GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

# Classification for Participants in Github based on their Behaviors

Final Project for Data Science and Big Data Analytics

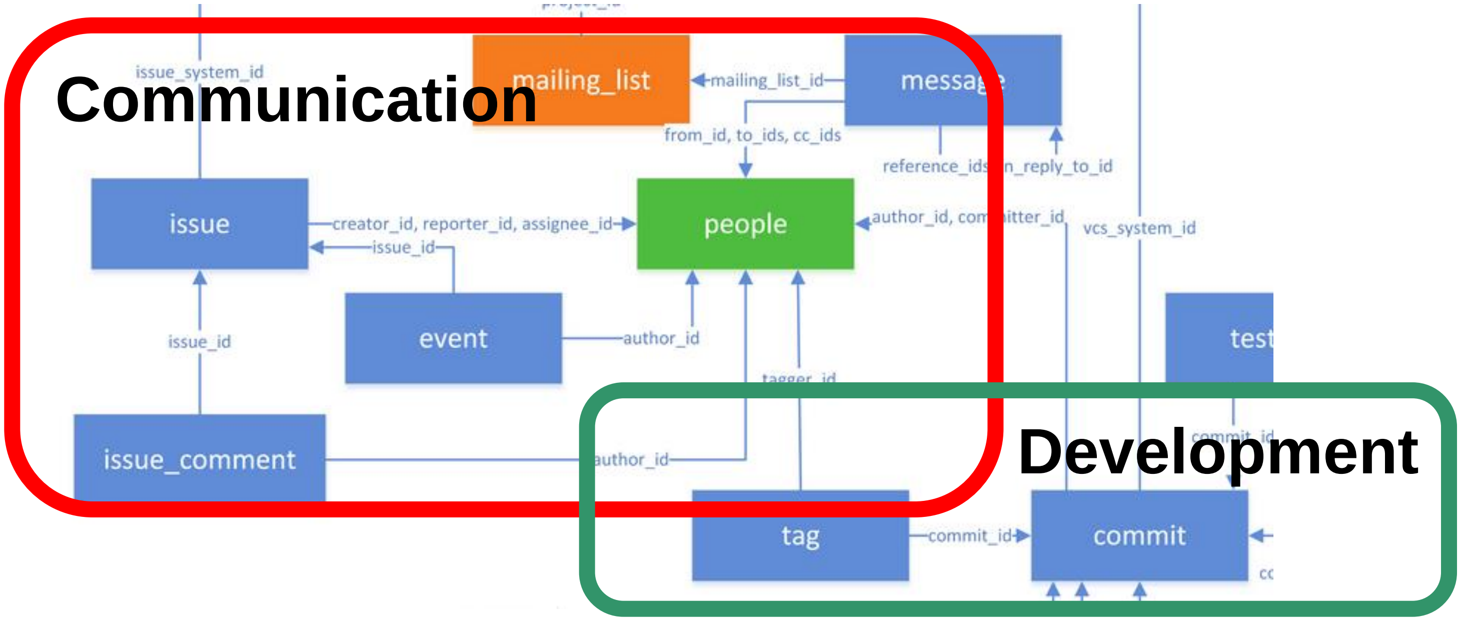
DENG, Fangqi
LIANG, Chencheng
LONG, Qinqin
ZHU, Chenfeng (Presenter)

# Situation & Project Goals

- Situation:
  - Zookeeper is a project in Github (anonymous).
  - There are thousands of participants.
- Goals:
  - Detect categories of the participants.
  - Other, user, developer.

# Main Findings & Approach

- Activities
  - Communication
  - Development
- Method & Tech
  - Cluster/Classification
  - Develop model in R



# Model Description

- Model
    - K-Means Model
    - Naive Bayes Model
    - Decision Tree Model
- Dependent variable
    - Numbers of commit, issue, event, tag and email
- Sampling (4583)
    - Training sample: 3000 participants
    - Testing sample: 1583 participants

# K-Means

- Simple & Direct

| K-means | Other | User | Developer |
|---|---|---|---|
| **Other** | 4318 | 1 | 13 |
| **User** | 110 | 0 | 0 |
| **Developer** | 141 | 0 | 0 |

  - Accuracy=94.22%, accuracy is different every time(some time accuracy=2.42%).
  - K=3.
  - Used attributes(7): issue_create_total, issue_report_total, etc.
  - Who has weak connection (other) with this project and who has strong connection (user and developer).

# K-Means

- Advanced

| K-means | User | Developer |
|---|---|---|
| **User** | 0 | 110 |
| **Developer** | 1 | 140 |

– Category “other” is deleted. (not so accurate.

- Accuracy=56.57%
- K=2.
- Used attributes(3): commit_commit_total, issue_create_total, issue_report_total.

# Naive Bayes

| naiveBayes | Other | User | Developer |
|---|---|---|---|
| **Other** | 0 | 0 | 0 |
| **User** | 1188 | 43 | 44 |
| **Developer** | 306 | 2 | 1 |

- Accuracy=2.78%. Accuracy is constant every time.
- Used attributes(7): issue_create_total, issue_report_total, etc.

# Decision Tree

| ctree | Other | User | Developer |
|---|---|---|---|
| **Other** | 1494 | 0 | 0 |
| **User** | 45 | 0 | 0 |
| **Developer** | 45 | 0 | 0 |

- Accuracy=94.32%.  Accuracy is constant every time.
- Used attributes(7): issue_create_total, issue_report_total, etc.
- Who has weak connection (other) with this project and who has strong connection (user and developer).

# Model Details

- Data Preparation
    - Load data from DB (by R and Robomongo).
    - Calculate the values of different variables(attributes).
    - Transform data into a new data table. (variables as the columns and row for participants)
- Data Analysis
    - Create the training set and testing set.
    - Analysis with kmeans, naiveBayes and ctree in R.

# References

- http://smartshark2.informatik.uni-goettingen.de/documentation/
- https://cran.r-project.org/manuals.html
- https://docs.mongodb.com/manual
- https://github.com/sampig/DataScience/blob/master/DataScience/final_project.R

Questions & Answers

**THANKS!**