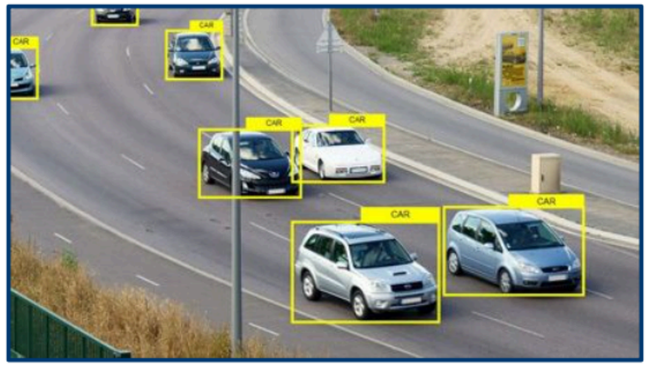
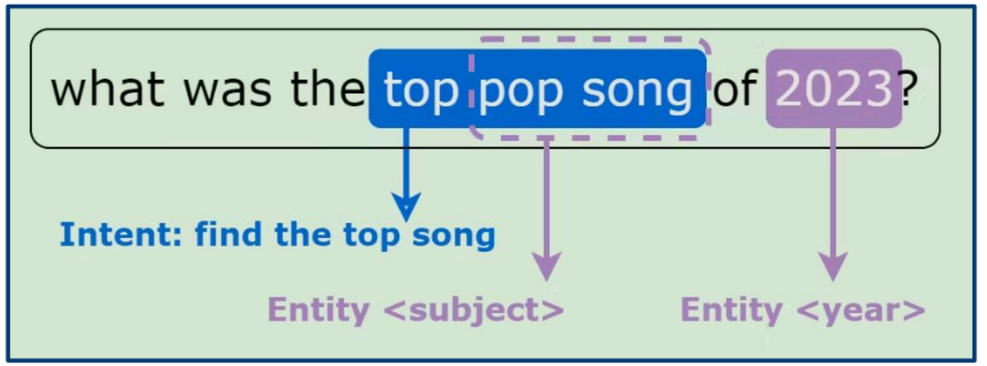
Data annotation is the process of adding meaningful labels to a dataset enabling models to recognize patterns and make accurate predictions

Bounding boxes for images



Text classification



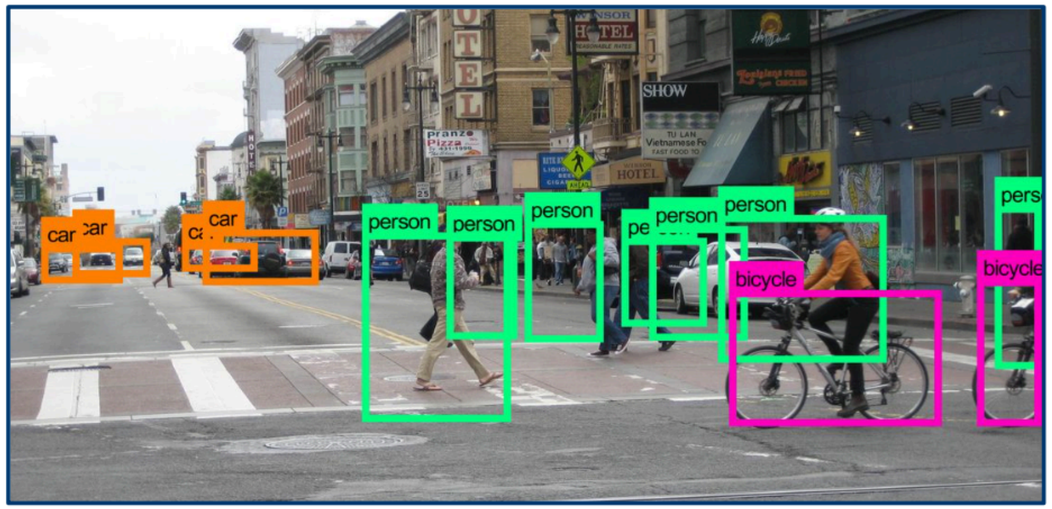
Annotations, aligned with defined labeling guidelines, enrich datasets with metadata that enhances the value and performance of models

Data annotation in model development

Training	**Evaluation**	**Tuning**
Data curation	*Output benchmarking*	*Improvement and contexting*

