# Crossing the Data Chasm and Finding the Data You Need for Machine Learning Models: From Data Augmentation to Synthetic Data and Data-Centric AI

Robert L. Grossman

Analytic Strategy Partners
and
Center for Translational Data Science, University of Chicago

May 13, 2022

# 1. Data Chasms

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] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32]

“… 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

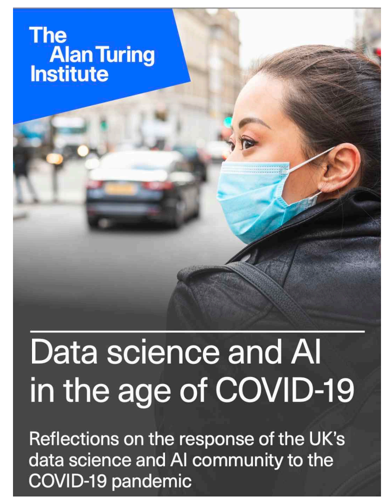
Featured Topics Newsletters Events Podcasts Sign in Subscribe

ARTIFICIAL INTELLIGENCE

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

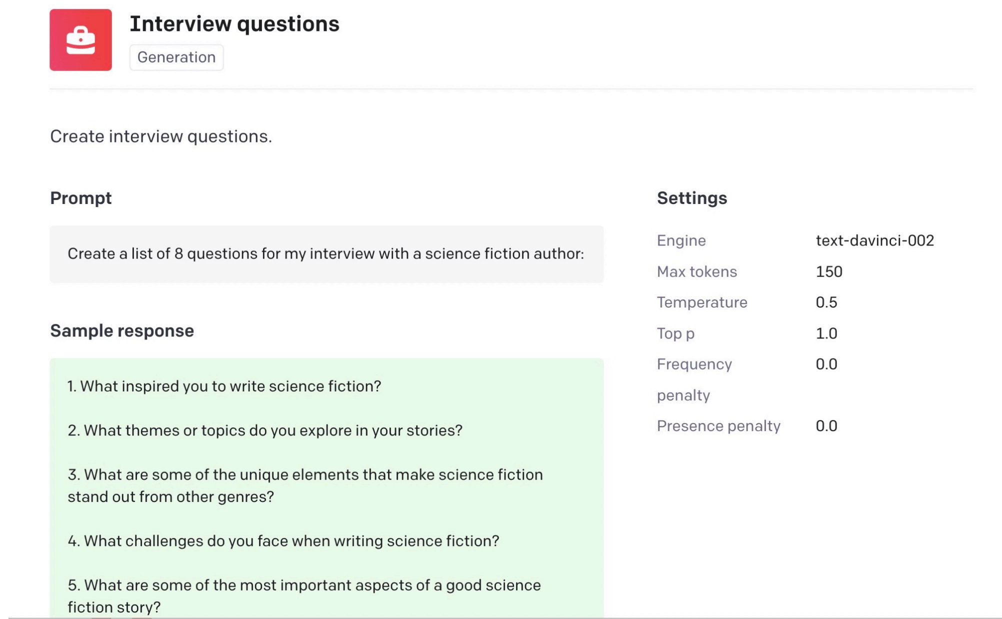
Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven July 30, 2021




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.

- GPT-3 large language model
- It was developed by OpenAI and is available via an API
- Number of learned weights – 175 billion (compared to next largest at the time 17 billion)
- Training cost for a single run c. $4,000,000
- There are similar projects that are open source



