Data Management Needs in Cancer Research

Arfath Pasha
MSKCC, March 2020

Computational Oncology at MSK

- Brings together computer science, machine learning and data engineering to advance the major questions in oncology, cancer biology and clinical care

https://componcmsk.org/

Data

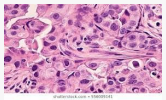
Studies

lung
ovary
breast
colorectal
...

Genomics

single cell whole genome DNA
single cell RNA
single cell TCR
bulk whole genome DNA
cell free DNA

Pathology



hematoxylin and eosin (H&E)
multiplexed imaging (mpIF)

Radiology

CT scans
MRI scans

Clinical

electronic medical records
labs, treatments, diagnoses

Cohorts

500-600 patients per study

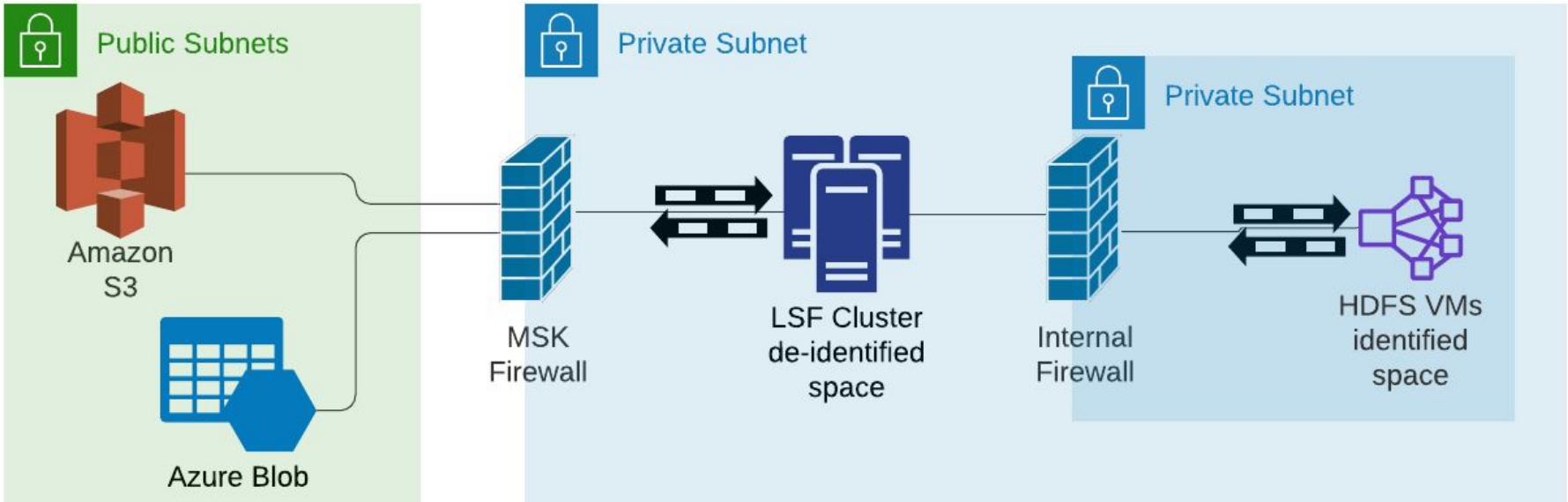
100-200Tb per cohort

Time
~ 5 years per study

Data Management Needs

Hardware Infrastructure:

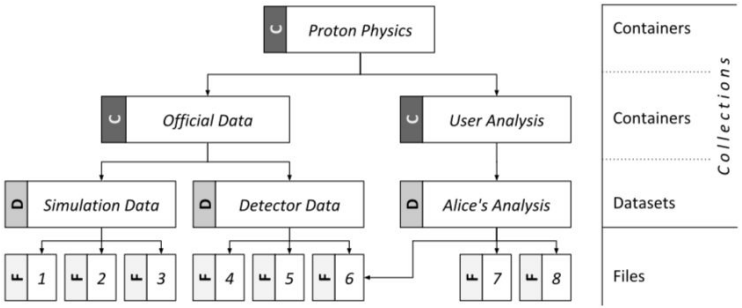
- manage data across on-premises and cloud infrastructures
- can push/pull data from inside private subnets but not from outside



Data Management Needs

fine grained permissions:

- satisfy HIPAA constraints
- researchers naturally want to protect data until publication

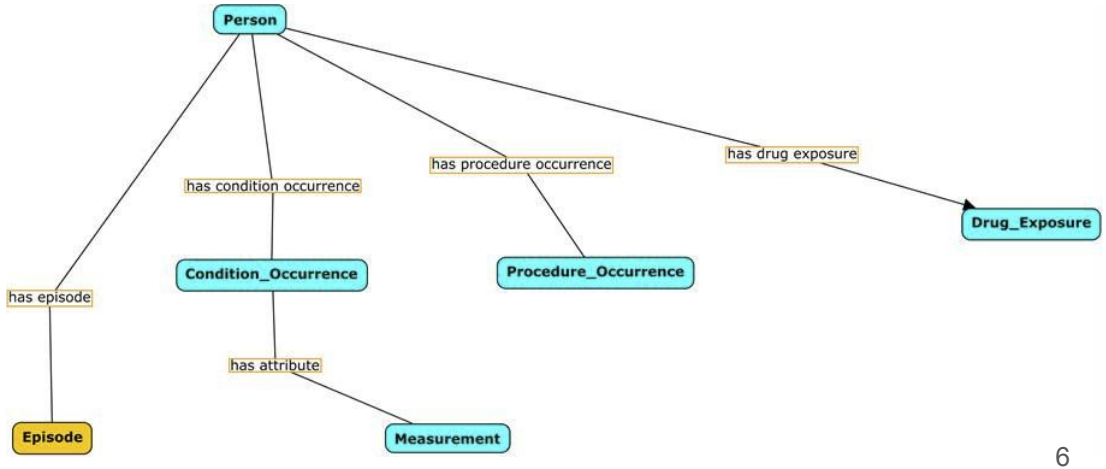


Barisits M. et al. Rucio: Scientific Data Management. Computing and Software for Big Science (2019) 3:11

Data Management Needs

Data Evolution:

- multi-year data collection effort
- complex datasets and relationships
- a push towards standardization

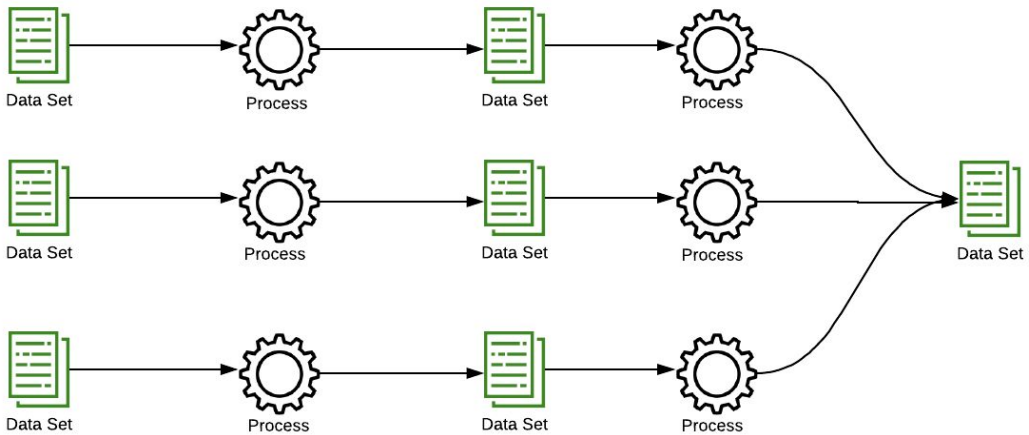


https://www.ohdsi.org/data-standardization/

Data Management Needs

Data Provenance:

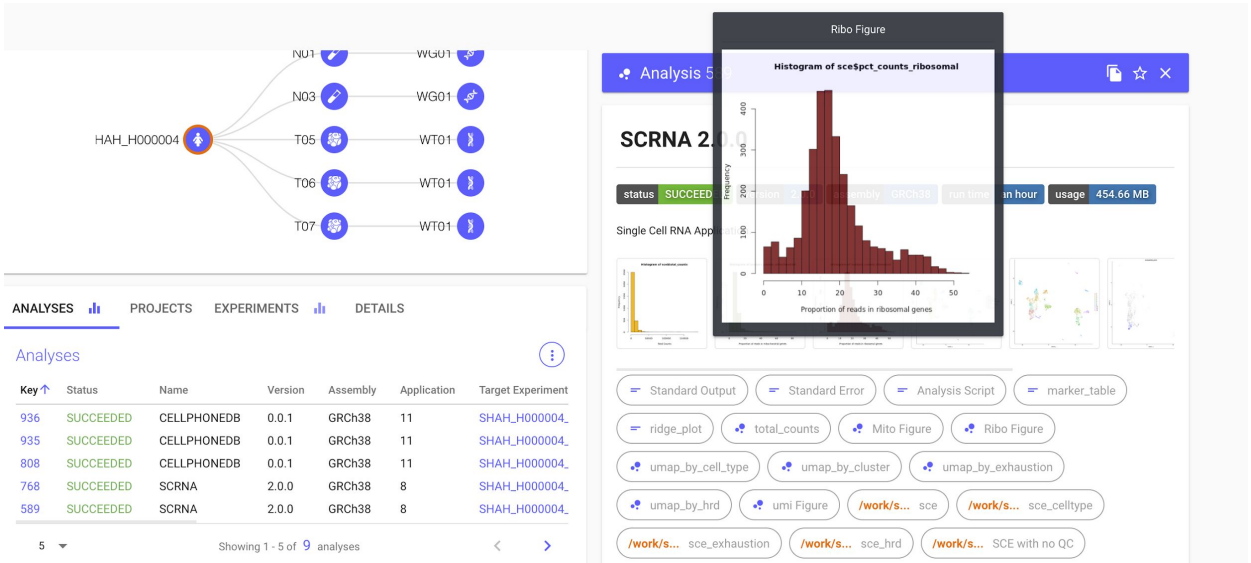
- track lineage of data generated from pipelines
- allow for reproducibility



Data Management Needs

User Interface:

- spot check results



https://isabl.io/

Roadmap

2019: Launched our first study on high grade serous ovarian cancer using existing MSK resources

2020: Prototype a new system using open source technologies

2021: Productionize the system