OSC (Common Fund)

Concept Clearance: New Common Fund Program

TITLE: Artificial Intelligence for BiomedicaL Excellence (AIBLE)

Objective: Generate new biomedically relevant data sets amenable to machine learning analysis at scale

1. Convert ML-friendliness attributes into rubrics and standards that allow planning and evaluation.
2. Create software and hardware to speed annotation and structuring
3. Immediately initiate collaboration with existing projects
4. Generate large multimodal, metadata-complete, available data that exemplify ML-friendliness
5. Use the rubrics to assess and improve select public data sets of biomedical importance.

Funds Available $23M avg cost per year

Program Duration: 7 years

Council Action: Vote on support of Program

Artificial Intelligence for BiomedicaL Excellence (AIBLE)

Draft Common Fund Concept in Response to the Recommendations of the ACD AI Working Group

Program Co-Chairs:

Patti Brennan, NLM
Eric Green, NHGRI
Bruce Tromberg, NIBIB

NIH National Institutes of Health
Office of Strategic Coordination - The Common Fund

Starting point: the ACD WG recommendations…

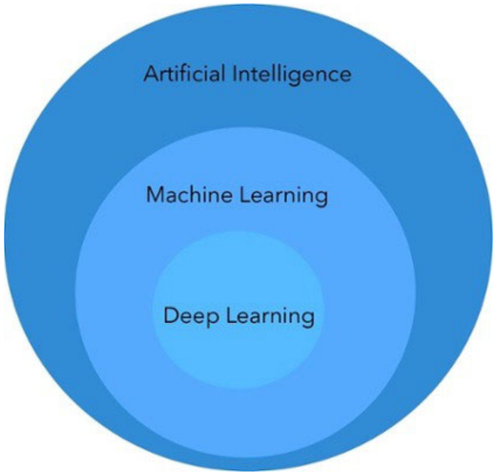
NIH The Common Fund

1 Support flagship **data generation** efforts to propel progress by the scientific community.
2 Develop and publish criteria for **ML-friendly** datasets.
3 Design and apply **“datasheets”** and **“model cards”** for biomedical ML.
https://modelcards.withgoogle.com/about
4 Develop and publish **consent and data access standards for** biomedical ML.
5 Publish **ethical principles for** the use of ML in biomedicine.
6 Develop **curricula to** attract and train ML-BioMed experts.
7 Expand the pilot for ML-focused **trainees and fellows**.
8 Convene cross-disciplinary collaborators.

…and “anti-recommendations” (considered by the WG and rejected):

“We discussed, but are *not* recommending”:

- NIH investment in improving general-purpose ML techniques
- Additional focus on continued use of existing ML tools on existing data
- Investment in scalable secure cloud infrastructure for biomedical data



Additional context: what does the ACD WG mean by "ML-friendly"?

Clear provenance
Well-described
Accessible
Large
Multimodal
Contains perturbations
Longitudinal (time is a perturbation)
Actively learning (data set changes)

Business-as-usual research thinking is *an impediment* to doing this properly.

How do we operationalize the recommendations?

The Common Fund

Core of a Common Fund program. BUT…

1 Support flagship **data generation** efforts to propel progress by the scientific community.

…NEED THIS FIRST

2 Develop and publish criteria for **ML-friendly** datasets.

3 Design and apply **“datasheets”** and **“model cards”** for biomedical ML. https://modelcards.withgoogle.com/about

…AND THIS,

4 Develop and publish **consent and data access standards for** biomedical ML.

…AND THIS.

