Kevin Jenkins

Synthetic Data and Public Policy

supporting real-world policymakers with algorithmically generated data

Abstract

Good policy is best developed by drawing on a wide array of high-quality evidence. The rapid growth of data science and the emergence of big datasets has materially advanced the supply and use of quantitative evidence. However, some key constraints remain, including that available datasets are still not big enough for some analytical purposes. There are also privacy and data security risks. Synthetic data is an emerging area of data science that can potentially support policy decision making through enabling research to work faster and with fewer errors while also ensuring privacy and security.

Keywords synthetic data, data science, public policy, privacy, AI

'Synthetic data' – data that is algorithmically generated to approximate the real world – can potentially improve and expand the research and evidence necessary for sound public policy. It can be valuable when real data is limited or when privacy concerns limit access to real datasets.

Kevin Jenkins is a founder of professional services firm MartinJenkins.

This article explains what synthetic data is and the key benefits it offers, and briefly summarises the methods and tools used to generate it (called 'synthesis'). The article discusses the rapid development and expanding use of synthetic data for different purposes, and considers the relevance of this new technology for public policy by looking at some public sector use cases, including in Aotearoa.

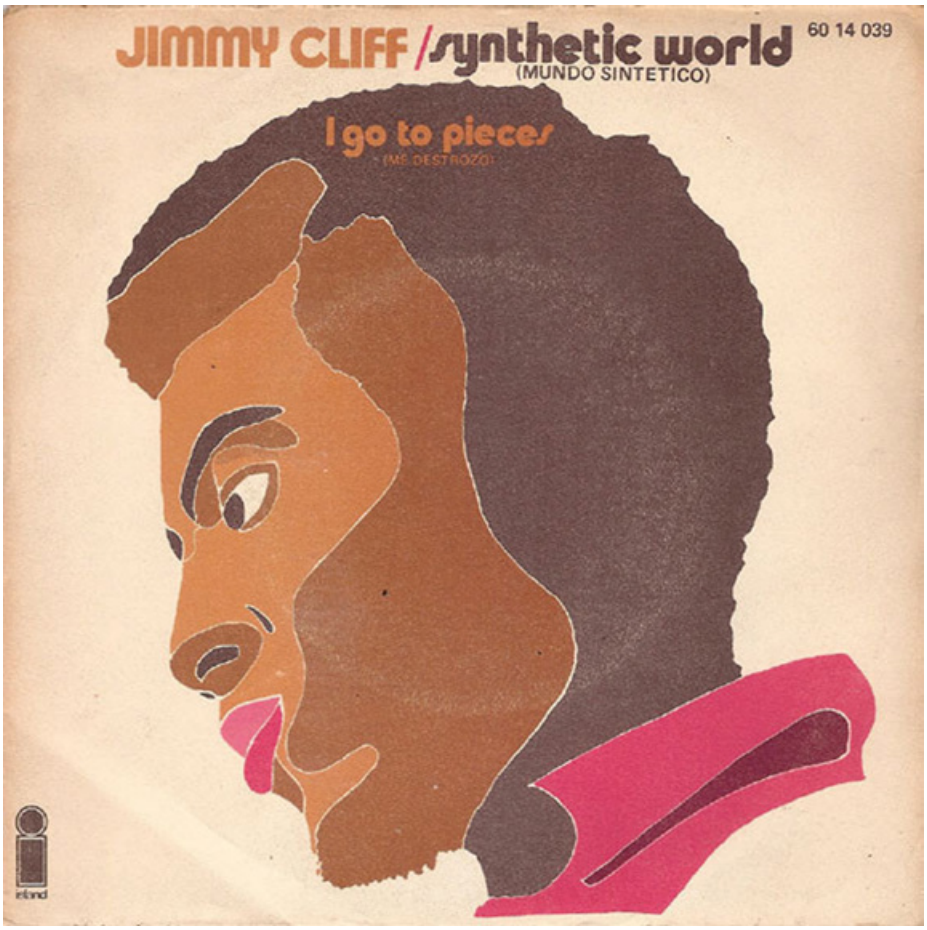
Finally, the article looks at some challenges and risks, and discusses how they could be addressed and how public sector use of synthetic data could be facilitated.

Being synthetic

'Synthetic world' is a great Jimmy Cliff song from the early 1970s, apparently centred on fake, two-faced friends (and drug abuse):

> So you see, my patience is growin' thin
> With this synthetic world we're livin' in.

'Synthetic data', on the other hand, is something altogether more faithful to the



real world than the fakery that Cliff was complaining of, and so is something well worth getting excited about. It can lead to positive changes for people in the real world through ensuring that policy decisions are better informed and can be made more quickly, and with less error, while also ensuring privacy and security. However, it is not yet a mature technology, and it faces methodological, ethical and philosophical challenges, with obstacles to acceptance and uptake.

Artificial intelligence (AI) is developing exponentially and bringing lots of opportunities for improved services, but also lots of regulatory challenges. Turbocharging that development through adding synthetic data may also turbocharge those regulatory challenges. But AI is only one domain where synthetic data is likely to upend traditional approaches to data-driven insights, meaning yet more regulatory challenges.

What exactly is 'synthetic' data?

Synthetic data is data that has been generated from real data and that has some or all of the same statistical properties as the real-world dataset it stands in for (MIT Laboratory for Information and Decision Systems, 2020). Data scientists refer to the process of generating synthetic data as 'synthesis'. The basic idea is simple: you use a model to capture the relationships in the real-world dataset, and then you use the model to generate synthetic data that preserves those relationships (Emam, Mosquera and Hoptroff, 2020).

Unlike 'dummy data', which is randomly generated fake data used to test systems before they go live with real data, synthetic data is generated to preserve the statistical relationships and patterns of the original real-world dataset. An analyst working with a synthetic dataset should therefore get results similar to what they would get with real data. As Paul Calcraft and colleagues explain, a synthetic dataset is:

> generated at random but made to follow the structure and some of the patterns of the original data set. Each piece of information in the [synthetic] data set is meant to be plausible (e.g., an athlete's height will usually be between 1.5 and 2.2 meters, and would never be 1 kilometer), but it is chosen randomly from the range of possible values, not by pointing to any original individual in the data set.

Data that is generated in this way reveals very little, if anything, about any individual in the original data set, but still represents the data well as a whole. (Calcraft et al., 2021)

As Calcraft et al. make clear, the 'randomness' of the selection from within the possible value range is only partial and relative – this depends on the extent to which the synthetic dataset preserves the relationships and patterns in the original dataset. This article will come back to this point in its discussion of 'lo-fi' and 'hi-fi' synthetic data – the higher the degree of fidelity, the more relationships are preserved, and the less random is the selection process.