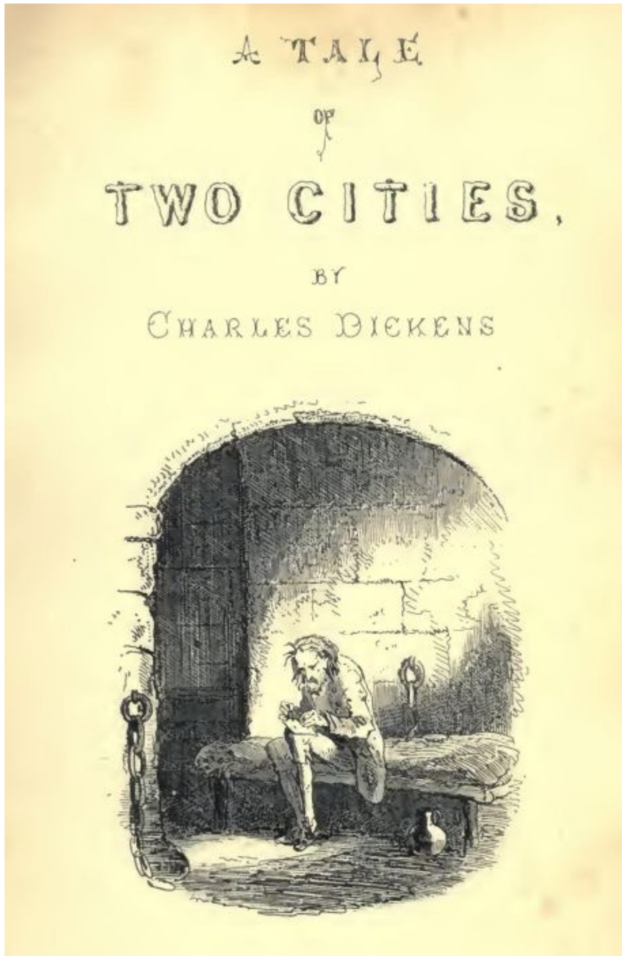

Prompt: Create a list of 8 questions for my interview with a science fiction author.

Example of response: What are some of the unique elements that make science fiction stand out from other genres?

Source: OpenAI, retrieved from https://beta.openai.com/examples/default-interview-questions on April 10, 2022.

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair …"

Source: The Project Gutenberg eBook of A Tale of Two Cities, by Charles Dickens, retrieved from https://www.gutenberg.org/files/98/98-h/98-h.htm on April 2, 2022.



It’s the age of data abundance, it’s the age of data scarcity. It’s the age of model fairness, it’s the age of the model bias.  It’s the age of the model explainability, it’s the age of black-box models.  It’s the age of the automatic generation of features, it’s the age of hand generated features.

| | | |
|---|---|---|
| Short responses from GPT-3 are cannot be easily distinguished from those written by humans. | vs | “Hundreds of AI Tools have been built to catch COVID. None of them helped.*” |
| The world generates exabytes of data per day and we can use this data to build large language models of unprecedented utility. | vs | We didn’t have enough data of high enough quality to build useful COVID models during the pandemic. |

* 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/

Top Ten Reasons Analytic Projects Fail*

1. You never get the data that you need. (Crossing the "data chasm")

2. The model never gets deployed. (ML Ops)
3. The model does not return the value expected / promised.

…

Topic for Today: What makes COVID models so different than large language models and how can we learn to recognize the difference?

Source: Robert L. Grossman, Developing an AI Strategy: a Primer, Open Data Press 2020, https://analyticstrategy.com/books/

# 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 |
| Common Crawl | 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.

arXiv:2108.07258v2 [cs.LG] 18 Aug 2021

# 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 Kuditipudi
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*[1]

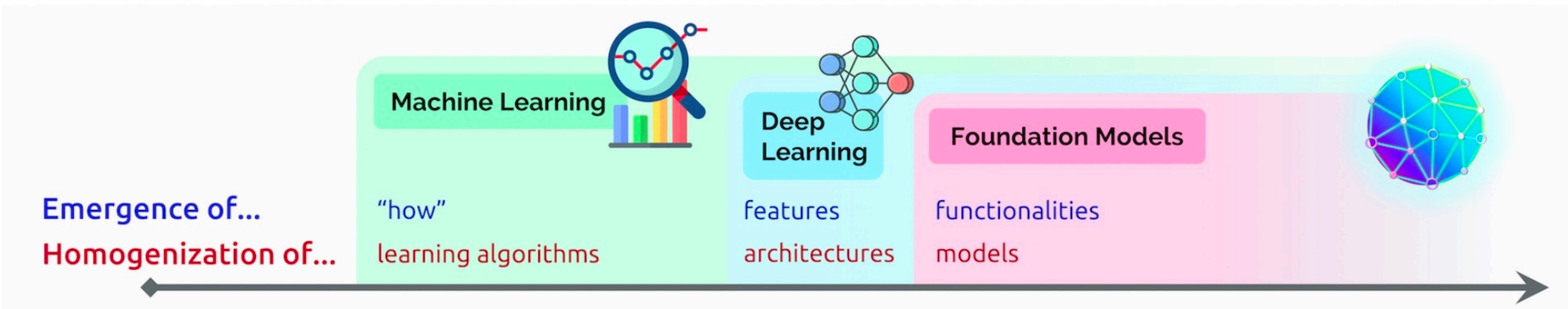
Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

