

# Exploring Identical Users on GitHub and Stack Overflow

Takahiro Komamizu, Yasuhiro Hayase,  
Toshiyuki Amagasa, Hiroyuki Kitagawa

University of Tsukuba, Japan

# Mining Software Repository (MSR)

---

- MSR is a data mining field
  - Analysis of software in the repositories
    - How the software are used?
    - What are the popular software?
    - Which part of the codes can be reused in other software?
    - ...
  - User behaviour analysis
    - Who are the professionals of a particular language (e.g., Java, Python, Scala)?
    - Who are suitable for solving issues on projects?
    - Who can give advices for improving software in some aspects (e.g., performance, usability)?
    - ...
  - ....



# User Behaviour Analysis on MSR

---

- User profiling
  - Mainly based on users' activities on repositories
    - e.g., commits, bug fixing
  - For instance,
    - Users who commits lots of Java codes can be regarded as Java professionals.
    - Users who solve lots of issues can be regarded as good issue solvers.
- Problem: lack of information
  - newly registered users
  - users having few activities on repositories

# Approach: Cross-platform Analysis

---

- In cooperating with other platforms





- Expectation
  - Users' activities in other platforms can be imported as supplemental facts of the users.
  - For instance,
    - Users answering questions about Java programming are professional of Java.
    - Users asking questions about some libraries may be interested in participating their developments.

