Deploying Complex Crowdsourcing Tasks Without Fear: An Introduction to Crowd_Frame

Michael Soprano, Kevin Roitero, and Stefano Mizzaro

michael.soprano@uniud.it, University of Udine (DMIF)

The 3rd Italian Conference on Big Data and Data Science

Pisa - September 17, 2024

UNIVERSITÀ DEGLI STUDI DI UDINE

hic sunt futura

Introduction

- **Human Computation**: Leveraging human intelligence to solve computational problems or complete tasks that machines cannot easily handle alone
  - Larger workflows, where human input is needed
- **Crowdsourcing**: a modality to leverage human intelligence at scale [1]
- Not all crowdsourcing qualifies as human computation

[1] Lease, M., and Alonso, O. *Crowdsourcing and Human Computation*. 2014. Introduction.

Introduction

- **Crowdsourcing** = *crowd* + *outsourcing* [2]
- Outsourcing some kind of task
- Performed internally, for instance by employees
- To an **undefined** and (generally) **large** crowd of people
- In the form of an **open call** to work

[2] Jeff Howe, The Rise of Crowdsourcing, Wired, 2006, https://www.wired.com/2006/06/crowds/

Introduction

- **Individuals who need to conduct a crowdsourcing task**
- Crowdsourcing platforms are **marketplaces for human labor**
- Requesters publish **open calls for work**
- Tasks are allocated to a **crowd of workers**
- Crowd workers perform the assigned tasks and receive a **reward**

Terminology

- **Human Intelligence Task (HIT)**: a single, self-contained, virtual work unit allocated to and performed by an individual
- **Requester**: an employer who recruits employees (*workers*) from a platform for the execution of HITs in exchange for a wage (*reward*)
- **Worker**: an individual who joins a crowdsourcing platform to perform and complete HITs published by requesters
- **Element**: item that a individual evaluates, uses, addresses within a HIT
  - A HIT is composed of a set of elements
- **Batch**: a set composed of multiple HITs published by a single requester
- **Task**: all the batches of work units published

Partly defined by Paolacci, G. Chandler, J. and Ipeirotis, P. *Running Experiments on Amazon Mechanical Turk*. 2010.

Introduction

Various crowdwsourcing platforms:

- **Amazon Mechanical Turk**: a general-purpose human workforce
- **Prolific**: workers are explicitly recruited for participation in research tasks
- **Toloka**: primarily focused on data labeling tasks, but not exclusively

Introduction

- A requester needs to choose the **most suitable platform**
- **Each platform supports** crowdsourcing task workflows **in different ways**
  - This workflow can be **complex** and present several challenges
- `Crowd_Frame` allows for **simplifying parts of this workflow**
- What must a requester do to **design and publish a crowdsourcing task**?
- Sample platform: *Amazon Mechanical Turk*
- Afterwards, we will turn to `Crowd_Frame`

https://www.mturk.com/

Amazon Mechanical Turk

The task design and deployment workflow involves three phases:

- **Project definition**
  - The requester chooses the type of task and sets parameters
- **Task interface design**
  - Development of a web application
- **Customization** and **publication**
  - The requester can define multiple pools of workers to recruit

Task interface design

- **Web application:** HTML + CSS + JavaScript
- The platform provides the **Crowd Elements**
- **Superset of HTML**
  - Custom tags that include and hide CSS styles and JS scripts
  - Attributes to modify their behavior
  - Task body: `<crowd-form>`
  - Instructions: `<crowd-instructions>`
  - ...

Amazon Mechanical Turk

```
<crowd-form answer-format="flatten-objects">

  <crowd-instructions link-text="View instructions" link-type="button">
    <short-summary>
      <p>Provide a brief instruction here</p>
    </short-summary>

    <detailed-instructions>
      <h3>Provide more detailed instructions here</h3>
      <p>Include additional information</p>
    </detailed-instructions>
  </crowd-instructions>

  <div>
    <p>What is your favorite color for a bird?</p>
    <crowd-input name="favoriteColor" placeholder="example: pink" required></crowd-input>
  </div>

</crowd-form>
```

Amazon Mechanical Turk

Some considerations:

- The **whole codebase** in a single box
- HTML + CSS + Javascript
- Presentation and logic are mixed
- Data storage: formd fields of type `hidden`
  - `JSON` objects as values
- ...

