Data as the New Oil: Parallels, Challenges, and Regulatory Implications

Chiara Farronato
Harvard University, CEPR, NBER

January 2025

The impact of technological innovation often relies on harnessing previously overlooked or underutilized resources, as illustrated by the historical progression of oil and the contemporary role of data. Just as the development of drilling and refining technologies revolutionized the utility of oil in the second half of the 19th century, a series of technological developments in past decades, most recently exemplified by the advent of generative artificial intelligence (GenAI), are unlocking new potential of data. These series of innovations turned raw inputs—crude oil and raw data—into valuable outputs, fundamentally reshaping industries. However, the emergence of such technologies also raises critical questions about property rights and other forms of regulation. The metaphor "data is the new oil" often arises in discussions, both within academic circles—such as at the NBER conference for this handbook—and beyond. In this short article, I delve into this analogy, exploring the similarities and differences between oil and data as transformative resources.

The association between data and oil is widely attributed to Clive Humby, a British mathematician and data scientist.[1] In 2006, Humby emphasized that data, like crude oil, is only valuable in combination with technologies that refine it into useful outputs, be it consumer insights for product improvements or response suggestions to customer service inquiries. This metaphor underscores the shared trajectory of oil and data: both became transformative resources only through significant technological advancements that unlocked their potential. However, this parallel also highlights an important challenge. Just as oil's extraction and use brought about complex regulatory and societal issues, the emergence of data as a valuable resource introduces equally difficult questions about regulatory and societal implications.

The fact that innovation brings about regulatory uncertainty is not new. In fact, new businesses often take advantage of legal gray areas to drive innovation. There are many such examples in the most recent digital age. Electronic commerce enjoyed a tax advantage in the early days because a 1992 Supreme Court ruling[2] required businesses to have a physical presence in a state for the sales taxes to be applicable. Companies like Amazon and eBay (Einav, Knoepfle, et al. 2014) could thus charge lower prices compared to brick-and-mortar retailers, contributing to their growth. Eventually in 2018, the Supreme Court overturned its initial ruling,[3] thus allowing states to collect sales taxes from online retailers. More recently, ride-sharing and short-term rental platforms also entered markets while avoiding the regulations that govern incumbents. At least initially, Uber and Lyft avoided requirements such as taxi medallions and commercial licensing and insurance. Similarly, Airbnb avoided lodging taxes and did not have to comply with zoning laws. Eventually, regulators caught up with the new service platforms by applying the existing regulation to the new providers, creating new regulations, or banning the new providers altogether.

These examples have in common the fact that the innovators' approach relied on exploiting gaps and ambiguity in several existing laws (Einav, Farronato, and Levin 2016), of which Beraja and Yuchtman 2025 in this handbook emphasize property rights. In these contexts, it is often impossible to clarify the applicable regulation ex ante. First, the parties with whom the innovator would have to contract are many and fragmented: states in the case of sales taxes, and local jurisdictions in the case of lodging taxes and taxi medallions. Second, regulators can often be influenced by incumbents (Callander and Li 2025 in this

[1] https://www.forbes.com/sites/nishatalagala/2022/03/02/data-as-the-new-oil-is-not-enough-four-principles-for-avoiding-data-f
[2] Quill Corp. v. North Dakota (1992), https://supreme.justia.com/cases/federal/us/504/298/
[3] South Dakota v. Wayfair, Inc. (2018), https://www.supremecourt.gov/opinions/17pdf/17-494_j4el.pdf.

handbook) with more lobbying resources and a strong interest not to be disrupted. Finally, regulation can take a long time, increasing investment costs and the risks of innovation obsolescence. As an example, news aggregators such as Google News, which will be discussed later, emerged in the early 2000s, whereas regulation over the aggregators' use of snippets and hyperlinks from original publishers was not approved until at least 2019 in the European Union and even later in other countries.

Oil and data are alike in that radical innovations have leveraged these inputs. However, it is worth highlighting that it was not a single technological innovation that unleashed their potential. For oil, the drilling of oil wells was not enough. In fact, shortly after 1859, when Edwin Drake drilled his first well in Titusville, Pennsylvania, it was shut down because it was not profitable. The price of crude oil, mostly only usable as lamp fuel in its raw form, was too low to justify the project.[4] The oil well was followed by a long series of technological advancements over the following several decades—from distillation to cracking to more complex refining processes—which allowed for oil to be transformed into usable (and very valuable) products, such as gasoline or plastics (Yergin 2011).

In this respect, the "data as oil" metaphor is well fitting. Humby's 2006 quote, over a decade before the current GenAI revolution, was likely motivated by the early applications of data science and machine learning to electronic commerce. Google's PageRank algorithm was one of the earliest applications of using information about webpages to rank them in search results (Page et al. 1999). Amazon pioneered collaborative filtering to suggest products based on customers' own behavior and the behavior of similar users (Linden, Smith, and York 2003). Like oil, the data refining supply chain is a complex set of processes (Iansiti and Lakhani 2020) that, over the past decades, has been experiencing a long sequence of technological innovations. Most recently, that is exemplified by the introduction of generative adversarial networks (Goodfellow et al. 2020) and transformers (Vaswani 2017) in combination with computing technology—such as graphics processing units (GPUs)—that allows for parallelization and scaling up of massive models. Such evolution has led data to generate more and more insights across an increasing number of applications. There is no reason to think today is the end of the journey.

That is where the similarities between oil and data stop. The two differ in several important dimensions that will affect how property rights and other forms of regulation will be defined in the age of artificial intelligence (AI). When it comes to data, the task of regulating property rights is made more challenging than with oil by the fundamental difficulty of pricing information, also known as Arrow's information paradox (Arrow 1962). Data generates information whose value can only be assessed after knowing what kind of information is generated. Ex ante, it is very hard to know whether an AI model would provide information that is actually valuable. For example, even now that large language models have been available for some time, hallucinations (instances when the models produce incorrect or nonsensical output that appear plausible) are a critical area of investigation (Huang et al. 2023).

Further, data comes in at least two forms: data that tends to be static or slowly changing, like a person's date of birth or address; and data that frequently and dynamically changes, like a person's recent online searches or purchases. The first type is more similar to oil than the second type. Static data maintains its relevance to generate insights over time, making its value (and thus its transaction price) easier to determine. Dynamic or behavioral data, on the other hand, can be quite difficult to value because its relevance is often tied to specific contexts or time periods, and it quickly loses significance as behaviors or circumstances change. This distinction highlights a core challenge in defining property rights for data: static data, being more enduring, aligns slightly better with traditional notions of ownership and transferability, similar to tangible assets like oil. Behavioral or dynamic data, however, is ephemeral, context-dependent, and often co-created by multiple entities, such as platforms and users, often in interconnected ways.

The recombinant nature of AI models, especially GenAI, further complicates issues by making it difficult to assess the contribution of individual data sources. GenAI learns simultaneously from multiple fragmented inputs, which are often small and diverse, leading to challenges in determining the value of each independent source. This difficulty is exacerbated by the uncertain substitutabilities and complementarities between these sources, as their combined effects may differ significantly from their individual impacts.

Despite their intrinsic differences, both static and dynamic data defy traditional notions of ownership and transferability because they are inherently non-rivalrous and can be replicated, shared, and reused without diminishing their original value. This makes it difficult to assign exclusive ownership rights in the same way

[4] `https://timesmachine.nytimes.com/timesmachine/1934/07/22/110042820.html?pageNumber=133`.

that we do for physical assets. In the recent context of GenAI, the most relevant form of property right is copyright, which protects creative works such as newspaper articles, images, and music. Judging from the large number of lawsuits against GenAI companies,[5] it is clear that current copyright owners like The New York Times and Getty Images consider the use of their works to train GenAI models to fall outside of the fair use doctrine.[6] (The fact that incumbents will fight to maintain the status quo is very eloquently described in Callander and Li (2025) in this handbook.)

Although an evaluation of whether fair use applies in the context of GenAI is beyond the scope of this commentary, it is important to emphasize the economic effects that such innovations can have on the market for creative works (Waldfogel 2012). On one hand, it may reduce the returns to creative works because some demand will substitute towards the output from GenAI models. As an example, some people may rely on ChatGPT instead of The New York Times to find information about past events. On the other hand, it may also reduce the costs of producing and distributing creative works in the first place (Kretschmer and Peukert 2020). Indeed, The New York Times writers may be able to write an article about current events faster with the help of ChatGPT. Or somebody searching for past events may be directed to The New York Times articles via ChatGPT. Together, the two forces have opposite effects on the incentives to produce creative works given the current stringency of copyrights, making it uncertain whether copyrights should be strengthened or weakened to maintain the current level of creative works (assuming such level is even optimal).[7]

As previously mentioned, news aggregators were the subject of similar speculations a few years ago. Copyright issues with news aggregators like Google News arose from their use of headlines, snippets, and links to display news content without compensating publishers. Publishers argued that these practices infringed on copyright and reduced traffic to their websites, thus undermining their ad revenue and subscriptions. Aggregators claimed fair use, asserting that snippets were factual and their services increased traffic to original sources, especially small publishers (Athey, Mobius, and Pal 2021). Eventually, legislation caught up with the business model of news aggregators. Among the new regulations, the European Union's 2019 copyright directive allowed for unlicensed hyperlinks and short snippets, whereas Australia in 2021 and Canada in 2023 opted for the compensation of publishers by news aggregators through "link taxes."[8]

The fact that data is endogenously produced (unlike oil, whose quantity is fixed) creates an additional challenge. Although the first iteration of AI models often use "naturally occurring" data, like text from classic books or paintings from the Renaissance, that is unlikely to be the case in the future. If AI models are used to create new creative works, a circularity of input-output relationships emerges between the data used for training AI models and the use of those models to generate even more data. Subsequent rounds of model deployments mean that previous versions of AI models are influencing (and changing) the data generating process of data that is the input to future model iterations. This feedback loop raises critical questions (Sühr, Samadi, and Farronato 2024) about the quality and diversity of data used in training subsequent iterations of AI models, as well as additional property rights questions over who owns that data.

There is one final commonality between data and oil worth highlighting, which involves externalities. For oil, it is pollution; for the use of data, it is the risk of erosion of privacy and potential misuse, as in the case of deep fakes. Just as the widespread use of oil has led to environmental degradation and climate change, the extensive collection and use of data may generate significant societal costs. These include the risk of surveillance, data breaches, and the perpetuation of biases through algorithmic decision-making. In both cases, these externalities are often borne by individuals and communities that are distinct from the entities that profit from the resource, creating an imbalance that requires regulatory intervention and innovative solutions to address.

While the analogy between oil and data provides useful insights into their transformative potential, it is essential to recognize their fundamental differences. Unlike oil, which is finite and primarily tied to physical

[5] `https://blogs.gwu.edu/law-eti/ai-litigation-databa`.

[6] The fair use doctrine in copyright law allows for use of copyrighted material without the permission of the copyright owner under a limited set of circumstances. Fair use is established based on four factors: the purpose and character of the use (e.g., transformative or educational), the nature of the copyrighted work (e.g., fictional or non-fictional), the amount of original work used, and the effect on the market for the original work. Common examples of fair use include criticism, commentary, research and teaching, and news reporting. `https://www.copyright.gov/title17/92chap1.html#107`.

[7] This is a delicate balance that affects virtually all intellectual property rights, including copyrights (Waldfogel 2012) and patents (Romer 1990, Scotchmer 2004), and that sometimes is resolved with public investment in innovation (Greenstein 2015).

[8] `https://www.forbes.com/sites/taxnotes/2023/10/02/whats-going-on-with-digital-link-taxes/`.

extraction, data is an endogenously generated and non-rivalrous resource. Its value is heavily contingent on the technologies and contexts in which it is applied. In addition, oil's externalities are predominantly environmental, whereas data's societal risks involve privacy, fake news, surveillance, and algorithmic biases. These distinctions highlight that the challenges and solutions for regulating data and assigning property rights must diverge significantly from those applied to oil.

In reflecting on the rhetoric surrounding data and AI, it is important to question other popular claims that often circulate in industry and policy conversations, such as "AI will democratize access to knowledge" or "AI scales effortlessly." These statements may sound compelling, but they risk ignoring the complexities of how data is collected, governed, and used effectively. Taking a more realistic view that acknowledges both the potential and limitations of data-driven technologies is essential for developing meaningful approaches to regulation, ownership, and societal impact.

References

Arrow, Kenneth (Dec. 1962). "Economic Welfare and the Allocation of Resources for Invention". In: *The Rate and Direction of Inventive Activity: Economic and Social Factors*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 609–626.

Athey, Susan, Markus Mobius, and Jeno Pal (2021). "The impact of aggregators on internet news consumption". In: *NBER Working Paper No. 28746*.

Beraja, Martin and Noam Yuchtman (2025). "Generalized Disruption: Society, Work, and Property Rights in the Age of AI". In: *Handbook*.

Callander, Steven and Hongyi Li (2025). "Regulating an Innovative Industry". In: *Handbook*.

Einav, Liran, Chiara Farronato, and Jonathan Levin (2016). "Peer-to-peer markets". In: *Annual Review of Economics* 8.1, pp. 615–635.

Einav, Liran, Dan Knoepfle, et al. (2014). "Sales taxes and internet commerce". In: *American Economic Review* 104.1, pp. 1–26.

Goodfellow, Ian et al. (2020). "Generative adversarial networks". In: *Communications of the ACM* 63.11, pp. 139–144.

Greenstein, Shane (2015). *How the internet became commercial: Innovation, privatization, and the birth of a new network*. Vol. 16. Princeton University Press.

Huang, Lei et al. (2023). "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions". In: *ACM Transactions on Information Systems*.

Iansiti, Marco and Karim R Lakhani (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Harvard Business Press.

Kretschmer, Tobias and Christian Peukert (2020). "Video killed the radio star? Online music videos and recorded music sales". In: *Information Systems Research* 31.3, pp. 776–800.

Linden, Greg, Brent Smith, and Jeremy York (2003). "Amazon. com recommendations: Item-to-item collaborative filtering". In: *IEEE Internet computing* 7.1, pp. 76–80.

Page, Lawrence et al. (1999). "The PageRank Citation Ranking: Bringing Order to the Web". In: *The Web Conference*.

Romer, Paul M (1990). "Endogenous technological change". In: *Journal of political Economy* 98.5, Part 2, S71–S102.

Scotchmer, Suzanne (2004). *Innovation and incentives*. The MIT Press.

Sühr, Tom, Samira Samadi, and Chiara Farronato (2024). "A Dynamic Model of Performative Human-ML Collaboration: Theory and Empirical Evidence". In: *arXiv preprint arXiv:2405.13753*.

Vaswani, A (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems*.

Waldfogel, Joel (2012). "Copyright research in the digital age: Moving from piracy to the supply of new products". In: *American Economic Review* 102.3, pp. 337–342.

Yergin, Daniel (2011). *The prize: The epic quest for oil, money & power*. Simon and Schuster.