Cultivating High-Performance microbial genomics with microBioRust

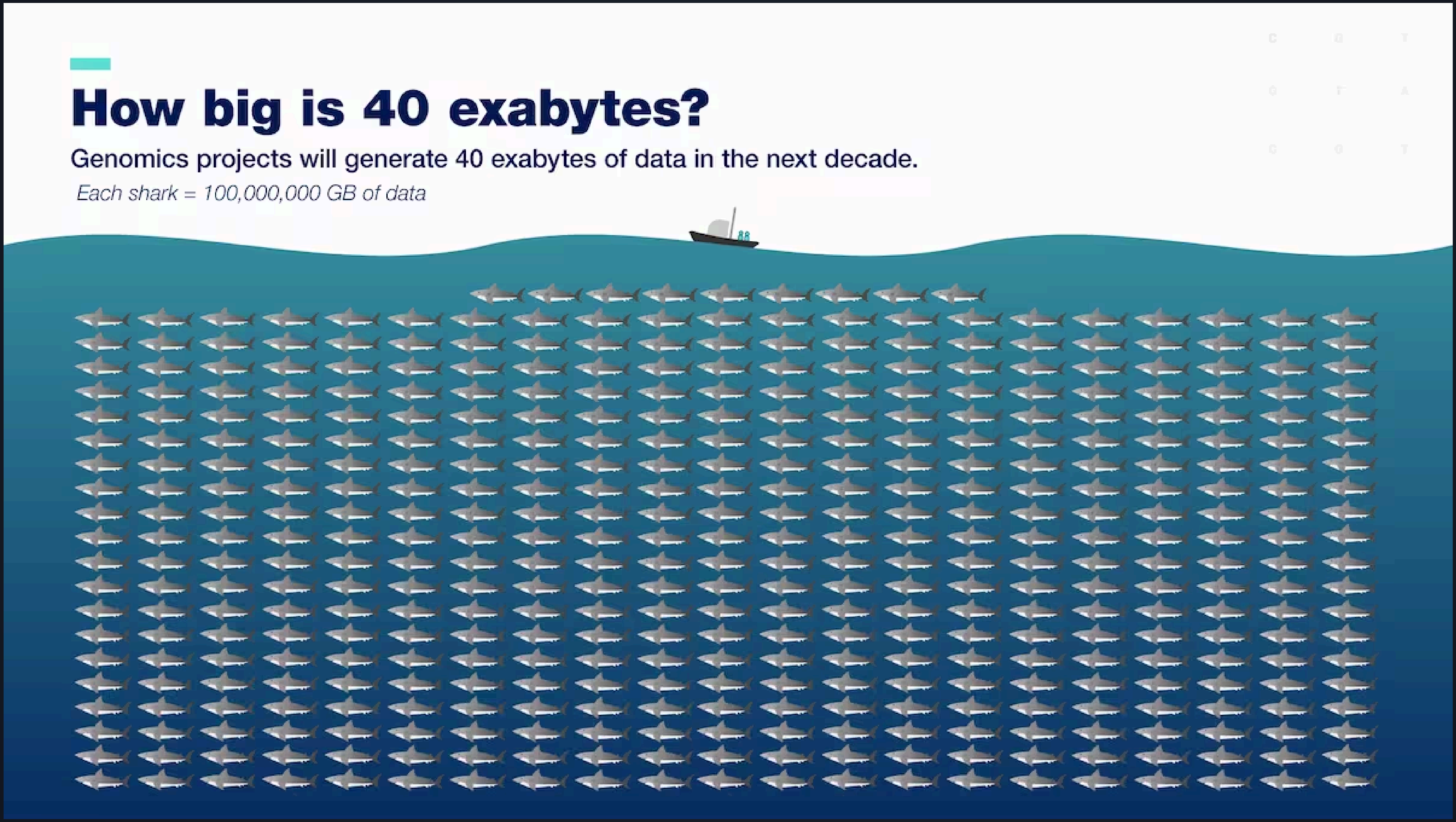
explore this»

Dr Lisa Crossman
SequenceAnalysis.co.uk
Phylonoe

image from @CodingArtist

Norfolk County Council
Go Digital
UEA
University of East Anglia
norwich research park
Interreg
France (Channel Manche) England
C-Care
European Regional Development Fund

Genomic data is expanding



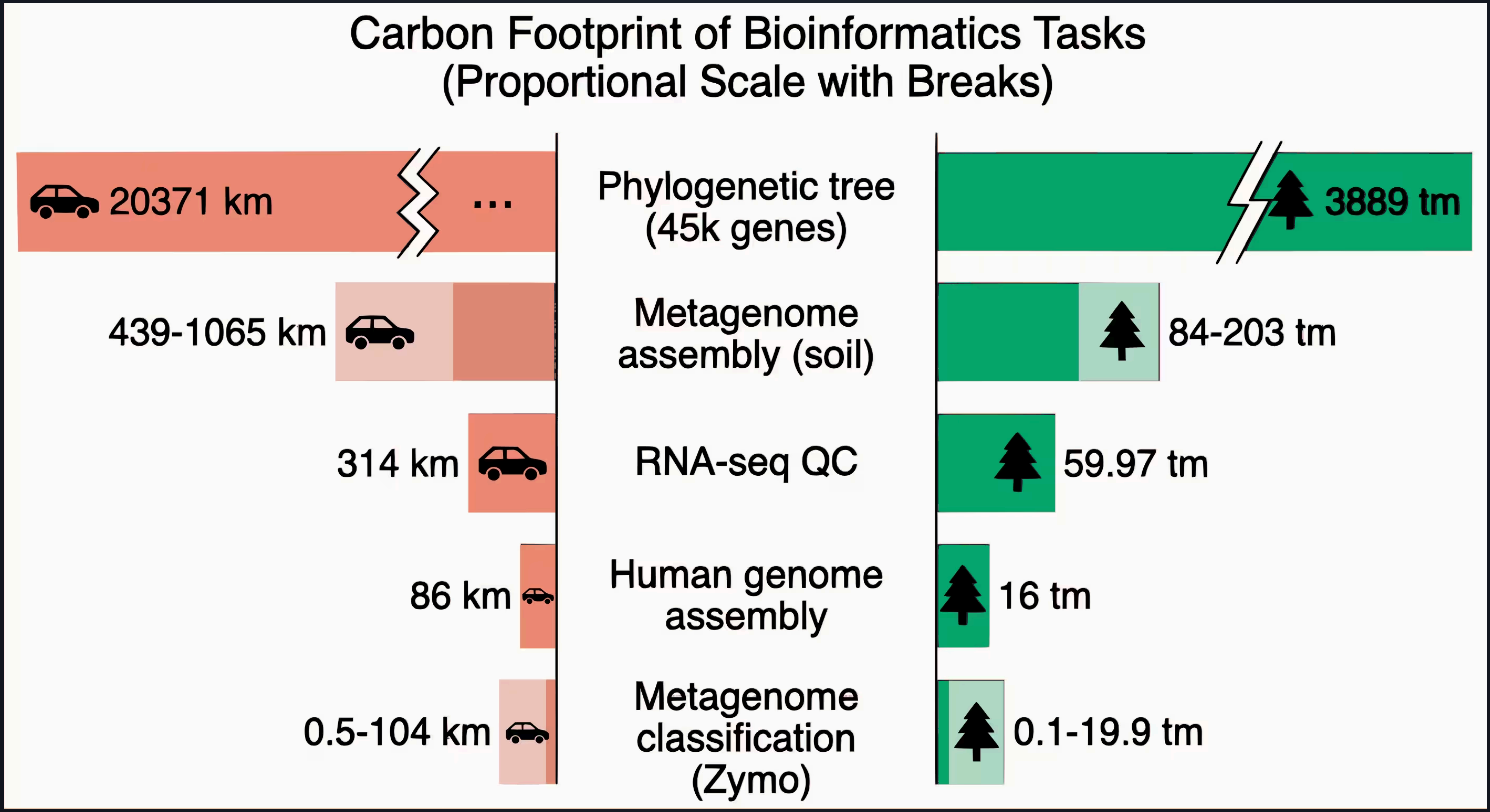
How big is 40 exabytes?

Genomics projects will generate 40 exabytes of data in the next decade.

Each shark = 100,000,000 GB of data

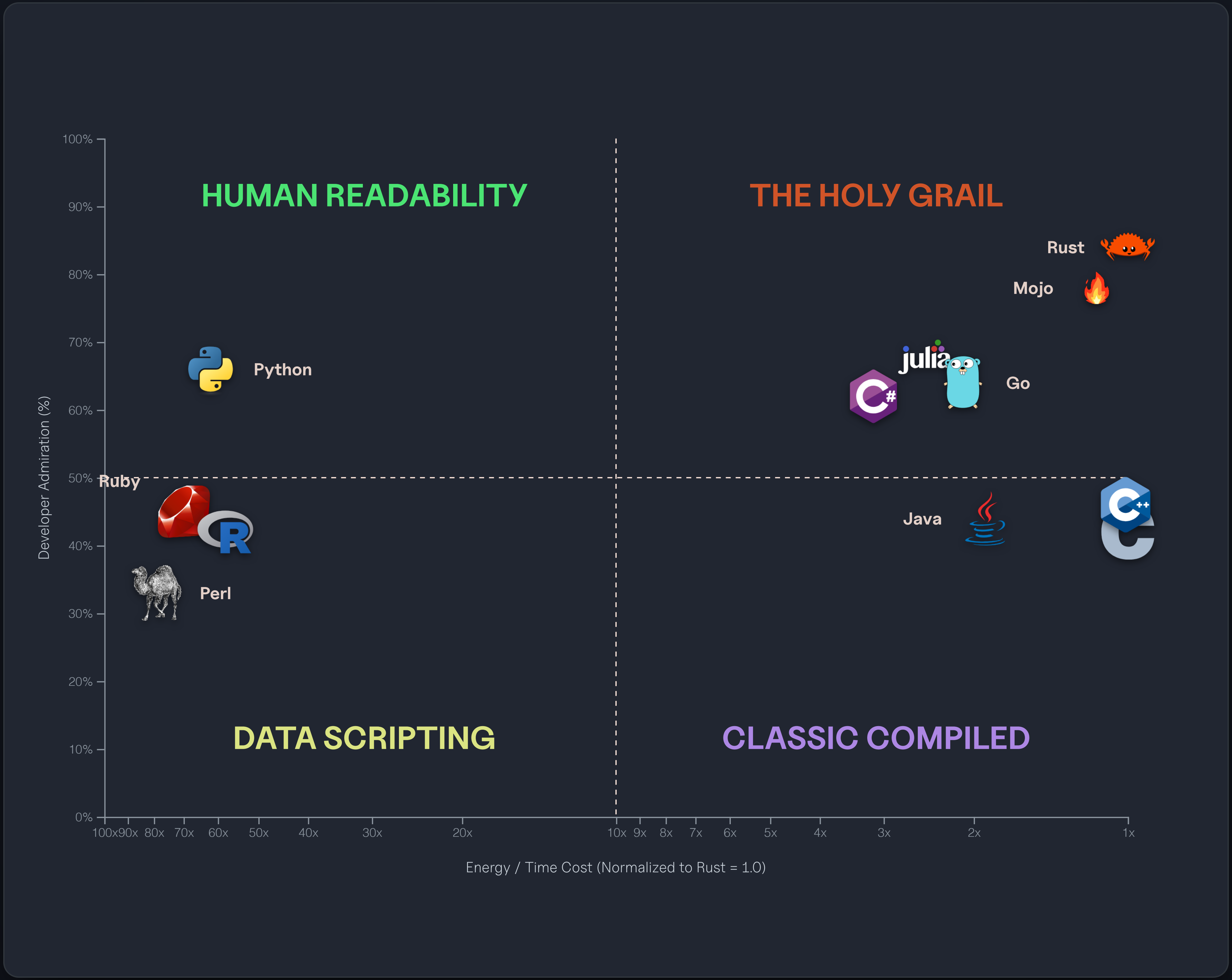
The Carbon footprint of bioinformatics

Results were compared with distances travelled by car (km) and carbon sequestered by trees (Grealey *et al.*, 2022)



