Workforce and Research Needs: Biomedical Big Data Science

Valerie Florance, PhD
Associate Director for Extramural Programs
National Library of Medicine, NIH/DHHS
florancev@mail.nih.gov
Topics for Today

- NLM Today
- New Vision for NLM’s Role at NIH
- A Vision for the Global Digital Research Enterprise
- Data Science at NIH Today
- Precision Medicine Initiative
- Other Funding Opportunities
When you think of NLM...

- PubMed and PubMed Central
- MedLine Plus
- Clinical Trials.gov

- Intramural Research Fellowships
- UMLS
- Biomedical Informatics Short Course
$400 million, 240 active areas of research & service
For Librarians: Associate Fellowship Program
http://www.nlm.nih.gov/about/training/associate/index.html

Associate Fellowship Program for Librarians

The National Library of Medicine Associate Fellowship Program is a one-year postgraduate training fellowship at the NLM in Bethesda, Maryland, with an optional second year program component. The program is designed to provide a broad foundation in health sciences information services, and to prepare librarians for future leadership roles in health sciences libraries and in health services research.

During the program, Associates will:
- Learn about the collections, databases, terminologies, research and development, training,
Short Course in Biomedical Informatics
http://www.nlm.nih.gov/about/training/bioinformatics.html

NLM Georgia Biomedical Informatics Course

The Robert B. Greenblatt, M.D. Library, Georgia Regents University hosts the NLM Georgia Biomedical Informatics Course, funded by the National Library of Medicine. The course offers participants a week-long immersive experience in biomedical informatics and provides continuing education to health care professionals interested in the application of computer technologies to medicine. The Georgia course will be held at Brasstown Valley.

Biomedical administrators, faculty and others who can become change agents for their institutions are strongly encouraged to apply. All costs for the course including travel, housing, and per diem are supported by NLM. The application is open to US citizens and US permanent residents.

The course will provide attendees a diverse set of skills and experiences incorporating concepts, theories and building blocks of biomedical informatics; ability to use informatics for solving current health care challenges; application and policies related to computer technologies and information science; hands-on experience during evening workshops; and networking with nationally known bioinformatics educators and thought leaders.
141 active extramural research grants & training awards, $51 million
NLM Informatics Training

- NLM Pre- and Post-doctoral training: 14 Universities produce graduates who are prepared for research careers in biomedical informatics
- NLM intramural fellowships
- Individual NRSA Pre-doctoral Fellowships F30 and F31
NLM Grant Supplements for Informationist Services

• Funds an NIH-funded scientist to integrate an information specialist into the research team
  – Informationists are information specialists, usually health sciences librarians, who have graduate training and practical experience in biomedical, behavioral or biological sciences and in library and information sciences/informatics
  – Informationists who work as team members with research scientists and health professionals are sometimes called in-context or ‘embedded’ information specialists.
  – 31 awards since 2012, with 50 informationists

Schmidt, Brian
Role of proteases and peptides in cancer pain
New York University, 3 R01 DE-019796-03S1 2012 award

Aileen McCrillis, Rich McGowan and Alisa Surkis, New York University Health Sciences Library, worked with 10 investigators in New York and Boston. They assisted with improving the team’s search queries for NCBI data sources; identifying a reference management approach for use by the multiple participants; and developing a web-based data management system for each molecule identified, to enhance workflow between the 2 sites.

Data Sharing and Management Snafu in 3 Short Acts
• https://www.youtube.com/watch?v=N2zK3sAtr-4
NLM Informationist Grant Supplements – 2014 award

• Liebold, Lori J
  Susceptibility to and Release from Masking in Infancy and Childhood UNC-CH

• 3 R01 DC-011038-04S1

• Barbara Renner, Barrie Hayes and Kathleen McGraw, UNC Health Sciences Library will conduct systematic reviews on auditory and visual development to find research gaps; assess current data management processes; introduce new tools & best practices; and facilitate improved study recruitment through the use of communication technologies.
2016 NLM Funding Opportunities

- Re-competition of NLM’s Institutional Research Training Programs for Biomedical Informatics
- Re-issuance of RFAs for the Administrative Supplements for Informationist Services and Grants for Information Resources to Reduce Health Disparities
- Regular deadlines for other, including R01 and R21 research, Fellowships and Career awards
New Vision for NLM

• Donald A.B. Lindberg, MD retired as NLM Director in March 2015 after 30 yrs

• NIH Advisory Committee to the NIH Director formed a Working Group on NLM – to propose a vision for the NLM’s future

