FUJITSU

AI and Synthetic Data

Is Synthetic Data the Enabler for Wider AI Adoption?

The prospective benefits and opportunities presented by artificial intelligence, or AI, are unequivocal, with the potential to impact almost every corner of society. Through the adoption of various AI technologies, we all stand to gain from quicker, more informed, accurate decision-making leading to more effective allocation of resources that will improve our everyday lives.

Yet despite the huge advances in AI algorithmic innovation, wider adoption and deployment is still somewhat constrained by the data limitations caused by today's incredibly data-hungry AI models. So, could the use of synthetic data overcome this challenge and facilitate the wider adoption of AI systems?

The ingredients for effective AI adoption

For AI to become widely adopted into our daily working practises it will rely on the three pillars of algorithms (models), computing (processing power) and data. Although all three pillars are required to facilitate successful AI adoption, the most pressing challenge facing organisations in today's landscape remains data, especially data collection, annotation and cataloguing.

To deliver insights, AI algorithms need to be trained on massive datasets and validated on even larger ones. Data enables AI algorithms to perform better, learn faster, and become more robust. So, it's imperative that organisations seeking to successfully adopt AI into their decision-making processes today must address the issue of data, and in particular these four key criteria:

Data quality:

The performance of any AI system is completely reliant on the quality and integrity of the data that's fed into it. If poor quality data is ingested, then data becomes the primary cause of project failure. Assessing the quality of data, and taking action to improve it, should be the first step of any AI project. This would typically include checking data for consistency, accuracy, completeness, duplicity, missing values, data corruption and compatibility.

Data bias:

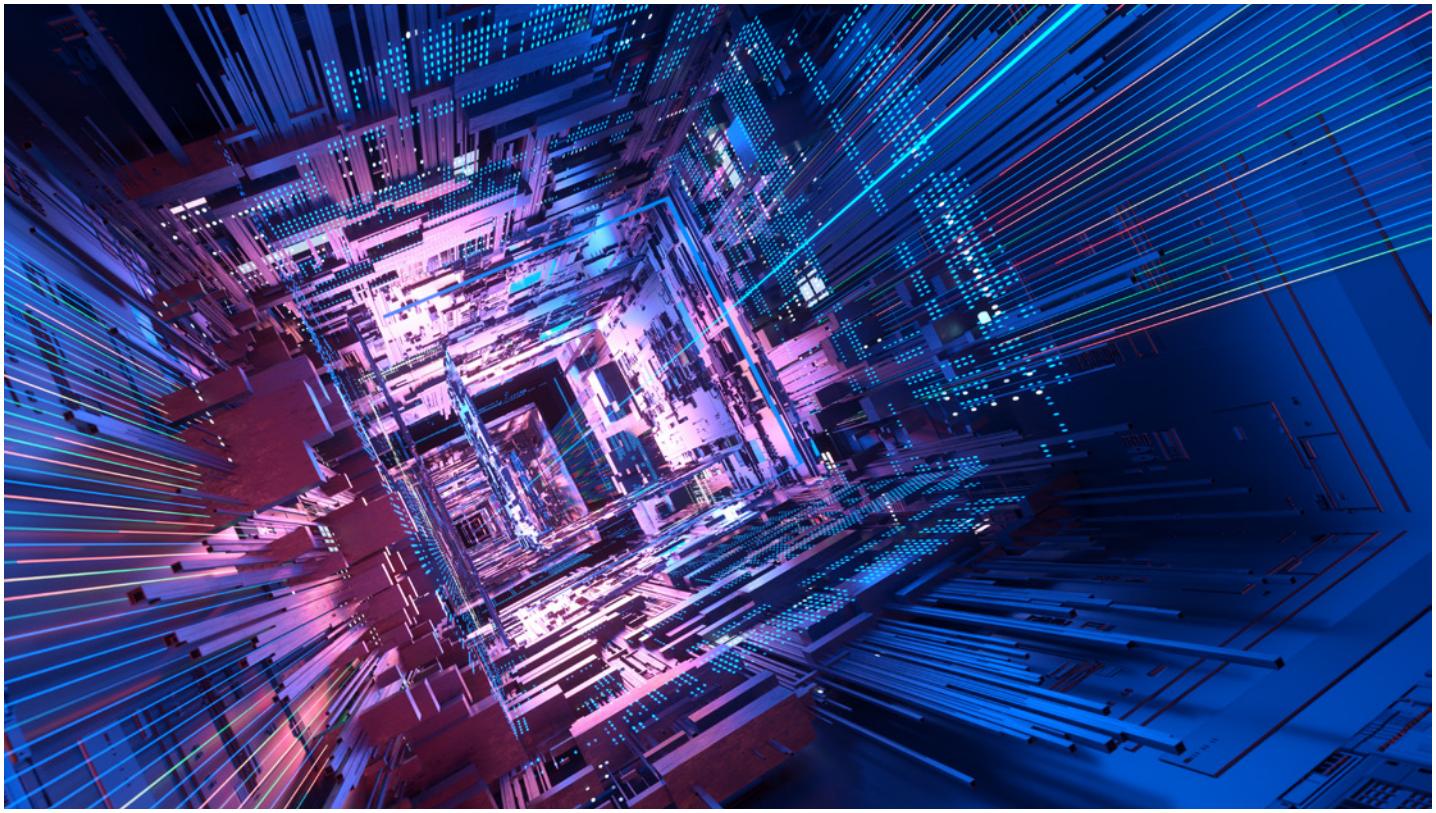
AI systems can only make decisions based on the available data. Biases can occur as a result of the way data is collected (e.g. a marketing survey). In such cases, we can't always say that the dataset is a true representation of the entire population we are trying to model[2].

Data labelling:

AI systems can always be trained and optimised to reach a given level of accuracy with time. However, most AI systems do not generalise well to a given real-world scenario if they haven't been trained with enough real-world examples. This naturally lends itself to the burden of getting enough generic labelled data to prevent unintentional consequences[1].

Data quantity:

AI models are incredibly data hungry and rely on huge volumes to establish accurate outputs. The quantity of data required can be assessed by leveraging a common data infrastructure with shared standards across the organisation to not only capture the right data, but also to create the necessary management of metadata (catalogued data) and provenance. The benefit of a consolidated repository allows for improve data visibility, suitability and any constraints in answering specific questions.



1. www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/
2. fortune.com/2019/10/08/why-did-google-offer-black-people-5-to-harvest-their-faces-eye-on-a-i/

Tackling the AI data challenge

So, it is clear that organisations seeking to deploy AI effectively will need to have access to large volumes of relevant, clean, well-organised data that can be trusted. Tackling this AI data challenge is a much bigger issue for some organisations than it is for others. Large tech firms like Google, Apple, and Amazon for instance have an almost limitless supply of diverse data streams, acquired through their products and services, creating the perfect ecosystem for data scientists to train their algorithms.

For small-medium sized organisations - including public sector departments - acquiring data at scale is a much greater challenge. Their data is often proprietary; it's restricted for use due to contractual agreements; they lack common data standards for sharing; and the data is time-consuming for people to manually prepare, making it expensive. The end result is that data becomes a barrier to innovation and wider AI adoption.

Aside from the big tech firms with their endless access to, and supply of data, organisations have looked to approach the AI data challenge in three main ways, but without much clear, measurable success to date[3]:

- **Crowdsourcing or building an in-house data team:** This approach is costly in terms of management time and other resources and can take a number of years to implement. In addition, significant expertise is required to select and apply the appropriate data management tools to maximise reliability and ensure proper data governance procedures are followed. But if done correctly, it has the potential to provide the necessary scalability demanded by effective AI.
- **Outsourcing to third parties:** This can also be extremely expensive and naturally creates difficulties in terms of compliance and security of data for many industries, such as Defence and National Security.
- **Robotic Process Automation:** Where data can be acquired from open public sources, such as Wikipedia or social media sites, robotic process automation (RPA) offers a definable, repeatable and rules-based approach to scrape and crawl this data. However, using RPA becomes limited in scope if certain data polices are in play, restricting its use. This quickly becomes an intensive management task when scaled up, in order to ensure that correct policies are being adhered to, which naturally introduces additional data governance challenges and overheads.

So could synthetic data be the answer?

The reality is that the cost of quality data acquisition is high, and this is acting as a barrier preventing many from considering AI deployment. To tackle this challenge, organisations are increasingly looking towards synthetic data to address the data shortfall that is preventing AI adoption.

But what is synthetic data?

In its purest form, synthetic data is generated programmatically by mimicking real-world phenomena. For example, realistic image examples of objects in arbitrary scenes can be created using video game engines or audio examples being generated by speech synthesis engines from known text[4]. Currently, synthetic data has started to make an impact in clinical and scientific trials to avoid privacy issues related to healthcare data[5]. Likewise, within software development it can be used for agile development and DevOps in order to speed up testing of software while improving quality assurance cycles[6].

While synthetic data generation has been around since the 1990s, renewed interest is now emerging with the massive advances in computing power, coupled with lower storage costs and the advent of new algorithms such as Generative Adversarial Networks (GANs). The data generated can also be anonymised and created based on user-specified parameters so that it's as close as possible to the properties experienced from real-world scenarios.

In this way, the main advantage of using synthetic data becomes scalability and flexibility, allowing AI developers to programmatically generate as much data as they need to train algorithms and improve model performance and accuracy. Synthetically generated data can also assist organisations and researchers to build reliable data repositories needed to train and even pre-train AI models. Similar to how a scientist might use synthetic material to complete experiments at low risk, organisations can now leverage synthetic data to minimise time, cost and risk.

A real-world example of such usage can be found in the autonomous driving community which is an early adopter of synthetic data repositories. Google's Waymo self-driving AI car is said to complete over three million miles of simulated driving every day, with synthetic data enabling Waymo's engineers to test any improvements within a safe, synthetic simulation environment before being tested in the real-world[7].

3. dzone.com/articles/the-challenges-of-ai-adoption
4. www1.fle.fujitsu.com/private/innovation-focus-creating-a-smarter-anomaly-detection-product-with-ai-stream-processing-and-advanced-analytics/
5. mdclone.com/
6. www.softwaretestingnews.co.uk/when-to-use-production-vs-synthetic-data-for-software-testing/
7. waymo.com/

Synthetic data applications

In addition to autonomous driving, the use cases and applications of synthetic data generation are many and varied from rare weather events, equipment malfunctions, vehicle accidents or rare disease symptoms[8]. In the modelling of rare situations, synthetic data maybe the only way to ensure that your AI system is trained for every possible eventuality. Apart from rare event cases other possible indications for including synthetic data in AI development projects might be found if:

- **Analysis reveals an imbalanced data set:** One common issue that happens when you have imbalance in your training data is overfitting and poor classification performance. Applications such as fraud and anomaly detection suffer from this characteristic, creating unreliable outcomes when confronted with real-world usage. Synthetically generated datasets can provide a reliable and cost-effective way to correct these issues and guarantee a well-balanced dataset.
- **Labelled data is scarce:** With synthetic data, organisations can rapidly develop large-scale labelled data sets in line with requirements for testing purposes. Furthermore, this data can then be modified and improved through iterative testing to provide organisations with the highest likelihood for success in subsequent data collection operations.
- **Safety net for testing algorithms is needed:** A good example is visual based AI systems that need to understand the world around them but are prone to bias. AI developers for visual systems can use synthetic data to produce proof-of-concepts to justify the time, validity and expense of AI initiatives. They can also demonstrate that a specific combination of algorithms can, in principle, be modified to achieve the desired results (for ethical reasons), providing assurance that costs related to a full development cycle will not be wasted.

