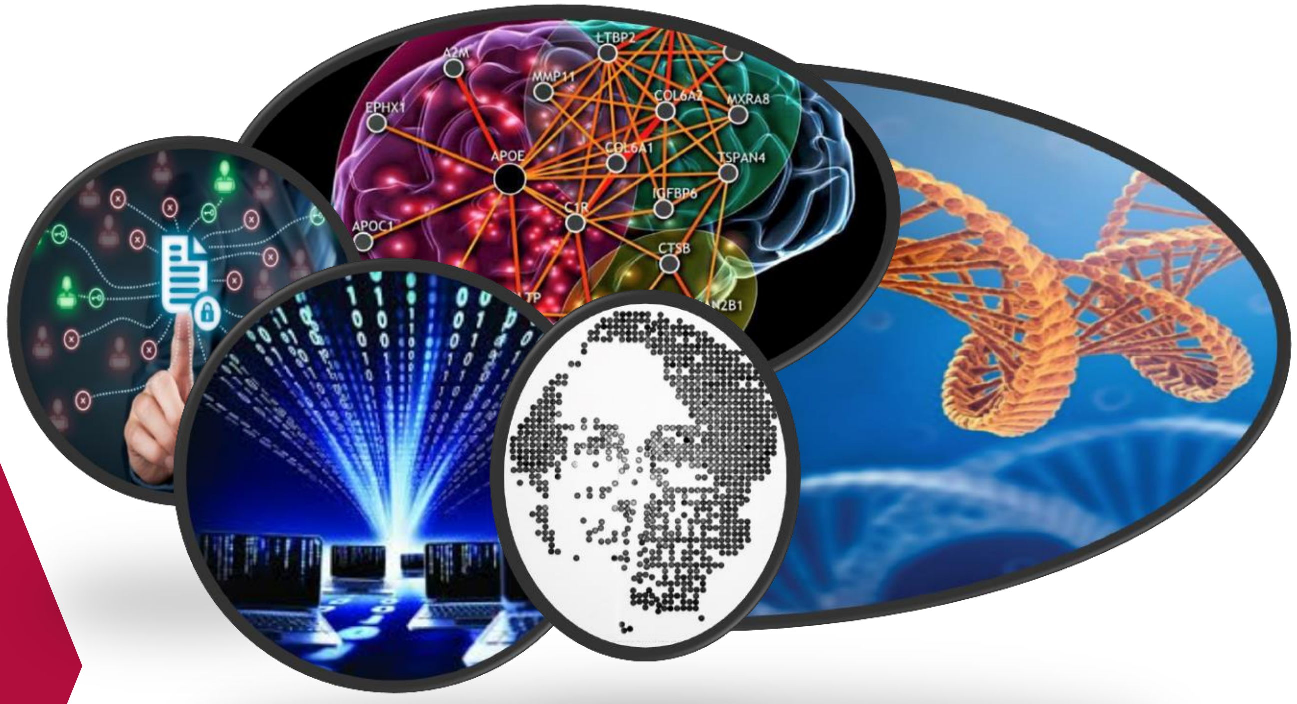
YOU and Big Data in Data Wonderland: It's NOT a Data Jabberwocky



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

Deputy Director, Office of Data Sharing
Center for Biomedical Informatics & Information Technology (CBIIT)
National Cancer Institute-NIH

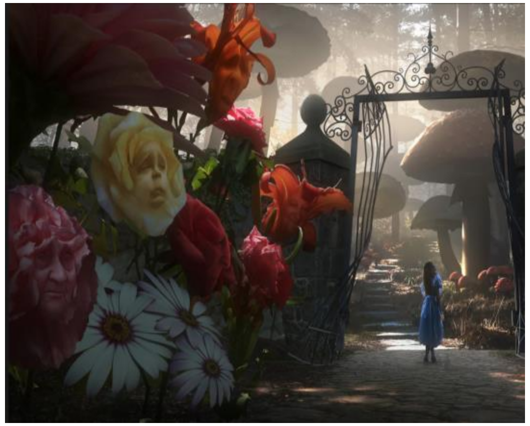
Professional Development Workshop and Mock Review
03 June 2019

NIH NATIONAL CANCER INSTITUTE

Agenda

Data Wonderland

Big Data
Human Rights
Open Science



Jabberwocky

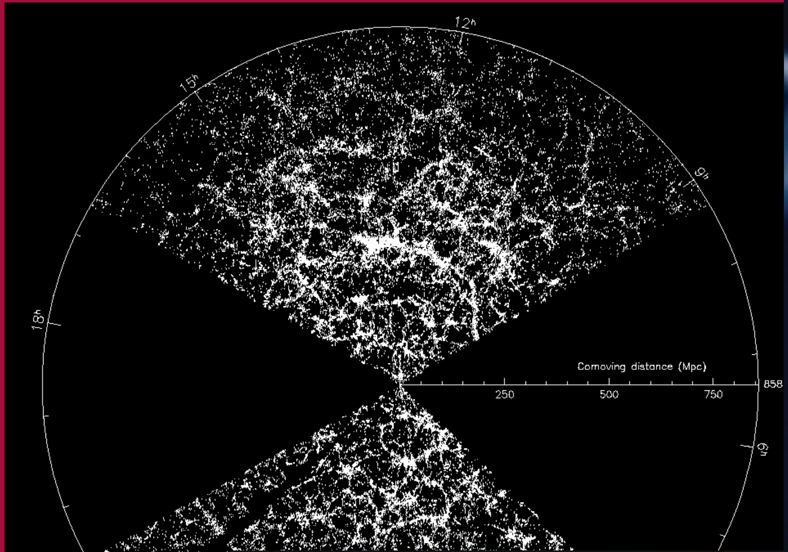
The Policies
Office of Data Sharing

Science

The Data
The People
The Ethics

Data Wonderland

Big Data
Open Science
Human Rights



Exploring the performance of transgender athletes p. 322

A threat to innovation in biomedicine p. 336

How social amoebae manage a microbiome p. 402

Science

$15
27 JULY 2018
sciencemag.org

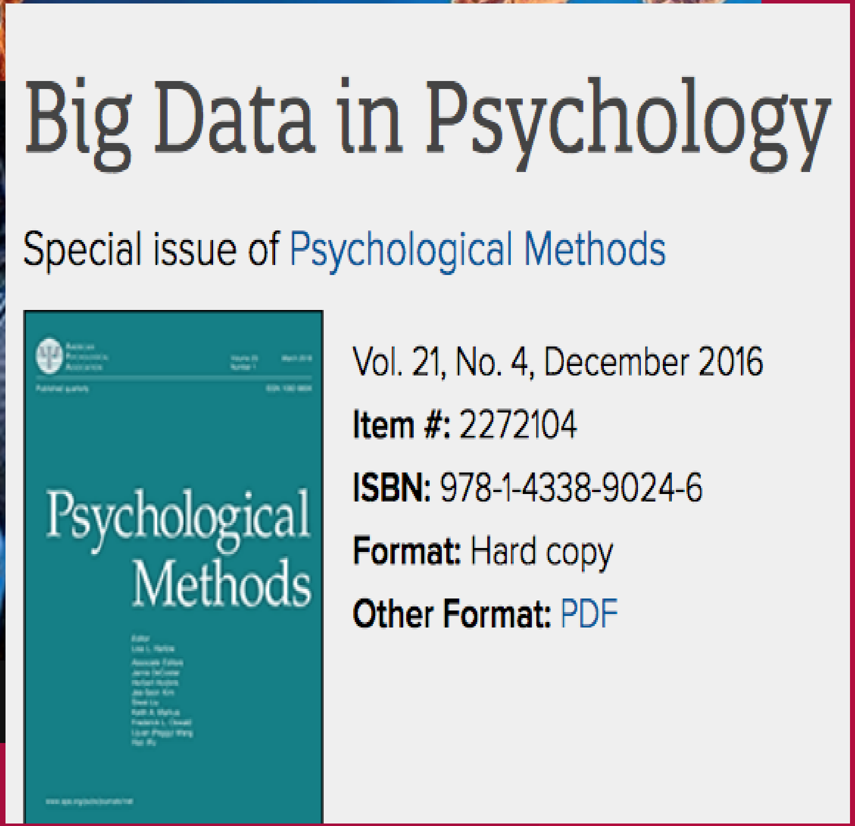
AAAS

3

SPECIAL ISSUE

FRONTIERS IN COMPUTATION

VOLUME 361, ISSUE 6400
26th July 2018



Big Data in Psychology

Special issue of Psychological Methods

Psychological Methods

Vol. 21, No. 4, December 2016

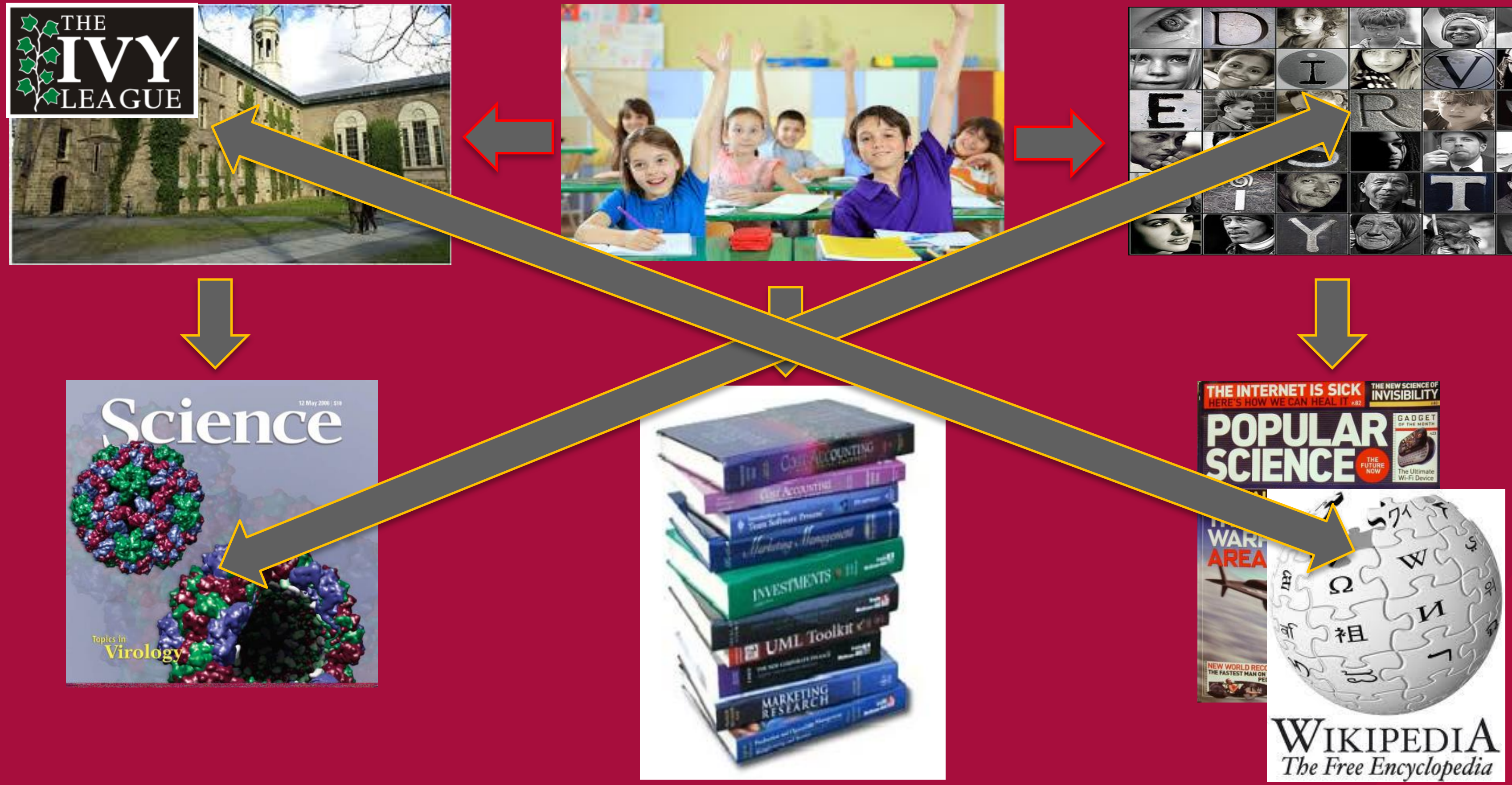
Item #: 2272104

ISBN: 978-1-4338-9024-6

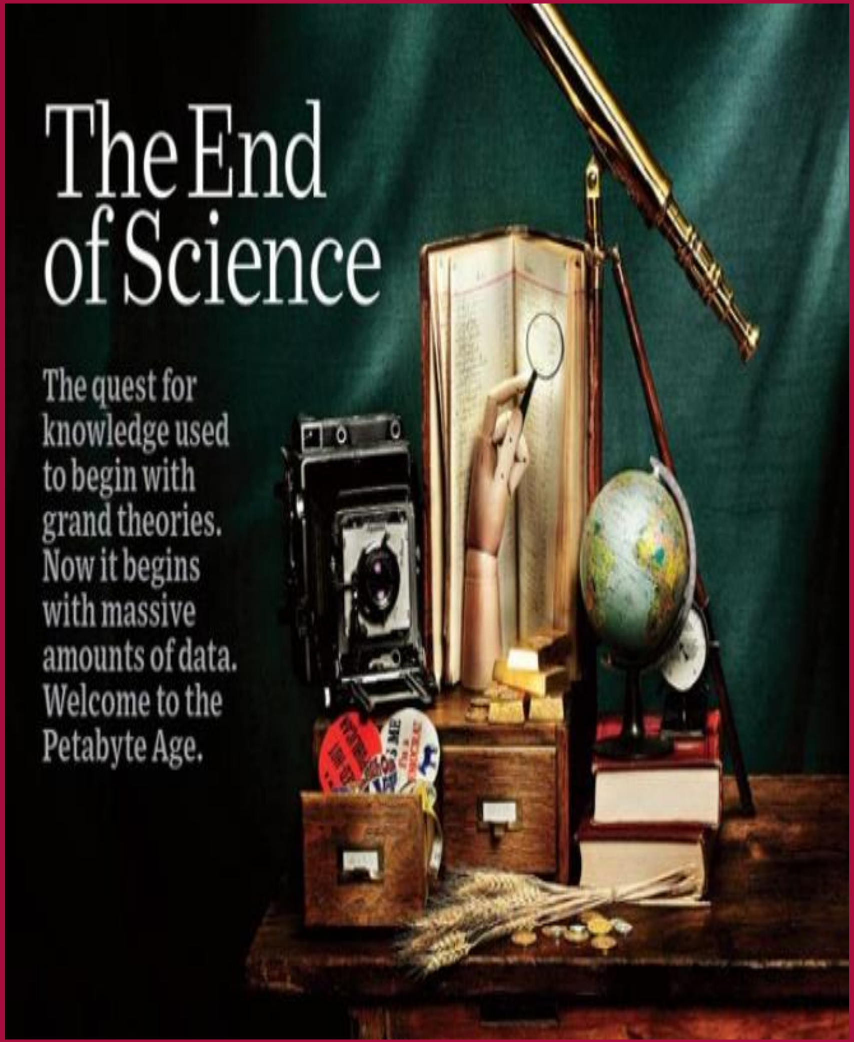
Format: Hard copy

Other Format: PDF

Big Data: Human Rights and the Democratization of Knowledge



Big Data

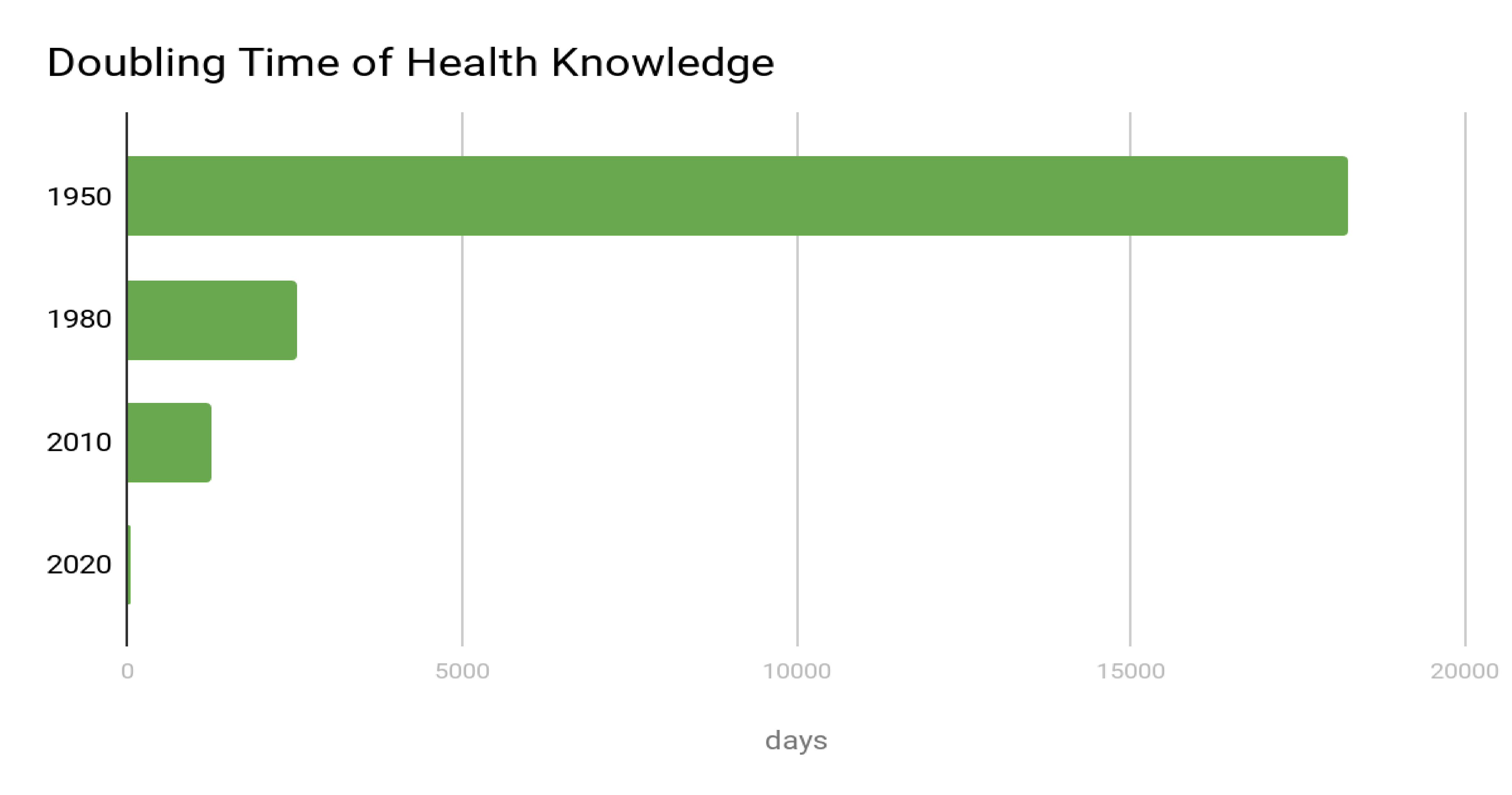


The End of Science

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.

