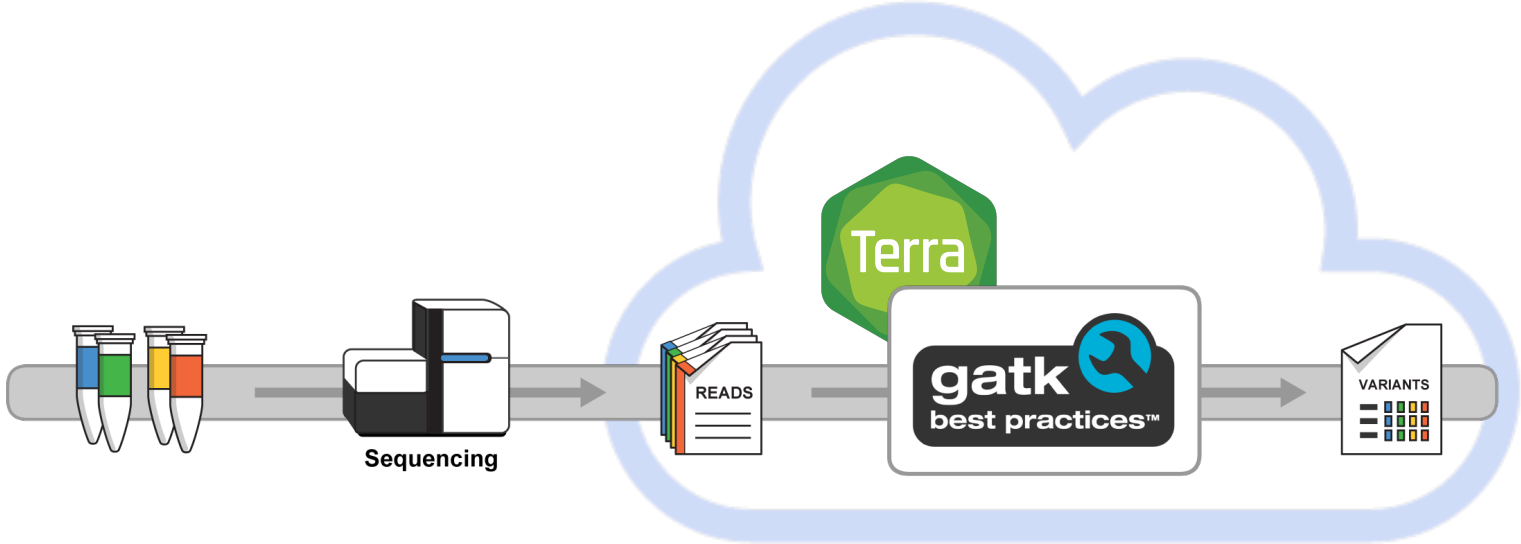
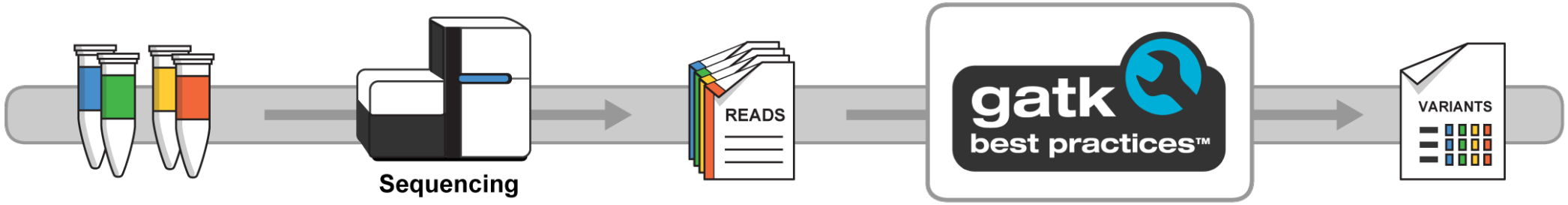
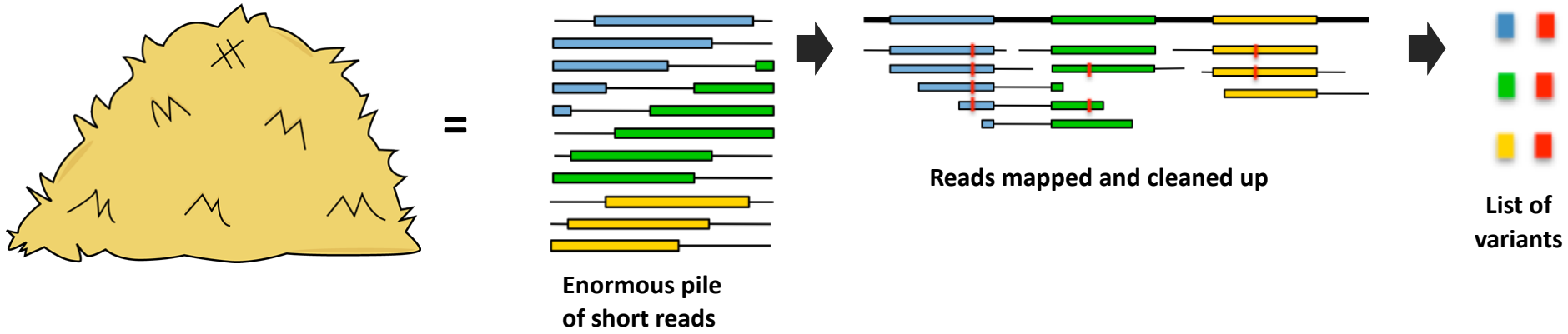
GATK & Terra

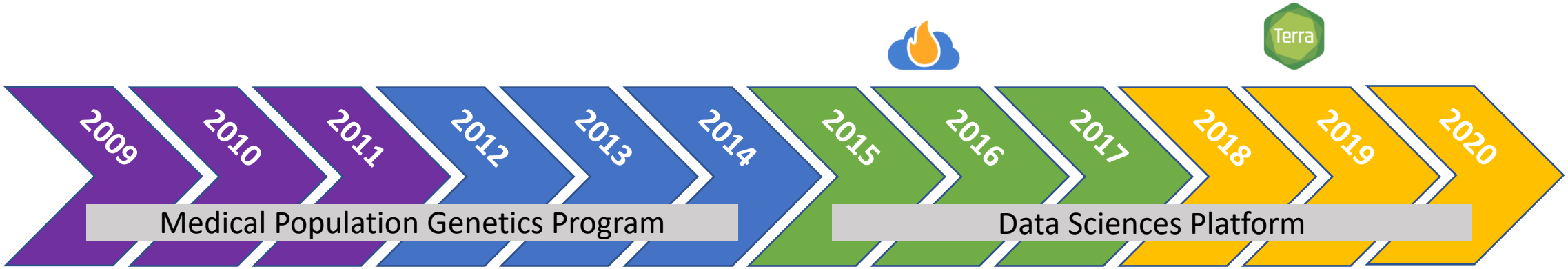
A brief history



GATK = Genome Analysis Toolkit







GATK 1.0
framework

1000 Genomes Project

Open source (MIT)

GATK 2.0
toolkit rebranding

NHGRI iSeqTools

Hybrid licensing with partner (Appistry)

GATK 3.0
scaling up

ExAC + gnomAD

Hybrid licensing @Broad (all in-house)

GATK 4 BETA
re-engineering

GATK 4.0
cloud support
& scope expansion

Open source (BSD-3)

Intel-Broad Center ----------

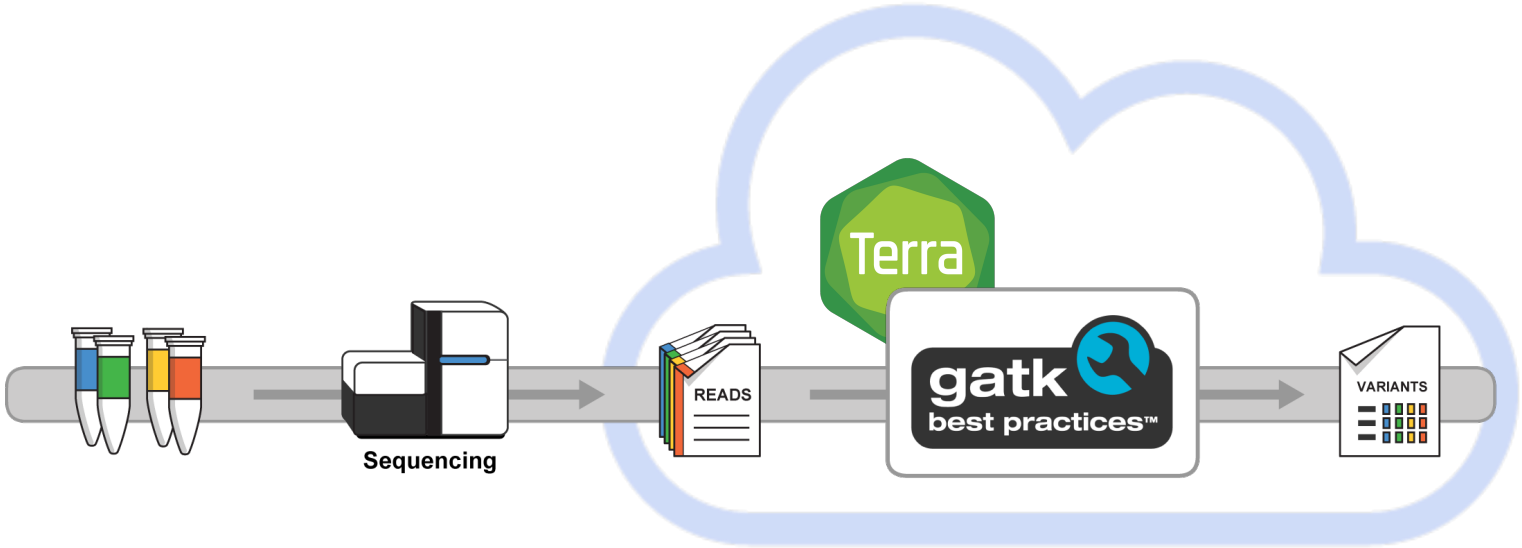
Google Cloud Partnership ----------

Cloud 6 Partnership
GCP, MSFT Azure, AWS, IBM, Cloudera, Intel

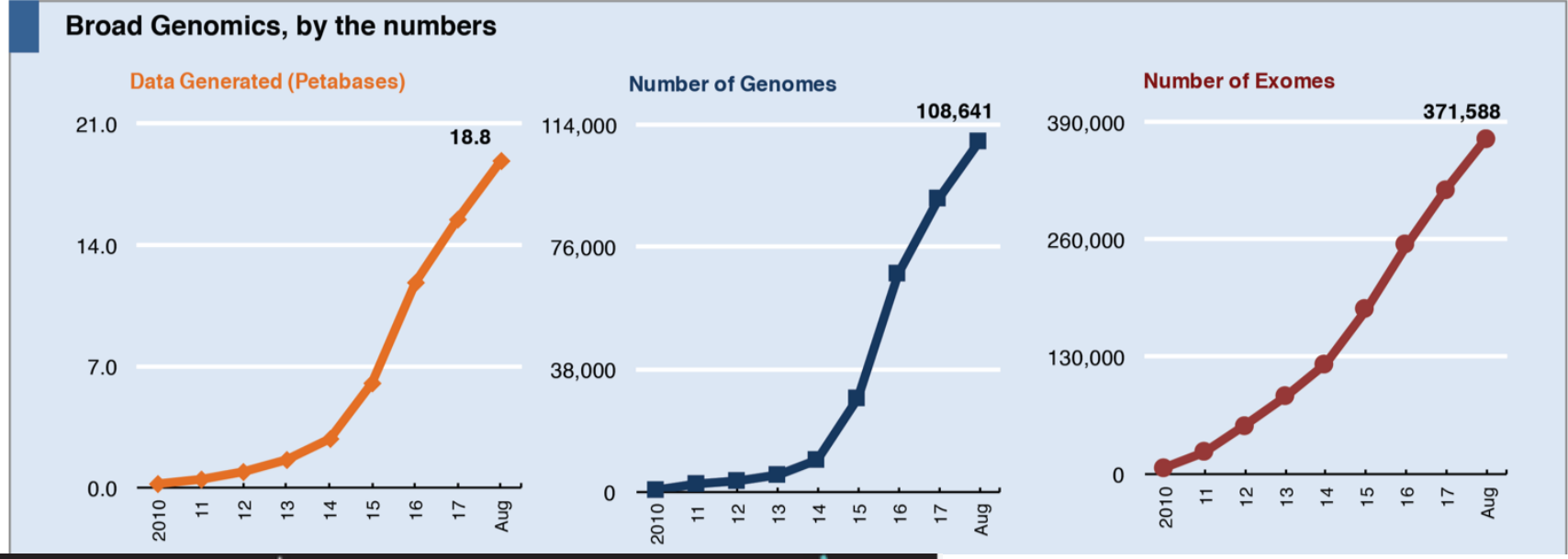
Alibaba Cloud

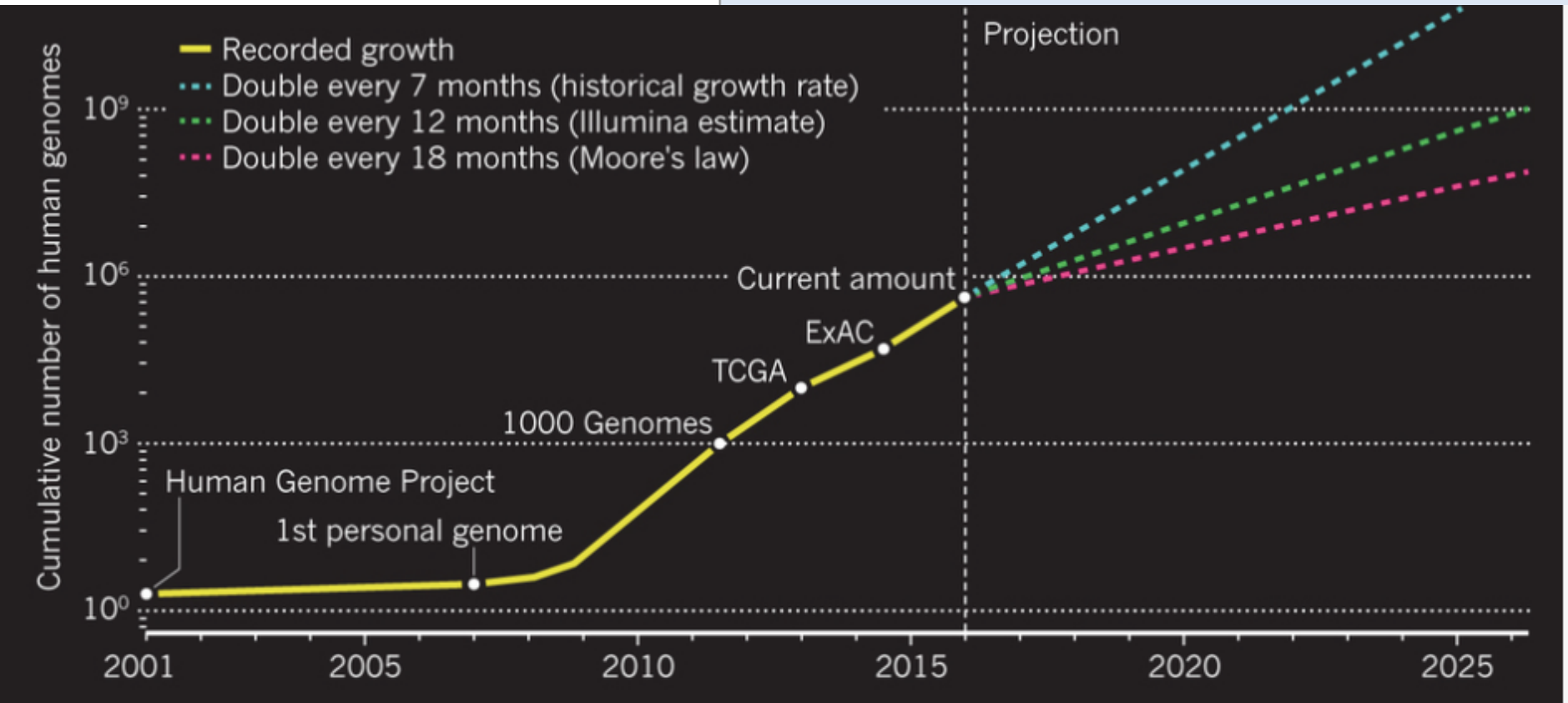
ILMN DRAGEN

The real story is in the cloud



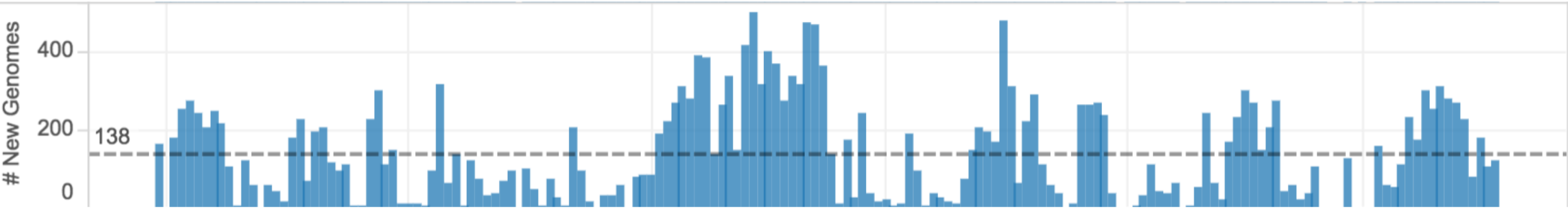
Explosive growth in data generation





Pipeline throughput is "spiky"

Genome processing requests per day over several months



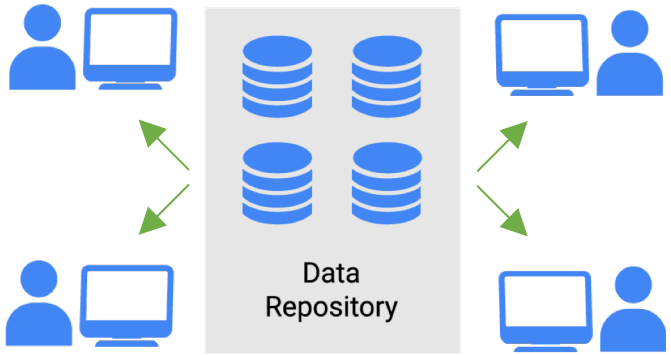
- **Advantages of cloud over on-premises computing:**
  - No need to pay for compute power when we aren't using it
  - Can tolerate spikes without being forced to maintain a backlog of "things to process once everything calms down"

+ No longer need to take over Allston, MA with giant datacenter

Data access and federation are key

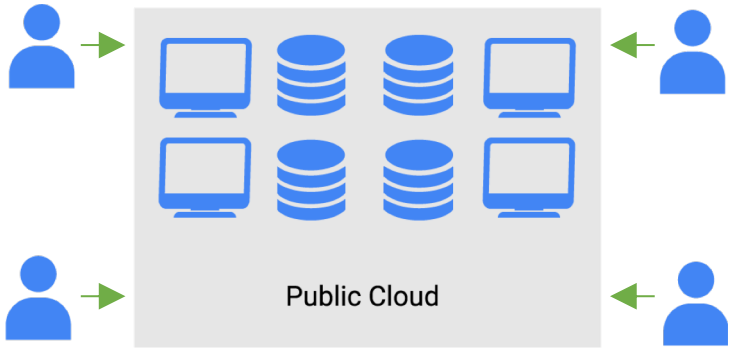
Traditional Way

Bring data to the researchers



Cloud Way

Bring researchers to the data



Problems

Data sharing = data copying
Requires big infrastructure at each site
Largely fixed compute
Individual security implementations

Solutions

True data sharing
Cloud provides the infrastructure
Elastic compute and storage
Centralized security implementation

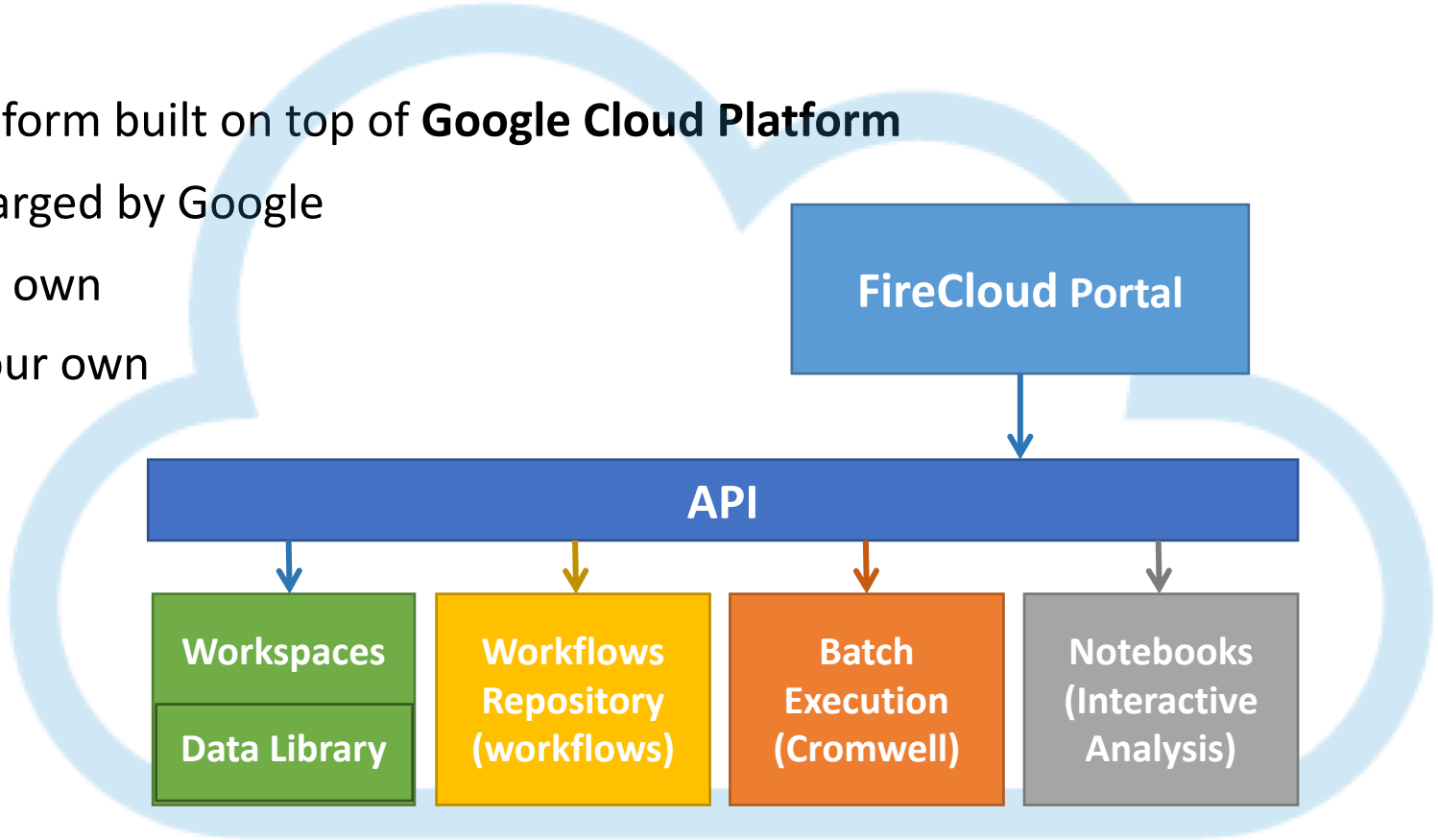
Data Sciences Platform developed FireCloud

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

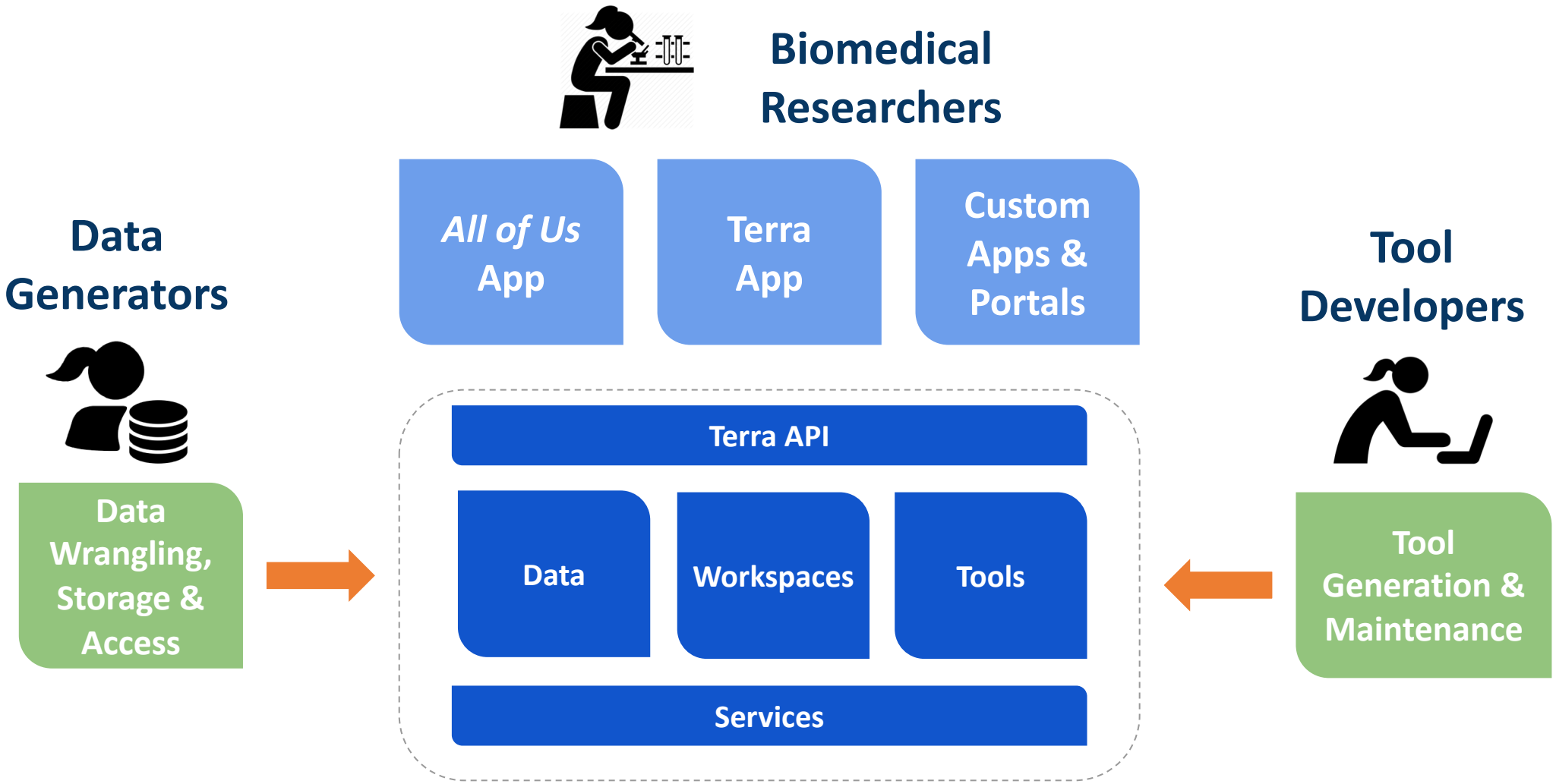
Cloud Pilots →

FireCloud

- Collaborative **cloud-based** analysis platform built on top of **Google Cloud Platform**
- **Free to access** / compute & storage charged by Google
- Access published **data** and/or add your own
- Access existing **methods** and/or add your own
- **Execute** analyses in auditable pipelines
- **Share** data, methods and results with collaborators



Next step: Generalize platform model for data access & analysis

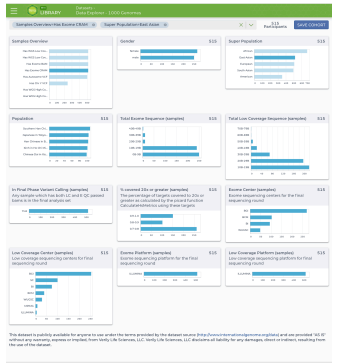


Data Sciences Platform develops Terra with verily

Terra supports the full lifecycle of data access & analysis

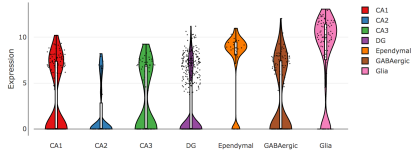
STORE DATA

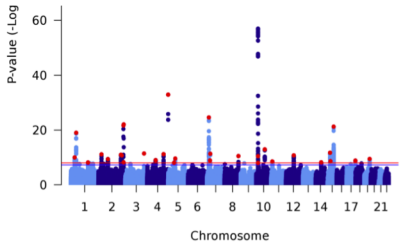
SEARCH & ACCESS



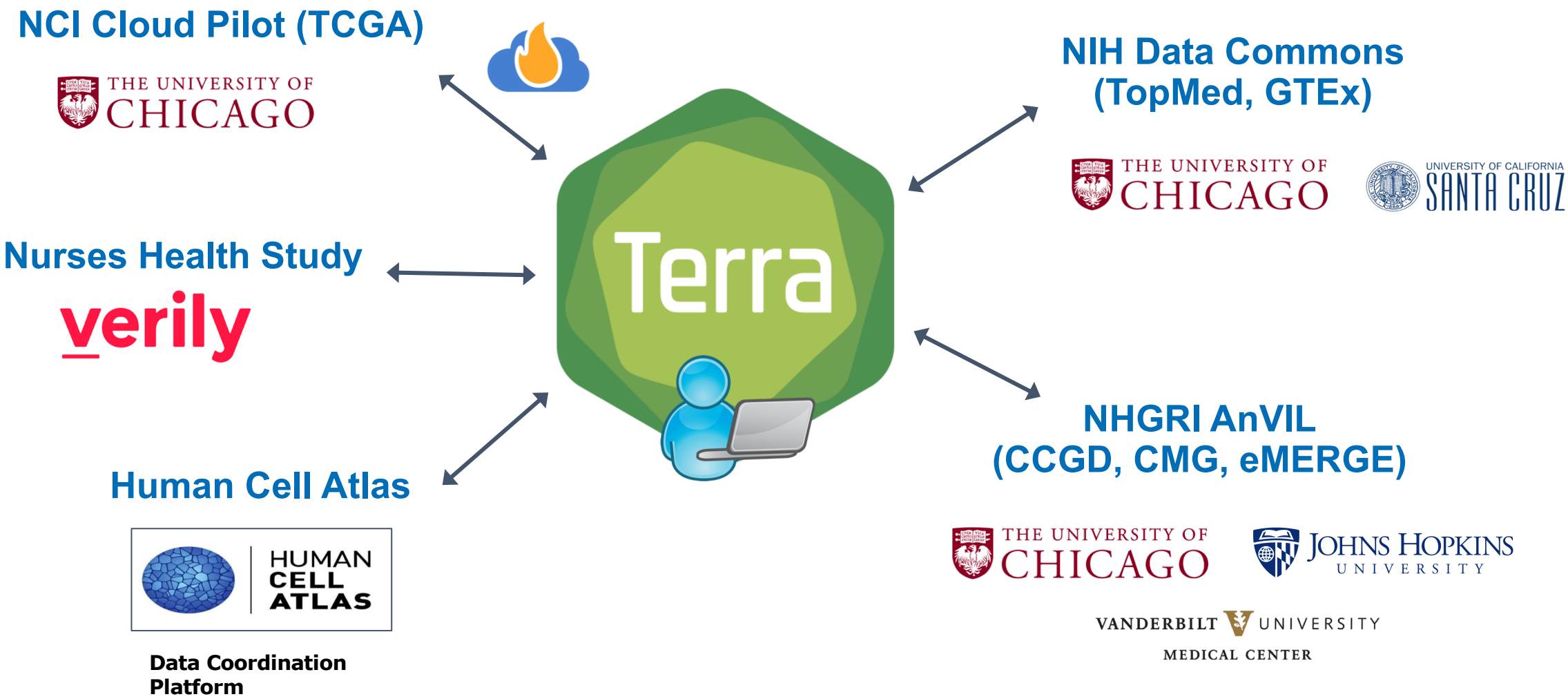
PROCESS | ANALYZE

VISUALIZE & INTERPRET





Terra connects to datasets hosted by many orgs / apps



Data Biosphere

A shared vision for an open ecosystem

- **Modular**
- **Open**
- **Community-driven**
- **Standards-based**

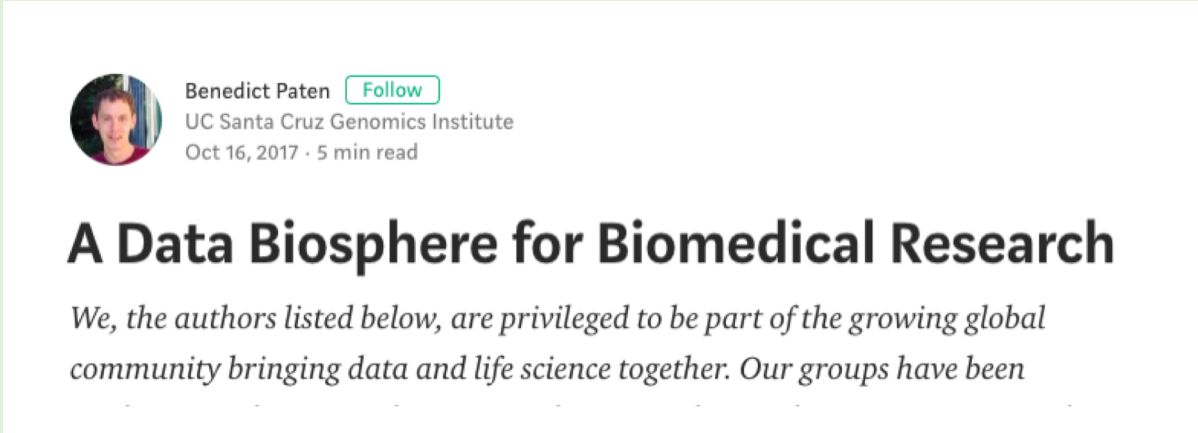
BROAD INSTITUTE

THE UNIVERSITY OF CHICAGO

UNIVERSITY OF CALIFORNIA SANTA CRUZ

VANDERBILT UNIVERSITY MEDICAL CENTER

verily



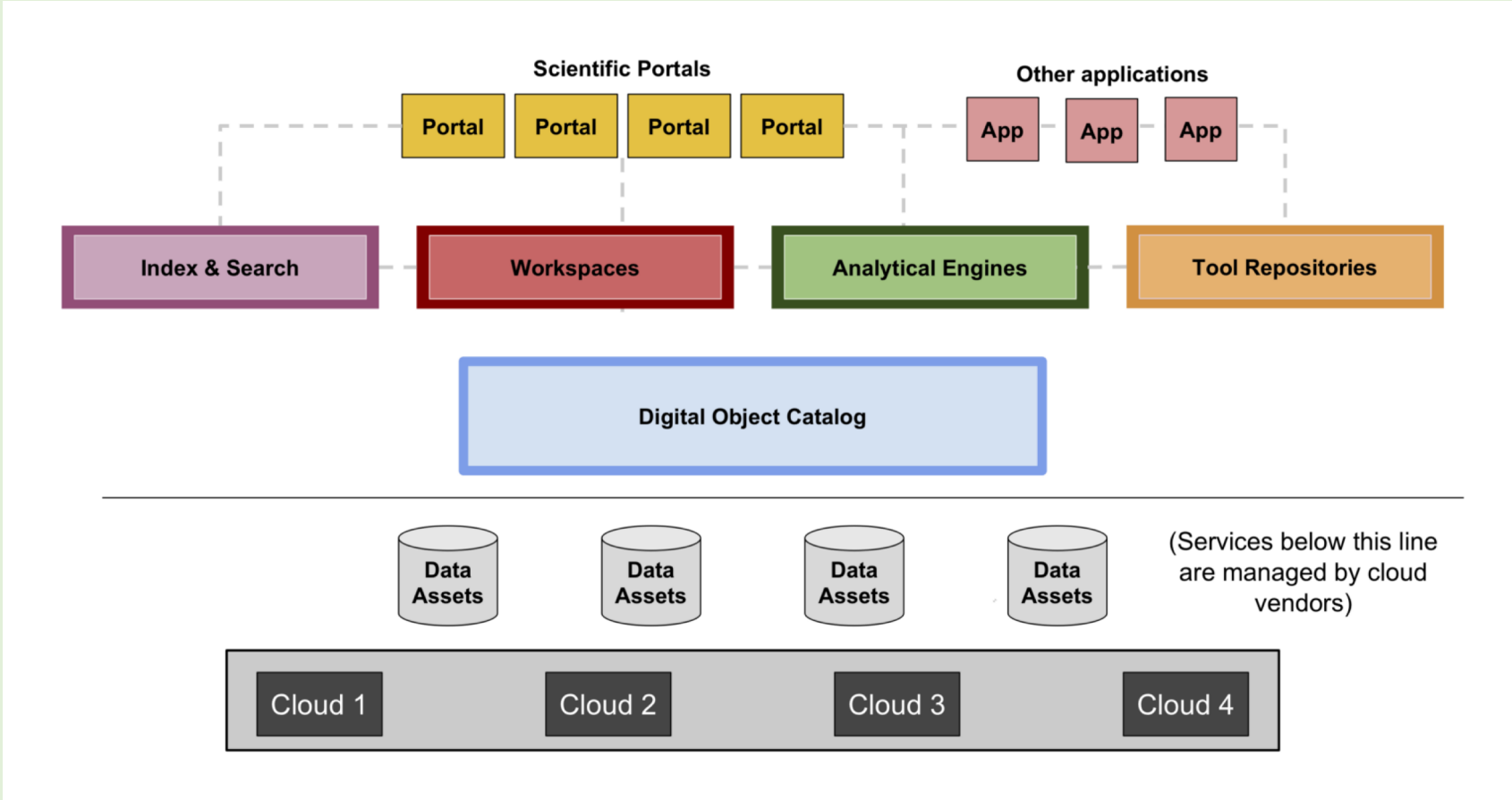
Benedict Paten Follow
UC Santa Cruz Genomics Institute
Oct 16, 2017 · 5 min read

A Data Biosphere for Biomedical Research

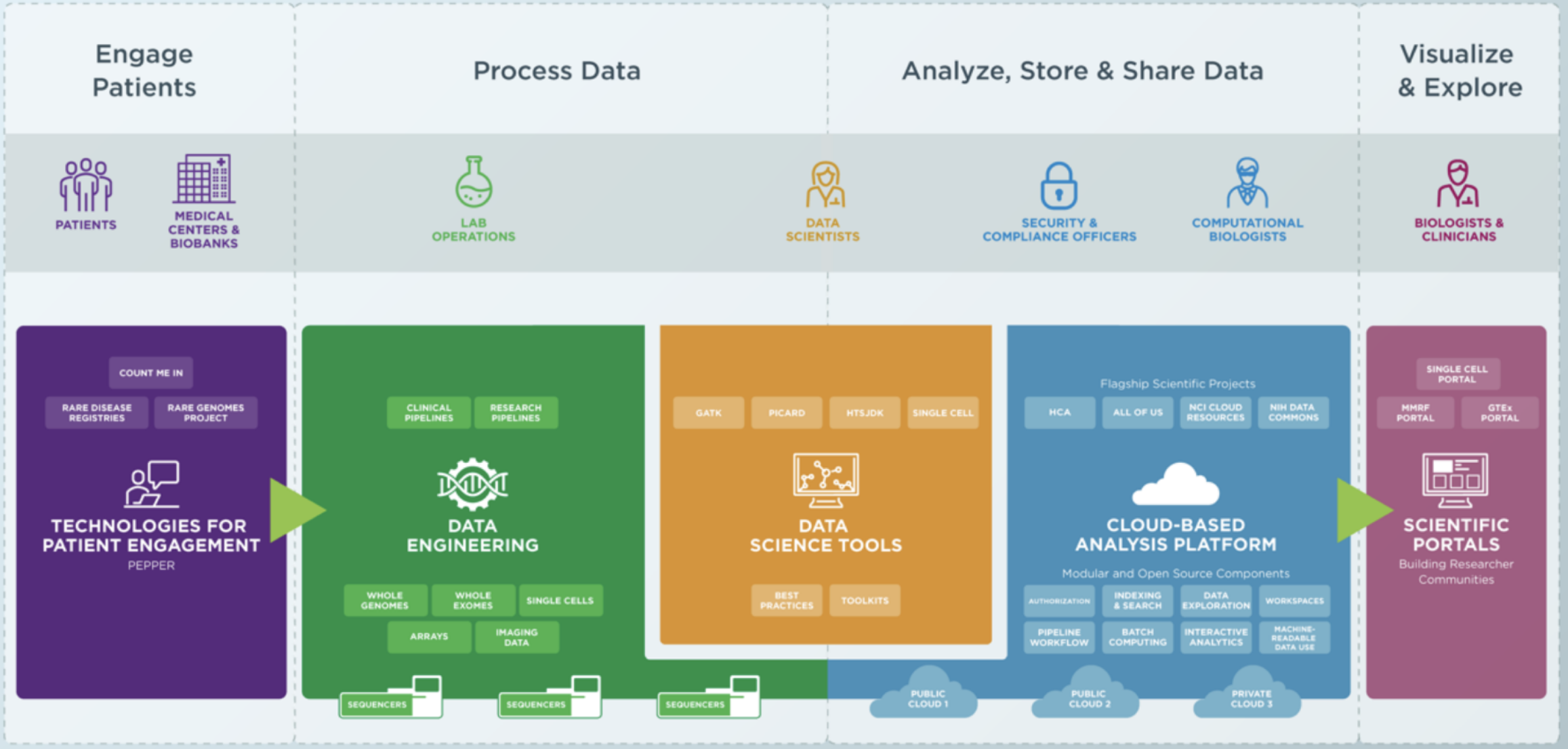
We, the authors listed below, are privileged to be part of the growing global community bringing data and life science together. Our groups have been

Read the manifesto on MEDIUM

broad.io/databiosphere



DSP Team and Mission



That’s the story so far…