



# The Data Gap in Machine Learning and AI: Why Much of Machine Learning and AI is Still Data Limited, and Some of the Options Available

Robert L. Grossman

Dept. of Medicine & Computer Science  
Center for Translational Data Science  
University of Chicago

April 13, 2022

# 1. The Translational Challenge

“Hundreds of AI Tools have been built to catch COVID. None of them helped.\*”

\* Will Douglas Heaven, Hundreds of AI tools have been built to catch covid. None of helped. MIT Technology Review, July 30, 2021, <https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/>

RESEARCH

OPEN ACCESS

Check for updates

**FAST TRACK**

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

Laure Wynants,<sup>1,2</sup> Ben Van Calster,<sup>2,3</sup> Gary S Collins,<sup>4,5</sup> Richard D Riley,<sup>6</sup> Georg Heinze,<sup>7</sup> Ewoud Schuit,<sup>8,9</sup> Marc M J Bonten,<sup>8,10</sup> Darren L Dahly,<sup>11,12</sup> Johanna A Damen,<sup>8,9</sup> Thomas P A Debray,<sup>8,9</sup> Valentijn M T de Jong,<sup>8,9</sup> Maarten De Vos,<sup>2,13</sup> Paula Dhiman,<sup>4,5</sup> Maria C Haller,<sup>7,14</sup> Michael O Harhay,<sup>15,16</sup> Liesbet Henckaerts,<sup>17,18</sup> Pauline Heus,<sup>8,9</sup> Michael Kammer,<sup>7,19</sup> Nina Kreuzberger,<sup>20</sup> Anna Lohmann,<sup>21</sup> Kim Luijken,<sup>21</sup> Jie Ma,<sup>5</sup> Glen P Martin,<sup>22</sup> David J McLernon,<sup>23</sup> Constanza L Andaur Navarro,<sup>8,9</sup> Johannes B Reitsma,<sup>8,9</sup> Jamie C Sergeant,<sup>24,25</sup> Chunhu Shi,<sup>26</sup> Nicole Skoetz,<sup>19</sup> Luc J M Smits,<sup>1</sup> Kym I E Snell,<sup>6</sup> Matthew Sperrin,<sup>27</sup> René Spijker,<sup>8,9,28</sup> Ewout W Steyerberg,<sup>3</sup> Toshihiko Takada,<sup>8</sup> Ioanna Tzoulaki,<sup>29,30</sup> Sander M J van Kuijk,<sup>31</sup> Bas C T van Bussel,<sup>1,32</sup> Iwan C C van der Horst,<sup>32</sup>

“... 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.”

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

The Alan Turing Institute

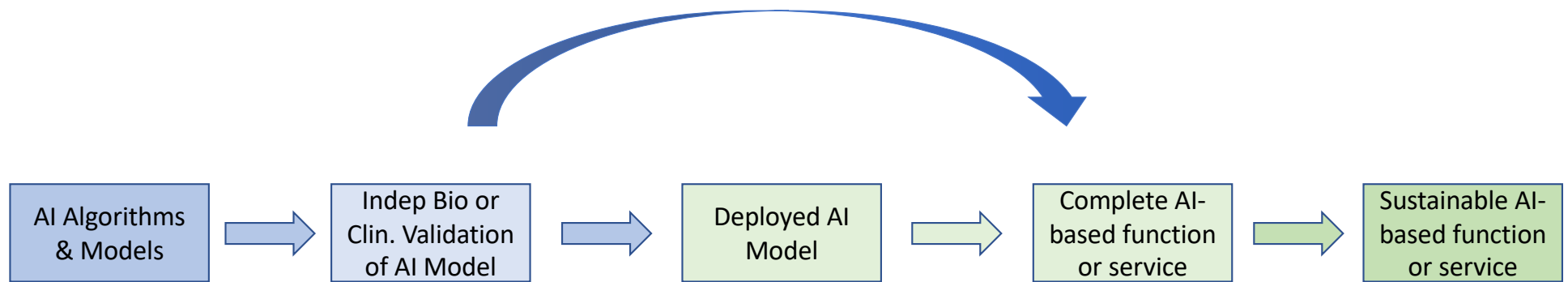


## Data science and AI in the age of COVID-19

Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic

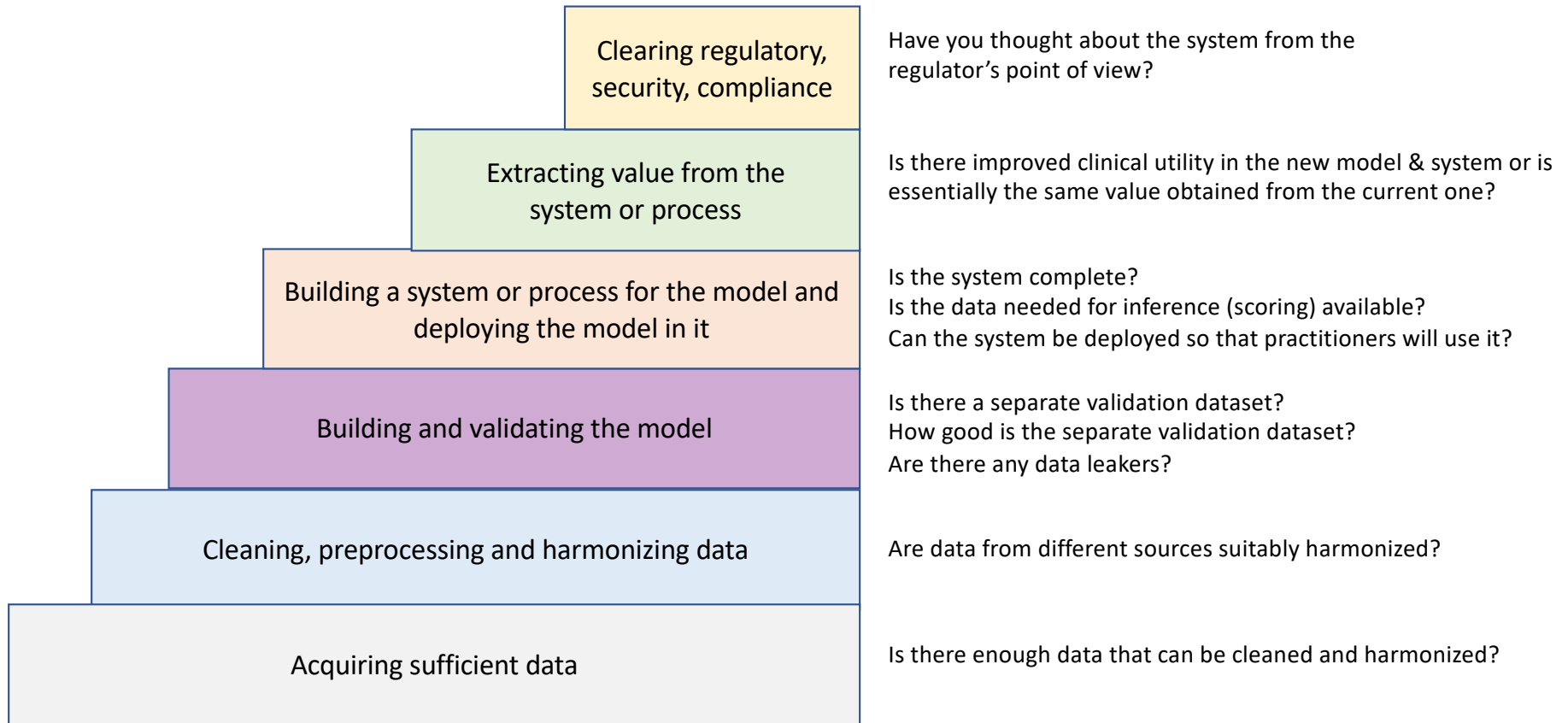
Inken von Borzyskowski, et. al., editors, Data science and AI in the age of COVID-19 – report, Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic, Turing Institute, 2021.

