Ground_Truth_Data for AI

Towards Sustainable AI

Dossier

Romero and Salmeron
June 27, 2024

EXECUTIVE SUMMARY

1. EXECUTIVE SUMMARY

In today's data-driven world, managing and processing vast amounts of data efficiently is crucial for organizations across industries. Our innovative data reduction technology addresses this challenge by offering a sustainable and highly efficient solution. By leveraging cutting-edge algorithms, our technology significantly reduces the size of datasets while maintaining data integrity and quality.

Key Benefits

- Sustainability: Our technology plays a pivotal role in environmental sustainability by reducing the need for extensive data storage infrastructure and minimizing energy consumption during data processing and transmission. This results in a reduced carbon footprint and aligns with global initiatives for sustainable business practices.
- Efficiency for AI Models: By employing advanced algorithms tailored for AI datasets, our technology enhances the efficiency of machine learning processes. It accelerates data access and reduces latency, enabling AI models to process information faster and deliver more responsive insights.
- Operational efficiency: Through advanced algorithms, our solution enhances operational efficiency. It enables faster data retrieval and processing speeds, thereby improving overall system performance and reducing operational costs associated with data management and processing.
- Scalability: Designed to scale seamlessly with growing data volumes, our technology caters to the evolving needs of diverse industries such as healthcare, finance, telecommunications, and specially data centers. It ensures that organizations can efficiently manage increasing data loads without compromising on performance or reliability.

Market Impact

Our data size reduction technology addresses a critical gap in the market for sustainable data management solutions. By enabling organizations to optimize their data storage and processing capabilities, it positions them to make informed decisions faster and gain a competitive edge in their respective markets. The technology's ability to reduce costs associated with data storage and energy consumption further enhances its attractiveness to businesses seeking operational efficiencies.

Our technology has a profound impact on the market by addressing fundamental challenges in data management while fostering sustainable business practices:

- Competitive Advantage: Organizations adopting our technology gain a competitive edge by optimizing their data management capabilities. They can leverage insights from streamlined data processes to make informed decisions faster, innovate product offerings, and respond swiftly to market demands.
- Cost Savings: Reduced data storage requirements and lower energy consumption lead to significant cost savings over time. Organizations can reallocate resources towards strategic initiatives such as research and development, customer experience enhancements, or expanding market reach.
- Compliance and Risk Management: Our technology helps businesses mitigate risks associated with data breaches and non-compliance with data protection regulations. By ensuring data security and integrity, organizations can avoid legal penalties and reputational damage.

Conclusion

In conclusion, our data technology represents a transformative advancement in data management practices. By offering sustainable solutions that reduce data footprint while enhancing operational efficiency and scalability, we empower organizations to unlock the full potential of their data assets. Embracing our technology not only supports environmental sustainability but also drives business growth and innovation in an increasingly digital landscape.

For further information on implementation strategies and how our technology can benefit your organization, please contact us. Together, we can embark on a journey towards smarter, more sustainable data management practices.

MARKET

2. MARKET

2.1. SUSTAINABILITY

In recent years, we have witnessed an extraordinary acceleration in the growth of artificial intelligence (AI), transforming the way we live, work, and interact with technology. AI algorithms impact sectors such as healthcare, finance, manufacturing, transportation, and entertainment. These advances are driving new chip and server technologies, resulting in extreme rack power densities and presenting new challenges in the design and operation of data centers to meet the massive demand for AI. As part of their Environmental, Social, and Governance (ESG) programs, data center operators are making commitments to environmental sustainability.

Data center operators should use a standard set of metrics. The Green Grid (TGG) proposed power usage effectiveness (PUE) in 2007, which was widely adopted and helped drive efficiency improvements across the industry. A global survey conducted by Uptime Institute in 2023 showed that the average annual PUE of large data centers improved from 2.5 to 1.58 since 2007.

Schneider Electric's growth predictions estimate that AI accounts for 8% of the total power consumption of data centers in 2023, rising to 15-20% by 2028. Given that data centers worldwide consumed over 500 TWh of electricity in 2023 and are projected to consume 815 TWh in 2028 (the equivalent of 16 New York Cities), AI's power consumption equates to 40 TWh in 2023 and is expected to increase to between 122 and 163 TWh by 2028.

2.2. GROUND TRUTH DATA SOLUTION

2.2.1. Approach

At Ground Truth Data, we're tackling energy consumption in data centers with our efficient data reduction technology. This approach not only cuts down on storage space for AI data but also reduces computing times and network loads, boosting overall system efficiency.

Our solution brings a range of benefits: better use of storage resources, cost savings, longer hardware lifespans, and optimized data processing. For AI applications, our techniques improve performance and scalability, allowing for faster training and deployment of AI models, especially deep learning.

Additionally, our technology speeds up backup and recovery processes, which is crucial for the finance and healthcare sectors, and enhances the efficiency of IoT devices in edge computing. For cloud service providers, it helps meet sustainability goals and lowers operational costs.

Using our data reduction strategies leads to significant energy savings, lower operational costs, and greater environmental sustainability, all while maintaining high performance and reliability. Our solution also supports AI development, making AI technologies more accessible and efficient across various industries.

2.2.2. Competitors

In the competitive landscape of data reduction technology, several key players have emerged, each offering different solutions to address the growing challenges of managing and optimizing large datasets.

- IBM: IBM offers various data compression and storage optimization solutions through its cloud services and software products. Their technologies focus on enhancing data efficiency and scalability for enterprises.
- Dell Technologies: Dell provides data reduction technologies as part of its storage and data management solutions. They emphasize deduplication, compression, and encryption to improve storage efficiency and performance.
- NetApp: NetApp specializes in data management and storage solutions, including data reduction technologies such as deduplication and compression. Their offerings are tailored for enterprise data centers and cloud environments.
- Pure Storage: Pure Storage offers flash storage arrays with built-in data reduction features, including deduplication and compression. Their solutions target high-performance computing and enterprise storage needs.
- Veritas Technologies: Veritas provides data management and protection solutions, including data reduction technologies for backup and storage optimization. They serve a wide range of industries with a focus on data integrity and compliance.

Our solution distinguishes itself from existing market solutions by offering a novel approach that does not rely on encryption or compression techniques. Instead, we **leverage advanced algorithms that fundamentally transform how datasets are managed**. By adopting this innovative methodology, our solution achieves a remarkable reduction in dataset size while meticulously **preserving the integrity, knowledge, and critical information** embedded within the data.

This preservation ensures that organizations **not only benefit from streamlined data storage and processing but also retain the full depth and value of their data assets**.

In essence, our approach sets a **new standard in data management efficiency and intelligence**, empowering businesses to optimize operations and drive informed decision-making with confidence.

2.3. RESULTS

Ground Truth Data's solution achieves data reductions of up to 86%, leaving only 14% of the original data. This reduction translates to a 36% savings in computing for AI algorithms on IoT devices and an 86% savings in data centers.

Considering these figures and the optimistic consumption forecasts for data centers in the coming years, we can illustrate the potential energy and cost savings for the industry in the following table:

Year	AI (TWh)	Elect. price ($)	AI Cost (B$)	AI E. saved (TWh)	Cost saved (B$)	CO_2 saved (Mt)
2023	40	0.17	7	34	6	7
2024	54	0.175	9	46	8	10
2025	72	0.18	13	62	11	13
2026	92	0.185	17	79	15	16
2027	115	0.19	22	99	19	20
2028	143	0.195	28	123	24	25
2029	175	0.2	35	151	30	31
2030	212	0.206	44	182	38	38
				776 (TWh)	151 (B$)	160 Mt CO_2

MARKET

TECHNOLOGY

3. DATA REDUCTION

3.1. THE GOAL

The aim of data reduction is to simplify a dataset while preserving its key information. This can typically be accomplished by either reducing the number of features or the number of samples. Our approach will concentrate on the more challenging task of reducing the sample size.

3.2. NEED FOR DATA SIZE REDUCTION

With data being collected at an unprecedented pace, data reduction plays a critical role in boosting training efficiency. By reducing the number of samples, we create a simpler yet representative dataset, which can alleviate memory and computation constraints. This not only enhances sustainability by lowering energy consumption but also contributes to significant energy savings.

3.3. REGULAR METHODS FOR REDUCING SAMPLE SIZE

Sample reduction is typically achieved through instance selection, which involves choosing a representative subset of data samples that retain the original dataset's properties. Existing methods can be categorized into wrapper and filter methods. Filter methods select instances based on scoring functions, such as selecting border instances that often shape the decision boundary. Wrapper methods, on the other hand, select instances based on model performance, considering their interaction with the model. Additionally, instance selection techniques can address data imbalance issues by undersampling the majority class, such as with random undersampling. Recent advancements have incorporated reinforcement learning to optimize undersampling strategies. However, these regular methods neither achieve the data size reduction nor the accuracy level that our data size reduction approach can deliver.

3.4. CHALLENGES

The challenges of data reduction are twofold. First, selecting the most representative data or projecting data into a low-dimensional space with minimal information loss is complex. While learning-based methods can partially address these challenges, they often require substantial computational resources, particularly with very large datasets. Consequently, achieving both high accuracy and efficiency is difficult. Second, data reduction can potentially amplify data bias, raising fairness concerns.

4. VALIDATION TESTS

4.1. INTRODUCTION

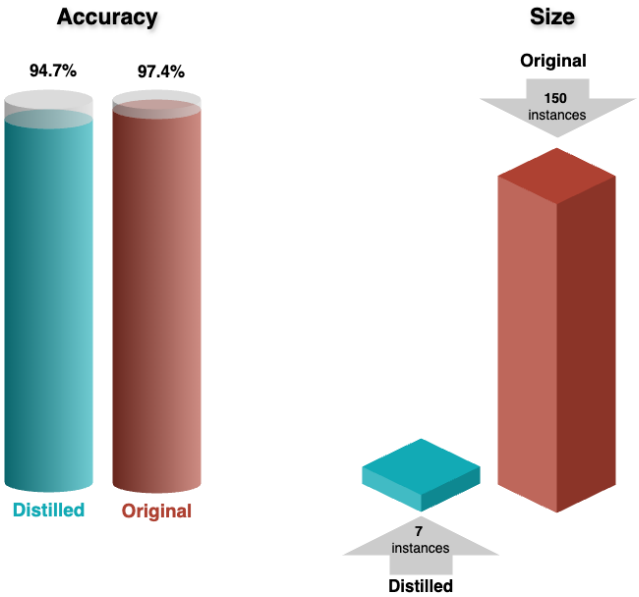
In the realm of data science and machine learning, the efficiency and effectiveness of data processing are often contingent upon the size and manageability of the datasets in use. The burgeoning volume of data necessitates innovative methods for reducing dataset size without compromising the integrity and utility of the data. This chapter focuses on validation tests for a revolutionary data size reduction method, applied to four well-known datasets: the Iris dataset, the Wine dataset, the Breast Cancer dataset, and the MNIST dataset.

4.2. TESTS DONE

4.2.1. Iris Dataset

The Iris dataset, introduced by Ronald A. Fisher in 1936, is a staple in the field of machine learning and statistics. It consists of 150 instances of iris flowers, each described by four features: sepal length, sepal width, petal length, and petal width. These features are used to classify the flowers into three species: Iris-setosa, Iris-versicolor, and Iris-virginica.

The simplicity and clarity of the Iris dataset make it an ideal candidate for demonstrating basic principles of data size reduction.



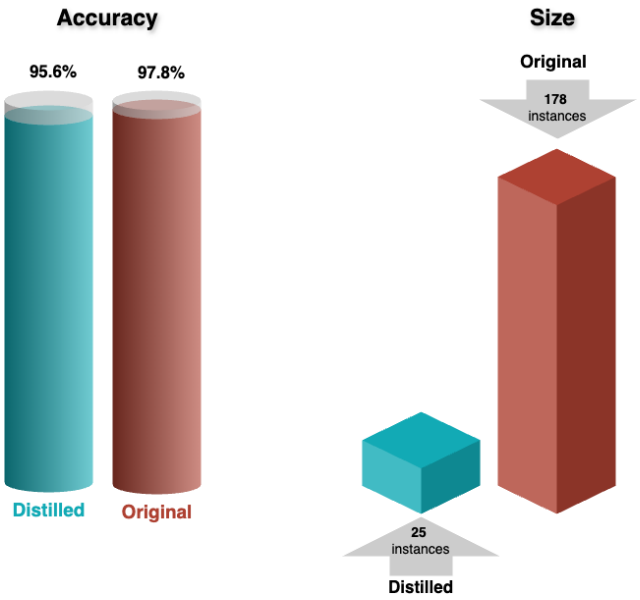
The figure illustrates the accuracy of a classifier trained on the original dataset, which contains 150 instances, achieving 97.4% accuracy. In comparison, a model trained on a reduced dataset of 7 instances achieves 94.7% accuracy.
This indicates that the proposed solution successfully reduces the original dataset by 95% while maintaining a high classification accuracy of 94.7%.

4.2.2. Wine Dataset

The Wine dataset is derived from the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. It contains 178 instances with 13 attributes including alcohol content, malic acid, ash, and others.

This dataset is often used for classification problems and serves as an excellent test bed for validating data reduction techniques due to its moderate size and complexity.

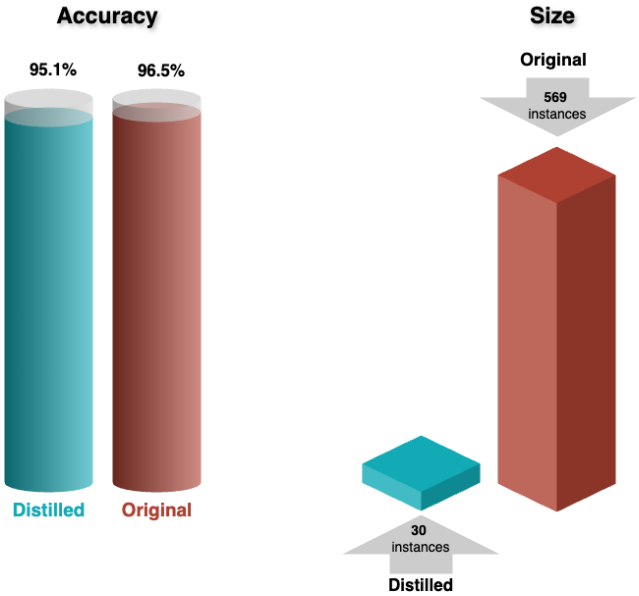
The figure shows the accuracy of a classifier trained on the original dataset of 178 instances, which achieves 97.8% accuracy. In comparison, a model trained on a reduced dataset of 25 instances achieves 95.6% accuracy.
This demonstrates that the proposed solution effectively reduces the original dataset by 86% while maintaining a high classification accuracy of 95.6%.



4.2.3. Breast Cancer Dataset

The Breast Cancer Wisconsin dataset, created by Dr. William H. Wolberg, is used for binary classification tasks in predicting the malignancy of breast cancer samples. It comprises 569 instances, each with 30 numeric features representing characteristics of cell nuclei present in a digitalized image of a fine needle aspirate of a breast mass.

This dataset is critically important for medical research and diagnostics, making the preservation of data quality essential during size reduction.



The figure displays the accuracy of a classifier trained on the original dataset of 569 instances, which achieves 96.5% accuracy. In comparison, a model trained on a reduced dataset of 31 instances achieves 95.1% accuracy.
This proofs that the proposed solution effectively reduces the original dataset by 95% while maintaining a high classification accuracy of 95.1%.

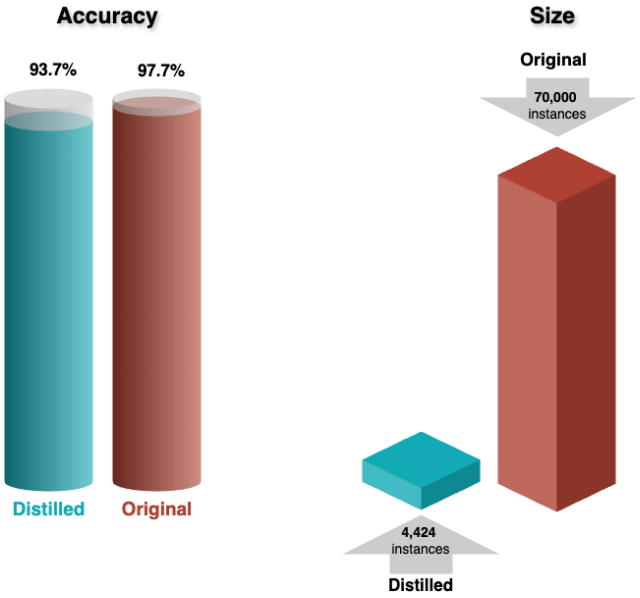
4.2.4. MNIST Dataset

The MNIST dataset is a large collection of handwritten digits, commonly used for training various image processing systems. It includes 60,000 training examples and 10,000 testing examples, each

represented by a 28x28 grayscale image of a digit (0-9).

The high dimensionality and substantial size of the MNIST dataset present significant challenges for data size reduction, thus providing a rigorous test for our proposed method.

The figure shows the accuracy of a classifier trained on the original dataset of 70,000 instances, achieving 97.7% accuracy. In comparison, a model trained on a reduced dataset of 4,424 instances achieves 93.7% accuracy.
This demonstrates that the proposed solution effectively reduces the original dataset by 93% while maintaining a high classification accuracy of 93.7%.



TEAM

5. TECHNOLOGY INVENTORS

5.1. MEMBERS

5.1.1. Ivan Romero, PhD

Dr. Ivan Romero-Ruiz has consistently balanced his career between academia and the corporate sector. He began as a Lecturer at the University of Barcelona, earning a prestigious FPI Doctoral Fellowship and authoring numerous geophysics publications while participating in Ministry of Science and Innovation projects. Transitioning to the corporate world, he became the Principal Data Scientist at Clevernet, where he authored patents on AI algorithms for network communications, mentored international students, and published related articles. Currently, he works on various AI projects in healthcare, personalized medicine, and non-convex optimization, showcasing his broad expertise and commitment to data science.

Contact: ✉ **ivanromeroruiz@gmail.com**

5.1.2. Jose L. Salmeron, PhD

Prof. Jose L. Salmeron, Ph.D., Eng., ACM Senior Lifetime Member, holds the position of (Catedrático) Professor of Computer Science and Artificial Intelligence at CUNEF University. He has also served as Principal AI & Quantum Scientist at the Hybrid Intelligence division of Capgemini. He has almost 30 years of experience in technology and research, including positions at several universities, consulting in the AI/IT industry, and a wide range of projects with private and public organizations such as Intel, Cisco, Vodafone, Gilead, Microsoft, Airbus, BBVA, and others. He has been awarded ACM Senior Membership. Prof. Salmeron has consistently been featured in Stanford's list of the world's top 2% scientists since its inception. He has authored or co-authored over 200 scientific articles, conference papers, and book chapters indexed in ISI Web of Science. Furthermore, his work has garnered almost 6,000 citations from researchers in 20 different countries, resulting in an h-index of 37 on Google Scholar. His primary fields of expertise encompass Distributed Artificial Intelligence, eXplainable Artificial Intelligence, Reservoir Computing, Quantum Computing, Causal Machine Learning, and Federated Learning.

Contact: ✉ **joseluis.salmeron@gmail.com**

ECONOMIC IMPACT

6. REVENUE

The revenue table forecasts profits through 2026. The AI (TWh) field shows the estimated energy demand for AI processes in data centers worldwide, based on predictions from Schneider Electric. The Electricity price ($) field predicts the average electricity price in the United States.

The AI cost (B$) field predicts the global cost of energy demand for AI processes in data centers. The AI energy saved (TWh) field estimates the energy savings from our solution. The Saved cost (B$) field represents the predicted cost savings due to our solution, and the CO2 saved (Mt) field shows the megatons of CO_2 emissions that will be prevented with our solution. SAM (%) indicates the percentage of market share defining the serviceable addressable market, while SAM (B$) represents the money saved in the SAM. SOM (M$) predicts the serviceable obtainable market, which is the estimated profit until 2026 (comission 10%). Cum. SOM (M$) is the cumulative predicted profit expected until 2026. Following the upward trend in average electricity prices and the growing demand in data centers, it is estimated that by 2030, SOM benefits will reach at least 606 million dollars.

Year	AI (TWh)	Electricity price ($)	AI cost (B$)	AI energy saved (TWh)	Saved cost (B$)	CO_2 saved (Mt)	SAM (%)	SAM (M$)	SOM (M$)	Cum. SOM (M$)
2023	;40	0.170	7	34	6	7	0.000	0	0	0
2024	54	0.175	9	46	8	10	0.000	0	0	0
2025	72	0.180	13	62	11	13	0.001	11	1.1	1
2026	92	0.185	17	79	15	16	0.007	105	10.5	12
2027	115	0.190	22	99	19	20	0.015	285	28.5	40

Energy demand prevision and costs

Given the increasing power demand in data centers and the anticipated scarcity of energy in the midterm future, our data size reduction solution should be considered a standard protocol. The adoption of our solution in the algorithmic AI processes of data centers leads to significant energy savings, which translates to substantial cost reductions and a decrease in CO2 emissions.

As energy demands continue to rise, the necessity for efficient energy solutions becomes more critical. Our solution offers a sustainable path forward, ensuring that data centers can meet the growing needs while minimizing their environmental impact and operational costs. The projected savings and reduced carbon footprint underscore the vital role our solution will play in the future of data center operations. Adopting our data size reduction as a standard protocol is essential for maintaining energy efficiency and sustainability in the face of increasing energy constraints.