Deep Learning Indaba 2024 1-7th September, University Amadou Makhtar Mbow. Senegal

Transformative power of Artificial Intelligence (AI) in Trusted Research Environments (TREs) for African Scientists

Prof Muliaro Wafula PhD, FCCS, FCSK

Lead Scientist AOSP EA Node, Kenya

Associate Professor Computing Department – JKUAT, Kenya

Chair CODATA, Kenya

AICAD, JKUAT and AOSP-EA

- African Institute for Capacity Development (AICAD), is a Regional Intergovernmental Organization established by a Charter subscribed to by the Governments of the Republic of Kenya, the United Republic of Tanzania and the Republic of Uganda. The Headquarter is in Nairobi, Kenya with its headquarters at the Jomo Kenyatta University of Agriculture and Technology (JKUAT) main campus in Juja town. AICAD's overriding objective is to develop human capacity critical in exploiting abundant resources available in the African region for poverty reduction and to be mindful of sustainability and perpetuity.
- AICAD currently has collaboration MoU with over 25 universities in the region.
- AICAD and JKUAT jointly won the bid to host the AOSP-EA region node.
- The Africa Open Science Platform Eastern Africa Node (AOSP-EA node) administration office are based at AICAD.
- The main objective of the AOSP)-EA node is to promote efforts aligned with implementing open science Programmes in the Eastern Africa region, strengthen knowledge networks and infrastructure access and enhance cooperation between regions and globally in support of AOSP's vision.
- JKUAT also the CODATA national member, contributes the data science, data engineering and data management expertise at AOSP EA Node and are key resource persons behind the proposal for the development of the DASSA platform proposal.

TREs to facilitate secure and ethical data sharing.

- TREs implement rigorous security measures, including data **anonymizati encryption**, and **controlled access**, to safeguard against unauthorized ac breaches.
- TREs enable researchers to access and use data in a controlled, transparent manner that adheres to **ethical guidelines**.
- TREs ensure that data is used responsibly, respecting participants' rights and mitigating the risk of misuse through **enforcing governance, and legal (National/International) frameworks** eg General Data Protection Regulation (GDPR) in the EU or the Health Insurance Portability and Accountability Act (HIPAA) in the U.S.
- TREs **foster collaboration** among researchers by providing secure access to shared datasets, while also ensuring that the integrity and confidentiality of the data are maintained,
- TREs **build trust among researchers, data providers, and the public**. This trust is crucial for encouraging data sharing and collaboration, which are essential for **addressing complex global challenges**
- TREs by protecting sensitive data and building trust among stakeholders, they enable collaborative research while **safeguarding individual privacy and data integrity**

Primary Safeguards for Privacy and Data Security in TREs

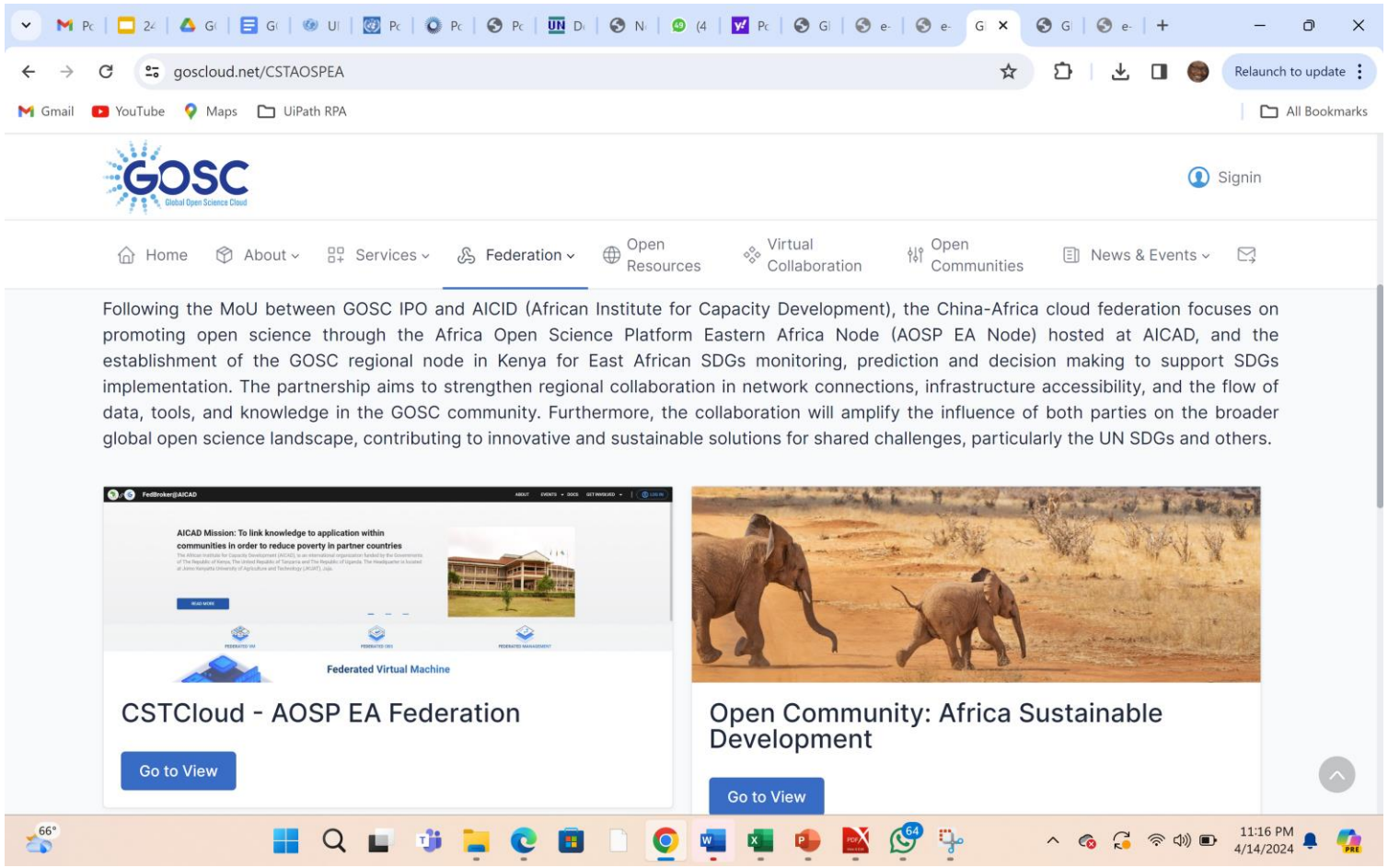
- ✓**Encryption** ensures that data is transformed into an unreadable format while in storage (at rest) and during transmission (in transit).
- ✓**Access Controls and Authentication**
  - **Role-based Access Control (RBAC)**: Users are given access based on their role within an organization
  - **Multi-factor Authentication (MFA)**: Requires users to provide multiple forms of verification (e.g., password and phone authentication)
- ✓**Data Anonymization and Pseudonymization**
  - **Anonymization**: Irreversibly removing personally identifiable information (PII) from datasets to ensure that individuals cannot be identified from the data.
  - **Pseudonymization**: Replacing PII with fictional identifiers (such as random IDs) so that data can still be re-linked to individuals by authorized parties if needed.

Primary Safeguards for Privacy and Data Security in TREs ….2

- ✓ **Audit Trails and Monitoring:**
  - Keeping detailed logs of who accessed the data, when, and for what purpose provides transparency and accountability

- ✓**Data minimization** involves collecting and storing only the data that is necessary for a specific purpose.
- ✓**Regular audits and penetration tests** are used to identify potential vulnerabilities in the system
- ✓**Privacy and Security Policies:** Establishing clear internal policies regarding data use, storage, sharing, and deletion ensures that employees understand their responsibilities regarding data privacy
- ✓**Legal and Regulatory Compliance**

Development of Trusted Research Environment



Challenges in African Research: Data and Collaboration Gaps

In the African context, the integration of AI technologies into **Trusted Research Environments (TREs)** holds immense potential for overcoming persistent challenges that have limited research and data-driven innovation

1. **Fragmented Collaboration**: Due to a lack of shared research infrastructure and collaborative frameworks, researchers in African institutions often work in silos, limiting interdisciplinary and cross-border collaboration (Akinola et al., 2022).
2. **Limited Data Access**: High-quality, actionable datasets are often scarce or difficult to access. Eg in critical fields such as public health, agriculture, and environmental sustainability (World Bank, 2021).
3. **Scarcity of Technical Resources**: Inadequate skilled data scientists to harness AI technologies for development (Njuguna & Ikoja-Odongo, 2020).

Potential of AI in TRE

- Through AI-based tools, data from multiple sources can be harmonized, making it easier for researchers to access, analyze, and derive insights from large datasets
- AI-powered search and query systems enable researchers to navigate vast amounts of data more efficiently, promoting faster and more reliable research outcomes (Shen et al., 2022).
- AI can analyze patterns across various disciplines, identifying commonalities and potential areas for cross-sector research.
- AI-enabled platforms also facilitate collaboration by recommending research partnerships and relevant datasets based on an individual researcher's work, thereby fostering a more connected research environment across the continent (Tchamyou et al., 2019).
- integrating AI allows for the development of data-driven solutions to local and regional challenges, accelerating progress toward the **Sustainable Development Goals (SDGs)** (UN, 2020).

Overcoming Barriers to AI Adoption

- **Limited Funding**: African research institutions often lack the financial resources to invest in cutting-edge AI technologies
- **Lack of Regulatory Frameworks**: Many African nations lack comprehensive data governance frameworks that support AI research

How to Leveraging AI in TREs

1. **Automated Data Curation and Preparation**: AI can automate data preprocessing tasks like cleaning, anonymizing, and structuring data - In healthcare TREs, AI could automatically de-identify patient data, categorize medical images, and structure clinical trial datasets for easy access

2. **Smart Querying and Data Discovery-** Using AI-driven search tools, researchers from different disciplines can quickly find shared datasets across disciplines like genomics, economics, or social sciences by asking questions in natural language rather than using complex query languages

3. **Federated Learning for Decentralized Data Access-** Multiple hospitals or institutions could collaborate to train an AI model on patient data without needing to centralize or transfer sensitive datasets, thus enhancing accessibility to the combined knowledge without data leaving its secure environment

4. **Collaborative Platforms Using AI-** An interdisciplinary team of biologists, data scientists, and social scientists can collaborate on public health data in a TRE, using AI tools to highlight patterns, make predictions, and visualize relationships across datasets that individual researchers may not notice on their own.

5. **AI-Driven Knowledge Graphs-** AI could link genetic research with environmental studies by identifying relationships between genomic variations and environmental factors in large datasets, thus promoting collaboration between geneticists and environmental scientists

6. **Natural Language Processing for Multidisciplinary Insights**: NLP technologies can help translate complex terminology across disciplines

AI frameworks supporting TREs

1. **TensorFlow- H**andle large datasets. Eg analyzing healthcare, genomic, or other research data. Its sca
for federated learning allow researchers to build models without moving sensitive data outside the s
environment.
2. **PyTorch** can be used to develop AI models that analyze structured or unstructured data, like medical images or large-scale datasets in bioinformatics, while adhering to privacy constraints
3. **Apache Spark**- Its distributed processing capabilities can handle large-scale datasets typical of research environments, ensuring data is processed efficiently within the boundaries of the TRE
4. **Federated Learning Frameworks (e.g., TensorFlow Federated, PySyft)-** ensure sensitive data never leaves the TRE. These frameworks can be mainstreamed into TREs to facilitate collaboration between institutions while preserving data privacy.
5. **XGBoost-** efficiency and strong performance in tasks such as classification, regression, and ranking make it suitable for TREs dealing with tabular datasets. It's often used in predictive modeling in healthcare and other data-driven research areas
6. **IBM Watson** - healthcare AI modules can be integrated into TREs to analyze medical research data .
7. **H2O.ai** - Has ability to handle large-scale machine learning models makes it useful for TREs that need to automate data analysis without exposing sensitive information
8. **Azure-** Has strong security protocols and compliance with global data standards make it an ideal solution for TREs.
9. **Scikit-learn** -Useful for TREs handling moderate-sized datasets. Its simplicity and ease of integration with other libraries (like pandas and NumPy) make it a good choice for research that doesn't require deep learning
10. **Privacy-Preserving AI Frameworks (e.g., OpenMined, CrypTen) -**critical in TREs where data confidentiality is paramount.

References

1. Akinola, O., et al. (2022). *Collaboration Challenges in African Research Institutions*. African Journal of Science, Technology, Innovation and Development.
2. Njuguna, J., & Ikoja-Odongo, R. (2020). *Data Access Challenges in African Research*. Research Policy and Planning.
3. Nkohkwo, Q., & Islam, M. (2013). *Fostering AI Capacity in African Universities*. Journal of African Higher Education.
4. OECD. (2020). *Trusted Research Environments and Their Role in Data Sharing*. OECD Policy Brief.
5. Shen, Y., et al. (2022). *AI and Data Accessibility in Research*. Journal of Computational Science.
6. Tchamyou, V., et al. (2019). *AI, Data, and Interdisciplinary Collaboration in Africa*. Development and Policy Review.
7. Tisné, M. (2021). *Data Governance and AI Ethics in Africa*. Journal of Ethics and Data Innovation.
8. United Nations (UN). (2020). *Artificial Intelligence and the Sustainable Development Goals (SDGs)*. UN Development Report.
9. World Bank. (2021). *Challenges in Data Utilization in African Research*. World Bank Policy Paper.

Thank you!

muliaro@icsit.jkuat.ac.ke