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

Big Data


Doubling Time of Health Knowledge
1950
1980
2010
2020
0
5000
10000
15000
20000
days

Open Science

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

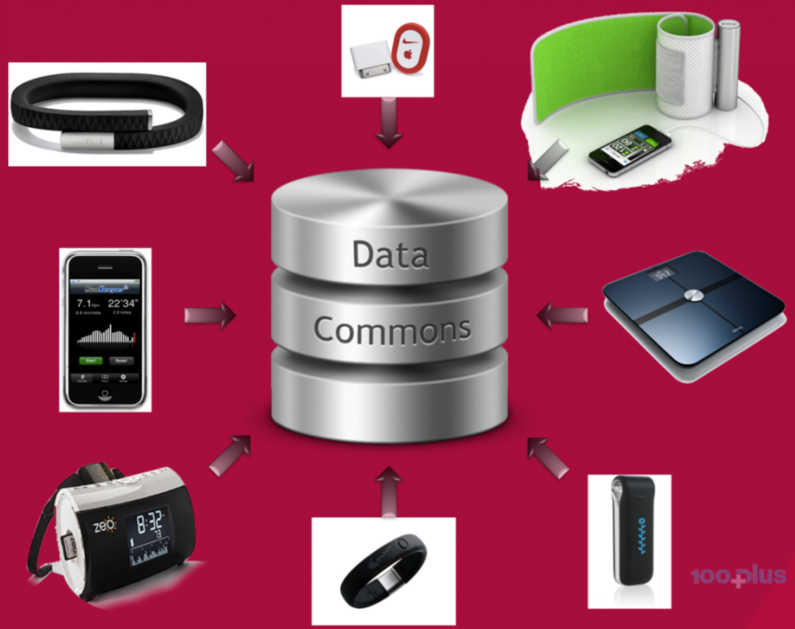
DATA SHARING AND INNOVATION

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

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.

Open Science and Data Sharing

Facilitates innovation of research tools and methods



NATIONAL CANCER INSTITUTE

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

Open Science and Data Sharing

Facilitates innovation of research tools and methods

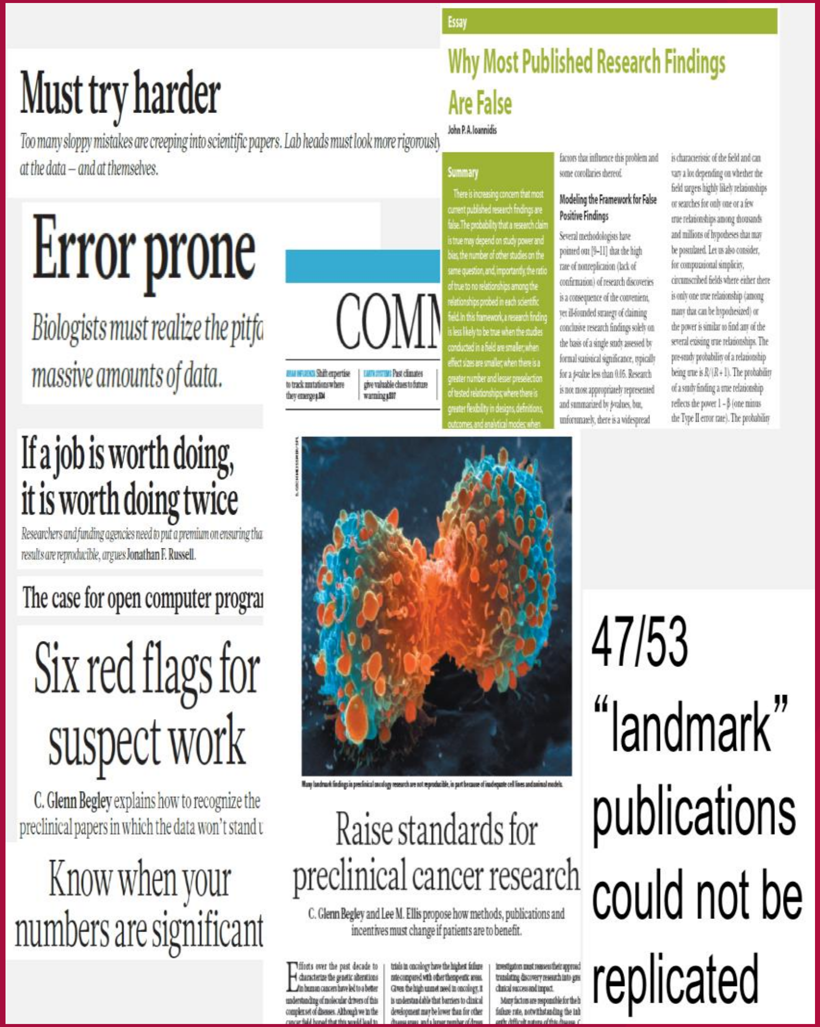
- Increases statistical power

NIH NATIONAL CANCER INSTITUTE

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

Open Science and Data Sharing

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



Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data – and at themselves.

Error prone

Biologists must realize the pitfa massive amounts of data.

COMM

Shift expertise to track mutations where they emerge p.314 | Past climates give valuable clues to future warming p.317

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring tha results are reproducible, argues Jonathan F. Russell.

The case for open computer progra

Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand u

Know when your numbers are significant

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs

investigators must reassess their approa translating discovery research into gre clinical success and impact.

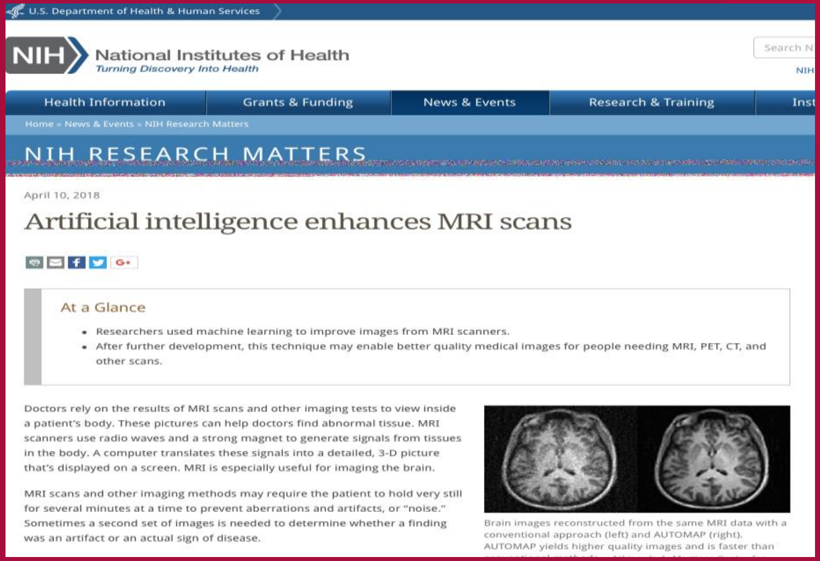
Many factors are responsible for the h failure rate, notwithstanding the inh ently difficult nature of this disease. C

47/53 "landmark" publications could not be replicated

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

Open Science and Data Sharing

Increases scientific value and analyses by enabling data from multiple studies to be combined and explored


U.S. Department of Health & Human Services
NIH National Institutes of Health
Turning Discovery Into Health
Search N
NIH
Health Information
Grants & Funding
News & Events
Research & Training
Inst
Home » News & Events » NIH Research Matters

NIH RESEARCH MATTERS

April 10, 2018

Artificial intelligence enhances MRI scans

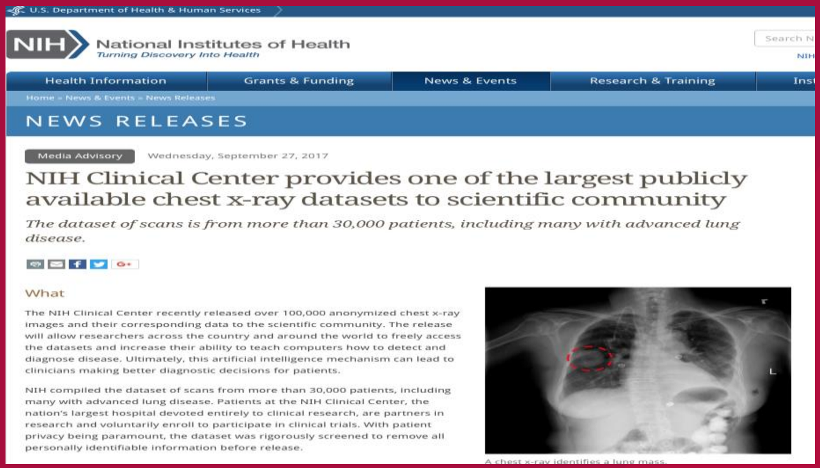
At a Glance

- Researchers used machine learning to improve images from MRI scanners.
- After further development, this technique may enable better quality medical images for people needing MRI, PET, CT, and other scans.

Doctors rely on the results of MRI scans and other imaging tests to view inside a patient's body. These pictures can help doctors find abnormal tissue. MRI scanners use radio waves and a strong magnet to generate signals from tissues in the body. A computer translates these signals into a detailed, 3-D picture that's displayed on a screen. MRI is especially useful for imaging the brain.

MRI scans and other imaging methods may require the patient to hold very still for several minutes at a time to prevent aberrations and artifacts, or "noise." Sometimes a second set of images is needed to determine whether a finding was an artifact or an actual sign of disease.

Brain images reconstructed from the same MRI data with a conventional approach (left) and AUTOMAP (right). AUTOMAP yields higher quality images and is faster than


U.S. Department of Health & Human Services
NIH National Institutes of Health
Turning Discovery Into Health
Search N
NIH
Health Information
Grants & Funding
News & Events
Research & Training
Inst
Home » News & Events » News Releases

NEWS RELEASES

Media Advisory Wednesday, September 27, 2017

NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community

The dataset of scans is from more than 30,000 patients, including many with advanced lung disease.

What

The NIH Clinical Center recently released over 100,000 anonymized chest x-ray images and their corresponding data to the scientific community. The release will allow researchers across the country and around the world to freely access the datasets and increase their ability to teach computers how to detect and diagnose disease. Ultimately, this artificial intelligence mechanism can lead to clinicians making better diagnostic decisions for patients.

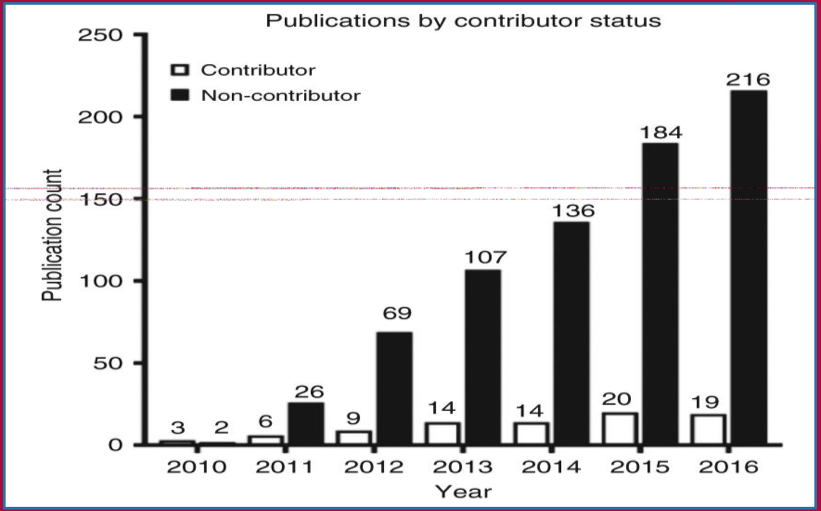
NIH compiled the dataset of scans from more than 30,000 patients, including many with advanced lung disease. Patients at the NIH Clinical Center, the nation's largest hospital devoted entirely to clinical research, are partners in research and voluntarily enroll to participate in clinical trials. With patient privacy being paramount, the dataset was rigorously screened to remove all personally identifiable information before release.

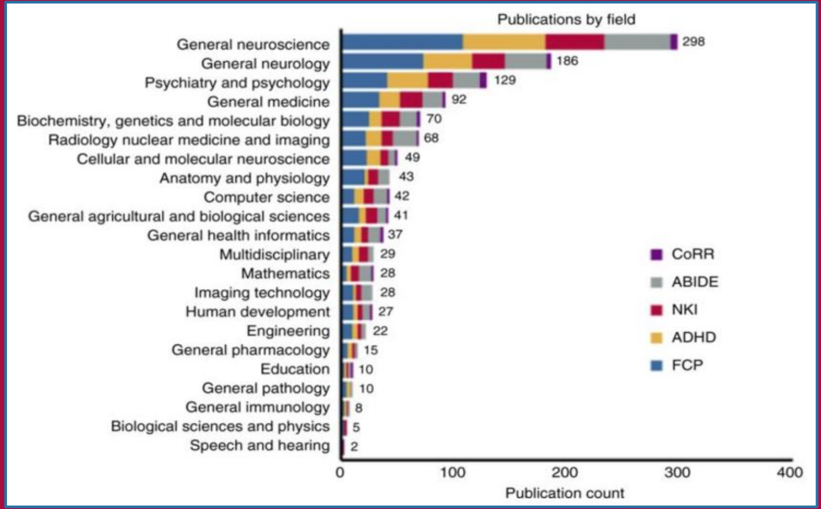
A chest x-ray identifies a lung mass.

