gretel

eBook

How to Unlock Your Data Projects with Synthetic Data

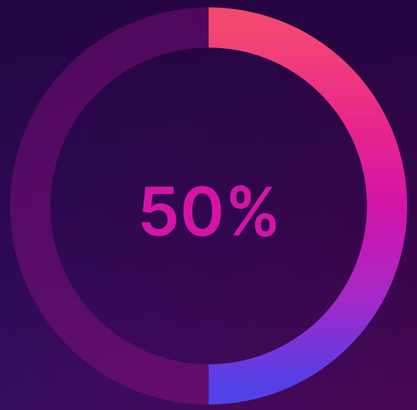
Contents

Introduction

If you've ever embarked on a data project and ended up stuck, you're not alone. According to McKinsey and other reports, only 15%-20% of businesses' ML projects get shipped, while the rest succumb to operational friction and never make it off the ground.[1,2]

This can stem from a variety of causes, but the primary culprit is often the data itself. A Kaggle survey[3] of over 16,000 developers found that gathering, cleaning, and productizing data takes more than 50% of an average project's time. That's a huge amount of precious time, but these are critical steps that cannot be skipped. That's also assuming the data is actually sufficient for the project. If there's not enough of it, it is unusable, or isn't up to quality standards, projects end up at a standstill.

Data is the lifeblood of every organization. It is a critical component of every business operation, transformation, new product development, and AI/ML project. But it's also prone to problems. Read on to learn how you can solve the most common ones.



16,000 developers found that gathering, cleaning, and productizing data takes **more than 50% of an average project's time.**

- Kaggle

3 Common Problems When
Building Data-driven Products

Data on-demand

Challenge

Limited or insufficient data

Many projects are stopped before they can even begin if there's no data to work with in the first place. Often dubbed "the cold start problem," this is usually the issue for startups and newer businesses.

Many companies have an idea for a new product or capability, but lack the necessary data to bring it to life. It's a catch-22: they need to collect a high volume of customer data in order to develop their product or service, but cannot collect that data until the business grows. This leaves the business unable to move forward.

Established businesses may also find themselves stuck at the starting line. Existing products or services and a healthy database aren't always enough to expand into new geographies or industry sectors. Often, they lack the necessary data for estimating ROI, adapting their offering, or furthering progress.

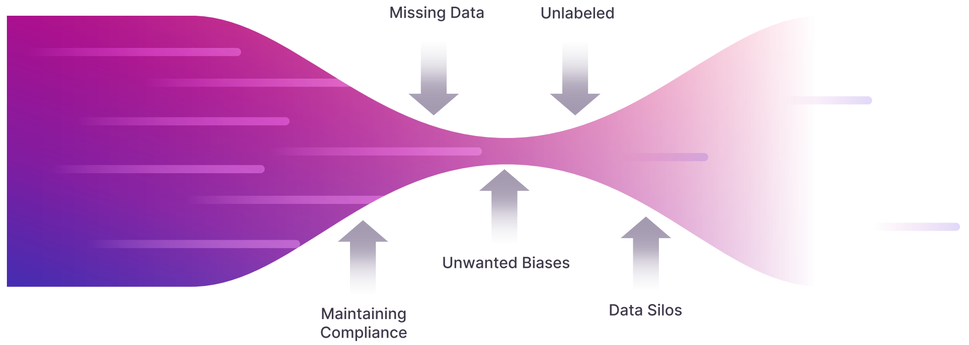


Figure 1: This image showcases the various data bottlenecks faced by organizations looking to operationalize data.

Solution

From no data to unlimited data on-demand

If your business needs a high volume of data to grow and scale but cannot obtain that data until it grows, synthetic data is your best option for getting started.

Relevant synthetic data can be generated on-demand using simple natural language prompts or schema definitions. This data is not randomly generated filler. It's **high-quality, context-specific, and informed by existing comparable datasets.** As a result, it empowers new businesses to get off the ground and expand into new markets quicker and easier.

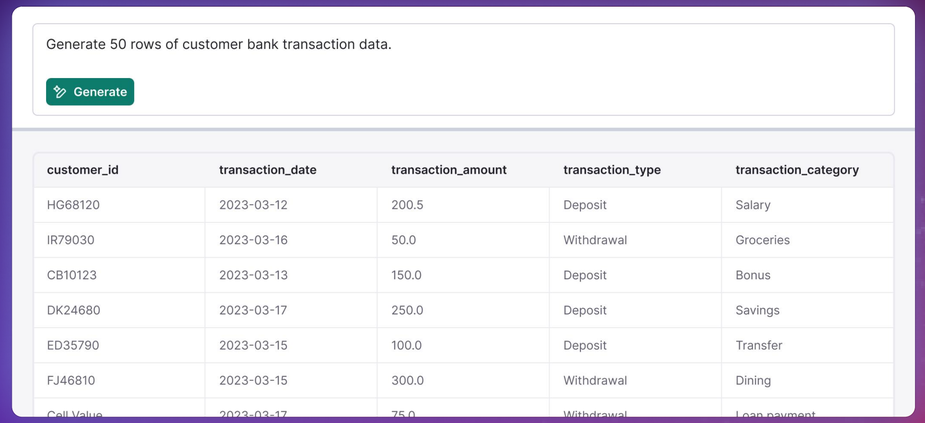
Solving the "cold-start" problem in seconds

A model can't make inferences from historic data if it did not have quality training data to begin with — synthetic data helps you generate data from scratch or fill in missing gaps.

Generate high-quality, domain-specific data in seconds

- Zero-shot prompting
- Generate from SQL schema
- Augment existing datasets

Generate data today →



3 Common Problems When
Building Data-driven Products

Data privacy

Challenge

Privacy and compliance

While other businesses will have all the data they need, they may be stuck at a standstill because their data is far too sensitive to actually use.

Enterprise data may contain proprietary information about the business, its products, trade secrets, and its customers that, if compromised, could cause reputational damage and potential harm. To minimize this risk, some organization's internal data privacy policies limit data sharing between business units, teams, and geographical locations. While this might safeguard information, it also creates silos and slowdowns, with some businesses stuck in months-long approval processes just to share data with their coworkers.

Externally, regulatory compliance can create a massive bottleneck. Legislation including GDPR, CCPA/CPRA, the EU AI Act, and more are increasingly limiting what data businesses can collect and use. In heavily regulated industries such as financial services and healthcare, there are even stricter requirements under the SEC's reporting mandates, HIPAA, and more. Violations of these regulations can expose businesses to punitive action and hefty fines.

Many organizations get caught up in this regulatory red tape and let the risk of non-compliance get in the way of continuing their data projects.

Solution

Privacy and compliance you can trust

Changing names in datasets is not a thorough enough method of de-identification to protect your customers' personal information. If you remove or change the name "John Doe" everywhere it appears in the data set as well as associated attributes like phone number, address, and other sensitive details, the data can still be vulnerable to reidentification by a simple detail like the timing of a doctor appointment they attended. It's not difficult for a bad actor to cross reference that information with other data sources to determine John's true identity from these seemingly benign details.

Synthetic data lacks the personal identifiers of real-world data. It removes all direct references to any sensitive information and simply mirrors the original data characteristics, and can be designed to be impossible to trace back to any individual customer.

Gretel offers synthetic data that is differentially private. Our proprietary method uses a technique called Differential Privacy Stochastic Gradient Descent (DP-SGD), which adds noise to the optimization process and clips gradients to prevent memorization of any single data example with minimal loss to accuracy. This provides mathematical guarantees that no individual's personal information can be traced or revealed while still allowing the model to learn trends, insights, and distributions from the real world data.

Enhancing the privacy of data helps facilitate data sharing between teams, business units, and business locations, which eliminates some of the silos and shortens the approval processes that slow down projects. And since synthetic data does not carry many of the attributes of real-world data that pique regulators' interest, using it mitigates the risk of non-compliance.

3 Common Problems When
Building Data-driven Products

Data quality

Challenge

Messy or incomplete data

Some businesses may have large datasets that are unusable for another reason entirely: they're a mess.

Real-world data is challenging and costly to collect and maintain, but even more expensive and time consuming to clean and label for use in model development and training. These tasks are also prone to human error. Manual entry errors can lead to datasets with missing fields or blank columns that render the information unusable for downstream applications. These gaps may be easily overlooked or labor-intensive to fix. In other words, the business may undergo all this effort just to receive inaccurate or lackluster results.

There's another issue that can arise: data that's unintentionally biased, that may not take into account underrepresented groups or rare occurrences. Using this data could result in an outcome that is not representative of reality or fair to those who may also be impacted.

Sometimes, a business's project attempts to solve a problem that is entirely new. A limited amount of data on these rare edge cases may exist, but it may not be enough to draw from. In novel cases, generating real-world data for this scenario may be too expensive, impractical, or difficult to obtain.

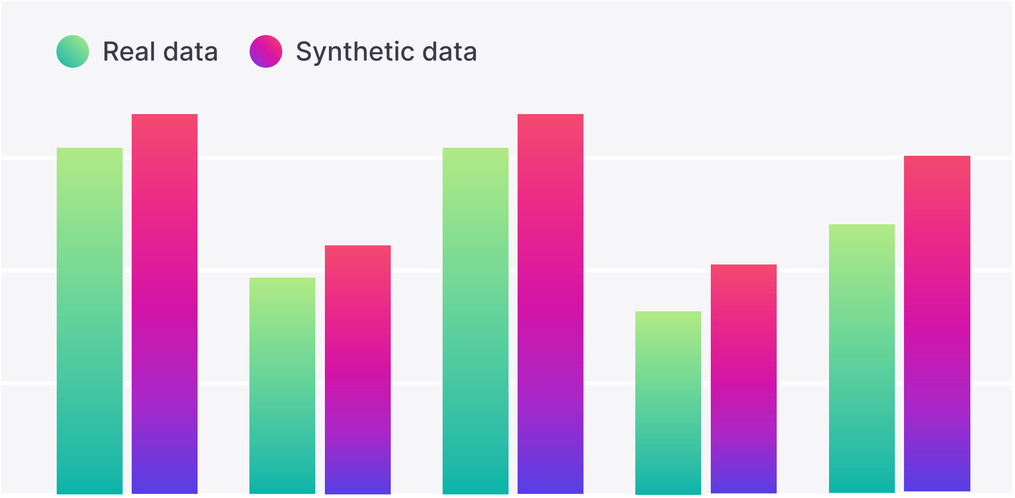


Figure 2: Even with increased privacy protection, models trained with synthetic data perform at similar — or even better than — accuracy levels as those trained with ground truth data.

Solution

Robust, high-quality data

When looking to build up incomplete or lackluster datasets, some may try to upsample or generate data using scripts, stored procedures, or mock service calls to increase volume. While this works in theory, the outcome is repetitive information that may not be entirely accurate, useful, or representative of reality.

Rather than simply duplicating what already exists, **synthetic data can learn the patterns of real-world data and generate new, relevant data to boost volume** based on the existing data as well as what can be found in other real-world environments.

Not only does this boost volume, it closes any gaps in the dataset by filling in the blanks with contextually relevant information. With synthetic data, businesses also can generate additional labeled examples for their training sets at a fraction of the cost. Synthetic data can be created at the speed of development, and automatically comes correctly labeled.

Being able to generate your own data means being able to account for more situational variances and consider more diverse groups, behavioral patterns, and data events. By boosting underrepresented classes to change distributions in the dataset, unwanted biases are removed and the business generates more responsible, equitable, and representative outcomes.

Benefits of using synthetic data to unlock data access:

1. Quickly generate new training data from scratch.
2. Ensure data is use-case and domain-specific.
3. Protect sensitive and private information.

Synthetic data models also enable businesses to test novel, remote, or hypothetical edge cases. Using pre-trained generative models, public data, and natural language prompts, synthetic data models can create simulations that enhance the model's robustness and generalizability across a wider range of potential scenarios. Even the most remote of possibilities can be accounted for, resulting in more thoroughly trained algorithms.

If data is the new oil, then synthetic data is the new renewable energy.

Synthetic data breaks the bottleneck

If data is the new oil, then synthetic data is the new renewable energy. It’s generated from computer simulations or algorithms as a substitute for real-world data and trained using real-world datasets to learn patterns, distributions, and behaviors so that it can generate representative versions that are true to the original. Synthetic data is not “fake” or “made up” data, but because of the nature of how it is created, synthetic data boasts several benefits that overcome the common challenges we discussed.

Synthetic data accelerates access to data in days rather than months so you can develop applications 10x faster and improve model performance by 50% or more.

But not all synthetic data is created the same. When selecting a vendor, it is important to consider privacy, accuracy, and scalability.

There's a good reason why **Gretel is the preferred synthetic data platform for developers who need to simulate real-world scenarios, safely train AI/ML algorithms, and protect customer data while deriving rich insights.** Our platform is composed of powerful enterprise features that make generating synthetic data safely and reliably at scale.

Unlock your data projects with synthetic data and get real results.

~~6 months~~
2 days
faster time to data access

10x
faster application development

50%
improvement in model performance

Validate your synthetic data with comprehensive, customizable reports

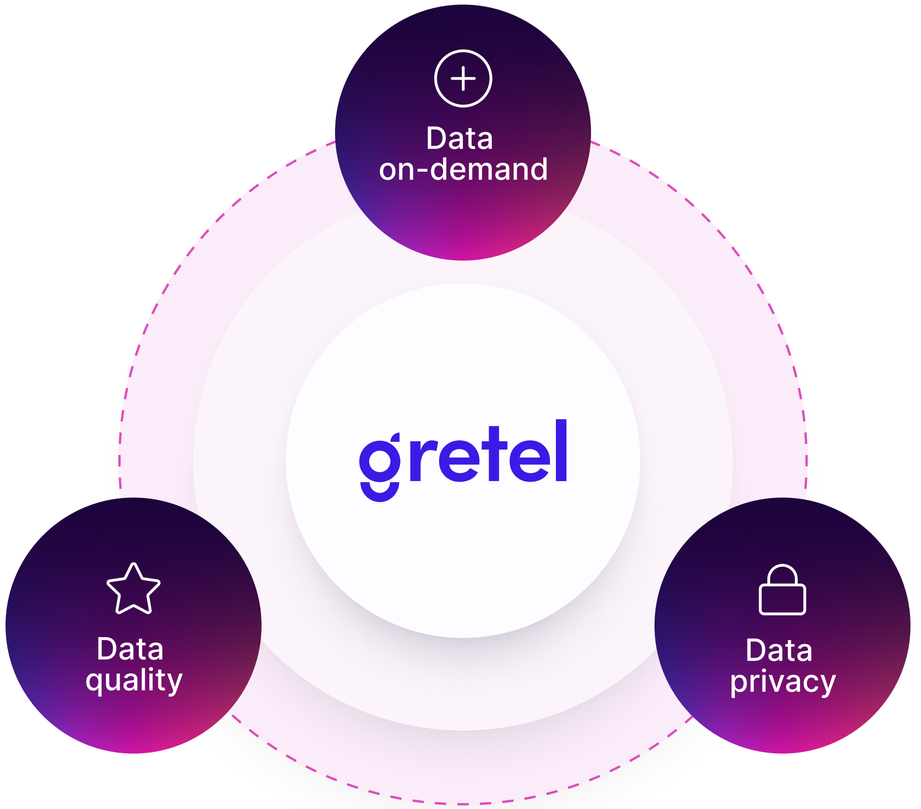
For every synthetic dataset you generate, Gretel provides a detailed summary and score evaluating the privacy and quality of your data, using the original training data as comparison. This can be viewed as a confidence score as to whether scientific conclusions drawn from the synthetic dataset would be the same as those drawn from the original dataset. The result is synthetic data you can trust every time.

Evaluate data performance

Understand how your data performs on key metrics and compare results side by side with Gretel Evaluate.

Explore how →

Accelerate your data-driven ML/AI projects at the speed of modern innovation with synthetic data. To get started, **reach out to our team** to learn how your organization can embark on its synthetic data journey.



About Gretel

Gretel is a multi-modal synthetic data platform for enterprise developers, fueled by advanced generative AI models and privacy-enhancing technologies. Users can generate a diverse array of synthetic data types, including text, relational tables, time series, and tabular.

With customizable privacy and accuracy settings, **Gretel empowers users to create secure, statistically accurate replicas of sensitive data, ideal for safe model training.** The platform also facilitates on-demand data generation to enrich limited machine learning training datasets. Serving a wide range of clients—from top financial institutions, healthcare and life sciences companies, to government and public sector organizations, as well as leading tech brands and startups — Gretel adheres to key privacy regulations such as GDPR, CCPA, and HIPAA. It's the go-to solution for organizations seeking fast, secure, data-driven decision-making, and responsible AI development.

gretel

Learn more at gretel.ai

Resources

1. https://www.mckinsey.com/capabilities/operations/our-insights/operationalizing-machine-learning-in-processes
2. https://hbr.org/2023/11/keep-your-ai-projects-on-track#:~:text=Sadly%2C%20beneath%20the%20aspirational%20headlines,increase%20the%20odds%20of%20success
3. https://www.google.com/url?q=https://www.kaggle.com/datasets/kaggle/kaggle-survey-2017/data&sa=D&source=docs&ust=1713370206741537&usg=AOvVaw1ecgLMO7D1U1agDENNe8Hk