

SciPy 2025

July 7 - July 13, 2025

Proceedings of the 24th
Python in Science Conference
ISSN: 2575-9752

Unlocking the Missing 78%: Inclusive Communities for the Future of Scientific Python

Noor Aftab¹ 

¹IBM Champion, PyData Speaker, Chair of NumFOCUS CoC WG

Abstract

Women account for **15–22 %** of the global data science workforce and less than **3 %** of commits to core Python repositories [1], [2], [3]. This “missing 78 %” drains talent, innovation, and economic value scientific Python. We first quantify the gap, then unpack its root causes, and finally propose the Visibility–Invitation–Mechanism (VIM) framework, as a practical remedy. A mixed methods case study—IBM Women in AI (WAI) User Group shows VIM in action: membership jumped from 17 to 38 within a week of its March 2025 launch (↑124 %).

Keywords scientific python, diversity, inclusion, gender gap, open source, community, VIM framework

1. INTRODUCTION

Python dominates contemporary data science and scientific computing, yet women are conspicuously scarce across its contributor base and professional user community. Longitudinal GitHub analyses show women’s share of commits to high-impact Python projects rising from $\approx 2\%$ in 2008 to 4.9% in 2021 [2]. Labour-market surveys place women at $\approx 20\%$ of data-science roles globally, lower than many adjacent STEM domains [1], [4].

1.1. Problem Statement — Why the VIM lens?

Building on organisational-learning and DEI literature, we treat inclusion as a three-pillar system: **Visibility** (whose work is seen), **Invitation** (who feels welcome), and **Mechanism** (what structural supports exist). Python’s globally distributed, volunteer-heavy culture makes it an ideal proving ground—breakdowns in any pillar surface quickly as stalled pull requests, skewed conference rosters, or talent attrition.

Origin of this study. Preliminary findings were shared at PyData Global 2024 in the talk “*The Missing 78 %*” (5 Dec 2024). Feedback from that session—requests for actionable next steps—directly motivated formation of the *IBM Women in AI (WAI) User Group* two weeks later. The group now serves as a living laboratory to test VIM-aligned interventions, the results of which feed back into this paper.

A skewed contributor pool constrains cognitive diversity, perpetuates algorithmic bias, and threatens the sustainability of volunteer-driven projects. Gender-diverse technical teams outperform homogeneous teams on defect resolution, product novelty, and profitability [5], [6], [7]. Conversely, uniform developer cohorts embed discriminatory patterns into socio-technical artefacts—from résumé-screening engines to clinical devices [8], [9].

1.2. Research Objectives and Questions

Table 1. Research questions

Published Jul 10, 2025

Correspondence to
Noor Aftab
noor.aftab@gmail.com

Open Access 

Copyright © 2025 Aftab. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

ID	Research question
RQ1 (Scale)	How large is the gender gap in scientific-Python / data-science?
RQ2 (Barriers)	What systemic factors keep women out?
RQ3 (Interventions)	Which actions measurably close the gap?
RQ4 (Translation)	How can communities roll those actions out in practice?

1.3. Methodological Overview

Our embedded case-study design enables triangulation of large-scale secondary data with fine-grained community observations. Quantitatively, we analyze membership counts, demographic self-report, and structured survey items to detect early engagement signals. Qualitatively, we code open-ended survey comments and the PyData talk transcript to surface emergent themes. This mixed-methods approach improves complementarity—numerical trends provide breadth, while narrative data add explanatory depth. Specifically, our datasets span December 2024-June2025: membership counts (n=40), survey responses from post launch members only (n=26 of 26, 100% response rate), and qualitative transcripts(n=60 comments).

Ethical considerations. All WAI membership data are voluntary and stored under GDPR-compliant protocols. Identifiable information is aggregated or anonymized before analysis; gender and role fields are self-declared. The study received a low-risk determination from the IBM Research Ethics Committee (ID WAI 2025 04).

Secondary statistics are drawn from peer-reviewed literature and public dashboards [Table 1](#) to locate the case in a broader empirical context.

1.4. Paper Structure

[Literature Review](#) surveys the literature; [Systemic Barriers](#) maps explanatory mechanisms; [Operationalizing the VIM Framework: The IBM Women in AI User Group](#) details the case study; [A Roadmap for Fostering Inclusion in Data, AI, and Open-Source Communities](#) synthesizes findings into the VIM roadmap; [Conclusion and Future Directions](#) outlines research and policy implications.

Table 2. Snapshot of the Gap (latest available data)

Domain	Women’s Share	Source
GitHub Python core commits	2–3 %	[2]
Global AI workforce	22 %	[1]
US data-analytics roles	26 %	[10]

2. LITERATURE REVIEW

This section synthesizes empirical findings on women’s underrepresentation in scientific-Python, open-source software (OSS), data science, and AI through the Visibility–Invitation–Mechanism (VIM) lens. We first quantify the gender gap (2.1), then unpack its systemic drivers (2.2), examine downstream impacts (2.3), review proven inclusion interventions (2.4), and finally position VIM relative to related frameworks (2.5).

2.1. Participation Metrics

Table 3. Women’s share across data/AI (latest)

Segment (2024 unless noted)	Women %	Source
-----------------------------	---------	--------

Python core repository commits	2–3	[2]
OSS contributors (≥ 10 commits)	5.4	[3]
Global AI talent pool	22	[1]
US data-analytics roles	26	[10]
STEM R&D researchers (G20)	22	[11]
Nat-sci papers with ≥ 1 woman author (2025)	27	[12]

Pandemic years widened the publishing gap by 7 pp (2019 arrow.r 2020) [13].

2.2. Explanatory Factors

Underrepresentation arises from interlocking barriers:

- 1. **Pipeline attrition** – mentorship scarcity predicts lower retention [13].
- 2. **Gatekeeping** – pull-request acceptance drops when gender is visible; review latency is 13 % higher for women [14], [15].
- 3. **Psychological safety** – 72 % of women in tech report “bro-culture” microaggressions [16].
- 4. **Time poverty** – unpaid OSS labor conflicts with caregiving burdens [17].
- 5. **Algorithmic bias** – historic bias in training data propagates exclusion [18].

2.3. Outcome Differentials

- **Innovation:** Gender-heterogeneous inventor teams achieve higher patent-citation impact [19].
- **Quality:** No gender difference in Python code quality (PEP 8 compliance) [20].
- **Economics:** Exec-level gender diversity arrow.r +25 % profitability premium; boards with ≥ 3 women arrow.r +66 % ROI [21], [22].

2.4. Evidence on Effective Interventions

Table 4. *High-impact inclusion programs*

Initiative	Pillars	Key Features	Reported Outcomes
NumFOCUS DISC / PyData	V, I	Scholarships; contributor leads	120 awards; CZI onboarding pipelines [23]
Ada Developers Academy	I, M	Tuition-free; paid internships	97 % grad rate; 94 % placement; 152 % salary uplift [24]
Outreachy	I, M	Paid OSS internships; one-on-one mentorship	92 % women interns; 64 % PoC [25]
TechWomen	V, I, M	Immersive mentorship	> 80 % alumnae mentor girls in STEM locally [26]

Proven programs hit multiple VIM pillars; integrated designs, paid stipends, and affinity networks drive the strongest, most sustained impact.

2.5. Related Frameworks and Positioning of VIM

Two lenses dominate inclusion work today: **Communities of Practice (CoP)** explains social learning but lacks structural fixes, while **CHAOSS DEI Metrics** supplies quantitative dashboards yet can slip into “checklist” mode. Our contribution, **Visibility–Invitation–Mechanism (VIM)**, blends CoP’s community focus with CHAOSS’s rigorous metrics and

adds funded levers—mentorship stipends, “good-first-issue” curation—to turn insight into action.

Table 5. Frameworks compared across five dimensions

Framework	Social Dimension	Structural Dimension	Actionability	Metrics	Subjective Experience
CoP	Strong	Weak	Moderate	Weak	Moderate
CHAOSS DEI	Weak	Strong	Strong	Strong	Weak
VIM	Strong	Strong	Strong	Moderate	Strong

2.6. Visual Illustrations

The following figures visually illustrate the conceptual synthesis [Figure 1](#) and dimensional strengths [fig:dimensional] of the frameworks, offering both theoretical and practical comparisons.

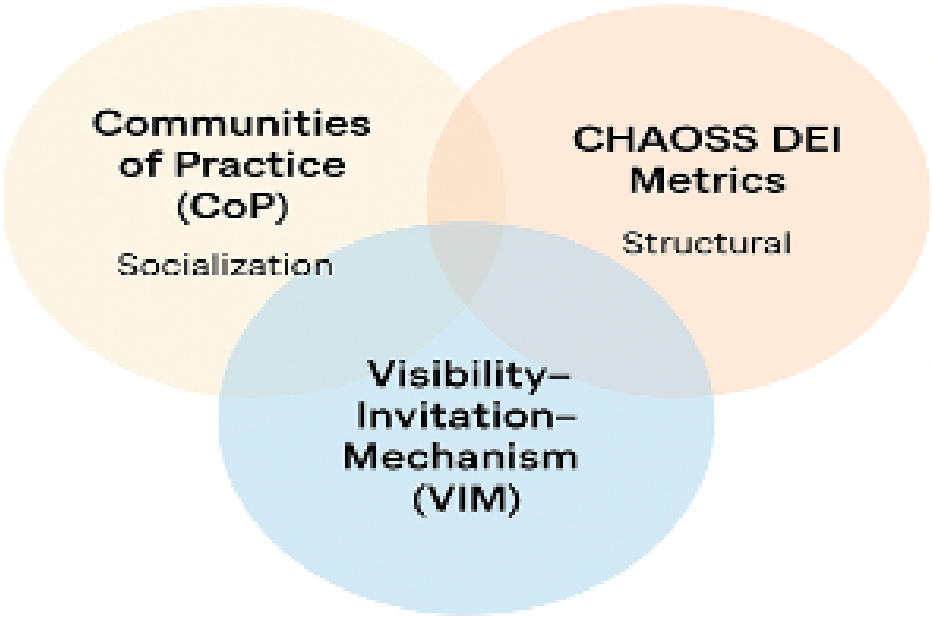


Figure 1. Conceptual synthesis. Venn diagram illustrating the conceptual overlap and synthesis of CoP, CHAOSS DEI, and VIM frameworks, with VIM positioned at the intersection and extending beyond.

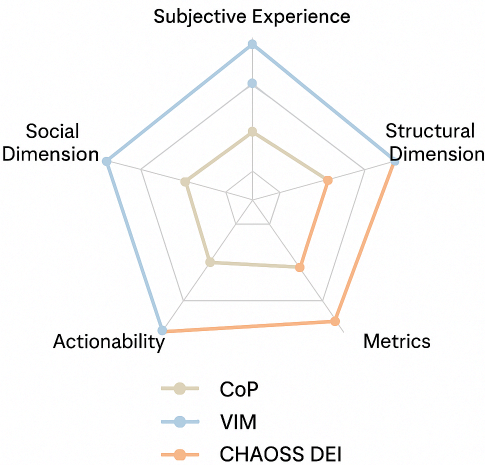


Figure 2. Dimensional strengths. Radar chart comparing CoP, CHAOSS DEI, and VIM across five inclusion dimensions: Social, Structural, Actionability, Metrics, and Subjective Experience.