Open Science and Data Sharing

- Increases scientific value and analyses by enabling data from multiple studies to be combined and explored

- 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*

Open Science and Data Sharing

The Economist

JANUARY 14TH–20TH 2017

Trump v the spooks

The stain of Guantánamo

Pop stars and patronage in Congo

Inflation's welcome return

Lifelong learning

How to survive in the age of automation

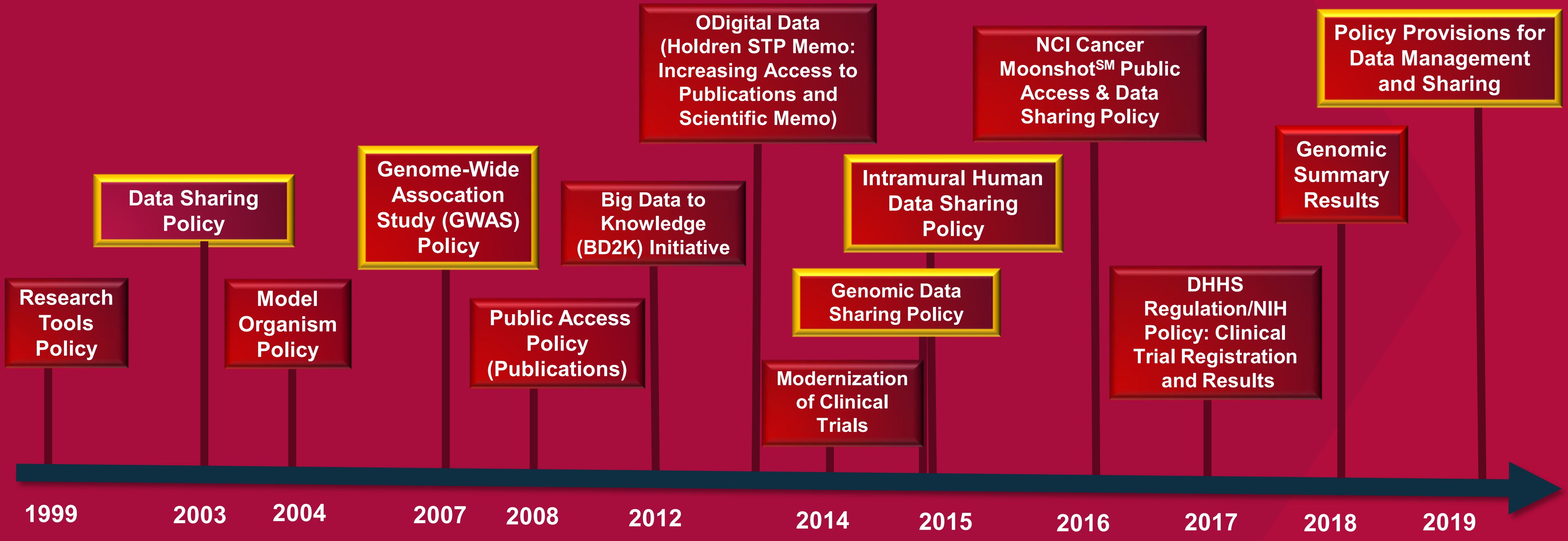
A SPECIAL REPORT

- Biology and Medicine are now *data intensive enterprises*
- *Rapidly changing* scale
- Technology, *data computing* and *information technology (IT)* are *pervasive* in the *lab, clinic, and home*

The Jabberwocky

Policies
NCI Office of Data Sharing

NIH Data Sharing Policies



Research Tools Policy

Data Sharing Policy

Model Organism Policy

Genome-Wide Assocation Study (GWAS) Policy

Public Access Policy (Publications)

Big Data to Knowledge (BD2K) Initiative

ODigital Data (Holdren STP Memo: Increasing Access to Publications and Scientific Memo)

Modernization of Clinical Trials

Genomic Data Sharing Policy

Intramural Human Data Sharing Policy

NCI Cancer Moonshot^{SM} Public Access & Data Sharing Policy

DHHS Regulation/NIH Policy: Clinical Trial Registration and Results

Genomic Summary Results

Policy Provisions for Data Management and Sharing

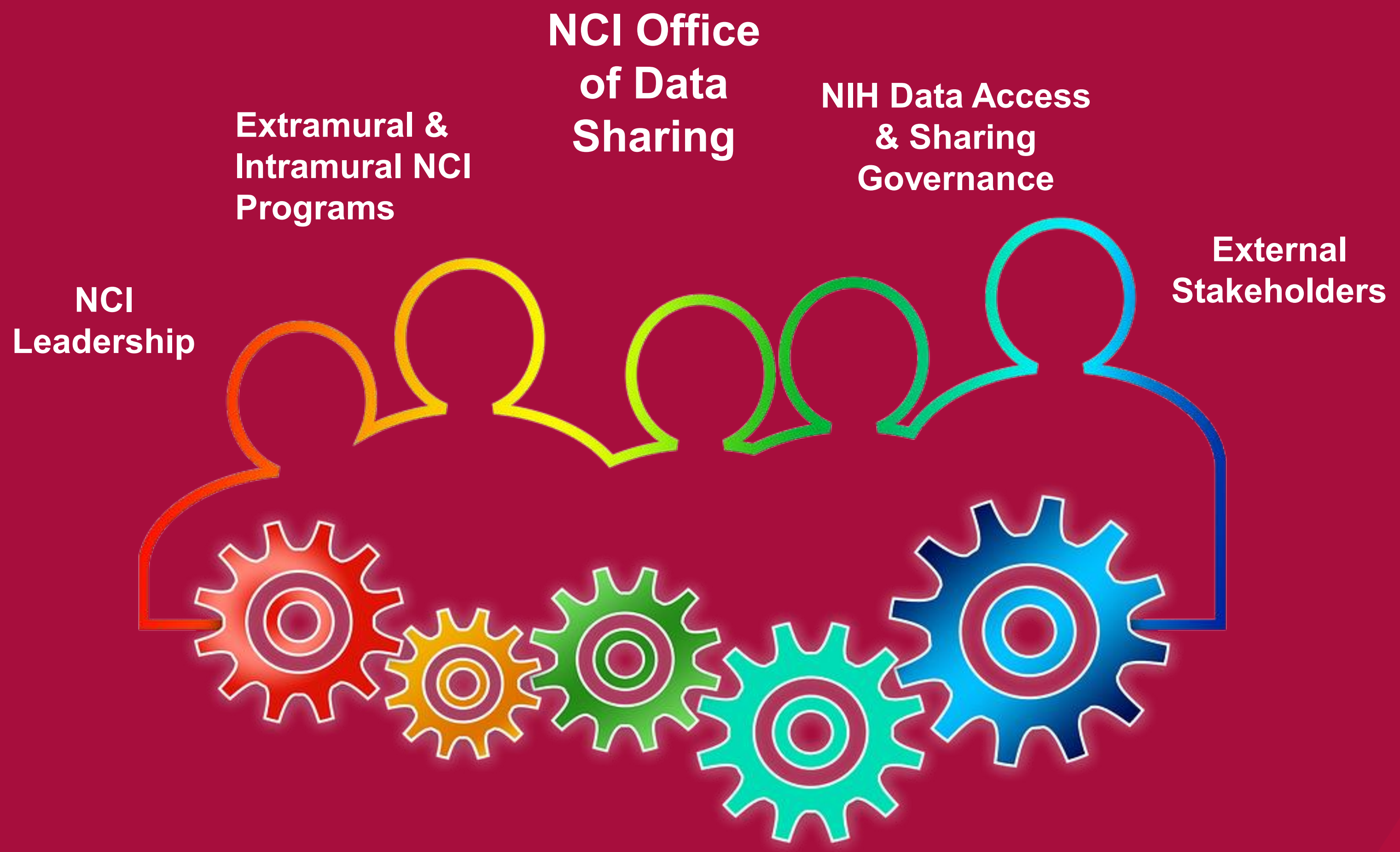
1999 2003 2004 2007 2008 2012 2014 2015 2016 2017 2018 2019

NIH NATIONAL CANCER INSTITUTE

(Courtesy: (Adapted from D.Paltoo and L.L. Rodriguez)

NCI Office of Data Sharing

nciofficeofdatasharing@nih.gov


NCI Office of Data Sharing
Extramural & Intramural NCI Programs
NIH Data Access & Sharing Governance
NCI Leadership
External Stakeholders

(Courtesy: (Adapted from J Guidry Auvil, NCI – NIH)

NCI Office of Data Sharing

nciofficeofdatasharing@nih.gov

Provide *leadership and guidance* to enhance data sharing for NCI and the cancer research community.

Coordinate registration, submission, and access procedures for NCI datasets/repositories.

Encourage *participation* in major data sharing initiatives.

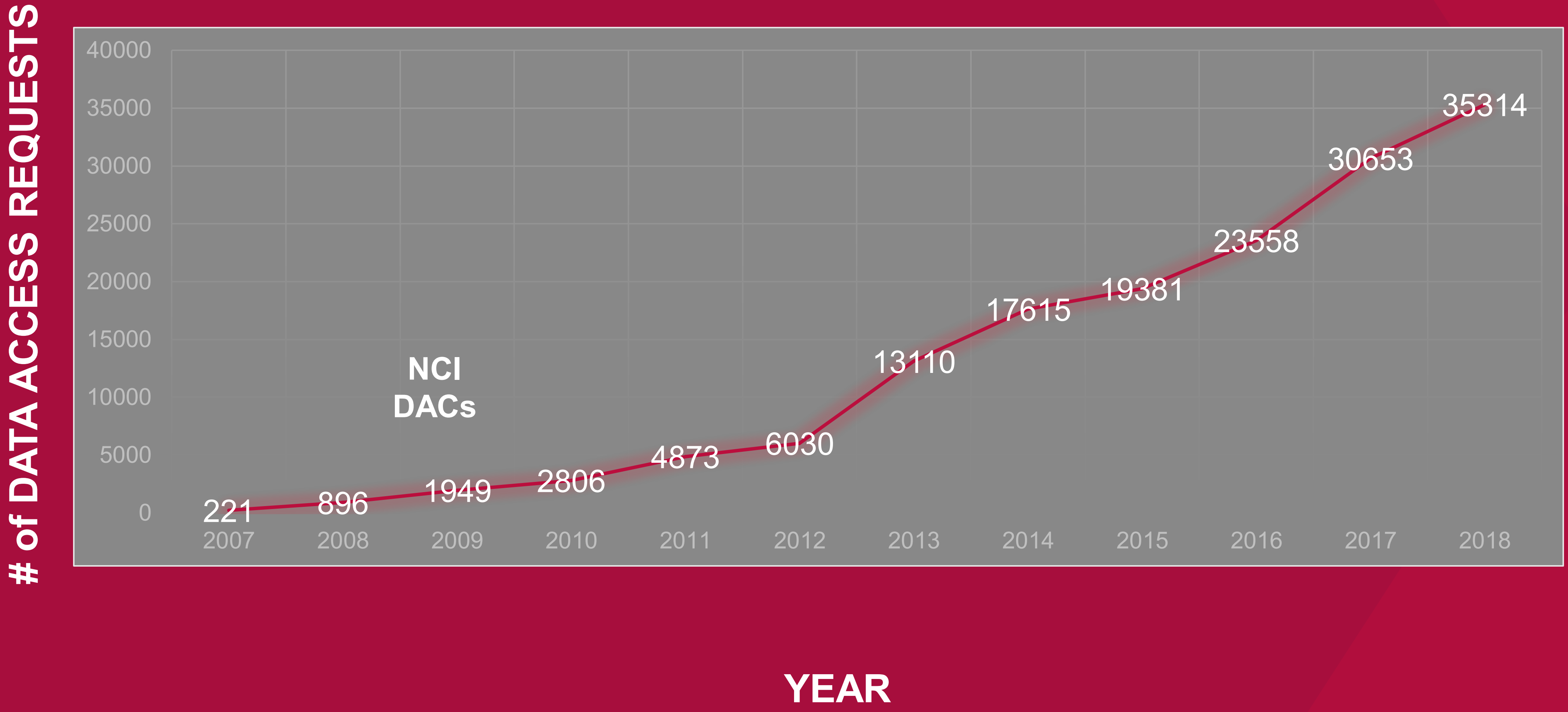
Guide NCI approach to *implementation and interpretation* of NIH and NCI data management and sharing policies.

Advise on considerations for *ethical and minority and health disparity issues* related to data access and sharing for the cancer community.

Create data sharing resources to inform and guide the cancer communities.

(Courtesy: (Adapted from J Guidry Auvil, NCI – NIH)

Total Number of Actual and Projected NCI dbGaP Data Access Requests By Year

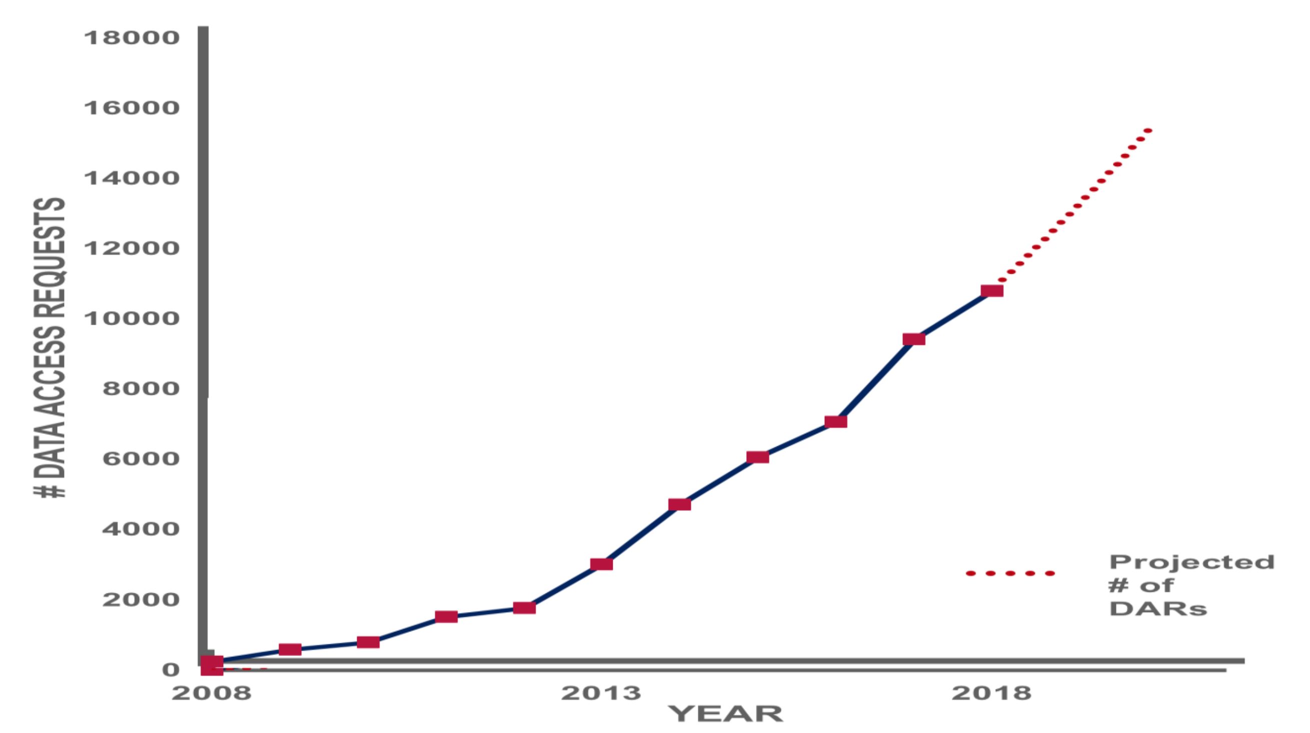


NIH NATIONAL CANCER INSTITUTE

(Courtesy: Adapted from J Guidry Auvil, NCI – NIH)

Total Number of Actual and Projected NCI dbGaP Data Access Requests By Year

# of DATA ACCESS REQUESTS

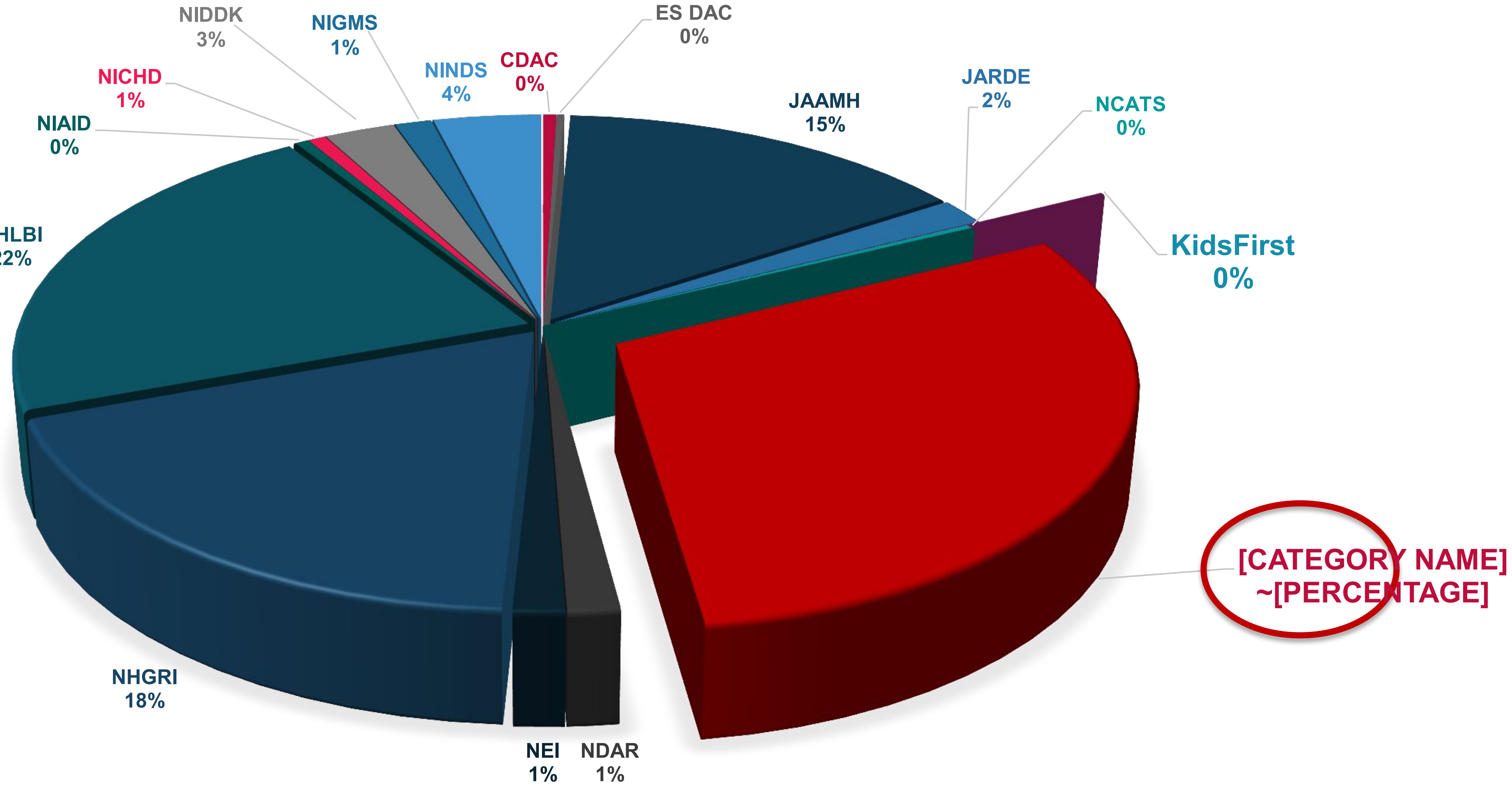


YEAR

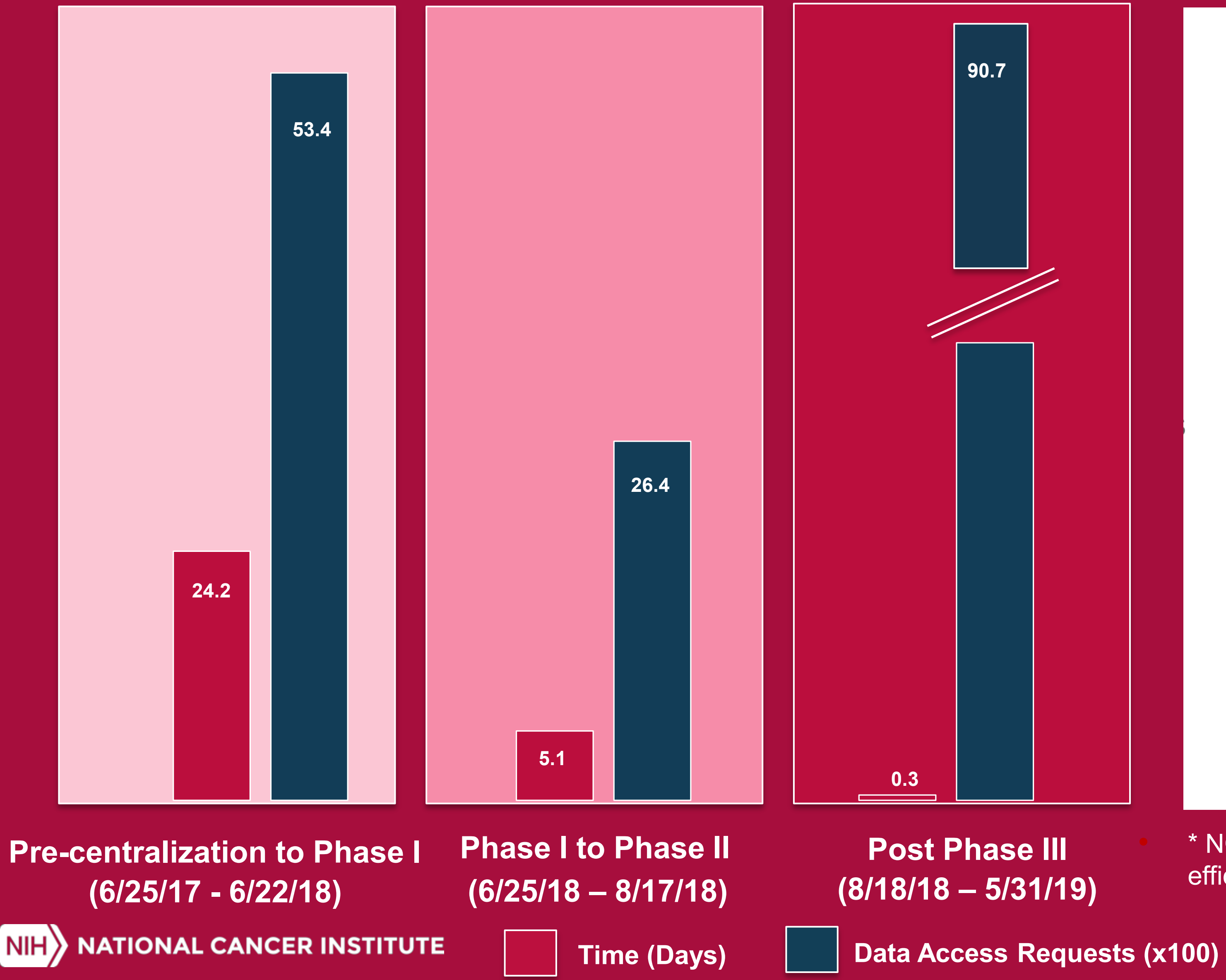
NIH NATIONAL CANCER INSTITUTE

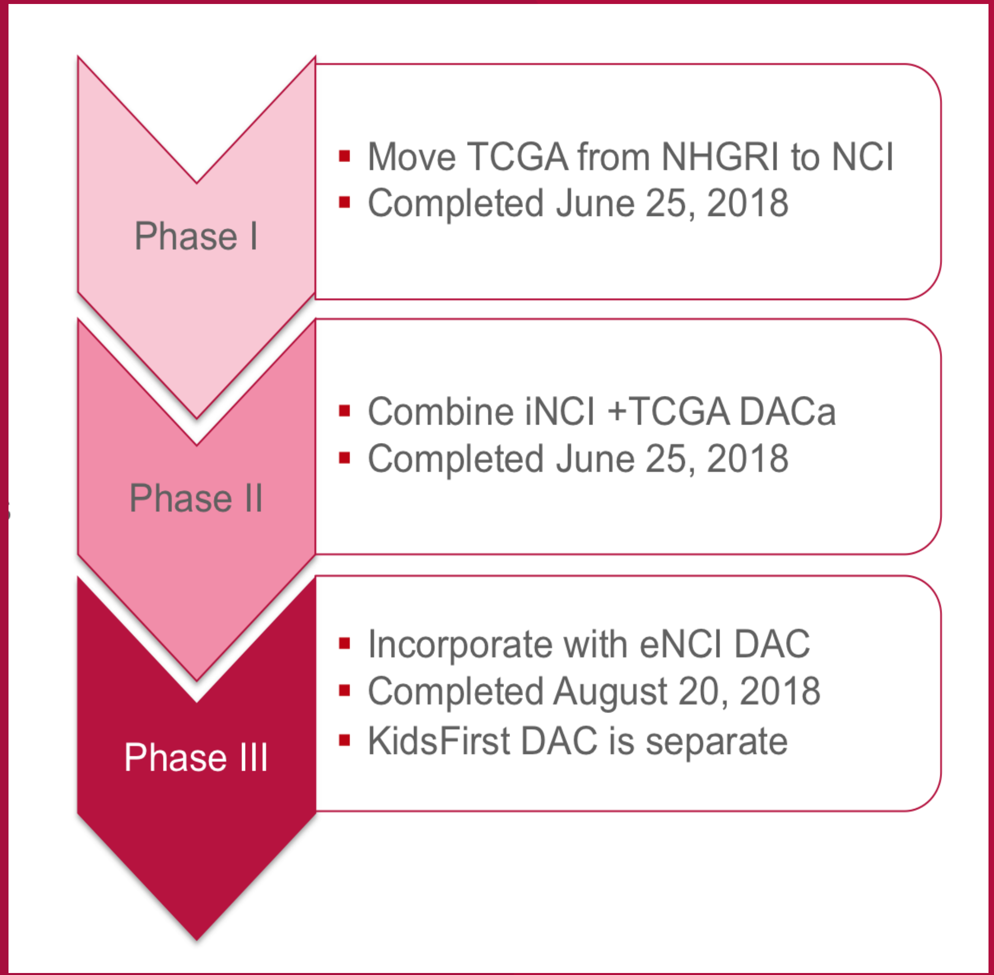
(Courtesy: Adapted from J Guidry Auvil, NCI – NIH)

Percentage of Data Access Requests By NIH I/Cs 01/01218 –05/31/2019



NCI Office of Data Sharing*





Phase I

- Move TCGA from NHGRI to NCI
- Completed June 25, 2018

Phase II

- Combine iNCI +TCGA DACa
- Completed June 25, 2018

Phase III

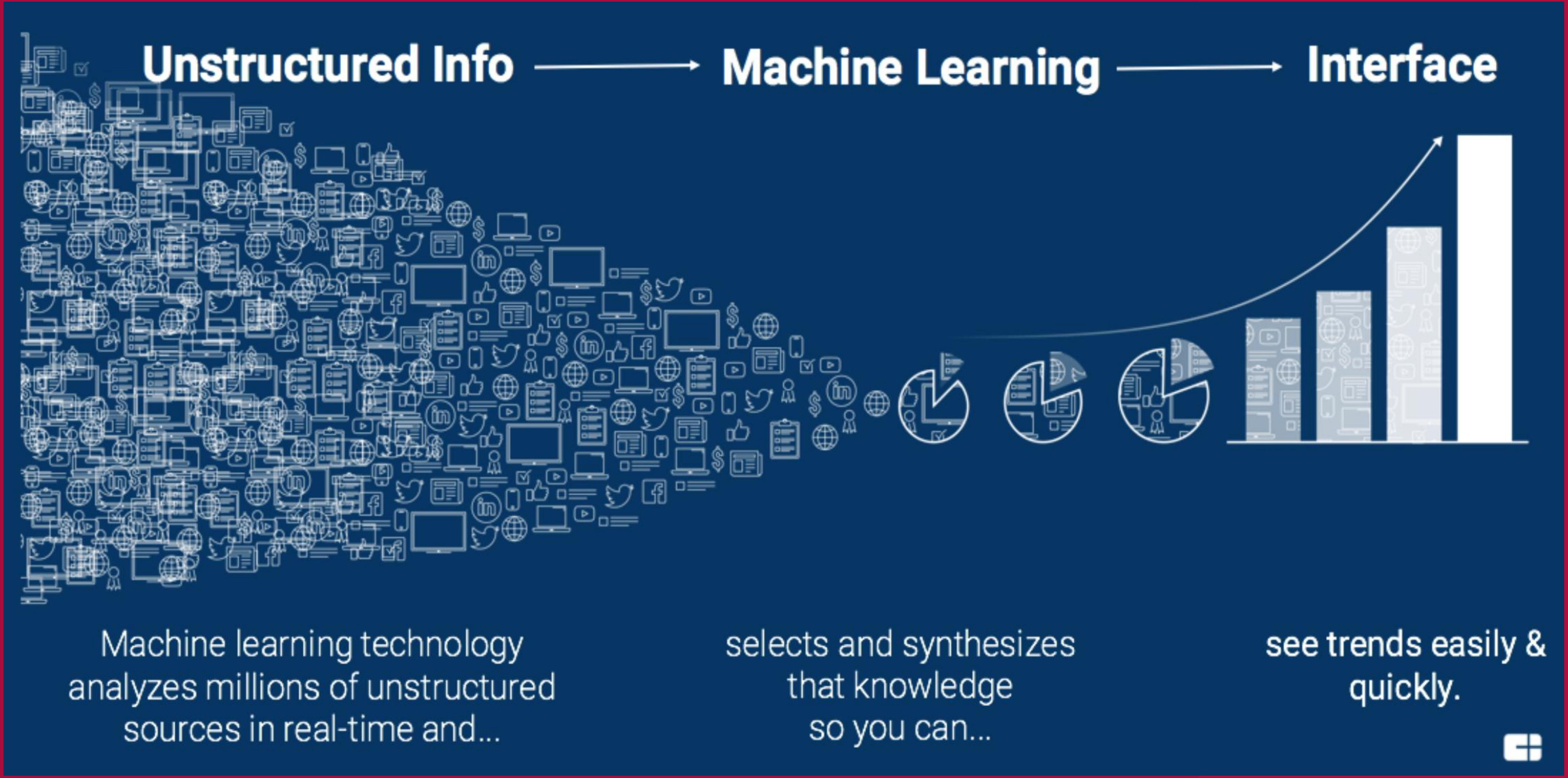
- Incorporate with eNCI DAC
- Completed August 20, 2018
- KidsFirst DAC is separate

* NCI DAC receives ~30% of DARs to dbGaP; results include significant efficiency in DAR processing times and eliminating 1000+ DAR backlog

he World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
TMI | JUL.2008
Do we need THEORY?
How do we find ARK MATTER?
Is Earth safe from ASTEROIDS?
Where will TERRORISTS strike next?
How can you win any LAWSUIT?
Who will be the next PRESIDENT?
How will we grow enough FOOD?
How can we protect BUSINESS from risk?
The End of Science
The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data. Welcome to the Petabyte Age.

The Science

The Data
The People
The Ethics



Unstructured Info → Machine Learning → Interface

Machine learning technology analyzes millions of unstructured sources in real-time and...

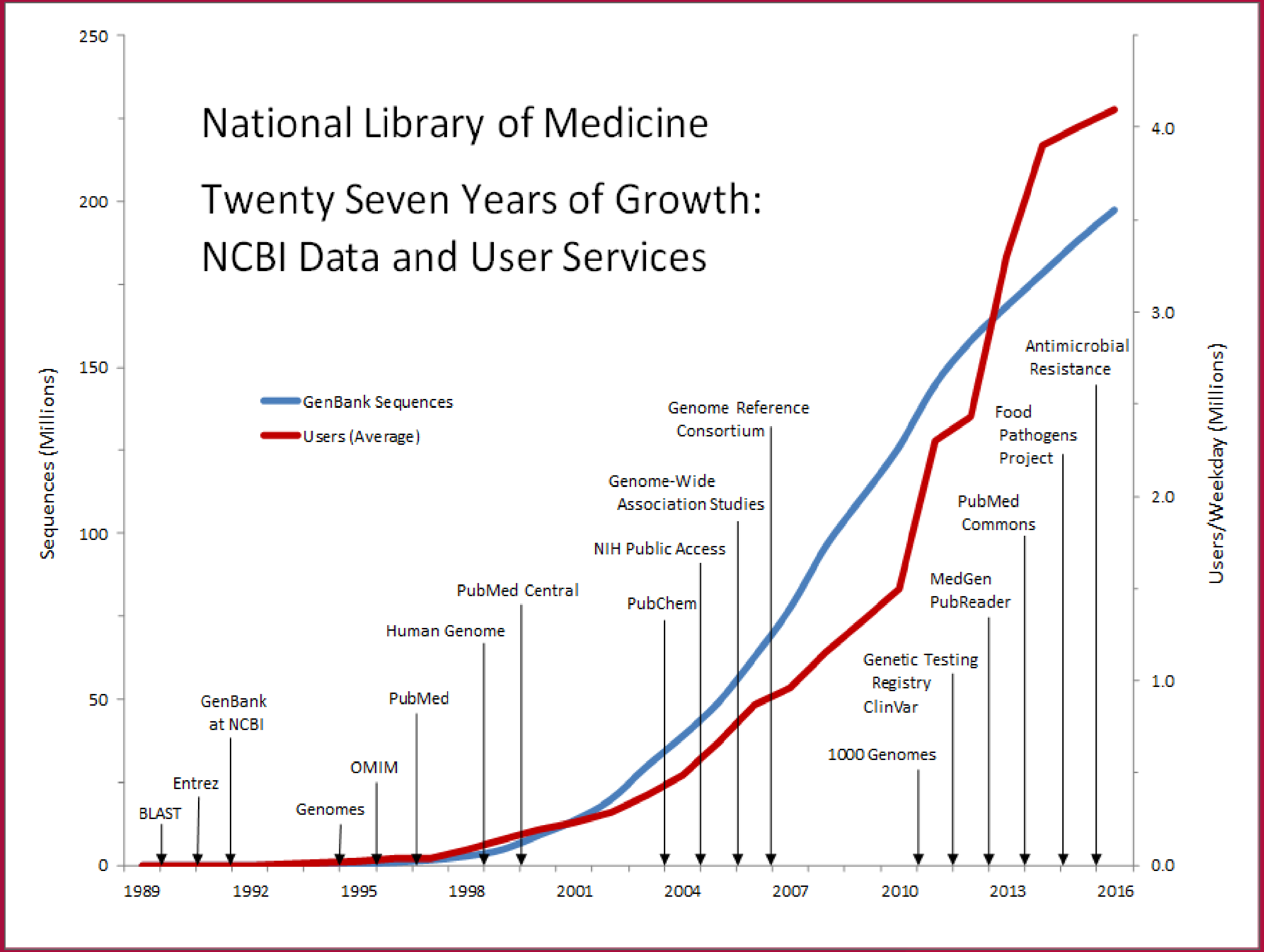
selects and synthesizes that knowledge so you can...

see trends easily & quickly.

