Yandex

Yandex

WSDM 2020 hands-on tutorial

Practice of Efficient Data Collection via Crowdsourcing: Aggregation, Incremental Relabelling, and Pricing

Alexey Drutsa, Valentina Fedorova, Dmitry Ustalov, Olga Megorskaya, Evfrosiniya Zerminova, Daria Baidakova

Introduction

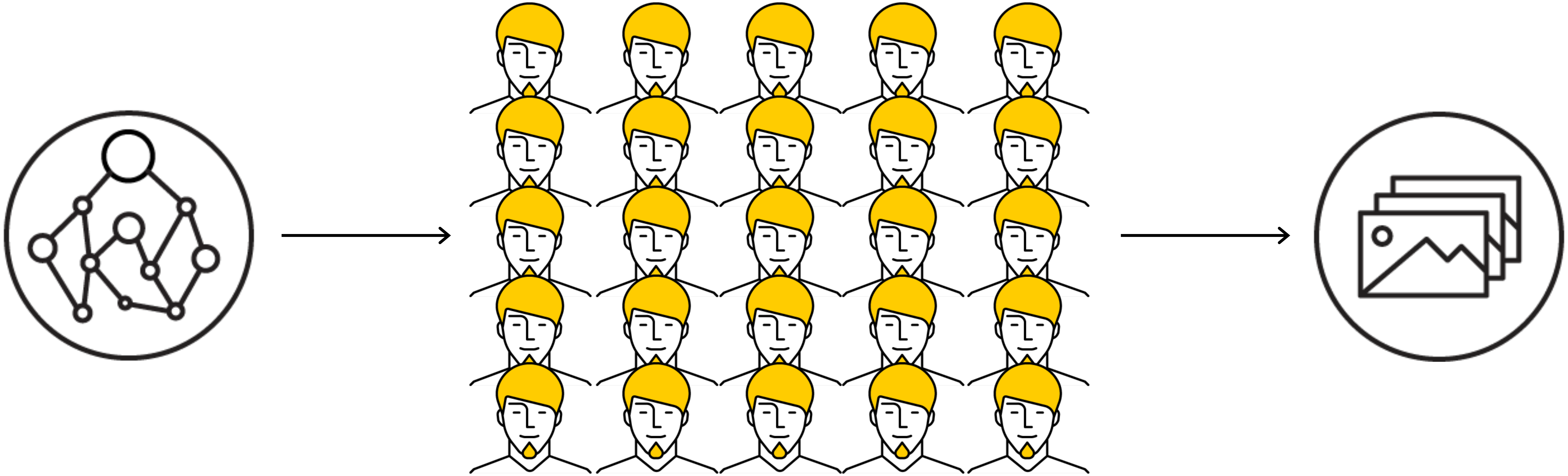
Olga Megorskaya,
Head of Crowdsourcing Department, Yandex

Yandex.Toloka is a service of Swiss company Yandex Services AG

Facebook chat for materials
wsdm2020crowd

Crowdsourcing

Specific way to design a business process



A big task

Cloud of performers

Result

Crowdsourcing applications: examples

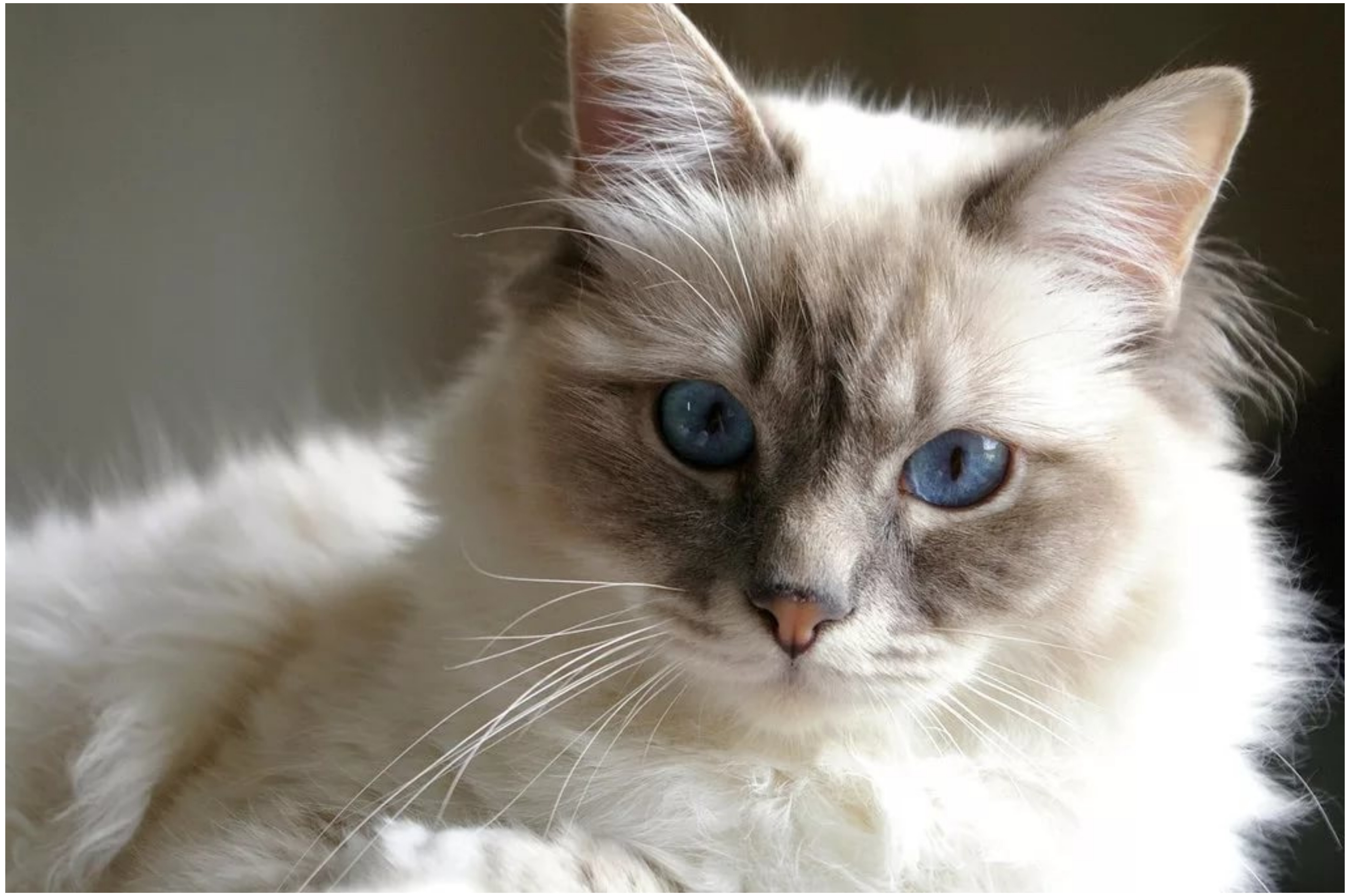
Task type	Where is used
Information assessment	Ranking of search results
Content categorization	Text and media moderation, data cleaning and filtering
Content annotation	Metadata tagging
Pairwise comparison	Offline evaluation, media duplication check
Object segmentation, including 3D	Image recognition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

Example: binary classification

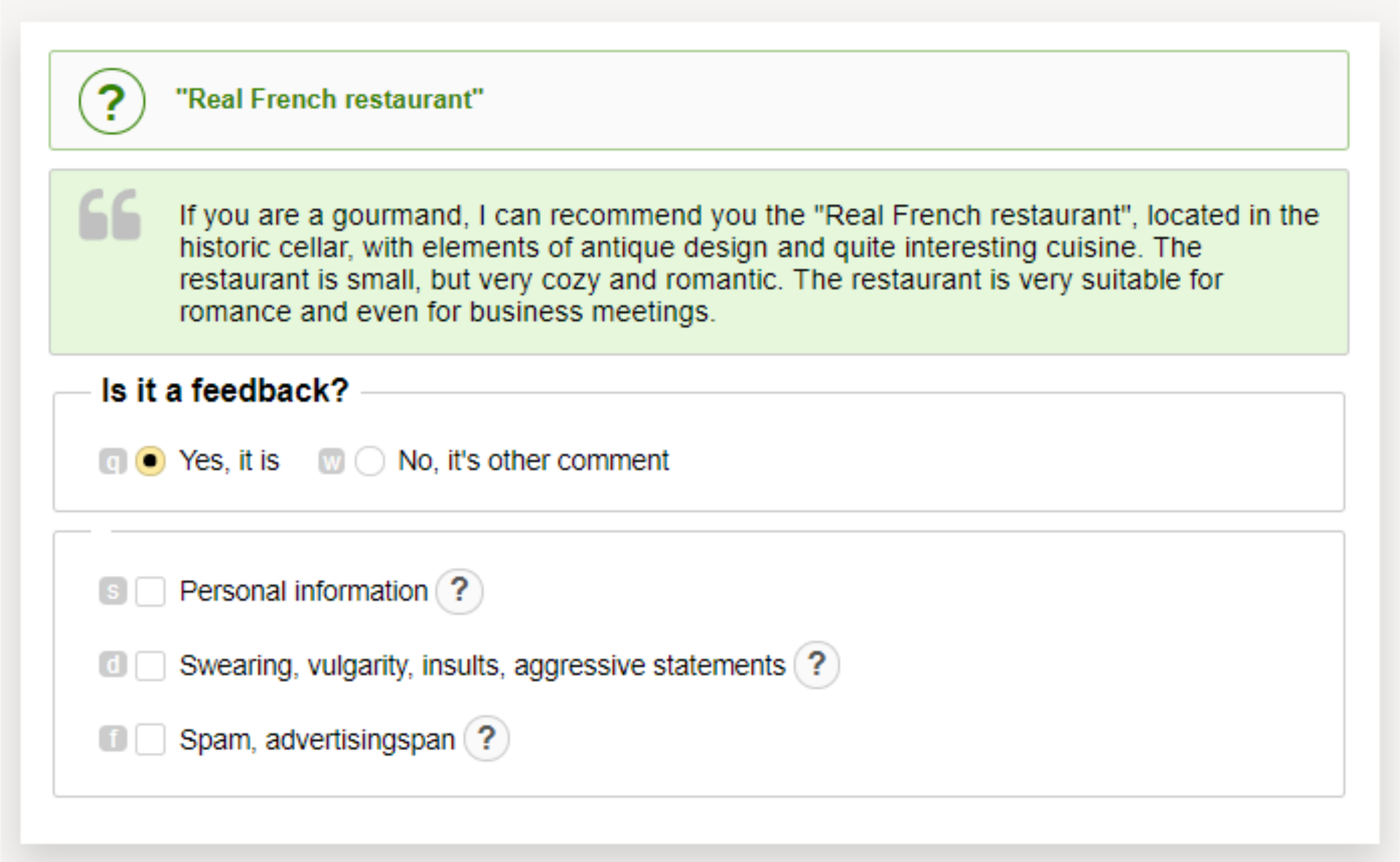
Is this cat white?

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?

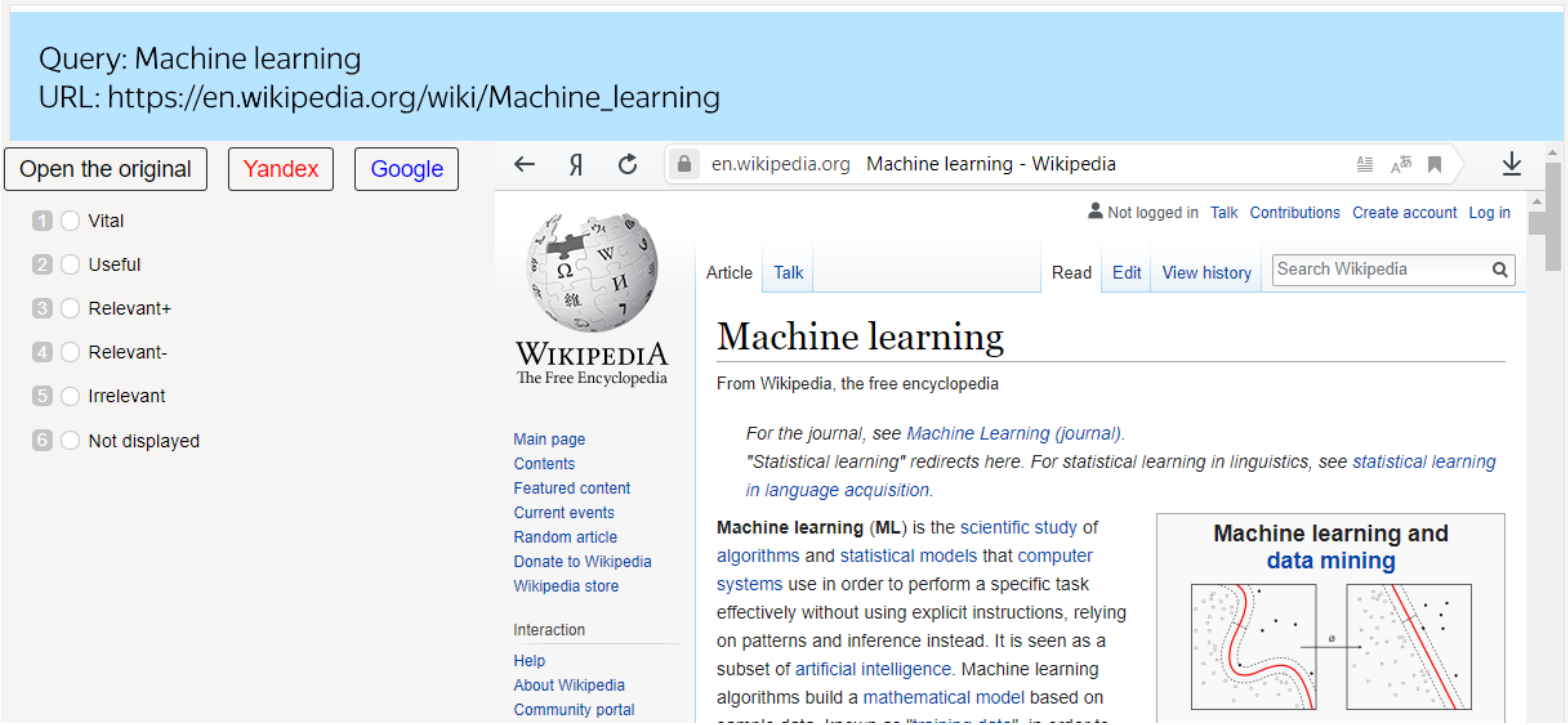
q ● Yes, it is   w ○ No, it's other comment

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

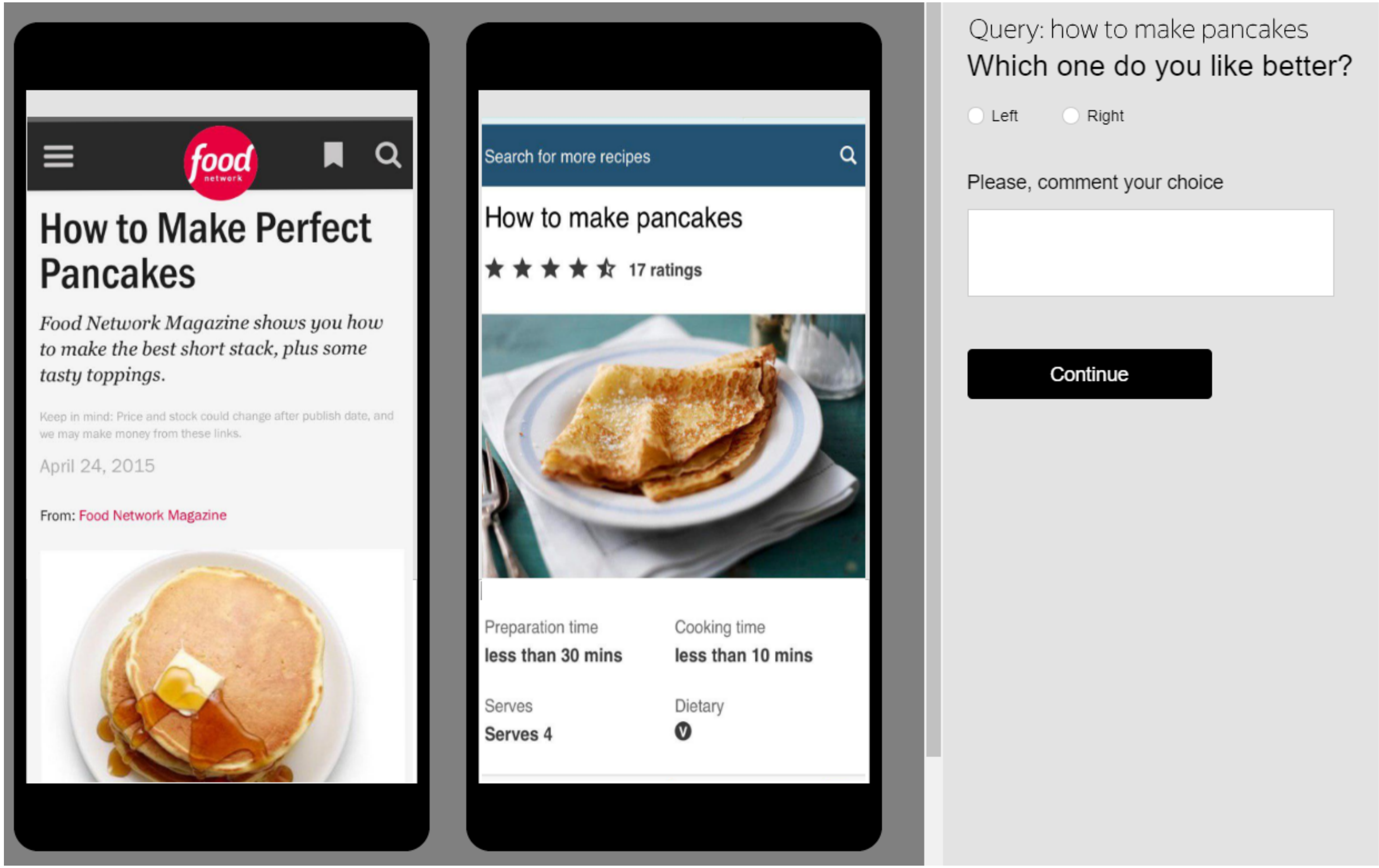
From Wikipedia, the free encyclopedia

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

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

How to make pancakes

★★★★⯪ 17 ratings

Preparation time less than 30 mins	Cooking time less than 10 mins
Serves Serves 4	Dietary V

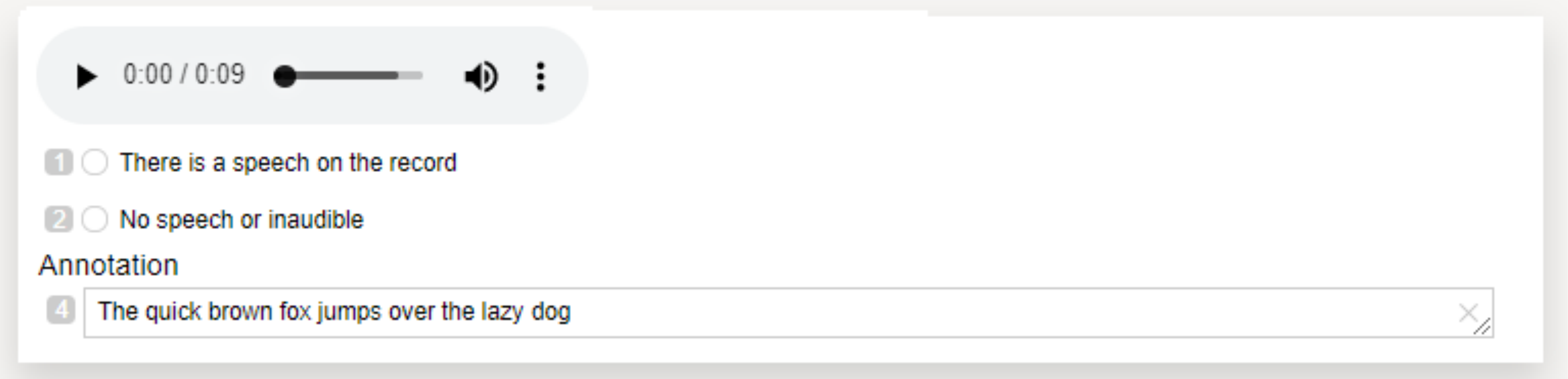
Query: how to make pancakes

Which one do you like better?

○ Left ○ Right

Please, comment your choice

Examples: transcription with textual answers



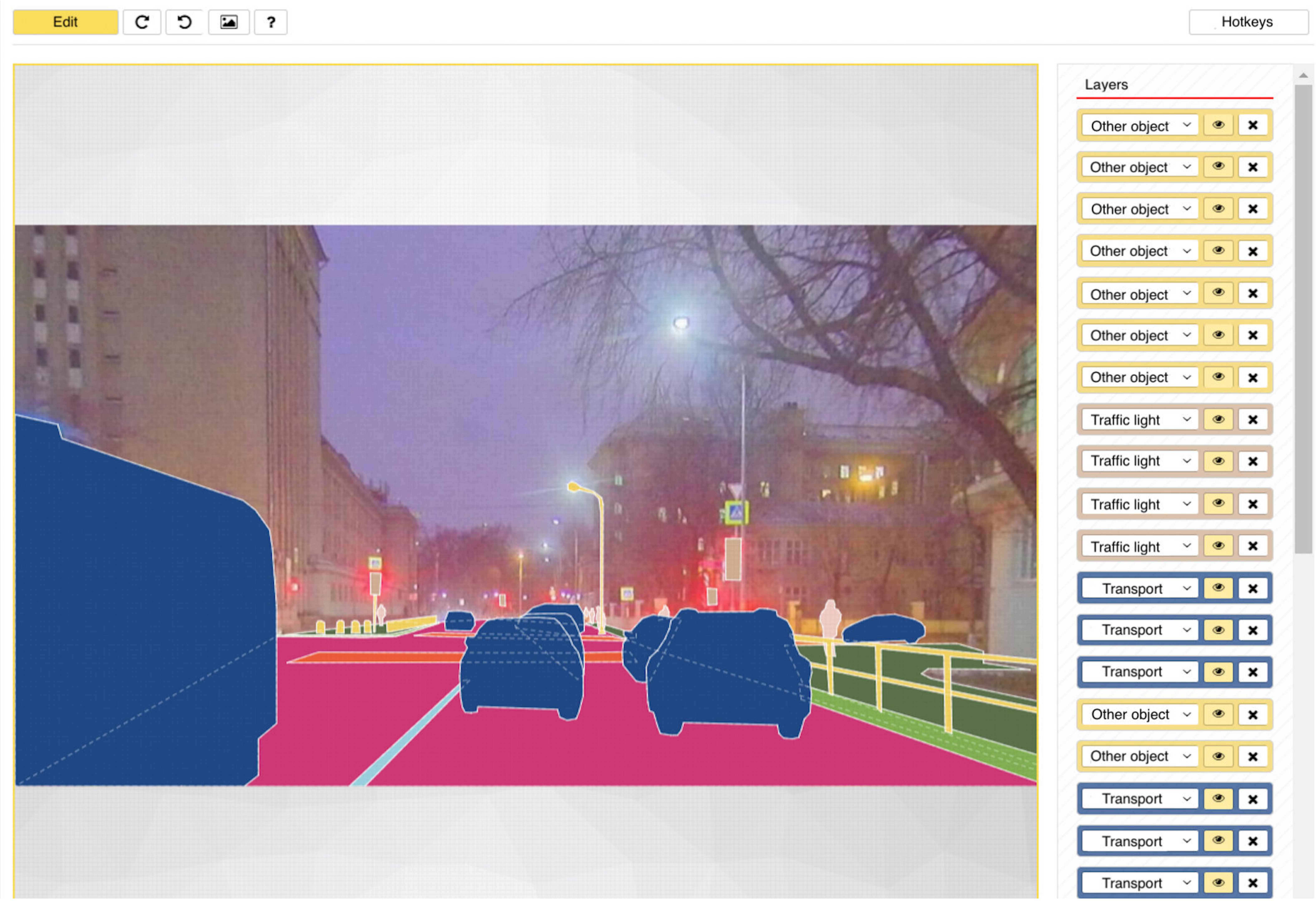
1 ○ There is a speech on the record

2 ○ No speech or inaudible

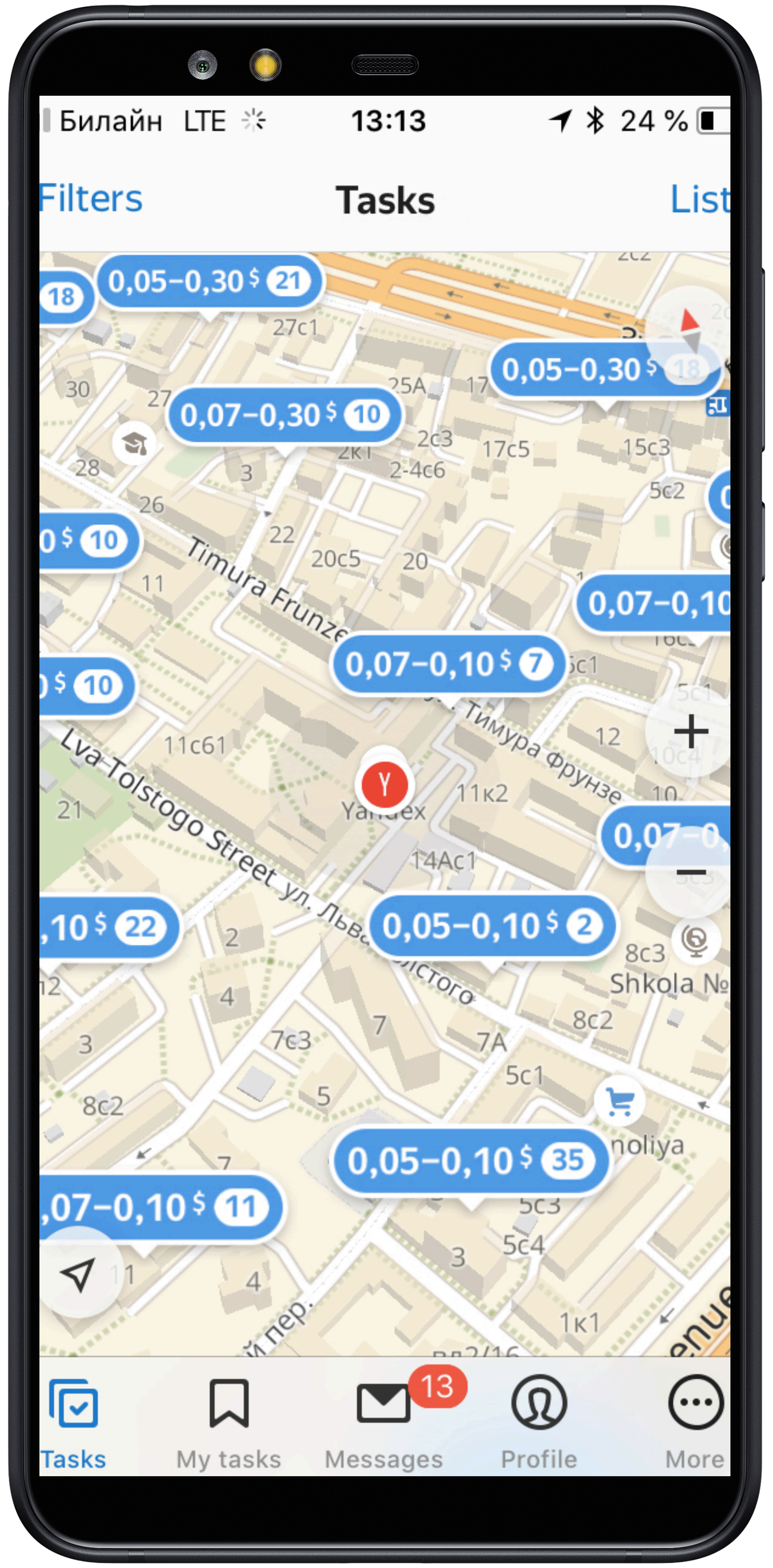
Annotation

4 The quick brown fox jumps over the lazy dog

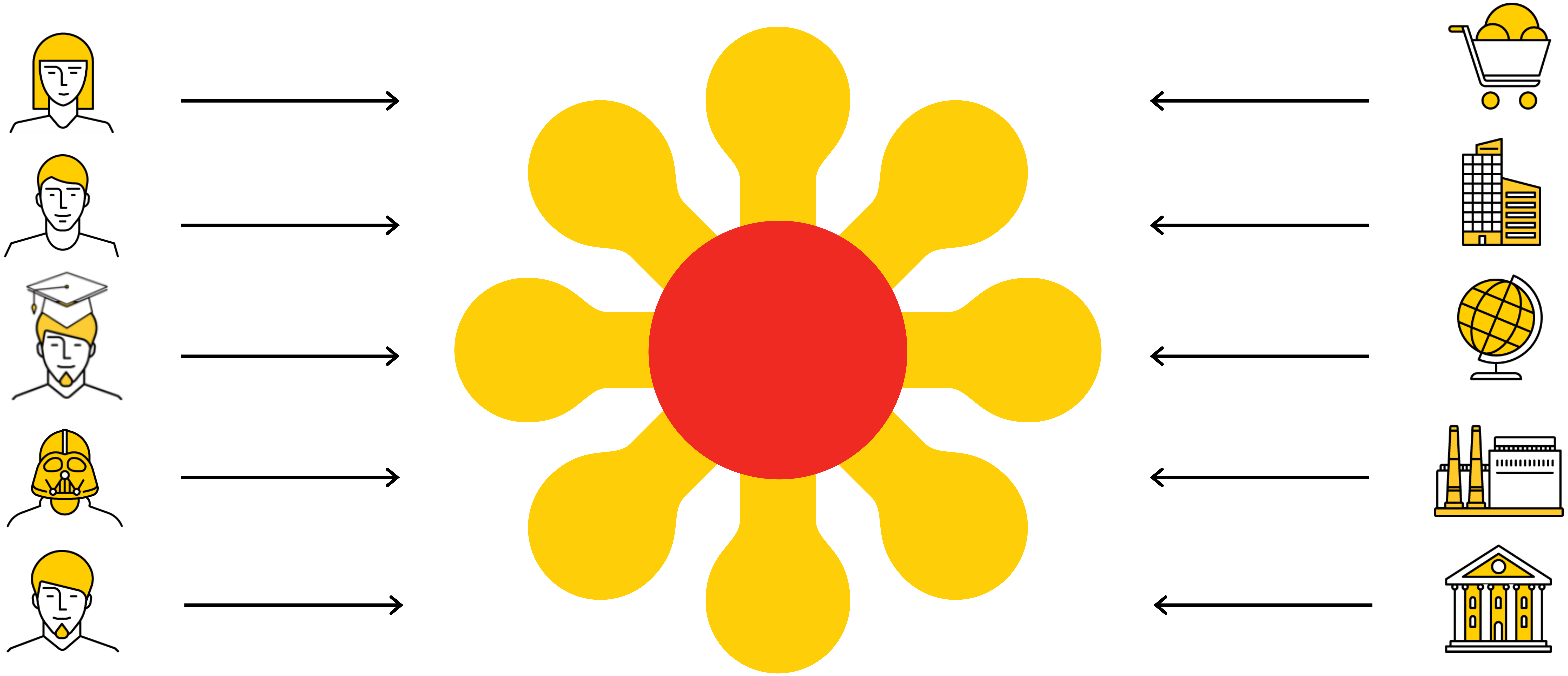
Examples: object segmentation



Examples: spatial crowdsourcing



A crowdsourcing platform: two-sided market



Performers

Requesters

Crowdsourcing platforms: examples

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

Pros of crowdsourcing platforms

24/7

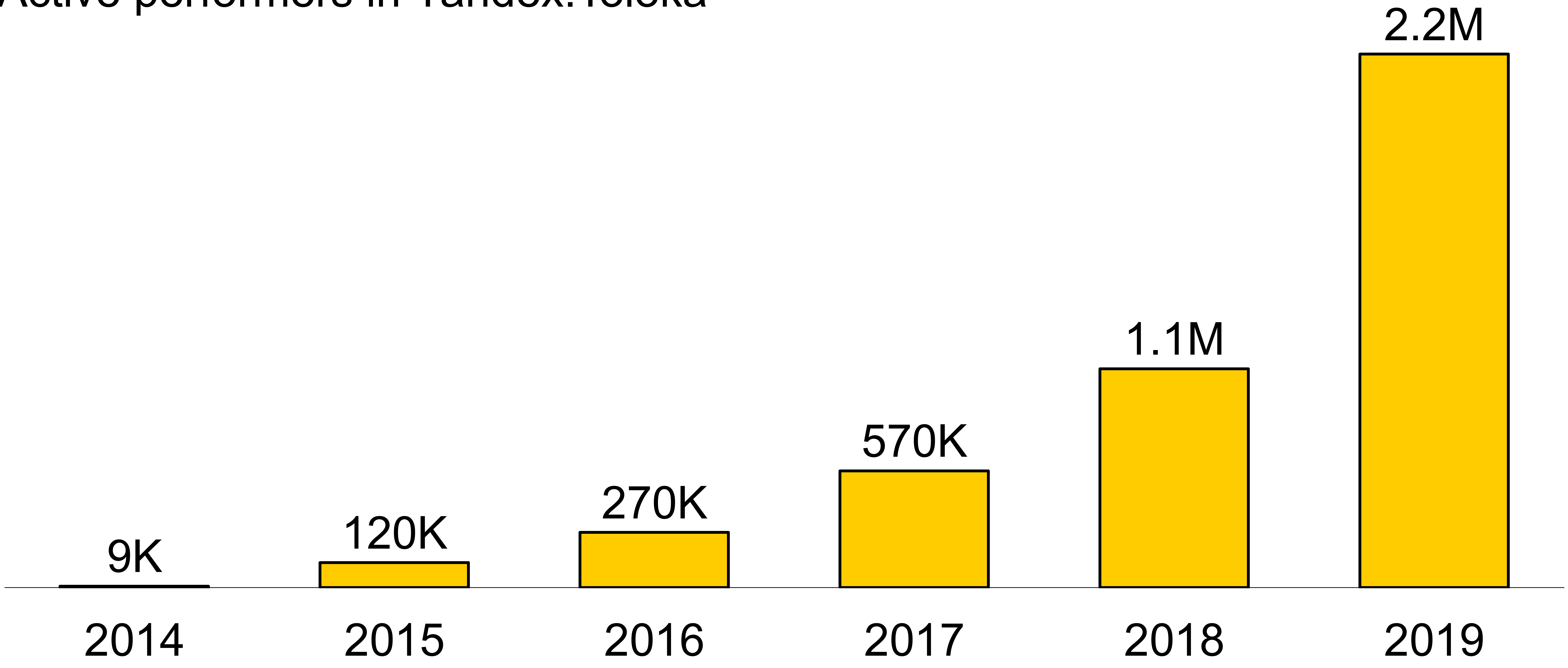
Variety of skilled performers

Vast region coverage

Ongoing processes

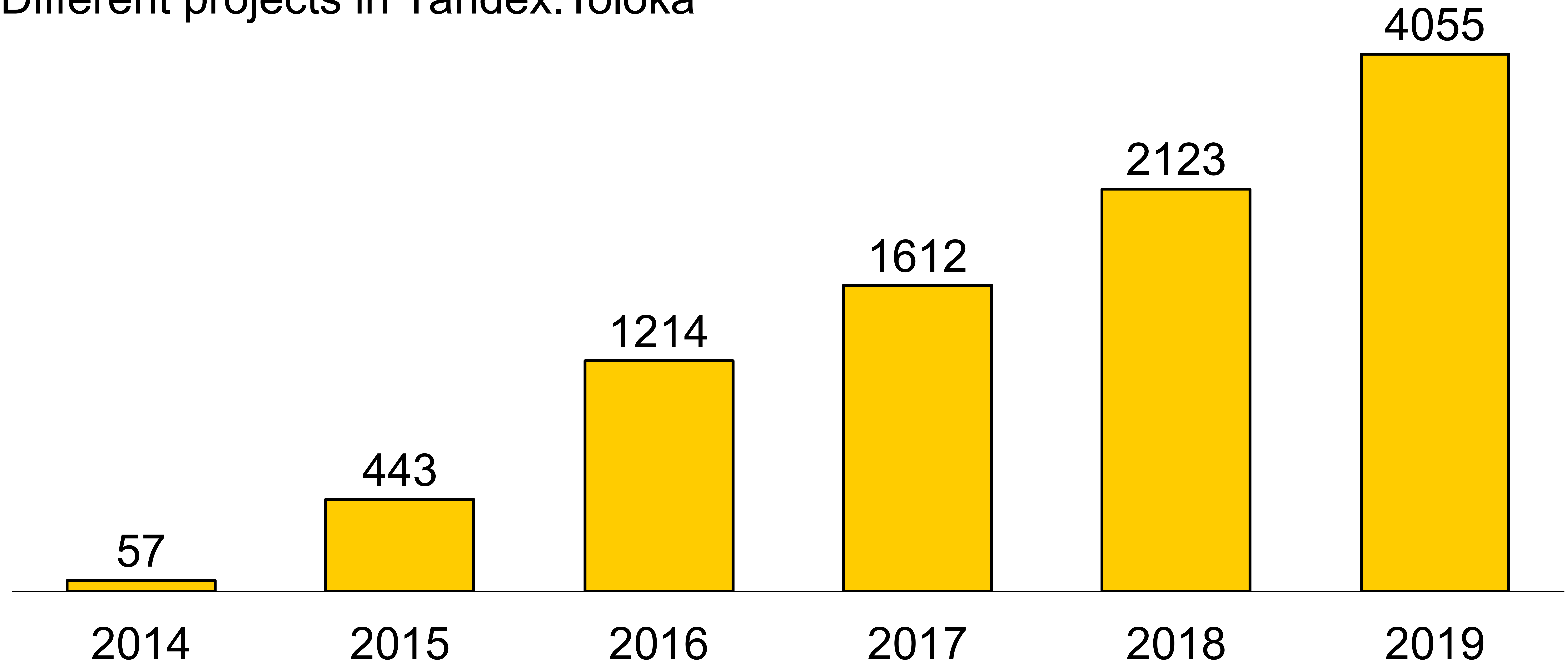
Crowdsourcing growth: Yandex experience

Active performers in Yandex.Toloka



Crowdsourcing growth: Yandex experience

Different projects in Yandex.Toloka



Everyday on Yandex.Toloka

500+ different projects

36K+ performers

12M+ tasks

Yandex.Toloka: real-life cases

Side-by-side object comparison
1,000 tasks

Done in 10 min
Cost: $2.4

Phrase generation for a chatbot
500 phrases for the same topic

Done in 15 min
Cost: $1

Object classification
1,000 photos

Done in 15 min
Cost: $1.2

Audio transcription
100 recordings 25 minute long

Done in 20 min
Cost: $6

Object segmentation
about 1,000 objects in 100 photos

Done in 6 h
Cost: $3.6

Video rating
10,000 videos

Done in 2 h
Cost: $10

Tutorial overview

Why this tutorial?

Practice

Part I: 30 min

Main components of data collection via crowdsourcing

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

Olga Megorskaya
Head of Crowdsourcing Department,
Yandex

Part II: 25 min

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

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

Daria Baidakova,
Project Manager,
Crowdsourcing Department, Yandex

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
Crowdsourcing Department, Yandex

Part IV:60 min

Setting up and running label collection projects (practice session)

You

- create
- configure
- run on real performers

data labelling projects in real-time

Daria Baidakova,
Project Manager,
Crowdsourcing Department, Yandex

Part V: 35 min

Interface & quality control

- Detailed examination of quality control techniques
- Comprehensive overview of best practices for creating a functional interface

Alexey Drutsa
Head of Efficiency and Growth Division
Crowdsourcing Department, Yandex

Part VI: 25 min

Theory on Aggregation

- Multiclass labels
- Pairwise comparisons

Valentina Fedorova
Researcher,
Research Department, Yandex

Part VII: 90 min

Setting up and running label collection projects cont. (practice session)

You

- create
- configure
- run on real performers

data labelling projects in real-time

Daria Baidakova,
Project Manager,
Crowdsourcing Department, Yandex

Part VIII: 20 min

Theory on efficient incremental relabelling and pricing

- Incremental relabelling
- Performance-based pricing

Valentina Fedorova
Researcher,
Research Department, Yandex

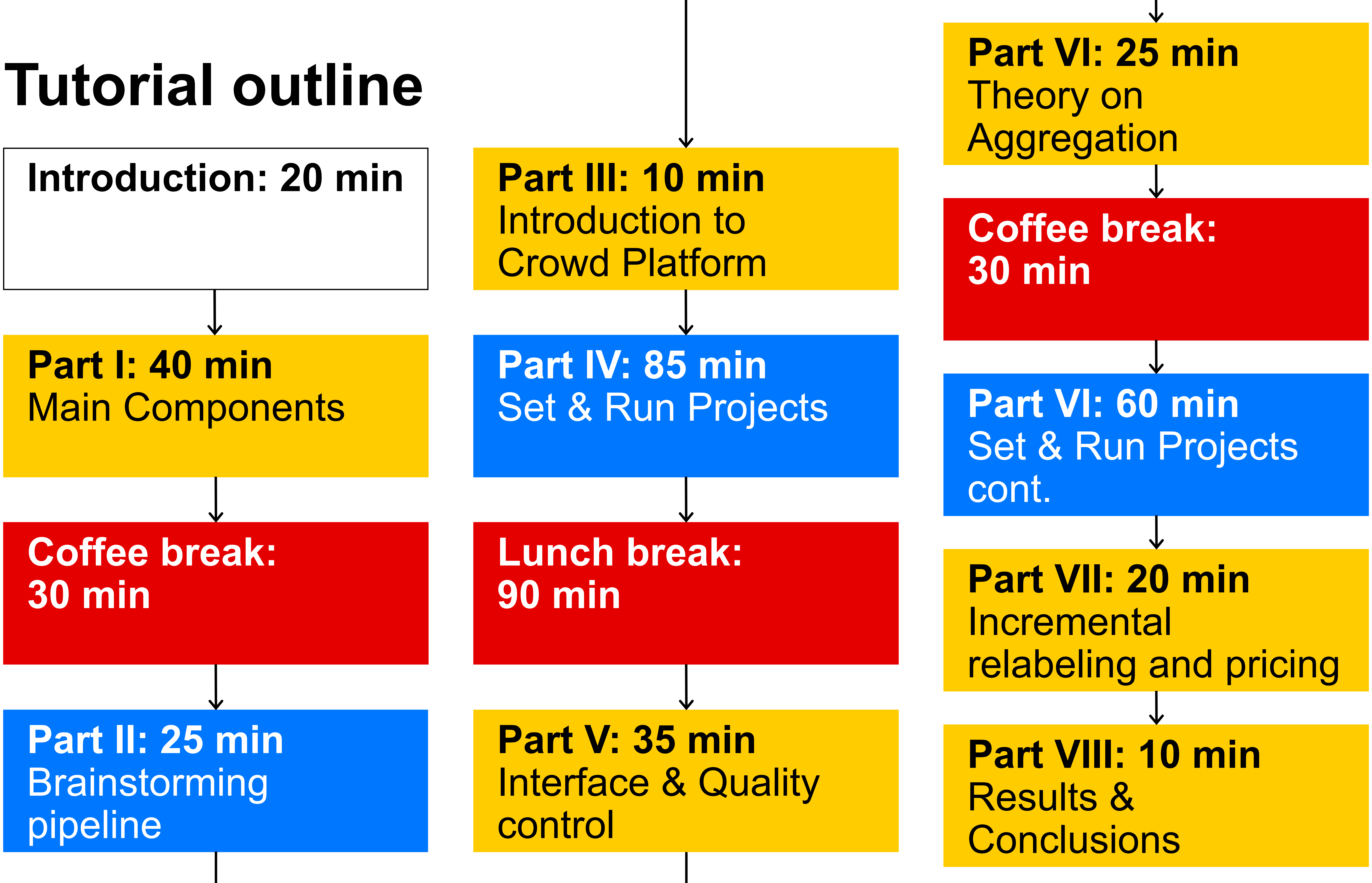
Part IX: 10 min

Discussion of results from the projects & conclusions

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

Alexey Drutsa
Head of Efficiency and Growth Division
Crowdsourcing Department, Yandex

Tutorial outline



Introduction: 20 min

Part I: 40 min
Main Components

**Coffee break:
30 min**

Part II: 25 min
Brainstorming pipeline

Part III: 10 min
Introduction to Crowd Platform

Part IV: 85 min
Set & Run Projects

**Lunch break:
90 min**

Part V: 35 min
Interface & Quality control

Part VI: 25 min
Theory on Aggregation

**Coffee break:
30 min**

Part VI: 60 min
Set & Run Projects cont.

Part VII: 20 min
Incremental relabeling and pricing

Part VIII: 10 min
Results & Conclusions

Yandex

Thank you!
Questions?

Olga Megorskaya

Head of Crowdsourcing Department

omegorskaya@yandex-team.ru

https://research.yandex.com/tutorials/crowd/wsdm-2020