AI AND DATA-DRIVEN COMPETITION:

WHAT IS THE RELEVANT PRODUCT MARKET FOR DATA?

ELENA PONTE[†]

ABSTRACT

Data is AI, and AI is only as capable as the data on which it is trained. The dynamic and multifaceted nature of training data requires a nuanced approach to its classification, especially in the context of antitrust analysis in AI and AI-adjacent markets. Not all training data functions the same, and distinctions in a dataset's many attributes—such as purpose, source, structure, processing, generation, sensitivity, and ownership—have important legal and competitive implications. Training data for one AI application—say, medical diagnostics—is vastly different from training data used for generative art or robotics, and their substitutability varies accordingly.

As a result, training data is too broad a category to serve as a coherent product market in antitrust analysis. A more granular approach is warranted for this key input in the development of AI. When defining the relevant product market, a one-size-fits-all approach to data may fail to capture important distinctions that impact the competitive landscape.

This Article examines the evolving data market and the rise of technologies like synthetic data. It proposes a framework antitrust practitioners may use to categorize training data based on its attributes and use cases. In addition to definitions and explanations, it aims to provide a practical tool for antitrust practitioners looking to understand how different types of data may influence competition in AI-driven industries. The proposed framework should equip antitrust practitioners and regulators working in this rapidly advancing field with analytical tools to address the complexities of training data in AI development.

[†] Intellectual Property Law and Policy LL.M. University of Washington, co-founder of the first Y Combinator-backed synthetic data startup. The views expressed herein are entirely those of the author do not necessarily reflect the views of my employer.

I. THE PRIMACY OF DATA IN AN AI WORLD 252
A. Training Data is a Key Input of AI 254
B. Key Trends for the Future of Training Data 255
1. Synthetic Data Revolution 257

II. THE RELEVANT PRODUCT MARKET 259

III. CLASSIFICATION OF THE TRAINING DATA MARKET 261
A. Based on Purpose 261
1. The Deep Learning Workflow 261
2. Attributes of Data Intended for Different Purposes 263
B. Based on Source 266
1. Collection Location 267
2. Collection Method 268
3. Collecting Entity 269
C. Based on Structure 271
D. Based on Cleaning and Processing Applied 272
E. Based on Synthetic Elements 273
1. Methods and Technologies to Synthesize Data 274
2. Advantages of Synthetic Data over Real-World Data 277
3. Limitations of Synthetic Data 287
F. Based on Sensitivity 290
G. Based on Ownership and Accessibility 292

IV. CONCLUSION 293

I. THE PRIMACY OF DATA IN AN AI WORLD

AI is only as capable as the data on which it is trained and fine-tuned.[1] Different AI applications require different types of data, with different attributes, at different junctures of the training workflow.[2] At the time of writing, generative AI has become ubiquitous across many different industries.[3] The future of AI will see AI integrated into human society, from AI agents[4] in the palms of our hands to AI-enabled robotics[5] accelerating the means of production. AI development will require increasingly task-specific, customized datasets gathered from multiple environments and leveraged in different ways across the training process.[6] As a critical input, data will become one of the most important differentiators between competing AI solutions in the marketplace.

"Data" alone is too broad and dynamic to be considered a product market. To undertake a successful antitrust analysis in an AI or AI-adjacent market, the datasets used to train the AI must be categorized and classified according to their particular attributes. Not all data is the same, and the distinctions in training data beget important differences. Regulation that promotes competition when the data used to train AI is real-world data scraped from the internet, for example, could have the opposite effect when models are trained on AI-generated synthetic data. Successful antitrust practice in an AI or AI-adjacent market needs a detailed classification that defines and accounts for the different attributes of the data on which the AI is trained.

While enforcers recognize that data confers competitive advantages,[7] current antitrust frameworks lack the vocabulary to define data markets with the precision necessary for rigorous competitive analysis.[8] Scholars have long debated whether a "market" for data exists at all when

---

[1] David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 677 (2017).

[2] *See id.* at 684; *see also* Maximilian Böther et al., *Modyn: A Platform for Model Training on Dynamic Datasets* (Dec. 11, 2023) (unpublished manuscript), https://arxiv.org/html/2312.06254v1/ [https://perma.cc/A4VV-BDPR].

[3] Generative AI refers to models like OpenAI's GPT for text and DALL-E for images, Google's Gemini, Anthropic's Claude, Amazon's AWS CodeWhisperer, or Stability AI's open-source Stable Diffusion. These models generate—hence the term generative AI—text, images, music, and more by learning from vast datasets. Generative models are excellent at synthesizing information and generating creative outputs but are largely reactive: they respond to inputs without long-term decision-making, planning, or goal-setting capabilities.

[4] Agentic AI makes decisions and takes actions autonomously in real time. Jensen Huang, CEO, NVIDIA, Keynote Address at the Consumer Electronics Show (Jan. 6, 2025) (recording available at https://www.nvidia.com/en-us/on-demand/session/ces25-keynote) (heralding the next iterations of AI, from our generative AI of today to agentic AI in 2025 and physical AI after that).

[5] Physical AI refers to autonomous machines like robots and self-driving cars that can perceive, understand, and perform complex actions. *See id*.

[6] *See* Khaled El Emam, *Accelerating AI with Synthetic Data: Generating Data for AI Projects*, O'REILLY MEDIA 5 (2020), https://www.nvidia.com/content/dam/en-zz/Solutions/deep-learning/resources/accelerating-ai-with-synthetic-data-ebook/accelerating-ai-with-synthetic-data-nvidia_web.pdf [https://perma.cc/9B6P-SH3F].

[7] When the European Commission reviewed Google's acquisition of Fitbit, regulators examined whether access to Fitbit's health and fitness data could strengthen Google's position in online advertising markets and digital healthcare. Commission Decision of Dec. 17, 2020, Case M.9660, *Google/Fitbit*, at § 7.2.1, 9.3.2, 9.3.3 (discussing data combination concerns in online advertising). The Commission imposed data silo commitments to prevent Google from using Fitbit data for advertising purposes, effectively acknowledging data's competitive significance. *See id.* at § 10.4.1.1 (describing the "Ads Commitment" requiring Google to maintain Fitbit health and fitness data in a separate data silo). Yet the Commission analyzed these concerns within the framework of existing product markets, online advertising and wearable devices, rather than defining a market for health data itself.

[8] *See* Cristina Caffarra & Tommaso Valletti, *Google/Fitbit Review: Privacy IS a Competition Issue*, VOXEU.ORG (Mar. 4, 2020), https://cepr.org/voxeu/columns/googlefitbit-review-privacy-competition-issue (arguing the Commission should have taken data

most data has not traditionally been traded as a stand-alone product.[9] Some have argued for recognizing "data markets" more explicitly,[10] while others caution that data's multifaceted nature resists treatment as a traditional product.[11] This theoretical uncertainty has translated into enforcement challenges, where data often appears as a factor within broader market analyses rather than as the market itself.

The data market is growing and changing just as fast as the AI technologies it enables. For example, synthetic, or artificially-created, data is a prominent new technology that is being used by data scientists in supervised and unsupervised settings to overcome limitations of real-world data.[12] Synthetic medical records used to train AI in the healthcare field, for instance, may reduce the risk of violating patient privacy by allowing the development of models trained on artificial medical records that are nearly indistinguishable from real ones.[13] Synthetic data can be used in conjunction with or as a substitute for real-world data, and it can be classified using similar criteria.[14] Its use, however, may have important implications for the competitive analysis.

This Article proposes a framework that antitrust practitioners may use to categorize training data based on its attributes and use cases. It provides an expansive view of how training data—that is, data that is provided to a model before the model is deployed—can be subdivided and classified. Data can be classified based on several different factors: purpose, structure, source, and sensitivity of the data. Datasets can also be defined by how they were generated, and whether the data is preprocessed or cleaned. Still another facet of data is the ownership of the data, which overlaps with concepts like the source or sensitivity in some contexts.

The purpose of the proposed framework is to provide a practical tool for antitrust practitioners looking to understand how different types of data may influence competition in AI-driven industries. Importantly, this framework aims to categorize datasets based on their attributes and characteristics, going beyond a simplistic description of what the data is. For example, it is more useful to understand that a structured financial dataset is data provided for fine-tuning, that it is real-world first-party data collected at the edge, and that it is highly sensitive. The

---

combination concerns more seriously in market definition); *see* Simon Vande Walle, *The European Commission's Approval of Google/Fitbit—A Case Note and Comment*, 12 J. EUR. COMPETITION L. & PRAC. 685, 686–88 (2021) (noting tensions in how the Commission addressed data-related theories of harm).

[9] *See* Ingae Graef, *Market Definition and Market Power in Data: The Case of Online Platforms*, 38 WORLD COMPETITION 473, 503 (Dec. 2015) (arguing that most online platforms do not trade data as a stand-alone product and therefore no supply and demand exists for a relevant product market for data under current competition law standards).

[10] *See* Pamela Jones Harbour & Tara Isa Koslov, *Section 2 in a Web 2.0 World: An Expanded Vision of Relevant Product Markets*, 76 ANTITRUST L.J. 769, 794–97 (2010) (proposing that competition authorities should consider "data markets" as relevant product markets in appropriate cases).

[11] *See* Gal & Lynskey, *supra* note 12, at 1115 (discussing difficulties in defining markets for data given its non-rivalrous nature and multiple uses).

[12] In particular, synthetic data is hailed as a solution, to the data scarcity problem. *See* Ilya Sutskever, *Sequence to Sequence Learning with Neural Networks: What a Decade, Test of Time Award Lecture*, Conference on Neural Information Processing Systems, YOUTUBE (Dec. 15, 2024), https://www.youtube.com/watch?v=sJE8qZmQWGU. *See also* SERGEY NIKOLENKO, SYNTHETIC DATA FOR DEEP LEARNING 12 (2021) ("Many problems of modern AI come down to insufficient data").

[13] *See* Mauro Giuffre & Dennis L. Shung, *Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy*, NPJ DIGIT. MED. 6, 186 (2023), https://www.nature.com/articles/s41746-023-00927-3 [https://perma.cc/L6A6-L4G7].

[14] *See id*.

characteristics of data inform how it is used and thus how it can be substituted in the marketplace.

The proposed framework should equip antitrust practitioners and regulators working in this rapidly advancing field with analytical tools to address the complexities of training data in AI development. By proposing a common framework for classifying training data, one that can be useful even as technologies like synthetic data enter the training data market, this Article aims to assist antitrust practitioners and regulators in analyzing AI-influenced industries for which training data is an essential input.

A. Training Data is a Key Input of AI

Before proceeding, some definitions may be helpful. AI is not one product—it is the overarching goal: to build systems that exhibit intelligence. Yann LeCun, a renowned AI researcher and pioneer in deep learning, defines AI as the creation of systems or machines capable of performing tasks that would normally require human intelligence, such as reasoning, understanding natural language, perception, and decision-making.[15] The idea behind AI is not radical; it is the recent advances in effectiveness and functionality that truly are incredible.[16]

To speak accurately about how AI works, a basic understanding of machine learning (ML) is helpful. Put simply, ML is a method for training AI. ML describes the process by which an AI system teaches itself—or "learns"—from data to improve its own performance at a given task. [17] In ML, the system identifies patterns in data and, rather than being programmed to return certain results, uses its recognition of those patterns and inferences from those patterns to make predictions or decisions.[18] LeCun highlights ML as one of the most successful and transformative techniques within AI, particularly with the advent of deep learning, which uses artificial neural networks to solve complex tasks.[19] ML, and specifically deep learning, dominates current AI development.

Training data is the data provided to the model for it to "learn" from. Training data is a key input of AI because ML and deep learning rely on vast amounts of training data to train models effectively.[20] The quality and the quantity of training data are important in determining the effectiveness of a model. For instance, at present, self-driving cars require millions of miles of driving data to identify road signs, detect pedestrians, and navigate diverse traffic conditions.[21] Without ample, accurate data, AI systems may fail to generalize effectively, leading to errors, incorrect predictions, and otherwise unreliable outcomes. Data is often the input leveraged to

---

[15] Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436, 444 (2015), https://www.nature.com/articles/nature14539 [https://perma.cc/NWU3-HEBC].

[16] N.Y. Times Events, *The Next Frontier: Sam Altman on the Future of A.I. and Society*, YOUTUBE (Dec. 4, 2024), https://youtu.be/tn0XpTAD_8Q?si=SKLMH9d-IFKhU3WI (wherein Sam Altman discusses the transformative power of AI that humanity has been able to harness).

[17] LeCun, Bengio & Hinton, *supra* note 16.

[18] *Id.*

[19] *Id.*

[20] *Id. See also* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017).

[21] *See* Alexey Dosovitskiy et al., *CARLA: An Open Urban Driving Simulator*, 78 PROC. 1ST ANN. CONF. ON ROBOT LEARNING, 1, 4–6 (2017), https://proceedings.mlr.press/v78/dosovitskiy17a.html [https://perma.cc/4MP2-WYSJ].

enable accelerated AI capabilities. Large language models (LLMs)[22] like OpenAI's GPT, for example, improve their performance as they access more diverse and expansive datasets. The hunger for quality data only grows with advances in AI. Future breakthroughs, including domain-specific AI or more human-like reasoning, will depend heavily on the availability of clean, rich, and well-labeled data, driving innovation and making AI systems safer and more capable.[23]

B. Key Trends for the Future of Training Data

Although the training data market is constantly evolving, a few key trends are likely to persist in the coming years. First is the move from centralized to decentralized data sources and data processing. Historically, training data was aggregated and processed from centralized sources—large data centers, public datasets, or single institutional repositories.[24] Increasingly, AI leverages decentralized data sources.[25] Edge AI techniques allow models to be trained on data from multiple distributed devices without consolidating it into a single location.[26] These approaches process data closer to the source, offering gains in speed, efficiency, and scalability.[27]

Another trend is a shift towards self-supervised and unsupervised learning methodologies, underscored by DeepSeek's meteoric rise to U.S. public awareness in late January 2025.[28] Training data serves as the foundational input for various ML paradigms, each of which uses this data

---

[22] Modern performance of LLMs relies on transformer architectures. Though, interestingly, transformers in the modern sense of ML models are the transformer architecture introduced in the seminal paper "Attention is All You Need" published in 2017 by researchers at Google Brain and Google Research, Richard Feyman, legendary physicist who died in 1988, long before the advent of such AI technologies, extensively discussed concepts related to electromagnetism and electrical transformers that are eerily similar. *See* RICHARD P. FEYNMAN, ROBERT B. LEIGHTON & MATTHEW SANDS, THE FEYNMAN LECTURES ON PHYSICS (Addison-Wesley, 1964).

[23] Jensen Huang, CEO of NVIDIA, heralds the next iterations of AI, from our generative AI of today to the agentic AI (AI systems that will be proactively capable of autonomous goal setting, decision-making, and execution) and physical AI (embodied AI, that is, systems that are not only intelligent by also capable of interacting dynamically with the real work in meaningful, goal-oriented ways) of tomorrow. *See e.g.*, John Werner, *Jensen Huang at CES: Agentic AI, Big Hardware, and More*, FORBES (Jan. 9, 2025), https://www.forbes.com/sites/johnwerner/2025/01/09/jensen-huang-at-ces-agentic-ai-big-hardware-and-more/ [https://perma.cc/2JW7-9M33]; Sean Kinney, NVIDIA CEO *Jensen Huang sees agentic AI and AI-enabled robotics as a multi-trillion dollar opportunity* RCRWIRELESS NEWS (Jan. 10, 2025), https://www.rcrwireless.com/20250110/ai-ml/nvidia-charts-a-course-from-agentic-ai-to-physical-ai [https://perma.cc/ZZ5E-WES4].

[24] ANTON DOLHOPOLOV, ARNAUD CASTELLTORT & ANNE LAURENT, TRICK OR TREAT: CENTRALIZED DATA LAKE VS DECENTRALIZED DATA MESH, International Conference on Management of Digital 303–16 (Springer Nature Switzerland, 2023), https://hal.umontpellier.fr/hal-04156120/document [https://perma.cc/8J9X-4VCX] (noting that most of the existing data architecture exists as 'data lakes', a centralized solution, that brings issues like "data silos, modules inter-dependencies, and long product delivery cycles", but there is a shift towards a decentralized model of 'data meshes.').

[25] *Id.*

[26] *See generally Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning*, GARTNER (Aug. 1, 2023), https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning [https://perma.cc/C8WQ-A8BY] ("Gartner predicts that more than 55% of all data analysis by deep neural networks will occur at the point of capture in an edge system by 2025, up from less than 10% in 2021.").

[27] Tobias Meuser et al., *Revisiting Edge AI: Opportunities and Challenges*, 28 IEEE INTERNET COMPUTING 49, 49–59 (2024) https://eprints.cs.univie.ac.at/8163/1/Revisiting_Edge_AI_Opportunities_and_Challenges.pdf [https://perma.cc/U9K3-E27X] (defining edge devices by "their closeness to the user and the locality of processed data" and explaining that "[e]dge intelligence represents a further paradigm shift from edge computing, integrating AI to enhance the processing capabilities at the edge of the network. This integration further reduces latency and alleviates the bandwidth demand on central servers, while also providing additional benefits, such as enhanced privacy due to distributed approaches for ML like federated learning and improved resilience due to local autonomy and decentralized control.").

