Norway grants

REFSA

Articles representations for the automation of the inclusion proces in SLR

Radosław Pytlak

Warsaw University of Technology

8th of June 2022

Agenda

- Project REFSA
- IT support for SLR
- Semantic representations of articles
- Active learning proces

Project REFSA

Machine Learning-based system for the automation of systematic literature reviews in food safety domain

Partners:

- **Warsaw University of Technology** (leader)
- **Oslo Metropolitan University**
- **Norwegian Institute of Public Health – Norwegian Scientific Committee for Food and Environment**
- **National Institute of Public Health**
- **Tecna, IT company**

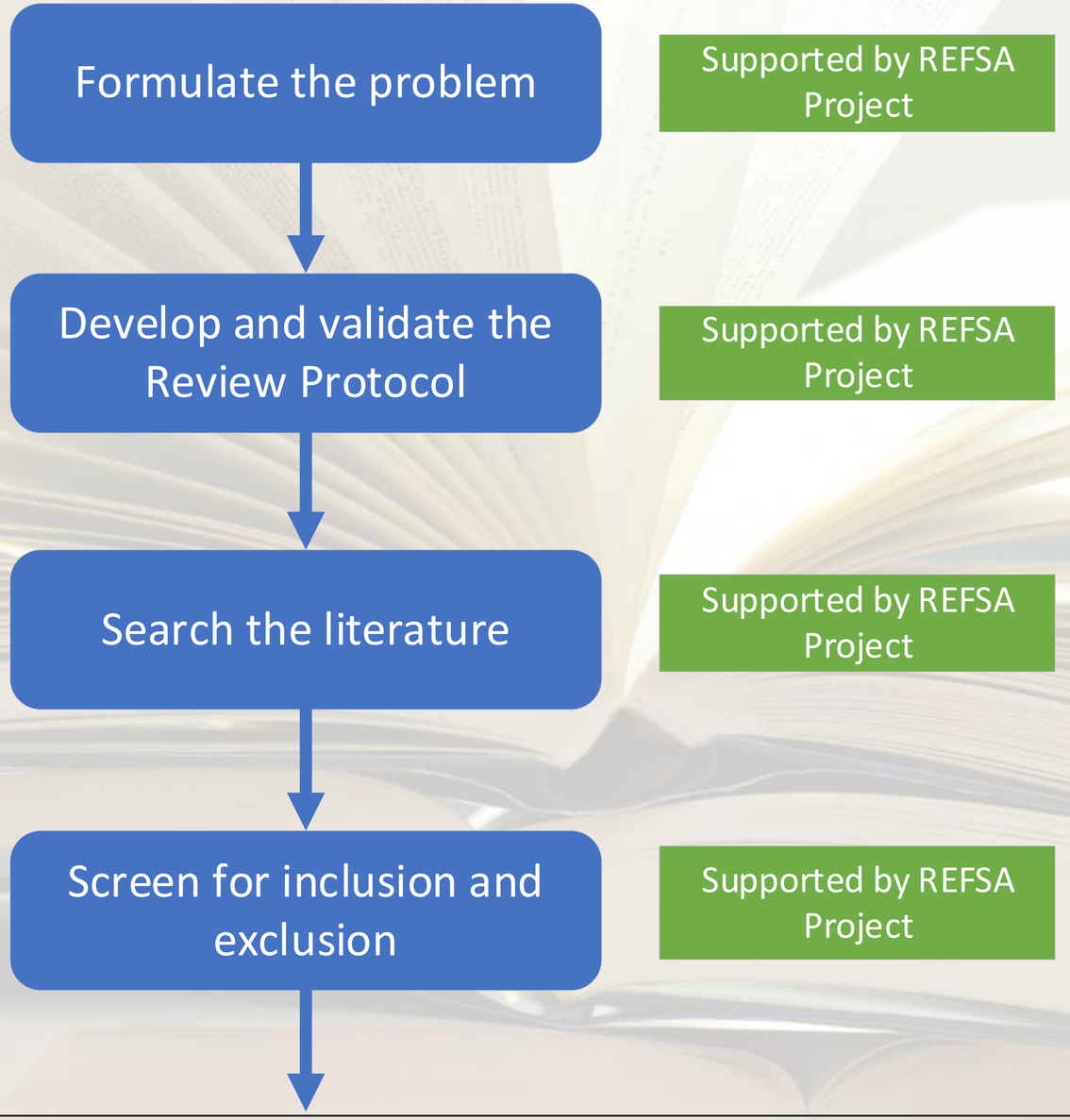
Systematic Literature Review

- PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) is a guideline for systematic reviews and meta-analyses of evaluations of health care interventions – published in 2008.

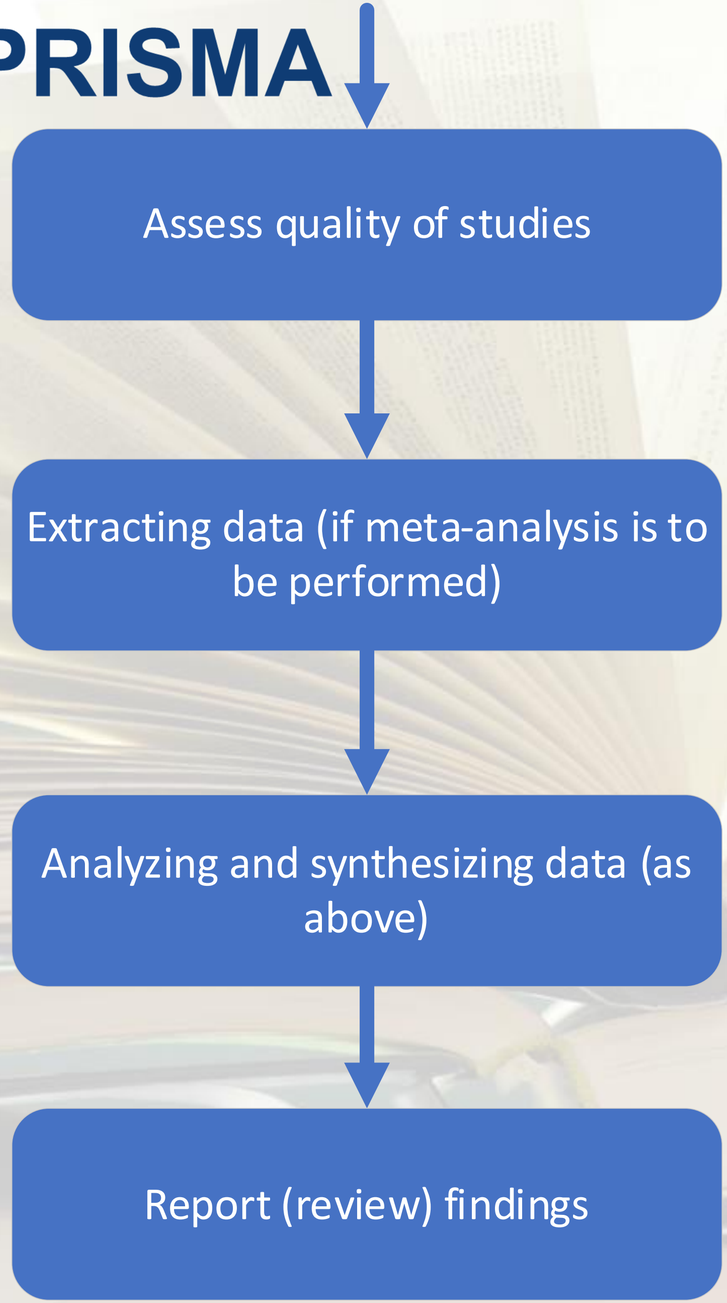
- The PRISMA Statement consists of a 27-item checklist and a four-phase flow diagram.

- The PRISMA Statement was developed by a group of 29 review authors, methodologists, clinicians, medical editors, and consumers.

Less detailed PRISMA

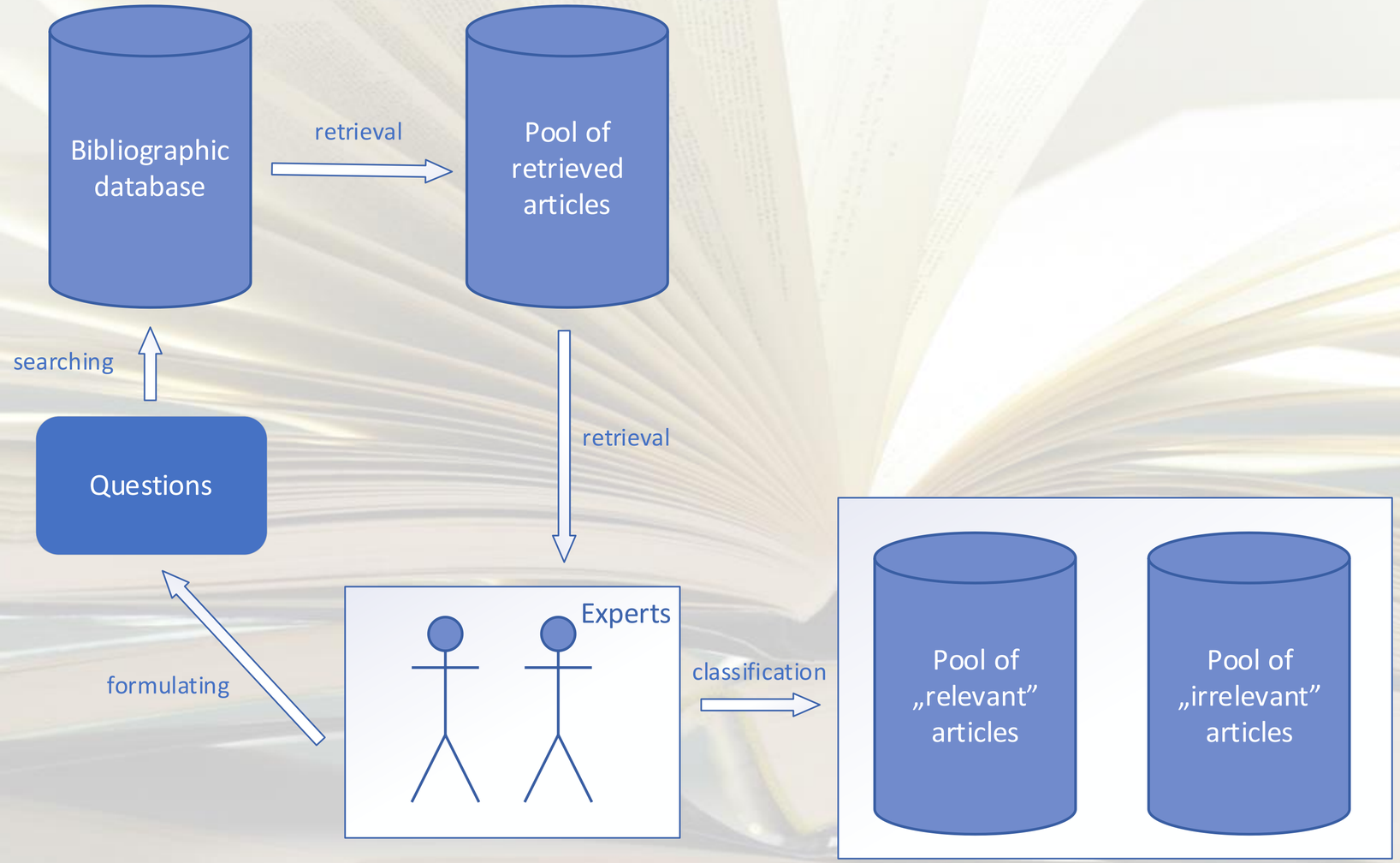


Less detailed PRISMA

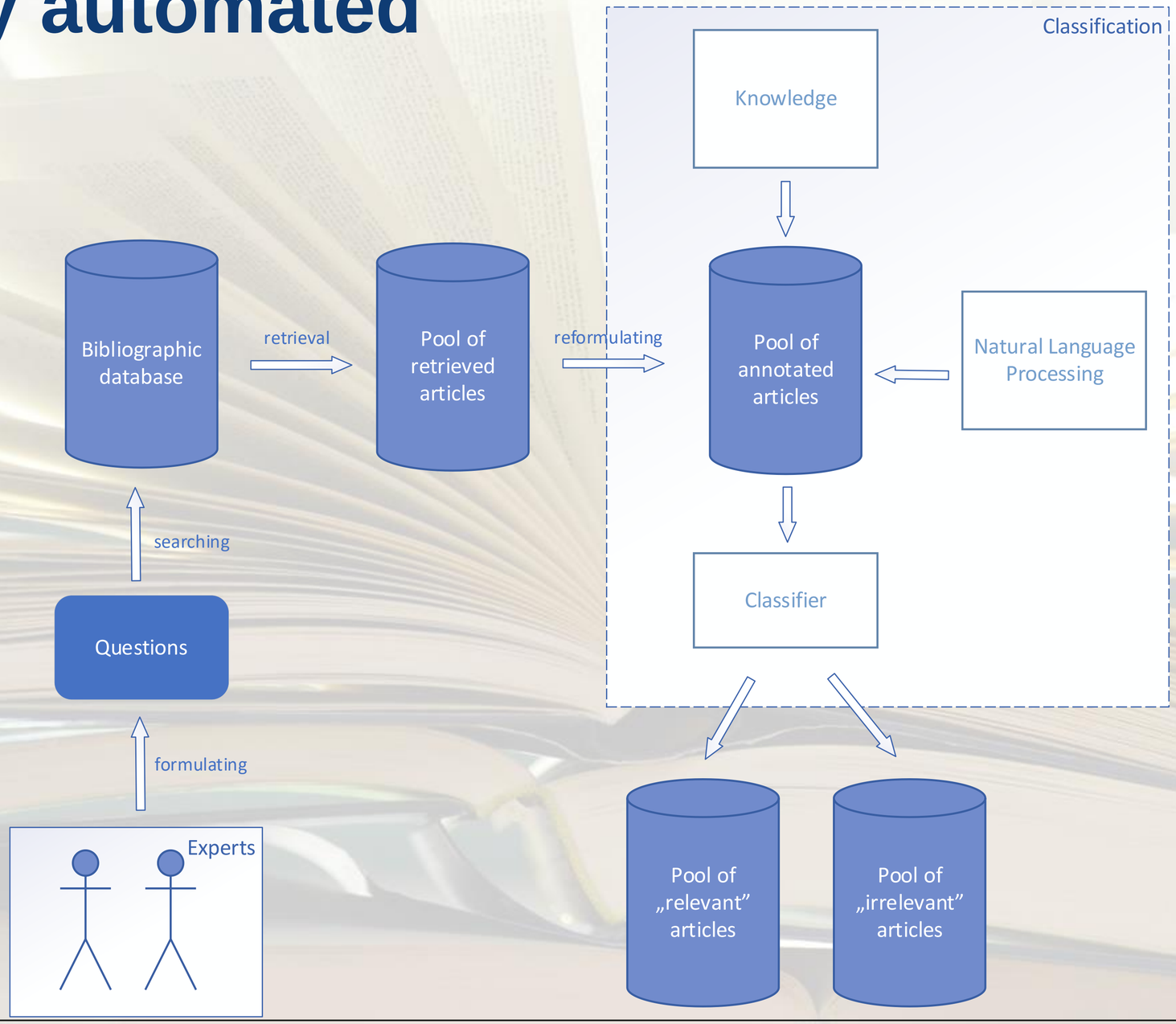


Supported by REFSA Project

SLR, manual screening process

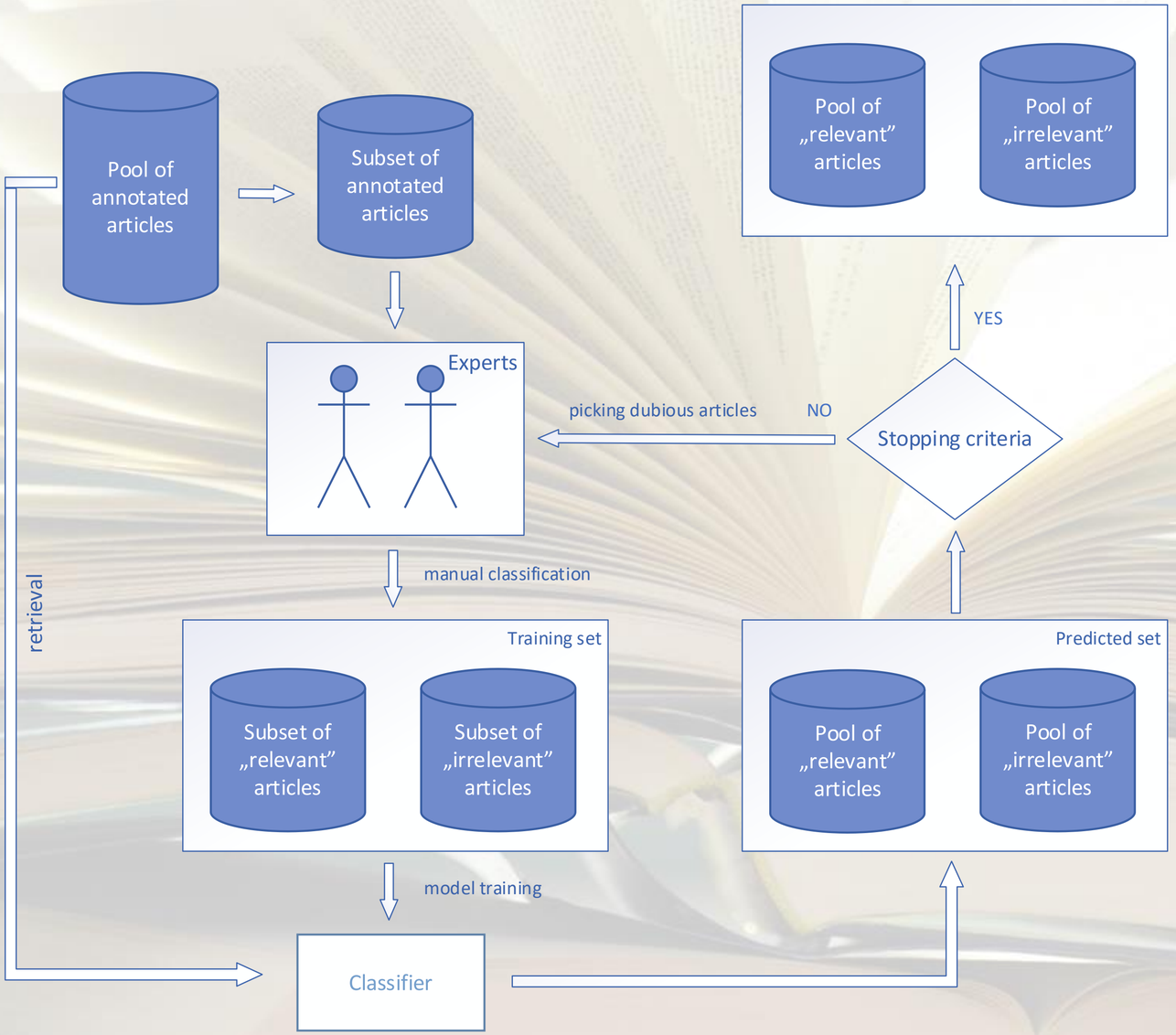


SLR – fully automated process

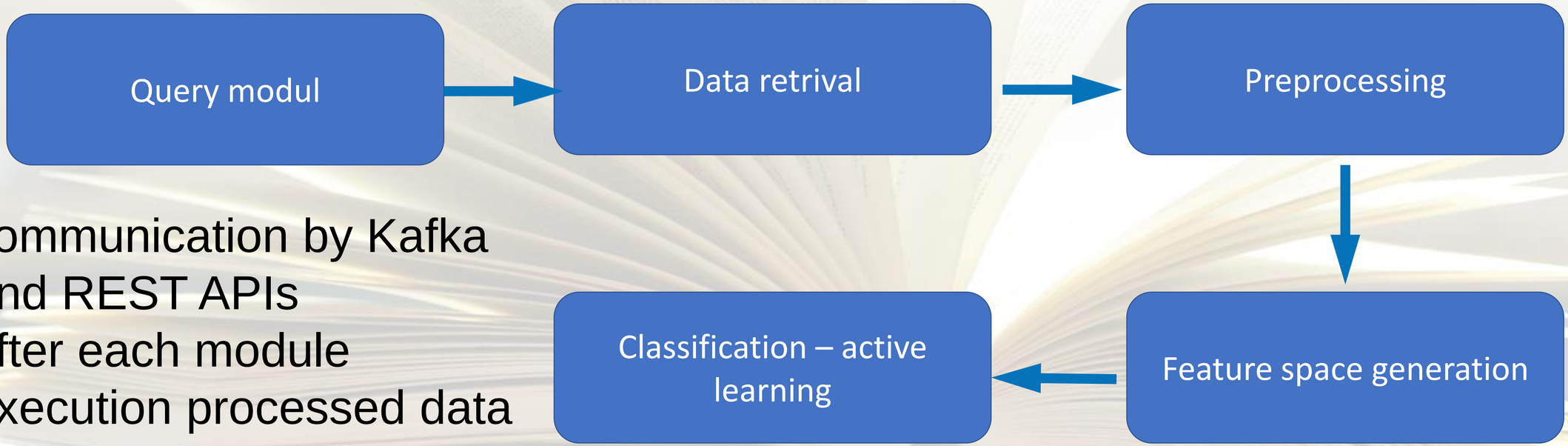


How to build a classifier?

