Human-Machine Information Extraction Simulator for Biological Collections

Ícaro Alzuru, Aditi Malladi, Andréa Matsunaga,
Maurício Tsugawa, and José A.B. Fortes

3rd IEEE HMData 2019
December 9th, 2019

HuMaIN is funded by a grant from the National Science Foundation's ACI Division of Advanced Cyberinfrastructure (Award Number: 1535086). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

NSF

UF UNIVERSITY of FLORIDA

AGENDA

- Digitization of Biological Collections
- HuMaIN Project
- Self-aware Information Extraction (SELFIE) Model
- The HuMaIN Simulator
- Experiments & Results
- Human-in-the-loop Example
- Conclusions

Digitization of Biocollections

- Information in biocollections can be used to understand pests, biodiversity, climate change, natural disasters, diseases, and other environmental issues.
- There are about 1 Billion specimens in Biocollections in the United States and about 3 Billion in the whole World (Estimated).
- NSF's Advancing Digitization of Biodiversity Collections (ADBC) program.

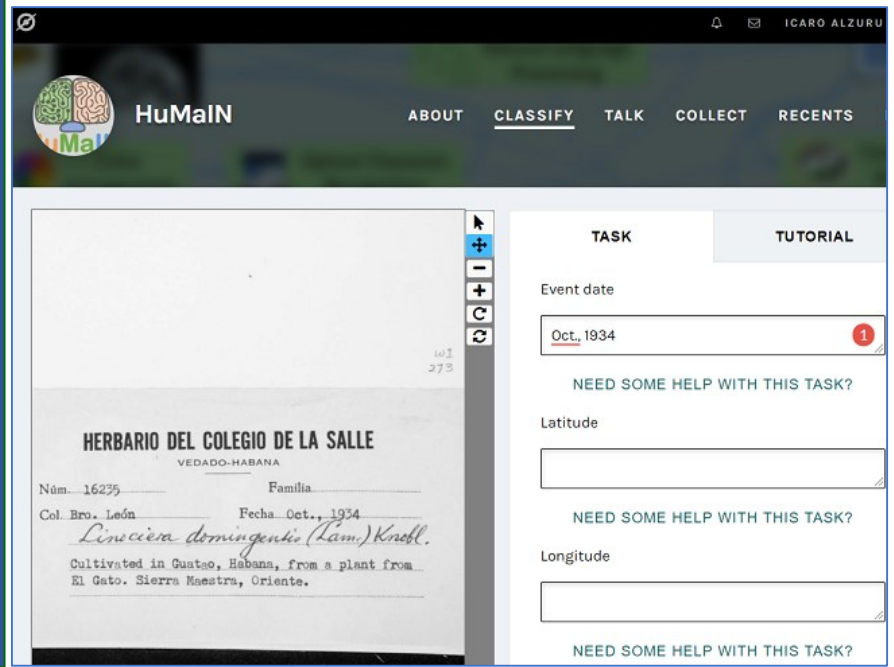
Photo by Chip Clark. U.S. National Herbarium at the Smithsonian Institution's National Museum of Natural History. Featured researchers: Dr. James Norris (right, front), research assistant Bob Sims (left, front), and associate researcher, Katie Norris (left, back).

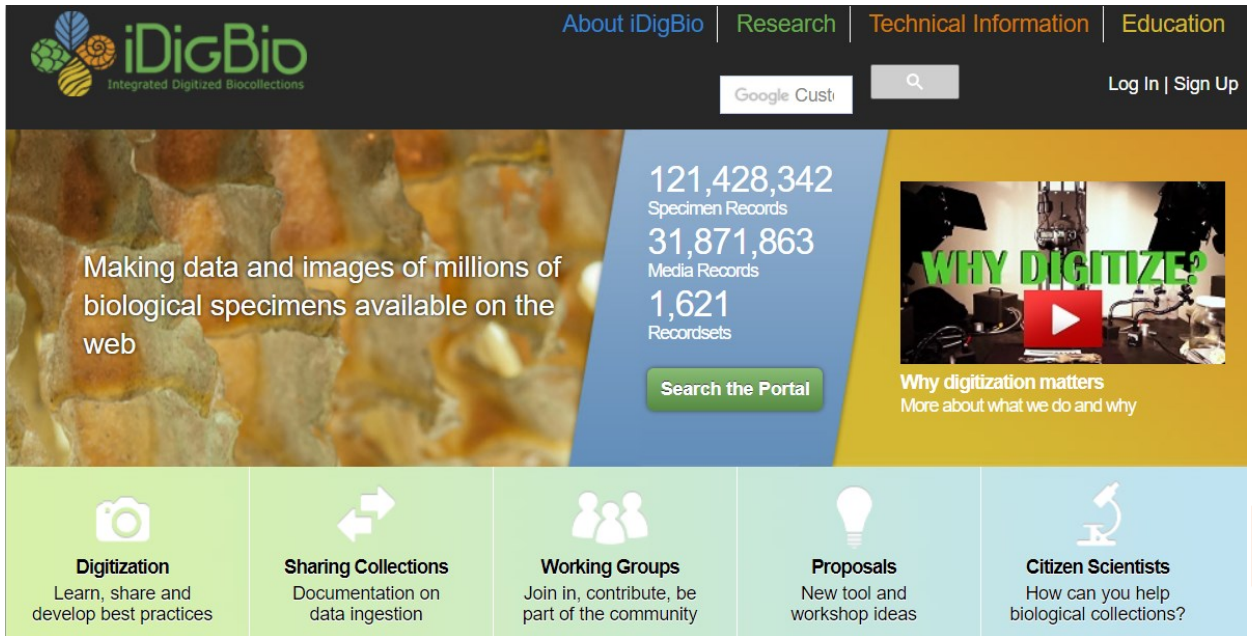
Photo by Chip Clark. Bird Collection, Dept. of Vertebrate Zoology, Smithsonian Institution's National Museum of Natural History. In the foreground: Roxie Laybourne, feather identification expert.

Current Digitization Process

Digitization:

1. Curators photograph each specimen together with their correspondent labels.
2. Transcription of the metadata in a database (commonly performed by **volunteers**)
3. Final metadata values are shared in a digitization repository.



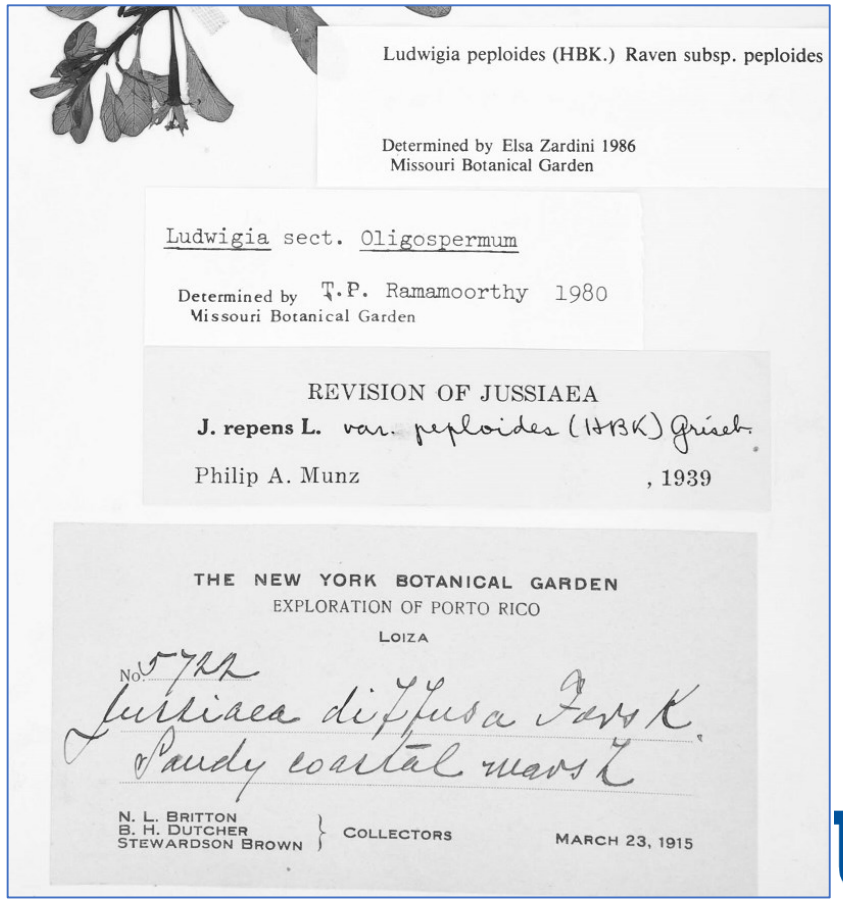


UNIVERSITY of FLORIDA

The Challenge of Automated Information Extraction

Automated IE: Optical Character Recognition + Natural Language Processing

- Biocollections' images are problematic for OCR engines
- OCR result is not perfect. Handwritten text is especially problematic.

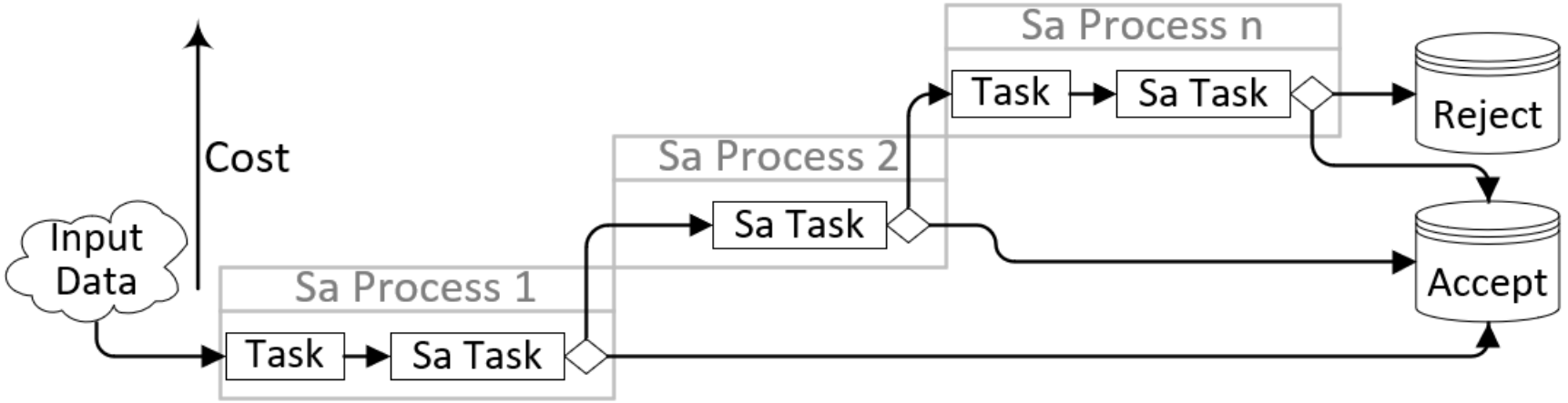


ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

HuMaIN – SELFIE Workflows

- **HuMaIN**: Human-Machine Intelligent Network of software components for the digitization of biocollections. http://humain.acis.ufl.edu
- We propose a **SELFIE** (Self-aware IE) workflow model for the transcription of biocollections' labels (https://doi.org/10.1109/eScience.2017.19)
- SELFIE organizes the IE alternatives (SaPs) in **incremental-cost order**.
- SaPs **estimate** the **confidence** of the extracted value, decide if it must be accepted or not, and send the unprocessed images to the next SaP.



The HuMaIN Simulator

- Problem: Human-Machine IE workflows require images, crowdsourcing interfaces, volunteers, ground-truth values, scripts to process data, etc. Researchers invest a lot of time and resources validating an idea.
- Objectives:
  - Promote and accelerate research in the area of IE from biocollections' labels.
  - Share the IE workflows, crowdsourced data, ground-truth data, and ideas.
  - Encourage biodiversity institutions and repositories to share their data in a more valuable format for data engineers.
- How does the simulation work?
  - Simulation: The execution of the task is emulated
    - The results from previous executed tasks are reused
  - Not all the tasks need to be simulated. But we recommend to execute only the task under study.

Configurable Components in a Simulation

Simulation Configuration

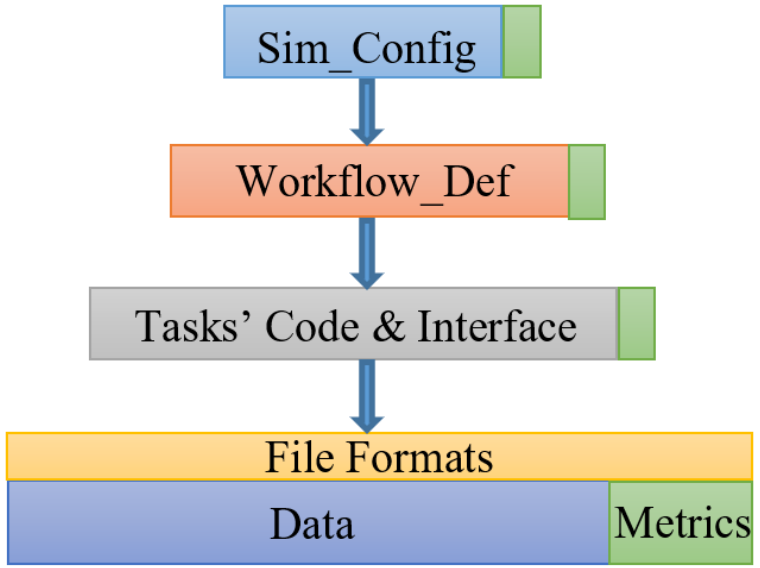
```
<tasks>
  …
  <task name="ocr_sim">
    <parameter name="ocr_input_dir">…/ocropus</parameter>
    …
    <parameter name="output_dir">…/ocr_sim</parameter>
  </task>
  …
</tasks>
<metrics>
  …
  <script name="quality_measure.py">
    <parameter name="accepted_f">…/regexp/accepted.tsv</parameter>
    …
    <parameter name="output_file">…/ed/quality.csv</parameter>
  </script>
  …
</metrics>
<post-processing>
```

Task's Interface

```
<task name="ocr_sim">
    <parameter name="ocr_input_dir" type="D_TXT"></parameter>
    <parameter name="include" type="STRING"></parameter>
    <parameter name="specimens_list" type="TXT"></parameter>
    <parameter name="metric" type="STRING"></parameter>
    <parameter name="output_dir" type="O_D_TXT"></parameter>
</task>
```

Workflow Definition

```
ocr_sim
ed_reg_expr_sim, ocr_sim
crowdsource_sim, ed_reg_expr_sim
consensus_sim, crowdsource_sim
```



File with the value to use for the parameter of every task

File that defines the order of simulation of the tasks

Interface and code of every simulated or executed task

Data files' internal format

Raw data, tasks' output and their metric values

Datatypes

```
INPUT_TYPES = ['INT', 'FLOAT', 'STRING', 'JPG', 'TXT',
'TSV', 'D_JPG', 'D_TXT', 'D_AR']
OUTPUT_TYPES = ['O_JPG', 'O_TXT', 'O_TSV', 'O_D_AR',
'O_D_JPG', 'O_D_TXT']
```

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

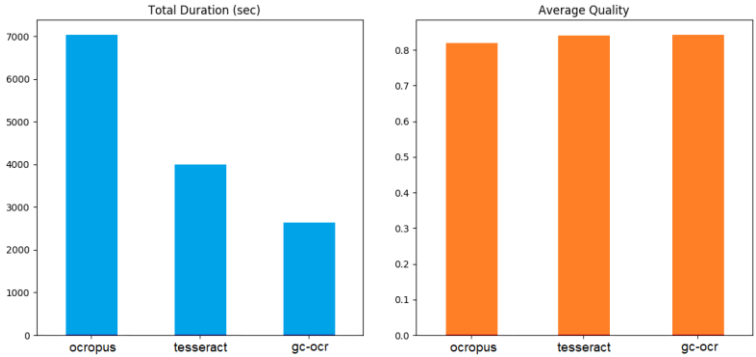
UF UNIVERSITY of FLORIDA

Features of the HuMaIN Simulator

- Simulation Engine and Features to facilitate the experimental validation of the IE from biocollections' images.
  - Simulation groups
  - Generation of tables, box plots, and bar graphs to visualize results and compare different simulations
  - Human-in-the-loop (iterative) simulations
- 3 IE workflows and their scripts (code) are shared
- Datasets with crowdsourced data, ground-truth values, and IE results
- 4 experiments on existent workflows are also shared
- Open Source: https://github.com/acislab/HuMaIN_Simulator

Experiments (1/2)

- IE workflow for the Event-date.
  - How to define workflows, tasks, and simulation-configuration files
  - The results for the quality and execution-time metrics are shown.
  - **Experiment** about how the quality of the OCR engine affects the final quality of the workflow.
    - Three different OCR engines are used. This experiment exemplifies how to compare tasks and implement groups of simulations.
  - **Experiment** shows how different crowds may affect the quality of the workflow's output.

