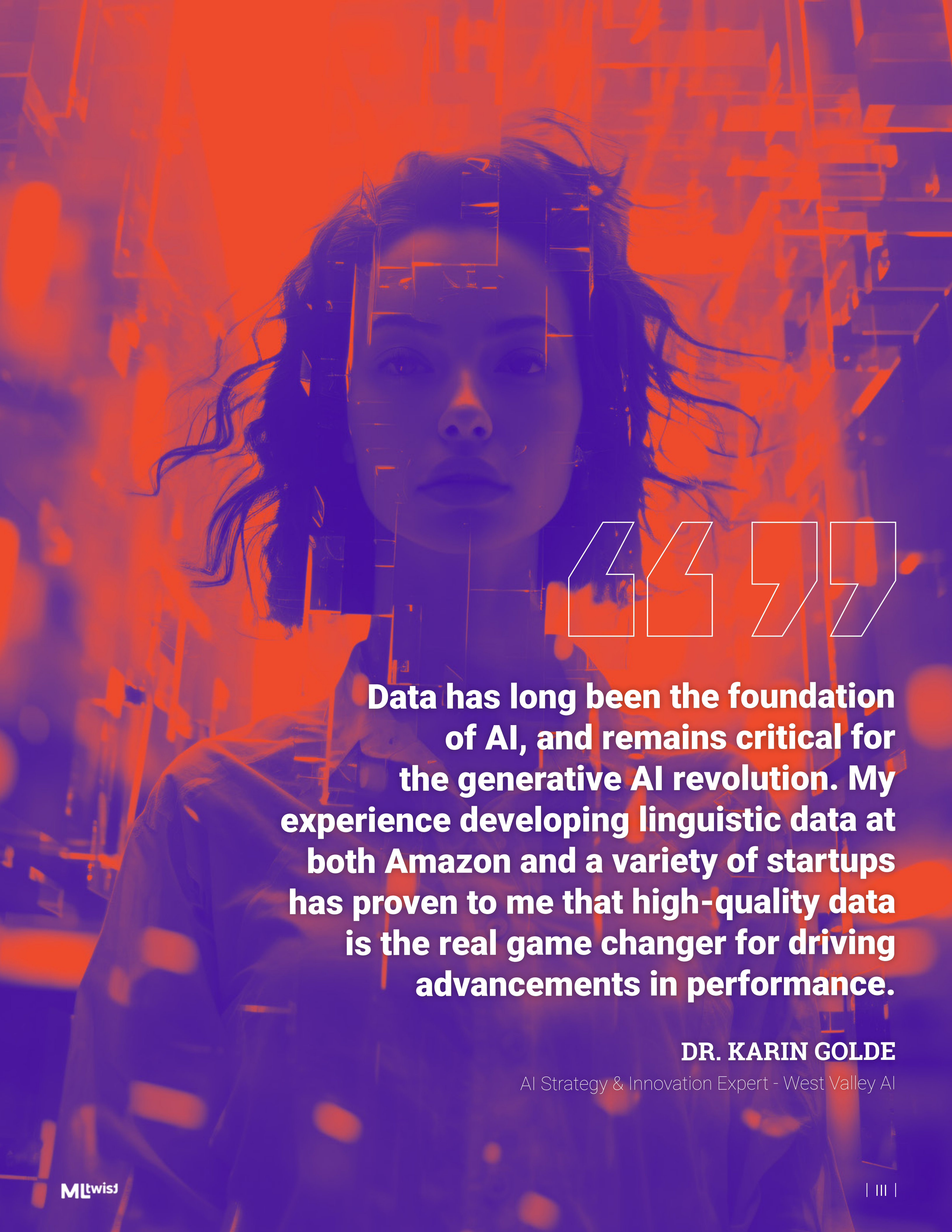
MLtwist

AI DATA PIPELINES

The Ultimate Guide

2024 Edition

Contents



> Data has long been the foundation of AI, and remains critical for the generative AI revolution. My experience developing linguistic data at both Amazon and a variety of startups has proven to me that high-quality data is the real game changer for driving advancements in performance.

DR. KARIN GOLDE

AI Strategy & Innovation Expert - West Valley AI

Introduction

The AI transformation has just begun. Already a $200 billion global industry, artificial intelligence is projected to reach $400 billion by 2025 and keep growing from there, revolutionizing whole sectors of the economy in the process. Its beating heart is data, and huge amounts of it. All AI models require a foundation of high-quality, high-volume, well-annotated data from which to learn with high confidence.

To get all this data into production AI models, businesses need data pipelines. They're the core plumbing that ships the data across a complex workflow of tools and services that transform it from raw to AI ready. This is true regardless of what kind of AI model your team is working with: supervised or unsupervised, traditional machine learning or natural language processing. The exact details of the workflow may change, but the data still has to be made AI ready. The AI data pipeline is always mission critical.

As the enterprise expands its use of AI, the problem becomes exponentially more complex. Each AI model requires a new pipeline for training, and without proper control and automation during pipeline creation, maintenance becomes a major risk for new projects.

Unfortunately, traditional ETL pipeline platforms and tools don't provide value in the AI world. These legacy tools shuttle structured data back and forth, but they don't perform any of the deep structuring, translation or asynchronous, bidirectional updating that AI models require. Due to these limitations, creating and managing pipelines for AI is highly manual and labor-intensive, with few ready-made tools for pipeline assembly and maintenance. As a result, up to 80% of a data scientist's time is spent on AI data pipeline creation and maintenance. Moreover, Gartner estimates that 80% of AI projects never reach deployment, partly because of issues in obtaining and processing sufficient amounts of high-quality data.

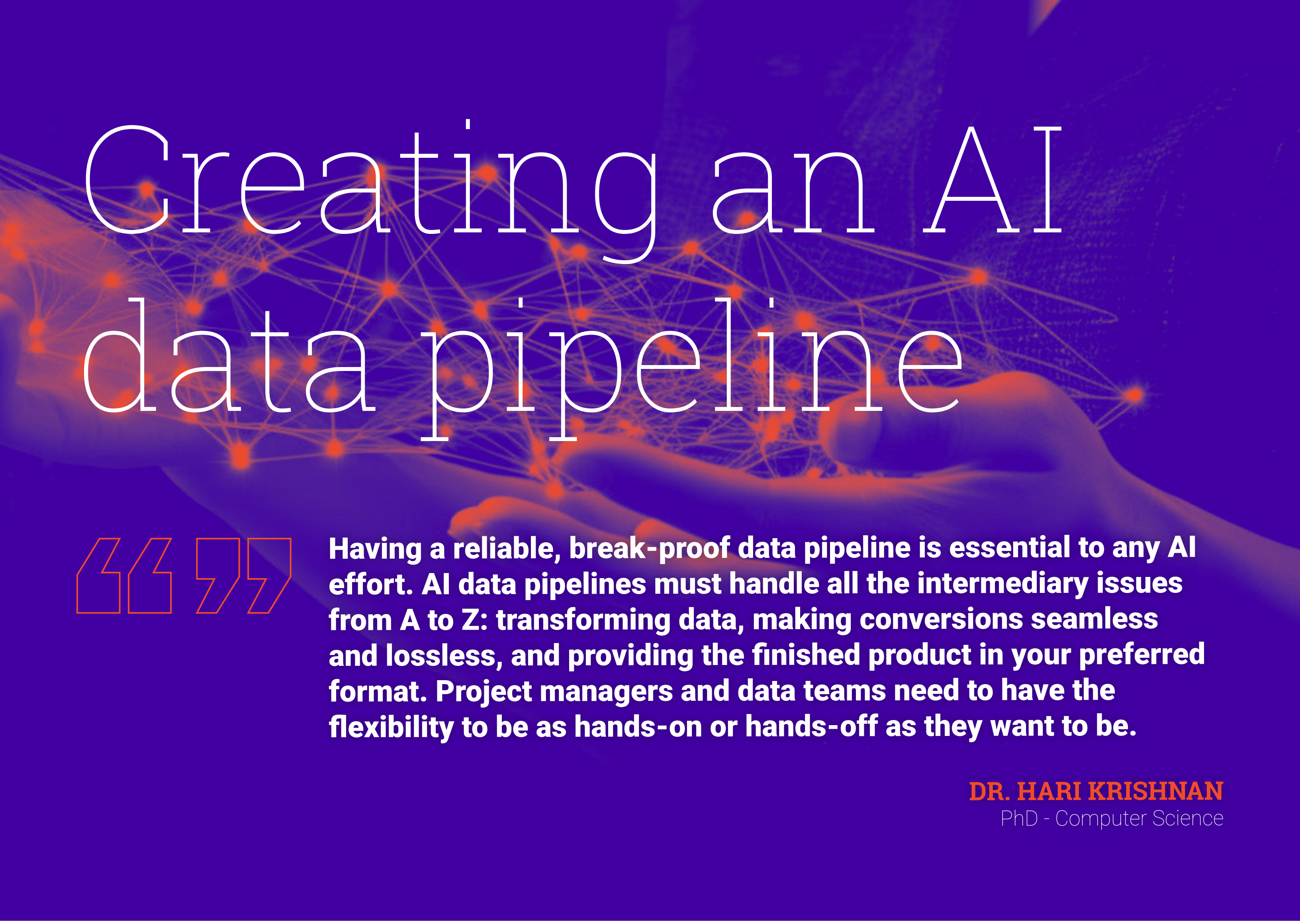
This Ultimate Guide gives executives as well as product, operations, engineering, and data science leaders a strong understanding of what AI data pipelines are, how they transform unstructured raw data to model-ready structured and annotated data, and which issues to consider when establishing a data workflow. You will get a clear sense of the steps involved in preparing data for models, the nature and volume of data preparation undertaken by your data science team, and where your AI project costs are going. Understanding this entire process will allow you to make informed decisions on how to deploy and manage AI data pipelines in your organization.

Why is this so hard?

> **Years of enterprise experience have taught me that models are only as good as their data. Data science teams spend on average over half their time cleaning and preparing data for processing, using fallible processes that impact project delivery and team morale.**
>
> **AVI ZUREL**
> Director of Infrastructure - Hippo Insurance
> Distinguished Engineer & Startup Advisor

The main difference between traditional enterprise data management and data management for AI is that AI workflows consume primarily unstructured data. Supervised AI models typically cannot usefully consume unstructured data in its raw form. This raw data needs processing in order to become readable and therefore useful to an AI model.

Unstructured data includes many different data types, including audio, video, image and text. A few examples are business documents, website pages, emails, social media feeds, 3D images, video, and more. According to Gartner, over 80% of enterprise data driving the AI revolution is unstructured. This number increases every year with the International Data Corporation predicting that unstructured data will grow to 175 zettabytes by 2025, with 30% of that data generated in real time.



Creating an AI data pipeline

> **Having a reliable, break-proof data pipeline is essential to any AI effort. AI data pipelines must handle all the intermediary issues from A to Z: transforming data, making conversions seamless and lossless, and providing the finished product in your preferred format. Project managers and data teams need to have the flexibility to be as hands-on or hands-off as they want to be.**
>
> **DR. HARI KRISHNAN**
> PhD - Computer Science

An AI data pipeline is the software infrastructure that moves raw data from its source to a target – a tool and/or service creating a workflow – so that it can be transformed into data usable by your model. When the data processing workflow is complete, the pipeline makes the data available to the model for production use. It also provides a number of important accountability functions, ensuring that your data processing is auditable, debuggable and compliant with the appropriate regulations and compliance/ethical standards.

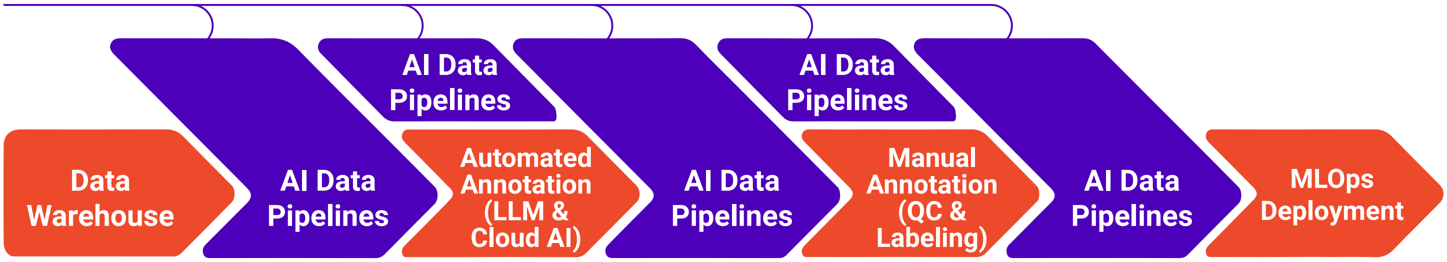


Illustration: AI Data Pipeline

> **When it comes to developing AI, one of the challenges that frequently crosses my radar is data. At first glance, pipelines seem simple. However, going even one layer in has shown us the dozens of different things that must go right in an AI data pipeline in order to deliver high quality AI.**
>
> **TED PRINCE**
> Group Chief Product Officer - Kantar

Today, at this early stage of the industry, AI data pipeline creation is complex. Data teams including data scientists, data operations teams, and decision makers must navigate a bewildering array of formats, workforces, and tools, each with their own features and limitations. We take you through each step in the process of creating and maintaining these pipelines.

Assembly & Ontology

Defining what data to include and how to categorize it

Before building a pipeline, your data scientists define the data requirements for the model and assemble a sufficient amount of data that meets those requirements. This data should be available from a single location and adhere to a common, well documented ontology. An ontology is a set of common definitions by which the data is logically structured. These include the features that the model is given to digest, as well as the labels and allowable label values for it to evaluate and populate. Data ontologies are often composed of multiple hierarchical tiers.

Ontology for a Data Labeling Project - Dog Breeds:

- **Dog Breeds Detection:**
  - Sporting Group:
    - Labrador Retriever
    - Golden Retriever
    - Pointer
  - Hound Group:
    - Beagle
    - Basset Hound
    - Bloodhound
  - Working Group:
    - German Shepherd
    - Rottweiler
    - Doberman Pinscher
  - Terrier Group:
    - Jack Russell Terrier
    - Yorkshire Terrier
    - Scottish Terrier
  - Non-Sporting Group:
    - Poodle
    - Bulldog
    - Dalmatian
  - Herding Group:
    - Border Collie
    - Australian Shepherd
    - Shetland Sheepdog
    - Shetland Sheepdog dog breed

Ontology for a Data Labeling Project - Horse Breeds:

- **Root: Horse Breeds Detection:**
  - Warmbloods:
    - Thoroughbred
    - Hanoverian
    - Dutch Warmblood
    - Westphalian
  - Coldbloods:
    - Percheron
    - Belgian Draft
    - Clydesdale
  - Gaits and Specialty Breeds:
    - Arabian
    - Morgan
    - Tennessee Walking Horse
    - Icelandic Horse

ATTRIBUTES:

Coat Color: Bay, black, chestnut, grey, palomino, pinto, appaloosa

Markings: Blaze, socks, stars, stripes

Build: Compact, cob-type, draft, light, muscular, slender

Height: Miniature, pony, horse, draft

Head shape: Straight, dished, Roman nose

Temperament: Energetic, docile, playful, spirited, calm

Having worked on several pioneering AI models, I am often reminded of the complexity involved in working with different types of data. The world ahead is multi-modal and technology which supports images, text and audio in several different data formats.

ANDREW COX

R&D Systems Analyst - Sandia National Laboratory

Pre-Processing

> **The world ahead is multi-modal and technology which supports images, text and audio in several different data formats. This is easy to say but difficult to implement, multi modal is not just popular data formats and as AI progresses, it pushes strongly into powerful formats like DICOS & HDF.**

EDWARD JIMENEZ
Principal R&D Optical Scientist and Engineer - Sandia National Laboratory

Making data clean, formatted, compliant, and auditable

Prior to any annotation or labeling of data, the pipeline needs to support several actions.

1. **STRUCTURE**
   Structuring your data files provides important cues that point your model in the right direction. For example, without structure a text file is simply a long sequence of characters. It's important to provide structural definition to the document which defines the headers, text body, tables, footnotes and other elements. By default, text files like PDFs are unstructured and require the data scientist to define this for the model. This is also important in other file formats, for example in defining ads (which can be safely ignored) within a video or audio file. In most cases this is done manually by your data science team (a long and labor-intensive task), but there is an emerging sector of automated services that help speed up this process. When using such services, the AI data pipeline needs to point the data to this service bidirectionally at the appropriate point in the workflow.

2. **HYGIENE**
   Most data needs to be cleaned before being further annotated and/or fed to an AI model. This includes accounting for missing fields, checking for obvious value inaccuracies and corrupt files. Today this is a highly manual process requiring a significant effort from data scientists.

3. **FORMAT**
   AI data comes in a wide array of formats encompassing text, image, video, image, audio and beyond. It's often necessary to convert all or a portion of your library of data files to a new or common format prior to their onward transmission, for example from PDF to DOC or MPG to WMV. There are existing services that perform these file conversions, and the AI data pipeline must be configured to get the files to the conversion service and back.

4. **DIGESTIBILITY**
   In order for large files and complex file formats to be properly reviewed, annotated and evaluated by models, it's often necessary to perform file format conversions, for example dividing a video file into frames or subsets of frames, a 3D image into a set of 2D images, and so forth. These requirements vary depending on the human review tool and other services being used to annotate the data.

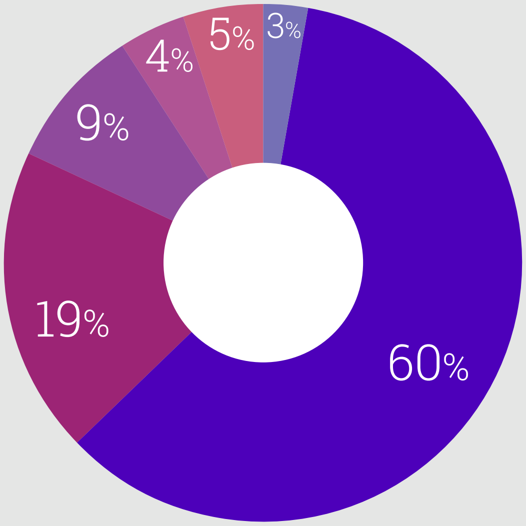
5. **JSON TRANSFORMATION**
JSON (JavaScript Object Notation) is a file standard that's used to communicate the ontology and field values for data files from one service to another. However, JSON is a generic file type and each data labeling platform requires the JSON files to be marshaled into a specific AI standard format. Some popular AI standard JSON formats are COCO, Pascal, VOC and YOLO. Given that the AI data industry is so new, the AI standard formats are often loosely followed by data labeling platforms and often require further processing to accommodate proprietary adjustments to the AI standard formats. This means that the data scientist who is creating the pipeline must create a new JSON file for each point in the workflow to accommodate the various data labeling platforms which they plan to utilize.

6. **PRIVACY REGULATIONS & STANDARDS**
There are a large number of global laws that apply to data files used to train AI models. The data science team needs to be aware of them and understand which ones apply to their project. Where they do apply, the team often needs to perform pre-processing to the data to comply. For example, the data may need to be sent to an anonymization service that strips out PII values prior to annotation, or a service that renders the data HIPAA compliant for health care use cases. If so, this involves more configuration work for the AI data pipeline. In addition, the environment in which the data is processed may need to comply, for example it may need to be located in a particular geographic location. Similar efforts need to be made in order to comply with industry standards around security such as SOC2 or IL.

7. **RESPONSIBLE AI**
As AI expands across all industries, companies need to ensure that their data use is ethical. Having traceability from the data sourcing phase to the training datasets helps reduce data bias and will encourage all actors to be more cautious on the how, the who and the what of AI to make sure technology is developed ethically and responsibly.

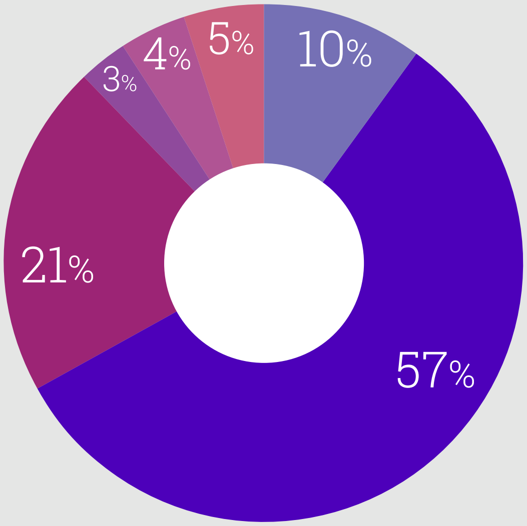
8. **AUDITABILITY**
The data scientists also need to provide sufficient auditability of the data as it is processed in order to be accountable for all changes. This involves creating data cards, which track every step in the process of data transformation, including which technologies and companies had access to the data. Today, data cards are uncommon among data vendors. Their usage will increase in the near future as regulations and AI ethical standards evolve to protect B2B and B2C users.

> **Data scientists spend approximately 80% of their time preparing data, 76% of them say it is the least enjoyable part of their work**



How Data Scientists Spend Their Time

- ■ Building training sets: 3%
- ■ Cleaning and organizing data: 60%
- ■ Collecting data sets: 19%
- ■ Mining data for patterns: 9%
- ■ Refining algorithms: 4%
- ■ Other: 5%



The Least Enjoyable Part of Data Science

- ■ Building training sets: 10%
- ■ Cleaning and organizing data: 57%
- ■ Collecting data sets: 21%
- ■ Mining data for patterns: 3%
- ■ Refining algorithms: 4%
- ■ Other: 5%

*Source

> Quality data is crucial for AI models and applications. It’s essential that this data is ethically sourced and responsibly managed. The importance of data ethics in AI for our future cannot be understated, and it’s likely to be increasingly regulated and subjected to third-party audits. Understanding your data’s origin, its access history, and its management is fundamental. Developing AI data pipelines that not only meet ethical standards but also align with upcoming legal requirements is vital for sustainable progress.

LAKE DAI

Adjunct Professor, Applied AI - Carnegie Mellon University

Annotation

As a Data Operations person, it is really cumbersome to find the right workforce and the right tool to deliver data annotations to the ML/Data science teams for each use case I'm working on. Having AI data pipelines that are connecting all the components I need, that are not breaking and can pivot as my needs evolve help me create high quality annotations at scale. This is crucial to excel at the 3 success metrics we measure on a daily basis: quality, speed and cost.

MARLEY JONES
Corporate VP, Data Capabilities, Generative AI - Insurance Industry

Labeling the data to train the model

Annotation of AI data files is traditionally performed by human raters before feeding the data into the model. In some cases, pre-annotation services offer automated annotation in order to speed the labeling process. These automated services typically require subsequent human rater review as well and have different success rates depending on data types and guidelines requirements. Pre-labeling can also be counterproductive, sometimes creating additional manual work to increase data annotation quality.

Preparing data for annotation is a labor intensive process. Data scientists must carefully choose:

1. **The human review tool** based on the type of annotations it allows, the file formats it supports, the allowable size of ontology, its ability to support concurrent work on the same datasets, JSON formats, ease of API integration and whether data is stored for review on a local machine or the cloud. Additional evaluation points are its level of optimization and efficiency, whether it puts the right signals in place at the right time in the review UI, whether reviews are pushed or pulled by the rater, whether there is reporting transparency on rater work, and cost.

   Every tool has its own strengths and weaknesses that will need to be audited and assessed. This process can take up to 3 months as every tool will be compared on ease of use, features offered, cost and flexibility. Once the tool is selected for a particular project, the user will commit to an annual license, most of the time. This locks the user's budget, sometimes forcing them to use an inadequate tool for their subsequent annotation needs.

Business stakeholders and data operations teams also need to understand the data rights implicit in any agreement with such a tool. Some tools, for example, claim data rights for any files they process, which might be used in the service of a competitor.

2. **The human review workforce** based on the required area of expertise, language support, cost, speed, flexibility, potential for workforce bias, conformity with local regulations and laws, location, and quality of ratings. Finding the right workforce solution can be a long and tedious process.

   As we enter a new AI phase with the adoption of Generative AI, many projects require human review workforces with specialized expertise in a particular area. If, for example, you are building a large language model (LLM) that can create computer code based on user queries, the workforce needs to be able to read and evaluate code in that language.

   Companies developing AI models choose between three workforce models, each with pros and cons:

   - An in-house team of annotators is a group of full time employees or contractors that work directly for the company developing the AI model. They are specifically trained and sometimes have a background that matches the industry the company belongs to. This is a very effective solution to have very highly skilled annotators that will be considered as domain experts. The downsides of in-house are high cost and an inability to quickly scale if an abnormally high volume of annotations is suddenly required quickly.

   - Crowdsourcing means accessing a group of individuals who work on demand on data annotation projects via services like Amazon Mechanical Turk. This can work when the use case is simple and requires no specialized knowledge. Cost is low and the effort can scale quickly. The downside is a lack of training, which can negatively impact annotation quality.

   - The third option is to contract with a workforce company. These are service companies that hire hundreds of annotators that will be trained on a specific project on demand. Workforce companies hire raters based on knowledge needs and train them until they reach a targeted level of quality. This allows their customers the flexibility to ramp volume up and down easily. Cost is typically higher than crowdsourcing but significantly lower than in-house.

   Once the tool and workforce have been selected, the data scientist needs to load the data files for review. Today this requires manual coding to account for JSON format, API requirements, API throttling (e.g. number of requests), file size limitations, and other limitations such as frame rate for video, audio time, and number of pages, among others. These limitations are not only time consuming but can also cause a cost increase if not handled properly.

   In parallel, the data operations team needs to define a policy with clear instructions to the rater workforce on how to conduct the review. This requires the establishment of a virtual cycle, where data ops train and test the workforce before they enter production to ensure that every single use case will be covered by the guidelines to minimize bias and/or errors that could hurt model performance.

   Once this infrastructure is in place, the raters enter the review tool and label batches of data according to the finalized policies. This labeling provides the ground truth on which the model will learn. For example, the rater may provide bounding boxes to detect a face or an object in a physical environment within video or 3D image frames, true/false values, etc.

Customer 1: AI Data Pipeline 1

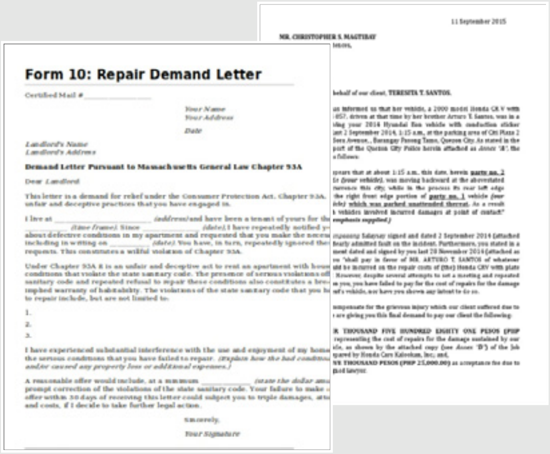
Extract 13 fields from different data sources (remittances, checks, freight, contracts, etc.)

Data type: Scans and Exports

Type: PDF, DOCX

Tool: Datasaur.ai

datasaur.ai



Form 10: Repair Demand Letter

Certified Mail # ________________

Your Name
Your Address

Date

Landlord's Name
Landlord's Address

Demand Letter Pursuant to Massachusetts General Law Chapter 93A

Dear *Landlord*

This letter is a demand for relief under the Consumer Protection Act, Chapter 93A.
unfair and deceptive practices that you have engaged in.

I live at ____________________ *(address)* and have been a tenant of yours for the
__________ *(time frame)*. Since ____________ *(date)*, I have repeatedly notified y
about defective conditions in my apartment and requested that you make the necess
including in writing on ________ *(date)*. You have, in turn, repeatedly ignored the
requests. This constitutes a willful violation of Chapter 93A.

Under Chapter 93A it is an unfair and deceptive act to rent an apartment with hous
conditions that violate the state sanitary code. The presence of serious violations of
sanitary code and repeated refusal to repair these conditions also constitutes a bre
implied warranty of habitability. The violations of the state sanitary code that you h
to repair include, but are not limited to:

1.

2.

3.

I have experienced substantial interference with the use and enjoyment of my home
the serious conditions that you have failed to repair. *(Explain how the bad condition*
and/or caused any property loss or additional expenses.)

A reasonable offer would include, at a minimum __________ *(state the dollar amo*
prompt correction of the violations of the state sanitary code. Your failure to make
offer within 30 days of receiving this letter could subject you to triple damages, atto
and costs, if I decide to take further legal action.

Sincerely,

Your Signature

11 September 2015

MR. CHRISTOPHER S. MAGTIBAY
lenos,

behalf of our client, **TERESITA T. SANTOS.**

as informed us that her vehicle, a 2000 model Honda CR V with
-857, driven at that time by her brother Arturo T. Santos, was in a
ding your 2014 Hyundai Eon vehicle with conduction sticker
last 2 September 2014, 1:15 a.m., at the parking area of Citi Plaza 2
Sora Avenue., Barangay Pasong Tamo, Quezon City. As stated in the
port of the Quezon City Police herein attached as *Annex "B"*, the
s follows:

ppears that at about 1:15 a.m., this date, herein **party no. 2**
(your vehicle), was moving backward at the abovestated
rameus this city, while in the process its rear left edge
the right front edge portion of **party no. 1** vehicle *(our*
icle) **which was parked unattended thereat.** As a result
h vehicles involved incurred damages at point of contact"
emphasis supplied.)

mpaunay Salaysay signed and dated 2 September 2014 (attached
learly admitted fault on the incident. Furthermore, you stated in a
ment dated and signed by you last 28 November 2014 (attached as
ou "shall pay in favor of MR. ARTURO T. SANTOS of whatever
uld be incurred on the repair costs of (the) Honda CRV with plate
However, despite several attempts to set a meeting and repeated
m you, you have failed to pay for the cost of repairs for the damage
ent's vehicle, nor have you shown any intent to do so.

compensate for the grievous injury which our client suffered due to
are giving you this final demand to pay our client the following:

R THOUSAND FIVE HUNDRED EIGHTY ONE PESOS (PHP
representing the cost of repairs for the damage sustained by our
icle, as shown by the attached copy (see *Annex "D"*) of the Job
pared by Honda Cars Kalookan, Inc.; and,
VE THOUSAND PESOS (PHP 25,000.00) as acceptance fee due to
ged lawyer.

Customer 1: AI Data Pipeline 2

Create bounding boxes on vehicles, boats & persons

Data type: Aerial Images

Type: JPG, PNG

Tool: Kili

KILI TECHNOLOGY

Customer 1: AI Data Pipeline 3

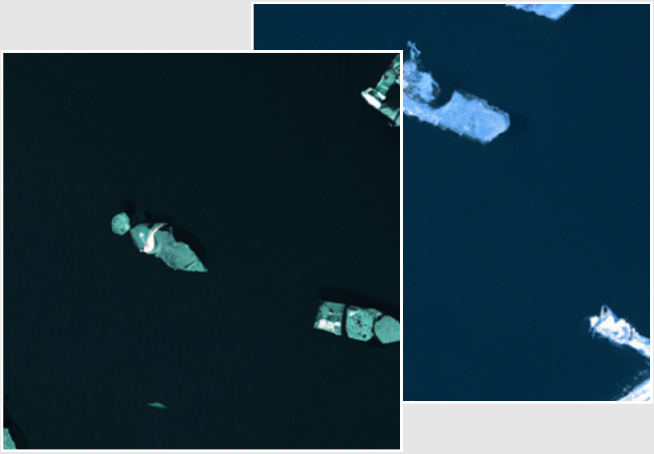
Create oriented bounding boxes on vehicles, boats & persons

Data type: SAR imagery

Type: PNG, TIF, JPG

Tool: Dataloop

Dataloop



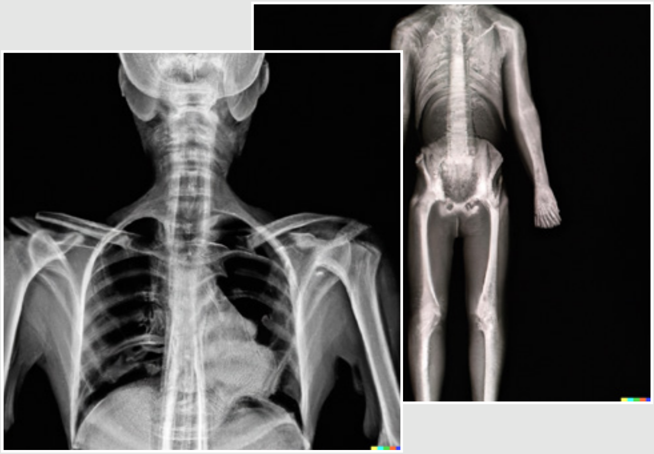
Customer 1: AI Data Pipeline 4

Create polygons at the voxel level on objects

Data type: 3D scans

Type: Dicom

Tool: Encord



Again, proper configuration of the AI data pipeline is necessary for transmitting newly labeled data back to its original location, with versioning and data cards for future debugging and auditability.

> **According to Gartner, only 4% of organizations reported that their data is AI-ready. Data pipelines are crucial for businesses of all sizes to unlock their unique, unstructured data for AI adoption.**

Quality Control

Evaluating data labeling results at scale

Once the rater workforce completes the initial set of ratings, the data ops team must review the data manually to identify any problems at scale. This includes any strange, inappropriate strings in the data, instances where a single rater has made a mistake at scale across multiple data files (for example, switching right and left arms in an image frame), and a wide variety of other potential challenges. Ideally, there should be an automated way to capture these mistakes at scale. However, today that's usually not the case.

There are multiple methods to measure data quality, for example:

- Send the same file to multiple reviewers and assess consensus of rating
- Send a "golden set" of pre-reviewed files to all reviewers and assess accuracy per rater

If the selected rater tool does not have its own quality control features, Data Ops will need to extract the data from the tool and convert it to a readable format before assessing quality, then retrain the workforce as needed.

According to a 2021 MIT study, data labeling mistakes are widespread enough that they destabilize machine learning benchmarks.

Model Evaluation Loop

Observing model performance and calibrating input data

Once the human labeling process is complete, it's time for the data science team to feed the annotated dataset to the model and determine next steps based on performance. The team may deem the data "good enough" to deploy in the model, send it back to the human rater team for reassessment given a set of identified problems, or throw it away if annotation problems are too systemic to easily fix. If this is the case, the entire process will restart. This may be the case for the entire data set or just a subset.

To support these efforts the AI data pipeline must be configured to render data files in a quality control tool, as well as in a natural environment (for example, a browser) for sanity checks.

Poor data quality destroys business value. Recent Gartner research has found that organizations believe poor data quality to be responsible for an average of $15 million per year in losses.

Extract & Reformat

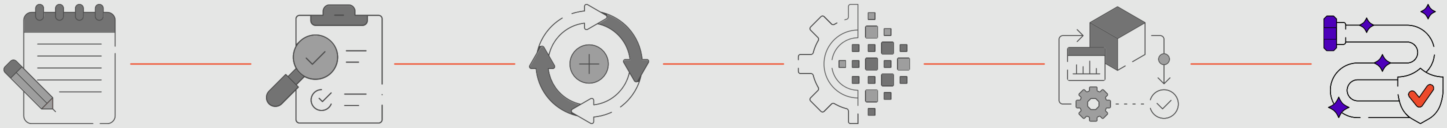
Preparing annotated and validated data for digestion

Each time a data batch is completed, it is assessed and approved (or not) as meeting quality targets defined by the data science team. Once approved, the feedback loop has been closed. It's time to extract the data and reformat it before sending it to the model for training, including JSON translation for the model. Today this is primarily a manual effort by the data science team.

> **The production of scientific data is increasing at an unprecedented rate. However, much of this data is not utilized by researchers because the available tools do not adequately support findability, accessibility, interoperability, and reusability. There is a growing need for improved tools and foundational software architectures that can effectively manage and make sense of this data, keeping pace with ongoing innovation.**

Model Assessment

Once the model has been trained, a further evaluation period begins that involves the AI data pipeline. The data science and ML Ops teams assess the success of the model and debug where appropriate, including reassessing the quality and accuracy of the data. For example, in NLP models query results are often piped back to a team of human reviewers for assessment of accuracy, toxicity and bias, among other measures. This involves the AI data pipeline transferring results from the model to the third party human review tool and workforce, as with the original data files.



Pipeline Maintenance

Keeping an eye on mission critical data flow

Once the pipeline is up and running, it is only a matter of time before it breaks. Knowing when it breaks through alerting and monitoring, along with an ability to quickly recover, is instrumental in running a successful, cost effective pipeline.

Pipeline Monitoring and Alerting

The reliability of data labeling pipelines is crucial, particularly when facing issues like broken third party APIs, corrupt data uploads, or the data uploaded in the wrong format. Monitoring and alerting are essential. Unfortunately, when the data scientist builds the pipeline themself, proper monitoring is not always put in place at the outset.

Key maintenance practices include:

1. **API Monitoring:** Continuous monitoring for data labeling platform APIs to detect breaks or changes that can disrupt the pipeline.
2. **Format Validation Checks:** Automated checks that ensure uploaded data matches an agreed-upon format.
3. **Adaptive Alerting Systems:** Dynamic alerting mechanisms that can adapt to changes in data formats and API specifications.
4. **Robust Logging and Reporting:** Comprehensive logs and reports for each stage of the pipeline to facilitate quick diagnosis and resolution.
5. **Regular API and Format Reviews:** Periodic reviews of API integrations and data formats to anticipate and mitigate future issues.

Pipeline Error Recovery

In the context of any data pipeline, implementing a strategic error recovery approach is a technical necessity and also a cost-effective measure. It ensures operational continuity and data integrity, which are critical for the precision of machine learning models. Key practices include:

1. **Data Checkpointing** at various stages of the pipeline. This allows restarting of the process from the point of error, rather than the beginning. This is significantly more cost-effective, as it reduces the time and computational resources required for reprocessing.
2. **Queue Systems with Dead Letter Queues:** A series of queues for each processing step, with dead letter queues for handling failures. This allows the team to retry only failed steps, preventing waste of resources on reprocessing successfully completed stages.
3. **Rollback and Roll Forward Strategies:** Rollback mechanisms revert to the last stable state, while roll forward strategies correct errors and proceed without losing progress. Roll forward is particularly cost-saving as it avoids the need for complete reprocessing, thus conserving resources and reducing operational cost.
4. **Regular Data Validation:** Routine validations within the pipeline to ensure data format correctness. Early detection of errors can prevent costly issues later in the pipeline.
5. **Testing and Simulation:** Regularly testing the pipeline with simulated errors helps fine-tune the recovery processes. This preparedness reduces the likelihood of long downtimes and extensive reprocessing in real-world scenarios, leading to cost savings. This is commonly referred to as chaos engineering.§

Pipeline Updates and Modifications

Maintaining a data processing pipeline, especially amidst frequent updates and modifications, demands a structured approach to ensure both operational efficiency and reliability. Embracing Infrastructure as Code (IaC) tools plays a pivotal role in this context, offering a robust framework for managing and tracking infrastructure changes.

1. **Version Control and Collaboration:** Utilizing IaC tools enables the entire infrastructure setup to be version-controlled, much like software code. By maintaining infrastructure configurations in a repository like GitHub, teams can collaborate effectively, with changes being tracked and reviewed. This approach mitigates the risk of a single point of failure, as no one engineer holds all the knowledge or control over the infrastructure.
2. **Seamless Updates and Testing:** With IaC, updating or modifying the pipeline becomes a controlled and transparent process. Teams can duplicate the entire pipeline in a pre-production environment using the same codebase, allowing for thorough testing of updates before they are rolled out in the production environment. This greatly reduces the risks associated with direct modifications to live pipelines.
3. **Documentation and History Tracking:** IaC provides a clear documentation of the infrastructure evolution. Every change, update, or modification is recorded, offering a comprehensive history. This transparency not only aids in troubleshooting but also in understanding the rationale behind past infrastructure decisions.
4. **Consistency and Standardization:** IaC ensures that the infrastructure is consistently deployed and managed. This standardization is crucial in large teams or when scaling operations, as it eliminates discrepancies that might arise from manual configurations.
5. **Disaster Recovery and Reproducibility:** In the event of a failure, IaC allows for rapid disaster recovery. The entire pipeline code can be rolled back to a previously known working version or the pipeline code can be rolled forward swiftly and accurately with a simple bug fix and redeployed through an automated CI/CD pipeline, ensuring minimal downtime.

> Automating AI data pipelines from Google to my own companies enabled our team to recapture over half our data science time which we used to increase model performance. I have run research showing that inaccurate data exponentially deteriorates model performance and as the model accuracy increases, the stakes become higher to ensure only good quality data continues to flow into the model.

DR. SOOHYUN BAE

Founder & CEO - Podonos

Conclusion

The AI data pipeline process today involves an incredible amount of bespoke manual labor on the part of data scientists and data engineers. The 2021 Wakefield State of Data Management Report, for example, states that companies are on average overspending in excess of $500K per year due to their reliance on the manual building and maintenance of data pipelines. Meanwhile, data scientists report spending over 50% of their time on this manual data janitorial work.

Faced with these challenges, some large enterprises choose to hand off this work to a third party service provider. But this approach has its own challenges. Large service providers are forced to use their own workforce and tools for the job, which is costly for them, and sometimes not the best option for a given job. They are often non-transparent, costly, slow and deliver data of mediocre quality.

The costly, labor intensive process outlined above is a symptom of a new industry that has not yet maximized efficiency in its data operations. Over time, customers building AI models can expect that this process will become increasingly automated and efficient. Automated AI data pipeline builders will increase scale and speed, eliminating the vast majority of the manual work required to maintain these pipelines. As these solutions scale, they will discover new use cases to solve and value to provide now that the unreasonable amount of grunt work involved in keeping these pipelines running has been removed. The result will be more innovation and scale, benefitting the entire AI ecosystem.

mltwist.com

contact@mltwist.com

linkedin.com/company/mltwist