ODSS Search Listening Tour Report

This presentation summarizes the Search Listening Tour done on behalf of NIH Office of Data Science Strategy by BioTeam conducted in the period between Sept. 2020-July 2021

BIOTEAM
Enabling Science

Overview/ Summary

- *This deck summarizes the contents of the Search Listening Tour from project planning to workshop planning*
- ***Project Overview** discusses the steps and goals of the Listening Tour*
- ***What We Heard** provides an introduction to the types of feedback we received from interviewees*
- ***Results: the Data** discusses the process that the data goes through during the search process from being gathered to being analyzed and then reused*
- ***Results: the Journey** discusses the ways that users intersect the data during the search process*
- ***Results: the User** presents use cases gathered during the Listening Tour*
- ***The Workshop** provides brief information on the upcoming Search Workshop*

BIOTEAM
Enabling Science

Table of Contents

BIOTEAM
Enabling Science

Project Overview

Objective: Gather Data to Coordinate a Search Strategy

- Search is a critical first step for any scientist looking to use data
- A more coordinated search strategy could
  - Facilitate access to more of NIH (and NIH-funded) sources of information
  - Save researchers time and money
  - Enable more data use and reuse (increasing ROI for NIH investments)
  - Support downstream analyses, analytics, etc.. etc..
- Methodology:
  - This project used a light-weight ‘Listening Tour’ approach to gain a sense of what ‘search’ means to the community and to identify areas for subsequent analysis.
  - Conduct a limited number of interviews across a broad cross-section of relevant stakeholders to identify high-level themes and get a sense of the key areas of interest.
  - Combine with follow-on workshop to explore key themes in more detail.

Project Goals

The overall goals of this project are to capture the broader gestalt perspective on ‘Search’, for example:

a. How broad is the landscape
b. Existing activities and approaches
c. Use cases
d. Needs and Requirements
e. Potential Strategies for next steps/implementation

First, let’s start with defining ”Search”

The word ‘Search’ covers a very broad set of actions in the area of biomedical research. The three lists opposite show some of the actions that researchers often perform that come under the heading of ‘Search’

It can be seen that many activities are essentially examples of trying to find, or Search, for different types of information, housed in different places. This gives some sense of ***What*** is being search for, we go into more detail about ***Who*** is doing the searching and ***Why*** in the use case section later in the document.

I want to...

- Get
- Find
- Browse
- Download
- Look for
- Track down
- See if we have
- Find out more about
- I.e. SEARCH….

for..

- Information on...
- Datasets like…
- Genes expressed in...
- Records for…
- Cohorts that have...
- Clinical data matching...
- Phenotype data for..
- Genomic data for...
- Bio samples that match...
- Who to talk to about...
- Things that look like...
- How to...
- Evidence for..
- Papers about..

by looking at...

- Google
  - Scholar
  - Dataset search
- Knowledgebase(s)
  - NLM/NCBI
  - AGR/MGD/RGD/
  - BDC, AnVIL, GCRC
  - Kids First, NDA
- Database(s)
- Local system(s)
- Distributed system(s)
- Systems worldwide
- etc.…….

Topics of interest for the interviews

The first phase of this project entailed interviewing a cross-section of scientists, technologists and other stakeholders in the area of Search. The types of topics discussed in these interviews are shown opposite.

- **Search Use Cases**
  - How do scientists search
  - What are they looking for and why
  - What does success look like
  - What types of data are involved
- **Metadata**
  - Metadata schemas
  - Metadata creation
  - Ontologies and terminologies
- **Search technologies**
  - Semantic search
  - Knowledge graphs
  - Federated Search approaches
- **Technological challenges**
- **Pain points and rate limiting steps**
- **UI/UX aspects of search**
- **Cultural challenges**
  - Data sharing plans
  - Encouraging effective curation and metadata annotation
- **Training and education**
  - Annotation and data curation
  - Use of appropriate metadata
- **Examples of effective search, approaches taken by other data-rich communities**

Historical Context

Improving access to data has been a topic of significant interest and investment for over a decade. Relevant tools and technologies have advanced significantly in this time (e.g. cloud computing, knowledge graphs and the semantic web) and there are important initiatives addressing the challenge of finding and using data (and other assets) across platforms.

That being said, data is being generated in greater amounts than ever before and the challenge of finding and leveraging this wealth of data still remains today.

- 2012 - Data and Informatics Working Group Report to The Advisory Committee to the Director
  - "The DIWG recommends that the NIH should invest in technology and tools needed to enable researchers to easily find, access, analyze, and curate research data"
  - Recommendation 1: Promote Data Sharing Through Central and Federated Catalogues
    - Recommendation 1a. Establish a Minimal Metadata Framework for Data Sharing
    - Recommendation 1b. Create Catalogues and Tools to Facilitate Data Sharing
    - Recommendation 1c. Enhance and Incentivize a Data Sharing Policy for NIH-Funded Data
  - Outcomes
    - BioCaddie was funded
    - DATS was developed as a minimal metadata framework
    - DataMed was created as a platform to facilitate data sharing
    - NIH recently updated its Data Sharing Policy
  - However
    - BioCaddie ended when BD2K ended
    - DataMed is no longer funded
    - DATS has been superseded by DATMM (at least for internal NIH usage)

WHAT WE HEARD

Observations from the interviews - Search

Search is the hardest problem...

Inability to search data across all NIH ICs; Multiple interfaces for Search are distracting and inefficient.

We have 8 different data domains [in our project], this helps develop a data ecosystem, but there are barriers if you want to search across domains

How to implement a somewhat consistent search functionality across diverse scientific domains?

“There should be a PubMed for data.”

“We want 'Vivo' for Datasets.”

Search, regardless of how we do it, is dependent on metadata that is consistent

Observations from the interviews - Metadata

Develop a simple system by which scientists would register their data, i.e., data registry, catalog, indexing, etc..

Data registries should include basic, high-level information, i.e., PI, start/end dates, demographics, etc..

How to come up with a manageable yet useful metadata schema, that can be broadly applied?

Metadata is a difficult issue, people have different meanings [for the same words]

Different groups have different ideas and different needs for what metadata fields are important to capture

Observations from the interviews - Technology

We can’t put everything in one spot, it doesn’t make sense. We need different repositories but have an open API over all those repositories. That is the key.

