Yandex

Yandex

NAACL-HLT 2021 hands-on tutorial

Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial

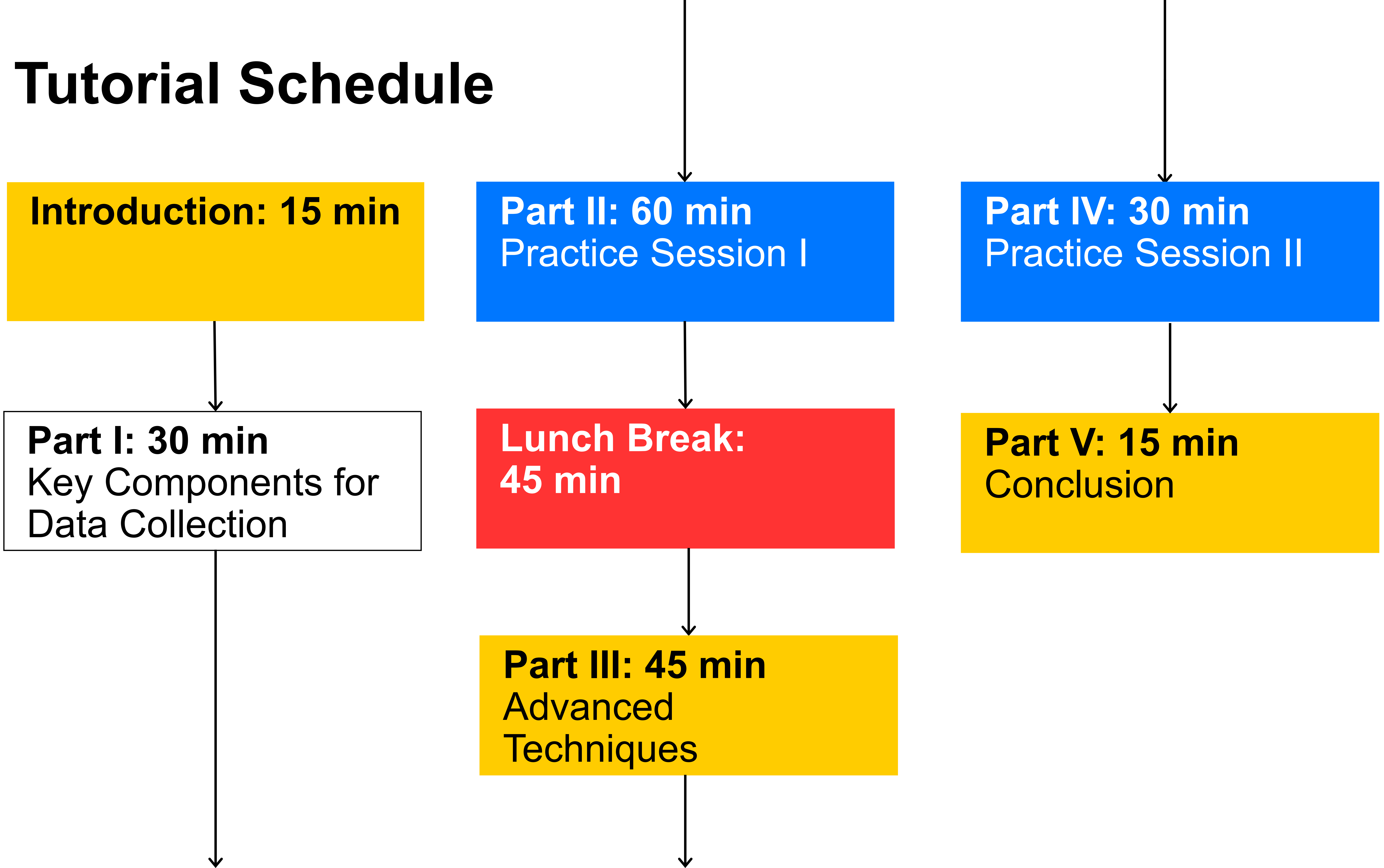
Alexey Drutsa, Dmitry Ustalov, Valentina Fedorova, Olga Megorskaya, Daria Baidakova

Part I

Key Components for Efficient Data Collection

Alexey Drutsa,
Head of Efficiency and Growth Division

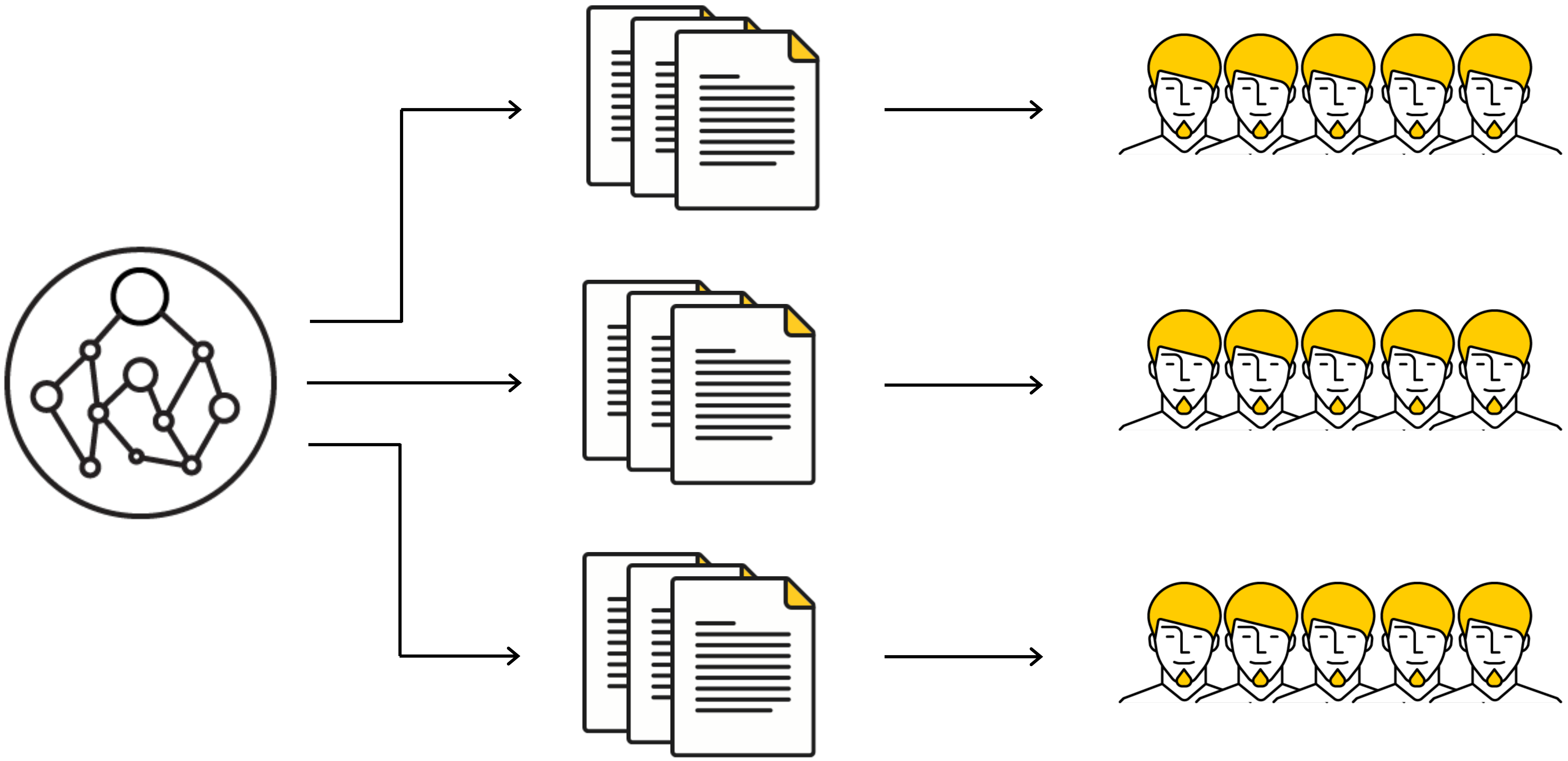
Tutorial Schedule


Introduction: 15 min
Part I: 30 min Key Components for Data Collection
Part II: 60 min Practice Session I
Lunch Break: 45 min
Part III: 45 min Advanced Techniques
Part IV: 30 min Practice Session II
Part V: 15 min Conclusion

Decomposition
Instruction
Task interface
Quality control
Aggregation
Incremental Relabeling
&
Pricing

Decomposition

Decomposition



Decomposition: why?

Performers are usually non-specialists in your specific task

The simpler a single task is:

- the more humans can perform your task
- the easier its instruction
- the better quality of performance

A way to:

- distinguish tasks with different difficulty
- control and optimize pricing
- control quality by post verification

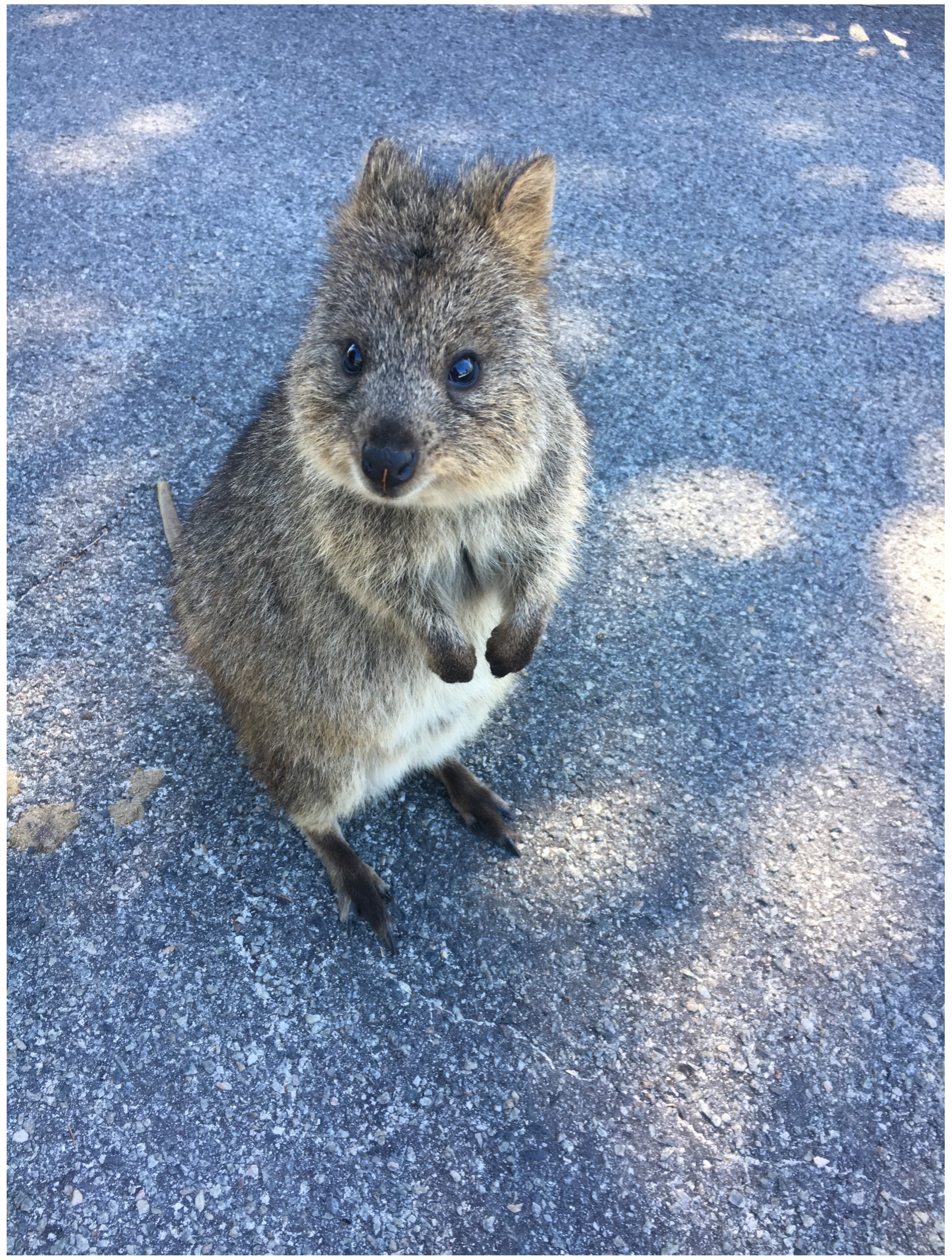
Decomposition: when?

If

- your task requires an answer selected among more than 3-5 variants
- your task has a long instruction hard to read

then your task requires decomposition

Case of decomposition: a lot of questions



All questions in one task

Bad practice

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

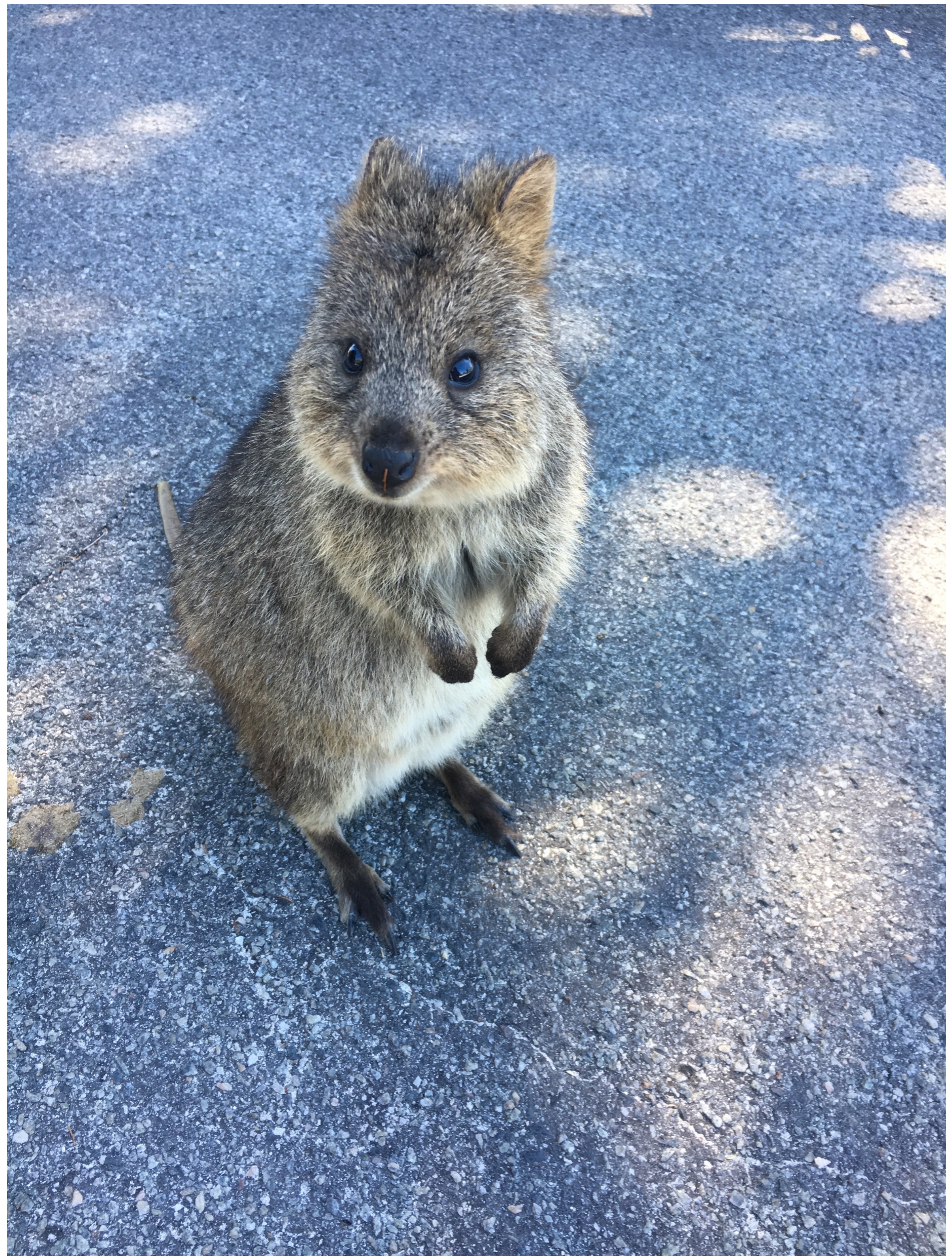
Is its tail visible?
- Yes
- No

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



Each question in a separate task Good practice

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

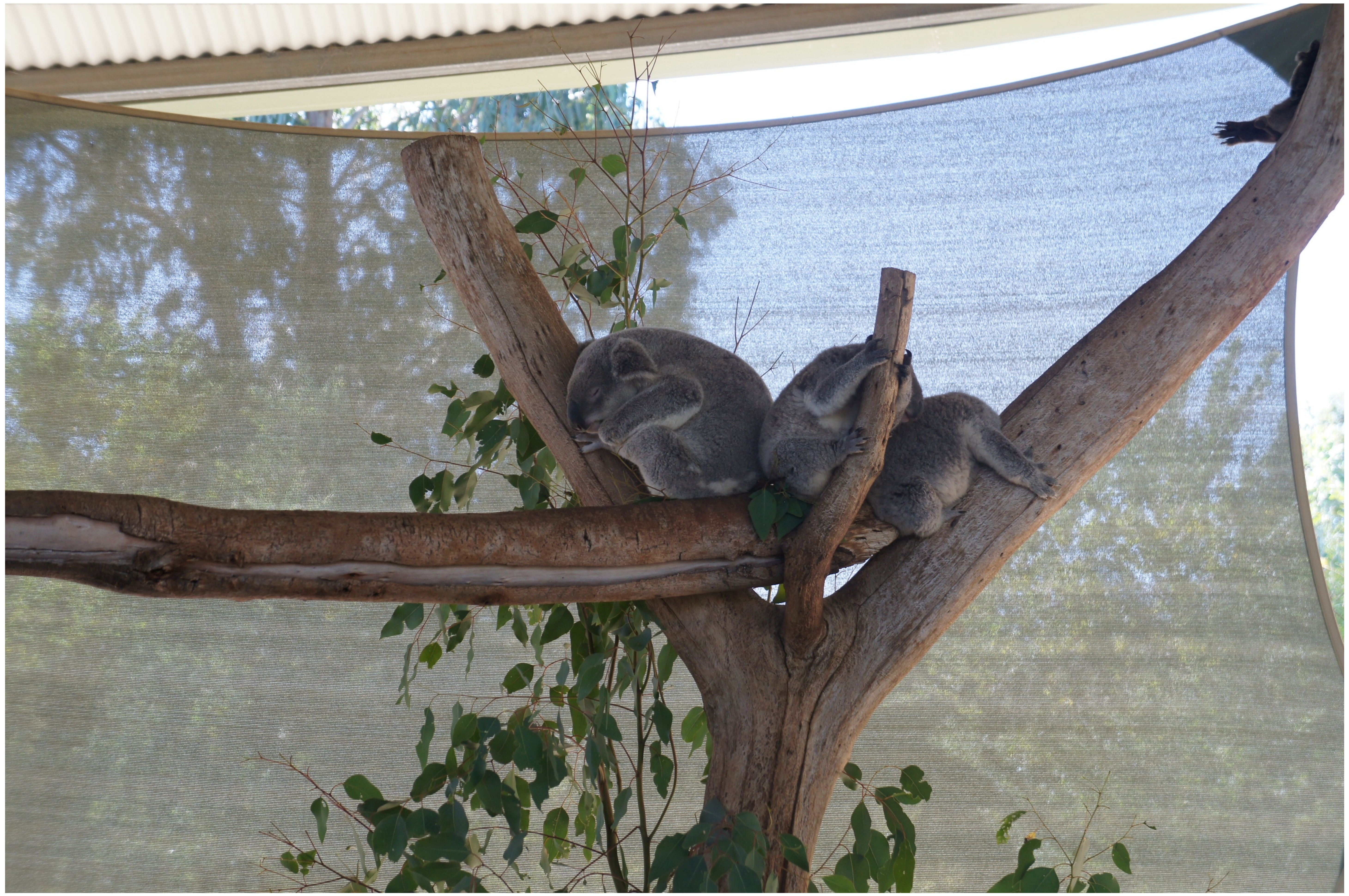
Is its tail visible?
- Yes
- No

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 make:

- comparison with control answers
- aggregation of answers from different performers

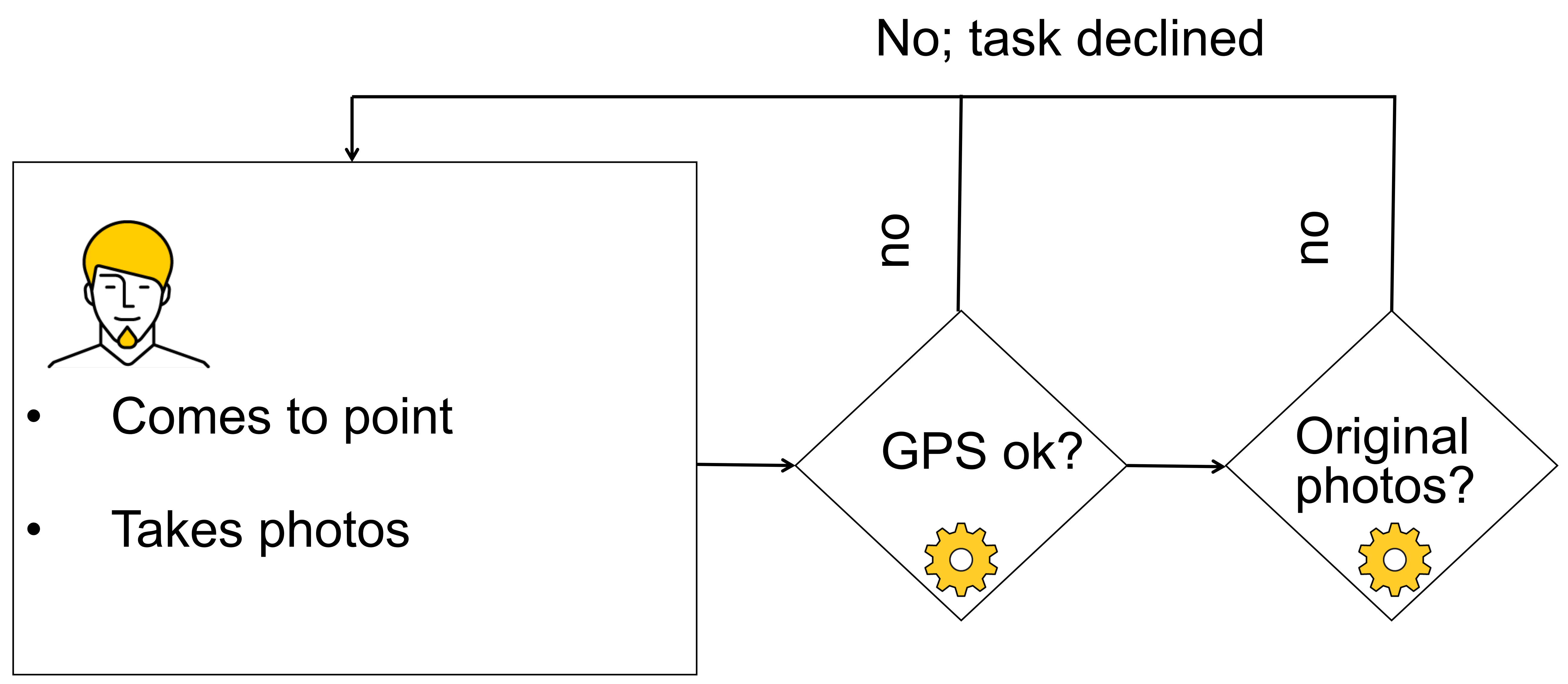
A good solution

A task for another performer:
Is the highlighting of all koalas made correctly?

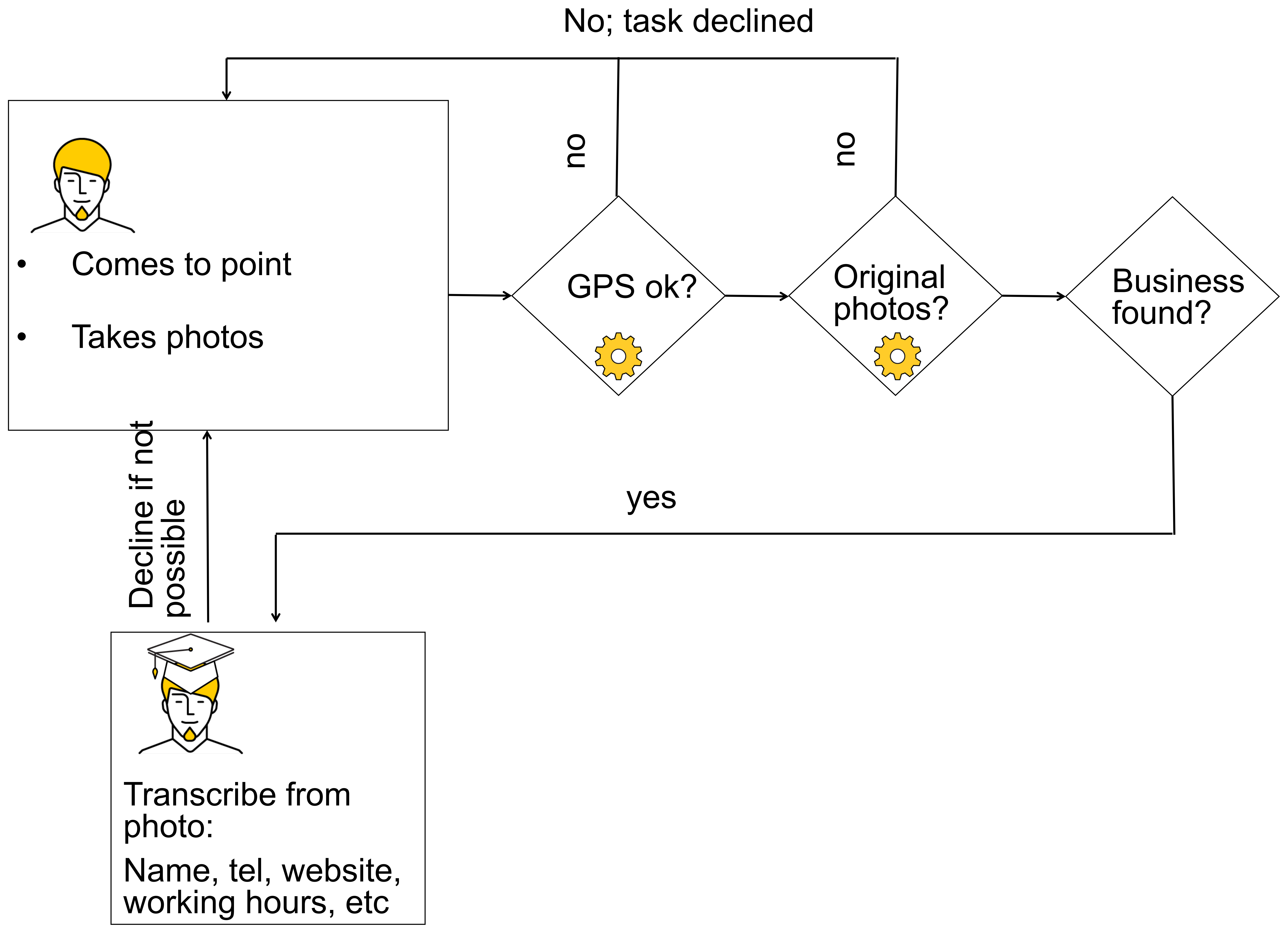
Real example:
decomposition for an offline data collection task

- Comes to point
- Takes photos

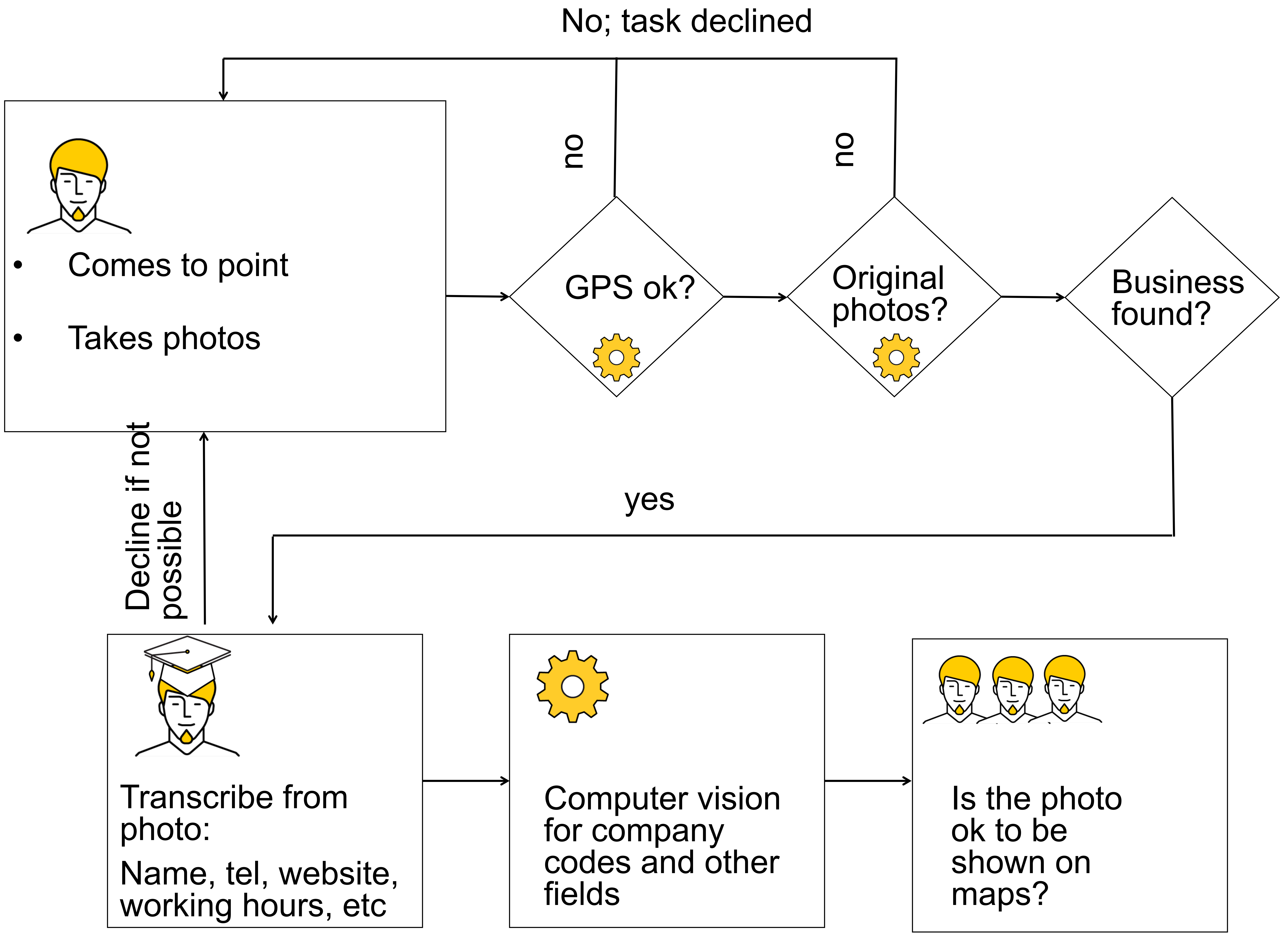
Final result: verified business



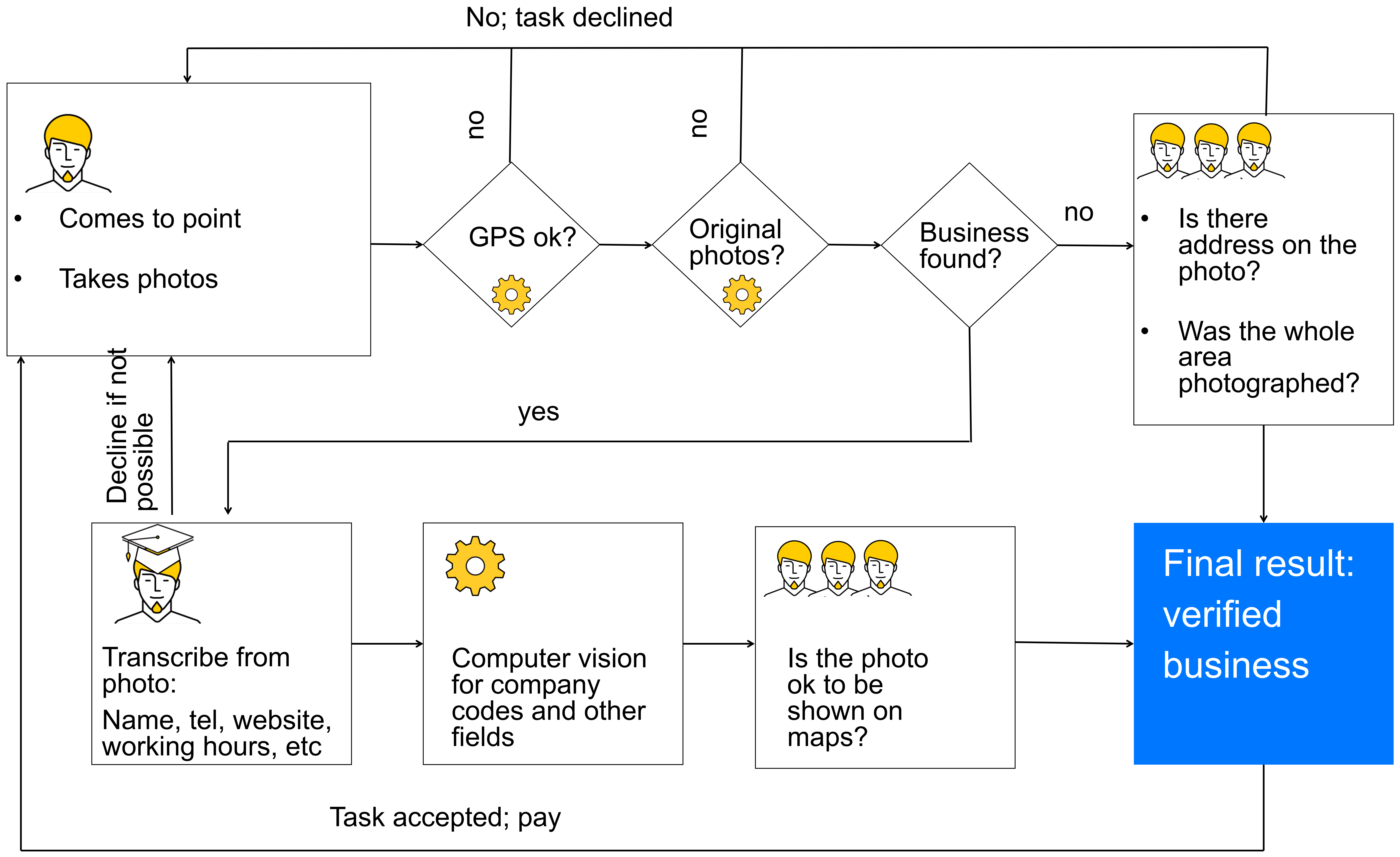
Final result: verified business



Final result: verified business



Final result: verified business


No; task declined
Comes to point
Takes photos
no
GPS ok?
no
Original photos?
Business found?
no
Is there address on the photo?
Was the whole area photographed?
yes
Decline if not possible
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

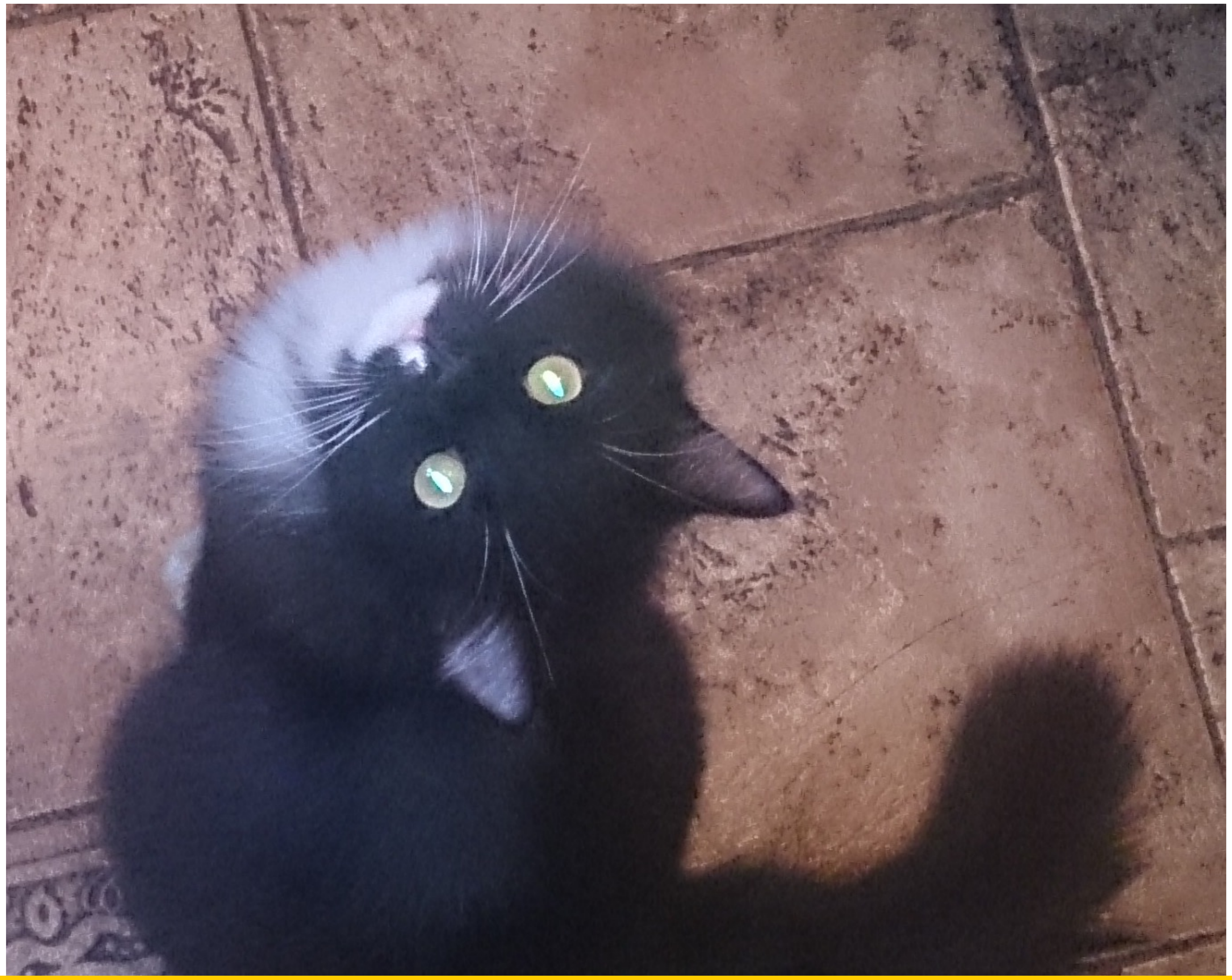
Most pitfalls are there

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No



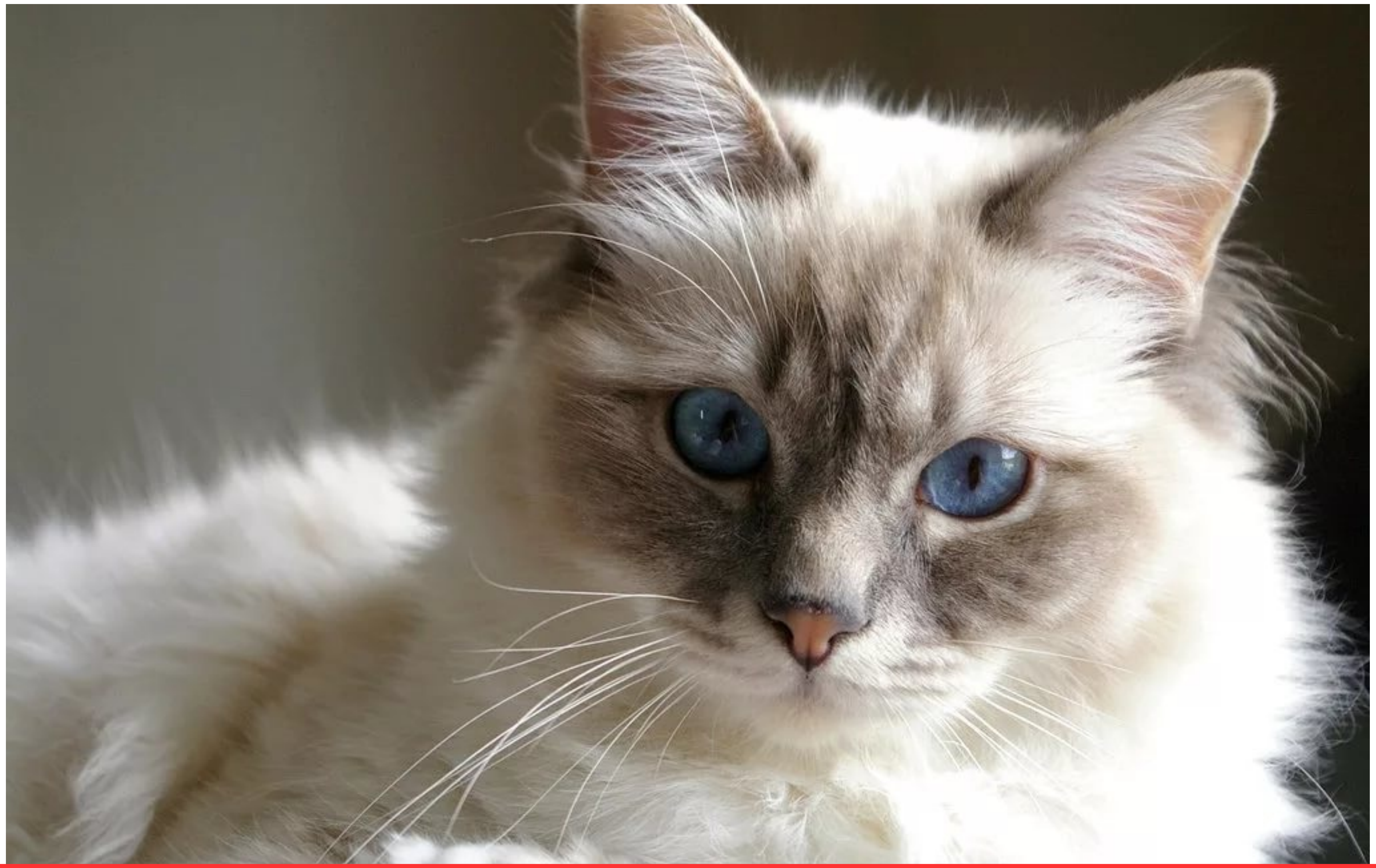
OK: the answer and the task seem clear

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No

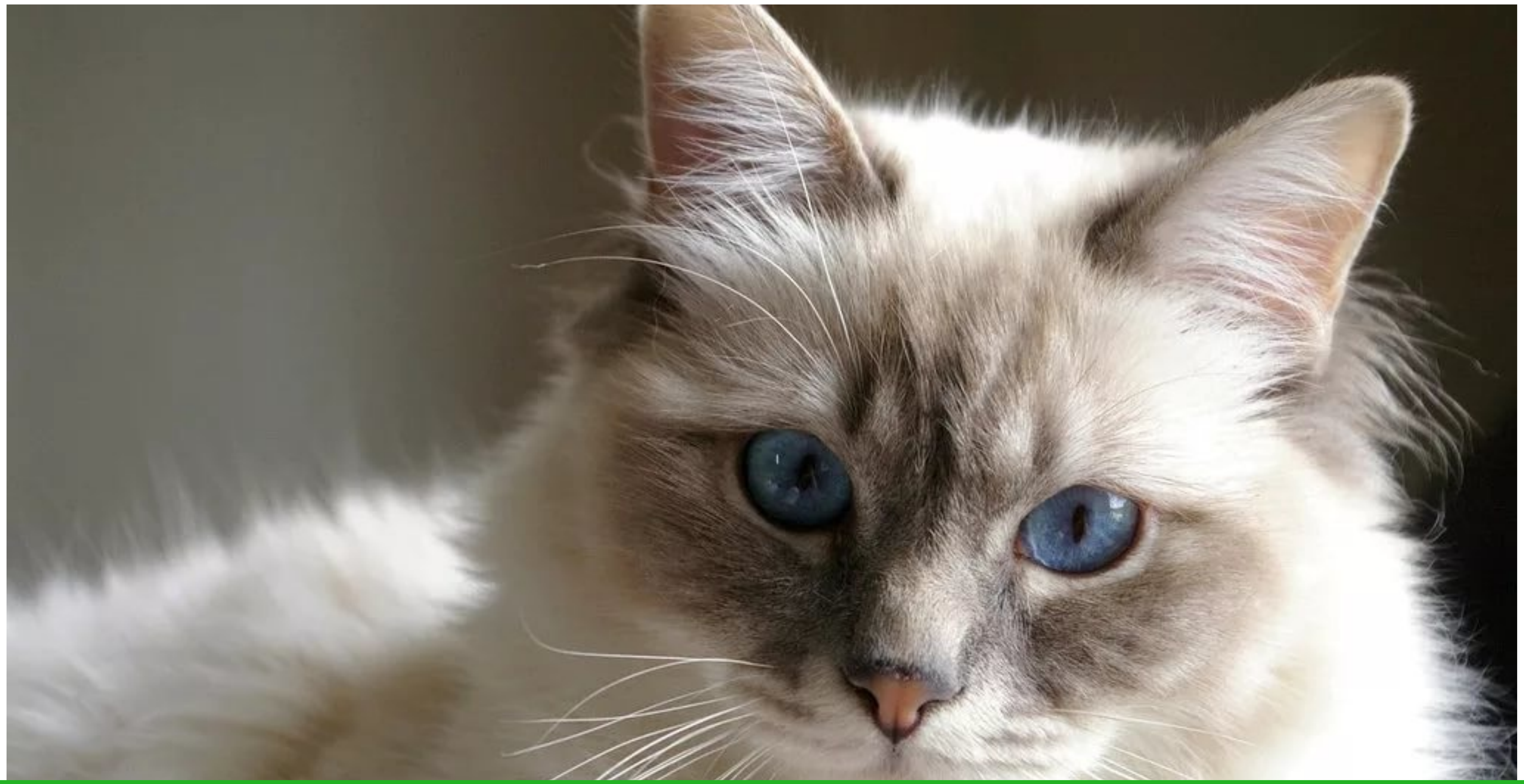


What is the correct answer?

Instruction ambiguity for a rare case: example

Is this cat white?

Yes
No



How to fix:

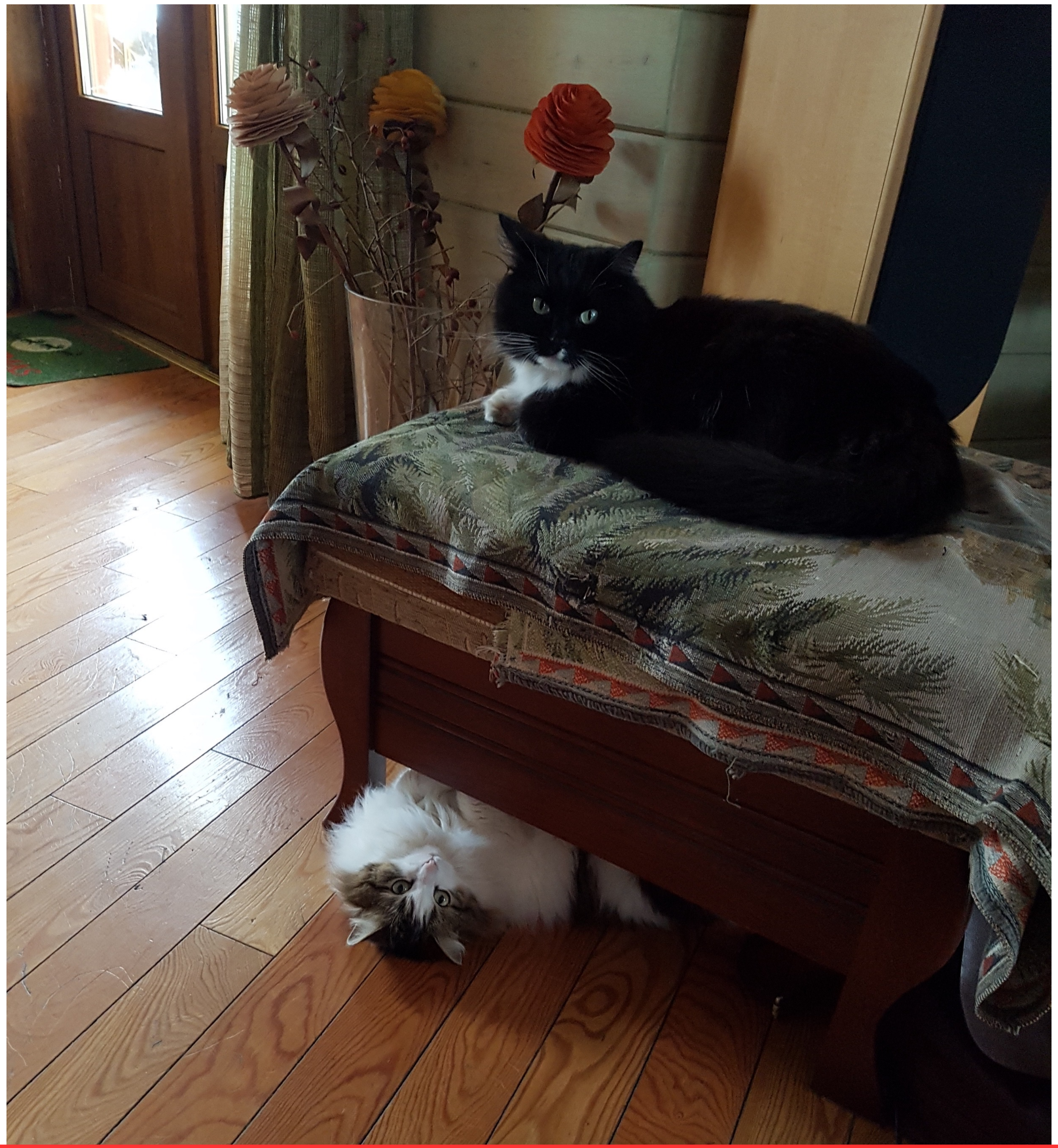
- In the instruction: clarify what you mean under “a white cat”

Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No



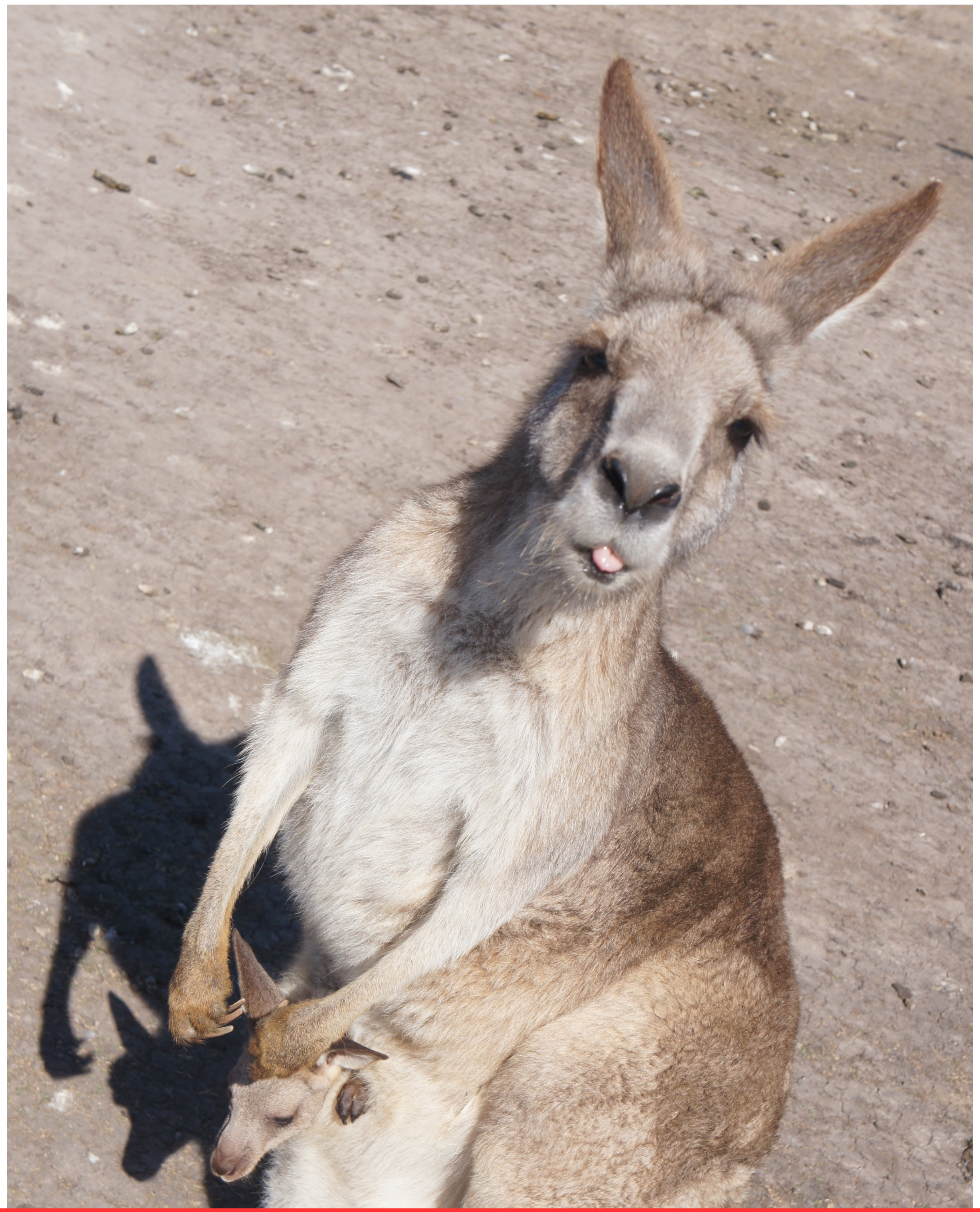
Rare case: many cats

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

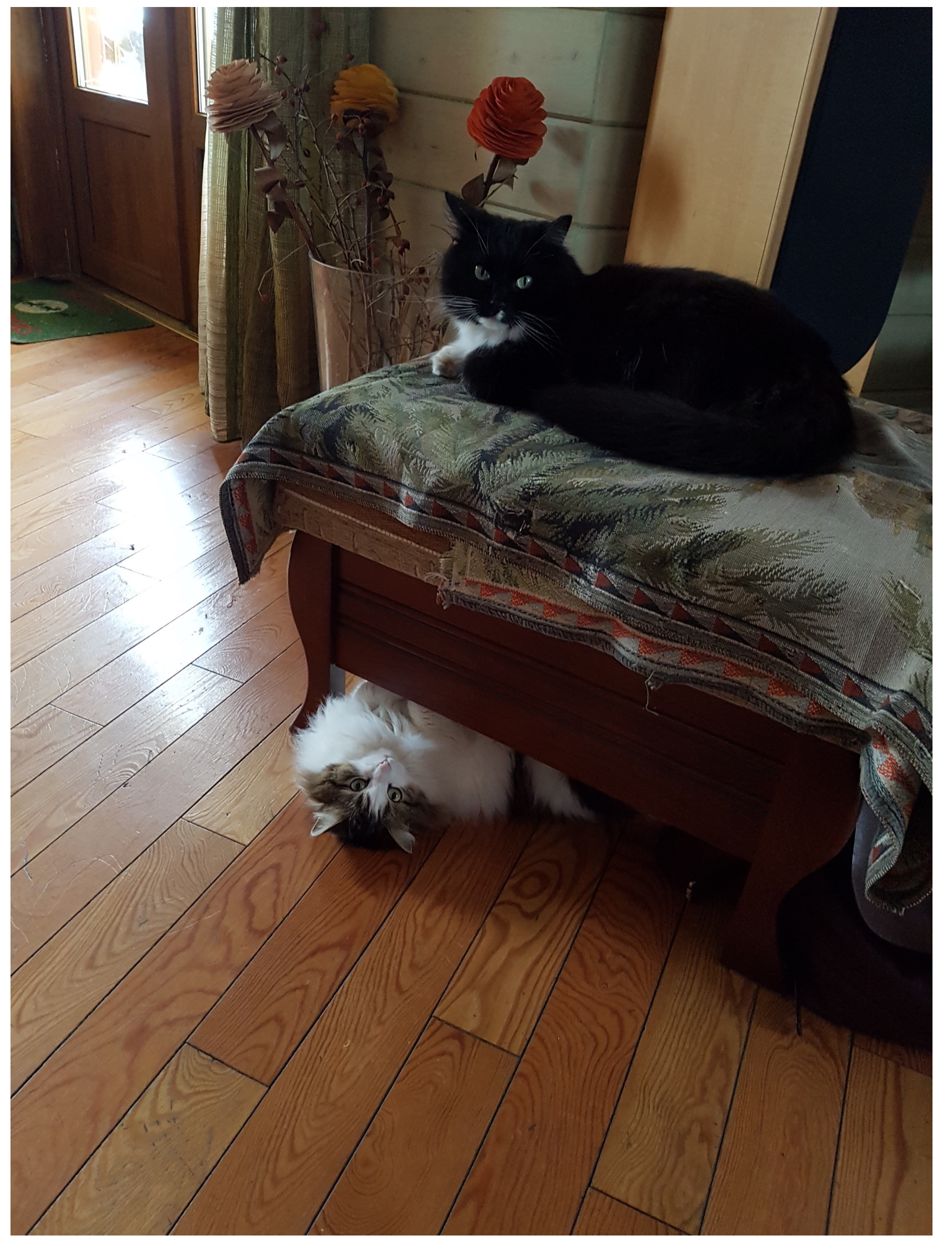
Rare case: image has not been shown

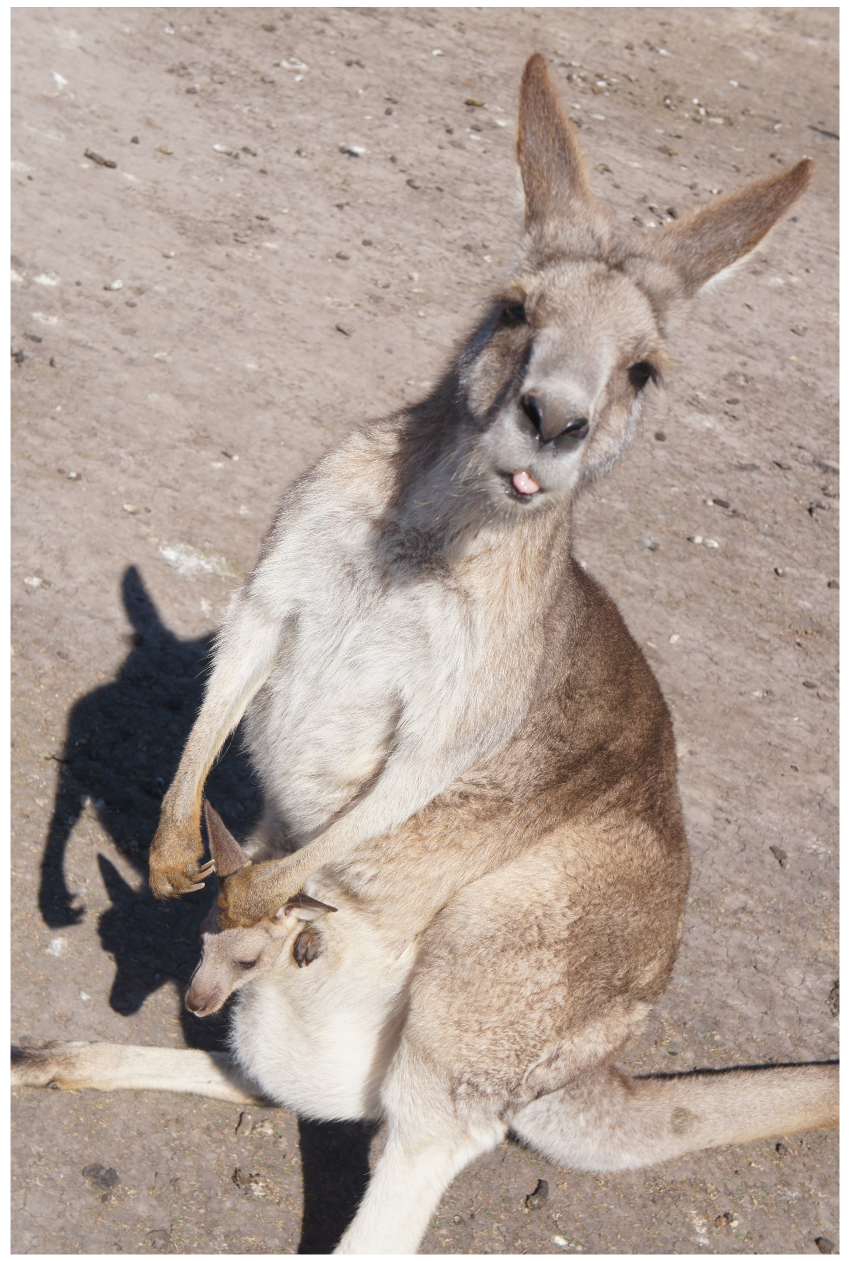
Instruction ambiguity for a rare case: example

Is this cat white?

Yes

No





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

Task interface: summary on best practices

For faster performance:

- Hot key combinations for checkboxes / radio buttons / buttons
- Reduce navigation to third-party sites
- Effective composition of a task template
- Optimal position of tasks on a page

For better quality and less errors:

- Dynamic interface (show/hide input controls depending on user actions)
- Adaptive interface (good view for any device and screen resolution)
- Always test your interface (template testing)
- Dynamic validation of input data (e.g. a text is less than 3 words)

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

Best practices

- 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)

Motivation

- Bonuses for a good quality within a period
- Gamification (e.g. achievements, leader boards, etc)
- Price depending 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 money only if his answer is accepted

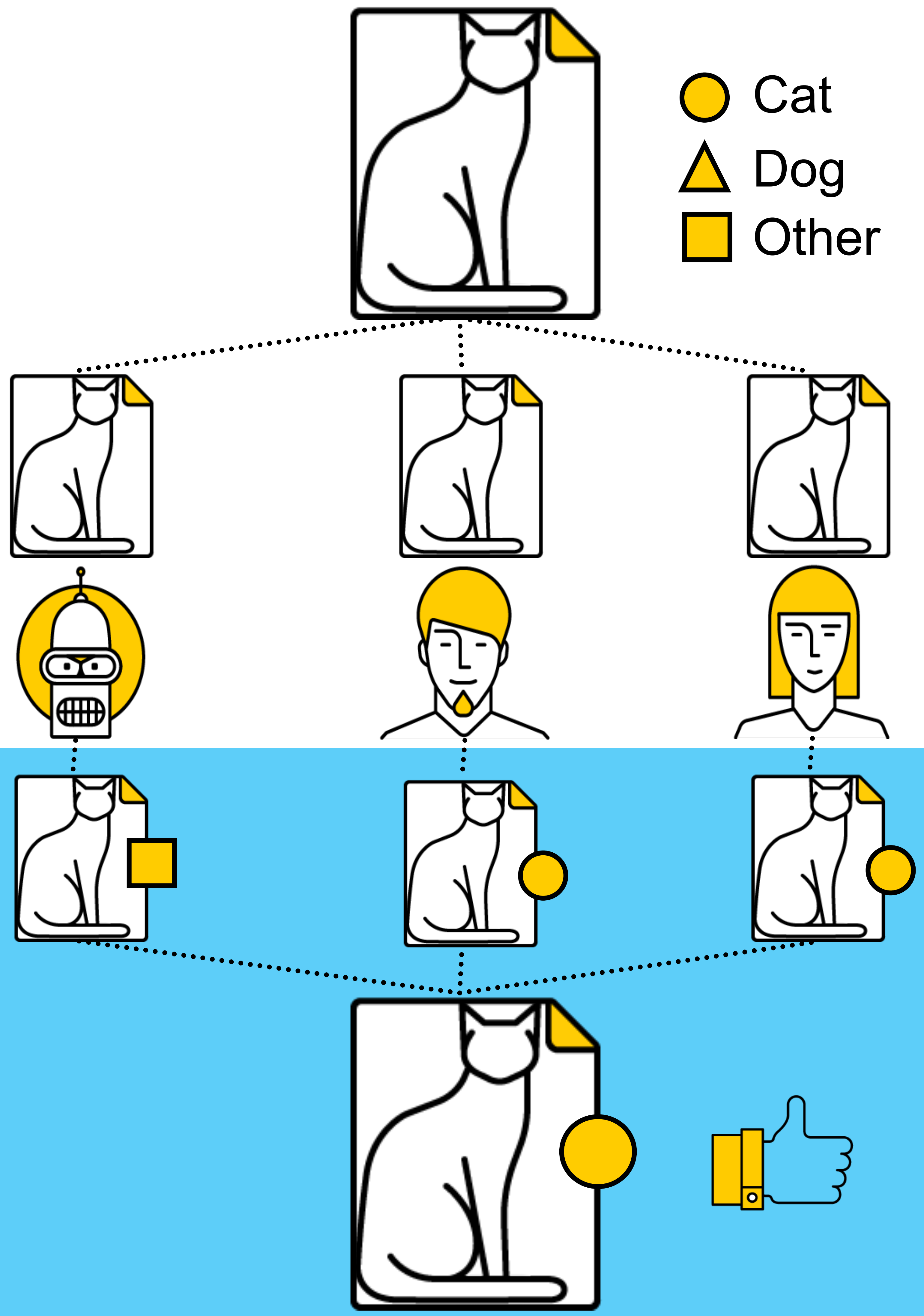
- Is used when a task is sophisticated
  (neither golden set nor consensus models work)

- 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)

Aggregation

Aggregation



Upload multiple copies of each object to label

Performers assign noisy labels to objects

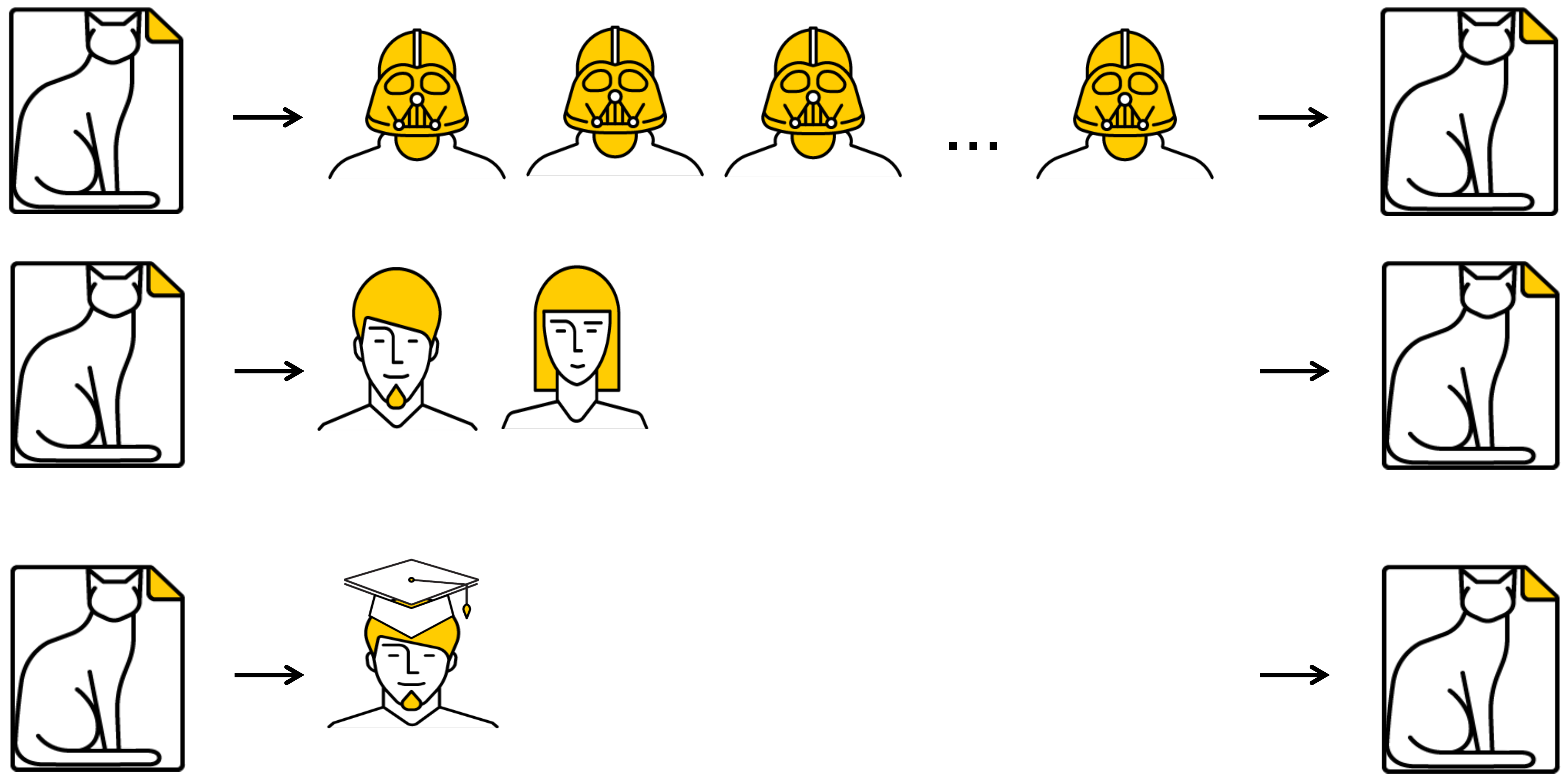
Aggregate multiple labels into a more reliable one

Will be discussed in Part III

Incremental Relabeling & Pricing

Incremental relabeling

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



Several unknown performers

A few performers with known good quality

One expert with high quality

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 accomplished tasks (latency)?

Result quality:

- Incentivize better performance by 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

Yandex

Thank you!
Questions?

Dmitry Ustalov

Analyst/Software Developer
Crowdsourcing Research Group

dustalov@yandex-team.ru

https://research.yandex.com/tutorials/crowd/naacl-2021