

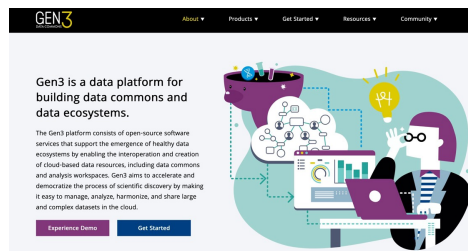
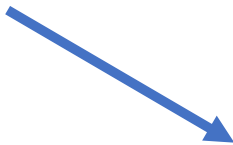
The Long Tail of Data Curation and its Impact on the Value of Data

Robert L. Grossman
University of Chicago

June 8, 2022

1. The Problem and some Background

Data
users



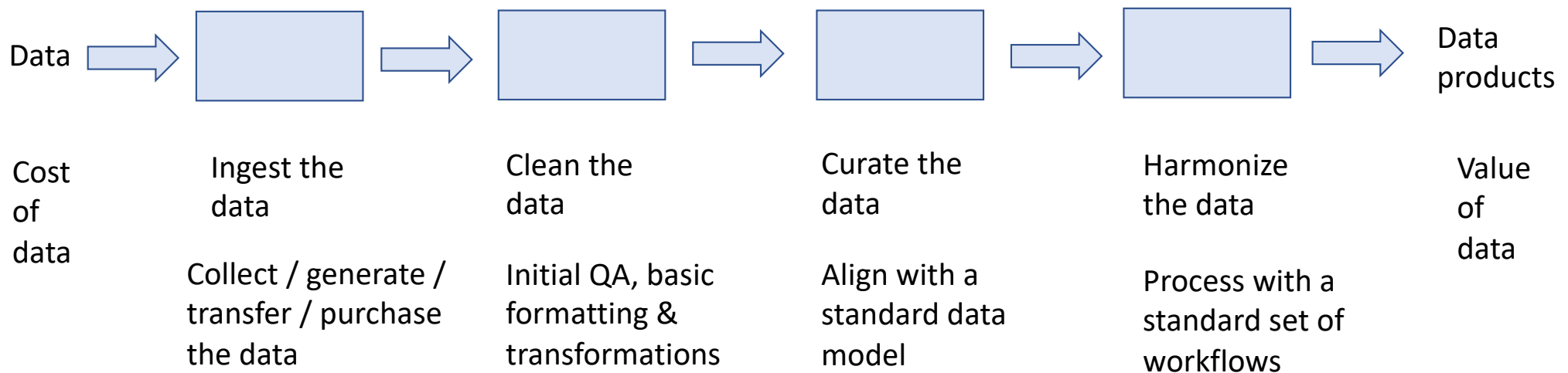
Data platform

Platform sponsor

Data
contributors



Typical Model for Data Cleaning / Curation / Processing

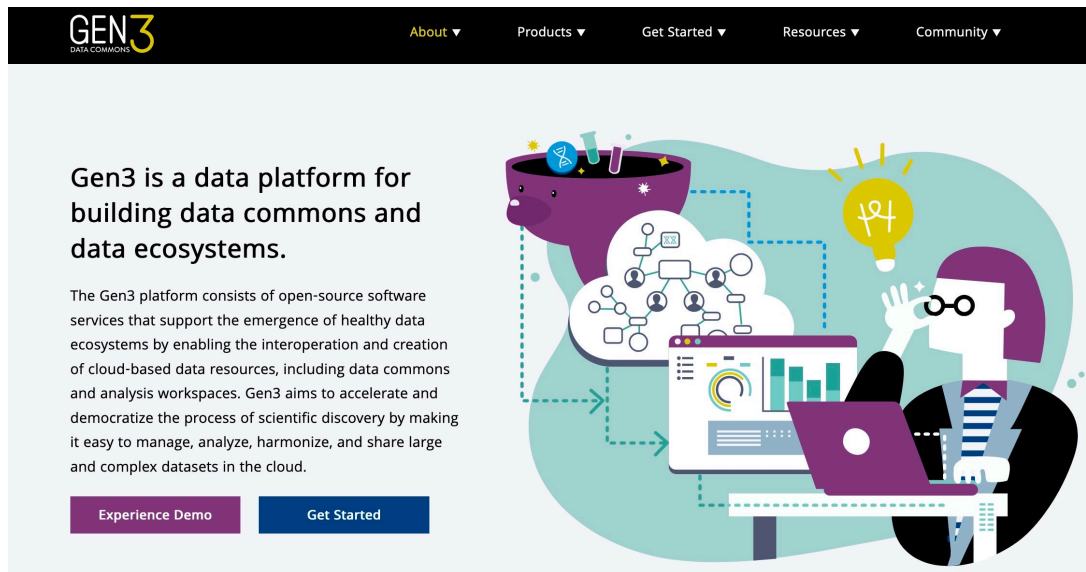


- For simplicity, in this talk, we call this process **data curation**.
- For many projects, data curation is one of the largest costs for data platforms

Questions

- What does data curation look like in the wild?
- How do we model data curation?
- How does data curation relate to data value?
- How do we lower or shift the costs of data curation?

Our Experience



Gen3.org

- We developed an open source data platform called Gen3.
- Gen3 includes the Data Ingestion, Integration & Release Management (DIIRM) services for data curation.
- We operate over 20 Gen3 platforms.

Data curation is one of the important drivers of data value creation, but it's expensive, hard to value, and most stakeholders are reluctant to pay for it.

GEN3 Data Mesh

1,445,041 **43,254,328** **14.58 PB**
Total Subjects Total Files Total File Size

 National Institute of
Energy and
Infectious Diseases
AccessClinicalData@NIAID

2,096 Subjects
147 Attributes
7 Files
Total Size **3.11 MB**

 **Veterans Affairs
Data Commons**

658,278 Subjects
1,606 Attributes
3 Files
Total Size **1054.46 GB**

 **NATIONAL CANCER INSTITUTE**
Cancer Research Data Commons

83,709 Subjects
622 Attributes
5,226,838 Files
Total Size **3.71 PB**

NIH
HEAL
INITIATIVE

153 Attributes
85 Files
Total Size **14.88 GB**

 **Justice Community Opioid
Innovation Network (JCOIN)**

237 Subjects
871 Attributes
368 Files
Total Size **1.27 GB**

 **CANINE**
Data Commons

1,499 Subjects
1,008 Attributes
3,802 Files
Total Size **1.88 TB**

 **CHICAGOLAND COVID-19 COMMONS**

53,728 Subjects
1,462 Attributes
285,227 Files
Total Size **117.64 TB**

 **GenoMEL**
the Melanoma Genetics Consortium

1,390 Subjects
387 Attributes
6,555 Files
Total Size **31.6 TB**

 **BDGC**

107,418 Subjects
786 Attributes
15,950 Files
Total Size **3.81 TB**

GEN3

Open Access Data Commons

1,366 Subjects
1,452 Attributes
1,584 Files
Total Size **13.77 TB**

 **ACCOUNT**

1,516 Subjects
985 Attributes
5,661 Files
Total Size **7.77 TB**

 **BioData
CATALYST**

240,460 Subjects
745 Attributes
645,520 Files
Total Size **3.57 PB**

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

4,839 Subjects
888 Attributes
34,946 Files
Total Size **32.7 TB**

 **Veterans
Precision Oncology
Data Commons**

163,695 Subjects
1,606 Attributes
352,785 Files
Total Size **2.18 TB**



The AnVIL
26,636 Subjects
551 Attributes
187,134 Files
Total Size **502.27 TB**

 **Kids First**
Pediatric Research Program
Data Resource Center

21,833 Subjects
776 Attributes
687,023 Files
Total Size **6.49 PB**

 **MIDRC**
MEDICAL IMAGING AND DATA RESOURCE CENTER

7,627 Subjects
510 Attributes
4,529,435 Files
Total Size **1.74 TB**

 **Environmental
Data Commons**

265 Attributes
31,271,405 Files
Total Size **92.17 TB**

Data Commons

Data commons are software platforms that co-locate: 1) **well-curated data**, 2) cloud-based computing infrastructure, and 3) commonly used software applications, tools and services to create a resource for managing, analyzing, integrating and sharing data with a research community.

