A data biosphere for all of us (and *All of Us*)

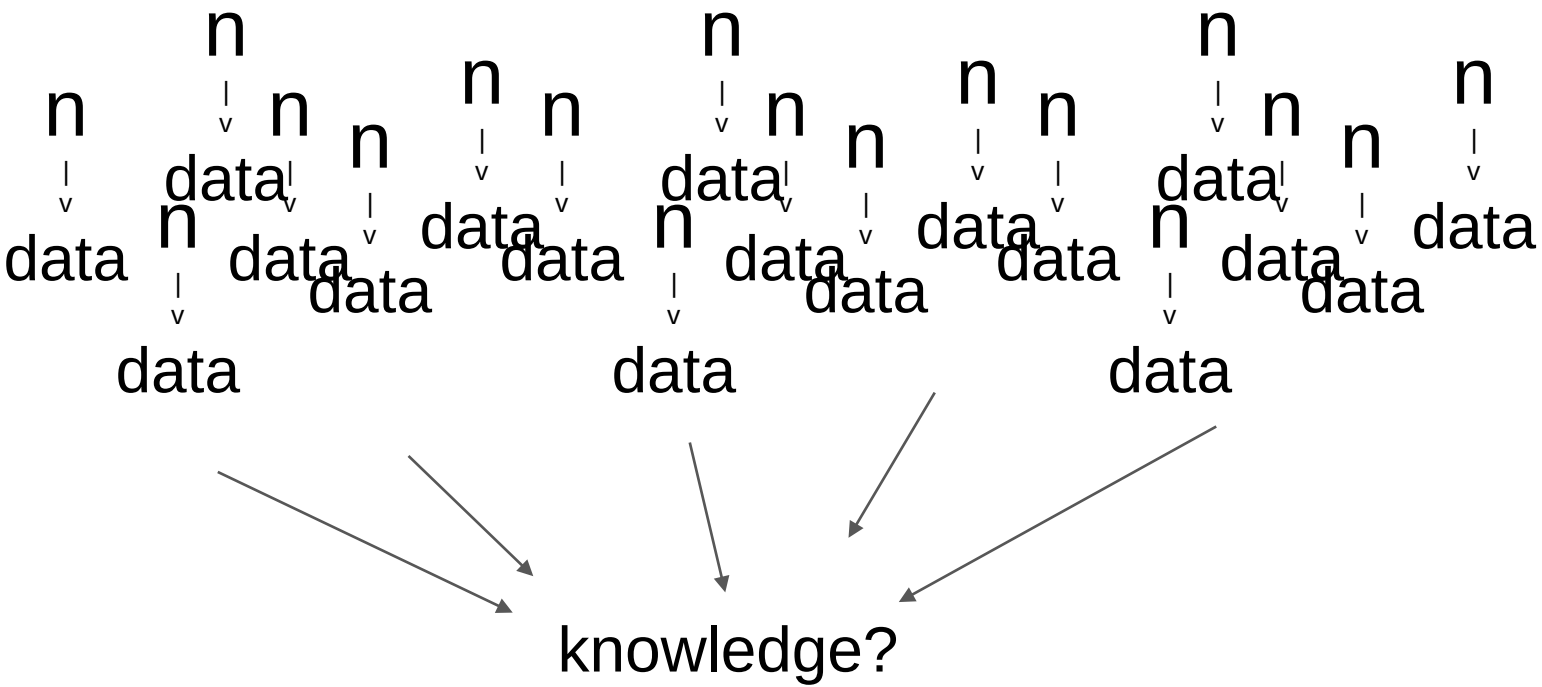
David Glazer, Verily

XLDB, Stanford, April 30, 2018

1. Why?
2. What?
3. How?
4. e.g.

Q: Why?
A #1: idealism

David Glazer
@GA4GH



David Glazer

@GA4GH

n+n+n+n+n+n+n+n+n+n+n+n+n+n+n+n+n+n+n+n+n=#bigN

#bigN*data=#bigData

#bigData+#dataScience=#bigKnowledge

#bigKnowledge*#bigN=#bigHealth

#GA4GH2017

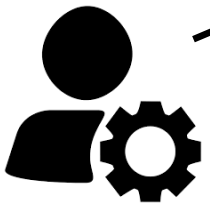
8:41 AM - 17 Oct 2017

Tweet your reply

Q: Why?
A #2: pragmatism

Human Cell Atlas Needs

- Generators upload to cloud-based data store
- Process data with analysis engine, and curate metadata
- Enable search & discovery
- Ecosystem of applications



