开源软件及其开发

周明辉

zhmh@pku.edu.cn

2023秋季学期

目录

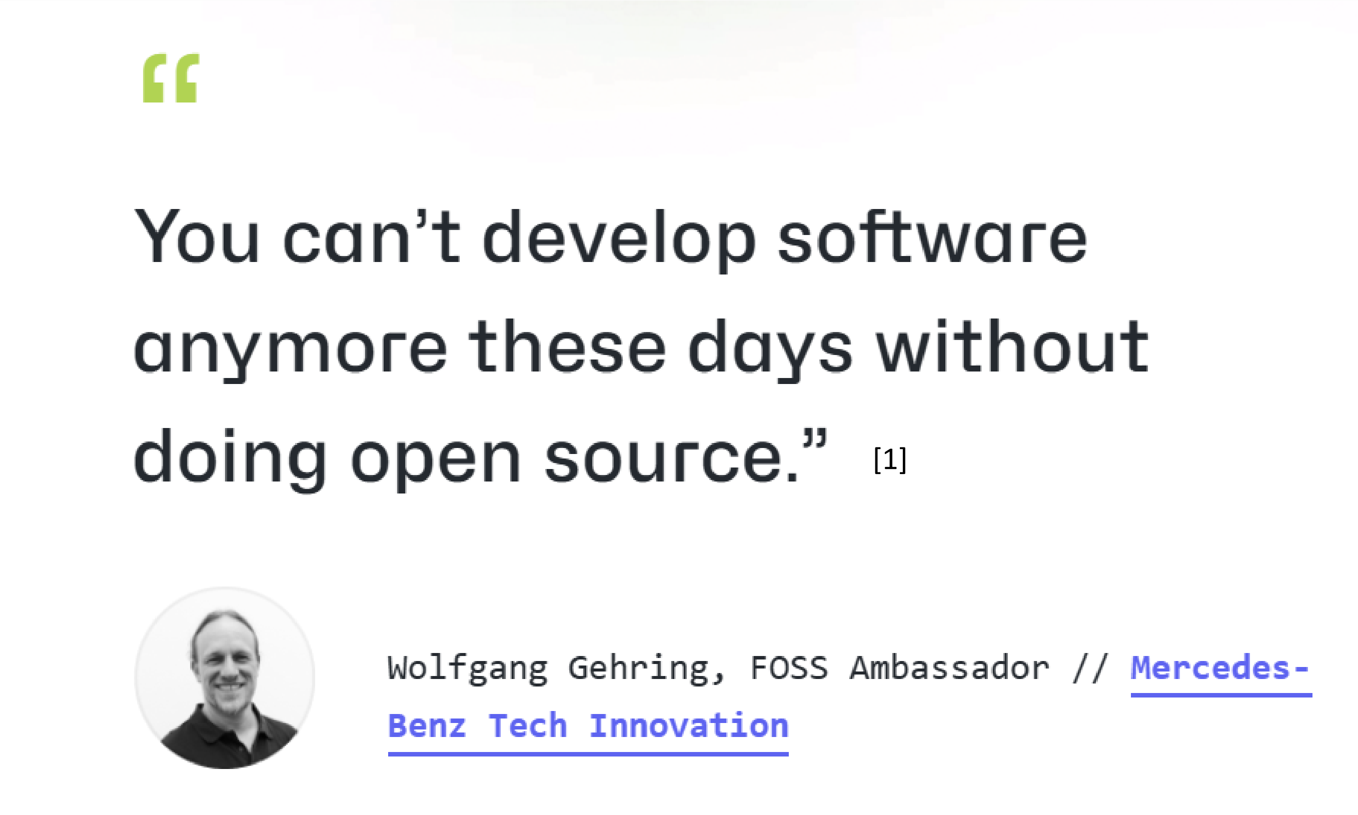
CONTENTS

01. 开源的必要性和正义性

02. 开源的挑战及研究

03. 开源的定义

开源创新：时代和社会发展的需要

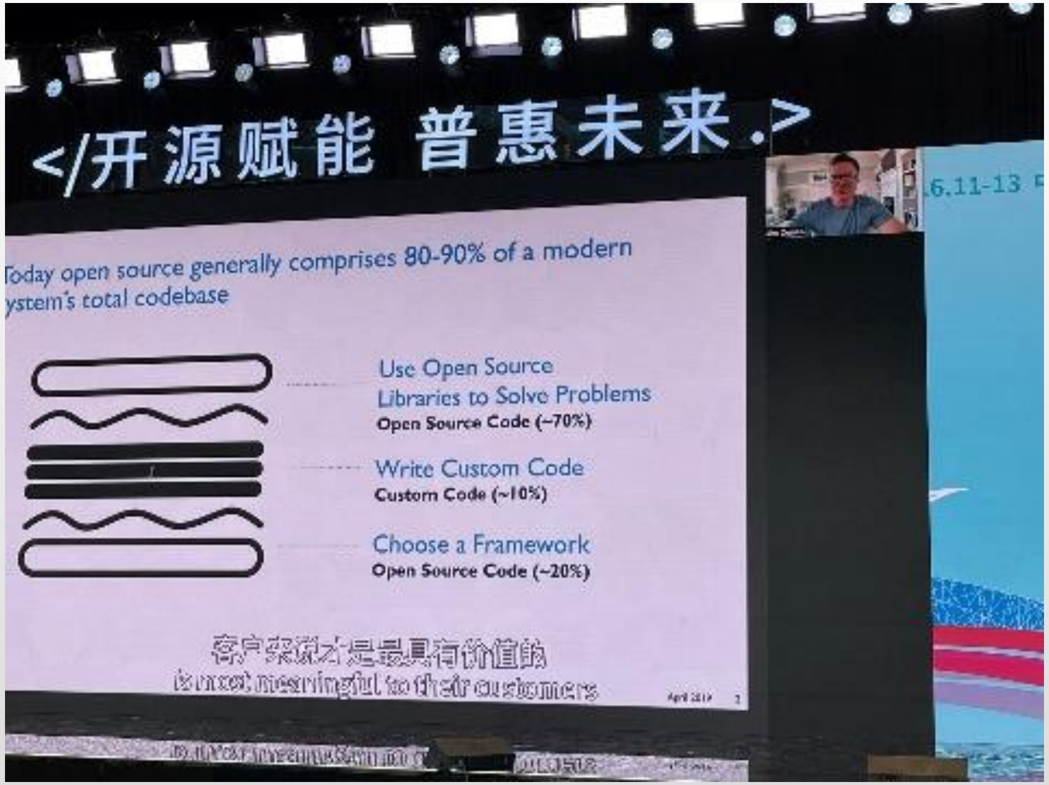


“

You can’t develop software anymore these days without doing open source.” [1]

Wolfgang Gehring, FOSS Ambassador // Mercedes-Benz Tech Innovation

Linux基金会:
现代系统中80~90%源代码来自开源



Today open source generally comprises 80-90% of a modern system's total codebase

Use Open Source Libraries to Solve Problems
Open Source Code (~70%)

Write Custom Code
Custom Code (~10%)

Choose a Framework
Open Source Code (~20%)

客户[illegible]才是最具有价值的
is most meaningful to their customers

在社会经济全球化退潮的今天，开源软件的全球化势不可逆

[1] https://octoverse.github.com/

为什么开源？开源的本质特点

→

用户创新：open user innovation [1]（用户自己实现想要的新特征)，极大地降低了创新成本，缩短了从生产者到消费者的距离

能者治理/才配其位：it suggests a broadly libertarian view of the proper relationship between individuals and institutions. [2]

宏观角度：开源具有推进人类命运共同体发展的正义性！

微观角度：我的贡献能被看见，我的贡献并入到开源项目，在下一次release后被所有用户下载并使用，我有成就感和自豪感！

[1] Hippel and Krogh. Open source software and the “private-collective” innovation model: Issues for organization science. 2003
[2] Eric Raymond. The Cathedral and the Bazaar. 1998

目录

CONTENTS

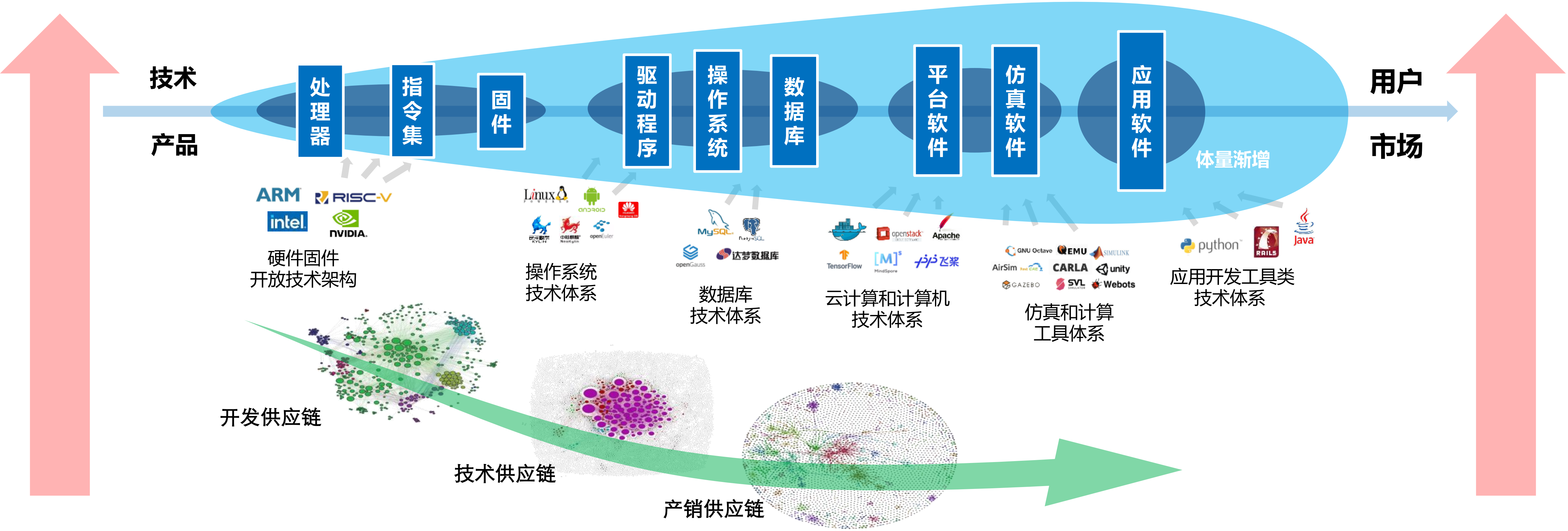
01. 开源的必要性和正义性

02. 开源的挑战及研究

03. 开源的定义

开源核心挑战：生态的形成和持续

- 开源成功的核心表征是生态，生态有两个维度：
  - 水平横向：软硬件全栈，各类供应链关系网络
  - 垂直纵向：项目社区的汇聚、协作、可持续



技术

产品

处理器

指令集

固件

驱动程序

操作系统

数据库

平台软件

仿真软件

应用软件

体量渐增

用户

市场

硬件固件
开放技术架构

操作系统
技术体系

数据库
技术体系

云计算和计算机
技术体系

仿真和计算
工具体系

应用开发工具类
技术体系

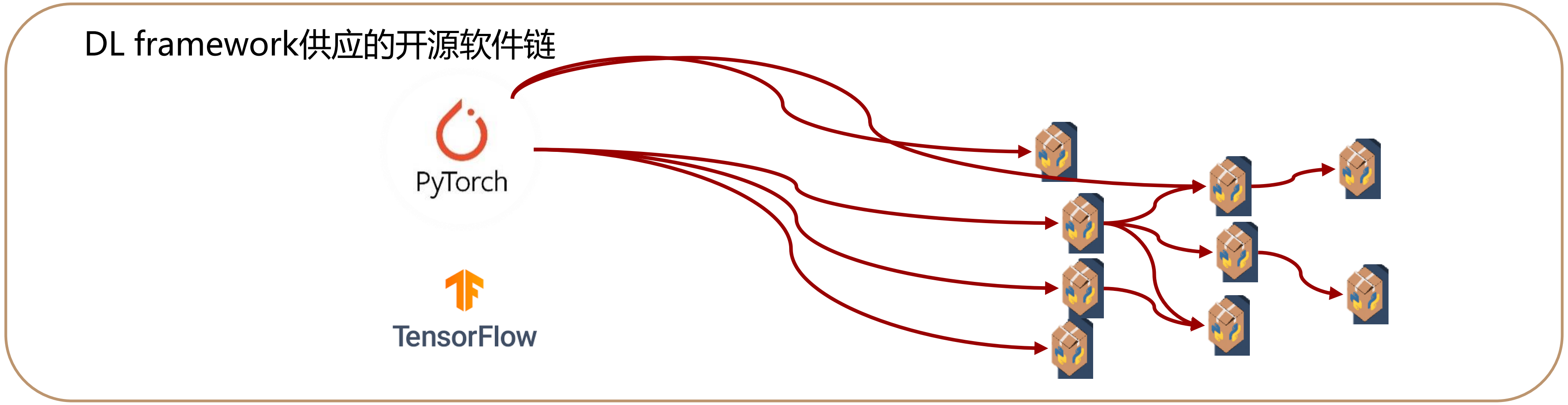
开发供应链

技术供应链

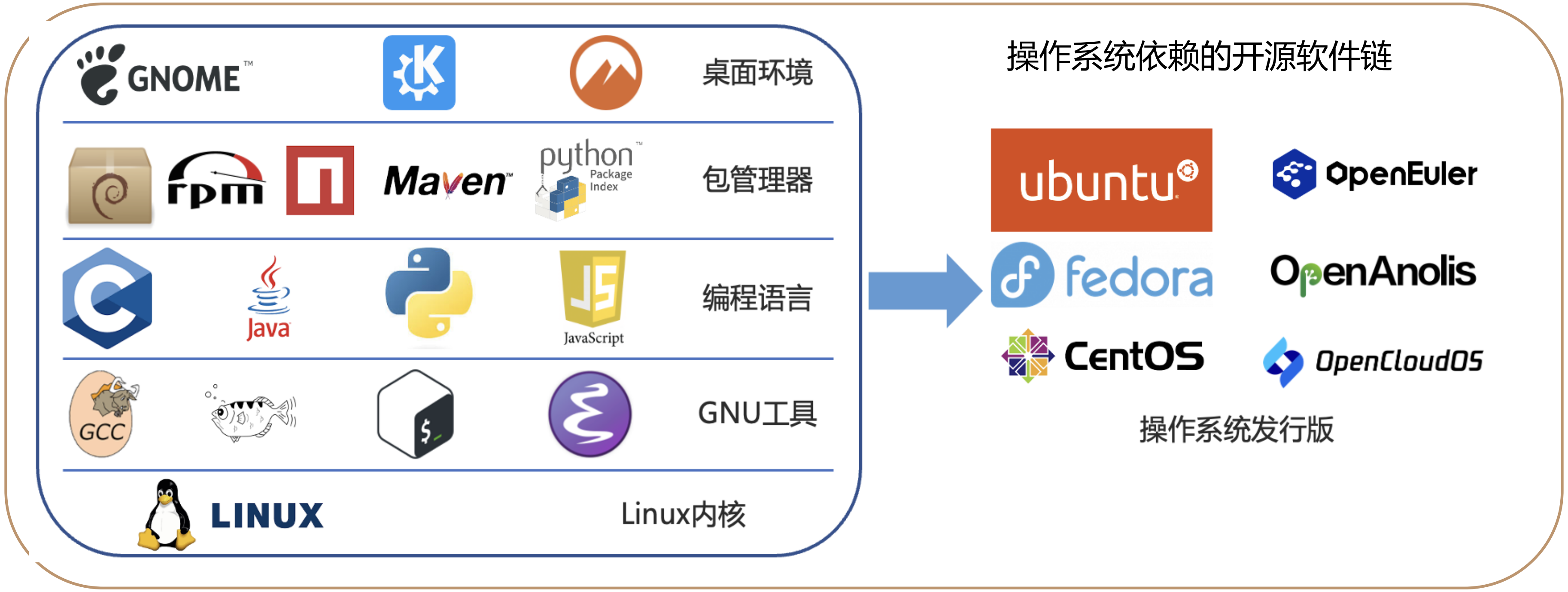
产销供应链

开源软件供应链：如何形成和持续？

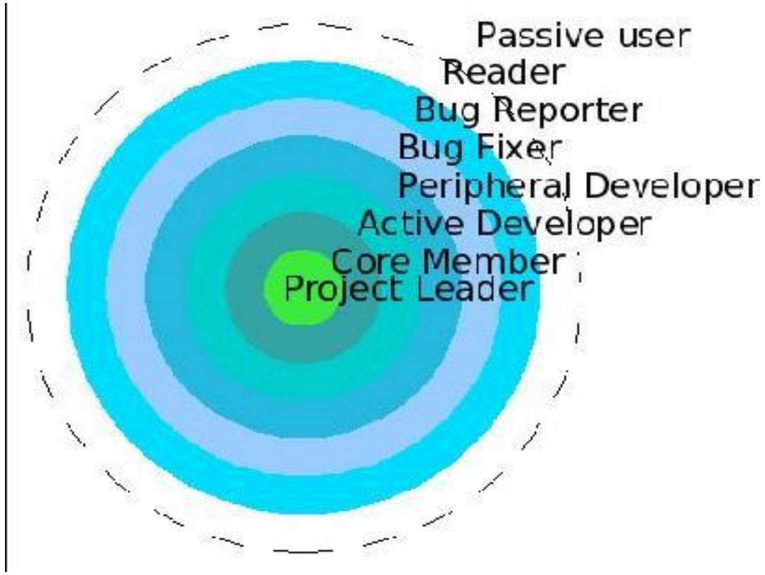
DL framework供应的开源软件链



