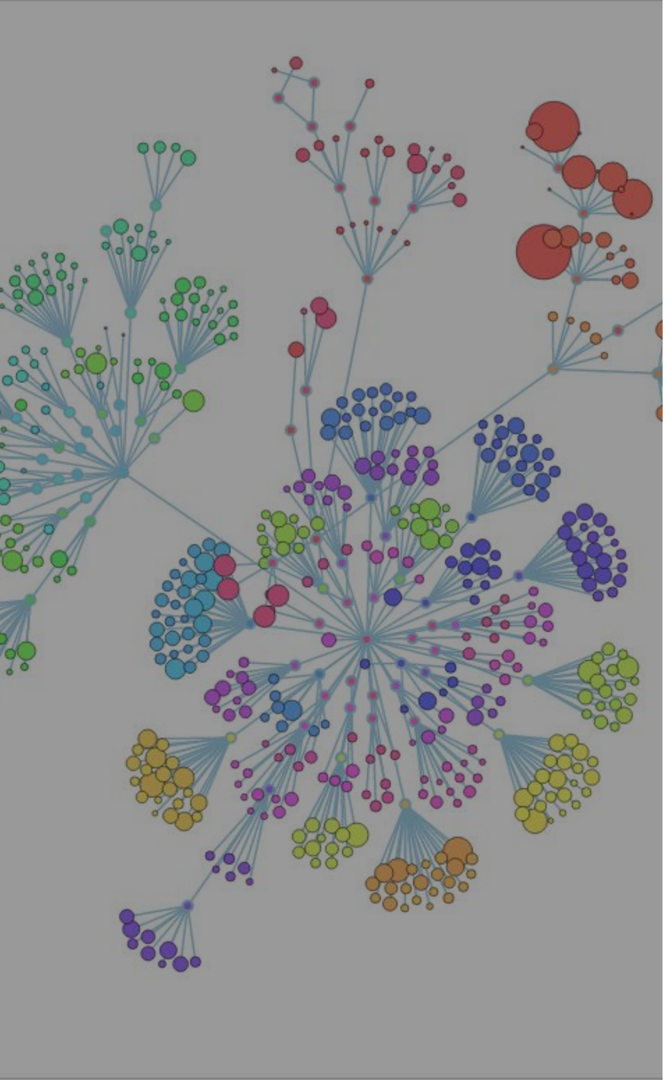
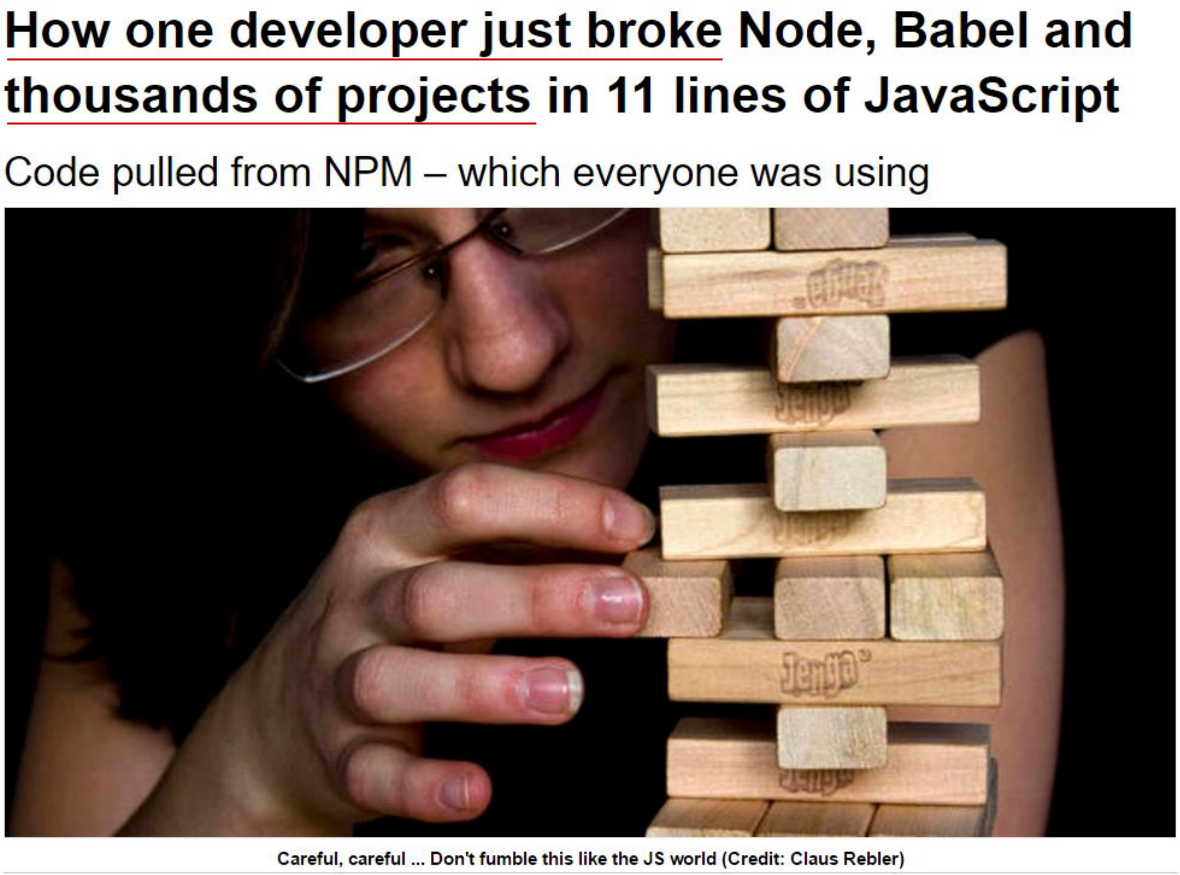
Repo Flower

