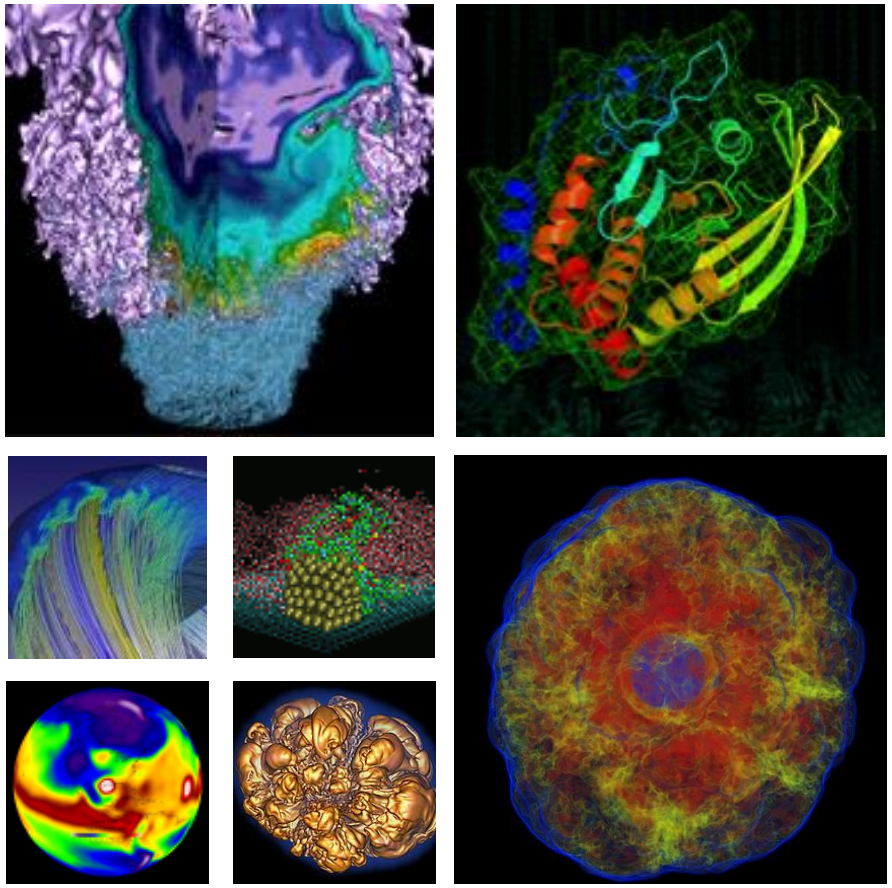
Big Data:
Genomics vs. Physics



Tony Wildish
HPC Consultant, NERSC/JGI

October 18, 2016

What is Big Data?

- **Volume, Velocity, Variety**
- **"any data that's too big to analyze on your desktop/laptop"**
- **Both Physics and Genomics have lots of data**
  - But how do they compare to each other?
  - (A: they're more different than they are alike)
  - Do the differences really matter?
  - (A: yes, a lot!)
  - And who has the biggest Big Data problem?
  - Compare the Large Hadron Collider with Genomics to find out

EMBL-EBI vs. the LHC

- **Large Hadron Collider**
  - 4 experiments: ALICE, ATLAS, CMS, LHCb
  - CMS alone has ~100 PB of data now
    - ~70 participating computing centers at labs/universities
    - Dedicated facilities, shared resources
    - Significant effort on managing, moving and processing data
- **EMBL-EBI**
  - Over 20 PB of data, growing fast
  - Several different data repositories
  - Serving several large communities
    - ~18 million web hits daily
- **Sounds similar enough, let's do a point-by-point comparison**

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
European Variation Archive
European Genome-phenome Archive
Ensembl
Ensembl Genomes
GWAS Catalog
Metagenomics portal

