软件工程中的大规模数据

周明辉
北京大学
zhmh@pku.edu.cn
http://sei.pku.edu.cn/~zhmh

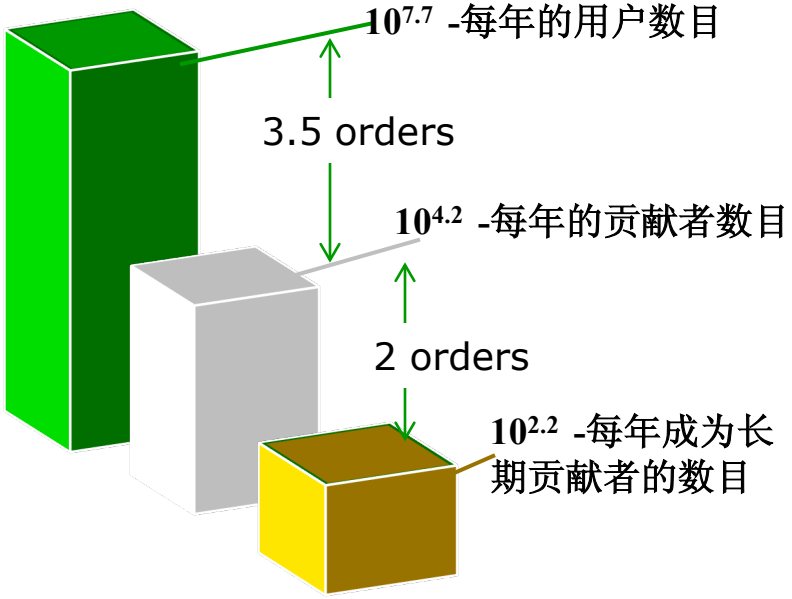
目录

- 研究中的一个例子
- 数据驱动的软件度量的挑战
- 关于大数据，我们可以做些什么？

开源社区怎么吸引长期贡献者？

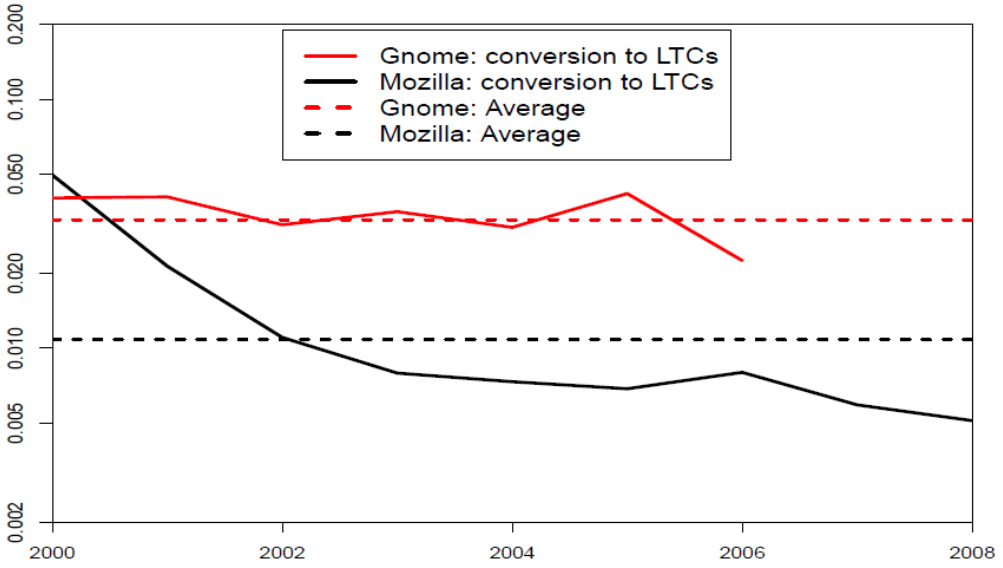
- 不同类型参与者的数目差距巨大，且比例日趋降低

Mozilla (Average over 2000-2008)



每年用户：千万级别
每年新的参与人数：万
每年成为长期贡献者的人数：百

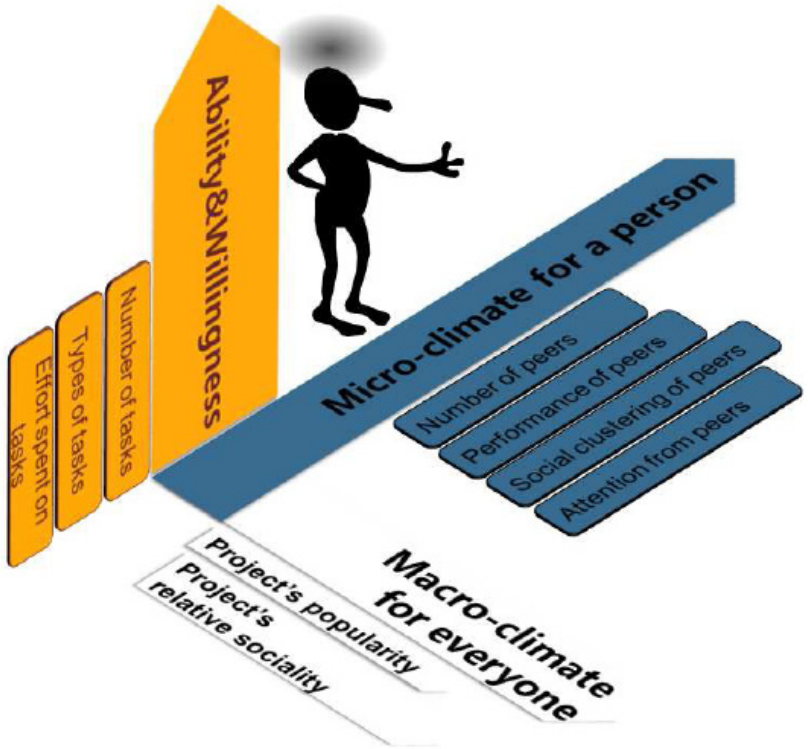
Gnome和Mozilla年贡献者与长期贡献者的比率



Project	Years	MLOC [7]	Domain	Cntrbtrs
Gnome	10	7.9	UI	156, 332
Evolution		.8	Calendar&Mailbox	21, 041
Nautilus		.1	File manager	17, 430
Epiphany		.1	Browser	3, 716
Mozilla	12	20	UI	187, 333
Firefox		5.3	Browser	47, 690
Thunderbird		1.1	Mailbox	12, 993
Calendar		.8	Calendar	4, 130

进入社区的第一个月发生了什么？

- **Issue Tracking System**
- **Gnome: 156,332个用户id，517,801个任务，6,398,475个动作**
- **Mozilla: 187,333个用户 id，620,511个任务，15,662,706个动作**



计算贡献者第一个月跟谁交互，他们的效率，交互的人数，

Intel 2GHz CPU x 32, 64G mem：4天

面向开源宇宙的计算：复用模式

- 数据量： 开源宇宙
  - 20万个工程
  - 6亿个文件
  - 24T原始代码数据和1T的版本信息数据（sourceforge 的10倍大小）
  - 70G文件名索引
  - 100个代码索引文件，总计1T大小，存储着压缩后的代码
  - 数据还在持续增长

- 计算量
  - Intel 2GHz CPU x 32, 64G mem, 30T disk server
  - 8进程提取压缩文件需200分钟
  - 20进程对每个文件和最可能相似的5个文件计算最长子串理论上需要3000小时

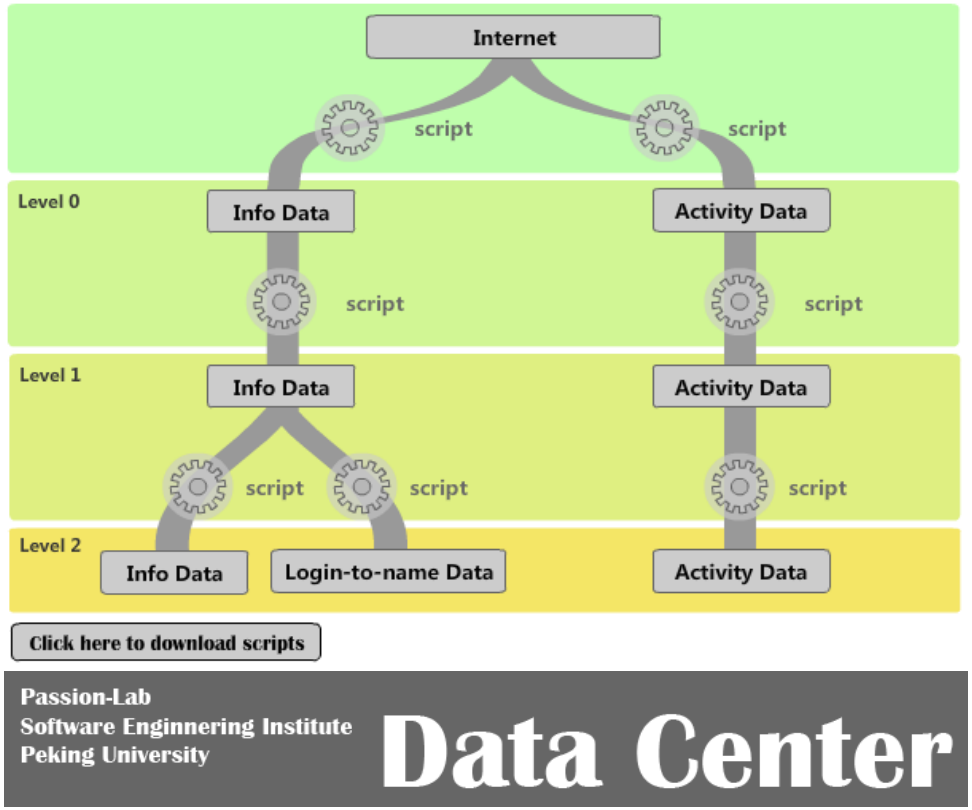
Forge	Type	VCSs	Files	File/Versions	Unique File/Versions	Disk Space
git.kernel.org	Git	595	12,974,502	97,585,997	856,920	205GB
SourceForge	CVS	121,389	26,095,113	81,239,047	39,550,624	820GB
netbeans	Mercurial	57	185,039	23,847,028	492,675	69GB
github.com	Git	29,015	5,694,237	18,986,007	7,076,410	154GB
...	...	...				

数据驱动的软件度量与分析 – 挑战

- 数据收集： 与系统管理员的斗智斗勇
- 数据规整化： 不同的系统，不同的格式
- 提出量度，观测现象： 项目差异性，个体差异性
- 联线分析：计算时空 -- 不要最优，但要高效
  - 横看成岭侧成峰

北大软工所的尝试

- 建立软件工程数据池
  - 收集和规整化数据，开源
  - http://passion-lab.org
- 可视化项目基本信息和基础量度
- 网构软件
  - 从软件的角度看互联网；系统化地研究开放、动态网络环境下的软件模型、开发方法和运行支撑平台



关于大数据，我们可以做的

- 建造一个数据共享平台

- 软件，开源，...(云计算)，大数据
  - 没数据？企业数据
  - 小规模数据？共享，开放数据访问、展示和分析
  - 实现从data.org到info.org 及knowledge.org的转变，最终指导决策

现代化的社会，它能够将整个的社会以数目字管理

--黄仁宇