See how GitHub Repos depend on one another

Insight Data Engineering Fellowship, Palo Alto
Victor Chen





How one developer just broke Node, Babel and thousands of projects in 11 lines of JavaScript

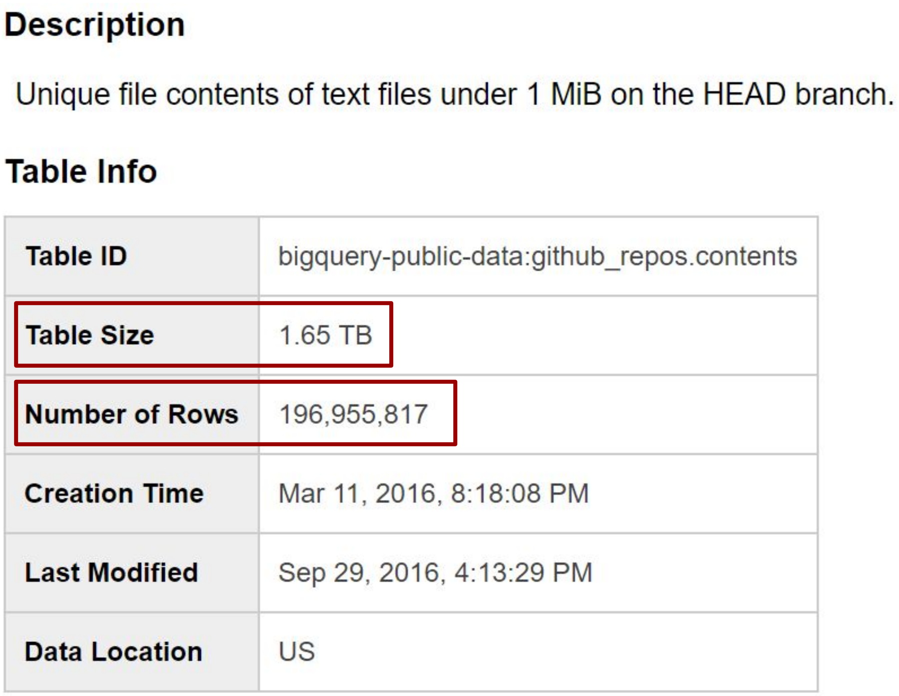
Code pulled from NPM – which everyone was using

Careful, careful ... Don't fumble this like the JS world (Credit: Claus Rebler)

http://www.theregister.co.uk/2016/03/23/npm_left_pad_chaos/

Data

git GitHub Archive Google BigQuery



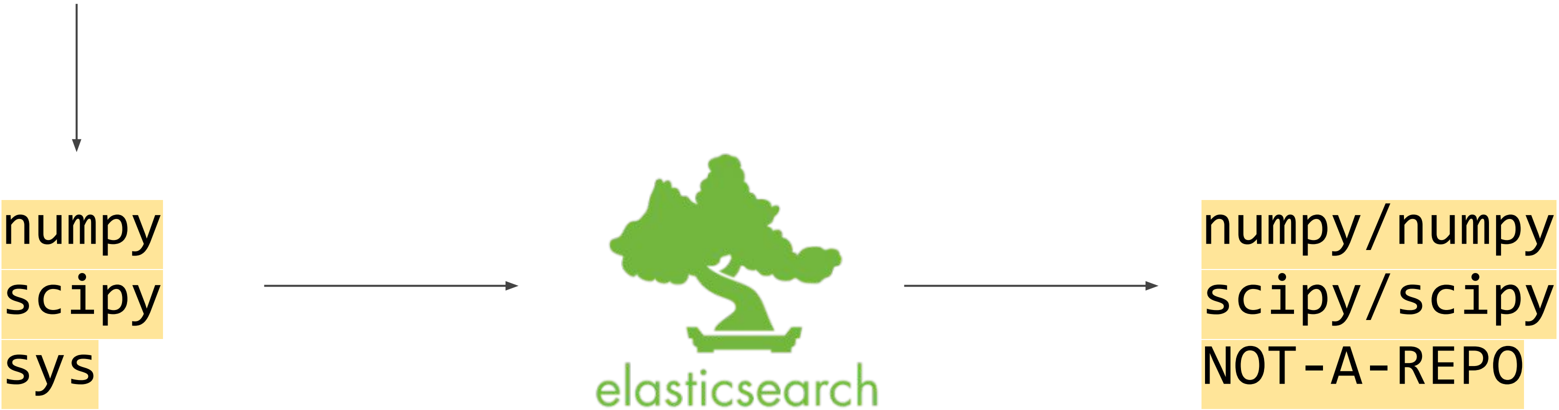
Description

Unique file contents of text files under 1 MiB on the HEAD branch.

Table Info

Table ID	bigquery-public-data:github_repos.contents
Table Size	1.65 TB
Number of Rows	196,955,817
Creation Time	Mar 11, 2016, 8:18:08 PM
Last Modified	Sep 29, 2016, 4:13:29 PM
Data Location	US

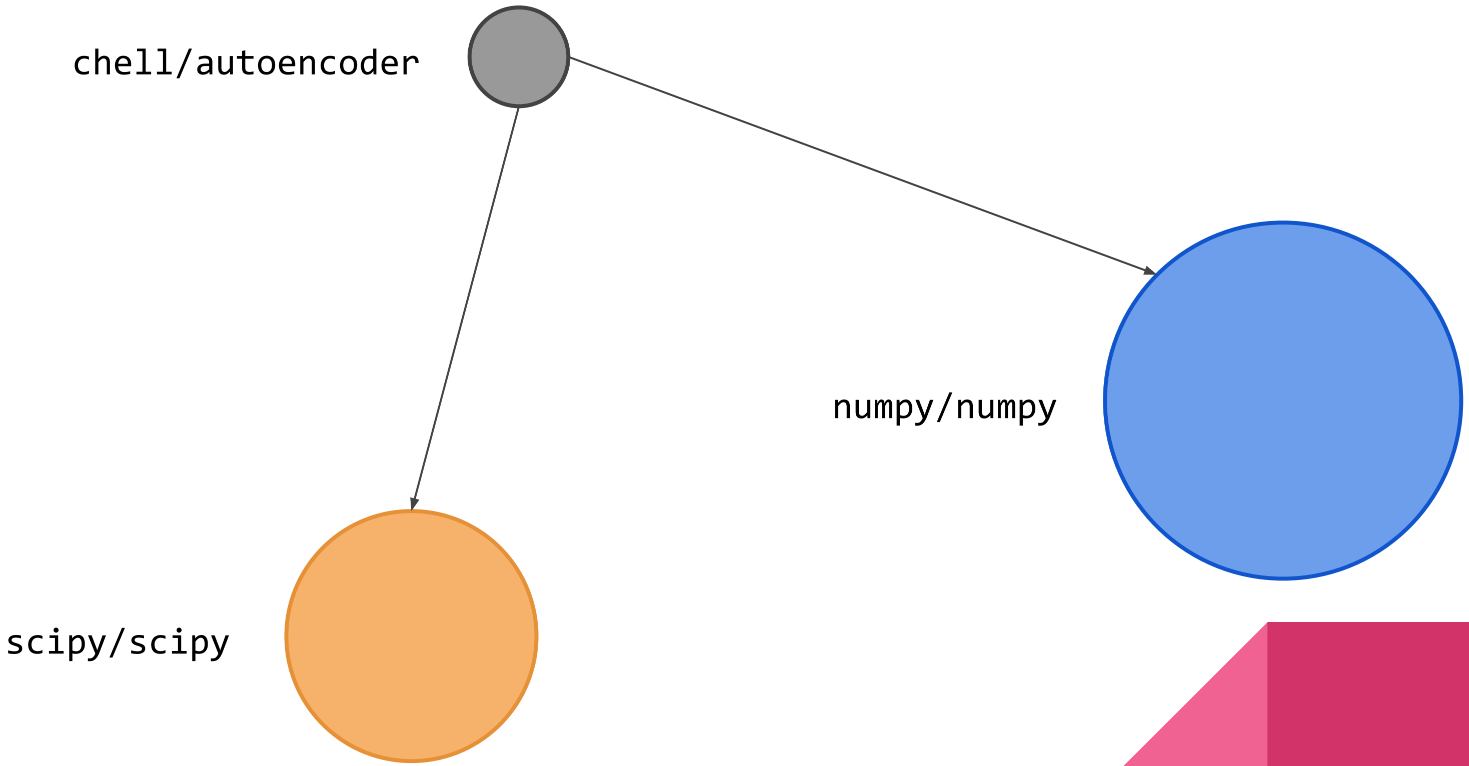
```
# github.com/chell/autoencoder/main.py
import numpy as np
from scipy.sparse import csr_matrix
import sys
if __name__ == ‘__main__’: ...
```

