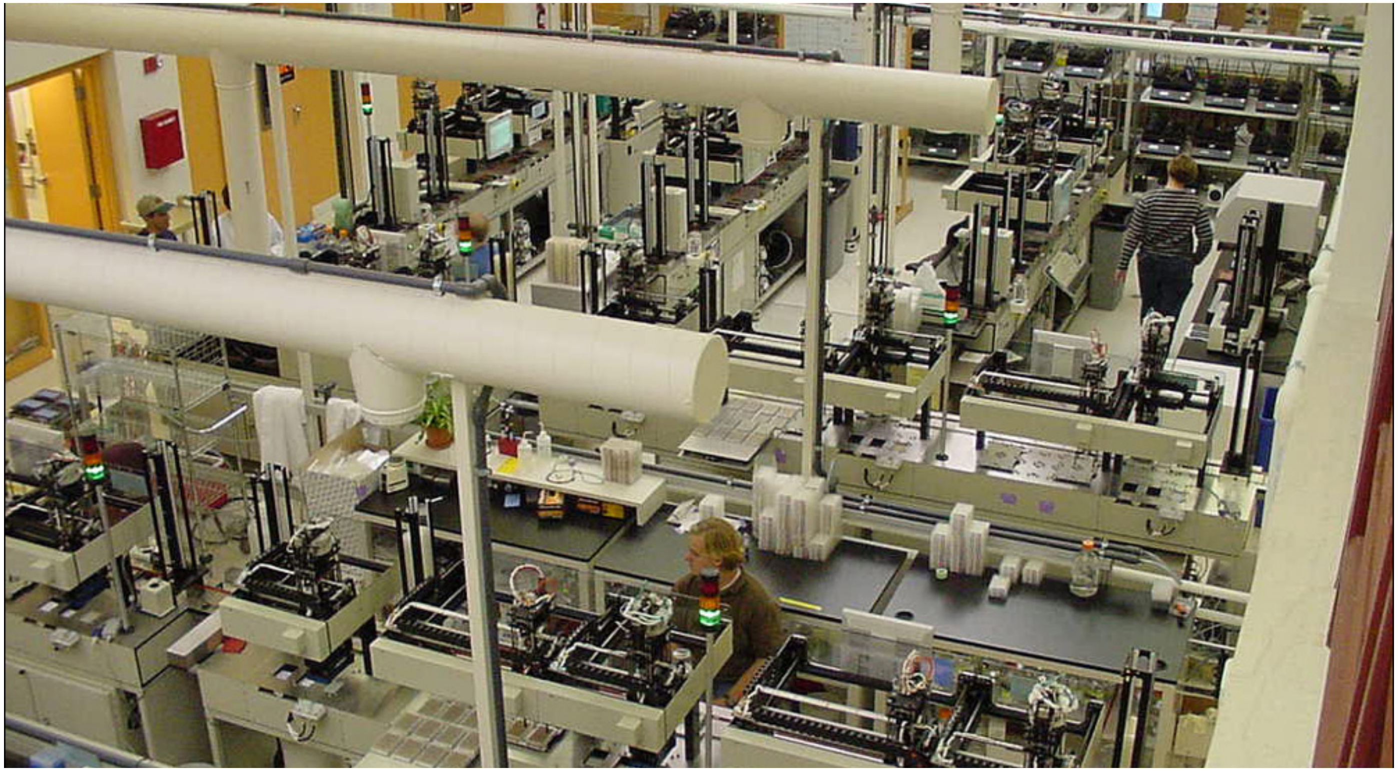
Data Commons for the Genomics Community

Allison Heath

Center for Data Intensive Science
University of Chicago

June 8, 2015

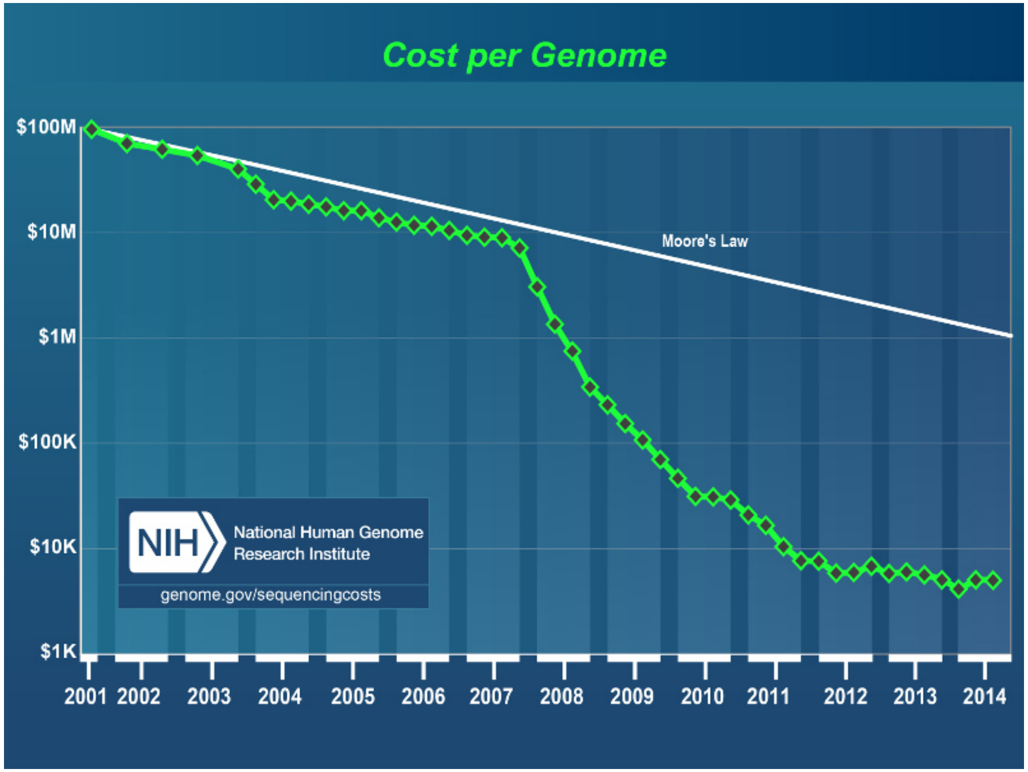
Explosion of Genomics Data