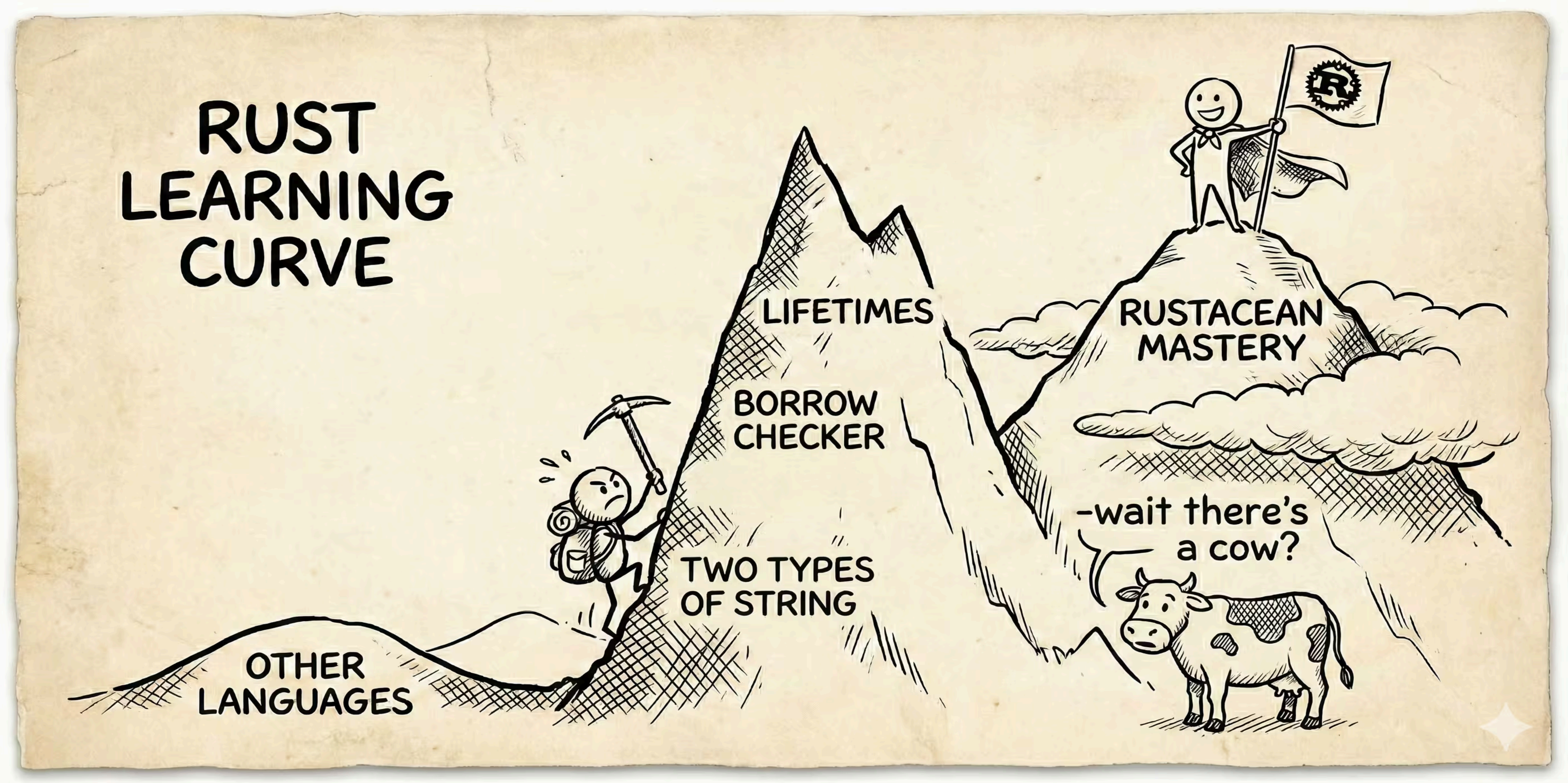
Energy efficient ranks

Efficiency (Benchmarks Game) vs Admiration (Stack Overflow)



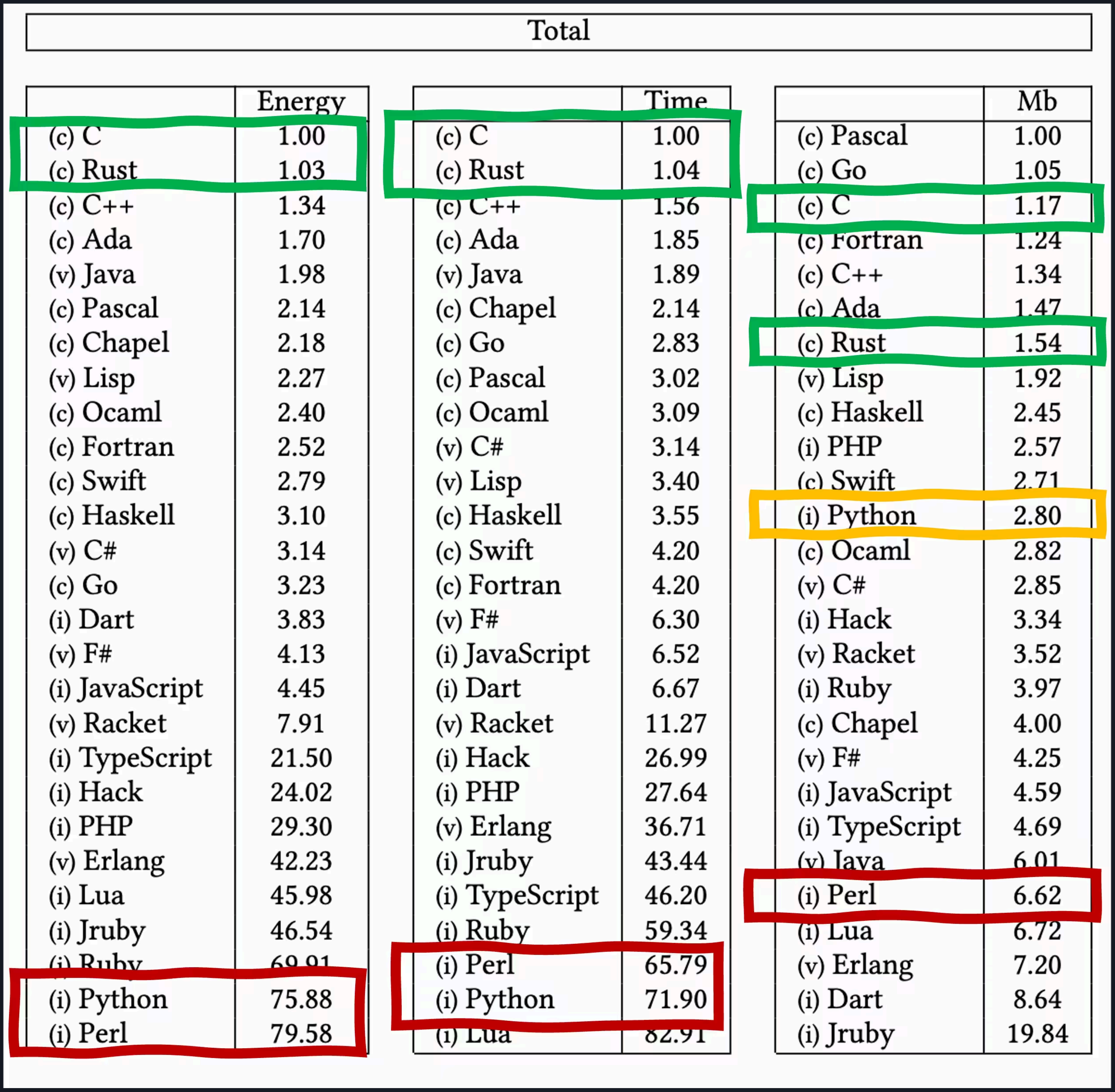
C and Rust consume up to 70% less energy (Green software foundation)

