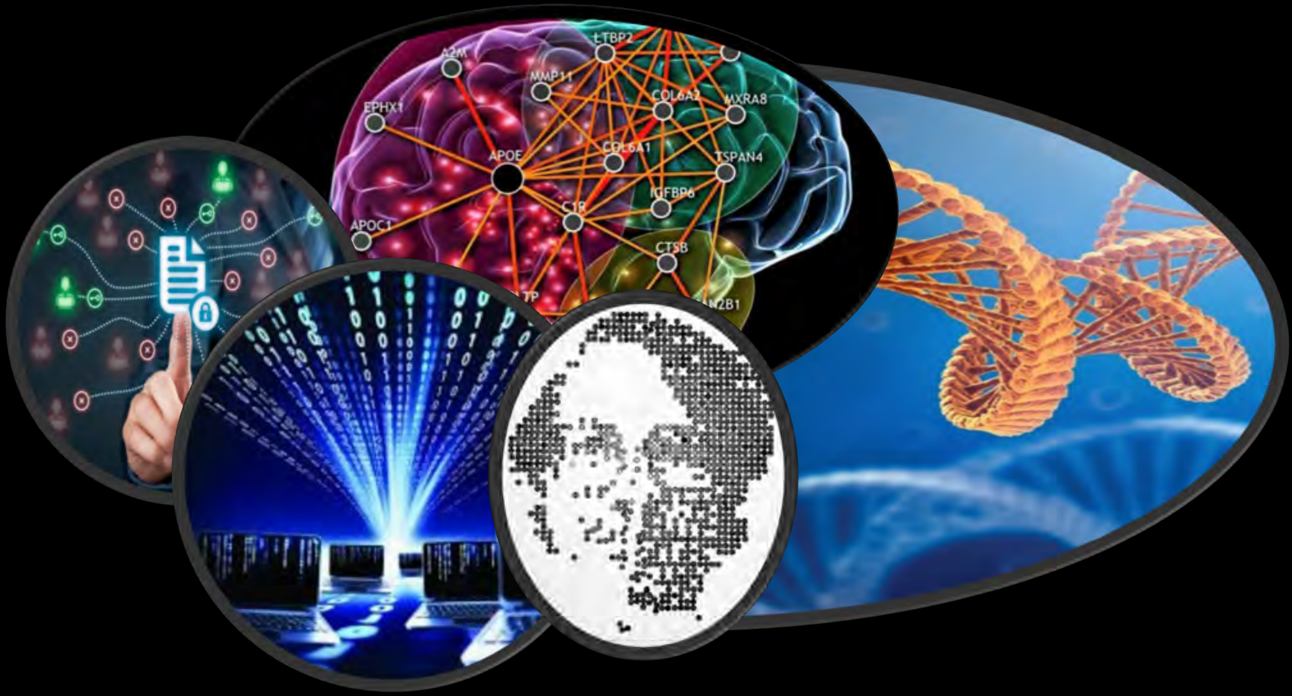
Advancing Knowledge Through Secondary Data Use



NIDDK Data Management & Sharing Webinar Series

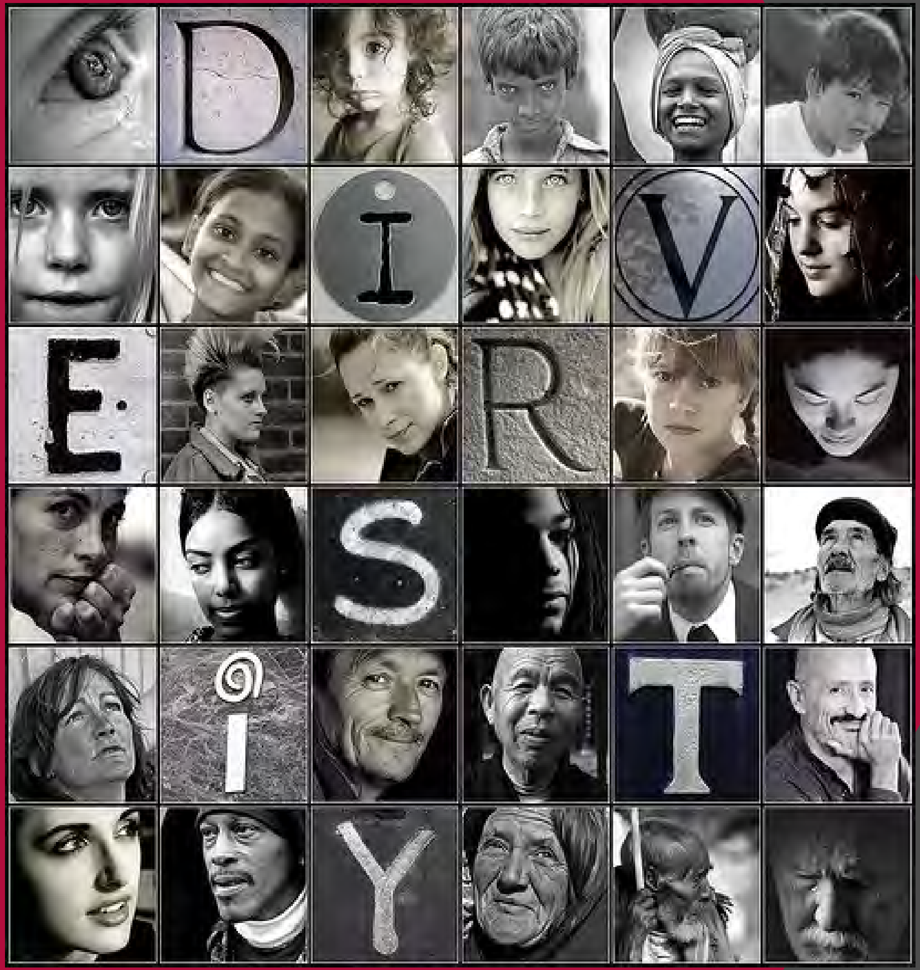
03 March 2023

Vivian OTA WANG, Ph.D., CGC, FACMG

Office of Data Science Strategy

DPCPSI/Office of the Director

National Institutes of Health, DHHS



- **The Drivers**

Human Rights and Open Science

- **The Data and Data Science**

Volumes of Varied and Complex Data

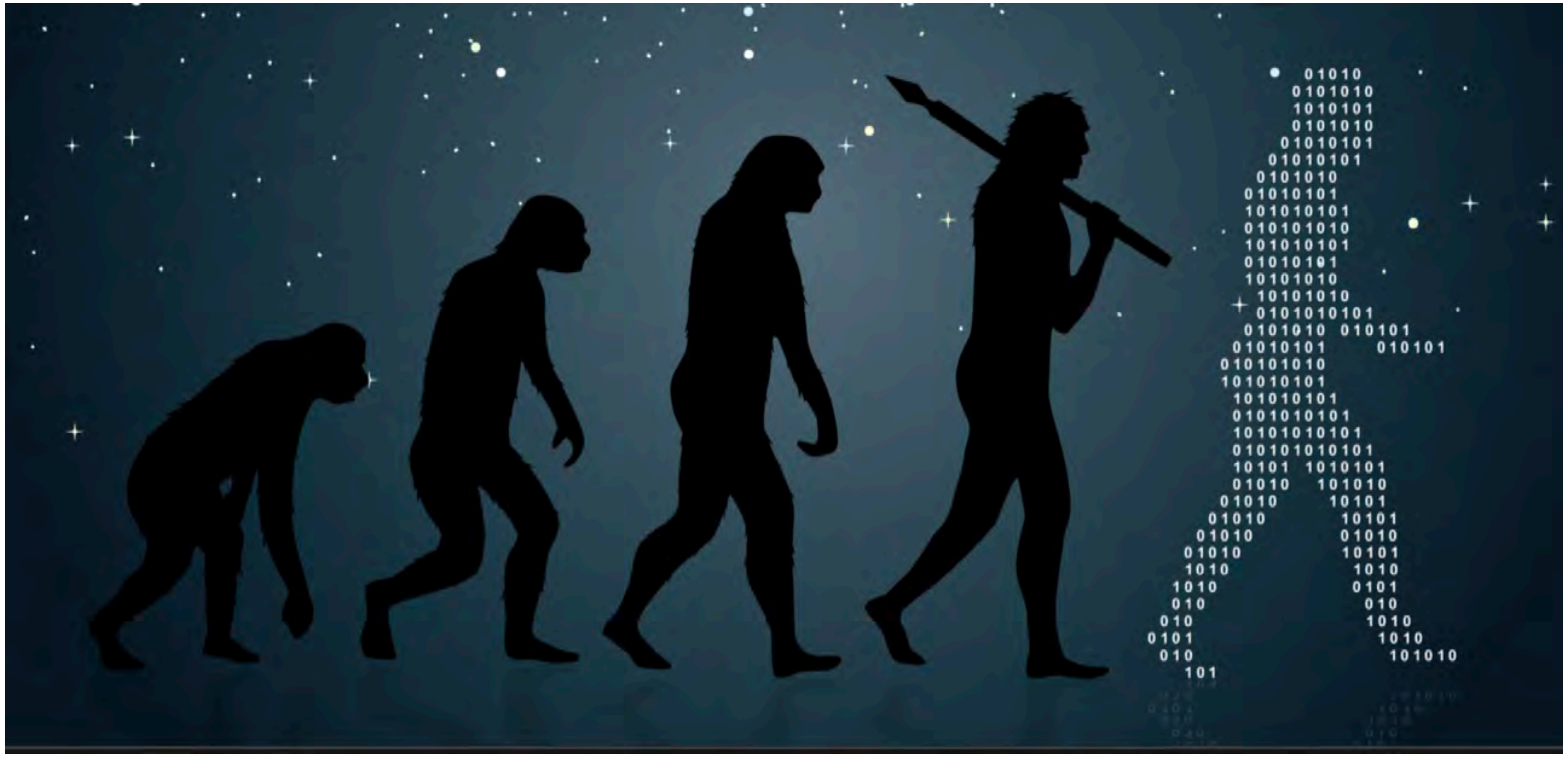
- **The Challenges**

Economic, Ethical, Legal, and Social Implications

- **Next Steps**

You

THE DRIVERS



Top Ten Leading Causes of Death *(2020-2022)*

	Category	Total deaths (Jan.-Sept. 2022)	Total deaths (2021)	Total deaths (2020)
1	Heart disease	572,336	767,937	764,512
2	Cancer	454,176	604,358	599,607
3	COVID-19	234,434	475,059	343,566
4	Accidents	170,166	226,987	203,033
5	Stroke	123,215	162,769	159,248
6	Chronic respiratory	107,559	141,906	152,051
7	Alzheimer	87,866	119,442	134,271
8	Diabetes	74,716	103,197	101,355
9	Other respiratory	50,635	66,381	66,053
10	Renal failure	42,596	53,057	51,221

Notes: For 2022, the total death sum for each category is for January 1 - September 30, 2022, except deaths from accidents and suicides are from January - September 2021. Chronic respiratory is chronic lower respiratory disease.

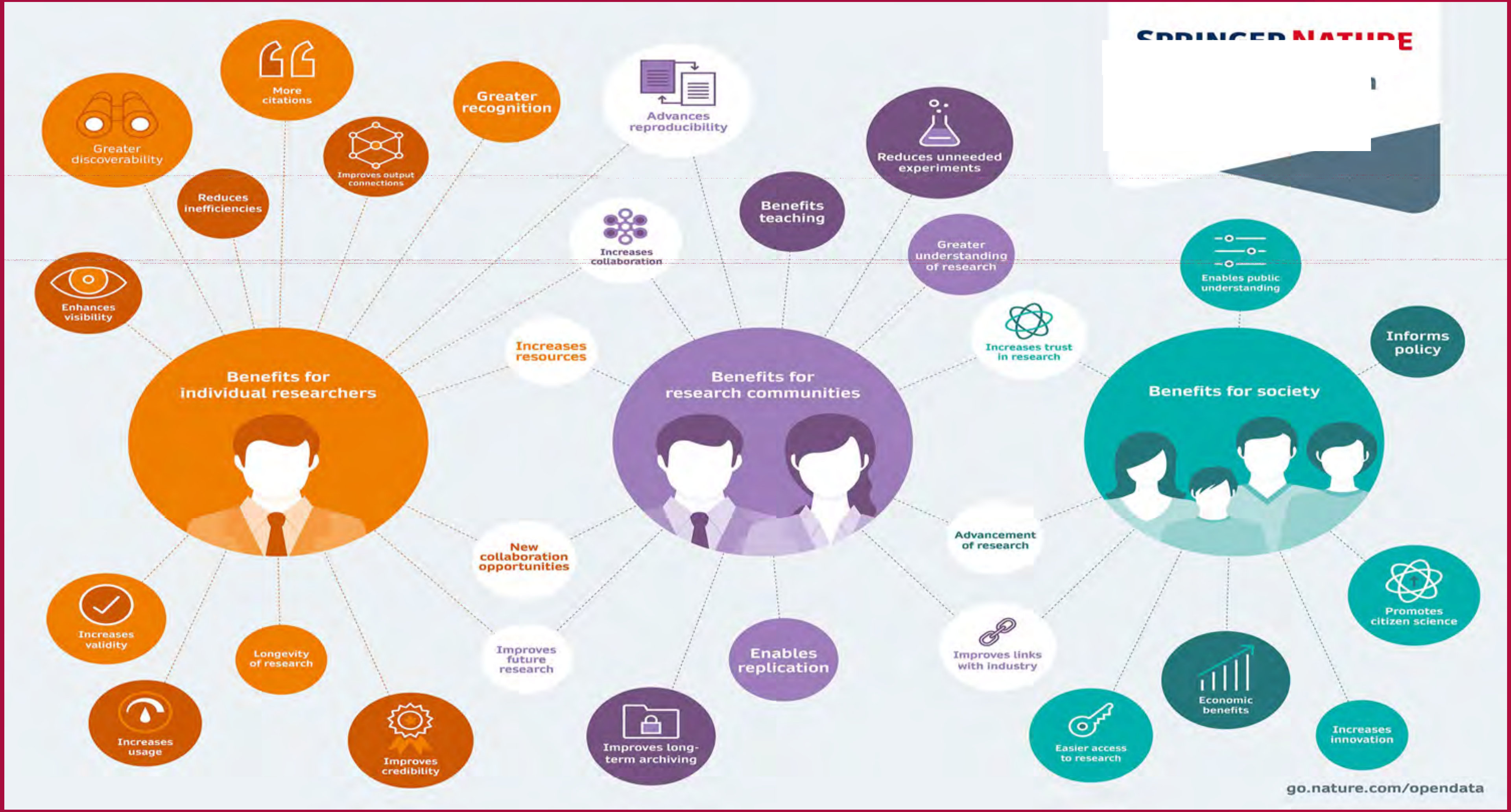
https://www.healthsystemtracker.org/brief/covid-19-leading-cause-of-death-ranking/

The People Factor: Top Ten Modifiable Behaviors Contributing to Mortality

- Tobacco use
- Diet
- Physical activity
- Alcohol misuse
- Microbial agents
- Toxic agents
- Firearms
- Sexual behavior
- Motor vehicle accidents
- Substance abuse


Big Data in Psychology
Special issue of Psychological Methods
Psychological Methods
Vol. 21, No. 4, December 2016
Item #: 2272104
ISBN: 978-1-4338-9024-6
Format: Hard copy
Other Format: PDF
Science
$15
27 JULY 2018
sciencemag.org
AAAS
SPECIAL ISSUE
FRONTIERS IN
COMPUTATION
BIG
DATA
STORAGE
VOLUME
PROCESSING
DECISION
VARIETY
VELOCITY
INFORMATION
SHARING
COLLECTION
OPTIMIZATION
RELATIONAL
SPEED
SOFTWARE
BUSINESS
DATABASE
SEARCH
SOURCES
PETABYTES
STATISTICS
VISUALIZATION

Human Rights and the Democratization of Knowledge



http://www.springernature.com/gp/authors/research-data-policy

Open Science

National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25116.

DATA SHARING AND INNOVATION

- Open access
  - Accessible research & data to all members of society *(e.g., public, citizen scientists, and professionals)*
- Open data
- Open sources

NIH

Open Science

- Facilitates innovation of research tools and methods
- Increases statistical power
- Improves research quality through validation and replication

Courtesy: (Adapted from Paltoo, Harris, Luetkemeier, OSP-Office of Science Policy – NIH)

Open Science

- Facilitates innovation of research tools and methods
- Increases statistical power
- Improves research quality through validation and replication

Courtesy: (Adapted from Paltoo, Harris, Luetkemeier, OSP-Office of Science Policy – NIH)

Open Science

- Facilitates innovation of research tools and methods
- Increases statistical power
- Improves research quality through validation and replication

Courtesy: (Adapted from Paltoo, Harris, Luetkemeier, OSP-Office of Science Policy – NIH)

Open Science and Data Sharing


Courtesy of P. Kuhn (USC)

Precision Health

- Accounts for social and cultural complexity influencing underlying biology
- Requires
  - Biological understanding
  - Inclusion of social, cultural, and psychological factors
  - Scientific methods advancements
  - Instrumentation advancements
  - Technology advancements
  - Data management and computation advancements
- -omic, imaging, clinical, laboratory, etc data
- Can *change* disease classifications and treatments

Open Science and Data Sharing

- Increases scientific value by exploring, combining, and analyzing data from multiple sources

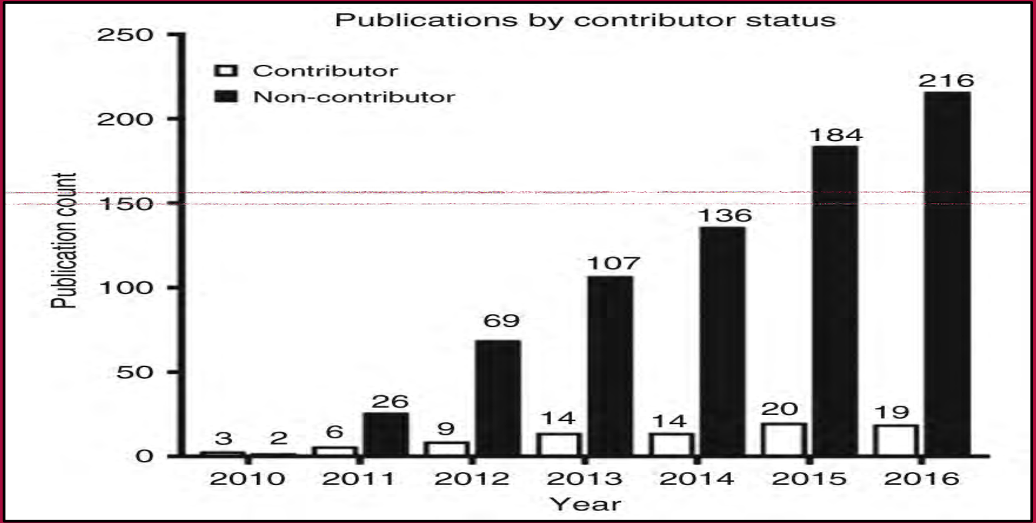
- Increases scale of studies, # publications, and types of scientists from a broader range of disciplines

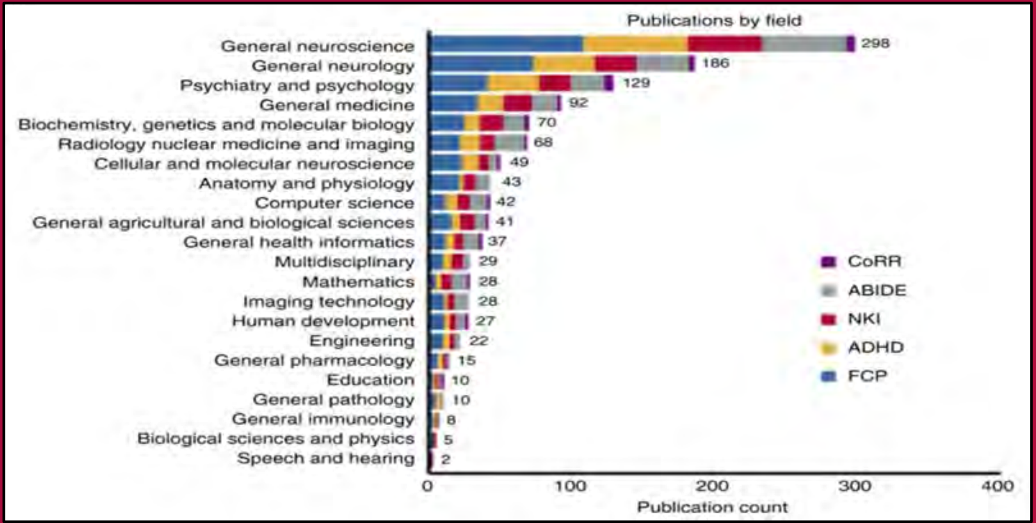
https://visual.ly/community/interactive-graphic/health/network-graph-human-diseases

Open Science and Data Sharing

- Increases scientific value by exploring, combining, and analyzing data from multiple sources

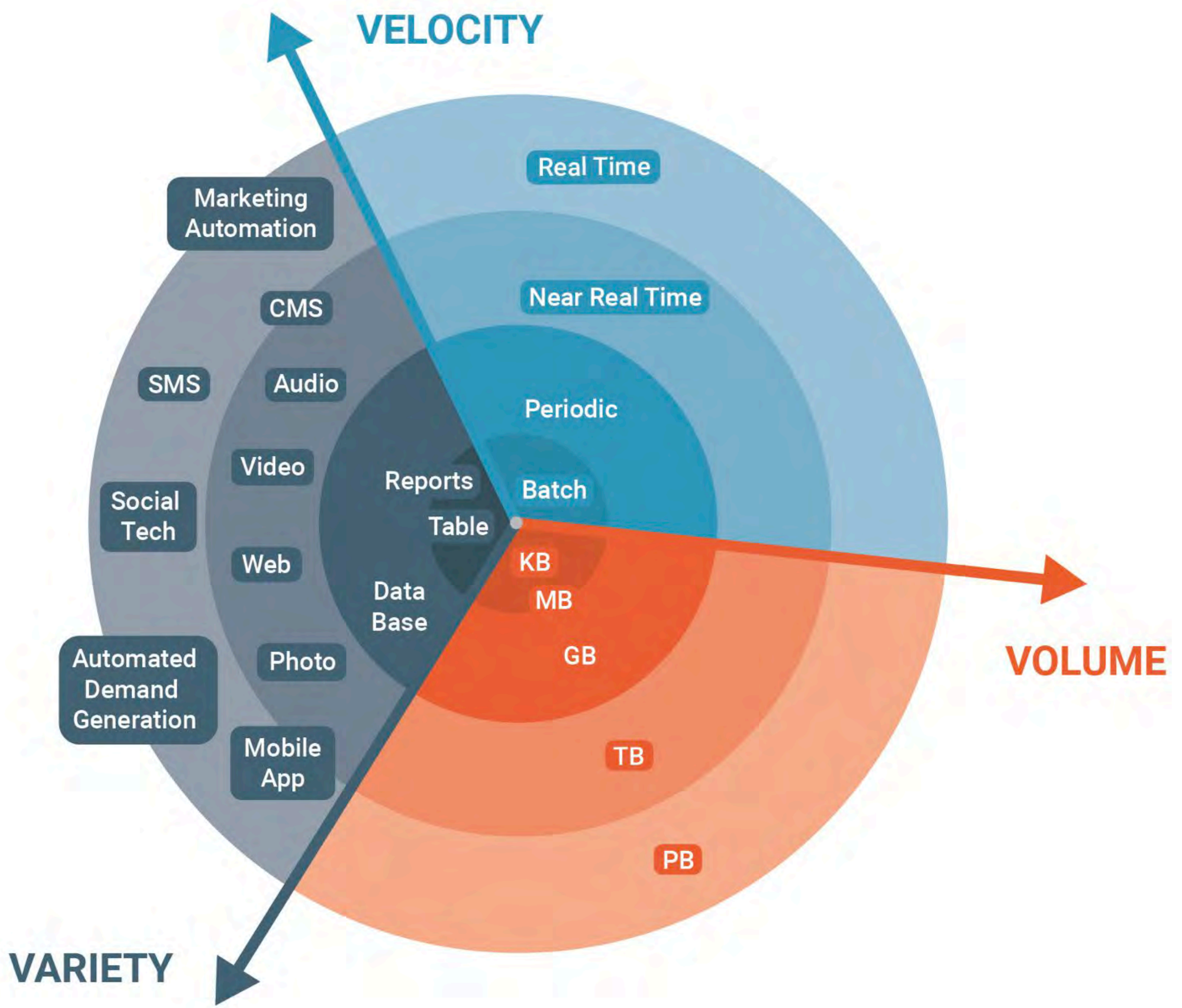
- Increases scale of studies, # publications, and types of scientists from a broader range of disciplines



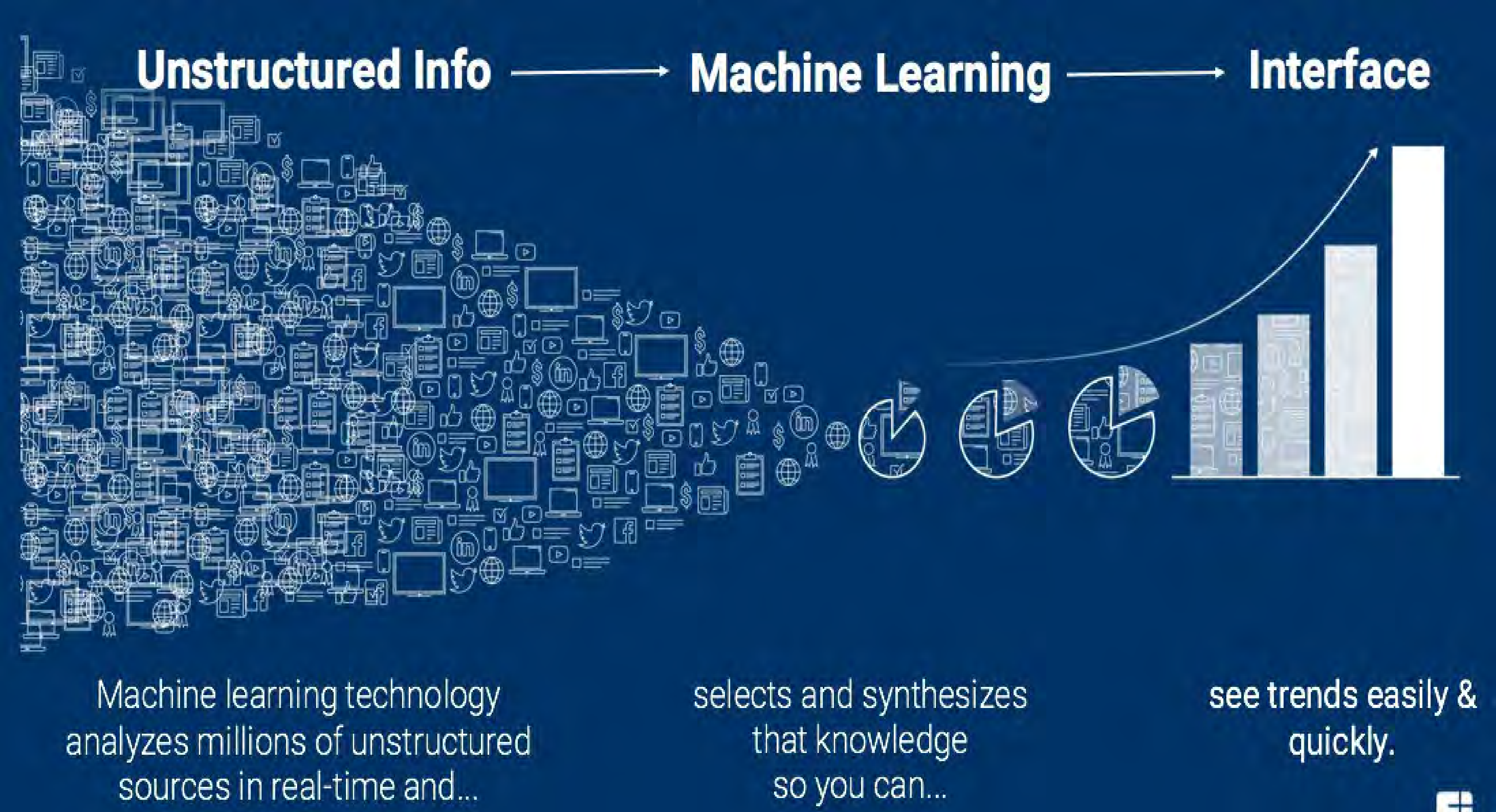


Milham, MP et al. (2018) Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications, 9, DOI: 10.1038/s41467-018-04976-1*

THE DATA



VELOCITY
Real Time
Near Real Time
Periodic
Batch
VOLUME
KB
MB
GB
TB
PB
VARIETY
Marketing Automation
CMS
SMS
Audio
Video
Social Tech
Web
Automated Demand Generation
Photo
Mobile App
Reports
Table
Data Base

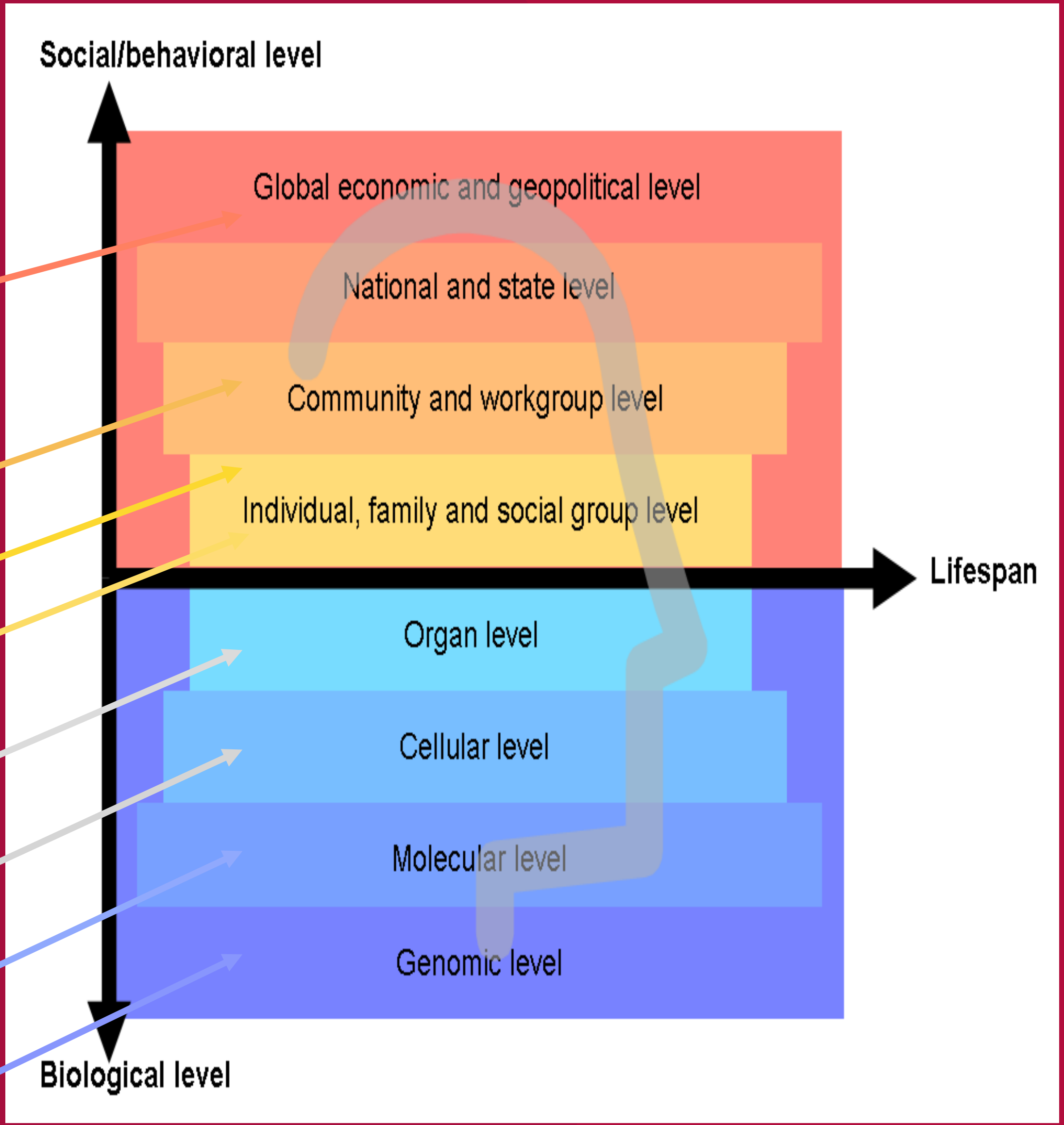


https://www.cbinsights.com/research/team-blog/data-network-effects/

BIOECOLOGICAL SYSTEM

Societal and cultural beliefs, attitudes and ideologies
Organizations, Schools
Neighborhoods, Communities
Family, group affiliations
Individual

Disease/health phenotypes
Stem, white blood cells, germline
Biomolecules (proteins, lipids)
DNA, RNA, gene expression



(Adapted from Glass, TA & McAttee, MJ. (2006) *Social Science & Medicine, 62, 1650-1671)*.

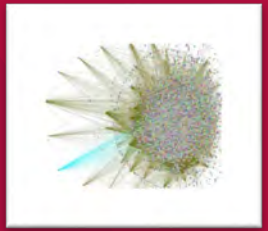
Data: Variety, Volume, Velocity, and Veracity

Real-time, Real-world Data Capacities

- Proteomics
- Metabolomics
- Microscopy
- Imaging
- Electronic Medical Records
- Mobile Devices
- Psychological/behavioral/ self-report
- Other technologies


Observational Phenotype

Genomic

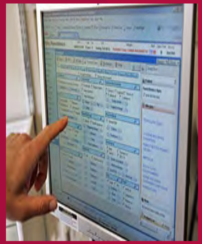


Other 'omic'

Location

Behavior

Imaging

Clinical/EMR

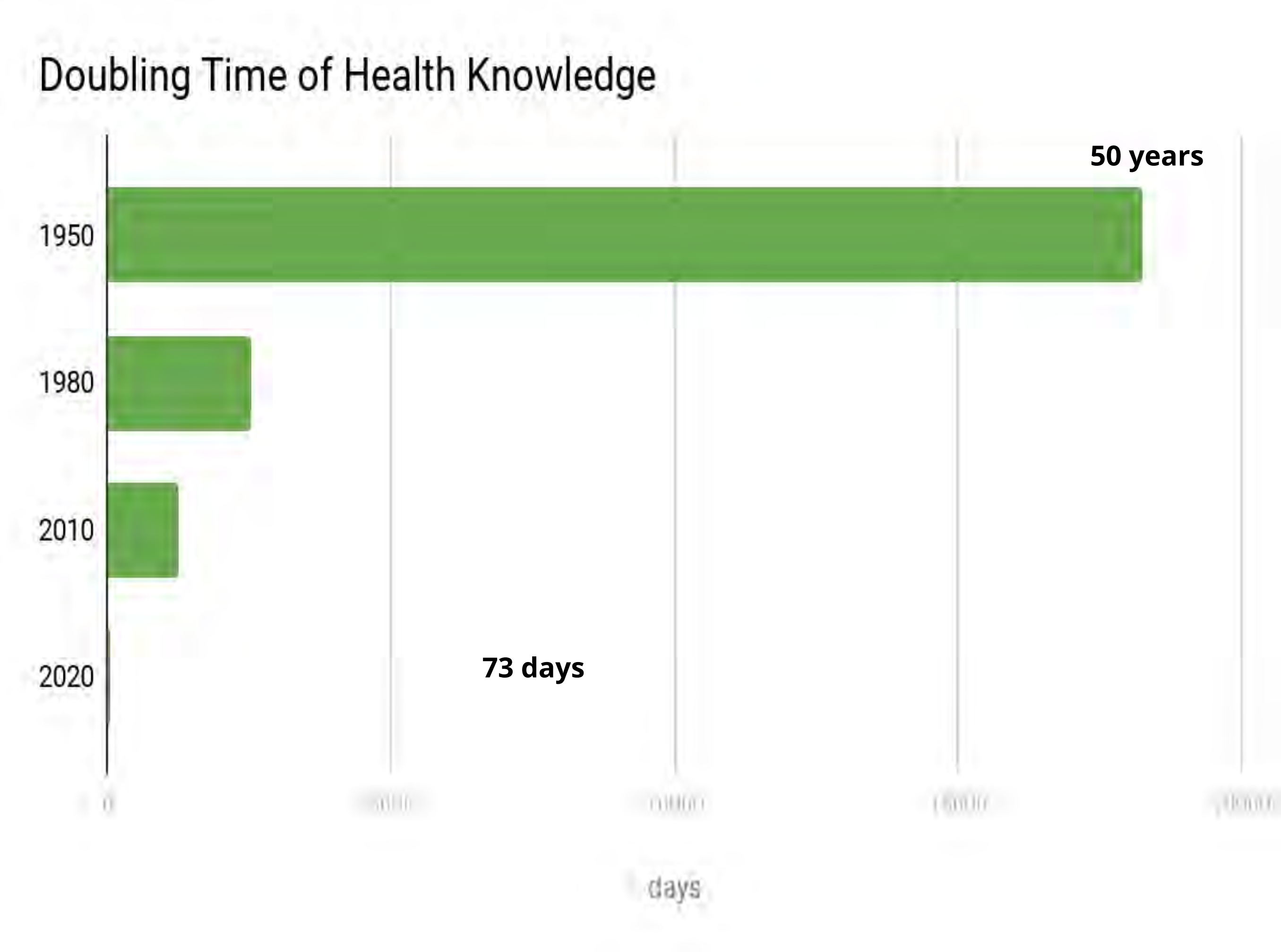
Exposures




Mobile Sensors

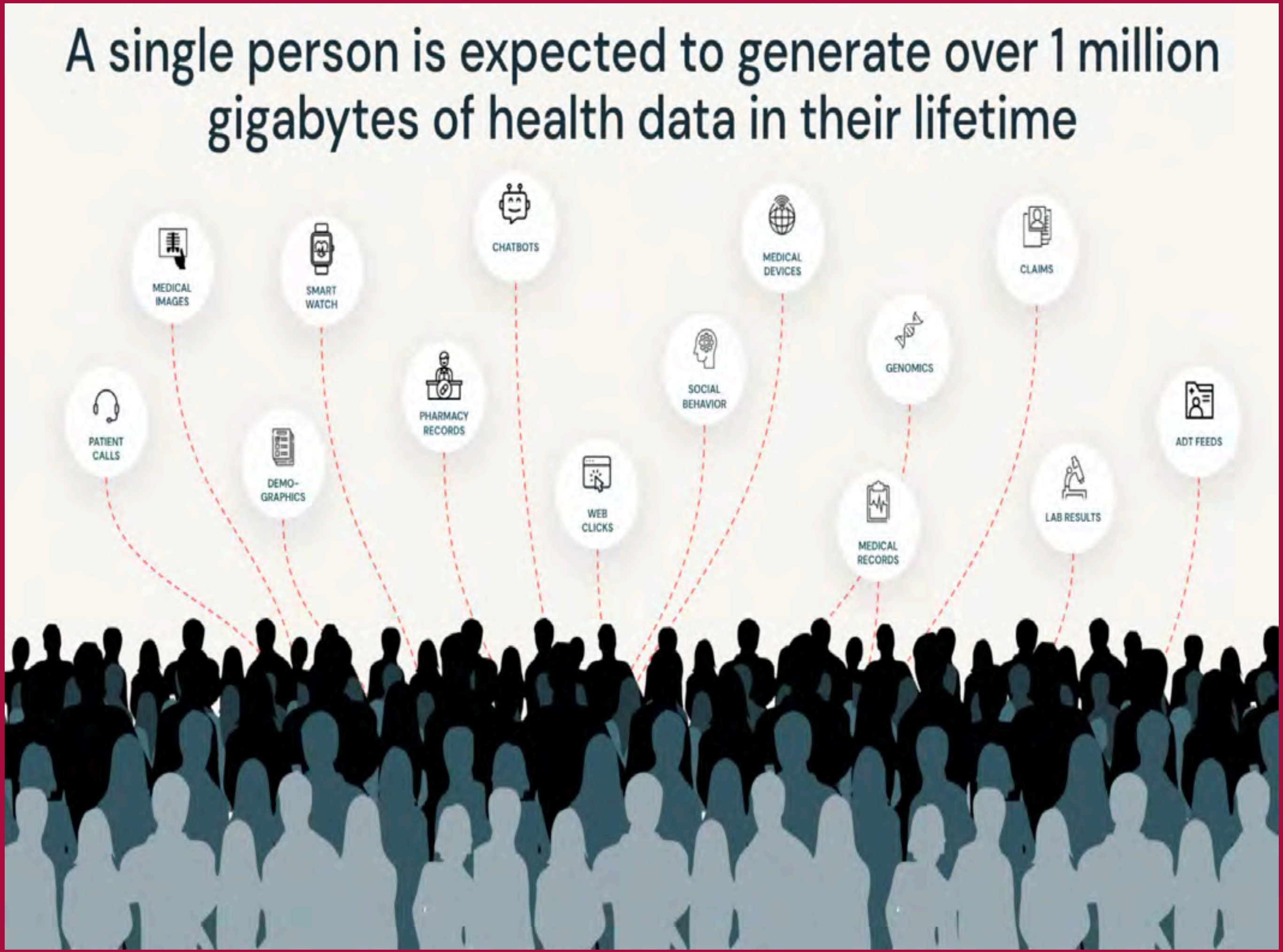
Data: Variety, Volume, Velocity, and Veracity

Health Knowledge Doubling Time



(Densen, P. *Trans Am Clin Climatol Assoc* (2011)

Data: Variety, Volume, Velocity, and Veracity
Personal Health Data



https://databricks.com/blog/2022/03/09/introducing-lakehouse-for-healthcare-and-life-sciences.html

Data: Variety, Volume, Velocity, and Veracity

Improvements to Store and Process Data

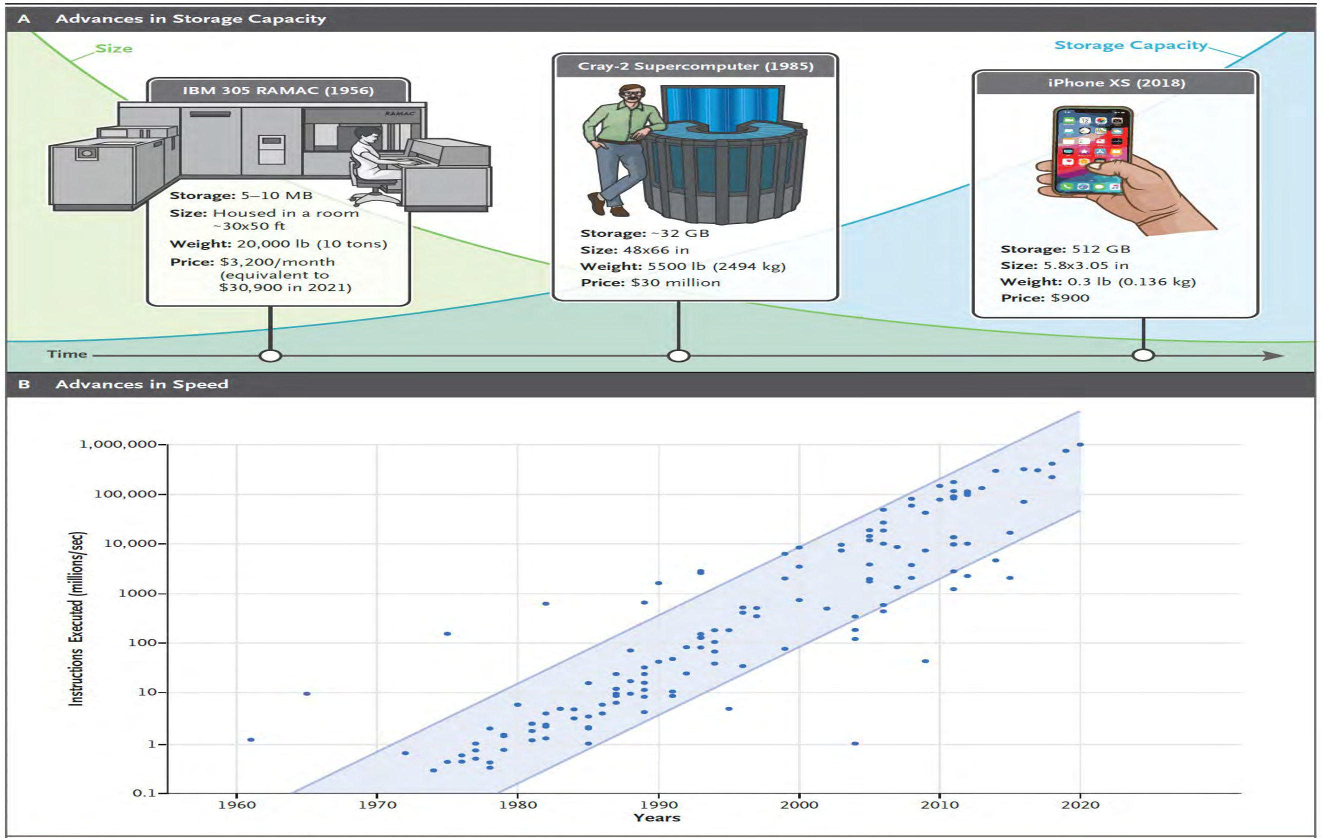


Figure 1. Improvements over 50 Years in the Ability of Computers to Store and Process Data.
Panel A shows advances in data storage, in terms of both physical size and cost per unit of storage. RAMAC denotes random access method of accounting and control. Panel B shows advances in the speed of computing. Each dot represents an individual machine type and the approximate year of its introduction. These improvements in storage and speed have allowed machine learning to progress from a dream to reality. Data in both panels are estimates from many types of system architecture and are derived from multiple public sources.

(N Engl J Med 2023;388:1201-8. DOI: 10.1056/NEJMra2302038)

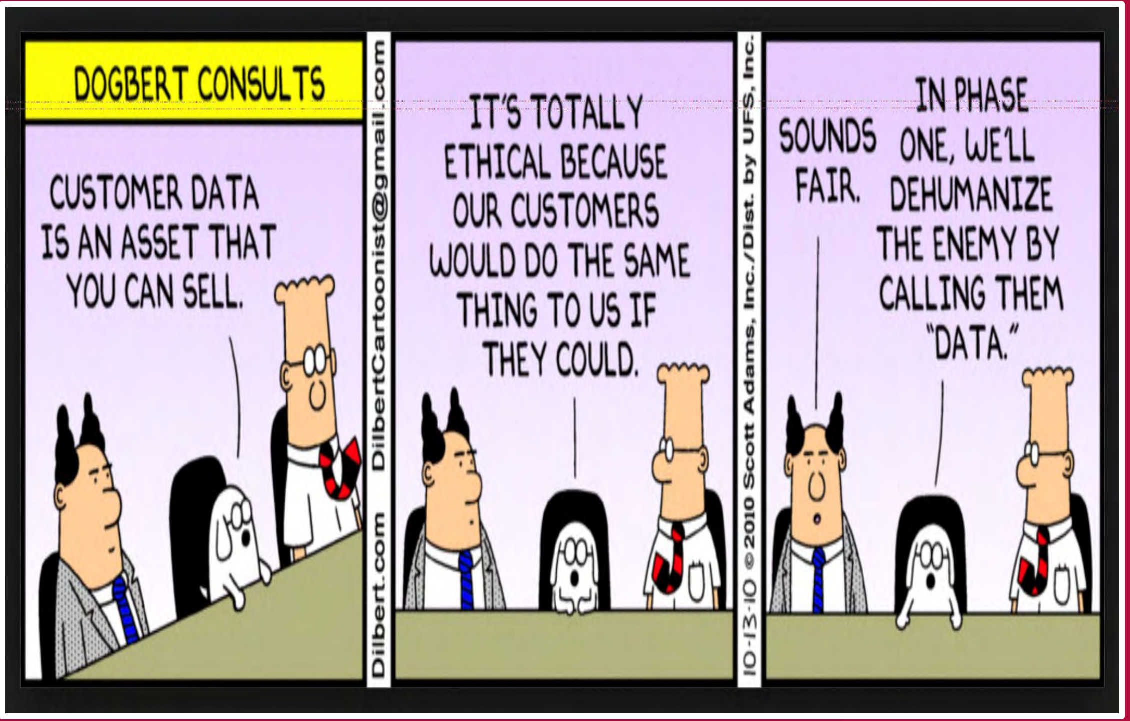
Data: Variety, Volume, Velocity, and Veracity

- Trustworthy
- Accurate
- Reliable

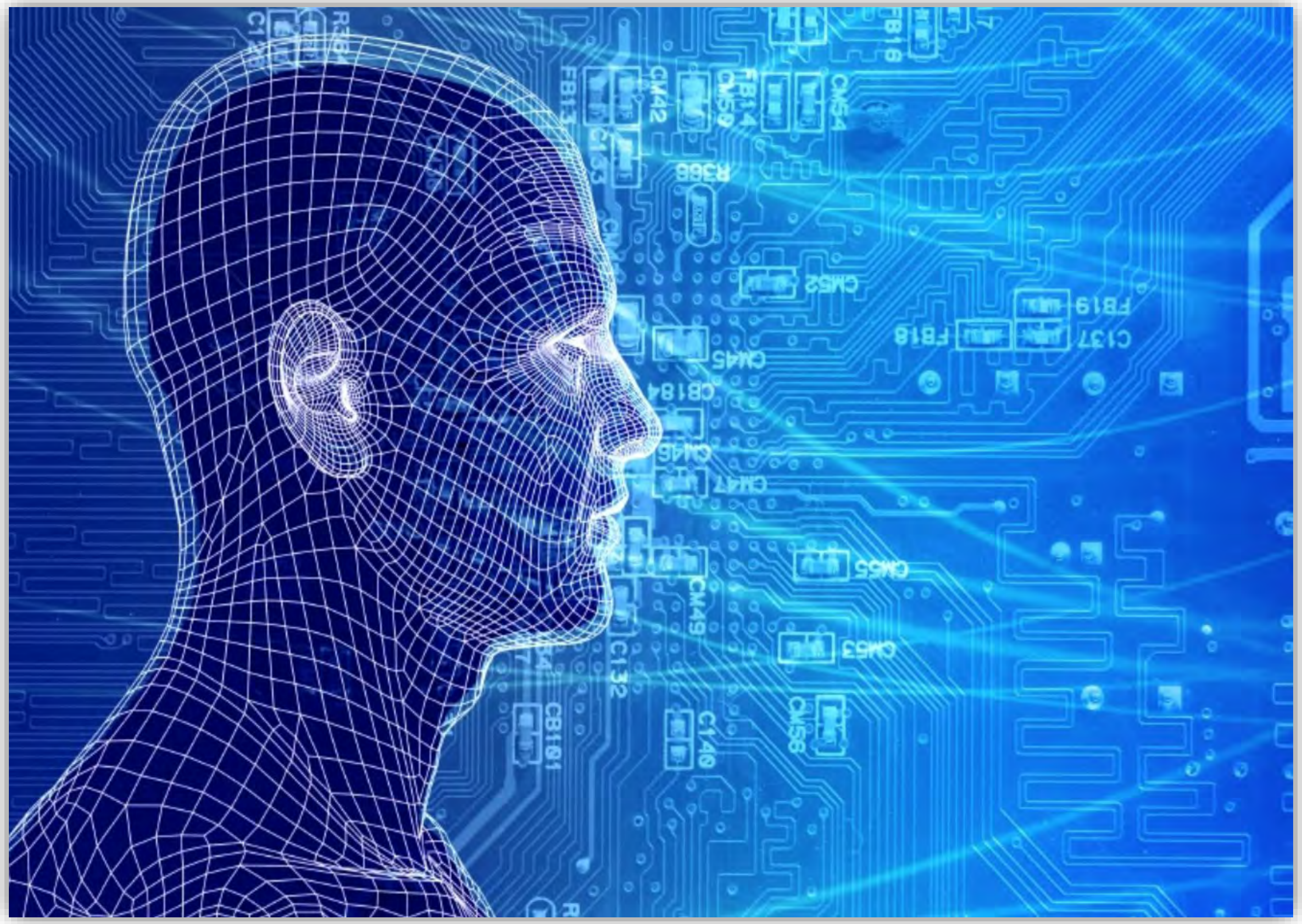
[Information, information uses], and "definitions belong to the definers, not the defined

- Toni Morrison

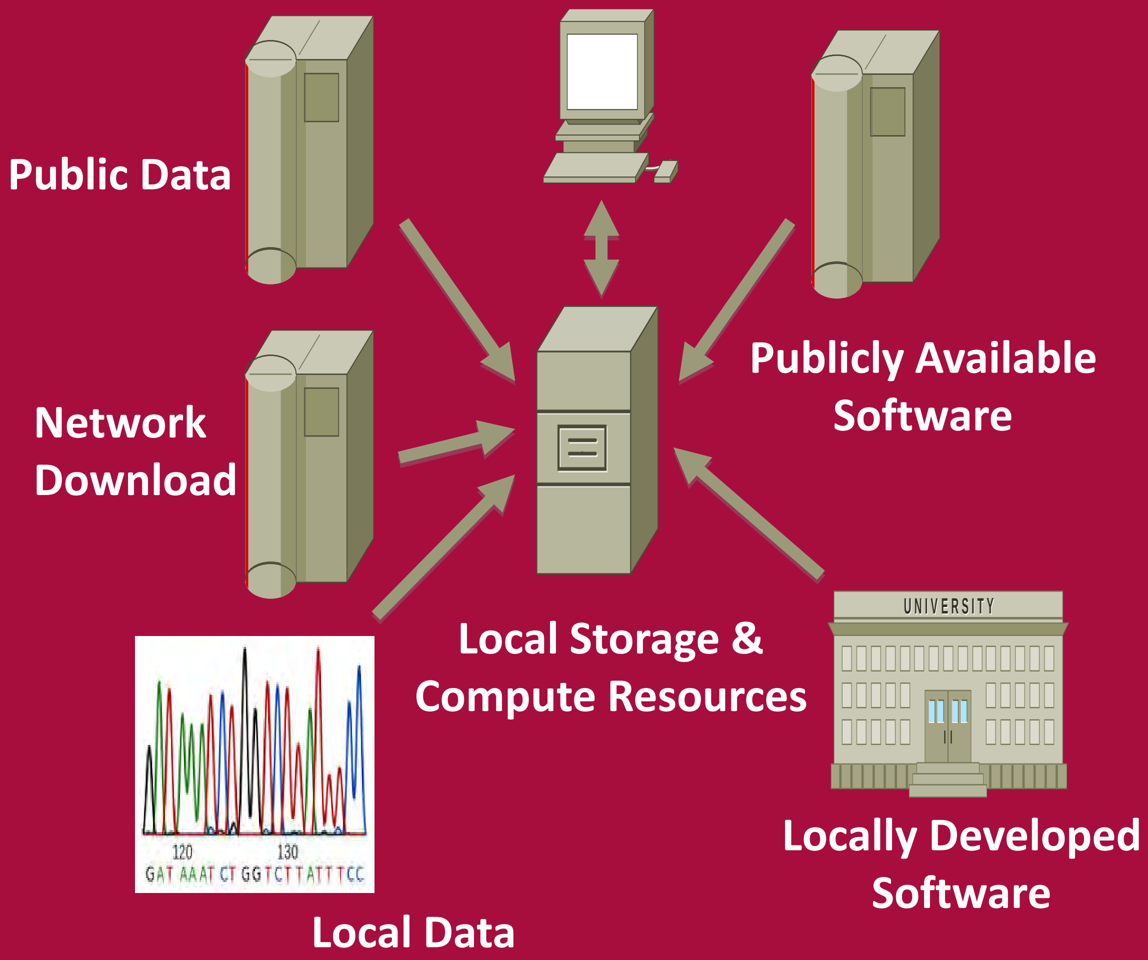
Beloved



SOME
SCIENCE
DATA SCIENCE



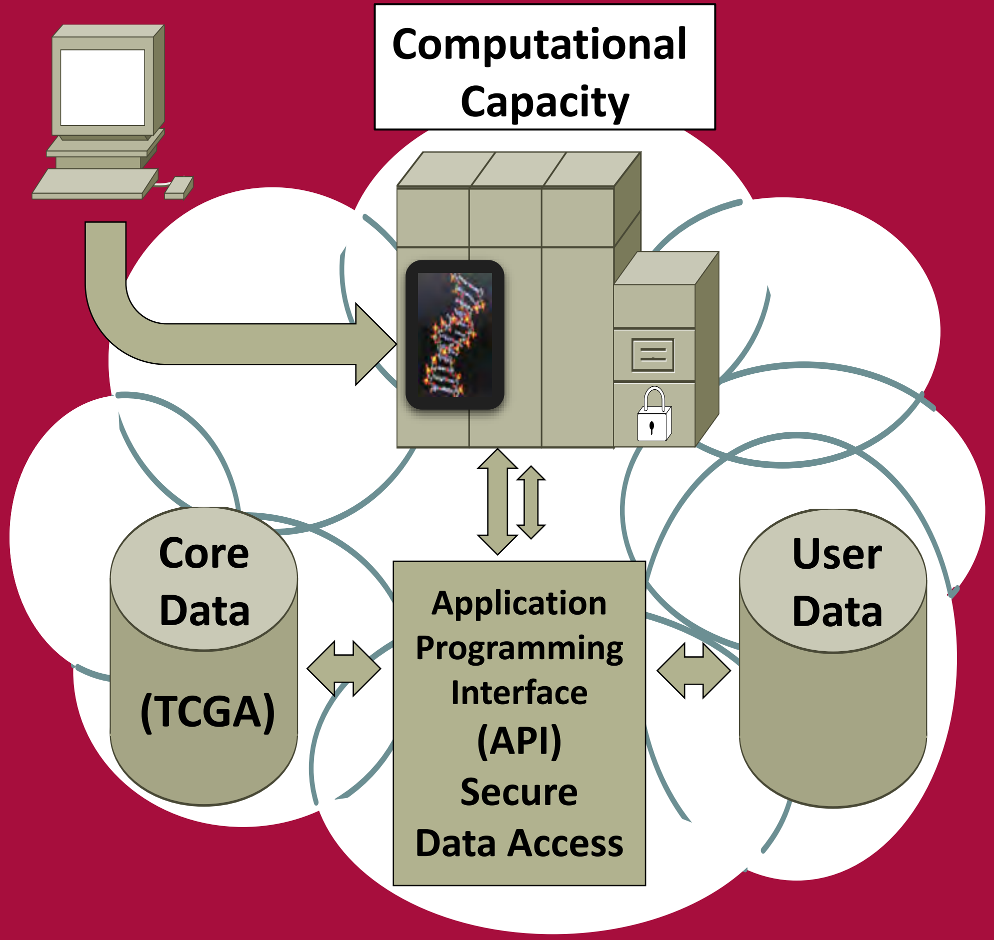
Download Computation Resources and Data
(circa ~pre-2014)



- Only large institutions had ability to utilize data
- Storage/data protection cost~$2M/yr
- Data download at 10 Gb/second (*23 days*)
- Increase rate of data generation

(Courtesy: Adapted from A. Kerlavage, NCI – NIH)

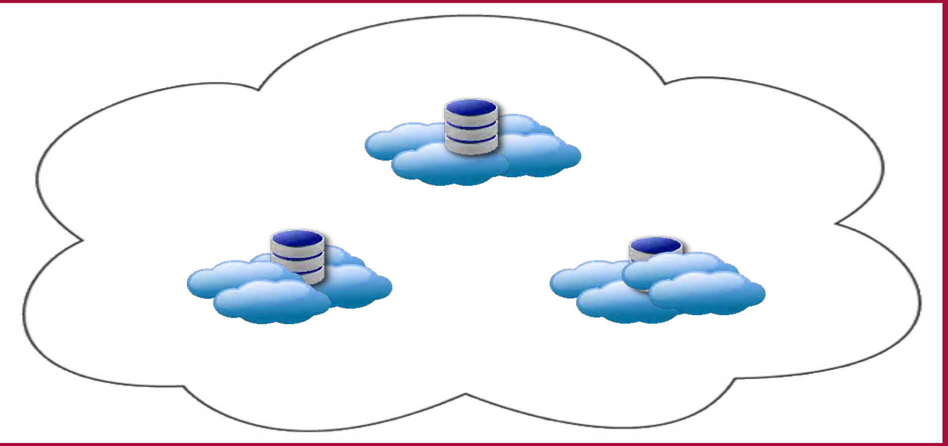
Co-Locate Computation Resources and Data
(circa ~2014-2016)

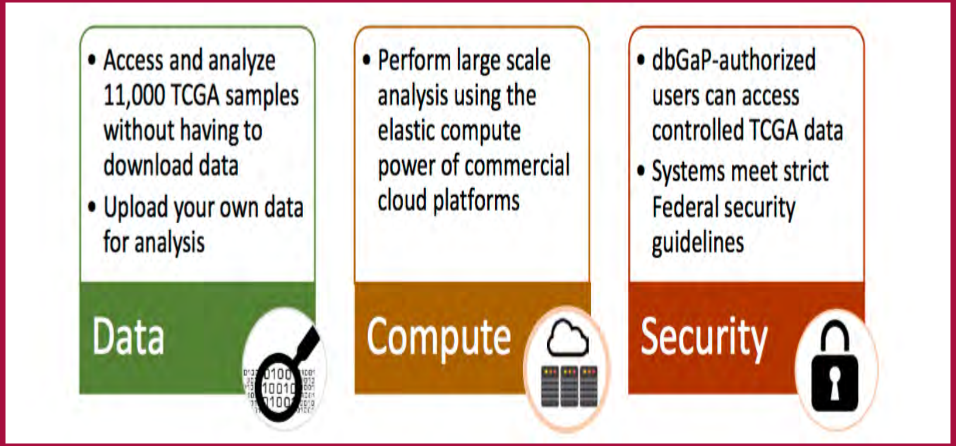


- Access large data sets without downloading data
- Bring tools and pipelines to data
- Combine own data and analyze with existing data
- Workspace to save, share, and analyze data

(Courtesy: Adapted from A. Kerlavage, NCI – NIH)

Co-Locate Computation Resources and Data
(circa 2017 – now)

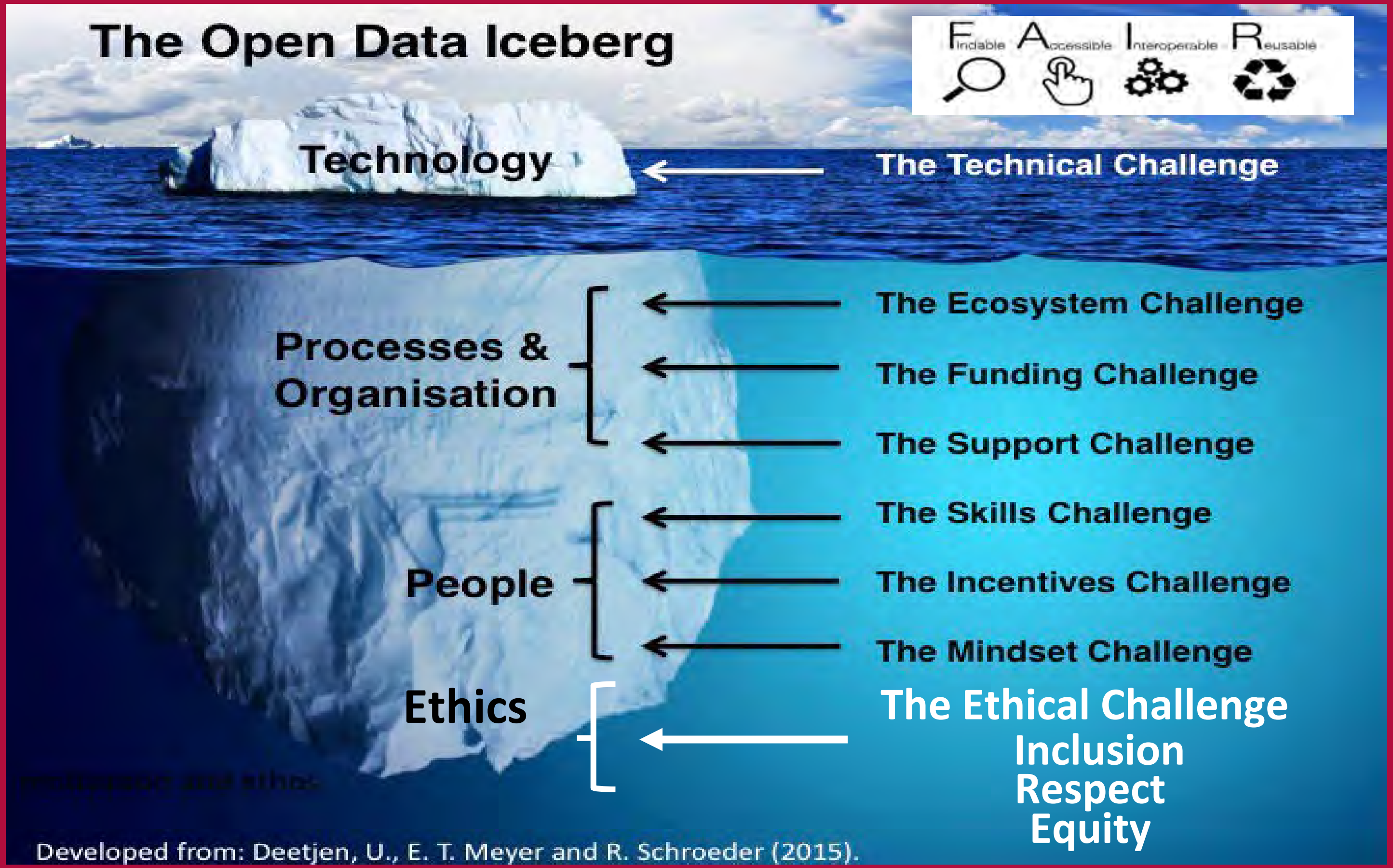




- Democratize data sharing and access
- Cost-effective Scalable Computational Capacity

(Courtesy: Adapted from A. Kerlavage, NCI – NIH)

Data Reuse:
Isn't Only a Data and Technology Challenge



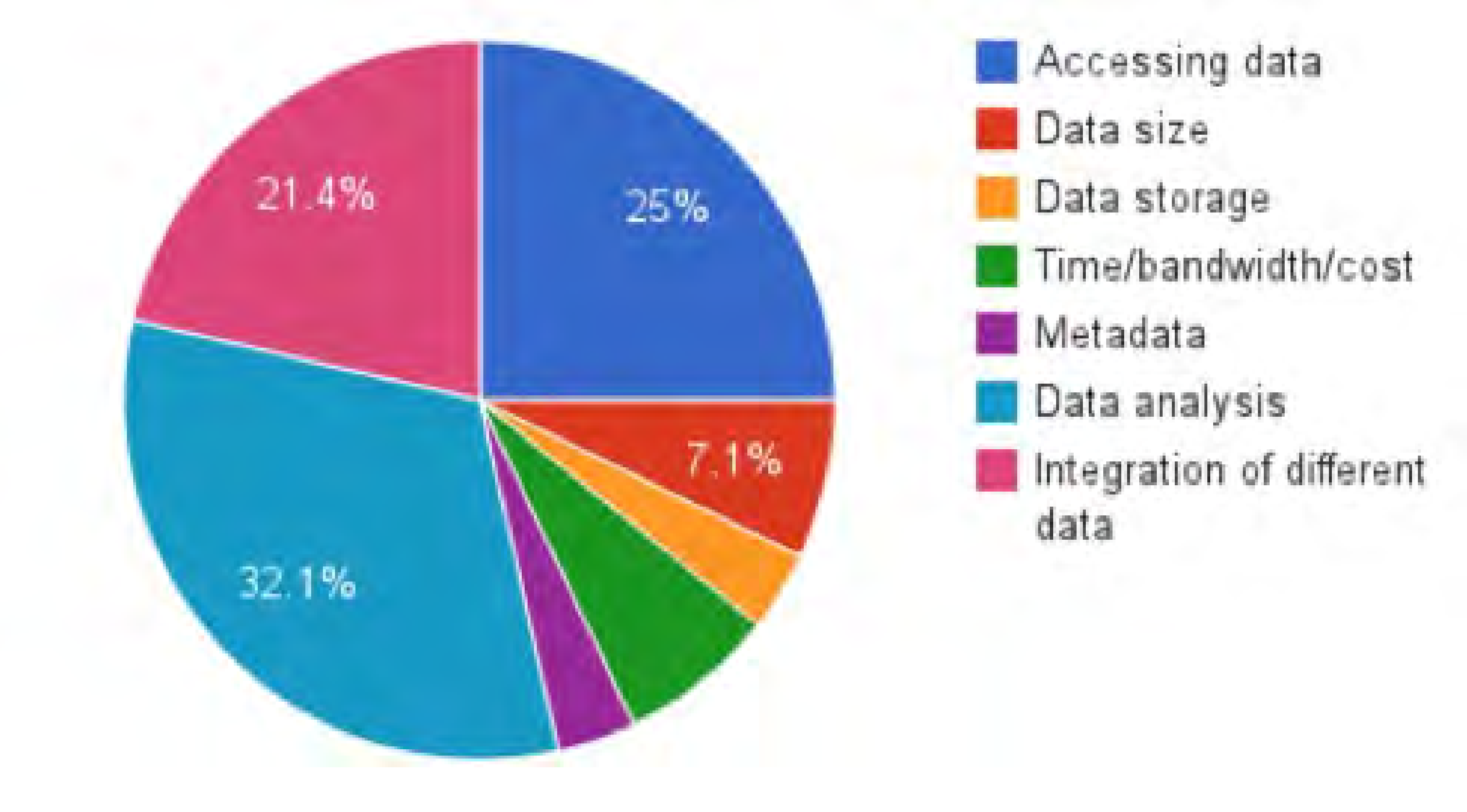
Wilkinson, M. D. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*3:160018 doi: 10.1038/sdata.2016.18

[1] F.A.I.R. https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pd

[2] C.A.R.E https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pdf

Data Reuse:
Isn't Only a Data and Technology Challenge

Difficulties Accessing, Analyzing, and Integrating Data

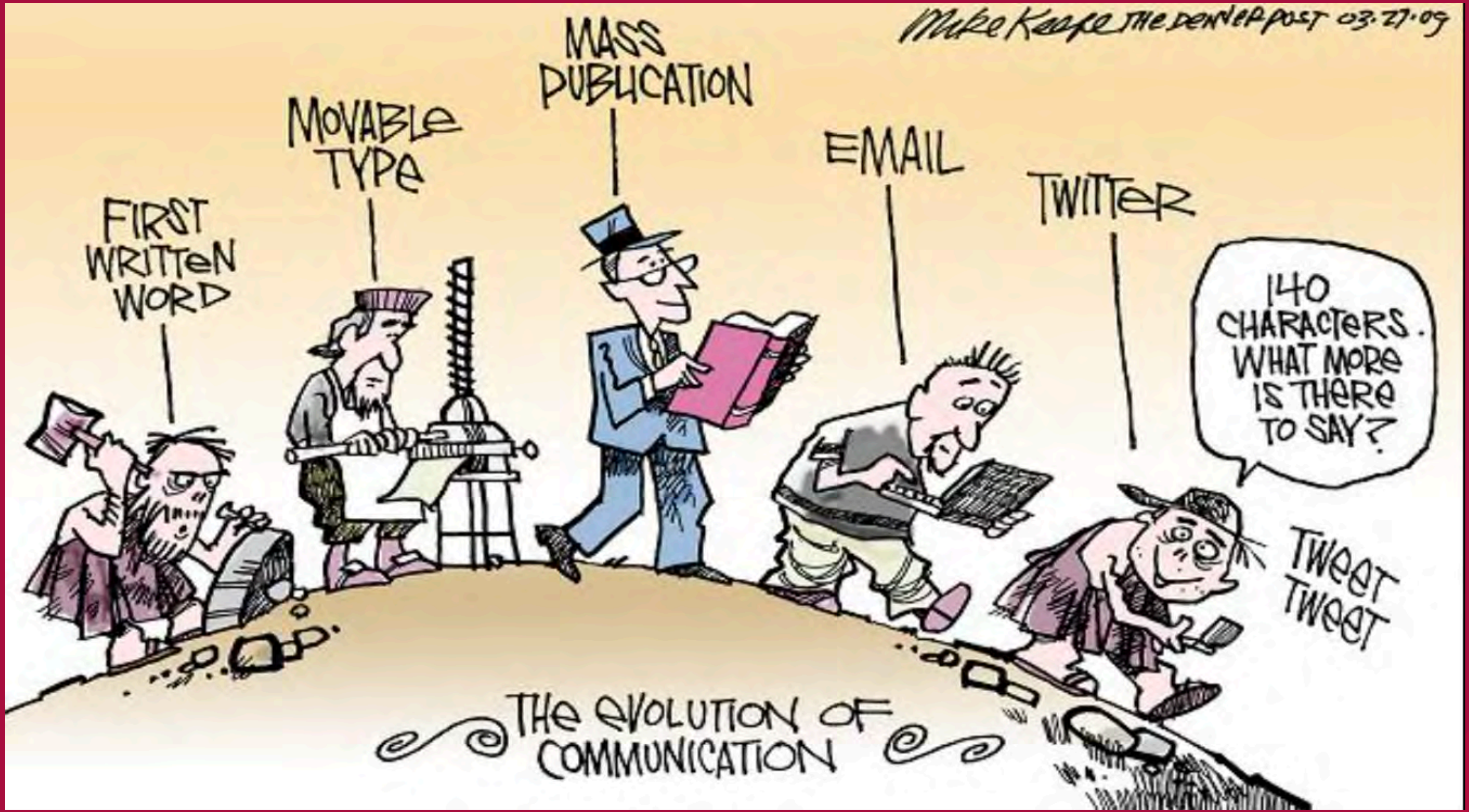


Data Reuse:
Isn't Only a Data and Technology Challenge

Difficulties Accessing, Analyzing, and Integrating Data

Data Reuse:
Isn't Only a Data and Technology Challenge

Communication

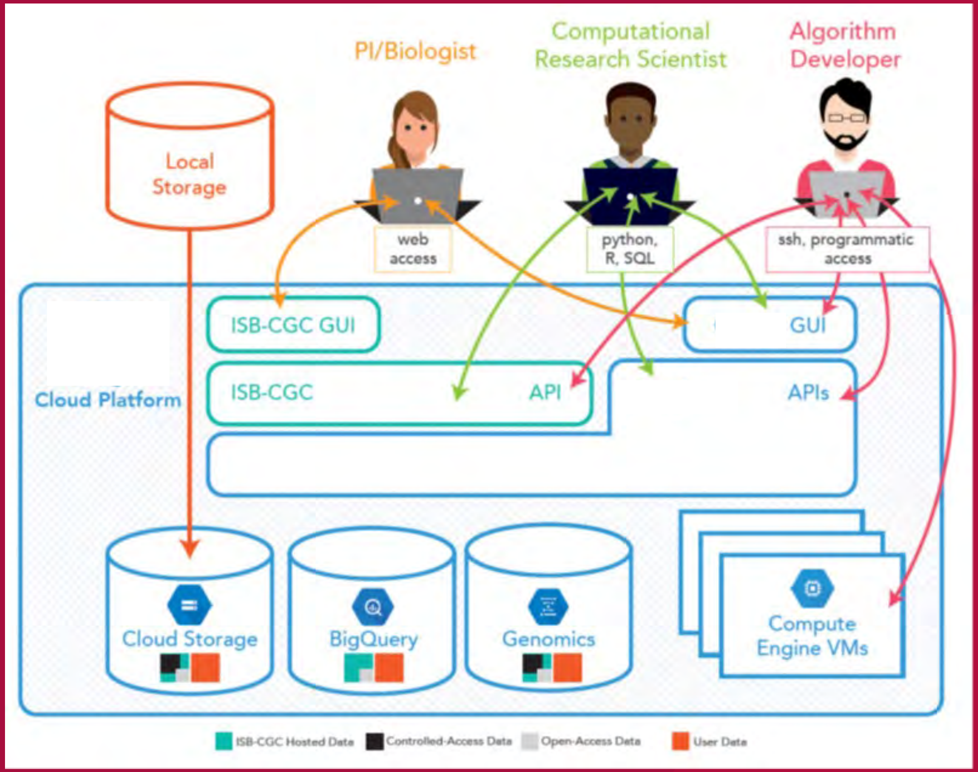


Data Reuse:
Isn't Only a Data and Technology Challenge

Multidisciplinary Teams with Diverse Expertise and Resources

Biology/Social/Psychology Researcher

- select a data subset based on clinical, molecular, clinical characteristics
- explore all data for a specific pathways or models
- compare one cohort to another
- upload a small private dataset to analyze in conjunction with existing dataset



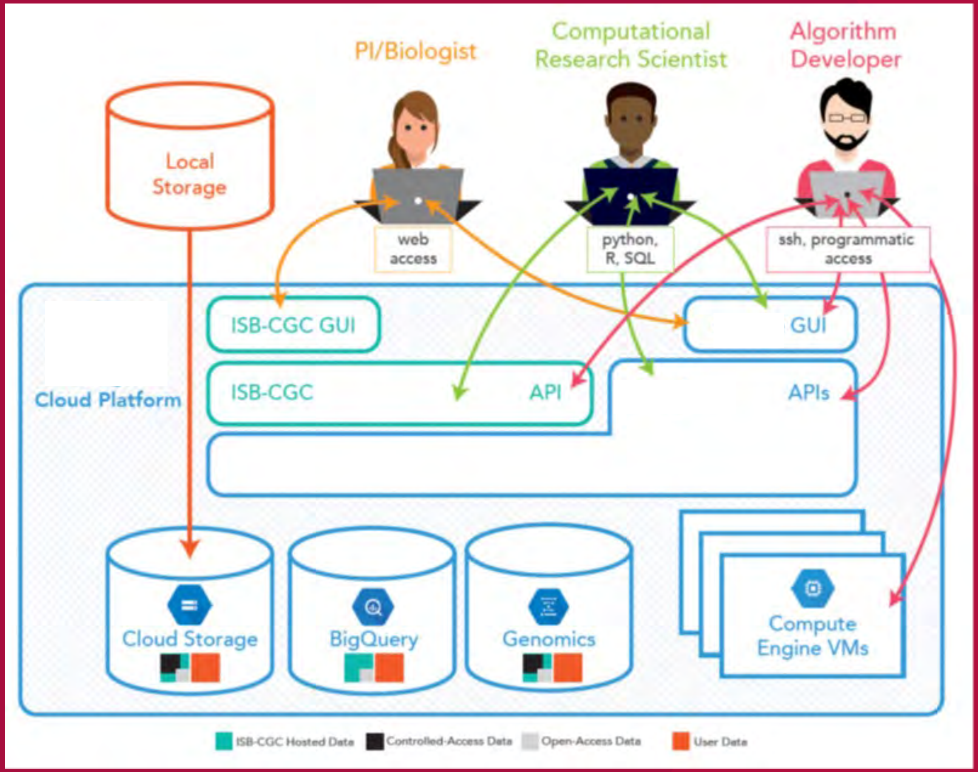
(Courtesy: Adapted from A. Kerlavage, CBIIT-NCI-NIH)

Data Reuse:
Isn't Only a Data and Technology Challenge

Multidisciplinary Teams with Diverse Expertise and Resources

Biology/Social/Psychology Researcher

Computational Scientist
- interactive data exploration
- use R or Python to perform custom analyses
- develop new tools
- share new tools
- publish new tools (including interactive)
- develop/customize pipelines



(Courtesy: Adapted from A. Kerlavage, CBIIT-NCI-NIH)

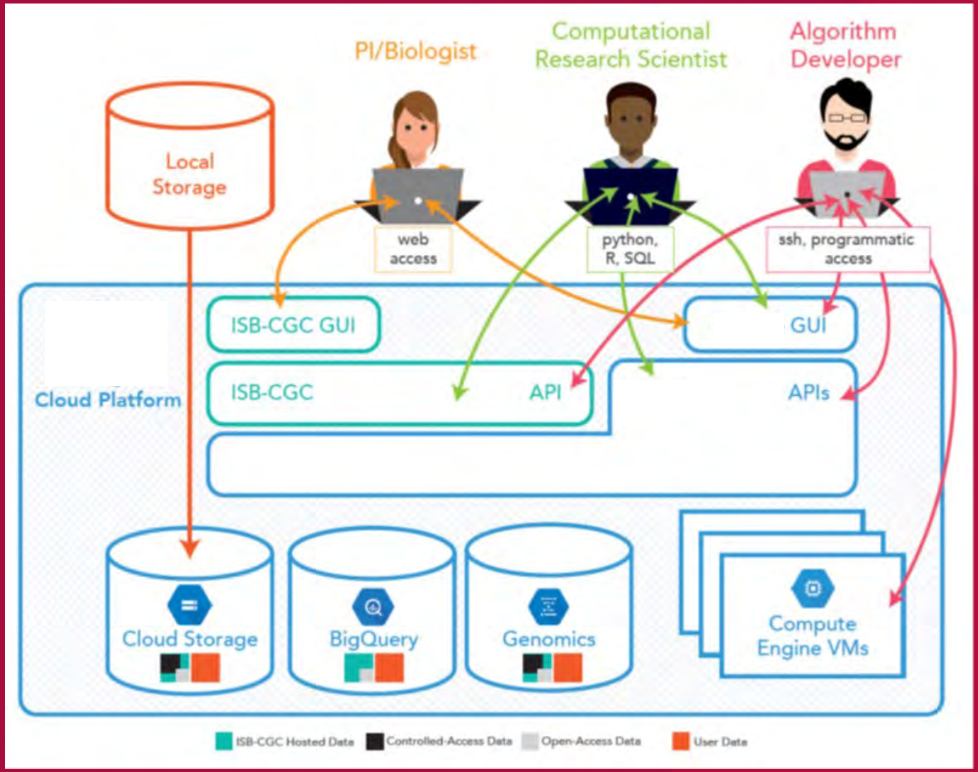
Data Reuse:
Isn't Only a Data and Technology Challenge

Multidisciplinary Teams with Diverse Expertise and Resources

Biology/Social/Psychology Researcher

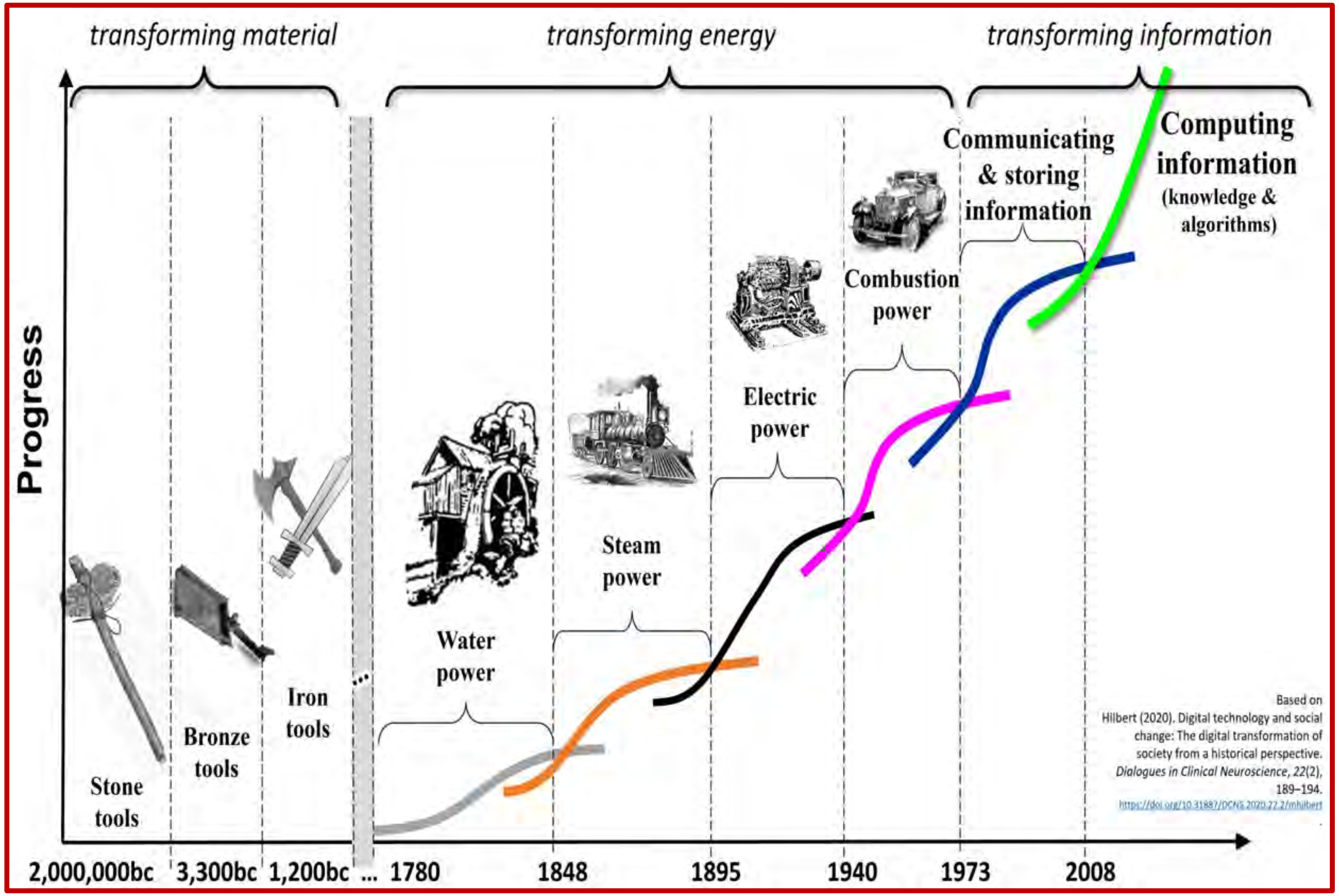
Computational Scientist

Algorithm Developer
- test new algorithm on hundreds or thousands of data files
- run novel image segmentation method across whole-slide images



(Courtesy: Adapted from A. Kerlavage, CBIIT-NCI-NIH)

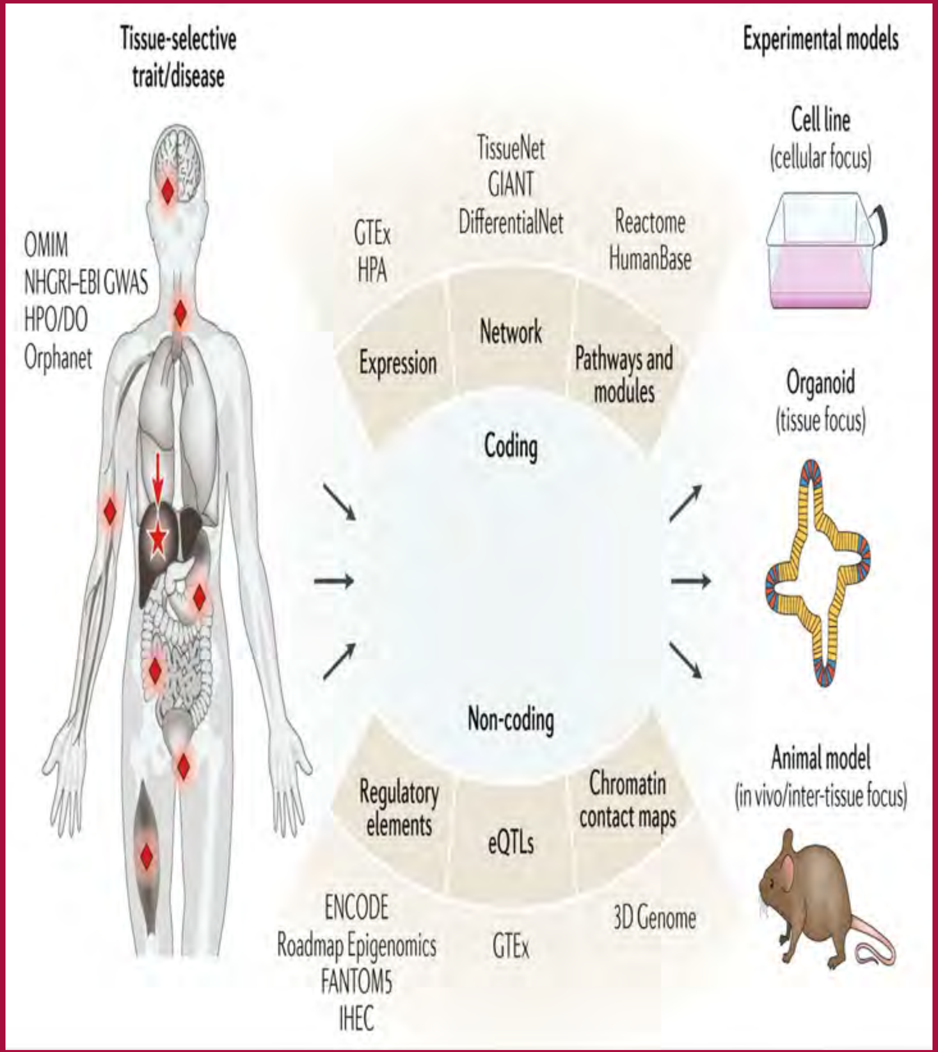
INNOVATION AND REVOLUTION



The Data Revolution

- Research is now a *data intensive enterprises*
- *Rapidly changing* scale
- Technology, *data computing & information technology (IT)* are *pervasive* in the *lab, clinics, and homes*

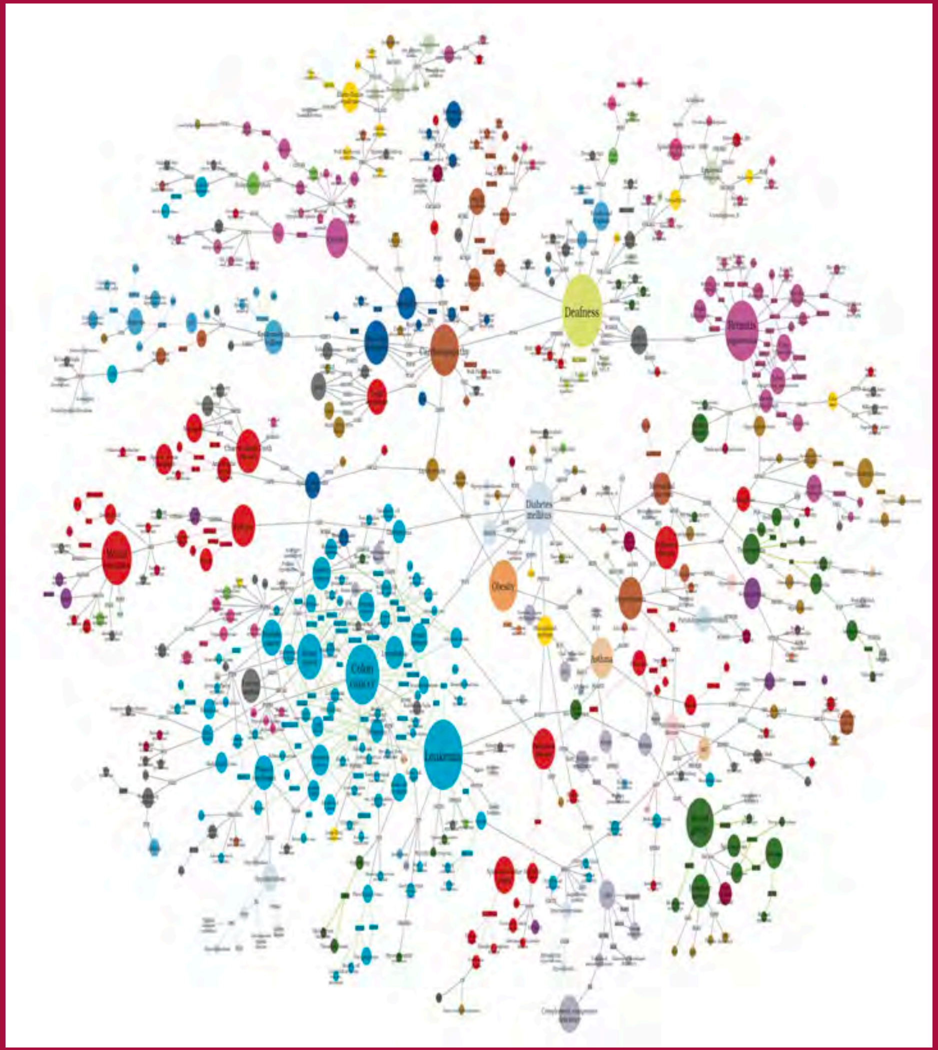
The Data Revolution



Hekselman, I. and Yeger-Lotem, E. (2020). *Nature Reviews Genetics,* (21) 137-150. https://doi.org/10.1038/s41576-019-0200-9

- *Changing functional roles of genes across tissues*
- *Relationships among diseases*
- *Relationships of behavior and health*

The Data Revolution

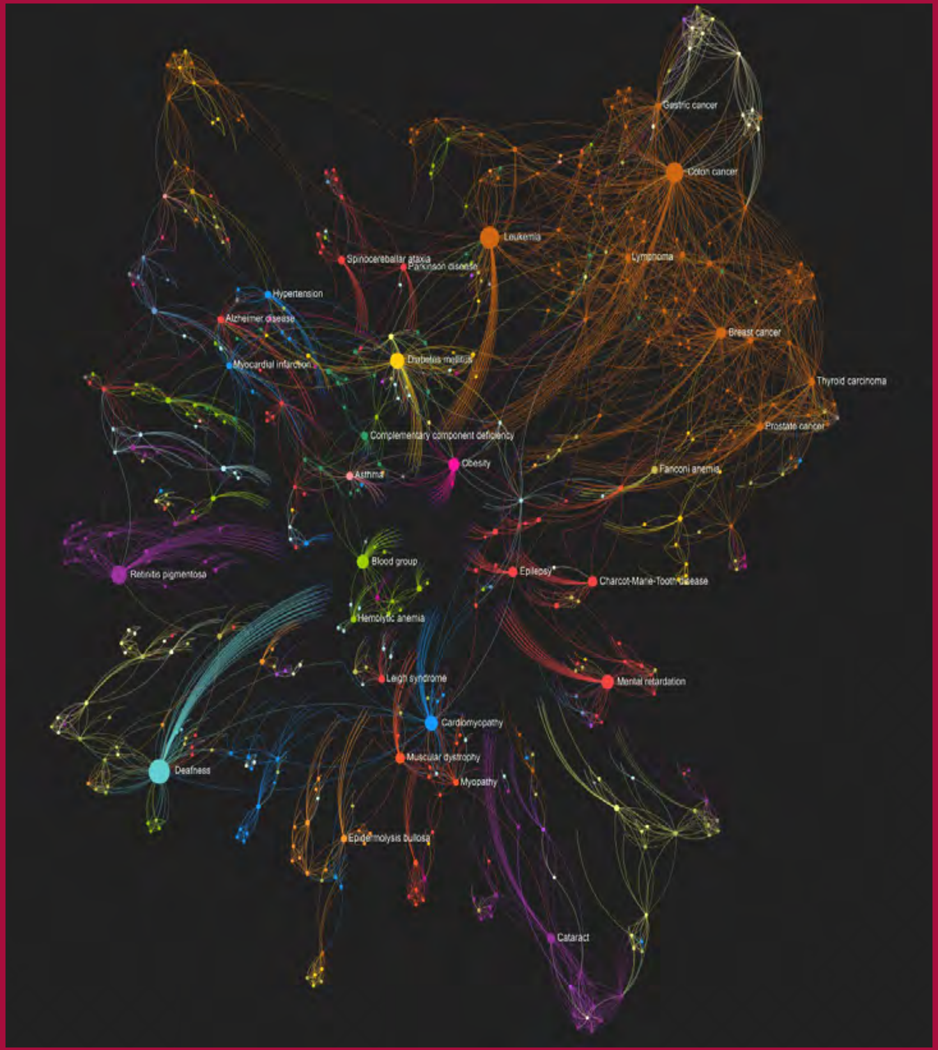


Goh, K., Cusick, ME, Valle, D., Childs, B., Vidal, M., and Barabási, A-L.
PNAS May 22, 2007 104 (21) 8685-8690; https://doi.org/10.1073/pnas.0701361104

- *Changing functional roles of genes across tissues*
- *Relationships among diseases*
- *Relationships of behavior and health*

NIH

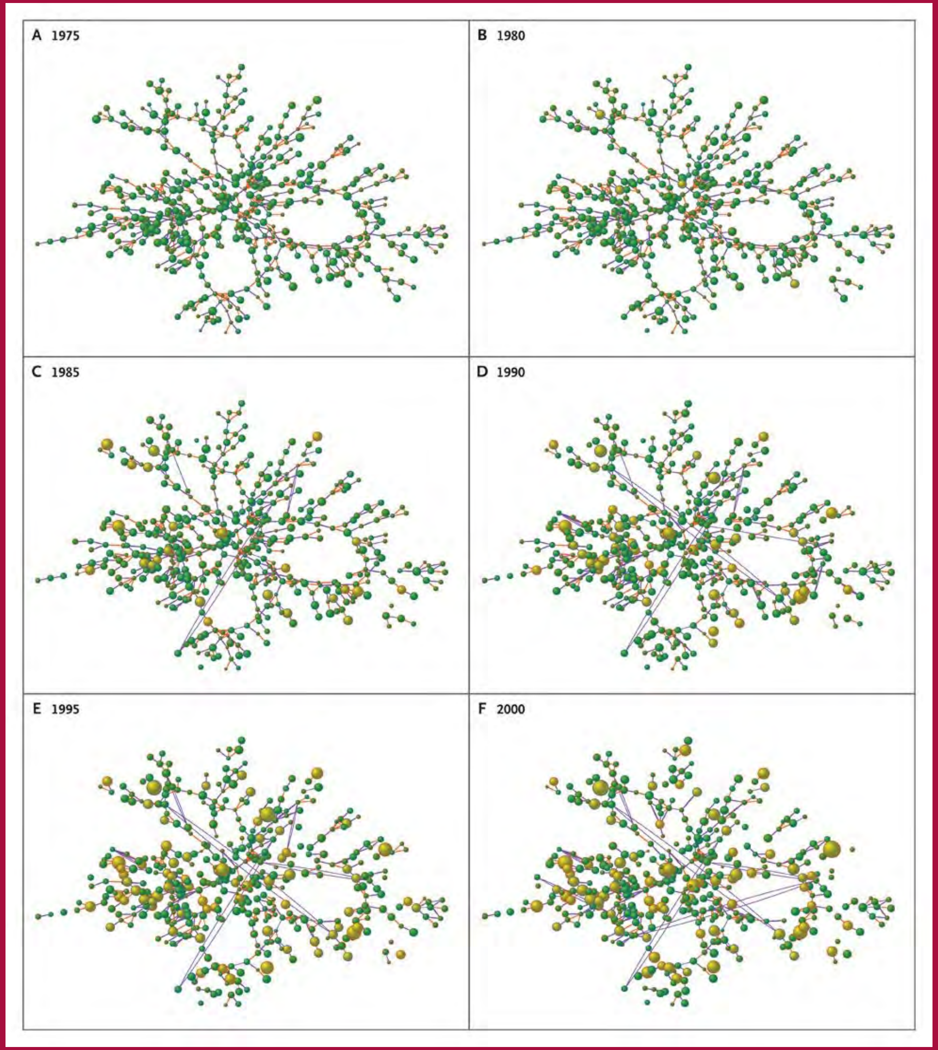
The Data Revolution



Goh, K., Cusick, ME, Valle, D., Childs, B., Vidal, M., and Barabási, A-L.
PNAS May 22, 2007 104 (21) 8685-8690; https://doi.org/10.1073/pnas.0701361104

- *Changing functional roles of genes across tissues*
- *Relationships among diseases*
- *Relationships of behavior and health*

The Data Revolution

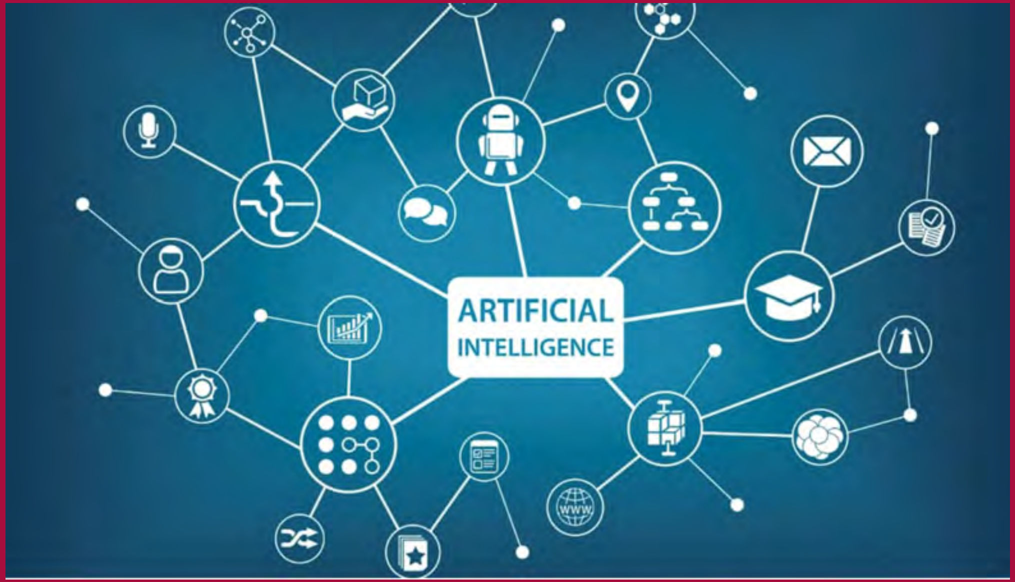


Christakis NA, Fowler JH. N Engl J Med 2007;357:370-379

- *Changing functional roles of genes across tissues*
- *Relationships among diseases*
- *Relationships of behavior and health*

NIH

The Data Revolution



- Large amounts of data and data types
  - *Mobile devices, tracking systems, RFID, sensor networks, social networks, Internet searches, electronic medial records, video archives, e-commerce*
- Secondary analyses of primary and derived data
- Identify trends
- Improve research quality

THE BIG DATA REVOLUTION


FiveThirtyEight

abcNEWS

Politics Sports **Science** Podcasts Video Interactives

Science Isn't Broken

It's just a hell of a lot harder than we give it credit for.

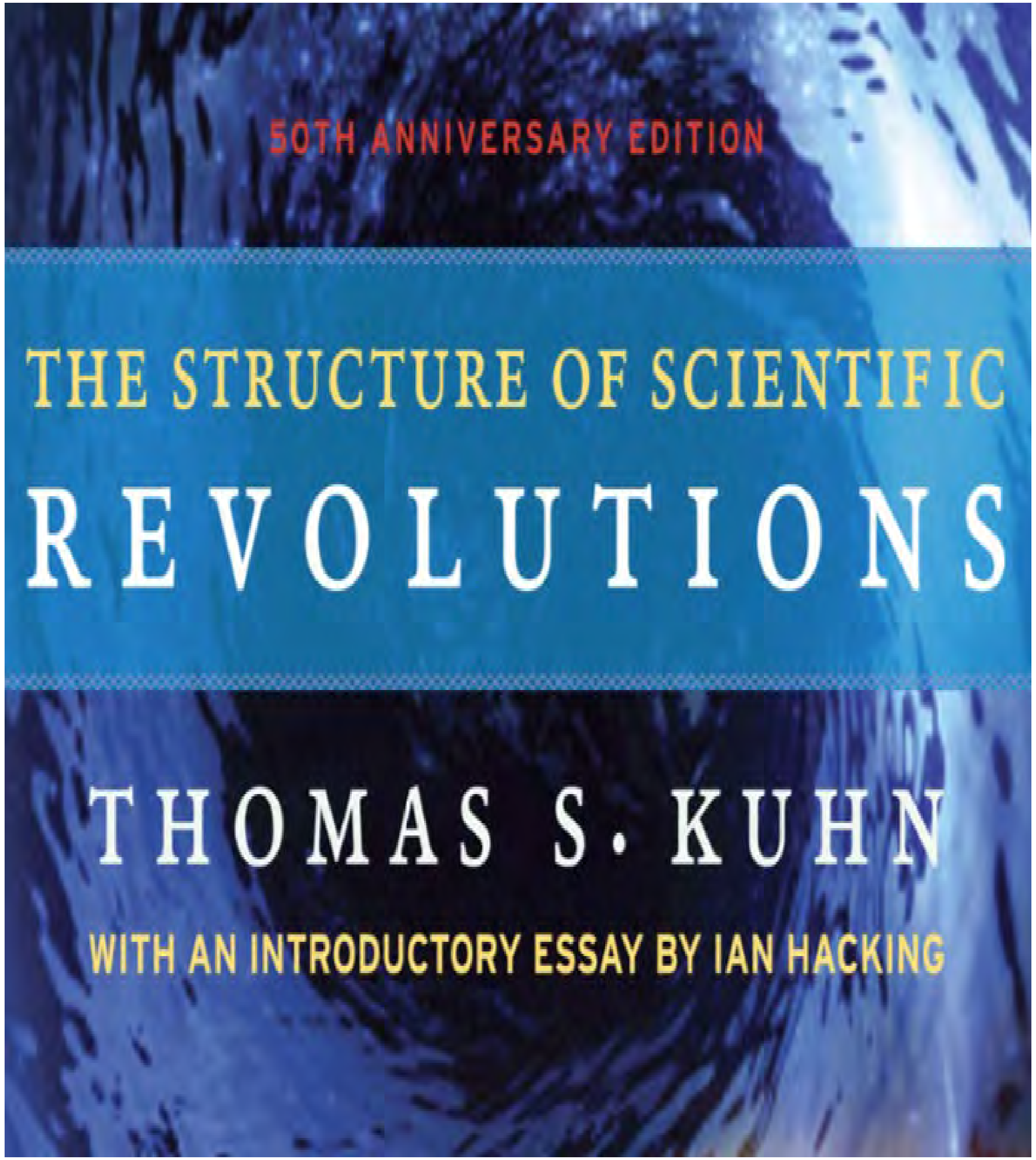
Graphics by Ritchie King

Filed under Scientific Method

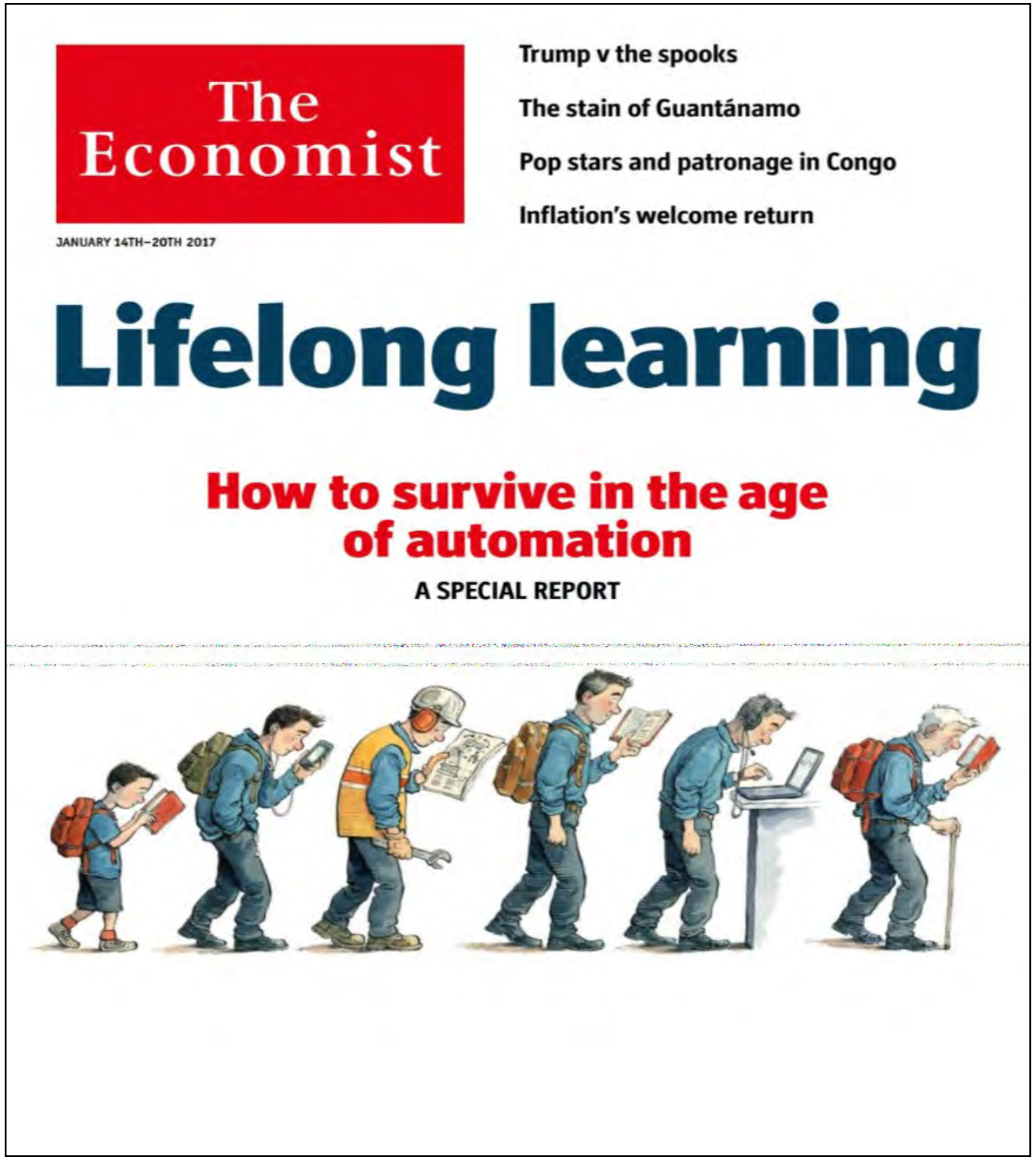
Published Aug. 19, 2015

ILLUSTRATION BY SHOUT

The Paradigm Shift



The Paradigm Shift


The Economist

Trump v the spooks

The stain of Guantánamo

Pop stars and patronage in Congo

Inflation's welcome return

JANUARY 14TH–20TH 2017

Lifelong learning

How to survive in the age of automation

A SPECIAL REPORT

The Paradigm Shift

Hypothesis Confirmation

e Paradigm Shift

TO

Hypothesis Generation

NIH Strategic Plan for Data Science

(2018)

VISION: A modernized, integrated, and F.A.I.R. biomedical data ecosystem

NIH STRATEGIC PLAN FOR DATA SCIENCE

Introduction

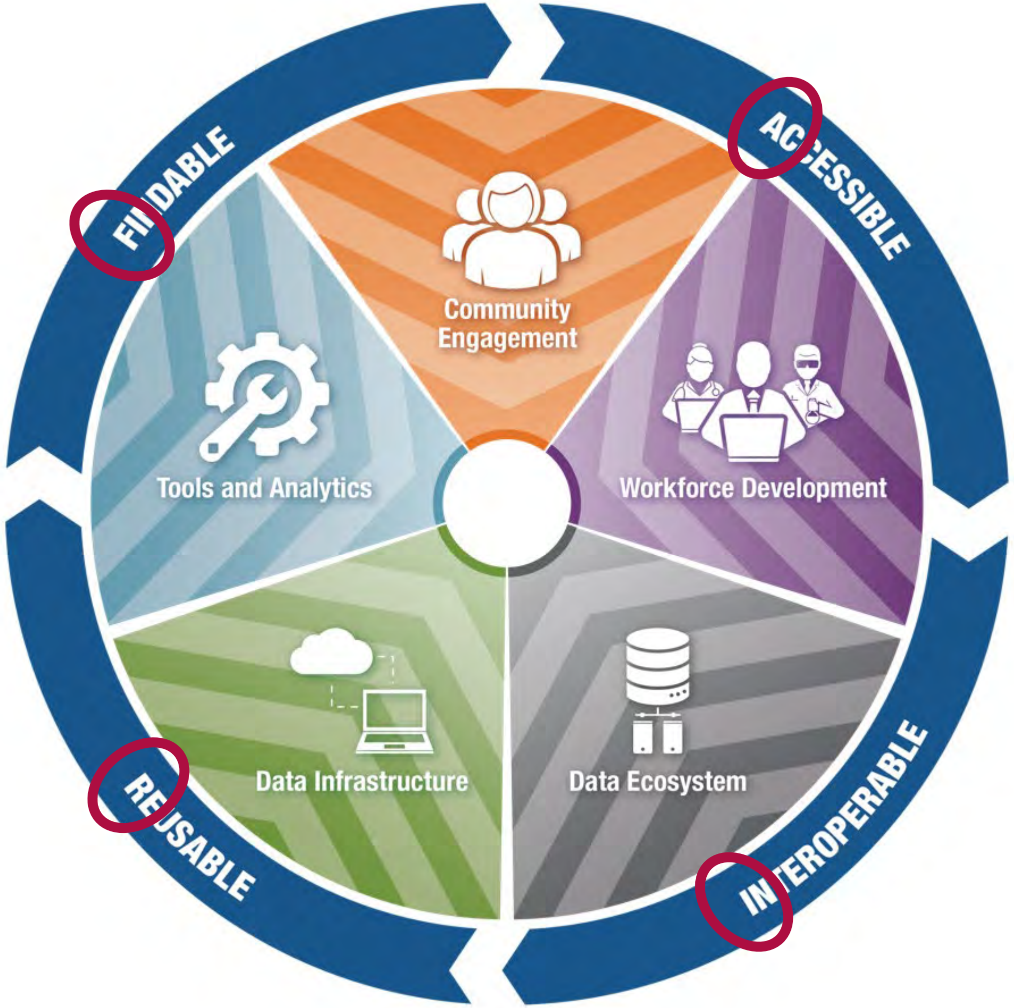
As articulated in the National Institutes of Health (NIH)-Wide Strategic Plan[1] and the Department of Health and Human Services (HHS) Strategic Plan,[2] our nation and the world stand at a unique moment of opportunity in biomedical research, and data science is an integral contributor. Understanding basic biological mechanisms through NIH-funded research depends upon vast amounts of data and has propelled biomedicine into the sphere of "Big Data" along with other sectors of the national and global economies. Reflecting today's highly integrated biomedical research landscape, NIH defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

NIH supports the generation and analysis of substantial quantities of biomedical research data (see, for example, text box "Big Data from the Resolution Revolution[3]"), including numerous quantitative and qualitative datasets emanating from fundamental research using model organisms (such as mice, fruit flies, and zebrafish), clinical studies (including medical images), and observational and epidemiological studies (including data from electronic health records and wearable devices). Metadata, "data about data," provides

Big Data from the Resolution Revolution

One of the revolutionary advances in microscope, detectors, and algorithms, cryogenic electron microscopy (cryoEM) has become one of the areas of

Data Ecosystem Approach



https://datascience.nih.gov/

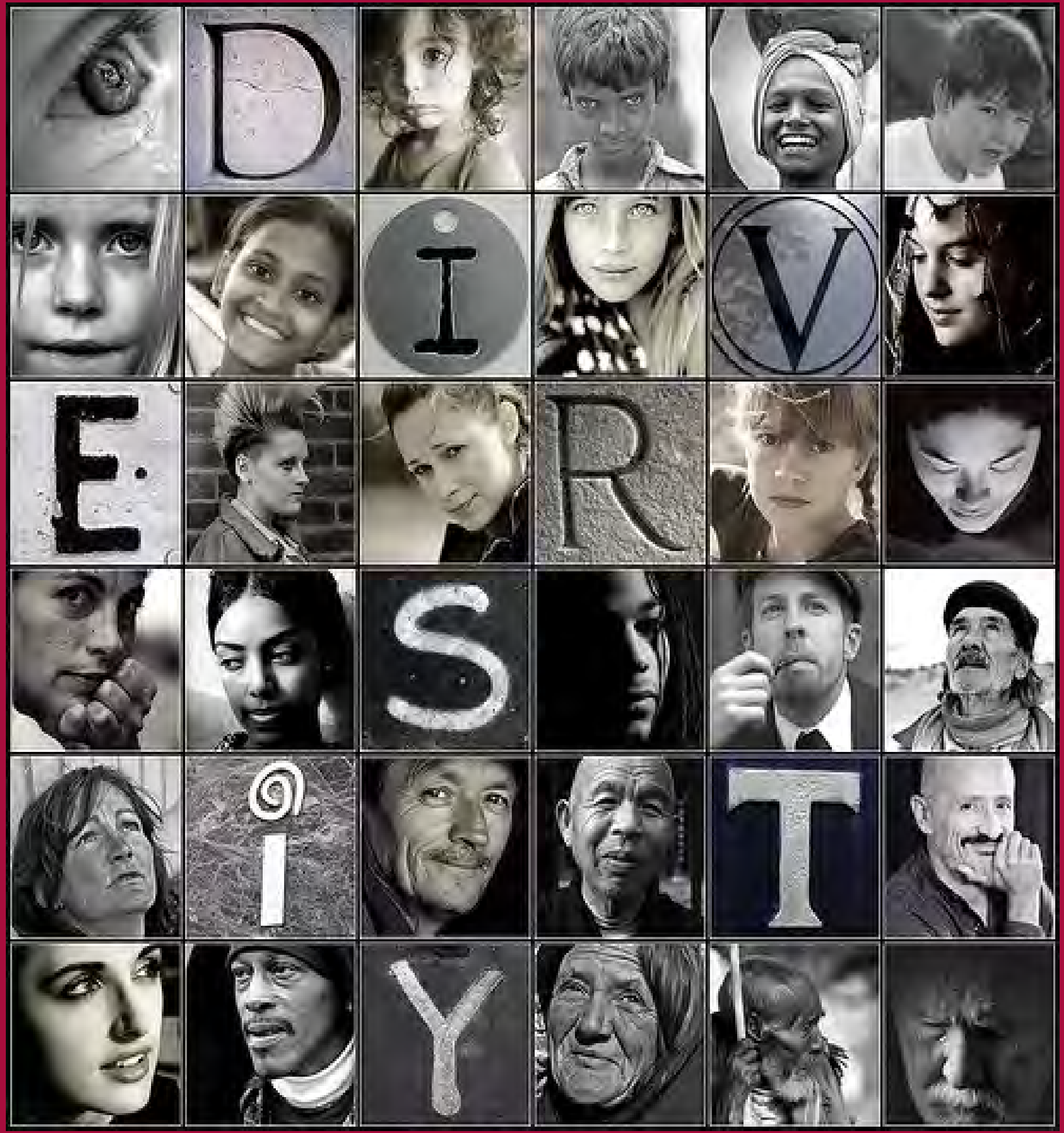
Ethical, Economic, Legal, Social Implications (EELSI)

- Nature vs Nurture
- Privacy and Confidentiality
- Research and other People Protections
- Informed consent
- Risk Assessment/Decision-making
- Benefits and Harms
- Predictive/Prognostic Screening/Testing
- Intellectual Property
- Dual Uses *(Forensics/Surveillance)*

Equity and Disparity Issues

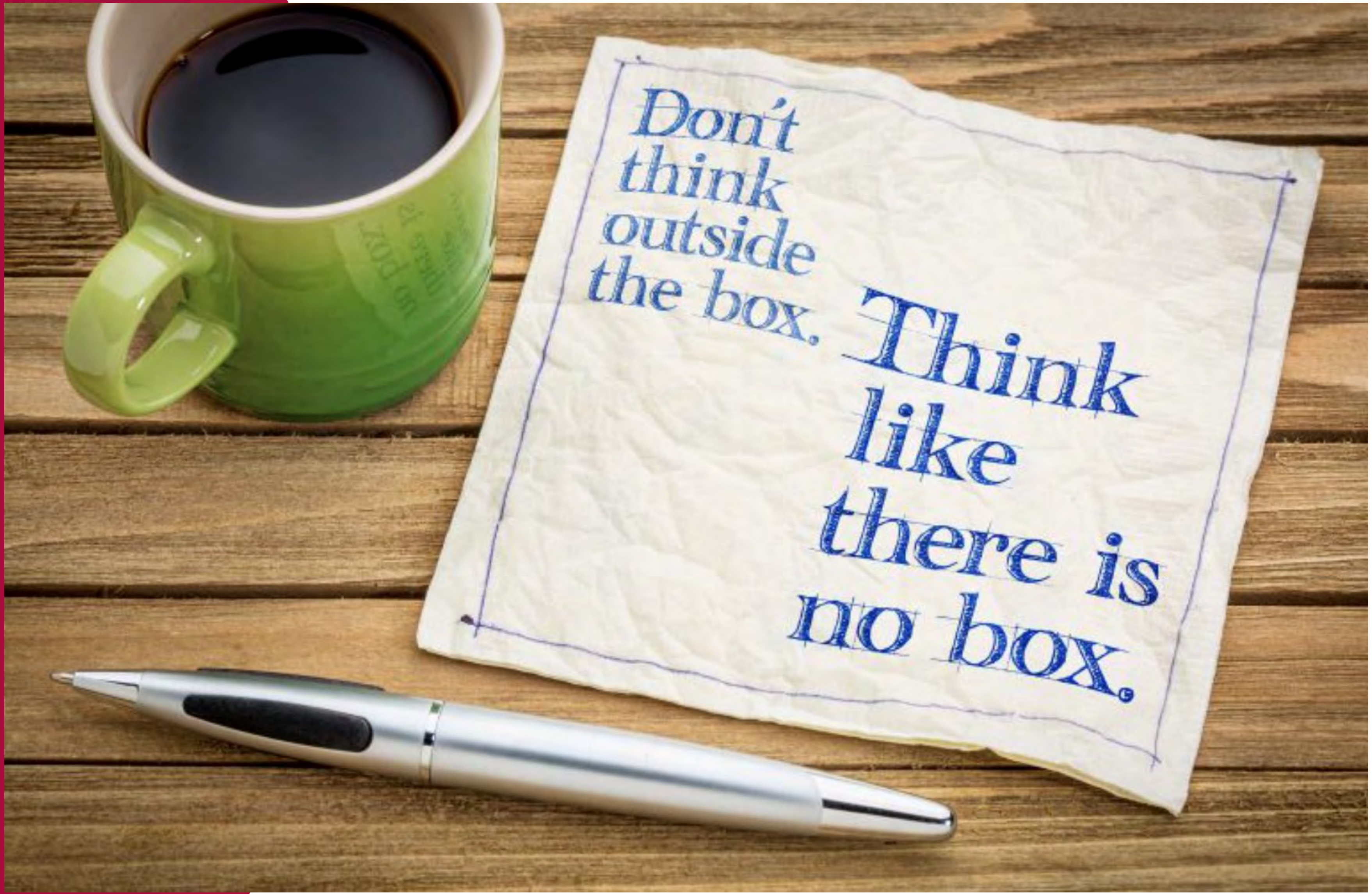
- Data and Information are not Neutral
- Stigma: *People/Groups/ Communities/Phenotypes*
- Inclusion: *Basic/Applied/Clinical Trial Research*
- Diversity and Workforce *Issues*
- Citizen Science *and Community Engagement*
- Inclusion, Equity, *and The Haves and Have Nots*

Data



If not designed
to address equity,
research (and data)
will perpetuate
disparities
and injustices

- me


Don't
think
outside
the box.
Think
like
there is
no box.