Data Generators

HCA

Explore
Store
Ingest

Portals

Analysis Engine | Methods Repo | Work-Spaces

Use in cloud

Download

Researchers

HCA Team:

- Broad
- UCSC
- EBI
- CZI

NCI Cloud Resources Needs

- Generators upload to cloud-based data store
- Process data with analysis engine, and curate metadata
- Enable search & discovery
- Ecosystem of applications

Researchers

Portals

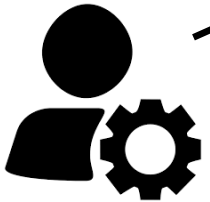
Analysis Engine | Methods Repo | Work-Spaces

Explore | Store | Ingest — GDC

Explore | Store | Ingest — HCA

Use in cloud

Download



Data Generators

NCI Team:

- Broad
- SBG
- ISB
- U. Chicago

***All of Us* Needs**

- Generators upload to cloud-based data store
- Process data with analysis engine, and curate metadata
- Enable search & discovery
- Ecosystem of applications

Researchers

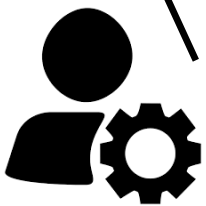
Portals

Analysis Engine | Methods Repo | Work-Spaces

Explore / Store / Ingest

AoU | GDC | HCA

Use in cloud



Data Generators

AoU Team:

- Broad
- Verily
- Vanderbilt
- U. Michigan
- Columbia

1 Why?

2.What?

3.How?

4.e.g.

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 working together in overlapping combinations during the past two years to drive the creation of ***data commons to support flagship scientific initiatives****. This document lays out our* ***evolving vision*** *for the next steps in that journey. Our hope is that others will join the effort to build momentum for an* ***open, compatible, and secure*** *approach to data within the larger research community. We welcome your feedback, and look forward to continuing this journey together.*

-- Josh Denny (Vanderbilt), David Glazer (Verily Life Sciences), Robert L. Grossman (University of Chicago), Benedict Paten (University of California at Santa Cruz), Anthony Philippakis (Broad Institute)

DataBiosphere.org

principles

code

services

Data Biosphere: Principles

1

modular, composed of functional components with well-specified interfaces

2

community-driven**,** created by many groups to foster a diversity of ideas

3

open**,** developed under open-source licenses that enable extensibility and reuse, with users able to add custom, proprietary modules as needed

