NO. 25

DISRUPTIVE PERSPECTIVES

# DATA'S ROLE FOR AI

**AUDUN WICKSTRAND IVERSEN**
Portfolio Manager

**EMILIE KRUTNES ENGEN**
Portfolio Manager

**LEO RUNDGREN OLSEN**
Junior Analyst

## The purpose of Disruptive Perspectives

When we analyze various themes, we spend a great deal of time and use many tools (quarterly reports, analyses, dialogue with the companies, company visits, Excel, calculators and word models). We often make small notes, and sometimes large notes that we think of as perspectives. We are old enough to know that there are rarely truths, often only different perspectives.

Disruptive Perspectives has only one purpose, namely to share our perspectives on themes that shape our future. These are not academic notes, encyclopedia entries or recommendations to do, buy or sell anything. Only good old-fashioned information sharing, to make visible how we view various themes at the time of publication. Perspectives do not become smaller, perhaps rather greater, when they are shared. With that as a starting point, enjoy the journey through our perspectives.

# Contents



## Disclaimer

> The content of this article is not intended as investment advice or recommendations. If you have any questions about the funds referred to, you should contact a financial advisor who knows you and your situation. Remember also that historical returns in funds are never a guarantee of future returns. Future returns will depend, among other things, on market developments, the manager's skill, the fund's risk, as well as the costs of purchase, management and redemption. Returns may also become negative as a result of price losses.

# 1.0 Data as the new oil in the AI era

In a time when AI is transforming developments within self-driving and humanoids, data is no longer merely a resource. It is the very driving force behind innovation and competitive advantage. As Larry Ellison, founder of Oracle, has said: «It's all public data from the internet ... So they're all basically the same. And that's why they're becoming commoditized so quickly». His statements illuminate the distinction between public and private data, and how this affects the value of AI models in the short and long term. To make this more concrete, we present a 2x2 matrix based on Jay Barney's Resource-Based View (RBV). According to this theory, sustainable competitive advantages arise from resources and capabilities that are valuable, rare, difficult to imitate and not easily substitutable. In academia this is known as the VRIN framework (or VRIO). It shows that private data can be a source of sustainable competitive advantage, while public data more often contributes to companies becoming more alike.

**Postcards From The Future**

## Data: Type of data and value

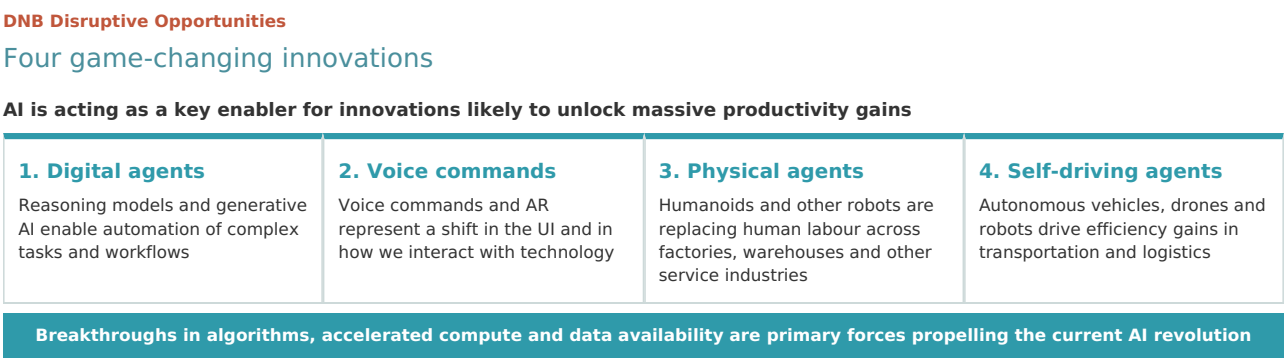
| | **Public data**<br>Shared & Generic | **Private data**<br>Proprietary, domain-specific |
|---|---|---|
| **Short Term**<br>Operational value | **High availability for fast model training (e.g. Wikipedia data for generalist AI); low cost, but limited differentiation** | **Limited use due to privacy, but immediate utility in niche applications (e.g. bank data for fraud detection)** |
| **Long Term**<br>Strategic value | **Leads to commoditization and convergence of models, as Ellison warns** | **Creates VRIN resources for disruption, e.g. via RAG for secure analyses** |

This matrix sets the foundation for our analysis, which builds on discussions of data categorization, Big Data principles and AI applications. We will go on to explore how data is «fed» to neural networks, the risk of «garbage in, garbage out» (GIGO), and an integrated framework for future value.

# 2.0 Theoretical Perspective

In the introduction we established data's role as the new «oil» in the AI era. This chapter deals with the overarching theoretical perspective for understanding the topic. In an exponential world, where technological growth follows Moore's law and doubles capacity every two years, linear business models, which assume that the past equals the future, are disrupted.

