Fostering Data Science Capacity and Engagement in Africa
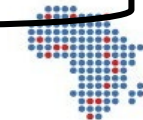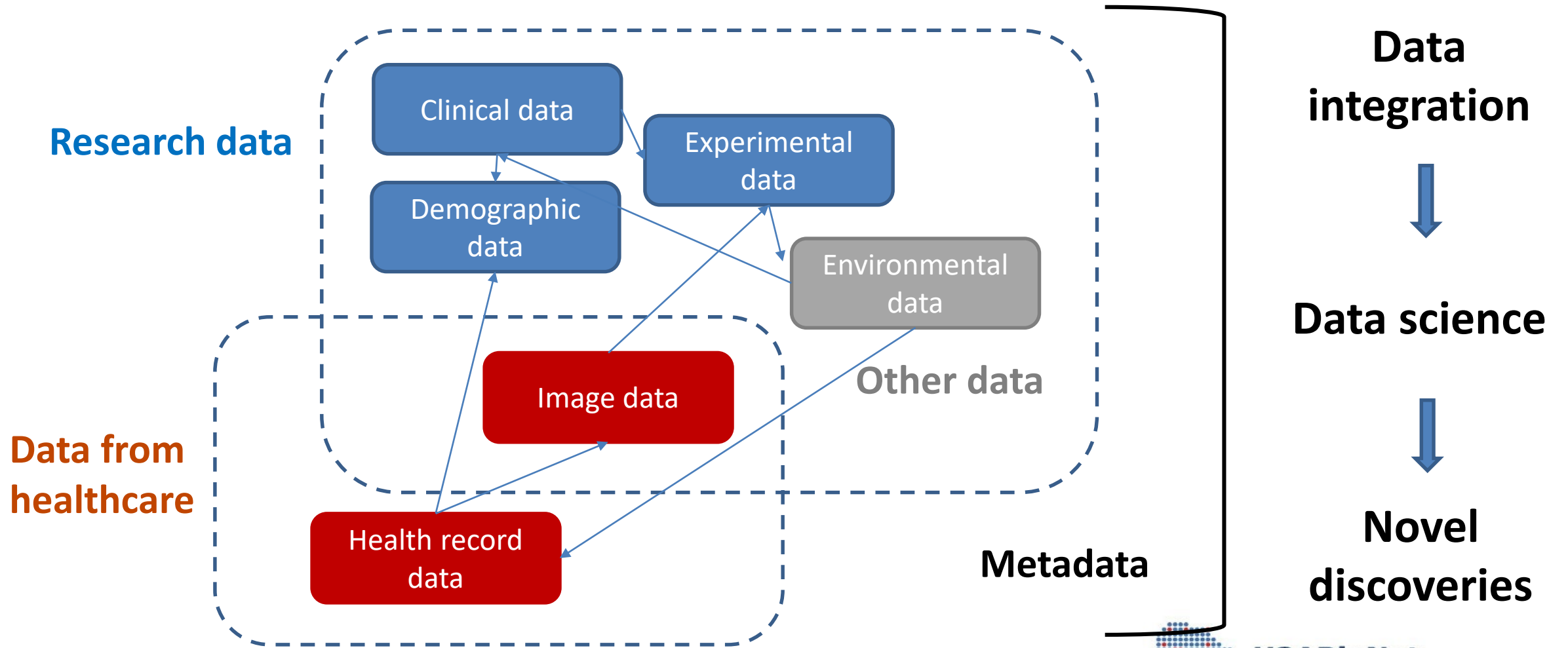
Prof Nicola Mulder

PI of H3ABioNet & eLwazi ODSP

Outline

- Data needs and challenges
- H3ABioNet capacity development
- Data Science capacity development
- Fostering engagement in public health

