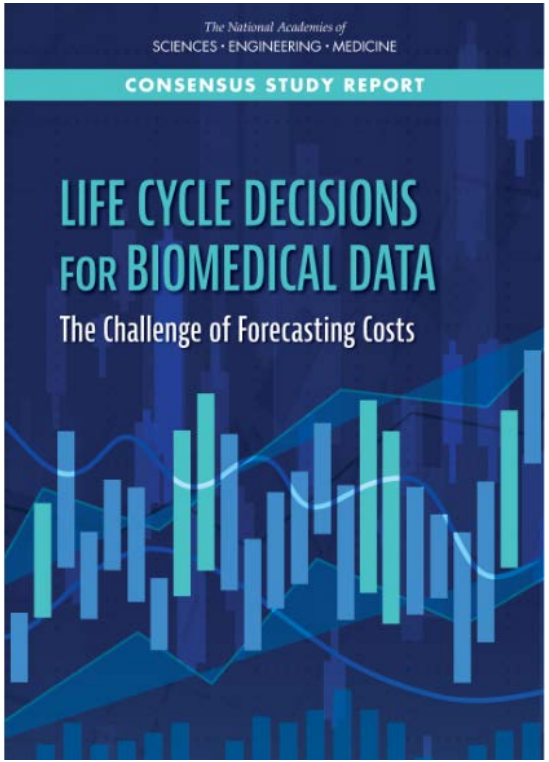
The National Academies of SCIENCES ENGINEERING MEDICINE

Board on
Mathematical Sciences & Analytics

LIFE CYCLE DECISIONS FOR BIOMEDICAL DATA

The Challenge of Forecasting Costs

BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS



Forecasting Data Costs for Researchers

Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs

Presented to the Public
August 13, 2020

Forecasting Data Costs for Researchers, Funders, and Storage Providers

August 2020 weekly webinar series, 12-1pm ET

August 13: *Forecasting Data Costs for Researchers*

August 20: *Forecasting Data Costs for Funding Institutions*

August 27: *Forecasting Data Costs for Data Storage*

This webinar series is sponsored by the ***National Library of Medicine of the National Institutes of Health***

Forecasting Data Costs *for Researchers*

Lars Vilhuber (Moderator)
Cornell University

Maryann Martone
University of California, San Diego

Lance Waller
Emory University

Robert Williams
University of Tennessee

Forecasting Data Costs
for Researchers

Maryann Martone

Professor Emerita

Forecasting Data Costs: Highlights for Researchers

Context

- Biomedical researchers generate, collect, and store more research data than ever.
- It is important to store those data in discoverable and accessible ways.
- Responsibility for data management can shift as the data moves through its life cycle.
- Proper data preservation generates costs that may be difficult to predict and allocate responsibility for.

Statement of Task

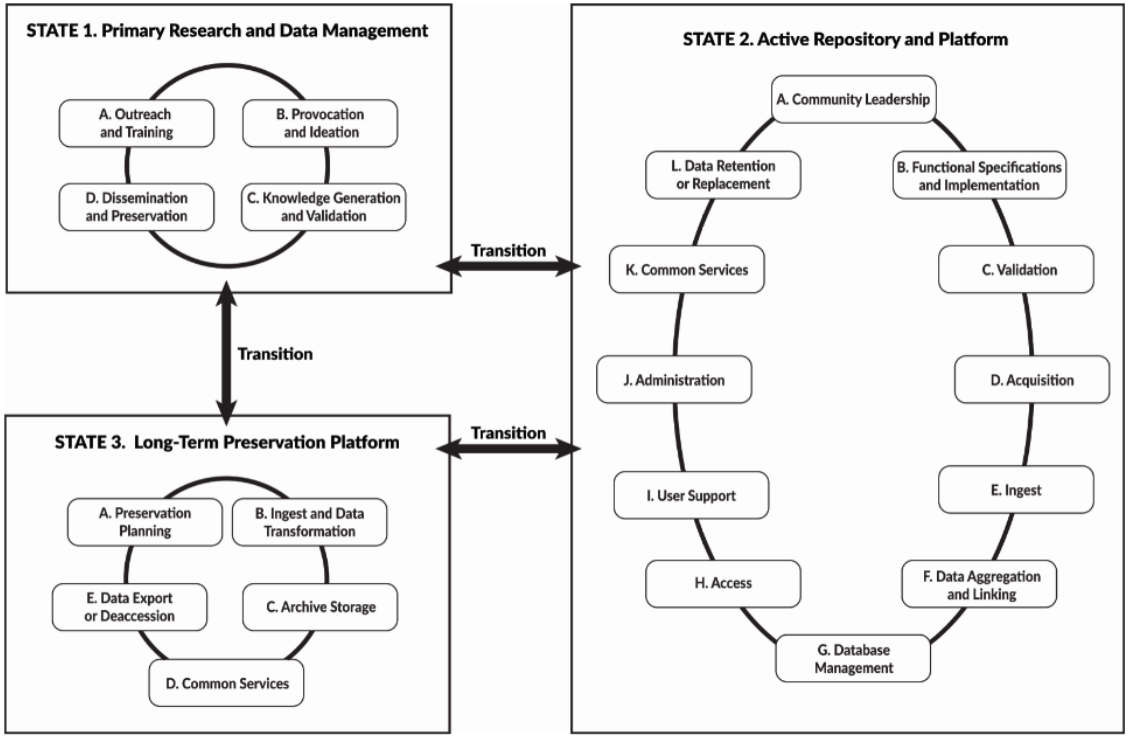
National Library of Medicine of the National Institutes of Health asked for a framework for forecasting long-term costs for preserving, archiving, and accessing biomedical data.

