





# Actively Blocking Scheme Learning for Entity Resolution

Jingyu Shao and Qing Wang

Research School of Computer Science  
The Australian National University

# Entity Resolution

- Refers to the process of identifying **records which represent the same real-world entity** from one or more datasets.

		Amazon.com					
#	Name	Color	Storage	Size			
	$r_1$	Apple iPhone X	Silver	64GB	5.8"		
	$r_2$	Apple iPhone X	Space Gray	256 GB	5.8"		
	$r_3$	iPhone X	Space Gray	256 GB	5.8"		
	$r_4$	Apple iPhone X	Space Gray	256 GB	-		
	$r_5$	Samsung Galaxy S9 Plus	Midnight Black	64GB	6.2"		
	$r_6$	Galaxy S9 Plus	Black	64GB	-		
	$r_7$	Galaxy S9 Plus	Coral Blue	-	-		

# Blocking

- Commonly applied to improve time efficiency in the ER process by **grouping potentially matched records** into the same block.
- Considering a dataset of 1,000 records:

Without Blocking:  
500,000 pairwise comparisons

With Blocking:  
≤ 50,000 pairwise comparisons,  
if the largest block contains  
100 records

- Using blocking schemes: (Which is better?)



How to learn a good blocking scheme?

# Related Work

- Using blocking schemes from:
  - i. Domain Expert: The scheme is assigned based on **experience**.
  - ii. Supervised Learning: Large numbers of records with **labels** are need for quality guarantee, which is hard to achieve in entity resolution.
  - iii. Unsupervised Learning: No quality guaranteed because pairwise records are often labeled in terms of their **syntactic similarity**.
- Limitation: Existing work on learning may either use a large number of labels or the blocking quality is hard to guarantee.



Active Learning: We aim to select biased samples in order to efficiently use labels with quality guaranteed.

# Two Challenges

- **Class Imbalance Problem**

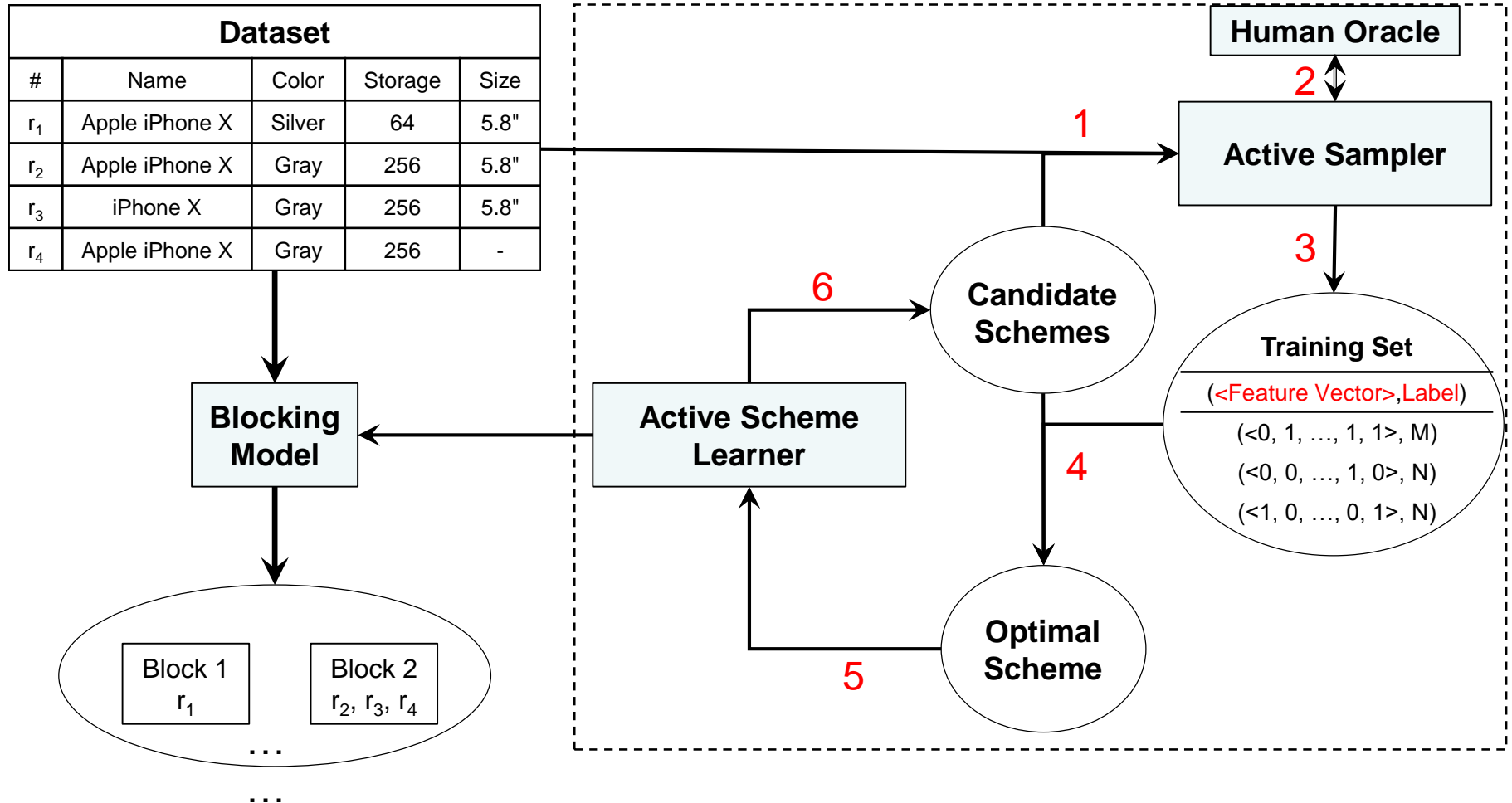
In blocking scheme learning, both matched and non-matched samples are necessary. However, if samples are selected randomly, there are usually much more non-matches than matches. To achieve enough matched samples, large numbers of labels are needed in existing work. We proposed **Active Sampling** to tackle it.

