# Estimating accuracy

Zihan Zhou, Kirin Hong, Mars Hao

May 8, 2019

**Abstract**

In this step, we'd like to estimate the precision and recall. We used the Prediction list, Candidate set, Table A, Table B downloaded from Cloud Master and 400 random sample pairs labeled manually to compute the confidence interval of precision and recall.

## 1 Methods

After we downloaded Candidate set **C**, we found that there are about 70,000 candidates in **C**, which is larger than 500. So we followed the instruction step by step and tried to add some blocking rules to reduce the size of **C**.

By checking the blocking rules and batch samples, we realized that the restriction on *year* could be stricter. At the first time, we only considered that matches whose year is identical. But in this case, we found it would drop some true positives. Thus, we only kept $|\text{Table A.year} - \text{Table B.year}| \leq 1$. In this case, we didn't lose true positives any more. And the candidate set was significantly reduced to **C'**.

Then, we sampled 50 pairs randomly from **C'** and the density here is large than 0.2. Finally, we started labeling the 400 tuple pairs.

## 2 Results

The result could be replicated via sample.ipynb and estimating_precison_recall.ipynb.

| Recall | Precision |
|---|---|
| 1.0 - 1.0 | 0.95 - 1 |

Table 1: The estimated recall and precision