# The Long Journey to Translational Impact (from data)



Adapted from: Robert L. Grossman, *Developing an AI Strategy: a Primer*, Open Data Press, 2020

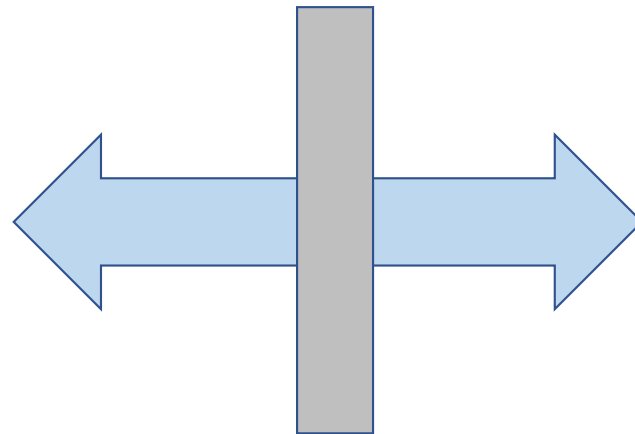
# Why do so many AI models fail? (The Staircase of failure)



Adapted from: Robert L. Grossman, Developing an AI Strategy: a Primer, Open Data Press, 2020.

# The Data Gap in Machine Learning and AI

The amount of global IP data traffic per month is estimated to be more than 270 EB per month.\*

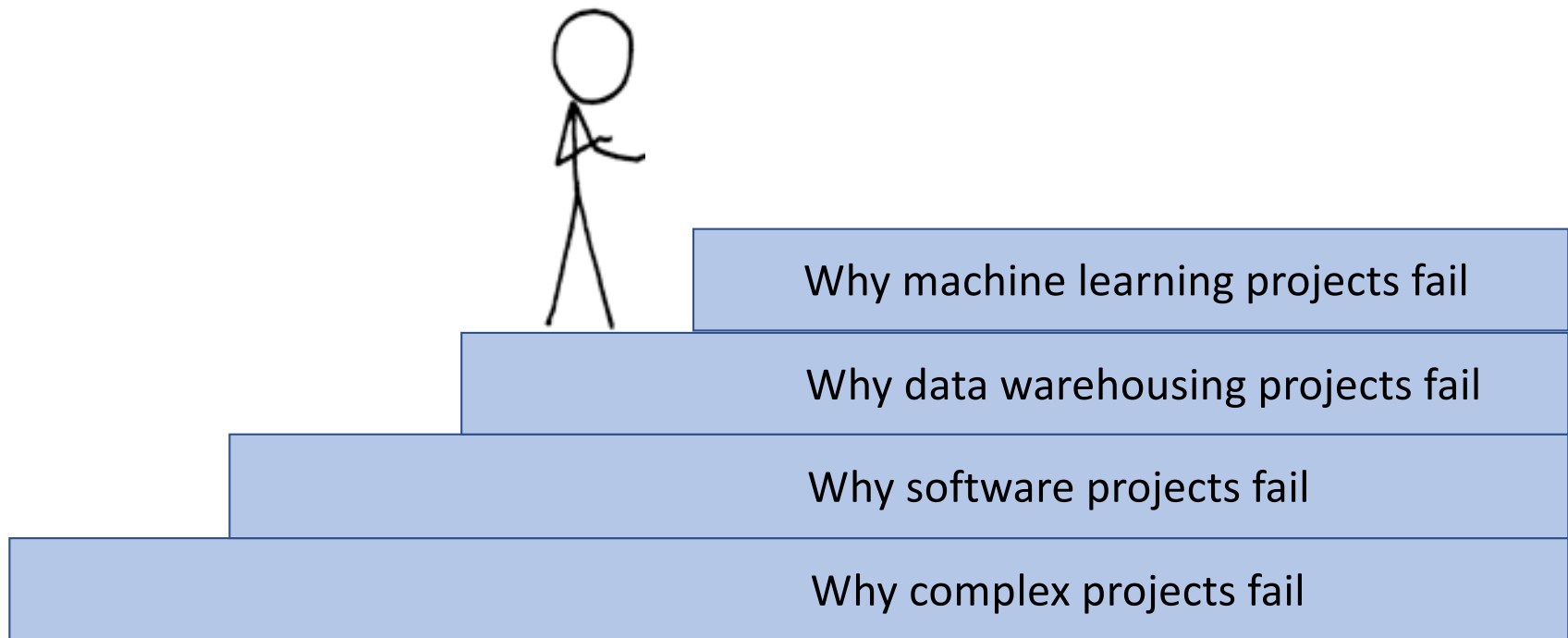


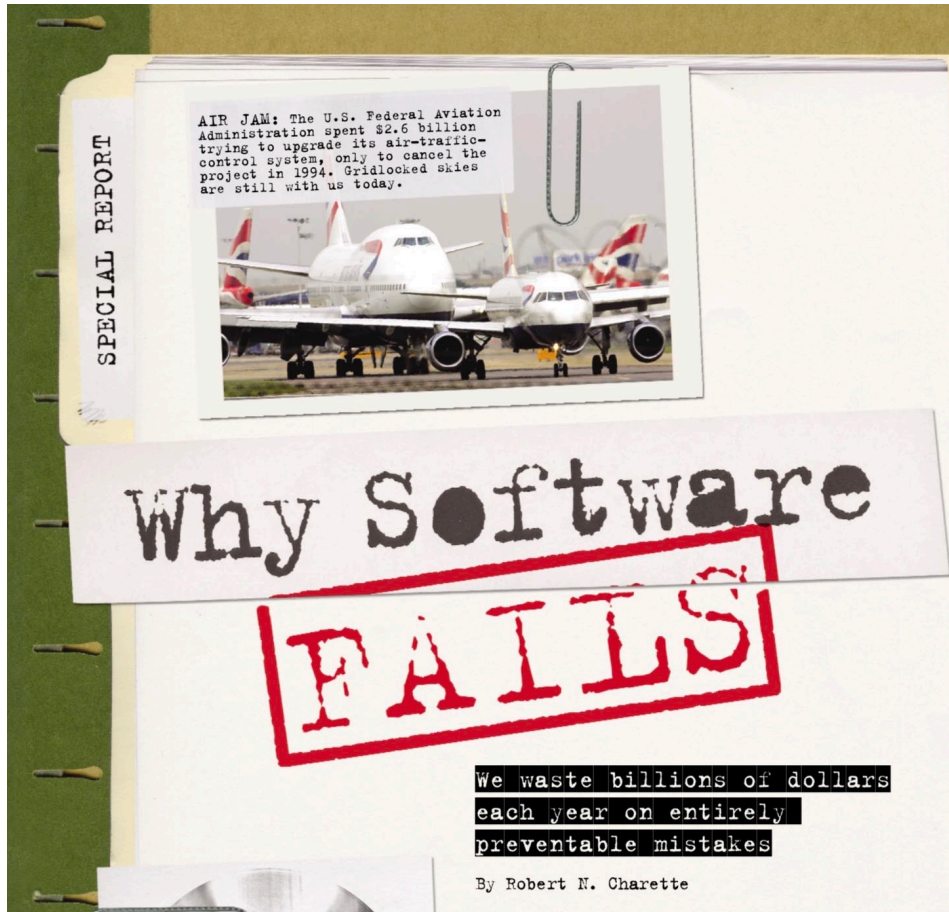
We are usually data-limited in (biomedical) data science (at least in terms of well-curated data) at the scale we need data for machine learning and AI.

The Data Gap

\*Cisco, VNI Complete Forecast Highlights, 2016, [https://www.cisco.com/c/dam/m/en\\_us/solutions/service-provider/vni-forecast-highlights/pdf/Global\\_2021\\_Forecast\\_Highlights.pdf](https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf)

# The Staircase of Failure



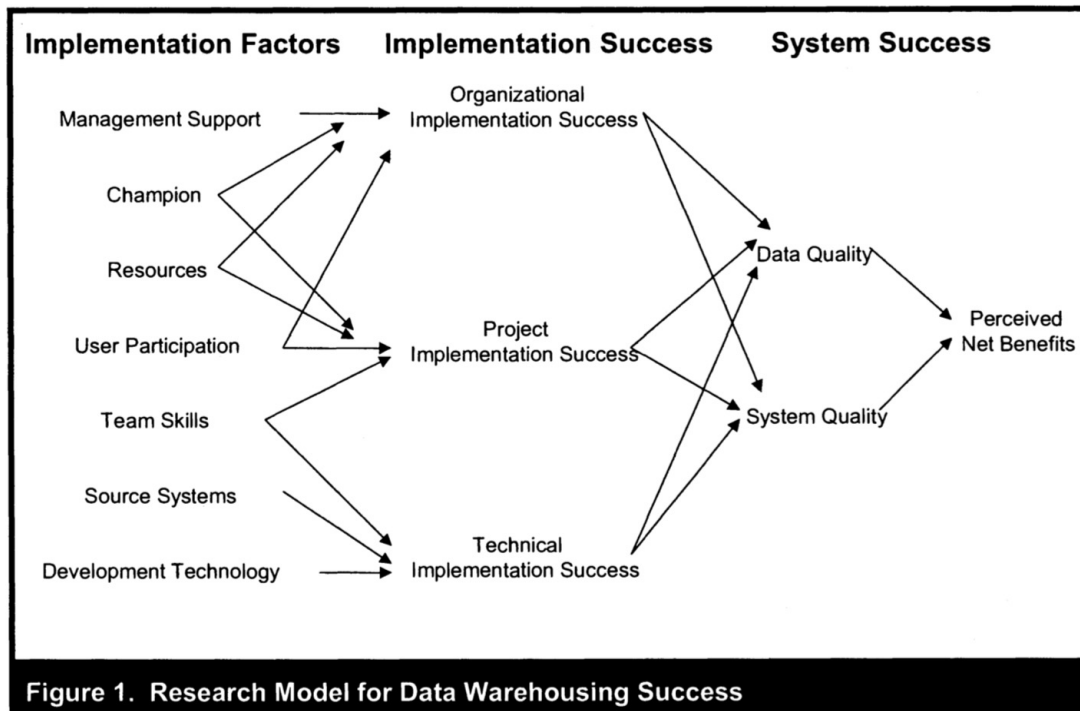


1. Unrealistic or unarticulated project goals
2. Inaccurate estimates of needed resources
3. Badly defined system requirements
4. Poor reporting of the project's status
5. Unmanaged risks
6. Poor communication among customers, developers, and users
7. Use of immature technology
8. Inability to handle the project's complexity
9. Sloppy development practices
10. Poor project management
11. Stakeholder politics
12. Commercial pressures

Source: Charette, R.N., 2005. Why software fails, IEEE spectrum, 42(9), pages 42-49 (896 citations).