[28] Eric Schmidt & Dhaval Adjodah, *Will China's Open-Source AI End U.S. Supremacy in the Field?*, WASH. POST (Jan. 28, 2025), https://www.washingtonpost.com/opinions/2025/01/28/china-deekseek-ai-us-supremacy/ [https://perma.cc/RV82-9E9Z] ("[I]n recent days, even those on the cutting edge of AI research were taken by surprise – by [DeepSeek]… DeepSeek demonstrated that large model training could be made more efficient by bypassing the traditional supervised fine-tuning stage. They even created R1-Zero, a model that omits this step in AI training, to challenge the research community's assumptions about fine-tuning's indispensability.").

differently to learn and make decisions. In self-supervised or unsupervised settings, training data encompasses the unlabeled data that provides the initial knowledge base for ML models, enabling them to learn general patterns and relationships.[29] It includes, for example, large-scale text corpora for large language models (LLMs) like OpenAI's GPT. In the supervised learning approach, models are trained on task-specific datasets used to adapt, or fine-tune, pre-trained models for particular applications.[30] These datasets develop model performance on domain-specific tasks by providing high-quality, targeted examples.[31] Generally, this data must be labeled and annotated and be of high quality.[32] Datasets used in supervised settings, for instance in post-training fine-tuning or validation, are smaller than pre-training datasets.[33] In reinforcement learning (RL), models learn by interacting with an environment, typically in simulation, to achieve a goal.[34] In complex applications like robotics, simulated environments act as synthetic data generators: the data is produced dynamically through continuous interaction, and its content is highly dependent on the model's exploration strategy.[35]

A significant breakthrough, or "Aha moment," for DeepSeek was the development of their R1-zero model, which was trained using pure reinforcement learning (RL) without relying on supervised fine-tuning (SFT).[36] This approach enabled the model to self-improve through interaction and feedback, reducing dependence on large, labeled datasets. The success of DeepSeek's R1 model exemplifies the potential of unsupervised learning techniques, particularly RL, in advancing AI capabilities efficiently and cost-effectively. As it becomes possible to successfully train models following different learning paradigms, data scientists require different types of training data for each different method they employ.

Dynamic and continuous data updates, features that, while not new, are increasingly possible thanks to more sophisticated data aggregation processes and fluid data architectures.[37] Traditionally, training data was often collected as a static snapshot—prepared once, used until a new dataset was created.[38] Training data is now more dynamic, reflecting a shift to more adaptable

---

[29] *See What's the Difference Between Supervised and Unsupervised Learning?,* AWS, https://aws.amazon.com/compare/the-difference-between-machine-learning-supervised-and-unsupervised/ [https://perma.cc/Y5HJ-BP3J].

[30] Ilya Sutskever, *An Observation on Generalization*, YOUTUBE (Aug. 14, 2023), https://www.youtube.com/watch?v=AKMuA_TVz3A [https://perma.cc/U5GQ-PHDL].

[31] *Id.*

[32] *Id.*

[33] *Id.*

[34] RICHARD S. SUTTON & ANDREW G. BARTO, REINFORCEMENT LEARNING: AN INTRODUCTION 1 (2015).

[35] *See* Charles Xu et al., *RLDG: Robotic Generalist Policy Distillation via Reinforcement Learning* (Dec. 13, 2024) (unpublished manuscript), https://arxiv.org/pdf/2412.09858 [https://perma.cc/86NS-JMNB].

[36] DeepSeek-AI, *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* 5 (Jan. 4, 2026) (unpublished manuscript), https://arxiv.org/pdf/2501.12948 [https://perma.cc/DCR3-YYTQ] (In the "aha moment," DeepSeek-R1-Zero learns to allocate more thinking time to a problem by reevaluating its initial approach. "[This moment] underscores the power and beauty of RL: rather than explicitly teaching the model how to solve a problem, we simply provide it with the right incentives, and it autonomously develops advanced problem-solving strategies.").

[37] DOLHOPOLOV, CASTELLTORT & LAURENT , *supra* note 25, at 303–16.

[38] SJ Kline, *Static Models vs. Online Learning in Artificial Intelligence: Key Differences and Use Cases*, TECHINY (Apr. 12, 2025), https://techiny.com/artificial-intelligence/static-model-vs-online-learning [https://perma.cc/ZU9L-K63A].

and evolving data sources, enhancing models' ability to handle real-world scenarios.[39] Dynamic data allows for continuous learning and adaptation, enabling models to remain relevant as new information becomes available.[40] Systems are updated regularly with fresh data to ensure models remain relevant in rapidly changing domains (e.g., news, financial markets, or real-time sensor data).[41] Continuous or incremental learning pipelines are increasingly common, enabling models to adapt without full retraining from scratch.[42]

1. Synthetic Data Revolution

Finally, and arguably most crucially, is the advent of synthetic data. Up to this point, most AI has been trained on traditional, real-world datasets.[43] Traditional training data comes from proprietary datasets gathered by companies from users of their products or from publicly available information, such as in datasets like Common Crawl,[44] an open repository of data scraped from the web.

Synthetic data is artificial data that can mimic these traditional, real-world data sources.[45] Synthetic data refers to data artificially generated through simulations, procedural methods, or generative models rather than collected from the real-world.[46] Synthetic data models the real world and typically has similar statistical properties to the relevant real-world data.[47] As models demand more diverse, higher-quality, and larger-scale datasets, synthetic data has become a common supplement, if not the sole source of training data.[48] Synthetic data can either supplement or replace traditional datasets with artificially created data that serves to train AI as well as, or better

---

[39] Nils Baumann et al., *Dynamic Data Management for Continuous Retraining*, PROC. OF MODELS 2022, WORKSHOP MDS INTEL. 1 (Oct. 2022), https://www.se-rwth.de/publications/Dynamic-Data-Management-for-Continuous-Retraining.pdf [https://perma.cc/AH9J-YUM5] (noting that " the training database continuously grows even after the first version of an ML model was designed, trained, and even deployed," this happens, for example with "fresh data" including updated user interactions and continuous sensor data.).

[40] Maximilian Böther et al., *supra* note 3 (arguing that "Models must incorporate this evolving training data to improve generalization, adapt to potential distribution shifts, and adhere to privacy regulations.").

[41] *Id.*

[42] *Id. (*Modyn's ML pipeline abstraction that enables users to continuously train a model on a growing dataset).

[43] LeCun, Bengio & Hinton, *supra* note 16, at 436–44.

[44] Common Crawl, https://commoncrawl.org/ (last visited Feb. 22, 2026) [https://perma.cc/A6F4-9Q7L] (stating that its corpus contains "petabytes of data" collected since 2008, including raw web page data, metadata extracts, and text extracts).

[45] *See* Rina Caballar, *What Is Synthetic Data?*, INT'L BUS. MACH. (last visited Feb. 22, 2026), https://www.ibm.com/think/topics/synthetic-data [https://perma.cc/K352-HCQ9].

[46] *See id.*

[47] *See* El Emam, *supra* note 7, at 2–4.

[48] *See* Adam Zewe, *3 Questions: The Pros and Cons of Synthetic Data in AI*, MASS. INT. OF TECH. COLL. OF COMPUTING (Sep. 3, 2025), https://computing.mit.edu/news/3-questions-the-pros-and-cons-of-synthetic-data-in-ai/ [https://perma.cc/T44N-YEL3]; Matthew Misele, *What's Synthetic Data and Why It's Important for AI Development*, MOVEWORKS BLOG (Sep. 20, 2024), https://www.moveworks.com/us/en/resources/blog/synthetic-data-for-ai-development [https://perma.cc/2LZ2-HRX8]. Ilya Sutskever, a renowned computer scientist who works at the frontier of AI and has made several seminal contributions to the field of deep learning, called data the "fossil fuel" of AI and leans on synthetic data as the solution to real-world high-quality data scarcity. *See* Automation with Wilt: Software & AI, *Ilya Sutskever Full Talk at NeurIPS 2024 Vancouver 15/12/2024*, YOUTUBE (Dec. 14, 2024), https://www.youtube.com/watch?v=sJE8qZmQWGU [https://perma.cc/CT6Q-H5S8]. Similarly, Sundar Pichai, CEO of Google, has remarked that the era of rapid advancements in AI will be more challenging if it is solely dependent on real-world training data. Hugh Langley, *Google CEO Sundar Pichai Says the Era of Rapid AI Advancements Is Hitting a Wall as the 'Low-Hanging Fruit Is Gone,'* BUS. INSIDER (Dec. 5, 2024), https://www.businessinsider.com/google-ceo-sundar-pichai-ai-wall-progress-harder-2024-12 [https://perma.cc/6SHT-47TF].

than, traditional datasets alone.[49]

Synthetic data is a technological revolution in training data.[50] Increasingly, developers are using simulated environments,[51] digital twins,[52] artificial medical records,[53] and even synthesized speech[54] to train more capable, task-specific ML models. Developers can design synthetic datasets to eliminate biases, improve diversity, and target specific use cases that real-world data often cannot address.[55] This intentional creation of data allows developers to mitigate risks such as perpetuating harmful stereotypes or exposing sensitive information.[56]

Importantly, synthetic data can lead to new and more accurate learnings.[57] This is what makes synthetic data a zero-to-one technology.[58] Take for example, AlphaFold, a 2020 Google DeepMind model that uses computer simulations based on background knowledge about proteins "to accurately and efficiently predict the 3D shape of an unknown protein using just its DNA or RNA source code."[59] AlphaFold and the database of over two hundred million predicted protein structures so far are a monumental scientific breakthrough. Another example of synthetic data's usefulness comes from the healthcare field, where the U.S. Food and Drug Administration is working with developers to explore the use of virtual patients in medical drug and device

---

[49] *See What Is Synthetic Data: Types, Techniques, Benefits & Use Cases*, QUINNOX BLOG (July 30, 2025), https://www.quinnox.com/blogs/synthetic-data/ [https://perma.cc/EN5X-WKLQ].

[50] *See 2022 10 Breakthrough Technologies,* MASS. INST. OF TECH. REV. (Feb. 23, 2022), https://www.technologyreview.com/2022/02/23/1045416/10-breakthrough-technologies-2022/#synthetic-data-for-ai [https://perma.cc/9ZCV-8QAE] (identifying synthetic data as one of the top 10 breakthrough technologies of 2022); Emma Keen, *Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning*, GARTNER (Aug. 1, 2023), https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning [https://perma.cc/UK3A-BDYM] ("By 2024, Gartner predicts 60% of data for AI will be synthetic to simulate reality, future scenarios and derisk AI, up from 1% in 2021.").

[51] Both Amazon and NVIDIA use simulation data to train robots in warehouses to recognize objects of different shapes and sizes in different conditions. *See* El Emam, *supra* note 7, at 34–36.

[52] An application in smart cities is Singapore's Virtual Singapore Project, an initiative that creates a dynamic 3D digital twin of the entire city to simulate urban planning, infrastructure management, and disaster response scenarios. *See 5 Things to Know About Virtual Singapore*, GOV'T TECH. AGENCY OF SING. (Mar. 28, 2017), https://www.tech.gov.sg/technews/5-things-to-know-about-virtual-singapore/ [https://perma.cc/KQ3U-436V].

[53] Synthetic images were used to supplement the training of a melanoma detection model to enable its accuracy on darker skin tones, since the collected real-world data mostly contained lighter skin tone images. *See* Timo Kohlberger & Yuan Liu, *Generating Diverse Synthetic Medical Image Data for Training Machine Learning Models*, GOOGLE RSCH. BLOG (Feb. 19, 2020), https://research.google/blog/generating-diverse-synthetic-medical-image-data-for-training-machine-learning-models/ [https://perma.cc/ZAP4-SC8H].

[54] Amazon used synthesized speech to train its digital assistant, Alexa, in Hindi, U.S. Spanish, and Brazilian Portuguese, languages for which Amazon faced a shortage of collected data. *See* Janet Slifka, *Tools for Generating Synthetic Data Helped Bootstrap Alexa's New-Language Releases*, AMAZON SCI. (Oct. 11, 2019), https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases [https://perma.cc/38PE-3TQA].

[55] *See* Misele, *supra* note 49.

[56] *See* Laura Cabrera et al., *Tempered Enthusiam by Interviewed Experts for Synthetic Data and ELSI Checklists for AI in Medicine*, 5 AI & ETHICS 3241, 3243 (2025).

[57] Gal & Lynskey, *supra* note 12 ("[S]ynthetic data can lead to new (or more accurate) learnings…synthetic data can also be used to enable new products or services, such as creating virtual spaces in the metaverse.").

[58] Zero-to-one is a concept popularized by entrepreneur and investor Peter Thiel. The phrase describes the act of creating something entirely new and groundbreaking, as opposed to merely building on or replicating existing ideas. *See generally* PETER THIEL & BLAKE MASTERS, ZERO TO ONE: NOTES ON STARTUPS, OR HOW TO BUILD THE FUTURE (Crown Business 2014).

[59] Bryan McMahon, *AI Is Ushering in a New Scientific Revolution*, GRADIENT (June 4, 2022), https://thegradient.pub/ai-scientific-revolution [https://perma.cc/VBT4-B56W].

development.[60] These virtual patients include completely artificial computer-generated persons that represent the full range of human variables, for whom decades' worth of real data can be created in seconds.[61]

Synthetic data is not static, but rather a dynamic, customizable service. Several large market players in the technology space are currently using synthetic data to train their AI.[62] In an emerging business model, there are companies that specialize in providing synthetic datasets, either by leveraging their proprietary databases or generating custom synthetic data on demand for specific customers.[63] As AI penetrates more market sectors and the tasks demanded of AI become more detailed and specific, developers will demand datasets with particular characteristics. Synthetic data technology allows data scientists to generate the data they need, "on demand, in whatever volume they require, tailored to their precise specifications."[64]

II. THE RELEVANT PRODUCT MARKET

In antitrust, a product market refers to a group of products or services that are considered reasonably interchangeable or substitutable by consumers, meaning that if the price of one product increases significantly, buyers would likely switch to another product within that market to meet their needs.[65] Identifying a product market is a key component in determining whether behavior by a company has the potential to have anticompetitive effects.[66] Defining the boundaries of the product market often can be determinative of whether a company has substantial market power and the potential to harm competition by raising prices, restricting output, or creating moats.[67]

---

[60] Allison Proffitt, *The Role of Virtual Twins in Clinical Trials*, CLINICAL RSCH. NEWS (July 13, 2020), https://www.clinicalresearchnewsonline.com/news/2020/07/13/the-role-of-virtual-twins-in-clinical-trials [https://perma.cc/BQ3U-6DEL].

[61] Gal & Lynskey, *supra* note 12, at 1107.

[62] These include, for example, Microsoft, Google, and OpenAI. *See Why computer-made data is being used to train AI models*, FIN. TIMES (July 18, 2023) https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de [https://perma.cc/L3KX-UM9V] (highlighting synthetic data as a core pillar in training ML and models).

[63] Some examples include Synthesis AI, Scale, Gretel AI, Mostly AI, and YData. Additionally, large incumbents like NVIDIA are building increasingly sophisticated synthetic data engines, like OMNIVERSE. *See NVIDIA Announces Omniverse Replicator Synthetic-Data-Generation Engine for Training AIs*, NVIDIA (Nov. 9, 2021), https://nvidianews.nvidia.com/news/nvidia-announces-omniverse-replicator-synthetic-data-generation-engine-for-training-ais [https://perma.cc/3BQ9-5A7A].

[64] Rob Toews, *Synthetic Data is About to Transform Artificial Intelligence*, FORBES (June 12, 2022), https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=2b8888347523 [https://perma.cc/X4MK-QJ8D].

[65] *See* ELEANOR FOX & DANIEL CRANE, CASES AND MATERIALS ON U.S. ANTITRUST IN GLOBAL CONTEXT 481–88 (4th ed. 2020).

[66] F.T.C. v. Swedish Match, 131 F. Supp. 2d 151, 156 (D.D.C. 2000) (citing Brown Shoe Co. v. U.S., 370 U.S. 294, 324 (1962)) ("Merger analysis begins with defining the relevant product market."); *id.* (quoting F.T.C. v. Cardinal Health, Inc., 12 F. Supp. 2d 34, 45 (D.D.C. 1998)) ("'Defining the relevant market is critical in an antitrust case because the legality of the proposed merger[] in question almost always depends upon the market power of the parties involved . . . .'"); *see also id.* (citing F.T.C. v. Staples, Inc., 970 F. Supp. 1066, 1073 (D.D.C. 1997)) (reasoning that the relevant market definition is often "the key to the ultimate resolution of this type of case because of the relative implications of market power"). "Plaintiffs who sue under federal antitrust laws must define a 'relevant market' for their allegations, a threshold issue that has become a stumbling block for high-stakes litigation in recent years." Katie Arcieri, *Defining a 'Market' Creates Headaches for Antitrust Plaintiffs*, BLOOMBERG LAW NEWS (Feb. 7, 2025), https://news.bloomberglaw.com/antitrust/defining-a-market-creates-headaches-for-antitrust-plaintiffs [https://perma.cc/D7UQ-BRUZ] (discussing the January 31, 2025 ruling which threw out the FTC's bid to block Tempur Sealy International Inc. from closing its acquisition of Mattress Firm after finding that the market it targeted, premium mattresses priced at $2,000 or above, was arbitrary).

[67] *See*, *e.g.*, F.T.C. v. Whole Foods Market, Inc., 548 F.3d 1028 (D.C. Cir. 2008) (considering whether the market should be defined as all supermarkets or just premium organic supermarkets); U.S. v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001) (considering whether the market should be defined as operating systems for all computers or just Intel-compatible PC operating systems).

Defining a relevant product market in antitrust law requires identifying the boundaries within which competition occurs. Courts and regulators focus on two key factors: product substitutability and cross-elasticity of demand.[68] Substitutability examines whether consumers view two products as reasonably interchangeable for the same purpose, while cross-elasticity considers whether price changes in one product cause significant shifts in demand for another. In *Brown Shoe Co. v. United States* (1962), the Supreme Court emphasized that market definition should consider the "reasonable interchangeability of use" and the preferences of consumers.[69] This practical, consumer-centric lens remains a cornerstone of antitrust analysis. The *Brown Shoe* decision identified several practical factors for defining markets, including industry recognition, distinct product characteristics, specialized vendors, price differences, and distinct consumer preferences.[70] These factors ensure market definition reflects real-world commercial practices rather than abstract theory.[71] For example, courts consider whether consumers view products as interchangeable, whether vendors specialize in specific goods, and whether price changes affect demand across products.[72]

Factors influencing market definition have evolved with changing technology and industry dynamics. Early cases like *United States v. E.I. du Pont de Nemours & Co.* (1956) introduced the "cellophane fallacy," cautioning against defining markets too broadly by relying solely on existing price competition.[73] Modern courts also consider supply-side factors, like the ease with which producers can switch to making competing products.[74] In technology markets, courts have grappled with platform ecosystems, as in *Ohio v. American Express Co.* (2018), where the Supreme Court recognized the unique dynamics of two-sided markets.[75] These decisions illustrate the need to adapt market definitions to account for network effects, innovation cycles, and barriers to entry.

Training data is too broad to be considered one product market. This is true today, and it will only become more apparent as AI and AI-adjacent industries develop and demand increases for the quantity and quality of data needed to train AI. AI is a collection of specialized tools and systems, and each requires different types of training data. Below, I propose a taxonomy of different data attributes to understand the nuances of training data and enable antitrust practitioners

---

[68] *Brown Shoe Co.*, 370 U.S. at 325 ("The outer boundaries of a product market are determined by the reasonable interchangeability of use or the cross-elasticity of demand between the product itself and substitutes for it.").
[69] *Id.*
[70] *Id.* at 325
[71] *Id.* at 326 (emphasizing that the Court's analysis was grounded in economic reality, considering how markets function in practice rather than in theoretical abstraction).
[72] *Id.* at 325.
[73] United States v. E.I. du Pont de Nemours & Co., 351 U.S. 377 (1956) (noting that focusing on existing competition between products could misrepresent the market's true scope when a monopolist already controls pricing, which is known as the *Cellophane* fallacy, and highlighted the need to examine competition in terms of price, quality, and substitution under competitive conditions).
[74] While *Brown Shoe* primarily addressed demand-side considerations, later antitrust cases like FTC v. Staples, Inc., 970 F. Supp. 1066, 1074–76 (D.D.C. 1997) integrated supply-side substitution as a relevant factor. Courts recognized that the ability of producers to retool and compete in the market impacts the proper definition of its boundaries.
[75] The Supreme Court in Ohio v. American Express Co., 585 U.S. 529, 534–37 (2018) held that two-sided markets require analyzing competition on both sides (e.g., merchants and cardholders) to define the relevant product market. This case established that traditional antitrust tools must adapt to digital platforms and ecosystems with indirect network effects.

to construct a robust definition of the data product market(s) that may be relevant in their analyses.[76]

III. CLASSIFICATION OF THE TRAINING DATA MARKET

A. Based on Purpose

There are many ways to define the product market boundaries of data. Data can be classified by its intended purpose, answering the question: **how is the data going to be used?** The purpose element is two faceted. On one hand, one can talk about the motivation of the dataset's creators for creating the dataset, and, on the other, one can discuss whether the users of the dataset have a specific task for the data.[77]

1. The Deep Learning Workflow

Purpose is frequently a reflection of where the data will be used in the deep learning workflow. The general-purpose question, how is the data going to be used, can also be asked: **where in the training and deployment workflow is the data going to be used?** To illustrate the process, consider an example, simplified deep learning training workflow broken down into five steps: data collection, pre-processing, pre-training, post-training, and optimization.[78]

FIGURE 1: Simplified Example Deep Learning Workflow for Training a Model [79]

---

[76] This Article rests on the shoulders of a comprehensive data documentation movement. Data documentation refers to the practice of creating clear, comprehensive records about datasets, including their origin, structure, content, usage, and limitations. It involves providing metadata, data dictionaries, and provenance records to ensure that users understand how to use and interpret the data. Proper documentation helps developers, analysts, and regulators assess the quality, reliability, and fairness of datasets. *How to Prevent Discriminatory Outcomes in Machine Learning*, White Paper, World Economic Forum Global Future Council on Human Rights 2016–2018 (2018), https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning [https://perma.cc/4WEM-V966]. Thorough documentation helps developers understand these risks and decide whether the dataset requires augmentation, further validation, or rejection. Sandra Wachter et al., *Bias Preservation in Synthetic Data: Challenges and Solutions*, 123 W. VA. L. REV. 735 (2021). In the U.S., organizations like the National Institute of Standards and Technology (NIST) have championed the importance of data documentation. *See* NAT'L INST. OF STANDARDS AND TECH., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (2023), https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 [https://perma.cc/BYF7-EJAM]. Frequently adopted documentation frameworks used by data scientists and developers to promote transparency in their AI workflows include model cards and data sheets for datasets. *See* Margaret Mitchell et al., *Model Cards for Model Reporting*, in *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220 (2019), https://dl.acm.org/doi/10.1145/3287560.3287596; Timnit Gebru et al., *Datasheets for Datasets* 5, 8 (Dec. 1, 2021) (unpublished manuscript), https://arxiv.org/abs/1803.09010 [https://perma.cc/JEP3-56GF].

[77] *See* Gebru et al., *supra* note 77 at 4.

[78] *See* Sebastian Raschka, *New LLM Pre-Training and Post-Training Paradigms*, AHEAD OF AI (Aug. 17, 2024), https://magazine.sebastianraschka.com/p/new-llm-pre-training-and-post-training [https://perma.cc/YKM3-ABNE].

[79] *See id.*

The pre-processing step involves data cleaning and curation, where data is preprocessed to remove noise, handle missing values, and ensure quality. More often now, at this stage, synthetic data can be used to augment a dataset for better training. The pre-processing step includes data cleaning and formatting.[80] Pre-training and post-training are terms which generally are used only in LLM training. In pre-training, the model learns general knowledge and foundational representations.[81] The model is then refined in post-training.[82] Post-training can leverage techniques such as supervised fine-tuning (SFT) and reinforcement learning (RL).[83] In SFT, the pre-trained model is, for instance, adapted to a specific task by incorporating labeled data examples and supervised learning.[84] In RL, the model is refined through feedback or reward signals.[85] RL can improve task-specific performance, especially for subjective tasks.[86] Post-training may be a long-term, ongoing process that can go beyond deployment.[87] Post-training occurs after the initial training and covers the monitoring, evaluation, fine-tuning, validation, and improvements required to adapt the model and keep it relevant in dynamic real-world scenarios.[88] Models may be periodically fine-tuned or retrained with newly collected data to address data drift or changing user needs.[89]

The workflow will change depending on the learning paradigm utilized to train a model. Supervised learning involves a workflow where labeled input data is used to train a model to predict outcomes for unseen data.[90] The process begins with collecting input data and corresponding labels, extracting relevant features, and feeding this labeled data into the model during training.[91] The model then learns patterns that map inputs to outputs, enabling it to make predictions on new data.[92] In contrast, self-supervised or unsupervised learning focuses on

---

[80] *Data Labeling: The Authoritative Guide*, SCALE AI GUIDES (visited Jan. 23, 2025), [https://perma.cc/F558-95B9].
[81] *Pre-Training,* NAT'L INST. OF STANDARDS & TECH., https://csrc.nist.gov/glossary/term/pre_training [https://perma.cc/3Q5R-RNC5].
[82] David Testuggine, *A Primer on LLM Post-Training,* PYTORCH (Aug. 26, 2025), https://pytorch.org/blog/a-primer-on-llm-post-training/ [https://perma.cc/29W4-9TZK].
[83] *Id.*
[84] *Id.*
[85] *Id.*
[86] AlphaGo, developed by DeepMind, is a landmark AI system that used RL to master the game of Go, a board game with near-infinite complexity. As a model for RL, AlphaGo combined supervised learning on expert games with self-play, where it iteratively improved by playing against versions of itself, optimizing strategies through trial and reward mechanisms. Notably, in its famous 2016 match against Lee Sedol, AlphaGo's move 37 in game two was an unconventional and seemingly counterintuitive move that surprised experts but ultimately demonstrated the model's ability to uncover strategies beyond human intuition. *See* Bill Parker, *Move 37*, MEDIUM (June 27, 2024), https://neuralocity.medium.com/move-37-ffff2969b7ed [https://perma.cc/JYE2-37QS] ("AlphaGo came up with this move not because it was a strategy directly programmed by the Google team, but because of reinforcement learning and self-play, it was able to see this move as the best move to meet [its] objective of winning."); *see also* Cade Metz, *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*, WIRED (Mar. 16, 2016), https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ [https://perma.cc/22TK-37HW].
[87] Xingkin Yin et al., *Experience Scaling: Post Deployment Evolution for Large Language Models* (Sept. 23, 2025), https://arxiv.org/abs/2509.18771 [https://perma.cc/7JBG-EPX8].
[88] Testuggine, *supra* note 83.
[89] Baumann et al., *supra* note 40.
[90] YOSHUA BENGIO, IAN GOODFELLOW & AARON COURVILLE, DEEP LEARNING, 134–39 (MIT Press 2016).
[91] *Id.*
[92] *Id.*

analyzing and finding hidden patterns in unlabeled data.[93] Here, the workflow involves inputting raw data, extracting features, and grouping or clustering data points based on similarities or patterns without pre-existing labels.[94] RL is another paradigm where agents learn by interacting with an environment and receiving rewards or penalties for actions.[95] Finally, transfer learning is another powerful technique that has become more prominent with the rise of large pre-trained models. In transfer learning, also referred to as knowledge transfer, knowledge (in the manner of features, weights, or representations) learned from a large general-purpose dataset is applied to more specific problems, often with limited data.[96]

2. Attributes of Data Intended for Different Purposes

A dataset may be aggregated to be used as pre-training data for a foundational model. Foundational models are AI models trained on vast, broad datasets that can be adapted to a wide range of downstream tasks and application.[97] Pre-training a foundational model will leverage unsupervised learning and require large-scale, typically unlabeled, diverse datasets. OpenAI's foundational model GPT-3 was trained on essentially the whole internet.[98] In computer vision, the foundational architecture ResNet[99] was pre-trained on ImageNet, a massive and diverse repository of images that is manually labeled and organized using the WordNet hierarchy.[100] The purpose of pre-training data is to provide the initial knowledge base for ML models, enabling them to learn general patterns and relationships.[101] Importantly, the key feature of a robust pre-training dataset seems to be size: almost always, the larger, more diverse dataset, will equate to better downstream performance.[102]

As above, post-training involves refining a pre-trained machine learning model to improve its performance on specific tasks, align it with desired behaviors, or adapt it to particular domains.[103] This step includes fine-tuning, reinforcement learning with human feedback (RLHF),

---

[93] *Id.* at 139–42.
[94] *Id.*
[95] *See* Xu et al., *supra* note 36.
[96] Sinno Jialin Pan & Qiang Yang, *A Survey on Transfer Learning*, 22(10) IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENG'G 1345–59 (2010).
[97] The key elements of a foundation model are scale, breadth, and adaptability. Through transfer learning, foundational models can apply knowledge from general training to specific applications. See Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ARXIV (Aug. 16, 2021), https://arxiv.org/abs/2108.07258 [https://perma.cc/C766-NBFR] (introducing the term "foundation models").
[98] More properly, GPT-3 was trained on a combination of Common Crawl, WebText2, and even Wikipedia. Tom B. Brown et al., *Language Models are Few-Shot Learners,* ARXIV (July 22, 2020), at 9 (tbl. 2.2), https://arxiv.org/pdf/2005.14165 [https://perma.cc/H8ME-62FU].
[99] Kaiming He et al., *Deep Residual Learning for Image Recognition,* ARXIV (Dec. 10, 2015), https://arxiv.org/abs/1512.03385 [https://perma.cc/XSU2-NL8X] (the extensive size and diversity of ImageNet allowed ResNet to learn robust features for image recognition tasks).
[100] *About*, IMAGENET https://www.image-net.org/about.php [https://perma.cc/WQ3D-6THP].
[101] *See generally* Rahim Entezari et al., *The Role of Pre-Training Data in Transfer Learning*, ARXIV (Mar. 1, 2023), https://arxiv.org/abs/2302.13602 [https://perma.cc/HG2W-FKNA] (analyzing the impact of pre-training data distribution, curation, and scale on transfer learning performance across various downstream tasks).
[102] *Id.* at 2, 7 (finding that curated data outperforms similarly sized noisy data, though scaling noisy data 15x-2000x matches curated performance).
[103] *See* BENGIO, GOODFELLOW & COURVILLE, *supra* note 91.

or model evaluation and calibration.[104] Developers must ensure the data used in post-training aligns with the goal of the refinement process.[105] To organize this process, developers can break post-training data needs into three categories: task adaptation data for fine-tuning, ensuring relevance to the target domain or problem;[106] behavioral alignment data, including human feedback for shaping outputs;[107] and evaluation and test data, designed to rigorously validate the model's performance against desired metrics and failure modes. Each dataset must be curated to ensure the model learns effectively while avoiding overfitting or unintended biases.[108]

Fine-tuning data, which may also be called task adaptation data, is a task-specific dataset used to adapt pre-trained models for particular applications.[109] The purpose of fine-tuning is further training to adjust the model weights to perform better in specific scenarios or on tasks that were not well covered during pre-training.[110] The new knowledge added during fine-tuning is more about enhancing the model's performance in specific contexts rather than broadly expanding its general knowledge.[111] This dataset thus enhances model performance on domain-specific tasks by providing high-quality, targeted examples.[112] The characteristics of fine-tuning data differ from those of pre-training data: fine-tuning data is directly aligned with the final task the model is intended to perform, is often meticulously labeled to ensure the model learns the desired outputs, and must be high-quality, since errors or noise in a fine-tuning dataset can negatively affect model performance.[113] Although the datasets must be higher in quality, validation typically requires less data compared to pre-training, since the model already has a foundational understanding, so these datasets generally can be smaller in volume.[114] For instance, ResNet's pre-trained models are widely used as backbones in transfer learning for various downstream tasks. To adapt ResNet to a task like medical image analysis, a ResNet pre-trained on ImageNet may be fine-tuned on a small,

---

[104] *See, e.g.*, Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, 35 ADVANCES IN NEURAL INFO. PROCESSING SYS. 27730, 27731–32 (2022) (detailing a post-training methodology that applies supervised fine-tuning followed by reinforcement learning from human feedback to align model behavior).

[105] Andrej Karpathy, *A Recipe for Training Neural Networks* (Apr. 25, 2019), https://karpathy.github.io/2019/04/25/recipe/ [https://perma.cc/6F6U-H9DH] (emphasizing the alignment of data and task).

[106] *See* Aditya Jain, Amir Maleki, & Nathalie Saade, *How to Fine-Tune: Focus on Effective Datasets*, META (Aug. 7, 2024), [https://perma.cc/QDQ7-CMEA].

[107] *See* Paul Christiano et al., *Deep Reinforcement Learning From Human Preferences* (Jun. 12, 2017) (unpublished manuscript), https://arxiv.org/abs/1706.03741 [https://perma.cc/6KAU-79WA] (last updated Feb. 17, 2023).

[108] Karpathy, *supra* note 72.

[109] *See, e.g.*, Huanxuan Liao et al., *DATA: Decomposed Attention-Based Task Adaptation for Rehearsal-Free Continual Learning,* ARXIV (Feb. 17, 2025), https://arxiv.org/pdf/2502.11482 [https://perma.cc/XVZ2-8ZR9] (defining task adaptation as the phase where pre-trained models are updated using task-specific datasets to address domain-specific applications).

[110] Rashmiranjan Pradhan, *RAG vs. Fine-Tuning vs. Prompt Engineering: A Comparative Analysis for Optimizing AI Models*, INT'L J. OF COMPUT. TECH. AND ELEC. COMMC'N 11326, 11328 (2025).

[111] *Id.*

[112] *Id.*

[113] *Fine-tuning LLMs and AI models*, GOOGLE CLOUD, https://cloud.google.com/use-cases/fine-tuning-ai-models [https://perma.cc/EGP5-84DZ] (last visited Feb. 19, 2026).

[114] *See* Jain, Maleki & Saade, *supra* note 107 (in the context of fine-tuning an LLM, "quality is paramount," highlighting that "a few thousand curated examples of the LIMA dataset [1,000 carefully curated prompts and responses] had better performance than the 50k machine-generated Alpaca dataset when fine-tuning.").

targeted, labeled and annotated dataset of lung tumor MRI pictures.[115] A customer who requires fine-tuning data could not interchange that dataset for a dataset intended for pre-training.

RL datasets enable agents to learn through sequential decision-making, where each action influences future states and rewards.[116] RL fits into the ML pipeline as a dynamic feedback loop, where agents explore, exploit learned behaviors, and continuously update their policies based on episodic or lifelong learning.[117] Unlike datasets in both supervised and unsupervised learning, which researchers typically compile as static collections of labeled or unlabeled examples, RL datasets form dynamically as agents interact with their environments.[118] These datasets include state representations, action choices, reward signals, and environment transitions, all of which shape the agent's policy. Most RL systems generate data within simulated environments, such as game engines in AlphaGo's self-play matches[119] or physics-based simulations in OpenAI's hide-and-seek experiments.[120] Since agents produce this data rather than collecting it from real-world observations, it is primarily synthetic. However, some RL applications rely on real-world data, such as robot control logs or historical trading records, though this approach, termed offline RL,[121] demands careful handling due to safety and sample efficiency concerns.[122] Effective RL datasets must capture long-term dependencies, delayed rewards, exploration-exploitation trade-offs, and high-dimensional state spaces.[123] For RLHF, developers need feedback data, often generated from comparisons or rankings provided by human annotators, to guide the model's behavior.[124] Researchers create RL datasets through self-play (e.g., AlphaGo refining its strategies against itself), curriculum learning (e.g., agents progressing through increasingly complex tasks), or human demonstrations (e.g., imitation learning in robotics).

