

# 开源社区数据分析



主讲人：赵生宇

# 内容

1 | 为什么  
Why

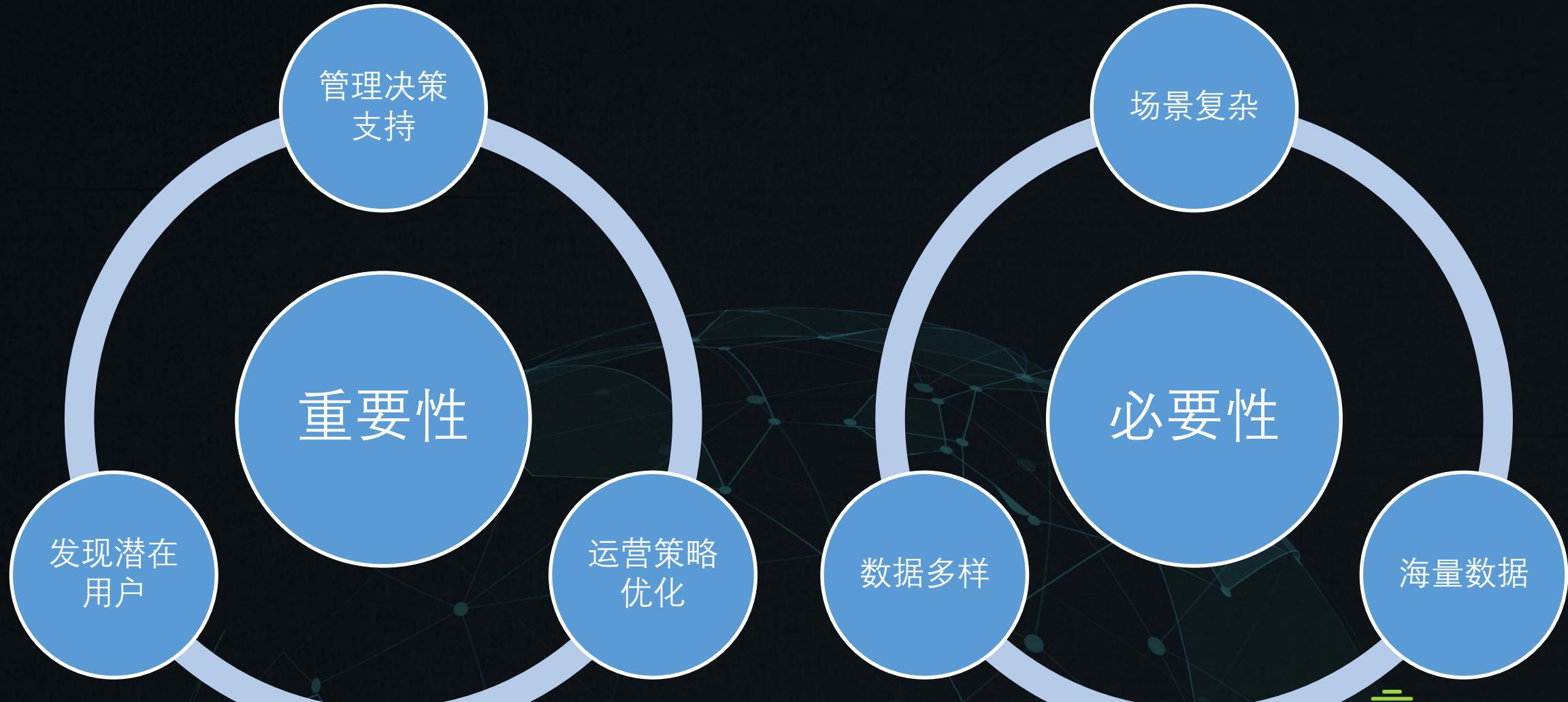
2 | 数据&可视化  
Data&Visualization

3 | 问题  
Problems

4 | 举个例子  
Example

# 01 / 为什么

Why



# 02 / 数据

## Data

### Open Source Alternatives to Reflector? [closed]

Asked 11 years, 2 months ago Active 4 years, 1 month ago Viewed 121k times

419

votes

156

Comments

Closed 8 years ago.

As it currently stands, this question is not a good fit for our Q&A format. We expect answers supported by facts, references, or expertise, but this question will likely solicit debate, polling, or extended discussion. If you feel that this question can be improved and possibly reopened, visit the help center for guidance.

Closed 8 years ago.

```
commit 7f0f8deb6d7ad99fd5160c6a93b529f5c359a06 (HEAD -> master, tag: n8n@0.117.0)
Author: Jan Oberhauser <jan.oberhauser@gmail.com>
Date: Sat Apr 24 21:41:33 2021 +0000

:bookmark: Release n8n@0.117.0

packages/cli/package.json | 2 ++
1 file changed, 1 insertion(+), 1 deletion(-)

commit 91aa4252016631228ccac26cca2a1d083fab00df
Author: Jan Oberhauser <jan.oberhauser@gmail.com>
Date: Sat Apr 24 21:41:32 2021 +0000

:arrow_up: Set n8n-editor-ui@0.87.0 and n8n-nodes-base@0.114.0 on n8n

packages/cli/package.json | 4 +++
1 file changed, 2 insertions(+), 2 deletions(-)

commit a6a19862f2abec1c23da89d481b56be567161d723 (tag: n8n-editor-ui@0.87.0)
Author: Jan Oberhauser <jan.oberhauser@gmail.com>
Date: Sat Apr 24 21:40:22 2021 +0000
```

#### hosted-git-info

15 days ago by GitHub package-lock.json #8

lodash

15 days ago by GitHub package-lock.json #7

handlebars

16 days ago by GitHub package-lock.json

node-notifier

22 Dec 2020 by GitHub package-lock.json

yargs-parser

11 Sep 2020 by GitHub package-lock.json

minimist

06 Apr 2020 by GitHub package-lock.json

kind-of

03 Apr 2020 by GitHub package-lock.json

CVE-2021-33026

The Flask-Caching extension through 1.10.1 for Flask relies on Pickle for serialization, which may lead to remote Python code.

CVE-2021-30183

Cleartext storage of sensitive information in multiple versions of Octopus Server where in certain situations when running a scheduled task, the password used to encrypt files is stored in plain text.

CVE-2021-29641

Directus 8 before 8.8.2 allows remote authenticated users to execute arbitrary code because file-upload permissions include write access to the main upload directory.

CVE-2021-29442

Nacos is a platform designed for dynamic service discovery and configuration and service management. In Nacos before version 1.4.1, the configOpsController lets the user.

CVE-2021-28653

The iOS and macOS apps before 1.4.1 for the Western Digital G-Technology ArmorLock NVMe SSD store keys insecurely. They choose a non-preferred storage mechanism.