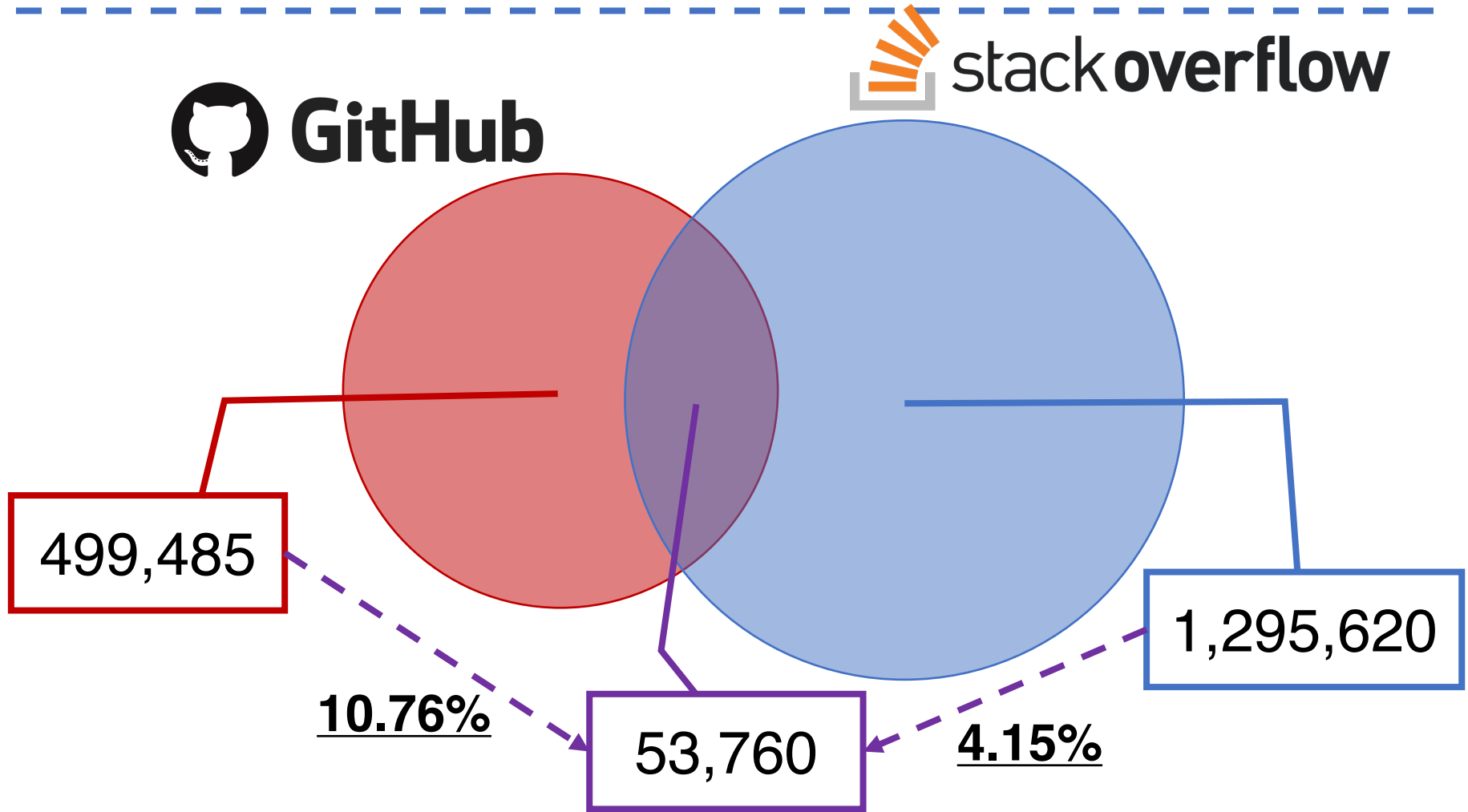
# User identification b/w GH and SO

---

- Users can be identified by hashed values of email addresses
  - Email addresses on Stack Overflow are hashed by MD5 function.
  - Those on GitHub are raw string.
  - Hashing email addresses on GitHub make it possible to match with those in Stack Overflow.

MD5 (  **GitHub** ) ==  **stackoverflow**

# How many the identifiable?



[1] G. Gousios, "The GHTorrent Dataset and Tool Suite," in *MSR 2013*, 2013, pp. 233–236.

[2] A. Bacchelli, "Mining Challenge 2013: Stack Overflow," in *MSR 2013*, 2013.

# Is email address only way to identify? – No.

---

- Same users can easily use other email addresses in various reasons.
  - A user changes her email address from service to service.
  - Another changed her email address caused by some reasons.
- Profile information have many commonality.
  - Similar / same user-name
  - Close locations (e.g., Pittsburg vs. PA)
- Users' activities also have commonality.
  - Projects and questions.