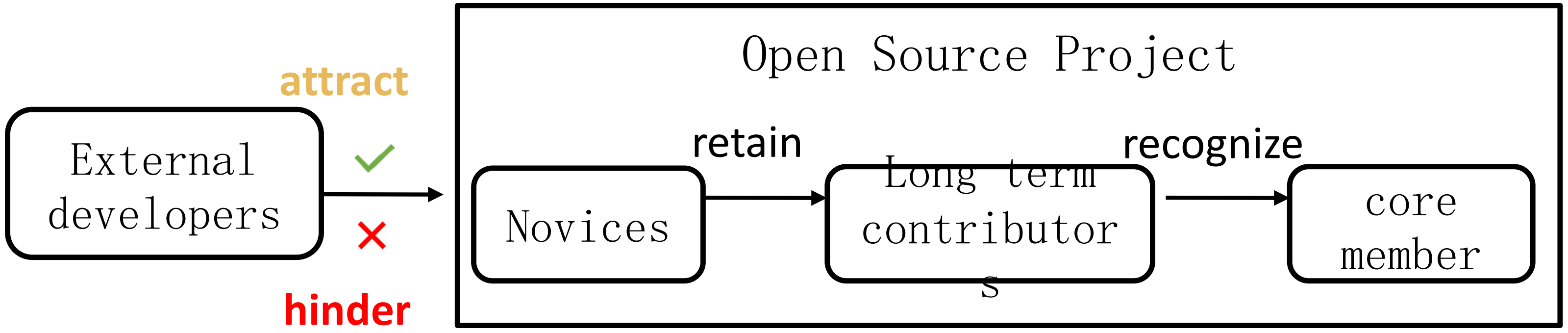
操作系统依赖的开源软件链



社区的汇聚:如何吸引和留存贡献者？



贡献者进入开源项目的流程[1,2]:



External developers

attract ✓

hinder ✗

Open Source Project

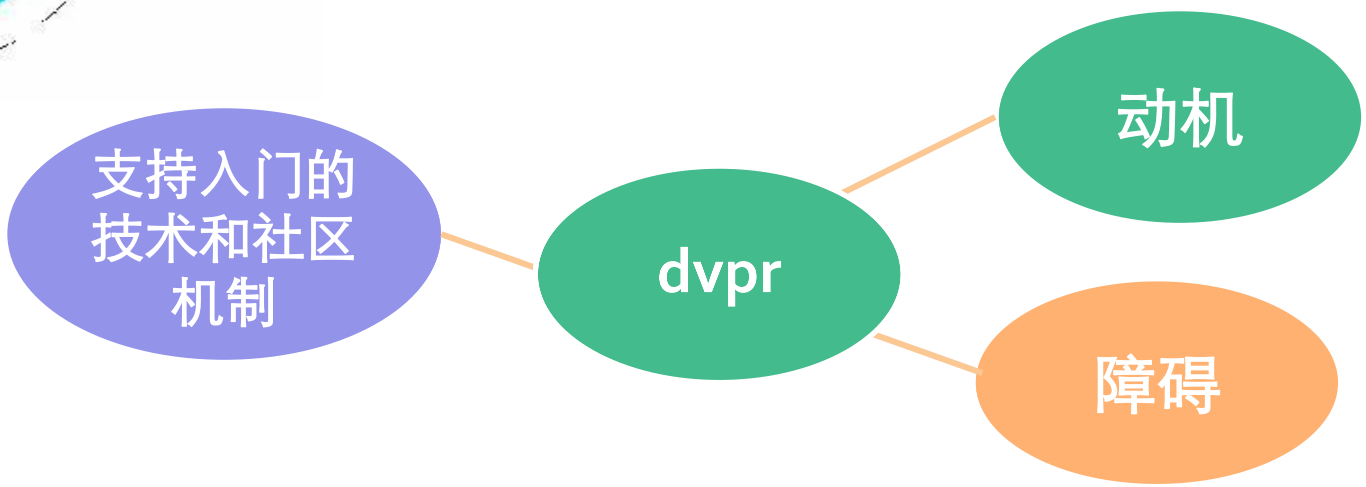
Novices

retain

Long term contributors

recognize

core member



支持入门的技术和社区机制

dvpr

动机

障碍

[1] Ye and Kishida. Toward an Understanding of the Motivation of Open Source Software Developers. ICSE'2003

[2] Zhou and Mockus. What make long term contributors: Willingness and opportunity in OSS community. ICSE'2012.

精细度量，以求理解，进而控制

Open the "black box"

Understand the complexity

Control risks & achieve success

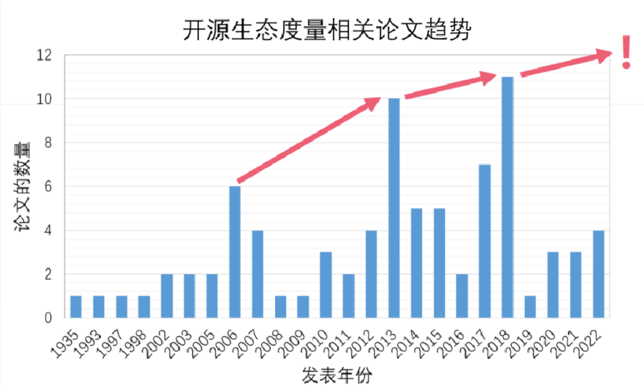
why精细度量？

- 开源供应链是个复杂系统，难以控制
- 人类行为的第一性原理：可解释性

开源度量: 现行方法和工具

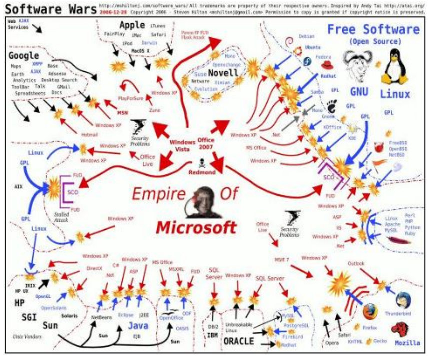
	Open Source Insights	Libraries.io	OSS Insight	OSS Compass
面向的集合	软件包		开源社区	
支持的平台	npm, Go, Maven, PyPI, NuGet, Cargo	npm, Go, Maven, PyPI, NuGet, Cargo, CRAN, conda等常见包管理平台	GitHub	GitHub, Gitee
数据源	包管理平台元数据并扩充部分存储库数据		GH Archive + GitHub API	常见源代码管理、问题跟踪系统、论坛等(by GrimoireLab)
度量的属性	软件包安全风险	软件包可靠性	开源社区流行度与活跃度	开源社区健康 (by CHAOSS)
度量的核心方法	OpenSSF Scorecard	SourceRank	Trending Score	Jensen等人多维、多层的开源生态系统健康度量模型和指标体系
度量方法概述	软件包安全最佳实践和行业标准	上下游依赖 社区活跃度	活动， 等计算流	社区服务，协作开发，活跃度等

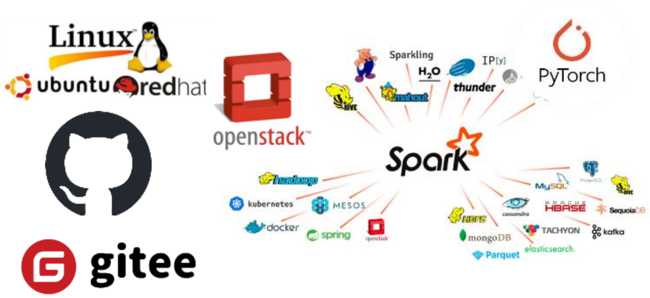
Gap 1: 软件和社区



Gap 2: 研究和实践

数据驱动的~开源软件及其社区的~
度量、预测和智能化支持





开源开发/复杂系统的机制机理、方法技术和支撑工具

开源大数据

海量成功案例和最佳实践

复杂系统原理和技术

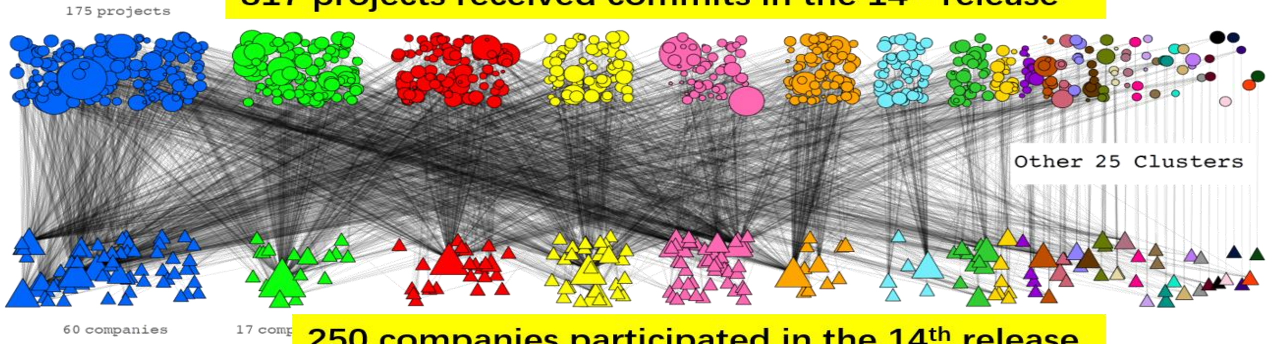
开源数字社会学/开源动力学

北京大学开源分析实验室：
https://osslab-pku.github.io

开源生态的度量：谁供应我的社区？

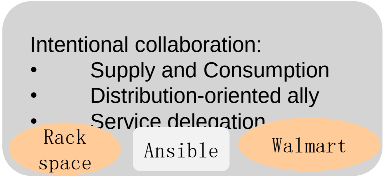
OpenStack 14th : 817code repo/34,192patch/2,439dvpr/250org

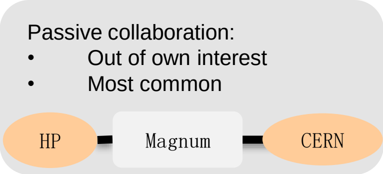
817 projects received commits in the 14th release

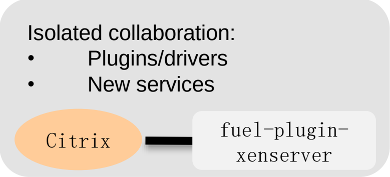


250 companies participated in the 14th release

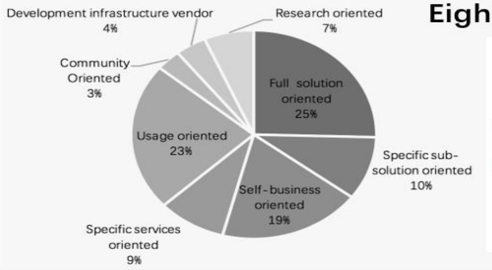
Three Collaboration Patterns:
- Intentional
- Passive
- Isolated

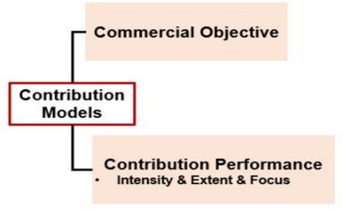

Intentional collaboration:
- Supply and Consumption
- Distribution-oriented ally
- Service delegation


Passive collaboration:
- Out of own interest
- Most common


Isolated collaboration:
- Plugins/drivers
- New services

Eight Contribution Models





Four Types of Task Preference:
- Correction Focused
- Feature Focused
- Reengineering Focused
- Mainly Correction and Feature

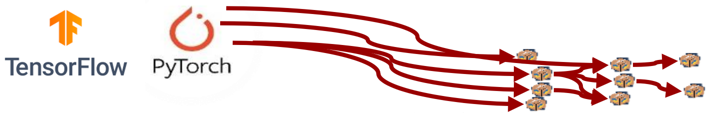
250个参与机构：8种贡献模式；4种任务选择模式；3种协作模式

- 提供完整解决方案获益: Rackspace, Huawei, 99cloud
- 提供部分解决方案获益: SwiftStack,高鸿信安 gohighsec
- 特定业务集成: Intel
- 提供补充服务获益: Codethink
- 面向使用: eBay, china mobile
- 面向社区: Linux foundation
- 开发基础设施供应: Google
- 面向研究: Peking University

Zhang et al. Companies'participation in oss development -- an empirical study of openstack. IEEE Transactions on Software Engineering, 2019.
Zhang et al. How do companies collaborate in open source ecosystems? ICSE'2020
Zhang et al. Corporate dominance in open source ecosystems: a case study of OpenStack. FSE'2022

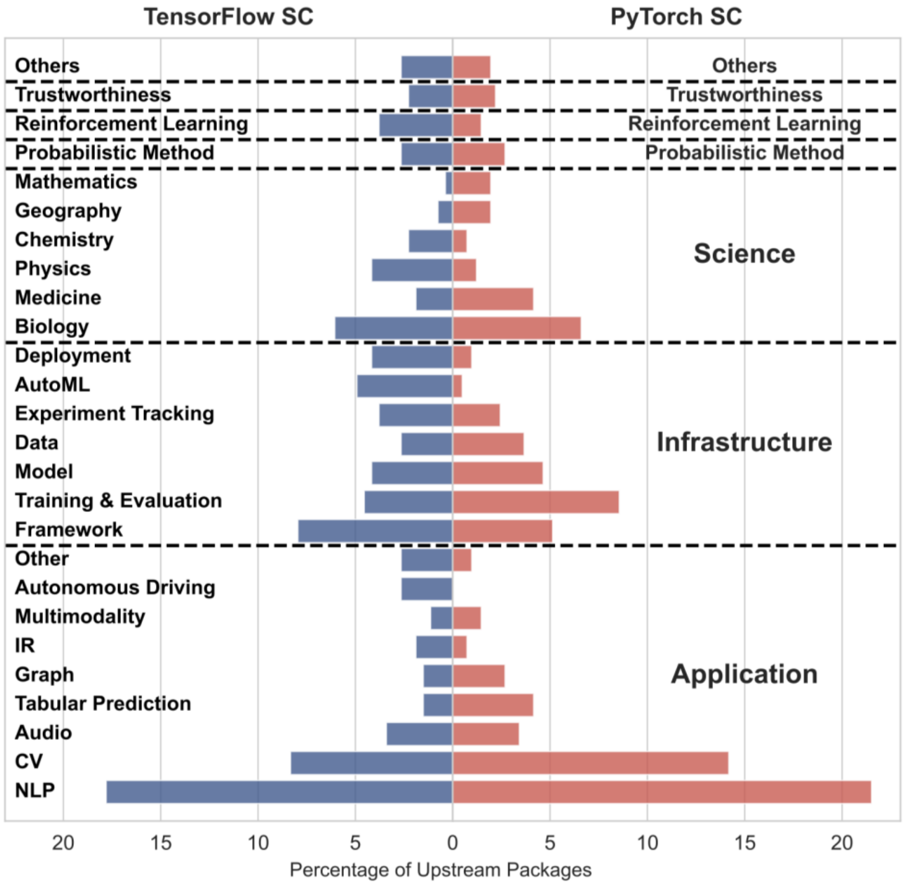
开源生态的度量：我能供应谁？供应链生态的持续性/培育问题

□ DL供应链：以TensorFlow和PyTorch为起点构建PyPI上软件包之间通过安装依赖形成的供应链

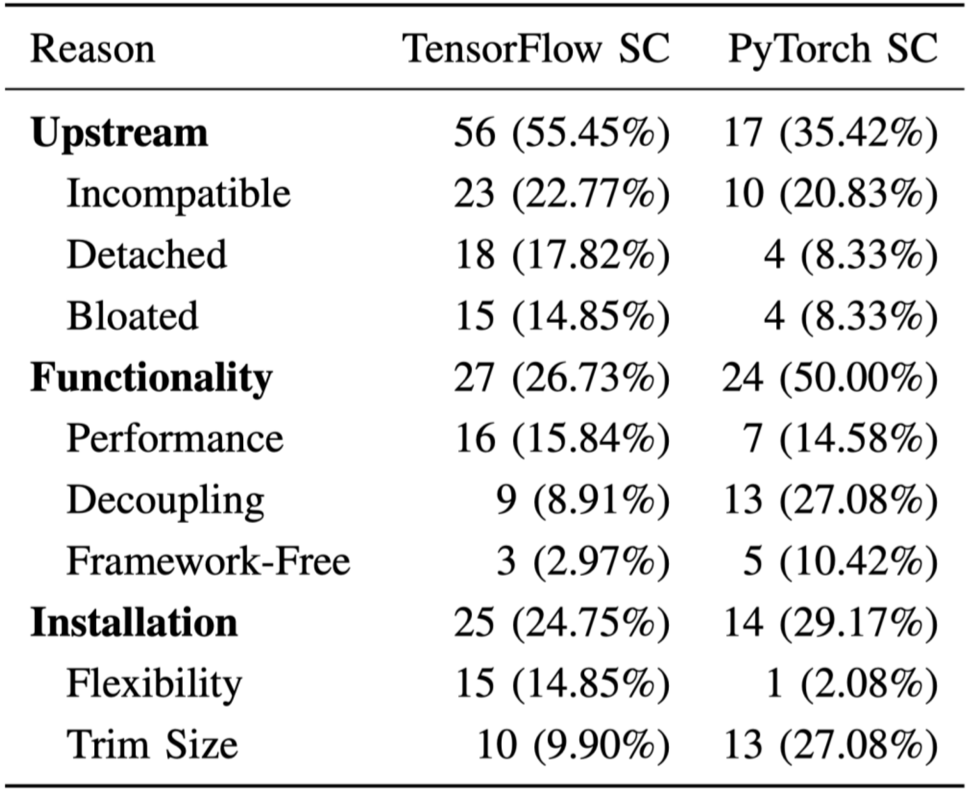


□ 观测：

- ➢ TensorFlow SC和PyTorch SC分别在通用开发工具和特定应用领域上形成了自己的核心竞争优势
- ➢ TensorFlow SC上软件包脱离的主要原因是上游依赖间的不兼容问题，PyTorch SC上软件包脱离的主要原因是为了降低安装包的大小和功能解耦



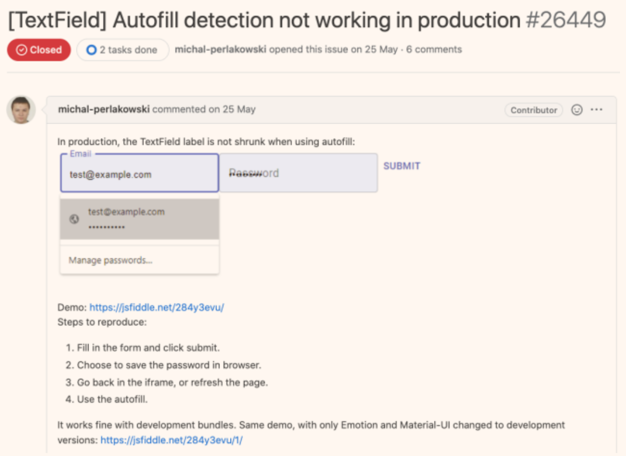
软件包脱离供应链的原因



Reason	TensorFlow SC	PyTorch SC
Upstream	56 (55.45%)	17 (35.42%)
Incompatible	23 (22.77%)	10 (20.83%)
Detached	18 (17.82%)	4 (8.33%)
Bloated	15 (14.85%)	4 (8.33%)
Functionality	27 (26.73%)	24 (50.00%)
Performance	16 (15.84%)	7 (14.58%)
Decoupling	9 (8.91%)	13 (27.08%)
Framework-Free	3 (2.97%)	5 (10.42%)
Installation	25 (24.75%)	14 (29.17%)
Flexibility	15 (14.85%)	1 (2.08%)
Trim Size	10 (9.90%)	13 (27.08%)

自动化新手任务推荐：GFI-Bot

- AI-Powered智能推荐: 为开源项目issue自动打标签（即判断其是否“新手”任务）



[TextField] Autofill detection not working in production #26449

Closed · 2 tasks done · michal-perlakowski opened this issue on 25 May · 6 comments

michal-perlakowski commented on 25 May · Contributor

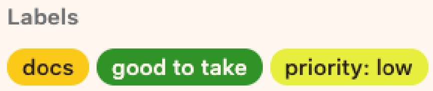
In production, the TextField label is not shrunk when using autofill:

Demo: https://jsfiddle.net/284y3evu/
Steps to reproduce:

1. Fill in the form and click submit.
2. Choose to save the password in browser.
3. Go back in the iframe, or refresh the page.
4. Use the autofill.

