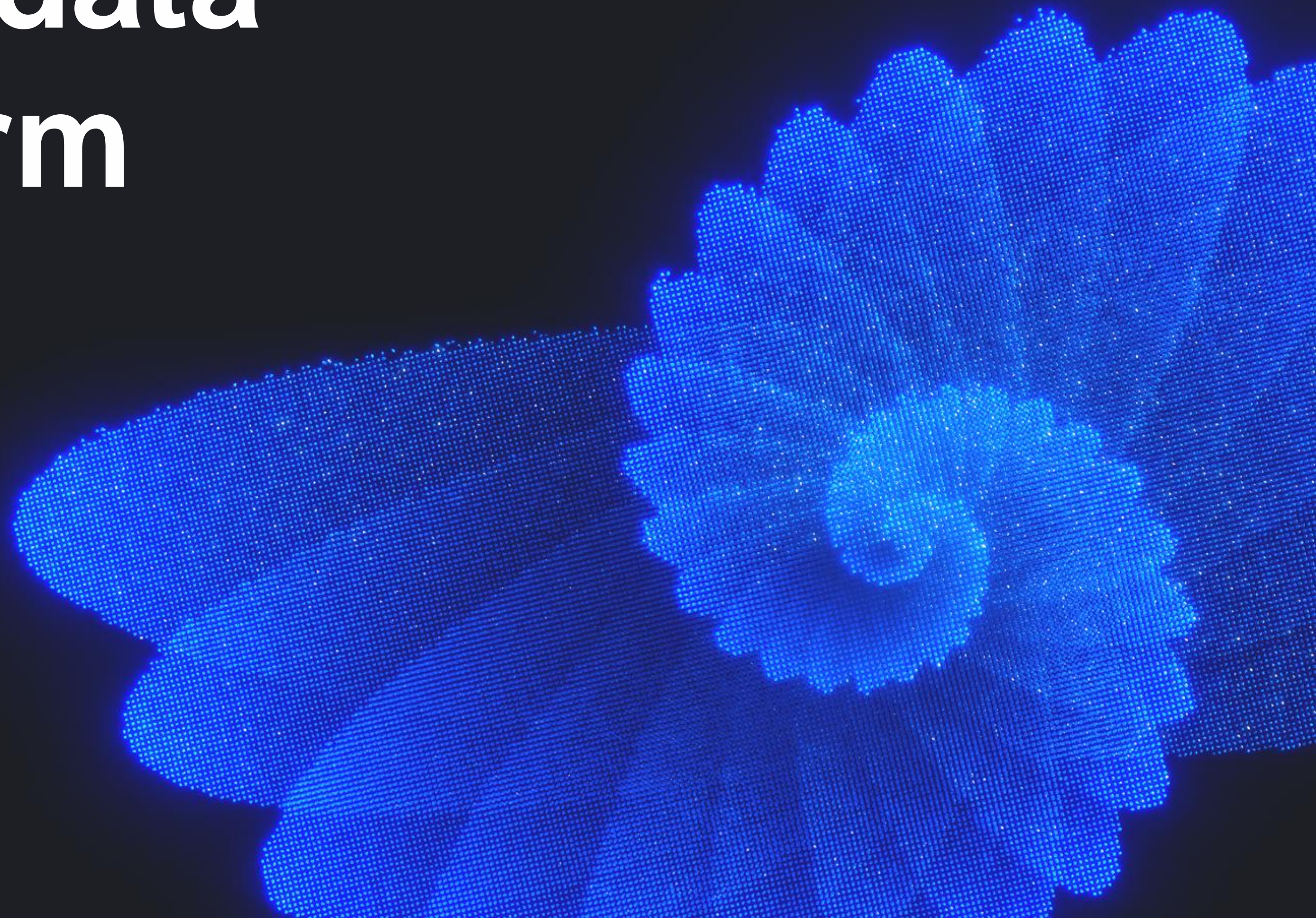




Your one-stop data labeling platform

Fast data iterations and easy scaling
to support AI/ML development

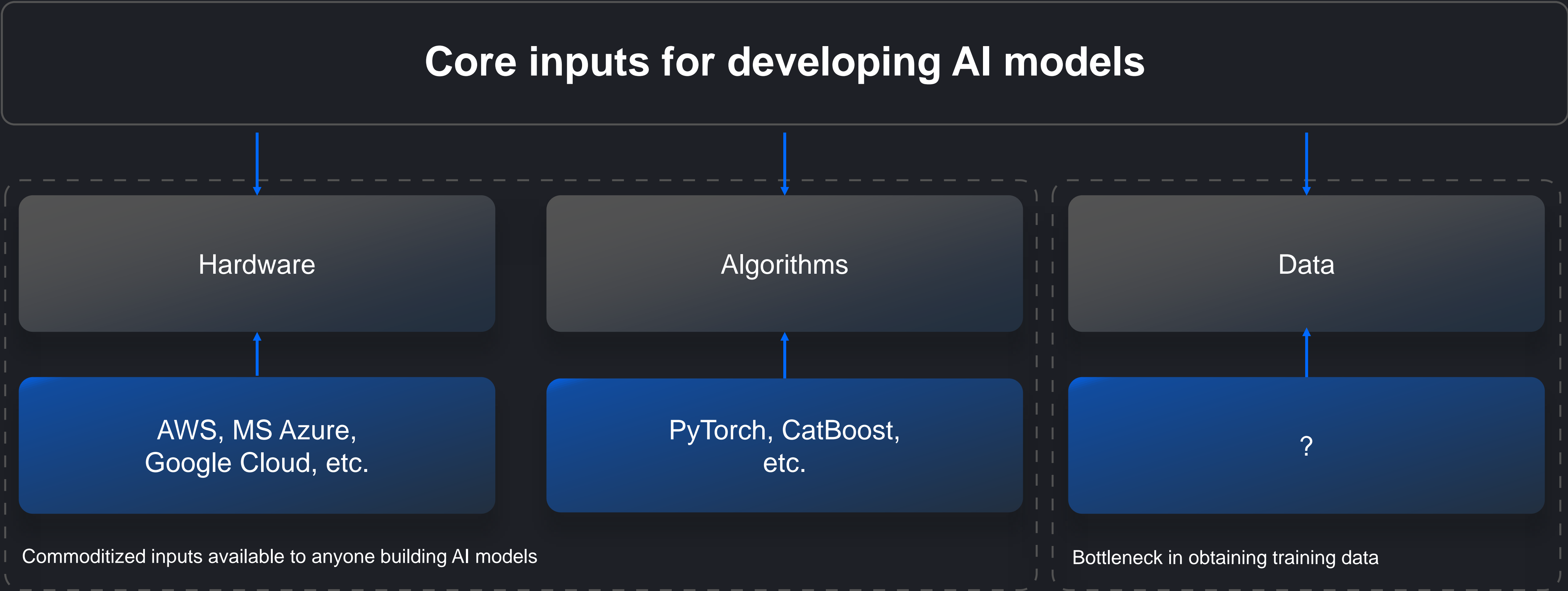
Powering data-centric AI development since 2014



