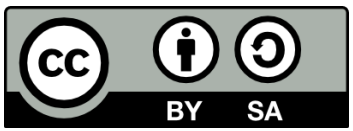
Applying Text Retrieval and NLP to Software Engineering Data

2016 SATToSE hackathon

Gregorio Robles

grex@gsyc.urjc.es
Universidad Rey Juan Carlos (Madrid, Spain)

Bergen, July 11th 2016



Who has been at the SATToSE hackathon this summer?

Goal of the hackathon

Apply the TR and NLP techniques learned in the tutorial to software engineering data from a FLOSS project

MetricsGrimoire Toolset

- Versioning: CVSAnalY (`metricsgrimoire.github.io/CVSAnalY`)
- Bug-tracking: Bicho (`metricsgrimoire.github.io/Bicho`)
- Mailing lists: MailingListStats (`metricsgrimoire.github.io/MailingListStats`)
- *et al.*

The complete suite can be accessed at
`metricsgrimoire.github.io`

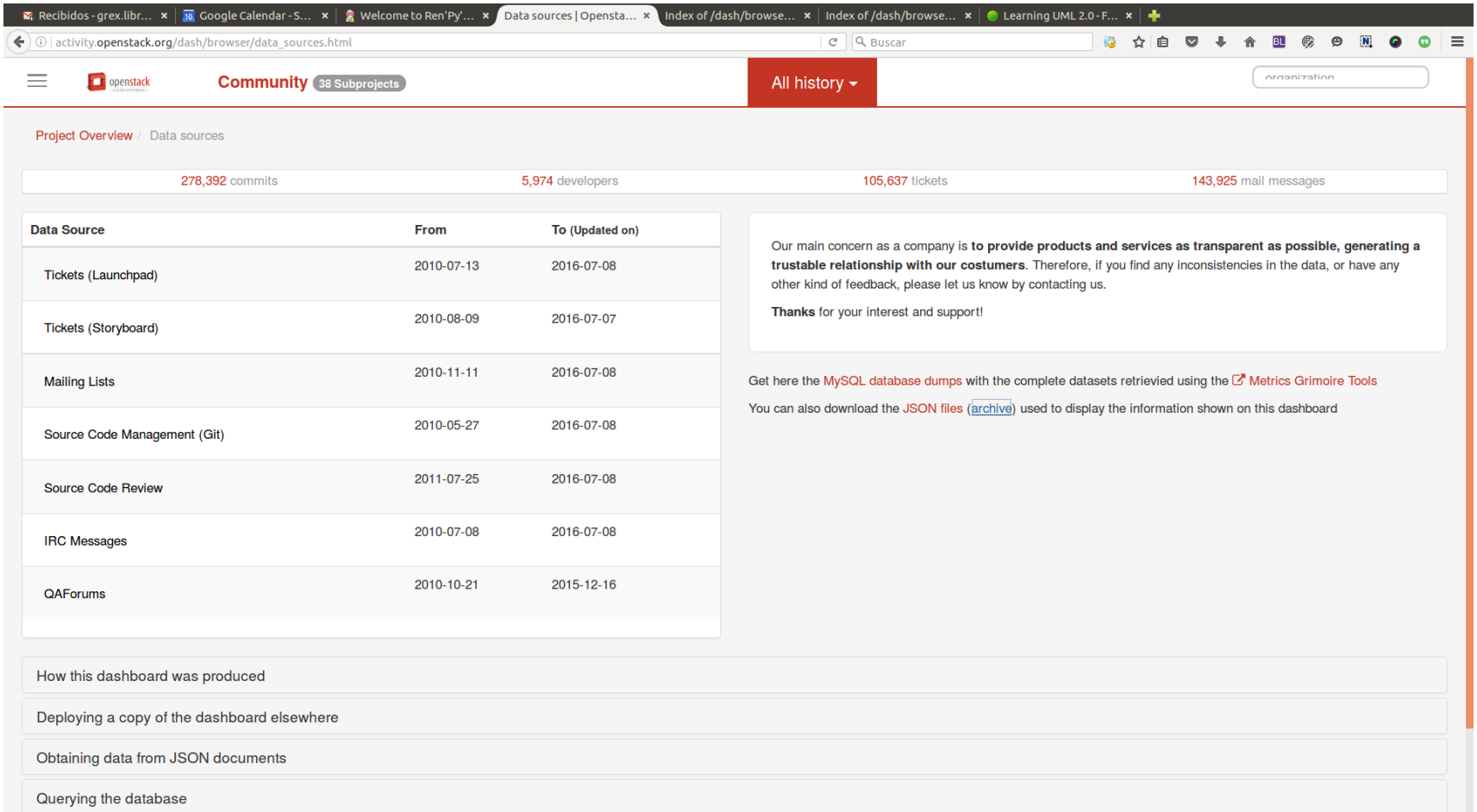
Bitergia Datasets

- Based on the MetricsGrimoire toolset, Bitergia analyzes FLOSS projects
- Results are offered in web-based dashboards
- Data used for creating these dashboards is publicly available
- ... and will be the basis of this hackathon!

OpenStack

- FLOSS platform for cloud computing
- Over 10K developers and/from 500 companies work on it
- Heavy use of code review and modern FLOSS development techniques
- Written mainly in Python

Bitergia's OpenStack Dashboard data sources



`activity.openstack.org/dash/browser/data_sources.html`

Data, data, data... (I)

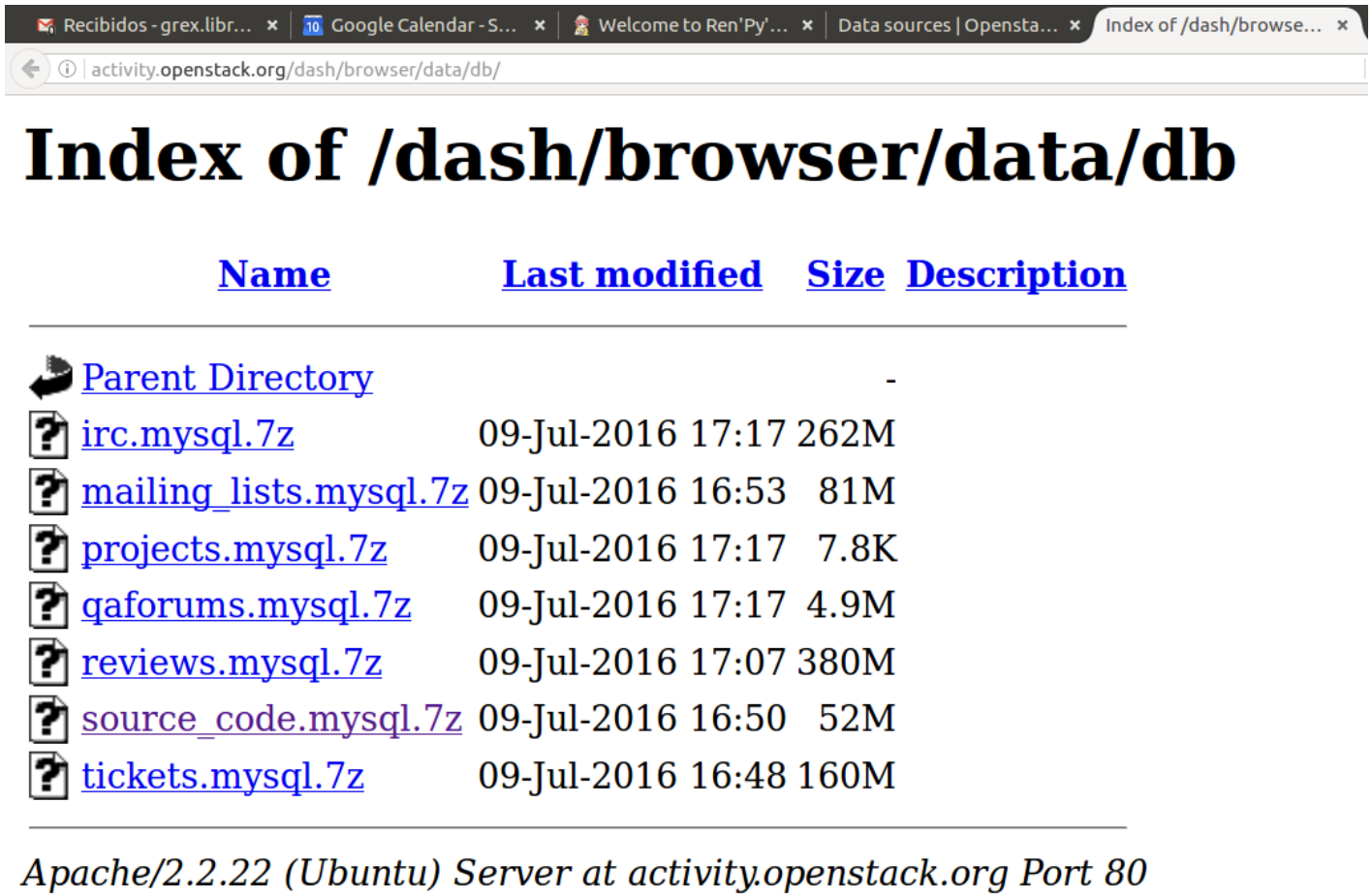
- Updated daily!
- Offered in two (really three) different formats:
  1. (compressed) SQL (as a MySQL dump)
  2. JSON
     - However contains only statistics
  3. (For the hackathon, I transformed the data partially to) CSV

Data, data, data... (and II)

1. 9,584 questions, 9,661 answers and 13,841 comments
2. 108,066 tickets and 519,066 comments on them
3. 150,963 mailing list messages (including content)
4. 272,912 reviews and 6,838,049 comments on them
5. 549,116 commits meta-data (including message)
6. 22,385,190 messages on IRC
7. ... but no code (you can clone it from `https://github.com/openstack`)

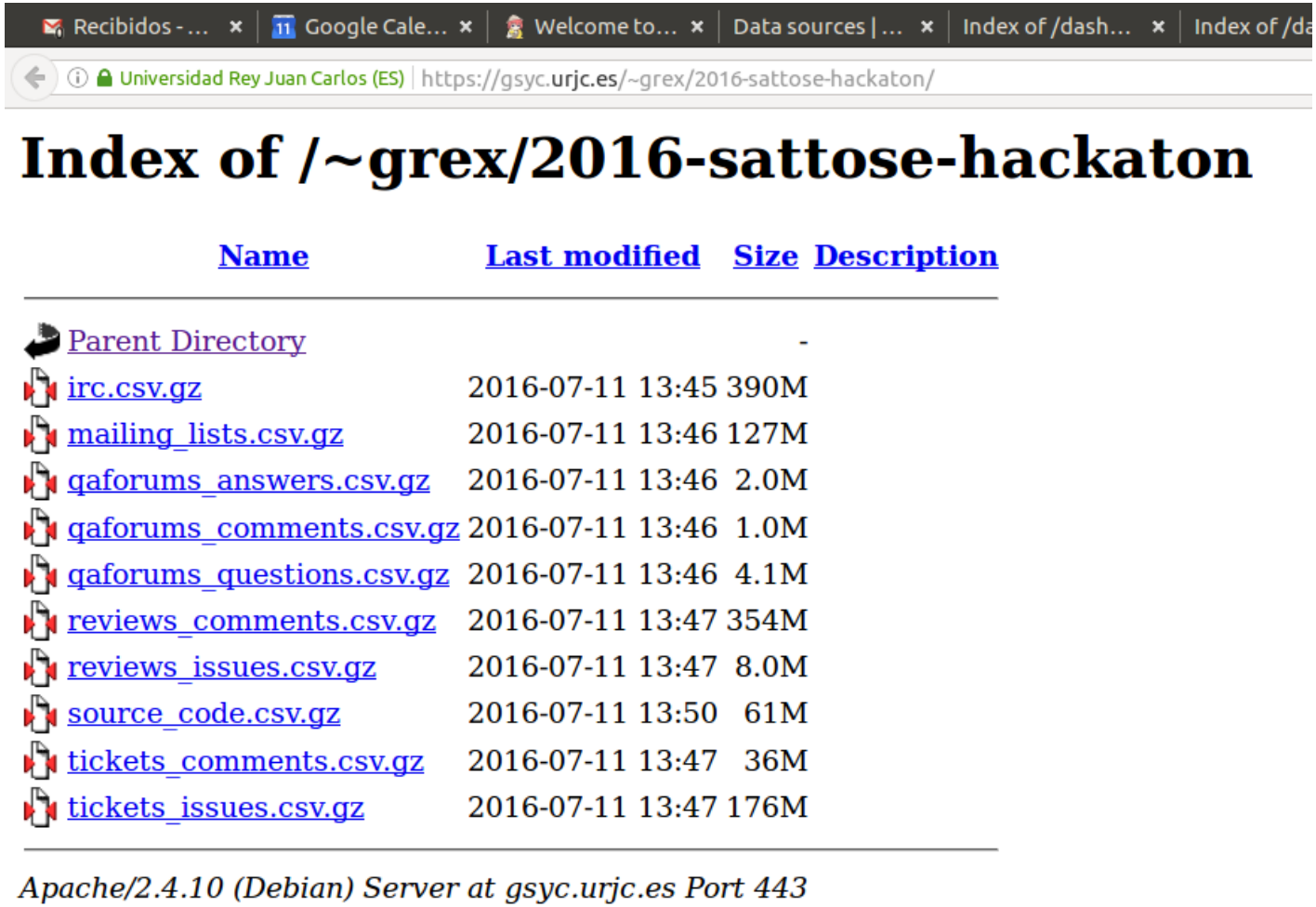
hackathon side activity?

SQL data



http://activity.openstack.org/dash/browser/data/db/

CSV data (I)



http://gsyc.urjc.es/~grex/2016-sattose-hackathon/

CSV data (and II)

To ease *playing* with the data and getting it known:

- First 10,000 lines of each CSV file
  - Really, header line + 9,999 lines with content
- Compressed and uncompressed
- All: all_10000s.tar.gz

`http://gsyc.urjc.es/~grex/2016-sattose-hackathon/10000s`

What shall I do? It depends!

Possible hackathon question #1

Extraction of Domain Concepts from the Identifiers

[parse source code, extract certain identifiers, process identifiers, apply POS tagging, VSM with TFIDF. Preprocessing may be required]

(Acknowledgement goes to S. L. Abebe, P. Tonella - WCRE, 2011)

Possible hackathon question #2

Inconsistencies among the name, type, and comment of a programming entity (methods and attributes)

[parse source code, map comments to entities, split identifiers, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)]

(Acknowledgement goes to V. Arnaoudova, M. Di Penta, and G. Antoniol. - JEMSE 2015)

Possible hackathon question #3

Do developers speak/cite/talk about scientific (i.e., research) literature in their development?

[parse irc/mailing lists/commit messages, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)... Preprocessing may be required]

(Acknowledgement goes to Tom Mens)

Possible hackathon question #4

Do developers speak/cite/talk about architecture and/or modeling in their development?

[parse irc/mailing lists/commit messages, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)... Preprocessing may be required]

(Acknowledgement goes to Michel Chaudron)

Possible hackathon question #5

How do concepts (specific methodologies, etc.) expand in a project?

[parse irc/mailing lists/commit messages, apply POS tagging (Stanford POS tagger) and semantic analysis (WordNet)... Preprocessing may be required]

(Acknowledgement goes to me ;)...)

Go hack!

(There will be a small present for participants!)

Hack! Links where to start from

Tools

http://metricsgrimoire.github.io/

Code

https://github.com/openstack

SQL

http://activity.openstack.org/dash/browser/data/db/

CSV

http://gsyc.urjc.es/~grex/2016-sattose-hackathon/

CSV - first 10,000 lines

http://gsyc.urjc.es/~grex/2016-sattose-hackathon/1000s

Applying Text Retrieval and NLP to Software Engineering Data

2016 SATToSE hackathon

Gregorio Robles

grex@gsyc.urjc.es
Universidad Rey Juan Carlos (Madrid, Spain)

Bergen, July 11th 2016