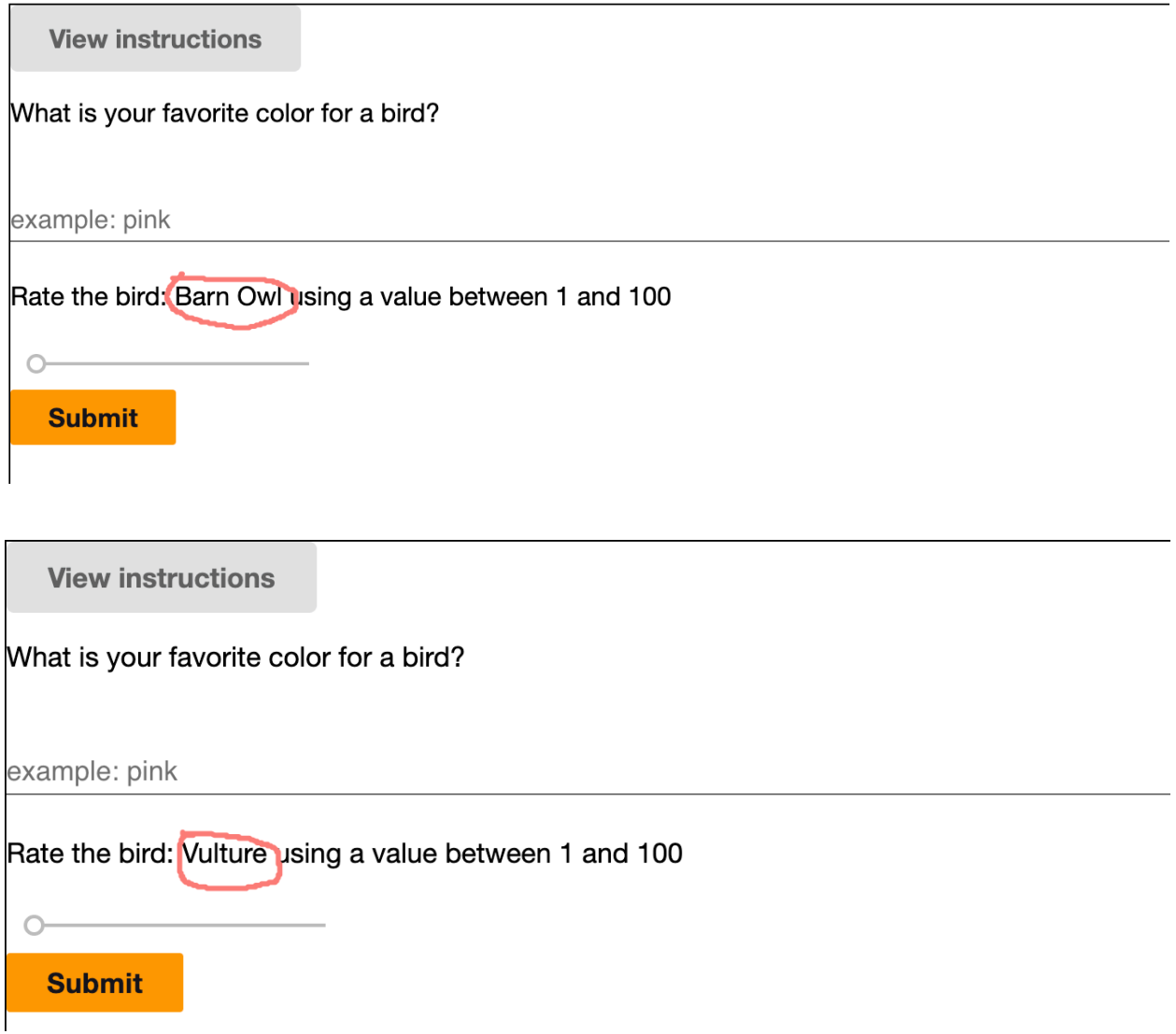
Customization and publish

How to **customize the task interface for each HIT** assigned to a worker?

- Define **variables** whose values can be **used in the codebase**
  - Syntax: `${VARIABLE_NAME}`
- **Initialize** them **during task publication**
  - CSV file-based mechanism
  - Columns: variable names, rows: values

```
<div>
 <p>Rate the bird: ${BIRD_NAME} using a value between 1 and 100</p>
  <crowd-slider name=“birdRating" min="1" max="100" step="1" required></crowd-slider>
</div>
```

Amazon Mechanical Turk



Two different HITs customized using the CSV-based variable passing mechanism

Amazon Mechanical Turk

Several difficulties emerge:

- The requester must have **advanced programming skills**
- The user interface is built by **mixing presentation and business logic**
- The **input data passing** mechanism is **cumbersome**
- **Data storage** for complex data **is not trivial** and falls on the requester
  - Logging...
  - Timestamps...

