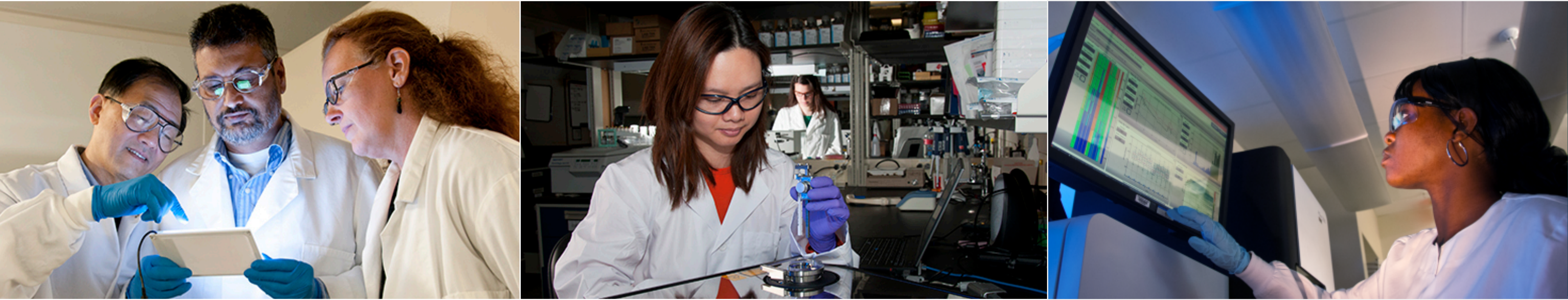
Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute



Data as the Driver of Biomedical Research: Why aren’t we going faster?

Jack R. Collins
Director, Advanced Biomedical/Computational Sciences, FNLCR/NCI

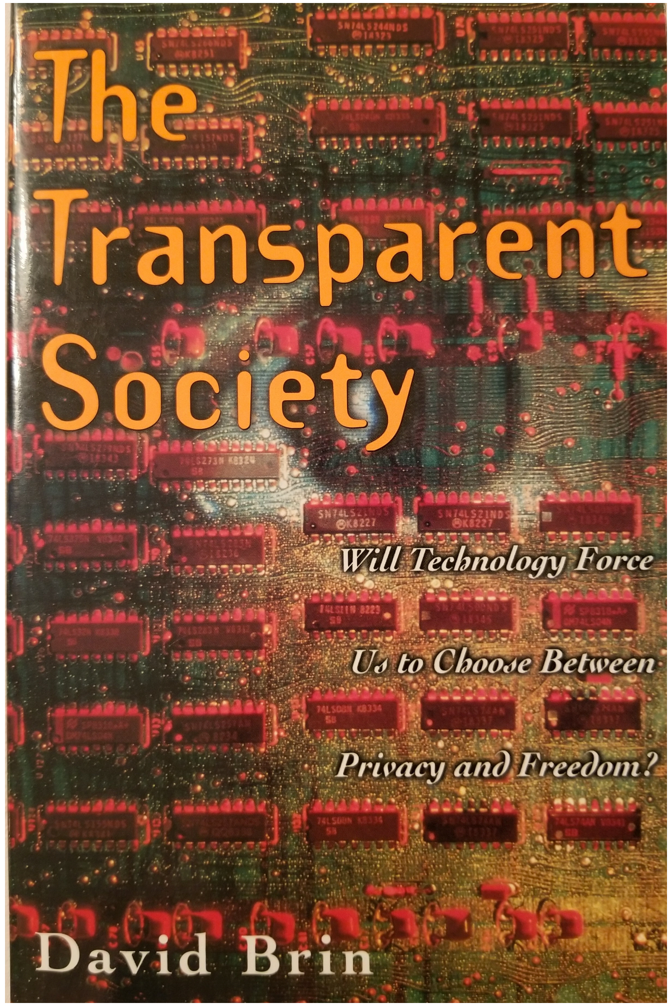
Date October 10, 2019

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

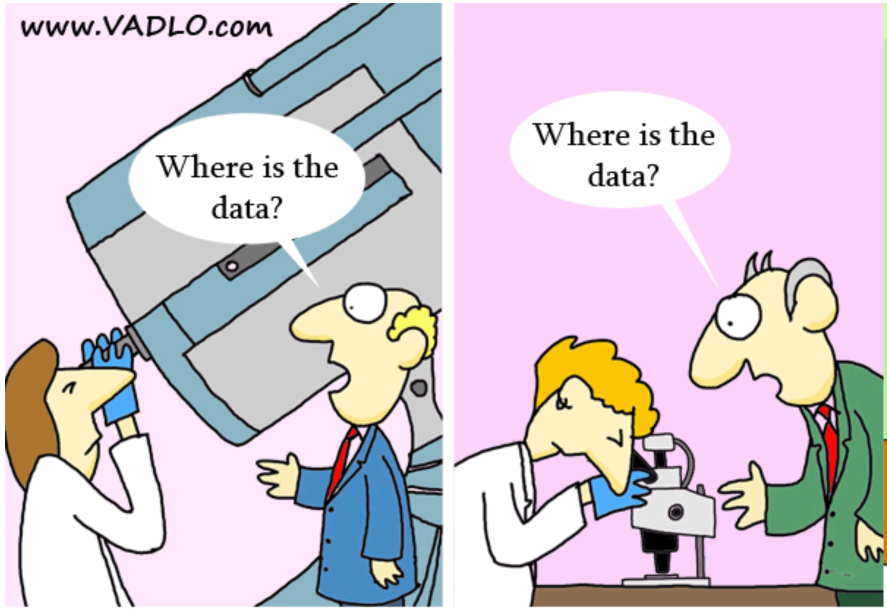
Acknowledgements and Disclaimer

- Curtis Lisle, KnowledgeVis
- Yanling Liu, ABCS
- Hyun Jung, ABCS
- Uma Mudunuri, ABCS
- Our Collaborators and Sponsors at FNLCR, NCI/NIH
- The CLSAC organizers for inviting me.


The Transparent Society
Will Technology Force Us to Choose Between Privacy and Freedom?
David Brin

Disclaimer: I am not an attorney or an expert in data privacy, the legal framework and rules associated with privacy, or in the overall polcy issues involved with data sharing. I do have experience trying to get data for biomedical research, navigating many of the rules, and observing and developing technologies that are outpacing policy. Opinions are, of course, my own.

Biology is Data Driven and is dependent on proper data management, integration/sharing, and analyses



Grand Unification of Sciences

“Data don’t make any sense, we will have to resort to statistics.”

- Genomic Sequencing
- RNA-Seq
- Image Analysis
- Microarray
- Protein-Protein Interactions
- Structural Biology
- Computational Oncology

- Machine Learning / AI
- Predictive Models
- Model Systems

Overview of Presentation

- Data sharing and privacy issues profoundly affect biomedical research, especially in the area of (era of) big data, computational/data sciences, machine learning/AI, and evaluating predictive models.
- We will address data sharing policies at NIH (rules vs reality), and issues that make data sharing (data reuse) challenging.
- We will also discuss the fact that data is often useless unless it’s properly annotated and documented so that it can facilitate productive use of the data.
- Data, and the analysis, is Global.
- Examples of real-world workflows, challenges, and lessons learned will be discussed.

Interplay between Technology, Science, Medicine, Society, and Data (Sharing/Privacy)

Complementary DNA sequencing
- - -
WWW goes public

Haemophilus influenzae Sequenced – 1st whole genome

HIPAA

Google launched

Human Genome Published

Solexa Genome Analyzer (NGS)

1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005

Common Rule

9/11

PNAS MD/DL Geno->Pheno

HIPAA – Genetic info is Health info

Deep Belief Networks Published

iPhone

Fitbit

GPU-enabled DL

DL ImageNet Victory

OPM Breach

EU Privacy Shield Decision

Cambridge Analytica Disclosure

CRISPR Human Tests

2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

1TB hard drive

GINA

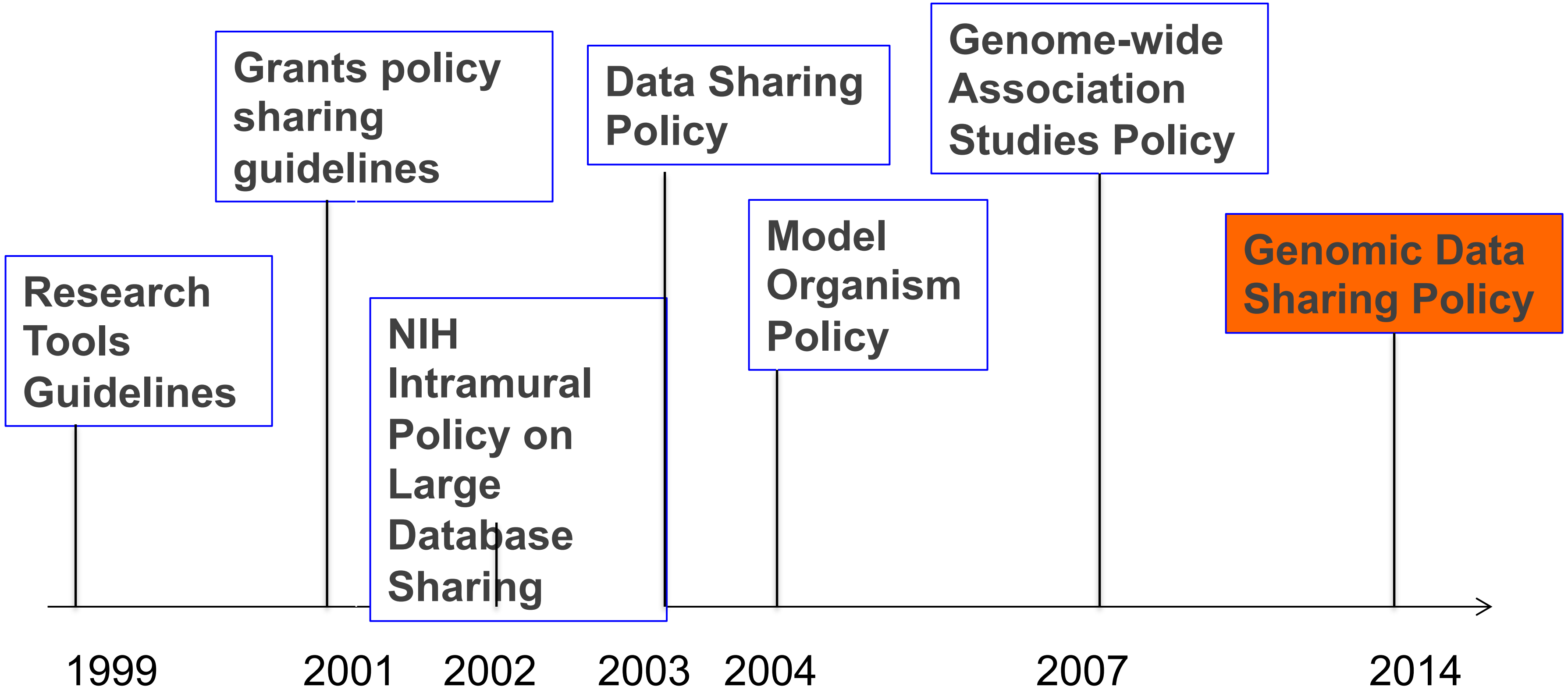
SSN Predicted from Public Data

Individuals re-identified from genomic, genealogy, public databases

rCR - Revised Common Rule “Final Rule” 2019 Enforce

1TB μSD card

NIH: A Culture of Sharing



1999 Research Tools Guidelines

2001 Grants policy sharing guidelines

2002 NIH Intramural Policy on Large Database Sharing

2003 Data Sharing Policy

2004 Model Organism Policy

2007 Genome-wide Association Studies Policy

2014 Genomic Data Sharing Policy

"Sharing research data supports the NIH mission"

NIH Genomic Data Sharing Policy, 2014

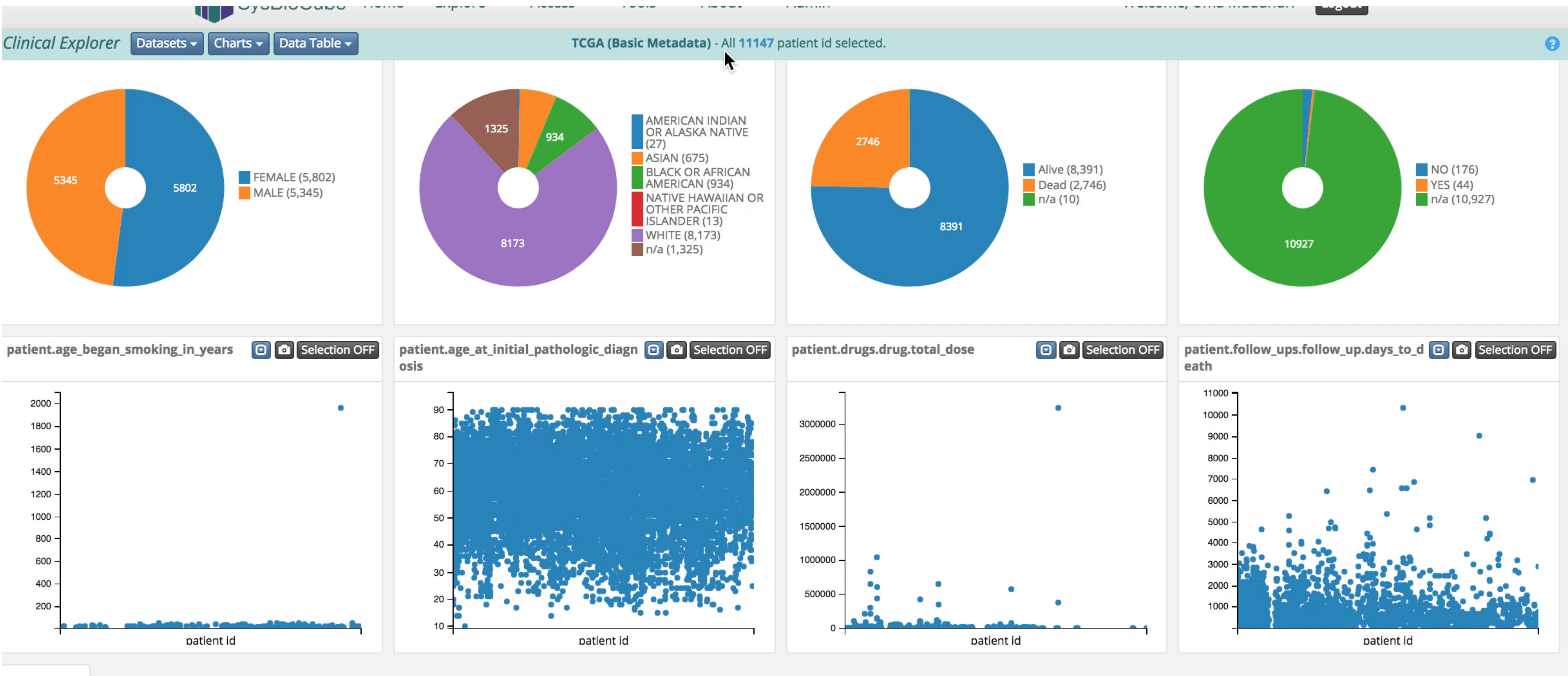
Data Management Challenges

Primarily Human Data

- **Need to know what you have**
- **Properly Annotated / Structured**
- **Ontologies play important role**
- **Determine what is PII / PHI**
- **Is it anonymized? Can it be re-identified?**
- **Who “owns” the data? Are there “special rules”? Is it published? Confidentiality?**
- **Is there an IRB / Consent associated?**
- **And can be harder than first thought –**
  - Especially if we’re reactive rather than proactive

Exploring and Understanding the Data

Choosing quality data (relevant cohorts)



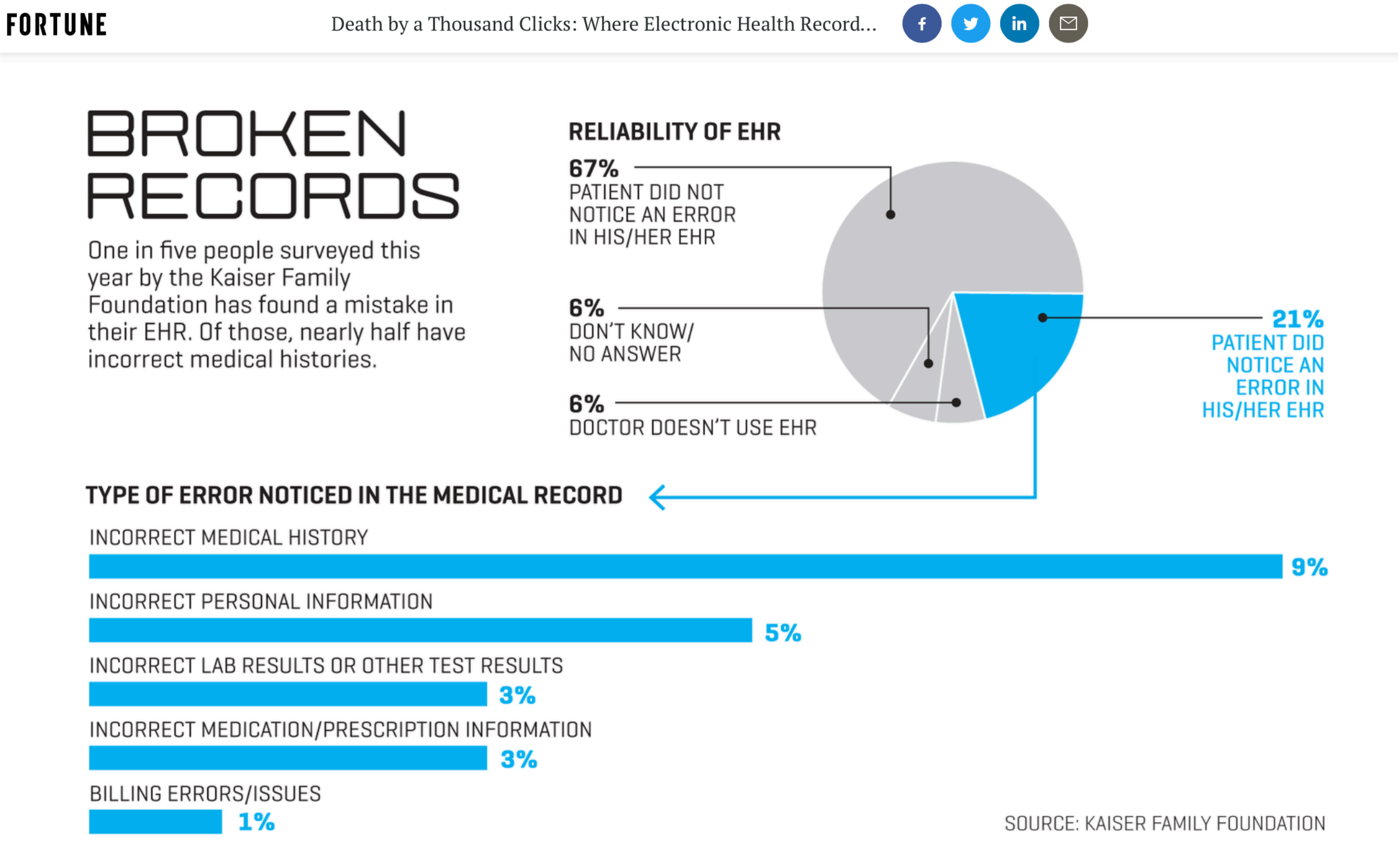
Clinical Explorer

TCGA (Basic Metadata) - All 11147 patient id selected.

