ABU DHABI GLOBAL MARKET ACADEMY

2023

ADGM Academy Research Centre

# A BRIEF ON **DATA AND AI**


AUTHORED BY

**SANJAY SARMA**
ASB and MIT

**SHIEN JIN ONG**
ASB

**SATISH VISWANATHAN**
Accenture


ASIA School of Business
in collaboration with MIT Sloan Management

# WELCOME NOTE

AI and data are crucial in virtually every aspect of our lives, both professional and personal.

The combination of AI and data help drive automation, improve efficiency and decision making, enable predictive analytics, personalisation and enhancement of user experiences…the list is endless. Ultimately, AI and data help facilitate innovation, which is the focus of the ADGM Academy Research Centre.

We are proud to publish this paper by Sanjay, Shien and Satish as they share their insights on the extraordinary amount of data available and continuously generated, and how AI can be utilised to improve the way people's lives, businesses, governments and national economies can evolve.

ADGM Academy
Research Centre

# INTRODUCTION

Facts, figures, and statistics are all what we know as data. Data has existed since record keeping began, and the history of data analysis goes back to when states – ancient China, Rome, and medieval Europe – began systematic collection of facts and figures or what we now term as statistics. Interestingly, the word “statistics” itself originates from these early efforts to systematize data for state governance.

Advances in computation and storage meant that data could–and did–become more copious over the course of the last century. Today, “big data” refers to the extraordinary amount of data generated by an unimaginable range of sources: from businesses, whether banking or farming; from customers, whether in retail or in travel; from processes, whether manufacturing or supply chain; from governments, whether productivity increases or poverty reduction; from industrial sectors, whether electricity output or construction activity; from social media, whether resumes or discussion threads; from automation systems, whether the pressure in a gas pipeline or the video captures from smart doorbells; from education, whether national testing scores or individual performance in an online course; from planetary issues, whether precipitation across the globe or the measurement of the earth’s magnetosphere; and much, much more. Bigdata is widely recognized to have great value.

Ask an executive about the value of this data and you are likely to hear platitudes. Dig deeper, and ask the how, the why, and the action plan for extracting and using this data, and you will see confusion about where to start. This is entirely understandable. Often, terms such as big data become catch-all banalities not because there is no substance in them, but because there is too much substance. The challenge is to slice and dice this substance to give managers and executives a better taxonomy of the field, and to make it actionable. And as of late, enter artificial intelligence (AI). Individually, the data and AI are bound to make a massive difference to the way people, businesses, governments, and indeed, national economies, operate and evolve. But they are inextricably tied, and together their impact will be profound. Here we briefly describe the engineering of data in the era of AI and point to the future.

## Data can be categorized in many ways.

First, data can be classified based on whether it is structured or unstructured. In other words, is it in a systematically tabulated and labelled form or is it disorganized? The former is called, unsurprisingly, structured data and the latter is called unstructured data. A spreadsheet is an example of structured data, and a scan of a pile of documents is an example of unstructured data. There is a spectrum between these extremes, though, and most data is semi-structured. A dump of files from a laptop is semi-structured. Many files may be individually structured – pictures, spreadsheets, text files, say – but there is no overall structure.

## There is data about data too and it is called metadata.

An image from a smart-phone camera, for example, has data from the pixels (compressed in JPEG but raw in TIFF), with metadata indicating the format, the time, the size of the image, and perhaps the GPS location of the image. Often all the data is bundled into a single file.

On smartphones, the format of choice is Exif – the Exchangeable image file format. A picture is an example of structured data, but a collection of scans of documents can also have metadata – a record of where the content came from, when it was scanned, and the resolution of the scanner, for example.

A second view is the scope of the data. Drawing inspiration from economics, we split data into two forms: macro-data and micro-data based on the perspective from which data is looked at. Macro-data will refer to the aggregate perspective. For example, the passenger traffic from Abu Dhabi to London every day – perhaps to analyse travel before, during and after the pandemic. Micro-data will refer to disaggregated perspective. For example, a single person's flight patterns, and all their touch points with an airline. Of course, data need not be about a person. Micro-data could be specific to a particular aircraft, and macro-data may apply to an entire fleet.