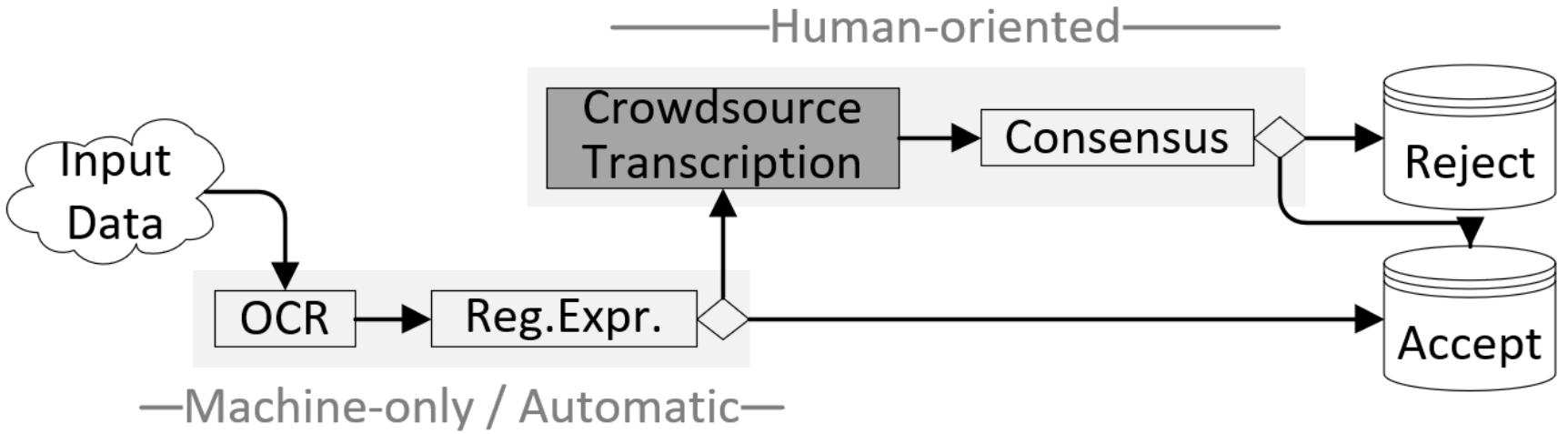


Total Duration (sec)

Zooniverse's Volunteers	Paid Students
7033.3	5960.4

Average Quality

Zooniverse's Volunteers	Paid Students
0.8185	0.8338

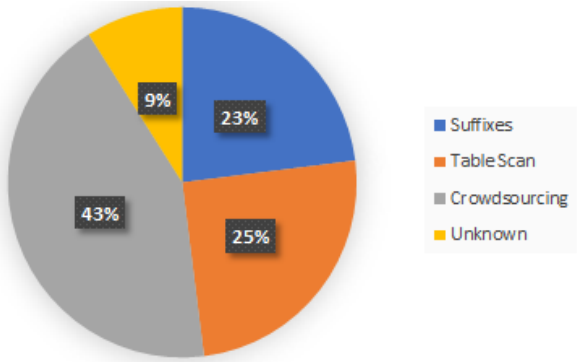


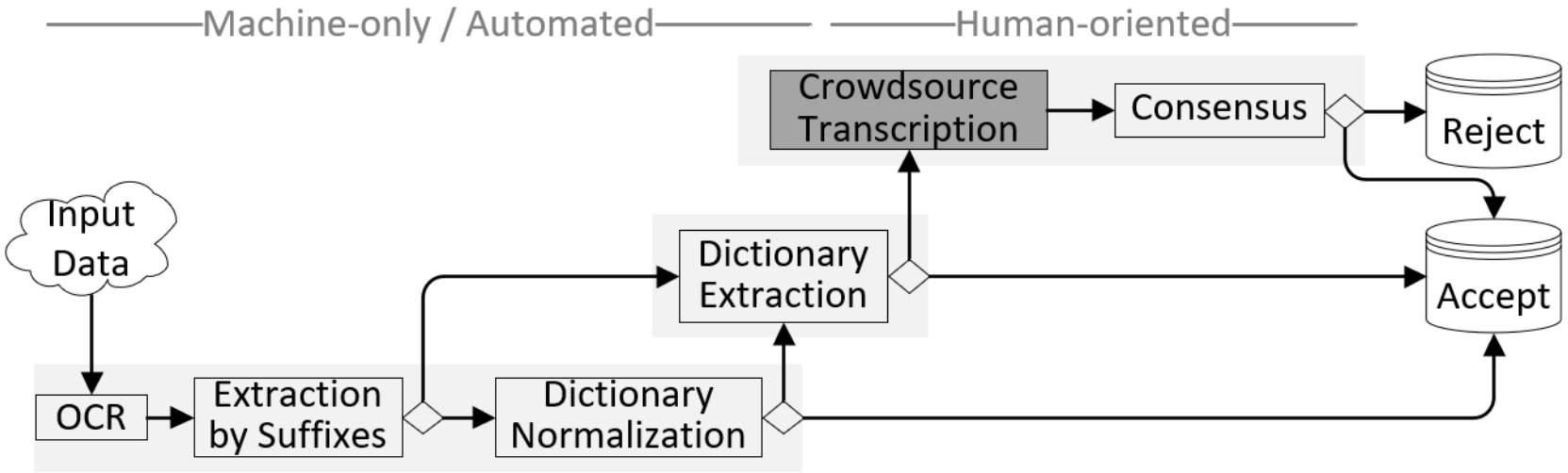
ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

Experiments (2/2)

- Workflow for the extraction of the Scientific-name.
  - The results for the quality and execution-time metrics are shown.
  - **Experiment** about how to tune a parameter: the similarity threshold that decides when to accept or reject an extracted value.
    - A group of simulations is utilized in this experiment.

Similarity Threshold	0.5	0.55	0.6 - 0.85	0.9 – 1.0
Number of Accepted Values	37	36	25	24
Avg. Similarity to Ground-truth	0.53	0.55	0.63	0.63