Evaluation and test data validate a model's performance by rigorously assessing its

---

[115] *See generally* Wanni Xu, You-Lei Fu & Dongmei Zhu, *ResNet and Its Application to Medical Image Processing: Research Progress and Challenges*, 240 COMPUT. METHODS & PROGRAMS IN BIOMEDICINE 107660 (2023) (research study examining ResNet's use in modern medical image processing).

[116] Volodymyr Mnih et al., *Human-level control through deep reinforcement learning*, NATURE 518, 529–533 (2015).

[117] *Id.*

[118] Bowen Baker et al., *Emergent Tool Use From Multi-Agent Autocurricula*, (Feb. 11, 2020), https://arxiv.org/pdf/1909.07528 [https://perma.cc/8LNZ-TVQS].

[119] David Silver, Thomas Hubert, & Julian Schrittwieser, *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm,* ARXIV (Dec. 5, 2017), 1–3, https://arxiv.org/pdf/1712.01815 [https://perma.cc/Y3J2-XMQZ].

[120] Baker et al., *supra* note 117 (where through training in a simulated hide-and-seek environment, agents discover progressively more complex tool use while playing a simple game of hide-and-seek, formulating a series of six distinct strategies and counterstrategies, some of which the researchers did not know the environment supported).

[121] Justin Fu et al., *D4RL: Datasets for Deep Data-Driven Reinforcement Learning* 2 (Feb. 6, 2021) (unpublished manuscript), https://arxiv.org/pdf/2004.07219 [https://perma.cc/2E8R-E9Z2[("The offline [RL] setting (also known as full batch RL), where a policy is learned from a static dataset, is compelling as progress enables RL methods to take advantage of large, previously-collected datasets, much like how the rise of large datasets has fueled results in supervised learning").

[122] *Id.* (introducing Datasets for Deep Data-Driven Reinforcement Learning (D4RL), "a suite of tasks and datasets for benchmarking progress in offline RL" to deal with the current efficiency gaps in offline RL.).

[123] *Id.*

[124] For using reinforcement learning from human feedback (RLHF), *see* Christiano et al., *supra* note 108; Nisan Stiennon et al., *Learning to summarize from human feedback* (Sep. 2, 2020) (unpublished manuscript), https://arxiv.org/pdf/2009.01325 [https://perma.cc/272W-J2WJ].

accuracy, robustness, and alignment with desired metrics and behaviors.[125] Developers design evaluation datasets to include diverse examples that reflect real-world scenarios, edge cases, and potential failure modes. For example, testing a language model might involve clean task-specific benchmarks like GLUE,[126] adversarial examples to measure robustness,[127] and fairness subsets to identify biases.[128] In RL, evaluation data might simulate diverse episodes or real-world conditions to assess generalization and long-term decision-making.[129] Developers also test for robustness to distribution shifts, such as domain-specific variations or temporal changes, using datasets like WILDS.[130] Effective evaluation data must align with the application's goals, whether testing classification accuracy, fairness in predictions, or the robustness of generated outputs.[131] By using targeted evaluation datasets, developers ensure models meet deployment requirements and adapt effectively to real-world demands.[132]

B. Based on Source

Datasets are often classified based on the provenance, or source, of the data.[133] However, source requires more specificity. Datasets can be classified by source in several categories: collection method, collection location, and collecting entity. Thus, bounding data by source requires answering the following three questions: **where was the data gathered? How was the data collected?** And **who collected the data?** Datasets also can combine data gathered utilizing multiple collection methods at multiple collection locations. There is a significant overlap between these three subsets of source, and all three are important categories informing the composition of

---

[125] Mahmoud Mohammadi et al., *Evaluation and Benchmarking of LLM Agents: A Survey* (July 29, 2025) (unpublished manuscript), https://arxiv.org/html/2507.21504v1 [https://perma.cc/U3AK-4K23]; Hoda Khoshvaght et al., *A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction* 13 J. ENV'T CHEM. ENG'G 6 (2025).

[126] Alex Wang et al., *GLUE: A multi-task benchmark and analysis platform for natural language understanding*, Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 353–355 (2018), https://doi.org/10.18653/v1/W18-5446 [https://perma.cc/9SQM-7J78].

[127] Robin Jia & Percy Liang, *Adversarial Examples for Evaluating Reading Comprehension Systems* (July 23, 2017) (unpublished manuscript), https://arxiv.org/abs/1707.07328 [https://perma.cc/77F7-E2HB] (proposing adversarial evaluation for NLP where the test dataset, a series of reading comprehension questions and answers, includes a distractor sentence in the prompt that does not answer the question but looks related).

[128] Vinodkumar Prabhakaran, Ben Hutchinson & Margaret Mitchell, *Perturbation Sensitivity Analysis to Detect Unintended Model Biases*, ASS'N FOR COMPUTATIONAL LINGUISTICS 5740, 5740–5745 (2019), https://aclanthology.org/D19-1578/ [https://perma.cc/GB9C-BWVV] (proposing an evaluation approach for fairness, since "NLP models are susceptible to learning incidental associations around named referents at a particular point in time, in addition to general linguistic meaning.").

[129] Dianzhao Li & Ostap Okhrin, *A platform-agnostic deep reinforcement learning framework for effective Sim2Real transfer towards autonomous driving*, NAT'L LIBRARY OF MEDICINE (2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC11487131/ [https://perma.cc/AR5E-NSWL].

[130] Pang Wei Koh et al., *WILDS: A Benchmark of in-the-Wild Distribution Shifts* (Dec. 14, 2020) (unpublished manuscript), https://arxiv.org/abs/2012.07421 [https://perma.cc/J9V4-R2KQ] (focusing on testing robustness to distribution shifts, that is, where training distribution differs from the test distribution in such a way that substantially degrade the accuracy of ML systems deployed across domains, by presenting WILDS, a benchmark of 10 datasets reflecting a diverse range of distribution shifts).

[131] Mitchell, *supra* note 77, at 224 (highlighting the need to use targeted evaluation datasets to measure fairness, bias, and application-specific metrics).

[132] *See* Kamya Shah, *A Step-by-Step Guide to Building Robust Evaluation Datasets for AI Agents*, MAXIM (Nov. 22, 2025), https://www.getmaxim.ai/articles/a-step-by-step-guide-to-building-robust-evaluation-datasets-for-ai-agents/ [https://perma.cc/2QFG-ZFRS] ("Quality evaluation datasets enable teams to run meaningful evaluations, catch regressions before deployment, and align agent behavior with user expectations.").

[133] *See generally* Tim Mucci, *What is data provenance?,* INT'L BUS. MACHINES (July 23, 2024), https://www.ibm.com/think/topics/data-provenance [https://perma.cc/T6GY-X5Y7] (highlighting the importance of data provenance and general purpose).

a dataset.

1. Collection Location

Where data is gathered provides information not only as to the geographic origin of the data, but also regarding the device which collected it.[134] Geographically, the location where data is collected can introduce biases and limitations in models.[135] Datasets predominantly sourced from certain regions may lack diversity, causing models to underperform when applied to data from areas underrepresented in the training dataset.[136] For instance, one study found that models trained mainly on Western data struggled to accurately geolocate images from African regions.[137] Cultural and socioeconomic contexts existing where the data is collected can similarly influence data characteristics.[138] In another study, researchers analyzed two popular public datasets, ImageNet and Open Images, datasets which are generally considered academic benchmarks, and found that the datasets exhibited an observable Americentric and Eurocentric representation bias.[139] In a third study, researchers found that the activity "cooking" was over 33% more likely to involve females than males in a labeled training dataset, and that a model trained on that dataset further amplified the disparity to 68%.[140] Thus, for example, a developer looking to train a facial recognition model for deployment in Africa may not obtain good results using a dataset composed of portraits taken in Norway.[141] These datasets likely would not be equivalent or substitutable.

The device used to collect the data may also influence the composition of a dataset. Data may be collected directly from an individual or a group of people through internal systems such as customer or patient databases, user surveys, field studies, or proprietary software logs. Data can also be collected by extracting information from online platforms or websites. It follows that access to collection platforms can be determinative of what the dataset represents; the data will be skewed to the userbase's attributes.[142] Data can also be gathered from sensors embedded in physical devices. Models built for applications involving physical environments or real-world interactions will necessitate data that covers the task or capability the AI endeavors to achieve. Data collected on the edge, by IoT devices, wearable technology, industrial machines, or autonomous vehicles is specific and not substitutable. This type of data may be challenging to collect and require

---

[134] *See What is Location Data? The Complete Guide for Data and Product Teams*, UNACAST (2025), https://www.unacast.com/what-is-location-data, [https://perma.cc/H9AF-B3P7].

[135] Ximena Salgado Uribe, Martí Bosch, & Jérôme Chenal, *Regional biases in image geolocation estimation: a case study with the SenseCity Africa dataset* (Apr. 3, 2024) (unpublished manuscript), https://arxiv.org/pdf/2404.02558 [https://perma.cc/Y3PR-BXLG].

[136] *Id.*

[137] *Id.* (highlighting the importance of including geographically diverse data to enhance model generalizability).

[138] *See* Shreya Shankar et al., *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World* (Nov. 22, 2017) (unpublished manuscript), https://arxiv.org/pdf/1711.08536 [https://perma.cc/2LJK-AJC7].

[139] *Id.*

[140] Jieyu Zhao et al., *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints* (July 29, 2017) (unpublished manuscript), https://arxiv.org/abs/1707.09457 [https://perma.cc/4ACA-V3R6].

[141] *See* Buolamwini & Gebru, *supra* note 143, at 77–91 (showing that because of underrepresentation in the composition of the training dataset, facial recognition models had a higher error rate for black women than for white men).

[142] Occupations, personality traits, and everyday situations, can be very gender, age, race, and geographical location dependent. *See* Ranjita Naik & Besmira Nushi, *Social Biases through the Text-to-Image Generation Lens* (Mar. 30, 2023) (unpublished manuscript), https://arxiv.org/pdf/2304.06034 [https://perma.cc/4R5P-Y5VW].

specialized hardware and infrastructure, making it more expensive.

2. Collection Method

Methods of obtaining data vary widely. As a subset of source, collection method looks to procedure (that is, **what procedure was used to collect the data?** And if the dataset is a sample from a larger population, **what was the sampling strategy?**) and temporal scope (that is, **over what timeframe was the data collected?**).[143] Collection procedures and temporal scope are critical to determine the reliability and applicability of the data for AI training and deployment. The sampling strategy employed, whether random, stratified, or convenience-based, directly influences whether the resulting dataset accurately represents the target population or introduces systematic biases that compromise model fairness and performance.

The methods used to select subsets of data from an entire population determine how representative the data is and directly affects the model's bias, diversity, and capabilities.[144] For instance, random sampling (where every data point in the population of that particular data has an equal chance of being selected) can produce datasets that capture population-level characteristics, making them ideal for generalizable AI models, while stratified sampling (where the population is divided into distinct subgroups and samples are taken proportionally from each) may ensure balanced representation of specific subgroups, which is essential for avoiding bias against underrepresented segments of the population.[145] In contrast, convenience sampling can in some instances introduce significant biases, limiting the dataset's utility for fair or accurate AI predictions depending on the application.[146] Understanding the collection method helps identify potential gaps, biases, or strengths in the data, enabling developers to align the dataset with the specific objectives and ethical requirements of the AI system being trained.[147] Without this understanding, developers risk using data that could compromise the model's performance, fairness, or generalizability.

Temporal scope refers to the timeframe over which data is collected.[148] This affects the relevance and adaptability of data to dynamic systems and evolving phenomena.[149] Generally, data may be static data (data collected at a single point in time or over a short period), time-series data (data collected continuously or at regular intervals over time), longitudinal data (data collected

---

[143] *See* Gebru et al., *supra* note 77, at 7.

[144] Line H. Clemmensen & Rune D. Kjærsgaard, *Data Representativity for Machine Learning and AI Systems* (Feb. 3, 2023) (unpublished manuscript), https://arxiv.org/pdf/2203.04706 [https://perma.cc/EG39-KFNJ] ("Biased results can arise if the sample does not accurately represent the target population, or if there is a lack of sufficient representation for subgroups within the data.").

[145] *Id.* (showing the results of stratified random sampling can closely mimic a population instead of generalizing to the population).

[146] *See* Aspen Hopkins et al., *Designing Data: Proactive Data Collection and Iteration for Machine Learning* 2–4 (unpublished manuscript) (Jul. 29, 2023), https://arxiv.org/pdf/2301.10319 [https://perma.cc/Y52S-GM85] (noting how difficulty of collection leads to compromises, which has the potential to inhibit robust ML).

[147] Margaret Mitchell et al., *Diversity and Inclusion Metrics in Subset Selection* 1 (unpublished manuscript) (Feb. 9, 2020), https://arxiv.org/pdf/2002.03256 [https://perma.cc/2Q2N-WHMU] (highlighting that failure to effectively recognize data quality and coverage needs can lead to biased ML models, which can, in turn, perpetuate systemic power and access differentials and lead to the deployment of inaccessible or defective product experiences).

[148] *See* Gebru et al., *supra* note 77, at 7.

[149] *Id.* at 11.

from the same subjects or entities over an extended period), real-time data (data collected and processed as events occur), or retrospective data (data collected from past records or historical datasets).[150] While static data, also referred to as cross-sectional data, may be suitable for baseline model training and applications with predictable environments, longitudinal data will be more useful to train models that try to predict user behavior or determine appropriate health treatment.[151] Economically, retrospective data is typically more cost-effective and readily available, while gathering and assembling real-time datasets entails expensive infrastructure and computational demands.[152]

The choice of temporal scope, too, is important in avoiding biases. For example, in the mid-2010s, Amazon developed an AI recruitment tool to automate the process of reviewing resumes.[153] The tool was trained on ten years of internal hiring data, which predominantly reflected the company's male-dominated workforce in the technology sector.[154] This historical bias in the original dataset became embedded in the AI's decision-making, resulting in the algorithm systematically penalizing resumes that included indicators of female candidates, such as mentions of women's colleges or activities.[155]

3. Collecting Entity

The individuals or organizations responsible for collecting data can significantly influence the dataset's attributes. Their choices regarding where the data is collected and how it is collected can introduce biases and affect data quality. Certain data documentation approaches try to capture the importance of the collecting entity by asking dataset creators to clearly identify **who created the dataset** (e.g., which team or research group), on behalf of which entity (e.g., company, institution, organization), and **who funded the creation of the dataset**.[156]

Datasets are frequently categorized as first-, second-, or third-party data.[157] First-party data is data collected directly by the entity developing and deploying the model, often from its own

---

[150] *See* THE CAMBRIDGE HANDBOOK OF RESEARCH METHODS AND STATISTICS FOR THE SOCIAL AND BEHAVIORAL SCIENCES: VOLUME 1: BUILDING A PROGRAM OF RESEARCH, 267–440 (John Edlund & Austin Lee Nichols eds., 2023).

[151] *See, e.g.,* Juan Zhao et al., *Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction*, SCI. REP. 9, 717 at 1 (Jan. 24, 2019), https://doi.org/10.1038/s41598-018-36745-x [https://perma.cc/NY6E-GAK2] (highlighting how longitudinal data was more effective to improve health treatment).

[152] Hans-Christian Thorsen-Meyer et al., *Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records*, 2 LANCET DIGIT. HEALTH 179, 180 (Mar. 12, 2020), https://doi.org/10.1016/S2589-7500(20)30018-2 [https://perma.cc/VAC4-ED8L] (noting the difficulties of gathering real-time data in the healthcare context).

[153] Robert Iriondo, *Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women*, TOWARDS AI NEWS (Sept. 17, 2024), https://towardsai.net/p/machine-learning/amazon-scraps-secret-ai-recruiting-engine-that-showed-biases-against-women [https://perma.cc/MEU9-SYCH].

[154] *Id.*

[155] *Id*.

[156] Gebru et al., *supra* note 77, at 8.

[157] *See* OECD DIRECTORATE FOR FIN. AND ENTER. AFFAIRS COMPETITION COMM., CONSUMER DATA RIGHTS AND COMPETITION – BACKGROUND NOTE 16–18 (2020) (comparing first-party versus third-party consumer data collection); *see also* Jackie Yeaney, *What is First Party, Second Party and Third Party Data*, SALESFORCE ASIA BLOG (Sep. 26, 2023), https://www.salesforce.com/ap/blog/first-party-customer-data/ [https://perma.cc/79SH-WHEB].

users, operations, or products.[158] This sort of data can provide proprietary insights that competitors may not have, and the compiled dataset can give a developer's model a unique edge.[159]

Second- and third-party data categorize the proximity of the data sharer to the entity deploying the model.[160] Second-party data is data shared by a partner organization, typically under a mutual agreement.[161] The last few years have seen the creation or expansion of partnerships between the largest cloud service providers (CSPs) and prominent AI developers, for example, Microsoft and OpenAI, and Google and Anthropic.[162] In January 2024, the Federal Trade Commission launched a market study into these partnerships using its 6(b) authority[163], finding in its final report that the partnerships could permit CSPs and AI developer partners to share or otherwise obtain training data.[164]

Third-party data is data purchased from external providers or aggregators[165], for example a market research dataset purchased from an analytics firm. When purchasing data, developers will consider a data collector's conscious or unconscious preferences and biases, as well as data quality, including any inconsistent labeling or errors during data collection, all of which could degrade model performance.[166] Additionally, whether the data collector is diligent about meeting ethical standards, legal obligations (including obtaining necessary licenses),[167] and privacy obligations (such as ensuring consent) will affect the model's usability.[168]

---

[158] First-party data can include user behavior data—such as page visits, time spent, and click-through rates), transactional data (purchase history, order frequency, product preferences), and also data gathered by employees through connected devices. *See* Yeaney, *supra* note 159.

[159] For example, Tesla uses trade secrets, which require the protected secret to be one that grants the holder a competitive edge, to protect its training data for autonomous vehicle systems. In 2019, Tesla sued several former employees who joined a competing startup, Zoox, alleging they took confidential documents related to Tesla's automation systems, which included its training processes and datasets. *See* Tesla, Inc. v. Zoox, Inc., No. 3:19-cv-01463 (N.D. Cal. 2019); and Sean O'Kane, *Former Tesla Employee Admits Uploading Autopilot Source Code to His iCloud*, THE VERGE (July 10, 2019, at 18:27 ET), https://www.theverge.com/2019/7/10/20689468/tesla-autopilot-trade-secret-theft-guangzhi-cao-xpeng-xiaopeng-motors-lawsuit-filing [https://perma.cc/526Y-DYT2].

[160] Yeaney, *supra* note 159.

[161] *Id.*

[162] *Behind the FTC's 6(b) Report on Large AI Partnerships & Investments*, FTC TECH. BLOG (Jan. 17, 2025), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2025/01/behind-ftcs-6b-report-large-ai-partnerships-investments [https://perma.cc/N569-MRRM].

[163] Section 6(b) of the Federal Trade Commission Act authorizes the FTC to conduct economic and market studies by gathering information about business practices and competitive conditions. 15 U.S.C. § 46(b). Unlike enforcement actions, § 6(b) investigations are non-adversarial inquiries designed to inform the Commission, Congress, and the public about market conditions and competitive dynamics.

[164] *Partnerships Between Cloud Service Provides and AI Developers FTC Staff Report on AI Partnerships & Investments 6(b) Study*, FTC, (Jan. 2025), https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf [https://perma.cc/N5H8-CM8P] ("AI developer partners may be able to access additional training data with the help of their CSP partner…[i]n at least one case, an AI model developer partner sought or planned to seek access to training data from the CSP partner's own products and services. One CSP partner noted that its first-party data would be used by its AI developer partner…").

[165] Yeaney, *supra* note 159.

[166] *Id.*

[167] The risk of copyright infringement may turn developers away from certain datasets. At the time of writing, the first wave of cases that will become landmark for analyzing the fair use doctrine in the AI context are still in the process of being decided by the federal courts. If the courts rule in favor of a narrow interpretation of fair use in the copyright infringement suits filed against companies like Stability AI, some argue that it could make web-scrapped datasets unusable such that data collectors may need to turn to obtaining licenses from copyright holders to use their data for training. *See, e.g.*, Andersen v. Stability AI, No. 3:23-cv-00201 (N. D. Cal. Jan. 13, 2023). *See also* Mark A. Lemley & Brian Casey, *Fair Learning*, 99 TEX. L. REV. 743, 776–77 (2021).

[168] *See* Thomas Dethmann & Jannis Spiekermann, *Ethical Use of Training Data: Ensuring Fairness and Data Protection in AI*, LAMARR INST. BLOG (Jul. 3, 2024), https://lamarr-institute.org/blog/ai-training-data-bias https://perma.cc/6KQS-3S26 (noting a lack of careful

C. Based on Structure

Training datasets can also be classified based on whether they are structured or unstructured[169]. This classification may overlap with the composition of a dataset, that is, what type of data is included in a dataset[170]. The structure of data impacts the usage and processing of the data.[171] Data is classified by structure by answering the question: **how is the data organized?**[172]

Structured data refers to highly organized data that is formatted into predefined, fixed schema, such as tables with rows and columns, that provide defined relationships between data points.[173] Financial data, medical records, and customer databases are examples of structured data[174]. Structured datasets are often labeled and quantitative, making it easier for supervised learning, as they can be directly fed into ML algorithms with minimal preprocessing.[175] However, structured datasets often fail to capture qualitative or complex relationships. Structured data also may be more sensitive, liable to privacy or security concerns, and more capable of passing on inherent biases.[176]

Unstructured data refers to data that lacks predefined organization.[177] Images, audio, videos, transcripts, and emails are all examples of unstructured data.[178] Unstructured data contains contextual and qualitative information, making it ideal for tasks like NLP and computer vision, as it often reflects real-world complexities, enhancing model generalization.[179] However, unstructured data requires extensive preprocessing, it is often unlabeled and requires annotations for supervised learning. Additionally, these datasets generally require significant storage infrastructure due to their size.[180]

There is also semi-structured data, data that does not conform to a strict schema but has some organizational properties. Examples include JSON (a lightweight data-interchange format

---

dataset curation can perpetuate societal inequalities and discrimination against minorities and can lead to malicious outputs and the disclosure of private information).

[169] Bernard Marr, *What's the Difference Between Structured, Semi-Structured, and Unstructured Data?*, FORBES (Dec. 10, 2021), https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/ [https://perma.cc/463C-W4KS].

[170] *Id*.

[171] *Id*.

[172] *Id.*

[173] *Data Labeling: The Authoritative Guide*, SCALE AI GUIDES, https://scale.com/guides/data-labeling-annotation-guide [https://perma.cc/L2UX-BXSA] (last visited Jan. 23, 2025).

[174] *Id.*

[175] Marr, *supra* note 171.

[176] *See* Andrea Gadotti et al., *Anonymization: The Imperfect Science of Using Data While Preserving Privacy*, 10 SCI. ADVANCES, no. 29 (2024) (noting structured data enables adversaries to define attributes for re-identification attacks). Structured data also encodes biases through categorical classifications that machine learning systems can amplify. *See* Tiago Palma Pagano et al., *Bias and Unfairness in Machine Learning Models* (Nov. 3, 2022) (unpublished manuscript), https://arxiv.org/abs/2202.08176 [perma.cc/3KV9-NFZ2].

[177] *Data Labeling: The Authoritative Guide*, SCALE AI GUIDES https://scale.com/guides/data-labeling-annotation-guide [perma.cc/6XS3-PDCX].

[178] *Id*.

[179] *Structured vs Unstructured Data: What's the Difference?*, IBM CLOUD EDUC. TEAM (June 29, 2021), https://www.ibm.com/think/topics/structured-vs-unstructured-data [perma.cc/27DW-4NM9].

[180] *Id.*

used in APIs and web applications), XML (used for data storage and communication in applications like web services), or CSV files (comma-separated values, that while generally considered structured, variations in column organization or missing fields can make it semi-structured). In each format, tags or markers define the data elements.[181] Email messages are also another good example: while the content is unstructured, an email contains structured data such as the email address of the sender and recipient, and the date and time sent.[182] Generally, semi-structured data is readable and interpretable without requiring extensive preprocessing but requires more computational resources than structured data to parse and analyze.[183]

D. Based on Cleaning and Processing Applied

It may be necessary to treat data before it can be used to train a model. In its original form, there is raw data: unprocessed, uncleaned data directly from the source.[184] This data generally comes in an unstructured or semi-structured form.[185] When approaching a dataset, a developer will ask the following questions to determine its suitability: *has the data been cleaned? Was any processing/transformation of the data done (e.g., tokenization, tagging, feature extraction, or removal of instances)? Is there a label or annotation associated with each instance that comprises the dataset?*[186]

Often, raw data will be processed in some way.[187] Data can be cleaned to handle missing values, detect and remove outliers, remove duplicates, or correct inconsistencies like typos. [188] Cleaning data helps prevent bias or redundancy in a model.[189]

Data is also often transformed. [190] Transforming data refers to the process of modifying or converting data into a format that is more suitable for analysis or input into a ML model. [191] Collected raw data may be encoded with categorical values (converting categories into numerical values using one-hot encoding or embeddings, for example),[192] or text data may be tokenized or

---

[181] Chad Sanderson, *Semi-Structured Data: Example and How to Manage*, GABLE (October 23, 2024), https://www.gable.ai/blog/semi-structured-data [https://perma.cc/U64G-QZ4M].
[182] *See*, Sanderson *supra* note 181; Marr, *supra* note 171.
[183] *See*, Sanderson *supra* note 181.
[184] Robb Gibson, *Data Basics: Structured, Unstructured, and Semi-Structured Data*, ALATION (Oct. 3, 2025), https://www.alation.com/blog/structured-unstructured-semi-structured-data/ [https://perma.cc/Q765-NZXQ].
[185] *See* Sanderson, *supra* note 181.
[186] *See* Gebru et al., *supra* note 77, at 8.
[187] *See* Chad Orzel, *What Does it Mean to Share 'Raw Data'?,* FORBES, https://www.forbes.com/sites/chadorzel/2017/07/20/what-does-it-mean-to-share-raw-data/ [https://perma.cc/BZ8Q-XMGU].
[188] *See* Ayushman Durgapal, *Data PreProcessing – Handling Duplicate Values and Outliers in a dataset,* MEDIUM, https://medium.com/@ayushmandurgapal/handling-duplicate-values-and-outliers-in-a-dataset-b00ce130818e [https://perma.cc/UY9H-XQ73].
[189] *See* Manping Guo et al. *Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data: Viewpoint,* 12 INTERACTIVE J. MED. RSCH. 1, 3, 8 (2023).
[190] *See What is Data Transformation? Techniques and Best Practices,* SNOWFLAKE, https://www.snowflake.com/en/fundamentals/unlocking-value-through-data-transformation-in-modern-pipelines/ [https://perma.cc/EF5E-36AK].
[191] *Id.*
[192] *A Comprehensive Guide to Data Cleaning*, KAGGLE (Jun. 2024) at §8 https://www.kaggle.com/code/darrylljk/data-cleaning

preprocessed to remove stop or stem words for natural language processing (NLP).[193] In this step, "features" refer to the measurable properties or independent variables that represent the input to the model.[194] In a dataset predicting house prices, for instance, features might include square footage, number of rooms, and location.[195] For text-based data, tokenization is crucial. Tokenization is the process of breaking down text into smaller units called tokens, which can be individual words, subwords, or even characters, depending on the context and model requirements.[196] This step transforms unstructured text into a structured form that models can process.[197]

Data may also need to be annotated or labeled.[198] Labels and annotations become crucial during supervised fine-tuning or task-specific training stages; they provide the ground truth and contextual information needed to train and guide models effectively. Pre-training data does not typically require labels because pre-training is often conducted using unsupervised learning or self-supervised learning methods. In these methods, the model learns directly from the data itself without requiring explicit labeled outputs.[199] High-quality labels and annotations ensure the model learns accurate patterns and minimizes bias, directly impacting its performance and generalization.[200] Poor-quality data can lead to incorrect predictions, reduced reliability, and potential ethical issues in real-world applications.[201]

E. Based on Synthetic Elements

Data can either be real-world, gathered data, or synthetic data.[202] The key question to be answered here is **what portion of the dataset is synthetically generated?** Distinguishing different types of synthetic data is best done with respect to how dependent the synthetic data is

---

[https://perma.cc/BVK6-ST98] ("Encoding categorical data is important because many machine learning algorithms can only work with numerical data. Categorical data, like gender and cities, needs to be turned into numbers so the algorithms can understand it. We do this encoding step before training our models. It helps the models learn from the data and make accurate predictions.").

[193] *See* Sheikh Muhammad Abdullah, *Text Processing in NLP Basis Steps to Preprocess the Textual Data,* KAGGLE, https://www.kaggle.com/code/abdmental01/text-preprocessing-nlp-steps-to-process-text [https://perma.cc/22L8-Z5C7].

[194] *Understanding Features and Labels in ML Data,* APXML, https://apxml.com/courses/introduction-to-machine-learning/chapter-2-fundamental-concepts/features-labels-explained [https://perma.cc/U6MJ-WRAJ].

[195] Ryan Holbrook, Alexix Cook & Vincent Roseberry, *Feature Engineering for House Prices*, KAGGLE, https://www.kaggle.com/code/ryanholbrook/feature-engineering-for-house-prices [https://perma.cc/9PRV-LZWU].

[196] Murilo Gustineli, *The Art of Tokenization: Breaking Down Text for AI*, TOWARDS DATA SCI. (Sep. 26, 2024), https://towardsdatascience.com/the-art-of-tokenization-breaking-down-text-for-ai-43c7bccaed25/ [https://perma.cc/V2UA-L49C].

[197] *See id.*

[198] While the terms labels and annotations are often used interchangeably in the context of training data for ML, they refer to slightly different concepts. Labels are the specific, structured outputs or targets associated with a data instance that a machine learning model is trained to predict. They represent the ground truth. Annotations are additional metadata or markings applied to raw data that help describe, identify, or interpret it. They provide context or instructions to guide machine learning models, often for complex tasks. *See* JAMES PUSTEJOVSKY & AMBER STUBBS, NATURAL LANGUAGE ANNOTATION FOR MACHINE LEARNING: A GUIDE TO CORPUS-BUILDING FOR APPLICATIONS 12–13 (2012).

[199] For example, in natural language processing (NLP), the model learns from raw text to predict the next word in a sentence (causal language modeling) or fill in missing words (masked language modeling).

[200] NIKOLENKO, *supra* note 13, at 12 (stressing that in supervised learning, data must be labeled, and this labeling must be near as possible to the pixel-level. Models learn better when the data provides a full understanding. This is often unachievable and a costly, laborious process done by hand by human labelers).

[201] *See* PUSTEJOVSKY & STUBBS, *supra* note 200.

[202] *See* Peter Lee, *Synthetic Data and the Future of AI*, 110 CORNELL L. REV. 1, 4 (2025).

on real-world data.[203] How reliant synthetic data is on collected data, or how artificial it is, correlates with the different methods and technologies used to synthesize data.[204] Moreover, each method and technology leveraged to synthesize data yields datasets with different attributes, different uses, and which are not substitutable products. Thus, the guiding classification question begets the sub-question: **how was the synthetic portion of the dataset generated?**

1. Methods and Technologies to Synthesize Data

i. Augmented Data

Synthetic data can be generated by modifying collected data. Generally, this is called data augmentation – data enhanced through transformations or manipulations to increase diversity and improve robustness.[205] Examples of augmented data can be simple, like rotated or flipped images for computer vision tasks or paraphrased sentences for NLP models.[206] Data may be manipulated by data perturbation, where a data scientist adds noise to a real-world dataset to improve model performance and accuracy.[207] The key characteristic of augmented data is that the synthetic dataset is created based on the determined statistical characteristics of the original, real-world dataset. Data scientists frequently combine real-world and synthetic data to boost the representation of under sampled groups.[208] Data augmentation works such that the synthetic data generated utilizing these methods is entirely based on the input data.[209]

ii. Model-based Generators

Another way to generate synthetic data is to use a model-based generator.[210] Synthetic data generators are similar to data augmentation methods in that they create data by learning the statistical patterns of real-world datasets.[211] The key distinction is that while data augmentation transforms and manipulates existing data points, this refers to completely artificial data.[212] For example, to create computer-generated data for a facial recognition system, the synthetic data company Datagen scanned hundreds of real people's faces to train a synthetic data generator that

---

[203] Legal scholarship is united in defining synthetic data based on how "synthetic" the data is, that is, how reliant it is on collected data. *See, e.g.*, *id.* at 17–18; Gal & Lynskey, *supra* note 12, at 1095.

[204] Lee, *supra* note 202, at 24.

[205] Leinar Ramos & Jitendra Subramanyam, *Maverick Research: Forget About Your Real Data - Synthetic Data Is the Future of AI*, GARTNER, at 6 (Jun. 24, 2021).

[206] *Id.*

[207] Allan Tucker et al., *Generating High-Fidelity Synthetic Patient Data for Assessing Machine Learning Healthcare Software*, NPJ DIGIT. MED. 3, 147 (2020), https://doi.org/10.1038/s41746-020-00353-9 [https://perma.cc/5W5V-M7BF] (showing that by adding noise they can achieve synthetic datasets that "are not significantly different from original ground truth data in terms of feature distributions, feature dependencies, and sensitivity analysis statistics when inferring machine learning classifiers…[and where] the risk of generating synthetic data that is identical or very similar to real patients is shown to be low.").

[208] *The Real Promise of Synthetic Data*, MIT NEWS (Oct. 16, 2020), https://news.mit.edu/2020/real-promise-synthetic-data-1016 [https://perma.cc/3XA7-Y87K] ("For example, if a particular group is underrepresented in a sample dataset, synthetic data can be used to fill in those gaps — a sensitive endeavor that requires a lot of finesse.").

[209] Gal & Lynskey, *supra* note 12, at 1098.

[210] Ramos & Subramanyam, *supra* note 207, at 9.

[211] Lee, *supra* note 202, at 24.

[212] *See id.* at 24.

can then create millions of new identities.[213]

iii. GANs

Generative adversarial networks ("GANs") are a groundbreaking ML framework that has played a pivotal role in generating synthetic data.[214] Introduced by Ian Goodfellow and his collaborators in 2014, GANs consist of two separate neural networks—the generator and the discriminator—that are trained simultaneously in a competitive process.[215] The Generator is a generative model that generates new images within the distribution of the training data, and the Discriminator is a model trained to return the probability that the input example from the Generator is synthetic.[216] Essentially, the Generator can be thought of as an art counterfeiter and the discriminator as an art evaluator.[217] The two neural networks learn from each other as adversaries, training each other to be better at their respective tasks.[218]

GANs are a prominent synthetic data generation method because they can close the sim-to-real gap.[219] The sim-to-real gap refers to the performance disparity between systems trained in simulated environments, or with artificial data, and their real-world deployment.[220] This is a critical challenge in robotics, for AVs, and in reinforcement learning.[221] Using GANs to create synthetic data increases the fidelity of synthetic data to real-world data.[222]

iv. Knowledge Distillation

Knowledge distillation utilizes synthetic data technologies to generate a compact, high-quality, and representative synthetic dataset from a larger dataset or directly from a data distribution.[223] This method has emerged as a practical method for generating datasets and enhancing the training of smaller, more efficient models.[224] The process involves using a larger,

---

[213] Sam Forsdick, *Artificial Advantage: Can Synthetic Data Make AI Less Biased?,* RACONTEUR (Aug. 1, 2022), https://www.raconteur.net/technology/artificial-advantage-can-synthetic-data-make-ai-less-biased [https://perma.cc/M3MS-GFN9].

[214] Nehaa Bansal & Samarth Tibdewal, *Synthetic Data Generation Using GANs*, IMPETUS (Oct. 2023), https://www.impetus.com/resources/blog/synthetic-data-generation-using-gans/ [https://perma.cc/BP2Y-NCBH].

[215] *See* Ian J. Goodfellow et al., *Generative Adversarial Nets* 2672, 2674 (Jun. 10, 2014) (unpublished manuscript), https://arxiv.org/pdf/1406.2661 [https://perma.cc/WY6H-J36J].

[216] Norman Ponte, *Short Intro to GANs*, MEDIUM (May 19, 2021), https://normponte.medium.com/short-intro-to-gans-9359c888b806 [https://perma.cc/S7EK-JKG2].

[217] *Id.*

[218] *See id.*

[219] *See id.*

[220] Jingzhi Cui & Yutao Ouyang, *Bridging the Sim-to-Real Gap for Efficient and Robust Robotic Skill Acquisition* 1 (Nov. 5, 2024), https://openreview.net/pdf?id=R5FR5H1fAk [https://perma.cc/V99X-6FEX] (defining the sim-to-real gap as "a critical challenge that hampers the effective transfer of learned behaviors from simulations to real-world scenarios due to discrepancies in sensor performance, environmental conditions, and material properties").

[221] *See* Haoran He et al., *Bridging the Sim-to-Real Gap from the Information Bottleneck Perspective* 1 (2024) (unpublished manuscript), https://arxiv.org/pdf/2305.18464. [https://perma.cc/J4H8-C84X].

[222] Ponte, *supra* note 218.

[223] *See* Geoffrey Hinton et al., *Distilling the Knowledge in a Neural Network* 1 (Mar. 9, 2015) (unpublished manuscript), https://arxiv.org/abs/1503.02531 [https://perma.cc/Y5M8-RJHN] (introducing this form of model compression); *see also* Tongzhou Wang et al., *Dataset Distillation* 1 (2020), (unpublished manuscript), https://arxiv.org/abs/1811.10959 [https://perma.cc/3ZT4-2Y9J].

[224] *See* Vishal Yadav & Nikhil Pandey, *Distillation: Turning Smaller Models Into High-Performance, Cost-Effective Solutions*, MICROSOFT FOUNDRY BLOG (Dec. 6, 2024), https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/distillation-turning-smaller-models-into-high-performance-cost-effective-solutio/4355029 [https://perma.cc/8H6T-ZRCC].

pre-trained model (the teacher) to generate outputs, labels, or synthesized data that guide the training of a smaller model (the student).[225] The aim is to reduce the size and complexity of the dataset while retaining its essential characteristics and statistical properties, enabling efficient and effective use for downstream tasks like training ML models or conducting analysis.[226]

There are a few key aspects to synthetic data distillation. For one, data distillation condenses data representations. Instead of storing or working with the entire original dataset, synthetic data distillation seeks to create a smaller synthetic dataset that captures the key patterns, relationships, and diversity of the original data.[227] Techniques like GANs, Variational Autoencoders ("VAEs"), or diffusion models are often employed to distill the statistical and structural properties of the dataset.[228] These models can then generate new samples that closely resemble the real-world data distribution. Synthetic data distillation reduces the dataset size, which can significantly lower storage requirements and computational overhead for model training.[229]

A popular example of data distillation is the Stanford Alpaca project.[230] Stanford Alpaca is a fine-tuned version of the LLaMA 7B language model, created with a distilled synthetic dataset derived from OpenAI's GPT models.[231] The core technique used in Alpaca is the "self-instruct" method, where the larger model generates new instructions based on a small seed set of human-written prompts, creating a large dataset of diverse instruction-response pairs.[232] Alpaca exemplifies how synthetic data distillation can be used to generate high-quality data without relying on extensive real-world data collection.[233]

Recent advances, such as DeepSeek, highlight the increasing sophistication of teacher models.[234] DeepSeek combines pre-trained language models with retrieval-augmented generation (RAG), allowing for more accurate and contextually rich data synthesis.[235] By leveraging such

---

[225] *See id*.

[226] *See* Ziyao Guo et al., *Towards Lossless Dataset Distillation Via Difficulty-Aligned Trajectory Matching* 1 (2024) (unpublished manuscript), https://arxiv.org/abs/2310.05773 [https://perma.cc/Q93C-ASAZ].

[227] *See* Wang et al., *supra* note 225, at 1.

[228] *See* Sam Bond-Taylor et al., *Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models* 1–7 (Mar. 28, 2022), https://arxiv.org/abs/2103.04922 [https://perma.cc/SS3B-B963].

[229] Jianyang Gu et al., *Efficient Dataset Distillation Via Minimax Diffusion* 1 (Mar. 25, 2024) (unpublished manuscript), https://arxiv.org/abs/2311.15529 [https://perma.cc/S2F9-Q3TZ].

[230] Rohan Taori et al., *Stanford Alpaca: An Instruction-following LLaMA Model*, GITHUB, https://github.com/tatsu-lab/stanford_alpaca [https://perma.cc/2EDC-HE54].

[231] Rohan Taori et al., *Alpaca: A Strong, Replicable Instruction-Following Model*, STAN. CTR. RSCH. FOUND. MODELS, https://crfm.stanford.edu/2023/03/13/alpaca.html [https://perma.cc/3ALV-D9A3] (Alpaca demonstrated the capability of generating high-quality datasets for instruction tuning using outputs from OpenAI's GPT-3.5 model as the teacher).

[232] *Id.* (citing Yizhong Wang et al., *Self-Instruct: Aligning Language Models with Self-Generated Instructions* 1–4 (Ann. Meeting of the Ass'n for Computational Linguistics, 2023)).

[233] Alpaca succeeded in distilling the knowledge of a large-scale model into smaller, more targeted datasets. Alpaca was "trained in less than 2 months for less than $600 on 52k question and answer pairs generated using OpenAI's [GPT-3.5]. At the time of release Alpaca boasted near comparable performance to GPT-3.5." Mikiko Bazeley, *A Pragmatic Introduction to Model Distillation for AI Developers*, LABELBOX: BLOG (Jan. 23, 2024), https://labelbox.com/blog/a-pragmatic-introduction-to-model-distillation-for-ai-developers/ [https://perma.cc/3UJH-ZP63].

[234] *See* DEEPSEEK-AI, *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, 645 NATURE 633, 633–34, 637 (2025) (discussing capabilities of DeepSeek).

[235] *See id.* at app. 13, 60, 72 (discussing capabilities of DeepSeek); *see also* Satyapriya Krishna et al., *Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation* 1, 9, 12 (2025) (unpublished manuscript), https://arxiv.org/abs/2409.12941 [https://perma.cc/993D-29FA].

advancements, knowledge distillation now extends beyond static datasets, enabling dynamic dataset generation tailored to specific tasks.[236] Models like DeepSeek can "teach" smaller models to integrate contextual understanding, effectively distilling not only knowledge but also task-specific strategies.[237]

v. Simulators

Simulators create entirely new virtual worlds from which synthetic data can be harvested.[238] They are computational tools used to model real-world processes and systems.[239] As such, simulators generate data without direct use of real-world data.[240] Rather, the data is generated based on a set of world rules.[241] Simulated environments create novel data that is "not extrapolated from an existing dataset."[242]

Today, simulators are employed in scenarios where real-world data is scarce, expensive, or challenging to collect.[243] Tools like CARLA, an open urban driving simulator, enables the generation of diverse driving scenarios for training and testing AVs.[244] Simulators allow data scientists to manipulate parameters and generate data under specific conditions, making it possible to study edge cases or rare events.[245] They are also cost-effective compared to collecting real-world data.[246] Simulated environments also enable repeatable experiments, generating data that would take many human years to collect.[247]

2. Advantages of Synthetic Data over Real-World Data

Synthetic data is an attractive training data product for developers as it can reduce costs and data barriers in the collection, preparation, and analysis stages of the ML workflow.[248] Table 1 distills some of the main advantages of synthetic data to illustrate where and in which use cases data customers may prefer synthetic data.

---

[236] RAG injects external information into a model dynamically. RAG can be used to fetch information that isn't available to the student or teacher model for a number of reasons, including that the data didn't exist during training. *See* Krishna et al., *supra* note 237, at 1, 9, 12.
[237] Yadav & Pandey, *supra* note 226.
[238] *See* Gal & Lynskey, *supra* note 12, at 1090, 1094, 1100–01, 1104.
[239] *See id.* at 1090, 1100–01, 1104.
[240] *See id.* at 1090, 1100.
[241] *See id.* at 1100.
[242] James Steinhoff, *Towards a Political Economy of Synthetic Data: A Data-Intensive Capitalism That is Not a Surveillance Capitalism?*, NEW MEDIA & SOC'Y 1, 8 (2022).
[243] El Emam, *supra* note 7, at 3, 8–9, 54.
[244] Dosovitskiy et al., *supra* note 22.
[245] El Emam, *supra* note 7, at 34–36 (exemplifying how engineers at NVIDIA compiled a training dataset by having a graphics rending engine to create images of dominoes in different positions, with different textures, and under different lighting conditions, specified by the engineers).
[246] Especially as they cut out the labor and cost of data labeling, enabling pixel-level labels that would be nigh unachievable through manual labeling. NIKOLENKO, *supra* note 13, at 12 ("[I]f you have a three-dimensional virtual scene complete with objects of interest, and images for the dataset are produced by rendering, it means that you automatically know which object every pixel belongs to, what are the 3D relations between them, and so on, and so forth. Producing new data becomes very cheap, and labeling becomes free.").
[247] EMANUEL TODOROV ET AL., *MuJoCo: A Physics Engine for Model-Based Control* 5026, 5033 (2012 IEEE/RSJ Int'l Cong. on Intelligent Robots & Sys. (IROS), (2012), https://homes.cs.washington.edu/~todorov/papers/TodorovIROS12.pdf [https://perma.cc/93Y4-HUMP] (introducing MuJoCo, a simulation platform for robotics research, focusing on dynamic and control problems).
[248] Gal & Lynskey, *supra* note 12, at 1102.

TABLE 1: TRANSFORMATIVE ADVANTAGES OF SYNTHETIC DATA TECHNOLOGY FOR THE DATA MARKET		
ADVANTAGE	PROBLEM IT SOLVES	EXAMPLES
Potentially infinite amounts of task-specific data	Synthetic data can help overcome data scarcity.[249] Because of the way it works, ML models perform better the more data they have to train on.[250] This is particularly useful in domains where real-world data is rare or expensive to collect.	Training medical imaging systems, such as those used for detecting cancer, rare genetic disorders, or uncommon heart conditions, requires large datasets of annotated medical images (e.g., X-rays, MRIs, or CT scans). The NIH Clinical Center used synthetic chest X-rays to train AI for detecting diseases like tuberculosis in populations with limited real-world data.[251]
High-quality labels and annotations	The necessary manual labeling and annotation of real-world data is a tedious and imprecise exercise that falls short of the detail and richness needed to enable the increasingly complex tasks demanded of AI. [252] Obtaining and labeling data often is an enormous time and labor expense, or even infeasible in some real-world settings. The current computer vision data pipeline requires images to be manually processed - humans have to either draw-on or at least verify and correct labels such as bounding boxes or segmentation masks).[253] Synthetic data makes possible higher quality labels: the	Training ML models for agricultural applications, such as crop disease detection, pest identification, or yield estimation, requires large datasets of labeled images.[255] However, annotating agricultural images (e.g., identifying diseased vs. healthy crops) requires domain experts like agronomists or botanists. PlantVillage, a project supported by the United Nations, uses synthetic data to improve the performance of AI-based tools for detecting plant diseases in

[249] Automation with Wilt: Software & AI, *supra* note 49 (showing Ilya Sutskever's full talk at the Conference on Neural Information Processing Systems about sequence to sequence learning with neural networks and explaining how data is the "fossil fuel" of AI).

[250] Joel Hestness et al., *Deep Learning Scaling is Predictable, Empirically* 6 (Dec. 2017) (unpublished manuscript), https://arxiv.org/abs/1712.00409 [https://perma.cc/6MP6-GYYX].

[251] *See NIH Clinical Center Provides One of the Largest Publicly Available Chest X-Ray Datasets to the Scientific Community*, NAT'L INST. HEALTH (Sept. 27, 2017), https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community [https://perma.cc/Q9FD-G5BE].

[252] *See* Yashar Behzadi, *A Community for Synthetic Data is Here and This is Why We Need It*, KDNUGGETS (Apr. 22, 2022), https://www.kdnuggets.com/2022/04/community-synthetic-data-need.html [https://perma.cc/P85C-RSRK].

[253] Azim Ahmadzadeh et al., *A Guide for Manual Annotation of Scientific Imagery: How to Prepare for Large Projects* 2 (Aug. 2025) (unpublished manuscript), https://arxiv.org/pdf/2508.14801v1 [https://perma.cc/P28N-TRL9].

[255] David. P. Hughes & Marcel Salathé, *An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics* 2 (2015) (unpublished manuscript), https://arxiv.org/pdf/1511.08060 [https://perma.cc/V8CB-8HLW] (introducing the PlantVillage dataset); *see also* Tom Miller, *PlantVillage: Supporting Farmers from the Ground Up*, NAT'L PHILANTHROPIC TR. (Apr. 18, 2022), https://www.nptrust.org/philanthropic-resources/philanthropist/plantvillage-supporting-farmers-from-the-ground-up/ [https://perma.cc/2Y9U-9TCL].

	computer itself creates the scene and the assets, which means perfect state knowledge. This in turn means perfect segmentation and rich, detailed annotations that capture not only what is visible, but also the underlying physics of the environment.[254]	underrepresented regions.[256]
Provides and captures edge cases	Edge cases are rare, unusual, or extreme scenarios that occur infrequently in real-world data but can significantly impact the performance of machine learning models.[257] Synthetic data provides a unique advantage in addressing these situations because it allows the deliberate and controlled creation of such scenarios.[258]	The key example here is the ability to generate diverse datasets for autonomous vehicles to train in varied weather and traffic conditions.[259] It is no mystery that the first adopters of early-stage synthetic data were AV companies.
More robust data	Synthetic data need not be similar to real-world data to replace it. Using synthetic data that does not reflect real-world conditions may produce more optimal results.[260] Synthetic data can ensure that the model isn't learning irrelevant parameters.[261] ML techniques like domain randomization improve model performance in new and unseen environments, improving what is defined as a model's generalization capability. These techniques mitigate the sim-to-real gap. Synthetic data does not have to look real to be effective. In fact, having crazy backgrounds and wacky colors, for example, often makes for more robust models.[262]	There are many seminal examples of this. In one, a model was trained to classify images as either huskies or wolves.[263] While it achieved high accuracy during testing, further analysis revealed that the model was not identifying the animals themselves but rather relying on the background features. The model associated snowy backgrounds with wolves and non-snowy backgrounds with huskies because wolves were predominantly photographed in snowy environments in the training dataset. By introducing variations in the background (e.g., snow, grass, urban environments), lighting, and other environmental

[254] *See* NIKOLENKO, *supra* note 13, at 77.
[256] *Id.*
[257] Hsin-Yu Chang et al., *A Survey of Data Synthesis Approaches* 2 (2024) (unpublished manuscript), https://arxiv.org/pdf/2407.03672 [https://perma.cc/QNY8-7QFC] (noting that "addressing edge cases, scenarios that occur at the extreme ends of data distributions, is crucial for achieving comprehensive and reliable performance," and that this is one of the objectives of generating synthetic data.).
[258] *Id.*
[259] Dosovitskiy et al., *supra* note 22.
[260] Chang et al., *supra* note 259, at 2 (categorizing the objectives of conducting data augmentation with synthetic data generation as: improving diversity, data balancing, addressing domain shift, and resolving edge cases).
[261] Wenqian Ye et al., *Spurious Correlations in Machine Learning: A Survey* 3–4 (2024) (unpublished manuscript), https://arxiv.org/html/2402.12715v1 [https://perma.cc/Q6FN-SAAU] (discussing various instances where models learn unintended associations, such as correlating background textures with object classes, leading to misclassifications when these backgrounds change).
[262] *See* Josh Tobin et al., *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World* 6 (2017 IEEE/RSJ Int'l Cong. on Intelligent Robots & Sys., 2017), https://arxiv.org/pdf/1703.06907 [https://perma.cc/3GKJ-6UXU].
[263] Marco Tulio Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* 8 (2016) (unpublished manuscript), https://arxiv.org/pdf/1602.04938 [https://perma.cc/6JAG-VX4M].

		features in synthetic training images, domain randomization ensures the model learns features related to the animals rather than the background.[264]
Mitigates privacy concerns	Real-world data may contain sensitive or personal information, leading to regulatory and ethical challenges. Laws such as GDPR, HIPAA, and CCPA impose strict requirements on handling personal data. Synthetic data mitigates these concerns by providing data that is statistically similar to the original while containing no real-world data.[265]	The privacy safeguards of synthetic data are particularly valuable in health and finance, two highly regulated industries.[266] MDClone, a healthcare data analytics company, generates synthetic patient data that maintains the statistical properties of real data while eliminating any direct ties to real individuals.[267]
Reduces bias	Real-world data may reflect societal or collection biases that could skew model predictions.[268] Synthetic data provides a significant advantage in reducing bias by enabling the deliberate creation of balanced, representative datasets. It allows developers to correct imbalances and ensure fair representation of all groups or scenarios, fostering more equitable AI systems. Synthetic data can be used to reflect norms of equality rather than real-world inequalities.[269]	Facial recognition systems often show biased performance across demographic groups due to imbalances in training datasets.[270] Systems may perform better on lighter-skinned individuals because the training datasets are disproportionately composed of images of lighter-skinned faces, leading to unfair results for underrepresented groups.[271] Bias mitigation efforts, powered by synthetic data, improve system fairness and enhance trust in AI systems deployed in real-world applications like security and authentication.[272] Companies like DataGen generate synthetic datasets of human faces with a wide range of attributes (e.g., skin tone, age, gender, and

[264] *Id.* at 9.

