Big Data – Big Deal? Big Problems? Big Promise?

Carlie Williams, PhD, MPH
Chief Epidemiology, BSP, DAIDS, NIAID
February 11, 2020
Unlocking the Power of Big Data in Health
University of South Carolina

Big Data is visualized in so many ways...all of them blue and with numbers and lens flare

Hobbled by our own success?

If we cannot get a handle on the scientific body of knowledge, how can we do science?

- We are generating so much data and databases, so many articles, so many theories, so many computer programs and tools… *(Big Data)*
- That we are losing the "systematic organization of knowledge" that is the foundation of the scientific enterprise. *(The problem that Big Data causes)*
- So we need help turning the morass of data back into a systematic organization of knowledge … thus **data science**. (*How to address the problem that Big Data causes)*

What is Data Science?

Our problem: Science is accumulating new data at a rate that exceeds our capacity to extract value from it.

- An interdisciplinary field that helps **extract knowledge & insights from data**.
  - Combines expertise in statistics, informatics, computer science, and data management.
  - Knowledge of biomedical research allows for useful knowledge management and mining.
- Data Science can help biomedical researchers benefit more from the Big Data we are creating.

Today's biomedical researcher

- Dr. Juanita Doe wants to identify possible shared mechanisms underlying dementia and hypertension.
  - Epidemiological studies and other data indicate co-occurrence and possible commons mechanisms.
  - She suspects common pathways, particularly cytokines, may play a major role.
- But the existing genomic and clinical data is fragmented, difficult to find, and difficult to integrate.
  - So she has to base her developing hypotheses on only part of the existing knowledge (that which is available to her)
  - This introduces gaps and biases in her hypotheses, which makes her grant applications and papers weaker in review.
  - And slows down possible identification of shared mechanisms and possible therapeutic targets.

But is she asking the right question?

Dementia and hypertension are fundamentally multifactorial

- Genetics - Intelligence
- Environment
  - Childhood abuse
  - Adult microclimates
- Educational attainment
- Diet
- Exercise
- Access to health care
- etc

You shouldn't control for, negate the influence, of confounding variables

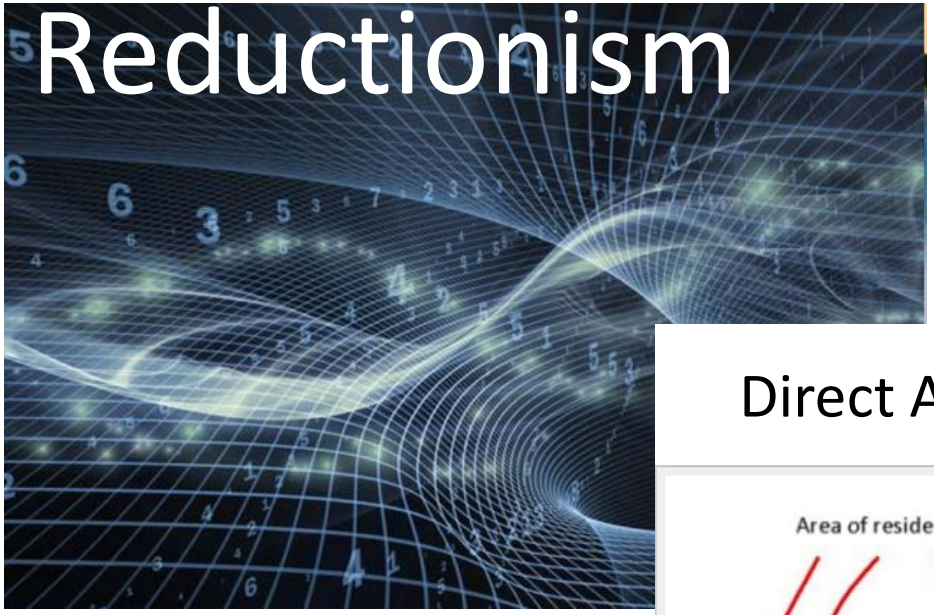
- Can you describe these factors sufficiently?
- Can you do the analysis you need to do?

The inter-dependence of covariates is the most interesting part

- Population level systems biology

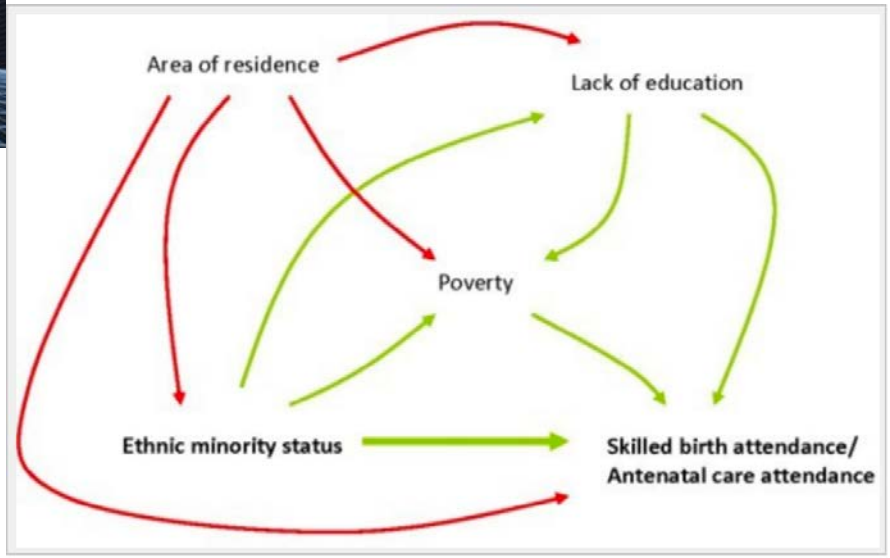
Complexity Demands Big Data Science

β x + intercept =Y

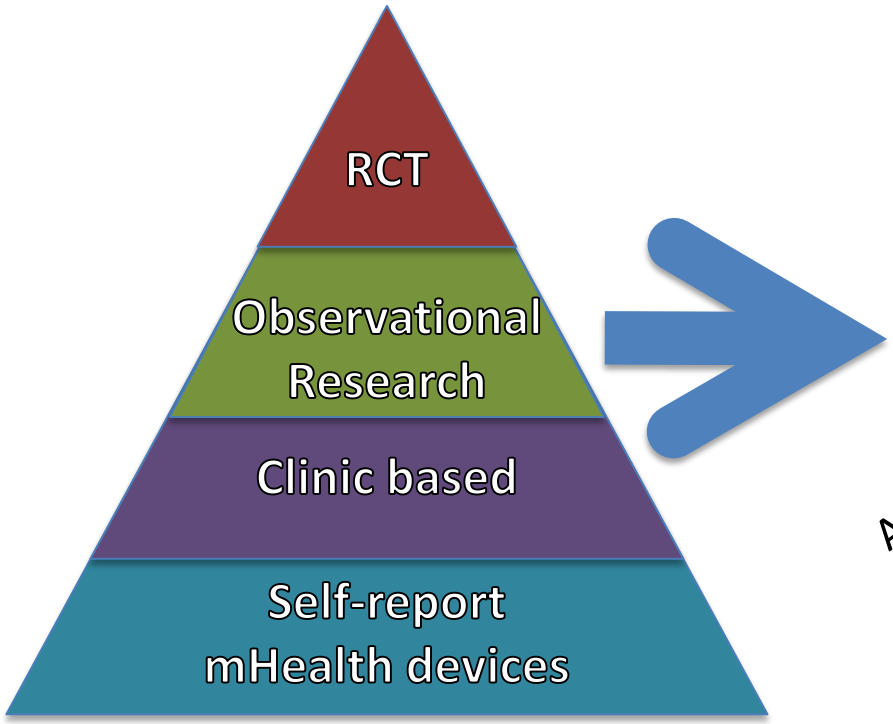


Holism

Direct Acyclic Graphs - DAGs



Debunking the Hierarchy of Data
What is your question?



Accurate, fixed interval data

Non-confounded randomized patients

Accurate, fixed interval, generalizable

physician optimized treatment

Standard of care

deeper data @ clinical events

data constantly contributed

Evolving expectations for data sharing by funders, publishers and researchers.

Annals of Internal Medicine

EDITORIAL

Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal Editors

Published online 20 January 2016

The NEW ENGLAND JOURNAL of MEDICINE

PLOS | MEDICINE

EDITORIALS

EDITORIAL

Can Data Sharing Become the Path of Least Resistance?

