Genomic Scale Data

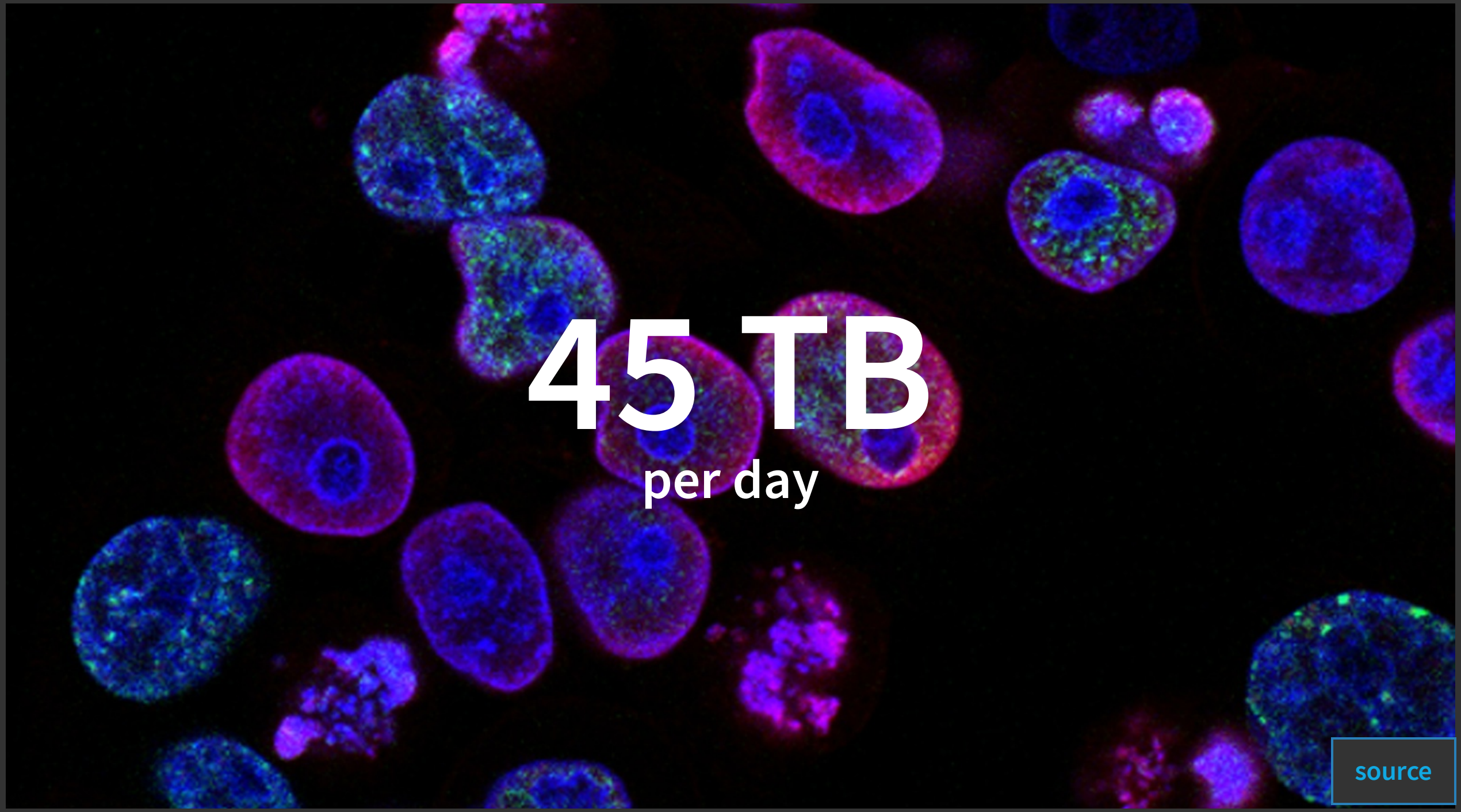
Midwest BioDefense Summit

Lynn Langit

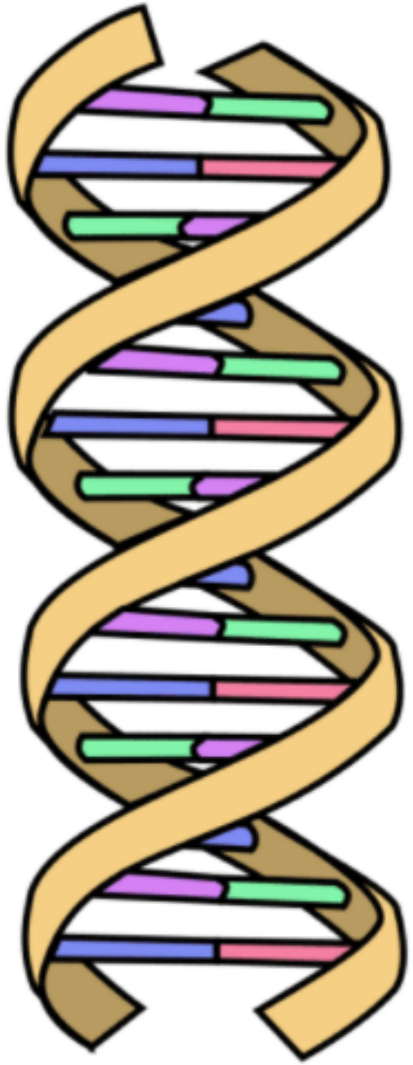
Big Data Considerations

Applied to Bioinformatics

Volume	Velocity	Variety	Veracity
How much?	How fast?	What kinds?	Quality?


45 TB
per day
source

VOLUME



DNA

=Adenine

= Thymine

= Cytosine

= Guanine

= Phosphate backbone

99.9% alike

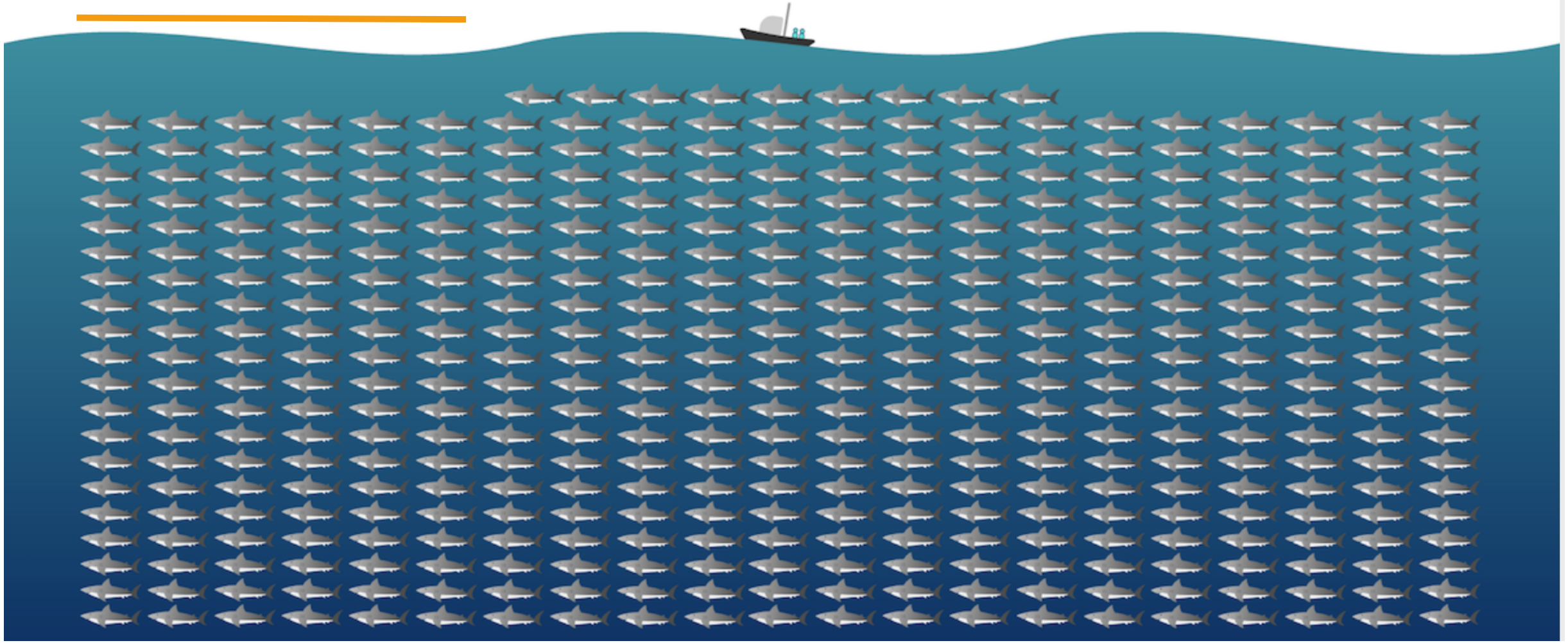
3 Billion *.1 ≈ 300 Million

source

How big is 40 exabytes?