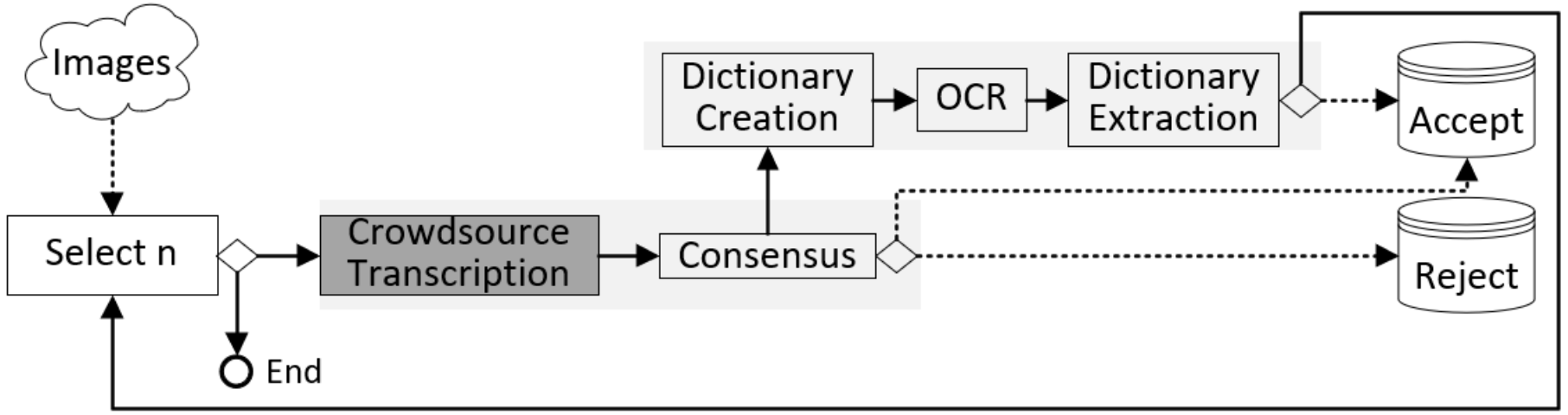


ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

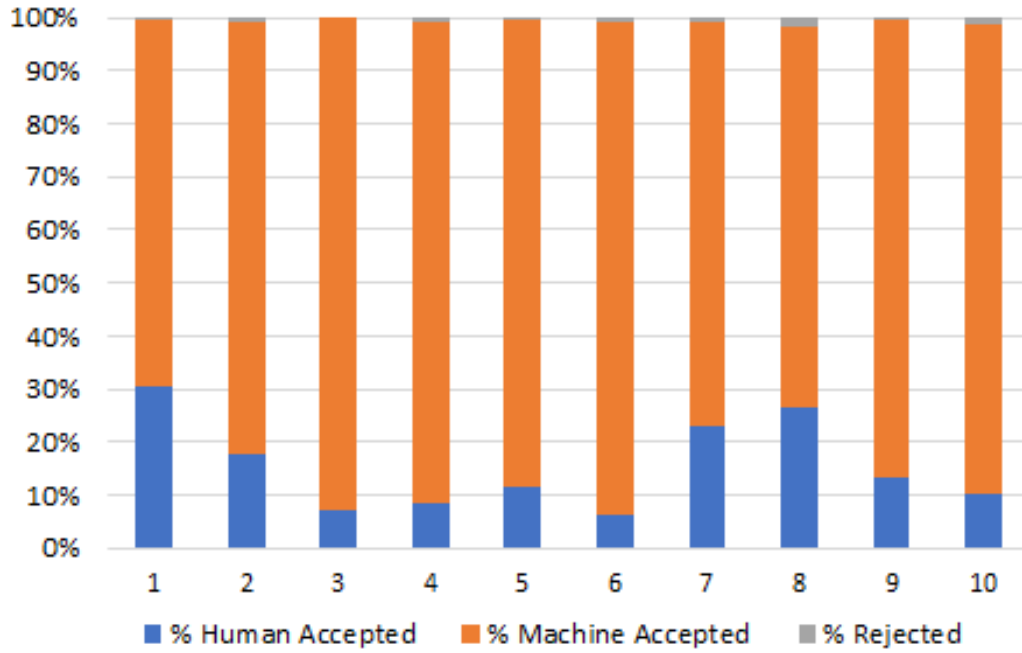
Example: Human-in-the-loop Experiment

Collector Extraction

(Recorded-by)



Collection	1	2	3	4	5	6	7	8	9	10
# Images	739	2880	1041	1639	2152	704	901	954	1252	1971
# Iterations	5	11	2	3	6	1	5	6	4	5
Human Accepted	224	504	73	136	252	45	207	253	169	205
Machine Accepted	511	2359	967	1489	1897	654	685	686	1077	1743
Rejected	4	17	1	14	3	5	9	15	6	23



ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

Conclusions

- This paper proposes a human-machine simulator for the extraction of the specimens' metadata.
- The IE workflows can include executed and simulated tasks. The simulated tasks reuse the output of tasks previously executed.
- The simulator permits to accelerate the experimental process by copying and reusing workflows, tasks, simulations, and data.
- Groups of simulations can be automatically generated by specifying different parameter values, while Human-in-the-loop capabilities allow running iterative simulations that incrementally improve automated tasks from the data generated by humans. Embedded graphical capabilities permit to generate tables, box plots, and bar graphs to easily visualize the results and compare different simulations.
- After implementing a workflow in the HuMaIN Simulator, several experimental scenarios can be easily explored: parameter tuning, tasks comparison, evaluation of IE approaches, and HITL workflows.
- The process of definition of the components of a workflow was detailed, while three workflows and four experiments were presented to exemplify the research process and potentiality offered by the HuMaIN Simulator.
- Available at https://github.com/acislab/HuMaIN_Simulator

Thank you

Questions?

This work is supported in part by the National Science Foundation (NSF) grants No. ACI-1535086 and No. EF-1115210, and the AT&T Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or the AT&T Foundation.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA