

开源生态大数据智能分析

——从数据到知识，从知识到智能

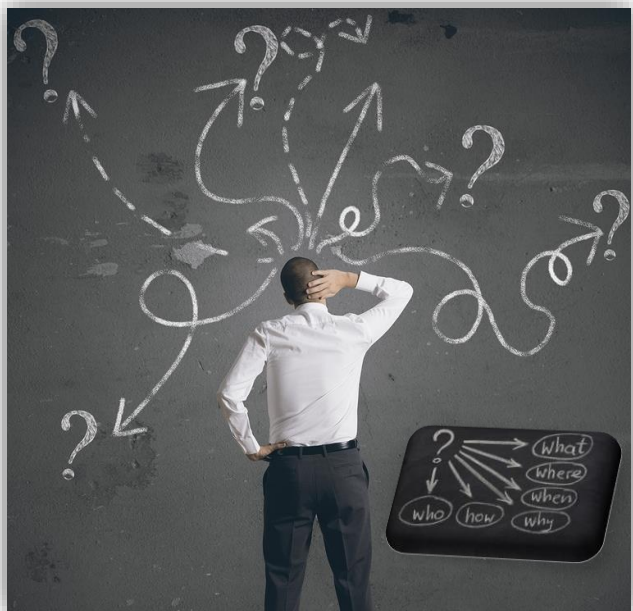
报告人：余 跃
国防科技大学



国防科学技术大学
National University of Defense Technology



经验主义
Empiricism



经验?



猜想?

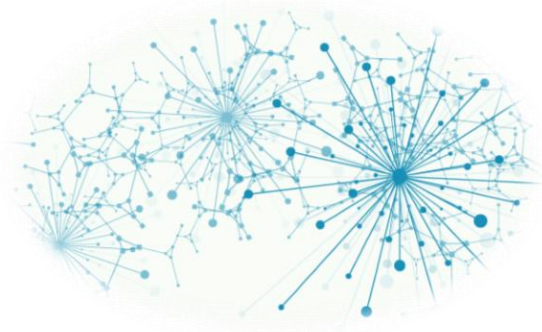
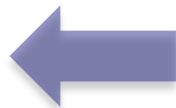
直觉驱动决策
intuition-driven



大数据时代，如何让数据释放智能？

- Software Intelligence: The Future of Mining Software Engineering Data, FoSER, 2010
- Tim Menzies, Data Science for Software Engineering, ICSE Tutorial, 2013

❖ 软件生态大数据

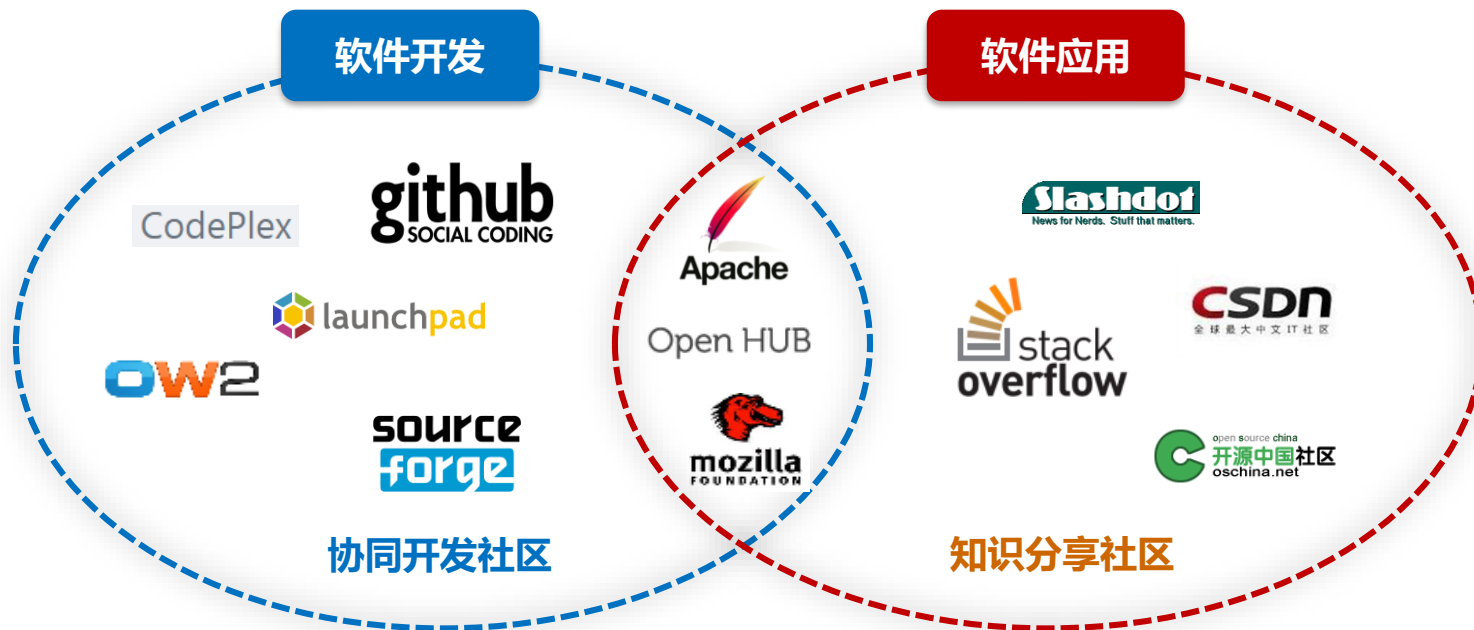


❖ 多维量化度量体系

❖ 智能化服务案例



从数据到知识，从知识到智能



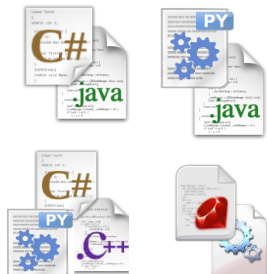
➤ OSSEAN: Mining Crowd Wisdom in Open Source Communities, SOSE 2015



数据的抽象总结：协同开发社区和知识分享

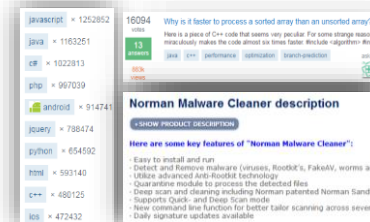
代码仓库

- ✓ 源代码
- ✓ 代码注释
- ✓ 配置文件
- ✓ 项目文档



Web语义

- ✓ 软件问答
- ✓ 功能介绍
- ✓ 社会化标签



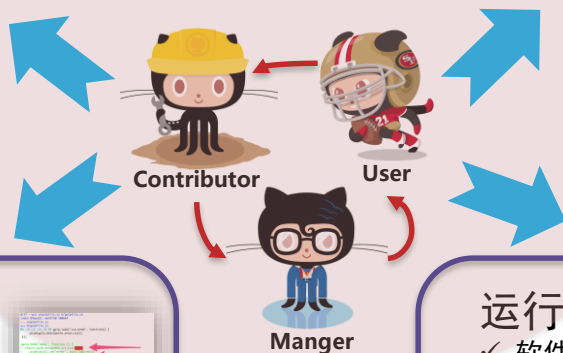
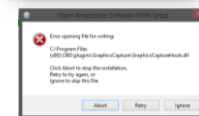
开发历史库

