R工程实践与 Data Scientist

阿稳@douban

2013.05.19 于第6届R语言会议

豆瓣douban

当R面临大数据量

Copy-on-modify semantics

- 案例

```
> address <- function(x) .Internal(inspect(x))
> x <- 1:10
> address(x)
@102b83e28 13 INTSXP g0c4 [NAM(2)] (len=10, tl=0) 1,2,3,4,5,...
> x[1] <- 2
> address(x)
@1029acff0 14 REALSXP g0c6 [NAM(2)] (len=10, tl=0) 2,2,3,4,5,...
>
```

- loop -> 向量化运算（apply, 矩阵运算）
- apply -> c/cpp

> The art of R performance improvement is to build up a good intuitions for what operations incur a copy, and what occurs in place.
>
> -- https://github.com/hadley/devtools/wiki/Profiling

稀疏矩阵M的每一列非零元素减去该列的均值

```
> M
2 x 4 sparse Matrix of class "dgCMatrix"

[1,] . . . 3
[2,] 1 2 . .
> str(M)
Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
  ..@ i       : int [1:3] 1 1 0
  ..@ p       : int [1:5] 0 1 2 2 3
  ..@ Dim     : int [1:2] 2 4
  ..@ Dimnames:List of 2
  .. ..$ : NULL
  .. ..$ : NULL
  ..@ x       : num [1:3] 1 2 3
  ..@ factors : list()
```

```
library(Matrix)
...
N = M
N@x = 1
S = N %*% Diagonal(colSums(M)/colSums(N))
M = M - S
```

cs = colMeans(M)

for (i in nrow(M))
    for (j in ncol(M))
        M[i, j]=M[i, j]-cs[j]

for (i in ncol(M))
    M[i, M[i,]!=0]=M[i, M[i,]!=0]-cs

R能不能做并行

- Rmpi, RHadoop
- snow, snowfall
- multicore
- foreach+iterator
- doParallel

Revolution Analytics benchmark

	Base R 2.14.2 64	Revolution R (1-core)	Revolution R (4-core)	Speedup (4 core)
Matrix Calculation	17.4 sec	2.9 sec	2.0 sec	7.9x
Matrix Functions	10.3 sec	2.0 sec	1.2 sec	7.8x
Program Control	2.7 sec	2.7 sec	2.7 sec	Not Appreciable

Speedup = Slower time / Faster Time - 1 Test descriptions available at http://r.research.att.com/benchmarks

	Base R 2.14.2 64	Revolution R (1-core)	Revolution R (4-core)	Speedup (4 core)
Matrix Multiply	124.4 sec	11.4 sec	4.4 sec	27.1x
Cholesky Factorization	18.0 sec	1.8 sec	.6 sec	29.8x
Singular Value Decomposition	37.8 sec	8.4 sec	4.6 sec	7.1x
Principal Components Analysis	141.2 sec	22.4 sec	11.0 sec	11.9x
Linear Discriminant Analysis	117.0 sec	39.8 sec	32.0 sec	2.7x

Speedup = Slower time / Faster Time - 1

他们选择的路径能说明一些问题：R本身的实现并不着重考虑性能，底层代码需要经过改造才能适应工程中大数据量的需求。而且是针对单机的实验。

最基本的问题在于

- 一切的并行都受限于单核时的速度
- 这门语言最初并不是设计用于工程用途的
- 社区的构成和关注点（基因）

```
At the heart of R is a tension between
interactive data analysis and programming.
                -- Hadley Wickham
```

可能是目前最合理的包搭配

- foreach: 任务分割方式定义
- iterator: 迭代器
- doParallel: R支持的并行后端

Rpark的探索

- 源自spark和dpark
- 弹性分布式数据集： rdd
- 惰性计算
- 一个示例

出路

- 定位：数据分析师而不是数据挖掘工程师的工具。（R不是万能的，做它适合做的事）
- 按需而行：如果不能提升整个框架，就提升我们所需要的功能。
- 需要与底层的接口：Rcpp值得使用

前景

- 越来越多有工程背景的人加入到这个社区，使得R的应用领域也在拓展。
- 从而改变这门原先由统计学家主导的语言。
- 2.13官方提供parallel包的支持
- Rcpp的发展
- O'Reilly 今年来出版越来越多R方面的书

Data Scientist

一个在实际工作中未被充分定义的角色

豆瓣douban

角色描述

- 管理数据：采集和整理数据，提供数据报表（常规）
- 回答问题：通过数学建模、数据分析，支持产品和运营决策（具体产品相关）
- 探索数据：从数据中发现问题、挖掘知识，影响长期的产品策略（探索工作）

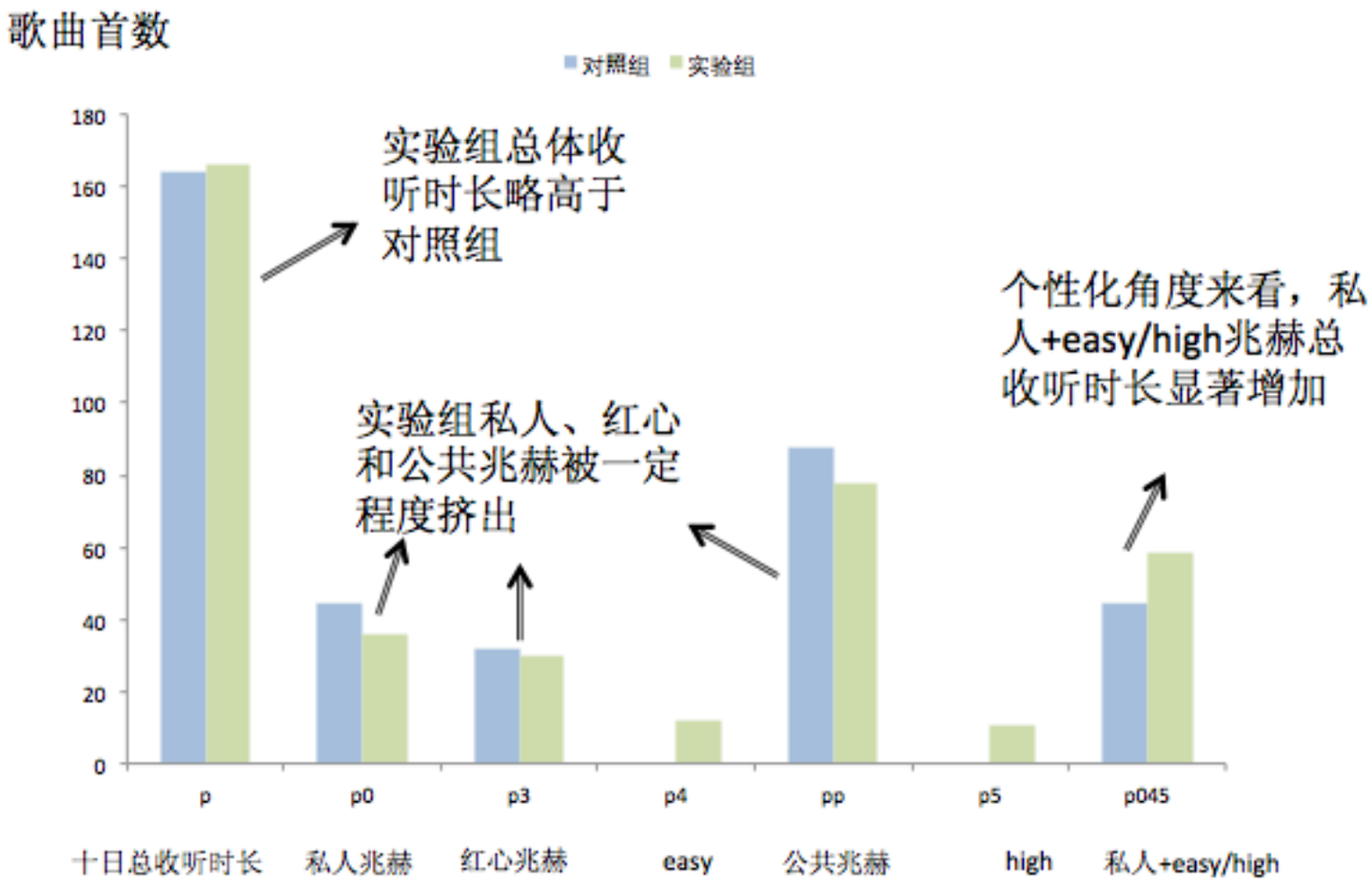
案例

- 豆瓣FM数据统计
- 豆瓣电影购票因素分析
- 小组分类研究

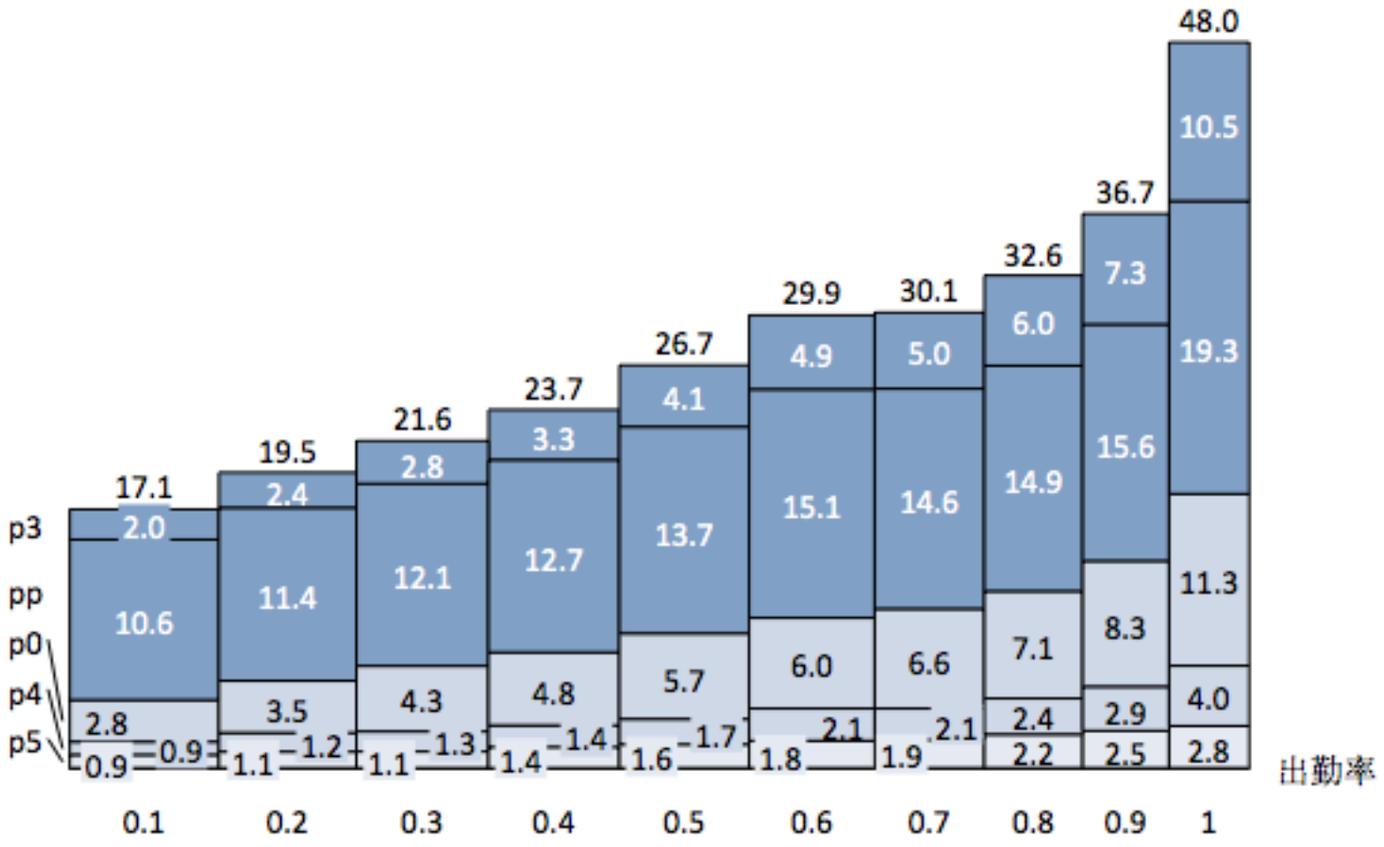
豆瓣FM数据统计

- 一段时间用户行为的统计描述
- 某项产品功能对用户行为的影响
- 对比预定义的不同用户群体行为

收听时长差别


歌曲首数
对照组
实验组
180
160
140
120
100
80
60
40
20
0
实验组总体收听时长略高于对照组
实验组私人、红心和公共兆赫被一定程度挤出
个性化角度来看，私人+easy/high兆赫总收听时长显著增加
p
p0
p3
p4
pp
p5
p045
十日总收听时长
私人兆赫
红心兆赫
easy
公共兆赫
high
私人+easy/high

不同群体的出勤率


p3
pp
p0
p4
p5
17.1
2.0
10.6
2.8
0.9
0.9
19.5
2.4
11.4
3.5
1.1
1.2
21.6
2.8
12.1
4.3
1.1
1.3
23.7
3.3
12.7
4.8
1.4
1.4
26.7
4.1
13.7
5.7
1.6
1.7
29.9
4.9
15.1
6.0
1.8
2.1
30.1
5.0
14.6
6.6
1.9
2.1
32.6
6.0
14.9
7.1
2.2
2.4
36.7
7.3
15.6
8.3
2.5
2.9
48.0
10.5
19.3
11.3
2.8
4.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
出勤率

电影购票因素分析

- 影响用户购票的各种因素及其因素中各种选择的效用值
- 为运营活动选片、选影院、定价等决策提供依据

问卷调查

CBC模型（Choice-Based Conjoint，基于选择的联合分析）

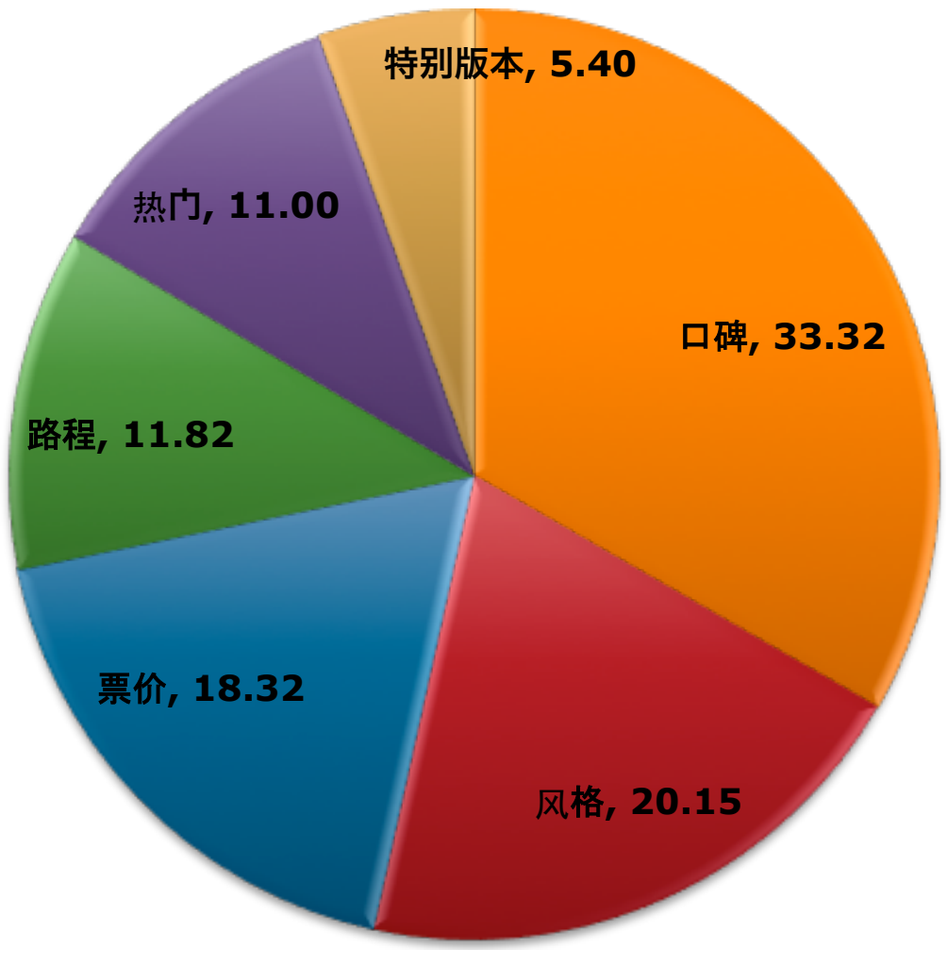
豆瓣电影

欢迎参加豆瓣电影调查

第2题 共8题 请从以下选项中选择您认为最好的一种观影方式。点击表格底部的按钮就可以自动换至下一题。

选择A	选择B	选择C
电影受关注程度一般 口碑很差 是一部恐怖电影 普通2D版本 花45分钟路程去影院 2张电影票共80元	电影很冷门 口碑不太清楚 是一部动作电影 IMAX版本 花90分钟路程去影院 2张电影票共40元	我都不喜欢
我喜欢这个电影	我喜欢这个电影	我都不喜欢

属性效用值

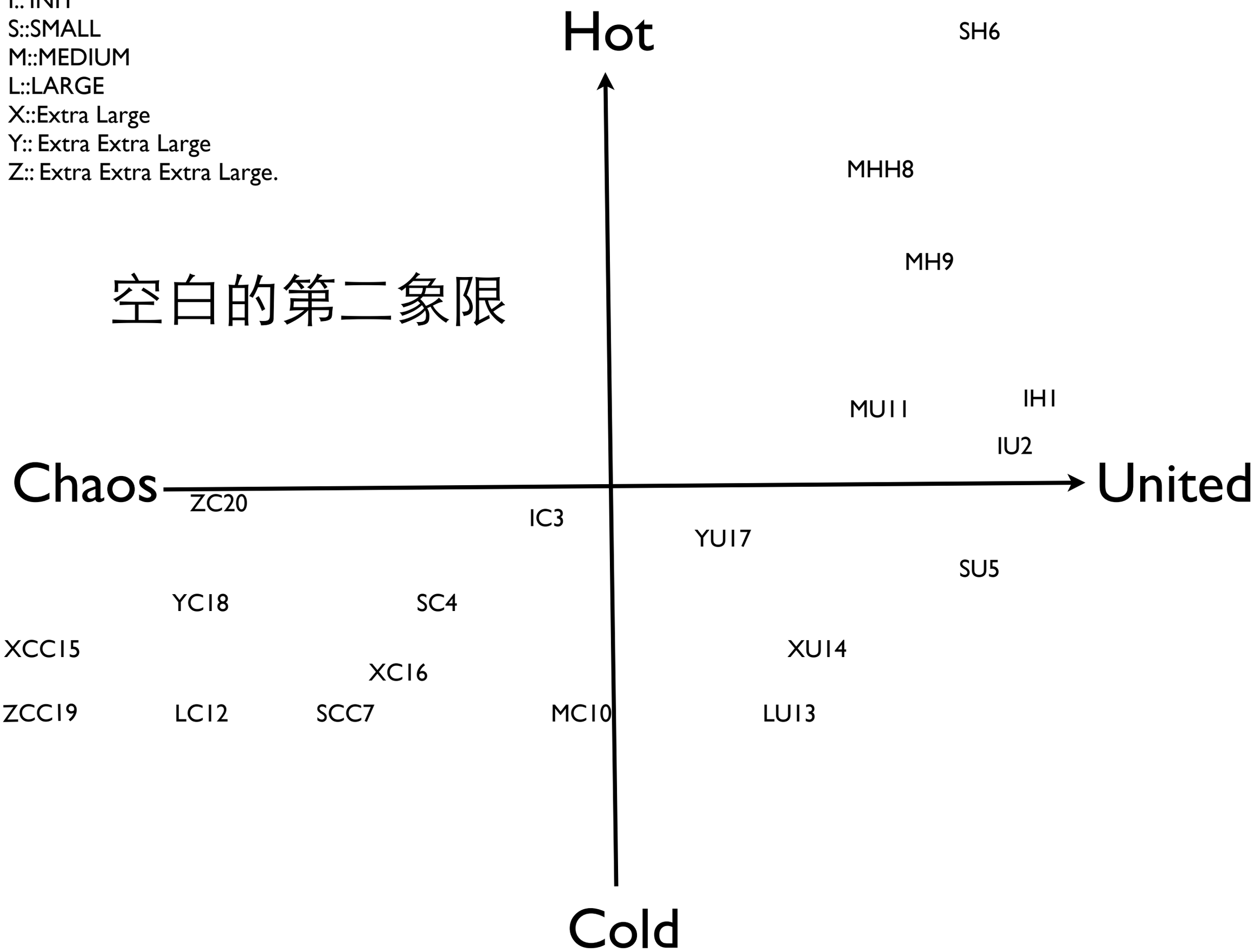


小组分类研究

- 小组事实是一个参差多态的生态系统
- 和其他系统一样需要明晰的分类学知识
  - 范例：林奈建立的动物分类学
- 基于分类的演化过程

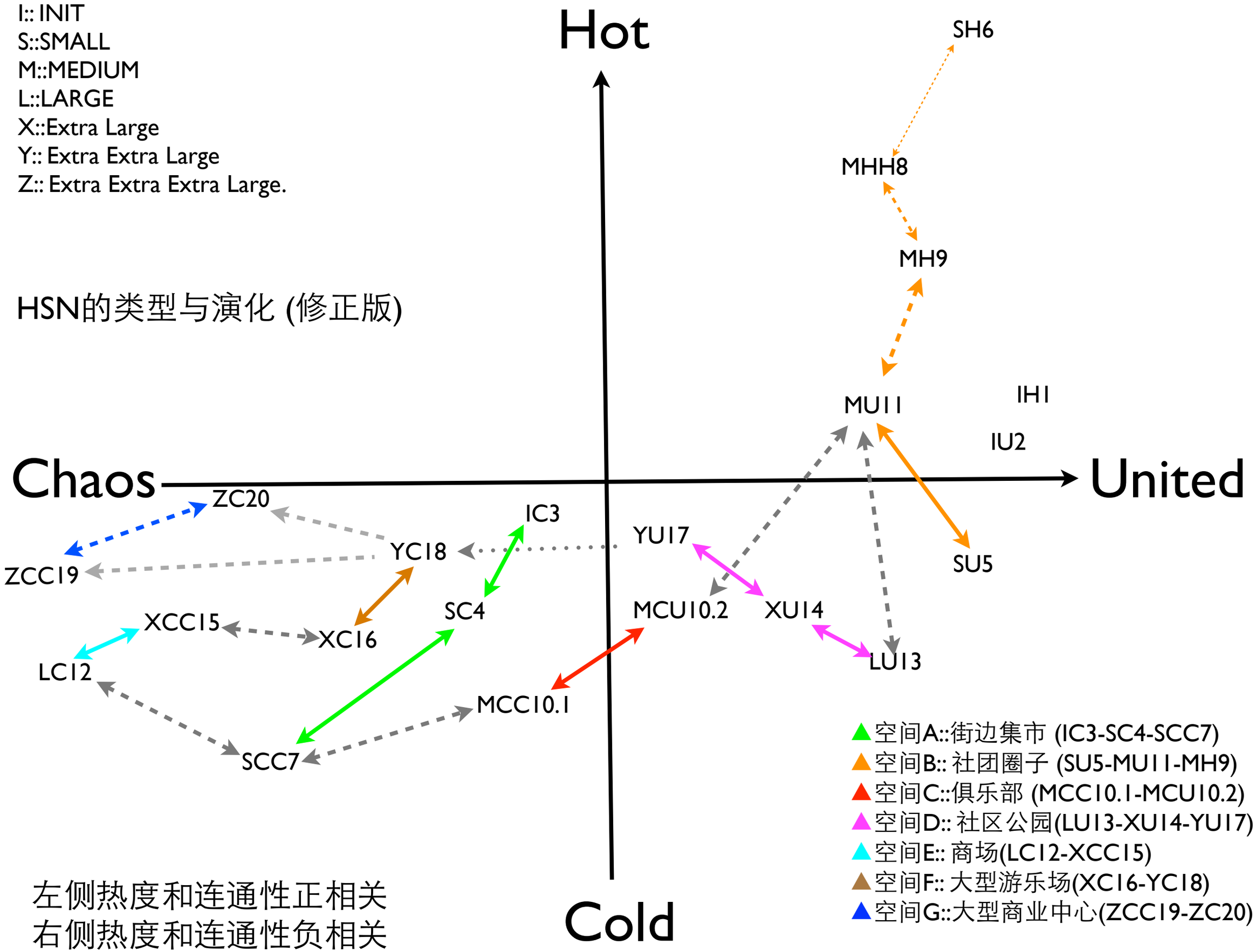
I:: INIT
S::SMALL
M::MEDIUM
L::LARGE
X::Extra Large
Y:: Extra Extra Large
Z:: Extra Extra Extra Large.

空白的第二象限



I:: INIT
S::SMALL
M::MEDIUM
L::LARGE
X::Extra Large
Y:: Extra Extra Large
Z:: Extra Extra Extra Large.

HSN的类型与演化 (修正版)



▲空间A::街边集市 (IC3-SC4-SCC7)
▲空间B:: 社团圈子 (SU5-MU11-MH9)
▲空间C::俱乐部 (MCC10.1-MCU10.2)
▲空间D:: 社区公园(LU13-XU14-YU17)
▲空间E:: 商场(LC12-XCC15)
▲空间F:: 大型游乐场(XC16-YC18)
▲空间G::大型商业中心(ZCC19-ZC20)

左侧热度和连通性正相关
右侧热度和连通性负相关

合格的人才符合职位，优秀的人才定义职位

有意于豆瓣Data Scientist职位者，请豆瓣上联系flycondor

豆瓣 douban