Appen Limited
Level 6, 9 Help Street
Chatswood NSW 2067
Tel: 02 9468 6300
www.appen.com
ACN 138 878 298

**ASX ANNOUNCEMENT**
20 May 2021

**INVESTOR TECHNOLOGY DAY PRESENTATION**

Appen Limited (ASX: APX) is holding an online Investor Technology Day event commencing at 11.00am AEST today.
To join the webcast, please pre-register at https://kapara.rdbk.com.au/landers/12ee87.html. Registration will close at 10.00am AEST today.

The presentation materials for the event are attached.

---

Authorised for release by the CEO.

**Please contact for more information:**

Linda Carroll
Investor Relations
+61 2 9468 6300
investorrelations@appen.com
www.appen.com/investors

**About Appen Limited**

Appen is a global leader in the development of high-quality, human annotated datasets for machine learning and artificial intelligence. Appen brings 25 years of experience collecting and enriching a wide variety of data types including speech, text, image and video. With expertise in 235 languages and access to a global crowd of over 1,000,000 skilled contractors, Appen partners with leading technology, automotive and ecommerce companies - as well as governments worldwide - to help them develop, enhance and use products that rely on natural languages and machine learning. www.appen.com

appen

# Appen Limited

## Investor Technology Day

20 May 2021

# Agenda

| | |
|---|---|
| 11:00am | **CEO's opening remarks and introductions**<br>Mark Brayan |
| 11:15am | **AI market update**<br>Ryan Kolln, VP Corporate Development |
| 12:00pm | **Break** |
| 12:20pm | **How our technology creates value for customers**<br>Wilson Pang, CTO |
| 1:20pm | **Q&A** |
| 1:55pm | **CEO's closing remarks**<br>Mark Brayan |
| 2:00pm | **End** |

# Introductions

**Mark Brayan**

CEO

**Wilson Pang**

CTO

**Ryan Kolln**

VP Corporate Development

# Appen is accelerating its transformation into an AI powered provider of AI data and solutions

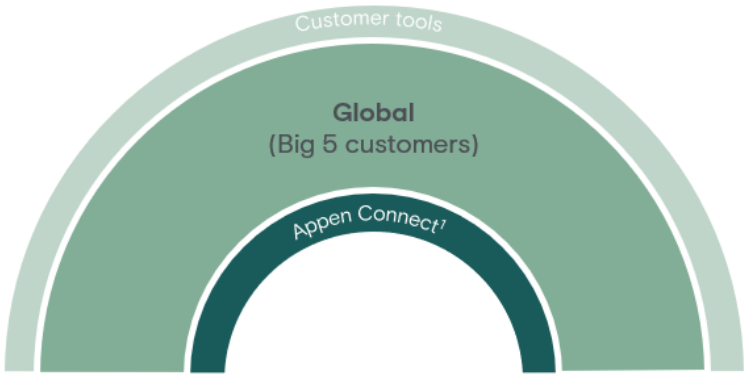
# Appen continues to evolve

| | From | To |
|---|---|---|
| **Data type** | Language data | AI data |
| **Delivery model** | Service led | Product led |
| **Revenue** | Project based | Committed |
| **Customer** | Concentration | Diversification |
| **Org structure** | Functional alignment | Customer alignment |
| **Reporting** | Data modality, AUD | Strategy led, USD |

# Appen's evolution

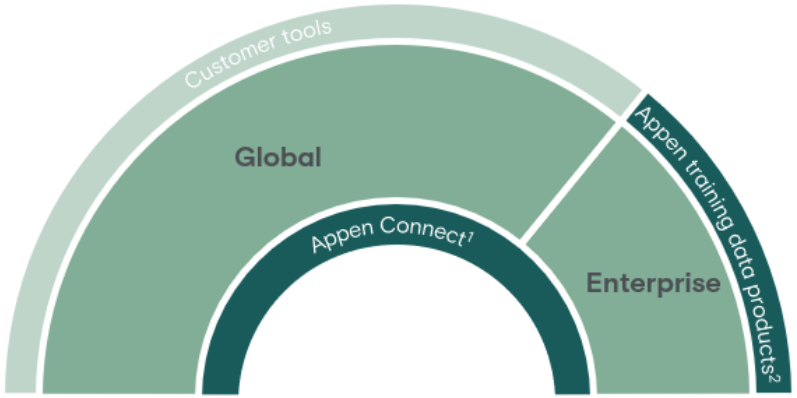
## Phase 1: Dominate the AI data annotation services market

Acquired Butler Hill 2010
Acquired Leapforce 2017



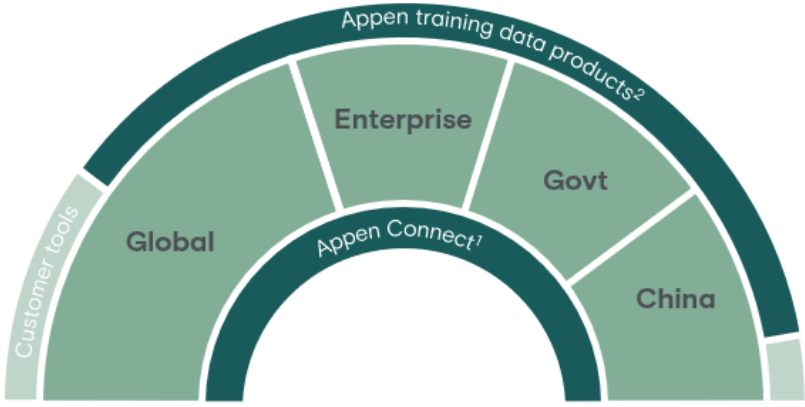

## Phase 2: Build leading data annotation products

Acquired Figure Eight 2019
Ongoing product investment
New Enterprise customer base




## Phase 3: Product-led expansion

Increased sales and marketing China and Govt BUs established
Global customers on Appen products



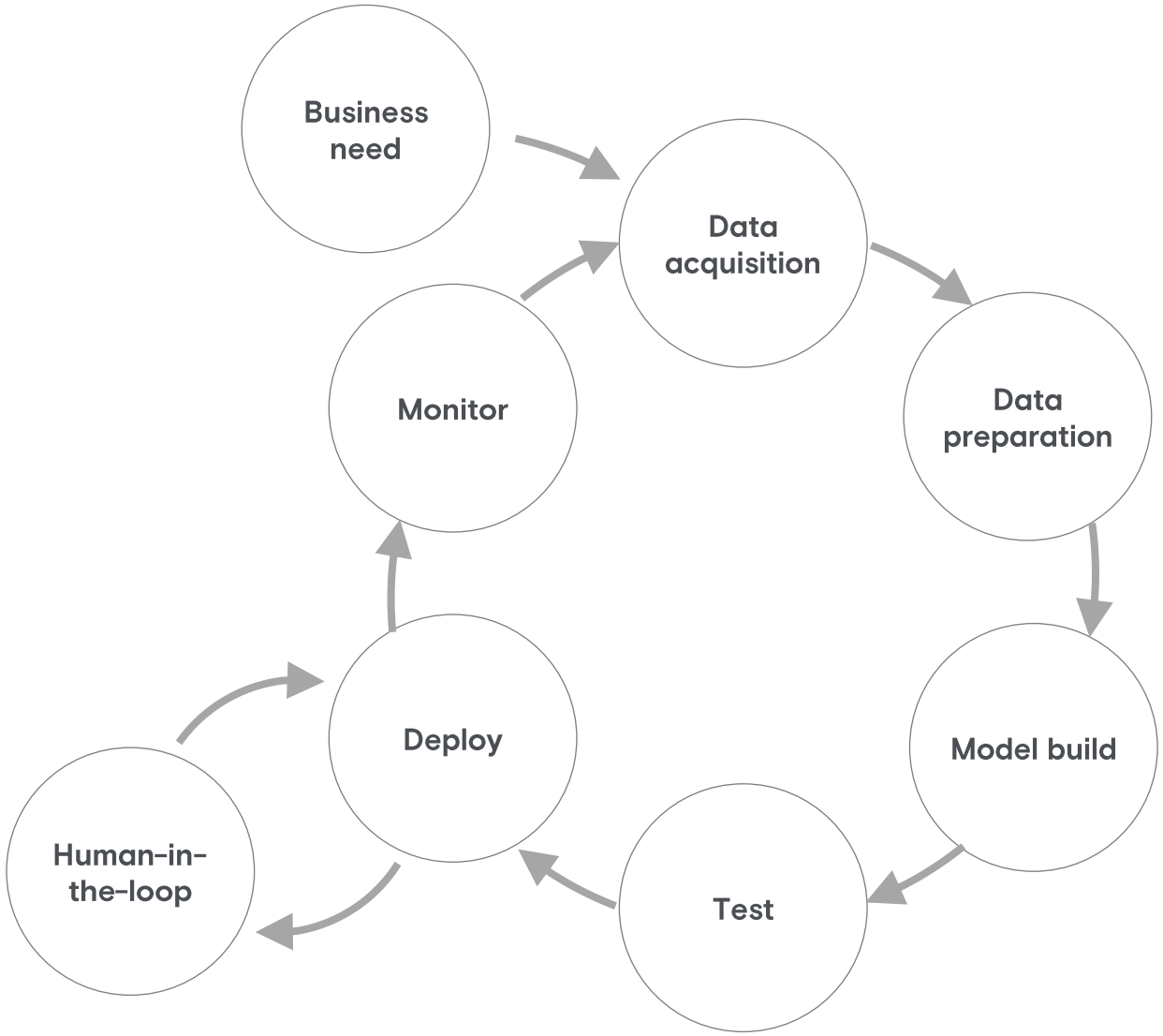



appen

# Our areas of focus

| | |
|---|---|
| **Product Led** | Our future is product led, enabling us to deliver high-quality training data, faster, at larger scale, with improved unit economics, and is a foundation for future capabilities |
| **Customer Centric** | We are aligning our operations to better support the needs of our target customer cohorts |

# Appen plays an important role in the AI application lifecycle

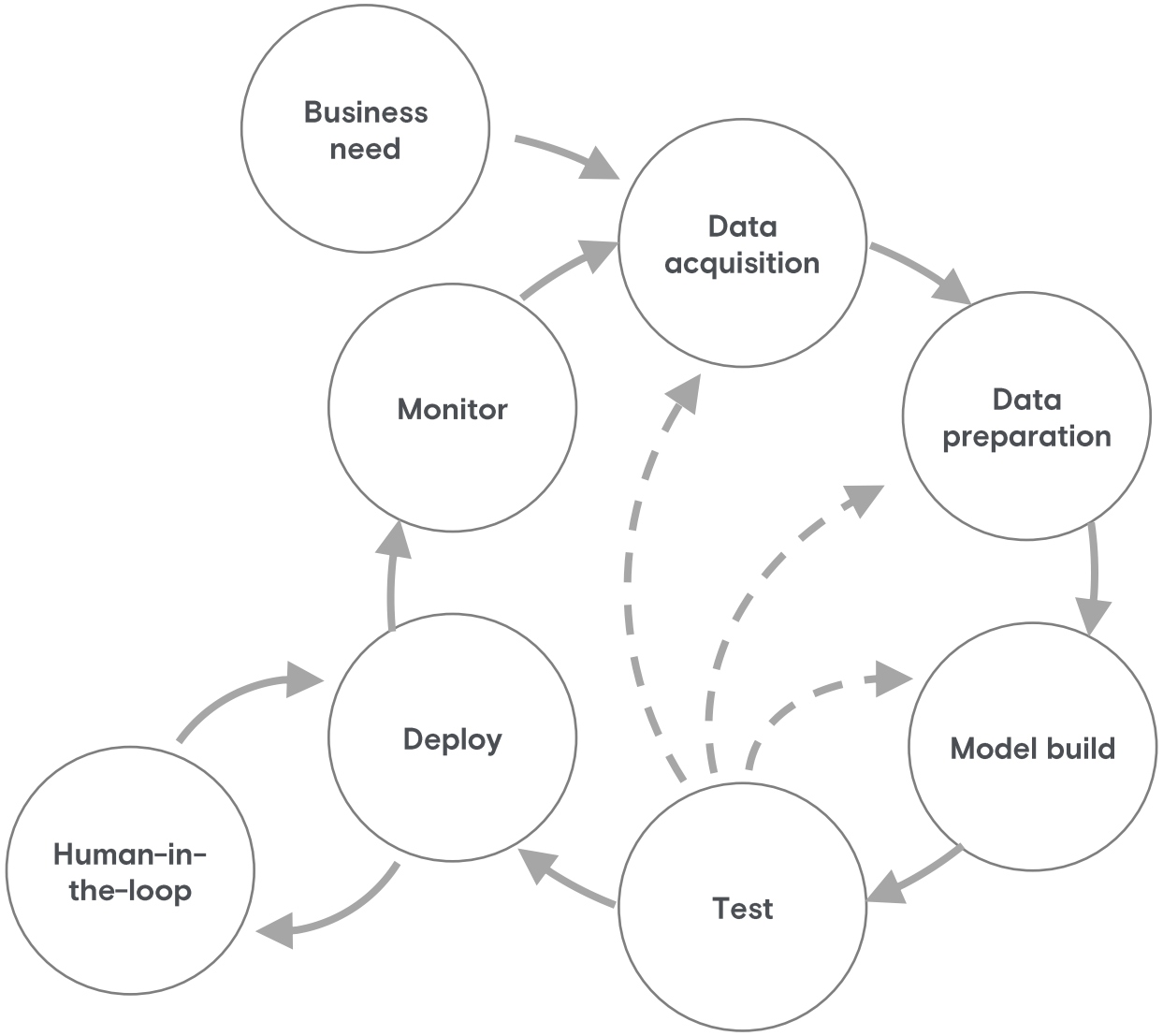



# AI lifecycle

An ongoing, iterative approach



appen




# AI lifecycle

An ongoing, iterative approach ... with multiple iteration loops



appen

AI model = Model instructions + Training Data

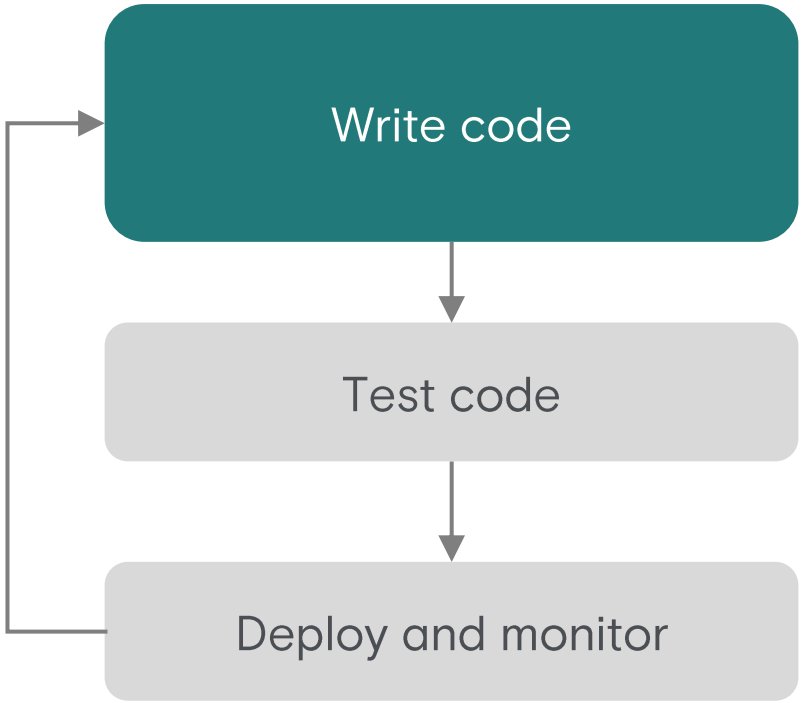
AI model = Model instructions + Training Data

Architecture for the model to learn – can be as little as 10 lines of code

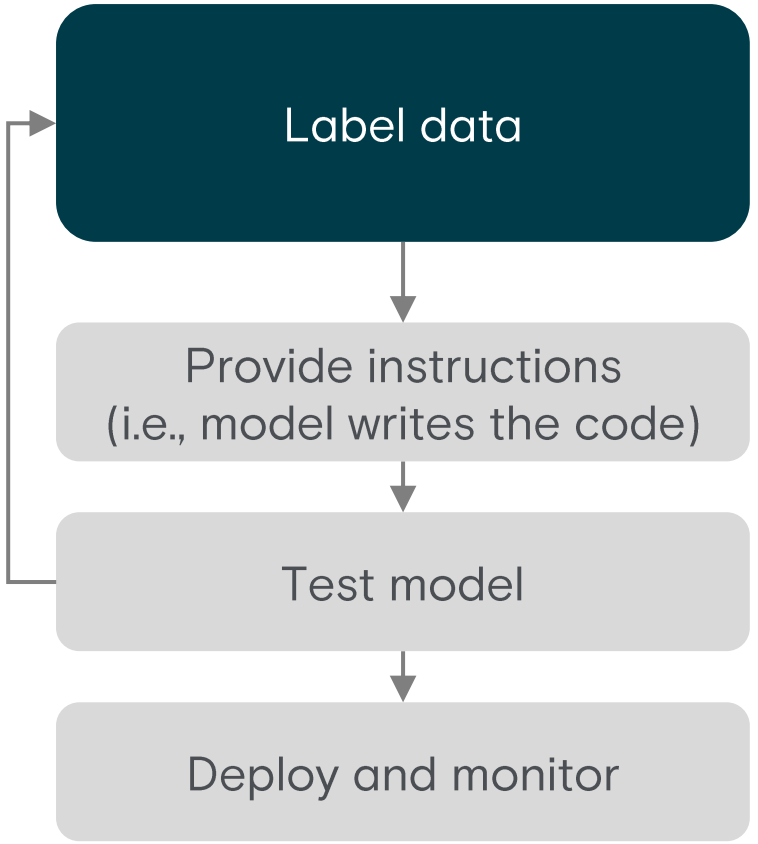
Examples used in the model training process, typically more the better

