ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

HuMaIN: Human- and Machine-Intelligent Network Software Elements

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

Introduction and Motivation

Data scientists spend extensive **time, effort, and resources** collecting, integrating, curating, transforming, and assessing data quality before actually performing discovery analysis.

Data is often in **non-structured** form and **incompatible** with analytics tools.

There are two main approaches to deal with these challenges:

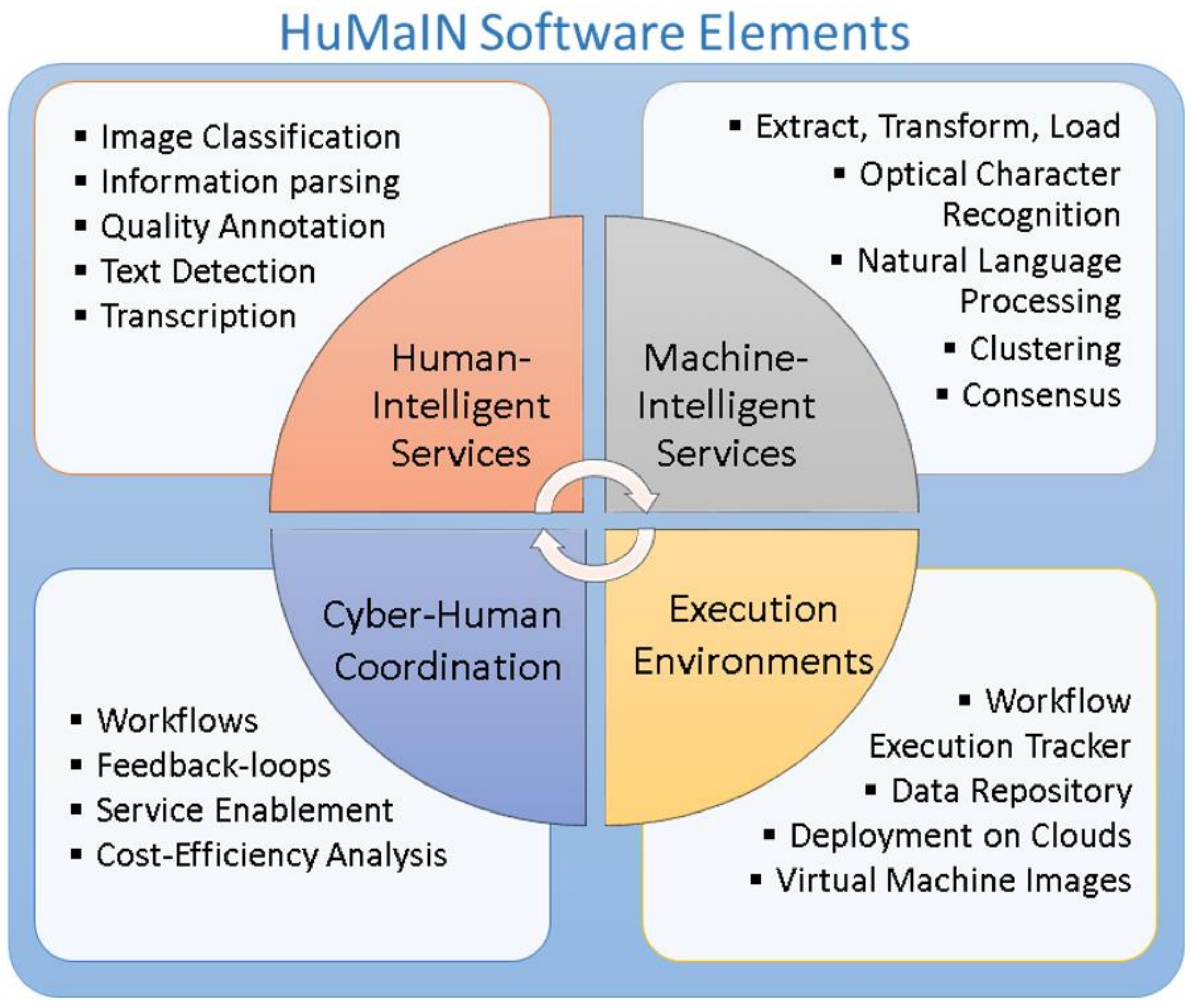
❑ **Crowdsourcing** (Human-Intelligent processes)

