
Toloka

Toloka

RecSys 2022 Tutorial

Improving Recommender Systems with Human-in-the-Loop

Dmitry Ustalov, Natalia Fedorova, Nikita Pavlichenko,
Maxim Kunakov, Fedor Zhdanov

Toloka

Part III
Human-in-the-Loop Essentials

Nikita Pavlichenko,
Researcher

Tutorial Schedule

Part I Intro: 10 min
Introduction

Part III: 20 min
Human-in-the-Loop Essentials

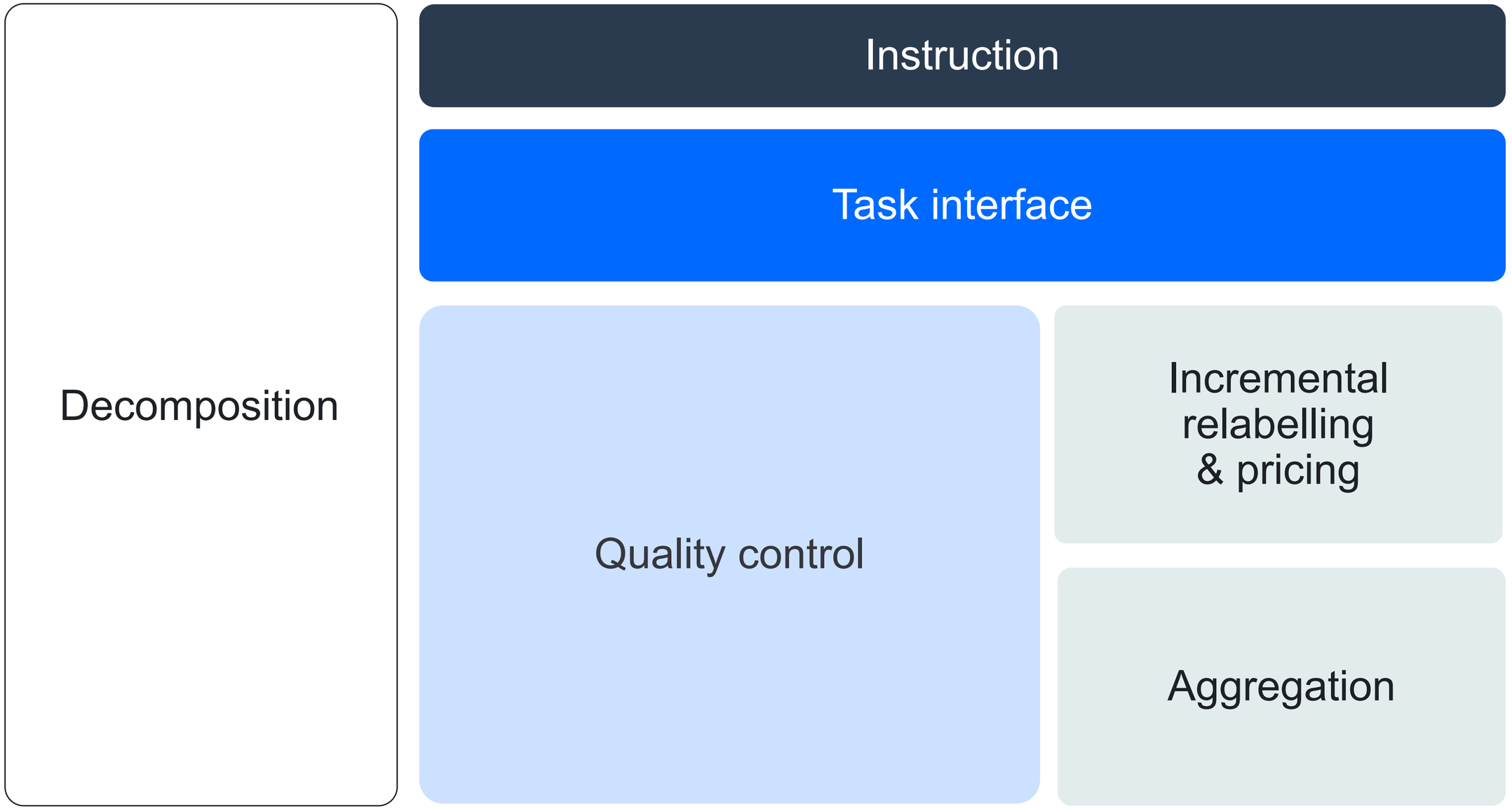
Part III: 20 min
Human-in-the-Loop Essentials

Part IV: 50 min
Hands-On Practice Session

Coffee Break :
10 min

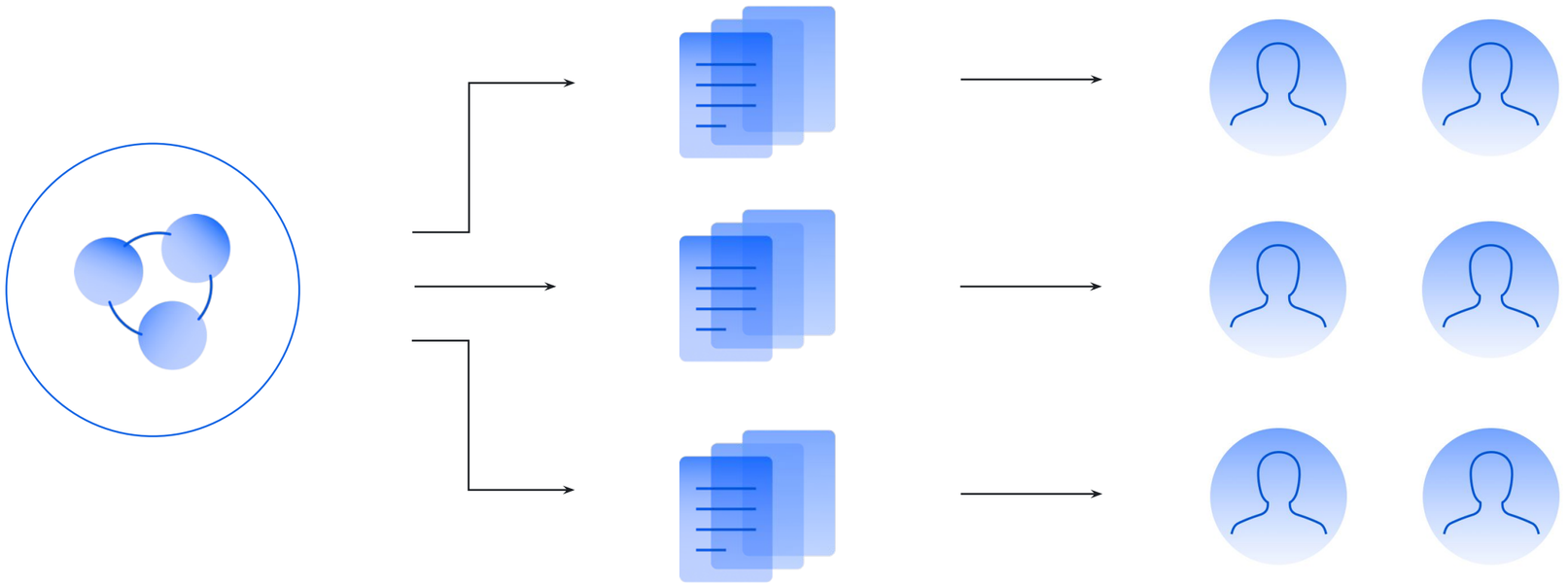
Part V: 30 min
From Human Labels to Ground Truth

Part VI: 10 min
Conclusion


Decomposition
Instruction
Task interface
Quality control
Incremental relabelling & pricing
Aggregation

Decomposition

Decomposition



A big task

Projects with microtasks
of different type

Cloud
of performers

Decomposition: why?

- Performers are usually non-specialists in your specific task
- The simpler a single task is:
  - The more people can perform your task
  - The easier its instruction
  - The better quality of performance
- A way to:
  - Distinguish tasks of different difficulty levels
  - Control and optimize pricing
  - Control quality by post verification

Decomposition: when?

- If
  - Your task requires an answer selected among more than 3–5 options
  - Your task has long instructions that are hard to read
- Then your task requires decomposition

Case of decomposition: a lot of questions

Bad practice: All questions in one task

What animal is on the photo?
•Cat
•Rabbit
•Bear
•Whale
•Koala
•None of the above

Is its tail visible?
- Yes
- No

Is it running?
- Yes
- No

What color is it?
- White
- Black
- Brown
- Red
- Other

Where is it situated?
- On the grass
- On a tree
- On a road
- It is flying
- None of the above

Case of decomposition: a lot of questions

Good practice: Each question in a separate task

What animal is on the photo?
- Cat
- Rabbit
- Bear
- Whale
- Koala
- None of the above

Is its tail visible?
- Yes
- No

Is it running?
- Yes
- No

What color is it?
- White
- Black
- Brown
- Red
- Other

Where is it situated?
- On the grass
- On a tree
- On a road
- It is flying
- None of the above

Case of decomposition: need to verify answers

The task: Highlight all koalas on the photo

Problem: **highlighting can be done in different ways**

Hence, it is difficult to:

— Compare with control answers

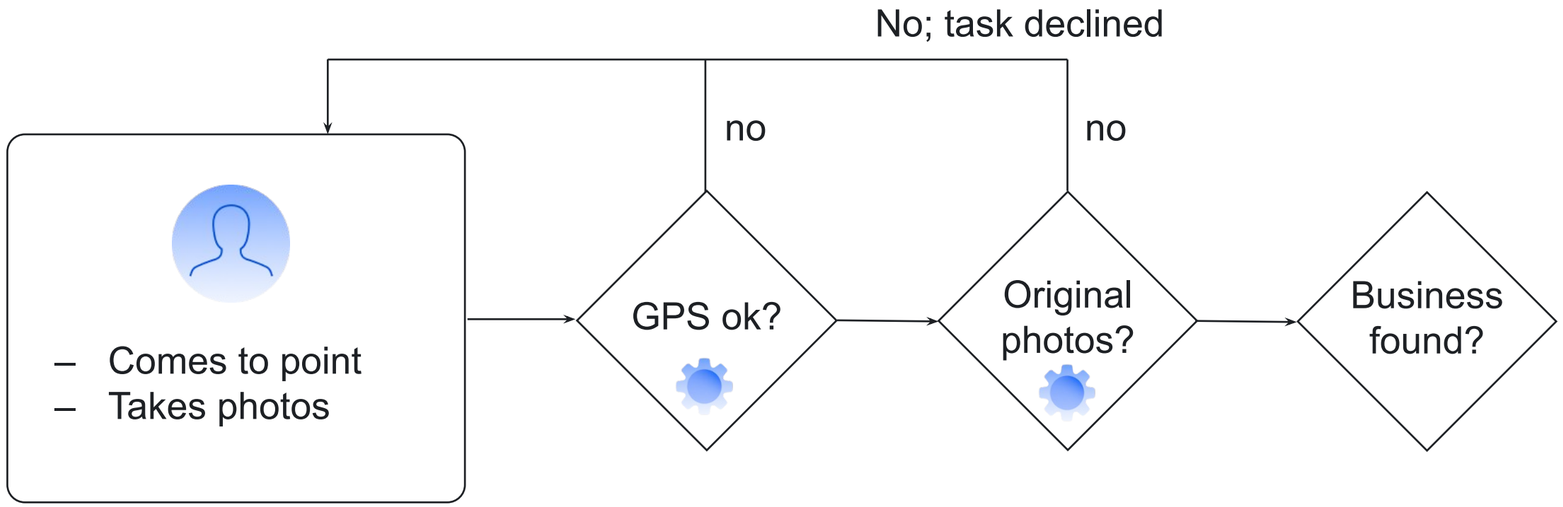
— Aggregate answers from different performers

A good solution
A task for another performer:
Have the koalas been highlighted correctly?

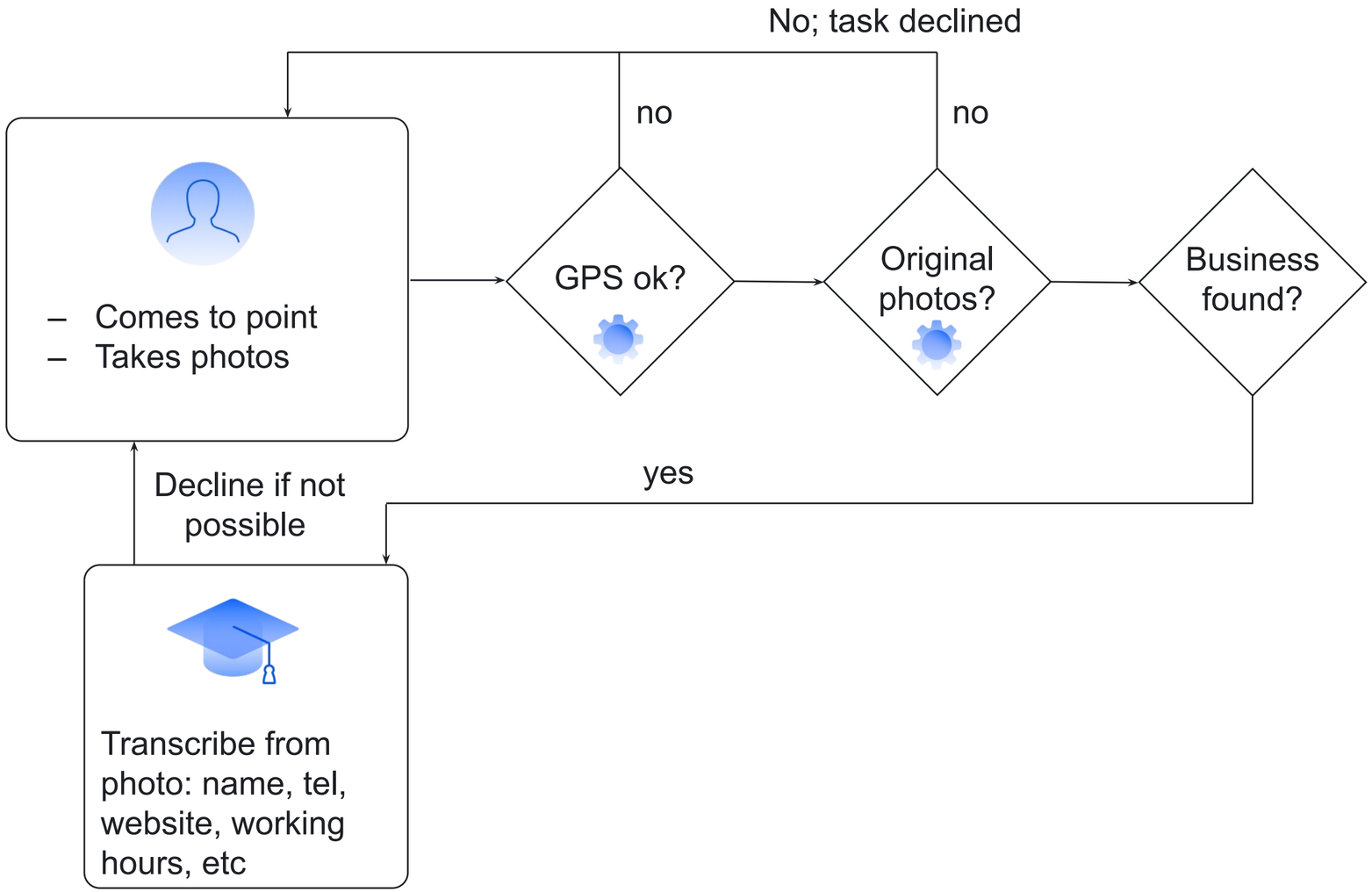
Real example: decomposition for an offline data collection task

- Comes to point
- Takes photos

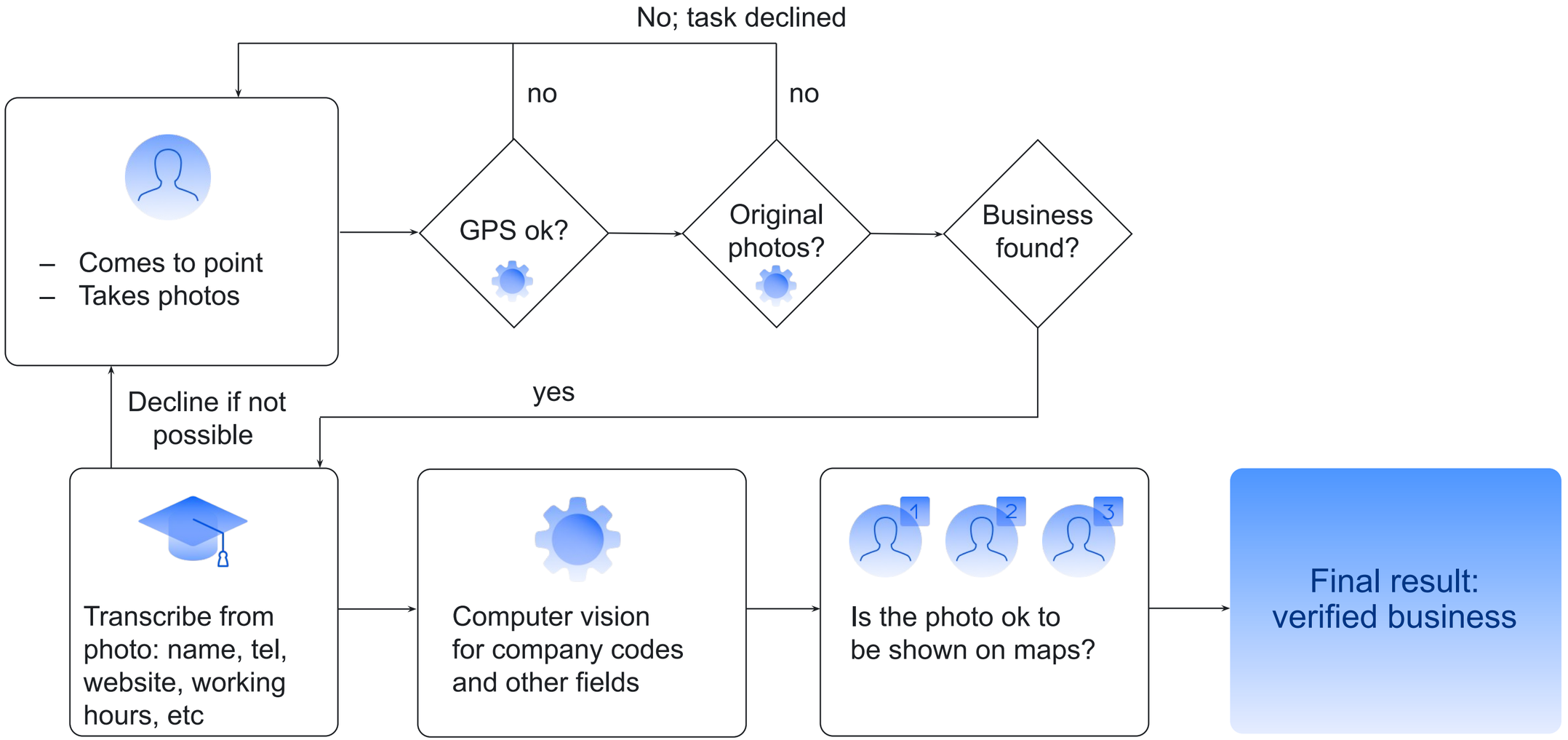
Final result:
verified business

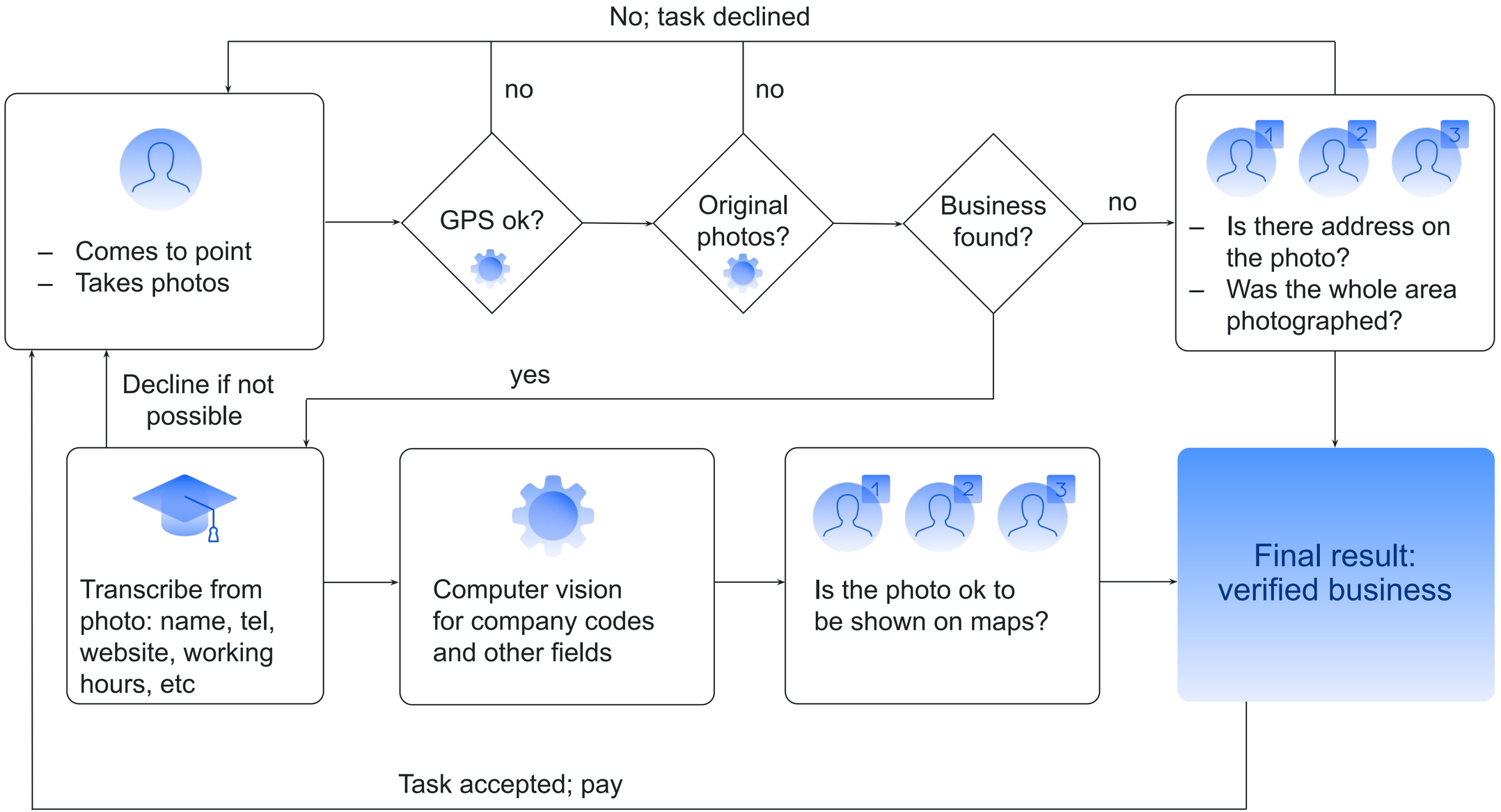


Final result:
verified business



Final result:
verified business


No; task declined
no
no
– Comes to point
– Takes photos
GPS ok?
Original photos?
Business found?
Decline if not possible
yes
Transcribe from photo: name, tel, website, working hours, etc
Computer vision for company codes and other fields
1
2
3
Is the photo ok to be shown on maps?
Final result: verified business


No; task declined
no
no
– Comes to point
– Takes photos
GPS ok?
Original photos?
Business found?
no
– Is there address on the photo?
– Was the whole area photographed?
Decline if not possible
yes
Transcribe from photo: name, tel, website, working hours, etc
Computer vision for company codes and other fields
Is the photo ok to be shown on maps?
Final result: verified business
Task accepted; pay

Instruction

Instruction: a typical structure

- Goal of the task to be done
- Interface description
- Algorithm of required actions
- Examples of good and bad answers
- Algorithm and examples for rare cases
- Reference materials

Most pitfalls are here

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No

OK: the answer and the task seem clear

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No

What is the correct answer?

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No

How to fix

– In the instruction: clarify what you mean under «a white cat»

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No

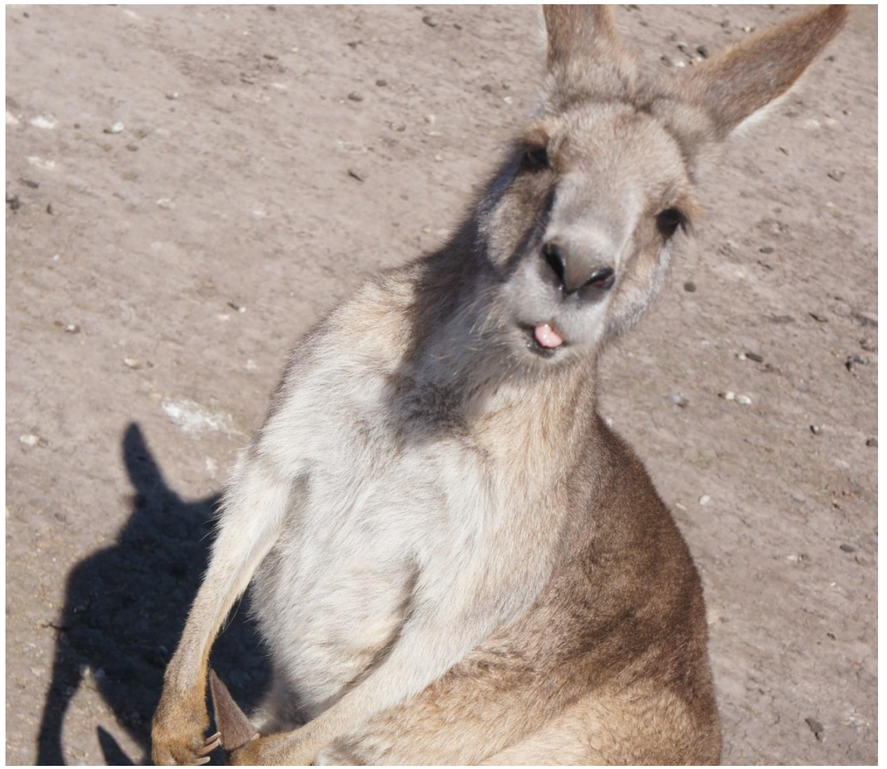
Rare case: many cats

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No



Rare case: not a cat

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No

404: Cannot download the image

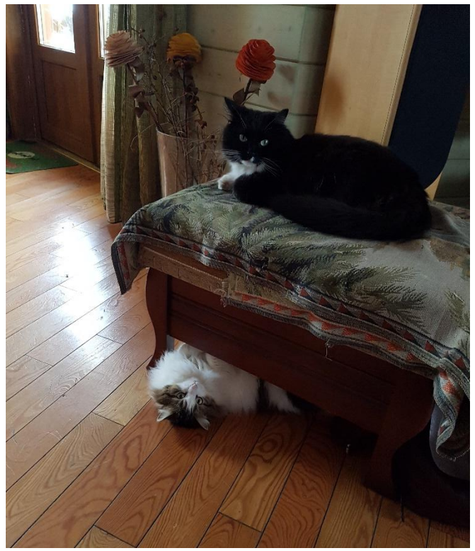
Rare case: image has not been shown

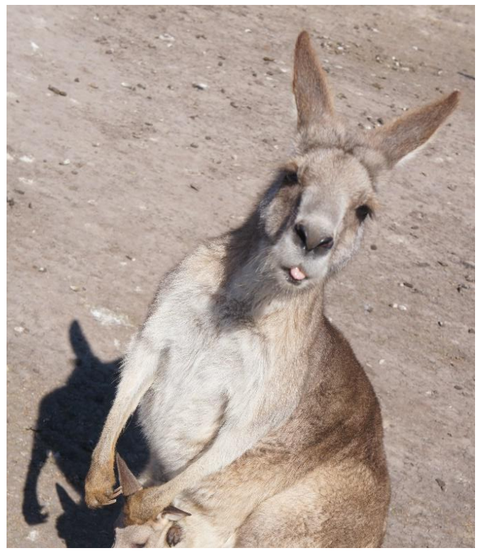
Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No





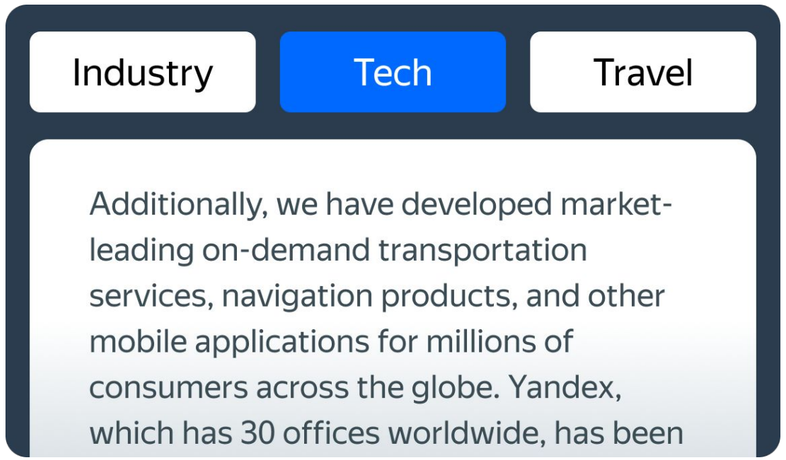
404: Cannot download the image

It is difficult to predict situations of any kind, but you can:
- In the instruction: clarify what should be done in a non-standard situation
- In the interface: add a text field to allow a performer to report the case

Task interface

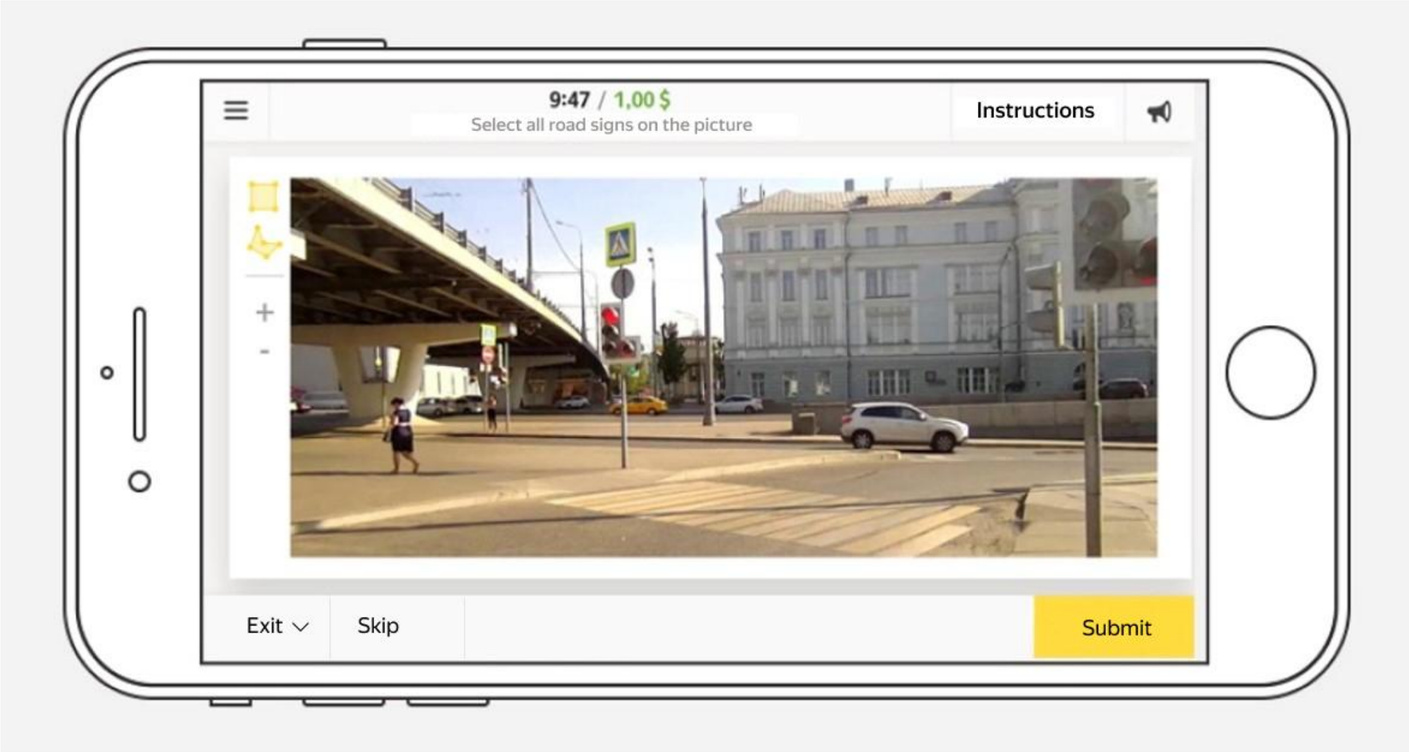
Put yourself in the performer’s shoes

- Many tasks at a time
- Earnings depend on the amount of tasks done
- Monotonous tasks
- Concentration required

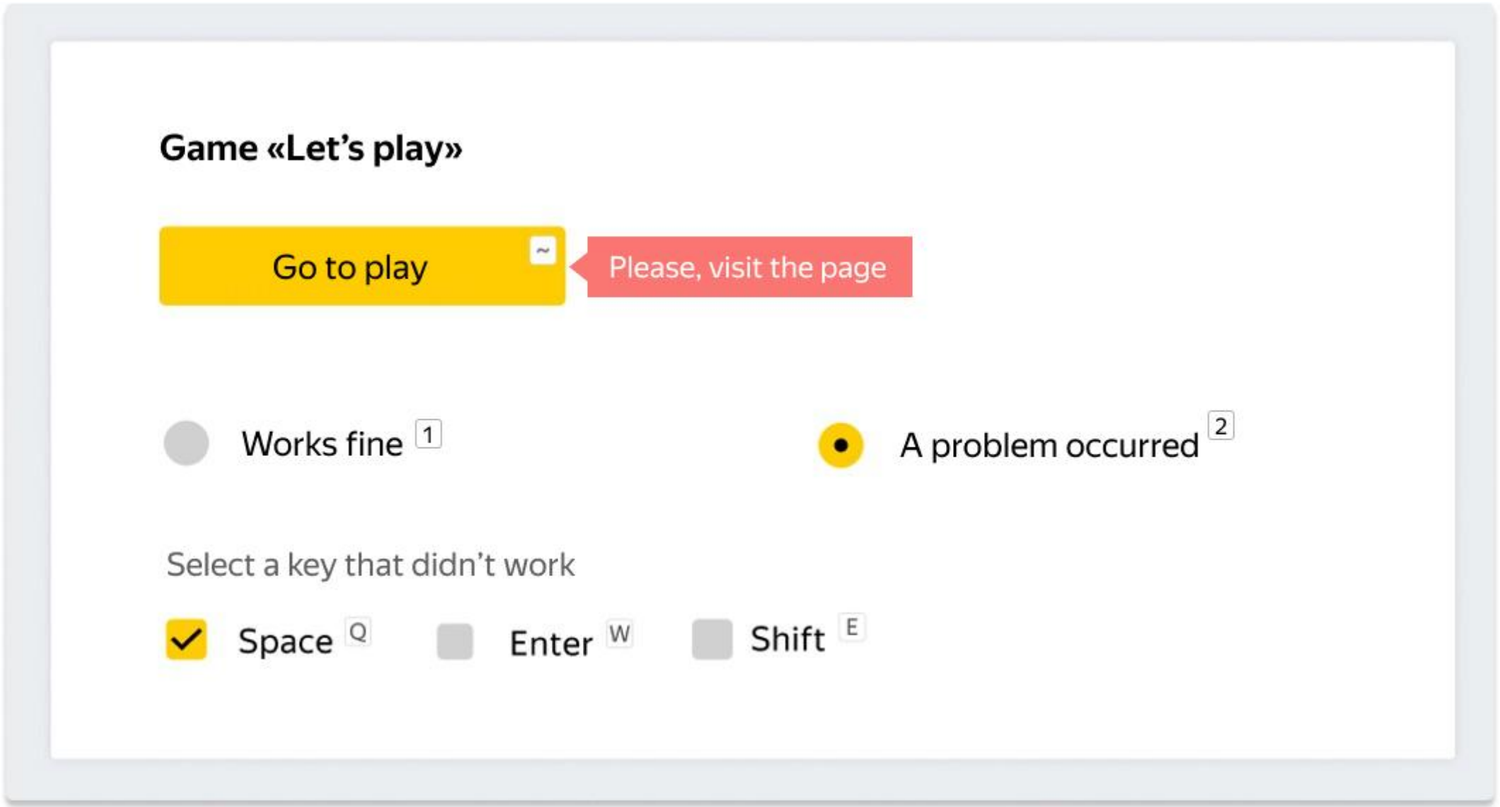


Some tasks are not suitable for cross-platform use

- Task is to find and encircle all road signs on the picture



Checking required actions



Minimum use of external resources

- Add screenshots
- Save data in your storage
- Check clicks on necessary links

Minimalistic design

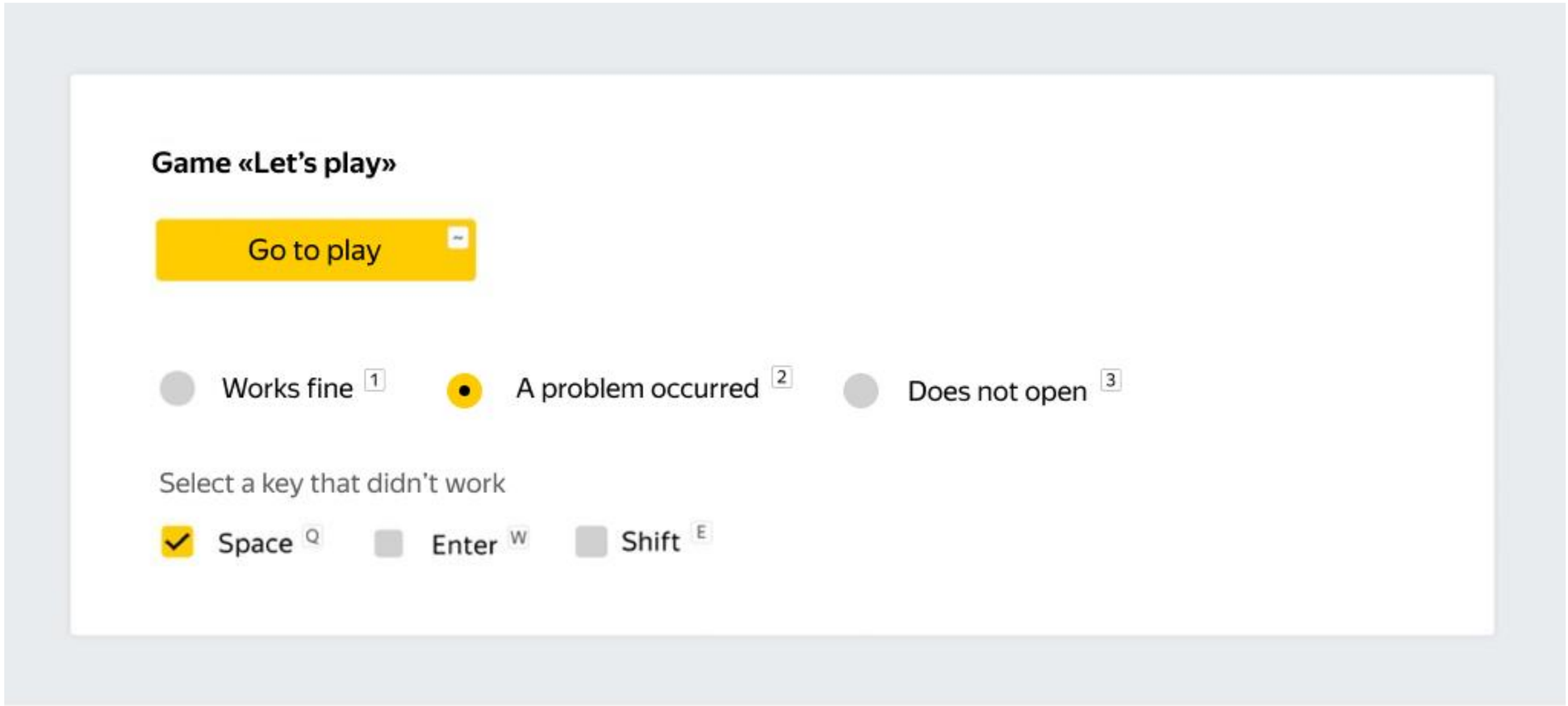
Too many nested blocks

Same font everywhere

Verdicts look like links

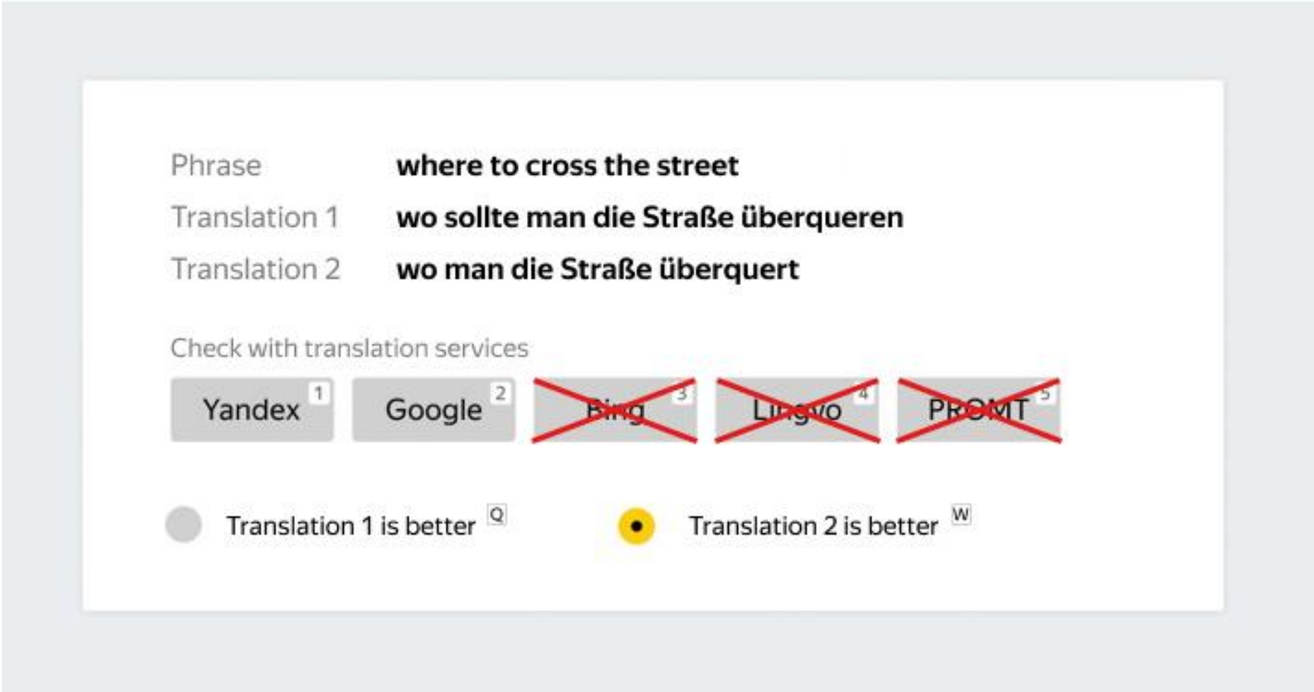
Color contrast too bright

Reasonable space usage



Only necessary elements

- Task is to evaluate which translation is better

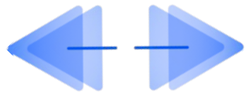


— Multiple tasks on one page help to save time on switching between pages.

— Put as many as can be completed in 5 minutes.

Task suite

Same width
of task blocks



Avoid empty
spaces between
blocks

2–3 task
in a row

— Test your task before launching!

Quality control

Quality control

- “Before” task performance
  - Selection of performers
  - Well-designed instruction
- “Within” task performance
  - Golden set (aka honey pots)
  - Well-designed interface
  - Motivation (e.g. performance-based pricing)
  - Tricks to remove bots and cheaters (e.g. quick answers)
- “After” task performance
  - Post verification (acceptance)
  - Consensus between performers and result aggregation

Selection of performers

- Filter by static properties (e.g. education, languages, citizenship, etc.)
- Filter by computed properties (e.g. browser, region by phone/IP, etc.)
- Filter by skills
  - To select proper specialization
  - To control quality level on your tasks
  - To get performers with best quality on past projects
- Educate to perform your tasks
  - Use training tasks to show how to perform tasks
  - Use exam tasks to evaluate education level

Golden set (aka honey pots)

- Tasks with known correct answer shown to performers to evaluate their quality

  - Distribution of answers in golden set = distribution in whole set of tasks
  - But should contain rare answer variants with higher frequency
  - Refresh your set of honey pots regularly to avoid bots and cheating
  - Automatic golden set generation via performers:
    - Tasks with answers of high confidence
    - (e.g. aggregation of answers from a large number of performers)

Best practices

Motivation

- Bonuses for a good quality within a period
- Gamification (e.g. achievements, leader boards, etc)
- Price depends on quality

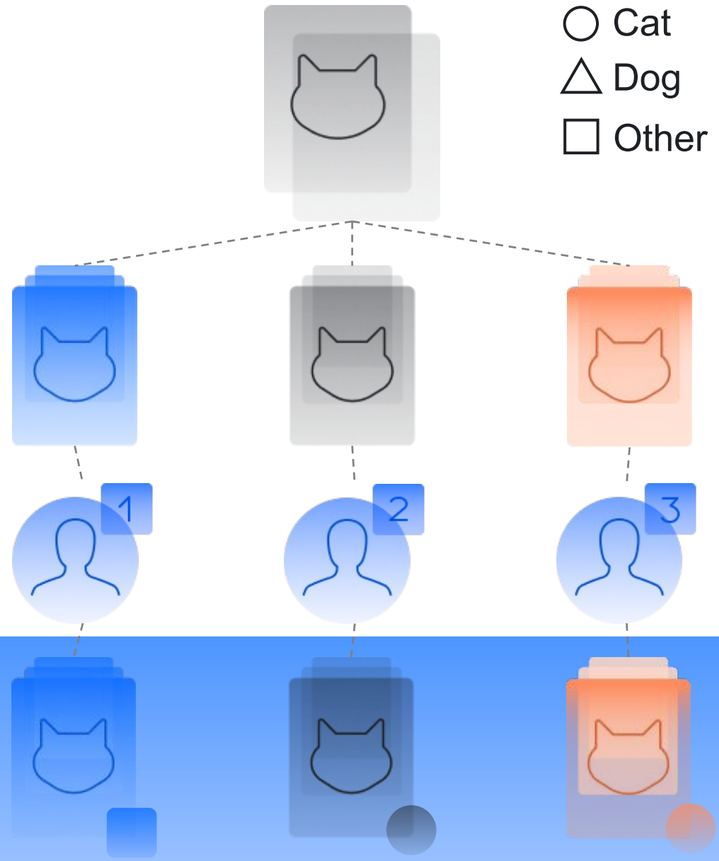
Tricks to remove bots and cheaters

- Control fast responses
- Check whether a link has been visited
- Check whether a video has been played
- etc

Post verification (acceptance)

- A performer gets paid only if his answer is accepted
  - Is used when a task is sophisticated (neither golden set nor consensus models work)
  - Can be performed on your own, but
- You can use other crowd performers via a task of different type
  - Thus, you deal with hierarchy of projects (you apply decomposition)

Consensus between performers



Upload multiple copies of each object to label

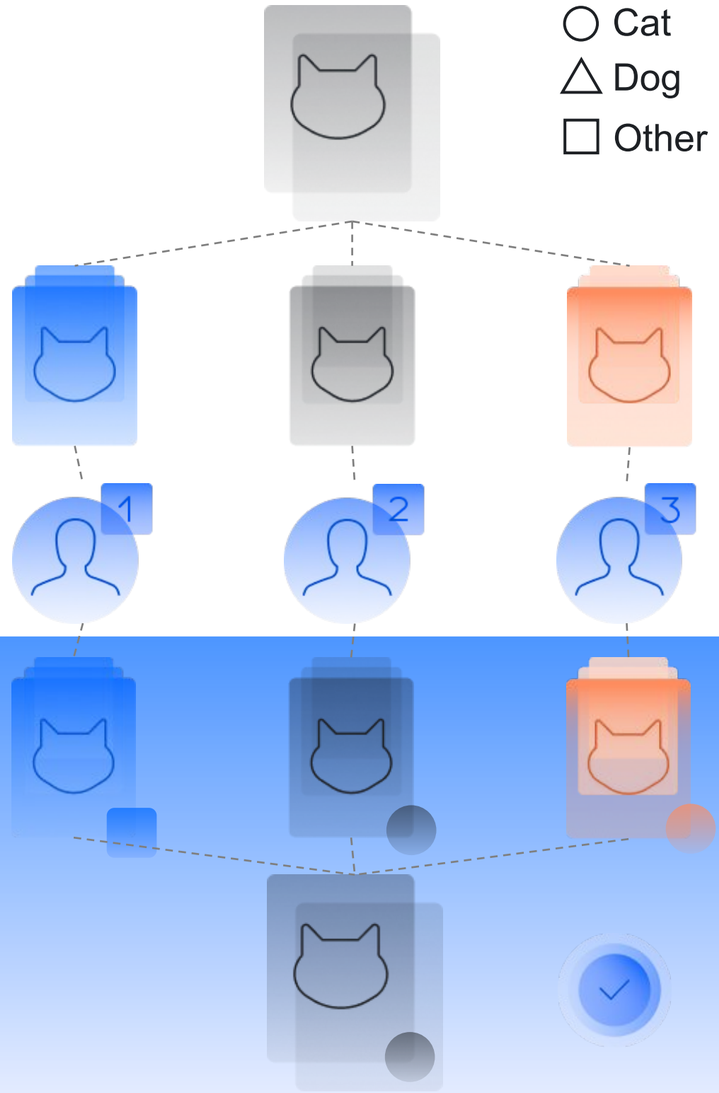
Performers assign noisy labels to objects

Derive quality from consensus

Works well only if most performers have good quality

Aggregation

Aggregation



Upload multiple copies of each object to label

Performers assign noisy labels to objects

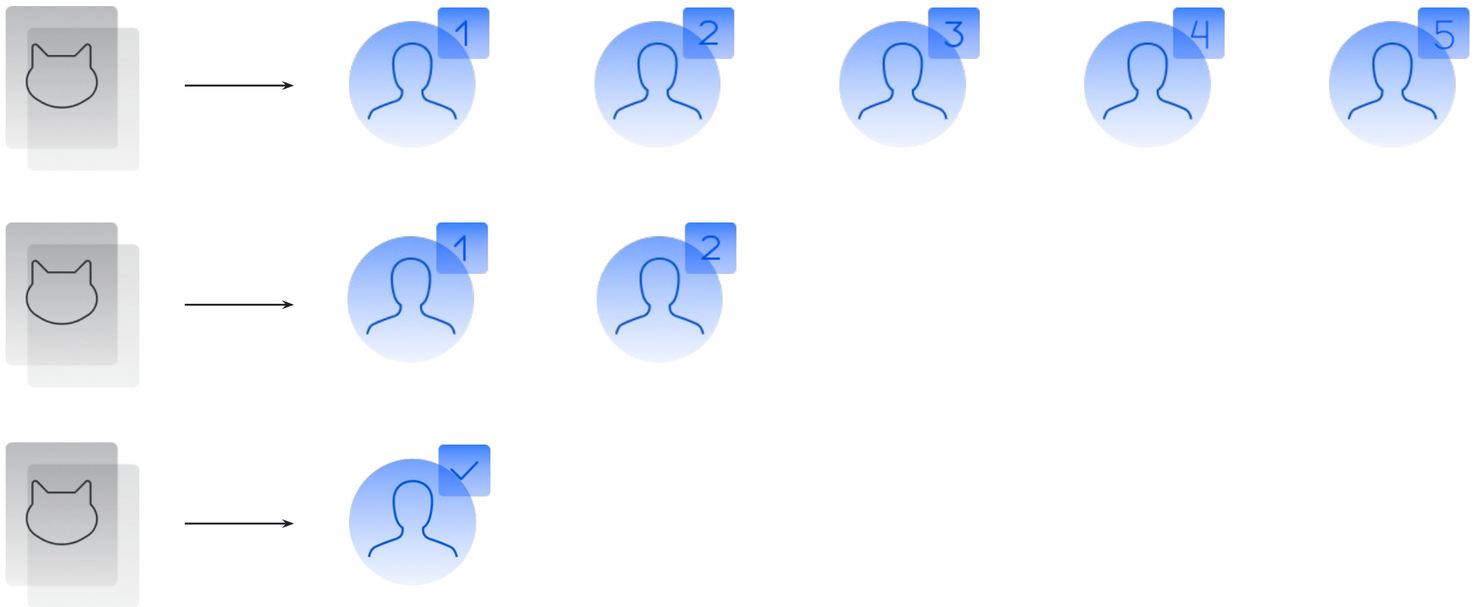
Aggregate multiple labels into a more reliable one

The simplest way:
–Assign the most popular answer (Majority Vote)
–There are more sophisticated methods

Incremental relabeling & pricing

Incremental relabeling

Obtain aggregated labels of a desired quality level using a fewer number of noisy labels



Pricing depends on

- Task design
  - Payment is made per a batch of microtasks (aka a task suite)
  - Time required to perform a task: control hourly wage
- Market economy aspects
  - The lower supply of performers is (e.g. due to specific skills), the higher price
  - How quickly do you need the accomplished tasks (latency)?
- Result quality
  - Incentivize better performance with a quality-dependent price

IF

Good decomposition

THEN

Simple instruction

Easy to use task interface

Performers do tasks with better quality

Easy to control quality

Standard aggregation models work well

Easy to control and optimize pricing

Questions

Join our Slack: **recsys_2022**

Nikita Pavlichenko

Researcher

pavlichenko@toloka.ai

https://toloka.ai/events/recsys-2022/

https://bit.ly/3eYIX2P