# Explore Data-Centric AI Solutions to Streamline AI Development

Published 9 June 2023 - ID G00782740 - 37 min read

By Analyst(s): Zain Khan

Initiatives: Analytics and Artificial Intelligence for Technical Professionals; Evolve Technology and Process Capabilities to Support D&A

Data and analytics technical professionals must prioritize generating, curating and preparing the right data to build successful AI models. This research provides curated information on data-centric AI solutions that focus on systematically improving data to build reliable and accurate AI systems.

## Overview

### Key Findings

- Data-centric AI is a shift away from a model-centric approach. With data-centric AI, the outcome of AI solutions is driven more by enhancing and enriching the training data, as opposed to tuning the model or the code.

- Time, cost and resources spent on manual data exploration, cleansing, feature extraction, data quality and automation continue to rise as data complexity and lack of accessibility increases. Poor quality data leads to inaccurate ML models with wasted time, effort and resources.

- Modern augmented and self-serve data preparation tools offer powerful capabilities like data lineage, automation, automatic feature extraction and data enhancement to help prepare high quality datasets for AI.

- It is difficult to find data enrichment, AI-assisted annotation, data labeling, data preprocessing, augmentation and no-code features all in a single tool. Differentiating features include data lineage tracking, automated quality checks, predictive transformation impact, 3D and virtual reality-based data exploration.

### Recommendations

Data and analytics technical professionals working on building AI systems should:

- Adopt a data-centric AI approach by using augmented and self-serve data preparation, as well as automated feature engineering tools and techniques to tackle complex data challenges.
- Leverage data enrichment, AI-augmented visual data exploration and enhancement tools to enhance dataset quality.
- Reduce time, cost and efforts spent on manual data labeling and annotation by leveraging state-of-the-art labeling techniques and AI-augmented annotation and data labeling tools.
- Evaluate the financial and technical implications of investing in self-serve tools versus exploring open-source options. Choose a blend that caters to your needs, keeping in mind the data and skill diversity.

## Strategic Planning Assumption(s)

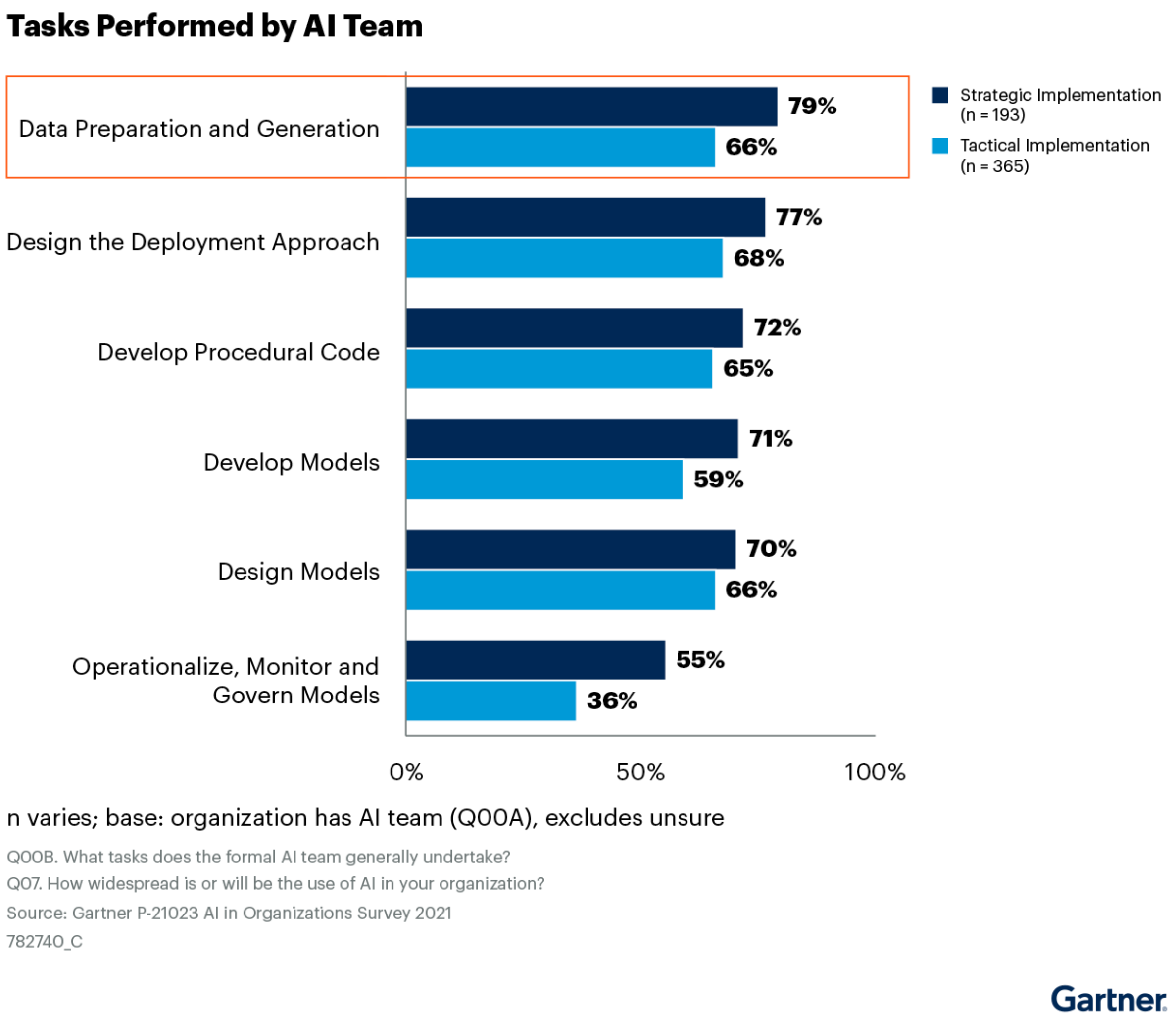
Through 2024, manual data management tasks will be reduced by over 50% through the addition of machine learning, leading to most of these tasks either being automated or augmented.

By 2025, use of synthetic data and transfer learning will reduce the volume of real data needed for AI by more than 50%.

## Analysis

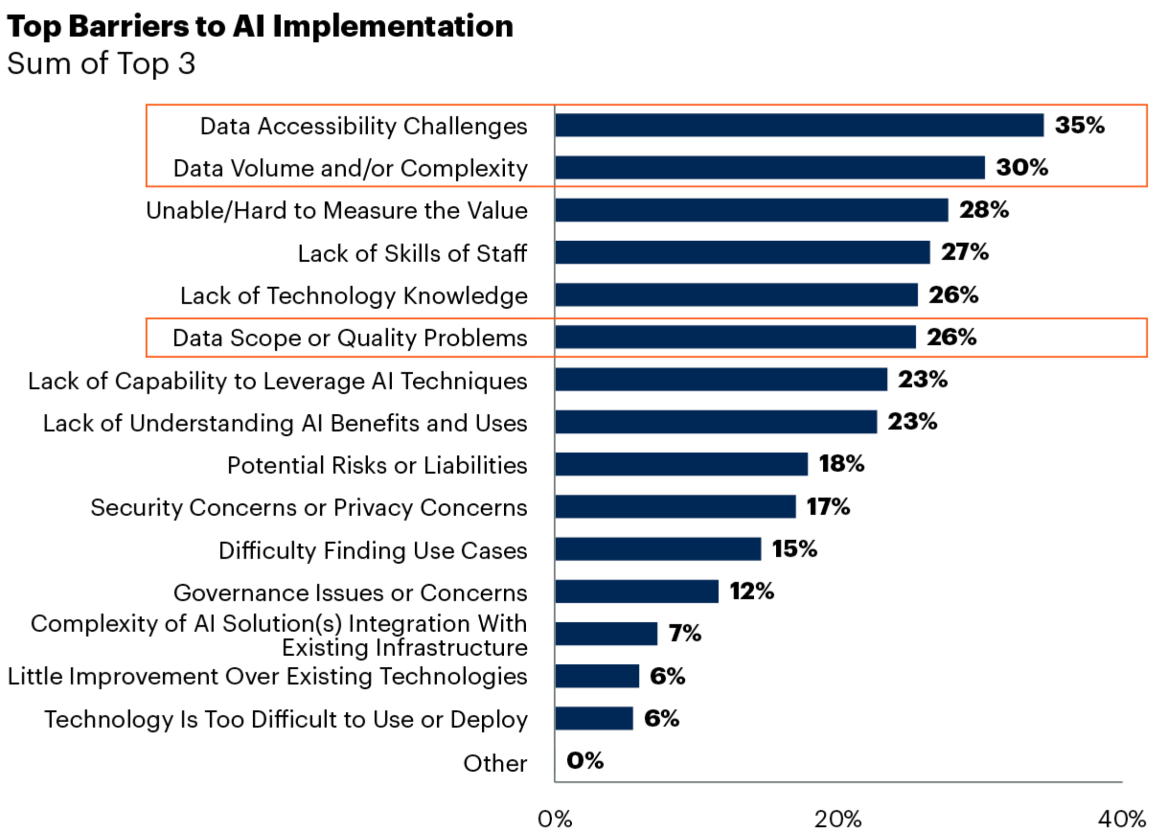
According to the 2021 Gartner in AI Organization's Survey, AI teams continually spend more time generating, curating and preparing data than developing and designing models (see Figure 1).

Figure 1: AI Team Task Distribution



Despite this, data challenges around accessibility, volume, complexity and quality continue to hinder AI implementations (see Figure 2).

Figure 2: Barriers to AI Implementation



n = 698; base: excludes not sure

Q18. What are or will be the top 3 barriers to the implementation of AI techniques within your organization?

Source: Gartner P-21023 AI in Organizations Survey 2021

782740_C

A question then emerges: why do data problems persist despite AI teams spending more time on data preparation?

## What is Data-centric AI and Why Is It Needed?

Data-centric AI focuses on systematically engineering data to build better AI systems and represents a shift from a model-centric and code-centric approach to a more data-focused approach. Traditional AI development has centered around refining and fine-tuning the algorithms or enhancing the code used to develop AI models, but the data-centric AI approach keeps the model and code constant while iteratively improving the data.

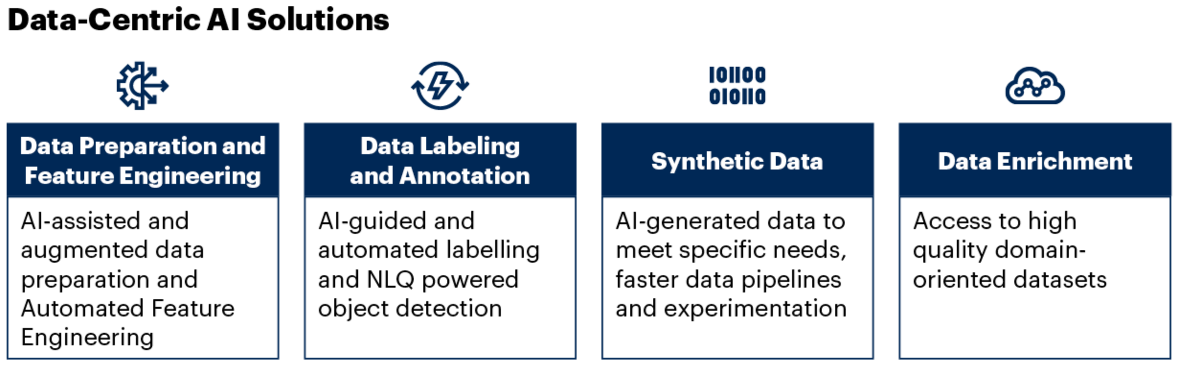
Bad data costs the U.S. around $3 trillion every year, and data quality issues are prevalent in every industry. [1] As datasets become larger and more complex, it becomes difficult to ensure their quality without using techniques and tools from both the data management and advanced analytics space. The recent rise of generative AI tools like ChatGPT rely on massive amounts of human labor to produce high-quality training data, and it may not always be feasible to use humans — as was the case with Kenyan workers getting exposed to explicit content while working to remove toxicity from training data. [2] As analytics and AI use cases continue to grow, automated methods, newer tools and systematic engineering approaches need to be revisited and strengthened to ensure maximum benefits from AI development.

> **Data-centric AI focuses on systematically engineering data to build better AI systems, and represents a shift from a model-centric and code-centric approach to a more data-focused approach.**

Figure 3 shows the different components of Data-centric AI:

- Data Preparation and Feature Engineering
- Data Labeling and Annotation
- Synthetic Data
- Data Enrichment

Figure 3: Data-Centric AI solutions



Data-Centric AI Solutions

**Data Preparation and Feature Engineering**
AI-assisted and augmented data preparation and Automated Feature Engineering

**Data Labeling and Annotation**
AI-guided and automated labelling and NLQ powered object detection

**Synthetic Data**
AI-generated data to meet specific needs, faster data pipelines and experimentation

**Data Enrichment**
Access to high quality domain-oriented datasets

Source: Gartner
782740_C

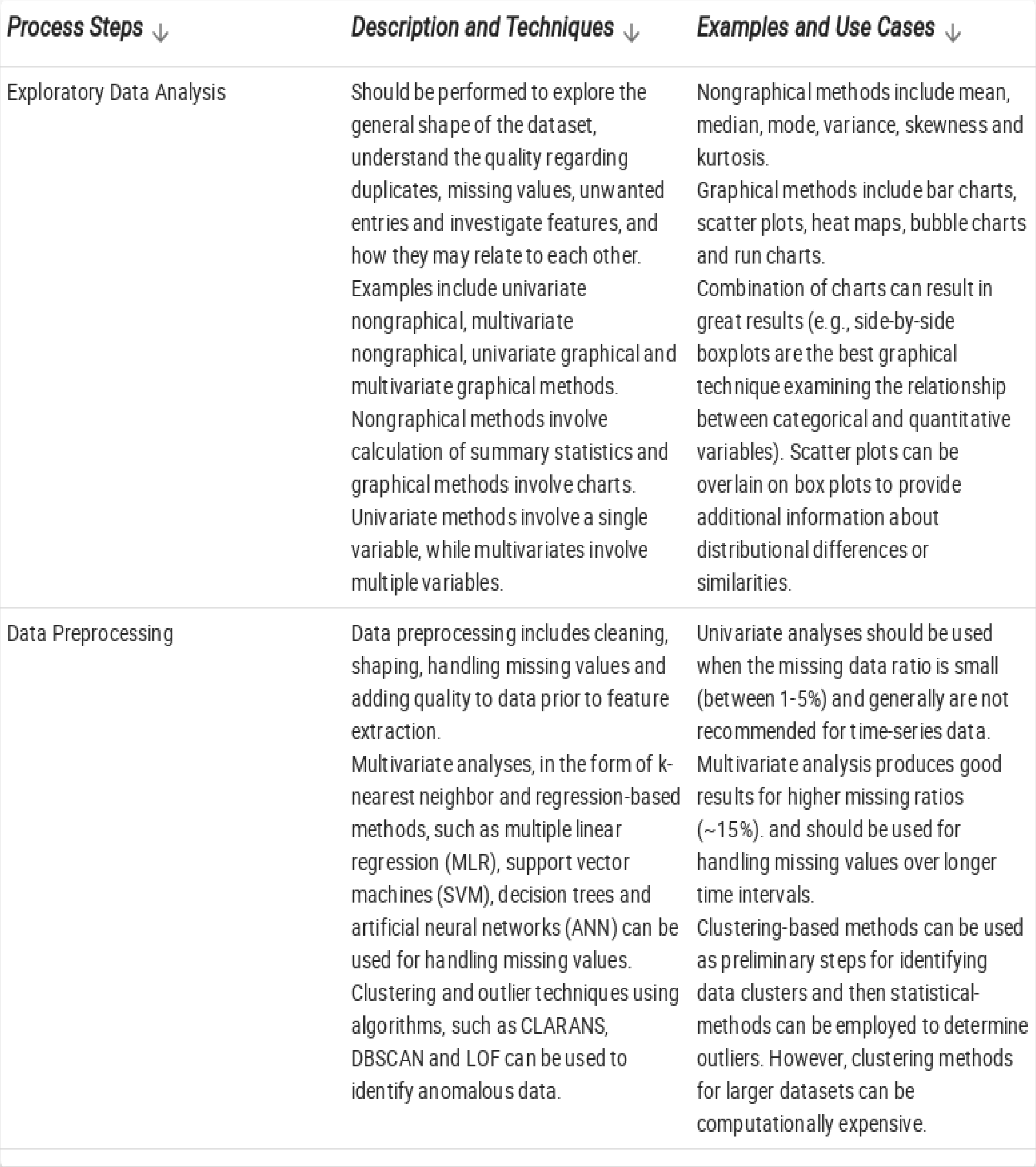
This research will explore each of these solutions and shed light on their features and how they can add more dynamism and help prepare high-quality data, as well as help democratize AI development.

### Data Preparation and Feature Engineering

Data preparation mostly focuses around exploratory data analysis (EDA), cleansing and transformation to prepare high-quality structured datasets for feature extraction and engineering. Both tasks are intertwined, but data preprocessing is mostly performed prior to feature engineering. Table 1 describes the data preparation steps.

Table 1: Data Preparation Processes, Steps and Techniques

(Enlarged table in Appendix)



| Process Steps ↓ | Description and Techniques ↓ | Examples and Use Cases ↓ |
|---|---|---|
| Exploratory Data Analysis | Should be performed to explore the general shape of the dataset, understand the quality regarding duplicates, missing values, unwanted entries and investigate features, and how they may relate to each other.<br>Examples include univariate nongraphical, multivariate nongraphical, univariate graphical and multivariate graphical methods.<br>Nongraphical methods involve calculation of summary statistics and graphical methods involve charts.<br>Univariate methods involve a single variable, while multivariates involve multiple variables. | Nongraphical methods include mean, median, mode, variance, skewness and kurtosis.<br>Graphical methods include bar charts, scatter plots, heat maps, bubble charts and run charts.<br>Combination of charts can result in great results (e.g., side-by-side boxplots are the best graphical technique examining the relationship between categorical and quantitative variables). Scatter plots can be overlain on box plots to provide additional information about distributional differences or similarities. |
| Data Preprocessing | Data preprocessing includes cleaning, shaping, handling missing values and adding quality to data prior to feature extraction.<br>Multivariate analyses, in the form of k-nearest neighbor and regression-based methods, such as multiple linear regression (MLR), support vector machines (SVM), decision trees and artificial neural networks (ANN) can be used for handling missing values.<br>Clustering and outlier techniques using algorithms, such as CLARANS, DBSCAN and LOF can be used to identify anomalous data. | Univariate analyses should be used when the missing data ratio is small (between 1-5%) and generally are not recommended for time-series data.<br>Multivariate analysis produces good results for higher missing ratios (~15%). and should be used for handling missing values over longer time intervals.<br>Clustering-based methods can be used as preliminary steps for identifying data clusters and then statistical-methods can be employed to determine outliers. However, clustering methods for larger datasets can be computationally expensive. |

Source: Gartner (June 2023)

Data preprocessing can be performed manually or through data preparation tools. Manual methods involve using open source programming languages like Python in the form of notebooks, and integrated development environments like PyCharm and Visual Studio Code for development. While common libraries like pandas, matplotlib and seaborn are relatively well-known, Table 2 provides some examples of powerful new and other well-known libraries that can help increase the speed and ease of data preparation.

Table 2: Open Source Tools for Data Preparation and Exploration

| *Python Library* ↓ | *Use Case* ↓ |
|---|---|
| Sweetviz, D-Tale, Vega-Altair, Bokeh, Plotly, bamboolib, Pandas-Profiling, Data-Purifier | Used in exploratory data analysis to create visualizations and charts for data exploration, statistics gathering, data distribution and completeness. Low-code nature can enable citizen data scientists as well. Data-Purifier is more suited to NLP applications. |
| dabl, Yellowbrick, Dash, pandas_ml, Petl, Data Measurements Tool | Used in data processing and transformations, but can also assist in model training. Petl can be used for ETL operations. Data Measurements Tool, by Hugging Face, is a no-code tool to help build, measure and compare NLP datasets. |
| | |

Source: Gartner (June 2023)

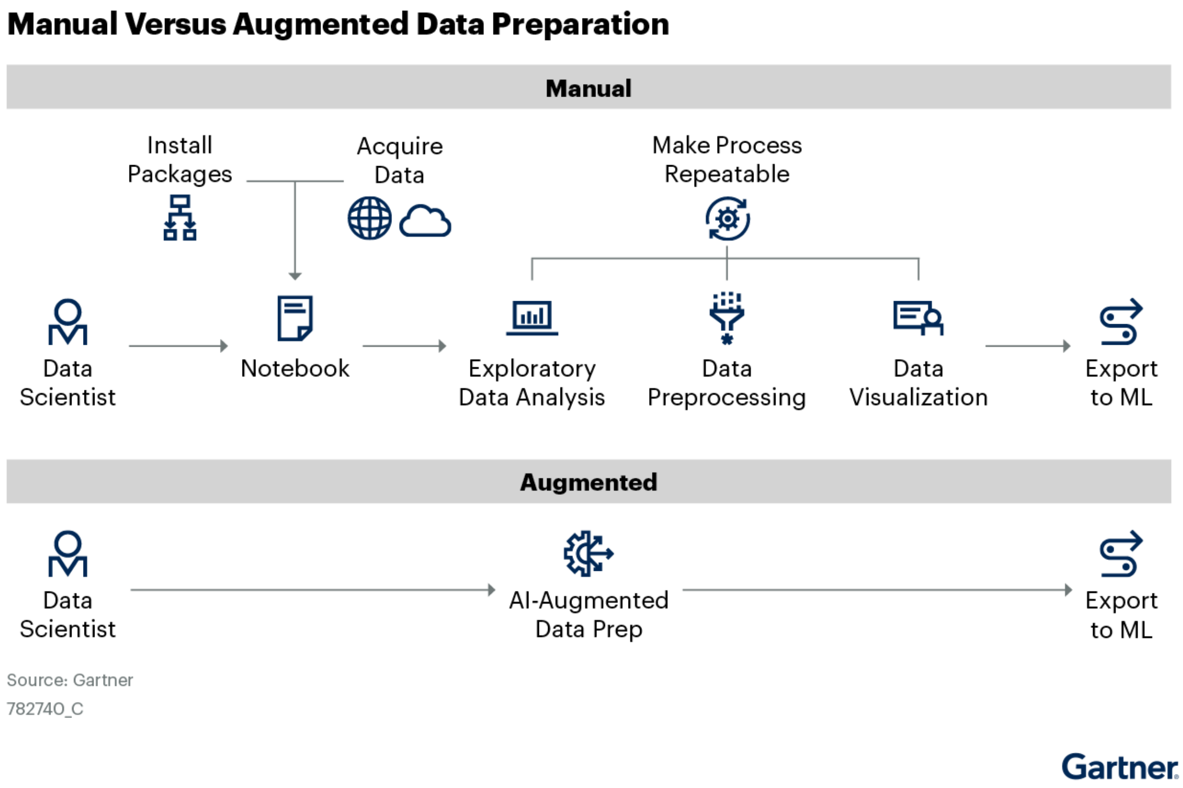
Data and analytics technical professionals should continue to explore open-source tools within the data-centric AI space.

Sometimes, manual data preparation can reduce the transparency or collaboration capabilities with other team members. The efforts can focus on solving one problem at a time, which prevents development of an end-to-end platform. Several components must be stitched together to create a platform-like appearance. This creates too many break points and requires stronger skill sets to manage the interactions between these components. Add in the complexity of infrastructure management (e.g., horizontal and vertical scaling of compute, selection of GPU versus CPU, automation, orchestration and DevOps) and the complexity of technical skills and cost management also go up.

To help offset this, modern data preparation tools provide augmentation in the form of no-code data preparation, cleansing, lineage, cataloging, as well as orchestration and operationalization capabilities. However, it should be noted there is no one-size-fits-all approach, and some data and analytics technical professionals prefer manual wrangling. As such, it is important to research the appropriate method — keeping in mind the availability of not only technical wrangling skills, but also IT resources.

Figure 4 shows the comparison of a manual data preparation workflow against an augmented workflow enhanced through self-serve data preparation tools.

**Figure 4: Manual Versus Augmented Data Preparation**

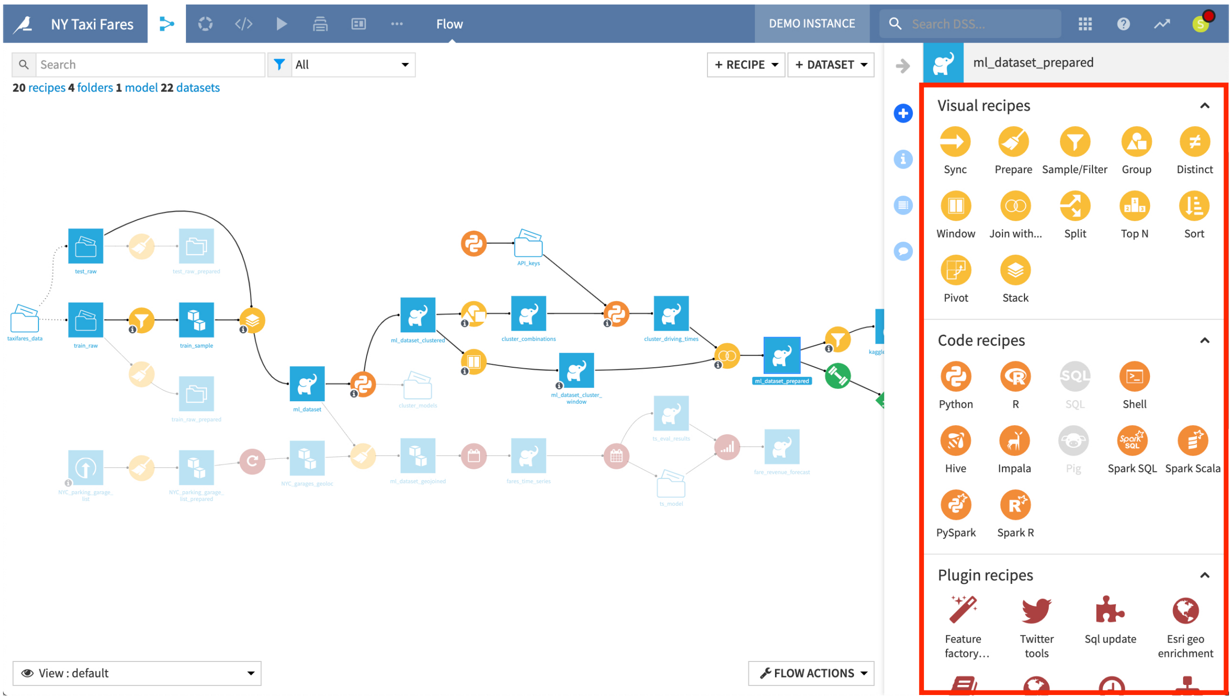


Source: Gartner
782740_C

Data and analytics professionals should carefully review the features of these tools and evaluate them according to their use cases:

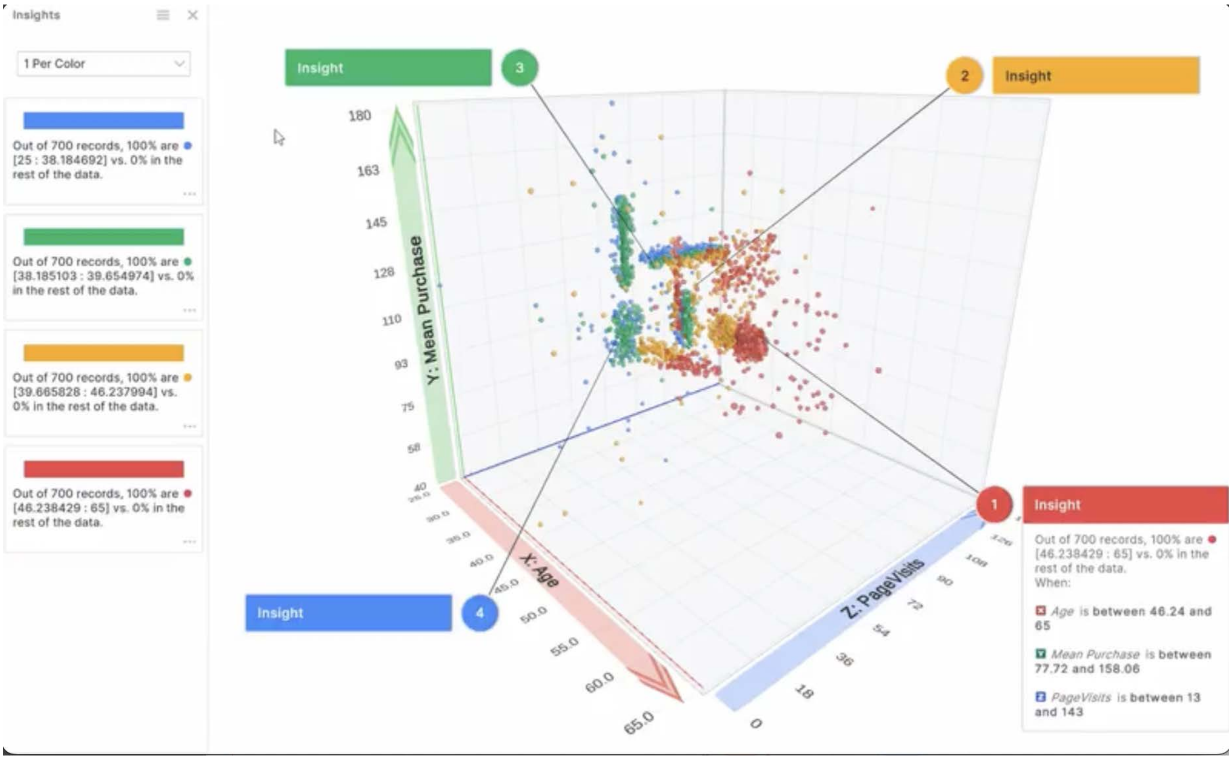
- **Low-code/no-code interface and code extensibility:** Most self-serve data preparation tools provide a visual interface for authoring transformation jobs called recipes. This ensures consistency in the method for preparation. These recipes can then be customized, (e.g., through Python, PySpark or R), while some tools also offer the ability to embed external scripts into the platform. Figure 5 shows the visual and code-based recipes on Dataiku for data preparation.

## Figure 5: Visual and Code Recipes in Dataiku for Data Preparation



- **AI-augmented and 3D exploratory data analysis:** 3D, NLQ and virtual reality infused visual exploration provides an in-depth view of the datasets and can reveal hidden relationships. Figure 6 shows an example from Virtualitics featuring AI-assisted insights.

Figure 6: AI-assisted Insights From Virtualitics



Out of 700 records, 100% are ● [25 : 38.184692] vs. 0% in the rest of the data.

Out of 700 records, 100% are ● [38.185103 : 39.654974] vs. 0% in the rest of the data.

Out of 700 records, 100% are ● [39.665828 : 46.237994] vs. 0% in the rest of the data.

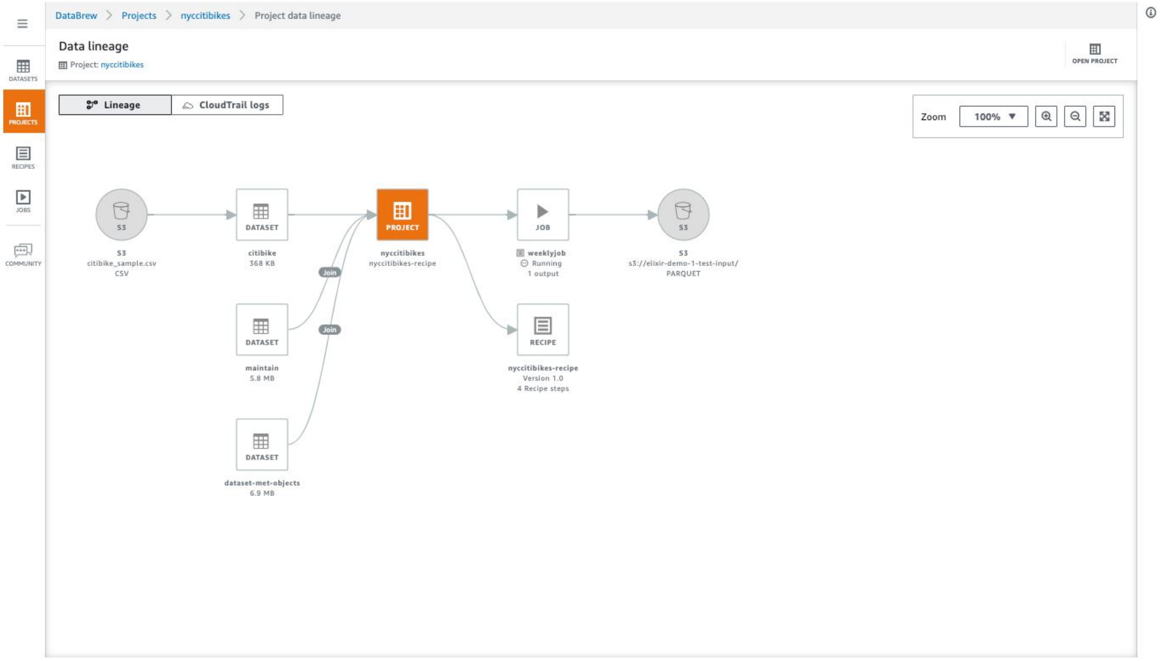
Out of 700 records, 100% are ● [46.238429 : 65] vs. 0% in the rest of the data.

Out of 700 records, 100% are ● [46.238429 : 65] vs. 0% in the rest of the data.
When:

- *Age* is between 46.24 and 65
- *Mean Purchase* is between 77.72 and 158.06
- *PageVisits* is between 13 and 143

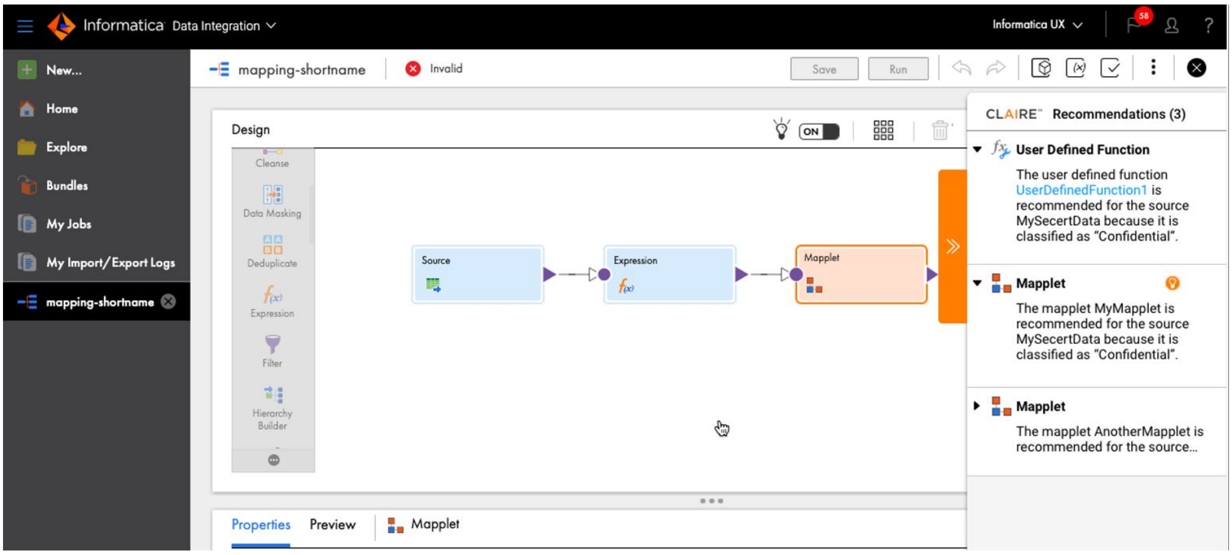
- **Data lineage and quality:** Visual authorship enables data lineage to be maintained and each step of the data preparation pipeline can be separately analyzed as opposed to scanning notebooks for code analyses. Augmented data quality checks automatically scan data for missing and erroneous values. Figure 7 shows a data lineage tracking and metadata collection example from AWS Glue DataBrew.

## Figure 7: Data Lineage in AWS Glue DataBrew



- **Predictive and Automated Transformations:** Impact of transformations can be predicted prior to application and tools can suggest transformations based on the metadata. Informatica's Data Preparation tool is powered by CLAIRE, a custom ML-engine used to suggest transformations as seen in Figure 8.

## Figure 8: AI-assisted Transformations in Informatica



User Defined Function

The user defined function UserDefinedFunction1 is recommended for the source MySecertData because it is classified as "Confidential".

Mapplet

The mapplet MyMapplet is recommended for the source MySecertData because it is classified as "Confidential".

Mapplet

The mapplet AnotherMapplet is recommended for the source...

Apart from the functionalities above, data preparation tools also help with:

- **Compute elasticity and scalability:** Compute scales automatically without requiring manual node and cluster addition or reconfiguration. Amazon SageMaker Data Wrangler allows the selection of EC2 machine types for particular jobs, enabling CPU or GPU intensive tasks to be handled accordingly. Google's Dataprep by Trifacta can select different engines based on the transformation type (e.g., for heavy workloads, it automatically selects BigQuery, while for smaller workloads it uses its internal in-memory engine).

- **Operationalization:** Enables continuous deployment with recipe imports/exports, flow parameters and custom configurations to automate software development life cycle and monitoring.

- **Comprehensive connectivity:** Reduce the need to set up custom connectors to SQL engines or data lakes by providing out-of-the-box connectors to cloud and on-premises data sources. Some tools are designed for specialized use cases (e.g., SAS Data Loader for Hadoop is primarily designed to load, ingest and transform data from the Apache Hadoop ecosystem).

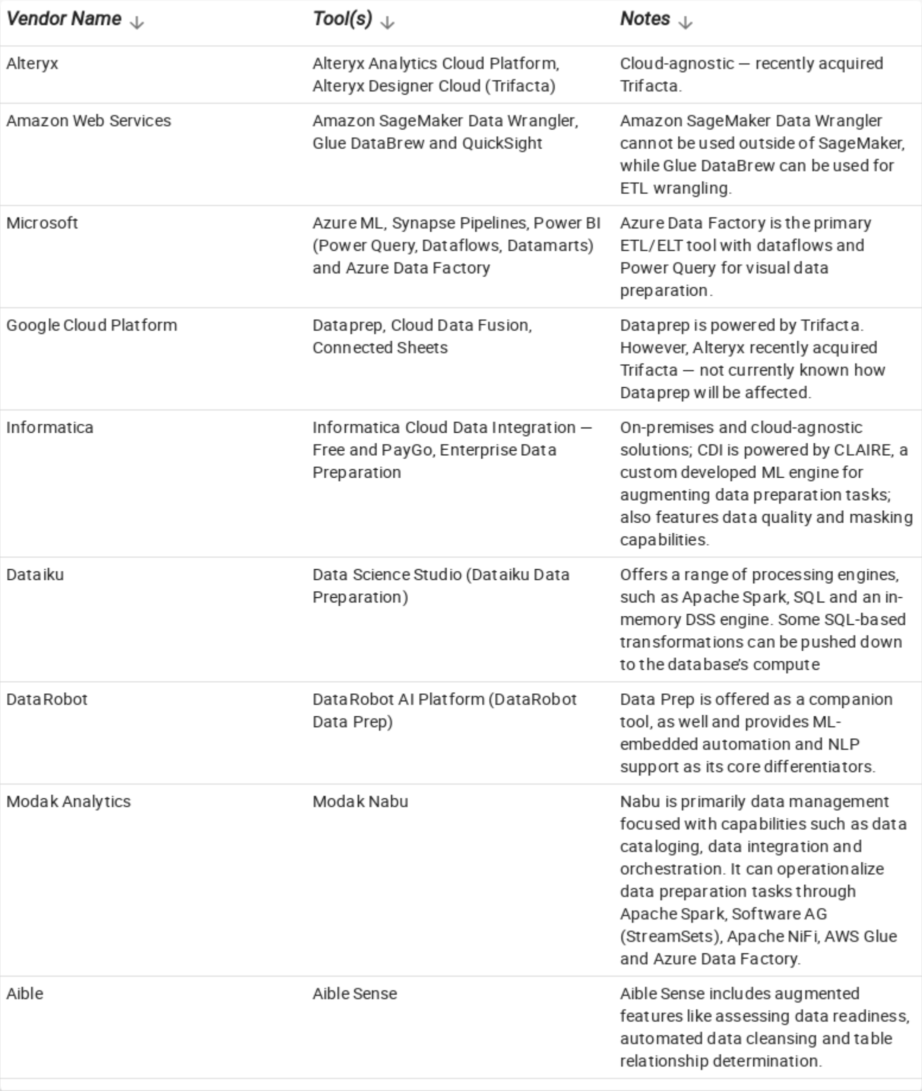
- **Security:** Provide individual data access control and can be combined with IAM rules on AWS, Azure and GCP to further restrict access to PII and PHI data.

For more information on developing a data quality framework for AI, read Overcoming Data Quality Risks When Using Semistructured and Unstructured Data for AI/ML Models.

Table 3 shows some illustrative examples, and is not meant as an exhaustive list.

## Table 3: List of Data Preparation Tools

(Enlarged table in Appendix)



| Vendor Name ↓ | Tool(s) ↓ | Notes ↓ |
|---|---|---|
| Alteryx | Alteryx Analytics Cloud Platform, Alteryx Designer Cloud (Trifacta) | Cloud-agnostic — recently acquired Trifacta. |
| Amazon Web Services | Amazon SageMaker Data Wrangler, Glue DataBrew and QuickSight | Amazon SageMaker Data Wrangler cannot be used outside of SageMaker, while Glue DataBrew can be used for ETL wrangling. |
| Microsoft | Azure ML, Synapse Pipelines, Power BI (Power Query, Dataflows, Datamarts) and Azure Data Factory | Azure Data Factory is the primary ETL/ELT tool with dataflows and Power Query for visual data preparation. |
| Google Cloud Platform | Dataprep, Cloud Data Fusion, Connected Sheets | Dataprep is powered by Trifacta. However, Alteryx recently acquired Trifacta — not currently known how Dataprep will be affected. |
| Informatica | Informatica Cloud Data Integration — Free and PayGo, Enterprise Data Preparation | On-premises and cloud-agnostic solutions; CDI is powered by CLAIRE, a custom developed ML engine for augmenting data preparation tasks; also features data quality and masking capabilities. |
| Dataiku | Data Science Studio (Dataiku Data Preparation) | Offers a range of processing engines, such as Apache Spark, SQL and an in-memory DSS engine. Some SQL-based transformations can be pushed down to the database's compute |
| DataRobot | DataRobot AI Platform (DataRobot Data Prep) | Data Prep is offered as a companion tool, as well and provides ML-embedded automation and NLP support as its core differentiators. |
| Modak Analytics | Modak Nabu | Nabu is primarily data management focused with capabilities such as data cataloging, data integration and orchestration. It can operationalize data preparation tasks through Apache Spark, Software AG (StreamSets), Apache NiFi, AWS Glue and Azure Data Factory. |
| Aible | Aible Sense | Aible Sense includes augmented features like assessing data readiness, automated data cleansing and table relationship determination. |

Source: Gartner (June 2023)

Some vendors have started to offer specialized capabilities, such as AI-assisted data exploration and data enrichment as well. Examples include Virtualitics, SparkBeyond and Explorium. Virtualitics integrates with Databricks, and offers NLQ and virtual reality capabilities.

For details on the AWS, Azure and GCP products read the following documents:

- Building an Analytics and AI Architecture Using Amazon Web Services
- Building an Analytics and AI Architecture Using Microsoft Azure
- Building an Analytics and AI Architecture Using Google Cloud Platform

For more details on vendors operating within this space, read the following documents:

- Cool Vendors in Data-Centric AI
- Market Guide for Data Preparation Tools
- Market Guide for Multipersona Data Science and Machine Learning Platforms

### Feature Engineering

Feature engineering is the construction of features to add nuance or meaning to datasets for improving model performance and accuracy. It is an iterative process where features are extracted from the refined and curated data, and used as inputs to the model. If the model needs to be retrained, the features are reengineered. It is one of the most resource intensive tasks since it requires a combination of domain and technical skills.

Feature engineering involves the following steps:

- **Feature creation:** involves creating features from available data and techniques include splitting, binning and one-hot encoding.
- **Feature transformation:** focuses on handling missing features and replacing them if they are not required. Common techniques include creating a cartesian product of features and creating domain-specific features.
- **Feature extraction:** includes dimensionality reduction techniques to reduce the amount of data to be processed. This also reduces the need for more compute resources.
- **Feature selection:** revolves around selecting a curated subset of features to select the most relevant features for model training.

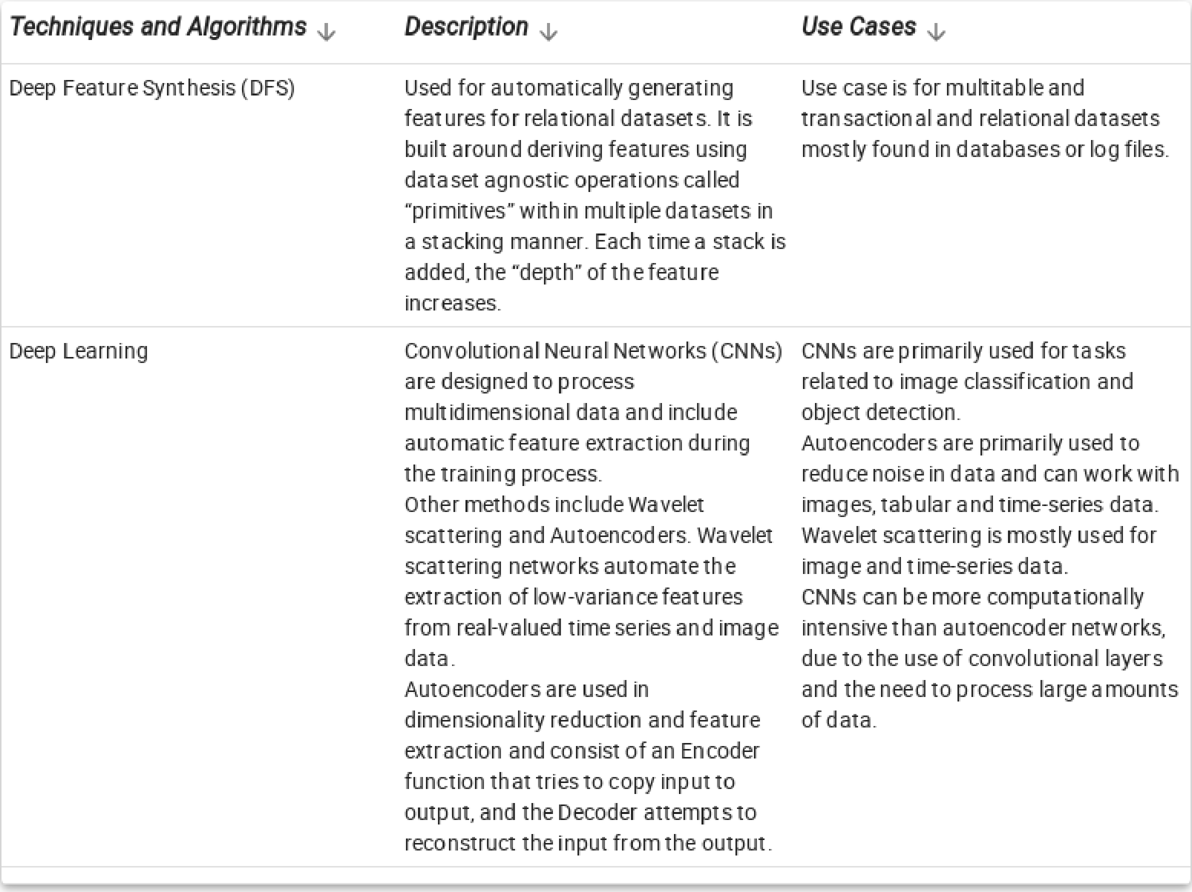
Manual feature engineering has revolved around building features one at a time using domain knowledge. However, this can be an error-prone, time-consuming and resource-intensive task, with the code often being use-case-specific and requiring a rewrite for new cases. Moreover, with the growth of datasets, there may be a large number of features to consider and not every potential feature is relevant to the modeling problem.

Automated Feature Engineering (AFE) focuses on automatically generating features using a framework that is use-case agnostic. It aims to not only reduce the resources spent on feature engineering, but also creates highly interpretable features to build models faster.

Table 4 shows powerful techniques and frameworks to enhance feature engineering, as well as application use cases.

Table 4: Automated Feature Engineering Techniques, Tools and Vendors

(Enlarged table in Appendix)



| Techniques and Algorithms ↓ | Description ↓ | Use Cases ↓ |
|---|---|---|
| Deep Feature Synthesis (DFS) | Used for automatically generating features for relational datasets. It is built around deriving features using dataset agnostic operations called "primitives" within multiple datasets in a stacking manner. Each time a stack is added, the "depth" of the feature increases. | Use case is for multitable and transactional and relational datasets mostly found in databases or log files. |
| Deep Learning | Convolutional Neural Networks (CNNs) are designed to process multidimensional data and include automatic feature extraction during the training process.<br>Other methods include Wavelet scattering and Autoencoders. Wavelet scattering networks automate the extraction of low-variance features from real-valued time series and image data.<br>Autoencoders are used in dimensionality reduction and feature extraction and consist of an Encoder function that tries to copy input to output, and the Decoder attempts to reconstruct the input from the output. | CNNs are primarily used for tasks related to image classification and object detection.<br>Autoencoders are primarily used to reduce noise in data and can work with images, tabular and time-series data.<br>Wavelet scattering is mostly used for image and time-series data.<br>CNNs can be more computationally intensive than autoencoder networks, due to the use of convolutional layers and the need to process large amounts of data. |

Source: Gartner (June 2023)

Some examples of open-source libraries for automated feature engineering include:

- Featuretools — uses Deep Feature Synthesis at transforming temporal and relational data. However, nlp_primitives can be installed for NLP use cases as well
- AutoFeat — uses AutoFeatRegressor and AutoFeatClassifier models to automatically generate and select features — use case restricted to Supervised Learning

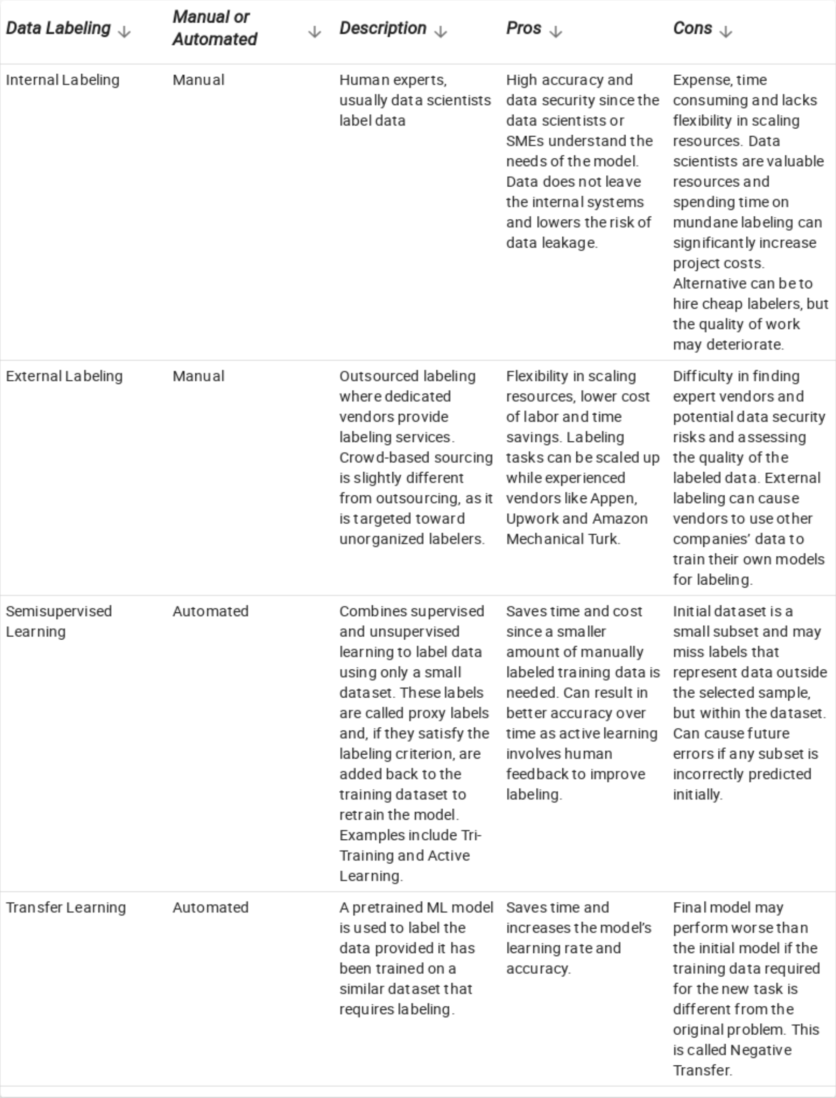
### Data Labeling and Annotation

Data labeling refers to the addition of metadata to unstructured data (images, text, videos and audio files) for identifying features for AI development. It is one of the most time-consuming and resource-intensive tasks in the AI development process and requires a very high degree of precision. Even a small case of incorrectly labeled data in a pharmaceutical use case, for example, can lead to product recalls, government fines and reputational damage.

Data labeling can be divided into manual and automated. Table 5 explains the differences, pros and cons of both manual and automated approaches:

### Table 5: Data Labeling Approaches

(Enlarged table in Appendix)



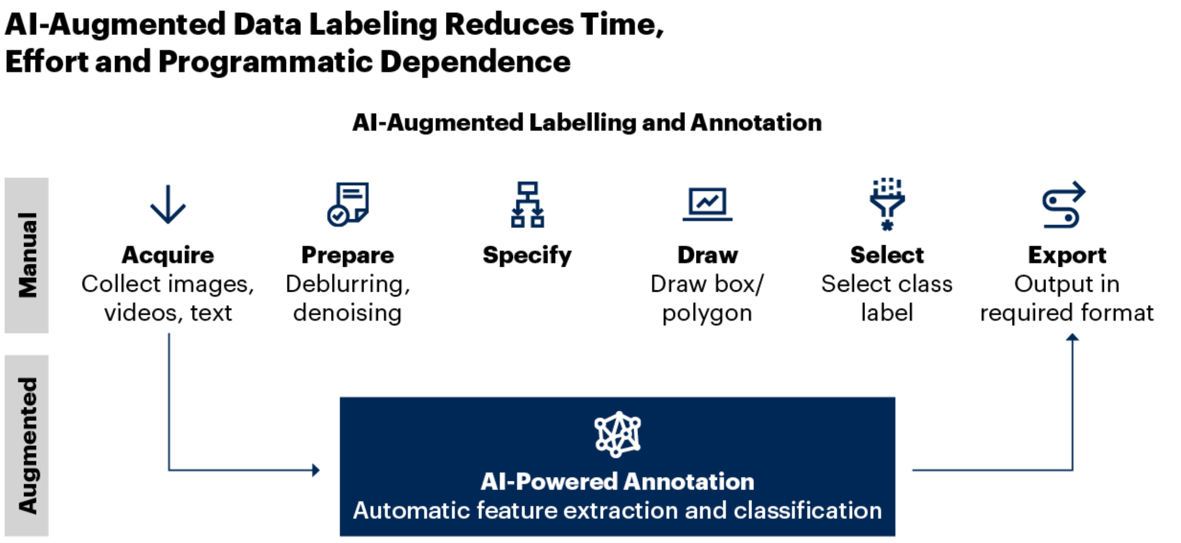
| *Data Labeling* | *Manual or Automated* | *Description* | *Pros* | *Cons* |
|---|---|---|---|---|
| Internal Labeling | Manual | Human experts, usually data scientists label data | High accuracy and data security since the data scientists or SMEs understand the needs of the model. Data does not leave the internal systems and lowers the risk of data leakage. | Expense, time consuming and lacks flexibility in scaling resources. Data scientists are valuable resources and spending time on mundane labeling can significantly increase project costs. Alternative can be to hire cheap labelers, but the quality of work may deteriorate. |
| External Labeling | Manual | Outsourced labeling where dedicated vendors provide labeling services. Crowd-based sourcing is slightly different from outsourcing, as it is targeted toward unorganized labelers. | Flexibility in scaling resources, lower cost of labor and time savings. Labeling tasks can be scaled up while experienced vendors like Appen, Upwork and Amazon Mechanical Turk. | Difficulty in finding expert vendors and potential data security risks and assessing the quality of the labeled data. External labeling can cause vendors to use other companies' data to train their own models for labeling. |
| Semisupervised Learning | Automated | Combines supervised and unsupervised learning to label data using only a small dataset. These labels are called proxy labels and, if they satisfy the labeling criterion, are added back to the training dataset to retrain the model. Examples include Tri-Training and Active Learning. | Saves time and cost since a smaller amount of manually labeled training data is needed. Can result in better accuracy over time as active learning involves human feedback to improve labeling. | Initial dataset is a small subset and may miss labels that represent data outside the selected sample, but within the dataset. Can cause future errors if any subset is incorrectly predicted initially. |
| Transfer Learning | Automated | A pretrained ML model is used to label the data provided it has been trained on a similar dataset that requires labeling. | Saves time and increases the model's learning rate and accuracy. | Final model may perform worse than the initial model if the training data required for the new task is different from the original problem. This is called Negative Transfer. |

Source: Gartner (June 2023)

Manual data labeling can be prone to human error and lack of standardization apart from being time- and resource-consuming. Data and analytics technical professionals should strive to follow a blended approach by using AI-augmented data labeling and annotation tools. Modern data labeling tools are powered by active learning to allow AI to detect images and create polygons and bounding boxes automatically. These tools also feature human-in-the-loop (HITL), thereby allowing data scientists and SMEs to validate the labels prior to consumption. Advances in generative modeling are also enabling new data-centric AI vendors, such as LandingLens, that use visual prompting to reduce labeling steps and, instead, focus more on image pattern recognition.

Figure 9 shows the difference between manual and AI-augmented data labeling.

Figure 9: AI-augmented Data Labeling Reduces Time, Effort and Programmatic Dependence


AI-Augmented Data Labeling Reduces Time, Effort and Programmatic Dependence
AI-Augmented Labelling and Annotation
Manual
Augmented
Acquire
Collect images, videos, text
Prepare
Deblurring, denoising
Specify
Draw
Draw box/ polygon
Select
Select class label
Export
Output in required format
AI-Powered Annotation
Automatic feature extraction and classification

Source: Gartner
782740_C

When selecting tools for annotation, analytics technical professionals should determine the use cases. Computer vision use cases involve polygons, polylines, semantic and instance segmentation. NLP use cases involve text annotation and include entity annotation, entity linking, text classification and sentiment annotation. Not all vendors provide capabilities for computer vision and NLP annotation within a single tool. Some vendors also offer human labeling services in addition to the human-in-the-loop validation, and specialize in certain domains such as medical imaging.

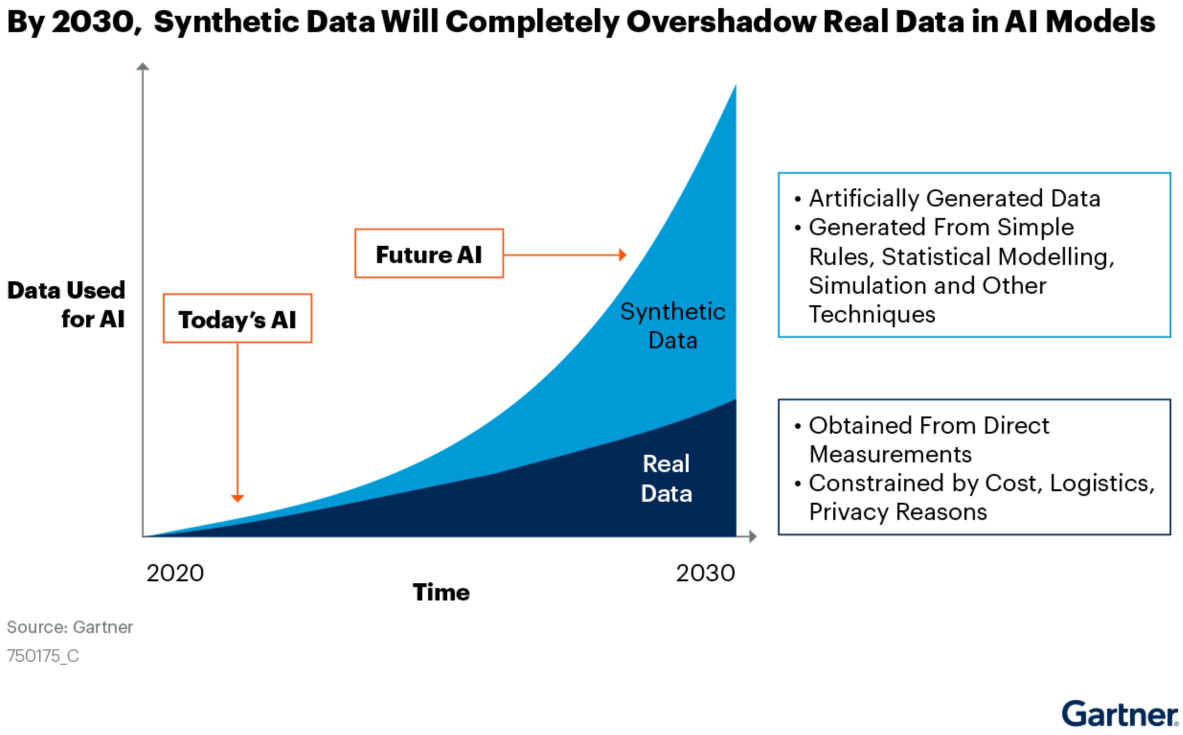
Some example vendors operating in this space include SuperAnnotate, ZERO Systems, Labelbox and Landing AI.

### Synthetic Data

Synthetic data is a class of generative AI that can optimize scarce data, mitigate bias or preserve data privacy. It is generated through different means such as statistical sampling from real data, semantic approaches, GANs or through Large Language Models. Synthetic datasets retain the statistical and behavioral aspects of real datasets. Most often, these are used when access to real-world data is expensive or not available, in the case of privacy concerns for customer data, or when internal data needs to be augmented or enhanced. It can be used as a replacement for sensitive production data in nonproduction environments.

Synthetic data has been used in computer modeling and simulations, but has recently emerged as an important aspect within AI development and is projected to overshadow real data in the future (see Figure 10).

Figure 10: Projected Growth of Synthetic Data



Source: Gartner
750175_C

Synthetic data can be structured or unstructured. Structured data refers to tabular data in the form of rows and columns and can mirror data stored in data warehouses. Unstructured data includes text, images, audio and video formats.

Synthetic data can be generated by:

- Using services offered by vendors, which reduce the technical complexity required to manually or programmatically generate synthetic data.
- Using synthetic data generation techniques such as statistical distribution, variational autoencoders (VAEs), generative adversarial networks (GANs) and other techniques like gaming engines.

Table 6 shows some illustrative vendors and some example open source Python libraries for synthetic data. For more details and an exhaustive list, see Innovation Insight for Synthetic Data and Market Guide for AI Trust, Risk and Security Management.

Table 6: Tools for Synthetic Data

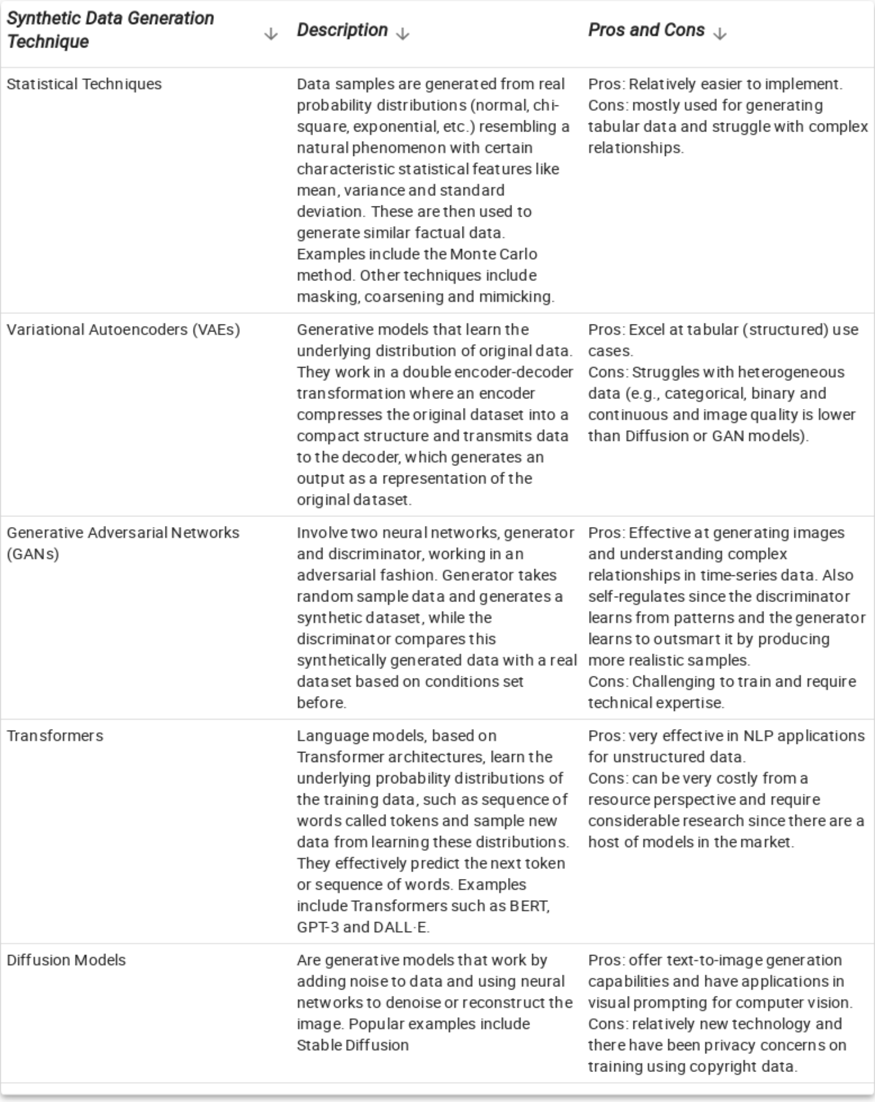
| *Synthetic Data* ↓ | *Vendors* ↓ | *Open Source* ↓ |
|---|---|---|
| Relational | Gretel AI, MOSTLY AI, Synth, Synthesized, IBM, Hazy, Facteus | Synthetic Data Vault, DataSynthesizer, Pydbgen, Mimesis |
| Nonrelational | MOSTLY AI, Bitext, EEdgecase.AI, Rendered.AI, Scale AI, Synthesis AI | zpy, genalog, Kubric, Blender |
| | | |

Source: Gartner (June 2023)

While using vendor solutions provide ease of use, they can also be costly. Some data and analytics technical professionals would prefer generating synthetic data manually using statistical or more advanced generative AI techniques. Table 7 shows techniques around synthetic data generation and their pros and cons.

## Table 7: Synthetic Data Generation Techniques

(Enlarged table in Appendix)



| *Synthetic Data Generation Technique* | *Description* | *Pros and Cons* |
|---|---|---|
| Statistical Techniques | Data samples are generated from real probability distributions (normal, chi-square, exponential, etc.) resembling a natural phenomenon with certain characteristic statistical features like mean, variance and standard deviation. These are then used to generate similar factual data. Examples include the Monte Carlo method. Other techniques include masking, coarsening and mimicking. | Pros: Relatively easier to implement.<br>Cons: mostly used for generating tabular data and struggle with complex relationships. |
| Variational Autoencoders (VAEs) | Generative models that learn the underlying distribution of original data. They work in a double encoder-decoder transformation where an encoder compresses the original dataset into a compact structure and transmits data to the decoder, which generates an output as a representation of the original dataset. | Pros: Excel at tabular (structured) use cases.<br>Cons: Struggles with heterogeneous data (e.g., categorical, binary and continuous and image quality is lower than Diffusion or GAN models). |
| Generative Adversarial Networks (GANs) | Involve two neural networks, generator and discriminator, working in an adversarial fashion. Generator takes random sample data and generates a synthetic dataset, while the discriminator compares this synthetically generated data with a real dataset based on conditions set before. | Pros: Effective at generating images and understanding complex relationships in time-series data. Also self-regulates since the discriminator learns from patterns and the generator learns to outsmart it by producing more realistic samples.<br>Cons: Challenging to train and require technical expertise. |
| Transformers | Language models, based on Transformer architectures, learn the underlying probability distributions of the training data, such as sequence of words called tokens and sample new data from learning these distributions. They effectively predict the next token or sequence of words. Examples include Transformers such as BERT, GPT-3 and DALL-E. | Pros: very effective in NLP applications for unstructured data.<br>Cons: can be very costly from a resource perspective and require considerable research since there are a host of models in the market. |
| Diffusion Models | Are generative models that work by adding noise to data and using neural networks to denoise or reconstruct the image. Popular examples include Stable Diffusion | Pros: offer text-to-image generation capabilities and have applications in visual prompting for computer vision.<br>Cons: relatively new technology and there have been privacy concerns on training using copyright data. |

Source: Gartner (June 2023)

It is challenging to find a single model that can work with multiple data types and can generate use-case agnostic data. Data and analytics technical professionals need to carefully evaluate which information they will work with to pick suitable approaches.

Synthetic data can be useful in the following scenarios:

- **Lack of data.** Real-world datasets may not represent extreme or edge cases, since they do not occur frequently. While these anomalous events may not seem important, they can have serious consequences when it comes to training self-driving cars, which may require training on unanticipated events.

- **Data quality.** Access to highly curated datasets can be restricted and watered-down due to anonymization. High-quality synthetic data can be generated to help data scientists gain access to data in such scenarios. For more details, read Overcoming Data Quality Risks When Using Semistructured and Unstructured Data for AI/ML Models.

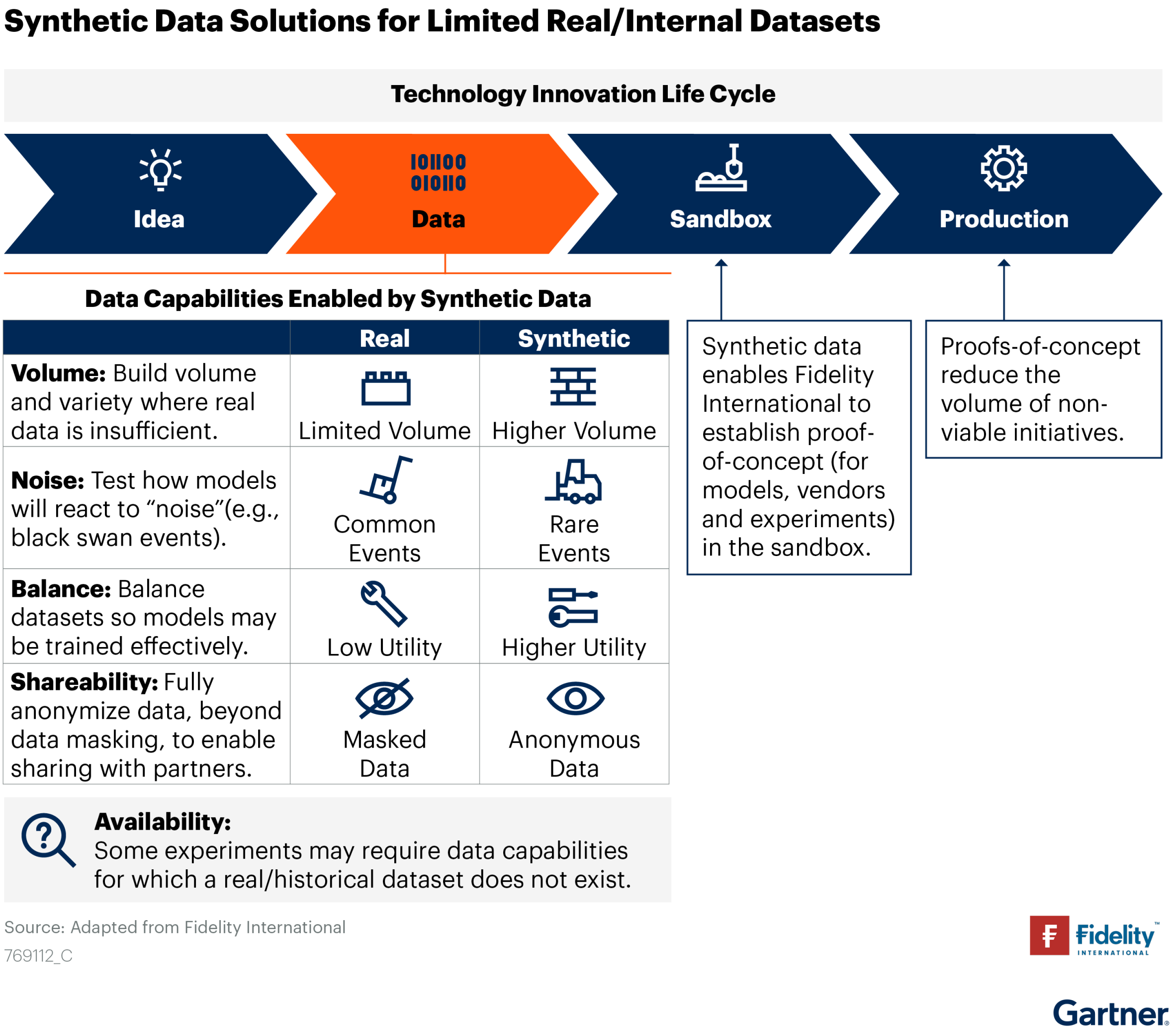
- **Data privacy.** First-party data can become difficult to collect, especially in the digital marketing space. It becomes difficult to find solutions that provide the required capabilities and protect consumer's privacy as well. Synthetic data can be used to generate datasets that would otherwise not be possible due to privacy restrictions.

- **Data labeling.** It is very time-consuming and resource intensive to gather real-world visual data, while allowing for diversity and respecting privacy. Accurate data labeling and annotation can take months. Synthetic data does not require manual data collection or annotation, and data scientists can begin working with accurate labeled datasets from the get-go.

In Case Study: Enable Business-Led Innovation with Synthetic Data (Fidelity International), Fidelity faced three most common obstacles when working with real internal datasets: access, anonymization and availability. Due to the sensitive nature of data, access to data was restricted as part of governance measures. Anonymization measures, such as data masking, were not enough to hide sensitive information and made distribution of data outside a central analytics team difficult. Lastly, some experiments required data volume and variety, for which internal datasets were simply not sufficient.

Synthetic data allowed Fidelity International to develop a robust data footprint that could accommodate a larger swath of business cases. Fidelity focused on volume, noise, balance and shareability as key capabilities of synthetic data, as shown in Figure 11.

Figure 11: Key Data Capabilities Enabled Through Synthetic Data



Synthetic Data Solutions for Limited Real/Internal Datasets

Technology Innovation Life Cycle

Data Capabilities Enabled by Synthetic Data

| | Real | Synthetic |
|---|---|---|
| **Volume:** Build volume and variety where real data is insufficient. | Limited Volume | Higher Volume |
| **Noise:** Test how models will react to “noise”(e.g., black swan events). | Common Events | Rare Events |
| **Balance:** Balance datasets so models may be trained effectively. | Low Utility | Higher Utility |
| **Shareability:** Fully anonymize data, beyond data masking, to enable sharing with partners. | Masked Data | Anonymous Data |

**Availability:** Some experiments may require data capabilities for which a real/historical dataset does not exist.

Synthetic data enables Fidelity International to establish proof-of-concept (for models, vendors and experiments) in the sandbox.

Proofs-of-concept reduce the volume of non-viable initiatives.

Source: Adapted from Fidelity International
769112_C

- **Volume**: Machine learning requires extensive amounts of data points to train models. In some experiments, for example, a model may require more data points as opposed to what Fidelity has available internally.

- **Noise**: Fidelity International wanted to be able to train models around hypothetical and/or rare events, such as geopolitical conflict; real/internal datasets may not contain instances of these events for testing.

- **Balance**: As a financial services company, Fidelity International often needs to develop solutions for emerging business problems for which existing datasets are incomplete or skewed. To train a model to identify, analyze and mitigate fraud, for example, Fidelity International would need to increase the volume and variety of data points.

- **Shareability:** Real data, even when anonymized, may be restricted internally and externally, which makes that data unusable for any internal experiment — or particularly, for any initiative involving a third party, such as vendors or ecosystem partners. Synthetic data is inherently anonymized and safe for sharing.

As is the case with any technology, data and analytics technical professionals should be aware of the following risks when considering synthetic data:

- Overall quality and reliability of the synthetic dataset depends on the foundation dataset. If this dataset changes, it will be necessary to regenerate synthetic data using new characteristics.
- Synthetic data may not be nuanced enough to mimic real-world data. For example, synthetic MRI images, which require a very high degree of detail, are still considered inferior to their real counterparts.
- Synthetic data can face skepticism and be mistaken for fake or inferior data and can reduce trust in the model predictions generated from synthetic data. Awareness needs to be spread that synthetic data should not be confused with fake data, since it is generated based on datasets relevant to the use case.
- It is an emerging technology, and adoption rate is still low in enterprises. Generating synthetic data, in the absence of ready-made solutions, can be expensive and requires highly skilled data scientists with expertise in deep learning.

### Data Enrichment

Searching for, and selecting, highly curated datasets aligned to the AI use cases is one of the key challenges plaguing data and analytics technical professionals. Data enrichment refers to augmenting internal data with domain-specific data from external data sources. Data enrichment tools can gather third-party data from the internet (among other sources) and organize, clean and aggregate the data from disparate sources.

Data elements can include basic contact information like phone and emails to driver licenses, passports, bank accounts and IP addresses. More capabilities continue to be added, such as calculating distances between two addresses, returning the user's name and address associated with a phone number, or finding correlations between individual and business entities.

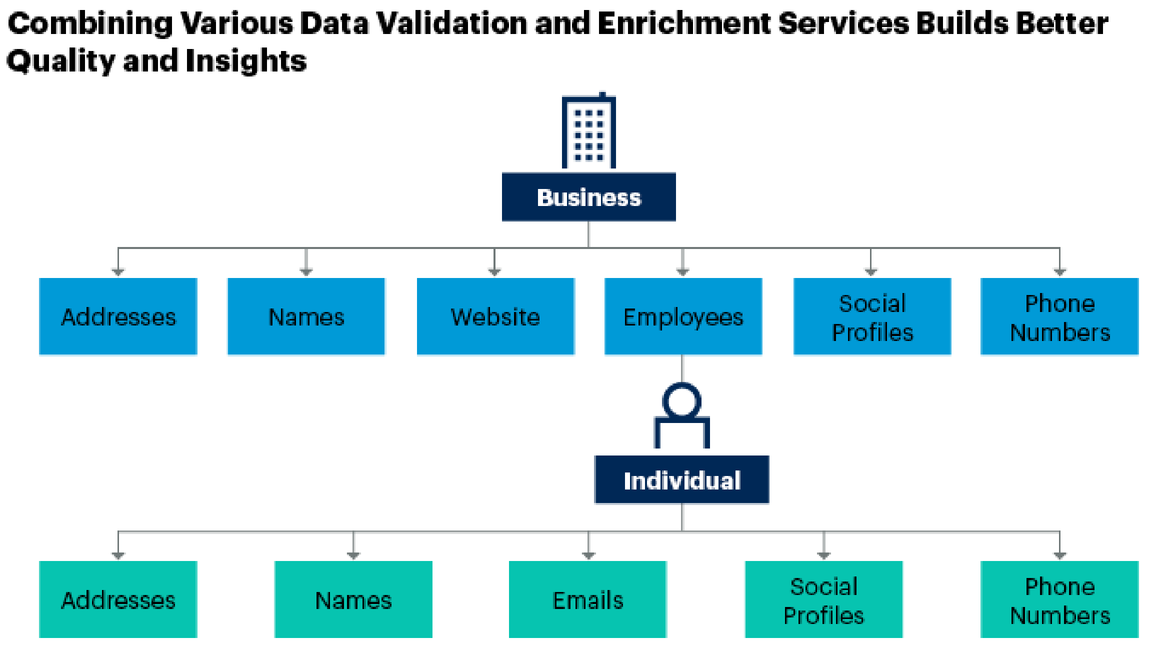
Data enrichment can help analytics professionals by:

- **Improving data accuracy.** Internal datasets may only provide a single view of a customer. Data enrichment tools can add missing data that can result in accurate feature creation. For more details on Feature Engineering, see Feature Stores for Machine Learning (Part 1): The Promise of Feature Stores.

- **Reducing the need to collect data.** Certain AI use cases, such as fraud detection, require access to data that users will be hesitant in submitting. For example, data enrichment tools can help by asking for email addresses and then comparing them with external databases to check for social media profiling and domain verification.

Figure 12 shows how data enrichment can be used to build enriched insights for individual data.

Figure 12: Data Enrichment for Better Insights



Combining Various Data Validation and Enrichment Services Builds Better Quality and Insights

Source: Gartner
736508_C

- **Postal addresses.** Enrichment services include address validation, look-up and geographical coverage for different formats. Addresses can be enriched with geolocation, business and property data. Common features in enrichment tools include verifying address information against authorized databases from postal authorities, converting zip codes to latitude and longitude coordinates, and providing auto-fill suggestions by detecting country and region.

- **Names.** Business entities can be validated by checking their registration status in states, countries and regions. This can help make sure that the company data being used is from legitimate ventures and not false fronts for corruption and money laundering. Features can include verifying the company's name and directorship status, registration number, physical address and enriching it with digital addresses, data of incorporation or counts of bankruptcy.

- **Emails.** Enrichment services can help validate if the emails are legitimate and deliverable, and exist with reliable domains. Features can include spell checks, checking email syntax, filtering for spams and autocomplete features.

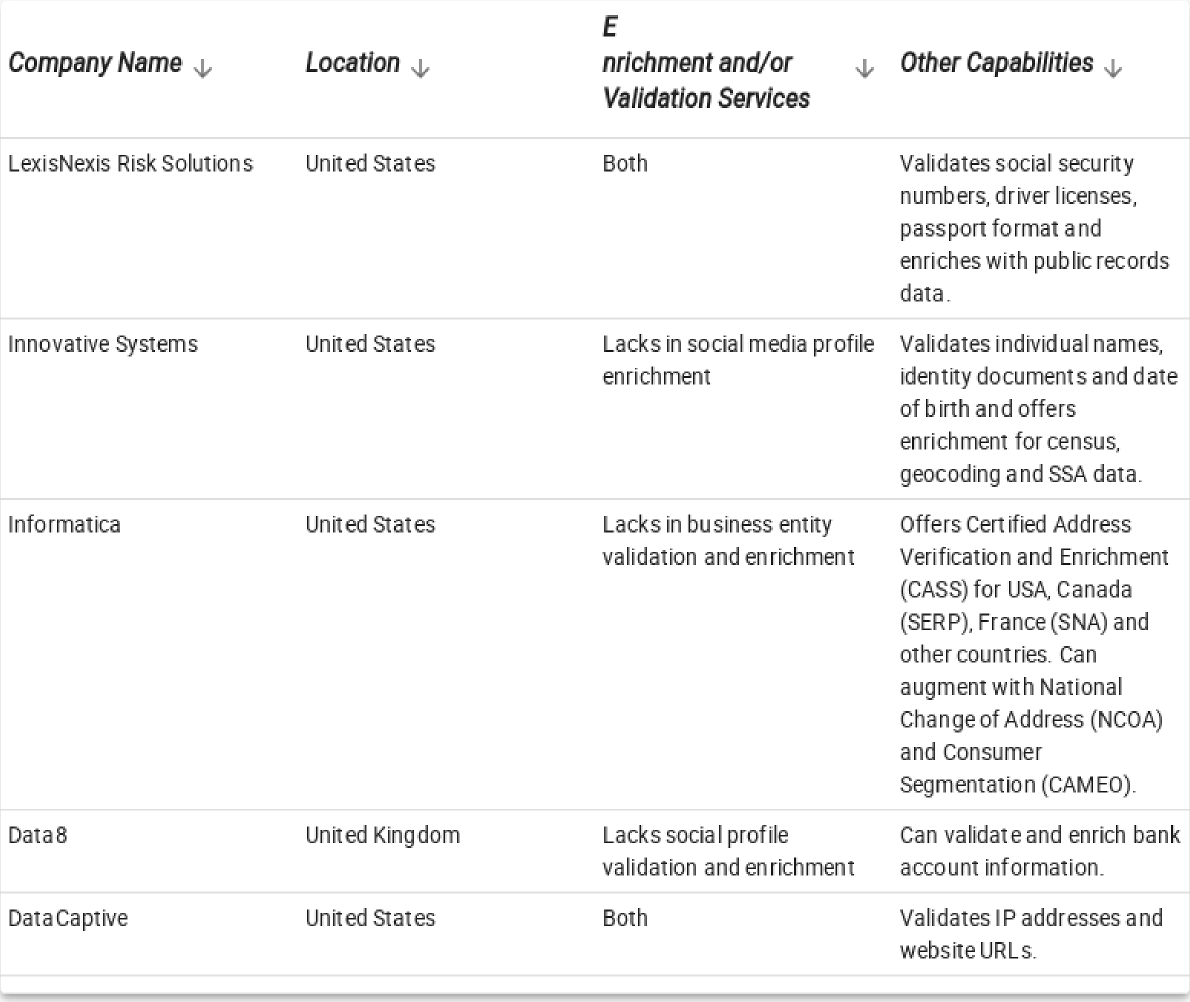
- **Social Profiles.** This can include validation social media profile URLs and is mainly used in social media analytics to provide accurate social insights, such as verifying names and titles of individuals and companies against various social networks.

- **Phone Numbers.** Enrichment tools help verifying if the number exists and is valid, as well as provide additional data such as line type (landline, mobile phone), carrier name and country. Other data enhancements can include associating the user's name and address associated with the number.

Table 8 shows a list of vendors that work within the data enrichment space. Note this is not an exhaustive list and covers vendors that offer data augmentation for most data attributes (as otherwise stated) discussed earlier. For more details, see Tool: Vendor Identification for Data Validation and Enrichment Services for Party Data.

Table 8: Sample Vendors for Data Enrichment and Validation Services

(Enlarged table in Appendix)



| Company Name ↓ | Location ↓ | E<br>nrichment and/or Validation Services ↓ | Other Capabilities ↓ |
|---|---|---|---|
| LexisNexis Risk Solutions | United States | Both | Validates social security numbers, driver licenses, passport format and enriches with public records data. |
| Innovative Systems | United States | Lacks in social media profile enrichment | Validates individual names, identity documents and date of birth and offers enrichment for census, geocoding and SSA data. |
| Informatica | United States | Lacks in business entity validation and enrichment | Offers Certified Address Verification and Enrichment (CASS) for USA, Canada (SERP), France (SNA) and other countries. Can augment with National Change of Address (NCOA) and Consumer Segmentation (CAMEO). |
| Data8 | United Kingdom | Lacks social profile validation and enrichment | Can validate and enrich bank account information. |
| DataCaptive | United States | Both | Validates IP addresses and website URLs. |

Source: Gartner (June 2023)

## Strengths

A shift toward data-centric AI can help data and analytics technical professionals and leaders in the following ways:

- Create greater efficiencies in AI development by focusing on improving and enhancing the data, rather than simply focusing on code and model tuning.
- Help lower compute, storage, resource, infrastructure and operations cost as code-centric tools require environment maintenance, compute elasticity and IT resources to manage and monitor the environments.
- Use low-code and self-serve data preparation tools that feature powerful AI-augmented capabilities, such as predictive transformations, data lineage, quality metrics and automatic feature extraction to help reduce the time to insights.

- Engage automated feature engineering techniques to reduce the manual labor, dependence on domain knowledge and increase the speed and accuracy of feature generation.
- Access external data enrichment tools to augment internal datasets and even open up avenues for data monetization, as internal datasets can be shared with external users. Data sharing is gaining speed and cloud data warehouse vendors like Snowflake now feature marketplaces to find and share curated data with both internal and external users.
- Generate high-quality training data with variety and balance with synthetic data, which can be useful for increasing the accuracy of predictions. Data scientists and analysts can reduce the dependence on internal and private datasets, as well as control the structure and format of data.
- Equip AI-assisted data labeling and annotation tools to greatly reduce the dependence on external and internal workforce and free up precious time for data scientists and analytics professionals. Most tools feature HITL for annotation validation to increase the quality of labeling.

### Weaknesses

While data-centric AI aims to shift focus toward data management for AI, there are challenges that data and analytics technical professionals and leaders might face:

- Synthetic data remains an emerging market, and the technology needs time to mature.
- Self-service capabilities can introduce large risks related to data governance, data quality, data loss, security and privacy. Incorrect data usage can result in false-positives or false-negatives and can allow mistakes to add up. There will be a need to balance for who has access to which data and in which context it is being used. Without a central IT team managing the platform, efficient guardrails on consumption and cost governance may require extra steps. A balance between control, offered by a central team against autonomy in decentralized teams, needs to be maintained.
- Level of skills can vary within analytics teams. Some data scientists might prefer a code-centric approach to data preparation and labeling, while others may prefer visual authoring. Carefully review code extensibility options while also considering a blended approach based on dataset variety and skill availability.

- Migrating from current to modern tools and methods may not always be smooth and can be costly. While newer tools are more cloud-focused and offer many new features, they may not replicate the same features as the older, on-premises offerings. Carefully assess the pros and cons of the traditional versus newer platforms.

- Accessing enrichment sources may result in access to PII data. Strong governance is required to ensure that privacy is maintained.

## Guidance

### Evaluate Maturity and Capabilities of Data-centric AI Solutions

Synthetic data generation, data preparation, annotation and enrichment tools have varying levels of maturity and capabilities, and should be carefully evaluated:

- **Synthetic data market is still immature:** According to Emerging Tech: Overcome Marketing and Go-to-Market Challenges for Tabular Synthetic Data, the greatest challenge currently faced by tabular synthetic data generation platforms is the low state of market maturity. This is due to constraints around privacy and accuracy — a perfectly accurate dataset will not be completely private, and vice versa. Analytics professionals should use their deep understanding of specific domains to help determine the trade-offs.

- **Supplement open source options with specialized tools:** AI development has been centered more around DSML and open source tools, but modern data preparation tools from the data management domain have powerful capabilities as well. These include data cataloging, lineage and monitoring, as well as predictive transformation capabilities. When evaluating data preparation tools also carefully consider support for diverse data sources, as well as data movement requirements. Carefully assess if the data will be moved out of the cloud into the platform's hosted cloud instance or can the transformation be pushed down to the data storage layer (see Working With Semistructured and Unstructured Datasets and Comparison of Data Stores to Support Modern Use Cases for more details).

- **Combine AI-augmentation with HITL:** While AI-augmented capabilities for data annotation and preparation are becoming more capable, verification and validation should still be the domain of data and analytics professionals.

- **Explore alternative methods for data enrichment:** While data enrichment tools offer access to curated datasets, analytics professionals should assess alternative enrichment methods such as web scraping, as well as exploring possibilities within the in-house tools. For example, if the enterprise is using Snowflake, they can explore its marketplace to find external datasets.

- **Ensure AI ethics compliance:** Generative modeling has made image, text, video and audio generation easier, but presents risks of abuse and false positives. Ensure that appropriate caution, risk mitigation and oversight measures are installed prior to using generative technologies. For more details, read Incorporate Explainability and Fairness Within the AI Platform.

- **Convergence of capabilities and the rise of data ecosystems:** Most data preparation and enrichment capabilities are converging into end-to-end platforms for data management and analytics. DSML platforms acquire competing platforms to augment their offerings. Examples include Alteryx acquiring Trifacta. In fact, data platforms are starting to offer ML-augmentation as well and are evolving into data ecosystems. For more information, read Innovation Insight: Data Ecosystems Will Reshape the Data Management Market. Analytics technical professionals should carefully evaluate the growth of the platforms and evaluate other capabilities being offered as part of the package.

### Conduct a Cost-Benefit Analysis

Opting for new tools and vendors to help with data augmentation, enrichment, annotation and data preparation can incur costs, while also reducing the need for dedicated data engineers and IT support. Data and analytics technical professionals and leaders should carefully evaluate the skills they are looking to automate, and the reduction in cost of doing so. While low-code/no-code tools will reduce the need for data engineers, they will still require data scientists to use the tool. However, the rise of the citizen data scientist role can help reduce the dependence on data scientists and help in democratization of data.

SureSparkle* conducted a cost-value comparison analysis to gauge the business value derived from self-serve tools against financial benefits, nonfinancial quantitative benefits and qualitative benefits. If the use case cannot demonstrate net value even after a collaborative comparison of cost and value, then SureSparkle would turn it down. For more details on this case study, read Case Study: Demand Management for Self-Service Data and Analytics Tools (SureSparkle*).

As opposed to commercial low-code platforms, open source, pro-code options can offer a greater degree of customization, but require more technical skills. However, most low-code platforms now also provide scripting options as well, providing appeal to both technical and nontechnical users.

## Implement Data Governance to Manage Self-Service

As proliferation of data and analytics tools increases, governance can become a problem. Data and analytics technical professionals will have to learn self-governance to ensure responsible use of their platforms — otherwise, IT dependence will continue and the full benefits of self-serve may not be realized. Successful management of self-serve analytics will emerge as a key challenge for data and analytics technical professionals — as well as leaders as data ecosystems and augmented analytics become common. Technical professionals should explore the different methods of governance, including a decentralized form of governance.

Decentralized governance involves a central unit that develops analytics governance policies, while the line of business carries out the enforcement. The central team could be the IT team or a center of excellence. Decentralized governance is suited to conglomerates that consist of several companies, large consulting firms and universities where colleges and departments operate independently.

However, the decentralized analytics governance approach can result in lack of standardization across functions. It can lead to redundancies, wasted license costs and a broader exposure to security risk. For more details, read Data and Analytics Governance Approaches for the Technical Professional.

## Acronym Key and Glossary Terms

| | |
|---|---|
| GANs | Generative Adversarial Networks |
| GPT | Generative Pretrained Transformer |
| BERT | Bidirectional Encoder Representations from Transformers |
| NLU | Natural Language Understanding |
| SSA | Social Security Administration |
| NLQ | Natural Language Querying |
| RPA | Robotic Process Automation |
| PII | Personal Identifiable Information |
| HITL | Human In The Loop |
| CLARANS | Clustering Large Applications based upon RANdomized Search |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| LOF | Local Outlier Factor |

## Evidence

**P-21023 2021 Gartner AI in Organizations Survey.** 2021 Gartner AI in Organizations Survey: This survey was conducted to understand the keys to successful AI implementations and the barriers to the operationalization of AI. The research was conducted online from October through December 2021 among 699 respondents from organizations in the U.S., Germany and the U.K. Quotas were established for company size and for industries to ensure a good representation across the sample. Organizations were required to have developed AI or intended to deploy AI within the next three years. Respondents were required to be part of the organization's corporate leadership or report into corporate leadership roles, and have a high level of involvement with at least one AI initiative. Respondents were also required to have one of the following roles when related to AI in their organizations. determine AI business objectives, measure the value derived from AI initiatives or manage AI initiatives development and implementation. The survey was developed collaboratively by a team of Gartner analysts and Gartner's Research Data, Analytics and Tools team.

Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

[1] Flying Blind: How Bad Data Undermines Business, Forbes.

[2] Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, TIME.

## Recommended by the Author

Some documents may not be available as part of your current Gartner subscription.

Overcoming Data Quality Risks When Using Semistructured and Unstructured Data for AI/ML Models

Incorporate Explainability and Fairness Within the AI Platform

Cool Vendors in Data-Centric AI

Working With Semistructured and Unstructured Datasets

Table 1: Data Preparation Processes, Steps and Techniques

| *Process Steps* ↓ | *Description and Techniques* ↓ | *Examples and Use Cases* ↓ |
|---|---|---|
| Exploratory Data Analysis | Should be performed to explore the general shape of the dataset, understand the quality regarding duplicates, missing values, unwanted entries and investigate features, and how they may relate to each other.<br>Examples include univariate nongraphical, multivariate nongraphical, univariate graphical and multivariate graphical methods. Nongraphical methods involve calculation of summary statistics and graphical methods involve charts. Univariate methods involve a single variable, while multivariates involve multiple variables. | Nongraphical methods include mean, median, mode, variance, skewness and kurtosis.<br>Graphical methods include bar charts, scatter plots, heat maps, bubble charts and run charts.<br>Combination of charts can result in great results (e.g., side-by-side boxplots are the best graphical technique examining the relationship between categorical and quantitative variables). Scatter plots can be overlain on box plots to provide additional information about distributional differences or similarities. |

| *Process Steps* ↓ | *Description and Techniques* ↓ | *Examples and Use Cases* ↓ |
|---|---|---|
| Data Preprocessing | Data preprocessing includes cleaning, shaping, handling missing values and adding quality to data prior to feature extraction.<br>Multivariate analyses, in the form of k-nearest neighbor and regression-based methods, such as multiple linear regression (MLR), support vector machines (SVM), decision trees and artificial neural networks (ANN) can be used for handling missing values. Clustering and outlier techniques using algorithms, such as CLARANS, DBSCAN and LOF can be used to identify anomalous data. | Univariate analyses should be used when the missing data ratio is small (between 1-5%) and generally are not recommended for time-series data. Multivariate analysis produces good results for higher missing ratios (~15%). and should be used for handling missing values over longer time intervals.<br>Clustering-based methods can be used as preliminary steps for identifying data clusters and then statistical-methods can be employed to determine outliers. However, clustering methods for larger datasets can be computationally expensive. |

Source: Gartner (June 2023)

### Table 2: Open Source Tools for Data Preparation and Exploration

| *Python Library* | *Use Case* |
|---|---|
| Sweetviz, D-Tale, Vega-Altair, Bokeh, Plotly, bamboolib, Pandas-Profiling, Data-Purifier | Used in exploratory data analysis to create visualizations and charts for data exploration, statistics gathering, data distribution and completeness. Low-code nature can enable citizen data scientists as well. Data-Purifier is more suited to NLP applications. |
| dabl, Yellowbrick, Dash, pandas_ml, Petl, Data Measurements Tool | Used in data processing and transformations, but can also assist in model training. Petl can be used for ETL operations. Data Measurements Tool, by Hugging Face, is a no-code tool to help build, measure and compare NLP datasets. |

Source: Gartner (June 2023)

### Table 3: List of Data Preparation Tools

| *Vendor Name* ↓ | *Tool(s)* ↓ | *Notes* ↓ |
|---|---|---|
| Alteryx | Alteryx Analytics Cloud Platform, Alteryx Designer Cloud (Trifacta) | Cloud-agnostic — recently acquired Trifacta. |
| Amazon Web Services | Amazon SageMaker Data Wrangler, Glue DataBrew and QuickSight | Amazon SageMaker Data Wrangler cannot be used outside of SageMaker, while Glue DataBrew can be used for ETL wrangling. |
| Microsoft | Azure ML, Synapse Pipelines, Power BI (Power Query, Dataflows, Datamarts) and Azure Data Factory | Azure Data Factory is the primary ETL/ELT tool with dataflows and Power Query for visual data preparation. |
| Google Cloud Platform | Dataprep, Cloud Data Fusion, Connected Sheets | Dataprep is powered by Trifacta. However, Alteryx recently acquired Trifacta — not currently known how Dataprep will be affected. |
| Informatica | Informatica Cloud Data Integration — Free and PayGo, Enterprise Data Preparation | On-premises and cloud-agnostic solutions; CDI is powered by CLAIRE, a custom developed ML engine for augmenting data preparation tasks; also features data quality and masking capabilities. |
| Dataiku | Data Science Studio (Dataiku Data Preparation) | Offers a range of processing engines, such as Apache Spark, SQL and an in-memory DSS engine. Some SQL-based transformations can be pushed down to the database's compute |

| *Vendor Name* ↓ | *Tool(s)* ↓ | *Notes* ↓ |
|---|---|---|
| DataRobot | DataRobot AI Platform (DataRobot Data Prep) | Data Prep is offered as a companion tool, as well and provides ML-embedded automation and NLP support as its core differentiators. |
| Modak Analytics | Modak Nabu | Nabu is primarily data management focused with capabilities such as data cataloging, data integration and orchestration. It can operationalize data preparation tasks through Apache Spark, Software AG (StreamSets), Apache NiFi, AWS Glue and Azure Data Factory. |
| Aible | Aible Sense | Aible Sense includes augmented features like assessing data readiness, automated data cleansing and table relationship determination. |

Source: Gartner (June 2023)

### Table 4: Automated Feature Engineering Techniques, Tools and Vendors

| *Techniques and Algorithms* ↓ | *Description* ↓ | *Use Cases* ↓ |
|---|---|---|
| Deep Feature Synthesis (DFS) | Used for automatically generating features for relational datasets. It is built around deriving features using dataset agnostic operations called "primitives" within multiple datasets in a stacking manner. Each time a stack is added, the "depth" of the feature increases. | Use case is for multitable and transactional and relational datasets mostly found in databases or log files. |
| Deep Learning | Convolutional Neural Networks (CNNs) are designed to process multidimensional data and include automatic feature extraction during the training process.<br>Other methods include Wavelet scattering and Autoencoders. Wavelet scattering networks automate the extraction of low-variance features from real-valued time series and image data.<br>Autoencoders are used in dimensionality reduction and feature extraction and consist of an Encoder function that tries to copy input to output, and the Decoder attempts to reconstruct the input from the output. | CNNs are primarily used for tasks related to image classification and object detection.<br>Autoencoders are primarily used to reduce noise in data and can work with images, tabular and time-series data.<br>Wavelet scattering is mostly used for image and time-series data.<br>CNNs can be more computationally intensive than autoencoder networks, due to the use of convolutional layers and the need to process large amounts of data. |

Source: Gartner (June 2023)

### Table 5: Data Labeling Approaches

| ***Data Labeling*** ↓ | ***Manual or Automated*** ↓ | ***Description*** ↓ | ***Pros*** ↓ | ***Cons*** ↓ |
|---|---|---|---|---|
| Internal Labeling | Manual | Human experts, usually data scientists label data | High accuracy and data security since the data scientists or SMEs understand the needs of the model. Data does not leave the internal systems and lowers the risk of data leakage. | Expense, time consuming and lacks flexibility in scaling resources. Data scientists are valuable resources and spending time on mundane labeling can significantly increase project costs. Alternative can be to hire cheap labelers, but the quality of work may deteriorate. |
| External Labeling | Manual | Outsourced labeling where dedicated vendors provide labeling services. Crowd-based sourcing is slightly different from outsourcing, as it is targeted toward unorganized labelers. | Flexibility in scaling resources, lower cost of labor and time savings. Labeling tasks can be scaled up while experienced vendors like Appen, Upwork and Amazon Mechanical Turk. | Difficulty in finding expert vendors and potential data security risks and assessing the quality of the labeled data. External labeling can cause vendors to use other companies' data to train their own models for labeling. |

| *Data Labeling* ↓ | *Manual or Automated* ↓ | *Description* ↓ | *Pros* ↓ | *Cons* ↓ |
|---|---|---|---|---|
| Semisupervised Learning | Automated | Combines supervised and unsupervised learning to label data using only a small dataset. These labels are called proxy labels and, if they satisfy the labeling criterion, are added back to the training dataset to retrain the model. Examples include Tri-Training and Active Learning. | Saves time and cost since a smaller amount of manually labeled training data is needed. Can result in better accuracy over time as active learning involves human feedback to improve labeling. | Initial dataset is a small subset and may miss labels that represent data outside the selected sample, but within the dataset.<br>Can cause future errors if any subset is incorrectly predicted initially. |
| Transfer Learning | Automated | A pretrained ML model is used to label the data provided it has been trained on a similar dataset that requires labeling. | Saves time and increases the model's learning rate and accuracy. | Final model may perform worse than the initial model if the training data required for the new task is different from the original problem. This is called Negative Transfer. |

Source: Gartner (June 2023)

Table 6: Tools for Synthetic Data

| *Synthetic Data* ↓ | *Vendors* ↓ | *Open Source* ↓ |
|---|---|---|
| Relational | Gretel AI, MOSTLY AI, Synth, Synthesized, IBM, Hazy, Facteus | Synthetic Data Vault, DataSynthesizer, Pydbgen, Mimesis |
| Nonrelational | MOSTLY AI, Bitext, EEdgecase.AI, Rendered.AI, Scale AI, Synthesis AI | zpy, genalog, Kubric, Blender |

Source: Gartner (June 2023)

Table 7: Synthetic Data Generation Techniques

| *Synthetic Data Generation Technique* | *Description* | *Pros and Cons* |
|---|---|---|
| Statistical Techniques | Data samples are generated from real probability distributions (normal, chi-square, exponential, etc.) resembling a natural phenomenon with certain characteristic statistical features like mean, variance and standard deviation. These are then used to generate similar factual data. Examples include the Monte Carlo method. Other techniques include masking, coarsening and mimicking. | Pros: Relatively easier to implement.<br>Cons: mostly used for generating tabular data and struggle with complex relationships. |
| Variational Autoencoders (VAEs) | Generative models that learn the underlying distribution of original data. They work in a double encoder-decoder transformation where an encoder compresses the original dataset into a compact structure and transmits data to the decoder, which generates an output as a representation of the original dataset. | Pros: Excel at tabular (structured) use cases.<br>Cons: Struggles with heterogeneous data (e.g., categorical, binary and continuous and image quality is lower than Diffusion or GAN models). |
| Generative Adversarial Networks (GANs) | Involve two neural networks, generator and discriminator, working in an adversarial fashion. Generator takes random sample data and generates a synthetic dataset, while the discriminator compares this synthetically generated data with a real dataset based on conditions set before. | Pros: Effective at generating images and understanding complex relationships in time-series data. Also self-regulates since the discriminator learns from patterns and the generator learns to outsmart it by producing more realistic samples.<br>Cons: Challenging to train and require technical expertise. |

| *Synthetic Data Generation Technique* ↓ | *Description* ↓ | *Pros and Cons* ↓ |
|---|---|---|
| Transformers | Language models, based on Transformer architectures, learn the underlying probability distributions of the training data, such as sequence of words called tokens and sample new data from learning these distributions. They effectively predict the next token or sequence of words. Examples include Transformers such as BERT, GPT-3 and DALL·E. | Pros: very effective in NLP applications for unstructured data.<br>Cons: can be very costly from a resource perspective and require considerable research since there are a host of models in the market. |
| Diffusion Models | Are generative models that work by adding noise to data and using neural networks to denoise or reconstruct the image. Popular examples include Stable Diffusion | Pros: offer text-to-image generation capabilities and have applications in visual prompting for computer vision.<br>Cons: relatively new technology and there have been privacy concerns on training using copyright data. |

Source: Gartner (June 2023)

Table 8: Sample Vendors for Data Enrichment and Validation Services

| *Company Name* | *Location* | *E<br>nrichment and/or Validation Services* | *Other Capabilities* |
|---|---|---|---|
| LexisNexis Risk Solutions | United States | Both | Validates social security numbers, driver licenses, passport format and enriches with public records data. |
| Innovative Systems | United States | Lacks in social media profile enrichment | Validates individual names, identity documents and date of birth and offers enrichment for census, geocoding and SSA data. |
| Informatica | United States | Lacks in business entity validation and enrichment | Offers Certified Address Verification and Enrichment (CASS) for USA, Canada (SERP), France (SNA) and other countries. Can augment with National Change of Address (NCOA) and Consumer Segmentation (CAMEO). |
| Data8 | United Kingdom | Lacks social profile validation and enrichment | Can validate and enrich bank account information. |
| DataCaptive | United States | Both | Validates IP addresses and website URLs. |

Source: Gartner (June 2023)