Intro

Context: Data-centric AI is trending in the ML world

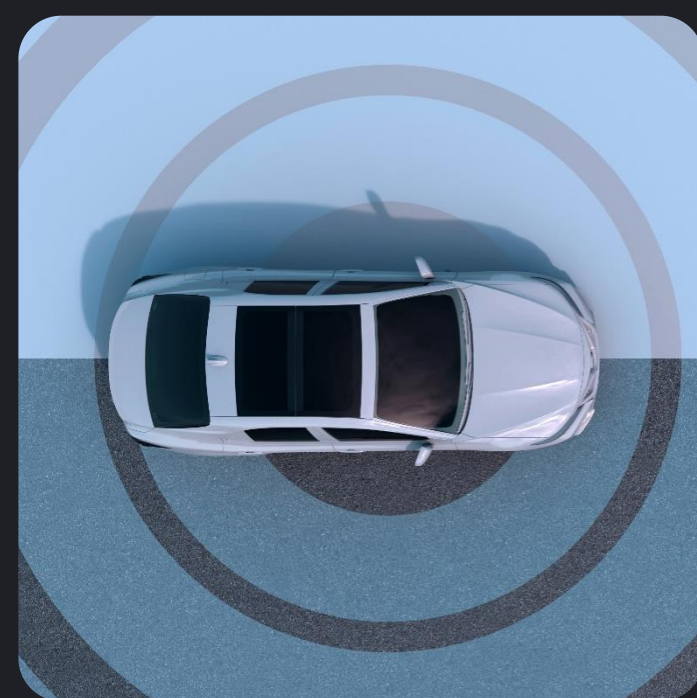
Everyone has access to hardware and algorithms.
Data is the key to gaining a competitive advantage.



