World of code:

Enabling a Research Workflow for mining and analyzing the universe of Open-source VCS data

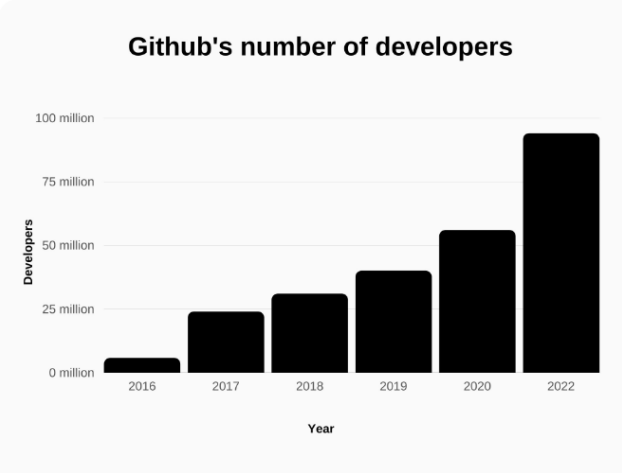
Original Authors: Yuxing Ma, Tapajit Dey, Chris Bogart, Sadika Amreen, Marat Valiev, Adam Tutko,David Kennard, Russell, Zaretzki, Audris Mockus

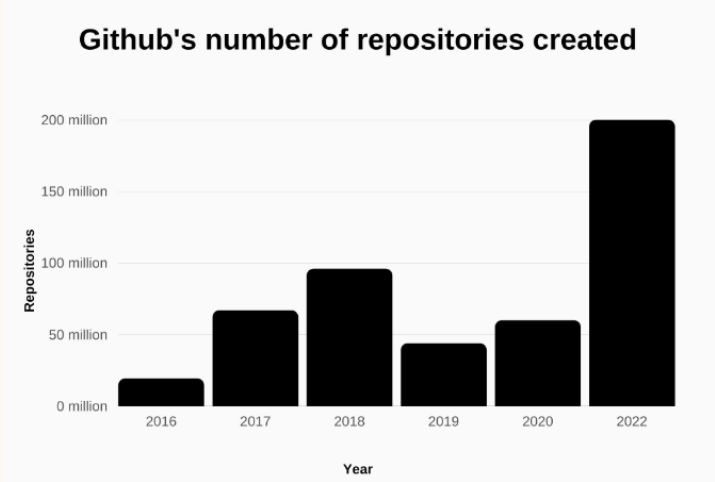
Presented by: Utkarsh Pratiush

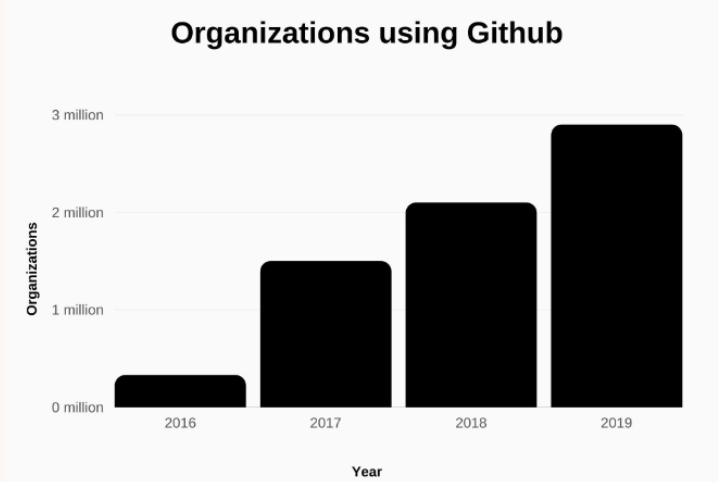
Structure

- Motivation
- Related work
- Background
- Methodology
- Application as discussed in paper
- Future work/challenges

Motivation







Want to answer questions like?

->**What are the top 10 most popular programming languages used in OSS?**

-> What are the largest open source projects by contributor count in GitHub?

At an organization level want to answer questions like?

->**Estimate productivity of employees**

->Assess potential vulnerability

https://www.usesignhouse.com/blog/github-stats

Related work and what WoC brings to table

several large-scale software mining efforts exist and may be roughly subdivided into attempts at preservation, data sharing for research purposes, and construction of decision support tools.

And many more...

WoC focuses on :
->Open source-related research
->Efficient update of this huge database that it builds on
->Preserve git objects and cross-references, so relationships between code and people is readily available.

background

- Git objects

objects:
1. Blobs
2. Trees
3. Commits
4. Tags

- Let’s take TensorFlow example

```
/my_tensorflow_project
    /models
        model.py
    main.py
    utils.py
```

Blobs: Each file (model.py, main.py, utils.py) is stored as a blob, containing the file's data.