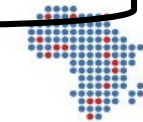# Data science for health
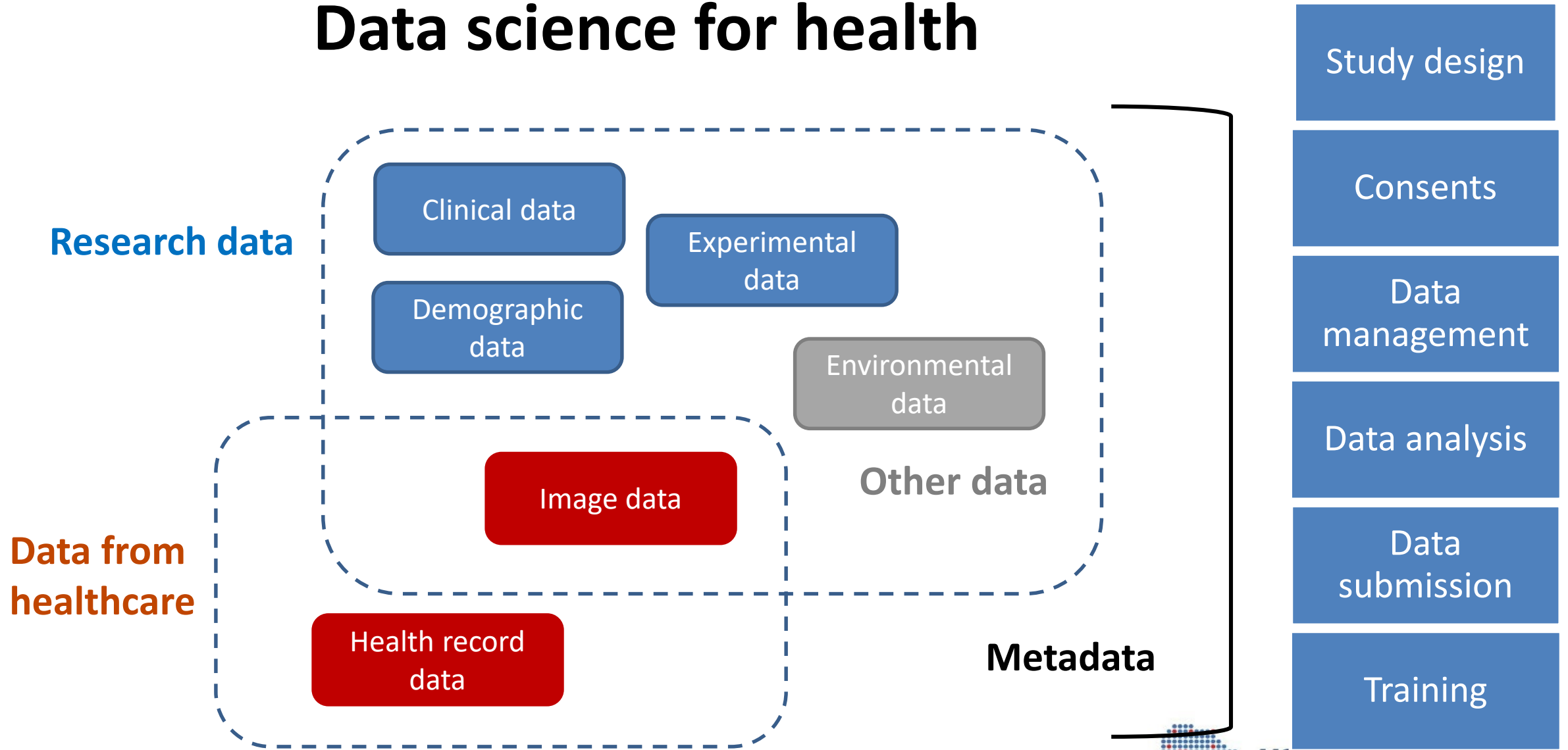
Human Heredity and Health in Africa (H3Africa)

Vision: to facilitate an African-based research approach to the study of genomic and environmental determinants of common diseases with the goal of improving the health of African populations

- ➢ **Capacity development** –human and research infrastructure
- ➢ Building **collaborative networks**
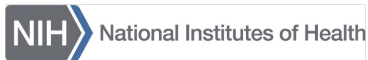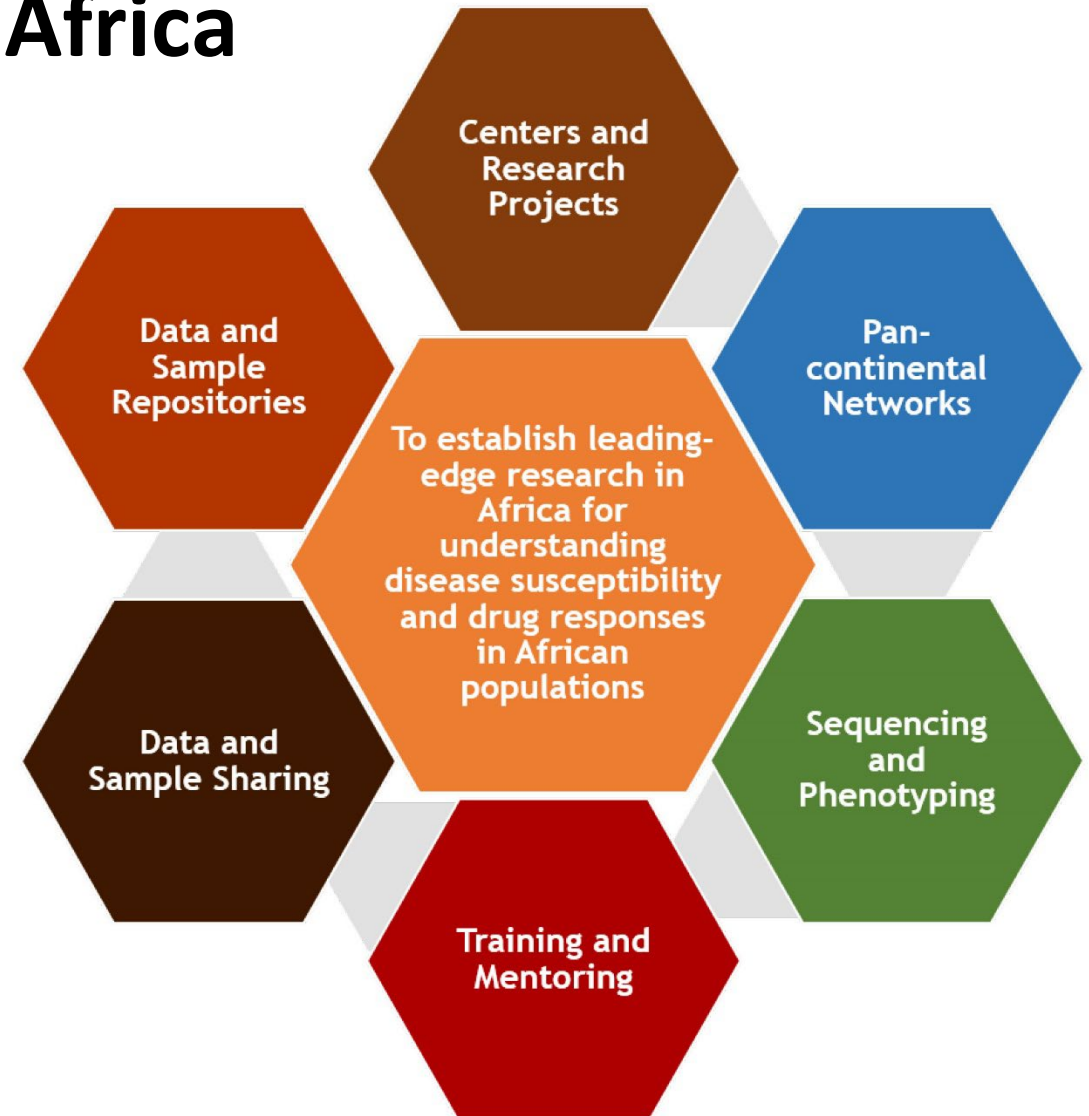
Bioinformatics/data needs

H3Africa data

Phenotype data

Genomics data

> **>100k participants**

Processed data

Computing & network infrastructure

Data management support

Data analysis tools and workflows

Data storage, submission and access

Training
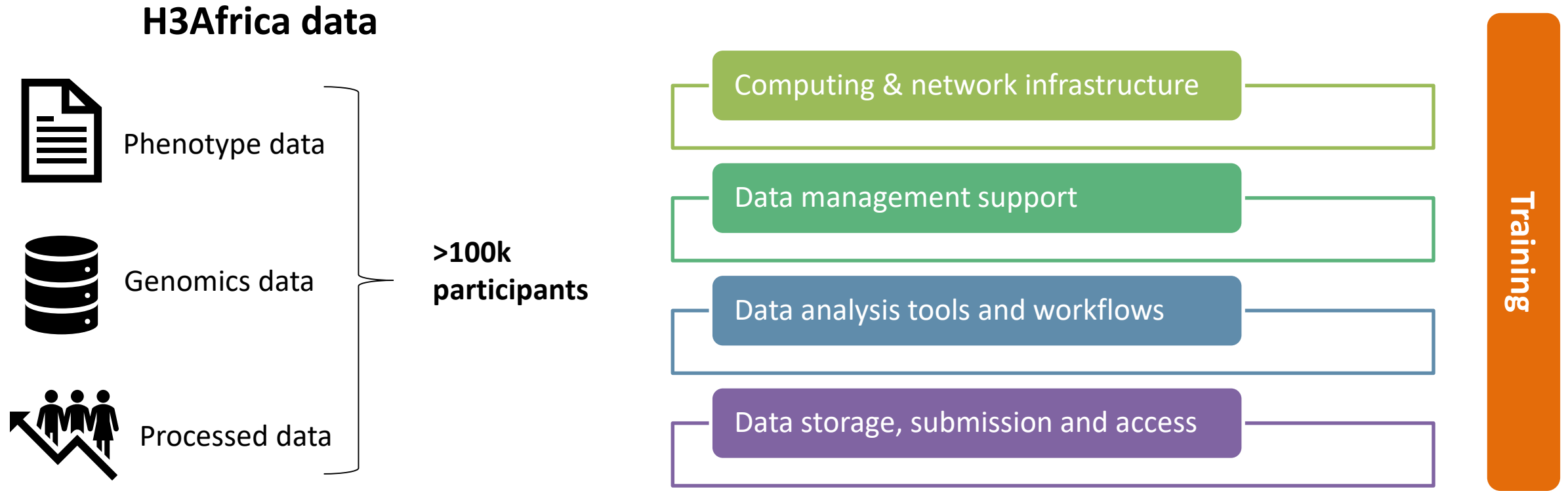
Needed to build infrastructure to enable and encourage large-scale genomics data management and analysis in Africa