Gene, protein & metabolite expression

RNA Central
Express
Array Expression Atlas
Metabolights
PRIDE

Literature & ontologies

Europe PubMed Central
BioStudies
Gene Ontology
Experimental Factor Ontology

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

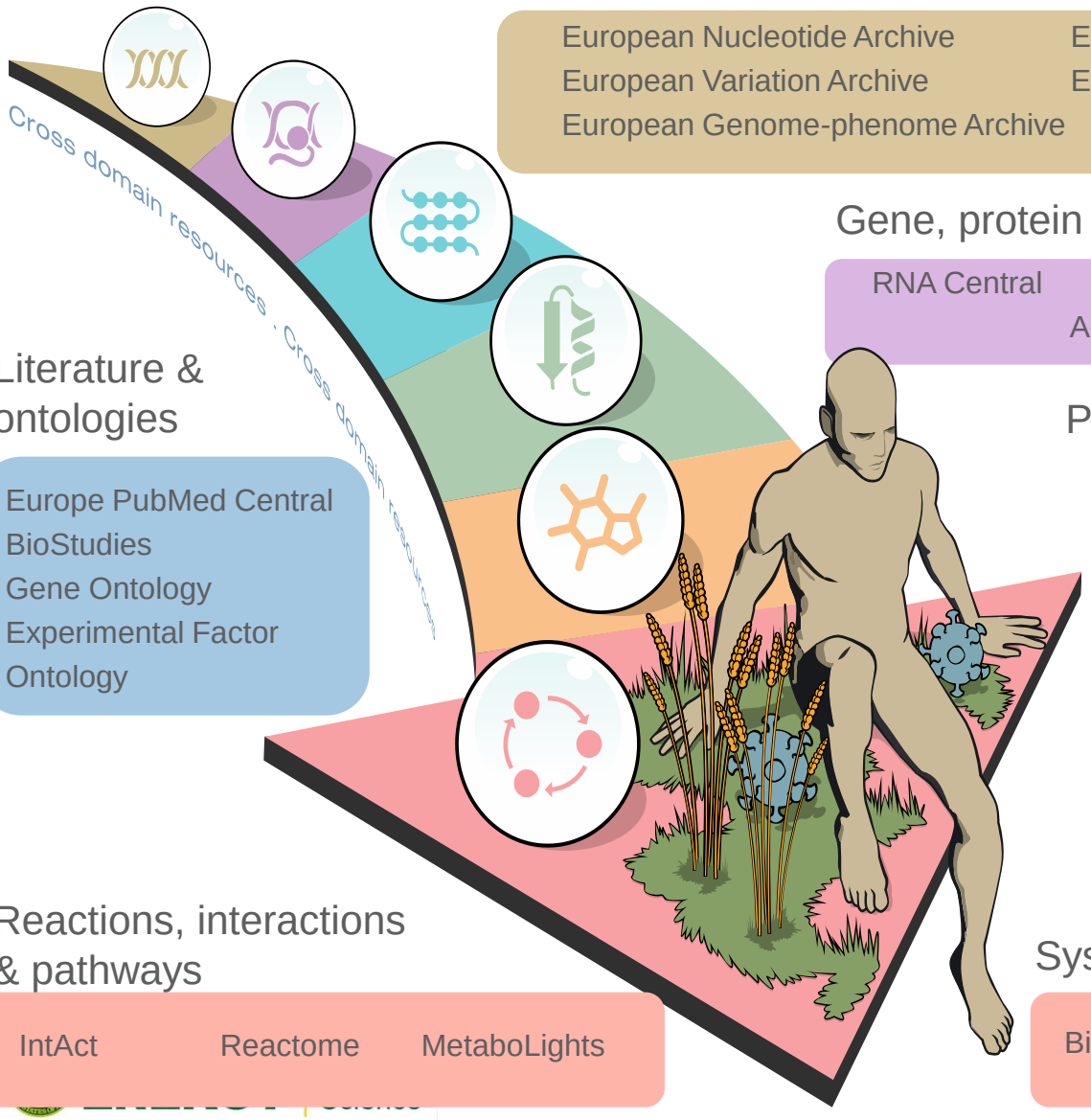
ChEMBL
SureChEMBL
ChEBI

Reactions, interactions & pathways

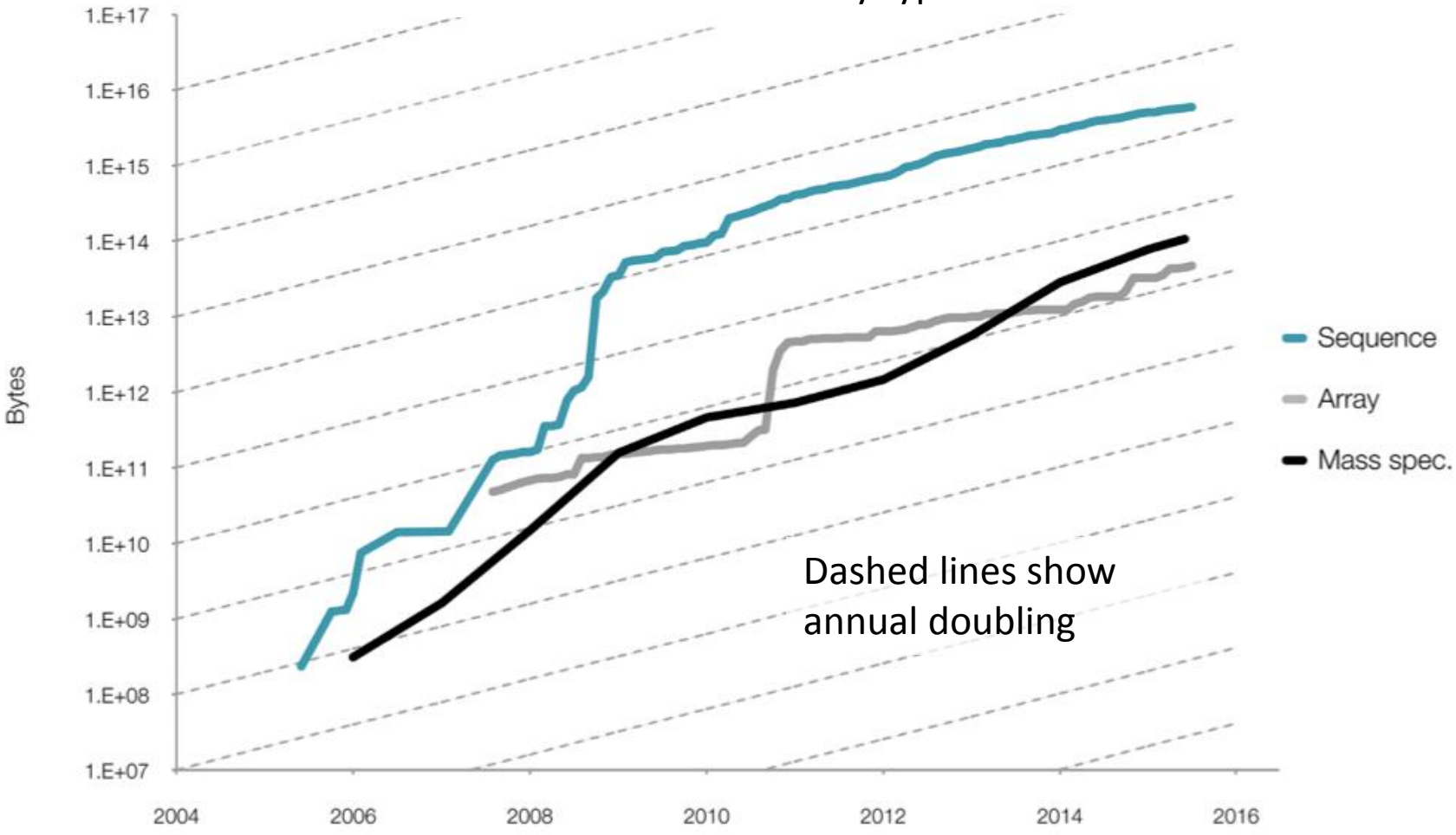
IntAct
Reactome
MetaboLights

Systems

BioModels
Enzyme Portal
BioSamples



EBI Data Accumulation by type



Volume, Velocity

- **Physics:**
  - 10-100 PB
- **Growth rate:**
  - ~10's PB per year
  - Essentially linear
- **Known years in advance**
  - fixed by accelerator & experiment design
  - Can't build the experiment without some idea what the data looks like

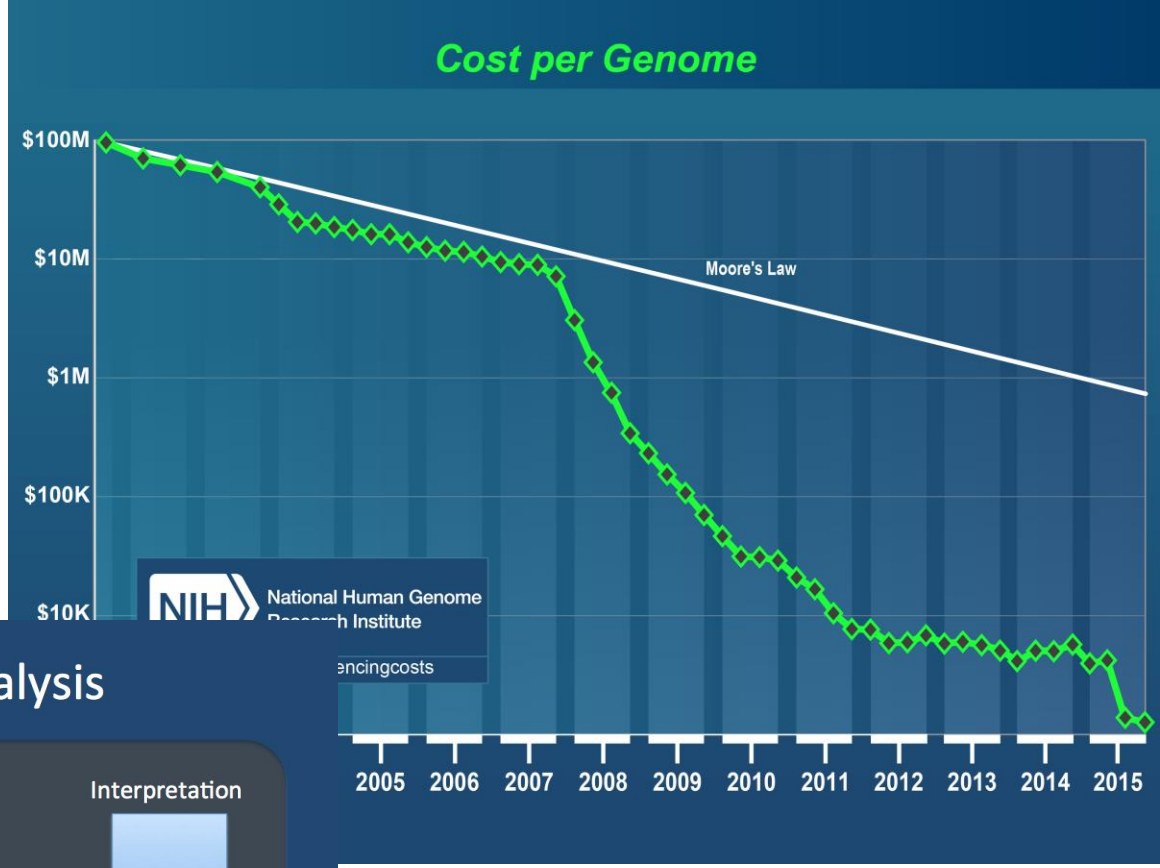
- **Genomics:**
  - 10-100 PB
- **Growth rate:**
  - Doubling every 12-18 months
- **Unpredictable future**
  - Cheaper, faster sequencing
  - New sequencing methods
  - Lower bound: scarily fast