• Request for Information issued, Feb 13–March 13, 2015

• Input requested on 5 questions

• 649 respondents
### Respondent Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librarian</td>
<td>38.7%</td>
</tr>
<tr>
<td>Researcher</td>
<td>11.1%</td>
</tr>
<tr>
<td>Emergency Personnel</td>
<td>9.4%</td>
</tr>
<tr>
<td>Educator</td>
<td>7.4%</td>
</tr>
<tr>
<td>Historian</td>
<td>7.6%</td>
</tr>
<tr>
<td>Healthcare Professional</td>
<td>5.1%</td>
</tr>
<tr>
<td>Environmental Science / Toxicology</td>
<td>5.2%</td>
</tr>
<tr>
<td>Other</td>
<td>4.8%</td>
</tr>
<tr>
<td>Unknown</td>
<td>3.7%</td>
</tr>
<tr>
<td>Public Health</td>
<td>2.8%</td>
</tr>
<tr>
<td>Advocate</td>
<td>2.9%</td>
</tr>
<tr>
<td>Publisher</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

N=649 respondents
## Institution Types

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>24.3%</td>
</tr>
<tr>
<td>Medical School</td>
<td>12.0%</td>
</tr>
<tr>
<td>Hospital</td>
<td>11.9%</td>
</tr>
<tr>
<td>Company</td>
<td>10.9%</td>
</tr>
<tr>
<td>State / Local Government</td>
<td>9.2%</td>
</tr>
<tr>
<td>Association</td>
<td>7.6%</td>
</tr>
<tr>
<td>Federal Government*</td>
<td>5.4%</td>
</tr>
<tr>
<td>K-12</td>
<td>4.0%</td>
</tr>
<tr>
<td>Unknown</td>
<td>3.9%</td>
</tr>
<tr>
<td>Public Library</td>
<td>2.6%</td>
</tr>
<tr>
<td>Military*</td>
<td>2.0%</td>
</tr>
<tr>
<td>NIH*</td>
<td>1.8%</td>
</tr>
<tr>
<td>International</td>
<td>1.7%</td>
</tr>
<tr>
<td>RML</td>
<td>1.5%</td>
</tr>
<tr>
<td>Community College</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

*N=649 respondents*
RFI Questions 1-4

Current NLM elements that are of the most, or least, value to...

1. the research community...
2. health professionals...
3. patients and the public...
4. libraries, publishers, organizations, companies, and individuals who use NLM data, software tools, and systems in developing and providing value-added or complementary services and products...

and future capabilities that would be needed...
Range of Comments

- Bibliographic
- Drug Information
- ClinicalTrials.gov
- Data Science/Sharing
- Standards/Vocabulary
- EHR / LHC
- Informatics Research
- Informatics Training
- Collections–Print, Digital
- Training: DB, Librarian, NN/LM
- History of Medicine
- Consumer Health
- Env/Tox/Disaster
- Health Services Research
- APIs, E-utils, LinkOut
- NLM Overall
### Value to Research Community
- PubMed
- HMD
- PMC
- NCBI
- ClinicalTrials
- MeSH
- DOCLINE
- NN/LM
- UMLS
- TOXNET

### Value to Health Professionals
- PubMed
- WISER
- TOXNET
- PMC
- MedlinePlus
- ClinicalTrials
- MedlinePlus

### Value to Patients & Public
- MedlinePlus
- PubMed
- ClinicalTrials
- NN/LM
- HMD
- PMC
- DB Training
- K-12
- HouseholdProducts
- PM Health

### Value to Librarians, Publishers, Developers
- PubMed
- NN/LM
- DOCLINE
- PMC
- MeSH
- MedlinePlus
- DB Training
- K-12
- ClinicalTrials
- HMD

**Top 10 mentions**
RFI Question 5 – “Future”

How NLM could be better positioned to help address the broader and growing challenges associated with:

• Biomedical informatics, “big data,” and data science;
• Electronic health records;
• Digital publications; or
• Other emerging challenges/elements warranting special consideration.
Themes for Future

Biomedical Informatics
- Increase funding greatly
- Broad and deep portfolio
- Fundamental and applied
- Balance clinical, bio-, and public health informatics
- Expand training significantly
- Concentrate at NLM and/or
- Distribute broadly at NIH

Big Data
- Repository hosting, linking
- Discovery index, data catalog
- Data and software citation
- Data fraud protection tools
- Partner in standards development, best practices for storing and linking data
- Partner in support of data management plans
Themes for Future, cont’d

**EHR Standards**
- Continue leadership in health data standards, harmonization, adoption
- Continue current projects for vocabulary standards and value sets
- Collaborate/partner for new standards development
- Explore private sector role in developing new vocabulary standards and value sets