"Some top down architecture choices may be needed, multiple parallel architectures won’t help"

How to come up with an appropriate technological framework that can support shared functionality across diverse environments?

"Need some sort of clearing house framework [that can handle] more types of data (not just omics)"

If we can solve searching via protected access metadata this would be a game changer.

Cross platform search requires that things are abstracted 'up a level', there is no clear business owner, groups have to find common goals and commit to take them on...

Observations from the interviews - Data sharing

Data sharing - People have a lot of excuses for holding onto the data.

Need a culture change on giving data and understanding how messy data can be.

Need a policy or condition for funding to make people fill out data registry/catalog forms.

How to enable reliable data sharing and the high quality metadata needed to support search?

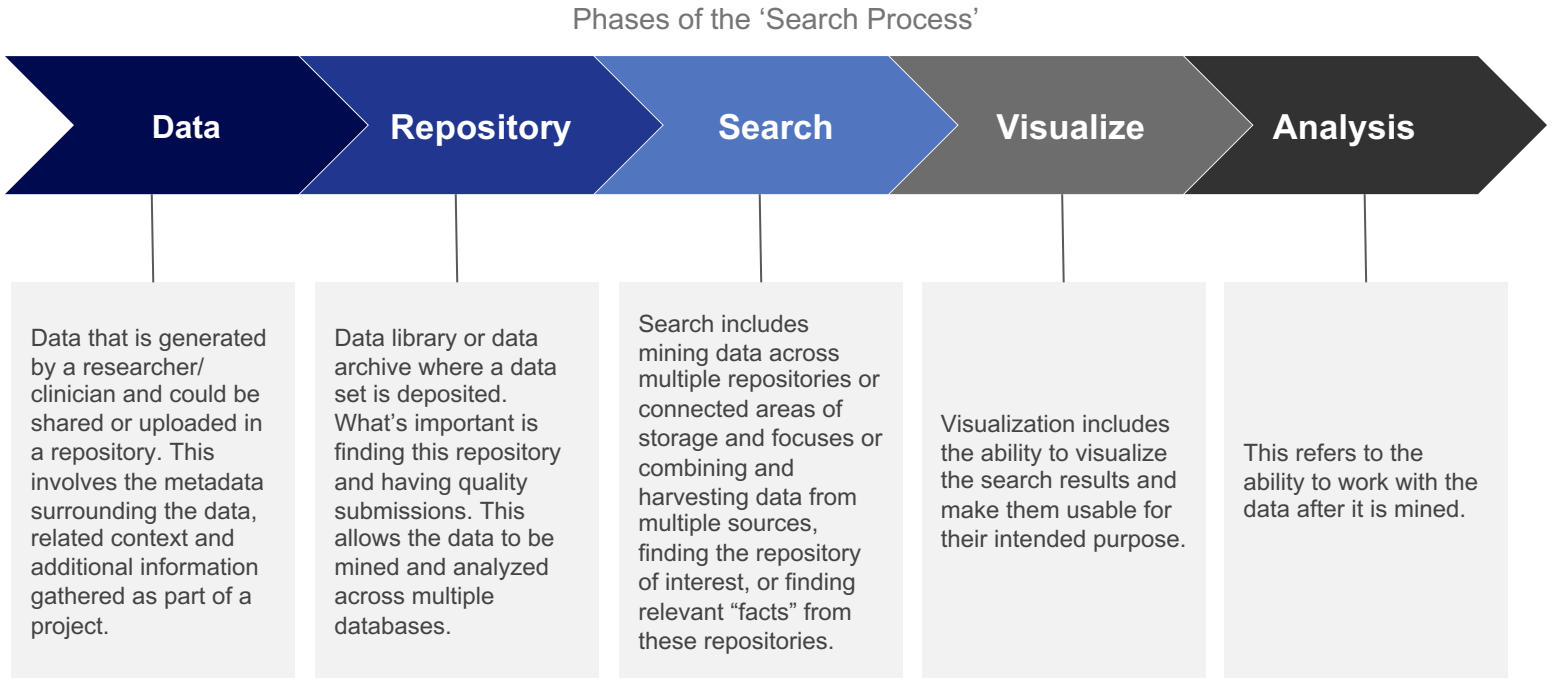
Researchers are having to do the same analysis over and over again because re-sharing data is so difficult.

The scientist isn't going to be the person who is going to do the annotations.

If the scientist's career will live and die by search (as it will eventually), he/she is going to want good annotation and rendering [of their data], they may not want to do it but they want what it enables.

RESULTS: THE DATA

A Data-Centric Perspective of the Search Process



Areas of Impact

- Through the analysis of interviews, we identified nine (9) areas that affect the search process
- Each of these 9 areas of impact affect parts of the data flow:
  - **Policy:** refers to a rule or plan of action which is set out by a governing body and is then followed by the members of that group or organization e.g. Data Sharing Standards
  - **Data Sharing Culture:** the collective behaviors and values of those who produce data that is then needed to further subsequent research
  - **UI/UX:** User experience and design is the process of supporting user behavior through usability, usefulness, and desirability of the search experience
  - **Data Quality & Management:** quality is a measure of how well suited a dataset is to serve its specific purpose and management refers to the process of storing data to make it useful
  - **Data Access:** refers to permissions and necessary access issues that need to be resolved for someone to have access to data
  - **Discoverability:** the effort that it takes to find the location of data and datasets related to a particular topic or the ease of exploring a new topic
  - **Metadata:** a set of data standards that describe and give information about other data
  - **Data Reuse:** either the reanalysis of datasets or combining different datasets to answer the same question with a new method
  - **Interoperability:** the ability to share information between multiple organizations, sites, repositories

Areas that Impact the Search Process

Each of the data stages faces its own set of challenges in the areas outlined below

	Data	Repository	Search	Visualize	Analysis
Policy	X	X	X	X	X
Data Sharing Culture	X	X			
UI/UX		X	X	X	X
Data Quality & Management	X	X	X		X
Data Access		X	X		
Discoverability	X	X	X	X	X
Metadata	X	X	X		X
Data Reuse	X				X
Interoperability		X	X	X	

Analysis of the Search Process

The interviews can be broken down by topics related to the search process outlined previously

- For each step in the search process, we have broken down the interview comments to highlight:
  - **Characteristics**: comments which speak to the current status each data-centered process
  - **Challenges**: comments which speak to the current roadblocks and challenges as they relate to this process
  - **Aspirations**: comments which highlight where interviewees would like to see the area evolve towards
  - **Summary**: an overall summary of interview findings

Data

Characteristics

“We can’t do search without data”

“People have different lenses through which they see data”

“Data is distributed and heterogeneous”

“Spent most career making data comparable and consistent”

“Costs money to submit data”

Challenges

“Need to know where data comes from”

“Annotation is an afterthought and not taken very seriously”

“People make up their own "ontologies"”

“Curation: there is no overall consensus on curation of data sets”

“Different communities converge around their own common language”

Aspirations

“NIH should make policies that support researchers to comply with the spirit and goals of data sharing”

“Annotation has to be a first rate priority: as important as discovering the molecule in the first place.”

