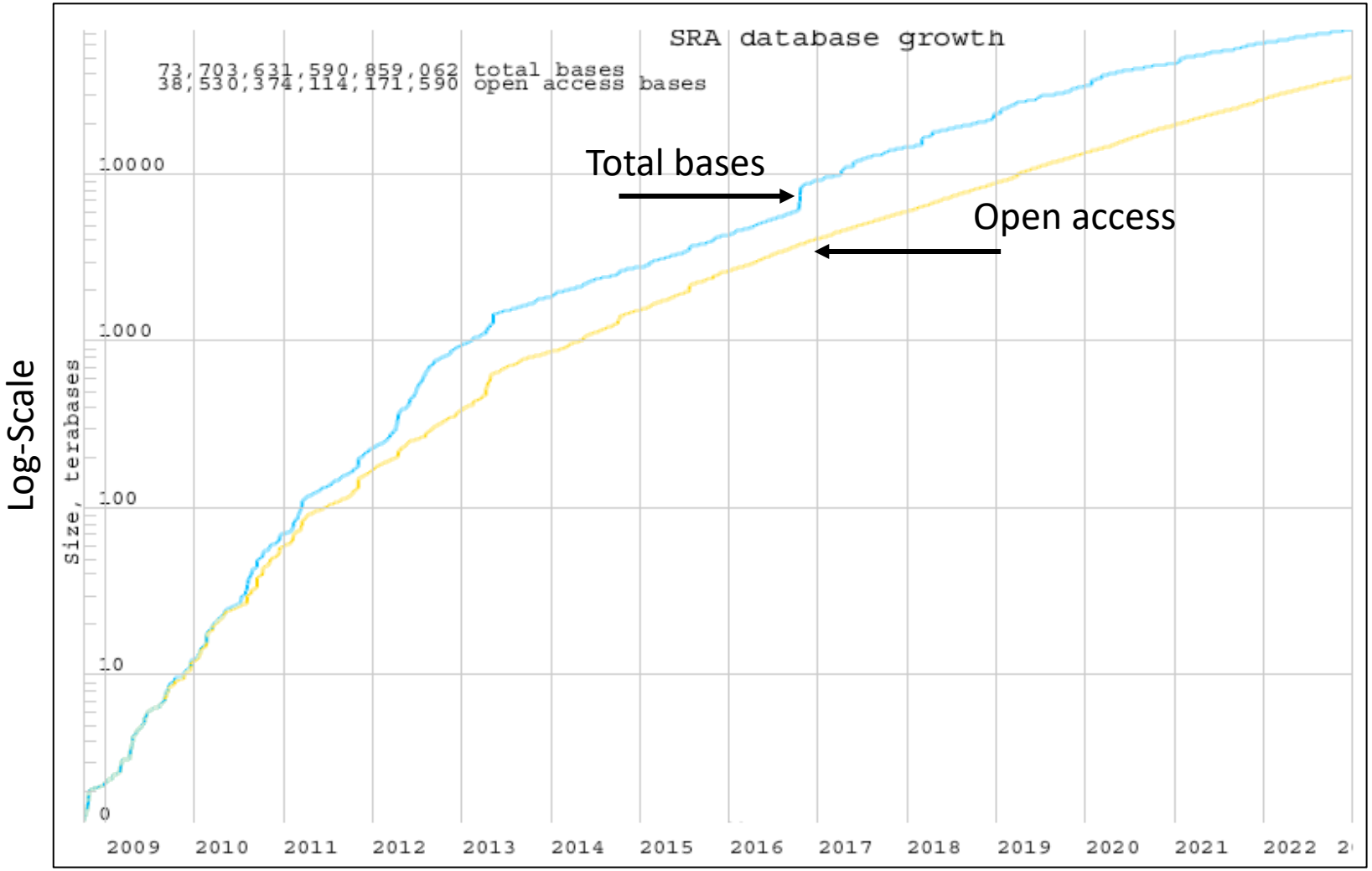
Scalability is a critical challenge in data science ft. Twitter

Professor Bioinformatics and Computational Biology
The University of Edinburgh

Professor Computer Science and Biology
Johns Hopkins University

Sequence Read Archive (SRA) growth

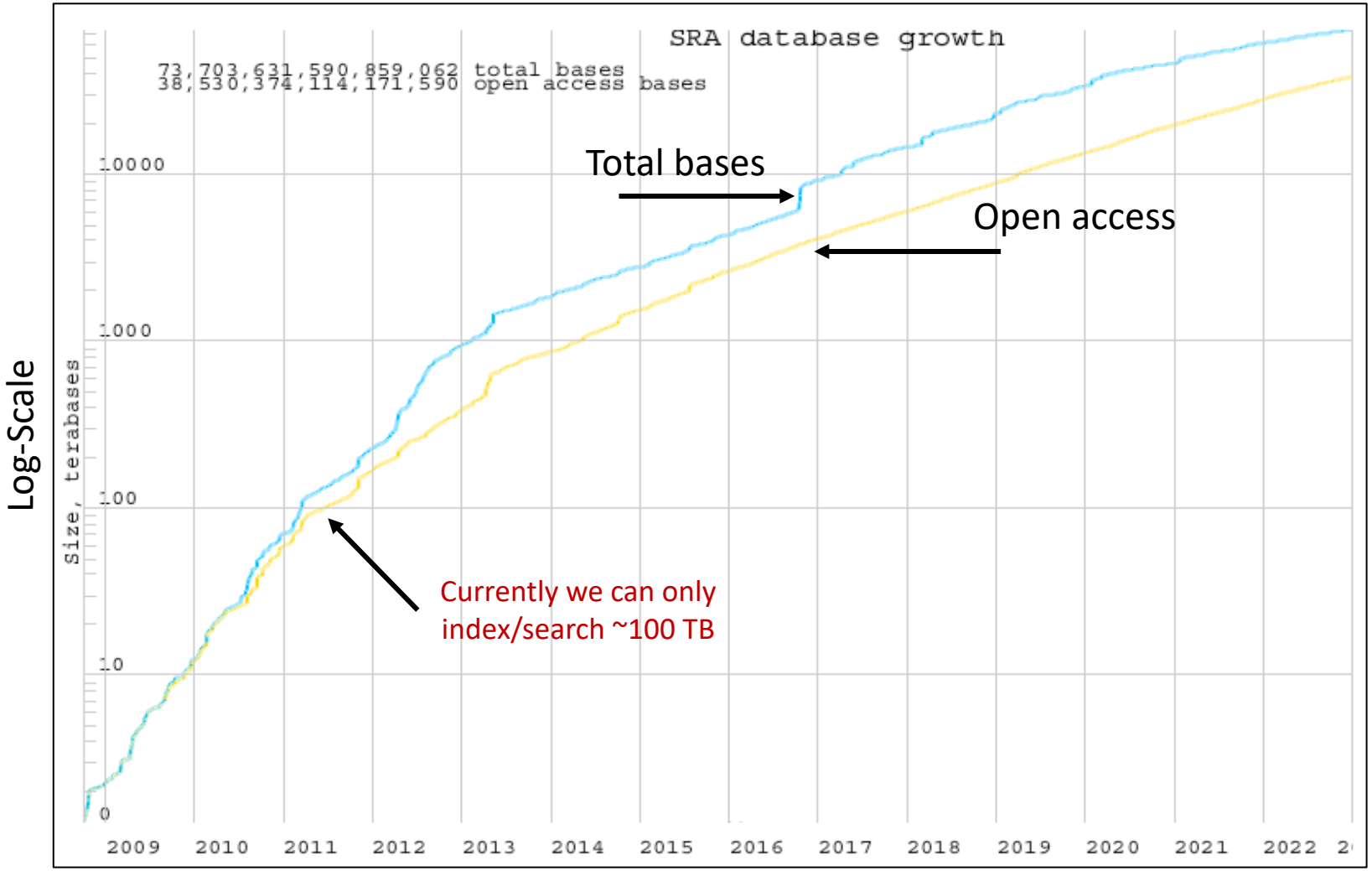
SRA contains a lot of *diversity information*



Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

Sequence Read Archive (SRA) growth

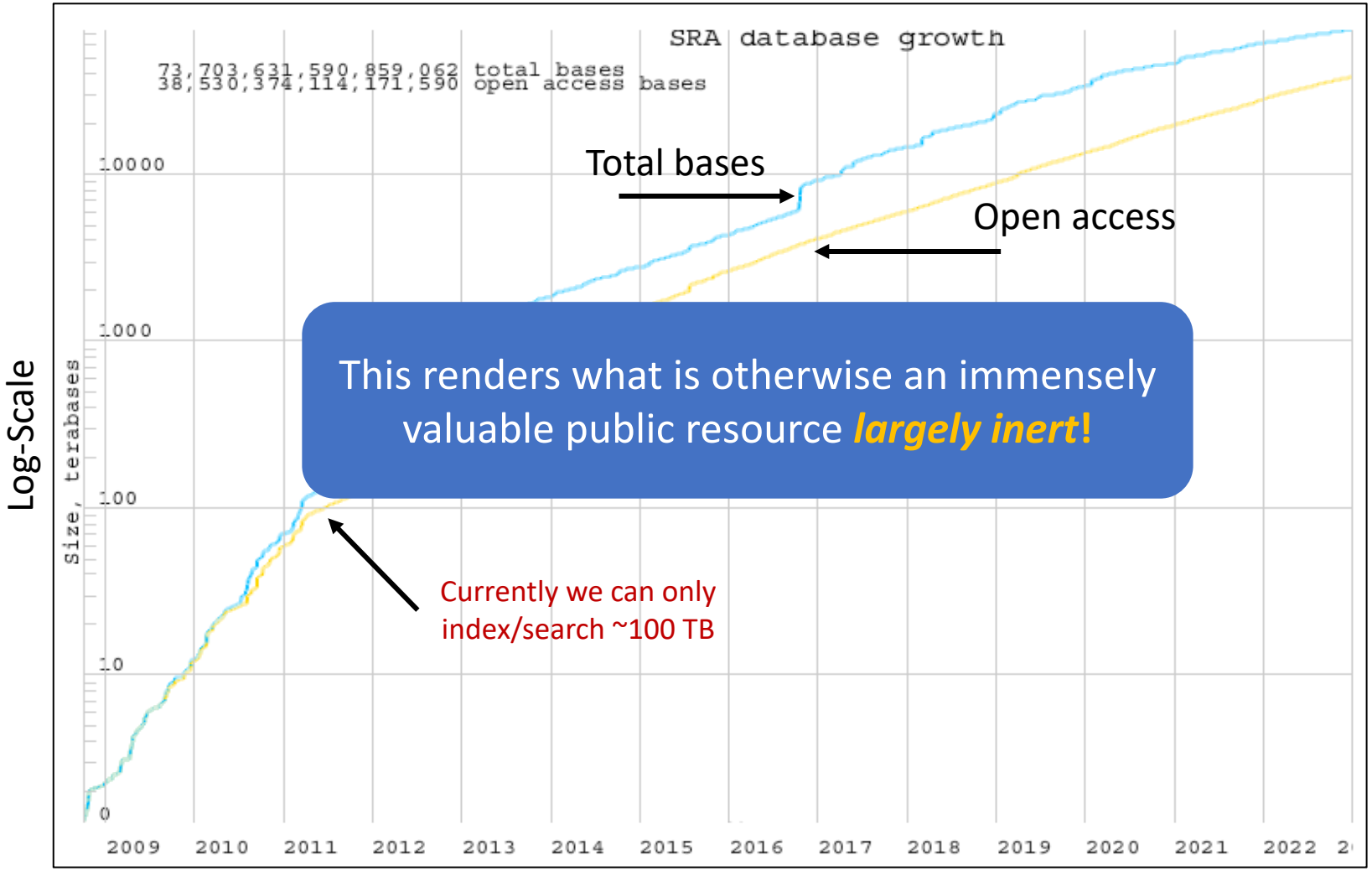
SRA contains a lot of *diversity information*



Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

Sequence Read Archive (SRA) growth

SRA contains a lot of *diversity information*



Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

Searches over SRA will enable complex biological analysis

- **What if I find a new putative disease-related transcript, and want to see if it appeared in other biological samples?**
- **What if I discover a new fusion event in a particular cancer subtype and want to know if it is common among samples with this subtype?**
- **What if I find an unexpected bacterial contaminant in my data; which other samples might contain this?**