Framework Foundation: Three Data States

State 1: Primary research/data management environment; data are captured and analyzed

State 2: Active repository and platform; data may be acquired, curated, aggregated, accessed, and analyzed

State 3: Long-term preservation platform



Cost Forecasting Framework

- *Helps forecaster identify major cost drivers*
- *Basis for a cost forecast (not a one-size-fits-all analysis tool)*
- *Will help forecaster identify decisions that impact short- and long-term costs and data value*
- *Forecaster necessarily focuses on costs of current funding period, but must be aware that early decisions affect long-term costs of data curation and use*

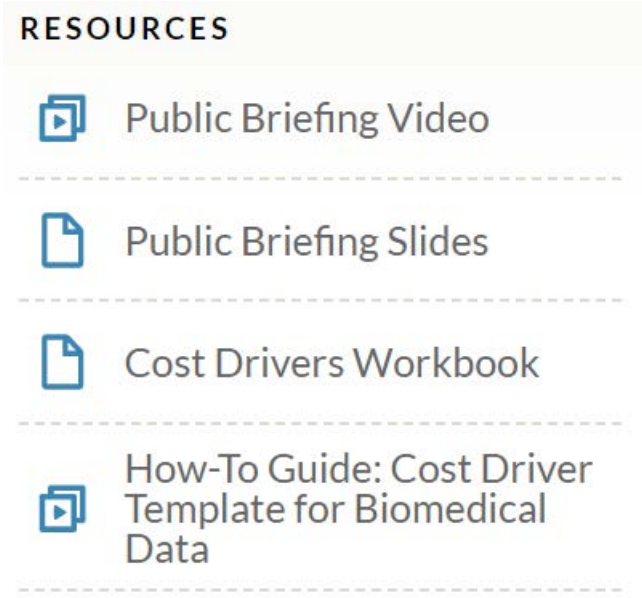
Framework Steps

These steps may occur concurrently or iteratively as new information is gathered.

1. **Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle.**
2. **Identify the characteristics of the data, the data contributors, and users.**
3. **Identify the current and potential value of the data and how the data value might be maintained or increased with time.**
4. **Identify the personnel and infrastructure likely necessary in the short and long terms.**
5. **Identify the major cost drivers associated with each activity, including how decisions might affect future data use and its cost**
6. **Estimate the costs for relevant cost components based on the characteristics of the data and information resource.**

Identify Major Cost Drivers

5. **Identify the major cost drivers associated with each activity, including how decisions might affect future data use and its cost.**



Cost Driver Categories

A. Content
B. Capabilities
C. Control
D. External Context
E. Data Life Cycle
F. Contributors and Users
G. Availability
H. Confidentiality
I. Maintenance and Operations
J. Standards, Regulatory, and Governance Concerns

The cost forecaster needs to rely on expertise within the host institution, the funding agency, and the research community to inform cost estimation.

Consult widely to develop a narrative for the long-term performance of an information resource.

Quantitatively assessing uncertainties may be difficult, but the cost forecaster should communicate concerns with decision makers, even if they cannot be precisely characterized.

Strategies for Researchers:
Fostering the Data Management Environment

- Create data environments that foster discoverability and interpretability through long-term planning and investment throughout the data life cycle.
- Incorporate data management activities throughout the data life cycle to strengthen data curation and preservation.
- Incorporate the expertise and resources needed to create and curate metadata throughout the data life cycle, and in the transition between data states, into the cost forecast.
- Weigh the benefits, risks, and costs (both short- and long-term) of data storage and computation options before selecting among them.

Forecasting Data Costs *for Researchers*

Lance Waller

Professor of Biostatistics and Bioinformatics; Co-Chair of Committee on Applied and Theoretical Statistics

Building Shoulders to Stand On

BUILDING SHOULDERS TO STAND ON

LANCE A. WALLER

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

ROLLINS SCHOOL OF PUBLIC HEALTH

EMORY UNIVERSITY

LWALLER@EMORY.EDU

MANY THANKS TO THE REPORT COMMITTEE

- Addresses tension between
  - Regulatory requirements for releasing data
  - Logistic and financial impacts of maintaining data availability within a FAIR perspective.

- So what does this mean to me and my research?

MY BACKGROUND AND RESEARCH INTERESTS

- Training: Math, Computer Science, Operations Research, (Veterinary) Epidemiology
- Career: Biostatistics, Spatial Statistics and Analysis in Public Health
- Research interests:
  - **Analytic tool development:** Spatial statistics, Geographic information systems
  - **Local exposures, local health:** Environmental health, Epidemiology
  - **Monitoring patterns, detecting hot spots:** Public health surveillance
    - Substance use patterns
    - Disease ecology
  - **Data Science in Public Health**

STATES OF DATA

- State 1: Primary research and data management
- State 2: Active repository and platform
- State 3: Long-term preservation

  - National Academies of Sciences, Engineering, and Medicine. 2020. *Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25639.

- Where do I fit in?

DATA SOURCES

- Administrative data
  - Census (decennial, American Community Survey), surveys
  - Geocoded point locations
  - Remote sensing images
  - Geolocated environmental monitoring
  - Population mobility
  - Physical landscape
  - Sociodemographic landscape

DATA FEATURES

- **Relevant data may come from States 1, 2, or 3**

- Multiple sources
- Typically, publicly available
- "De-identified" (nice discussion in pp. 7-6 through 7-8 of report)
- Aggregated to different levels (Census tract, county, state, region)
- Aggregated to different regions (Census tract, ZIP code)

RESEARCH OUTCOMES AND DATA PRODUCTS

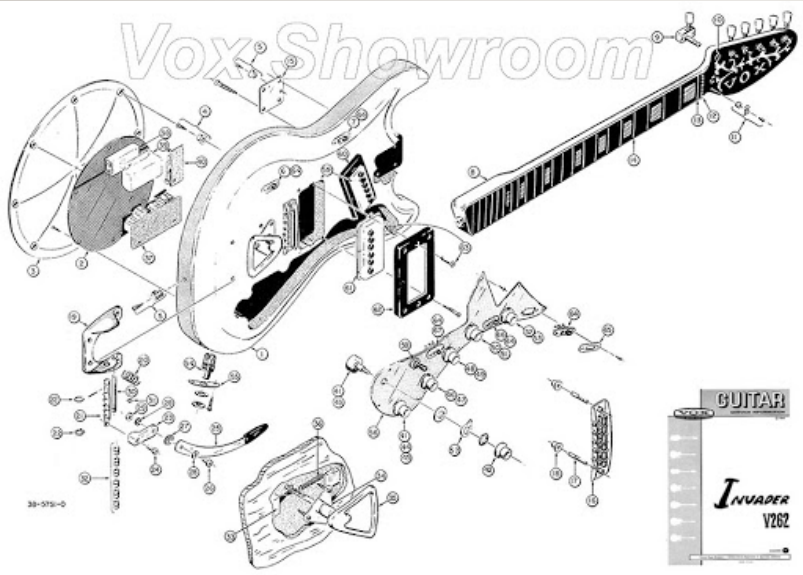
- New data set(s) linked from:
  - Multiple sources
  - Multiple types
- Outcomes
  - Derived variables/observations
  - Visualizations
  - Measures of associations
  - Measures of uncertainty
  - Imputed data for missing values

MOSTLY STAGE 2, LEADING TO MORE STAGE 2

- Big data? Maybe, but not always
- Interconnected data? Always

- Archiving challenges:
  - Do I want to keep connections? Always
  - Can I keep connections? Sometimes
  - What if some parts requires data use agreements and others don't?

DATA CITATION



- Discussed in detail in Chapter 4, Section B (Capabilities)
- What capabilities *inside* versus *outside* archive?
- FAIR Guidelines
- Dataverse examples
  - http://best-practices.dataverse.org/data-citation/
- Dryad examples
  - www.datadryad.org/)

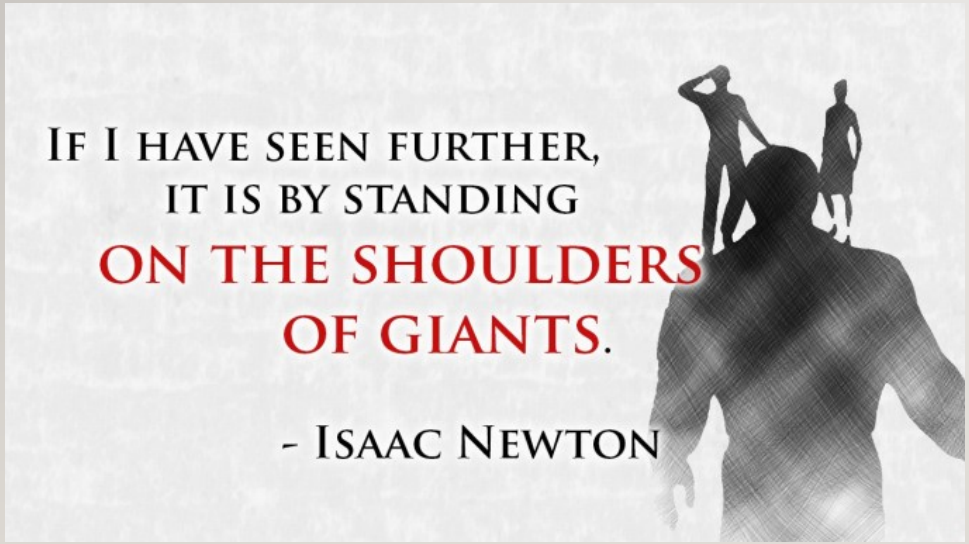
DATA CITATION VS. DATA ACCESS

- **Data citation:**
  - Identifies which data you used, which version, and defines exactly how to retrieve it.
    - Persistent identifiers (DOI, PID)
  - Can include restricted access data, *as long as you detail the process for obtaining access.*
- **Data access:**
  - Obtaining exactly the same data used by the original author.
  - *May require access or data use agreements.*
  - *May require additional costs.*

FROM CHAPTER

- "In the long term, it is more efficient to think early about how decisions may affect the costs of data management and access in future data states, the transitions to these states, and the future value of data to the scientific enterprise."

STANDING ON THE SHOULDERS OF A GIANT OR STANDING ON THE SHOULDERS OF GIANTS?



https://3starlearningexperiences.files.wordpress.com/2018/12/giants.jpg

https://www.cas.org/sites/default/files/styles/promo_hero/public/2018-07/shoulders_of_giants_0.jpg

Forecasting Data Costs *for Researchers*

Robert Williams

Professor of Genetics, Genomics and Informatics; UT-ORNL Governor's Chair in Computational Genomics

From the trenches, to tactics, to long-term strategies for data preservation

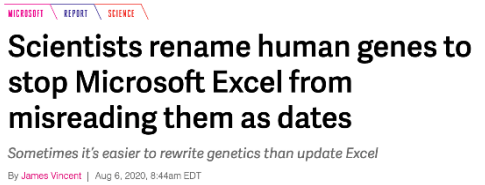
Make data live!

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The trenches: Where this started for me — the first decade

Quantitative neuroanatomy: Motivating questions: How many neurons in different brain regions as a function of age, sex, genotype? What gene variants exert control?

MS Excel, Excel, and more Excel: Excel is still the lingua franca/lowest common denominator. There is no way around it. This week, 27 human genes renamed to keep Excel happy. Excel loves to mess up: MARCH1 converted to 43891 or 1-Mar.


MICROSOFT EXCEL REPORT SCIENCE

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

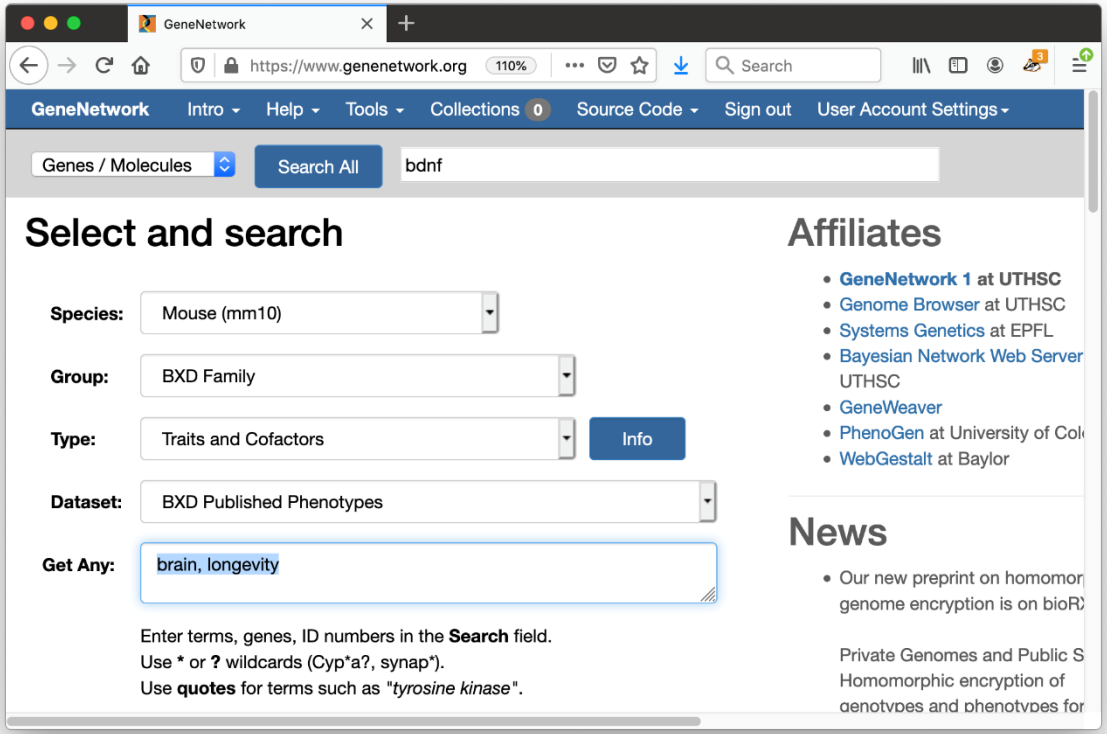
Simple to relational databases—FileMaker, Access, then MySQL, PostgreSQL This is a very big leap for many laboratories. The leap to FOSS databases requires significant long term investment—one postdoc equivalent of effort for the enitre duration—in my case, 20 years of support for programmers and DBAs

But databasing is not enough. Data need to be matched to analysis and integration software systems. In my case, this has required 20 years of support for a minimum of at least one programmer—preferably two—to develop webservices such as GeneNetwork.

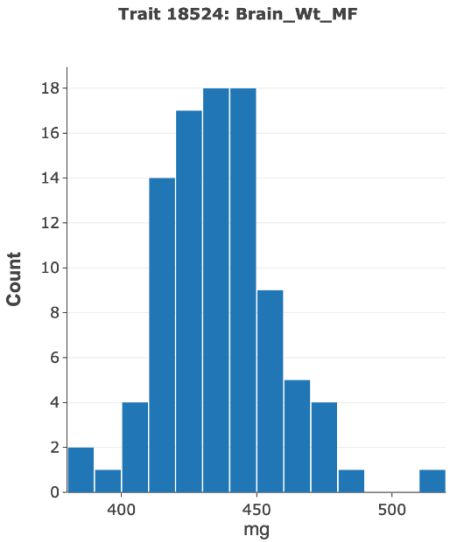
Make data live!

