A Strategic Framework for BioData Catalyst

V2.0 - 20200403

NIH National Heart, Lung, and Blood Institute

BioData CATALYST

A Strategic Framework for BioData Catalyst

V2.0 - 20200403

Document Status

Version

V2.0

Approvals

Signatures presented below denote review and approval of the BioData Catalyst Strategic Framework. These approvals are given based on the understanding that the Strategic Framework, and the information herein, will be revised at regular periods over the course of the program. It is the responsibility of the Principal Investigator (PI) of each funded team and select NHLBI program staff to add their name(s) in the indicated space below.

Approved Date

4/03/2020

PI Approvals:

PI	Team	Approval Date
Robert L. Grossman (University of Chicago)	Calcium	3/31/20
Anthony Philippakis (Broad Institute)		3/30/20
Benedict Paten (UCSC)		3/30/20
Paul Avillach	Carbon	03/31/20
Ashok Krishnamurthy	Helium	3//30/20
Brandi Davis-Dusenbery	Xenon+	3/31/20

NIH Approvals:

Responsible Person	NIH NHLBI BioData Catalyst Role	Approval Date
Jonathan Kaltman, NHLBI	Program Manager	4/3/2020
Alastair Thomson, NHLBI CIO	Information Security	4/3/2020

Next Review Date

4/03/2021

Document Owner

BDC3

Revision History

Date (YYYYMMDD)	Version Number	Revision Reviewed/ Approved By	Brief Description of Change
20191101	V0	N/A	Draft document created.
20190305	V0.1	Marcie Rathbun	In section 8, added link to Operationalization document: NHLBI DataSTAGE 60 Day o16n Plan v1-2
20190313	V0.2	Rebecca Boyles	User Narrative edits from consortia review incorporated
20190426	V1.0	NHLBI	V1.0 reviewed and approved by NHLBI Links & editing updates [Marcie]
20200403	V2.0		Changes based on annual consortium review: - re-branded as BioData Catalyst & updated relevant graphics - added a Design Principle: Implement rigorous testing and Quality Assurance measures for components and data - replaced all UNs after 1 & 2 with the MVP and a link out to the WP 3.0 RFC

TABLE OF CONTENTS

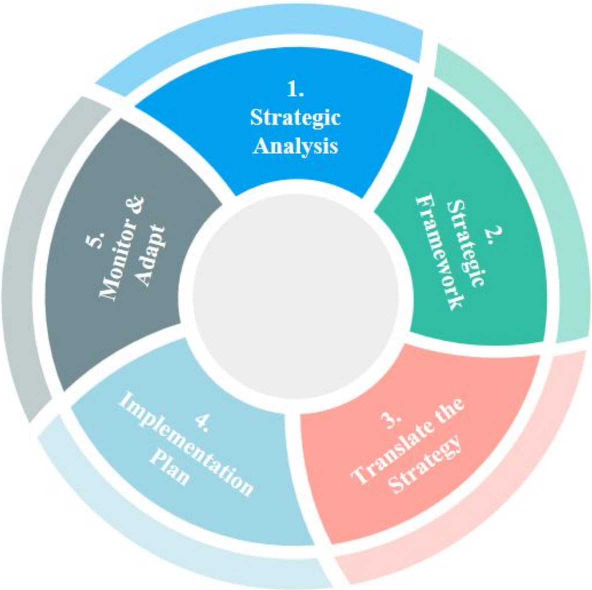
1 INTRODUCTION

1.1 PURPOSE OF THE STRATEGIC FRAMEWORK

The BioData Catalyst Strategic Framework identifies the mission and vision of the BioData Catalyst program and describes how the program will align across stakeholders to execute on common goals and how that performance will be measured. In the creation of this Framework and the complementary Implementation Plan, we focus the Consortium on a common goal, agree on actions, align resources, and prioritize needs.

2 EXECUTIVE SUMMARY

The purpose of the BioData Catalyst Strategic Framework is to articulate a forward-looking path for the BioData Catalyst Consortia and stakeholders to align across a complex Heart, Lung, Blood, and Sleep (HLBS) landscape of technologies, science, and data. The Framework is a culmination of an in-depth process that involved strategic analysis of the data, applicable methodologies, and needs with the key BioData Catalyst stakeholders. This analysis was then developed further into this Framework document. The Framework is evergreen and will be regularly amended to reflect new priorities. The Framework was envisioned and created with guidance from NHLBI and the BioData Catalyst Consortium.



