HuMaIN: Human- and Machine-Intelligent Network of Software Elements

Ícaro Alzuru, Andréa Matsunaga, Maurício Tsugawa , Jingchao Luan, 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:

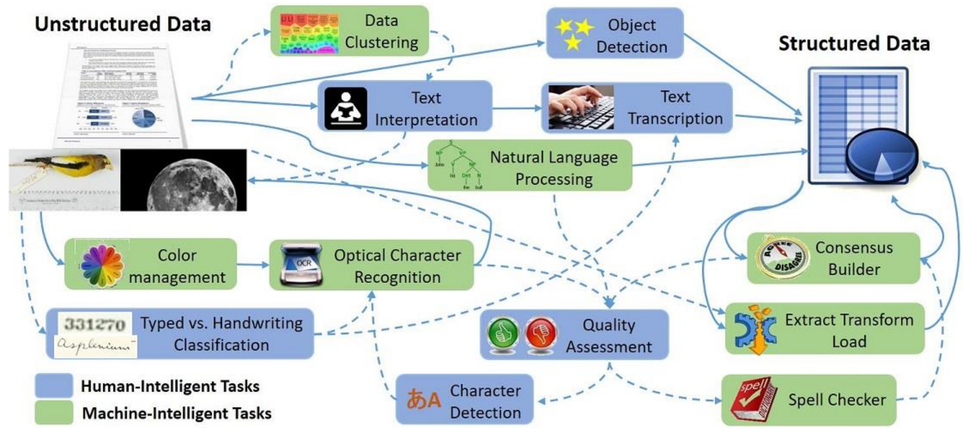
❑ **Crowdsourcing** (Human-Intelligent processes)

❑ **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 information extraction challenges present in the **Integrated Digitized Biocollections (iDigBio**, https://www.idigbio.org**)** project are used as a **motivating use case**.



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**.

Progress and Results

❖ Devised, developed and demonstrated hybrid human- and machine-intelligent approaches to digitize labels from biocollections specimens.

❖ Ongoing Studies and Experiments:
- ❖ Task Complexity in Crowdsourcing
- ❖ Self-aware Data Extraction

❖ Web applications and data processing tools developed to support crowdsourcing and machine intelligent processes.

https://github.com/acislab/HuMaIN_Collaborative_Data_Extraction

❖ Paper presented at the 2016 IEEE 12th International Conference on eScience: Cooperative Human-Machine Data Extraction from Biological Collections.

Cooperative Human-Machine Data Extraction

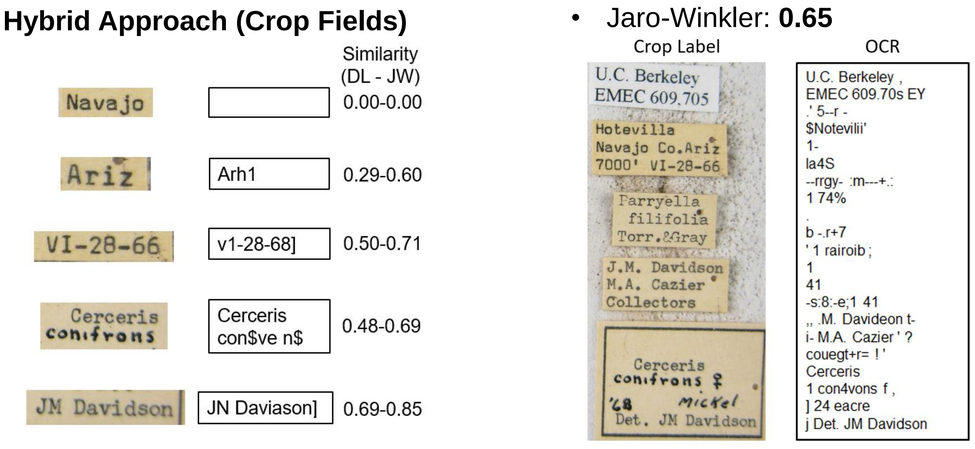
❖ **Cooperative Human-Machine Data Extraction from Biological Collections**

0. Human-only approach:
Extracted from the paper "Reaching Consensus in Crowdsourced Transcription of Biocollections Information", A. Matsunaga, A. Mast, and J. A.B. Fortes.
- Consensus found 86.7% of times with accuracy of 91.1% => **79%** correct