Tactics: The second and third decade

Data shepherding as a career: An unintended outgrowth of the Human Brain Project



https://www.genenetwork.org/show_trait?trait_id=18524&dataset=BXDPublish#redirect

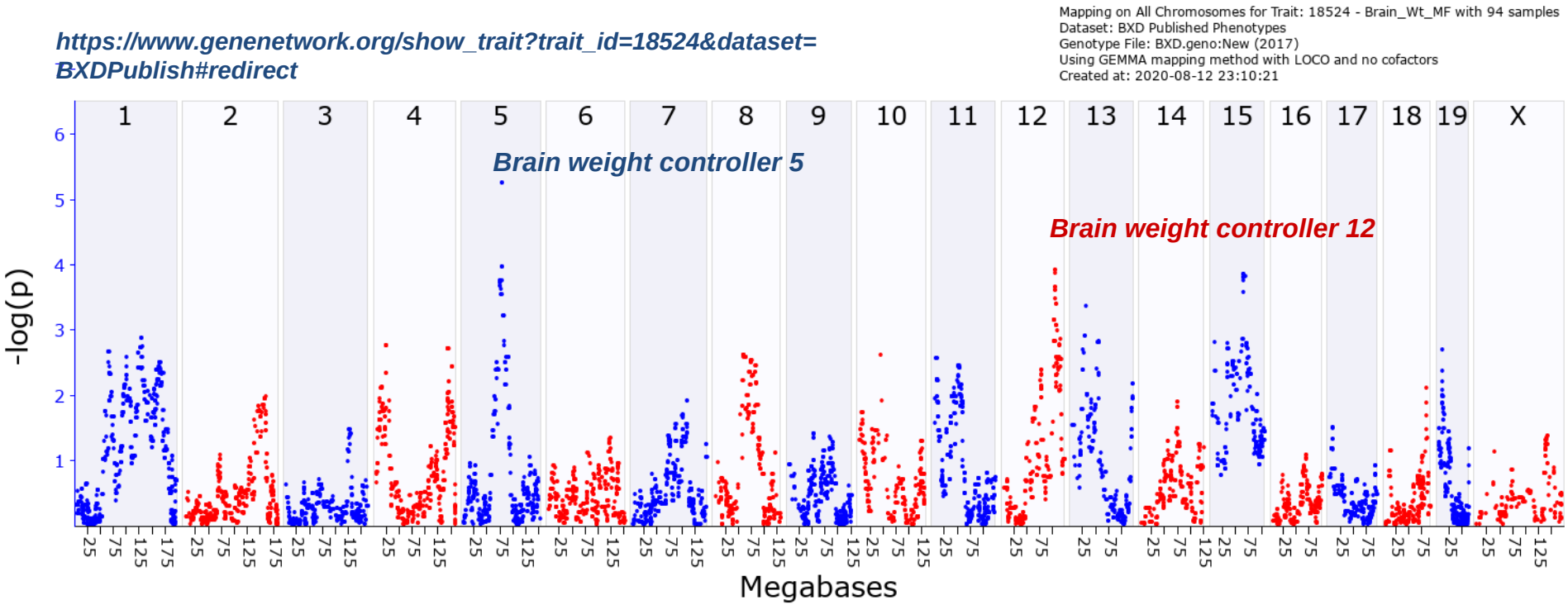


Make data live!

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Tactics: The 2nd and 3rd decades: building a community and a software stack

https://www.genenetwork.org/show_trait?trait_id=18524&dataset=BXDPublish#redirect

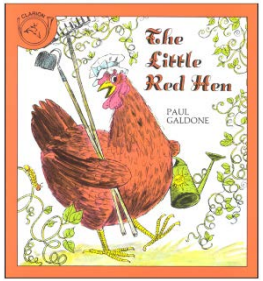


Data shepherding as a career: An unintended outgrowth of the Human Brain Project

Make data live!

Tactics: The 2nd and 3rd decades: What did I learn about data and costs?

1. You have to combine strong science with strong data/computational goals to get funded
2. Exploratory science can work. Support for programmers is not begrudged
3. Converting postdoc and grad students into programmers can work; but collaborations needed
4. Once the tools work well, expect to have many Little Red Hen moments
5. Once the tools work expect to be considered a super-tech, not a scientist
6. Getting barely adequate data is easy; getting good metadata — impossible
7. Software development only gets more complex and interesting with time
8. The code stack and database will be “junk” unless you work with professionals
9. A web service data team can be as small as two dedicated programmers; not less
10. There are now amazing strategic opportunities in data acquistion, FAIR-compliance, and new-wave data publication methods (R Shiny, Jupyter Notebooks, web services, and APIs



Make data live!

Strategies: The next decades: Where WE NEED TO GO

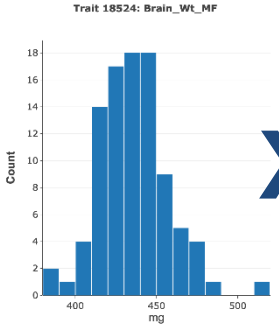
1. What is the WORTH of a data set?

Most data are additive, evaporative.

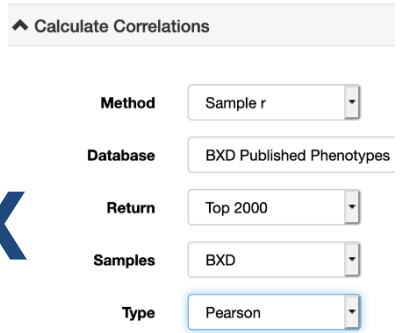
MULTIPLICATIVE data lives!

9 independent/siloed data sets

or

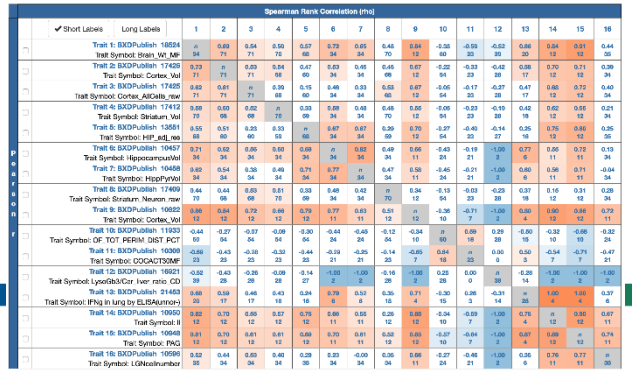


X



= 3M *rs*

There is a amazing paucity of stable multiplicative data—Framingham, WHS, ABA, GTEx, PDB, GeneNetwork... Almost none are directly computable.



Make data live!

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Strategies: The next decades: Where this should go

2. Who pays the price for data evaporation and obsolescence? We all do.

"Missed opportunity cost" is inestimable.

Upon this gifted age, in its dark hour,
Rains from the sky a meteoric shower
Of facts...they lie unquestioned, uncombined.
Wisdom enough to leech us of our ill
Is daily spun; but there exists no loom
To weave it into fabric;

—Edna St. Vincent Millay (Huntsman, What Quarry? 1939)

Rob Pike's 5 Rules of Programming

Rule 5. Data dominates. If you've chosen the right data structures and organized things well, the algorithms will almost always be self-evident. Data structures, not algorithms, are central to programming.

Write stupid code that uses smart objects

Fred Brooks: *The Mythical Man-Month*

3. As much as we need to preserved good data, we need a new approach to data generation based on statistical principles, causal analysis, and future AI needs.

Forecasting Data Costs *for Researchers*

Please submit questions using the Q&A button in the zoom menu.

Lars Vilhuber (Moderator)
Cornell University

Maryann Martone
University of California, San Diego

Lance Waller
Emory University

Robert Williams
University of Tennessee

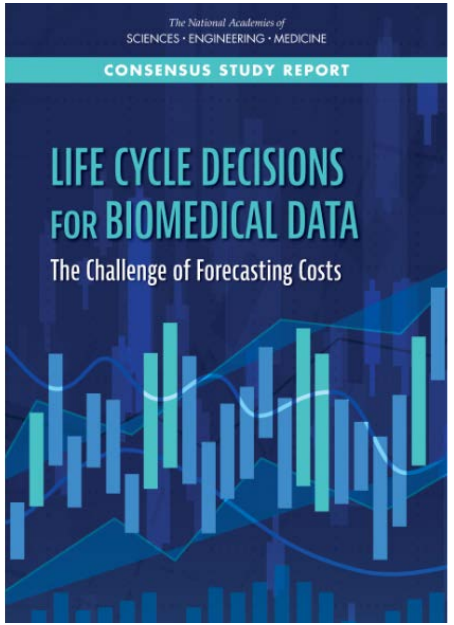
Forecasting Data Costs for Researchers, Funders, and Storage Providers

August 2020 weekly webinar series, 12-1pm ET

August 13: *Forecasting Data Costs for Researchers*

August 20: *Forecasting Data Costs for Funding Institutions*

August 27: *Forecasting Data Costs for Data Storage*



This webinar series is sponsored by the ***National Library of Medicine of the National Institutes of Health***

NIH U.S. National Library of Medicine

To register, visit http://biomed-data-costs.eventbrite.com/