"Rust is for systems engineers"



But should I learn Rust - is it even for data science?

The data that convinced me



Total

	Energy		Time		Mb
(c) C	1.00	(c) C	1.00	(c) Pascal	1.00
(c) Rust	1.03	(c) Rust	1.04	(c) Go	1.05
(c) C++	1.34	(c) C++	1.56	(c) C	1.17
(c) Ada	1.70	(c) Ada	1.85	(c) Fortran	1.24
(v) Java	1.98	(v) Java	1.89	(c) C++	1.34
(c) Pascal	2.14	(c) Chapel	2.14	(c) Ada	1.47
(c) Chapel	2.18	(c) Go	2.83	(c) Rust	1.54
(v) Lisp	2.27	(c) Pascal	3.02	(v) Lisp	1.92
(c) Ocaml	2.40	(c) Ocaml	3.09	(c) Haskell	2.45
(c) Fortran	2.52	(v) C#	3.14	(i) PHP	2.57
(c) Swift	2.79	(v) Lisp	3.40	(c) Swift	2.71
(c) Haskell	3.10	(c) Haskell	3.55	(i) Python	2.80
(v) C#	3.14	(c) Swift	4.20	(c) Ocaml	2.82
(c) Go	3.23	(c) Fortran	4.20	(v) C#	2.85
(i) Dart	3.83	(v) F#	6.30	(i) Hack	3.34
(v) F#	4.13	(i) JavaScript	6.52	(v) Racket	3.52
(i) JavaScript	4.45	(i) Dart	6.67	(i) Ruby	3.97
(v) Racket	7.91	(v) Racket	11.27	(c) Chapel	4.00
(i) TypeScript	21.50	(i) Hack	26.99	(v) F#	4.25
(i) Hack	24.02	(i) PHP	27.64	(i) JavaScript	4.59
(i) PHP	29.30	(v) Erlang	36.71	(i) TypeScript	4.69
(v) Erlang	42.23	(i) Jruby	43.44	(v) Java	6.01
(i) Lua	45.98	(i) TypeScript	46.20	(i) Perl	6.62
(i) Jruby	46.54	(i) Ruby	59.34	(i) Lua	6.72
(i) Ruby	69.91	(i) Perl	65.79	(v) Erlang	7.20
(i) Python	75.88	(i) Python	71.90	(i) Dart	8.64
(i) Perl	79.58	(i) Lua	82.91	(i) Jruby	19.84

Pereira *et al.*, 2017

Starting microBioRust

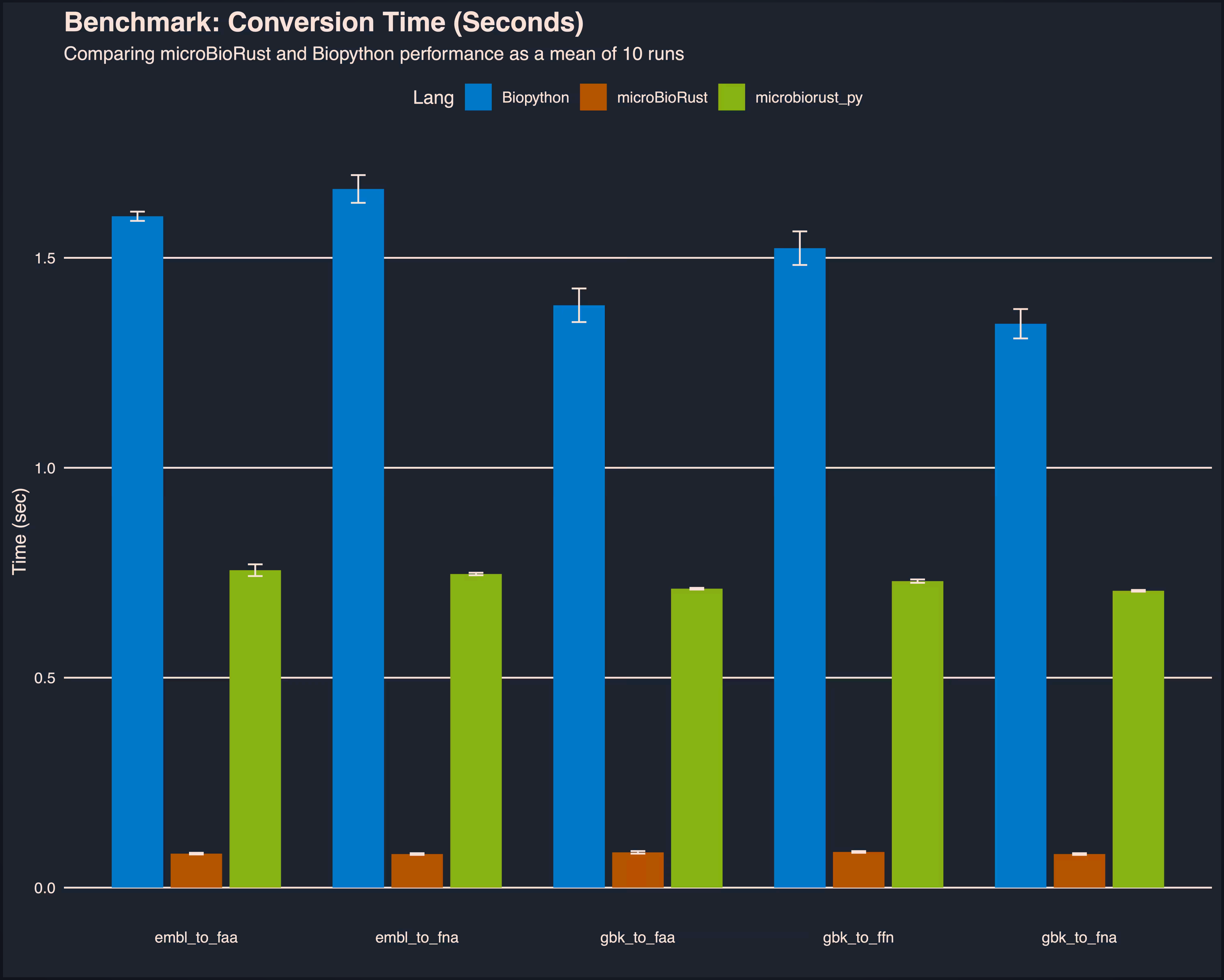
I had some code mounting up locally and I thought it could be useful to others

microBioRust - to help bioinformaticians carry out their data analysis with speed, security & simplicity

microbial genomics has different requirements

source: Bored Panda

Is it useful?


Benchmark: Conversion Time (Seconds)
Comparing microBioRust and Biopython performance as a mean of 10 runs
Lang
Biopython
microBioRust
microbiorust_py
Time (sec)
0.0
0.5
1.0
1.5
embl_to_faa
embl_to_fna
gbk_to_faa
gbk_to_ffn
gbk_to_fna