Digital and physical agents «eat» and grow on three essential resources. These are silicon (semiconductors), data and energy. We focus on data as the central resource, and integrate theory from Schumpeter, Christensen, Blue Ocean Strategy, Porter, network effects, the flywheel and Kuhn. These theories illuminate how data drives disruption, creates new markets and accelerates exponential growth, while challenging traditional linear approaches. We take the liberty of offering a brief summary.




### Schumpeter's Creative Destruction

Joseph Schumpeter introduced the term creative destruction in 1942 to describe how innovation tears down old structures and paves the way for new ones. For Schumpeter this was the very core of capitalism. In an exponential world this dynamic also applies to data. Public datasets (such as content on the open internet) are gradually weakened by data pollution from AI-generated material, while private and synthetic data form the basis for new value chains.

Linear models, which build on the assumption of stable growth and historical patterns, are challenged by exponential innovations. A good example is RAG technology, which stands for Retrieval-Augmented Generation. Simply explained, it means that the model does not only answer based on what it has learned in advance, but can also retrieve relevant information from a private data store in real time (for example internal documents, customer data or company knowledge). In this way, businesses can use their own data to make the AI far smarter and more precise, without the information having to be shared or trained openly into the model. This makes private data a strategic moat. Data thus becomes a central driving force in what Schumpeter called industrial mutation. AI agents grow on data as fuel, break up traditional industries and lay the foundation for new ecosystems.

### Christensen's Disruptive Innovation

Clayton Christensen defines disruption as a process where simpler, cheaper products start in low-end niches and move upward, challenging established leaders. In the context of data, this concerns the shift from public, generic datasets (incremental innovation) to private and synthetic data (disruptive). Linear models, which rely on historical data analysis, are disrupted by exponential technologies such as edge AI, where agents process data in real time without centralized training. For example, synthetic data starts in niches such as the simulation of less common drone scenarios, before scaling and rendering traditional data collection methods superfluous. This accelerates the demand growth for silicon (GPUs for generation) and energy (compute-intensive processing), and underscores that data is a disruptive force that reshapes markets.

### Blue Ocean Strategy

Blue Ocean Strategy is about creating new markets through differentiation and low cost, rather than competing in red oceans of bloody rivalry. Applied to data, this involves moving from commoditized public

datasets to new markets where private data is combined with synthetic data to create unique value, as in RAG-based systems for enterprise AI. Exponential growth disrupts linear models by turning data into a so-called «value innovation», for example by using multimodal data (video plus sensor) for humanoids, which opens new markets without direct competition. Agents grow here by using data as fuel, while silicon and energy support the scaling, creating sustainable growth without fighting over existing resources.

### Porter's Five Forces and competitive strategy

Porter's five forces analyze industry structure through threats from new entrants, suppliers, buyers, substitutes and rivalry among them, in order to understand competitive intensity and profit potential. In an exponential, data-driven world, the threat from new entrants increases (startups with synthetic data), while supplier power (databases such as Oracle) is strengthened by the private data moat. Linear models underestimate substitutes such as AI-generated data, which disrupts traditional collection. Porter's framework shows how data changes the competitive landscape. Agents that use data reduce rivalry by creating differentiated niches, while semiconductors and energy become critical inputs that affect supplier power and profit.

### Metcalfe's Law and Network Effects

Metcalfe's law states that a network's value is proportional to the number of users of the platform, raised to the second power ($n^2$). The point is exponential network effects. For data, this applies to ecosystems where more users (or agents) generate more data, which accelerates growth. A positive feedback loop. In the AI world, agents grow on data as network value. More sensors in drones create exponentially much better understanding, while semiconductors and energy support the scaling. This explains why data platforms such as Oracle or Tesla grow faster than linear competitors, and underscores data's role in creating «winner-takes-all» markets.

### Jim Collins' Flywheel Effect

The flywheel effect describes how small, consistent advances build momentum over time, like a heavy flywheel that accelerates exponentially. In a data context, it starts with the collection of primary data, which builds on itself to create moats. Like Amazon's flywheel, where more data leads to better AI, which attracts more data. Exponential growth disrupts linear models by turning data into a self-reinforcing cycle.

### Thomas Kuhn's Paradigm Shift

Kuhn described paradigm shifts as revolutions where old scientific frameworks are replaced by new ones when anomalies accumulate, often driven by social factors. Applied to data, the shift from linear, historical analyses to exponential, data-driven AI represents a paradigm shift. Old models collapse under anomalies such as model collapse (that is, that quality is weakened when models are increasingly trained on AI-generated data), while new paradigms (for example RAG) come to dominate.

Taken together, these theories integrate a perspective where data, in an exponential world, disrupts linear models, and positions our four agents as central actors that grow on silicon, data and energy. This chapter sets the scene for the subsequent analyses of data categorization and AI applications, and underscores that success requires embracing disruption rather than resisting it.

