COLLABORATIVE CLASSIFIER AGENTS

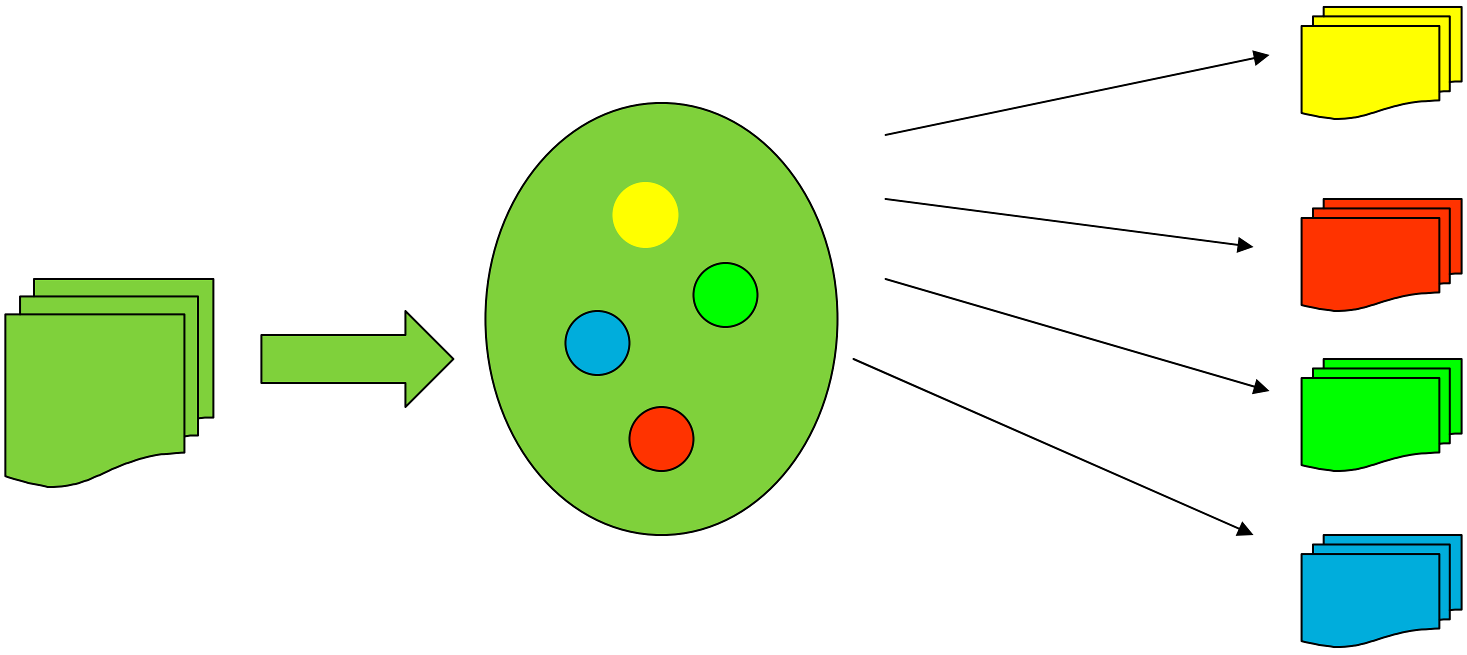
Studying the Impact of Learning in Distributed Document Classification

Weimao Ke, Javed Mostafa, and Yueyu Fu
{wke, jm, yufu}@indiana.edu
Laboratory of Applied Informatics Research
Indiana University, Bloomington

RELATED AREAS

- Text Classification (i.e. categorization)
  - Information Retrieval
  - Digital library
  - Indexing, cataloging, filtering, etc.
  - → Distributed Text Classification
- Multi-Agent Modeling
  - Machine Learning:
    - Learning Algorithms
    - Multi-Agent Modeling
  - → Modeling Agent Collaboration