5 Publish **ethical principles for** the use of ML in biomedicine.

6 Develop **curricula to** attract and train ML-BioMed experts.

7 Expand the pilot for ML-focused **trainees and fellows**.

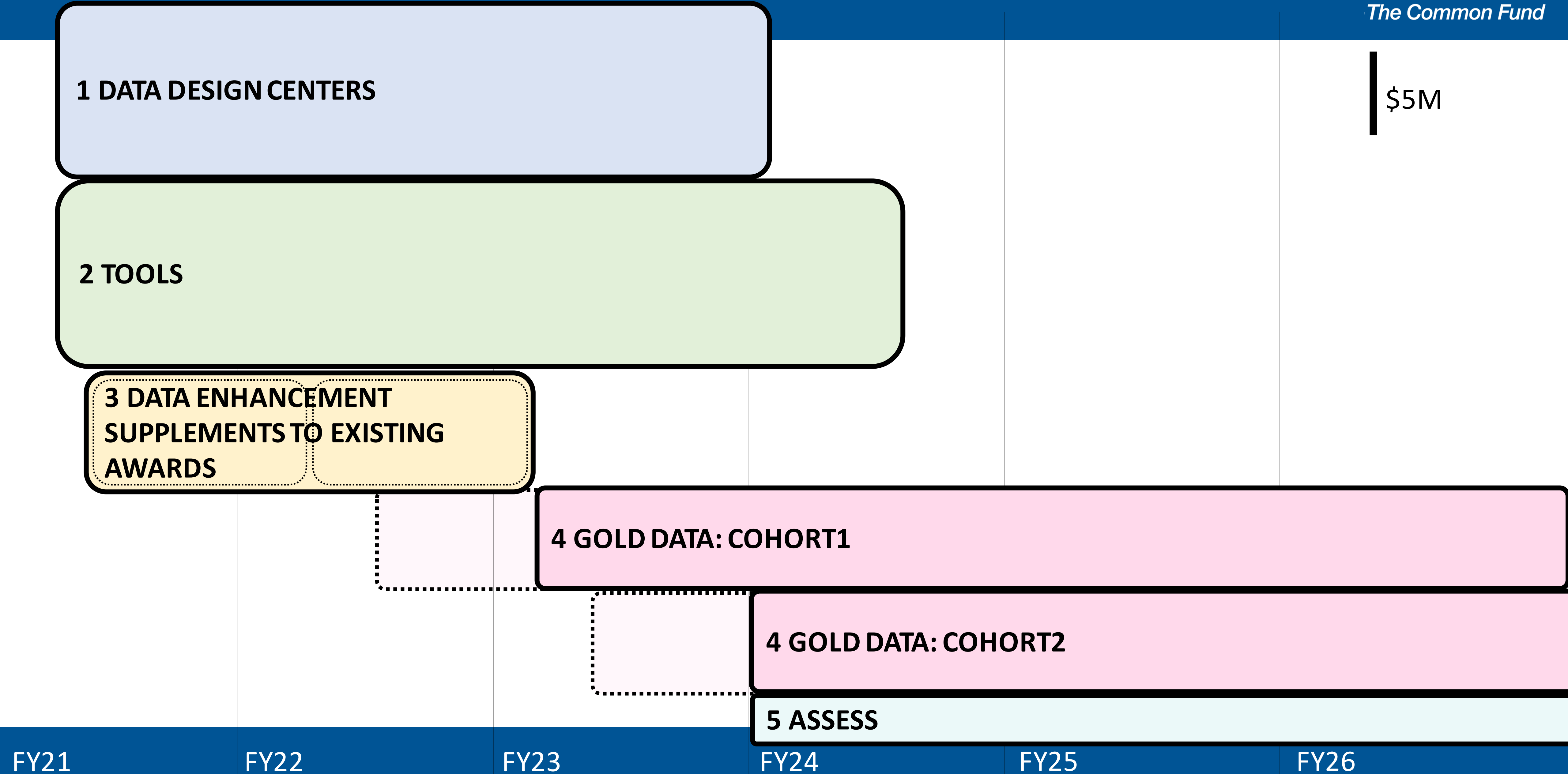
8 Convene cross-disciplinary collaborators.

WITHIN ODSS MANDATE TBD

Draft initiative table

Initiative	Notes	Admin IC
1 DATA DESIGN CENTERS Convert the ACD "ML-ability" recommendations into rubrics that allow evaluation of data sets and plans to generate data sets. Create infrastructure to disseminate tools, host and promote datasets.	Starts in year 1. Main point of contact(s) for NIH. ELSI expertise lives here. Key issues: data provenance, accessibility, representation, privacy.	NHGRI
2 TOOLS Software and firmware tools to accelerate AI-readiness. Instruments that generate AI-ready data, software that speeds annotation and metadata completion, new methods of scientific communication.	Starts in year 1	NIBIB/NLM
3 DATA ENHANCEMENT Immediately initiate new work with supplements to existing projects	Starts in year 1	NHGRI
4 GOLD DATA Generate gold-standard, multimodal, metadata-complete, human data sets that exemplify adherence to the rubrics.	Starts in year 2-3	TBD
5 ASSESS EXISTING DATA Use the rubrics to evaluate and update select existing public data of relevance to biomedical research.	Starts in year 4	NLM