# 3.0 Definitions and categorization of data

To understand why data has become one of the most important resources in the AI economy, we must first distinguish between different types of data. Not all data has the same value. Some data is easily accessible to everyone, other data is protected, proprietary and difficult to copy. Some data is tidy and structured, other data lies hidden in text, images, video, audio and sensors. And precisely these differences determine how useful the data is for artificial intelligence.

A fundamental division runs between public and private data. Public data is information that is freely available, often drawn from open sources such as Wikipedia, Reddit, public databases or the open internet. Such data has been crucial for the training of large language models, because it exists in enormous quantities and covers a broad range of topics. At the same time, it has clear limitations. It can be full of errors, noise, biases and repetitions. When many models are trained on the same open data sources, the risk also increases that the models become more alike. Data that everyone has access to rarely provides a lasting competitive advantage (moat).

Private data is different. This is data owned by individuals, companies or institutions, and which is often protected by laws, agreements or internal security requirements. Examples can be medical records, financial transactions, customer data, production data, supply chain data or internal documents. Such data is often more valuable because it is domain-specific, fresher and closer to real decision-making processes. It can give AI systems a deeper understanding of a concrete problem, a specific industry or a particular organization. At the same time, it is harder to use, precisely because it can be sensitive and subject to regulations such as GDPR or HIPAA.

Another important dimension is the difference between structured and unstructured data. Structured data is data organized in a fixed format, for example tables with rows and columns in a database. This can be sales data, inventory status, accounting figures or customer transactions. The advantage of structured data is that it is relatively easy to search, analyze and use in traditional models. It is well suited to precise calculations, forecasts and automated decisions.

Unstructured data is more chaotic, but often far richer. This is data that does not follow a fixed schema, such as text, images, video, audio, emails, PDFs, social media posts and sensor data. For humans, this type of information is often easy to interpret, but for machines it has historically been harder to exploit. This is where modern AI has truly changed the rules of the game. Large language models, computer vision and multimodal models make it possible to extract value from data that was previously too messy, fragmented or complex to analyze effectively. For self-driving cars, drones and humanoid robots, this is absolutely crucial. They must be able to interpret the world through cameras, radar, lidar, microphones and other sensors. In that case, neat spreadsheets are not enough. The models must understand unstructured and multimodal data from the physical world.

Between structured and unstructured data we find semi-structured data. This is data that does not fit fully into traditional tables, but which nonetheless has a certain organization. Examples are JSON files, XML files, log files and many types of IoT data. Semi-structured data has become increasingly important in modern software, because it combines the flexibility of unstructured data with some of the searchability of structured data.

Data can also be categorized by source. Primary data is data that an actor collects itself. It can be camera data from a car, sensor data from a drone, usage data from a digital platform or production data from a factory. The advantage of primary data is that it is often unique, relevant and directly tied to the actor's own products or services. It can thereby become a strategic resource. The disadvantage is that it requires investment in infrastructure, sensors, software and systems for collection and storage.

Secondary data is data that comes from external sources. It can be public datasets, purchased datasets, industry reports or aggregated data sources. Such data is often cheaper and faster to access, but it is rarely as unique. It can be useful as a supplement, but normally provides a weaker competitive advantage than data one has collected oneself.

Finally, data must be assessed by sensitivity and privacy. Some data is non-sensitive, such as weather data, traffic patterns or anonymized statistics. Other data is sensitive, such as health information, financial information, personal data or business-critical documents. The more sensitive the data is, the greater the requirements for security, access control and legal compliance. Many businesses therefore classify data in tiers such as public, internal, restricted and confidential.

This determines who can use the data, how it can be used, how valuable it is, and whether it can create lasting competitive advantages. In AI, the old principle of «garbage in, garbage out» (GIGO) still holds. Poor, biased or irrelevant data produces weak models. Good, relevant and proprietary data, by contrast, can become the very fuel that makes an AI model better, more precise and harder to copy.

As Geoffrey Moore observes in the context of Big Data:

> "*Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway.*"

This underscores how important it is to understand data along several dimensions, among other things how it is structured, where it comes from, and how sensitive it is. In practice, it concerns the difference between structured and unstructured data, primary data and secondary data, as well as non-sensitive and sensitive data. For AI systems such as self-driving cars, drones and robots, this is decisive. Video data, for example, is both unstructured and multimodal, because it combines image, motion, time and context. Such data must be of high quality and credibility for the model to learn correctly. If the input is poor, biased or incomplete, the output also becomes unreliable (GIGO).

In the AI context, these categories are absolutely central because neural networks learn from data. One can say that they «eat» data in order to understand patterns, relationships and anomalies. Public and unstructured data can give the models broad knowledge about the world, but private and more specific data is often what provides precision in concrete use cases. It can be diagnostics in healthcare, fraud detection in finance, predictive maintenance in industry or motion planning in robots. The better the data reflects the problem the model is meant to solve, the more useful the AI system becomes.

