



# Building a better AI data engine

Mark Ghannam



# Phases of AI development



## Phase 1

### Assembling

- Have an understanding of the problem to be solved
- Have internal champions (commitment)
- Has access and plans to obtain data



## Phase 2

### Building

- Possesses AI & ML talent
- Has models (in test or in production)
- Generating data from applications



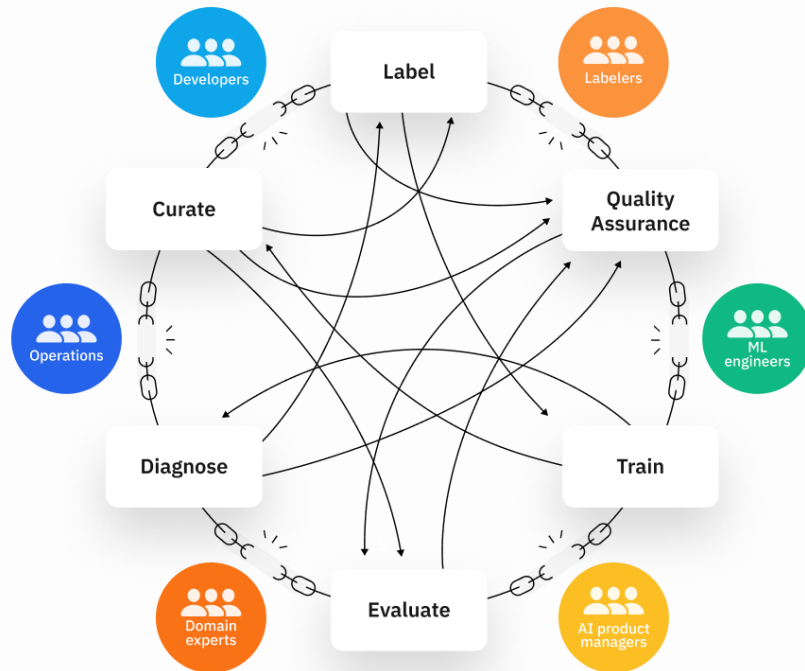
## Phase 3

### Productionizing

- Constantly training models (Refreshing data regularly)
- Labeling spend (dedicated budget and need for labeled data is exponential)
- Iterating & improving quality

Prototypes are easy.

**Production  
AI is *hard*.**



# Companies with the best training data produce the most performant models



**Free**

## **Algorithms**

State of the art deep learning algorithms are free and readily available.



**Commodity**

## **Compute**

Effective AI computation is doubling every 3-6 months.

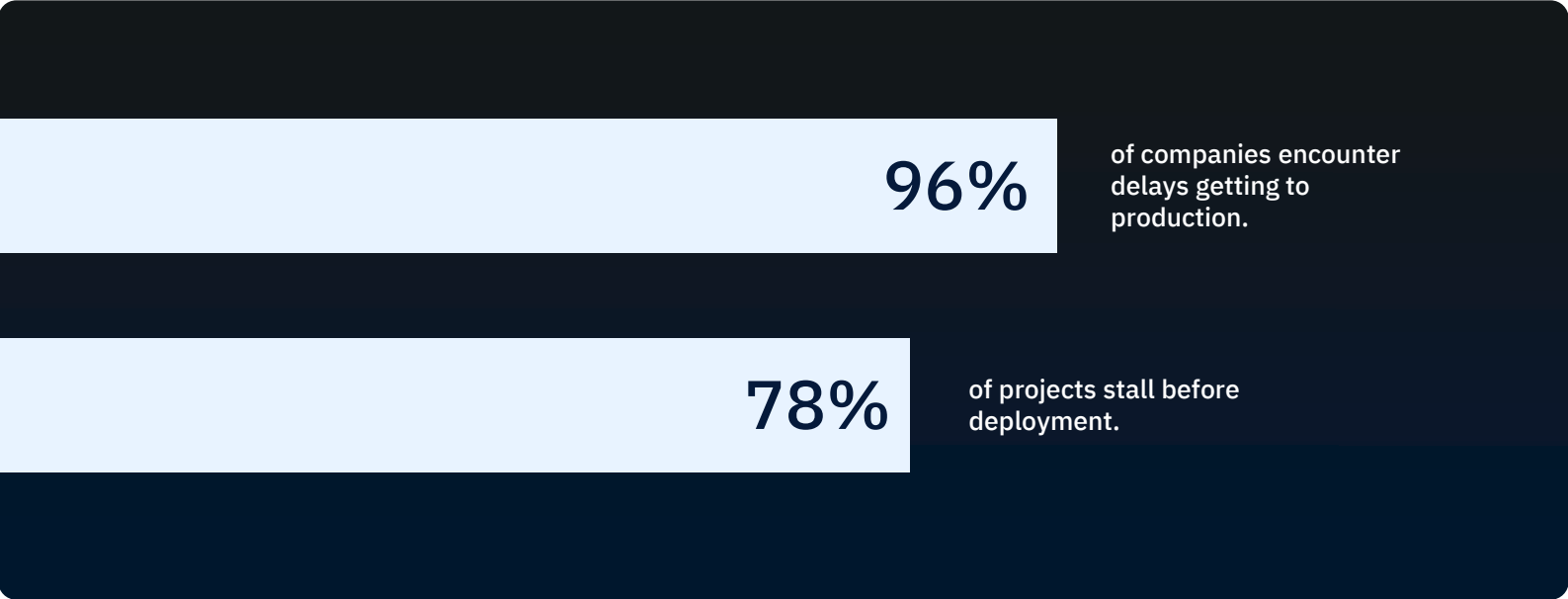


**Proprietary**

## **Data**

Training data is the new code. It is what makes AI work in the real world

# Converting proprietary data into revenue-generating AI is challenging



A horizontal bar chart with two bars. The top bar is light blue and contains the text '96%'. The bottom bar is also light blue and contains the text '78%'. The bars are set against a dark blue background.

Metric	Percentage
Companies encountering delays getting to production	96%
Projects stalling before deployment	78%

96%

of companies encounter delays getting to production.

78%

of projects stall before deployment.

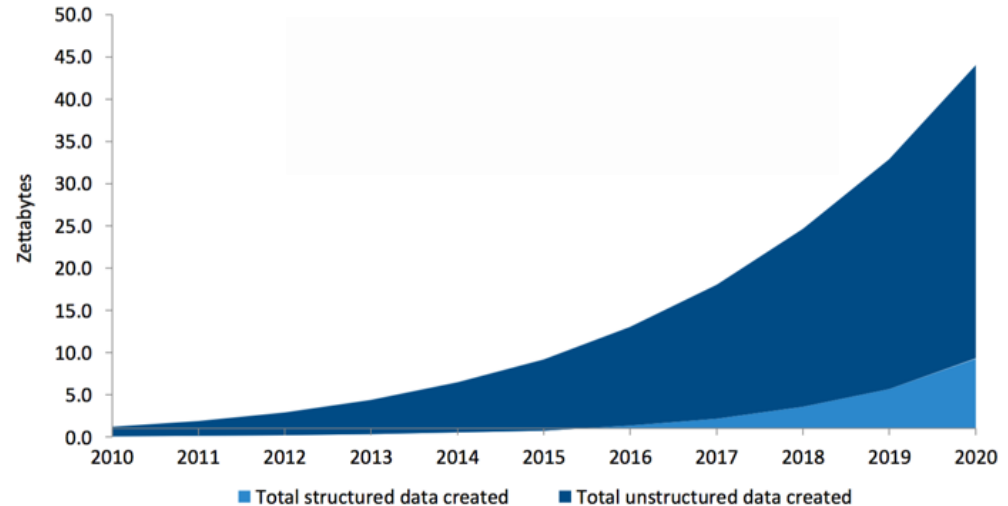
*Source: Algorithmia 2020 State of ML Survey*

# The Value of Unstructured Data

- Unstructured data makes up almost 80% of an enterprise's data
- Extracting the value from this data has been difficult as it involves complex and time consuming data analytics processes
- Change is afoot recent technologies make this data easier to leverage than ever:
  - Internet of things(IOT)
  - Machine Learning
  - Computer Vision
  - Document Understanding/ NLP

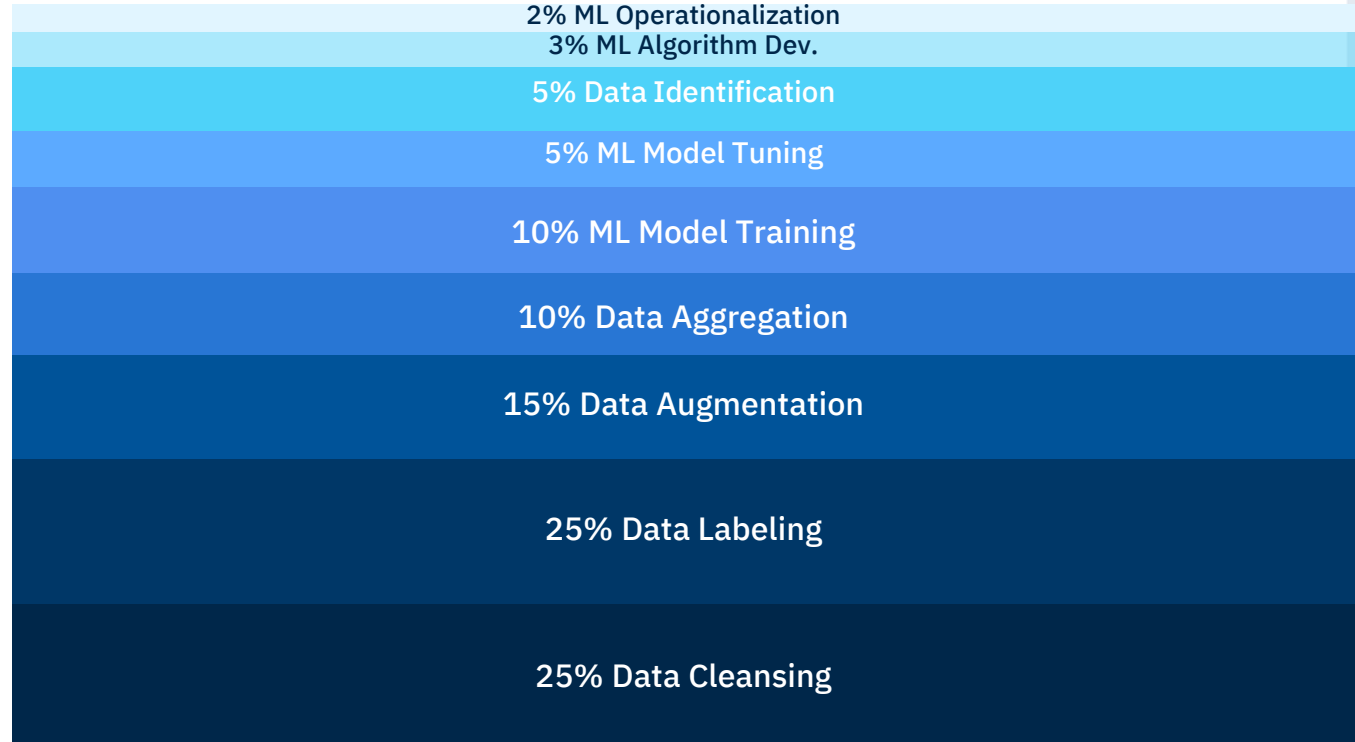
FIGURE 1

Capacity Growth by Data Type



Source: IDC, 2016

**Managing  
training data  
requires  
significant time  
& resources**

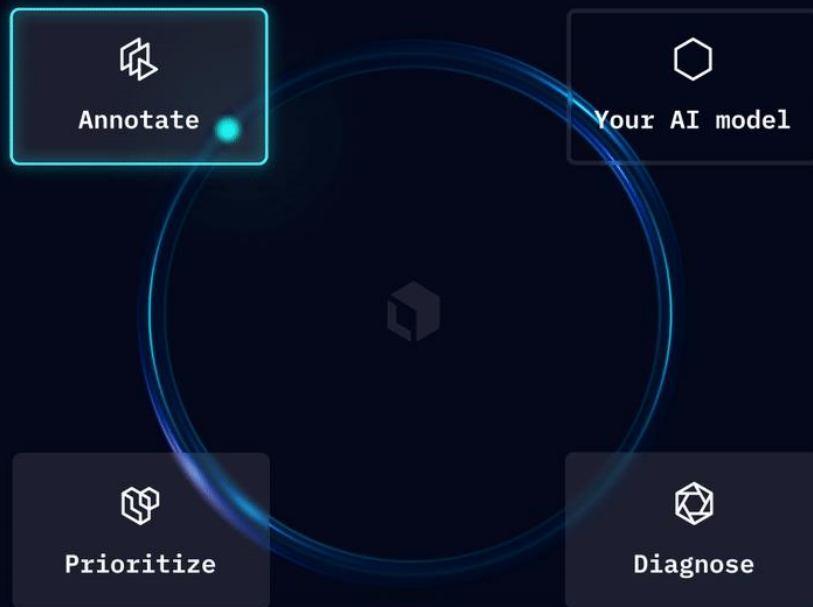


*Source: Cognilytica*





# Building an active learning **iteration loop**



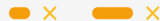
# Building your data engine

Data Sources



Data, metadata, and model predictions

Search



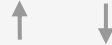
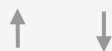
Explore



Visualize



Analyze



## Annotate

Label data across all data modalities

- AI assisted labeling
- Quality control
- Internal and external collaboration



## Model

Improve your data and your models

- Data versioning
- Experiment management
- Model training integration
- Model evaluation and testing



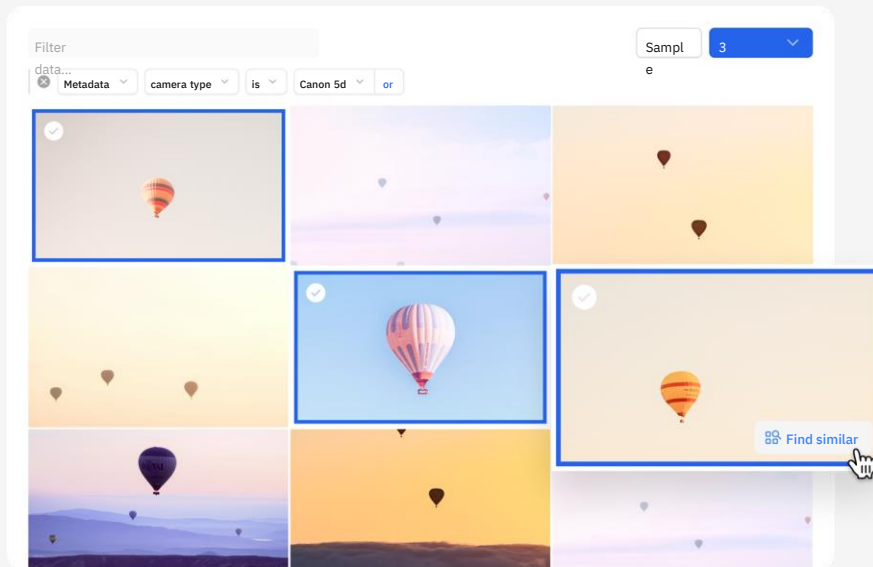
Catalog

# Manage unstructured data with precision

Visualize and search all  
your data in one place.

Find and prioritize the data  
that needs improvement.

Automatically assign custom  
metadata to assets in bulk.



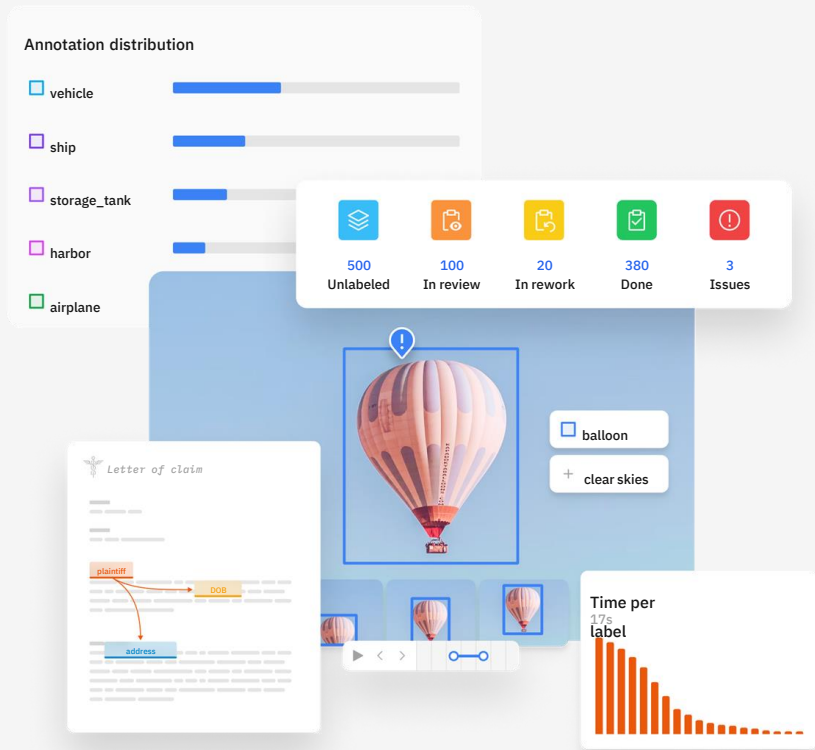


Annotate

# The most powerful labeling tools at your fingertips

Access a full suite of labeling, collaboration, and quality tools that give you full visibility and control.

Leverage automation and custom workflows to reduce costs and save time.



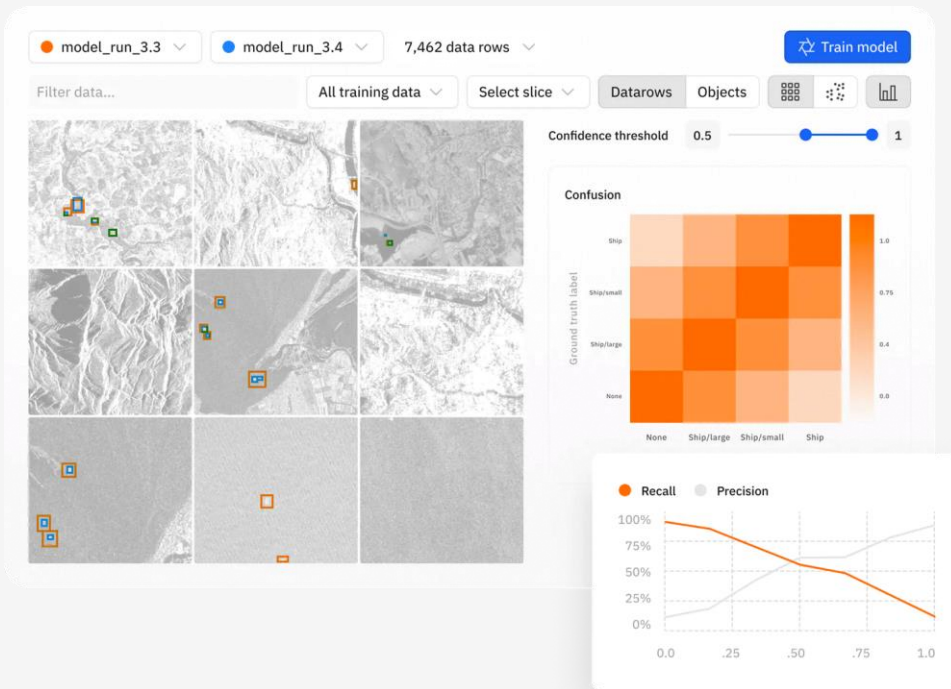


Model

# Train and evaluate AI models

Track model performance across versions with detailed metrics.

Find and fix model errors that will most dramatically improve performance.  
Manage model configurations and data selection in one place.



## Case Study

# Airbnb

### Problem

Flexibility is a key trend emerging from the pandemic as people become less tethered and the lines blur between travel, living and working

### Solution

Airbnb leveraged Labelbox's annotate, catalog, and automation products to rapidly tag/categorize around 6 million live listings.

### Result

Flexible Destinations will help to spread travel with many unique listings located outside of popular tourist destinations. It will also help to surface and drive demand to unique listings. Hosts who have collectively earned more than \$300 million globally since the start of the pandemic

