“Data is the new oil,” said Jaron Lanier in a recent op-ed for The New York Times. Lanier’s use of this metaphor is only the latest instance of what has become the dumbest meme in tech policy. As the digital economy becomes more prominent in our lives, it is not unreasonable to seek to understand one of its most important inputs. But this analogy to the physical economy is fundamentally flawed. Worse, introducing regulations premised upon faulty assumptions like this will likely do far more harm than good. Here are seven reasons why “data is the new oil” misses the mark:

1. **Oil is rivalrous; data is non-rivalrous**

If someone uses a barrel of oil, it can’t be consumed again. But, as Alan McQuinn, a senior policy analyst at the Information Technology and Innovation Foundation, noted, “when consumers ‘pay with data’ to access a website, they still have the same amount of data after the transaction as before. As a result, users have an infinite resource available to them to access free online services.” Imposing restrictions on data collection makes this infinite resource finite.
2. Oil is excludable; data is non-excludable

Oil is highly excludable because, as a physical commodity, it can be stored in ways that prevent use by non-authorized parties. However, as my colleagues pointed out in a recent comment to the FTC: “While databases may be proprietary, the underlying data usually is not.” They go on to argue that this can lead to under-investment in data collection:

Companies that have acquired a valuable piece of data will struggle both to prevent their rivals from obtaining the same data as well as to derive competitive advantage from the data. For these reasons, it also means that firms may well be more reluctant to invest in data generation than is socially optimal. In fact, to the extent this is true there is arguably more risk of companies under-investing in data generation than of firms over-investing in order to create data troves with which to monopolize a market. This contrasts with oil, where complete excludability is the norm.

3. Oil is fungible; data is non-fungible

Oil is a commodity, so, by definition, one barrel of oil of a given grade is equivalent to any other barrel of that grade. Data, on the other hand, is heterogeneous. Each person’s data is unique and may consist of a practically unlimited number of different attributes that can be collected into a profile. This means that oil will follow the law of one price, while a dataset’s value will be highly contingent on its particular properties and commercialization potential.

4. Oil has positive marginal costs; data has zero marginal costs

There is a significant expense to producing and distributing an additional barrel of oil (as low as $5.49 per barrel in Saudi Arabia; as high as $21.66 in the U.K.). Data is merely encoded information (bits of 1s and 0s), so gathering, storing, and transferring it is nearly costless (though, to be clear, setting up systems for collecting and processing can be a large fixed cost). Under perfect competition, the market clearing price is equal to the marginal cost of production (hence why data is traded for free services and oil still requires cold, hard cash).

5. Oil is a search good; data is an experience good

Oil is a search good, meaning its value can be assessed prior to purchasing. By contrast, data tends to be an experience good because companies don’t know how much a new dataset is worth until it has been combined with pre-existing datasets and deployed using algorithms (from which value is derived). This is one reason why purpose limitation rules
can have unintended consequences. If firms are unable to predict what data they will need in order to develop new products, then restricting what data they’re allowed to collect is per se anti-innovation.

6. Oil has constant returns to scale; data has rapidly diminishing returns

As an energy input into a mechanical process, oil has relatively constant returns to scale (e.g., when oil is used as the fuel source to power a machine). When data is used as an input for an algorithm, it shows rapidly diminishing returns, as the charts collected in a presentation by Google’s Hal Varian demonstrate. The initial training data is hugely valuable for increasing an algorithm’s accuracy. But as you increase the dataset by a fixed amount each time, the improvements steadily decline (because new data is only helpful in so far as it’s differentiated from the existing dataset).

7. Oil is valuable; data is worthless

The features detailed above — rivalrousness, fungibility, marginal cost, returns to scale — all lead to perhaps the most important distinction between oil and data: The average barrel of oil is valuable (currently $56.49) and the average dataset is worthless (on the open market). As Will Rinehart showed, putting a price on data is a difficult task. But when data brokers and other intermediaries in the digital economy do try to value data, the prices are almost uniformly low. The Financial Times had the most detailed numbers on what personal data is sold for in the market:

- “General information about a person, such as their age, gender and location is worth a mere $0.0005 per person, or $0.50 per 1,000 people.”
- “A person who is shopping for a car, a financial product or a vacation is more valuable to companies eager to pitch those goods. Auto buyers, for instance, are worth about $0.0021 a pop, or $2.11 per 1,000 people.”
- “Knowing that a woman is expecting a baby and is in her second trimester of pregnancy, for instance, sends the price tag for that information about her to $0.11.”
- “For $0.26 per person, buyers can access lists of people with specific health conditions or taking certain prescriptions.”
- “The company estimates that the value of a relatively high Klout score adds up to more than $3 in word-of-mouth marketing value.”
- “[T]he sum total for most individuals often is less than a dollar.”

Data is a specific asset, meaning it has “a significantly higher value within a particular transacting relationship than outside the relationship.” We only think data is so valuable because tech companies are so valuable. In reality, it is the combination of high-skilled labor, large capital expenditures, and cutting-edge technologies (e.g., machine learning) that makes those companies so valuable. Yes, data is an important component of these production functions. But to claim that data is responsible for all the value created by these
businesses, as Lanier does in his NYT op-ed, is farcical (and reminiscent of the labor theory of value).

**Conclusion**

People who analogize data to oil or gold may merely be trying to convey that data is as valuable in the 21st century as those commodities were in the 20th century (though, as argued, a dubious proposition). If the comparison stopped there, it would be relatively harmless. But there is a real risk that policymakers might take the analogy literally and regulate data in the same way they regulate commodities. As this article shows, data has many unique properties that are simply incompatible with 20th-century modes of regulation.

A better — though imperfect — analogy, as author Bernard Marr suggests, would be renewable energy. The sources of renewable energy are all around us — solar, wind, hydroelectric — and there is more available than we could ever use. We just need the right incentives and technology to capture it. The same is true for data. We leave our digital fingerprints everywhere — we just need to dust for them.