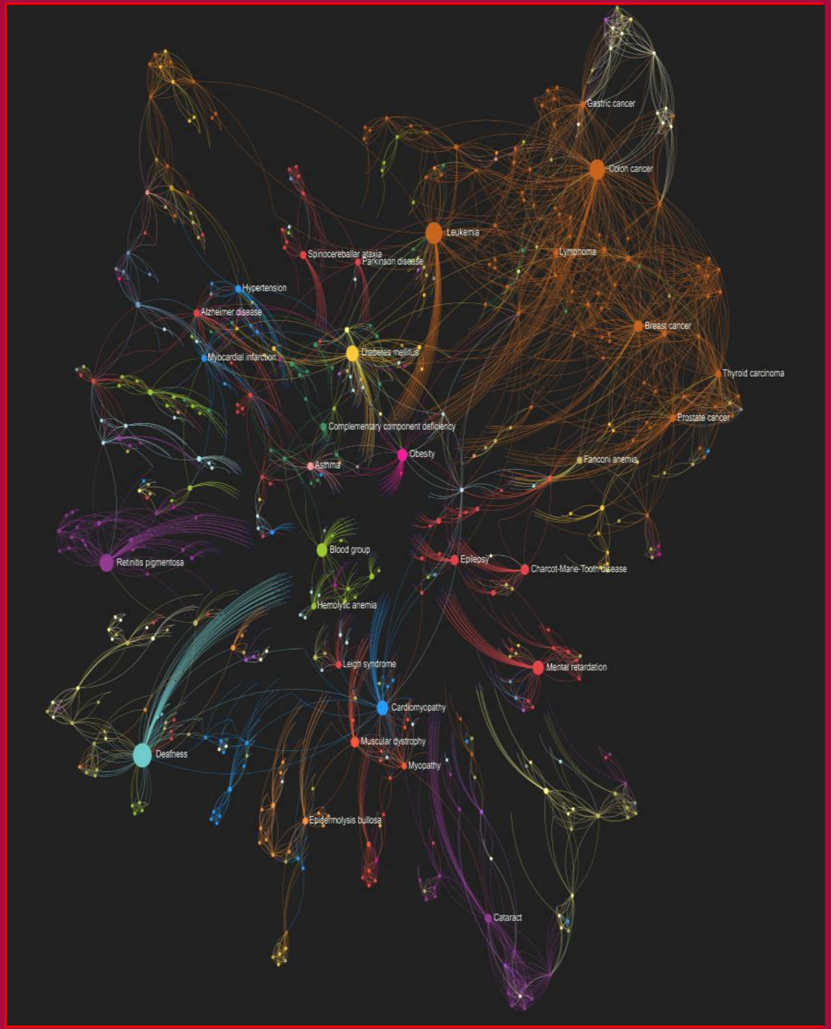
https://www.cbinsights.com/research/team-blog/data-network-effects/

Data: Variety, Volume, Velocity, and Veracity

- More scientific data domains are emerging with capacities to capture real time health information
  - Proteomics
  - Metabolomics
  - Microscopy
  - Medical imaging
  - Other various technologies



Human Disease Networks (2015)



- Identify disease gene-phenotype associations at higher cellular and organismal levels

Greene, CS, et al, (2015) *Nature Genetics*, 47, 569-576. https://doi.org/10.1038/ng.3259

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

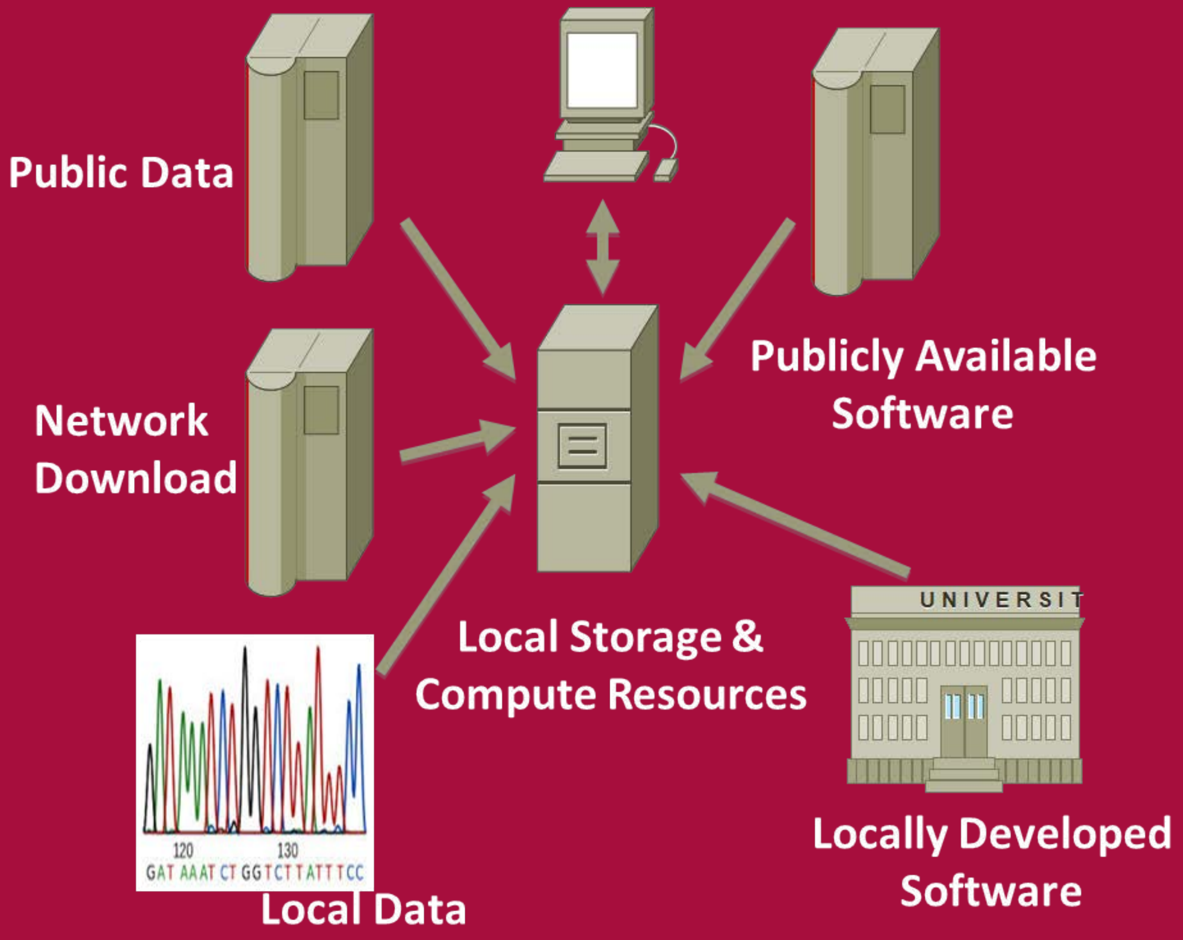
Courtesy of P. Kuhn (USC)

Precision Medicine

- Learning system that accounts for complexity of underlying biology
- Requires
  - Deep biological understanding
  - Advances in scientific methods
  - Advances in instrumentation
  - Advances in technology
  - Advances in data management and computation
- Can *change* disease classifications
- Genomic, imaging, clinical, and laboratory data

Standard Model of Computational Analyses

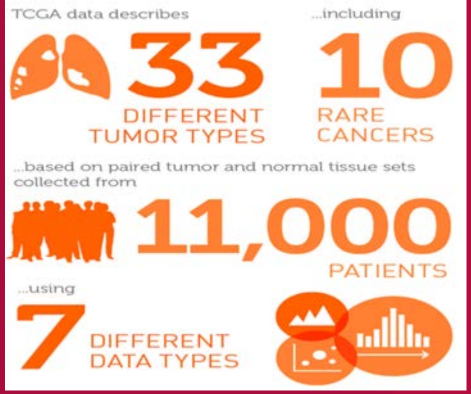
(circa 2014)


Public Data
Network Download
Local Data
120
130
GAT AAAT CT GGTCTTATTTCC
Local Storage & Compute Resources
Publicly Available Software
UNIVERSIT
Locally Developed Software

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

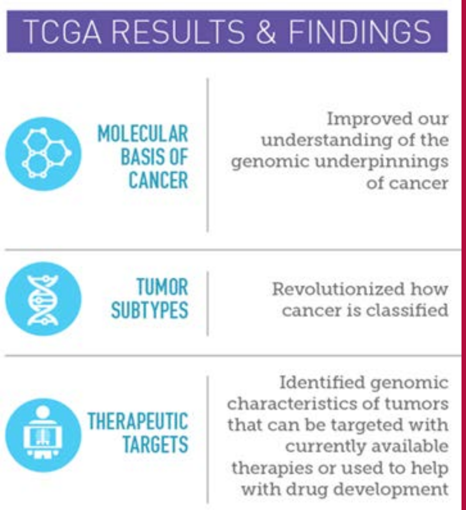
TCGA BY THE NUMBERS

TCGA produced over 2.5 PETABYTES of data


TCGA data describes 33 DIFFERENT TUMOR TYPES ...including 10 RARE CANCERS

...based on paired tumor and normal tissue sets collected from 11,000 PATIENTS

...using 7 DIFFERENT DATA TYPES


TCGA RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER — Improved our understanding of the genomic underpinnings of cancer

TUMOR SUBTYPES — Revolutionized how cancer is classified

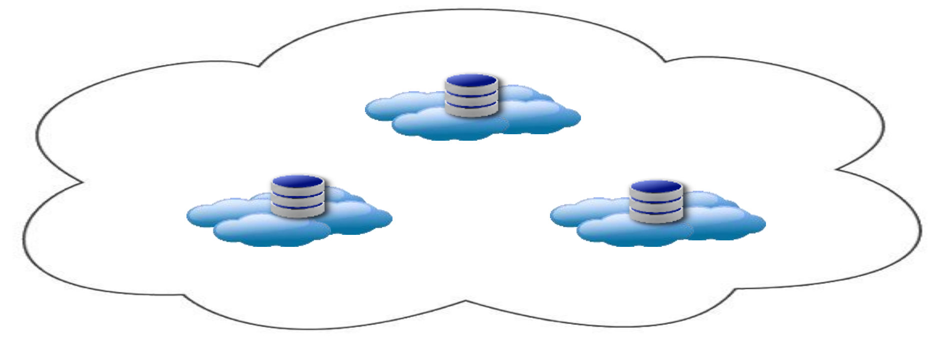
THERAPEUTIC TARGETS — Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

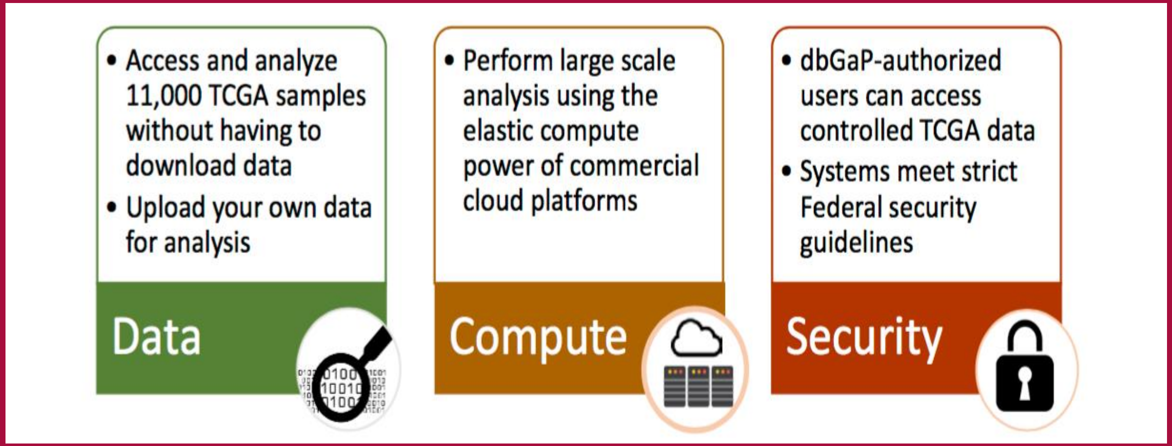
(Courtesy: Adapted from A. Kerlavage, NCI – NIH)

NCI Cloud Resources

- Democratize access to NCI-generated data

Create cost-effective scalable computational capacity





- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Data

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Compute

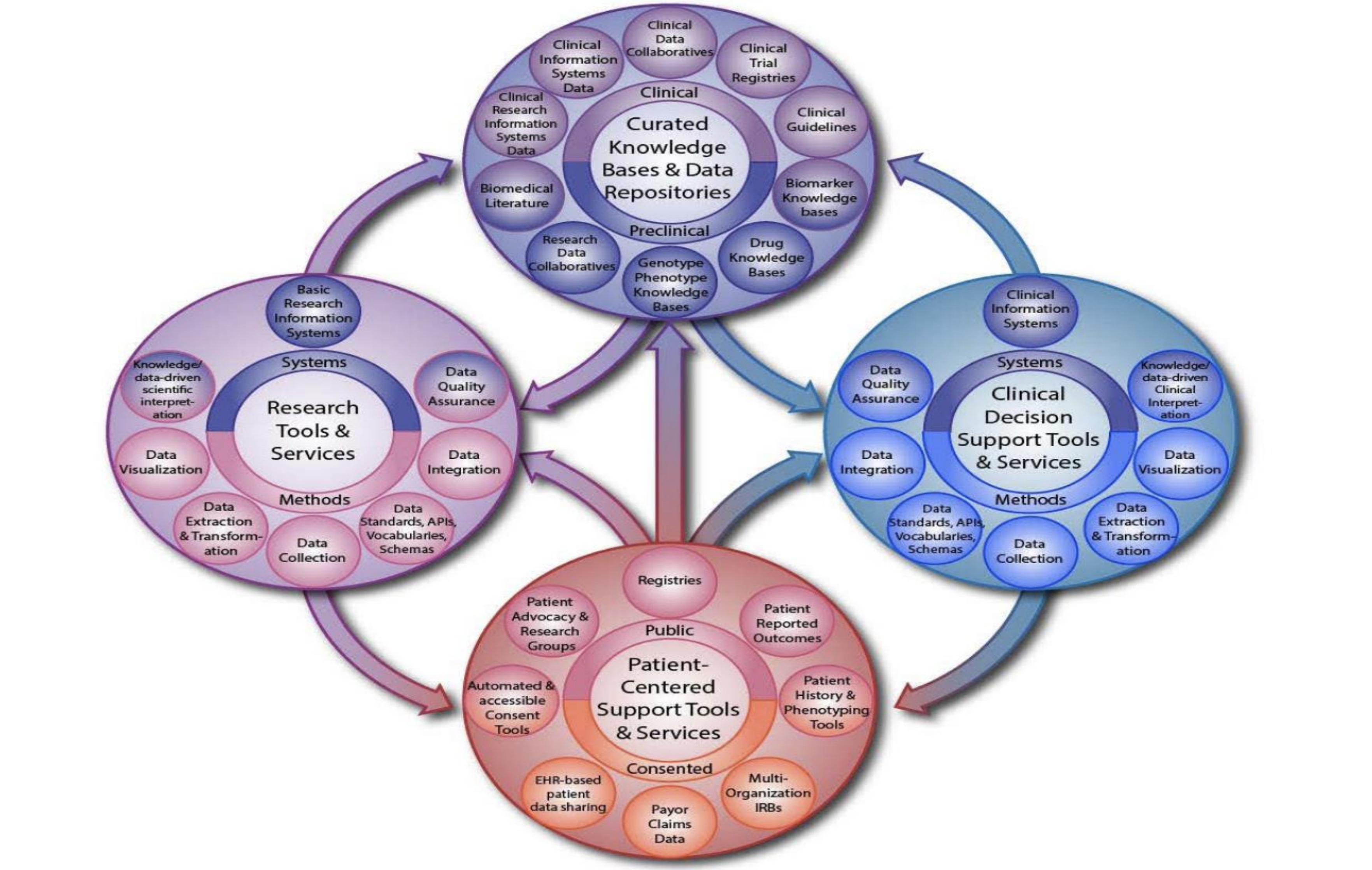
- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Security

- Access large data sets without downloading data
- Bring tools and pipelines to the data
- Bring and combine own data and analyze with existing data
- Workspace to save and share data and results of analyses

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

National Cancer Data Ecosystem


Clinical Information Systems Data
Clinical Data Collaboratives
Clinical Trial Registries
Clinical Research Information Systems Data
Clinical
Curated Knowledge Bases & Data Repositories
Preclinical
Clinical Guidelines
Biomedical Literature
Biomarker Knowledge bases
Research Data Collaboratives
Genotype Phenotype Knowledge Bases
Drug Knowledge Bases
Basic Research Information Systems
Knowledge/ data-driven scientific interpret-ation
Systems
Research Tools & Services
Methods
Data Quality Assurance
Data Visualization
Data Integration
Data Extraction & Transform-ation
Data Collection
Data Standards, APIs, Vocabularies, Schemas
Clinical Information Systems
Data Quality Assurance
Systems
Clinical Decision Support Tools & Services
Methods
Knowledge/ data-driven Clinical Interpret-ation
Data Integration
Data Visualization
Data Standards, APIs, Vocabularies, Schemas
Data Collection
Data Extraction & Transform-ation
Registries
Patient Advocacy & Research Groups
Public
Patient-Centered Support Tools & Services
Consented
Patient Reported Outcomes
Automated & accessible Consent Tools
Patient History & Phenotyping Tools
EHR-based patient data sharing
Payor Claims Data
Multi-Organization IRBs

