Big Data Challenges in the Life Sciences …

This speaks to something more fundamental that more data …

It speaks to new methodologies, new skills, new emphasis, new cultures, new modes of discovery …

New types of funding
Conversation Cards

- A brief historical perspective
- What could happen in the future
- The implications for this future
- NIH initiatives in this landscape
  - Big Data to Knowledge (BD2K)
  - Precision Medicine
Laying the Foundation for Open Access: HGP, Bermuda, 1996
The History of Computational Biomedicine According to Bourne

Searls (ed) The Roots in Bioinformatics Series *PLOS Comp Biol*

1980s 1990s 2000s 2010s 2020

**Discipline:**
- Unknown
- Expt. Driven
- Emergent
- Over-sold
- A Service
- A Partner
- A *Driver*

**The Raw Material:**
- Non-existent
- Limited /Poor
- More/Ontologies
- Big Data/Siloed
- Open/Integrated

**The People:**
- No name
- Technicians
- Industry recognition
- data scientists
- Academics
What could happen in the future?
We are at a Point of Deception …

- Evidence:
  - Google car
  - 3D printers
  - Waze
  - Robotics
  - Sensors

From: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies by Erik Brynjolfsson & Andrew McAfee

The 6 Ds of Exponentials: Digitalization, Deception, Disruption, Demonetization, Dematerialization, and Democratization

Source: Peter H. Diamandis, www.abundancehub.com
Example - Photography

- **Digitization**

  - Megapixels & quality improve slowly; Kodak slow to react

- **Deception**

  - Instagram, Flickr become the value proposition

- **Disruption**

  - Digital media becomes bona fide form of communication
  -phones replace cameras

- **Dematerialization**

  - Film market collapses; Kodak goes bankrupt

- **Demonetization**

  - Kodak slow to react

- **Democratization**

  - Volume, Velocity, Variety
We Are At a Point of Deception
The 6D Exponential Framework

- Digitization of Basic & Clinical Research & EHR’s
- Deception
- Disruption
- Demonetization
- Dematerialization
- Democratization
- Patient centered health care
- Open science
What Are Some Implications of Such a Future?

- Open collaborative science becomes of increasing importance
- The value of data and associated analytics becomes of increasing value to scholarship
- Opportunities exist to improve the efficiency of the research enterprise and hence fund more research
- Cooperation between funders will be needed to sustain the emergent digital enterprise
- Current training content and modalities will not match supply to demand
- Precision medicine is indeed a reality
What Are Some Implications of Such a Future?

- Open collaborative science becomes of increasing importance
- The value of data and associated analytics becomes of increasing value to scholarship
- Opportunities exist to improve the efficiency of the research enterprise and hence fund more research
- Cooperation between funders will be needed to sustain the emergent digital enterprise
- Current training content and modalities will not match supply to demand
- Precision medicine is indeed a reality
“And that’s why we’re here today. Because something called precision medicine … gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen.”

President Barack Obama
January 30, 2015
Precision Medicine Initiative

- **National Research Cohort**
  - >1 million U.S. volunteers
  - Numerous existing cohorts (many funded by NIH)
  - New volunteers

- Participants will be centrally involved in design and implementation of the cohort

- They will be able to share genomic data, lifestyle information, biological samples – all linked to their electronic health records
An Example of That Promise: Comorbidity Network for 6.2M Danes Over 14.9 Years

Jensen et al 2014 Nat Comm 5:4022
Conversation Cards

- A brief historical perspective
- What could happen in the future
- The implications for this future
- NIH initiatives in this landscape
  - Big Data to Knowledge (BD2K)
  - Precision Medicine
What is the NIH Doing to Fulfill That Promise?
ADDS Mission Statement

To foster an open ecosystem that enables biomedical* research to be conducted as a digital enterprise that enhances health, lengthens life and reduces illness and disability.

* Includes biological, biomedical, behavioral, social, environmental, and clinical studies that relate to understanding health and disease.
Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K)

Santosh Kumar, Ph.D.
Director, MD2K Center of Excellence
Professor & Moss Chair of Excellence in Computer Science
University of Memphis

https://datascience.nih.gov/bd2k/funded-programs/centers
MD2K Applications – CHF and Smoking

Detect → Predict → Adapt

- CHF
- Smoking

- Stress
- Trigger
- Stress

- Marlboro
- Cigarettes

- Smartphone
- Phone
- Craving

- Heart
- Stethoscope

- Money
- Wallet

- Pizza
- Fast food

NIH
Example: BD2K Center
Working Across Strategic Areas

- Research Objects in the Commons
- Over 100 Public Lectures
- Collaboration with a Minority Institution
- Voxel Wide Genome Scanning
- MRI standardization
- Genomic Data Sharing
- Policy
- 185 Institutions Involved
Elements of The Ecosystem

- Community
- Policy
- Infrastructure
  - Sustainability
  - Collaboration
  - Training
Elements of The Ecosystem

- Community
- Policy
- Infrastructure

Virtuous Research Cycle
- Sustainability
- Collaboration
- Training
Policies – Now & Forthcoming

- Data Sharing
  - Genomic data sharing announced
  - Data sharing plans on all research awards
  - Data sharing plan enforcement
    - Machine readable plan
    - Repository requirements to include grant numbers

Policies - Forthcoming

- Data Citation
  - Goal: legitimize data as a form of scholarship
  - Process:
    - Machine readable standard for data citation (done)
    - Endorsement of data citation for inclusion in NIH bib sketch, grants, reports, etc.
    - Example formats for human readable data citations
    - Slowly work into NLM/NCBI workflow
Infrastructure - The Commons
The Commons: Components

- Discoverability
  - Metadata / unstructured data
- Open APIs
  - data & tools
- Digital Objects
  - with unique identifiers
- Containers
  - for packaging applications
- Digital Object Compliance
- Computing Platform
  - Cloud and HPC
The Commons

*Digital Object Compliance: FAIR*

- **Attributes of digital objects in the Commons**
- **Initial Phase**
  - Unique digital object identifiers of some type
  - A minimal set of searchable metadata
  - Physically available in a cloud based *Commons* provider
  - Clear access rules (especially important for human subjects data)
  - An entry (with metadata) in one or more indices

- **Future Phases**
  - Standard, community based unique digital object identifiers
  - Conform to community approved standard metadata for enhanced searching
  - Digital objects accessible via open standard APIs
  - Are physically and logical available to the commons
# The Commons: Evaluation Pilots Underway

<table>
<thead>
<tr>
<th>Evaluation Criteria</th>
<th>Pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation</td>
<td>BD2K Centers</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Model organism databases</td>
</tr>
<tr>
<td>Computation on Big Data</td>
<td>HMP data and tools in the cloud</td>
</tr>
<tr>
<td>Multi-cloud accessibility</td>
<td>NCI cloud pilots &amp; genomic data commons</td>
</tr>
<tr>
<td>Business model</td>
<td>Supply and demand via credits</td>
</tr>
</tbody>
</table>
A Quick Word on Training....
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
I not only use all the brains I have, but all I can borrow.

– Woodrow Wilson
The Team
NIH... Turning Discovery Into Health

philip.bourne@nih.gov
https://datascience.nih.gov/