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
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
Starting point: the ACD WG recommendations...

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.
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?

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](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.
<table>
<thead>
<tr>
<th>Initiative</th>
<th>Notes</th>
<th>Admin IC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 DATA DESIGN CENTERS</strong></td>
<td>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.</td>
<td>NIHGRI</td>
</tr>
<tr>
<td><strong>2 TOOLS</strong></td>
<td>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.</td>
<td>NIBIB/NLM</td>
</tr>
<tr>
<td><strong>3 DATA ENHANCEMENT</strong></td>
<td>Immediately initiate new work with supplements to existing projects</td>
<td>NIHGRI</td>
</tr>
<tr>
<td><strong>4 GOLD DATA</strong></td>
<td>Generate gold-standard, multimodal, metadata-complete, human data sets that exemplify adherence to the rubrics.</td>
<td>TBD</td>
</tr>
<tr>
<td><strong>5 ASSESS EXISTING DATA</strong></td>
<td>Use the rubrics to evaluate and update select existing public data of relevance to biomedical research.</td>
<td>NLM</td>
</tr>
<tr>
<td></td>
<td>FY21</td>
<td>FY22</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Data design centers</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Data readiness hardware</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Data readiness software</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Data readiness supplements</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Gold data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assess data</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>21</td>
<td>21</td>
</tr>
</tbody>
</table>

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.
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

- Genomic data
- Proteomic data
- Radiological images
- Cellular images
- Patient-reported outcomes
- Screenomic data
- Location data
- Social determinants of health
- Citations
- Health outcomes
- Serology
- Height, weight
- Cellular electrophysiology
- Movement, kinematics
- Behavioral rating scales
- Nutrition
What will this program produce?

- 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

<table>
<thead>
<tr>
<th>Institute</th>
<th>Interagency Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIBIB</td>
<td>Bruce Tromberg</td>
</tr>
<tr>
<td></td>
<td>Grace Peng</td>
</tr>
<tr>
<td>NLM</td>
<td>Patti Brennan</td>
</tr>
<tr>
<td></td>
<td>Anna Calcagno</td>
</tr>
<tr>
<td>NHGRI</td>
<td>Eric Green</td>
</tr>
<tr>
<td></td>
<td>Carolyn Hutter</td>
</tr>
<tr>
<td></td>
<td>Shurjo Sen</td>
</tr>
<tr>
<td>NIDDK</td>
<td>Danny Gossett</td>
</tr>
<tr>
<td>NCMRR</td>
<td>Alison Cernich</td>
</tr>
<tr>
<td></td>
<td>Theresa Cruz</td>
</tr>
<tr>
<td>NCCIH</td>
<td>Helene Langevin</td>
</tr>
<tr>
<td></td>
<td>Wendy Weber</td>
</tr>
<tr>
<td>NIMH</td>
<td>Adam Thomas</td>
</tr>
<tr>
<td></td>
<td>Holly Lisanby</td>
</tr>
<tr>
<td>NIA</td>
<td>P. Bhattacharyya</td>
</tr>
<tr>
<td>OBSSR</td>
<td>Christine Hunter</td>
</tr>
<tr>
<td>NIMHD</td>
<td>Deborah Duran</td>
</tr>
<tr>
<td>OSC</td>
<td>Gene Civillico</td>
</tr>
<tr>
<td></td>
<td>Wendy Knosp</td>
</tr>
<tr>
<td></td>
<td>Kate Nicholson</td>
</tr>
<tr>
<td>ODSS</td>
<td>Susan Gregurick</td>
</tr>
<tr>
<td></td>
<td>Jess Mazerik</td>
</tr>
</tbody>
</table>

### Interagency interest

- **DOE**
  - Laura Biven
  - Health

- **FDA/CDRH Digital**

*Updated 5 May 2020*