Data Management has Consequences
(Fortune Magazine, April 2019)

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



FORTUNE Death by a Thousand Clicks: Where Electronic Health Record...

BROKEN RECORDS

One in five people surveyed this year by the Kaiser Family Foundation has found a mistake in their EHR. Of those, nearly half have incorrect medical histories.

RELIABILITY OF EHR
67% PATIENT DID NOT NOTICE AN ERROR IN HIS/HER EHR
6% DON'T KNOW/ NO ANSWER
6% DOCTOR DOESN'T USE EHR
21% PATIENT DID NOTICE AN ERROR IN HIS/HER EHR
TYPE OF ERROR NOTICED IN THE MEDICAL RECORD
INCORRECT MEDICAL HISTORY 9%
INCORRECT PERSONAL INFORMATION 5%
INCORRECT LAB RESULTS OR OTHER TEST RESULTS 3%
INCORRECT MEDICATION/PRESCRIPTION INFORMATION 3%
BILLING ERRORS/ISSUES 1%

SOURCE: KAISER FAMILY FOUNDATION

Data Sharing
Balance between Risks and Benefits

Risks

- Personal
- Family
- Societal
- …

Benefits

- Facilitates Research
- Can enable insights
  - Especially in instances of rare events
- …

Federal Regulations Governing Human Subjects Research

Federal Regulations Governing Human Subjects Research

Published in 1991, The Federal Policy for the Protection of Human Subjects-also known as the "Common Rule"- establishes the baseline standard of ethics for government-funded research in the United States. The Common Rule requires all federally funded research projects to obtain informed consent from each participant prior to their participation. Participants must be informed of all the potential risks of the particular study, including risks associated with release of their private information. Informed consents for genomic research should clarify the uses of research results, including with whom the information will be shared. It has been shown that, when given control over when and with whom their research data is shared, most individuals are eager to participate in research studies, fueling scientific discovery and medical progress. For further information about informed consent in genomics and guidance for researchers or IRB members, please see the Informed Consent for Genomics Research Resource.

January 19, 2017--Final rule implementing major revisions to the Common Rule

Summary : This final rule strengthens protections for people who volunteer to participate in research, while ensuring that the oversight system does not add inappropriate administrative burdens, particularly to low-risk research. It also allows more flexibility in keeping with today's dynamic research environment. The final rule will now generally expect consent forms to include a concise explanation – at the beginning of the document – of the key information that would be most important to individuals contemplating participation in a particular study, including the purpose of the research, the risks and benefits, and appropriate alternative treatments that might be beneficial to the prospective subject.

Federal Regulations Governing Human Subjects Research

Genetic Information Nondiscrimination Act (GINA)

The Genetic Information and Nondiscrimination Act of 2008 (GINA) protects the genetic privacy of the public, including research participants. The passage of GINA makes it illegal for health insurers or employers from requesting or requiring genetic information of an individual or of family members (and further prohibits the discriminatory use of such information).

HIPAA

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule establishes protections to maintain the confidentiality of patients' individually identifiable health information. Such information held by entities covered by HIPAA, such as a health care provider or insurance company, is defined as Protected Health Information (PHI) and there are limits on when and with whom PHI may be shared. In 2013, as required by the passage of the Genetic Information Nondiscrimination Act, the Privacy Rule was modified to establish that **genetic information is health information** protected by the Privacy Rule to the extent that such information is individually identifiable, and that HIPAA covered entities may not use or disclose protected health information that is genetic information for underwriting purposes. There are no such restrictions on the use or disclosure of PHI that has been de-identified.

Genomic Data Sharing and Privacy @NIH

- To advance genomics research, NIH houses a number of databases through which researchers can share de-identified genomic data. Given the need to consider participant privacy, it is important to minimize the possibility that any research participants are identified. Indeed, a **study** published in 2013 demonstrated that it is possible to re-identify research participants using genomic data from one such database alongside genealogical databases and public records. NIH therefore controls access to sensitive or potentially identifiable information held in these databases to ensure that the privacy of the research participants is respected. (See Genomic Data Sharing Policy below.) In addition, NIH issues **Certificates of Confidentiality** to enable NIH-funded researchers to limit access to research participant information held at grantee institutions.
- People have a right to keep their medical information, and that of their dependents, private. Yet medical records are a rich source of research data, and it is in the interest of medical research, and thus everyone's health and well-being, that scientists have access to large numbers of participants and quantities of data. How do we strike the proper balance between scientific progress and patient privacy? Federal laws, like the **Common Rule** and the **Health Insurance Portability and Accountability Act** (HIPAA) aim to strike that delicate balance.
- **https://www.genome.gov/about-genomics/policy-issues/Privacy**

Anonymization

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute

The following identifiers of the individual or of relatives, employers, or household members of the individual must be removed to achieve the “safe harbor” method of de-identification: (A) Names; (B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of Census (1) the geographic units formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000; (C) All elements of dates (except year) for dates directly related to the individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older; (D) Telephone numbers; (E) Fax numbers; (F) Electronic mail addresses: (G) Social security numbers; (H) Medical record numbers; (I) Health plan beneficiary numbers; (J) Account numbers; (K) Certificate/license numbers; (L) Vehicle identifiers and serial numbers, including license plate numbers; (M) Device identifiers and serial numbers; (N) Web Universal Resource Locators (URLs); (O) Internet Protocol (IP) address numbers; (P) Biometric identifiers, including finger and voice prints; (Q) Full face photographic images and any comparable images; and ® any other unique identifying number, characteristic, or code, except as permitted for re-identification purposes provided certain conditions are met. In addition to the removal of the above-stated identifiers, the covered entity may not have actual knowledge that the remaining information could be used alone or **in combination with any other information to identify an individual** who is subject of the information. 45 C.F.R. § 164.514(b)

https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

Identifiable populations

Ethnically, geographically, and linguistically identifiable populations present particular concerns with regard to privacy, stigmatization, and discrimination, since the ability to protect the privacy of these individuals or groups participating in the research is diminished. For example, members of an identifiable population may be stigmatized or discriminated against if research reveals that the group is at high risk of having a genetic variant associated with a particular disease. For some communities, close family relationships also may make it especially challenging to protect participants' privacy, **even if research samples are de-identified**.

https://www.genome.gov/about-genomics/policy-issues/Privacy

Data are …

- Ubiquitous
- Cheap to generate
- Valuable to utilize
- Mobile
- “Big” and growing exponentially
- Often Aggregated/Integrated
- Mutable
- Potentially Beneficial (can power progress)
- Potentially Harmful (“weaponized”)
- Protected (in some cases)
- In need of Interpretation.
- Not Generally Self-Annotating

“ipsa scientia potestas est”

“knowledge itself is power” – Sir Francis Bacon
Commodification = “and money”

- Within a capitalist economic system, **Commodification** is the transformation of goods, services, ideas and people into commodities or objects of trade. A commodity at its most basic, according to Arjun Appadurai, is "anything intended for exchange," or any object of economic value.[1]
- Commodification is often criticised on the grounds that some things ought not to be treated as commodities—for example water, education, data, information, knowledge, human life, and animal life.
- The word *commodification*, which describes assignment of economic value to something not previously considered in economic terms
- Information Economy Impacts Data Sharing and Privacy
  - Maximizing profit in an information economy often conflicts with transparency, data sharing, and privacy.
  - Policy is keeping up with technological changes – reactive, in general

https://en.wikipedia.org/wiki/Commodification
Accessed Oct 2, 2019

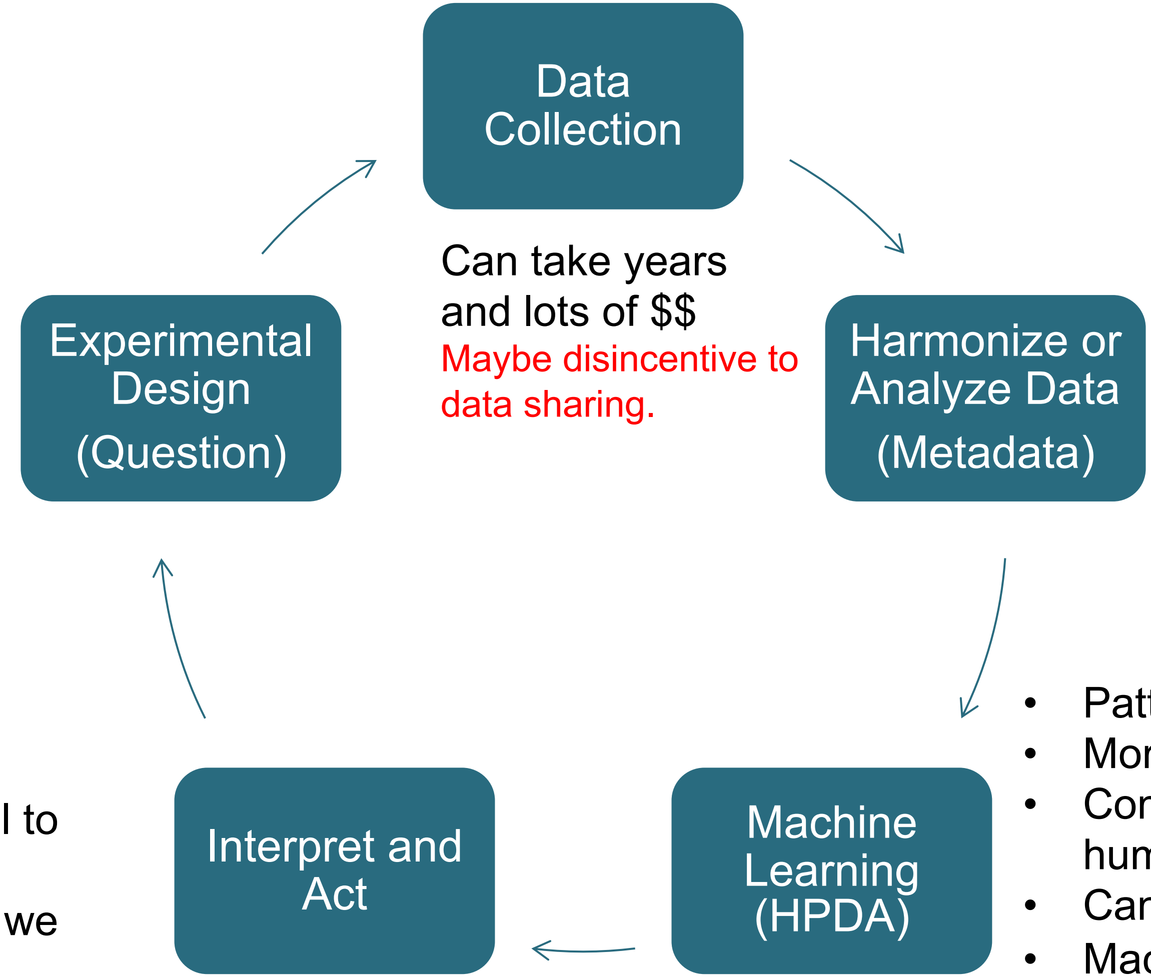
Technologies are …

- Data Hungry
- Computationally Scalable
- Facilitating Large-Scale Aggregation/Integration of Data
- Changing Faster than Laws and Policy

Machine Learning is – ALL ABOUT DATA (And speeding up and automating analysis)

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



Data Collection

Harmonize or Analyze Data (Metadata)

Machine Learning (HPDA)

Interpret and Act

Experimental Design (Question)

Can take years and lots of $$ Maybe disincentive to data sharing.

- May include HPC for analysis and generate even more data (integrated with simulation).
- Summaries and/or data Integration from multiple sources
- May involve data reduction.
- May also be expensive $$

- Pattern recognition
- More data usually better
- Computer ”sees” data differently than human
- Can be highly compute intensive
- Machines do what we tell them to do so be careful
- May lead to proprietary insights and competitive advantage..

- Interpretation is critical to intelligent action.
- Must understand how we got to the solution.
- Business behavior may not align or be subject to federal policy.

Critical Step – That must be informed by all of the other steps in the process (Think hard here.) Balance between science, policy, and privacy.

Data and Sampling Bias are Critical Issues

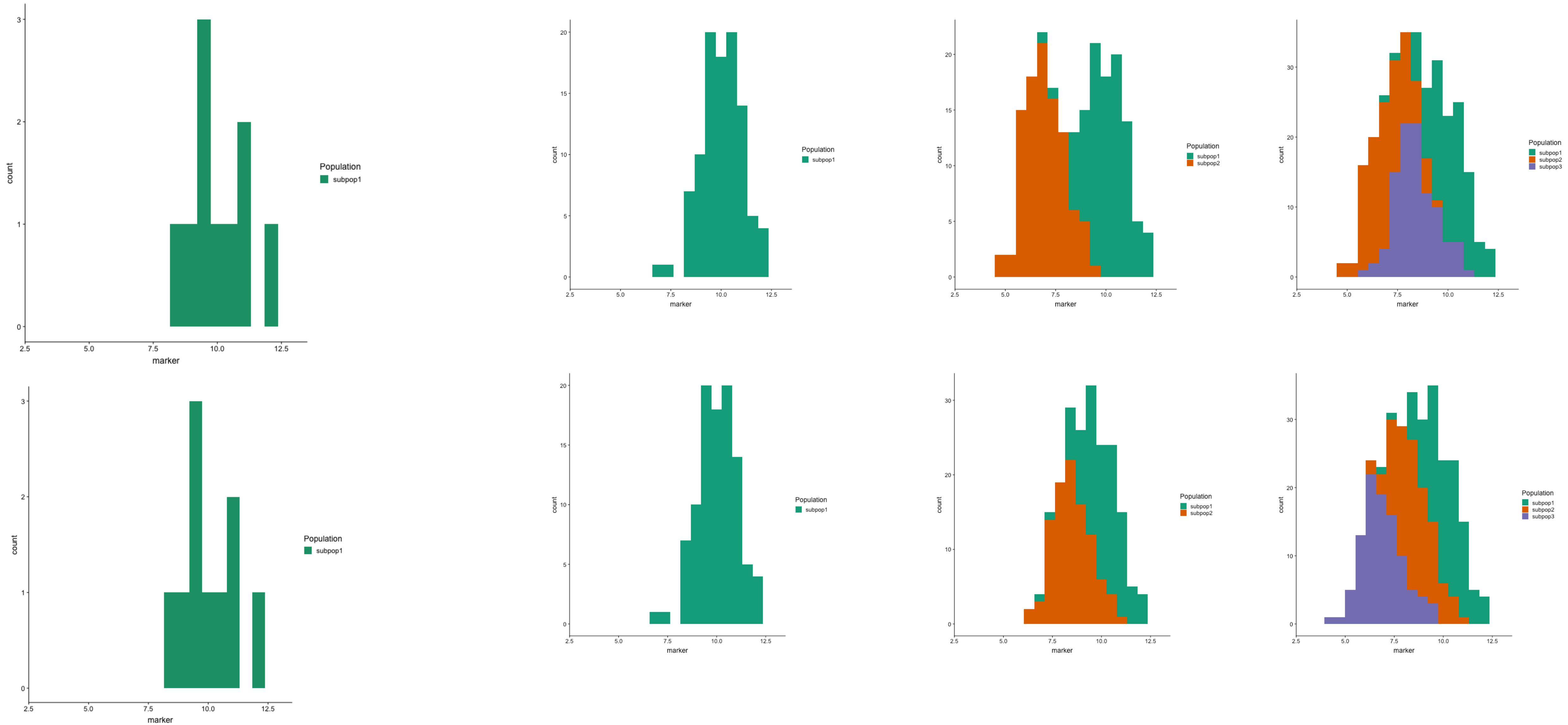
Sources and Impact (some algorithms can magnify bias)

- **Male/Female**
- **Caucasian/non-Caucasian (minority groups, ethnicity, geographic)**
- **Human/non-Human (we can generally cure mice)**
- **Adult/Children**
- **Sequencing/Genomics (Panel, Exome, Whole Genome, RNA) – Proteomics**
- **Imaging (yes/no, quantitative/qualitative), EHRs (controlled vocabulary)**
- **Socioeconomic Status (sensors, wearables, access to diagnostics and treatment, …)**
- **Environment and Culture**

Sampling (Bias) influences Machine Learning
Especially if it’s automated and not repeatedly assessed and questioned

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



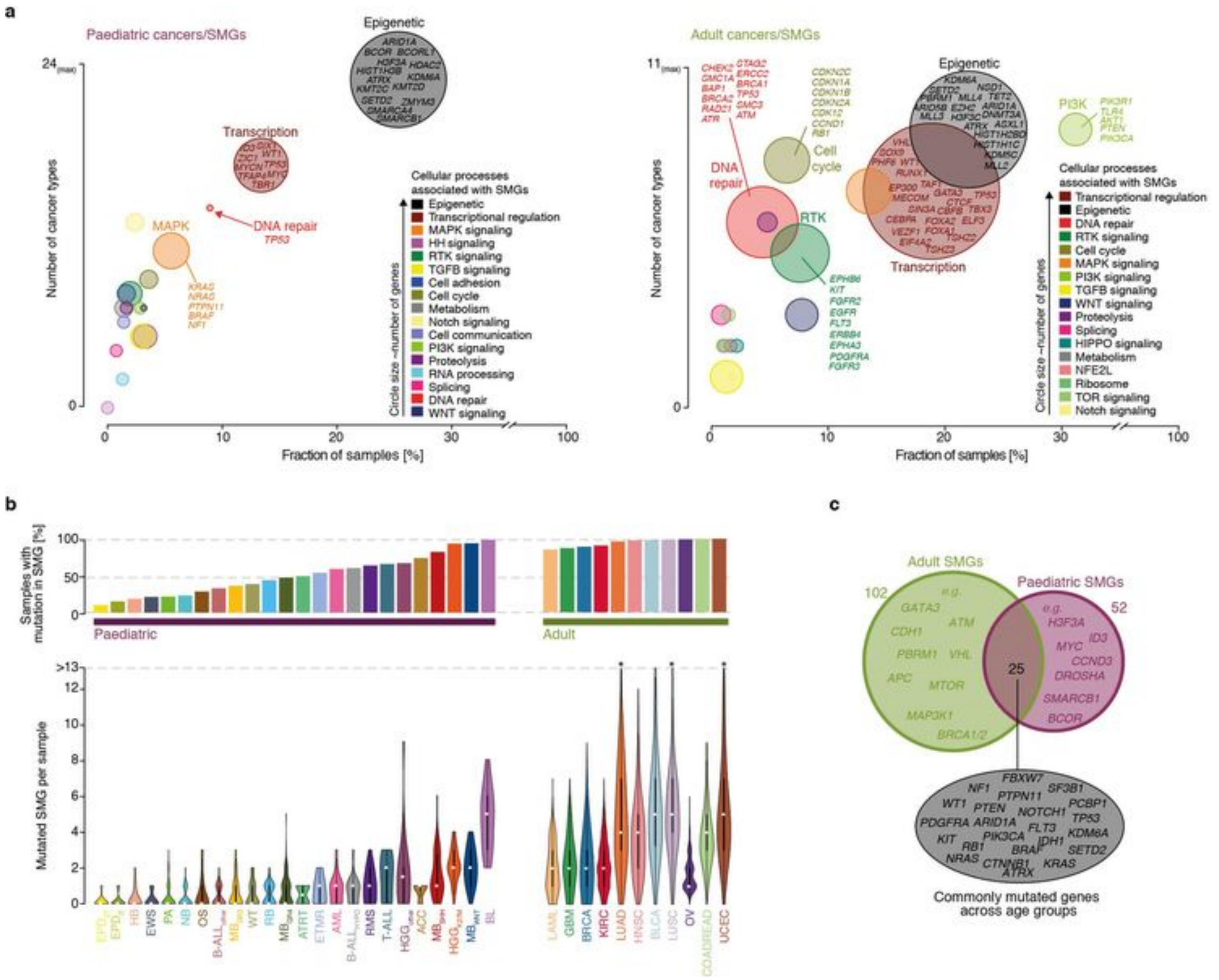
Adult Cancer is Different from Pediatric Cancers