*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

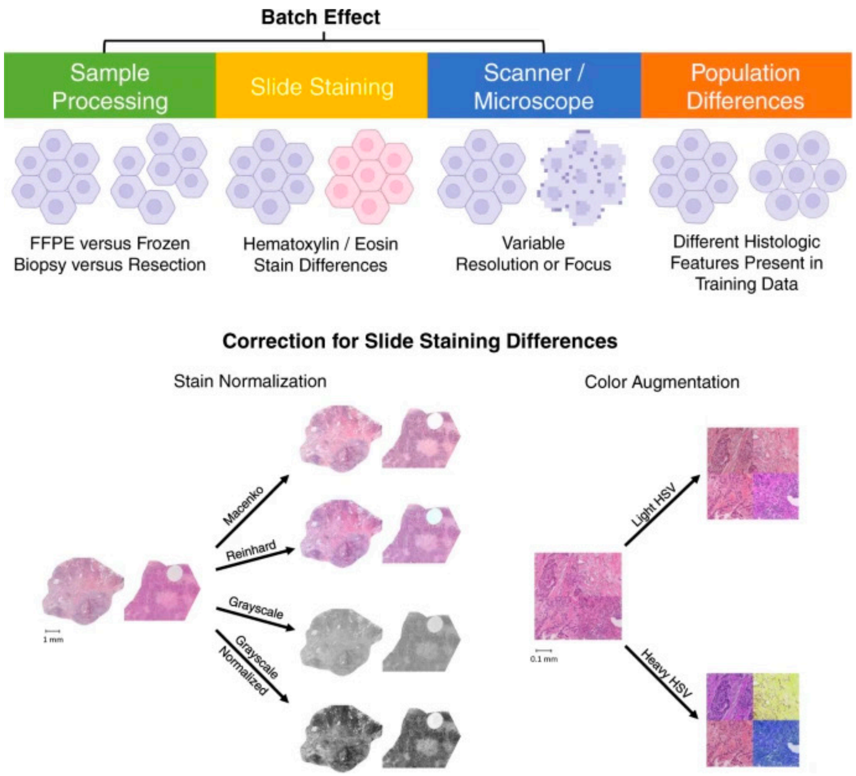



The story of AI has been one of increasing emergence and homogenization. With the introduction of machine learning, how a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced 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 Reasons for Data Gaps

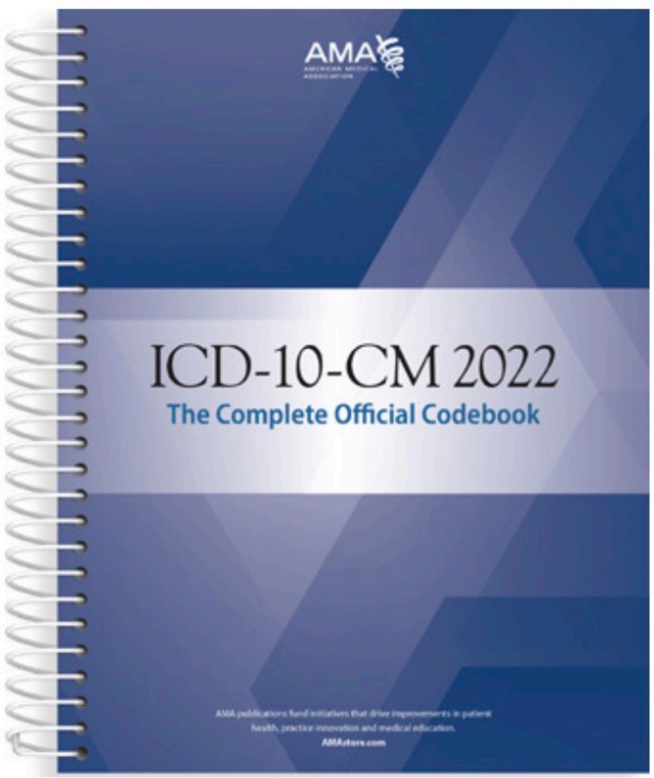
# DG1: Are there (data) source differences that creates data gaps?




- **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 process, scanner / microscope, etc. are the most important components that ML/DL models identify.

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. Grossman1, 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.

# DG2: Are there data coding & data curation differences that create data gaps?




- Think of coding as a map from a analog event or measurement to a numeric / symbolic representation of it.
- For example, 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.

# DG3: Are there data contributor differences that create a data gaps?

| | |
|---|---|
| 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. |

Publicly available medical data can be quite different than data within a healthcare system.

# DG4: Are there differences in the data elements in the training data versus the data elements in the real-world data?

| Training: | Many features + curated outcome |
|---|---|

**vs**

| Deployed: | Some features |
|---|---|

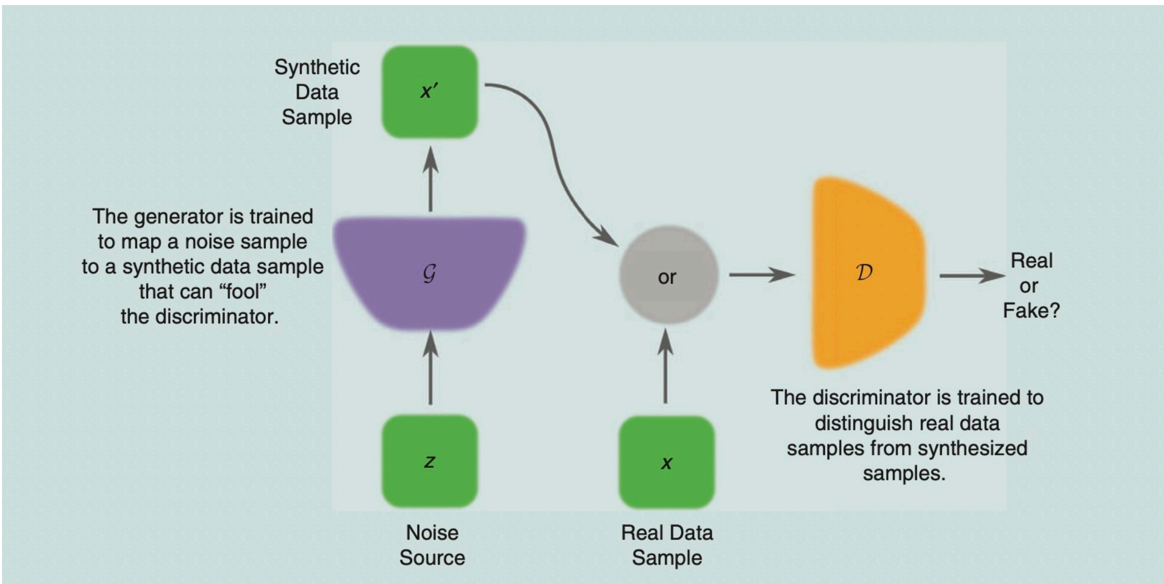
# 5. Five Techniques to Close the Data Gap

# 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)



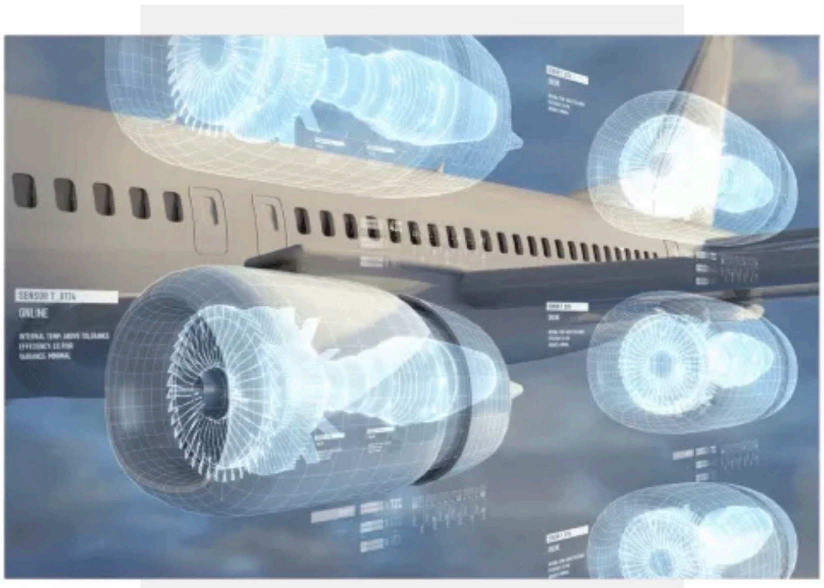

Two models that are learned during the training process for a GAN:

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.*

Example: each Tesla car has a digital twin.

