Data with a human touch

High-quality data for machine learning, enhanced by human interaction

appen.com
Why the Right Data is Essential for Success

Investment in AI in 2016 was in the range of $26 billion to $39 billion*. Yet, even with the huge advances made in AI solutions in the last decade, and the growing number of them on the market and in our lives, this basic truth holds: AI is only as good as the machine learning data that trained it.

To build a successful solution, you need the right data, and a lot of it. You can have the most appropriate algorithm, but if you train your machine on bad data, then it will learn the wrong lessons, will not meet expectations, and not work as you (or your customers) expect. Your success is almost entirely reliant on your data.

As enterprises race to stand out from the competition and do more with the same resources, structured, human-annotated data can help optimize their machine learning efforts to achieve their business goals.

The Appen Difference

The Scale
With a crowd of 1,000,000+ people in over 130 countries working in more than 180 languages and dialects, we help our clients quickly ramp up their teams so they can respond to ever-changing customer needs.

Quality
We’ve built a strong reputation providing high-quality data that is tailored to the specific needs of our clients to fuel their machine learning and artificial intelligence initiatives.

Experience
With over 20 years of experience working with global firms in various industries, we have a proven track record of solving a wide variety of data challenges.

Content Relevance
Train your algorithms with high-quality data to give users more relevant results. In an environment where users expect faster, more relevant search results—and increasingly personalized experiences, you need to ensure that your machine learning algorithms are being trained on a regular basis with high-quality, human-annotated data. To scale to a global user base, you also need experienced local users who can provide the right level of cultural understanding.

Appen provides a suite of content relevance services across web search, eCommerce site search and social media sites to support your machine learning initiatives and provide an enhanced user experience for your customers.

Use Our Data to Train Your Machine Learning Model for a Variety of Use Cases

Search Relevance
Interpret queries, assess the user intent, and ensure that relevant and useful results are returned to the user.

Categorization & Tagging
Organize and tag your data to increase conversion rates.

Personalization
Personalized data collection to enhance social media feeds and eCommerce search results.

Content Moderation
Human moderation of data such as product reviews to ensure quality control.


Leaders in Machine Learning are adopting the following use cases to improve the end user experience for web search, eCommerce sites and social media platforms:

**eCommerce Site Search**
- Auto Fill Evaluation
- Content Moderation
- Defect Testing
- Entity Tagging
- Field Testing
- Query Relevance
- Side by Side Evaluation
- UX Testing

**Web Search**
- Ads Relevance
- Adult Search
- Caption Evaluation
- Categorization
- Document Quality
- Field Testing
- Fresh/News Evaluation
- Map Verification
- Media Search
- Multimedia Search
- Query Relevance
- Related Search
- Side by Side Evaluation
- Social Search
- Spam Junk Page Evaluation
- Spelling
- Whole Page Measurement

**Social Personalization**
- Ads Evaluation
- Text Data Collection
- Content Annotation
- Entity Evaluation
- Field Work
- Image Categorization
- Media Search
- News Feed Evaluation
- News Search Evaluation
- Offensive Content Review
- Personalized Results Evaluation
- Categorization
- Query Classification
- Query Relevance
- Refined Search Evaluation
- UX Testing
- Video Categorization
Secure Services
In scenarios where you are working with confidential customer information such as PII, financial or government records, you’ll want to be sure that your data is handled in a secure environment by individuals with the appropriate security clearance. Whether you need this data to be transcribed, annotated or translated, it is critical to ensure that the partner you choose has the resources to work with sensitive data.

Our secure facilities and transcription operations are ISO 9001 and ISO 27001 certified so you can rest assured your data will remain protected and quality controlled. We also hold the Cyber Essentials certification as additional independent assurance that we have the appropriate levels of protection in place when working with secure data.

What is ISO 27001?
ISO 27001 is a specification for an information security management system (ISMS). An ISMS is a framework of policies and procedures that includes all legal, physical and technical controls involved in an organization’s information risk management processes.

What is ISO 9001?
ISO 9001 is the international standard that specifies requirements for a quality management system (QMS). Organizations use the standard to demonstrate the ability to consistently provide products and services that meet customer and regulatory requirements.

The Benefits of ISO 27001
- Protects critical and sensitive information
- Provides comprehensive, risk-based approach to secure information and data
- Demonstrates credibility, trust, satisfaction, and confidence with stakeholders, partners, and customers
- Adheres to security status per internationally accepted criteria

The Benefits of ISO 9001
- Provides effective quality management system
- Increases staff awareness and operating efficiency while reduced costs
- Demonstrates commitment to quality, customers and continuous improvement
- Ensures compliance with legislation and regulatory standards
- Demonstrates credibility, trust, satisfaction and confidence with stakeholders, partners and customers

Data with a Human Touch
Data Collection
Machine learning models require large volumes of high-quality data to be trained effectively.
Scale your data collection efforts across multiple file formats including text, image, video and speech.

Data Annotation
Annotated data enables richer and more valuable machine learning-based products. Appen’s curated crowd allows you to get the high-quality data you need to develop better products for your customers.
Are you equipped to source the large amount of language data needed to train and test your machine learning model? Do you have the resources to scale your data collection to meet customer needs in all your target markets, including in low-resource languages?

Appen can help. We understand the complex needs of today’s organizations. For the past 20 years, Appen has delivered the highest quality linguistic data and services, in over 180 languages and dialects, to government agencies and the world’s largest corporations.

**Data Collection**

You build the technology – Appen hand-crafts the data for you.

- Automatic speech recognition (ASR)
- Text-to-speech (TTS) engines
- Natural language processing
- Machine translation
- Computer vision & AI platforms

**We help you train and evaluate these systems so you can build these applications:**

- Virtual assistants
- Home automation devices
- Gaming consoles & smart TVs
- Call-center systems
- In-car infotainment systems

**Data Collection**

- Automatic speech recognition (ASR)
- Text-to-speech (TTS) engines
- Natural language processing
- Machine translation
- Computer vision & AI platforms

**Looking for speech data? Appen’s got it covered.**

- Scenario-based natural language text collections
- Handwriting and keyboard input
- Non-speech audio recordings (e.g. gunshots, shattering glass)
- Image capture
- Video recordings

**Here’s why we’re different:**

- Unrivaled breadth of data collection coverage
- Global community of over 400,000 pre-screened speakers, covering 180+ languages
- 20 years of experience running collections of all sizes in 90+ countries
- No limit to languages, geography, collection modes and data volumes
- Experienced project managers, language/domain specialists and audio engineers

**Environment**

- Home
- Office
- Studio
- Outdoor
- In-car

**Recording Technology**

- Smartphone collection app
- Telephony server
- Single/multi-channel professional audio recording kit
- Personal computer with a built-in or USB mic
- Proprietary customer devices

**Speech Modality**

- Scripted
- Scenario-based responses
- Role-plays
- Spontaneous conversations

**Participants**

- Single
- Pairs
- Groups

**Quality Speech and Language Data to Power Your Applications**
Transcription and Annotation

Are you working on a cutting-edge acoustic model? Appen’s transcribed data can take it to the next level.

Our proprietary web-based transcription interface - Appen Global - accelerates ramp up time so we can deliver large volumes of high-quality transcription and annotation data to you.

What makes us unique:

• 20+ years of experience in human transcription and annotation
• 40 million audio files transcribed in 2016 alone
• 180+ languages supported
• Proprietary web-based transcription interface (Appen Global)
• All native alphabets and orthographies
• Speech, text, video and handwriting supported
• ISO 27001 and ISO 9001 certified for highly secure transcription

Here’s how we guarantee quality:

• Screening & training of in-country transcribers
• Automated spelling checks & rigorous post-processing by senior team members
• Extensive experience developing spelling standardization in low-resource languages

And don’t forget to ask about our premium extras:

<table>
<thead>
<tr>
<th>Orthographic transcription</th>
<th>Annotation of transcribed data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Automated or manual segmentation</td>
<td>• Foreign / accented speech</td>
</tr>
<tr>
<td>• Time alignment</td>
<td>• Foreground / background speech</td>
</tr>
<tr>
<td>• Acoustic tagging</td>
<td>• Mispronunciations and speech disfluencies</td>
</tr>
<tr>
<td>• Speaker ID / tracking</td>
<td>• Background / overlapping speech</td>
</tr>
<tr>
<td>• Silence detection</td>
<td>• Slot value identification</td>
</tr>
</tbody>
</table>

Translation and Localization

Got a great local product? Appen helps you take your technology into new global markets with:

• Localization of human-machine interfaces into 50+ languages
• Evaluation and tuning of machine-translation technology
• Post-editing and end-to-end support for all MT-related needs
• Traditional document translation, software and website localization services
• Global crowd coverage to meet the demand for low-resource languages

Our expertise covers a variety of domains, including:

• Conversational speech translation
• Speech-to-speech MT evaluation
• Capture of persona and intent
• Live-chat, military and disaster relief
• Navigation, multimedia and voice search
• BNF grammar development

Image Annotation

Our image annotation services accurately describe image content for use in training object recognition systems, at scale. Appen’s in-house platform provides bounding boxes, polygons and much more, with built-in QA processes to make sure your training data is top of the line.
Custom Linguistic Solutions
Appen’s high-end linguistic annotation support helps you fine-tune your ASR, TTS and NLP systems and expand them into new markets and domains.

Licensable Resources
Do you need immediate access to a complete speech and language database? Appen has an extensive catalog of off-the-shelf, licensable resources ready to ship. We even cover low-resource languages, including dialects from West and North Asia, the Middle East and Africa.

Ask for a copy of our catalog today!
Appen is a global leader in the development of high-quality, human-annotated datasets for machine learning and artificial intelligence. Appen brings over 20 years of experience capturing and enriching a wide variety of data types including speech, text, image and video. With deep expertise in more than 180 languages and access to a global crowd of over 1 million skilled contractors, Appen partners with 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.

Coverage for over 180 languages & dialects

Access to a global crowd of over 1,000,000

20+ years of experience providing high-quality, human-annotated data for machine learning for leading solutions in speech, search, eCommerce and more