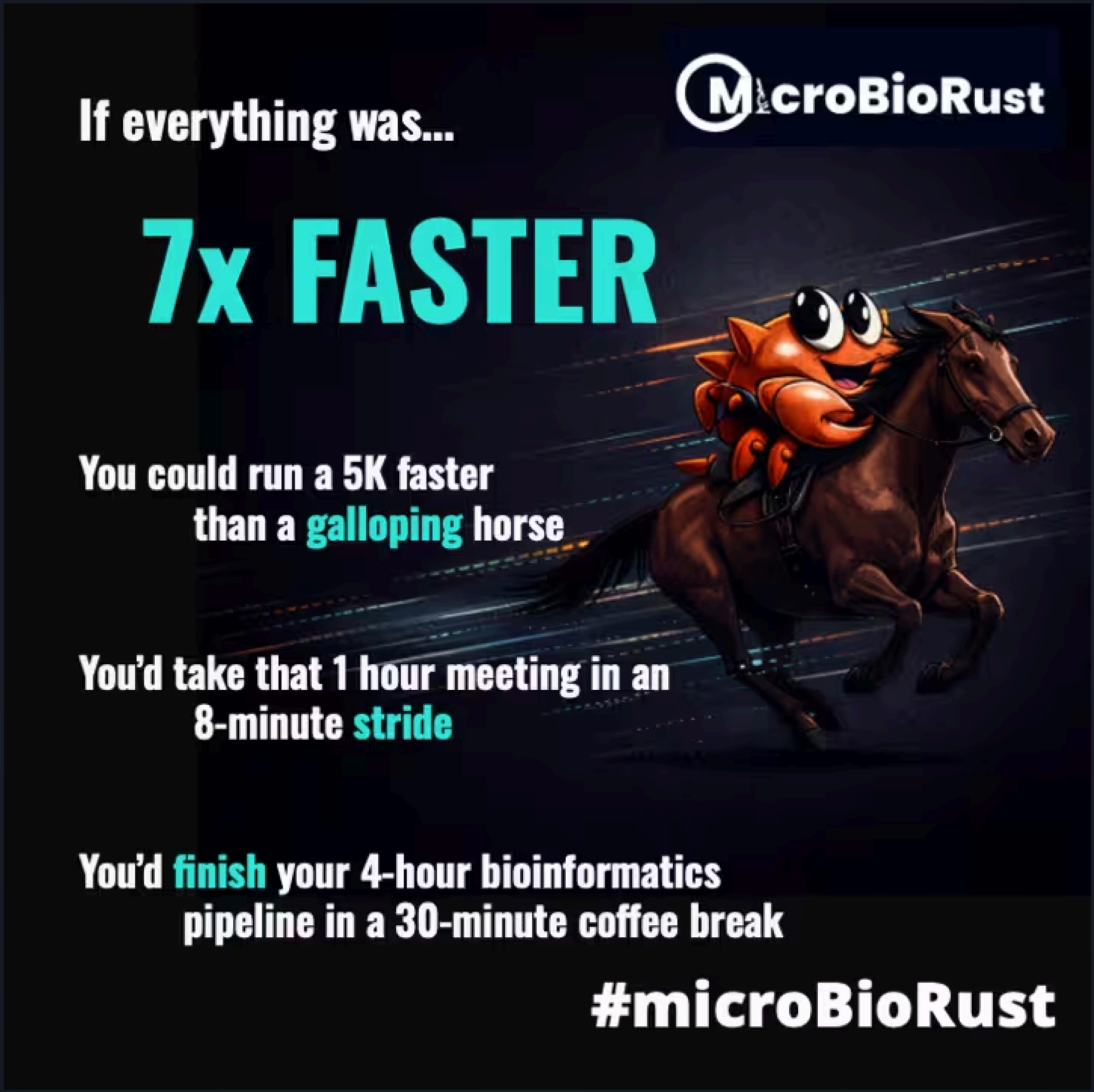
A joint Rust + Python strategy

Python Interop

```rust
fn main() -> Result<(), anyhow::Error> {
   let records = genbank!(&args.filename);
   for record in records.iter() {
      for k in record.cds.attributes.keys() {
          if let Some(seq) = record.seq_features.get_sequence_faa(k) {
              println!(">{}|{}\n{}", &record.id, &k, seq);
              }
          }
      }
   Ok(())
   }
```

```python
from microbiorust import gbk
result = gbk_to_faa("Ecoli.gbk")
```

Speed



MicroBioRust

If everything was...

