EXTRACTING DATA FROM YOUR OPEN SOURCE COMMUNITIES

Dawn M. Foster

PhD Student, University of Greenwich
Consultant, The Scale Factory

@geekygirldawn
dawn@fastwonder.com
fastwonderblog.com

WHOAMI

- **Geek, traveler, reader**
- **20 year tech career. Past 15 years doing community & open source (Intel, Puppet Labs, etc.)**
- **PhD student at University of Greenwich researching Linux kernel**
- **Community and open source consultant at The Scale Factory**

Photos by Josh Bancroft, Don Park

I 💖 METRICS GRIMOIRE

MailingListStats aka MLStats

CVSAnalY - repos

Bicho - bugs

More

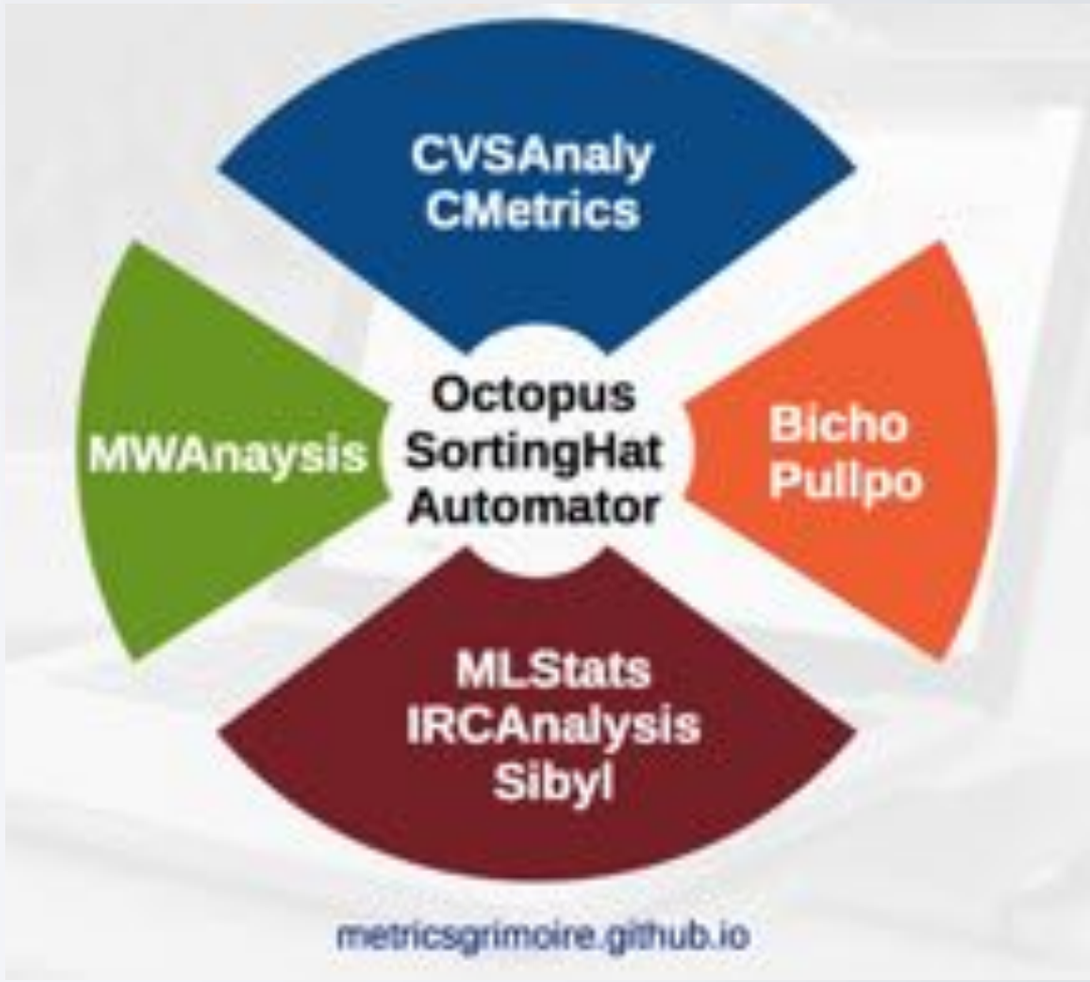


Photo by Bitergia

http://metricsgrimoire.github.io/

MLSTATS AND CVSANALY

a) Install

```
$ python setup.py install
```

b) Create database

```
mysql> create database mlstats;
mysql> create database cvsanaly;
```

c) Import data

```
$ mlstats http://URLOFYOURLIST
$ cvsanaly2 /path/to/repo
```

MLSTATS: EXTRACT DATA

Top 100 messages (most replied to threads):

```
SELECT subject, COUNT(*) as total
FROM messages
GROUP BY subject
ORDER by total DESC
LIMIT 100;
```