When individuals are involved, privacy becomes an issue. Microdata can be used to help customers – consider an airline frequent flyer account – but is burdened from a privacy perspective. Aggregation can be helpful from a systemic perspective but aggregation into macro- data does not eliminate privacy concerns. Precisely how the data was collected, what representation was made to the individual, permissions granted, national laws, human rights, and of course, ethics, are all key considerations that become more and more important as the world becomes more data centric. There is a large and valuable body of academic literature, best-practice guidelines, and regulations on this topic – ranging from differential privacy to the EU's General Data Protection Guidelines (GDPR).


**A third approach to categorization relates to the types of data.**

One usually thinks about numeric data or coded data – for example, a spreadsheet from a receiving dock with dates, times, weights and perhaps a label to say undamaged or damaged. With the growth of telemetry, recording equipment, sensors, and particularly IoT, a large range of new data types is becoming available: images, audio recordings, videos, GPS traces, text inputs, comments, time series – of temperatures, chemical compositions, numbers, vehicle speeds – and so on. These fall somewhere on the spectrum between structured and unstructured, and the list will likely grow. In fact, the greater the variety of data, the more the body behaves as if it were semi-structured because the analysing party must parse through all the options.

# THE ASCENT OF DATA

Data is valuable; this is beyond argument today. The growth of the telegraph was driven partly by the value of data such as stock and commodity prices[1]. According to research firm Markets and Markets, the global big data market was thought to be worth $162.6 billion in 2021 and could become as large as $273.4 billion by 2026. The CAGR is expected to be 11.0% from 2021 to 2026[2]. But precisely how to extract this value requires more illumination.

1. Lamoreaux, N. R. (1986). Information costs and the organization of the grain trade. The Journal of Economic History, 46(2), 423-436.
2. https://www.marketsandmarkets.com/Market-Reports/big-data-market-1068.html

# RAW DATA:

In its basic undigested form, there is “raw data.” This could simply be a list of a passenger’s flights on a particular airline or the numbers of passengers between two airports. This information would presumably be valuable to an acquiring airline and would be part of the value proposition in the purchase. But raw data is seldom valuable by itself. The real value is the insights it yields. Raw data is more valuable when it is interpreted. It is the ore while the value is in the gold that is extracted from it.

Raw data has had a second coming of sorts in the era of Large Language Models (LLM’s – think GPT), which we will talk about later in this article. Platforms such as X (nee Twitter) and Reddit, which are replete with user conversations that are continuously being refreshed, have found themselves the target of learning bots seeking fresh training data. This is widely considered to be the primary reason why X and Reddit have throttled use[3, 4].

3. https://innotechtoday.com/reading-limits-imposed-at-twitter-to-stop-ai-data-scraping/
4. https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html

# DATA INSIGHTS:

Data insights are profound and can drive behaviours in companies at tactical, operational, or strategic levels. Examples abound, and we give a few here. Consider tactical actions: traffic data can indicate congestion and lead to a new route for a delivery company on a minute-by-minute basis. Another example is the turning on of generators by a power company – so-called spinning reserves – to absorb demand spikes based on expected weather and the varying levels of renewable power available. A third example is surge-pricing by ride-share companies based on high demand (for example, after a sporting event or during inclement weather).

Operational use of data is very common too. A striking example occurs with the Tesla Powerwall, a home battery system for electric energy storage. When a weather event such as a storm is expected, in which power might go out, Tesla increases the charge level on the Powerwall battery pack to provide energy to the home during the outage. It is no surprise that IBM purchased the Weather Company in 2016 – weather is a key variable which affects much of what we do. Airlines, hotels, and travel and online services providers decide operational parameters such as capacity and prices based on insights gleaned from data.

Strategic use of insights is rarer but perhaps most dramatic. One example is Netflix, which creates entire shows based on the data it collects. The hit show "House of Cards" is a prime example of the role of big data in entertainment.[5] Using data, the satellite TV company DirectTV realized customers are most likely to switch when they move. Armed with this strategic insight, DirectTV created a new operational plan. Data from the US Postal Service informs them when customers move homes, and DirectTV markets to these recent movers.

5. https://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html

But extracting data insights requires a significant amount of bespoke effort. First, the infrastructure for wrangling massive amounts of data (so-called big data) is sophisticated and requires advanced skills. It may involve clean-up, distributed storage and computing on large clusters. The cloud may be involved because it enables almost unbounded expandability. Sometimes the data is used to confirm a hypothesis that came from a person. Or the insight might be discovered through a combination of statistical analysis, some forms of AI, advanced clustering algorithms, and visualizations. The process is complicated and has led to a field of study that is only about 15 years old: data science.

Sophisticated intermediaries who have the data are in pole position. In the travel and hospitality industries, it may be the aggregation sites such as Google, Expedia, Kayak, Booking.com, TripAdvisor, or Agoda. Or it may be back-end systems such as Amadeus or SABRE. In entertainment, it could be aggregators such as Netflix, Amazon, or Apple (though these intermediaries are now becoming producers – a strategic shift drive, as the earlier example shows, by the very data they have). Amazon, particularly, has strong insights into consumer behaviour in retail. Many such companies use recommendation systems to further improve their sales outcomes.[6] Original producers – media companies, airlines and others cannot sit back. Data-driven disintermediation is changing the way value is delivered, and they must adapt or lose relevance.

# SUPRADATA


AI is now ubiquitous – whether it is in voice assistants such as Siri or face recognition system such as at airport immigration gates. Generative AI systems have been much in the news recently because "transformer" based systems such as large language models (LLM's or Generative pre-trained transformer) are an entirely new frontier. LLM's were seen merely as glorified predictive typing systems, Generative pre-trained transformer) are an entirely new frontier.

6. Hardesty, L. (2019). The history of Amazon's recommendation algorithm. Amazon Science, 22.

LLM's were seen merely as glorified predictive typing systems, except extended to entire paragraphs and even documents, but something strange happened: they turned out to capture knowledge beyond just semantics and syntax. GPT has surprised and alarmed in equal measures with its unexpected abilities: solving exam problems, writing code, writing entire articles, and even generating and executing new workflows (suggest an evening plan including a dinner recipe, placing a delivery order with a grocery store for ingredients, arranging rides for guests, etc.).

These capabilities were so surprising that LLM's received a whole new name in 2021: foundation models.[7] Theory of the mind thinkers, philosophers and cognitive scientists are still puzzling over the emergent capabilities of GPT. Is this the emergence of artificial general intelligence? Does this emergent power of language tell us something about how the mind works?[8] Related Generative AI systems such as DALL-E, which generate images from natural language, have been equally startling.

At its core, trained AI systems are simply a very large set of parameters which correspond loosely to the "the strengths of the synapses" in a particular neural network-based architecture. The architecture itself is often widely known. It is the parameters that are the new gold. In the case of GPT-3, a large corpus of text was analysed to generate a model with 175 GB of trainable parameters.[9] Microsoft's GitHub Copilot is based on the data from an enormous code repository now owned by Microsoft called GitHub. Copilot is trained on "159 gigabytes of Python code sourced from 54 million public GitHub repositories" according to Wikipedia.[10] We call this collection of parameters supradata – a term we coin here to distinguish it from the training data. In some sense supradata can be thought of as the gist, or the essence, of the much larger body, and is the ultimate value of the data.

The value of these parameters is great, but hard to comprehend. And who owns it? Facebook parent, Meta, developed a model called LLaMA (Large Language Model Meta AI) and open sourced it to the world. What they held back was the parameters it had generated with great effort by digesting massive amounts of data.[11] Startlingly, those parameters were leaked to the world through a message board called 4chan a week later.[12] Besides the loss to Meta, this leak immediately created worries that LLaMA could be weaponized by criminals to do ultraprecise spear phishing using their own installations of LLaMA. The supradata genie had escaped the bottle, potentially reanimating LLaMA in the wild.

7. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
8. Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083.
9. Brown et al. (2020). Language models are few-shot learners. Advances in neural information processing systems 33, 1877-1901.
10. The original reference for the Wikipedia article is here: https://www.infoq.com/news/2021/08/openai-codex/
11. The prefix "supra" comes from the Latin word "suprā," which means "above" or "over." Since "metadata" and "hyperdata" are taken, there really is no choice!
12. https://www.reddit.com/r/ChatGPT/comments/11mracj/metas_llama_llm_has_leaked_run_uncensored_ai_on/

# COPYRIGHT AND OWNERSHIP ISSUES

The supradata in CoPilot was generated from public GitHub repositories, and the former CEO of GitHub claimed the training set is "fair use".[13] The fair use doctrine is a legal principle in United States copyright law that permits limited use of otherwise copyrighted material without first seeking the permission of the rights holder. However, this interpretation has been disputed by copyright owners and there is now a class action lawsuit against GitHub.[14] The lawsuit is essentially a dispute about the monetization of supradata.

Imagine if you mined paintings from your favourite artist – say Hockney – and then generated new Hockney-like art. Does David Hockney have any rights over this material? Copyright laws will need to be rewritten in the coming years to ascribe value to supradata. The music industry has already grappled with this issue for decades. The hit song "Blurred Lines" was determined by the courts to have infringed on Marvin Gaye's work, and the artists who created "blurred Lines" were ordered to pay $5M to Gaye's family.[15] Recently the musician Ed Sheeran dodged a similar fate.

One could make the same argument against GPT- 3 and GPT-4. Could my writing be contributing to the value generated by GPT users? If so, does the supradata not have my intellectual data in it, and therefore must I not be compensated? If AI could be used to compose music in Marvin Gaye's style, does the Gaye family not have some rights to the proceeds?

13. https://news.ycombinator.com/item?id=27678354
14. https://www.infoworld.com/article/3679748/github-faces-lawsuit-over-copilot-coding-tool.html
15. https://www.nbcnews.com/pop-culture/music/robin-thicke-pharrell-williams-pay-5-million-marvin-gaye-estate n947666

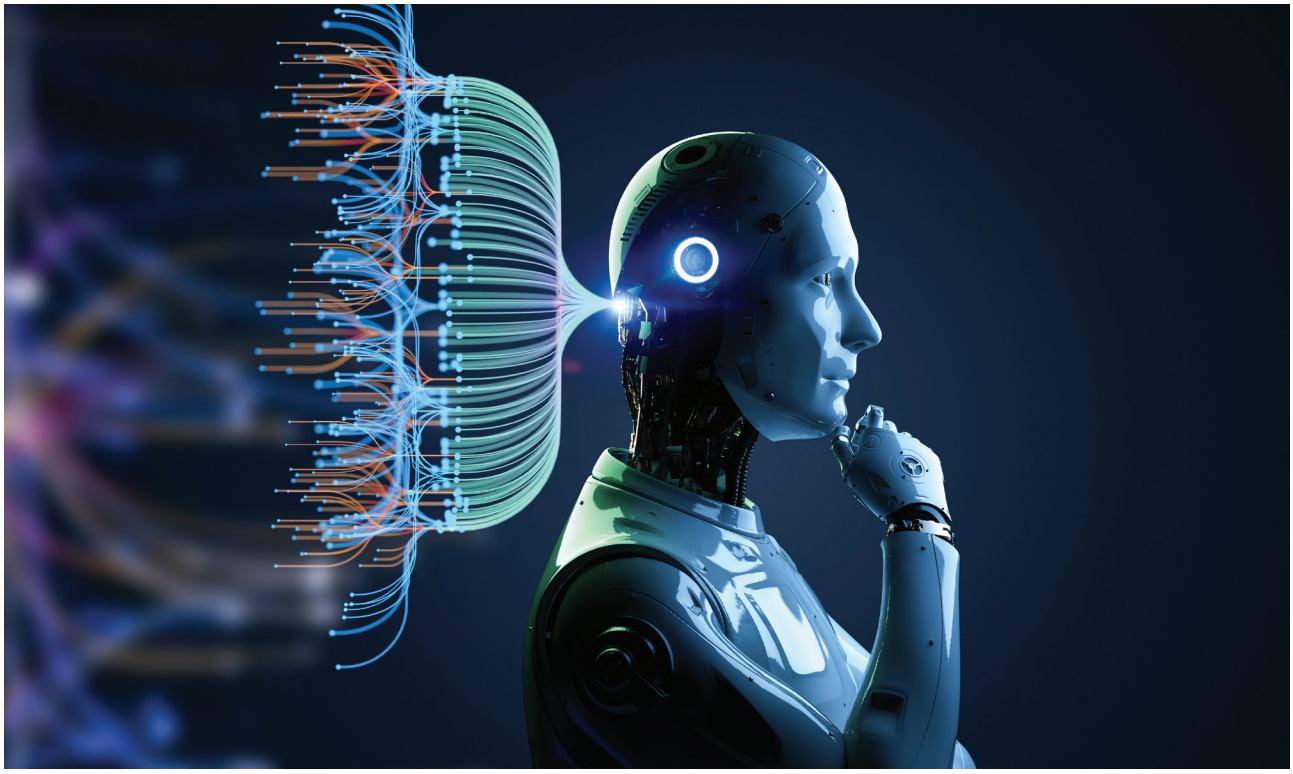
# OPERATIONAL MONETIZATION OF SUPRADATA


LLM's have an additional feature called fine-tuning.[16] It is possible to take a model pre-trained on a large, generic data set and then fine-tune it with a smaller, task-specific data set. Fine-tuning can be used to create knowledge bases which can be queried using natural language. This could lead to (and it is happening) a rebirth of the field of knowledge management. For example, fine-tuning could be used to create a chatbot for customer service by mining previous manual chats. Or it can be used to create a self-help system for employees using internal documentation on such topics as employee benefits.