# In AI, labelling data is *the most* important step

**Traditional software development**



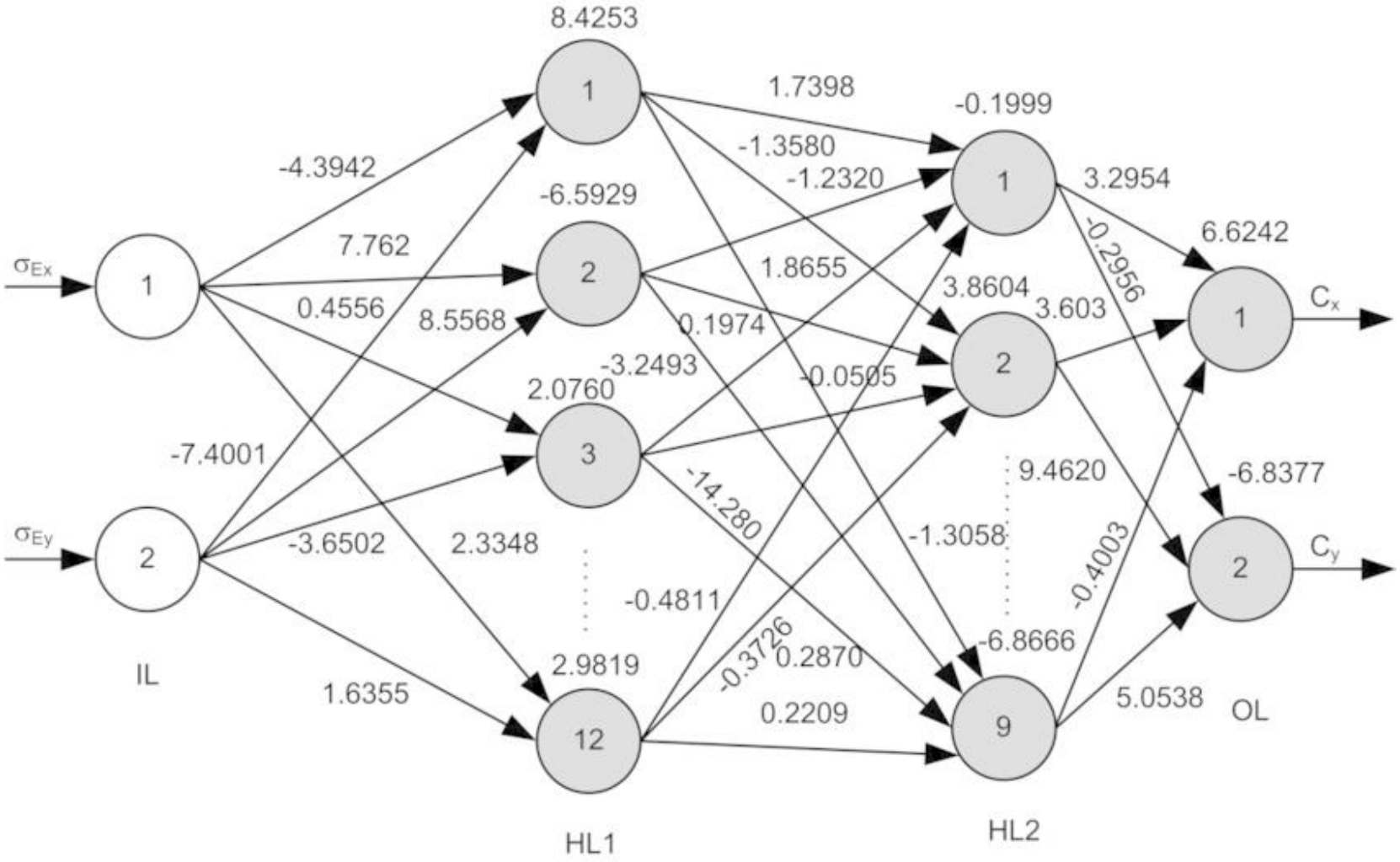

**AI model development**






appen

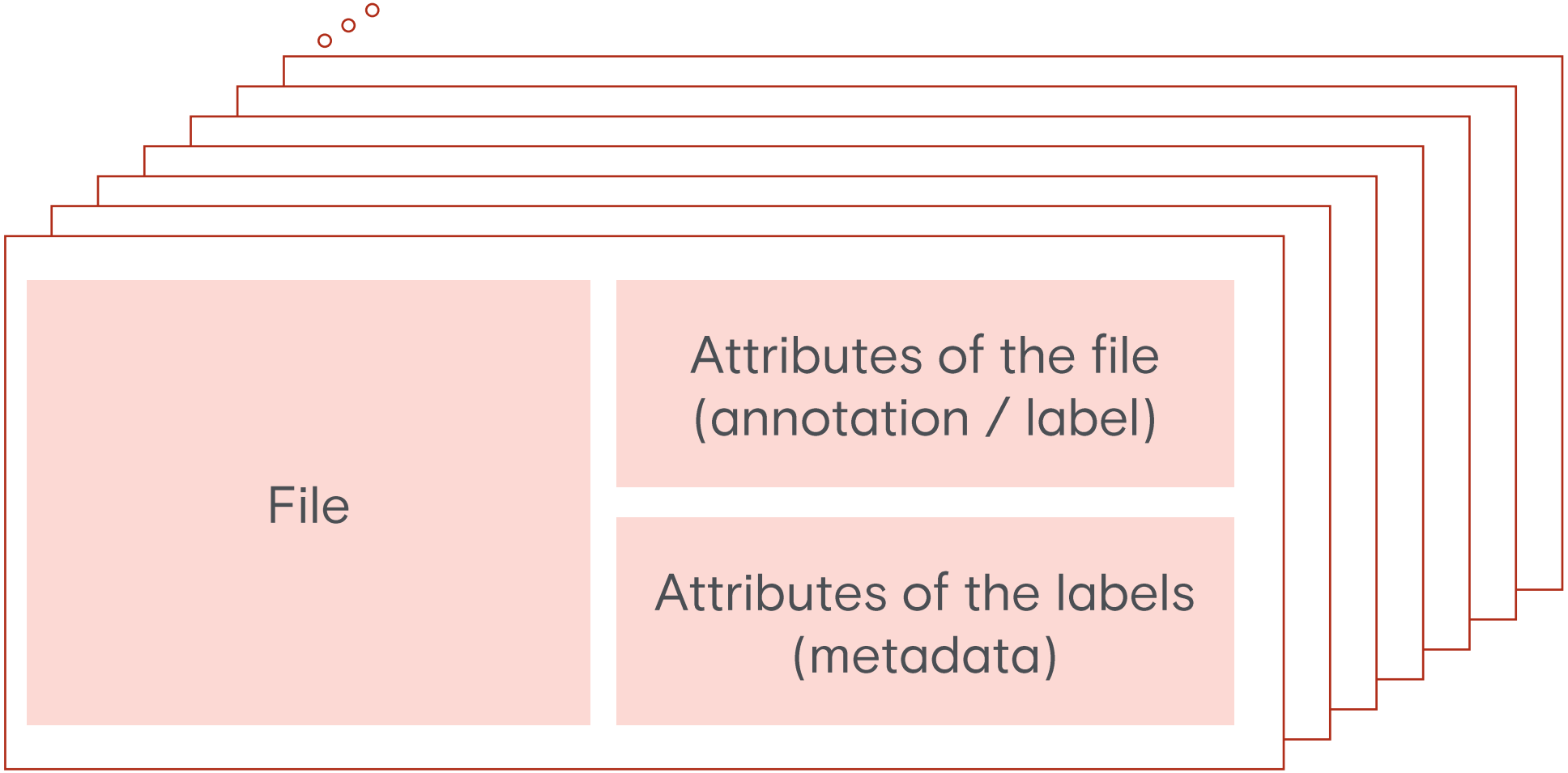
# AI models are a series of nodes, weights and biases






appen

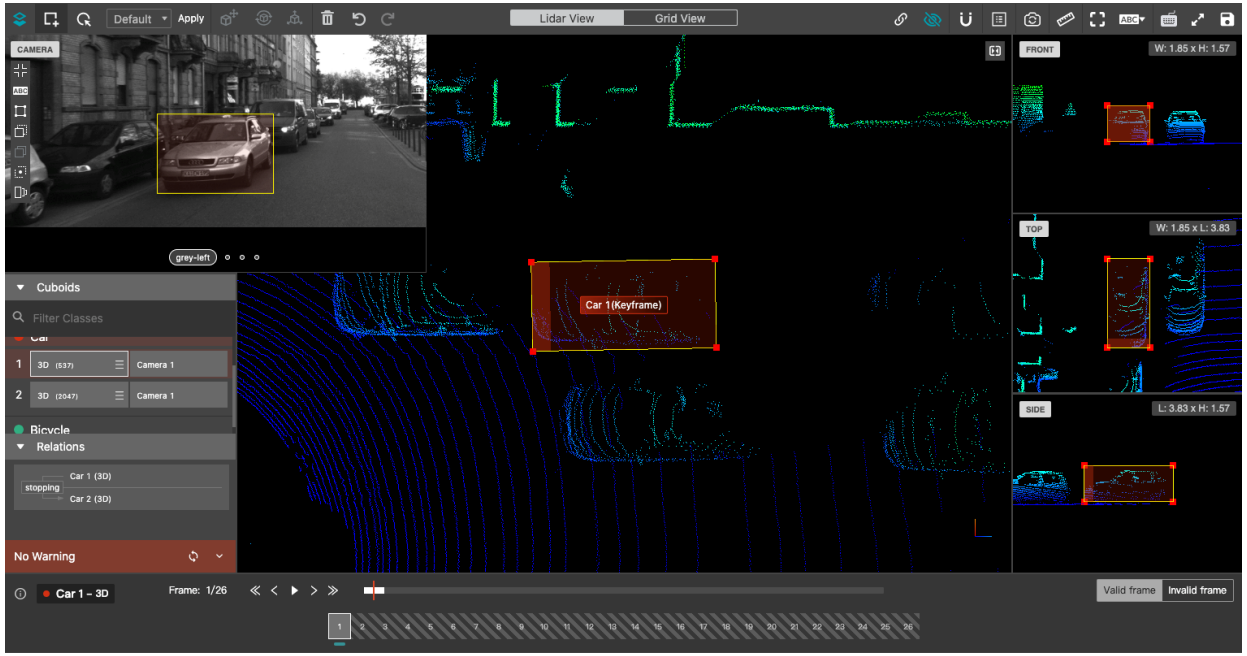
# Training data are the examples that “teach” the model






appen

# LIDAR training data example

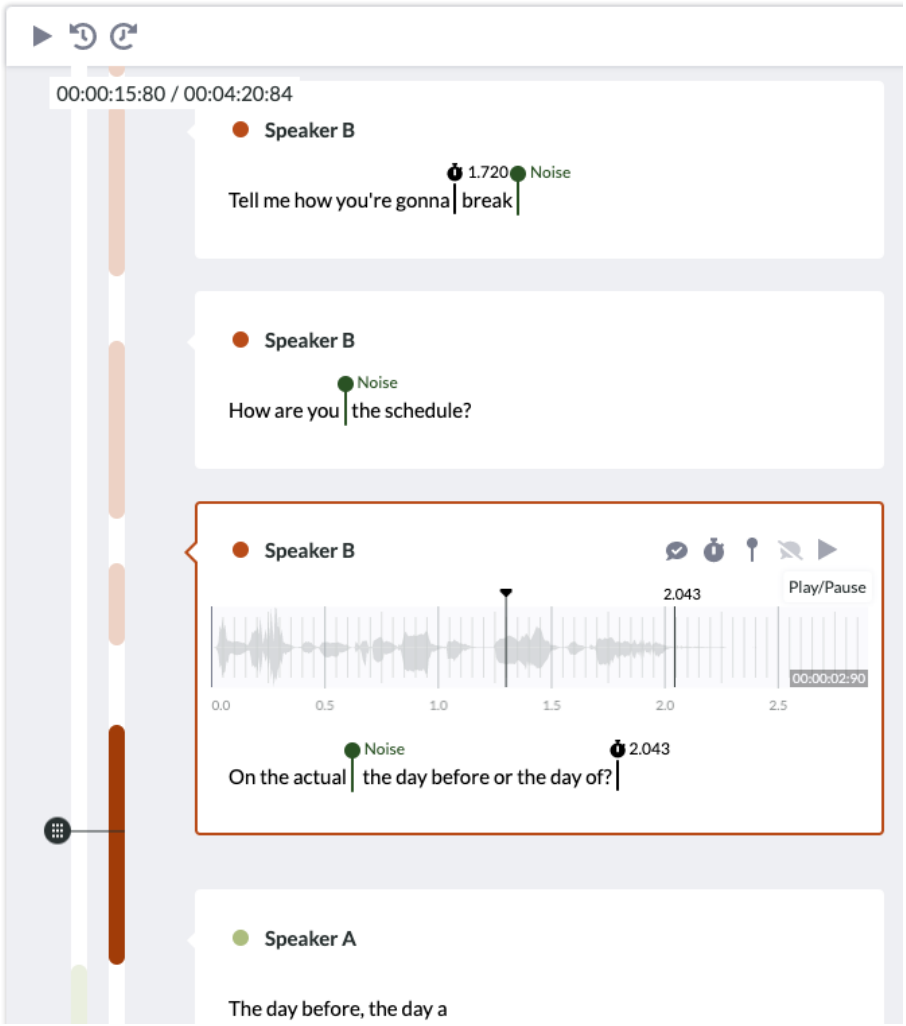



| | |
|---|---|
| "baseUrl": "https://sample-folder.com/", | folder containing point cloud source data |
| "frames": [ | list of frames in a scene |
| "items" : [ | array of objects in the specified frame id |
| "id": "13f222fd-065c-...", | unique identifier (uuid) of object |
| "category": "Car", | object class |
| "number": 8, | object instance number |
| "position": { | x,y,z position of center of cuboid |
| "x": 66.49787120373375, | |
| "y": -37.28758690422451, | |
| "z": -4.426572264322248 | |
| "rotation": { | rotation of cuboid in radians |
| "x": 0, | |
| "y": 0, | |
| "z": -1.5804235113355598 | |
| "dimension": { | full height, width, depth of a cuboid in meters |
| "x": 1.86, | |
| "y": 4.43, | |
| "z": 1.86 | |
| "locked": null, | not used, please ignore |
| "interpolated": true, | if the cuboid is a interpolated cuboid and never manually adjusted |
| "labels": {"size": "small", "occluded": "no"}, | Cuboid attribute information |
| "isEmpty": false, | optional parameter to indicate if the cuboid is an empty cuboid |
| "pointCount": 120 | count of points inside the cuboid |
| "isValid" : true, | To mark a frame as valid or not for labelling |
| "image" : "/image_00/image_.png", | file path to image |
| "items" : [ | array of 2D annotations |
| "id": "13f222fd-065c-...", | UUID of object, matches UUID of cuboid if it is the same object |
| "number": 1, | object instance |
| "category": "Car", | object class |
| "type": "RECT" | format of 2D annotation, a rectangle in this example |
| "position": { | top left corner of box |
| "x": 7.678934984761854, | |
| "y": 151.025760731091 | |
| "dimension": { | full width and height of the rectangle in pixels |
| "x": 311.2317472514253, | |
| "y": 184.7346955567628 | |
| "labels": {"size": "small", "occluded": "no"}, | 2D annotation attribute information |
| "isManual": true | the same meaning as the interpolated for 3D cuboid |



appen

# Speech training data example





```
{
  "id": "5",
  "layerId": "88620127-e109-4c6c-85a1-e92e39f0ff98",
  "startTime": 9.85503125,
  "endTime": 12.013343749999999,
  "ontologyName": "Speaker B",
  "metadata": {
    "transcription": {
      "text": "Tell me how you're gonna <11.575/> break <noise/> ",
      "annotatedBy": "human"
    },
    "original_text": "Tell me how you're going, they're great."
  },
  "nothingToTranscribe": false
},
{
  "id": "6",
  "layerId": "88620127-e109-4c6c-85a1-e92e39f0ff98",
  "startTime": 12.55671875,
  "endTime": 13.54221875,
  "ontologyName": "Speaker B",
  "metadata": {
    "transcription": {
      "text": "How are you <noise/> the schedule?",
      "annotatedBy": "human"
    },
    "original_text": "How are you the schedule?"
  },
  "nothingToTranscribe": false
},
{
  "id": "7",
  "layerId": "88620127-e109-4c6c-85a1-e92e39f0ff98",
  "startTime": 14.50578125,
  "endTime": 17.404906249999996,
  "ontologyName": "Speaker B",
  "metadata": {
    "transcription": {
      "text": "On the actual <noise/>  the day before or the day of? <16.549/> ",
      "annotatedBy": "human"
    },
    "original_text": "And after the day before the day of dutiful with me."
  },
  "nothingToTranscribe": false
},
```




appen

Training data quality is important

Low quality training data leads to poor performing models

Poor quality data is not always obvious

# Three types of problems we see with training data

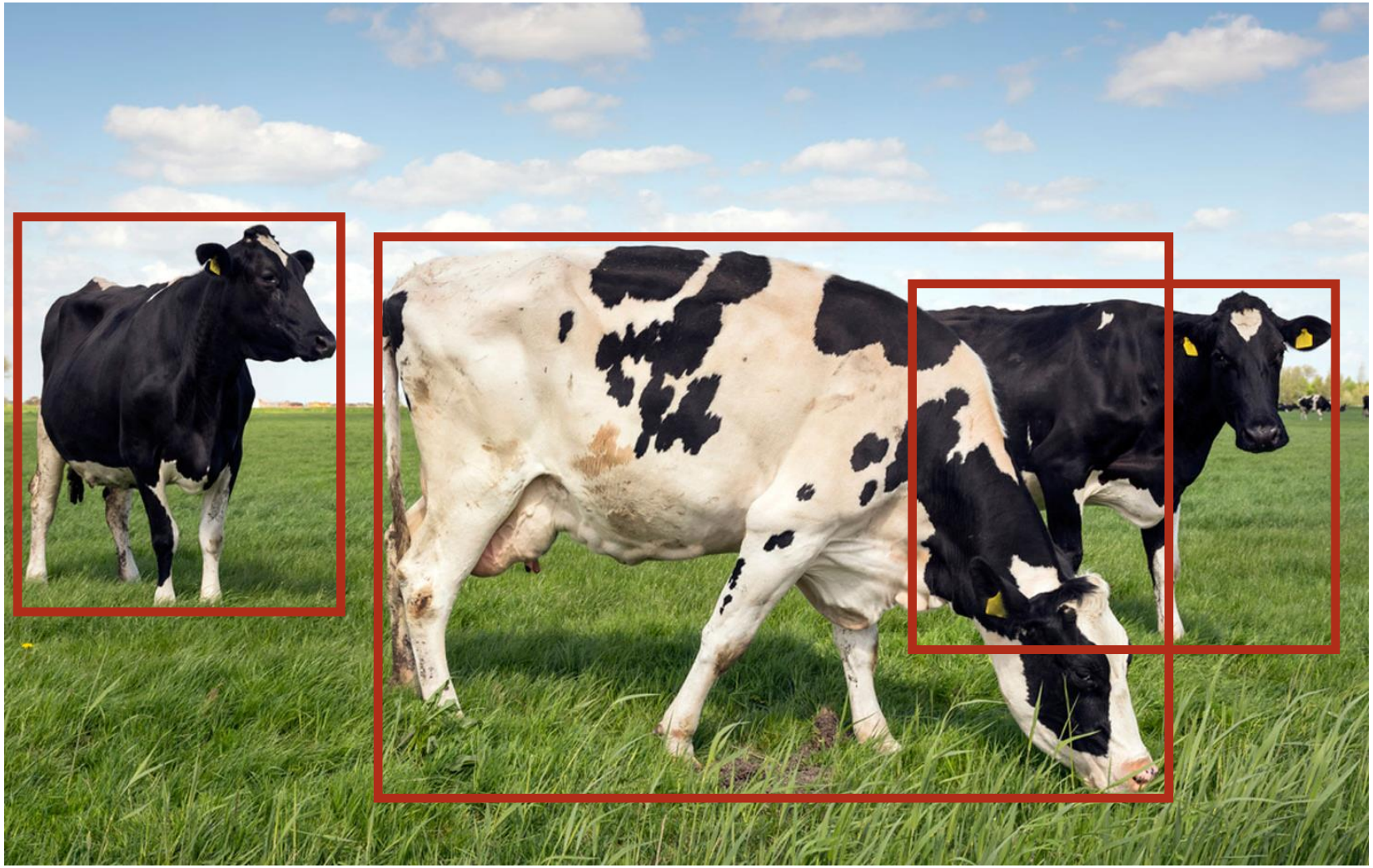
1. Labelling errors

2. Unbalanced training data

3. Bias in labelling process

# Task instructions: Draw a box around the cows

# Task instructions: Draw a box around the cows



Intended output – a tight bounding box around each of the visible portions of the cows

