

Big Data in Practice: Women (and Men) in Global Science

Center for Global Higher Education (CGHE)
University of Oxford, April 4, 2023



Professor Marek Kwiek
Institute for Advanced Studies in Social Sciences and
Humanities (IAS), Director
UNESCO Chair in Institutional Research and Higher
Education Policy
AMU University of Poznan, Poland, and
Visiting Researcher, German Center for Higher Education
Research and Science Studies DZHW, Berlin, Germany
kwiekm@amu.edu.pl

Twitter: @Marek_Kwiek

2. Exploratory Study - Introduction


- **Exploration of global bibliometric data sources** to study the **changing academic profession and its demographics** (Scopus data; 38 OECD countries).
- **What we can know.**
- **Assessing usefulness** of global data sources (gender, age, discipline, and time).
- **Measuring** demographic changes in global science **using new data sources.**
- Traditional approaches **inadequate**: national statistics (**OECD, UNESCO, Eurostat**) and surveys.
- **Discussing pros and cons** of using global publication and citation databases in academic profession studies.
- **Moving from bibliometrics (papers) to global academic profession studies (academics); and from publications to scientists.**



3. Poznan CPPS Team 2023: Global Academic Profession Research

- **Alicja Laskowska**, CPPS intern, Data Collection
- **Jakub Szymkowiak**, CPPS intern, Data Analysis & Visualizations
- **Dr. Wojciech Roszka**, Statistics, Observatory of Polish Science Dataset
- **Lukasz Szymula**, doctoral student, Big Data Analytics, Scopus Dataset, collaboration with ICSR Lab (May-November 2023: Boulder, Colorado, with Aaron Clauset's Lab)
- **Prof. Dominik Antonowicz**, Polish National Academic Profession Survey 2023 (and 2010)
- **Dr. Marcin Byczynski**, Projects Coordination
- **Prof. Marek Kwiek** (Head)

GENDER DISPARITIES IN INTERNATIONAL RESEARCH COLLABORATION: A STUDY OF 25,000 UNIVERSITY PROFESSORS

Marek Kwiek* 



Center for Public Policy Studies, University of Poznan,
Poland

Wojciech Roszka 

Institute of Informatics and Quantitative Economics, Poznan University of
Economics and Business
Poznan, Poland

Scientometrics (2022) 127:1697–1735
<https://doi.org/10.1007/s11192-022-04308-7>

Are female scientists less inclined to publish alone? The gender solo research gap

Marek Kwiek¹  · Wojciech Roszka² 



Gender-based homophily in research: A large-scale study of man-woman collaboration

Marek Kwiek^a, Wojciech Roszka^b

^a Institute for Advanced Studies in Social Sciences and Humanities (IAS), UNESCO Chair in Institutional Research and Higher Education Policy, Adam Mickiewicz University in Poznan, Poland

^b Poznan University of Economics and Business, Poznan, Poland



arXiv > cs > arXiv:2301.06196

Computer Science > Digital Libraries

[Submitted on 15 Jan 2023 (v1), last revised 9 Feb 2023 (this version, v2)]

Young Male and Female Scientists: A Quantitative Exploratory Study of Changing Demographics of the Global Scientific Workforce

Marek Kwiek, Lukasz Szymula



5. Why Women Leave Academic Science? (1/2)

- Both men and women leave academic science – but women leave it earlier (postdoctoral stage before creating own labs) and in larger proportions.
- Leaky pipeline vs. glass ceiling metaphors.
- Three theories (empirically tested).
 - (1) **The chilly climate theory:** a **hostile or unwelcoming** work environment in STEM fields can **discourage** women from pursuing and persisting in these fields.
 - (2) **The self-selection theory:** women are underrepresented in STEM fields because they are **less interested** in pursuing careers in these fields due to **societal and cultural factors** that discourage them.
 - (3) **The leaky pipeline theory:** a significant **loss of talent at every stage** of the academic career pipeline due to **systemic barriers** such as bias and discrimination.

6. Why Women Leave Academic Science? (2/2)

- Huge numbers of empirical studies. **Reasons in STEM** (sometimes: HUM and SOC) include:
- **Gender bias & discrimination**: in hiring, promotions, and pay; a hostile work environment.
- **Lack of work-life balance**: long working hours, high pressure to publish, secure grants; family responsibilities, caregiving.
- **Fewer role models and mentors**: isolation in male-dominated fields, limited support of female role models and mentors.
- **Limited opportunities for career advancement**: barriers to promotion and leadership positions.
- **Lack of institutional support**: insufficient institutional resources to help women overcome the challenges (child care, flexible work arrangements, mentoring programs).

7. Introduction (1/3): Digital Traces in Academic Profession Studies

- New opportunities for **collecting & analyzing data about academics**; offers new data **sources to study academic careers**.
- **Academics leave traces in their indexed publications. No other reliable traces (globally) today!**
- We can **combine** the digital traces with biographical, demographic, administrative (registries) & related data, both national & international.
- **Tracing academics & their careers** (longitudinal; countries; teams; men & women; juniors & seniors; disciplines).
- **Remarkable level of detail: measuring the academic profession with ever more precision possible!** (with some limitations)
- Demand for **more detailed, faster, and larger sample data** from researchers and policymakers.

8. Introduction (2/3): Academic Profession Studies and Structured Big Data


- Big Data **repurposed** for research from non-research sources.
- **Enormous and complex data** available today (high access costs).
- Big Data useful for new research questions and **testing old theories**.
- Extract **useful information** about academics from large datasets.
- Hundreds of millions of cells provide insight into academic profession (CPPS: 1.43 billion cited references).
- **Structured** data preferred (Scopus, WoS, national registries, CRIS systems).
- Big Data dramatically increases **insight into academic profession**.

9. Introduction (3/3): What to Explore Using Structured Big Data?

- **New** possibilities for **exploring the micro-level of individuals**, unimaginable a decade ago.
 - **Research productivity**
 - **Collaboration**
 - Citations (**scholarly impact**).
 - **Academic mobility**, national, cross-national, and cross-sectoral.
 - **All scholarly activities recorded in publications metadata, admin and biographical datasets (research mission only!)**.
- By **gender, age, academic seniority, and disciplines**.
- Both **statically** (e.g. 2022) and **dynamically**, over time (e.g. 2000-2022).
- **Longitudinal study designs**: (1) **"Once Highly Productive, Forever Highly Productive"**? (**Poland 2023; OECD 2023**) (2) **"The Young and the Old, the Fast and the Slow"**

Higher Education
<https://doi.org/10.1007/s10734-023-01022-y>

Once highly productive, forever highly productive? Full professors' research productivity from a longitudinal perspective

Marek Kwiek^{1,2,3}  · Wojciech Roszka^{2,4} 

Accepted: 7 March 2023
© The Author(s) 2023

arXiv > cs > arXiv:2211.06319

Computer Science > Computers and Society

[Submitted on 3 Nov 2022]

The Young and the Old, the Fast and the Slow: Age, Productivity, and Rank Advancement of 16,000 STEMM University Professors

Marek Kwiek, Wojciech Roszka

10. Studying Academic Careers and Access to Digital Databases (Our Current Approach at CPPS)

- **Transition needed:** from global publication metadata (bibliometrics) to global metadata on scholars (global academic profession studies).
- Combining **national-level data and Big Data**.
- Focus on **scientists** (and their attributes) rather than publications (and their properties).
- Large-scale and longitudinal approaches possible **increasing access to digital databases**.
- Databases: **national & global, commercial & noncommercial**, labor, workforce, administrative, bibliometric, and others.
- Examples of databases: Web of Science, Scopus, Microsoft Academic Graph, OpenAlex, Academic Analytics, DBLP, CRISTIN, POL-on.
- **Scholars have their attributes: gender, age, collaboration patterns, international mobility patterns, changing affiliations etc. Influence of all factors!**

11. The Changing Demographics of the Global (OECD) Academic Profession

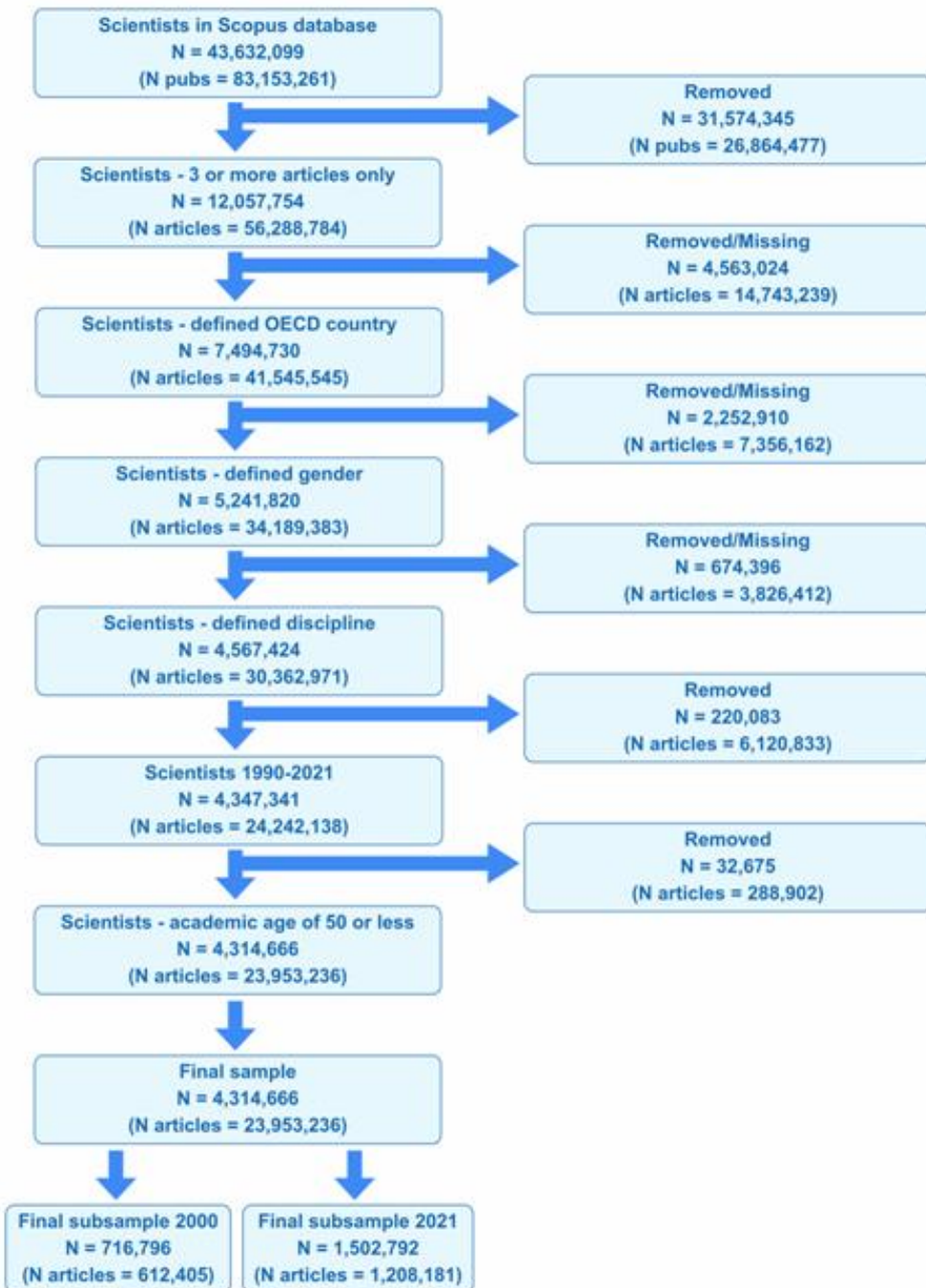
- **Rationale:** explore **changing demographics of global academic profession** using available bibliometric data sources.
- Focus on **four dimensions:** gender, age, discipline, and time (trends).
- **Testing how demographic transformations of the global academic profession can be measured using new data.**
- Move **beyond traditional national statistics aggregation** in OECD, UNESCO, and Eurostat datasets.
 - **Related reports:** *She Figures 2021, Diversity and STEM* (NSF 2023), *Gender in the Global Research Landscape* (Elsevier 2017), *The Research Journey Through a Gender Lens* (Elsevier 2021).
- Our focus: **young women scientists across STEMM disciplines (and trends over time).**
- **Further reading:** Marek Kwiek & Lukasz Szymula, "Young Male and Female Scientists: A Quantitative Exploratory Study of the Changing Demographics of the Global Scientific Workforce", <https://arxiv.org/abs/2211.06319> (revisions in *Quantitative Science Studies*).

12. Flowchart in Our Ongoing Global (OECD) Studies: Stages in Constructing the Sample

(1) Productivity Classes Lifetime; (2) Gender Self-Citation Gap; (3) Aging of the Academic Profession

(Cutting the Scopus publishing universe into slices):

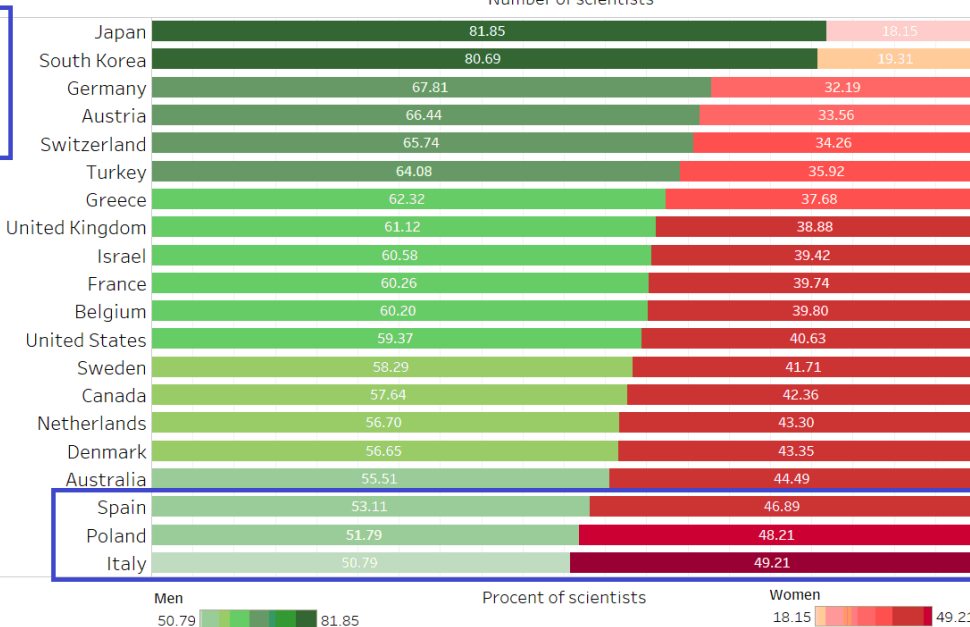
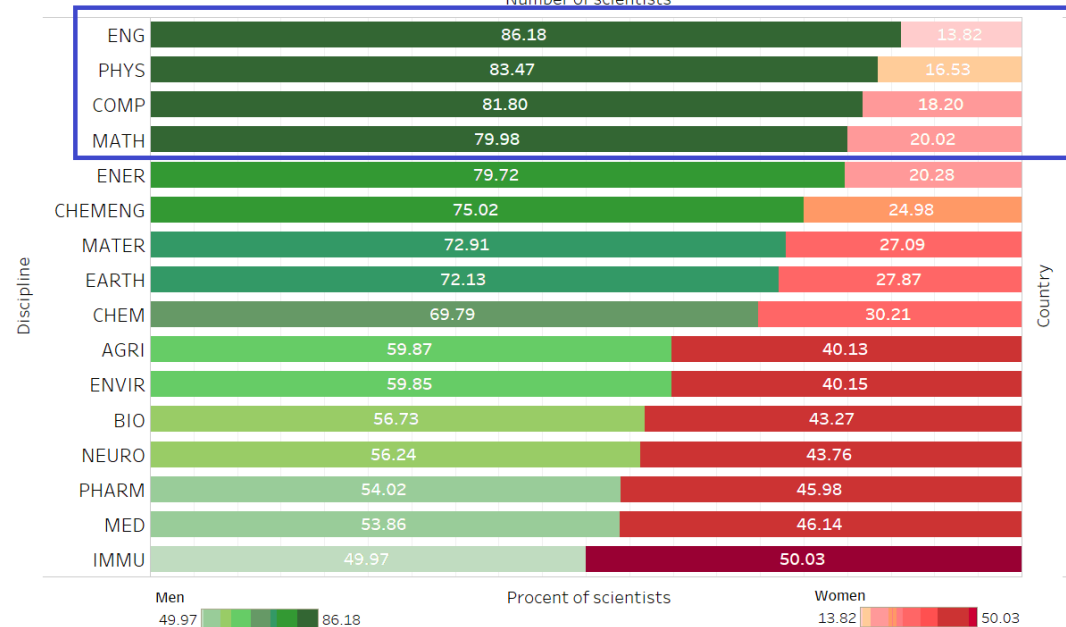
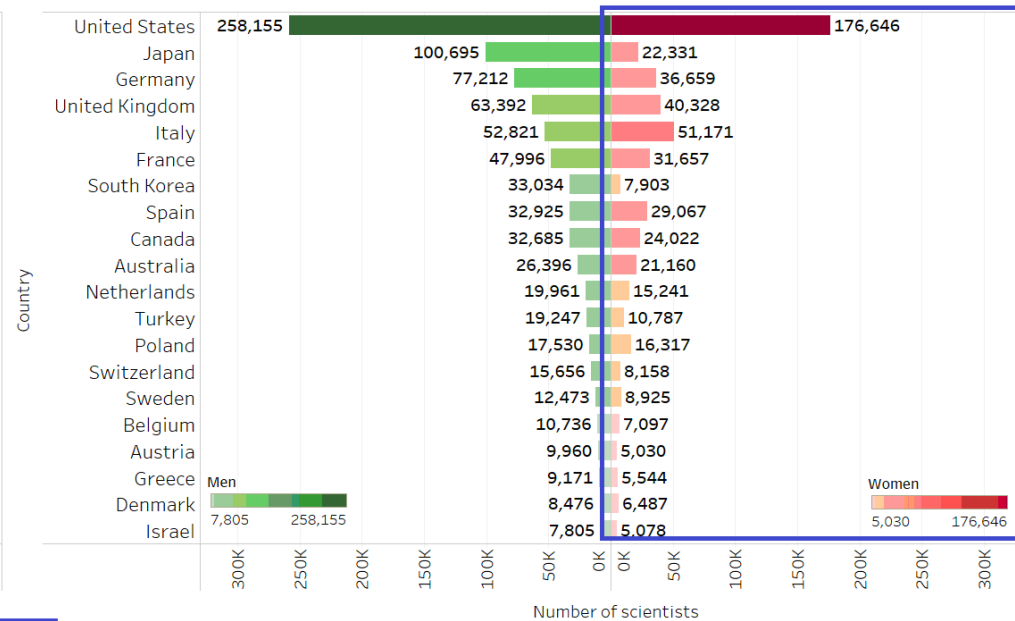
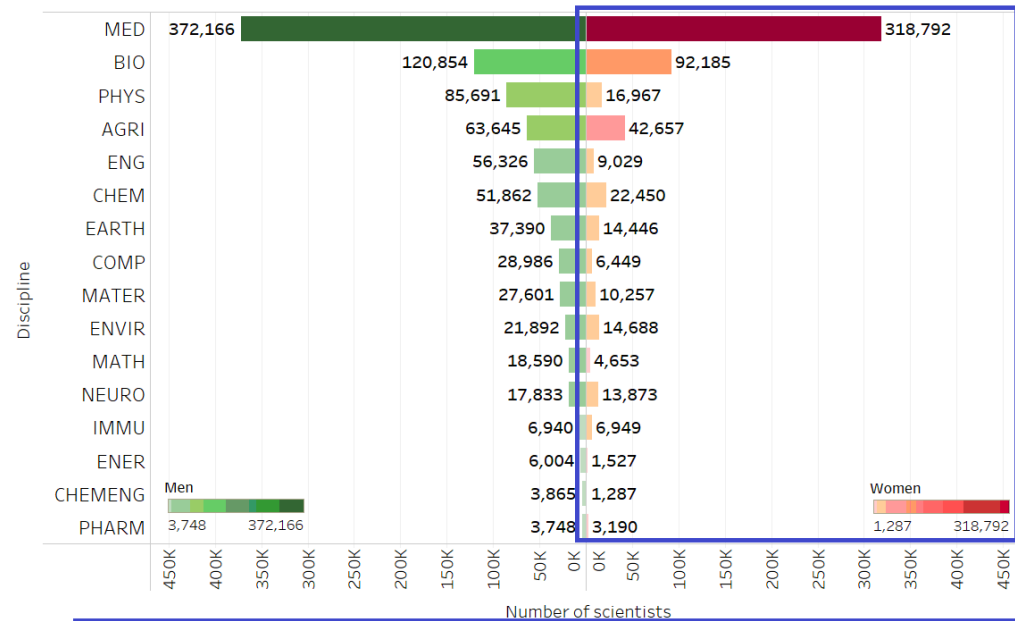
- **Gender** determination
- **Discipline** determination
- Determining the **country** of affiliation
- Determining scientists' **non-occasional status**
- Determining **academic age**
- **43M > 4,3M scientists (with 24M articles).**
- We used **raw data from the Scopus dataset** because our research heavily relied on **author identifiers**. Scopus provides bibliometric data with a **precision of 98.1% and recall of 94.4%** (Baas et al., 2020).
- **Research with Lukasz Szymula from Poznan CPPS Team & ICSR Lab.**



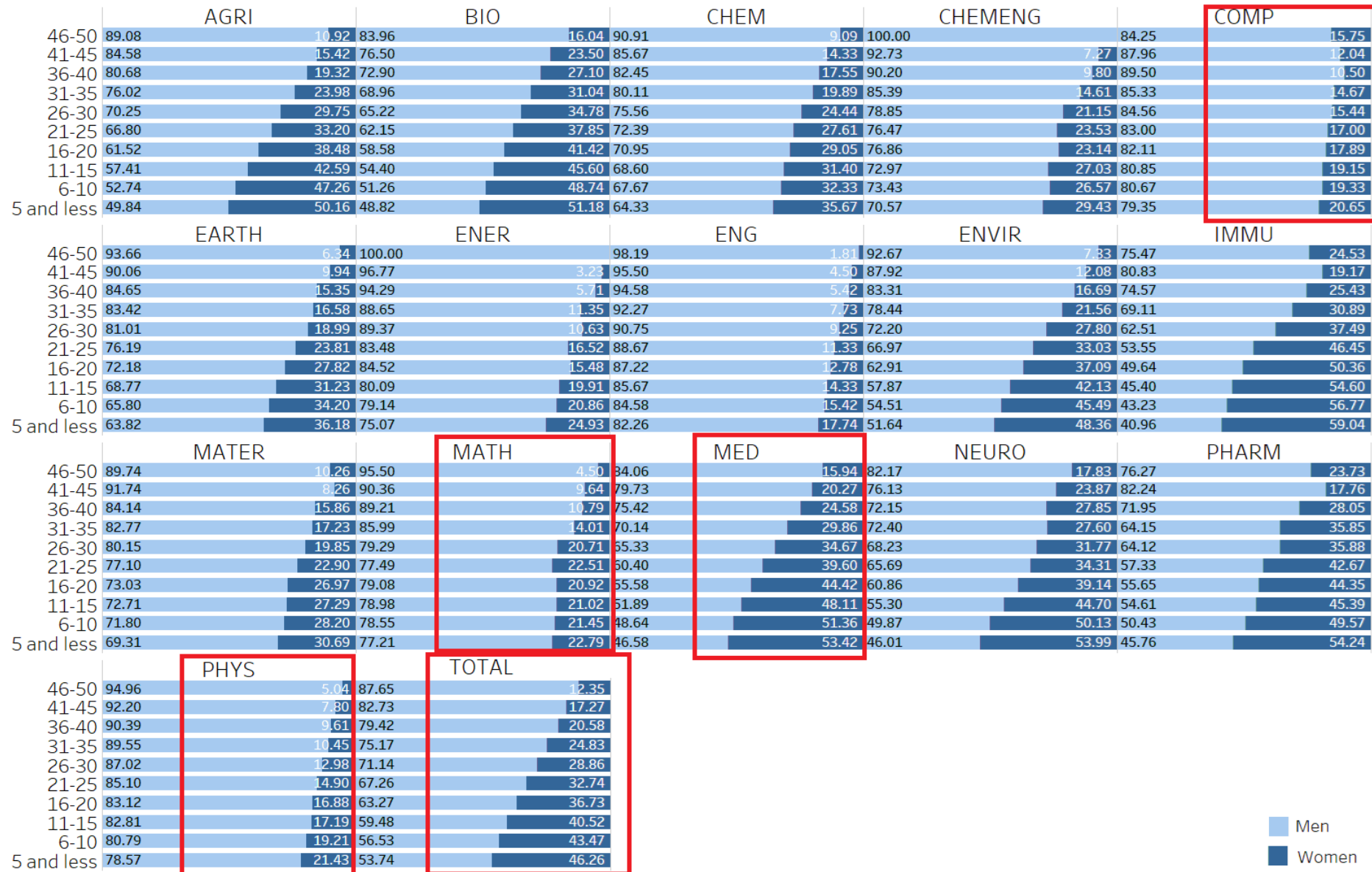
13. Our Methodological Approach at CPPS: Focus on Individual Attributes

- **Gender determination:** based on author's first name, last name, and first country in Scopus dataset.
- **Discipline determination:** modal value of the discipline with the highest number of cited references for each author.
- **Country of affiliation:** modal value of the country with the highest number of occurrences.
- **Nonoccasional status:** scientists with at least three research articles in their output.
- **Academic age:** based on the year of first and last publication, assigned to an age group according to **10 ranges (5 and less years, 6-10... 46-50 years)**.
- **Examining how men/women proportions change over time: a special case of publishing nonoccasional scientists from the OECD area, publishing in Scopus-indexed journals.**
- Probably **the only approach feasible (and cost effective) today** to have a more rigorous, global view of trends.

14. Cross-sectional view (2021). The number of publishing nonoccasional STEMM scientists by discipline and gender (left top) and by OECD country and gender (right top). The share by discipline and gender (left bottom) and by country and gender (right bottom) (in %) (N = 1.5 million)

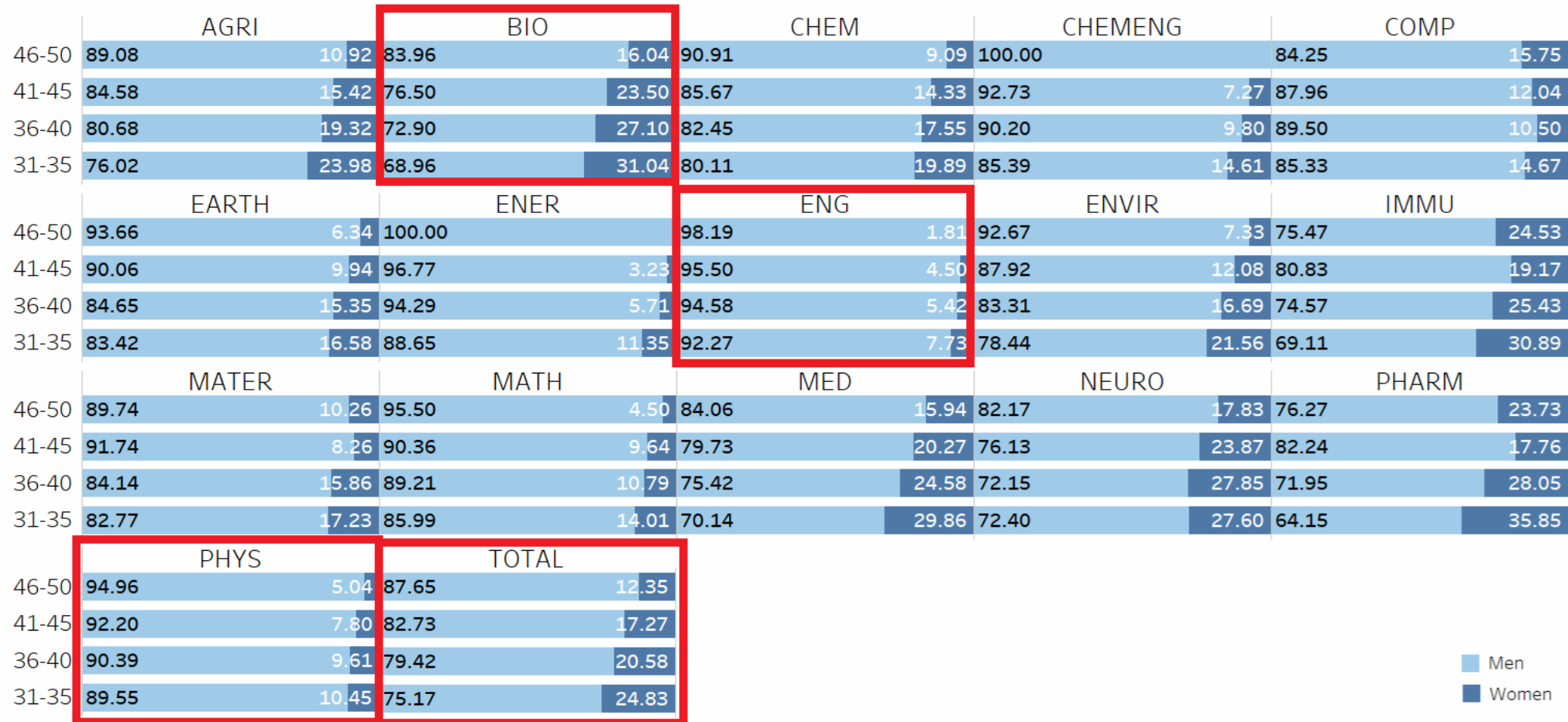


15. Cross-sectional view (2021). Ever-increasing participation of women in younger generations of scientists, with a few exceptions (the Big Four). Horizontal approach: distribution of publishing nonoccasional STEMM OECD scientists by discipline, age group, and gender (row percentages: 100% horizontally) (N = 1,5 million)



16. Cross-sectional view (2021). Zooming in on Old Scientists: Age Cohorts and Women Participation:

More old men than old women in all disciplines. Horizontal approach: zooming in on old scientists only (**academic age 31–50 years**). Distribution of young publishing nonoccasional OECD STEMM scientists by discipline, age group, and gender (row percentages: 100% horizontally) (N = 146,090)



17. Cross-sectional view (2021). Zooming in on Young Scientists: Age Cohorts and Women Participation

More young women than young men in six disciplines. Horizontal approach: zooming on young scientists only (academic age 10 years and less). Distribution of young publishing nonoccasional OECD STEMM scientists by discipline, age group, and gender (row percentages: 100% horizontally) (N = 666,355)

