

Introduction and Motivation

Data scientists in all communities spend the majority of their **time and effort** collecting, integrating, curating, transforming, and assessing quality before actually performing discovery analysis.

Data is often in **non-structured** form, **not compatible** with analytics tools.

Two main approaches to deal with these challenges:

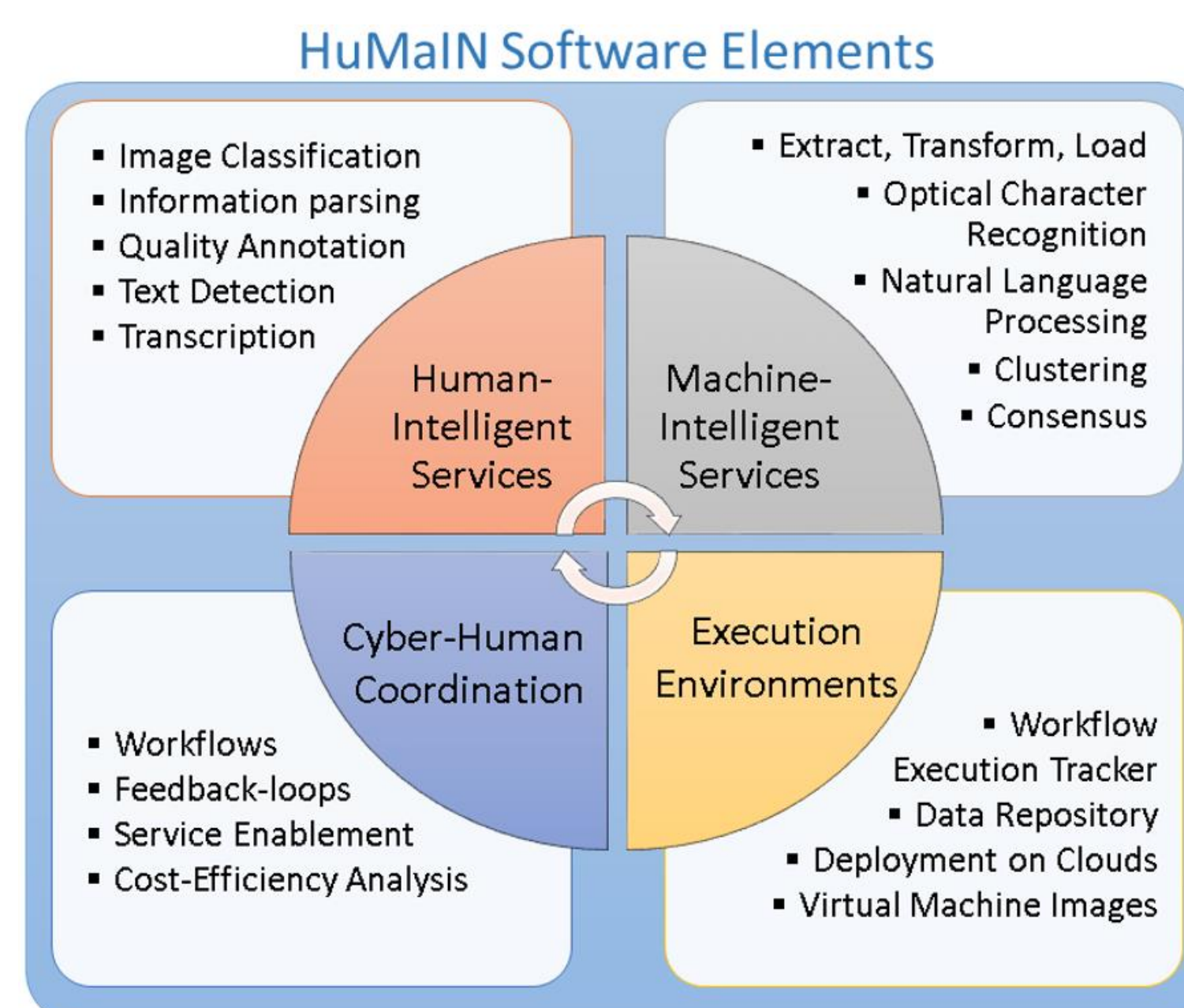
- ☐ **Crowdsourcing** (Human-Intelligent processes)
- ☐ **Machine Learning** (Machine-Intelligent processes)

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

Usually, implementing crowdsourcing or machine learning solutions demand a lot of **time** and **resources**.

The vision of the Human- and Machine-Intelligent Network (**HuMaIN**) project is to accelerate scientific digitization through fundamental advances in the **integration and mutual cooperation between human and machine processing** in order to handle practical hurdles and bottlenecks present in data digitization.