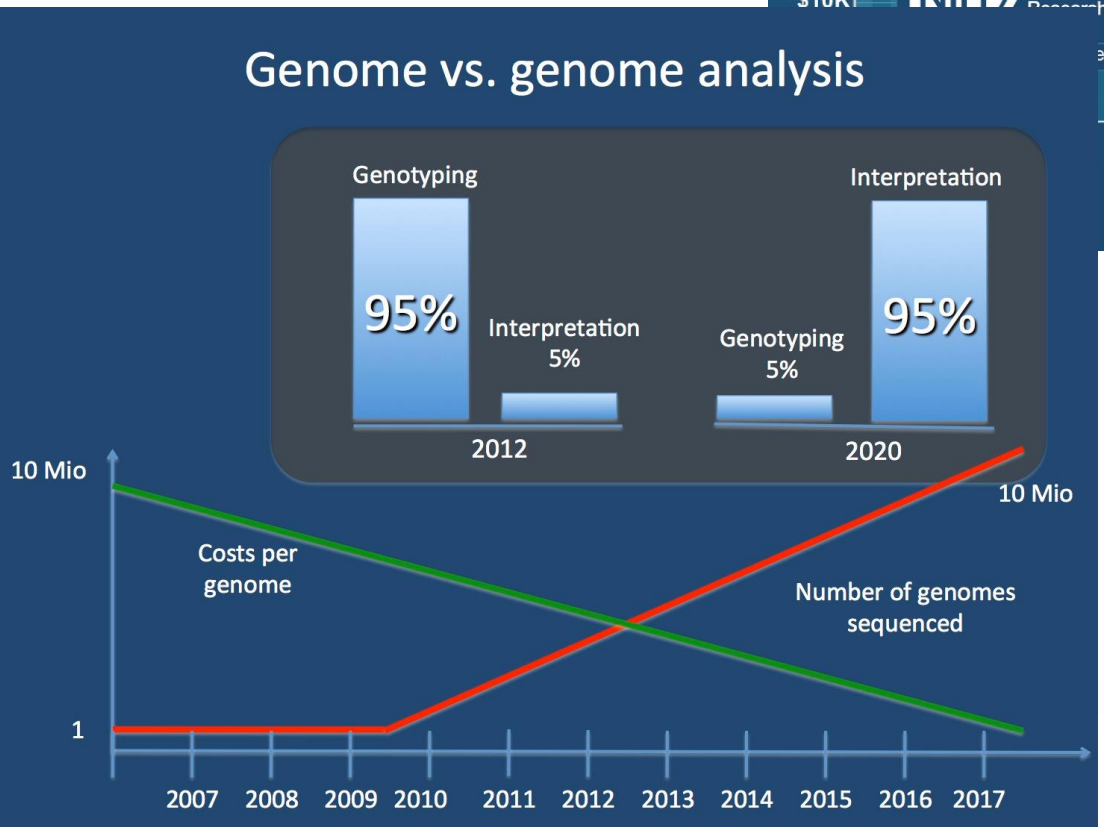
Velocity, Variety

- **Processing latency**
  - Urgent need to verify quality while running
  - Produce physics results fast, guide future operations
  - ‘Discovery’ machines
- **Not always as urgent**
  - Metagenomics
  - Population studies
    - ‘100,000 genomes project’
- **Variety: limited**
  - Data evolves little over life of experiment
  - Few data types, few formats
- **Variety: considerable**
  - Sequencing methods
    - WGS, scRNA…
  - Different data types
  - Different data formats

