African Population and Health Research Center

Data Science Without Borders

2025

DATA ANOMALY DETECTION & RECORD LINKAGE HACKATHON REPORT

Prepared by :

Armauer Hansen Research Institute (AHRI)

www.dswb.africa

Executive Summary

The Data Science Without Borders (DSWB) through Armauer Hansen Research Institute (AHRI), held a hackathon, from April 24 to May 22, in a hybrid format. The hackathon brought together 30 young and motivated AI and data science professionals to tackle two critical health data challenges: retrospective record linkage and data anomaly detection. Organized through the collaborative efforts of AHRI and its DSWB consortium partners, the hackathon aimed to develop robust, scalable, and context-sensitive solutions for fragmented health data systems commonly found in low- and middle-income countries.

Over three intensive weeks, two virtual and one in-person, the hackathon engaged participants through a structured program featuring a kickoff session, ongoing mentorship from local and international experts, regular check-ins, and final pitching. A total of 15 teams, formed based on complementary skills, competed by designing and building AI-driven models using real-world-inspired datasets. The event emphasized practical innovation, technical rigor, and interdisciplinary collaboration.

The final judging showcased high-quality, deployable solutions that addressed challenges like missing data, quasi-identifiers, and noisy datasets. Winning teams were recognized for excellence in innovation, clean coding, and user interface design, with prizes totaling 200,000 ETB and offers of continued mentorship and research collaboration. Participant and technical team feedback highlighted the event's strong learning impact, collaborative spirit, and areas for enhancement in future iterations. Overall, the hackathon demonstrated the potential of youth-led innovation to solve pressing health data challenges in resource-constrained settings.

Acknowledgment

We extend our heartfelt gratitude to everyone who contributed to the success of the DSWB_AHRI_Hackathon_2025.

First and foremost, we thank all participants for their creativity, resilience, and outstanding commitment throughout the competition. Your energy brought the hackathon to life.

Our deepest appreciation goes to the technical mentors and judges from Armauer Hansen Research Institute (AHRI), the African Population and Health Research Center (APHRC), London School of Hygiene and Tropical Medicine, U. of London (LSHTM) and OSPO Now, for their time, insights, and dedication. Your guidance played a vital role in shaping high-impact, context-aware solutions.

We acknowledge the unwavering support of the broader DSWB consortium leadership for their vision and encouragement. We are also grateful to the legal, and administrative teams at AHRI for their behind-the-scenes efforts in drafting agreements and managing logistics.

A special thank you goes to our funder, the Wellcome Trust, whose generous support made this hackathon possible. Your investment in data science innovation and global health research is deeply appreciated and continues to inspire transformative work.

Finally, we thank all partner institutions who enabled this initiative and supported our mission to foster open-source innovation and data science capacity in Ethiopia and beyond.

Table of Contents

Data Science Without Borders

1. Overview of the Report

This report will cover details on the DSWB_AHRI_Hackathon_2025 which was held from April 24 to May 22 in hybrid form online and in person on the two problem areas of Data anomaly detection and record linkage. Problem area selection, participant selection, running the hackathon, results, feedback and future plan is included in this report.

2. Thematic area selection

The themes for this hackathon were selected through the process of data mapping and prioritization done with local partners of DSWB_AHRI. The presence of various data anomalies and challenges in linking records across different databases for the same subjects were identified as technical issues. To address these, a hackathon was planned to engage young researchers, computer scientists, statisticians and data scientists to rapidly develop a viable solution within a short time frame.