It works fine with development bundles. Same demo, with only Emotion and Material-UI changed to development versions: https://jsfiddle.net/284y3evu/1/

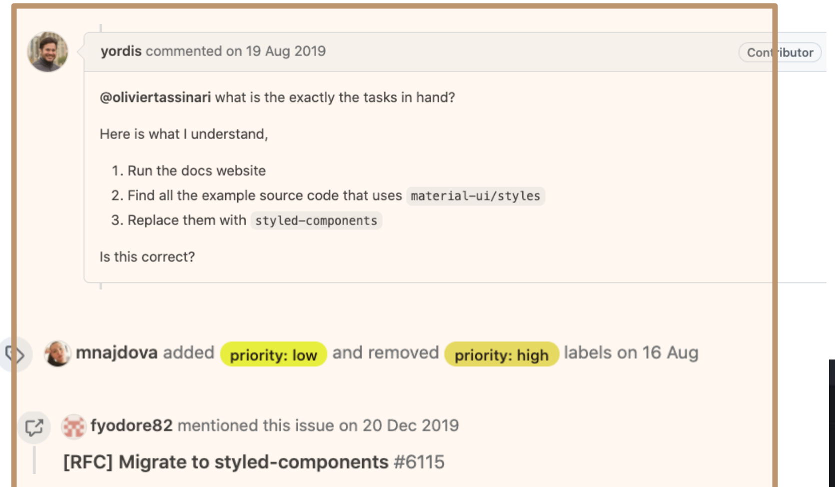
Issue的标题和描述: 是否清晰，是否提供了足够的信息。



Labels

docs · good to take · priority: low

Issue的标签:中低难度类, Doc类更可能是GFI;
Difficult/Important, Coding, Build类最不可能是GFI。



yordis commented on 19 Aug 2019 · Contributor

@oliviertassinari what is the exactly the tasks in hand?

Here is what I understand,

1. Run the docs website
2. Find all the example source code that uses `material-ui/styles`
3. Replace them with `styled-components`

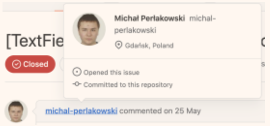
Is this correct?

mnajdova added priority: low and removed priority: high labels on 16 Aug

fyodore82 mentioned this issue on 20 Dec 2019

[RFC] Migrate to styled-components #6115

Issue是否有项目维护者的参与: 讨论、打标签、提及等。



[TextFie

Closed

Michał Perlakowski michal-perlakowski

Gdańsk, Poland

Opened this issue

Committed to this repository

michal-perlakowski commented on 25 May

背景信息: issue报告者的经验。
在该项目中具有少量经验的开发者提出的issue更可能是GFI。

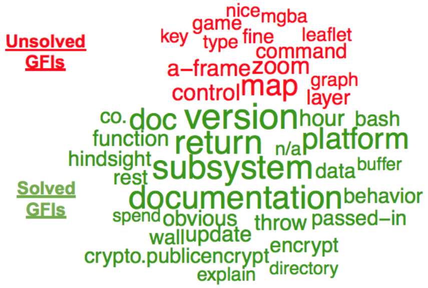
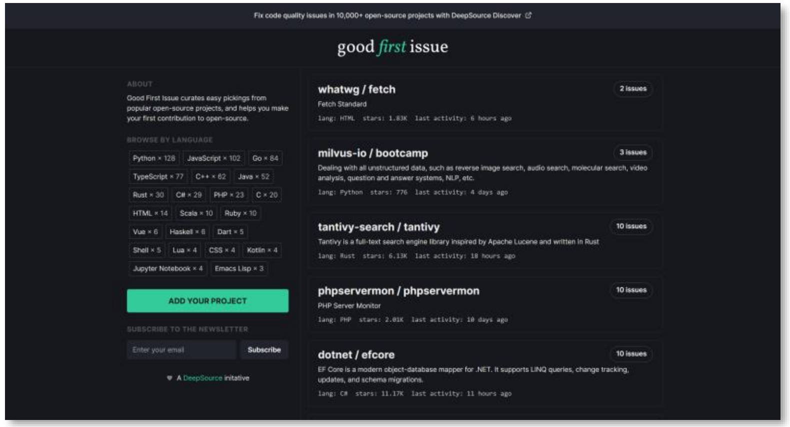


Figure 10: Comparison of solved GFIs and unsolved GFIs. The top 50 keywords with the highest frequency are shown.



Fix code quality issues in 10,000+ open-source projects with DeepSource Discover

good *first* issue

ABOUT

Good First Issue curates easy pickings from popular open-source projects, and helps you make your first contribution to open-source.

BROWSE BY LANGUAGE

Python × 128 · JavaScript × 102 · Go × 84 · TypeScript × 77 · C++ × 62 · Java × 52 · Rust × 30 · C# × 29 · PHP × 23 · C × 20 · HTML × 14 · Scala × 10 · Ruby × 10 · Vue × 6 · Haskell × 6 · Dart × 5 · Shell × 5 · Lua × 4 · CSS × 4 · Kotlin × 4 · Jupyter Notebook × 4 · Emacs Lisp × 3

ADD YOUR PROJECT

SUBSCRIBE TO THE NEWSLETTER

Enter your email · Subscribe

A DeepSource initative

whatwg / fetch · 2 issues
Fetch Standard
lang: HTML stars: 1.83K last activity: 6 hours ago

milvus-io / bootcamp · 3 issues
Dealing with all unstructured data, such as reverse image search, audio search, molecular search, video analysis, question and answer systems, NLP, etc.
lang: Python stars: 776 last activity: 4 days ago

tantivy-search / tantivy · 10 issues
Tantivy is a full-text search engine library inspired by Apache Lucene and written in Rust
lang: Rust stars: 6.13K last activity: 18 hours ago

phpservermon / phpservermon · 10 issues
PHP Server Monitor
lang: PHP stars: 2.01K last activity: 10 days ago

dotnet / efcore · 10 issues
EF Core is a modern object-database mapper for .NET. It supports LINQ queries, change tracking, updates, and schema migrations.
lang: C# stars: 11.17K last activity: 11 hours ago

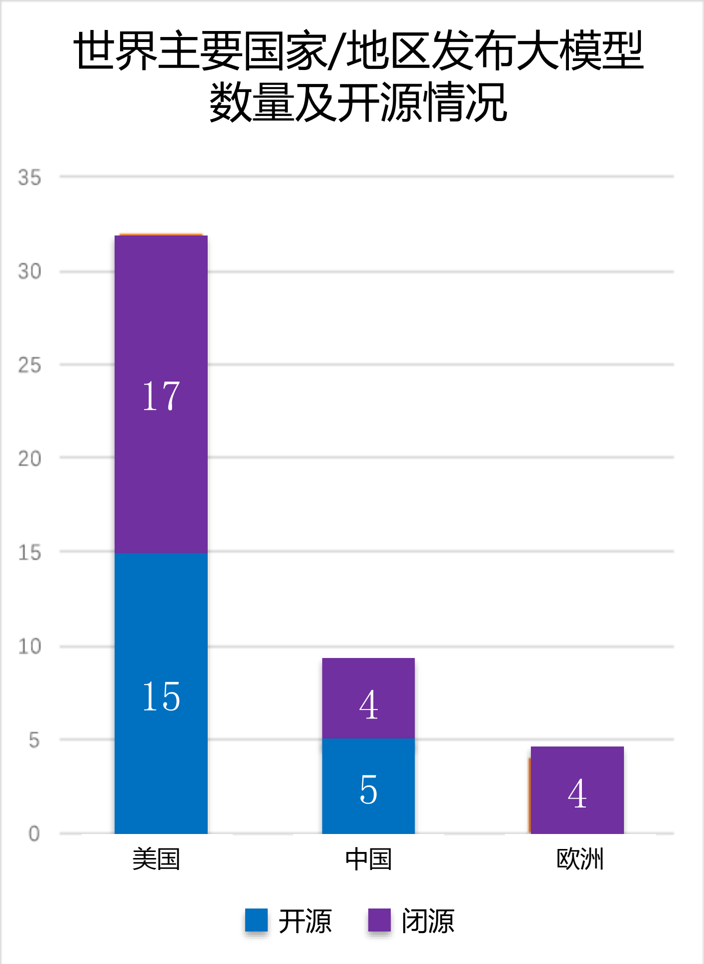
https://github.com/osslab-pku/gfi-bot

模型(f): 回归模型(XGBoost等)：$p=f(x)$
训练集: 历史issue,令被新人解决的issue的p=1,其余issue的p=0
测试点: 任一未被分配的open issue