Quality AI demands huge amounts of fresh data for ML production

Accelerate time-to-value with scalable human data labeling

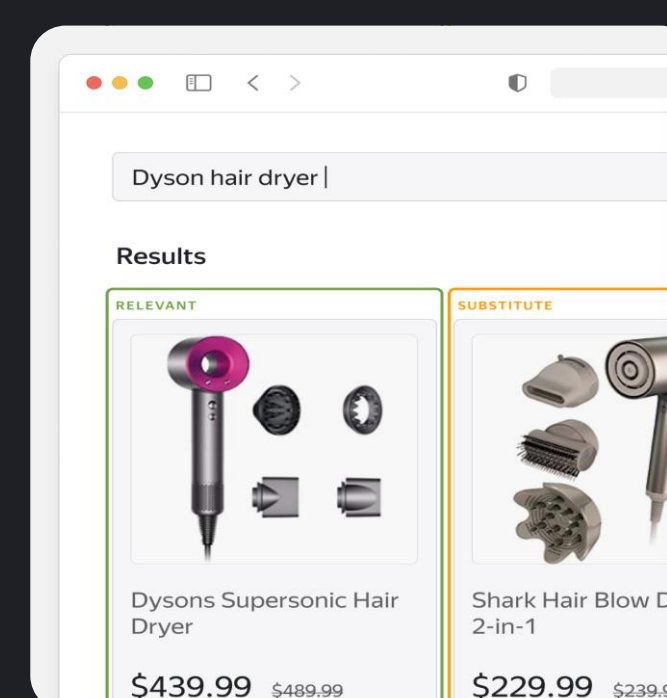
- **High-impact solutions are powered by manual data labeling**