⚒ A set of tools to index and search large-scale biological data

Our lab: enabling scalable data science

_K_-mer processing
Squeakr
[Bioinformatics 2017]

Weighted de Bruijn graphs
deBGR
[ISMB 2017, Bioinformatics 2017]

Colored de Bruijn graphs
Rainbowfish
[WABI 2017]

Sequence search over SRA

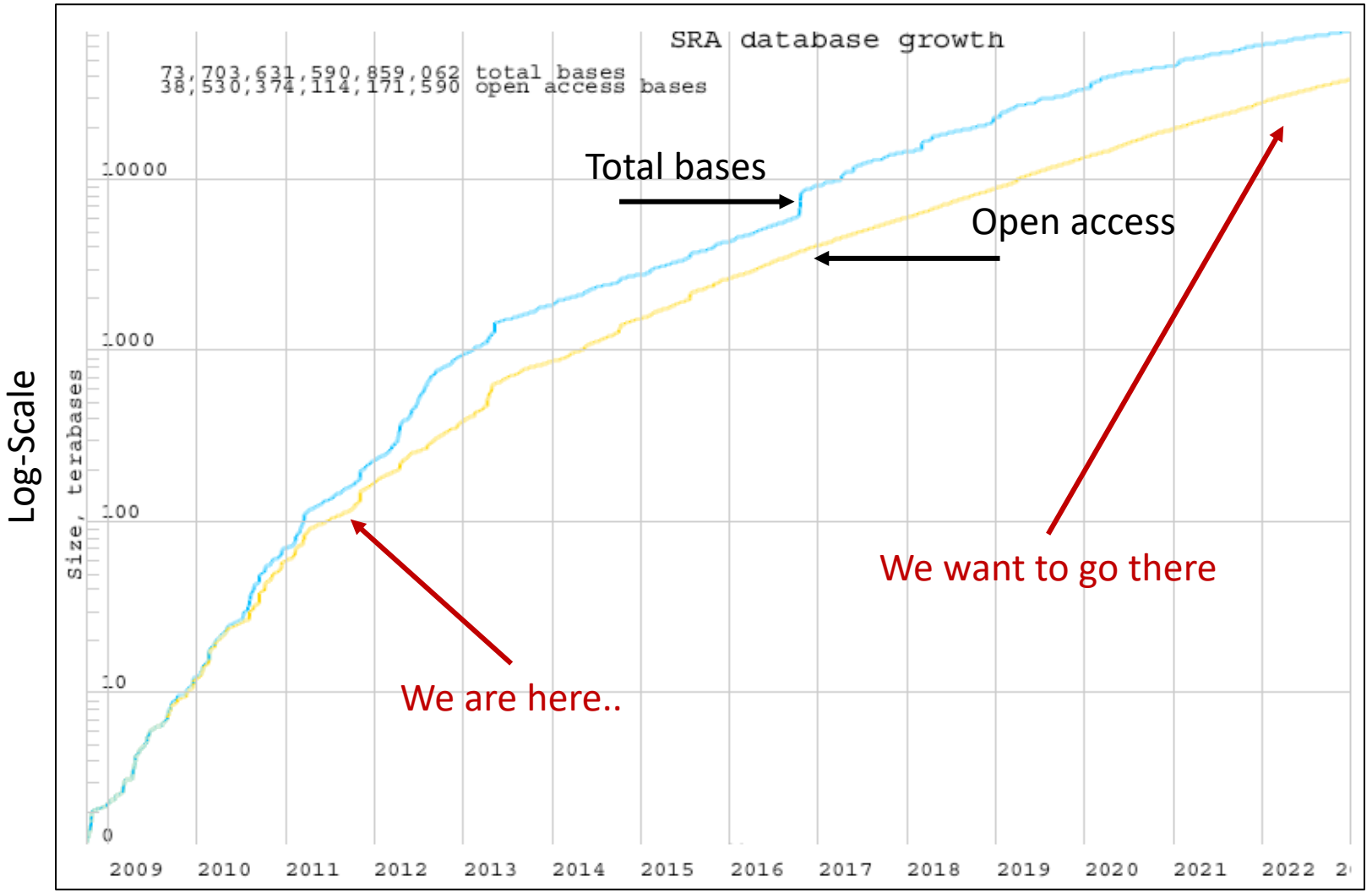
- Mantis [RECOMB 2018, Cell Systems 2018]
- Mantis-MST [RECOMB 2019, JCB 2020]
- Mantis-LSM [Bioinformatics 2022]

