Session 3: Data science for public health tools

Kirsty Lee Garson, Fatma Guerfali, Aquillah Kanzi, Amadou Diallo

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

wellcome connecting science

COG-TRAIN
COVID-19 GENOMICS GLOBAL TRAINING

Session outline

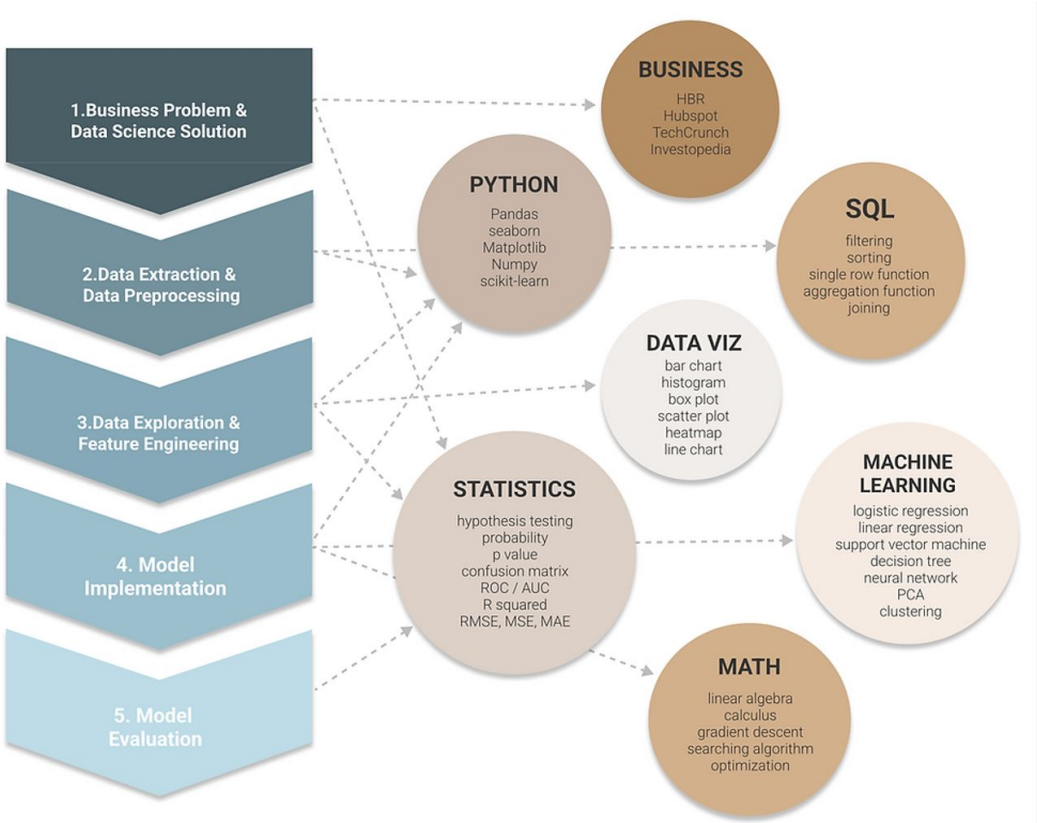
- Key components of data science, the role of data science in public health
- Why is sharing data and ensuring that data is FAIR so crucial in public health?
    - The growth in magnitude and complexity of disease surveillance data
    - Generating findable, accessible, interoperable and reusable data
- Public repositories, community standards and other resources which facilitate shared progress
- Applying open science principles in training and capacity development
- Group activity: scenarios for discussion

Session outcomes

At the end of this session, participants will be able to:

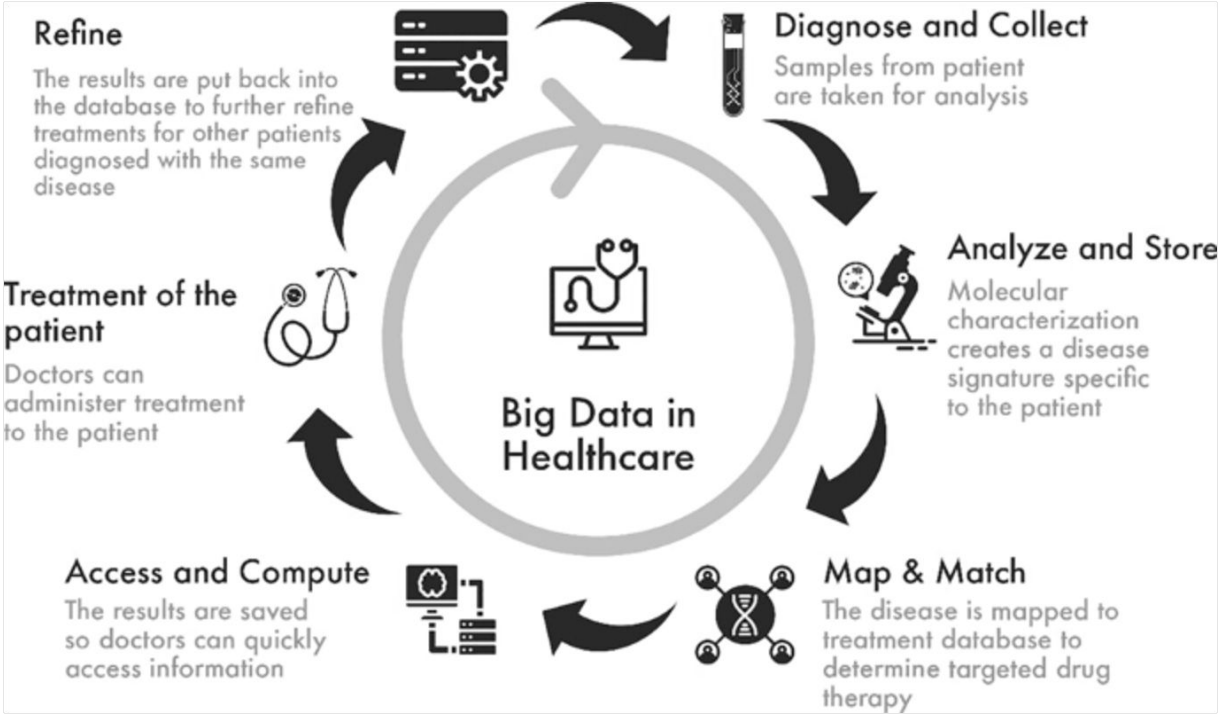
- Understand the principles of open science and its capacity to enable higher quality, more equitable knowledge generation
- Describe the elements of FAIR data and the importance of ensuring that public health data can be accessed and integrated timeously
- Recognise the value of applying and sharing best practices, established protocols, and code/workflows
- Describe some of the challenges and considerations involved in data sharing
- Identify ways to improve datasets by locating and applying appropriate standards for metadata

Key components of Data Science



https://www.visual-design.net/post/how-to-self-learn-data-science

Key components of Data Science in Public Health



Subrahmanya et al. *Ir J Med Sci* **191**, 1473–1483 (2022).

Approaches to working with data in pathogen surveillance

In the past decade, the use of Next Generation Sequencing (NGS) has become increasingly applied in the areas of:

- Diagnostics
- Surveillance
- Research of infectious diseases

"Humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data."

The value of any disease surveillance dataset can't be fully realised apart from **current and historical data for the regional, national and global context**

The ability to quickly access and re-use data is **vital**

Findable

Accessible

Interoperable

Reproducible

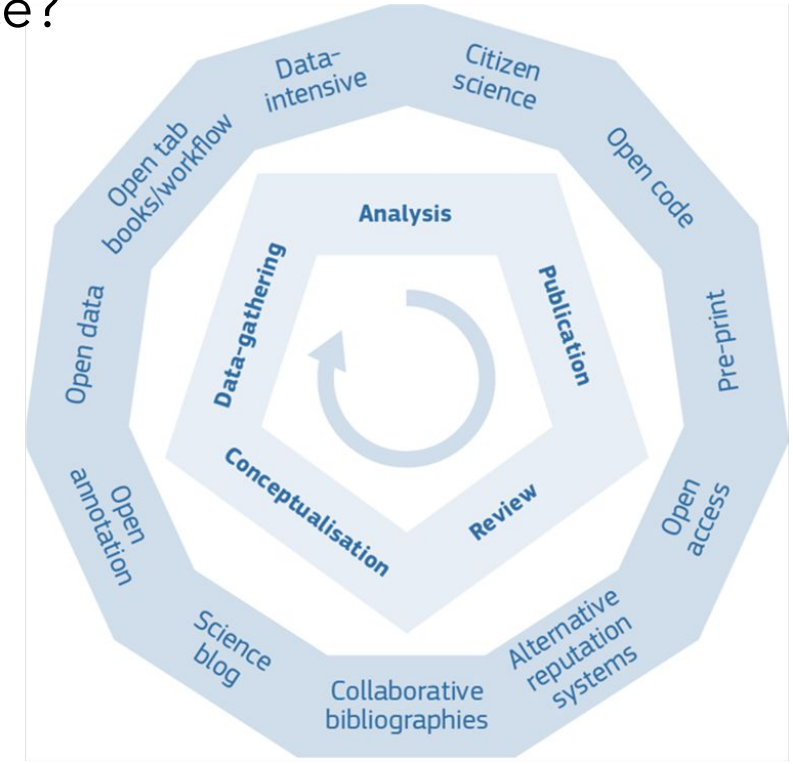
Kemp M and Maiers M (2022) Editorial: Application of Next Generation Sequencing (NGS) in Infection Prevention. Front. Public Health 10:945595. doi: 10.3389/fpubh.2022.945595; https://www.go-fair.org/fair-principles/

Why Open Science?

