TiD 2017
质量竞争力大会

下一代软件工程中的数据智能研究和实践

DJ Ning
教授级高工
智慧城市研究中心、大数据实验室主任
中国科学院上海高等研究院

下一代软件研发
SOFTWARE DEVELOPMENT

必然

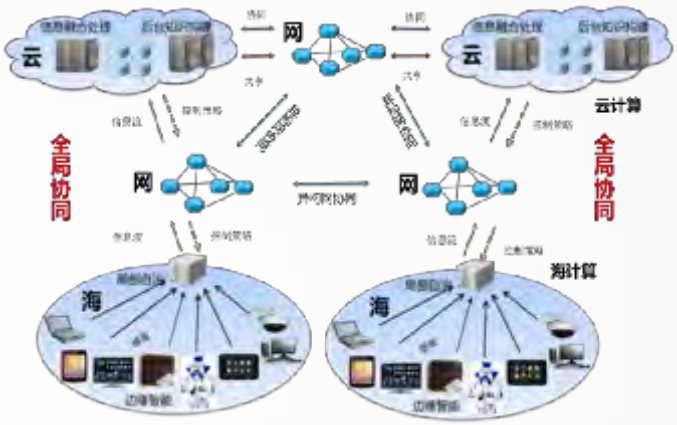
智慧经济：
模式：物联 + 数据智能 + 云服务

体验经济：
模式：交互 + VR + AI

共享经济：
模式：连接+使用权流动 + 信用体系

世界信息化发展的主要趋势

人机物三元融合



海网云协同计算

手工工具
土地
机械化

16世纪

机器装备
能源
自动化

18世纪

互联网络
信息
信息化

20世纪

IOE
数据
智能化

21世纪

时间

下一代软件工程的主要趋势

软件工程发展的主要趋势

1. 从严格的遵从软件开发过程到面向目标的智能汇聚
2. 软件定义、设计、开发、运营和业务价值交付之间的界限逐渐退化
3. 软件产品的成功交付建立在团队动态互动和协作的群体智能基础上
4. 基于开放社区的软件创新生态环境的形成
5. 基于人工智能/数据智能的软件工程技术逐渐兴起

汇报提纲

- 下一代软件工程（NSE）
- 软件工程中的数据智能
- 未来的软件工程大脑

TiD2017

下一代软件工程（NSE）概念图

驱动力：互联网+ 大数据 AI 物联网 移动计算 云计算

过程

DevOps
敏捷开发过程
敏捷实践

螺旋发展

技术

AI技术
微服务技术
物联网开发技术
移动开发技术
大数据技术

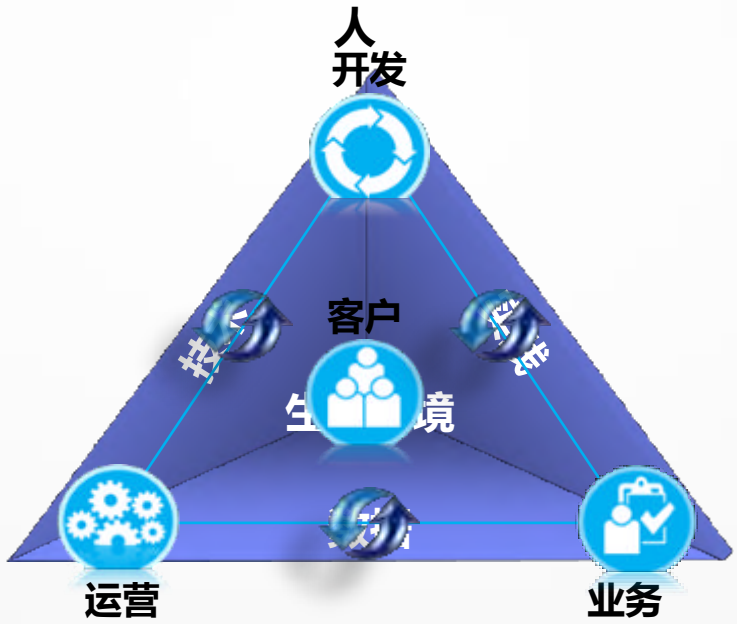
工具

互联网+工具
互联网+云平台
(BlueMix & Predix)

