seqera

Data in life sciences

Views and perspectives on challenges in the field

seqera

Data in life sciences

Views and perspectives on challenges in the field

Join at slido.com

#3201 014

Data in life sciences

- Background and introduction
- Big data keeps getting bigger
- Artificial intelligence
- Understanding and trust
- On the importance of being open
- Future challenges
- Conclusion

Background and introduction

Background and introduction

Phil Ewels, PhD

Product Manager for Open Source

phil.ewels@seqera.io

seqera

Background and introduction

PhD in epigenetics 2008-12

University of Cambridge

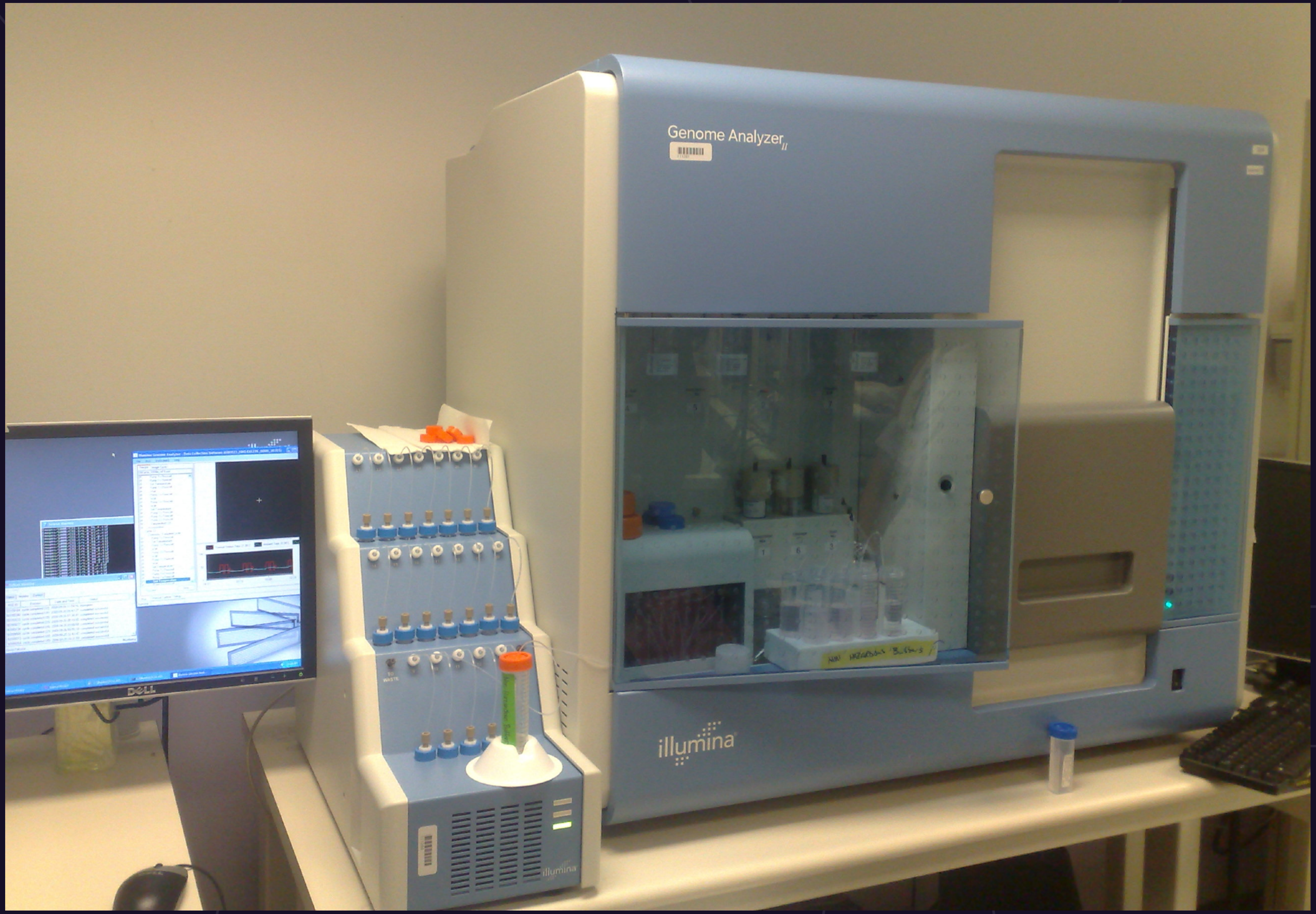
Dawn of NGS:

Big data came to biology

1 × 35bp

22M reads / run

Introduction to bioinformatics



Illumina Genome Analyzer IIx

Background and introduction

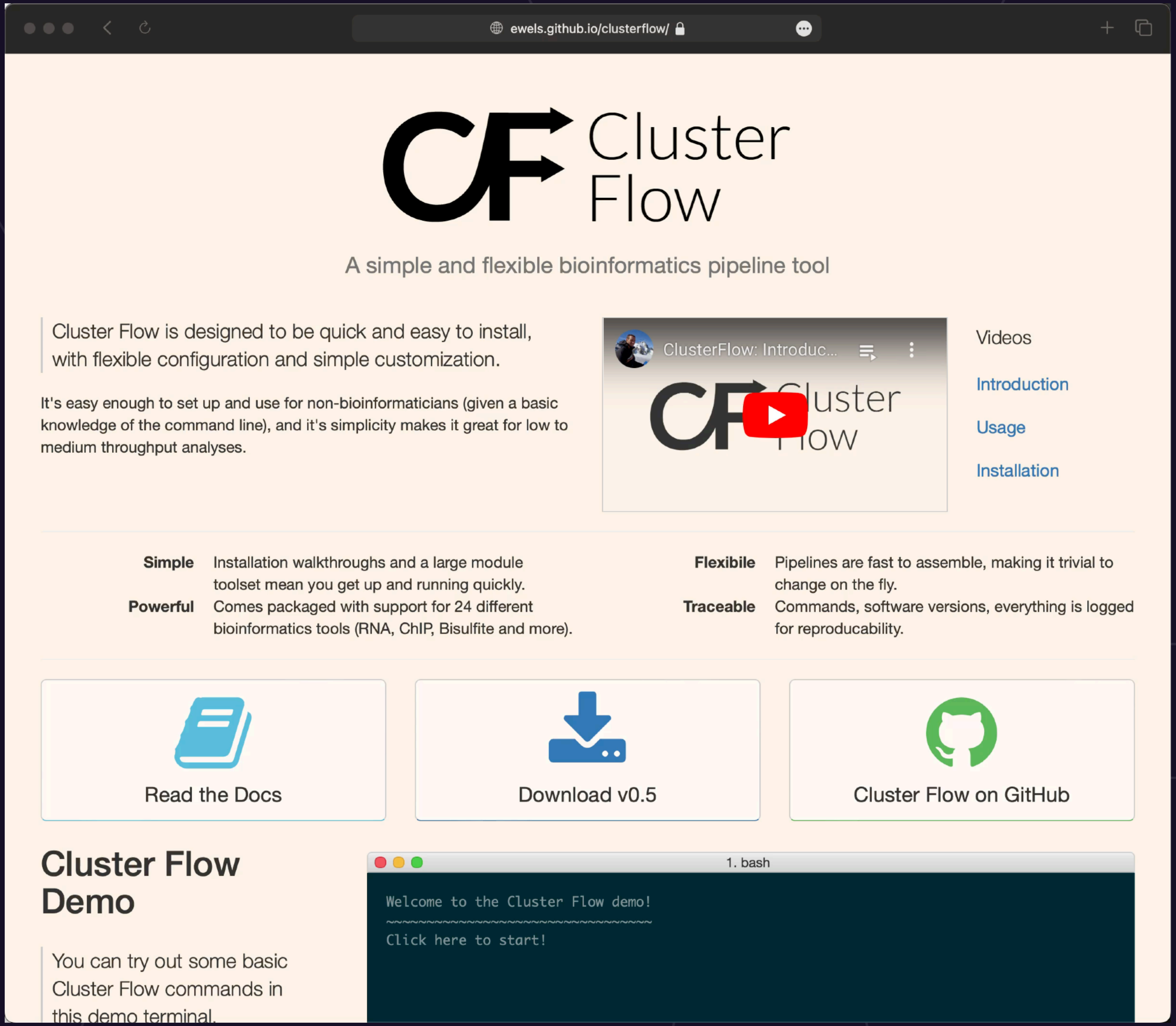
Postdoc in bioinformatics

Babraham Institute, Cambridge

Got the bug for building open-source software

Wrote my own workflow tool

Started writing data-vis scripts



Cluster Flow

A simple and flexible bioinformatics pipeline tool

Cluster Flow is designed to be quick and easy to install, with flexible configuration and simple customization.

It's easy enough to set up and use for non-bioinformaticians (given a basic knowledge of the command line), and it's simplicity makes it great for low to medium throughput analyses.

Videos

Introduction

Usage

Installation

Simple Installation walkthroughs and a large module toolset mean you get up and running quickly.

Powerful Comes packaged with support for 24 different bioinformatics tools (RNA, ChIP, Bisulfite and more).

Flexibile Pipelines are fast to assemble, making it trivial to change on the fly.

Traceable Commands, software versions, everything is logged for reproducability.

Cluster Flow Demo

You can try out some basic Cluster Flow commands in this demo terminal.

```
1. bash
Welcome to the Cluster Flow demo!
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Click here to start!
```