The *PLOS Medicine* Editors*

Published: January 26, 2016

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D. JANUARY 21, 2016

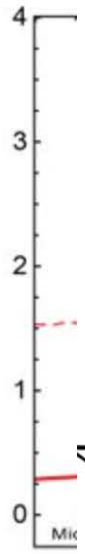
#IAmAResearchParasite

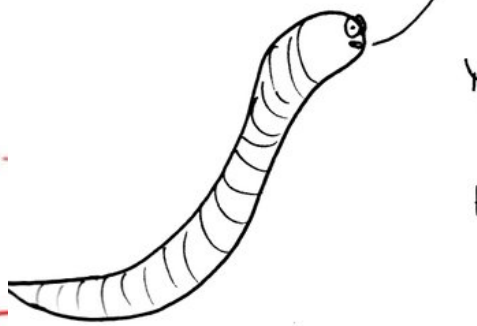
The continuing adventures of Sucky, the research parasite!

Hey guys. I was just trying to use this data that you already published for my own dastardly and nefarious purposes and I noticed that it looks like you did your normalization wrong, may have mapped all these reads to the wrong genome, and that all your conclusions are misleading. What? I should just what with my what now? Co-authorship? Really?

@redpenblackpen

The Adventures of Sucky, the research parasite





Hey! You gonna use that data?

Yes.

You done with it now?

No.

How about now?

No.

Now?

No.

That's real nice data.

Gah!

@redpenblackpen

ience: buff.ly/20iAGTZ #IAmAResearchParasite

Standard Practice

- Analysis based on your data

PLUS

- Your analysis in another data set

PLUS

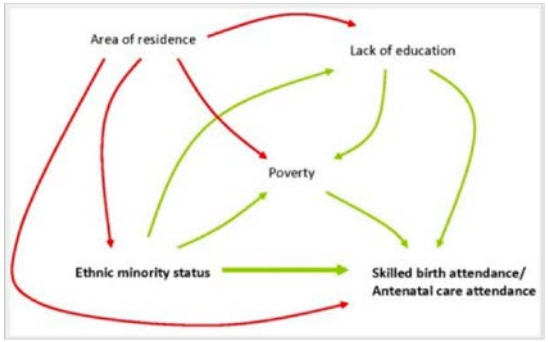
- References to published literature

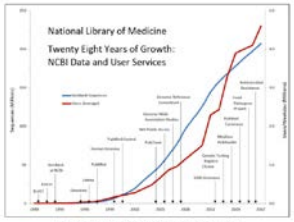
PLUS

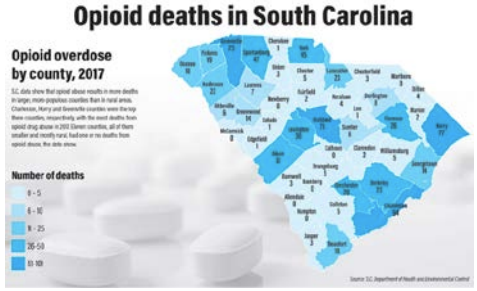
- Your data accessible to others

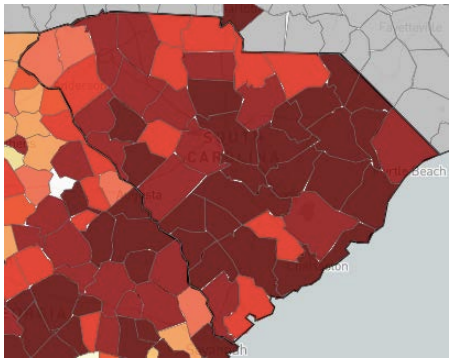
PLUS

- Your code in an accessible place









Big Data – can it work for people?

Investigators

- Credit for data creation: citation approaches, academic credit towards tenure
- Period of exclusive use
- Types of data outside of sharing requirements

Participants and Privacy

- Faster answers a trade off for less privacy? Is this okay? Desirable?
- What data are sensitive?
- Can you opt out of sharing?

Commercial Entities

- Intellectual property concerns
- Cost of data

Steve Kearney, SAS
"Consumers want privacy
Patients want good outcomes"

Big Data – Technical challenges

Secure environments

- HIPPA compliant environments
- Clouds, Limited access servers

Data architectures and interoperability

- Gen3, FHIR, OMAP, CDISC, I2B2. PhenX, OMAP, EMDI
- Huge need for cross cutting, multidisciplinary expertise to build these

Business Intelligence and visualization

- It is nothing if ***we*** can't visualize and use the data

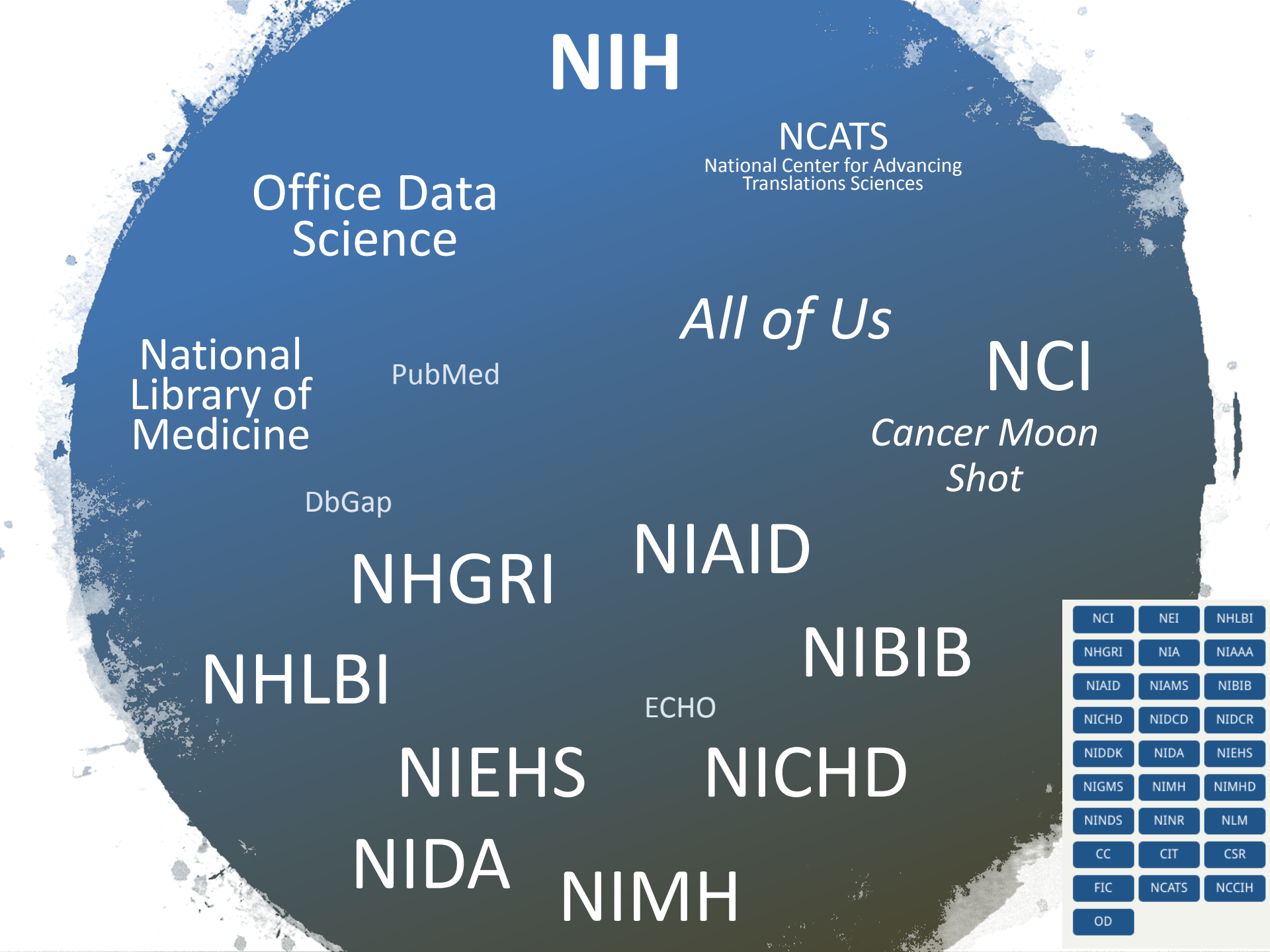
Analytics

- Causal inference, Network analysis, Machine learning
- Jupyter notebooks with Python, R, R studio
- Correct inference from complex data

How to foster an open digital ecosystem for biomedical research?

