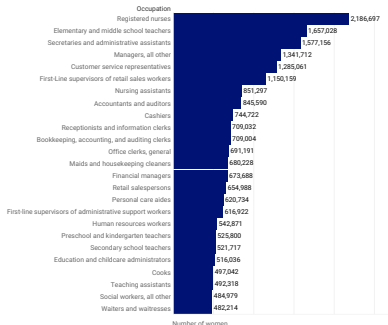Earnings are in 2017 CPI-U-RS adjusted dollars.
Source: 1961-2018 Annual Social and Economic Supplements, Current Population Survey, U.S. Census Bureau.
Graph by the Women's Bureau, U.S. Department of Labor

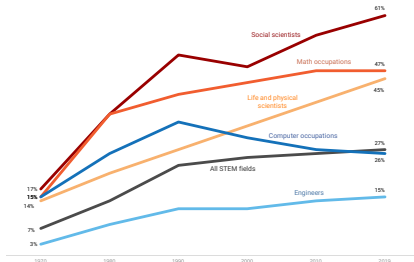**Percent distribution of workers employed full- and part-time by sex**

Select Population Group
☑ Women
☑ Men

☑ Full-time
☑ Part-time

Highlight Population Group
Women
Men

Notes: Based on persons in the civilian noninstitutional population 16 years of age and older. Employed persons are classified as full- or part-time workers based on their usual hours at all jobs regardless of the number of hours they are at work during the reference week. Persons absent from work also are classified according to their usual status. Full time is 35 hours or more per week; part time is less than 35 hours.

Data: U.S. Bureau of Labor Statistics, Current Population Survey 2020
Graphic: U.S. Department of Labor, Women's Bureau

Most Common Occupations for Women in the Labor Force

Occupation	Number of women
Registered nurses	2,186,697
Elementary and middle school teachers	1,657,028
Secretaries and administrative assistants	1,577,156
Managers, all other	1,341,712
Customer service representatives	1,285,061
First-Line supervisors of retail sales workers	1,150,159
Nursing assistants	851,297
Accountants and auditors	845,590
Cashiers	744,722
Receptionists and information clerks	709,032
Bookkeeping, accounting, and auditing clerks	709,004
Office clerks, general	691,191
Maids and housekeeping cleaners	680,228
Financial managers	673,688
Retail salespersons	654,988
Personal care aides	620,734
First-line supervisors of administrative support workers	616,922
Human resources workers	542,871
Preschool and kindergarten teachers	525,800
Secondary school teachers	521,717
Education and childcare administrators	516,036
Cooks	497,042
Teaching assistants	492,318
Social workers, all other	484,979
Waiters and waitresses	482,214

Note: Full-time, year-round civilian employed 16 years and older. Occupations with at least 100 sample observations.
Data: U.S. Census Bureau, American Community Survey 2019
Graphic: U.S. Department of Labor, Women's Bureau.

Percentage of science, technology, engineering, and math (STEM) workers who are women

Social scientists — 61%
Math occupations — 47%
Life and physical scientists — 45%
Computer occupations — 27%
All STEM fields — 26%
Engineers — 15%

1970: 17%, 15%, 14%, 7%, 3%

Note: STEM occupations are classified according to the Standard Occupational Classification STEM recommendations for presentation of government data available at https://www.bls.gov/soc/Attachment_C_STEM_2018.pdf
Source: U.S. Census Bureau, decennial Census 1970-2000 and American Community Survey public use microdata 2010 and 2019.
Graphic by the Women's Bureau, U.S. Department of Labor

Riccardo Corona

POLITECNICO MILANO 1863

- *The 'Glassdoor Method'*: a framework for evaluating gender pay gap which relies on **linear regression**. *(Chamberlain, 2017)*

Linear regression

Preprocessing technique used to smooth out noise or to find patterns within a dataset, which attempts to model the relationship between two or more variables by fitting data to a linear equation.

$$y = \beta_0 + \beta_1 x + \epsilon$$
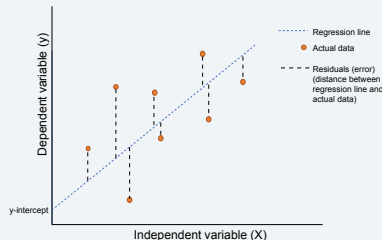
- *FAIR-DB*: an algorithm to detect bias in data based on **functional dependencies** and the related evaluation metrics. *(Azzalini, Criscuolo, and Tanca, 2021)*

Functional Dependencies (FDs)

Constraint involving two (sets of) attributes of the same relation in which the first *uniquely determines* the second.