4

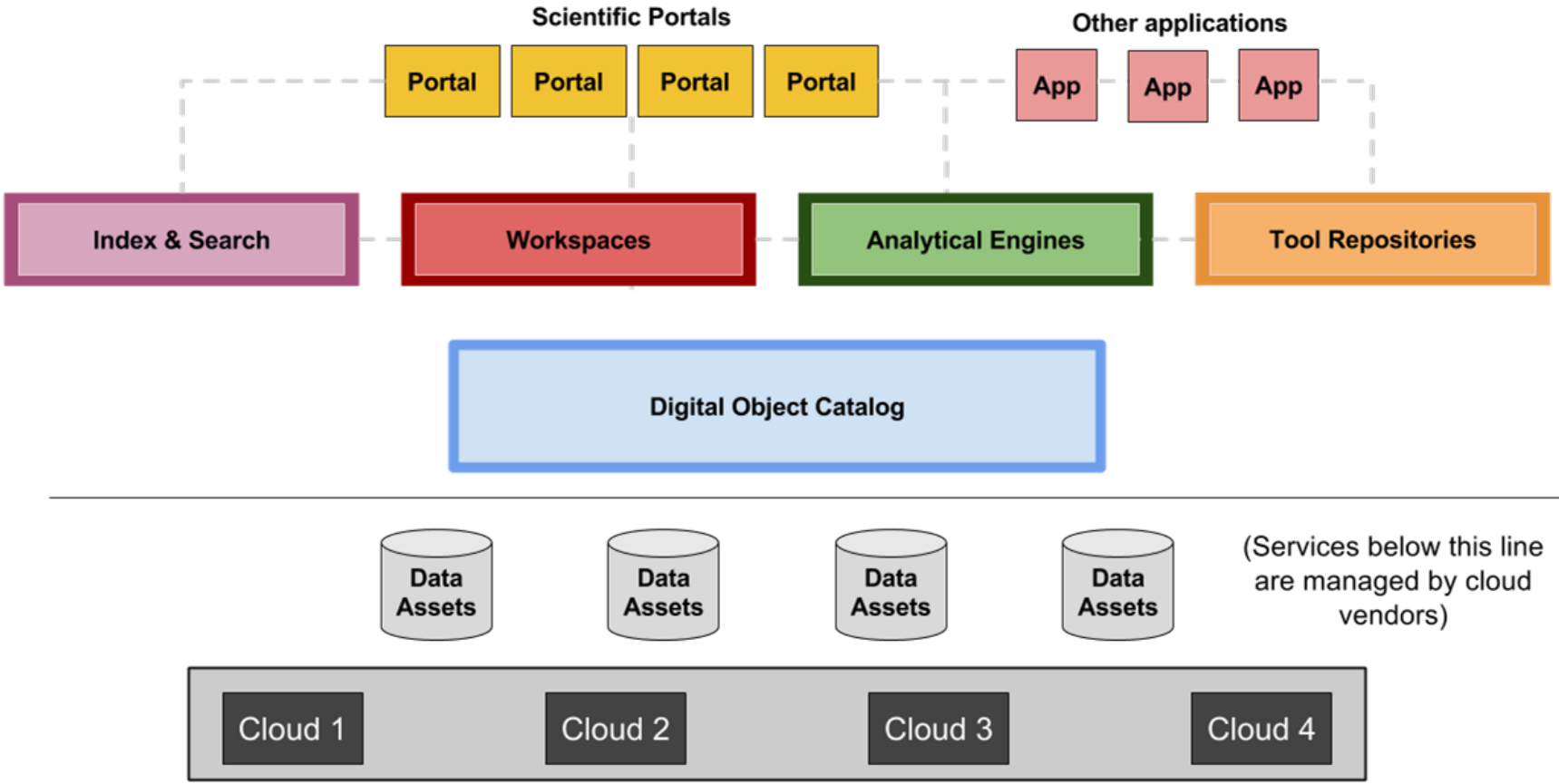
standards-based, consistent with standards developed by coalitions such as the GA4GH