Quality

Efficiency

Reproducibility



Democratizing access and participation

https://www.fosteropenscience.eu/learning/what-is-open-science/#/

Developing consensus on best practices, sharing protocols, making findings comparable across environments

protocols.io

A secure platform for developing and sharing reproducible methods

https://www.protocols.io/

Public Health Alliance for Genomic Epidemiology

Open. Consensus. Practice. Tools.

Improving Openness And Interoperability In Public Health Bioinformatics

A Global Coalition.

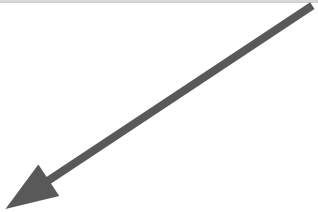
An international consortium working to develop shared standards in public health

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

The importance of informative and harmonizable metadata

PHA4GE SARS-CoV-2 Contextual Data Specification Package:
https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification



SARS-CoV-2 contextual data specification package

Spreadsheet-based (.xlsx) collection template

It contains the following items (tabs in the spreadsheet):

1. **a template for populating the complete set of contextual data;**

The collection template contains "required" (colour-coded yellow), "strongly recommended" (colour-coded purple) and "optional" (colour-coded white) fields.

2. **guidance for populating the template;**

The reference guide aims to facilitate the use of the collection template. It contains field definitions, further guidance/instructions, and examples of structured data.

3. **ontology-mapped controlled vocabulary for the picklists.**

Lists of controlled vocabulary, agreed upon by PHA4GE, are provided here for populating the template.

https://pha4ge.org/

The importance of informative and harmonizable metadata

Example of metadata standards for public health:

H3ABioNet Case Report Form Standards:
https://h3abionet.org/data-standards/phenotype-data-collection-standard

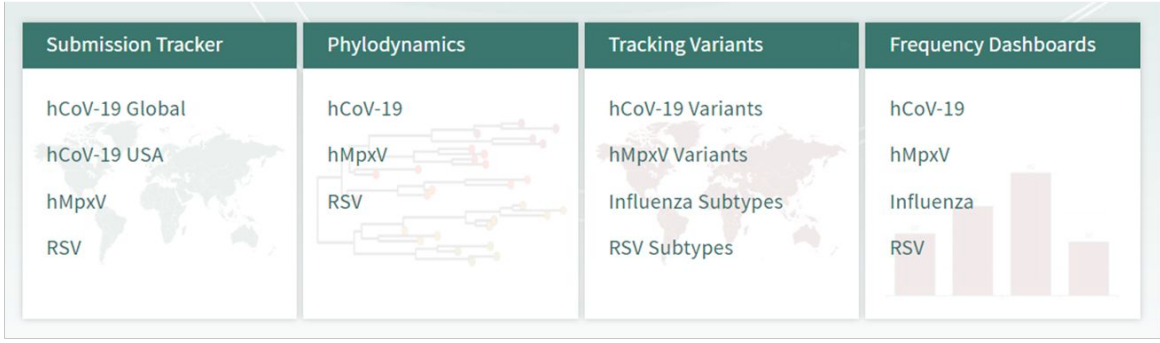
H3ABioNet

- COVID-19: REDCap Project Template; Data Dictionary; User Guideline

Data repositories and tools for disease surveillance data

"Rapid sharing of data from all influenza viruses and the coronavirus causing COVID-19"

Emphasis: "acknowledging the **Originating Laboratories** providing the specimens, and the **Submitting Laboratories** generating sequence and other metadata"



https://gisaid.org/

Data repositories and tools for disease surveillance data

International Nucleotide Sequence Database Collaboration (INSDC)

NCBI

ENA European Nucleotide Archive

DDBJ DNA Data Bank of Japan

Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

https://www.insdc.org

Additional Resources for Data Science in Public Health

ViralAI, a global network for genomic surveillance and infectious disease research
https://viral.ai/collections

European COVID-19 Data Portal
https://www.covid19dataportal.org/

eLwazi Open Data Science Platform
https://www.elwazi.org/

The value of training materials which are FAIR

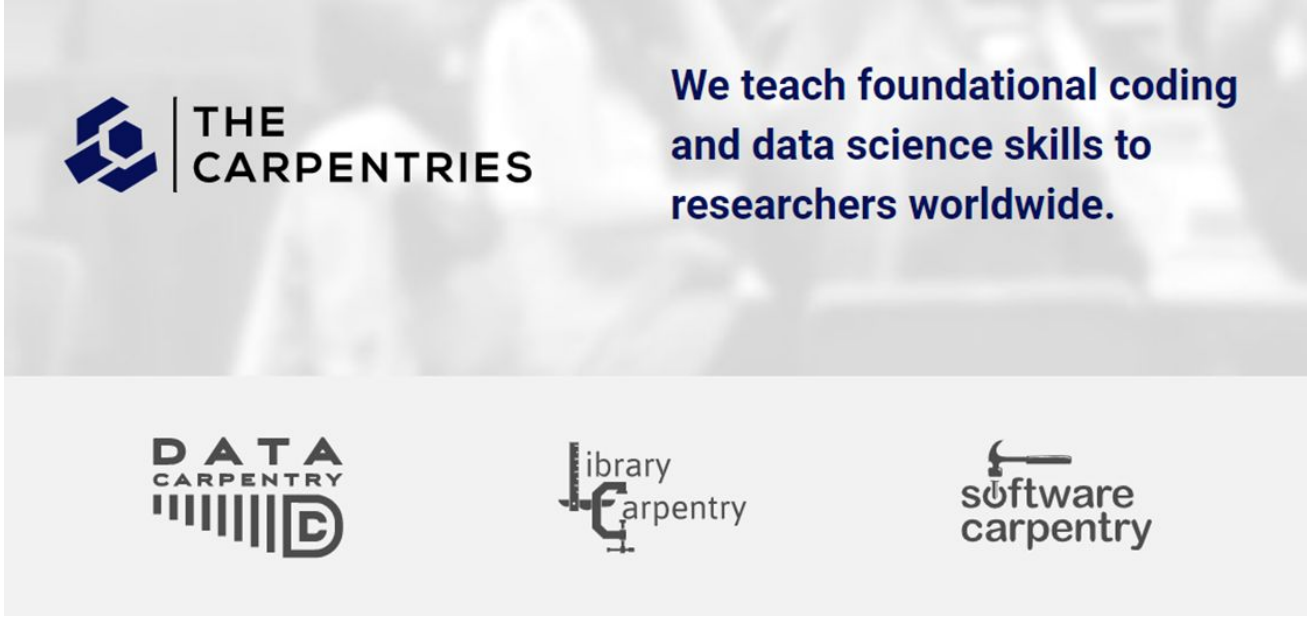
https://uct-cbio.github.io/ngs-academy/courses

https://redcap.h3abionet.org/redcap/surveys/?s=FCA9MTKHPH

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

The value of training materials which are FAIR



https://carpentries.org/

Africa PGI Data Management and Exchange Platform

Dr Gerald Mboowa

Africa Centres for Disease Control and Prevention

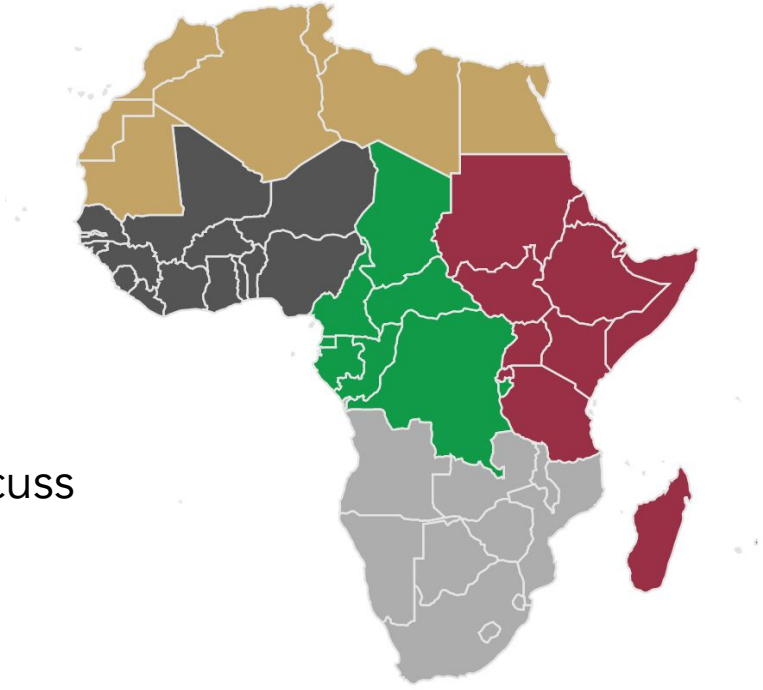
Christoffels, A *et al*. **A pan-African pathogen genomics data sharing platform to support disease outbreaks.** *Nat Med* (2023)**.**

https://doi.org/10.1038/s41591-023-02266-y

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Group activity



Work with participants from your regional context to discuss the following questions

Group activity

Discussion:

- What concerns and considerations shape how, whether and where we share data? (consider the differences between sharing complete metadata vs. selected metadata variables vs. raw reads etc.)
- How can we work to facilitate more effective data sharing in our local contexts?
- How can we model, practice and negotiate for changes to facilitate open science practices in resource-limited environments?