The Strategic Framework consists of a mission, vision, and values, as well as overarching User Narratives and the orthogonal work streams that comprise the types of work needed to execute the BioData Catalyst program.

A separate Implementation Plan, which maps goals, objectives, and strategies into specific Features, accompanies the Strategic Framework. The Implementation Plan, coupled with the Project Management Plan, establishes priorities, accountabilities, success indicators, and timeline and resources for projects. To create project priorities and transparency, the Implementation Plan uses the following guiding principles: availability of resources, impact on the NHLBI mission, return on investment, the utilization of technologies that maximize data security and integrity, and the implementation of cost-effective solutions.

3 BACKGROUND

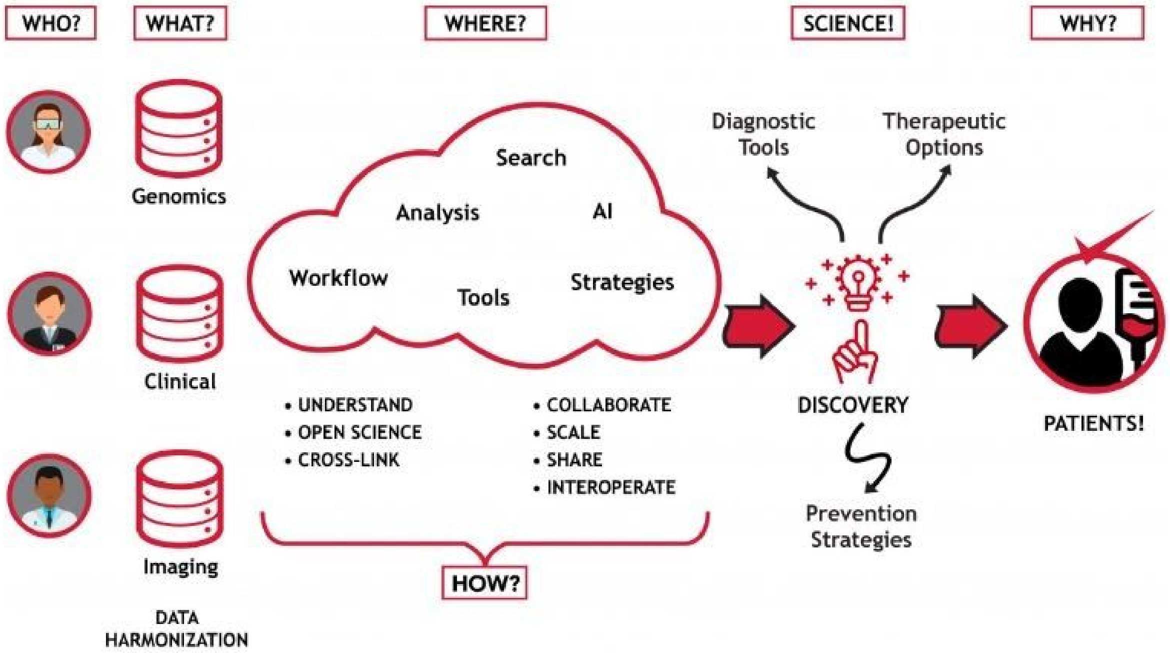
3.1 PROBLEM STATEMENT

Much has been written about the explosion of biomedical data that has been largely driven by the genomic revolution (Collins, Morgan, and Patrinos 2003; Green, Watson, and Collins 2015). In addition to the increasing availability and volume of genomic data, researchers and clinicians have seen a dramatic increase in data through the adoption of high-throughput assays and high-resolution imaging technologies. The need to leverage these data resources through the application of emerging data science approaches is recognized in the NHLBI Strategic Vision (National Heart, Lung, and Blood Institute, and Others 2016).

Modern HLBS research must now operate across a diverse data landscape that includes large data resources in high-throughput genomic, proteomic, metabolomic, microbiome, imaging, personal wearable, behavioral, and clinical domains. To support this work, advancements are needed in our ability to provide researchers with cost-effective and rigorous storage, management, tooling, and computation within their current workflows while upholding the NIH's responsibility to appropriately manage human subject data.

