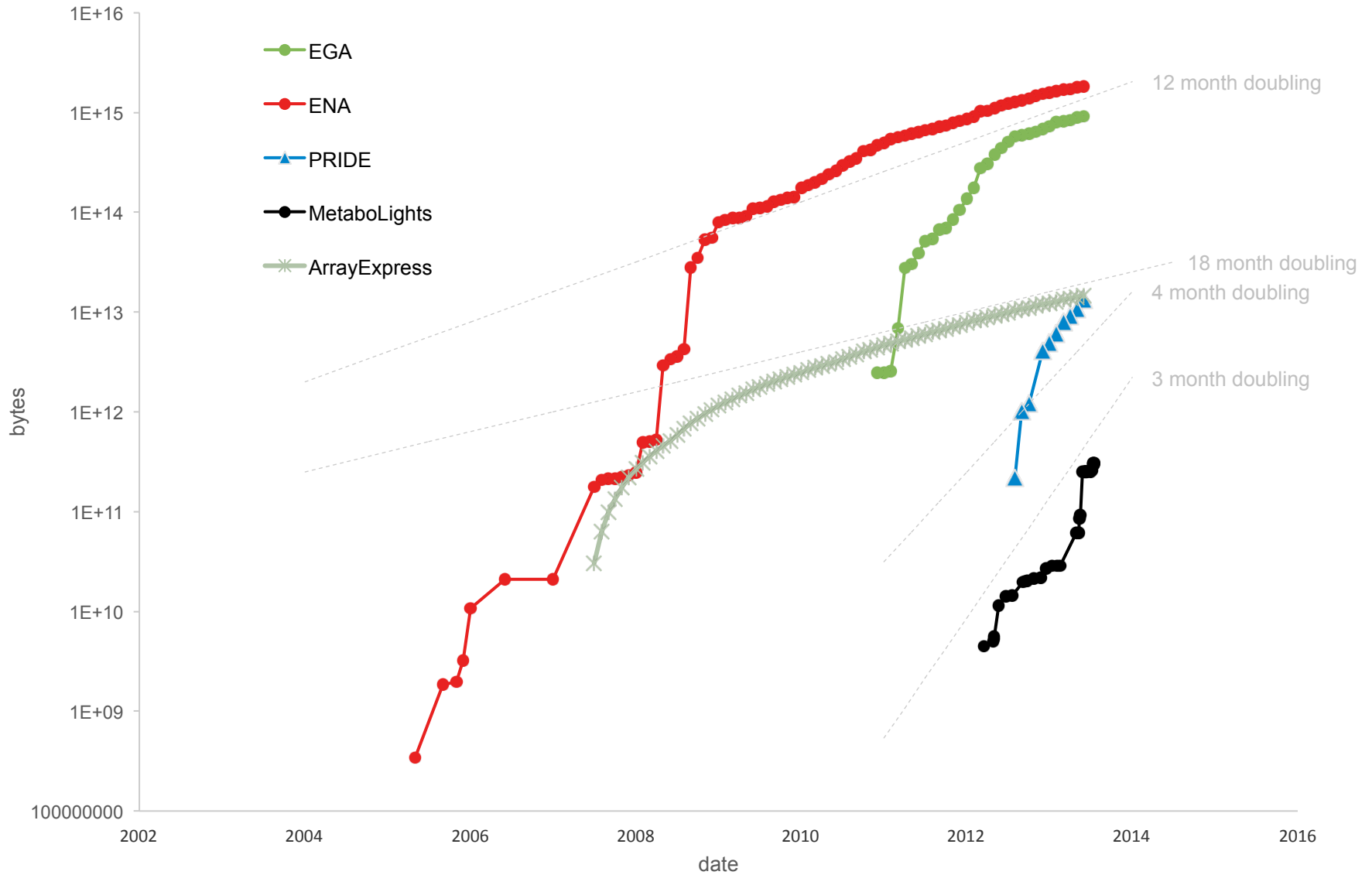
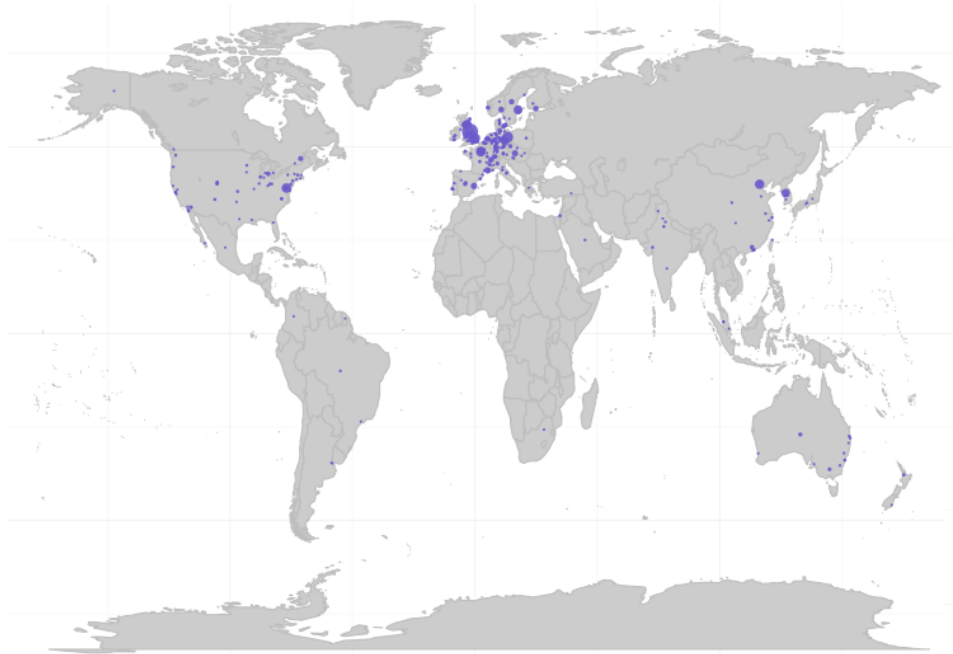


Landscape

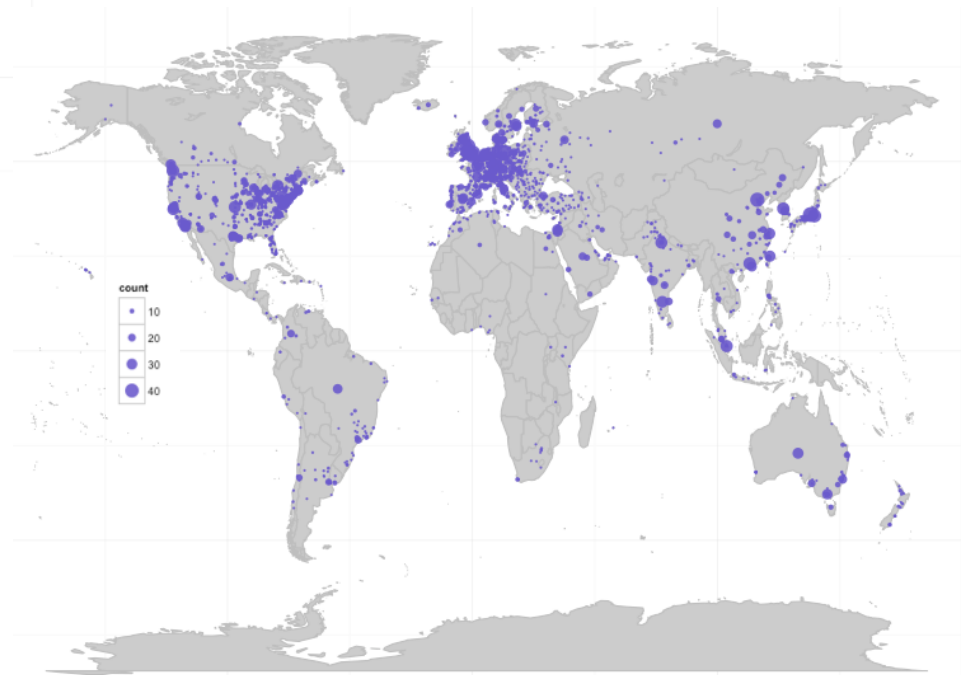
Growing data



Dispersed science



Data production

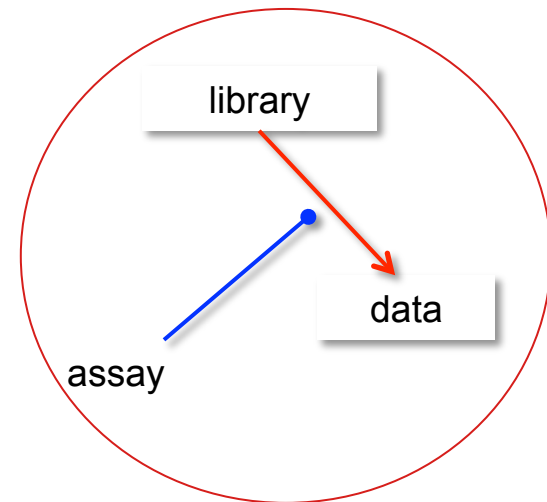


Data consumption

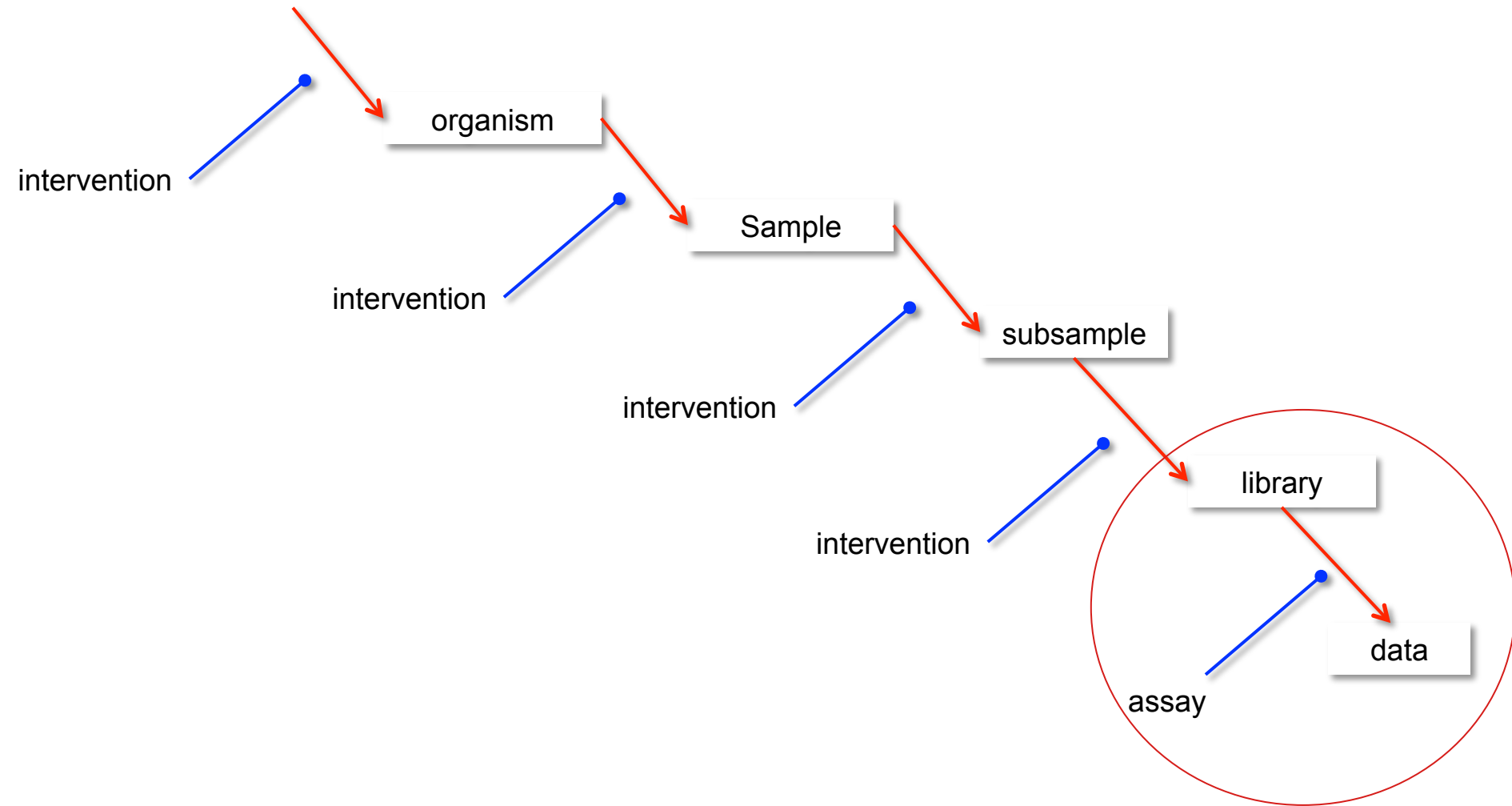
Why data?

- Direct requirements
 - Aggregation for processing and analysis
 - Statistical power
 - Method development/tuning
- Indirect requirements
 - Reuse/repurposing
 - Scrutiny
 - Scientific record

True cost of data generation



True cost of data generation

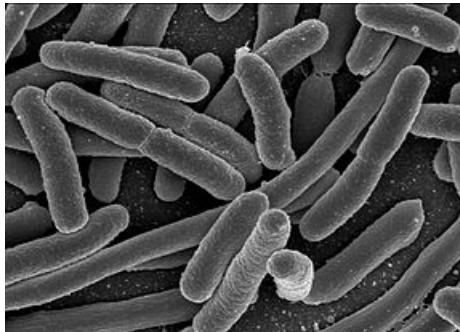


Sample cost



Western English Channel Observatory L4 site

- 10 year monthly samples amplicon sequenced
- Some shotgun metagenomic data
- Oceanographic data
- Remotely sensed data



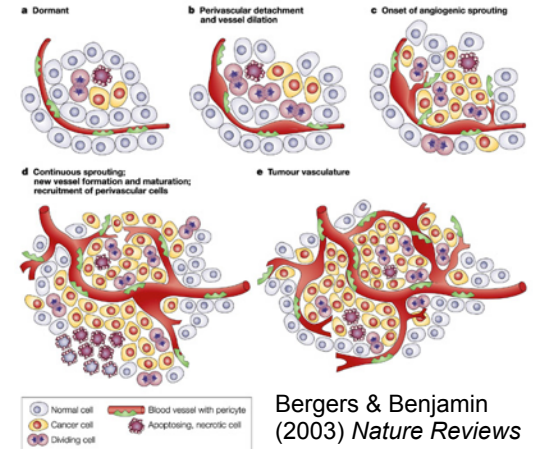
Rocky Mountain Laboratories, NIAID, NIH, http://en.wikipedia.org/wiki/File:EscherichiaColi_NIAID.jpg

Cultured stock

- Genetically stable
- Cheaply accessible

Cancer genomics

- Tumour/normal pairs
- Tumour sub-sampling
- Time series
- Medical history
- Ethical practice
- Regulatory/legal constraints



Data fluidity



Cloud computing



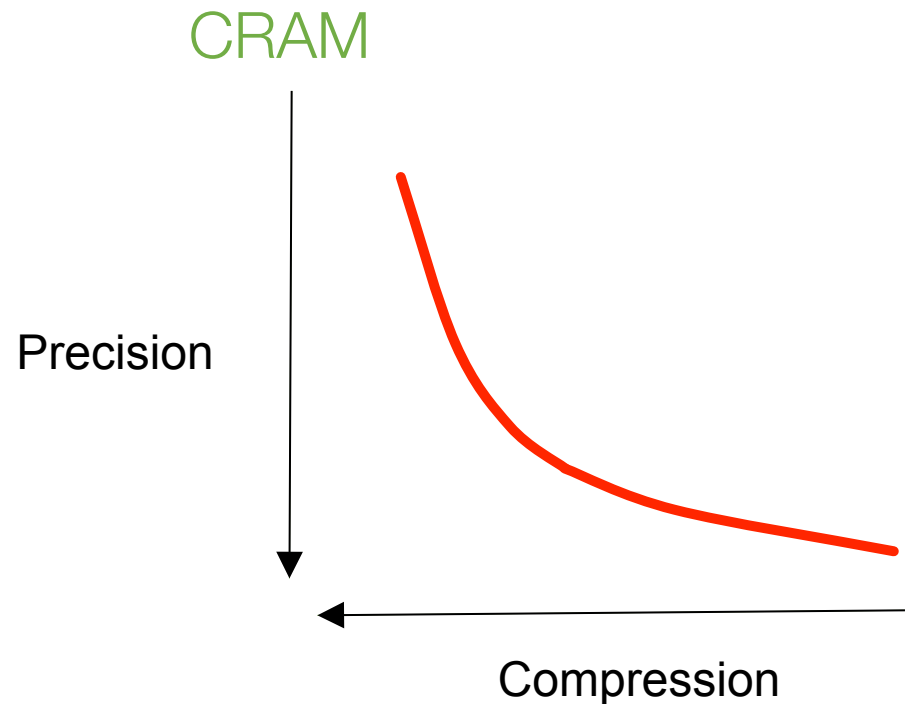
Data fluidity

Data compression

- Efficient representation
- Capacity for controlled data reduction
- Affordable transformations
- Tool chain

Data compression

- Efficient representation
- Capacity for controlled data reduction
- Affordable transformations
- Tool chain



•Fritz, M.H. Leinonen, R., et al. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40

•Cochrane G., Cook C.E. and Birney E. (2012) The future of DNA sequence archiving. *GigaScience* 2012, 1:2

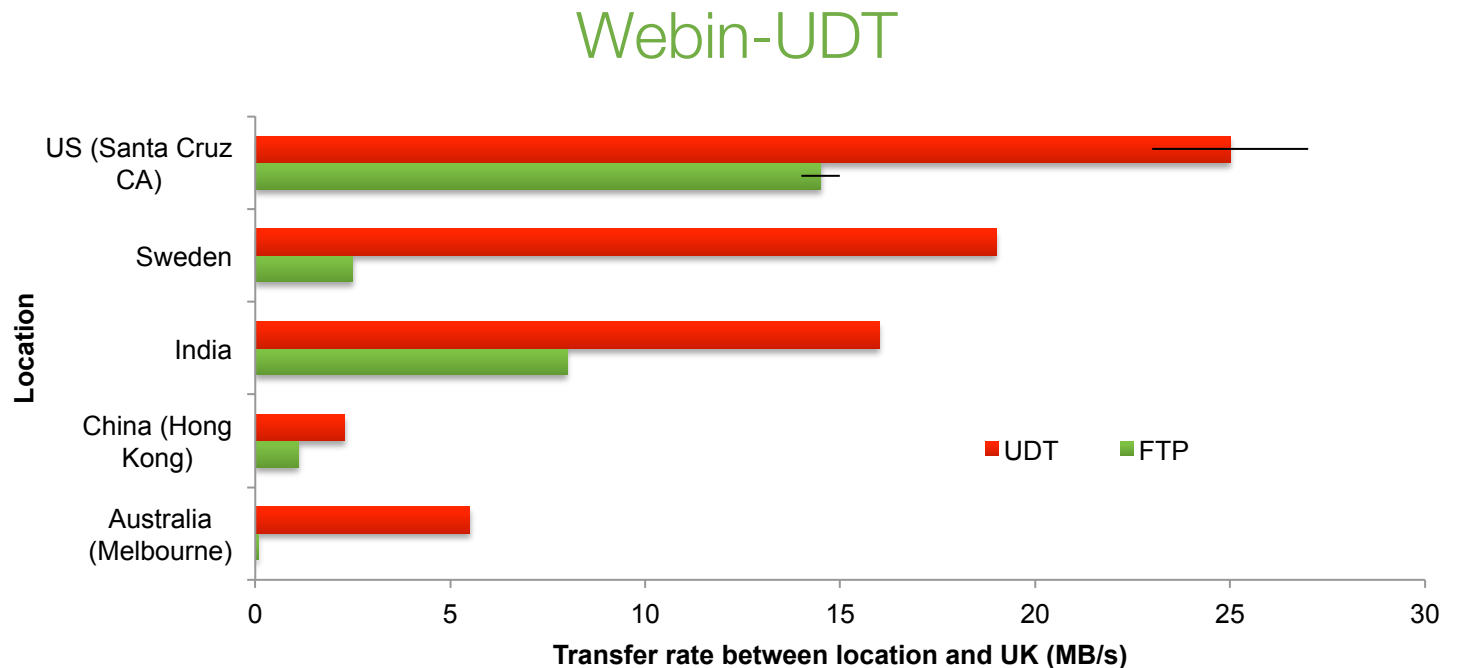
•http://www.ebi.ac.uk/ena/about/cram_toolkit

Network protocol

- Optimal use of available bandwidth
- Support and tool chain

Network protocol

- Optimal use of available bandwidth
- Support and tool chain



<https://github.com/enasequence/webin-data-streamer-UDT>

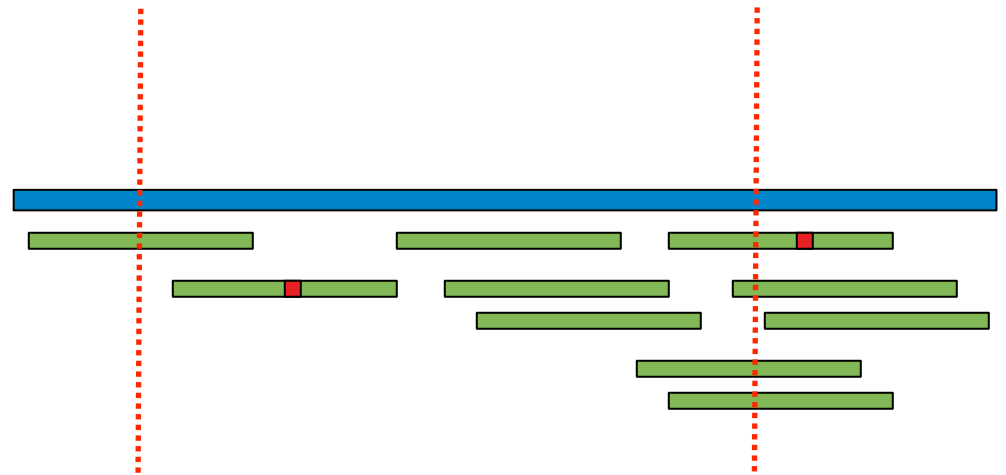
Data partitioning

- Organisation of data around biological concepts
- Indexing system around these concepts
- Support for requests for partitions along this index

Data partitioning

- Organisation of data around biological concepts
- Indexing system around these concepts
- Support for requests for partitions along this index

Reference-oriented indexing



Acknowledgements

Data growth statistics

EBI service teams

Reference-based compression

Markus Fritz, **James Bonfield**, Ewan Birney, Vadim Zalunin, Rasko Leinonen

Webin-UDT

Alexander Senf, Rasko Leinonen