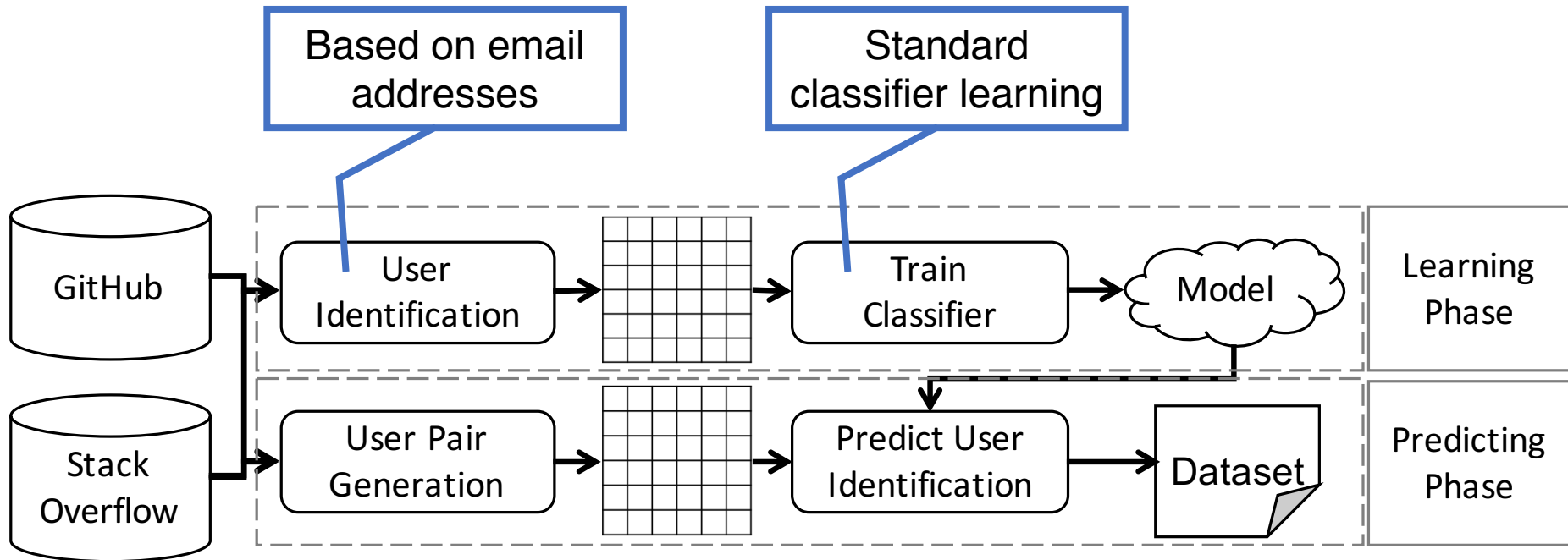
# This paper

---

- Purpose: increase the num. of the identifiable
  - The more information about users, the more evidence for cross-platform analyses.
- Attempt: identify users
  - Identify users from other information than email addresses.
- Contributions
  - Classification-based user identification mechanism
  - Examining standard classification methods
  - Public datasets and tools
    - [https://github.com/Taka-Coma/PJD\\_GHSO](https://github.com/Taka-Coma/PJD_GHSO)



# The framework



- **Issues**

- Attributes selection
- Label skewness

# Attributes selection & similarity

---

Attributes on GitHub	Attributes on Stack Overflow	
users.name	users.display_name	3gram-based cosine sim.
users.location	users.location	TFIDF-based cosine sim.
users.created_at	users.creation_date	Inverse of time diff.
projects.description	users.about_me	TFIDF-based cosine sim.
projects.description	posts.body	
projects.description	posts.tags	
projects.description	posts.title	
projects.description	comments.comments	

$$\text{Cosine}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{|\mathbf{v}_1| |\mathbf{v}_2|}$$

$$\text{DateSim}(\text{date}_1, \text{date}_2) = \frac{1}{|\text{date}_1 - \text{date}_2|}$$

# Skewness problem

---

- Quite small number of positive samples comparing with that of negative samples
  - Positive: the identifiable via email addresses
  - Negative: other pairs (combinations of users)
  - In the dataset
    - #pos = 53,760
    - #neg = 96.5 billion
- If highly skewed, classifier always answers labels of majority (i.e., negative).
- Approach: Down sampling the negatives

