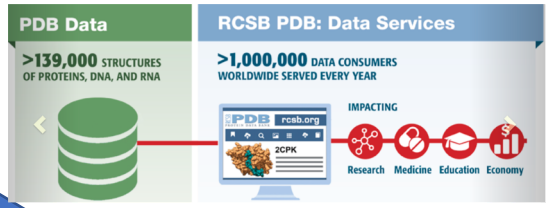
What Is It Going To Cost And What Is In It For Me?

Philip E. Bourne PhD, FACMI
Stephenson Chair of Data Science
Director, Data Science Institute
Professor of Biomedical Engineering
peb6a@virginia.edu
https://www.slideshare.net/pebourne
@pebourne

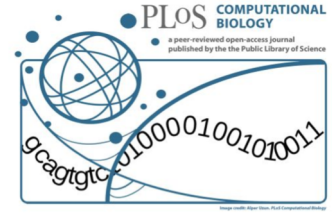
UNIVERSITY OF VIRGINIA
DATA SCIENCE INSTITUTE

BIOMEDICAL ENGINEERING

My Perspective



The cost to replicate the contents of the PDB archive is estimated at
$14 billion



PLOS' New Data Policy: Public Access to Data

NIH
National Institutes of Health

BD2K

Why This Title –

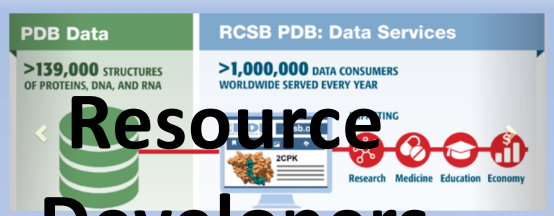
What Is It Going To Cost And What Is In It For Me?

Because whomever is considering questions of data management/preservation/access these are the only questions that seem to matter

Consider the problem from the perspective of stakeholders in a supply chain ….

Stakeholders in Supply Chains

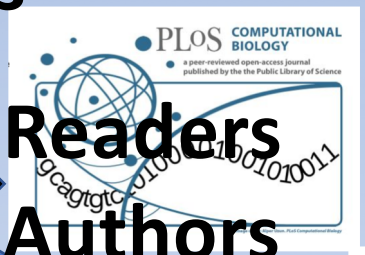
Funders



Resource Developers

The cost to replicate the contents of the PDB archive is estimated at **$14 billion**

Publishers



Readers Authors

PLOS' New Data Policy: Public Access to Data

NIH Directors Congress

NIH Researchers

National Institutes of Health

BD2K

Deans Presidents

Faculty Students

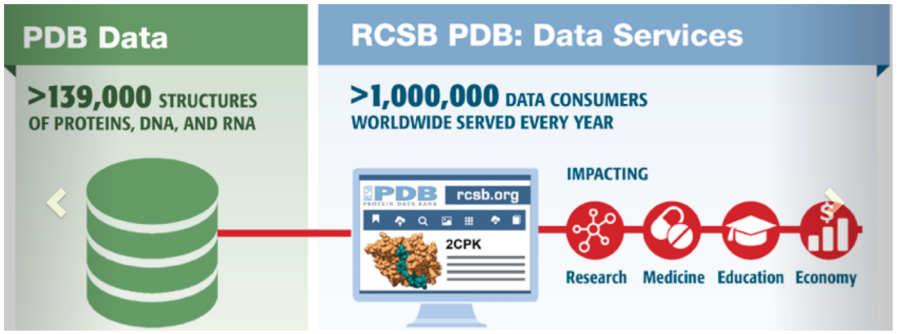
When it comes to data …

None of these supply chains is sustainable in its current form ….

Funders

Resource Developers

The cost to replicate the contents of the PDB archive is estimated at $14 billion



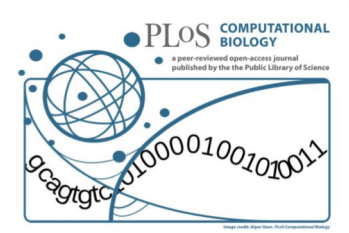
- Even for a resource so heavily used 5-year funding cycles are not assured
- There is little international cooperation at the funder level
- Funders have ownership issues too
- Developers are reluctant to seek private funding as they fear it will impact their federal funding

Publishers

PLOS' New Data Policy: Public Access to Data

- Only large publishers have the means to sustain a data ecosystem – they are large because they are profit making
- Lack of expertise

Readers
Authors



- Authors want to publish their next paper not deposit high quality data because there is little reward
- Data are only accessed a small fraction of the time
- Data are move valuable in aggregate

NIH Directors
Congress

Researchers

- The distinction between data science and data management is not clear
- Experimental mindedness
- Need to support alternative business models
- Need to put teeth into data management plans

- Need to think business models
- Need to move beyond a sense of entitlement

Deans
Presidents

- May not appreciate the value of data – think its free
- Have yet to realize how data are critical to the future of the institution

Faculty
Students

- Lack appropriate access even to their own data

Moreover, “Forecasting Costs” whether you believe the system is sustainable or unsustainable is very difficult.. and becoming more difficult …

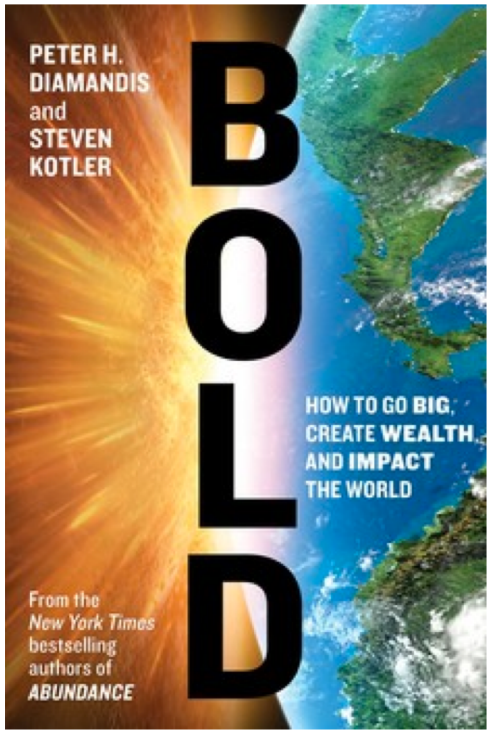
Here is why …

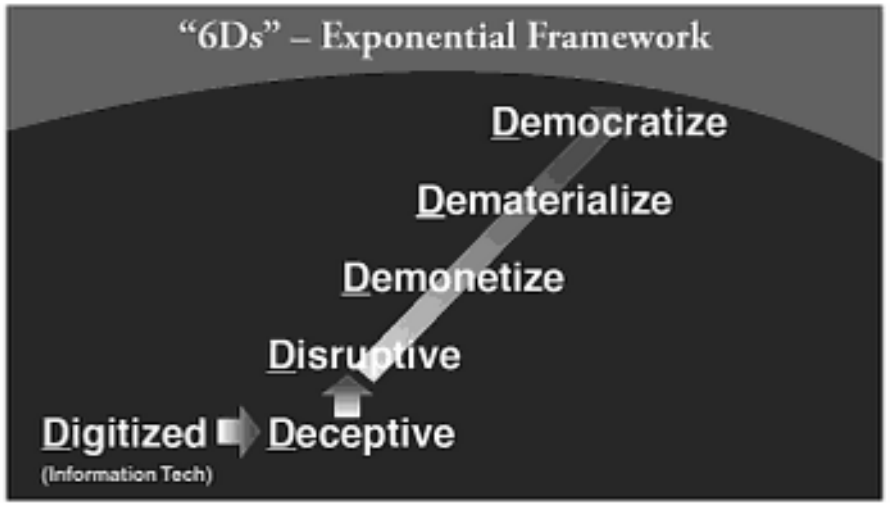
Story of the Trauma Surgeon …

- What does this story tell us?

  - It's the promise of things to come
  - Data integration by new types of researchers leading to important biomedical outcomes
  - Suddenly biomedical data is only part of the story to be told
  - That data must be preserved collectively if the story is to be reproduced
  - There is no repository as suitable support for this story
  - It's the tip of an iceberg

How Disruptive Could this Be? (with Apologies)



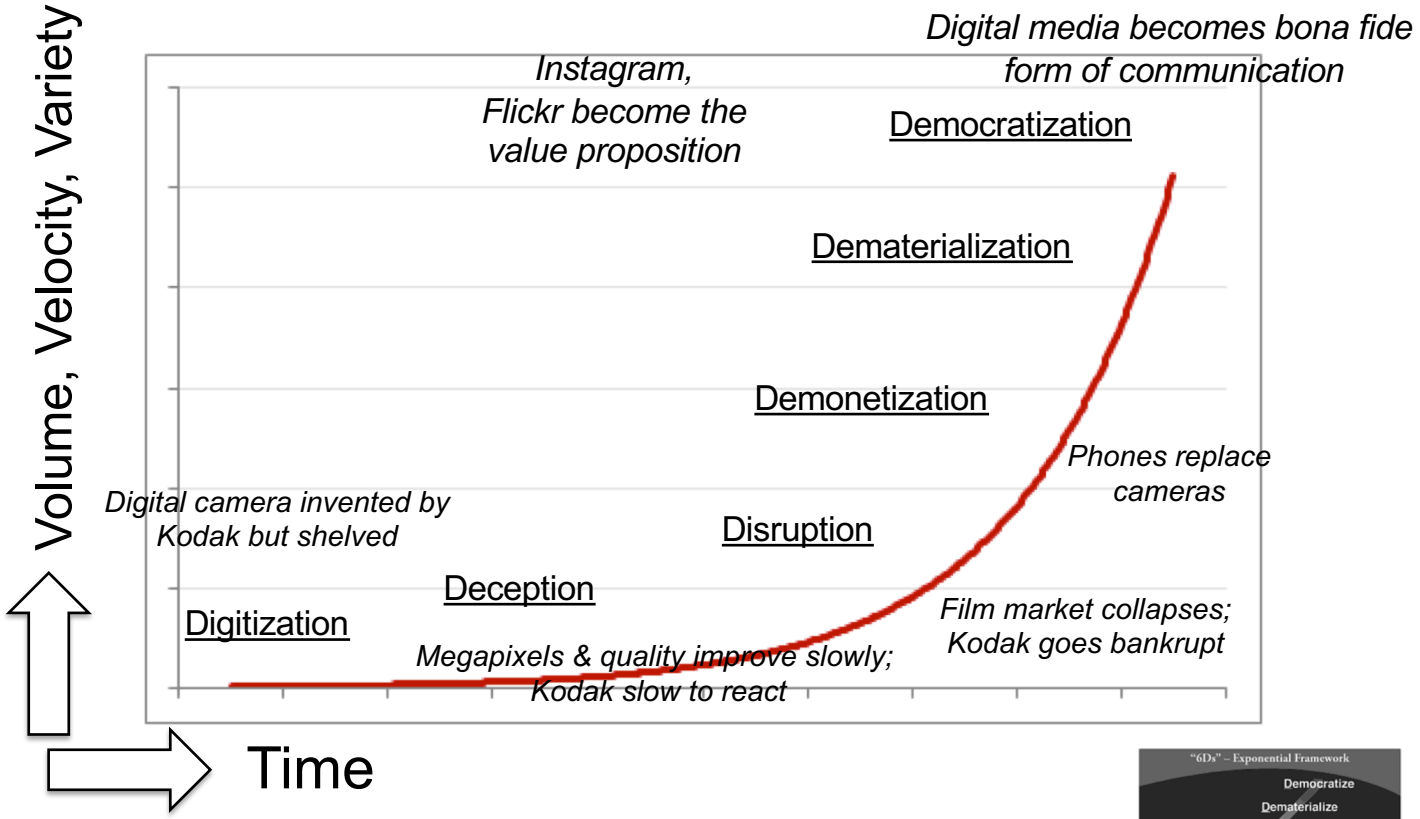


The 6 Ds of Exponentials: Digitalization, Deception, Disruption, Demonetization, Dematerialization, and Democratization

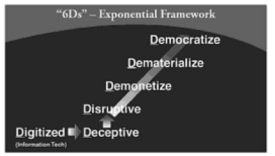
Source: Peter H. Diamandis, www.abundancehub.com

From a 2015 presentation to the Advisory Board to the NIH Director

Example - Photography



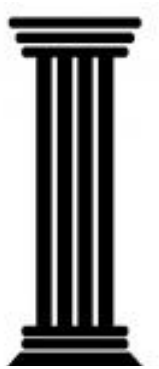
From a 2015 presentation to the Advisory Board to the NIH Director



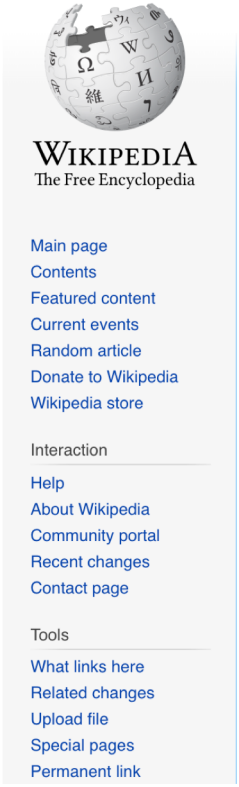
The 6 Ds of Exponentials: Digitalization, Deception, Disruption, Demonetization, Dematerialization, and Democratization

Source: Peter H. Diamandis, www.abundancehub.com

Yet Another Wake Up Call



Machine Learning
& Analytics



Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link

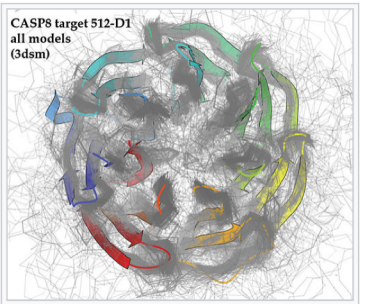
CASP

From Wikipedia, the free encyclopedia

Critical Assessment of protein Structure Prediction, or **CASP**, is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994.[1] CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users. Even though the primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence, many view the experiment more as a "world championship" in this field of science. More than 100 research groups from all over the world participate in CASP on a regular basis and it is not uncommon for entire groups to suspend their other research for months while they focus on getting their servers ready for the experiment and on performing the detailed predictions.

Contents [hide]

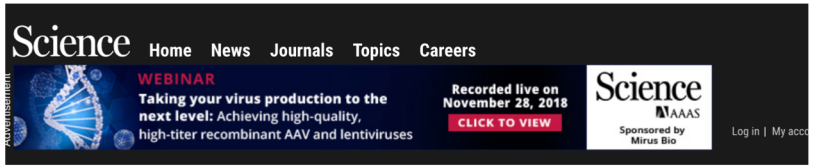
1 Selection of target proteins
2 Evaluation
3 See also
4 References
5 External links
5.1 Result Ranking



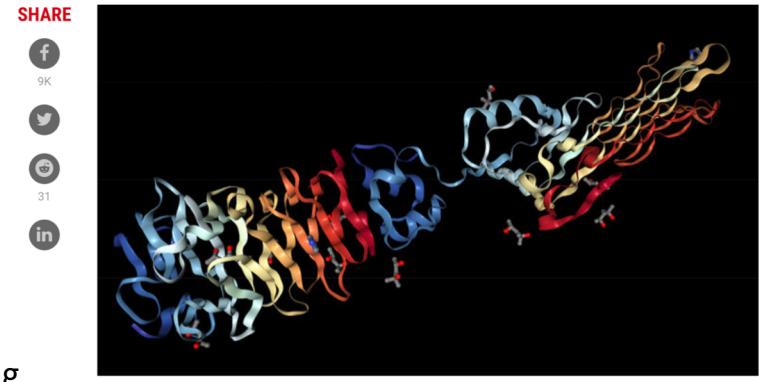
A target structure (ribbons) and 354 template-based predictions superimposed (gray Calpha backbones); from CASP8

Yet Another Wake Up Call

Machine Learning
& Analytics



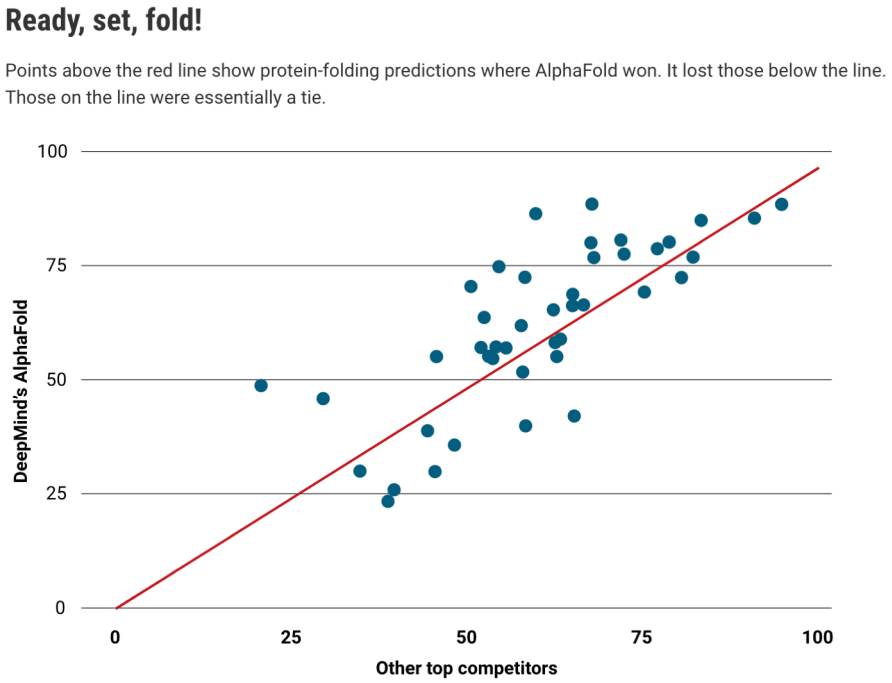
SHARE



Complex of bacteria-infecting viral proteins modeled in CASP 13. The complex contains four separate subunits that were modeled individually. PROTEIN DATA BANK

Google's DeepMind aces protein folding

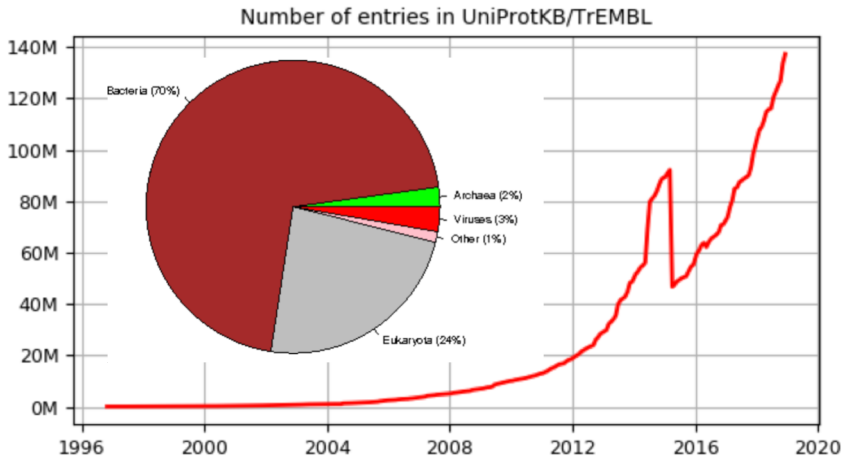
By Robert F. Service | Dec. 6, 2018 , 12:05 PM



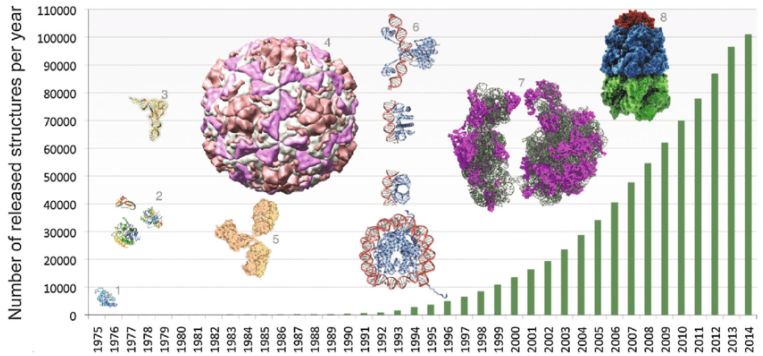
https://www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding
https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/

Further Drivers of Change (ELSI Notwithstanding)

- Training data is doubling every two years
- Robust and reusable tools in Python and R
- More advanced tools e.g., Deep Artificial Neural Networks (DNNs)
- New computing power e.g., GPUs, the cloud
- Advances coming from the private sector NOT academia
- Successful integration into lifestyles - patients will demand it



https://www.ebi.ac.uk/uniprot/TrEMBLstats



Contents of the Protein Data Bank

Pastur-Romay et al. 2016 doi:10.3390/ijms17081313

Lets summarize with respect to our original questions…

What is it going to cost?
As much as you are willing to spend

What is in it for me?
A significant part of the future of biomedical research proportional to your spend

These answers are not very satisfactory to say the least…

Let us consider possible solutions at least at the academic institution level

One institution with an important opportunity

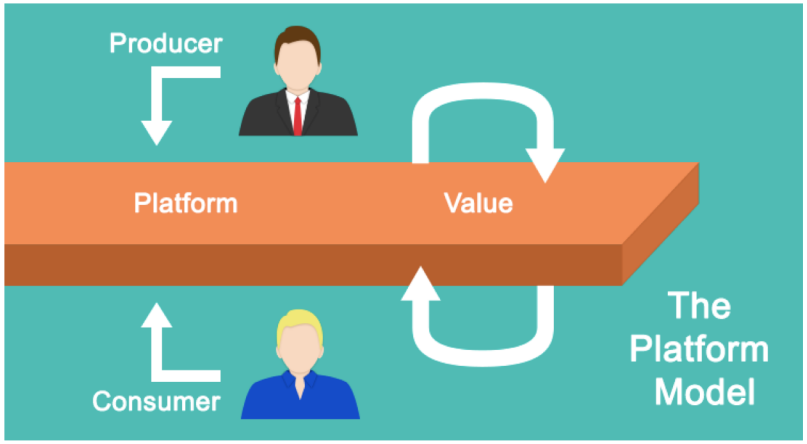
We would not exist if not for open data

We Need to Change the Institutional Culture Surrounding Data

- We need use cases of “eat your own dog food” to show value
- We need to embrace the institutional libraries role as one beyond data preservation to that of analyst
- We need to reward reproducible science and open science where data plays a major role:
  - Part of the faculty/staff handbook
  - Part of the hiring process
  - Part of the promotion process
- We need better data governance

We need the institutional infrastructure for data …

We need to move from pipes to platforms



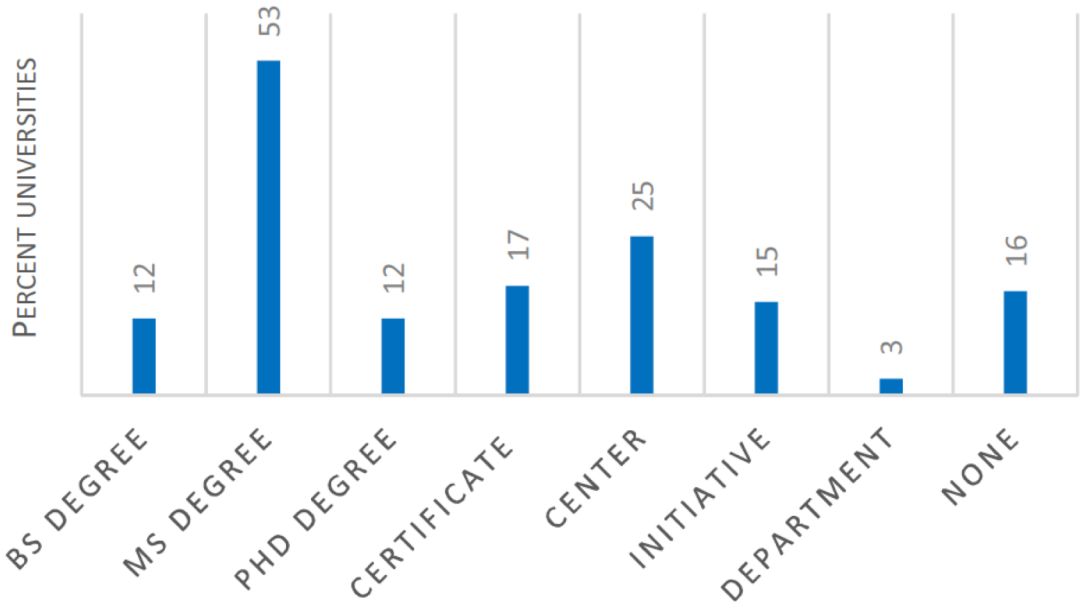
https://blog.lexicata.com/wp-content/uploads/2015/03/platform-model-750x410.png

We Need a Realistic Business Model

- Tuition
  - Students use and reuse data and hence should pay for that quality data
- Federal Funding
  - It's a part of the solution, but not the whole solution, it will not scale
- Philanthropy
  - Most philanthropists are not aware of the importance of data in what they give money to support – Advancement offices need to be educated first
- Public Private Partnership
  - Funding agencies should encourage this – it is more than SBIRs – witness capstones

We Are Not Alone

Data Science Offerings at Research Universities (n=116)



Source: Moore Foundation, 2017.

2019 N> 160

What Should be Done?

- A data deluge and opportunities lost are what happens when you are forecasting costs
- Demand (science) far outweighs supply (data resources) – support those resources that make the most strategic sense
- Broaden the responsibility for data to include academic institutions and the private sector
  - Develop incentives to support institutional data resources that impact the culture
  - Resource institutional/biomedical libraries
  - Foster public private partnerships that support public data

My answers to the original questions…

What is it going to cost?
Less if we consider data as part of a broader ecosystem with many stakeholders

What is in it for me?
Improved research and healthcare outcomes

Conversation Cards

- What role do you think institutions should play in support of data?
- Does the emergence of data as a science – data science – present opportunities?
- What role should the private sector play?