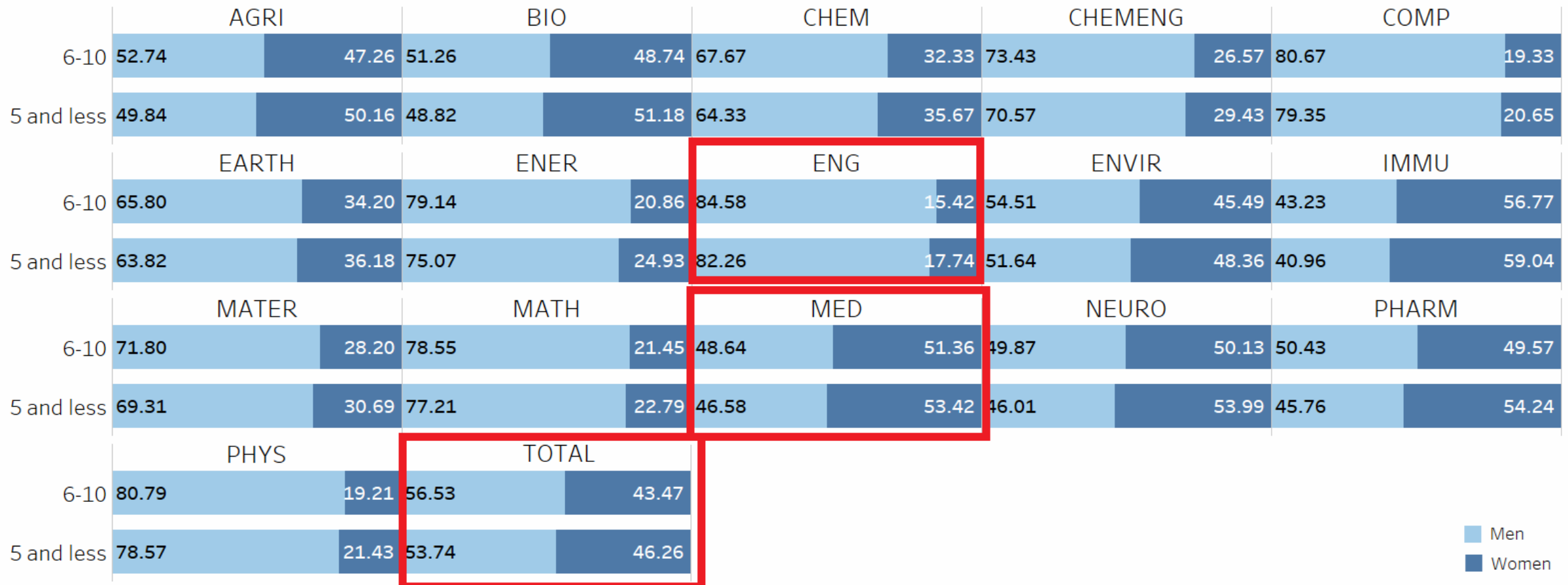


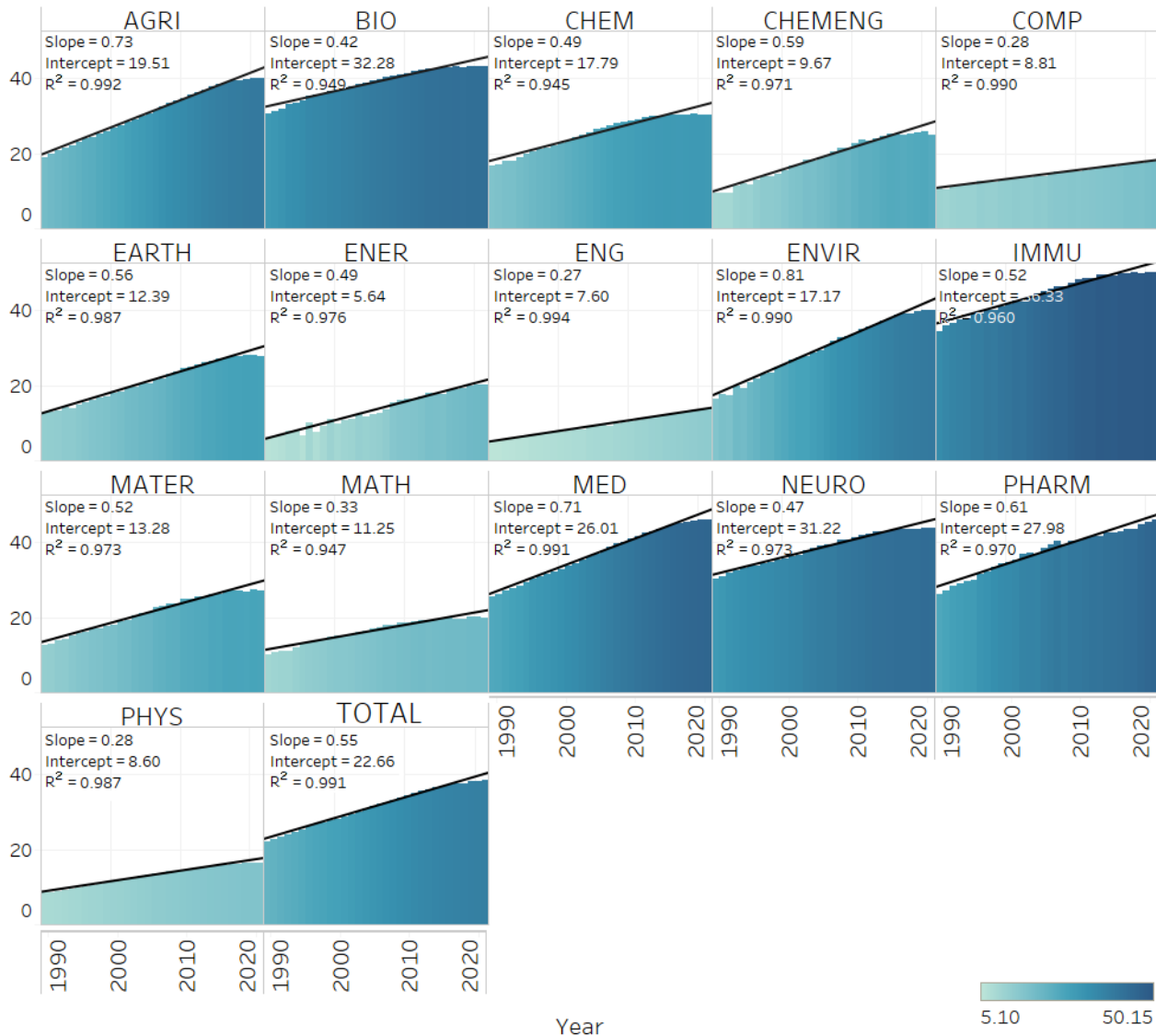
Table 7. Gender- and age-disaggregated data: distribution of non-occasional publishing STEM scientists by selected academic age groups and gender, 2021

Discipline	Gender	5 years and less	6-10 years	Total young cohorts	31-35 years	36-40 years	41-45 years	46-50 years	Total Old cohorts
AGRI	Female	9,714	10,675	20,389	1,238	647	244	77	2,206
	Male	9,652	11,913	21,565	3,925	2,702	1,338	628	8,593
BIO	Female	21,139	23,394	44,533	3,463	1,757	887	315	6,422
	Male	20,161	24,601	44,762	7,692	4,726	2,888	1,649	16,955
CHEM	Female	6,793	5,601	12,394	693	380	176	64	1,313
	Male	12,253	11,721	23,974	2,792	1,785	1,052	640	6,269
CHEMENG	Female	377	330	707	32	15	4		51
	Male	904	912	1,816	187	138	51	28	404
COMP	Female	1,049	1,469	2,518	231	76	26	20	353
	Male	4,030	6,130	10,160	1,344	648	190	107	2,289
EARTH	Female	2,732	3,631	6,363	534	335	118	39	1,026
	Male	4,820	6,985	11,805	2,686	1,848	1,069	576	6,179
ENER	Female	557	456	1,013	16	4	1		21
	Male	1,677	1,730	3,407	125	66	30	10	231
ENG	Female	2,316	2,429	4,745	198	84	19	6	307
	Male	10,739	13,324	24,063	2,362	1,466	403	326	4,557
ENVIR	Female	3,807	3,951	7,758	277	150	40	11	458
	Male	4,065	4,734	8,799	1,008	649	291	139	2,087
IMMU	Female	1,617	1,653	3,270	249	104	51	26	430
	Male	1,122	1,259	2,381	557	305	215	80	1,157
MATER	Female	3,397	2,706	6,103	193	98	20	12	323
	Male	7,670	6,891	14,561	927	520	222	105	1,774
MATH	Female	829	1,006	1,835	193	112	62	19	386
	Male	2,808	3,684	6,492	1,185	926	581	403	3,095
MED	Female	80,100	83,904	170,004	9,217	4,005	1,865	688	15,775
	Male	75,065	79,455	154,520	21,655	12,289	7,338	3,628	44,910
NEURO	Female	3,520	3,880	7,400	369	227	111	51	758
	Male	3,000	3,860	6,860	968	588	354	235	2,145
PHARM	Female	985	756	1,741	128	62	19	14	223
	Male	831	769	1,600	229	159	88	45	521
PHYS	Female	3,817	4,034	7,851	705	396	190	79	1,370
	Male	13,998	16,968	30,966	6,040	3,726	2,247	1,489	13,502
TOTAL	Female	148,749	149,875	298,624	17,736	8,432	3,833	1,421	31,422
	Male	172,795	194,936	367,731	53,682	32,541	18,357	10,088	114,668

18. Decreasing Isolation, a Generational Perspective: from ten-fold difference to five-fold difference.

The presence of women in the four disciplines, young vs. old generations (2021)

19. The trend in the percentage of female scientists by discipline, 1990–2021 (N = 4.3 million)



- Analyzed **women's participation in science over time** to test the claim that female scientists' inflow into science was **differentiated by discipline**.
- Compared **the starting points** and **growth** of women's participation in various disciplines.
- Used slope (a) and intercept (b) to measure **average change** and level of the phenomenon in the zero period.
- Women's participation in some disciplines was **high with strong growth** (MED and PHARM), high with weak growth (BIO), and **low with weak growth** (COMP, ENG, MATH, and PHYS).
- Identified a **cluster of disciplines with low share and weak growth, including math-intensive ones (COMP, ENG, MATH, and PHYS)**.
- Compared **the Big Four disciplines** with **the rest**.

20. Trends in the Percentage of Female Scientists by STEMM Discipline, 1990–2021

Gender Parity (50%/50%), Gender Balance (40%/60%);
 For all vs. for young scientists (for 6 – achieved)?
 Under-representation (below 40%)

Discipline	Slope	Intercept	Time needed to a 1 p.p. change (in years)	Hypothetically, time needed to achieve gender parity (women 50%) in years, and the date
ENVIR	0.81	17.17	1.24	13.5 (2035)
AGRI	0.73	19.51	1.37	16.1 (2038)
MED	0.71	26.01	1.41	40.6 (2062)
PHARM	0.61	27.98	1.64	5.4 (2027)
CHEMENG	0.59	9.67	1.70	6.6 (2028)
EARTH	0.56	12.39	1.78	39.4 (2061)
IMMU	0.52	36.33	1.92	0 (achieved)
MATER	0.52	13.28	1.94	44.4 (2066)
ENER	0.49	5.64	2.02	60.0 (2081)
CHEM	0.49	17.79	2.05	40.6 (2062)
NEURO	0.47	31.22	2.15	13.4 (2035)
BIO	0.42	32.28	2.39	16.1 (2038)
MATH	0.33	11.25	3.02	90.5 (2112)
COMP	0.28	8.81	3.55	112.9 (2134)
PHYS	0.28	8.60	3.55	118.5 (2140)
ENG	0.27	7.60	3.69	133.5 (2155)
TOTAL	0.55	22.66	1.82	-

- Male-female parity in the Big Four disciplines is expected to be reached in about a century from 2021.
- 90.5 years for MATH, 112.9 years for COMP, 118.5 years for PHYS, and 133.5 years for ENG.
- To calculate the date for gender parity in any discipline, the percentage points missing to reach 50% parity in 2021 were multiplied by the time needed to reach 1 p.p. change.
- Predictive analytics was outside the scope of the analysis.
- This exercise is purely hypothetical.

21. Finally (1/4): Global Datasets and Their Limitations

- **Bibliometric sources allow to assess global distributions** by gender, discipline, and age groups cross-sectionally or longitudinally.
- **Individual scientific careers** can be studied by focusing on publications, but it has limitations.
- New knowledge comes at **a methodological price** (needs assessment).
- Global bibliometric datasets (Big Data) require **new algorithmic techniques** for useful information extraction.
- Limitations of bibliometric datasets well known (language and STEM focus, Anglo-Saxon bias, and article-only content).
- Our use of Scopus **to define individual attributes** of the global academic profession shows **new limitations (next slide)**.

22. Finally (2/4): New Limitations to Tackle, Our Own Research

- (1) **Gender determination**: algorithms work better for **some countries** than others; gender-unknown cases were removed from analysis.
- (2) **Discipline determination**: a **commercial journal classification** used as a proxy for nationally-defined disciplines; Scopus publication history used to determine single attribute of discipline, suppressing changes over time.
- (3) **Determining country of affiliation**: single dominant value used, suppressing individual **migration** histories.
- (4) **Determining non-occasional status**: 3-article **threshold arbitrary**; higher threshold would decrease sample, underplaying role of scientists in early careers.
- (5) **Determining academic age**: first publications appear at **different times in different disciplines**; publishing patterns change over time.
- (6) Big datasets require **new statistical assumptions (samples vs. populations)**, different from traditional assumptions) (Big Data analytics).

23. Finally (3/4): National vs. Global Studies

- **Nationally, bibliometric** data can be **merged with administrative** and biographical data
- **National studies** can use **available datasets for a few countries only** (e.g., USA, Norway, Poland, Italy).
- **Globally, biographical information like gender, date of birth, national discipline classifications, and employment history are unavailable.**
- **Global studies:**
 - use **proxies** to examine biological age,
 - **infer** gender,
 - use **proxies** of academic ranks (first publication)
- In global studies, **all scientists registered nationally** vs. **publishing-only scientists indexed by Scopus** (or WoS).
- **Real scientists** (with national IDs) vs. **Scopus Author IDs.**
- **Perfect national** admin and biographical data vs. **inferred data or proxies.**
- **Trade-offs needed to test new ideas!**
- Both global and national studies are useful for **moving beyond national analytical containers** and toward **disciplines (globally).**

24. Summary & Conclusions

- New data **from governments and corporations** can **complement** traditional academic **surveys and interviews** in examining the academic profession.
 - New data require **repurposing** and have their **own limitations**.
 - **Longitudinal data** on academic careers offer **great promise** for discovering **imperceptible patterns**.
 - **Big Data** such as bibliometric datasets **can sharpen insights into** the academic profession.
 - **Remarkable precision and detail**.
 - **Globalization, globalization of science**, global academic profession studies: still a new kid on the block – with a potential to discuss...
-
- **Questions or comments?** Contact kwiekm@amu.edu.pl or [@Marek_Kwiek](https://twitter.com/Marek_Kwiek) on Twitter.
 - **Thank you!**