Approximate Conditional Functional Dependencies (ACFDs)

FDs holding on a subset of tuples (Approximate) which use conditions on attribute values to specify the subset on which they hold (Conditional).

$Status =$' F ', $Gender =$' female ' $\rightarrow AnnualSalaryBin =$'$\leq 90K$ '

Riccardo Corona

POLITECNICO MILANO 1863

- *Ranking Facts*: an application built on the idea of **ranking** which makes use of three statistical measures to evaluate fairness. *(Yang et al., 2018)*
  - **FA*IR**: compares the number of protected elements in every prefix of the ranking (i.e., the top-i positions, with $i \in [1, k]$) with the expected number of protected elements if they were picked at random using Bernoulli trials with success probability p.
  - **Proportion**: statistical measure based on the concept of *(two sample) z-test* – particular type of hypothesis test which allows to compare two proportions to check whether they are the same or not.
  - **Pairwise**: compares options in pairs and determines which is the preferred choice or has the highest level of importance based on defined criteria. At the end of the comparison process, each option has a rank or relative rating as compared to the others.

- **Data preprocessing**: 20,309 tuples, of which 16,146 males and 4,163 females, and with 35 distinct *Job Title* values and 20 distinct *Department* values.

- **The 'Glassdoor Method'**: 24.2% 'unadjusted' pay gap; 0.4% 'adjusted' pay gap $\rightarrow$ no evidence of a systematic gender pay gap.
- **FAIR-DB**: 6 final functional dependencies; 11.4% of the dataset 'problematic' $\rightarrow$ dataset quite fair.
- **Ranking Facts**: dataset fair for both males and females, for each statistical measure.

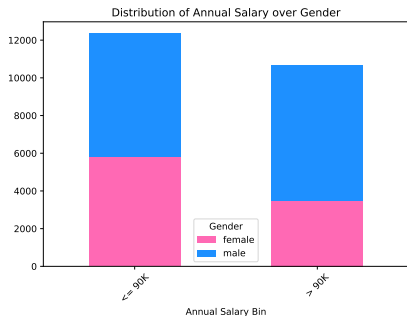Distribution of Annual Salary over Gender

Case Study 2: San Francisco
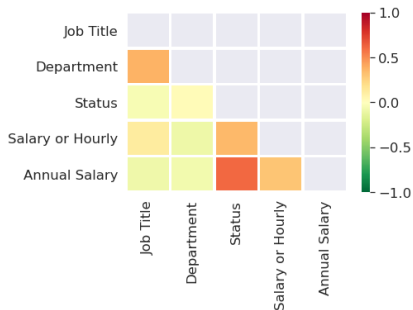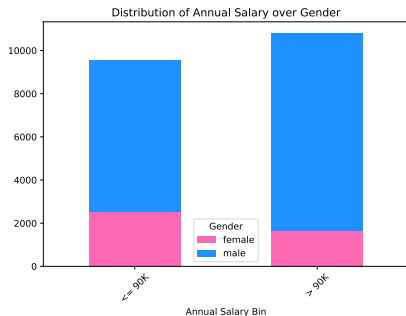
- **Data preprocessing**: 22,996 tuples, of which 13,688 males and 9,308 females, and with 81 distinct *Job Title* values.

- **The 'Glassdoor Method'**: 30.4% 'unadjusted' pay gap; -5% 'adjusted' pay gap $\rightarrow$ no evidence of a systematic gender pay gap.
- **FAIR-DB**: 10 final functional dependencies; 24.3% of the dataset 'problematic' $\rightarrow$ dataset quite fair because of the low values of *difference* ('unfairness level') and *support* (number of tuples involved), but for higher-paying jobs men seem to have an economic advantage over women.
- **Ranking Facts**: dataset fair for males and unfair for females, for each statistical measure $\rightarrow$ proportion of women in the top-k ranking effectively very low.

