Fosscomm 2020

A Personalized Artificial-Intelligence-enabled Method for Efficient Research in Ethnopharmacology

A.Kontogiannis E.Axiotis G.Giannakopoulos

Overview

- *Ethnopharmacology*: the scientific study of substances used medicinally, especially folk remedies, by different ethnic or cultural groups
- We propose an AI system that helps researcher find documents, relevant to ethnopharmacology topics
- Relevant ethnopharmacology topics are predefined by researcher
- Focused search with interaction between system and researcher

Motivation

- Documentation of indigenous knowledge on the use of plants is difficult
- Focused search of ethnopharmacological references related to certain places and plant species is a very challenging task
- Big Data challenges

The vision

- An **open science community**
- collaborating on an **open source system**
- supporting their research **through AI**

Helping researchers do research

- **Researcher:**

  - define their relevant topics,

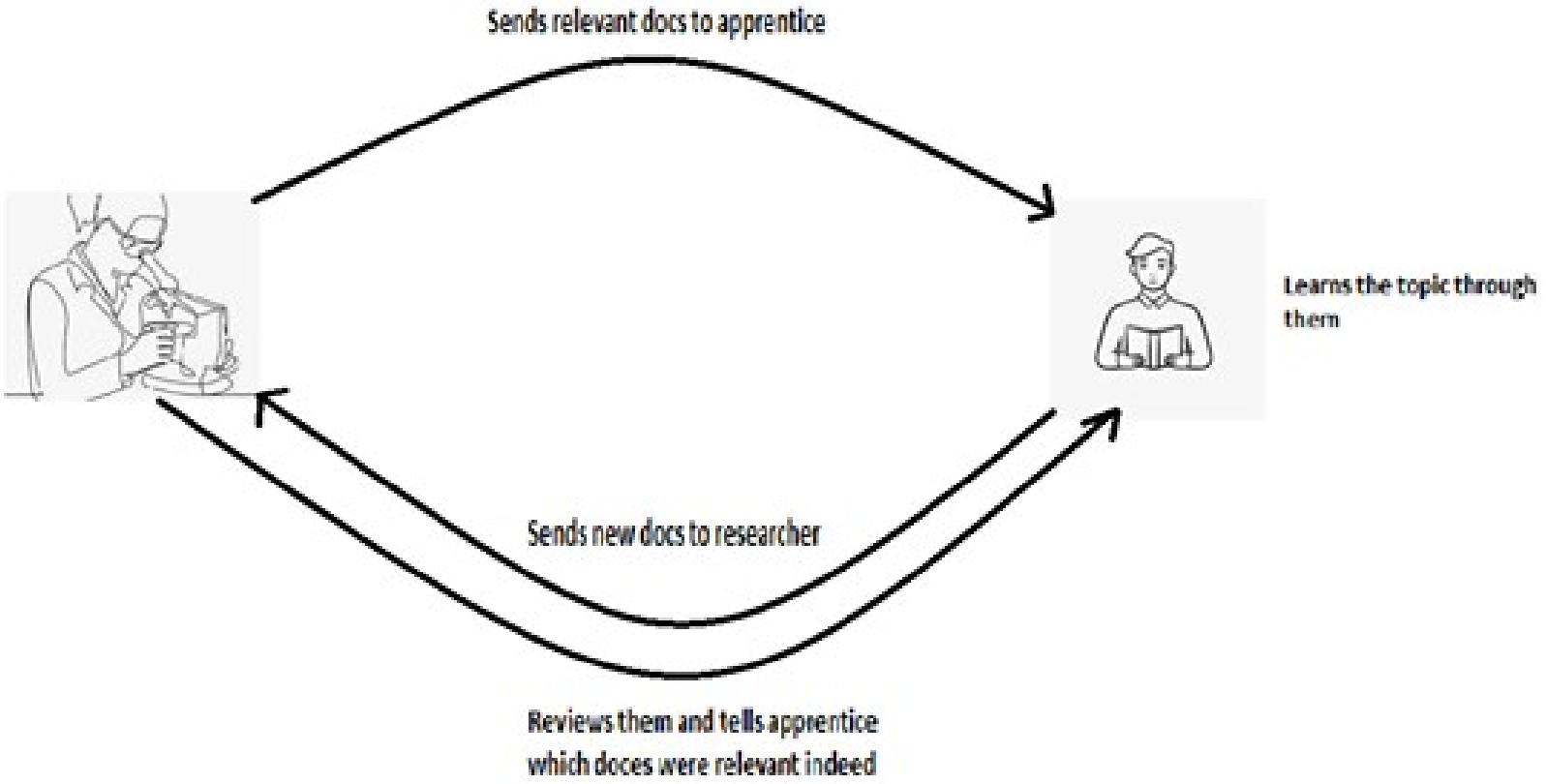
  - tell these topics to the system

- **System:**

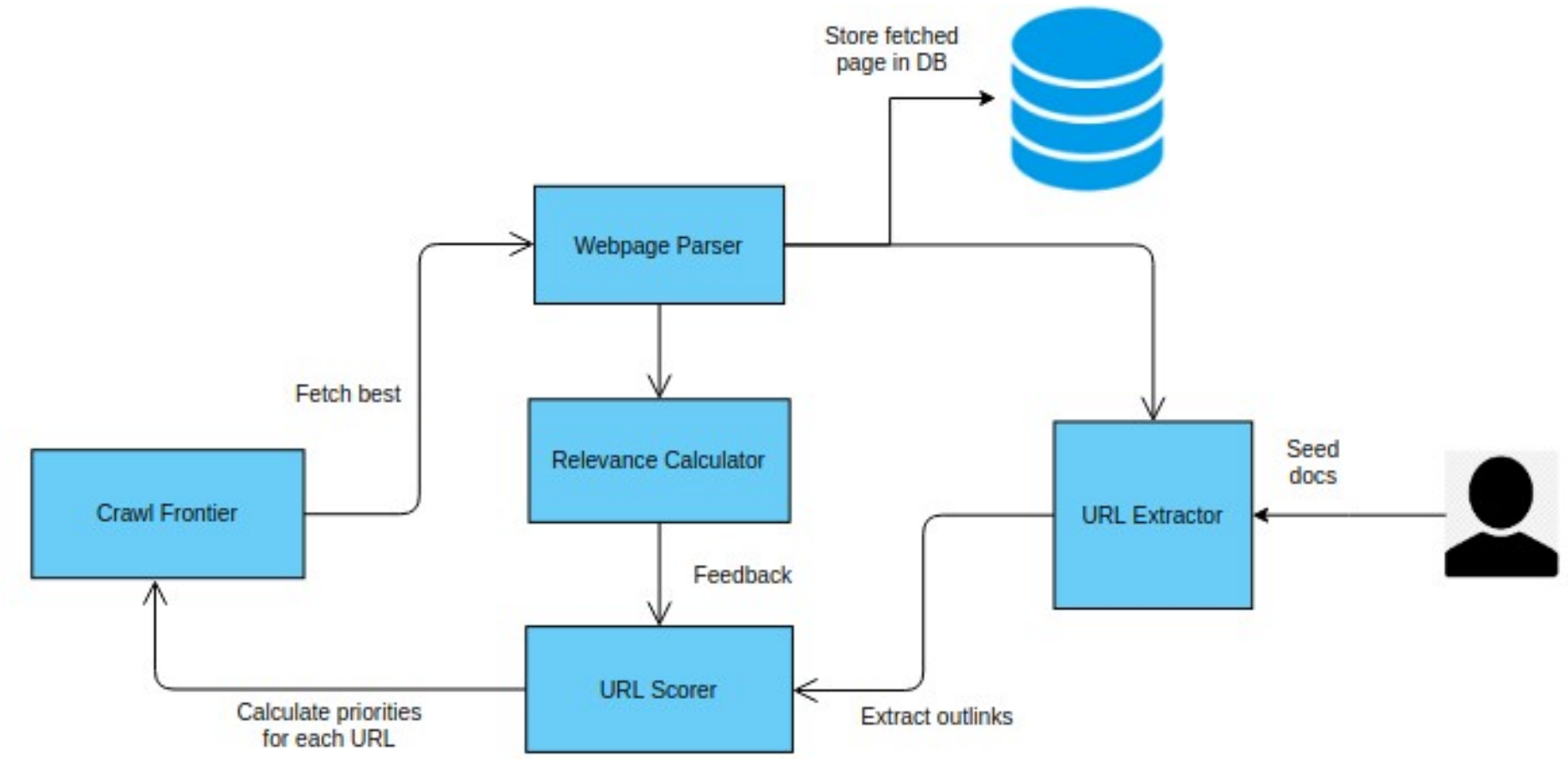
  - output as many relevant url references as possible

Our Intuitive Approach

Researcher – Apprentice Paradigm



AI as an apprentice: Focused Crawler



Our Two-Stage Focused Crawler

- **Researcher teaching apprentice**:

  - The AI learns what is interesting and relevant

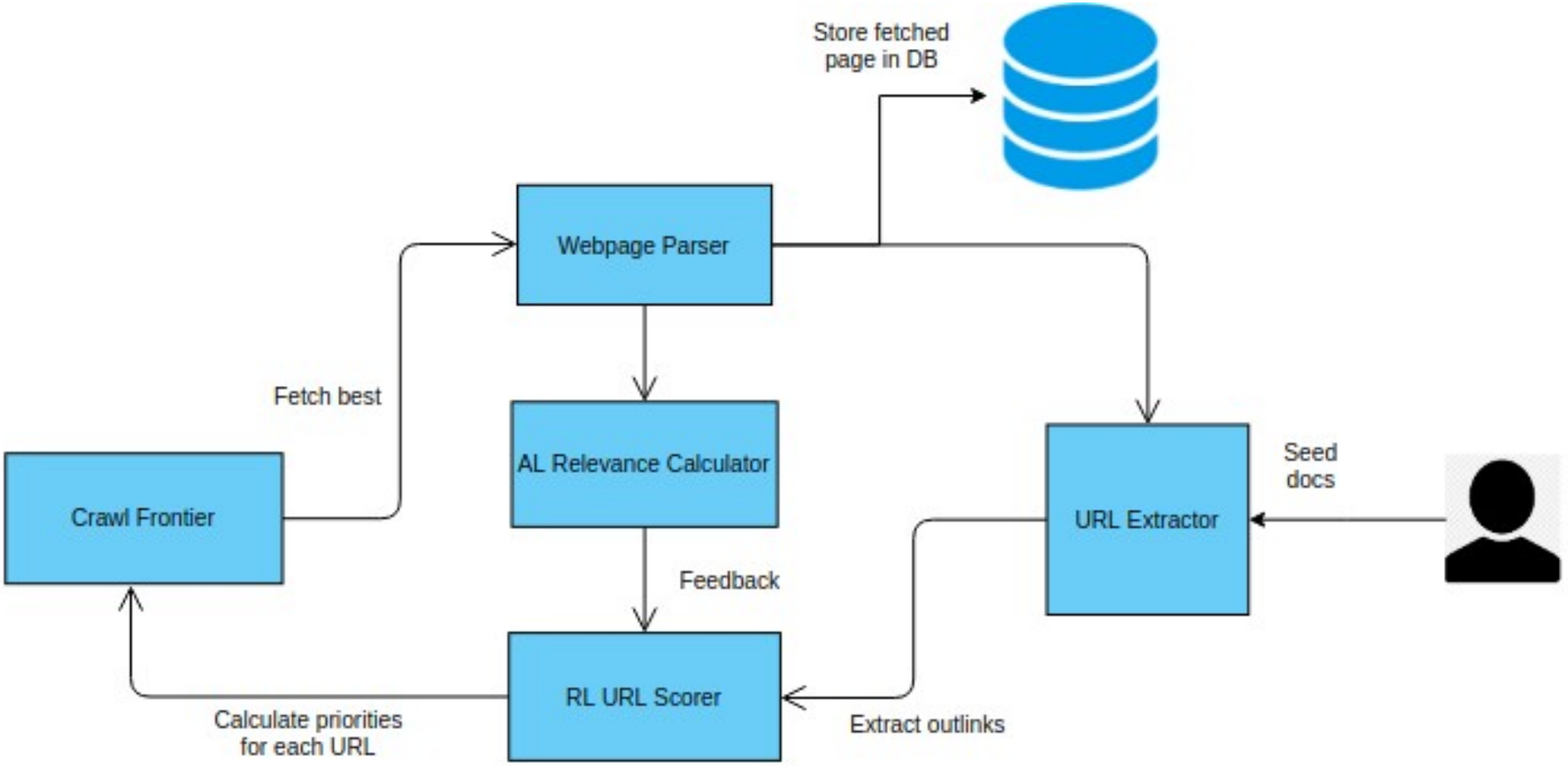
  - supervised learning → *active learning*

- **Apprentice learning how to search best**:

  - The AI learns how to be more efficient when searching for relevant, interesting items

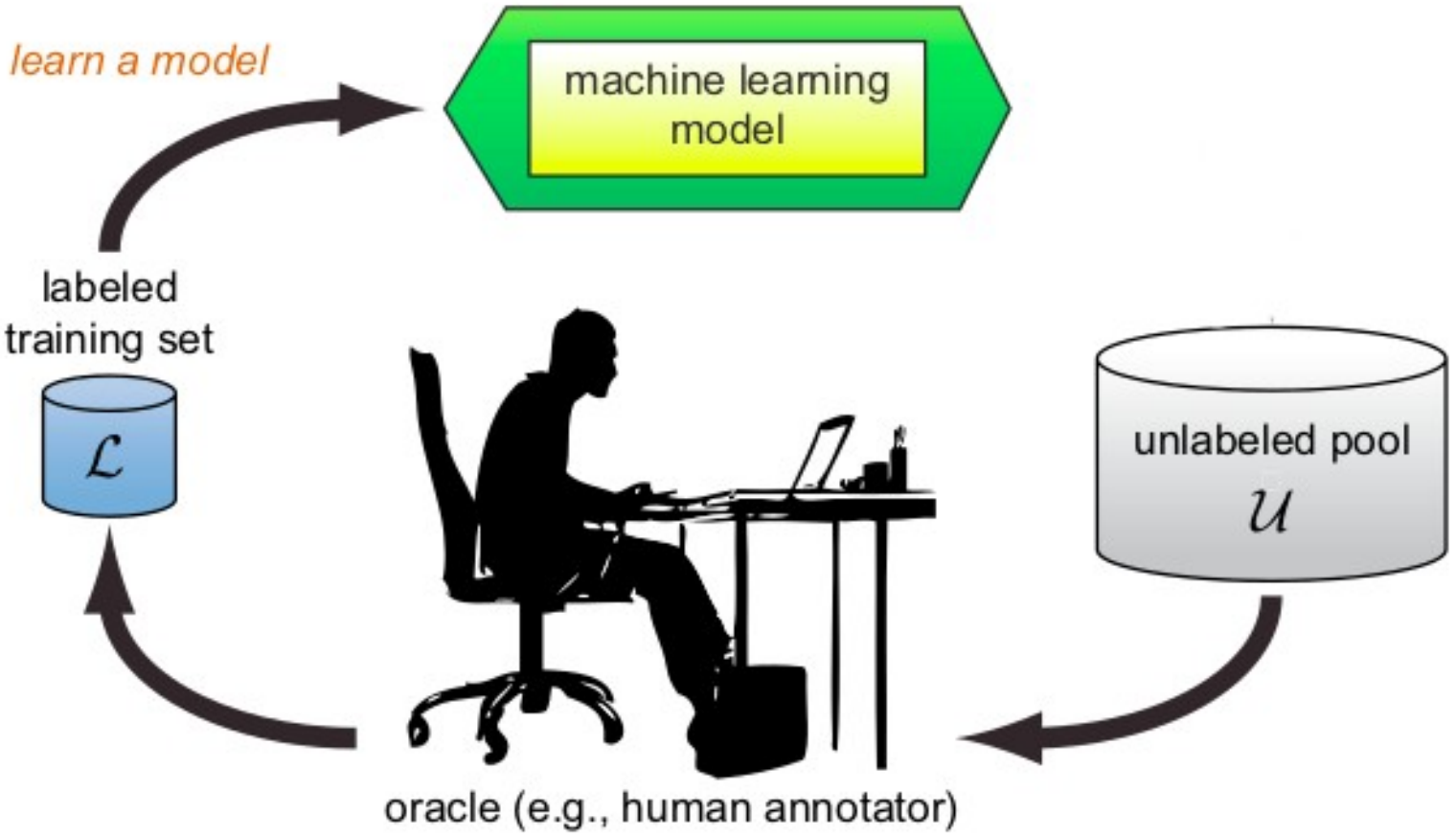
  - *reinforcement learning*

Components


Store fetched page in DB
Webpage Parser
Fetch best
AL Relevance Calculator
Crawl Frontier
URL Extractor
Seed docs
Feedback
RL URL Scorer
Calculate priorities for each URL
Extract outlinks

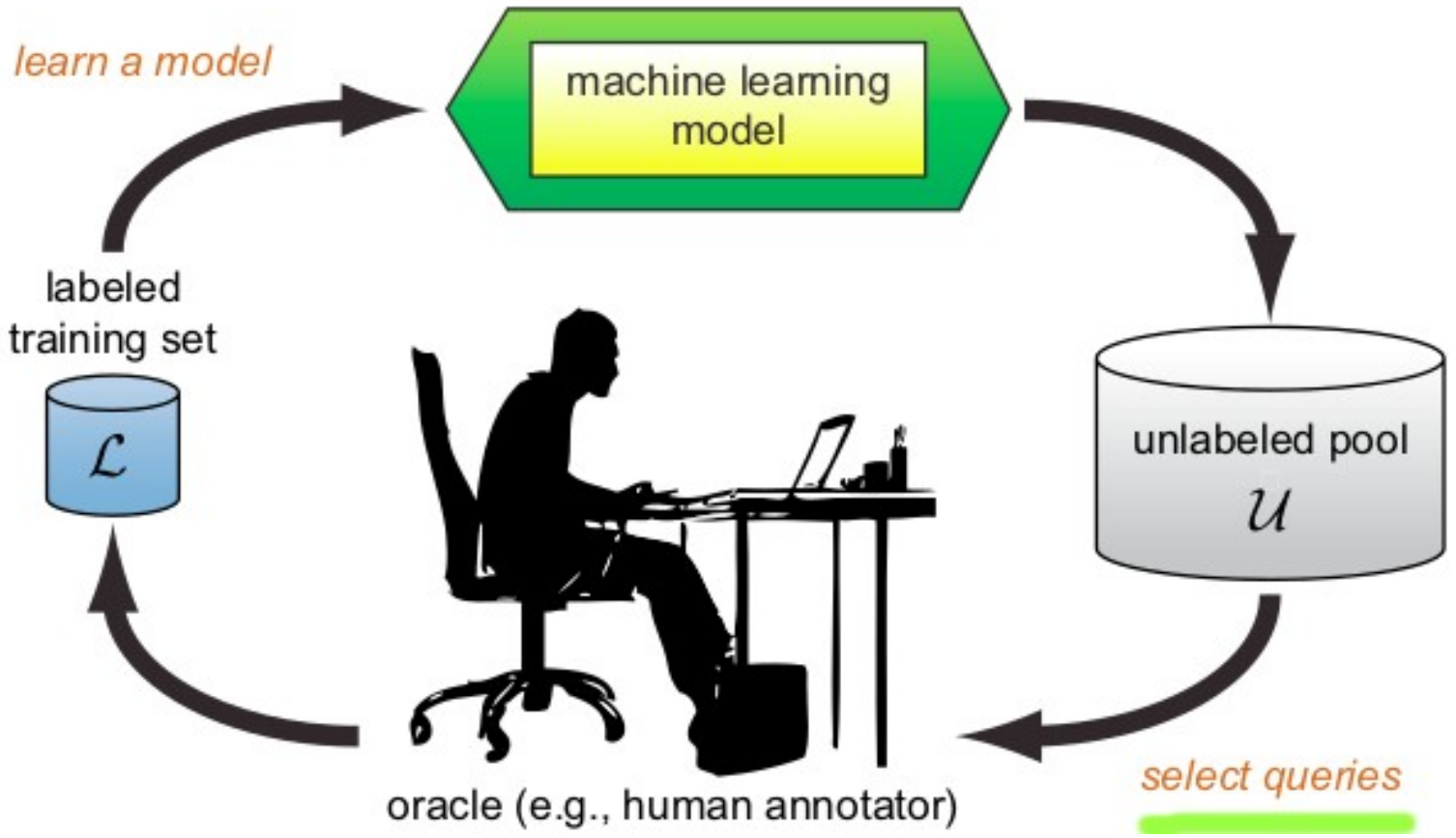
Researcher teaching apprentice

Supervised Learning



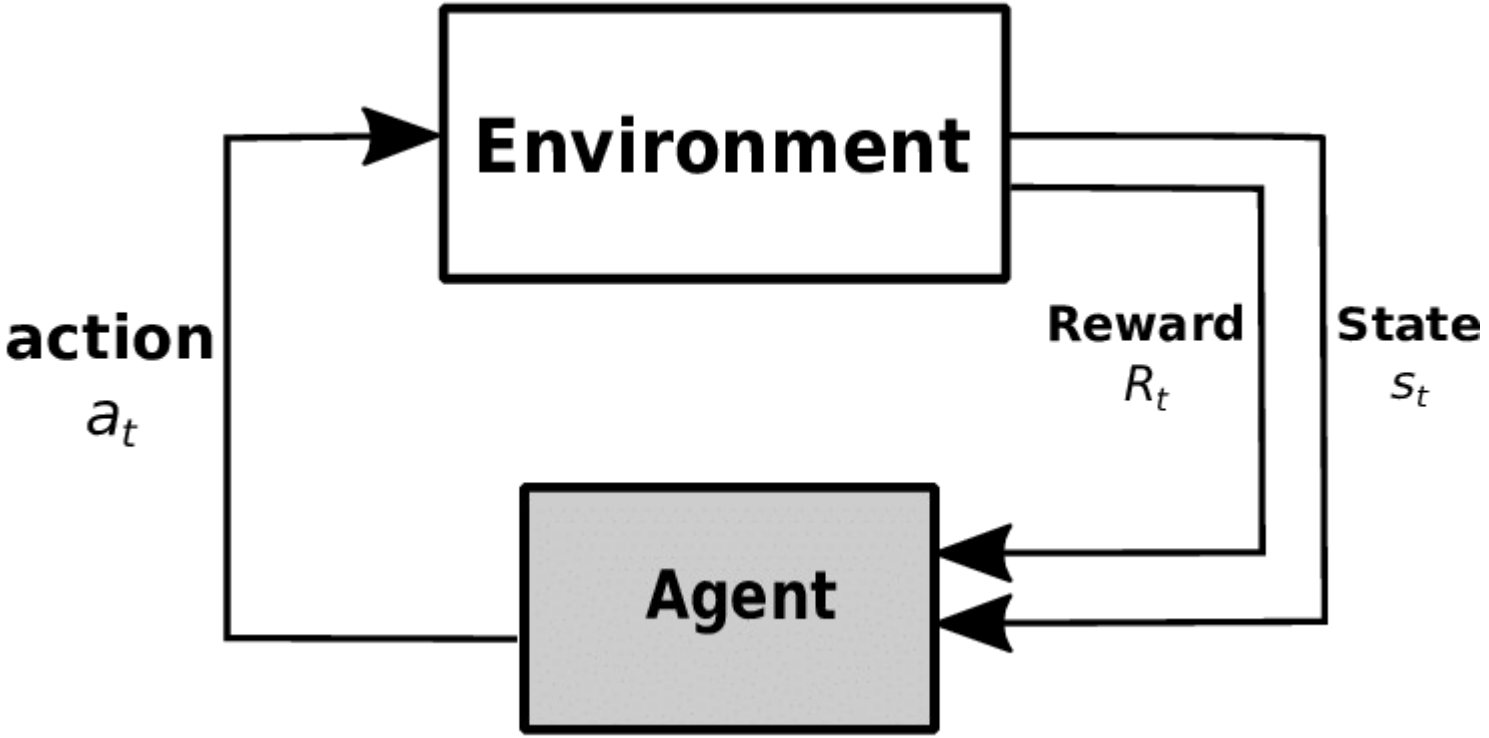
Researcher teaching apprentice

Pool-based Active Learning (AL)



Apprentice learning how to search best

Reinforcement Learning (RL)



- Agent goal:

To maximize the accumulated reward

Our Crawler with RL

- State: current fetched page (abstract)
- Actions: outlinks of fetched page (titles)
- Reward:

  - Relevant doc: 1

  - Irrelevant doc: 0

Text Representation Layer

- Crawling MEDLINE database through PubMed
- word2vec pretrained embeddings [1]
- Mean-Max pooling of embeddings for each doc (abstract / title)

[1] https://bio.nlplab.org/#word-vector-tools

Python Libraries / Modules

- **Parsing:** re, pubmed_parser, requests
- **Text Processing:** nltk
- **AI:** tensorflow, keras, gym, sklearn, numpy, math
- **Visualization:** matplotlib
- **Memory:** pickle, joblib, pandas

Experimental Setup: Pool Creation

- **Random focused crawling:**
  - Starting from 25 relevant seed urls
  - Crawl 1000 urls selecting each time at random
  - Crawl 200 urls, that is the outlinks of the seed urls

- **Create a pool of urls:**
  - 1200 unlabeled urls

Experimental Setup

- **Researcher's Defined Topics:**
  - Highly Relevant: Ethnopharmacology related to specific plant families in Greece, Anatolia or Balkan countries
  - Relevant: Ethnopharmacology in Greece, Anatolia or Balkan countries
  - Partially Relevant: Ethnopharmacology in general
  - Not Relevant: Otherwise

Experimental Setup: Active Learning

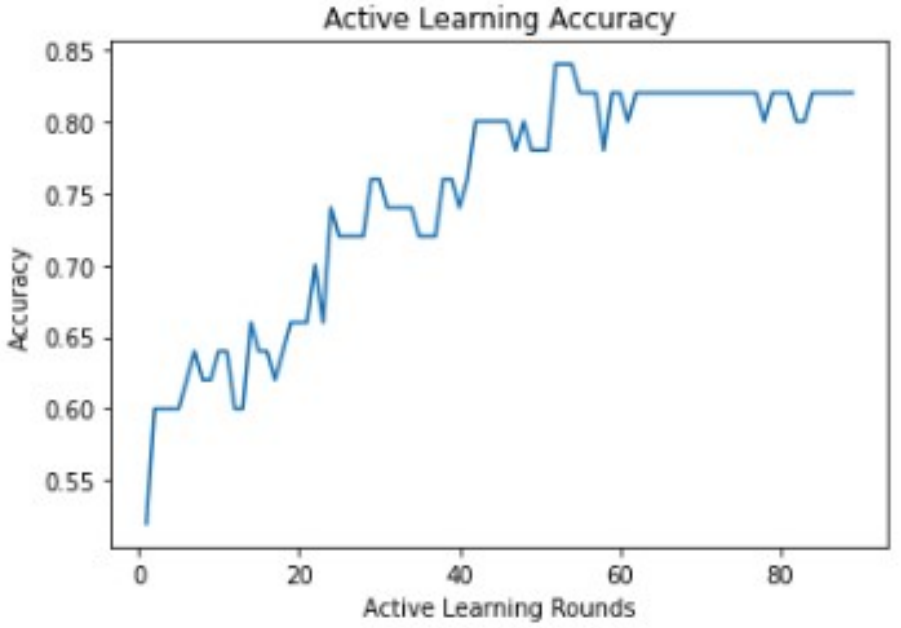
- **Classification problem of *four* discrete labels**
- **Pool is highly imbalanced (too many irrelevant)**
- **Initial Dataset**
  - 50 labeled urls for training set
  - 1200 unlabeled urls for Pool

- **Evaluation of performance on held-out test set**
  - 50 labeled urls

Evaluation Metrics:

- *Accuracy*; percentage of correct predictions
- *Recall* (relevant); percentage of interesting docs fetched

Does the AI learn ?



Best Model

	accuracy
all	0.84
	recall
relevant	0.67

- **Yes we do!**

Teaser!!!
We can do better

- Previous scores

	accuracy
all	0.84
	recall
relevant	0.67

- Filtering with keywords 

	accuracy
all	0.9
	recall
relevant	0.92

Only Keywords – No AI

- **Relevant docs only those with keyword(s)**
- **Is recall better ?**

	Recall
Relevant	0.75

- **0.75 < 0.92**
- 

Summary

- An **open science community**
- Collaborating on an **open source system**
- Supporting their research **through AI**
- A *researcher–apprentice* view on AI
- AL to train the apprentice
- RL to help him search better – ongoing work

Future Steps

- Design the overall system
- Integrate AI
- Optimize AI
- Run pilot experiments
- Further classification of results based on plant species as well as organs of the human body

Fosscomm 2020

Thanks for watching
Questions ?

Fosscomm 2020

Training Oracle with AL

- Random Sampling with SVM
  - At random
- Margin Sampling with SVM
  - Minimum Distance from decision boundary
- Uncertainty Sampling with Deep Learning
  - Minimum certainty probabilities produced by softmax

DQN Algorithms

- Q-Learning
- Deep Learning
- Experience Replay
- Q-Network
- Target Q-Network

DQN (Mnih et al, 2015)

DDQN (van Hasselt et al, 2015)