Here, data classification also plays an important role. With the help of machine learning, raw data can be tagged, sorted and organized automatically based on content, context and sensitivity. This makes it easier to separate useful signals from noise, protect sensitive information and build more reliable datasets for training and analysis. Raw data in itself is rarely enough. Value arises only when the data is structured, understood and made applicable.

# 4.0 Big Data's four V's

As a natural continuation of the data categorization, Big Data can be understood through four central dimensions. In English they are often referred to as Volume, Velocity, Variety and Veracity. In Norwegian we can translate them as volum, hastighet, mangfold and pålitelighet.

This framework provides a useful way of assessing data, not only by how much data one has, but also by how quickly the data arises, how varied it is, and how much one can trust it. In an AI context this is absolutely decisive. Neural networks and language models do not become better merely because they are fed more data. They become better when the data is relevant, varied, up to date and reliable.

> "*Big data is at the foundation of all the megatrends that are happening*" — *Chris Lynch*

### 1. Volume

This concerns the large quantities of data created, stored and processed every day. It is important because neural networks often need large datasets in order to learn patterns and become more precise.

One example is the healthcare sector, where genomic datasets can be used to identify genetic markers for disease and to develop more tailored treatment. Another example is self-driving, which collects large quantities of sensor and camera data to improve object recognition and driving behaviour.

The challenge is that high data volume requires storage, computing power and good structure. Without this, the data can create more noise than value.

### 2. Velocity

This concerns how quickly data is created, collected and analyzed. In many AI systems this must happen in real time or near real time.

This is decisive in use cases that require an immediate response. In healthcare, AI can analyze patient data continuously to detect signs of acute situations. For drones and self-driving cars, sensor data must be processed within milliseconds to avoid collisions. Digital agents also use continuous data streams from users to adapt answers and actions instantly.

High velocity often requires technology such as edge computing, where data is processed closer to the source (in the end products). This reduces latency (the delay), but also makes data handling more complex.

### 3. Variety

This concerns the fact that data comes in many different forms, such as structured data in tables, semi-structured data such as JSON files, and unstructured data such as text, images, audio and video.

In healthcare, for example, electronic records, medical images and data from smartwatches can be combined to provide more holistic diagnoses. For drones and robots, video data, sensor data and text-based instructions can be linked together to understand situations better.

A good example of this is the difference between camera and lidar in self-driving cars, drones and robots. Cameras collect visual data, roughly the way humans see the world, with colours, signs, road markings, objects and movements. This provides rich data for AI models that need to understand context, but cameras can be vulnerable to poor light, rain, snow and glare. Lidar works differently. It emits laser pulses and measures the distance to objects, thereby creating a precise three-dimensional map of the surroundings. Lidar provides very good depth understanding, but less semantic information than a camera.

This is the core of multimodal AI, where models learn from several types of data at once. The challenge is that different data formats must be interpreted, cleaned and linked together in a way that actually makes sense.

### 4. Veracity

This concerns how accurate, consistent and credible the data is. In AI this is decisive, because models learn from the data they receive. If the data is wrong, biased or full of noise, the model may also produce weak or misleading results.

One example is AI in healthcare, where medical data must be validated carefully before it is used for diagnostics. Another is robots and cobots, where sensor data must be cleaned and checked to avoid faulty actions in the physical world.

This becomes even more important with synthetic data. Such data can be useful, but if it does not reflect reality well enough, it can weaken the model over time. High veracity therefore requires good data governance, quality assurance and continuous verification.

Finally, the framework is often expanded with a fifth V, namely value. This concerns turning data into insight, decisions and concrete improvements.

In AI we see this clearly in personalized recommendations, such as those of Netflix and Amazon. By analyzing user data, they can suggest films, series or products that feel more relevant to each individual user. The same applies in finance, healthcare and industry, where good data can provide better risk models, more precise diagnoses or more efficient operations.

Without value, the other V's are of little use. Large, fast, varied and reliable data means little if it cannot be used for something. It is only when data improves products, decisions or business models that AI creates real value.

With the four V's (plus Value) as a foundation, we move toward the differences in training basis and model convergence, where theoretical frameworks such as Disruptive Innovation help us understand data's transformative potential.

### Examples of synthetic data

Synthetic data is artificially generated data that imitates real data. It is often used in AI when real data is hard to obtain, too sensitive to share, or too biased to produce good models. Such data can be created through statistical models, simulations or generative models.

In healthcare, synthetic patient records can be used for research and AI training without exposing real patient data. They can simulate symptoms, treatments and outcomes, and thereby make it possible to develop better diagnostic models within the bounds of privacy regulations such as GDPR and HIPAA. Synthetic health data can also be used to fill gaps in datasets, for example for very rare diseases or for groups that are poorly represented in existing data.

In finance, synthetic data is used to simulate transactions, customer behaviour and market conditions. This makes it possible to test new solutions, train models for fraud detection and analyze risk without using sensitive customer data. Banks can thereby build and test AI systems in safe environments before they are put into use in the real world.

For self-driving cars, drones and robots, synthetic data is especially important. Instead of waiting for rare situations to occur in reality, one can simulate them. These can be dangerous traffic events, poor weather, darkness, unexpected obstacles or complex urban environments. In this way, AI systems can be trained on far more scenarios than would be practical or safe to collect physically.

Synthetic data is also used within media and image processing. Artificially generated images and videos can help computer vision models recognize objects, people, backgrounds or movements. In the same way, synthetic text and audio can be used to train language models, voice assistants and systems for translation or customer service.

The point is that synthetic data can make AI development faster, cheaper and more privacy-friendly. At the same time it requires high quality. If the synthetic data does not resemble reality closely enough, the model can learn the wrong patterns. In that case synthetic data goes from being a strength to becoming a source of noise. The balance between scale, realism and reliability therefore becomes absolutely decisive.

# 5.0 Differences in training basis and model convergence

After having looked at different data categories and Big Data's four V's, the next question is what kind of data AI models are actually trained on. This is where the distinction between public and private data becomes decisive. Public data can produce broad and useful models, but because many actors use the same data sources, the models can also begin to resemble one another more closely. Private data works differently. It is harder to obtain, more domain-specific and often far more valuable. It can therefore become a strategic moat.

The training basis is the dataset used to develop and improve AI models. Large language models such as ChatGPT, Gemini and Llama are to a large extent built on enormous quantities of publicly available data from the internet, such as websites, Wikipedia, books, code, forums and open datasets. This gives the models a broad understanding of language and the world, but it also makes them generalists. They can write, summarize, reason and answer many types of questions, but they often lack deep insight into internal processes, specific customers, medical patient histories, industrial sensor data or a company's unique decision-making basis.

When several models are trained on much of the same open internet material, a form of model convergence arises. The models learn many of the same patterns, acquire similar strengths and weaknesses, and over time produce results that can feel quite alike. Over time this can make general AI models more like commodities or infrastructure. They become useful, powerful and cheap, but harder to differentiate. A little like cloud storage or payment infrastructure. Important for everyone, but not necessarily unique in itself.

This does not mean that public data is unimportant. On the contrary, it has been the very foundation for the first wave of modern AI. But public data primarily provides breadth. Private data provides depth. In healthcare it can be patient records, image diagnostics and treatment history. In finance it can be transactions, customer behaviour and risk data. In industry it can be production data, maintenance logs and sensor data from machines. Such data is often more precise, more relevant and more closely tied to concrete decisions.

This is the reason why private data can create divergence between AI systems. Two companies can use the same base model, but obtain entirely different results if one of them has access to better, more relevant and more structured data. In practice, the competitive advantage then moves from the model itself to the data around the model. The model becomes the engine, but the data determines how well it runs in a concrete terrain.

Here RAG technology becomes important. Retrieval-Augmented Generation makes it possible to connect an AI model to private documents, databases or knowledge sources without the model necessarily having to be trained directly on them. A bank can use AI to analyze internal lending data, compliance documents and customer history. A hospital can use AI to retrieve relevant information from patient records and medical guidelines. An industrial company can let a model retrieve insight from maintenance logs, sensor data and manuals. In this way, general models can become domain-specific without all private information having to be exposed or built into the model itself.

Synthetic data reinforces this picture further. It can be used to fill gaps where real data is hard, expensive or risky to collect. For self-driving cars and drones, one can simulate rare events such as accidents, extreme weather or unexpected obstacles. For robots, one can train in virtual environments before the systems are tested physically. Synthetic data can thereby function as a bridge between public and private data, but only if the quality is high. If the synthetic data does not reflect reality well enough, the model can learn the wrong patterns.

Disruption is not necessarily about the most advanced technology from the start. It often begins in niches, with solutions that are simpler, cheaper or more tailored to a concrete need (Christensen's Disruptive Innovation). Public data and general AI models drive much of the incremental improvement at the large technology companies. Private data, RAG and domain-specific AI systems, by contrast, can open the way for more disruptive solutions in industries where precision, safety and context are more important than general language understanding.

For AI, this can mean that value is gradually moved from having the largest base model to having the best data foundation, the best data governance and the strongest connection between model and real use. For digital agents, autonomous systems, robots and health technology, this becomes decisive. Whoever owns the most relevant data can build systems that learn faster, act more precisely and become harder to copy.

Thus, model convergence and data divergence become two sides of the same development. On the surface, the base models become more alike. Beneath the surface, the difference between the actors grows larger, because private, proprietary and high-quality data determines who actually manages to create value. This is where data goes from being a raw material to becoming a strategic resource.