Riccardo Corona   **POLITECNICO** MILANO 1863

Distribution of Annual Salary over Gender

- **Part-time employees removal**: most of the tuples removed related to women (Chicago); excessive amount of tuples removed (San Francisco).
- **FAIR-DB: discretization using more bins**: less and different final dependencies detected (Chicago and San Francisco).
- **FAIR-DB: choice of different dependencies**: 85.6% (Chicago) and 92.5% (San Francisco) of the dataset 'problematic'.
- **Grouping of job titles**: overturning of the outcomes for Ranking Facts (Chicago dataset unfair for males and fair for females, for each statistical measure).
- **Voluntary introduction of bias**: results from each tool oriented toward unfair Chicago dataset, in which women are discriminated against (retaining 50%, 75%, and 90% of the *Annual Salary* value of female employees).

- Strengths and weaknesses of the tools highlight their non-exhaustiveness and complementarity.
- Tools practically fail in capturing the several facets of *equity*.

Equity

The idea that people should have access to resources (possibly of a different nature and to a different extent) in order to be able to reach the same condition.

- **Representation problem**: disproportion in the percentage of women employed in different sectors.
- **Part-time problem**: higher number of women employed in part-time jobs, typically less paid than full-time ones.

- Fairness is a multifaceted concept which cannot be exhausted by providing a single definition and pursuing that specific definition experimentally.
- Tools are susceptible to decisional choices, and therefore users must be properly trained on the specific area of analysis.
- Double perspective on the gender pay gap issue emphasized the importance of multidisciplinarity, especially when dealing with problems of an ethical and sociological nature.

- Sociological
  - ▶ Non-specific literature
  - ▶ No direct insights from U.S. workers
- Technical
  - ▶ Original datasets already partial or possibly grouped
  - ▶ gender-guesser (package to infer employees' gender)
- Design
  - ▶ Removal of job titles with less than 100 occurrences
  - ▶ FAIR-DB: parameter values; manual selection of rules; number of bins

- Combine all the tools in a unique, more complete instrument, possibly trying to encompass even more facets of equity, or more definitions of fairness.
- Support analyses of this kind by sociological research.
- Enrich sociological research by conducting an interview with workers and HR practitioners of the cities under study.
- Retrieve further information in support of the mere data and create effective documentation, possibly pointing at **context-awareness** (provide the tools with knowledge on the context of use).

Thank you!

Any question?

POLITECNICO MILANO 1863

Julia Angwin et al. "Machine Bias". In: *ProPublica* (2016).
https://www.propublica.org.

Fabio Azzalini, Chiara Criscuolo, and Letizia Tanca. "FAIR-DB: FunctionAl
Dependencles to discoveR Data Bias". In: *EDBT/ICDT Workshops*.
http://ceur-ws.org/Vol-2841. 2021.

John J Beggs. "The Institutional Environment: Implications for Race and Gender
Inequality in the U.S. Labor Market". In: *American Sociological Review* 60.4
(1995), pp. 612–633. DOI: 10.2307/2096297.

Cambridge Advanced Learner's Dictionary. ""Gender gap"". In: *Cambridge University
Press* (2013). https://dictionary.cambridge.org. Accessed 28 June 2021.

Andrew Chamberlain. "How to Analyze Your Gender Pay Gap: An Employer's Guide".
In: *Glassdoor Economic Research* (2017). https://www.glassdoor.com.

Riccardo Corona   **POLITECNICO** MILANO 1863

Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* (2018). https://www.reuters.com.

Nancy Folbre. "Gender inequality and bargaining in the U.S. labor market". In: *Economic Policy Institute* (2021). https://www.epi.org/209716.

András Tilcsik. "Statistical Discrimination and the Rationalization of Stereotypes". In: *American Sociological Review* (2021). DOI: 10.31235/osf.io/gf7ry.

Ke Yang et al. "A Nutritional Label for Rankings". In: *Proceedings of the 2018 International Conference on Management of Data*. Association for Computing Machinery, 2018, pp. 1773–1776. DOI: 10.1145/3183713.3193568.

Riccardo Corona   **POLITECNICO** MILANO 1863