Public institutions such as courts could expose their content in an easily accessible form to external parties using this approach. Perhaps someday public authorities will permit citizens to mine information freely under the right to information act in pursuit of better governance. Operational benefits such as these have significant financial and societal benefits.

16. https://en.wikipedia.org/wiki/Fine-tuning_(machine_learning)

# A NOTE ON THE SOCIETAL IMPLICATIONS OF DATA AND AI



It would be unforgivable not to mention the societal and ethical implications of data science and AI in any article on the two topics. It would also be arrogant to claim to do the topics justice in a brief note. But the implications are vast. Simply put, humans have explicit and implicit biases. One hopes that people can be educated to rise above them – in other words, to show humanity. But AI runs the risk of implementing them soullessly. Whether it is a resume sorting system[17] or a face recognition system[18], examples abound, and much has been written about the biases in AI systems. Garbage in, garbage out, the expression goes. In AI, bias in, bias out. Biased supradata is a massive threat.

The implications of generative AI on jobs constitute another topic that deserves mention. Traditionally, labour economists have argued that new technologies generally reshape rather than eliminate jobs. The classic example is automatic teller machines (ATM's), which were expected to eliminate bank teller jobs. What really happened was the tellers began to do more cognitively demanding jobs that were also more human-centric – selling mortgages, for example.[19] The fly in this ointment, however, is the pace of change. Can employees be retrained fast enough to respond to the breath-taking pace at which technologies such as generative AI have flooded the thinking of companies? This is a matter for another article, but it calls for a new approach to education – something the Asia School of Business is deeply involved in.[20]

17. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
18. https://www.bbc.com/news/technology-52978191
19. David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: an empirical exploration. National Bureau of Economic Research, 2001.
20. Sarma, Sanjay, and Luke Yoquinto. Grasp: The Science Transforming How We Learn. Anchor, 2021.

# ENTERING A NEW WORLD

Data – and its sibling, AI – are becoming central to the future of businesses and organizations. Learning to master data and AI, to operationalize them, and to monetize them, is a core competency that every business must master.

It will not be easy. As Niccolò Machiavelli says in The Prince, “There is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things.” But to quote Canadian ice hockey great Wayne Gretzky on the other hand, “You miss 100 percent of the shots you don’t take.” Data and AI are in your future – for you to have a future!

## ABOUT ADGM ACADEMY

ADGM Academy is part of Abu Dhabi Global Market (ADGM), an International Financial Centre (IFC) located in the capital city of the United Arab Emirates. The Academy has been established with the vision of becoming one of the leading academies in the region, providing world-class financial research and training services.

Delivering world-class financial education and literacy, ADGM Academy will help to position Abu Dhabi as a leading global financial centre. This will be achieved through globally recognised educational and experiential programmes on a range of topics and qualifications in banking, finance, leadership, entrepreneurship, technical and soft skills.

## ABOUT RESEARCH CENTRE

The ADGM Academy Research Centre brings together an ecosystem of academics, financial industry practitioners, government and technology experts to unlock the shared potential to improve the financial environment in MENA and beyond.

The financial industry continues to transform at a rapid pace with new technologies, disruptors, threats and opportunities appearing all the time. Independent research is crucial to be able to understand and utilise this transformation for the benefit of your business, your customers and society in general.

The Research Centre provides that understanding through insights developed in collaboration with the academic community.

*STAY UP TO DATE WITH ADGM ACADEMY RESEARCH CENTRE.*

adgmacademy.com research@adgm.com

**FOLLOW US**
ON OUR SOCIAL NETWORKS

ADGM Academy
Research Centre

# A BRIEF ON DATA AND AI


ADGM Academy Abu Dhabi Global Market
Level 20, Al Maqam Tower
ADGM Square, Al Maryah Island
PO Box 111999 – Abu Dhabi, UAE

**T** : +971 2 333 8500