NIH NATIONAL CANCER INSTITUTE

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

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

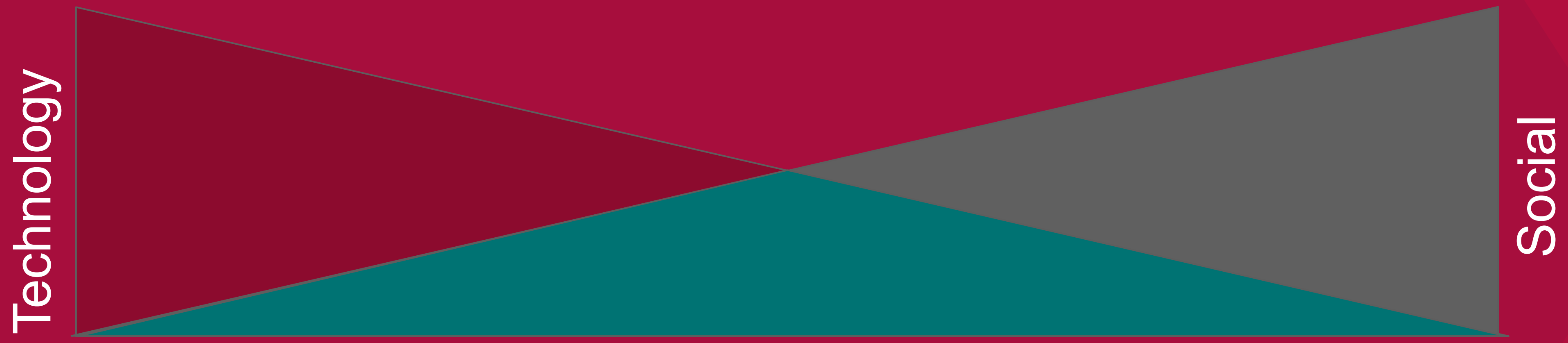
How to survive in the age of automation

A SPECIAL REPORT

The People

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

Data Access and Sharing: Isn't Only a Technology Challenge



Scalable & Secure Environments

Data Harmonization & Organization

Data Sharing & Collaboration

Data Analysis Fluency

Data Access and Sharing: Isn’t Only a Technology Challenge

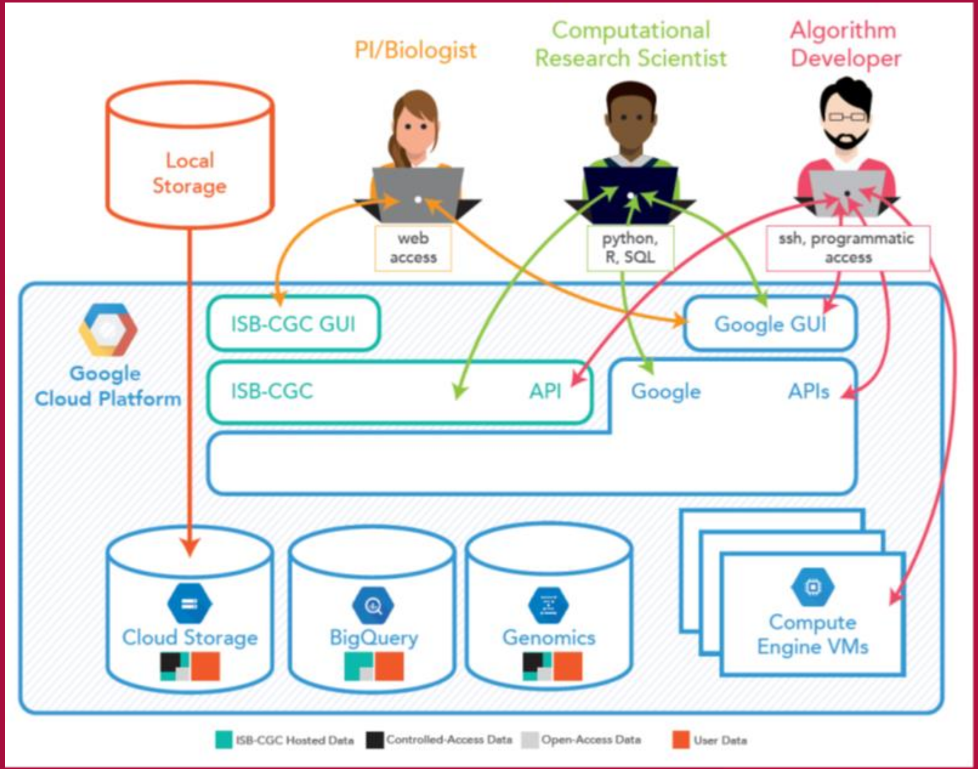
Multidisciplinary Teams with Diverse Expertise and Resources

Biology/Social/Psychology Researcher

Computational Scientist

Algorithm Developer

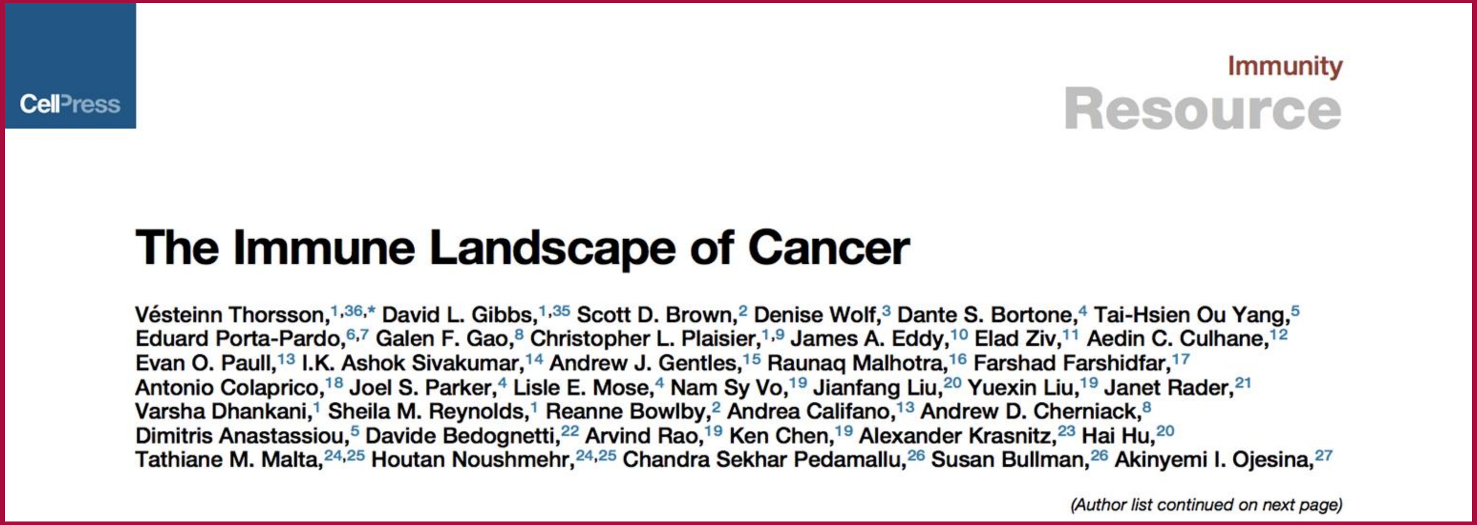
- test new algorithm on hundreds or thousands of BAM or FASTQ files
- run novel image segmentation method across whole-slide images



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

Data Access and Sharing: Isn't Only a Technology Challenge

Multidisciplinary Teams with Diverse Expertise and Resources



CellPress

Immunity

Resource

The Immune Landscape of Cancer

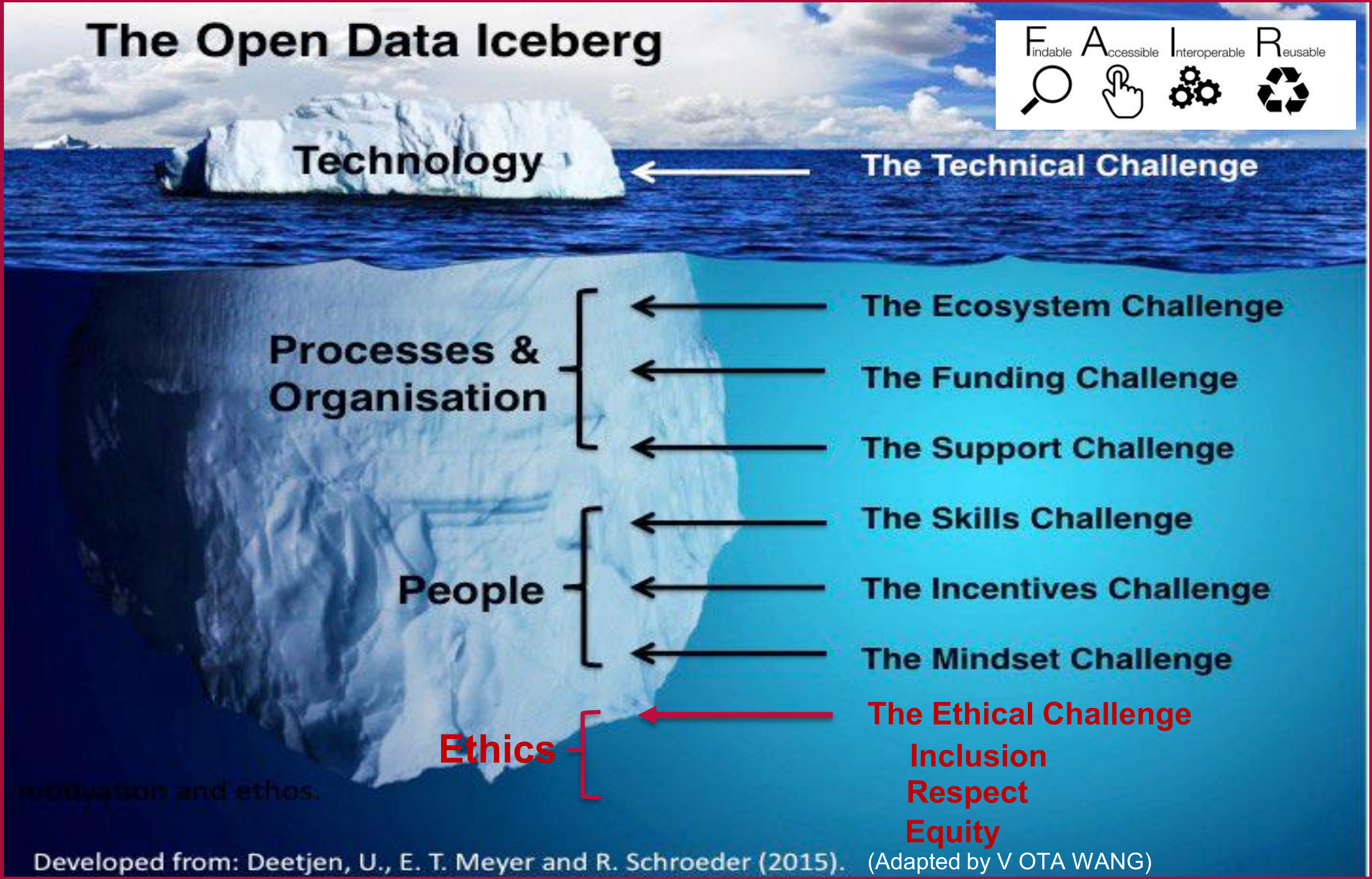
Vésteinn Thorsson,[1,36,*] David L. Gibbs,[1,35] Scott D. Brown,[2] Denise Wolf,[3] Dante S. Bortone,[4] Tai-Hsien Ou Yang,[5] Eduard Porta-Pardo,[6,7] Galen F. Gao,[8] Christopher L. Plaisier,[1,9] James A. Eddy,[10] Elad Ziv,[11] Aedin C. Culhane,[12] Evan O. Paull,[13] I.K. Ashok Sivakumar,[14] Andrew J. Gentles,[15] Raunaq Malhotra,[16] Farshad Farshidfar,[17] Antonio Colaprico,[18] Joel S. Parker,[4] Lisle E. Mose,[4] Nam Sy Vo,[19] Jianfang Liu,[20] Yuexin Liu,[19] Janet Rader,[21] Varsha Dhankani,[1] Sheila M. Reynolds,[1] Reanne Bowlby,[2] Andrea Califano,[13] Andrew D. Cherniack,[8] Dimitris Anastassiou,[5] Davide Bedognetti,[22] Arvind Rao,[19] Ken Chen,[19] Alexander Krasnitz,[23] Hai Hu,[20] Tathiane M. Malta,[24,25] Houtan Noushmehr,[24,25] Chandra Sekhar Pedamallu,[26] Susan Bullman,[26] Akinyemi I. Ojesina,[27]

(Author list continued on next page)

55 authors (+ TCGA network)
34 author affiliations
2 cloud resources + local infrastructures

NIH NATIONAL CANCER INSTITUTE

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



The Open Data Iceberg

Technology ← The Technical Challenge

Processes & Organisation
- ← The Ecosystem Challenge
- ← The Funding Challenge
- ← The Support Challenge

People
- ← The Skills Challenge
- ← The Incentives Challenge
- ← The Mindset Challenge

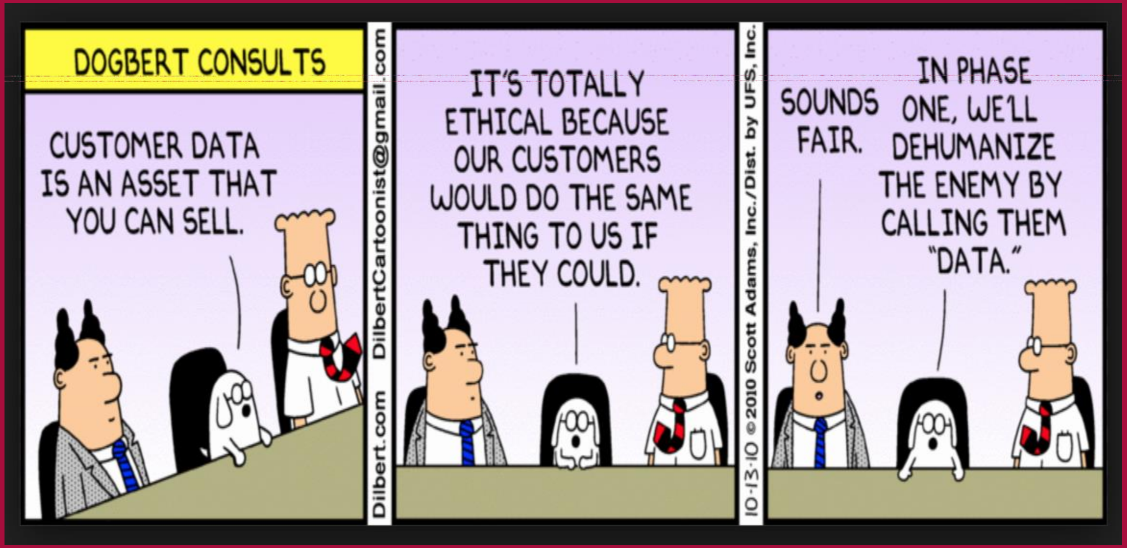
Ethics ← The Ethical Challenge
- Inclusion
- Respect
- Equity

Developed from: Deetjen, U., E. T. Meyer and R. Schroeder (2015). (Adapted by V OTA WANG)

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

datasharing Ethical, Economic, Legal, Social Implications (dEELSI)

- Data and Information are not Neutral


DOGBERT CONSULTS
CUSTOMER DATA IS AN ASSET THAT YOU CAN SELL.
Dilbert.com DilbertCartoonist@gmail.com
IT'S TOTALLY ETHICAL BECAUSE OUR CUSTOMERS WOULD DO THE SAME THING TO US IF THEY COULD.
10-13-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.
SOUNDS FAIR.
IN PHASE ONE, WE'LL DEHUMANIZE THE ENEMY BY CALLING THEM "DATA."

datasharing Ethical, Economic, Legal, Social Implications (dEELSI)

- Data and Information are not Neutral
- Identity, Phenotypes, and Bias
- Identifiability and Privacy
- Bias –Machine Learning and Artificial Intelligence
- Incidental Findings and Return of Results
- Informed Consent and Broad Data Uses
- Governance, and Trustworthiness

Minority and Health Disparity Issues

- Data and Information are not Neutral
  - Stigma: People/groups/ communities/diagnoses/phenotypes
  - Inclusion: Data collected in basic/applied/clinical trial research
  - Diversity and Workforce issues
- Citizen Science and *community and patient engagement*
- Inclusion, Equity, and Data Access – The Haves and Have Nots

Challenges

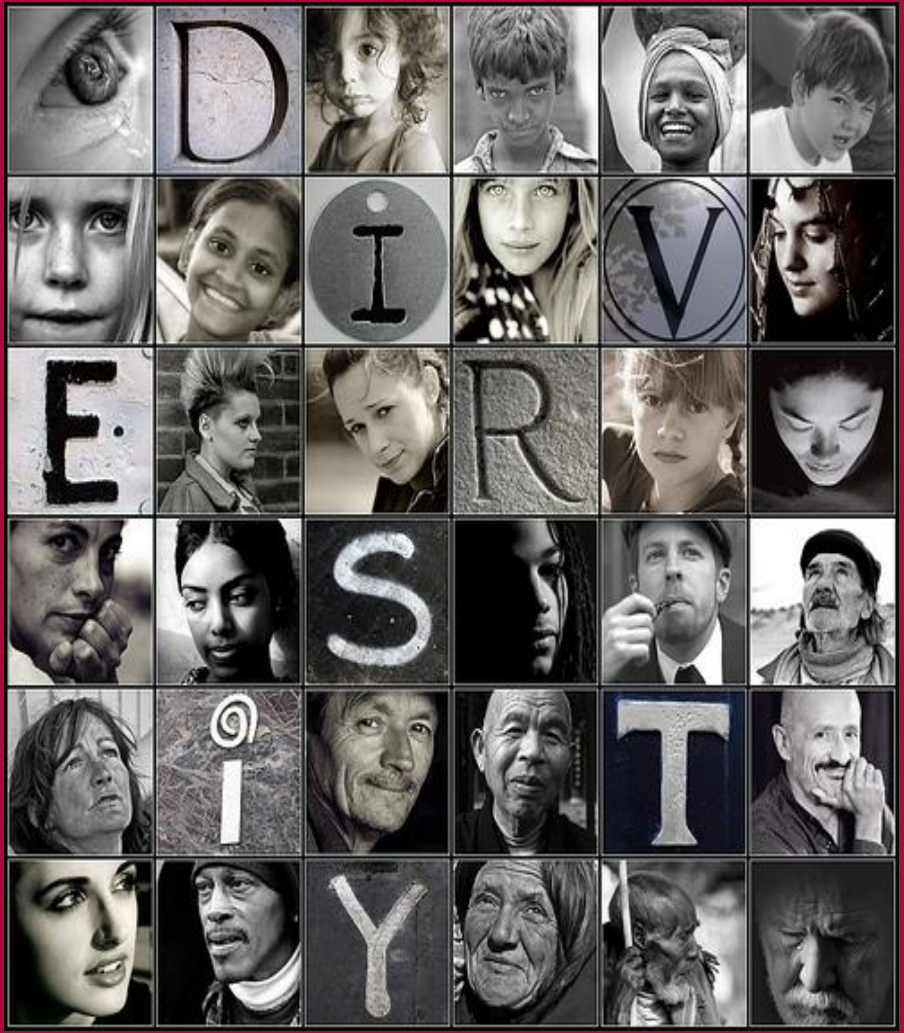
The World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
WIRED — 16.07
JUL.2008
Do we need THEORY?
How do we find DARK MATTER?
Is Earth safe from ASTEROIDS?
Where will TERRORISTS strike next?
How can you win any LAWSUIT?
Who will be the next PRESIDENT?
How will we grow enough FOOD?
How can we protect BUSINESS from risk?
The End of Science
The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data. Welcome to the Petabyte Age.

