Entity Resolution with Active Learning

Jingyu Shao

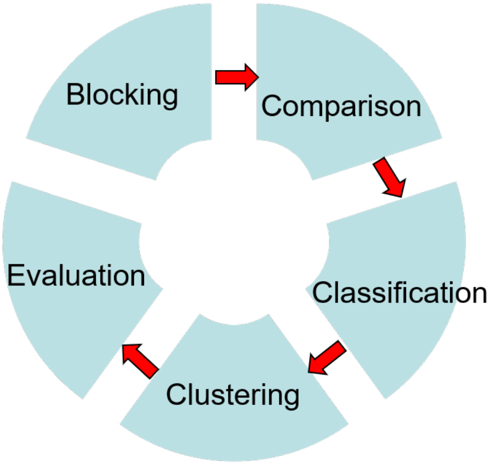
School of Computing
Australian National University

May 24, 2021

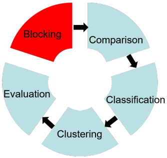
Agenda

* Introduction and challenges
* How to build a set of "optimal" blocking schemes efficiently?
* How to design an AL approach under various data distributions?
* How to alleviate the overfitting problem for powerful models?
* Conclusion

The process of identifying records which represent the same real-world entity from one or more datasets

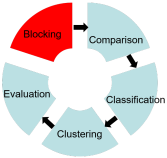


Blocking



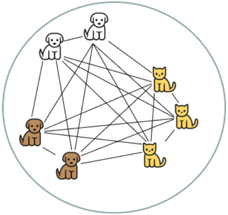
Reduce the number of record pairs to be compared by grouping potentially matched records into the same block.
E.g., millions of pairs in real life.

Blocking

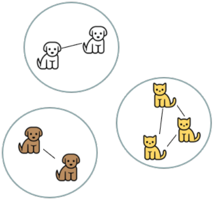


Reduce the number of record pairs to be compared by grouping potentially matched records into the same block.
E.g., millions of pairs in real life.

Without blocking:
7 records with 21 pairs



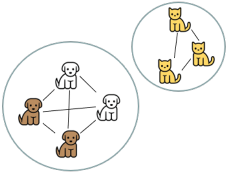
With blocking:
7 records with 5 pairs



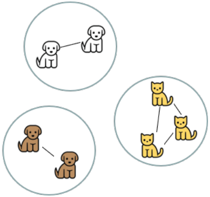
Blocking Schemes

Using blocking schemes: (Which is better?)

Name



Color

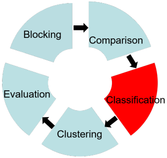


Blocking Schemes

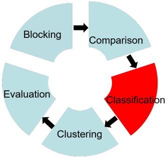
Using blocking schemes: (Which is better?)

How to learn a good blocking scheme?

– Millions of record pairs, with highly imbalanced labels hard to obtain.
– The search space for all possible blocking schemes is large.

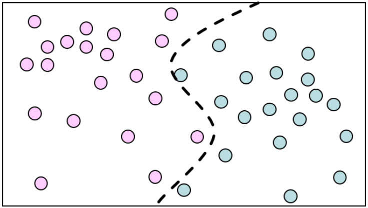


A classifier is used to categorize samples into matches and non-matches.



A classifier is used to categorize samples into matches and non-matches.

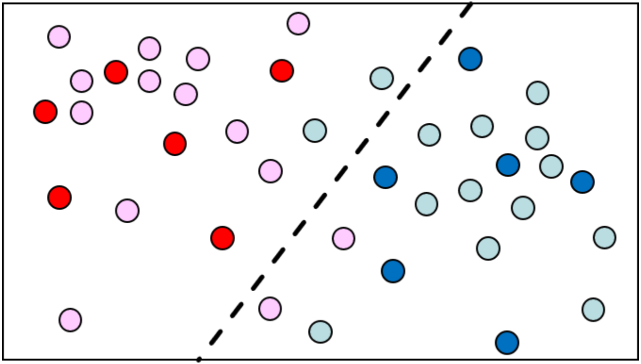
Considering we have samples within a block, and they are mapped into a feature space shown as below:



The red and blue points refer to matches and non-matches

Training a Classifier

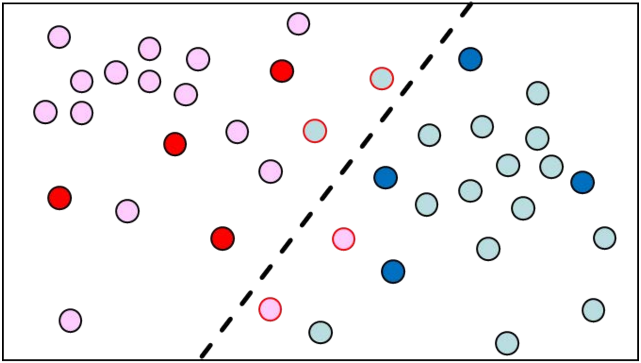
Sufficient number of samples are necessary for training, but obtaining their labels for learning is costly.



Accuracy: 90% * Red: 6/20 * Blue: 6/20

Initialization: random seed samples

Select the most uncertain instances

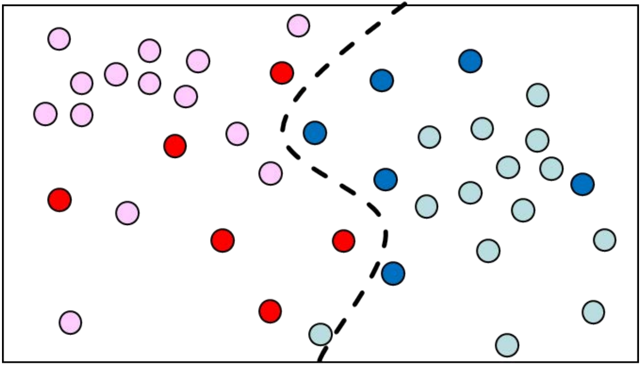


Accuracy: 90%   * Red: 4/20   * Blue: 4/20 [1]

[1] B. Settles, Active learning literature survey, 2010

Australian National University

4 more samples are labeled



Accuracy: 97.5%

* Red: 6/20

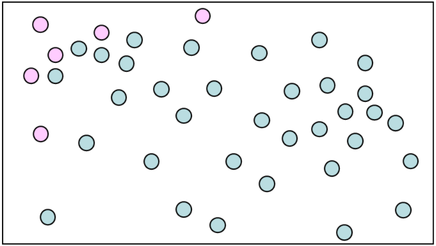
* Blue: 6/20

Challenges

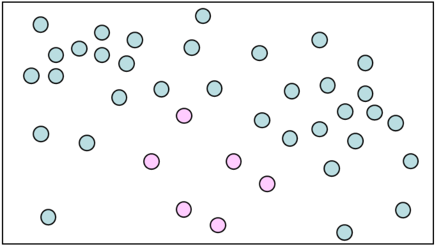
The distribution of matches and non-matches is highly imbalanced.

A small number of samples are labeled.

- Various strategies: different datasets



Sample distribution



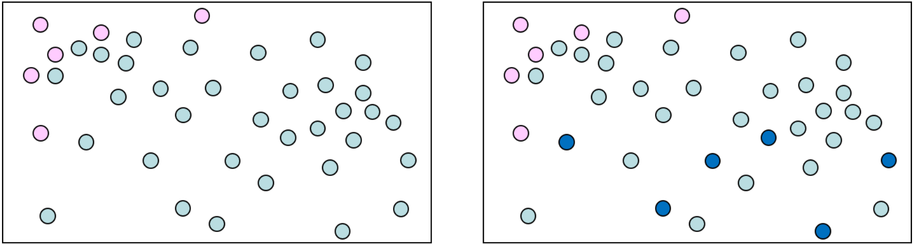
Sample distribution

Challenges

The distribution of matches and non-matches is highly imbalanced.

A small number of samples are labeled.

- Various strategies: different datasets
- Cold start: imbalanced ER data distribution



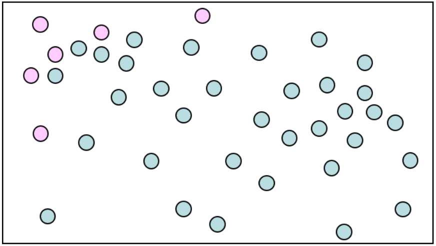
Sample distribution Cold Start

Challenges

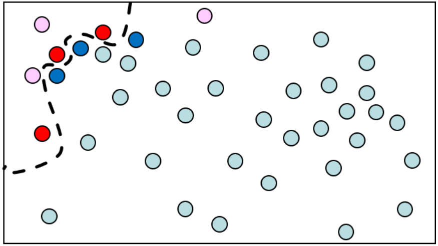
The distribution of matches and non-matches is highly imbalanced.

A small number of samples are labeled.

- Various strategies: different datasets
- Cold start: imbalanced ER data distribution
- Overfitting: powerful models



Sample distribution



Overfitting

Agenda

Australian National University

* Introduction and challenges
* How to build a set of "optimal" blocking schemes efficiently?
  – Active scheme learning and scheme skyline learning [1][2]
* How to design an AL approach under various data distributions?
* How to alleviate the overfitting problem for powerful models?
* Conclusion

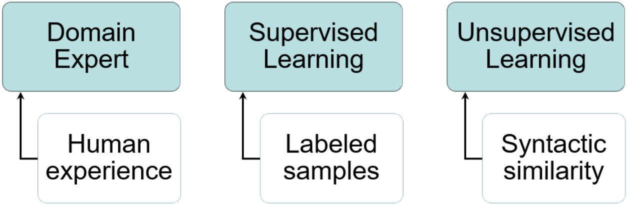
[1] J. Shao and Q. Wang. Active Blocking Scheme Learning for Entity Resolution. PAKDD'18.

[2] J. Shao, Q. Wang and Y. Lin. Skyblocking for Entity Resolution. IS'19.

Schemes for Blocking

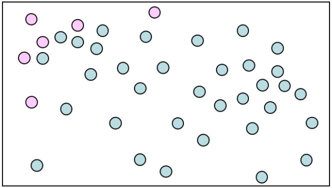
Disjunction of conjunction of attributes

Blocking schemes are built from:



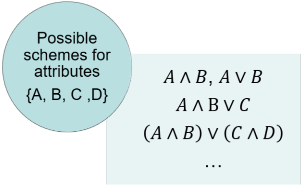
Challenges for Scheme Learning

Class Imbalance Problem:



Sample distribution

Large Search Space $2^{\binom{n}{\lfloor n/2 \rfloor}}$:



Balanced Sample Selection

Our observation: similar attribute values $->$ matches

How to select attributes to build schemes?
Some values are frequent but useless, e.g. year.
Balanced samples for all possible attributes and select!

Balance Rate $\gamma(s, X)$ describes the balance degree for a given scheme s under a sample set X

Balanced Sample Selection

Our observation: similar attribute values $->$ matches

How to select attributes to build schemes?
Some values are frequent but useless, e.g. year.
Balanced samples for all possible attributes and select!

Balance Rate $\gamma(s, X)$ describes the balance degree for a given scheme s under a sample set X

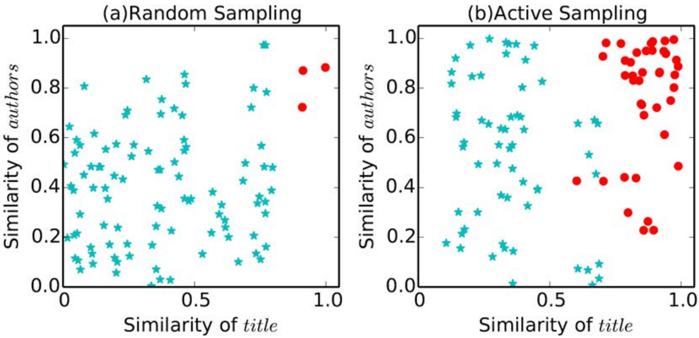
E.g. if $s = A \wedge B$, $X = \{x_1, x_2\}$, then $s(x_1) = true$, $s(x_2) = false$
Thus $\gamma(s, X) = \frac{1(\#true) - 1(\#false)}{2} = 0$ (balanced):

	A	B	C	D
x_1	1	1	0	1
x_2	0	1	0	1

Active Sampling

Select samples to minimize the balance rate for a given set of schemes:

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



Active Branching

Reduce the search space by extending “proper” schemes w.r.t. a specific criterion, e.g. Pair Completeness (Recall) and Pair Quality (Precision).

