Toloka

KDD’2019 hands-on tutorial

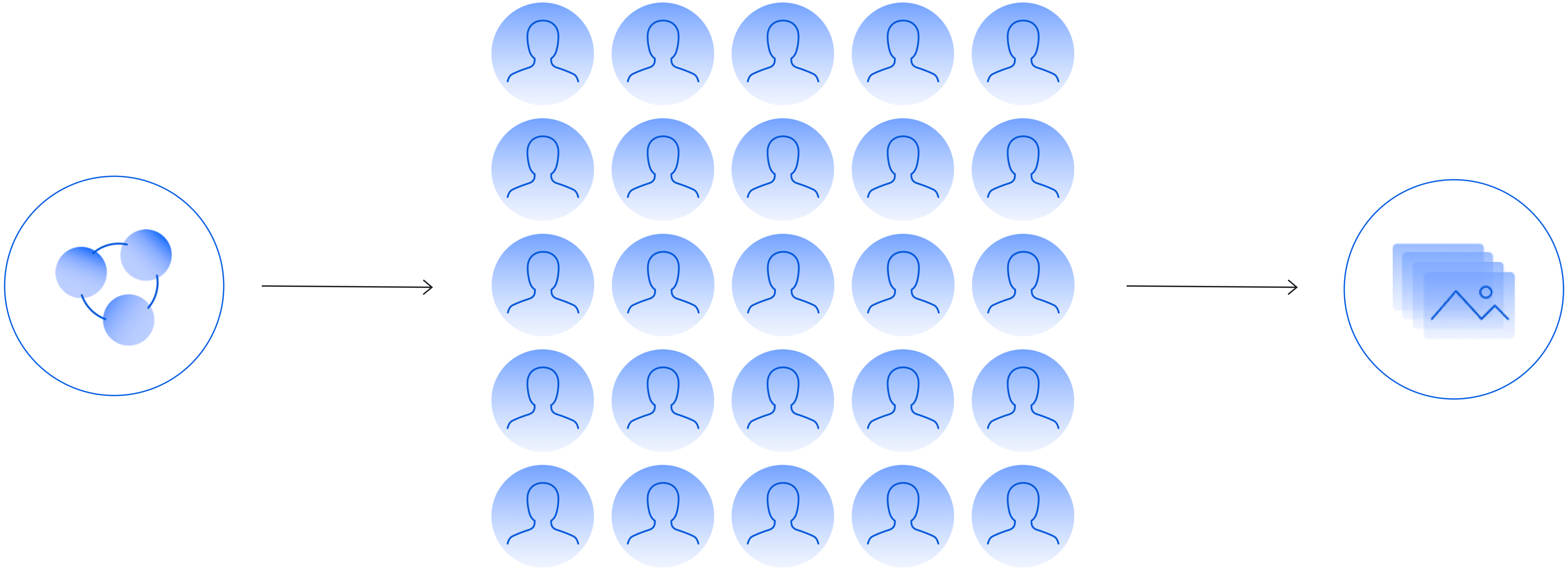
Practice of Efficient Data Collection via Crowdsourcing at Large-Scale

Alexey Drutsa, Valentina Fedorova, Olga Megorskaya,
Evfrosiniya Zerminova

Introduction

Alexey Drutsa,
Head of Efficiency and Growth Division, Toloka

Crowdsourcing: specific way to design a business process



A big task

Cloud of performers

Result

Crowdsourcing applications: examples

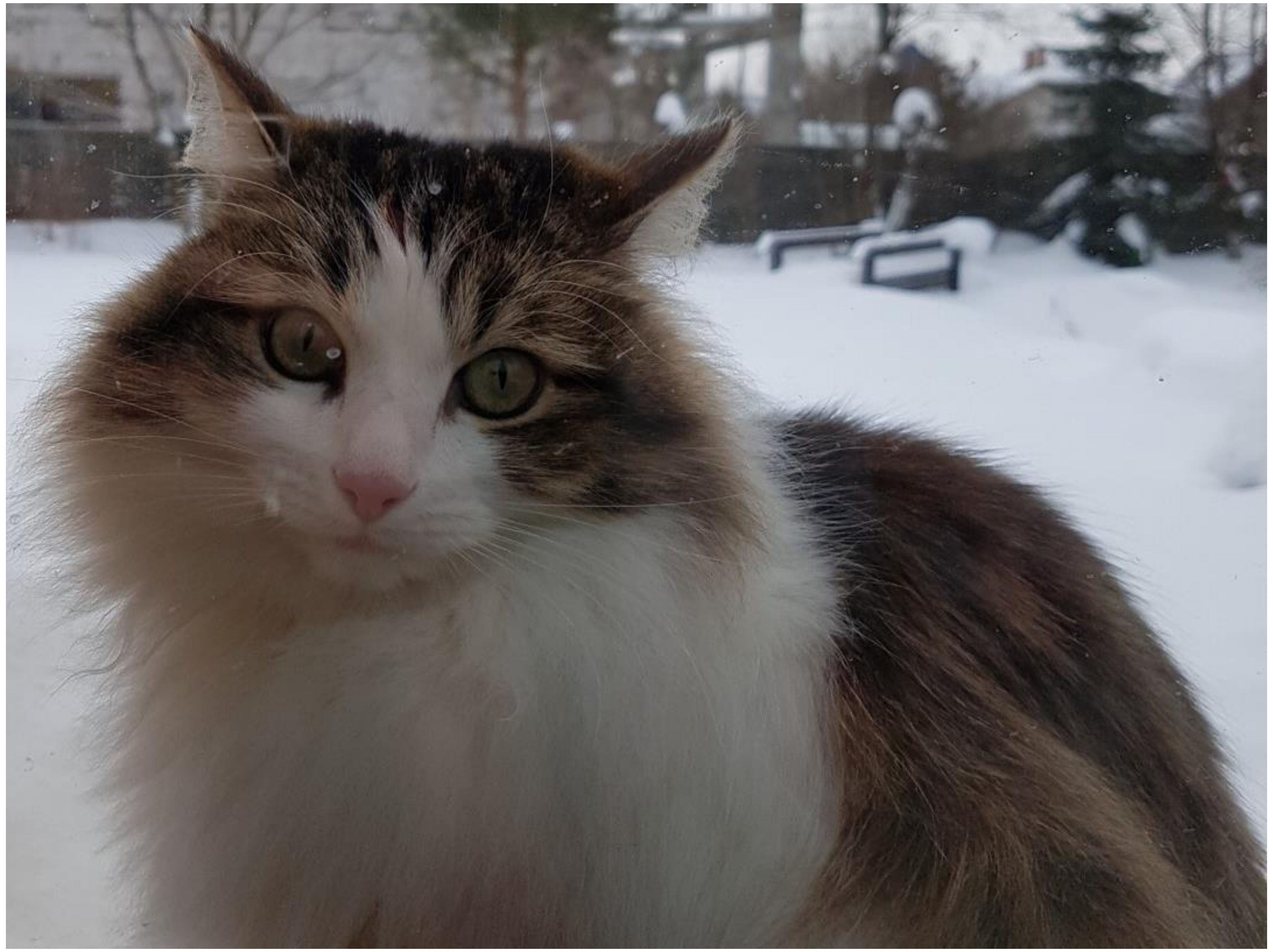
Task type	Where is used
Information assessment	Ranking of search results
Content categorization	Text and media moderation, data cleaning and filtering
Content annotation	Metadata tagging
Pairwise comparison	Offline evaluation, media duplication check
Object segmentation, including 3D	Image recognition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

Example: binary classification

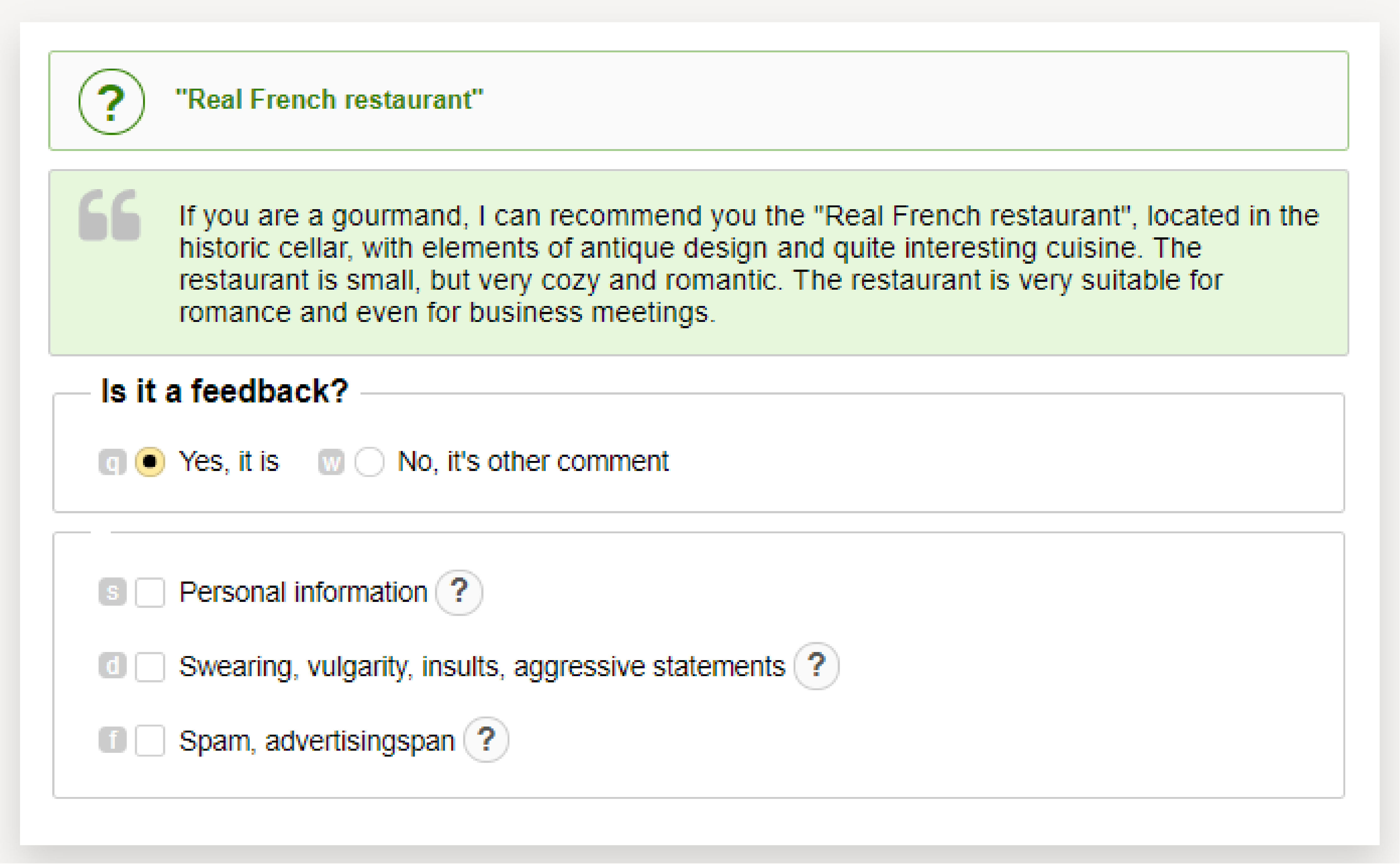
Is this cat white?

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?

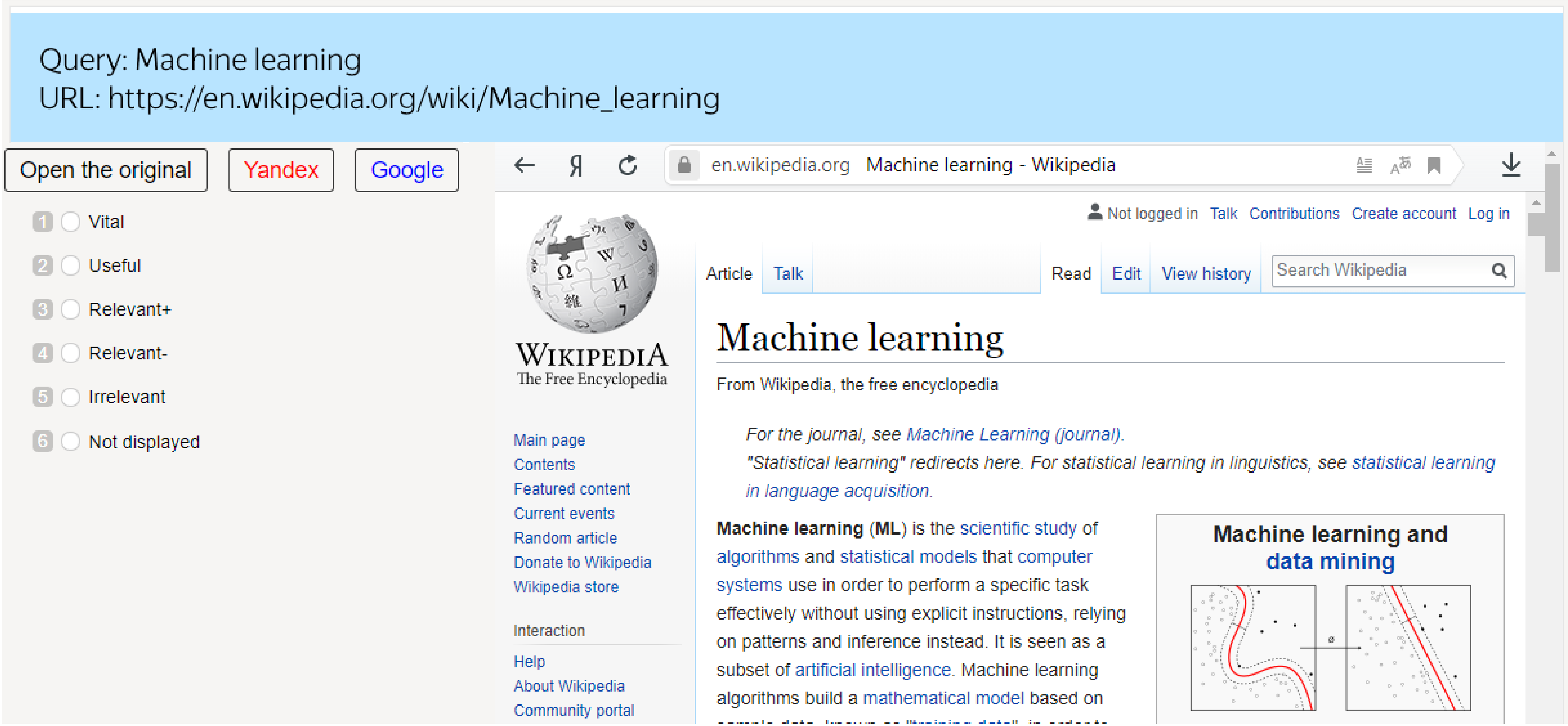
q ● Yes, it is   w ○ No, it's other comment

s □ Personal information ?

d □ Swearing, vulgarity, insults, aggressive statements ?

f □ Spam, advertisingspan ?

Example: multi classification with ordered labels



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

1 ○ Vital

2 ○ Useful

3 ○ Relevant+

4 ○ Relevant-

5 ○ Irrelevant

6 ○ Not displayed

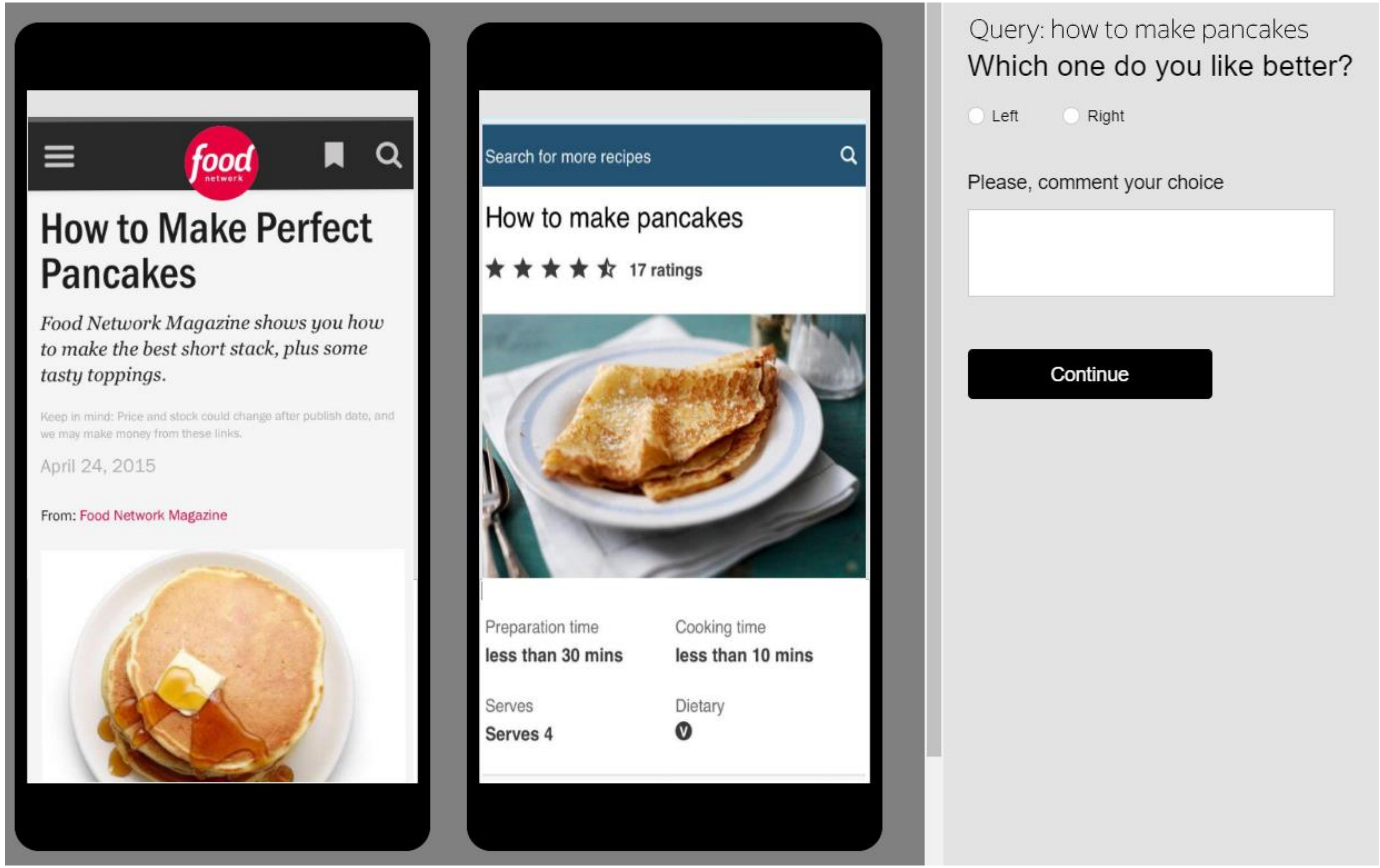
Machine learning

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

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

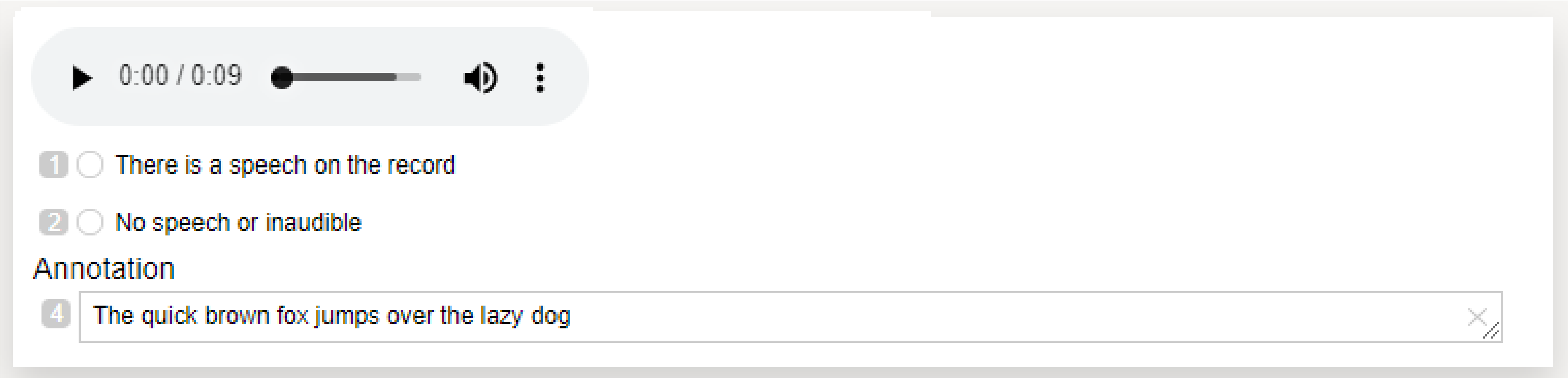
Which one do you like better?

○ Left ○ Right

Please, comment your choice

Continue

Examples: transcription with textual answers



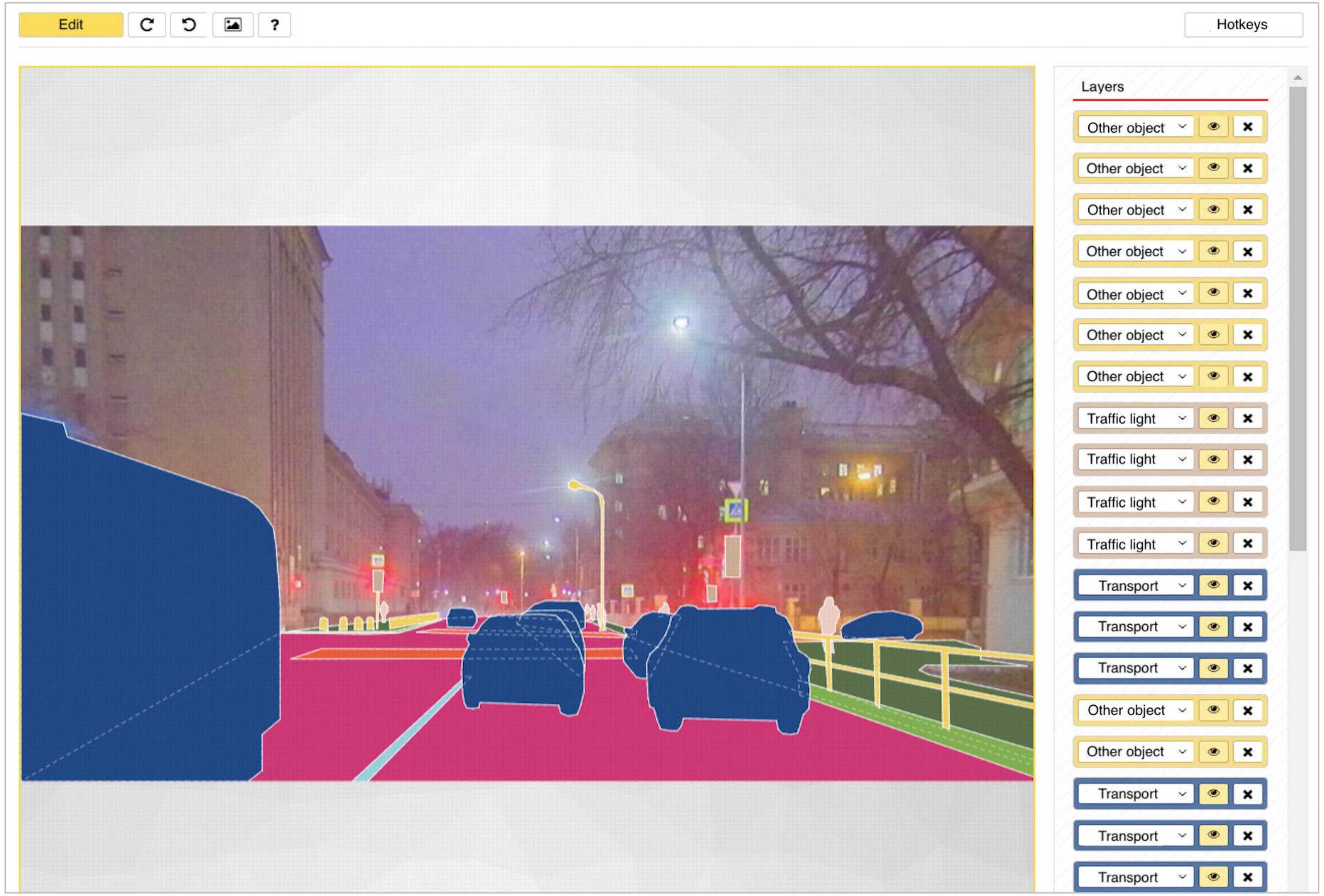
1 ○ There is a speech on the record

2 ○ No speech or inaudible

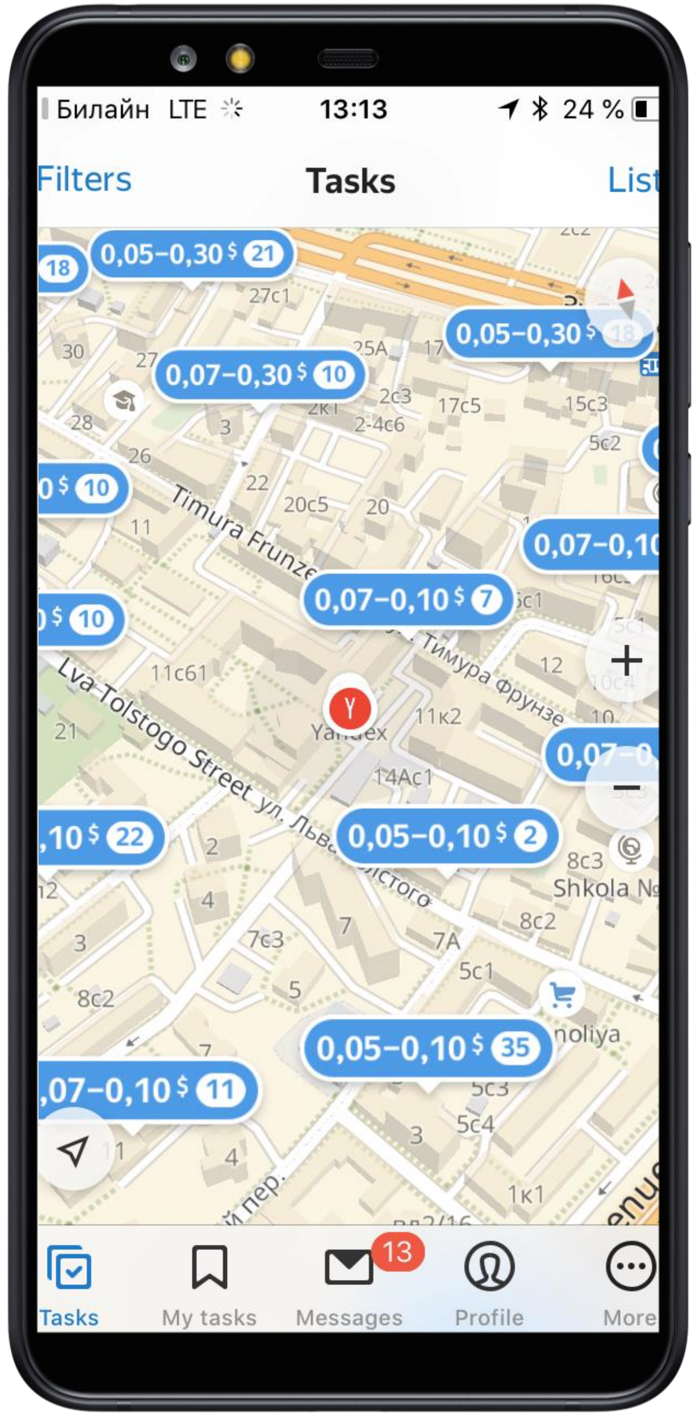
Annotation

4 The quick brown fox jumps over the lazy dog

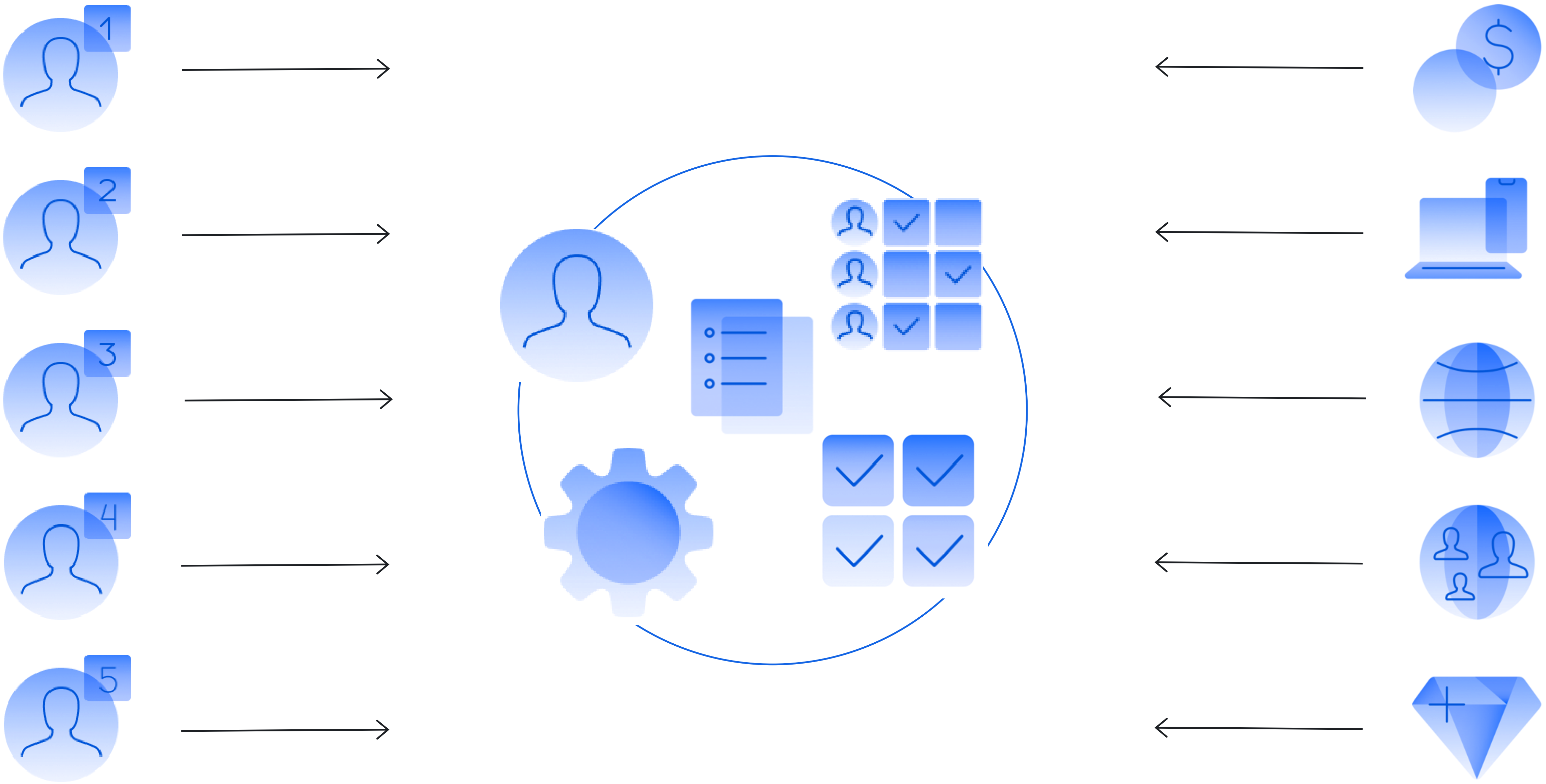
Examples: object segmentation



Examples: field surveys



A crowdsourcing platform: two-sided market



Performers

Platform

Requesters

Crowdsourcing platforms: examples

- Amazon Mechanical Turk
- Toloka
- Microworkers
- Gigwalk
- ClickWorker
- CloudFactory
- Figure Eight
- CrowdSource
- DefinedCrowd
- …

Pros of crowdsourcing platforms

24/7

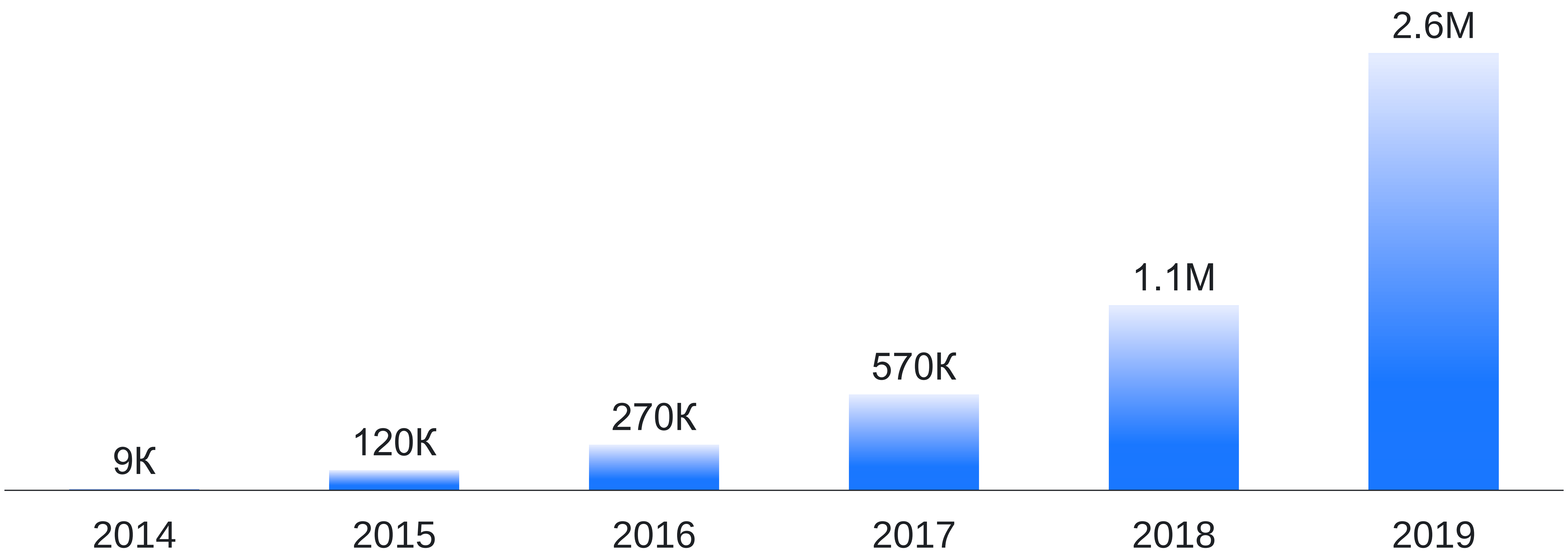
Variety of skilled performers

Vast region coverage

Ongoing processes

Crowdsourcing growth: our experience

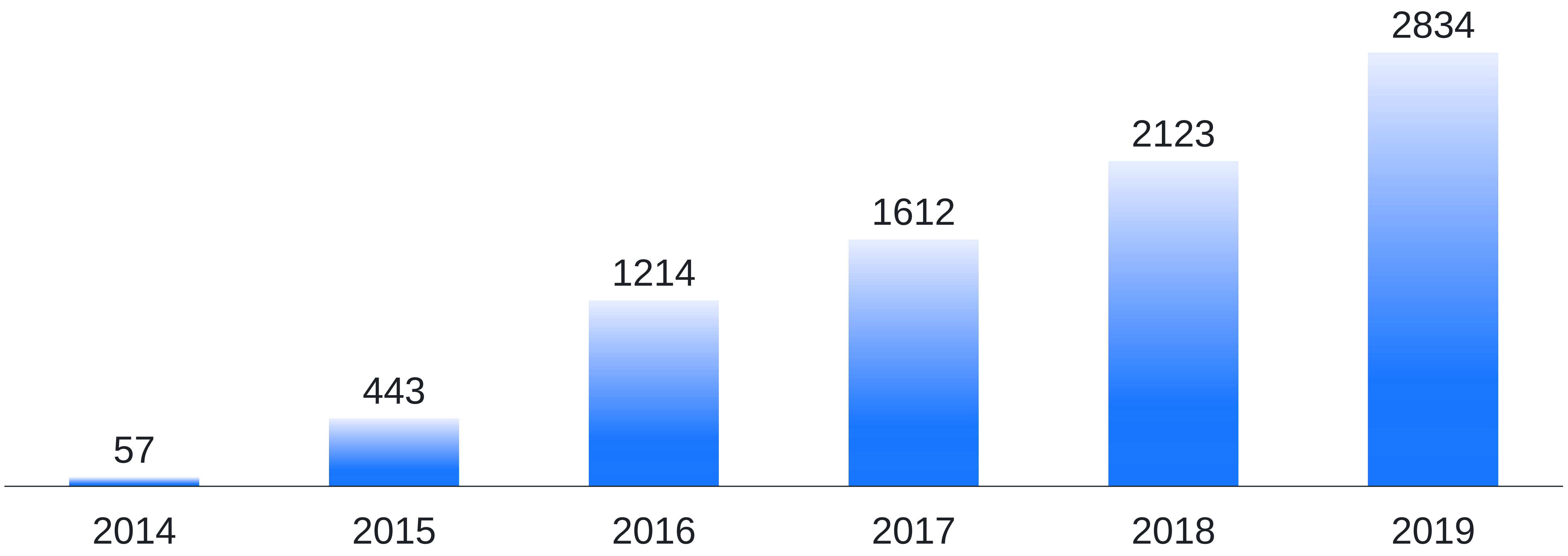
Active performers in Toloka



* An extrapolation based on the first 7 months of 2019

Crowdsourcing growth: our experience

Different projects in Toloka



* An extrapolation based on the first 7 months of 2019

Everyday on Toloka

500+
different
projects

25K
performers

6M+
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

Part I: 20 min

Main components of data collection via crowdsourcing

- Decomposition for effective pipeline
- Task instruction & interface: best practices
- Quality control techniques

Alexey Drutsa

Head of Efficiency and Growth Division
Crowdsourcing Department, Toloka

Part II: 25 min

Analysis of label collection projects to be done (practical session)

- Dataset and required labels
- Discussion: how to collect labels?
- Data labelling pipeline for implementation

Olga Megorskaya

CEO, Toloka

Part III: 10 min

Introduction to the crowdsourcing platform Yandex.Toloka for requesters

- Main types of instances
- Project: creation & configuration
- Pool: creation & configuration
- Tasks: uploading & golden set creation
- Statistics in flight and download of results

Evfrosiniya Zerminova

Head of Data Analysis and Research Group, Toloka

Part IV: 70 min

Setting up and running label collection projects (practical session)

Olga Megorskaya

CEO, Toloka

You

- ► Create
- ► Configure
- ► Run on real performers

Data labelling projects in real-time

Part V: 25 min

Theory on efficient aggregation, incremental relabelling, and pricing

- Aggregation models
- Incremental relabelling to save money
- Performance-based pricing

Valentina Fedorova

Researcher, Toloka

Part VI: 15 min

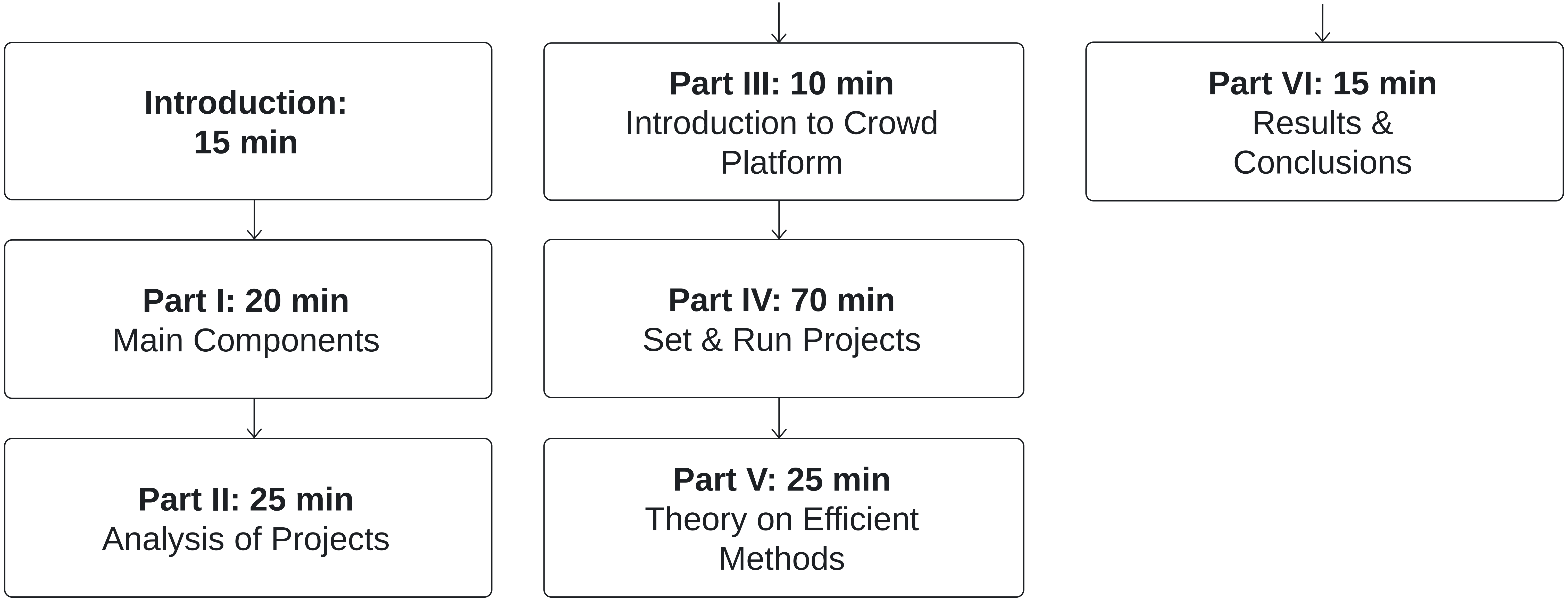
Discussion of results from the projects & conclusions

- Results of your projects
- Extensions to work on after tutorial

Olga Megorskaya

CEO, Toloka

Tutorial outline



Introduction:
15 min

Part I: 20 min
Main Components

Part II: 25 min
Analysis of Projects

Part III: 10 min
Introduction to Crowd Platform

Part IV: 70 min
Set & Run Projects

Part V: 25 min
Theory on Efficient Methods

Part VI: 15 min
Results & Conclusions