Much of our current knowledge and treatments have targeted Adults

The landscape of genomic alterations across childhood cancers

Susanne N. Gröbner
, Barbara C. Worst
[…]
Stefan M. Pfister

Nature **volume 555**, pages 321–327 (15 March 2018)



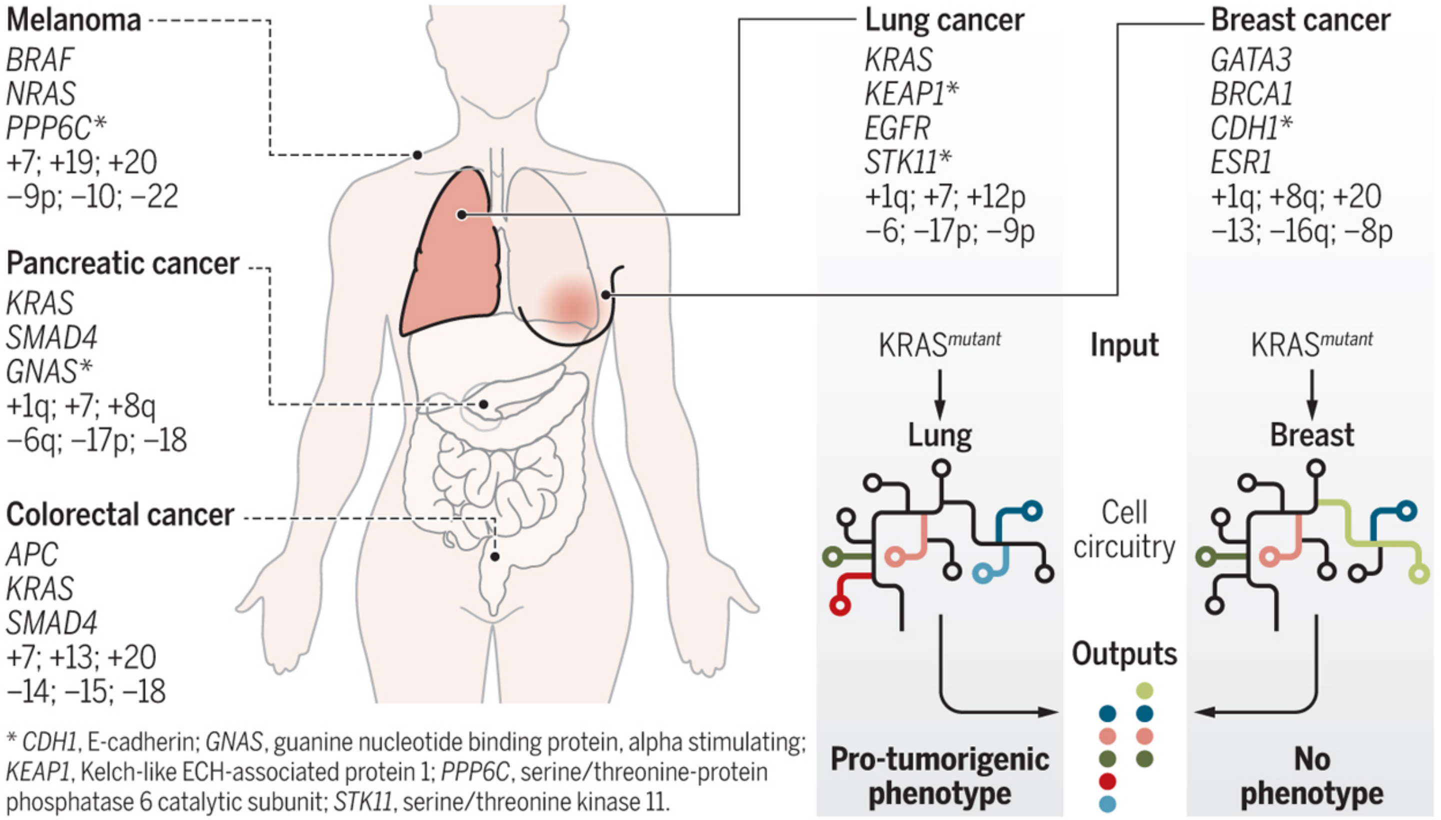
Tissue-specific genetic alterations and differential responses

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute

Tissue-specific genetic alterations and differential responses

Different human cancers contain a subset of recurring cancer driver gene mutations and chromosome copy number alterations that are specific for, or enriched in, that tumor type. The underlying tissue-specific epigenetic architecture may differentially determine the responsiveness to oncogenic signals and thus the propensity to acquire alterations that lead to cancer.



* *CDH1*, E-cadherin; *GNAS*, guanine nucleotide binding protein, alpha stimulating; *KEAP1*, Kelch-like ECH-associated protein 1; *PPP6C*, serine/threonine-protein phosphatase 6 catalytic subunit; *STK11*, serine/threonine kinase 11.

Tissue-specificity in cancer: The rule, not the exception

Kevin M. Haigis, Karen Cichowski, and Stephen J. Elledge

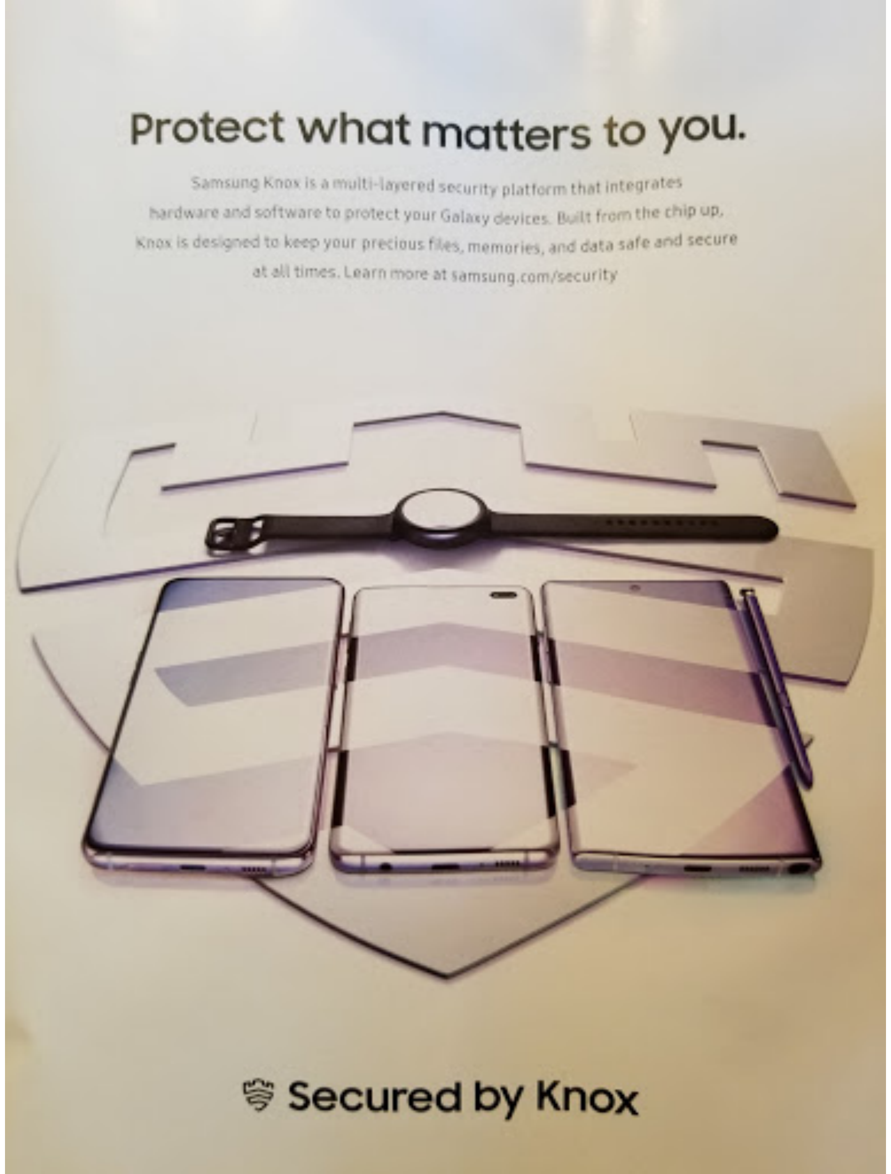
Science
Volume 363(6432):1150-1151
March 15, 2019

Published by AAAS

Science AAAS

Security and Privacy are …

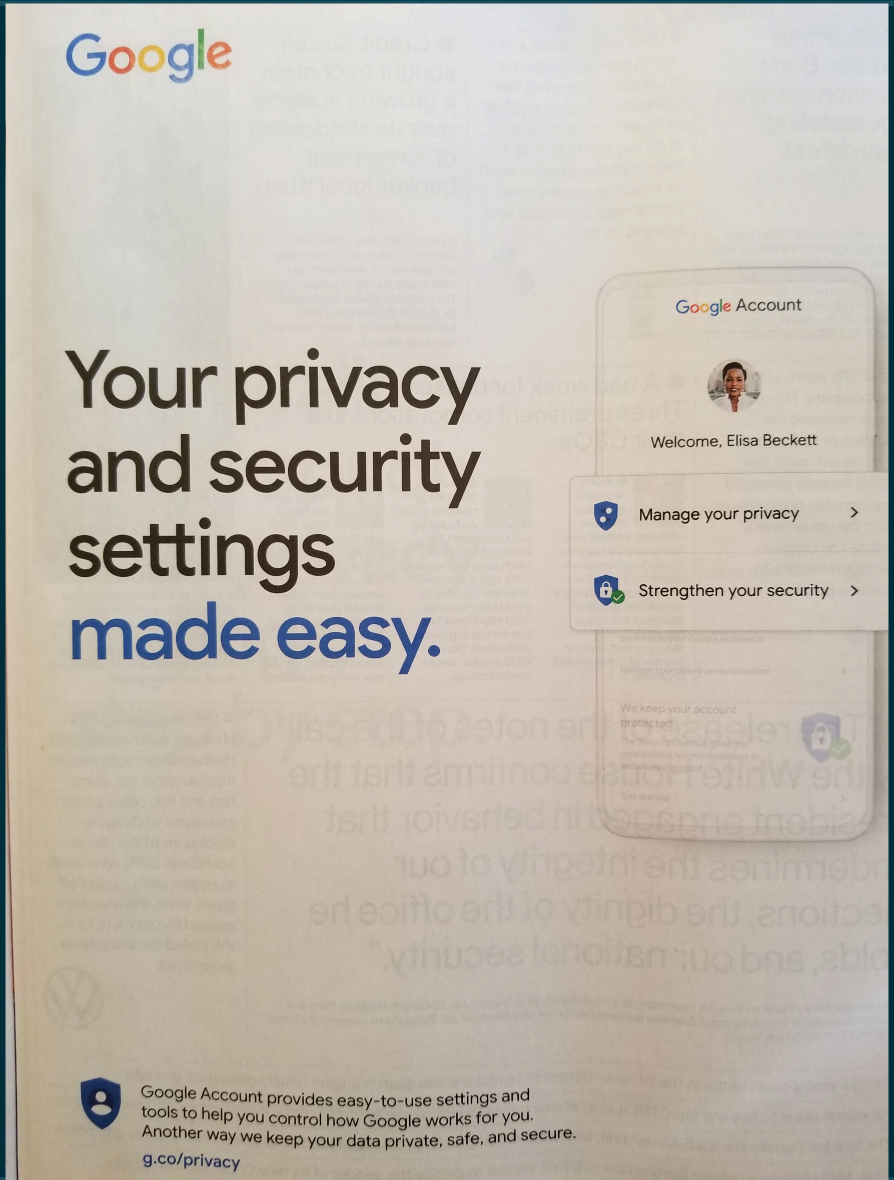
- Misunderstood by many (researchers, general public, …)
- Risks vs. Benefits
- Often Assumed
- Becoming “commoditized” (“get what you pay for”)
- Increasingly difficult
- Breaches can be expensive
- Informed by cultural norms
- Different to many people



Samsung Knox is a multi-layered security platform that integrates hardware and software to protect your Galaxy devices. Built from the chip up, Knox is designed to keep your precious files, memories, and data safe and secure at all times. Learn more at samsung.com/security

Who controls your data?
Who profits from your data?

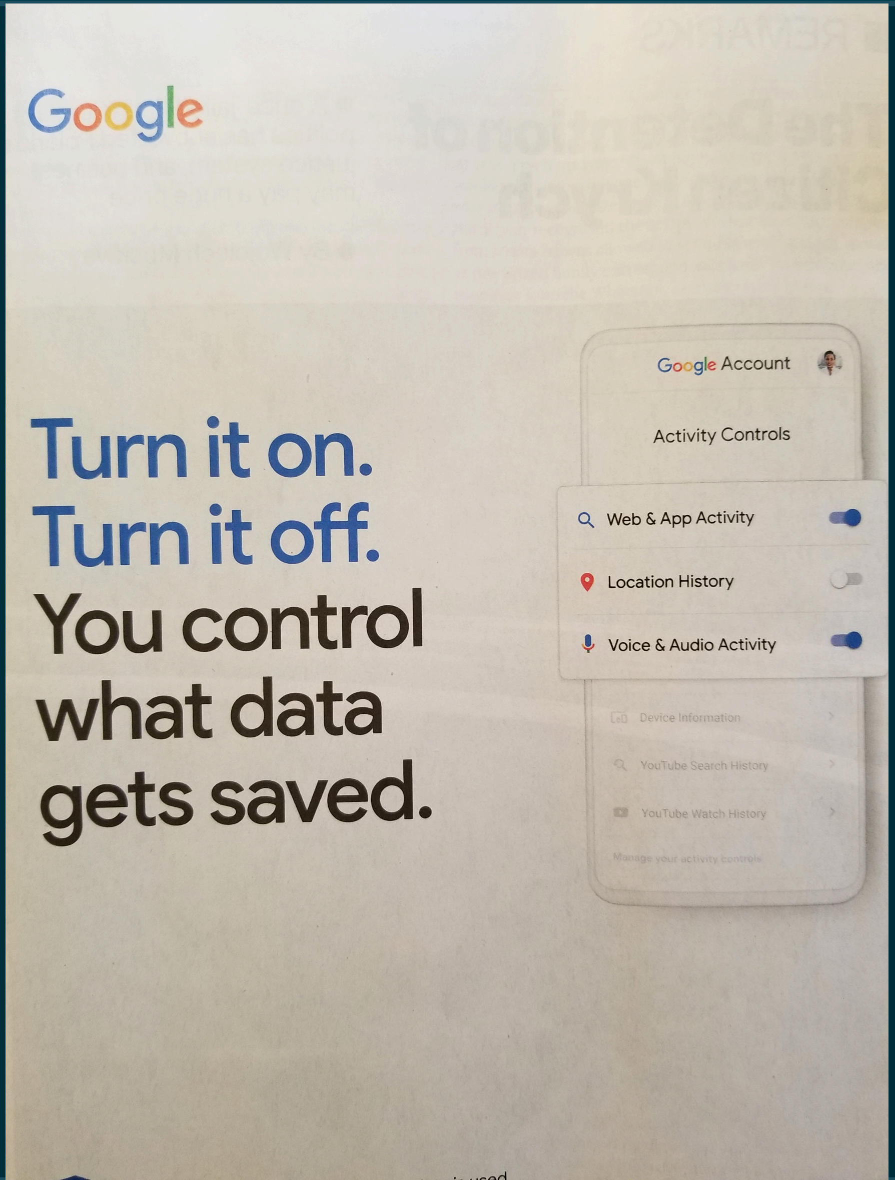
Healthcare, generally, has not kept up with technology.



Google

Your privacy and security settings made easy.

Google Account provides easy-to-use settings and tools to help you control how Google works for you. Another way we keep your data private, safe, and secure.
g.co/privacy



Google

Turn it on. Turn it off. You control what data gets saved.

Drivers for Data Sharing

- Scale of data production has grown dramatically; large scale data are generated, processed and analyzed at a significant cost.

- Open access allows published claims to be verified

- Large-scale data can be used to address scientific issues distinct from the original research problem

- Current state of IT allows data to be transferred, stored, analyzed and disseminated at a much larger scale

- Data sharing facilitates the development of novel research methods and tools

NIH Genomic Data Sharing (GDS) Policy:
Overarching Principles

- Data sharing promotes maximum public benefit from federally funded genomics research.

- Genomic data to be shared with timely data release through broadly accessible and open or, if more appropriate, controlled access data repositories

- Systems to ensure human subject protection and oversight of research conduct, data quality, data management, data sharing, and data use are critical to effective data sharing policies

GDS Policy Expectations

- **Large scale data, both human and non-human, are expected to be shared, irrespective of funding level or mechanism**
  - Examples of large-scale data include: whole genome/exome sequencing, transcriptome, epigenome and single-nucleotide polymorphism array data

- **Metadata and annotations necessary to interpret study and replicate results are to be shared**

- **Data is to be submitted to a NIH-designated repository**
  - Examples: dbGAP, dbSNP, dbVar, SRA, GEO, GenBank, ClinVar, Genomic Data Commons (GDC), ICGC, etc.

Are your data accessible and in a format that can be readily shared upon request?

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute

Contents of a manuscript must be available to readers, journals, institutions, if requested. (raw data, data-sets, reagents, etc)

“May I have the data?”

Balancing Risks vs Benefits / Rewards

Integration of data is key to both

“Reidentification of Anonymized Data” from Public Sources

Predicting Social Security numbers from public data

Alessandro Acquisti and Ralph Gross
PNAS July 7, 2009 106 (27) 10975-10980;
https://doi.org/10.1073/pnas.0904891106

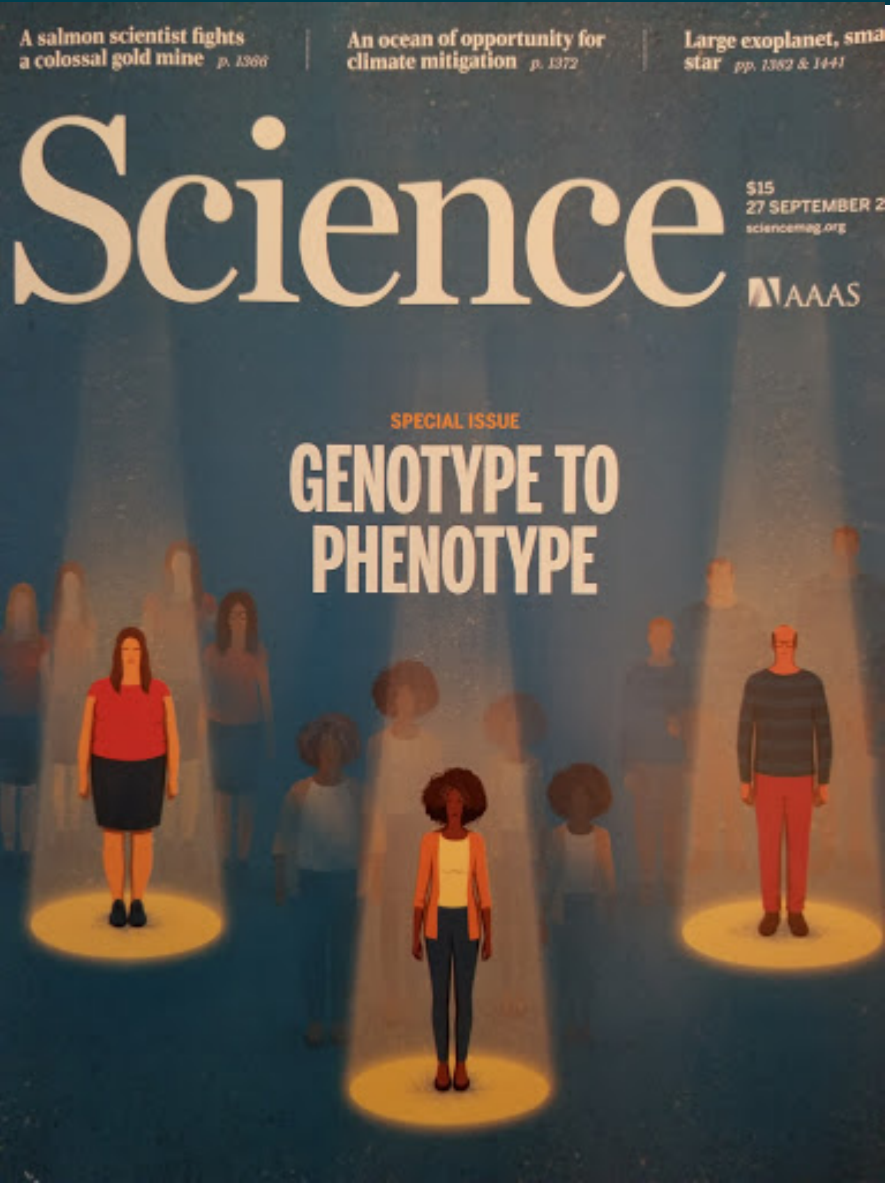
Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites. Our results highlight the **unexpected privacy consequences of the complex interactions among multiple data sources in modern information economies** and quantify privacy risks associated with information revelation in public forums.

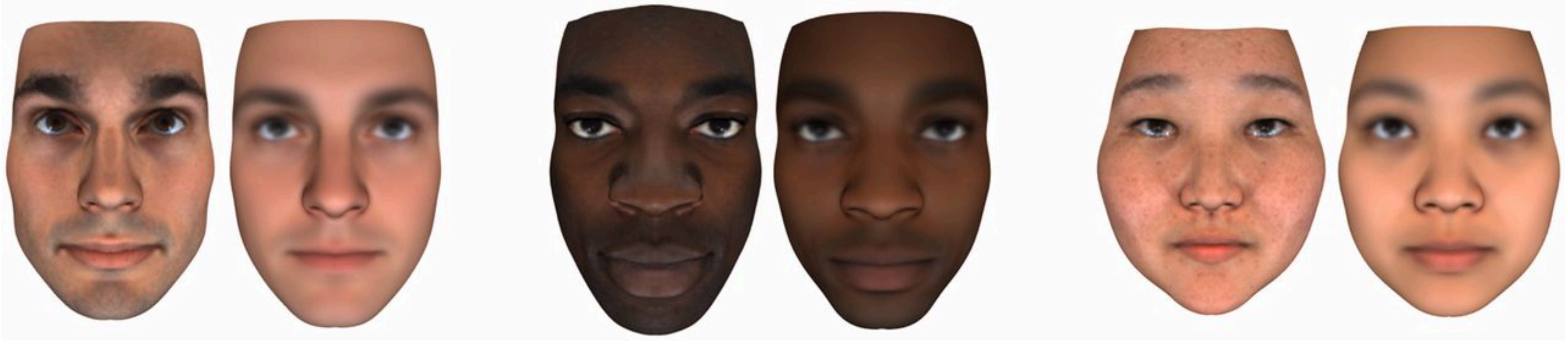
Identifying Personal Genomes by Surname Inference

Melissa Gymrek, Amy L. McGuire, David Golam, Eran Halperin, Yaniv Erlich
***Science* 18 Jan 2013:**
Vol. 339, Issue 6117, pp. 321-324
DOI: 10.1126/science.1229566

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it **entirely relies on free, publicly accessible Internet resources**. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Genome => Phenotype





Examples of real (Left) and predicted (Right) faces.

Christoph Lippert et al. PNAS
doi:10.1073/pnas.1711125114
2017

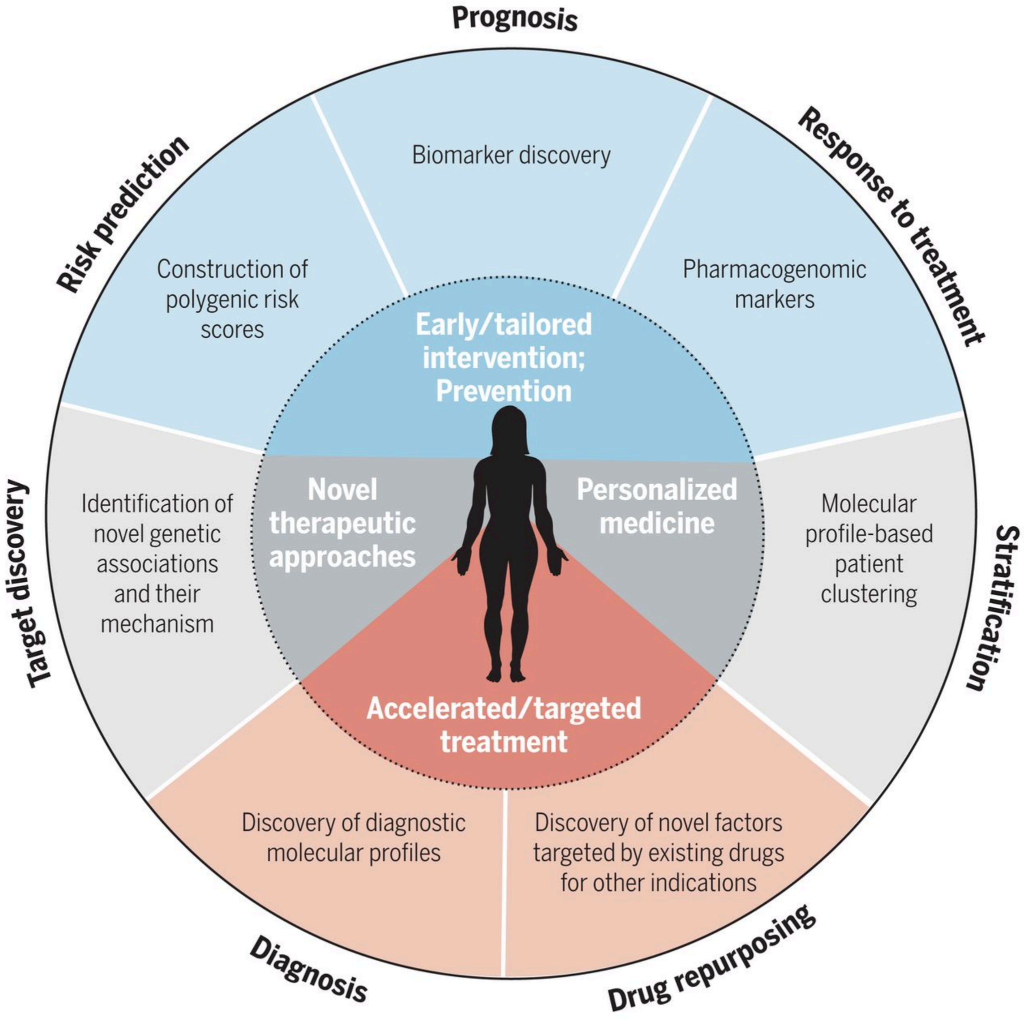
Significance

By associating deidentified genomic data with phenotypic measurements of the contributor, this work challenges current conceptions of genomic privacy. It has **significant ethical and legal implications on personal privacy, the adequacy of informed consent, the viability and value of deidentification of data,** the potential for police profiling, and more. We invite commentary and deliberation on the implications of these findings for research in genomics, investigatory practices, and the broader legal and ethical implications for society. Although some scholars and commentators have addressed the implications of DNA phenotyping, this work suggests that a deeper analysis is warranted.

The translational potential of complex disease genomics.

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute


Prognosis
Biomarker discovery
Response to treatment
Pharmacogenomic markers
Risk prediction
Construction of polygenic risk scores
Early/tailored intervention; Prevention
Novel therapeutic approaches
Personalized medicine
Target discovery
Identification of novel genetic associations and their mechanism
Stratification
Molecular profile-based patient clustering
Accelerated/targeted treatment
Discovery of diagnostic molecular profiles
Discovery of novel factors targeted by existing drugs for other indications
Diagnosis
Drug repurposing

Eleftheria Zeggini et al. Science 2019;365:1409-1413

Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works

Science AAAS

Use Cases - Examples

Cancers / Rare Cancers
Precision/Personalized Medicine (N of 1)
Rare Diseases
Potential of AI

Rare Diseases
Moving toward genomically defined disease

- **National Center for Advancing Translational Science (NCATS/NIH)**
  - Rare Disease - In the United States, a rare disease is defined as a condition that affects fewer than 200,000 people. This definition was created by Congress in the Orphan Drug Act of 1983. In the European Union, a disease is defined as rare when it affects fewer than 1 in 2,000 people.
  - ~7000 Rare Diseases Defined (prevalence of risk often unknown)
  - Rare Diseases, by definition, have few data points (Need to aggregate/integrate)
  - Potentially Identifiable Population (often children and significant consequences)
  - Symptoms often fall on a Spectrum
  - Diseases often share phenotypes (and, presumably, genotypes)
  - Farber’s Disease (Case 1) ~80+ reported cases worldwide
  - Proper and rapid diagnosis is currently an issue. Benefits to understanding disease at molecular level
  - Phenyketonuria – (low or defective PAH gene -> decreased metabolism of phenylalanine)
    - Significant effects can be mitigated by early intervention and strict diet. Significant developmental issues otherwise.

Rare Cancers
MyPART

- **MyPART info**
  - MyPART is the My Pediatric and Adult Rare Tumor network. It is a group of scientists, patients, family members, advocates, and healthcare providers who want to help find treatments for rare cancers. We are working on childhood, teen, and young adult solid rare tumors that have no cures.
  - ask patients, their family members, and healthcare providers about how the rare tumor affects patients’ lives.
  - collect samples like blood, saliva, and biopsy tissue from people with rare tumors to study how rare tumors grow and how we could treat them.
  - share data from rare cancer samples with scientists around the world.
  - hold workshops with patients, advocates, doctors, and scientists to talk about how to improve patients’ lives and find new treatments.
  - build new ways of testing new treatments.
  - use what we learn to design new clinical trials for rare cancers.
  - teach the public about how we are trying to find new treatments.
  - share research results with individual patients.

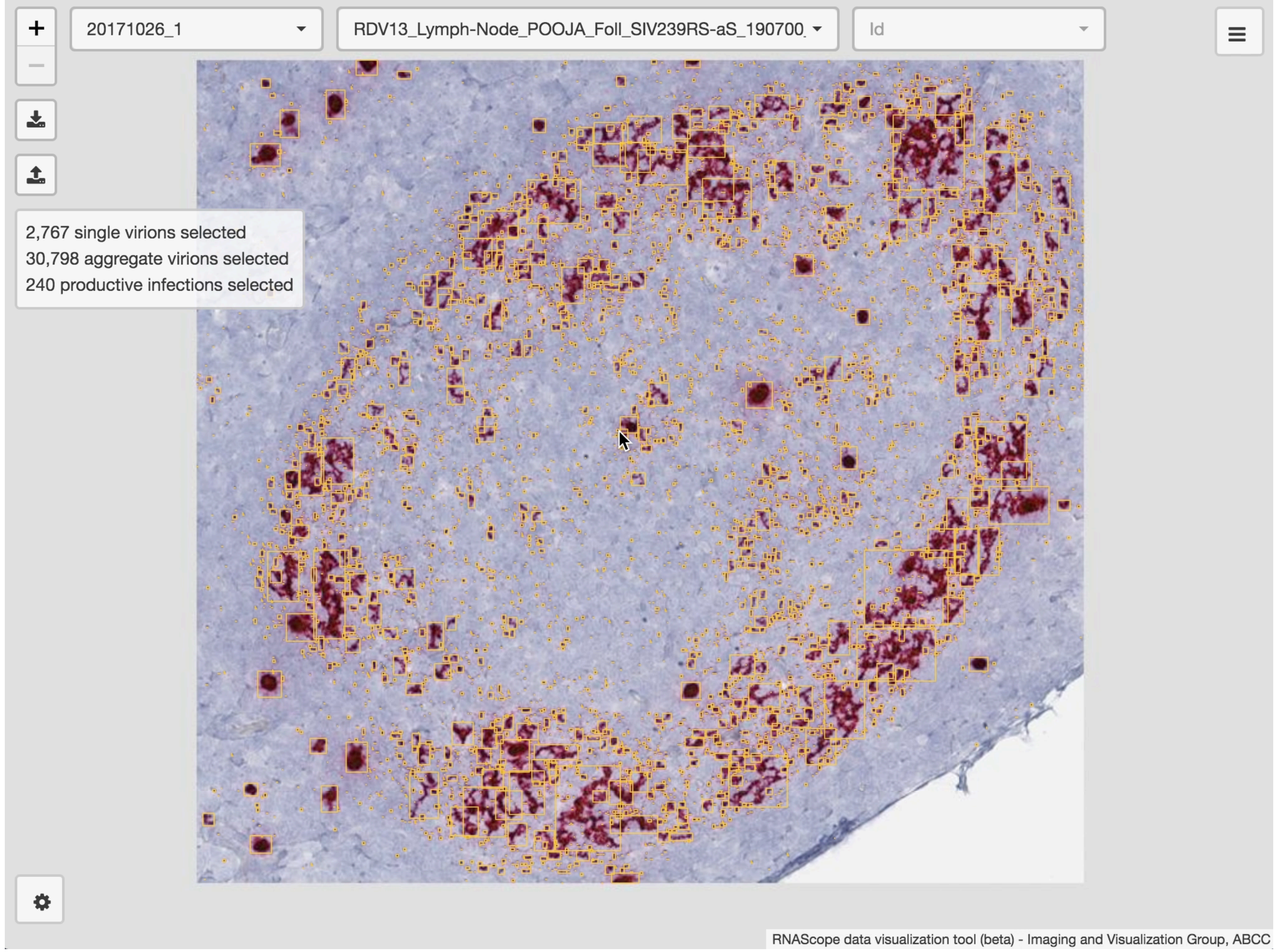
Building Tools for Pathologists to Explore, Understand, and Interpret ML/AI Results in the Context of their Data

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute

Dr. Christian Suloway

- **Assist the pathologist by allowing them to focus attention on the important decision points. Keep the human in the loop.**
- **Automate the tedious.**
  - Like Counting Infected Cells
  - Like Sorting by Feature
- **Tools that help the scientists (in this case pathologists) understand the ML/DL results in the context of their data are critical.**
- **Acceptance and usage depend on demonstrating the value and keeping the human in the loop on actionable recommendations.**
- **Models and workflows must be as transparent as possible (a challenge with ML/DL systems).**
- **Don't move the data – facilitate access.**



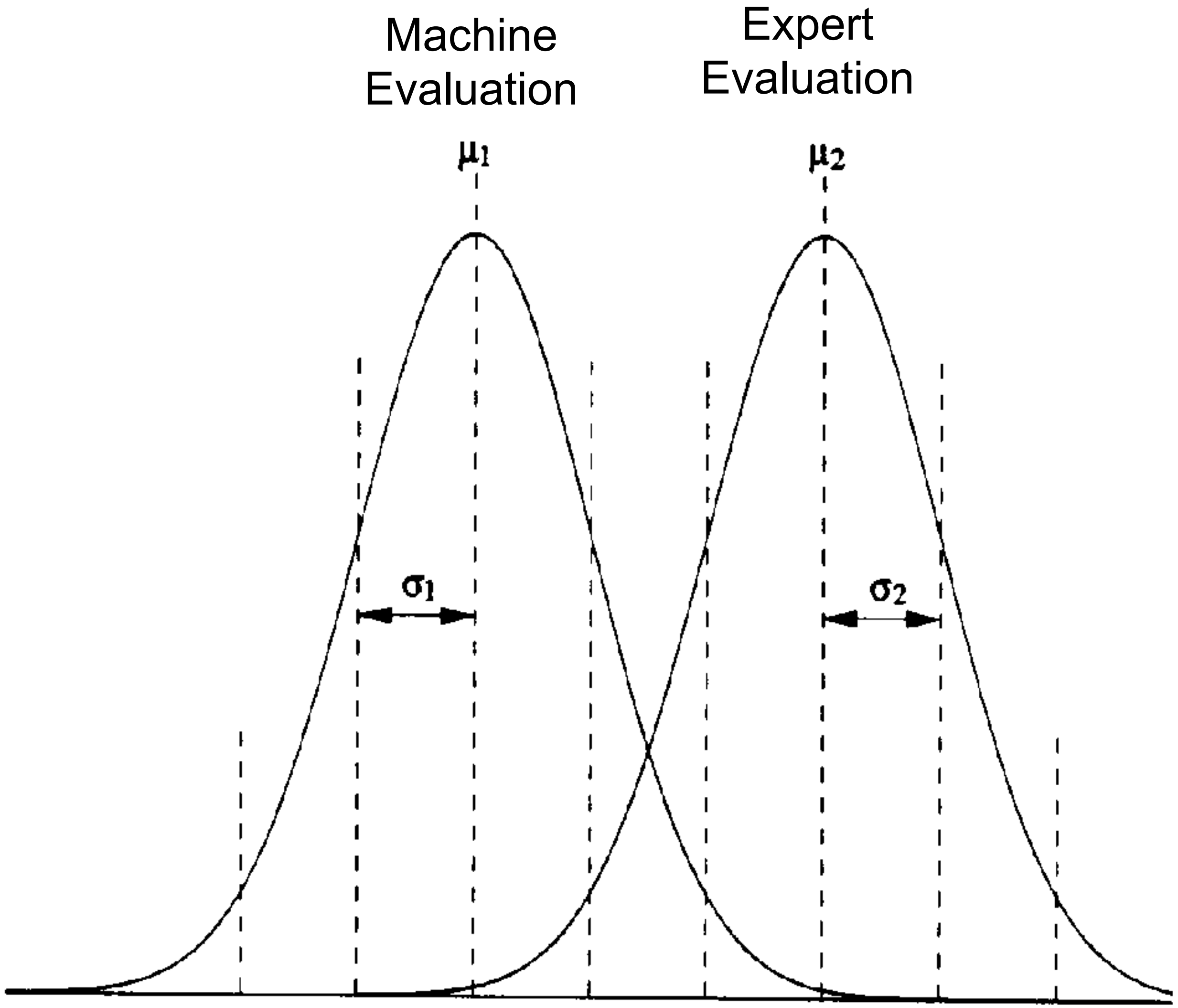
2,767 single virions selected
30,798 aggregate virions selected
240 productive infections selected

RNAScope data visualization tool (beta) - Imaging and Visualization Group, ABCC

Video showing quantification results using the IVG developed web application

What is the “truth”?
How can we tell? Or – How do we define?

- **If we could evaluate/assess images/data using multiple human evaluators (Expert Evaluation) with different conditions and different machine learning models (Machine Evaluation), then we could compare the distributions and estimate the probability that they represent the same distributions.**
- **HPC/HPDA/Advanced Computing can enable this type of evaluation. – if we have the data**



SCIENCE FUTURE

DATA DRIVING DISCOVERY

Liberating, Sharing, Aggregating Data

Technology Enabling Science