# 6.0 The Data Value Pyramid (DVP)

As a result of our discussions, we introduce the Data Value Pyramid. This is a hierarchical framework that categorizes data from the base (source, structure) to the top (use). At the base we find companies such as Oracle (ORCL) for structured private data, while the middle level includes Nvidia (NVDA) for synthetic multimodal data within robotics. At the top, value is assessed. Tesla (TSLA) provides a moat via primary video data for humanoids. RBV supports this, as Barney notes:

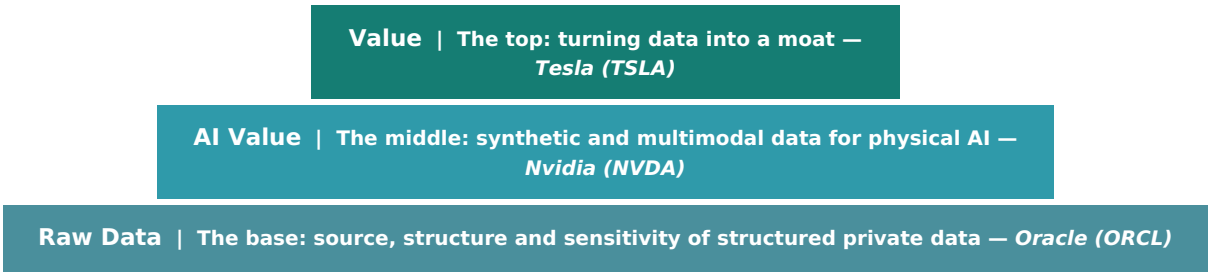
> "*To attract the kinds of resources that can generate profits, managers must recognize that stakeholders... have claims on the profits.*"

The future points toward AI factories where synthetic data dominates, as Ellison predicts:

> "*The future lies in leveraging private enterprise data.*"

Whether this is correct we do not know, but it is a powerful perspective. And that is precisely what we are looking for in our perspective notes.

The Data Value Pyramid (DVP)




At the bottom of the pyramid we find the fundamental categories. Here data is assessed by source, structure and sensitivity. Is the data primary or secondary? Is it structured, semi-structured or unstructured? Is it public, internal or sensitive? A company such as Oracle is relevant in this part of the pyramid, because much of its value lies in the storage, organization and use of structured, private enterprise data. This is not the most spectacular part of AI, but it is often the foundation on which everything else is built.

In the middle of the pyramid we find the AI-specific extension. Here the questions become more advanced. Is the data real or synthetic? Is it text-based, visual, sensor-based or multimodal? How high is the quality, and how well does the data reflect reality? This level becomes especially important for robotics, drones, self-driving cars and humanoids. Nvidia is a good example of an actor that fits in here, because the company does not only deliver computing power, but also builds ecosystems for simulation, synthetic data and physical AI. When robots can be trained in virtual environments before they are tested in reality, the entire learning curve changes.

At the top of the pyramid it is no longer only about what kind of data one has, but about what the data can actually be used for. Does it create better models? Lower costs? Faster product development? Stronger customer insight? Or a moat that competitors cannot easily copy? Tesla is a good example of this part of the pyramid. If the company's enormous quantities of video and sensor data from cars can be used to train self-driving, humanoids and physical AI, the data can become a long-term strategic resource. The value lies not only in the cars, but in the learning loop between products in use, new data, better models and better products.

This can also be understood through the Resource-Based View. In RBV, the most valuable resources are those that are rare, difficult to imitate and well organized within the company. Proprietary data of high quality can fulfill precisely these criteria. Public data can be used by everyone. Private, fresh and domain-specific data is far harder to copy.

Today, data is used primarily to train, fine-tune and improve AI models. Self-driving cars use camera, radar and lidar data for object recognition and motion planning. Robots use motion data and sensor data to learn

interaction with the physical world. Digital agents use text, documents and user history to understand context and act more precisely. The value lies in better decisions, more automation and higher operational efficiency.

Going forward, the value can become even greater. Toward 2030, the development points toward AI factories, where data is not only used to train models once, but enters into continuous learning loops. Fresh data from real interactions can become more important than old datasets. Synthetic data can fill gaps in rare scenarios, such as accidents, extreme weather or dangerous robot situations. Edge AI can make it possible to process data closer to the source, with lower latency and better privacy.

At the same time, this is not without risk. If synthetic data is used uncritically, the models can learn the wrong patterns. If private data is handled poorly, it can create privacy problems and regulatory risk. If the data is incomplete or old, it can produce models that appear intelligent, but make poor decisions.

The DVP therefore shows that it is not enough to have a lot of data. One must have the right data, reliable data, fresh data and the ability to transform data into better AI systems. The winners of the future will not necessarily be those with the largest models, but those with the best learning loops between data, models and real products. This is where data goes from being a digital raw material to becoming a strategic resource.

**Postcards From The Future**

### Data: Real vs. Synthetic Data Matrix (GIGO and applications)

| | **Real Data**<br>Authentic, high veracity | **Synthetic Data**<br>Generated, flexible |
|---|---|---|
| **High Value**<br>Long Term, Strategic | **Proprietary video from humanoids: builds a moat, but with privacy risk** | **GAN-generated data, e.g. for drones, fills gaps, disrupts the market (80% of AI data in 2028?)** |
| **Low Value**<br>Short Term, Operational | **Public text data: commodity, convergence** | **Low-quality synthetic data: risk of model collapse in agents** |

**Postcards From The Future**

### Data: Modality vs. Time Horizon Matrix (extension with video)

| | **Unimodal**<br>Text & Numbers | **Multimodal**<br>Video + Other |
|---|---|---|
| **Present** | **Basic training for digital agents: effective, but limited** | **Video for self-driving cars: rich, but processing-intensive. Edge AI processing is the enabler** |
| **Future**<br>Long term | **Commodity in AI factories** | **Dominant for humanoids & cobots: real-time video integration via edge AI, high market growth** |

This framework provides an analytical lens for maximizing the value of data, while it addresses risks. For implementation, prioritize synthetic data for testing and video for vision-based systems. The future belongs to those who cultivate «good» data.

## 6.1 Relevant companies and stocks

With the DVP as a framework, we can place different companies and stocks according to what role they play in the AI ecosystem. The list is not meant to be complete, but gives examples of actors that are relevant for data collection, data storage, data quality, AI training and practical applications within self-driving cars, drones, humanoids, cobots and digital agents.

The point is not only who is «doing AI», but where in the value chain they create value. Some companies control fundamental data infrastructure. Others collect proprietary data from products in use. Some make the

data applicable through storage, structuring and governance. Others use the data to train models, simulate scenarios or build new AI products. The DVP framework helps us distinguish between these roles.

### The Base

At the bottom of the pyramid we find companies that handle data at the foundational level. This concerns source, structure, sensitivity and processing. Here lies much of the foundation for Big Data's V's, in particular volume, velocity, variety and veracity.

Tesla (TSLA) is an example of an actor with large quantities of primary data. The company collects video and sensor data from cars in use, which can be used for the training of self-driving and, in time, physical AI such as Optimus. The value lies in the data being proprietary, new and drawn from real situations.

Figure AI (not listed) is a humanoid company and is dependent on large quantities of primary data from physical interaction with the world. When the robots move, grip objects, cooperate with humans and carry out tasks in factories or warehouses, valuable sensor, video and motion data is collected. Such data can be used to improve the robots' motor skills, decisions and ability to handle new situations. The value lies in the data being proprietary, practical and drawn from real working environments, which can give Figure an important training basis for physical AI over time.

Alphabet and Waymo (GOOGL) also have valuable primary data from self-driving cars. Waymo's fleets collect data from cameras, lidar and sensors in complex traffic environments. This provides a strong training basis for autonomous systems.

Snowflake (SNOW) and Amazon (AMZN) are more relevant on the secondary and aggregation side. Snowflake helps businesses collect, organize and use data across cloud environments. Amazon uses enormous quantities of e-commerce, logistics and customer interaction data, while AWS is a central platform for companies building AI solutions.

Oracle (ORCL) sits close to structured enterprise data through databases and business-critical systems. This is perhaps less visible than the large AI models, but very important for RAG-based solutions, where private enterprise data is connected to AI models.

Salesforce (CRM) also fits in here, because the company sits on structured CRM data tied to customers, sales and workflow. Such data can become valuable when used for agentic AI in sales, service and customer dialogue.

For unstructured data, Meta (META) is a good example. The company has enormous quantities of text, images, video and social interaction, which is relevant for multimodal models. Seagate (STX) can be placed in the same layer from a different angle, as a supplier of storage capacity for the enormous data quantities AI requires.

Within sensitive data and data security, SentinelOne (S) and Rubrik (RBRK) are relevant examples. SentinelOne uses AI on cybersecurity data to detect threats, while Rubrik is exposed to backup, data protection and governance. In a world where AI is connected more closely to private data, this type of security and control becomes increasingly important.

Palantir (PLTR) fits particularly well into the category of data processing and refinement. The company works to collect, clean, structure and make complex datasets applicable, especially in defence, government, industry and healthcare. This is a good example of how value often lies in making raw data decision-ready.

### The Middle

In the middle of the pyramid we find companies that are more directly tied to AI training, synthetic data, multimodality and physical AI. Here it is not only about storing or structuring data, but about making it useful for neural networks.

Nvidia (NVDA) is the clearest actor at this level. The company does not only deliver GPUs to data centres, but also builds ecosystems for simulation, robotics and synthetic data. Through platforms such as Omniverse, Isaac and Cosmos, Nvidia can help create virtual training environments for robots, drones and autonomous systems. This makes the company central in the transition from pure data processing to physical AI.

Ambarella (AMBA) is relevant within multimodal processing, especially for camera and video-based AI systems. The company's semiconductors are used in solutions where visual information must be processed efficiently, such as in drones, security cameras, robots and autonomous vehicles.

Ouster (OUST) is an example of a company that delivers sensor technology and physical data to autonomous systems. Its lidar and sensor data can be used for mapping, object recognition and navigation. This is especially important in physical AI applications where the model must understand the world in three dimensions.

Apple (AAPL) fits into multimodality through the combination of text, speech, image, video, sensors and user data in its ecosystem. The company has a unique position because it sits close to the end user and can integrate AI directly into devices.

ServiceNow (NOW) represents a more enterprise-oriented form of AI data. The company uses workflow data and process data to build digital agents that can automate tasks internally within businesses. Here the value is not necessarily multimodal video, but contextual understanding of how organizations actually work.

Symbotic (SYM) combines sensor data, motion data and operational data to optimize logistics and robot control.

### The Top

At the top of the pyramid we assess which companies can transform data into lasting competitive advantages. Here the question becomes whether the data is rare, difficult to copy and tightly integrated into the company's products and business model.

Tesla (TSLA) is one of the clearest examples of a possible long-term data moat. If the company manages to use video and sensor data from the car fleet to improve self-driving and physical AI, it can create a strong flywheel. More products in use provide more data, more data provides better models, better models provide better products, and better products can in turn provide more users.

Nvidia (NVDA) has a different type of moat. Here the value lies in the infrastructure around AI training, simulation and accelerated data processing. If the AI of the future is increasingly trained in virtual environments and AI factories, Nvidia can become a central platform for both real and synthetic data.

TSMC (TSM) and Arm (ARM) lie further down the physical value chain, but are nonetheless strategically important. TSMC produces many of the most advanced chips that the AI ecosystem depends on. Arm delivers architecture that is important for energy-efficient data processing, especially on edge devices, mobile devices and, in time, robotics.

Microsoft (MSFT) and Amazon (AMZN) have strong positions through cloud, enterprise customers and AI platforms. They control not only infrastructure, but also distribution channels into businesses. This gives them an important role in how private data is connected to models and agents.

Palantir (PLTR) can also be placed high in the pyramid, because the company focuses on making complex, sensitive and often fragmented data operational. In industries such as defence, healthcare and industry, this can provide significant value if the AI systems become part of critical decision-making processes.

### Implications

The framework shows that AI value is not created only in the model itself. It is created in the entire chain around the model. From data collection and storage to structuring, quality assurance, simulation, training and practical application.

Tesla illustrates the value of primary and proprietary real-world data. Nvidia illustrates the value of synthetic data, simulation and AI infrastructure. Oracle, Snowflake and Salesforce illustrate the value of structured enterprise data. Palantir and Rubrik show the importance of data governance, security and applicability. Ambarella and Ouster show how important sensor and video data becomes for physical AI.

The most important point is that the AI winners of the future will not necessarily be only those with the largest models. It can just as well be those who own the best data, have the strongest learning loops and manage to

turn data into concrete products, better decisions and lasting competitive advantages.

# 7.0 The Future

This note has navigated through the evolution of data in AI, from fundamental categories to strategic frameworks, and underscored that good data is essential for robust models in applications such as drones and cobots. As Christensen summarizes:

> "*First, disruptive products are simpler and cheaper. They generally promise lower margins, not greater profits.*"

We look toward a future where synthetic and private data disrupt traditional paradigms. To visualize this, we present a concluding 2x2 matrix that integrates Disruptive Innovation Theory with data types, and shows the potential for innovation while we warn against risks such as bias and commoditization.

**Postcards From The Future**

## Data: Data Type vs. Disruptive Potential

| | **Real Data**<br>Authentic, High Veracity | **Synthetic Data**<br>Generated, Flexible |
|---|---|---|
| **Low Disruptive Potential**<br>Sustaining Innovation | **Public real data: improves existing models, but leads to convergence (e.g. general AI agents)** | **Low-quality synthetic data: fills temporary gaps, but risks model collapse** |
| **High Disruptive Potential**<br>New-Market Disruption | **Private real data: creates moats in niches (e.g. video for humanoids, cars, AI glasses and drones)** | **Disrupts by simulating rare scenarios (e.g. drone training), potentially 80% of future AI data** |

To conclude, the value of data lies in its ability to develop ethical and effective AI. By investing in quality data and theoretical frameworks, we can navigate toward a future where AI is not only smart, but also sustainable.

Thank you for your time.

**We'll see you in the future.**