Other queries:

- # of messages from a specific person
- # of messages per person from email domain
- Find all messages with specific word in subject line (patch)

CVSANALY: EXTRACT DATA

Number of commits per person by email domain:

```
SELECT p.name, p.email,
COUNT(distinct(s.id)) as num_commits
FROM people p, scmlog s
WHERE email like "%company.com"
AND p.id=s.author_id
GROUP BY email
ORDER BY num_commits DESC;
```

Other queries:

- Top commit authors all time
- # of commits for specific person

OTHER GRIMOIRE OPTIONS

Bug data

Wikis

IRC

Aggregate across tools

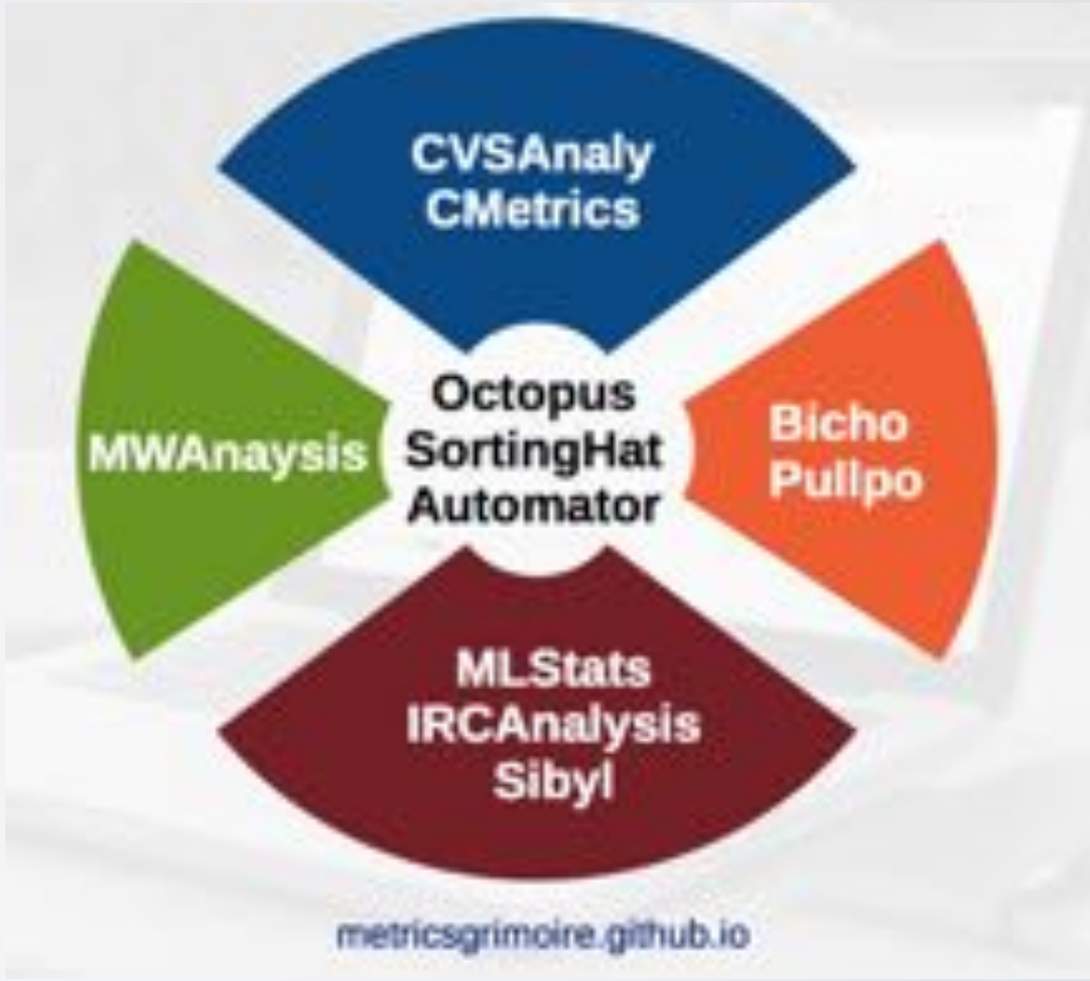
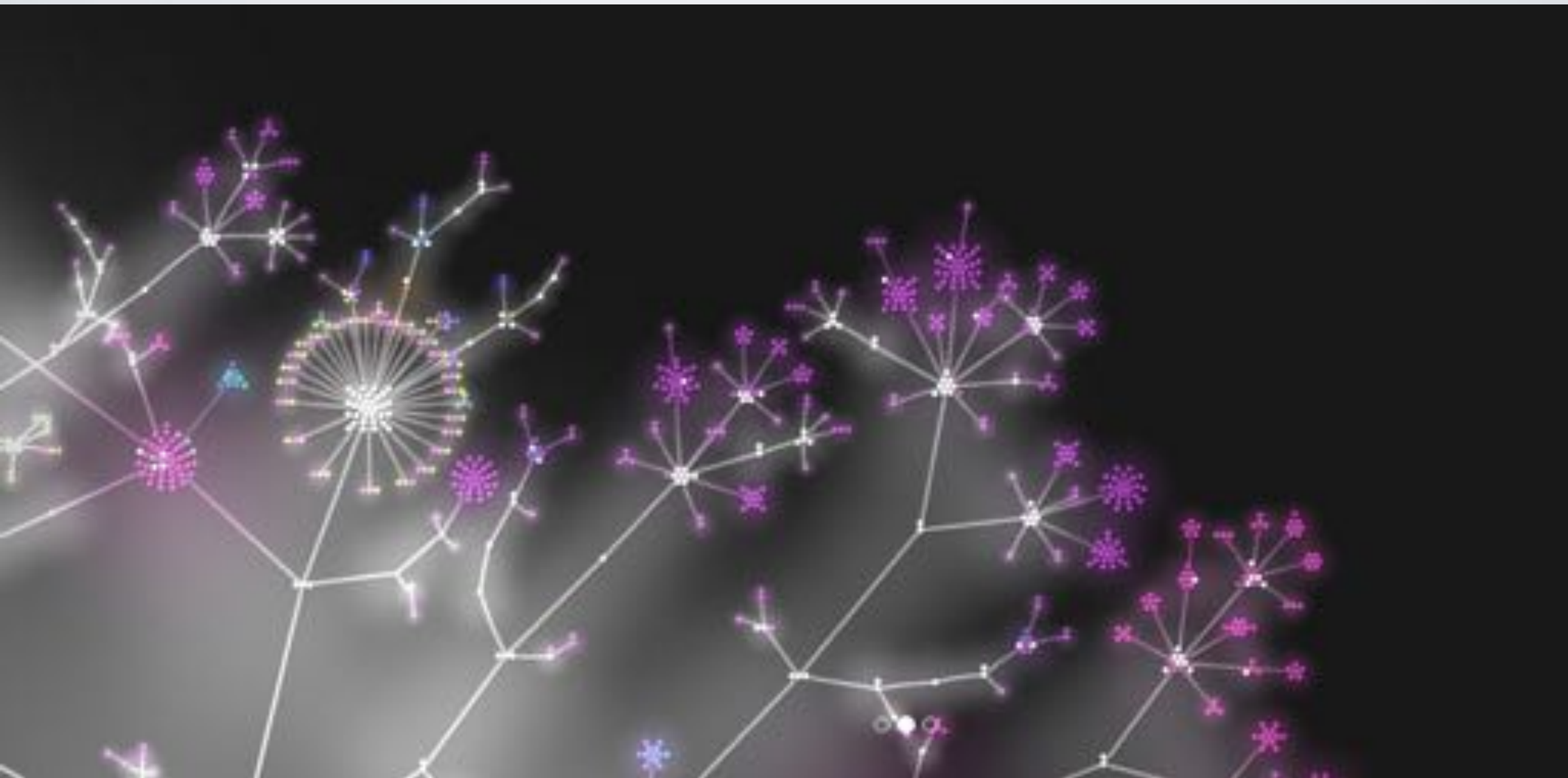


Photo by Bitergia

GOURCE

Visualize repository data using Gource

http://gource.io/



THANK YOU

Dawn Foster
PhD student, University of Greenwich
Consultant, The Scale Factory

@geekygirldawn, dawn@dawnfoster.com
fastwonderblog.com