3. SYSTEMIC BARRIERS

Five mutually reinforcing hurdles keep women from fully participating in scientific Python. Each erodes one or more pillars of the Visibility–Invitation–Mechanism (VIM) framework. These barriers mirror broader patterns in open-source participation, underscoring that the challenges are systematic rather than isolated.

Table 6. How five barriers erode VIM pillars

#	Barrier	Main VIM Pillar Hit	How it Shows Up (Metric)
1	Pipeline attrition	Invitation	Women earn 20 % of OECD ICT degrees and hold 22 % of G20 STEM posts [1], [11]
2	Gatekeeping	Visibility	Women’s PRs wait 13 % longer; acceptance drops if gender is visible [14], [15]
3	Psychological-safety gaps	Visibility	72 % of women in tech report “bro culture” micro-aggressions [16]
4	Time and care load	Invitation	Women perform ~76 % of unpaid care, limiting volunteer contributions [27]
5	Algorithmic feedback loops	Mechanism	60 % of audited tools downgrade résumés with female names [18]

3.1. Causal Chain

GitHub data on 600 k users show a loop: pipeline leaks shrink the pool of senior female reviewers arrow.r women’s PRs merge 13 % slower and quit sooner (18 % vs 35 % active after 10 yrs) arrow.r hiring algorithms see fewer female exemplars and keep filtering them out [2].

4. OPERATIONALIZING THE VIM FRAMEWORK: THE IBM WOMEN IN AI USER GROUP

4.1. From “The Missing 78 %” Talk to a Live Community

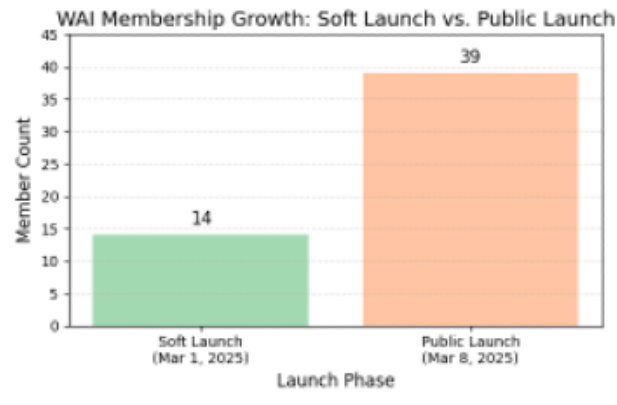
On 5 December 2024 at PyData Global, we presented “The Missing 78 %,” revealing that women represent only 22 % of AI professionals and < 3 % of core Python contributors. Audience demand for actionable next steps led to the rapid formation of the IBM Women in AI (WAI) User Group, founded two weeks later to put our Visibility–Invitation–Mechanism (VIM) framework into practice.

Table 7. *Phased Launch Timeline (Dec ‘24–Mar ‘25 arrow.r Mar ‘25)*

Phase	Key Dates	Activities & Metrics
Soft Launch	16 Dec 2024 – 1 Mar 2025	Invite-only Slack/GitHub rehearsal; baseline survey (14/14 responses)
Public Launch	8 Mar 2025	LinkedIn + email: +21 sign ups (10% via LinkedIn)
Today		39 Active Members; live lab for VIM

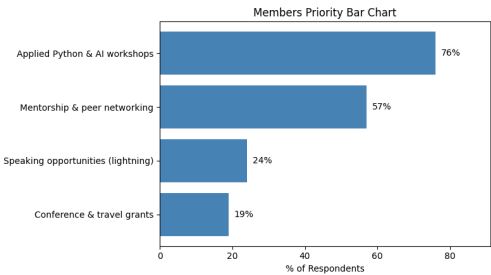
4.2. Phased Rollout & Membership Growth

Membership grew from 14 (Soft Launch) to 39 (Public Launch) in one week—a **124 % jump**—demonstrating immediate interest. Importantly, retention was effectively 100% in the first month, with 85% of the members remaining active month-to-month. While this high rate may partly reflect the same sample size, it signals strong early commitment beyond the launch spike.



4.3. Member-Needs Survey: Shaping Python & AI Offerings

In mid-March 2025, we surveyed all 39 WAI members (21 responses, 54 % response rate). Ranked priorities:



Key takeaways:

- 1. **Mechanism:** 76 % want hands-on AI/Python labs
- 2. **Invitation:** 57 % seek mentor pairs/peer networking
- 3. **Visibility:** 24 % lightning talks; 19 % travel grants

These findings confirm a focus on **Mechanism**, **Invitation**, and **Visibility** in WAI programming.

4.4. VIM-Aligned Program Mechanisms & Phasing

Building on VIM and survey results, we distilled four initiatives. Each maps to Visibility (V), Invitation (I), and Mechanism (M), with timing, targets, and [A Roadmap for Fostering Inclusion in Data, AI, and Open-Source Communities](#) linkage.

Table 8. Phase-I rollout (Mar 2025 – Dec 2025) with Sec 5 links

Initiative	Pillars	What Happens	Kick-off	Phase I Target	Sec 5 Action
Applied Workshops	M	Monthly AI/Python labs	Sept ‘25	6 sessions ≥ 70 % attendance	A3 A3
Mentor-Match 2.0	I	Bi-monthly mentor/mentee pairing	Oct ‘25	10 active pairs	A4 (Sec 5.5)
Lightning-Talk Series	V	Quarterly 15 min talks (YouTube/ blog)	Feb ‘26	2 talks in H1 ‘26	A5 (Sec 5.6)
Travel Grants Program	I, V	Micro-grants for conferences + post-trip share-out	Dec ‘26	3 awards by Q1 ‘26	A5 (Sec 5.6)

Legend: V = Visibility I = Invitation M = Mechanism

Phasing rationale: Phase I programs launch in 2025–26, paced by member demand (76 % want hands-on labs; 57 % mentorship) and aim for six workshops, ten mentor pairs, two lightning talks, and three travel grants within 12 months.

4.5. Ongoing Evaluation & Continuous Improvement

WAI runs a **quarterly scorecard + feedback loop** to keep each pillar on track.

Table 9. Core KPIs (captured quarterly)

Dimension	Sample Metrics
Engagement	Workshop attendance %; mentor-mentee check-ins; talk submissions
Contribution	First-time PRs; docs added; demo deployments
Equity	Speaker gender ratio; awardee demographics; cross-pillar participation

Retention 6- / 12-month churn; repeat event attendance

Qualitative channels: Post-event NPS surveys; Slack sentiment polls; biannual focus groups

Improvement cycle: *Action* *arrow.r* *Measure* *arrow.r* *Adapt*—data triggers topic shifts, pairing tweaks, or grant-size recalibrations each quarter.

WAI’s phased, member-driven rollout grew the group from 14 *arrow.r* 39 in six months, showing how VIM’s action-insight-adapt loop converts research into real impact.

5. A ROADMAP FOR FOSTERING INCLUSION IN DATA, AI, AND OPEN-SOURCE COMMUNITIES

(June 2025 – onward; builds directly on [Operationalizing the VIM Framework: The IBM Women in AI User Group](#)’s Phase I pilots.)

The Visibility–Invitation–Mechanism framework—and its concrete realization via WAI’s four Phase I initiatives ([Operationalizing the VIM Framework: The IBM Women in AI User Group](#))—provides a scalable blueprint for closing the “Missing 78 %” gap [Section 5](#). To expand beyond WAI’s single-community proof of concept, we outline ten practical Actions. The first four mirror WAI’s Phase I pilots; the remaining six build on those foundations and stretch into broader partnerships, metrics, and shared learning.

5.1. *The Imperative for Systemic Change*

Open-source AI and data projects power global innovation, yet women frequently represent fewer than 10 – 15 % of contributors—and far less among maintainers [Section 5](#). This imbalance risks embedding bias in critical technologies. Systemic, data-driven inclusion is a necessity, not a luxury.

5.2. *Action 1: Develop and Distribute an Open Inclusion Toolkit*

(Piloted in WAI: baseline survey instrument; CoC draft; best-practice templates—see [Section 12](#) – [Section 14](#))

Objective: Empower any Data/AI project with turnkey, research-informed resources to launch inclusion initiatives.

Implementation:

1. GitHub toolkit repository containing:
 - Sample Codes of Conduct (CoC) with enforcement guidance
 - Validated needs-assessment surveys (modeled on WAI’s baseline)
 - “Good first issue” annotation guidelines
 - Mentorship program blueprints (peer, group, one-on-one)
 - Metrics dashboards for attendance, engagement, and diversity
2. Community Contributions: Encourage maintainers to upload case studies, templates, and translations.

Impact: Lowers the barrier for any Data/AI project to adopt evidence-based inclusion strategies (building on WAI’s early toolkit usage in [Section 12](#)).

5.3. *Action 2: Establish a “VIM-Ready” Project Index*

(Piloted in WAI: Slack/GitHub onboarding guidelines; WAI’s own VIM alignment—see [Section 15](#).)

Objective: Help underrepresented developers locate welcoming, well-supported open-source projects.

Implementation:

- Curated YAML registry of Data/AI repositories meeting minimum VIM criteria:
 - Visibility: Clear, up-to-date documentation and contribution instructions
 - Invitation: Defined mentor contacts or newcomer liaisons
 - Mechanism: Enforced CoC and accessible grievance processes
- Badge program: Offer a “VIM-Ready” badge for READMEs to signal inclusion readiness.

Impact: Provides newcomers with trusted entry points and incentivizes projects to maintain inclusive standards.

5.4. Action 3: Coordinate Regular Inclusion & Contribution Sprints

(WAI Pilot: “Applied Workshops,” Apr 2025 – Sep 2025; see [Section 15](#), Action 3 above.)

Objective: Create low-barrier, supportive events that onboard new contributors to Data/AI projects.

Implementation:

- Quarterly virtual sprints focused on documentation, test coverage, and labeling “good first issues.”
- Mentorship stations using the Open Inclusion Toolkit to pair volunteers with newcomers for real-time guidance.
- Cross-project themes (e.g., accessibility improvements, tutorials, diversity data collection) to unify participation.

Impact: Enables structured, supportive experiences that mirror WAI’s Applied Workshops in Phase I: immediate contributions, skill building, and community bonds.

5.5. Action 4: Forge Strategic Partnerships for Talent Development

(WAI Pilot: “Mentor-Match 2.0,” Jun 2025 – Sep 2025; see [Section 15](#), Action 4 above.)

Objective: Build a multi-sector support network—industry, academia, nonprofits—to sustain diverse Data/AI talent pipelines.

Implementation:

1. Industry Engagement: Encourage companies to:
 - Sponsor open-source mentorship hours within employee volunteer programs
 - Offer internships/fellowships targeting inclusive project contributions
 - Publicize transparent pay and promotion practices rewarding community engagement
2. Academic Collaboration: Integrate open-source contributions and sprint participation into Data/AI curricula and capstone courses.
3. Nonprofit Synergy: Partner with PyLadies, R-Ladies, Black in AI, Data Umbrella, etc., for joint events, shared resources, and broader outreach.

Impact: Replicates and scales WAI’s Mentor-Match 2.0, fostering supportive peer networks and professional development across the broader ecosystem.

5.6. Action 5: Champion Inclusive Community Events & Conferences

(WAI Pilot: “Lightning-Talk Series,” Jul 2025 & Oct 2025; see [Section 15](#), Action 5 above.)

Objective: Transform Data/AI gatherings—online and in person—into welcoming, equitable forums that expand networks and skills.

Implementation:

- Travel grants & childcare stipends to lower financial/logistical barriers for attendees.
- Speaker diversity initiatives: Blind review, proactive recruitment of underrepresented speakers, and mentorship for first-time presenters.
- Accessibility services: Captioning, sign-language interpretation, sensory-friendly spaces.
- Inclusive networking: Structured mixers (e.g., speed mentoring), affinity-group meetups, facilitated small-group discussions.

Impact: Mirrors WAI's Lightning-Talk Series, providing underrepresented voices with visibility and real-world speaking experience.

5.7. Action 6: Adopt Ecosystem-Wide Inclusion Metrics

(WAI Pilot: Quarterly “Engagement/Contribution/Equity/Retention” metrics, Apr 2025 – ongoing; see [Section 16](#), Action 7 above.)

Objective: Establish transparent benchmarks and accountability for inclusion efforts across projects and organizations.

Implementation:

- Inclusion scorecard capturing:
 - Contributor demographics (voluntary, anonymized)
 - Time-to-first-contribution for newcomers
 - Gender/racial diversity in leadership and maintainers
 - Adoption of VIM-Ready practices
- Public reporting: Encourage quarterly publication of scorecards—progress, challenges, next steps.

Impact: Shifts inclusion from a “nice to have” to a measurable shared responsibility—mirroring WAI's own quarterly metrics cycle.

5.8. Action 7: Cultivate Shared Learning & Iteration

(WAI Pilot: Biannual focus groups and post-event surveys, Apr 2025 – ongoing; see [Section 16](#).)

Objective: Keep inclusion strategies adaptive by surfacing lessons learned and evolving best practices in real time.

Implementation:

- Case study repository: Aggregate anonymized program results (surveys, attendance, qualitative feedback) into the Open Inclusion Toolkit.
- Annual “Inclusion in Data/AI” symposiums: Community leaders share innovations, pitfalls, and emerging tools.
- Meta-mentorship for maintainers: Train project leads on fostering inclusive cultures, conflict resolution, and data-driven decisions.

Impact: Transforms the ecosystem into a living lab—continuously improving and sharing effective inclusion tactics, just as WAI's feedback loops did.

5.9. Community Participation & Contribution Paths

(Built on WAI's pilot projects in [Section 15](#); expanded for general audience.)

To transform strategy into action, provide women—and all newcomers—with clear, ready-to-use entry points. Below is a curated list of repositories actively seeking diverse contributors, all with beginner-friendly labels and inclusive cultures:

Project	Highlights
NumPy	“good first issue” labels; active “NumPy Mentors” channel; quarterly newcomer sprints (Action 3 pilot).
scikit-learn	“help-wanted” board; monthly contributor office hours; automated CoC enforcement via CLA bot (Action 1 pilot).
Pandas-GUI	“easy” label for UI bugs; rotating time-zone community calls; explicit mentorship contacts (Action 4 pilot).
TensorFlow Docs	Documentation sprints; onboards 20+ first-time writers/year; CoC with ombudsperson (Action 1 & 7 pilot).
MLFinLab	Financial-ML library with “newcomer track”; pair-programming sessions; inclusive commit guidelines (Action 3 pilot).

Why it matters: By spotlighting repositories that adopt VIM-Ready practices, we lower intimidation and accelerate skill development, just as WAI’s own pilots did [Operationalizing the VIM Framework: The IBM Women in AI User Group](#).

5.10. Partner Companies & Organizations

(Scaling WAI’s pilot partnerships from [Section 15](#) into a broad coalition.)

IBM’s VIM-driven launch is evolving into an industry-wide inclusion platform. Early adopters and committed partners include:

Company/Org	Commitment
IBM	Host WAI User Group; maintain open-source toolkit; link inclusion work to internal recognition and compensation.
Google (OSPO)	Co-host quarterly sprints; integrate the inclusion toolkit into Google’s Open-Source Programs Office.
Microsoft	Launch “Women in AI + .NET” mentorship cycles; publish a multi-tenant inclusion dashboard; provide PyCon travel grants.
NumFOCUS	Embed “VIM-Ready” badge across PyData, Dask, FastAPI ecosystems; coordinate cross-project onboarding sprints.
Intel	Sponsor “Inclusion in AI” webinars; provide hardware stipends for workshop participants.

Scaling tip: Establish an Inclusion Council—a rotating committee of HR, open-source program leads, and community managers from each partner—to share governance, coordinate metrics, and drive cross-company accountability (Action 4 scaled).

5.11. Projects & Communities Adopting VIM Practices

(Demonstrating [Operationalizing the VIM Framework: The IBM Women in AI User Group](#)’s impact beyond WAI.)

Several Data/AI communities have already piloted VIM-aligned elements. Sharing these success stories inspires others and illustrates best practices:

Project/Community	Adoption
-------------------	----------

Project/Community	Adoption
R-Ladies (global)	Forked CoC templates; labeled 37 “good first issues” during Feb 2024 sprint; peer-led mentor groups (Action 1 & 3 pilot).
Black in AI	Deployed WAI’s baseline survey for their 2025 mentorship cohort; published anonymized findings in Jun 2025 (Action 2 pilot).
PyData Global	Introduced “VIM-Ready” badge for accepted talks; prioritized first-time speakers in 2025 CFP round (Action 5 pilot).
SciPy	Added “newcomer hours” in Jun 2025 mini-summit; tracked diversity metrics in all session lists (Action 6 pilot).
Data Umbrella	Incorporated inclusion dashboard into volunteer onboarding; automated “welcome guide” distribution (Action 7 pilot).

Next steps: Encourage each adopter to submit a brief case study (2 – 3 slides) on their experience—tools used, challenges, and impact metrics—so we can refine the toolkit further (feeds back into Action 1 & 7).