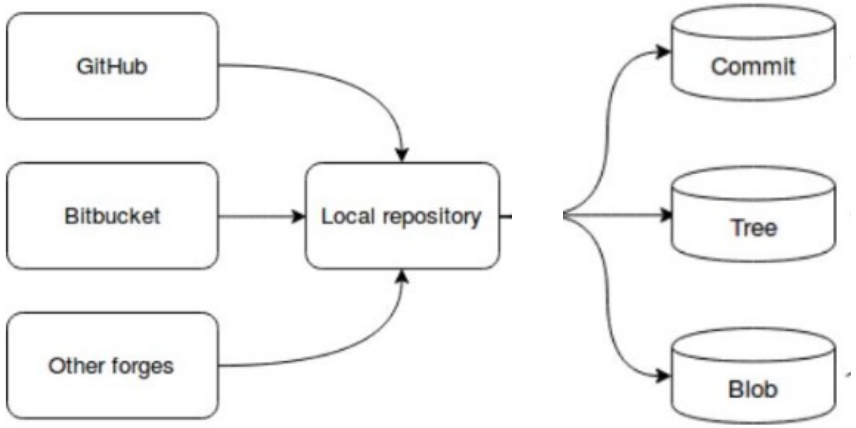
Trees: represent directories.
•Each tree object can point to blobs and other trees.
•models' tree:
•Root tree:

Commits: Current snapshot of Tensorflow project.

Tags: Tensorflow version 1.0: Specifies the release

More info: https://git-scm.com/book/en/v2/Git-Internals-Git-Objects

Methodology



Discover and retrieve

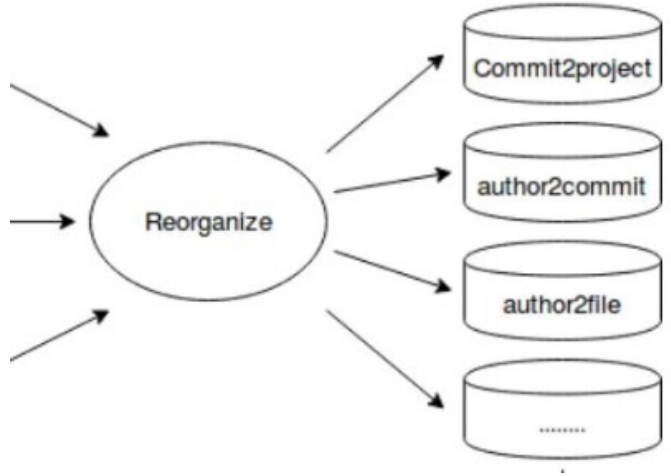
Extract: git objects

Git object retrieval
1.5PB →80TB
-> 95% reduction

Discovery:
a) Manual search
b) Html Parsing
c)Use of REST/GraphAL API's
d) Git project url

Once project discovered it can be retirved by cloning the repo. Due to huge number of repos
a) Need : High Network bandwidth and storage
b) Run multiple threads per server

We have stored the objects but how do we get info about project names, Authors and relationship?

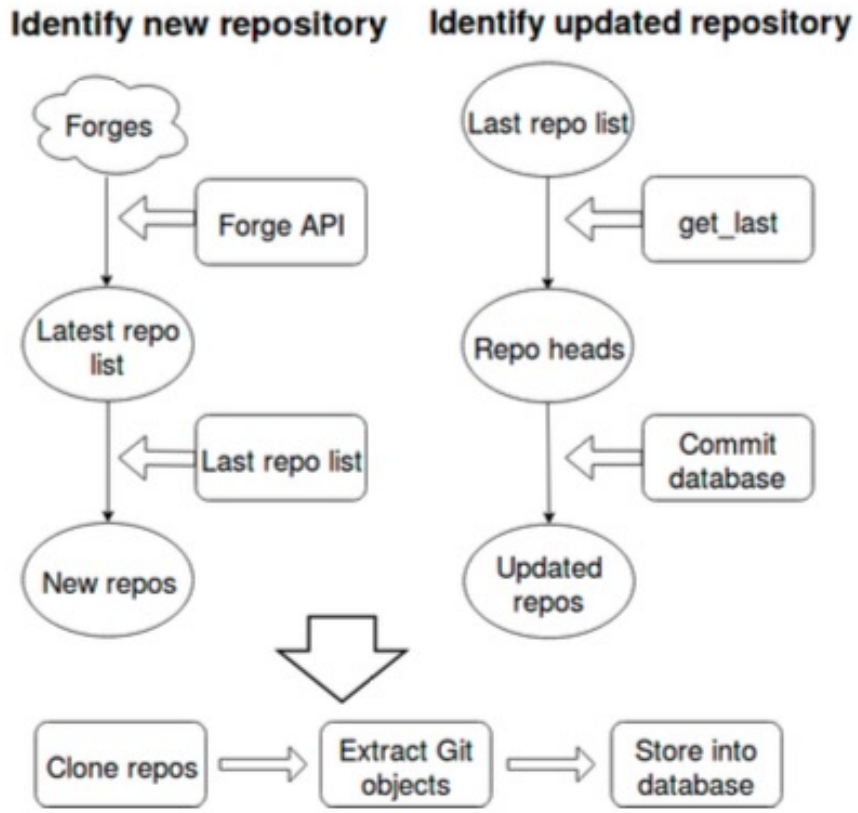


Reorganize

Results in a database with **12** mappings.

Data update

– Identify new repositories, clone and extract Git objects
– Identify updated repository and retrieve only newly added Git objects



Applications as described in paper

1. Use of programming languages

2. Correcting Developer Identity Errors

3. Cross-ecosystem comparison studies
    - File cloning across ecosystems
    - Developer migration across ecosystems

4. Python ecosystem analysis

5. Repository filtering tool

6. A number of research publications have utilized the WoC database

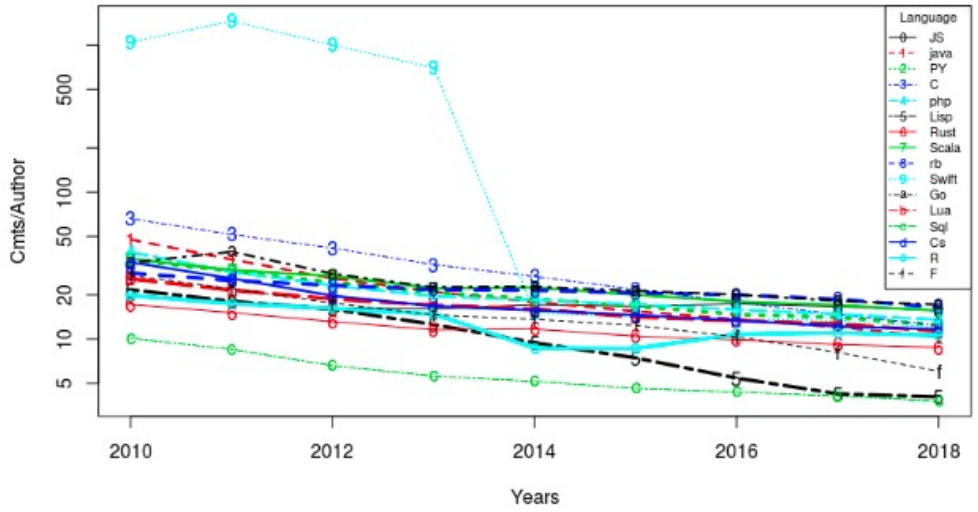


Fig. Productivity by Language

Further challenges/directions

1. Availability of API

2. Accommodate massive queries needs powerful hardware

3. Additional methods/procedures to improve acquisition: as project discovery procedure is only an approximation.

4. Reliably clean, correct and augment the collected data

Our Project

Server	OS	Memory GB	Storage TB	Storage Used TB	File System	Cores	CPU	CPU Codename
DA0	RHEL 7.9	384	35	31	/export/data	12	E5-2630	Sandy Bridge
DA1	RHEL 7.9	384	35	33	/export/data	12	E5-2630	Sandy Bridge
DA2	RHEL 7.9	384	35	35	/export/data	12	E5-2630	Sandy Bridge
DA3	RHEL 7.9	384	70	61	/export/data	8	E5-2623	Haswell
DA3			15	14	/fast			
DA4	RHEL 7.9	768	90	77	/data	8	E5-2623	Haswell
DA4			15	12	/fast			
DA5	RHEL 7.9	1280	123	115	/export/data	40	Intel Xeon Gold 6148	Skylake
DA5			56	54	/fast			
DA6	RHEL 7.9	256	90	77	/data	40	Intel Xeon Gold 6148	Skylake
DA7	Ubuntu 20.04.6	256	1700	464	/mnt/corrino/…	16	Intel Xeon Silver 4215R	Cascade Lake
DA8	Ubuntu 22.04.3	384	1700	306	/mnt/ordos/…	16	Intel Xeon Silver 4215R	Cascade Lake
ISAAC Legacy	RHEL 8.7		200	192	/lustre/haven/…			
ISAAC Next Gen	CentOS 7.6.1810		350	293	/lustre/isaac/…			
UT-StorR archival storage			TBD	0	/ut-storr/…			

Credits: Victor Hazlewood

Thanks