SLR, how to use experts to build classifiers



The pipeline of ML modules

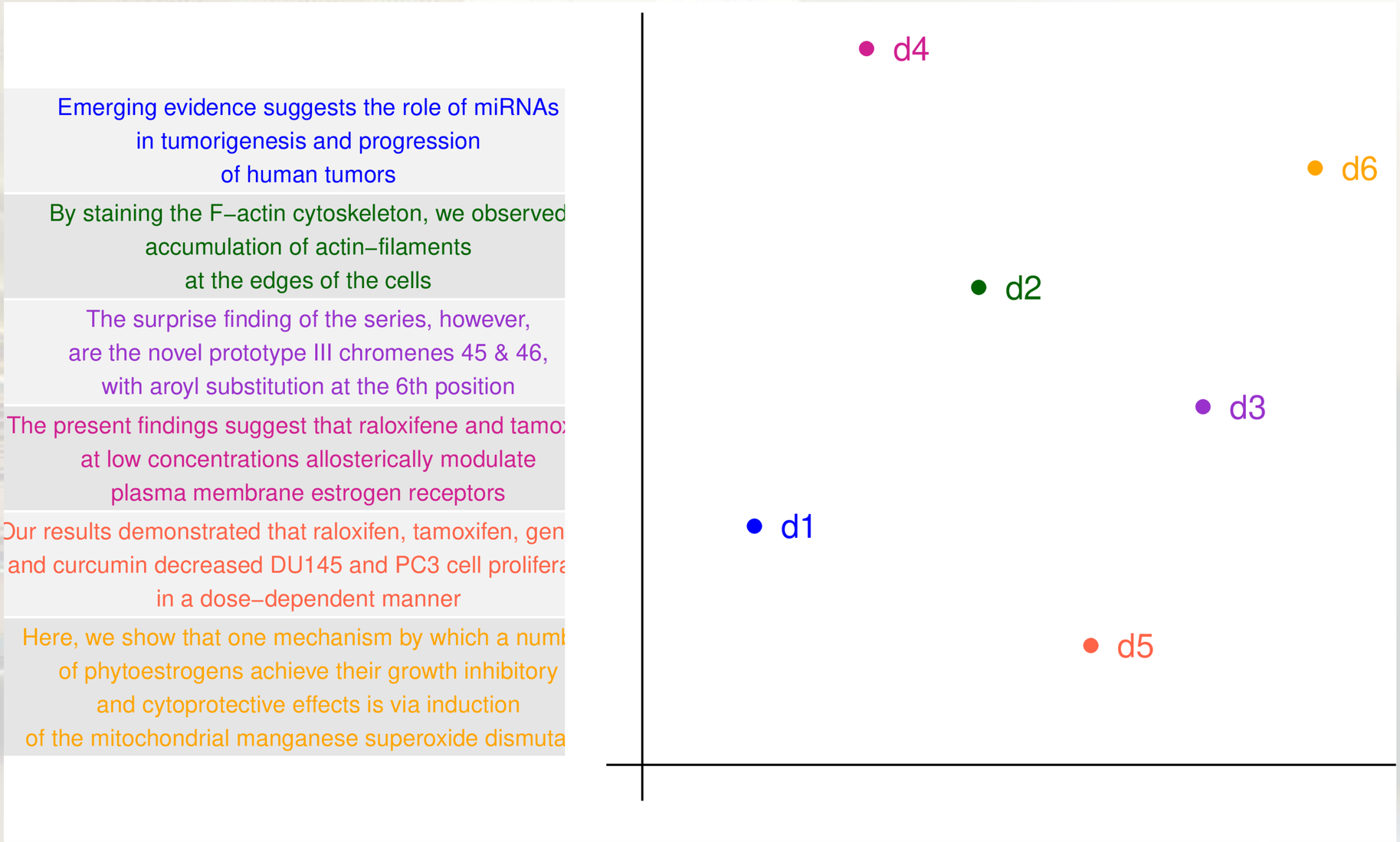


- communication by Kafka and REST APIs
- after each module execution processed data are stored in database
- all processing are carried out using cloud computing technology

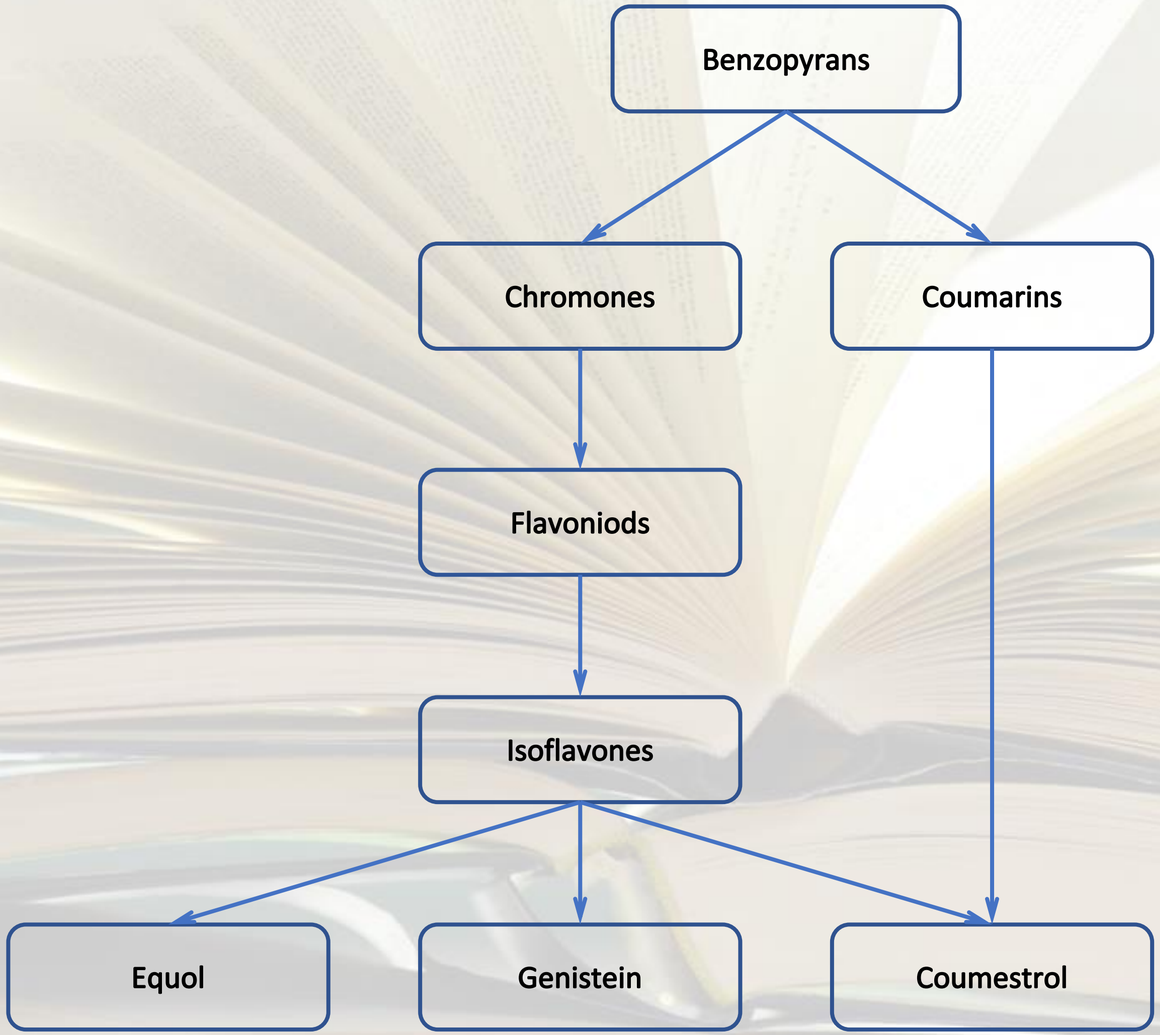
- modules can be substituted by other modules with the same functionatlity
- most modules written in Python, some in Java

Text Representation

- Documents mapped to vectors in a multidimensional numeric space
- Several approaches investigated:
  - bag of words
  - bag of concepts - annotation based on ontology
  - word or text embeddings – semantic representation based on BioBERT



Bag of Concepts – ontology



Bag of Concepts – ontology

Objective. Genistein, a naturally occurring isoflavenoid abundant in soy products, has anti-neoplastic activity in multiple tumor types. There are several mechanisms reported for genistein's anti-neoplastic activity. In the present Study, we studied the mechanism of genistein-induced cell death in ovarian cancer cells. Methods. The effect of genistein on the induction of apoptosis, autophagy, and inhibition of glucose uptake in ovarian cancer cells was determined. The effect of genistein on the expression of phosphorylated Akt was determined by immunoblotting. Results. Genistein is cytotoxic to ovarian cancer cells. The mechanism of genistein-induced cell death includes both apoptosis and autophagy. Because autophagy is typically an adaptive response to nutrient starvation, we hypothesized that genistein could induce a starvation-like signaling response. We show here that genistein treatment results in caspase-independent cell death with hallmarks of autophagy. Genistein treatment dramatically inhibits glucose uptake in ovarian cancer cells, and methyl pyruvate, a cell-permeable 3-carbon substrate for oxidative phosphorylation and fatty acid synthesis, rescues cells from genistein-induced autophagy.

Concept ID	mesh:D019833			
Synonyms	genistein			
Count	11			

Bag of Concepts – ontology

Objective. Genistein, a naturally occurring isoflavenoid abundant in soy products, has anti-neoplastic activity in multiple tumor types. There are several mechanisms reported for genistein's anti-neoplastic activity. In the present Study, we studied the mechanism of genistein-induced cell death in ovarian cancer cells. Methods. The effect of genistein on the induction of apoptosis, autophagy, and inhibition of glucose uptake in ovarian cancer cells was determined. The effect of genistein on the expression of phosphorylated Akt was determined by immunoblotting. Results. Genistein is cytotoxic to ovarian cancer cells. The mechanism of genistein-induced cell death includes both apoptosis and autophagy. Because autophagy is typically an adaptive response to nutrient starvation, we hypothesized that genistein could induce a starvation-like signaling response. We show here that genistein treatment results in caspase-independent cell death with hallmarks of autophagy. Genistein treatment dramatically inhibits glucose uptake in ovarian cancer cells, and methyl pyruvate, a cell-permeable 3-carbon substrate for oxidative phosphorylation and fatty acid synthesis, rescues cells from genistein-induced autophagy.

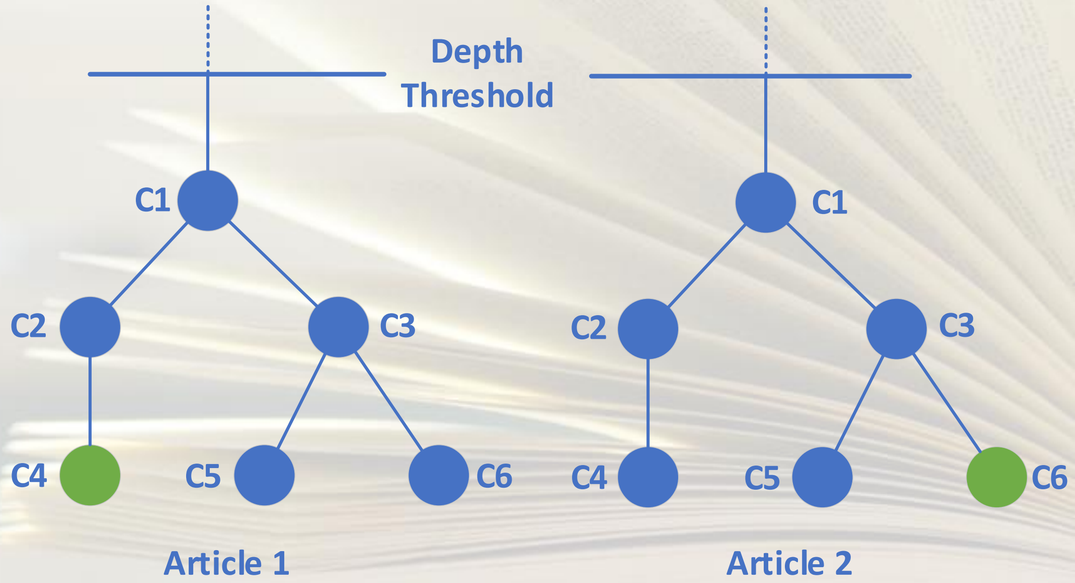
Concept ID	mesh:D019833	mesh:D009369		
Synonyms	genistein	neoplastic tumor		
Count	11	3		

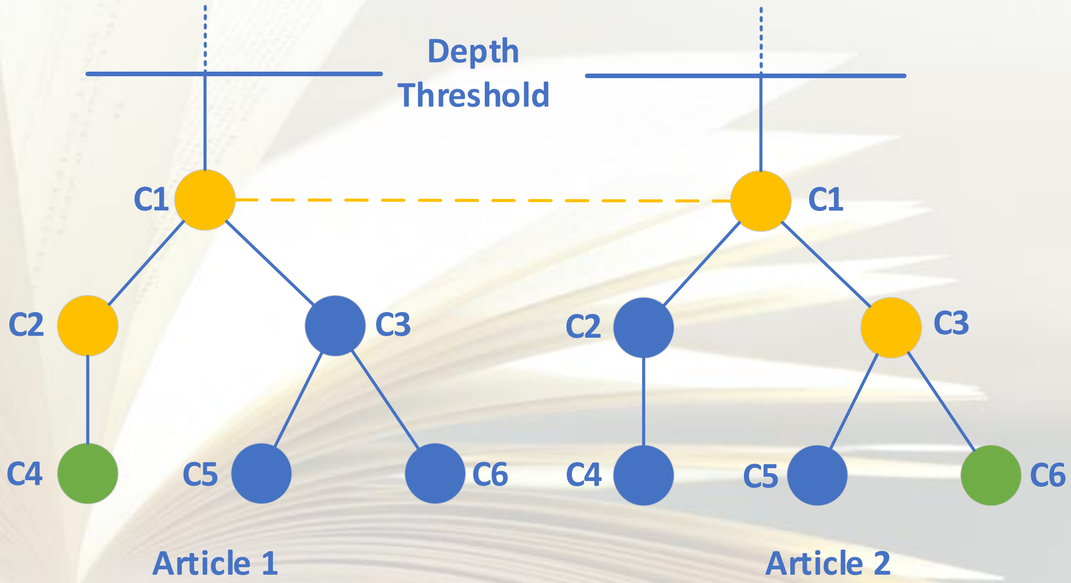
Bag of Concepts – ontology

Objective. Genistein, a naturally occurring isoflavenoid abundant in soy products, has anti-neoplastic activity in multiple tumor types. There are several mechanisms reported for genistein's anti-neoplastic activity. In the present Study, we studied the mechanism of genistein-induced cell death in ovarian cancer cells. Methods. The effect of genistein on the induction of apoptosis, autophagy, and inhibition of glucose uptake in ovarian cancer cells was determined. The effect of genistein on the expression of phosphorylated Akt was determined by immunoblotting. Results. Genistein is cytotoxic to ovarian cancer cells. The mechanism of genistein-induced cell death includes both apoptosis and autophagy. Because autophagy is typically an adaptive response to nutrient starvation, we hypothesized that genistein could induce a starvation-like signaling response. We show here that genistein treatment results in caspase-independent cell death with hallmarks of autophagy. Genistein treatment dramatically inhibits glucose uptake in ovarian cancer cells, and methyl pyruvate, a cell-permeable 3-carbon substrate for oxidative phosphorylation and fatty acid synthesis, rescues cells from genistein-induced autophagy.

Concept ID	mesh:D019833	mesh:D009369	mesh:D010051	…
Synonyms	genistein	neoplastic tumor	ovarian cancer	…
Count	11	3	4	…

Extended Bag of Concepts created with the help of ontology tree

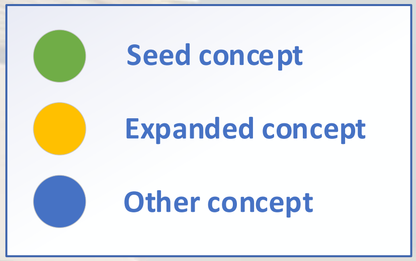




After expansion

11	0	…	3	0	4	0	0	0

11	0.34	…	3	0.733	4	0.45	0	0



Semantic text representations

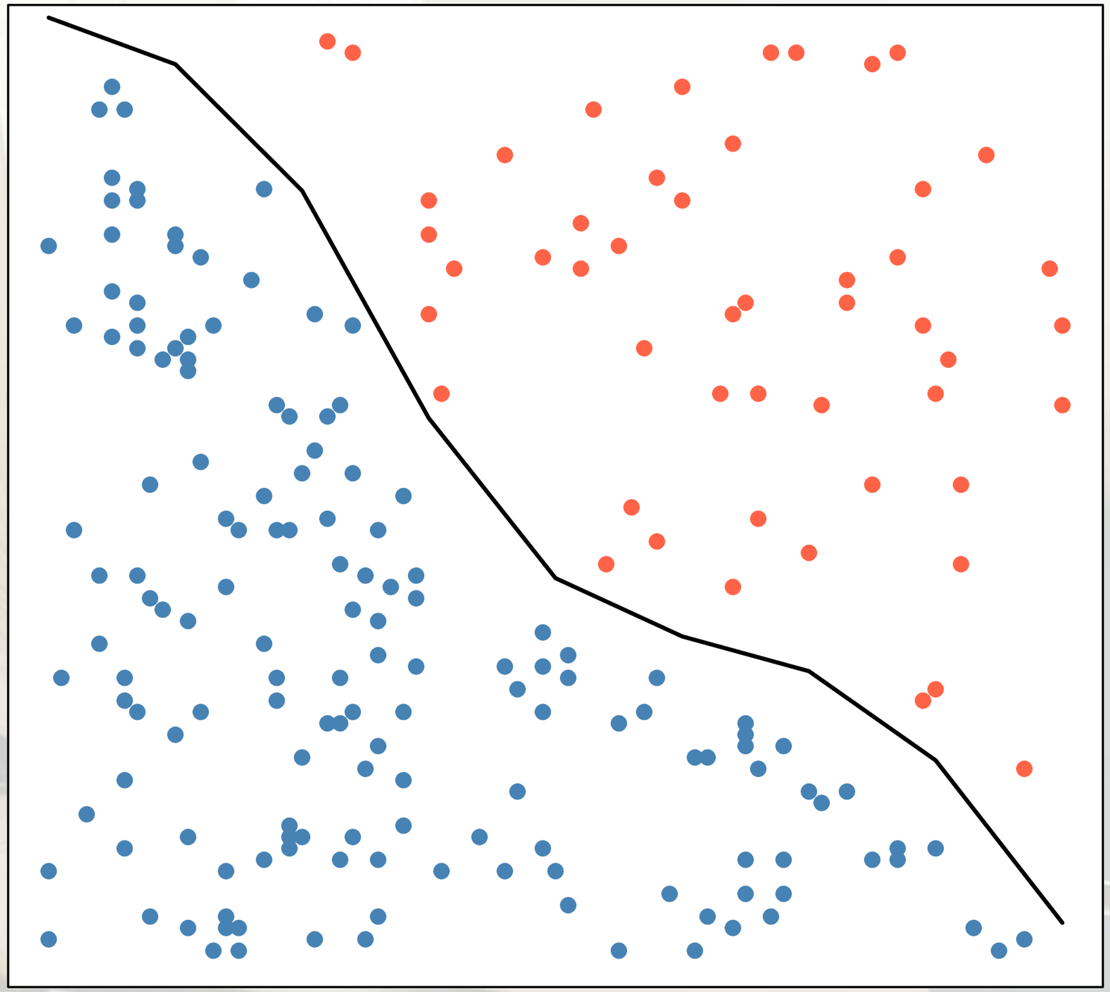
Experiments with annotators based on recently developed NER models based on Deep Learning - Transformers and Long-Short Term Memory Network:

- considerable improvements in accuracy
- require explicit supervised training
- experiments with BioBERT reveal high computing time – currently experiments using GPU (8x NVIDIA A100 40 GB) to transform abstracts to numerical vectors

Recently the beginning of experiments with unsupervised approach to NER - it uses unsupervised machine learning to disambiguate entities.

Classification Models

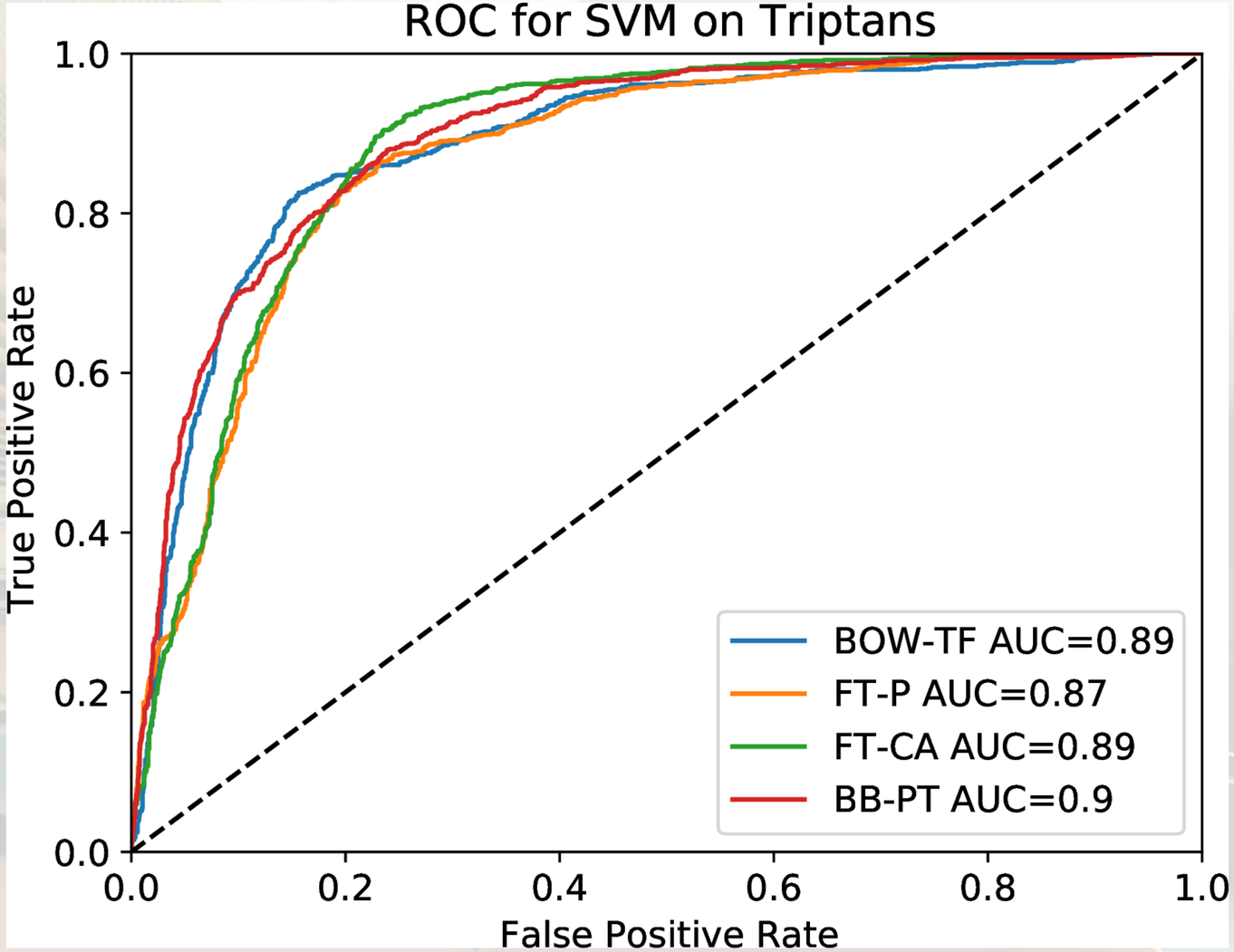
- Trained on documents labeled as relevant or relevant by a human expert
- Representing a decision boundary based on training data generalization
- Providing predictions of:
  - class labels
  - class probabilities
  - decision values (scores)
- Several algorithms investigated:
  - naïve Bayes, logistic regression, SVM, random forest,



Classification Quality

ROC curve – tradeoff between:

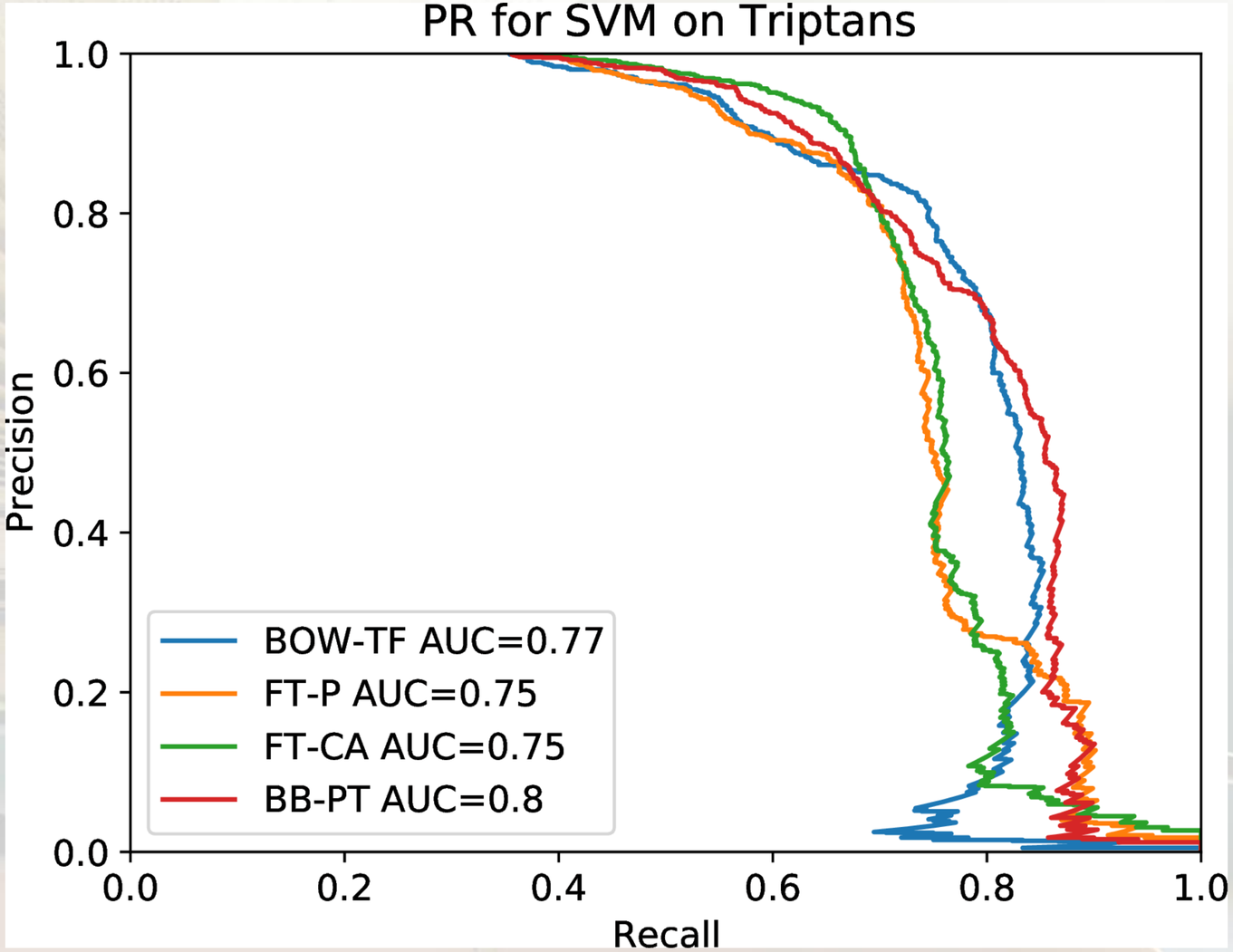
- **true positive rate** – the share of relevant documents correctly classified by the model
- **false positive rate** – the share of irrelevant documents incorrectly classified by the model



Classification Quality

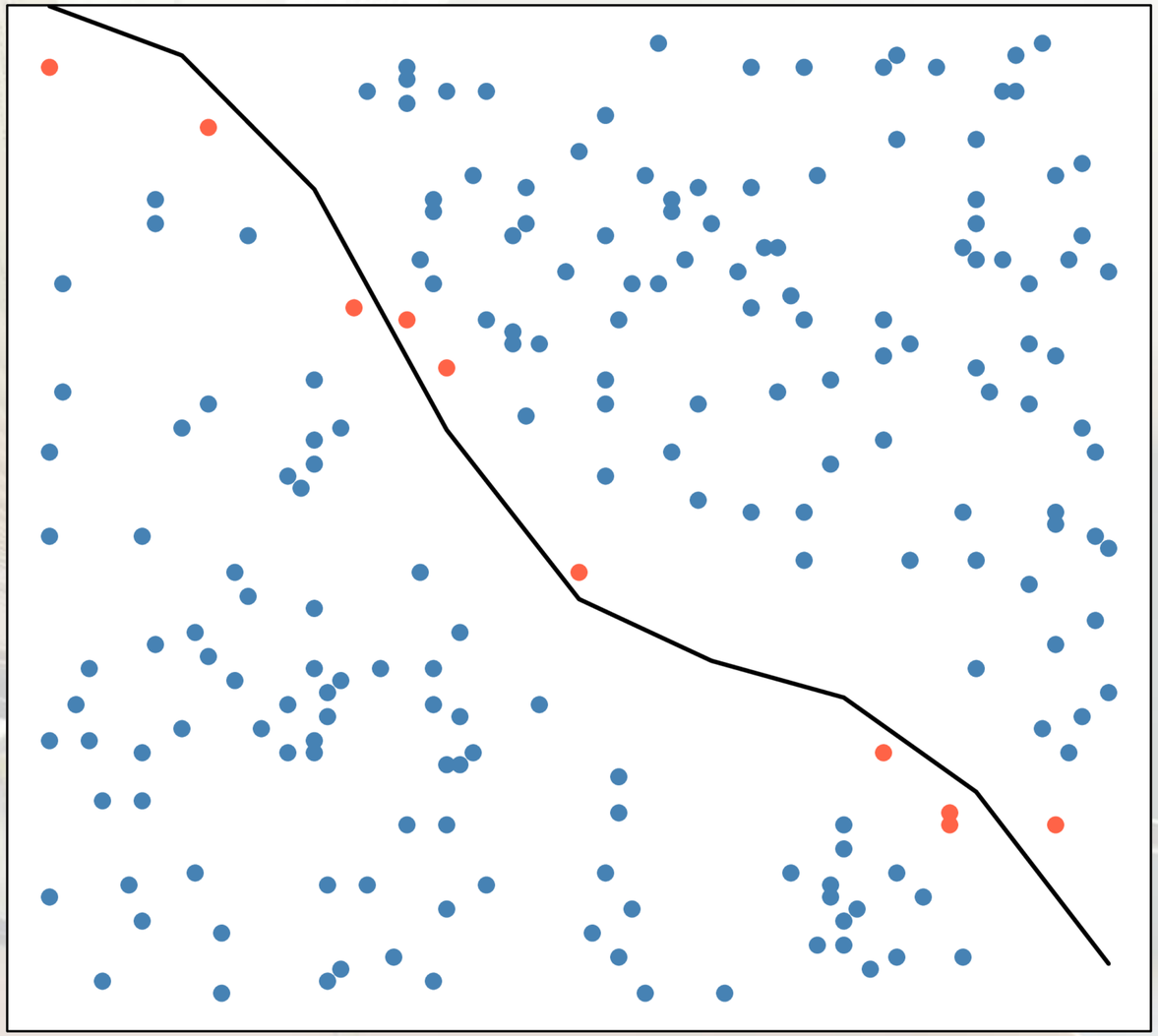
Precision-recall curve – tradeoff between:

- **precision** – the share of documents classified by the model as relevant that are truly relevant
- **recall** – the share of relevant documents correctly classified by the model

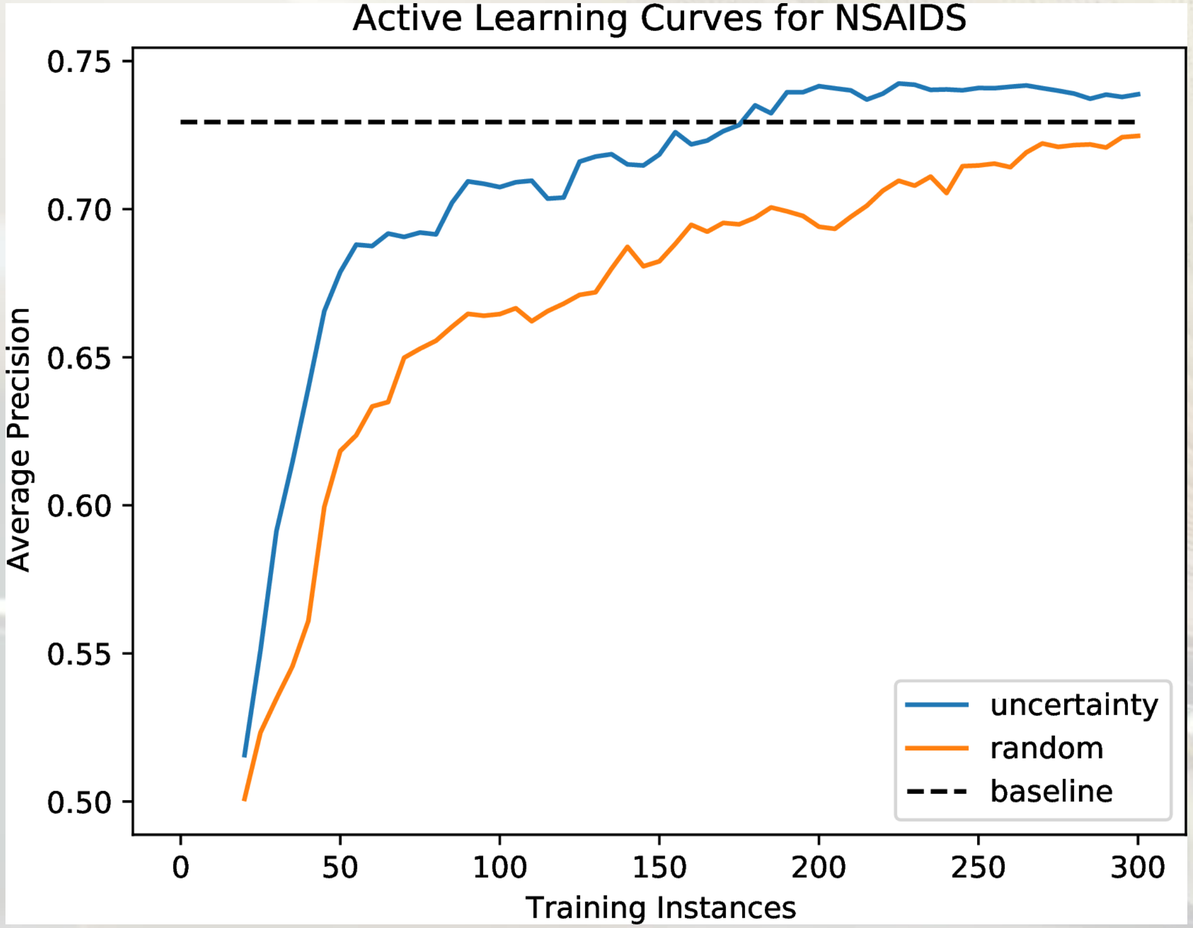


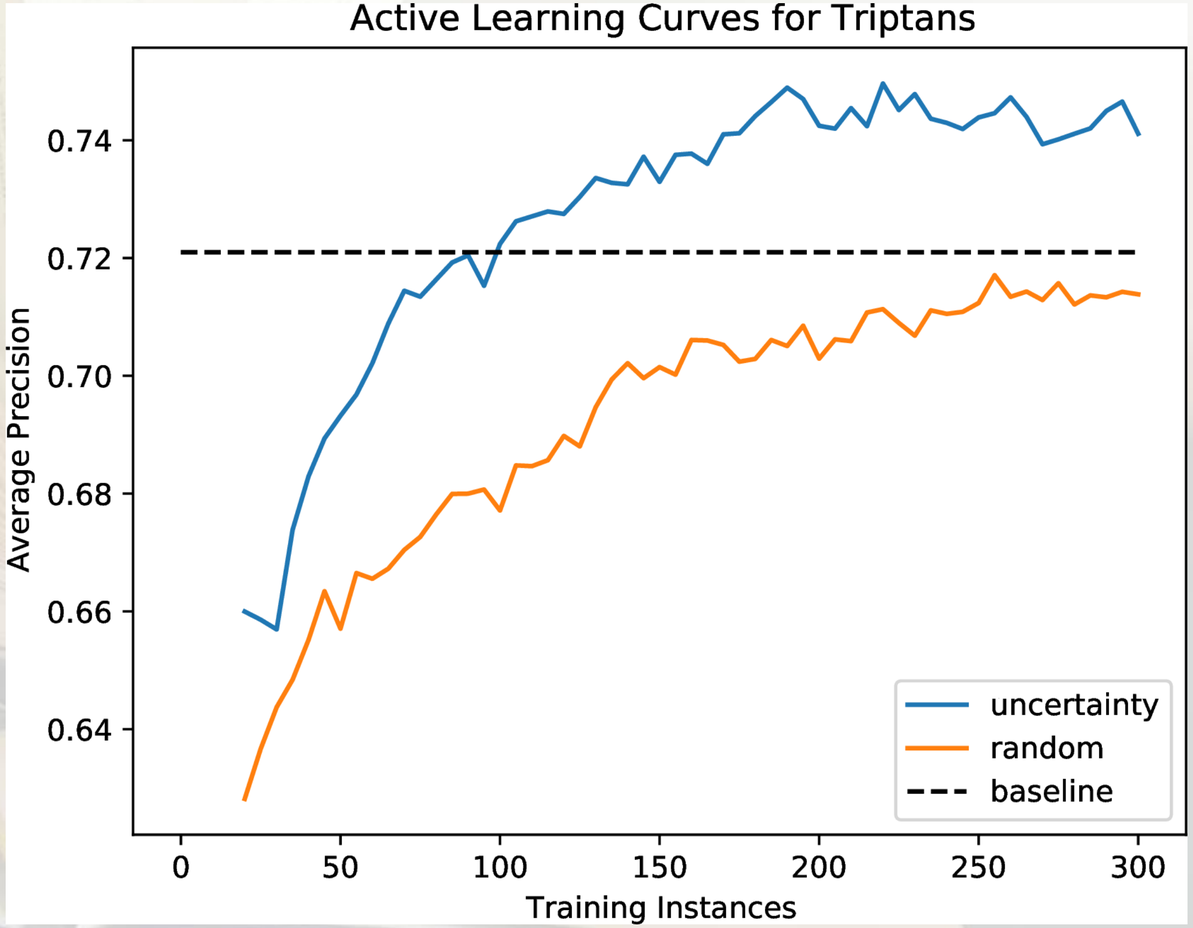
Active Learning

- Reduce human effort needed to provide labeled training data
- **Initialization:** start from a small initial training set and create the first model
- **Querying:** repeatedly request labeling a small number of documents about which the current model is the most unconfident, add them to the training data, and create the next model



Active Learning





Further Work

- **Research:**
  - improving active learning procedure: class imbalance compensation methods; active learning stop criteria
  - improving text representation methods
  - combining models using different text representations
- **Implementation:**
  - improved modules integration
  - testing the REFSA application with experts in food safety domain

Norway grants

REFSA

Thank you!