EBNet
Environmental Biotechnology

Why Bioinformatics Training is Important

EBNet Working Group: Bioinformatics Training for Microbial Environmental Biotechnologies

11th May

Outline of this seminar

EBNet
Environmental Biotechnology

Introduction by Working Group lead James Chong

Giacomo Peru, Ed-DaSH - The University of Edinburgh Data Science training in Health and Bioscience

Xenia Perez Sitja, DASH/ELIXIR training and the data stewardships

Emma Rand, CloudSPAN - Cloud-based High Performance Computing for SPecialised ANalyses on environmental ‘omics

Sarah Forrester, Software Sustainability Institute - Developing metagenomic bioinformatics training materials

Annabel Cansdale, Why Bioinformatics training is important - An eight terabyte case study

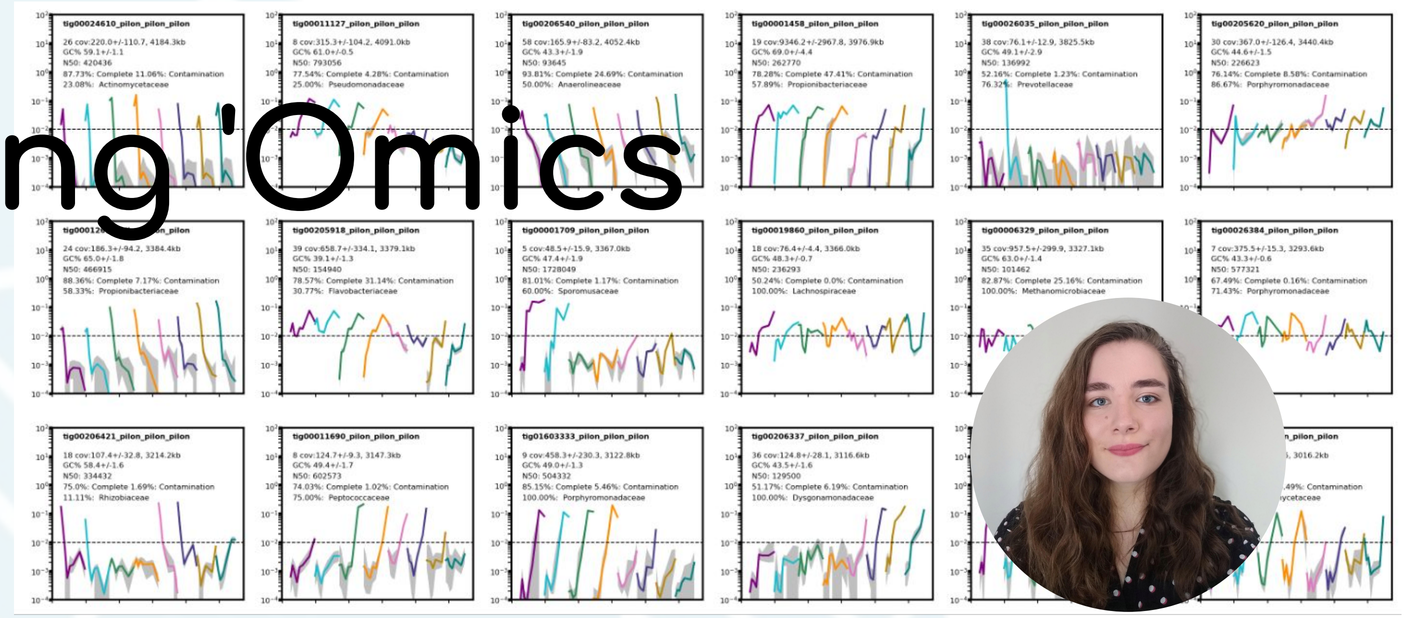
Evelyn Greeves, FAIR- Why making data science reproducible is important

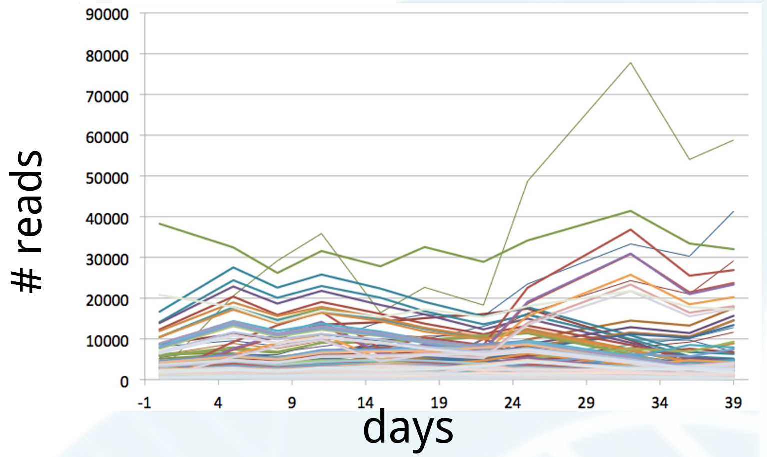
Panel with all speakers, we will have time for lots of questions so save these for this section

Why Bioinformatics Training is Important

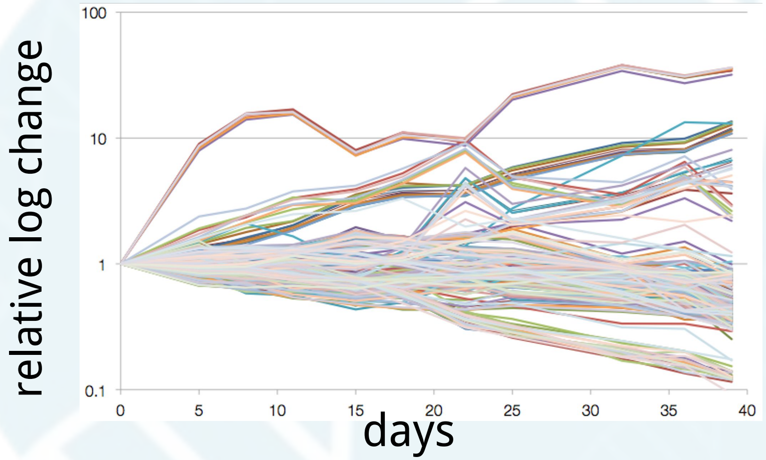
James Chong

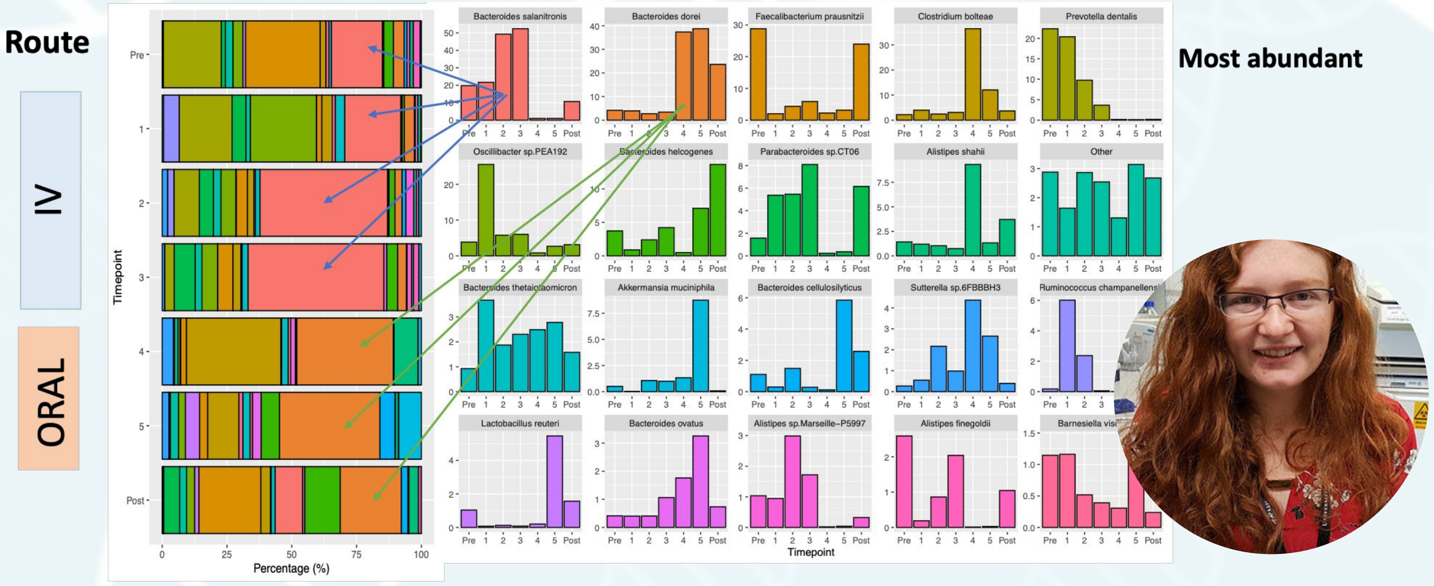
Analysing 'Omics





Annabel Cansdale



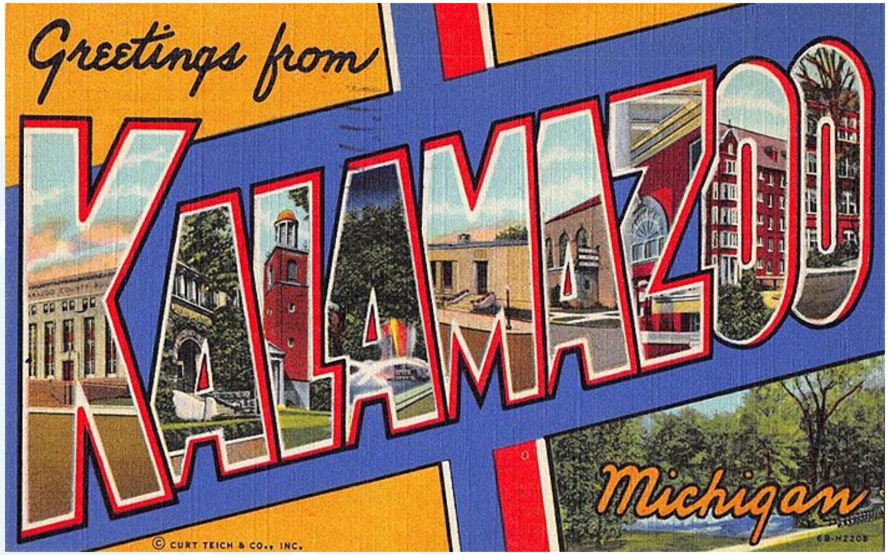


Sarah Forrester

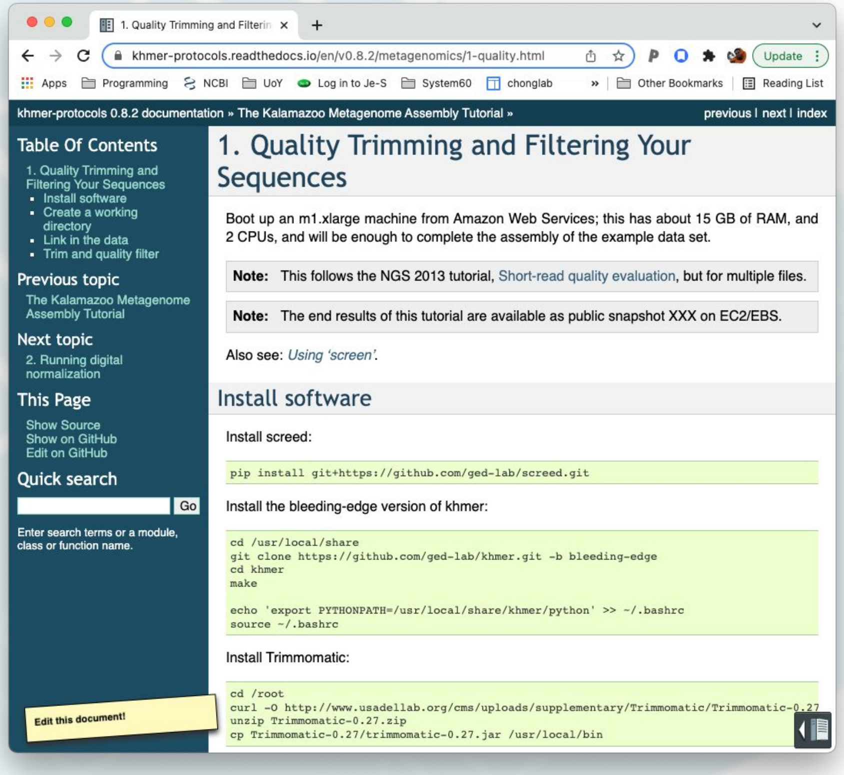
Challenges in Environmental 'Omics

Software	Operating system / scheduler / permissions
Hardware	CPU / GPU / RAM / storage may all be "limited"
Skills	Installing software / specifying resources / research software support
Time	Brain time / wall time / result time

Software



The Kalamazoo Metagenome Assembly Tutorial (experience)



khmer-protocols 0.8.2 documentation » The Kalamazoo Metagenome Assembly Tutorial »

previous | next | index

Table Of Contents

1. Quality Trimming and Filtering Your Sequences
   - Install software
   - Create a working directory
   - Link in the data
   - Trim and quality filter

Previous topic

The Kalamazoo Metagenome Assembly Tutorial

Next topic

2. Running digital normalization

This Page

Show Source
Show on GitHub
Edit on GitHub

Quick search

Enter search terms or a module, class or function name.

1. Quality Trimming and Filtering Your Sequences

Boot up an m1.xlarge machine from Amazon Web Services; this has about 15 GB of RAM, and 2 CPUs, and will be enough to complete the assembly of the example data set.

Note: This follows the NGS 2013 tutorial, Short-read quality evaluation, but for multiple files.

Note: The end results of this tutorial are available as public snapshot XXX on EC2/EBS.

Also see: *Using 'screen'*.

Install software

Install screed:

```
pip install git+https://github.com/ged-lab/screed.git
```

Install the bleeding-edge version of khmer:

```
cd /usr/local/share
git clone https://github.com/ged-lab/khmer.git -b bleeding-edge
cd khmer
make

echo 'export PYTHONPATH=/usr/local/share/khmer/python' >> ~/.bashrc
source ~/.bashrc
```

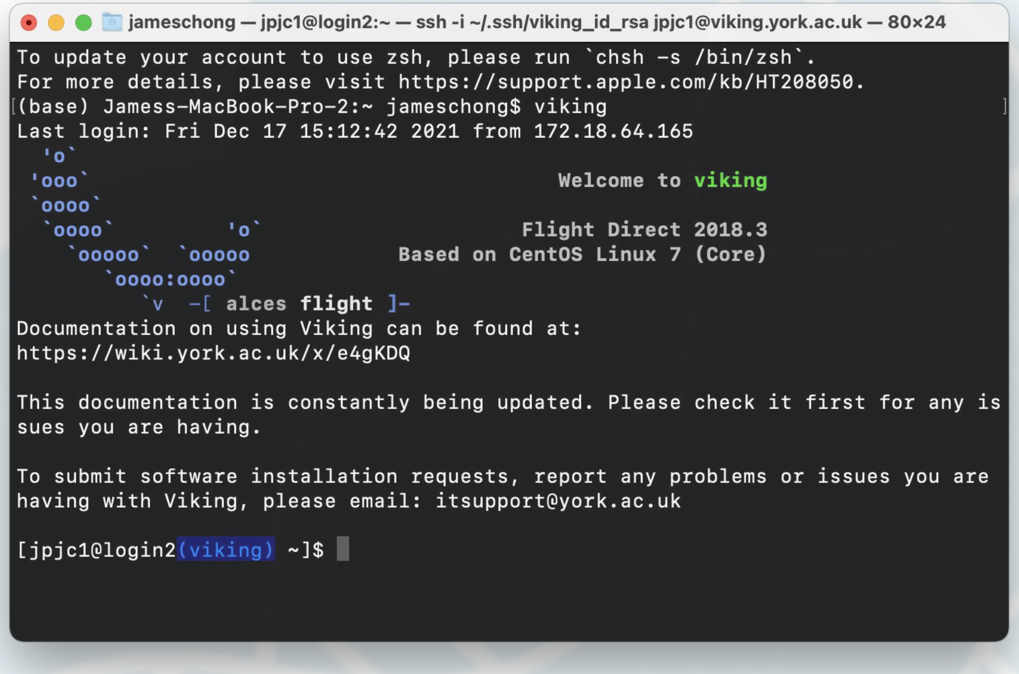
Install Trimmomatic:

```
cd /root
curl -O http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.27
unzip Trimmomatic-0.27.zip
cp Trimmomatic-0.27/trimmomatic-0.27.jar /usr/local/bin
```

Operating system / permissions

Hardware

Compute resources



jameschong — jpjc1@login2:~ — ssh -i ~/.ssh/viking_id_rsa jpjc1@viking.york.ac.uk — 80×24

```
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Jamess-MacBook-Pro-2:~ jameschong$ viking
Last login: Fri Dec 17 15:12:42 2021 from 172.18.64.165
  'o`
 'ooo`                                      Welcome to viking
 `oooo`
  `oooo`         'o`                     Flight Direct 2018.3
    `ooooo`  `ooooo            Based on CentOS Linux 7 (Core)
       `oooo:oooo`
          `v   -[ alces flight ]-
Documentation on using Viking can be found at:
https://wiki.york.ac.uk/x/e4gKDQ

This documentation is constantly being updated. Please check it first for any is
sues you are having.

To submit software installation requests, report any problems or issues you are
having with Viking, please email: itsupport@york.ac.uk

[jpjc1@login2(viking) ~]$
```

York

c2d2

YARCC

Viking

N8

Archer

Bede

National Cloud

CLIMB

CLIMB-BIG-DATA

AWS

Google

Azure

Skills and knowledge

Research computer literacy

Unix command line
Shell scripts
Resources / queues / scratch
Package installation (Conda / Mamba / PIP / modules)
Dependencies and updates / containers
Coding (Python / Snakemake)

Biomedical pipelines
Bespoke analysis

Time

Brain time	thinking, coding, problem solving
Queue time	extends trouble-shooting, queuing priorities
Wall time	resource quantity and availability
Result time	visualisation, interpretation, parameter optimization

Edinburgh Data Science Training for Healthcare & Biosciences

Ed-DasH

School of Biological Sciences
School of Mathematics
College of Medicine and Veterinary Medicine

Edinburgh Carpentries

Ed-DaSH Training Programme

Objective: to develop and deliver data science training using The Carpentries methodology.

Topics:

- Computational workflows: Conda, Nextflow/Snakemake
- Open science, FAIR principles, and data management:
  - Hands on Open Science, FAIR principles, and data management
- Statistics:
  - Basic and intermediate statistical skills
  - High dimensional statistics
  - Machine learning

Computational Workflows

Introduction to Conda

https://edcarp.github.io/2022-08-31_ed-dash_conda/

Workflows with Snakemake

https://edcarp.github.io/2022-09-06_ed-dash_workflows-snakemake/

Workflows with Nextflow

https://edcarp.github.io/2022-05-31_ed-dash_workflows-nextflow/

Stats & Machine Learning

Introduction to Statistics with R

https://edcarp.github.io/2022-05-03_ed-dash_intro-statistics/

High-Dimensional Statistics with R

https://edcarp.github.io/2022-05-17_ed-dash_high-dim-stats/

Introduction to Machine Learning with Python

https://edcarp.github.io/2022-05-24_ed-dash_machine-learning/

FAIR Data Management

FAIR in (Biological) Practice

https://edcarp.github.io/2022-06-14_ed-dash_fair-bio-practice/

The Carpentries

Mission: The Carpentries builds global capacity in essential data and computational skills for conducting efficient, open, and reproducible research. We train and foster an active, inclusive, diverse community of learners and instructors that promotes and models the importance of software and data in research.

Acknowledgements

Co-Investigators

Alison Meynert, MRC Human Genetics Unit

Alex Twyford, School of Biological Sciences

Catalina Vallejos, MRC Human Genetics Unit

Edward Wallace, School of Biological Sciences

Programme Coordinator

Giacomo Peru, EPCC

Administrative & Finance Support

MRC Institute of Genetics and Cancer

Steering Group

Chair: Neil Chue Hong, Software Sustainability Institute

Nicola Cuthbert, Institute of Academic Development

Karin Halliday, Dean of Systematic Inclusion, CSE

Malcolm McLeod, Academic Lead for Research Improvement and Research Integrity

Mick Watson, Roslin Institute

Development teams

Computational workflows

Tim Booth, Edinburgh Genomics

Graeme Grimes, MRC Human Genetics Unit

Nathan Medd, Edinburgh Genomics

Flic Anderson, School of Biological Sciences

Open science, FAIR principles & data management

Andrew Millar, School of Biological Sciences

Andrew Romanowski, School of Biological Sciences

Tomasz Zielinski, School of Biological Sciences

Statistics

Ailith Ewing, MRC Human Genetics Unit

Alan O’Callaghan, MRC Human Genetics Unit

Gail Robertson, Statistical Consultancy Unit, School of Mathematics

School of Biological Sciences
School of Mathematics
College of Medicine and Veterinary Medicine

MRC Human Genetics Unit

THE CARPENTRIES

Edinburgh Carpentries

EIDF Edinburgh International Data Facility

ELIXIR-UK DaSH: A Fellowship of Data Steward Ambassadors

Xènia Pérez Sitjà

Data Stewardship Community Manager
Faculty of Life Sciences
Univeristy of Bradford (ELIXIR-Uk)

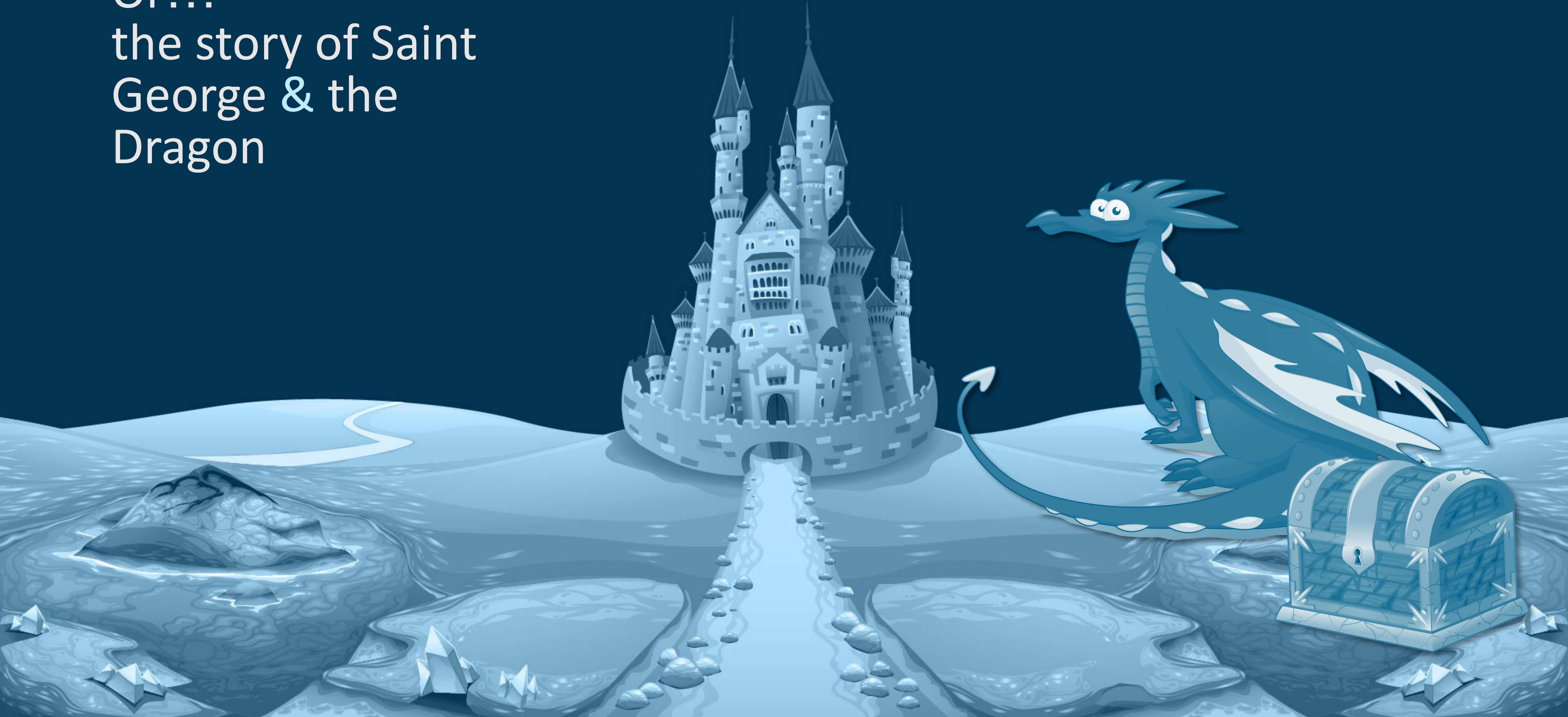
UKRI Medical Research Council

Funded by the ELIXIR-UK: FAIR Data Stewardship training UKRI award (MR/V038966/1)

elixir UNITED KINGDOM

UNIVERSITY OF BRADFORD

Or…
the story of Saint George & the Dragon



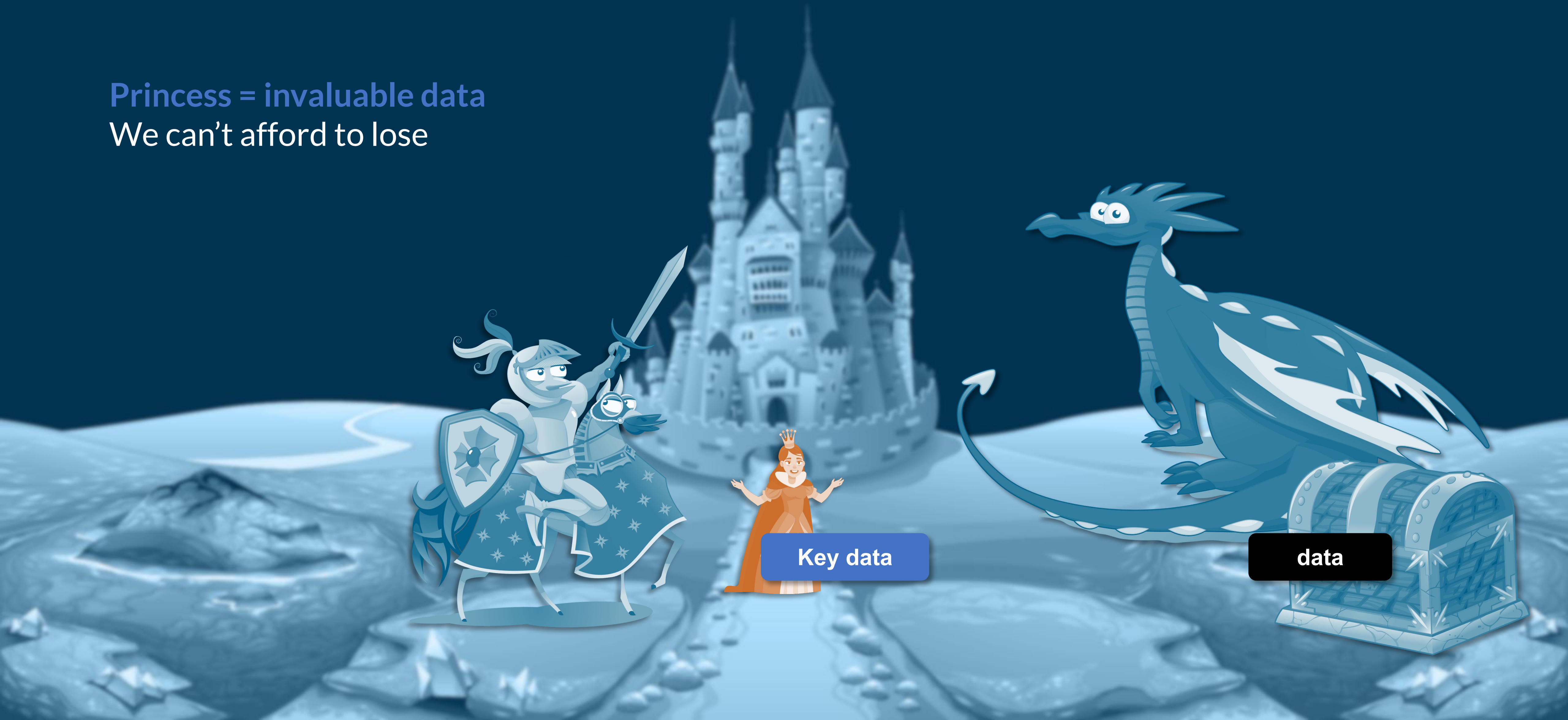
Castle = your organisation

Tribute = data

Data from projects, research…

Princess = invaluable data

We can’t afford to lose



Dragon = hurdles

Time, funding, lack of capacity, skills, organisation…

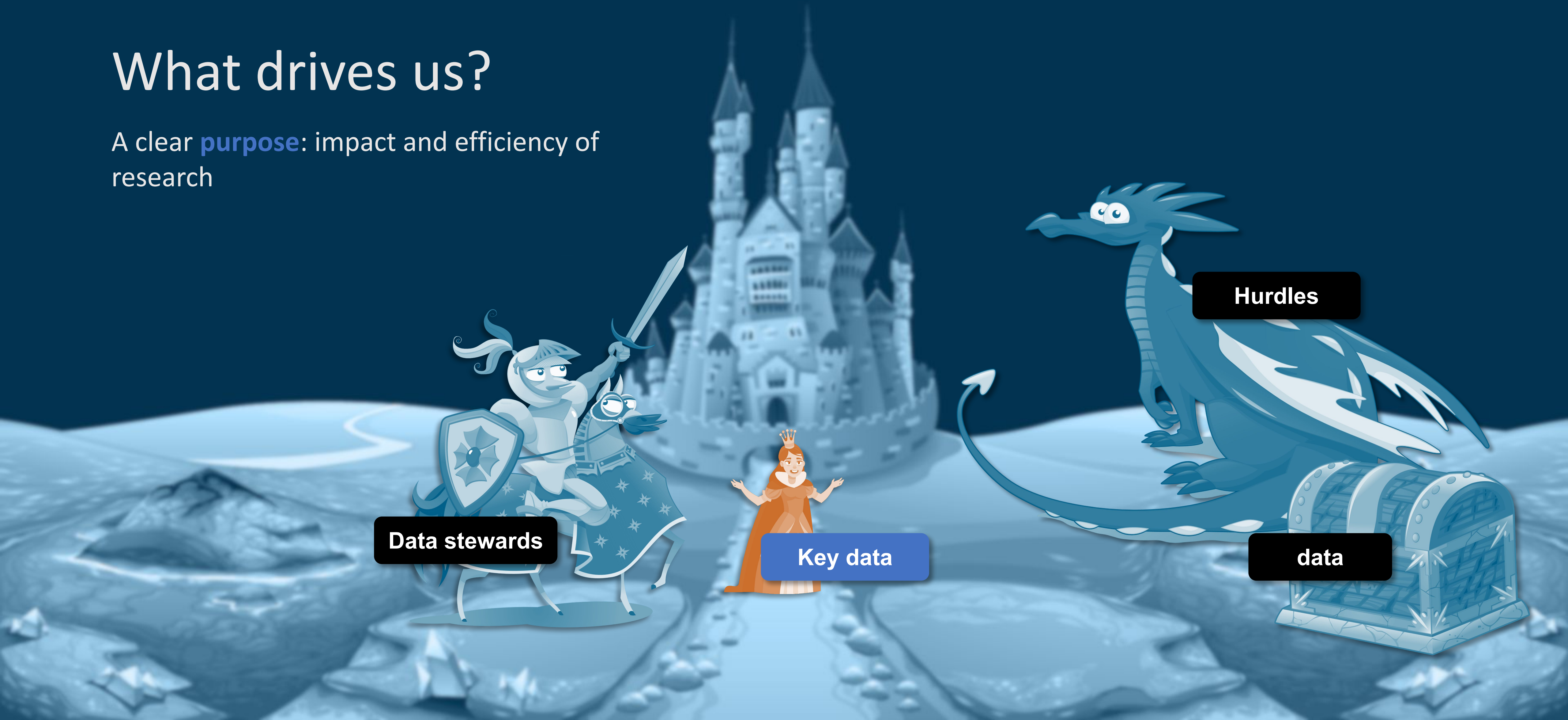
Saint George = data stewards

Those skilled people that can kill the dragon and take the data back to make it reusable

Hurdles
Data stewards
Key data
data

What drives us?

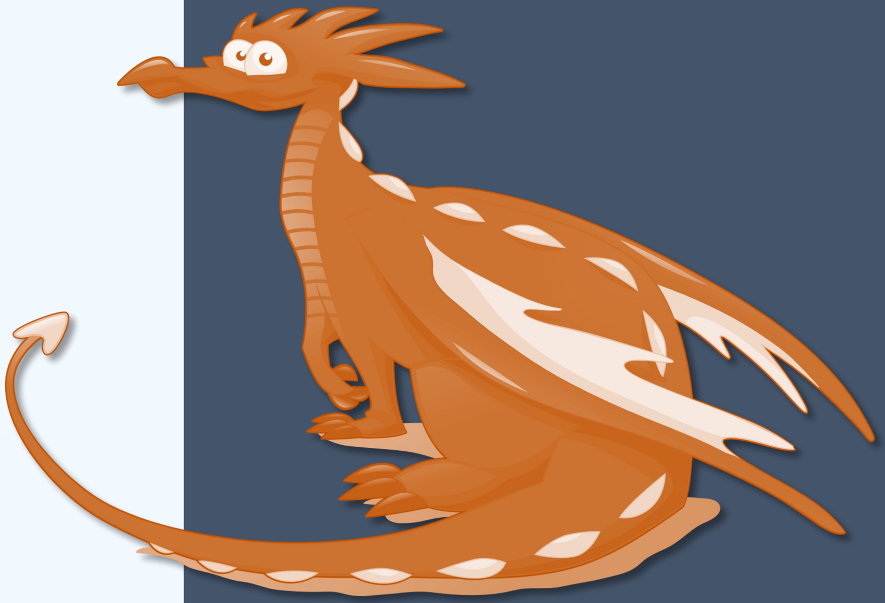
A clear **purpose**: impact and efficiency of research



The three great hurdles

What is stopping you?

- Time & funding
- Buy-in
- Skills



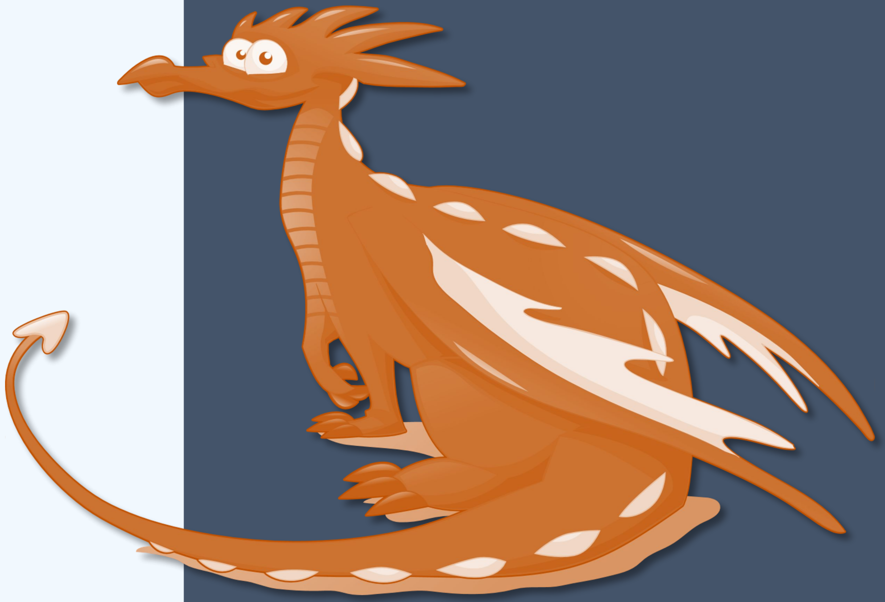
The three great hurdles

What is stopping you?

Time & funding

Buy-in

Skills



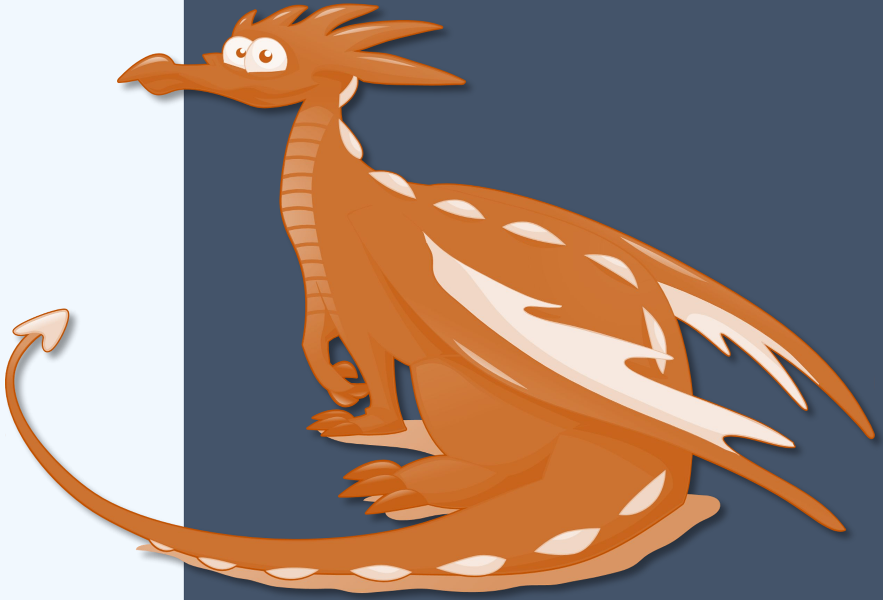
The three great hurdles

What is stopping you?

Time & funding

Buy-in

Skills



Data stewards

Our innovative proposition

Come to rescue you?

Data stewards

Equip you, not rescue you

A Fellowship of data stewards – of ambassadors

Data stewards

Skills

Equip Data Stewards with training and make them trainers

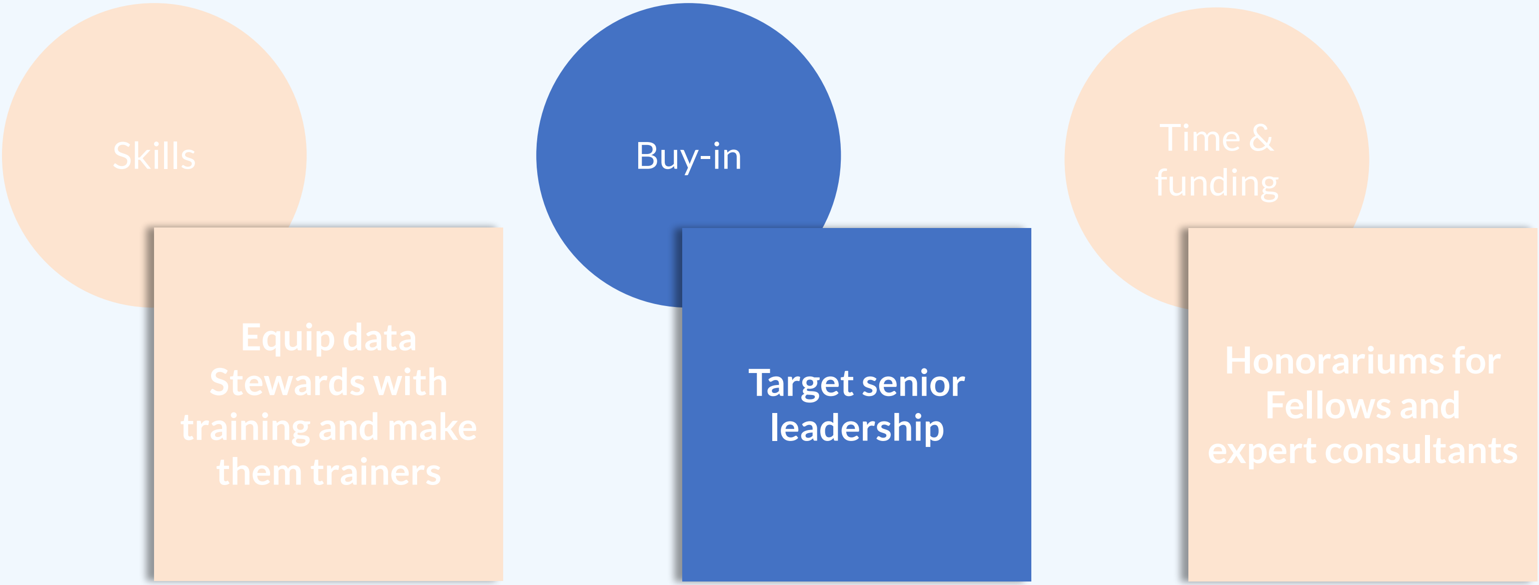
Buy-in

Target senior leadership

Time & funding

Honorariums for Fellows and expert consultants

Data stewards



Skills

Equip data Stewards with training and make them trainers

Buy-in

Target senior leadership

Time & funding

Honorariums for Fellows and expert consultants

Data stewards

Skills

Equip data Stewards with training and make them trainers

Buy-in

Target senior leadership

Time & funding

Honorariums for Fellows and expert consultants

The Fellows

Associate Professor
University of Southampton

Senior Research Fellow
Oxford Brookes University

Research Fellow
University of York

Research Associate
Alan Turing Institute

Research Fellow
University of Warwick (Medical School)

Data stewards

The Fellows

Bioinformatician
University of York

Data scientist
University of Manchester

Research Data Manager
University of Edinburgh

Data stewards

The Fellows

PhD student
University of Bradford

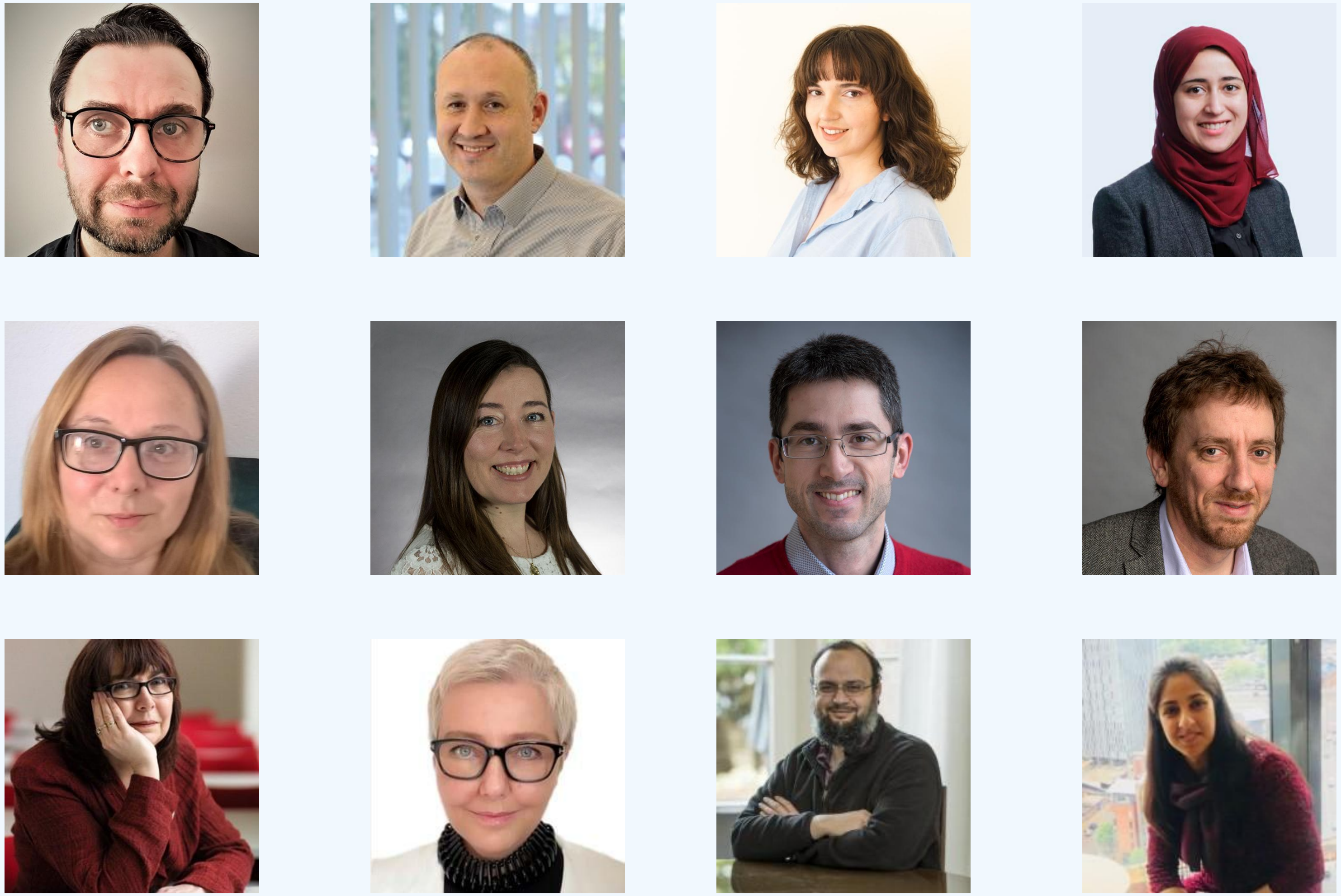
elixir UNITED KINGDOM

Data stewards

2nd Cohort is open now!

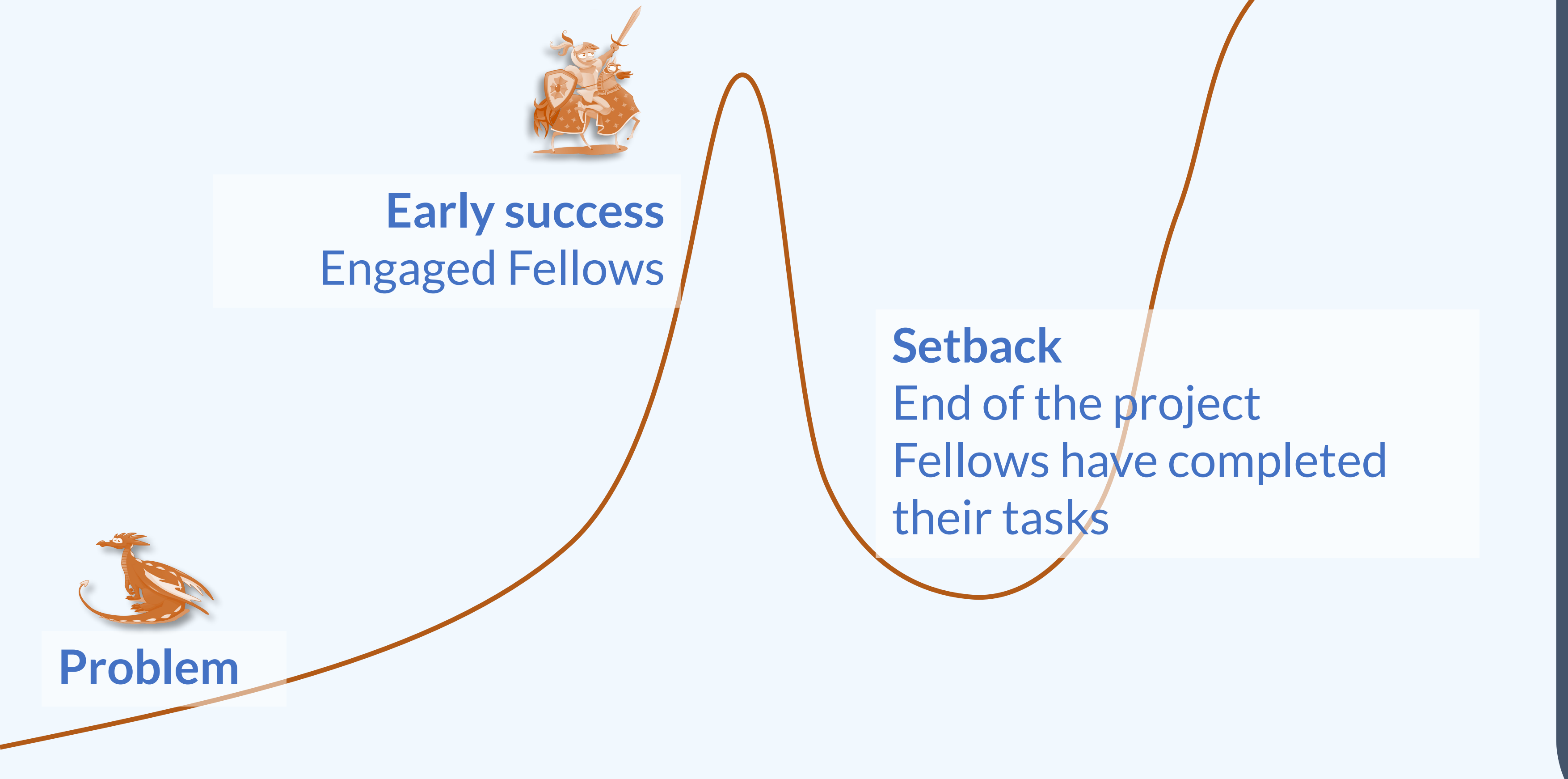
The team behind the project

The team behind the project



No easy way

What is realistic?



Problem

Early success
Engaged Fellows

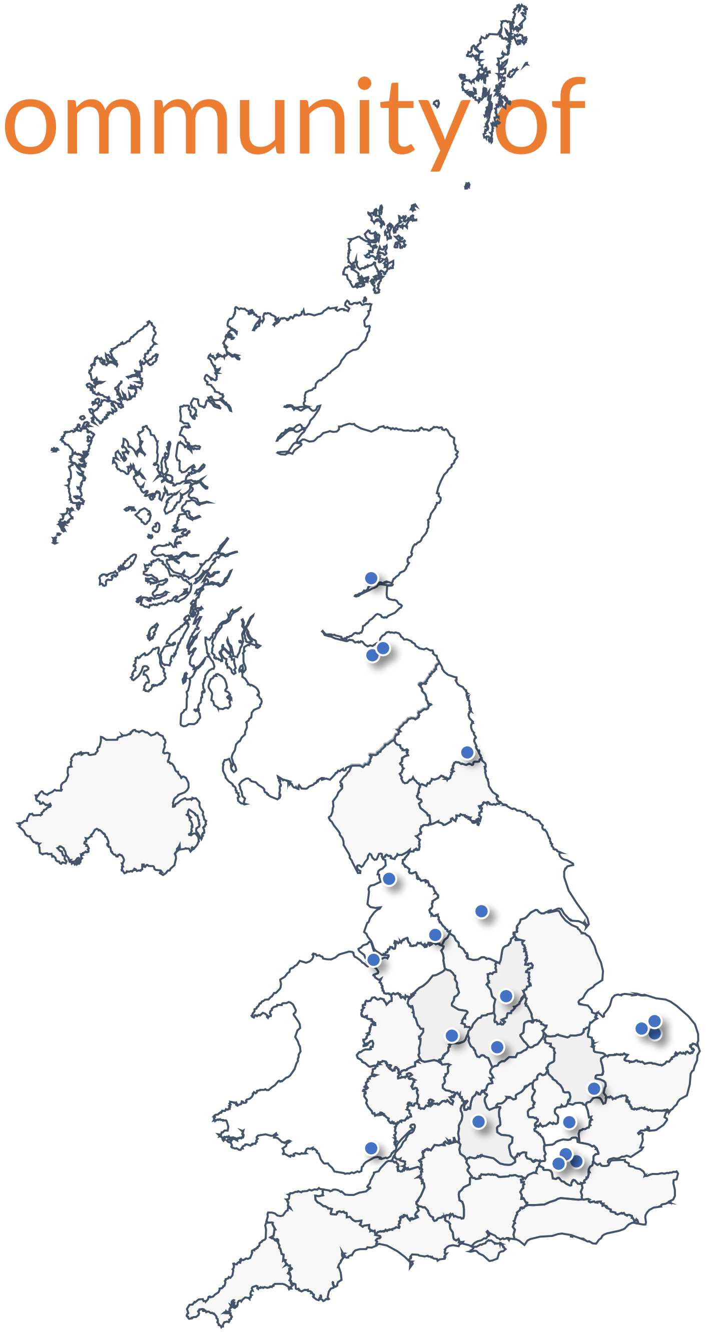
Setback
End of the project
Fellows have completed their tasks

Sustainability

Part of a larger network of experts and successful communities of practice.

We've partnered with the **Sustainable Software Institute** (SSI). A successful community of practice with 10 years of expertise.

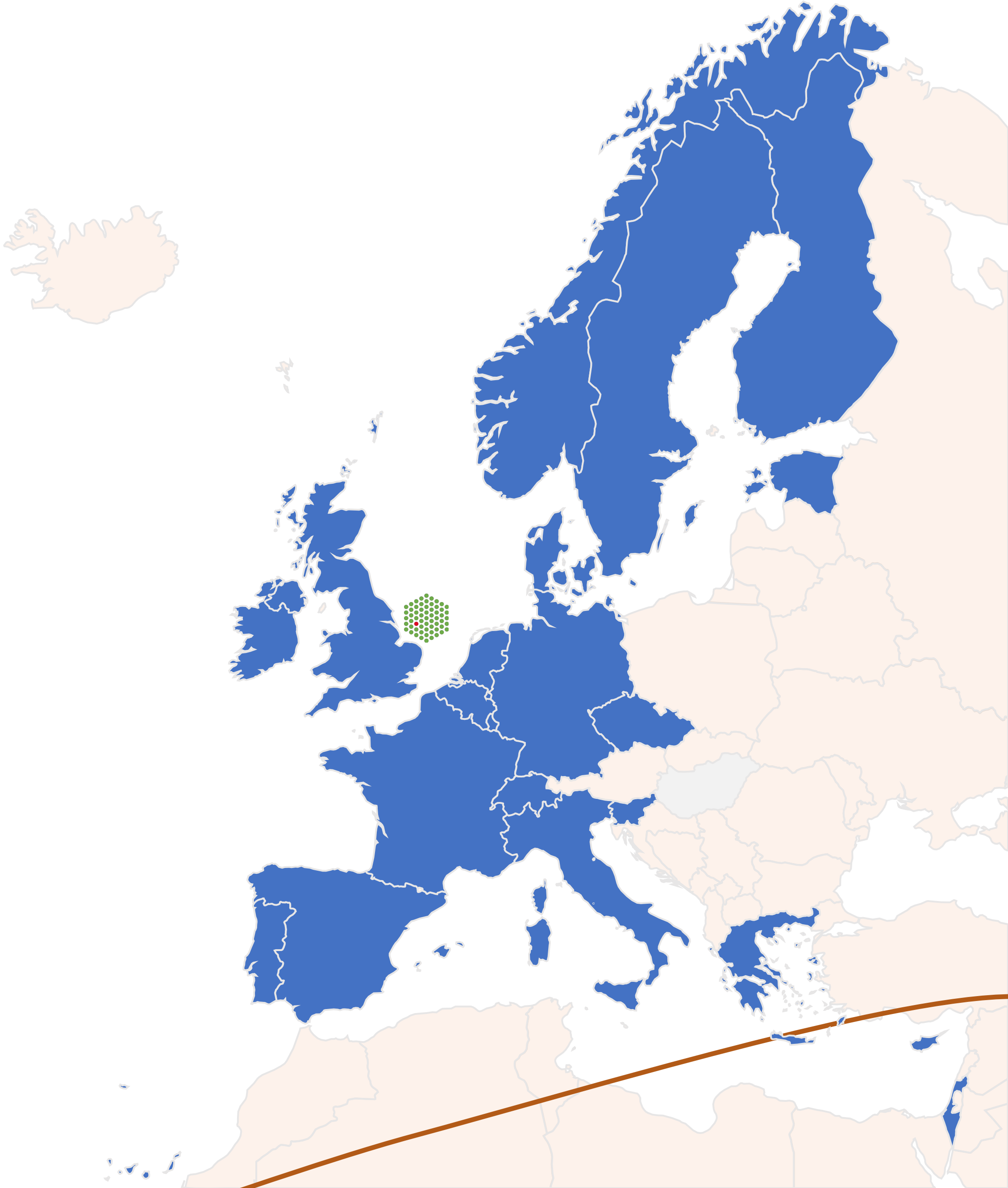
An active UK community of experts ELIXIR-UK



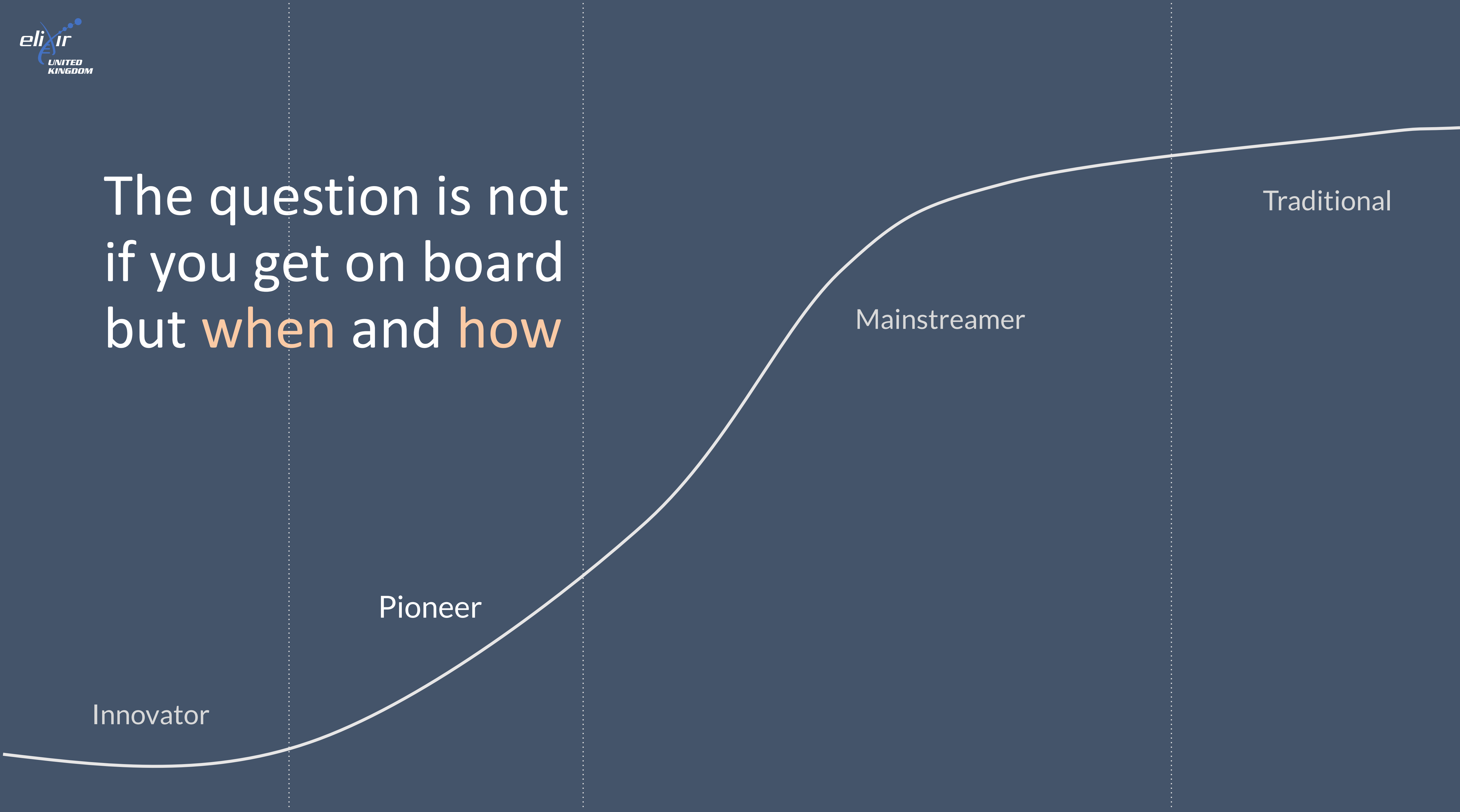
Access to an even bigger ELIXIR Europe community

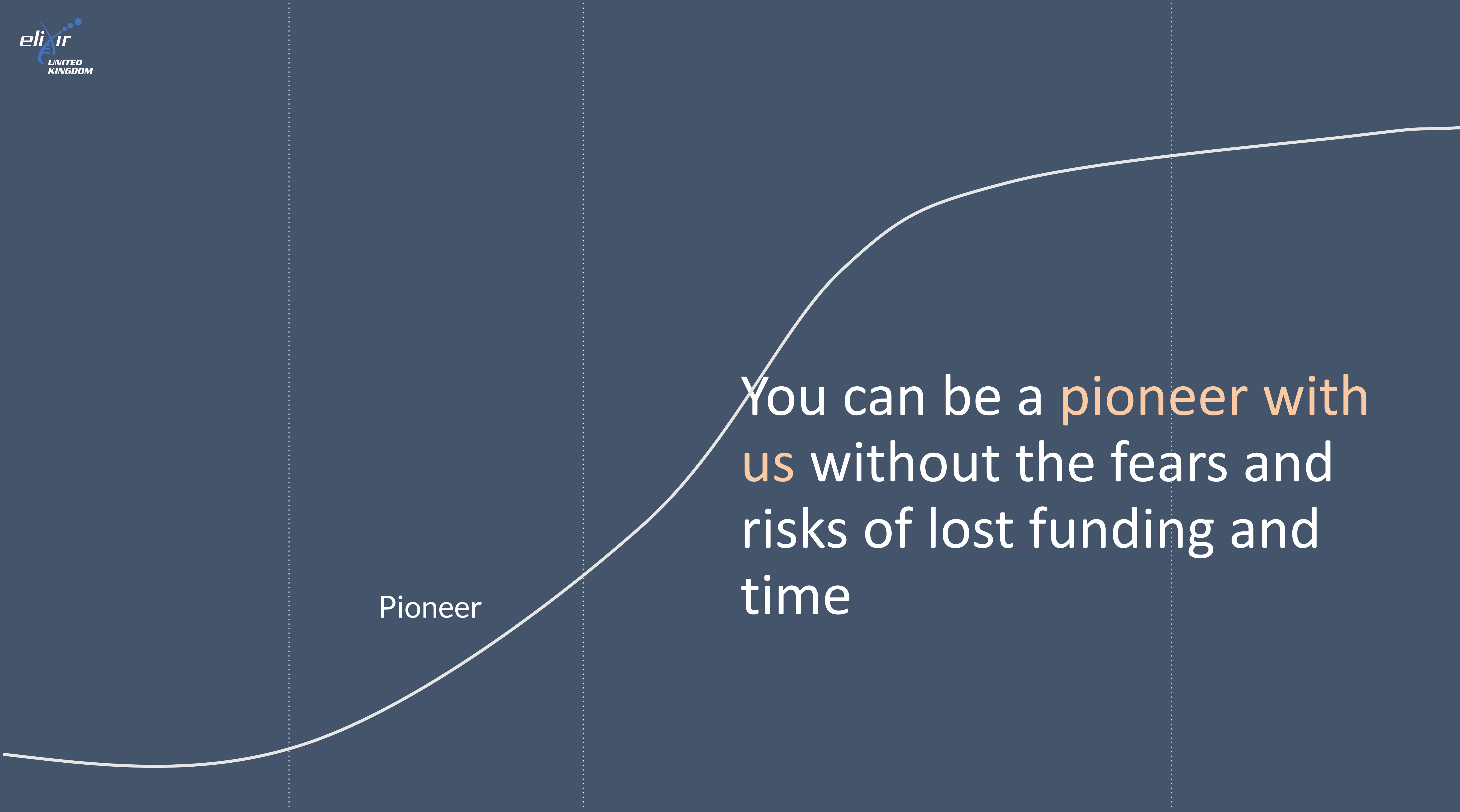
22 Countries + EMBL-EBI

elixir-europe.org



The question is not if you get on board but *when* and *how*





You can be a pioneer with us without the fears and risks of lost funding and time

Thanks!!!

Xènia Pérez Sitjà
Data Stewardship Community Manager

✉ x.perezsitja@bradford.ac.uk

£

becoming a Fellow

Application open until 10 July

elixiruknode.org/activities/elixir-dash-fellowship

Medical Research Council

Funded by the ELIXIR-UK: FAIR Data Stewardship training UKRI award (MR/V038966/1)

UNIVERSITY OF BRADFORD

Cloud-SPAN: Training Researchers in Cloud-based High Performance Computing for SPecialised ANalyses on environmental 'omics

Emma Rand & James Chong

UK Research and Innovation

MR/V038680/1

Software Sustainability Institute

University of York

✉ cloud-span-project@york.ac.uk

http://cloud-span.york.ac.uk/

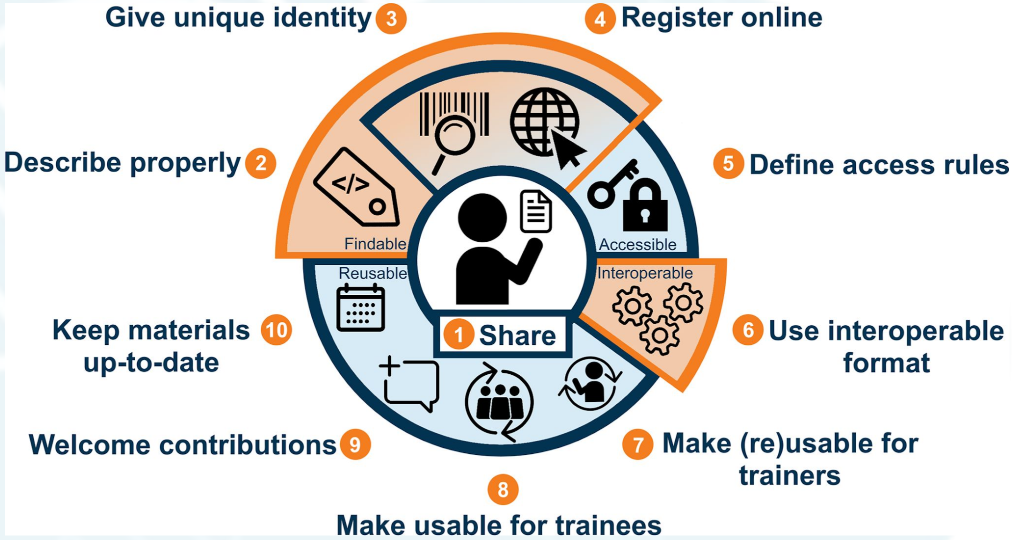
@SpanCloud

Cloud-SPAN

For researchers and
Research support teams

Cloud-based, Containerised

Findable Accessible
Interoperable Reusable



Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
Garcia, L., et al., 2020. Ten simple rules for making training materials FAIR. PLoS Comput. Biol. 16, 1–9. https://doi.org/10.1371/journal.pcbi.1007854

art designer at PLoS, CC BY-SA 3.0 via Wikimedia Commons.

Wiegers, Luc, & van Gelder, Celia W. G. (2019). Illustration for "Ten simple rules for making training materials FAIR" (1.0). Zenodo. https://doi.org/10.5281/zenodo.3593258

What are we teaching?

Foundational

Advanced

Train trainer

Content

[Genomics]

[Prenomics]

[Self assessment]

What are we teaching?

Foundational

Advanced

Train trainer

Content

- Create your own AWS instance
  - Self-study courses
- Metagenomics
- Experimental design
- Scheduling, automating analyses

What are we teaching?

Foundational

Advanced

Train trainer

Content

- Documentation
- Cloud Administration

  Guides for institutional RSE, RDM or HPC Teams

How are we teaching?

- In-person & online, small groups
- Free, Diversity Scholarships
- Code retreats
- Community of practice

The team!

Sarah Forrester

Annabel Cansdale

Emma Barnes

Evelyn Greeves

Jorge Buenabad-Chávez

Sarah Dowsland

Coming next

- Prenomics November 22nd and 24th, 10:00-13:00. Online
- Genomics December 6th and 7th, 09:30-16:30 In-person at York
- Genomics by self study *Soon!*
- Create Your AWS Instance
- Code Retreats *Soon!*
- Metagenomics, Autumn - Spring 2023 TBC

https://cloud-span.york.ac.uk/home

Software sustainability institute Fellowship - Developing metagenomic bioinformatics training materials

Sarah Forrester

Bioinformatician/ PDRA
University of York

Software sustainability institute

Taken from the SSI website:

”Since 2010, the Software sustainability Institute has facilitated the advancement of software in research by cultivating better, more sustainable, research software to enable world-class research (“Better software, better research”)

BETTER SOFTWARE BETTER RESEARCH

One of the ways in which they facilitate this is through their fellowship programme. Many fellowship applications involve the development of training materials

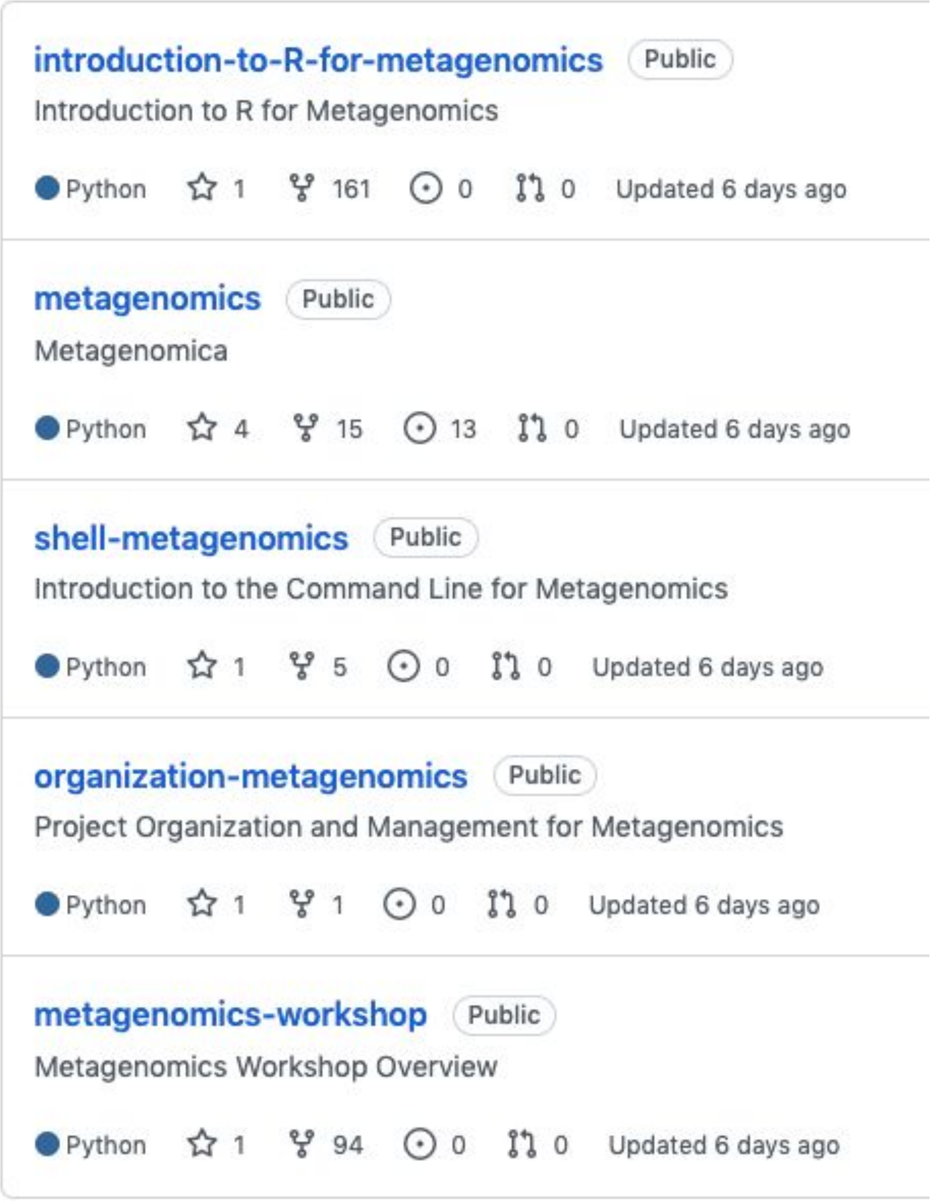
The SSI also has a partnership with The Carpentries - which has a suite of programmes for essential data management and analysis skills

Caveat: I only became an SSI Fellow in 2022 I wouldn’t consider myself a representative for the SSI - this is purely to show how the SSI fellowship has enabled me to apply my metagenomic knowledge to develop training material

What bioinformatics training will my fellowship involve

Developing metagenomics course

- Adapt carpentry lessons currently in the carpentries incubator
- Place for carpentry lessons that are actively being developed
- There are currently 5 ”metagenomics” courses available
- 4 of these are lessons based on the intro to genomics course
    - Intro to command line
    - Intro to R
    - How to organize data
- Delivery to be done alongside CloudSPAN



introduction-to-R-for-metagenomics Public

Introduction to R for Metagenomics

● Python ☆ 1 ⑂ 161 ⊙ 0 ⇅ 0 Updated 6 days ago

metagenomics Public

Metagenomica

● Python ☆ 4 ⑂ 15 ⊙ 13 ⇅ 0 Updated 6 days ago

shell-metagenomics Public

Introduction to the Command Line for Metagenomics

● Python ☆ 1 ⑂ 5 ⊙ 0 ⇅ 0 Updated 6 days ago

organization-metagenomics Public

Project Organization and Management for Metagenomics

● Python ☆ 1 ⑂ 1 ⊙ 0 ⇅ 0 Updated 6 days ago

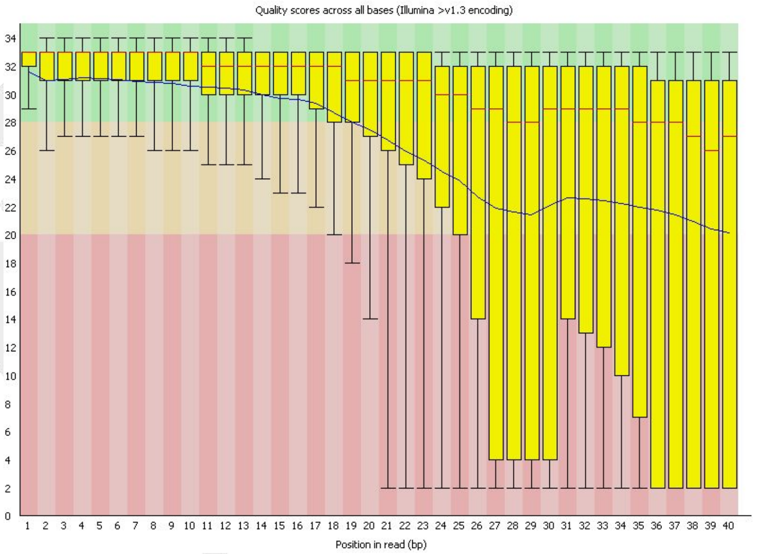
metagenomics-workshop Public

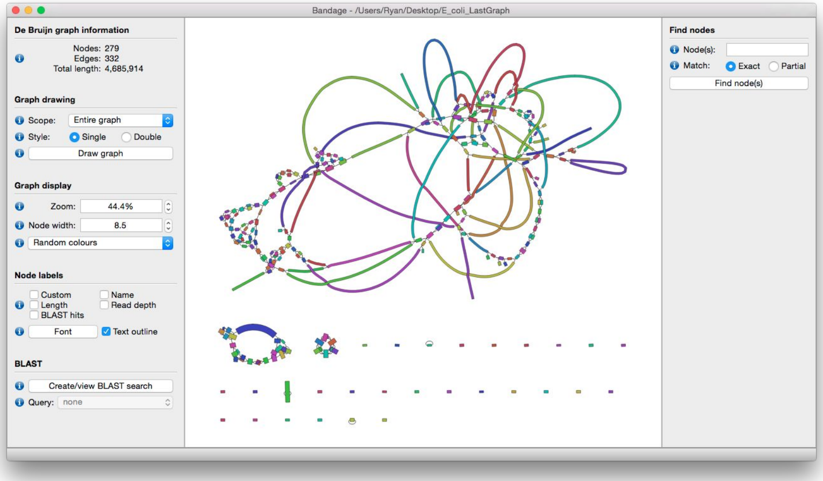
Metagenomics Workshop Overview

● Python ☆ 1 ⑂ 94 ⊙ 0 ⇅ 0 Updated 6 days ago

What content are we going to cover on the course?

- QC raw data
- Generating an assembly
- QC this assembly
- Binning the assembly into MAGs (organisms)
- Identifying what is the taxonomy of these MAGs
- Functional information (what metabolisms might be present)




Bandage - /Users/Ryan/Desktop/E_coli_LastGraph

De Bruijn graph information

Nodes: 279
Edges: 332
Total length: 4,685,914

Graph drawing

Scope: Entire graph
Style: ● Single ○ Double

Graph display

Zoom: 44.4%
Node width: 8.5
Random colours

Node labels

□ Custom □ Name
□ Length □ Read depth
□ BLAST hits

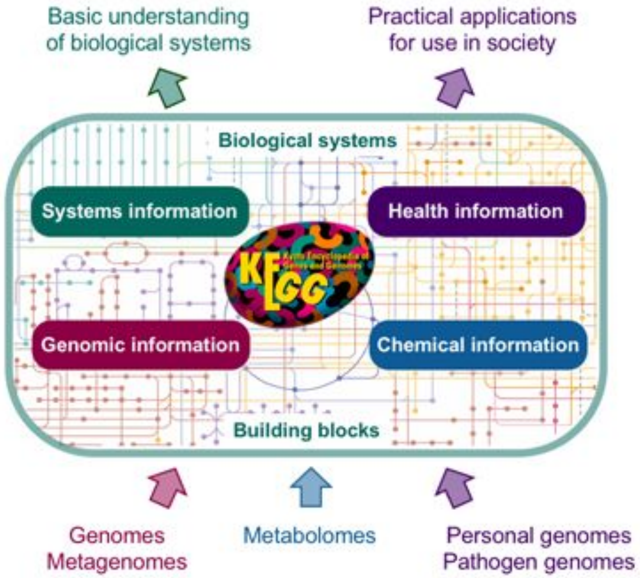
☑ Text outline

BLAST

Query: none

Find nodes

Node(s):
Match: ● Exact ○ Partial


Basic understanding of biological systems

Practical applications for use in society

Biological systems

Systems information

Health information

Genomic information

Chemical information

Building blocks

Genomes
Metagenomes

Metabolomes

Personal genomes
Pathogen genomes

What bioinformatics training will my fellowship involve

Additional content:

- Long read sequencing methods (assembly and QC)
- Reduce non metagenome specific content
- Database selection and its effect on your taxonomic annotations
- How to perform these stages using AWS
- **The importance of making datasets publicly available**

Course to be delivered alongside Annabel Cansdale via CloudSPAN in Autumn - Spring 2023

EBNet: Why bioinformatics training is important

An **eight terabyte** case study

Annabel Cansdale

Bioinformatician, Chong Lab, University of York

DNA Sequencing

- It is getting faster and cheaper to generate large amounts of data

- It is especially easy to generate very large amounts of sequencing data with metagenomics!

Quick case study

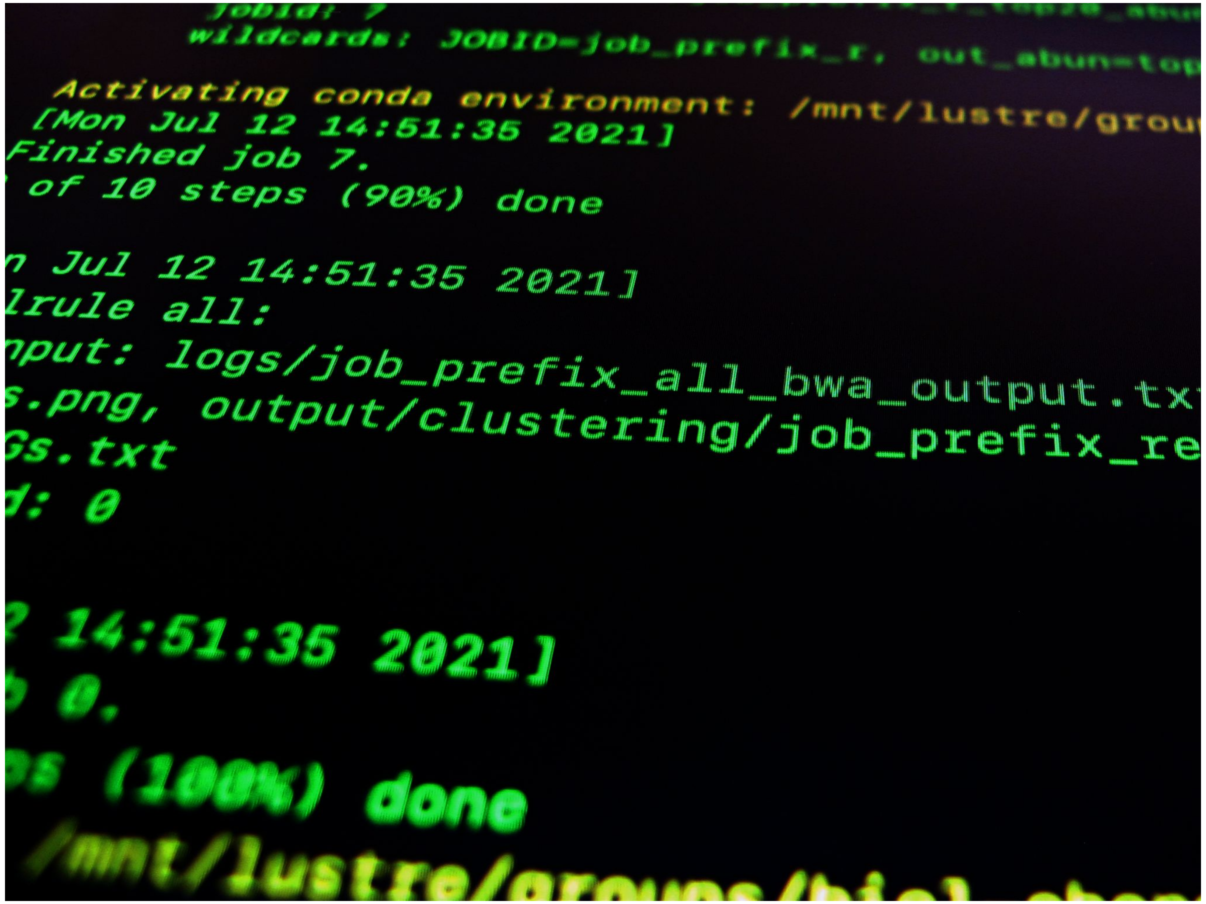
- Anaerobic digestion metagenomic time-series dataset
- Combination of Nanopore and Illumina sequencing
- Received **8TB** of raw data!
- Had to think about storage and backup of this and how we would do the analysis

Image credit: wasatchresourcerecovery.com

Computational power

The analysis of this ended up taking a lot of computational power!

Just the initial assembly/polishing used >500GB RAM and generated ~700GB of intermediary files



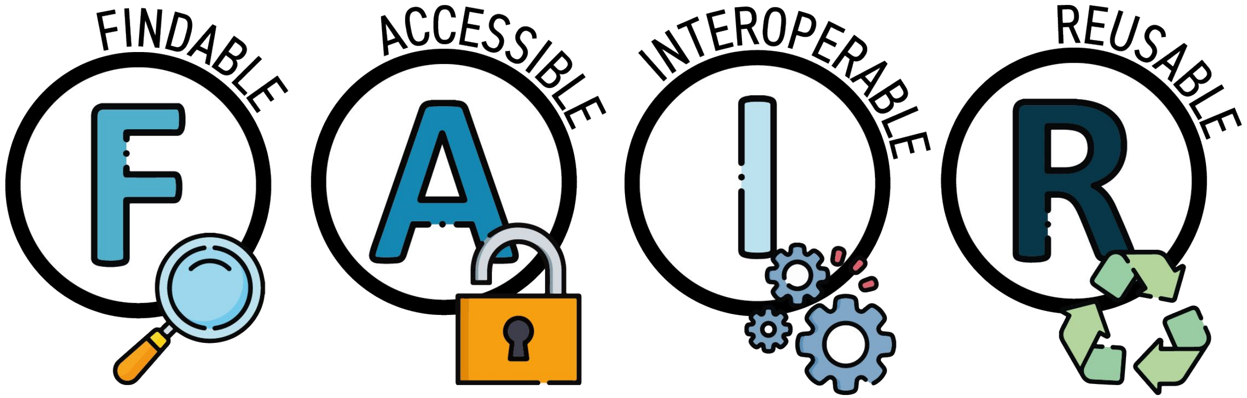
```
    jobid: 7
    wildcards: JOBID=job_prefix_r, out_abun=top
 Activating conda environment: /mnt/lustre/grou
[Mon Jul 12 14:51:35 2021]
Finished job 7.
 of 10 steps (90%) done

n Jul 12 14:51:35 2021]
lrule all:
nput: logs/job_prefix_all_bwa_output.tx
s.png, output/clustering/job_prefix_re
Gs.txt
d: 0


2 14:51:35 2021]
b 0.
ps (100%) done
/mnt/lustre/groups/biol
```

Takeaways

- This is an extreme example! but:
  - Datasets are becoming larger!
  - & We need people with the skills to be able to deal with them!
  - (P.s. don’t forget about storage)
- Training early on is key so you
  - Can plan your experiments effectively
  - Don’t panic when you receive large amounts of data
  - Can identify where you might get computational bottlenecks
- This is where the projects we’re hearing about today come in!

FAIR – Why making data science reproducible is important

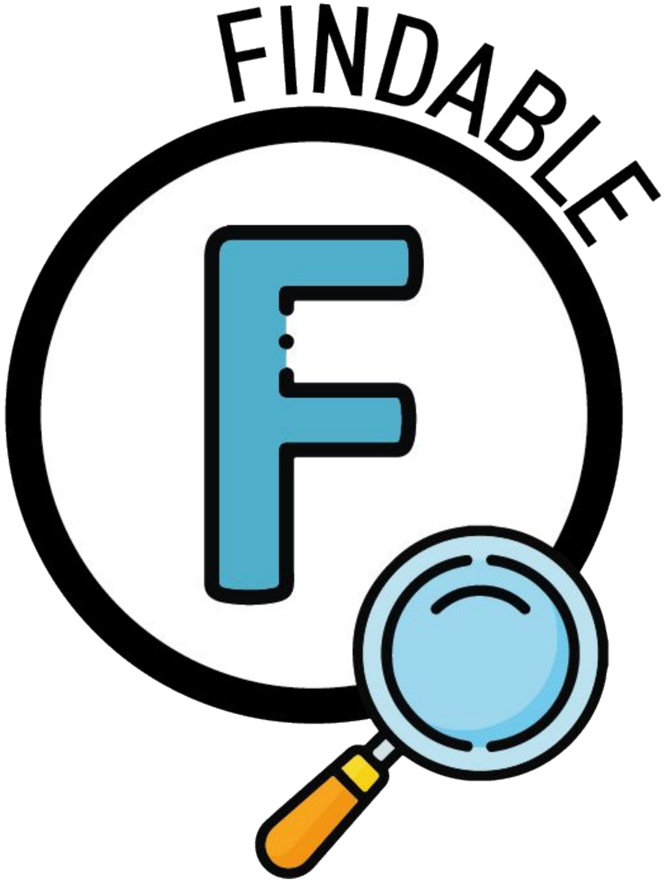
Evelyn Greeves



METADATA

Do I know what this data is?

- Metadata = data *about* data
- Gives an overview of dataset/resource
- Allows tagging, tracking and indexing in a registry



PERSISTENT IDENTIFIERS

Do I know where to find it?

- e.g. DOI, ORCID iD
- Long-lasting and unique to dataset/resource
- Prevent "link rot"

Why is it important?

People can't re-use your data if they don't know it exists or can't find it.

RETRIEVABLE DATA

Do I know how to get the data?

- No special tools needed to get hold of the (meta)data
- Authentication and authorization may be needed to access the data itself

PERSISTENT METADATA

Will there be a record of the data if it disappears?

- Metadata persists after data no longer available
- Allow tracking down of those associated with original research

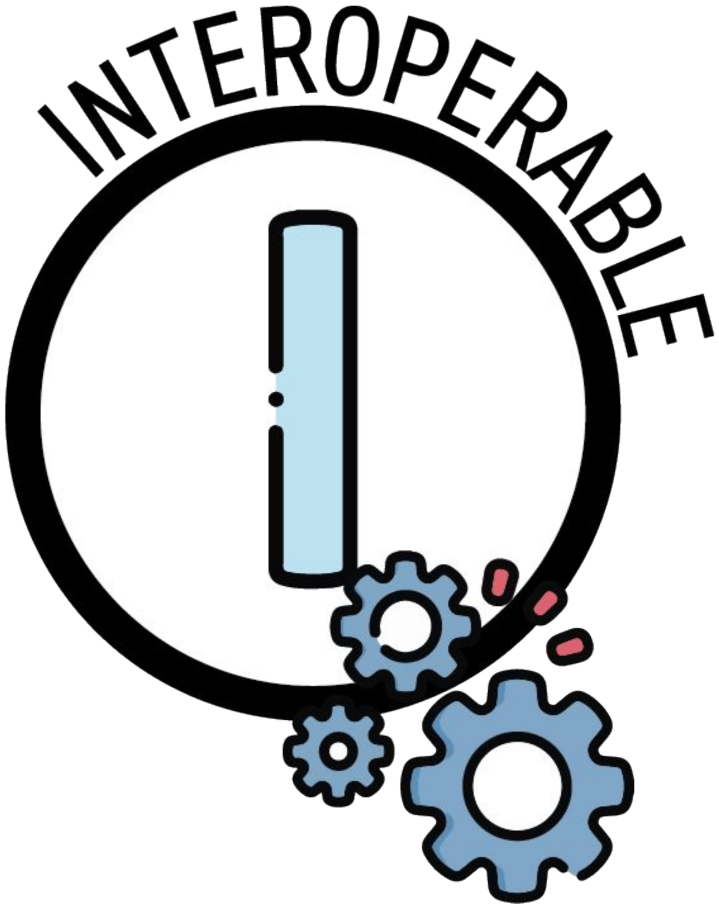
Why is it important?

People won’t re-use your data if it isn’t easy to get hold of.

OPEN FORMATS

Can I open this data?

- Use standardised and open source formats
- e.g. .PPTX instead of .PPT
- Conform with field-specific standards



COMMON VOCABULARIES

Is it easy for a computer to categorise this data?

- Enable better organization of knowledge
- Conform with field-specific ontologies

Why is it important?

People can’t re-use your data if they can’t open it or don’t know what it’s about.

RICH METADATA

Do I understand this data's context?

- Tells story about context of data generation
- As much information as possible included

USAGE LICENSES

Who can reuse this data?

- Clarify how data can be remixed and reused
- e.g. CC-BY license allows free reuse with credit

Why is it important?

Allows maximum benefit to be extracted from your data by helping other researchers re-use it.

Panel discussion

What questions do you have for us?