TRADITIONAL CENTRALIZED CLASSIFICATION

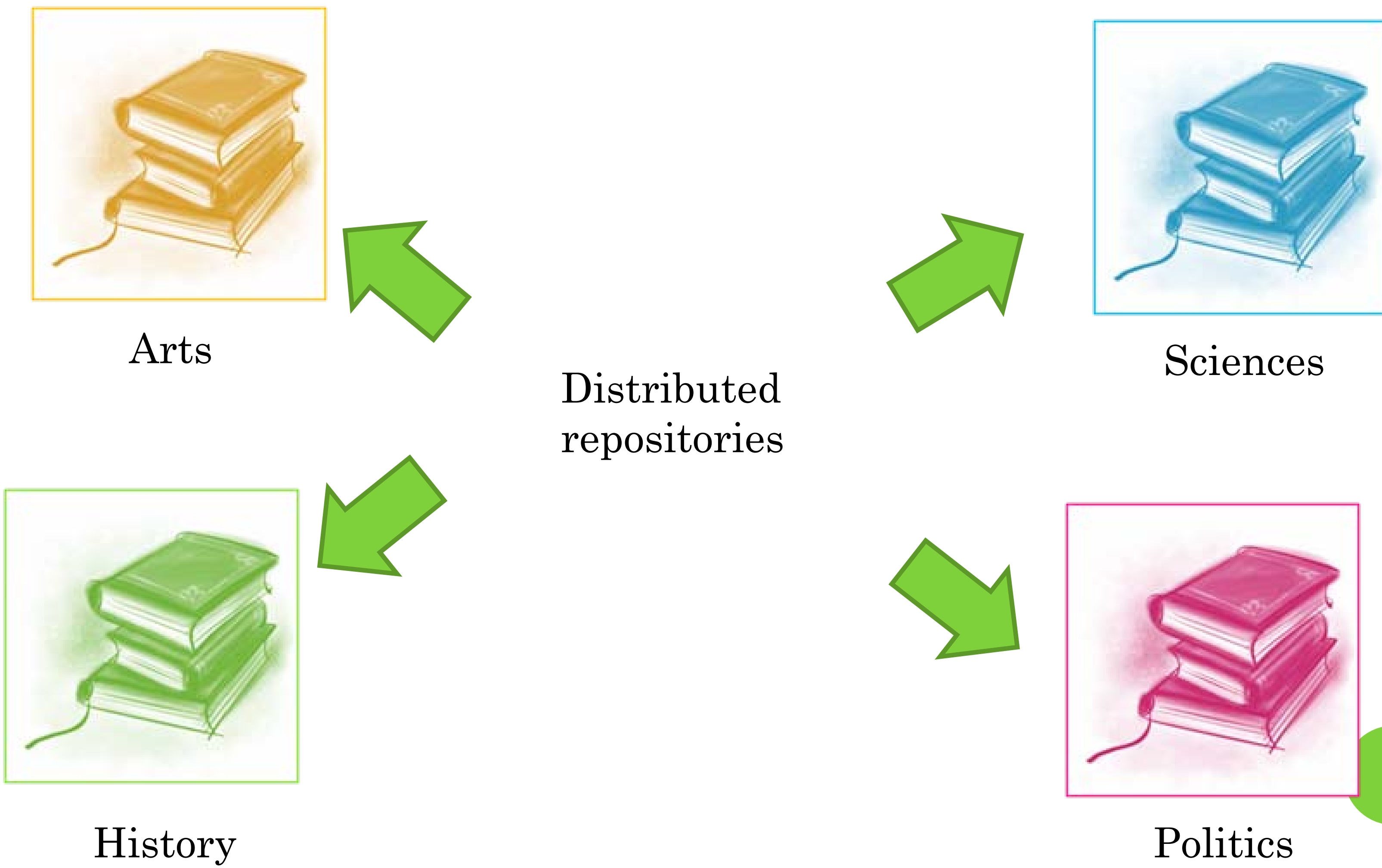


A Centralized Classifier

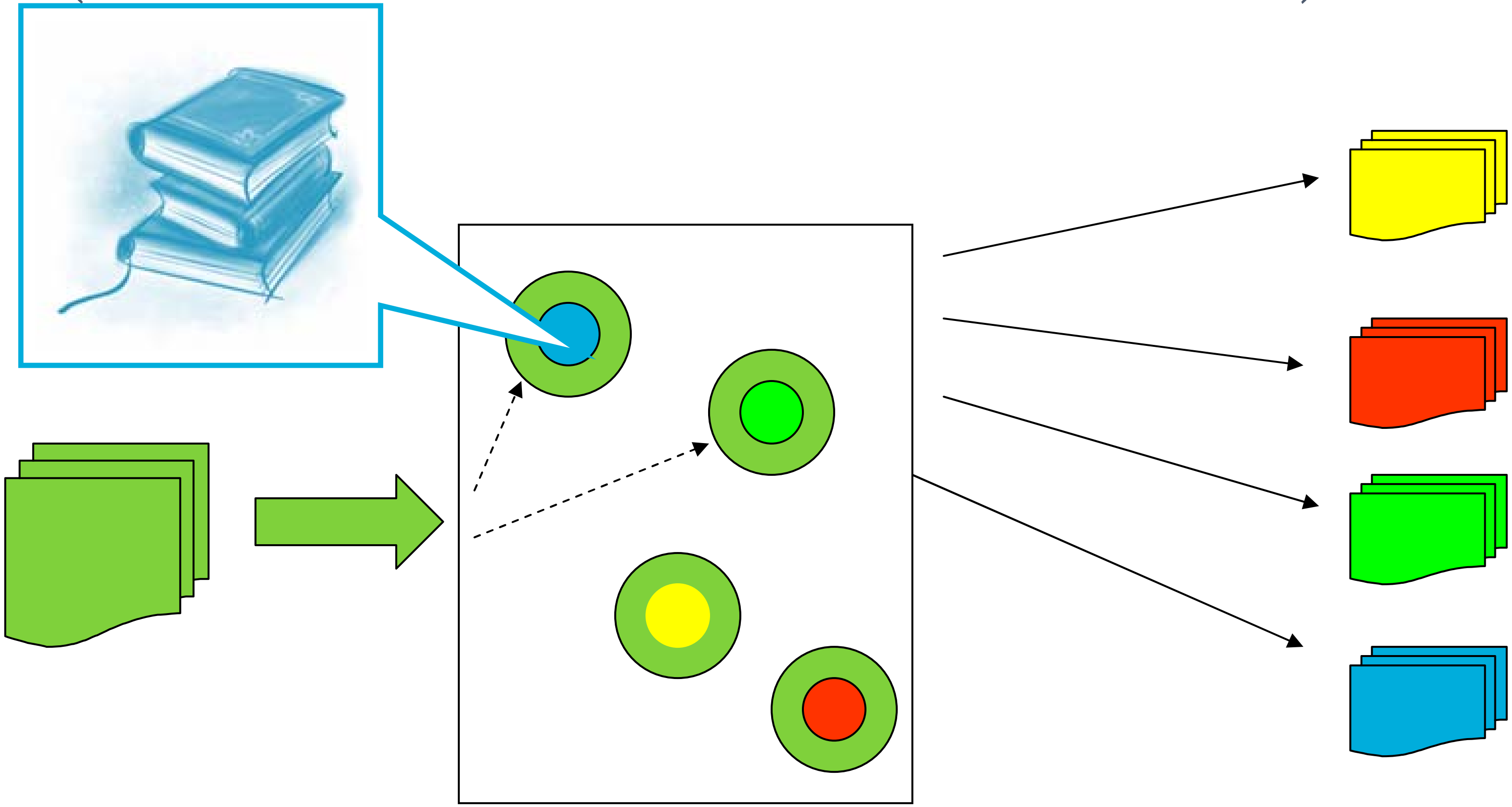
CENTRALIZED?

- Global repository is rarely realistic
  - Scalability
  - Intellectual property restrictions
  - …

KNOWLEDGE DISTRIBUTION



DISTRIBUTED CLASSIFICATION (WITHOUT COLLABORATION)

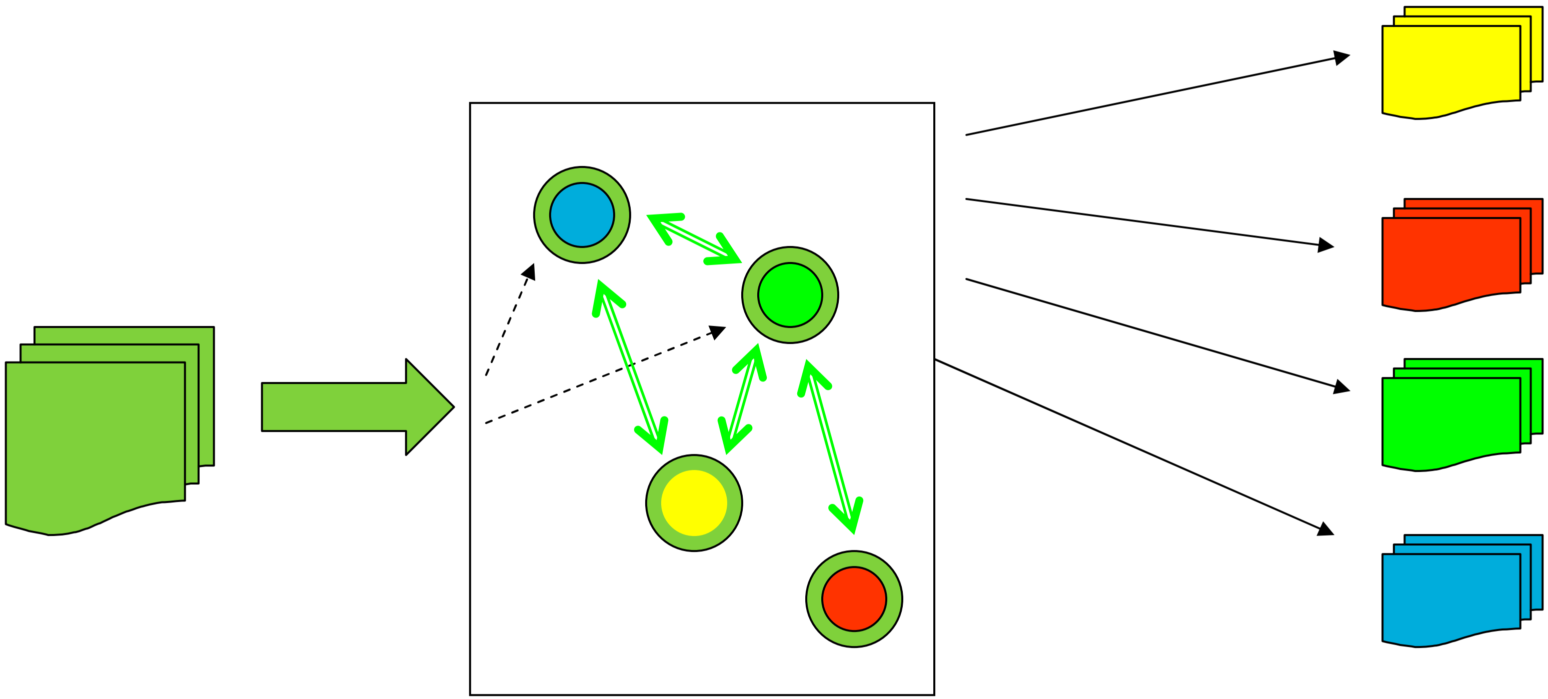


Distributed Classification without Collaboration

Classifier Knowledge

Doc Assignment

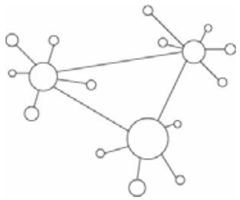
DISTRIBUTED CLASSIFICATION WITH COLLABORATION



Distributed Classification with Collaboration

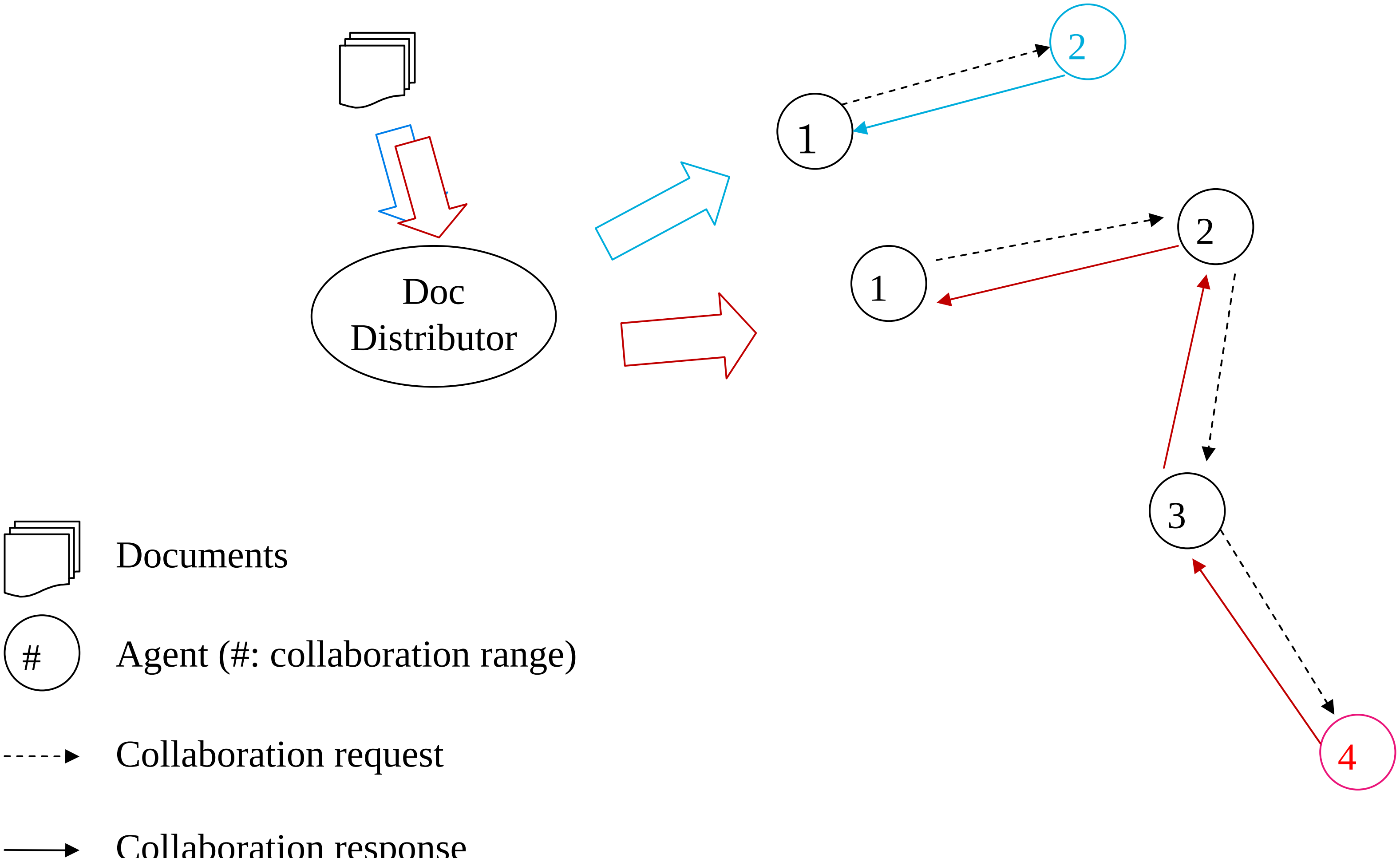
RESEARCH QUESTIONS

- Motivation: Why distributed text classification?
  - Knowledge is distributed
    - No global knowledge repository
      - e.g. individual digital libraries
  - Advantages of distributed methods:
    - fault tolerance, adaptability, flexibility, privacy, etc.
- Agents simulate distributed classifiers



- Problems/Questions
  - Agents have to learn and collaborate. But how?
  - Effectiveness and efficiency of agent collaboration?