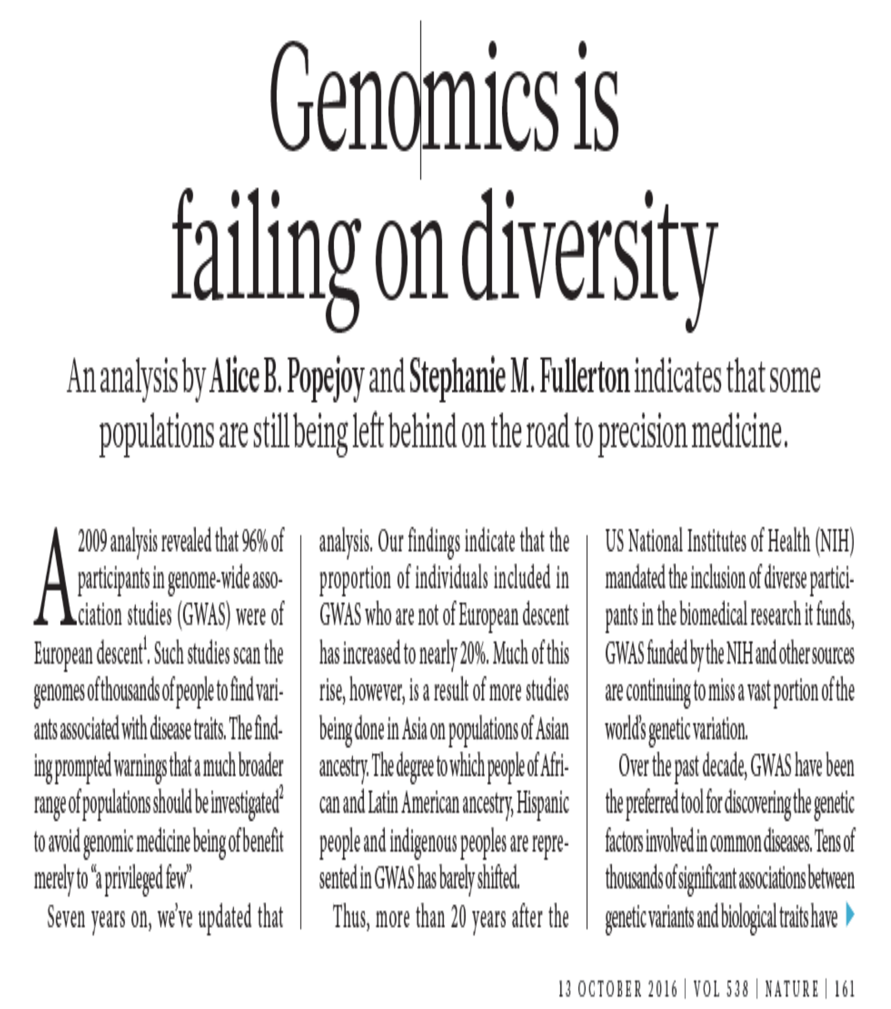
Data are difficult to

- *collect*
- *store*
- *delete*
- *search*
- *share*
- *visualize*
- *curate*
- *process*
- *analyze*

with current available databases and tools

Challenges





Genomics is failing on diversity

An analysis by **Alice B. Popejoy** and **Stephanie M. Fullerton** indicates that some populations are still being left behind on the road to precision medicine.

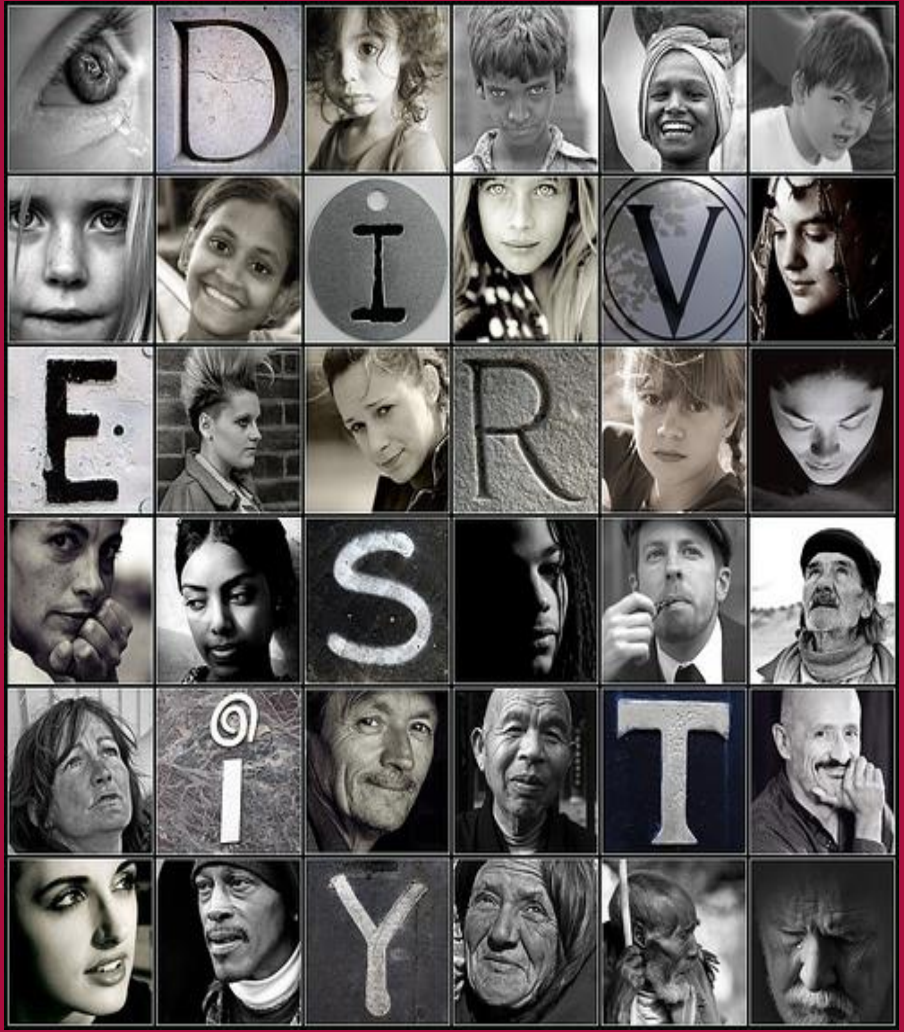
A 2009 analysis revealed that 96% of participants in genome-wide association studies (GWAS) were of European descent[1]. Such studies scan the genomes of thousands of people to find variants associated with disease traits. The finding prompted warnings that a much broader range of populations should be investigated[2] to avoid genomic medicine being of benefit merely to "a privileged few".

Seven years on, we've updated that analysis. Our findings indicate that the proportion of individuals included in GWAS who are not of European descent has increased to nearly 20%. Much of this rise, however, is a result of more studies being done in Asia on populations of Asian ancestry. The degree to which people of African and Latin American ancestry, Hispanic people and indigenous peoples are represented in GWAS has barely shifted.

Thus, more than 20 years after the US National Institutes of Health (NIH) mandated the inclusion of diverse participants in the biomedical research it funds, GWAS funded by the NIH and other sources are continuing to miss a vast portion of the world's genetic variation.

Over the past decade, GWAS have been the preferred tool for discovering the genetic factors involved in common diseases. Tens of thousands of significant associations between genetic variants and biological traits have ▸

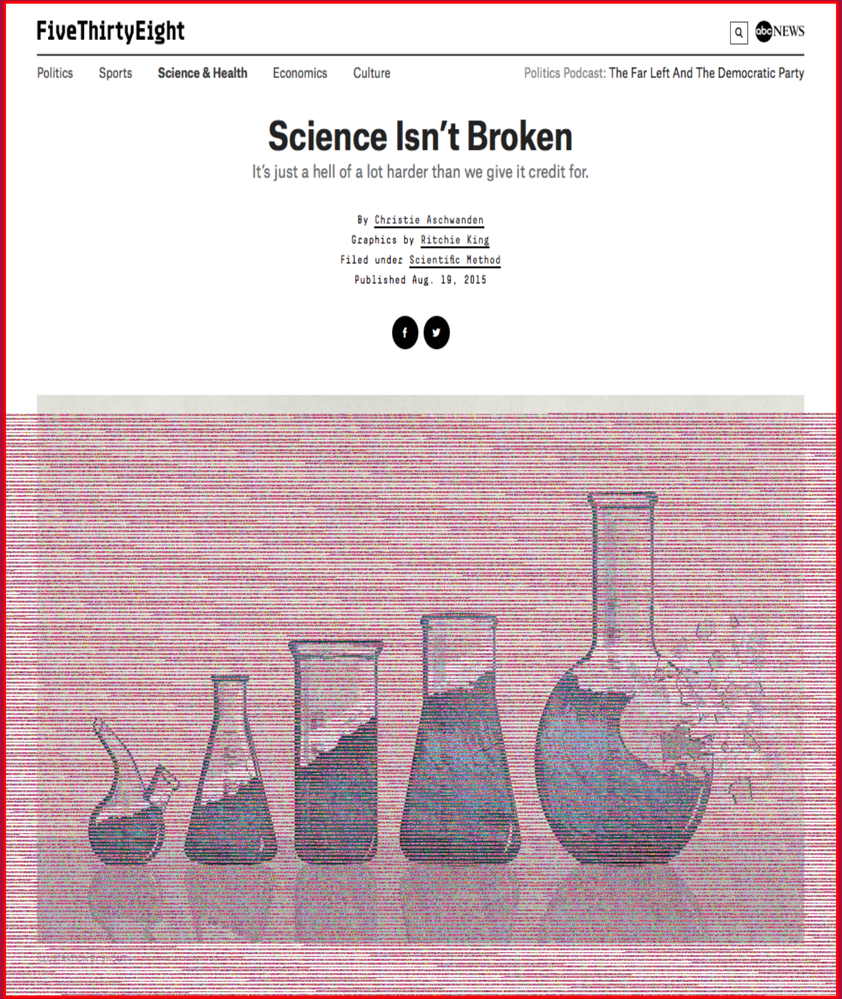
Challenges



How to have integrity
in [data
and data sharing]
in a world
that does not
affirm everyone's
humanity

- Adapted from Thomas A. Parham

Challenges



Science Isn't Broken

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

By Christie Aschwanden
Graphics by Ritchie King
Filed under Scientific Method
Published Aug. 19, 2015

Science Isn't Broken

"…it's just…a lot harder than we give it credit for…"

Challenges

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

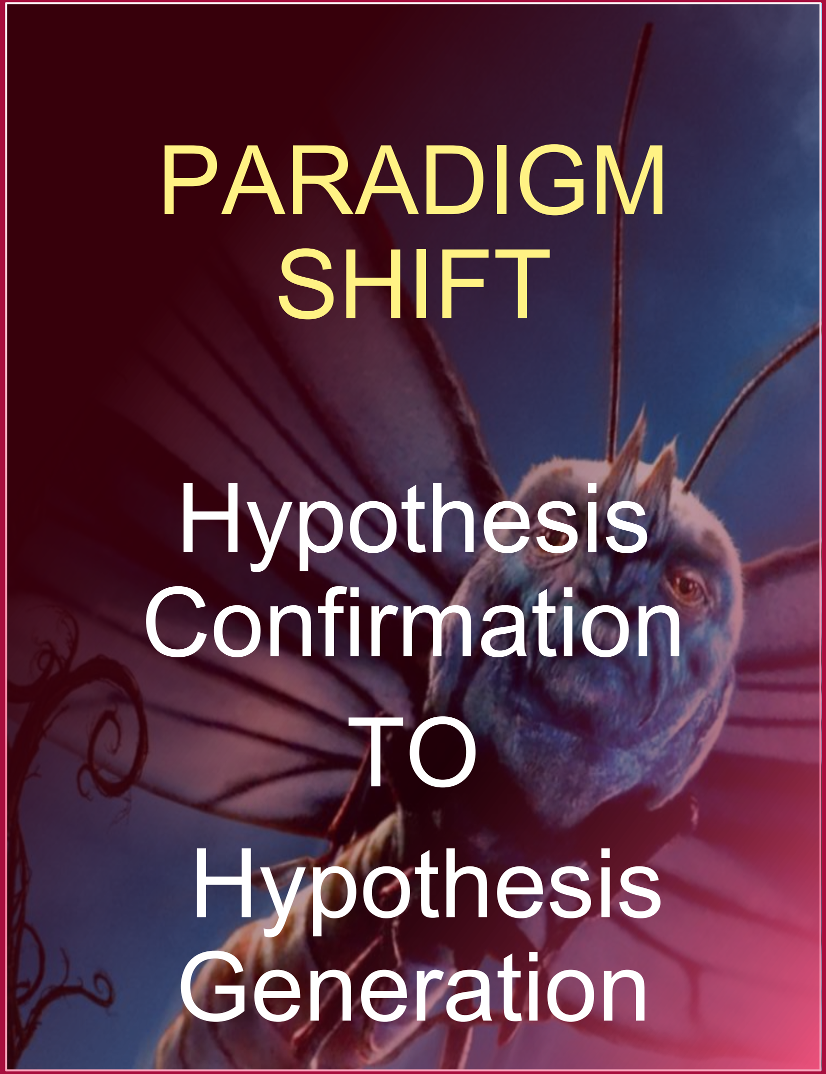
https://www.economist.com/leaders/2013/10/21/how-science-goes-wrong

Science Isn't Broken

"…it's just…a lot harder than we give it credit for…"

"Now it needs to change itself…"

