Toloka

Toloka

ECIR 2023 Tutorial

Crowdsourcing for Information Retrieval

Dmitry Ustalov, Natalia Fedorova, Nikita Pavlichenko,
Alisa Smirnova, Daniil Likhobaba

Toloka

Part I
Introduction

Dr. Dmitry Ustalov,
Head of Ecosystem Development Unit at Toloka

Information Retrieval (IR)

- **IR Research** relies on evaluation and training datasets for studying search, relevance, user behaviour

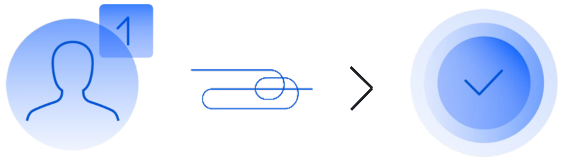
- **IR Applications** require up-to-date and accurate information about human preferences

- In this tutorial, we will demonstrate **how to gather IR datasets** using crowdsourcing and **how to train machine learning models** based on crowdsourced data

Data Labeling Techniques

In-house “expert”: managing people

Directly managing in-house crowd

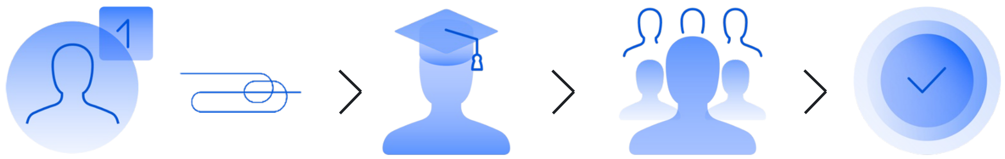


- Easy to setup
- Expensive

- Unmeasurable
- Impossible to scale

BPO / vendor

Access to a crowd via third-party BPO who manages them

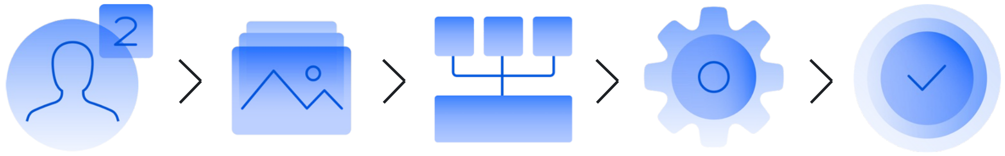


- Quick access to crowd
- Expensive

- Unmeasurable
- Hard to scale

Crowdsourcing

Technologically managing the crowd as yet another computing power

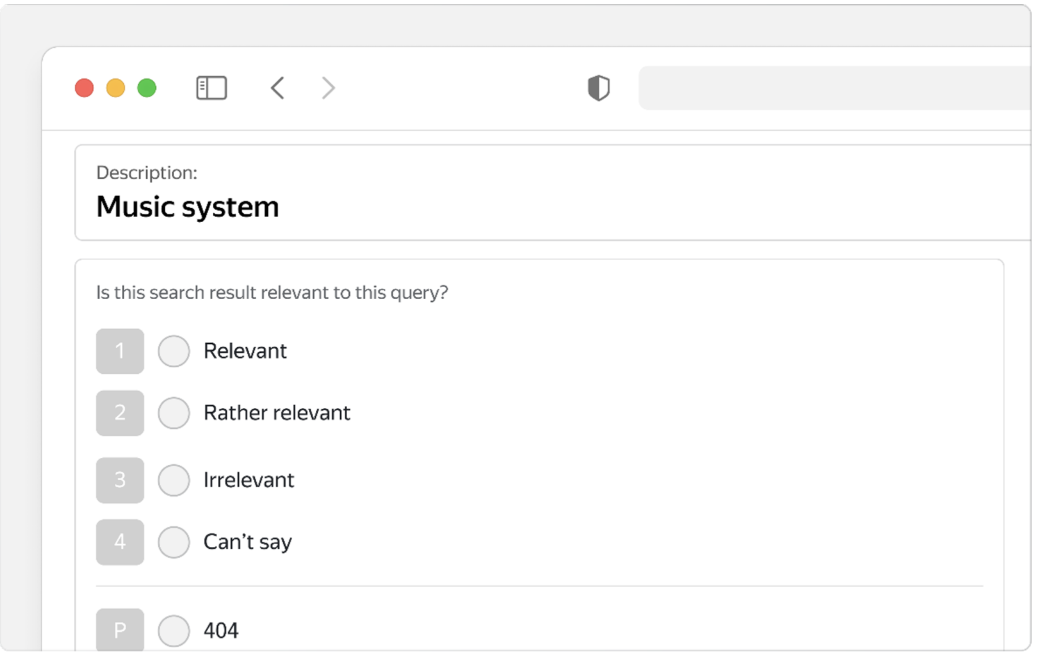


- Measurable
- Scalable

- Manageable
- Requires advanced tech

Crowdsourcing for IR

Product +
search query



Description:
Music system

Is this search result relevant to this query?

1 ○ Relevant

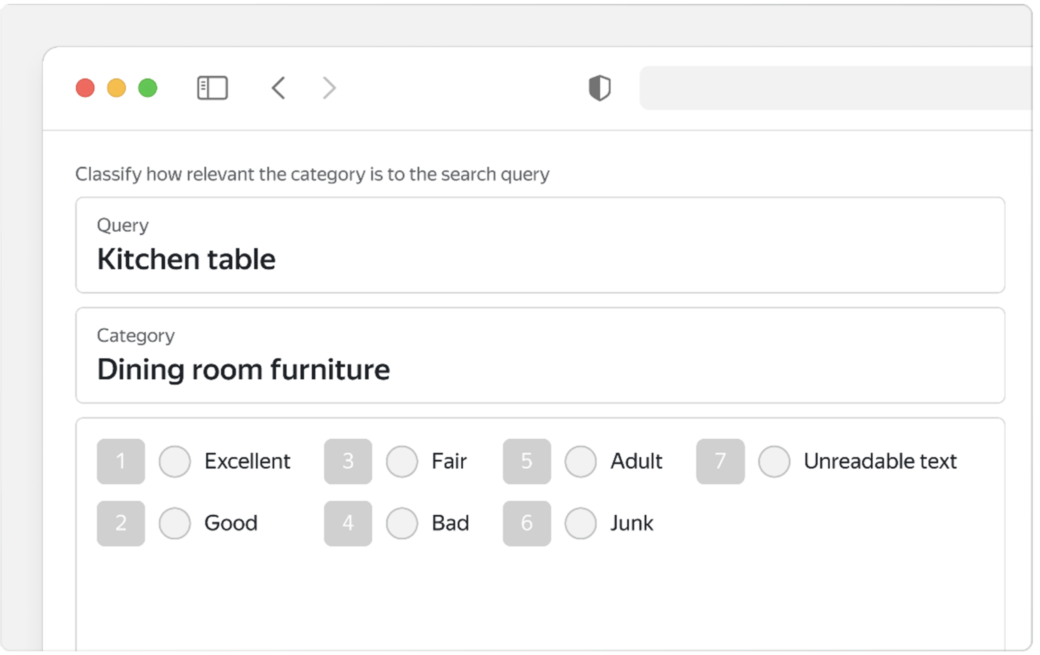
2 ○ Rather relevant

3 ○ Irrelevant

4 ○ Can't say

P ○ 404

Category +
search query



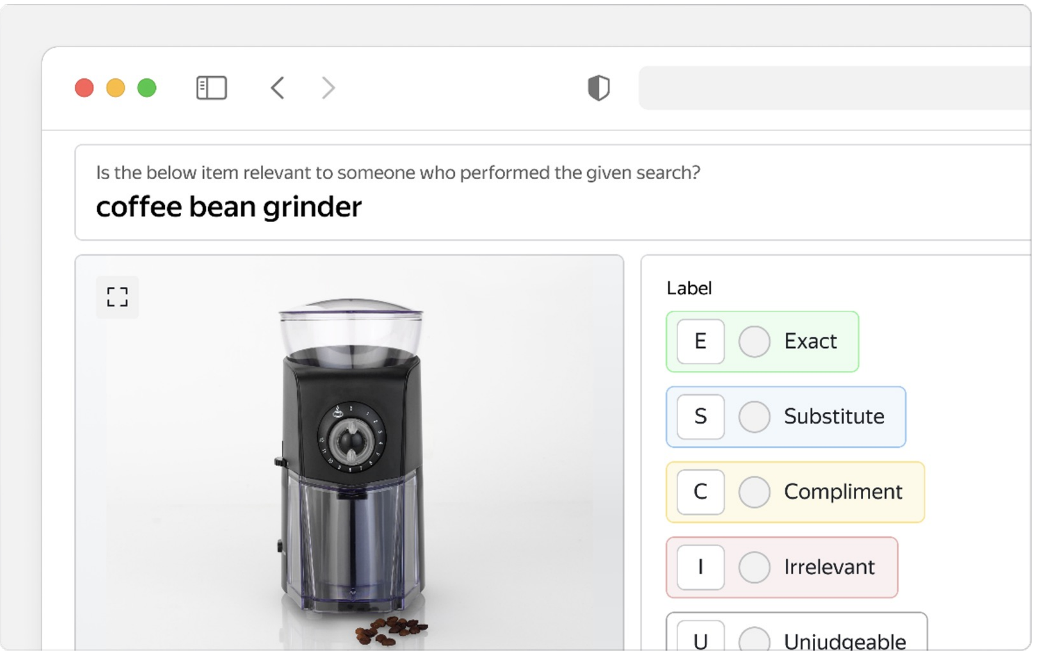
Classify how relevant the category is to the search query

Query
Kitchen table

Category
Dining room furniture

1 ○ Excellent
2 ○ Good
3 ○ Fair
4 ○ Bad
5 ○ Adult
6 ○ Junk
7 ○ Unreadable text

Image +
search query



Is the below item relevant to someone who performed the given search?
coffee bean grinder

Label

E ○ Exact

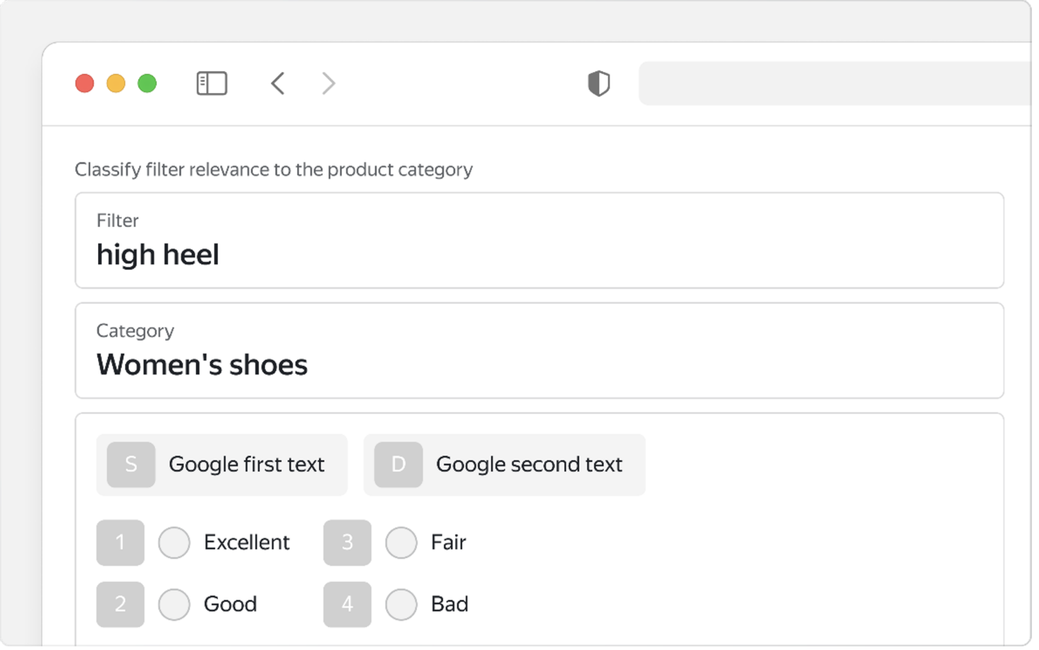
S ○ Substitute

C ○ Compliment

I ○ Irrelevant

U ○ Unjudgeable

Filters +
category



Classify filter relevance to the product category

Filter
high heel

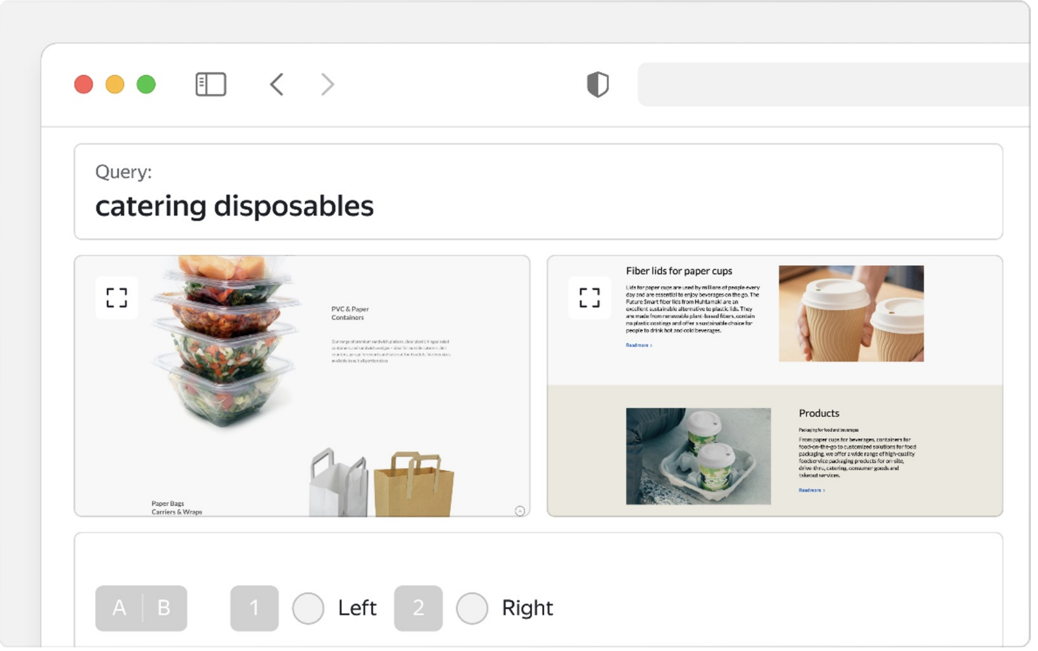
Category
Women's shoes

S Google first text
D Google second text

1 ○ Excellent
2 ○ Good
3 ○ Fair
4 ○ Bad

Crowdsourcing for IR

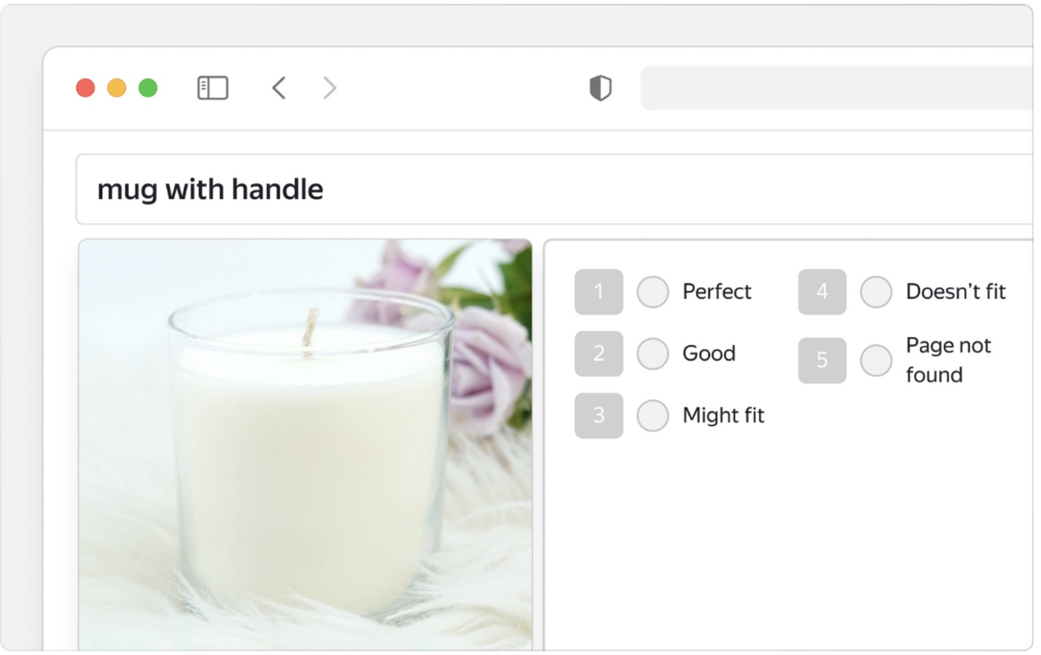
Side-by-side comparison of search results



Query:
catering disposables

A | B   1 ○ Left   2 ○ Right

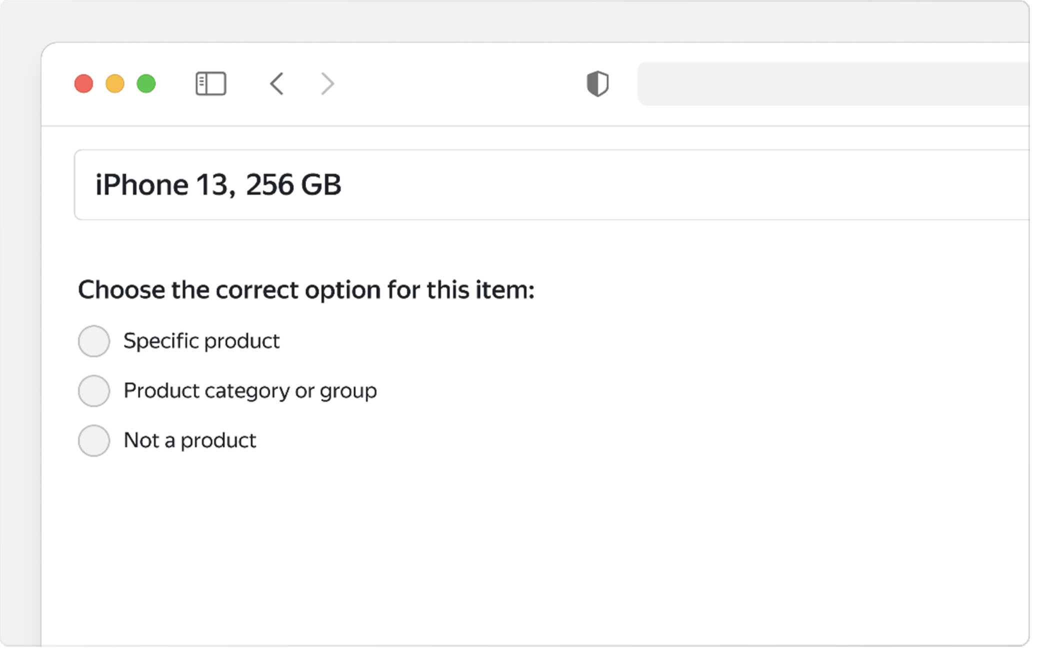
Identify spam or irrelevant matches



mug with handle

1 ○ Perfect
2 ○ Good
3 ○ Might fit
4 ○ Doesn't fit
5 ○ Page not found

Classify type of search query (broad vs narrow)



iPhone 13, 256 GB

Choose the correct option for this item:

○ Specific product
○ Product category or group
○ Not a product

Why this tutorial?
Practice!

Learning outcomes

Theory

- Crowdsourcing essentials
- Aggregation and learning from crowds

Practice

- Build scalable data labeling pipelines
- Set up crowdsourcing projects with real annotators
- Run human-in-the-loop via open source Python libraries (Toloka-Kit and Crowd-Kit)

Tutorial Schedule

Part I Intro: 15 min
Introduction

Part II: 45 min
Crowdsourcing Essentials

Part III: 30 min
Hands-On Practice Session

**Coffee Break :
20 min**

Part III: 30 min
Hands-On Practice Session

Part IV: 45 min
Learning from Crowds

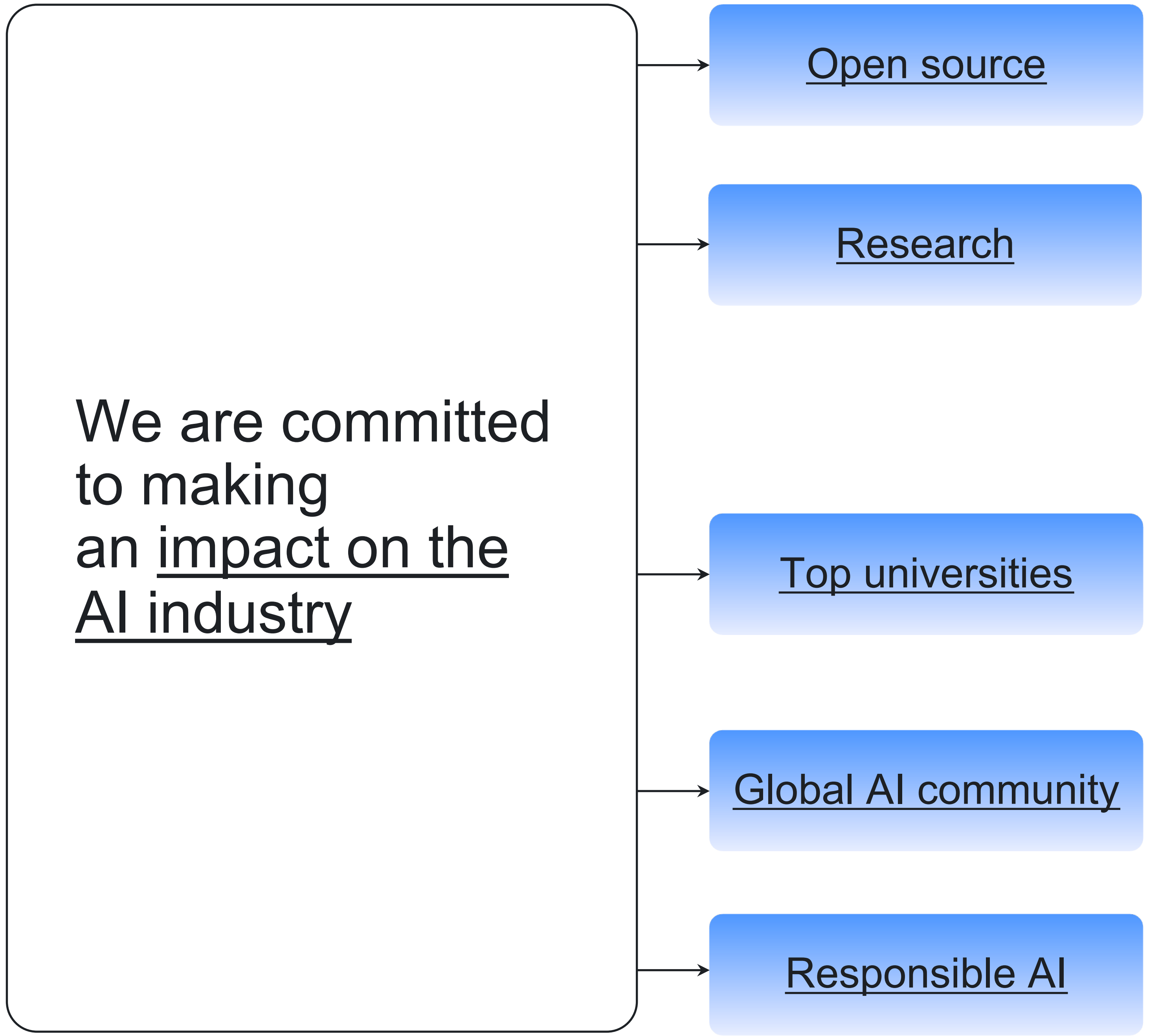
Part V: 15 min
Conclusion

Toloka Research Grants Program

- We encourage the use of crowdsourcing for research purposes
- Recipients of the grant are awarded up to $500 in credit to fuel their research

https://toloka.ai/grants/

Our team helps the AI industry



We are committed to making an impact on the AI industry

Open source

We encourage collaborative development on open-source projects that make it easier to work with crowdsourced data

Research

We have established our expertise among industry leaders with research papers and workshops at top-tier AI conferences:

KDD, SIGMOD, CVPR, NeurIPS, ICML,
NAACL-HLT, WSDM, VLDB, ECIR, SIGIR

Top universities

We share our advanced crowdsourcing techniques in open datasets, online courses on data labeling, and collaboration with top universities

Global AI community

We support a global community of AI practitioners for open exchange of ideas and know-how

Responsible AI

We consider it a privilege to contribute to the AI community with responsible data production that supports ethical approaches to training, testing, and monitoring AI

Thank You!

Dr. Dmitry Ustalov

Head of Ecosystem Development at Toloka

dustalov@toloka.ai

https://toloka.ai/

LinkedIn https://www.linkedin.com/company/toloka/

https://twitter.com/tolokaai

slack tolokacommunity.slack.com

GitHub https://github.com/Toloka