“Make your data discoverable, part of the economy of data submission”

“No one won a Nobel prize for annotation”

Summary:
We are hearing people talking about lack of incentives for sharing data, the value of harmonizing data and submitting quality data. We have heard multiple mentions of the desire for NIH to provide firm requirements for data sharing and stronger enforcement of the data sharing policies. Data diversity seems to be accepted as a given, but quality submissions should be promoted to facilitate the linking of different datasets. Additionally, there is general consensus that the NIH could work to improve the generation of high quality data by addressing the associated cultural, educational, and training considerations.

Repository

Characteristics

“Search or query is fundamentally dependent on metadata for reliable retrieval”

“Often people just want something that’s big enough to point their tool at [so they can find enough data for their analysis independent of the specific qualities of data]”

“I made a dataset but it’s not big enough so I need one that’s similar”

Challenges

“Difficult to find open access data; publicizing open access data is even more difficult.”

“People don't know where to begin.”

“Not possible to segment out data”

“No information on how useful a dataset will be”

“No institutional awareness for libraries”

Aspirations

“Just-in-time or ready-to-use data is extremely valuable”

“Develop a data registration portal” to help in dataset discovery

“Create the next gen of class library services which can help find data with the most relevant findings first”

“Build a transactional layer that traffics in queries and responses and augment with AI”

Summary:

It’s very hard to find what datasets are available, where they are, and the quality of data that is within them. Having multiple places to search is difficult in this scenario and there’s significant interest in a one stop shop to find relevant repositories.

Search

Characteristics

Searches for: projects/studies/datasets, concepts like disease, donor attributes, genes, mutations, workflows

Search using: keywords, gene symbols, anatomical information, phenotype, demographics, image files or non-text queries

“We gain trust in the data because we can track how something was done”

Challenges

Data quality, metadata management, level of data access, education, indexing and cataloguing

“Communities do not understand the process of search and how effective search happens...educate”

“Scalability is challenging. Many interoperability efforts are pairwise but this doesn't scale. We are not converging on a single standard”

Aspirations

Search “built for serendipity” would promote discoverability

“A top down architecture instead that allows harmonization and variation along the layers” would provide structure to incoming data

“Start with a ‘simple metadata standard;”

“Socially it will be helpful to know what ‘search’ means”

Summary:

There seems to be in interest in a more standardized approach with supporting policies and a focus on clean data, data sharing standards, and UI/UX with open access APIs built in (or opportunity for community to develop) that can handle federated repositories. There are additional requests for minimal metadata standards. There are additional points around consent, security, and data access that are more regulatory in nature.

Visualize

Characteristics

Visualization is an important aspect of the results that are derived from the search

Challenges

“No rich system to visualize the data”

“Difficult to extract what people are searching for. Diversity in users. Hard to create multiple interfaces for diverse groups”

“There is to a learning curve [to learning how to generate and visualize result and] it is a barrier for new investigators”

Aspirations

There is a desire to “Leveraging the hierarchy” to produce a way to search along the hierarchy of data

Various tools such as knowledge graphs would allow to “analyze in many dimensions”

Visualizations should keep the context that surrounds the data for better interpretability

Summary:

UI/UX played a large role here with the need to explore the data from various perspectives and levels of complexity. We have heard that the technology is not the problem, whether its knowledge graphs or tables, it needs to provide the ability to filter data. We also heard that as long as there is quality data, there is the potential to open the development of visualization APIs to the community.

Analysis

Characteristics

Currently there is a lot of downloading from databases to analyze the data

Context around data is just as important as the data itself

“Different players need different levels of complexity”

Challenges

How to hand off search results to compute

“How to work with data in notebooks or other environments”

“How to track what type of analysis was done across what databases”

“I don’t want someone else to waste their time doing my same analysis with my same dataset”

Aspirations

“The need to download data and analyze it where the tools are”

“Easy to download and work in platform of choice”

“There needs to be a means to...conduct secondary analysis/ compute however you want…[then] upload or indicate the work that was done”

Summary:

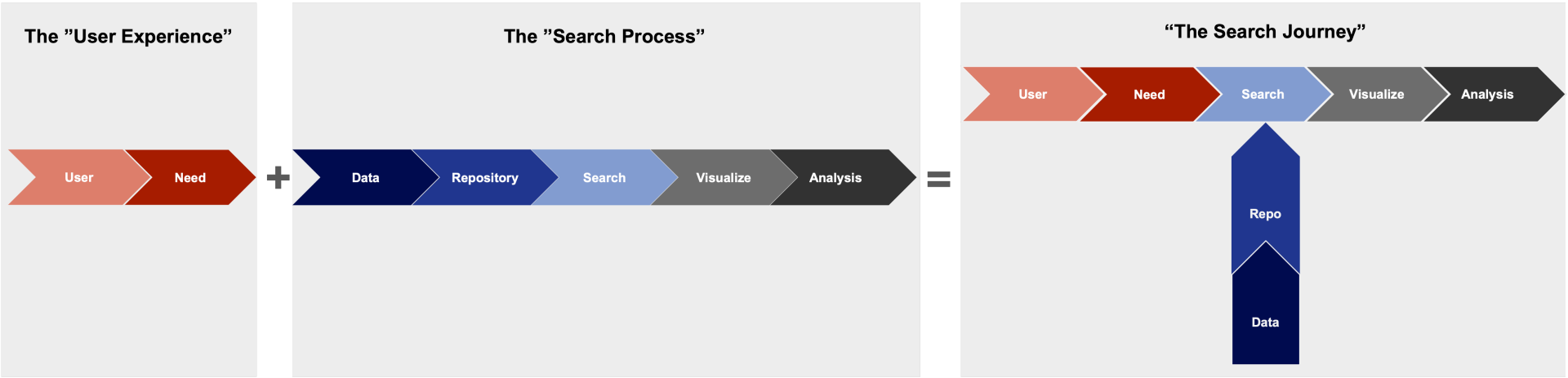
In terms of analysis, it seems to be important to meet different communities and users where they are. In general, linking data and publishing workflows and dataset ID’s is a common theme. Currently there’s no way to do analysis of data but then share it again.

RESULTS:
THE JOURNEY

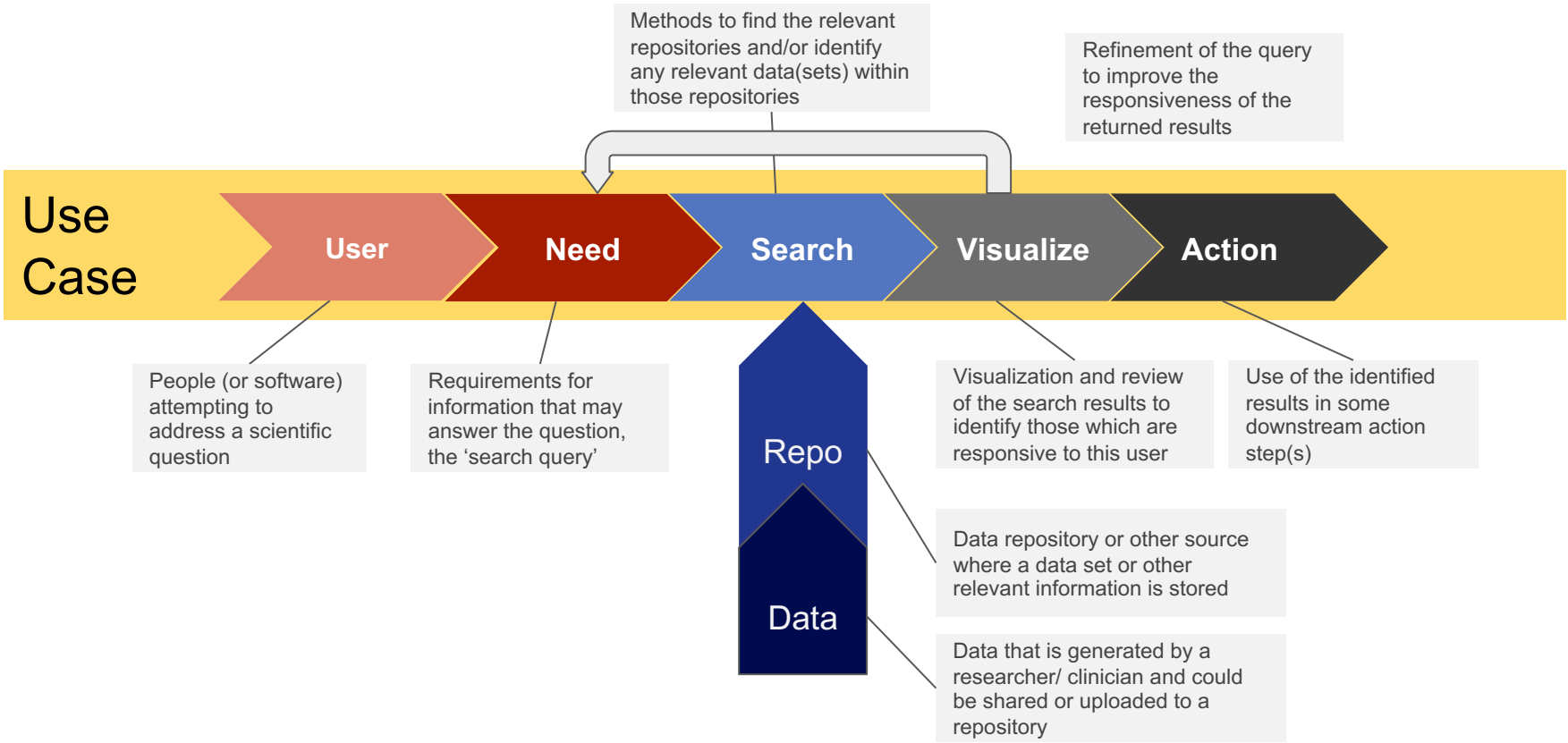
From “Search Process” to “User Experience”

BIOTEAM
Enabling Science

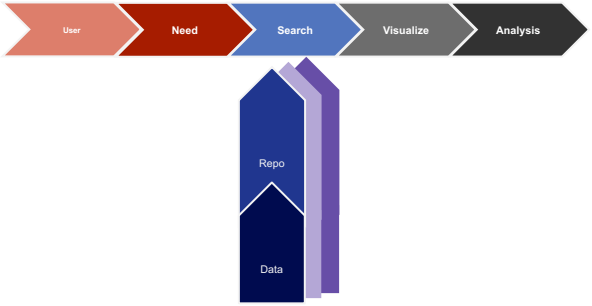
- While the initial results focused on the process of data flow, it is critical to understand how the user and the user’s journey intersects the data
- User needs and use cases were extracted from interviews and incorporated into the data ”Search Process”



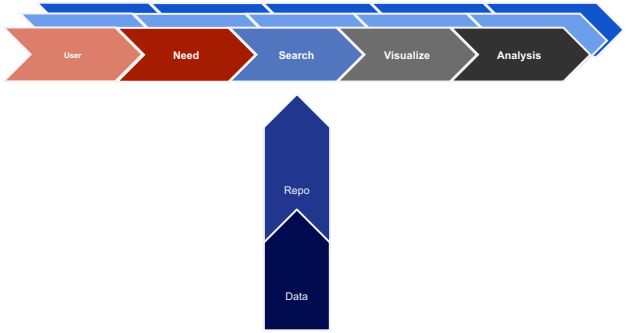
Overview of the Full Search Process



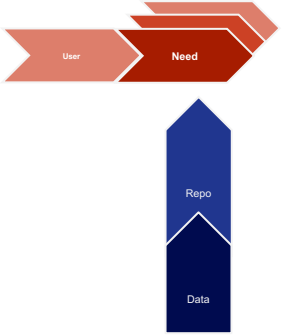
Diversity in the Search Process:



A single use case often needs to query multiple data sources

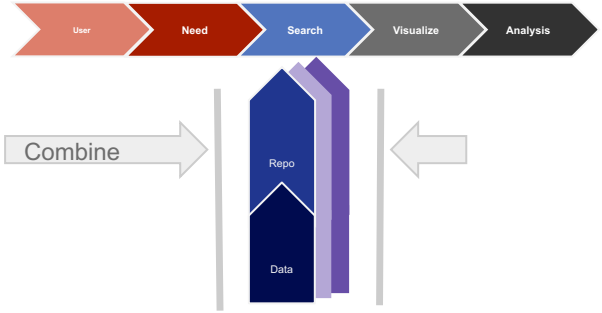


A single data source can be relevant to multiple use cases



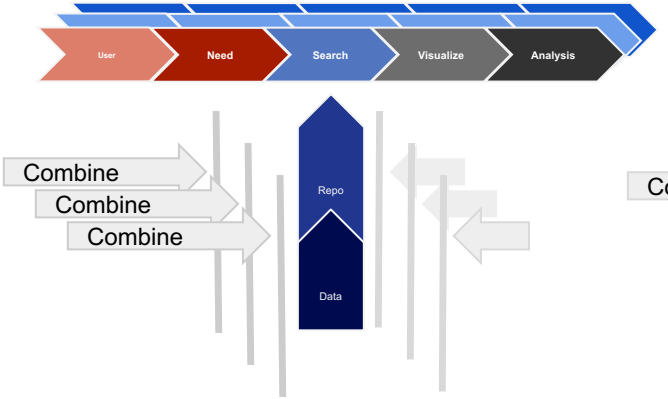
A single user can have multiple needs in order to answer a research question

Use Cases: Search as a way to **combine**



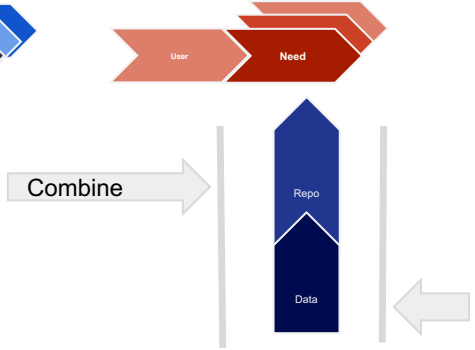
A single use case often needs to query multiple data sources:

Need to combine different sources: image and GWAS



A single data source can be relevant to multiple use cases

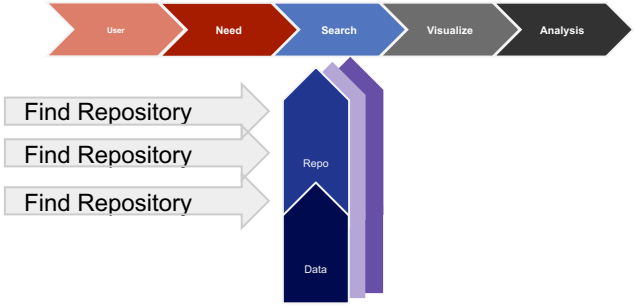
Different users need to combine in different ways: bioinformatician vs clinical researcher



A single user can have multiple needs in order to answer a research question

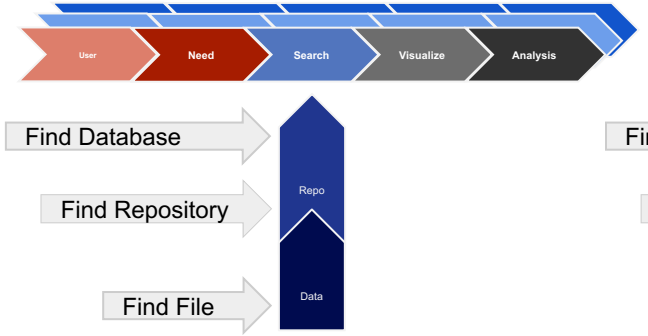
Need to combine across various levels of the data: database to formulate hypothesis and file to test it

Use Cases: Search as a way to **find/answer**



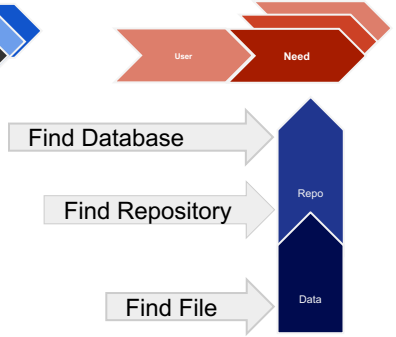
A single use case often needs to query multiple data sources:

Some questions will require access to a variety of data to compile an answer



A single data source can be relevant to multiple use cases at various “depths”:

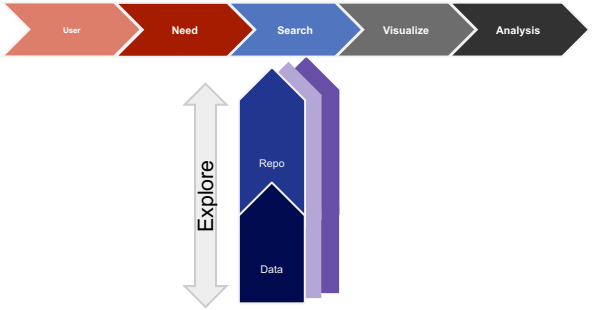
One “database” may need to be accessed at various depths by different users: summaries, files, answers



A single user can have multiple needs in order to answer a research question:

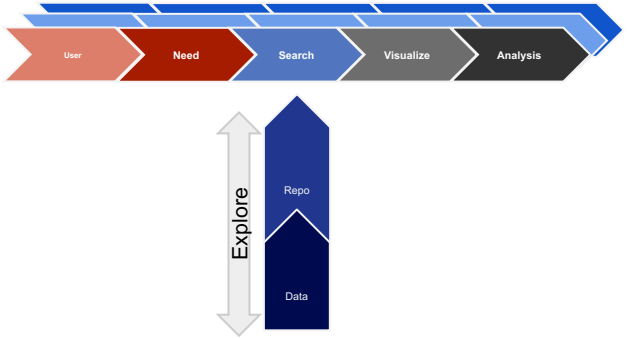
Throughout the project, each user will likely need different depths of access

Use Cases: Search as a way to **explore/discover**



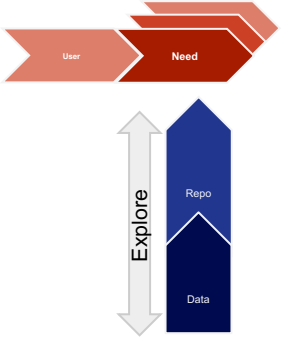
A single use case often needs to query multiple data sources:

Need to explore up and down databases for ideas and connections



A single data source can be relevant to multiple use cases at various “depths”:

Different researchers will need to explore different levels of data



A single user can have multiple needs in order to answer a research question:

During the course of a project, each user’s needs will change with analysis needs

Use Cases: Search as a way to explore **non-textual concepts**

Although Search is often thought of in terms of textual inputs, there are many variations on search inputs to include:

- Search by **Images**
- Search by **Structures**
- Search by **Similarity**
- Search by **Graph**
- Search by **Gene**
- Search by **Geography/Location**
- Search by **Scale**
- Search by **Social Determinants of Health**
- Search by **Spectra** (combination of pictures and text)

For further use case exploration of these concepts, refer to Annex

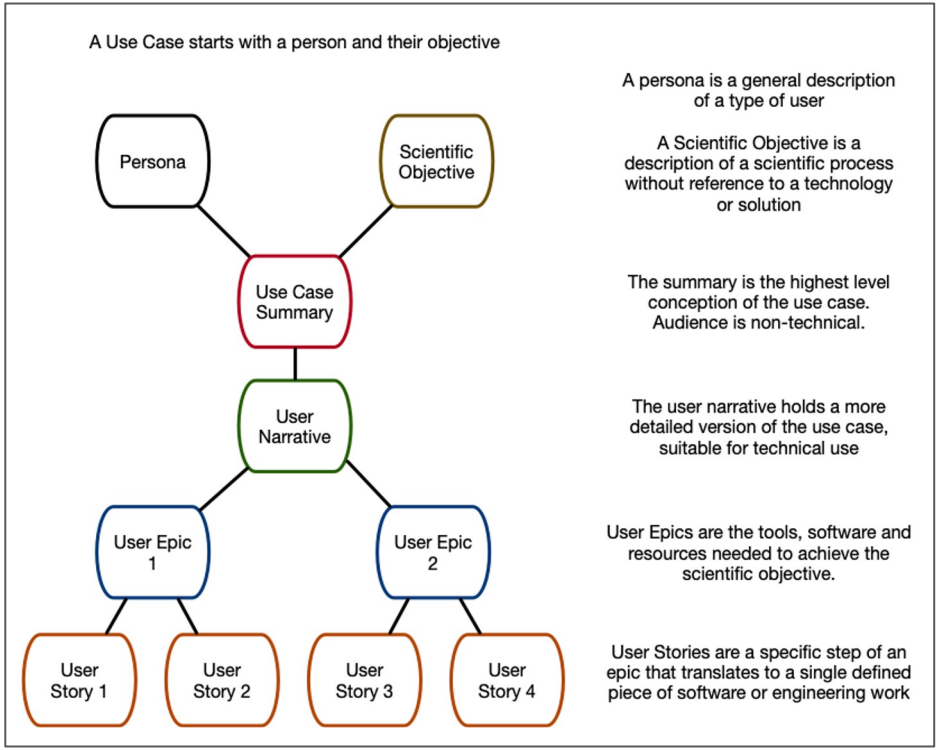
RESULTS:
THE USE CASES

Use Cases

‘Use Cases’, scenarios, stories, epics - they can mean different things to different groups

The Data Commons Pilot Phase Consortium (DCPPC) developed an excellent online use case library which also contained a diagram containing definitions of Personas, use cases, narratives and other terms (see opposite).

To reduce confusion we have chosen to follow the same approach. This document contains **Use Case Summaries**, high level, non-technical descriptions that link a particular persona (e.g. ‘Clinical Researcher’) to a specific Scientific Objective.



Overview of Use Case Synthesis

The Use Cases provide a set of high-level summaries that capture essential scientific objectives and places them in the context of the larger mission

- ~50 use cases were identified during the course of the interviews
- These were analyzed to identify various categories and trends in the interview feedback:
  - Did the interview describe a Use Case or a Feature
    - Is it a way that users are looking for things or
    - Is it a feature that makes searching easier
  - What category of search use case did they describe (these categories arose from the Search workshop planning group's discussions:
    - Cohort Building, Dataset Discovery, Results-based Search
- Example queries were documented
- High-level gaps in current search functionality were identified

Define the three types of use cases?

1. **Dataset Discovery:** this type of search focuses on trying to find datasets (typically one or more files) on a particular topic. Finding relevant dataset(s) is usually not an end in itself but is a key intermediate step in the integration and reuse of existing data to inform new scientific hypotheses.
2. **Cohort Building:** In many use case scenarios, this is the process of finding and pulling together sets of human subjects that can be used for some type of comparative analysis. We also identified other ‘cohort building-like’ use cases that do not refer to patient-based cohorts but rather to the nature of pulling together data from multiple sources into one source for subsequent analysis.
3. **Results- Based Search:** this type of search focuses on identifying facts and pieces of connected “knowledge” on a specific topic in a quick and low-effort manner that does not require additional data manipulation and analysis.

General Overview

During the interview process we identified different user requests and examples that we categorized into use cases for search and features for search technology:

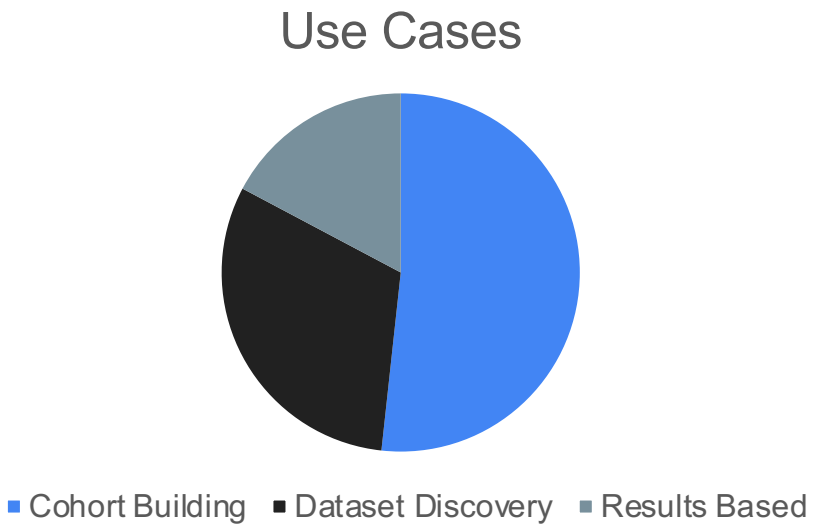
Use Case:

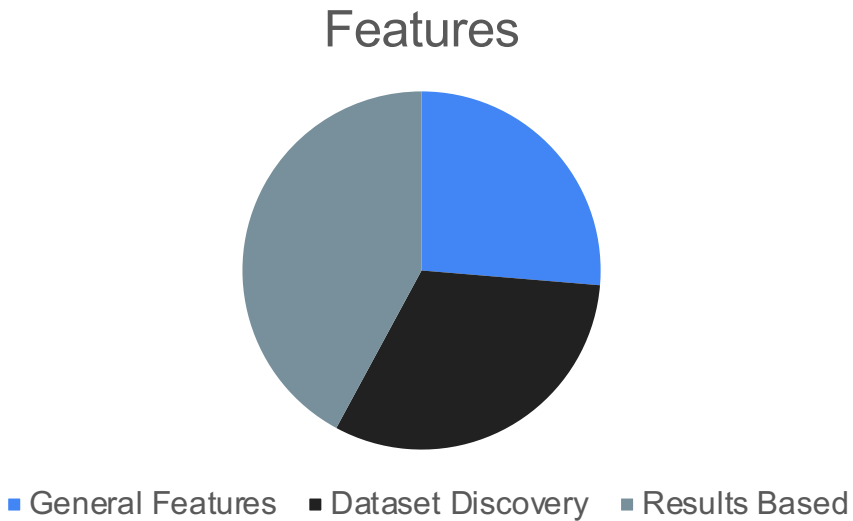
A user with a particular search in order to fulfill a particular goal and end in mind

Feature:

When doing a specific type of search, this feature would be very useful for that type of search

Each use case can include aspects of expansion for hypothesis generation or funneling down for fact finding





Cohort Building

Using search to understand landscape

Clinical Researcher

As a Clinical Researcher, I want to **find data across different portals using patient attributes** (female, less than 30) in order **to get and overview** of a specific trait in a portion of a population.

Core Motivations

- I need to aggregate data across portals using donor attributes
- Want to explore the data in an “Amazon” like way
- Data catalogues would provide 80% of value

Example Query

- Find all females under age of 30 who have a specific rare mutation
- Find all males who were admitted to hospital for COVID19 in specific region
- Are there enough patients that look like this phenotype

Classifications

- Cohort Building
- Exploring Data

Gaps

- Harmonized metadata
- Data catalogues
- Interactive interfaces

Using search to build cohorts

Researcher

As a Researcher, I **want to be able to create a cohort of patients for clinical analysis** with clinical data and imaging data, in **order to make inferences**, i.e., COVID positivity based on lung image analysis.

Core Motivations

- Cohort for clinical analysis
- While collecting imaging data with clinical data, want to see lung images to determine covid positivity

Classifications

- Cohort Building
- Combining Data

Example Query

- Find me patients with lung images
- Find me patients who were tested for COVID

Gaps

- Ability to apply the search criteria across multiple databases
- Ability to quickly capture if a dataset is useful
- Harmonized metadata
- Data Access

Using search to refine cohorts

Clinical Researcher

As a Researcher, **I want to further refine a large set of subjects** identified through an initial search, in **order to create a cohort of patients** for subsequent analyses that meets my needs (e.g. sufficient numbers of subjects, compatible data collection or data analysis methods used, etc.)

Core Motivations

- Start with 500K patients and then whittle down
- I need to reduce to a quality cohort

Classifications

- Cohort Building
- Find Specific Data

Example Query

- Find me patient information across time (temporal events)
- Find all patients in this dataset that had blood based metabolite analysis
- Filter the dataset only to patients who completed the trial

Gaps

- Ability to filter down by same parameters in different repositories
- Metadata harmonization
- Data Access

Dataset Discovery

Using search to find all relevant repositories

Principal Investigator

As a Principle Investigator, I **want to be able to find all social health data repositories** related to COVID-19 in **order to have a comprehensive view of all available research and data conducted** for the RADx Underserved Population Initiative, to date.

Core Motivations

- My first question is where do I find data for a new topic
- I need to know I'm not missing something that can be useful

Classifications

- Dataset Discovery
- Exploring Data

Example Query

- Find me all social health data related to COVID19 patients in the past year
- Find me all COVID19 data in the RADxUP Initiative

Gaps

- Catalogue of repositories
- Updating indexes to keep abreast with new topics

Using search to look for specific datasets

Early Career Scientist

As an Early Career Scientist, I **want to find data linked to a publication** I am reading in order to **analyze the data in a specific way**.

Core Motivations

- Quick search to find data I know exists
- Want to read a publication and find the data directly
- May need it at the file level

Classifications

- Dataset Discovery
- Find Specific Data

Example Query

- Find the characteristics of the dataset analyzed in this article
- Find the location of the dataset analyzed in this article
- Find the workflow for the data that was analyzed in this article

Gaps

- Linked Data
- Data Identifiers
- DataMed
- Easy interface for someone who is not skilled in bioinformatics

Using search to build ML/AI models

Data Scientist

As a Data Scientist, I want to **find relevant datasets of sufficient quality and quantity** in order to **train ML/AI models**.

Core Motivations

- I need a landscape of data to play with without it being disease or hypothesis specific
- I need the data to have specific parameters for my technical application

Classifications

- Dataset Discovery
- Combining Data

Example Query

- Find all datasets with more than 50 patients and WGS for each patient
- Find all datasets where the sequencing was long-read
- Find any dataset where patients are tracked over 10 or more years

Gaps

- Ability to search on unique parameters

Results Based Search

Using search to quickly access answer

Principal Investigator

As a Principal Investigator, I want an **easy way to answer a specific question** without the additional burden of data download or analysis in order **to avoid the long data-discovery process**.

Core Motivations

- Want a user-friendly, Google-like way to see the data
- Don't want to download
- Need a way to get information easily for important decision making

Classifications

- Results Based Search
- Find Specific Data

Example Query

- What disorders is Gene ABC1 related to
- Find me the 3D structure of the COVID-19 protein
- What's the incidence of Alzheimer's in the Native American population

Gaps

- UI interfaces for searching ease
- Underlying integrated knowledge bases

Using search to find concrete answers

Physician

As a Physician in a clinic, I want to **search for information about [a drug, or disease] and receive quick results** that are relevant to my current needs, in **order to efficiently use this information to make decisions for patient care.**

Core Motivations

- I need a reputable answer to a question
- I do not have time to analyze details and need concrete and actionable answers

Classifications

- Results Based Search
- Find Specific Data

Example Query

- Find all side effects of Hydroxychloroquine
- Find most common symptoms of COVID19

Gaps

- UI/UX to present/visualize answers
- Ways to quantify data quality
- Integration of biological datasets/information into Google

THE WORKSHOP

Overall goals and drivers

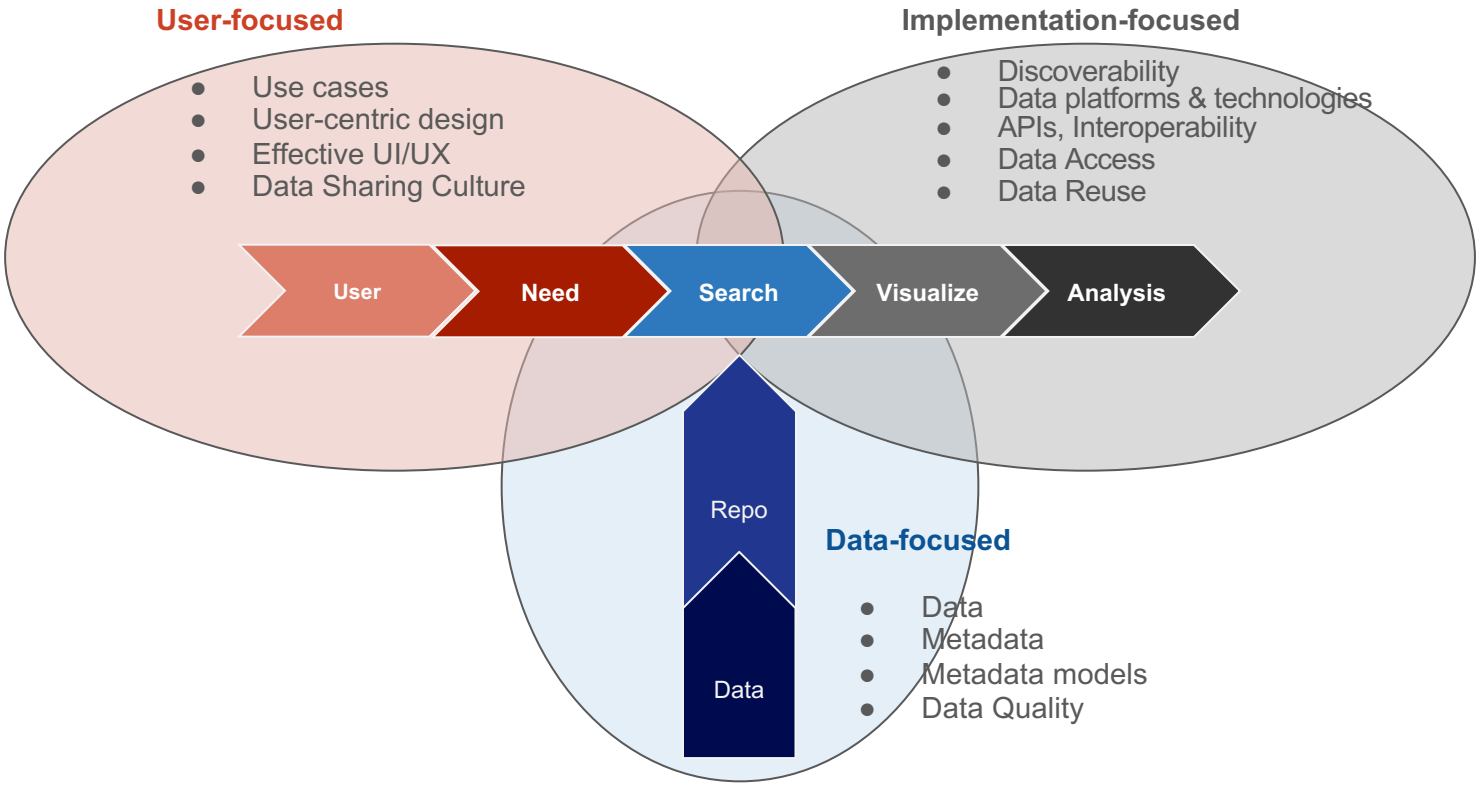
Goal: Explore current capabilities, gaps and opportunities for global data search across the data ecosystem. Workshop will explore selected **science drivers across** three main themes:

- **User expectation** and experience, including UI/UX
- **Data focused** capabilities and context, including metadata for search and data sharing
- **Implementation opportunities**, including discoverability across platforms, repositories and considering data access and data reuse in analytics

Science Drivers:

- **Using search to build cohorts:** find data across different platforms/repositories using patient attributes in order to create a cohort of patients for clinical analysis
- **Using search to find relevant data & repositories:** find data & repositories, links to data in the publications and analyze the data in a specific way or to create computational models.
- **Using search for (complex) information retrieval**: specific question without the additional burden of data download or analysis

High-level themes for the workshop



Workshop logistics

- **Anticipated Date:** January 19th and 20th 2022
- **Format:** Virtual workshop over two half days
- **Keynote Speakers:** danah boyd and Sir Nigel Shadbolt

- More information will be published on the ODSS website (https://datascience.nih.gov) as it becomes available.

The End