- ✓ 提交、测试历史
- ✓ 问题解决 (bug, feature)
- ✓ 交流讨论:
 - 代码审查
 - 需求获取



运行时日志

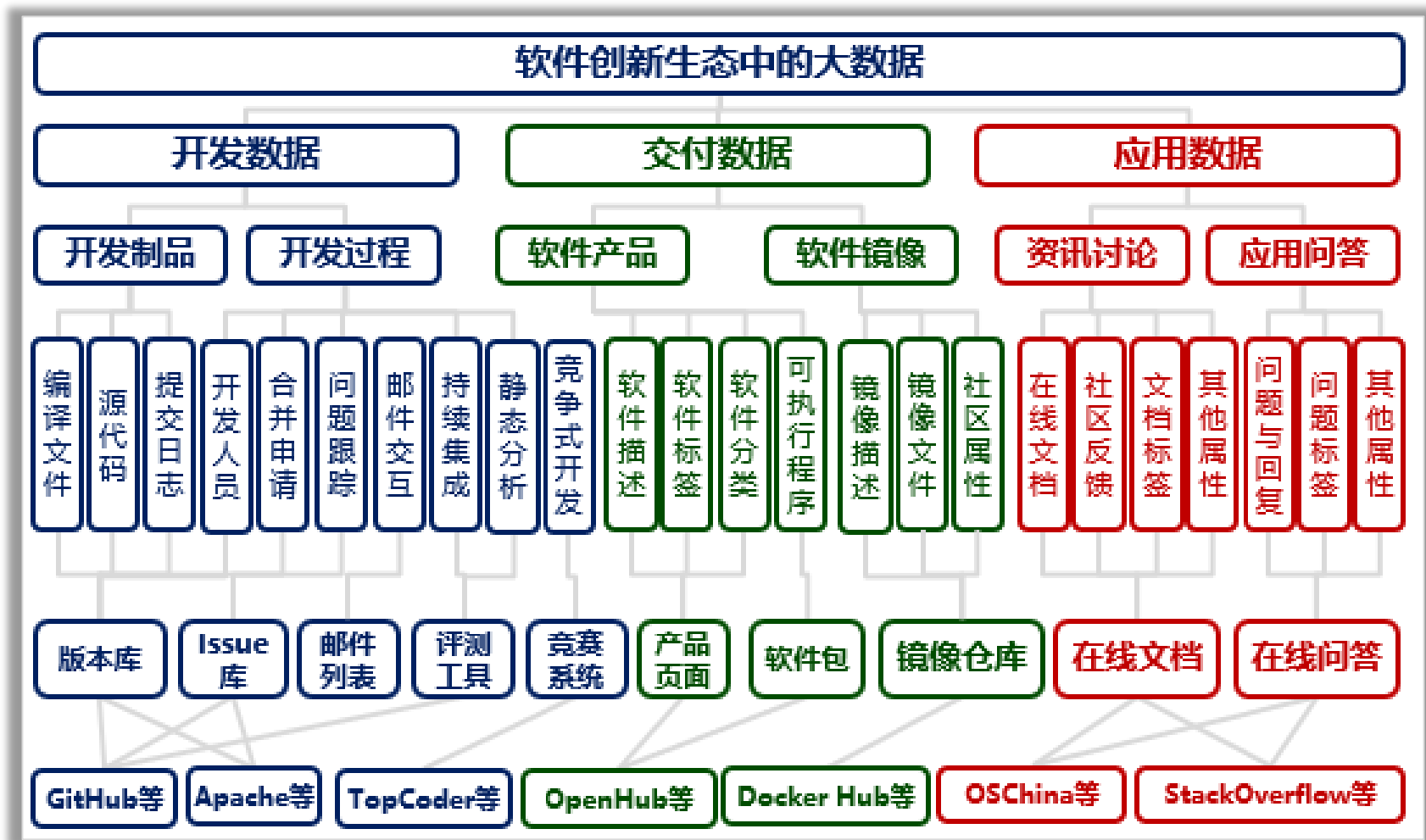
- ✓ 软件崩溃
- ✓ 内存峰值
- ✓ 访问频率



Web网页爬虫

RESTful JSON API

版本控制系统



碎片化原始数据



工作量 workload

变更复杂度 size, complexity

开发者信息 expertise, interest

目前html5 中 stream module不支持在跨域cors情况下带cookie, 目前不管zepto和jquery都有这个配置, 希望可以加上, 原生客户端不需要此配置, 但是h5得场景下有需求, 希望可以合进去, 谢谢!

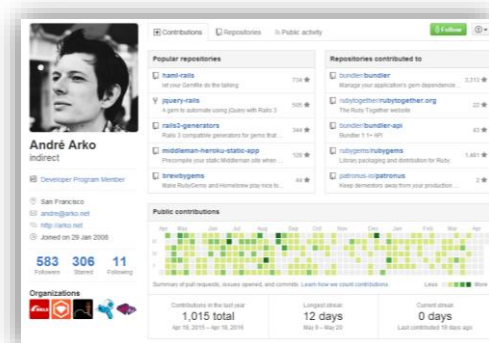


变更细节 details

```
6  html5/browser/extend/api/stream.js
@@ -69,6 +69,11 @@ function xhr (config, callback) {
69  xhr.responseType = config.type
70  xhr.open(config.method, config.url, true)
71
72  + // cors cookie support
73  + if (config.withCredentials === true) {
74  +   xhr.withCredentials = true
75  + }
76  +
77  const headers = config.headers || {}
78  for (const k in headers) {
79    xhr.setRequestHeader(k, headers[k])
@@ -177,6 +182,7 @@ const stream = {
177  * - headers {obj}
```

变更目地 purpose

- 1) 改动了哪些文件
- 2) 具体的代码修改信息
- 3) 代码的上下文环境



Public contributions

583 306 11

Contributions in the last year: 1,015 total

Largest streak: 12 days

Current streak: 0 days

精准数据采集-案例

C

首页 开源项目 问答

开源项目

全部项目分类

编程语言

Web应用开发

手机/移动开发

iOS代码库

程序开发

开发工具

jQuery 插件

建站系统

企业应用

服务器软件

数据库相关

应用工具

插件和扩展

游戏/娱乐

管理和监控

其他开源

反编译工具 (29)

Git开源工具 (190)

PHP开发工具 (133)

Python开发工具 (132)

项目构建 (209)

GUI 测试工具 (31)

代码混淆和加密 (31)

安装制作工具 (67)

程序文档工具 (141)

持续集成系统 (44)

Java开发工具 (206)

C/C++开发工具 (111)

Perl开发工具 (27)

单元测试工具 (206)

性能测试和优化 (216)

UML/模型工具 (132)

编译器 (176)

语法解析工具 (48)

SQL注入工具 (19)

.NET开发工具 (51)

Ruby/Rails开发工具 (78)

BUG跟踪管理 (78)

测试工具 (407)

程序调试工具 (191)

界面原型设计工具 (28)

代码管理分析/审查/优化 (237)

汇编开发工具 (13)

软件资讯

共 46392 个项目

更多

推荐

基于 IntelliJ 生成插件 Easy

所有编程语言

高性能网络 HP-Socket 是 件和Agent组件

GUI 测试工具 StoryText

StoryText (前身是 PyUseCase) 是一个使用 PyGTK、Tkinter、wxPython、Swing、SWT 和 Eclipse RCP 编写的 GUI 测试工具

Previously unseen actions; provide names for the interesting ones

Widget Type	Identified By	Action Performed	Usecase Name
TreeView	Type=TreeView	clicked on row	select bug
ToggleButton	Label=VERI	unchecked	hide verified bugs
ToggleButton	Label=RESO	unchecked	hide resolved bugs
TreeView	Type=TreeView	clicked column header 'description'	sort bugs by description
ToggleButton	Label=RESO	checked	show resolved bugs
Window	Title=Who needs jira anyway?	closed	close application

Current Usecase Preview
select bug 123014
hide verified bugs
hide resolved bugs
sort bugs by description
show resolved bugs
close application

上次更新: 2014年05月08日 收藏 12

GUI 自动化测试框架 PyAutoGUI

PyAutoGUI是一个简单易用, 跨平台的可以模拟键盘鼠标进行自动操作的python库. PyAutoGUI中文文档: 《PyAutoGUI——让所有GUI都自动化》 <https://m...>

收藏 8

持续开发 k8s 应用的命令 具 Skaffold

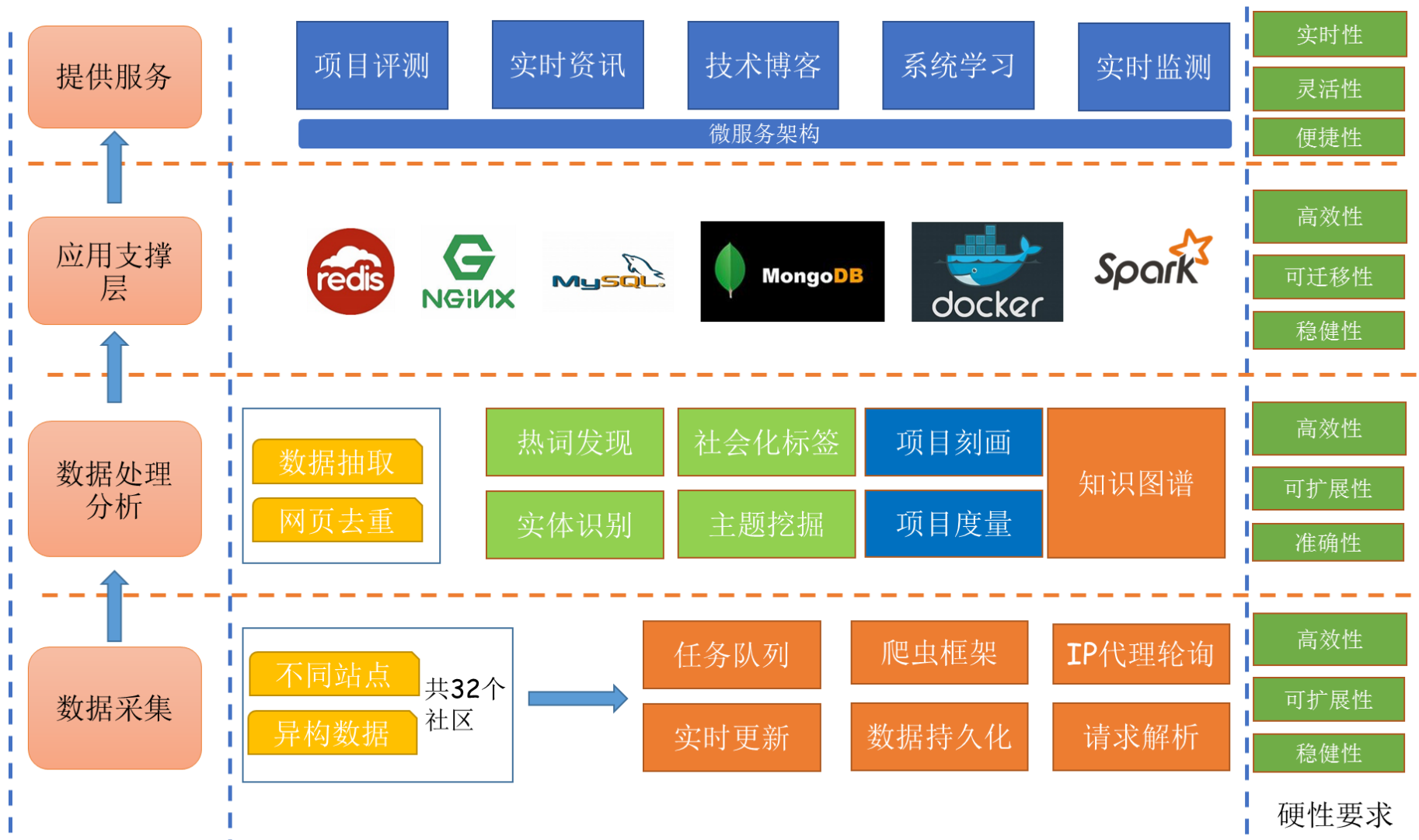
客户端组 共 C...

精准数据采集-案例



特点：面向多个站点的定向模块

开源大数据采集-体系结构

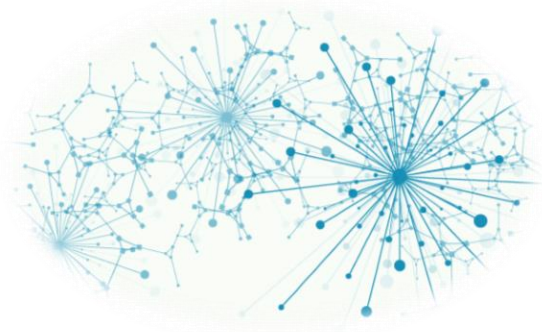


❖ 软件生态大数据

❖ 多维量化度量体系



❖ 智能化服务案例



从数据到知识，从知识到智能

多维量化-度量体系



社区影响力

该项目在开发与传播过程中的社会影响力

watchers
关注人数

stars
收藏人数

forks
派生人数

Twitter
关注度

Facebook
影响力

Reddit
关注度

项目开发（托管）社区指标

用户社区（社交网站）指标



开发活跃度

该项目代码开发和维护的增长动态

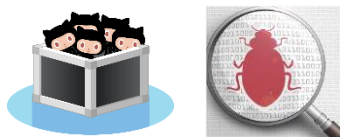
代码提交增长量
commits

合并请求增长量
pull requests

开发任务增长量
issues

新增版本数量
releases

涵盖代码提交、开发任务和软件成型三个方面的指标



软件健康度

该项目在特定阶段的软件质量情况

静态分析
static issues

动态分析
dynamic bugs

缺陷修复率
bug-fix ratio

缺陷修复速度
bug-fix time

项目源代码的客观质量指标

项目开发过程的主观质量指标



该项目开发团队的合理性与稳定性

团队健康度

团队稳定性
stability

新人增长率
newcomer

任务平衡度
balance

团队中持续贡献者比例、新增贡献者趋势和开发任务是否分配均衡等指标



该项目开发活动与影响力传播的未来趋势

发展趋势

任务增长速率
issues

开发增长速率
commits、PR

开发社区影响力趋势
stars、watchers、forks

用户社区影响力趋势
Twitter、Facebook、Reddit

项目开发的增长趋势指标

项目在开发社区和用户社区的流行趋势指标



该项目发展到特定阶段时的成熟程度

项目成熟度

累计任务数量
issues

累计开发活动
commits、PR

项目有效开发年龄
project age

开发社区和用户社区累计关注人数
Twitter、Facebook、Reddit

项目开发与维护活动的成熟度指标

项目传播角度的成熟度指标

开源软件排行榜

Machine Learning (166)

tensorflow/tensorflow	1.00
Microsoft/CNTK	0.41
pytorch/pytorch	0.38
fchollet/keras	0.38
scikit-learn/scikit-learn	0.38
BVLC/caffe	0.36
lisa-lab/pylearn2	0.27
Theano/Theano	0.25
torch/torch7	0.20
shogun-toolbox/shogun	

Database (442)

antirez/redis	1.00
Elasticsearch	0.92
apache/cassandra	0.90
apache/hive	0.80
apache/hbase	0.78
apache/lucene-solr	0.70
neo4j/neo4j	0.68
memcached/memcached	0.63
apache/couchdb	0.61
influxdata/influxdb	

Testing Tools (404)

junit-team/junit4	1.00
jasmine/jasmine	0.96
sebastianbergmann/phpunit	0.95
wg/wrk	0.91
cloudera/hue	0.90
karma-runner/karma	0.80
mockito/mockito	0.79
andresriancho/w3af	0.70
jshint/jshint	0.70
locustio/locust	

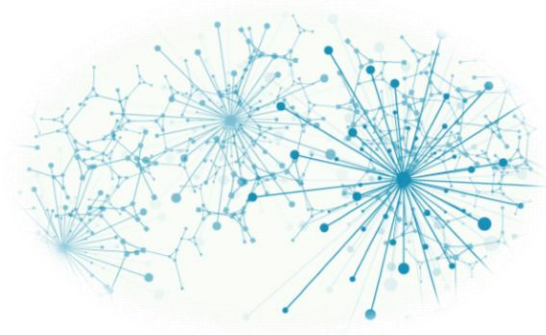
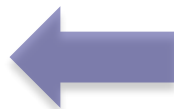
Block Chain (100)

Bitcoin	1.00
Ethereum	0.73
Cardano	0.73
Waves	0.72
Litecoin	0.71
EOS	0.69
PIVX	0.68
Zcash	0.67
Bitcoin Gold	0.66
Emercoin	

❖ 软件生态大数据

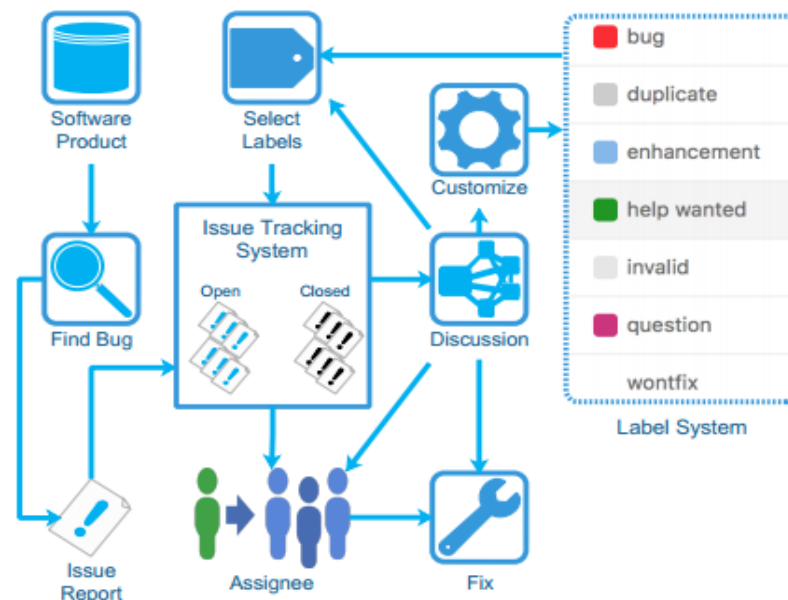
❖ 多维量化度量体系

❖ 智能化服务案例



从数据到知识，从知识到智能

❖ 轻量级缺陷管理系统



The screenshot shows a GitHub issue form. It includes a 'Title' field, a 'Write' button, a 'Preview' button, and a 'Leave a comment' section. The form also has a 'Submit new issue' button. The form is styled with a dashed border and a light gray background.

This block shows the 'Labels', 'Milestone', and 'Assignee' dropdowns from the GitHub issue form. The 'Labels' dropdown shows 'None yet'. The 'Milestone' dropdown shows 'No milestone'. The 'Assignee' dropdown shows 'No one—assign yourself'.

■ Labels

■ Assignee

■ Content


贡献指南

Contributing

code helpers 1186


We encourage you to contribute to Ruby on Rails! Please check out the [Contributing to Ruby on Rails guide](#) for guidelines about how to proceed. [Join us!](#)

现实情况




[lazonixon](#) commented on 19 Feb 2015 + 🗨️


Write fixtures that have counter_cache columns is tedious and complicated, I tried use method reset_counters in test_helper.rb but it didn't work. Same way to put it working ?



[rafaelfranca](#) commented on 19 Feb 2015 Ruby on Rails member + 🗨️

Please don't ask questions in the issue tracker. Use stack overflow or a mailing list

🚫  [rafaelfranca](#) closed this on 19 Feb 2015



[rafaelfranca](#) commented on 19 Feb 2015 Ruby on Rails member + 🗨️

Please don't ask questions in the issue tracker. Use stack overflow or a mailing list

EX: “Currently, auto-archiving cannot be used if Piwik’s authentication is configured to use the CAS plugin. I ran into this problem with authentication when running archive.php with CAS plugin enabled on my site... Add a feature to auto-archiving, so that it can succeeds when Piwik uses CAS for authentication instead of the default Login module.”

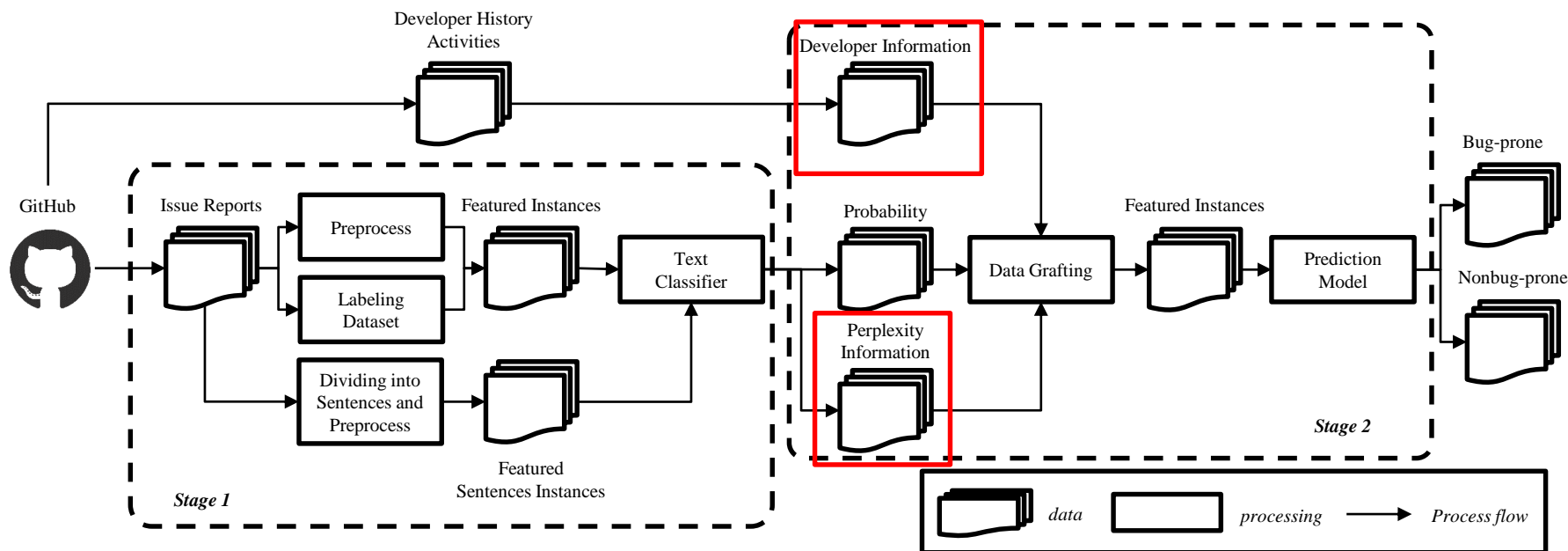
缺陷语义



特征语义

从描述中获取语义混乱度特征

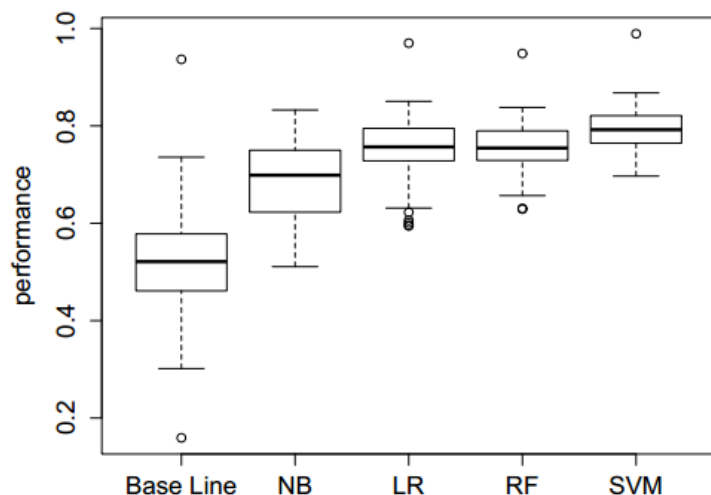
❖ 基于语义混乱度的自动分类方法



大数据分析提取
语义混乱特征



语义混乱特征加入机
器学习分类器



SVM performs best among 4 different classifiers

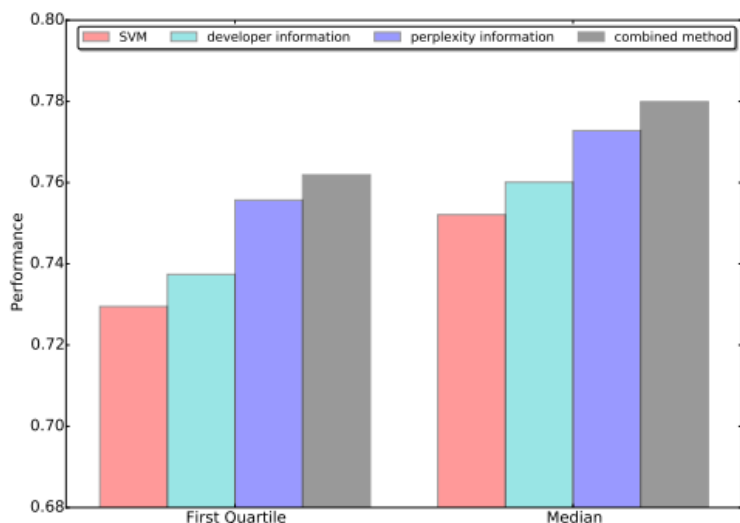
	Coeffs	Sum Sq.
(intercept)	-3.37543*	
$\log(\text{star} + \text{watch})$	0.06316	0.197
$\log(\text{issue_num})$	1.90440***	30.484***
$\log(\text{contributors})$	-0.03135	0.022
$\log(\text{age} + 0.5)$	-0.22421*	0.288
$\log(\text{commits})$	-0.34256	0.842*
$\log(\text{confuse_count} + 0.5)$	-1.83346***	134.623***
$\log(\text{med_word_count})$	0.12505	0.067
marginal R-squared	0.6798150	
conditional R-squared	0.9251896	

signif.: $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'

The number of confused issues is the major factor influencing the performance of classifier

All 2-stage methods perform better than SVM

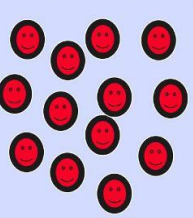
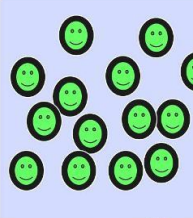
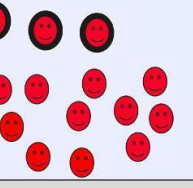
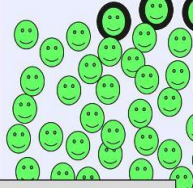
Combined method performs best among 4 methods



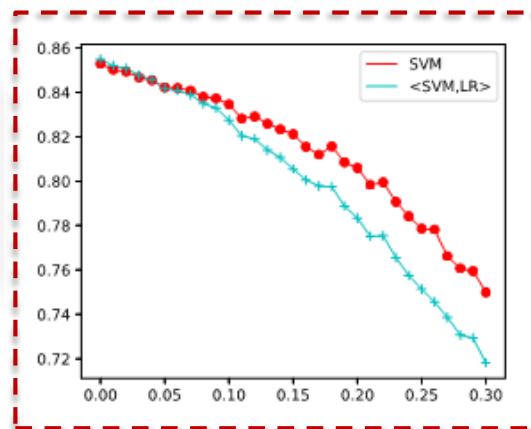
❖ 智能方法的鲁棒性提升

EX: “Currently, auto-archiving cannot be used if Piwik’s authentication is configured to use the CAS plugin. I ran into this problem with authentication when running archive.php with CAS plugin enabled on my site... Add a feature to auto-archiving, so that it can succeeds when Piwik uses CAS for authentication instead of the default Login module.”

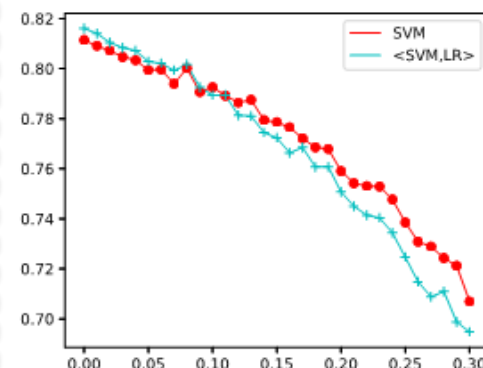


	Diseased	Not Diseased
Exposed		
Not Exposed		

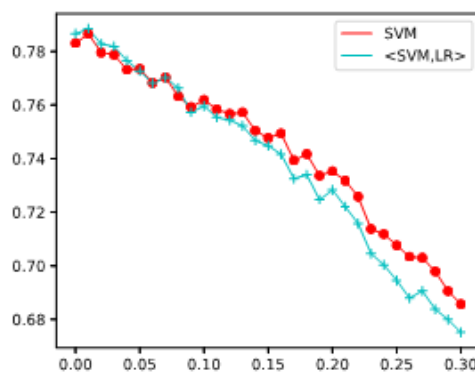
Nondifferential Misclassification of Exposure #1



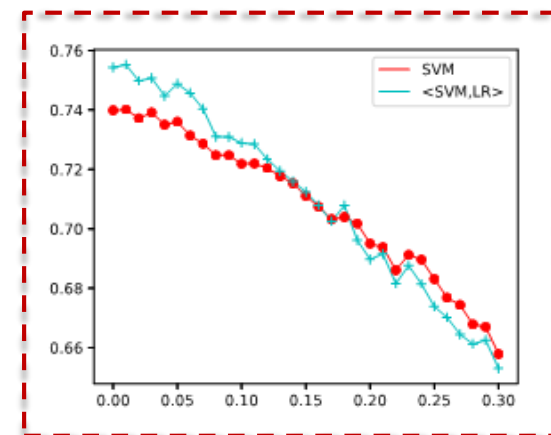
(a) Part I



(b) Part II



(c) Part III

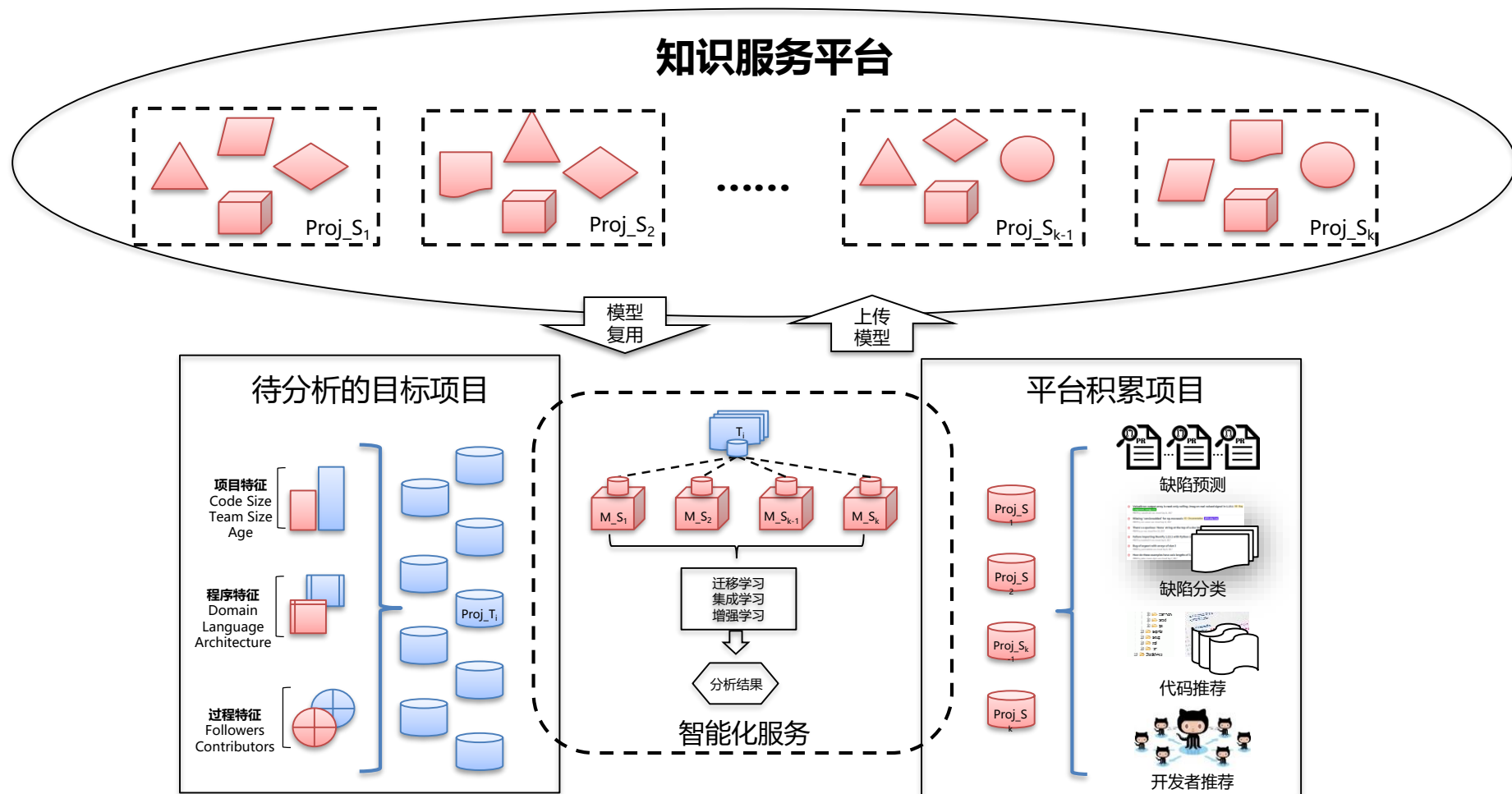


(d) Part IV

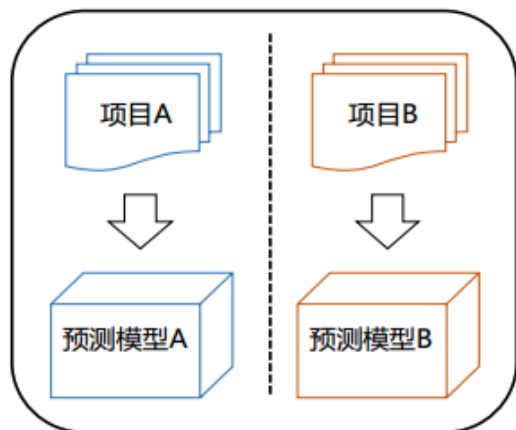
Where Is the Road for Issue Reports Classification Based on Text Mining?

❖ 目标：项目无关的知识服务

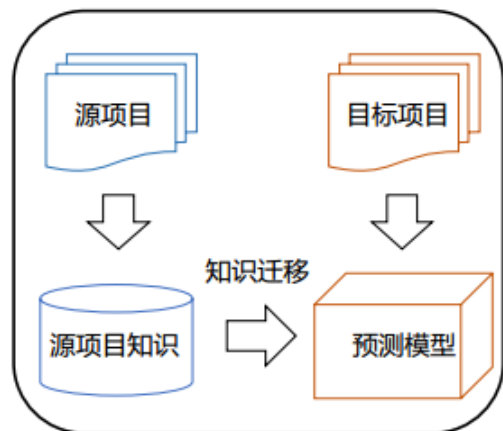
- 将跨项目、冷启动问题一网打尽



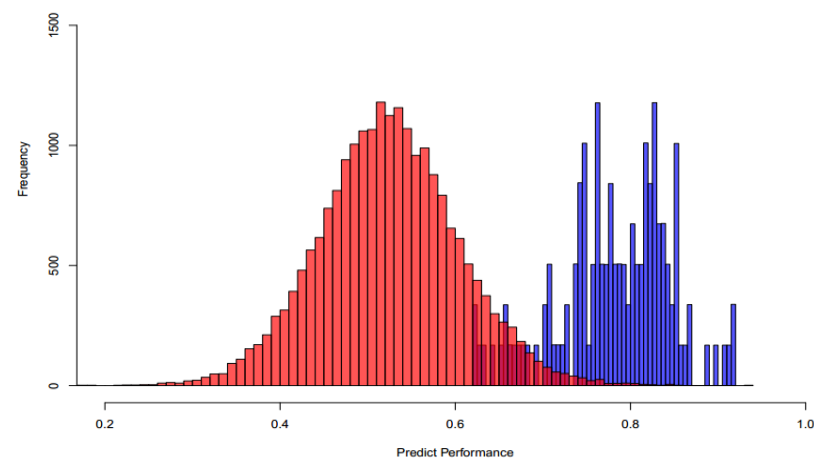
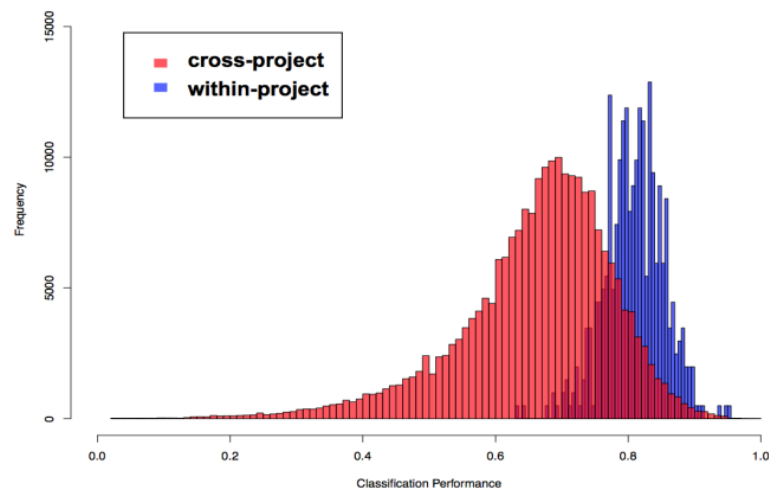
❖ 工作1：迁移性（可复用性）分析



(a) 项目内预测

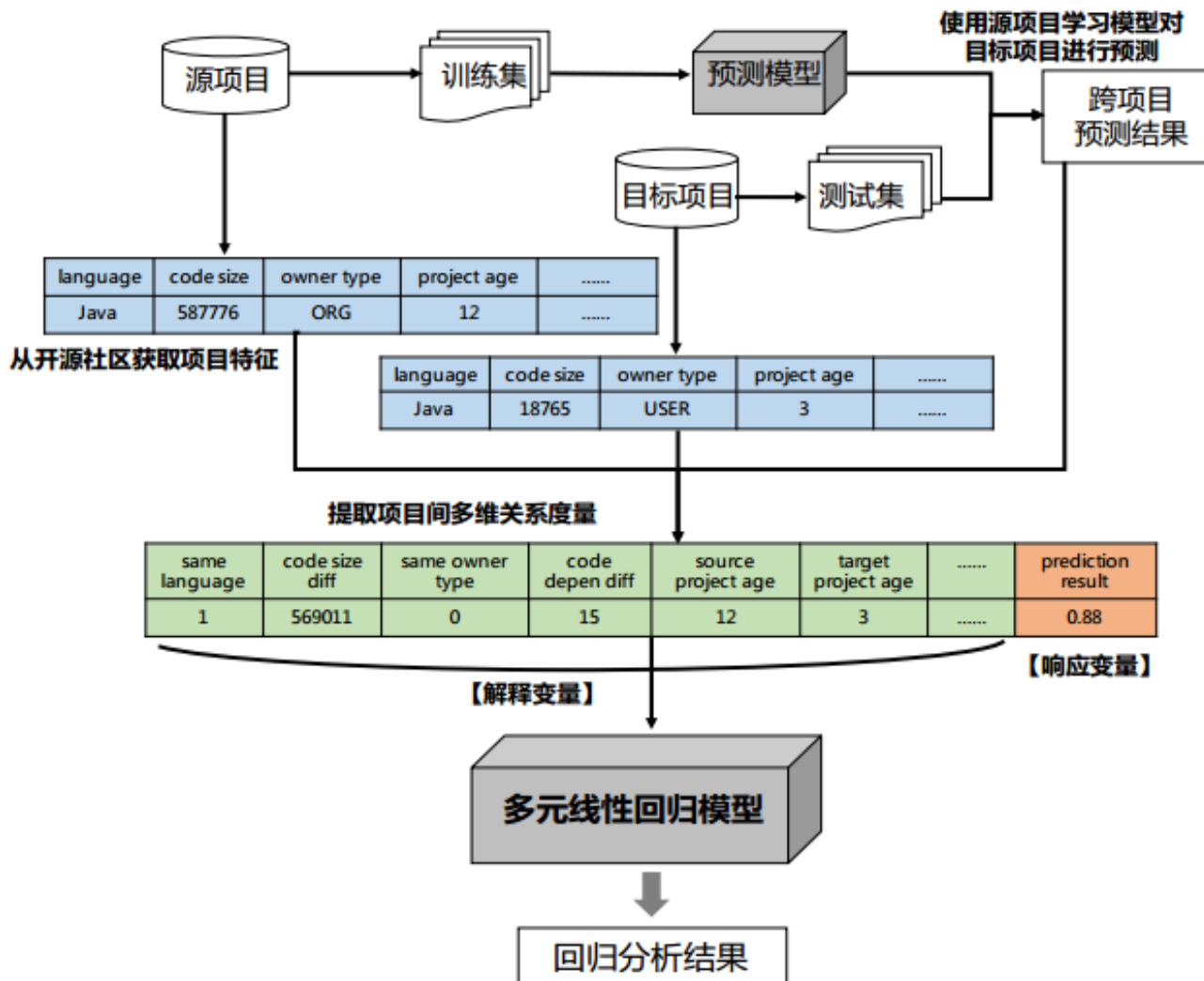


(b) 跨项目预测



❖ 工作1：迁移性（可复用性）分析

- 两个场景：缺陷分类、缺陷预测



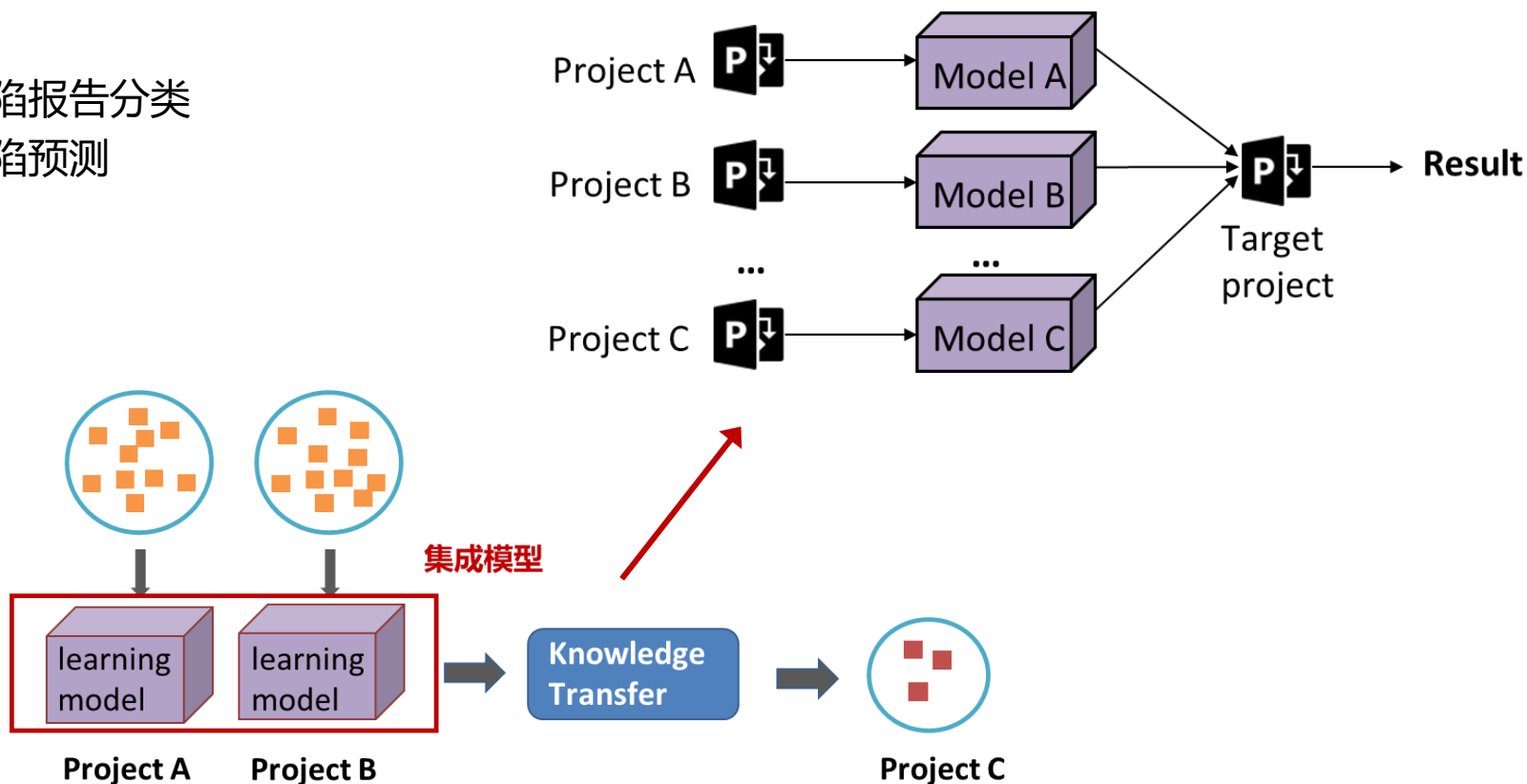
❖ 工作2：基于集成学习的知识迁移方法

方法：

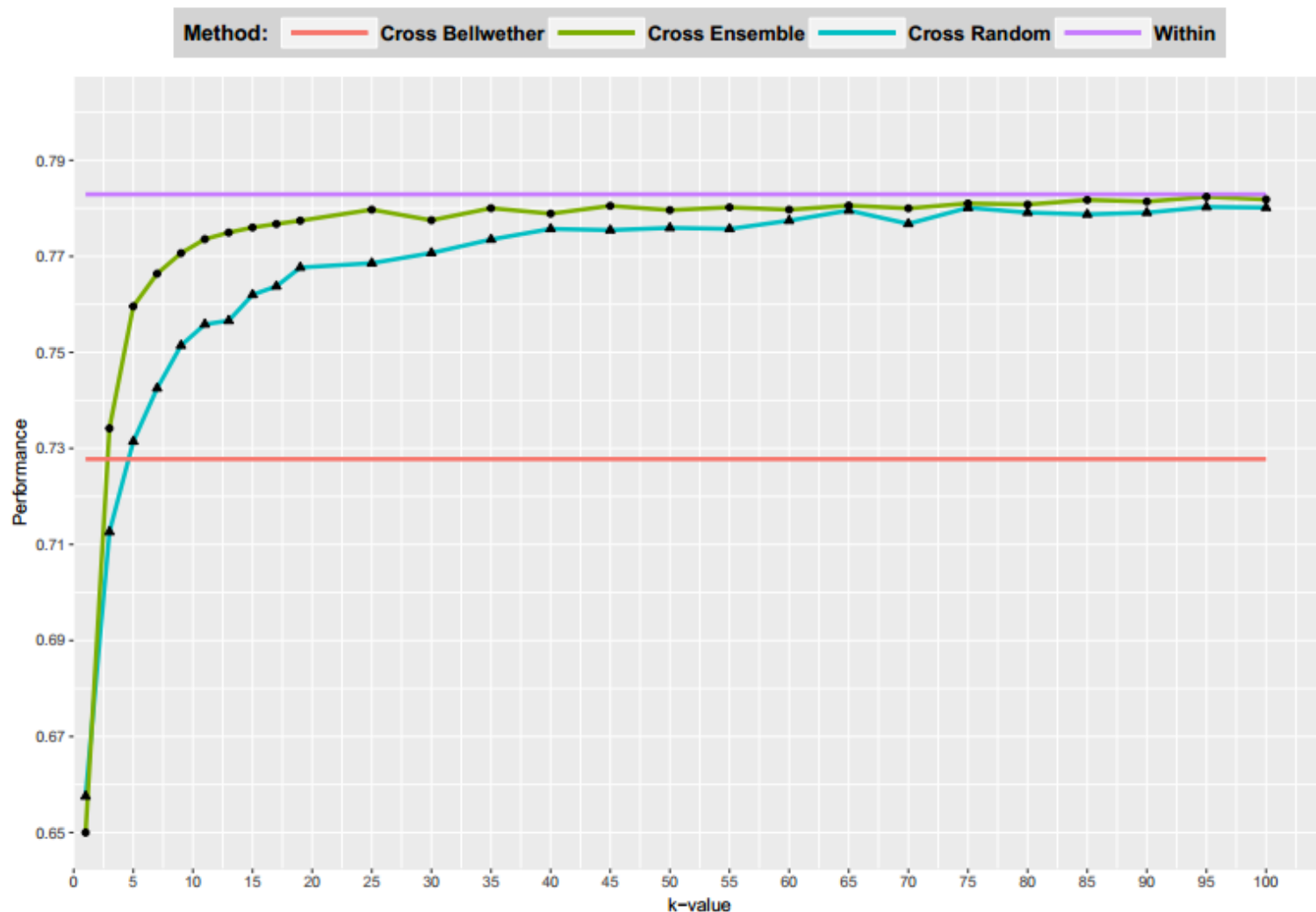
- 基于可迁移分析结果找到最相关的K个项目
- 集成多个相关项目的学习模型进行预测

应用：

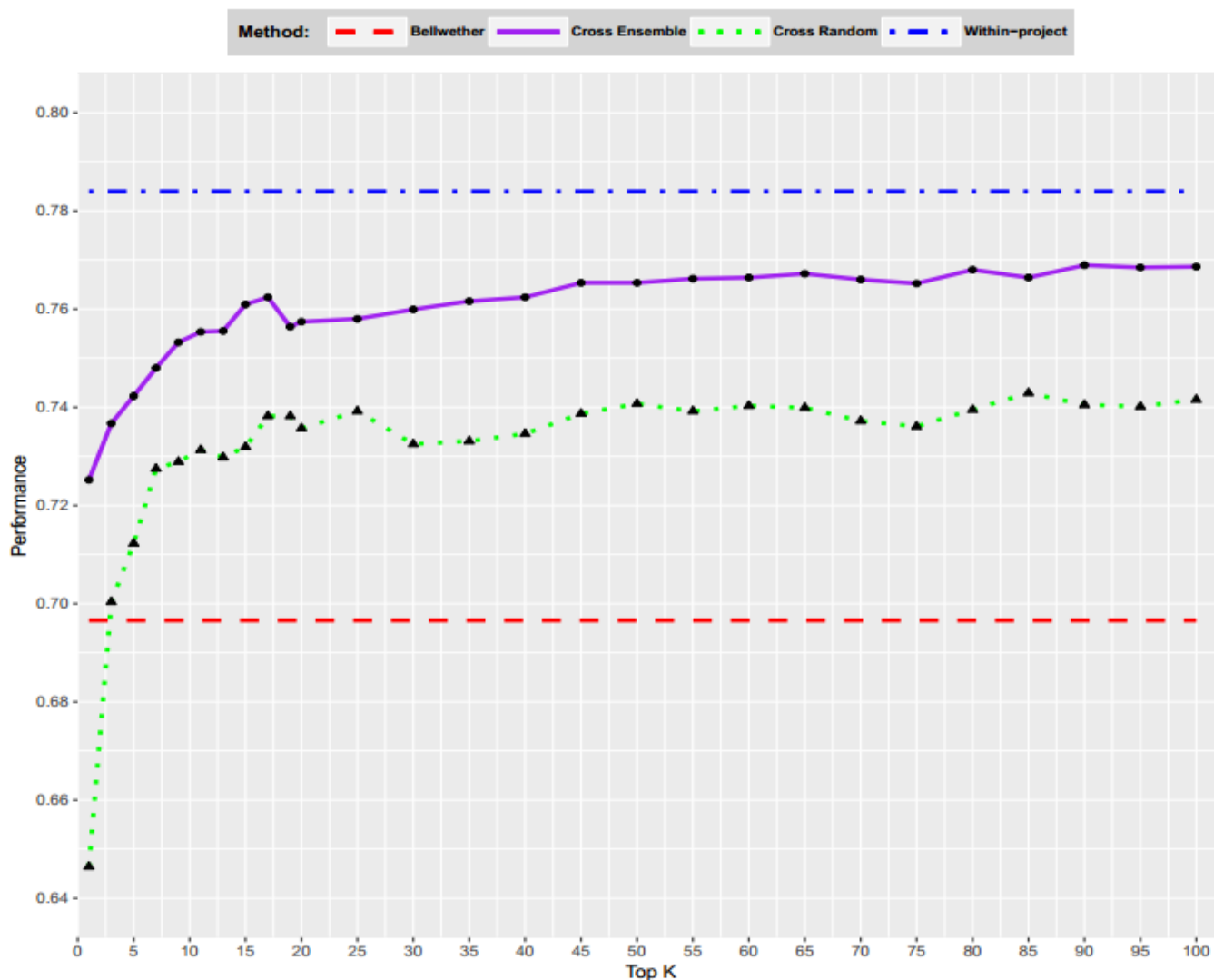
- 缺陷报告分类
- 缺陷预测



❖ 工作2：基于集成学习的知识迁移方法-缺陷分类



❖ 工作2：基于集成学习的知识迁移方法-缺陷预测



谢谢各位老师同学！

敬请批评指正！

<http://yuyue.github.io>



国防科学技术大学
National University of Defense Technology



开源社区

开源改变世界

支持开源社区生态建设，加强开源软件开发和群体化学习

开通企业版

开通高校版

最新开源资讯

最新教育动态

创新项目

创新课堂

Apache Dubbo 2.6.5 发布，分布式 RPC 服务框架

最新开源资讯

今天 15:48

GO 开源网关 API-Gateway v2.5.0-beta，提供 WEBUI

最新开源资讯

今天 15:48

GuiLite 1.1 发布：“大力”优化底层，CPU占用率低至：0% ~ 3%

最新开源资讯

今天 09:17

通知公告



全国高校绿色计算大赛
——项目挑战

全国高校绿色计算大赛——开源标注

百万专业开发资源，就在开源众包

塑造服务品牌、打造接单明星

<https://toschina.trustie.net/>