[265] Hasan Chowdhury et al., *AI Improvements Are Slowing Down Companies Have a Plan to Break Through the Wall*, BUS. INSIDER (Nov. 27, 2024, at 4:00 ET), https://www.businessinsider.com/generative-ai-wall-scaling-laws-training-data-chatgpt-gemini-claude-2024-11 [https://perma.cc/3YUG-NLFN].

[266] Lee, *supra* note 202, at 30.

[267] MDCLONE, *Frequently Asked Questions*, https://mdclone.com/faqs/ [https://perma.cc/73XE-ZLBQ].

[268] *See, e.g.*, Uribe, Bosch, & Chenal, *supra* note 137; Shankar et al., *supra* note 140, at 1; Jieyu Zhao et al., *supra* note 142, at 1; Buolamwini & Gebru, *supra* note 143, at 77.

[269] Paul Tiwald et al., *Representative & Fair Synthetic Data* 1 (2021) (unpublished manuscript), https://arxiv.org/abs/2104.03007 [https://perma.cc/X5T3-XDNE] ("We consider representative and fair synthetic data a promising future building block to teach algorithms not on historic worlds, but rather on the worlds that we strive to live in.").

[270] *See* William Crumpler, *The Problem of Bias in Facial Recognition*, CSIS.ORG (May 1, 2020), https://www.csis.org/blogs/strategic-technologies-blog/problem-bias-facial-recognition [https://perma.cc/JEJ3-HY6V]; *see also* Buolamwini & Gebru, *supra* note 143, at 78.

[271] Buolamwini & Gebru, *supra* note 143, at 78.

[272] Forsdick, *supra* note 215.

		hairstyles).[273] These datasets are used to improve fairness in facial recognition models. [274]
Enhanced access to relevant data	Regulatory or competitive barriers may restrict access to valuable real-world datasets.[275] Synthetic data provides enhanced access to relevant data by circumventing barriers associated with real-world data.[276] By enabling access to relevant data that would otherwise be restricted, or unavailable, synthetic data unlocks opportunities for innovation and development across industries. This advantage is particularly critical in fields where real-world data is either highly sensitive, scarce, or costly to obtain.[277]	Financial institutions need vast datasets to train machine learning models for detecting fraudulent activities.[278] Fraudulent transactions are rare compared to legitimate transactions, leading to class imbalance in datasets.[279] Competitive and regulatory constraints limit cross-institution collaboration on fraud detection.[280] Using synthetic data that mimic real-world patterns of legitimate and fraudulent activities can enhance fraud detection systems.[281]
Avoids copyright infringement risk	Using real-world data for ML purposes sometimes risks violating intellectual property laws or contractual obligations, particularly when data is used without proper authorization.[282] Licensing	Training generative AI models (e.g., text-to-image or text-to-video models) requires large datasets of creative works, such as images, videos, and music.

[273] *Id.*

[274] *Id.*

[275] Ellison Anne Williams, *Overcoming Barriers To Unlock Data Value*, TECHFINITIVE (Aug. 28, 2024), https://www.techfinitive.com/opinions/overcoming-barriers-to-unlock-datavalue/#:~:text=In%20this%20article%2C%20we'll,data%20transfers%20and% 20data%20confidentiality [https://perma.cc/8HYB-QT9M].

[276] *See* Lee, *supra* note 202, at 24.

[277] Khaled El Emam, *Could Synthetic Data Be the Future of Data Sharing?*, CPO MAGAZINE (Aug. 25, 2021), https://www.cpomagazine.com/data-privacy/could-synthetic-data-be-the-future-of-data-sharing/ [https://perma.cc/2ZJM-8C7H] (noting that synthetic data can mitigate a number of data access challenges, where data exists but it is rendered inaccessible by social, rather than technical, factors including privacy concerns to proprietary ownership to regulation).

[278] Samuel Assefa et al., *Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls*, JP MORGAN CHASE (June 23, 2020), https://ssrn.com/abstract=3634235 [https://perma.cc/8JYJ-8EUA] (highlighting the growing need for effective synthetic data generation in the financial domain, where data is often siloed for various reasons including regulatory requirements and business needs). *See also* PayPal, *Harnessing Machine Learning Fraud Detection Technologies*, PAYPAL BUS. RES. CTR. (Nov. 5, 2024), https://www.paypal.com/us/brc/article/payment-fraud-detection-machine-learning [https://perma.cc/92YC-VEG9].

[279] Andrea Dal Pozzolo et al., *Learned Lessons in Credit Card Fraud Detection From a Practitioner Perspective*, 41 EXPERT SYS. WITH APPLICATIONS 4915, 4916 (2014).

[280] Prince Grover et al., *Fraud Dataset Benchmark and Applications* 1 (2023) (unpublished manuscript), https://arxiv.org/abs/2208.14417 [https://perma.cc/W4B9-DRFW] (underscoring that, as compared to other well researched fields, fraud detection has unique challenges).

[281] Kevin Levitt, *How Is AI Used in Fraud Detection*, NVIDIA CO. BLOG (Dec. 13, 2023), https://blogs.nvidia.com/blog/ai-fraud-detection-rapids-triton-tensorrt-nemo [https://perma.cc/Z3A4-BTZ6] ("Generating synthetic data is [an] important application of generative AI for fraud prevention. Synthetic data can improve the number of data records used to train fraud detection models and increase the variety and sophistication of examples to teach the AI to recognize the latest techniques employed by fraudsters.").

[282] For example, several copyright infringement suits allege that AI companies used copyrighted content in their training datasets without consent or compensation to the rights holders. *See, e.g.*, Complaint, Doe I v. Github, Inc., No. 4:22-cv-06823 (N.D. Cal. Nov. 3, 2022); Complaint, Andersen et al. v. Stability AI et al., No. 3:23-cv-00201 (N. D. Cal. Jan. 13, 2023); Complaint, Getty Images (US), Inc. v. Stability AI, No. 1:23-cv-00135 (D. Del. Feb. 3, 2023). *See also* Lee, *supra* note 202, at 19–20.

	restrictions on training data can prevent companies from commercially using models trained on real-world data.[283] Synthetic data mitigates these risks by providing an alternative that does not rely on copyrighted material, offering a clear advantage for businesses and researchers.[284] Even in close cases, synthetic data may be considered transformative because it re-creates the properties of the original data in a new form, without including the original content.[285]	Synthetic data provides a viable alternative by creating training datasets that replicate the statistical and stylistic features of real-world content without copying copyrighted works.[286]
Cheaper Data	Collecting real-world data can be very expensive, whether this be gathering first-party data or purchasing third-party data from an aggregator.[287] Synthetic data can also reduce storage costs: maintaining large databases as a competitive moat becomes superfluous; synthetic data need not be stored, as it can be easily recreated; and there is no need to store data for long periods of time before enough is accumulated for a meaningful analysis.[288] Of course, depending on the use case, synthetic data could also be expensive. Additionally, the synthetic data is still quite young, but it is possible that the generation of synthetic data could become concentrated in the hands of a few	Retail companies rely on large-scale transaction data to develop and refine systems like recommendation engines to suggest relevant products to customers.[289] However, gathering and maintaining real-world transaction data across multiple stores or platforms can be expensive.[290] Amazon uses synthetic datasets to simulate customer behaviors, enabling the development of robust recommendation engines without requiring vast amounts of costly,

[283] For example, copyleft licenses (e.g., CC BY-NC-SA 4.0) on some open-source datasets impose restrictions that prevent developers from commercializing models trained on those datasets. *See* Kacper Szkalej & Martin Senftleben, *Mapping the Impact of Share Alike/Copyleft Licensing on Machine Learning and Generative AI*, OPEN FUTURE FOUND. (June 12, 2024), https://openfuture.eu/wp-content/uploads/2024/06/Share-Alike-and-ML-Report-FINAL.pdf [https://perma.cc/TF9Q-T7ZY].

[284] Gareth Kristensen et al., *Training Artificial Intelligence Models on Synthetic Data: No Silver Bullet for Intellectual Property Infringement Risk in the Context of Training AI Systems (Part 1 of 4)*, CLEARY GOTTLIEB (Dec. 12, 2023), https://www.clearyiptechinsights.com/2023/12/training-ai-models-on-synthetic-data-no-silver-bullet-for-ip-infringement-risk-in-the-context-of-training-ai-systems-part-1-of-4/ [https://perma.cc/BHQ2-V9UA].

[285] *Id.* ("Synthetic data . . . should not (in theory) give rise to the same IP infringement risks as using real-world data. Synthetic data mimics real-world data and, if properly developed, should be technically and statistically indistinguishable from such data for the purpose of training AI models.").

[286] *Id.* In December 2022, Stability AI announced plans to allow artists to opt out of having their work included in the training dataset for the upcoming Stable Diffusion 3.0 release. Artists can register and manually flag their images in the LAION database to exclude them from future training. This decision came amid growing concerns from artists about the use of their work in AI training without consent. *See* Benj Edwards, *Stability AI Plans to Let Artists Opt Out of Stable Diffusion 3 Image Training*, ARS TECHNICA (Dec. 15, 2022, at 17:42 ET), https://arstechnica.com/information-technology/2022/12/stability-ai-plans-to-let-artists-opt-out-of-stable-diffusion-3-image-training/ [https://perma.cc/U658-DUCK].

[287] *See* Behzadi, *supra* note 254.

[288] *See* Gal & Lynskey, *supra* note 12, at 1103.

[289] *See* Gopi Kandukuri, *Personalization and Recommendation Engines: The Power of Data in Retail*, SAXON (Jan. 2, 2024) https://saxon.ai/blogs/personalization-and-recommendation-engines-the-power-of-data-in-retail/ [https://perma.cc/7AQ7-TF93].

[290] Gal & Lynskey, *supra* note 12, at 1103.

	market players and drive up prices for training datasets.	real-world data.[291] It generates synthetic scenarios for rare events, such as promotions or sales spikes, to enhance demand forecasting models.[292]
Iterative data	Synthetic data provides an iterative advantage in the development and refinement of machine learning models and analytics systems by enabling continuous, flexible, and efficient improvements in the data generation and utilization process.[293] Humans learn in a similar way, consuming previously generated data and using it to create new knowledge. Learning happens in a continuous, compounded manner. Unlike real-world data, synthetic data can be generated, modified, and augmented repeatedly to suit evolving project needs, fostering rapid iteration cycles.[294]	Robotic manipulation requires training models to handle complex tasks, such as grasping and manipulating objects with precision.[295] However, collecting real-world data for every possible scenario costs significant time and effort, especially for rare or intricate tasks.[296] OpenAI's Dactyl project addresses this by generating synthetic data to simulate a variety of object interactions, including solving Rubik's cubes.[297] Engineers iteratively retrain models using these synthetic datasets and refine them based on feedback to better capture edge cases, such as objects slipping or becoming partially obstructed.[298] This approach helps robots generalize their skills and perform

[291] *See* Lukas Wenzel, David Boldt & Johannes Langer, *Generate Synthetic Data for Evaluating RAG Systems Using Amazon Bedrock*, AWS MACH. LEARNING BLOG (Sep. 23, 2024), https://aws.amazon.com/blogs/machine-learning/generate-synthetic-data-for-evaluating-rag-systems-using-amazon-bedrock/ [https://perma.cc/46TC-56LC] (explaining how the Amazon Bedrock service allows developers to create synthetic data to simulate user queries); *see also* Alak Eswaradass & Xin Huang, *Deep Demand Forecasting with Amazon SageMaker* (June 28, 2022), https://aws.amazon.com/blogs/machine-learning/deep-demand-forecasting-with-amazon-sagemaker/ [https://perma.cc/6UGH-PBSB] (explaining that by incorporating synthetic scenarios into their training data, Amazon's forecasting models can better anticipate and respond to anomalies like sudden sale spikes or promotions, leading to more accurate inventory and supply chain management).

[292] *See id.*

[293]Reyhane Askari Hemmat et al., *Feedback-guided Synthesis for Imbalanced Classification* 1 (Sep. 9, 2024) (unpublished manuscript), https://arxiv.org/pdf/2310.00158 [https://perma.cc/RCP2-F2Y4] (introducing a framework that uses feedback from classifiers to guide the generation of synthetic data, allowing for the continuous refinement of training data and thus enhancing model performance on underrepresented classes and enabling flexible adjustments to evolving data distributions).

[294] *7 Best Practices for Synthetic Data Generation*, YDATA (Aug. 16, 2024), https://ydata.ai/resources/synthetic-data-generation-best-practices [https://perma.cc/8VDY-67J4] ("Synthetic data generation is an iterative process. Initial datasets may require refinement to improve their utility, accuracy, or realism. Continuous feedback and iteration are key to achieving high-quality synthetic data.").

[295] OpenAI et al., *Learning Dexterous In-Hand Manipulation* (Jan. 18, 2019) (unpublished manuscript), https://arxiv.org/abs/1808.00177 [https://perma.cc/2PTA-X3SJ]. *See also* Josh Tobin et al., *Learning Dexterity*, OPENAI (July 30, 2018), https://openai.com/index/learning-dexterity/ [https://perma.cc/6GKP-56PL].

[296] *See id.*

[297] *Id.*

[298] *Id.*

		effectively in unpredictable real-world environments.[299]

3. Limitations of Synthetic Data

Although advantageous in many circumstances, synthetic data is not a cure-all for the limitations and risks of real data.[300] Synthetic data, like real data, can be inaccurate and can enhance bias or perpetuate misrepresentations.[301] Additionally, depending on how it is deployed, privacy concerns[302] and IPR risks may persist despite the use of synthetic data.[303]

i. Inaccuracy

Synthetic data can be inaccurate.[304] Synthetic data generation techniques, such as GANs, often aim to approximate the statistical properties of the original data.[305] There are at least two consequences. First, incorrect or misrepresentative input data or background information may result in models that fail to capture the complex patterns and relationships present in real-world datasets. Second, without conscientious design in the generation of synthetic data (for example, making erroneous assumptions about correlations among the relevant variables), the data created may fail to represent rare events or subtle associations, leading to inaccuracies that undermine the data's utility for critical applications.

Synthetic data that is too different from reality, that is, data that fails to bridge the sim-to-real gap, can result in skewed models and predictions.[306] Even state of the art simulators may fail to capture real-world complexities like friction, lighting, or sensor noise.[307] There are also issues of overfitting: models trained in simulations may exploit unrealistic or idealized features that don't exist in the real world.[308]

---

[299] *Id.*
[300] *See* Gal & Lynskey, *supra* note 12, at 1109–10.
[301] *Id.*
[302] *Id.* at 1125.
[303] *See id.* at 1115.
[304] *See id.* 1109.
[305] *See* El Emam, *supra* note 7, at 2–5; *see also* Ponte, *supra* note 218.
[306] *See* Patrick Yin et al., *Rapidly Adapting Policies to the Real World via Simulation-Guided Fine-Tuning* 1–2 (Int'l Conf. on Learning Representations, 2025), https://arxiv.org/pdf/2502.02705 [https://perma.cc/X2UB-LJZ9] (explaining how synthetic or "simulation-generated data is not a silver bullet, as it provides cheap but ultimately *off domain data*.").
[307] *Id.* ("[P]hysics engines are fundamentally misspecified approximations to reality . . . [which] makes direct zero-shot transfer from simulation to reality challenging" . . . and "[i]n these regimes, *no choice of parameters for the physics simulator accurately capture the realworld dynamics.*").
[308] *See* Yin et al., *supra* note 308, at 2 ("[D]espite impressive performance for many tasks, methods that transfer policies from simulation to reality zero-shot (citation omitted) run into failure modes when they encounter novel dynamics not covered by the simulator"); *see also* El Emam, *supra* note 7, at 4, 12–13 (noting that overfitting can create nontrivial privacy problems, where the synthetic dataset is indistinguishable from the real data).

IBM Watson Health's early challenges in healthcare AI provide a notable real-world example of the limitations of synthetic data and its impact on accuracy.[309] Watson Health was designed to assist medical professionals in diagnosing and recommending treatments for cancer patients.[310] However, when applied in real-world hospital settings, Watson Health struggled to handle diverse, real-life cases that deviated from the patterns present in the training data.[311] Watson Health's AI was trained on synthetic patient scenarios and curated data that did not adequately represent the complexity and variability of real-world clinical environments; the synthetic training data primarily reflected idealized or theoretical scenarios.[312] In some cases, Watson Health recommended treatments that were deemed inappropriate or unsafe by doctors.[313] This potential inaccuracy may lead developers to steer away from choosing synthetic datasets to train their models in exacting fields.

ii. Amplifying bias or perpetuating misrepresentations

Synthetic data can enhance biases or perpetuate misrepresentations.[314] Where synthetic data is a derived product, biases present in the original data can be amplified during the generation process, particularly if the generation algorithm lacks robustness.[315] If real-world data disproportionately represents certain demographics, the synthetic data may inherit these biases, leading to models that reflect existing inequities.[316] Synthetic data generation cannot inherently correct biases without explicit intervention, underscoring the importance of careful bias auditing and validation when using synthetic data in AI-driven decision-making processes.[317]

iii. Not a solve-all for privacy

Synthetic data does not eliminate all privacy risks. Deidentification processes, like

---

[309] *See* Angela Chen, *IBM's Watson Gave Unsafe Recommendations for Treating Cancer*, THE VERGE (July 26, 2018, 16:29 ET), https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science [https://perma.cc/B5JH-ZMQ5]; Sandeep Konam, *Where Did IBM Go Wrong with Watson Health?*, QUARTZ (July 20, 2022), https://qz.com/2129025/where-did-ibm-go-wrong-with-watson-health [https://perma.cc/8FSN-XF6P].

[310] *See id.*; *see* Meg Bryant, *STAT: IBM's Watson Gave 'Unsafe and Incorrect' Cancer Treatment Advice*, HEALTHCARE DIVE (July 26, 2018), https://www.healthcaredive.com/news/stat-ibms-watson-gave-unsafe-and-incorrect-cancer-treatment-advice/528666/ [https://perma.cc/5QLX-2YCL].

[311] *See id.*

[312] Konam, *supra* note 311 ("[Watson Health's cancer diagnostics tool] was not trained with real patient data, but instead with hypothetical cases provided by a small group of doctors in a single hospital. Hand-crafted or synthetic data aren't necessarily bad, but Watson didn't seem to account for the fact that this data reflected the doctors' own biases and blindspots and wasn't necessarily generalizable to all patient cases. As a result, the tool was accused of making inaccurate and unsafe recommendations, leading high-profile hospital partners to cancel their collaborations with Watson.").

[313] *See* Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ [https://perma.cc/P9S6-KQ5G].

[314] Forsdick, *supra* note 215 ("Synthetic data also isn't a silver bullet for AI bias…any biases that are present in the data-generation process – whether intentionally or unintentionally – will be picked up by models trained on it.").

[315] *See* Aman Gupta et al., *Transitioning From Real to Synthetic Data: Quantifying the Bias in Model* 4 (2021) (unpublished manuscript), https://arxiv.org/abs/2105.04144 [https://perma.cc/8WHW-TPYK] (studying the bias in synthetic datasets generated by different methods).

[316] *Id.* at 2–6.

[317] *Id.* at 2.

anonymization or pseudonymization, can be reversed.[318] Poorly designed generation techniques may inadvertently retain patterns or correlations from the original dataset, enabling re-identification attacks when combined with auxiliary data sources.[319] Additionally, if sensitive features are overrepresented or improperly generalized, adversaries may still infer private information about individuals indirectly.[320] This vulnerability undermines the intended privacy benefits of synthetic data.

iv. Enduring IP risks

To the extent that synthetic data is derived from real-world datasets, which may themselves be subject to copyright, trade secret, or data ownership laws, there may be enduring intellectual property (IP) rights risks. Although synthetic data is generated to be distinct from the original data, it can sometimes retain underlying patterns or features that are too similar to the source material, raising questions about whether it infringes on the IP rights of the original dataset.[321]

The ongoing *Andersen v. Stability AI*[322] case may preview future IP issues with synthetic data. There, Stability AI is accused of training its model on copyrighted artworks without consent, leading to outputs resembling those works.[323] Though not about synthetic data, it is important to remember that synthetic data can be model-generated data. As such, the *Andersen* case illustrates a critical concern with synthetic data: although the data is technically new, when it is derived from existing copyrighted material, courts still could find that it violates copyright laws if it closely resembles the original works or reproduces distinctive elements.[324] The challenge lies in determining when synthetic data crosses the line from inspiration to infringement, especially when models produce outputs that mimic the style or content of copyrighted materials.[325]

Finally, synthetic data complicates the attribution of ownership and control. For instance, if synthetic data is created from multiple datasets with varying ownerships, determining who holds rights over the derivative data becomes challenging.

---

[318] *See* Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets* 1 (2008) (unpublished manuscript), https://arxiv.org/pdf/cs/0610105 [https://perma.cc/8YAR-JHEZ] (In fact, advances in quantum computing are demonstrating that no encryption method is truly unbreakable); *see* Tammy Xu, *What Are Quantum-Resistant Algorithms—And Why Do We Need Them?*, MIT TECH. REV. (Sep. 14, 2022), https://www.technologyreview.com/2022/09/14/1059400/explainer-quantum-resistant-algorithms [https://perma.cc/F5YB-UNNP].

[319] *See* Bristena Oprisanu, Georgi Ganev & Emiliano De Cristofaro, *On Utility and Privacy in Synthetic Genomic Data* 2 (Proc. of the 29th Network and Distributed Sys. Sec. Symp., 2022), https://www.ndss-symposium.org/wp-content/uploads/2022-92-paper.pdf. [https://perma.cc/PGG8-NB58] (for example, even synthetic genetic data may not fully protect study participants from privacy attacks).

[320] Luc Rocher, Julien Hendrickx & Yves-Alexandre de Montjoye, *Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models,* NATURE COMMC'N, July 2019, at 1, 6 (finding that even heavily sampled anonymized datasets may not meet modern anonymization standards, challenging the effectiveness of traditional de-identification methods).

[321] *See* Emanuel Maiberg, *OpenAI Can't Fix Sora's Copyright Infringement Problem Because It Was Built With Stolen Content*, 404 MEDIA (Nov. 12, 2025, at 10:44 ET), https://www.404media.co/openai-cant-fix-soras-copyright-infringement-problem-because-it-was-built-with-stolen-content/ [https://perma.cc/U3P7-ZHG9] (highlighting the IP risks posed by generative models trained off of datasets able to be recompiled into their original forms even past the synthetic, model generated veil).

[322] No. 3:23-cv-00201 (N.D. Cal. filed Jan. 13, 2013).

[323] Complaint at 1–2, Andersen v. Stability AI, No. 3:23-cv-00201 (N.D. Cal. filed Jan. 13, 2013).

[324] *See* Zach Schor, *Andersen v. Stability AI: The Landmark Case Unpacking the Copyright Risks of AI Image Generators*, NYU J. INTELL. PROP. & ENT. L. (Dec. 2, 2024), https://jipel.law.nyu.edu/andersen-v-stability-ai-the-landmark-case-unpacking-the-copyright-risks-of-ai-image-generators/ [https://perma.cc/BH3U-YKW9].

[325] *See id.*

F. Based on Sensitivity

The product market boundaries of data can also be defined based on the sensitivity of the data. The sensitivity of data refers to the potential harm or risks associated with unauthorized access, misuse, or exposure of the data.[326] Sensitivity should be understood as a separate characteristic than ownership. The key questions for developers when considering data are: *does the dataset contain data that may be considered confidential? Does the dataset contain data that may be considered sensitive in any way (from data that may reveal personally identifiable information (PII) to political opinions)?* A data's sensitivity helps ensure developers implement the right controls and safeguards when training a model. Data may be public data, internal or confidential data, or protected data.[327]

Public data or non-sensitive data is data that is publicly available and does not pose any risks if accessed or used by anyone.[328] Public data can include government-released open datasets, such as census data, open-source datasets like ImageNet[329] or Common Crawl[330], and publicly available research papers or statistics. Because of its artificial nature, synthetic data will generally also be non-sensitive data.[331] When the dataset required has sensitive attributes, developers may choose to utilize synthetic data instead, which is often used to mimic sensitive data without exposing real data or represent generalized patterns without mirroring sensitive information.[332]

Internal or confidential data is sensitive data and includes data that is exclusive to an organization and provides strategic value and more generally, data where unauthorized disclosure can lead to legal disputes or loss of a competitive advantage.[333] This data is generally non-personal

---

[326] Tim Herr, *Sensitive Data Classification Explained: A Security and Compliance Guide*, FORCEPOINT (Dec. 4, 2025), https://www.forcepoint.com/blog/insights/sensitive-data-classification%E2%80%AF [https://perma.cc/4SGS-PMJZ].

[327] The definition of public, confidential, and protected data varies across contexts (i.e., legal, organizational, and technical), but the core principles are consistent. Notably, the National Institute of Standards and Technology (NIST) and the International Organization for Standardization (ISO) provide frameworks for assessing and classifying data based on sensitivity and potential impact, guiding individuals and organization in implementing appropriate security measures for data. *See* ERIKA MCCALLISTER, TIM GRANCE & KAREN SCARFONE, NAT'L INST. OF STANDARDS & TECH., NIST SPECIAL PUBL'N 800-122, GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII) §3 (Apr. 2010), https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf [https://perma.cc/R8BD-RS3K]; ISO, *ISO/IEC 27001:2022 Information Security, Cybersecurity, and Privacy Protection – Information Security Management Systems – Requirements*, at Annex A, Control 5.12 (2022), https://www.iso.org/standard/27001 [https://perma.cc/LVC3-EYJV].

[328] *See* MCCALLISTER, GRANCE & SCARFONE, *supra* note 330, at § 3 (the NIST addresses the classification of sensitive data through the determination of PII Confidentiality Impact Levels. These levels—low, moderate, and high—indicate the potential harm that could result to individuals or organizations if PII were inappropriately accessed, used, or disclosed. The potential impact is low when data is public or non-sensitive).

[329] IMAGENET, http://www.image-net.org [https://perma.cc/5L4S-WWCM] (last visited Dec. 10, 2025); *see* Jia Deng et al., *ImageNet: A Large-Scale Hierarchical Image Database* 1 (2009 IEEE Conf. on Comput. Vision & Pattern Recognition, 2009), https://ieeexplore.ieee.org/document/5206848 [https://perma.cc/KR5P-6D27].

[330] COMMON CRAWL, https://commoncrawl.org [https://perma.cc/F6V9-R9BW] (last visited Feb. 22, 2026) (open repository containing petabytes of web crawl data collected regularly since 2008, including raw web page data, metadata extracts, and text extracts).

[331] *See* Lee, *supra* note 202, at 23–32.

[332] *Id.*

[333] For the purposes of this Article, data that is protected by legal or regulatory protections is classified as protected data, not confidential data. In common categorization methods, confidential data may refer to data that must remain private and protected at all times and internal data is data that isn't intended for public release but may be accessed via legislative regimes like FOIA. *See, e.g.*, Paula Smith & Adam Roberts, *Guide To Data Classification: Sensitive vs. Public* , RECORDPOINT BLOG (Jan. 21, 2025), https://www.recordpoint.com/blog/a-guide-to-data-classification-confidential-vs-sensitive-vs-public-information [https://perma.cc/3NS6-G6EK]; Chris Brook, *5 Common*

data that is sensitive for organizational or business reasons.[334] Trade secrets, business strategies, and intellectual property are examples of confidential data.[335]

Protected data is highly sensitive.[336] This type of data is subject to legal or regulatory protections and includes personal data, health data, financial data, biometric data, and children's data.[337] Personally Identifiable Information (PII) and more specifically biometric data applies to any data that can identify an individual, directly or indirectly.[338] Health data is data specifically related to an individual's physical or mental health and will be subject to sector-specific regulations, like HIPAA in the United States.[339] Children's data, or data related to individuals under the age of 13 in the United States per COPPA, commands significant protection.[340] Protected data will also include highly sensitive restricted data like classified government documents. When protected data is necessary to train a model, for example, a medical treatment AI, the need to comply with the appropriate regulations and apply anonymization or pseudonymization where possible will be an important attribute considered by a developer in the market.

G. Based on Ownership and Accessibility

Finally, data can be classified based on ownership. Depending on its owner, access to training data can be difficult to impossible, legally uncertain, or even ethically unsound.[341] Moreover, data distributed under license or other IP restrictions may come with downstream distribution and commercialization restrictions. Data ownership and accessibility is a spectrum, with important caveats for licenses and IP rights when and where they may arise. At one end of the spectrum are open datasets; at the other, proprietary data. The key questions developers ask include: **is the dataset distributed under a copyright or other IP license? What terms of use govern the dataset? Are there any restrictions to downstream use of the dataset?**[342]

Open data is freely available for anyone to access, use, and share without significant restrictions. Data may be fully open, with no restrictions, or that openness may be conditional, as in cases where use requires attribution or adherence to specific licenses, like Creative Commons

---

Data Classification Types, FORTRA BLOG (Nov. 23, 2022), https://www.digitalguardian.com/blog/5-common-data-classification-types [https://perma.cc/S53Q-NURD].

[334] *See* Smith & Roberts, *supra* note 336.

[335] *See id.*

[336] *See id.*

[337] Brook, *supra* note 336.

[338] *See* MCCALLISTER, GRANCE & SCARFONE, *supra* note 330, at §2.1–2.2 ("PII is ―any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.").

[339] The Health Insurance Portability and Accountability Act (HIPAA) protects a specific type of data called Protected Health Information (PHI). PHI includes any individually identifiable health information that is created, received, maintained, or transmitted by a healthcare provider, health plan, or healthcare clearinghouse. 45 C.F.R. §§ 160, 162, 164 (2023).

[340] The Children's Online Privacy Protection Act (COPPA) protects the personal information of children under the age of 13 who use online services, including websites, apps, and other digital platforms. It applies to operators of commercial websites or online services directed at children or that knowingly collect personal information from children. 16 C.F.R. § 312 (2023).

[341] Steinhoff, *supra* note 244.

[342] *See* Gebru et al., *supra* note 77, at 9.

or Open Data Commons.[343] Open datasets include public datasets provided by governments and open-source projects like the COCO[344] dataset for computer vision.[345]

Proprietary data is data owned by an individual, an organization, or an entity, and often comes with restricted access or usage.[346] As with open data, there are levels of proprietary restrictions.[347] Some data may be strictly proprietary, where the data cannot be shared or accessed outside the owning entity, other data may be licensed for use, that is, accessible under specific terms, such as a subscription or contract.[348]

When using data to train AI, licensing becomes a critical factor that can significantly impact downstream applications.[349] Licensing agreements dictate how data can be used, shared, and modified, and failing to comply with these terms can lead to legal and ethical complications.[350] For example, proprietary data or datasets under restrictive licenses may prohibit commercial use, redistribution, or derivative works, limiting the scope of models trained on such data.[351] Moreover, unclear or ambiguous licensing terms can create risks of unintentional misuse, especially when datasets combine open and proprietary sources.[352] Developers must carefully evaluate the licensing terms of training data to ensure compliance and to avoid restricting the usability or legality of their AI systems in the future.[353]

Dataset ownership will also affect the maintenance of a dataset. Training a dataset is an iterative process. Further, datasets may need to be aggregated, processed and collected in the cloud.[354] For developers, the following questions will be important: *who is hosting and maintaining the dataset? How will the dataset be updated? Will older versions of the dataset*

---

[343] Himel Mondal, *Copyright and Creative Commons Licenses in Scholarly Publishing: A Practical Guide for Researchers*, J. OF CLINICAL IMAGING SCI., Sep. 2025, at 1, 3.

[344] COMMON OBJECTS IN CONTEXT, https://cocodataset.org/#home [https://perma.cc/J5YC-WETU] (COCO is a large-scale object detection, segmentation, and captioning dataset).

[345] For example, the National Institutes of Health (NIH) provides access to biomedical and genomic data to foster innovation in healthcare. *See Accessing Data & Publications*, NAT'L INST. OF HEALTH, https://sharing.nih.gov/accessing-data [https://perma.cc/KV7D-NZGN].

[346] *What is Proprietary Data? Understanding Its Importance and Context*, INITIAL DATA OFFERING (Aug. 1, 2025), https://initialdataoffering.com/blog/what-is-proprietary-data-understanding-its-importance-and-context/ [https://perma.cc/LNU9-NB7N].

[347] *What Are the Four Levels of Data Classification?*. CAL. LEARNING RES. NETWORK (July 2, 2025), https://www.clrn.org/what-are-the-four-levels-of-data-classification/ [https://perma.cc/RB6F-24G4].

[348] *See id.*

[349] John Lanza, *Eight Tips for Navigating the Complex World of AI Licensing*, IP WATCHDOG (Jan. 16, 2025, at 11:15 EST), https://ipwatchdog.com/2025/01/16/eight-tips-navigating-complex-world-ai-licensing/ [https://perma.cc/JK72-TZRL].

[350] *See, e.g.*, Jeffrey Neuburger & Jonathan Mollod, *Data License Restrictions In the AI Spotlight: Careful Drafting Is More Important Than Ever*, JDSUPRA (Dec. 16, 2025), https://www.jdsupra.com/legalnews/data-license-restrictions-in-the-ai-7860639/ [https://perma.cc/D9AY-8BN3] ("A recently-filed federal court complaint tests the enforceability of restrictive terms in a data license against the use of licensed data for generative AI purposes.").

[351] *See id.*

[352] *See id.*

[353] *See id.*

[354] *See* Nick Jones, *Cloud Computing RFI: What We Heard and Learned* n.13, FED. TRADE. COMM'N TECH BLOG (Nov. 16, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/11/cloud-computing-rfi-what-we-heard-learned#ftn13 [https://perma.cc/PHG7-8PY8] ("Generative AI products are heavily reliant on cloud providers."). *See also Generative AI Raises Competitive Concerns*, FED. TRADE. COMM'N TECH BLOG (June 29, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns [https://perma.cc/PR9K-L6T6] (noting the relationship between cloud providers and large AI training databases).

continue to be supported?[355]

IV. CONCLUSION

The idea that training data is a monolithic entity is not useful for a successful competitive analysis involving data in an AI or AI-adjacent industry. Characteristics of the datasets used to train AI—such as purpose, source, structure, processing, generation, sensitivity, and ownership—are important considerations when defining the relevant product market. These attributes may, for example, affect competitors' ability to access the same or similar data, obtain or create equivalent datasets, or otherwise compete when training a competing AI system.

As AI adoption continues to grow in many or most industries, the role of training data as a critical input cannot be overstated. The emergence of synthetic data further complicates the legal landscape, offering new opportunities for AI training while challenging traditional notions of data substitution and market dynamics. Not all data is created equal, and the use of myriad types of datasets carries profound implications for both AI development and antitrust analysis. Regulators and practitioners need a framework for discussing data, and that framework must capture the details that may matter in a competitive analysis.

By embracing a classification system that accounts for the attributes of training data, practitioners can better evaluate competitive concerns, understand the substitutability of different types of data, and craft policies that align with the realities of the modern AI ecosystem. In doing so, they will help pave the way for a future where AI thrives as a transformative force, powered by a diverse and competitive data landscape.

---

[355] *See* Gebru et al., *supra* note 77, at 10.