1. Machine-only Approach:
OCR the whole image with OCRopus and Tesseract. Obtained similarity:
- Damerau-Levenshtein: **0.36**
- Jaro-Winkler: **0.59**

2. Hybrid Approach (Crop Label):
- Damerau-Levenshtein: **0.49**
- Jaro-Winkler: **0.65**

3. Hybrid Approach (Crop Fields)



Average similarity and improvement with respect to A1

	Entomology	Bryophyte	Lichen
1. Machine-only	0.27	0.38	0.64
2. Hybrid (Crop Label)	**0.52 – 93%**	0.61 – 61%	**0.66 – 3%**
3. Hybrid (Crop Fields)	0.43 – 59%	**0.67 – 76%**	0.64 – 0%

- Hybrid approaches improve similarity with respect to the machine-only approach (1) up to a factor of 1.93.
- No improvement for Lichens, because their images contain only text (a Label)
- Cropping fields eliminate the need of NLP, adding interpretation.

Time, Cost, and Similarity

Approach	Human + Machine (Time in years)	Cost ($ in Millions)	Recognition rate or Similarity
0. Human-only	17123 + 0 (17123)	1500.00	**0.79**
1. Machine-only	0 + 1202 (1202)	**3.61**	0.43
2. Hybrid (Crop Label)	580 + 422 (**1002**)	52.10	0.60
3. Hybrid (Crop Fields)	6342 + 1218 (7560)	559.21	0.58

Task Complexity in Crowdsourcing (in progress)

❖ **Participate at: http://humain.acis.ufl.edu/complexity**

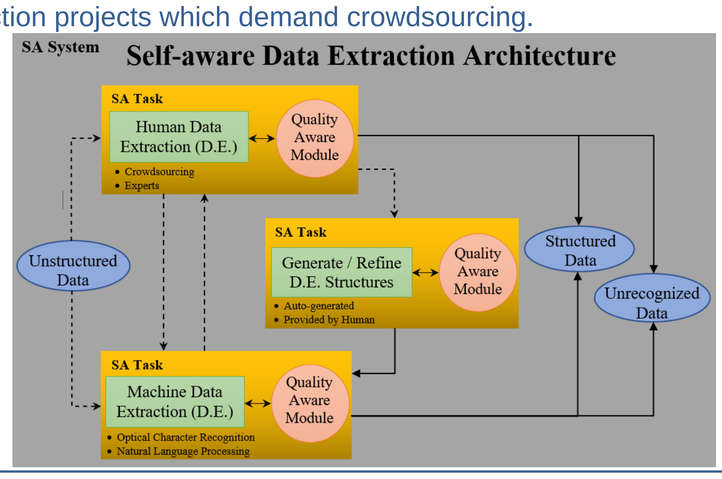
❖ **Goals:**
- Finding what task size generates the best quality and completion rate.
- Studying three types of interfaces: transcribing, selecting, and cropping, and their perceived complexity, productivity, and friendliness.

❖ **Experiments progress:**
- 16 volunteers of the Florida Museum of Natural History participated in the experiments
- 6 members of the ACIS Lab evaluated and took part in the experiments
- 10 volunteers used the cropping webapp during the 2016 WeDigBio Transcription Blitz.
- New participants will be paid for completing one hour experiment's sessions.
- A Zooniverse project was created and has collected data from hundreds of volunteers: Participate at **https://www.zooniverse.org/projects/ialzuru/humain**

Self-aware Data Extraction (in progress)

❖ **Participate in our data collection process, at: http://humain.acis.ufl.edu/aware**

❖ **Goals:**
- Designing Data Extraction processes able to self-evaluate and take smart decisions.
- Reducing the amount of crowdsourcing required and therefore, the time and cost of data extraction projects which demand crowdsourcing.



Summary and Conclusions

❖ The combined execution of human- and machine-intelligent techniques addresses weaknesses found in crowdsourcing-only or machine-only 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.

❖ Self-aware machine intelligence will have the ability to determine when it is necessary to engage human intelligence. This is part of ongoing work.

NSF SI2 PI Workshop : February 21-22, 2017

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.