- **Large Search Space**

Searching all possible schemes is a heavy work. Existing work on reducing the search space, such as ranking attributes as part of a scheme by ad-hoc method, is not reliable.

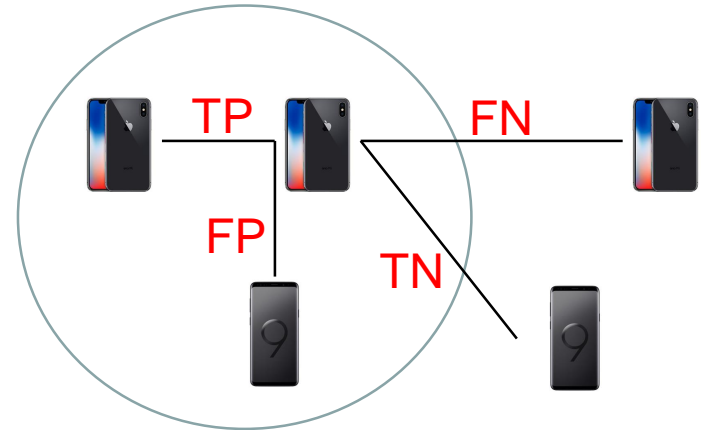
We proposed **Active Branching** to tackle it.

# Framework



# Preliminaries

- True/False positive/negative



- A **predicate**  $\langle a_k, h_{a_k} \rangle$  is associated with an attribute and a blocking function  
E.g.  $\langle \text{Name}, \text{Soundex} \rangle$
- A **blocking scheme** is a disjunction of conjunctions of predicates  
E.g.  $(\langle \text{Name}, \text{Soundex} \rangle \wedge \langle \text{Color}, \text{Exact} \rangle) \vee (\langle \text{Name}, \text{Soundex} \rangle \wedge \langle \text{Storage}, \text{Exact} \rangle)$

# Problem Definition

- A good blocking scheme should: yield blocks that contain **minimum number of FPs and FNs**.
- Given a human oracle  $\xi$ , and an error rate  $\varepsilon \in [0, 1]$ , the **active blocking problem** is to learn a blocking scheme  $s$  in terms of the following objective function, through actively selecting a training set  $T$

$$\begin{aligned} & \text{minimize} \quad |fp(B_s)| \\ & \text{subject to} \quad \frac{|fn(B_s)|}{|tp(B_s)|} \leq \varepsilon, \text{ and } |T| \leq \text{budget}(\xi) \end{aligned}$$

where  $B_s$  is the blocks generated by blocking scheme  $s$



# Active Sampling

- Based on our observation and the notification of some related work, we assume that a similar record pair is more likely to be a match than a dissimilar record pair.
- We define **Balance Rate**  $\gamma(s, X)$  to describe the sample distribution in the feature vector set  $X$  under scheme  $s$ . (Details of function can be found in the paper)

E.g. if  $s = p_1 \wedge p_2$ , then  $s(x_1) = \text{true}$ ,  $s(x_2) = \text{false}$

$$\gamma(s, X) = \frac{|\{x_i \in X | s(x_i) = \text{true}\}| - |\{x_i \in X | s(x_i) = \text{false}\}|}{|X|} = 0$$

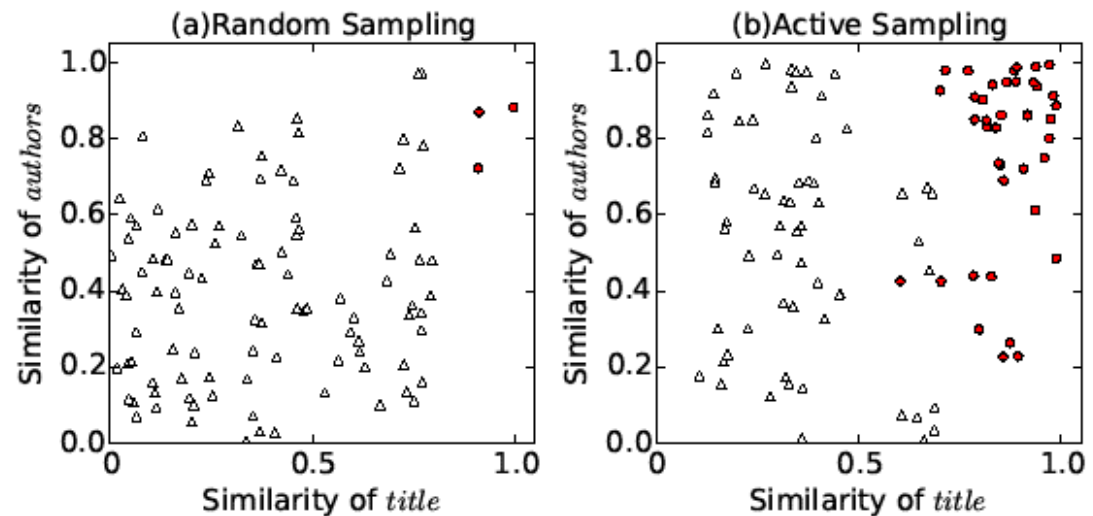
	$p_1$	$p_2$	$p_3$	$p_4$
$x_1$	1	1	0	1
$x_2$	0	1	0	1

# Active Sampling

We tackle the class imbalance problem by transferring it into the **Balanced Sampling Problem**. It selects a set of feature vectors that can minimize the balance rate for all candidate schemes:

$$\text{minimize } \sum_{s_i \in S} \gamma(s_i, X)^2$$

E.g. Sampling examples in terms of *Cora* dataset



# Active Branching

- Active Branching aims to reduce the search space by extending the existing scheme.
- The total number of possible blocking schemes is known as **Dedekind Number**, which is  $2^{\binom{n}{[n/2]}}$ ,  $n$  is the number of predicates.

E.g. The number of possible blocking schemes of a dataset with 5 attributes, each attribute is associated with 4 blocking functions can be  $2^{\binom{20}{10}}$ .

- Our strategy can reduce it to at most  $n^2$ .

# Lemmas for Active Branching

- To minimize fp by conjunction:

$$|fp(B_{S_i})| \geq |fp(B_{S_1 \wedge S_2})|, \text{ where } i = 1, 2$$

Applied when  $\frac{|fn(B_{S_i})|}{|tp(B_{S_i})|} \leq \varepsilon$  holds

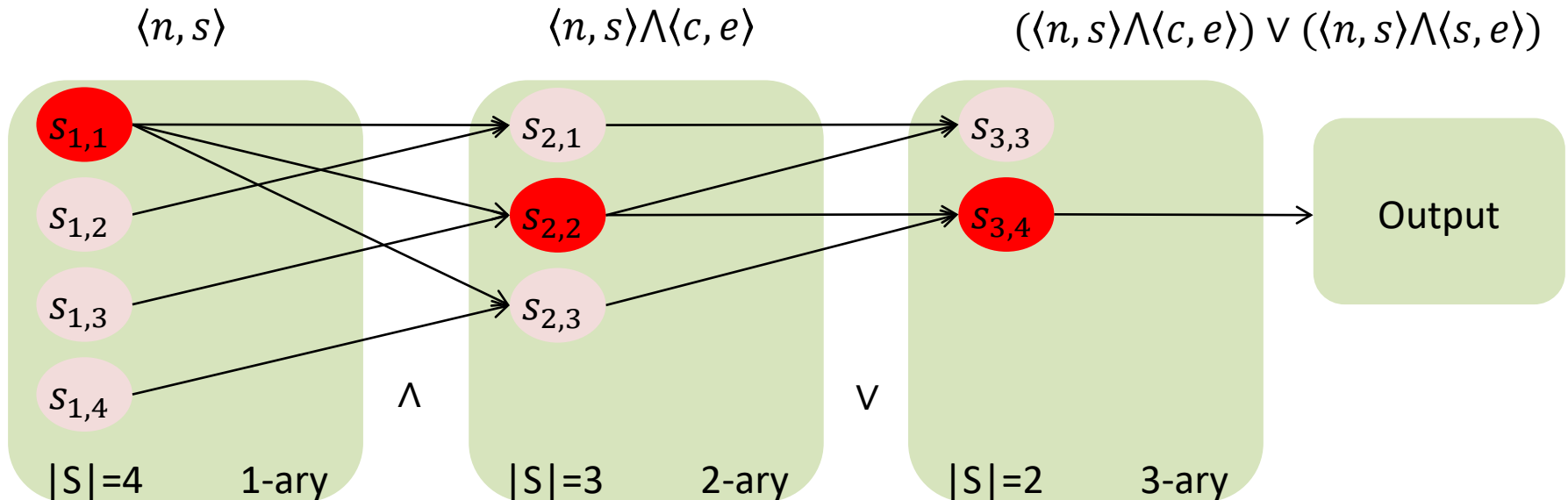
- To reduce the rate of fn and tp by disjunction:

$$\frac{|fn(B_{S_i})|}{|tp(B_{S_i})|} \geq \frac{|fn(B_{S_1 \vee S_2})|}{|tp(B_{S_1 \vee S_2})|}, \text{ where } i = 1, 2$$

Applied when  $\frac{|fn(B_{S_i})|}{|tp(B_{S_i})|} \leq \varepsilon$  does not hold

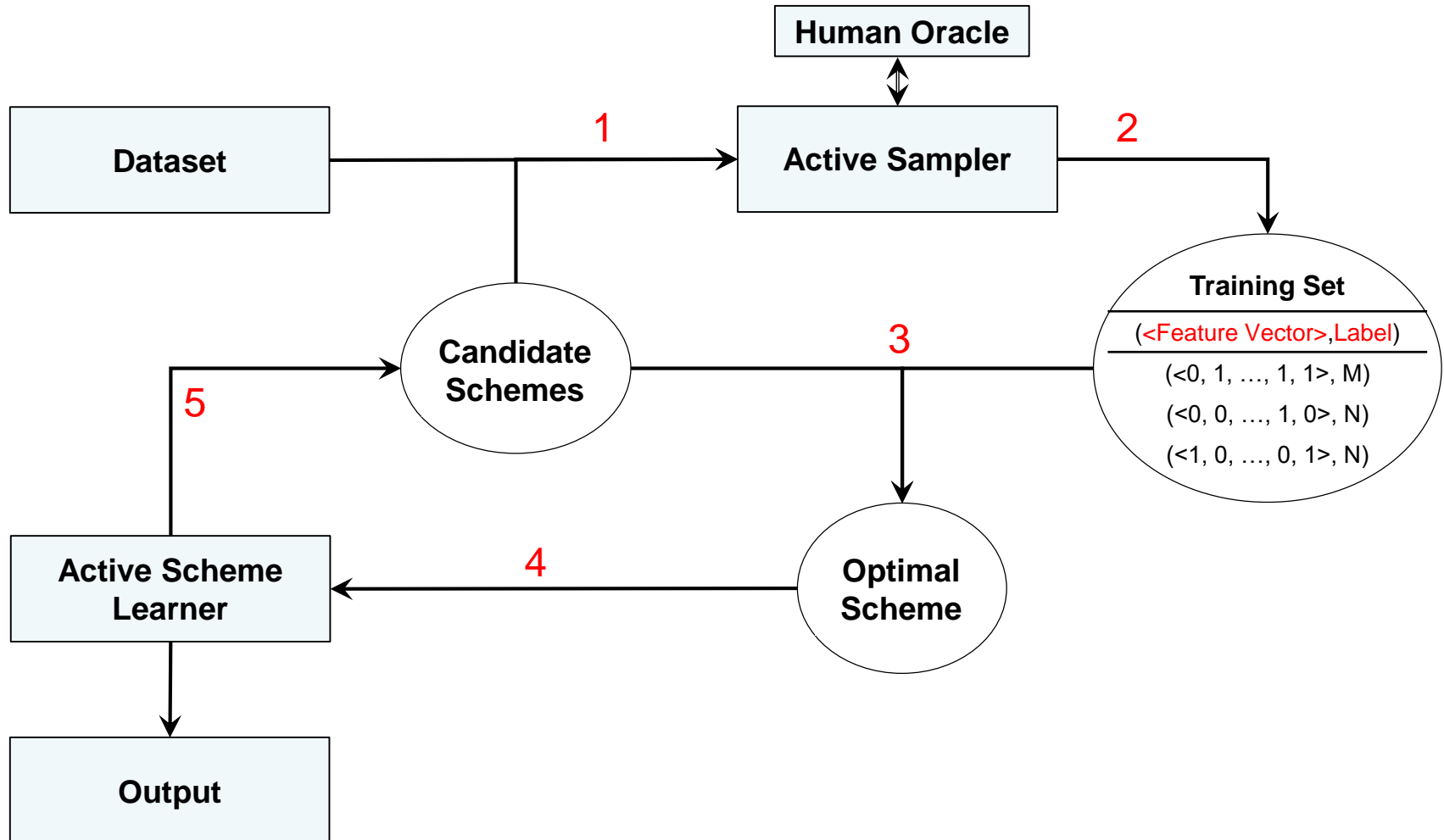
# Active Branching Example

Given four predicates as initial schemes:  $\langle \mathbf{name}, \mathbf{soundex} \rangle$ ,  $\langle \mathbf{name}, \mathbf{exact} \rangle$ ,  $\langle \mathbf{color}, \mathbf{exact} \rangle$ ,  $\langle \mathbf{storage}, \mathbf{exact} \rangle$



Terminate: when budget is used out or all the predicates are contained in  $s$

# Approach Presentation



# Experimental Setup

- Datasets

Dataset	# of Attributes	# of Records	Class Imbalance Ratio
Cora	4	1,295	1:49
DBLP-Scholar	4	2,616/64,263	1:31,440
DBLP-ACM	4	2,616/2,294	1:1,117
NCVR	18	267,716/278,262	1:2,692

- Baselines

ASL	Fisher	TBlo	RSL
Our approach	The state-of-the-art unsupervised approach	The schemes are assigned by domain experts	It is similar to ASL but uses random sampling strategy

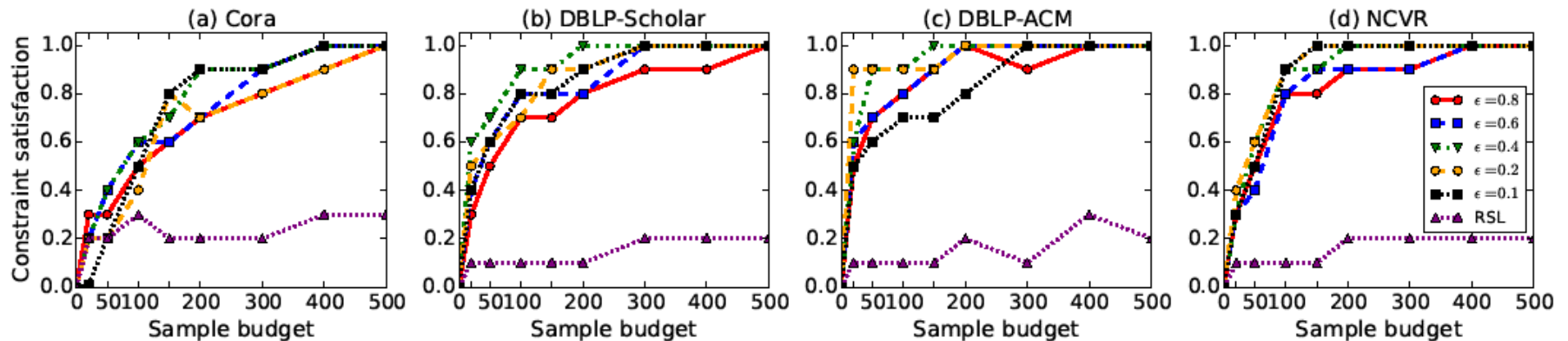
# Measures

Reduction Ratio (RR)	$RR = \frac{tp + fp}{tp + fp + tn + fn}$
Pair Completeness (PC)	$PC = \frac{tp}{tp + fn}$
Pair Quality (PQ)	$PQ = \frac{tp}{tp + fp}$
F-Measure (FM)	$FM = \frac{2 \times PC \times PQ}{PC + PQ}$
Constraint Satisfaction (CS)	$CS = \frac{N_s}{N} \times 100\%$ <p>Where <math>N_s</math> is the number of times to have <math>s</math> as output, <math>N</math> is the number of times to run the algorithm</p>



# Constraint Satisfaction

This experiment aims to evaluate the performance under different error rates and label budgets. The higher CS is, the more stable and consistent results it learns.



Conclusion: our algorithm can learn stable blocking schemes with a budget of 500 in most cases.

# Label Cost

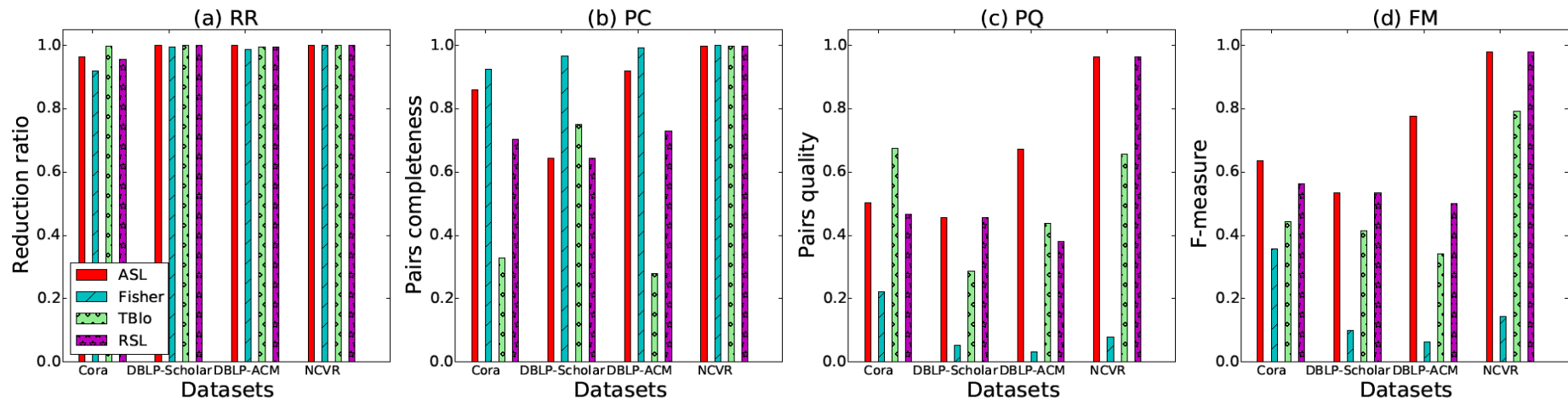
This experiment aims to evaluate the label cost comparing with random sampling under different error rates

Error Rate	Cora	DBLP-Scholar	DBLP-ACM	NCVR
0.8	600	500	300	300
0.6	400	350	200	350
0.4	450	250	150	250
0.2	550	300	200	200
0.1	500	250	300	250
RSL	8,000	10,000+	2,500	10,000+

Conclusion: Our approach uses significantly less labels than random sampling.

# Blocking Quality

This experiment gives an overview performance on blocking quality comparing with baselines in terms of RR, PC, PQ and FM.



Conclusion: all approaches can generate high RR blocks. Furthermore, our approach can learn schemes that generates highest FM blocks, i.e. both high PC and PQ.

# Blocking Efficiency

This experiment aims to evaluate the blocking efficiency comparing with baselines by evaluating the number of record pairs generated by different approaches.

	TBlo	Fisher	ASL	RSL
Cora	2,945	67,290	29,306	17,974
DBLP-Scholar	6,163	1,039,242	3,328	3,328
DBLP-ACM	25,279	69,037	3,043	17,446
NCVR	932,239	7,902,910	634,121	634,121

Conclusion: Our approach can conduct concise blocks in most cases except Cora, where TBlo is better but it discards more matches.

# Conclusion

- We propose an approach which uses active learning techniques for blocking scheme learning to reduce the label cost with quality guarantee.
- Two strategies are used in our approach:
  - i. Active Sampling to reduce the label cost
  - ii. Active Branching to reduce the search space
- Experimental results show that the proposed approach can highly reduce the label cost while outperform the baselines with a trade-off of PC and PQ.



# Thank you!

## Q & A

[Jingyu.shao@anu.edu.au](mailto:Jingyu.shao@anu.edu.au)