3.2 WHAT IS BIODATA CATALYST?

In 2007, Jim Gray famously described a Fourth Paradigm of Science, in which science of the future would leverage interoperable knowledge and data online (Hey et al. 2009). The NHLBI BioData Catalyst is an instance of a Data Commons, where HLBS researchers can go to find, search, access, share, store, crosslink, and compute on large scale data sets. It will be a cloud-based platform that has, at its foundation, a Commons that provides controlled access to data, tools, applications, and workflows to enable these capabilities in secure workspaces. BioData Catalyst will accelerate efficient biomedical research and maximize community engagement, productivity and discovery.



4 MISSION

Critical to the success of the BioData Catalyst program is consensus through a mission statement that articulates what BioData Catalyst will provide for the user, developer, and programmatic communities.

The NHLBI BioData Catalyst's *mission* is to develop and integrate advanced cyberinfrastructure, leading-edge tools, and FAIR data to support the NHLBI research community and accelerate discovery.

5 VISION

The BioData Catalyst Consortium is jointly working towards a common future vision that will drive our implementation and management actions.

The *vision* for BioData Catalyst is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.

6 CONSORTIUM VALUES

In all of our work as the BioData Catalyst Consortium, we remain committed to a set of values that guides our thinking and ideas as an organization.

The BioData Catalyst Consortium is committed to:

- Engagement with stakeholders to inform development;
- Responsible stewardship of NHLBI data assets and resources;
- Respect for the study participant and individual consent;
- Service to scientific advancements and HLBS health; and
- Alignment with the NHLBI Strategic Vision, NIH Data Science Strategic Plan, and related emerging data-intensive initiatives;
- Training and development of the next generation of health data scientists.

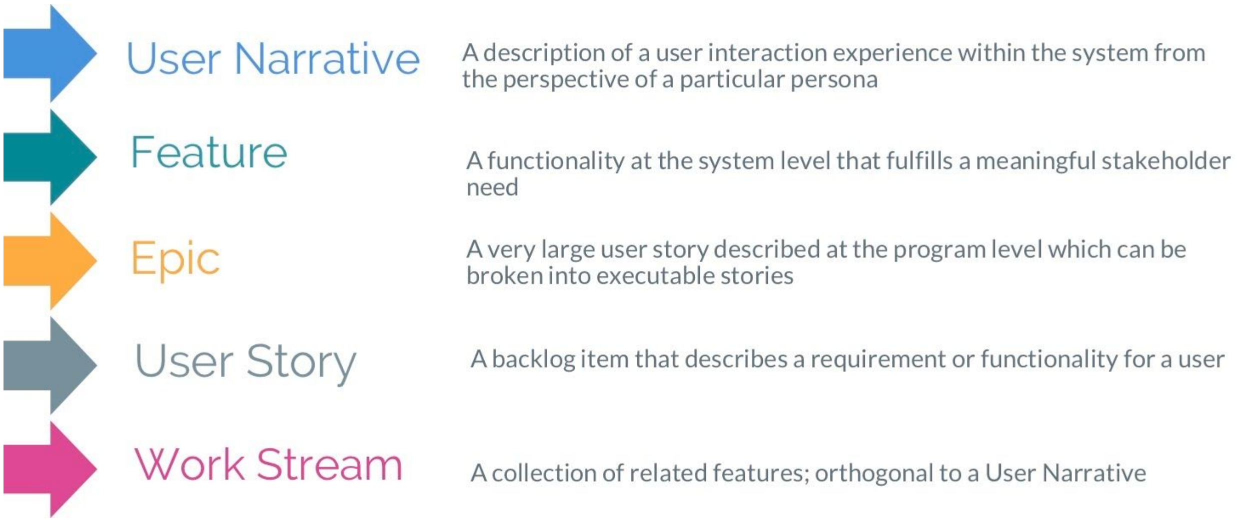
7 DESIGN PRINCIPLES

Design Principles are common guidelines or considerations that inform the approach to the BioData Catalyst development. Here we highlight a number of high-level Design Principles that are cross-cutting across the User Narratives for BioData Catalyst.