- **Ensure there are people who can make it happen**
  - People and/or teams who combine biomedical/behavioral/clinical and data science expertise
- **Develop necessary infrastructure and tools**
  - make open, accessible digital resources (data, software, etc)
  - that are findable, accessible, interoperable, and reusable (FAIR)
- **Invest in data science research applied to biomedical research challenges**
  - prove its utility and push the frontiers

Goal: foster a new culture and new capabilities


NIH
NCATS
National Center for Advancing Translations Sciences
Office Data Science
All of Us
National Library of Medicine
PubMed
NCI
Cancer Moon Shot
DbGap
NIAID
NHGRI
NIBIB
NHLBI
ECHO
NIEHS
NICHD
NIDA
NIMH
NCI
NEI
NHLBI
NHGRI
NIA
NIAAA
NIAID
NIAMS
NIBIB
NICHD
NIDCD
NIDCR
NIDDK
NIDA
NIEHS
NIGMS
NIMH
NIMHD
NINDS
NINR
NLM
CC
CIT
CSR
FIC
NCATS
NCCIH
OD

From published to unstructured

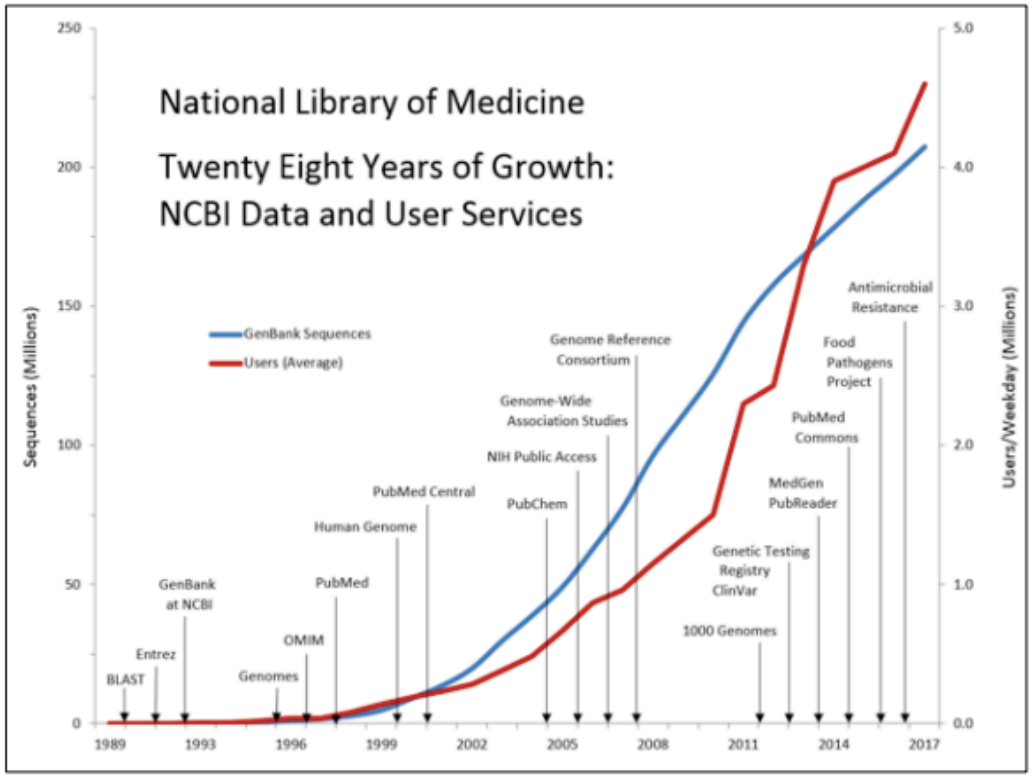


Figure 1. Growth of NCBI Data and Services, 1989-2017 Credit: NCBI

NIH STRATEGIC PLAN FOR DATA SCIENCE

https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

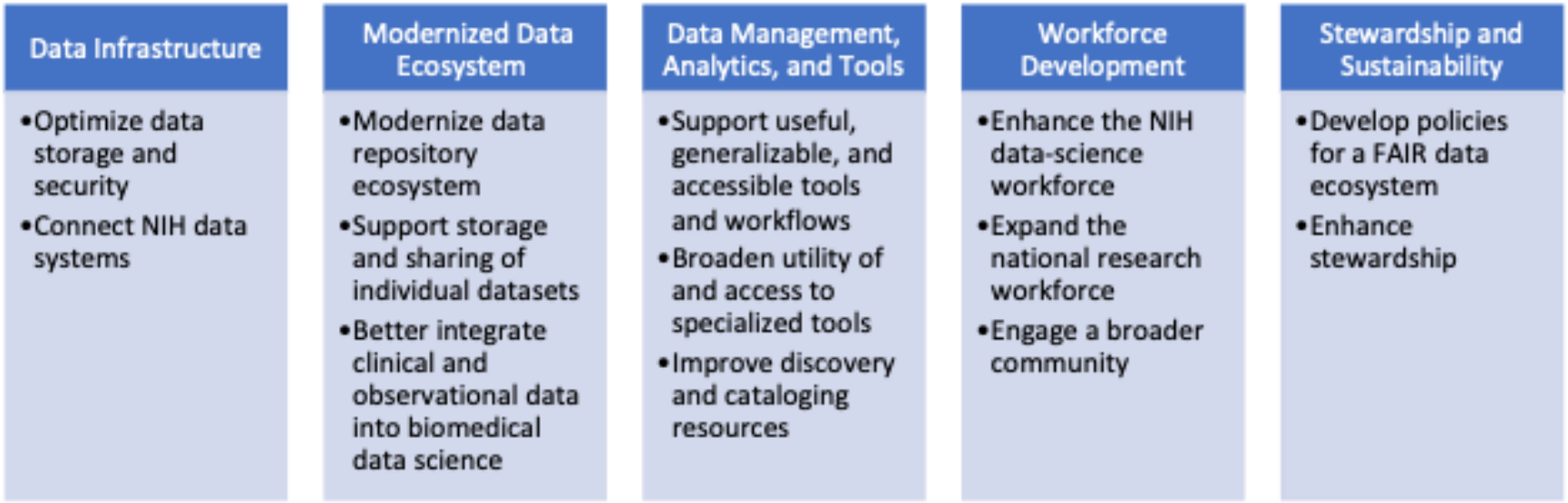


Figure 2. NIH Strategic Plan for Data Science: Overview of Goals and Objectives

Training Programs

Emerging Leaders in Data Science Fellowship

- Master's or Doctoral Degree received in last 60 months or by 9/16/2020 11:59:00 PM.
- Computer Sciences, Engineering, Life Health/Medical Sciences, Mathematics/Statistics

Mentored Research Scientist Development Award (K01)

- **Applicant Profile** health-professional doctorate.
- Research plan in epidemiology and/or data science ONLY
- Duration of three to five years, not renewable.
- Salary up to $75,000, Research support of $25,000 each year.
- Minimum nine person months (or 75 percent) effort required each year.

Mentored Clinical Scientist Career Development Award (K08)

- **Applicant Profile** clinical doctoral degree (e.g., M.D., D.V.M., or O.D.); licensed to practice working in biomedical or behavioral research, including translational research.