Data meshes integrate multiple data commons and other data resources.

Data
users

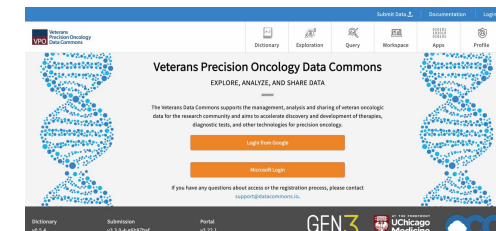
- Data curation
- Data harmonization
- Security & compliance

Governance

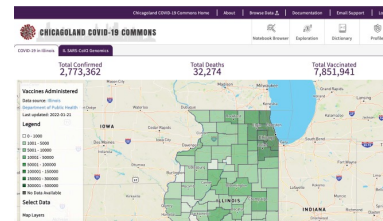
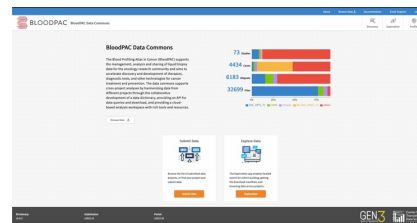
- Data
- Commons
- Consortium

Data
contributors

- Standard software apps
- Custom software development



Platform (open source)



Supported by an organization (not-for-profit, government, ...)

Robert L. Grossman, (2019). Data lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data. Trends in Genetics, 35(3), 223-234.

- Legal agreements
- Liability management

One of the Dirty Secrets of Data Curation

- Everyone talks about using automation, AI and machine learning for data curation, but behind the scenes there are usually large teams of human data curators doing the work.

2. An Example of Large-Scale Data Curation

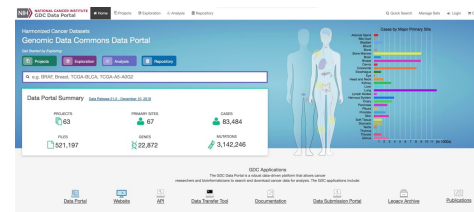
Data Curation at Different Scales



Small scale –
custom work by
individuals &
small projects

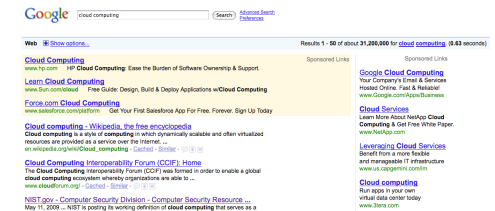
\$1,000's

Mid scale – scripts
and/or commercial
software with
shared staff



Large scale – dedicated
systems, software and
staff; continuous process
improvement &
investment; maintained
over years

\$1,000,000's

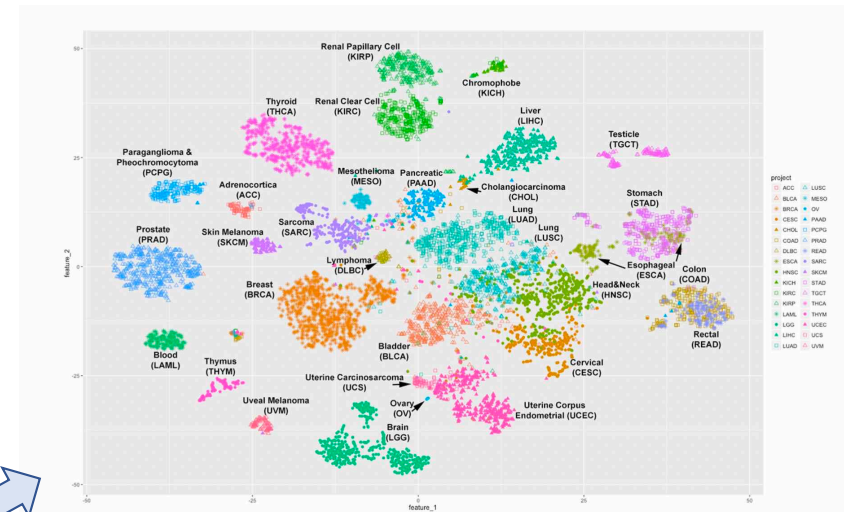
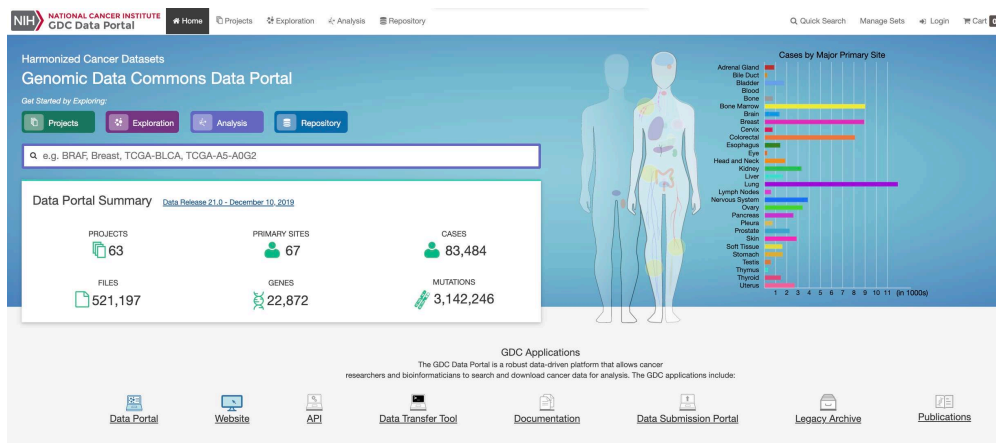


Very large scale –
hyperscalers with
dedicated systems
and staff and deep
experience

\$1,000,000,000's

Ex. 1: NCI Genomic Data Commons (GDC)

- 2012: Prototype
- 2014: GDC project starts
- 2016: Initial **data curation & analysis** finishes & GDC launches
- 2021: c. 60,000 users / month; c. 2 PB of data accessed/ month



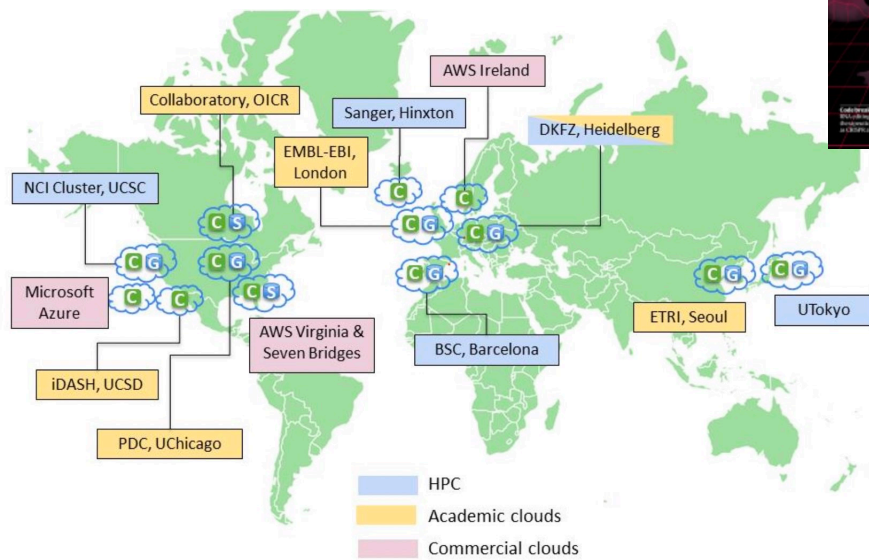
CTDS pan-cancer molecular
subtyping using GDC data

Zhang, Z., Hernandez, K., Savage, J., Li, S., Miller, D., Agrawal, S., ... & Grossman, R. L. (2021). Uniform genomic data analysis in the NCI Genomic Data Commons. *Nature Communications* 2021; 12(1), pp 1-11.

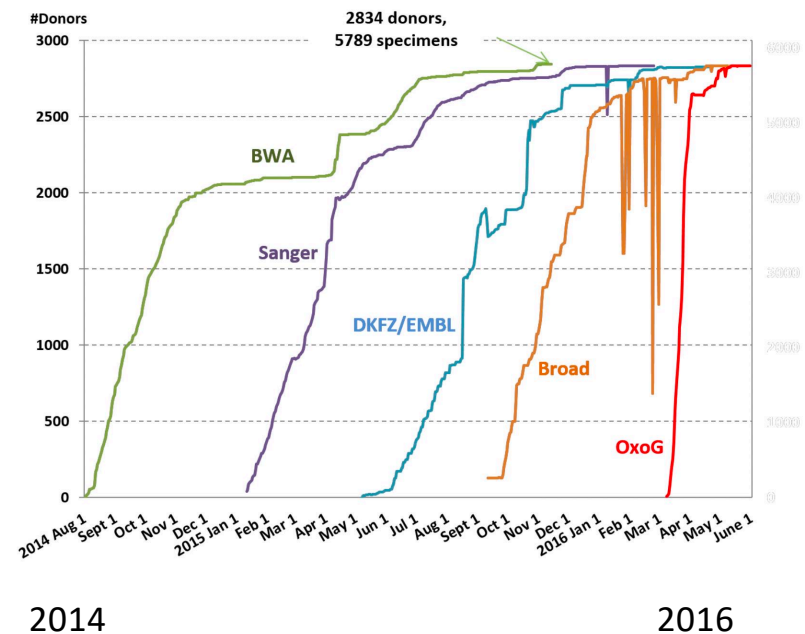
Heath AP, Ferretti V, ... Grossman RL, The NCI Genomic Data Commons, *Nature Genetics* 2021 Mar; 53(3), pp 257-262.

Ex 1 (cont'd): TCGA-ICGC Pan-Cancer Analysis Variant Calling (2014—2016) – a world-wide federated computation.

Federated Machine Learning



Nature special Issue: 6 Feb 2020



ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Concosortium, Pan-cancer analysis of whole genomes, Nature 578, no. 7793 (2020): pages 82-93. PMID: 32025007 PMCID: PMC7025898 DOI: 10.1038/s41586-020-1969-6

Introduction

GDC Data Model

Data Security

File Format: MAF

File Format: VCF

Bioinformatics Pipeline: DNA-Seq
Analysis

Bioinformatics Pipeline: mRNA
Analysis

Bioinformatics Pipeline: miRNA
Analysis

Bioinformatics Pipeline: Copy
Number Variation Analysis

Bioinformatics Pipeline: ...

Data Release Notes

Version	Date
v33.1	May 31, 2022
v33.0	May 3, 2022
v32.0	March 29, 2022
v31.0	October 29, 2021
v30.0	September 23, 2021
v29.0	March 31, 2021
v28.0	February 2, 2021
v27.0-fix	November 9, 2020
v27.0	October 29, 2020
v26.0	September 8, 2020
v25.0	July 22, 2020

Source: GDC Documentation, retrieved from
https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/ on June 1, 2022.


GDC Data Release 32 Updated GENCODE From v22 to v36



[Human](#) [Mouse](#) [How to access data](#) [FAQ](#) [Documentation](#) [About us](#)


HUMAN

GENCODE 40 (11.04.22)



MOUSE

GENCODE M29 (11.04.22)

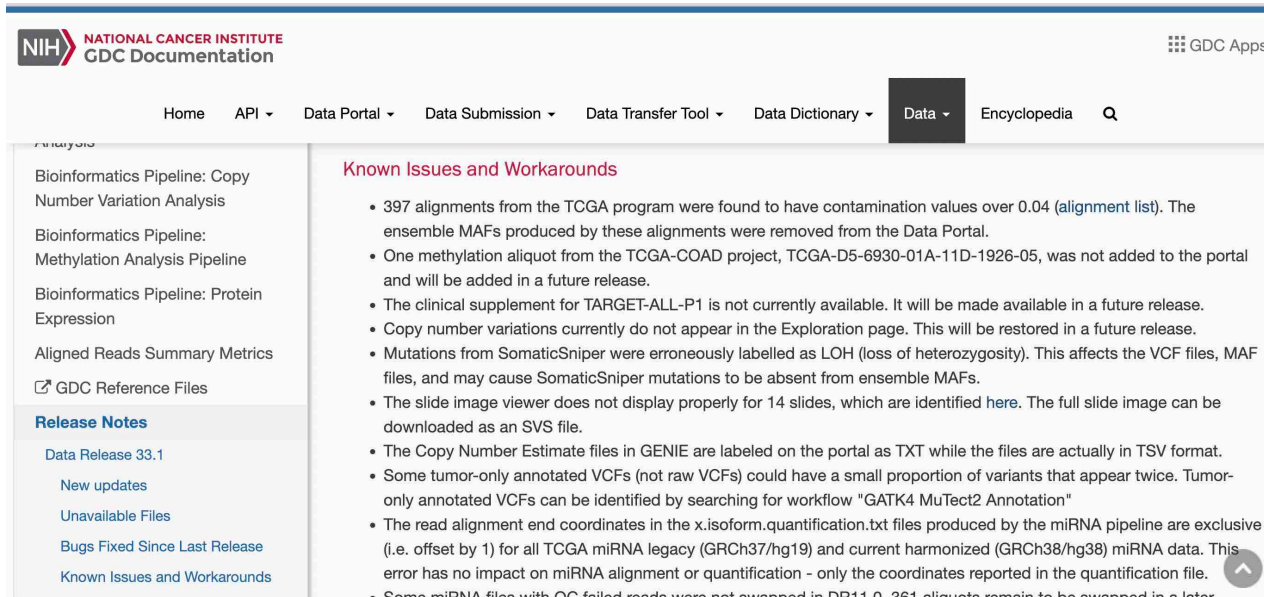


- The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence.

The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation.

Source: www.gencodegenes.org

GDC Release 32



- About 700 TB new data
- 158 Data Sets across 70 GDC projects
- About 107,000 data objects were released
- Six data types
 - New annotated variants
 - RNA-Seq
 - Whole Genome Seq
 - Whole Exome Seq
 - Lifter Updates
 - Methylation Data
- GDC as a whole is about 4.3 PB of released data with about 17 million data objects and about 1 TB of structured data.

Source: GDC documentation, retrieved from https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-320 on June 1, 2022.

Sample GDC Data Release 32 Notes

- GDC contains ~ 10 million data objects
- Rel 32 contains ~ 100,000 data objects

- 397 alignments from the TCGA program were found to have contamination values over 0.04 (alignment list). The ensemble MAFs produced by these alignments were removed from the Data Portal.
- One methylation aliquot from the TCGA-COAD project, TCGA-D5-6930-01A-11D-1926-05, was not added to the portal and will be added in a future release.
- The slide image viewer does not display properly for 14 slides, which are identified here. The full slide image can be downloaded as an SVS file.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- The raw and annotated VarScan VCF files for aliquot TCGA-VR-A8ET-01A-11D-A403-09 are not available. These VCFs files will be replaced in a later release.

Source: GDC documentation, retrieved from https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-320 on June 1, 2022.