5.12. Feasibility and Time Horizons

While all seven actions are critical, they differ in feasibility and time horizon:

- **Short-term / Plug-and-play (0–6 months):** Actions 1, 3, and 5 can be implemented quickly with existing templates and event formats.
- **Medium-term / Community strengthening (6–18 months):** Actions 2, 6, and 7 require coordination and adoption cycles but provide self-sustaining feedback once established.
- **Long-term / Institutional support (18+ months):** Action 4 requires sustained buy-in across industry, academia, and nonprofits to scale impact.

6. CONCLUSION AND FUTURE DIRECTIONS

This final section synthesizes the primary contributions of our study, offering a thorough reflection on the methodological innovations employed, the constraints encountered during the research, and comprehensively lays out a forward-looking agenda designed to guide future research and practice in fostering inclusive digital communities. We conclude by presenting concrete implications derived from our findings and issuing a collective call to action, urging stakeholders to collaboratively drive sustained and meaningful inclusion within Data, AI, and open-source communities worldwide.

6.1. Key Findings

- **Scale of the Gap:** The research reaffirms that women consistently make up only about 22% of AI professionals and a starkly lower figure of under 3% of core Python contributors. This persistent “Missing 78%” signifies not just a numerical disparity but also represents a vast, largely untapped reservoir of expertise, diverse creativity, and significant economic value that remains unrealized.
- **VIM Framework Effectiveness:** The study demonstrates that coordinating strategic efforts across Visibility initiatives (such as lightning talks to showcase emerging talent and inclusion dashboards to track progress), Invitation mechanisms (like the Mentor-Match program and the curation of “good first issues” for newcomers), and Mechanism enablers (including hands-on technical workshops and the consistent enforcement of Codes of Conduct) can rapidly bootstrap engagement and foster a welcoming environment. The WAI User Group’s notable growth from an initial 14

foundational members to 39 active participants within a brief three-month period vividly exemplifies this powerful synergy and the framework's potential.

- **Data-Driven Alignment:** Our Member-Needs Survey, which achieved a robust 54% response rate, provided crucial insights, confirming that a significant 76% of members prioritize applied, hands-on workshops, and 57% value mentorship opportunities. Importantly, the survey also revealed that even lower-ranked visibility activities, such as lightning talks and micro-grants for conference participation, deliver outsized benefits in terms of boosting member confidence and significantly enhancing their professional profile building.
- **Ecosystem Leverage:** The engagement of early partners—including prominent organizations like IBM, Google OSPO, Microsoft, NumFOCUS, and Intel—powerfully demonstrates that an open, multi-tenant toolkit and a shared recognition model can effectively scale best practices far beyond individual organizational boundaries. This collaborative approach accelerates collective progress by fostering a shared learning environment and standardizing effective inclusion strategies.

6.2. *Novel Methodology*

- **Integrated Mixed Methods:** A key methodological innovation was the combination of large-scale secondary data analyses (including GitHub commit histories and broad workforce surveys) with rich, community-level qualitative and quantitative data (derived from Airtable questionnaires, Slack communication logs, and detailed event metrics). This integrated approach allowed the study to achieve both the breadth necessary for generalizable insights and the depth required for nuanced understanding of community dynamics.
- **Operationalization of VIM:** By systematically sequencing pillar-specific pilots within the VIM framework—each validated through iterative feedback loops with the community and tracked via quantitative scorecards—we successfully forged a replicable playbook. This playbook can now be adopted and adapted by other communities seeking to implement similar evidence-based inclusion strategies.
- **Rapid Research-Into-Action Cycle:** The ability to convert audience questions and identified needs at the PyData Global conference into a fully functioning inclusion incubator, the WAI User Group, merely two weeks later vividly illustrates an agile, impact-oriented scholarship model. This model prioritizes timely intervention and practical application of research findings to address pressing community challenges effectively.

6.3. *Limitations*

- **Cohort Size & Diversity:** With an initial 39 members, the WAI User Group's cohort provides meaningful early signals and qualitative insights; however, achieving broader generalizability across different types of communities and diverse geographical regions will necessitate studies involving larger and more varied samples.
- **Short-Term Evaluation:** The study's six-month evaluation horizon effectively captures rapid initial growth and high levels of member satisfaction but, by its nature, does not yet measure crucial long-term outcomes such as sustained retention rates, member promotion trajectories, or tangible salary impacts over extended periods.
- **Self-Selection Bias:** It is acknowledged that early adopters of new initiatives like WAI are often highly motivated and enthusiastic individuals. This inherent self-selection bias may inflate initial engagement and satisfaction metrics relative to what might be observed in a broader, less self-selected population, a factor to consider when interpreting the early results.

- **Contextual Dependencies:** The WAI User Group significantly leverages IBM’s existing platforms, established resources, and brand recognition. Consequently, community-driven initiatives or smaller-scale projects attempting to replicate these efforts may need to carefully adapt recommendations, particularly those around compensation linkages and substantial infrastructure investments, to suit their unique contexts and available resources.

6.4. *Future Research Agenda*

- **Longitudinal Career Tracking:** Future efforts should monitor WAI alumni and VIM-Ready contributors over an extended period of 1–3 years to rigorously assess the long-term impacts of participation on key career indicators such as job promotions, salary increases, academic publication records, and continued engagement in open-source software participation. This will provide deeper insights into the sustained value of VIM interventions.
- **Cross-Ecosystem Replication:** It is important to test and refine the VIM framework in adjacent technical domains—including burgeoning fields like DevOps, cloud-native development, data visualization, and the scientific R programming community—to validate the relevance of its core pillars and to surface any necessary domain-specific adaptations for optimal effectiveness.
- **AI-Enabled Personalization:** The development of ML-driven tools presents an exciting opportunity to optimize mentor-mentee pairing (e.g., through advanced skill clustering, granular interest matching, and personality compatibility assessments) and to build predictive models that can identify early disengagement risks among members, thereby enabling proactive and targeted interventions.
- **Deep Qualitative Studies:** There is a need to conduct in-depth ethnographic and interview-based research to carefully unpack how micro-cultural factors within communities (such as the sense of identity safety, the prevalence and impact of microaggressions, and implicit biases) interact with structural VIM interventions, informing more nuanced and culturally sensitive Mechanism designs.
- **Compensation & Incentives Experiments:** Future research should pilot and rigorously evaluate various financial reward structures—such as direct stipends, integration of inclusion work into performance bonuses, and targeted micro-grants for contributions—to precisely quantify their effects on sustained member participation, individual and team performance, and overall equity outcomes within communities.

6.5. *Implications for Maintainers and Organizations*

- **Open-Source Maintainers:**
 - **Visibility:** Actively adopt “VIM-Ready” badges to signal an inclusive environment, consistently curate “good first issues” specifically for newcomers, and make it a practice to spotlight newcomer success stories in release notes and community communications to foster encouragement and role modeling.
 - **Invitation:** Offer regular, accessible office hours for Q&A, establish peer-mentorship circles to provide support, and implement automated welcome bots to streamline initial interactions and significantly reduce onboarding friction for new members.
 - **Mechanism:** Clearly articulate and consistently enforce an accessible Code of Conduct; further, appoint an independent ombudsperson or a neutral third-party for conflict resolution and transparently share summaries of incident resolutions to build and maintain community trust.
- **Corporate Stakeholders:**

- ▶ **Performance & Compensation:** Formally integrate inclusion-related contributions into employee performance evaluations—awarding tangible career credit, corresponding merit increases, and financial bonuses for activities such as workshop leadership, active mentorship, and valuable open-source software contributions. This signals that such work is valued.
 - ▶ **Resource Allocation:** Actively sponsor multi-tenant inclusion councils that span across departments or even companies, provide necessary resources like cloud credits or hardware stipends for contributors, and fund cross-company sprints or hackathons focused on inclusion to effectively lower financial and logistical barriers to participation.
 - ▶ **Transparency & Accountability:** Commit to publishing quarterly inclusion dashboards that benchmark VIM-aligned metrics (e.g., diversity in leadership, contributor engagement from underrepresented groups) against peer organizations or industry standards to foster healthy competition, shared accountability, and collective learning across the ecosystem.
- **Academic & Nonprofit Partners:**
 - ▶ **Curriculum Integration:** Proactively embed open-source contribution projects, the use of inclusion toolkits, and active participation in community sprints directly into Data Science and AI courses, as well as capstone projects, to strategically align academic skill acquisition with real-world impact and collaborative practices.
 - ▶ **Joint Programming:** Collaborate with community-focused nonprofits (e.g., PyLadies, R-Ladies, Black in AI, and similar organizations) to co-market and co-host inclusion sprints, workshops, and events, thereby maximizing outreach to diverse talent pools and optimizing the sharing of valuable resources.
 - ▶ **Funding & Fellowship:** Establish and offer targeted scholarships and internship fellowships that are explicitly tied to contributions to open-source software and demonstrated leadership in inclusion initiatives, thereby directly linking academic credentials and opportunities with meaningful community engagement and impact.

7. CALL TO ACTION

The challenge posed by the “Missing 78%” is not an insurmountable or permanent void but rather a compelling call to action—an opportunity to mobilize our collective expertise, supportive structures, and sustained willpower to effect lasting change. We invite all stakeholders—individuals, communities, corporations, and institutions—to actively participate in this transformation:

- **Adopt & Contribute to the Toolkit:** Fork the Open Inclusion Toolkit available on GitHub, thoughtfully tailor it to your specific organizational or community context, and importantly, share back any enhancements made—such as new survey instruments, improved onboarding guides, or refinements to codes of conduct—to benefit the wider ecosystem.
- **Display the “VIM-Ready” Badge:** By displaying this badge, signal clearly to diverse contributors that your project or organization demonstrably meets established inclusive standards across visibility, invitation, and mechanism pillars, doing so in a transparent and verifiable manner that builds confidence.
- **Join the Inclusion Council:** Become an active part of a rotating, cross-company and cross-community consortium. This council will play a key role in governing shared inclusion metrics, coordinating collaborative sprints, and broadly amplifying success stories to inspire wider adoption of best practices.

- **Host or Participate in Sprints:** Commit to organizing or participating in regular quarterly inclusion sprints, dedicated mentorship days, and engaging lightning-talk events—these can be held either within your own community or in broader collaboration with other organizations to maximize impact and knowledge sharing.
- **Publish Your Metrics & Case Studies:** Transparently report your organization's or community's inclusion scorecards and provide detailed, honest reflections on what strategies worked, what didn't, and the lessons learned along the way, thereby enriching the shared evidence base for effective inclusion practices.

Together, by systematically and thoughtfully weaving the principles of Visibility, Invitation, and Mechanism (VIM) into every layer of the Data, AI, and open-source ecosystems—from the daily practices of individual contributors to the strategic imperatives of global corporations—we can collectively strive to turn the “Missing 78%” not just into a closed gap, but into a historic and sustained overrepresentation of diverse, empowered innovators who are driving ethical, robust, and equitable technologies forward for the benefit of all.

REFERENCES

- [1] Boston Consulting Group, “What's keeping women out of data science?,” 2020. [Online]. Available: <https://www.bcg.com/publications/2020/what-keeps-women-out-data-science>
- [2] D. Rossi and S. Zacchiroli, “Worldwide gender differences in public code contributions,” *arXiv*, 2022, [Online]. Available: <https://arxiv.org/abs/2202.07278>
- [3] Idaho State University, “1 in 20 Programmers on Open-Source Software Projects Are Women, ISU Study Says.” [Online]. Available: <https://localnews8.com/isu/2023/06/24/1-in-20-programmers-on-open-source-software-projects-are-women-isu-study-says>
- [4] The Alan Turing Institute, “Women in Data Science and AI.” [Online]. Available: <https://www.turing.ac.uk/research/research-programmes/public-policy/public-policy-themes/women-data-science-and-ai>
- [5] R. Lorenzo, N. Voigt, M. Tsusaka, M. Krentz, and K. Abouzahr, “How Diverse Leadership Teams Boost Innovation,” 2018. [Online]. Available: <https://www.bcg.com/publications/2018/how-diverse-leadership-teams-boost-innovation>
- [6] C. Díaz-García, Á. González-Moreno, and F. J. Sáez-Martínez, “Gender Diversity within R&D Teams: Its Impact on Radicalness of Innovation,” *Innovation: Management, Policy & Practice*, vol. 15, no. 2, pp. 149–160, 2013, doi: [10.5172/impp.2013.15.2.149](https://doi.org/10.5172/impp.2013.15.2.149).
- [7] M. M. Fuentes-Fuentes, A. M. Bojica, and M. Ruiz-Arroyo, “Gender Diversity, Inclusive Innovation and Firm Performance,” *Sustainable Development*, vol. 31, no. 5, pp. 3622–3638, 2023, doi: [10.1002/sd.2615](https://doi.org/10.1002/sd.2615).
- [8] G. Desvaux, S. Devillard, and S. Sancier-Sultan, “Women Matter: Gender Diversity, a Corporate Performance Driver,” 2007. [Online]. Available: <https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/gender-diversity-a-corporate-performance-driver>
- [9] B. Trinkenreich *et al.*, “Pots of Gold at the End of the Rainbow: What is Success for Open Source Contributors?,” *IEEE Transactions on Software Engineering*, vol. 48, no. 10, pp. 3940–3953, 2022, doi: [10.1109/TSE.2021.3108032](https://doi.org/10.1109/TSE.2021.3108032).
- [10] Harnham, “Diversity in Data Report,” 2023. [Online]. Available: <https://www.harnham.com/did-deep-dive>
- [11] UNESCO, “G20: UNESCO Calls for Greater Gender Equality in Science.” [Online]. Available: <https://www.unesco.org/en/articles/g20-unesco-calls-greater-gender-equality-science>
- [12] “Gender gap in research publishing is improving — slowly,” *Nature*, 2025, doi: [10.1038/d41586-025-00553-x](https://doi.org/10.1038/d41586-025-00553-x).
- [13] T. C. Dennehy and N. Dasgupta, “Female Peer Mentors Early in College Increase Women's Positive Academic Experiences and Retention in Engineering,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 23, pp. 5964–5969, 2017, doi: [10.1073/pnas.1613117114](https://doi.org/10.1073/pnas.1613117114).
- [14] J. Terrell *et al.*, “Gender Bias in Open Source: Pull Request Acceptance of Women Versus Men,” *PeerJ Computer Science*, vol. 3, p. e111, 2017, doi: [10.7717/peerj-cs.111](https://doi.org/10.7717/peerj-cs.111).
- [15] S. Sultana, A. Turzo, and A. Bosu, “Code Reviews in Open Source Projects: How Do Gender Biases Affect Participation and Outcomes?,” *arXiv*, 2022, [Online]. Available: <https://arxiv.org/abs/2210.00139>
- [16] WomenTech Network, “Women in Tech Stats 2025.” [Online]. Available: <https://www.womentech.net/women-in-tech-stats>
- [17] S. Gammage, “Gender Equality and the Costs of Unpaid Caregiving,” *IMF Finance & Development*, 2019, [Online]. Available: <https://www.imf.org/en/Publications/fandd/issues/2019/03/gender-equality-and-costs-of-unpaid-caregiving-gammage>

- [18] Built In, “AI Bias: What It Is and How to Prevent It.” [Online]. Available: <https://builtin.com/artificial-intelligence/ai-bias>
- [19] A. Maddi and Y. Gingras, “Gender Diversity in Research Teams and Citation Impact in Economics and Management,” *arXiv*, 2020, [Online]. Available: <https://arxiv.org/abs/2011.14823>
- [20] S. Brooke, “Programmed Differently? Testing for Gender Differences in Python Code Quality,” *Journal of Computer Mediated Communication*, vol. 29, no. 1, p. zmad49, 2023, doi: [10.1093/jcmc/zmad049](https://doi.org/10.1093/jcmc/zmad049).
- [21] V. Hunt, S. Prince, S. DixonFyle, and L. Yee, “Delivering Through Diversity,” 2018. [Online]. Available: <https://www.mckinsey.com/featured-insights/diversity-and-inclusion/delivering-through-diversity>
- [22] “The Tipping Point: Women on Boards and Financial Performance,” 2016. [Online]. Available: <https://www.msci.com/www/research-paper/the-tipping-point-women-on/0474298723>
- [23] PyData, “Diversity Scholarships.” [Online]. Available: <https://pydata.org/diversity-inclusion/>
- [24] CIO Magazine, “Ada Developers Academy: A No-Cost Career Program that Aims to Diversify IT.” [Online]. Available: <https://www.cio.com/article/308608/ada-developers-academy-a-no-cost-career-program-that-aims-to-diversify-it.html>
- [25] Outreachy, “Internships Supporting Diversity in Tech.” [Online]. Available: <https://www.outreachy.org/>
- [26] “Evaluation Report: TechWomen,” 2020. [Online]. Available: https://eca.state.gov/files/bureau/final_report__v2_techwomen140d0420f0211-for_web.pdf
- [27] Organisation for Economic Co-operation and Development, “Unpaid Care Work: The Missing Link in Gender Equality,” 2020. doi: [10.1787/1f3fd03f-en](https://doi.org/10.1787/1f3fd03f-en).