O'REILLY®

Practical Synthetic Data Generation

Balancing Privacy and the Broad Availability of Data

Khaled El Emam,
Lucy Mosquera &
Richard Hoptroff

Practical Synthetic Data Generation
by Khaled El Emam, Lucy Mosquera, and Richard Hoptroff

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Jonathan Hassell
Development Editor: Corbin Collins
Production Editor: Christopher Faucher
Copyeditor: Piper Editorial
Proofreader: JM Olejarz
Indexer: Potomac Indexing, LLC
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Jenny Bergman

May 2020: First Edition

Revision History for the First Edition
2020-05-19: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781492072744* for release details.

978-1-492-07274-4

[LSI]

CHAPTER 1

Introducing Synthetic Data Generation

We start this chapter by explaining what synthetic data is and its benefits. Artificial intelligence and machine learning (AIML) projects run in various industries, and the use cases that we include in this chapter are intended to give a flavor of the broad applications of data synthesis. We define an AIML project quite broadly as well, to include, for example, the development of software applications that have AIML components.

Defining Synthetic Data

At a conceptual level, synthetic data is not real data, but data that has been generated from real data and that has the same statistical properties as the real data. This means that if an analyst works with a synthetic dataset, they should get analysis results similar to what they would get with real data. The degree to which a synthetic dataset is an accurate proxy for real data is a measure of ***utility***. We refer to the process of generating synthetic data as ***synthesis***.

Data in this context can mean different things. For example, data can be ***structured*** data, as one would see in a relational database. Data can also be ***unstructured*** text, such as doctors' notes, transcripts of conversations or online interactions by email or chat. Furthermore, images, videos, audio, and virtual environments are types of data that can be synthesized. Using machine learning, it is possible to create realistic pictures of people who do not exist in the real world.

There are three types of synthetic data. The first type is generated from actual/real datasets, the second type does not use real data, and the third type is a hybrid of these two. Let's examine them here.

Synthesis from Real Data

The first type of synthetic data is synthesized from real datasets. This means that the analyst has some real datasets and then builds a model to capture the distributions and structure of that real data. Here ***structure*** means the multivariate relationships and interactions in the data. Once the model is built, the synthetic data is sampled or generated from that model. If the model is a good representation of the real data, then the synthetic data will have statistical properties similar to those of the real data.

This is illustrated in Figure 1-1. Here we fit the data to a generative model first. This captures the relationships in the data. We then use that model to generate synthetic data. So the synthetic data is produced from the fitted model.

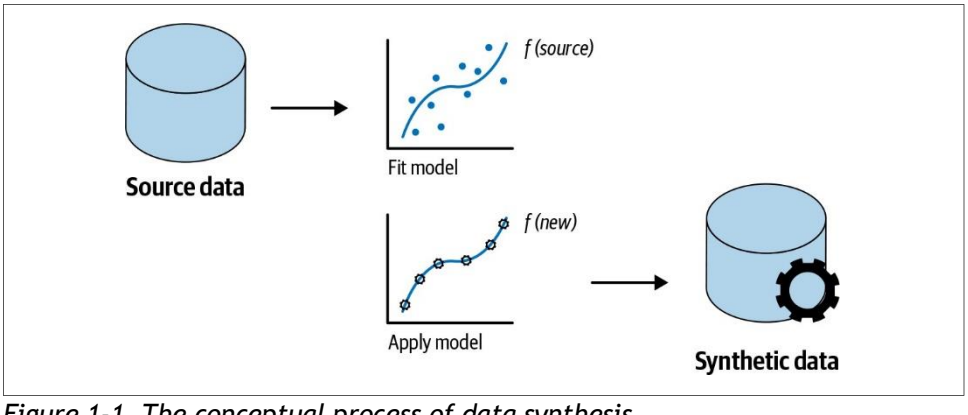

Figure 1-1. The conceptual process of data synthesis

For example, a data science group specializing in understanding customer behaviors would need large amounts of data to build its models. But because of privacy or other concerns, the process for accessing that customer data is slow and does not provide good enough data on account of extensive masking and redaction of information. Instead, a synthetic version of the production datasets can be provided to the analysts to build their models with. The synthesized data will have fewer constraints on its use and will allow them to progress more rapidly.

Synthesis Without Real Data

The second type of synthetic data is not generated from real data. It is created by using existing models or the analyst's background knowledge.

These existing models can be statistical models of a process (developed through sur- veys or other data collection mechanisms) or they can be simulations. Simulations can be, for instance, gaming engines that create simulated (and synthetic) images of scenes or objects, or they can be simulation engines that generate shopper data with

particular characteristics (say, age and gender) for people who walk past a store at different times of the day.

Background knowledge can be, for example, knowledge of how a financial market behaves that comes from textbook descriptions or the movements of stock prices under various historical conditions. It can also be knowledge of the statistical distribution of human traffic in a store based on years of experience. In such a case, it is relatively straightforward to create a model and sample from background knowledge to generate synthetic data. If the analyst's knowledge of the process is accurate, then the synthetic data will behave in a manner that is consistent with real-world data. Of course, the use of background knowledge works only when the analyst truly understands the phenomenon of interest.

As a final example, when a process is new or not well understood by the analyst, and there is no real historical data to use, then an analyst can make some simple assumptions about the distributions and correlations among the variables involved in the process. For example, the analyst can make a simplifying assumption that the variables have normal distributions and "medium" correlations among them, and create data that way. This type of data will likely not have the same properties as real data but can still be useful for some purposes, such as debugging an R data analysis program, or some types of performance testing of software applications.

Synthesis and Utility

For some use cases, having high utility will matter quite a bit. In other cases, medium or even low utility may be acceptable. For example, if the objective is to build AIML models to predict customer behavior and make marketing decisions based on that, then high utility will be important. On the other hand, if the objective is to see if your software can handle a large volume of transactions, then the data utility expectations will be considerably lower. Therefore, understanding what data, models, simulators, and knowledge exist, as well as the requirements for data utility, will drive the specific approach for generating the synthetic data.

A summary of the synthetic data types is given in Table 1-1.

Table 1-1. Different types of data synthesis with their utility implications

Type of synthetic data	Utility
Generated from real nonpublic datasets	Can be quite high
Generated from real public data	Can be high, although there are limitations because public data tends to be de-identified or aggregated
Generated from an existing model of a process, which can also be represented in a simulation engine	Will depend on the fidelity of the existing generating model
Based on analyst knowledge	Will depend on how well the analyst knows the domain and the complexity of the phenomenon

Type of synthetic data	Utility
Generated from generic assumptions not specific to the phenomenon	Will likely be low

Now that you have seen the different types of synthetic data, let's look at the benefits of data synthesis overall and of some of these data types specifically.

The Benefits of Synthetic Data

We will highlight two important benefits of data synthesis: providing more efficient access to data and enabling better analytics. Let's examine each of these in turn.

Efficient Access to Data

Data access is critical to AIML projects. The data is needed to train and validate models. More broadly, data is also needed for evaluating AIML technologies that have been developed by others, as well as for testing AIML software applications or applications that incorporate AIML models.

Typically, data is collected for a particular purpose with the consent of the individual—for example, for participating in a webinar or a clinical research study. If you want to use that same data for a different purpose, such as to build a model to predict what kind of person is likely to sign up for a webinar or to participate in a clinical study, then that is considered a secondary purpose.

Access to data for secondary purposes, such as analysis, is becoming problematic. The Government Accountability Office[1] and the McKinsey Global Institute[2] both note that accessing data for building and testing AIML models is a challenge for their adoption more broadly. A Deloitte analysis concluded that data-access issues are ranked in the top three challenges faced by companies when implementing AI.[3] At the same time, the public is getting uneasy about how its data is used and shared, and privacy laws are becoming stricter. A recent survey by O'Reilly highlighted the privacy concerns of companies adopting machine learning models, with more than half of companies experienced with AIML checking for privacy issues.[4]

Contemporary privacy regulations, such as the US Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) in

1 US Government Accountability Office, "Artificial Intelligence: Emerging Opportunities, Challenges, and Implications for Policy and Research" (March 2018) *https://www.gao.gov/products/GAO-18-644T*.

2 McKinsey Global Institute, "Artificial intelligence: The next digital frontier?", June 2017. *https://oreil.ly/pFMkl*.

3 Deloitte Insights, "State of AI in the Enterprise, 2nd Edition" 2018. *https://oreil.ly/EiD6T*.

4 Ben Lorica and Paco Nathan, *The State of Machine Learning Adoption in the Enterprise* (Sebastopol: O'Reilly, 2018).

Europe, require a legal basis to use personal data for a secondary purpose. An example of that legal basis would be additional consent or authorization from individuals before their data can be used. In many cases this is not practical and can introduce bias into the data because consenters and nonconsenters differ on important characteristics.[5]

Given the difficulty of accessing data, sometimes analysts try to just use open source or public datasets. These can be a good starting point, but they lack diversity and are often not well matched to the problems that the models are intended to solve. Furthermore, open data may lack sufficient heterogeneity for robust training of models. For example, open data may not capture rare cases well enough.

Data synthesis can give the analyst, rather efficiently and at scale, realistic data to work with. Synthetic data would not be considered identifiable personal data. Therefore, privacy regulations would not apply and additional consent to use the data for secondary purposes would not be necessary.[6]

Enabling Better Analytics

A use case where synthesis can be applied is when real data does not exist—for example, if the analyst is trying to model something completely new, and the creation or collection of a real dataset from scratch would be cost-prohibitive or impractical. Synthesized data can also cover edge or rare cases that are difficult, impractical, or unethical to collect in the real world.

Sometimes real data exists but is not labeled. Labeling a large amount of examples for supervised learning tasks can be time-consuming, and manual labeling is error-prone. Again, synthetic labeled data can be generated to accelerate model development. The synthesis process can ensure high accuracy in the labeling.

Analysts can use the synthetic data models to validate their assumptions and demonstrate the kind of results that can be obtained with their models. In this way the synthetic data can be used in an exploratory manner. Knowing that they have interesting and useful results, the analysts can then go through the more complex process of getting the real data (either raw or de-identified) to build the final versions of their models.

For example, if an analyst is a researcher, they can use their exploratory models on synthetic data to then apply for funding to get access to the real data, which may require a full protocol and multiple levels of approvals. In such an instance, efforts

5 Khaled El Emam et al., "A Review of Evidence on Consent Bias in Research," *The American Journal of Bioethics* 13, no. 4 (2013): 42–44.

6 Other governance mechanisms would generally be needed, and we cover these later in the book.

with the synthetic data that do not produce good models or actionable results would still be beneficial, because they will redirect the researchers to try something else, rather than trying to access the real data for a potentially futile analysis.

Another scenario in which synthetic data can be valuable is when the synthetic data is used to train an initial model before the real data is accessible. Then when the analyst gets the real data, they can use the trained model as a starting point for training with the real data. This can significantly expedite the convergence of the real data model (hence reducing compute time) and can potentially result in a more accurate model. This is an example of using synthetic data for transfer learning.

The benefits of synthetic data can be dramatic—it can make impossible projects doable, significantly accelerate AIML initiatives, or result in material improvement in the outcomes of AIML projects.

Synthetic Data as a Proxy

If the utility of the synthetic data is high enough, analysts are able to get results with the synthetic data that are similar to what they would have with the real data. In such a case, the synthetic data plays the role of a proxy for the real data. Increasingly, there are more use cases where this scenario is playing out: as synthesis methods improve over time, this proxy outcome is going to become more common.

We have seen that synthetic data can play a key role in solving a series of practical problems. One of the critical factors for the adoption of data synthesis, however, is trust in the generated data. It has long been recognized that high data utility will be needed for the broad adoption of data synthesis methods.[7] This is the topic we turn to next.

Learning to Trust Synthetic Data

Initial interest in synthetic data started in the early 1990s with proposals to use multiple imputation methods to generate synthetic data. *Imputation* in general is the class of methods used to deal with missing data by using realistic data to replace the missing values. Missing data can occur, for example, in a survey in which some respondents do not complete a questionnaire.

Accurate imputed data requires the analyst to build a model of the phenomenon of interest using the available data and then use that model to estimate what the imputed value should be. To build a valid model the analyst needs to know how the data will eventually be used.

---

7 Jerome P. Reiter, "New Approaches to Data Dissemination: A Glimpse into the Future (?)," *CHANCE* 17, no. 3 (June 2004): 11–15.

With multiple imputation you create multiple imputed values to capture the uncertainty in these estimated values. This results in multiple imputed datasets. There are specific techniques that can be used to combine the analysis that is repeated in each imputed dataset to get a final set of analysis results. This process can work reasonably well if you know in advance how the data will be used.

In the context of using imputation for data synthesis, the real data is augmented with synthetic data using the same type of imputation techniques. In such a case, the real data is used to build an imputation model that is then used to synthesize new data.

The challenge is that if your imputation models are different than the eventual models that will be built with the synthetic data, then the imputed values may not be very reflective of the real values, and this will introduce errors in the data. This risk of building the wrong model has led to historic caution in the application of synthetic data.

More recently, statistical machine learning models have been used for data synthesis. The advantage of these models is that they can capture the distributions and complex relationships among the variables quite well. In effect, they discover the underlying model in the data rather than requiring that model to be prespecified by the analyst. And now with deep learning data synthesis, these models can be quite accurate because they can capture much of the signal in the data—even subtle signals.

Therefore, we are getting closer to the point where the generative models available today produce datasets that are becoming quite good proxies for real data. But there are also ways to assess the utility of synthetic data more objectively.

For example, we can compare the analysis results from synthetic data with the analysis results from the real data. If we do not know what analysis will be performed on the synthetic data, then a range of possible analyses can be tried based on known uses of that data. Or an "all models" evaluation can be performed, in which all possible models are built from the real and synthetic datasets and compared.

Synthetic data can also be used to increase the heterogeneity of a training dataset to result in a more robust AIML model. For example, edge cases in which data does not exist or is difficult to collect can be synthesized and included in the training dataset. In that case, the utility of the synthetic data is measured in the robustness increment to the AIML models.

The US Census Bureau has, at the time of writing, decided to leverage synthetic data for one of the most heavily used public datasets, the 2020 decennial census data. For its tabular data disseminations, it will create a synthetic dataset from the collected individual-level census data and then produce the public tabulations from that

synthetic dataset. A mixture of formal and nonformal methods will be used in the synthesis process.[8]

This, arguably, demonstrates the large-scale adoption of data synthesis for one of the most critical and heavily used datasets available today.

Beyond the census, data synthesis is being used in a number of industries, as we illustrate later in this chapter.

Synthetic Data Case Studies

While the technical concepts behind the generation of synthetic data have been around for a few decades, their practical use has picked up only recently. One reason is that this type of data solves some challenging problems that were quite hard to solve before, or solves them in a more cost-effective way. All of these problems pertain to data access: sometimes it is just hard to get access to real data.

This section presents a few application examples from various industries. These examples are not intended to be exhaustive but rather to be illustrative. Also, the same problem may exist in multiple industries (for example, getting realistic data for software testing is a common problem that data synthesis can solve), so the applications of synthetic data to solve that problem will therefore be relevant in these multiple industries. Because we discuss software testing, say, only under one heading does not mean that it would not be relevant in another.

The first industry that we examine is manufacturing and distribution. We then give examples from healthcare, financial services, and transportation. The industry examples span the types of synthetic data we've discussed, from generating structured data from real individual-level and aggregate data, to using simulation engines to generate large volumes of synthetic data.

8 Aref N. Dajani et al., "The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau" (paper presented at the Census Scientific Advisory Committee meeting, Suitland, MD, March 2017).

Manufacturing and Distribution

The use of AIML in industrial robots, coupled with improved sensor technology, is further enabling factory automation for more complex and varied tasks.[9] In the warehouse and on the factory floor, these systems are increasingly able to pick up arbitrary objects off shelves and conveyor belts, and then inspect, manipulate, and move them, as illustrated by the Amazon Picking Challenge.[10]

However, robust training of robots to perform complex tasks in the production line or warehouse can be challenging because of the need to obtain realistic training data covering multiple anticipated scenarios, as well as uncommon ones that are rarely seen in practice but are still plausible. For example, recognizing objects under different lighting conditions, with different textures, and in various positions requires training data that captures the variety and combinations of these situations. It is not trivial to generate such a training dataset.

Let's consider an illustrative example of how data synthesis can be used to train a robot to perform a complex task that requires a large dataset for training. Engineers at NVIDIA were trying to train a robot to play dominoes using a deep learning model (see Figure 1-2). The training needed a large number of heterogeneous images that capture the spectrum of situations that a robot may encounter in practice. Such a training dataset did not exist, and it would have been cost-prohibitive and very time-consuming to manually create these images.

9 Jonathan Tilley, "Automation, Robotics, and the Factory of the Future," McKinsey, September 2017. *https://oreil.ly/L27Ol*.

10 Lori Cameron, "Deep Learning: Our No. 1 Tech Trend for 2018 Is Set to Revolutionize Industrial Robotics," IEEE Computer Society, accessed July 28, 2019. *https://oreil.ly/dKcF7*.

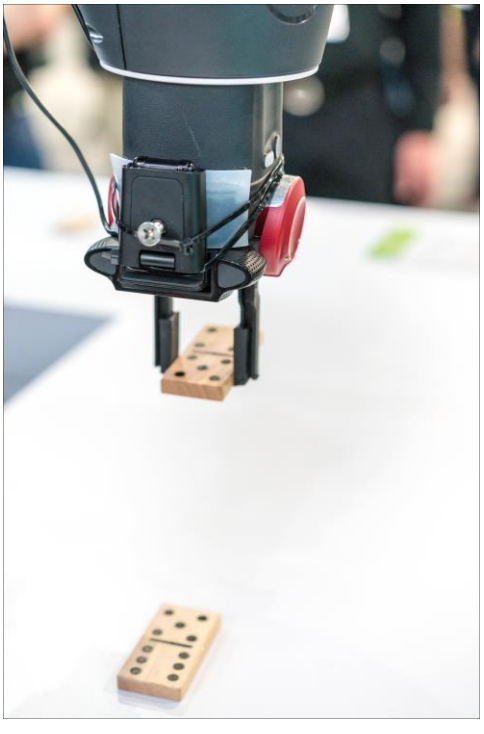

Figure 1-2. The dominoes-playing robot

The NVIDIA team used a graphics-rendering engine from its gaming platform to create images of dominoes in different positions, with different textures, and under different lighting conditions (see Figure 1-3).[11] No one actually manually set up dominoes and took pictures of them to train the model—the images that were created for training were simulated by the engine.

11 Rev Lebaredian, "Synthetic Data Will Drive Next Wave of Business Applications" (lecture, GTC Silicon Valley, 2019). *https://bit.ly/2yUefyl*.

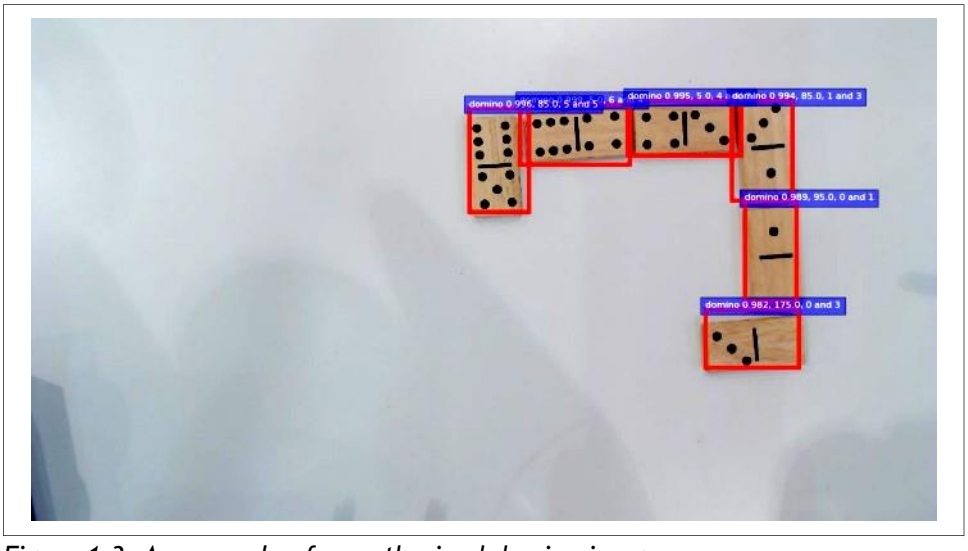

Figure 1-3. An example of a synthesized domino image

In this case the image data did not exist, and creating a large enough dataset manually would have taken a lot of people a long time—not a very cost-effective option. The team used the simulation engine to create a large number of images to train the robot. This is a good example of how synthetic data can be used to train a robot to recognize, pick up, and manipulate objects in a heterogeneous environment—the same type of model building that would be needed for industrial robots.

Healthcare

Getting access to data for building AIML models in the health industry is often difficult because of privacy regulations or because the data collection can be expensive. Health data is considered sensitive in many data-protection regimes, and its use and disclosure for analytics purposes must meet a number of conditions. These conditions can be nontrivial to put in place (e.g., by providing patients access to their own data, creating strong security controls around the retention and processing of the data, and training staff).[12] Also, the collection of health data for specific studies or analyses can be quite expensive. For instance, the collection of data from multiple sites in clinical trials is costly.

The following examples illustrate how synthetic data has solved the data-access challenge in the healthindustry.

12 Mike Hintze and Khaled El Emam, "Comparing the Benefits of Pseudonymisation and Anonymisation under the GDPR," *Journal of Data Protection and Privacy* 2, no. 1 (December 2018): 145–58.

Data for cancer research

There are strong currents pushing governments and the pharmaceutical industry to make their health data more broadly available for secondary analysis. This is intended to solve the data-access problem and encourage more innovative research to understand diseases and find treatments. Regulators have also required companies to make health data more broadly available. A good example of this is the European Medicines Agency, which has required pharmaceutical companies to make the information that they submitted for their drug approval decisions publicly available.[13] Health Canada has also recently done so.[14]

Medical journals are also now strongly encouraging researchers who publish articles to make their data publicly available for other researchers to replicate the studies, which could possibly lead to innovative analyses on that same data.

In general, when that data contains personal information, it needs to be de-identified or made nonpersonal before it is made public (unless consent is obtained from the affected individuals beforehand, which is not the case here). However, in practice it is difficult to de-identify complex data for a public release.[15] There are a number of reasons for this:

- Public data has few controls on it (e.g., the data users do not need to agree to terms of use and do not need to reveal their identities, which makes it difficult to ensure that they are handling it securely). Therefore, the level of data transformations needed to ensure that the risk of re-identification is low can be extensive, which ensures that data utility has degraded significantly.
- Re-identification attacks on public data are getting more attention by the media and regulators, and they are also getting more sophisticated. As a consequence, de-identification methods need to err on the conservative side, which further erodes data utility.
- The complexity of datasets that need to be shared further amplifies the data utility problems because a lot of the information in the data would need to be transformed to manage the re-identification risk.

Synthetic data makes it feasible to have complex open data. Complexity here means that the data has many variables and tables, with many transactions per individual.

---

13 European Medicines Agency, "External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use," September 2017. *https://oreil.ly/uVOna*.

14 Health Canada, "Guidance Document on Public Release of Clinical Information," April 1, 2019. *https://bit.ly/33JzHnY*.

15 Khaled El Emam, "A De-identification Protocol for Open Data," IAPP Privacy Tech, May 16, 2016. *https://bit.ly/33AetZq*.

For example, data from an oncology electronic medical record would be considered complex. It would have information about, for instance, the patient, visits, treatments, drugs prescribed and administered, and laboratory tests.

Synthesis can simultaneously address the privacy problem and provide data that is of higher utility than the incumbent alternative. A good example of this is the synthetic cancer registry data that has been made publicly available by Public Health England. This synthetic cancer dataset is available for download and can be used to generate and test hypotheses, and to do cost-effective and rapid feasibility evaluations for future cancer studies.

Beyond data for research, there is a digital revolution (slowly) happening in medicine.[16] For example, the large amounts of health data that exist with providers and payers contain many insights that can be detected by the more powerful AIML techniques. New digital medical devices are adding more continuous data about patient health and behavior. Patient-reported outcome data provides assessments of function, quality of life, and pain. And of course genomic and other -omic data is at the core of personalized medicine. All this data needs to be integrated into and used for point-of-care and at-home decisions and treatments. Innovations in AIML can be a facilitator of that.

In the next section we examine how digital health and health technology companies can use synthetic data to tap into this innovation ecosystem. And note that more traditional drug and device companies are becoming digital health companies.

Evaluating innovative digital health technologies

Health technology companies are constantly looking for data-driven innovations coming from the outside. These can be innovations from start-up companies or from academic institutions. Typical examples include data analysis (statistical machine learning or deep learning models and tools), data wrangling (such as data standardization and harmonization tools, and data cleansing tools), and data type detection tools (that find out where different types of data exist in the organization).

Because adopting new technologies takes resources and has opportunity costs, the decision to do so must be made somewhat carefully. These companies need a mechanism to evaluate these innovations in an efficient way to determine which ones really work in practice, and, more importantly, which ones will work with their data. The best way to do that is to give these innovators some data and have them demonstrate their wares on that data.

16 Neal Batra, Steve Davis, and David Betts, "The Future of Health," Deloitte Insights, April 30, 2019. *https://oreil.ly/4v_nY*.

Some large companies get approached by innovators at a significant pace—sometimes multiple parts of an organization are approached at the same time. The pitches are compelling, and the potential benefits to their business can be significant. The large companies want to bring these innovations into their organizations. But experience has told them that, for instance, some of the start-ups are pitching ideas rather than mature products, and the academics are describing solutions that worked only on small problems or in situations unlike the companies'. There is a need to test these innovations on their own problems and data.

In the pharmaceutical industry, it can be complex to provide data to external parties because much of the relevant data pertains to patients or healthcare providers. The processes that would be needed to share that data would usually include extensive contracting and an audit of the security practices at the data recipient. Just these two tasks could take quite some time and investment.

Sometimes the pharmaceutical company is unable to share its data externally because of this complexity or because of internal policies, and in that case it asks the innova- tor to come in and install the software in its environment (see "Rapid Technology Evaluation" for an example). This creates significant complexity and delays because now the company needs to audit the software, address compatibility issues, and figure out integration points. This makes technology evaluations quite expensive and uses up a lot of internal resources. Plus, this is not scalable to the (potentially) hundreds of innovations that the company would want to test every year.

These companies have started to do two things to make this process more efficient and to enable them to bring innovations in. First, they have a standard set of synthetic datasets that are representative of their patient or provider data. For example, a pharmaceutical company would have a set of synthetic clinical trial datasets in various therapeutic areas. These datasets can be readily shared with innovators for pilots or quick proof-of-concept projects.

Rapid Technology Evaluation

Cambridge Semantics (CS), a Boston company developing a graph database and various analytics tools on top of that, was planning to do a pilot with a large prospect in the health space to demonstrate how its tools can be used to harmonize pooled clinical trial data. To do this pilot, it needed to get data from the prospect. That way CS could demonstrate that its tools worked on real data that was relevant for the prospect—there are few things more compelling than seeing a problem solved in an elegant way on your own data.

The initial challenge was that to get data from the prospect, CS would need to go through an audit to ensure that it had adequate security and privacy practices to handle personal health information. That process would have taken three to four months to complete.

An alternative that was considered was for CS to install its software on the prospect's private cloud and then run it there using real data. However, the complexities of introducing new software into a regulated computing environment are not trivial. Furthermore, giving CS staff access to the internal computing environment would have required additional checks and processes. This also would have taken three to four months.

The team landed on a synthetic data solution whereby a number of synthetic datasets were created and given to CS to demonstrate how it would solve the specific problem. The pilot was completed in a few days.

The second process that is used is competitions. The basic idea is to define a problem that needs to be solved and then invite a number of innovators to solve that problem, using synthetic data to demonstrate their solutions. These can be open or closed competitions. With the former, any start-up, individual, or institution can participate, such as by organizing public hackathons or datathons. With the latter, closed competitions, specific innovators are invited to participate.

With public hackathons or datathons, entrants are invited to solve a given problem with a prize at the end for the winning individual or team. The main difference between such public events and the competitions described previously is that the innovators are not selected in advance; rather, participation tends to be more open. The diversity in these competitions means that many new ideas are generated and evaluated in a relatively short period of time. Synthetic data can be a key enabler under these circumstances by providing datasets that the entrants can access with minimal constraints.

A good example of an open competition is the Heritage Health Prize (HHP). The HHP was notable for the size of the prize and the size of the dataset that was made available to entrants. At the time of the competition, which lasted from 2011 to 2013, the availability of synthetic data was limited, and therefore a de-identified dataset was created.[17] Because of the challenges of de-identifying open datasets that were noted earlier, it has been more common for health-related competitions to be closed. However, at this point in time there is no compelling reason to maintain that restriction. Synthetic data is now being used to enable such competitions as described in "Datathons Enabled by Synthetic Data."

In practice, only a small percentage of those evaluations succeed when given a realistic dataset to work with. The innovators that make it through the evaluation or competition are then invited to go through the more involved process to get access to real

17 Khaled El Emam et al., "De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset," *Journal of Medical Internet Research* 14, no. 1 (February 2012): e33. *https://www.jmir.org/2012/1/e33*.

data and do more detailed demonstrations, or the company may decide to license the innovation at that point. But at least the more costly investments in the technology evaluation or adoption are performed only on candidates that are known to have an innovation that works.

Datathons Enabled by Synthetic Data

The Vivli-Microsoft Data Challenge was held in June 2019 in Boston. The goal of the competition was to propose innovative methods to facilitate the sharing of rare disease datasets, in a manner that maintains the analytic value of the data while safeguarding participant privacy. Rare disease datasets are particularly difficult to share while maintaining participant privacy because they often contain relatively few individuals, and individuals may be identified using only a handful of attributes.

This event gathered 60 participants on 11 teams from universities, hospitals, and pharmaceutical, biotech, and software companies. Each team had five hours to plan and propose a solution, then five minutes to present the solution to the judges. The solutions combined new and existing technologies in interesting ways that were tailored for use in rare disease datasets. Unsurprisingly, the winning team proposed a solution built around the use of synthetic data.

Synthetic data was critical to this event's success as it allowed all participants to "get their hands dirty" with realistic clinical trial data, without needing to use costly secure computational environments or other control mechanisms. The synthetic data grounded the competition in reality by providing participants with example data that their solutions would need to be able to accommodate. Groups that built demos of their solutions were also able to apply their methods to the synthetic data as a proof of concept.

Data challenges like this depend on providing high-quality data to participants, and synthetic data is a practical means to do so.

Another large consumer of synthetic data is the financial services industry. Part of the reason is that this industry has been an early user of AIML technology and data-driven decision making, such as in fraud detection, claims processing, and consumer marketing. In this next section we examine specific use cases in which synthetic data has been applied in this sector.

Financial Services

Getting access to large volumes of historical market data in the financial services industry can be expensive. This type of data is needed, for example, for building models to drive trading decisions and for software testing. Also, using consumer financial transaction data for model building, say, in the context of marketing retail

banking services, is not always easy because that requires the sharing of personal financial information with internal and external data analysts.

The following use cases illustrate how synthetic data has been used to solve some of these challenges.

Synthetic data benchmarks

When selecting software and hardware to process large volumes of data, financial services companies need to evaluate vendors and solutions in the market. Instead of having each company evaluate technologies from innovative vendors and academics one by one, it is common to create standardized data benchmarks.

A data benchmark would consist of a dataset and a set of tests that would be performed on that dataset. Vendors and academics can then use their software and hardware to produce the outputs using these data as inputs, and they can all be compared in a consistent manner. Creating a benchmark would make the most sense in situations where the market is large enough and the community can agree on a benchmark that is representative.

In competitive scenarios where multiple vendors and academics can supply solutions to the same set of problems, the benchmarks must be constructed in a manner that ensures that no one can easily game the system. With a standard input dataset, the solutions can just be trained or configured to produce the correct output without performing the necessary analytic computations.

Synthetic data benchmarks are produced from the same underlying model, but each vendor or academic gets a unique and specific set of synthetic data generated from that model. In that way, each entity running the benchmark will need to produce different results to score well on the benchmark.

An example is the STAC-A2 benchmark for evaluating software and hardware used to model financial market risk. The benchmark has a number of quality measures in the output that are assessed during the computation of option price sensitivities for multiple assets using Monte Carlo simulation. There is also a series of performance/scaling tests that are performed using the data.

When financial services companies wish to select a technology vendor, they can compare the solutions on the market using a consistent benchmark that was executed on comparable data. This provides a neutral assessment of the strengths and weaknesses of available offerings without the companies having to perform their own evaluations (which can be expensive and time-consuming) or relying on vendor-specific assess- ments (which may be biased toward that vendor).

Software testing

Software testing is a classic use case for synthetic data. This includes functional and performance testing of software applications by the software developers. In some cases large datasets are needed to benchmark software applications to ensure that they can perform at certain throughputs or with certain volumes. Extensions of the testing use case are datasets for running software demos by a sales team, and for training users of software on realistic data.

Software testing is common across many industries, and the problems being addressed with synthetic data will be the same. In the financial services sector there are two common use cases. The first is to test internal software applications (e.g., fraud detection) to ensure that they perform the intended functions and do not have bugs. For this testing, realistic input data is needed, and this includes data covering edge cases or unusual combinations of inputs. The second is to test that these applications can scale their performance (for example, response times in automated trading applications are important) to handle the large volumes of data that are likely to be met in practice. This testing must also simulate unusual situations—for example, when trading volumes spike due to an external political or environmental event.

In most software engineering groups, it is not easy to obtain production data. This may be because of privacy concerns or because the data contains confidential business information. Therefore, there is reluctance to make that data available to a large group of software developers. The same applies to making data available for demos and for training purposes. Furthermore, in some cases the software is new and there is insufficient customer data to use for testing.

One alternative that has been used is to de-identify the production data before making it available to the test teams. Because the need for test data is continuous, the de-identification must also be performed on a continuous basis. The cost-effectiveness of continuous de-identification versus that of synthetic data would have to be considered. However, a more fundamental issue is the level of controls that would need to be in place for the software developers to work with the de-identified data. As will be noted later on, re-identification risk is managed by a mix of data transformation and security and privacy controls. Software development groups are accustomed to working with lower levels of these controls.

The data utility demands for software testing are not as high as they are for some of the other use cases that we have looked at. It is possible to generate synthetic data from theoretical distributions and then use them for testing. Another approach that has been applied is to use public datasets (open data) and replicate those multiple times to create larger test datasets or resample with replacement (draw samples from the dataset so that each record can be drawn more than once).

There are more principled methods for the generation of synthetic data for testing, demos, and training. These involve the generation of synthetic data from real data

using the same approaches that are used to generate data for building and testing AIML models. This will ensure that the data is realistic and has correct statistical characteristics (e.g., a rare event in the real data will also be a rare event in the synthetic data), and that these properties are maintained if large synthetic datasets are generated.

The next industry that we will consider is transportation. Under that heading we will consider data synthesis for planning purposes through microsimulation models and data synthesis for training models in autonomous vehicles.

Transportation

The use of synthetic data in the transportation industry goes back a few decades. The main driver is the need to make very specific planning and policy decisions about infrastructure in a data-limited environment. Hence the use of microsimulation models became important to inform decision making. This is the first example we consider. The second example is the use of gaming engines to synthesize virtual environments that are used to train AIML models, which are then embedded in the autonomous vehicles.

Microsimulation models

Microsimulation environments allow users to do "what-if" analyses and run novel scenarios. These simulation environments become attractive when there is no real data available at all, and therefore synthetic data needs to be created.

In the area of transportation planning it is, for example, necessary to evaluate the impact of planned new infrastructure, such as a new bridge or a new mall. Activity-based travel demand models can use synthetic data to allow planners to do that.

A commonly used approach to creating synthetic data for these models combines aggregate summaries—for example, from the census, with sample individual-level data that is collected from surveys. Census data would normally provide information like household composition, income, and number of children. The aggregate data would normally cover the whole population of interest but may not have all the needed variables and not to the level of granularity that is desired. The survey data will cover a sample of the population but have very detailed and extensive variables.

Synthetic reconstruction then uses an iterative process such as iterative proportional fitting (IPF) to create synthetic individual-level data that plausibly generates the aggregate summaries and uses the sample data as the seed. The IPF procedure was developed some time ago and has more recently been applied to the data synthesis

problem.[18,19] IPF has some known disadvantages in the context of synthesis—for example, when the survey data does not cover rare situations. More robust techniques, such as combinatorial optimization, have been developed to address them.[20]

The next step is to use other data, also collected through surveys or directly from individuals' cell phones, characterizing their behaviors and movements. This data is used to build models, such as the factors that influence an individual's choice of mode of transportation.

By combining the synthetic data with the models, one can run microsimulations of what would happen under different scenarios. Note that the models can be cascaded in the simulation describing a series of complex behaviors and outcomes. For example, the models can inform decisions concerning the impact on traffic, public transportation usage, bicycle trips, and car usage caused by the construction of a new bridge or a new mall in a particular location. These microsimulators can be validated to some extent by ensuring that they give outputs that are consistent with reality under known historical scenarios. But they can also be used to simulate novel scenarios to inform planning and policy making.

Let's now consider a very different use case for synthetic data in the context of developing AIML models for autonomous vehicles. Some of these models need to make decisions in real time and can have significant safety impacts. Therefore, the robustness of their training is critical.

Data synthesis for autonomous vehicles

One of the key functions on an autonomous vehicle is object identification. This means that the analysis of sensor data needs to recognize the objects in the vehicle's path and surroundings. Cameras, lidar systems, and radar systems provide the data feeds to support object identification, as well as speed and distance determination of these objects.

Synthetic data is essential to train the AIML models that process some of these signals. Real-world data cannot capture every edge case, or rare or dangerous scenario— such as an animal darting into the vehicle's path or direct sunlight shining into a camera sensor—that an autonomous vehicle could encounter. Additionally, the cap-

---

18 W. Edwards Deming and Frederick F. Stephan, "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," *Annals of Mathematical Statistics* 11, no. 4 (1940): 427–44.

19 Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay, "Creating Synthetic Baseline Populations," *Transportation Research Part A* 30, no. 6 (1996): 415–29.

20 Zengyi Huang and Paul Williamson, "A Comparison of Synthetic Reconstruction and Combinatorial Optimization Approaches to the Creation of Small-Area Micro Data" (working paper, University of Liverpool, 2002); Justin Ryan, Hannah Maoh, and Pavlos Kanaroglou, "Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms," *Geographical Analysis* 41 (2009): 181–203.

tured environment is fixed and cannot respond to changes in the system's behavior when it is run through the scenario multiple times.

The only way to address these gaps is to leverage synthetic data. By generating customizable scenarios, engineers can model real-world environments—and create entirely new ones—that can change and respond to different behaviors. While real-world tests provide a valuable tool for validation, they are not nearly exhaustive enough to prove that a vehicle is capable of driving without a human at the wheel.

The synthetic data used in simulation is generated using gaming technology from video games or other virtual worlds. First, the environment must be created. It can either replicate a location in the real world, like New York City, using actual data, or be an entirely synthetic place. In either case, everything in the environment must accurately simulate the same material properties as the real world—for example, the reflection of light off of metal or the surface of asphalt.

This level of fidelity makes it possible to accurately re-create how a car sees the environment it is driving in, simulating the output from camera, radar, and lidar sensors. The processors on the car then receive the data as if it is coming from a real-world driving environment, make decisions, and send vehicle control commands back to the simulator. This closed-loop process enables bit-accurate, timing-accurate hardware-in-the-loop testing. It also enables testing of the functions of the vehicle under very realistic conditions.

Of course, the computing capacity needed to perform hardware-in-the-loop testing can be quite significant: achieving the fidelity necessary for autonomous vehicle validation is incredibly compute-intensive. First, a detailed world has to be generated. Then the sensor output must be simulated in a physically accurate way—which takes time and massive amounts of compute horsepower.

Summary

Over the last few years we have seen the adoption of synthetic data grow in various industries, such as manufacturing, healthcare, transportation, and financial services. Because data-access challenges are not likely to get any easier or go away anytime soon, the applicability of data synthesis to more use cases is expected to grow.

In this chapter we started with an overview of what synthetic data is and discussed its benefits. We then looked at a number of industries where we have seen how synthetic data can be applied in practice to solve data-access problems. Again, a characteristic of these use cases is their heterogeneity and the plethora of problems that synthesis can solve. Ours is not a comprehensive list of industries and applications, but it does highlight what early users are doing and illustrate the potential.

The examples we gave in this chapter cover multiple types of data. Our focus in this book is on structured data. Many of the concepts we will cover, however, are generally applicable to other types of data as well. In the next chapter we cover important implementation considerations, starting with ensuring that data synthesis is aligned with your organization's priorities. This is followed by a description of the synthesis process and deploying synthesis pipelines. We close with programmatic considerations as you scale data synthesis within the enterprise.