- Duration of three to five years, not renewable.
- Salary up to $100,000, Research support of $50,000 each year.
- Minimum nine person months (or 75 percent) effort required each year

Harnessing Big Data to Stop HIV

http://grants.nih.gov/grants/guide/pa-files/PA-18-764.html

Promote research that transforms understanding of HIV transmission, the HIV care continuum, and HIV comorbidities using Big Data Science (BDS). These approaches should include projects to assemble big data sources, conduct robust and reproducible analyses, and create meaningful visualization of big data.

Research Objectives

collaborations in epidemiology, bioinformatics, mathematical modeling, statistics, social and behavioral sciences, HIV prevention and care, and bioethics, among others, to address *both* of the following objectives:

- Improve our understanding of HIV risk and health seeking behaviors and the complex contextual environment in which they occur.
- Develop and advance the ethical framework to evaluate Big Data methods in the constantly changing environment of available digital data. Projects should explore and address relevant ethical challenges in conducting big data research including privacy concerns, questions regarding access to specific types of data, communication among users of data and the research community.

DAID's Data Portfolio

DAIDS funded cohorts

- Multicenter AIDS Cohort Study (MACS)
- Women's Interagency HIV Study (WIHS)
- CFAR Network of Integrated Clinical Systems (CNICS)
- International epidemiology Databases to Evaluate AIDS (IeDEA)

Clinical trials

- ACTG, HPTN, HVTN
- Independent investigators

Analytics

- Modelers
- Statisticians
- Epidemiologists
- Data Scientists

Call us
we love to chat

Rosemary McKaig

Epidemiologist

Data science, epidemiology, analysis, training awards

rmckaig@niaid.nih.gov

Carlie Williams

Chief Epidemiology

cwilliams@niaid.nih.gov