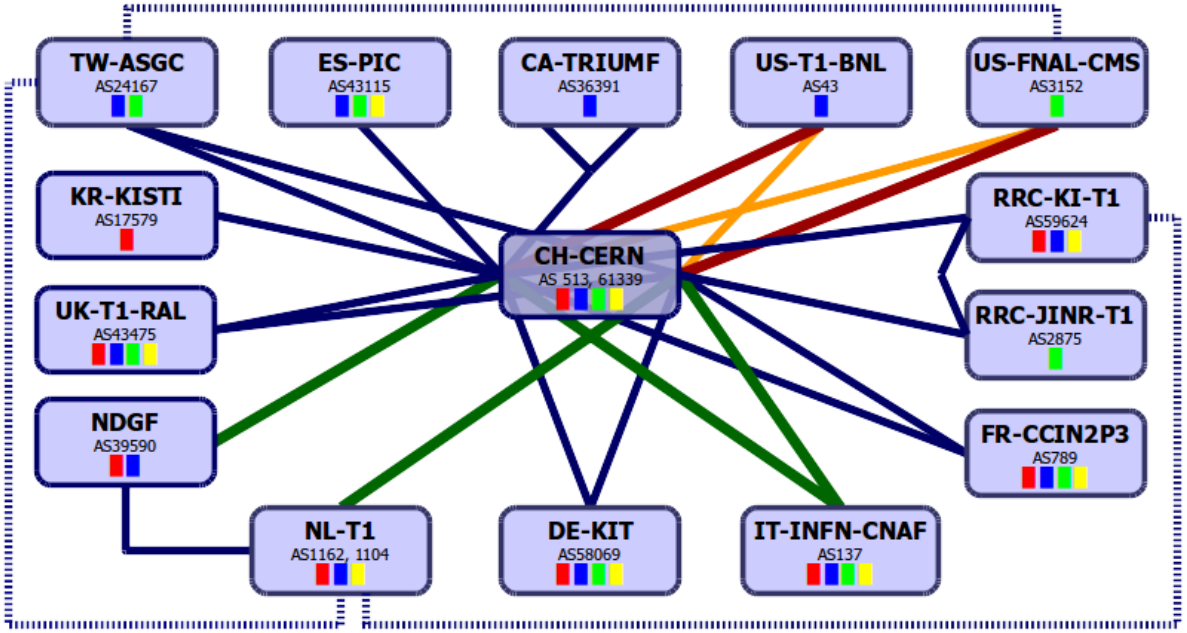
- **Cost: very expensive**
- **Production:**
  - Single-source
  - Tightly co-ordinated
  - Real-time quality control
- **Distribution:**
  - Dedicated networks
  - Experiment-controlled
  - Scheduled, managed

- **Cost: falling**
- **Production:**
  - Multiple sources
  - Increasing in number
  - Not all so well controlled
- **Distribution:**
  - Upload from anywhere
  - Distribute at-will
  - On-demand, reactive

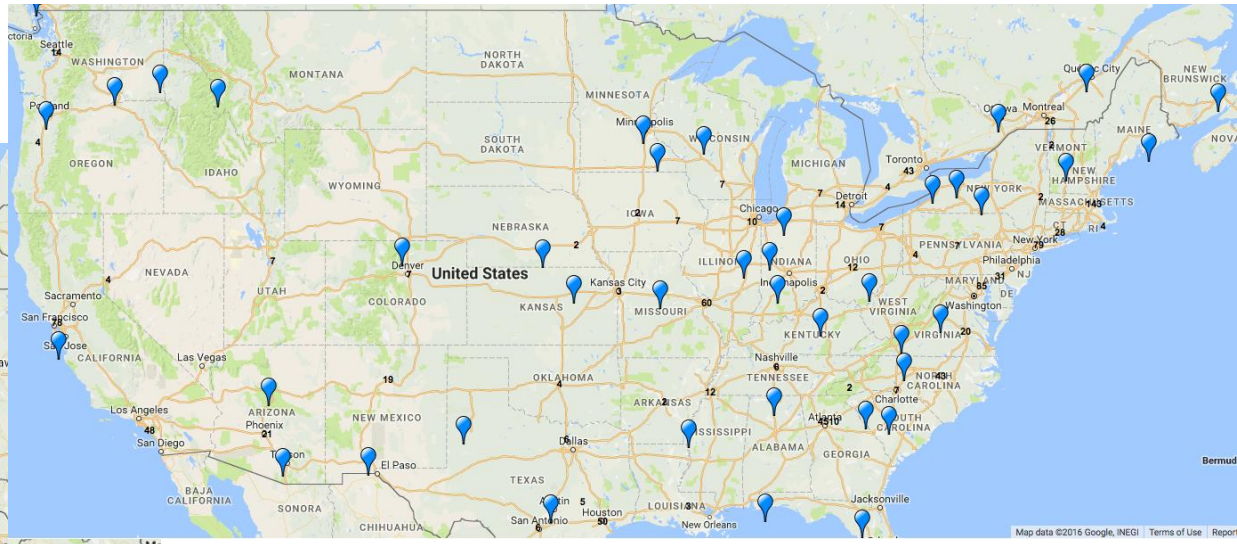


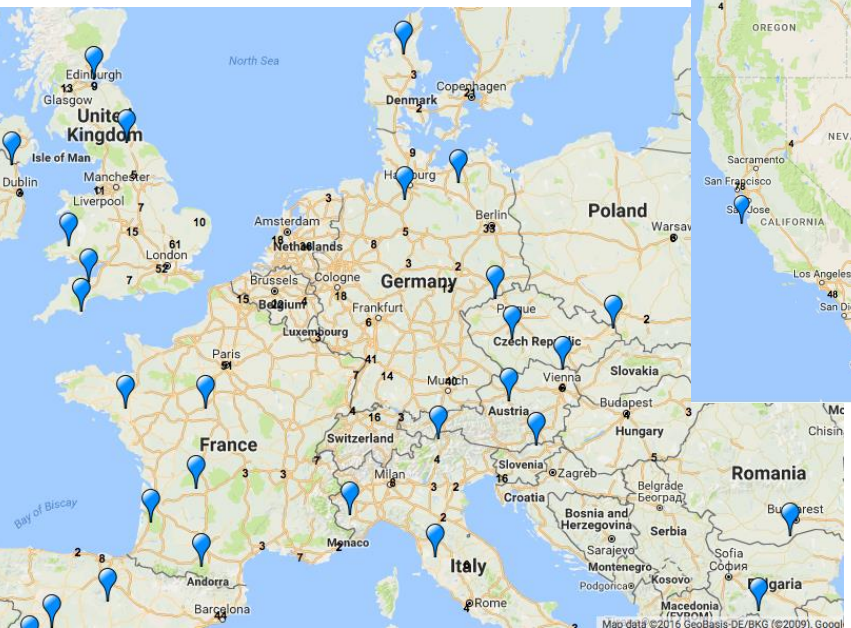


BERKELEY LAB



LHCOPN: dedicated data-distribution infrastructure for LHC experiments





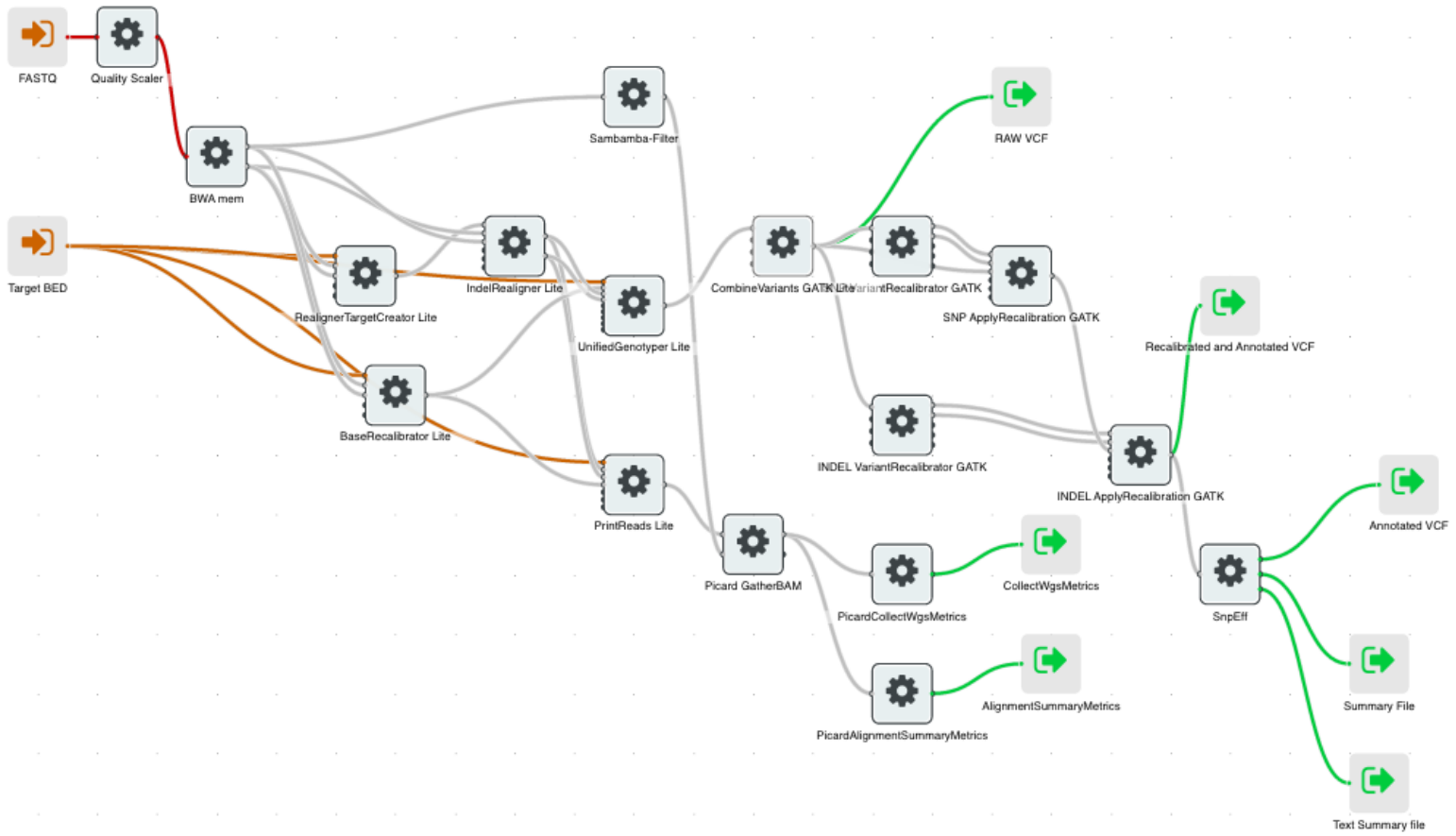
Omicsmaps.org: location of DNA sequencers, any/all of which can upload to EBI, NCBI, DDBJ

BERKELEY LAB

Data-pipelines

- **Physics: standard pattern**
  - Apply calibration and alignment
  - Reconstruct physics objects
  - Reduce data volume
  - Simple, few steps
- **Genomics: similar steps**
  - Filtering (read quality, short reads, contaminants, repeats)
  - Alignment, assembly
  - May not reduce data volume
    - Intermediate results may themselves be valuable
  - Larger numbers of steps and tools involved

Whole Genome Analysis pipeline (GATK)



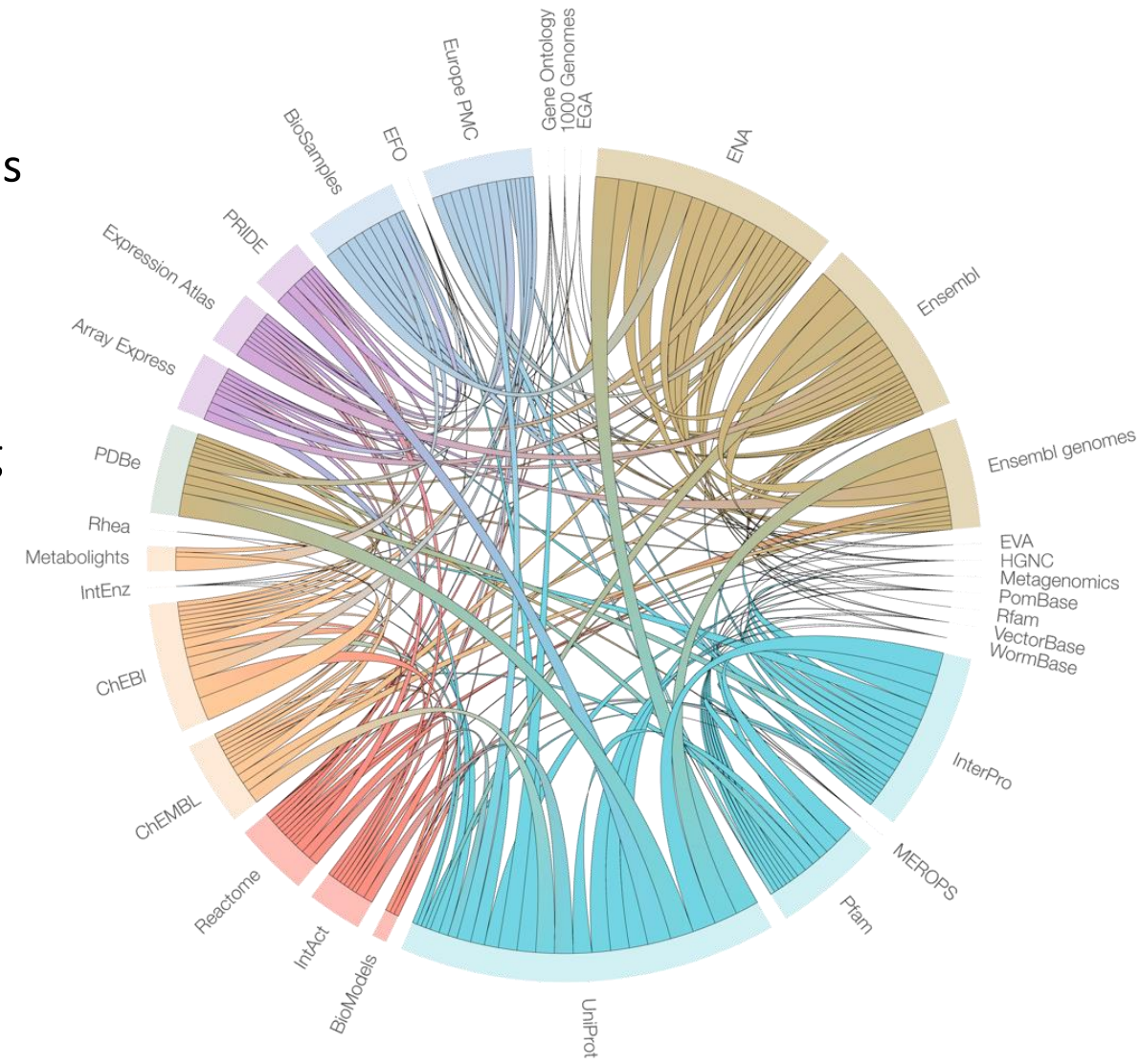
Data-pipelines

- **Dataset structure:**
  - Defined by data-taking conditions: beam, trigger…
  - Stable, consistent, reproducible
- **Metadata**
  - Calibration, alignment, beam conditions
  - Produced alongside raw data
  - ‘first-class’ data in its own right