principles

code

services

Data Biosphere: Modular Components

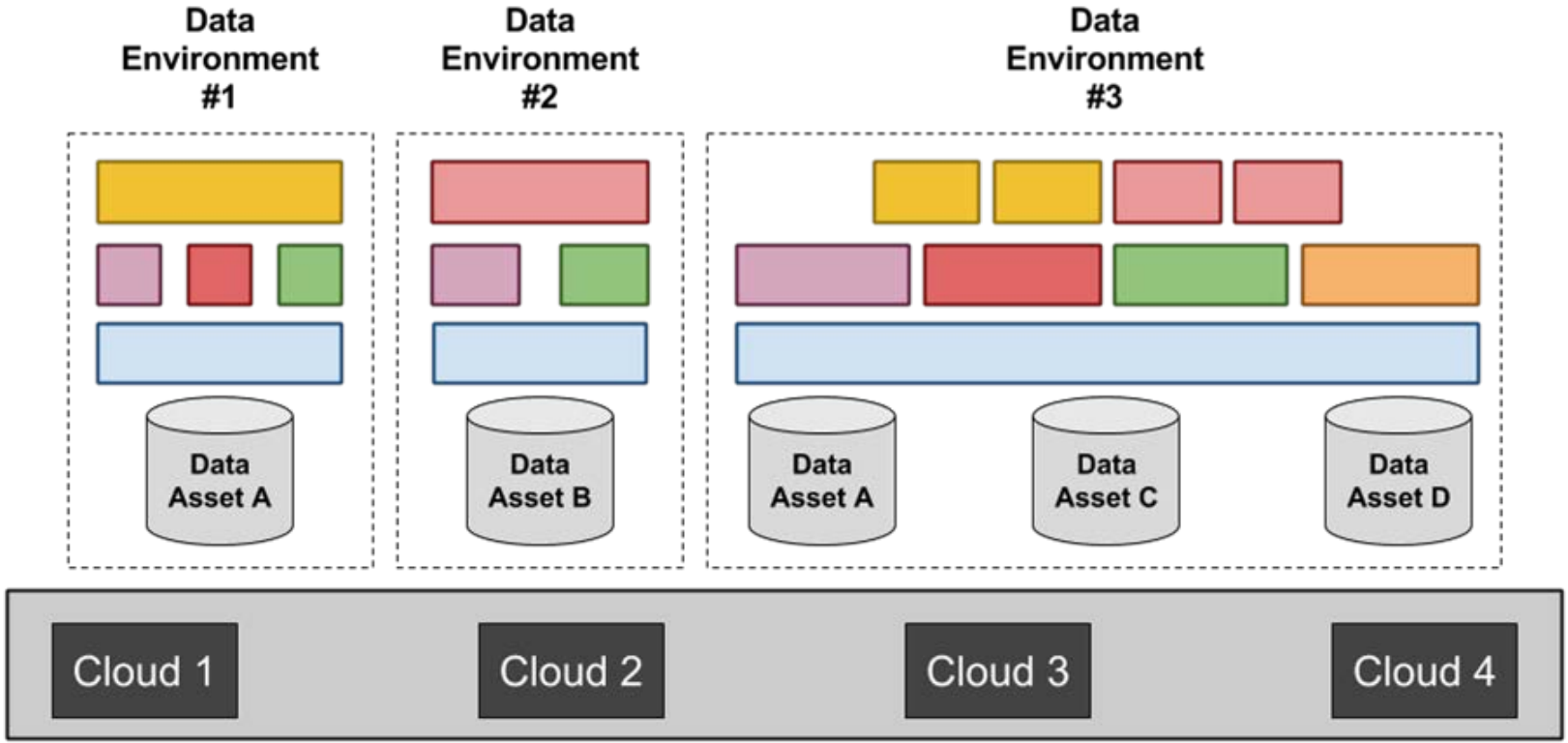


principles

code

services

Data Biosphere: Data Environments



Data Biosphere: Path Forward

Building a Data Biosphere to propel progress in biomedicine will require a **community working together**, including laboratory groups generating data, software developers creating Biosphere Components, and technical teams assembling and operating Data Environments.

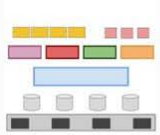
The vision described here will surely evolve with experience, but the fundamental principles provide a solid foundation for **using information to improve human health**. Together with the global community, we will **make the vision real**.

1 Why?

2.What?

3.How?

4.e.g.



Data Biosphere

Repositories 30 | People 67 | Teams 14 | Projects 0 | Settings

Search repositories… | Type: All | Language: All | Customize pinned repositories | New

leonardo

Notebook service

Scala ★ 4 ⑂ 4 Updated 7 hours ago

bagit-firecloud-lambda

Upload TOPMed metadata into a FireCloud workspace

BSD-3-Clause Updated a day ago

dos-resolver-lambda

Top languages

Python Jupyter Notebook JavaScript Java CSS

People 67 ›

github.com/DataBiosphere

Data Biosphere: Work in Progress

OSS components include:

- `boardwalk, data-explorer` -- faceted search data browsing, for exploration and building sub-cohorts
- `leonardo` -- service for launching and managing Jupyter notebooks to analyze data from a project
- `cromwell, toil` -- services for running pre-defined batch workflows
- `job-manager` -- UI for managing asynchronous batch workflows

… and more

1 Why?

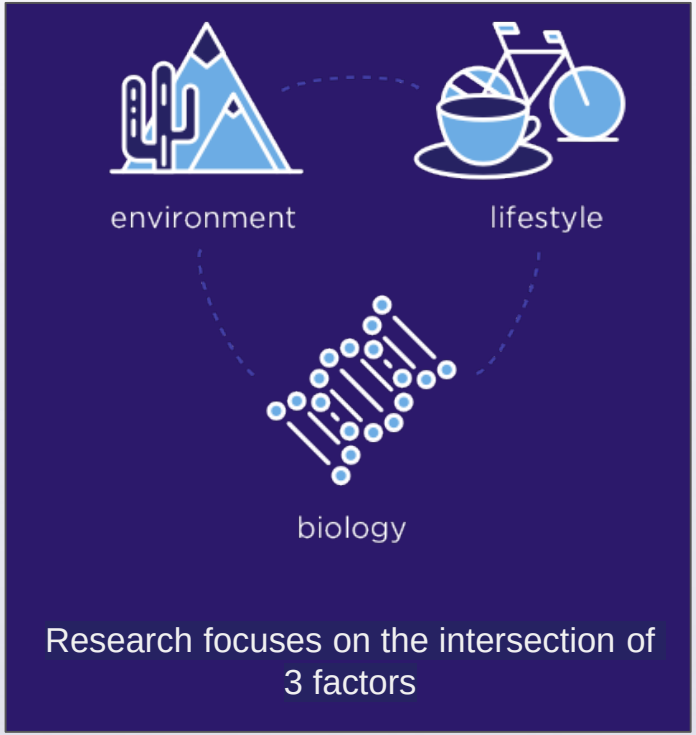
2.What?

3.How?

4.e.g.

All of Us Research Program

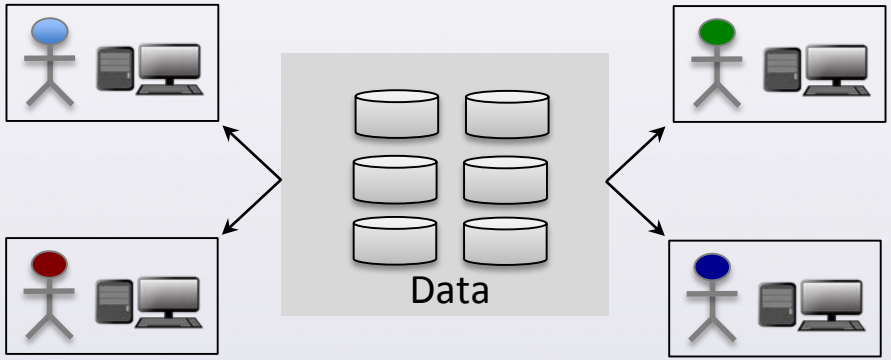
- Engage **1,000,000 or more** U.S. research participants
- Share **tissue** samples, **genetic** data, **lifestyle** information, electronic **health records**
- Pioneer a new model of research that emphasizes **engaged research participants**, **responsible data sharing**, and **privacy protection**
- Alpha launched May 2017; **national launch Spring 2018**; researcher launch 2019



Centralized data to **enhance security** and **improve usefulness**

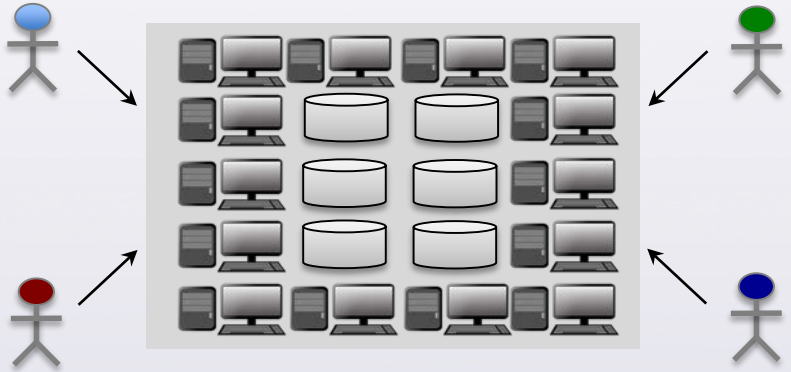
Traditional Approach

Bring data to researchers



AoU Approach

Bring researchers to the data



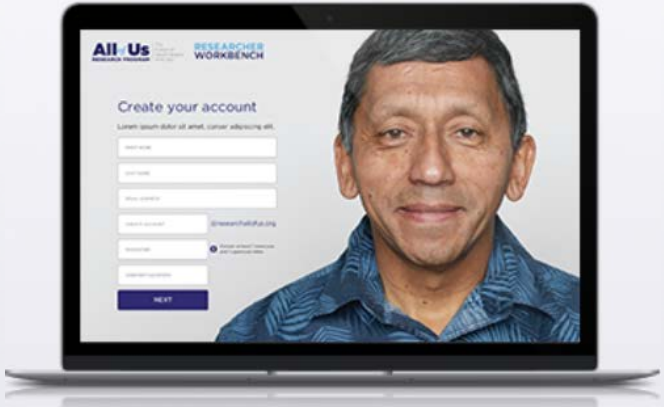
Problems

- *Data sharing = data copying*
- *Decreased security (data lots of places)*
- *Huge infrastructure needed*
- *Encourages siloed research*

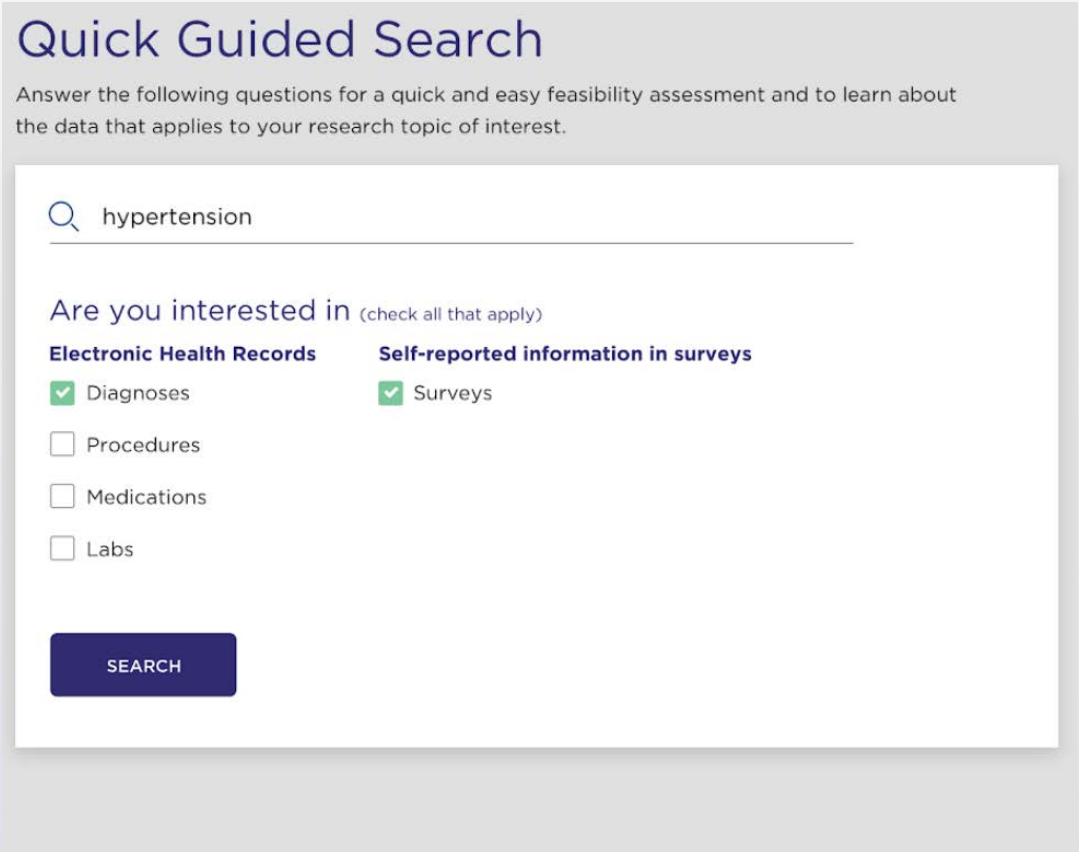
Advantages

- *Improved security and auditing*
- *Increased accessibility to researchers*
- *Shared compute*
- *Facilitates collaboration*

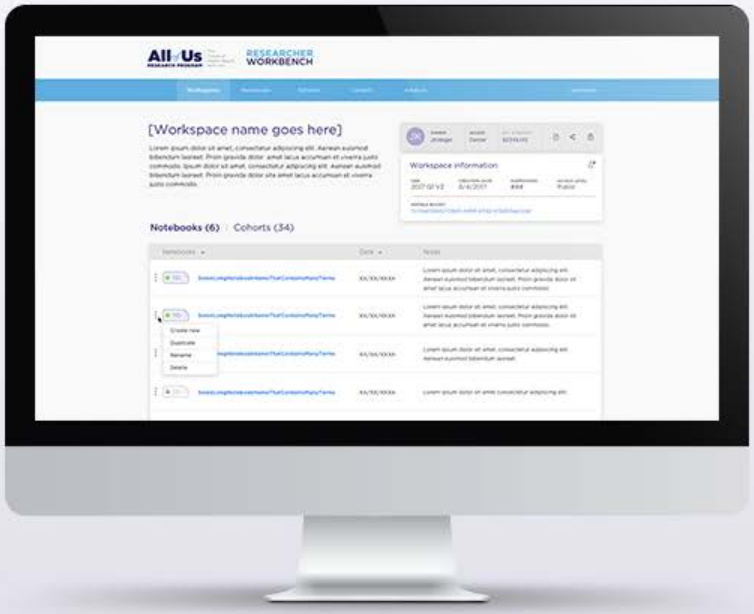
Registration [wireframes]

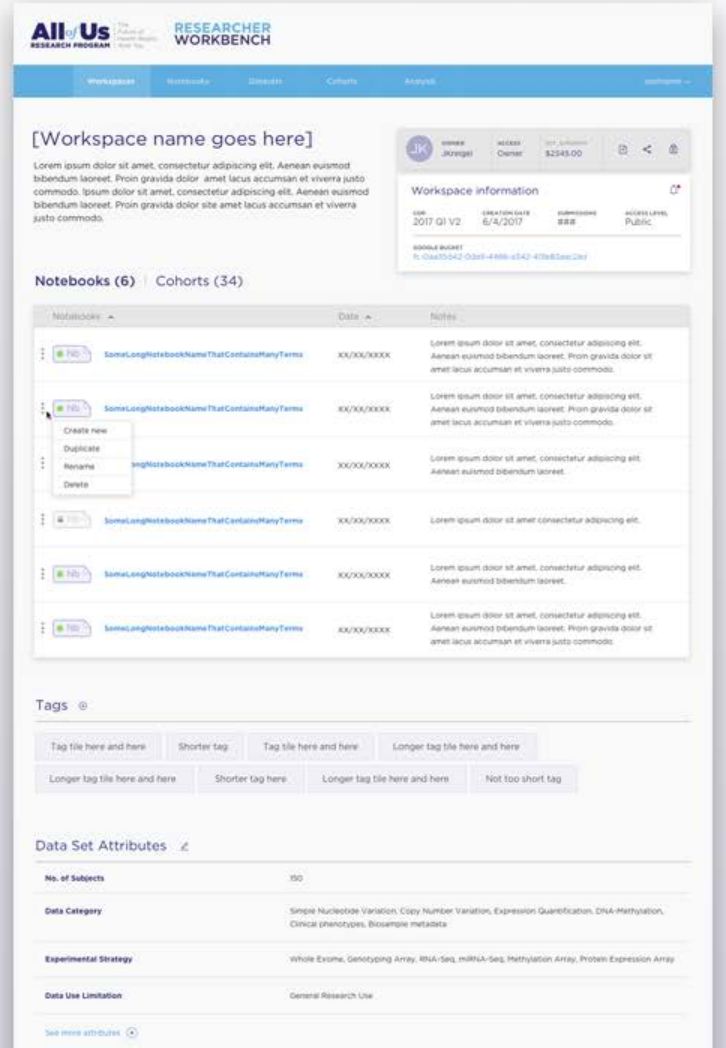


Data Browsing [wireframes]



Workspaces [wireframes]





Questions?