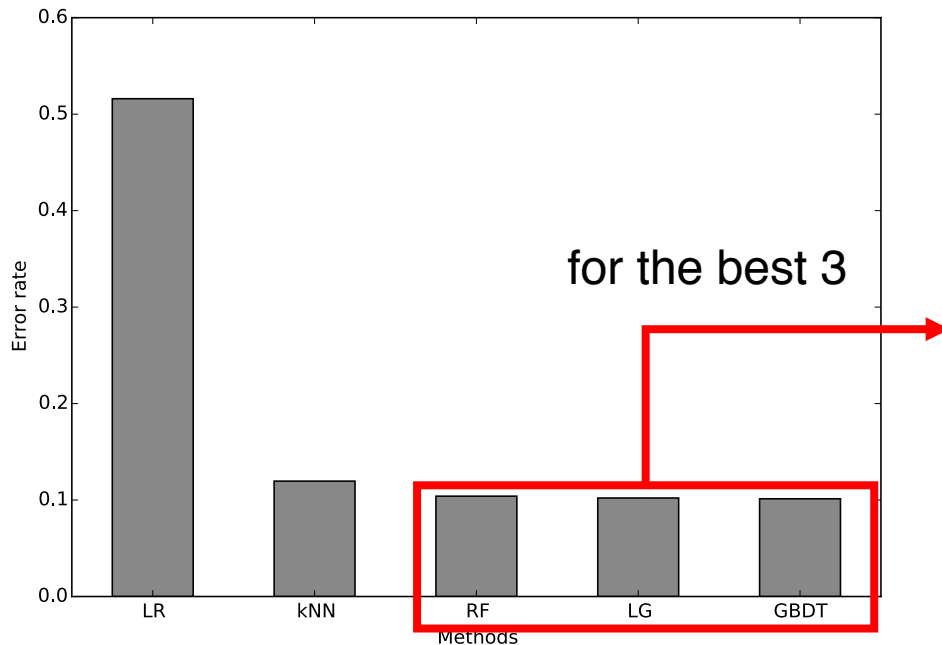
# User identification examination

---

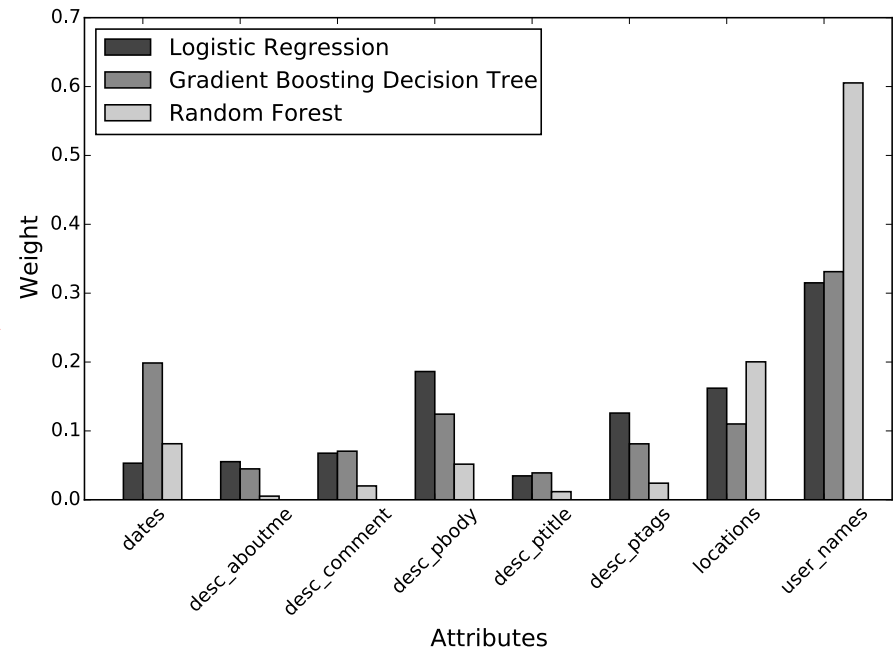
- Datasets:
  - MSR challenge datasets: 2013[2] and 2014[1]
- Classification methods
  - Linear regression (LR)
  - k-nearest neighbor (kNN)
  - Random forest (RF)
  - Logistic regression (LG)
  - Gradient boosting decision tree (GBDT)
- 10-fold cross validation

# Evaluations

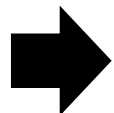
## Error rate



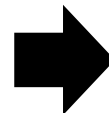
## Weights for attributes



10% error



More sophisticated classification methods have chance to improve.



- Users have similar names on GH and SO.
- Locations equally contribute in these classifications.

# Conclusion

---

- User identification problem b/w GH and SO
- Formulate as classification problem
  - Attribute selections
  - Skewness problem → down sampling
  - Standard classification methods
- Evaluations
  - 10% classification error
  - Attributes differently contribute on different classification methods
- Future work
  - Improve with more sophisticated classification methods.