Background and introduction

Moved to Sweden in 2014
Joined NGI at SciLifeLab

Started building software to
handle the scale of data

SciLifeLab

NATIONAL
GENOMICS
INFRASTRUCTURE

Background and introduction

Started building software to handle the scale of data

Wrote and released MultiQC
Adopted Nextflow, started nf-core

SciLifeLab

NATIONAL GENOMICS INFRASTRUCTURE

nextflow

multiqc

nf-core

Background and introduction

Joined Seqera in 2022 as employee #21

Set up the community team, now product manager for OSS*

seqera

the modern biotech stack

nextflow

workflow management

multiqc

reporting and analytics

fusion

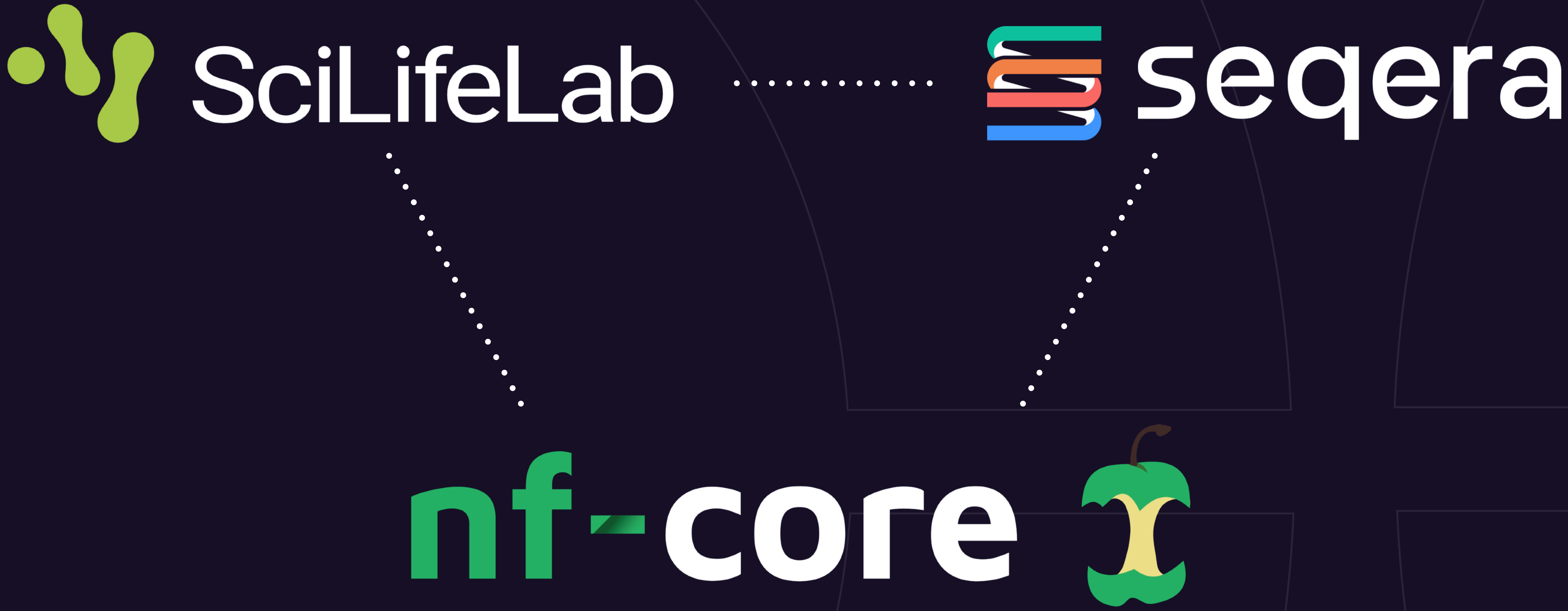
cloud native file-system

wave

container provisioning

nf-core

Background and introduction



SciLifeLab

seqera

nf-core

Background and introduction

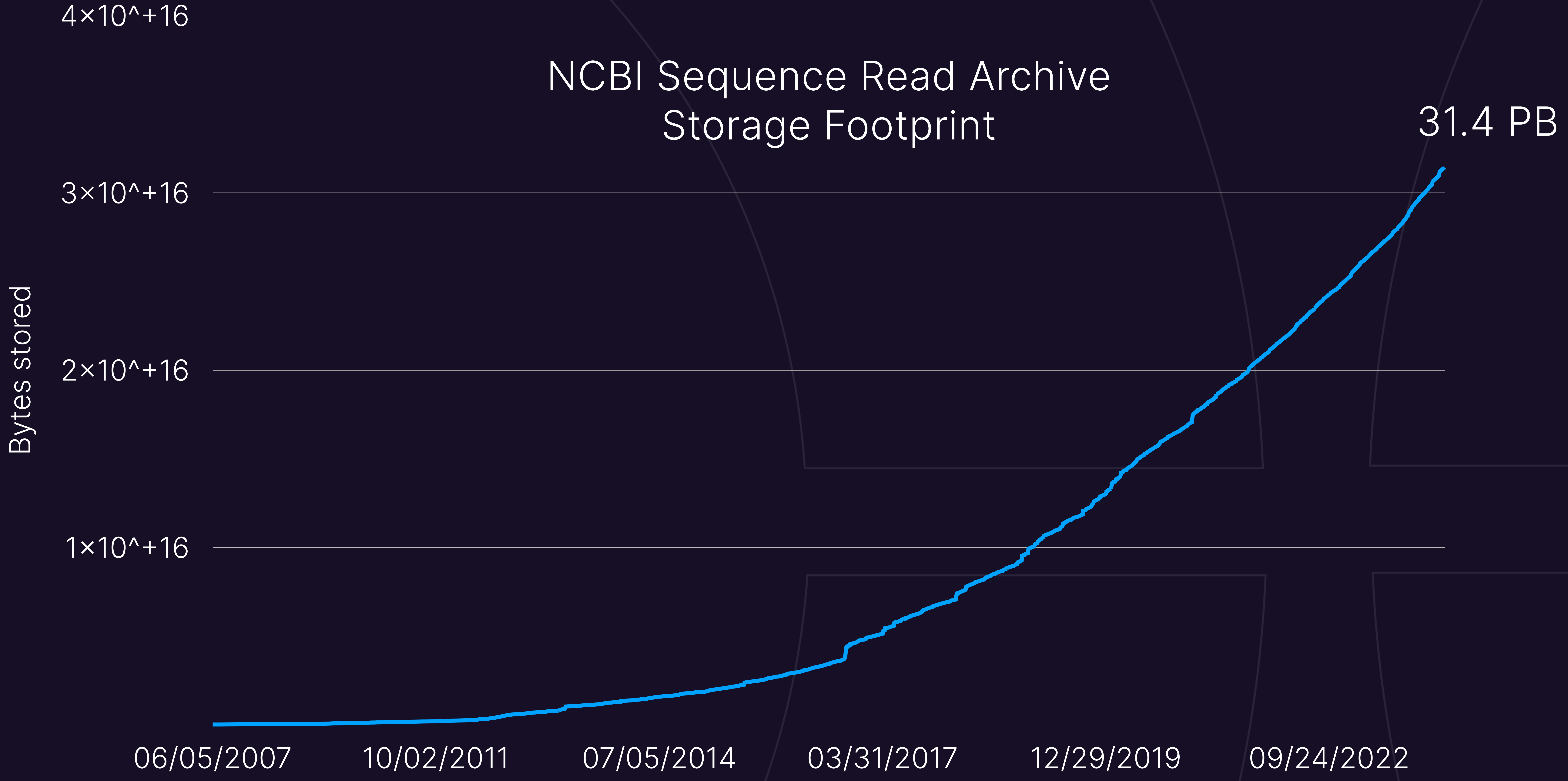
SciLifeLab

seqera

nf-core

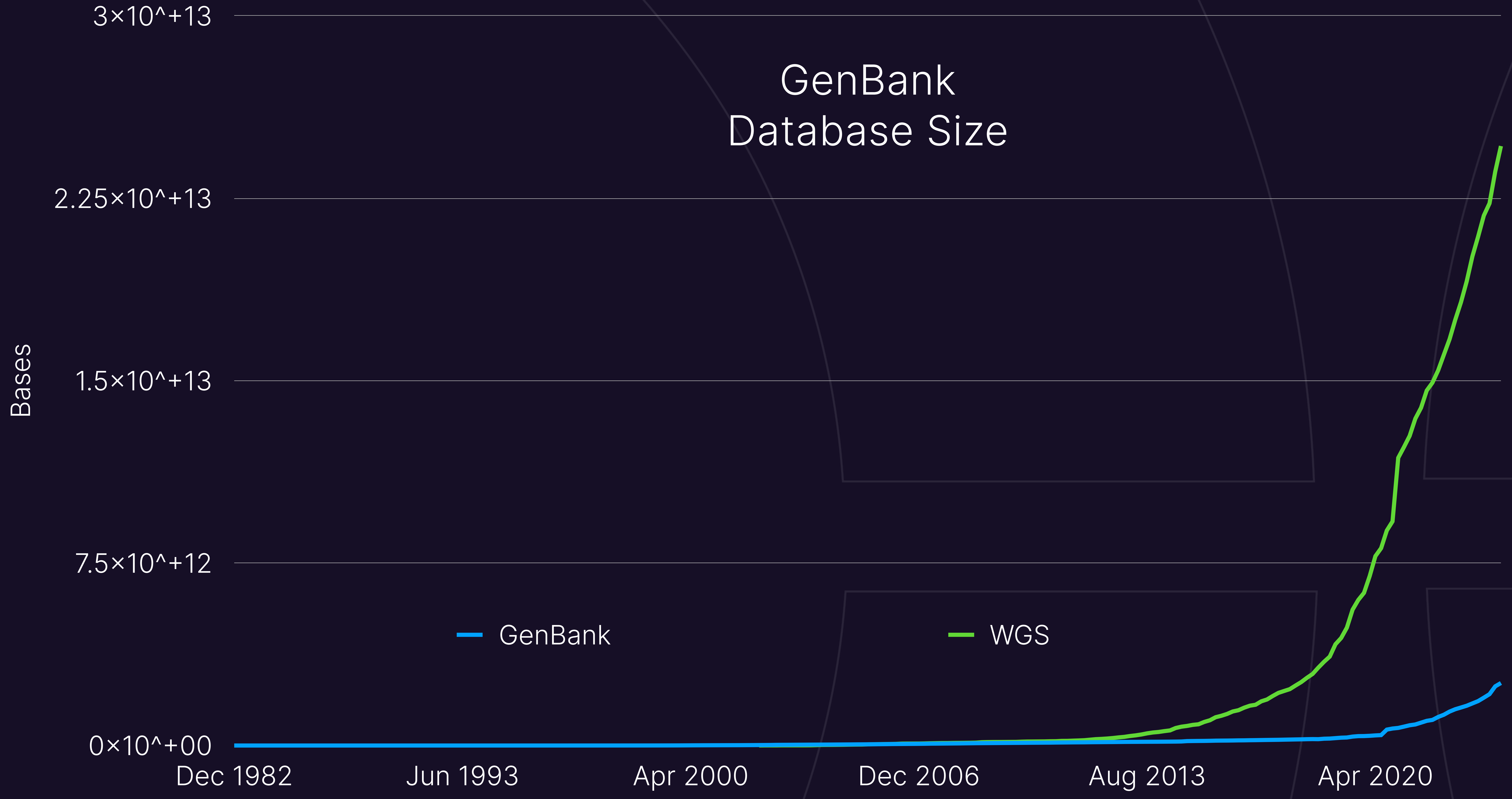
Big data keeps on getting bigger

Big data keeps getting bigger



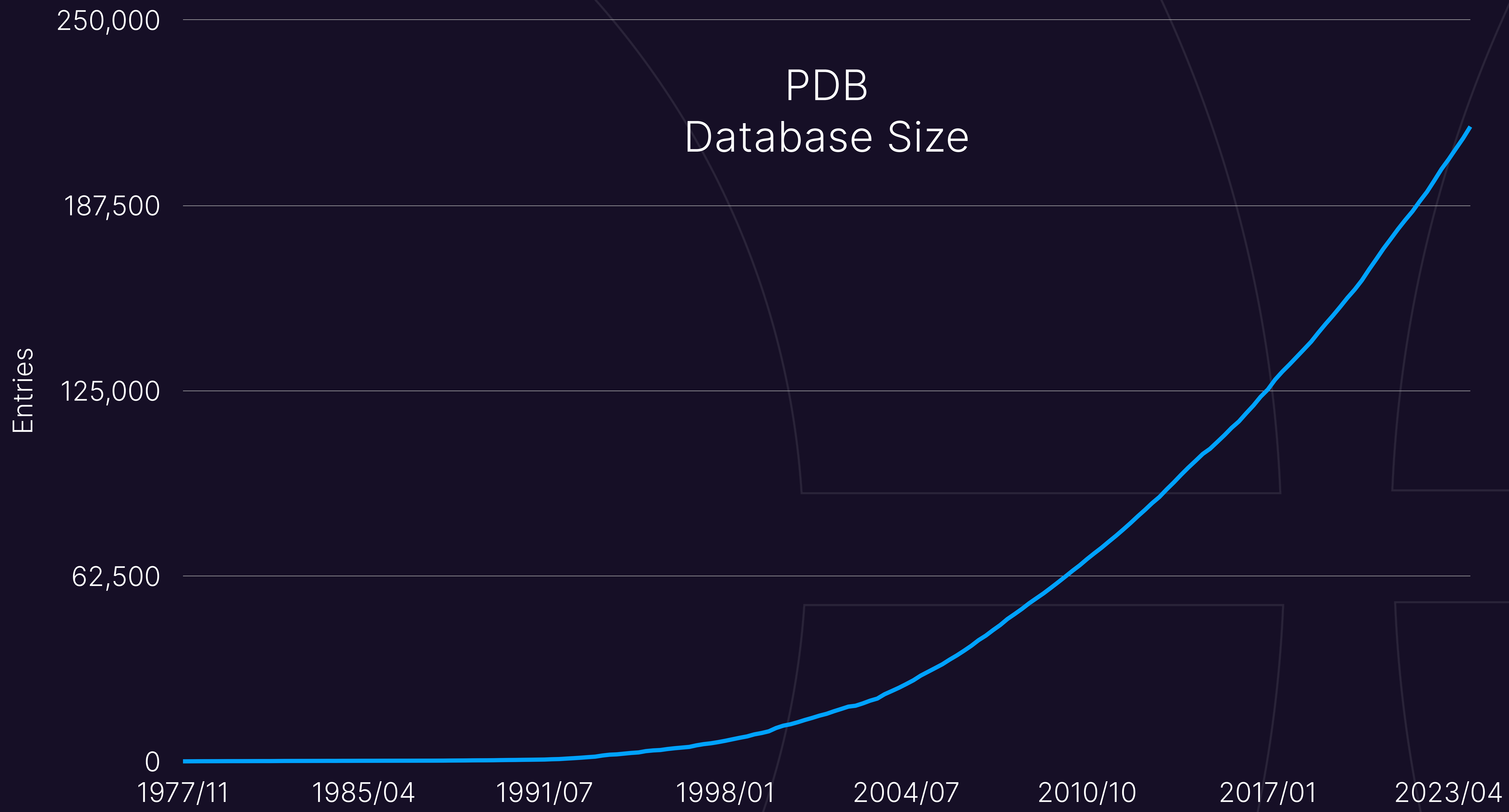
https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Big data keeps getting bigger