CVE-2021-28361

An issue was discovered in Storage Performance Development Kit (SPDK) before 20.01.01. If a PDU is sent to the iSCSI target with a zero length (but data is expected), i

CVE-2021-28192

The specific function in ASUS BMC#8217;s firmware Web management page (Remote video storage function) does not verify the string length entered by users, resulting in a denial of service.

CVE-2021-27452

A vulnerability has been found in multiple revisions of Emerson Rosemount X-STREAM Gas Analyzer. The affected products utilize a weak encryption algorithm for storage c

CVE-2021-27400

HashiCorp Vault and Vault Enterprise Cassandra integrations (storage backend and database secrets engine plugin) did not validate TLS certificates when connecting to Cassa

众志成城，共抗新型肺炎！

阅读 6.2万 文章已于2020/01/25修改



赞



在看 654

安全合规数据

代码演化数据

可行性

内容运营数据

活动运营数据

平台协作数据

软件供应链数据

开源数据

云原生数据

容器化数据

微服务数据

大数据数据

边缘计算数据

物联网数据

区块链数据

隐私计算数据

联邦学习数据

其他

A screenshot of a GitHub repository page. At the top, there's a large blue circle containing the text "可行性". Five lines radiate from this central circle to five smaller blue circles, each containing one of the six data types listed above. The GitHub interface shows a list of issues, with the first few being:

- hosted-git-info (moderate severity)
- lodash (high severity)
- handlebars (critical severity)
- node-notifier (moderate severity)
- yargs-parser (low severity)
- minimist (low severity)
- kind-of (low severity)
- CVE-2021-33026 (The Flask-Caching extension through 1.10.1 for Flask relies on Pickle for serialization, which may lead to remote Python code.)
- CVE-2021-30183 (Cleartext storage of sensitive information in multiple versions of Octopus Server where in certain situations when running a scheduled task, the password used to encrypt files is stored in plain text.)
- CVE-2021-29641 (Directus 8 before 8.8.2 allows remote authenticated users to execute arbitrary code because file-upload permissions include write access to the main upload directory.)
- CVE-2021-29442 (Nacos is a platform designed for dynamic service discovery and configuration and service management. In Nacos before version 1.4.1, the configOpsController lets the user.)
- CVE-2021-28653 (The iOS and macOS apps before 1.4.1 for the Western Digital G-Technology ArmorLock NVMe SSD store keys insecurely. They choose a non-preferred storage mechanism.)
- CVE-2021-28361 (An issue was discovered in Storage Performance Development Kit (SPDK) before 20.01.01. If a PDU is sent to the iSCSI target with a zero length (but data is expected), i
- CVE-2021-28192 (The specific function in ASUS BMC#8217;s firmware Web management page (Remote video storage function) does not verify the string length entered by users, resulting in a denial of service.)
- CVE-2021-27452 (A vulnerability has been found in multiple revisions of Emerson Rosemount X-STREAM Gas Analyzer. The affected products utilize a weak encryption algorithm for storage c
- CVE-2021-27400 (HashiCorp Vault and Vault Enterprise Cassandra integrations (storage backend and database secrets engine plugin) did not validate TLS certificates when connecting to Cassa

In the bottom right corner of the GitHub interface, there are statistics: 777 报名 (Registrations), 300+ 到场 (Attendance), and 6607 在线 (Online).

A screenshot of a GitHub repository page. At the top, there's a large blue circle containing the text "可行性". Five lines radiate from this central circle to five smaller blue circles, each containing one of the six data types listed above. The GitHub interface shows a list of issues, with the first few being:

- improve design (closed 31 minutes ago by AlexandroGonSan)
- fail in collapse "brackets" (opened 35 minutes ago by AlexandroGonSan)
- Failed to compile the git master version of vscode Ubuntu 20.04. (opened 38 minutes ago by hongyi-zhao)
- Corrupt ZIP: end of central directory record signature not found (opened 1 hour ago by aiman-mumtaz)
- [c/c++, etc] Hide compilation errors hints (opened 2 hours ago by Raspreval)
- Displaying imports in js/ts as table in text/code editor view (opened 3 hours ago by Kamil93)
- Insert cursors using keyboard only? (opened 3 hours ago by septsea)
- pytest test discover failed (opened 4 hours ago by browncrane)
- tooltip text grabbing broken (opened 4 hours ago by netanel-haber)
- why ?? i can't install extension from vs code .. always i install from vsix , how can i make the internet work in vs (opened 4 hours ago by Ouaimai)

On the right side, there's a "Key User Info" section with the following data:

用户类别	数量
研发/算法工程师	71%
技术主管/CTO	18%
学生	1%
其他	1%

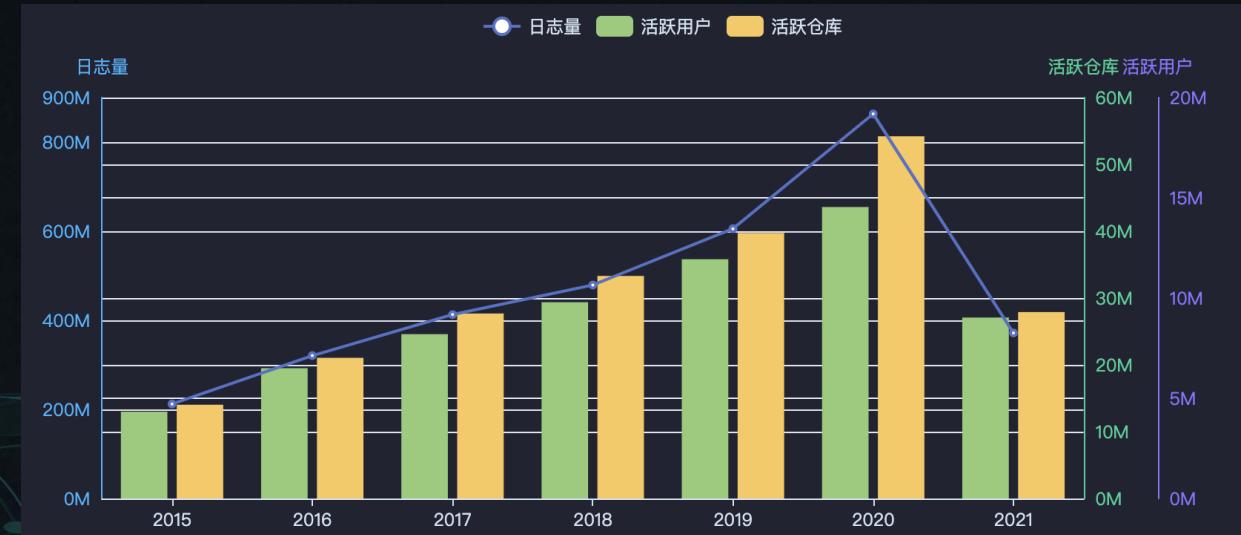
\*有效反馈 854 份

# 02 / 数据

## Data

### GitHub 事件日志

- 2015 至今共计 30 亿+ 条行为日志记录
- 覆盖全域所有项目的所有数据
- 覆盖 star、fork、issue、pull request、release、push、wiki 等主要协作数据
- 用数据说话，才能知道真相
  - GitHub 数据
    - Octoverse 2020：用户总量 5600W，新建仓库 6000W
    - 2020 日志数据：活跃用户 1454W，活跃仓库 5421W
    - TiDB(2W+ star) vs uni-app(3W+ star)



wanganxp commented on 20 Feb

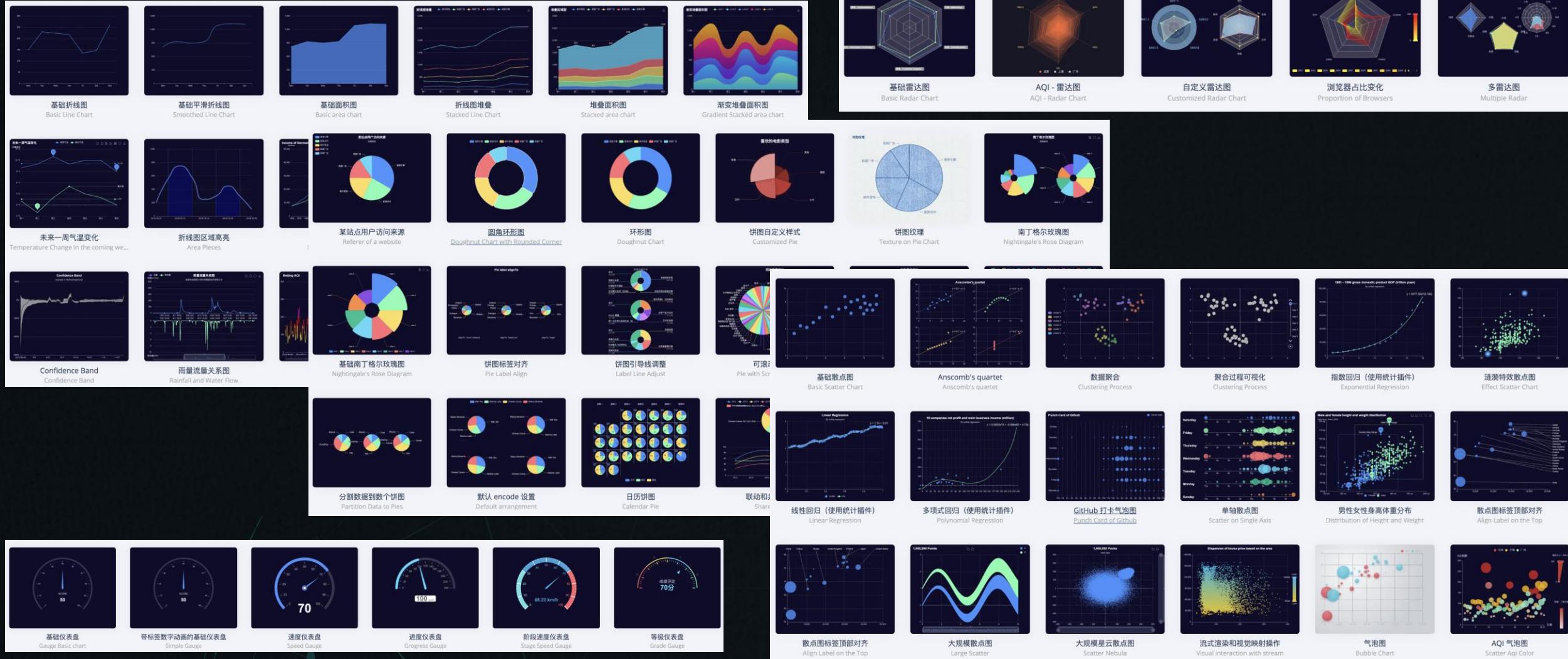
uni-app的用户量远超你家报告里列的其他开源项目。pinggap和用户量和star都和uni-app没法比

#	name	language	activity	developer_count	issue_comment	open_issue	open_pull	review_comment	merge_pull	commits	additions	deletions
3	pingcap/tidb	Go	2013.85	261	22443	729	1554	3161	1235	9191	115905	91389
40	dcloudio/uni-app	JavaScript	346.85	194	640	115	12	0	8	9	18	14

# 02 / 可视化

## Visualization

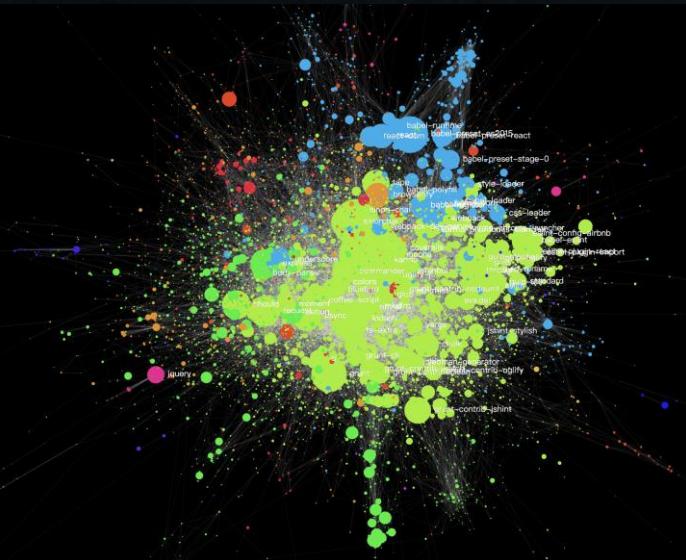
### 传统图表: echarts, AntV



# 02 / 可视化

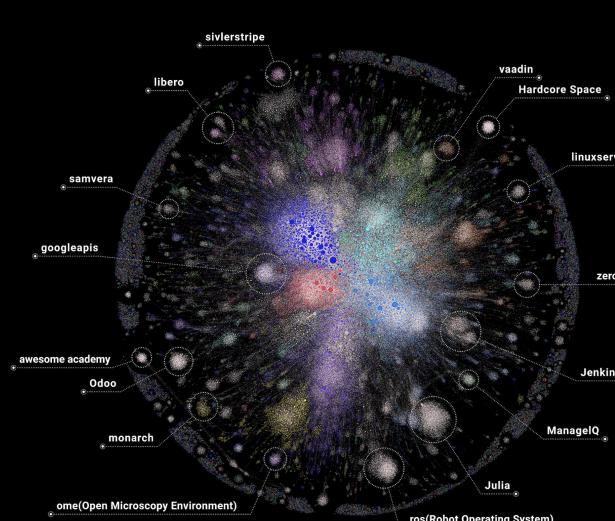
## Visualization

### 图 (Graph) 类型



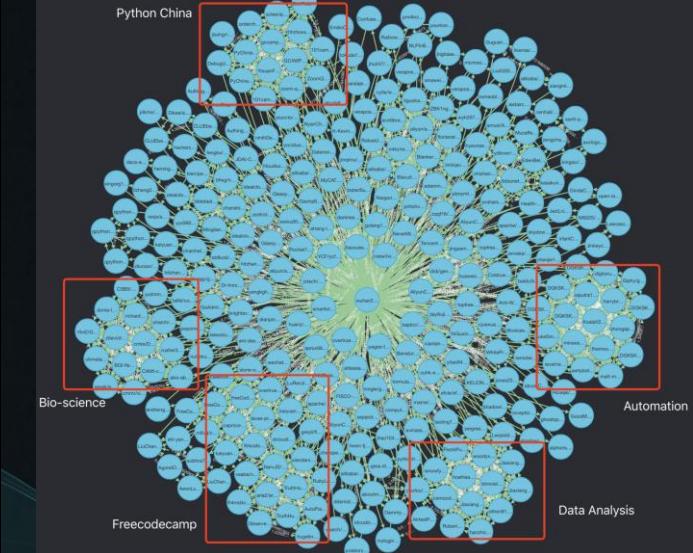
npm 包依赖网络

echarts GraphGL  
54,000 节点  
10,000 边  
npm 总量 181W



OpenGalaxy 2019

Gephi  
171,141 节点  
2,811,489 边  
GitHub 总量



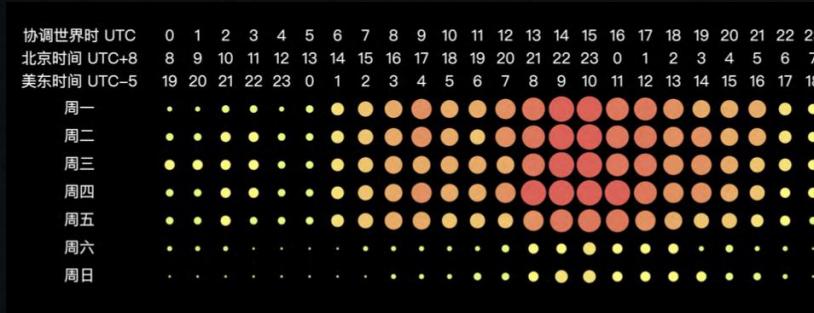
Wuhan2020 协作项目关系图

Neo4j  
300 节点  
862 边

# 02 / 可视化

## Visualization

### 自定义可视化

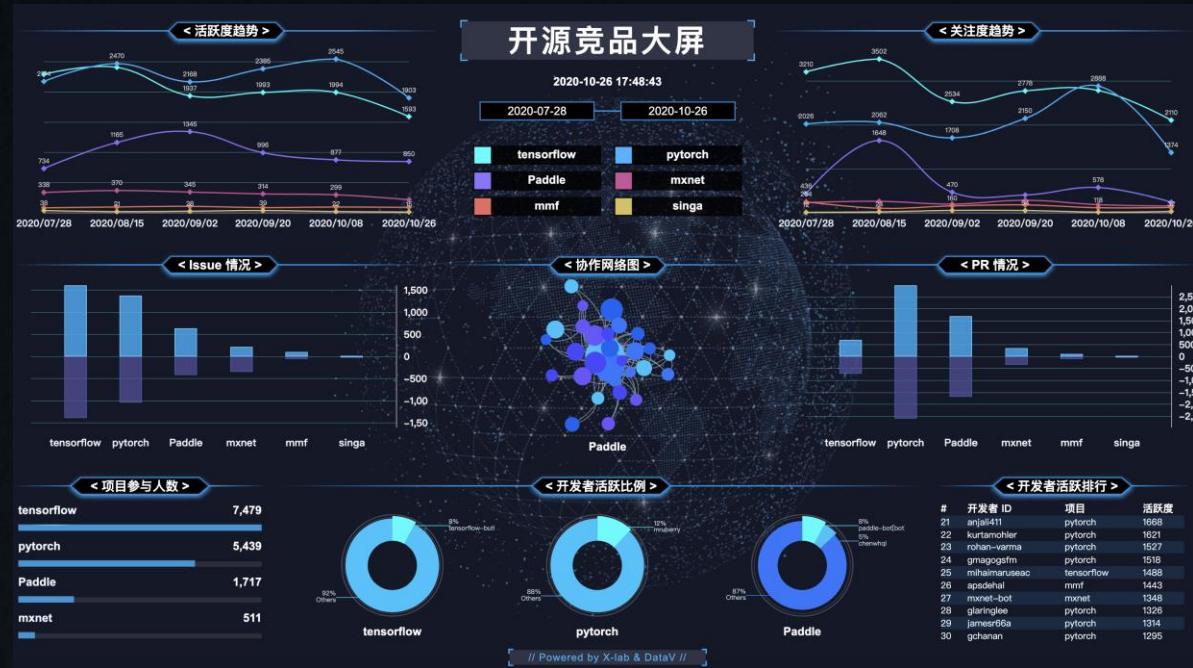


日志时间分布打孔图



词云

人员分布时区柱状图



2019 销售骑行数据



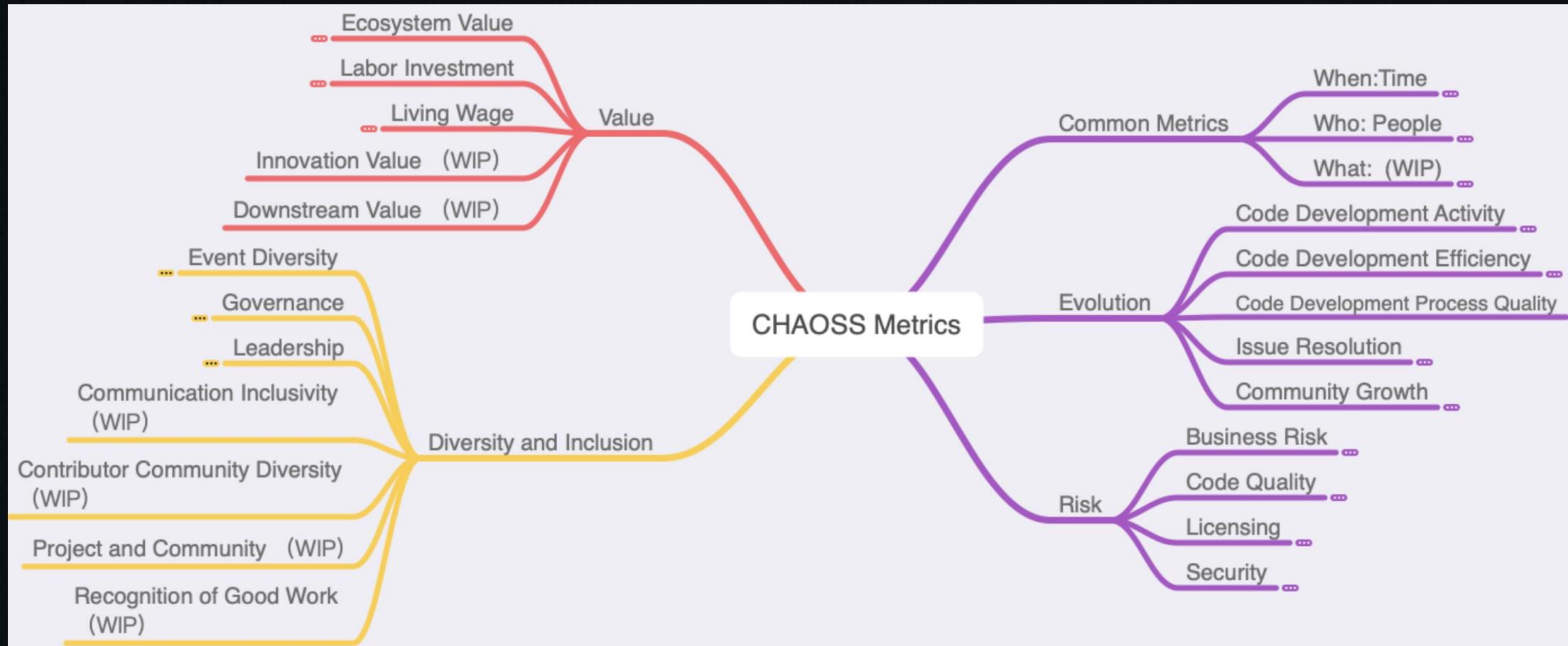
聚合数据大屏

# 03

## 问题

Problems

### 指标体系——项目健康度



# 03

## 问题

Problems

全域

- 项目聚类分类
  - 开源项目分布在哪些领域
  - 给定一个项目，如何预测其所属领域
  - 根据项目类型，可描述开发者画像
- 项目中心度
  - 开源世界中，哪些项目更加重要
  - 项目选型/项目价值预估

项目

- 项目活跃度
  - 项目活跃情况如何？趋势如何？
- 开发者画像
  - 那些参与项目的人都都是谁，他们是做什么的？
  - 谁在关注该项目，有什么诉求？
- 研发行为监测
  - 项目是否存在数据异常，有哪些需要注意的风险点？

全域

项目

研发行为监测

开发者画像

项目活跃度

项目中心度

项目聚类分类

# 03 / 问题

## Problems

基于 GitHub 的基本操作：开发者活跃于项目 Who-When-What

→ 开发者在项目中活跃， 提 Issue, 回复 Issue, 提 PR, 回复 PR, review PR, PR 合入

→ 开发者在具体实例上的活跃度  $A_{di} = \sum w_{ai} C_{ai}$ ,  $w_{ai} = 2,1,3,1,4,5$  ??? 权重

→ 开发者在项目上的活跃度  $A_{dr} = \sqrt{\sum A_{di}}$

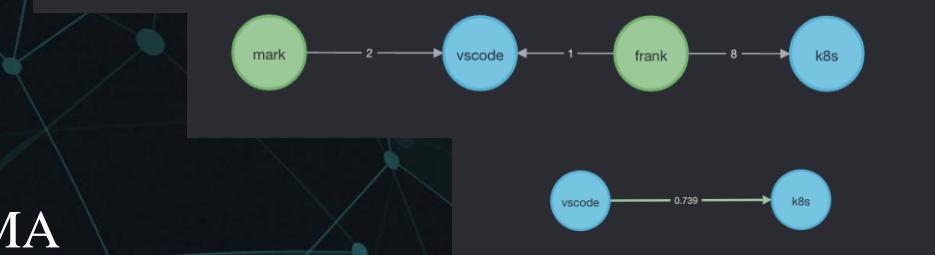
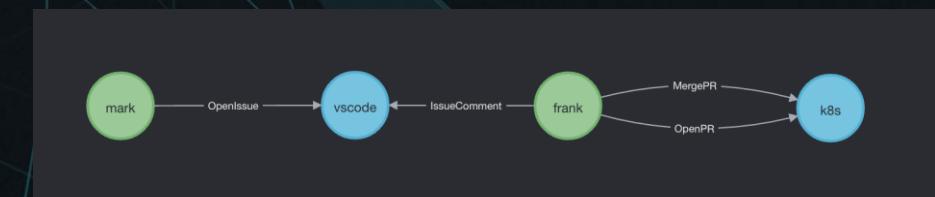
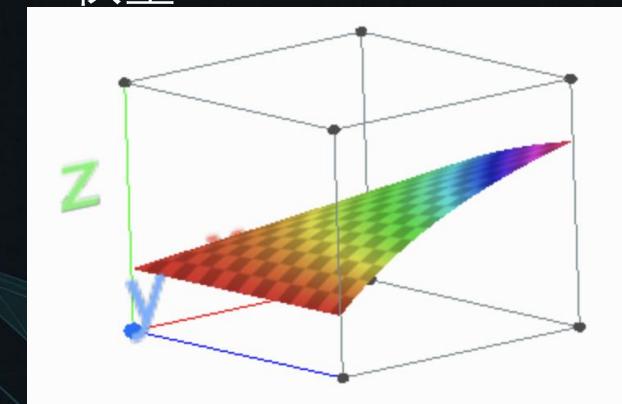
→ 项目的总体活跃度  $A_r = \sum A_{dr}$

→ 同一项目中开发者之间的协作关联度  $RDP_{ab} = \sum \frac{A_{ai}A_{bi}}{A_{ai}+A_{bi}}$

→ 开发者在全域上的协作关联度  $RDG_{ab} = \sum RDP_{ab}$

→ 不同项目间的协作关联度  $RP_{ab} = \sum \frac{A_aA_b}{A_a+A_b}$  ??? 聚合

→ 上述均为一段时间内的统计数据，某时间点：180 天 EMA

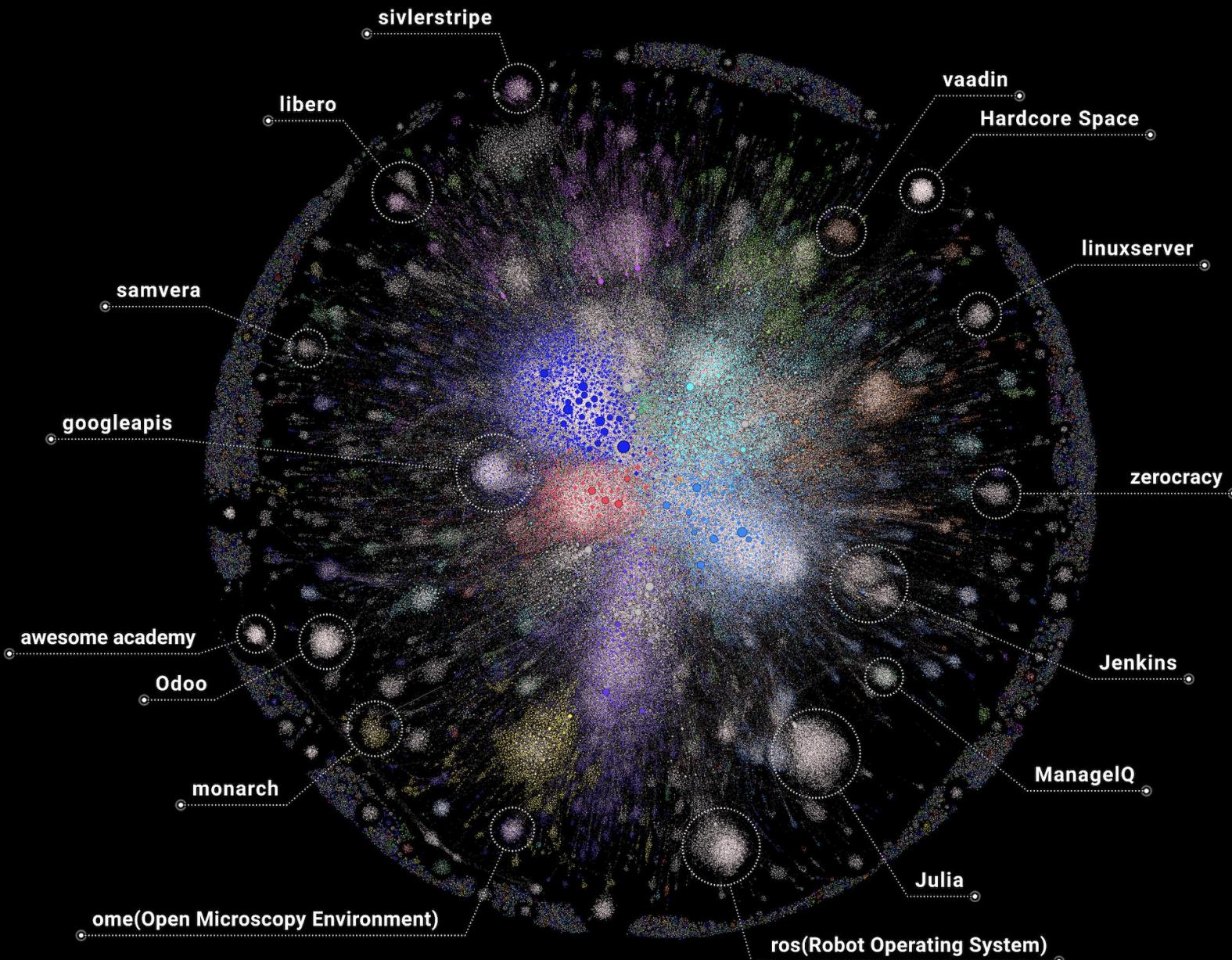


# OpenGalaxy 2019

OpenGalaxy is generated by collaboration network of all active GitHub repos in 2019. This graph contains 171,141 nodes and 2,811,489 edges. The generate method can be found in here [1] and the data is from GHArchive [2].

OpenGalaxy 是通过 GitHub 2019 年全域所有活跃项目的协作网络生成的。本图共包含 171,141 个节点和 2,811,489 条边。具体生成方法请参见这里 [1]，数据来自于 GHArchive [2]。

Area/领域	Top Repos/顶级项目	Count/项目数量
ts & frontend	VSCode, TypeScript, react, jest	23,254
cloud native	kubernetes, go, helm, ansible	15,787
AI libs	pandas, numpy, conda, openjournals	14,971
tools	rust, nextcloud, godotengine	13,361
PHP	symfony, laravel, wordpress, magento	8,158
Microsoft	azure-docs, AspNetCore, WSL	6,276
system	homebrew, systemd	6,193
biotech	rstudio, bioconda	6,102
blockchain	bitcoin, ethereum, ipfs	5,141



[1] [http://blog.frankzhao.cn/open\\_rank\\_and\\_open\\_galaxy/](http://blog.frankzhao.cn/open_rank_and_open_galaxy/)

[2] <http://www.gharchive.org/>

# 04 / 举个例子

Example

## VSCode

- 开源世界的核心，最流行 IDE
- 准备工作：获取项目的地址、创建时间、数据库 ID

The screenshot shows the GitHub repository page for `microsoft / vscode`. The repository has 325 branches and 192 tags. A specific commit by `mjbvz` is highlighted, showing a GraphQL query in the code editor. The query retrieves information about the repository's creation date and database ID.

```
query {
  repository(owner:"microsoft", name:"vscode") {
    createdAt
    databaseId
  }
}
```

The right side of the image shows the results of the GraphQL query, which includes the repository's creation date and database ID.

```
{ "data": { "repository": { "createdAt": "2015-09-03T20:23:38Z", "databaseId": 41881900 } }}
```

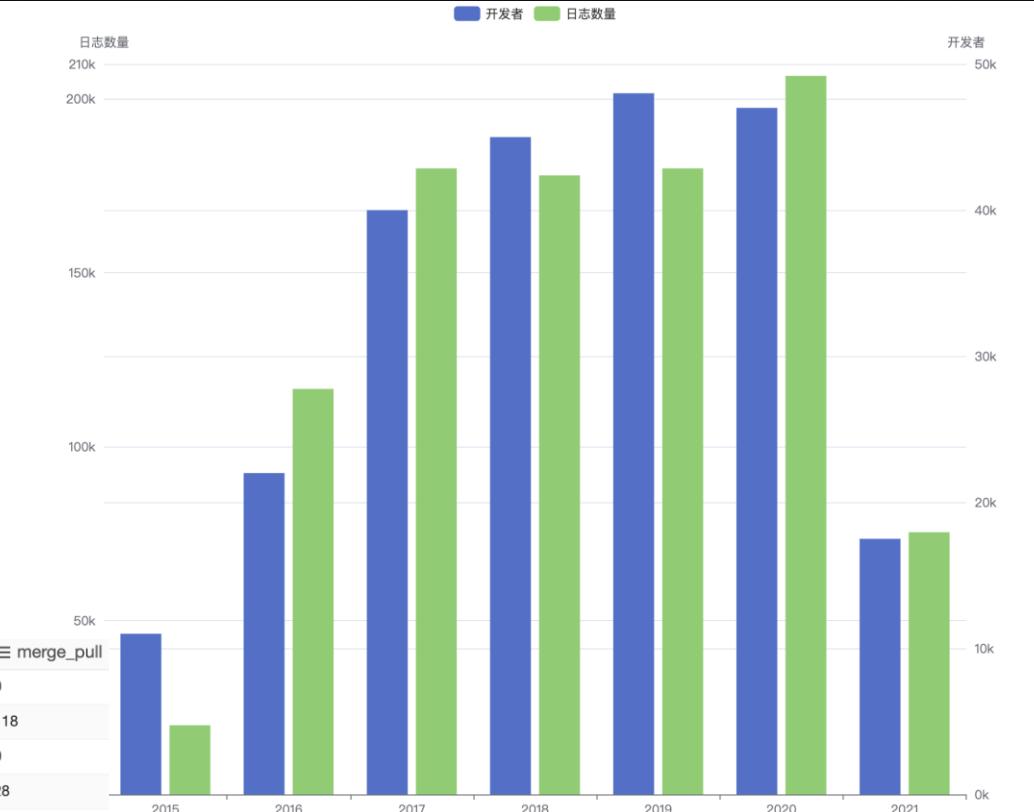
# 04

## 举个例子 Example

### VSCODE 案例分析

VSCODE 2020 开发者活跃度

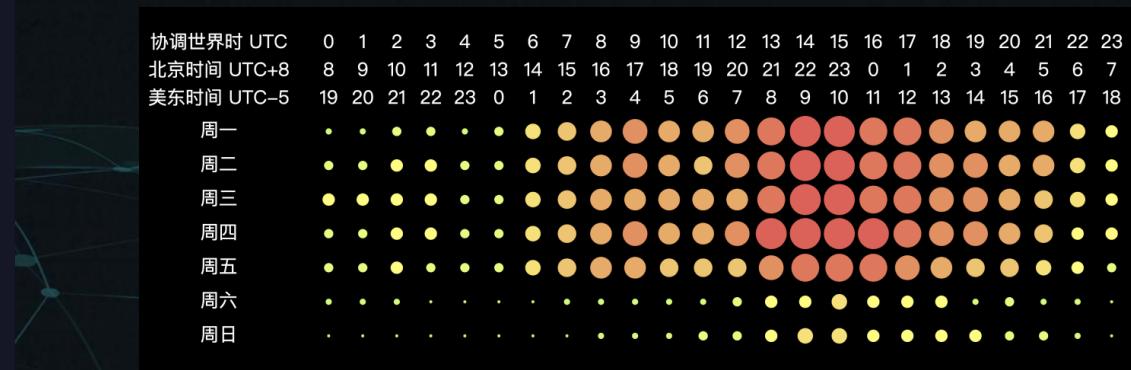
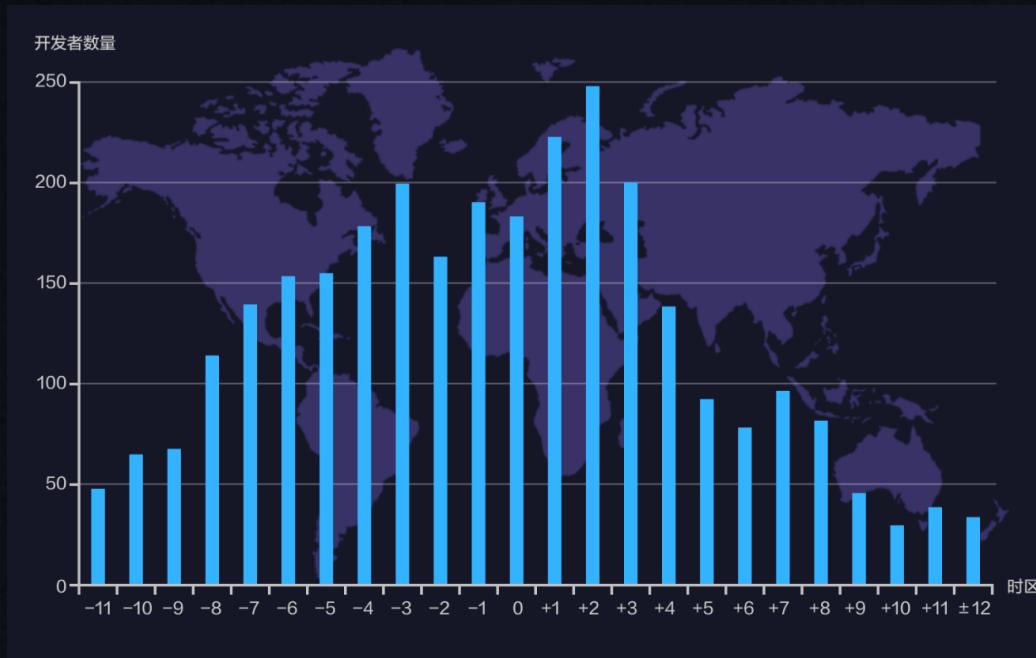
序号	login	activity	issue_comment	open_issue	open_pull	review_pull	merge_pull
1	vscode-triage-bot	9703	9703	0	0	0	0
2	bpasero	7836	4175	700	133	318	118
3	vscodebot[bot]	6912	6912	0	0	0	0
4	isidorn	5526	4231	282	33	123	28
5	mjbvz	4643	3337	152	75	113	65
6	jrieken	4587	2969	411	45	114	41
7	joaomoreno	4064	2767	238	49	111	46
8	sandy081	3875	2737	319	39	52	35
9	Tyriar	3469	2326	337	41	39	38
10	roblorens	3207	2085	289	29	83	25
11	connor4312	2788	2000	193	41	21	39
12	alexroo	2610	1424	154	81	70	71
13	alexdimar	2499	1817	106	46	33	40
14	gjsjohnmurray	2387	2065	51	18	24	14
15	rebornix	2366	1502	216	45	23	41
16	aeschli	2246	1626	108	45	16	41
17	deepak1556	2076	1470	32	45	53	39
18	JacksonKearl	1771	1012	253	24	24	17
19	weinand	1593	1295	125	2	8	2
20	github-actions[bot]	1539	1537	1	0	0	0



# 04 / 举个例子

Example

## VSCode 案例分析



# 04

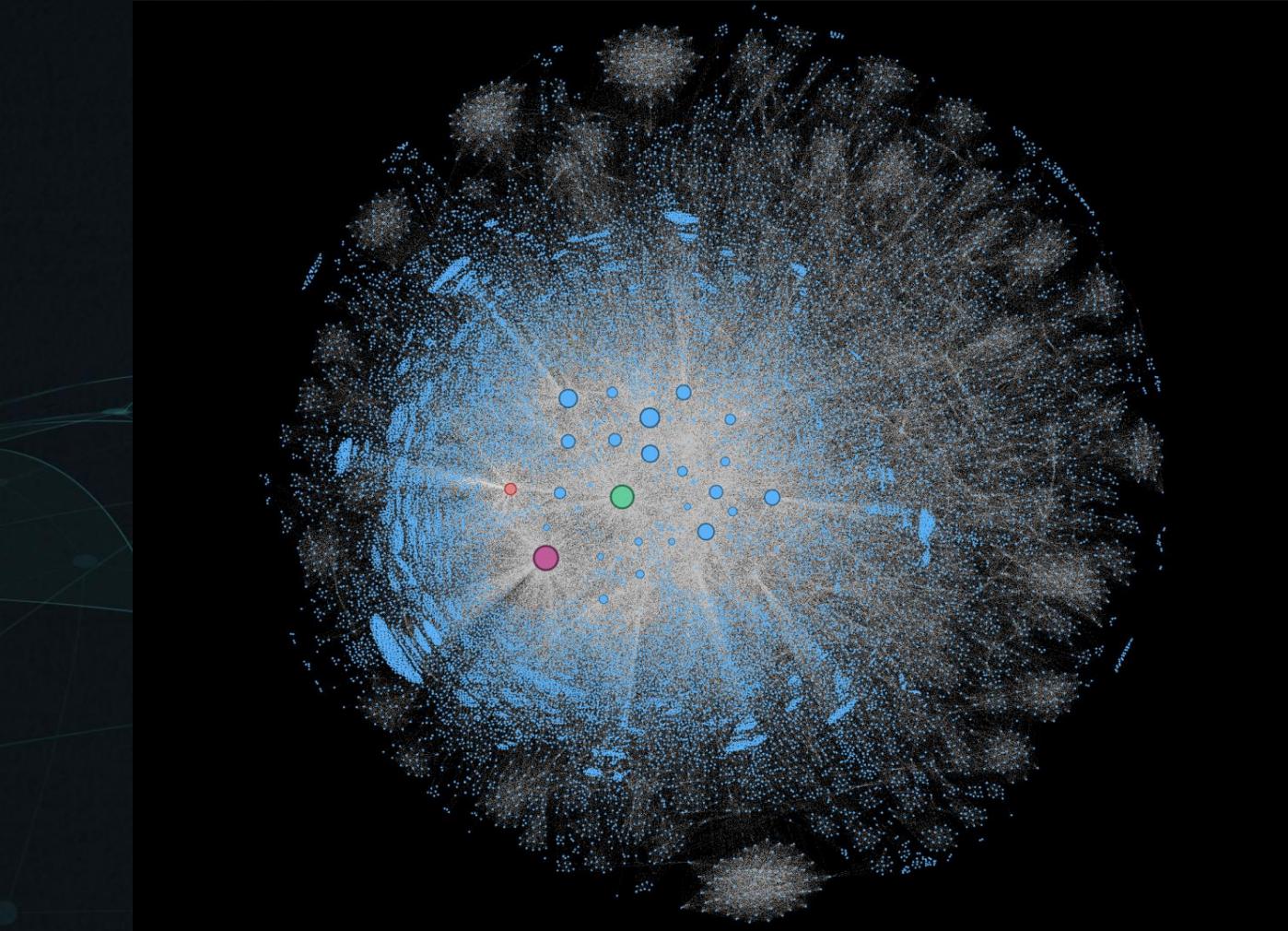
## 举个例子

Example

### VSCode 案例分析

序号	domain	count
1	gmail.com	150
2	users.noreply.github.com	91
3	microsoft.com	54
4	qq.com	7
5	hotmail.com	6
6	google.com	5
7	fb.com	5
8	outlook.com	4
9	me.com	3
10	yahoo.com	3
11	icloud.com	2
12	umich.edu	2
13	kdragOn.dev	2
14	github.com	2
15	googlemail.com	2
16	foxmail.com	2
17	163.com	2

VSCode 2020 开发者邮箱后缀分布



VSCode 2020 开发者协作网络

# 04 / 举个例子

Example

## VSCode 案例分析

#	Repo	Relation	PageRank
1	microsoft/TypeScript	799	544
2	microsoft/vscode-remote-release	594	162
3	microsoft/vscode-python	458	157
4	DefinitelyTyped/DefinitelyTyped	410	564
5	Microsoft/vscode-cpptools	360	102
6	microsoft/terminal	323	243
7	electron/electron	255	256
8	flutter/flutter	236	645
9	microsoft/vscode-docs	227	40
10	gatsbyjs/gatsby	220	504

VSCode 2020 项目协作关联 Top10

#	repo_name	resolve_period_avg	response_period_avg	resolve_period_median	response_period_median	count
1	microsoft/vscode	7d3h	1d17h	1d2h	2h33m	6154

VSCode 2021 Issue 响应解决周期

# 04

## 举个例子

Example

### VSCode 案例分析之社区流程

- build-chat: 将构建信息发送到 Slack 中
- classifier/classifier-deep: Issue 自动打标/基于机器学习
- copycat: 跨仓库 Issue 拷贝
- english-please: 非英文开 Issue 提示使用英文
- locker: Issue 关闭一段时间后自动锁定
- needs-more-info-closer: 需要用户反馈的 Issue 若一段  
    时间没有回复自动关闭
- regex-labeler: 根据 Issue 描述中正则匹配结果打标签
- topic-subscribe: 根据 label 提醒某些账户关注当前 Issue

另一些注重流程的有趣社区: Kubernetes、React、OpenDigger

 JacksonKearl	Merge pull request #25 from microsoft/dependabot/npm_and_yar...	...	ec0c76e 12 days ago	242 commits
 .vscode	Add assignee classifier		13 months ago	
 api	Let bot notify all assignees when issue needs attention		19 days ago	
 author-verified	Move to new update server endpoint		2 months ago	
 blob-storage-test	Add '*' to allowUsers in commands		8 months ago	
 build-chat	Ensure sort order		3 months ago	
 classifier-deep	Dont add assignees when in debug mode		2 months ago	
 classifier	Don't assign when in debug mode		2 months ago	
 commands	remove exclusive test		6 months ago	
 common	Move to new update server endpoint		2 months ago	
 copycat	Remove even more logging		7 months ago	
 english-please	Remove even more logging		7 months ago	
 feature-request	Don't auto-backlog-candidate issues authored by team members		3 months ago	
 latest-release-monitor	Remove even more logging		7 months ago	
 locker	Fix typo in how inputs are read		6 months ago	
 needs-more-info-closer	Let bot notify all assignees when issue needs attention		19 days ago	
 new-release	Move to new update server endpoint		2 months ago	
 regex-labeler	Remove even more logging		7 months ago	
 release-pipeline	Remove release checking from author-verified, use release-pipeline i...		4 months ago	
 test-plan-item-validator	Remove even more logging		7 months ago	
 topic-subscribe	Add topic-subscribe bot		9 months ago	

# 05 / 大作业

Assignment

## 某开源项目的 2020 年深入数据分析

### 数据类

- 基础的统计数据分析、可视化
- 开发者数据统计、可视化
- 关联数据的分析，如协作关联度高的其他项目
- 其他任意想做的数据分析

### 流程类

- 项目的日常协作流程调研
- 开发者参与流程调研
- 项目 CI/CD 的流程调研

## 课程提供

- 全域数据的 Clickhouse 数据库只读访问能力
- 预置常用数据统计的 SQL
- 分析过程中的指导答疑

<https://github.com/X-lab2017/open-digger>

---

THANK YOU

---