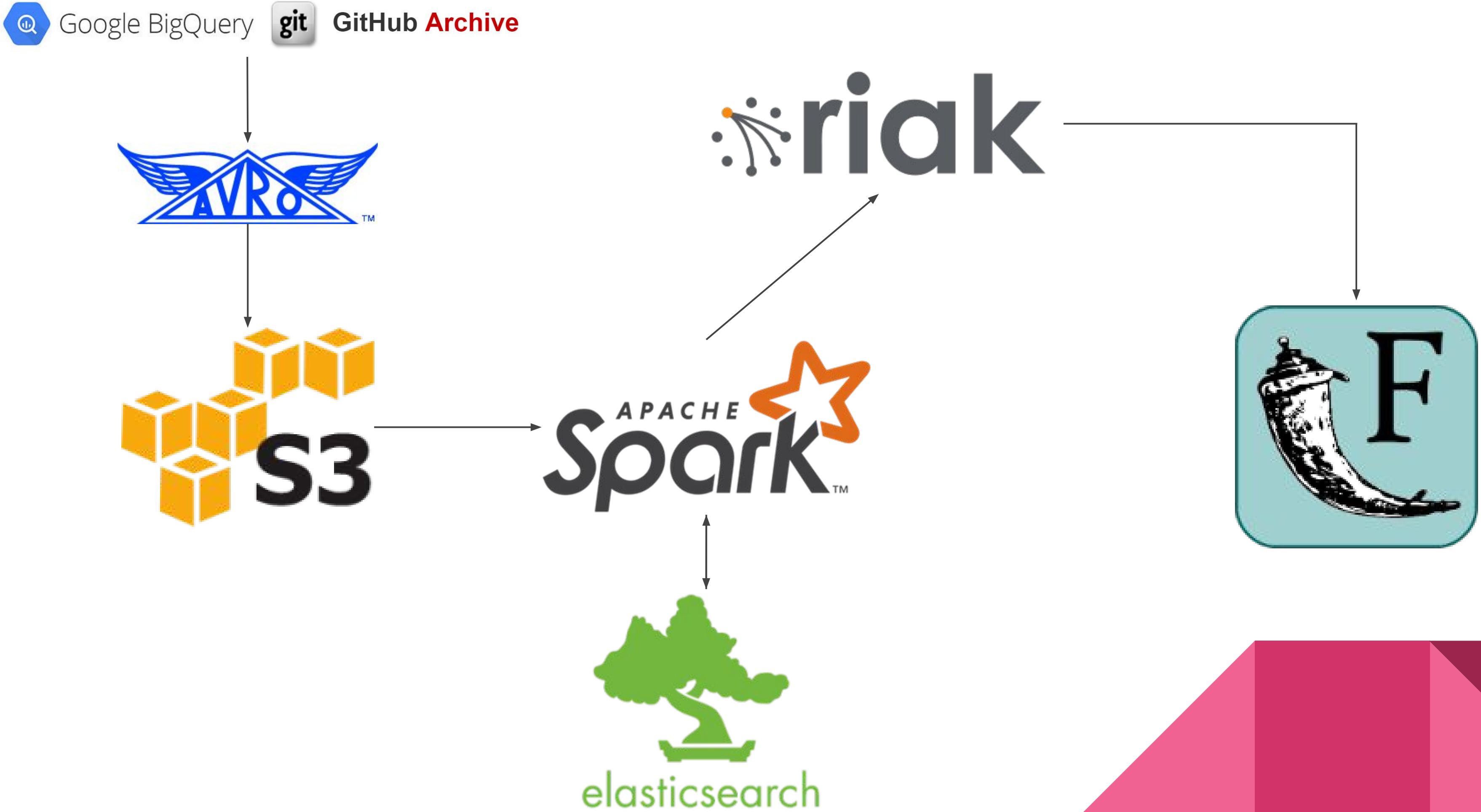


numpy
scipy
sys

numpy/numpy
scipy/scipy
NOT-A-REPO

(chell/autoencoder, [numpy/numpy, scipy/scipy])




Google BigQuery
git
GitHub Archive
AVRO ™
S3
APACHE Spark ™
riak
F
elasticsearch

Performance

- 4 mil (33GB) Python files
- ~30 min processing

- 1 mil (7GB) Go files
- ~10 min processing

- 4 node Spark
- 3 node Elasticsearch
- 3 node Riak
- m4.xlarge

Challenge: Entity Resolution

```
# victor/my_repo: test.py
from mock import ...
    # referring to
    #  victor/my_repo: mock.py?
    #  victor/my_repo: mock/?
    #  some_random_gal/mock?
if __name__ == ‘__main__’:
    ...
```

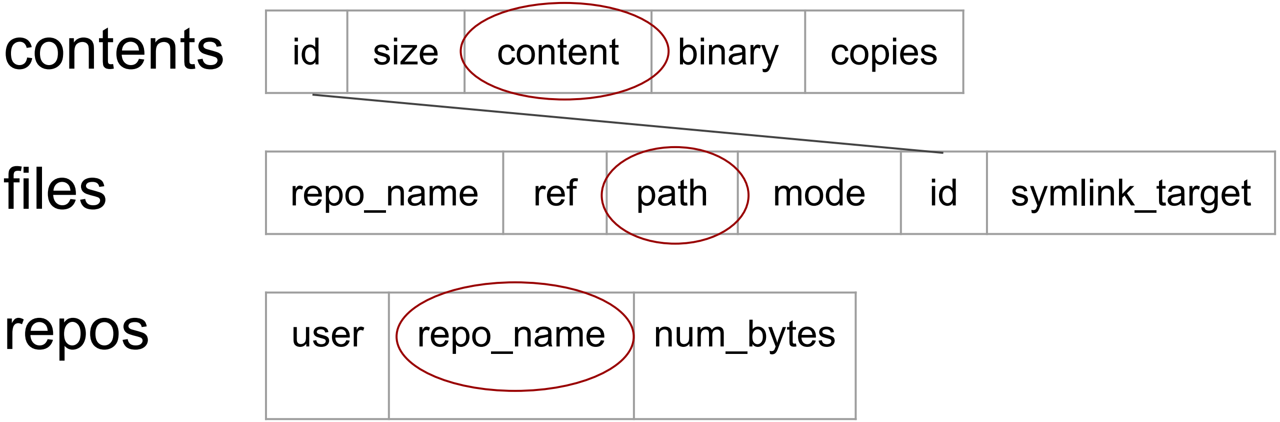
Challenge: Entity Resolution

```
# victor/my_repo: test.py
from mock import ...
    # referring to
    #  victor/my_repo: mock.py?
    #  victor/my_repo: mock/?
    #  some_random_gal/mock?
if __name__ == ‘__main__’:
    ...
```

- Ignore “self-references”

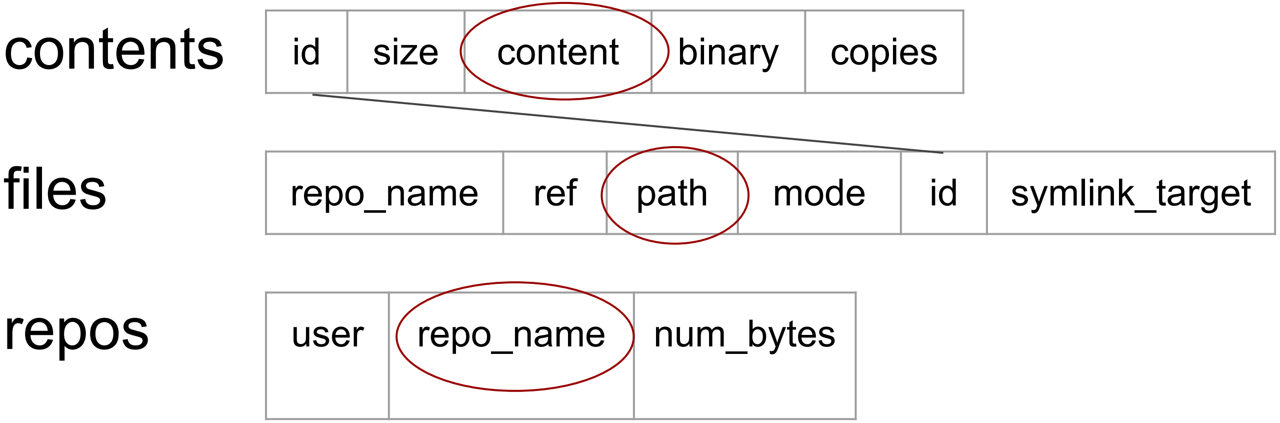
- Rank repos by star count

Challenge: Data Processing



contents	id	size	content	binary	copies	
files	repo_name	ref	path	mode	id	symlink_target
repos	user	repo_name	num_bytes			

Challenge: Data Processing



contents

id	size	content	binary	copies

files

repo_name	ref	path	mode	id	symlink_target

repos

user	repo_name	num_bytes

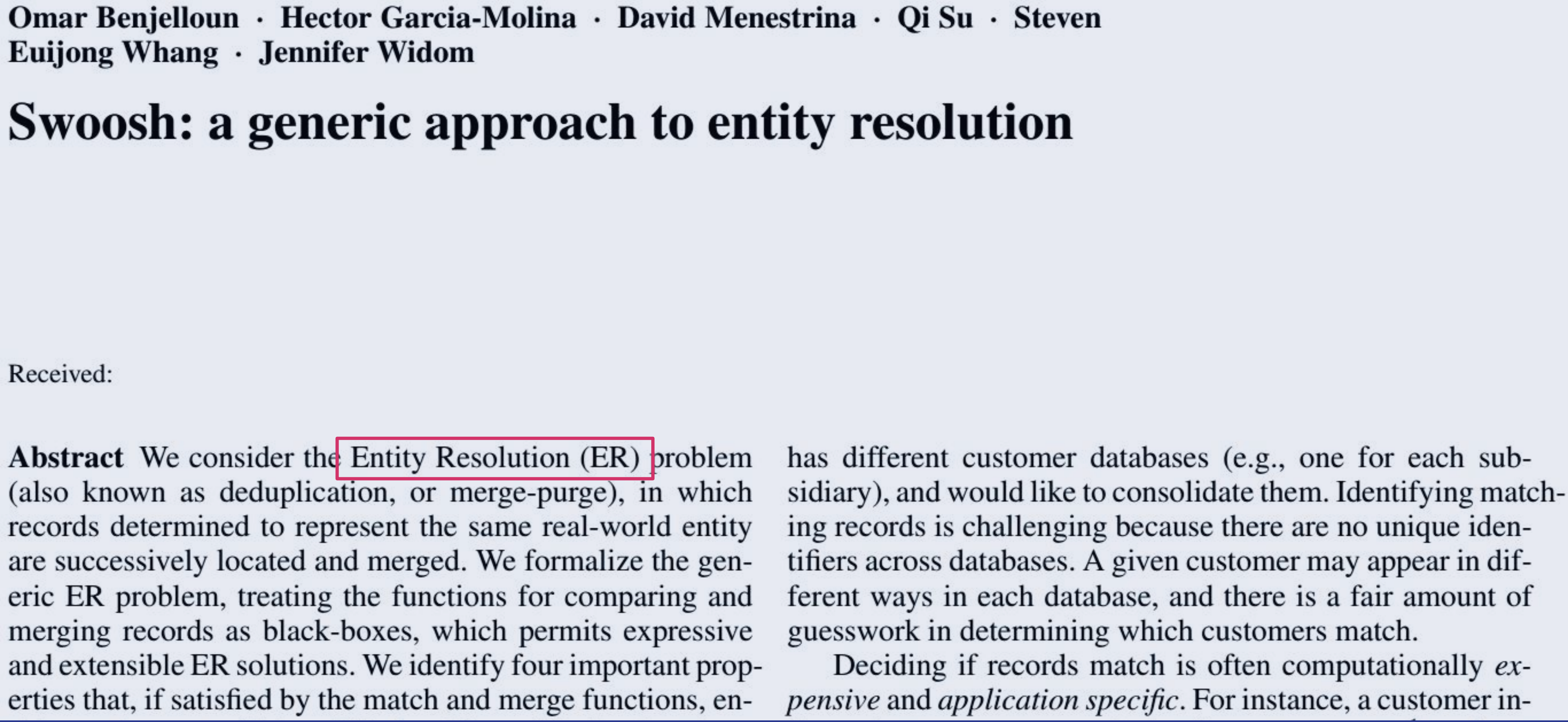
Repo-to-repo adjacency lists: e.g.

```
(chell/autoencoder, [numpy/numpy, scipy/scipy])
```

- Victor Chen
- University of California, Berkeley 2013
  - B.S. Engineering Physics
- Samsung Electronics 2014-2016
  - Image Processing Engineer
- I love exploring new places and activities!

Airflow to schedule hourly or daily batch jobs over the newly added/deleted source files.

Possible Improvements



Omar Benjelloun · Hector Garcia-Molina · David Menestrina · Qi Su · Steven Euijong Whang · Jennifer Widom

Swoosh: a generic approach to entity resolution

Received:

Abstract We consider the Entity Resolution (ER) problem (also known as deduplication, or merge-purge), in which records determined to represent the same real-world entity are successively located and merged. We formalize the generic ER problem, treating the functions for comparing and merging records as black-boxes, which permits expressive and extensible ER solutions. We identify four important properties that, if satisfied by the match and merge functions, en-

has different customer databases (e.g., one for each subsidiary), and would like to consolidate them. Identifying matching records is challenging because there are no unique identifiers across databases. A given customer may appear in different ways in each database, and there is a fair amount of guesswork in determining which customers match.

Deciding if records match is often computationally *expensive* and *application specific*. For instance, a customer in-

Possible Improvements