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



Goals

- ❖ Research and development of HuMaIN software elements in four main areas:
 - ☐ **Human-Intelligent services**
 - ☐ **Machine-Intelligent services**
 - ☐ **Cyber-Human Coordination**
 - ☐ **Execution Environments**
- ❖ Providing a **platform** for reusing the HuMaIN software elements as RESTful **services** in other projects.

Experimental Progress and Results

- ❖ The hardware platform, software, and web site for the HuMaIN Software Elements project was **setup**: <http://humain.acis.ufl.edu>
- ❖ **OCROpy** (<https://github.com/tmbdev/ocropy>) is being tested as the OCR software for the HuMaIN project
 - ☐ Several **scripts** have been created to automatize the process, detecting the language of the text, and extracting some fields: date, country.
 - ☐ **Cropping** the text area of the image has importantly improved the quality of the OCR result.
- ❖ These first tries of the OCR process made us decide beginning by the **5th step of the Development Plan**:
 - ☐ Human-only and machine-only workflows were setup for digitizing the label of scientific data from the iDigBio project.
 - ☐ 2 Hybrid workflows were also prepared and we expect to demonstrate these perform better than the human-only or machine-only approaches.
 - ☐ Anybody can **help us** to complete the crowdsourcing tasks at: <http://humain.acis.ufl.edu/app.html>



OCR

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.

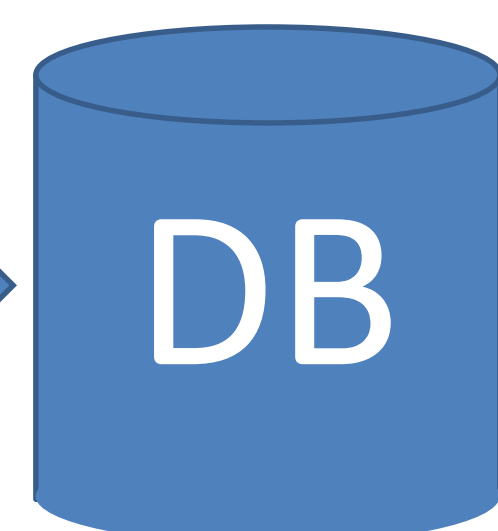
Coll. R. A. & E. S. Howard 8726 Sept. 2, 1946

NEW YORK BOTANICAL GARDEN

NEW YORK BOTANICAL GARDEN
00617450

2984

Fields
extraction



Development Plan and Deliverables

1. Machine-Intelligent Components

- ☐ Adding an interface to OCROpy tool to manage training sets for different fonts
- ☐ Expose OCROpy to a set of alternative methods to deal with noise
- ☐ Selecting and integrating Carrot² clustering algorithms and parameters

2. Human-Intelligent Components

- ☐ Create a set of Javascript sensors to detect the number, time, and sequence of user interactions
- ☐ Extending PyBossa to support configurable and reusable microtasks

3. Machine-Intelligent Services Enablement

- ☐ Implementing the 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 with management of batches of tasks and user qualification
- ☐ Set of complex tasks making use of multiple developed micro-tasks
- ☐ Evaluation of alternative human-intelligent workflows using sensors from step 2

5. Workflows with Human- and Machine-Intelligent Services

- ☐ Build a workflow with only machine-intelligent services (image binarization, OCR, and NLP)
- ☐ Build a workflow with only human-intelligent services (image selection, text interpretation, and transcription)
- ☐ Build a workflow where human- and machine-intelligent services improve machine-only and human-only processes.

6. Feedback-loops between Human- and Machine-Intelligent Services

- ☐ Online feedback loop workflow with CrowdConsensus controlling a multi-step text interpretation workflow.
- ☐ Online feedback loop workflow with OCR errors triggering need for additional training sets.
- ☐ Online feedback loop 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-effect comparative analysis of 1. and 7.
- ☐ Surveys with selected customers of HuMaIN.

Challenges

- ❖ **OCR (Optical Character Recognition)**: Text mixed with other elements (cropping), different fonts and sizes, handwritten text, different languages, underlined text, overlapped text, OCR performance.
- ❖ **Information extraction**: Data cleaning, multiple formats, incomplete data / completion, natural language processing, field value standardization, consensus, process efficiency, deduplication, ambiguity, spelling errors, dictionaries, abbreviations / data truncation.

Future Work

- ❖ The 5th step of the Development Plan will be completed
- ❖ Based on the gained experience, we are going to continue with steps 2 and 1 of the Development Plan, making these software components reusable and the improving the workflows, environment, and efficiency.

Summary and Conclusions

- ❖ Human- or machine-only approaches for information extraction have weaknesses that the integration of human- and machine-intelligent processes can improve.
- ❖ Information extraction is a complex problem with imperfect solutions that can rescue the knowledge buried in not digitized formats.
- ❖ HuMaIN project will provide a platform of reusable services for the human- and machine-intelligent processing.

Acknowledgements

- ❖ **iDigBio (Integrated Digitized Biocollections)**: National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210).