Gender-diversity analysis of technical contributions

(In the Hadoop Ecosystem)

ApacheCon, Sevilla 2016

Daniel Izquierdo Cortázar
@dizquierdo
dizquierdo at bitergia dot com
https://speakerdeck.com/bitergia

Outline

Introduction

First Steps

Some numbers and method

Conclusions

Introduction

A bit about me

Why this analysis

What we have so far

/me

CDO in Bitergia, the software development analytics company

Lately involved in understanding the gender diversity in some OSS communities

Involved in some analytics dashboards: OPNFV, Wikimedia, Eclipse...

Disclaimer: *not involved in any working group, own analysis and interest, I may have missed some stuff...*

Why this study

Diversity matters

I attended some (Women of OpenStack) talks in the OpenStack Summit (Tokyo and Austin)

Produced some numbers that gained some attention: OpenStack and Linux Kernel

In the end this is all about **transparency** and improvement

We need data to make decisions

What we have so far

Diversity strategies ideas (from the ASF wiki)

Expected outcomes: Increase , retain and ***monitor*** diversity

Potential actions:

- *Reach out and attract new contributors*
- *Ensure people feel safe and appreciated*
- *Culture of inclusiveness and openness*

https://cwiki.apache.org/confluence/display/COMDEV/Diversity+Strategy+Ideas

Bitergia

What we have so far

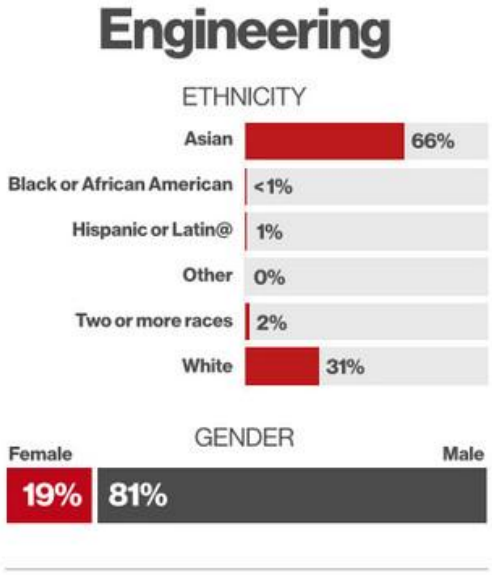
FOSS Survey in 2013:

- http://floss2013.libresoft.es/results.en.html
- 11% of women answered the survey

The Industry Gender Gap by the World Economic Forum.

- 5% for CEOs, 21% for Mid-level roles, 32% of Junior roles

Some companies



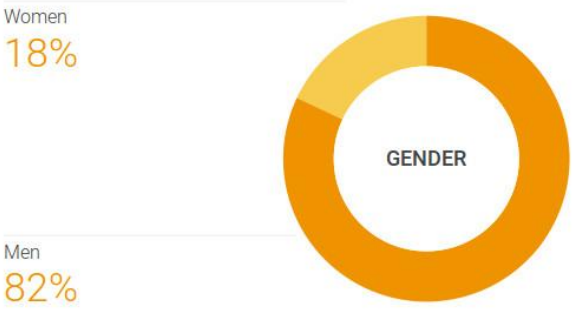
Pinterest Engineering focused employees.

https://blog.pinterest.com/en/our-plan-more-diverse-pinterest

Bitergia

Some companies

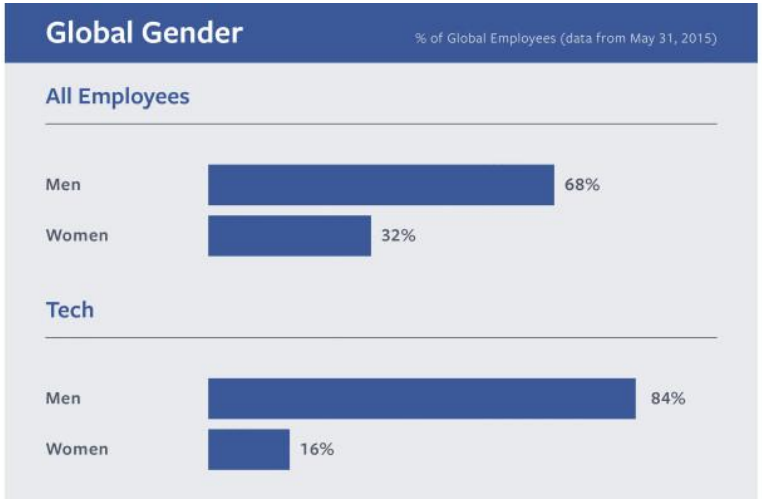
Google Tech focused employees.



http://www.google.com/diversity/

Bitergia

Some companies



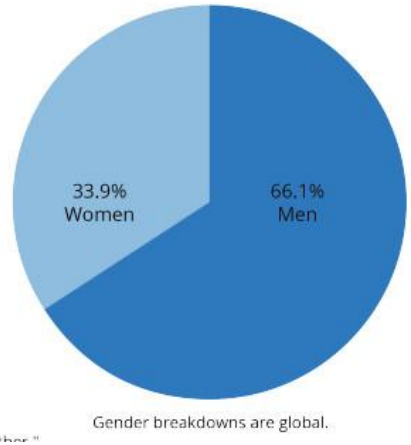
Facebook Tech focused employees.

http://newsroom.fb.com/news/2015/06/driving-diversity-at-facebook/

Bitergia

Some companies

Dropbox all employees.

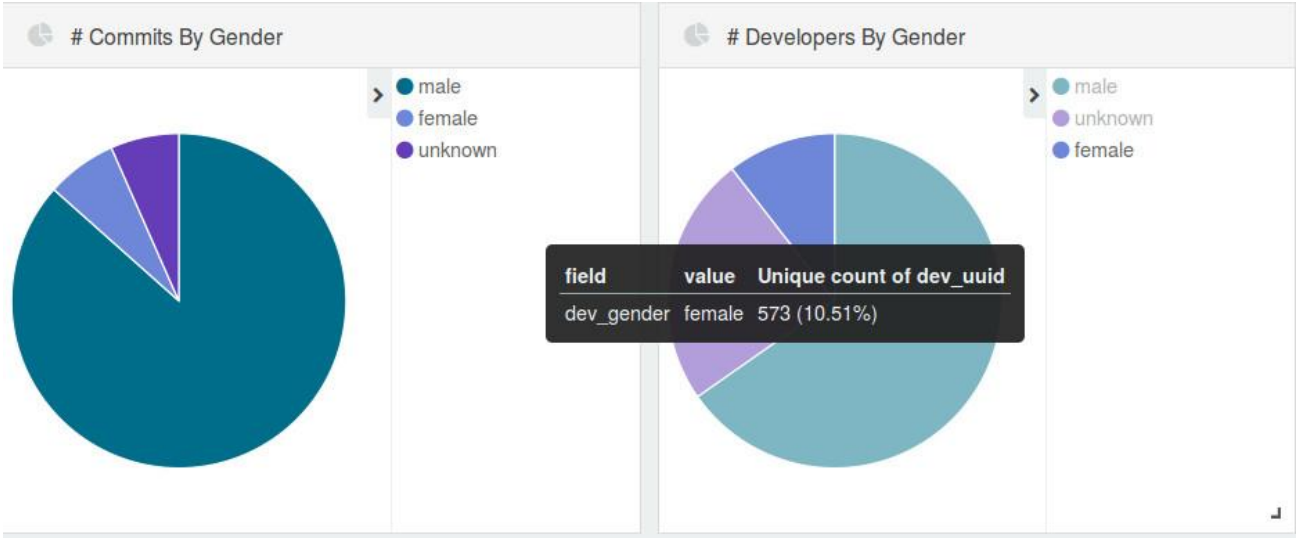


https://blogs.dropbox.com/dropbox/2014/11/strengthening-dropbox-through-diversity/

Bitergia

OpenStack (Austin) numbers

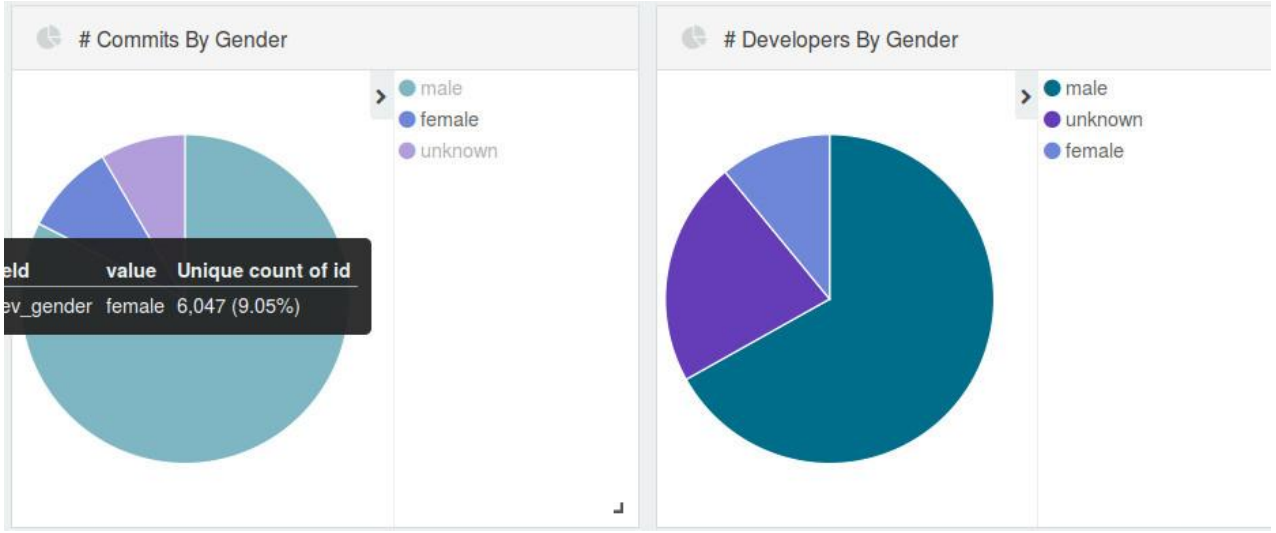
Women activity (**all of the history**):
~ 10,5% of the population (~ 570 developers)
~ 6,8% of the activity (>=16k commits)



Bitergia

OpenStack (Austin) numbers

Women activity (**last year**):
~ 11% of the population (~ 340 active developers)
~ 9% of the activity (>=6k commits)



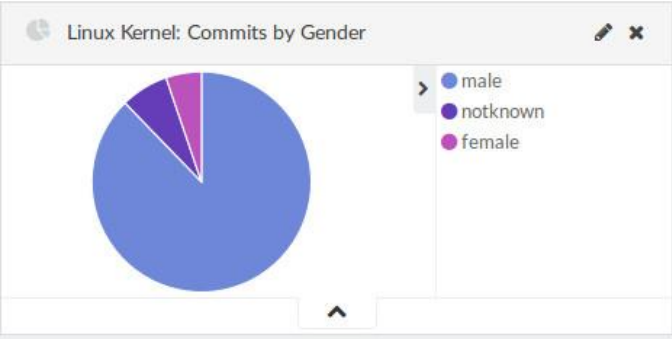
Bitergia

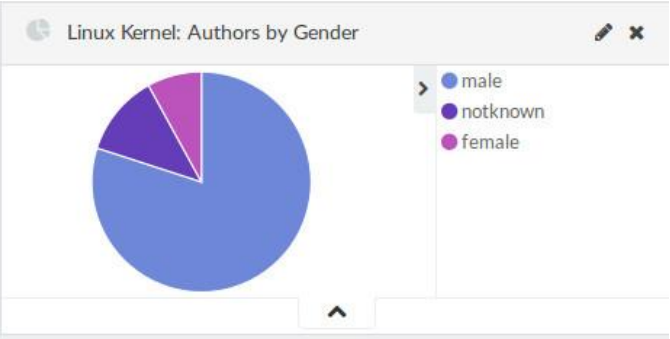
Linux Kernel Numbers

Women activity (since 2005):

~ 5.2% (> 31K commits)

~ 8% of the population (~ 1,15K developers)





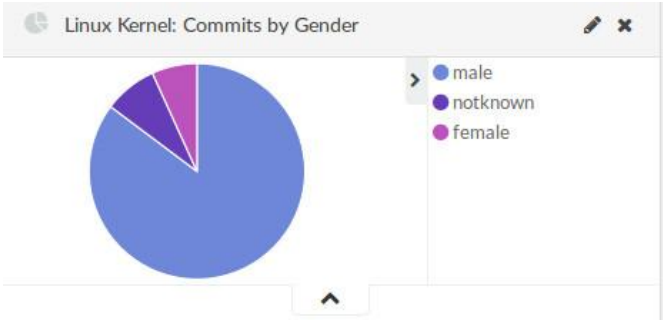
Bitergia

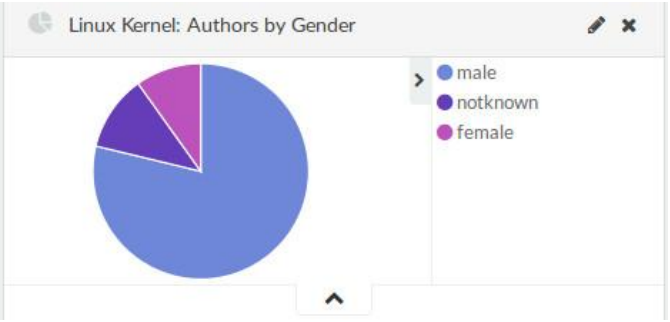
Linux Kernel Numbers

Women activity (last year):

~ 6.8% of the activity (~ 4k commits)

~ 9.9% of the population (~ 330 active developers)





Bitergia

Summary

Conclusions not representative, but:

- Women represents around 30%/40% of the workforce in tech companies.
- And between 10% and 20% if focused on tech teams.
- OpenStack shows a 11% of the population
- Linux Kernel shows a 10% of the population
- What about some projects in the ASF?

First Steps

Some Definitions

Contribution: commit

Other potential metrics: diversity by company, fairness in the code review among organizations and genders, transparency in the process

Available but sensitive info: affiliation, countries, time to review

Focus on the Hadoop ecosystem

First Steps

Names databases

Genderize.io

Manual analysis

Focus on main developers

Architecture

Original Data Sources

Mining Tools

Perceval

Info Enrich.

Genderize.io

Pandas

Manual work

Viz

ElasticSearch
+
Kibana

Bitergia

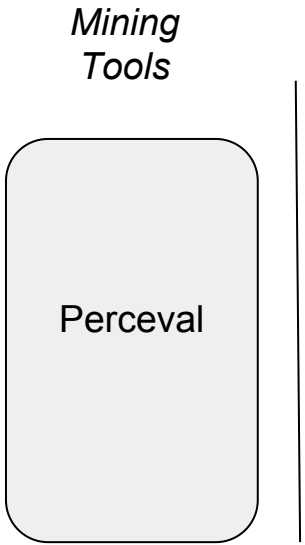
Architecture

Original Data Sources

- Git
- 14 projects:
- > 190K commits
- > 1.7K developers
- Info from Hadoop and related projects (http://hadoop.apache.org/)

Bitergia

Architecture



- Produces JSON documents from the usual data sources in OSS
- Part of the GrimoireLab toolchain
- grimoirelab.github.io

Bitergia

Architecture

Info
Enrich.

Genderize.io

Pandas

Manual work

- Genderize.io: name database
- Pandas: data analysis lib.
- Ceres library (dicortazar/ceres @ github)
- Manual work:

Bitergia

Architecture

Viz

ElasticSearch
+
Kibana

- ElasticSearch: Schemaless db
- Kibana: works great with ES
- This tandem helps a lot to verify info
- Drill down capabilities
- Extra info available (but not displayed)

Bitergia

Validation: manual work

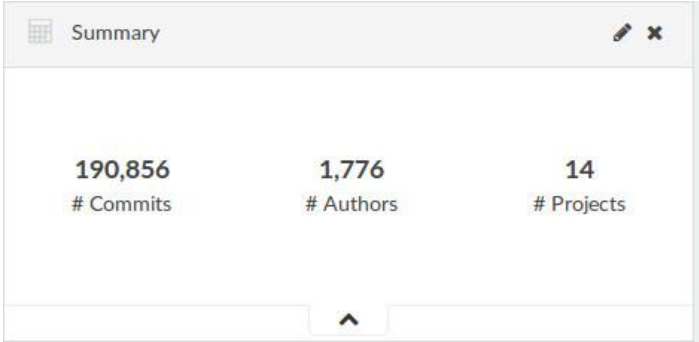
Check main contributors by hand

Asian names hard to check (u_u) (help needed!)

Some numbers

Git Contributions

Git Overview



- Aggregated historical data

Bitergia

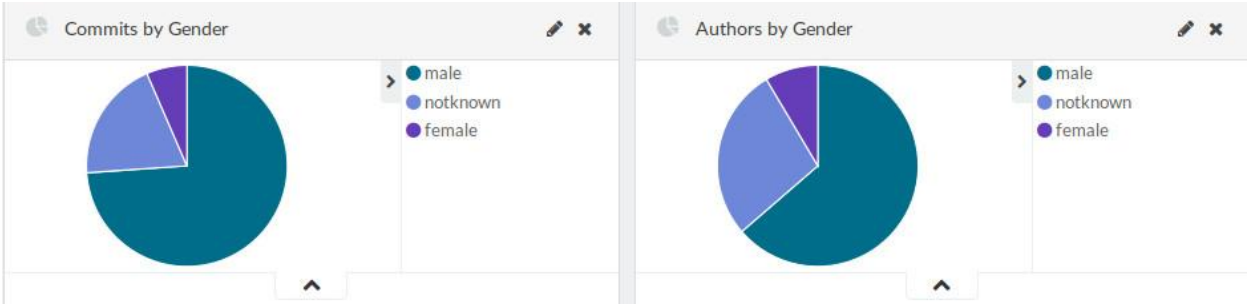
Git Activity and Population

Women activity (**all history)**:
8.8K commits (4.6% of activity)
129 (7.5% of population)

Bitergia

Git Activity and Population

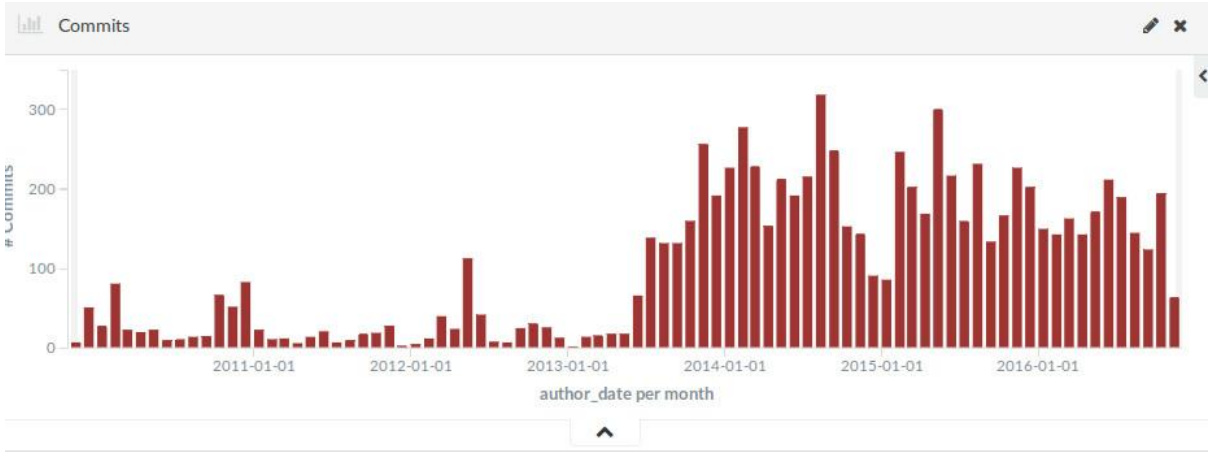
Women activity (**last year**):
~2K commits (6.5% of the activity)
71 developers (8.5% of the population)



Bitergia

Git Activity Women Evolution

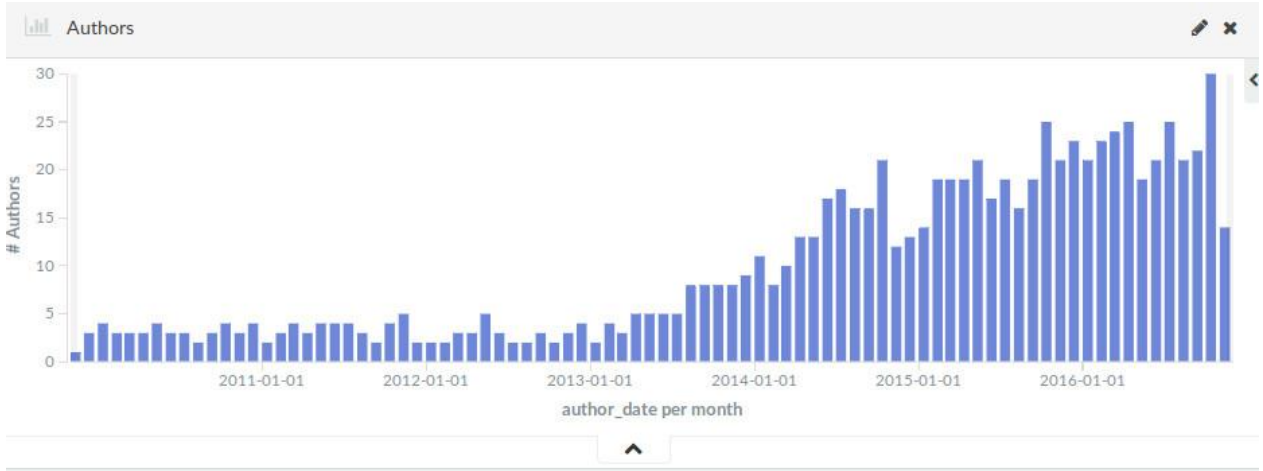
- In line with the general trend: stable and small activity till mid 2013, then a jump and stable in 2016



Bitergia

Git Authors Women Evolution

- Continuous increase after 2013
- Interesting pattern: peak of authors in October 2014, 2015 and 2016
  - Any idea?



Bitergia

Where are they based?

- Mainly in the west coast and then Europe
- Asia may be under represented

Bitergia

The most diverse projects

- Interesting to look for the best practices and learn from those
- This may be biased by external factors I'm not aware of (eg: version control system migrations…)

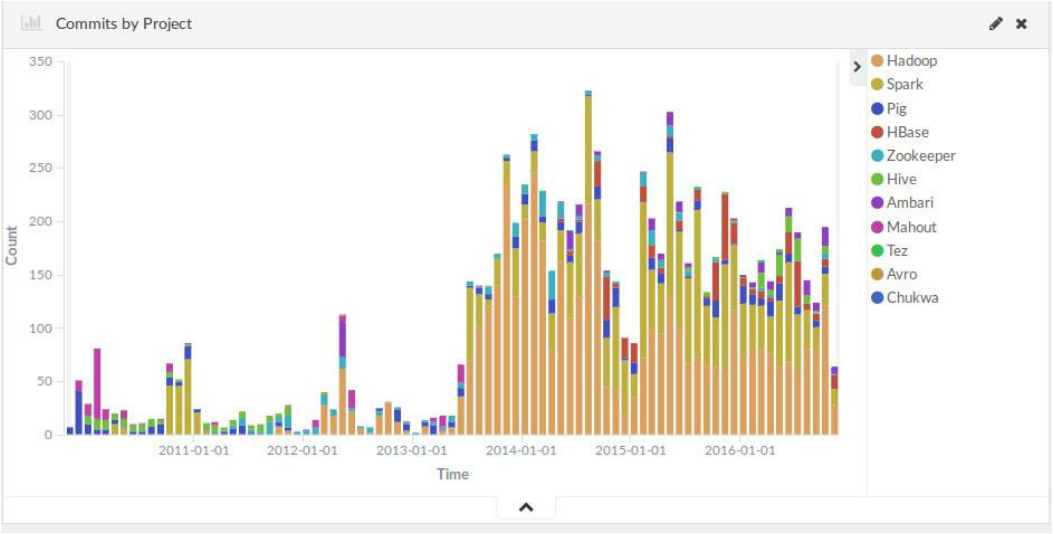
Projects

Project	Gender	# Commits	# Authors
Hadoop	female	4,474	17
Spark	female	2,396	84
Pig	female	647	3
HBase	female	412	15
Zookeeper	female	327	3
Hive	female	241	9
Mahout	female	211	3
Ambari	female	205	12
Tez	female	6	1
Avro	female	5	1
Chukwa	female	3	1

All Contributors:
Hadoop
HBase
Ambari
Spark
Hive
Pig
Mahout
Tez
ZooKeeper
Avro
Chukwa

The most diverse projects

- The jump in the activity after 2013 is due to mainly Hadoop and Spark



Bitergia

The most diverse projects

- Well, we should look at the relative numbers...

Zookeeper: 13.6%
Pig: 13.5%
Spark: 8.3%
Mahout: 5.5%
Hadoop: 5.3%
Hive: 1.8%
HBase: 1.5%

The rest of them < 1%

The most diverse projects

- So Zookeeper, Pig and Spark are the champions in diversity
- What can we learn from them?
- Are there specific policies focused on diversity in these projects?
- Is this more a matter of the community or the companies involved in the project?

Conclusions

Comparison with OpenStack/Kernel

Data to Make Decisions

Open Paths

OpenStack/Kernel/Hadoop Eco.

Last year women activity in OpenStack
~ 9% of the activity (>=6k commits)
~ 11% of the population (~ 340 active developers)

Last year women activity in the Linux Kernel
~ 6.8% of the activity (~ 4k commits)
~ 9.9% of the population (~ 330 active developers)

Last year women activity in the Hadoop ecosystem
~ 6.5% of the activity (~ 2K commits)
~ 8.5% of the population (~ 70 active developers)

How can be this used?

From the *diversity strategy ideas* wiki:

Go to where our potential new contributors are (Outreachy, GSoC, Women in Big Data, ...)

- Are you measuring success and retention in Outreachy?

This data may help to measure attraction and retention rate

The analysis can be extended to all of the ASF projects

How can be this used?

From the *diversity strategy ideas* wiki:

Make communities welcoming and inclusive (help newcomers, acknowledge contributions, there are several ways to contribute)

- How do you measure this? How to you make a distinction between a first email and a first piece of code? (identities identification issues)

Demographics study may help with this challenge

Other questions to have in mind

Organizations are a great way to bring women to the community, foster their participation and help them to be more diverse and inclusive.

Keep in touch with developers that used to work in the community. I'd say this is as important as welcoming newcomers!

Further Work

Sensitive info: dashboard still private

Extra analysis: time to merge **fairness**, **companies** women %, **Outreachy** follow ups, **quarterly** reports, updated data, specific policies **ROI** and others.

This [hopefully] helps to have a better picture

Other minorities analysis could be done

Gender diversity is not binary

Conclusions

Room for improvement of the dataset

This provides some initial numbers about the current status

Hopefully useful for the ASF

Gender-diversity analysis of technical contributions

(In the Hadoop Ecosystem)

ApacheCon, Sevilla 2016

Daniel Izquierdo Cortázar
@dizquierdo
dizquierdo at bitergia dot com
https://speakerdeck.com/bitergia