**EHR Enhancement**
- Create test bed EHR system
- Host a database for protected EHR data
- Create central, de-identified linked patient databases for research
- Partner with outside groups that have high-performance computing capabilities to analyze data for patterns
- Develop and fund clinical decision support systems integrated with EHR
Acknowledgments for RFI Analysis

• Barbara Rapp, Ph.D., OHIPD
• Dana Casciotti, Ph.D., OHIPD
• Maria Collins, Library Operations
• Mary Ann Leonard, OCPL
• Kathel Dunn, Ph.D., Library Operations

• Becky Baltich-Nelson, Associate Fellow
• Loan Nguyen, Associate Fellow
• Tyler Nix, Associate Fellow
New Vision for NLM: 6 Recommendations

1. NLM must continually evolve to remain a leader in assimilating and disseminating accessible and authoritative biomedical research findings and trusted health information to the public, healthcare professionals, and researchers worldwide.
2. NLM should **lead efforts to support and catalyze open science, data sharing, and research reproducibility**, striving to promote the concept that biomedical information and its transparent analysis are public goods.
New Vision for NLM

3. NLM should be the intellectual and programmatic epicenter for data science at NIH and stimulate its advancement throughout biomedical research and application.
4. NLM should strengthen its role in fostering the future generation of professionals in biomedical informatics, data science, library sciences, and related disciplines through sustained and focused training efforts.
5. NLM should **maintain, preserve, and make accessible the nation’s historical efforts in advancing biomedical research and medicine**, thereby ensuring that this legacy is both safe and accessible for long-term use.
6. New NLM leadership should **evaluate what talent, resources, and organizational structures are required** to ensure NLM can fully achieve its mission and best allocate its resources.
My Dream Job at NIH
3. NLM should be the intellectual and programmatic epicenter for data science at NIH and stimulate its advancement throughout biomedical research and application.
Recommendation 3, con’t.

• **NLM should**
  
  – *become the programmatic and administrative home for the BD2K Initiative*
  
  – *lead the coordination of data science programs (and programs with large data science components) conducted at other NIH Institutes/Centers, in order to maximize synergies and minimize redundancies.*
  
  – *promulgate intramural and/or extramural expertise, knowledge generation and dissemination, and leadership in areas of data science that are critical to the NIH mission [e.g., precision medicine]*
• September 17, Report on the Precision Medicine Initiative Cohort Program (PMI-CP) accepted by NIH Director
  • [http://acd.od.nih.gov/meetings.htm](http://acd.od.nih.gov/meetings.htm)

• See videocasts of workshop meetings that framed the ideas, press releases and other relevant resources at the NIH Precision Medicine Initiative page at

• ...the PMI-CP will enroll one million or more volunteers that are inclusive of the diversity of the U.S. population, and will follow their health and clinical outcomes over time.

• The PMI-CP...will comprise an accessible resource for researchers and participants to work in partnership to accelerate our understanding of health and disease. Thousands of researchers will be able to explore the underpinnings of disease, while participants will have the opportunity to use their health data in innovative ways for their own betterment.
Data Science at NIH

• Advisory Committee to the Director of NIH Working Group on Data and Informatics Working Group (DIWG) June 2012

• Recommended that NIH
  – Promote data sharing
  – Support development, implementation, evaluation, maintenance and dissemination of informatics methods and applications
  – Build capacity by training the workforce in the relevant quantitative sciences such as bioinformatics, biomathematics, biostatistics and clinical informatics
  – Provide a serious, substantial and sustained funding commitment to enable these recommendations

• [http://acd.od.nih.gov/diwg.htm](http://acd.od.nih.gov/diwg.htm)
“It is clear that modern interdisciplinary team science requires an infrastructure and a set of policies and incentives to promote data sharing, and it needs an environment that fosters the development, dissemination, and effective use of computational tools for the analysis of datasets whose size and complexity have grown by orders of magnitude in recent years.

Confidentiality issues, as well as fundamental differences between basic science and clinical investigation, create real challenges for the successful integration of molecular and clinical datasets.” (p. 9).
Findable, Accessible, Interoperable, Reusable Data

• NIH Associate Director for Data Science, Phil Bourne has described a connected ecosystem of biomedical data, information and knowledge. At the heart of his long-term vision is the idea of a Commons.

• “... a pilot experiment in the efficient storage, manipulation, analysis, and sharing of research output, from all parts of the research lifecycle.”
  – Storing data sets along with metadata about them, tools for analyzing and visualizing the data and compute power.
  – Assigning unique identifiers that can be used for finding them, for attribution and for linking to relevant resources. [https://pebourne.wordpress.com/, Oct 7 2014].
Vision for Sustainable Ecosystem

Investigator Managed Biomedical Data sets & Data resources

Community Managed Data Resources that serve large communities

NIH Data Resources

Policies for Data Sharing, Preservation

NIH Investment Models for Different Preservation Levels

Long Term Archival

Other

Findable, Accessible, Interoperable and Reusable
https://datascience.nih.gov/
Strategic Areas & Resource Distribution

FY16 $97M

Leadership: 24%
Sustainability: 19%
Workforce Development & Diversity: 19%
Policy & Process: 57%
Discovery & Innovation: 57%
Goal: To enable major scientific discovery and innovation through the BD2K Initiative

Sustainability
Workforce Development & Diversity
Discovery & Innovation
Policy & Process
Leadership

Funded in FY14:
- 12 BD2K Centers of Excellence
- Data Discovery Index Consortium

Funding in FY15:
- Centers Coordinating Center
- Targeted software awards

FY16 or beyond:
- Building Collaborations to harness new skill sets
- Standards Coordination Center
- Software index
## NIH BD2K Centers of Excellence

<table>
<thead>
<tr>
<th>Center Grant Title</th>
<th>Center Primary Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data for Discovery Science</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>Center for Big Data in Translational Genomics</td>
<td>University of California Santa Cruz</td>
</tr>
<tr>
<td>Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data</td>
<td>Pittsburgh</td>
</tr>
<tr>
<td>Center for Expanded Data Annotation and Retrieval (CEDAR)</td>
<td>Stanford University</td>
</tr>
<tr>
<td>Center for Predictive Computational Phenotyping</td>
<td>University of Wisconsin – Madison</td>
</tr>
<tr>
<td>Center of Excellence for Mobile Sensor Data-to-Knowledge</td>
<td>University of Memphis</td>
</tr>
<tr>
<td>Community Effort to Translate Protein Data to Knowledge: An Integrated Platform</td>
<td>University of California Los Angeles</td>
</tr>
<tr>
<td>ENIGMA Center for Worldwide Medicine, Imaging, and Genomics</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>KnowEng, a Scalable Knowledge Engine for Large-Scale Genomic Data</td>
<td>University of Illinois Urbana-Champaign</td>
</tr>
<tr>
<td>National Center for Mobility Data Integration to Insight</td>
<td>Stanford University</td>
</tr>
<tr>
<td>Patient-Centered Information Commons</td>
<td>Harvard University Medical School</td>
</tr>
</tbody>
</table>
Goal: To strengthen the ability of a diverse biomedical workforce to develop and benefit from data science

- Strengthening a diverse biomedical workforce to utilize data science
  - BD2K funding of Short Courses and Open Educational Resources
- Building a diverse workforce in biomedical data science
  - BD2K Training programs and Individual Career Awards

**Discovery of Educational Resources**

- **BD2K Training Coordination Center**
  - Fostering Collaborations
    - BD2K Training Coordination Center, NSF/NIH IDEAs Lab
  - Expanding NIH Data Science Workforce Development Center
    - Local courses, e.g. Software Carpentry
Training and Learning Resources

• Predoctoral Training – 6 awards to date
  – 4 new awards including Kosorok, MR and Forest, MG. UNC-CH. Big Data to Knowledge Training Program T32 CA201159
  – 2 awards are supplements to existing NLM Training Programs
  – More to be reviewed and awarded in FY 2016

• Postdoctoral
  – K01 Career Awards – 20 awards to date. Postdocs, MDs and mid-career investigators
Courses and Resources

Data Science – 19 awards to date
• Biomedical Data Science Online Curriculum (Harvard)
• MOOCs for training in neuroimaging and genomic big data analysis (JHU)
• OHSU Informatics Analytics BD2K Skill Course
• Summer Institute for Statistics of Big Data (U Washington)

Research Data Management
• MOOC: Demystifying Biomedical Big Data: User’s Guide (Georgetown)
• MOOC: Best Practices in Research Data Management (U Mass)
• Enabling Data Science in Biology (Rutgers)
• Preparing Medical Librarians to Understand and Teach RDM
• Training and Tools for Informationists to Facilitate Sharing of Next-Gen Sequence Data
BD2K Upcoming Events

• Call for hackathon event proposals related to biomedical big data, inviting proposals to host public hackathon events to speed development & application of big data software tools. Applications are due Oct 15, 2015.
  

• Center for Predictive Computational Phenotyping, U Wisconsin-Madison. **Privacy Symposium: Big Privacy: Policy Meets Data Science** Oct 15, 2015 1:00-5:00pm CDT
  
  – Biomedical research data may be subject to different privacy laws and regulations depending on the type of institution holding or using the data, the type of data, who funds the research, the state in which the research is conducted, and other factors. A live webcast of this event will be available.
2016 Funding

• BD2K Development of Software Tools and Methods in Targeted Areas of High Need – closing date Oct 06, 2015

• Courses for Skills Development in Biomedical Big Data Science – closing date April 1, 2016

• Open Educational Resources – closing date April 1, 2016
Activities in 2016

New BD2K Activities

- **Reference Datasets**
  - Will move important, FAIR, digital resources into the cloud to support increased access and utility.
  - Will release an RFI to inform details of an FOA.
  - Will follow up and formalize activities started in the Commons administrative supplements.

- **Sustainability of Data Repositories**
  - BD2K sustainability group is doing a financial and portfolio analysis of digital data repositories across NIH, and has been studying current sustainability approaches of such repositories.
  - BD2K is working in conjunction with NIGMS, NHGRI, and BISTI to develop FOA on this topic.
  - Will follow up on initial activities of the interoperability supplements.

- **Software Hardening Resource**
  - BD2K is standing up a group to develop a proposal to ensure useful software can develop from useful academic grade prototypes to more robust, commons-compliant tools.

*FAIR = Findable, Accessible, Interoperable, Reusable*

New BD2K Workshops

- **NIH Common Data Elements (CDE) Workshop**
  - in conjunction with NCI, NLM, and BMIC

- **Academic deans and data science career paths workshop**
  - in conjunction with NSF

- **Data Science of Citizen Science**
  - Possibly in conjunction with the Heart of Data Science (Ping) Center

- **National Academies (CATS) workshop on big data inference**
  - in conjunction with NSF/CISE

- **National Academies (CSTB) workshop on data science curriculum development**
  - in conjunction with NSF/CISE
NIH Common Fund – High Risk High Reward Research Grant Program

• **Pioneer Awards**
  – substantially different scientific directions from those already being pursued
  – 5 years, $700K direct costs + F&A

• **New Innovator Awards**
  – any scientific area relevant to the mission of NIH (biological, behavioral, clinical, social, physical, chemical, computational, engineering, and mathematical sciences), early stage investigators
  – 5 YEARS, $300K direct costs + F&A

https://commonfund.nih.gov/highrisk/index
UNC-CH: 1000+ active awards, $552 Million
UNC-CH “informatics” 100+ active awards, $83 Million
Finding out more about NLM and BD2K Grants

NLM Grant Program Contacts

- Dr. Hua-Chuan Sim – Clinical and public health informatics
  simh@mail.nih.gov
- Dr. Alan VanBiervliet – Consumer health informatics, information sciences & special programs – alan.vanbiervliet@nih.gov
- Dr. Jane Ye – Bioinformatics and translational informatics - yej@mail.nih.gov
- Dr. Valerie Florance – NLM and BD2K training
  florancev@mail.nih.gov

NIH Grants Information

- [http://projectreporter.nih.gov](http://projectreporter.nih.gov) - expenditures & results for NIH funded grants
- [http://datascience.nih.gov/bd2k](http://datascience.nih.gov/bd2k) - lists awards, events and open funding announcements
- [http: www.nlm.nih.gov/ep](http: www.nlm.nih.gov/ep) - NLM grants site, with lists of awards and active grant programs
Vacancy Announcement

Department of Health and Human Services (DHHS)
National Institutes of Health (NIH)

DIRECTOR

National Library of Medicine

This position offers a unique and exciting opportunity for an exceptional leader to serve as the chief visionary for NLM and lead all aspects of this highly complex organization. The Director, NLM, is responsible for the management and direction of the world’s largest biomedical library, and will lead the organization in championing data science efforts. Scientists, health professionals, and the general public rely on robust, uninterrupted access to NLM’s digital resources. As an international leader in information innovation, the NLM also supports and conducts computational biology, data science and standards, biomedical communications, and health information technology.

The Director, NLM, serves as the principal advisor to the Director, NIH, concerning matters related to biomedical informatics and access to biomedical information with responsibilities for program planning, implementation and evaluation, keeping the NIH Director abreast of NLM developments, accomplishments, and needs as the relate to the overall mission of NIH. The Director, NLM, maintains continuous contact with outside industry and private interest groups, members of Congress, officials of other Federal agencies, and officials of foreign governments for the purpose of explaining the program, mission, and goals of informatics research and information access at NIH.

National Library of Medicine