H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet Informatics network

H3ABioNet is a Pan African Informatics Network, to provide bioinformatics infrastructure and support for the H3Africa consortium

Includes 28 partners in 17 countries, >200 members

Developed:

- Bioinformatics and data analysis skills

- Infrastructure for data transfer, analysis (compute, tools, workflows),

- Data management and harmonization tools
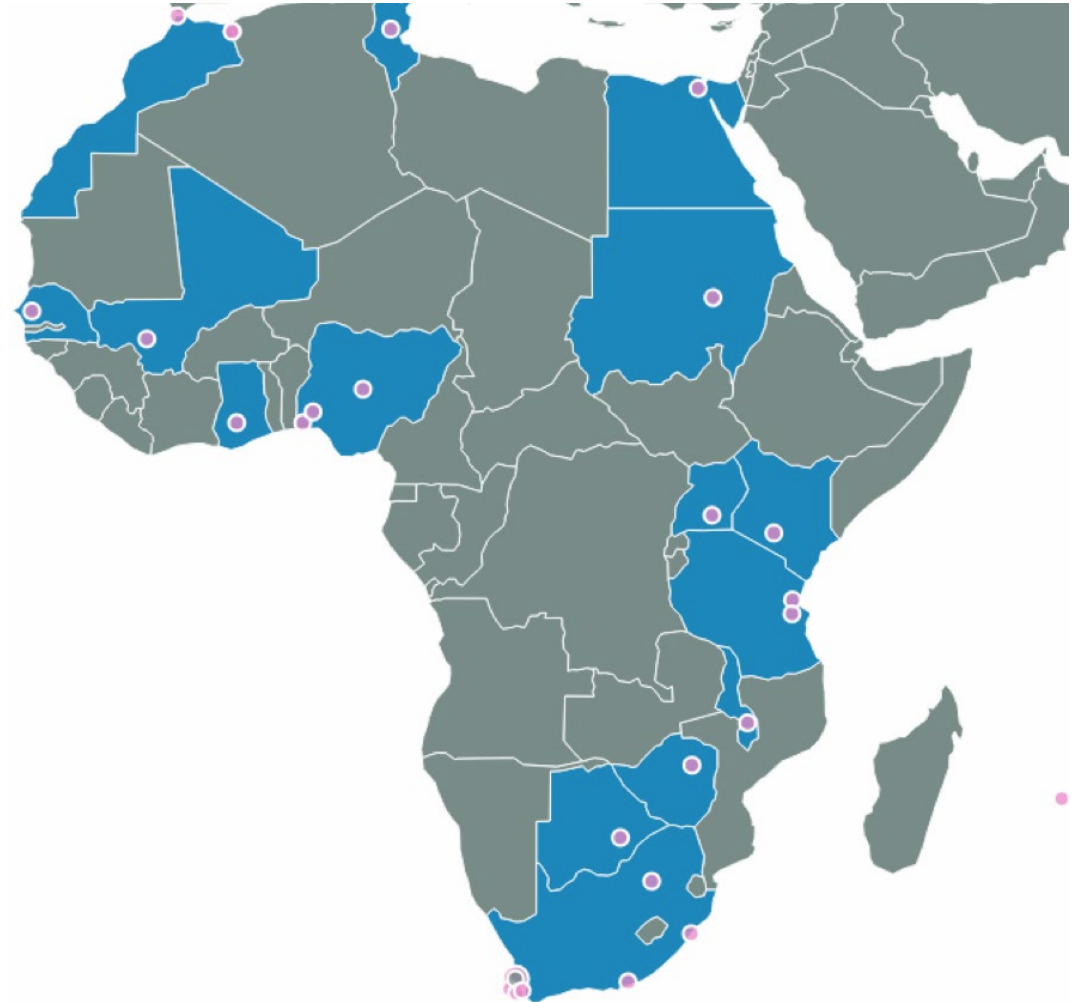
http://h3abionet.org

Role of data science in genomics

Data generation

Data collection
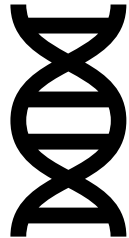
Data transfer, curation, storage

Data processing, analysis, interpretation

Data integration and translation

Data infrastructure	Data transfer protocols	Data standards	Data storage	Data security	Compute clusters	Analysis tools	Workflows

Data related skills	Data curation	Workflow development	Information technology	Data analysis	Data science	Biostatistics

Pan African Bioinformatics Network for H3Africa

Data Infrastructure –storage and computing
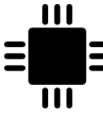
| Data infrastructure | Data transfer protocols | Data standards | Data storage | Data security | Compute clusters | Analysis tools | Workflows |

Compute Infrastructure for Genomics

- 3,348 Cores
- 8,872 RAM GB
- 3 Petabytes Storage
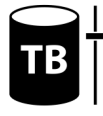- ~40 Million Core Compute hours used

2x40Gb/s SN2700 200Gb/s SN2700 200Gb/s SN2700

50 Gb/s per node

Ilifu 1   Ilifu 2   Ilifu 3   Ilifu 4

Multiple 10Gb/s   10Gb/s   6x10Gb/s   UCT Edge   40Gb/s per switch   10Gb/s

ARC Cluster   ARC Cluster   SANRen

Trained Sys Admins on basics and bioinformatics setup
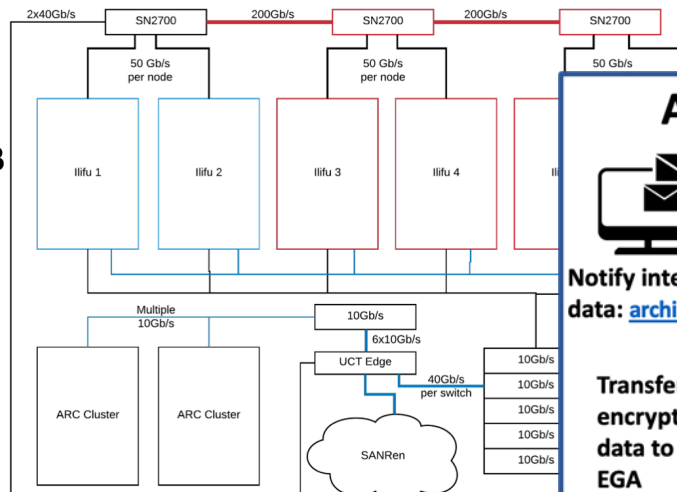
Archive for African Genom

Assist in preparing data for submission

Notify intent to submit data: archive@h3abionet.org

Register on Dashboard

Transfer encrypted data to EGA

european genome-phenome archive

Transfer encrypted data to the landing area

Move data to vault, decrypt and validate

Passes validation, prepare EGA XML files, re-encrypt and move to cold storage

- 14 African Genomic Datasets in the archive
- 134.9 Terabytes of African Genomics data
- 77.3 Terabytes transferred for storage at the EGA

SOFTWARE   Open Access

Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics

Ran hackathons to develop workflows

Workflows used for training workshops

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Data management

Data infrastructure	Data transfer protocols	Data standards	Data storage	Data security	Compute clusters	Analysis tools

Workflows

- Support for data management
- Developed std CRFs, ontology mapping

Harmonizing African Genomics data

PHWG has developed set of core phenotypes, standard CRF Extending these to disease specific stds

Mapping CRFs to ontologies, e.g. phenotype or disease ontology

Mapping ethics consent info to **Data Use** ontology

Mapping ethnic and demographic data to **ethnolinguistic** ontology

Mapping genomics data to **Experimental Factor** ontology

Mapping biobank data to **OMIABIS** ontology

REDCap
Log In

Biorepositories

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Training in REDCap and data harmonization

H3Africa Standard Case Report Form

Lifestyle Factors

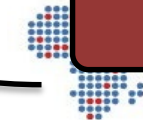Environmental Factors

Family History

Sickle Cell Disease

Kidney Disease

Mental Health

COVID-19

Pediatric options

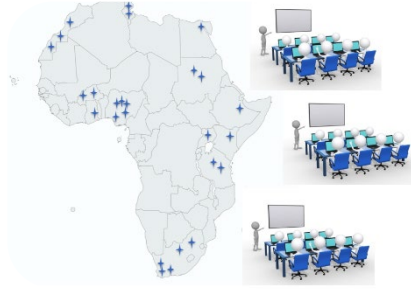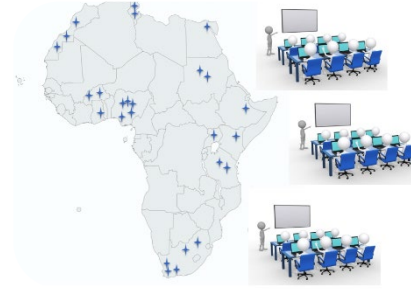Pan African Bioinformatics Network for H3Africa

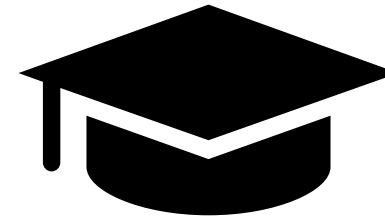# H3ABioNet human capacity development
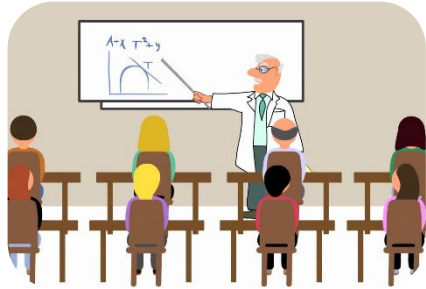
Hackathons and Courses