Common methods for generating synthetic data

Interestingly, methods for synthetic data generation can be found in the field of machine learning itself. Three general strategies for building synthetic data include:

- **Over-sampling:** Various methods that aim to increase the number of instances from the underrepresented class in the data set, for instance, when one data feature dominates the other. Over-sampling helps to create a more balanced representation of data by taking the difference between the sample under consideration and its nearest neighbour, before adding it to the feature vector.
- **Drawing samples from a statistical distribution:** By observing real-world data samples and generating a statistical distribution to describe it allows 'example' data to be drawn from it, to create a randomised synthetic data set. This technique works well when we know there is 'certainty' about the relationship between selected data features and the response variable, or relationships among predictors. These relationships can be extended by combining multiple distribution types to create variable synthetic data sets, to mimic variation in real-world observations.
- **Agent-based modelling:** Agent-based models attempt to capture the behaviour and interaction of individual entities to recreate complex dynamics within a synthetic environment. These interactions are test cases for different scenarios and can be captured as a set of data points to be used for later historical analysis.

Synthetic data enables faster product development

Synthetic data technology has the potential to create new product categories and open new markets rather than merely optimise existing business lines[9]. Simulated data will also enable faster, more agile product development as well as help to level the playing field between the big tech companies and smaller firms that don't have access to the same kinds of real-world data, by enabling:

- Organisations to take their data warehouses and create synthetic versions of them without breaching the privacy of their users and compliancy agreements, allowing AI developers to continue ongoing work without involving sensitive data;
- Organisations and researchers to build data repositories needed to train and even pre-train machine learning models for the future;
- Data to be created on-demand, based on specifications rather than needing to wait and collect data once it has occurred in reality;
- Real-world data to be complemented with synthetic data to allow testing for possible scenarios even if there isn't a good example to draw from the real-world data set. This allows organisations to accelerate performance testing and training of new AI systems.

8. www.datagen.tech/
9. gss.civilservice.gov.uk/blog/synthetic-data-innovation-for-public-good/

Synthetic data is not always the perfect solution

Despite its obvious advantages and benefits, we need to consider that synthetic data is still a replica of specific properties of a real data set. A model looks for trends to replicate, so some of the random behaviours might be potentially missed. Therefore, synthetic data is not always the perfect solution. Indeed, synthetic data is not always suited for all machine learning applications, especially in sensitive areas such as cyber security and more so if datasets are also considered too complex to 'synthesise' correctly, without proper ethics and governance[10].

Essentially, the right to privacy must be respected and individuals should have the ability to opt-out and control the usage of their data. Furthermore, using synthetic data can also lead to misunderstandings during the development phase about how AI models will perform with intended data once in the real world. Although significant progress is being made in this field, one challenge that persists is guaranteeing the accuracy of synthetic data. We must ensure that the statistical properties of synthetic data are matched accurately with the properties of the original dataset, and this very much remains an active research topic.

Synthetic data applications in Defence & National Security

Taking all of this into consideration, synthetic data has widespread potential applicability within the Defence & National Security sector as an enabler to:

- **Providing timely data:** Within fast-paced environments where the key is to be one step ahead of the competition effective data management and agility are imperative. This becomes complicated if data to serve operational needs is siloed and difficult to acquire, creating a 'knowledge gap' between what we have and what we know. If the gap becomes too wide this could lead to potential shortcomings in decision-making capability. The potential to use synthetic data to fill this gap is an opportunity as simulated data is algorithmically created, offering flexibility to create as much of the data you need to train algorithms and maintain appropriate readiness levels in support of operations.
- **Lowering training costs for operational planning:** The production of synthetic data can be utilised to create synthetic environments for simulation purposes. The Simple Synthetic Environment training demonstrator (SSE-TD) initiative is a good example[11]. Furthermore, by incorporating AI 'agents' within a simulated environment provides an ideal sandbox for training and learning-based support. A simulated mock-up of an operational scenario would allow AI 'agents' to self-explore and to learn new tasks/policies, which could be used to provide explanations in decision-making for scenario planning within a safe test environment. This potentially lowers the cost or need to physically perform multiple real-world operational training scenarios until an optimal decision point is found, improving agility for training-based outcomes.
- **Simulating edge cases:** Situational understanding provided by intelligence, surveillance, and reconnaissance (ISR) underpins all military operations, allowing informed decisions to be made and effects to be understood. This understanding can become limited in depth if certain events are difficult to capture in real-life or in some cases may not be worth capturing at all. In this scenario, synthetic data can be produced to solve problems for events that rarely occur to prevent AI algorithms 'behaving' in a biased and non-performant way.

10. blog.global.fujitsu.com/fgb/2020-01-21/the-rise-of-ai-and-the-risks-of-relinquishing-responsibility/
11. www.scmagazineuk.com/new-uk-strategic-command-drive-integration-multi--domain-effect/article/1667949

Positioning statement

Fujitsu is acutely aware that the use of any form of synthetic data for AI transformation activities will depend on the sensitive nature of project requirements. Fujitsu maintains that the data requirements necessary for AI training should be as random as possible, and that any synthetic data applications should be used to verify and test possible outcomes and not to confirm what is already known. Fujitsu is engaged with industry, academia and regulators as they continue to investigate and develop good practise measures and guidelines to ensure correct use of synthetic data in AI solutions across a wide range of industry applications.

What next?

To keep the conversation going, please share your experiences of implementing AI within your organisation. What lessons have you learned? What will you do differently next time?

Please share your insights with us online at:

twitter.com/search?q=fujitsu_Defence&src=typd

www.linkedin.com/showcase/fujitsu-in-defence-and-national-security/

www.facebook.com/fujitsuuk

To find out more about Fujitsu's approach please visit: www.fujitsu.com/uk/solutions/industry/defence-national-security/offerings-and-capabilities

About the Author

Dr. Darminder Ghataoura has over 15 years' experience in the design and development of AI systems and services across the UK Public and Defence sectors as well as UK and international commercial businesses. Darminder currently leads Fujitsu's offerings and capabilities in AI and Data Science within the Defence and National Security space, acting as Technical Design Authority with responsibility for shaping proposals and development of integrated AI solutions. He also manages the strategic technical AI relationships with partners and UK government.

Darminder holds an Engineering Doctorate (EngD) in Autonomous Military Sensor Networks for Surveillance Applications, from University College London (UCL).

Why Fujitsu?

For over 50 years we have innovated with the MOD, Government Departments and intelligence communities, co-creating new technologies and capabilities. As a result, Fujitsu has around 4,000 security cleared staff and the experience to deliver and manage both generic industry offerings and those tailored to specialist needs at OFFICIAL, SECRET and ABOVE SECRET classifications.

Enabling *Your* Information Advantage

In today's complex, digital operational environment, never before has information been such a key asset in securing operational advantage. Fujitsu's vision is to provide customers with the means to translate complex data into useful information upon which to base critical decisions and actions. Transforming this ever-increasing pool of data into meaningful, useful information through analytics, automation and genuine Artificial Intelligence is critical to achieving this goal.

Fujitsu is fully committed to working closely with our customers, and through the use of co-creation will seek to enhance capability both through the acceleration of existing processes, and also through the delivery of truly new capabilities and ways of working. Our approach is based upon maximising both existing investment and best-in-class innovation, delivering the full spectrum of capabilities needed to enable your information advantage.

Contact
Telephone: +44 (0)870 242 7998
Email: askfujitsu@uk.fujitsu.com
Ref: 3988
uk.fujitsu.com