IDC MarketScape

IDC MarketScape: Worldwide Data Labeling Software 2023 Vendor Assessment

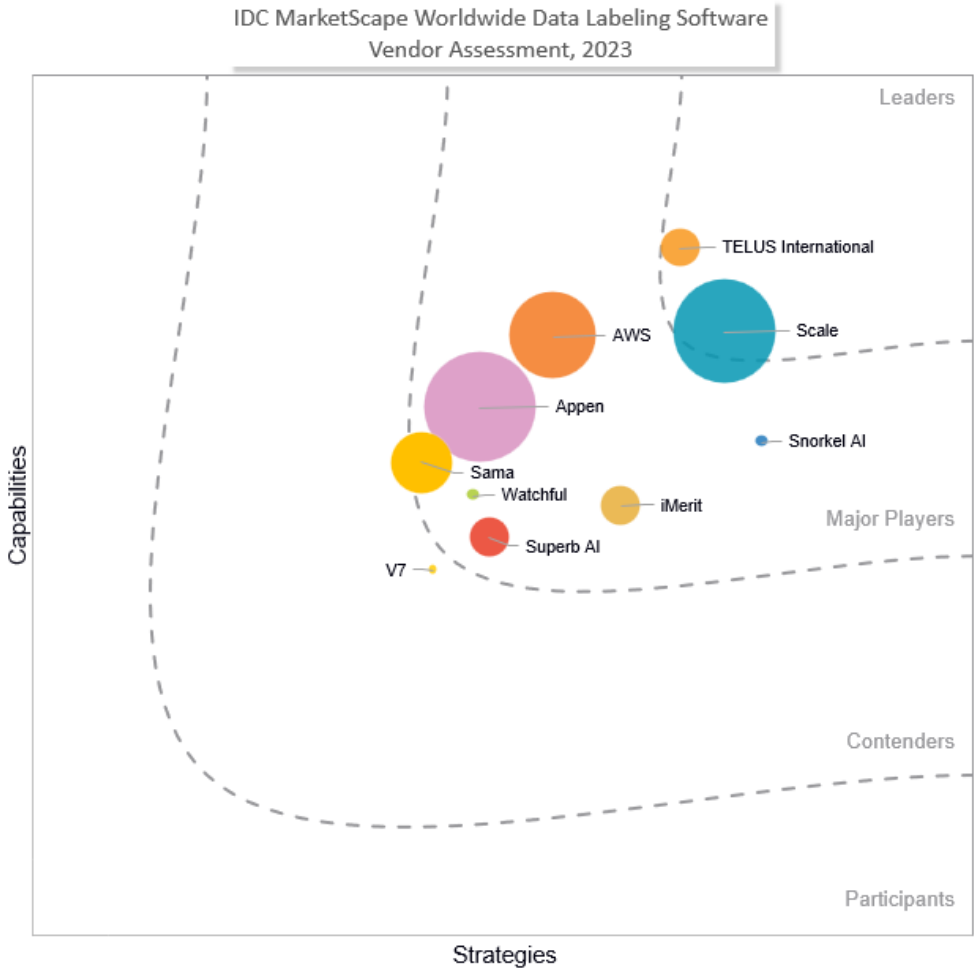
Kathy Lange

THIS IDC MARKETSCAPE EXCERPT FEATURES TELUS INTERNATIONAL

IDC MARKETSCAPE FIGURE

FIGURE 1

IDC MarketScape Worldwide Data Labeling Software Vendor Assessment



Source: IDC, 2023

November 2023, IDC #US50010523e

Please see the Appendix for detailed methodology, market definition, and scoring criteria.

IN THIS EXCERPT

The content for this excerpt was taken directly from IDC MarketScape: Worldwide Data Labeling Software 2023 Vendor Assessment (Doc# US50010523). All or parts of the following sections are included in this excerpt: IDC Opinion, IDC MarketScape Vendor Inclusion Criteria, Essential Guidance, Vendor Summary Profile, Appendix and Learn More. Also included is Figure 1.

IDC OPINION

Overview

In the face of uncertainty and market conditions (higher interest rates, inflation, layoffs, fears of recession, increased cost of goods and services), businesses continue to make substantial investments in artificial intelligence (AI) and machine learning (ML) technologies to remain competitive and realize the benefits of improved operational efficiency, increased innovation, improved customer experience (CX), and employee productivity.

IDC forecasts that the worldwide artificial intelligence life-cycle software platforms market will grow from $6.4 billion in 2022 to $27.0 billion in 2027 at a compound annual growth rate (CAGR) of 33.6%. Growth in this market continues to be driven by the increased adoption of AI technologies for digital transformations, the democratization of AI capabilities, and ever-expanding use cases for traditional ML, such as customer acquisition and retention, fraud detection, and recommendations, and growing demand for generative AI (GenAI).

The data labeling submarket of the AI/ML life-cycle tools and technologies is broad, with a variety of incumbents and many small to midsize vendors entering the market. It is evolving from purely services that use armies of human labelers to AI-assisted software solutions that are designed for high-context tasks. Along with this evolution, established data labeling vendors are redirecting much of their expertise, experience, and operations toward supporting enterprise generative AI endeavors. Drawing from their background in assisting prominent tech companies develop early foundation models, they are now pivoting to serve enterprise customers exploring the fine-tuning of their own foundational models for generative AI use cases.

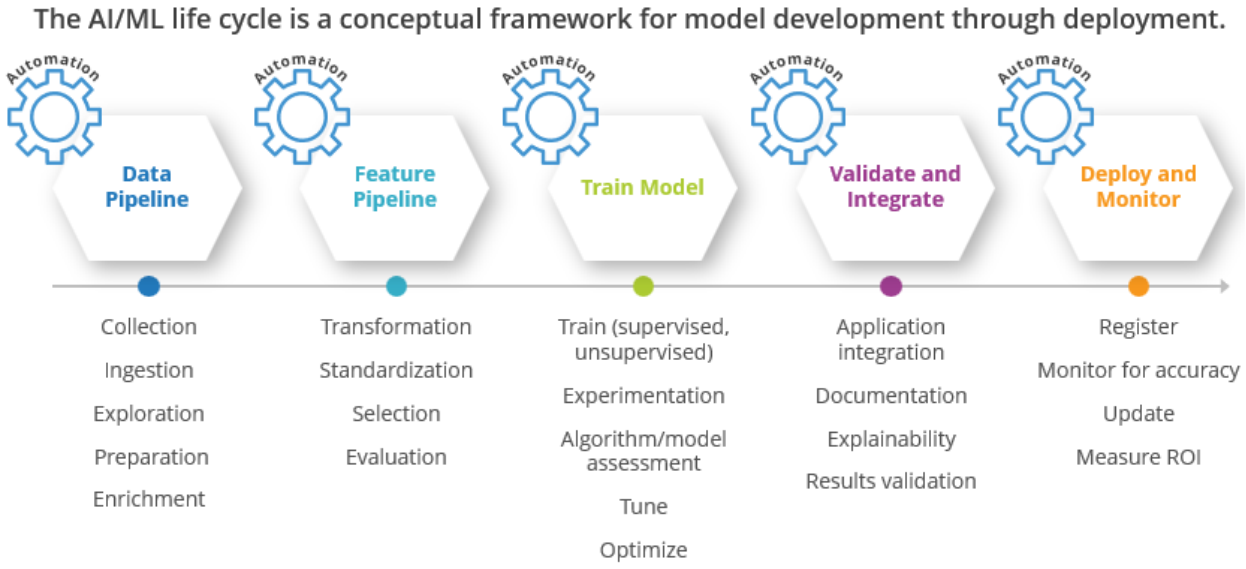
Data Labeling Within the Artificial Intelligence and Machine Learning Life Cycle

Within the AI/ML life cycle, data labeling is part of the "data pipeline" phase (see Figure 2).

New AI/ML applications using unstructured data, such as conversational AI (chatbots), speech recognition, sentiment analysis, object detection in autonomous driving and robotics applications, medical imaging, facial recognition, and generative AI, will fuel additional AI/ML market expansion. Organizations have vast treasure troves of unstructured data in the form of text, images, and video that are largely unused. Unstructured data is a resource for enterprises to mine, extract actionable insights, and improve business outcomes. IDC expects unstructured data to constitute as much as 90% of all business data generated by organizations. And while AI/ML is critical to unlocking the value of previously untapped unstructured data; it must be enriched with high-quality labels that feed AI/ML algorithms.

FIGURE 2

Artificial Intelligence and Machine Learning Life Cycle



Source: IDC, 2023

Prior to training AI/ML models, there are many data preparation activities (collection, ingestion, exploration, preparation, and enrichment). The data pipeline is the stage in the machine learning life cycle where data is enriched with labels. As the number and complexity of AI/ML models skyrocket, so do the data preparation difficulties. Large data volumes for AI/ML models precipitate managing, storing, processing, and versioning challenges. The lack of high-quality, unbiased labeled data for detecting patterns and predicting outcomes impacts model velocity (the pace of moving through the AI/ML life cycle). Approximately one-third of survey respondents to IDC's May 2022 *AI Strategies View* cited that manual data labeling was one of their top data preparation challenges.

Data labeling, also known as data annotation, data tagging, or data classification, turns unlabeled data into labeled data used to train supervised AI/ML models. Organizations developing AI/ML applications across industries will benefit from faster, more accurate, and less costly data labeling. The quality of training data is critical to the success of an AI/ML project.

Data labeling software tools support one or more data types – image, video, audio, text, time series, satellite imagery, LiDAR, or RADAR. The annotation types may vary, including segmentation and object detection for image data, named entity recognition (NER) and sentiment detection for text data, and transcription and emotion recognition for speech annotation. A growing number of projects require a combination of different data types, which are referred to as multimodal applications.

A Typical Human-Centered Data Labeling Process

A typical labeling effort involves the following broad steps:

- Define the project, collect unlabeled data, and prepare data.
- Set up the project workflow, categories, and quality metrics.

- Hire, outsource, or crowdsource annotators.
- Train the annotators based on instructions and guides.
- Execute the project, validate results, and measure quality. (Iterate with the annotators and instructions until quality measures are achieved.)
- Transform and format the labeled data for use in feature pipelines for AI/ML model training.

Challenges of Data Labeling

Humans are currently the primary source for labeling data; however, manual human data labeling can be slow and costly. Manual human data labeling challenges abound in hiring, training, and managing the labeling workforce. A growing number of labeling projects require experts in a specific field, such as lawyers, healthcare professionals, or medical imaging experts. The work can be tedious and time consuming, and there is often inconsistency across labelers. Rework or relabeling is often an issue. As projects evolve and new labels, classes, or categories are introduced within a project, humans may need to relabel the entire data set mid-project, which can add significant labor costs and timeline delays.

Buyers' critical concerns for data labeling include the quality and accuracy of results, costs, faster time to value, and data security. To address these, software vendors have created data labeling platforms with integrated machine learning tools to assist human labelers, often referred to as human-in-the-loop (HITL) offerings.

An emerging technology for data labeling, called programmatic labeling, data programming, or weakly supervised learning, appears in some of the vendors' offerings evaluated in this study. The technique extracts knowledge from human subject matter experts (SMEs) and uses AI-assisted tools to develop noisy, imprecise, or overlapping heuristics (weak labeling functions). The system reconciles these weak labels through probabilistic models and uses them to make inferences about the unlabeled data. These techniques are particularly useful for labeling projects with evolving taxonomies where relabeling is common or projects require scarce subject matter specialists rather than general-purpose annotators.

The purpose of this document is to identify and evaluate AI-assisted data labeling software-based platforms that feed labeled data into the AI/ML life cycle.

IDC MARKETSCAPE VENDOR INCLUSION CRITERIA

The inclusion criteria for this study are as follows:

- The offering must be commercially available for over a year prior to January 2023.
- The offering must provide either human-in-the-loop software or services.
- The offering must provide AI-assisted (machine learning) software for pre-labeling.
- The offering must support cross-industry labeling applications.
- The offering must provide annotation for at least two data types.
- The offering must be based on the vendor's intellectual property (IP).
- The offering must have at least 15 customers in production as of January 2023.

ADVICE FOR TECHNOLOGY BUYERS

Scope Your Data Labeling Project Completely

Develop a full-blown project plan for labeling projects. Think carefully about the project scope and volume of data. Large data volumes can be very costly. Evaluate your label concept and classes thoroughly from the beginning to avoid mid-project changes that can increase costs significantly. Experienced labeling service providers (SPs) can offer valuable guidance and expertise to assist you in shaping your project plan. Consider starting with a small project and using an agile approach to scaling out. Gain value at each step to maintain internal funding and stakeholder support. Talk to your provider about its ability to streamline the labeling project with third-party data sets or foundation models.

Focus on Effective Training of Data Labelers

Invest ample time and careful consideration into training labelers and project managers about your labeling needs. They may lack the context or understanding of your data and objectives. It's important to remember that the annotators come from diverse backgrounds and regions, which can sometimes result in varying levels of familiarity with the content that requires labeling. Developing a well-thought-out training plan is essential for effective onboarding. Additional feedback and coaching of annotators may be necessary to improve labeling outcomes.

Evaluate your Data Quality and Accuracy Needs

Each use case may have different requirements for data quality and accuracy, depending on costs or risks. Evaluate your needs and the metrics that the vendor provides to monitor, track, and evaluate data quality and accuracy. Many vendors rely on consensus for measuring inter-annotator agreement, but others provide human auditing, statistical evaluations, or AI-assisted functions to identify potential label errors. It is imperative to maintain ongoing monitoring and transparency of quality and accuracy throughout the entire labeling process rather than solely focusing on them at the project's conclusion, as midcourse corrections may be necessary.

Choose the Right Vendor for Your Needs

This IDC MarketScape highlights many vendors, all of which are successful at various aspects of data labeling and the broader AI/ML life cycle. Carefully consider your intended use cases and identify your requirements for data types, human annotators, price, security, deployment options, and geographic needs. There are significant upfront training costs to bring human annotators up to speed. Decide whether your project requires in-house experts, dedicated outsourced resources, or a rotating pool of annotators. Understand any specialized requirements that you might have, such as language or regional support, that will impact the labeling process, resources, and costs. Finding a good fit and alignment between the vendor and customer teams is essential for success.

VENDOR SUMMARY PROFILES

This section briefly explains IDC's key observations resulting in a vendor's position in the IDC MarketScape. While every vendor is evaluated against each of the criteria outlined in the Appendix, the description here provides a summary of each vendor's strengths and challenges.

TELUS International

TELUS International is positioned in the Leaders category in this 2023 IDC MarketScape for worldwide data labeling software.

TELUS International is a global digital customer experience company that designs, builds, and delivers next-generation artificial intelligence and content moderation solutions. Its proprietary data labeling platform is Ground Truth (GT) Studios.

GT Studios provides a cross-industry human-in-the-loop labeling technology stack with integrated project pipeline management software, AI-assisted annotation tools, people management tools, and a quality assurance framework. TELUS International provides an end-to-end AI data solution, including data collection, data management, data selection, and data annotation. TELUS International acquired Lionbridge AI, with its managed training data and data annotation services, in 2020 and acquired Playment, specializing in computer vision tools and services for 2D and 3D image, video, and LiDAR, in 2021. The 2023 acquisition of WillowTree bolstered its arsenal of machine learning application developers.

GT Studios provides an extensible and configurable no-code/low-code engine for operationalizing project workflows. TELUS International's strong suit is annotation for computer vision projects. It boasts a large, long-tenured customer base, deep project experience, and technology for use cases across industries. The company supports many challenging data labeling projects for autonomous driving. It provides data annotation tools, including 3D point cloud annotation, semantic segmentation, 2D/3D bounding boxes, polygons, polylines, key points, landmarks, object tracking, and sensor fusion.

TELUS International provides both in-house expert human labelers for quality auditing and onsite annotators within a secure BPO facility that can be scaled up or down, with access to a community of more than 1 million global workers to meet project demand. Additional points of note:

- **AI-assisted automation.** Automation capabilities are delivered through a series of tools, including ML-assisted functions. Autopilots enable zero-touch annotations for pre-labeling. Copilots work in conjunction with human annotators to find repeatable task methods such as interpolation. Both enhance label accuracy and improve speed. Semiautomated quality control helper bots smartly identify tasks for review based on error-prone items. GT Studios also integrates with a Segment Anything Model–based object detection model for image detection.
- **Quality and accuracy.** GT Studios provides detailed analytics of data sets and worker-level quality metrics. Statistical evaluations are used to determine quality and accuracy. Accuracy is based on a maker, checker, and master framework that ensures annotations go through three sets of checks rather than determining accuracy on a traditional consensus basis. Ongoing training for its dedicated BPO workers on TELUS International's in-house software improves familiarity and better long-term accuracy.
- **Security.** GT Studios uses the following security and privacy methods to protect a clients' data and user information; automated PII identification and redaction, data masking, data loss prevention, encryption, role-based access control, data backup and recovery, network security, real-time performance monitoring, and cloud platform infrastructure compliance for SOC-2 and TISAX.
- **GenAI integration within the labeling workflow.** GT Studios uses the Meta Segment Anything Model foundation model as a first pass in its auto-pilot feature to automatically assign labels to image segments. These segments are further refined using TELUS International's proprietary deterministic and probabilistic models.

Quick facts:

- **Year founded:** 2005
- **Headquarters:** TELUS International is headquartered in Vancouver, British Columbia, Canada, and is publicly held.
- **Total number of employees:** 75,000+
- **Deployment options:** GT Studios software is embedded as part of a fully managed service and is not separately licensable. The service is hosted in the AWS cloud.
- **Pricing model:** TELUS International offers annual or multiyear contracts with the potential for volume-based discounts. Pricing is consumption based (per label).
- **Annotation data types:** Image, text, video, LiDAR, RADAR, time series, and multimodal (text and image, text and video, sensor fusion)
- **Related products/services:** TELUS International provides custom data creation and collection, a "workforce only" solution using third-party software, and model-building services with on-staff data scientists.

Strengths

- **Enterprise features.** In addition to its AI-assisted labeling capabilities, the offering is highly adaptable and configurable for clients' unique workflow requirements. It contains comprehensive data management, workflow orchestration, integrated analytics, quality control performance dashboards, a host of security features, and more. All data annotation types are supported in a single tool, allowing dedicated workers to become more familiar with the interface and capabilities.
- **Global data collection.** TELUS International's broad geographic presence helps companies source and consolidate text, image, audio, and video data across many countries and regions. It has collected hundreds of thousands of kilometers of sensor data for its automotive and consumer electronics clients.
- **Scalability.** TELUS International's platform supports large data-intensive projects to support petabytes of data and a wide range of data and annotation types. The scalability of its global workforce also supports large, complex, single or multimodal projects.

Challenges

- **Price.** Customers perceive the cost of the solution to be high, particularly with projects that require large amounts of data.
- **Deployment options.** GT Studios is only offered as a fully managed service. Clients desired flexibility in deployment options, such as on-premises and software-only options.

Consider TELUS International When

Consider TELUS International for large, complex, and mission-critical computer vision data labeling projects, including multimodal projects, such as autonomous driving, where a secure data environment is paramount. TELUS International has a large customer footprint in Europe-based companies. Based on the scalability of the workforce, the company should be considered for time-sensitive projects that require fast turnaround.

APPENDIX

Reading an IDC MarketScape Graph

For the purposes of this analysis, IDC divided potential key measures for success into two primary categories: capabilities and strategies.

Positioning on the y-axis reflects the vendor's current capabilities and menu of services and how well aligned the vendor is to customer needs. The capabilities category focuses on the capabilities of the company and product today, here and now. Under this category, IDC analysts will look at how well a vendor is building/delivering capabilities that enable it to execute its chosen strategy in the market.

Positioning on the x-axis, or strategies axis, indicates how well the vendor's future strategy aligns with what customers will require in three to five years. The strategies category focuses on high-level decisions and underlying assumptions about offerings, customer segments, and business and go-to-market plans for the next three to five years.

The size of the individual vendor markers in the IDC MarketScape represents the market share of each individual vendor within the specific market segment being assessed.

IDC MarketScape Methodology

IDC MarketScape criteria selection, weightings, and vendor scores represent well-researched IDC judgment about the market and specific vendors. IDC analysts tailor the range of standard characteristics by which vendors are measured through structured discussions, surveys, and interviews with market leaders, participants, and end users. Market weightings are based on user interviews, buyer surveys, and the input of IDC experts in each market. IDC analysts base individual vendor scores, and ultimately vendor positions on the IDC MarketScape, on detailed surveys and interviews with the vendors, publicly available information, and end-user experiences in an effort to provide an accurate and consistent assessment of each vendor's characteristics, behavior, and capability.

Market Definition

Data Labeling Software

Data labeling software — also known as training data, data annotation, data tagging, or data classification software — provides a toolset for businesses to turn unlabeled data into labeled data and build corresponding artificial intelligence algorithms. Within these tools, the user inputs a given data set, and the software provides a label through machine learning-assisted labeling, a human task force, or the user themselves. Some platforms allow for the combination of the three, giving the user (or the system itself) the ability to choose who or what is doing the labeling, based on factors such as price, quality, and speed.

Data labeling tools differ as they relate to the types of data (e.g., image, video, audio, and text) as well as the subsets of those types (e.g., satellite imagery and LiDAR) they support. The annotation types also vary and include image segmentation and object detection for image data, named entity recognition (NER) and sentiment detection for text data, and transcription and emotion recognition for speech annotation. To assess the quality of the labels, most tools use metrics like consensus and ground truth.

This software can often integrate with data science and machine learning platforms, whereby the labeled data from the data labeling software helps train an AI/ML algorithm.

LEARN MORE

Related Research

- *Is the High Cost of Real Data for Large AI/ML Models Driving Synthetic Data Adoption?* (IDC #US51186223, September 2023)
- *Market Analysis Perspective: Worldwide AI Life-Cycle Software, 2023* (IDC #US50008423, July 2023)
- *Worldwide AI Life-Cycle Software Forecast, 2023-2027* (IDC #US50858923, June 2023)
- *Worldwide AI Life-Cycle Software Market Shares, 2022: A Focus on Governance and MLOps* (IDC #US50858823, June 2023)
- *State of Artificial Intelligence and Machine Learning Spending and Implementations* (IDC #US50605023, May 2023)
- *IDC's Worldwide Software Taxonomy, 2023* (IDC #US50513623, April 2023)
- *IDC Market Glance: Machine Learning Life-Cycle Tools and Technologies, 1Q23* (IDC #US50007623, February 2023)

Synopsis

This IDC study evaluates vendors that offer data labeling software technologies and capabilities. New AI/ML applications using unstructured data for conversational AI, speech recognition, sentiment analysis, object detection in autonomous driving and robotics applications, medical imaging, facial recognition, and more will fuel the machine learning life-cycle market expansion. Unstructured data is a resource for enterprises to mine, extract actionable insights, and improve business outcomes, but it must be enriched with high-quality labels to unlock its value within AI/ML algorithms.

"Data labeling can be expensive, slow, and labor intensive, but it's necessary to derive value from AI/ML models using unstructured data, such as text, images, videos, and audio data," said Kathy Lange, research director for AI Software at IDC. "New AI-assisted software capabilities can dramatically reduce the effort traditionally required by human labelers. New human-in-the-loop solutions provide dramatically reduced time and costs for data labeling while delivering similar or improved quality and accuracy of results."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com