DISTRIBUTED + COLLABORATION



METHODOLOGY

- Compare
  - Traditional/centralized approach (upper-bound)
  - Distributed approach without collaboration (lower-bound)
  - Distributed approach with collaboration
    - Two learning/collaboration algorithms:
      - Algorithm 1: Pursuit Learning
      - Algorithm 2: Nearest Centroid Learning
    - Two parameters
      - r: Exploration Rate
      - g: Maximum Collaboration Range
- Evaluation Measure
  - Effectiveness: precision, recall, F measure
  - Efficiency: time for classification

EXPERIMENT & EVALUATION

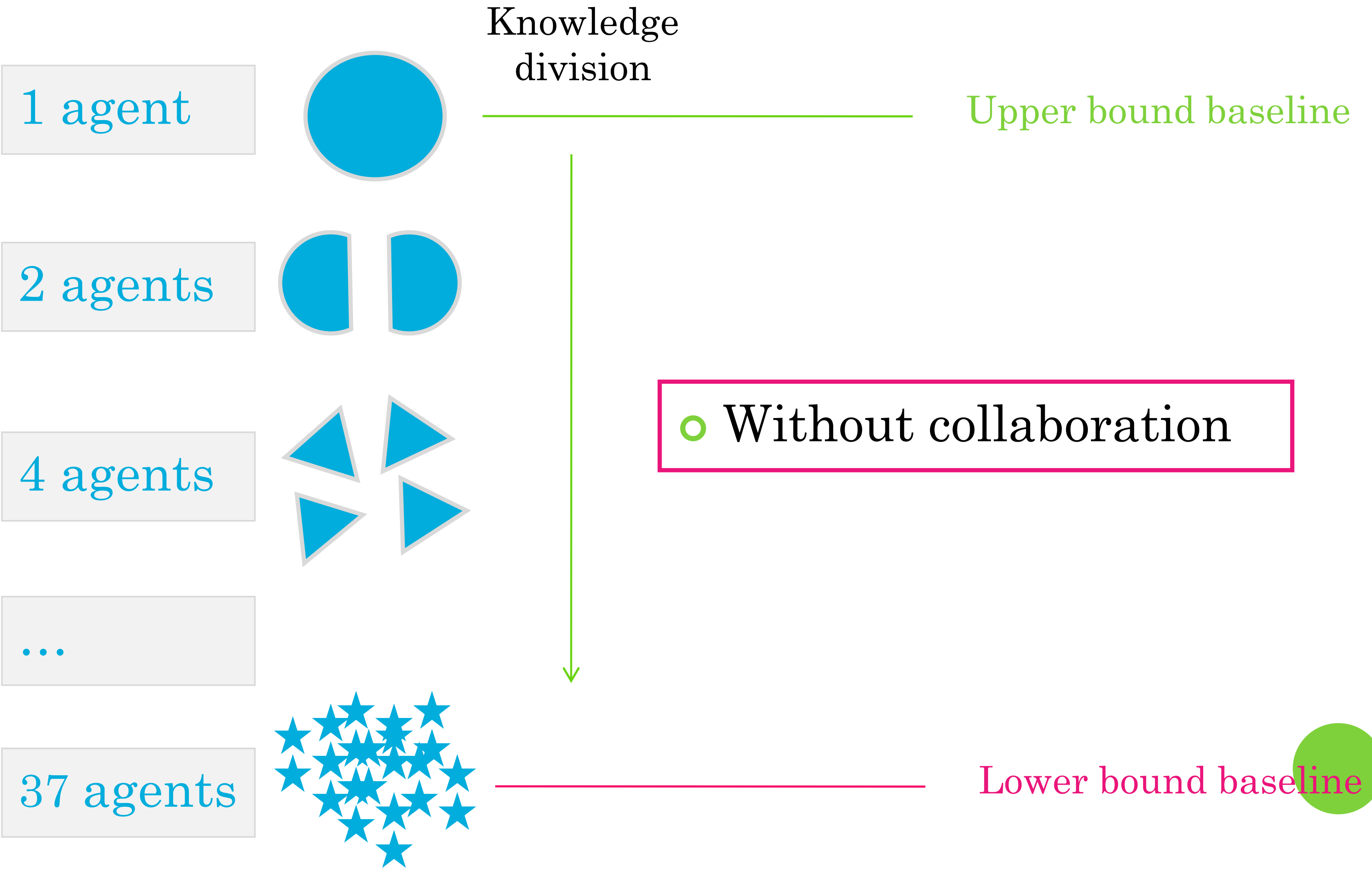
- Reuters Corpus Volumes 1 (RCV1)
- Training set: 6,394 documents
- Test set: 2,500 documents
- Feature selection: 4,084 unique terms
- Evaluation measures

Table 1: A contingency table

	Expert Says Yes	Expert Says No
System Says Yes	a	b
System Says No	c	d

- $Precision = a / (a + b)$
- $Recall = a / (a + c)$
- $F_1 = 2 * P * R / (P + R)$

EXPERIMENTAL RUN



RESULTS - EFFECTIVENESS BASELINES WITHOUT COLLABORATION

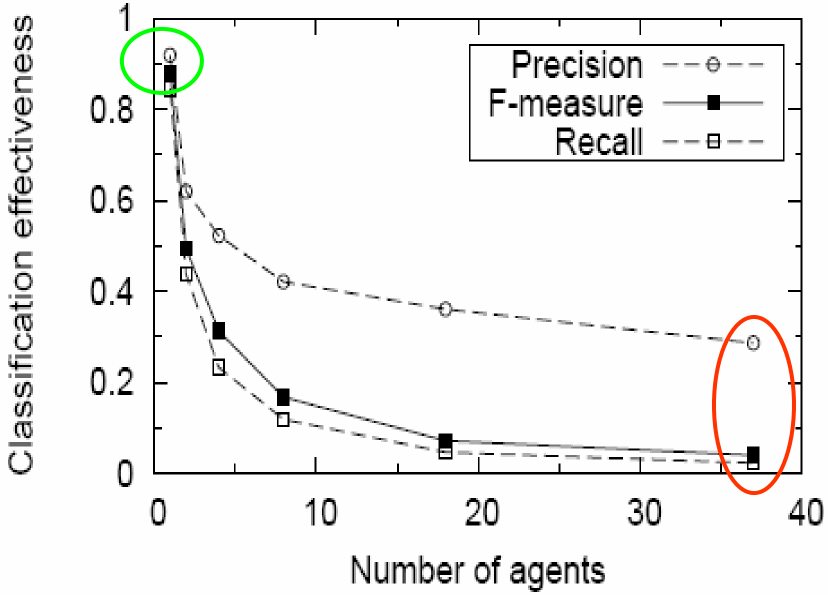


Figure 2: Classification effectiveness baseline

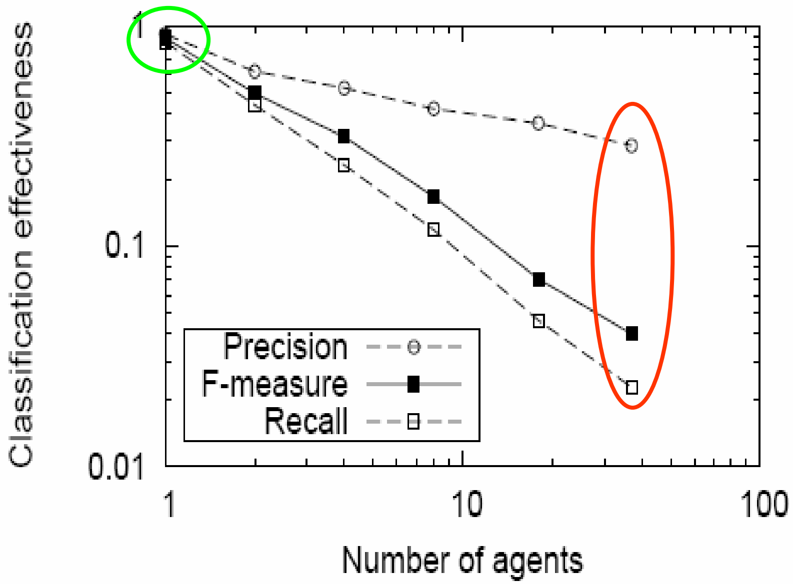
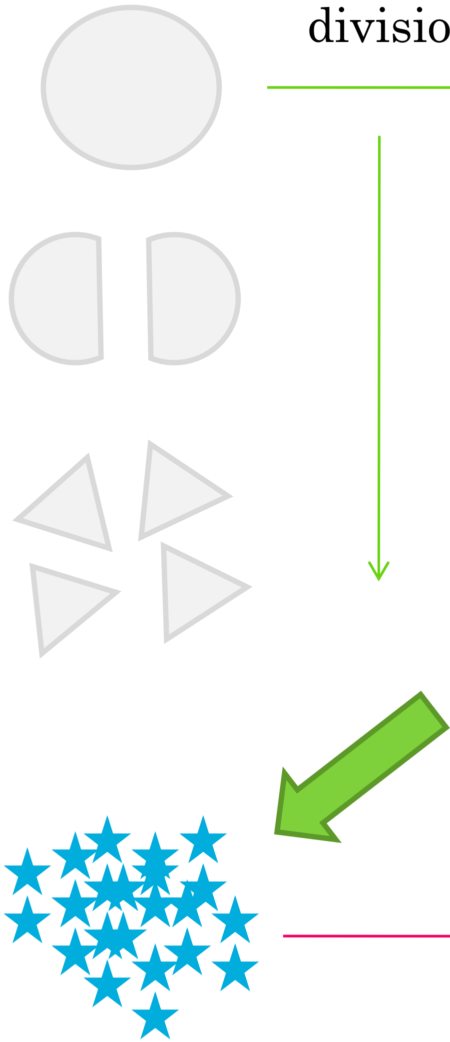


Figure 3: Effectiveness baseline (log/log)

Table 2: Baselines and some results for 37 agents

Method	g	r	R	P	F	Time (s)
Centralized			0.845	0.919	0.880	839
Non-collab			0.023	0.286	0.040	841
PL	8	.05	0.544	0.806	0.645	1022
NCL	8	.0	0.596	0.771	0.670	4205
PL	32	.1	0.681	0.753	0.715	1142
NCL	32	.1	0.558	0.719	0.628	7444

EXPERIMENTAL RUN



1 agent

2 agents

4 agents

...

37 agents

Knowledge division

Upper bound baseline

Lower bound baseline

- With collaboration
  - Pursuit leaning
  - Nearest centroid learning
  - Parameters
    - Exploration rate
    - Max collaboration range

RESULTS – EFFECTIVENESS OF LEARNING

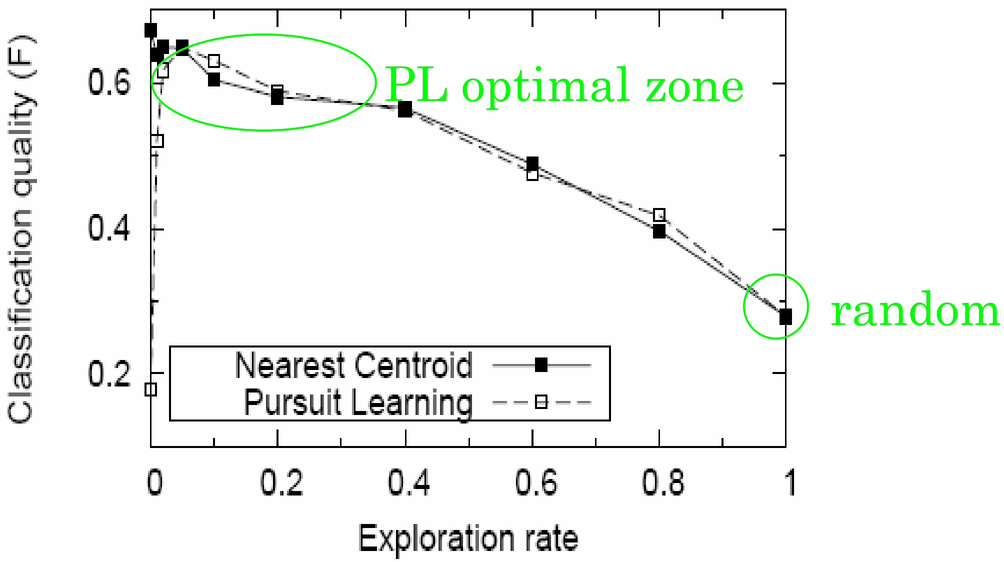


Figure 5: Classification effectiveness vs. exploration rate ($\#agents = 37, g = 8$ while $r \in [0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0]$)

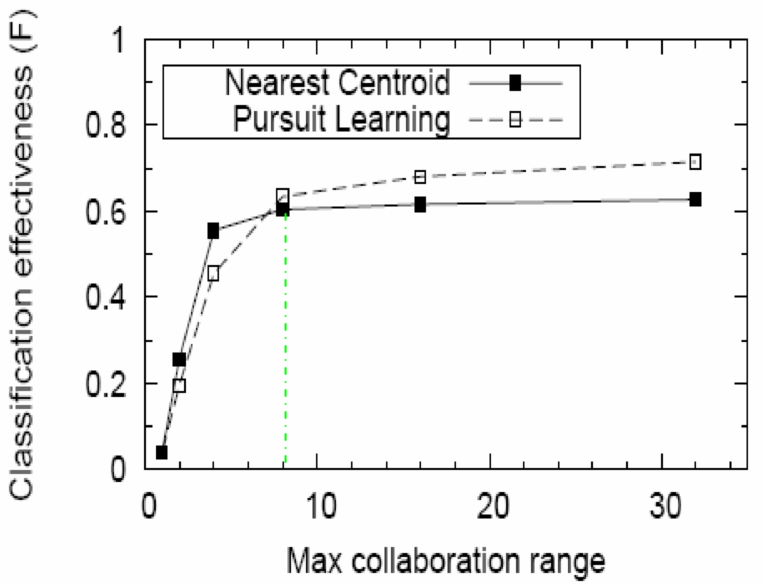


Figure 4: Classification effectiveness vs. max collaboration range ($\#agents = 37, r = 0.1$ while $g \in [2^0, 2^1, .., 2^5]$)

RESULTS – CLASSIFICATION EFFICIENCY

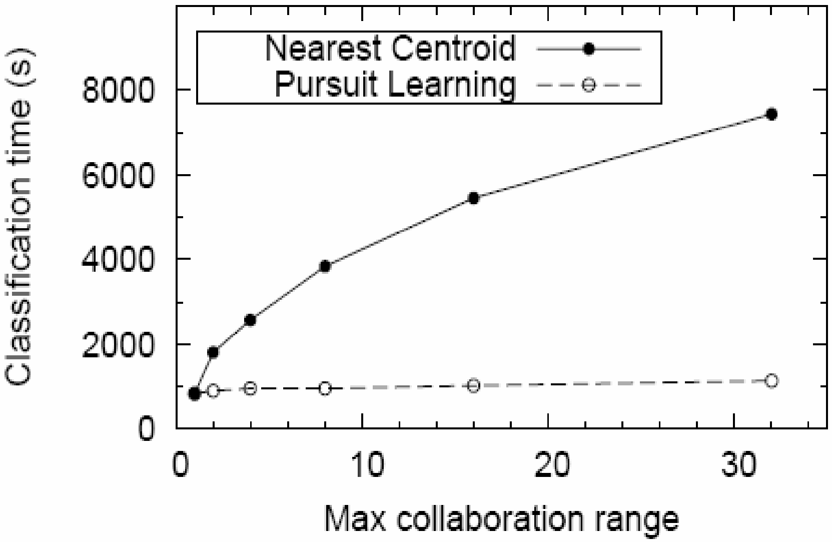


Figure 11: Classification efficiency vs. Max collaboration range ($\#agents = 37, r = 0.1$)

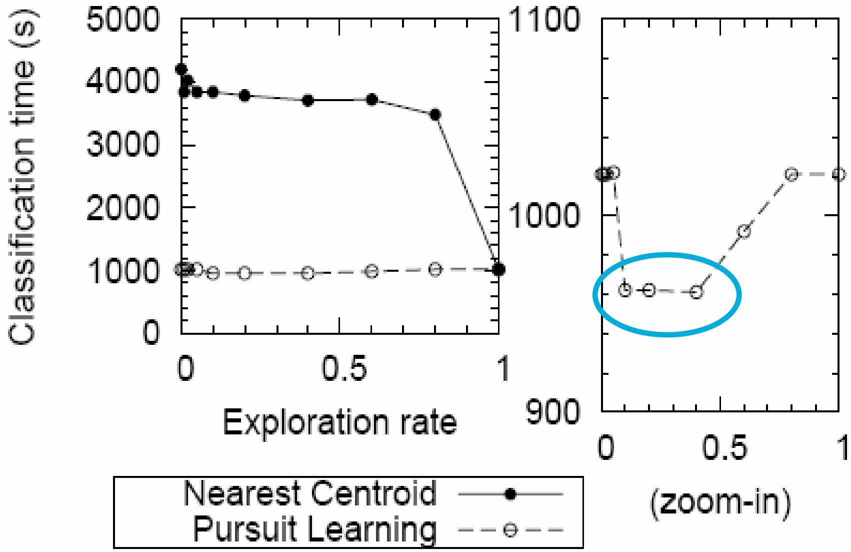


Figure 12: Classification efficiency vs. Exploration rate ($\#agents = 37, g = 8$)

Right: zoom-in of the Pursuit Learning curve.

RESULTS – EFFICIENCY VS. EFFECTIVENESS

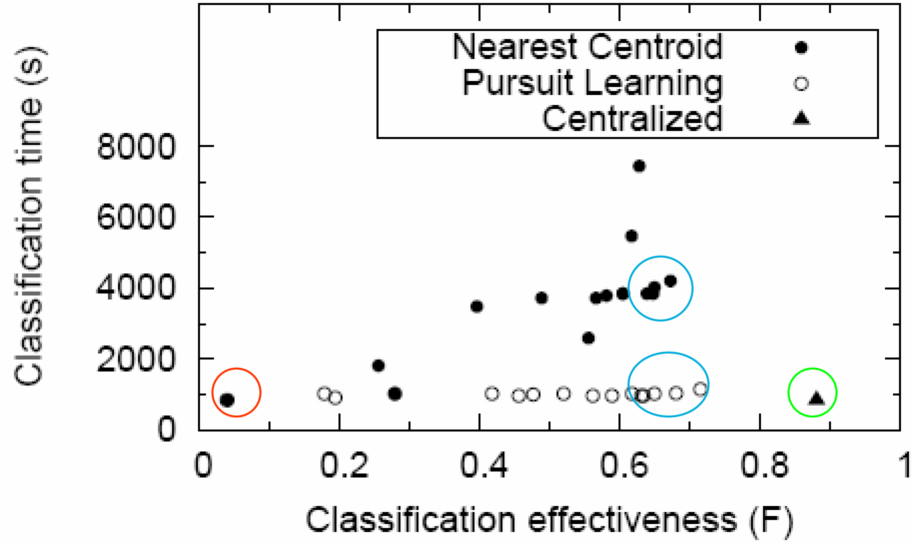


Figure 13: Classification efficiency vs. Effectiveness

SUMMARY

- Classification effectiveness decreases dramatically when knowledge becomes increasingly distributed.
- Pursuit Learning
  - Efficient – without analyzing contents
  - Effective, although not content sensitive
  - "*The Pursuit Learning approach did not depend on document content. By acquiring knowledge through reinforcements based on collaborations this algorithm was able to construct/build paths for documents to find relevant classifiers effectively and efficiently.*"
- Nearest Centroid Learning
  - Inefficient – analyzing content
  - Effective
- Future work
  - Other text collections
  - Knowledge overlap among the agents
  - Local neighborhood