Convei-Lab Seminar

2021-08-03

Wonpyo Hong

Contents

- Concept Review: Novelty Detection
- Goal of the Research
  - Big Picture
  - Insights & Motivation
- Approach
  - Problems and Solutions
- Future Work

Concept Review

- Novelty Detection

TASK \ SETTING	TRAINING	TESTING	GOAL
Traditional Classification	Known known classes	Known known classes	Classifying known known classes
Classification with Reject Option	Known known classes	Known known classes	Classifying known known classes & rejecting samples of low confidence
One-class Classification (Anomaly Detection)	Known known classes & few or none outliers from KUCs	Known known classes & few or none outliers	Detecting outliers
One/Few-shot Learning	Known known classes & a limited number of UKCs' samples	Unknown known classes	Identifying unknown known classes
Generalized Few-shot Learning	Known known classes & a limited number of UKCs' samples	Known known classes & unknown known classes	Identifying known known classes & unknown known classes
Zero-shot Learning	Known known classes & side-information[1]	Unknown known classes	Identifying unknown known classes
Generalized Zero-shot Learning	Known known classes & side-information[1]	Known known classes & unknown known classes	Identifying known known classes & unknown known classes
Open Set Recognition	Known known classes	Known known classes & unknown unknown classes	Identifying known known classes & rejecting unknown unknown classes
Generalized Open Set Recognition	Known known classes & side-information[2]	Known known classes & Unknown unknown classes	Identifying known known classes & cognizing unknown unknown classes

Concept Review

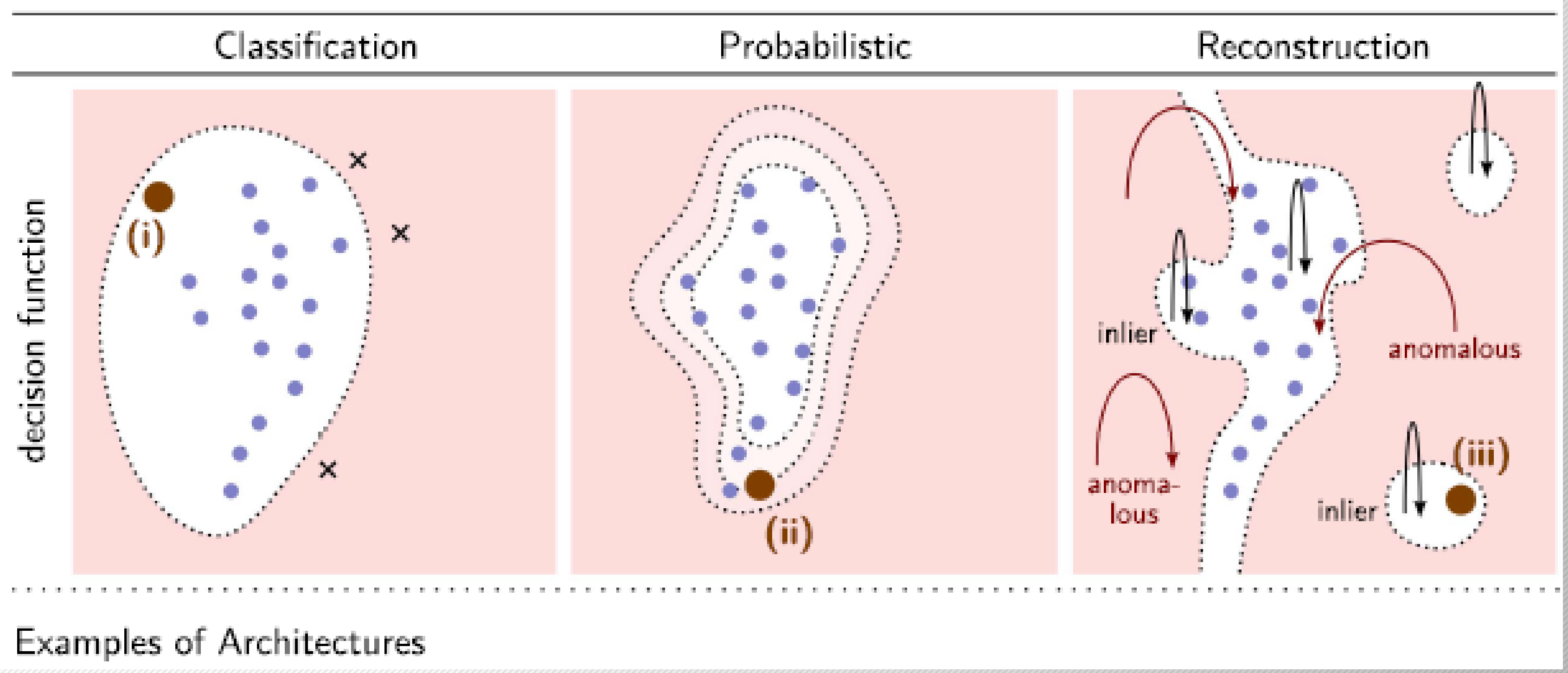
- Novelty Detection

Novelty detection is the identification of new or unknown data or signals that a machine learning system is not aware of during training.

용어	비정상 sample
Novelty Detection	지금까지 등장하지 않았지만 충분히 등장할 수 있는 sample
Outlier Detection + Anomaly	지금까지 등장하지 않았고 앞으로도 등장할 가능성이 없는, 데이터에 오염이 발생했을 가능성이 있는 sample

Concept Review

- Novelty Detection

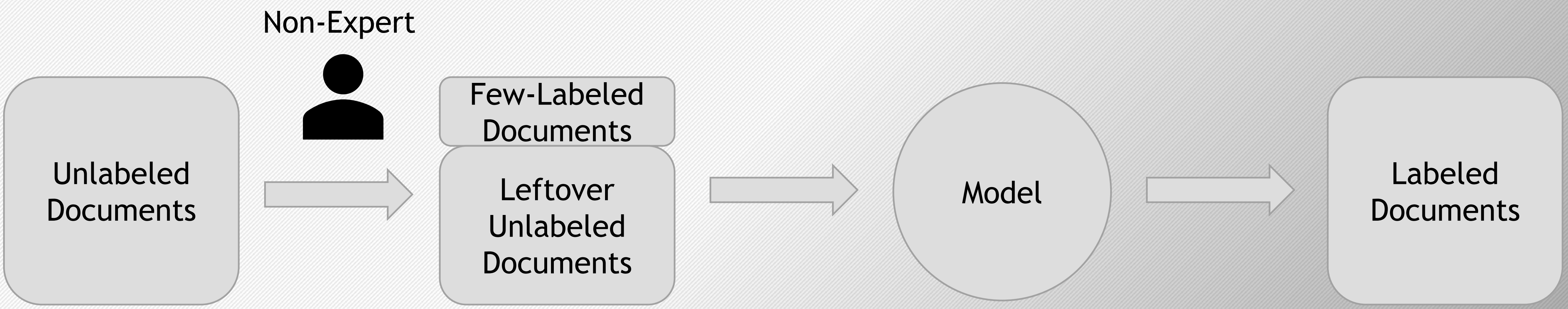


Goal of the Research

- ~~Open Set Recognition~~
  - ~~Identify Unknown Classes (Missing Labels)~~

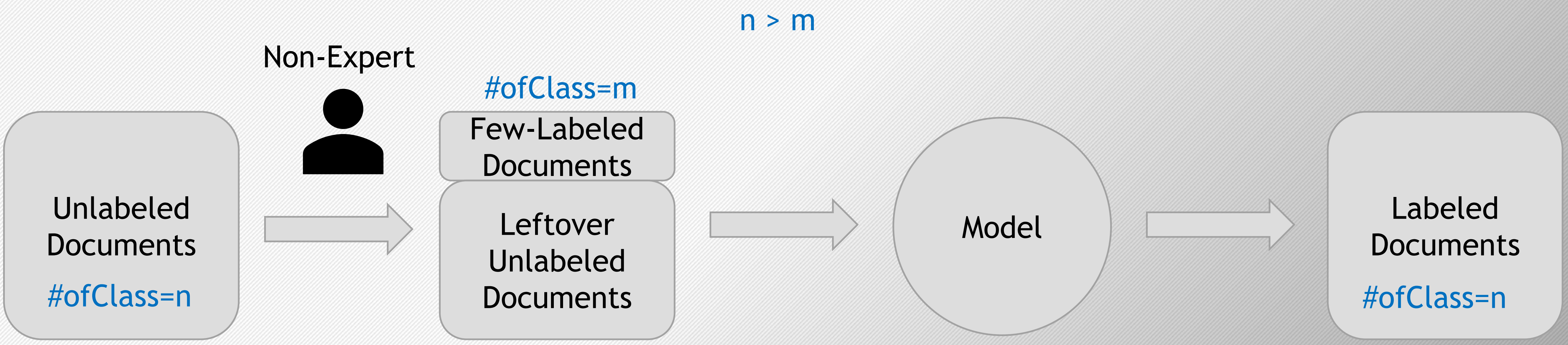
- Iterative Semi-Supervised Classification with Novelty Detection
  - Identify Unknown Classes (Unknown Labels)

Big Picture



What is done: Few-Shot Semi-Supervised Classification

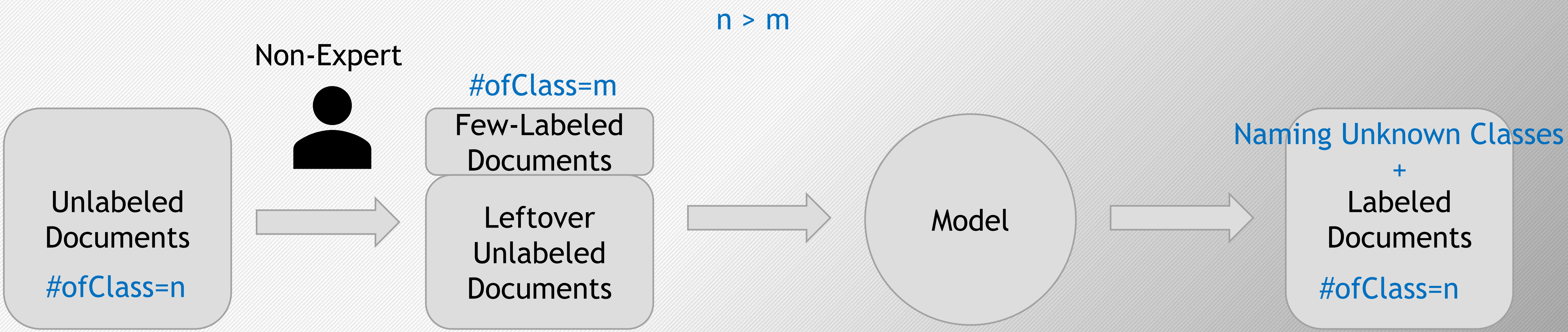
Big Picture



What I'm doing: Anomaly Detection

What is done: Few-Shot Semi-Supervised Classification

Big Picture



What I can do in future: Identification

What I'm doing: Anomaly Detection

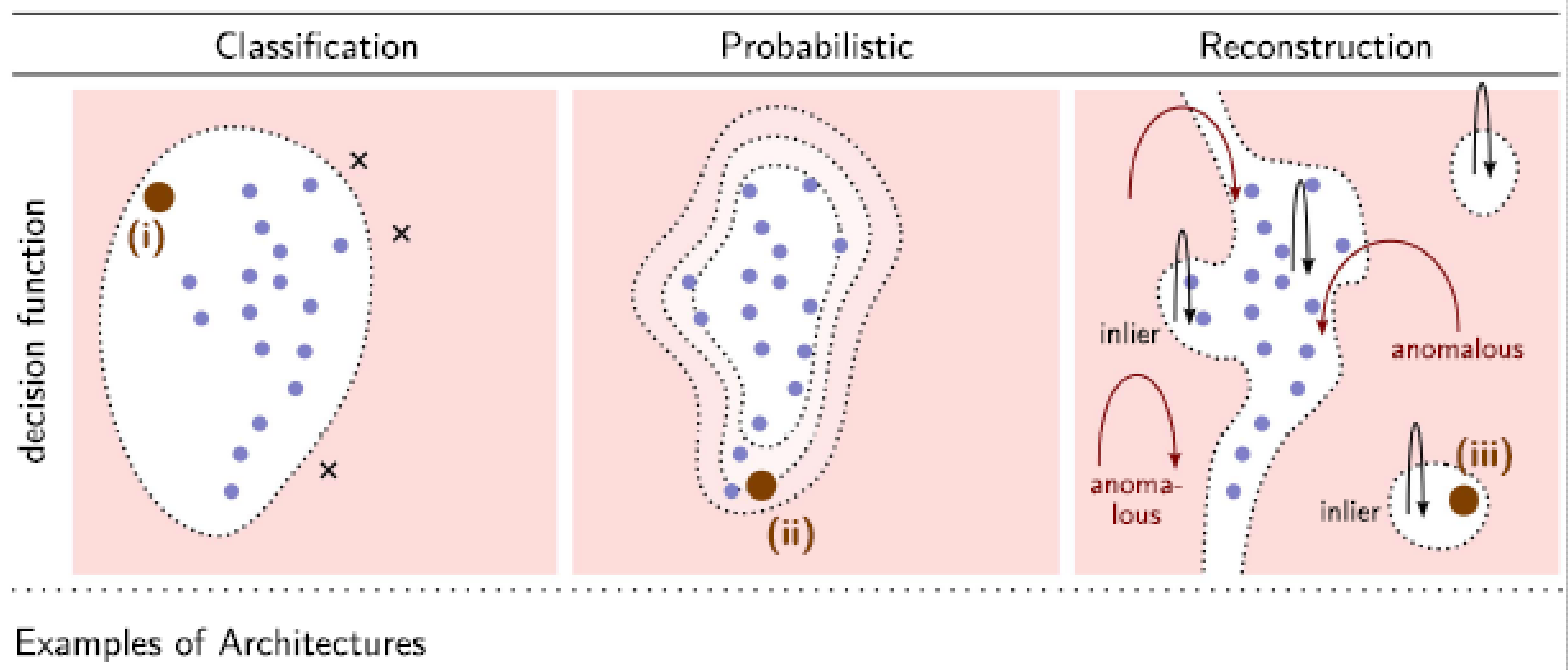
What is done: Few-Shot Semi-Supervised Classification

Challenges

- Unknown Class Detection
  - How to Detect Unknown Classes in SSC environment?
- Class Identification
  - How to appropriately Group unknown documents, and then Name them?

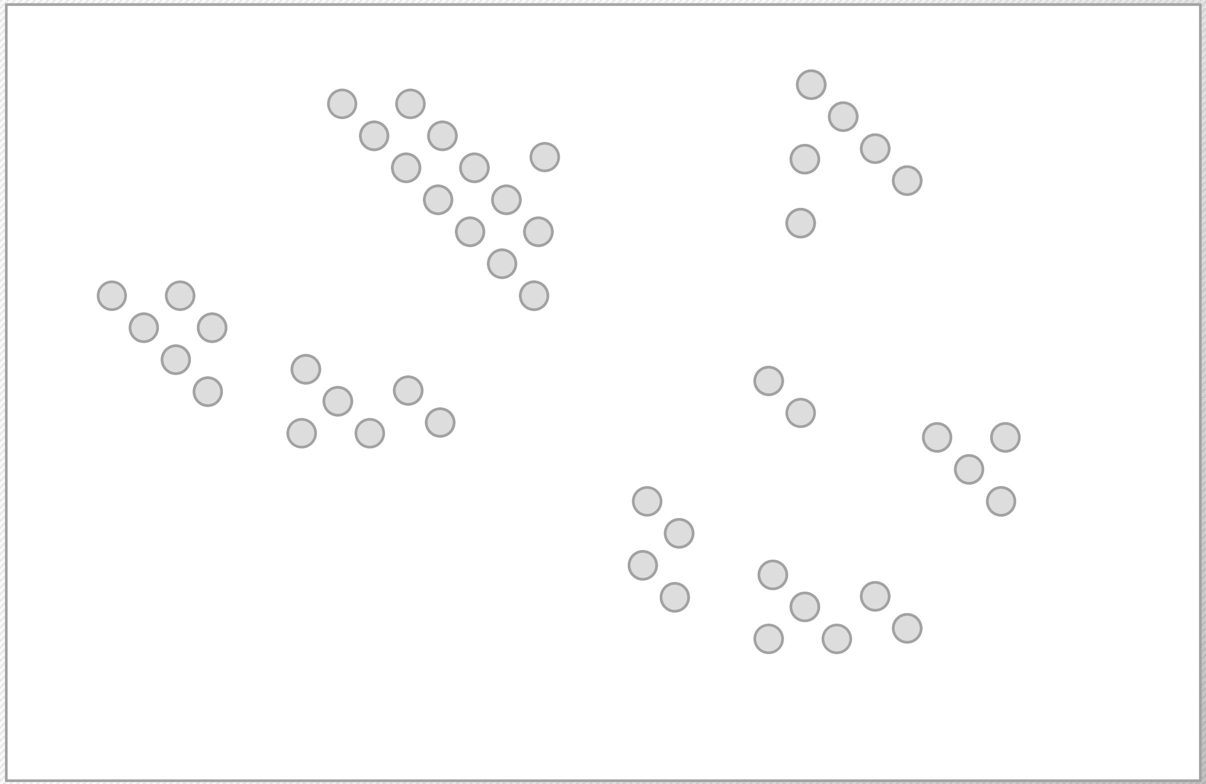
Insight & Motivation

- New Architecture: Iterative Removal of Known-Classes Data



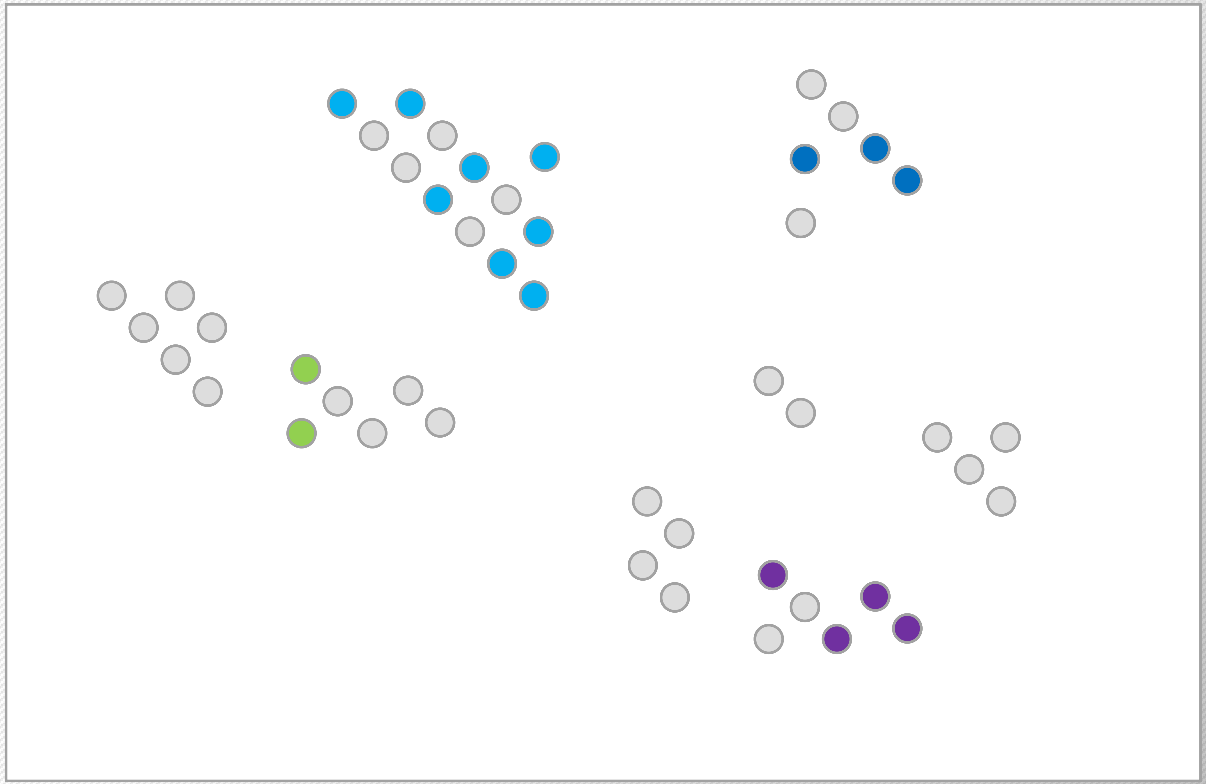
Insight & Motivation

- New Architecture: Iterative Removal of Known-Classes Data



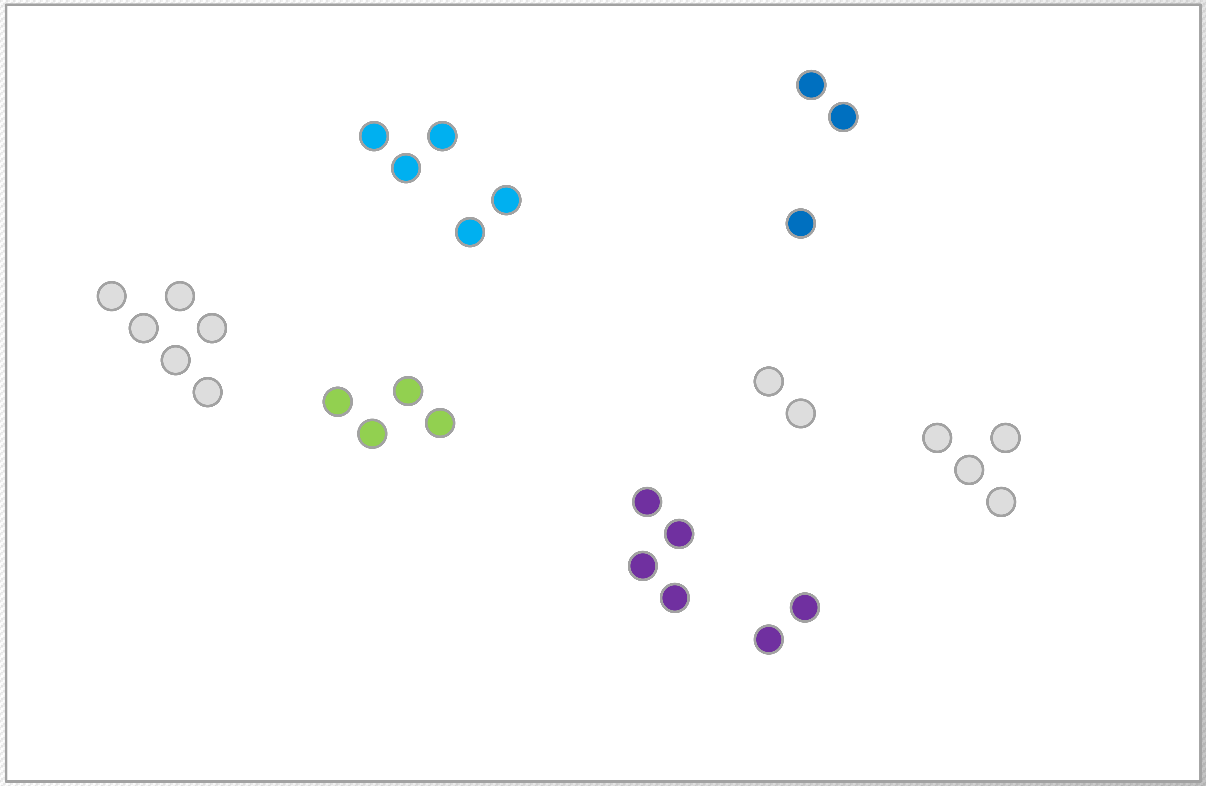
Insight & Motivation

- New Architecture: Iterative Removal of Known-Classes Data



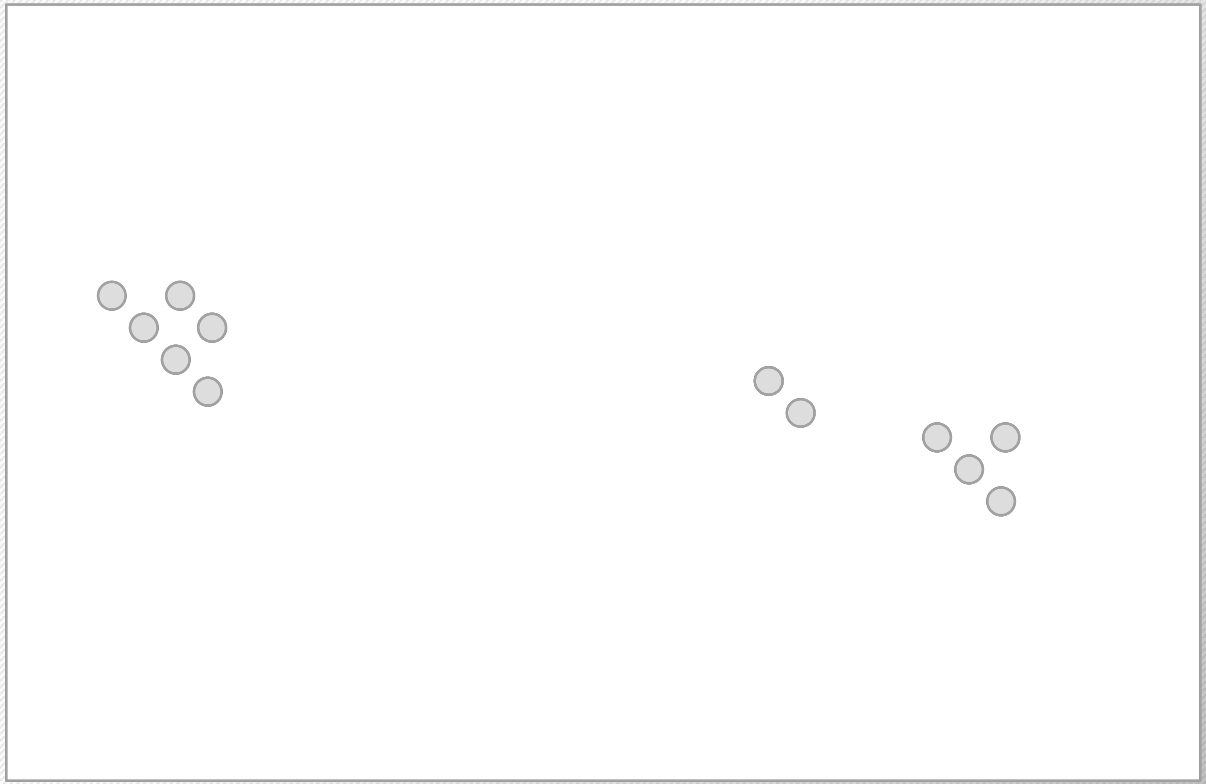
Insight & Motivation

- New Architecture: Iterative Removal of Known-Classes Data



Insight & Motivation

- New Architecture: Iterative Removal of Known-Classes Data



Insight & Motivation

- Lexical Synthesis
  - Set to Set Similarity

Assumption:

Each Lexicon of a class contains similar structure to one another.

Insight & Motivation

- Lexical Synthesis

    Assumption:

    Each Lexicon of a class contains similar structure to one another.

Insight & Motivation

- Lexical Synthesis

  Assumption:

  Each Lexicon of a class contains similar structure to one another.

Basketball
Lexicon$_1$
Basketball, Basket, Score, NBA, Assist, Rebound, Backboard, …

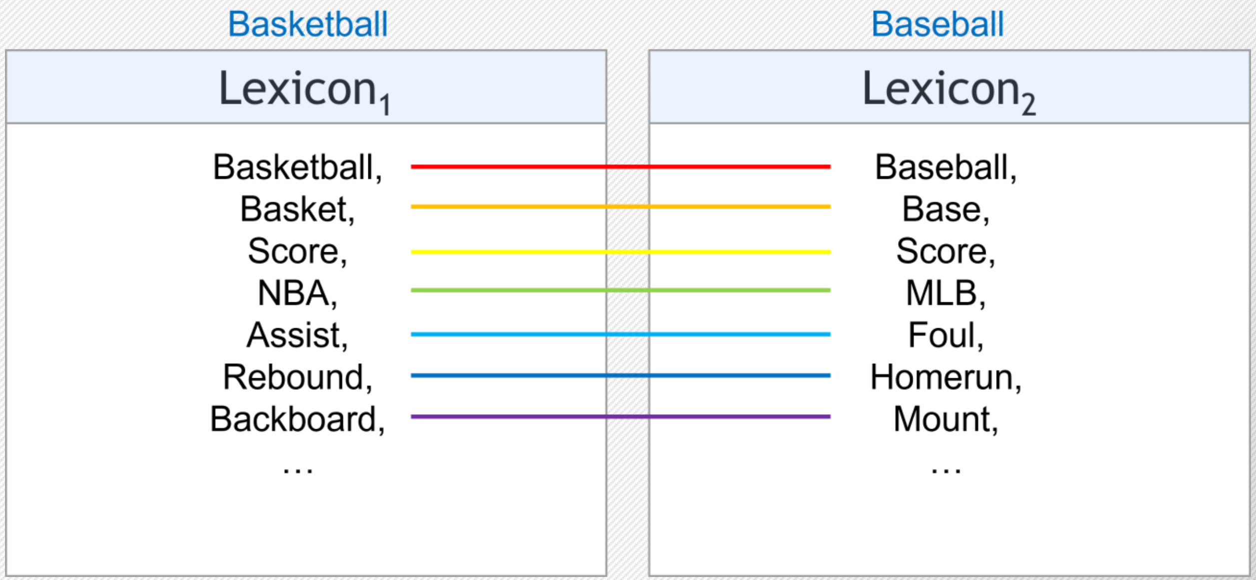
Baseball
Lexicon$_2$
Baseball, Base, Score, MLB, Foul, Homerun, Mount, …

Insight & Motivation

- Lexical Synthesis

  Assumption:

  Each Lexicon of a class contains similar structure to one another.

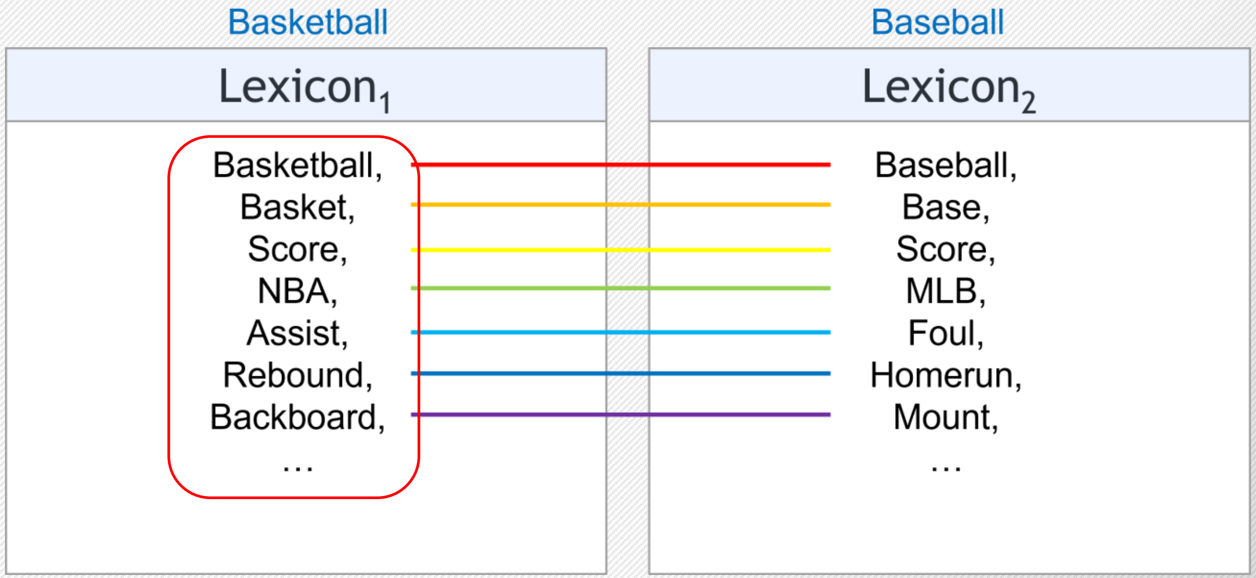


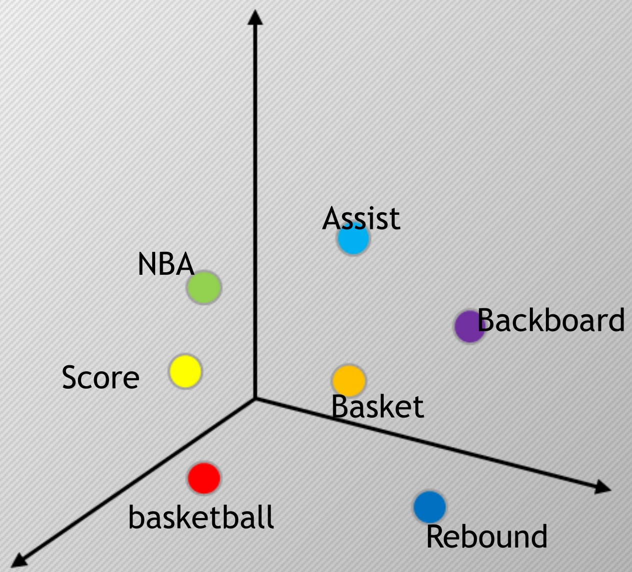
Insight & Motivation

- Lexical Synthesis

  Assumption:

  Each Lexicon of a class contains similar structure to one another.



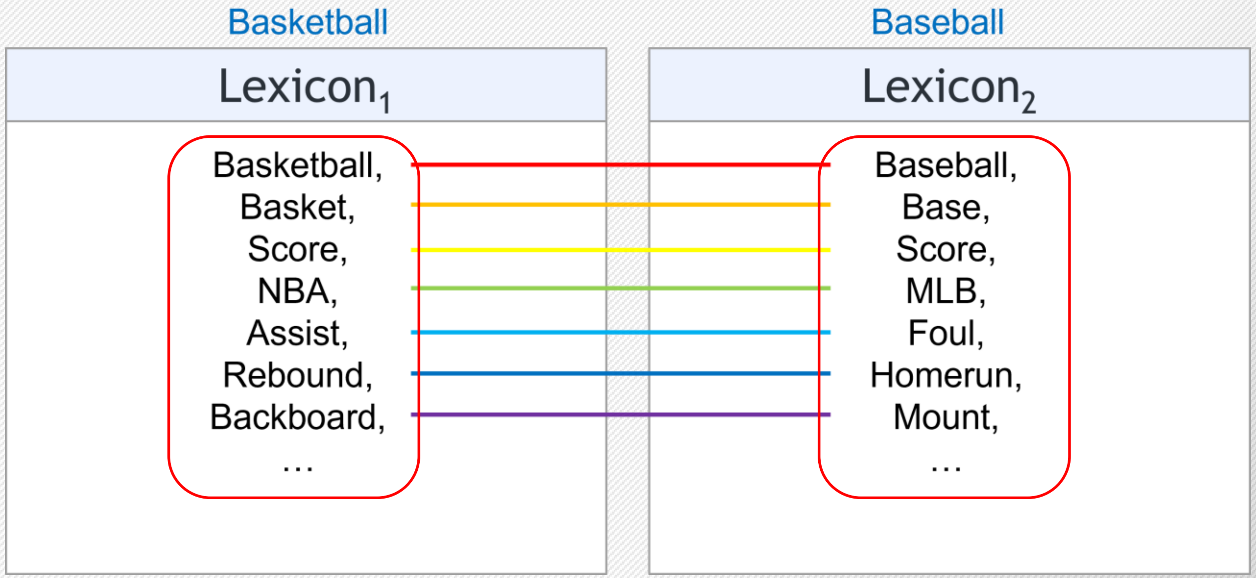


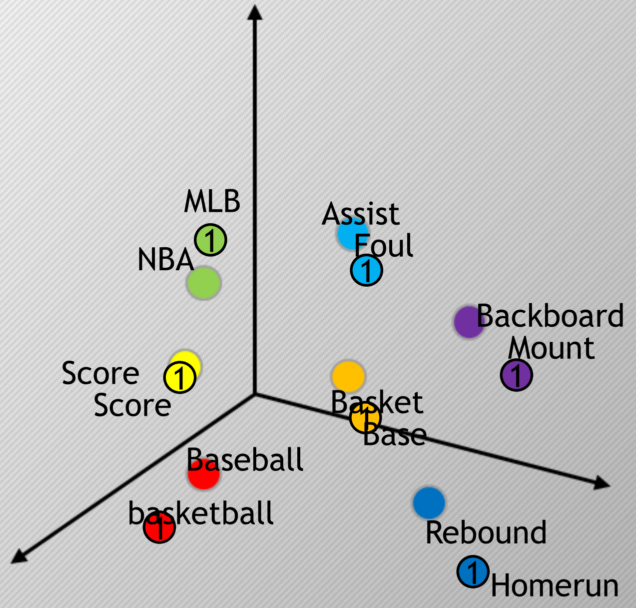
Insight & Motivation

- Lexical Synthesis

  Assumption:

  Each Lexicon of a class contains similar structure to one another.



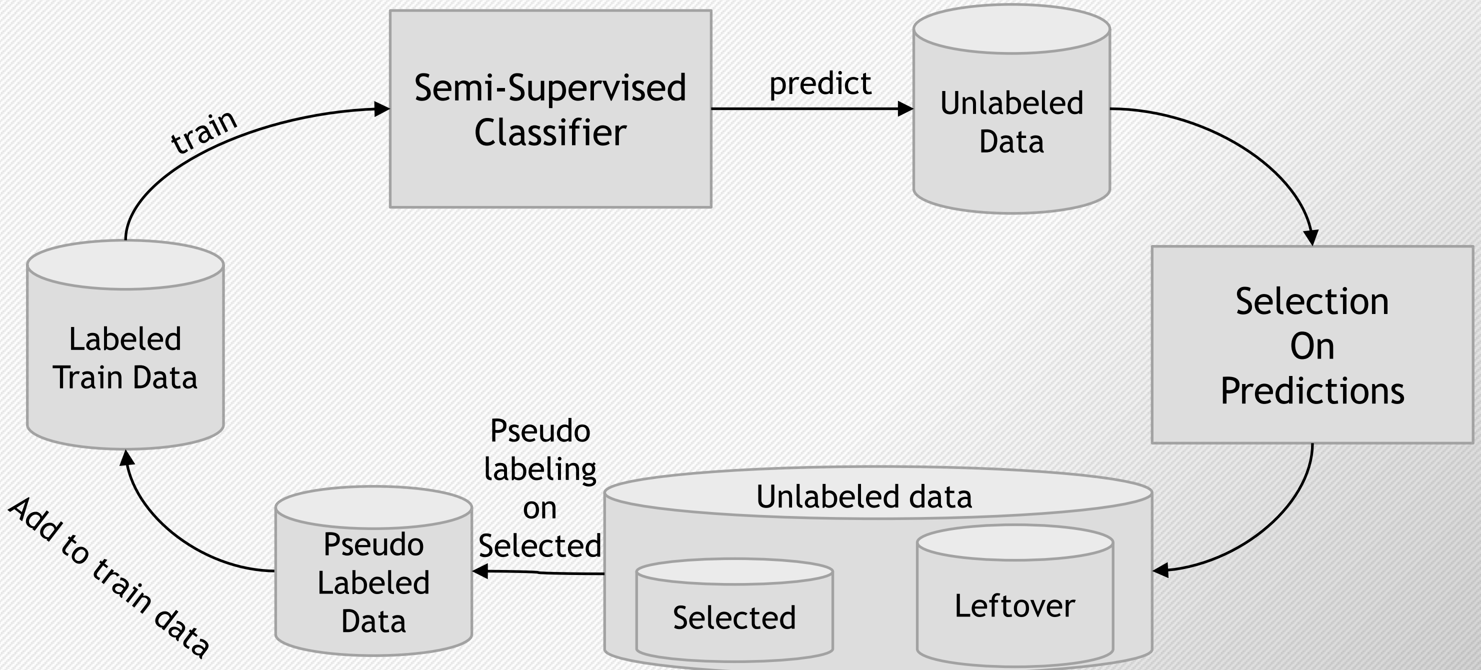


Solutions

- Unknown Class Detection
  - Iterative Semi-Supervised Classification
- Class Identification
  - Use Lexical Representation instead of Document Representation
  - Inference of Lexical Synthesis

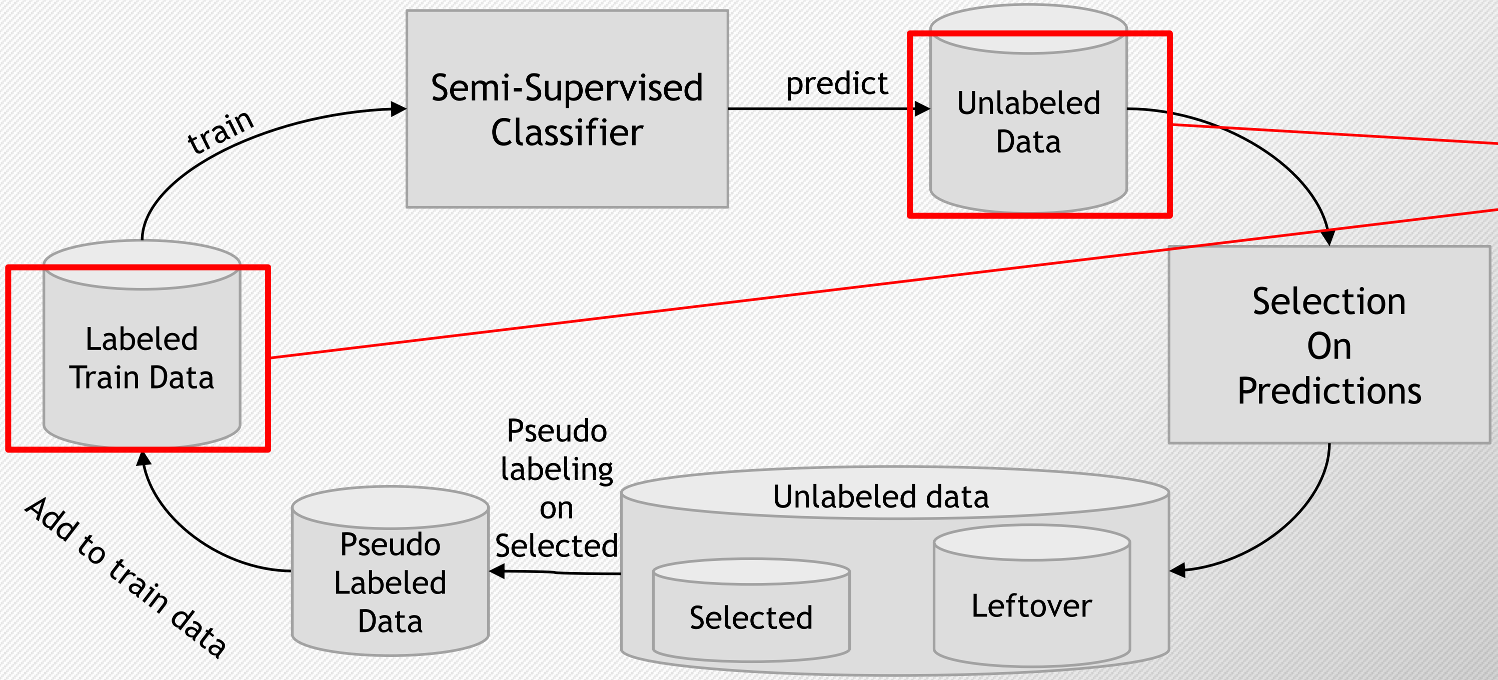
Solutions

- Unknown Class Detection
  - Iterative Semi-Supervised Classification



Solutions

- Unknown Class Detection
  - Iterative Semi-Supervised Classification

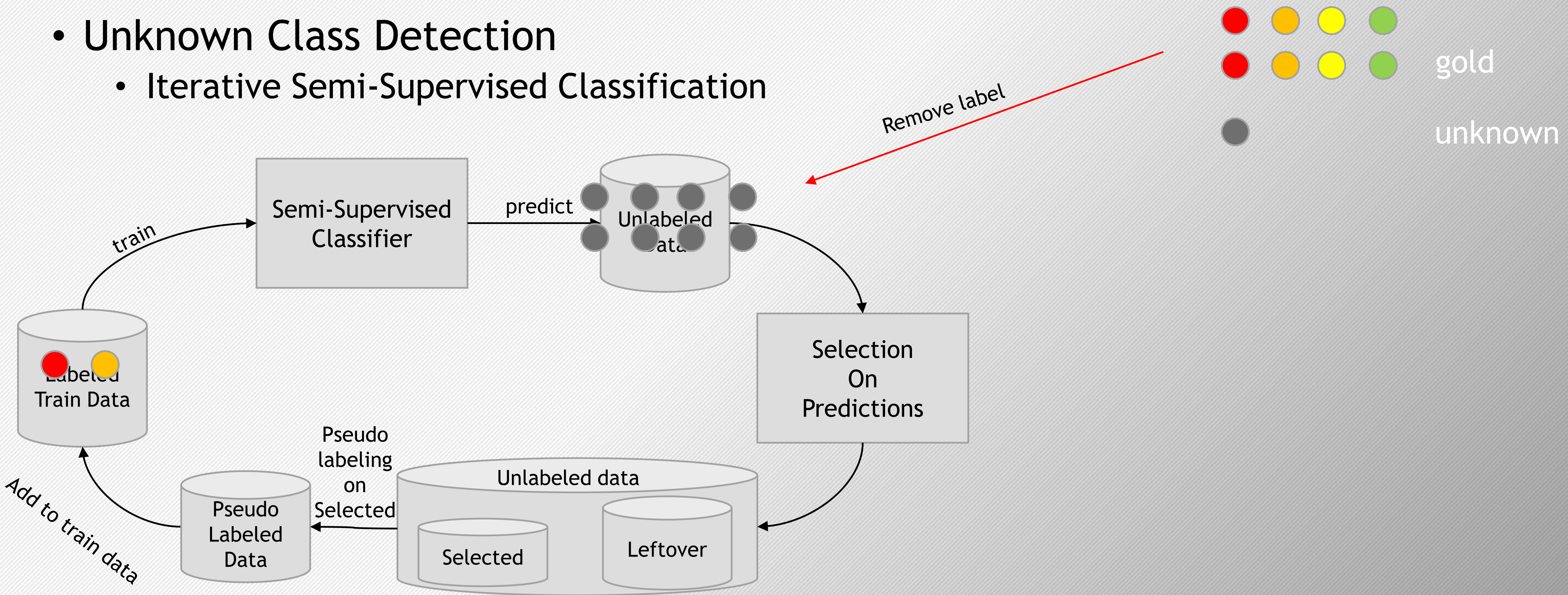


Both come from Original Training Data

(SALNet: 1% vs 99%)

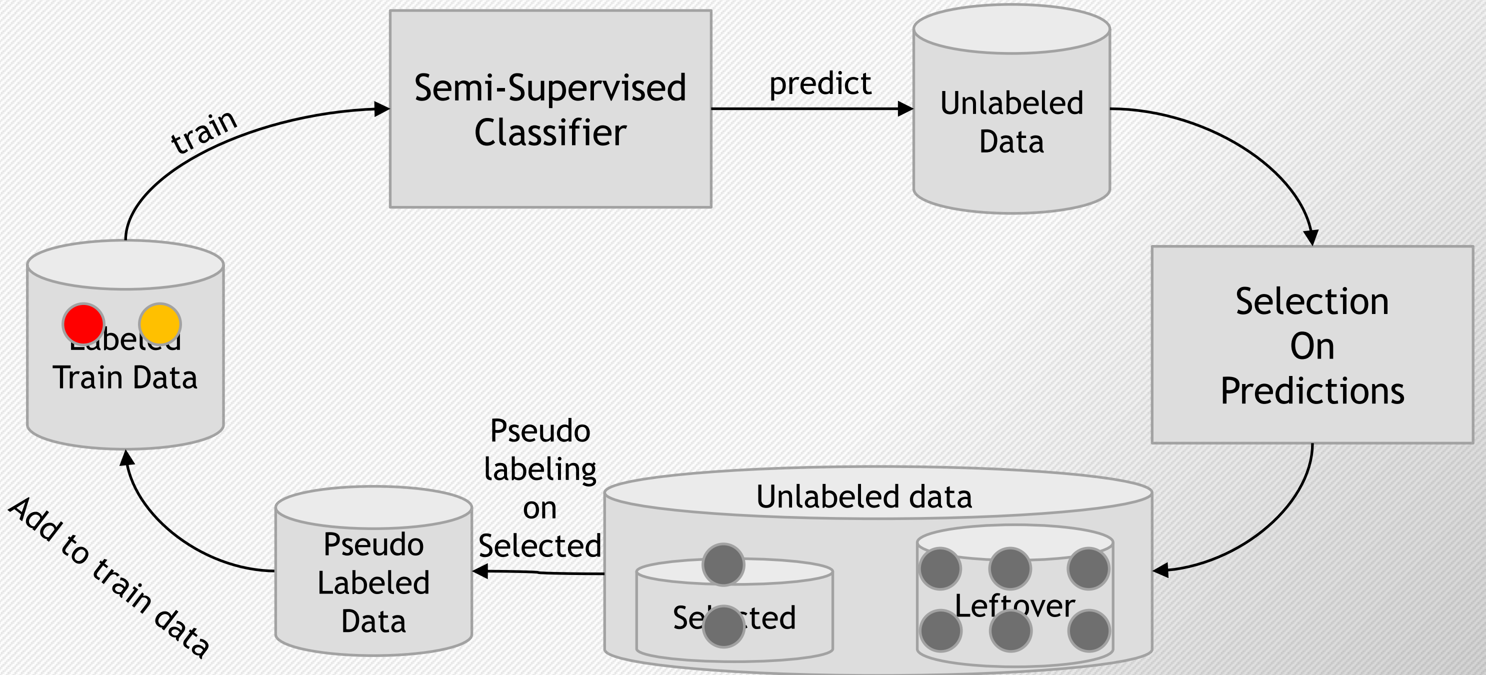
Solutions

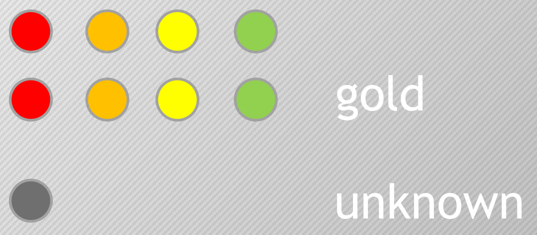
- Unknown Class Detection
  - Iterative Semi-Supervised Classification



Solutions

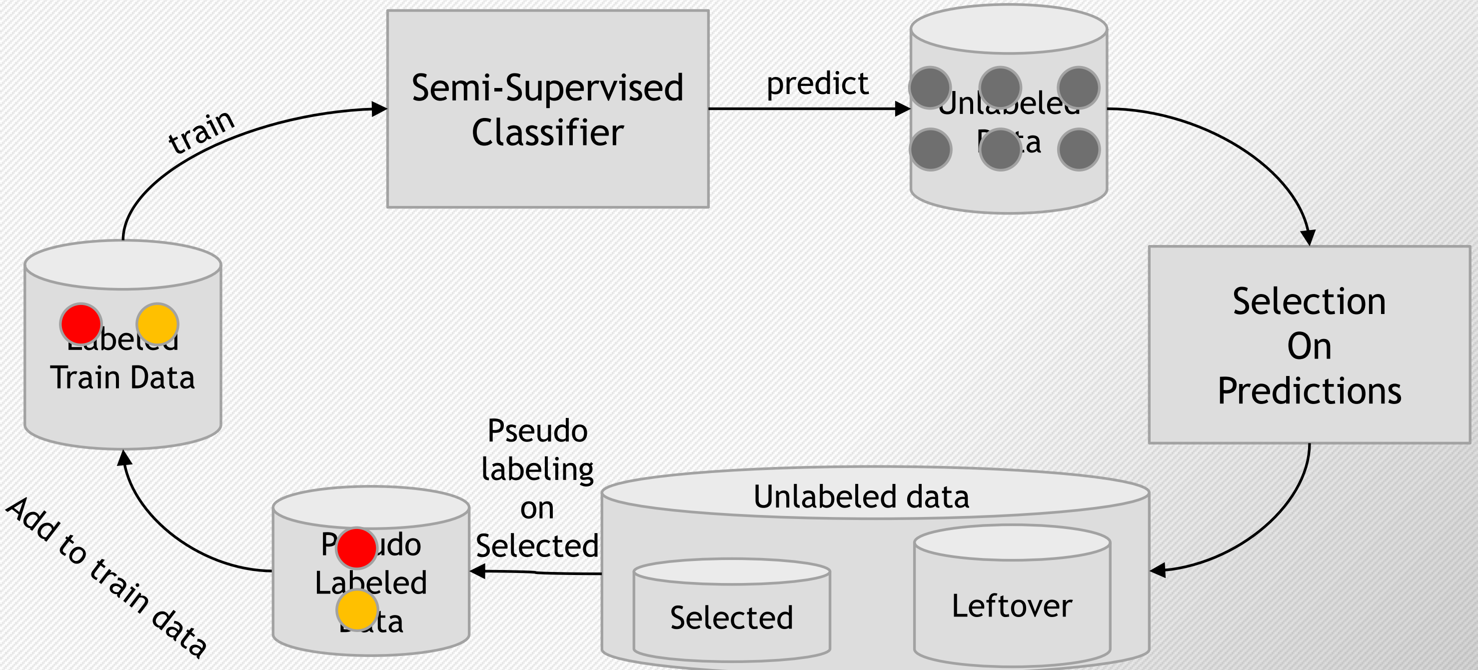
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

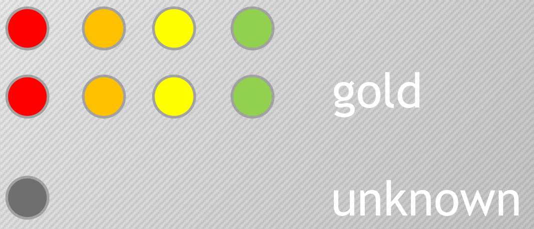




Solutions

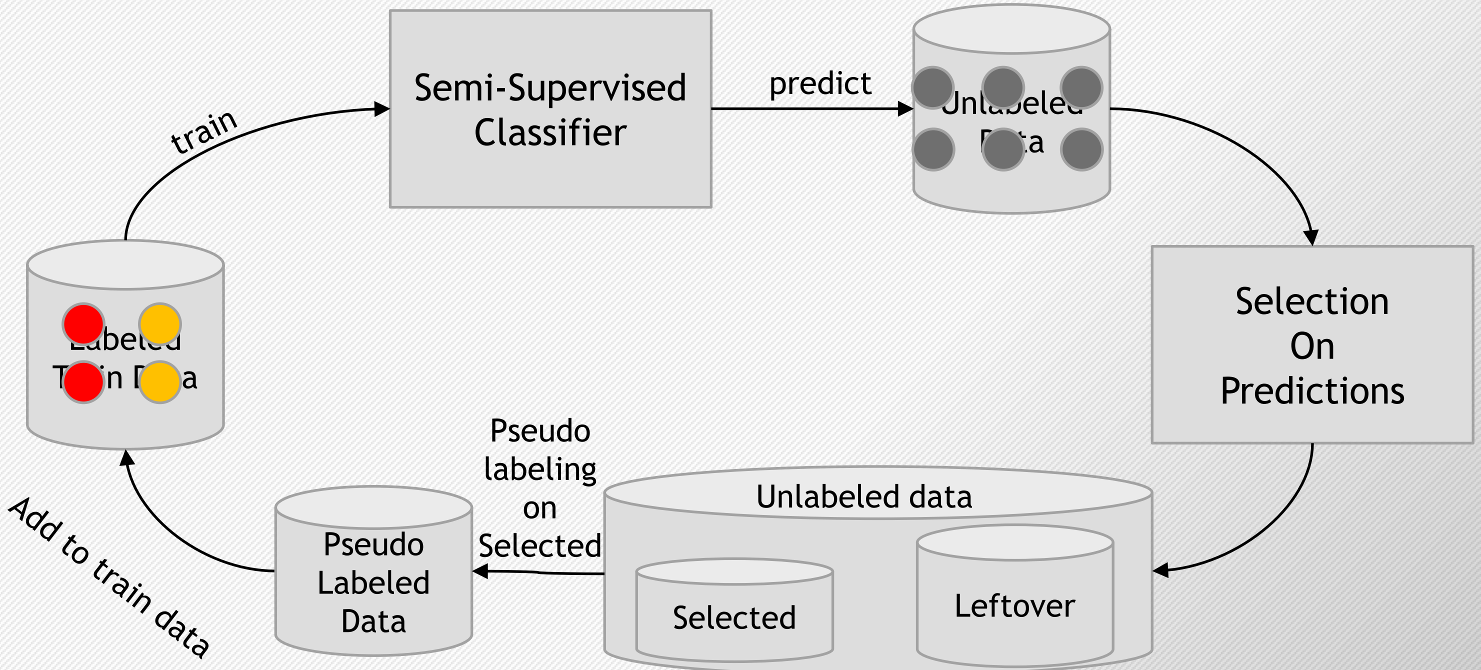
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

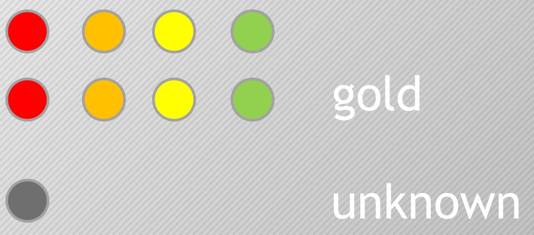




Solutions

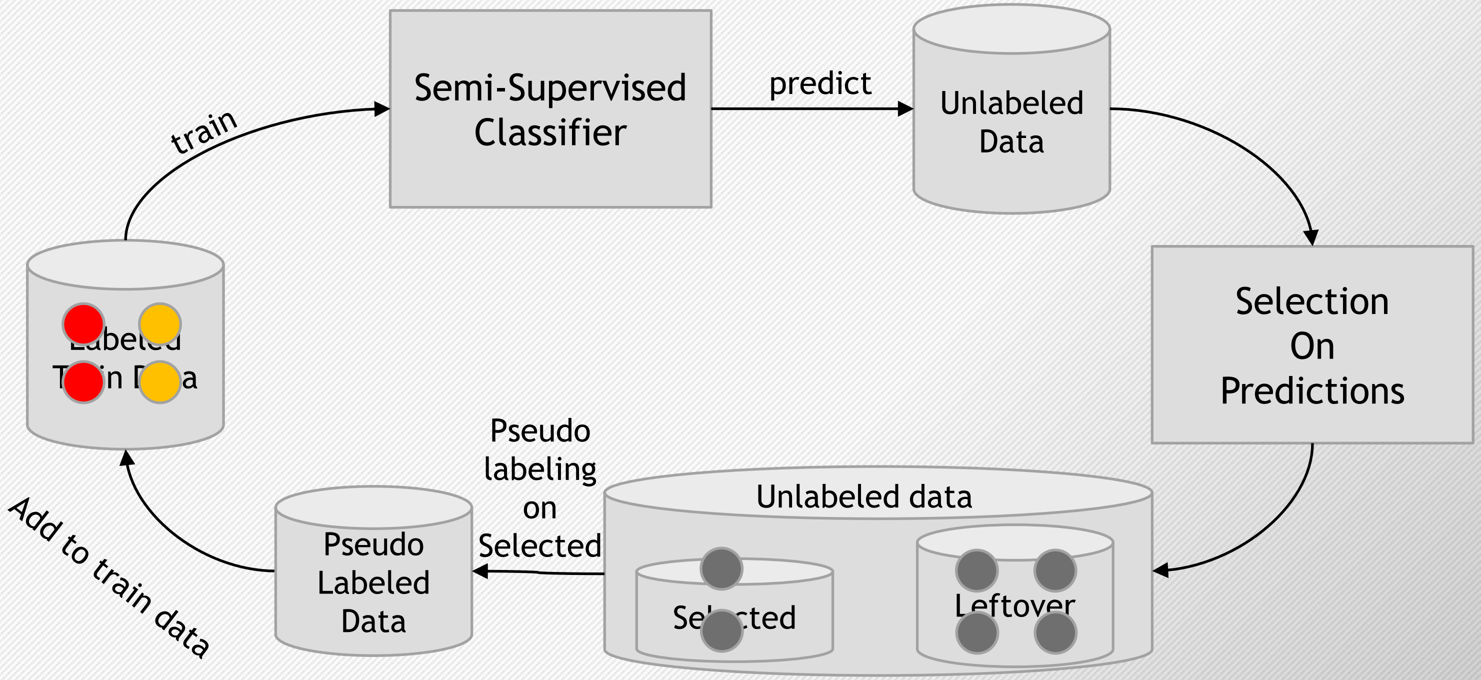
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

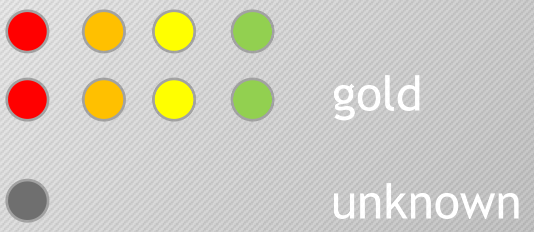




Solutions

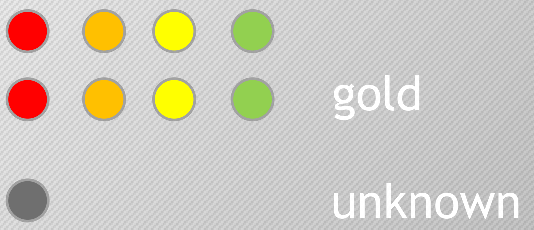
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

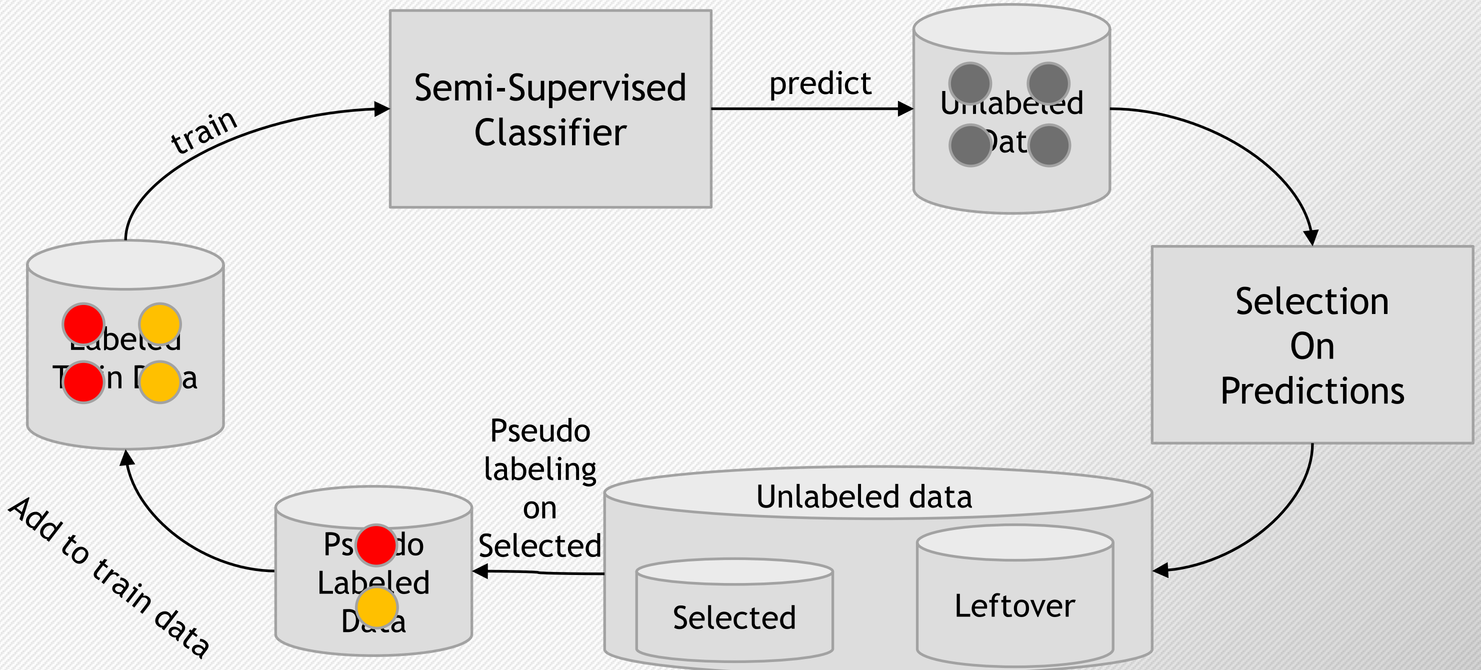




Solutions

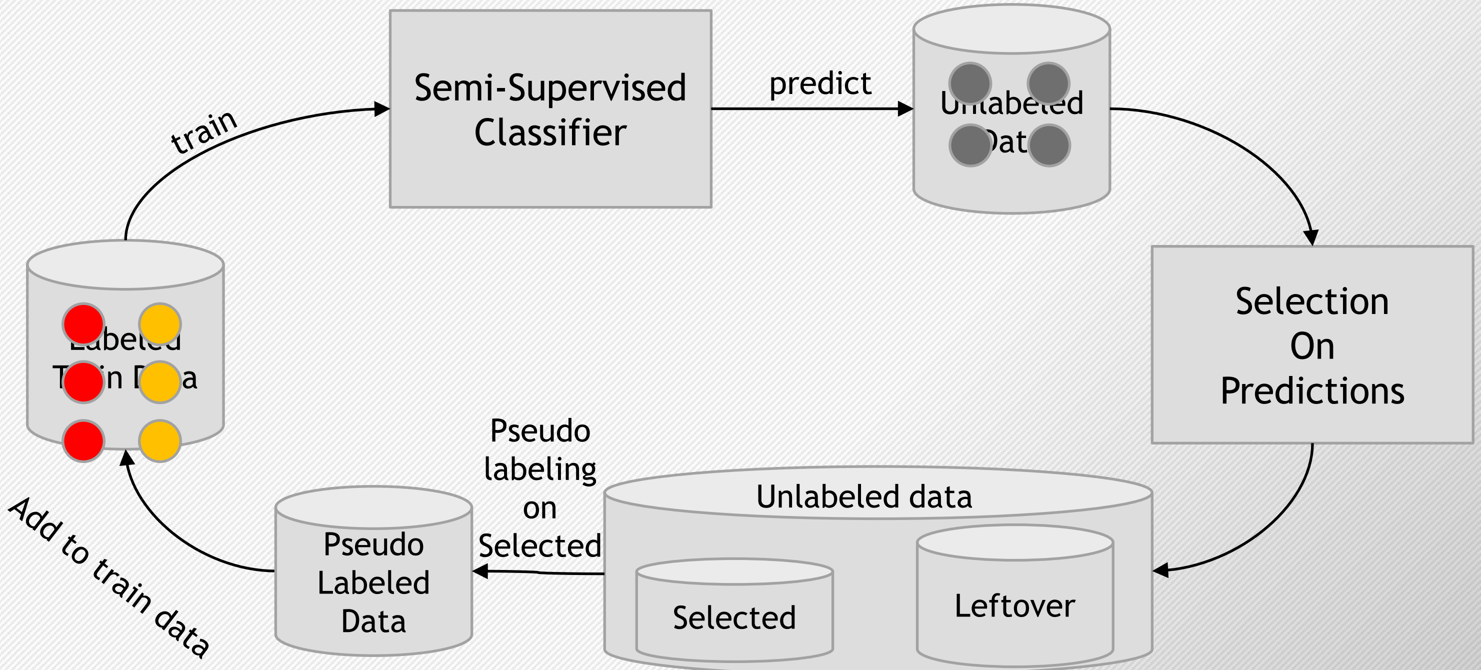
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

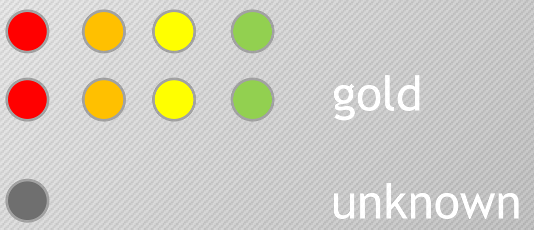




Solutions

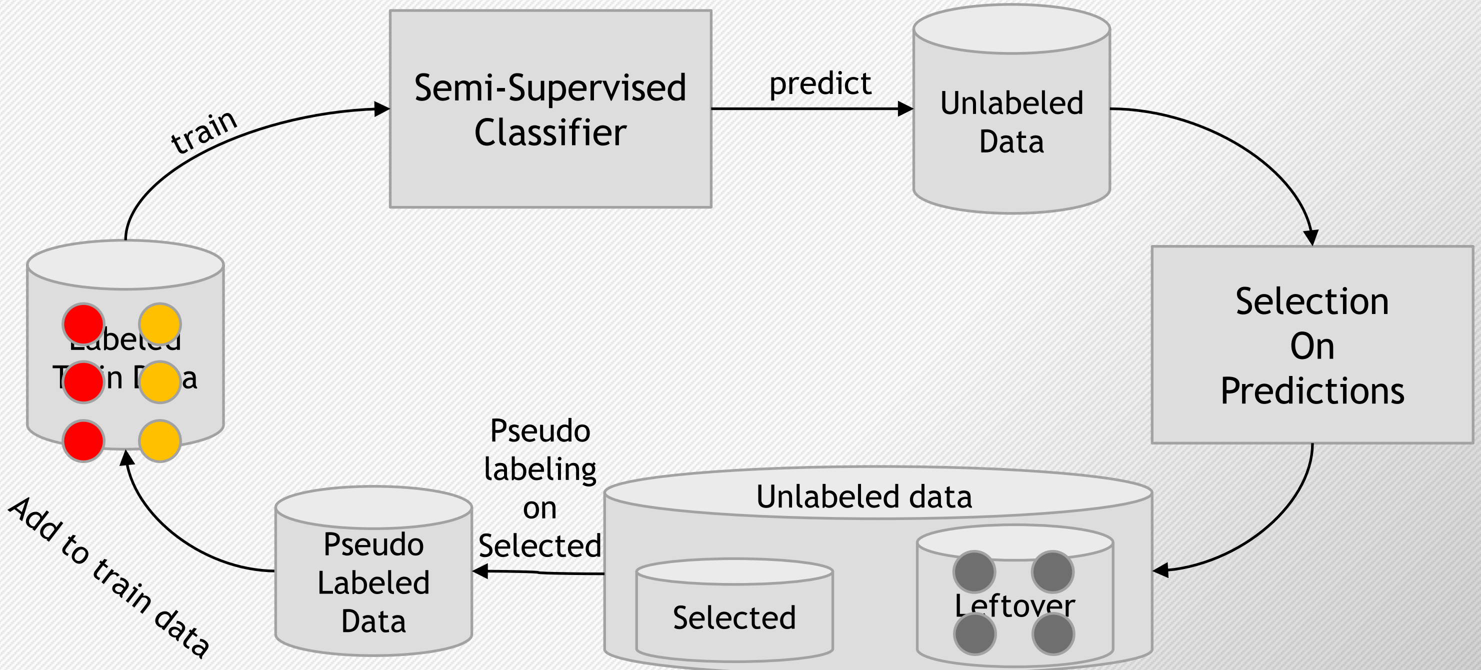
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

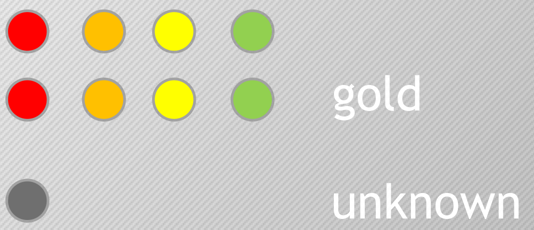




Solutions

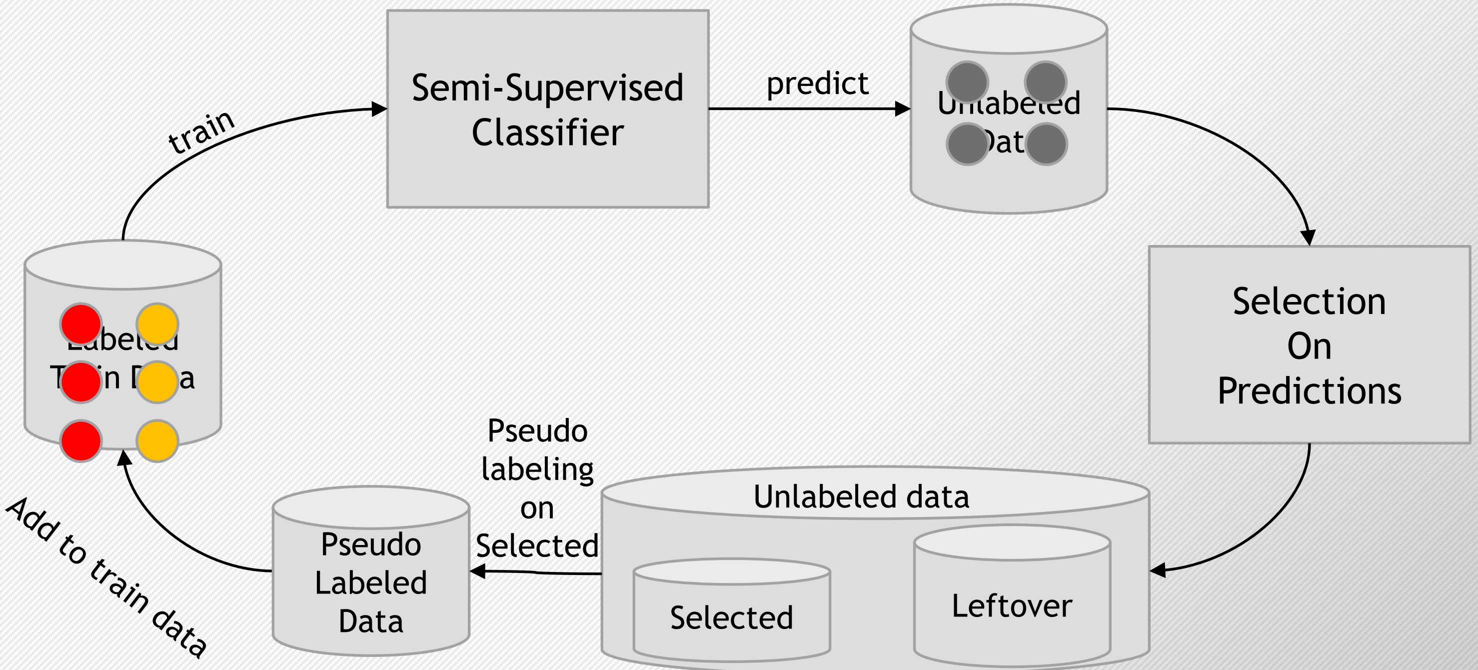
- Unknown Class Detection
  - Iterative Semi-Supervised Classification

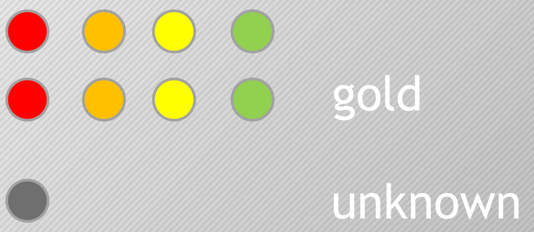




Solutions

- Unknown Class Detection
  - Iterative Semi-Supervised Classification





Problem

- Misunderstanding of SALNet

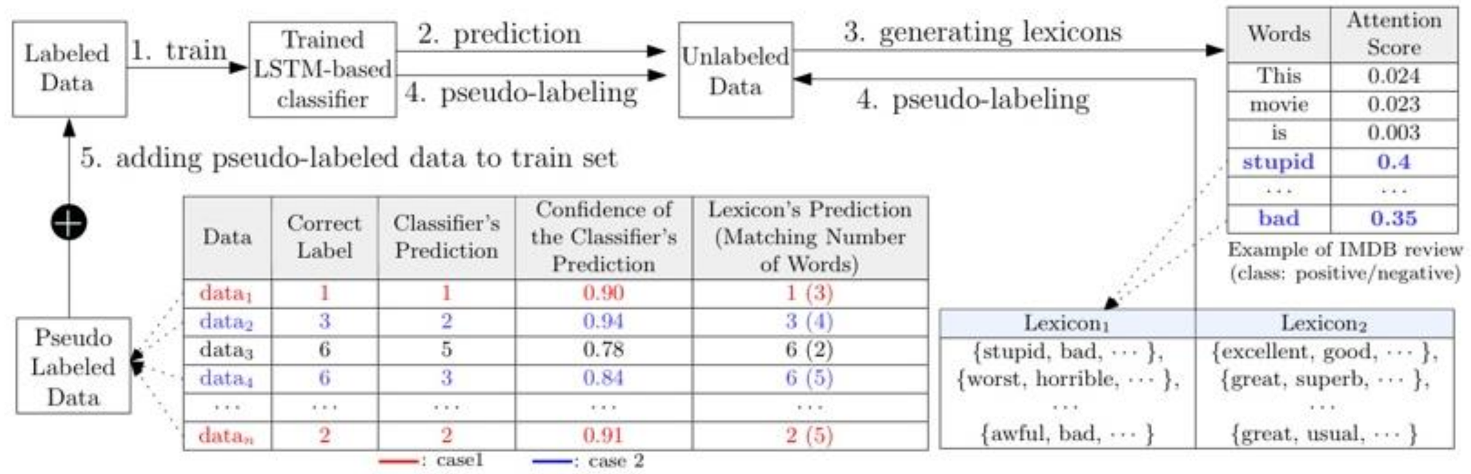


Figure 2: Our proposed method, using both attention-based classifier and lexicons. We set the t_1= 3 and t_2 = 4 in our method and set 0.9 as the threshold to verify the high confidence.

Problem

- Misunderstanding of SALNet

If we use more than two classifiers in SALNet, we can add the pseudo-labeled data by repeating the process of Case 1 and Case 2 for the additional classifiers. Once we obtain a new dataset after pseudo-labeling, the new dataset may have a different number of data in each class. This imbalance may make a classifier overfit to larger classes. We avoid this problem by selecting the same number of data from each class, which is the number of data in the smallest class, for the next training step.

Problem

- Misunderstanding of SALNet

	Class 0	Class 1	Class 2	Class 3	
Count	1000	800	400	100	
					-50
1	950	750	350	50	
					-30
1	920	720	320	20	
					-20
3	900	700	300	0	

Problem

- Misunderstanding of SALNet

	Class 0	Class 1	Class 2	Class 3
Count	1000	800	400	100
				-50
1	950	750	350	50
				-30
1	920	720	320	20
				-20
3	900	700	300	0

Solution Options

1. Give up SALNet?

Solution Options

1. Give up SALNet?

Iterative Process

Incremental Removal of KKC from { KKC + UUC }

Lexical Synthesis for Naming UUC, further contribution of Novelty Detection

A Unifying Review of Deep and Shallow Anomaly Detection IEEE 2021.2

This article deals with application of deep learning techniques to anomaly detection. Furthermore, connections between classic "shallow" and novel deep approaches are established, and it is shown how this relation might cross-fertilize or extend both directions.

By Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, *Member IEEE*, Marius Kloft, *Senior Member IEEE*, Thomas G. Dietterich, *Member IEEE*, and Klaus-Robert Müller, *Member IEEE*

First to approach Novelty Detection through iterative approach using Attention

Solution Options

1. Give up SALNet?

2. Removal of finished classes

Solution Options

1. Give up SALNet?

2. Removal of finished classes

	Class 0	Class 1	Class 2	~~Class 3~~
Count	1000	800	400	~~100~~
1	950	750	350	~~50~~ (-50)
1	920	720	320	~~20~~ (-30)
3	900	700	300	~~0~~ (-20)

Solution Options

1. Give up SALNet?

2. Removal of finished classes

	Class 0	Class 1	Class 2
Count	1000	800	400
1	950	750	350
1	920	720	320
3	900	700	300

Solution Options

1. Give up SALNet?

2. Removal of finished classes

	Class 0	Class 1	Class 2
Count	900	700	300
1	-	-	-
1	-	-	-
3	-	-	-

Solution Options

1. Give up SALNet?

2. Removal of finished classes

	Class 0	Class 1	Class 2
Count	900	700	300
			-200
1	700	500	100
			-70
1	630	430	30
			-30
3	600	400	0

Solution Options

1. Give up SALNet?
2. Removal of finished classes

	Class 0	Class 1	Class 2
Count	900	700	300
			-200
1	700	500	100
			-70
1	630	430	30
			-30
3	600	400	0

Solution Options

1. Give up SALNet?

2. Removal of finished classes

3. Proportional Selection

Solution Options

1. Give up SALNet?
2. Removal of finished classes
3. Proportional Selection

	Class 0	Class 1	Class 2	Class 3
Count	1000	800	400	100
				-50
1	950	750	350	50
				-30
1	920	720	320	20
				-20
3	900	700	300	0

Solution Options

1. Give up SALNet?
2. Removal of finished classes
3. Proportional Selection

	Class 0	Class 1	Class 2	Class 3
Count	1000	800	400	100
		10:8:4:1		
1	700	560	280	70
1	500	400	200	50
3	350	280	140	35

Up to this point...

1. ~~Give up SALNet?~~

2. **Removal of finished classes**

3. ~~Proportional Selection~~

Class Imbalance & Accuracy

Inaccurate Lexicons

HockeyTeam, RugbyClub, Basketballteam, CricketTeam, CyclingTeam

sports hockey team	hockey games team league
hockey team champions cup	hockey team ice championship
hockey team arena	teams hockey games club
hockey team ice torino	sponsored trophy team cricket
rugby sevens team	matches club cricket
teams league captained cricket	teams team tournament cricket
team pakistan cricket	league team premier cricket
league premier sporting cricket	rugby club tournament
rugby union tournament cup	rugby union tournament
t20 stadium team cup	stadium pakistan team t20
coach matches team cricket	games: team zealand cricket
cup team cricket	bangladesh cricketers bowling
team england cricket	t20 twenty20 cup
stadium twenty20 cricket	cricketer league team cricket
tournaments club cricket	rugby league cricket
league team finals cricket	rugby team league
lanka league team cricket	football league club cricket

Up to this point…

~~1. Give up SALNet?~~

2. Removal of finished classes

~~3. Proportional Selection~~

Class Imbalance & Accuracy

Inaccurate Lexicons

HockeyTeam, RugbyClub, Basketballteam, CricketTeam, CyclingTeam

basketball team league
star) basketball club blue
basketball team club
basketball in club championship
basketball league club championship
basketball multi-sports league
basketball womens club based
competitions cup handball
basketball developmental team league
dukla basketball club
basketball sponsorship) club
team club basketball)
basketball men club
league) basketball thunders arena
sports kabaddi team league
liga games play
cup champions league

basketball club) league
(serbian basketball team
montenegrin basketball womens team
basketball championship club league
basketball club league
basketball cup league
kosovan basketball team
spelling: basketball club
basketball monferrato team pallacanestro
league) basketball samsung yongin
sport colours club
basketball team fiba
basketball sponsorship clubs club
mursa basketball womens club
amateur club (sports
kabaddi stadium team league
flags union players national

Future Work

- Removal of Finished Classes
  - What is the most optimal Lexicon? At which iteration should the optimal Lexicon be created?
    - Optimal Lexicon can be created from known documents.
      (if trained data is stacked enough)
  - How do we know ‘no more remaining known documents’ in test-case?
    - Confidence and attention weights of Lexicons