

Efficient Data Annotation for Self-Driving Cars via Crowdsourcing on a Large-Scale

Alexey Drutsa, Denis Rogachevsky, Olga Megorskaya,
Daria Baidakova, Ivan Semchuk



Introduction

Search

Machine translation

Personal assistant

Self-Driving

Maps

Ads

E-commerce

Speech technologies

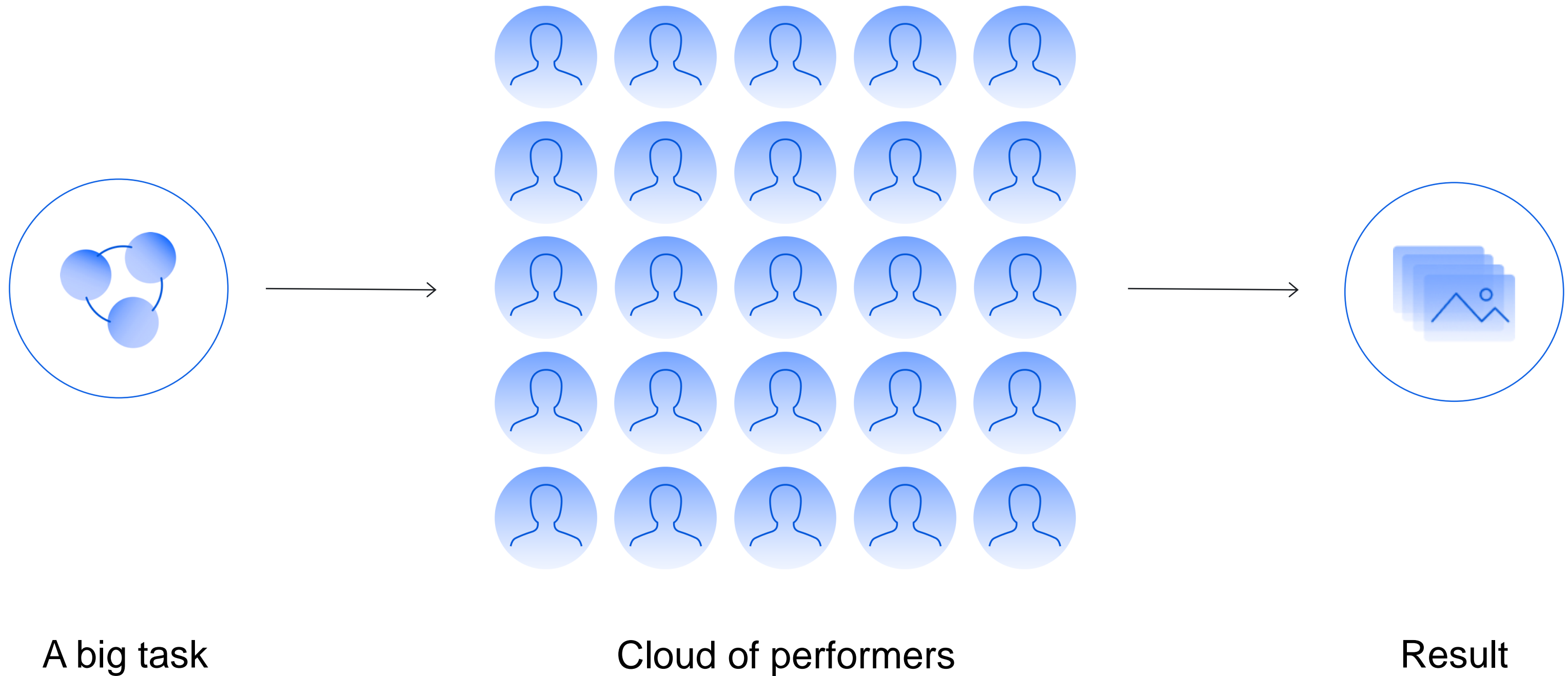
Majority of ML-based solutions require training data labelled by human



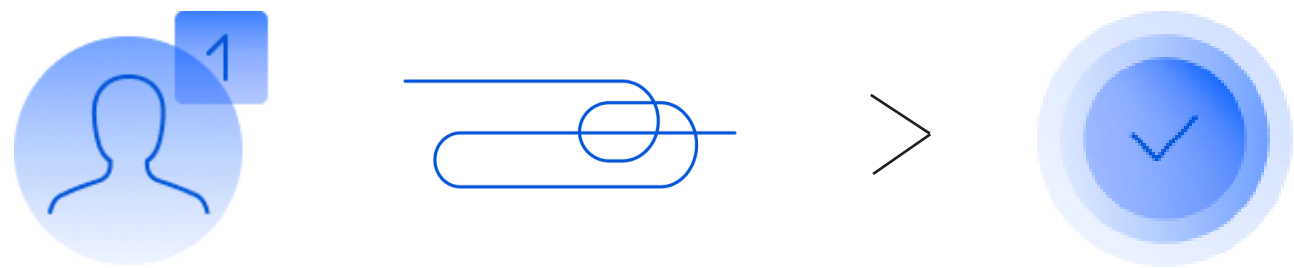
...at a large scale



Crowdsourcing: specific way to design a business process

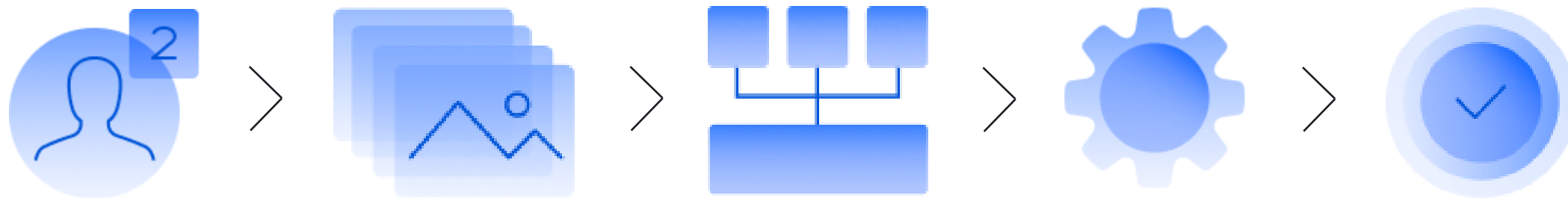


Crowdsourcing: require less from performer, more — from manager



Expert approach: rely on an expertise of a particular performer

- ▶ Expensive
- ▶ Unmeasurable
- ▶ Hard to scale



Crowdsourcing approach

- ▶ Measurable
- ▶ Scalable
- ▶ Manageable

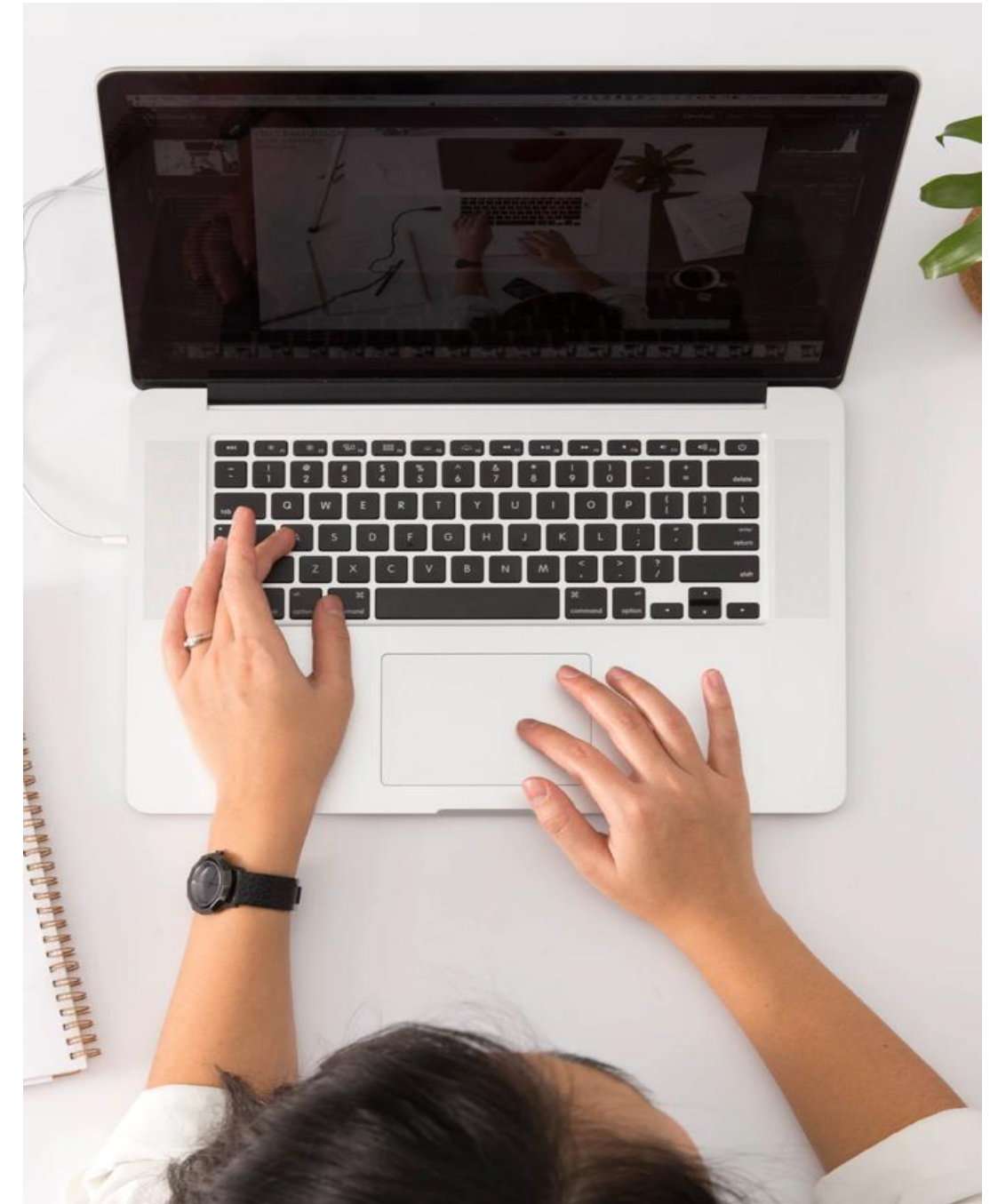
XX century — style management



- ▶ Routine tasks
- ▶ Regular work
- ▶ No ability to choose tasks

It can be different

- ▶ Flexibility to choose from hundreds of tasks
- ▶ No requirements in regularity
- ▶ Switch to another task when bored



Crowdsourcing can provide maximal flexibility to performers if

- ▶ On a platform side, efficient tools for quality management are available for requester
- ▶ **Requester knows how to build smart crowdsourcing pipelines** resistant to single performer's mistakes

Crowdsourcing applications: examples

Task type	Used in
Information assessment	Ranking of search results
Content categorization	Text and media moderation, data cleaning and filtering
Content annotation	Metadata tagging
Pairwise comparison	Offline evaluation, media duplication check
Object segmentation, including 3D	Image recognition for self-driving car
Audio and video transcription	Speech recognition for voice-controlled virtual assistant
Spatial crowdsourcing	Verify business information and office hours

Example: binary classification

Is this cat white?

Yes

No



Example: multi classification



"Real French restaurant"



If you are a gourmand, I can recommend you the "Real French restaurant", located in the historic cellar, with elements of antique design and quite interesting cuisine. The restaurant is small, but very cozy and romantic. The restaurant is very suitable for romance and even for business meetings.

Is it a feedback?

☒ Yes, it is ☐ No, it's other comment

☐ Personal information ?

☐ Swearing, vulgarity, insults, aggressive statements ?

☐ Spam, advertisingspan ?

Example: multi classification with ordered labels

Query: Machine learning
URL: https://en.wikipedia.org/wiki/Machine_learning

Open the original Yandex Google

1 ☐ Vital
2 ☐ Useful
3 ☐ Relevant+
4 ☐ Relevant-
5 ☐ Irrelevant
6 ☐ Not displayed

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal

en.wikipedia.org Machine learning - Wikipedia

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Machine learning

From Wikipedia, the free encyclopedia



*For the journal, see [Machine Learning \(journal\)](#).
"Statistical learning" redirects here. For statistical learning in linguistics, see [statistical learning in language acquisition](#).*

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of [artificial intelligence](#). Machine learning algorithms build a [mathematical model](#) based on sample data, known as "[training data](#)", in order to

Machine learning and data mining



Examples: pairwise comparison



How to Make Perfect Pancakes

Food Network Magazine shows you how to make the best short stack, plus some tasty toppings.

Keep in mind: Price and stock could change after publish date, and we may make money from these links.


April 24, 2015

From: **Food Network Magazine**

Search for more recipes

How to make pancakes

★★★★☆ 17 ratings



Preparation time
less than 30 mins

Cooking time
less than 10 mins

Serves
Serves 4

Dietary
V

Query: how to make pancakes
Which one do you like better?

☐ Left ☐ Right

Please, comment your choice

Continue

Examples: transcription with textual answers

▶ 0:00 / 0:09

🔊

⋮

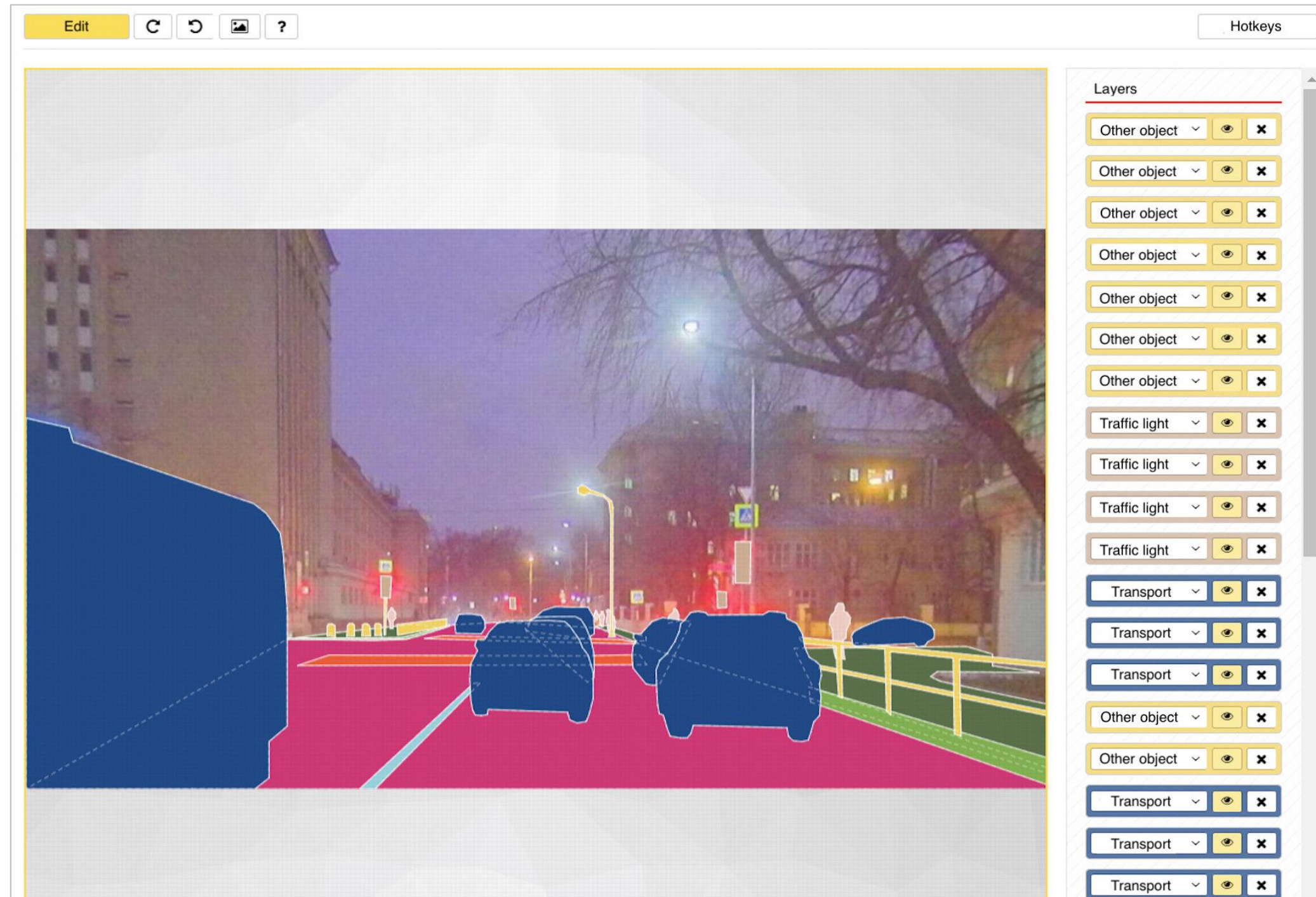
1 ☐ There is a speech on the record

2 ☐ No speech or inaudible

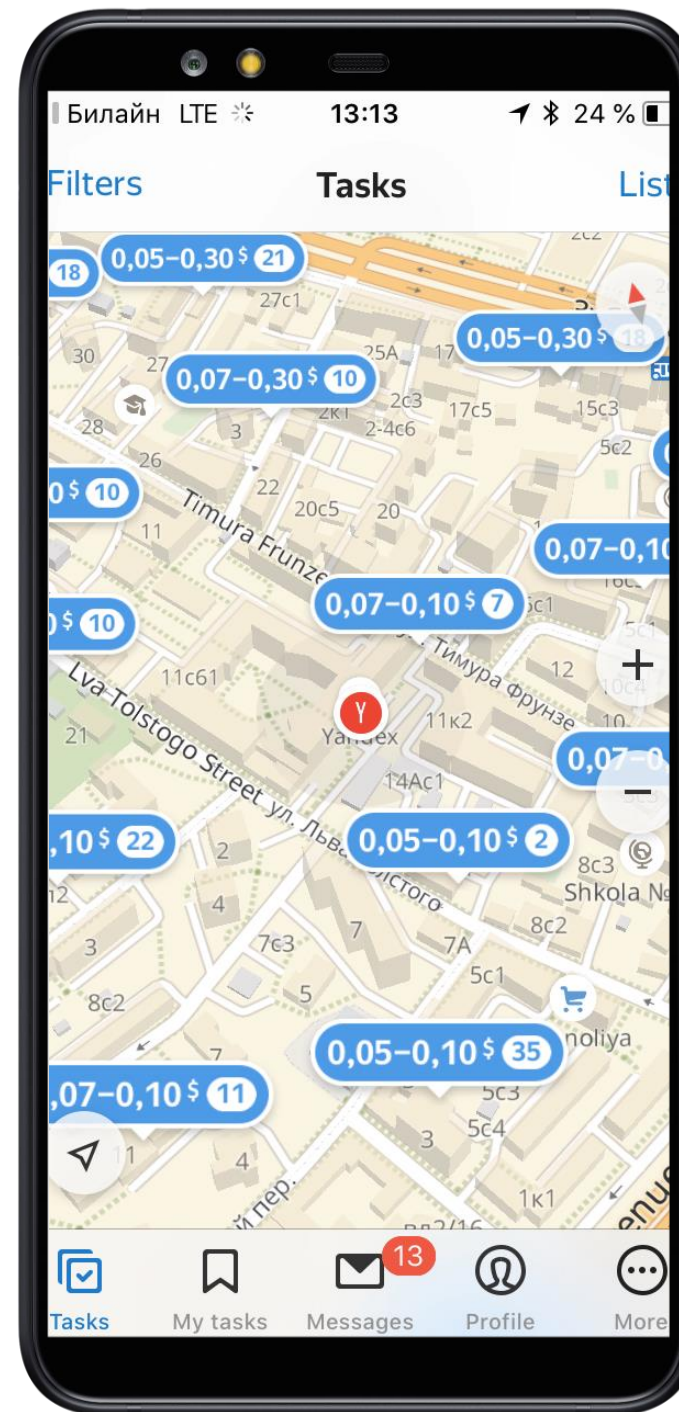
Annotation

4

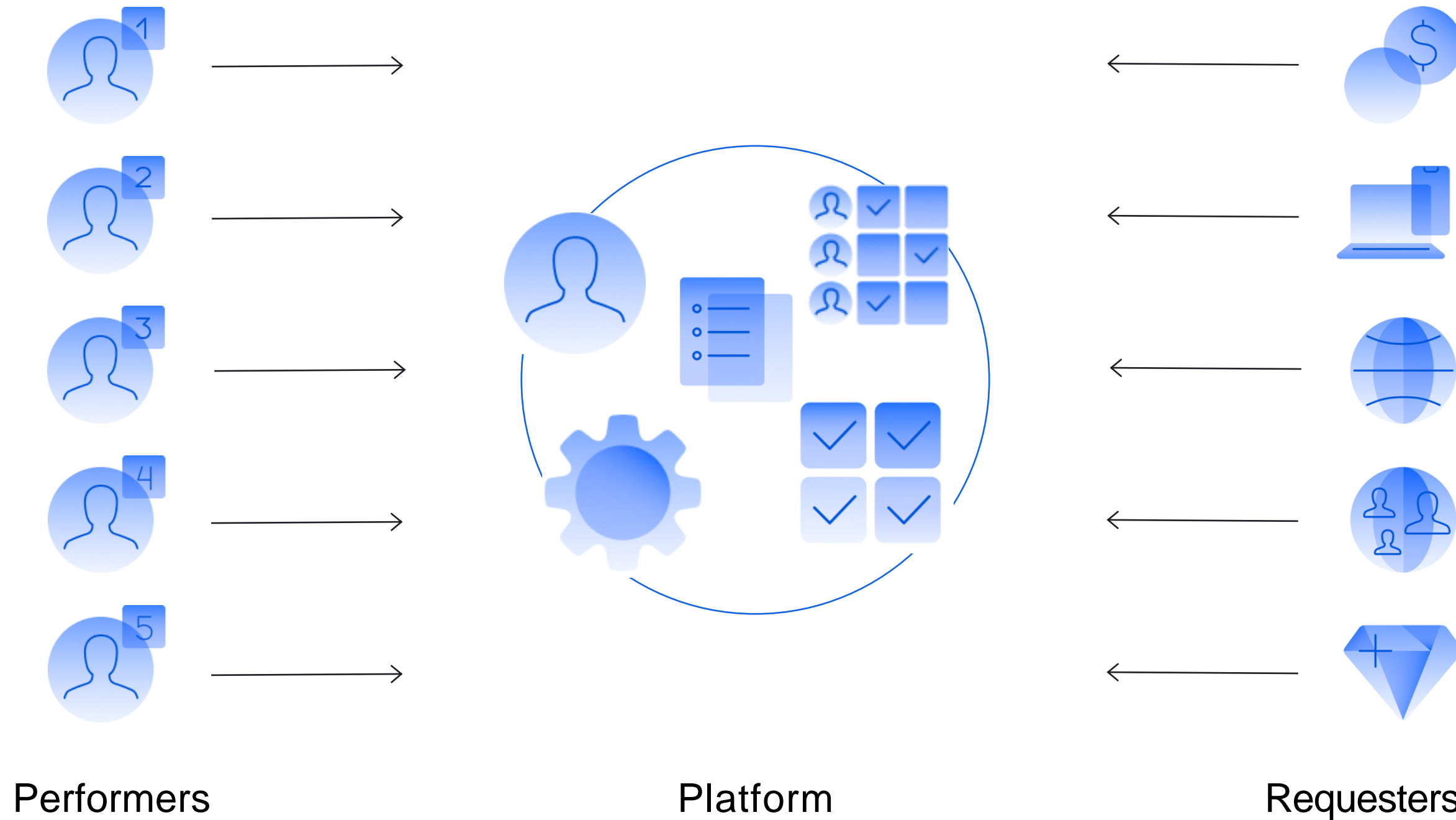
Examples: object segmentation



Examples: spatial crowdsourcing



A crowdsourcing platform: two-sided market



Crowdsourcing platforms: examples

- ▶ Amazon Mechanical Turk
- ▶ Toloka
- ▶ Microworkers
- ▶ Gigwalk
- ▶ ClickWorker
- ▶ CloudFactory
- ▶ CrowdSource
- ▶ DefinedCrowd
- ▶ ...

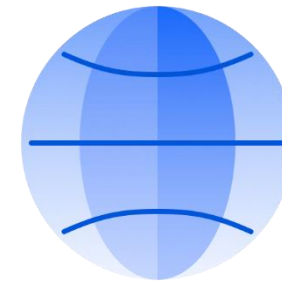
Pros of crowdsourcing platforms



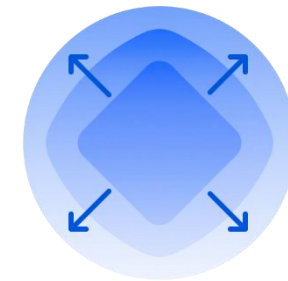
24/7



Variety
of skilled
performers



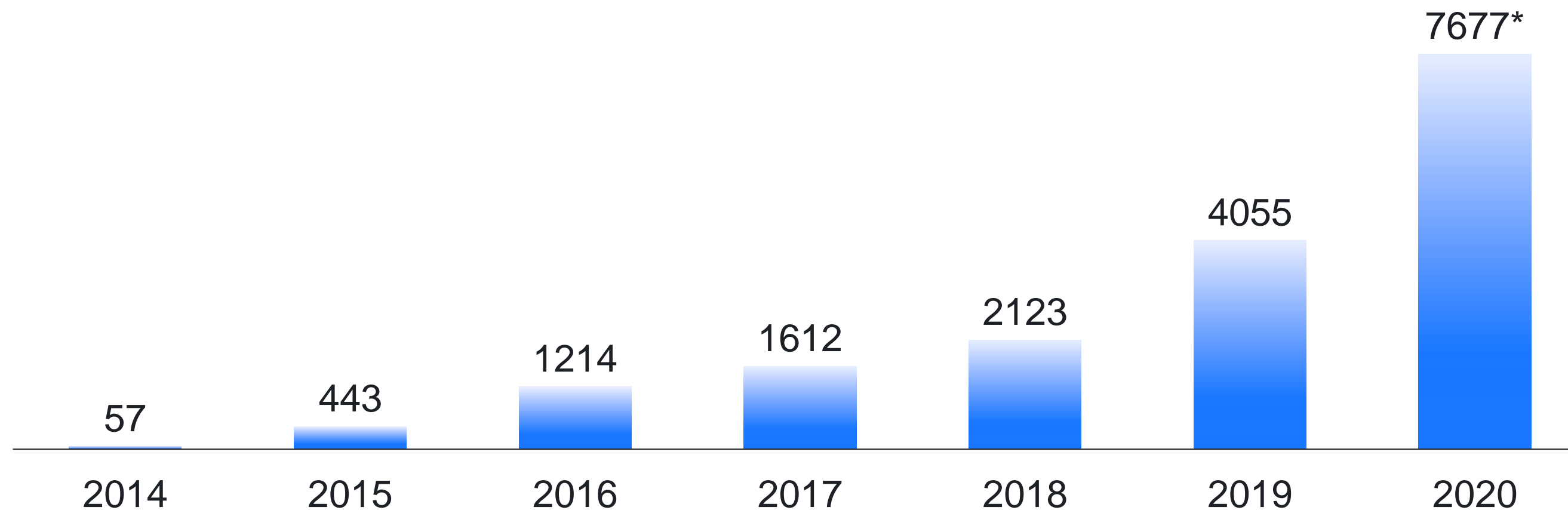
Vast
region
coverage



Ongoing
processes

Crowdsourcing growth: our experience

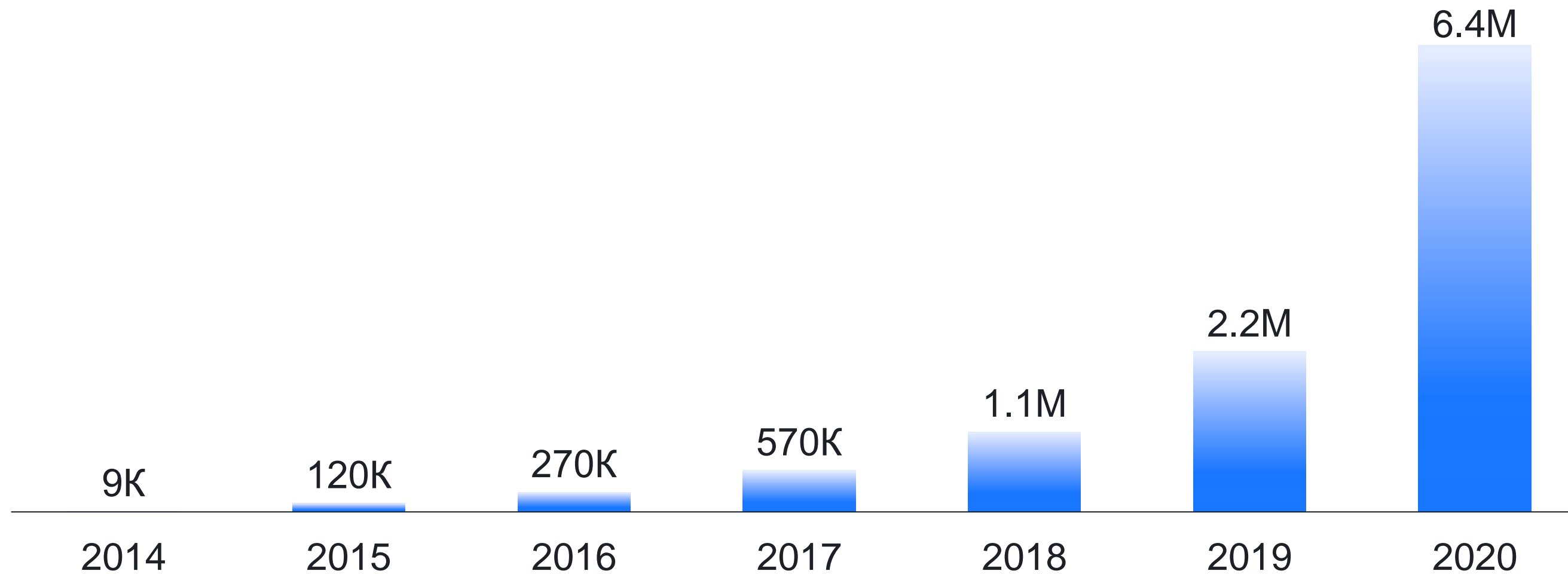
Different projects in Toloka



* An extrapolation based on the first 3 months of 2020

Crowdsourcing growth: our experience

Active performers in Toloka



* An extrapolation based on the first 3 months of 2020

Everyday on Toloka



500+
different
projects



37K+
performers



13M+
tasks

Toloka: real-life cases

Case	Tasks	Done in	Cost
Side-by-side object comparison	1,000 tasks	10 min	\$2.4
Object classification	1,000 photos	15 min	\$1.2
Object segmentation	About 1,000 objects in 100 photos	6 h	\$3.6
Phrase generation for a chatbot	500 phrases for the same topic	15 min	\$1
Audio transcription	100 recordings 25 minute long	20 min	\$6
Video ranking	10,000 videos	2 h	\$10

Tutorial overview

Why this tutorial?

Practice

Tutorial schedule