2.1 Thematic Area

2.1.1 Record linkage

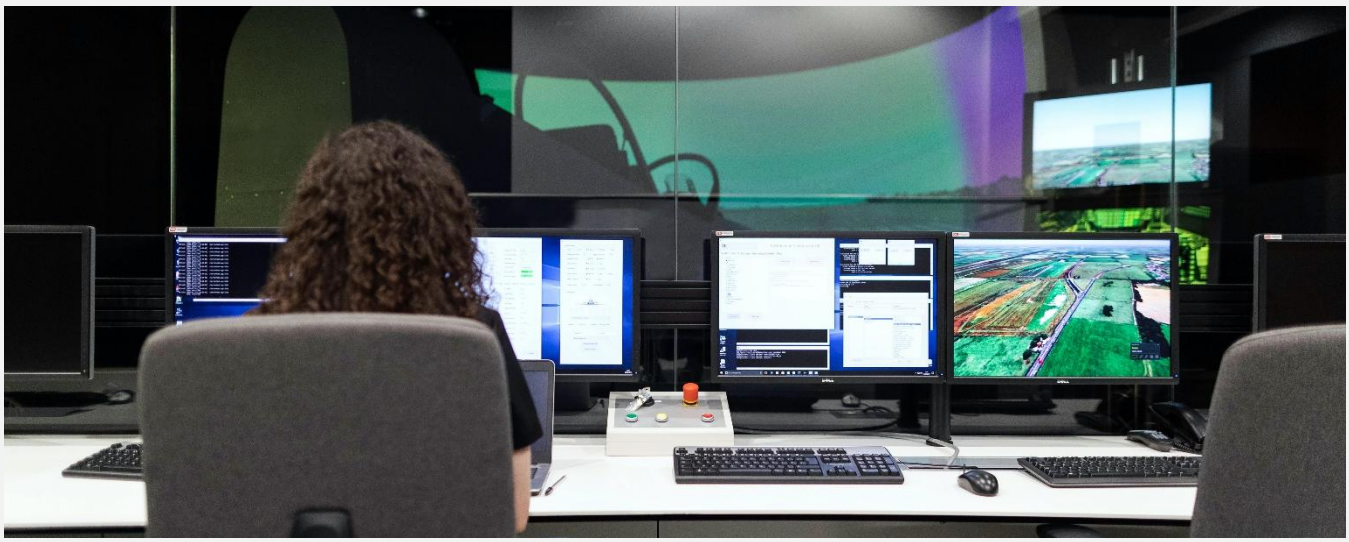
Record Linkage (RL) in this context refers to the process of identifying and connecting retrospective records that pertain to the same real-world entity, such as a person, household, or event, across multiple datasets not originally designed to interoperate.

Unlike prospective systems that systematically collect and maintain unique identifiers (e.g., national or patient IDs) or allow real-time validation, retrospective RL typically operates in environments where such identifiers are absent. This challenge is particularly acute in health and demographic research, such as linking Health and Demographic Surveillance System (HDSS) data with clinic or health records.

These datasets are often managed in isolation, lacking shared identifiers and exhibiting significant variability in structure, completeness, and quality. In the absence of unique IDs, researchers must rely on quasi-identifiers such as names, dates of birth, sex, and location.

However, these variables are frequently incomplete, inconsistently formatted, or poorly standardized: Names may be misspelled, inconsistently recorded, or influenced by local naming conventions. Dates of birth may be approximate, defaulted (e.g., "01/01/1970"), or missing. Location data may vary in granularity or reference different administrative boundaries.

Given the critical role of record linkage (RL) in enabling longitudinal analysis, population health monitoring, and policy evaluation, there is a need for robust, transparent, and context-aware linkage methodologies that can operate effectively under variable-poor and high-uncertainty conditions.



2.1.2 Data anomaly detection

Anomaly detection is a critical challenge in data analysis, aiming to identify patterns that deviate significantly from expected behavior. Anomalies can be missing data, point anomalies like outliers and Inconsiderateness which are usually referred to as contextual or collective anomalies.

These challenges are common in health data sets like electronic medical records, health and demographic data sets, medical device generated data and others. Data can be missing in three ways: completely at random (with no identifiable pattern), at random (with a pattern related to observed data), or not at random (where the missingness is related to unobserved data).

Point anomalies, extreme values that deviate from the norm like a weight of 5kg for an adult can be present. Temporal, clinical and device inconsistencies can also happen where patient discharge before admission, pregnancy diagnosis for male patient, pulsometer reading zero for some time can be examples respectively.

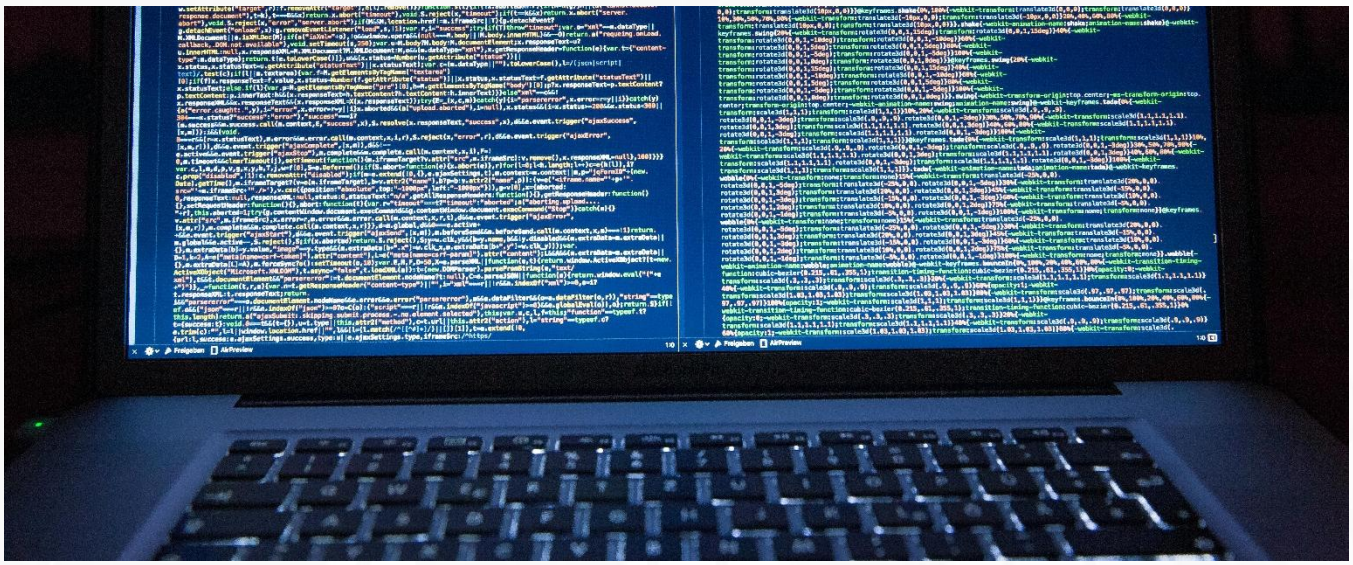
Considering data anomalies in the health sector mentioned above like missing data, outliers, and inconsistencies pose significant risks to clinical decision-making, research, and public health analytics, there is a need for robust AI driven solutions that can automate and enhance data anomaly detection. These problems are particularly pronounced in low- and

3. The Hackathon

The Hackathon was conducted with the aim of developing Artificial Intelligence (AI) / Machine Learning (ML) based innovative, scalable, robust and practically useful solutions/models which are

1. Capable of performing accurate and scalable retrospective record linkage (RL) in the context of fragmented, low-structure datasets often encountered in population and health research.

Participants were expected to design and train algorithms capable of linking records that refer to the same real world entity across disparate datasets. These models must learn from quasi-identifiers such as names, dates of birth, sex, and location and intelligently combine multiple signals to infer high confidence matches. Emphasis were placed on probabilistic, deep learning, and hybrid approaches to structured data environments. The participants were expected to come up with suggestions on additional variables that can improve the quality of RL and demonstrate the same.



2. Capable of identifying anomalies, unusual patterns ,in real world data sets addressing noise, imbalanced data and dynamic environments.

It was anticipated that participants will develop robust anomaly detection systems capable of identifying missing data, point anomalies (outliers), and contextual/collective anomalies in real-world health datasets. The solutions must handle noisy, incomplete clinical data while distinguishing true medical anomalies from artifacts or normal variations.

3.1. Pre-Hackathon

3.1.1 Call for Applications

The hackathon call was announced on March 25 and remained open until April 10. To be able to reach a wider range of applicants, the call was announced through AHRI university network, Computer science and data science affiliated universities, AHRI website, AHRI and DSWB'S LinkedIn and social media.

Applicants were required to submit a CV, motivation letter, and a summary of their educational, technical background, programming and AI related skills and related previous work.

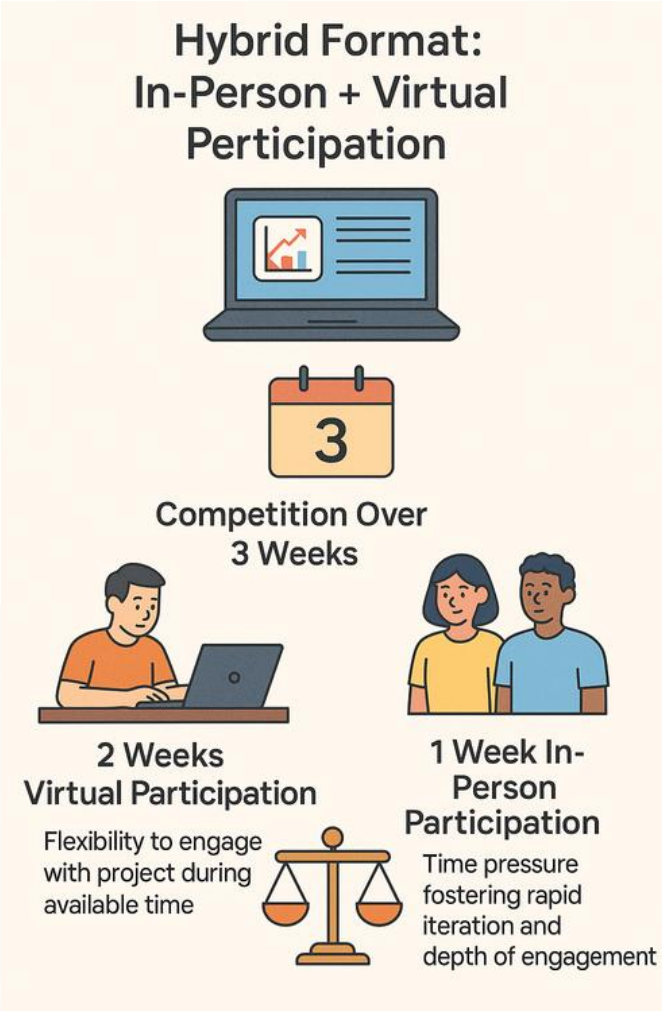
A total of 33 individuals applied where some applied as teams, while others applied individually. Those who applied individually were paired by the organizing committee based on complementary backgrounds and were accepted by the participants. Twenty five of the applicants were male and five women and among all 26 held bachelor's degrees either in Computer Science, data science or related field

3.1.2 Application form overview

Interested participants were required to complete an application form that collected essential information to assess their suitability and motivation for the event. The form annexed with this report included: (Annex 1)

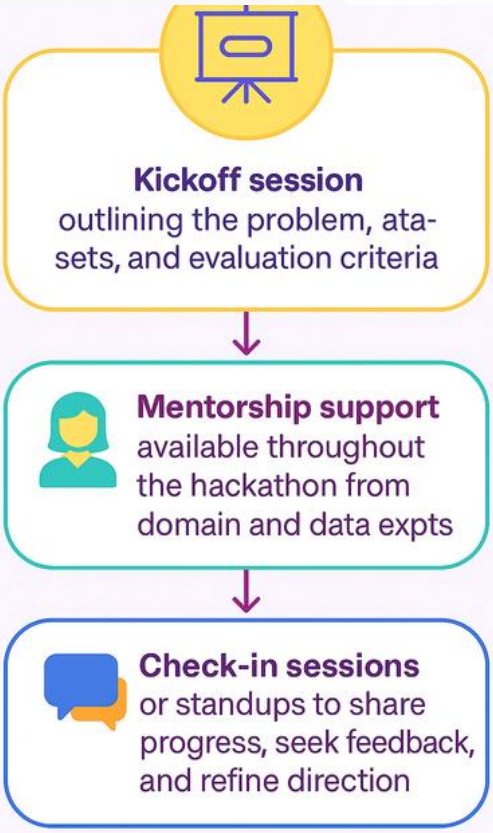
- **Personal Information:** Participants provided their email addresses, gender, age, and educational background to help organizers understand the demographic composition.
- **Academic and Professional Background:** Details about the participants' institutions and fields of study were gathered to ensure a diverse mix of expertise.
- **Technical Skills & Expertise:** Applicants were asked to indicate their proficiency in areas such as AI/ML, web development, data science, cyber security, UI/UX design, and embedded systems, the type of dataset they worked on, the model they have trained previously with the Github link. This information was crucial for team formation and resource allocation.
- **Motivational Letter:** Participants were encouraged to submit a brief statement explaining their reasons for wanting to participate and what they hoped to achieve from the hackathon. This section provided insights into their passion and commitment to the event's objectives.

3.1.3 Competition Format



The Hybrid Competition unfolded over three weeks where two weeks were virtual, and one week in person, designed to strike a balance.

Participants had access to - >



Team Formation and Thematic Assignment

Twelve applicants
who applied as teams remained paired

Remaining individuals
grouped based on complementary skills

15
teams

Data Anomaly Detection

Eight teams randomly assigned

Record Linkage

Seven teams randomly assigned

3.2 Platform selection

A joint discussion between DSWB project partners resulted in the recommendation to use Discord as the primary communication platform. A team member was assigned to set up and manage the channels. A technical committee comprising six members from the participating institutions was established to mentor and evaluate hackathon participants. Additional communication channels included email and WhatsApp.

The Discord server was launched shortly thereafter. During the first technical team meeting, members reviewed potential platforms for managing the hackathon and agreed on GitHub for code submission and project tracking. Python was selected as the primary programming language, with flexibility to incorporate R and other tools as needed.

Participant Guidelines, Agreements, and Evaluation Tools

Building on the hackathon guidelines shared by LSHTM, a customized participant guide was developed and distributed to all participants **(Annex 2).**

A Hackathon and Non-Disclosure Agreement, drafted by the AHRI legal team, was reviewed by consortium partners and signed by all participants **(Annex 3)**.

Evaluation forms were also created to assess participant progress at each phase of the hackathon (**Annex 4).**

4. Running the Hackathon

Phase 1: Virtual Phase

Kickoff and Team Formation

The event begun with a virtual opening
ceremony to welcome participants, outline the purpose and goals of the hackathon, and introduce the organizing team and mentors after receiving their selection via email on April 17.

Participants were notified of the team formation process, designed to create interdisciplinary groups that combine technical, analytical, and domain expertise.

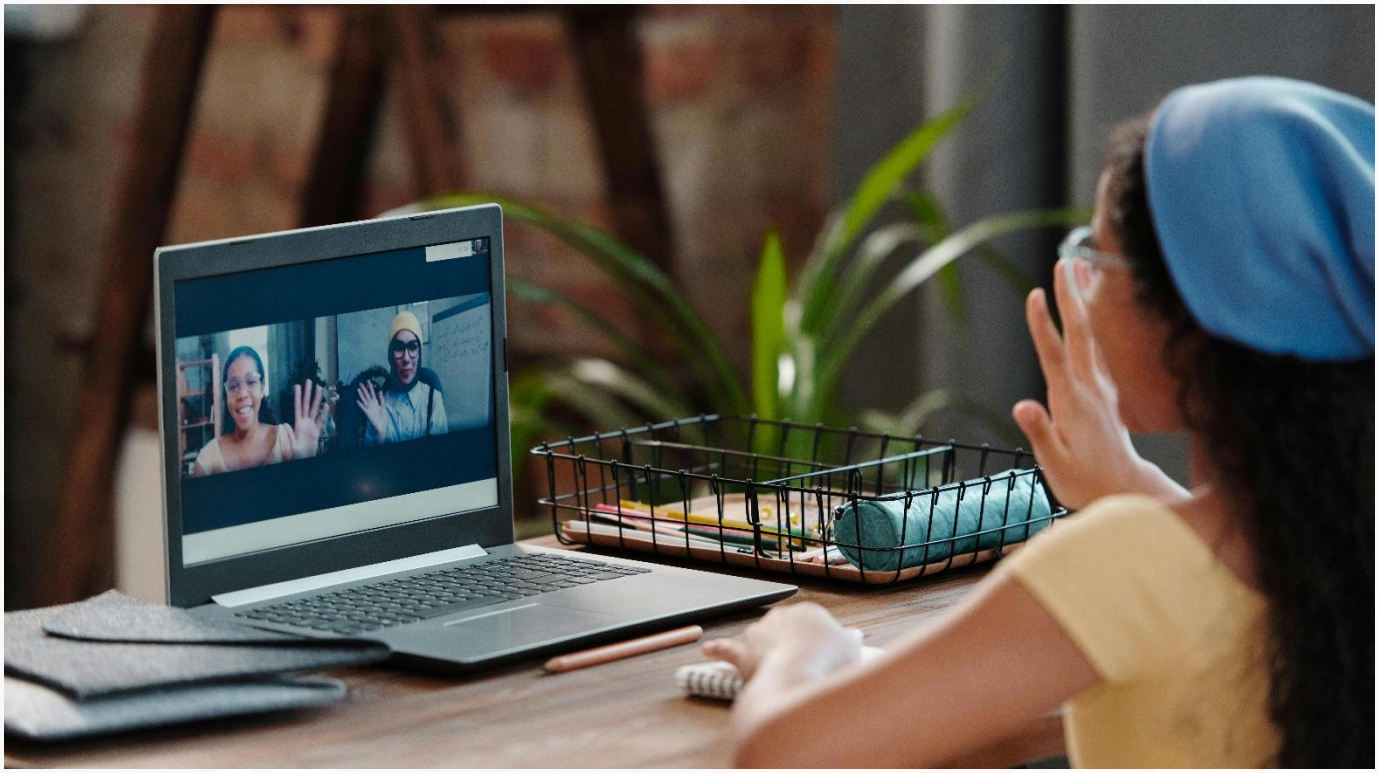
Online work session

After teams were finalized, synthetic datasets were distributed on May 12 along with documentation on data format, evaluation metrics, and submission guidelines.

Required documents, design documents and milestone code, were submitted on May 13 and the first evaluation was conducted on May 14 using the following criteria:

Three teams were selected from each thematic area which is a total of 6 teams to proceed to the in person session.

The first evaluation was conducted on May 14 using the following criteria:



Phase II - In person Phase

The in person phase began on the morning of May 15, with participants traveling to Lake Bishoftu Resort.

The opening session included a welcome address by one of the co-leads of the Ethiopia team, followed by presentations on open science and GitHub by a representative from OSPO Now, and an overview of data record linkage by a representative from LSHTM.

Onsite mentorship was provided and regular progress was monitored daily. Regular check-ins and optional presentations (such as lightning talks) provide opportunities for peer feedback, visibility of diverse approaches, and real-time mentorship. Late-stage support was made available through communication platforms and dedicated help desks, ensuring teams have access to technical assistance as needed.

4.1 Deliverables and Judging

After the hackathon, each team was required to submit the following:

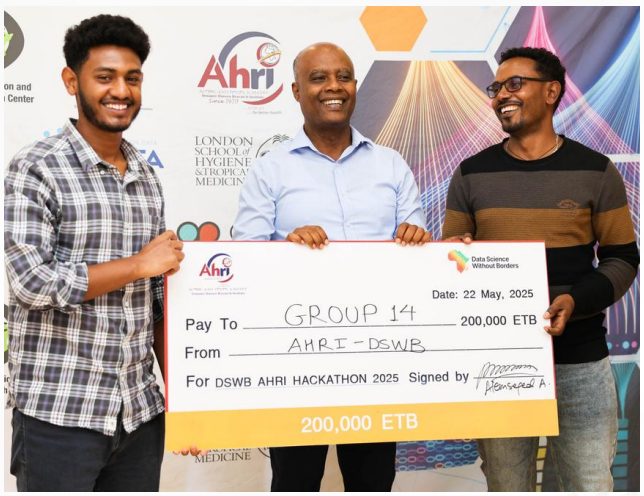
- A functional prototype or model addressing their assigned theme, either anomaly detection or record linkage, along with the corresponding code and reproducible outputs.
- A brief technical report outlining their methodology, challenges faced, and evaluation metrics used.
- A 15-minute pitch or demo summarizing their approach and presenting their results.

On the final day, teams presented their projects to a panel of judges, demonstrating their approaches, highlighting innovations, and discussing the strengths and challenges of their solutions.

4.2 Awards and Recognition

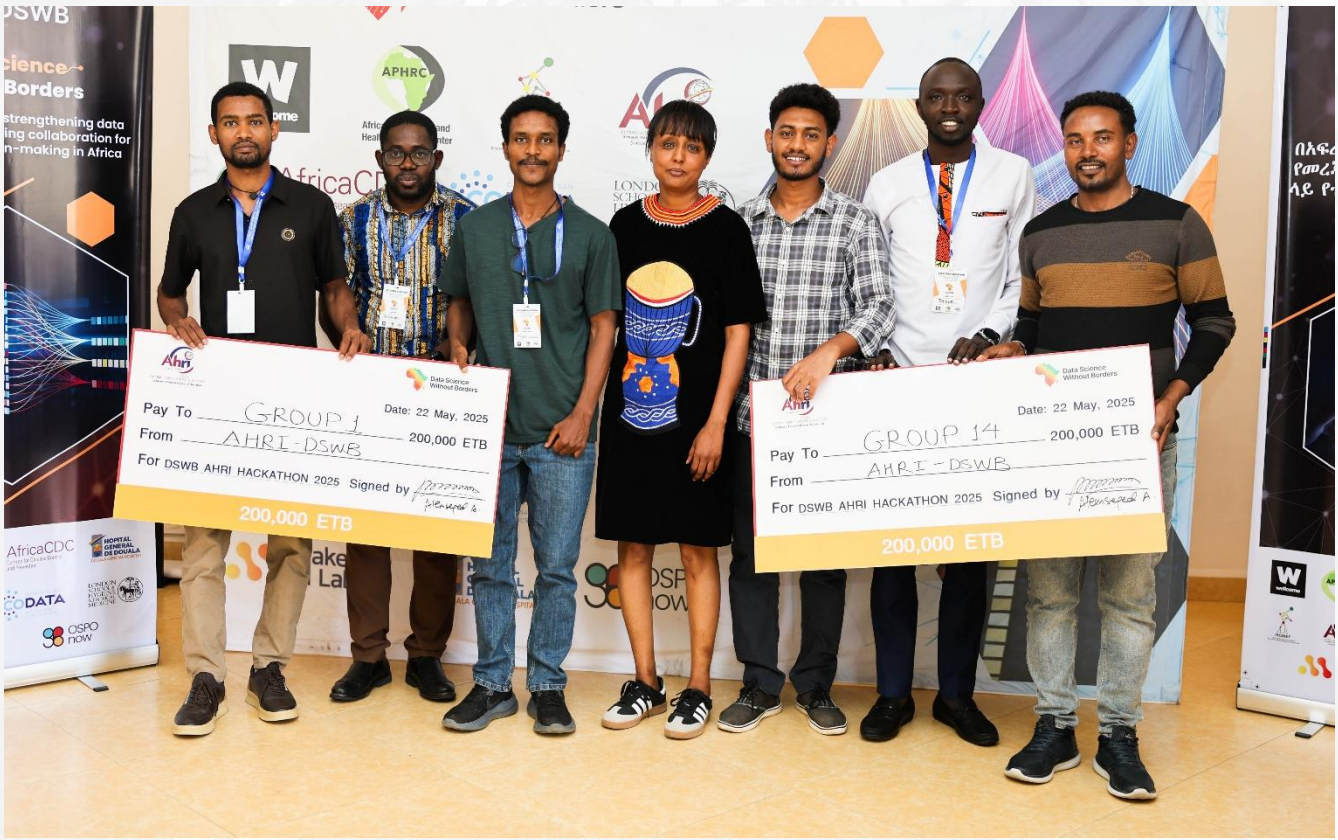
The closing ceremony, held on May 22, welcomed guests from various partner institutions. Closing remarks were delivered by Dr. Agnes Keraga, the project lead and Dr. Alemseged Abdissa, the AHRI PI & Site Lead. Each team presented their work and received feedback.

Awards of 200,000 ETB were given to the top teams in each thematic area. Additional recognitions included **Best Innovation**, **Best Clean Code**, and **Best UI**, with each receiving 10,000 ETB. Winning teams were also offered mentorship and potential opportunities for future collaboration.



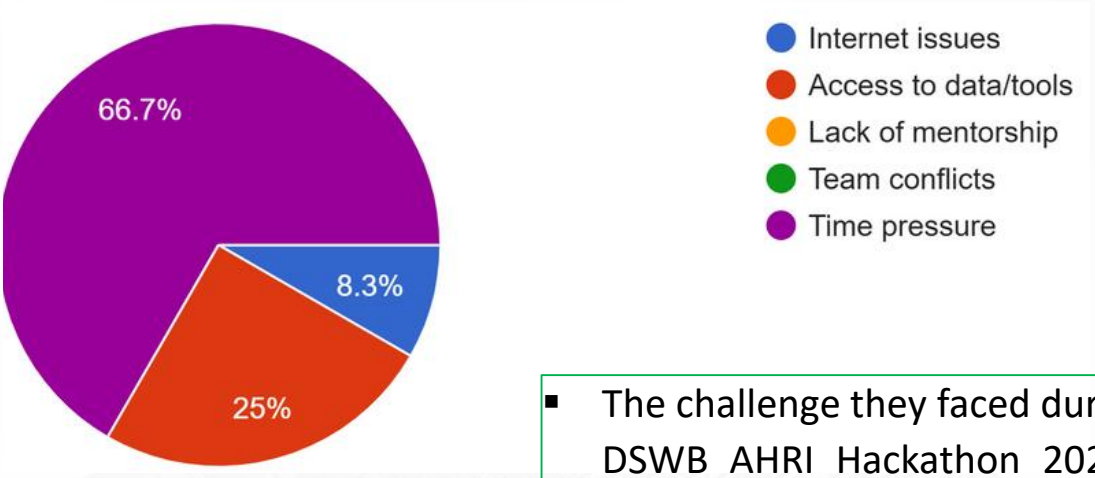
The Record Linkage platform is a no-code, modular system for record linkage, guiding users from data upload to matched results. It handles data cleaning, blocking, similarity scoring, and classification, with tools for model evaluation, active learning, and result exposure supporting both batch and real-time use.

The **Data Anomaly Detector** uses a hybrid approach combining rule-based logic, ML models (Isolation Forest, LOF, One-Class SVM), and deep learning (Autoencoder) to detect anomalies in health data. An LLM filters out irrelevant features, and results are explained and visualized via a Streamlit dashboard.

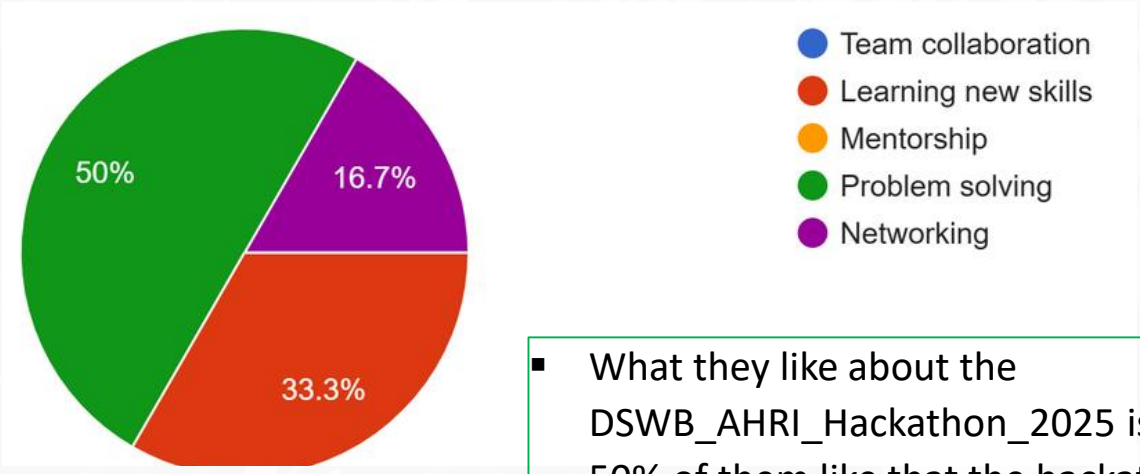


5. Feedback

Participants Feedback: 100% of the participants fill the form.



- The challenge they faced during the DSWB_AHRI_Hackathon_2025, 66 % said have only one week to actually build their products.



- What they like about the DSWB_AHRI_Hackathon_2025 is 50% of them like that the hackathon is problem solving.

Technical team Feedback

What Went Well?

- Excellent Venue and Cultural Experience: The choice of venue provided a conducive environment for collaboration.
- Dedicated Attendees: Participants demonstrated dedication, with mentors needing to encourage them to take breaks and enjoy the process.
- Growth in Confidence: Some teams started with unclear ideas but, through mentorship and support, developed strong, confident presentations, showcasing significant progress.
- Effective Coordination

Challenges Encountered

- Technical Difficulties: Online Mentors experienced issues with demonstrations, sound, and screen sharing, impacting their engagement.
- Short Notice for activities: OSPO Now received limited time to prepare for the GitHub session, affecting the quality of the topic delivered.

Areas for Improvement

- Pre-Hackathon Training: Implement standardized training on tools like Git and Docker to ensure all participants are on the same level.
- Extended Duration: Consider lengthening the in-person session to allow more time.
- Hybrid Setup Testing: Thoroughly test hybrid modes beforehand, ensuring consistent connectivity and clear audio-visuals.

Additional Recommendations

- Enhanced Planning Involvement: Involve more stakeholders, in the planning stages to ensure comprehensive preparation.
- Resource Accessibility: Provide necessary resources and training materials in advance to participants.
- Continuous Engagement: Foster ongoing collaboration among participants and technical partners post-event.

6. Conclusion

The hackathon successfully brought together diverse teams to solve challenges in data anomaly detection and record linkage. The virtual phase enabled wide participation and idea development, while the in-person phase provided focused collaboration and mentorship.

However, challenges were encountered such as timeline issues, delayed data availability, containerization and technical difficulty among some participants.

Future events should prioritize better communication, early preparation, and timely data access. Participants valued the teamwork and knowledge sharing, highlighting the event's overall success in capacity building and innovation.

7. Next Steps

Weekly meetings are currently ongoing, with the winning and second runner-up teams to consolidate additional feedback and the core functionality of the products has been finalized. Once the products are finalized, they will be deployed on the AHRI server for testing. A post-hackathon workshop will then be organized to gather feedback from relevant stakeholders. Based on this feedback, final refinements will be made, and the solutions will subsequently be made publicly available through the designated platform at the Armauer Hansen Research Institute (AHRI).

8. Annex

Annex 1: Participants Application Form

Annex 2: Hackathon Guideline

Annex 3: Nondisclosure Agreement of the Hackathon

Annex 4: Assessment Forms

- Phase_1 Evaluation Form
- Phase_2 Evaluation Form

Data Science Without Borders

Bridging data gaps, strengthening data systems, and enhancing collaboration for data-driven decision-making in Africa

www.dswb.africa