GPU-accelerated distributed _k_-mer processing
DEDUKT
[IPDPS 2021]

Pangenome index
VariantStore
[Genome Biology 2021]

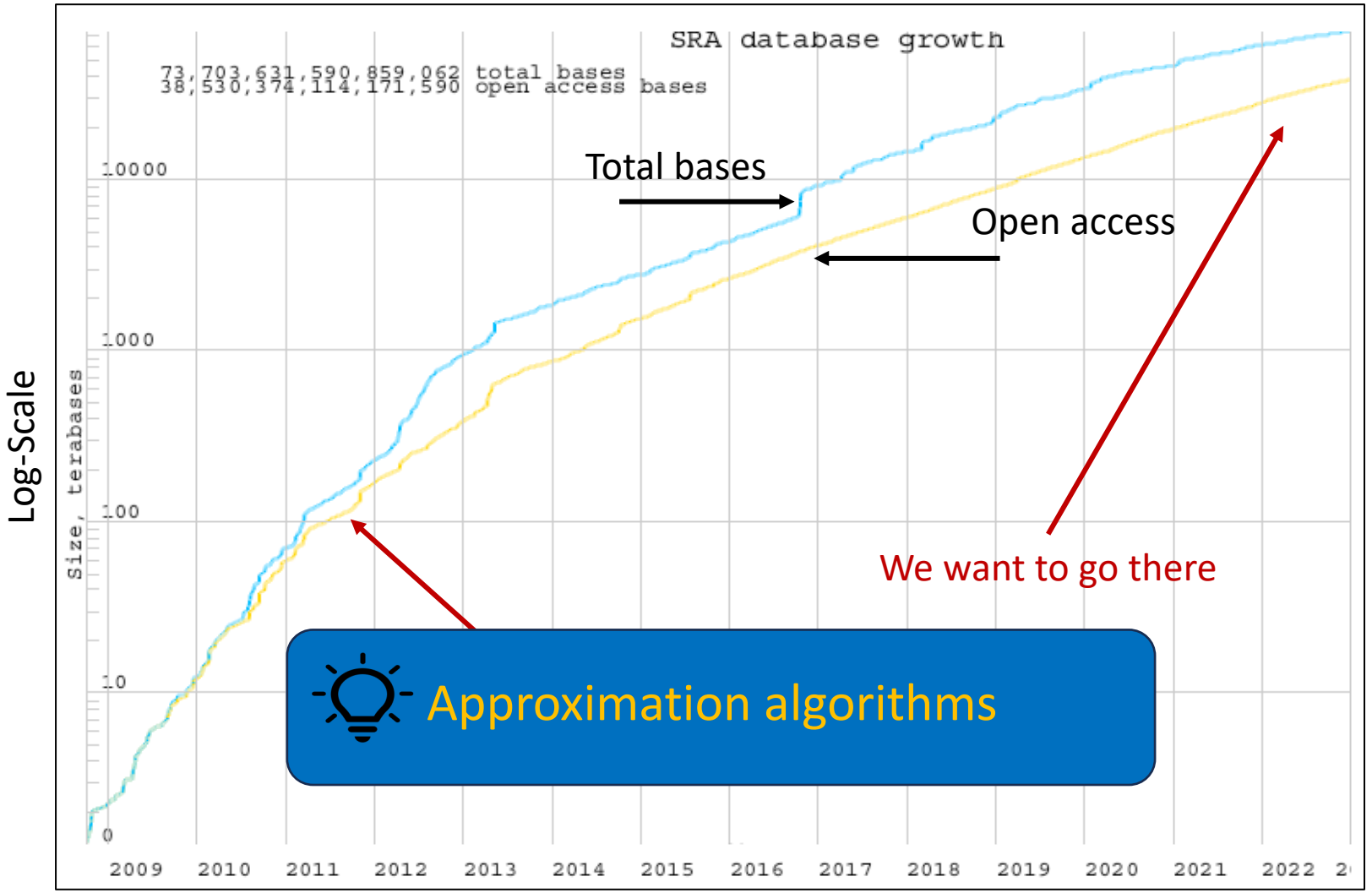
Exascale metagenomic assembly
MetaHipMer*
[PPoPP 2023, ACDA 2023]

What's next



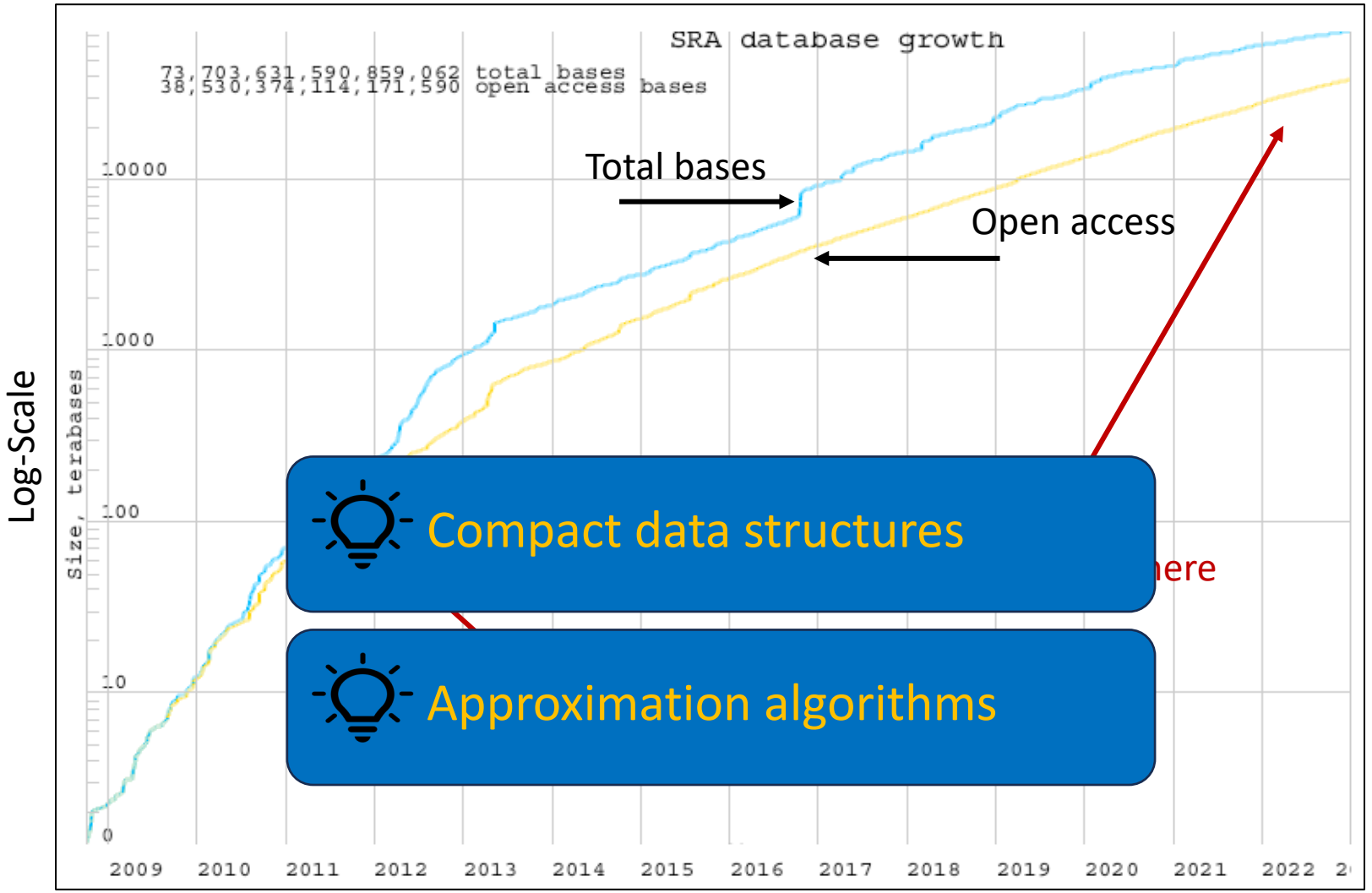
Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

What's next



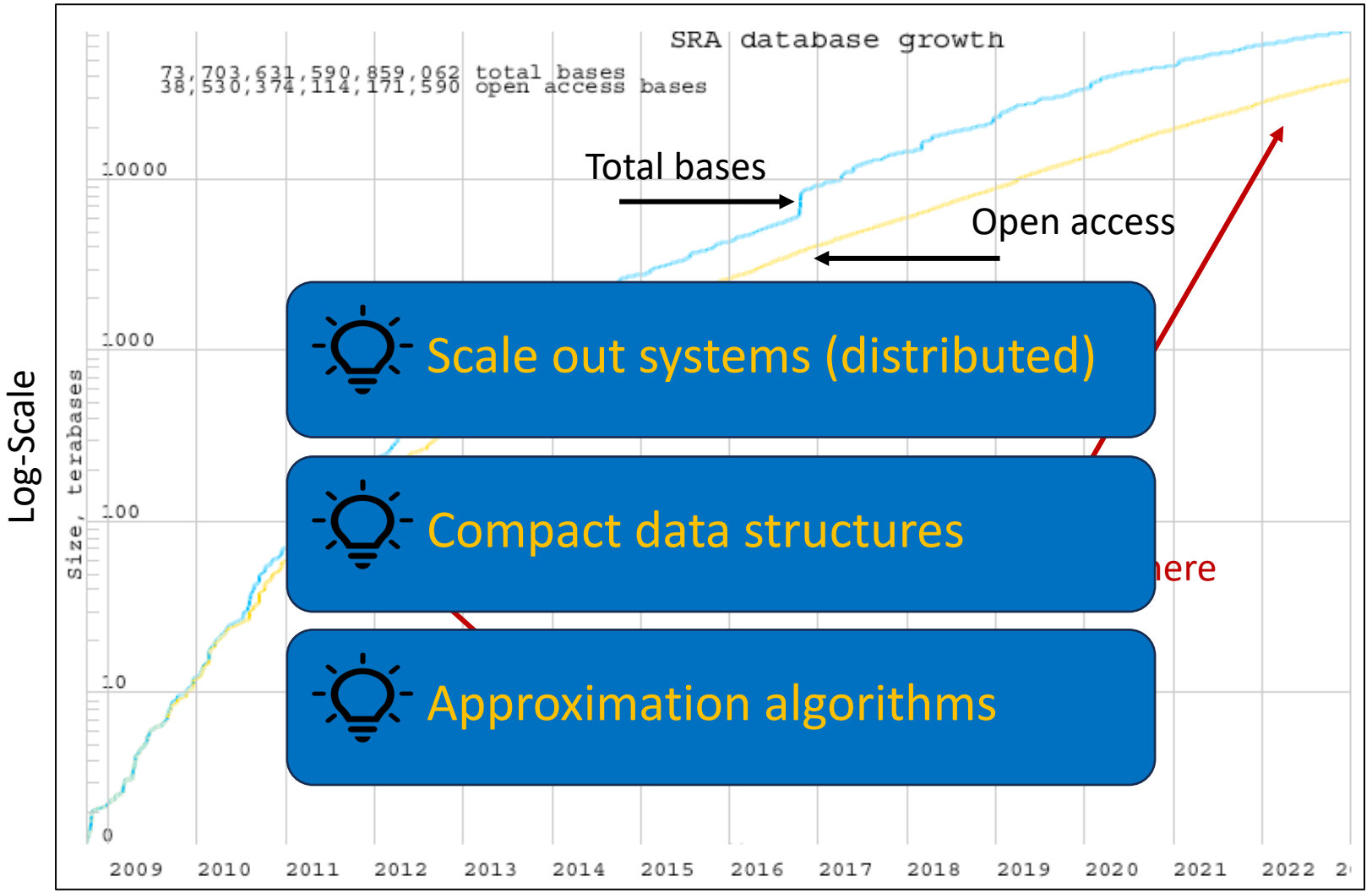
Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

What's next



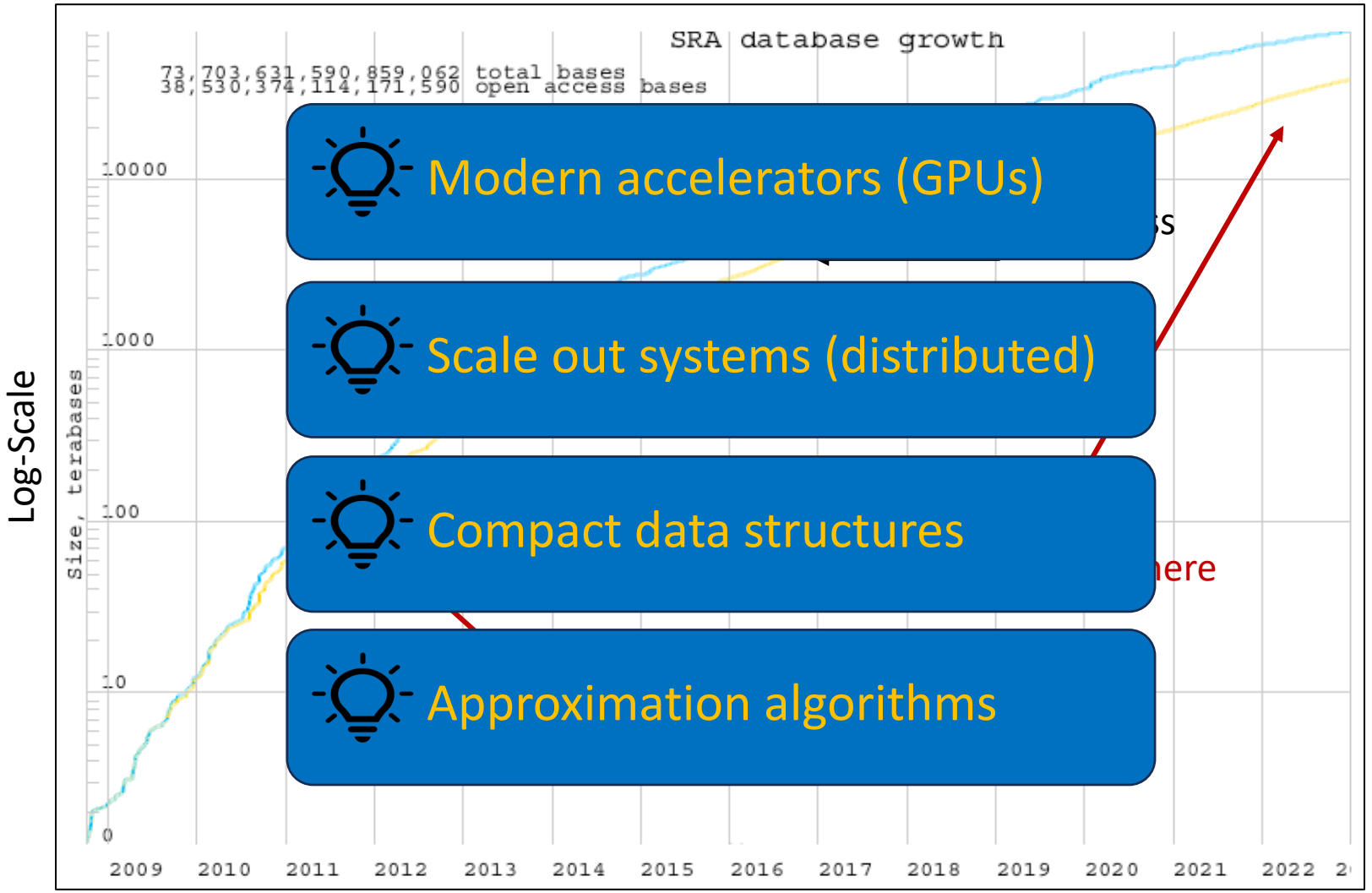
Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

What's next



Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

What's next



Data from: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi