Data labeling, or data annotation, is part of the preprocessing stage when developing a machine learning (ML) model. It requires the identification of raw data (i.e., images, text files, videos), and then the addition of one or more labels to that data to specify its context for the models, allowing the machine learning model to make accurate predictions.

Data labeling underpins different machine learning and deep learning use cases, including computer vision and natural language processing (NLP).

Companies integrate software, processes and data annotators to clean, structure and label data. This training data becomes the foundation for machine learning models. These labels allow analysts to isolate variables within datasets, and this, in turn, enables the selection of optimal data predictors for ML models. The labels identify the appropriate data vectors to be pulled in for model training, where the model, then, learns to make the best predictions.

Along with machine assistance, data labeling tasks require “human-in-the-loop (HITL)” participation. HITL leverages the judgment of human “data labelers” toward creating, training, fine-tuning and testing ML models. They help guide the data labeling process by feeding the models datasets that are most applicable to a given project.

- Labeled data is used in supervised learning, whereas unlabeled data is used in unsupervised learning .

- Labeled data is more difficult to acquire and store (i.e. time consuming and expensive), whereas unlabeled data is easier to acquire and store.

- Labeled data can be used to determine actionable insights (e.g. forecasting tasks), whereas unlabeled data is more limited in its usefulness. Unsupervised learning methods can help discover new clusters of data, allowing for new categorizations when labeling.

Data labeling approaches

Data labeling is a critical step in developing a high-performance ML model. Though labeling appears simple, it’s not always easy to implement. As a result, companies must consider multiple factors and methods to determine the best approach to labeling. Since each data labeling method has its pros and cons, a detailed assessment of task complexity, as well as the size, scope and duration of the project is advised.

Here are some paths to labeling your data:

- **Internal labeling** - Using in-house data science experts simplifies tracking, provides greater accuracy, and increases quality. However, this approach typically requires more time and favors large companies with extensive resources.

- **Synthetic labeling** - This approach generates new project data from pre-existing datasets, which enhances data quality and time efficiency. However, synthetic labeling requires extensive computing power, which can increase pricing.

- **Programmatic labeling** - This automated data labeling process uses scripts to reduce time consumption and the need for human annotation. However, the possibility of technical problems requires HITL to remain a part of the quality assurance (QA) process.

- **Crowdsourcing** - This approach is quicker and more cost-effective due to its micro-tasking capability and web-based distribution. However, worker quality, QA, and project management vary across crowdsourcing platforms. One of the most famous examples of crowdsourced data labeling is Recaptcha. This project was two-fold in that it controlled for bots while simultaneously improving data annotation of images. For example, a Recaptcha prompt would ask a user to identify all the photos containing a car to prove that they were human, and then this program could check itself based on the results of other users. The input of from these users provided a database of labels for an array of images.

Benefits and challenges of data labeling

The general tradeoff of data labeling is that while it can decrease a business’s time to scale, it tends to come at a cost. More accurate data generally improves model predictions, so despite its high cost, the value that it provides is usually well worth the investment. Since data annotation provides more context to datasets, it enhances the performance of exploratory data analysis as well as machine learning (ML) and artificial intelligence (AI) applications. For example, data labeling produces more relevant search results across search engine platforms and better product recommendations on e-commerce platforms. Let’s delve deeper into other key benefits and challenges:

Benefits

Data labeling provides users, teams and companies with greater context, quality and usability. More specifically, you can expect:

- **More Precise Predictions:** Accurate data labeling ensures better quality assurance within machine learning algorithms, allowing the model to train and yield the expected output. Otherwise, as the old saying goes, “garbage in, garbage out.” Properly labeled data provide the “ground truth” (i.e., how labels reflect “real world” scenarios) for testing and iterating subsequent models.

Challenges

Data labeling is not without its challenges. In particular, some of the most common challenges are:

- **Expensive and time-consuming:** While data labeling is critical for machine learning models, it can be costly from both a resource and time perspective. If a business takes a more automated approach, engineering teams will still need to set up data pipelines prior to data processing, and manual labeling will almost always be expensive and time-consuming.

- **Prone to Human-Error:** These labeling approaches are also subject to human-error (e.g. coding errors, manual entry errors), which can decrease the quality of data. This, in turn, leads to inaccurate data processing and modeling. Quality assurance checks are essential to maintaining data quality.

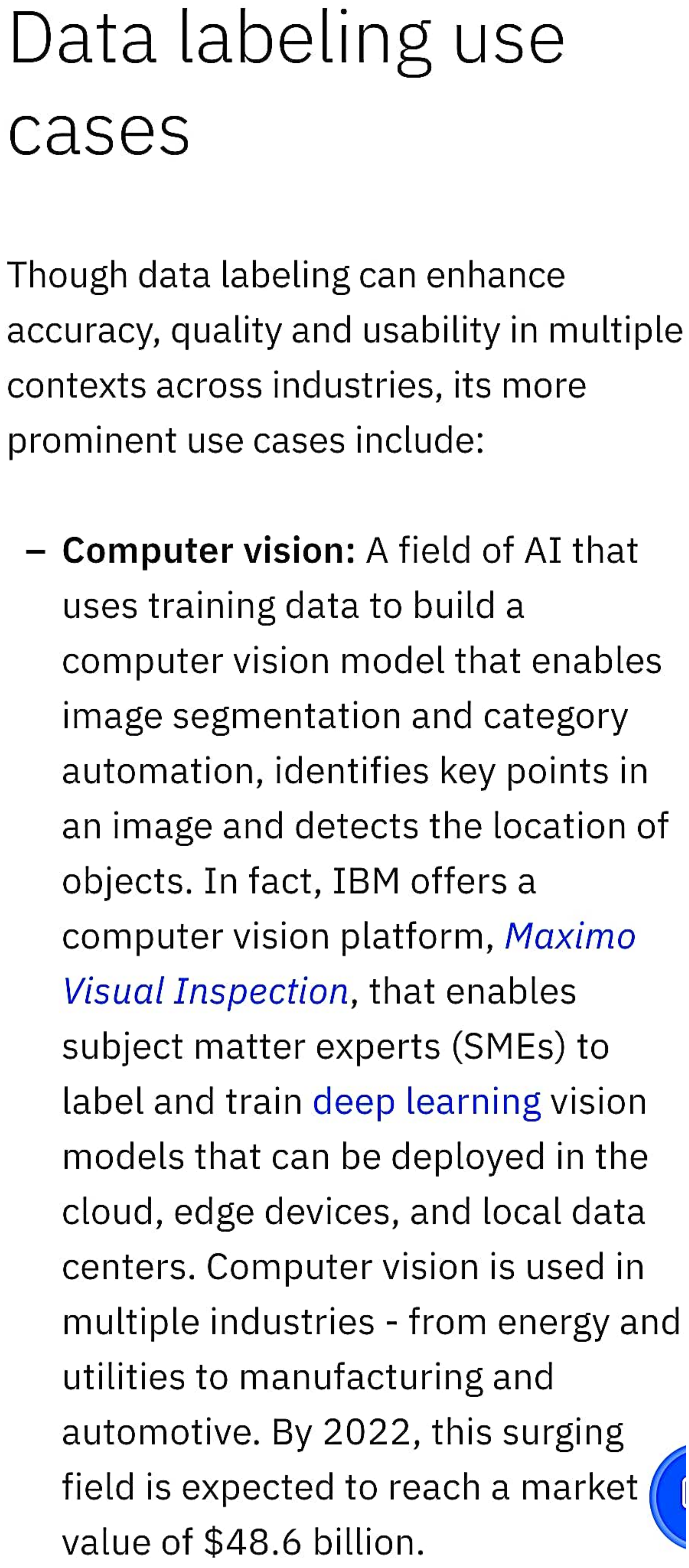
Data labeling best practices

No matter the approach, the following best practices optimize data labeling accuracy and efficiency:

- **Intuitive and streamlined task interfaces** minimize cognitive load and context switching for human labelers.

- **Consensus:** Measures the rate of agreement between multiple labelers(human or machine). A consensus score is calculated by dividing the sum of agreeing labels by the total number of labels per asset.

- **Label auditing:** Verifies the accuracy of labels and updates them as needed.



Data labeling use cases

Though data labeling can enhance accuracy, quality and usability in multiple contexts across industries, its more prominent use cases include:

- **Computer vision:** A field of AI that uses training data to build a computer vision model that enables image segmentation and category automation, identifies key points in an image and detects the location of objects. In fact, IBM offers a computer vision platform, *Maximo Visual Inspection*, that enables subject matter experts (SMEs) to label and train deep learning vision models that can be deployed in the cloud, edge devices, and local data centers. Computer vision is used in multiple industries - from energy and utilities to manufacturing and automotive. By 2022, this surging field is expected to reach a market value of $48.6 billion.

IBM offers more resources to help transcend data labeling challenges and maximize your overall data labeling experience.

- IBM Cloud Annotations - A collaborative open-source image annotation tool that uses AI models to help developers create fully labeled datasets of images, in real time, without manually drawing the labels.

- IBM Cloud Object Storage - Encrypted at-rest and accessible from anywhere, it stores sensitive data and safeguards data integrity, availability and confidentiality via Information Dispersal Algorithm (IDA) and All-or-Nothing Transform (AONT).

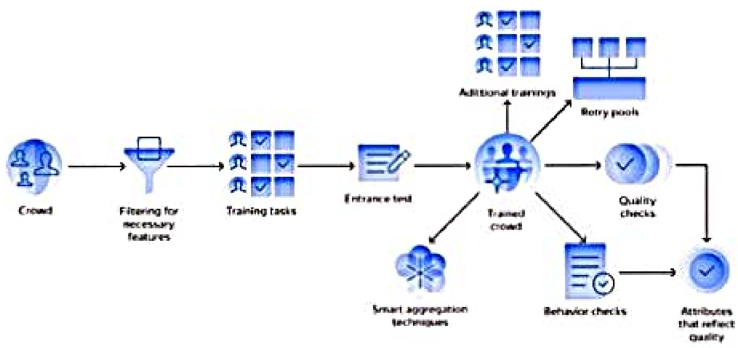
IBM and data labeling

IBM offers more resources to help transcend data labeling challenges and maximize your overall data labeling experience.

- IBM Cloud Annotations - A collaborative open-source image annotation tool that uses AI models to help developers create fully labeled datasets of images, in real time, without manually drawing the labels.

- IBM Cloud Object Storage - Encrypted at-rest and accessible from anywhere, it stores sensitive data and safeguards data integrity, availability and confidentiality via Information Dispersal Algorithm (IDA) and All-or-Nothing Transform (AONT).

Type of Data:	Most, including text, image, audio, and video
Availability:	Proprietary
Approach:	Crowdsourcing
Industries served:	eCommerce, Retail, Automotive, Cybersecurity, Banking, Sports. Legal Tech, Research, Manufacturing, Healthcare.
Price-to-quality ratio:	Above average
Time-effectiveness:	High
Instructions for beginners:	Yes (Knowledge base)
Learning curve for the user:	Average to Fast

Mechanisms for quality control:	Asynchronous quality control: assignment review, smart aggregation methods.  Crowd Filtering for necessary features Training tasks Entrance test Trained crowd Aditional trainings Retry pools Quality checks Smart aggregation techniques Behavior checks Attributes that reflect quality *Source: Toloka website*
Computational power:	None needed for the client

Type of Data:	Most
Availability:	Proprietary
Approach:	Crowdsourcing
Industries served:	Most industries
Price-to-quality ratio:	Above average
Time-effectiveness:	High
Instructions for beginners:	Yes (tutorials and support)
Learning curve for the user:	Average to Fast
Mechanisms for quality control:	Qualifications allow you to select or create specific Worker eligibility requirements for your projects. MTurk offers three types of pre-defined Qualifications — Masters, System, and Premium Qualifications.
Computational power:	None needed for the client