June 30, 2017
Recruit Holdings Co.,Ltd.

# Recruit's Artificial Intelligence Laboratory Releases BigGorilla: An Open-source Data Integration and Data Preparation Ecosystem

The Recruit Institute of Technology (hereinafter referred to as "RIT"), the artificial intelligence research laboratory of Recruit Holdings Co., Ltd. (Headquarters: Chiyoda, Tokyo; Representative Director and CEO: Masumi Minegishi; hereinafter referred to as "Recruit"), released BigGorilla, an open-source data integration and data preparation ecosystem in Python that will help reduce the time data scientists spent on preparing their data for analysis.

## 1. Background and Objective

Artificial Intelligence in businesses relies on quality data to derive meaningful insights. Data scientists obtain quality data through a process that typically involves several steps, including data acquisition from different sources, extraction of structured data from unstructured data, and cleaning and integrating heterogeneous data into a format consumable by downstream algorithms that will turn data into insights. Experts estimate that data scientists often spend "*50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.*"

The goal of BigGorilla is to reduce the time it takes to prepare unruly data so that data scientists can spend more time on the more interesting work of data analysis.

## 2. BigGorilla Summary

The BigGorilla project began in September 2016 jointly by the RIT team (led by Dr. Wang-Chiew Tan) and Professor AnHai Doan from University of Wisconsin at Madison. RIT and Professor Doan envisioned an open-source ecosystem centered around the different tasks of data integration and preparation. The project was named "BigGorilla" because data integration and preparation is a big, hairy, and nasty problem. For the different steps taken by data scientists in integrating and preparing data, BigGorilla documents existing technologies and also points to desired technologies that could be developed by the community.

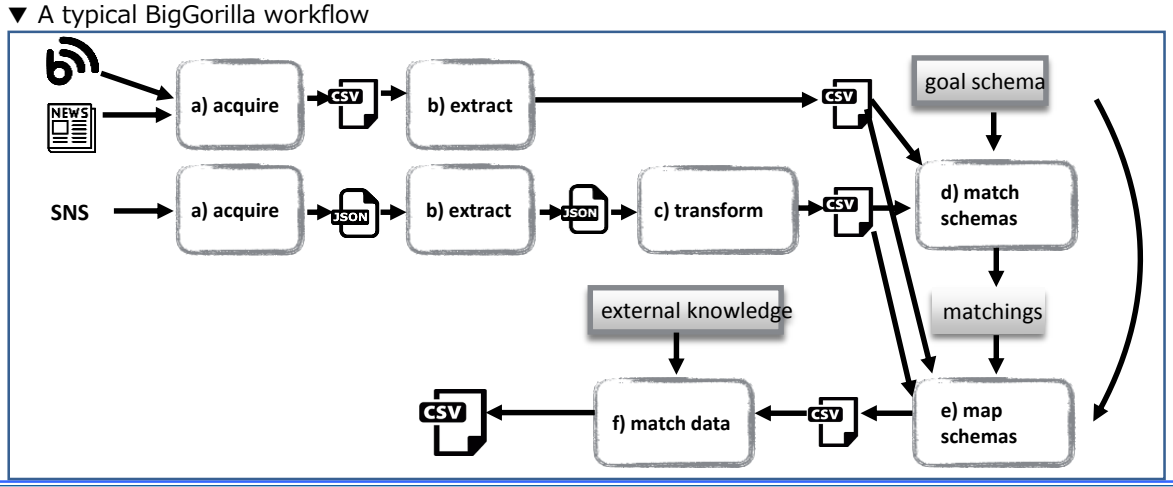
**▼BigGorilla Components**

a) Data acquisition: Acquire or scrape data from external sources, such as databases, web pages, social networks, etc.
b) Data extraction: Extract structured data, such as person or organization names, from unstructured sources, such as text.
c) Data transformation: Reshape data or convert data from one format to another, such as from CSV to JSON.
d) Schema matching and merging: Match attributes of different schemas and merge the schemas.
e) Schema mapping: Generate code (from schema matchings) that can be executed to combine data into a specific format.
f) Data matching and merging: Identify data that should be related, such as "Recruit" and "Recruit Holdings Inc."

To date, RIT contributed the KOKO and FlexMatcher packages in components (b) and (d) respectively. The Magellan package for entity matching from Professor Doan and his team can be found in (f).

**▼BigGorilla Official Site (English)** : http://www.biggorilla.org

▼ A typical BigGorilla workflow



## 3. BigGorilla Use Case and Effectiveness

As of today, 12 groups across 8 companies within Recruit are already actively using or in the process of considering BigGorilla.

Our experience showed that BigGorilla is effective across the company's diverse range of businesses: from the extraction of store names (or person names and location information) from unstructured data and web pages, the de-duplication of multiple variants of store names (or company names, property names), to the conversion of medical prescription data and merging of lists from multiple data sources. For example, with BigGorilla, we obtained 98.9% accuracy on the task of de-duplicating approximately 10,000 store names. BigGorilla achieves the best results when compared with existing systems offering similar services. BigGorilla is also efficient and is good at extracting structured data from web pages; it can de-duplicate about 100,000 store names in about 30 minutes and achieved about 70% accuracy when extracting information about stores within a specific region from a web site. For the latter task, BigGorilla extracted the information with a 98% reduction in man-hours when compared to the same task being performed manually.

Overall, our experience demonstrates that BigGorilla is highly promising in facilitating the integration and preparation of good quality data at a fraction of the time and effort that is normally needed. In addition to savings that will translate into the company's bottom line, BigGorilla will also allow data scientists to focus on the more crucial task of deriving meaningful insights from data that will help drive critical company decisions.

RIT will continue to develop tools to accelerate the overall process of integrating and preparing data. RIT's broader vision is to create environments that will facilitate the overall process of deriving insights from data so that businesses can quickly analyze, develop, test hypotheses, and ultimately contribute to user satisfaction and business growth.


## 4. Future Prospects

BigGorilla looks forward to contributing further to the open-source community and is also eager to collaborate with other members of the open-source community and with universities. If you are interested in getting in touch with BigGorilla, please feel free to contact us with the following email.

▼ For inquiries about BigGorilla, please contact:
*thebiggorilla.team@gmail.com*

Recruit Holdings will continue to provide people with useful information about the areas of work, learning, home, marriage, child rearing, travel, cars, hobbies and lifestyle. By offering services which provide users with new encounters and opportunities, we hope to be the one to help each of our users find their next big chance.

Inquiries about this press release:
http://www.recruit-rgf.com/support/