# 1. Labelling errors – missing labels

Cow on the right was not labelled

# 1. Labelling errors – incorrect fit

Bounding box is not “tight” enough around the cows

# 1. Labelling errors – misinterpreting instructions

One large box instead of three

# 1. Labelling errors – handling occlusion

Box is placed around the expected size of the occluded cow

# 2. Unbalanced training data – class imbalance

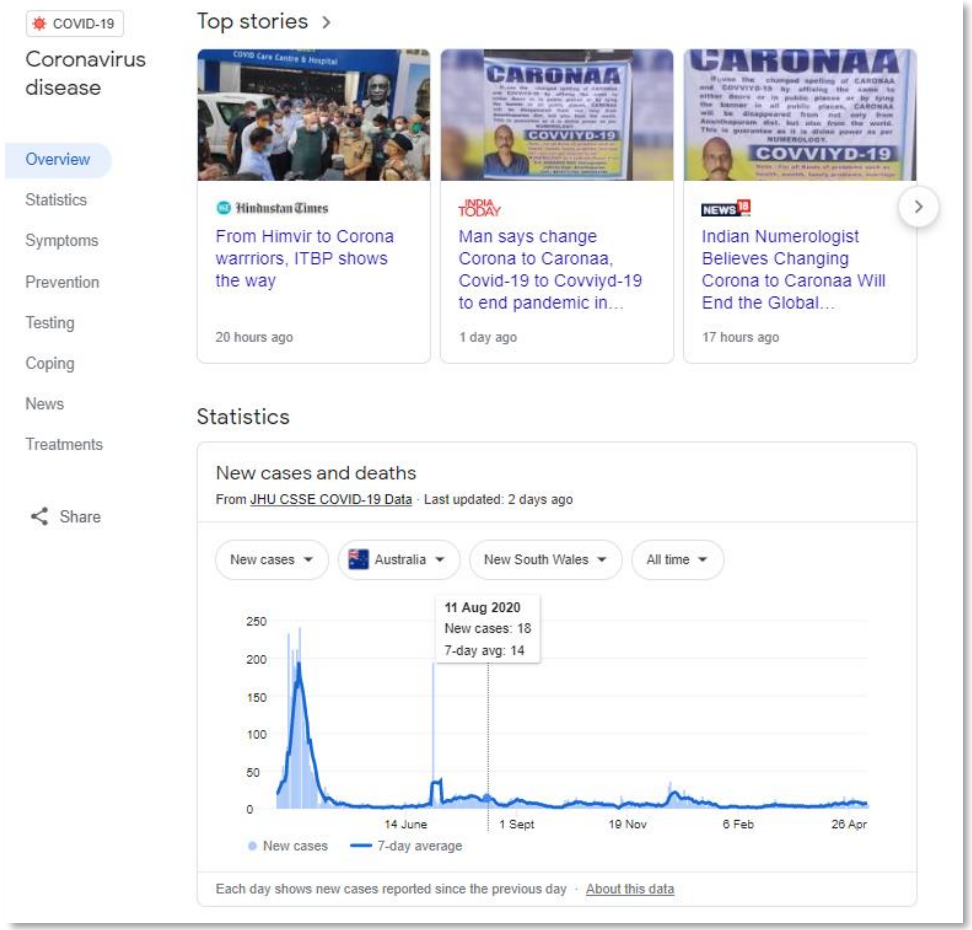







# 2. Unbalanced training data – data recency

## “Corona” search results – May 2021




## “Corona” search results – April 2019

https://www.liquorland.com.au › beer › corona-can-35...

**Beer & Cider - Corona Can 355Ml 12Pack | Liquorland**

27 Apr 2019 — Browse our selection of Beer & Cider - **Corona** Can 355Ml 12Pack available online and in-stores across Australia. Click through to find out more.

https://www.amazon.com › Corona-Posters-Prints-Wall-...

**Corona. - Posters & Prints / Wall Art: Home ... - Amazon.com**

27 Apr 2019 — purl zither Tin Sign Vintage Chic Art Decoration Poster **Corona** Beer Pinup Girl Sexy Bikini Girl for Store Bar Home Cafe Farm Garage or ...

https://www.amazon.com › Corona-Mens-Bottle-Label-...

**Corona Mens Bottle Label Boardshorts | Amazon.com**

26 Apr 2019 — Buy **Corona** Mens Bottle Label Boardshorts and other Board Shorts at Amazon.com. Our wide selection is elegible for free shipping and free returns.

https://www.medicrossclinic**corona**.com › corona-ca › r...

**Rashes and Burns Treatment | Urgent Care in Corona, CA**

26 Apr 2019 — Rashes and Burns Treatment**Corona**, CA. There are numerous reasons an individual may have developed a rash or suffered a burn, and as such, it can often be ...

https://www.medicrossclinic**corona**.com › corona-ca › i...

**Insect Bites and Stings Treatments | Urgent Care in Corona, CA**

26 Apr 2019 — Insect Bites and Stings Treatments**Corona**, CA. Insect bites and stings can have a wide variety of effects on an individual, ranging from an annoying nuisance to ...



appen

# 3. Bias in labelling process

Black Pudding

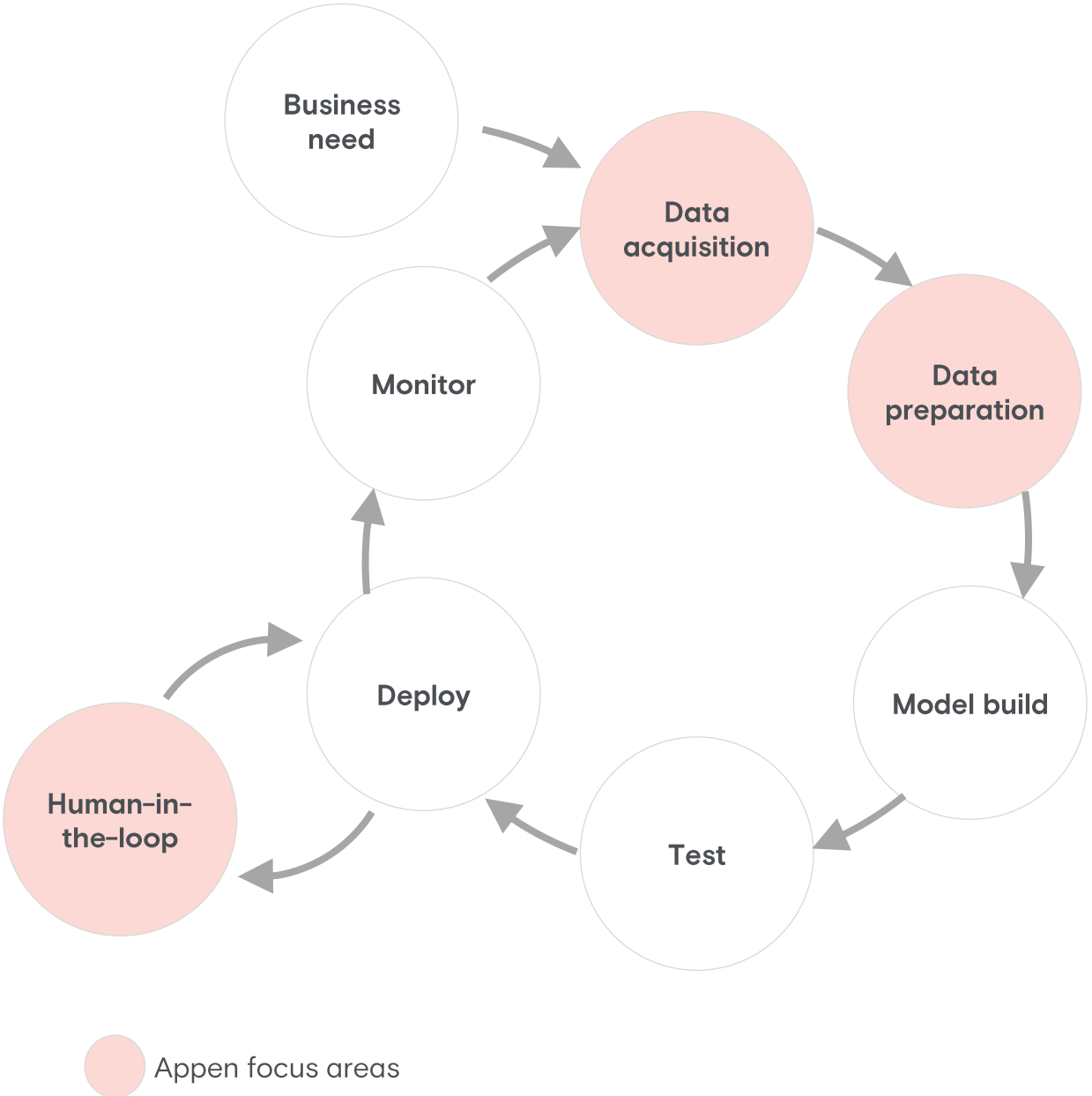
Hagelslag
(Sprinkles on toast)

Vegemite

**Some datasets require specific knowledge / context for accurate labelling**

AI model = Model instructions + Training Data

Good AI model = Model instructions + High Quality Training Data




# Our role in the AI lifecycle is to deliver high-quality training data



appen

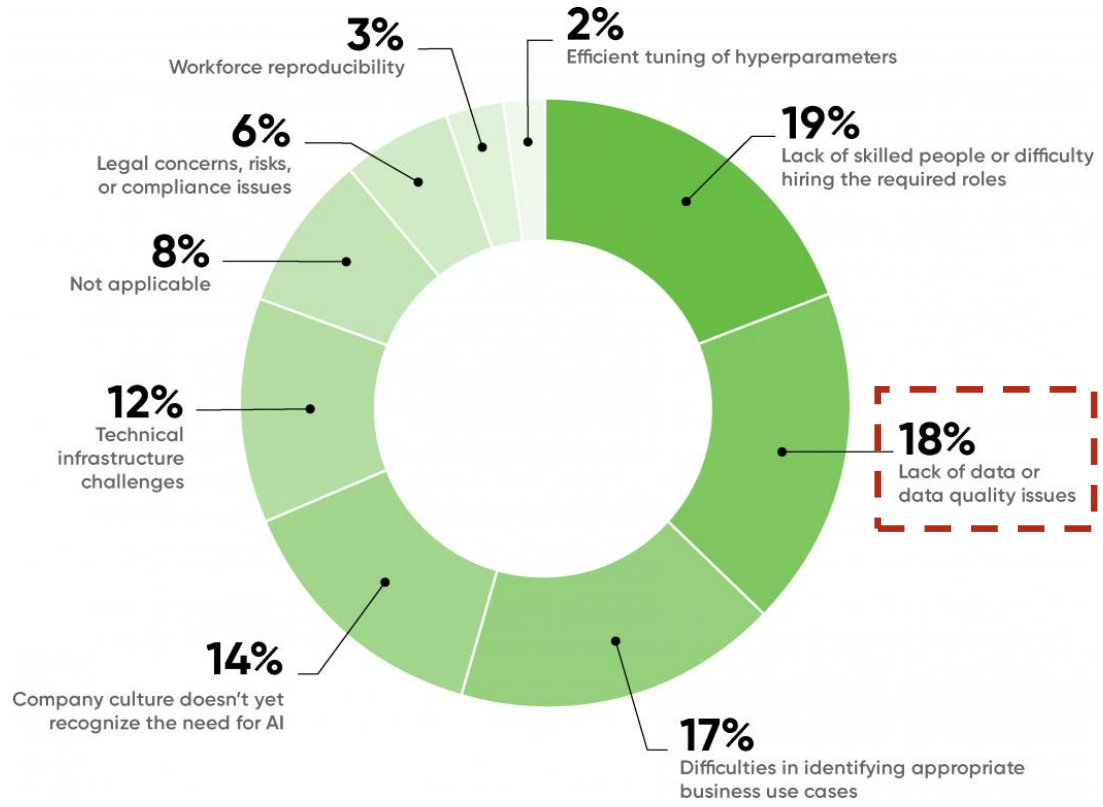
# The AI training data market is rapidly evolving

# Five market trends

1. Data remains a major roadblock for AI
2. AI use cases becoming narrower
3. Shift from model-centric to data-centric AI
4. Emerging need for training data operations
5. AI assisted annotations increasing

# 1 Data remains a major obstacle for AI

In a recent survey, 18% of respondents see lack of data or data quality as the major bottleneck to AI




O'Reilly: AI Adoption in the Enterprise 2021. Survey n = 3,574

Many AI practitioners spend most of their time collecting and preparing data

*"... Airbnb have discovered that nearly **70% of the time a data scientist spends** developing machine learning models – is not doing the actual modelling, but **collecting data and feature engineering**."*

https://read.hyperight.com/ml-powered-accommodation-hunt-airbnb-approach/



appen

# 2 AI is becoming narrower

***Real-world examples we have supported***

## Biz-speak

- AI model built to suggest improvements to common “biz-speak”
- Challenge: Biz-speak is highly nuanced
- Our task: Suggest alternative phrases for common biz-speak terms

## Body movements

- AI models built to automate personal training
- Challenge: Movement profile changes with age
- Our task: capture and annotate video of seniors doing somersaults

## Long-tail languages

- Covid information needs to be shared globally
- Challenge: Translation tech doesn’t support all languages
- Our task: Data collection/annotation for languages including Dari, Dinka, Hausa, Luganda, Pashto, and Zulu

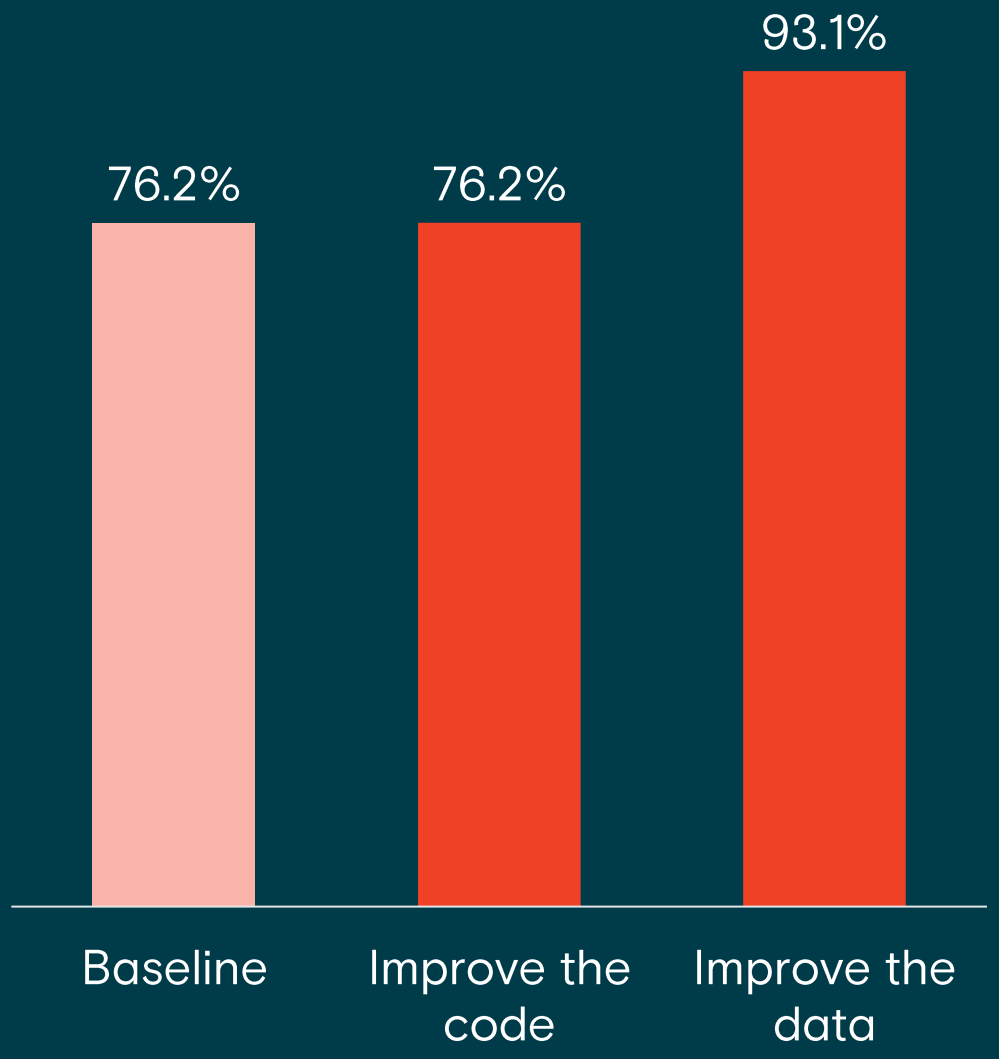
Good AI model = Model instructions + High Quality Training Data

# Is it better to improve code or training data?



appen

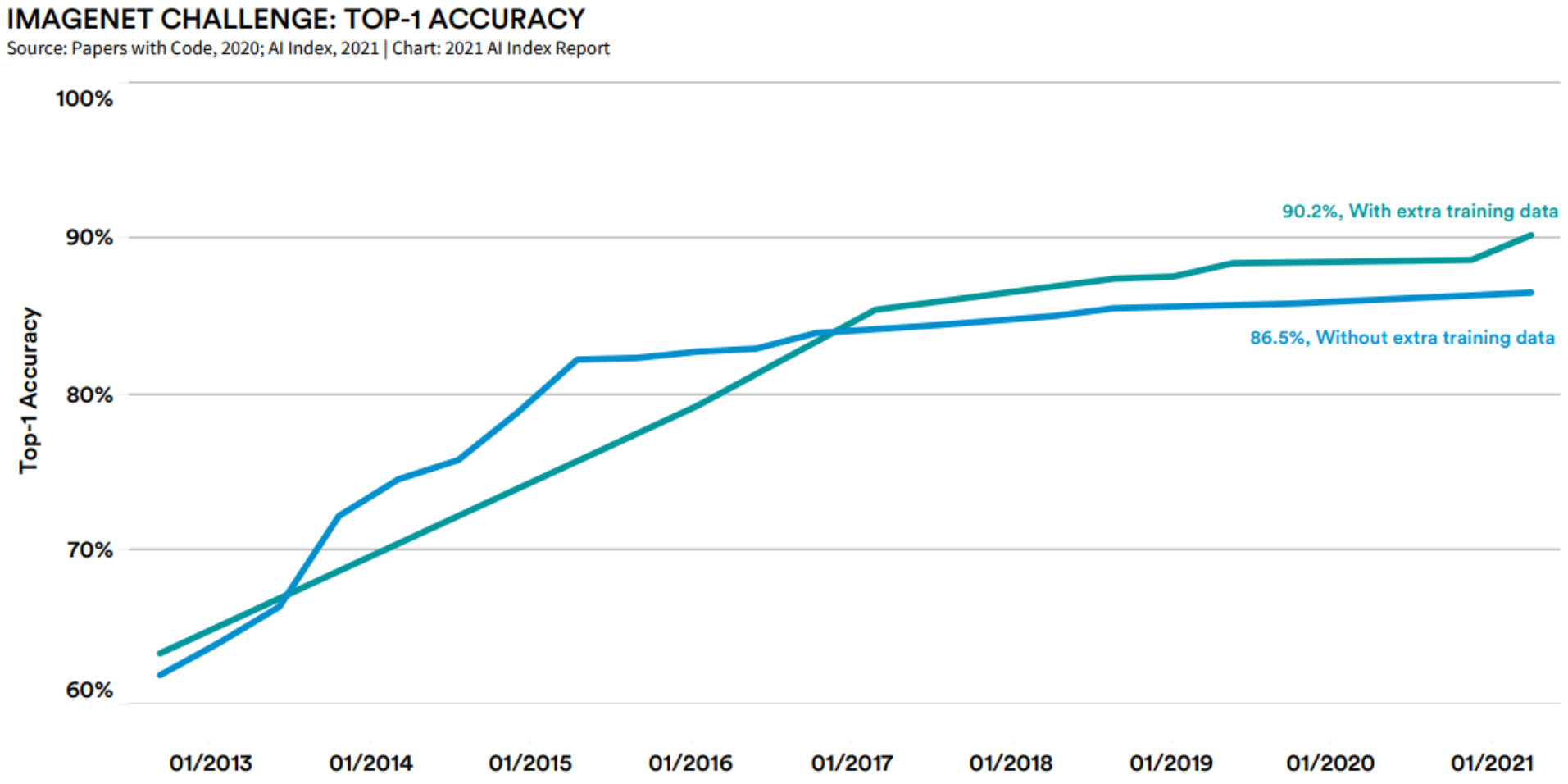
## Steel defect detection model




Source: Landing.ai

appen

# ③ ImageNet performance has capped without extra training data




Source: Artificial Intelligence Index Report 2021

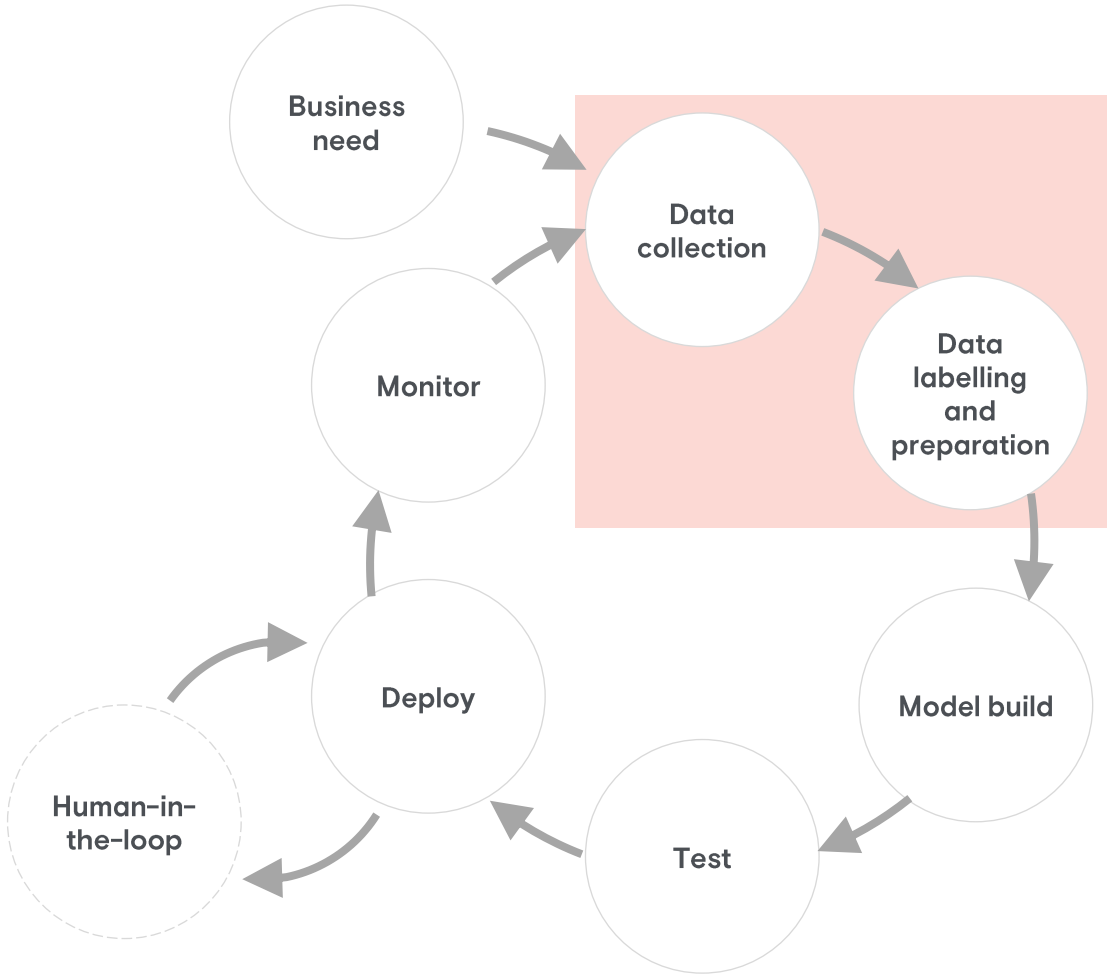


appen

# ③ There is a shift from “model-centric” to “data-centric” AI

| Model-centric | Data-centric |
| --- | --- |
| **Use available data**<br><br>Develop a model that compensates for noise in the data | **Improve volume and/or quality of the data**<br><br>Try different models to achieve the best performance |
| **Hold the data fixed and interactively improve the model** | **Hold the model fixed and interactively improve the data** |

# 4 Need for training data management emerging




- Data collection
- Data ingestion
- Data exploration
- Data labelling
- Data validation
- Data preparation

- Version control
- Training data security
- Access controls
- Data pipeline monitoring
- Collaboration
- Meta-data insights



appen

# 5 Automation critical to the labelling process

## Pre-labelling

- AI performs an initial ‘best guess’ of the annotation
- Crowd workers check and correct the pre-label (if required)
- Significantly reduces annotation time

## Speed labelling

- AI models that assist crowd workers by automating slow and/or manual tasks
- Works similar to an auto-complete function
- Significantly reduces annotation time

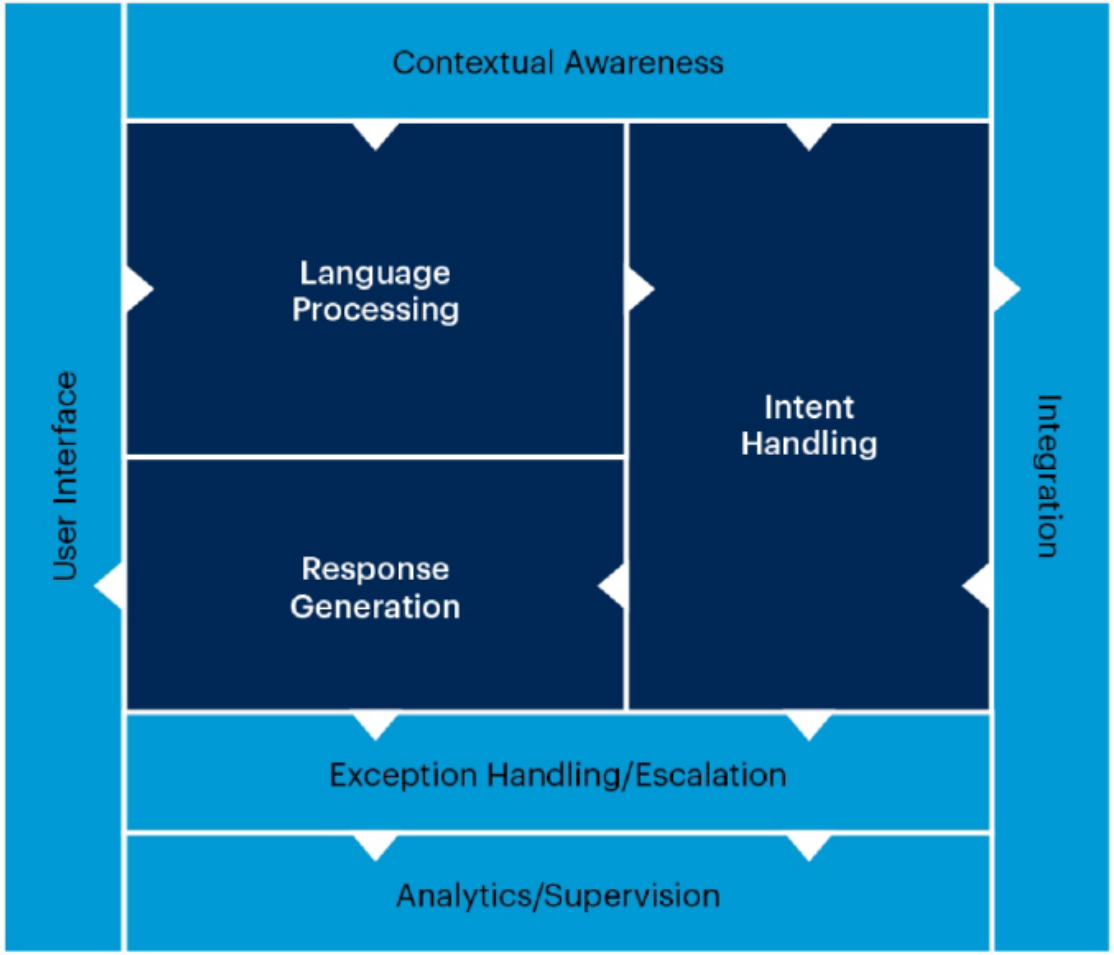
## Smart Validators

- AI models that verify crowd output before they are submitted
- Crowd get notified if input is not within expected thresholds
- Improves data quality and overall worker performance

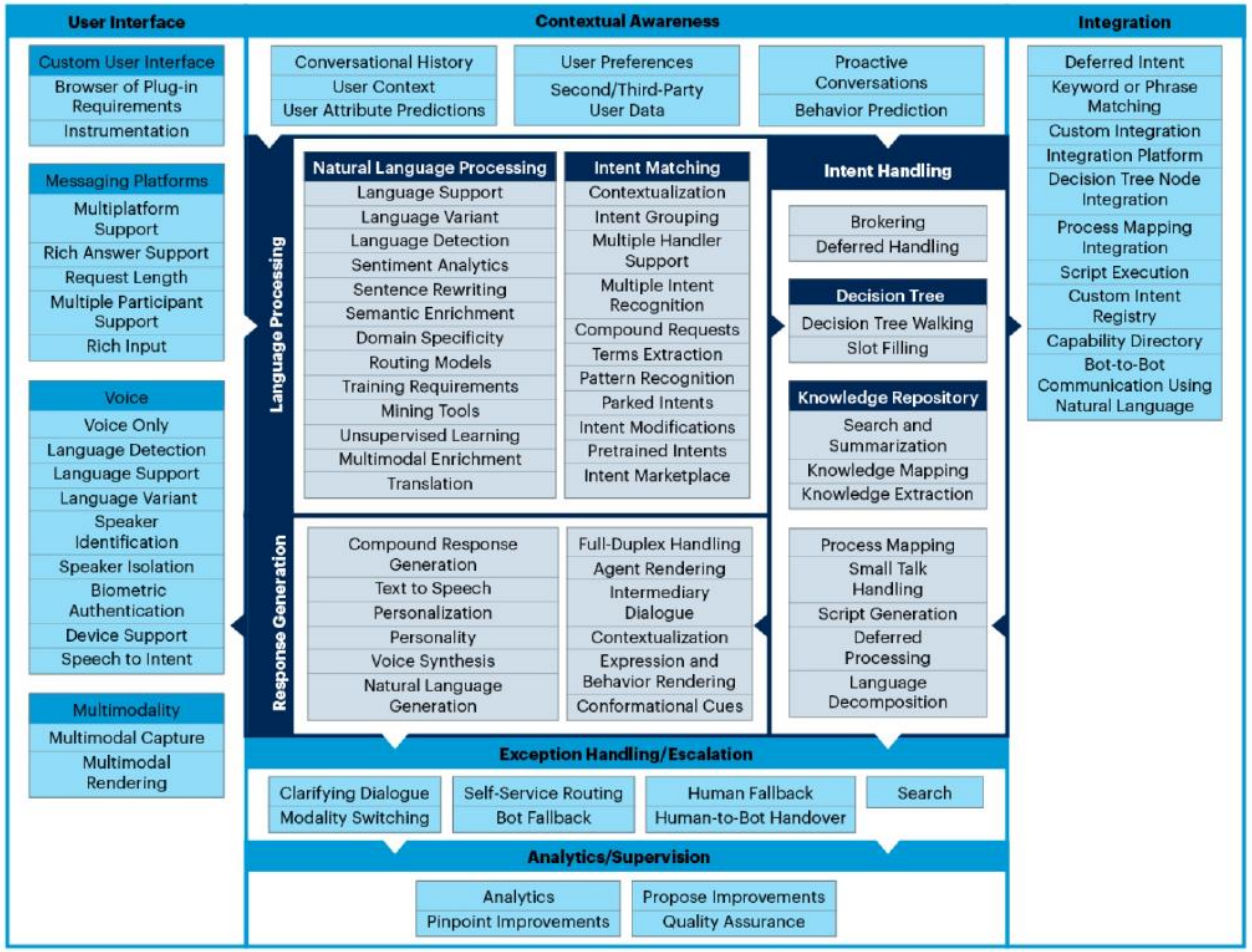
# AI models rely on a variety of training data techniques

# AI enabled applications typically involve many different models

Voice interface architecture



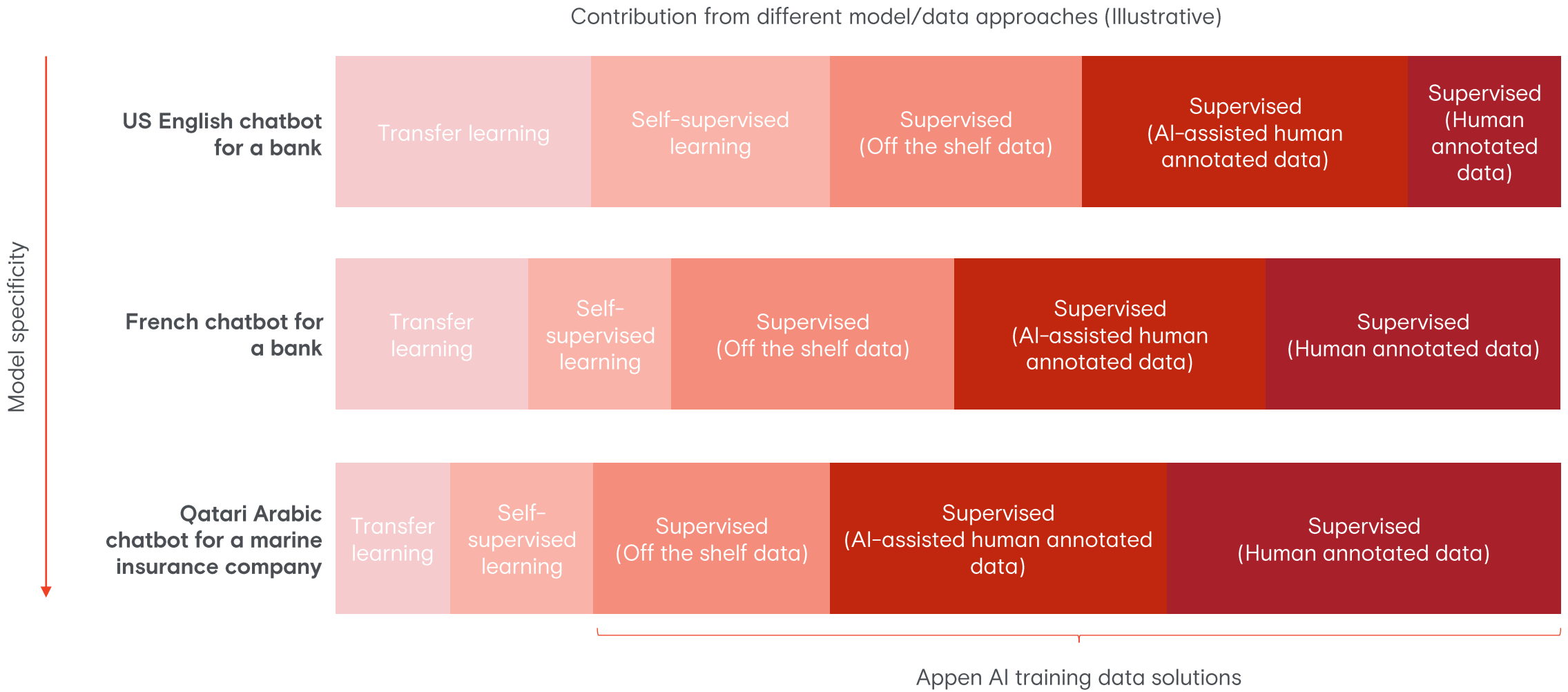

Detailed steps – many models and interfaces






appen

# Real-world models typically use a combination of data-sources and techniques

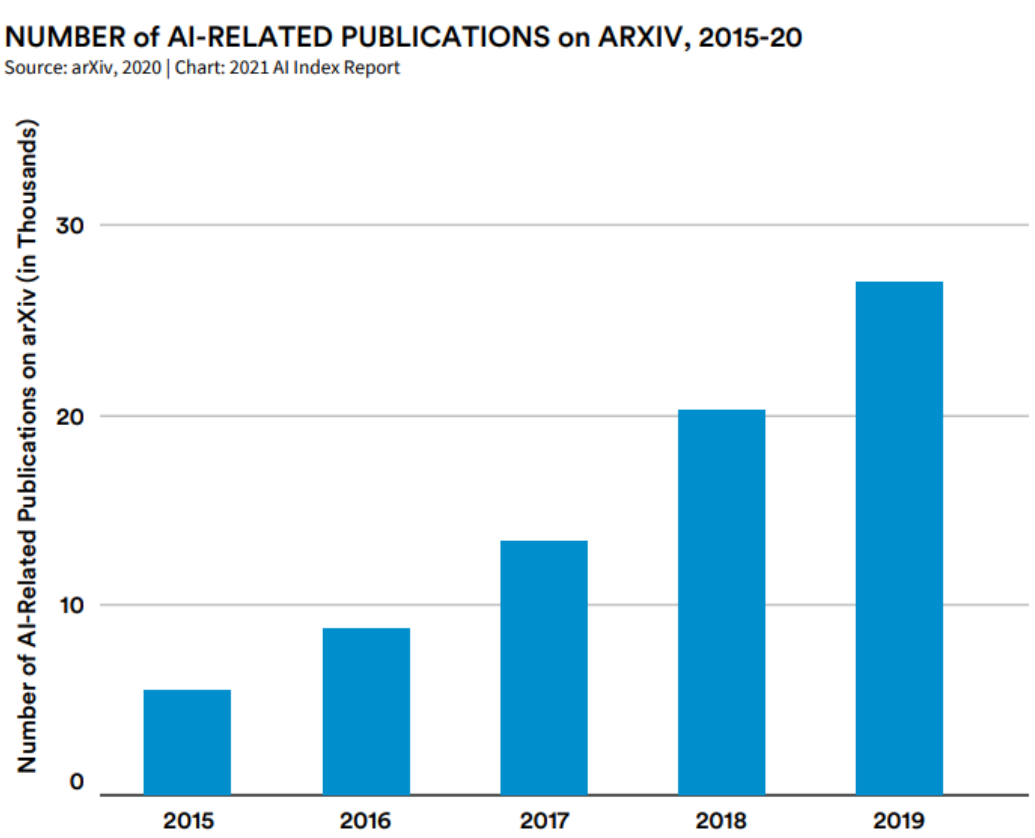





appen

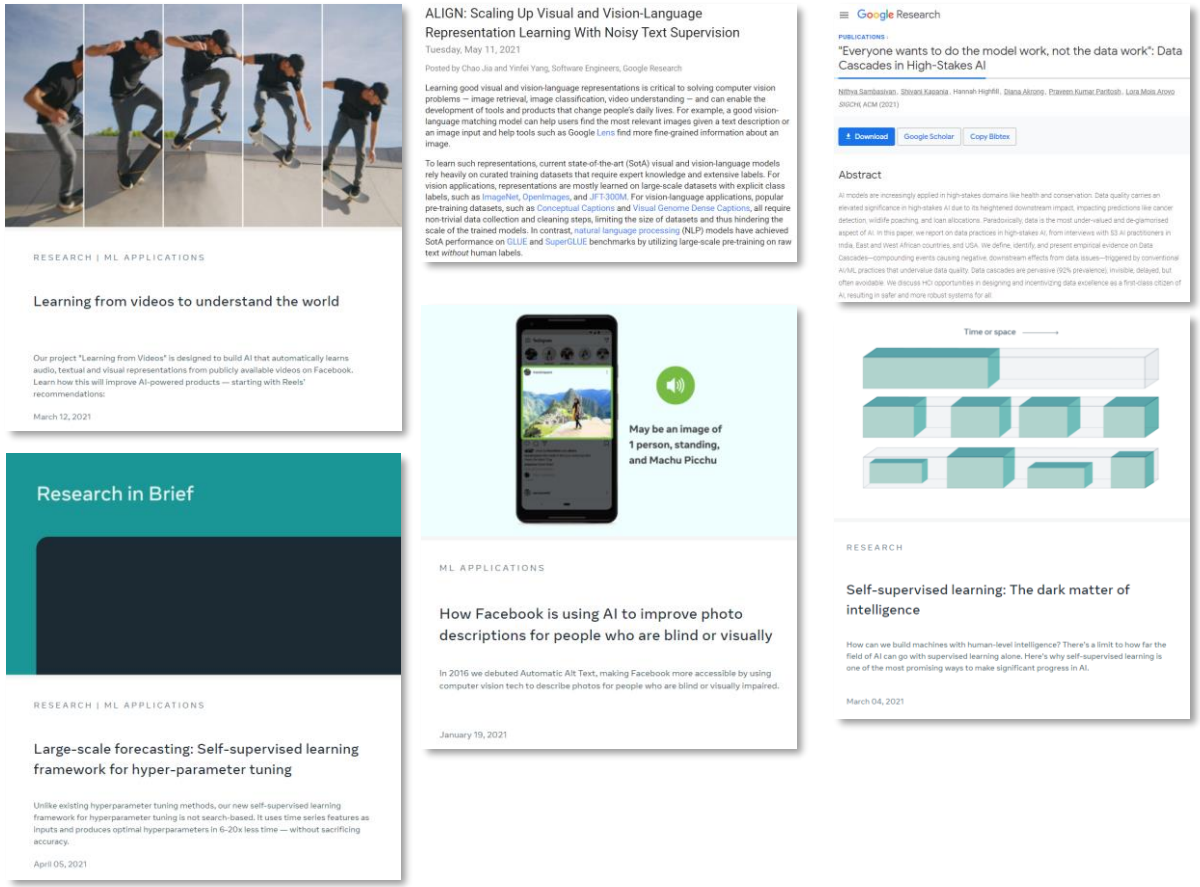
# New AI approaches continue to emerge

## ~30k AI related publications on ARXIV in 2019




Source: Artificial Intelligence Index Report 2021

## Many research teams exploring new areas

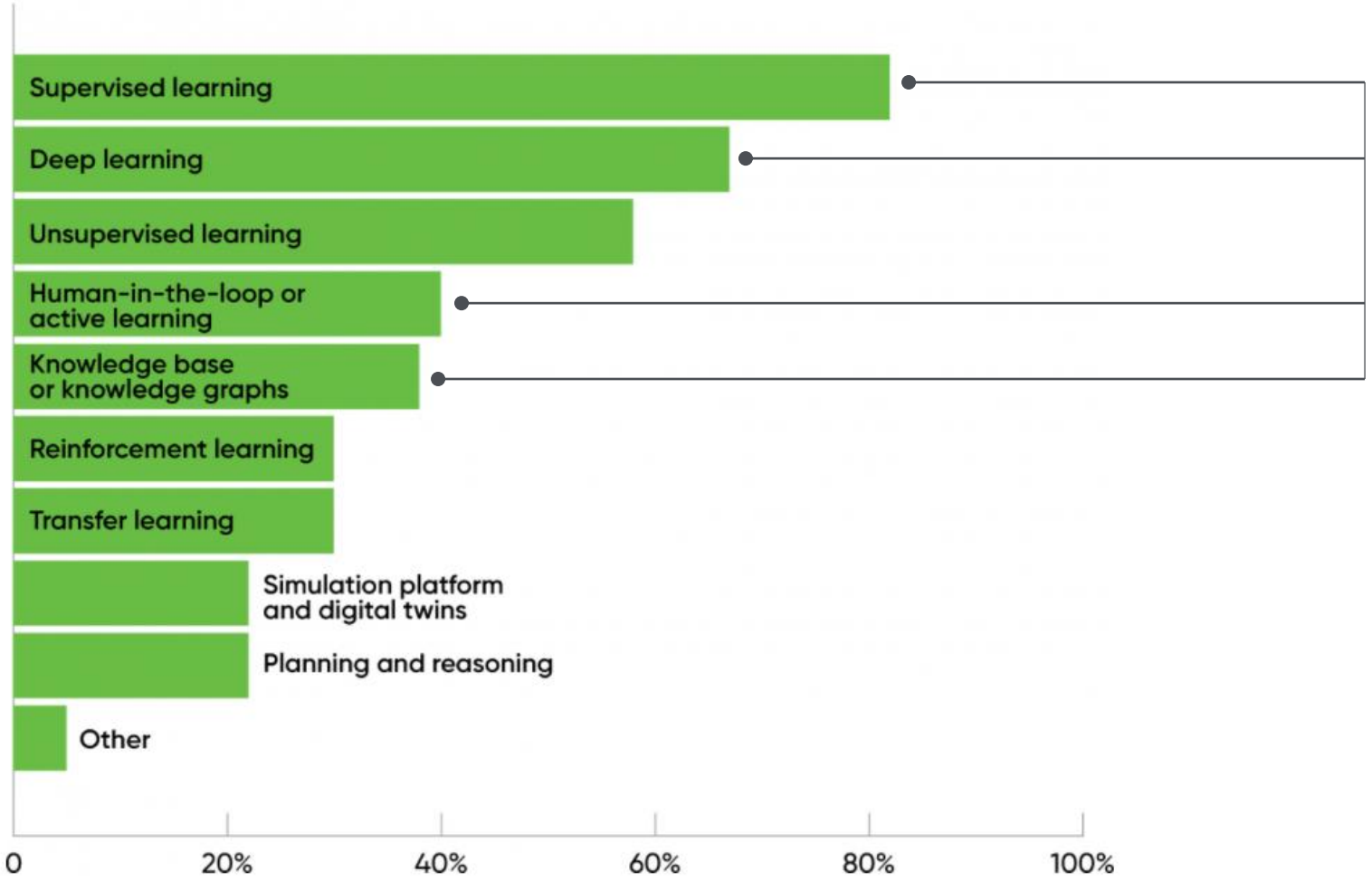





appen

# Popular AI techniques rely on human involvement

## AI technologies used in mature practices




Techniques that typically require some level of human annotation and/or preparation

Source: O'Reilly: AI Adoption in the Enterprise 2021. Survey n = 3,574



appen

# Appen is accelerating its transformation into an AI powered provider of AI data and solutions

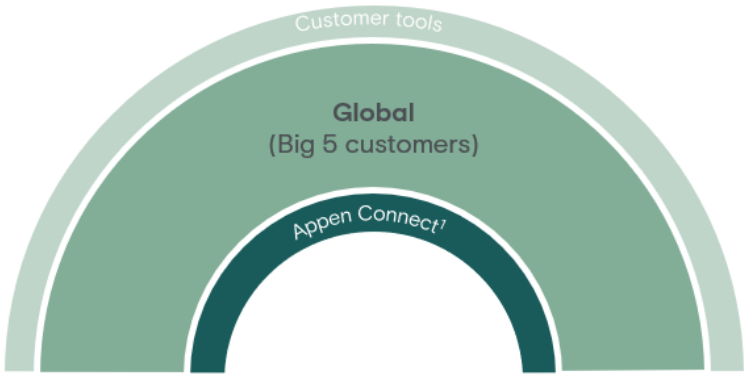
# Appen continues to evolve

| | From | To |
|---|---|---|
| **Data type** | Language data | AI data |
| **Delivery model** | Service led | Product led |
| **Revenue** | Project based | Committed |
| **Customer** | Concentration | Diversification |
| **Org structure** | Functional alignment | Customer alignment |
| **Reporting** | Data modality, AUD | Strategy led, USD |

# Appen's evolution

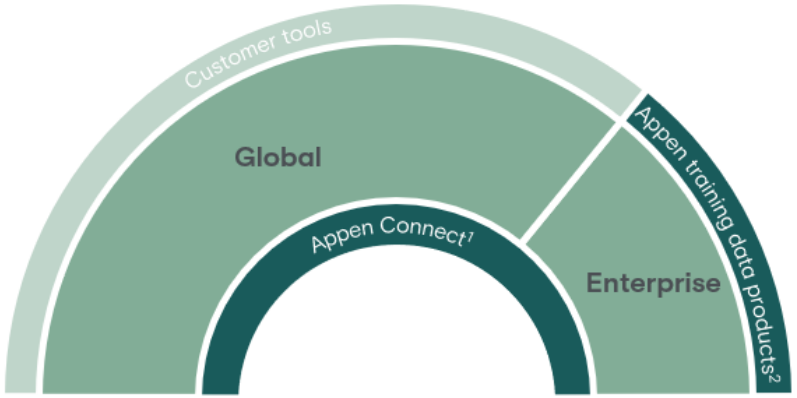
## Phase 1: Dominate the AI data annotation services market

Acquired Butler Hill 2010
Acquired Leapforce 2017



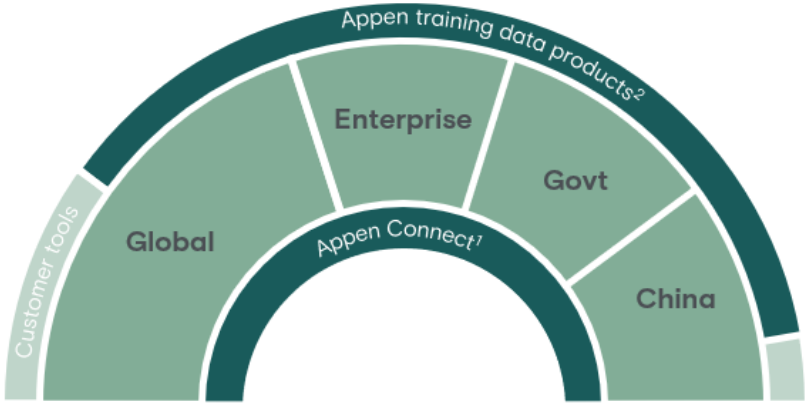

## Phase 2: Build leading data annotation products

Acquired Figure Eight 2019
Ongoing product investment
New Enterprise customer base




## Phase 3: Product-led expansion

Increased sales and marketing China and Govt BUs established
Global customers on Appen products






appen

# We combine our industry leading technology with a crowd of over 1m and deep internal expertise




appen

# The Appen Product Suite

New products

## Appen Connect

Match our global crowd workforce to annotation tasks

## Appen Data Annotation Platform

Collect and annotate training data

## Appen Intelligence

Powers Appen products with proprietary ML models

## Appen In-Platform Audit

Organize and analyze training data to identify quality, distribution & bias

## Appen Mobile

Engages, enables and expands crowd



appen

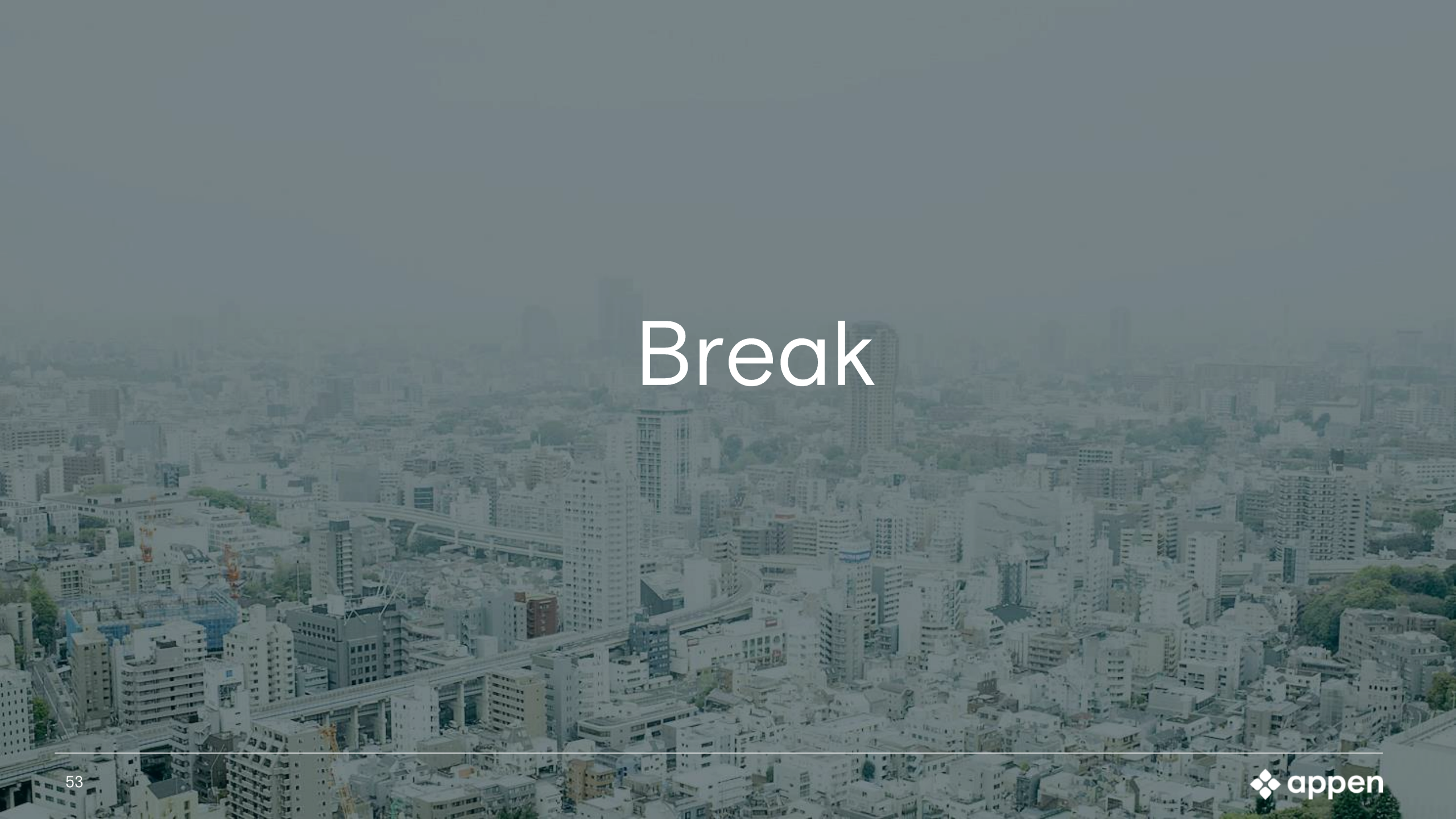
# Value created from our products

New products

| Appen Connect | Appen Data Annotation Platform | Appen Intelligence | Appen In-Platform Audit | Appen Mobile |
|---|---|---|---|---|
| AI-augmented data collection and labelling | AI-enabled crowd management | Embedded expertise | In-built crowd management features | Native integration with our crowd |
| **Improves speed, quality, scale and unit economics** | **Increases internal productivity and crowd experience** | **Delivers high quality annotation** | **Reduces risk for customers** | **Creates competitive differentiation** |



appen






appen

# Appen Product Suite

Enables the creation of high-quality training data, faster and at scale, using innovative technology

# The Appen Product Suite

New products

## Appen Connect

Match our global crowd workforce to annotation tasks

- Assigns tasks to our global workforce of 1m+ working in 170+ countries and 235 languages
- Pays 70k+ crowd workers per month
- Quality management and fraud detection

## Appen Data Annotation Platform

Collect and annotate training data

- Full suite of data collection and annotation tools
- AI-enabled/automation capabilities
- Ability to support complex, multi-step annotations
- Wide array of data security options

## Appen Intelligence

Empower Appen products with proprietary ML models

- Over 20 proprietary models to improve crowd productivity and data labelling quality
- Used extensively to automate internal crowd management tasks

## Appen In-Platform Audit

Organize and analyze training data to identify quality, distribution & bias

- Enables our customers to navigate and troubleshoot training data

## Appen Mobile

Engages, enables and expands crowd

- Simple and intuitive user experience
- Crowd can interact with Appen anywhere and anytime
- Reaches people who don't have access to a desktop or prefer mobile
- Enables location-based data collection



appen

# Appen Connect

**Appen Connect**

Match our global crowd workforce to annotation tasks



Appen Data Annotation Platform

Collect and annotate training data



Appen Intelligence

Powers Appen products with proprietary ML models



Appen In-Platform Audit

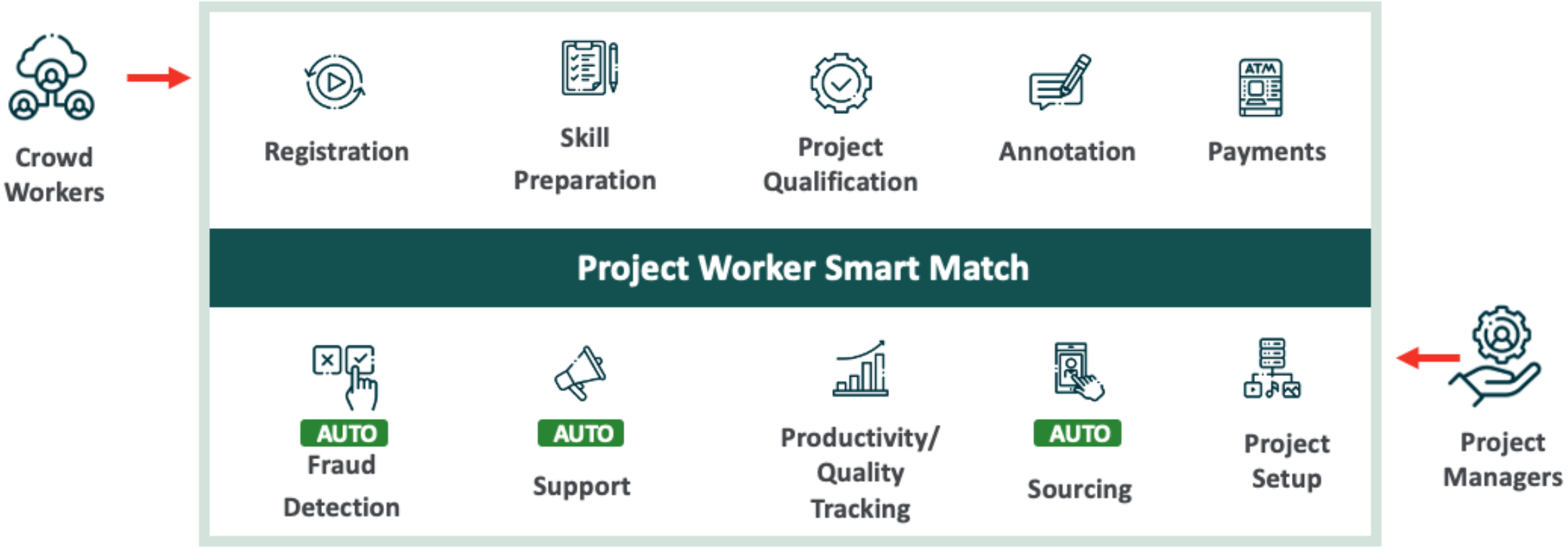
Organize and analyze training data to identify quality, distribution & bias



Appen Mobile

Engages, enables and expands crowd

# Appen Connect – an intelligent marketplace to match our crowd with customer tasks

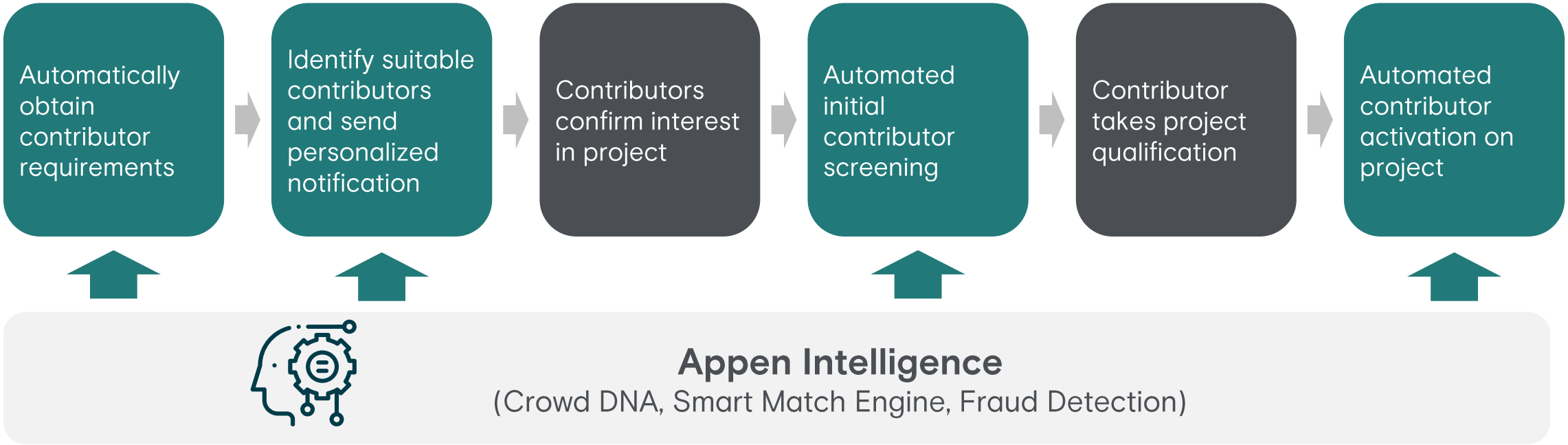





appen

# Highly automated process to match contributors to projects

## Process to match crowd contributors to projects

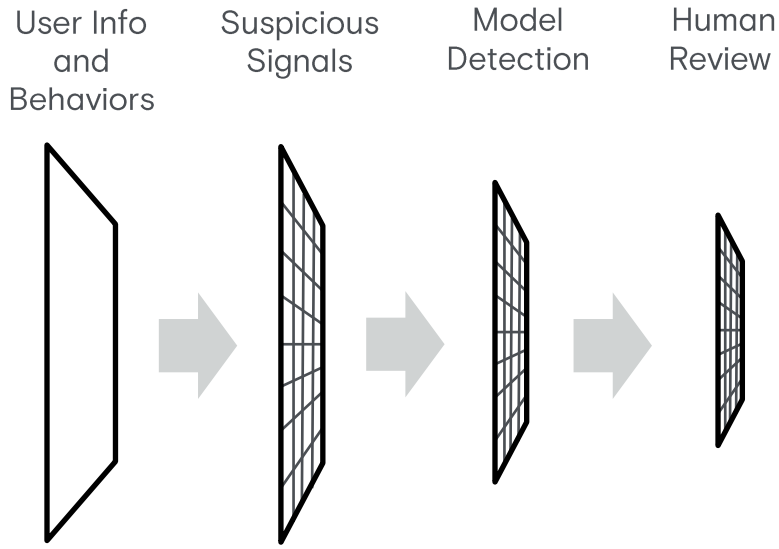


# Automatic identification of fraudulent contributors

## Example

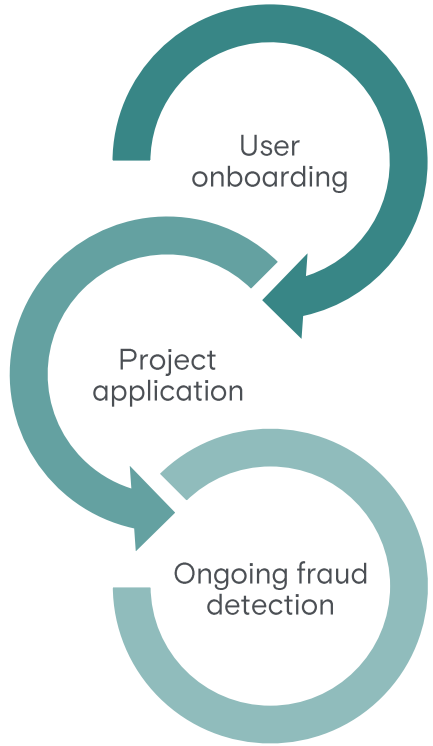
| Account | Country | IP | Active Time |
|---|---|---|---|
| 10029xx | Brazil | 10.21.368.xx | 8am–6pm |
| 40483xx | US | 10.21.368.xx | 3am–5am |

## Methodology




- Processes 1 million+ profiles every day
- Leverages 200+ signals
- Results in ~95% precision

## Scenarios






appen

# Appen Connect
## Customer benefits and future direction

### Benefits for our customers

- Connects customers to our global crowd of contributors
- Reduces project overhead cost by automating internal effort
- Supports future growth with scalability enabled by AI

### Future direction

- Continued optimisation of user experience
- Further automation of project management effort

# Appen Data Annotation Platform



Appen Connect

Match our global crowd workforce to annotation tasks

**Appen Data Annotation Platform**

Collect and annotate training data



Appen Intelligence

Powers Appen products with proprietary ML models



Appen In-Platform Audit

Organize and analyze training data to identify quality, distribution & bias



Appen Mobile

Engages, enables and expands crowd



appen

# Our annotation tools support a wide variety of AI use cases

## Data Collection

- Telephony
- Studio recording
- In-car recording

## Content Relevance

- Search relevance
- Sentiment analysis
- Data categorization
- Data validation

## Audio & Speech

- Voice transcription
- Translation

## Text & Language

- Name entity extraction
- Text relationship

## Image & Video

- Image transcription
- Image annotation
- Pixel level semantic segmentation
- Video annotation

## 3D Point Cloud

- LiDAR annotation
- LiDAR semantic annotation

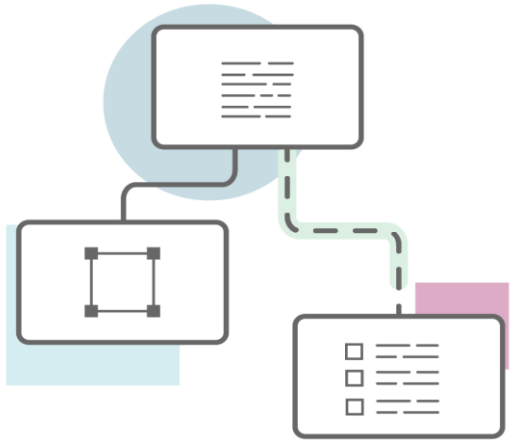
**Job Designer / CML (Custom Markup Language)**



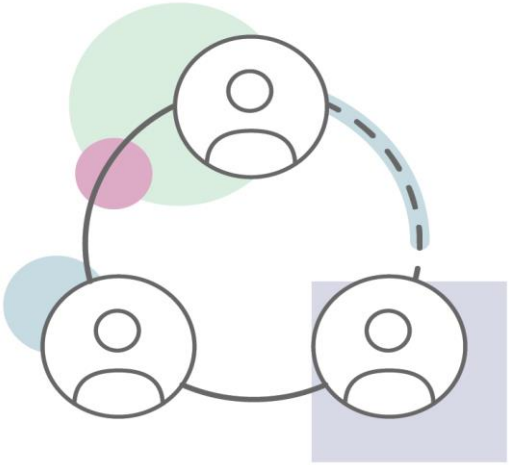
appen

# Appen Workflows enables complicated AI data use cases

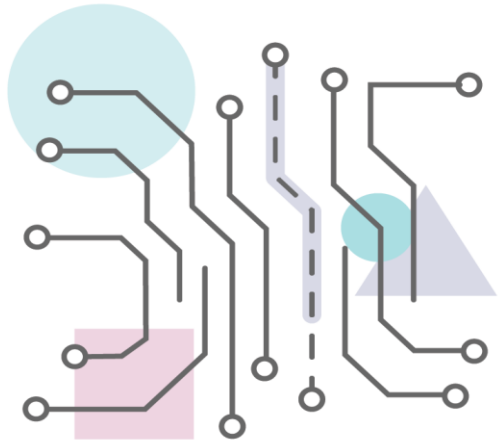
## Simple user interface to build and automate multi-step data annotation projects



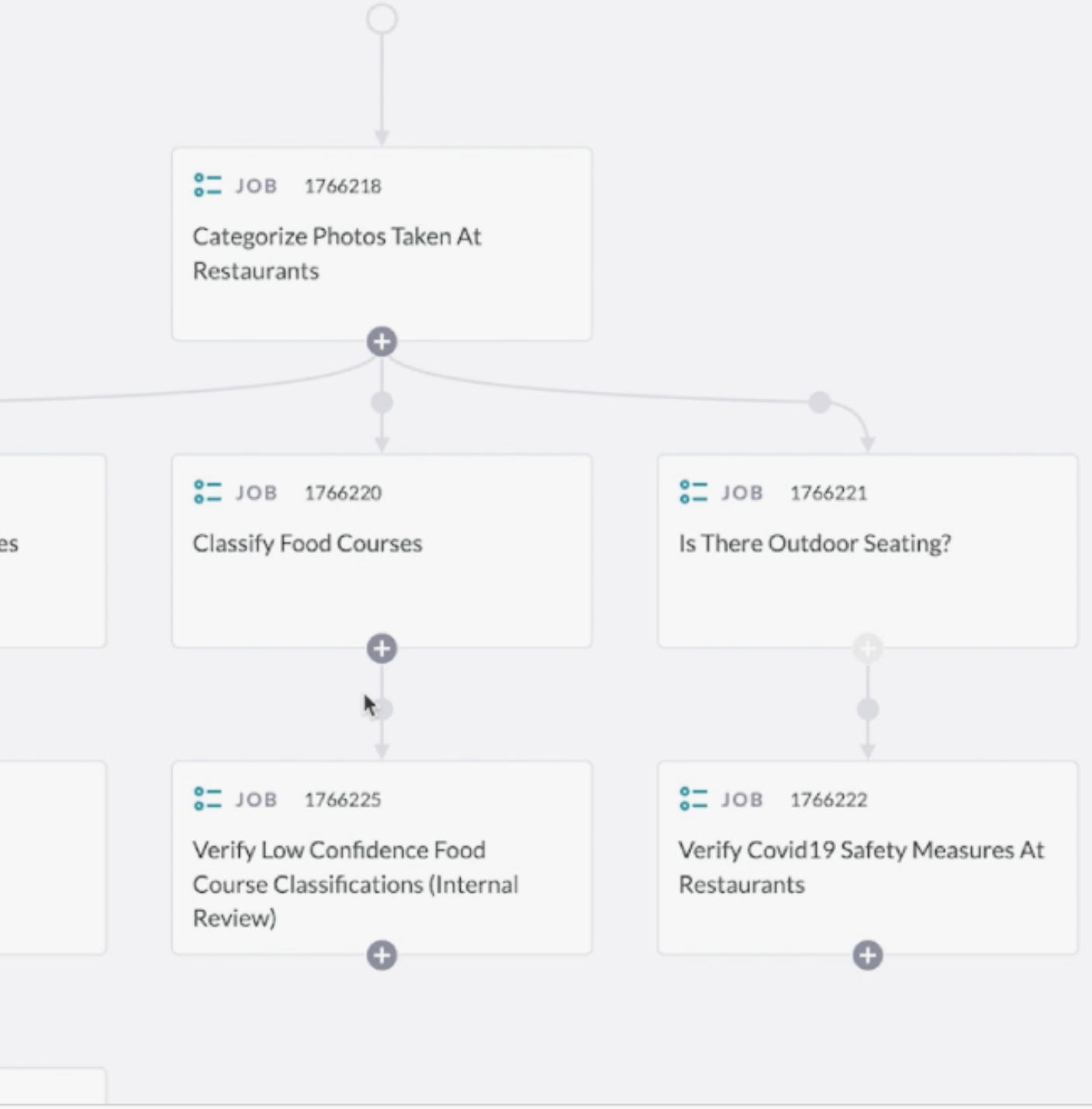
Simple user interface



Human & Machine Intelligence



Flexible settings




# Demo 1

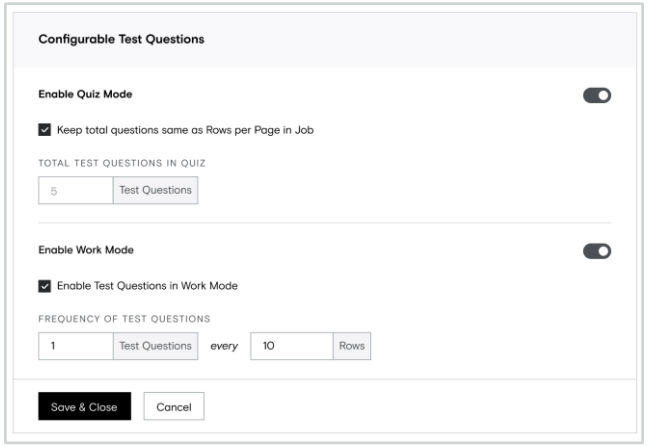
## Appen Workflows

appen

# Built-in quality controls

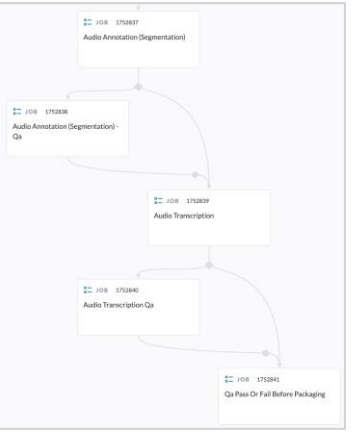
## Test Questions

Ground truth that tests contributors' understanding of a job both before and while they work



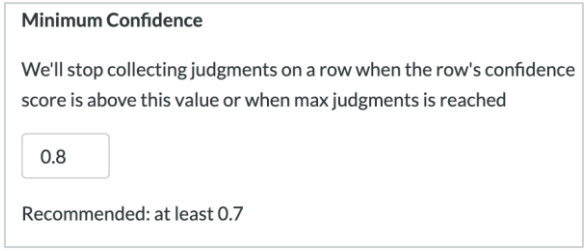

## Quality Assurance

Highly qualified contributors review and correct annotations within the workflow



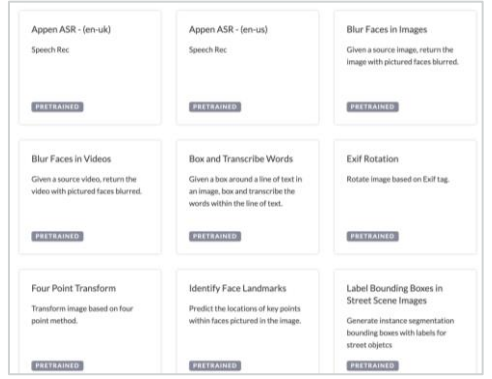

## Dynamic Judgments

Leverage majority vote, achieve high confidence annotations while minimizing the number of required judgments per unit




## ML Validation

Leverage machine learning model predictions to validate human results

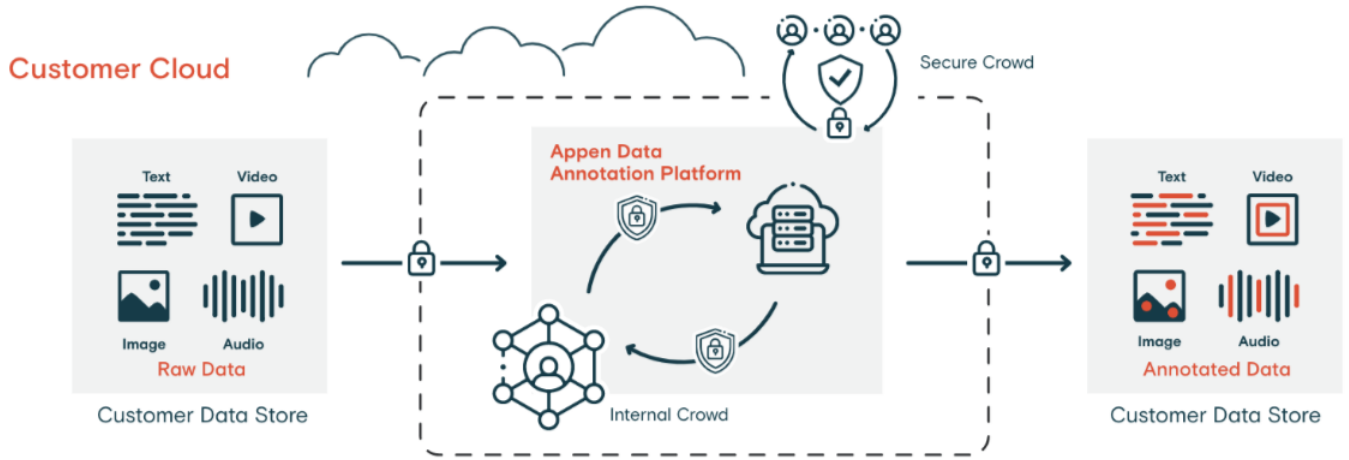


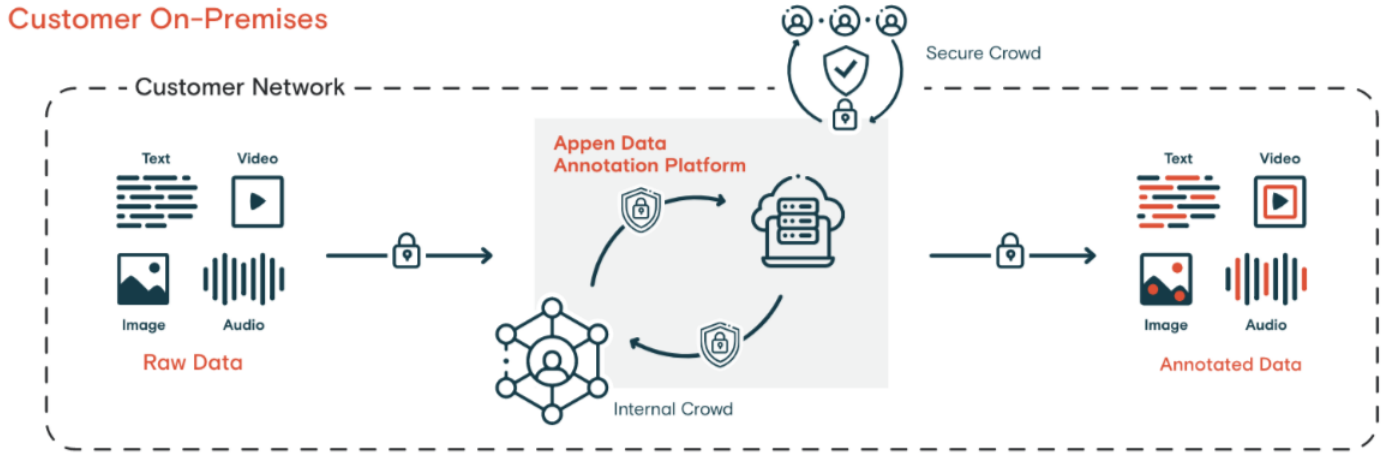



appen

# Industry leading security solutions

## Flexible Deployment Options




### Compliance & Certification

Comply with legal obligations related to data security and privacy to protect the health, safety and welfare of customers.

**Standards:** SOC 2, GDPR, HIPAA

### Secure Data Access

Customers can use secure data buckets (s3 and azure) to completely own and govern their own data. Appen Data Annotation Platform only renders the data when contributor works on the annotation.



appen

# Appen Data Annotation Platform
## Customer benefits and future direction

### Benefits for our customers

- Full suite of data annotation tools and flexibility to customise tasks
- Appen Workflows supports complex, multi-step annotation
- Built-in quality control to guarantee high quality
- Security options to protect customer data

### Future direction

- Ongoing evolution of annotation tools
- Tighter integration with customers' internal systems
- Improved Workflows features to support more complex annotation tasks
- Ongoing evolution of quality and security features

# Appen Intelligence



## Appen Connect

Match our global crowd workforce to annotation tasks



## Appen Data Annotation Platform

Collect and annotate training data

## Appen Intelligence

Powers Appen products with proprietary ML models



## Appen In-Platform Audit

Organize and analyze training data to identify quality, distribution & bias



## Appen Mobile

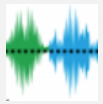
Engages, enables and expands crowd



appen

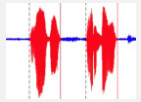
# Appen Intelligence - automated pre-labelling and quality control

## Audio & Speech

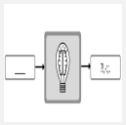

Speaker diarization


Language Detection


Speech Recognition

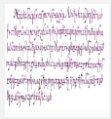

Audio Segmentation

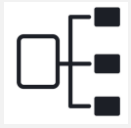

Gender Detection


Punctuation Restoration

## Text & Language


Text Tokenization


Gibberish Detection


Text Classification


Name Entity Extraction

Grammar & Spell Checking


Computer-Aided Translation

## Image & Video

OCR Transcription


Object Detection

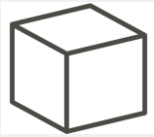

Object Tracking


Semantic Segmentation

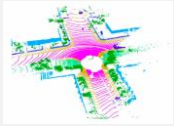

Blurring


Face Landmark

## 3D Point Cloud


3d Object Detection

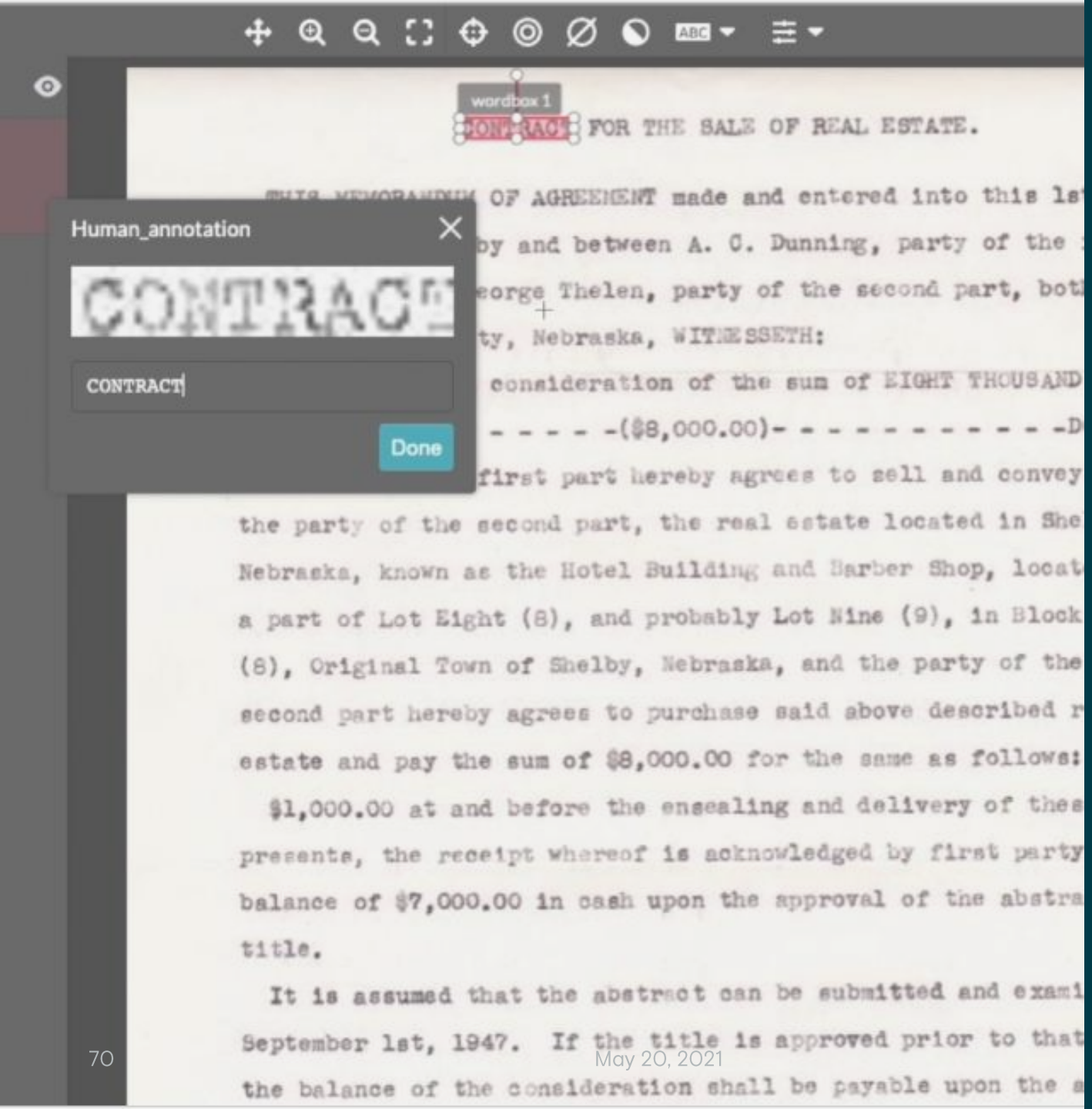

3d Object Tracking


3d Semantic Segmentation


3d Lane Detection



appen




# Demo 2

## Optical Character Recognition (OCR) Transcription

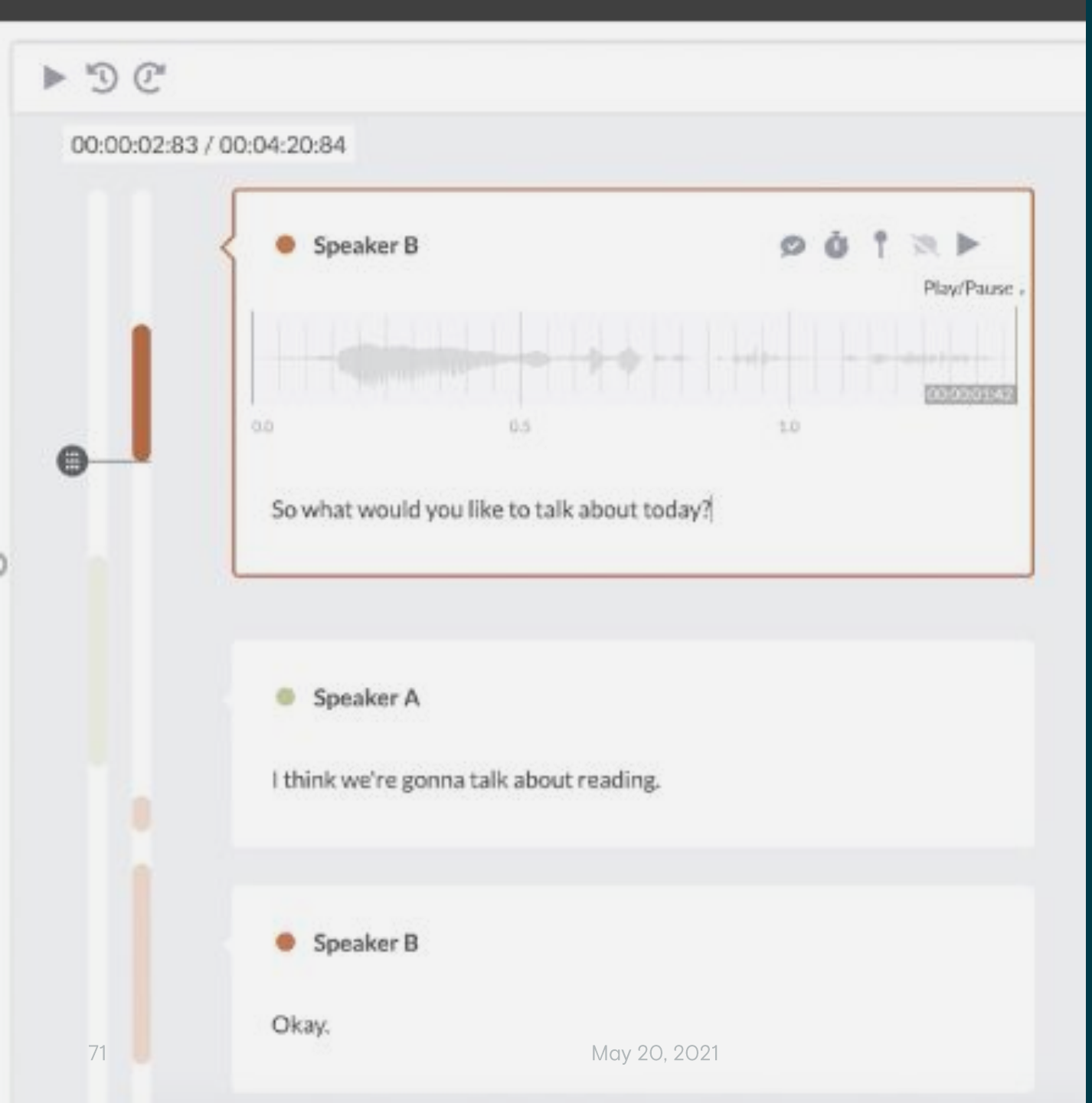




appen

# Demo 3

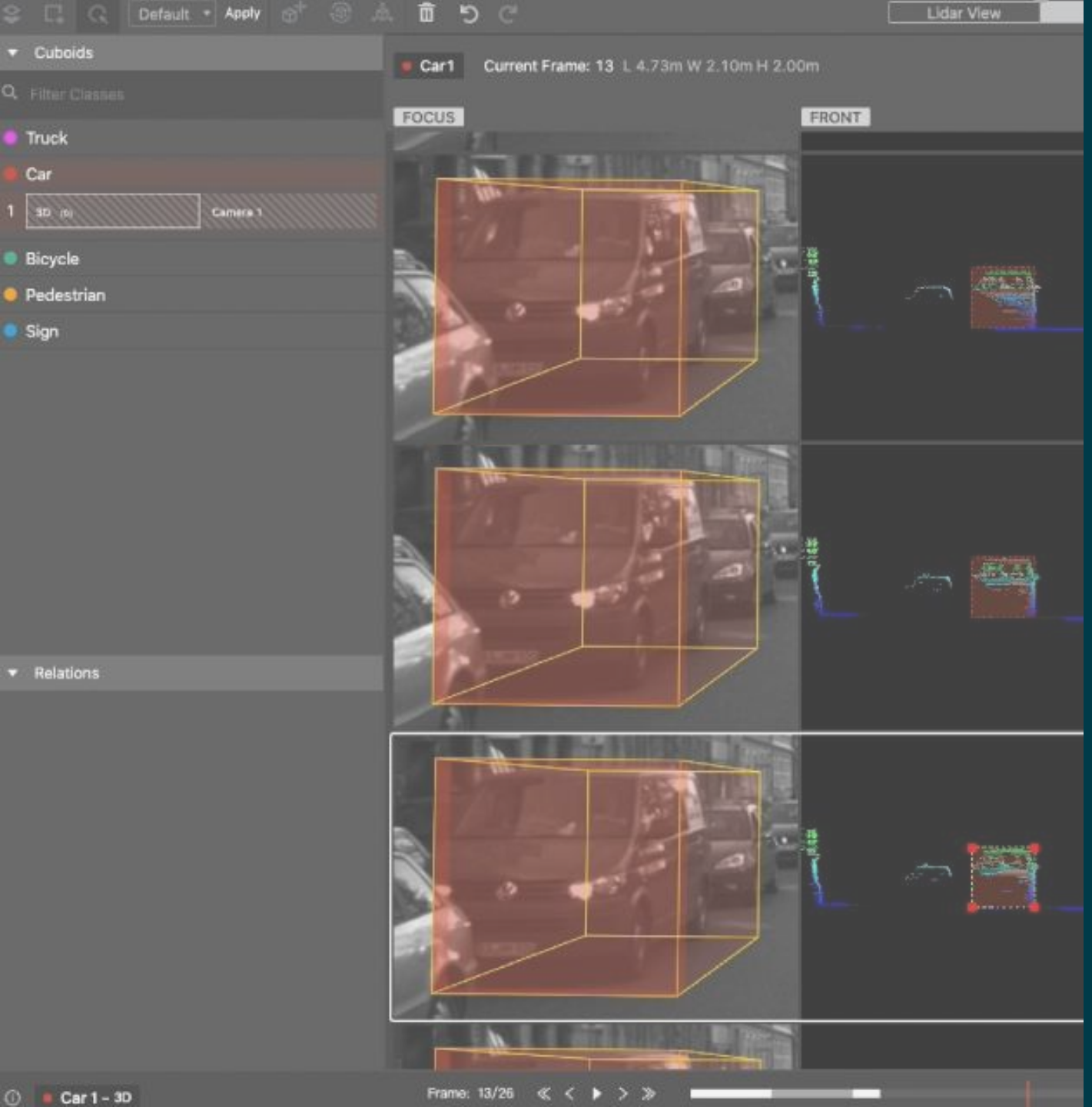
## Audio Transcription







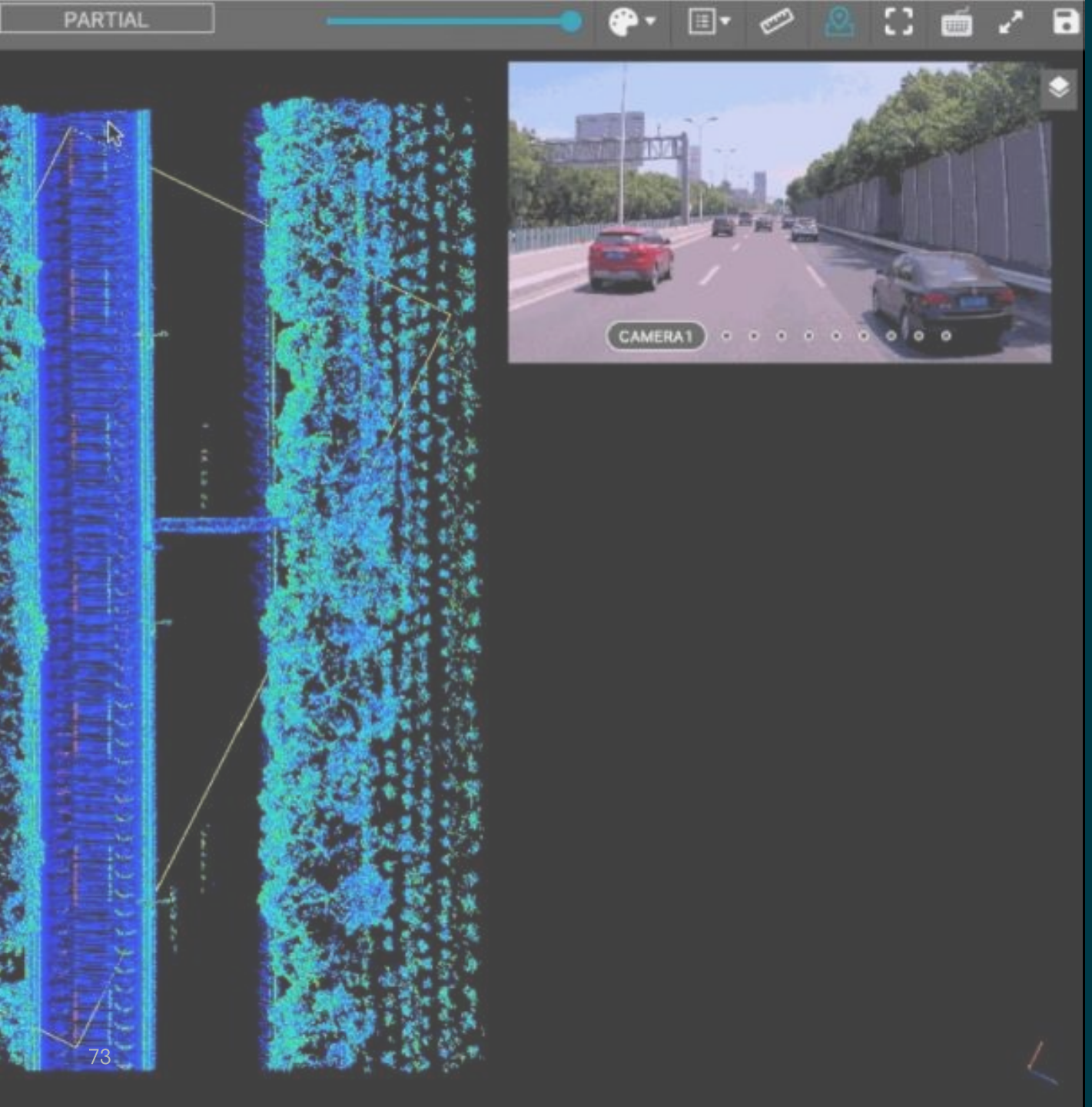

appen




# Demo 4

## LiDAR Annotation

appen




# Demo 5

## LiDAR Lane Line Segmentation



appen

# ML assisted annotation improves annotation time by up to 600%

| | | ML assisted annotation time improvement vs human annotation |
|---|---|---|
| **Audio & Speech** | Audio Transcription | **Up to 90%** |
| | Translation | **Up to 160%** |
| **Image & Video** | 2D Image Bounding Box | **Up to 30%** |
| | 2D Image Semantic Segmentation | **Up to 90%** |
| | Video Annotation | **Up to 500%** |
| | OCR Transcription | **Up to 600%** |
| **3D Annotation** | 3D Point Cloud Bounding Box | **Up to 400%** |
| | 3D Semantic Segmentation | **Up to 600%** |
| **Content Relevance** | Content Relevance | Tasks are subjective and difficult to automate |

*Note: These are results observed from A/B testing and certain production projects. The results can vary based on the actual project data.*

# Appen Intelligence
## Customer benefits and future direction

### Benefits for our customers

- Improved speed of annotation enables customers to obtain more data at lower unit cost
- Quality is typically improved compared to standalone human annotation
- Faster time to obtain training data shortens model iteration cycles
- Project management overhead is saved and crowd is managed by platform

### Future direction

- Expansion of models to support a wider variety of automation use cases
- Tighter integration with other products

# Appen In-Platform Audit



Appen Connect

Match our global crowd workforce to annotation tasks



Appen Data Annotation Platform

Collect and annotate training data



Appen Intelligence

Powers Appen products with proprietary ML models

**Appen In-Platform Audit**

Organize and analyze training data to identify quality, distribution & bias



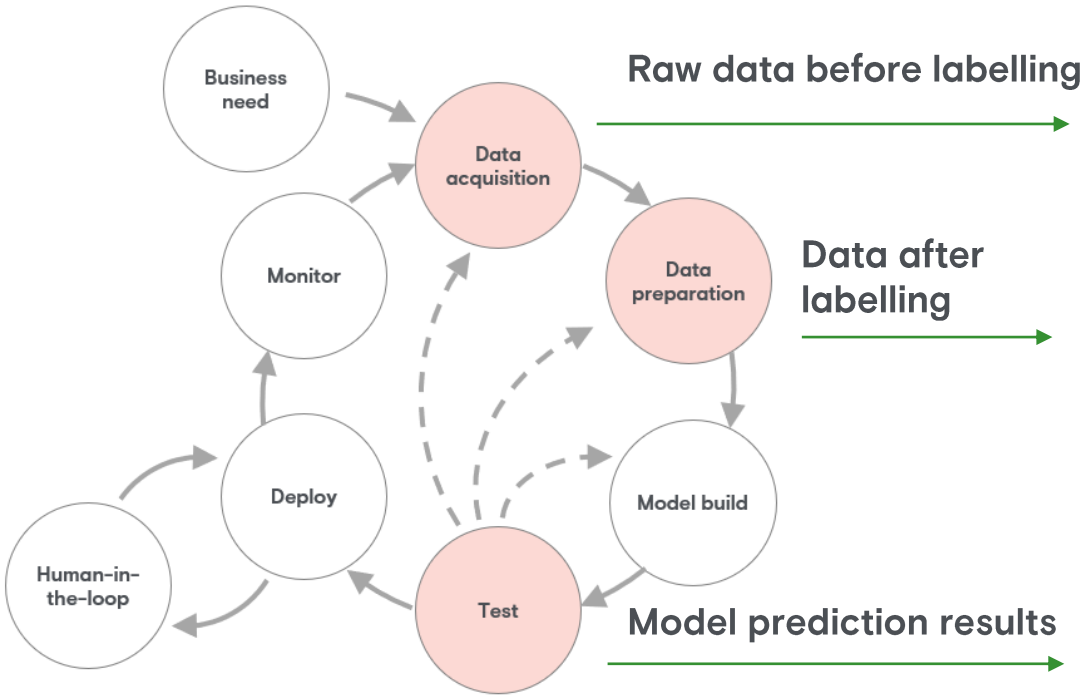
Appen Mobile

Engages, enables and expands crowd



appen

# In-Platform Audit helps to analyze and debug data




**Appen In-Platform Audit**

Analyse training data or ML model through human or machine learning

**Class imbalance**

- Number of examples in the raw dataset for each class is not balanced

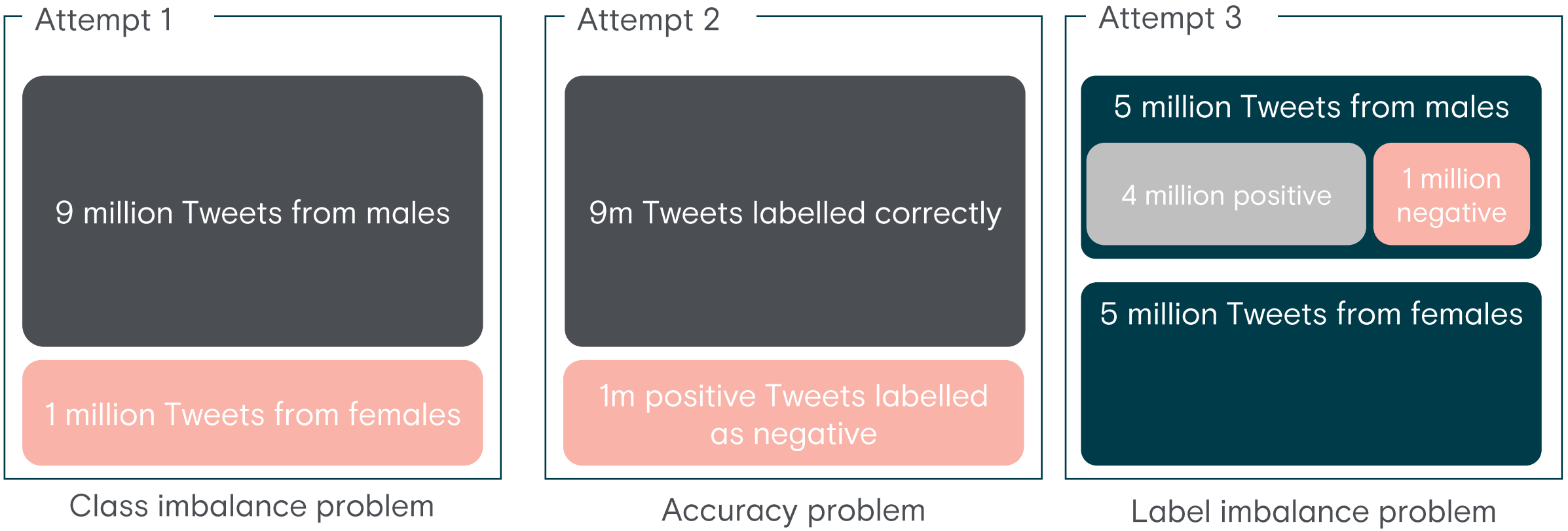
**Accuracy**

- Training dataset label accuracy

**Label imbalance**

- Number of examples in the training dataset for each class label is not balanced

# Example benefit of training data insights






appen

# Appen In-Platform Audit
## Customer benefits and future direction

### Benefits for our customers

- Ability to visually explore data to identify labelling errors
- Enables identification of training data composition and imbalances
- Support to debug and improve training data to improve AI performance

### Future direction

- Evaluate model performance
- Support more training data management needs

# Appen Mobile



Appen Connect

Match our global crowd workforce to annotation tasks



Appen Data Annotation Platform

Collect and annotate training data



Appen Intelligence

Empower Appen products with proprietary ML models



Appen In-Platform Audit

Organize and analyze training data to identify quality, distribution & bias

**Appen Mobile**

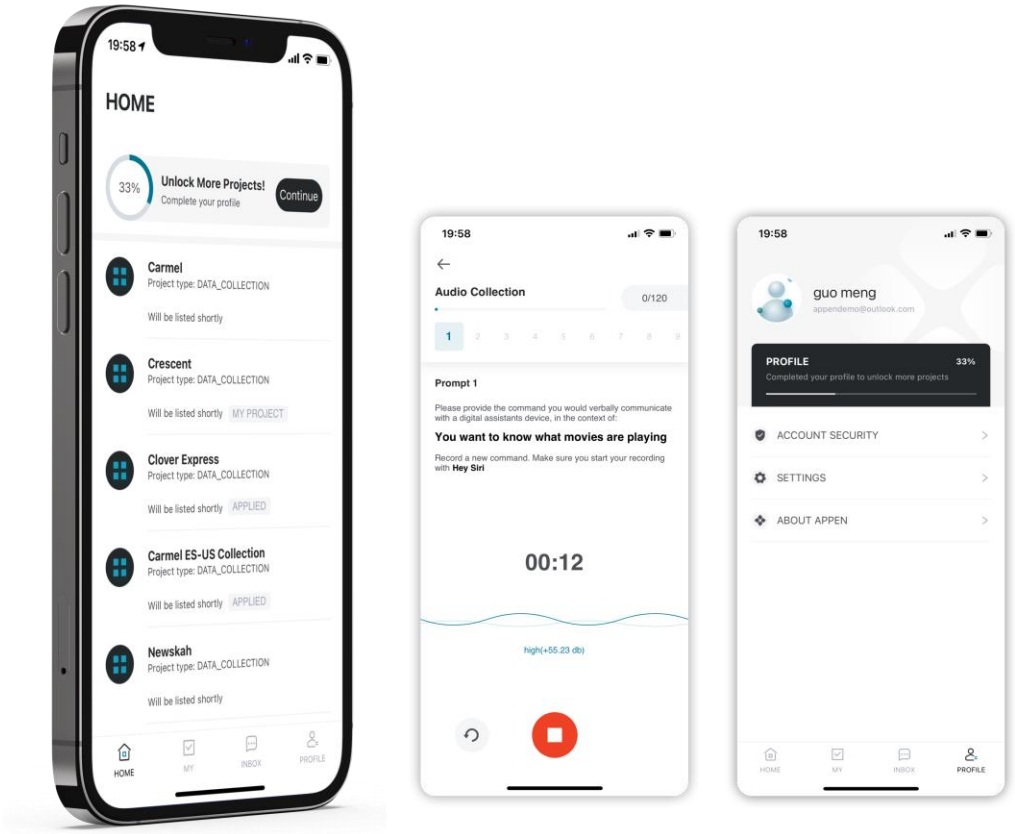
Engages, enables and expands crowd

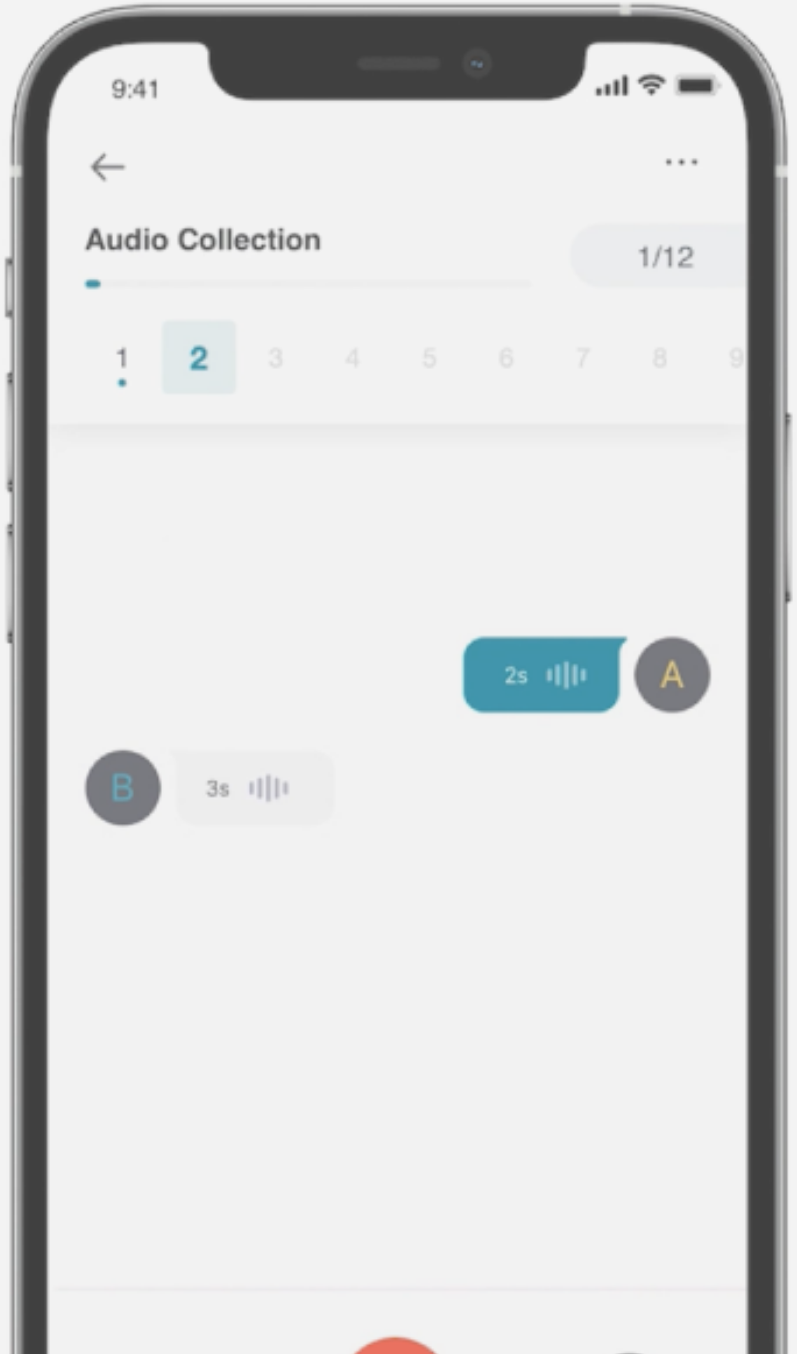


appen

# Appen Mobile provides an enhanced crowd experience and enables new use cases

- Simple and intuitive user experience
- Supports collecting many data modalities including audio, video, text, conversation, eye-tracking, and hand-written text
- Enables location-based data collection, heightened during pandemic
- Allows access to a broader set of mobile-only crowd workers – 65% of Global 2020 Internet traffic






# Demo 6

## Appen Mobile

appen

# Appen Mobile
## Customer benefits and future direction

### Benefits for our customers

- Improved sign-on and task experience for contributors leads to a broader crowd base
- Access to mobile-only workforce
- Mobile form factor provides enhanced data collection capabilities

### Future direction

- Collaborate with customers to identify new data collection techniques
- Support annotation tasks beyond data collection
- Actively expand crowd to mobile-only areas, supporting diversity and impact sourcing

# Recap and closing

# The Appen Product Suite

New products

## Appen Connect

Match our global crowd workforce to annotation tasks

## Appen Data Annotation Platform

Collect and annotate training data

## Appen Intelligence

Empower Appen products with proprietary ML models

## Appen In-Platform Audit

Organize and analyze training data to identify quality, distribution & bias

## Appen Mobile

Engages, enables and expands crowd



appen

# Value created from our products

New products

| Appen Connect | Appen Data Annotation Platform | Appen Intelligence | Appen In-Platform Audit | Appen Mobile |
|---|---|---|---|---|
| AI-augmented data collection and labelling | AI-enabled crowd management | Embedded expertise | In-built crowd management features | Native integration with our crowd |
| **Improves speed, quality, scale and unit economics** | **Increases internal productivity and crowd experience** | **Delivers high quality annotation** | **Reduces risk for customers** | **Creates competitive differentiation** |



appen

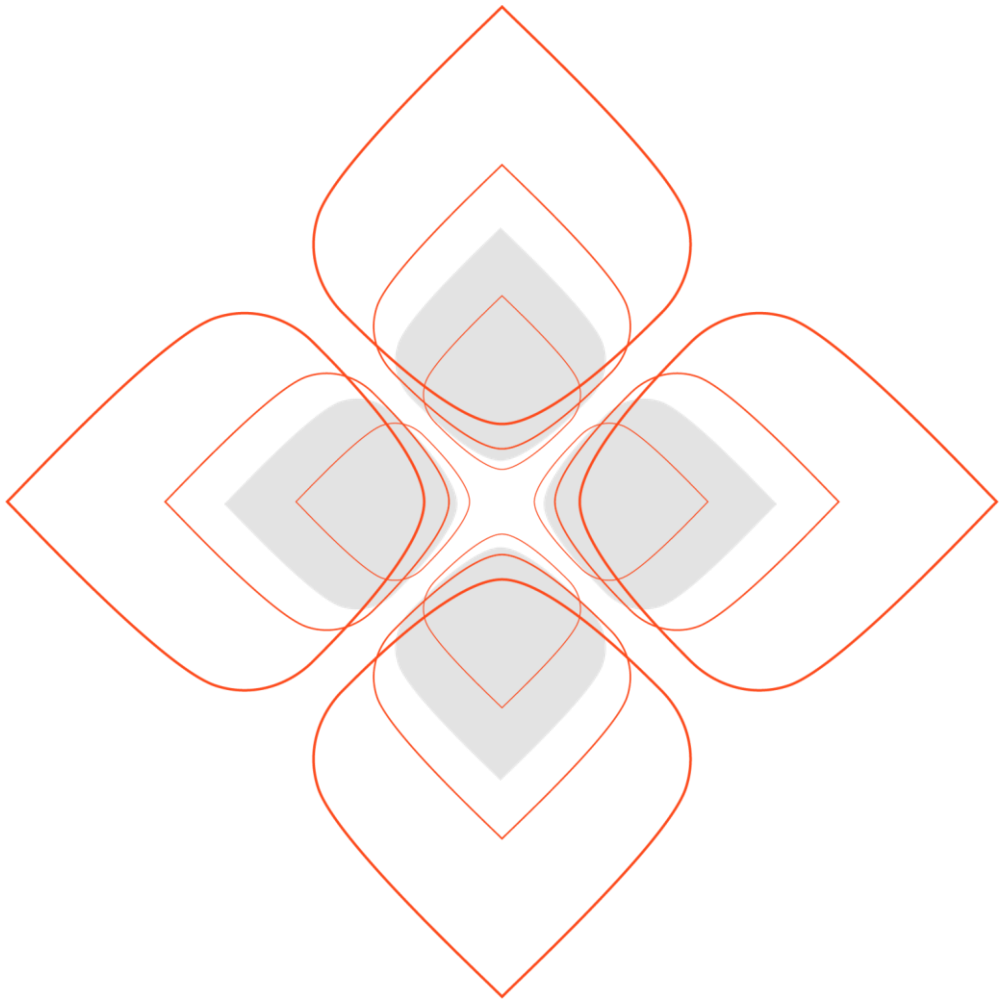
# Appen has the unique combination of industry leading technology, crowd of over 1m and deep internal expertise




appen

# Q&A

# Appen is accelerating its transformation into an AI powered provider of AI data and solutions



# End



appen