The Challenge

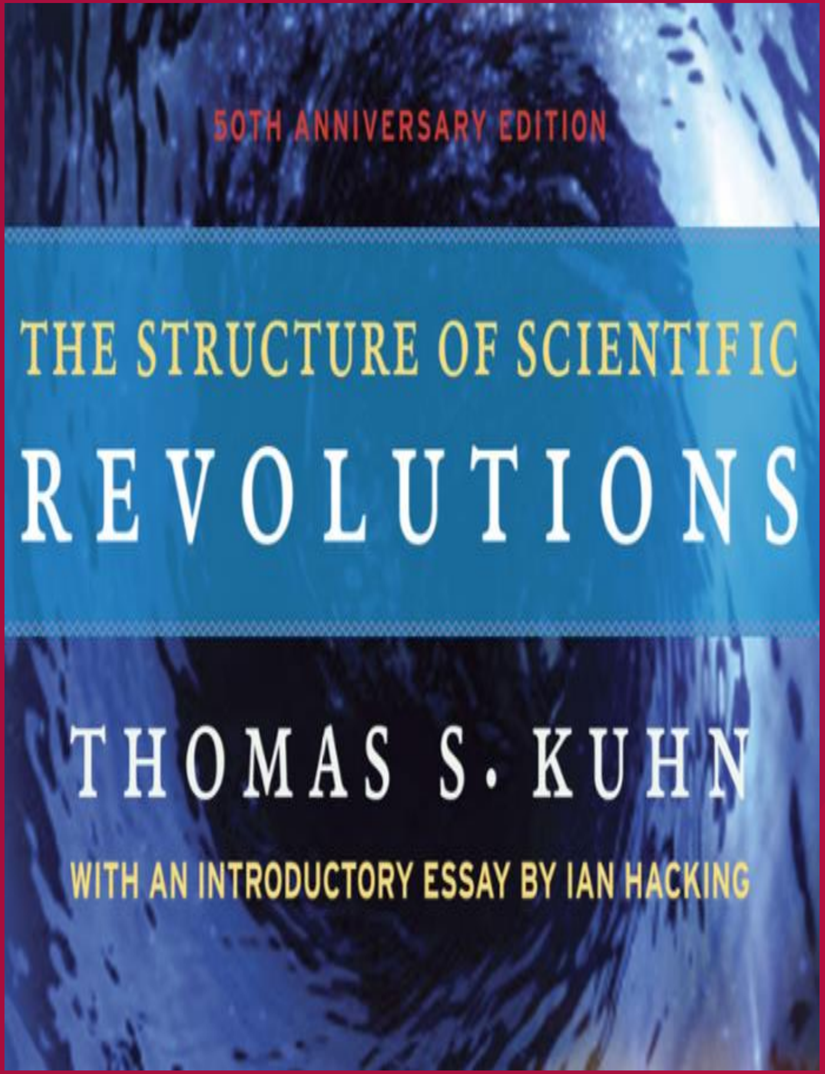


PARADIGM SHIFT

Hypothesis Confirmation

TO

Hypothesis Generation



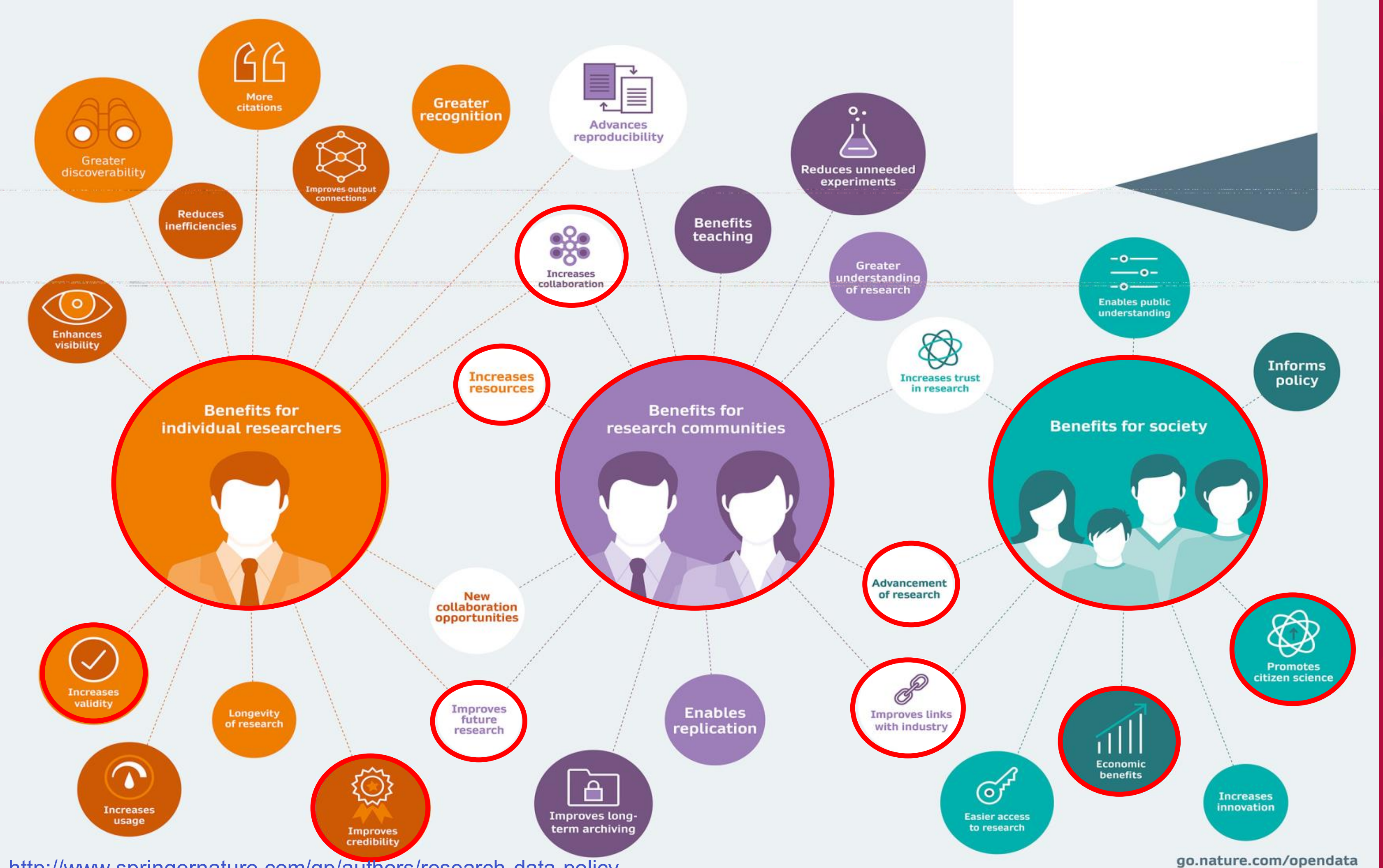
50TH ANNIVERSARY EDITION

THE STRUCTURE OF SCIENTIFIC REVOLUTIONS

THOMAS S. KUHN

WITH AN INTRODUCTORY ESSAY BY IAN HACKING

The Challenge

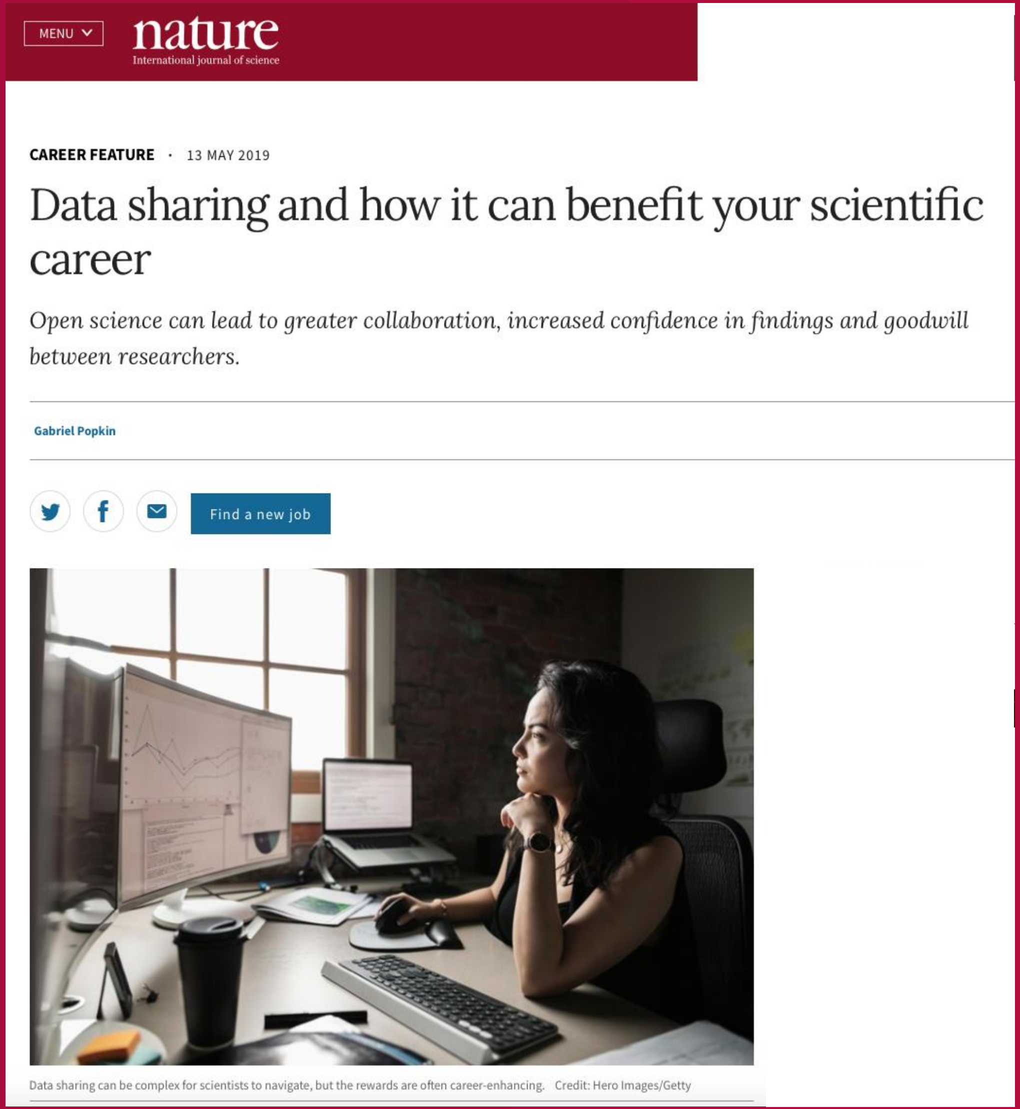


http://www.springernature.com/gp/authors/research-data-policy

Glass and MCAtee (2006).

Setting You and your Data Free

- Catalyze new collaborations
- Increase confidence in results
- Larger projects (scope & sample size)
- Local vs global (>external validity)
- Generate greater recognition
- Credit
  - Digital Object Identifier (DOI) enables independent discoverable citability for researcher credit
  - Data tracking on the impact of research
  - Journal data availability statements
  - Open data badges



CAREER FEATURE · 13 MAY 2019

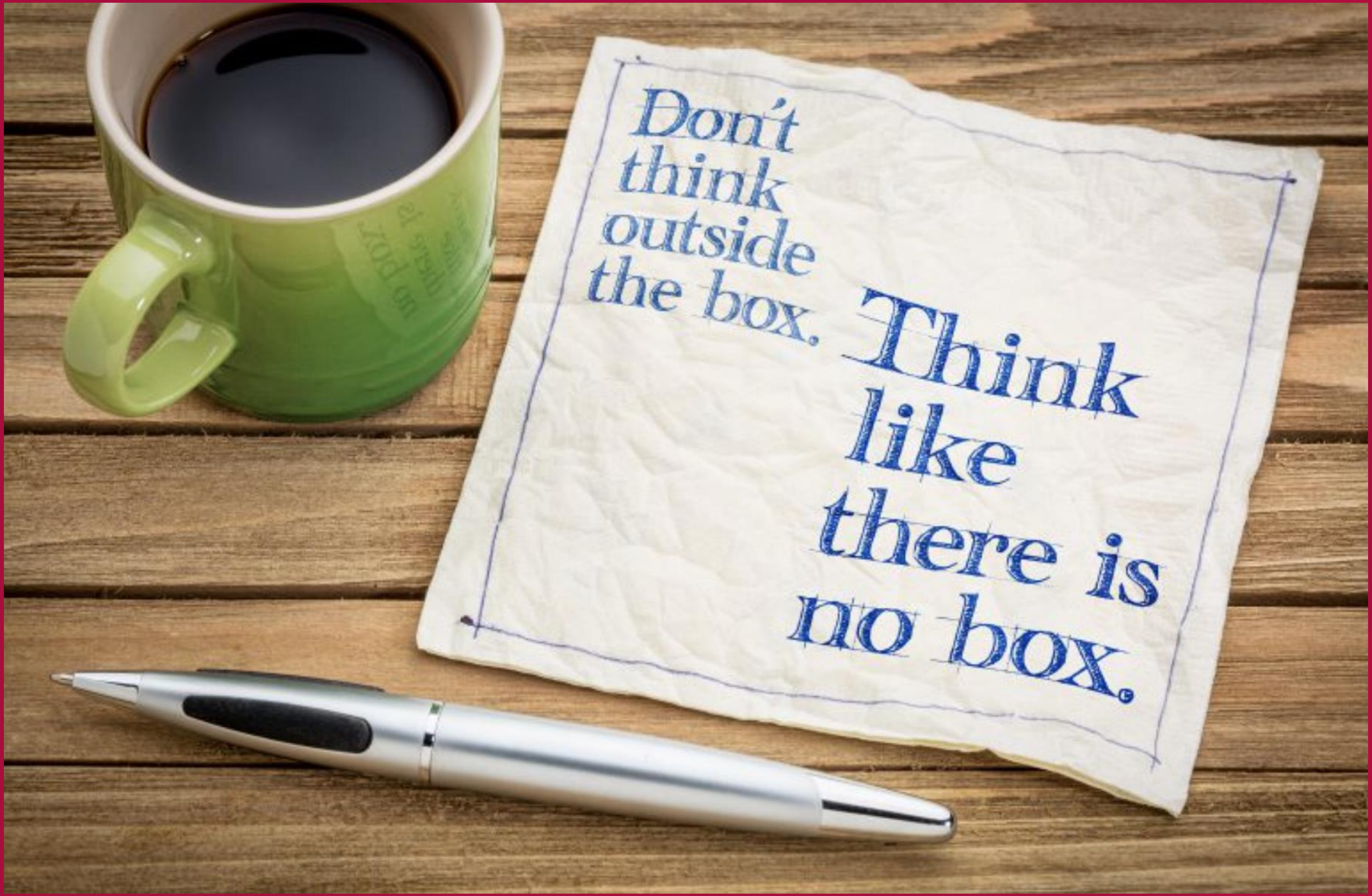
Data sharing and how it can benefit your scientific career

Open science can lead to greater collaboration, increased confidence in findings and goodwill between researchers.

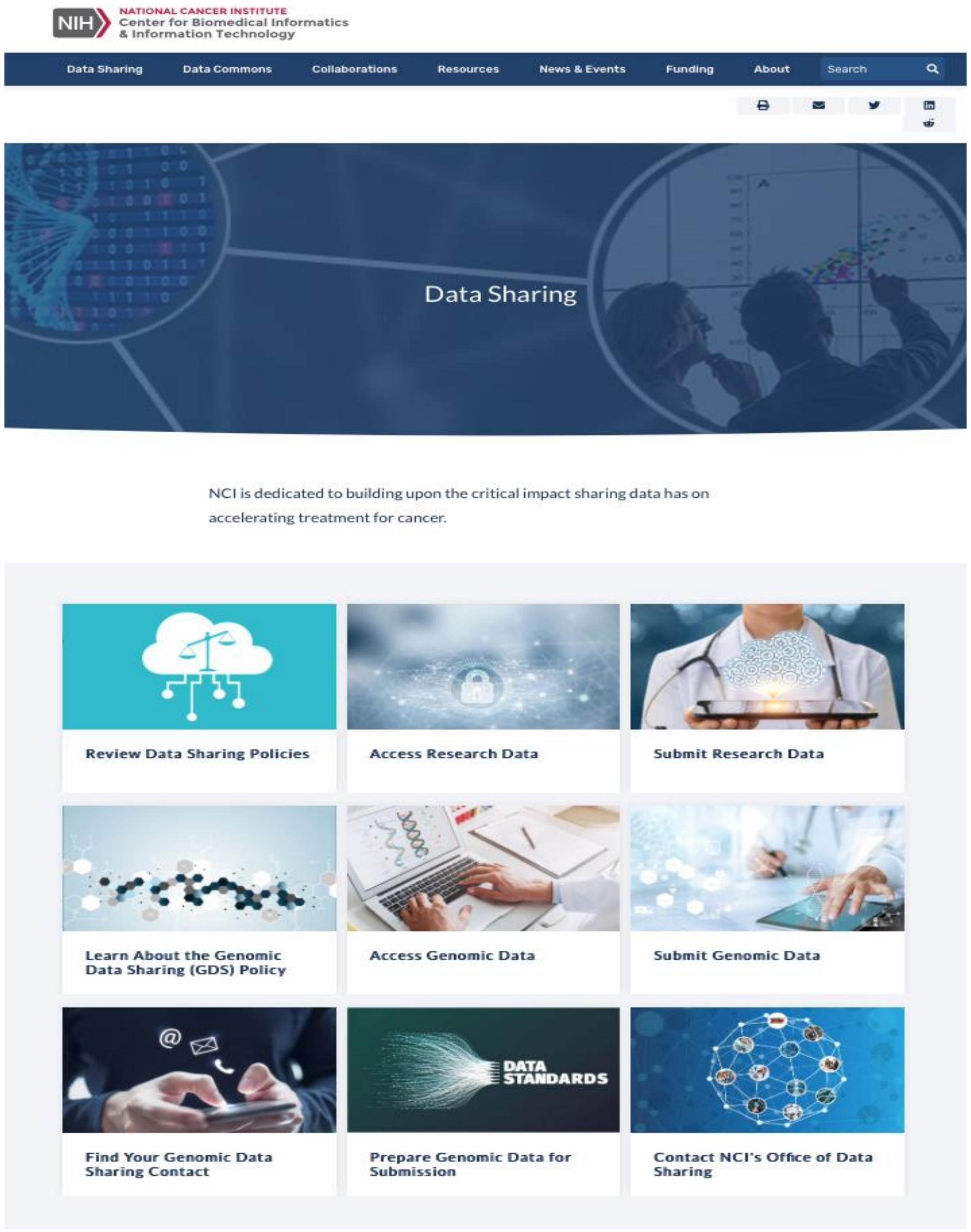
Gabriel Popkin

Data sharing can be complex for scientists to navigate, but the rewards are often career-enhancing. Credit: Hero Images/Getty

Popkin (2019). Data sharing and how it can benefit your scientific career. *Nature, (569, 446-7.*


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

https://datascience.cancer.gov/data-sharing



Data Sharing

NCI is dedicated to building upon the critical impact sharing data has on accelerating treatment for cancer.

Review Data Sharing Policies

Access Research Data

Submit Research Data

Learn About the Genomic Data Sharing (GDS) Policy

Access Genomic Data

Submit Genomic Data

Find Your Genomic Data Sharing Contact

Prepare Genomic Data for Submission

Contact NCI's Office of Data Sharing

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol