开源社
kaiyuanshe

2016中国开源年会

China Open Source Conference 2016

开源指南针

基于深度学习的源代码及开发人员的大数据分析

目录

1. 项目简介
2. 演示
3. 技术讨论
4. 开源社区排行榜
5. 计划
6. 加入我们吧！

开源项目有利的情况

种类数量

开源项目种类繁多，不同的实现方式、复杂度、考虑角度等让使用者可以依据自身条件享有丰富的选择

开发质量

开源项目多由国内外开发水平高、对架构与解决方式有独到见解的开发人员建构

开发速度

开源项目的积累与改动可能比使用者独立开发更加快速，测试应用场景也更加丰富

学习使用

项目的深度与广度比初接触实现领域的开发人员具有参考学习使价值

开源项目面临的困难

种类数量

开源项目种类繁多、没有妥善分类搜索，不容易有效找到同类的项目于以比较优劣和评估是否适合使用

开发质量

开源开发人员参差不齐，可能形成代码品质、文档说明、维护工作等没有保障的情况

开发速度

开源项目多半针对广大需求进行实现，使用者必须对此项目足够了解才能使用或修改代码用于自身项目

学习使用

开源代码量大，缺乏有效的文档说明，学习代码的时间长，没有效率

开源指南针

- 透过对开源项目信息和其与他项目的关联所得的综合分析，提供给使用者以分类、搜索、推荐等功能在数量庞大的开源项目中快速定位到相关的项目。
- 运用分析开源项目包括项目说明、代码、贡献者等信息和其与他项目的关联程度，提供搜索、分类、排名、评比、可视化等功能。

The code galaxy

name | type (user / repo) | search | reset | back

ronreiter

Ron Reiter (null) from:Israel
public repos: 42; followers:72
updated/created: 2016-02-27/2011-04-02

name type (user / repo) search

dspezia

Didier Spezia (Amadeus) from:French Riviera
public repos: 4; followers:10
updated/created: 2016-02-13/2011-03-05

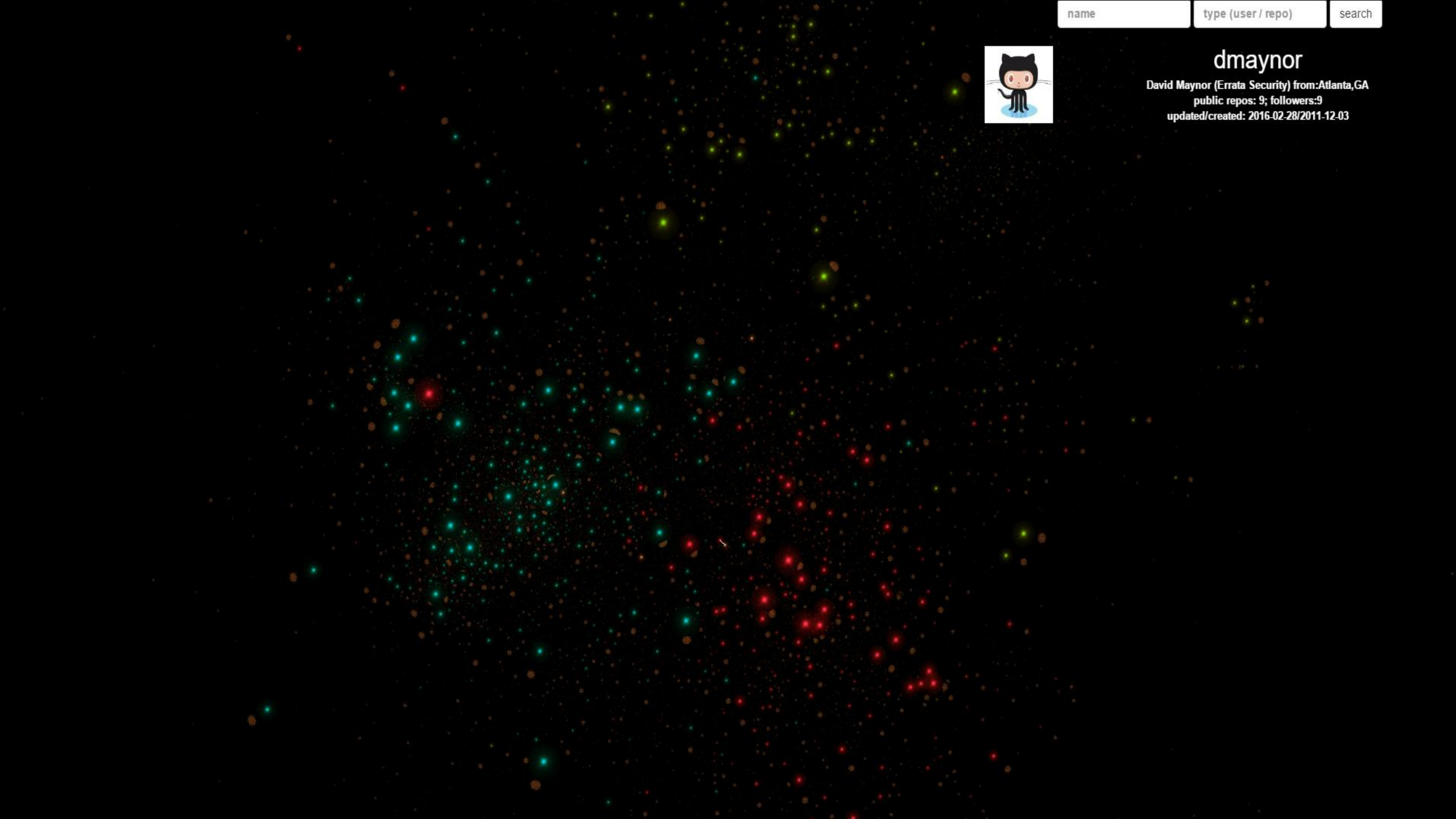
详细信息

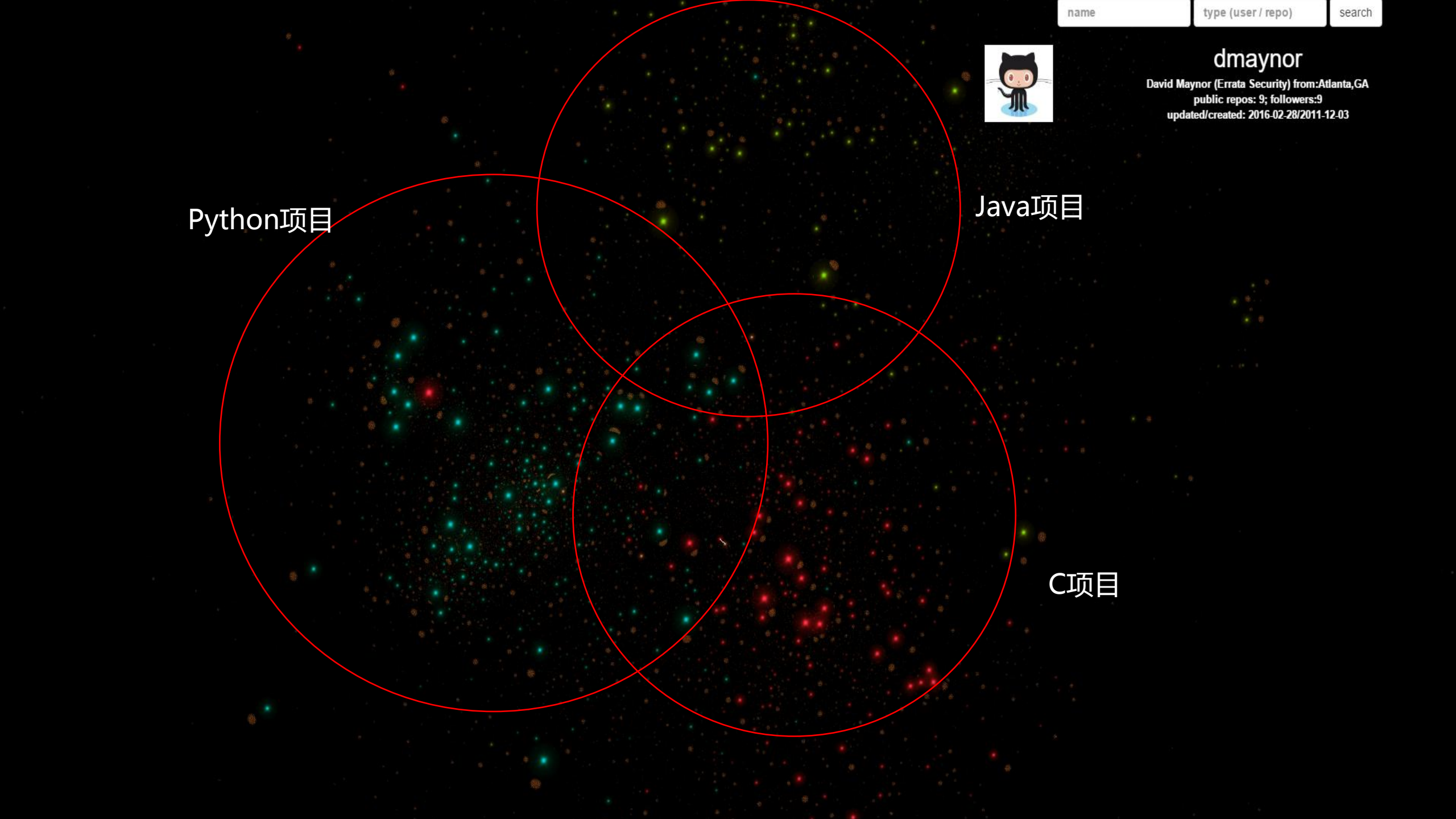
Java项目

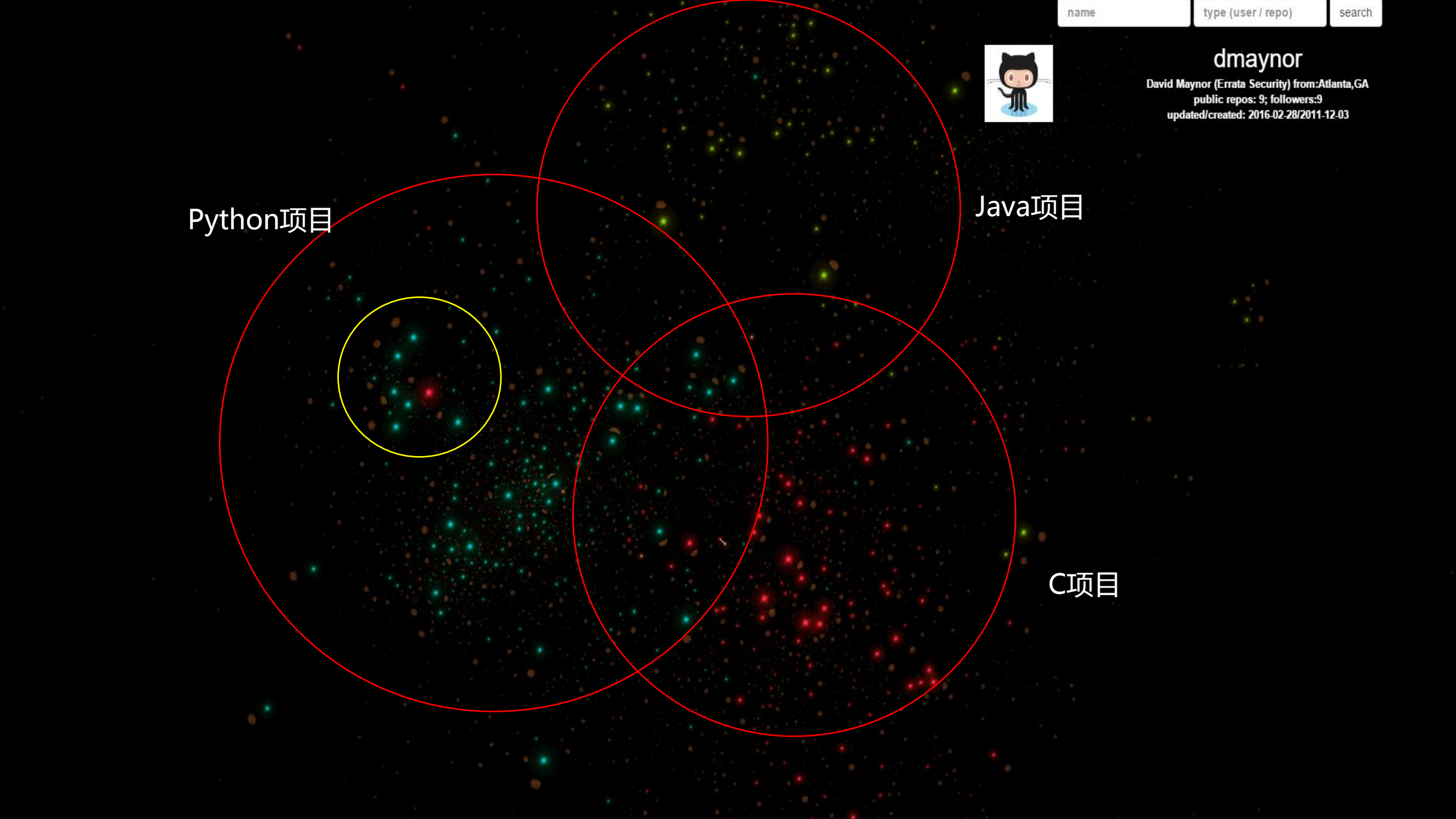
开发者

Python项目

C项目


name
type (user / repo)
search
dmaynor
David Maynor (Errata Security) from:Atlanta,GA
public repos: 9; followers:9
updated/created: 2016-02-28/2011-12-03


name
type (user / repo)
search
dmaynor
David Maynor (Errata Security) from:Atlanta,GA
public repos: 9; followers:9
updated/created: 2016-02-28/2011-12-03
Python项目
Java项目
C项目

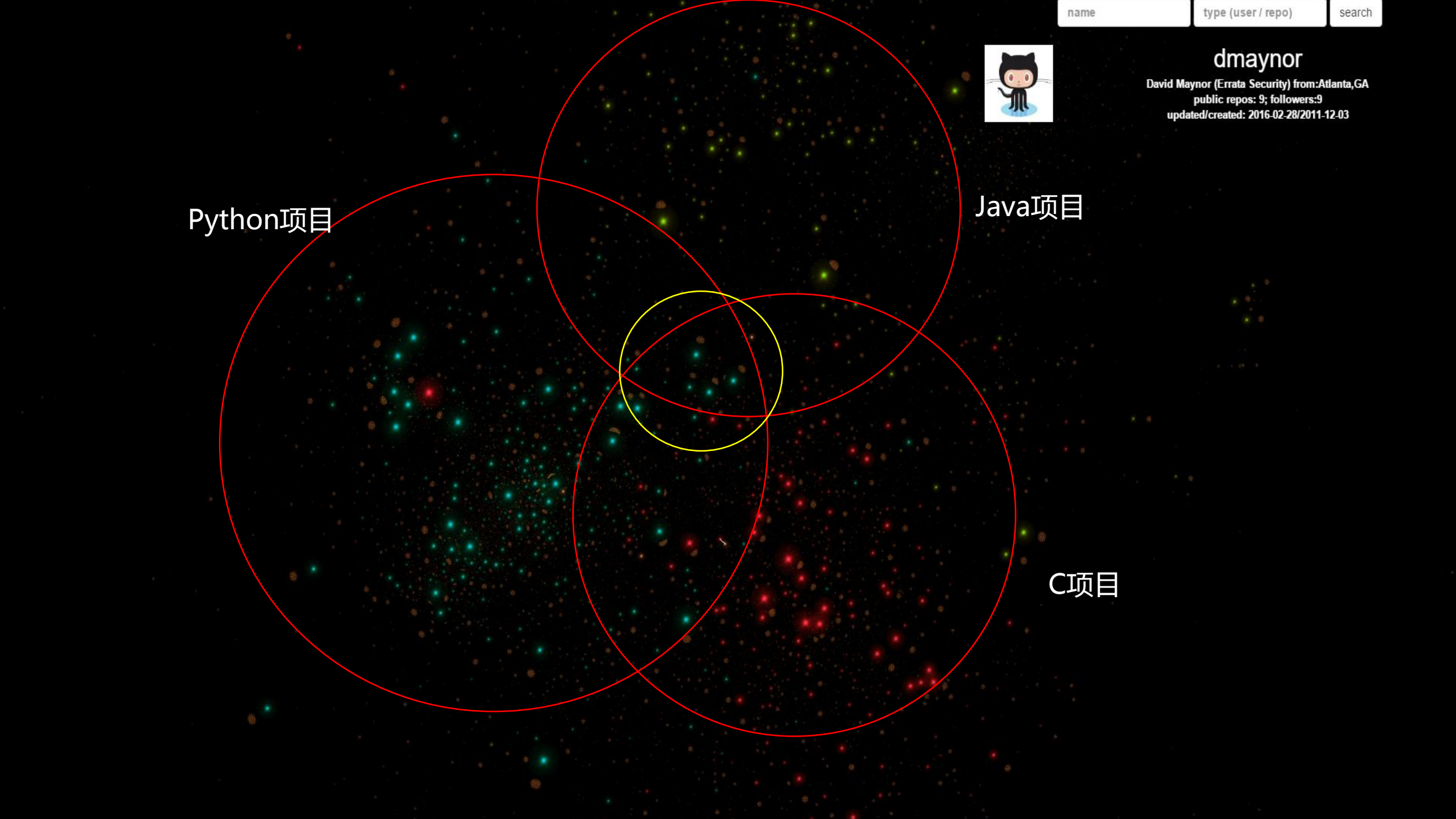

name
type (user / repo)
search
dmaynor
David Maynor (Errata Security) from:Atlanta,GA
public repos: 9; followers:9
updated/created: 2016-02-28/2011-12-03
Python项目
Java项目
C项目

Guess the red one

name | type (user / repo) | search | reset | back

sagi

Sagi Kedmi (IBM Security) from:null
public repos: 17; followers:11
updated/created: 2016-03-05/2013-12-31


name
type (user / repo)
search
dmaynor
David Maynor (Errata Security) from:Atlanta,GA
public repos: 9; followers:9
updated/created: 2016-02-28/2011-12-03
Python项目
Java项目
C项目

The center of the universe

cicku | user | search | reset | back

probablycorey/seriously

The Objective-C HTTP library that Apple should have created, seriously.

language:C size:529

forks:36 subscribers:5 watchers:463

updated/created:2016-03-04/2010-06-28

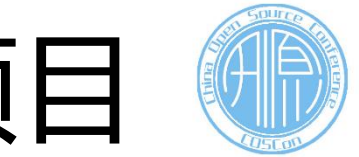
技术讨论

开源指南针
Compass

工作原理

- 使用网络爬虫技术获取开源代码库相关的数据如原代码，项目信息等。
- 将项目的静态信息如项目名称、起始日期等与动态信息如项目下载次数、客户喜爱程度等储存在数据结构中。
- 处理原始数据与结构化的处理单元，生成如分类、统计、排序、标签等综合数据。
- 使用机器学习、自然语言处理、大数据分析等技术，综合所有数据提供如分类、搜索、推荐、评比等应用服务。

应用场景-以Java 版本搜索开源项目



Github代码库没有java版本信息，使用者需要下载整个项目使用特定的java版本进行编译，才知道此项目是否符合版本的需求。

本项目分析java语言的特性，在github管理的代码加注java的版本，用户可以java的版本作为搜索的条件，找出合条件的项目供用户使用。

文本数据挖掘 vs 定量数据挖掘

- 常见的文本聚类算法：Lemur, BOW toolkit
- 定量数据聚类算法经过调整可以用于文本聚类：
  - 文本数据维度很大，但是基础数据是稀疏的.
  - 文本词汇量非常大，但是其中的概念数量小于特征空间.
  - 不同文本中词汇的数量会有很大不同，所以数据处理之前必须进行标准化.

文本聚类的特征选择

- 基于文档频率的选择
- Term Strength
- 排序
- Term Contribution

文本聚类的特征转换算法

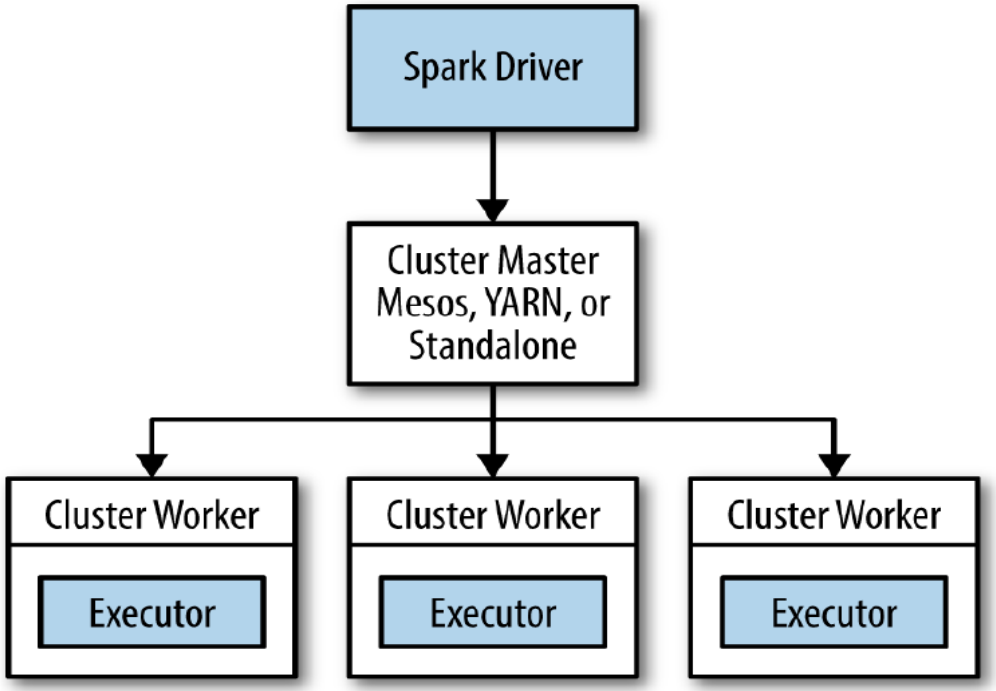
- 潜在语义索引 (LSI)
  - 针对主成分分析 (PCA) 或奇异值分解 (SVD)
  - 利用聚类对文本概念分解
- 概率潜在语义分析 (PLSA)
- 非负矩阵分解 (NMF)

自然语言处理

- 工具库: NLTK
- 主要用于:
  - 生成项目的简介 / README 等文件的文本语料库.
  - 项目数据及代码的预处理.
  - 处理并记录用户搜索的输入.
  - 构建结果的描述报告.

弹性处理技术

- 利用python加快项目构建速度和获得更快的文本处理速度.
- 可以在多核上并发的执行外部库和程序.
- 有效的利用多线程和多进程加快处理速度.
- 有效的利用矢量化技术.



开源社区排行榜

- Hall of fame top 50
- Game Changer top 50
- Ranking Limitations and exceptions

展示您项目生态 - 定制化Compass

- 可视化您的开源项目生态.
- 实时反映社区发展情况.
- 实例: Top 5 organization Ecosystems

Facebook

Apache

Mozilla

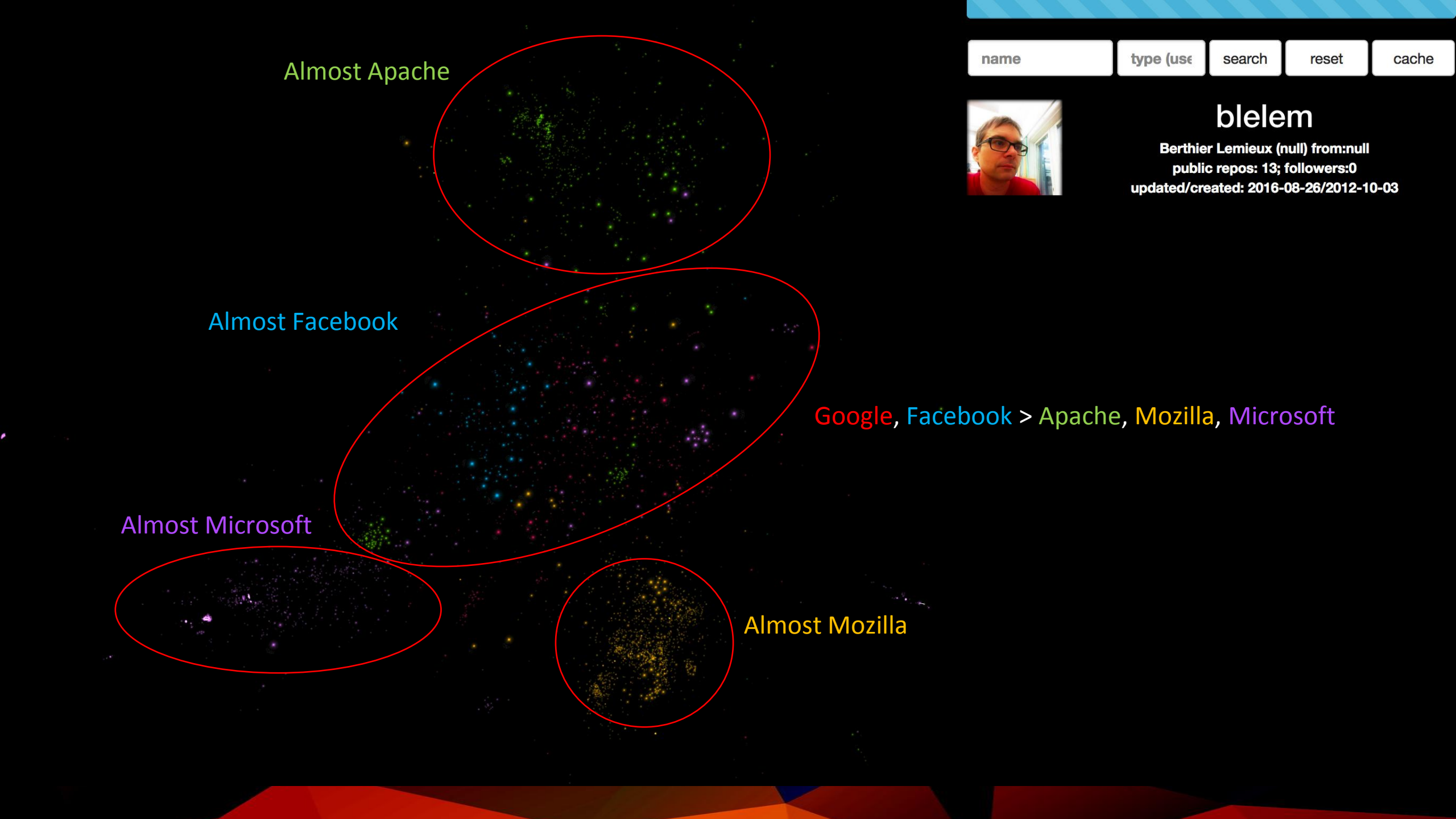
Google

Microsoft

name | type (use | search | reset | cache

blelem

Berthier Lemieux (null) from:null
public repos: 13; followers:0
updated/created: 2016-08-26/2012-10-03


Almost Apache
Almost Facebook
Almost Microsoft
Almost Mozilla
Google, Facebook > Apache, Mozilla, Microsoft
name
type (use
search
reset
cache
blelem
Berthier Lemieux (null) from:null
public repos: 13; followers:0
updated/created: 2016-08-26/2012-10-03

计划

- 更快: 编写更好的算法，获得更多的计算能力 / 计算资源，以更快的处理更多的数据和需求.
- 更深: 更深层次的挖掘，探索更深入的知识.
- 更多: 添加更多有用的新特性，提供更丰富的功能.
- 更广: 扩展到更多的站点 / 仓储库，添加更多的开源项目.

What is Linus doing lately

info@codecompass.net

msabramo | user | search | reset | back

moceap

Mosaab Alzoubi (null) from:null
public repos: 15; followers:14
updated/created: 2016-02-17/2013-05-05

加入我们吧!

- 加入这个社区，使这个项目变的更强更大，可以提供更多更好的服务！
- 如果是企业合作伙伴，您可以提供一些的计算资源，例如提供一台用于计算的虚拟机等.
- 加入核心团队，我们还有一些困难的问题需要一起攻克.
- 为开源社区贡献正能量！

联系我们

- http://www.codecompass.net
- 邮箱: info@codecompass.net
- 微信二维码: