Yandex

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

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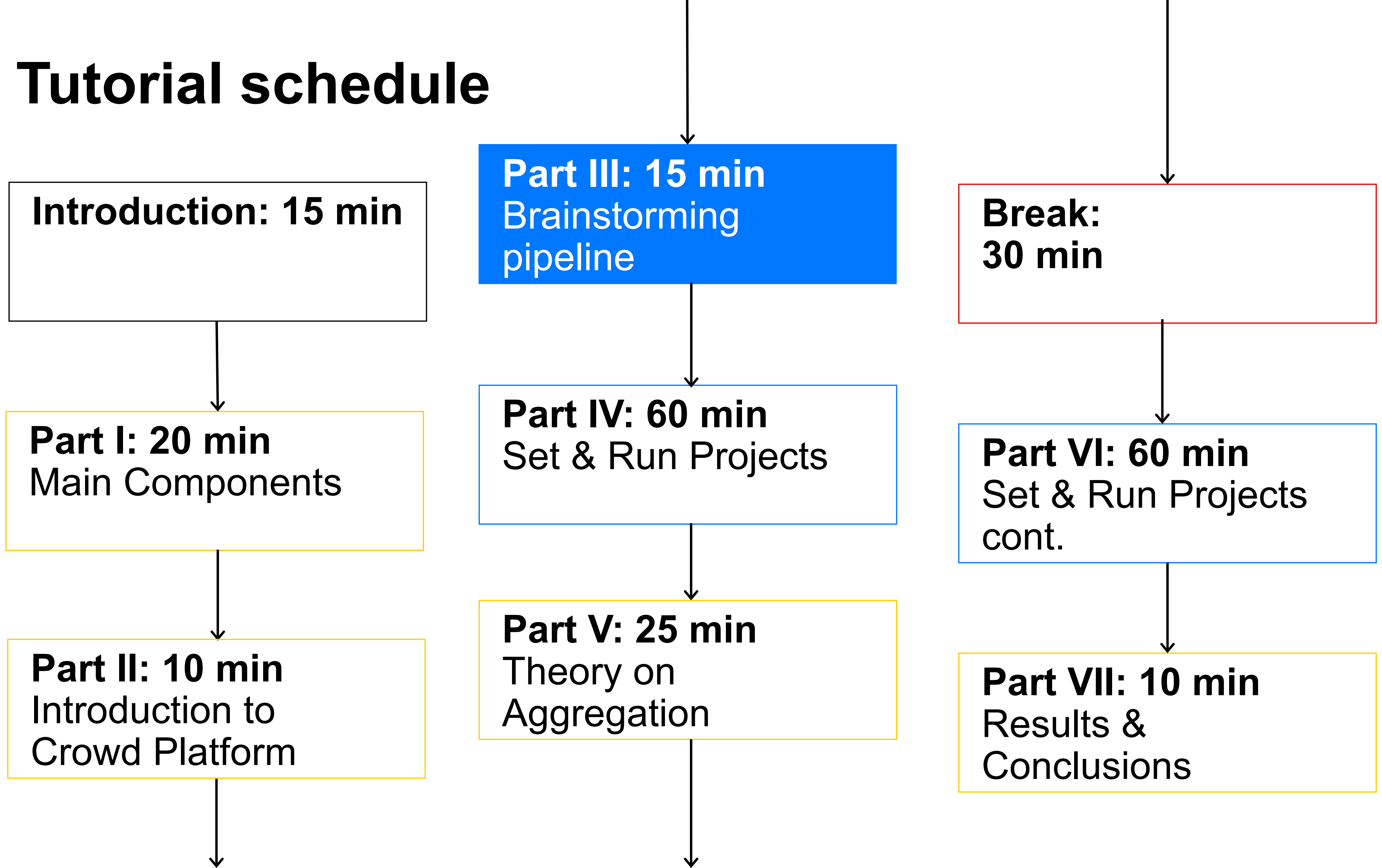
Part III:

Brainstorming the pipeline

Daria Baidakova,
Project Manager

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

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: 60 min
Set & Run Projects
cont.

Part VII: 10 min
Results &
Conclusions

Practice session

Our practice session will consist of three parts:

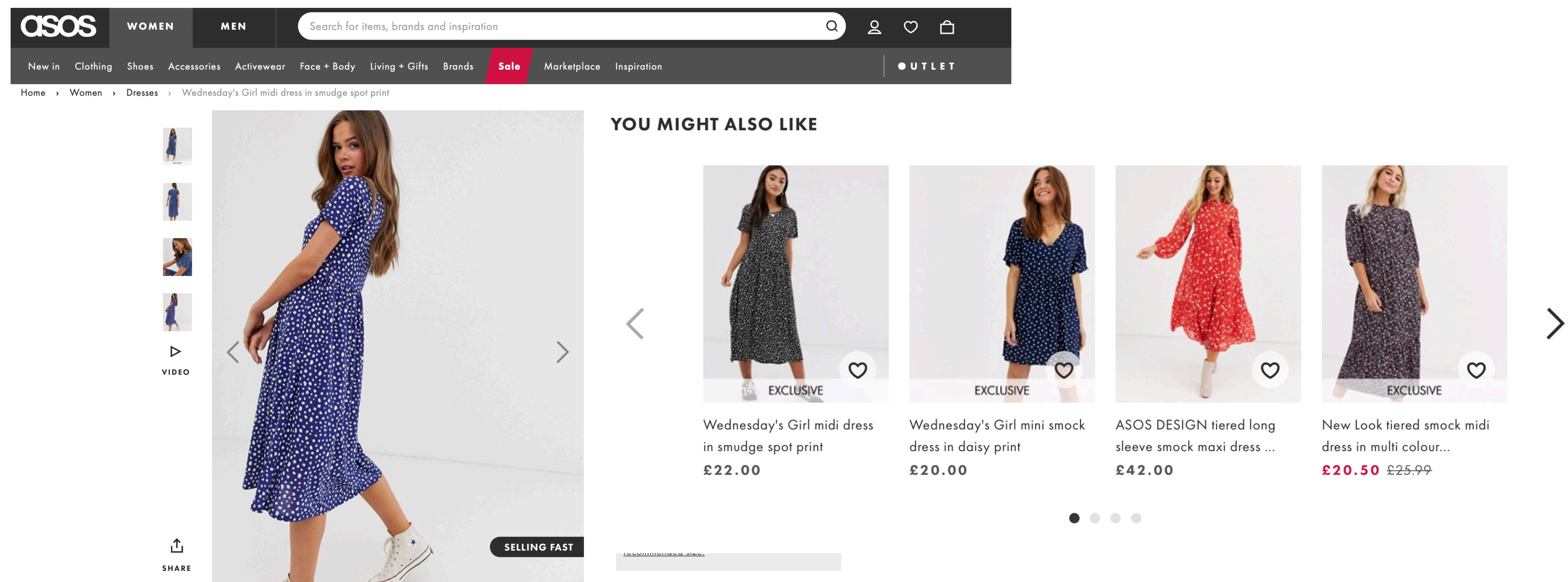| Part I (now)	Part II (in 15 mins)	Part III (in 120 min)
Think and discuss how you would design a crowdsourcing pipeline	Run the best-practice pipeline on a real crowd on Yandex.Toloka	Complete the pipeline on Yandex.Toloka

Practice session: scope

Imagine that you develop a machine learning pipeline
to help improve search quality at an online store to find substitutes

Practice session: scope

**Imagine that you develop a machine learning pipeline
to help improve search quality at an online store to find substitutes**

> You have a dataset of pictures with people wearing different clothes

> You need to find a better substitute for the initial item in an image

> These collected data will further be used to train a search algorithm

It is your goal during the practice session of our tutorial

Dataset under study: pictures of people wearing different clothing items

Items to be matched in photos

Each photo may contain clothing items of different types, for example:

› Hats
› Shirts
› Jackets
› Coats
› Jeans
› Pants (trousers)
› Bags
› Sunglasses
› Other items

During your practice:

Choose
one type of items
you want to find substitutes for
in the photos.

For example: Shoes

Formal setup: find the best substitute item

› Each clothing item of a selected type
› in each photo from the dataset
› needs to be matched by a substitute item

**Let us do it via crowdsourcing**

During your practice:

Think how you would design a crowd pipeline to find the best substitute!

Example: I decided to find the best substitute for the shoes, so my pipeline would be like..
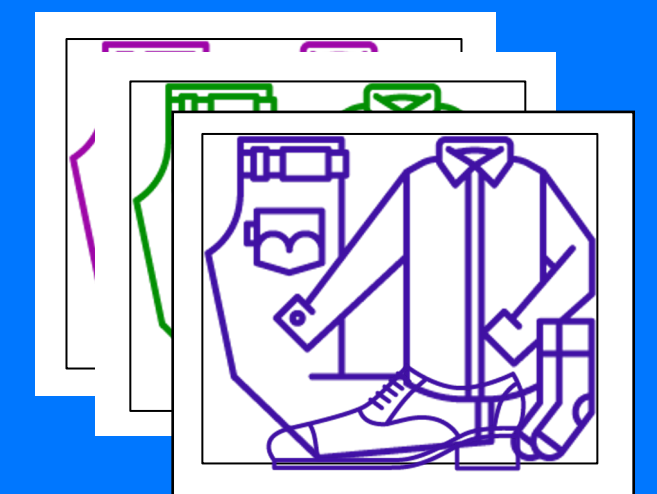
Practice: you have 7 mins now

1. Assume we need to find a best matching substitute to a clothing item on every photo
2. Design a crowdsourcing pipeline for it

How to get desired substitutes from a crowd?
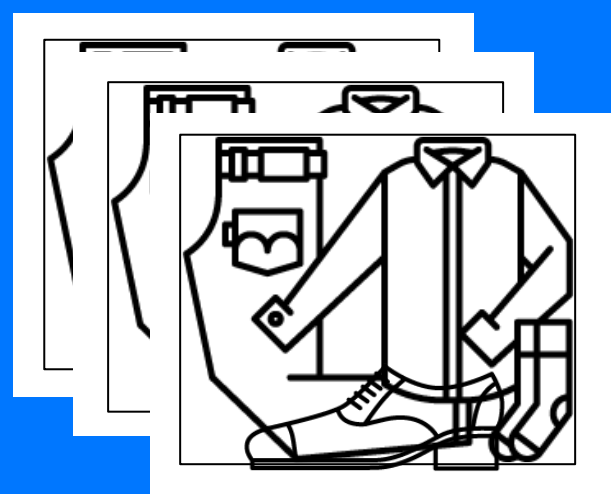
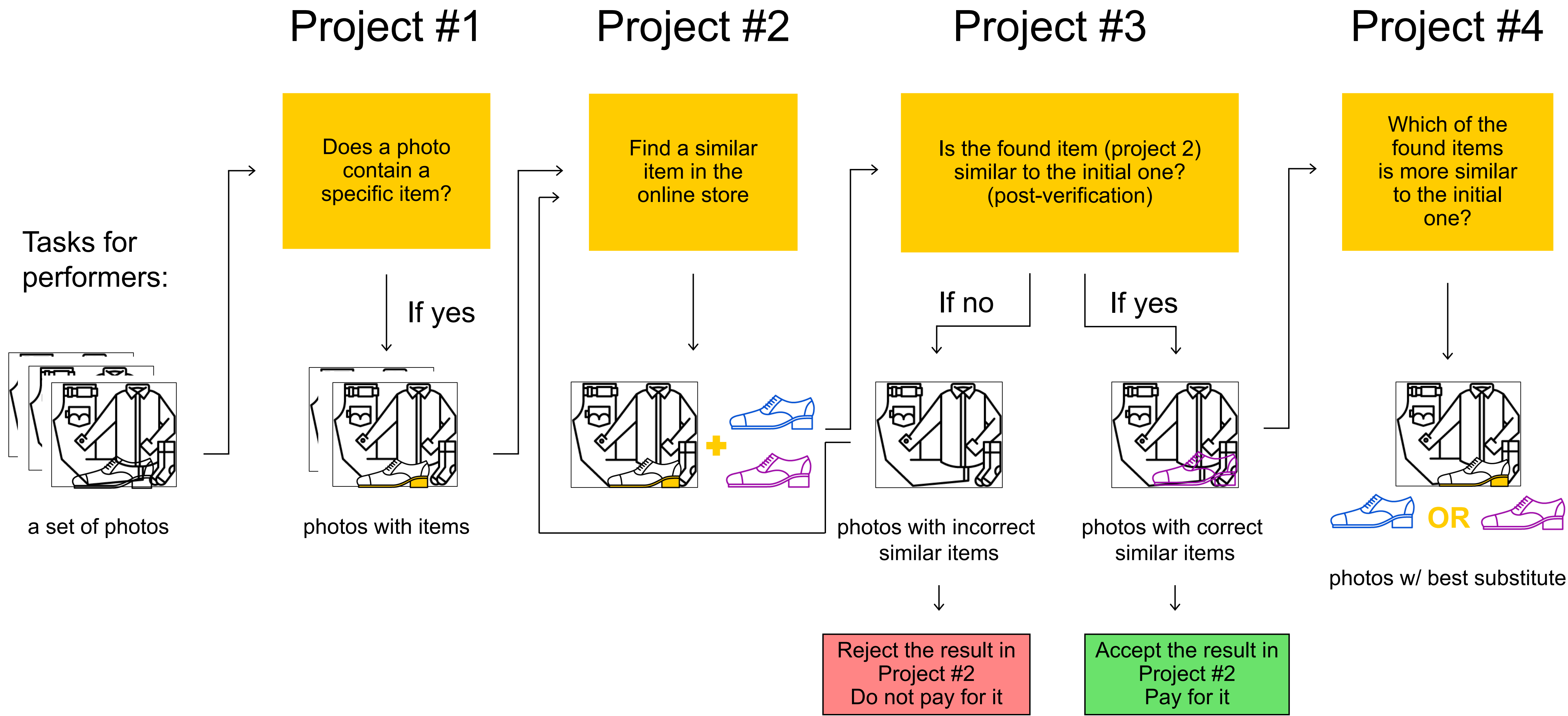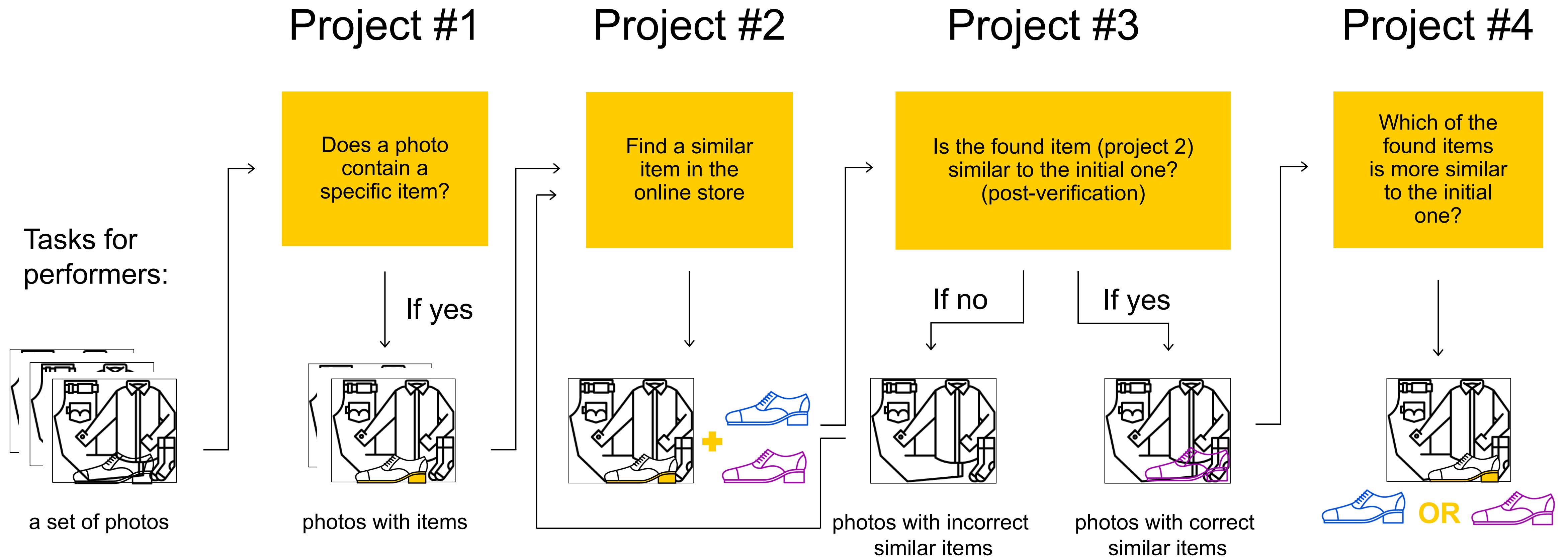a set of photos with clothing items   a set of different substitutes

Hint: discuss in groups how to
> decompose the problem
> make tasks simple
> control quality

Suggested pipeline

We suggest the following pipeline

Project #1

Project #2

Project #3

Project #4

Tasks for performers:

Does a photo contain a specific item?

Find a similar item in the online store

Is the found item (project 2) similar to the initial one? (post-verification)

Which of the found items is more similar to the initial one?

If yes

If no

If yes

a set of photos

photos with items

photos with incorrect similar items

photos with correct similar items

OR

photos w/ best substitute

Reject the result in Project #2 Do not pay for it

Accept the result in Project #2 Pay for it

We suggest the following pipeline

Project #1   Project #2   Project #3   Project #4

Tasks for performers:

Does a photo contain a specific item?

Find a similar item in the online store

Is the found item (project 2) similar to the initial one? (post-verification)

Which of the found items is more similar to the initial one?

If yes

If no   If yes

a set of photos   photos with items   photos with incorrect similar items   photos with correct similar items

OR

During the practical session we will help you implement and run this pipeline

Project #1: Filter out photos without objects

Task

› Does a photo contain an item of desired type?

Key setting

› Type: classification
› Quality control: golden set
› Overlap: 3 answers per photo
› Pay: $0.01 per a suite of 10 photo

Why?

› Save money: no need to process further photos without desired objects

Are there **shoes** in the picture?

○ Yes   ○ No   ○ Picture not found

Project #2: Searching for similar items on the online store

Task

› Find a similar item on the internet

Key setting

› Type: product photo search

› Quality control: post verification

› Overlap: 3 answers per photo

› Pay: $0.02 per 1 photo

Peculiar properties

› Hard to use golden set and consensus
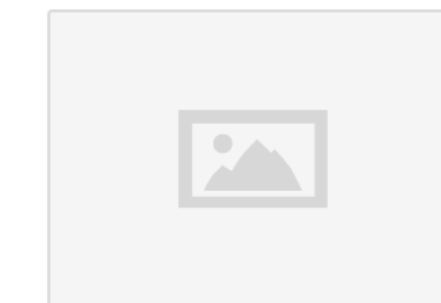
› Results will be verified in Project #3

Find the same **shoes** on Marks and Spencer

Marks and Spencer

Shoes must be the same color and the same style.

Paste the link here

Upload the image here
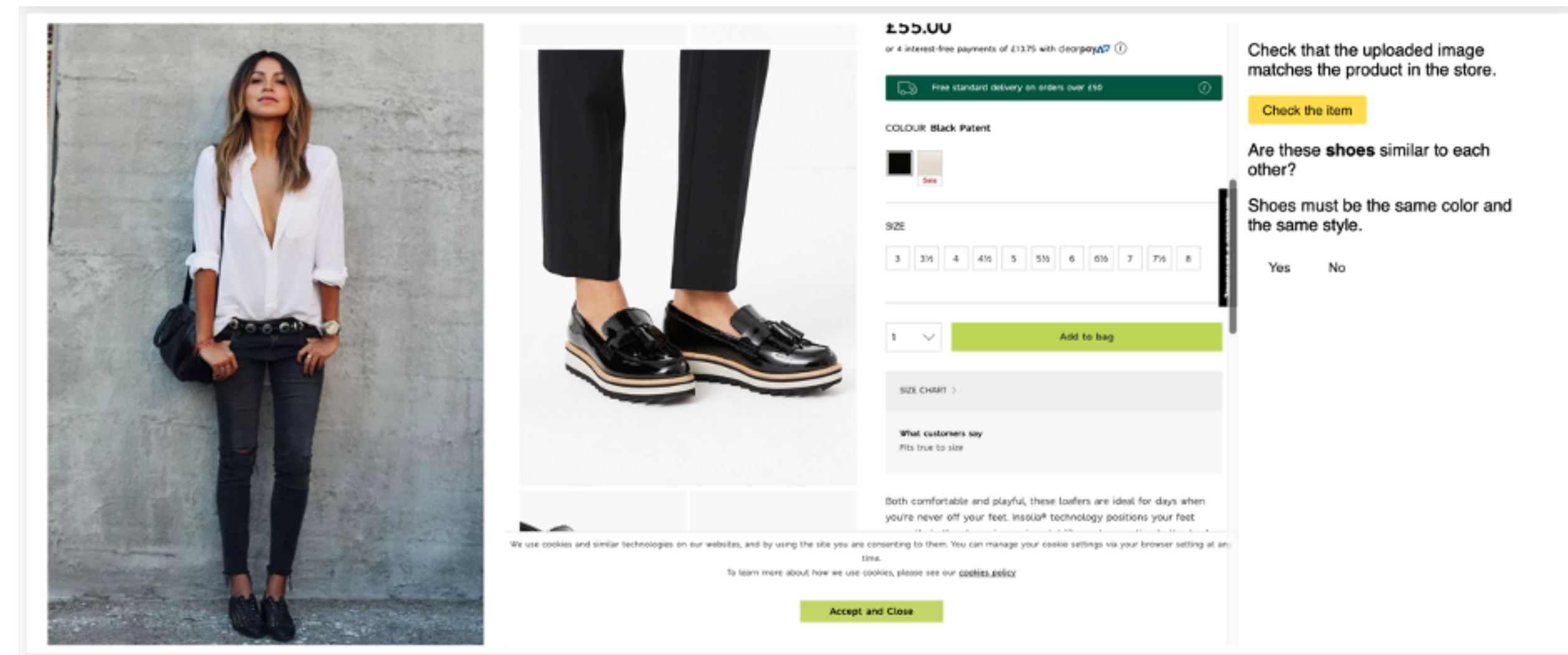
Project #3: Accept correctness of items found

Task

› Does an image contain a requested item?

Key setting

› Type: classification

› Quality control: consensus

› Overlap: 3 answers per photo

› Pay: $0.01 per a suite of 10 photo

Why?

› Need to verify the results obtained from Project #2
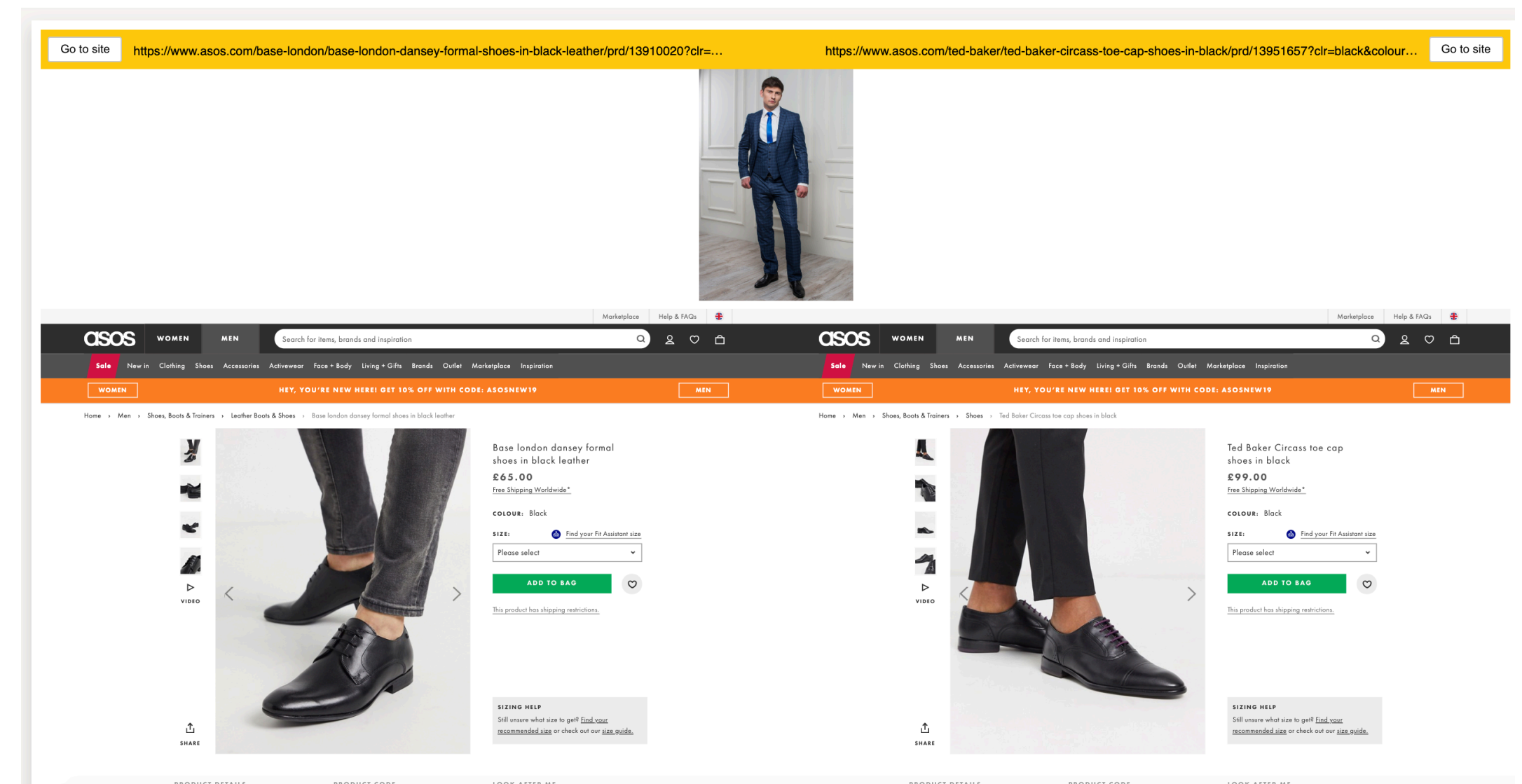
Project #4: Decide which substitute works best

Task

› Which of the items is similar to the initial one?

Key setting

› Type: side-by-side image comparison
› Quality control: consensus
› Overlap: 3 answers per photo
› Pay: $0.01 per a task suite of 10 photo

Why?

› Need to understand which substitute fits best

Yandex

Thank you!
Questions?

Daria Baidakova

Project manager

✉ dbaidakova@yandex-team.ru

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