Ex 2. Data About COVID-19 & Pandemic Response Commons

RESEARCH

OPEN ACCESS

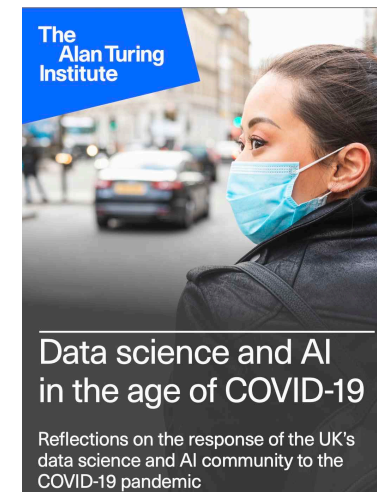
Check for updates

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,^{1,2} Ben Van Calster,^{2,3} Gary S Collins,^{4,5} Richard D Riley,⁶ Georg Heinze,⁷ Ewoud Schuit,^{8,9} Marc M J Bonten,^{8,10} Darren L Dahly,^{11,12} Johanna A Damen,^{8,9} Thomas P A Debray,^{8,9} Valentijn M T de Jong,^{8,9} Maarten De Vos,^{2,13} Paula Dhiman,^{4,5} Maria C Haller,^{7,14} Michael O Harhay,^{15,16} Liesbet Henckaerts,^{17,18} Pauline Heus,^{8,9} Michael Kammer,^{7,19} Nina Kreuzberger,²⁰ Anna Lohmann,²¹ Kim Luijken,²¹ Jie Ma,⁵

“... the single most consistent message across the workshops was the importance – and at times lack – of robust and timely data. Problems around data availability, access and standardization spanned the entire spectrum of data science activity during the pandemic. The message was clear: better data would enable a better response.”



Inken von Borzyskowski, et. al., editors, Data science and AI in the age of COVID-19 – report, Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic, Turing Institute, 2021.

MIT Technology Review

Featured Topics Newsletters Events Podcasts

Sign in Subscribe

ARTIFICIAL INTELLIGENCE

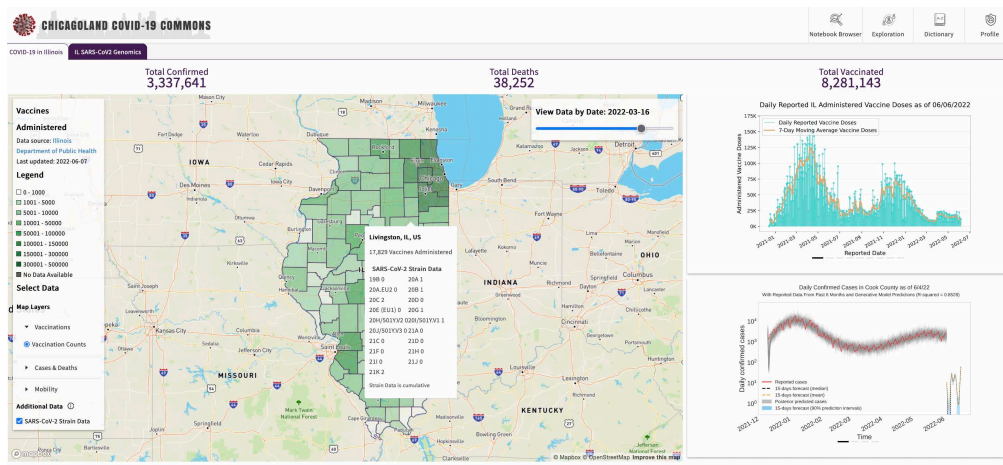
Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

Pandemic Response Commons



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES



Racial/Ethnic Disparities in the Observed COVID-19 Case Fatality Rate Among the U.S. Population

L. Philip Schumm, Mihai C. Giurcanu, Kenneth J. Locey, Jean Czerlinski Ortega, Zhenyu Zhang, Robert L. Grossman

doi: <https://doi.org/10.1101/2022.03.01.22271708>

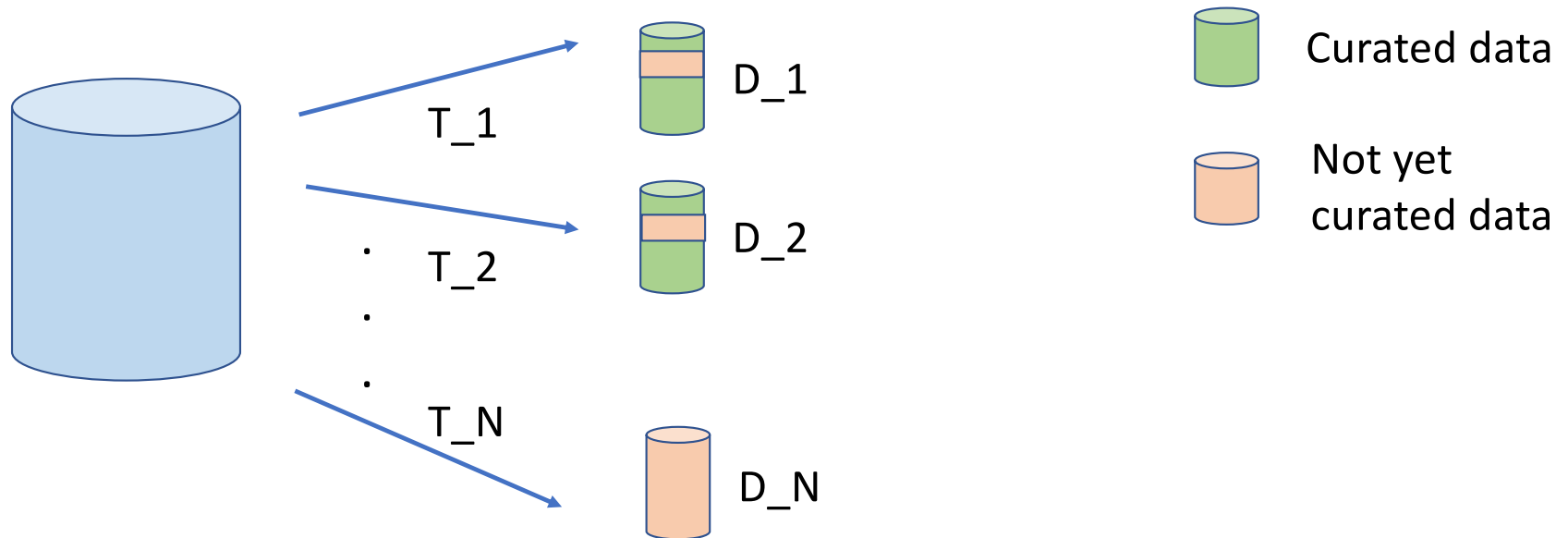
This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

Source: L. Philip Schumm, Mihai C. Giurcanu, Kenneth J. Locey, Jean Czerlinski Ortega, Zhenyu Zhang, Robert L. Grossman, Racial/Ethnic Disparities in the Observed COVID-19 Case Fatality Rate Among the U.S. Population, medRxiv 2022.03.01.22271708;

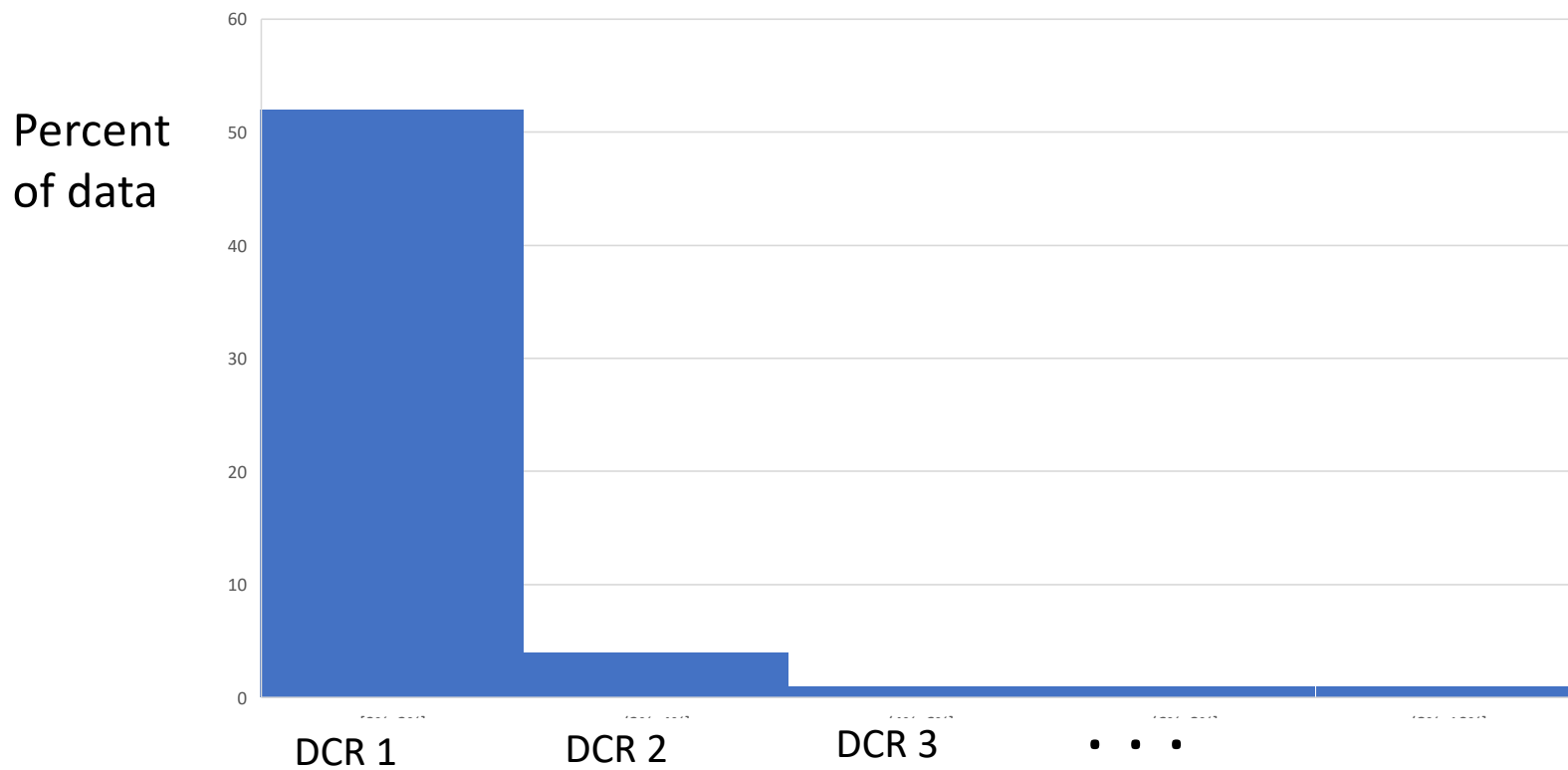
3. Data Curation Regimes: Towards a Model of the Long Tail of Data Curation

Introduce: Data Curation Regimes (DCR)

- We model data curation as consisting of data to be curated with the data divided into **data curation regimes**, D_1, D_2, \dots, D_N with separate rules or processes T_i used for each different data curation regime.



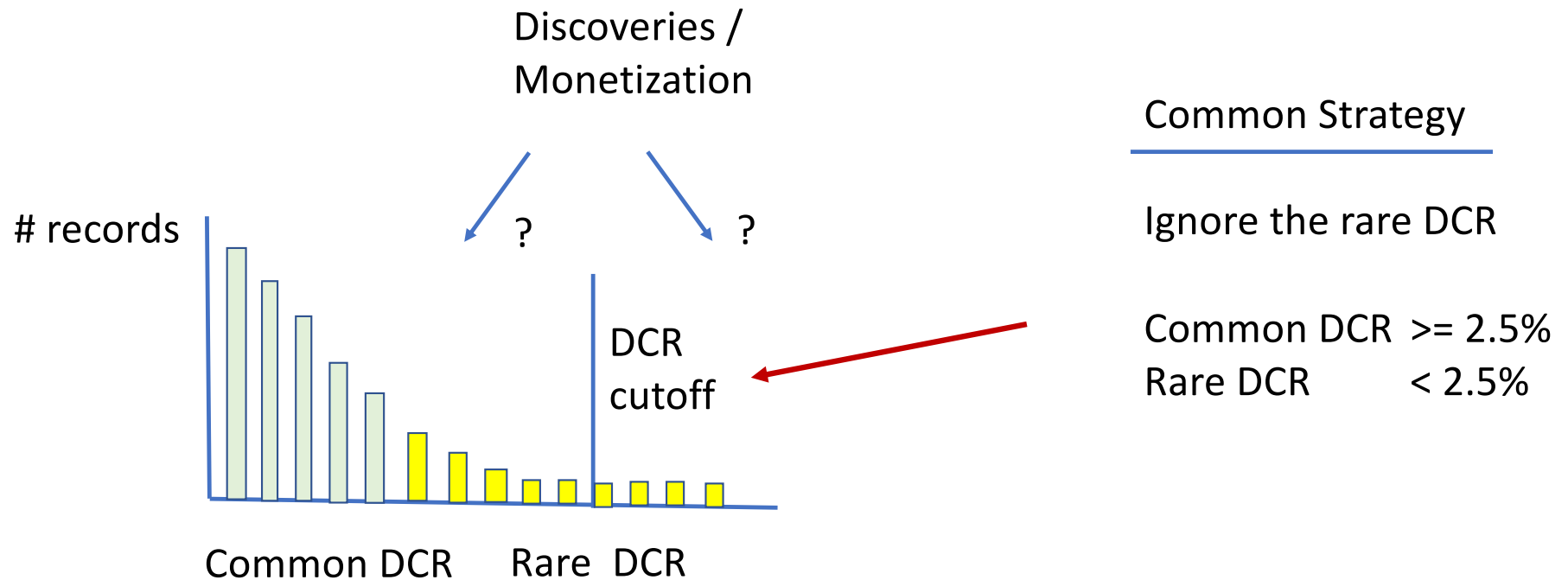
DCR Example 1. Percent of data by DCR



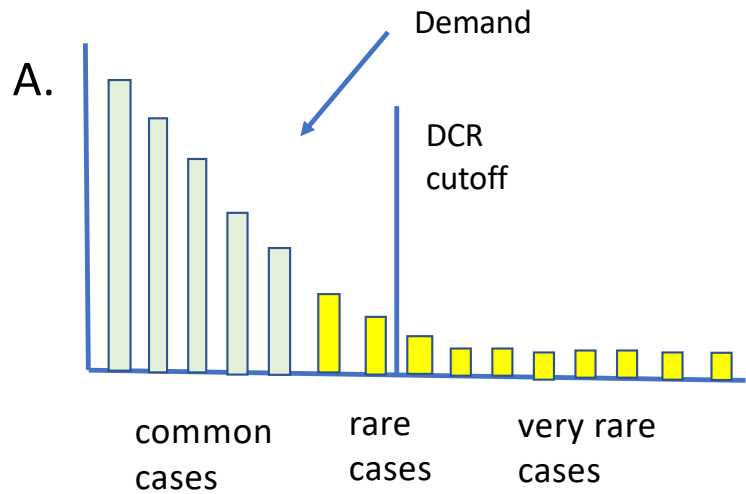
DCR Example 2. Amount of uncurated data by DCR

				Failure Rate	Success Rate
DCR 1		0	693	0%	100%
DCR 2		0	195	0%	100%
DCR 3		1	5048	0%	100%
		27	1978	1%	99%
•		19	1224	2%	98%
•		34	2122	2%	98%
•		182	10582	2%	98%
		26	1006	3%	97%
		39	1446	3%	97%
		621	21091	3%	97%
		768	22810	3%	97%

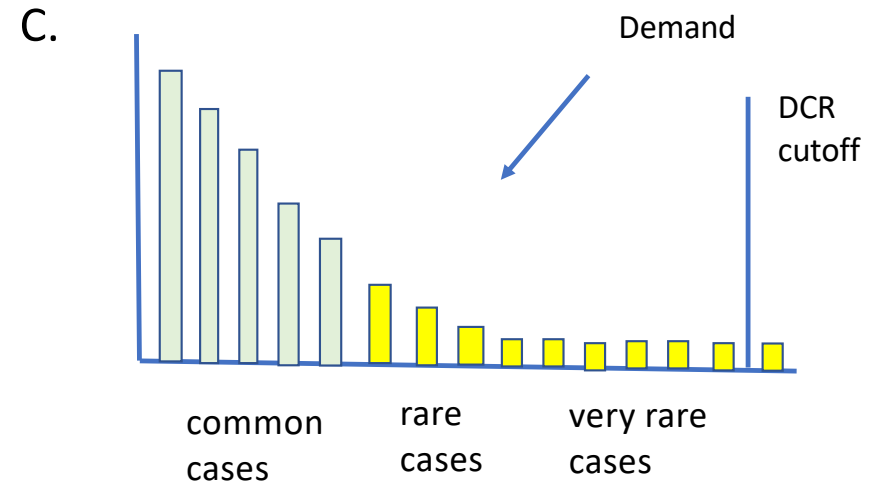
The Long Tail of Data Curation: Where is the Payoff?



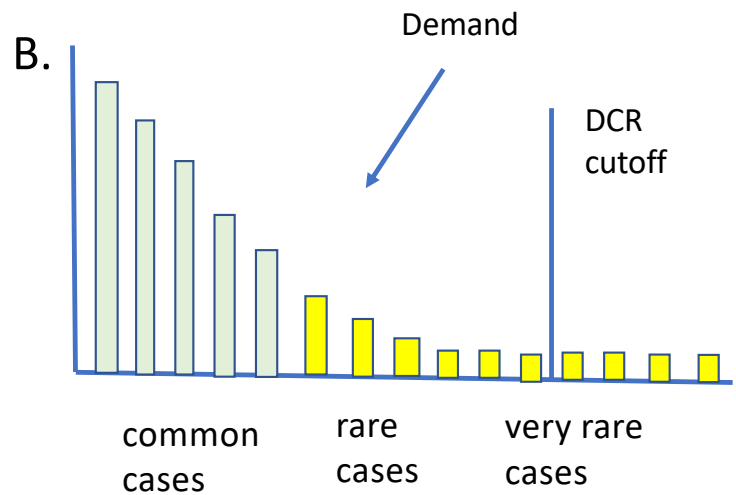
Adapted from: Robert L. Grossman, The Long Tail of Data Curation.



- Marketing applications
- AdTech



- Biomedical applications
- Particle physics



- Credit card fraud
- Electronic Medical Records

4. A Very Simple Model

Maximum Entropy Formalism, Fractals, Scaling Phenomena, and $1/f$ Noise: A Tale of Tails

Elliott W. Montroll^{1,2} and Michael F. Shlesinger^{1,2}

Received February 1, 1983

In this report on examples of distribution functions with long tails we (a) show that the derivation of distributions with inverse power tails from a maximum entropy formalism would be a consequence only of an unconventional auxiliary condition that involves the specification of the average value of a complicated logarithmic function, (b) review several models that yield log-normal distributions, (c) show that log normal distributions may mimic $1/f$ noise over a certain range, and (d) present an amplification model to show how log-normal personal income distributions are transformed into inverse power (Pareto) distributions in the high income range.

KEY WORDS: Clusters; self-similarity; Lévy distributions; log-normal distribution; random walks.

- Old observation about effects that vary multiplicatively based upon other effects

Source: Montroll, E.W. and Shlesinger, M.F., 1983. Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: a tale of tails. *Journal of Statistical Physics*, 32(2), pp.209-230.

Observation: Many interacting weak effects give risk to long-tailed DCR distributions

- Assume we effects E_1, E_2, \dots, E_3 , with probabilities p_1, p_2, p_3 , and are interested in phenomena that occur with probability

$$p = p_1 p_2 p_3 \dots$$

- Then

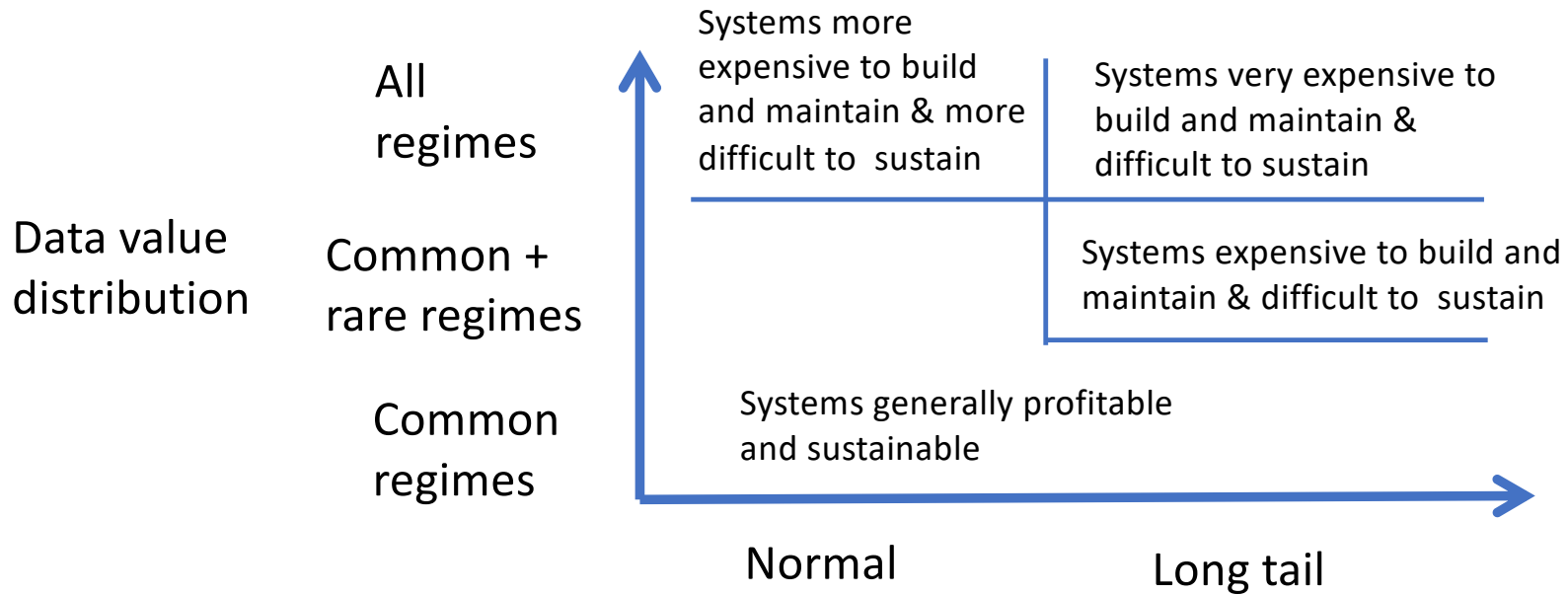
$$\log p = \log p_1 + \log p_2 + \log p_3 \dots$$

and in the limit, the distribution is normal.

This gives us a log normal distribution with a long tail.

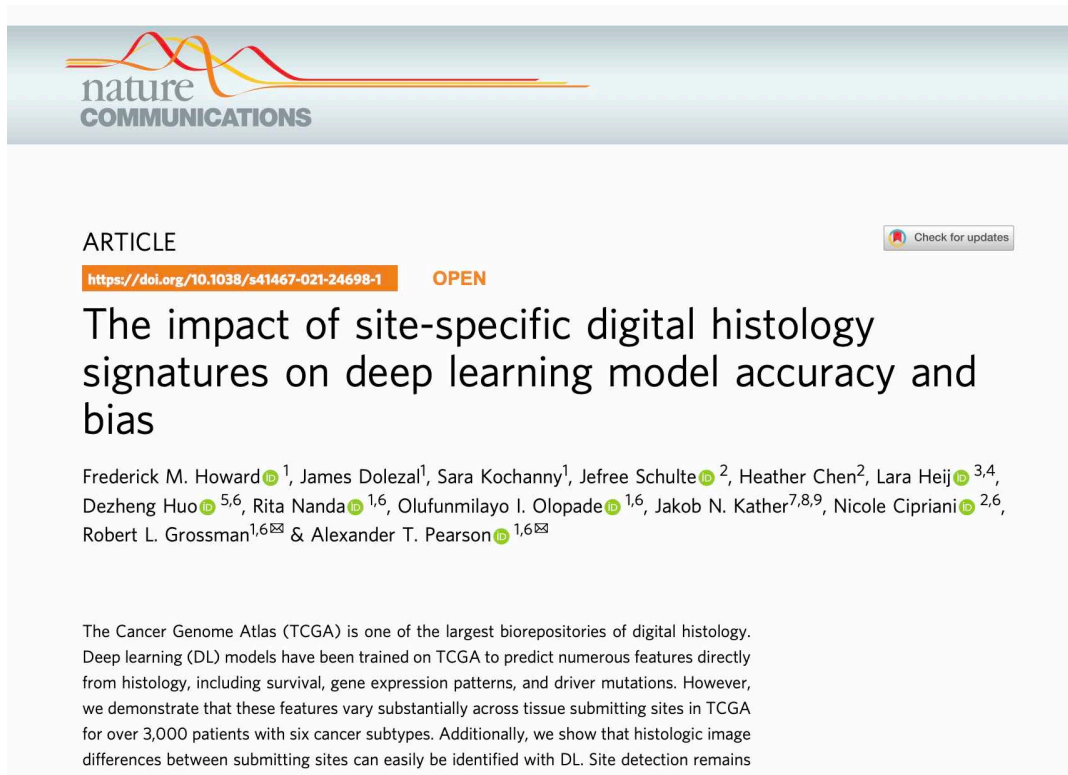
- Broadly corresponds to a phenomena which arise from interactions of many weak effects, which are common with lots of applications, including biomedical data.

Different



Data Curation Regime Distribution

How does this compare to Bias / Fairness?

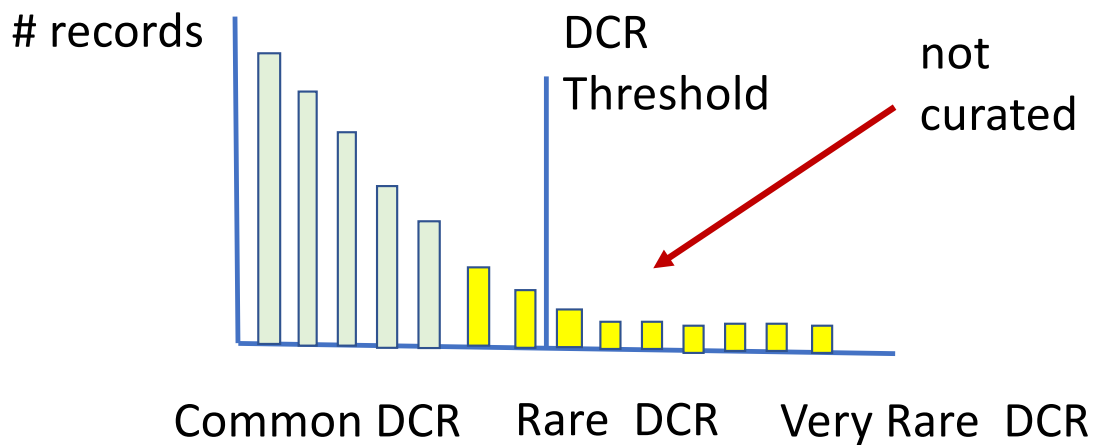


- Source of data is one obvious contributor to DCR
- But it is not the only one
- Each site tends to have its own DCR

Source: Frederick M. Howard et al, The impact of site-specific digital histology signatures on deep learning model accuracy and bias, Nature Communications 12, no. 1, 2021, pages 1-13.

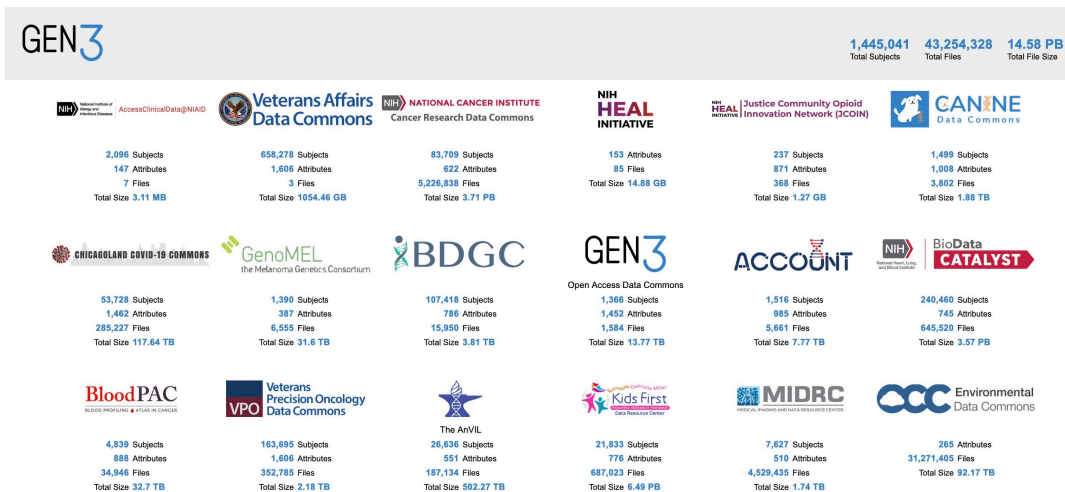
5. Some Ways That Platforms Deal with Data Curation

1. Ignore Rare DCR



- Ignore rare data curation regimes.
- This is the most common approach and in practice is almost always done, just with different DCR thresholds.
- Work well if the interest in the data is to the left of the DCR Threshold.

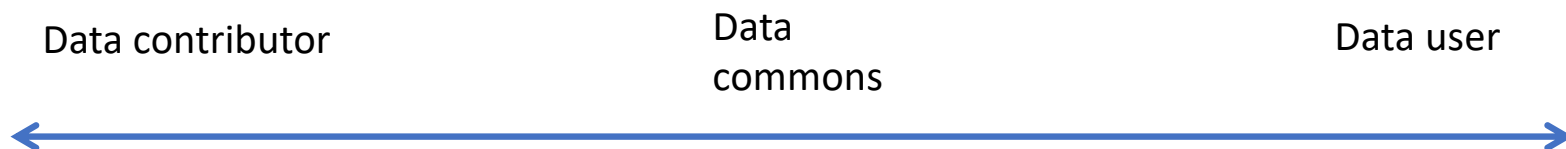
2. Leverage Economies of Scale



- Use a data mesh to centralize
 - Data curation
 - Data ingestion
 - Commons operations
 - Security & compliance

3. Handoff Data Curation to Others

- Require data contributors submit data through an API aligned with the data model (used by the GDC)
- Provide the data contributor different data curation tools
- Shift the responsibility of curation to intermediaries
- Shift the responsibility of curation to end users
- Shift the responsibility of curation to later

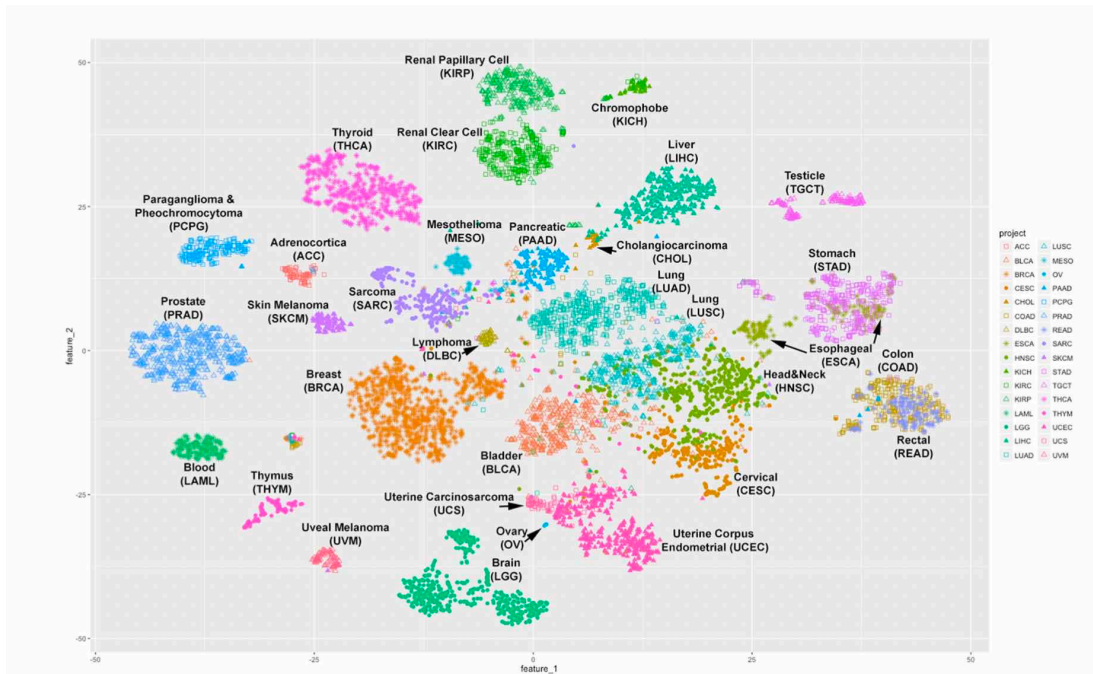


6. Why Does This Always Seem to Come Up in Practice

Why is the data for some models so much harder to curate than others?

1. There are many different data contributors, and very little standardization in practice.
2. There are many distinct subtypes, each of which requires a different model.
3. Curating data at the level required for a good model requires working through many rare DCR.
4. There are many weak interacting contributing causes vs a few strong contributing causes so there are many rare DCR.

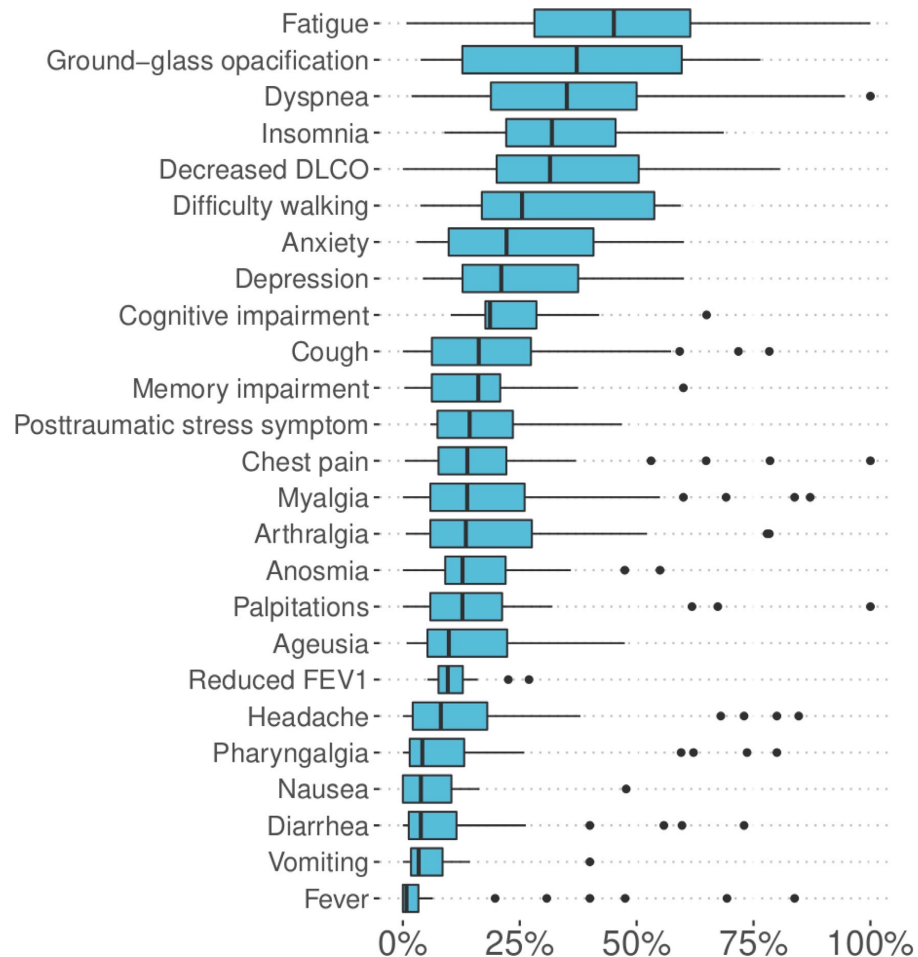
Many Distinct Subtypes, which Require Different Models



- Cancer is not a single disease but has many different subtypes
- Site of origin (breast, ovarian, pancreatic, etc.) is a very simple way to classify different cancers.
- Molecular subtyping based on the DNA sequence of the tumor is a more useful way to classify cancers and to build predictive models for them.

Zhang, Z., Hernandez, K., Savage, J., Li, S., Miller, D., Agrawal, S., ... & Grossman, R. L. (2021). Uniform genomic data analysis in the NCI Genomic Data Commons. *Nature Communications* 2021; 12(1), pp 1-11.

Many Distinct Subtypes, which Require Different Models

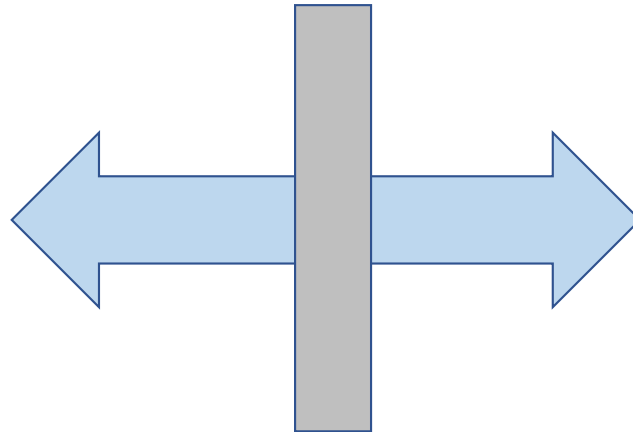


- Long covid is a complex disease that is still being characterized.
- The plot shows what symptoms are present

Deer, Rachel R., Madeline A. Rock, Nicole Vasilevsky, Leigh Carmody, Halie Rando, Alfred J. Anzalone, Marc D. Basson et al. "Characterizing long COVID: deep phenotype of a complex condition." *EBioMedicine* 74 (2021): 103722.

The Data Gap in Machine Learning and AI

The amount of biomedical data is growing exponentially.



We are usually data-limited in (biomedical) data science (at least in terms of **well-curated data**) at the scale we need data for machine learning and AI.

The Data Gap