Online IBT course

Intermediate online course

Data analysis accreditation

Developed foundational data-related skills

Software and Data Carpentry

Competencies mapping

Train the trainer program

Developed applied data analysis skills

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Increasing the reach: live online remote classroom training

- Local classrooms with teaching assistants (trained)
- Virtual classrooms
- pre-recorded lectures compiled by expert trainers
- Hands on practicals and assignments

Materials are open access, lecture on YouTube
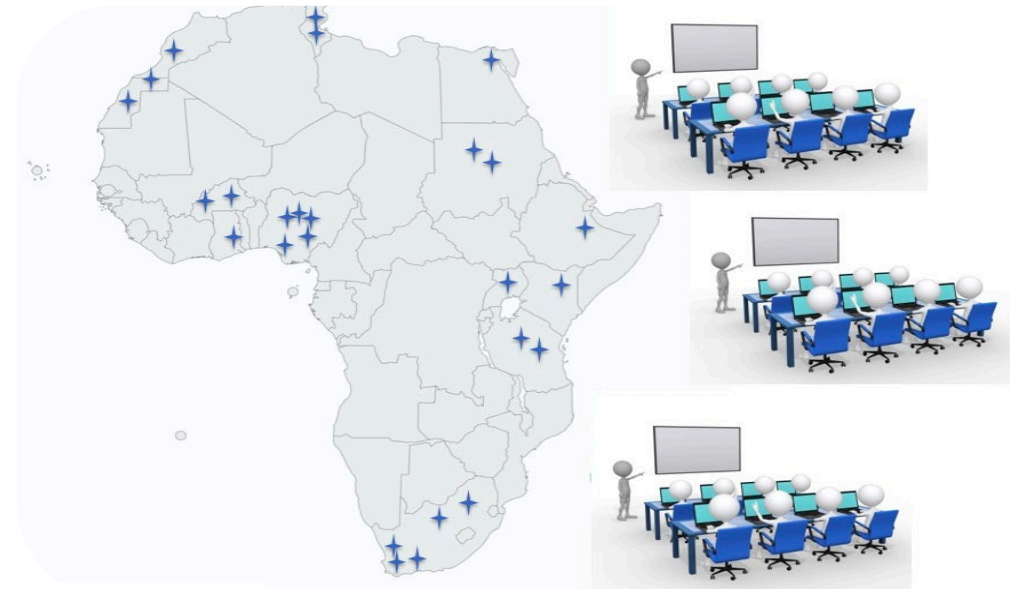
H3ABioNet IBT 2018 ▶ PLAY ALL

11:15	11:25	9:06	15:02	4:58
Staff training day 1 part 1	Staff training Day 1 Part 2	Staff training Day 1 Part 3	Staff training Day 1 Part 4	Staff training Day 1 Part 5
H3ABioNet H3ABioNet	H3ABioNet H3ABioNet	H3ABioNet H3ABioNet	H3ABioNet H3ABioNet	H3ABioNet H3ABioNet
2.9K views • 10 months ago	286 views • 10 months ago	259 views • 10 months ago	179 views • 10 months ago	140 views • 10 months ago

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Who have we trained and how?

| Trainer | User | Scientist | Engineer | Sys Admin |

Trainer
- IBT and IntBT
- Train-the-Trainer
- Carpentries
- Internships
- Career/Professional Development

User
- IBT and IntBT
- Carpentries
- F2F Courses
- Internships
- Internships
- Career/Professional Development

Scientist
- IntBT
- Carpentries
- F2F courses
- Hackathons
- Internships
- Career/Professional Development

Engineer
- IntBT
- F2F Courses
- Hackathons
- Internships
- Career/Professional Development

Sys Admin
- IBT and IntBT
- F2F Courses
- Hackathons
- Internships
- Career/Professional Development

H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet training approach outputs
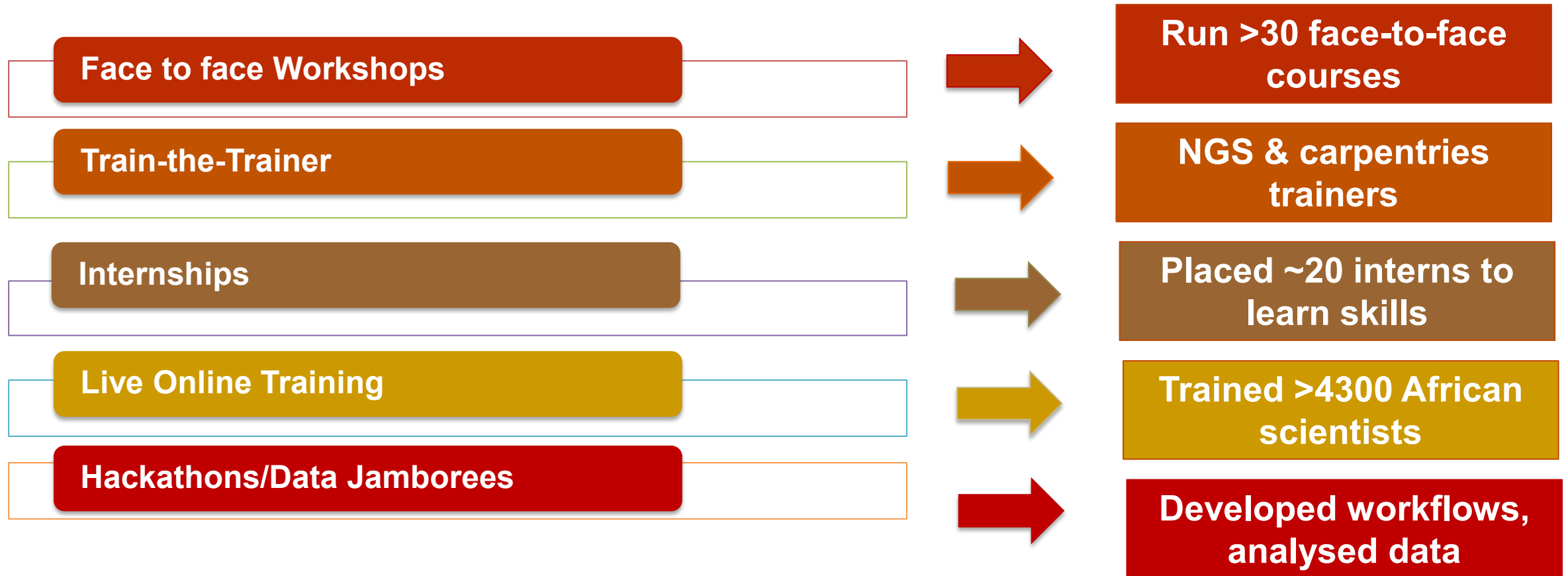
Approach	Output
Face to face Workshops	**Run >30 face-to-face courses**
Train-the-Trainer	**NGS & carpentries trainers**
Internships	**Placed ~20 interns to learn skills**
Live Online Training	**Trained >4300 African scientists**
Hackathons/Data Jamborees	**Developed workflows, analysed data**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

NIH DS-I Africa

- Harnessing Data Science for Health Discovery and Innovation in Africa
  - **Research Hubs:** Advance and demonstrate feasibility of data science research and innovation to improve health in Africa
  - **Training:** Increase capacity for data science research in Africa
  - **ELSI Research:** Explore Ethical, Legal, and Social Implications of data science research from an African perspective and contribute to policy discussion on the continent
  - **Open Data Science Platform & Coordination Center:** Facilitate the development of a trans-African network of data scientists

DS-I Africa capacity development & engagement

- **Data types are diverse**, need to manage data sharing, access, harmonization, linkage, integration and data science analysis

- Infrastructure: develop **African Open Data Science Platform** (eLwazi) to enable the implementation of data science for health, that is relevant to the African context

- **Engagement**: facilitate the development of a trans-African network of data scientists

- **Human capacity development**: train diverse audiences on data science techniques

H3ABioNet
Pan African Bioinformatics Network for H3Africa

eLwazi platform

Project 1 data → Project 2 data → Project 3 data

Platform

Metadata Catalogue

Cloud

Welcome to eLwazi

The eLwazi Open Data Science Platform is a project powered by Terra for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

Find how-to's, documentation, video tutorials, and discussion forums ✑

View Workspaces
Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely. →

View Examples
Browse our gallery of showcase Workspaces to see how science gets done. →

Browse Data
Access data from a rich ecosystem of data portals. →

eLwazi POWERED BY | Privacy Policy | Terms of Service | Security ✑ | Documentation ✑

Hosting sites across Africa

Workspaces bring together metadata, data, and compute environments through a web-based user interface. Access for beginner and advanced users

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Log into project work space

↓

Identify datasets from registry

↓

Extract relevant data

↓

Select tools from registry

↓

Run in Cloud
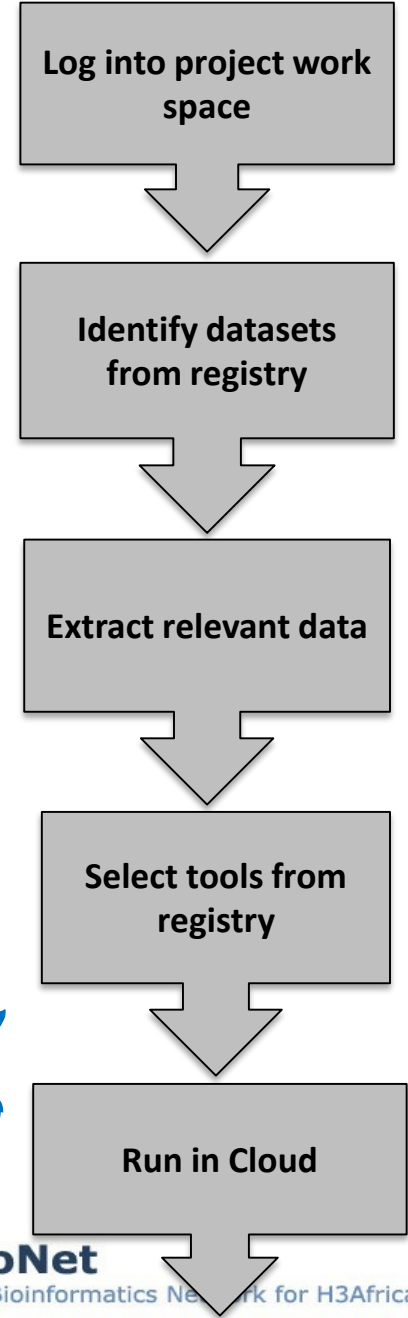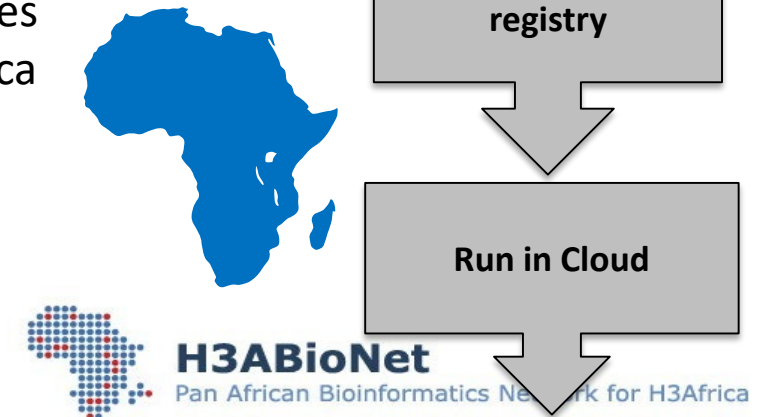
Data science training

- Training grants are developing data science degrees with specific priority focus areas
- Working on developing/collecting data science for biomedical sciences competencies
- Will map these to degree curricula
- Will develop curriculum for an online data science short course
- Specific training on use of eLwazi and components
- Focus on training fellows to develop next generation of research leaders

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Public health training

- Working with Africa CDC to train public health institutes in the use of NGS for pathogen surveillance

Training
- Practical training for wet lab
- Bioinformatics training

Curriculum development
- Standardized curriculum
- Modules –foundational & applied

Stakeholder engagement
- Landscaping exercise
- Trainer communities (Trainer DB), trainee support (helpdesk)

- Building underlying foundational skills
- Building trainer communities with better coordination
- Aiming for long term support for trainees
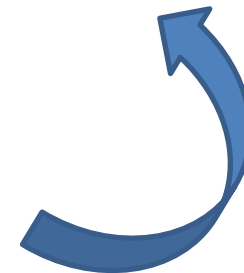
H3ABioNet
Pan African Bioinformatics Network for H3Africa

Take home messages

- Significant capacity development has occurred in bioinformatics and more recently, data science –academia and public health
- These have been done through **funded** pan African networks
- Networks have engaged with African and international efforts
- Once built, networks can be leveraged
- Interest in maintaining the networks post-funding –how?
- Discussions about an African Bioinformatics Institute or African Data Science Institute to continue to foster engagement and capacity development
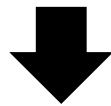
Skills and techniques can be applied to many areas of biomedical and other life sciences

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Acknowledgements

H3ABioNet

- **H3ABioNet Central Team**
- **WP chairs and co-chairs**
  - **Training**: Nicola Mulder, Shaun Aron
  - **Health Informatics**: Nicki Tiffin, Katherine Johnston
  - **Data & Standards**: Judit Kumuthini, Raphael Sangeda
  - **Pipelines & Computing**: Chris Fields, Shakuntala Baichoo
  - **Tools & Web Services**: Scott Hazelhurst, Itunu Isewon
  - **Databases & Resources**: Faisal Fadlemola, Kais Ghedira
  - **Data Integration**: Cheikh Loucoubar
  - **Outreach & Sustainability**: Jonathan Kayondo, Amel Ghouila

ODSP Partners

- University of Cape Town
- University of the Witwatersrand, PI Scott Hazelhurst
- University of the Western Cape/Ilifu, PIs Mattia Vaccari, Rob Simmonds, Russ Taylor
- University of Stellenbosch, PI Tulio de Oliviera
- University of Mauritius, PI Shakuntala Baichoo
- Uganda Virus Research Institute, PIs Jonathan Kayondo, Daudi Djjingo
- University of Khartoum, PI Faisal Fadlemola
- USTTB, University of Bamako, PI Seydou Doumbia
- Broad Institute, PI Brian O'Connor
- EMBL European Bioinformatics Institute, PI Tony Burdett
- University of California, Santa Cruz, PI Benedict Paten
- University of Chicago, PI Robert Grossman

H3ABioNet
Pan African Bioinformatics Network for H3Africa