7x FASTER

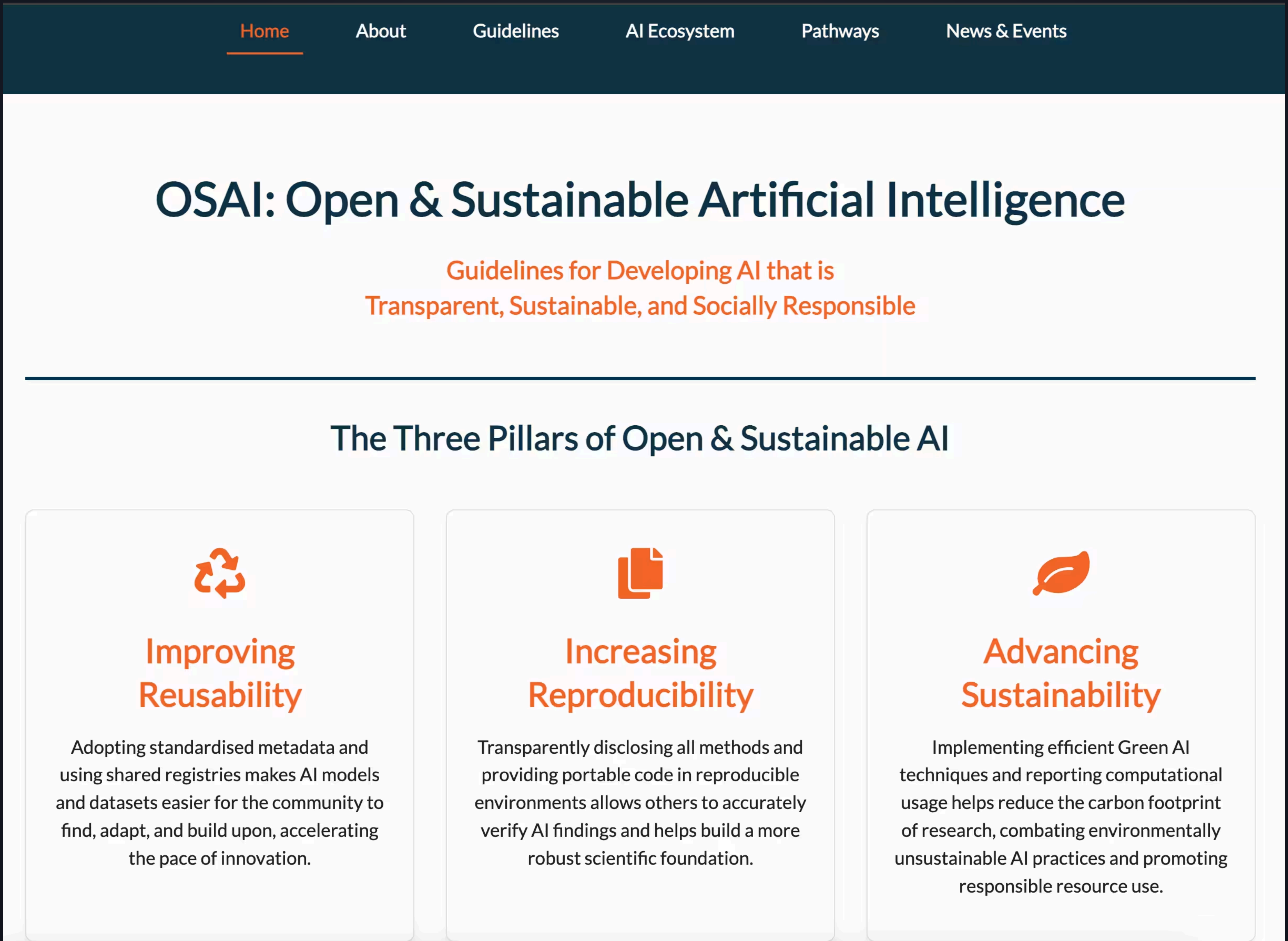
You could run a 5K faster
than a galloping horse

You'd take that 1 hour meeting in an
8-minute stride

You'd finish your 4-hour bioinformatics
pipeline in a 30-minute coffee break

#microBioRust

AI/ML are on the horizon



OSAI: Open & Sustainable Artificial Intelligence

Guidelines for Developing AI that is
Transparent, Sustainable, and Socially Responsible

The Three Pillars of Open & Sustainable AI

Improving Reusability

Adopting standardised metadata and using shared registries makes AI models and datasets easier for the community to find, adapt, and build upon, accelerating the pace of innovation.

Increasing Reproducibility

Transparently disclosing all methods and providing portable code in reproducible environments allows others to accurately verify AI findings and helps build a more robust scientific foundation.

Advancing Sustainability

Implementing efficient Green AI techniques and reporting computational usage helps reduce the carbon footprint of research, combating environmentally unsustainable AI practices and promoting responsible resource use.

Our Code, Our Climate

micro**BioRust** aims:

🚀

Fast Performance

Blazing-fast and optimized performance

🔒

Memory efficient and secure

🌱

Sustainable

Eco-friendly & easy-to-use

https://github.com/microBioRust

SequenceAnalysis.co.uk
Phylonoe

Norwich Research Park
Norwich, UK

COMMUNITY

ABOUT

about us

Bioinformatics consulting

Bioinformatics tools website

THANKS TO:

Mordecai Etukudo, Khalid Hussein, Godfrey Ogembo, Sreeram Peela microBioRust community

ELIXIR Machine Learning Focus group

Women in Rust, UK

J. Mo [rig] & Shuttle Team, now Neptune