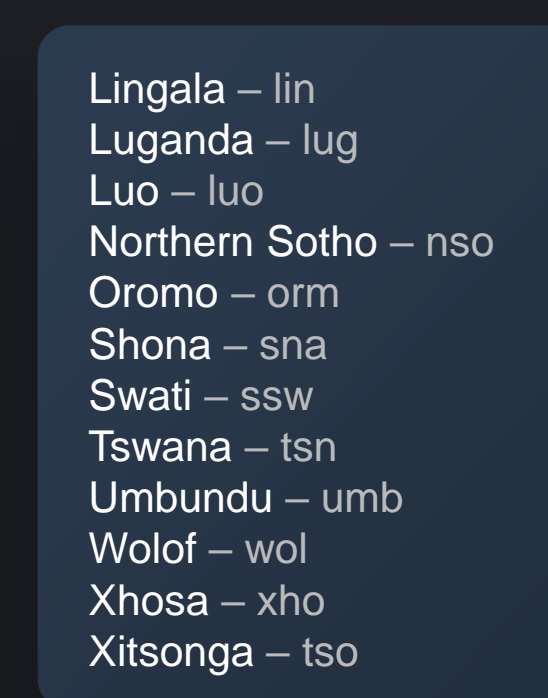
Autonomous Driving



Speech Technologies

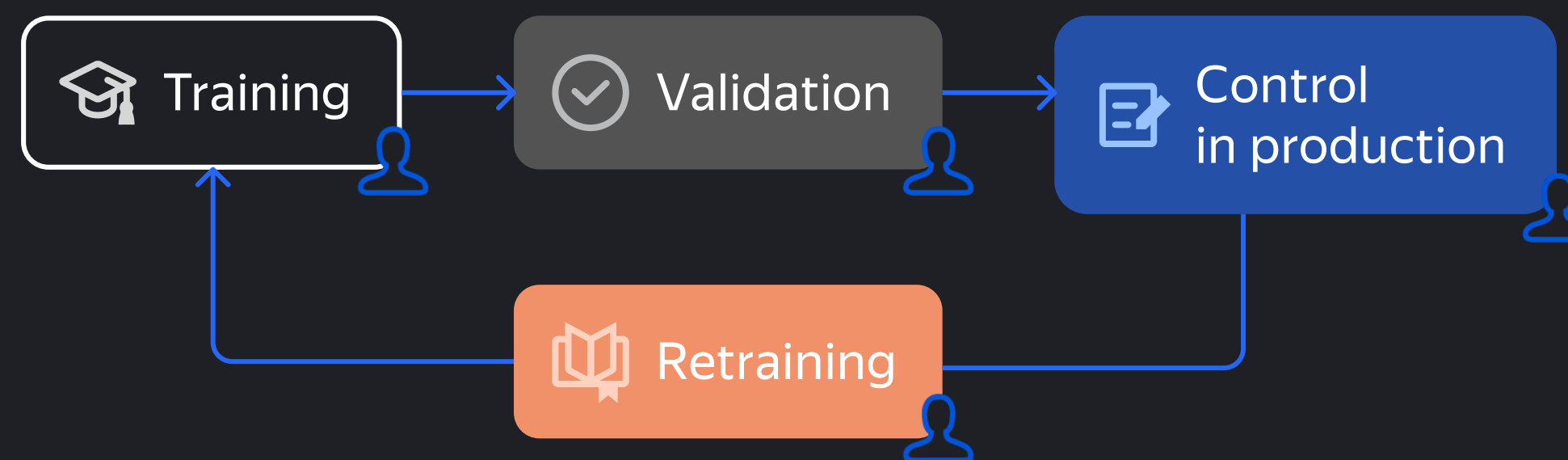


Search



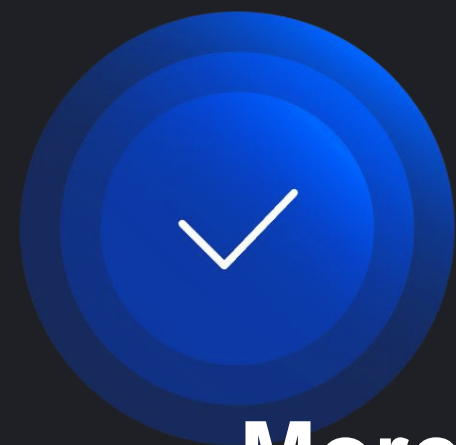
Machine Translation

- ML development requires human-labeled data at every step

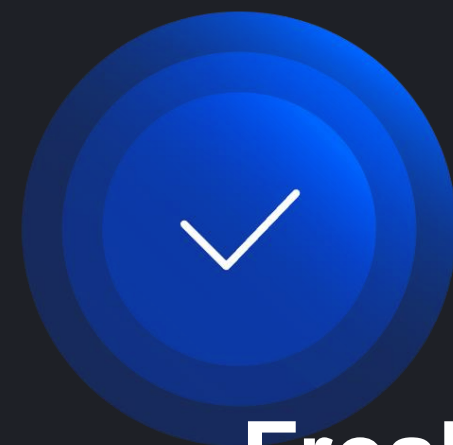


Improve quality of AI-based products by focusing on data

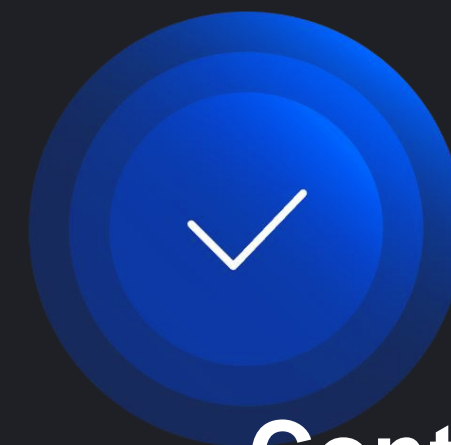
A scalable solution can produce larger datasets with shorter timeframes and faster feedback on model performance



**More training
data leads to
better model
accuracy**



**Fresh data for
continual
retraining helps
prevent data drift**



**Continual human
feedback
accelerates
validation
and deployment of
new models**

Challenges of working with data

- Sourcing the data
- Finding resources to label large datasets
- Managing and automating data collection and labeling
- Choosing the right tools for in-house labeling
- Scaling up and down with fluctuating data volumes
- Optimizing data processes with the right combination of automated and manual data labeling

Global trends for data labeling and collection



Crowdsourcing platforms
(self-service)



Data labeling vendors
(turnkey solutions)



In-house data labeling

Toloka: delivery models



Open crowdsourcing platform

Self service

Wide range of quality control tools and state-of-the-art technologies. Highly customizable for all your data workflows

☑ On-demand global crowd

☑ In-house workforce

- Design your own data flow or use preset projects
- Solution maintenance fully on Client side
- 24/7 tech support
- API + Python libraries

- Flexible price per task + platform fee
- Pay as you go, no data minimums



Bespoke solutions to support the ML lifecycle

Managed service

Individual solutions developed on request to support data-related processes across the entire machine learning lifecycle

- Provide the project specs and get a custom solution that's ready to use
- Solution maintenance fully on Toloka side
- Dedicated manager
- API

- Fixed price per label
- Annual commitments with volume-based pricing

About Toloka

Data labeling solutions to transform your AI

Founded in 2014 after years of research and experimentation, Toloka is a global tech company that develops a platform and environment to support data-related processes.

Designed by engineers for engineers, Toloka enables data scientists and ML teams to get ML solutions to production faster by:

- testing hypotheses
- boosting the success rate of prototypes
- building optimal data production pipelines

10+ years

of industry experience, scientific research,
and contributions to the AI community

Trusted by leading global ML & AI teams

AliExpress

bestplace

BIOTRONIK

FAANG

Handl

IDR&D

JET
BRAINS

LEROYMERLIN

NAVER
LABS

PLUGANDPLAY

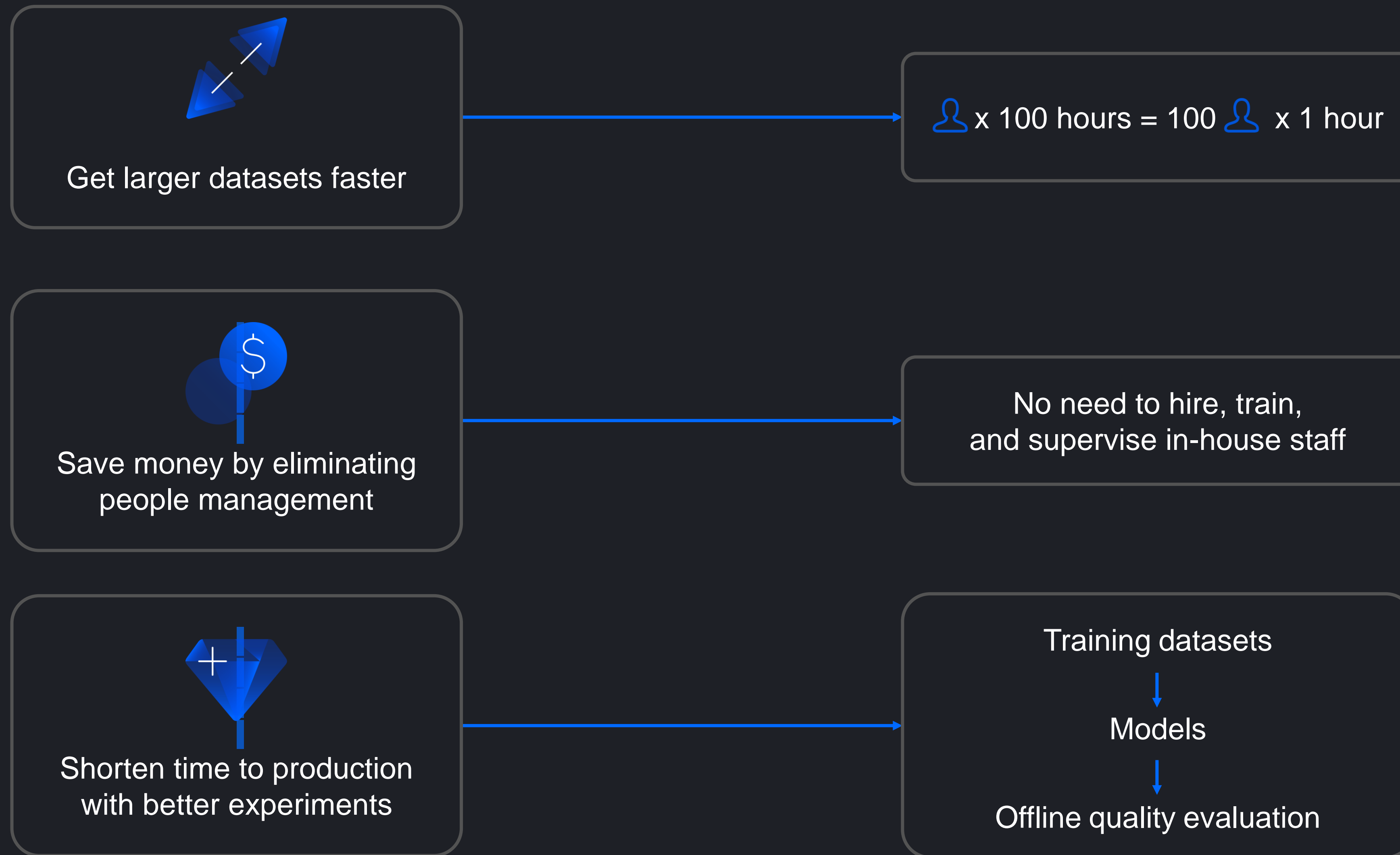
Replika

SAMSUNG

trivago®

Yandex

Why choose a highly scalable data solution like Toloka



Our data labelling platform is purpose-built for scaling



Loved by leading ML teams



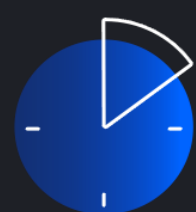
Fast scalability



Short turnaround time



Wide range of quality control tools



Real-time data labeling



API and Python SDK for easy integrations



Clear pricing



“We were really impressed with how fast we got our project done in Toloka — 10,000 ads were reviewed in just 12 hour.”
Special Projects Team



“Thanks to Toloka, we’re able to run numerous data projects on a regular basis. What we gain is a dependable approach to data labeling.”
Crowd Solutions Architect



“With Toloka we were able to resolve even the most difficult cases of recognizing handwritten text in documents for our customers.”
Founder and CTO of Dbrain, Y Combinator alum



“We chose Toloka because of the fast turnaround time and the active participation of performers.”
Data engineer

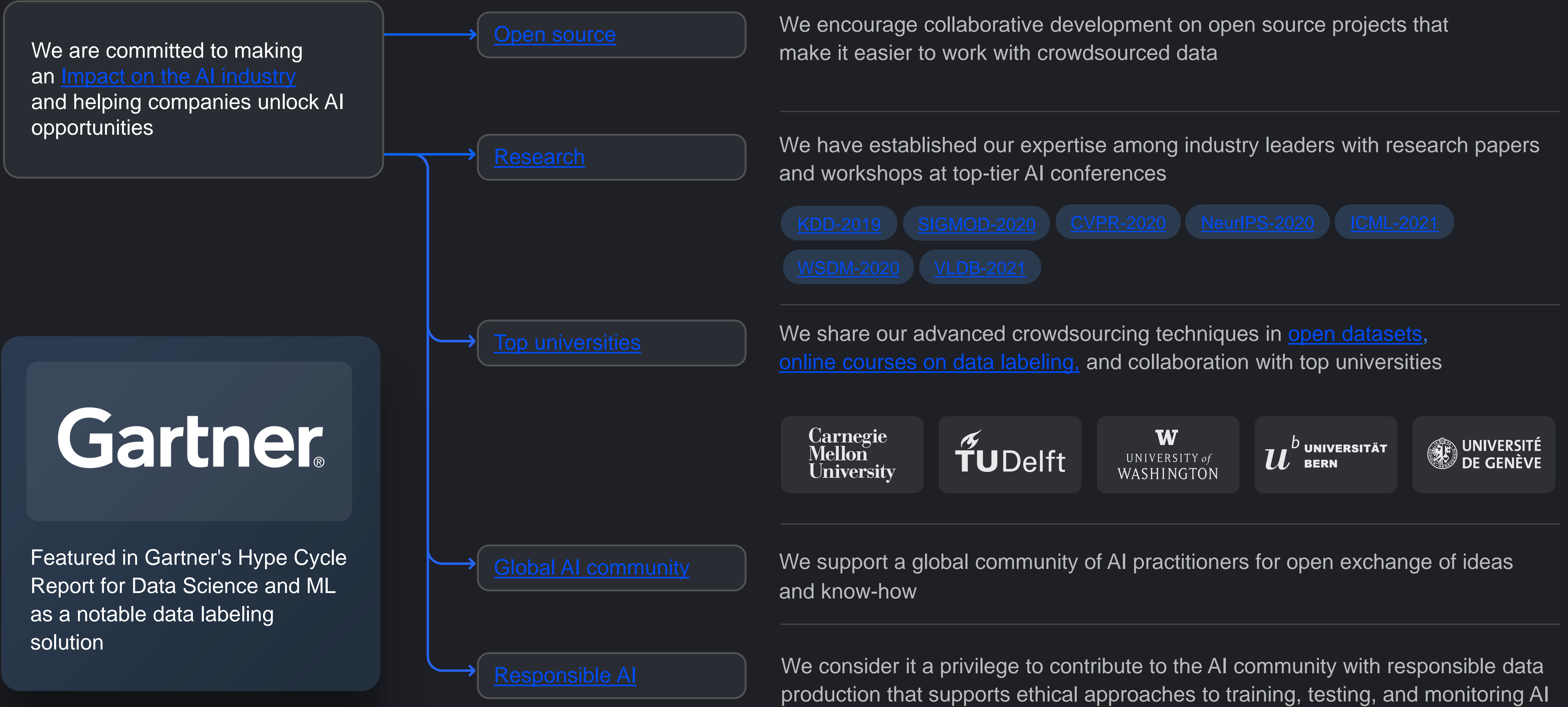


“We choose Toloka because of high throughput for large data volumes, – we collected the world’s largest database of 200,000 unique photos and videos.”
Science Director and Co-founder



“Toloka is the first place we go to prepare data for AI. We get a full set of quality control tools and it’s 10 times cheaper than our previous solution.”
Head of Technologies






Recognized as an industry advocate








Why Toloka is the platform of choice

Toloka’s capabilities are rooted in Intelligent technologies, a diverse global crowd, and infrastructure to support the most demanding projects






State-of-the-art technologies

-  Multiple quality control methods
-  Adaptive selection of performers
-  Smart matching mechanisms
-  Autolabeling (on demand)
-  API and open-source libraries for seamless integration

The largest global crowd coverage

-  **100+** Countries
-  **40+** Languages
-  **200k+** Monthly active Tolokers
-  **800+** Daily active projects
-  **24/7** Continuous data labeling

Robust infrastructure

-  **499M+** Tasks per month with exceptional throughput
-  Privacy-first, GDPR-compliant focus on data protection
-  Information security management system (ISMS) is ISO 27001 — certified
-  Multiple data storage options, including Microsoft® Azure cloud for world-class protection
-  Dedicated antifraud team monitors adherence to business policies

The largest global crowd coverage

For multilingual projects
and fast scaling

100+

countries with
active Tolokers

40+

languages
spoken

Top languages

· English · Spanish · Arabic · Portuguese · Russian
· Ukrainian · French · German · Italian · Polish · Hindi
· Latvian · Bulgarian · Czech · Turkish · Vietnamese
· Japanese · Chinese · Korean · Indonesian

State-of-the-art technologies

Transform the crowd into computing power with advanced technologies for quality management



Multiple quality control methods

Toloka offers different approaches to achieve the best quality for each project

-  Post-verification
-  Task-based crowd training and testing
-  Golden sets (honeypots) to monitor quality
-  State-of-the-art aggregation tools
-  Platform-wide anti-fraud system



Adaptive selection of Tolokers

Multi-stage selection of a distributed crowd

-  Audience filters by language, age, gender, interests, location, real-time ranking, and more
-  Training, exams, and retraining to find Tolokers for your exact task

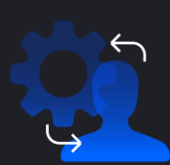

Smart matching mechanisms

Patent-pending matching system that honors the preferences of requesters and Tolokers for mutual benefit

-  Invite Tolokers to a project who are most qualified to handle it
-  Offer Tolokers personalized recommendations of interesting projects they will enjoy

Autolabeling & verification (on demand)

Autolabeling and pretrained models with quality control built in

-  Automated prelabeling. Results are verified by human Tolokers for high accuracy.
-  Human in the loop workflows

Designed for versatile use cases

E-commerce

Recommendation systems
Search relevance
Product page translation
Online product categories
Moderating reviews
Price extraction
Generating product descriptions
Marketing surveys
Product search by image
Monitoring support quality
Social media monitoring

Field data collection

Verifying addresses
Verifying business hours
Monitoring products on retail shelves

NLP

Search relevance
Text classification
Sentiment analysis
Intent classification
Utterance collection
Named entity recognition

Voice assistants

Speech to Text & Text to Speech
Verifying voice assistant responses
Recording activation phrases

Computer vision

Object detection
Image segmentation
Image classification
Image transcription
Side-by-side comparison
Image and video collection

Social science

Qualtrics surveys

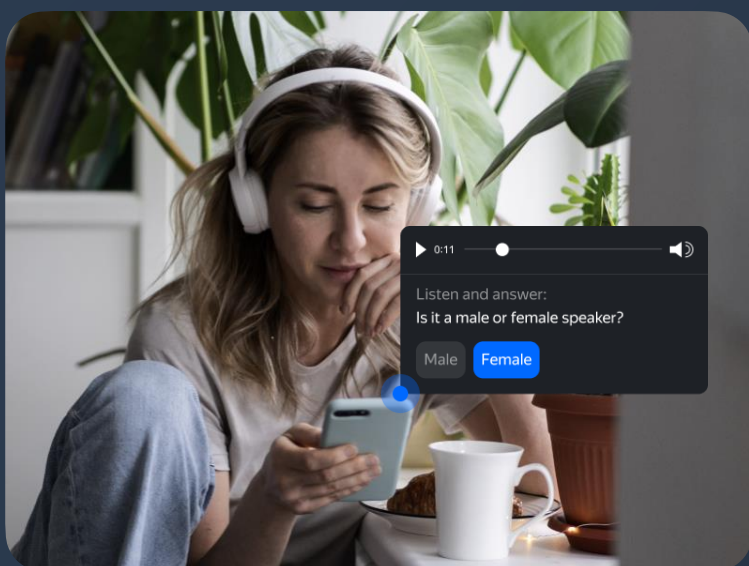
Turn-around time and cost in real use cases



Object classification

1000 photos

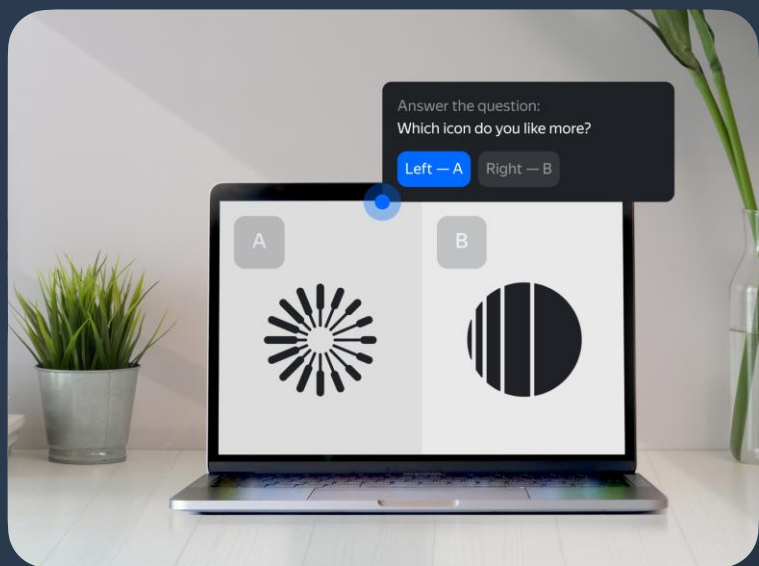
🕒 15 min 💰 \$2.4



Audio transcription

100 recordings
(25 min each)

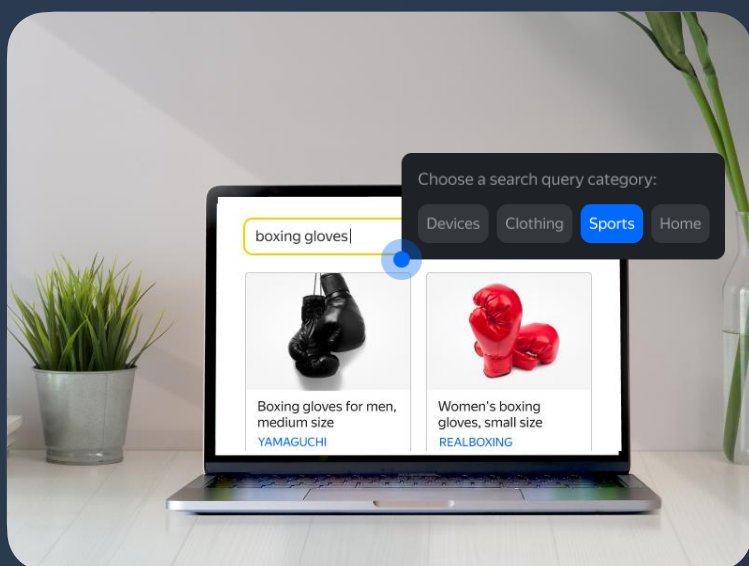
🕒 20 min 💰 \$12



Side-by-side comparison

1000 photos

🕒 10 min 💰 \$4.8



Object segmentation

1000 objects in
100 photos

🕒 5 hrs 💰 \$7.2

Case studies

Powering high-impact AI projects

Trusted by leading ML & AI teams

AliExpress

Improved crowdsourced translations of product descriptions

Results:

17%

budget reduction while achieving optimal quality

[Learn more](#)

Yandex

Enhanced performance of a recommendation engine

Results:

6x

reduction of errors in the product recommender model

[Learn more](#)

bestplace

Improved the accuracy of a predictive tool using local shopping patterns

Results:

30%

improvement in app accuracy after data collection, reaching 95%

[Learn more](#)

Neatsy

Tuned a 3D foot sizing app for better accuracy

Results:

12%

improvement in app accuracy with accelerated time to market

[Learn more](#)



Conclusion

Conclusion

- **AI products and data-driven business decisions require stable and fast inflows of high-quality data**
- **Crowdsourcing is an effective tool for collecting and labeling data used for products and decision-making**
- **Toloka provides the infrastructure, tools and crowdsourcing methodology for data collection and labeling pipelines**