螺旋发展

语言

面向互联网的语言和框架
(HTML5, Spring boot & Cloud)
面向大数据的语言
（Python & R）



过程

工具

TiD2017

下一代软件工程（NSE）概念图

驱动力：互联网+ 大数据 AI 物联网 移动计算 云计算

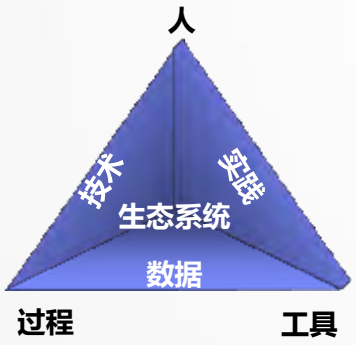
过程

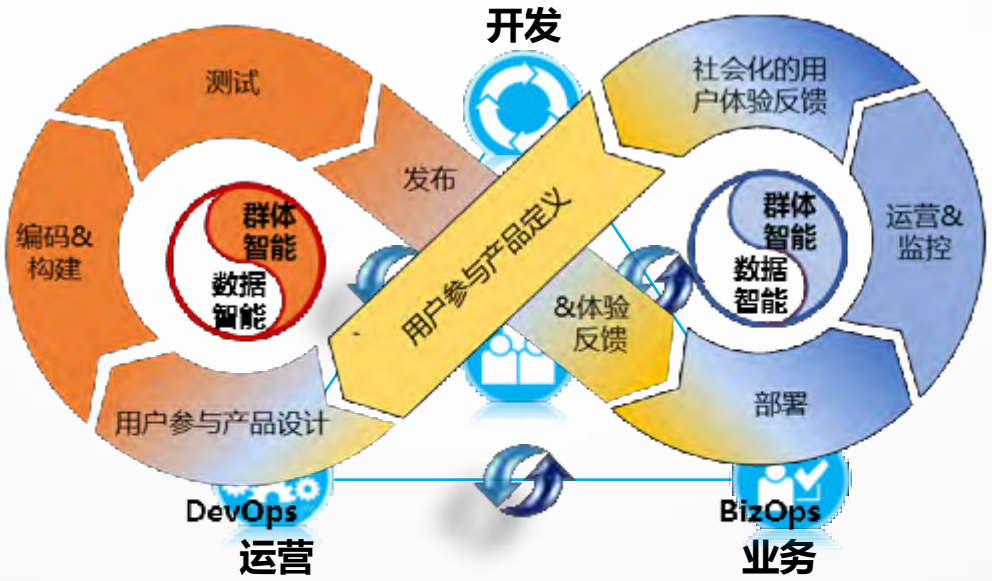
DevOps
敏捷开发过程
规范软件过程
定制化过程

螺旋发展

工具

互联网+工具
互联网+云平台
(BlueMix & Predix
CloudFoundry、Azure)





围绕价值流，整合数据流；
基于数据流，激活创新流；
关键在于系统级数据智能。

微服务技术
物联网开发技术
移动开发技术
数据分析技术

技术

面向互联网的语言和框架(HTML5, Node.js, Go,Spring)
面向大数据的语言（Python & R）

语言

TiD2017

下一代软件工程（NSE）概念图

驱动力：互联网+ 大数据 AI 物联网 移动计算 云计算

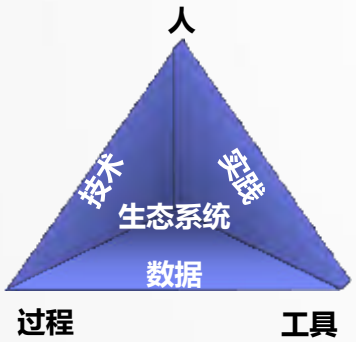
过程

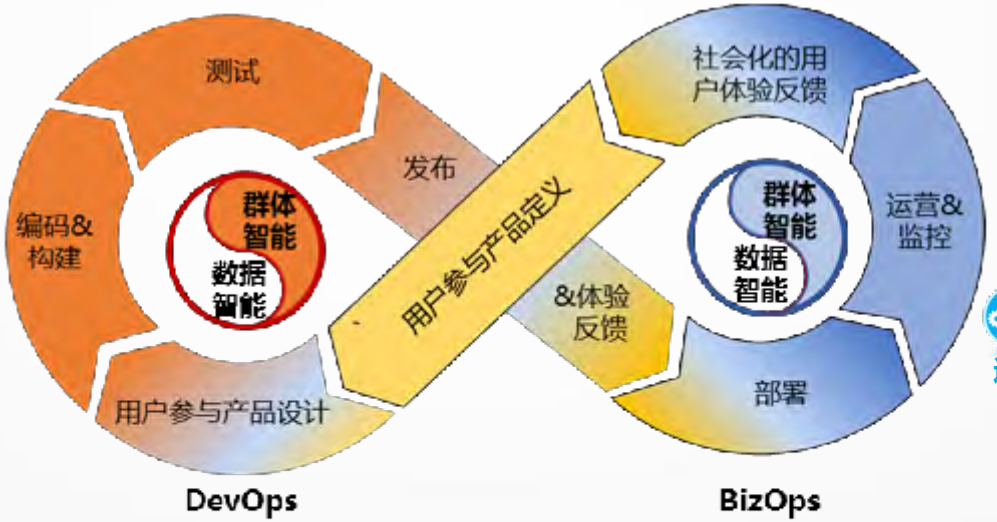
DevOps
敏捷开发过程
规范软件过程
定制化过程

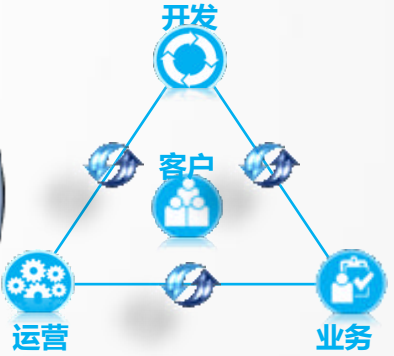
螺旋发展

工具

互联网+工具
互联网+云平台
(BlueMix & Predix
CloudFoundry、Azure)







微服务技术
物联网开发技术
移动开发技术
数据分析技术

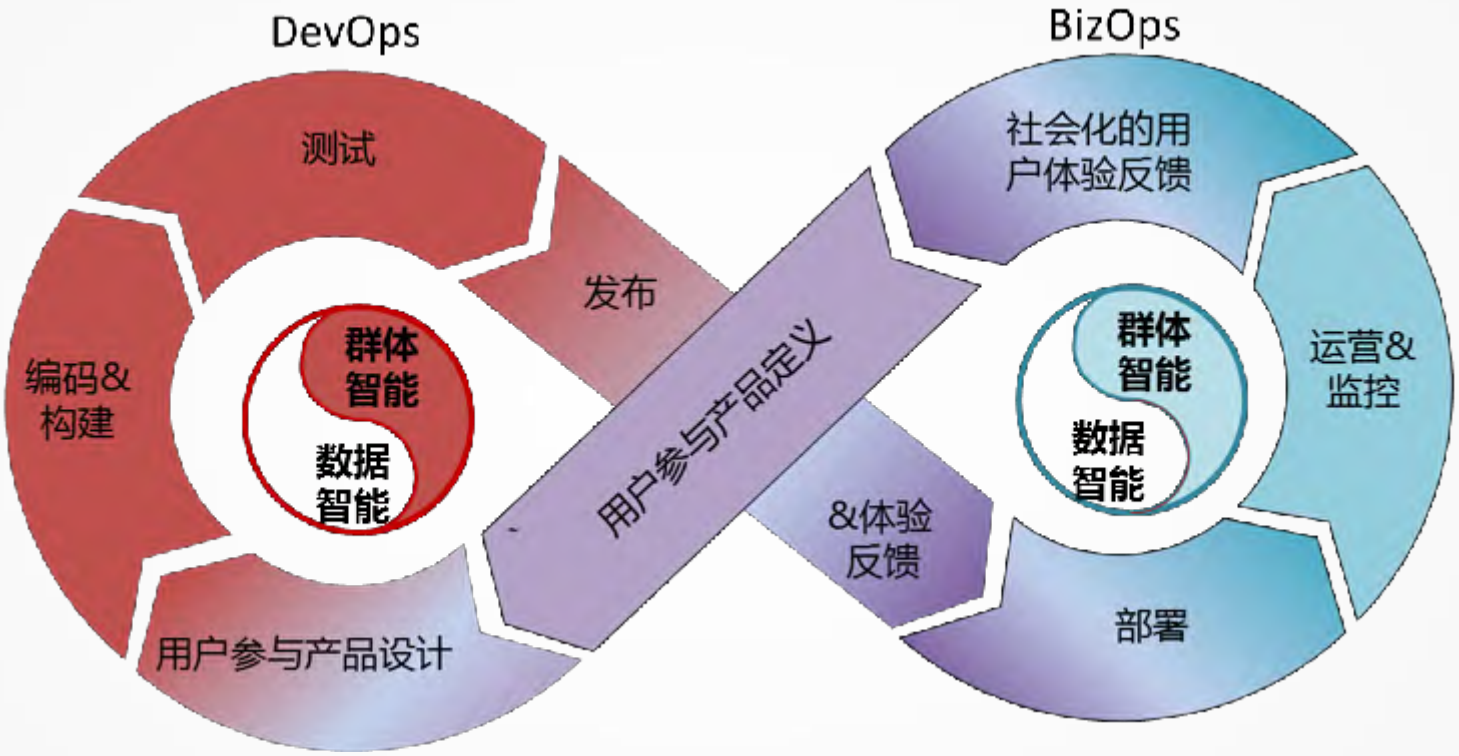
敏捷化 **服务化** **智能化** **生态化**

面向互联网的语言
和框架(HTML5,
Node.js,
Go,Spring)
面向大数据的语言
（Python & R）

技术

语言

NSE开发过程模型



两个主要创新点：

1. **NSE模型**中的面向协同创新的**群体智能**
2. **NSE模型**中的面向全局优化的**数据智能**

四大趋势：敏捷化、服务化、智能化、生态化

下一代软件工程方法：NSE

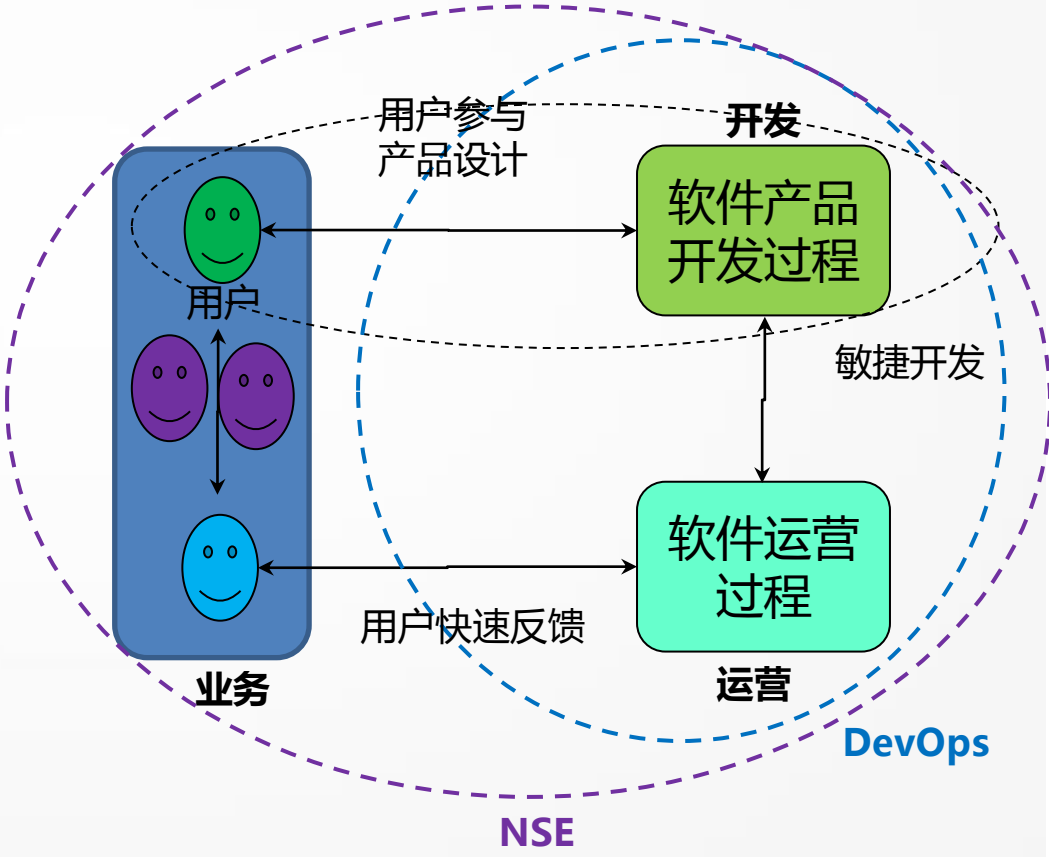
敏捷开发：更加关注个体与交互、可用的软件、用户参与和快速响应变化

DevOps: 通过打通开发运营,实现软件产品的快速发布和反馈。

NSE: 通过打通开发、运营、业务，围绕用户需求开展人机物协作创新。

智能化趋势：

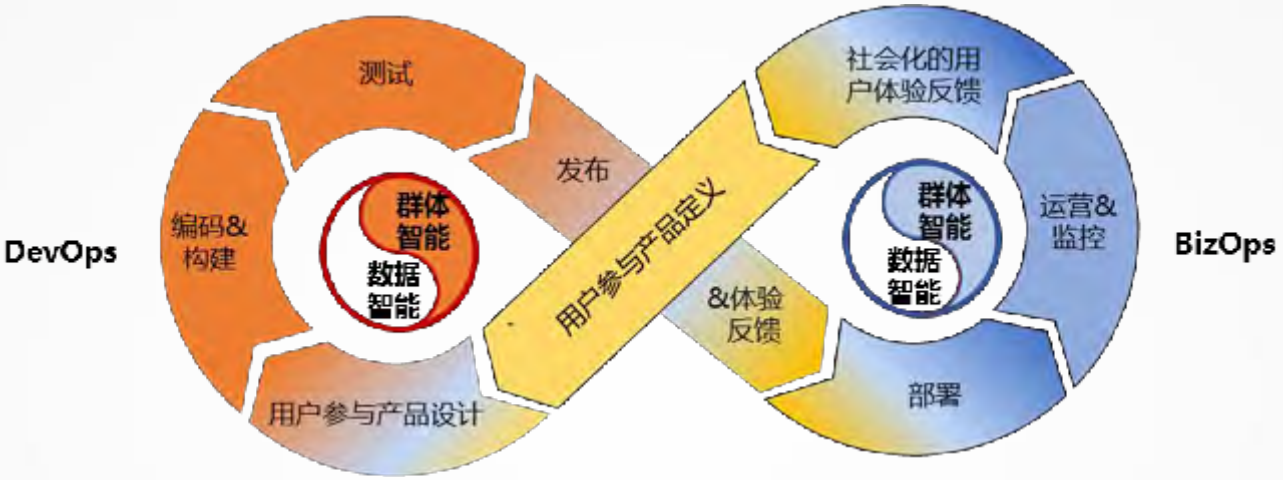
- **群体智能** – Crowd intelligence
- **数据智能** – Data Intelligence
- **人工智能** – Artificial Intelligence



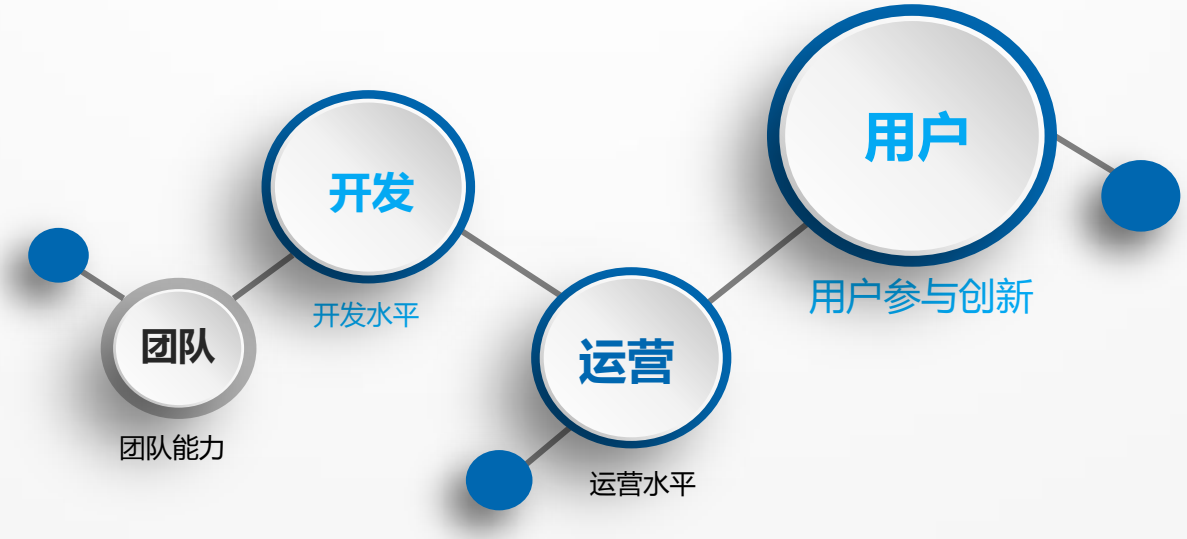
汇报提纲

- 下一代软件工程（NSE）
- 软件工程中的数据智能
- 未来的软件工程大脑

NSE中的数据智能研究



度量什么？项目、过程、产品？软件成功？



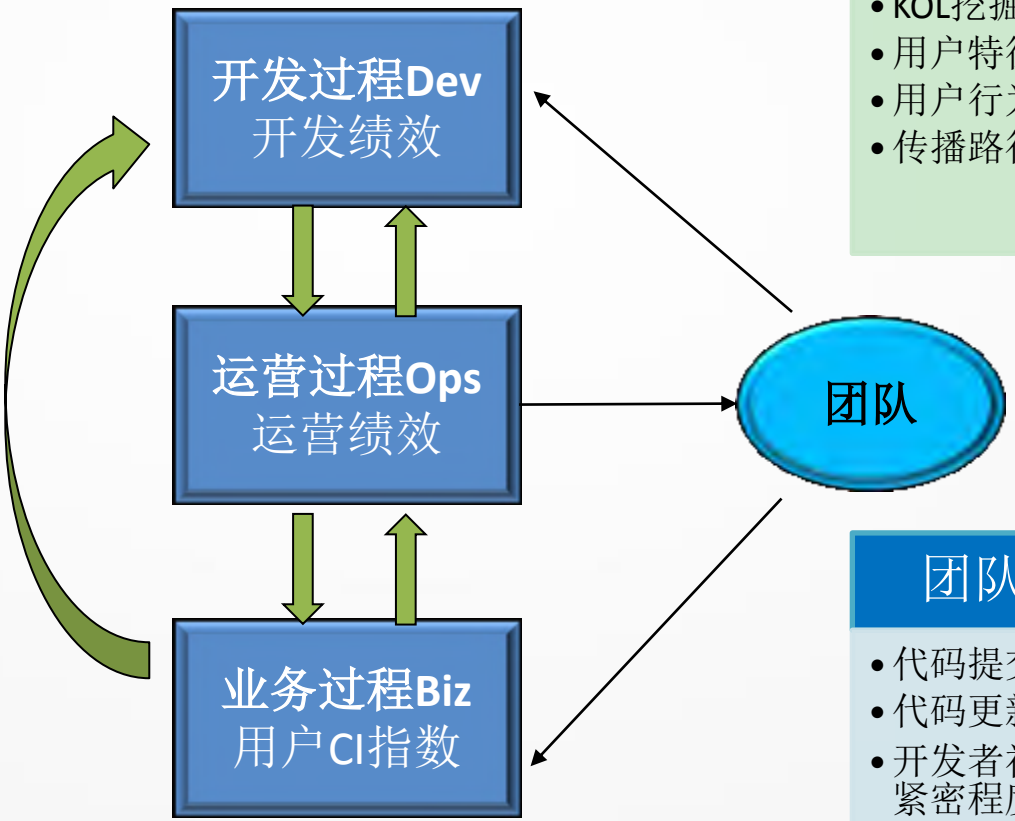
NSE中的数据智能研究

开发绩效

- 代码提交量(行数)
- 文档提交量
- 缺陷修复速率
- ……

运营绩效

- 用户增长速度
- KOL挖掘
- 用户特征挖掘
- 用户行为分析
- 传播路径分析



用户参与

- 软件用户数
- 软件收益
- 软件社区用户数
- 软件更新速度
- ……

团队活性

- 代码提交数
- 代码更新率
- 开发者社区协作紧密程度
- 团队自组织程度
- ……

NSE中的数据智能研究

	全 github 社区	followers 数前 1000 人		本次抓取的样本集	
		数值	占总体百分比	数值	占总体百分比
总项目数	82,000,000+	32,438	0.04%	270,287	0.33%
stars 数超过 1000	10,271	1,192	11.61%	8,763	**85.32%**
stars 数超过 500	20,800	2,027	9.75%	16,875	**81.13%**
stars 数超过 100	82,646	5,107	6.18%	49,355	**59.72%**

我们获得了27万条存储库信息，通过数据预处理，包括异常值处理，人工剔除异常记录、聚类等处理手段，最终保留下9万多条存储库记录。

NSE中的数据智能研究

- **主成分PCA 分析**

  确定开源软件的主要影响因素：用7个主成分代表原有的14个字段，概括原始变量所包含信息的82.74%

- **实际意义**

  结合软工实践经验，赋予各影响因素实际业务意义

指标定义	主成分构成	关键因子
开发绩效	第一主成分	codesize
		mlangsize
用户参与度	第二主成分	stars
		watchers
		folks
运营绩效	第三主成分	open-issue
		close-issue
		close-pull
	第七主成分	forks
		open-pull
团队活性	第四主成分	create-diff-record
		push-diff-record
	第五主成分	push-diff-record
		update-diff-record
	第六主成分	Contributors
		mainbranch_commits

NSE中的数据智能研究

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.294	23.526	23.526	3.294	23.526	23.526	2.257	16.124	16.124
2	2.493	17.809	41.335	2.493	17.809	41.335	2.123	15.167	31.290
3	1.725	12.323	53.658	1.725	12.323	53.658	2.044	14.601	45.892
4	1.195	8.535	62.193	1.195	8.535	62.193	1.306	9.330	55.221
5	1.048	7.487	69.680	1.048	7.487	69.680	1.294	9.245	64.466
6	1.009	7.204	76.885	1.009	7.204	76.885	1.284	9.170	73.636
7	.821	5.864	82.749	.821	5.864	82.749	1.276	9.113	82.749
8	.623	4.449	87.198						
9	.426	3.045	90.244						
10	.411	2.938	93.182						
11	.380	2.714	95.895						
12	.353	2.521	98.416						
13	.210	1.502	99.918						
14	.011	.082	100.000						

Extraction Method: Principal Component Analysis.

开发绩效：f_{dev}； 用户参与度：f_{par}； 运营绩效：f_{op}；团队活性：f_{vit}

$f_{op} = \frac{w3}{w3+w7} * X_3 + \frac{w7}{w3+w7} * X_7$, $X_1 \sim X_7$ 取PCA分析结果中主成分的特征值

W=贡献度/累计贡献度

NSE中的数据智能研究

f 软件成功度量

维度	指标定义	形式化表示	权重形式化表示	计算公式	主成分构成	形式化表示	权重表示
开发要素	开发绩效	f_{dev}	w_{dev}	$f_{dev} = x_1$	第一主成分	x_1	w_1
用户参与	流行度	f_{pop}	w_{pop}	$f_{pop} = x_2$	第二主成分	x_2	w_2
运营要素	运营绩效	f_{op}	w_{op}	$f_{op} = w3/(w3+w7)*x3 + w7/(w3+w7)*x7 = 0.6*x3 + 0.4*x7$	第三主成分	x_3	w_3
					第七主成分	x_7	w_7
团队活性	活跃度	f_{vit}	w_{vit}	$f_{vit}=w4/(w4+w5+w6)*x4 + w5/(w4+w5+w6)*x5 +w6/(w4+w5+w6)*x6=0.36*x4 + 0.36*x5 + 0.28*x6$	第四主成分	x_4	w_4
					第五主成分	x_5	w_5
					第六主成分	x_6	w_6

$$F_{success}=W_{dev} * f_{dev} + W_{pop} * f_{pop} + W_{op} * f_{op} + W_{vit} * f_{vit}$$

NSE中的数据智能研究

开源软件成功度量模型：

$$F_{success}=W_{dev} * f_{dev} + W_{par} * f_{\mathrm{par}} + W_{op} * f_{op} + W_{vit} * f_{vit}$$

验证：

1. f_{par} 和Github中用于标识项目流行度的Star数相关性很高。

2. 根据我们提出的模型，2016年官方出具的15个最受关注的项目有11个位于成功度排名前100的项目中，所有15个项目都落在了排名的前1%中。

汇报提纲

- 下一代软件工程（NSE）
- 软件工程中的数据智能
- 未来的软件工程大脑

智慧建筑是什么

智慧建筑

1

全面感知和永远在线的“生命体”

嵌入式传感器和各种智能感知设备
人工智能和各种创新技术的普遍使用

2

拥有大脑的自进化智慧平台

虚拟现实和增强现实将会成为人类和建筑交互的主要方式
人工智能会是提升用户体验和智慧化感知交互的重要手段

3

人机物深度融合的开放生态系统

人、机、物融合系统（HCPS）
人类的追求和用户体验

来源：阿里巴巴集团《智慧建筑白皮书》

软件工程大脑

软工大脑基本模式 ： = 物联 + 数据智能 + 类人服务

软件工程大脑

优化

智能咨询师：自适应的软件开发过程，实践、技术等按需推荐
基于数据闭环的自学习算法

预测

需求获取 – 用户行为分析
软件成功度
开发绩效、运营绩效

推荐

推荐HR – 用户画像、专家标签
推荐开源软件 – 类似需求或描述推荐架构 、推荐代码

分类 聚类

客户理解 – 用户画像、市场分析
软件理解 – 软件/项目画像、分类排名


THANKS
点燃成功
点燃梦想
谢谢