https://www.ncbi.nlm.nih.gov/genbank/statistics/

Big data keeps getting bigger



https://www.kanehisa.jp/en/db_growth.html

Big data keeps getting bigger

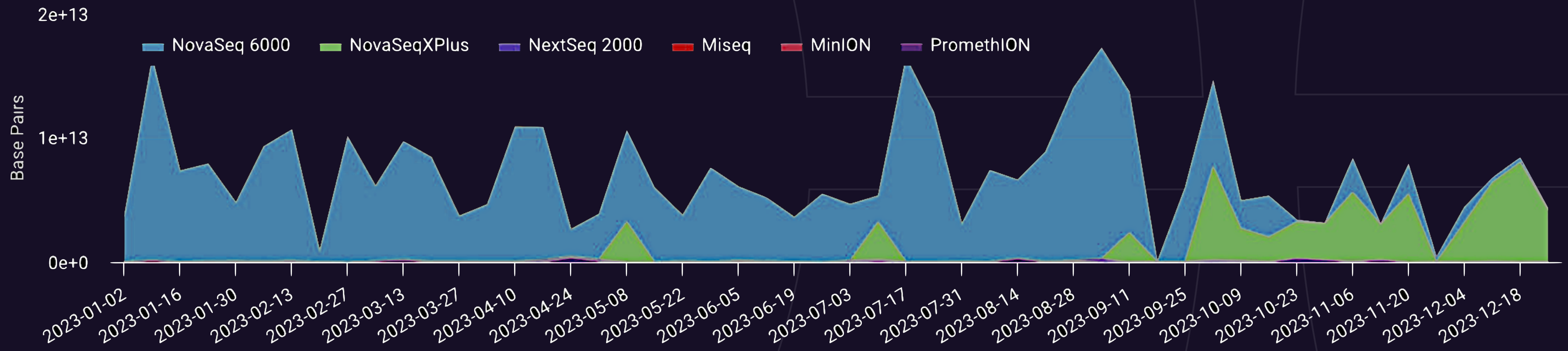
Reproducible analysis of genomics data at scale

SciLifeLab

NATIONAL GENOMICS INFRASTRUCTURE

~1 Tbp sequencing per day in 2023 *(from the Stockholm site only)*

Average for 52 weeks: 1020 Gbp per day
(1 Human genome equivalent every 4.57 minutes)



https://ngisweden.scilifelab.se/resources/ngi-stockholm-status/

Big data keeps getting bigger

Reproducible analysis of
genomics data at scale

SciLifeLab

NATIONAL
GENOMICS
INFRASTRUCTURE

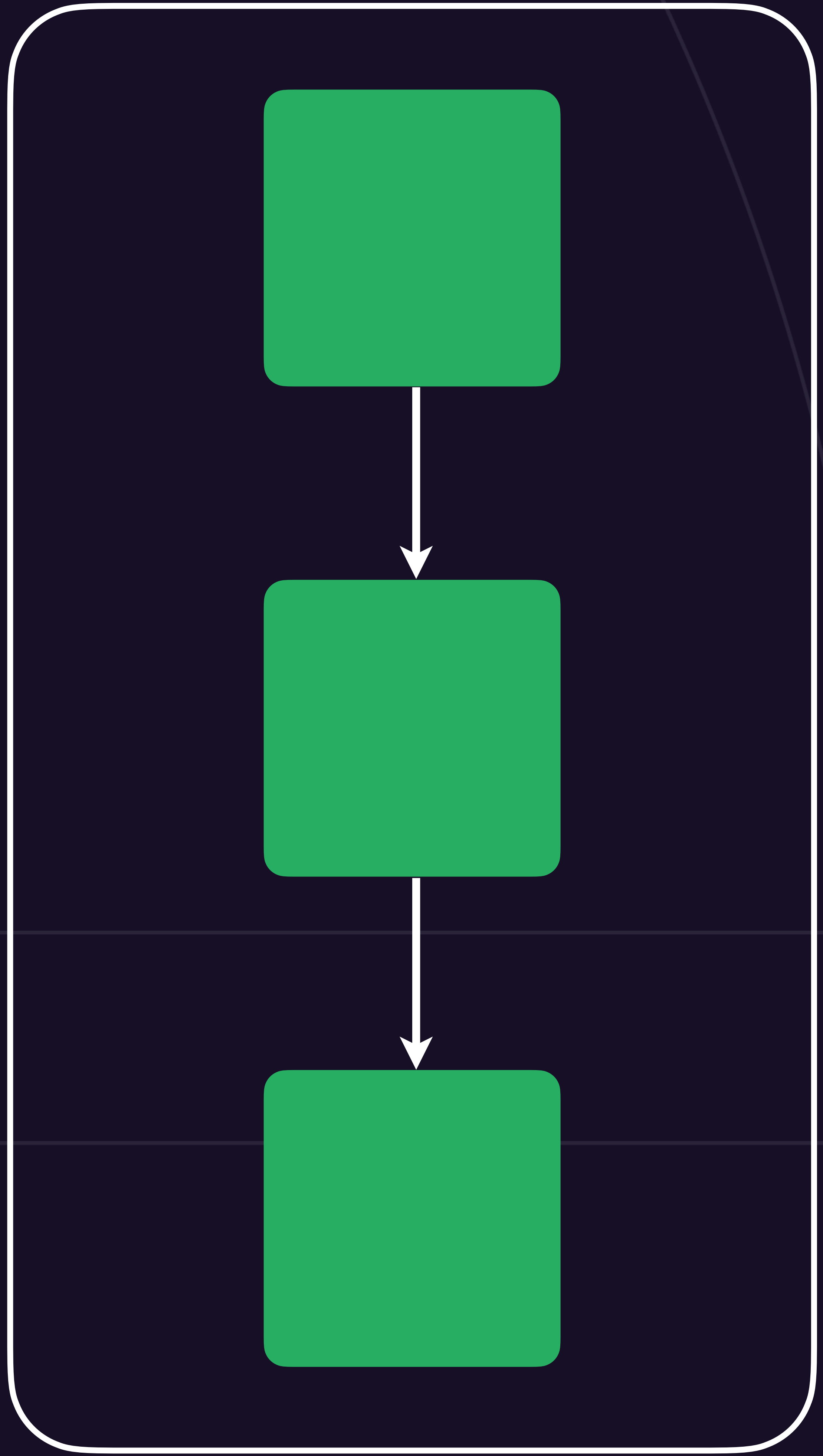
~1 Tbp sequencing per day in 2023 *(from the Stockholm site only)*

nextflow

Big data keeps getting bigger

Reproducible analysis of
any kind of data at scale

nextflow



Processes

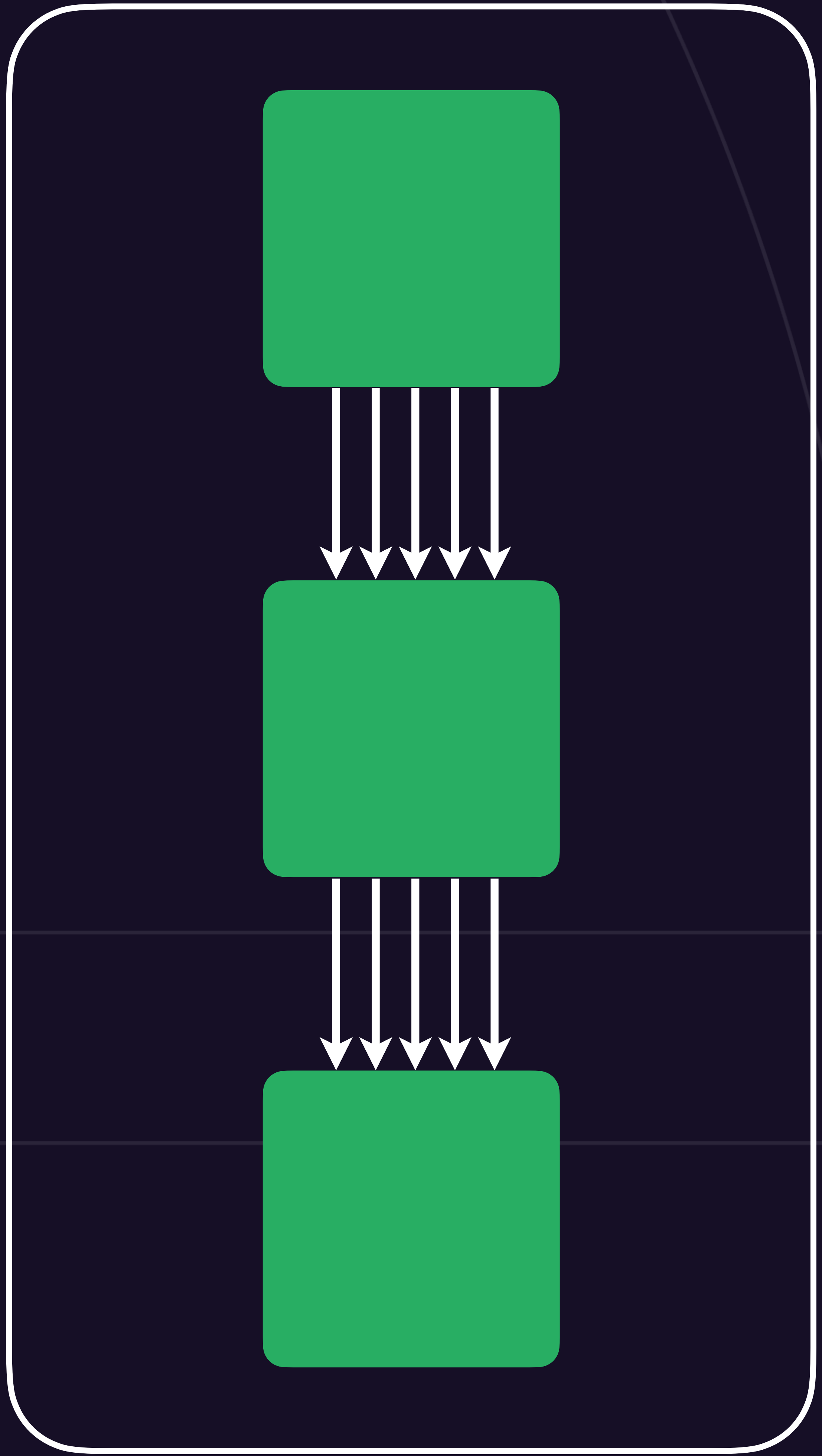
Channels

Workflows

Big data keeps getting bigger

Reproducible analysis of
any kind of data at scale

nextflow



Parallelisation

Reentrancy

(resume partial runs)

Reusability

Big data keeps getting bigger

Reproducible analysis of
any kind of data at scale

nextflow

Compute platforms

slurm

grid

kubernetes

Google GKE

Amazon EKS

AWS BATCH

Azure Batch

Google Life Sciences

Altair | PBS Works™

IBM Spectrum LSF

Big data keeps getting bigger

Reproducible analysis of
any kind of data at scale

nextflow

Compute platforms

slurm

grid

kubernetes

Google GKE

Amazon EKS

AWS BATCH

Azure Batch

Google Life Sciences

Altair | PBS Works™

IBM Spectrum LSF

Storage and data

NFS

aws | S3

Azure Cloud Storage

Google Cloud

SQL

Big data keeps getting bigger

Reproducible analysis of
any kind of data at scale

nextflow

Scalable

Portable

Reproducible

Compute platforms

slurm

grid

kubernetes

Google GKE

Amazon EKS

AWS BATCH

Azure Batch

Google Life Sciences

Altair | PBS Works™

IBM Spectrum LSF

Storage and data

NFS

aws | S3

Azure Cloud Storage

Google Cloud

SQL

Container technologies / SCM

docker

CONDA

Singularity

GitHub

Bitbucket

Artificial Intelligence

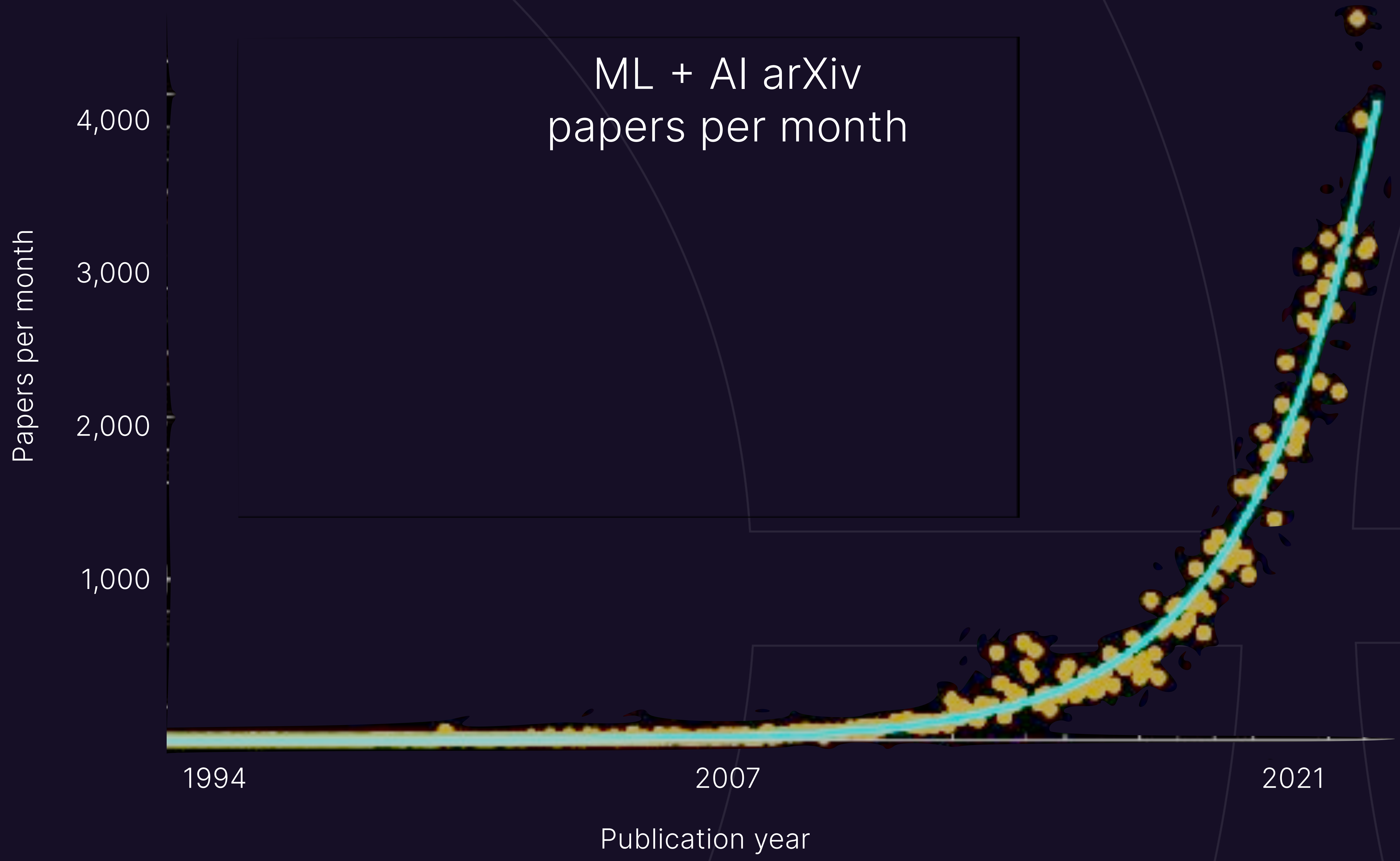
seqera

Artificial Intelligence

Join at slido.com

#3201 014

Artificial Intelligence

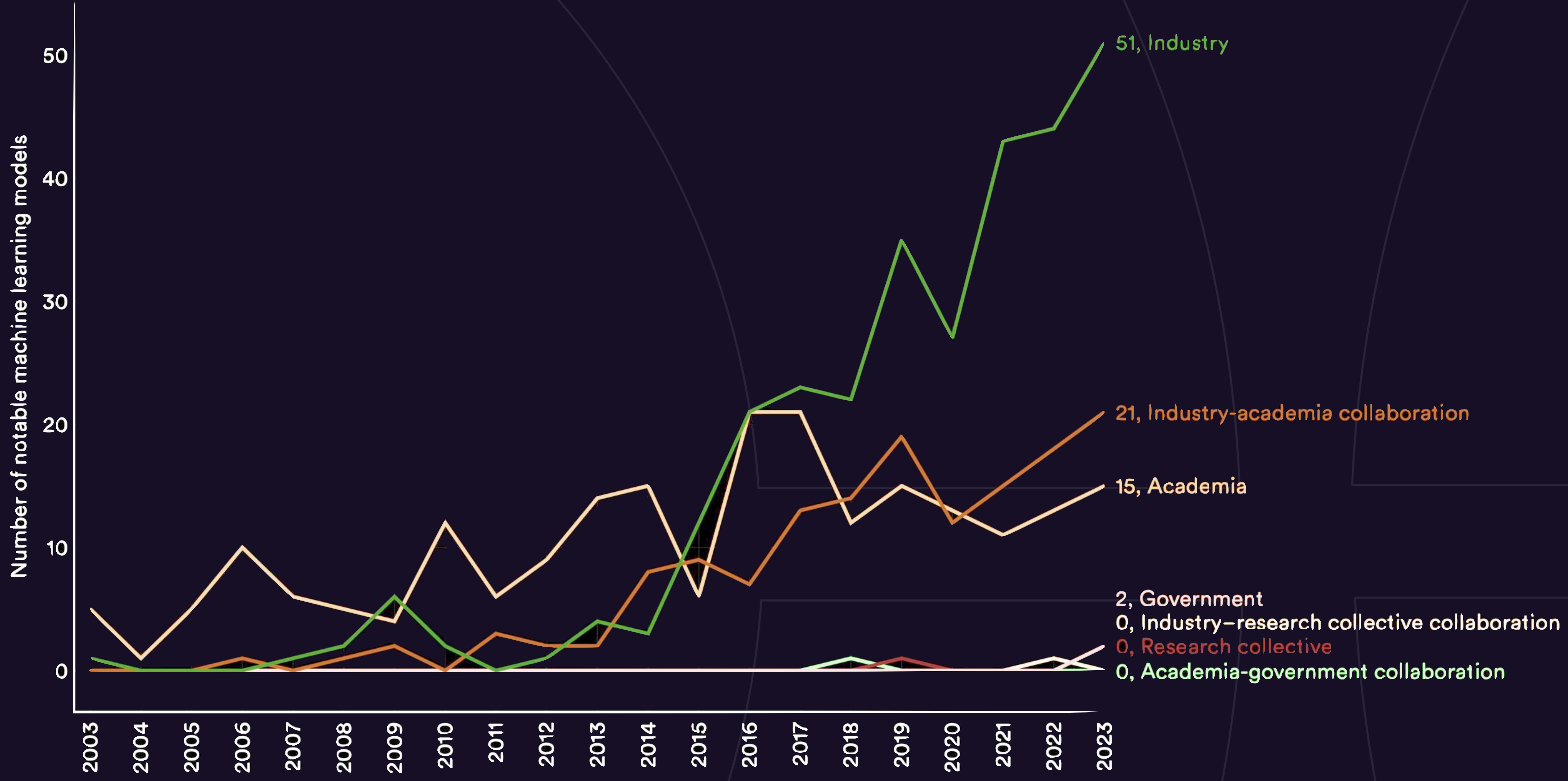


https://arxiv.org/abs/2210.00881

Artificial Intelligence

Number of notable machine learning models by sector, 2003–23

Source: Epoch, 2023 | Chart: 2024 AI Index report



https://aiindex.stanford.edu/report/

Artificial Intelligence

nextflow

nextflow-io/nf-gpt

```
def query = '''
        Extract information about a person from In 1968, amidst the fading echoes of Independenc
        a child named John arrived under the calm evening sky. This newborn, bearing the surname
        marked the start of a new journey.
        '''

def response = gptPromptForData(query, schema: [firstName: 'string', lastName: 'string', birthDa

println "First name: ${response[0].firstName}"
println "Last name: ${response[0].lastName}"
println "Birth date: ${response[0].birthDate}"
```

Artificial Intelligence

nextflow

Future: Explanation of errors?

The error message `command not found` with an exit status of **127** typically indicates that the command (in this case, `fastqc`) is not found in the system's **PATH**. This means the system does not know where to locate the `fastqc` executable to run it.

This is likely because you have not specified a container or environment profile with the `-profile` option when running the pipeline. This is essential as it ensures that all the required tools, such as `fastqc`, are available and correctly configured in a reproducible manner.

The best way to fix this is by specifying a profile that suits your system's capabilities. Commonly used profiles include `docker`, `singularity`, `podman`, or `conda`. Using one of these profiles ensures that the necessary software environments are set up automatically, avoiding the need to manually manage dependencies and **PATH** configurations.

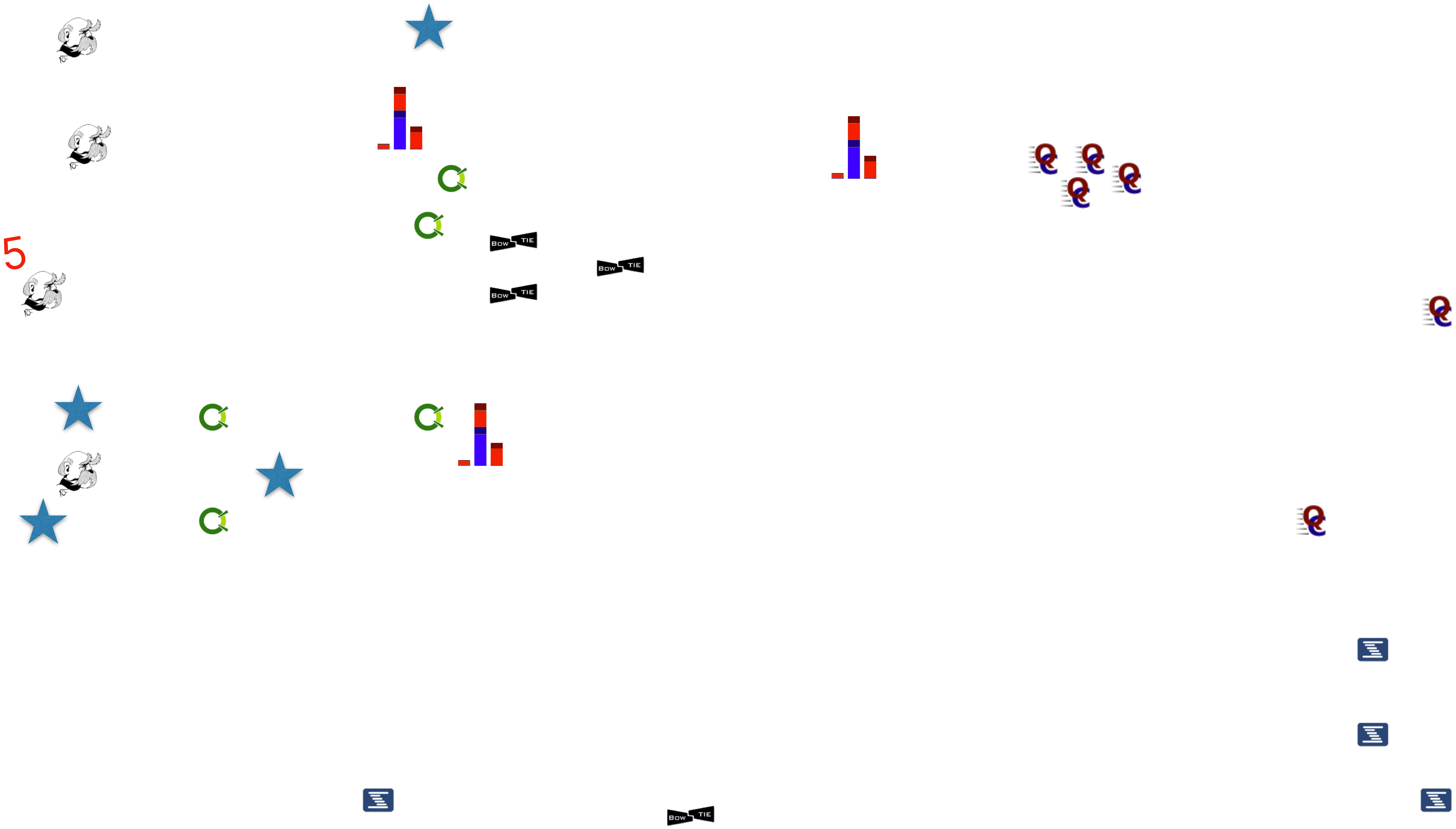
For example, if you have Docker installed, you can modify your command like this:

```
nextflow run nf-core/rnaseq --input mysamples.csv --outdir ./results -profile docker
```

Big data + AI results are only useful if they can be understood and trusted

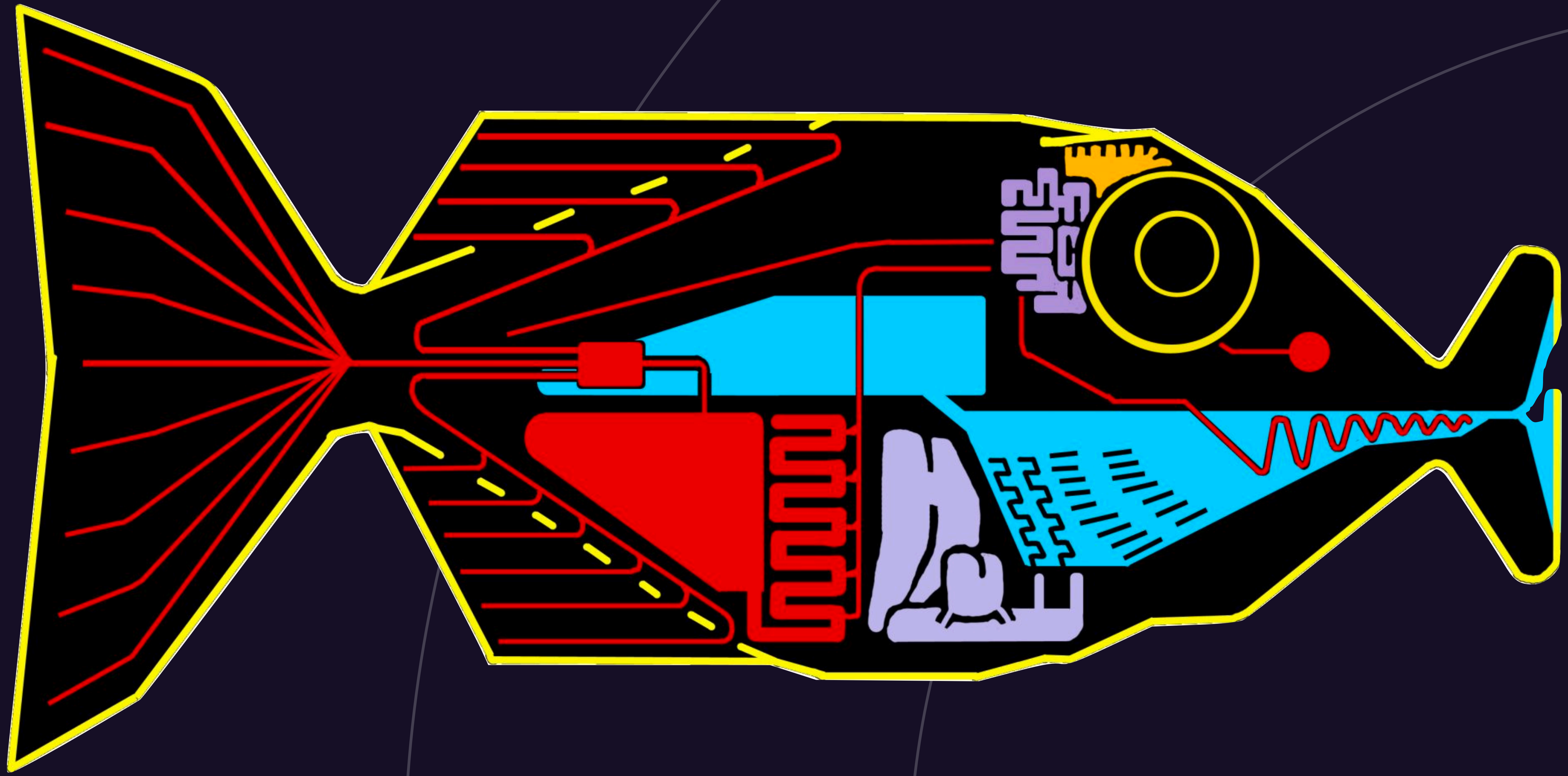
HELP
DNA-club 2015
WHERE'S WALLY?
WHERE'S WALDO

DNA-club 2015



Understanding and trust

Bioinformatics log outputs



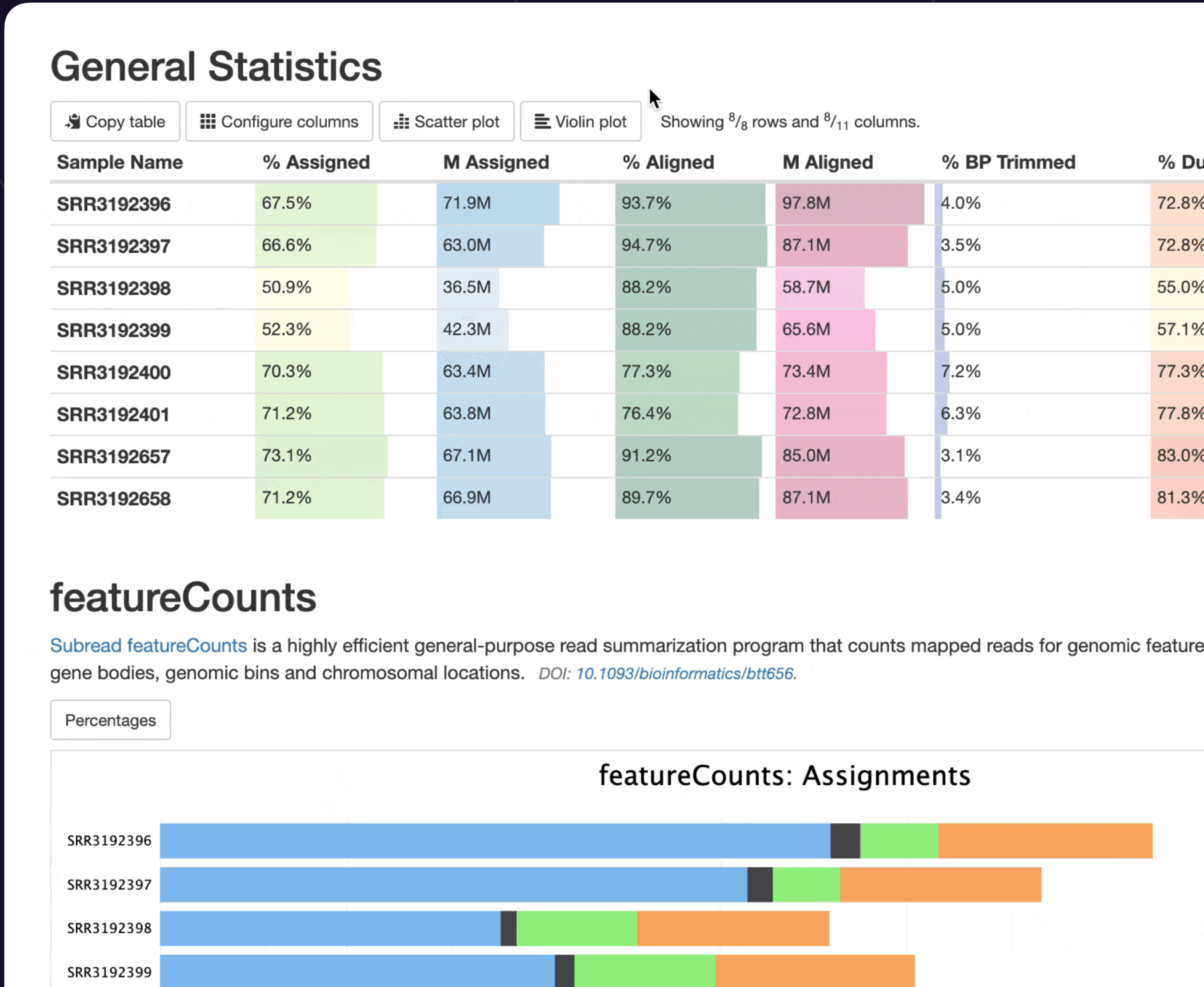
→ Human-readable report

multiqc

Understanding and trust

multiqc

Visualises metrics across many tools and many samples



General Statistics

Copy table | Configure columns | Scatter plot | Violin plot

Showing $^{8}/_{8}$ rows and $^{8}/_{11}$ columns.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% BP Trimmed	% Du
SRR3192396	67.5%	71.9M	93.7%	97.8M	4.0%	72.8%
SRR3192397	66.6%	63.0M	94.7%	87.1M	3.5%	72.8%
SRR3192398	50.9%	36.5M	88.2%	58.7M	5.0%	55.0%
SRR3192399	52.3%	42.3M	88.2%	65.6M	5.0%	57.1%
SRR3192400	70.3%	63.4M	77.3%	73.4M	7.2%	77.3%
SRR3192401	71.2%	63.8M	76.4%	72.8M	6.3%	77.8%
SRR3192657	73.1%	67.1M	91.2%	85.0M	3.1%	83.0%
SRR3192658	71.2%	66.9M	89.7%	87.1M	3.4%	81.3%

featureCounts

Subread featureCounts is a highly efficient general-purpose read summarization program that counts mapped reads for genomic feature gene bodies, genomic bins and chromosomal locations. *DOI: 10.1093/bioinformatics/btt656.*

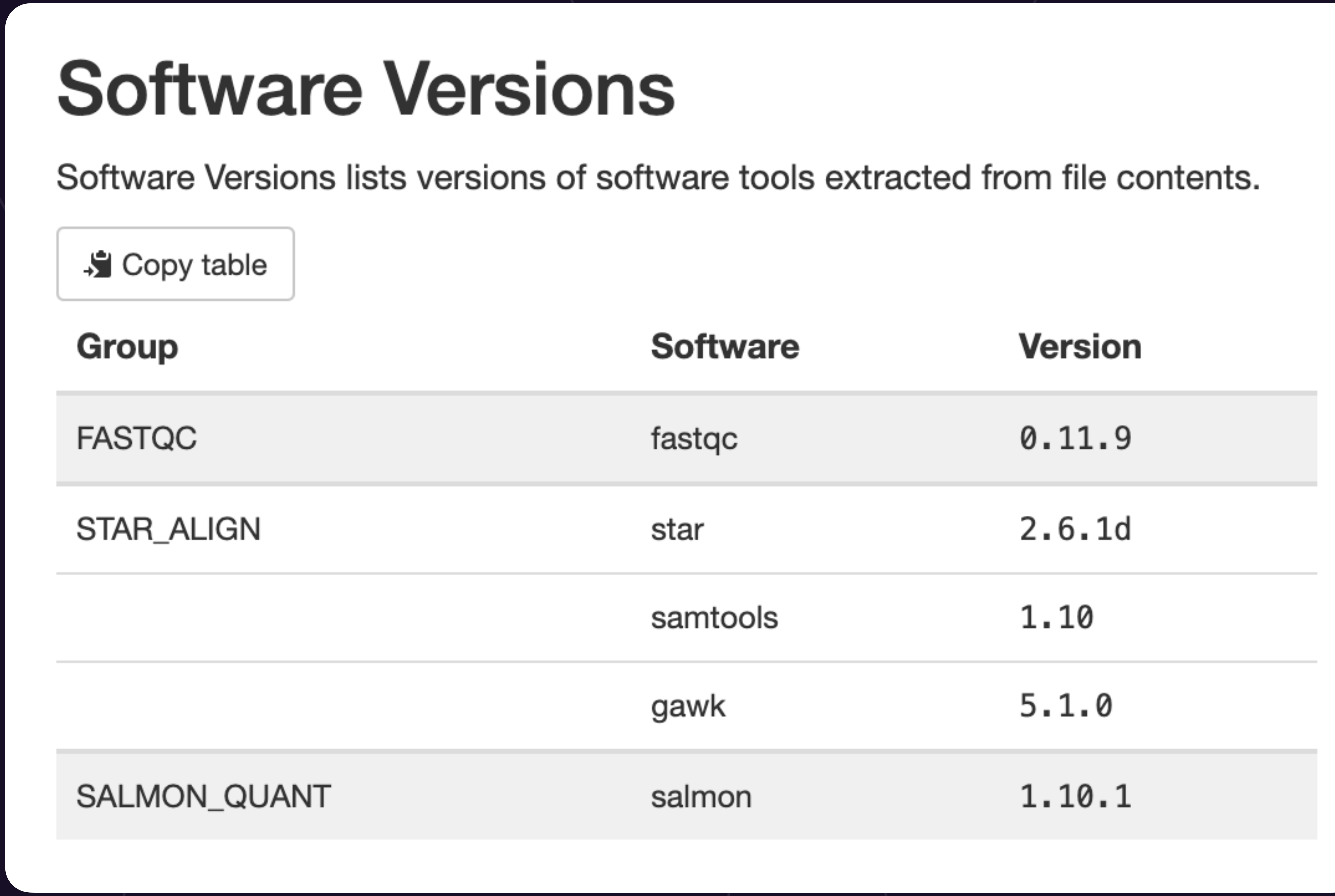
Percentages

Understanding and trust

multiqc

Visualises metrics across many tools and many samples

Collects software versions automatically



Software Versions

Software Versions lists versions of software tools extracted from file contents.

Group	Software	Version
FASTQC	fastqc	0.11.9
STAR_ALIGN	star	2.6.1d
	samtools	1.10
	gawk	5.1.0
SALMON_QUANT	salmon	1.10.1

Understanding and trust

SWEDAC

FDA U.S. FOOD & DRUG ADMINISTRATION

SAFE DATA SAFE PROJECTS SAFE PEOPLE SAFE SETTINGS SAFE OUTPUTS

Five Safes

RO-Crate

BioCompute Objects

Understanding and trust

nextflow

nextflow-io/nf-prov

Automatically generate standards-compliant provenance reports

RO-Crate

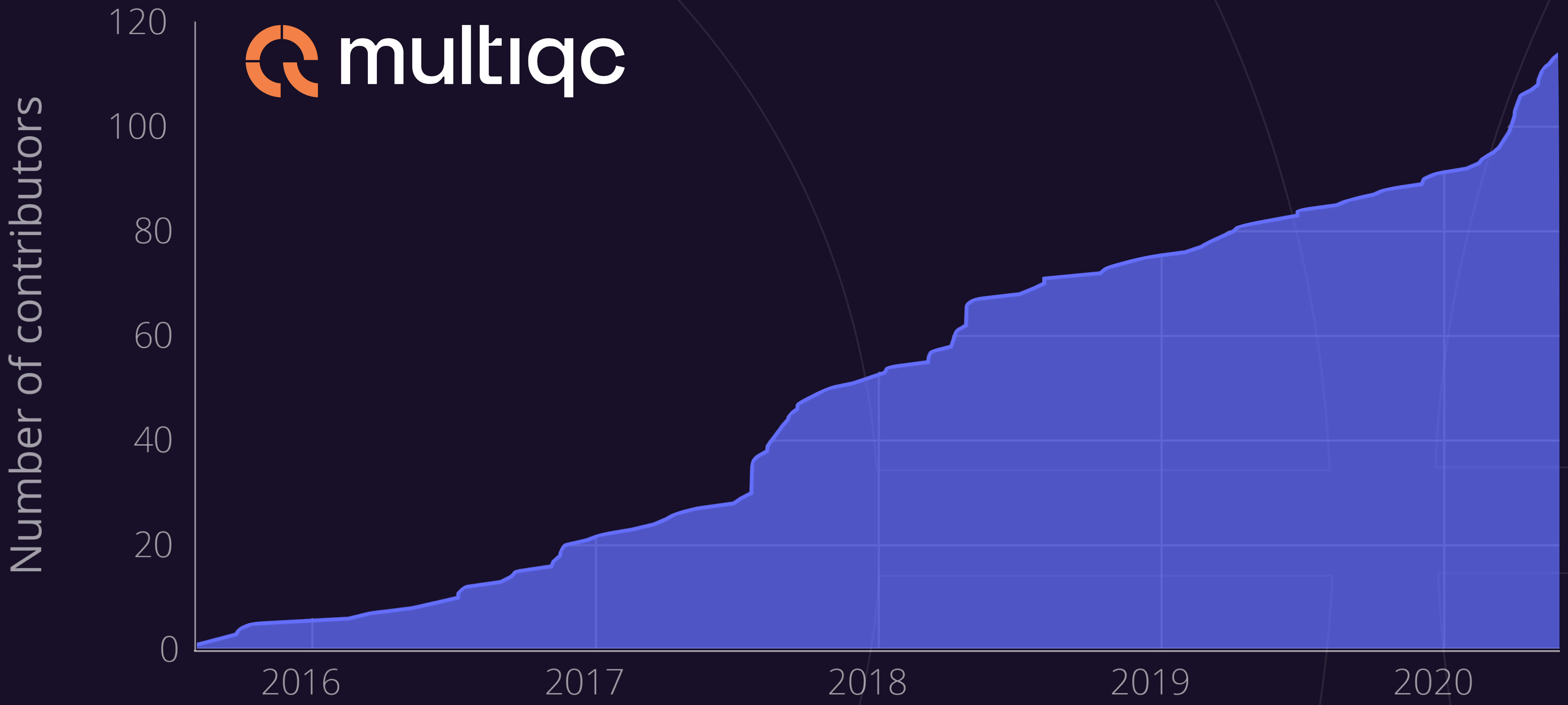
nf-core

BioCompute Objects

Automatically find contributors and link to GitHub and ORCiD

On the importance of being open

On the importance of being open



On the importance of being open

multiqc

0.45
per day

⊙ Open issue

0.23
per day

Open pull request

On the importance of being open

Reproducible analysis of genomics data at scale

Founded the nf-core community - by removing institutional branding

Now 9000 community members on Slack and still growing fast

SciLifeLab

NATIONAL GENOMICS INFRASTRUCTURE

nextflow

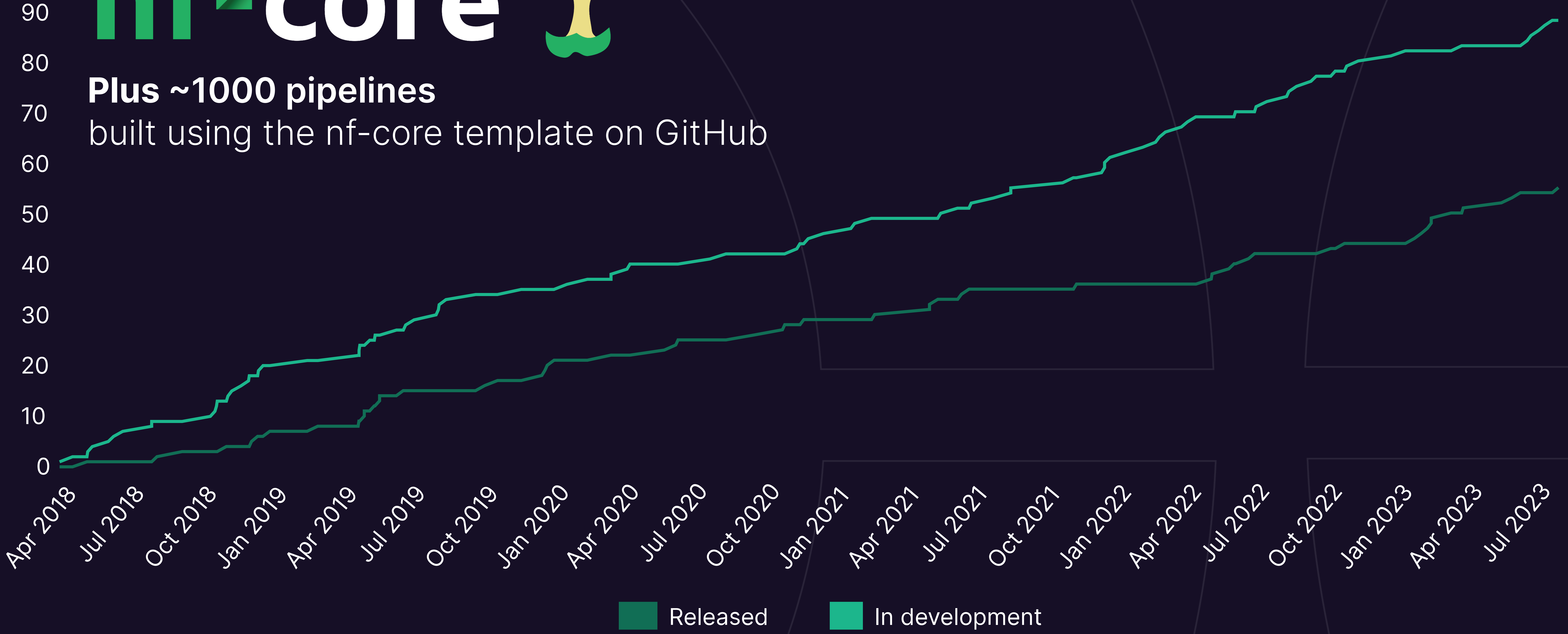
nf-core

On the importance of being open

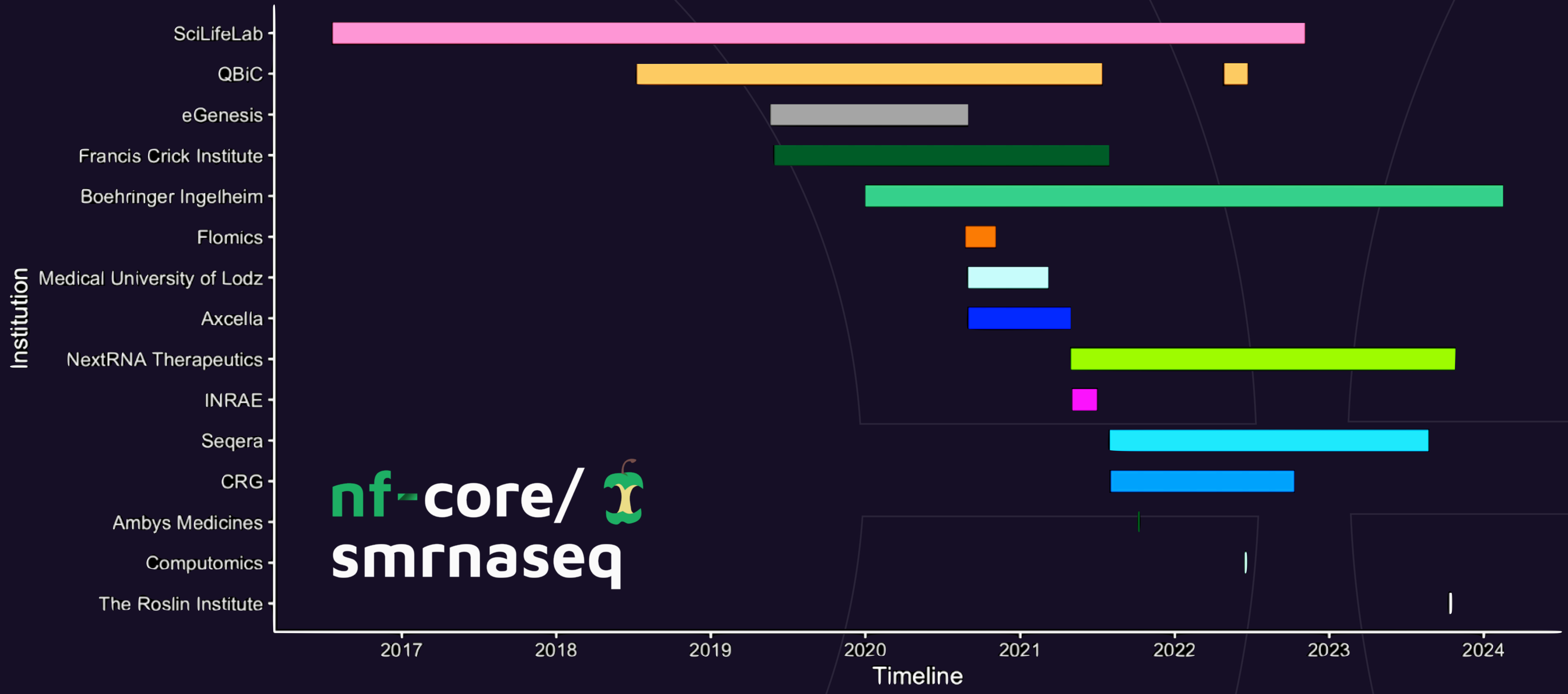
nf-core

Plus ~1000 pipelines

built using the nf-core template on GitHub



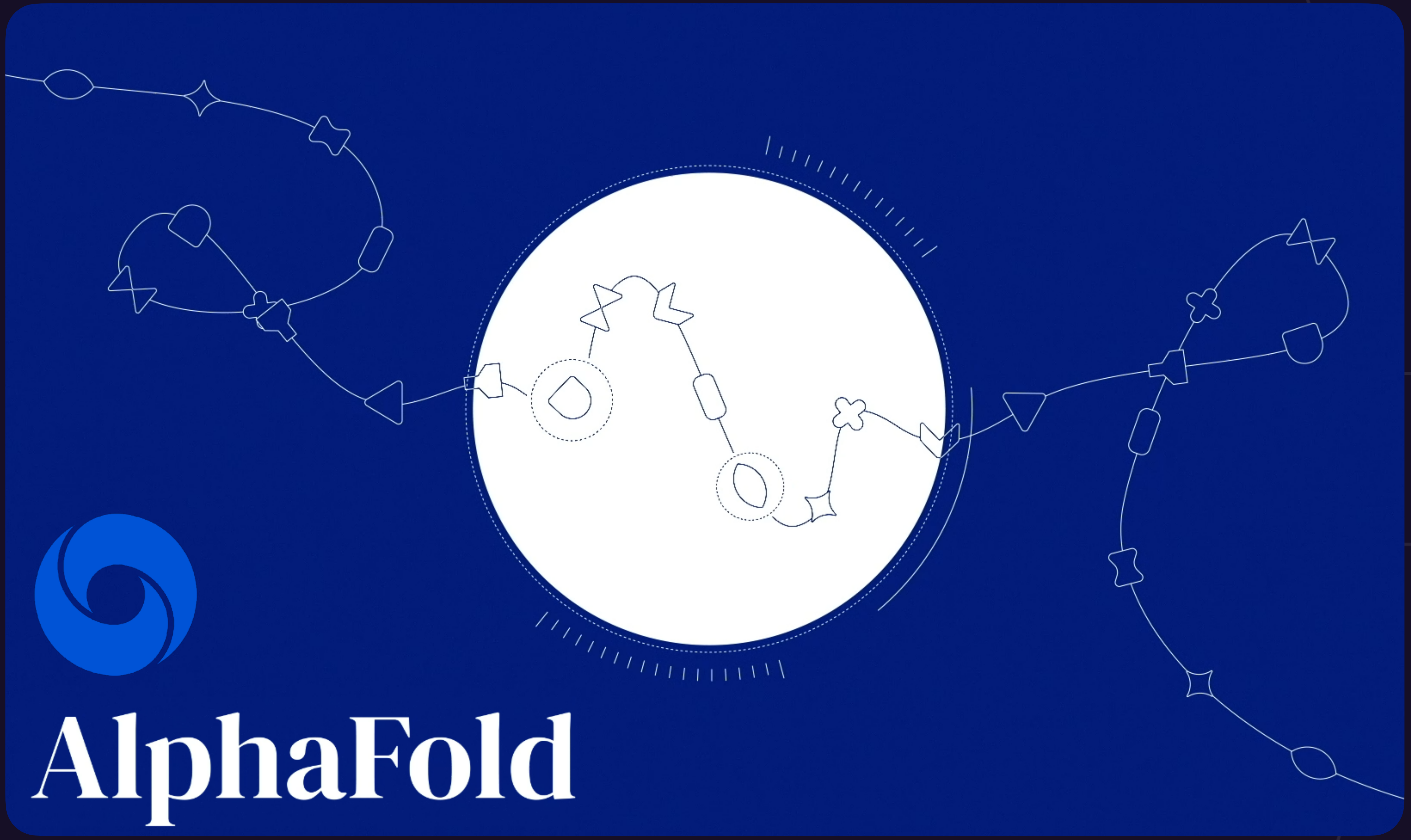
On the importance of being open



nf-core/ smrnaseq

On the importance of being open

Hugging Face



SciLifeLab
Serve

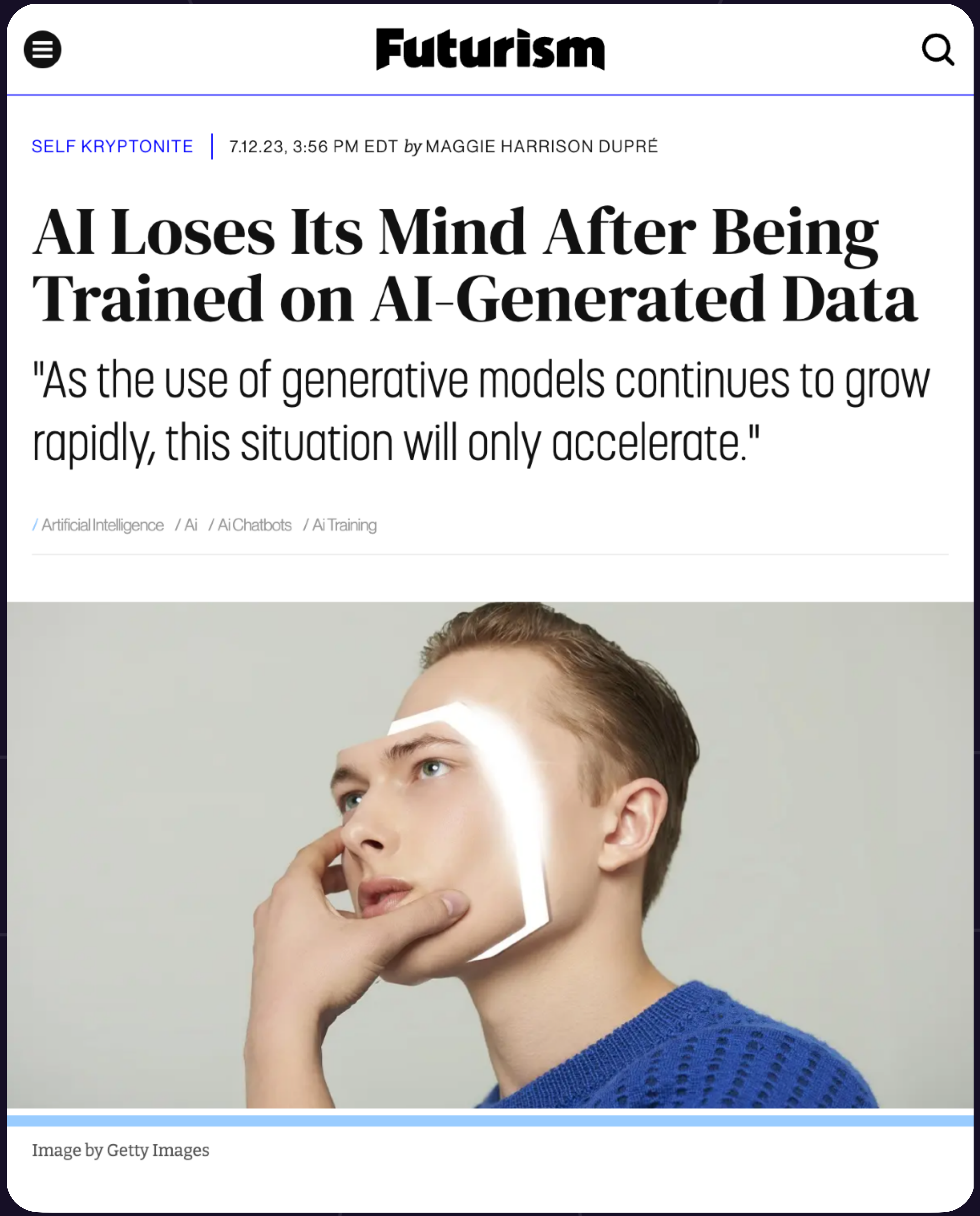
On the importance of being open

nf-core

Domain experts are more important than ever

New for nf-core:
Special interest groups

Inter-disciplinary collaboration can yield incredible results



Futurism

SELF KRYPTONITE | 7.12.23, 3:56 PM EDT *by* MAGGIE HARRISON DUPRÉ

AI Loses Its Mind After Being Trained on AI-Generated Data

"As the use of generative models continues to grow rapidly, this situation will only accelerate."

/ Artificial Intelligence / Ai / Ai Chatbots / Ai Training

Image by Getty Images

https://futurism.com/ai-trained-ai-generated-data

Future challenges

Future challenges

Accessible to anyone

Intuitive and easy to use for people from any background.

Suitable for any use case

Generalist interface that can be used with any data type.

Powerful at any scale

Can be run on your laptop or scaled to a production cluster with millions of samples.

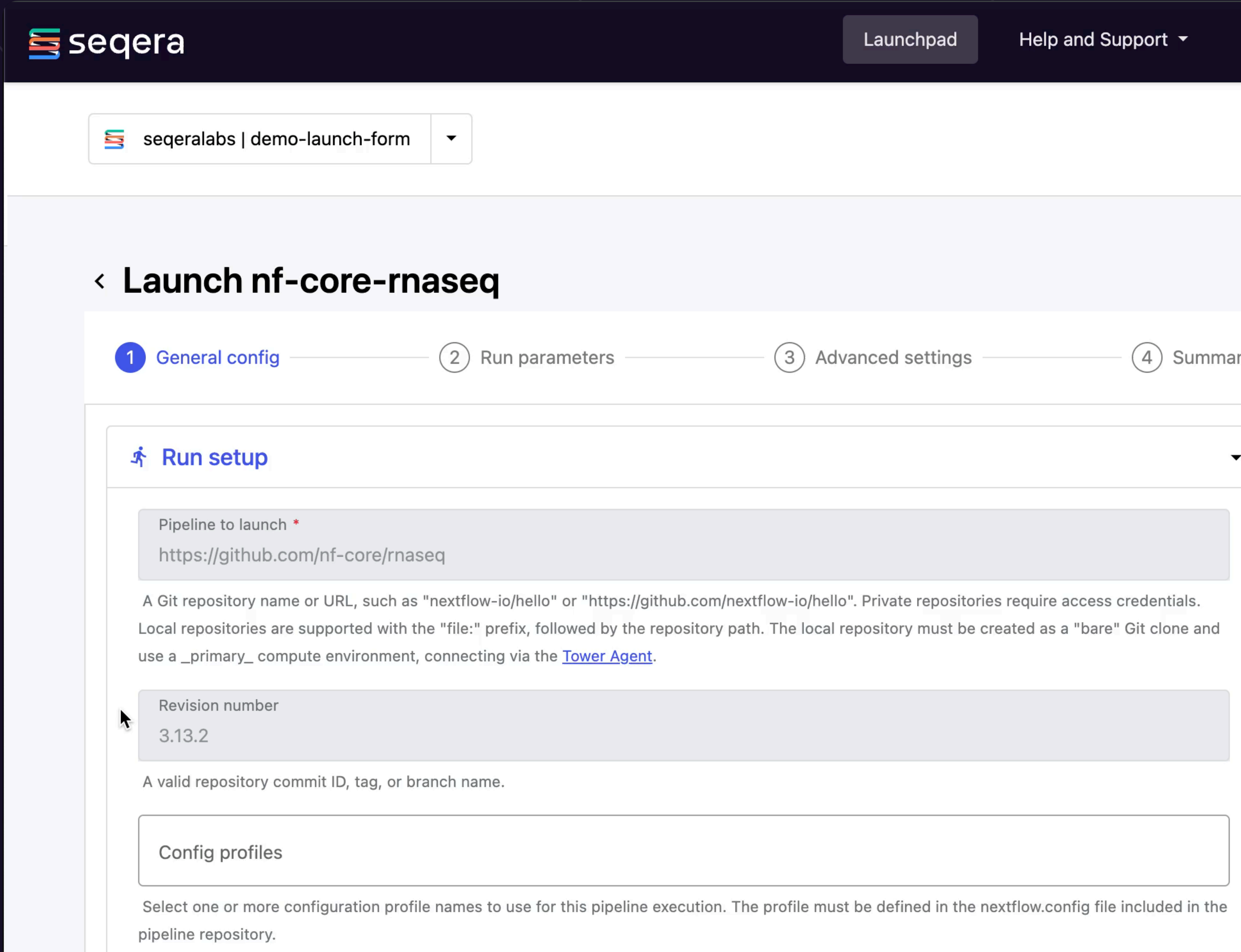
(choose two?)

Future challenges

Accessible to anyone

Suitable for any use case

Powerful at any scale



Launch nf-core-rnaseq

Run setup

Pipeline to launch *
https://github.com/nf-core/rnaseq

A Git repository name or URL, such as "nextflow-io/hello" or "https://github.com/nextflow-io/hello". Private repositories require access credentials. Local repositories are supported with the "file:" prefix, followed by the repository path. The local repository must be created as a "bare" Git clone and use a _primary_ compute environment, connecting via the Tower Agent.

Revision number
3.13.2

A valid repository commit ID, tag, or branch name.

Config profiles

Select one or more configuration profile names to use for this pipeline execution. The profile must be defined in the nextflow.config file included in the pipeline repository.

Future challenges

Generally useful

Useful for the majority of people running with this type of data / analysis.

Specific

Analysis that can be applied to a specific research question.

Maintainable

Clean code base without excessive logic or parameter space.

The "final mile" of analysis

Future challenges

Generally useful

Specific

Maintainable

nf-core

Modularity of components

Chaining of workflows

Importing and extending workflows

Future challenges

Generally useful

Specific

Maintainable

nf-core

multiqc

"Custom content"

Use as library / use within notebooks

Future challenges

Generally useful

Specific

Maintainable

nf-core

multiqc

seqera

Data Studios - eg. Notebooks, but any container:
interactive environments for downstream analysis

Future challenges

Lossy storage

What can we afford to throw away?

Green computing

Justifying the cost of your data centre

Heterogenous data

Mixing and matching data types for new science

How much is enough?

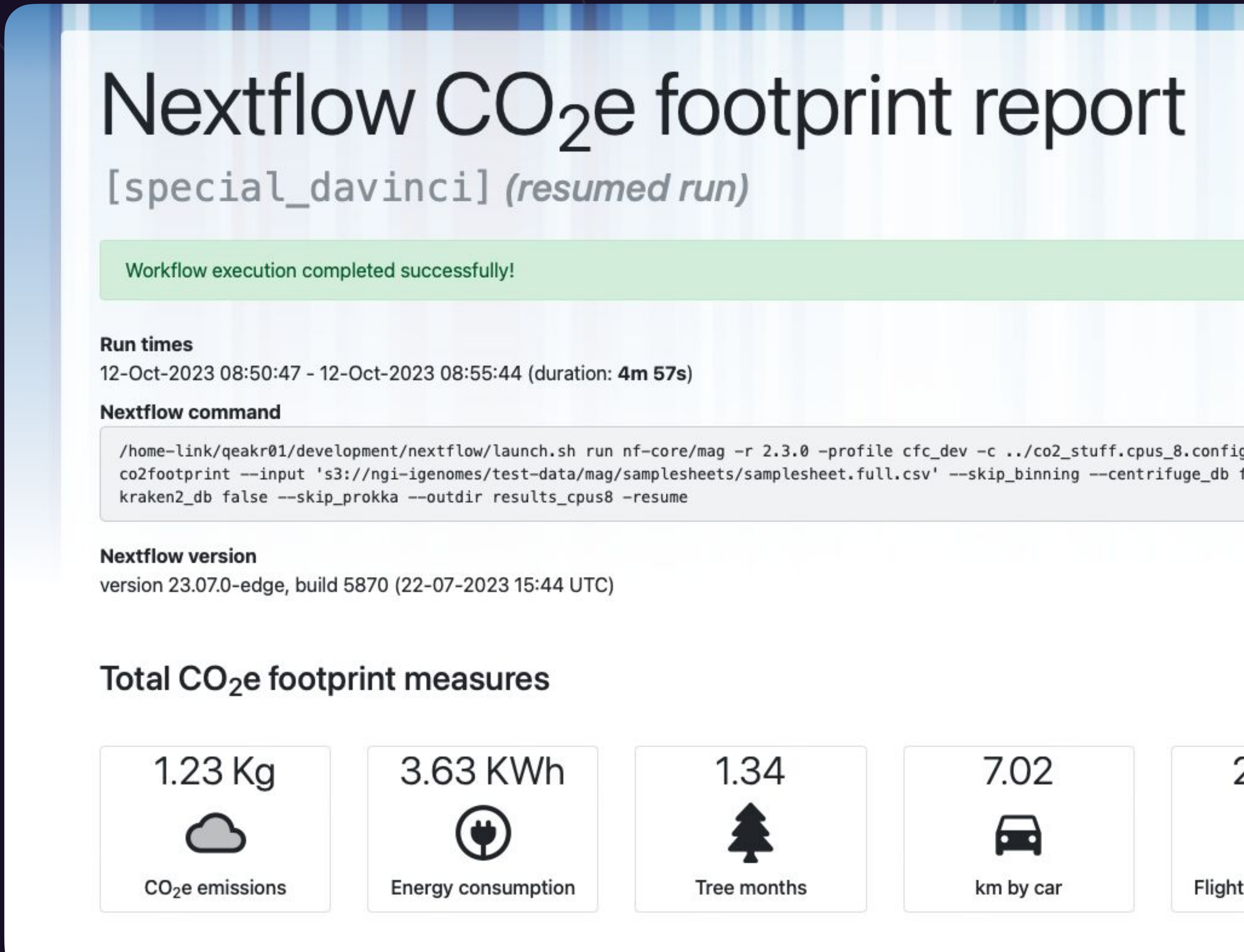
Future challenges

Loïc Lannelongue, Sabrina Krakau
green-algorithms.org
nextflow-io / nf-co2footprint

Lossy storage

Green computing

Heterogenous data



Nextflow CO$_2$e footprint report

[special_davinci] *(resumed run)*

Workflow execution completed successfully!

Run times

12-Oct-2023 08:50:47 - 12-Oct-2023 08:55:44 (duration: **4m 57s**)

Nextflow command

```
/home-link/qeakr01/development/nextflow/launch.sh run nf-core/mag -r 2.3.0 -profile cfc_dev -c ../co2_stuff.cpus_8.config
co2footprint --input 's3://ngi-igenomes/test-data/mag/samplesheets/samplesheet.full.csv' --skip_binning --centrifuge_db f
kraken2_db false --skip_prokka --outdir results_cpus8 -resume
```

Nextflow version

version 23.07.0-edge, build 5870 (22-07-2023 15:44 UTC)

Total CO$_2$e footprint measures

1.23 Kg	3.63 KWh	1.34	7.02	2
☁	⏻	🌲	🚗	
CO$_2$e emissions	Energy consumption	Tree months	km by car	Flight

Conclusion

Conclusion

Join at slido.com

#3201 014

Conclusion

Work in the open

Find your next collaborator online!

Join initiatives

Don't assume you need to DIY

Build for the future

A tool is for life, not just for Christmas

Phil Ewels, PhD

Product Manager for Open Source

phil.ewels@seqera.io

seqera