The cross-cutting Design Principles are:

- Meet user needs and incorporate feedback
- Leverage existing tools and infrastructure, when feasible
- Do not duplicate infrastructure components
- Duplicate functionality when intentional and reasonable
- Architect interoperability with relevant systems
- Encounter a seamless experience, regardless of underlying components
- Leverage cost-advantageous cloud resources
- Support scalability and extension of functionality
- Have an early impact on computational-driven HLBS science
- Enable consistent, easy access to applications and tools for users across BioData Catalyst
- Provide systems security for hosting identifiable data
- Implement rigorous testing and Quality Assurance measures for components and data

We will use the vocabulary below to discuss the various levels of work breakdown for BioData Catalyst.



User Narrative — A description of a user interaction experience within the system from the perspective of a particular persona

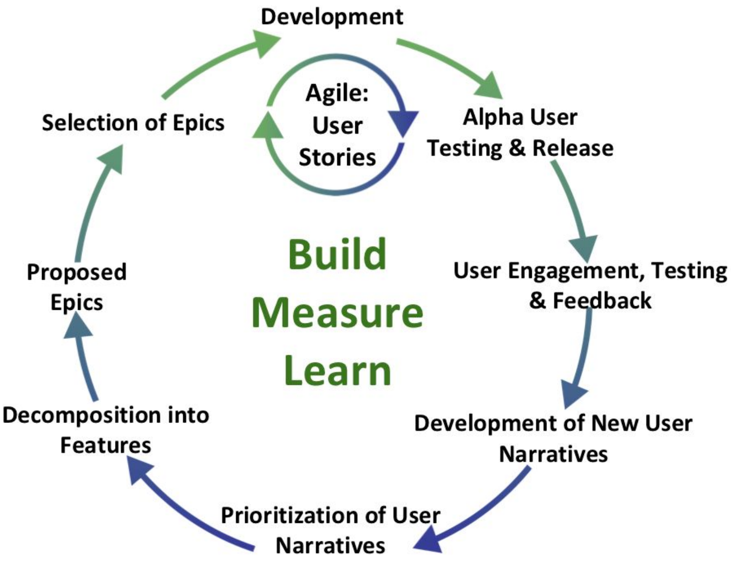
Feature — A functionality at the system level that fulfills a meaningful stakeholder need

Epic — A very large user story described at the program level which can be broken into executable stories

User Story — A backlog item that describes a requirement or functionality for a user

Work Stream — A collection of related features; orthogonal to a User Narrative

These terms are drawn directly from the Agile literature in consultation with NHLBI, but many Agile methods use alternative terminology. Additional details can be found in the BioData Catalyst Implementation Plan. Further, we are using a "Build Measure Learn" design cycle, as shown below. This provides us with a substantial degree of flexibility with respect to the User Stories, while maintaining the discipline and collective focus on the overall objective of BioData Catalyst through less frequent modifications of the User Narratives.


Development
Alpha User Testing & Release
User Engagement, Testing & Feedback
Development of New User Narratives
Prioritization of User Narratives
Decomposition into Features
Proposed Epics
Selection of Epics
Agile: User Stories
Build Measure Learn

7.1 USER NARRATIVES

One way to describe the intended outcome of the BioData Catalyst program is through User Narratives. User Narratives are descriptions of a user interaction experience within the system from the perspective of a particular persona. Within Jira, our project management tool, these User Narratives will further be broken into Features, Epics, and User Stories, as appropriate.

As is further described in the BioData Catalyst Implementation Plan, stakeholder feedback and user testing will drive the prioritization and further development of BioData Catalyst User Narratives. Overall this approach will provide a flexible, coordinated framework to drive BioData Catalyst development while integrating user needs. Tasks necessary to accomplish User Narratives can also be organized into Work Streams, which are orthogonal to a User Narrative. Work Streams group similar activities together to present an alternative view of progress towards the BioData Catalyst vision and map User Narratives to broader Work Streams to help the Consortium meet objectives.

Additional detail on the structure of the work hierarchy can be found in the BioData Catalyst Implementation Plan.

Critically, User Narratives will be used to benchmark progress towards the BioData Catalyst vision by working with users towards specific outcomes and documenting trouble spots and other potential improvements. These identified issues will be funneled into the program's development backlog. The testing will also note positive elements in order to identify potential features or practices to promote across the platform.

User Narratives offer an opportunity to engage potential users in the development process, with regular feedback opportunities (e.g., sprint demos) to ensure that BioData Catalyst is executing towards the vision, but also meeting future users' needs, even as they evolve over time. It is anticipated that new User Narratives will be refined through this process and will be incorporated into the Strategic Framework and Implementation Plan materials. BioData Catalyst will remain connected to the user community, remain agile, and will intentionally evolve to drive future development efforts forward. User Narratives may represent near-term partial solutions that are deployed in stages to solicit user feedback and allow for rapid development.

The below are abbreviated User Narratives formulated in rough six-month timelines with more detail in the near term and less detail further out, with plans to refine as BioData Catalyst progresses. The complete User Narratives will be maintained in the BioData Catalyst User Narratives, Features, and Epics document and will be regularly reviewed and edited through a change control board-led process..

June 2019

1. a) A pre-approved group of computationally savvy researchers can create a user profile on the BioData Catalyst environment using their eRA commons identity. b) They can access molecular and phenotypic data from selected TOPMed datasets for which they have approval and can use a simple cohort search to find and explore phenotypes using an interface that returns counts of study subjects within a single study matching search criteria. c) They can use a visual or programmatic (API) interface to perform computationally demanding (alignment/variant calling, etc.) batch analysis Jupyter notebooks or Rstudio to perform interactive analysis. d) They can collaborate on analyses (including sharing scripts, results, etc.) with other approved researchers.

2. a) A pre-approved group of computationally savvy users can test *proof of concept* functionality to understand and optimize cloud costs associated with running large scale computing or interactive analysis. b) As Alpha Users, they will be able to browse phenotypic and annotations of genomic variables within a single TOPMed study and view standard statistics on returned results.

   NOTE: User Narratives 1 & 2 scheduled for June 2019 have been completed as part of Workplan 2.0.

March 2020

3. **Go live Minimum Viable Product**: Sophisticated and new users, including non-technical highly motivated researchers (both command-line and GUI users) with data approvals and current TOPMed researchers, can: (a) start from the BioData Catalyst landing page to identify and select their platform of choice. They are directed to the relevant platform and can log in using the same username and password across all systems; based on their credentials and the Data Access Committee approval in the system, the user can access data; (b) search PIC-SURE for any phenotypes from selected cohorts in BioData Catalyst consistent with policy; (c) browse list of harmonized TOPMed data using phenotype keywords. User can retrieve list of studies related to one of the 44 harmonized variables; (d) Find and access genomic and reference data files (e.g., genotype files, kinship matrix, annotation files); (e) Supply their own phenotype and computed genomic and reference data files (bring your own data); (f) Analyze in the context of an association study, data using tools, services, and applications available in the platforms (e.g., phenotype data files); (g) Execute GWAS single variant and gene-based rare variant analyses and access sufficient support documentation for this workflow; (h) Enable users to understand and monitor costs associated with their analysis in those cases when the user is responsible for the costs.

September 2020 - March 2022

See RFC-4: Workplan 3.0 User Narratives:
https://docs.google.com/document/d/11nee2ZxOrRjjxKbBVjrzV_Lm8WoBgut7UXwAEHtnlw0/edit

See also: FINAL_Delivery Timeline (2020) BioData Catalyst_slides:
https://docs.google.com/presentation/d/1pXxtBGBQAeP3s4aU-6KmJxtyzBOyg6FlWrpWYShkdbI/edit#slide=id.g6c8b9250a5_1_59

8 REFERENCE DOCUMENTS

- Implementation Plan
- Project Management Plan
- NHLBI DataSTAGE 60 Day o16n Plan v1-2 (drafted by the Operationalization Tiger Team)
- DataSTAGE User Narratives, Features, and Epics
- STAGE-RFC-2_DataSTAGE_Strategic_Planning_Nomenclature
- STAGE-RFC-4_Workplan 3.0_DataSTAGE User Narratives
- MVP_as_UNv1

APPENDIX A: REFERENCES

Collins, Francis S., Michael Morgan, and Aristides Patrinos. 2003. “The Human Genome Project: Lessons from Large-Scale Biology.” *Science* 300 (5617): 286–90.

Green, Eric D., James D. Watson, and Francis S. Collins. 2015. “Human Genome Project: Twenty-Five Years of Big Biology.” *Nature* 526 (7571): 29–31.

Hey, Tony, Stewart Tansley, Kristin M. Tolle, and Others. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Vol. 1. Microsoft research Redmond, WA.

National Heart, Lung, and Blood Institute, and Others. 2016. “Charting the Future Together: The NHLBI Strategic Vision.” *Bethesda, MD: NHLBI.*