- **Dataset structure:**
  - No clear analog
  - Study, experiment, project
    - Protein, gene, population, organism, cell-type, disease
  - Data associated with a particular published paper
    - May come from other sources
- **Metadata**
  - Data related across projects or exp’ts (e.g. TARA, OSD)
  - Contextual data may be poor quality
  - Annotation, manual curation: no analog in HEP

Database interactions

- Internal interactions between data resources as determined by the exchange of data.
- Width of each internal arc weighted according to the number of different data types exchanged.



Pipelines and data value

- **Raw data is inconvenient**
  - Event-rate from CMS detector is 40 MHz
  - Online trigger: reject most data forever, immediately
  - Keep only what passes trigger: 1 kHz
  - Avoid re-reading raw data after initial reconstruction
  - Only reconstructed data used for analysis
- **Future value of today's data not known**
  - Delete raw reads and keep only assembled genome?
    - Assembly algorithms constantly improving
  - Delete everything and keep only the biological sample?
  - Reproducibility of statistical properties vs. of published analyses vs. ???

Analysis

- **Communities:**
  - ~3K scientists per exp't
  - ~12K CERN users
- **Data structure:**
  - Discrete events
  - Analyze independently

- **Communities:**
  - ~100K life-scientists in Europe alone
- **Data structure:**
  - Multiple overlapping reads
  - Hard to control locality and depth of coverage
  - Sample contamination may be an issue

Analysis

- **Reconstruction, analysis**
  - Always event-by-event
  - Modest resource needs
  - Embarassingly parallel
- **Analysis requirements**
  - Per event, decreases over time
  - ‘AOD’ well-tuned formats
  - Can predict analysis requirements accurately
- **More complex analyses**
  - De-novo assembly, multiple alignments
  - Metagenomics
  - Genome re-arrangements
  - Population studies
  - No analogy in HEP
- **Analysis requirements**
  - Often big, and increasing...
  - Hard to predict runtimes
    - “no free lunch”: runtime depends on the particular data you have

Analysis

- **Can ask types of questions that don't exist in physics**
  - More data, more types of data, more accurate data...
  - ...more complex questions
  - Big-data paradigm shift:
  - Hypothesis driven -> data-driven
  - Hard to predict computing needs for questions you haven't asked yet!

Software, tools

- **Strong computing community within HEP**
  - Professional s/w engineers
- **Domain-wide effort**
  - Shift from FORTRAN to C++ took ~10 years
- **Relatively few tools**
  - Shared simulation & analysis libraries
  - Highly tuned, scalable
- **Development not so consolidated**
- **Larger user community**
  - Harder to re-educate
- **Many tools**
  - Different authors, different priorities, orphan code
    - JGI: >50 assemblers
  - Many will not scale well

Software, tools

- **Detector technology changes slowly**
  - Iterative process, not disruptive
  - Need s/w development to drive technology design
  - More lead time for development for analysis
- **New technology can be very disruptive, very fast**
  - NGS created a need for new assembly algorithms
  - Long-read sequencing has higher error rate
    - Nanopore: much current effort to get good data
  - Takes time for computing to catch up with data
    - Older tools can become entrenched through familiarity

Data structures, access

- **Data structures well tuned**
  - Object-oriented -> data-oriented
  - Parallelism, distributability
  - Major efforts on code efficiency, ongoing
- **Flexible data access**
  - POSIX -> socket, streaming
  - Read over LAN with fallback to WAN
- **Profusion of data formats**
  - Lots of good work going on
  - CRAM, lossy compression, split-file formats
  - Analysis often harder to optimise for data access
    - E.g. often need random access to reference genome

Infrastructure

- **Physics community accustomed to owning infrastructure**
  - Each experiment tunes their own infrastructure
  - Infrastructure, software, and physicists co-evolve
  - Dedicated networking for relatively small number of sites
- **Infrastructure often provided as a platform**
  - Typically web-based or batch farm
  - Expertise unevenly distributed
  - Harder to move community to a new platform
  - Hard to provide dedicated network for such a distributed community

Summary

- **Physics and genomics both have Big Data**
  - But in almost every respect, they differ greatly from each other
- **Genomics data:**
  - More complex and variable, used in more demanding ways
  - Growth is accelerating faster than physics data
  - Greater uncertainty on short timescales => less time to respond
  - Less community-wide investment in s/w and infrastructure
- **Conclusion:**
  - Life-scientists have a much harder problem than physicists
  - …so they should have a lot more fun ☺

National Energy Research Scientific Computing Center