Annotation needs vary by use case and complexity

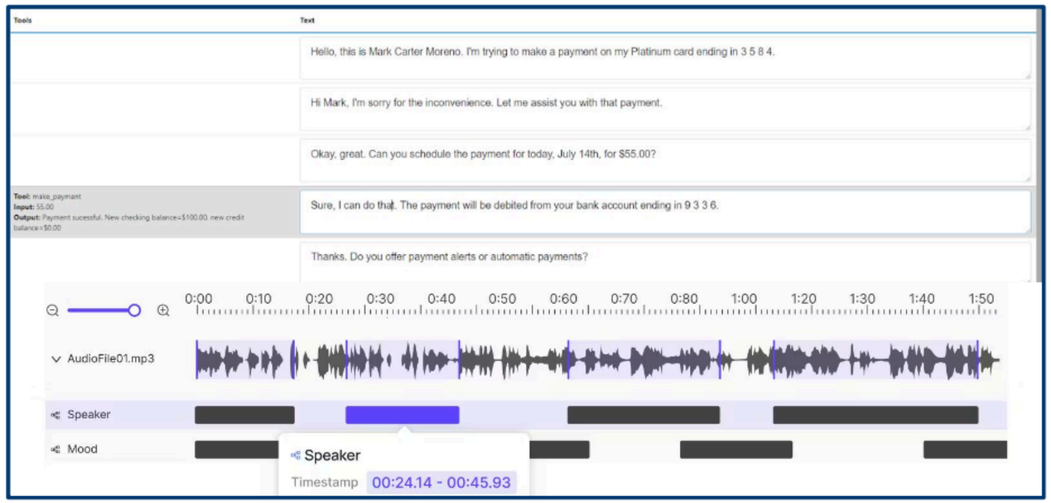
Annotating a Road Image for a Self-Driving Car



This process is **linear and objective**. The rules are concrete and the output is a set of geometric shapes with clear labels.

Complexity: Low. The annotator's task is primarily object recognition and spatial marking with little to no ambiguity. A car is a car, and a pedestrian is a pedestrian. The rules are clear and consistently applied.

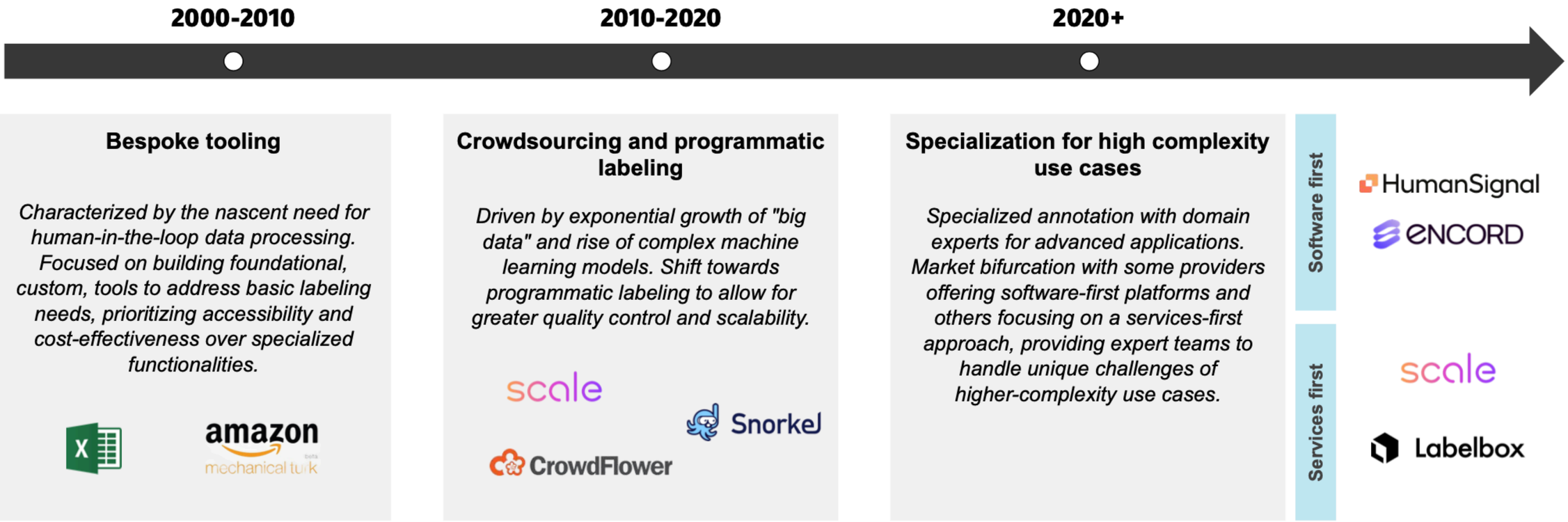
Annotating a Card Customer Voice Call Transcript



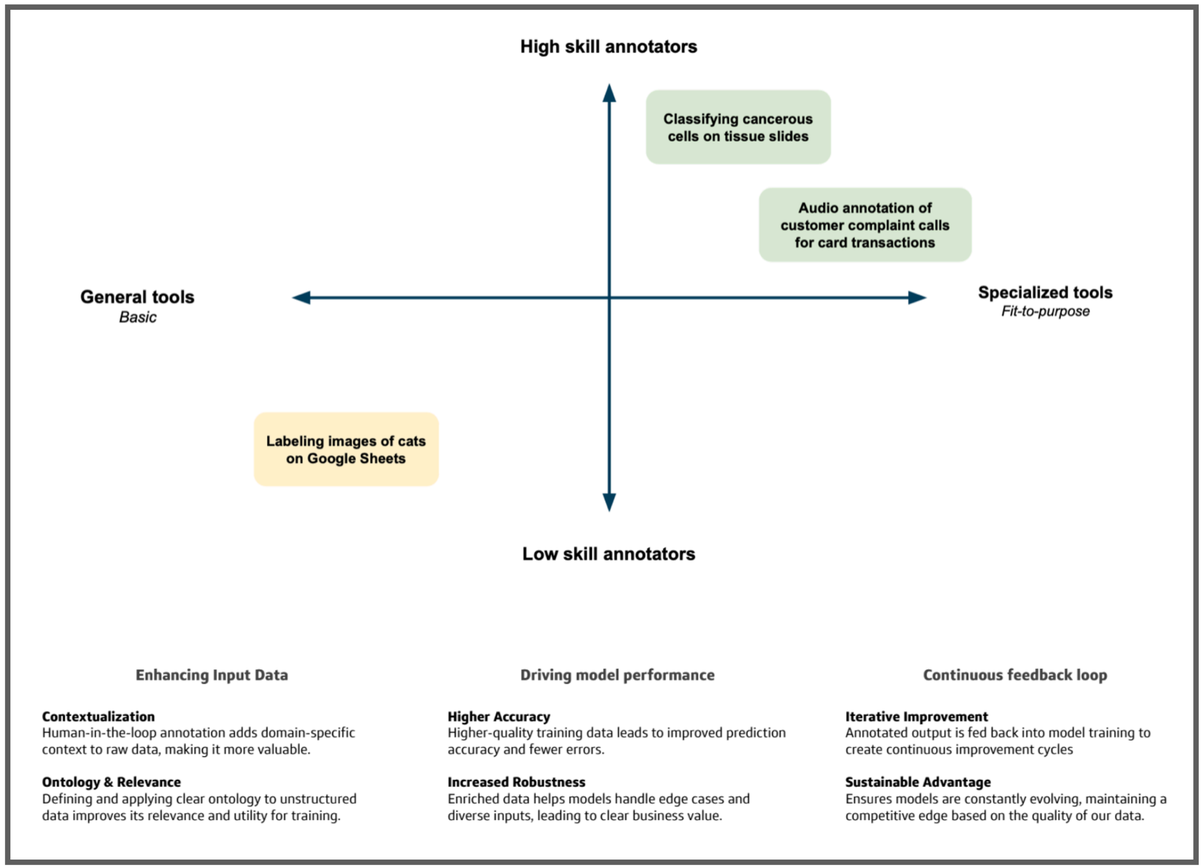
This process is **complex and interpretive.** The annotator must understand context, subtext, and implied meaning, with multiple decision points and subjective judgments required at each step.

Complexity: High. Requires annotators to make nuanced judgments, handle ambiguity, and synthesize information to accurately identify and re-frame a customer's true needs.

Eras of data annotation have been shaped by evolving market demand



In today’s market we are seeing a separation between providers who focus on the services offering versus those emphasizing a software offering



Annotation Solutions

Software-first

Self-service tool for teams to manage their own data annotation projects and workforce

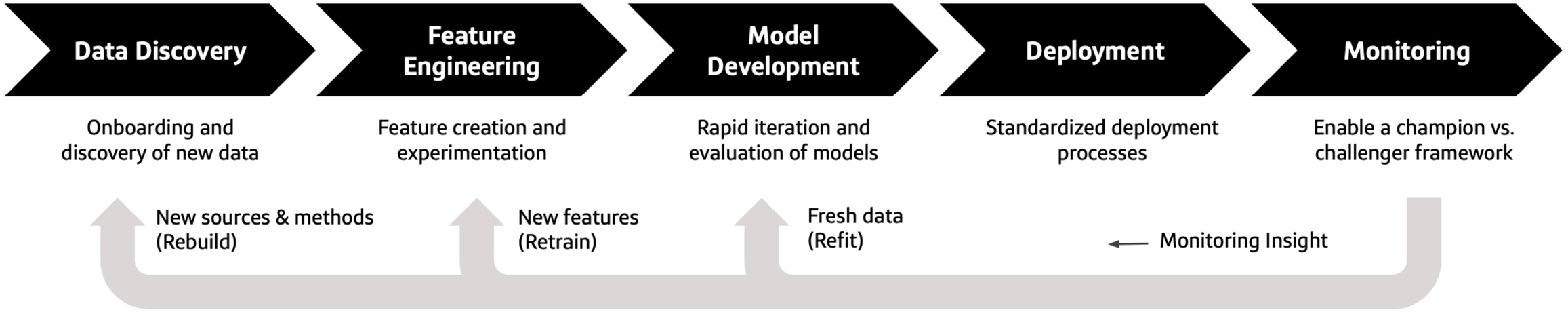
HumanSignal · ENCORD · Snorkel · SuperAnnotate · Datature · V7

Services-first

Workforce of trained annotators to label a customer's data as a managed service

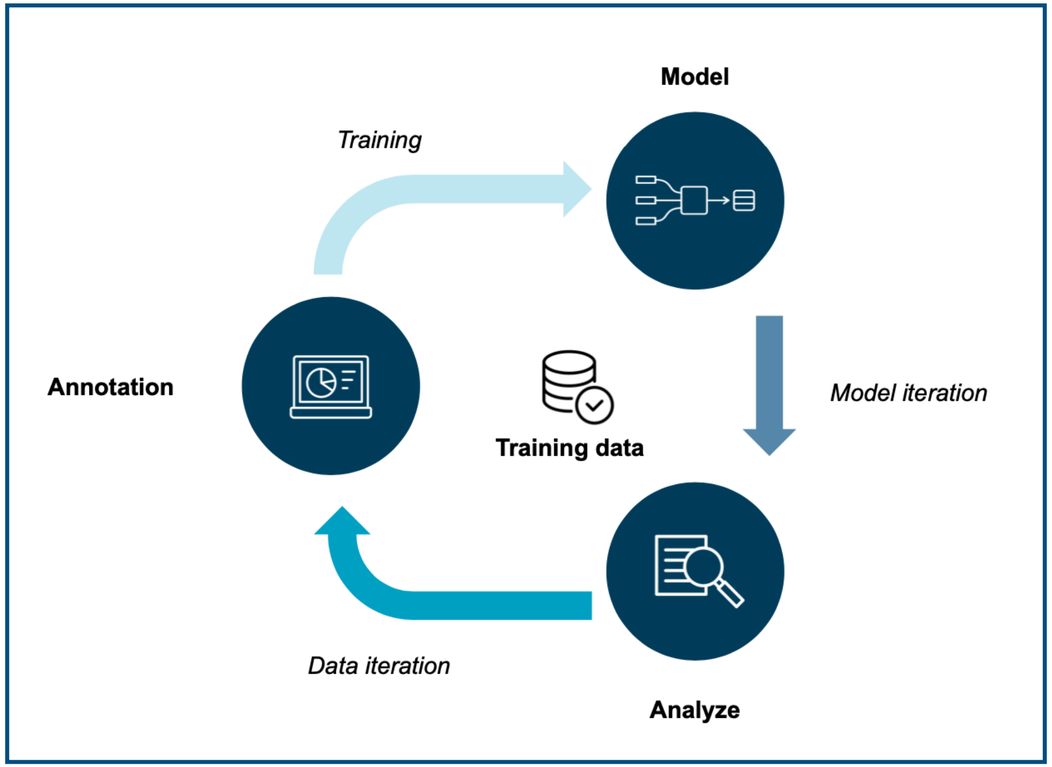
Labelbox · scale · surge AI · MERCOR · Appen · KEYMAKR

Annotation supports model development tasks on structured and unstructured data



Design Time	Training Time	Run Time
Data scientist	**Annotator**	**Product manager**
"I want to explore a small sample of call recordings to test new hypotheses that might lead to model improvements"	*"I need to mark the timestamps when customer complained during a call, over several hundred recorded calls"*	*"A model in production seems to be underperforming, so I need to examine model input/output pairs to diagnose"*

For enterprise applications, **data annotation platforms** will allow subject experts to fine-tune models in production to achieve higher performance and accuracy.



Training
Data curation

Evaluation
Output benchmarking

Tuning
Improvement and contexting

1. Training Models

Humans apply a label for the model to predict. The original record + the label is fed to a model to “teach” it to associate features with the target variable.

- **Structured Data:** Data that can be read as tables
- **Unstructured Data:** Data requiring advanced tools to apply bounding boxes to image/doc text, highlight strings, or label audio

2. Fine-Tuning Models

Adjusting parameters of the pre-trained model using a smaller, task-specific annotated datasets to improve its performance on a particular task.

3. Model Validation

A validation dataset is annotated by a human, and fed to the model to measure the difference between the human label and the system prediction.