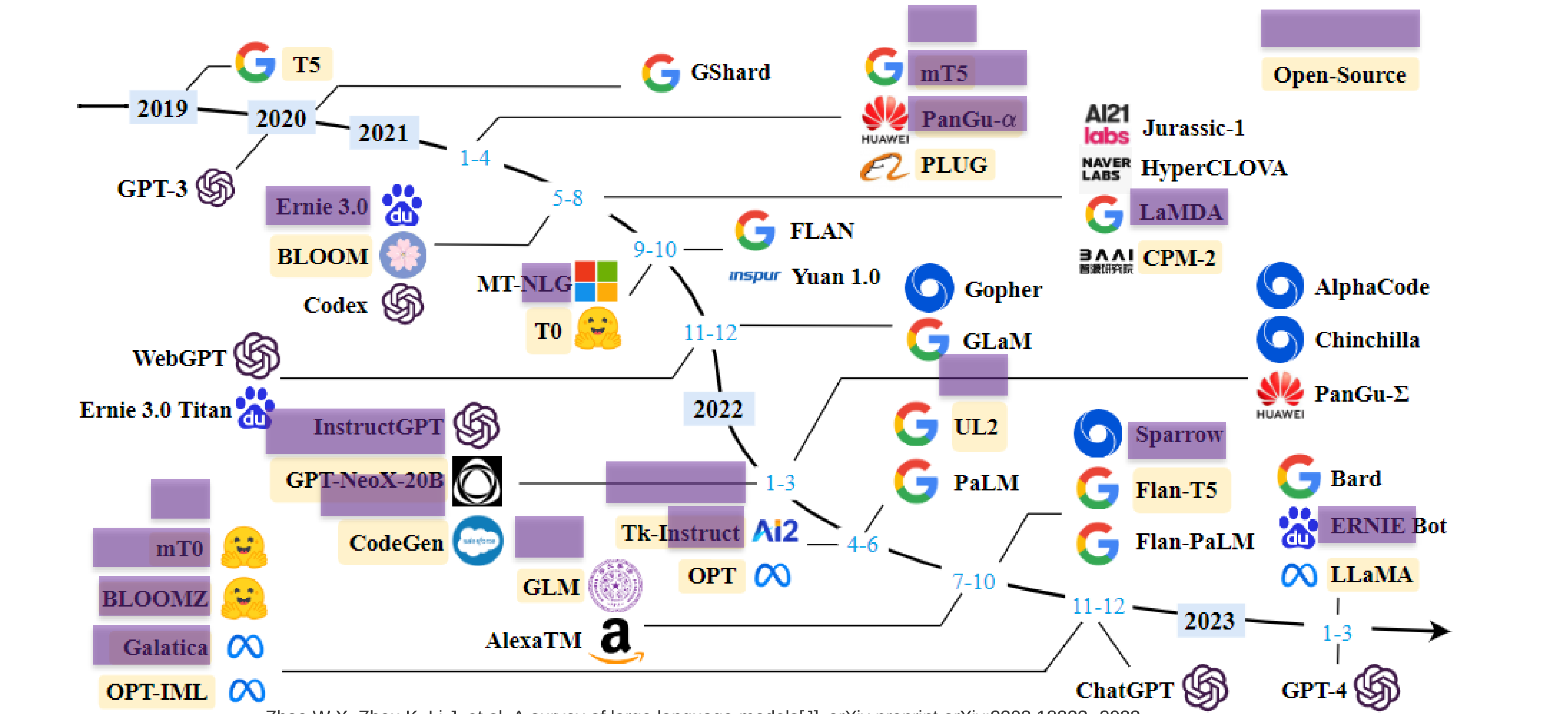
输出(p): 一个issue是GFI的概率

He et al. GFI-Bot: Automated Good First Issue Recommendation on GitHub. ESEC/FSE’2022
Xiao et al. Recommending good first issues in GitHub OSS projects. ICSE’2022
Tan et al. A First Look at Good First Issues on GitHub. FSE’2020

ChatGPT颠覆所有--每个任务都会被问: ChatGPT能做吗?


世界主要国家/地区发布大模型数量及开源情况

百亿参数以上大模型发布时间线



Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

“就像工业革命一样，大模型将会被各行各业广泛应用，带来生产力的巨大提升，并深刻改变我们的生活方式。”

“在未来，人工智能大模型将为「万物互联」提供解决方案，任意的设备都可以像人一样能听会说、能理解会思考，将极大推动万物互联、大家公认的IT产业发展的第六次浪潮，人工智能一定会以解决人类刚需而更深刻地载入史册。”

目录

CONTENTS

01. 开源的必要性和正义性

02. 开源的挑战及研究

03. 开源的定义

开源软件

- 开源软件是一种**源代码可以自由获取**的计算机软件。

- 发布开源软件需要附带**开源许可证**：
  - 开源许可证是对开源软件的知识产权进行规范和约束的法律合同：甲方是软件版权所有者，乙方是软件用户。
  - 软件的版权持有人在开源许可证的规定之下允许用户使用、修改以及分发该软件。许可证定义了开源软件用户的权利和义务。
  - 软件知识产权：版权，专利，商标。

- 开源许可证通常符合**开源的定义**的要求。

开源的定义

- **开源不仅意味着可以访问源代码，开源软件的分发条款必须满足下述条件:**

（1）自由再发行；

（2）程序必须包含或方便取得源代码；

（3）许可证必须允许更改和派生程序；

（4）保护作者源代码的完整性；

（5）无个人或团体的歧视；

（6）无领域歧视；

……

（8）许可证不能限制其他软件；

（10）许可证需要是技术中立的。

open source initiative®

https://opensource.org/osd

The Open Source Definition

-- https://opensource.org/osd

Introduction

Open source doesn't just mean access to the source code. The distribution terms of open-source software must comply with the following criteria:

1. Free Redistribution

The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

2. Source Code

The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost, preferably downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

3. Derived Works

The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

4. Integrity of The Author's Source Code

The license may restrict source-code from being distributed in modified form only if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

5. No Discrimination Against Persons or Groups

The license must not discriminate against any person or group of persons.

6. No Discrimination Against Fields of Endeavor

The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

7. Distribution of License

The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

8. License Must Not Be Specific to a Product

The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

9. License Must Not Restrict Other Software

The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.

10. License Must Be Technology-Neutral

No provision of the license may be predicated on any individual technology or style of interface.

Last modified, 2007-03-22

开源许可证需要符合开源的定义的要求

- 开源社区存在大量不同类型的许可证，OSI 认证的开源许可证已有 126 个[1]，在开源项目中的使用率 > 80%。
  - 宽松型，Apache/MulanPSL，关键特点：分发可以闭源
  - 传染型，GPL/MulanPubL，关键特点：分发必须开源

木兰宽松许可证(宽松型许可证)：http://license.coscl.org.cn/MulanPSL2
木兰公共许可证(传染型许可证)：http://license.coscl.org.cn/MulanPubL-2.0

均采用中英文双语表述，以中文为准

【1】 https://spdx.org/licenses/?ivk_sa=1024320u, 2021/03/07

木兰许可证系列：中英文双语

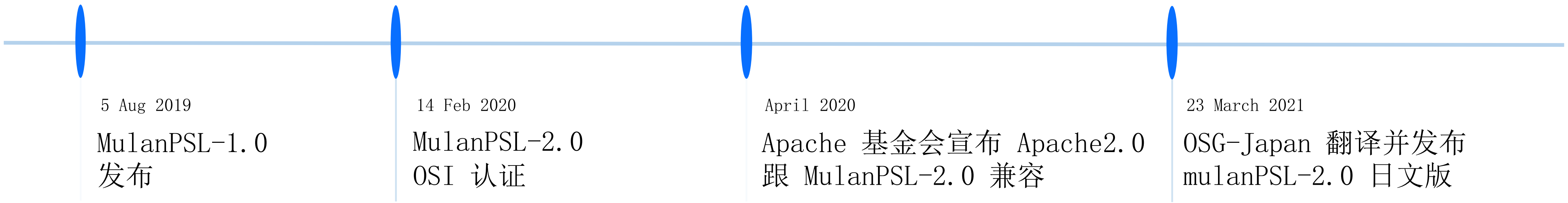
大背景：

本土企业需求，中文开源社区的发展和成长需求。

具体需求：

中文解释权，少/无风险的开源项目和产品的发展需求。

第一个获得OSI认可的本土开源许可证：木兰宽松许可证MulanPSL2



5 Aug 2019

MulanPSL-1.0
发布

14 Feb 2020

MulanPSL-2.0
OSI 认证

April 2020

Apache 基金会宣布 Apache2.0
跟 MulanPSL-2.0 兼容

23 March 2021

OSG-Japan 翻译并发布
mulanPSL-2.0 日文版

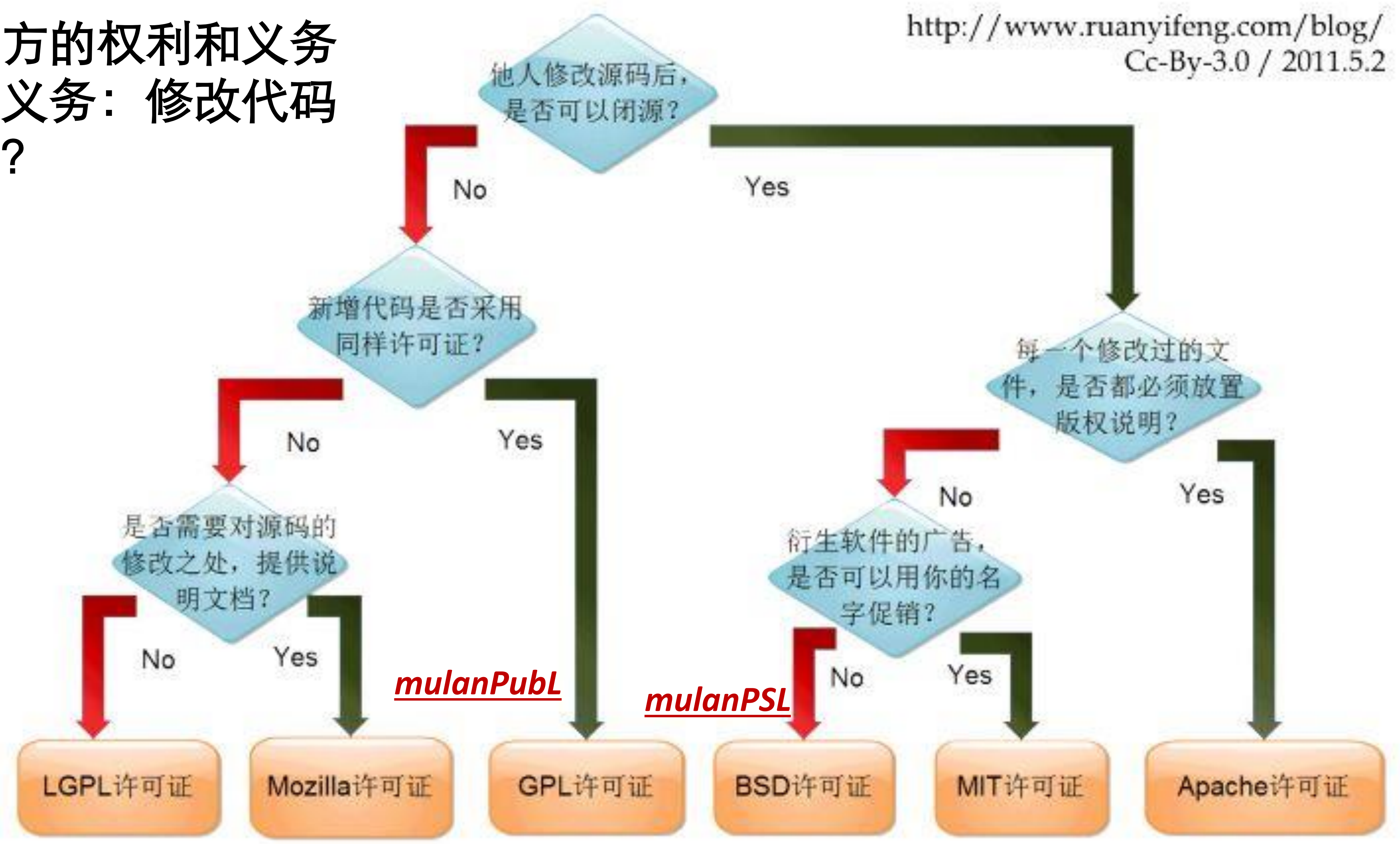
国内社区广泛支持

木兰开源社区 MULAN OPEN SOURCE　gitee　确实社区　iHub　开源社

100k+ 项目采用了 MulanPSL-2.0，覆盖云计算、大数据、AI、OS 等，例如，openEuler，openGauss，香山(RISK-V处理器)。

不同开源许可证的使用

- 合同：定义乙方的权利和义务
- 主要区别在于义务：修改代码是否需要开源?



http://www.ruanyifeng.com/blog/
Cc-By-3.0 / 2011.5.2
他人修改源码后，是否可以闭源？
No
Yes
新增代码是否采用同样许可证？
No
Yes
每一个修改过的文件，是否都必须放置版权说明？
No
Yes
是否需要对源码的修改之处，提供说明文档？
No
Yes
衍生软件的广告，是否可以用你的名字促销？
No
Yes
mulanPubL
mulanPSL
LGPL许可证
Mozilla许可证
GPL许可证
BSD许可证
MIT许可证
Apache许可证

建议阅读的文献

- *The Cathedral & the Bazaar (*大教堂与集市*)*. Raymond, E.S. (1999). O'Reilly Retrieved from http://www.catb.org/~esr/writings/cathedral-bazaar/
- 人月神话. 弗雷德里克·布鲁克斯. 出版社: 清华大学出版社. 译者: 汪颖. 出版年: 2002-11. ISBN: 9787302059325
- 开源的成功之路（The Success of Open Source），史蒂文（美国），外语教学与研究出版社， 2007-6，ISBN: 9787560066363.
- A. Mockus, R. T. Fielding, and J. Herbsleb, “Two case studies of open source software development: Apache and Mozilla,” ACM Trans. Softw. Eng. Methodol. vol. 11, no. 3, pp. 1–38, Jul. 2002.
- Minghui Zhou and Audris Mockus. Who Will Stay in the FLOSS Community? Modelling Participant's Initial Behaviour. IEEE Transactions on Software Engineering. vol.41, no.1, pp.82-99, Jan. 1 2015.
- 周明辉,张伟,尹刚. 开源软件的量化分析.中国计算机学会通讯.第12卷,第2期. 2016年2月.