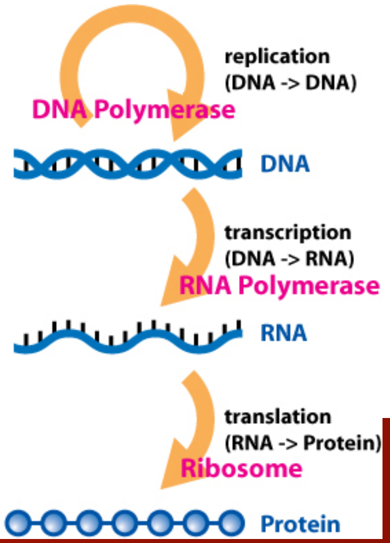


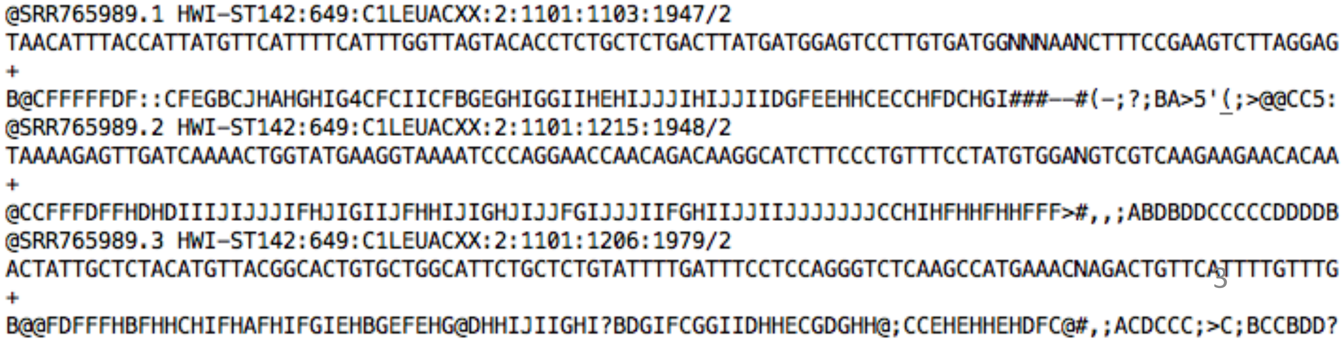
Sequencers at the Broad Institute of MIT and Harvard.

Current Sequencing Capabilities

- DNA-Seq, RNA-Seq
- Large scale studies of genomic variation
- Sequencing data is sensor data





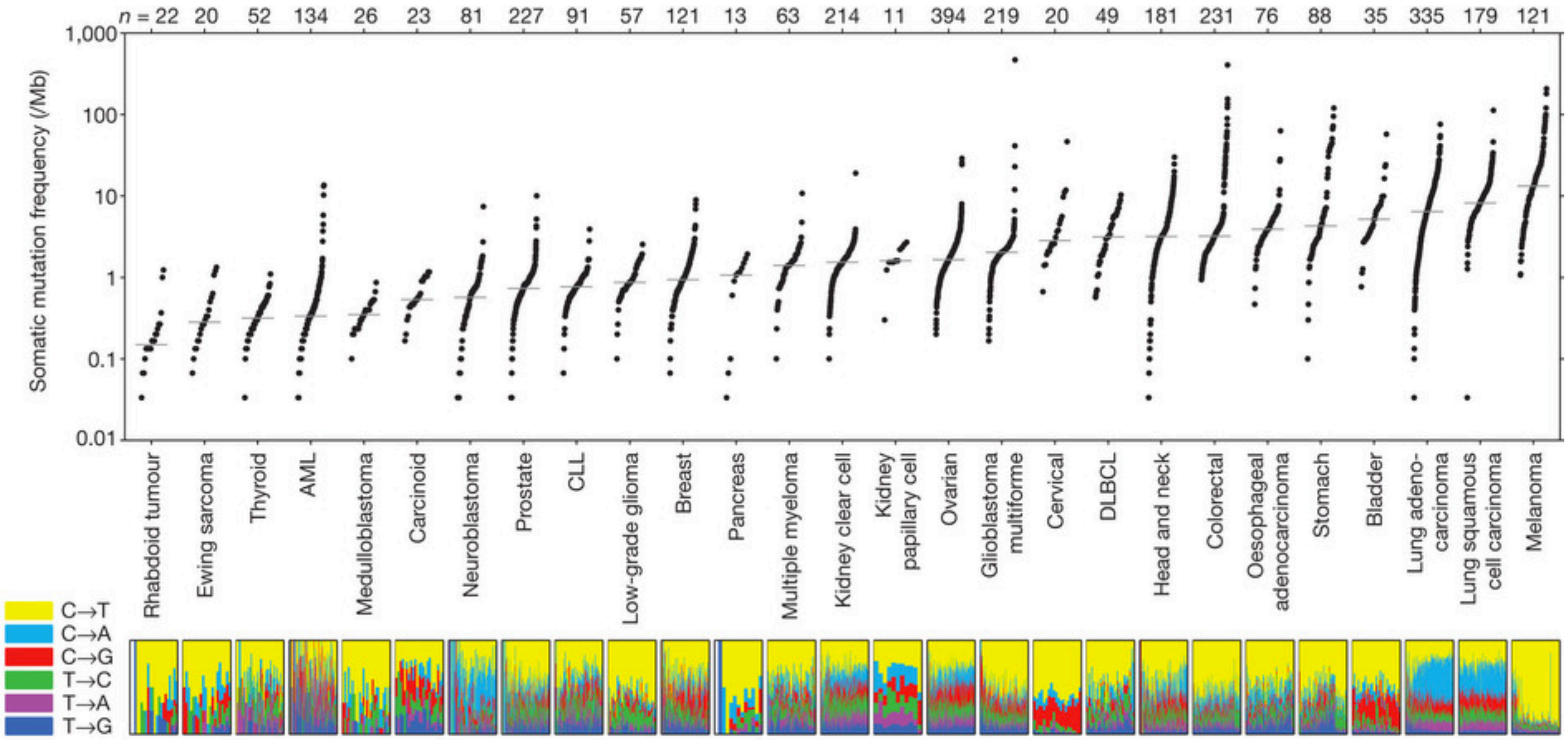


```
@SRR765989.1 HWI-ST142:649:C1LEUACXX:2:1101:1103:1947/2
TAACATTTACCATTATGTTCATTTTCATTTGGTTAGTACACCTCTGCTCTGACTTATGATGGAGTCCTTGTGATGGNNNAANCTTTCCGAAGTCTTAGGAG
+
B@CFFFFFDF::CFEGBCJHAHGHIG4CFCIICFBGEGHIGGIIHEHIJJJIHIJJIIDGFEEHHCECCHFDCHGI###--#(-;?;BA>5'(;>@@CC5:
@SRR765989.2 HWI-ST142:649:C1LEUACXX:2:1101:1215:1948/2
TAAAAGAGTTGATCAAAACTGGTATGAAGGTAAAATCCCAGGAACCAACAGACAAGGCATCTTCCCTGTTTCCTATGTGGANGTCGTCAAGAAGAACACAA
+
@CCFFFDFFHDHDIIIJIJJJIFHJIGIIJFHHIJIGHJIJJFGIJJJIIFGHIIJJIIJJJJJJJCCHIHFHHFHHFFF>#,,;ABDBDDCCCCDDDDB
@SRR765989.3 HWI-ST142:649:C1LEUACXX:2:1101:1206:1979/2
ACTATTGCTCTACATGTTACGGCACTGTGCTGGCATTCTGCTCTGTATTTTGATTTCCTCCAGGGTCTCAAGCCATGAAACNAGACTGTTCATTTTGTTTG
+
B@@FDFFFHBFHHCHIFHAFHIFGIEHBGEFEHG@DHHIJIIGHI?BDGIFCGGIIDHHECGDGHH@;CCEHEHHEHDFC@#,;ACDCCC;>C;BCCBDD?
```

Cohort of One Million

- Fundamentally change the way we understand genomic variation
- The genomic data for a patient is about 1 TB
  - Tumor and normal tissue
- One million genomes is about 1000 PB or 1 EB
  - With compression, it may be about 100 PB
- At $1000/genome, the sequencing would cost about $1B

Mutational Heterogeneity in Cancer



3,083 exome tumor/normal pairs

Lawrence et. al, Nature 2013
doi:10.1038/nature12213

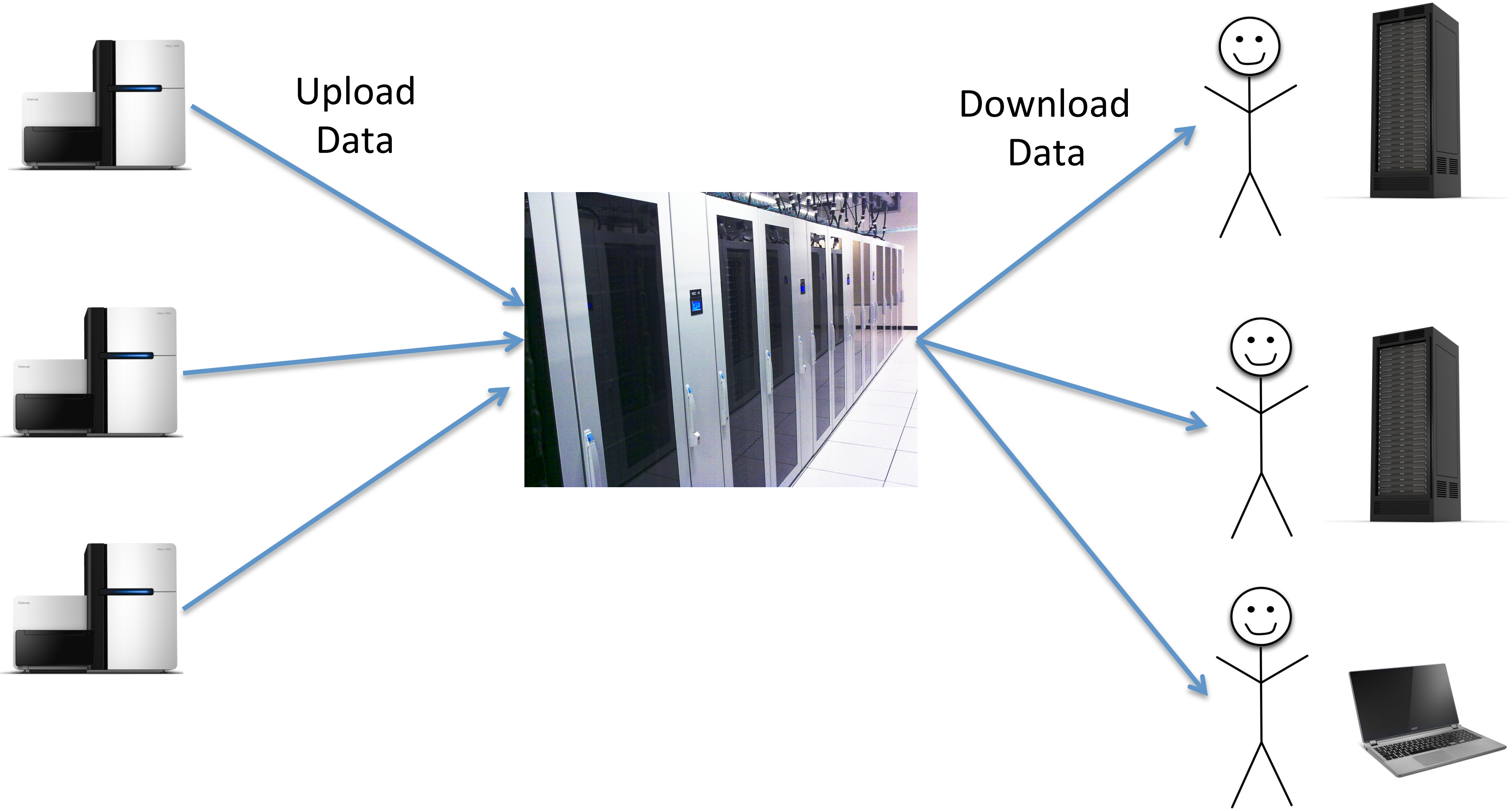
1,000,000 patients
1,000 PB
$1B

10,000 patients
10 PB
$10M

100,000 patients
100 PB
$100M

1000 patients

Standard Practice

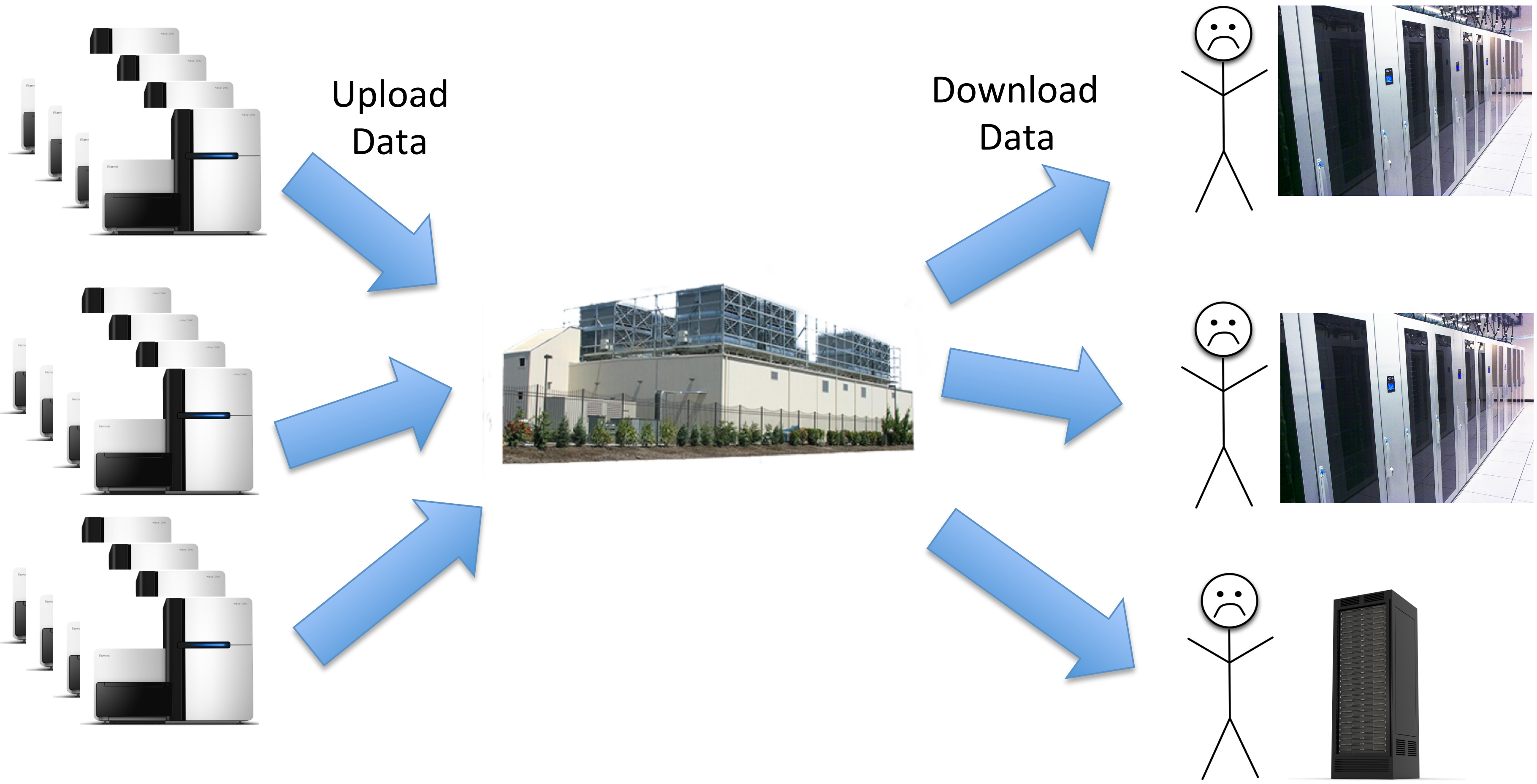


Upload Data

Download Data

Perform Analysis

Standard Practice Breaks

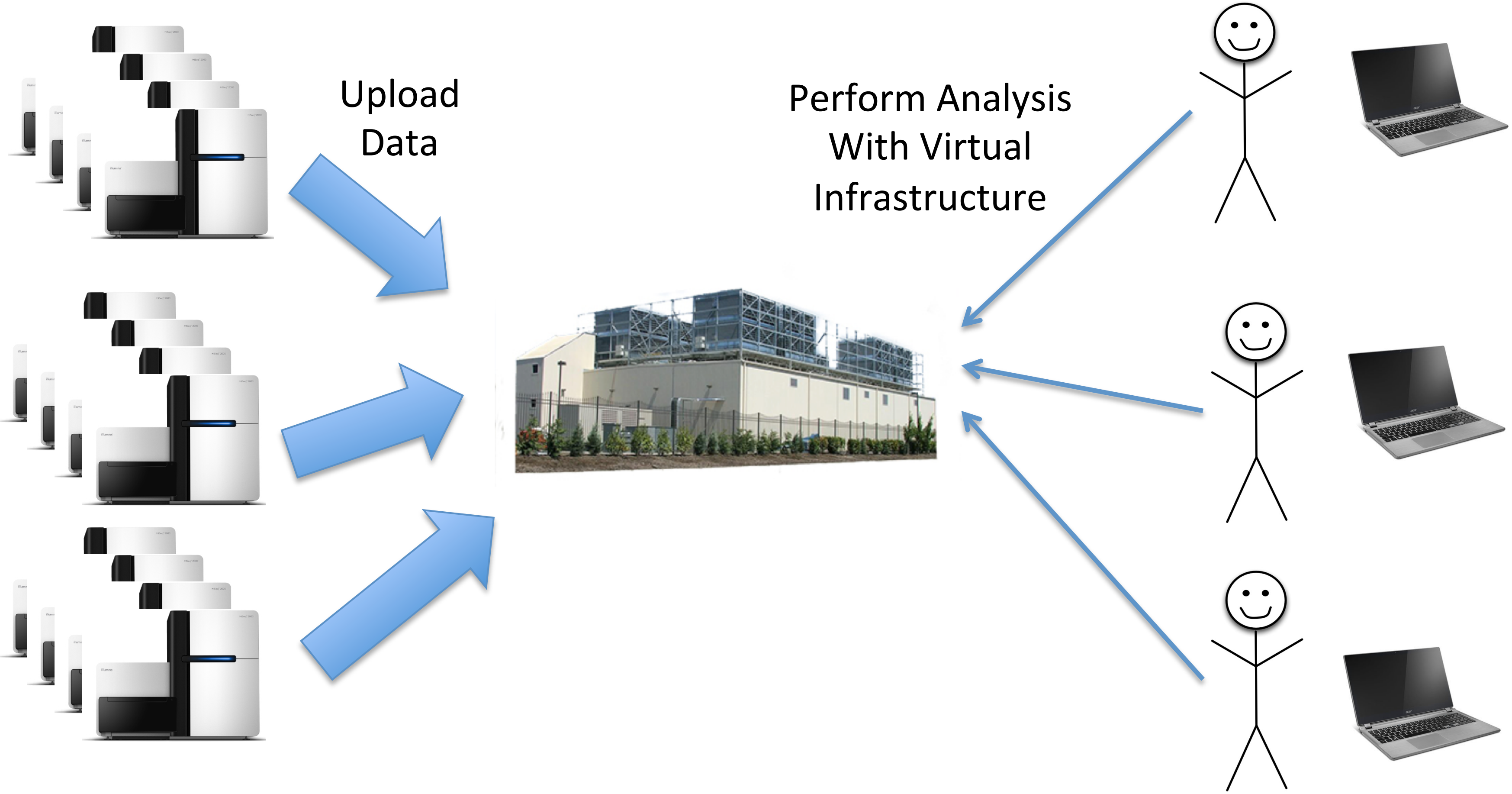


Upload Data

Download Data

Perform Analysis

Cloud Computing Enables New Model

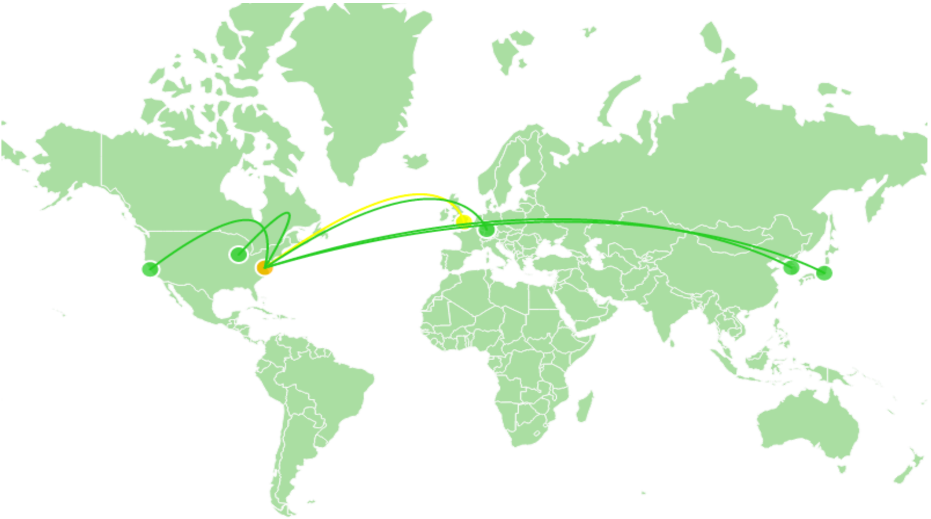


Upload Data

Perform Analysis With Virtual Infrastructure

International Cancer Genome Consortium (ICGC) PCAWG

- Paired tumor/normal whole genomes with >=25X coverage
- Utilize cloud infrastructures across the world to uniformly align and call variants
- Many lessons learned
- http://pancancer.info/



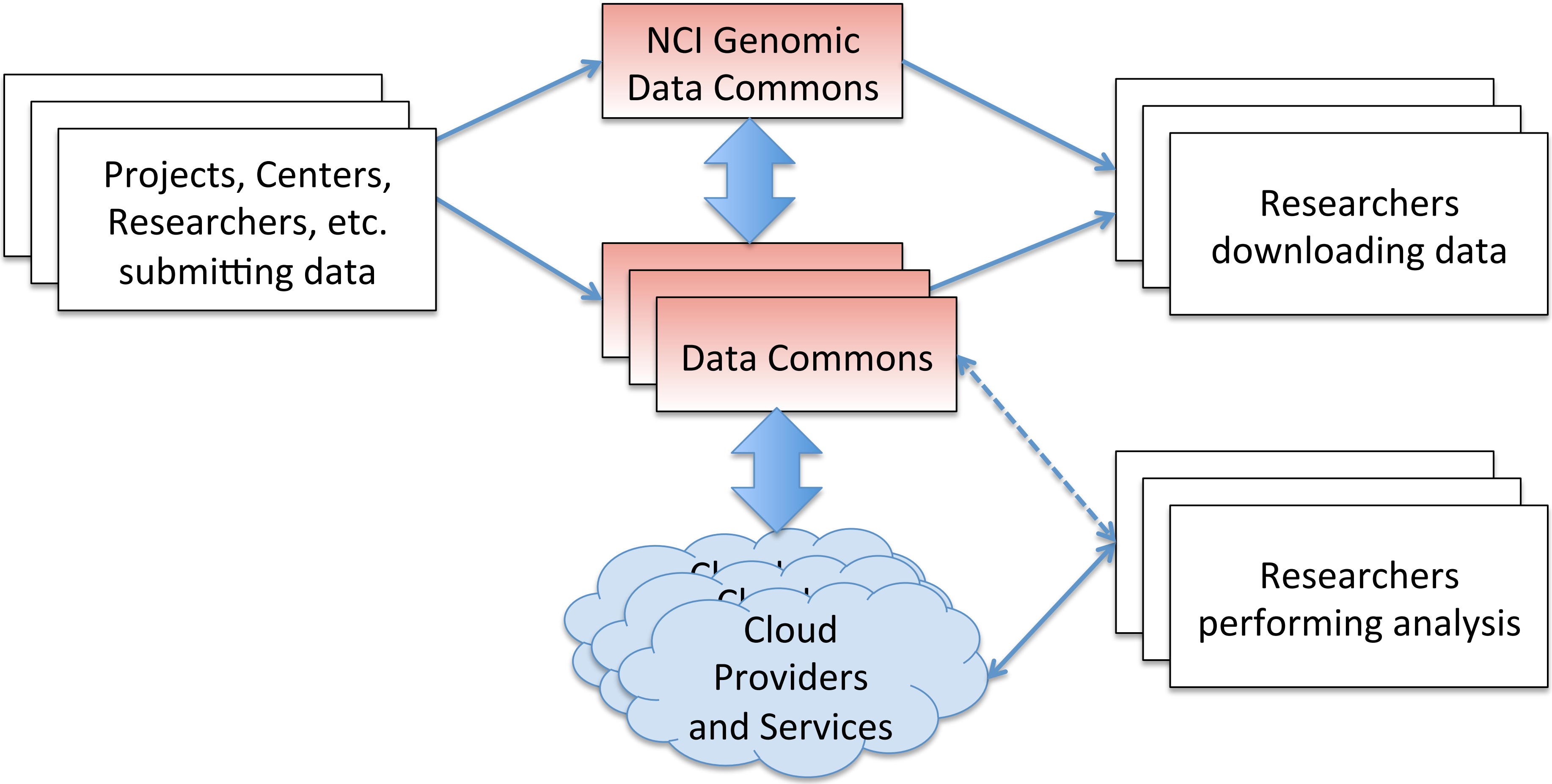
Data Commons

- A shared community driven data resource
- Cloud (virtualized) infrastructures enable bring analysis to the data
- Data management
- Interoperability

NCI Genomic Data Commons

- Two genomics projects with > 1 PB data sets
  - Many future projects slated
- The Cancer Genome Atlas (TCGA)
  - Over 11,000 patients across 25 cancers
- Therapeutically Applicable Research to Generate Effective Treatments (TARGET)
  - 5 childhood cancer types
- Store, harmonize, analyze, distribute
- Platform for democratizing data

Many Commons, Many Clouds



Projects, Centers, Researchers, etc. submitting data

NCI Genomic Data Commons

Data Commons

Cloud Providers and Services

Researchers downloading data

Researchers performing analysis

Lung Cancer Classification

- Demo!

This Should be Easier

- Data commons would have speed up this analysis

- Raw data was hosted on Bionimbus PDC
  - Months to gather and understand metadata, including "hidden" annotations on data

- Difficult to replicate legacy TCGA pipelines

Data Commons Principles

- Harmonized meaningful metadata
- Digital identifiers for data
- APIs
- Computational capacity for analyses
- High performance, wide area networks
- Goal: Scientific discovery and real world impact

Metadata and Data Models

- Need for minimal mandatory set of metadata
  - Traditionally complex XML
  - JSON
  - RDF / JSON-LD
- Flexible data model
  - Traditionally rigid normalized schemas for relational databases

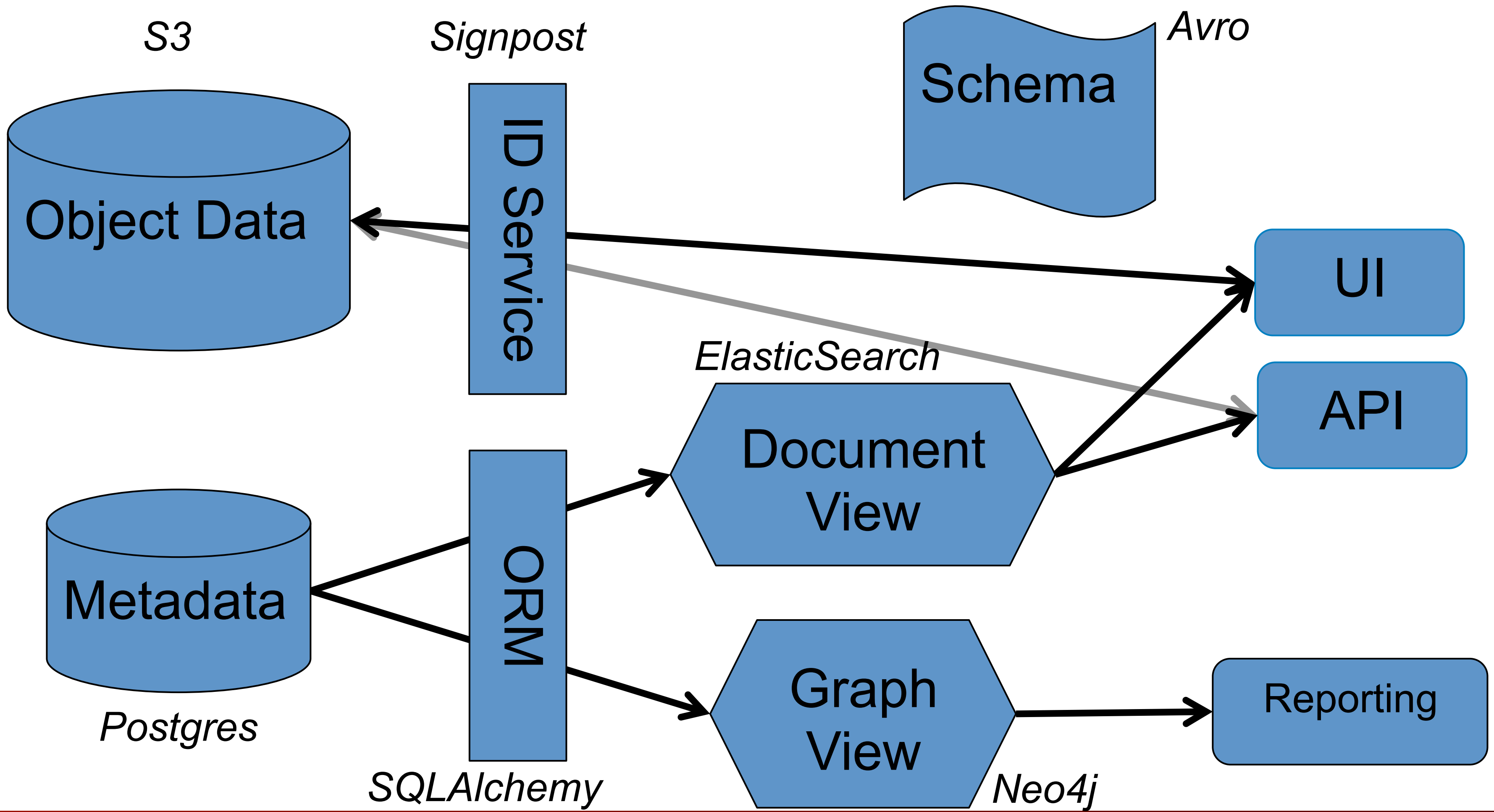
Metadata and Data Models

- Continue to accept standard XML formals
- Developing minimal set of metadata, encoding with Apache Avro
  - Following efforts of the Global Alliance for Global Health (GA4GH)
  - Rethinking
- Data model stored as a property graph
  - Evaluated graph databases, not ready
  - Persist node/edge data in Postgres, export to graph databases for advanced querying
    - Properties stored as jsonb, validating using Avro schemas

Digital IDs – "Signpost"

- Very lightweight "DNS" for data
  - Maps a UUID to a list of URLs
  - ACLs for ownership and updates of URLs
  - RESTful API
- Separates data from metadata
  - Registered data is immutable
- Planning:
  - Discovery and namespaces
  - Client optimizations based on data location

Framework – Tech Choices



Realignment and Higher Level Analysis

- New reference genomes and new algorithms cause a need for periodic reprocessing
- Computationally demanding
- Requires workflow and resource management
- Lesson learned from ICGC and other projects:
  - Creating virtual clusters that look like HPC environments is not a good idea or effective use of resources
  - Developing lightweight and fault-tolerant system for managing analyses in cloud environment

Summary

- National system to store, harmonize, analyze and distribute existing cancer genomics data
  - Currently roughly 2 PB and growing to 10 PB
- First step toward the development of a "Knowledge System" for cancer
  - Originally outlined in the Institute of Medicine Report entitled "Toward Precision Medicine."
- Built on open-source cloud computing technologies
- One template for future data commons

GDC Portal

- Demo!