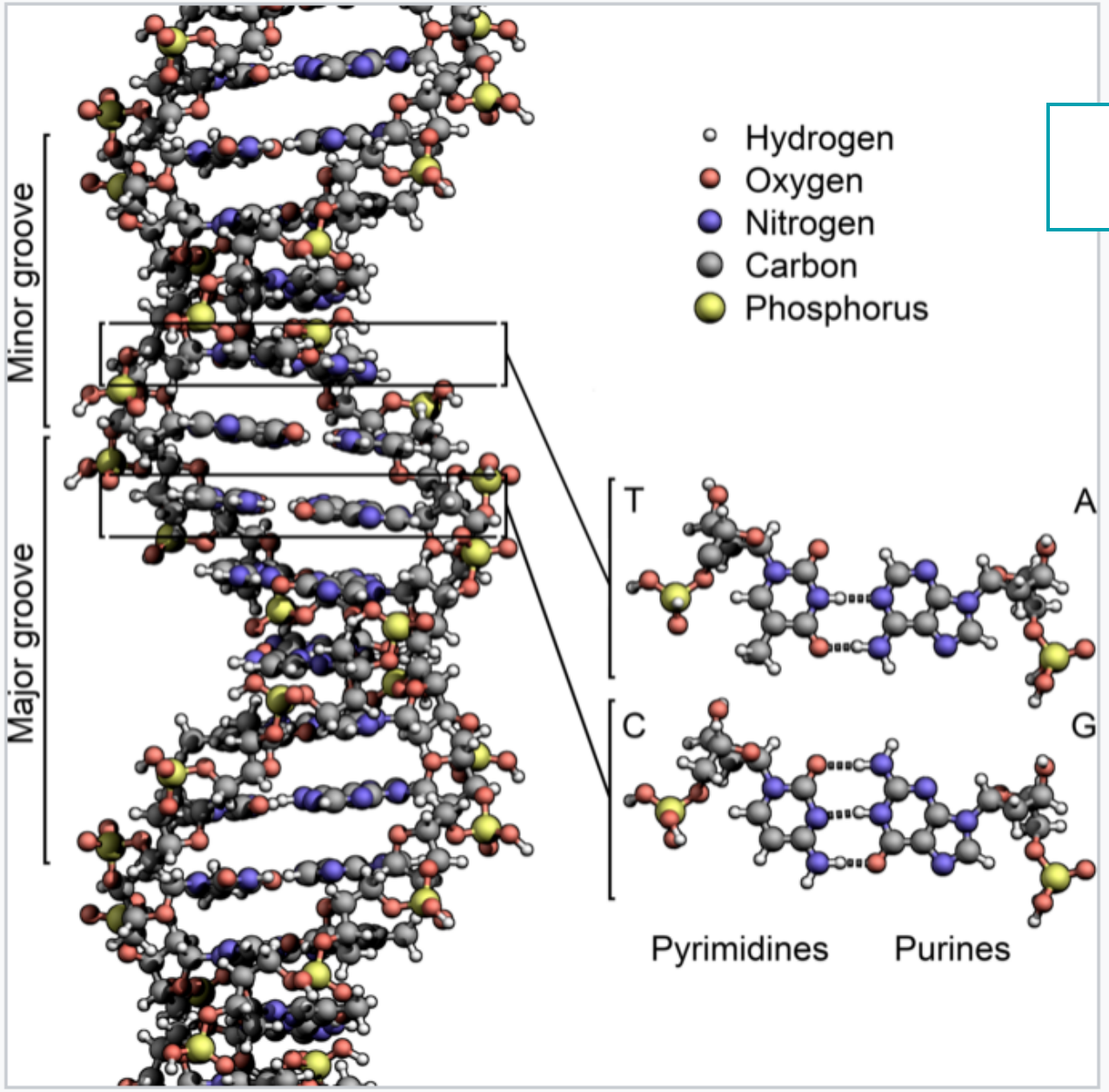
Genomics projects will generate 40 exabytes of data in the next decade.

Each shark = 100,000,000 GB of data



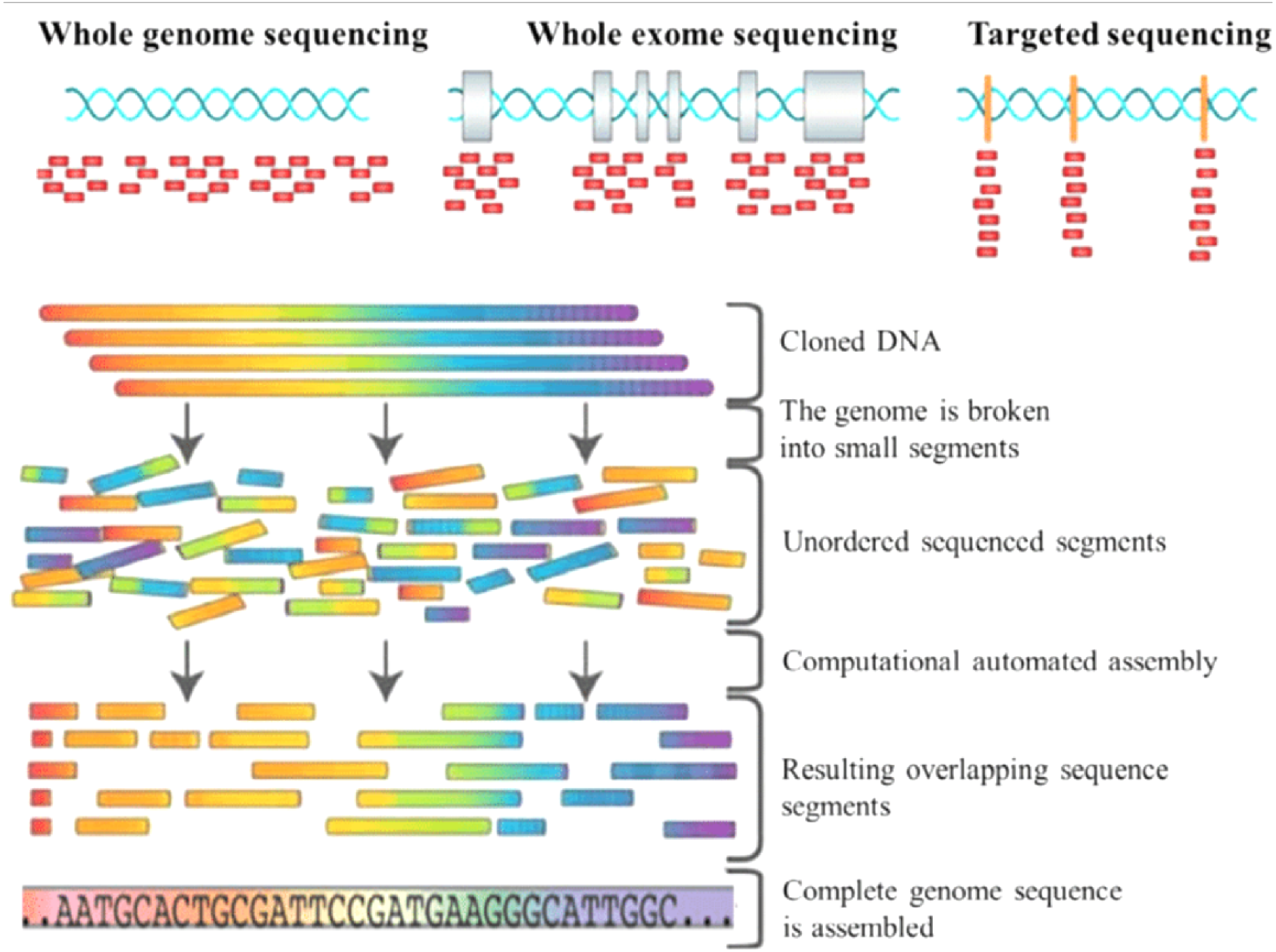
VELOCITY

source



VARIETY

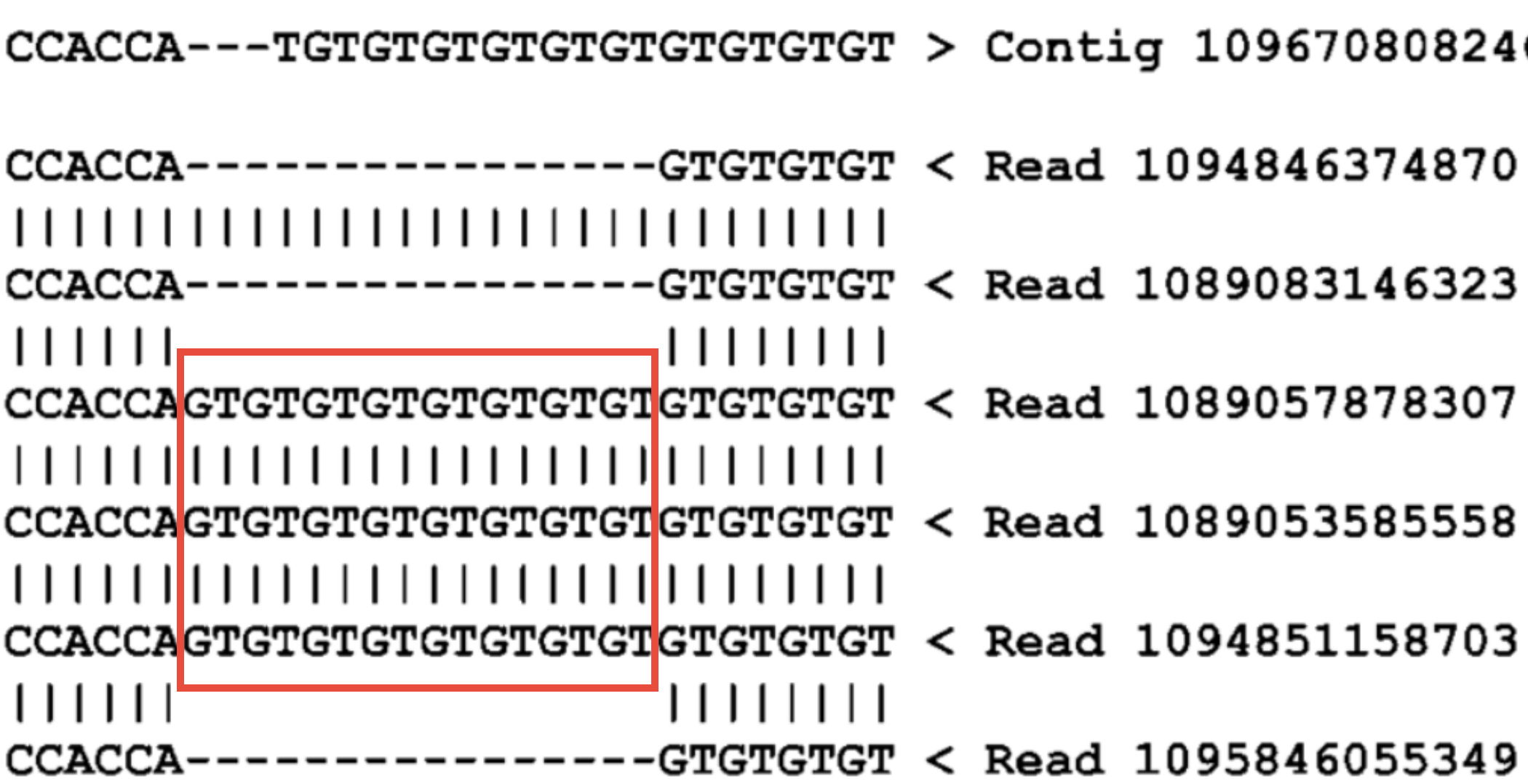
source



VARIETY

source

VERACITY



source

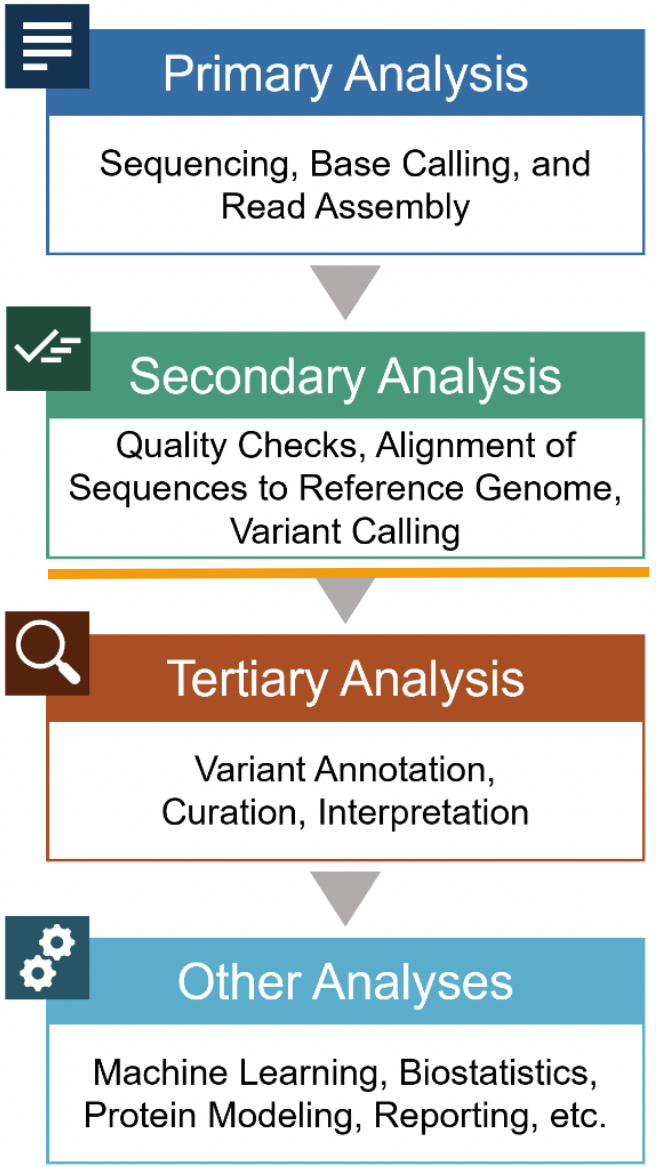
Research Challenges

1. Analysis takes too long!
2. Emergency Scalability!
3. Out of storage space!*

*and more...


500
Hours

Analysis Phases



A high-level overview of NGS data processing

Base calling
(vendor tool)

FASTQ file:
raw NGS reads

Alignment
or assembly

SAM/BAM file:
aligned NGS reads

Variant calling

VCF file:
genomic variation

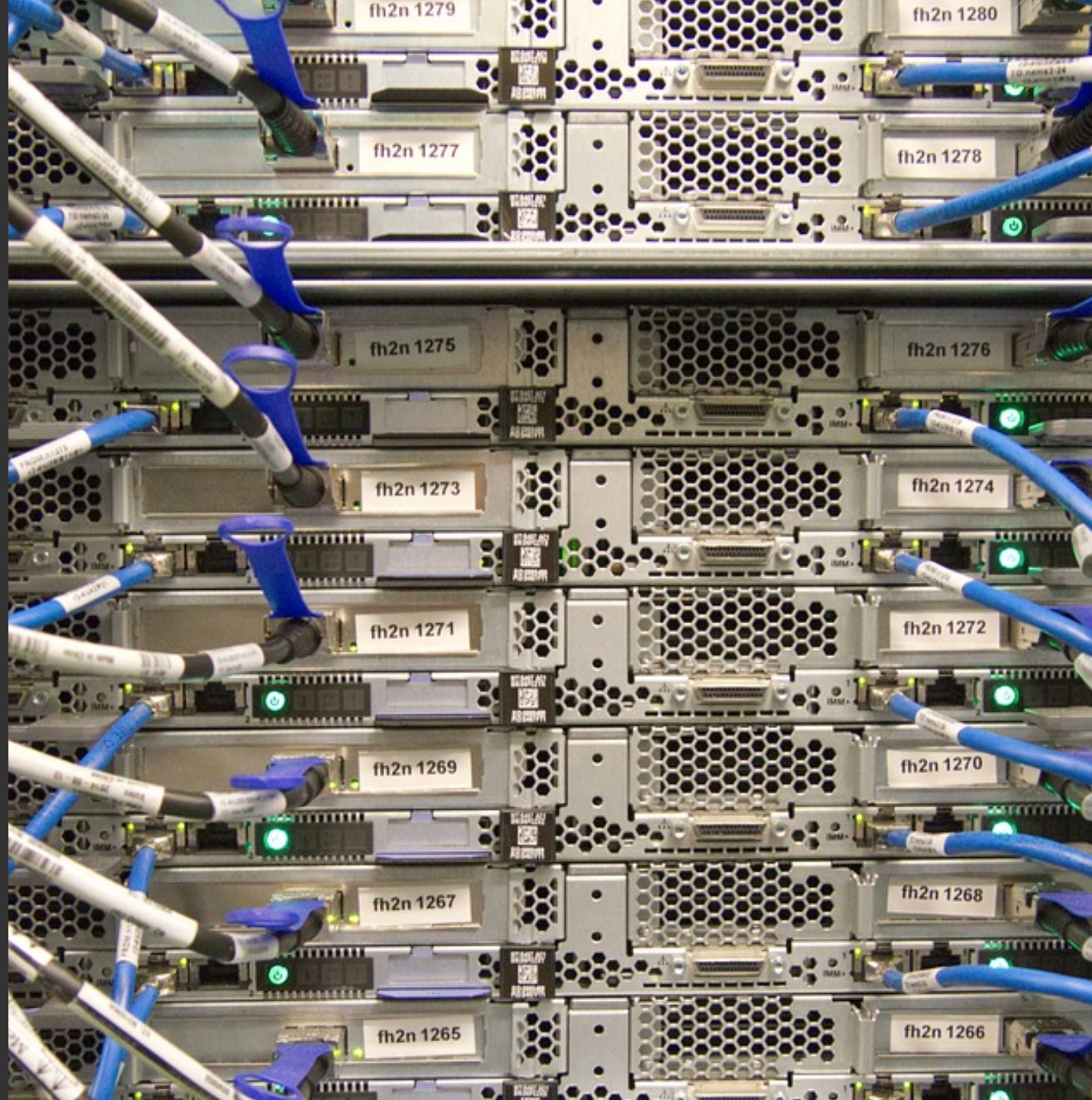
.sam file: Uncompressed text file
.bam file: Compressed and indexed file

VCF Files can be block-compressed and indexed.

source

Use the Cloud

- Google Cloud: GCP
- Amazon: AWS
- Microsoft Azure



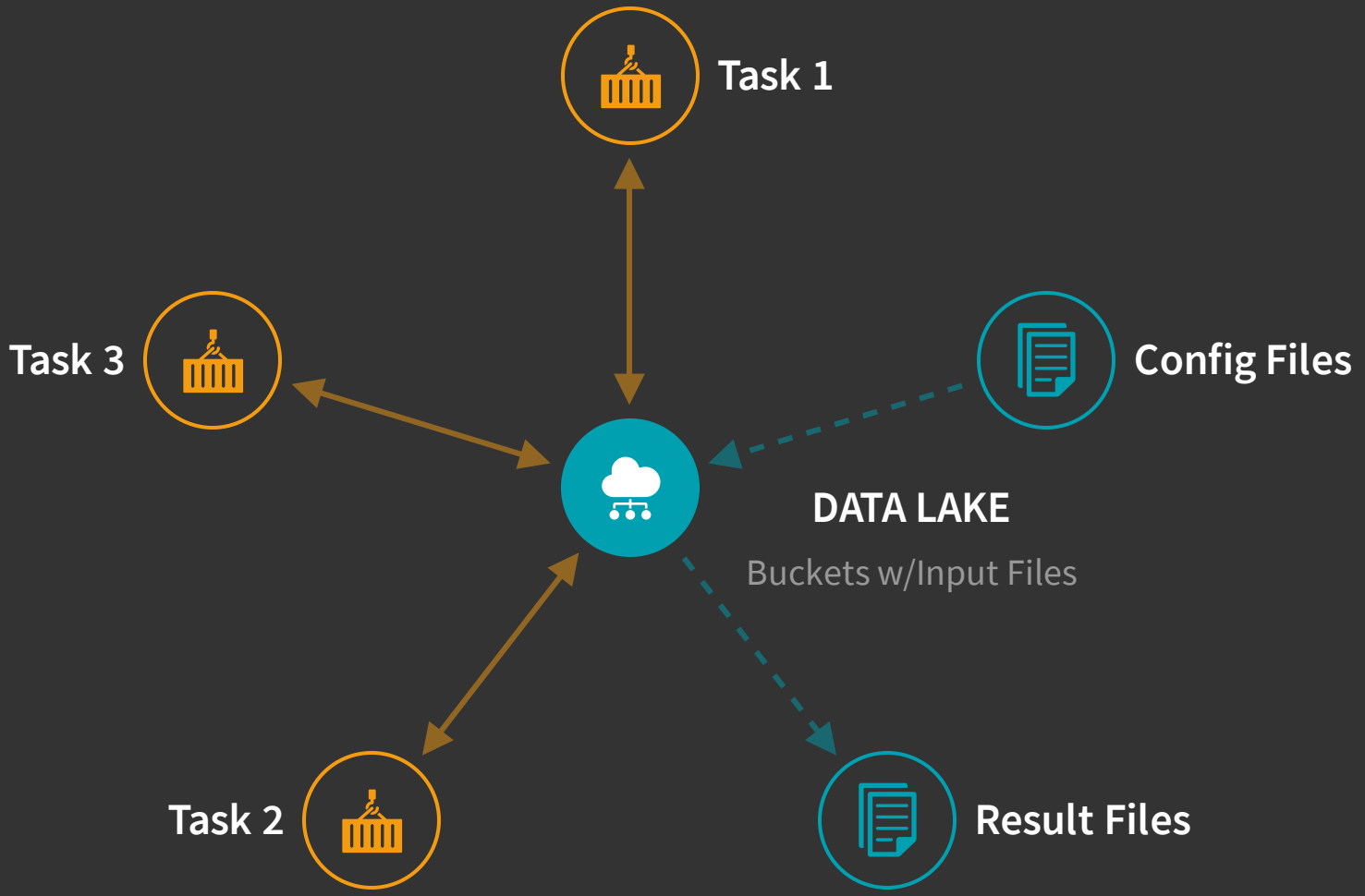
Practical Considerations for Developers

Dev Env

Sample Data

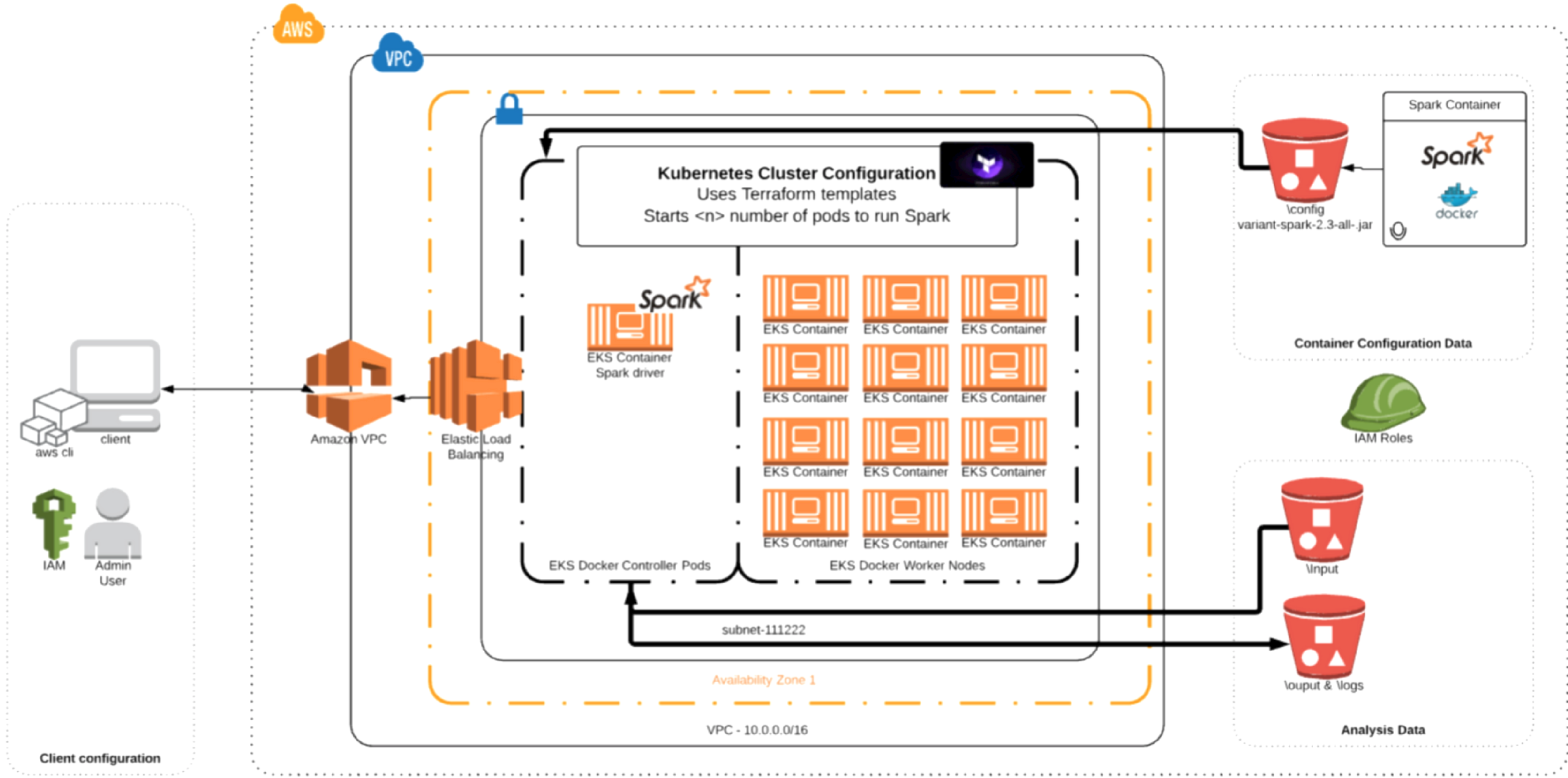
Readiness

Pattern 1: Cloud-Native Data Lake



AWS EKS SPARK

Lynn Langit



Traditionally the columns (features) of large “raw” datasets needed to be scaled down to allow advanced analytics to take place. This process of eliminating information potentially biases the result and impacts on the sensitivity of the analysis.
VariantSpark is different, it can deal with the raw data directly resulting in a more accurate output.

Variant Spark was tested on datasets with 3000 samples, each containing 80 Million features, in either unsupervised clustering approaches (e.g. k-means) or supervised applications (e.g. Random Forests) with target/truth values that are categorical (classification) or continuous (regression). It can cluster data according to required profiles or identify predictive markers of events in just 30 minutes.
By building a model on the full dataset, this allows, for the first time, to identify sets of markers that together have a stronger predictive power than the previously identified independent markers.

Source

Results

1. **Speed (30,000 -> 30 minutes)**
2. **Predictable Costs**
   - Cloud Data Lake
   - Reproducibility / Collaboration

Source

Population-based study reveals strong genetic links with multiple cardiometabolic diseases and traits

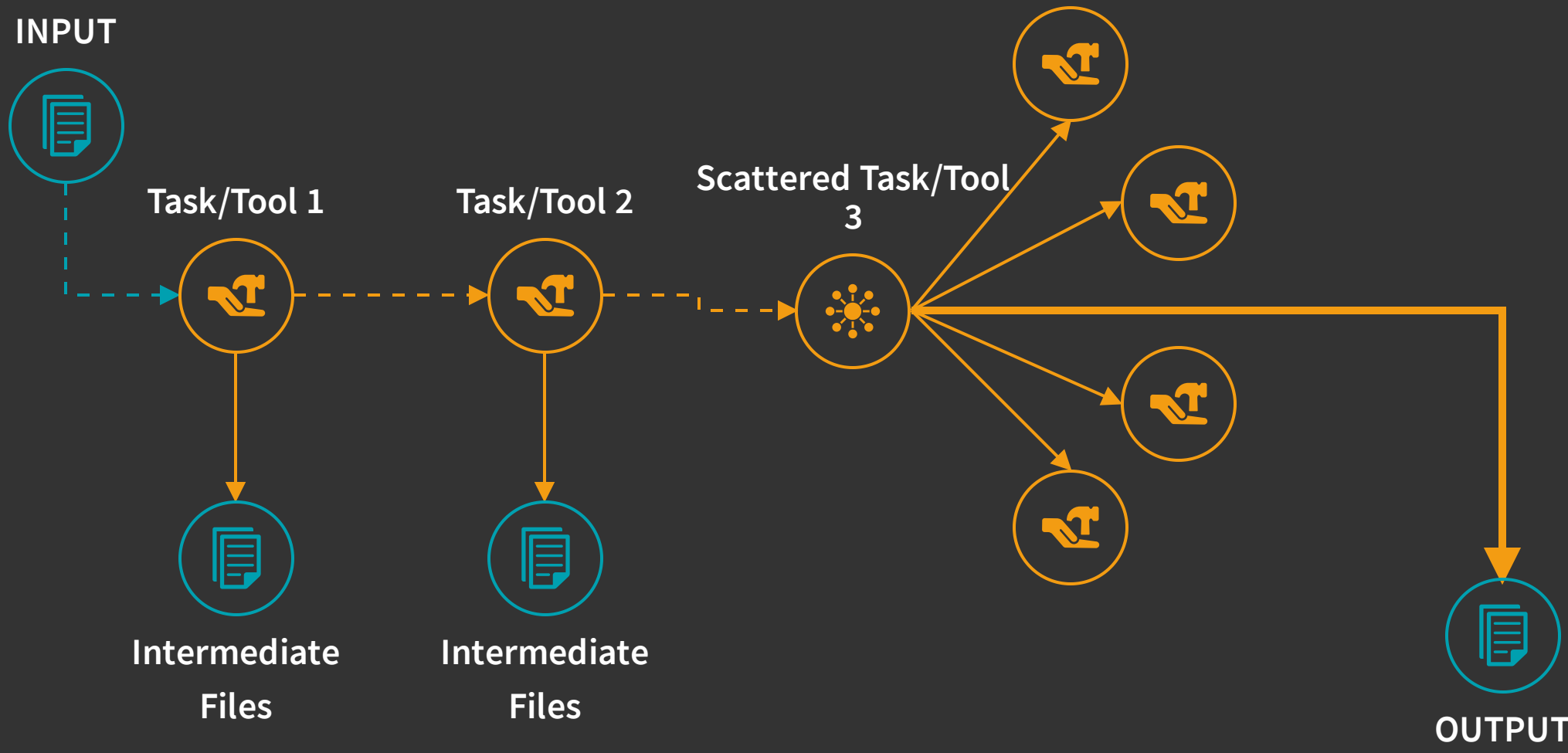
By Nora Bradford

Broad scientists analyzed sequencing data from more than 200,000 people and found rare genetic variants associated with diseases like diabetes and heart failure.

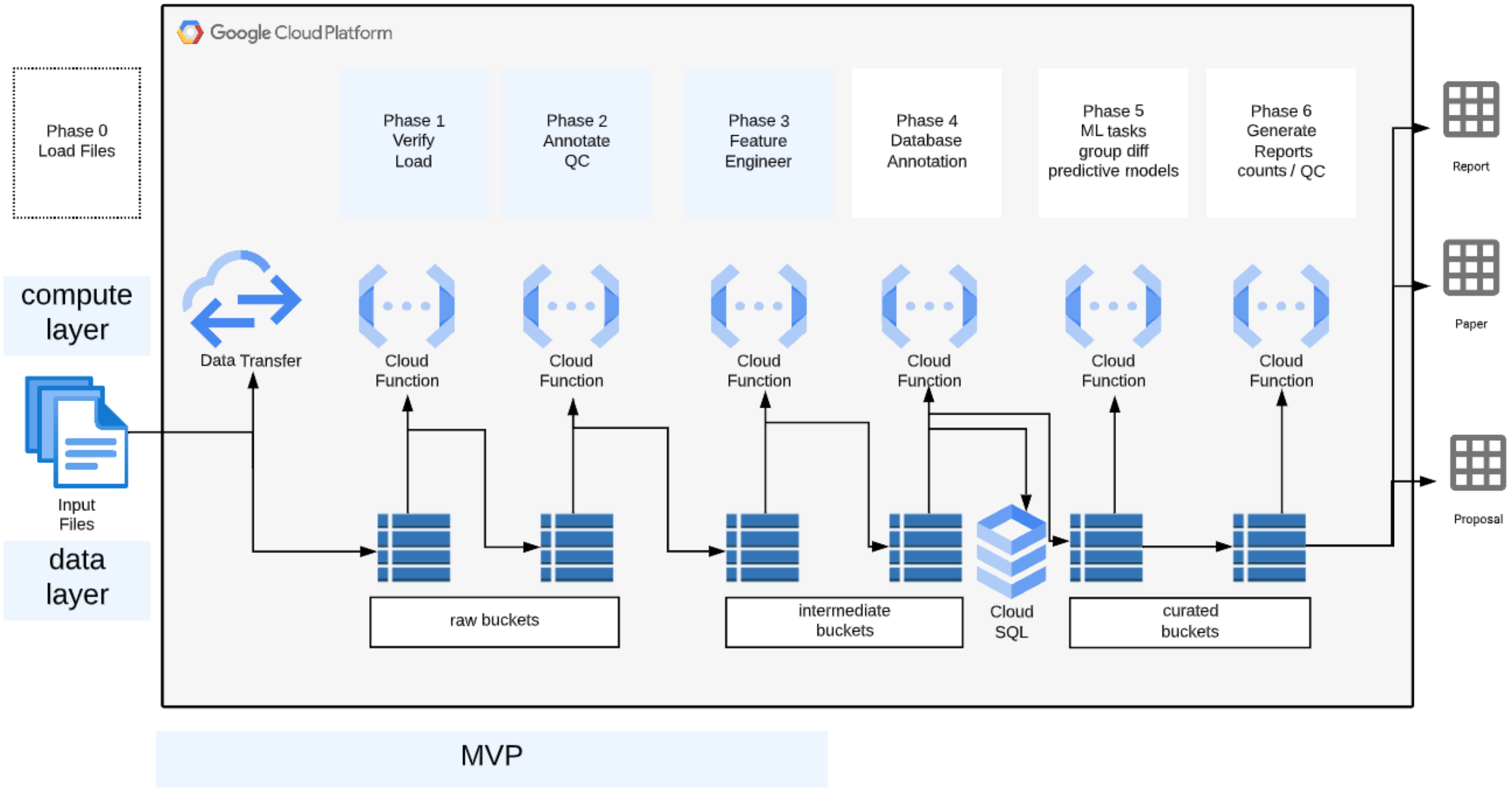
source


0 to ???
Emergency!

Pattern 2: Serverless Data Analysis Pipeline

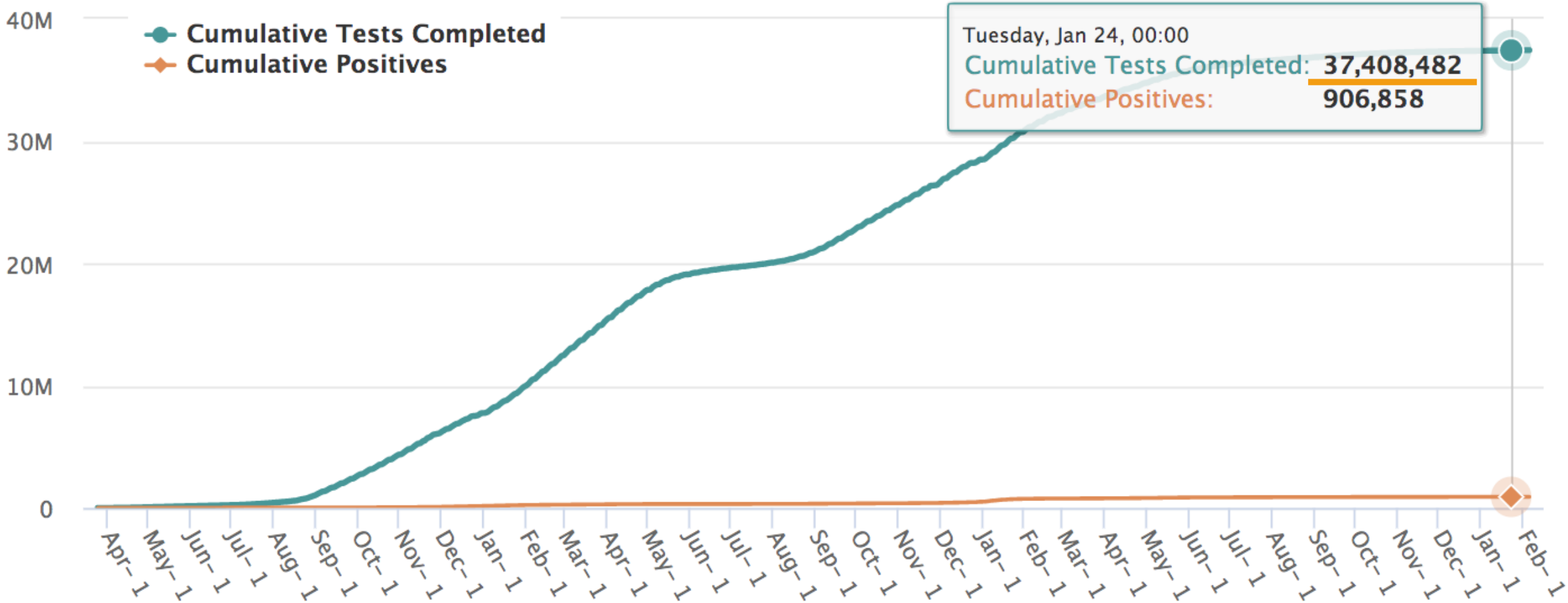


Serverless Pipeline

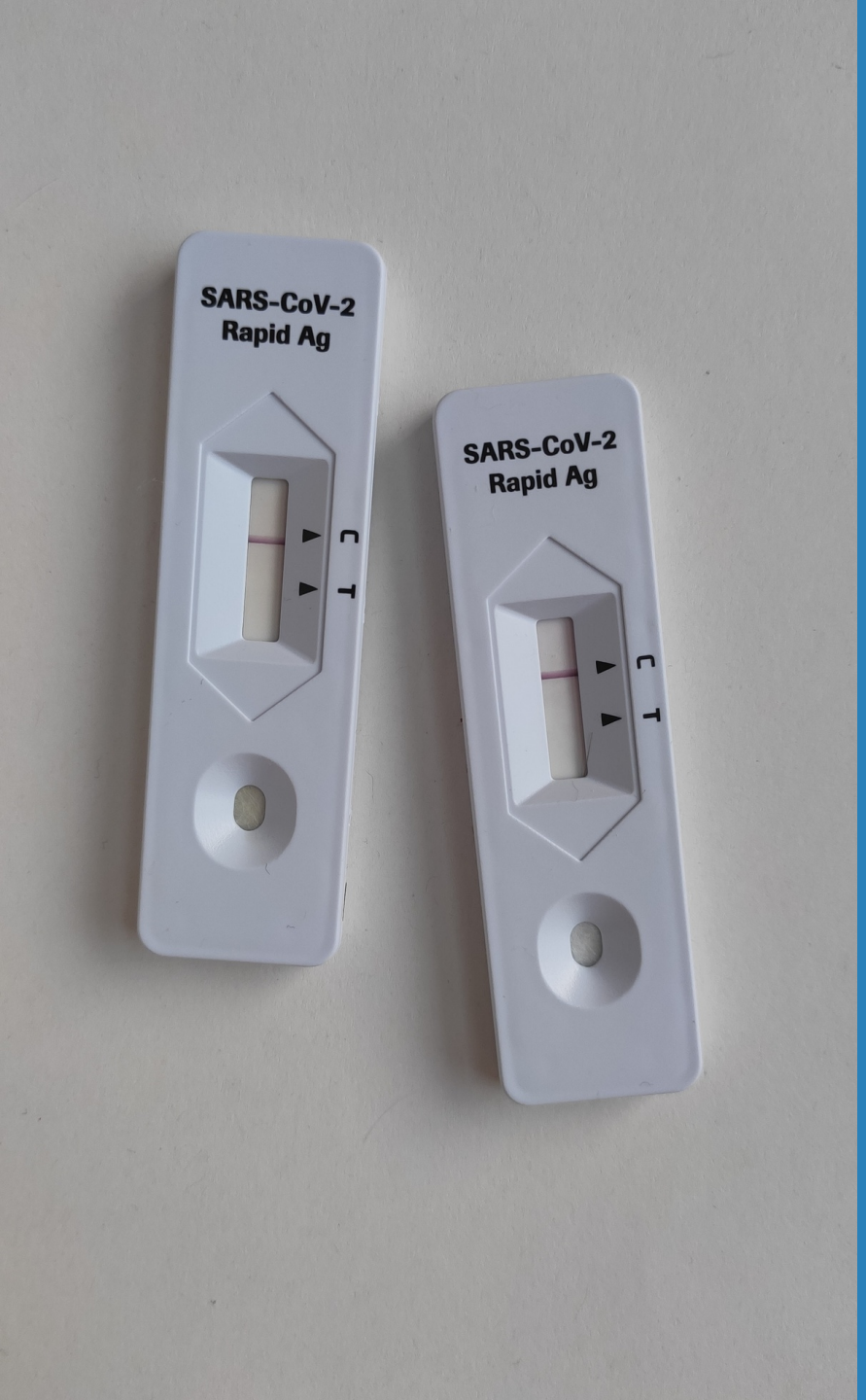


1W | 1M | 3M | 6M | All

Cumulative Volume



Source



Results

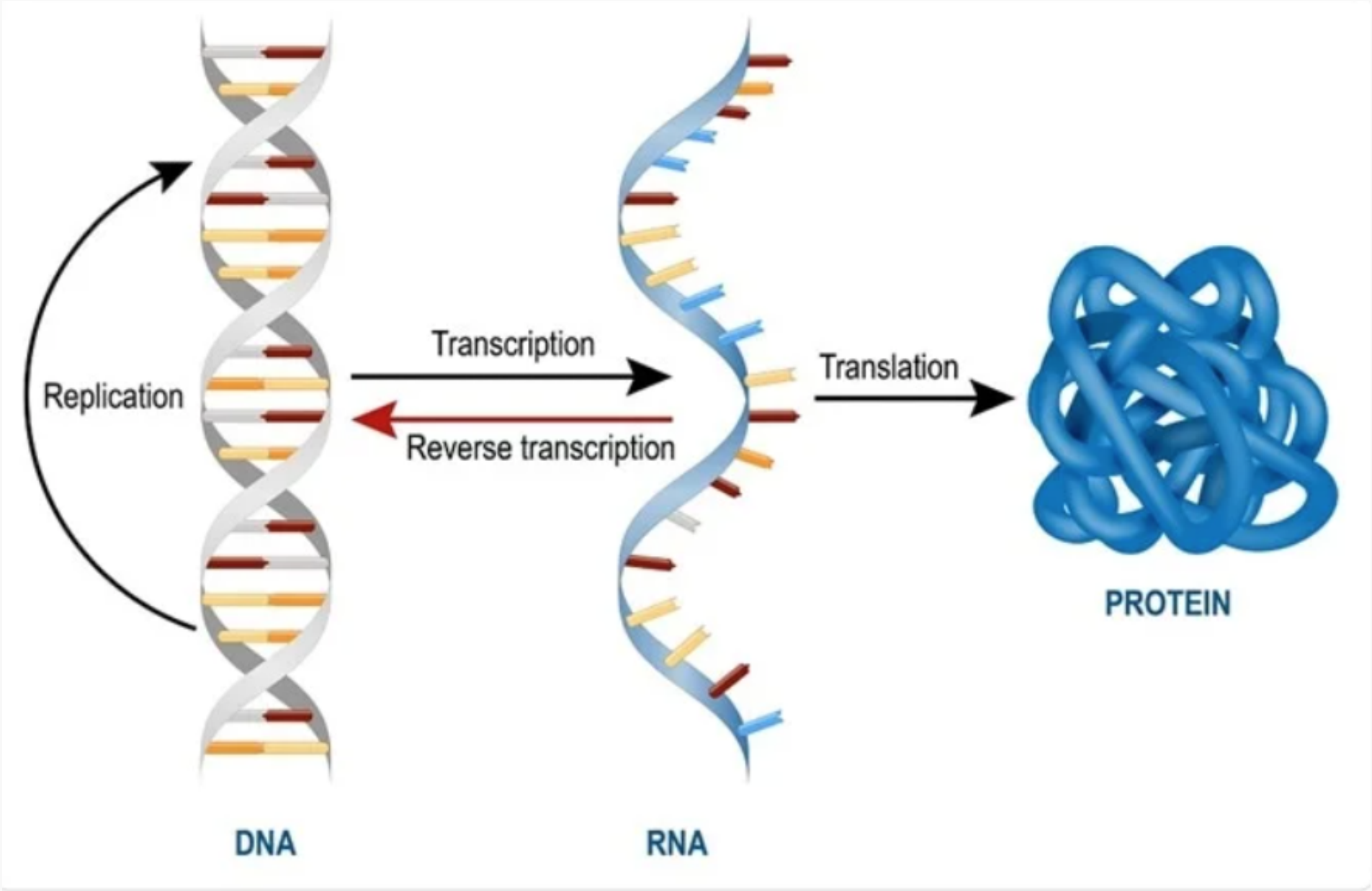
1. **Scalability (0-> 37 M Tests)!**
2. **Zero Down-time**

   Predictable Costs

   Serverless Pipeline

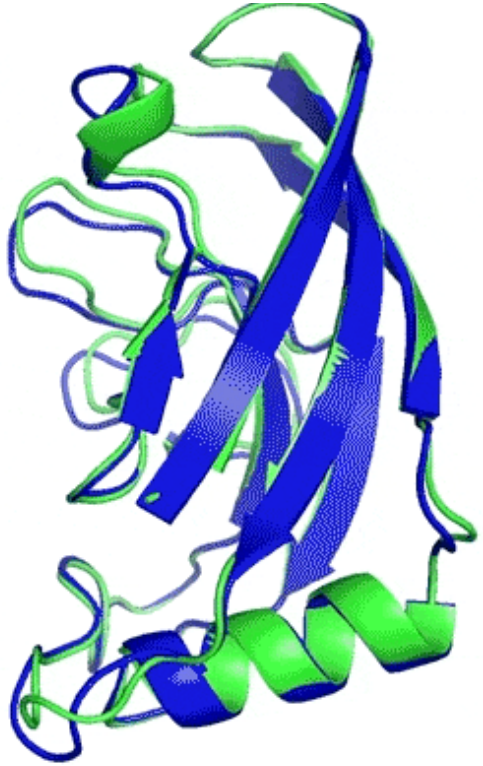
Source


Analyze
Everything!


Replication
Transcription
Reverse transcription
Translation
PROTEIN
DNA
RNA

source

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



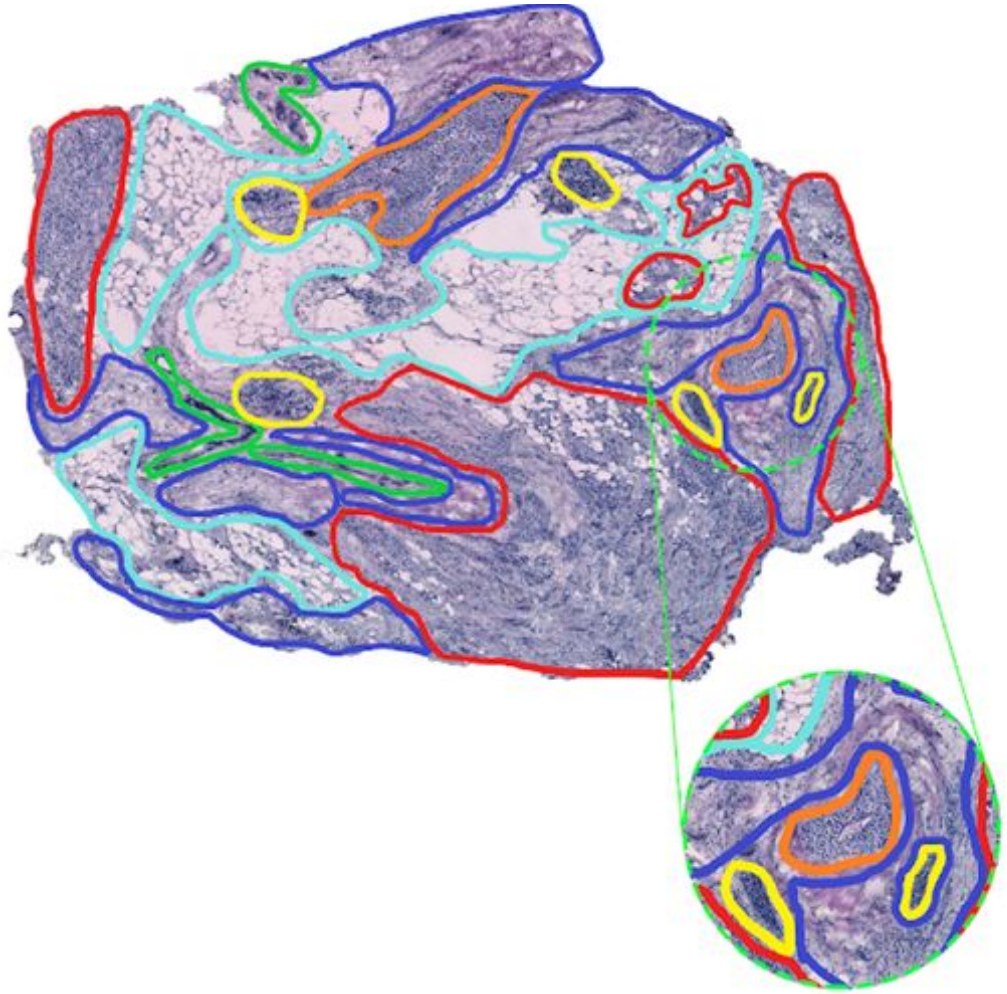
T1049 / 6y4f
93.3 GDT
(adhesin tip)

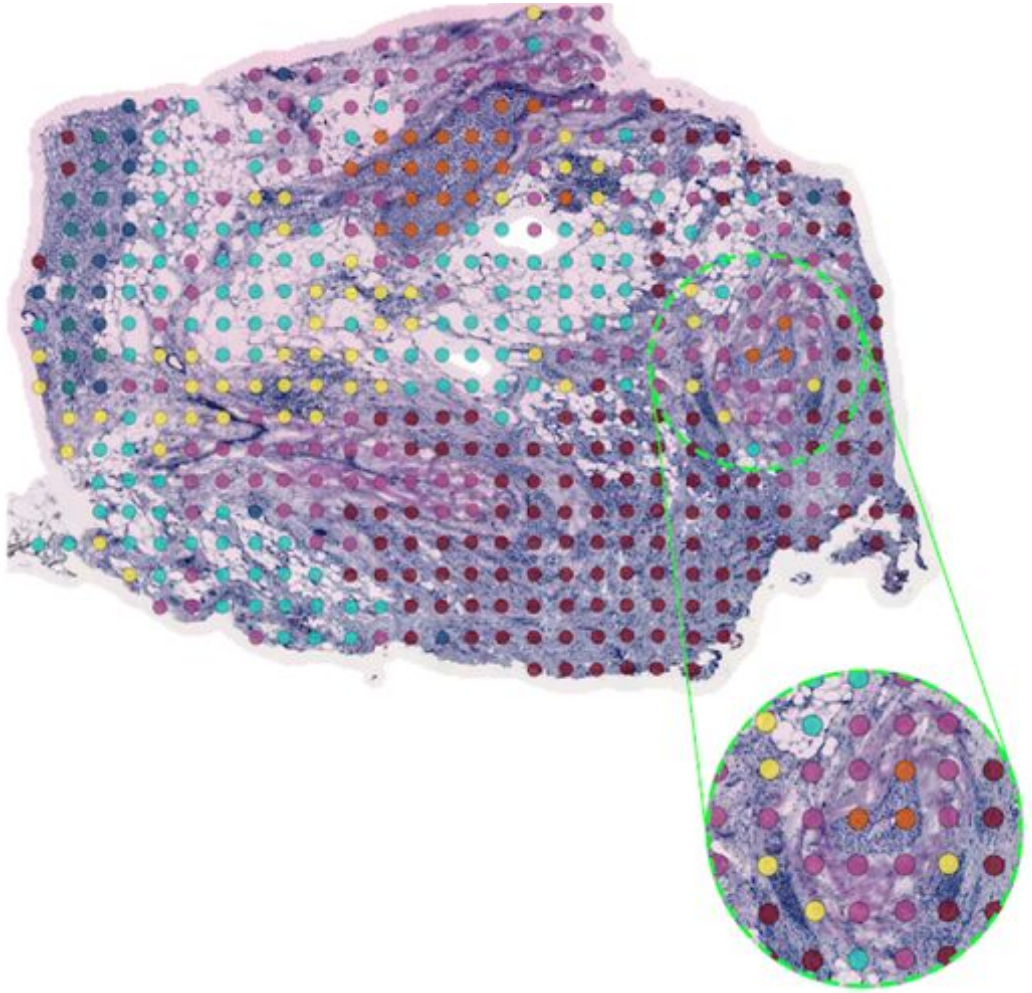
Experimental result

Computational prediction

source

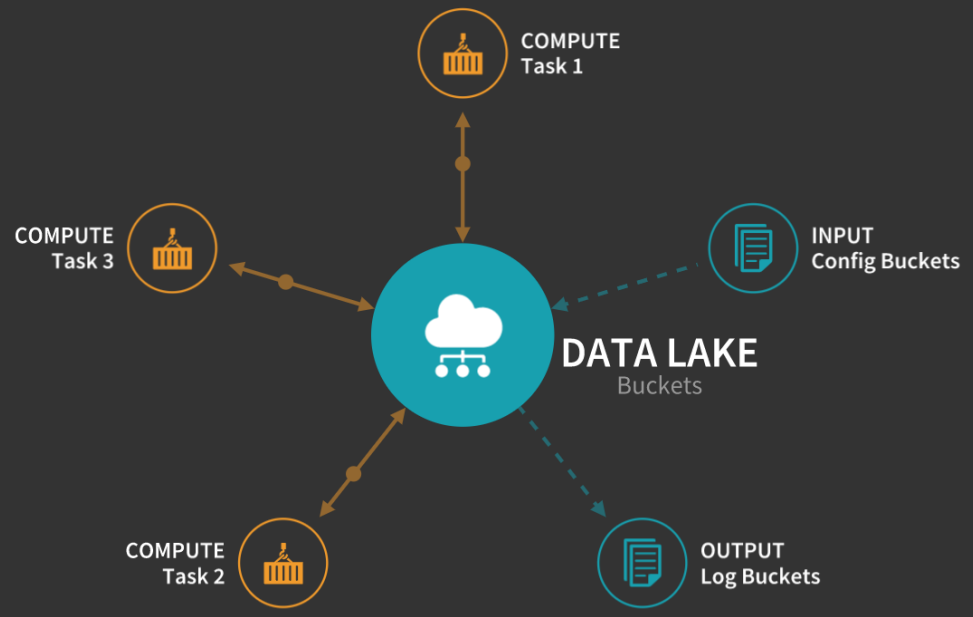
Spatial Multiomics





source

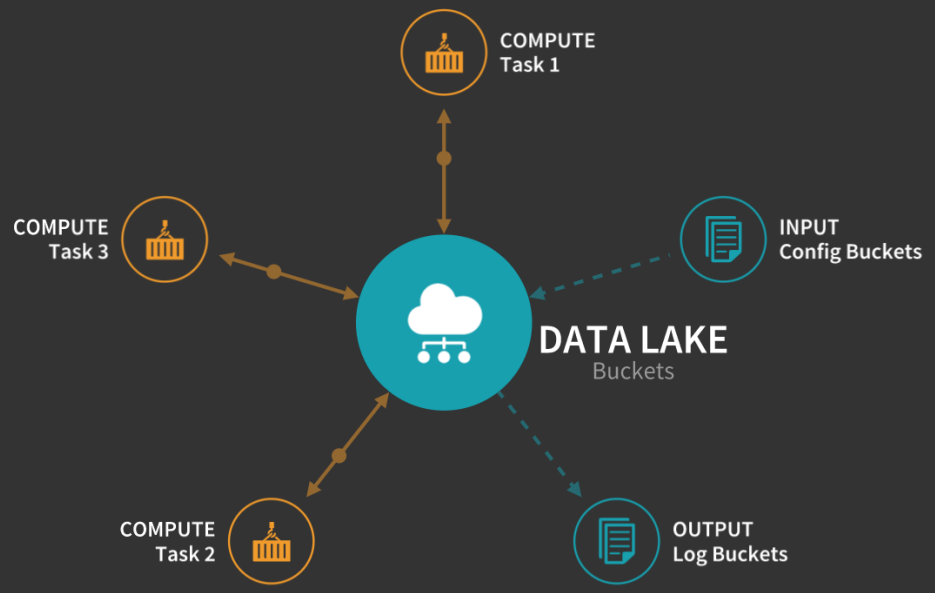
Pattern 3: Data Lakes/Data Pipelines and...



Research Lake

Source

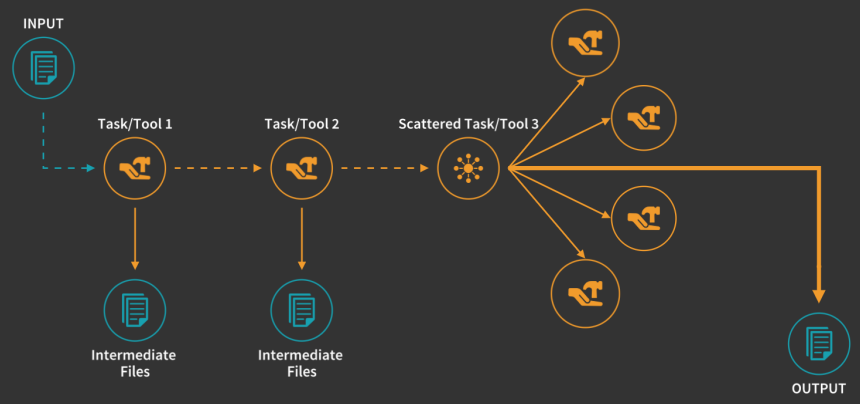
Pattern 3: Data Lakes/Data Pipelines and...



Clinical Lake

Source

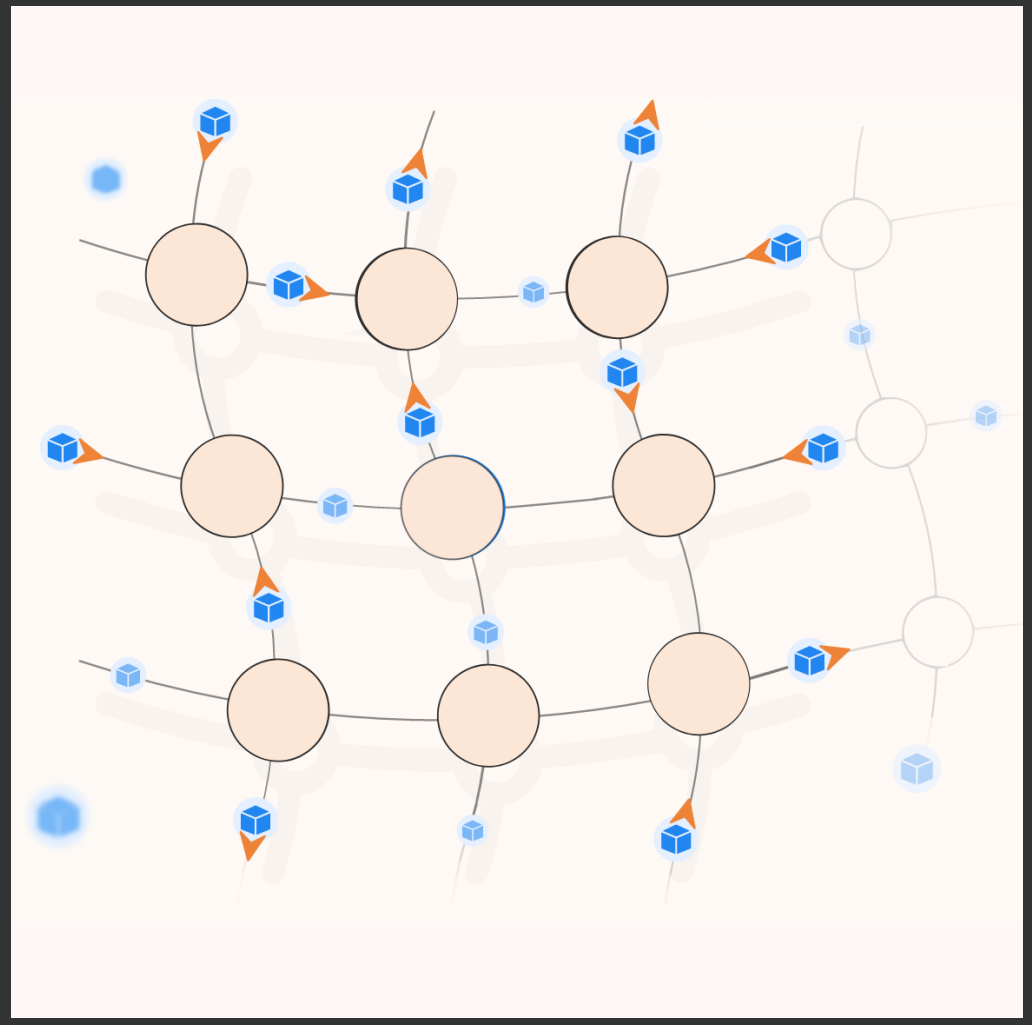
Pattern 3: Data Lakes/Data Pipelines and...



Pipelines

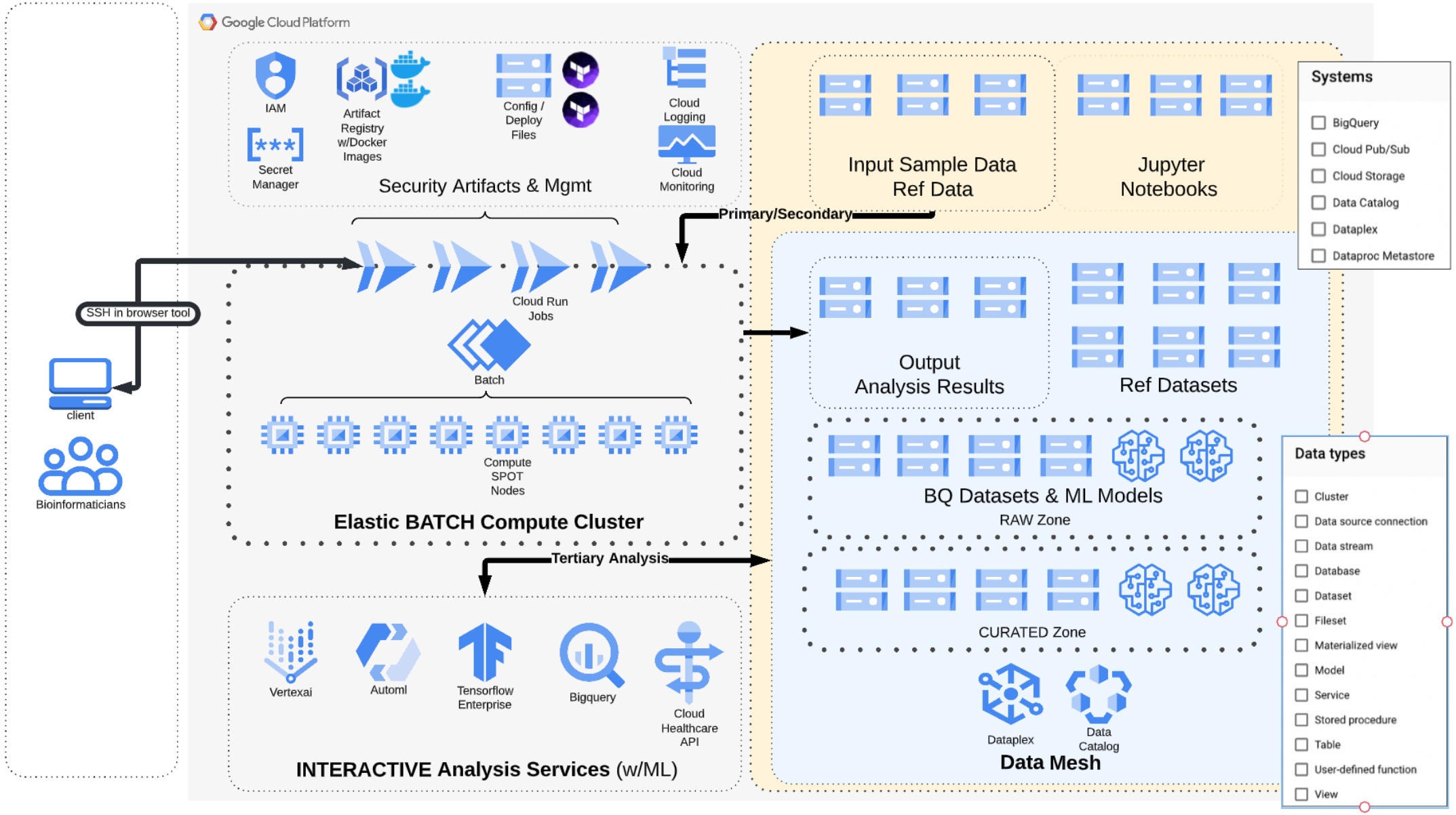
Source

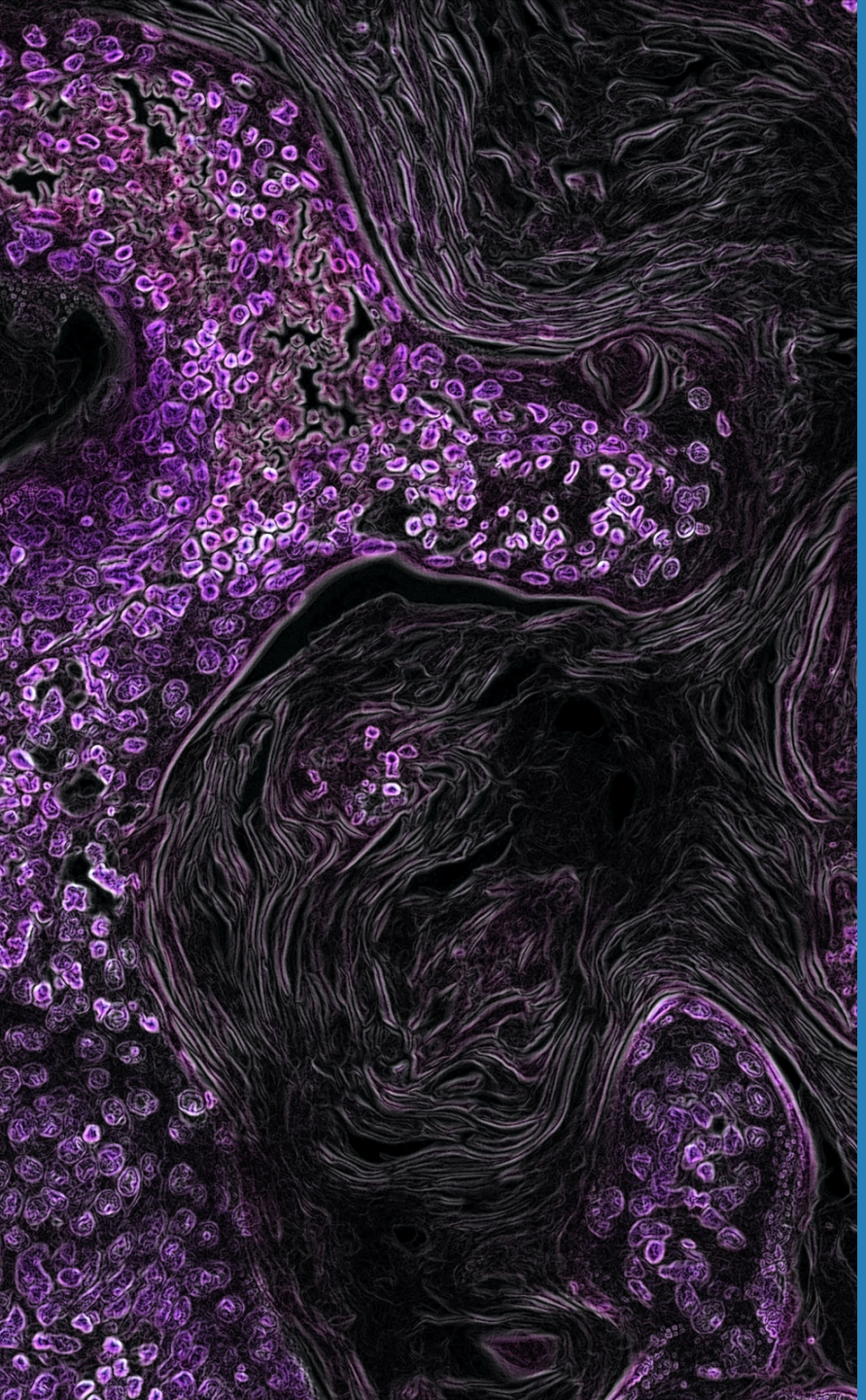
Pattern 3: Data Lakes/Data Pipelines and...



Data Mesh

Source


Google Cloud Platform
IAM
Secret Manager
Artifact Registry w/Docker Images
Config / Deploy Files
Cloud Logging
Cloud Monitoring
Security Artifacts & Mgmt
Input Sample Data
Ref Data
Jupyter Notebooks
Primary/Secondary
SSH in browser tool
client
Bioinformaticians
Cloud Run Jobs
Batch
Compute SPOT Nodes
Elastic BATCH Compute Cluster
Tertiary Analysis
Output
Analysis Results
Ref Datasets
BQ Datasets & ML Models
RAW Zone
CURATED Zone
Dataplex
Data Catalog
Data Mesh
Vertexai
Automl
Tensorflow Enterprise
Bigquery
Cloud Healthcare API
INTERACTIVE Analysis Services (w/ML)
Systems
BigQuery
Cloud Pub/Sub
Cloud Storage
Data Catalog
Dataplex
Dataproc Metastore
Data types
Cluster
Data source connection
Data stream
Database
Dataset
Fileset
Materialized view
Model
Service
Stored procedure
Table
User-defined function
View



Results - Space + Speed

1 **File Storage**

Data Lakes

Grouping Lakes into DataMesh

2 **Experiments / Analytics**

Serverless Pipelines

2 weeks from idea to deployment

Source

Learn on GitHub

Notes, Code and Videos

- TeamTeri Repo
- GCP-for-Bioinformatics Repo
- AWS-for-Bioinformatics Repo
- Learn Quantum Programming Repo
- Learn Apache Spark Repo