Sensors in the car provide real time data to the digital twin.

*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

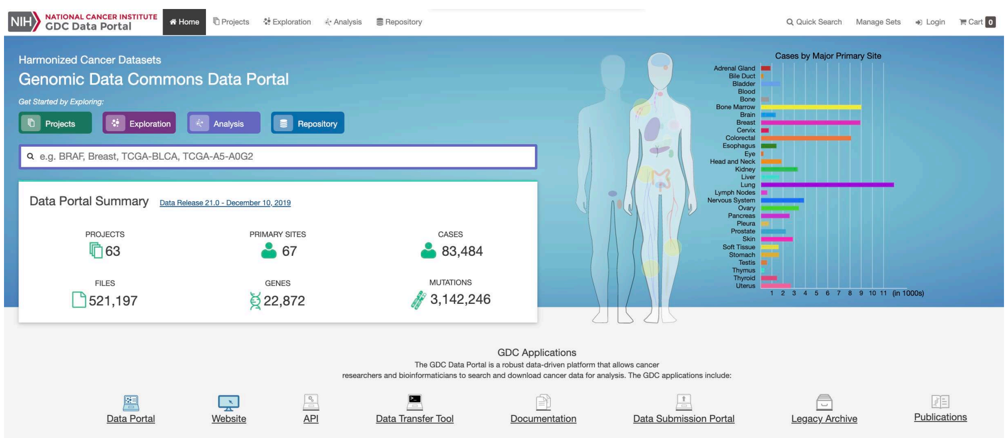

“In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn.”

Andrew Ng, CEO & Founder, Landing AI

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

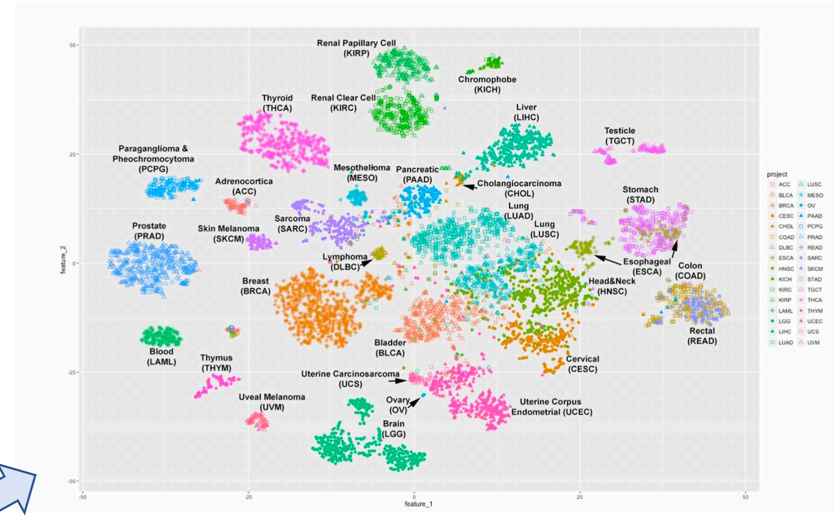
# 5. Data Commons

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




The NCI Genomic Data Commons (GDC) is one of the world's largest collections of curated cancer genomics data.

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




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.

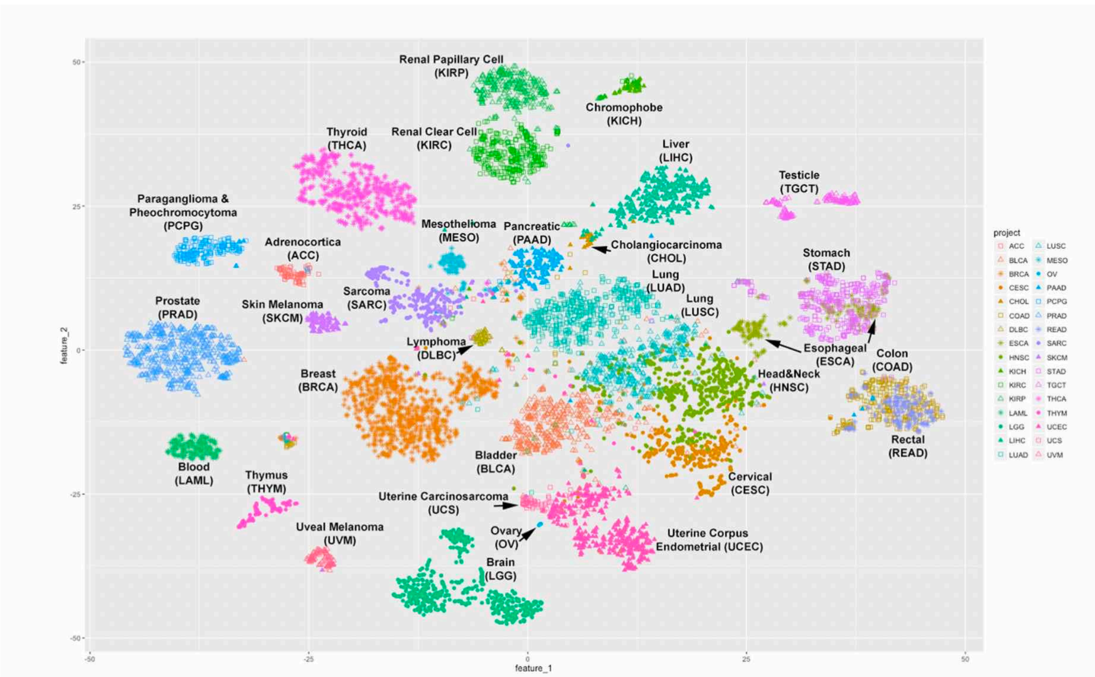
5. Why are some AI and predictive models so much harder than other predictive models?

| Meta | Google | Large Language Models | COVID |
|---|---|---|---|
| Generate your own data | Build your own data collection system | Use third party data from a single source | Use third party data from a multiple sources |
| platform operators | crawl the web | satellite data | COVID data |

# Why are some models harder than others?

1. There are many distinct subtypes, each of which requires a different model.
2. There are very few positive cases to observe and label.
3. There are many weak interacting contributing causes vs a few strong contributing causes.
4. Curating data at the level required for a good model requires working through many, many edge cases.
5. Modeling proximate causes are easier than identifying and modeling root causes.

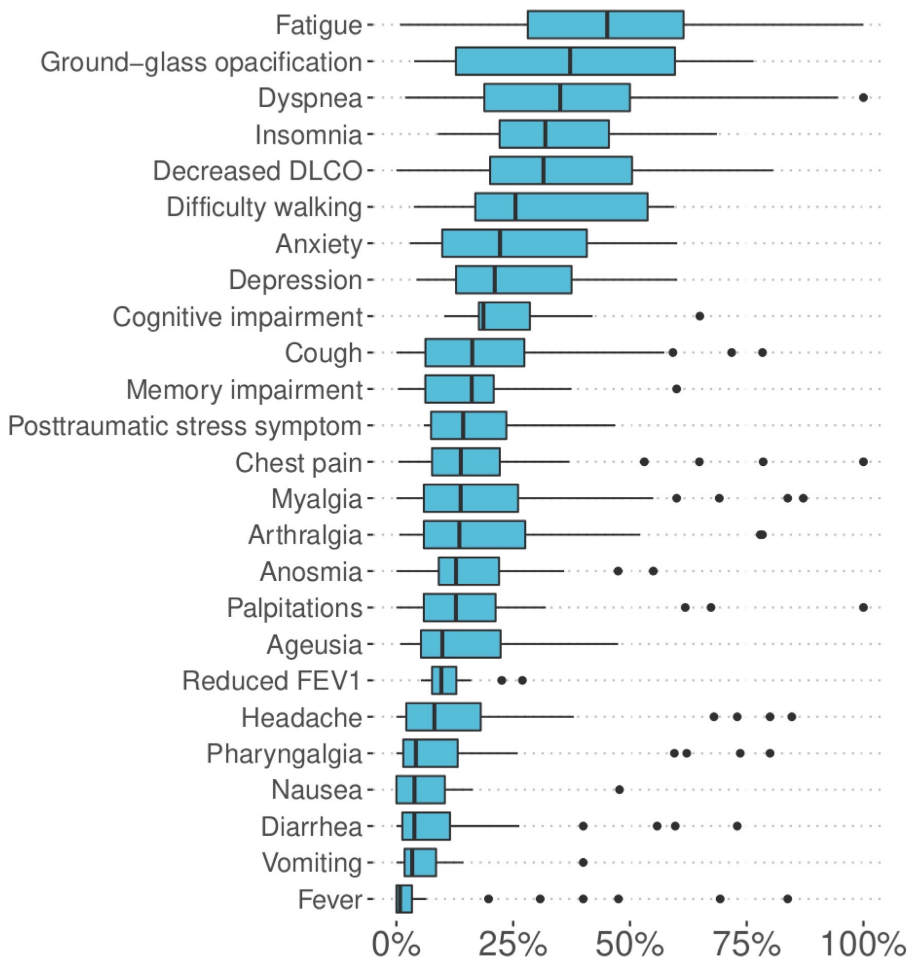
# Many Distinct Subtypes, which Require Different Models




- Cancer is not a single disease but has many different sub-types
- 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

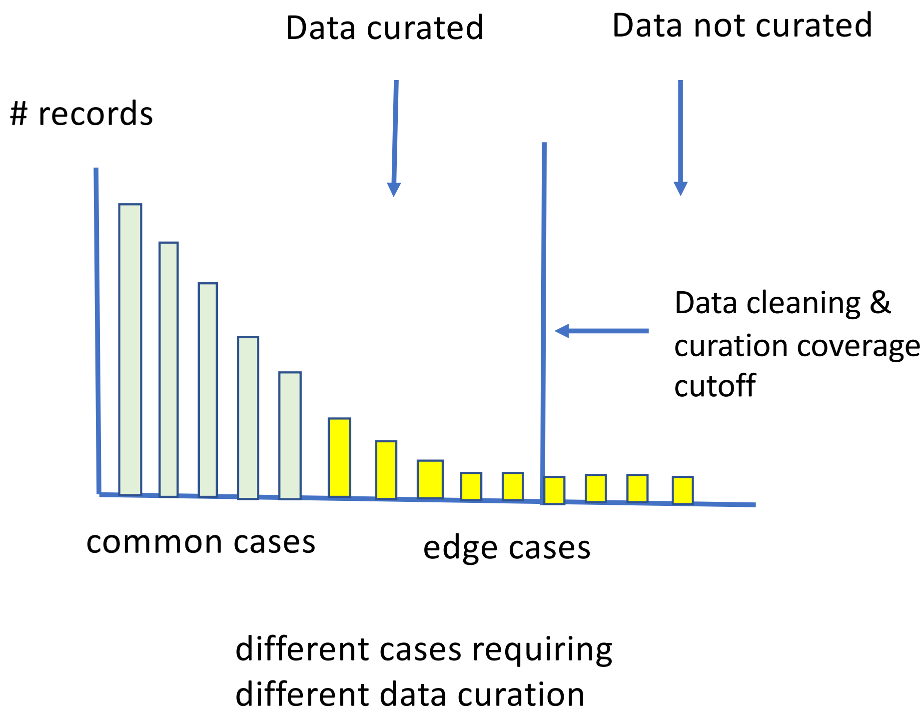



- Long covid is a complex disease that is still being characterized.
- The plot shows what symptoms are present

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 Long Tail of Data Curation

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




- It can be very expensive to clean and curate all the data and generally there is a cut off and a number of edge cases are not carefully curated.
- As a consequence, the model performs poorly on these edge cases.

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

# 5. Summary

# Three Rules

1. Build a simple model over data you understand as a baseline model and a sanity check.
2. Red team any large third-party data you use to identify any data gaps and any opportunities for reputational risk.
3. Use foundation models for applications that are forgiving.

# Summary

1. Most AI and machine learning models fail before achieving sustainable deployment that delivers value.
2. The biggest reason is a data gap and not getting enough well curated data to build a useful model.
3. There are some good techniques (e.g. GANs, digital twins, "hand engineering," data commons), but it still requires deep domain knowledge and experience to build useful models in many domains.

For more information:


Robert L. Grossman
rgrossman.com
@BobGrossman


analyticstrategy.com

ctds.uchicago.edu

**Abstract:** Although foundational models have changed the way we do large-scale AI and machine learning of text and images, developing and deploying AI models that provide value and limit bias is still quite difficult in many application areas due to the lack of suitable data. 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 a Partner at Analytic Strategy Partners LLC, which he founded in 2016. Prior to that, from 2002 to 2015, he was the Founder and Managing Partner at Open Data Group, which built and deployed AI and machine learning models over big data in financial services, insurance, healthcare and IoT. He is also 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 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 a Director of the Open Commons Consortium, a not-for-profit that develops and operates data commons to support research in science, medicine, health care, and the environment.