Yandex

Yandex

SIGMOD 2020 hands-on tutorial

# Crowdsourcing Practice for Efficient Data Labeling: Aggregation, Incremental Relabeling, and Pricing

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

# Introduction

Olga Megorskaya,
Head Yandex.Toloka

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

Search

Machine translation

Self-Driving

Personal assistant

Ads

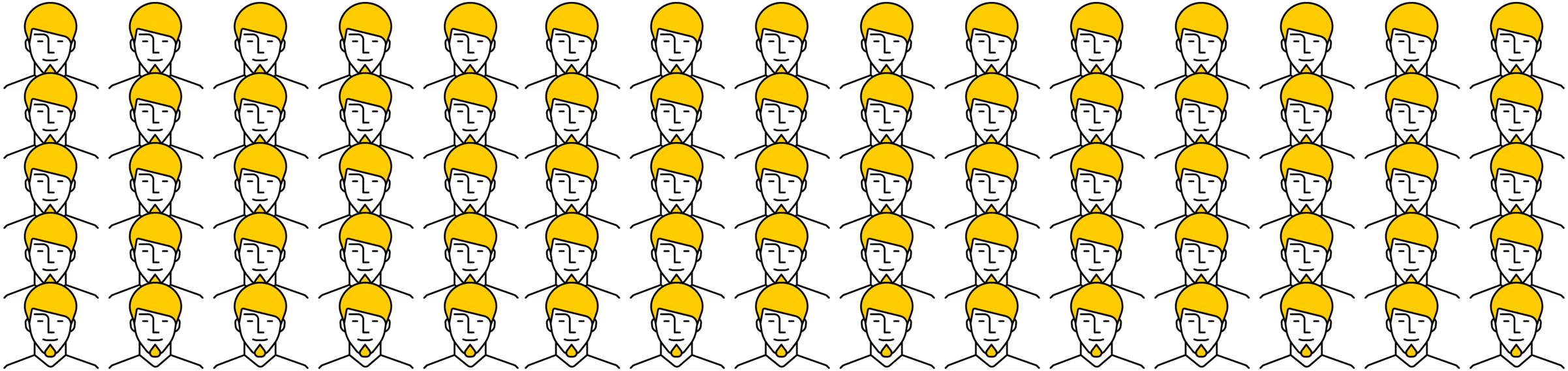
Maps

Speech technologies

E-commerce

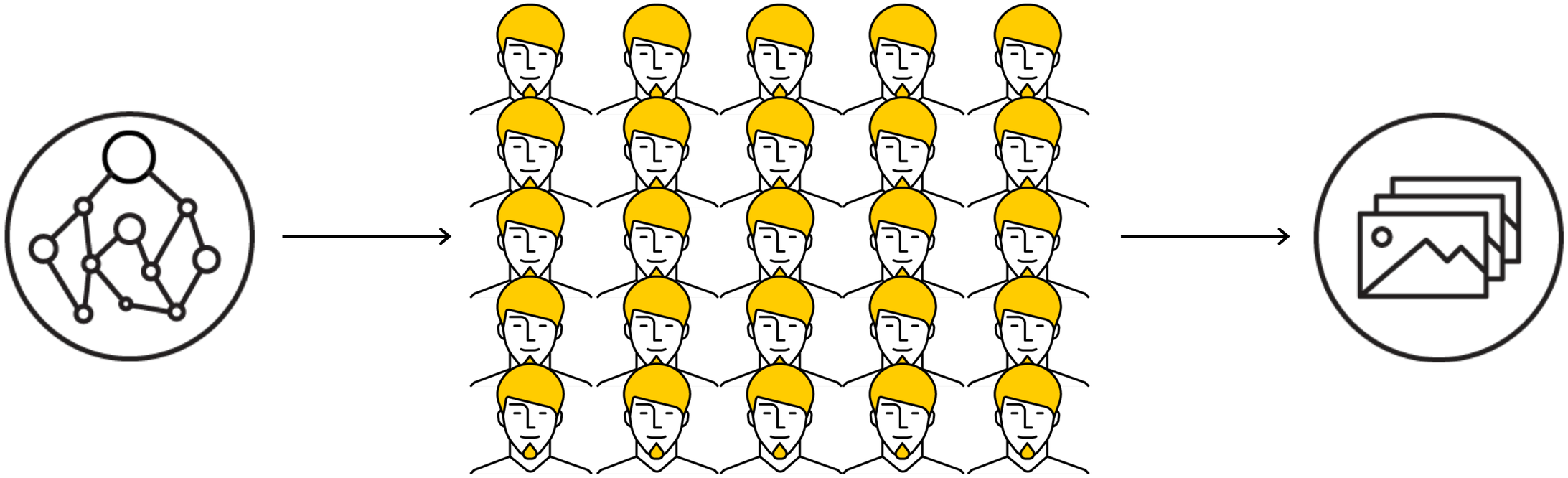
# Majority of ML-based solutions require training data labelled by human

# …at a large scale



# Crowdsourcing

Specific way to design a business process

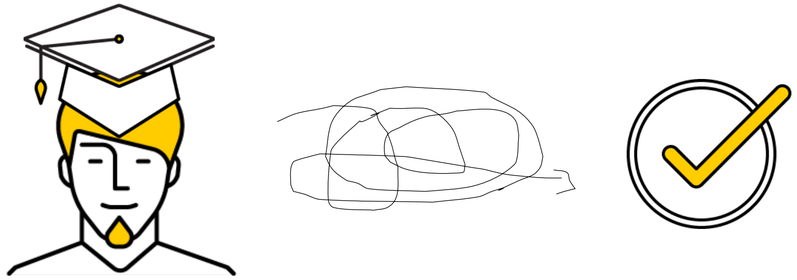


A big task

Cloud of performers

Result

# Crowdsourcing: require less from performer, more – from manager

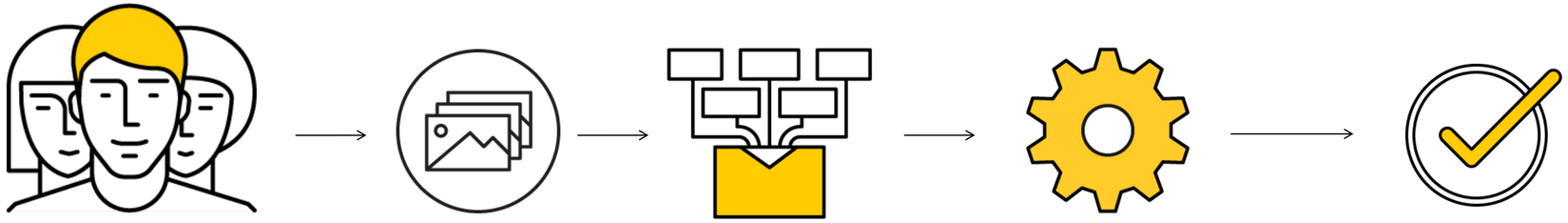


Expert approach: rely on an expertise of a particular performer:

-expensive

-unmeasurable

-hard to scale



Crowdsourcing approach:

-measurable

-scalable

-manageable

# XX century – style management

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

https://media.glassdoor.com/l/6e/c3/ee/17/perfectserve-global-partner-cubicles.jpg

# 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

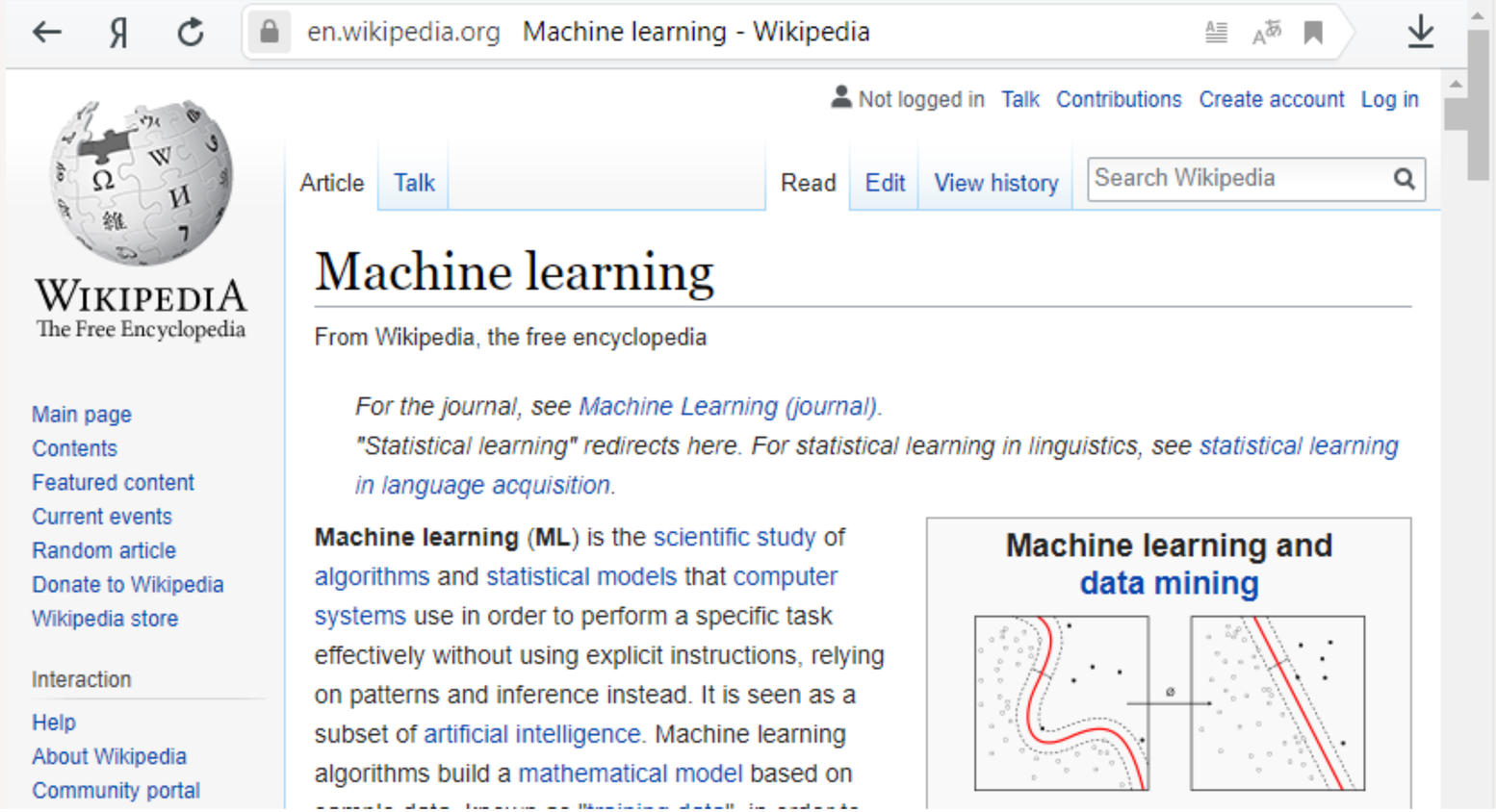
# Example: multi classification

# Example: multi classification with ordered labels

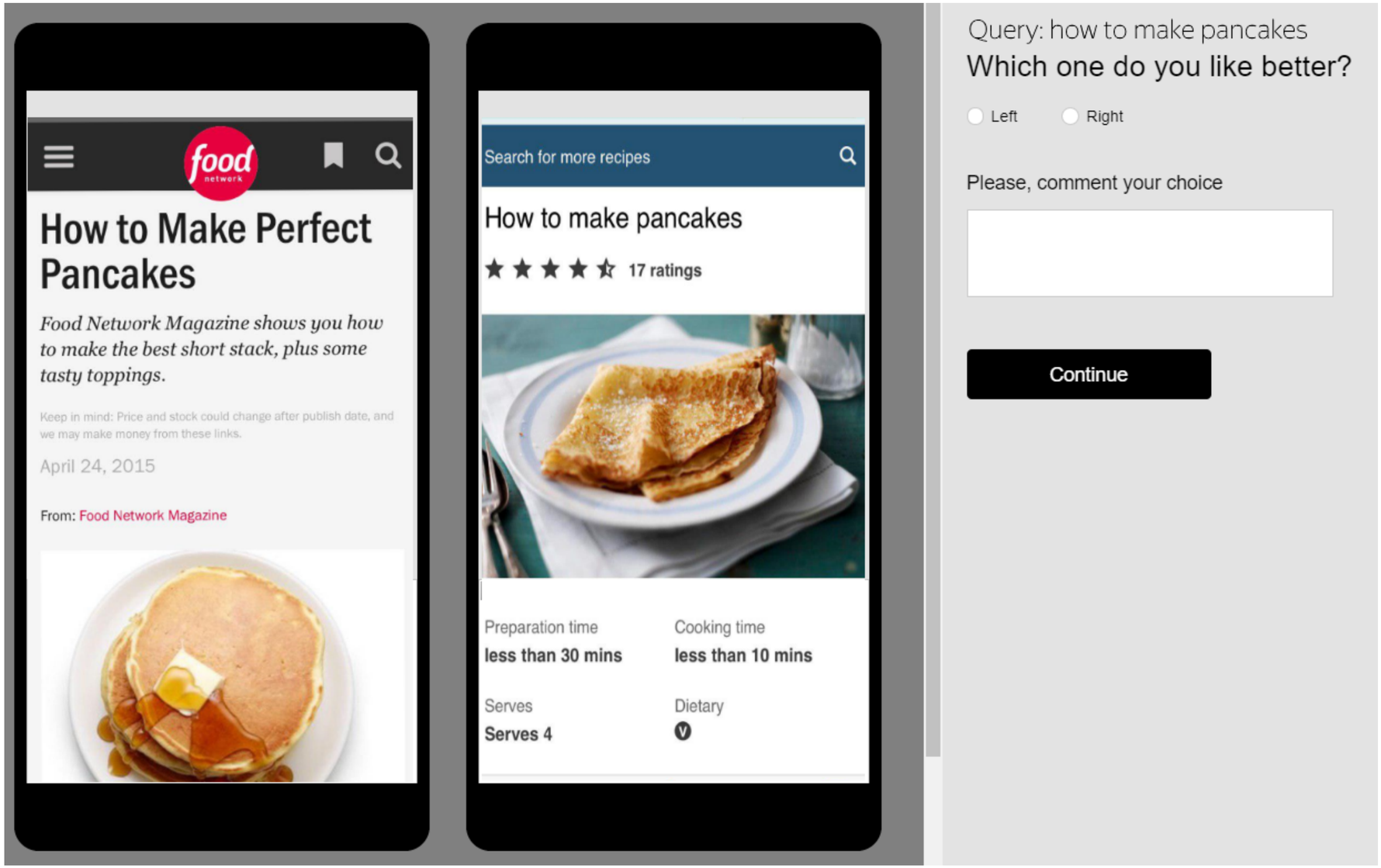
Query: Machine learning
URL: https://en.wikipedia.org/wiki/Machine_learning

Open the original | Yandex | Google

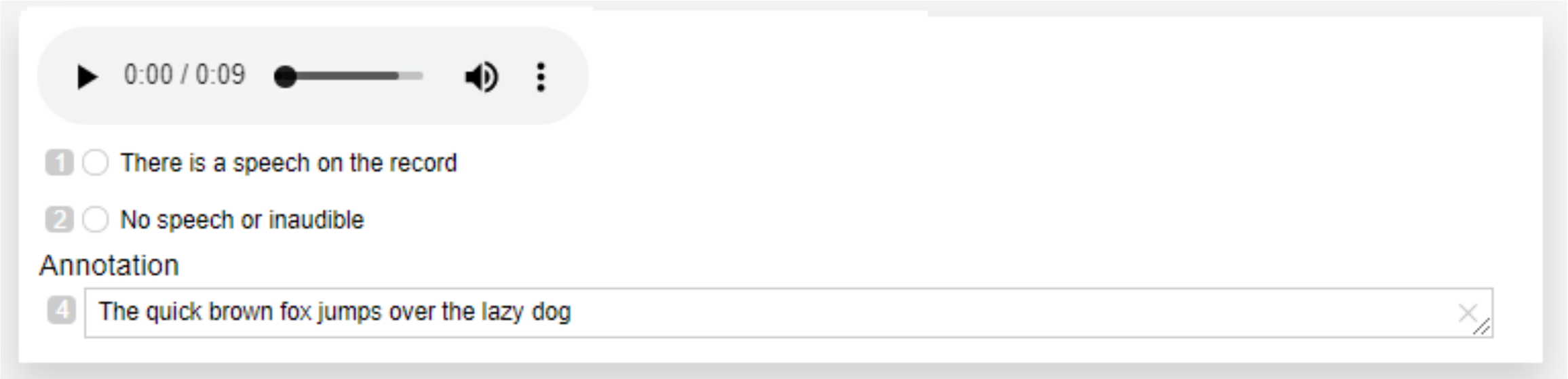
1 Vital
2 Useful
3 Relevant+
4 Relevant-
5 Irrelevant
6 Not displayed



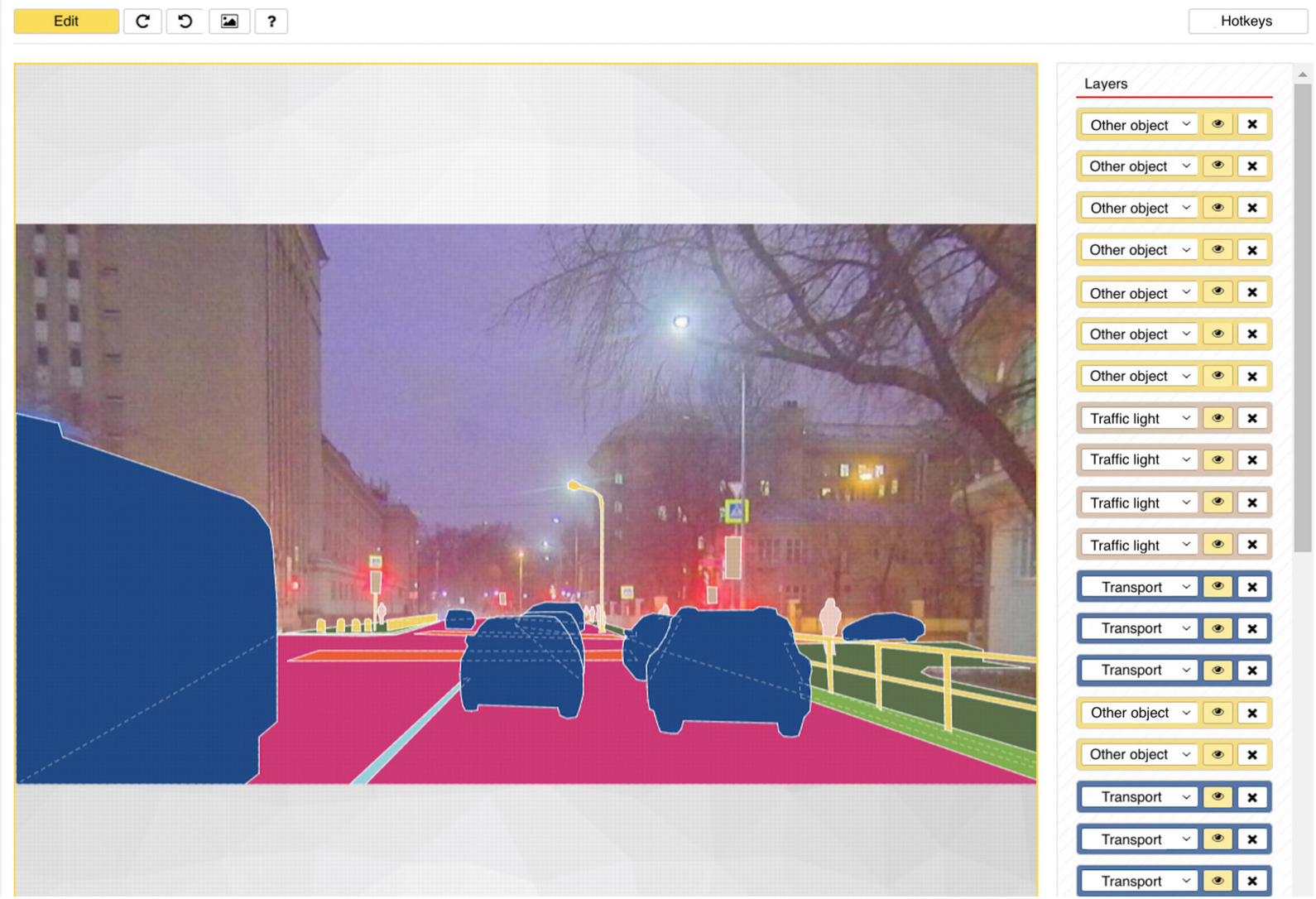
# Examples: pairwise comparison



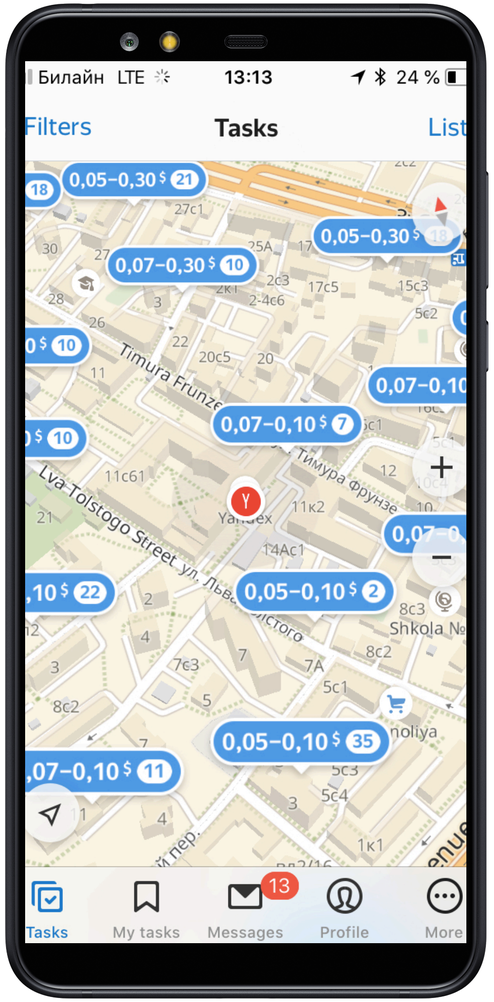
# Examples: transcription with textual answers



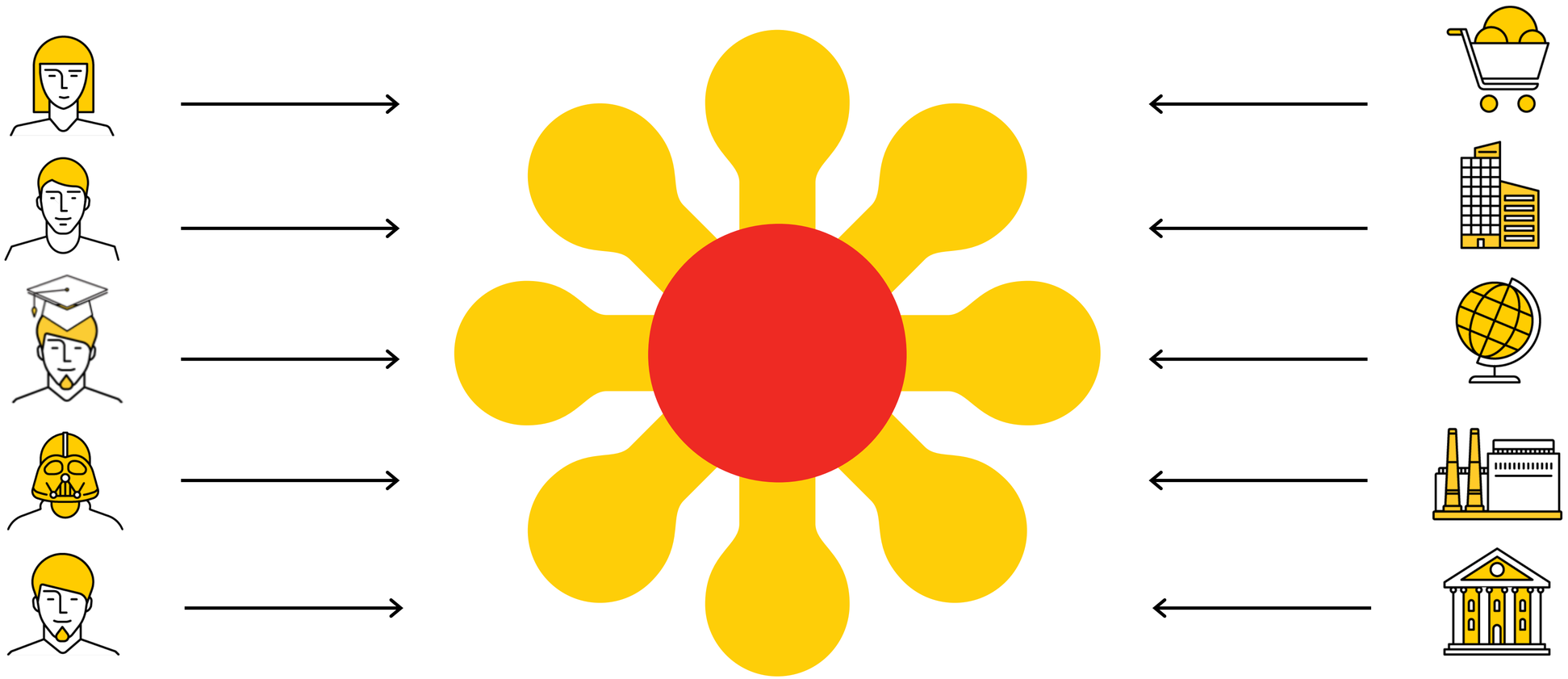
# Examples: object segmentation



# Examples: spatial crowdsourcing



# A crowdsourcing platform: two-sided market



Performers

Requesters

# Crowdsourcing platforms: examples

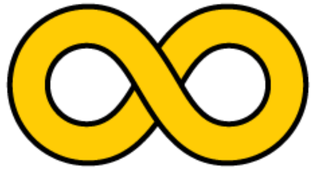
- Amazon Mechanical Turk
- Yandex.Toloka
- Microworkers
- Gigwalk
- ClickWorker
- CloudFactory
- CrowdSource
- DefinedCrowd
- …

# Pros of crowdsourcing platforms

24/7

Variety of skilled performers

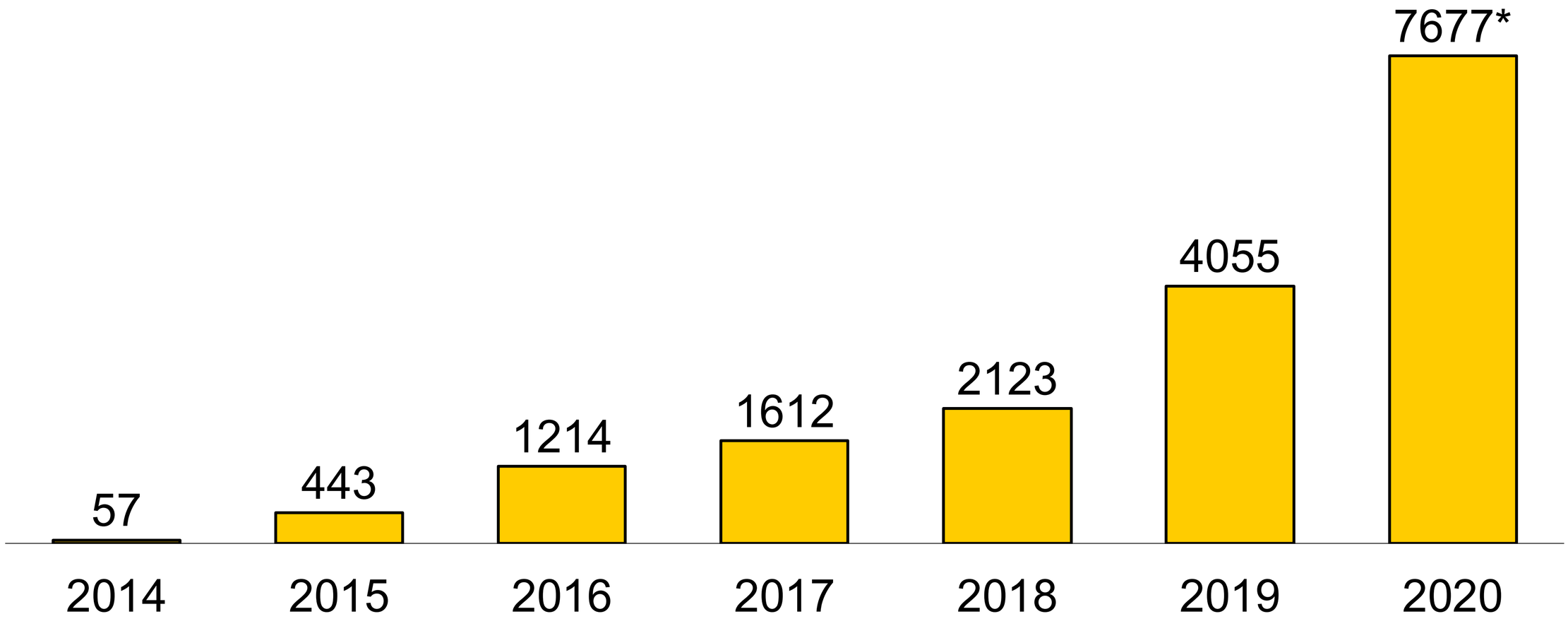
Vast region coverage



Ongoing processes

# Crowdsourcing growth: Yandex experience

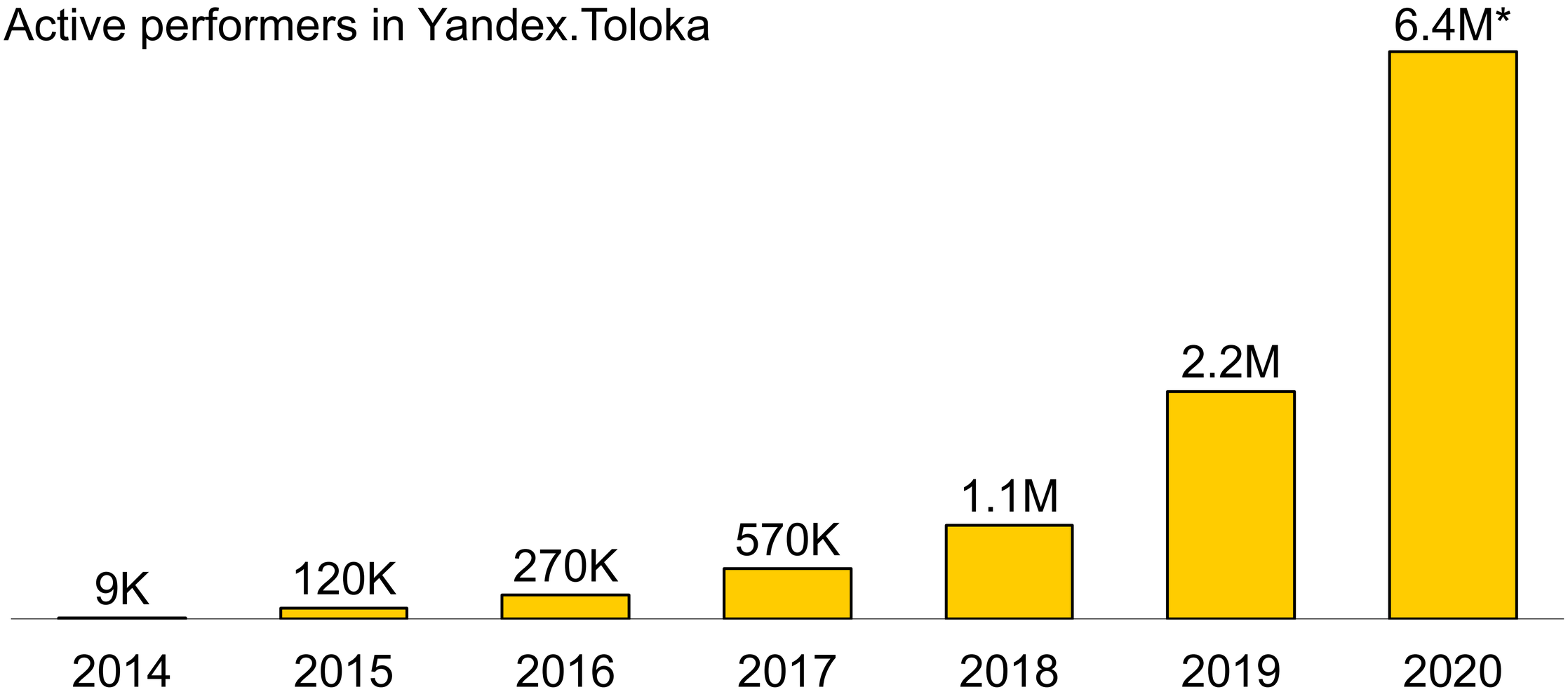
Different projects in Yandex.Toloka




* An extrapolation based on the first 3 months of 2020

# Crowdsourcing growth: Yandex experience

Active performers in Yandex.Toloka




* An extrapolation based on the first 3 months of 2020

# Everyday on Yandex.Toloka

500+ different projects

37K+ performers

13M+ 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 ranking**
10,000 videos

Done in 2 h
Cost: $10

# Tutorial overview

# Why this tutorial?

Practice

# Tutorial schedule

**Introduction: 15 min**

**Part I: 20 min**
Main Components

**Part II: 10 min**
Introduction to Crowd Platform

**Part III: 15 min**
Brainstorming pipeline

**Part IV: 60 min**
Set & Run Projects

**Part V: 25 min**
Theory on Aggregation

**Break:**
**30 min**

**Part VI: 20 min**
Set & Run Projects cont.

**Part VII: 10 min**
Results & Conclusions

Yandex

# Thank you!
# Questions?

**Olga Megorskaya**

Head of Crowdsourcing Department

omegorskaya@yandex-team.ru

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