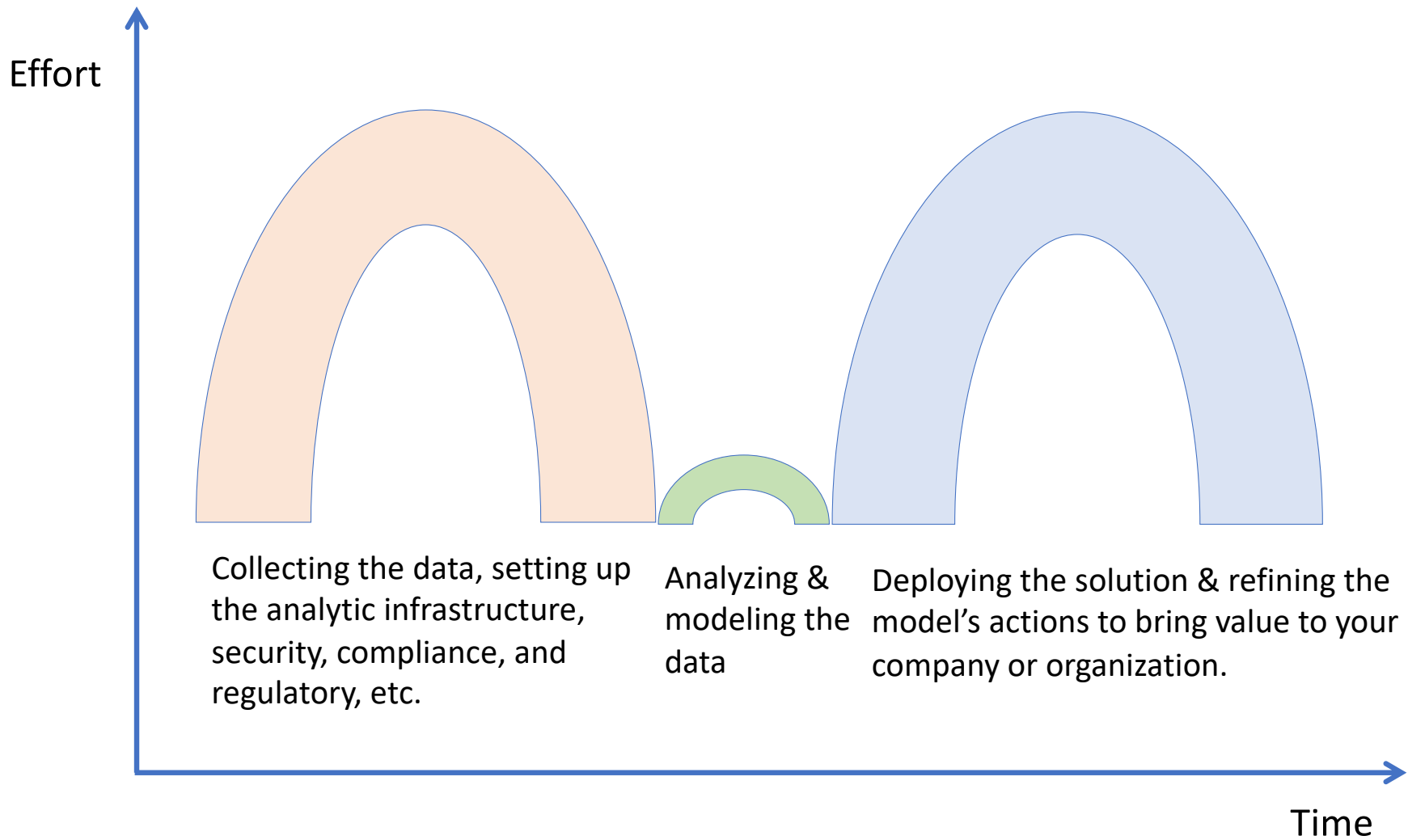
# Why Data Warehouse Projects Fail



## Implementation Factors

1. Management support
2. Champion
3. Resources
4. User participation
5. Team skills
6. Source systems
7. Development technology

Source: Wixom, Barbara H., and Hugh J. Watson. "An empirical investigation of the factors affecting data warehousing success." MIS quarterly (2001): 17-41 (1647 citations)



## Top Ten Reasons Analytic Projects Fail

1. You never get the data that you need. (The “data gap”)
2. The model never gets deployed.
3. The model does not return the business value expected / promised.
4. You do not have the data scientists required to manage the data, build the models, deploy the models, and advocate for the business value delivered.
5. You do not have an automated testing, deployment, and evaluation environment for improving the model.

## Top Ten Reasons Analytic Projects Fail

6. You cannot build the analytic infrastructure required.
7. There is no senior analytic leader who can effectively organize the analytic efforts.
8. You have the right model, but the wrong actions.
9. You chose the wrong analytic opportunity to pursue.
10. You bring in consultants who do not deliver what you need.

## 2. The Emerging Role of Foundation Models

# What is a Foundational Model?

“In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.”\*

- Examples of foundational models: GPT-3
- Foundation models include large language models that can answer questions or generate text from a prompt.
- These models can be very impressive and seem to show new emergent capabilities.
- These types of models can go very wrong. GPT-3 responses have been very polarizing and biased, reflecting the data it was built on.
- “These models are really castles in the air, they have no foundation whatsoever.”\*\*

\* Developing and understanding responsible foundation models, The Center for Research on Foundation Models (CRFM),, retrieved from <https://crfm.stanford.edu/> on March 10, 2022.

\*\* Jitendra Malik, Professor of computer science, UC Berkeley

## Example: GPT-3 Training Data

GPT-3 Training Data		
Dataset	# Tokens	Weight in Training Mix
<a href="#">Common Crawl</a>	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

- GPT-3 was trained on about 500 billion tokens.
- GPT-3 has 175 billion machine learning parameters.
- Available through open API
- GPT-J is an open-Sources 6 Billion Parameter GPT-3 Clone developed by EleutherAI.

## On the Opportunities and Risks of Foundation Models

Rishi Bommasani\* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora  
 Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill  
 Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji  
 Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue  
 Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh  
 Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman  
 Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt  
 Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain  
 Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani  
 Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kudipudi  
 Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent  
 Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning  
 Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan  
 Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan  
 Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech  
 Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren  
 Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh  
 Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin  
 Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu  
 Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia  
 Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou  
 Percy Liang\*<sup>1</sup>

Center for Research on Foundation Models (CRFM)  
 Stanford Institute for Human-Centered Artificial Intelligence (HAI)  
 Stanford University

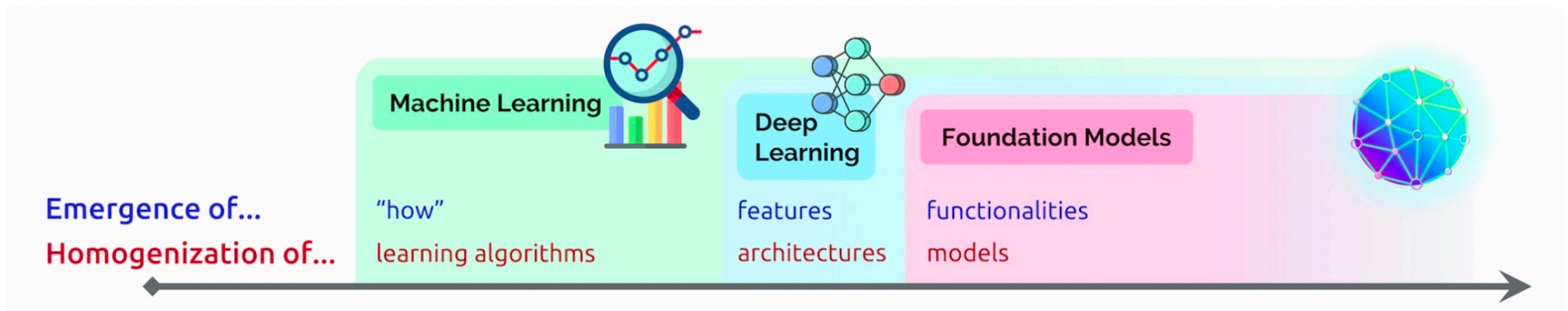
*AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles*

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character.\*

\*Bommasani, Rishi, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).



# One Narrative: ML → Deep Learning → Foundation Models

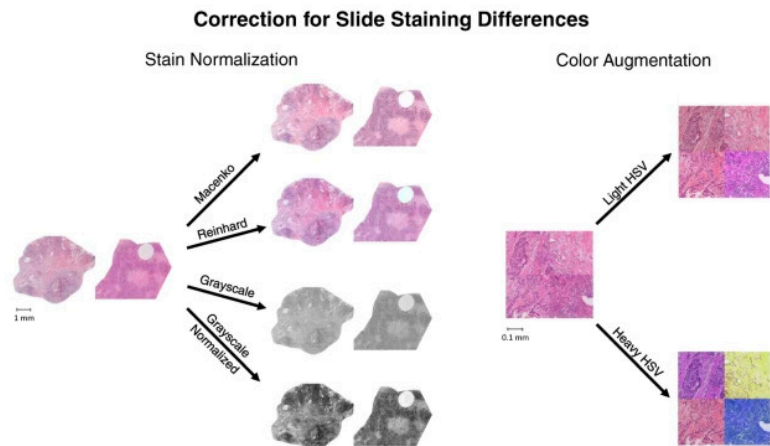
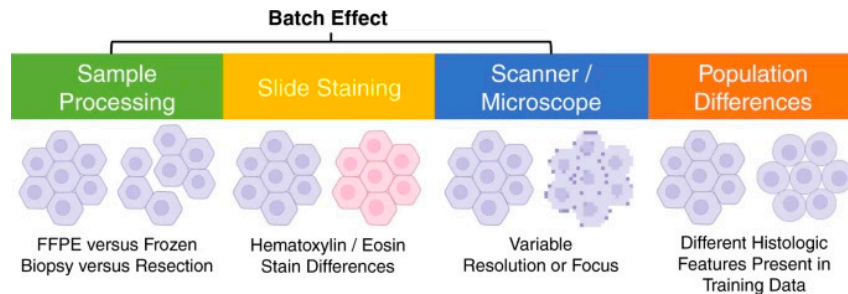


The story of AI has been one of increasing emergence and homogenization. With the introduction of machine learning, how a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3)

Source of figure and caption: Bommasani, Rishi, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).

### 3. On (some of) the Varieties of Data & Model Bias

## Q1: What is the (data) source bias?



- Different sources of data have different biases that can be hard to identify and to get rid of.
- For example, in histopathology images, the sample preparation process, staining, scanner / microscope, or leave artifacts that ML/DL models pick up.

Howard, Frederick M., James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman<sup>1</sup>, and Alexander T. Pearson, The impact of site-specific digital histology signatures on deep learning model accuracy and bias, Nature Communications 12, no. 1, 2021, pages 1-13.

## Q2: What is the Coding Bias?



- Think of coding as a map from a analog event or measurement to a numeric / symbolic representation of it.
- Medical coding is designed for medical reimbursement, not to support medical research.
- There are over 70,000 ICD-10-PCS procedure codes and over 69,000 ICD-10-CM diagnosis codes, compared to about 3,800 procedure codes and roughly 14,000 diagnosis codes found in the previous ICD-9-CM.

### Q3: What is the Contributor Bias?

Training: Text and images on the internet.

Deployed: Text and images on the internet.

VS

Training: Medical data available on the internet.

Deployed: Medical data available within a hospital's EMR or other operational system.

Q4: What Elements of the Training Data are Available in the Real World Data?

Training: Many features +  
curated outcome

vs

Deployed: Some features

## 4. Four Techniques to Close the Data Gap

# 1. Data Augmentation

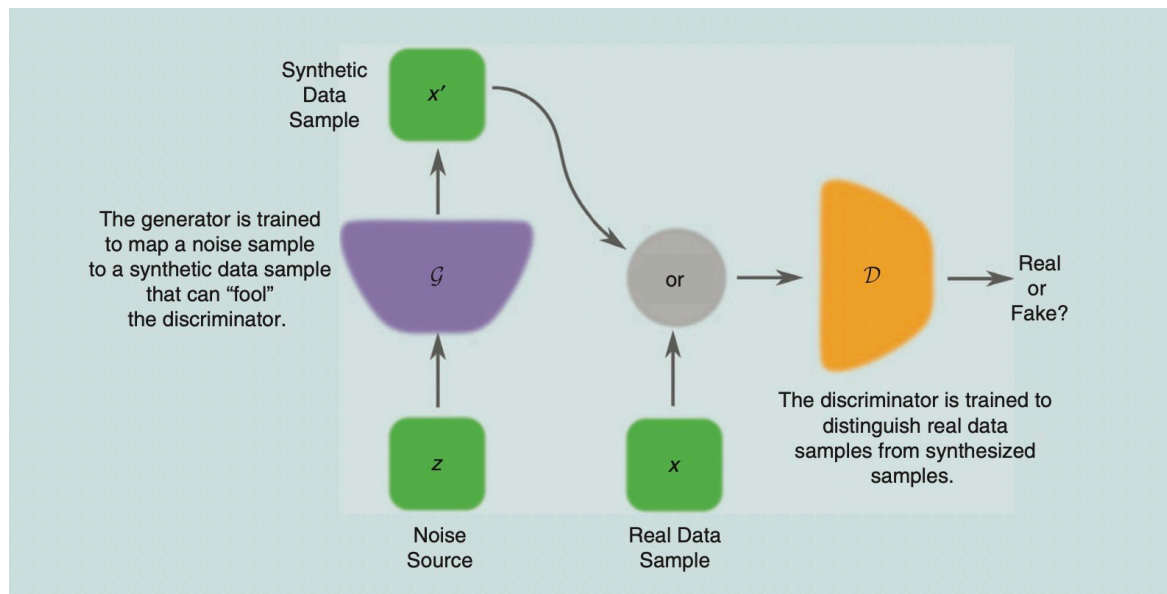


Creating more artificial images by taking an image and cropping, rotating, flipping, changing hues and colors, and mixing is quite effective

Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of big data, 6(1), pp.1-48.



## 2. Generative Adversarial Networks (GAN)



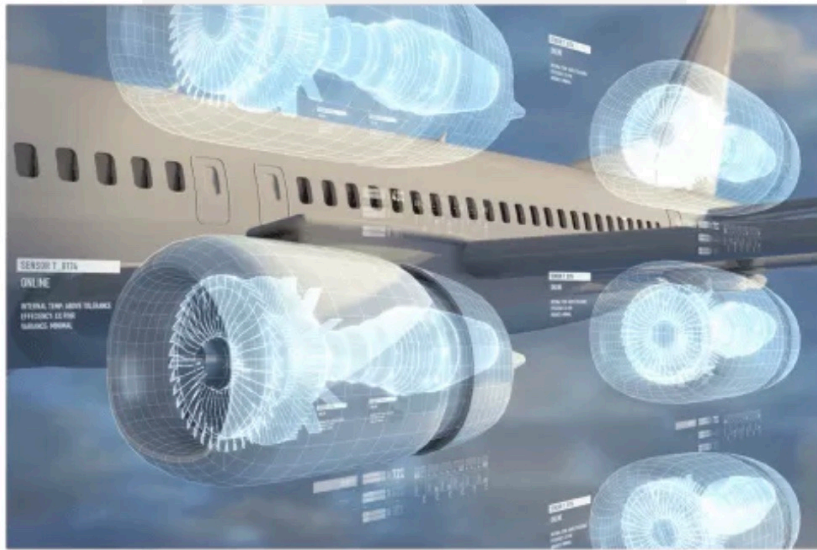
Two models that are learned during the training process for a GAN:

- 1) Generator
- 2) Discriminator

Typically, the models are deep neural networks.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1), pp.53-65.

### 3. Digital Twins (or other simulated data)



Visualization of GE's digital twins for jet engines\*\*

A digital twin is the electronic representation -- the digital representation -- of a real-world entity, concept, or notion, either physical or perceived.\*

\*Voas, J., Mell, P. and Piroumian, V., 2021. Considerations for Digital Twin Technology and Emerging Standards (No. NIST Internal or Interagency Report (NISTIR) 8356 (Draft)). National Institute of Standards and Technology.

\*\*Source of image: <https://www.ge.com/digital/applications/digital-twin>

## 4. “Hand Engineered” Data

Andrew Ng: Farewell, Big Data

The AI pioneer says it's time for smart-sized, “data-centric” solutions to big issues

[Eliza Strickland](#)

09 Feb 2022

10 min read



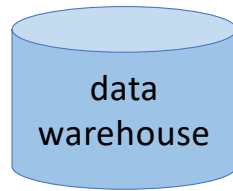
“In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn.”

Andrew Ng, CEO & Founder, Landing AI

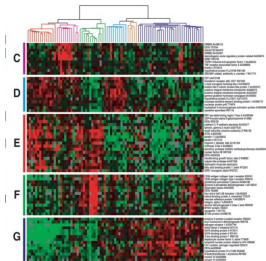
Strickland, Eliza. IEEE Spectrum (2022). Andrew Ng: Farewell Big Data.

## 5. Closing the Data Gap with Data Commons and Data Meshes

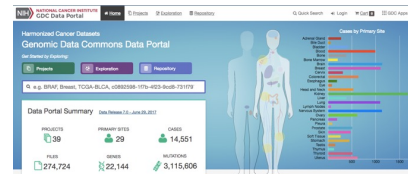
Data commons = data curation + data analysis  
+ data products available via FAIR APIs that  
support interactive notebooks + computational  
pipelines



**Data warehouses**  
organize the data for an  
**organization** (1990's)



**Databases** organize data  
around a **project** (1970's)



**Data commons** organize  
the data for a scientific  
**discipline** or field (2010's)



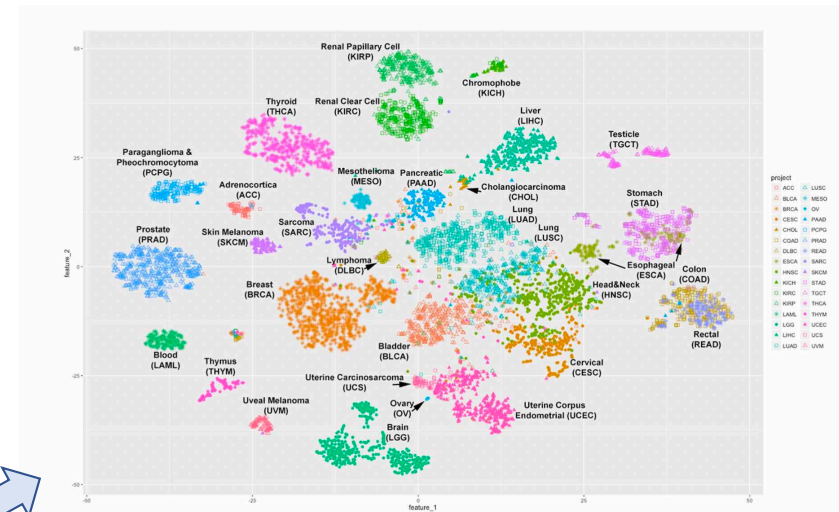
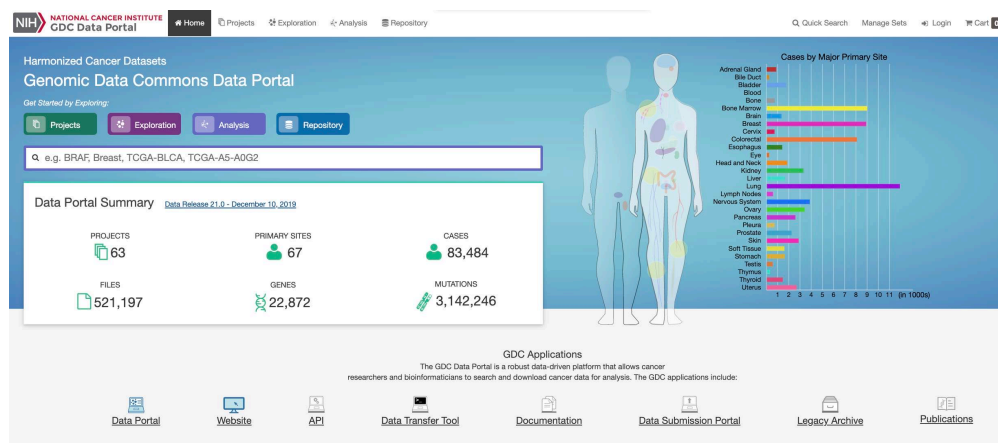
**Data meshes (aka data ecosystems)** enable discoveries  
**across multiple commons**  
**operated by different organizations.** (2020's)

**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.

## 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



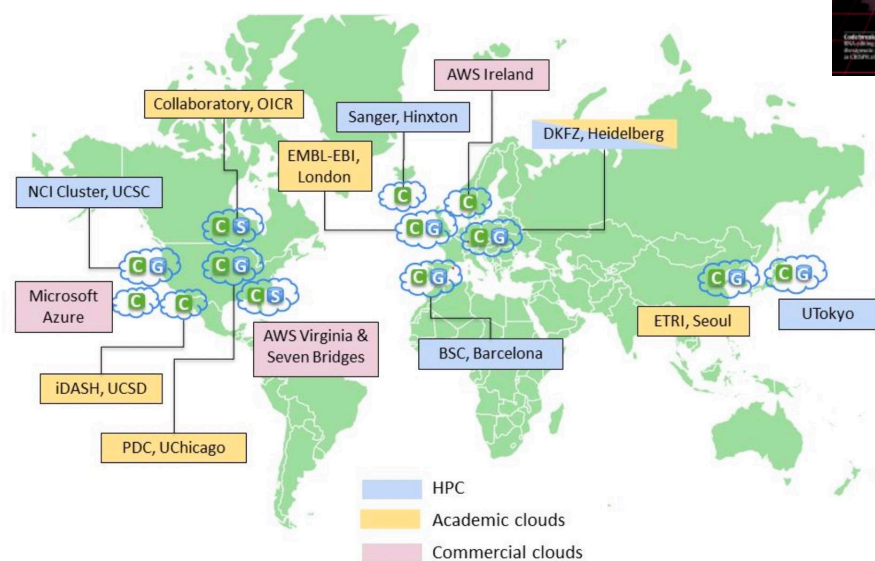
CTDS pan-cancer molecular  
subtyping using GDC data

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.

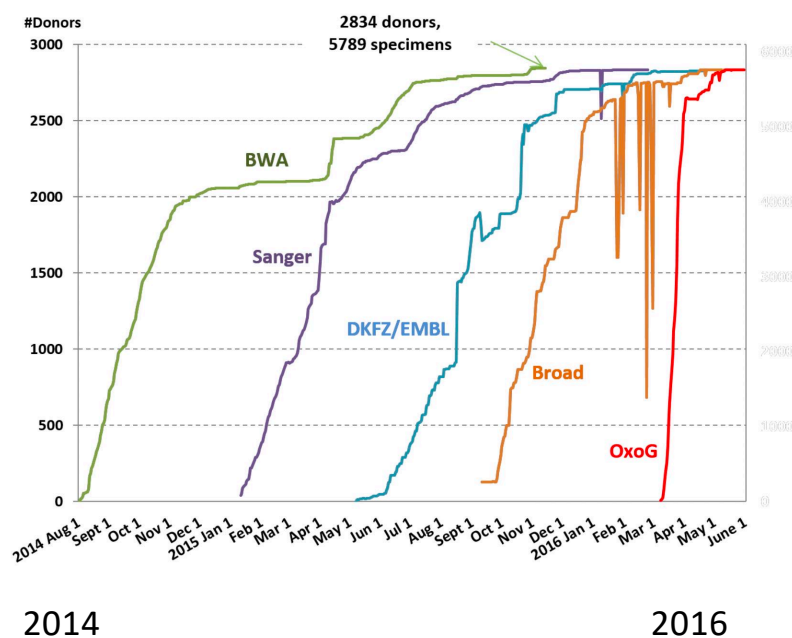
Heath AP, Ferretti V, ... Grossman RL, The NCI Genomic Data Commons, *Nature Genetics* 2021 Mar; 53(3), pp 257-262.

# Ex 1 (cont'd): TCGA-ICGC Pan-Cancer Analysis Variant Calling (2014—2016)

## Federated Machine Learning



Nature special Issue: 6 Feb 2020





## Ex. 2: Pandemic Response Commons

Figure 2. State-specific CFRs, separately by age group and racial/ethnic group (CDC data)

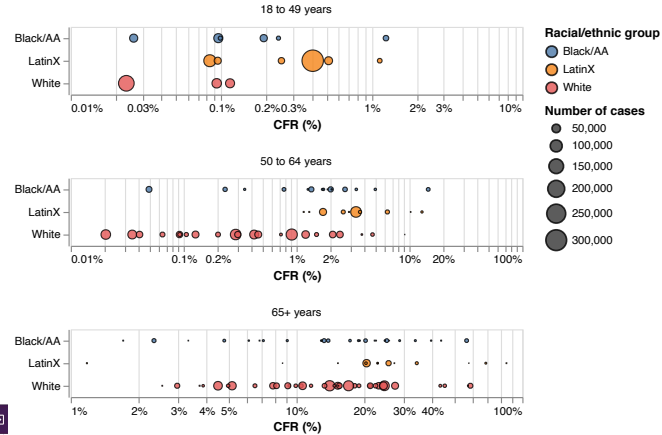
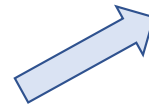
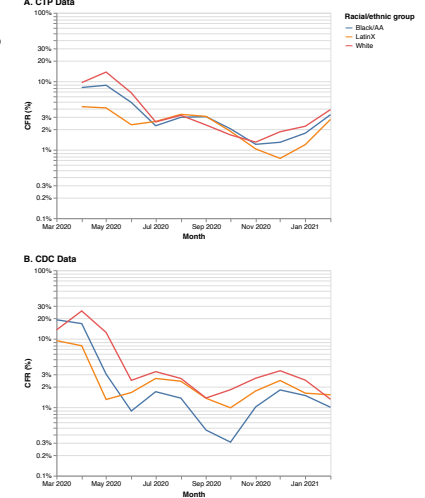


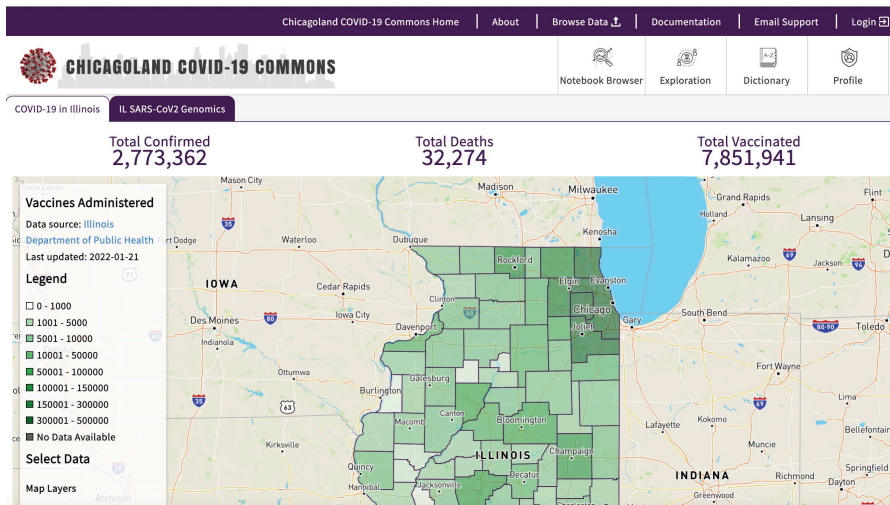
Figure 1. CFR by month and racial/ethnic group, CTP and CDC datasets



Simpson's Paradox is the observation that the direction of an association at the population-level may be reversed within the subgroups comprising that population.

As a start to understanding health disparities in COVID-19 data, we looked at case fatality rate.

- Overall, the case fatality rate, as computed from data provided by CDC & COVID Tracking Project, is higher for Whites, but the association is reversed when you look at subgroups by ages (simplified version)



Schumm, L.P., Giurcanu, M.C., Locey K.J., Ortega, J.C., Zhang, Z., Grossman R.L., Racial/Ethnic Disparities in the Observed COVID-19 Case Fatality Rate Among the U.S. Population, submitted for publication.



# GEN<sup>3</sup> Data Mesh

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



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



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



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



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



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



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



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



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



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



Open Access Data Commons  
**1,366** Subjects  
**1,452** Attributes  
**1,584** Files  
Total Size **13.77 TB**



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



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



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



**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**



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

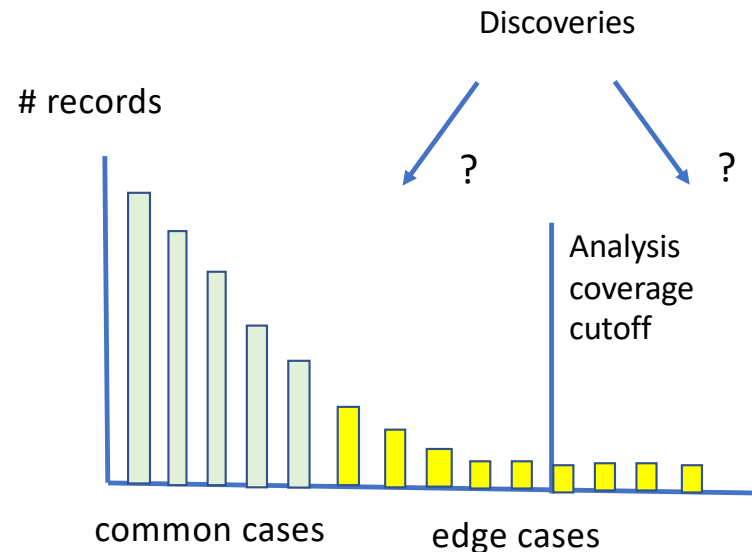


**7,627** Subjects  
**510** Attributes  
**4,529,435** Files  
Total Size **1.74 TB**



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

# The Long Tail of Data Curation



The challenges of cleaning & curating data at scale (the long tail of data curation)

- Many biomedical problems requiring understanding very weak effects resulting from many complex interacting mechanisms.
- This requires data scale, careful curation, coverage of edge cases, and a process for updating models as more evidence accumulates.

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

## 6. Summary

## Summary

1. Most AI and machine learning models fail before achieving sustainable deployment.
2. The biggest reason is a data gap.
3. Data commons are a good approach for closing this data gap and creating well curated collections of data serving a particular community.



**Abstract:** Although large amounts of online text, images and audio have provided enough data that deep learning models can be developed that significantly improve language translation, image recognition, speech recognition and related applications, developing and deploying machine learning and AI models that provide value and limit bias is still quite difficult in many application areas due to the lack of suitable data. This is especially the case in biology, medicine and health care. We discuss some of the reasons that many important AI problems are still data-limited and some of the approaches that have been taken to address this challenge. We use case studies from machine learning models in COVID-19 and cancer to illustrate some of the challenges and some of the options available.

**Biographical sketch:** Robert L. Grossman is the Frederick H. Rawson Distinguished Service Professor of Medicine and Computer Science and the Jim and Karen Frank Director of the Center for Translational Data Science (CTDS) at the University of Chicago. CTDS is a research center that focuses on data science and its applications to problems in biology, medicine, health care and the environment. He is also a Partner at Analytic Strategy Partners LLC, which helps companies develop machine learning and AI strategies to advance their mission.