Synthetic data can play a role even when our understanding of the underlying relationships is more tenuous. For example, synthetic data can be generated when real data is unavailable but we have a theory about the relationship between variables. There can also be a hybrid, where we have *some* historical data and we make some basic assumptions about the distributions and correlations within that data.

Synthetic data is a fast-growing, critical technology

An early use of synthetic data was in 1993 with a synthetic version of the United States census, which allowed the Census Bureau to release samples without disclosing the microdata (Kaloskampis, 2019). Since then, technological advances have led synthetic data to become enormously more sophisticated.

Synthetic data isn't widely talked about outside data science circles, but that's probably about to change. AI commentator Rob Toews believes this new technology is approaching 'a critical inflection point in terms of real-world impact. It is poised to upend the entire value chain and technology stack for artificial intelligence, with immense economic implications' (Toews, 2022). The tech research and consulting firm Gartner predicts that over the next ten years synthetic data will start to massively overshadow real data in AI models (Dilmegani, 2021, and see Figure 1). By 2024, Gartner projects, 60% of data

used for AI and machine learning will be synthetic data (White, 2021).

Rob Toews claims that 'the rise of synthetic data will completely transform the economics, ownership, strategic dynamics, even (geo)politics of data' (Toews, 2022). He cites Ofir Zuk, CEO and founder of synthetic data startup Datagen, claiming that the total addressable market of synthetic data and the total addressable market of data will converge.

'A substantial missed opportunity'?

Governments are looking at synthetic data and there are now a growing number of public sector use cases. However, some commentators are arguing that we should pick up the pace. Paul Calcraft writes that synthetic data 'is not yet a widely known technology in government, even among government analysts and researchers … this is a substantial missed opportunity' (Calcraft, 2022).

Stefanie James and colleagues illustrate how the technology for creating synthetic data has matured at a faster rate than the rate at which it has been adopted within organisations (James et al., 2021 – see Figure 2).

So synthetic data is here and growing fast, but what is it good for?

From scarcity to abundance

Cem Dilmegani of tech industry analysts AI Multiple summarises a central problem that synthetic data can address:

> Despite its success in a wide range of tasks, deep learning has an important limitation: its data-hungry nature. Collecting and labeling huge data with desired properties is costly, time-consuming, or unfeasible in some applications. (Dilmegani, 2021)

Synthetic data can replace data scarcity with abundance. It can augment real-world data when simply more volume is needed, and also balance real-world data when specific kinds of data is needed. As Rob Toews writes, 'synthetic data technology enables practitioners to simply digitally generate the data that they need, on demand, in whatever volume they require, tailored to their precise specifications' (Toews, 2022).

The development of autonomous vehicles is a good example. Given the risks they pose for all road users and pedestrians, the equivalent of hundreds of years of driving is needed to encompass a sufficiently wide set of scenarios. Already by 2016 Waymo had generated 2.5 billion miles of simulated driving data compared to 3 million miles of real-world driving data; by 2019 it had simulated 10 billion miles.

Big synthetic datasets can also better account for rare outlier events – 'edge cases' – by including them in the dataset at appropriate frequencies, and can also simulate conditions that have not yet been encountered (Dilmegani, 2022).

In general, more data can lead to better predictions (Krenchal and Cury, 2022), and so, for governments, more effective policy.

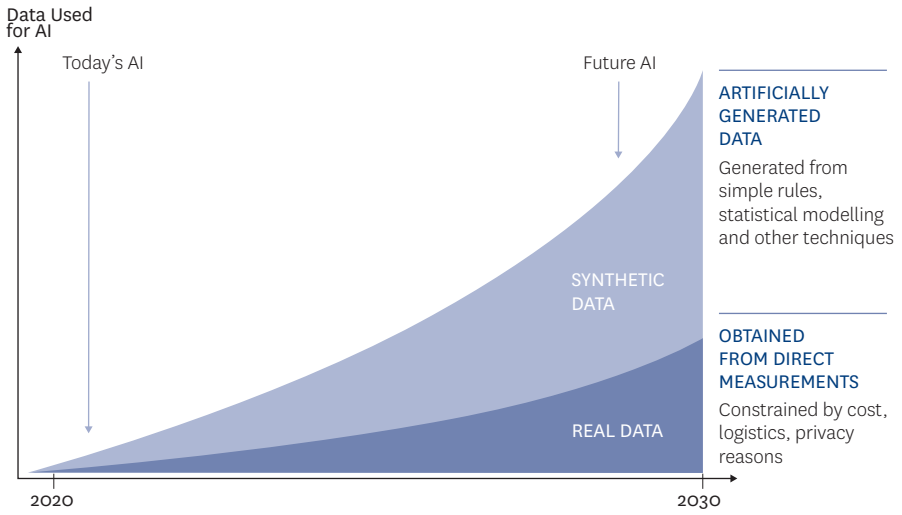
Analysis and insight without infringing privacy

Synthetic data can also represent real data when confidential information is involved. The US Census Bureau has used the technology for this purpose: it provides high-fidelity synthetic data built on a linked underlying dataset which combines the real-world census data with administrative tax and benefit data.

For the 2020 US census, the Census Bureau decided to release high-fidelity synthetic data that incorporated a form of 'differential privacy'. This is an advanced technique to further reduce the risk of an individual being identified, basically through adding random values – 'noise' – to the dataset at controlled levels. Notably, differential privacy allows government census agencies to precisely quantify the probability of an individual being identified through the synthetic dataset (Calcraft et al., 2021).

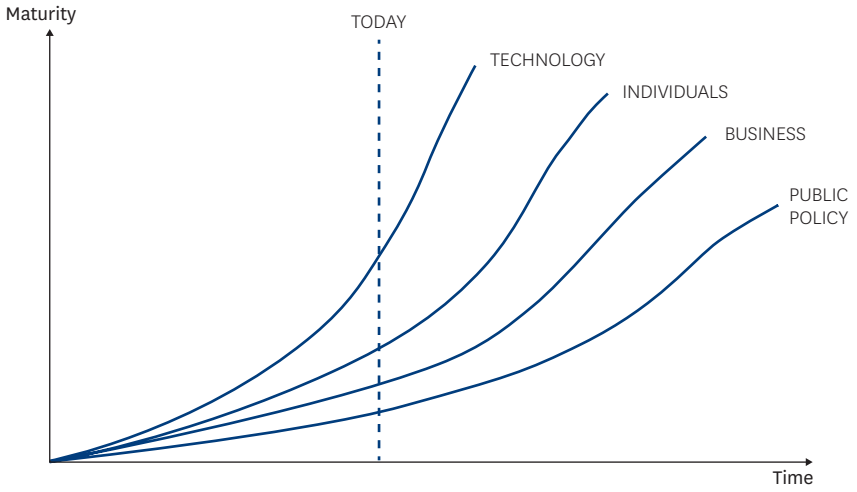
Using synthetic data addresses several different kinds and levels of privacy risks: 'singling out' – the possibility of

Figure 1: Projection of relative use of synthetic and real data in the AI sector



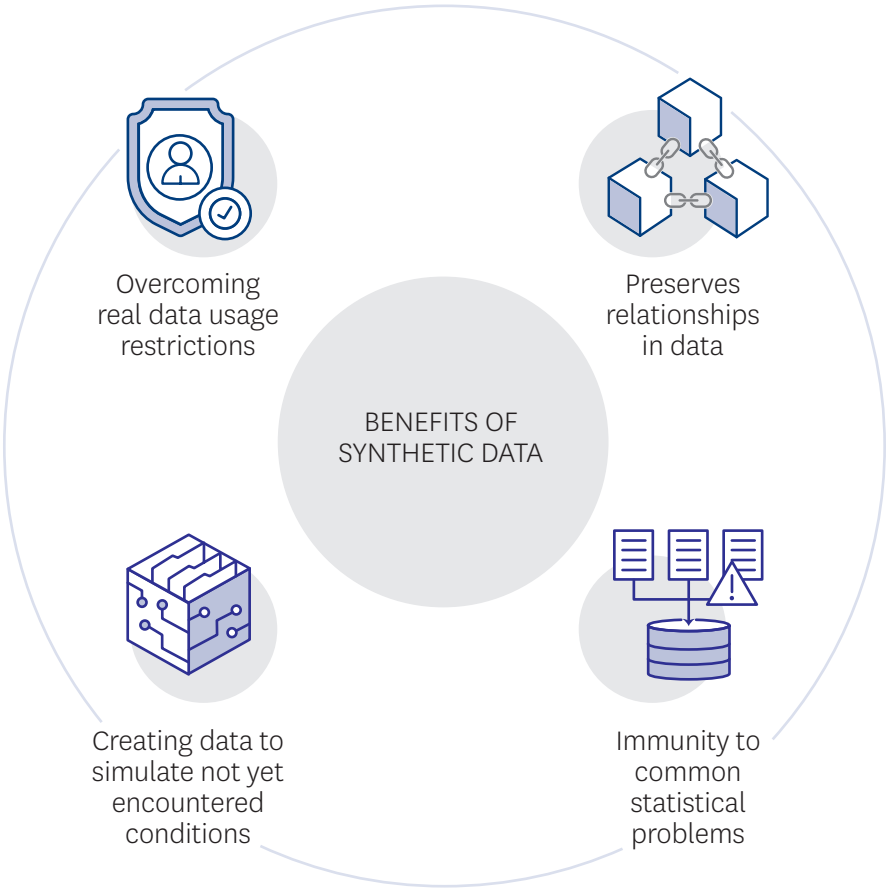
Source: Adapted from Dilmegani, 2021

Figure 2: Rate of development of synthetic data technology versus rate of adoption



Source: Adapted from James et al, 2021

Figure 3: Benefits of synthetic data



Source: Adapted from Dilmegani, 2018

distinguishing and identifying individual people; 'linkability' – the ability to link two or more data points concerning the same data subject within one or more datasets; and 'inference' – the possibility of deducing, with significant probability, the value given to other attributes within the dataset (Article 29 Data Protection Working Party, 2014). This can remove constraints in various situations, including allowing long-term research to continue when regulations limit the length of time that data can be stored.

Synthetic data technology could therefore have implications for the use of Māori data, as for indigenous populations elsewhere, by providing greater privacy protections. As Karaitiana Taiuru writes:

> Māori communities are especially vulnerable to privacy-related risks that come with (for example) the collection and storage of data on individual persons. The risk of individuals and whānau being re-identified through anonymised data is heightened when dealing with minority groupings and with sparsely distributed populations such as Māori. (Taiuru, 2020, p.8)

Synthetic datasets could potentially allow for meaningful analysis of data on Māori communities and individuals while better protecting privacy and Māori sovereignty over their data as taonga.

As an example, in an early Australian use case of synthetic data, Yogi Vidyattama and colleagues addressed the problem of a lack of data on indigenous disadvantage. They explained that 'spatial micro-simulation' techniques had usually been used to derive small area estimates of various social and economic indicators, with these estimates in turn used to help allocate government and community programmes for indigenous communities.

> However, for previous applications, a record unit file from a survey dataset has always been available on which to conduct the spatial microsimulation. For the case of indigenous disadvantage, this record unit file was not available due to the scarcity of the Indigenous population in Australia, and concerns from the ABS [Australian Bureau of Statistics] about confidentialising the file. (Vidyattama, Tanton and Biddle, 2013.

As a solution, Vidyattama and colleagues built a synthetic unit record file containing the same number of observations as the real-world survey file, and then applied spatial microsimulation to that synthetic dataset in order to generate the necessary small area estimates.

Reducing error and bias

Synthetic data can also sometimes be more faithful to the real world than real-world data, when the real-world dataset contains known sources of error and the synthetic data is corrected to remedy this.

One assessment has some 85% of the algorithms currently in use as error-prone, largely due to bias, which is in turn often due to samples under-representing women, non-white people, and other groups (Krenchel and Cury, 2022). Synthetic data could be part of the answer to this bias, because it can analyse real-world data and observe and compensate for bias, and it can generate much larger datasets that can better accommodate smaller groups and edge cases (Brouton Lab, 2022).

New Zealand's census provides an example. We learned from our last census that we under-sample minority groups in Aotearoa, and so any analysis of the census data will carry over that bias. But a synthetic dataset based on the census could add in a correction so that the synthetic data is more representative than the original, by adding records to make the synthetic dataset more in proportion to what we expect the data to contain.

With real-world data, the challenges involved in protecting privacy and combating bias can also sometimes be related, and inversely so: de-identification to protect privacy tends to amplify bias by removing minorities that could be re-identified. By contrast, synthesising data reduces the need for de-identification in the first place (see Box 1). What's more, it allows the option of generating synthetic data from the de-identified real-world dataset by creating extra records, as a compensatory virtual over-sampling.

Democratising data?

Finally, synthetic data can potentially have major implications for the relationships that Meta/Facebook and the other digital behemoths have with the rest of us, as their commercial and social power rests on their command of and ready access to oceans of customer data. Synthetic data can potentially enable lots of AI and other startups to drive innovation. People other than data scientists would be able to readily build dashboards, and synthetic data also lends itself more to crowdsourcing innovation (Kohli, 2021).

So, synthetic data can potentially level the playing field, which would in turn present another wave of public policy and regulatory challenges; but that's a topic for another article.

Generating synthetic data

The process of generating synthetic data from a real-world dataset is called, logically enough, 'synthesis', and there are different techniques.

A key group of techniques are 'deep generative models' – or DGMs – which Lars Ruthotto and Eldad Haber describe as one of the 'most hotly researched fields' in AI in recent years. These are 'neural networks' that are trained to analyse samples and recognise and approximate complicated probability distributions involving a large number of different dimensions and variables. 'When trained successfully, we can use the DGM to estimate the likelihood of each observation and to create new [that is, synthetic] samples from the underlying distribution' (Ruthotto and Haber, 2021).

GANs and VAEs

One of the most popular deep generative models for synthesising tabular data (as opposed to images or text) is 'generative adversarial networks', or GANs. 'Adversarial' here refers to the fact that GANs pit two neural networks against each other in a contest.

The first network is called the 'generator', and, in the original application of GANs to images, it would create new images, such as human faces that are similar to real faces. The second network is called the 'discriminator': it looks at images of both real and created faces without being told which are which. The generator keeps trying to fool the discriminator and the discriminator keeps trying to see through the deception. Over time the discriminator's success rate drops below 50% – in other words, no better than guessing at whether an image is real or synthetic.

Data scientist Alex Wang says that while deep generative models have been shown to work for images, audio and molecular synthesis, their application to tabular data is still at an early stage, with various unresolved challenges.[1] GANs, and also another type of model called a VAE (variational auto encoder), can work well with tabular data, and in some cases the two types – GANs and VAEs – have been combined. These approaches to generating synthetic data have demonstrated quite high 'utility values' (that is, a high degree of fidelity to the relationships in the real data) working from complex datasets and are a very active area of research (Emam, Mosquera and Hoptroff, 2020).

Language AI is of course a fast-moving area. Before ChatGPT burst on the scene in late 2022, Daniel Yogatama from the AI firm DeepMind, a next-generation synthetic data technology involving 'massive foundation models' can generate unstructured text at a new level of 'realism, originality, sophistication and diversity', and often indistinguishable from human-written text:

> This new type of synthetic data has been successfully applied to build a wide range of AI products, from simple text classifiers to question-answering systems to machine translation engines to conversational agents. Democratizing this technology is going to have a transformative impact on how we develop production AI models. (Toews, 2022)

Choosing the right synthesis method

Data scientist Marianna Pekar says that there is no one right way of synthesising data, and that it always depends on the underlying dataset:

> As a rule of thumb, the generation method should be suited to the complexity of the underlying data. Machine learning and deep learning models are the only real practical techniques for handling high data complexity, but on the other hand deep learning models can perform poorly on simple datasets.[2]

Data scientist Marianna Pekar

Marianna adds that, as with just about any human activity, an element of subjectivity creeps into the choice of method: different analysts choose a method they prefer and continuously optimise it (Emam, Mosquera and Hoptroff, 2020).

Verifying a synthetic dataset's 'utility'

The original 1970 vinyl of Jimmy Cliff's 'Synthetic world' would probably have advertised it as 'hi-fi'. Fidelity is a central property of synthetic data too. Unlike completely random dummy data used simply to test new systems, synthetic data is useful because it is faithful, in key respects, to the original data.

Data scientists use the term 'utility' to describe the value and usefulness of synthetic data. In turn, utility depends centrally on the fidelity, or similarity, of the synthetic dataset to the real dataset.

In assessing and measuring a synthetic dataset's utility and the degree of fidelity, data scientists apply various empirical tests: for example, testing for 'prediction accuracy', which assesses the ability of the synthetic data to replicate the results of a prediction analysis performed on real data.

Marianna Pekar emphasises that using the right synthesis techniques is crucial for achieving a high degree of fidelity and utility while also minimising the risk of re-identification. High fidelity and utility does not necessarily mean a greater risk of re-identification and therefore of a privacy

breach, but it *may* do if you haven't applied the right synthesis techniques and tools.

Hi-fi and lo-fi synthetic data

Consider a dataset with the height and weight of a group of athletes: low-fidelity synthetic data would represent the patterns of height and weight, but it would provide no information about the relationship between the heights and the weights – for example, whether the taller people tend to be heavier. High-fidelity synthetic data, by contrast, would include that relationship. The data in the high-fidelity dataset is partially random, in that it doesn't relate to any real data points, but it is generated around the line that represents that height–weight relationship – and potentially many other relationships within the data.

But it's not that high fidelity is good and low is bad. It may be that you don't need your synthetic dataset to be faithful to many of the statistical relationships in the real-world dataset, and that low fidelity meets your purposes perfectly.

Use cases: synthetic data in the real world

The concept of synthetic data was first applied commercially at scale in the autonomous vehicle sector in the mid-2010s (Toews, 2022). Use cases in other sectors quickly followed, including robotics, geospatial imagery, banking, and genome studies into diseases.

Although the biggest users and innovators continue to be in the autonomous vehicle sector, a distinct synthetic data sector is growing quickly too. One example is the Synthetic Data Vault established by Data to AI Lab (DAI), which has links to the Massachusetts Institute of Technology (MIT). This is an open-source and scalable collection of libraries offering the latest tools to all, whether students or large organisations (MIT Laboratory for Information and Decision Systems, 2020).

Public sector examples in the Anglosphere

Since the groundbreaking US census example from the early 1990s, there have been several US public sector examples in the area of health records. Here's one high-profile use case:

> The National Institutes of Health used synthetic data to replicate their database of more than 2.7 million COVID-19 patient records, creating a dataset with the same statistical properties but none of the identifying information that could be quickly shared and studied by researchers the world over. The aim was to help identify better treatments without infringing on the privacy of the people involved. (Krenchel and Cury, 2022)

The US Office of the National Coordinator for Health Information Technology is also using publicly available health data to generate a synthetic dataset which will be used for testing and refining analyses, as a prerequisite for researchers being granted access to real data.

In the UK, this was also the intention driving low-fidelity synthetic data being produced for the Ministry of Housing, Communities and Local Government and its 'Troubled Families' project. I also referred earlier to the creation of a synthetic dataset in Australia as a solution to a lack of available data, because of small populations and privacy concerns, on social and economic indicators for indigenous populations (Vidyattama, Tanton and Biddle, 2013).

Stefanie James et al. cite the 'Simulacrum', a synthetic dataset project from the UK health sector:

> The Simulacrum imitates data held by Public Health England's National Cancer Registration and Analysis Service. Scientists get access to Simulacrum synthetic data[;] once the scientific query is refined scientists are able to submit a request to Public Health England to run queries on the real data. Public Health England will provide aggregate and anonymous data back to the scientist. Scientists are able to publish results based on the synthetic data. (James et al., 2021)

[New Zealand use cases of synthetic data] demonstrate how this new technology can contribute to addressing a variety of critical issues for Aotearoa – here, climate change and dependency on fossil fuels, the housing affordability crisis, and social services and wellbeing.

Three synthetic data use cases in Aotearoa

The following three examples do not exhaust the list of public sector use cases in this country. However, they demonstrate how this new technology can contribute to addressing a variety of critical issues for Aotearoa – here, climate change and dependency on fossil fuels, the housing affordability crisis, and social services and wellbeing.

Modelling the impact of wind farms on New Zealand's national grid

Back in 2009, in an early use of synthetic data here, NIWA and MetService created synthetic ten-minute wind datasets at 15 actual or potential wind farm sites across the country, to help the Electricity Commission model the impact of wind farms on the national grid (NIWA, 2009, n.d.). Wind data at ten-minute to hourly time scales is a key factor in modelling the performance of wind farms. However, little of this data is publicly available, whether for existing or proposed sites, and so it was decided to simulate it.

The project team first developed an hourly synthetic dataset, drawing on several years of archived wind data for the

whole of New Zealand based on a 12km grid (called NWP wind data, for 'numeral weather prediction').

> By developing a robust statistical relationship between these hourly NWP winds and hourly speeds observed at hub-height at wind farms it was possible to produce an hourly synthetic wind dataset which preserved the statistical properties of the hourly observed data. To then obtain a ten-minute synthetic dataset with all the desired properties of the ten-minute wind farm observations, realistic ten-minute fluctuations in wind speeds for these wind farm sites were then superimposed on the hourly time-series. (NIWA, n.d.)

Particular attention was paid to accurately simulating the frequencies of wind speeds that are outside the operating ranges of the turbines.

The outcome was a realistic synthetic dataset for uses such as preparing generation scenarios during storms, calculating wind power's contribution to total capacity, and estimating seasonal variations in wind-power generation.

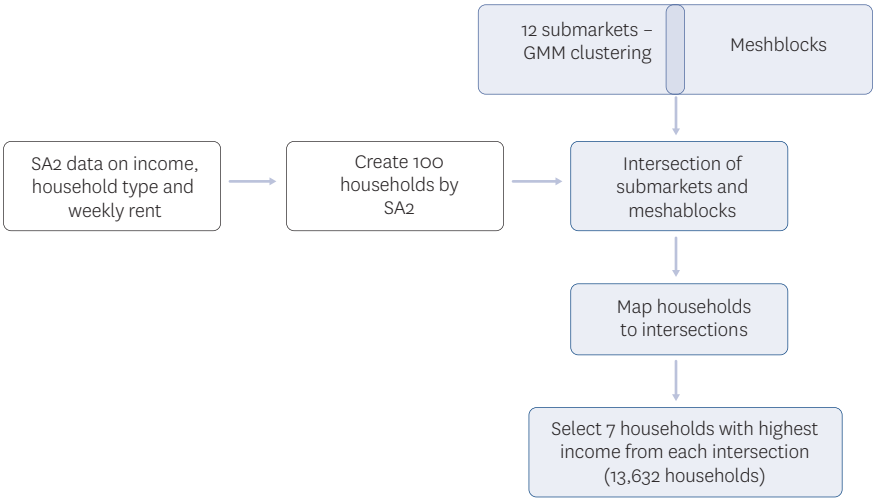
Policy responses to a housing affordability crisis in Auckland

More recently, Mario Fernandez and three colleagues used synthetic data to simulate some of the levers that local and central government could use to affect housing prices and affordability in Auckland, such as direct intervention on the supply side and subsidies (Fernandez et al., 2022). Specifically, they simulated a retention-and-targeting programme (where houses are temporarily retained for sale to households earning below an income threshold), and subsidies to raise deposits through shared ownership.

Fernandez and colleagues (all employed by or affiliated with Auckland Council at the time) wanted to address three questions: what annual rate of growth of affordable housing would solve the affordability crisis; consequently, how long would it take to solve the crisis; and how much would that policy package cost?

The team constructed a sample of about 13,000 synthetic households, representing households searching for and bidding for a new dwelling in Auckland. The model worked by running two rounds of bidding: first, households bid for a dwelling in their local submarket; second, if they were outbid locally, they then bid in two adjacent submarkets above and below.

Figure 4: Construction of the sample of synthetic households



Source: Adapted from Fernandez et al., 2022

This simulation was run in two different supply scenarios, each with 6,000 dwellings: a 'competitive' market scenario, reflecting the current housing stock with an average price of $1.5 million, and an 'affordable' market scenario with an average price of $833,363. The aim was to simulate market behaviour and estimate the rate of housing take-up in each scenario, and to explore whether the distributions of prices set by developers and the income of households lead to more affordable housing.

The simulation included a number of variables, including latitude and longitude; distances to the nearest beach, waterway, road, open space, school and CBD; and sales price, floorspace, slope and elevation.

The approach gave the authors confidence to identify a possible package of policies to materially improve the affordability of housing in Tāmaki Makaurau. They wrote:

> Results in this paper should be interpreted as the boundaries of what is feasible and realistic in the realm of affordability policies … Its scope is a blueprint for the design of policies in other cities where unaffordability has become extreme. (ibid.)

Synthetic data meets social services and wellbeing

Marianna Pekar, whom I mentioned earlier, is currently working with VUW-based data scientist Alex Wang on an exciting three-year research project involving New Zealand's Integrated Data Infrastructure (IDI). The research is funded by the Informatics for Social Services and Wellbeing Programme | Te Rourou Tātaritanga, through the MBIE Endeavour Fund, and its aims include evaluating the synthesising of datasets in a key area of public policy – social services and wellbeing. This particular research, which builds on previous investigations, is supervised by Professor Binh Nguyen of the School of Mathematics and Statistics at Victoria University.

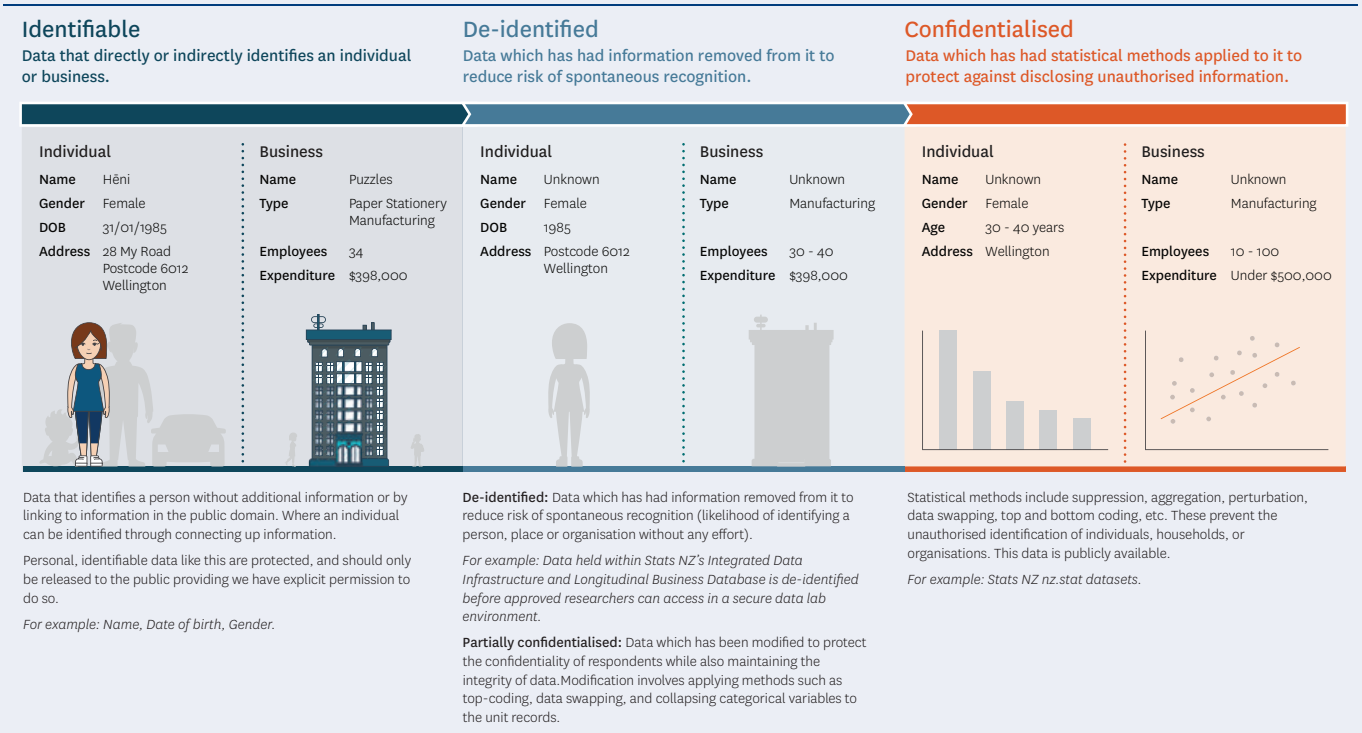
The Integrated Data Infrastructure is a large research database which holds de-identified microdata about people and households. It covers life events and use of government services like education, income support, justice and health. The data comes from government agencies, Statistics New Zealand surveys and NGOs. The data is linked and integrated together to form the IDI. The IDI is therefore a powerful tool for evidence-based policymaking in Aotearoa. Researchers use it mostly for cross-sector research that provides insights into our society and economy.

So, why would one want to synthesise data in the IDI? Well, even though the data is de-identified, it's still too granular to be made public, because there would be a high

BOX 1 Privacy and the IDI: understanding 'de-identification' and 'confidentialisation'

Statistics New Zealand gives a simple example of these different terms (Statistics New Zealand, 2018). 'De identifying' data about a named individual would remove their name and give just their gender, the year of birth rather than their precise birth date, and just their postcode and city or area. But after that de-identification there would still be some risk of spontaneous recognition – recognition without any effort. By contrast, 'confidentialising' this item of data would give just the person's gender, a ten-year age bracket, and the city or region where they live.

Figure 5: Degrees of identification of data



Source: Stats NZ, 2018, www.data.govt.nz, Creative Commons 4.0, (brief introductory text removed)

re-identification risk. So IDI security is tight: researchers are vetted and must use the data onsite, where it is stored in secure locations. And before research results can be published, the researchers have to 'confidentialise' the data – for example, by aggregating it and suppressing small counts.

The shared research environment of the IDI is unique in the world and is a taonga of New Zealand (Jones et al., 2022). However, it is not the right environment for applying resource-intensive methods that take in microdata inputs. Data synthesis can bring benefits for the use of IDI data by potentially allowing for more tabulations at more granular levels (for example, lower levels of geography) that aggregation rules currently prohibit, and by also allowing for known sources of error to be corrected.

Pekar and Wang's research project is looking at relevant use cases to assess the advantages and disadvantages of using different methods of synthetic data generation for different purposes. This includes using low-fidelity synthetic data generated outside the IDI for training and to demonstrate methods. It also includes using high-quality synthetic datasets generated inside the IDI environment to assess the advantages of using advanced machine-learning methods outside the IDI.

An additional phase of the research looks at the tests and requirements that synthetic data should need to pass before being released from the secure IDI environment. The project team is working closely with the statistical methods team from Statistics New Zealand to determine a selection of suitable statistical tests that strike the balance between fidelity and mitigating the risk of re-identification.

Synthetic datasets generated from the IDI also have potential benefits that go beyond the scope of this particular research. Researchers would be allowed to leave the secure IDI environment (the data labs operated by Statistics New Zealand) and work remotely. After completing their research, they would also be able to make the data more broadly available for others to test reproducibility and for secondary analysis.

Researchers would also have the freedom to apply resource-intensive methods with microdata as input: for example, Explainable AI (XAI) techniques to detect bias and hidden relationships between inputs, models and outputs, and agent-based micro-simulation to model future outcomes (with micro-simulation, users do 'what-if' analyses and run novel scenarios) (Emam, Mosquera and Hoptroff, 2020).

Challenges and risks involved with using synthetic data

There are sceptics about the use of synthetic data. For example, Neil Raden,

an actuary, has concerns around privacy and anonymisation: he suggests that anonymising data does not work when some personal information is necessary for the model to draw inferences – for example, in medical research (Raden, 2021). He is also concerned that anonymisation might sometimes be reversible.

Other challenges that commentators have pointed to include variable user acceptance, because it is a new technology and users are still learning its limitations and learning to trust it (Dilmegani, 2018). Calcraft et al. (2021) add structural barriers like lack of knowledge, technical capability, and legal concerns within public sector bodies.

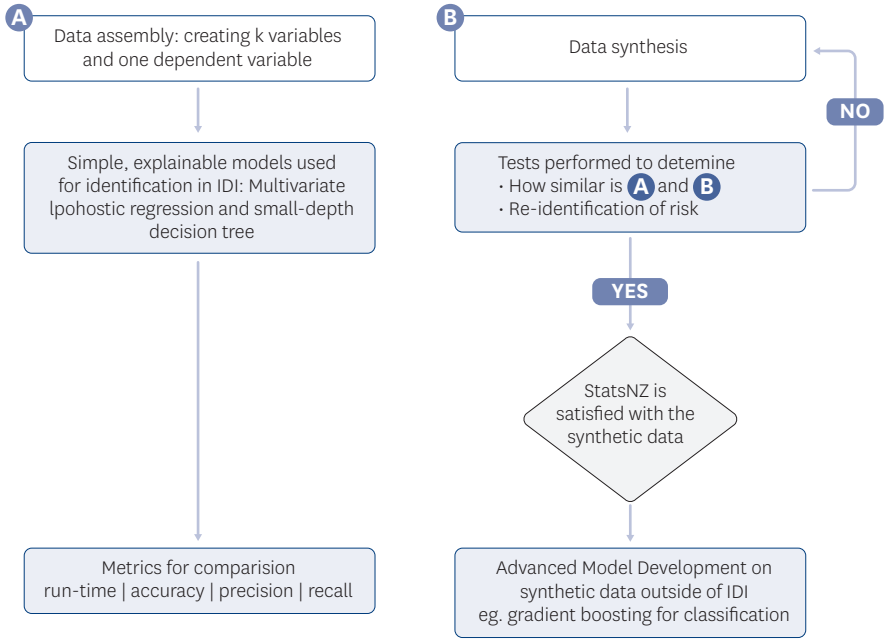
Raden (2021) also reminds us of the undeniable point that the quality of synthetic data relies on the quality of the real-world data it's based on, as well as on the quality of the model that generated it.

Of course, synthetic data technology cannot magically generate knowledge and insights that we would not otherwise have access to through conventional real-world research and analytical techniques. If we have no data on a particular variable, then synthetic data technology will not create it. Similarly, if the real-word data contains systematic biases, those biases will be carried over into the new synthetic dataset unless those biases are known and corrected for.

Mikkel Krenchel and Maria Cury (2022) see the answer to these challenges as being transparency and data literacy:

> [W]e believe the social and human sciences ought to get involved. The input most crucial to making sure the synthetic data revolution does not simulate low-quality reflections of the world we live in (or worse, create worlds we didn't intend) is small, not big, data. In a synthetic data world, the quality of the initial, small dataset from which the synthetic data is derived, is absolutely paramount. And so is a deeply contextualized understanding of that dataset itself – where it came from, what it can be used for, what it explains, and what it doesn't. This is the kind of context that is difficult to obtain, make sense of, or relate to underlying structures and biases.

Figure 6: A simple flow diagram of the joint research project 'The Impact of Synthetic Data Generation Techniques in IDI Research'



Source: Marianna Pekar

Krenchel and Cury wonder if the future will see an 'AI dance' between human imagination and intuition and its machine counterpart. They are justifiably definitive, though, that

> [t]he stakes are too high to leave these important decisions to data scientists alone – social scientists and philosophers (as well as policymakers) have a role to play. Otherwise, the effects of this data revolution could be disastrous.

How to facilitate the role of synthetic data in helping develop sound public policy

Paul Calcraft and colleagues in the UK have looked at how synthetic data can accelerate public policy research without privacy risks, and they made some recommendations to the government partnership body, Administrative Data Research UK (Calcraft et al., 2021). These included using lo-fi synthetic data across government and researchers to reveal whether data for a given policy is available and usable; for writing and testing code before access to real-world data is available; and to provide quicker access where there are data security issues.

They also recommended the development of a cross-government synthetic data repository accessible to accredited researchers and government policy analysts (reminiscent of the rules of access to the IDI here). This would assist with the discovery of available data, with the design of more informed research questions and plans, and with establishing a semi-automated pipeline generating lo-fi synthetic data at the end of each project.

Stefanie James et al. (2021) also discuss some technical and organisational measures to guide the effective, efficient and economical use of synthetic data. They emphasise that whoever synthesises the data needs not just technical capability but also knowledge and understanding of privacy requirements and risks. In order for the synthetic dataset to be trusted, documentation about its utility should also be embedded within it, and the organisation needs to ensure transparency and an audit trail.

The organisation also needs to have the necessary infrastructure, tools and data-sharing processes, whether in-house or bought as a service. James et al. also recommend that organisations take the opportunity to build end-to-end pipelines so that synthetic datasets can be used for multiple purposes, not all of which will be known in the design phase.

Proper training of their people would presumably be critical to organisations successfully managing the risks and

opportunities presented by synthetic data, training that covers not just technical issues but also the ethical ones.

Many rivers to cross

'Many rivers to cross', the title of another Jimmy Cliff hit, aptly describes the challenges that will need to be overcome for synthetic data to meet the optimistic predictions. However, the potential gains are considerable. By effectively allowing for important data to be made public, synthetic datasets allow for public policy researchers and analysts to subject each other's work to the same scrutiny and checking for reproducibility that goes on in the natural sciences. This potentially means not just better, more effective policy, but also greater transparency and therefore greater trust that significant policy decisions are based on sound evidence.

The rise of synthetic data is an international phenomenon that has now seen several notable use cases in Aotearoa. There is, however, an opportunity here to increase the pace of adoption and ensure that the full benefits of this new area of data science are realised.

1 Conversation with the author, September 2022.
2 Conversation with the author, September 2022.
3 Conversation with the author, September 2022.

Acknowledgements

The author thanks Marianna Pekar for bringing the subject of synthetic data and some key books and articles to his attention, for contributing material for the article, and for comments on an earlier draft. He also thanks Alex Wang for contributing material and for comments on the draft.

For their comments on earlier drafts the author would also like to thank: Professor Colin Simpson, associate dean research, Wellington Faculty of Health, VUW; Sarah Lomas, manager spatial analysis and modelling, Auckland Council; Mario Fernandez, formerly senior researcher at Auckland Council, now principal economist at DairyNZ; Anna McDowell, general manager data services, Statistics New Zealand; Kevin Ross, CEO of Precision Driven Health; Ken Quarrie, chief scientist for New Zealand Rugby; and an anonymous reviewer.

The author would also like to thank Marcus Pawson for editing and comments on drafts, and Sharyn Jones for the infographics.

References

AltexSoft (2022) 'Synthetic data for machine learning: its nature, types, and means of generation', blog post, 22 March, https://www.altexsoft.com/blog/synthetic-data-generation/

Article 29 Data Protection Working Party (2014) 'Opinion 05/2014 on anonymisation techniques', adopted 10 April, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Asiamah, N., H. Mensah and E. Oteng-Abayie (2017) 'General, target and accessible population: demystifying the concepts for effective sampling', *The Qualitative Report*, 22 (6), pp.1607–21

Benedetti, M. (2018) 'The advantages and limitations of synthetic data', blog post, 24 January, https://www.sama.com/blog/2018-01-24-the-advantages-and-limitations-of-synthetic-data/

Breiman, L., J. Friedman, C. Stone and R. Olshen (1984) *Classification and Regression Trees*, Milton Park: Taylor & Francis

Brouton Lab (2022) 'How can synthetic data solve the AI bias problem?', blog post, https://broutonlab.com/blog/ai-bias-solved-with-synthetic-data-generation

Calcraft, P. (2022) 'Accelerating public policy research with easier & safer synthetic data', blog post, Behavioural Insights Team, 2 March, https://www.bi.team/blogs/accelerating-public-policy-research-with-easier-safer-synthetic-data/

Calcraft, P., I. Thomas, M. Maglicic and A. Sutherland (2021) 'Accelerating public policy research with synthetic data', 14 December, Behavioural Insights Team, https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf

Chawla, N., K. Bowyer, L. Hall and W. Kegelmeyer (2002) 'Smote: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp.321–57, https://www.jair.org/index.php/jair/article/view/10302/24590

Dilmegani, C. (2018) 'What is synthetic data? What are its use cases & benefits?', 19 July (updated 1 September 2022), AI Multiple, https://research.aimultiple.com/synthetic-data/

Dilmegani, C. (2021) 'Synthetic data to improve deep learning methods', 28 May (updated 18 April 2022), AI Multiple, https://research.aimultiple.com/synthetic-data-for-deep-learning/

Dilmegani, C. (2022) 'Top 20 synthetic data use cases & applications in 2022', 10 July (updated 11 February), AI Multiple, https://research.aimultiple.com/synthetic-data-use-cases/

Emam, K. (2020) Accelerating *AI with Synthetic Data: generating data for AI projects*, O'Reilly Media

Emam, K., L. Mosquera and R. Hoptroff (2020) *Practical Synthetic Data Generation: balancing privacy and the broad availability of data*, O'Reilly Media

Fernandez, M., J. Joynt, C. Hu and S. Martin (2022) 'Sorting (and costing) the way out of the housing affordability crisis in Auckland, New Zealand', *International Journal of Housing Markets and Analysis*, https://doi.org/10.1108/IJHMA-04-2022-0061

James, S., C. Harbron, J. Branson and M. Sundler (2021) 'Synthetic data use: exploring use cases to optimize data utility', *Discover Artificial Intelligence*, 1 (15)

Jones, C., A. McDowell, V. Galvin and D. Adams (2022) 'Building on Aotearoa New Zealand's Integrated Data Infrastructure', *Harvard Data Science Review*, 4 (2), https://hdsr.mitpress.mit.edu/pub/9dkr3v8v/release/1

Kaloskampis, I. (2019) 'Synthetic data for public good', 21 February, Data Science Campus, https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

Kohli, S. (2021) 'Why synthetic data could be useful for a government department', DWP Digital blog post, Department for Work and Pensions, 18 June, https://dwpdigital.blog.gov.uk/2021/06/18/why-synthetic-data-could-be-useful-for-a-government-department/

Krenchel, M. and M. Cury (2022) 'We should all be worried about synthetic data', *IAI News*, 20 May, https://iai.tv/articles/we-should-all-be-worried-about-synthetic-data-auid-2138

MIT Laboratory for Information and Decision Systems (2020) 'The real promise of synthetic data', *MIT News*, 16 October, https://news.mit.edu/2020/real-promise-synthetic-data-1016

Nikolenko, S. (2021) *Synthetic Data for Deep Learning*, Springer Cham, https://doi.org/10.1007/978-3-030-75178-4

NIWA (2009) *Multi-Year Ten-Minute Synthetic Wind Speed Time-Series for 15 Actual or Proposed New Zealand Wind Farms*, NIWA Client Report: WLG2009-43 June, https://niwa.co.nz/sites/niwa.co.nz/files/import/attachments/NIWA-Synthetic-Wind-Study-report.pdf

NIWA (n.d.) 'Generating synthetic wind data', https://niwa.co.nz/environmental-information/research-projects/synthetic-wind-data

Nowok, B., G. Raab and C. Dibben (2016) 'synthpop: bespoke creation of synthetic data in R', *Journal of Statistical Software*, 74 (11), https://www.jstatsoft.org/article/view/v074i11

Peter, D. (2021) 'The promise and pitfalls of synthetic data', *University Affairs*, 13 December

Raden, N. (2021) 'Synthetic data for AI modelling? I'm still not convinced', *Diginomica*, 13 October, https://diginomica.com/synthetic-data-ai-modeling-im-still-not-convinced

Roelofs, F. et al. (2020) 'SYMBA: an end-to-end VLBI synthetic data generation pipeline: simulating event horizon telescope observations of M 87', *Astronomy and Astrophysics*, 636, A5, https://doi.org/10.1051/0004-6361/201936622

Ruthotto, L. and E. Haber (2021) 'An introduction to deep generative modeling', *GAMM-Mitteilungen*, 44 (2), https://onlinelibrary.wiley.com/doi/10.1002/gamm.202100008

Snoke, J., G. Raab, B. Nowok, C. Dibben and A. Slavkovic (2018) 'General and specific utility measures for synthetic data', *Journal of the Royal Statistical Society: Series A* (Statistics in Society), 181 (3), pp.663–88, https://rss.onlinelibrary.wiley.com/doi/10.1111/rssa.12358

Statistics New Zealand (2018) 'Data confidentiality principles and methods report', https://www.data.govt.nz/assets/Uploads/data-confidentiality-principles-methodology-report-oct-2018.pdf

Taiuru, K. (2020) 'Treaty of Waitangi/Te Tiriti and Mäori ethics guidelines for: AI, algorithms, data and IOT', http://www.taiuru.Maori.nz/TiritiEthicalGuide

Toews, R. (2022) 'Synthetic data is about to transform artificial intelligence', *Forbes*, 12 June, https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=4aedbdbd7523

Vidyattama, Y., R. Tanton and N. Biddle (2013) *Small Area Social Indicators for the Indigenous Population: synthetic data methodology for creating small area estimates of Indigenous disadvantage*, working paper 13/24, National Centre for Social and Economic Modelling, University of Canberra

White, A. (2021) 'By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated', blog post, 24 July, https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/