- $\wedge$ reduce block size, increase PQ
- $\vee$ increase block size, increase PC

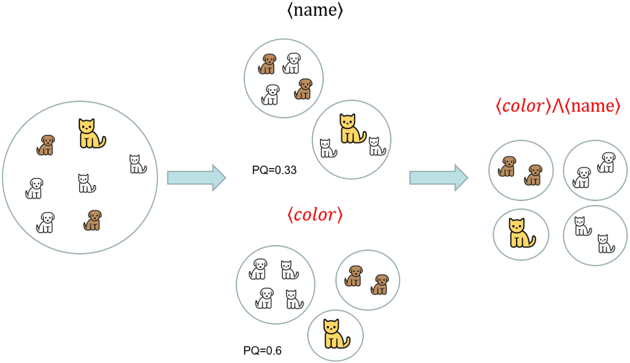
Active Branching

Reduce the search space by extending "proper" schemes w.r.t. a specific criterion, e.g. Pair Completeness (Recall) and Pair Quality (Precision).

$\wedge$ reduce block size, increase PQ

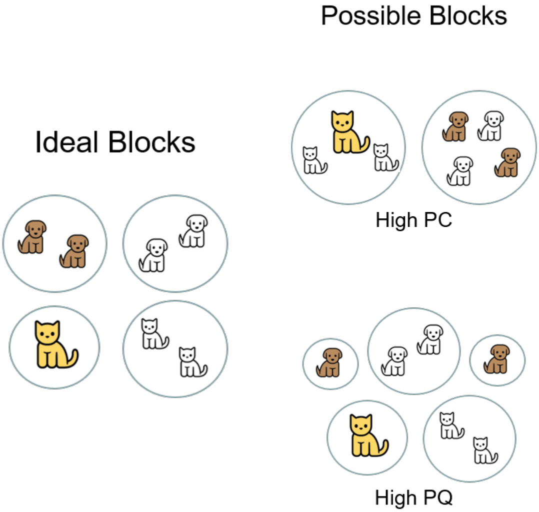
$\vee$ increase block size, increase PC

Example: to learn a blocking scheme with two attributes: $\langle name \rangle$, $\langle color \rangle$, w.r.t. $PQ = 0.8$



Limitation: Trade-offs

- The higher, the better (PC and PQ values)
- More records in one block (high PC threshold)
- Less records in one block (high PQ threshold

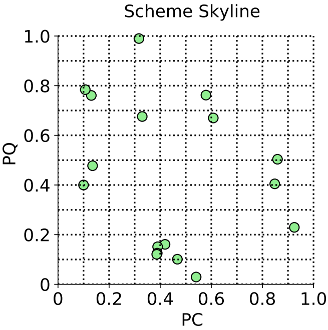


Skyblocking Approach

Skyline queries under a set of blocking schemes:

Map schemes into a measure space

Blocking scheme	PC	PQ
s_1	0.13	0.76
s_2	0.31	0.99
s_3	0.58	0.76
s_4	0.84	0.40
s_5	0.86	0.50
...	...	...

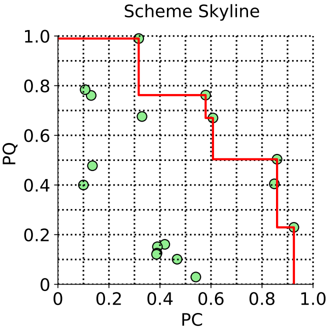


Skyblocking Approach

Skyline queries under a set of blocking schemes:

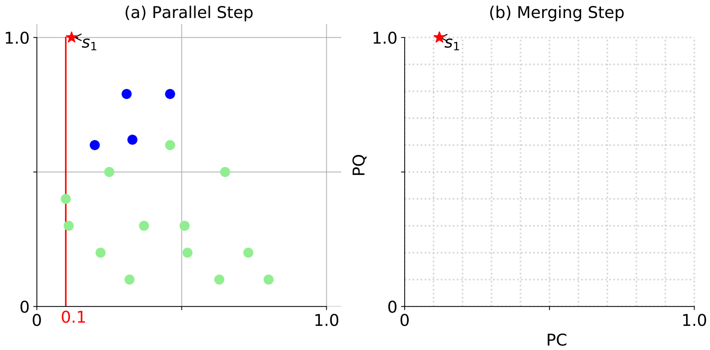
Dominated VS Dominating schemes

Blocking scheme	PC	PQ
s_1	0.13	0.76
s_2	0.31	0.99
s_3	0.58	0.76
s_4	0.84	0.40
s_5	0.86	0.50
...	...	...



A naïve way to learn scheme skyline:

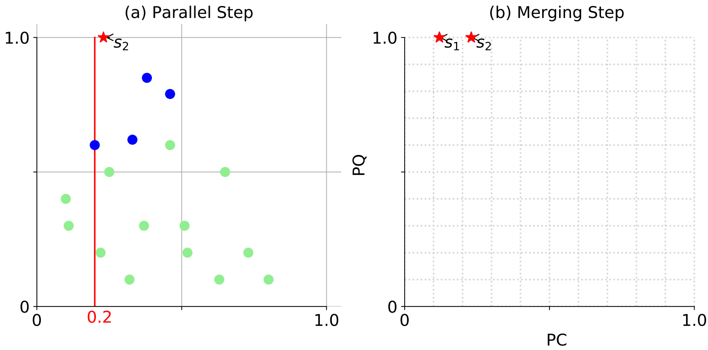
- Learn "optimal" schemes w.r.t. different thresholds



Naïve Skyline Learning (Naïve-Sky)

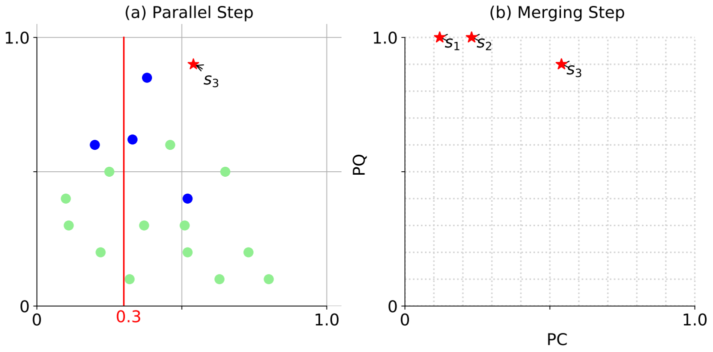
A naïve way to learn scheme skyline:

- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



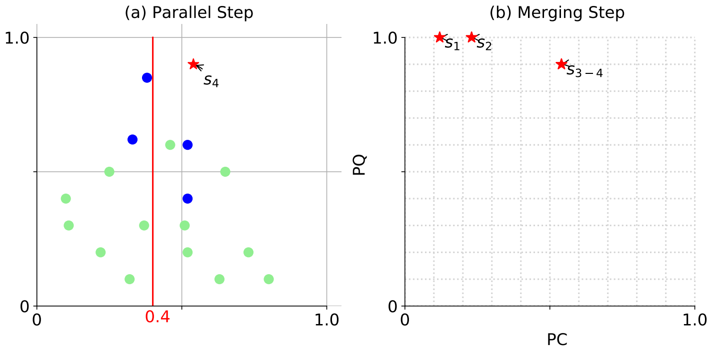
A naïve way to learn scheme skyline:

- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



A naïve way to learn scheme skyline:

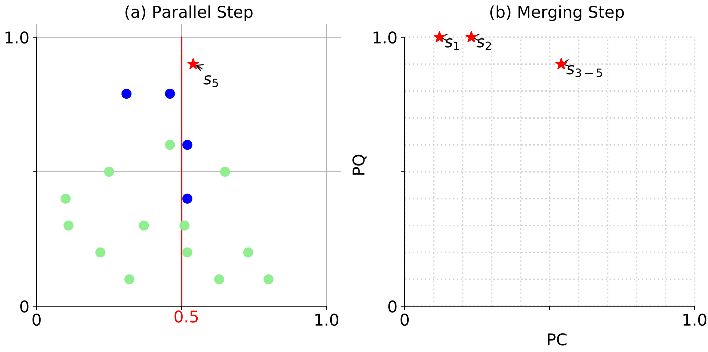
- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



Naïve Skyline Learning (Naïve-Sky)

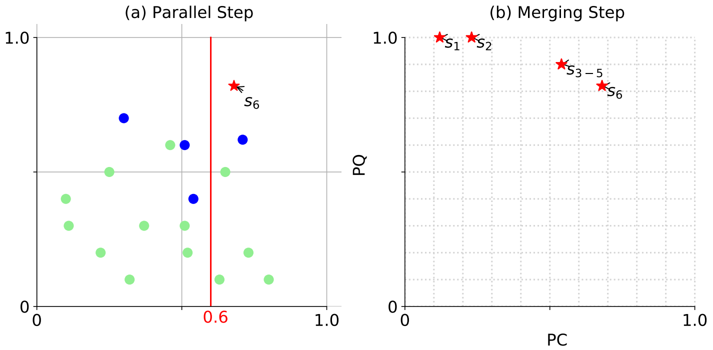
A naïve way to learn scheme skyline:

- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



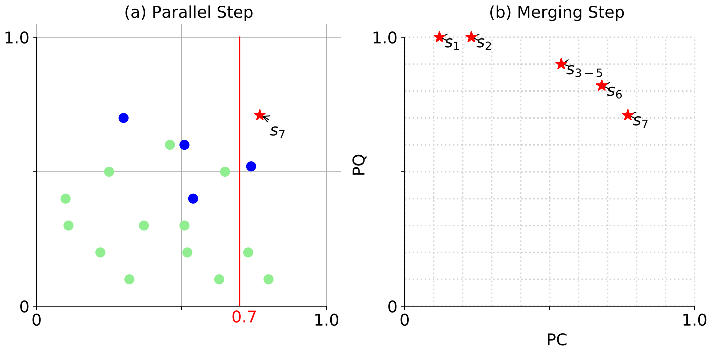
A naïve way to learn scheme skyline:

- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



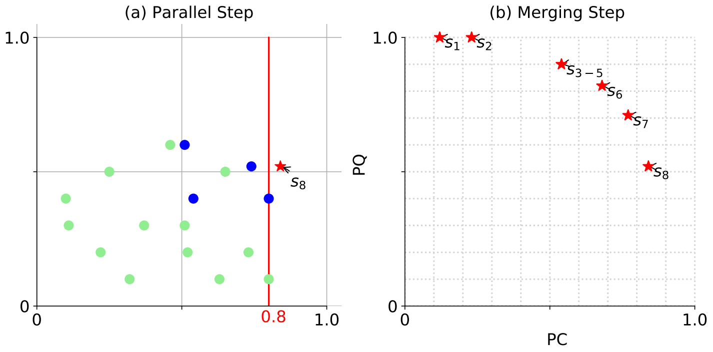
A naïve way to learn scheme skyline:

- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



A naïve way to learn scheme skyline:

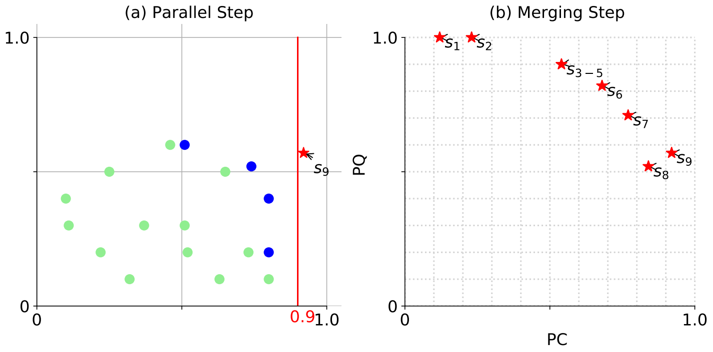
- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



Naïve Skyline Learning (Naïve-Sky)

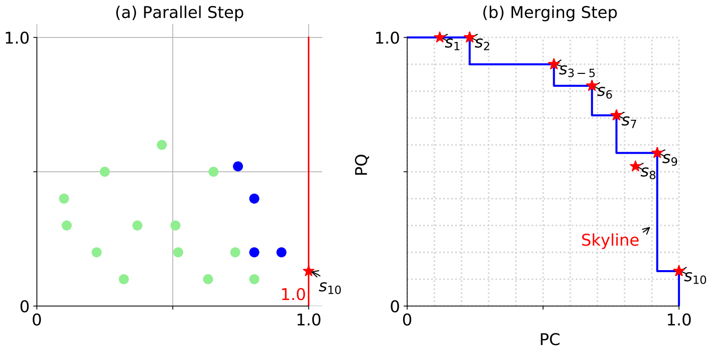
A naïve way to learn scheme skyline:

- New threshold: current one plus a threshold interval, e.g. $\Delta = 0.1$



A naïve way to learn scheme skyline:

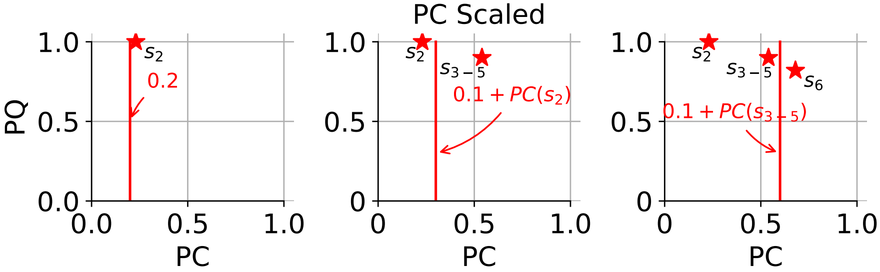
- Merge them for skyline



Adaptive Skyline Learning (Adap-Sky)

Observation: some are redundant under different thresholds: e.g. s_{3-5}

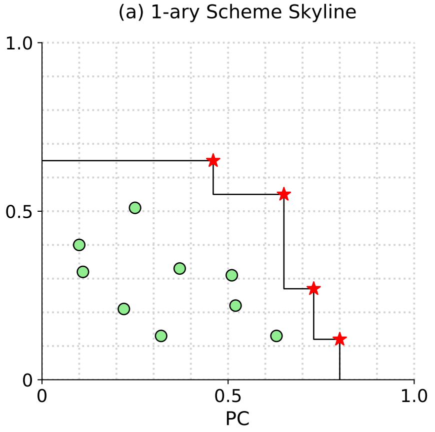
New threshold: PC/PQ value of current scheme plus a threshold interval



Progressive Skyline Learning (Pro-Sky)

Unnecessary label cost in Adap-Sky: samples are independently selected and may be duplicated under different thresholds.

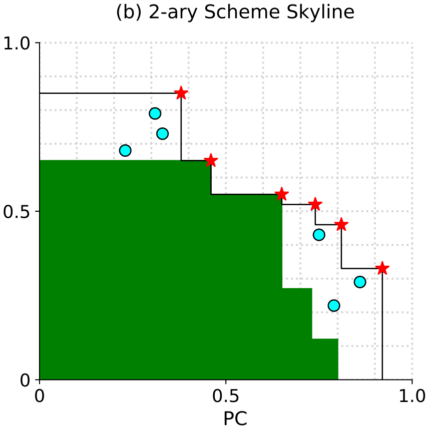
Pro-Sky with scheme extension:



Progressive Skyline Learning (Pro-Sky)

Unnecessary label cost in Adap-Sky: samples are independently selected and may be duplicated under different thresholds.

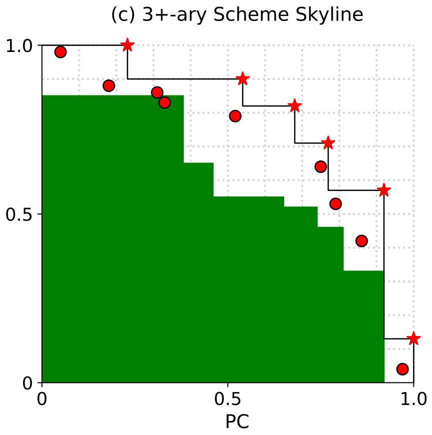
Pro-Sky with scheme extension:



Progressive Skyline Learning (Pro-Sky)

Unnecessary label cost in Adap-Sky: samples are independently selected and may be duplicated under different thresholds.

Pro-Sky with scheme extension:



Experimental Setup

Datasets

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

Baselines

Domain Expert

TBlo

Supervised Learning

RSL

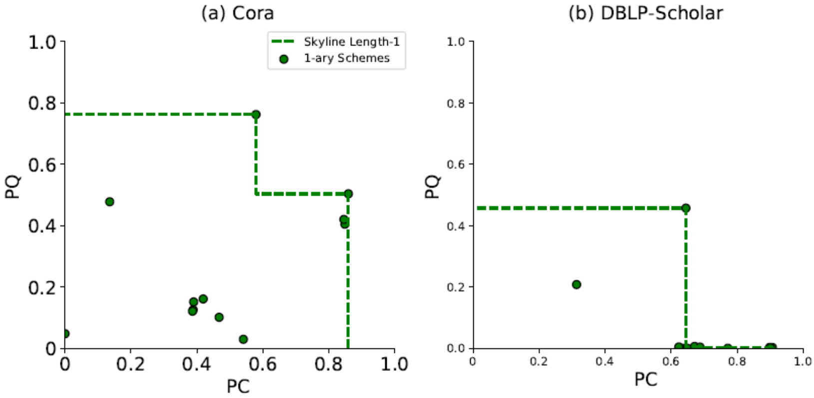
Unsupervised Learning

Fisher

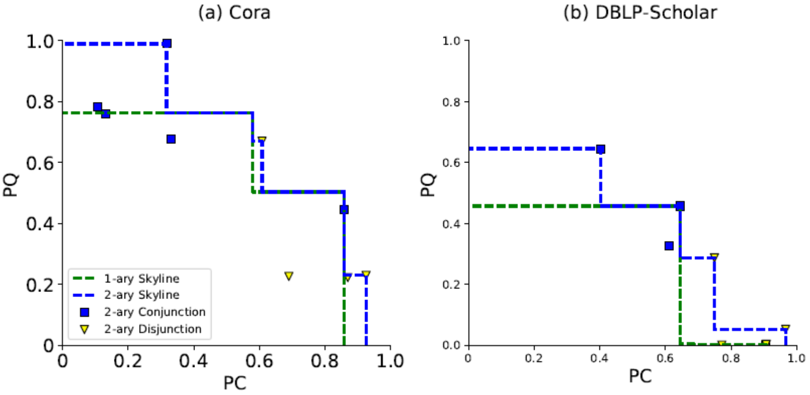
Active Learning

ASL

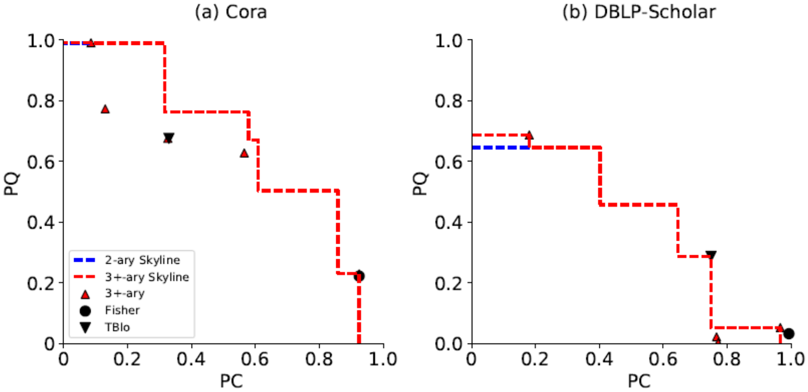
Scheme Skylines (Pro-Sky) in Experiments



Scheme Skylines (Pro-Sky) in Experiments



Scheme Skylines (Pro-Sky) in Experiments



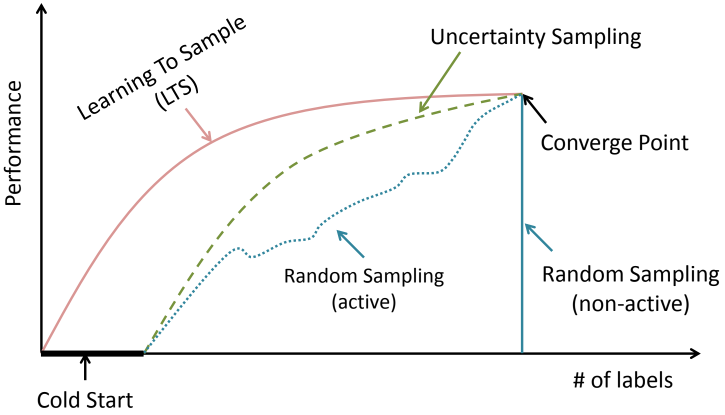
Agenda

* Introduction and challenges
* How to build a set of “optimal” blocking schemes efficiently?
* How to design an AL approach under various data distributions?
  – Learning based active learning for ER [1]
* How to alleviate the overfitting problem for powerful models?
* Conclusion

[1] J. Shao, Q. Wang and F. Liu. Learning To Sample: an Active Learning Framework. ICDM’19.

Objective

To build an active learning framework:



Challenges

* No one-fit-all: the "best" active learning strategy varies due to different datasets and machine learning models.
* Cold start problem: occurs under limited highly imbalanced samples.

Challenges

* No one-fit-all: the "best" active learning strategy varies due to different datasets and machine learning models.
* Cold start problem: occurs under limited highly imbalanced samples.

Solution

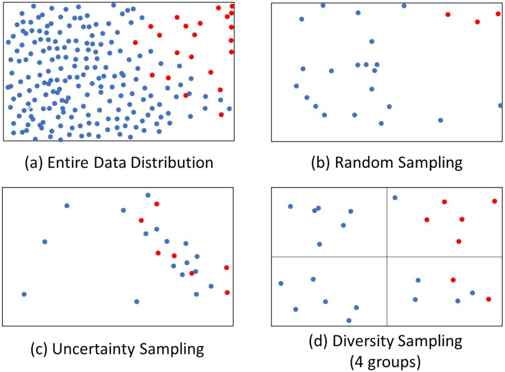
– Dynamical estimation of model performance (learning-based)

– Uncertainty and diversity of samples

Sample Uncertainty and Diversity

Uncertainty sampling: function-based uncertainty measures

Diversity sampling: considering sample distribution (feature values)



(a) Entire Data Distribution

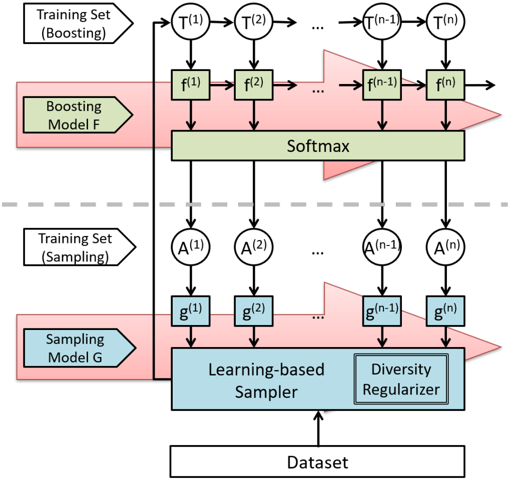
(b) Random Sampling

(c) Uncertainty Sampling

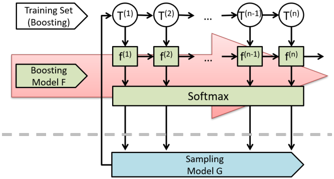
(d) Diversity Sampling (4 groups)

Framework: Learning to Sample (LTS)

Two models dynamically learn from each other in iterations for performance improvement.



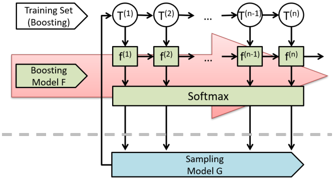
Boosting Model F



The boosting model F is a set of classifiers $\langle f^{(1)}, \ldots, f^{(n)} \rangle$.

A classifier $f^{(t)} \in F$ at the t-th iteration is trained by minimizing:

Boosting Model F



The boosting model F is a set of classifiers $\langle f^{(1)}, \ldots, f^{(n)} \rangle$.

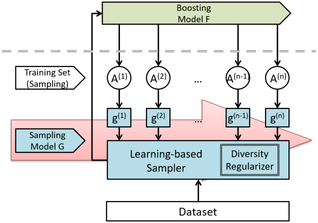
A classifier $f^{(t)} \in F$ at the t-th iteration is trained by minimizing:

$$\sum_{(x_i, y_i) \in T^{(t)}} \ell_1\left(\hat{y}_i^{(t-1)} + f^{(t)}(x_i), y_i\right) + \Omega_1(f^{(t)})$$

where:

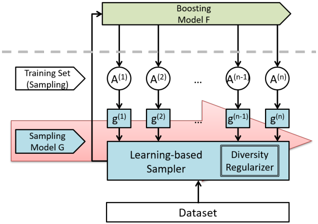
- $T^{(t)}$: training set;
- $\hat{y}_i^{(t-1)} = \sum_{k=1}^{t-1} f^{(k)}(x_i)$: predicted label of x_i;
- ℓ_1: a differentiable loss function;
- $\Omega_1(f^{(t)})$: the complexity penalty for $f^{(t)}$.

Sampling Model G



The sampling model G actively selects a set $\Delta^{(t)}$ of uncertainty and diversity samples at the t-th iteration by:

Sampling Model G



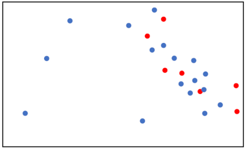
The sampling model G actively selects a set $\Delta^{(t)}$ of uncertainty and diversity samples at the t-th iteration by:

$$\textbf{maximize} \quad \sum_{i=1}^{k} v_i g^{(t)}(x_i) + \alpha \times \Gamma(\mathrm{v})$$

$$\textbf{subject to} \quad ||\mathrm{v}||_1 = |\Delta^{(t)}|$$

where $\mathrm{v} = (v_1, ..., v_k)^T \in \{0, 1\}^k$, k is the number of samples, and α is a parameter.

- A regressor $g^{(t)}(x_i)$ for uncertainty sampling
- A regularizer $\Gamma(\mathrm{v})$ for diversity sampling.



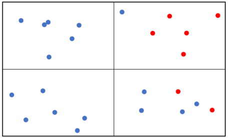
Uncertainty Sampling

A regressor is trained to predict the uncertainty of samples by minimizing:

$$\sum_{(x_i, z_i^{(t)}) \in A^{(t)}} w_i^{(t)} \ell_2(g^{(t)}(x_i), z_i^{(t)}) + \Omega_2(g^{(t)})$$

where:

- $A^{(t)} = \{(x_i, z_i^{(t)}) | x_i \in T^{(t)}, z_i^{(t)} \in [0,1]\}$: uncertainty sample set;
- $z_i^{(t)}$: the uncertainty of x_i;
- $w_i^{(t)}$: the weights of x_i;
- ℓ_2: a differentiable loss function;
- $\Omega_2(g^{(t)})$: the complexity penalty for $g^{(t)}$.



Diversity Sampling
(4 groups)

The diversity $\Gamma(\mathrm{v})$ is defined using a $l_{2,1}$-norm function:

$$\Gamma(\mathrm{v}) = ||\mathrm{v}||_{2,1} = \sum_{j=1}^{b} ||\mathrm{v}_j||_2$$

where:

- The sample space v with b groups $\{\mathrm{v}_1, \ldots, \mathrm{v}_b\}$;
- The vector $\mathrm{v}_j \in \{0, 1\}^m$ indicates samples selected in a group;
- Sample size $m = |X_j^{(t)}|$ in a group.

Results under Different Label Budgets

Dataset	Label Budget ζ (% of $\|X\|$)	CART	XG	XG+RS	XG + US $\alpha = 0$	XG+LTS $\alpha = 1$	XG + DS $\alpha \to \infty$
Cora	0.01	0	0	0	0	0.857	**0.878**
	0.05	0.741	0.763	0.750	0.827	0.864	**0.885**
	0.1	0.788	0.796	0.787	0.823	0.862	0.886
	0.5	0.848	0.835	0.835	0.873	**0.900**	0.893
	1	0.868	0.878	0.880	0.870	**0.902**	0.894
	5	0.878	0.897	0.892	0.907	**0.915**	0.898
NCVoter	0.01	0	0	0	0	0.324	**0.875**
	0.05	0	0	0	0	0.954	0.991
	0.1	0	0	0	0	**0.994**	0.993
	0.5	0	0	0	0	**0.994**	0.991
	1	0.334	0.379	0.398	0	0.993	**0.994**
	5	0.993	0.993	0.994	0.993	**0.997**	0.993
DBLP-ACM	0.1	0	0	0	0	0	**0.397**
	0.5	0	0	0	0	0.702	0.632
	1	0.348	0.347	0.279	0	**0.878**	0.721 3
	2	0.599	0.767	0.680	0.403	**0.884**	0.783
	5	0.870	0.850	0.803	0.874	**0.931**	0.833
	10	0.903	0.911	0.890	0.926	**0.981**	0.899
DBLP-Scholar	0.1	0	0	0	0	0.723	**0.731**
	0.5	0.378	0.54	0.498	0.555	0.773	**0.780**
	1	0.562	0.669	0.659	0.738	0.804	**0.792**
	2	0.772	0.806	0.771	0.807	**0.815**	0.801
	5	0.773	0.822	0.803	**0.836**	**0.836**	0.818 8
	10	0.808	0.835	0.830	**0.865**	0.851	0.829

* Introduction and challenges
* How to build a set of "optimal" blocking schemes efficiently?
* How to design an AL approach under various data distributions?
* How to alleviate the overfitting problem for powerful models?
  – A generative model with adversarial nets[1]
* Conclusion

[1] J. Shao, Q. Wang, A. Wijesinghe and E. Rahm. ErGAN: Generative Adversarial Networks for Entity Resolution. ICDM'20.

Challenges

* The imbalanced class problem: ER tasks
* The overfitting problem: powerful models

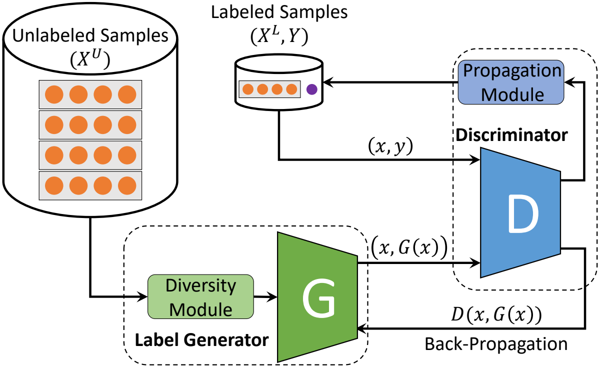
Challenges in ER with Limited Samples

Challenges

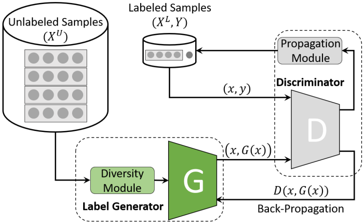
* The imbalanced class problem: ER tasks
* The overfitting problem: powerful models

Solution

- Label generator G: only have access to unlabeled samples, consider diverse samples
- Discriminator D: provide feedback to train G, limited labels used with propagation

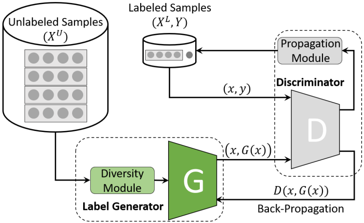

Unlabeled Samples
(X^U)
Labeled Samples
(X^L, Y)
Propagation Module
Discriminator
(x, y)
D
$(x, G(x))$
Diversity Module
G
Label Generator
$D(x, G(x))$
Back-Propagation

Label Generator G and Diversity Module



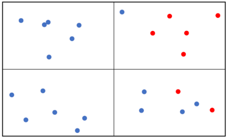
Generate pseudo labels for unlabeled samples

Learn a conditional distribution $p_g(Y|X^U) \approx p(Y|X^U)$



Generate pseudo labels for unlabeled samples

Learn a conditional distribution $p_g(Y|X^U) \approx p(Y|X^U)$



Diversity Sampling
(4 groups)

A minibatch of m samples is selected from X^U according to the following objective function:

$$\textbf{maximize} \quad ||\mathbf{v}||_{2,1} \quad \text{s.t.} \sum_{i,j} v_i^j = m$$

Objective Function of G

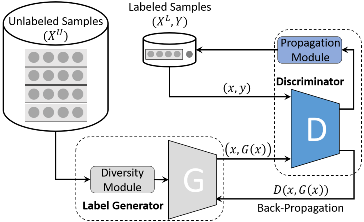
G updates its parameters according to:

$$\mathcal{L}_G = \min_G \quad \mathbb{E}_{x \sim p(X_i^U)}[\log(1 - D(x, G(x)))] \tag{1}$$

where:

- $G(x_i)$ is the pseudo label of x_i generated by G;
- $(x_i, G(x_i))$ is a pseudo labeled sample sent to the discriminator D;
- $D(x, G(x))$ is the feedback from the discriminator D.

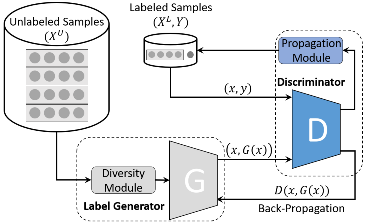
Discriminator D



Distinguish samples with pseudo labels from samples with real labels

Learn a joint distribution $p(X, Y)$

Discriminator D



Distinguish samples with pseudo labels from samples with real labels

Learn a joint distribution $p(X, Y)$

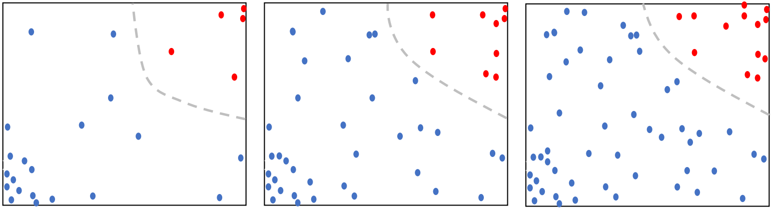
The objective function of D at the t-th iteration of propagation is:

$$\begin{aligned} \mathcal{L}_D = \max_D \quad & \mathbb{E}_{x \sim p(X_i^U)} \log[(1 - D(x, G(x)))] \\ & + \lambda \mathbb{E}_{(x,y) \sim (X^*, Y)^t} \log[D(x, y)] \end{aligned} \tag{2}$$

where:

- λ refers to a weighted term.
- $(X^*, Y)^t$ refers to the labeled samples in t-th iteration.

Propagation Module



(a) Seed Samples (b) 1-st Iteration (c) 2-nd Iteration

The *propagation module* selects a minibatch of $|\Delta X^t|$ high-quality pseudo labeled samples for training D:

$$\underset{\Delta X^t \subseteq X^t}{\operatorname{argmax}} \sum_{x \in \Delta X^t} D(x, G(x))$$

Experimental Setup: Baselines

Australian National University

* *Unsupervised:* **Two-Steps** and **Iterative Term-Entity Ranking and CliqueRank** (**ITER-CR**).
* *Semi-supervised:* **Semi-supervised Boosted Classifier** (**SBC**).
* *Fully supervised:* **Magellan** and **eXtreme Gradient boosting** (**XGboost**).
* *Deep Learning based:* **DeepMatcher** (**DM**) and **Deep Transfer Active Learning** (**DTAL**).
* Ablation Study: **ErGAN+WE**, **ErGAN-D**, **ErGAN-P**, and **ErNN**.

Method	Datasets			
	Cora	DBLP-ACM	DBLP-Scholar	NCVoter
2S	62.69	91.43	68.78	98.96
ITER-CR*	89.00	–	–	–
SBC	85.71	97.09	85.47	99.78
SVM	88.95	97.19	85.71	98.48
LR	80.25	95.56	83.84	99.37
XGBoost	91.34	97.20	86.63	**100**
ErGAN	**93.03**	**98.23**	**88.32**	**100**
DM	98.58	98.29	94.68	**100**
DTAL*	$98.68_{\pm 0.26}$	$98.45_{\pm 0.22}$	$92.94_{\pm 0.47}$	–
ErGAN+WE	$\mathbf{98.72}_{\pm 0.15}$	$\mathbf{98.51}_{\pm 0.23}$	$\mathbf{94.73}_{\pm 0.35}$	**100**

Results: Ablation Study

Datasets	Cora				DBLP-ACM			
	0.1%	1%	20%	60%	0.1%	1%	20%	60%
ErNN	84.46	90.67	91.43	92.78	88.05	95.68	98.20	98.22
ErGAN-D	79.87	85.14	91.27	92.97	0	93.30	97.16	98.21
ErGAN-P	85.18	90.76	91.42	**93.03**	92.67	95.96	98.21	**98.23**
ErGAN	**87.45**	**91.07**	**91.54**	**93.03**	**96.89**	**96.93**	**98.22**	**98.23**
Datasets	**DBLP-Scholar**				**NCVoter**			
	0.1%	1%	20%	60%	0.1%	1%	20%	60%
ErNN	82.76	83.17	86.71	87.73	99.39	**100**	**100**	**100**
ErGAN-D	0	78.85	83.43	88.29	0	99.58	**100**	**100**
ErGAN-P	83.43	85.34	86.55	**88.32**	99.39	99.79	**100**	**100**
ErGAN	**84.23**	**85.85**	**86.86**	**88.32**	**99.45**	**100**	**100**	**100**

In summary, we have proposed four approaches for ER:

* ASL: an active scheme learning approach
* Skyblocking: scheme skyline learning under different blocking criteria
* LST: A learning-based active learning framework
* ERGAN: a generative model with adversarial nets

Thank You!

Q & A

Email: Jingyu.shao@anu.edu.au