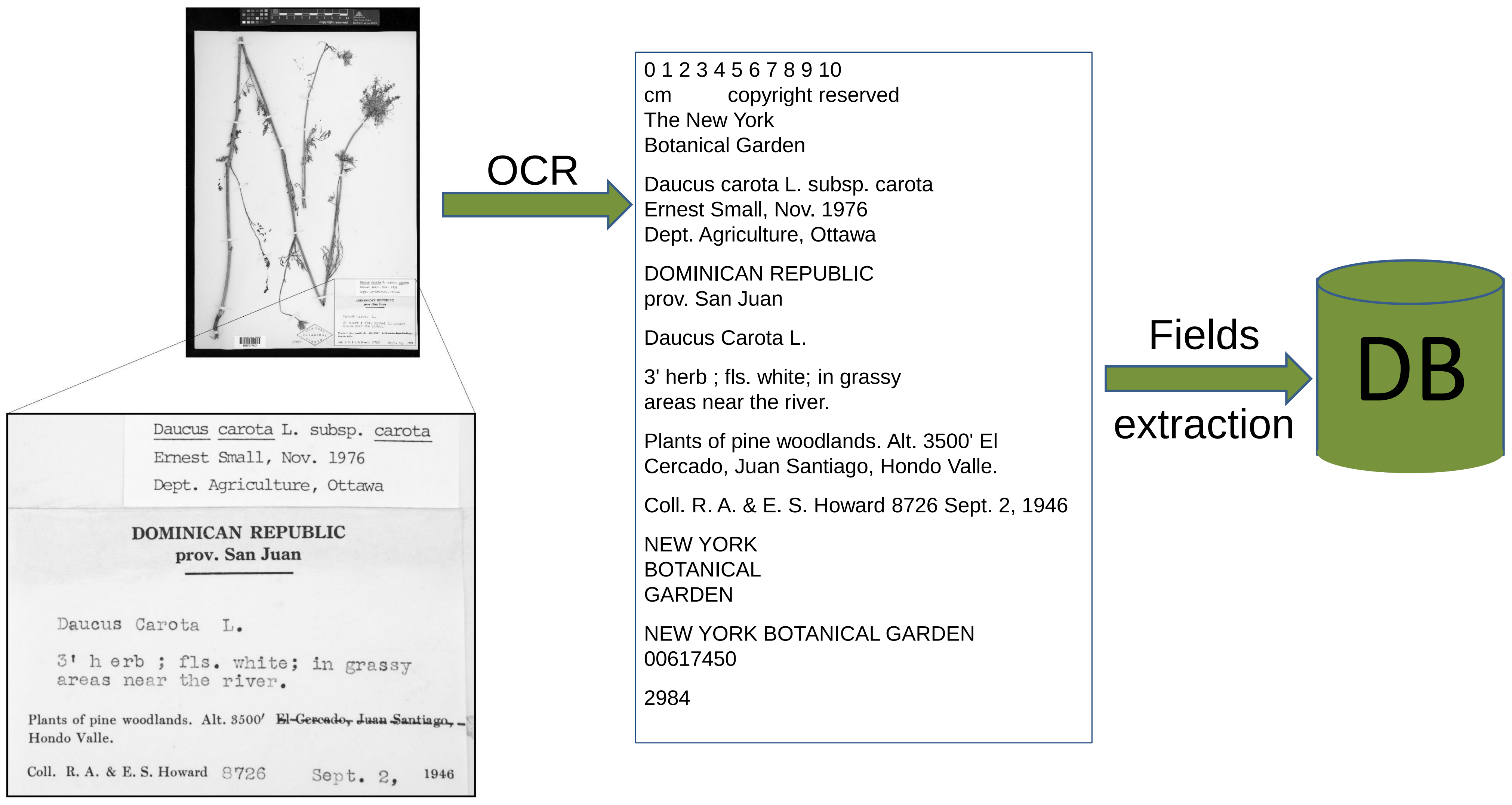
❑ **Machine Learning** (Machine-Intelligent processes)

Each method has its strengths and weaknesses. However, very little has been done to combine and **simultaneously** take advantage of both types of solutions.

The goal of the Human- and Machine-Intelligent Network (**HuMaIN**) project is to accelerate scientific digitization through the **integration and synergistic cooperation of human and machine processing** in order to overcome hurdles and bottlenecks found in data digitization.

The data collected in the **Integrated Digitized Biocollections (iDigBio)** project is used as a use case or **motivating example** for information extraction. https://www.idigbio.org





Daucus carota L. subsp. carota
Ernest Small, Nov. 1976
Dept. Agriculture, Ottawa
DOMINICAN REPUBLIC
prov. San Juan
Daucus Carota L.
3' h erb ; fls. white; in grassy
areas near the river.
Plants of pine woodlands. Alt. 3500' ~~El Cercado, Juan Santiago,~~
Hondo Valle.
Coll. R. A. & E. S. Howard 8726 Sept. 2, 1946

0 1 2 3 4 5 6 7 8 9 10
cm copyright reserved
The New York
Botanical Garden

Daucus carota L. subsp. carota
Ernest Small, Nov. 1976
Dept. Agriculture, Ottawa

DOMINICAN REPUBLIC
prov. San Juan

Daucus Carota L.

3' herb ; fls. white; in grassy
areas near the river.

Plants of pine woodlands. Alt. 3500' El
Cercado, Juan Santiago, Hondo Valle.

Coll. R. A. & E. S. Howard 8726 Sept. 2, 1946

NEW YORK
BOTANICAL
GARDEN

NEW YORK BOTANICAL GARDEN
00617450

2984

Development Plan and Deliverables

1. **Machine-Intelligent Components**
   - ❑ Interface to OCRopy tool to manage training sets for different fonts
   - ❑ Set of alternative methods for OCRopy to deal with noise
   - ❑ Selecting and integrating Carrot^2 clustering algorithms and parameters
2. **Human-Intelligent Components**
   - ❑ Javascript sensors to detect the number, time, and sequence of user interactions
   - ❑ PyBossa extension to support configurable and reusable microtasks
3. **Machine-Intelligent Services Enablement**
   - ❑ Automatic generation of RESTful services using CLAWS (Command-Line Application Wrapper service)
   - ❑ Extending PyBossa to support configurable and reusable microtasks
4. **Human-Intelligent Services Enablement**
   - ❑ PyBossa extension to allow management of batches of tasks and user qualification
   - ❑ Enabling complex tasks by composing micro-tasks developed by this project
   - ❑ Evaluation of alternative human-intelligent workflows using sensors from step 2
5. **Building workflows with Human- and Machine-Intelligent Services**
   - ❑ Using only machine-intelligent services (image binarization, OCR, and NLP)
   - ❑ Using only human-intelligent services (image selection, text interpretation, and transcription)
   - ❑ Using both human- and machine-intelligent services that improve machine-only and human-only processes
6. **Online feedback-loops between Human- and Machine-Intelligent Services**
   - ❑ Workflow with CrowdConsensus controlling a multi-step text interpretation workflow
   - ❑ Workflow with OCR errors triggering need for additional training sets
   - ❑ Workflow with chain of user expertise controlling the need for assessment of a worker
7. **Execution Environments**
   - ❑ Dedicated private compute-and-storage cloud for HuMaIN research and development.
   - ❑ Middleware to support workflows and feedback loops
   - ❑ Tutorials and how-to documents
8. **Cyber-Human System Cost Efficiency**
   - ❑ Cost-efficiency comparative analysis of 1. and 7.
   - ❑ Surveys of selected users of HuMaIN

Goals

❖ Research and development of HuMaIN software elements in four main areas:

- ❑ **Human-Intelligent services**
- ❑ **Machine-Intelligent services**
- ❑ **Cyber-Human Coordination**
- ❑ **Execution Environments**

❖ **Platform** for reusing the HuMaIN software elements as RESTful **services**.

Challenges

❖ OCR (Optical Character Recognition): Text mixed with other elements (cropping required), different fonts and sizes, handwritten text, different languages, underlined text, overlapped text, OCR performance.

❖ Information extraction: Data cleaning, multiple formats, incomplete data, data completion, natural language processing, field value standardization, consensus, process efficiency, deduplication, ambiguity, spelling errors, dictionaries, abbreviations / data truncation.

Progress and Results

❖ Hardware platform, system software, and web site:

http://humain.acis.ufl.edu

❖ **OCRopy** (https://github.com/tmbdev/ocropy) tested and selected as the OCR software for the HuMaIN project

- ❑ **Scripts** to automate the OCR process, detection of the text language, and fields extraction (date, country).
- ❑ **Cropping** the text area of the image improves the OCR result.
- ❑ Without training or cropping the text areas, **OCRopy identifies only 42% of the characters** of images hosted by iDigBio.

❖ Started the **5th step of the Development Plan** to address observed OCR limitations:

- ❑ Human-only and machine-only workflows were setup for digitizing the label of scientific data from the iDigBio project.
- ❑ Two hybrid workflows prepared to demonstrate that these perform better than the human-only or machine-only approaches.
- ❑ Public access to developed crowdsourcing tasks and progress at: **http://humain.acis.ufl.edu/app.html**

Summary and Conclusions

❖ Discovering information in non-digital records via digitization and information extraction remains a challenging problem with imperfect solutions.

❖ Combined human and machine approaches address weaknesses found when independently applying each of these approaches.

❖ Long term goal of HuMaIN project is to provide a platform of reusable services for combined human- and machine-intelligent to improve the processing of digitized biocollections.

NSF

NSF SI2 PI Workshop : February 16-17, 2016

This material is based in part upon work supported by the National Science Foundation under Grant No. 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.

UF UNIVERSITY of FLORIDA