Wrap Up

What about **other crowdsourcing platforms**?

- Let us consider, for instance, *Prolific* [4]
- The requester has to to **set some parameters**
- **External URL** to the task interface
- No way to **design the task "in-house"**

What to do, then?

[4] Palan, S. and Schitter, C. *Prolific.ac—A subject pool for online experiments*. 2018. Journal of Behavioral and Experimental Finance. Vol. 17

Idea

- Rely on **crowdsourcing platforms** only to (eventually) **recruit the crowd of workers**
- The **workers access a task designed and deployed using external software**
- They **perform the task** and **return on the platform to receive the reward** (if needed)
- This external software is *Crowd_Frame* [5]
  - https://github.com/Miccighel/Crowd_Frame

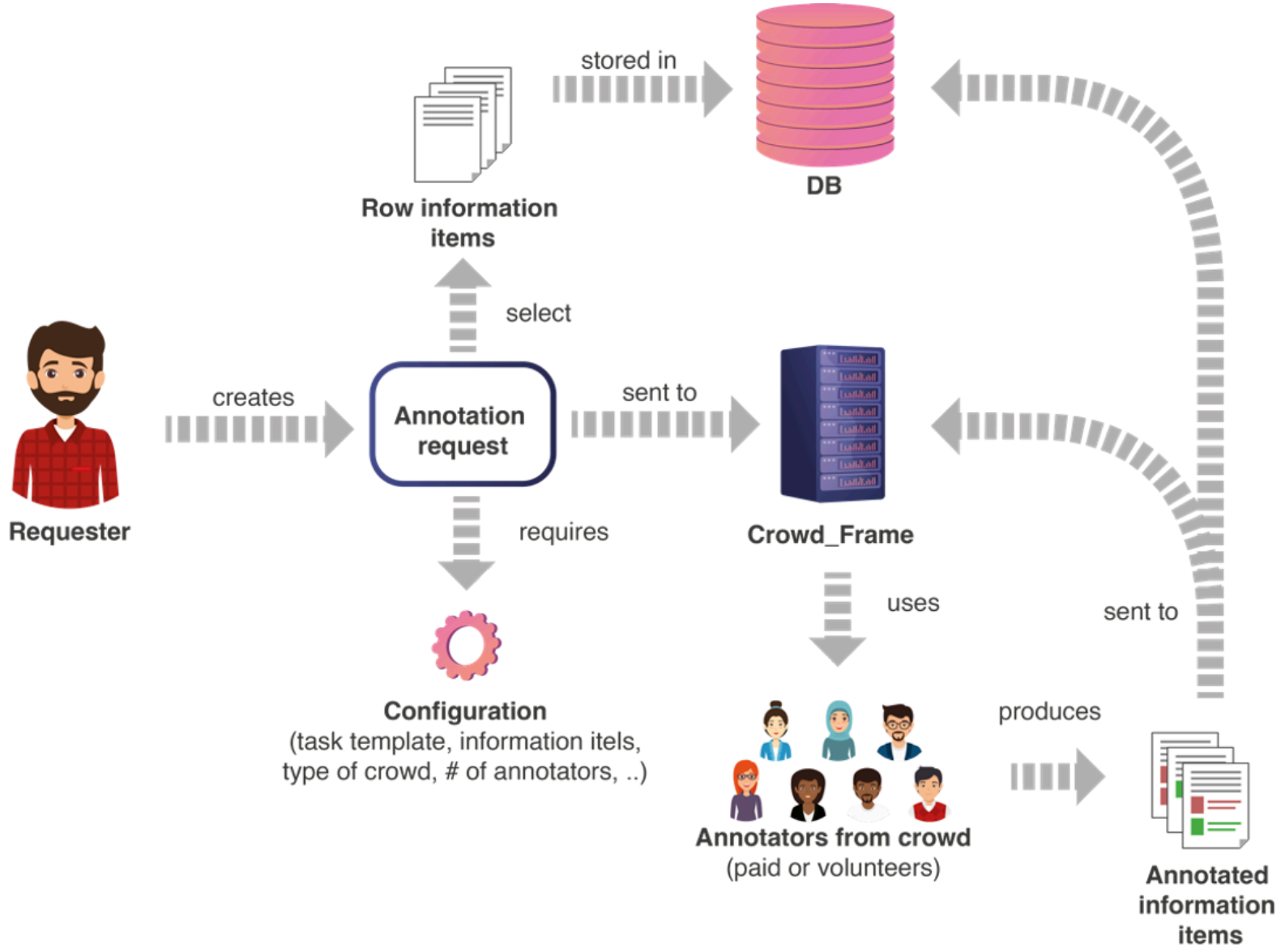
[5] Soprano M., Roitero K., Bombassei De Bona F., and Mizzaro S. *Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-shelf*. 2022. Proc. of the 15th ACM WSDM Conference

Aims

> *A software system that allows to easily design and publish diverse types of crowdsourcing tasks*

- Build **a variety of crowdsourcing tasks**
- A **customizable and controllable** environment
- Simplify **task design and publishing workflow**
- **Regardless of the crowdsourcing platform** chosen
- **Freely available to the research community**

Crowd_Frame



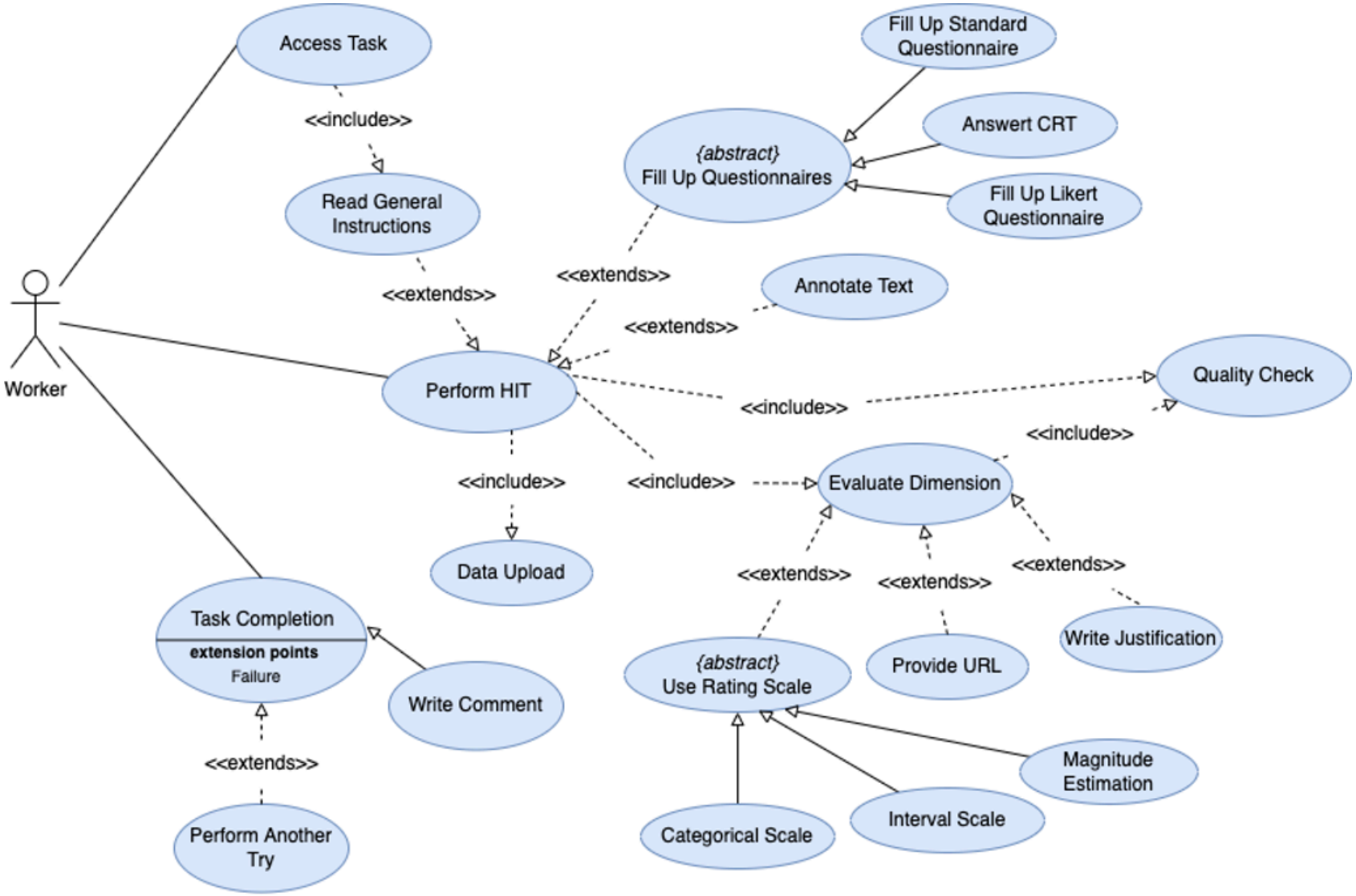
Architecture of Crowd_Frame

In a Nutshell

- The requester defines a set of **annotation requests** (HITs) composed of information items
- **Task configuration** is represented as a **set of JSON objects**
- **Modifying the configuration** results in **a different task**
  - *Crowd_Frame*: an Angular application
  - *DB*: based on Amazon Web Services
- Crowd workers **interact with the application**
- They **produce annotated information items**
- Results are **downloaded in a well-known format**

https://angular.io/

Crowd_Frame



High-level overview of possible worker interactions during an annotation

Crowd_Frame

- **Detailed instructions** available **in the repository**
- Follow the instructions, then focus on the `build/task` folder

File	Description
`hits.json`	Contains the full set of HITs
`questionnaires.json`	Defines the questionnaires
`dimensions.json`	Defines the evaluation dimensions
`instructions_general.json`	Provides general instructions
`instructions_evaluation.json`	Provides specific evaluation instructions
`search_engine.json`	Configures the custom search engine
`task.json`	Contains general task settings
`workers.json`	Configures worker access settings

Hands-on

- Let us **design a simple task** using `Crowd_Frame`
- **Fact-checking** process:
  - Check-worthiness, evidence retrieval, truthfulness assessment, ...
- "Most successful" use of `Crowd_Frame`
- **Truthfulness assessment** crowdsourcing task:
  - **Dataset**: 3 information items
    - `text`, `speaker`, and `date`
  - **General instructions**
  - **Workers**: 3 (i.e., 3 HITs)
  - **Evaluation dimension**: 1 (`truthfulness`)
  - **Assessment time**: 1 hour

Hands-on

- *PolitiFact*: an **organization that fact-checks information items**
  - US politicians, political parties, other public figures, and social media
- More than 24,000 fact-checks available
- Uses a six-level assessment scale
- Let us sample **3 information items**

Social Media

stated on September 10, 2024 in social media posts:

Vice President Kamala Harris wore NOVA H1 audio earrings in the Sept. 10, 2024, presidential debate.

NATIONAL | DEBATES | FACEBOOK FACT-CHECKS | SOCIAL MEDIA

https://www.politifact.com/

Our Dataset

Speaker	Date	Text	Ground Truth
Donald Trump	September 10, 2024	In Springfield (Ohio), they're eating the dogs, the people that came in, they're eating the cats. They're eating, they're eating the pets of the people that live there.	`pants-on-fire`
Kamala Harris	September 10, 2024	As of today, there is not one member of the United States military who is in active duty in a combat zone, in any war zone around the world, for the first time this century.	`mostly-false`
Elizabeth Warren	August 25, 2024	JD Vance actually sent a letter last year to the Department of Justice saying, ‘enforce the Comstock Act.’	`true`

Hands-on

Objects with **captions and text**, where the **text can contain HTML tags**

```
[
    {
        "caption": "Task Instructions",
        "text": "<p>Your main task is to <b>assess the truthfulness of 3 information items</b> by...</p>"
    }
]
```

HITs Allocation

1. There must be an **array of HITs** (also referred to as *units*)
2. Each HIT must have a *unique* **input token attribute**
3. Each HIT must have a *unique* **output token attribute**
4. The **number of elements in each HIT** must be specified
5. Each element must have an **attribute named** `id`
6. Each element can have an **arbitrary number of attributes**

Hands-on


```
[
        {
                "unit_id": "unit_0",
                "token_input": "INPUT1",
                "token_output": "OUTPUT1",
                "documents_number": 3,
                "documents": [
                        {
                                "id": "statement_1", "text": "In Springfield (Ohio) ...",
                                "date": "September 10, 2024", "speaker": "Donald Trump"
                        },
                        {
                                "id": "statement_2", "text": "As of today, ...",
                                "date": "September 10, 2024", "speaker": "Kamala Harris"
                        },
                        {
                                "id": "statement_3", "text": "JD Vance actually...",
                                "date": "September 10, 2024", "speaker": "Elizabeth Warren"
                        }
                ]
        }, ...
]
```

Hands-on

```
[
    {
        "name": "truthfulness",
        "scale": {
            "type": "categorical",
            "mapping": [
                { "label": "Pants-on-Fire", "description": "", "value": "0" },
                { "label": "False", "description": "", "value": "1" },
                { "label": "Mostly-False", "description": "", "value": "2" },
                { "label": "Half-True", "description": "", "value": "3" },
                { "label": "Mostly-True", "description": "", "value": "4" },
                { "label": "True", "description": "", "value": "5" }
            ],
            "instructions": {
                "label": "A",
                "caption": "Evaluate the truthfulness of the statement",
                "text": false
            }
        },
        "style": {
            "type": "list",
            "position": "middle",
            "orientation": "vertical",
            "separator": false
        }
    }
]
```

Evaluation dimensions configuration (`dimensions.json`)

Hands-on

- **Time available to assess**, **minimum time required** for each element, ...

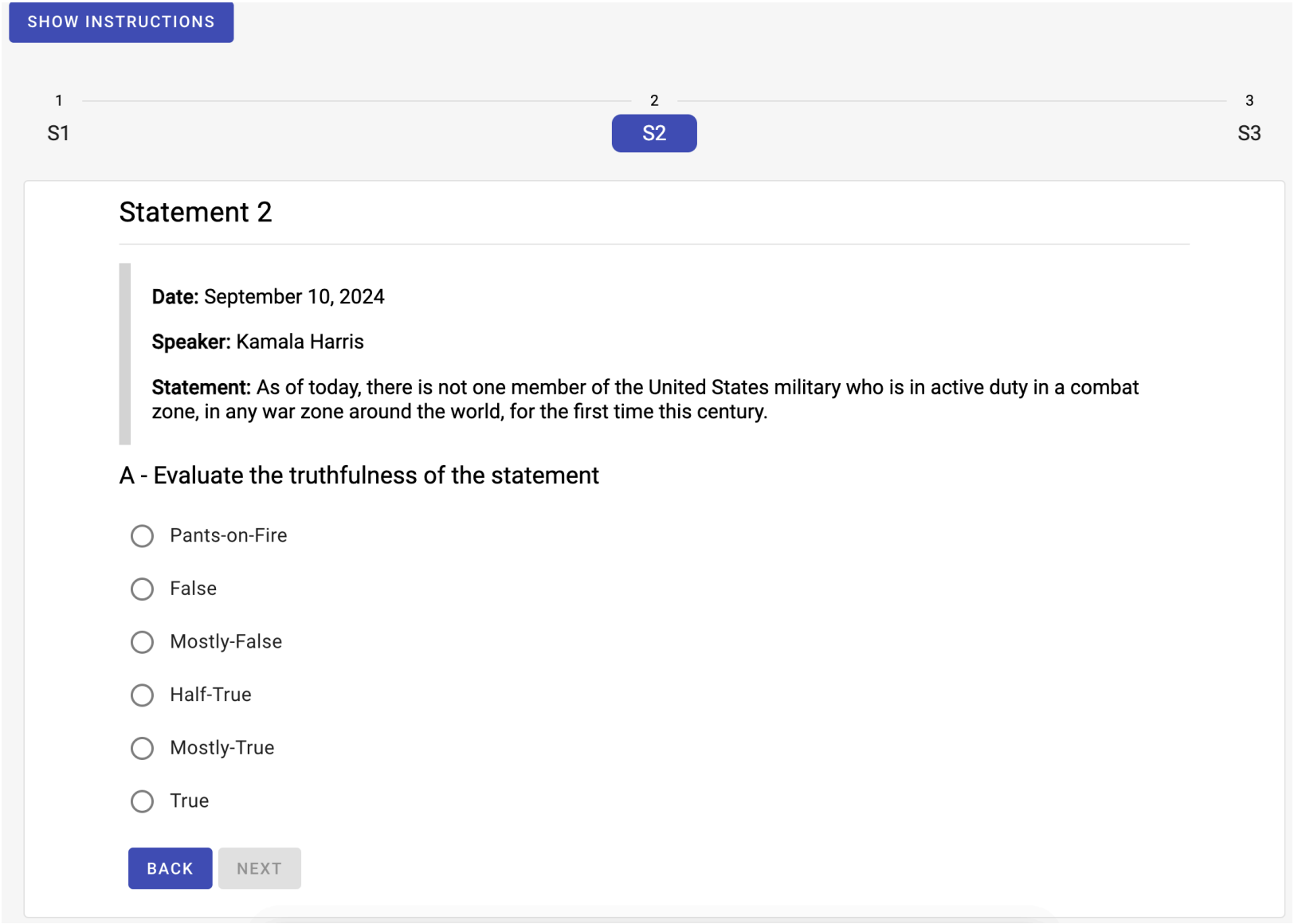
```
{
    "modality": "pointwise",
    "allowed_tries": 10,
    "time_assessment": 1,
    "time_check_amount": 3,
    "attributes": [
        { "name": "id", "name_pretty": false, "show": false, "annotate": false, "required": false},
        { "name": "date", "name_pretty": "Date", "show": true, "required": false, "annotate": false},
        { "name": "speaker", "name_pretty": "Speaker", "show": true, "required": false, "annotate": false},
        { "name": "text", "name_pretty": "Statement", "show": true, "required": false, "annotate": false}
    ],
    "element_labels": {
        "main": "Statement",
        "main_short": "S"
    },
    ...
}
```

Hands-on

- Remaining **configuration objects are left empty**
- Let's **take a look at the task**
  - https://bit.ly/itadata2024-demo-base
- Download **the sample configuration**
  - https://bit.ly/itadata2024-demo-configuration

Examples are available in the repository: https://github.com/Miccighel/Crowd_Frame/tree/master/examples

Hands-on



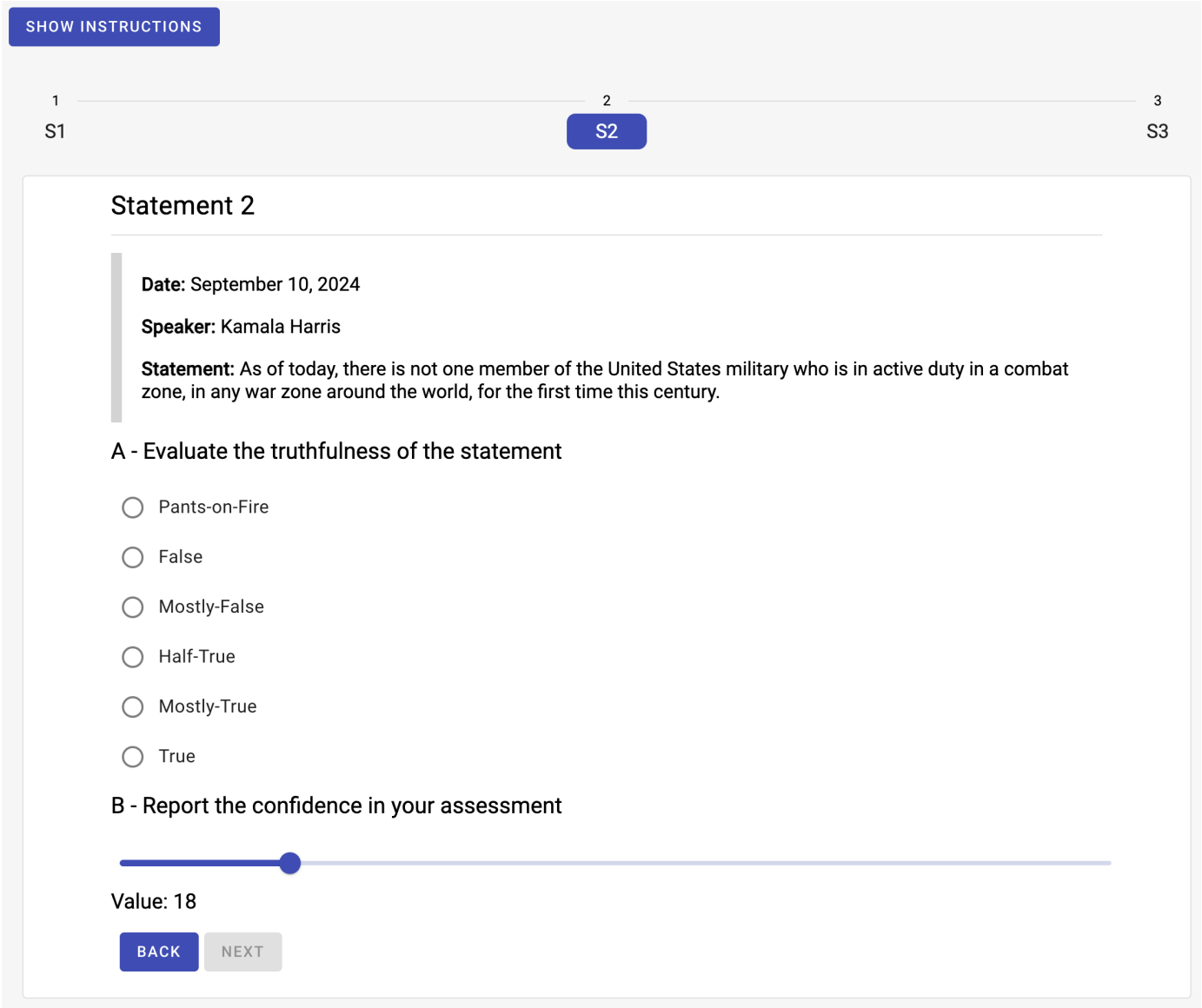
Crowdsourcing task interface

Hands-on

- What if we want to add **another evaluation dimension**?
- Ask workers to report **confidence in their assessments**

```
{
    "name": "confidence", "task_type": [ "Main" ],
    "scale": {
        "type": "interval", "min": 1, "max": 100,
        "instructions": {
            "label": "B",
            "caption": "Report the confidence in your assessment",
            "text": false
        }
    },
    "style": {
        "type": "list", "position": "middle",
        "orientation": "vertical", "separator": true
    }
}
```

Hands-on



Crowdsourcing task interface with the additional dimension

Hands-on

- We set up our truthfulness assessment task
- Now... what about **results**?
- **Download script** to **produce the final results**
  - `build/download.py`
- **Snapshots of the raw data** produced by workers during the task
- **Raw data** is **refined into a set of structured DataFrames**
  - The number of DataFrames depends on the task
  - **Different granularities** for rows and columns
- **Backup of task configuration**

Sample of **collected assessments** (`worker_answers.csv`)

```
worker_id,paid,task_name,batch_name,unit_id,try_last,try_current,action,time_submit,time_submit_parsed,
document_id,document_index,truthfulness_value,truthfulness_label,truthfulness_index,confidence_value,
time_start,time_end,time_elapsed,time_start_parsed,time_end_parsed,accesses

VAIL5CK0P6PD30,True,ITADATA2024,Demo,unit_0,1,1,Next,"Fri, 13 Sep 2024 08:56:01",2024-09-13 08:56:01,
statement_1,0,5,true,5,29.0,1726217754.0,1726217761.45,7.45,2024-09-13 08:55:54,2024-09-13 08:56:01,1

VAIL5CK0P6PD30,True,ITADATA2024,Demo,unit_0,1,1,Next,"Fri, 13 Sep 2024 08:56:06",2024-09-13 08:56:06,
statement_2,1,4,mostly-true,4.0,31,1726217761.45,1726217766.8,5.36,2024-09-13 08:56:01,2024-09-13 08:56:06,1

VAIL5CK0P6PD30,True,ITADATA2024,Demo,unit_0,1,1,Finish,"Fri, 13 Sep 2024 08:56:12",2024-09-13 08:56:12,
statement_3,2,2,mostly-false,2.0,39,1726217766.8,1726217772.47,5.67,2024-09-13 08:56:06,2024-09-13 08:56:12,1
```

Assessments of the 3 information items by a single worker (1 HIT)

Hands-on

- Now, you know a bit more about `Crowd_Frame`
- Try to perform a **real truthfulness assessment crowdsourcing task**:
  - https://bit.ly/itadata2024-demo-full
- Download the **full configuration**:
  - https://bit.ly/itadata2024-demo-full-configuration
- We conducted it in 2023 [6]
- It should take about **15-20 minutes**

[6] La Barbera D., Maddalena E., Soprano M., Roitero K., Demartini G., Ceolin D., Spina D., and Mizzaro S.. *Crowdsourced Fact-checking: Does It Actually Work?*. 2024. Information Processing & Management, Vol. 61, No. 5.

Limitations

- Implementation of **automatic allocation** of elements in HITs
- An interface for **internally monitoring the status** of tasks
- `Crowd_Frame` is (still) a **research tool**, not a product
- **Development skills are still required**…
  - …and a bit of patience

Conclusions

- `Crowd_Frame` is **designed to support crowdsourcing-based approaches**.
- It **simplifies the task design and publication** workflow.
- We used it to collect data for several studies
- If you need to conduct a crowdsourcing task...
- **...there is a good chance I might help you**

Any questions!?

Acknowledgments

This research is supported by the European Union's NextGenerationEU PNRR M4.C2.1.1 – **PRIN 2022 project** \lq\lq 20227F2ZN3 *MoT-The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness* (CUP G53D23002800006)