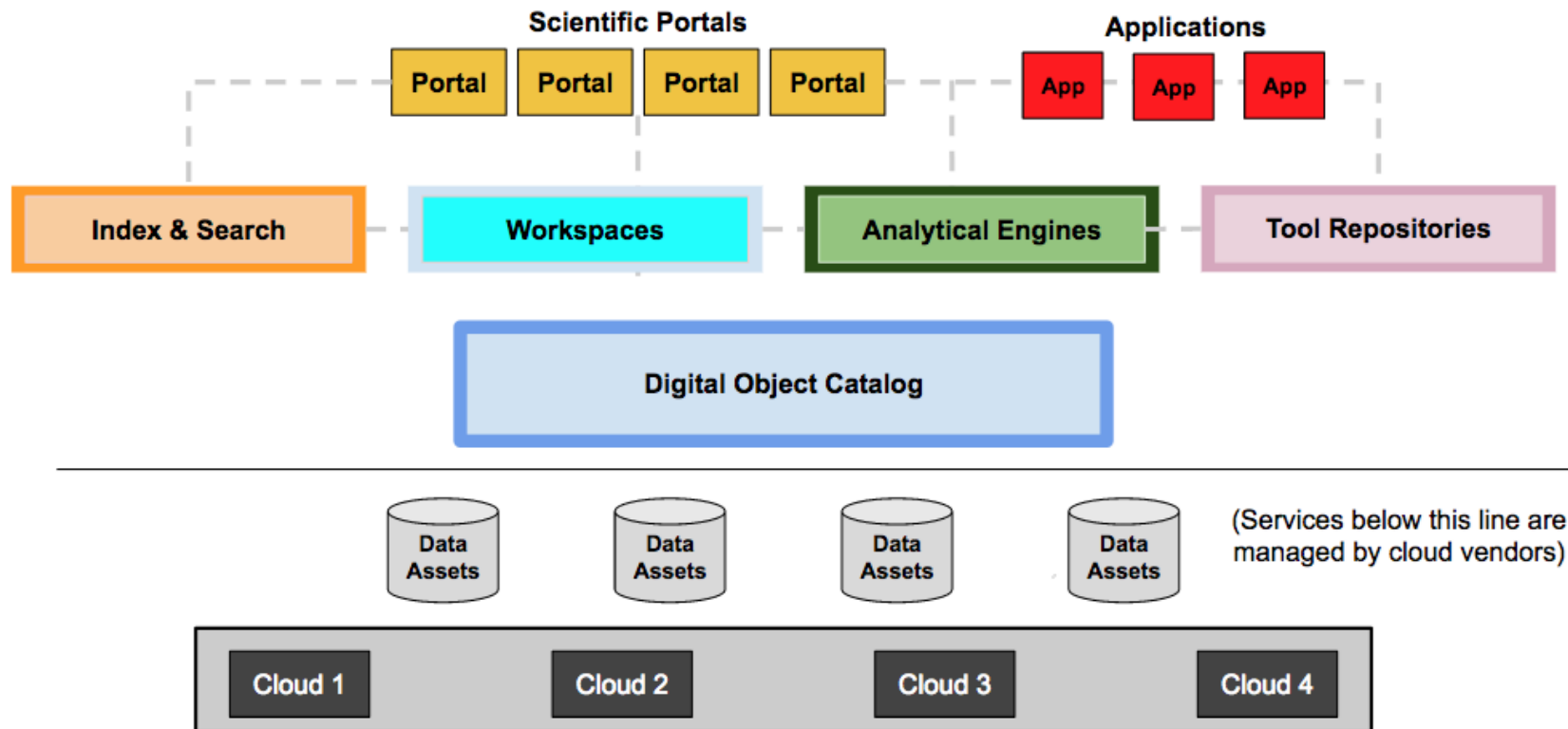# Terra Architecture and Design Overview

March 2020

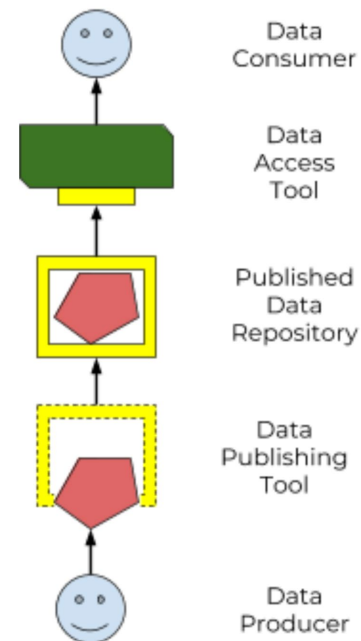Alex Baumann - abaumann@broadinstitute.org

BROAD INSTITUTE

Driven by Data Biosphere vision to enable an ecosystem of interoperable components across orgs

GA4GH standards used wherever possible

- Tool Repository Service (TRS)
  - enable libraries of pre-packaged shared methods

- Workflow Execution Service (WES)
  - federated analysis: run reproducible workflows

- Data Repository Service (DRS)
  - federated storage: access data, wherever it's stored

- Data Use (DURI)
  - data passport: match access rules to researchers

- Researcher Identity (DURI)
  - data passport: make researcher credentials portable

- and more …

Data Consumer

Data Access Tool

Published Data Repository

Data Publishing Tool

Data Producer

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Terra's architecture primarily serves 3 roles

Biomedical Researchers

Data Generators

| All of Us Workbench | Community Workbench | Single Cell Portal | Custom Apps & Portals |

Tool Developers

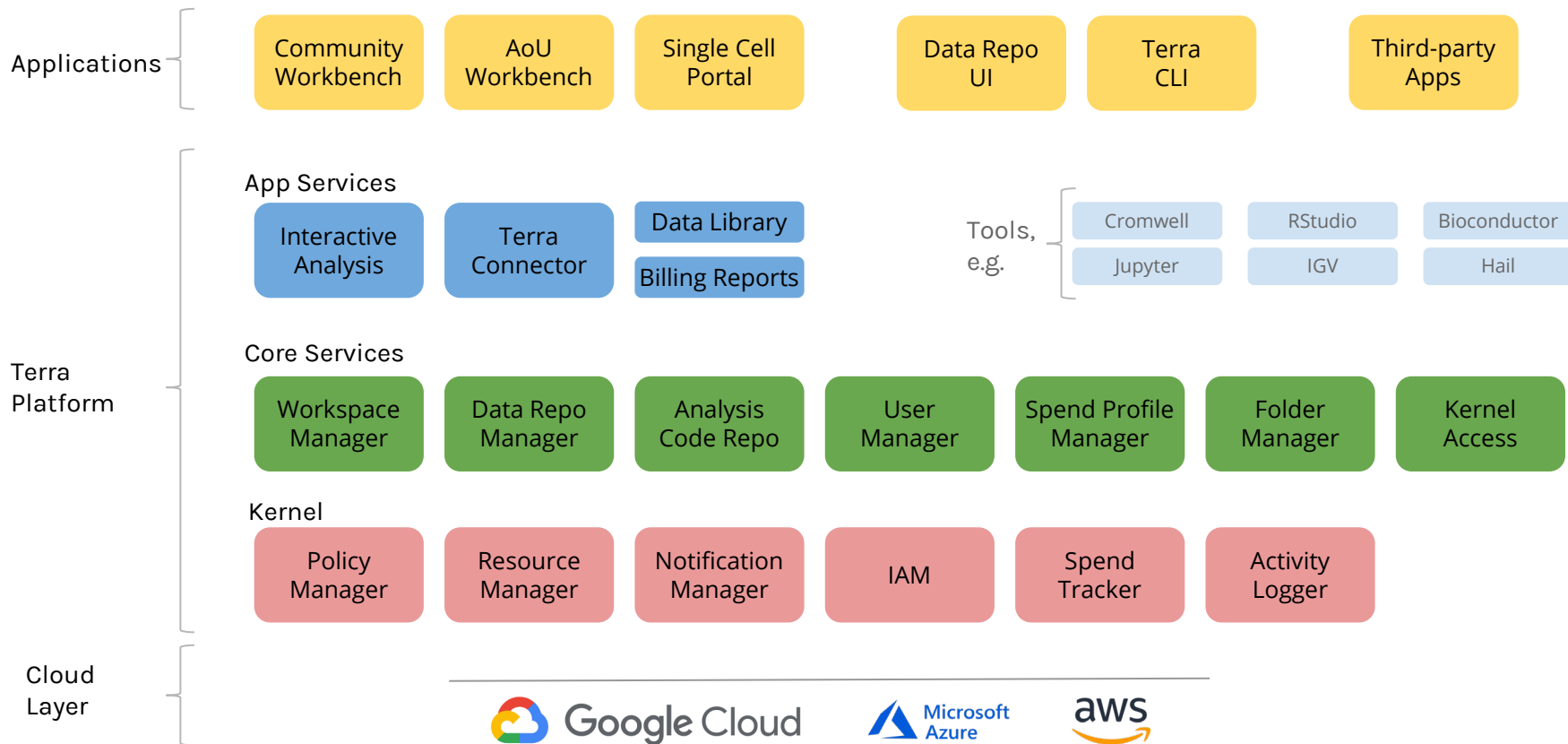Data Production & Curation

Terra Platform API

| Data Mgmt | Workspaces | Tools Mgmt |

Tool Creation & Publication

Cloud Services

Rearchitecture in progress to support multiple clouds and integration of new resources

Applications

| Community Workbench | AoU Workbench | Single Cell Portal | | Data Repo UI | Terra CLI | | Third-party Apps |

Terra Platform

App Services

| Interactive Analysis | Terra Connector | Data Library |
| | | Billing Reports |

Tools, e.g.

| Cromwell | RStudio | Bioconductor |
| Jupyter | IGV | Hail |

Core Services

| Workspace Manager | Data Repo Manager | Analysis Code Repo | User Manager | Spend Profile Manager | Folder Manager | Kernel Access |

Kernel

| Policy Manager | Resource Manager | Notification Manager | IAM | Spend Tracker | Activity Logger |

Cloud Layer

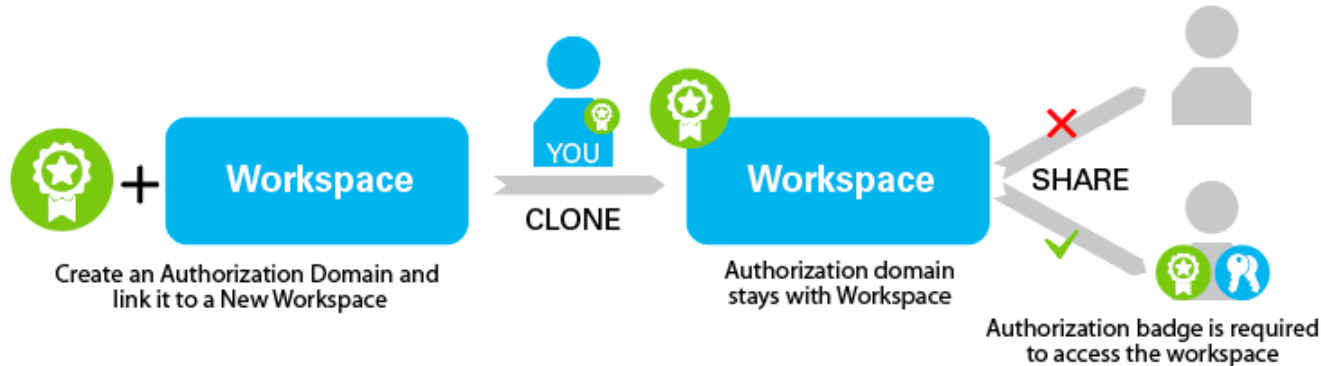Google Cloud   Microsoft Azure   aws

Many datasets co-analyzable according to research use

Design principles and considerations

- Not involved in billing now - user brings billing, we manage resources
- Scale is critical (per month: ~2k active users, ~500 users running workflows and notebooks, ~500k workflows per month, >10m CPU hours per month)

- Provenance is critical to reproducible data science

- Authorization Domains to protect derived results

Create an Authorization Domain and link it to a New Workspace

YOU CLONE

Authorization domain stays with Workspace

SHARE

Authorization badge is required to access the workspace

Questions?