



Toloka

# Web Engineering with Human-in-the-Loop

Dmitry Ustalov, Nikita Pavlichenko, Boris Tseytlin,  
Daria Baidakova and Alexey Drutsa

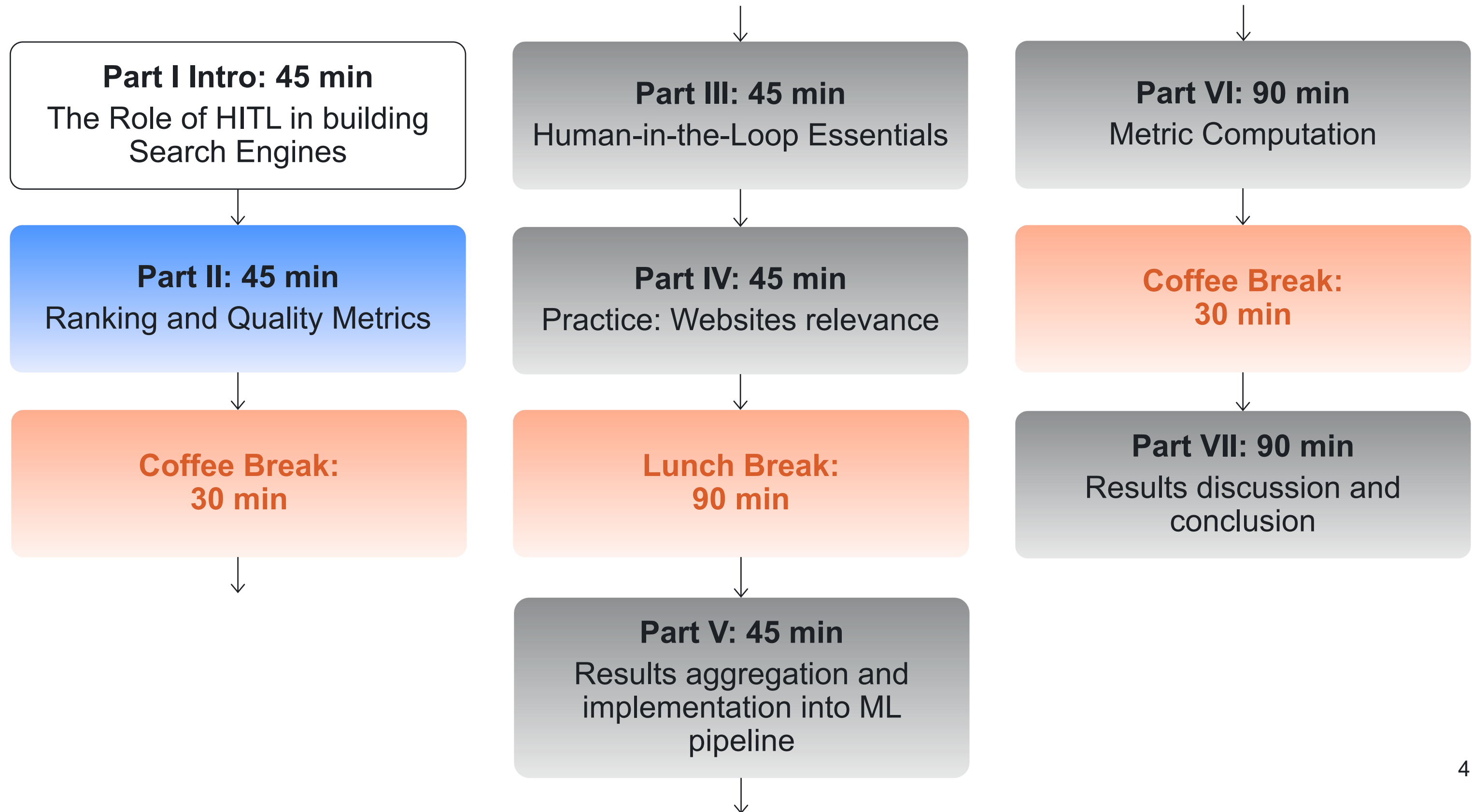


# Part I

# Introduction

Alexey Drutsa,  
deputy CEO, COO at Toloka

# Tutorial Schedule



# **Data-driven web services**

The background features a series of overlapping, curved, semi-transparent shapes in various shades of blue, ranging from a deep navy to a bright, vibrant blue. These shapes create a sense of depth and movement, resembling stylized waves or a modern architectural design. The overall aesthetic is clean, professional, and tech-oriented.

# Cases of leveraging data to provide modern web services and products

## Systems for web services

Recommendation systems

Search relevance

Translation to different languages

Reviews moderation

## Voice Assistants

Speech to Text & Text to Speech

Verifying voice assistant responses

Recording activation phrases

## NLP tasks

Text classification

Sentiment analysis

Intent classification

Named entity recognition

## Field data collection

Verifying addresses

Verifying business hours

Monitoring products on retail shelves

## Computer vision tasks

Object detection

Image segmentation

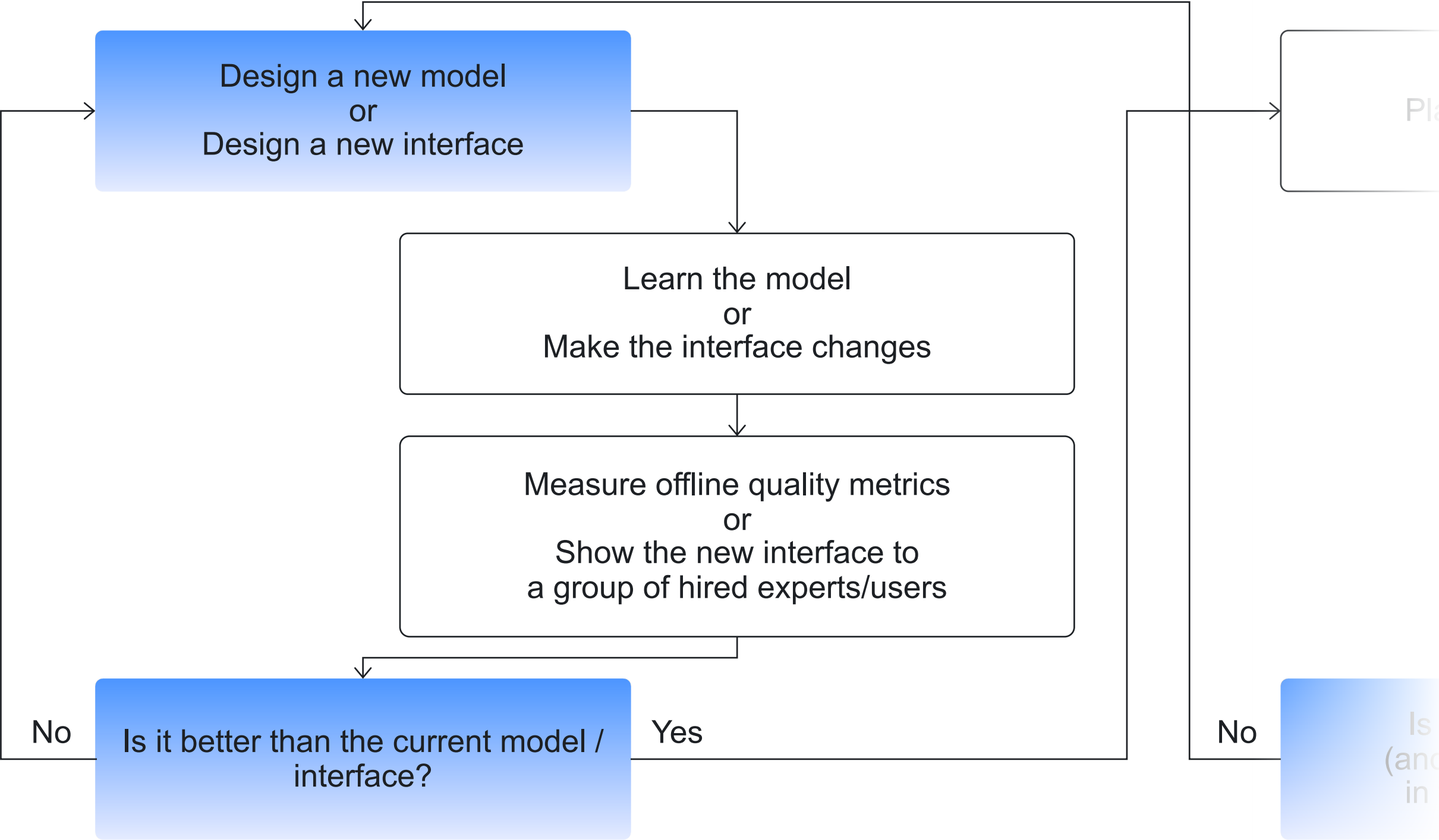
Image classification

Image transcription

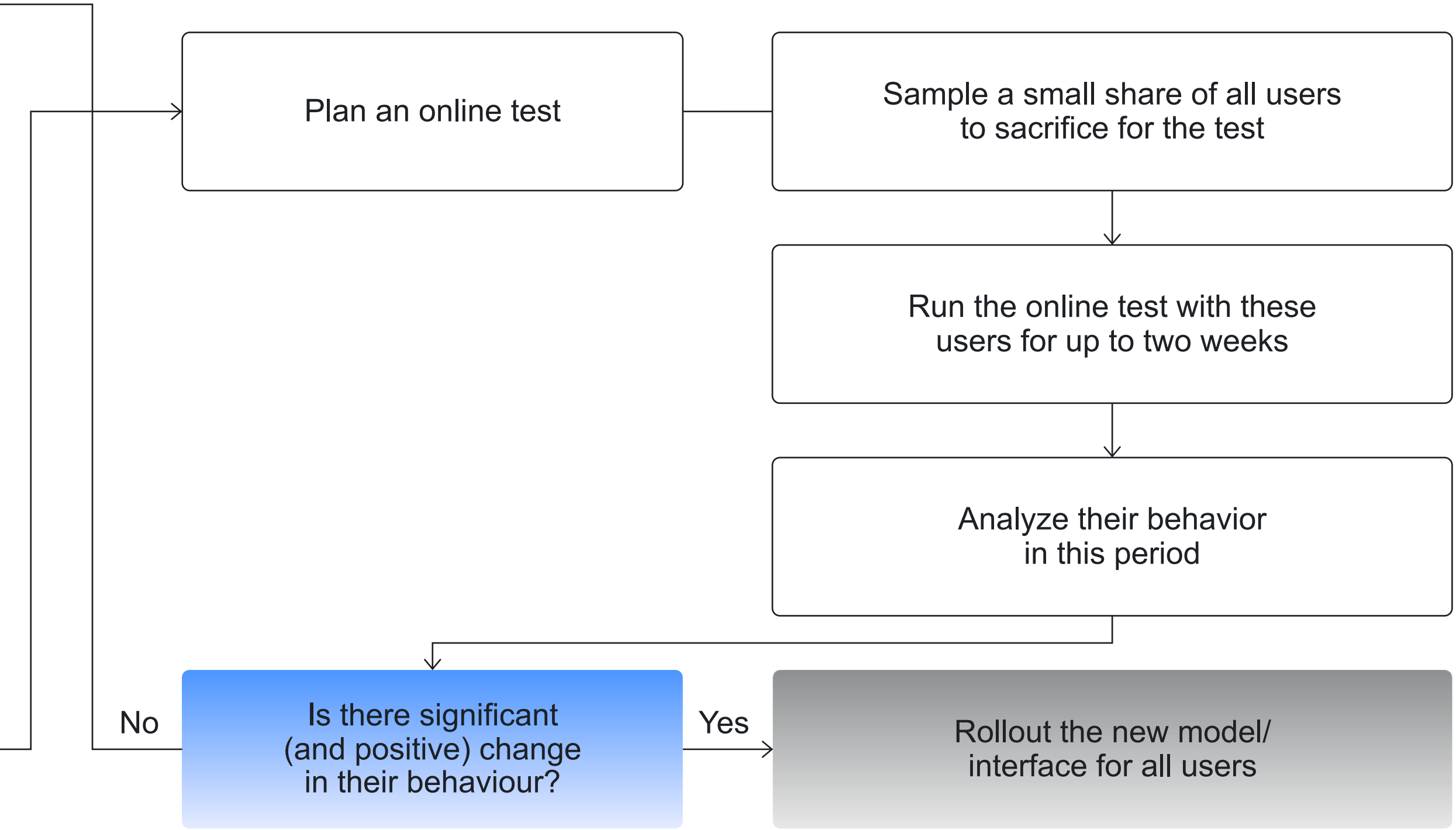
Side-by-side comparison

Image and video collection

# Scheme of web service evaluation



# Scheme of web service evaluation





# Pros and cons of online and offline metrics

## How labeled data is used

1. Calculating offline metrics to evaluate how a model is performing
2. Training ML models and choosing the best model version

## Offline metrics measured with data labeling

### Pros

- + Clear signal
- + Measures designated product characteristics
- + Fast results

### Cons

#### Fast results

- Not actual users (not always a representative sample)
- Can't measure business metrics

## Online metrics measured with A/B tests

### Pros

- + Results from real users
- + Measures business metrics (clicks, dwell time, leakage)

### Cons

- Implicit signal
- Delayed response
- Slow results (long experiments)
- Clickbait
- Fraud

# Evaluation

Offline eval

Online eval

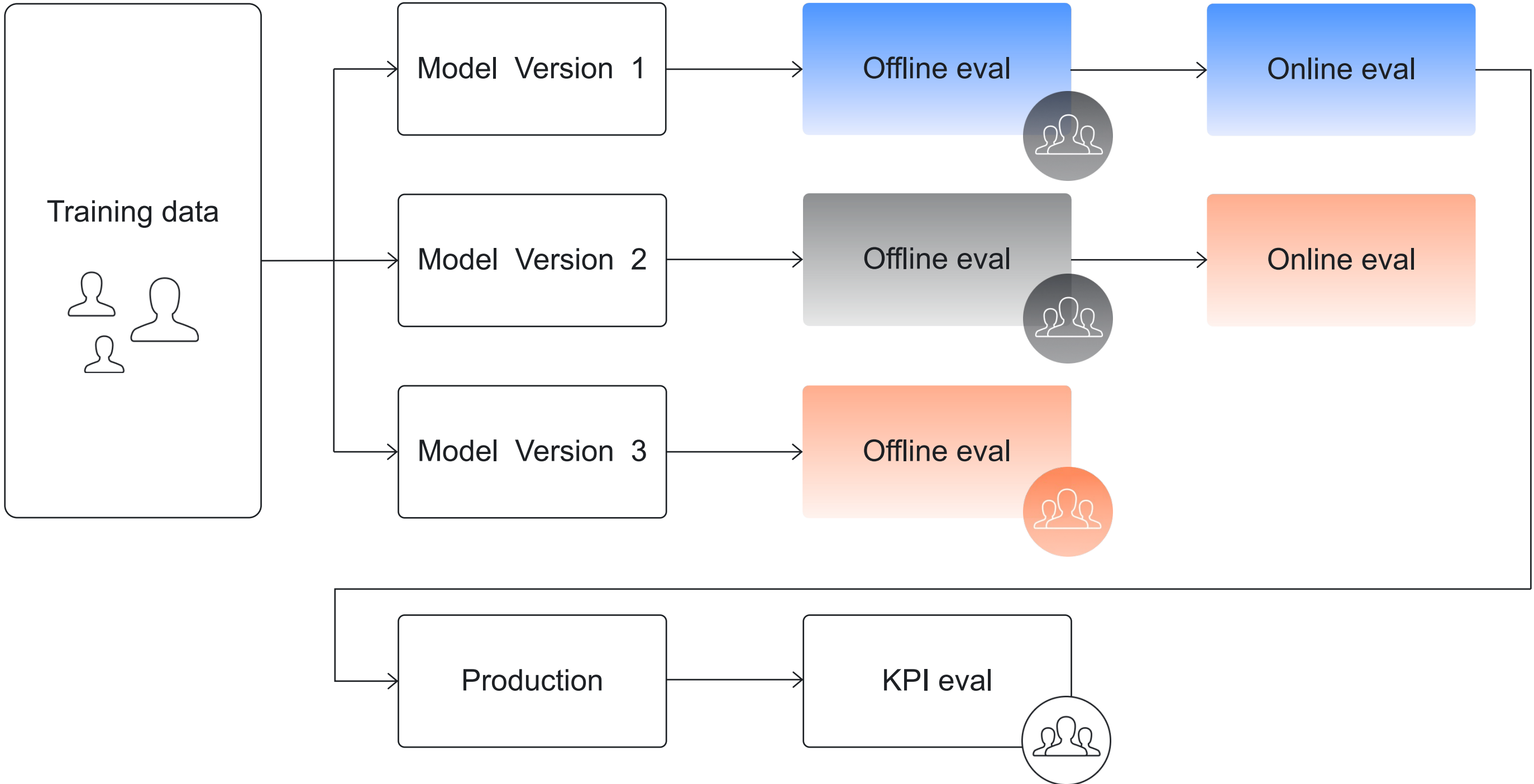
Covered in  
this tutorial

**E.g., see our tutorials**

- ▶ At TheWebConf 2018
- ▶ At KDD 2018
- ▶ At SIGIR 2019

# ML production pipeline: Humans are in the Loop!

Evaluation metrics are available within days or even hours



The faster you iterate to improve the processes  
and the pipeline the faster you improve quality

To accelerate web service improvement, you need

To accelerate experiments, you need

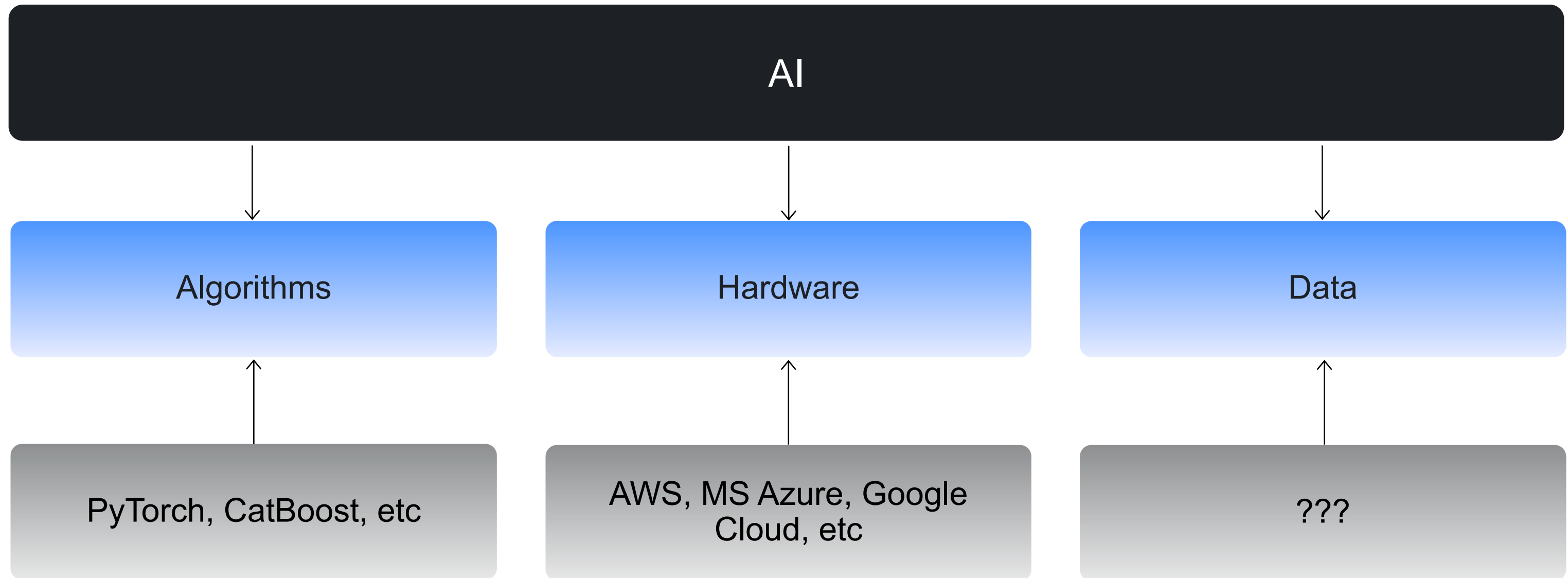
To have fast data iterations, you need ...

# The critical needs of an engineer to build effective human-in-the-loop processes

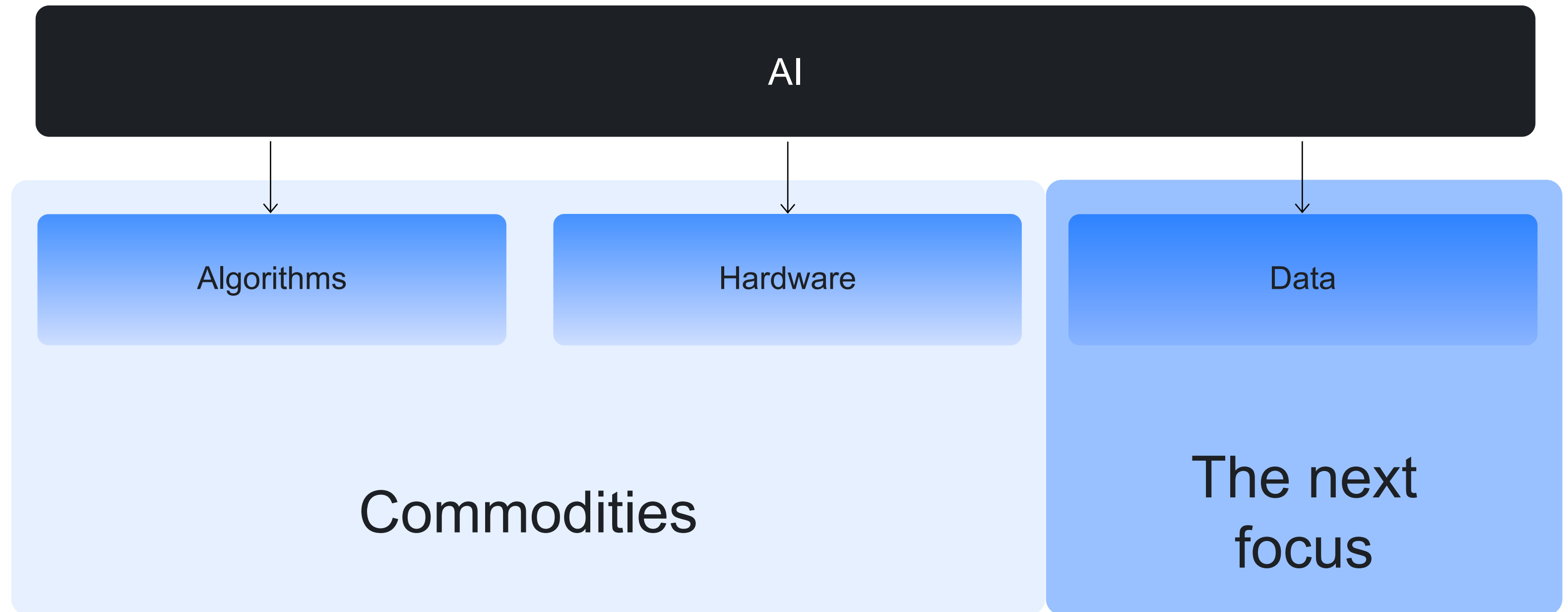
1. Direct hands-on access to make changes and improve processes
2. Integrated environment to work both with algorithms and humans
  - ▶ Ability to use the same programming language (e.g., Python and libraries)
  - ▶ Ability for easy integration (API)

# Automation: long-term effects and trends

# Three pillars of Artificial Intelligence (AI)

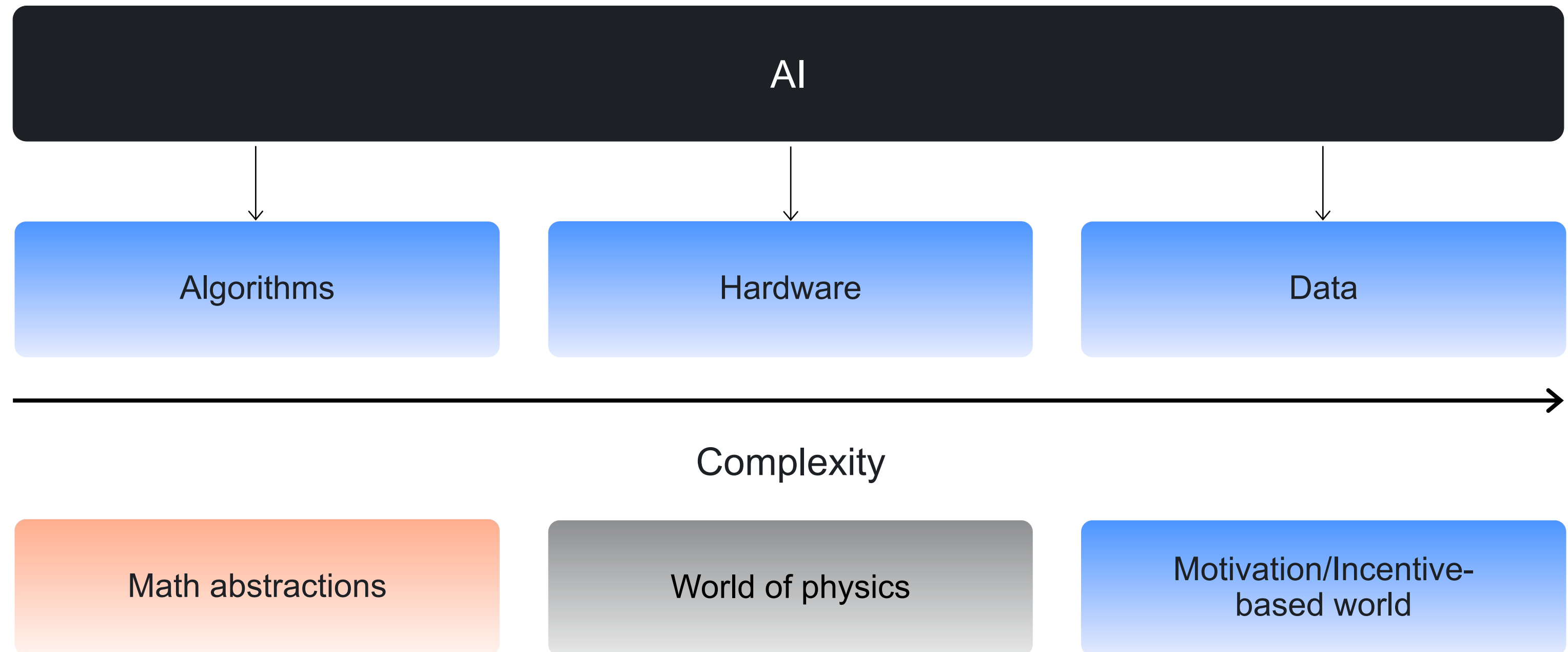


# Three pillars of AI: next revolution

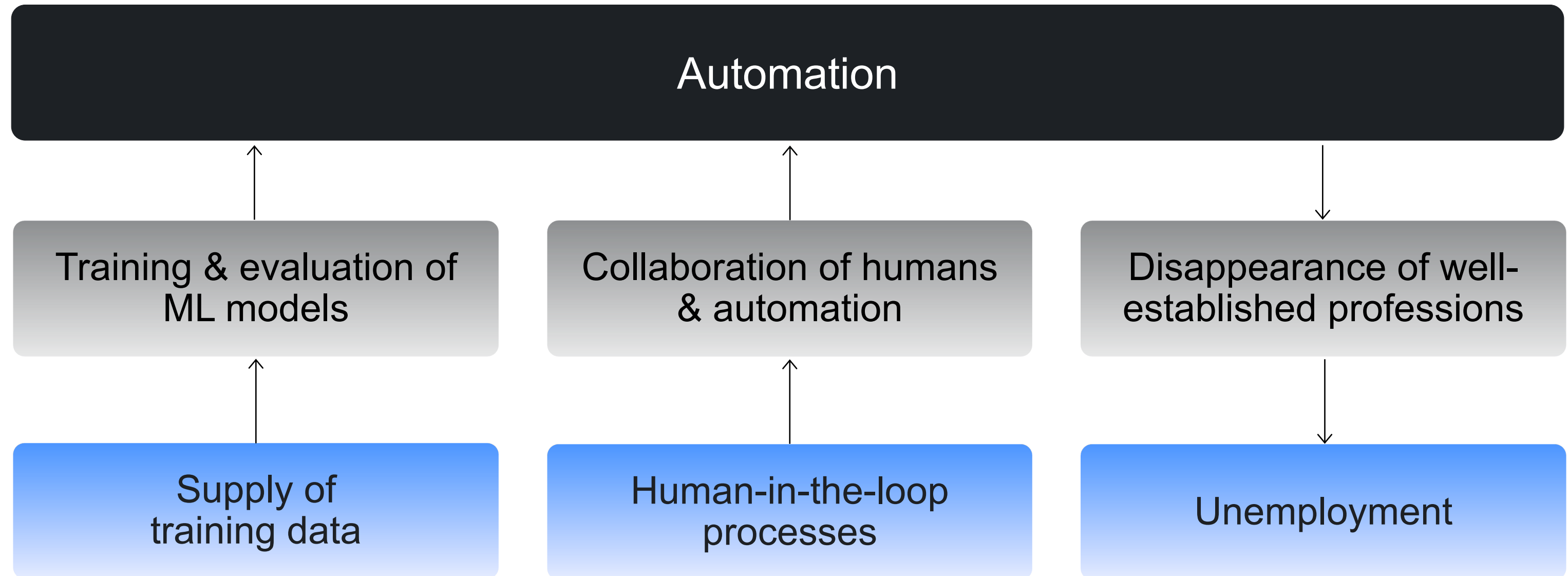




# Three pillars of AI: the order is natural



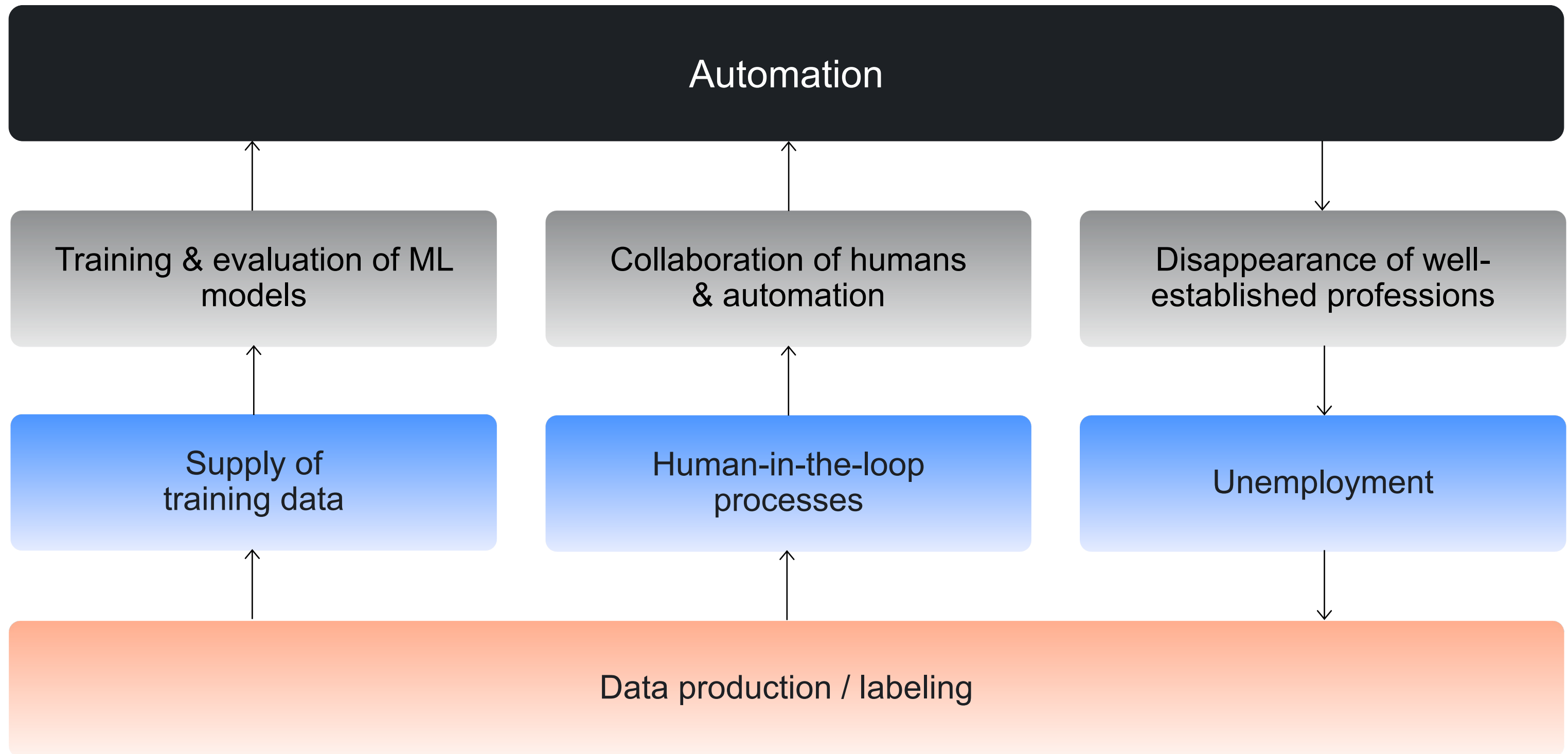
# Long-term trends




# Automation: pitfalls for humans

- ▶ Disappearances of well-established mass professions
- ▶ The ways how data production / labeling are organized

# Automation: pitfalls meet each other



The background features a series of overlapping, curved, blue shapes that create a sense of depth and movement, resembling a stylized fan or a series of concentric arcs. The colors range from a deep navy blue to a vibrant, bright blue.

# **SUMMARY:** **Requirements** **from two sides**

Majority of data-driven web services and products require training data labelled by human (annotators)



...at a large scale



# REQUEST-1: From community of industry engineers

Requirement for a layer that allows work with tasks for humans as with yet another computational cluster

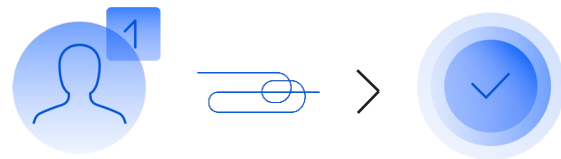




# Most popular non-engineer-oriented approaches

## In-house “expert”: managing people

Direct managing in-house crowd



- ▶ Easy to setup
- ▶ Expensive

- ▶ Unmeasurable
- ▶ Impossible to scale



## BPO / vendor

Access to crowd via third-party BPO who manage them



- ▶ Quick access to crowd
- ▶ Expensive

- ▶ Unmeasurable
- ▶ Hard to scale

# 20<sup>th</sup> century — style management

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



# It can be different

- ▶ Flexibility to choose from hundreds of tasks
- ▶ No requirements in regularity
- ▶ Switch to another task when bored



# REQUEST-2:

## From community of data annotators

- ▶ Prefer freedom in terms of place and time
- ▶ Doubt the availability and choice of tasks
- ▶ Need fair and ethical task assignment
- ▶ Need fair compensation and growth opportunities

# Key challenge

Address both requirements by properly organizing data production.

**Is crowdsourcing the solution?**



**Crowdsourcing**  
as a powerful technology  
for the data-driven era of the Web

# Crowdsourcing: specific way to design a business process



Huge task

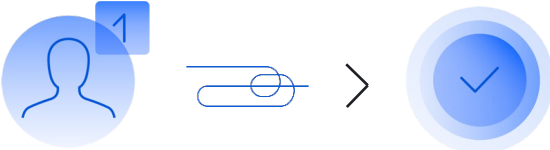
Cloud of annotators

Result

# Crowdsourcing as engineer-oriented approach

**In-house “expert”:** managing people


Direct managing in-house crowd



- ▶ Easy to setup
- ▶ Expensive
- ▶ Unmeasurable
- ▶ Impossible to scale

**BPO / vendor**


Access to crowd via third-party BPO who manage them



- ▶ Quick access to crowd
- ▶ Expensive
- ▶ Unmeasurable
- ▶ Hard to scale

**Crowdsourcing**

Technologically managing crowd as yet another computing power



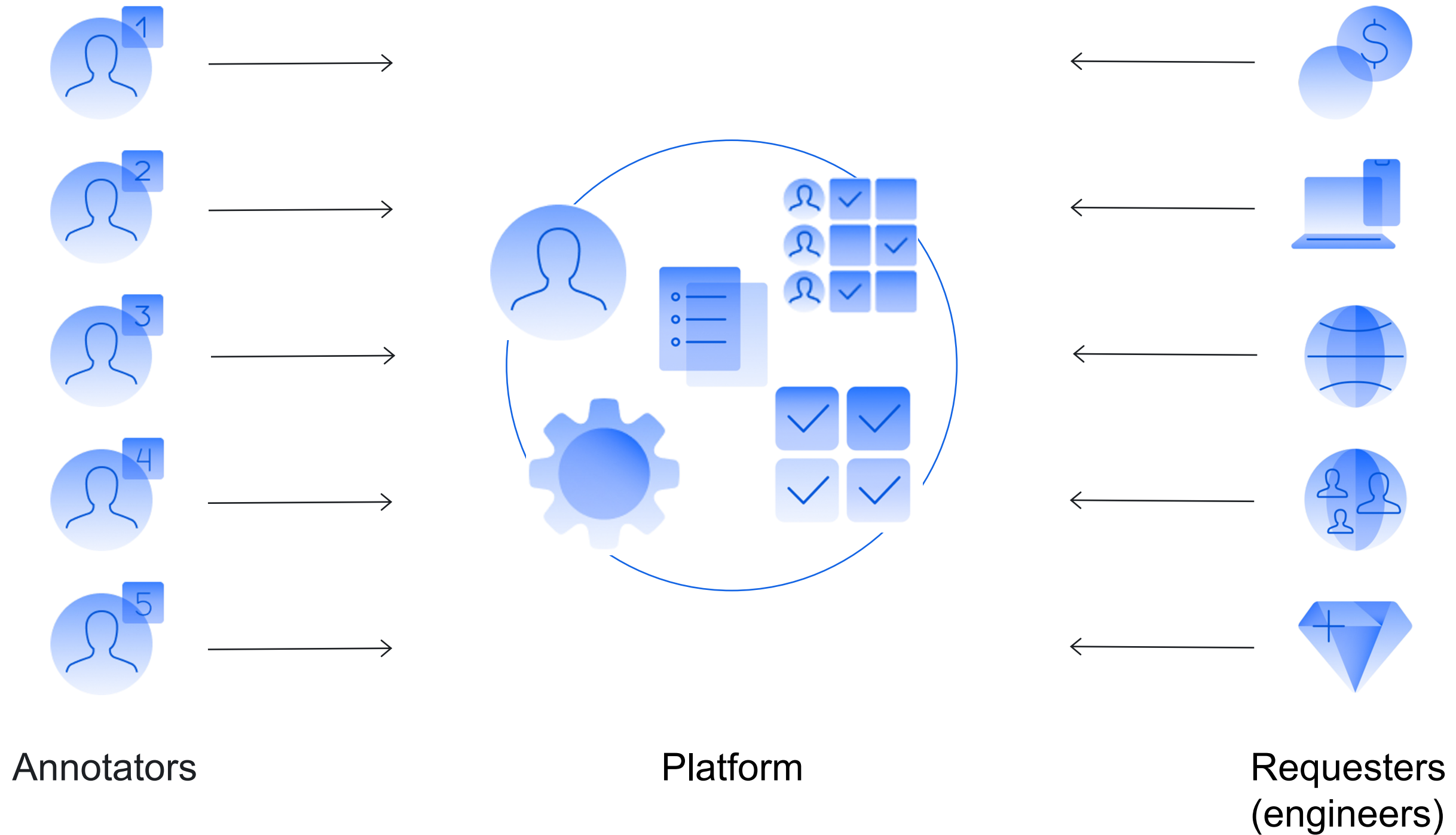
- ▶ Measurable
- ▶ Scalable
- ▶ Manageable
- ▶ Requires advanced tech



# Crowdsourcing can provide maximal flexibility to annotators if

- ▶ On a platform side, efficient tools for **quality management** are available for a requester-engineer
- ▶ **Requester-engineer knows how to build smart crowdsourcing pipelines** resistant to single annotator's mistakes

# A crowdsourcing platform: two-sided market



# Open crowdsourcing platforms: examples

- ▶ Toloka
- ▶ Amazon  
Mechanical Turk
- ▶ ClickWorker

# Pros of crowdsourcing platforms



**24/7**

Continuous  
data labeling

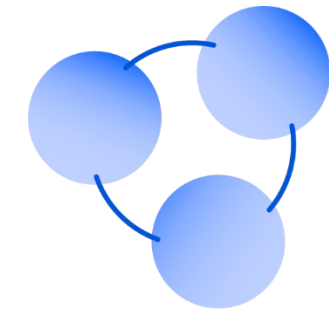


**Variety  
of skilled  
performers**



**Vast region  
coverage**

(40+ of most of  
languages and  
100+ countries)









**Ongoing  
processes**

# State-of-the-art crowdsourcing platforms offer





1. Self-service: Direct hands-on access to do changes and improve processes
2. Offer Python libraries that allow work with human processes in the same code base and environment as with algorithms and ML models
3. Ability for easy integration via API
4. Robust infrastructure, i.e., fault-tolerant high-load system for processing of millions of tasks per day

# Methodology behind efficient crowdsourcing


## Choose the crowd

-  Language & Region
-  Age
-  Gender
-  Device & OS
-  Individual quality rating
-  Subject matter experts




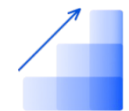
## Train the crowd

-  Training and practice
-  Testing skills
-  Skilled annotators
-  Quality-based pricing

## Control quality

-  Behavioral checks
-  Anti-robotic tools
-  Hidden control tasks
-  Consensus of multiple users
-  Verification of assignments

## Get high-quality results

-  System-level antifraud
-  Multiple aggregation models
-  Results-based performer selection
-  Real-time insights

# Crowdsourcing examples: use cases to improve search relevance

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 3  Fair 5  Adult 7  Unreadable text

2  Good 4  Bad 6  Junk

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  Uniuudqeable

Filters + category

Classify filter relevance to the product category

Filter  
**high heel**

Category  
**Women's shoes**

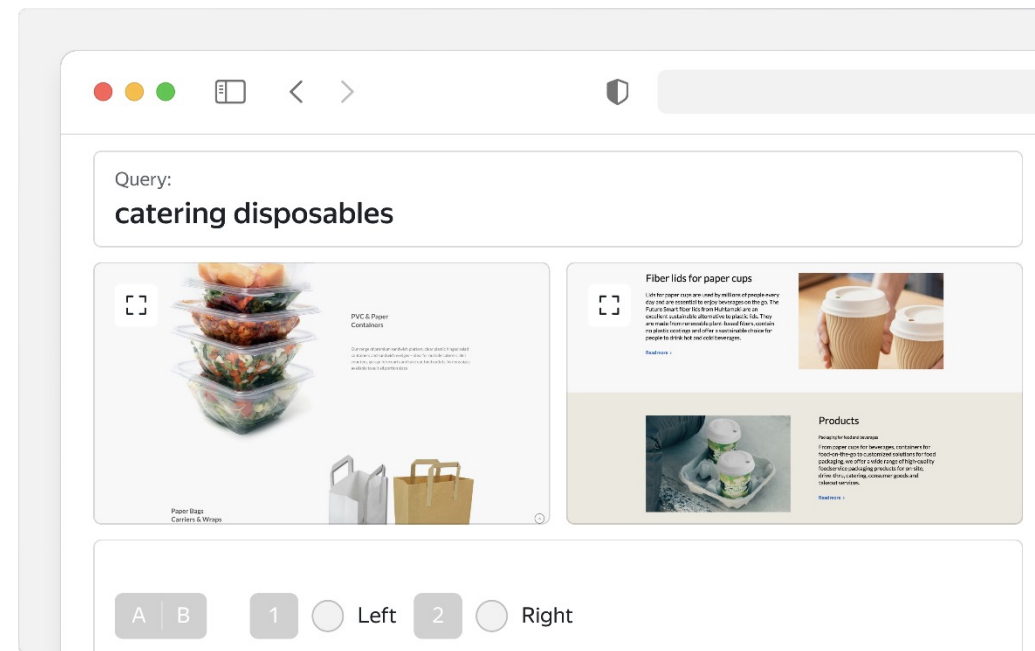
S Google first text D Google second text

1  Excellent 3  Fair

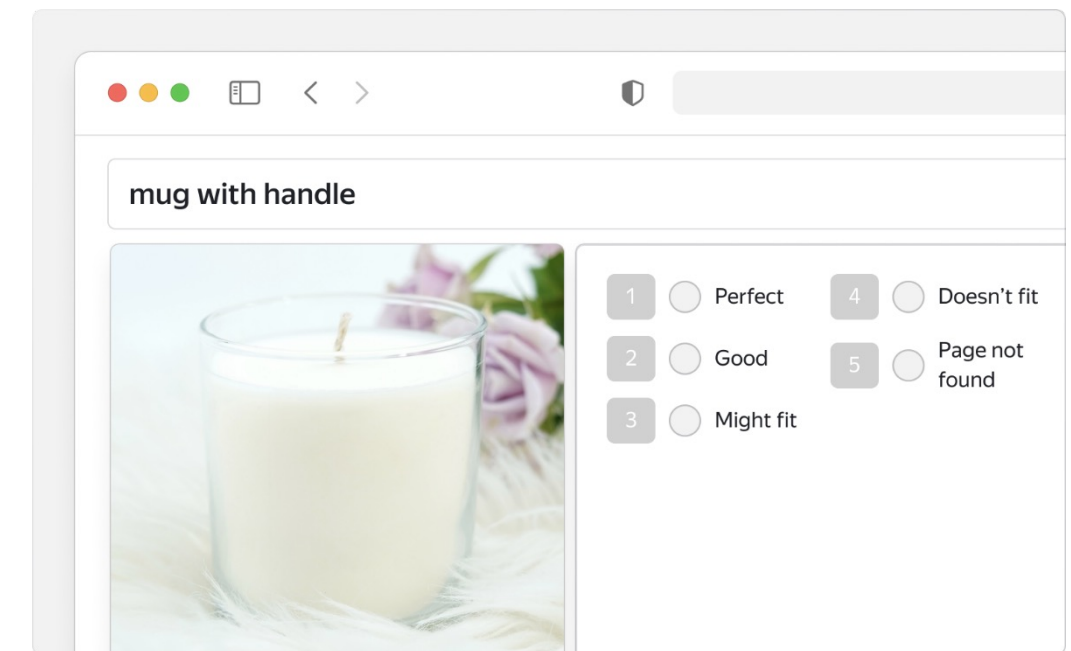
2  Good 4  Bad

# Crowdsourcing examples: use cases to improve search relevance

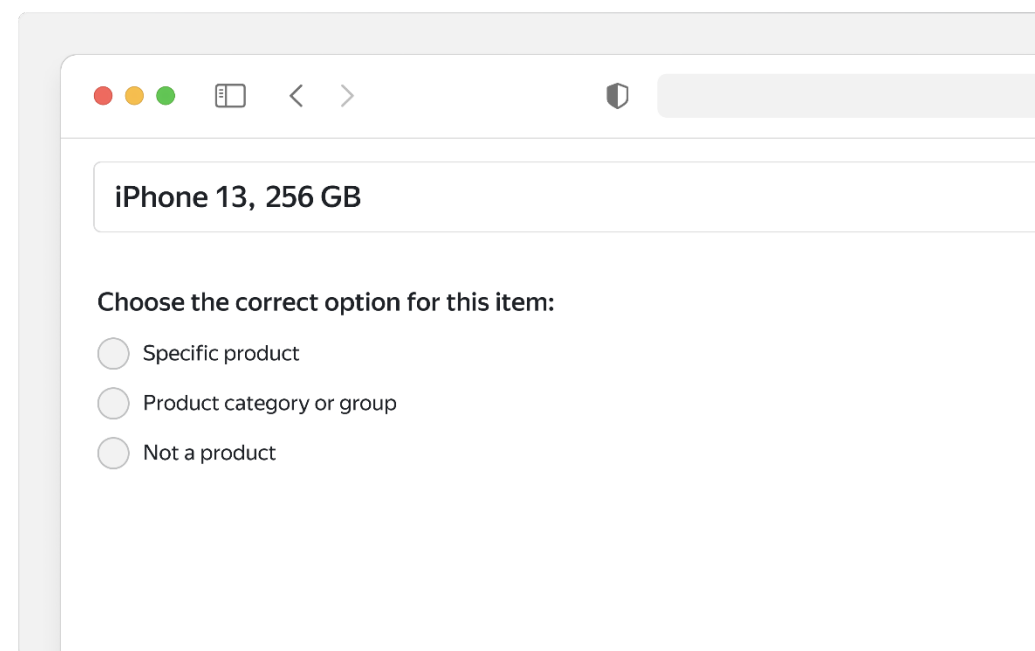
Side-by-side comparison of search results



Identify spam or irrelevant matches



Classify type of search query (broad vs narrow)



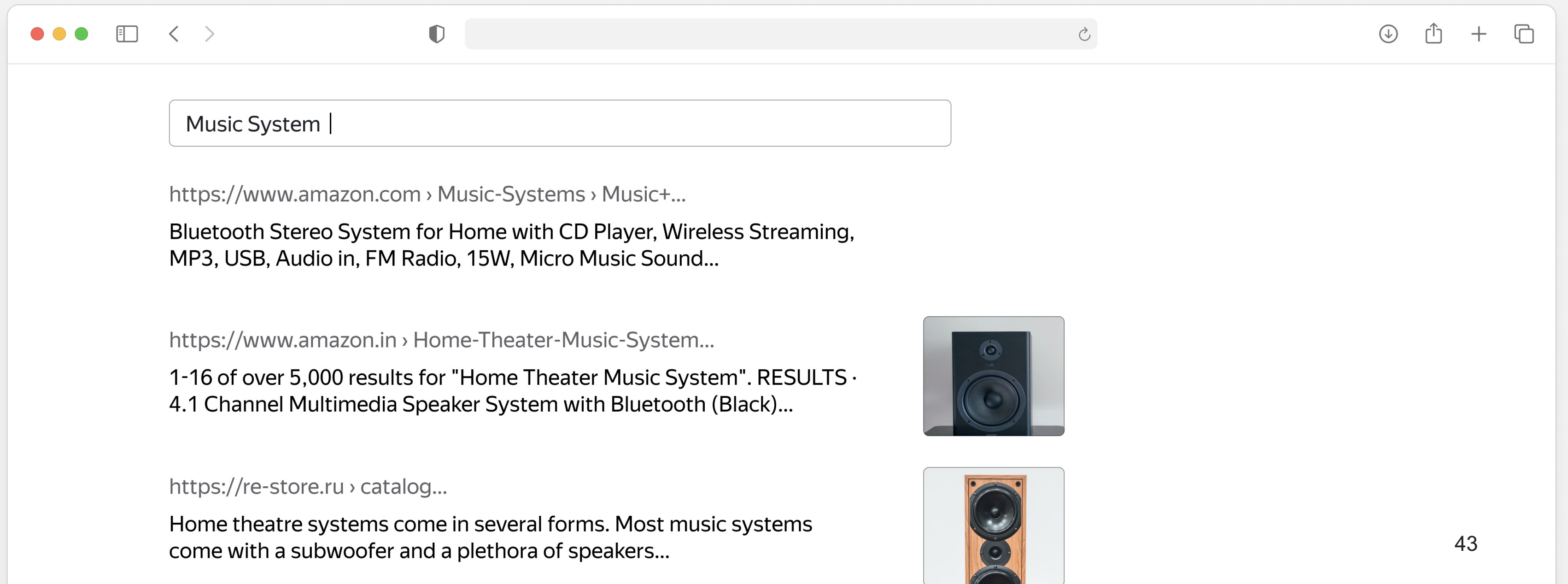


**Why this tutorial?**  
**Practice!**

# **We will learn on the example of ranking and offline evaluation**

**(however, the learned skills can be easily applied to  
other web services and applications)**

# Evaluation of a ranking



The screenshot shows a web browser window with a search bar containing the text "Music System |". Below the search bar, three search results are displayed. The first result is from "https://www.amazon.com" and is titled "Bluetooth Stereo System for Home with CD Player, Wireless Streaming, MP3, USB, Audio in, FM Radio, 15W, Micro Music Sound...". The second result is from "https://www.amazon.in" and is titled "1-16 of over 5,000 results for 'Home Theater Music System'. RESULTS · 4.1 Channel Multimedia Speaker System with Bluetooth (Black)...". To the right of this second result is a small image of a black speaker. The third result is from "https://re-store.ru" and is titled "Home theatre systems come in several forms. Most music systems come with a subwoofer and a plethora of speakers...". To the right of this third result is a small image of a wooden speaker.


Music System |

<https://www.amazon.com> › Music-Systems › Music+...

Bluetooth Stereo System for Home with CD Player, Wireless Streaming, MP3, USB, Audio in, FM Radio, 15W, Micro Music Sound...


<https://www.amazon.in> › Home-Theater-Music-System...

1-16 of over 5,000 results for "Home Theater Music System". RESULTS · 4.1 Channel Multimedia Speaker System with Bluetooth (Black)...



<https://re-store.ru> › catalog...

Home theatre systems come in several forms. Most music systems come with a subwoofer and a plethora of speakers...



# How can crowdsourcing be used?

1

As the **main instrument** for measuring context relevance and ranking ads

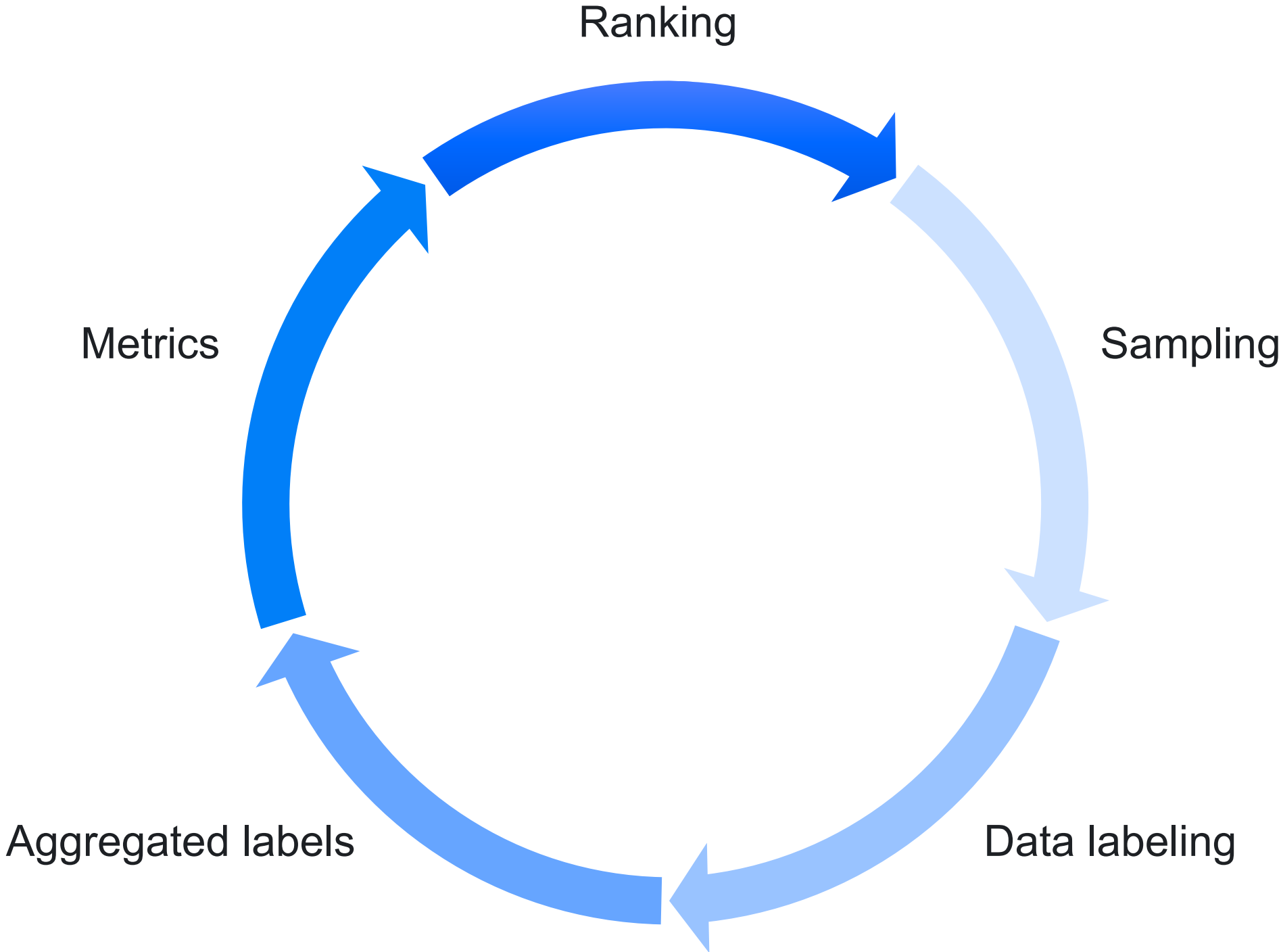
2

As a source for building large **datasets for ML models** (training, validation, test)

3

As a tool for **monitoring ML solutions**

# Human-in-the-Loop Pipeline for Offline Metrics



# Why crowdsourcing?

You can adapt to:

- ▶ The **type** of content (text, video, links, images, etc.)
- ▶ The **language** of the content
- ▶ Search results representation on different types of **devices**
- ▶ The fast-changing **people's needs**
- ▶ The emerging **content**

# Learning outcomes

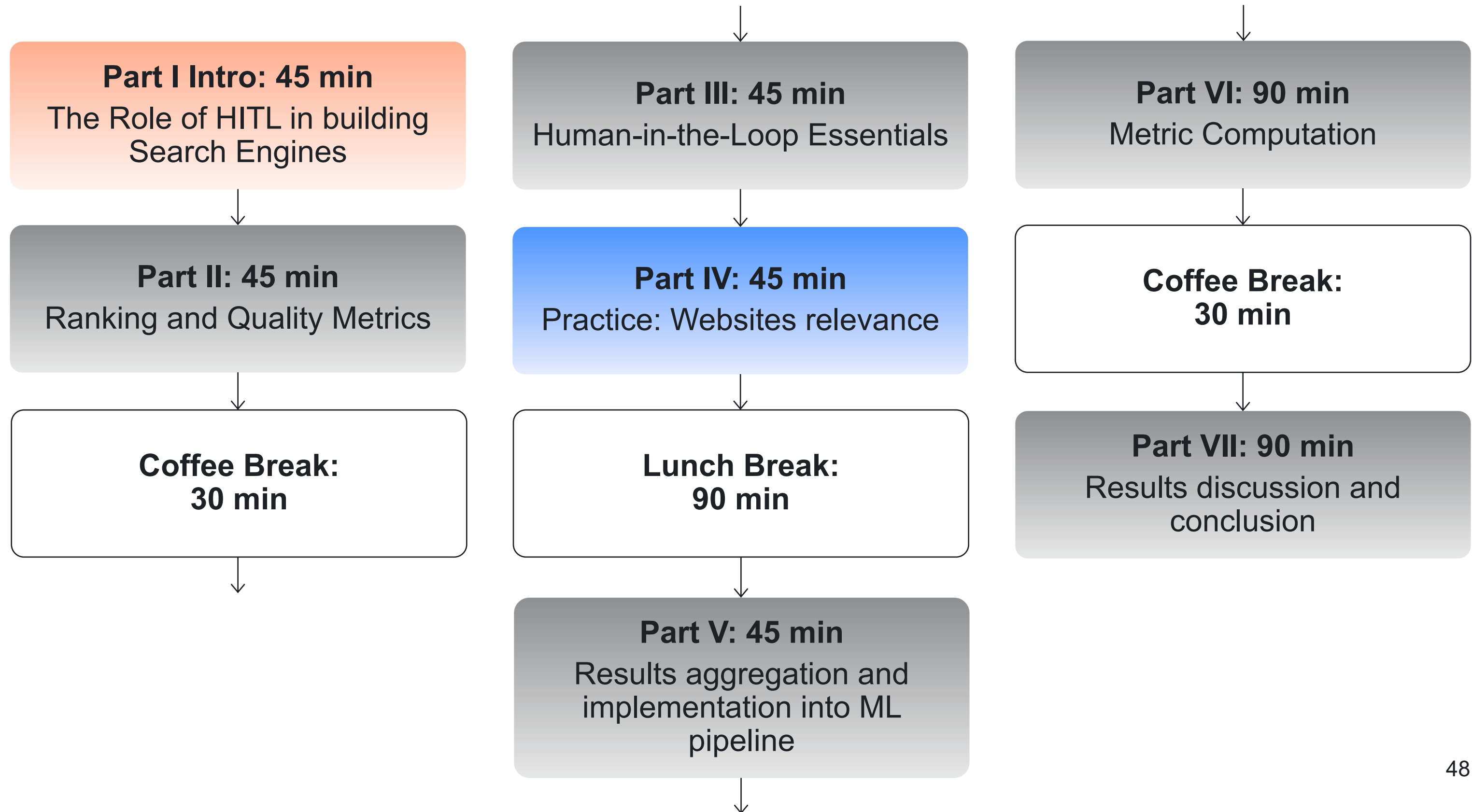
## Theory

- ▶ Offline approach for ranking evaluation
- ▶ Use crowdsourcing for industrial applications

## Practice

- ▶ Build scalable data labeling pipelines
- ▶ Run crowdsourcing projects with real annotators
- ▶ Program Human-in-the-Loop via public Python libraries (Toloka-Kit)

# Tutorial Schedule





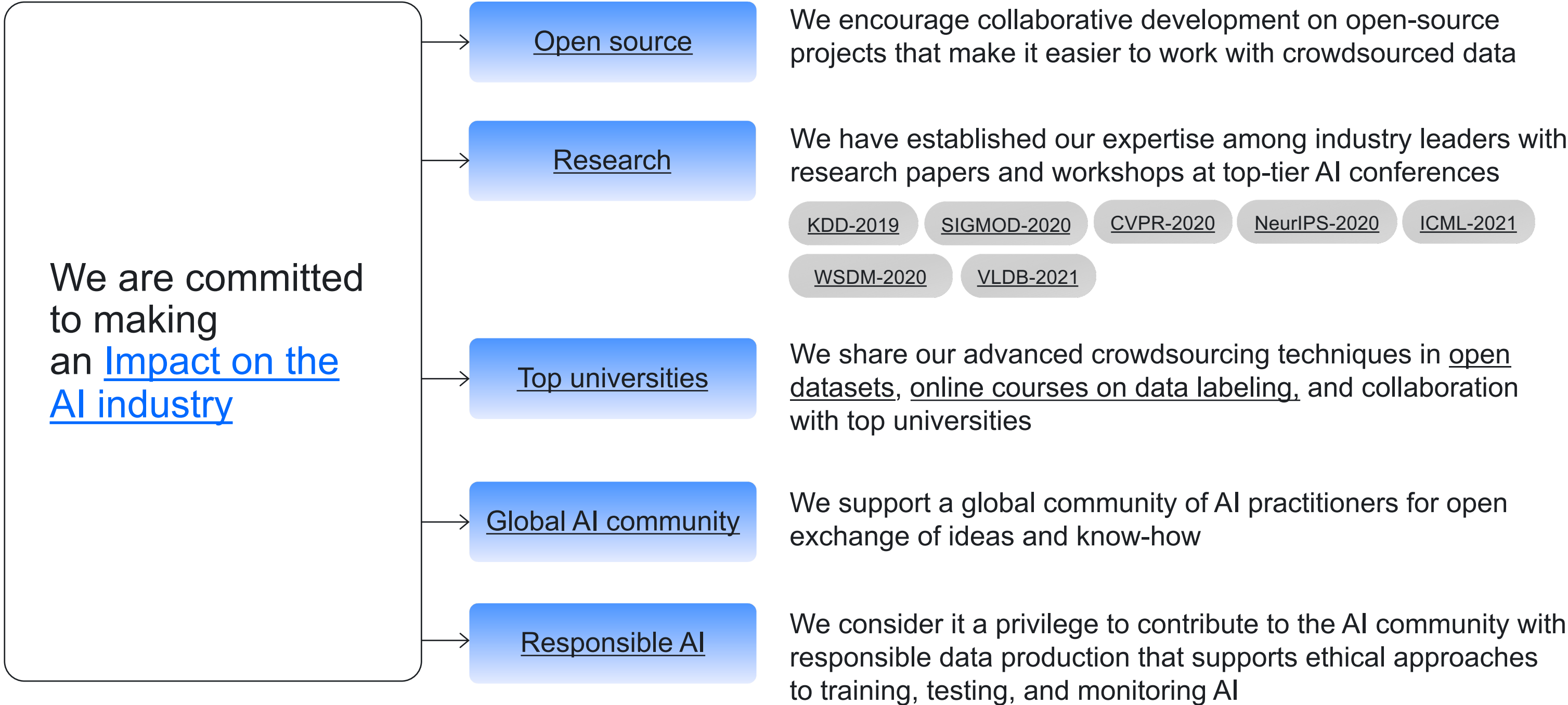
# Toloka Research Grants Program

- ▶ 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



# Join our Slack: icwe\_tutorial channel

**Alexey Drutsa**

deputy CEO, COO at Toloka



[adrutsa@toloka.ai](mailto:adrutsa@toloka.ai)



<https://www.linkedin.com/company/toloka/>



<https://tolokacommunity.slack.com/ssb/>



<https://toloka.ai/events/icwe-2022/>



<https://twitter.com/tolokaai>



<https://github.com/Toloka>