Draft initiative map



1 DATA DESIGN CENTERS

2 TOOLS

3 DATA ENHANCEMENT SUPPLEMENTS TO EXISTING AWARDS

4 GOLD DATA: COHORT1

4 GOLD DATA: COHORT2

5 ASSESS

High-level budget overview

	FY21	FY22	FY23	FY24	FY25	FY26	FY27
Data design centers	10	10	10	10			
Data readiness hardware	6	6	6				
Data readiness software	3	3	3				
Data readiness supplements	2	2					
Gold data			10	20	20	20	10
Assess data				2	2	2	2
TOTAL	21	21	29	32	22	22	12

Overall total: $160M over 7 years

Draft initiative 1 details: Data Design Centers

First year: FY21
Issuing IC: NHGRI

Functions:

- Main point(s) of contact for NIH WG for program.
- Convert the ACD "ML-ability" recommendations into rubrics that allow evaluation of data sets and plans to generate data sets.
- Create/endorse and maintain interoperable knowledge structures (controlled vocabularies/ontologies) for supported data types
- Create infrastructure to disseminate tools, host and promote datasets.
- Agree on and disseminate best practices
- Publish standards for data attributes enabling ethical use of data
- Continually transmit lessons learned

Notes:

RFA should encourage applications to specialize in one or a few fundamental data types.

If appropriate spread doesn't come out of first call, repeat call with strong encouragement to fill gaps.

Draft initiative 2 details: Tools

First year: FY21
Issuing IC: NLM/NIBIB

Functions:

- Create hardware, software, and firmware tools to accelerate generation of AI-ready data.
    - Research instruments that generate annotated data
    - Software that speeds annotation and metadata completion at the point of capture
    - Linking/mapping between new and established ontologies (e.g. SNOMED, LOINC, others)
    - New methods of scientific communication

Draft initiative 3: Data enhancement supplements to existing awards

First year: FY21
Issuing IC: various

Functions:

- Provide dedicated support to existing NIH awardees to build higher-quality data products from their existing raw data
- Support personnel to attend meetings and trainings at the Data Design Centers
- These personnel test and provide feedback on the Tools being created in initiative 2.

Draft initiative 4: Gold Data

First year: FY22 or 23
Issuing IC: TBD

Functions:

- Generate gold-standard, multimodal, metadata-complete, human data sets that exemplify adherence to the rubrics.

  NB: Output of Precision Nutrition program should be aligned with these standards.

- Awardees participate in twice-annual open progress meetings, convened by the Data Design Centers, to share pain points across disciplines and contribute to a common general framework/

- In keeping with the ACD recommendations, data generation plans must be reviewed according to the data design rubrics and not according to *a priori* research goals. Data-forward, not hypothesis-forward.

- Data generation to be balanced to ensure broad utility of the data to biomedical problems.

Draft initiative 5: Assess existing data

First year: FY21
Issuing IC: NLM

Function:

- Use the rubrics to assess and improve select public data sets of biomedical importance.

Partial cloud of data types of interest

The Common Fund

Location data

Genomic data

Social determinants of health

Cellular electrophysiology

Proteomic data

Citations

Movement, kinematics

Radiological images

Health outcomes

Behavioral rating scales

Cellular images

Serology

Patient-reported outcomes

Nutrition

Height, weight

Screenomic data

What will this program produce?

Acadia National Park, Maine

© Greg A. Hartford, AcadiaMagic.com

Rubrics that allow evaluation of datasets (and plans to generate datasets) for ML-readiness

Tools to accelerate the creation of ML-ready data sets (intelligent annotators, metadata-filling instruments)

Infrastructure to host, disseminate, and promote tools and datasets

A group of AI-ready datasets, ethically sourced, clean and available

AIBLE Working Group

NIBIB
Bruce Tromberg
Grace Peng

NLM
Patti Brennan
Anna Calcagno

NHGRI
Eric Green
Carolyn Hutter
Shurjo Sen

NIDDK
Danny Gossett

NCMRR
Alison Cernich
Theresa Cruz

NCCIH
Helene Langevin
Wendy Weber

NIMH
Adam Thomas
Holly Lisanby

NIA
P. Bhattacharyya

OBSSR
Christine Hunter

NIMHD
Deborah Duran

OSC
Gene Civillico
Wendy Knosp
Kate Nicholson

ODSS
Susan Gregurick
Jess Mazerik

Interagency interest

DOE
Laura Biven

FDA/CDRH Digital Health

Updated 5 May 2020