

数据科学与 R 语言

李舰



堡力山
Parra Mountain Intelligence Group

厦门大学数据科学与实验教学系列讲座

2015 年 06 月 27 日

目 录

1 数据科学简介

- 什么是数据科学
- 数据科学的误区

2 数据科学与 R

3 如何成为数据科学家

4 案例介绍

目 录

1 数据科学简介

- 什么是数据科学
- 数据科学的误区

2 数据科学与 R

3 如何成为数据科学家

4 案例介绍

什么是数据科学？

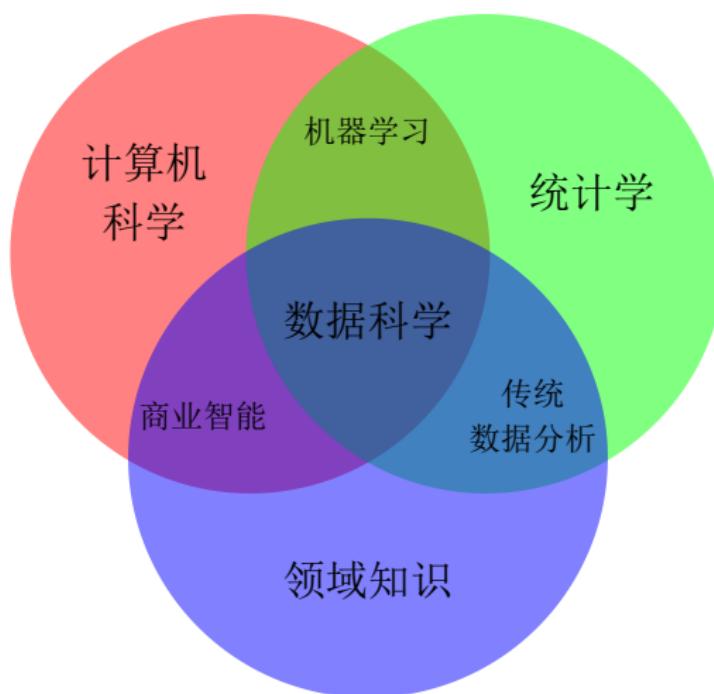
● 数据科学的来历

- Wikipedia 上目前最早考据到上个世纪 60 年代 Peter Naur 提出了这个概念
- 郁彬教授认为上个世纪 40 年代 Turner 和 Carver 等人就提出了数据科学的思想
- C.F. Jeff Wu 于 1997 年非常旗帜鲜明地提出了“Statistics = Data Science?”
- 从 2008 年 DJ Patil 和 Jeff Hammerbacher 把他们在 LinkedIn 和 Facebook 的工作职责定义为“数据科学家”的那段时期开始，数据科学开始在业界流行起来

● 定义

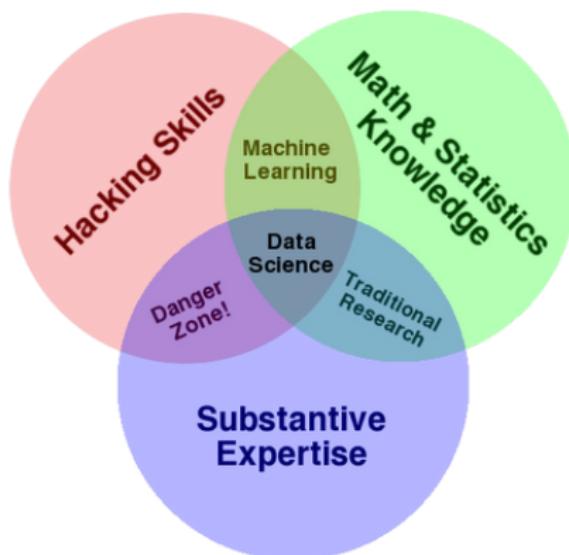
- 数据科学是使用科学方法从数据中获取知识的学科
- Wikipedia 上的定义：数据科学是一门利用数据学习知识的学科，其目标是通过从数据中提取出有价值的部分来生产数据产品

什么是数据科学？



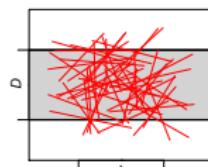
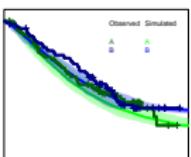
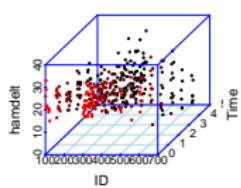
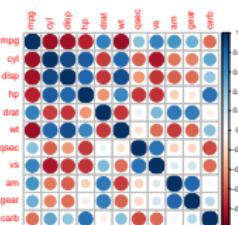
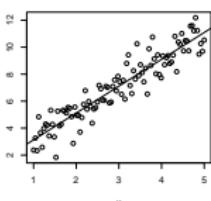
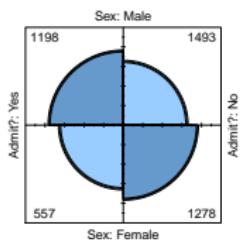
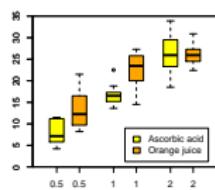
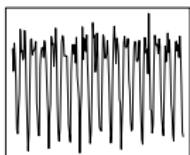
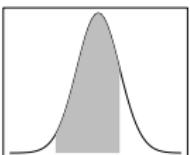
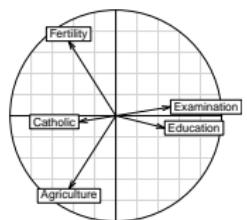
另一张流传很广的韦氏图

- 注意其中的 Danger Zone^a

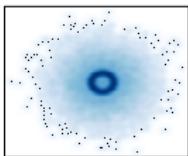
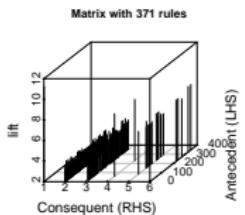
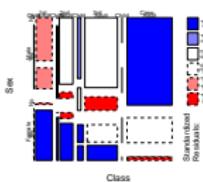
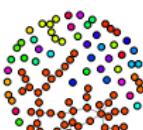
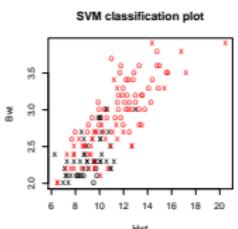
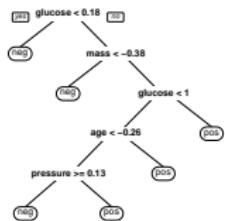
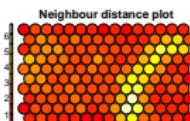
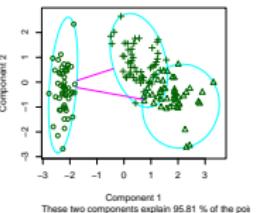
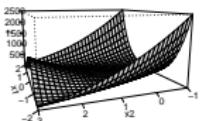
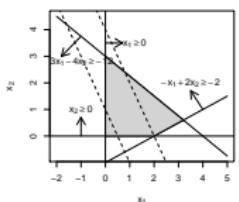


^a作者是 Drew Conway，最早发布于 <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

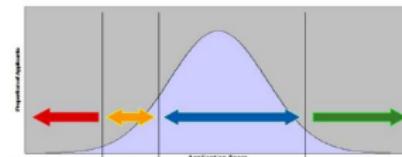
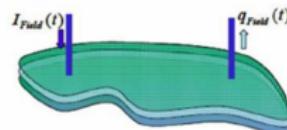
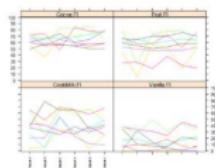
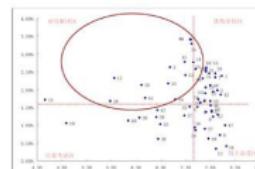
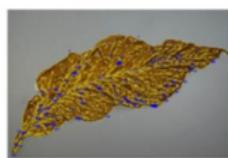
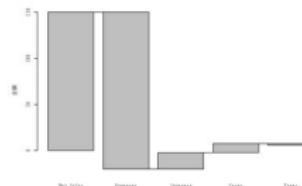
基于统计的分析



基于计算机的分析



基于领域的分析



数据科学与传统分析

● 传统分析是业界中数据应用的主流

- 传统分析主要处理结构化数据
- 传统分析最常用的方法是描述统计、SQL 语句、回归、聚类、相关分析等
- 传统分析容易被工具所限

● 数据科学可以满足深层次的分析需求

- 数据科学分析结构化数据和非结构化数据
- 数据科学完全以数据和需求为导向，不受制于任何技术、工具和方法

数据科学与大数据

● 大数据侧重于解决方案

- 业界流行的数据挖掘、商业智能、云计算等概念的目标都是提供平台和解决方案
- 大数据摆脱了对具体产品和厂商的依赖，通过开源软件和低价的硬件可以搭建顶级的平台
- 大数据的应用中仍然倾向于通过标准化的解决方案解决大量的实际问题

● 数据科学侧重于人的能力

- 数据科学的核心是数据科学家
- 数据科学并不提供标准的解决方案，也不依赖于任何工具
- 数据科学强调综合运用各种分析方法和工具来解决实际的数据问题的能力

目 录

1 数据科学简介

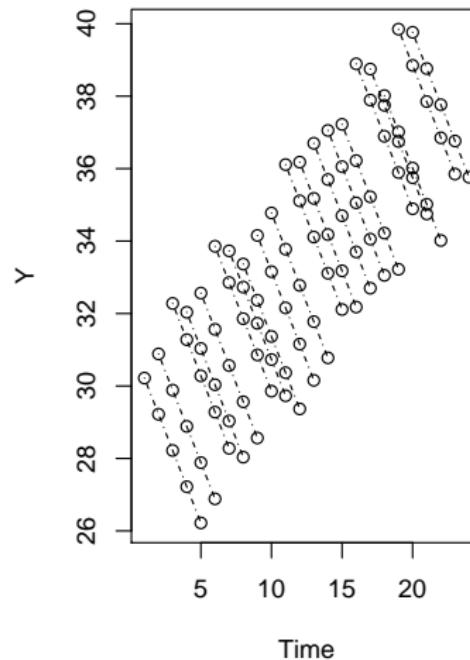
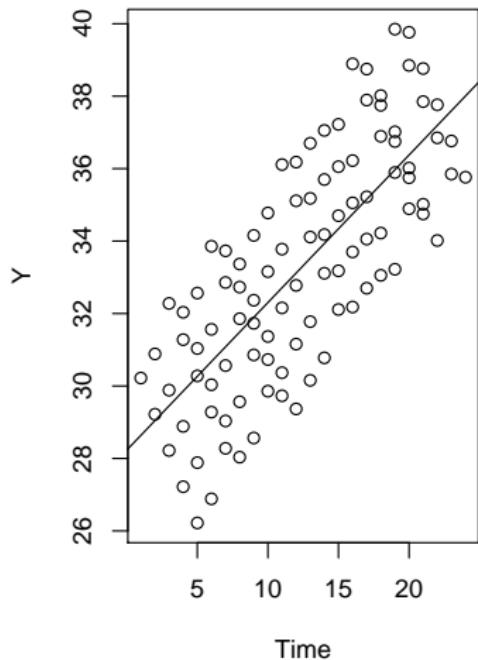
- 什么是数据科学
- 数据科学的误区

2 数据科学与 R

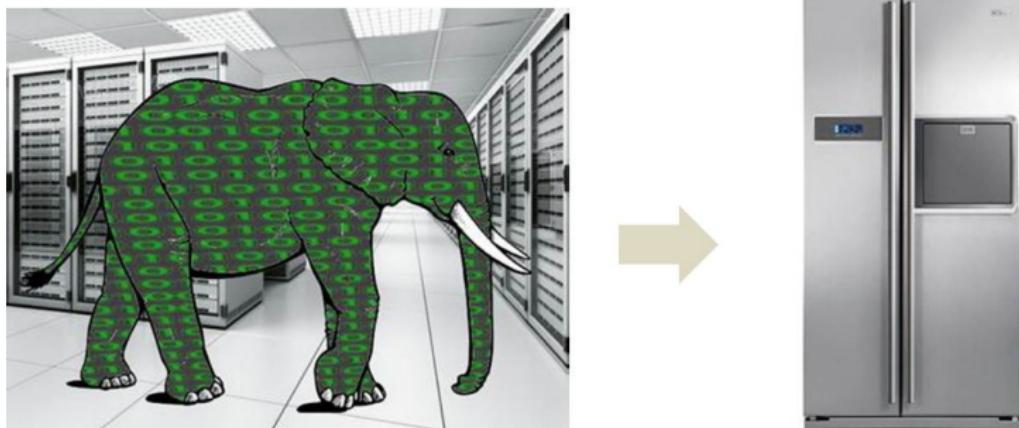
3 如何成为数据科学家

4 案例介绍

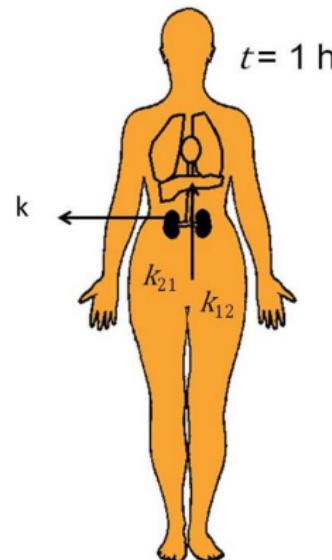
只会套用方法而不理解数据的不是数据科学



寄希望于“黑箱”工具的数据科学



无视数据而信仰“科学”的不是数据科学



目 录

1 数据科学简介

2 数据科学与 R

- 什么是 R
- 数据科学中的 R

3 如何成为数据科学家

4 案例介绍

目 录

1 数据科学简介

2 数据科学与 R

- 什么是 R
- 数据科学中的 R

3 如何成为数据科学家

4 案例介绍

什么是 R?

● R 语言

- 国人习惯于把能编程的东西成为语言
- R 最大的特色是 R 语言
- 通常用 R 语言来指代 R

● R 软件 / R 軟體

- 从应用的角度来说 R 是一个软件
- 不同的程序包 (package, 也称套件) 可以解决不同的问题

● R 环境

- R 的官方定义是一个统计计算和绘图的环境
- R 既是语言、也是软件、也是开发和应用环境

R 的历史 (I)

● S 语言是 R 语言的前身

- S 语言诞生于贝尔实验室统计研究部。1976 年 Chambers 和他的同事用 Fortran 实现了初步想法，称为“S1”。
- 到 1992 年的时候，已经发展到了“S3”。
- 1993 年，S 语言的许可证被 MathSoft 公司买断，S-PLUS 成为其公司的主打数据分析产品，1995 年发展到了“S4”。
- 1998 年美国计算机学会（ACM）授予了 S 语言的主要设计者 Chambers “软件系统奖”。
- 2008 年，TIBCO 收购了已改名成 Insightful 的原 MathSoft 公司，目前的 S-PLUS 已经纳入了 Spotfire 平台。

● R 语言吸收了很多 Scheme 语言的特性

- Scheme 语言 1975 年诞生于 MIT，是 LISP 的一个方言。
- 很久以前，有一次 R 语言的作者 Ross 准备用 Scheme 向别人演示词法作用域的时候，由于手边没有 Scheme，就用 S 来演示却失败了，这让他萌生了改进 S 语言的想法。

R 的历史 (II)

● R 语言诞生新西兰

- Ross Ihaka 和 Robert Gentleman 在奥克兰大学成为同事，他们最初希望在 Mac 环境下开发一个统计计算软件，于是模仿 Scheme，使用 C 开发了一个解释器，并采用 S 的语法。
- 1993 年，Ross 和 Robert 将 R 的部分二进制文件放到了卡耐基·梅隆大学统计系的 Statlib 中，并在 S 语言的新闻列表上发布了一个公告。
- 1995 年 6 月，在很多人的建议下，R 终于在 GPL 协议下作为开源软件发布了。

● 1997 年，R 核心团队成立

- 1997 年第一批核心团队的成员数目为 11 位；
- 2008 年 R 核心团队成员数目增加到了 19 位；
- 2011 年至今，R 核心团队成员数目达到 20 位。

R 的特点

- John M. Chambers 在 2009 年第一期《R Journal》上对 R 的定义：
 - an interface to computational procedures of many kinds (各类计算过程的接口);
 - interactive, hands-on in real time (具有可交互性，可以实时手动操作);
 - functional in its model of programming (函数式编程模式);
 - object-oriented, “everything is an object” (面向对象，“所有东西都是对象”);
 - modular, built from standardized pieces (模块化，由标准化块构建);
 - collaborative, a world-wide, open-source effort (协作性，全球范围的开源力量)。

R 的优势

● 灵活的语言

- 为数据而生的程序设计语言
- 一个设计理念：人的时间永远比机器的时间宝贵

● 混搭的平台

- 基于 S/R 语言进行数据操作、建模和绘图
- 类似 Scheme 的词法作用域和内存管理机制
- 调用 C 或 Fortran 进行底层运算
- 早期版本大量使用 Perl 进行系统的交互
- 在业界中常作为运算和绘图引擎嵌入到 JAVA 系统中

● 强大的社区

- CRAN 上已包含超过 6000 个第三方包
- R 社区最大的特点是来自具体领域的数据科学家数目众多
- R 的使用者可以分为“用户”和“开发者”，这在编程语言中是非常少见的

R 的性能问题

● 固有的缺陷

- 无法多线程

● 权衡的牺牲

- 解释型语言与交互式环境，可以使用 compiler 包来弥补
- 免费工具与代数运算库，可以换用商业 BLAS/LAPACK
- 统计建模与计算机编程，可以使用 C 或者 Fortran 开发对性能要求高的函数（注意，要多使用内置函数）

● 并非特有的缺点

- “基于内存的计算是个缺陷”。除了 SAS 等少数分析工具可以隐式执行内存外分析之外，数据量大都会撑爆内存。只是其他语言的用户不大容易遇到大数据分析的问题。解决办法都是内存外运算或并行，可以参考 R 中的 bigmemory 和 parallel 等包
- “难以处理大数据”。数据大到一定程度之后就不是编程语言的问题了，业界通常是借助于数据库或者并行系统来解决，可以参考 R 中的 ORE 或者 Rmpi、RHadoop 等包

目 录

1 数据科学简介

2 数据科学与 R

- 什么是 R
- 数据科学中的 R

3 如何成为数据科学家

4 案例介绍

哪些公司在使用 R?

Google™



GlaxoSmithKline

KPMG

KRAFT

syngenta

LLOYD'S



NOVARTIS



SONY



National Nuclear Laboratory

DIAGEO



Allianz

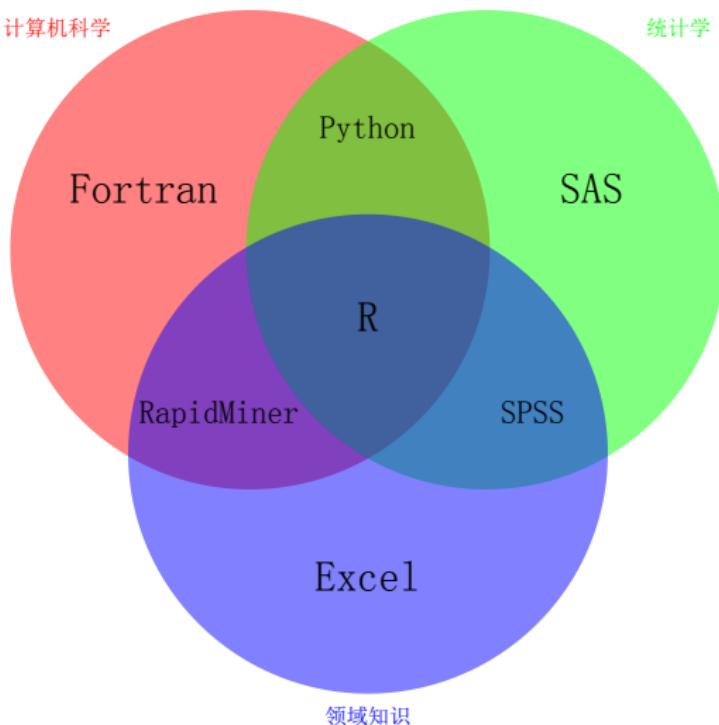


Henderson
Global Investors

KDNuggets 的调查

- 2014 年你在数据分析/数据挖掘/数据科学工作中使用过的编程语言或者统计语言有哪些?
 - 该项调查于 2014 年 8 月进行，共有 719 人参与了投票
 - R 语言得票率 49%，排名第一
 - SAS 排名第二；Python 排名第三；SQL 排名第四；Java 排名第五
- 在过去的一年里你在实际项目中用到的数据分析/数据挖掘/数据科学软件或工具有哪些?
 - 该项调查于 2014 年 5 月进行，共有 3285 人参与了投票
 - R 语言得票率 38.5%，排名第二
 - RapidMiner 得票率为 44.2%，排名第一
 - Excel 排名第三；SQL 排名第四；Python 排名第五

数据科学中各部分的常用工具



R 与工程开发



目 录

1 数据科学简介

2 数据科学与 R

3 如何成为数据科学家

- 什么是数据科学家
- 数据科学家的必备技能

4 案例介绍

目 录

1 数据科学简介

2 数据科学与 R

3 如何成为数据科学家

- 什么是数据科学家
- 数据科学家的必备技能

4 案例介绍

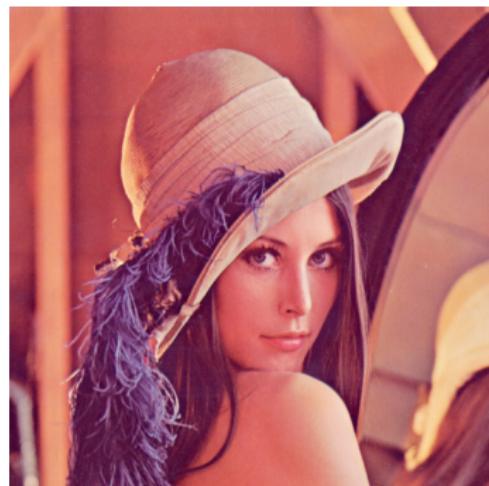
数据科学家是 21 世纪最性感的职业

- Data Scientist: The Sexiest Job of the 21st Century^a



^aHarvard Business Review

谁更性感？



谁更性感？

```
[,1]      [,2]      [,3]      [,4]  
[1,] 0.8862745 0.8862745 0.8862745 0.8862745  
[2,] 0.8862745 0.8862745 0.8862745 0.8862745  
[3,] 0.8745098 0.8745098 0.8745098 0.8745098  
[4,] 0.8745098 0.8745098 0.8745098 0.8745098  
[5,] 0.8862745 0.8862745 0.8862745 0.8862745  
[6,] 0.8862745 0.8862745 0.8862745 0.8862745  
[7,] 0.8941176 0.8941176 0.8941176 0.8941176  
[8,] 0.8901961 0.8901961 0.8901961 0.8901961  
[9,] 0.8901961 0.8901961 0.8901961 0.8901961  
[10,] 0.8823529 0.8823529 0.8823529 0.8823529  
[11,] 0.8941176 0.8941176 0.8941176 0.8941176  
[12,] 0.8823529 0.8823529 0.8823529 0.8823529  
[13,] 0.8745098 0.8745098 0.8745098 0.8745098  
[14,] 0.8862745 0.8862745 0.8862745 0.8862745  
[15,] 0.8745098 0.8745098 0.8745098 0.8745098  
[16,] 0.8666667 0.8666667 0.8666667 0.8666667  
[17,] 0.8666667 0.8666667 0.8666667 0.8666667  
[18,] 0.8666667 0.8666667 0.8666667 0.8666667  
[19,] 0.8705882 0.8705882 0.8705882 0.8705882  
[20,] 0.8705882 0.8705882 0.8705882 0.8705882
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
[1,]	5.1	3.5	1.4	0.2
[2,]	4.9	3.0	1.4	0.2
[3,]	4.7	3.2	1.3	0.2
[4,]	4.6	3.1	1.5	0.2
[5,]	5.0	3.6	1.4	0.2
[6,]	5.4	3.9	1.7	0.4
[7,]	4.6	3.4	1.4	0.3
[8,]	5.0	3.4	1.5	0.2
[9,]	4.4	2.9	1.4	0.2
[10,]	4.9	3.1	1.5	0.1
[11,]	5.4	3.7	1.5	0.2
[12,]	4.8	3.4	1.6	0.2
[13,]	4.8	3.0	1.4	0.1
[14,]	4.3	3.0	1.1	0.1
[15,]	5.8	4.0	1.2	0.2
[16,]	5.7	4.4	1.5	0.4
[17,]	5.4	3.9	1.3	0.4
[18,]	5.1	3.5	1.4	0.3
[19,]	5.7	3.8	1.7	0.3
[20,]	5.1	3.8	1.5	0.3

数据科学家是如何性感地处理图像的？

- 亮度、对比度、伽玛系数



Figure: lena, lena1, lena2, lena3

- 图像处理与矩阵运算

```
lena1 <- lena + 0.5  
lena2 <- 3 * lena  
lena3 <- (0.2 + lena)^3
```

目 录

1 数据科学简介

2 数据科学与 R

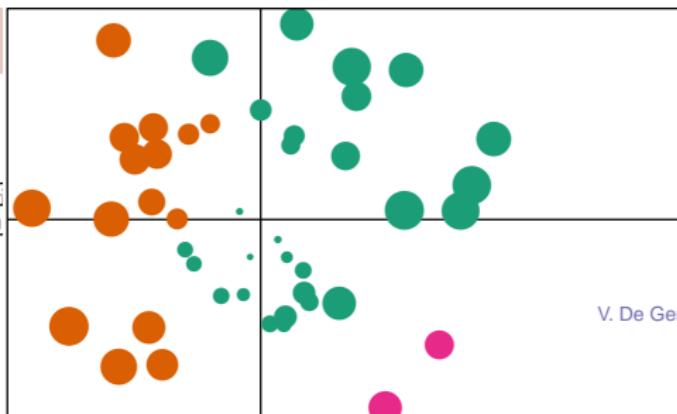
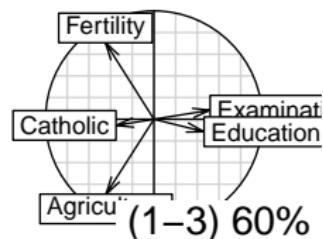
3 如何成为数据科学家

- 什么是数据科学家
- 数据科学家的必备技能

4 案例介绍

统计学方法

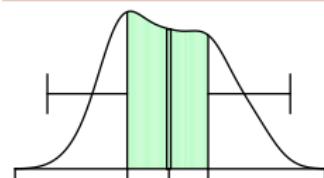
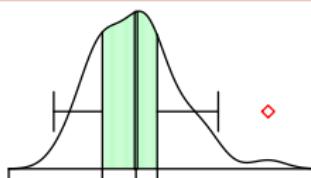
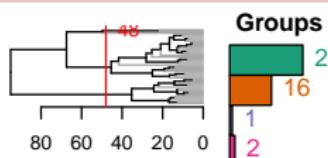
PCA 5 vars
`princomp(x = data, cor = cor)`



Clustering 4 groups

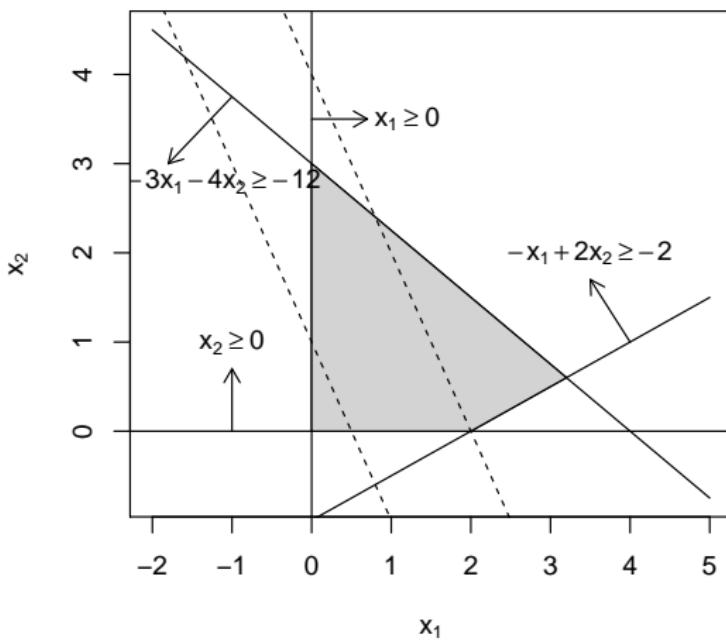
Factor 1 [41%]

Factor 3 [19%]



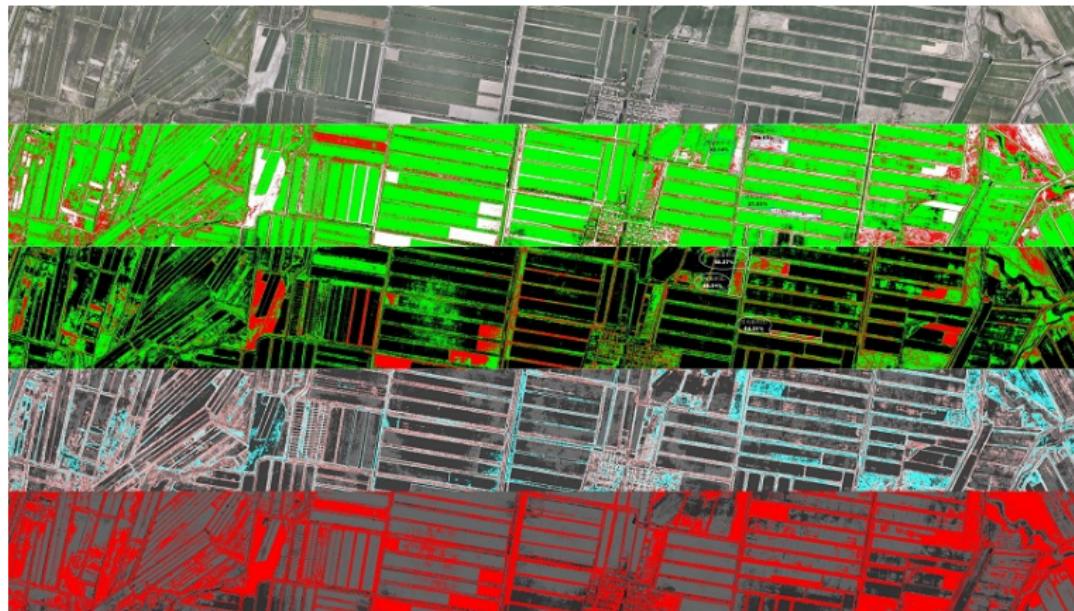
机器学习方法

最优化方法



蒙特卡罗方法

图像数据分析



空间数据分析

points



lines



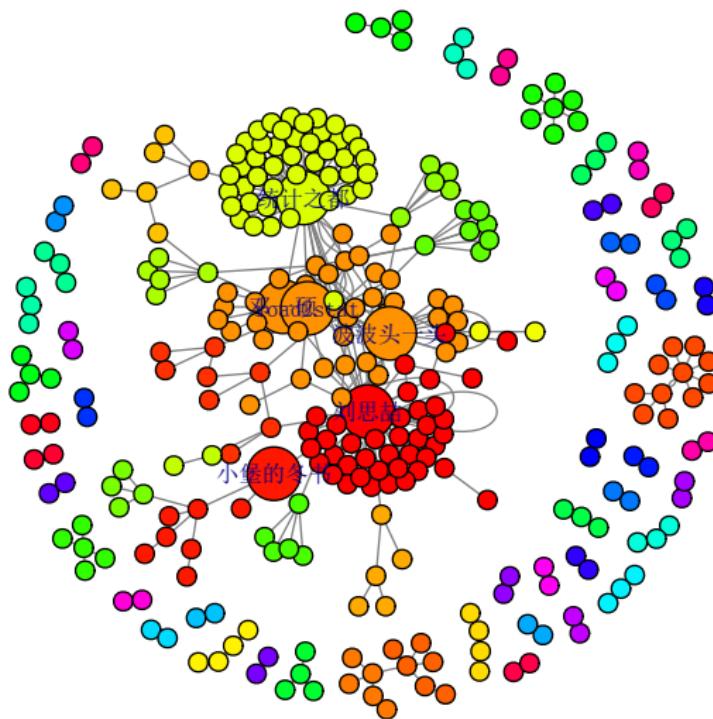
polygons



grid



社交网络分析



目 录

1 数据科学简介

2 数据科学与 R

3 如何成为数据科学家

4 案例介绍

- 制药
- 食品
- 零售和快消
- 其他

目 录

1 数据科学简介

2 数据科学与 R

3 如何成为数据科学家

4 案例介绍

- 制药

- 食品

- 零售和快消

- 其他

制药业概况

● 什么是药？

- 通常是指西药，主要是基于化合物的药
- 药物的研发阶段主要研究药物作用到人体后的各种反应
- 中药是很复杂的混合物，研究的难度比较高

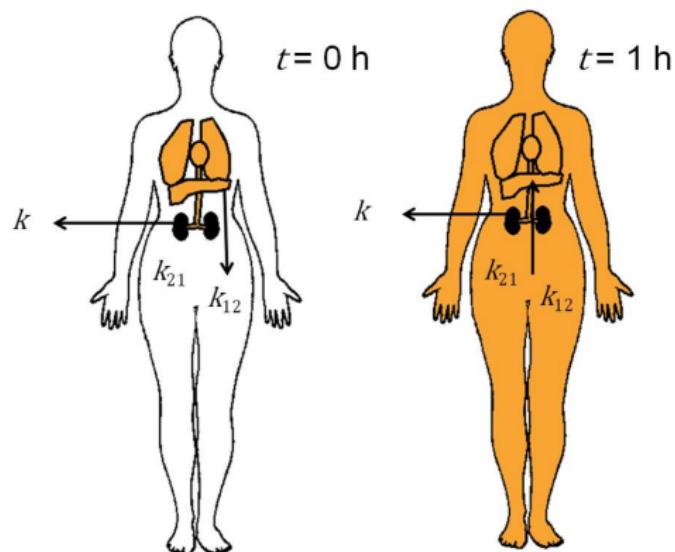
● 药的市场

- 整个制药行业，每年的销售额大约为 6 万亿元
- 每年新药研发花费的成本约为 1 万亿元
- 每款能成功面市的新药的平均研发时间是 12 年
- 平均每款药物的研发成本约为 50 亿元
- 实验室中筛选的化合物只有大约 $1/1000$ 能够进入到人体试验阶段

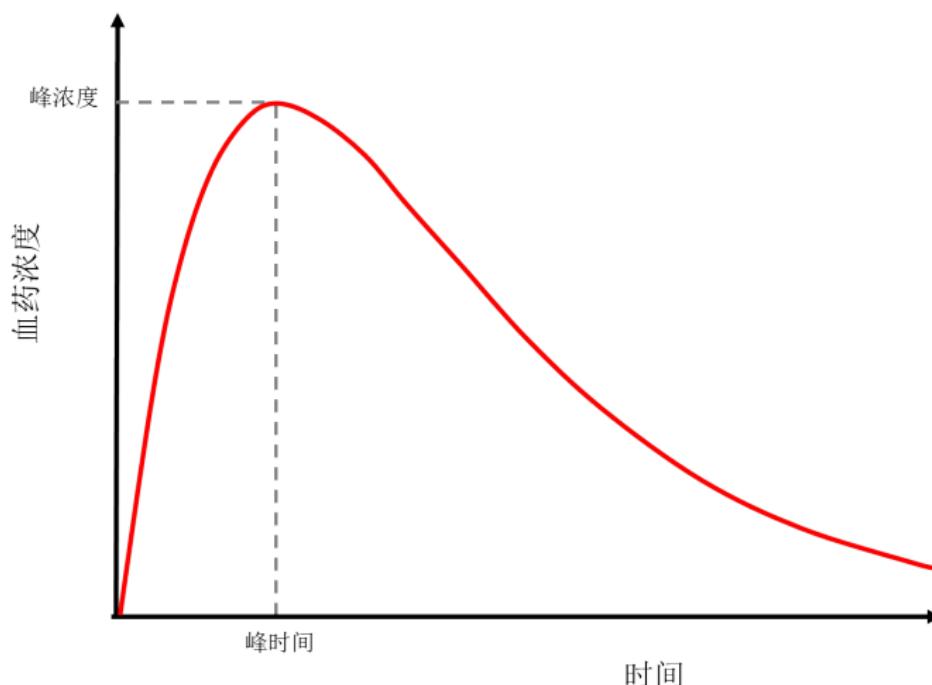
临床试验各阶段

	年数	试验对象	目的	% 成本
Preclinical	3.5	实验室研究和动物实验	研究生理学的反应和安全性	35
Phase I	1	20-80 健康的志愿者	确定药物安全和决定剂量	15
Phase II	2	100-300 病人	评估有效性和发现副作用	40
Phase III	3	1000+ 病人	验证有效性和长期的不良反应	10
FDA	2.5		审核和批准	
Phase IV		所有群体	FDA 要求的面市后的测试	

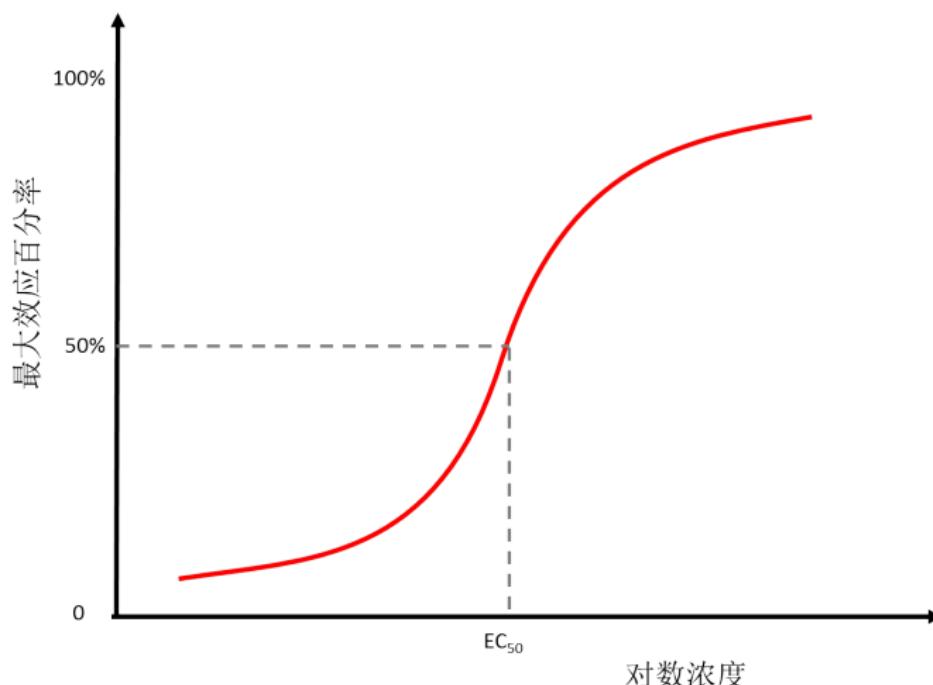
新药研发中的 PK/PD



药动学 (Pharmacokinetics, 简称 PK) 模型

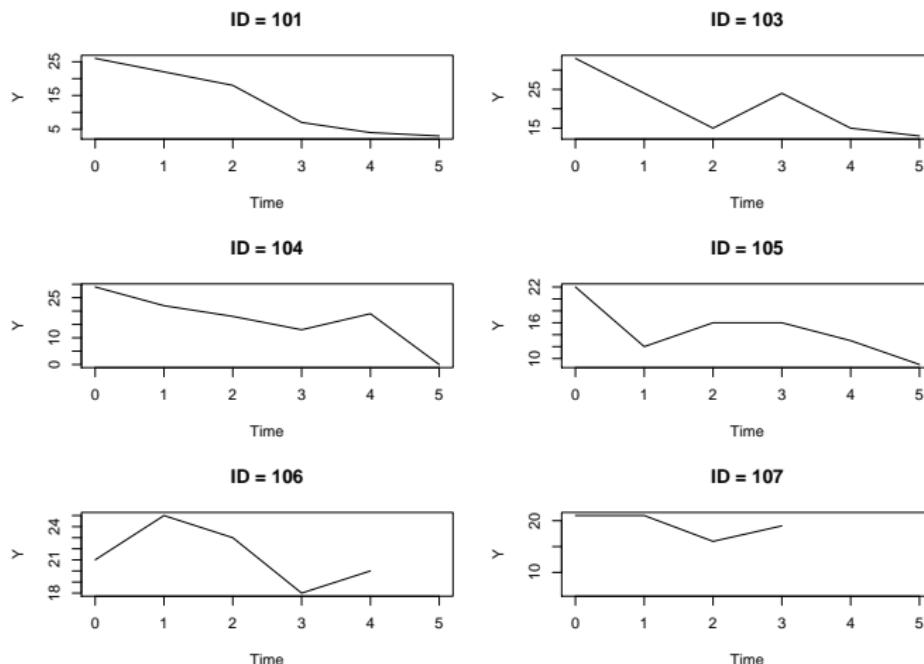


药效学 (Pharmacodynamics, 简称 PD) 模型



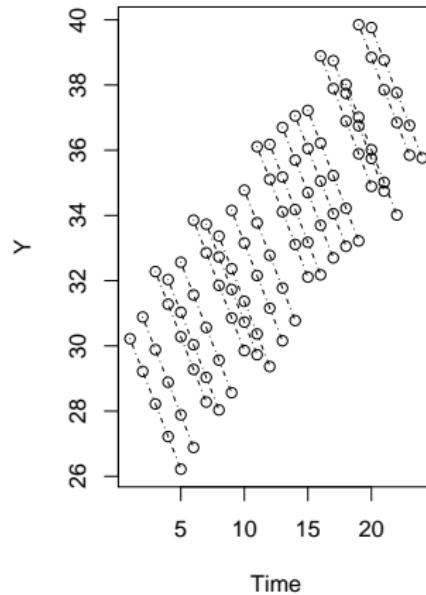
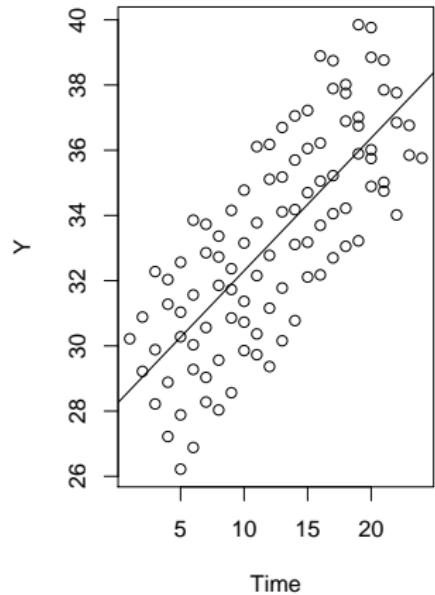
群体 PK/PD：可以尝试对个体分别分析

● 时间序列，数据量太小



群体 PK/PD：可以尝试对群体进行直接分析

- 有时候个体趋势和群体趋势甚至相反



NONMEM, 非线性混合效应模型

```

$PROB Example for PGRD

$INPUT SID SEX AGE RACE HT SMOK HCTZ PROP CON AMT WT TIME SECR DV DROP=RATE
$DATA data5.csv
$PK
    TVCL=THETA(1)
    TVV=THETA(2)
    TVKA=THETA(3)
    CL=TVCL*EXP(ETA(1))
    V =TVV * EXP(ETA(2))
    KA=TVKA*EXP(ETA(3))
    DV=V + KA + CL
$THETA 18.7 87.3 2.13
$OMEGA .128 .142
$OMEGA 1.82
$SIGMA 0.0231
$TABLE ID TIME IPRED IWRES NOPRINT ONEHEADER FILE=sdtabl
$TABLE ID CL V KA NOPRINT ONEHEADER FILE=patabl
$TABLE ID AGE HT WT SECR NOPRINT ONEHEADER FILE=cotab1
$TABLE ID SEX RACE SMOK HCTZ PROP CON NOPRINT ONEHEADER FILE=catab1
$TABLE ID OCC TIME IPRED IWRES NOPRINT ONEHEADER FILE=mutab1
$TABLE SID NOPRINT ONEHEADER FILE=mytab1

```

Inputs to NONMEM
Data on which to base model
Model: Conc = f(...)
Initial Estimates of parameters
Initial Estimates of parameter errors
Initial Estimates of overall error
Specification of output data

使用 R 建立简单的药动学模型

```
library(nlme)
head(Dosing)

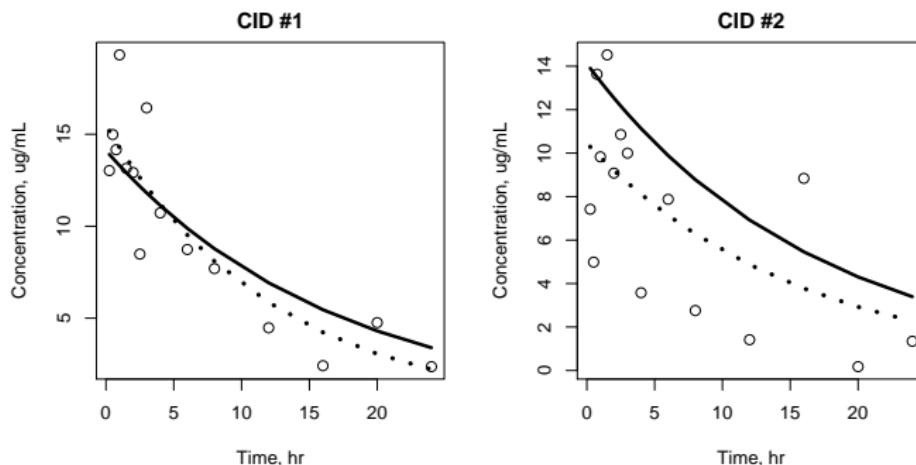
##   ID TIME CONC AMT DOSE MDV AGE WT ISM CLCR
## 1  1  0.00   NA 100 100   1 34.8 38.2 0 42.6
## 2  1  0.25 13.0  NA 100   0 34.8 38.2 0 42.6
## 3  1  0.50 15.0  NA 100   0 34.8 38.2 0 42.6
## 4  1  0.75 14.2  NA 100   0 34.8 38.2 0 42.6

Dosing.fit <- nlme(CONC ~ phenoModel(ID, TIME, AMT, lCl, lV),
                     fixed = lCl + lV ~ 1, random = pdDiag(lCl + lV ~ 1),
                     data = Dosing.grp,
                     start = c(lCl = -5, lV = 0),
                     weight = varConstPower(const = 1,
                     fixed = list(power = 1)), na.action = function(x) x,
                     naPattern = ~!is.na(CONC))

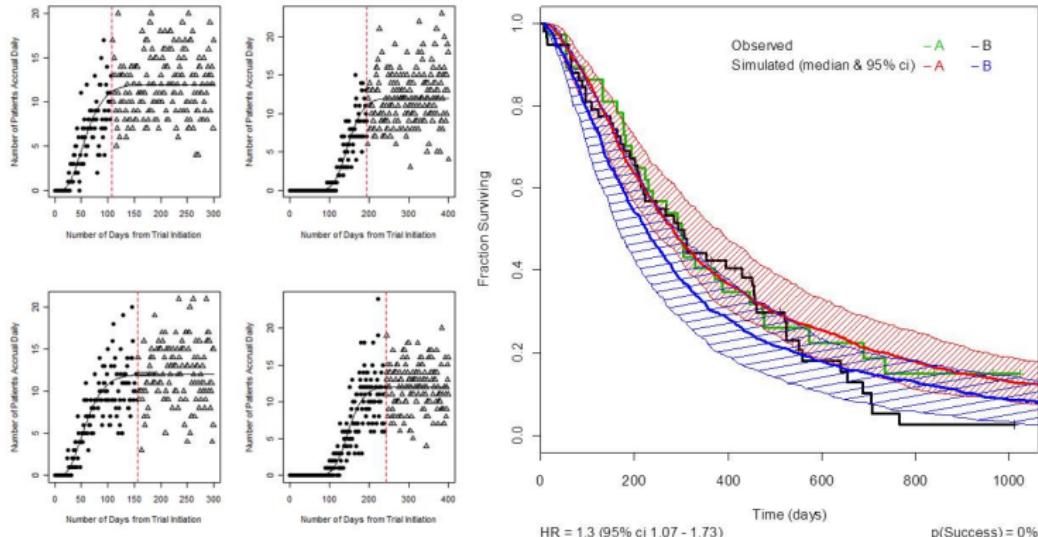
exp(fixed.effects(Dosing.fit))

##      lCl      lV
## 0.421 7.084
```

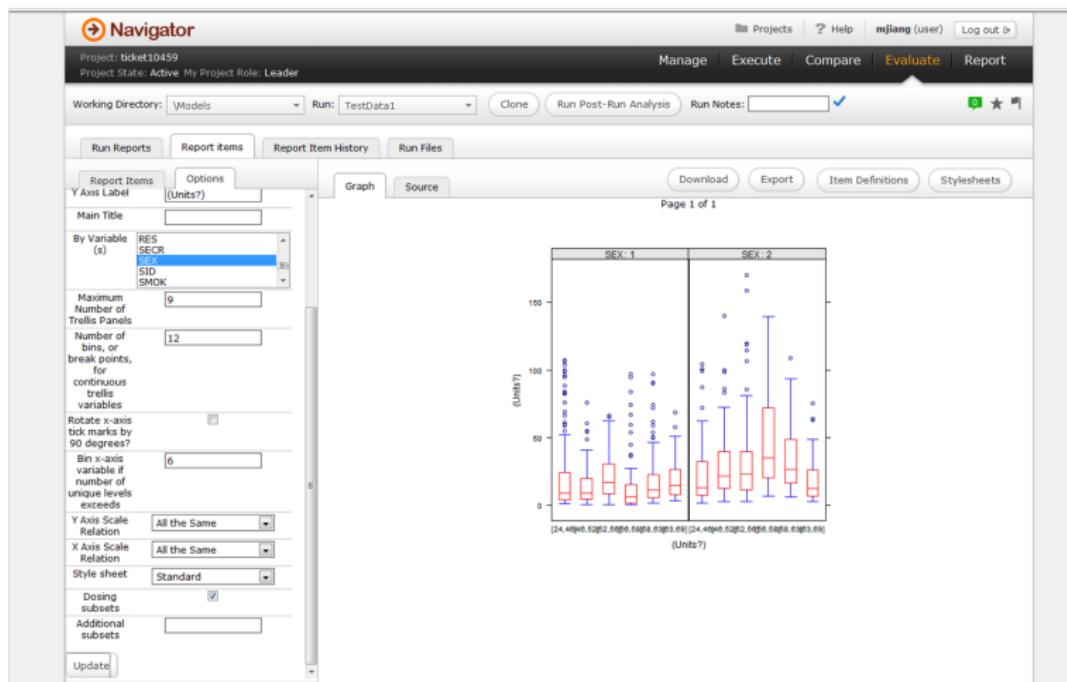
血药浓度的预测结果



统计模型和模拟



系统的实现



目 录

1 数据科学简介

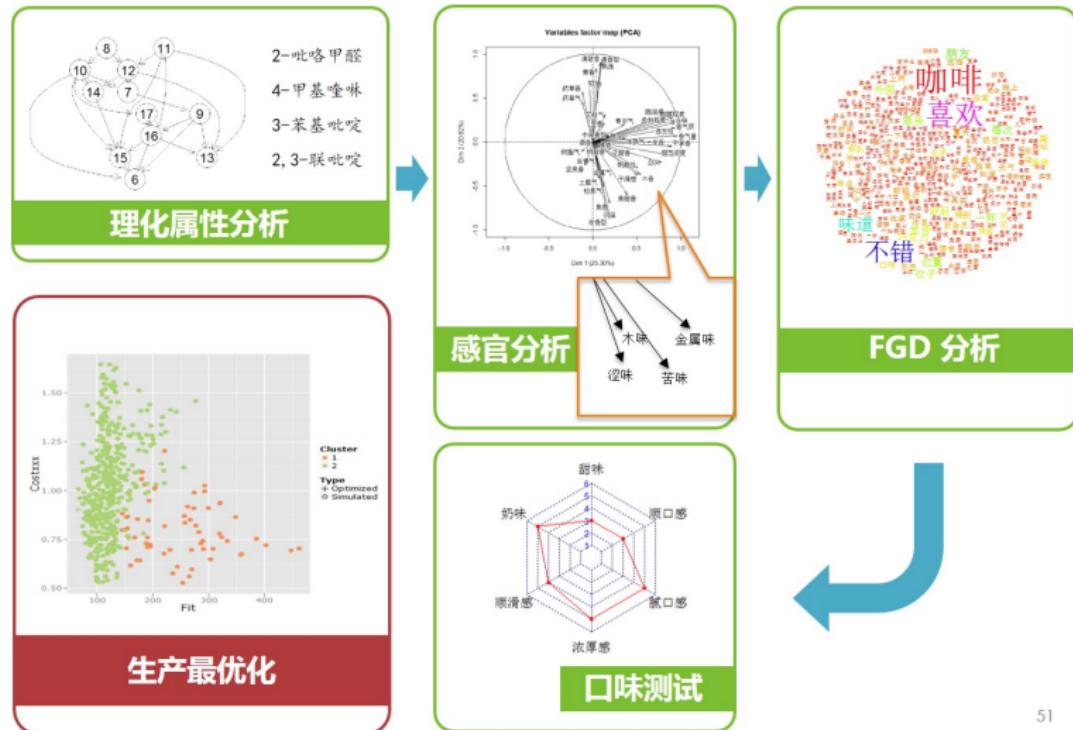
2 数据科学与 R

3 如何成为数据科学家

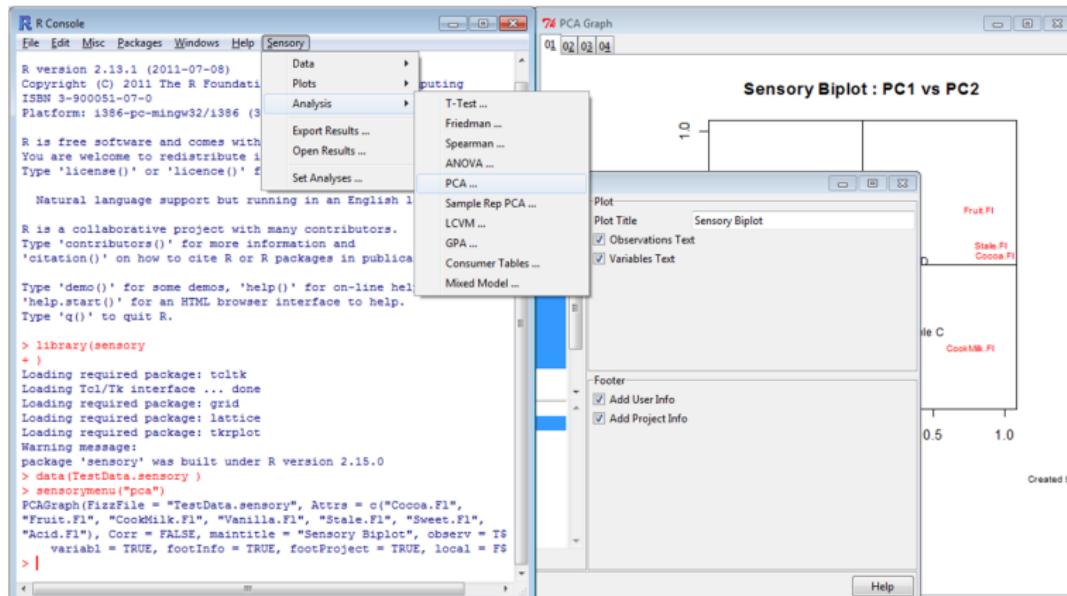
4 案例介绍

- 制药
- 食品
- 零售和快消
- 其他

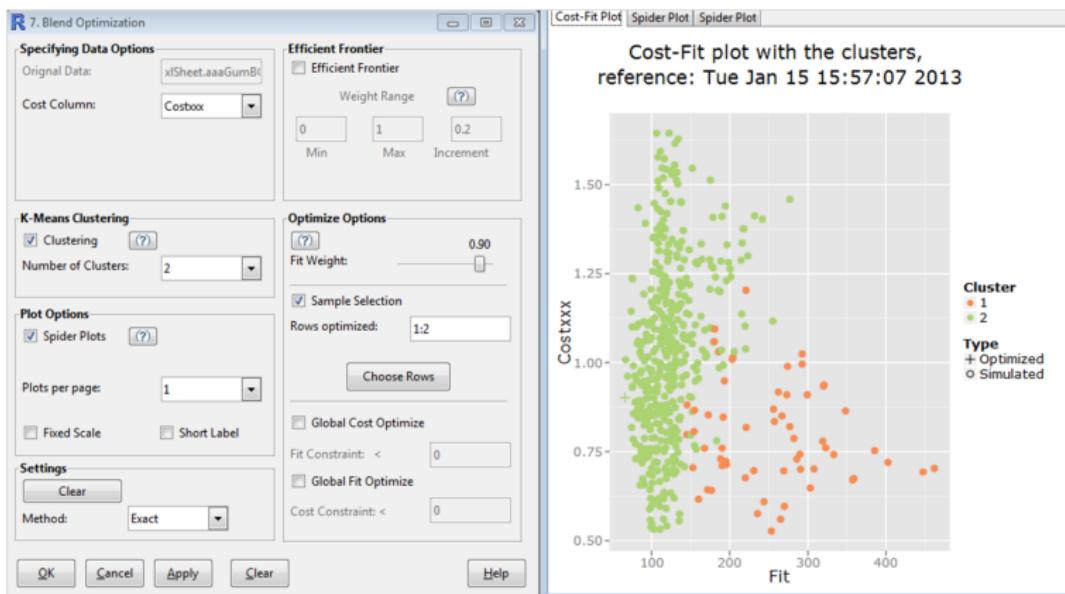
食品研发的流程



感官分析示例



配方优化示例



目 录

1 数据科学简介

2 数据科学与 R

3 如何成为数据科学家

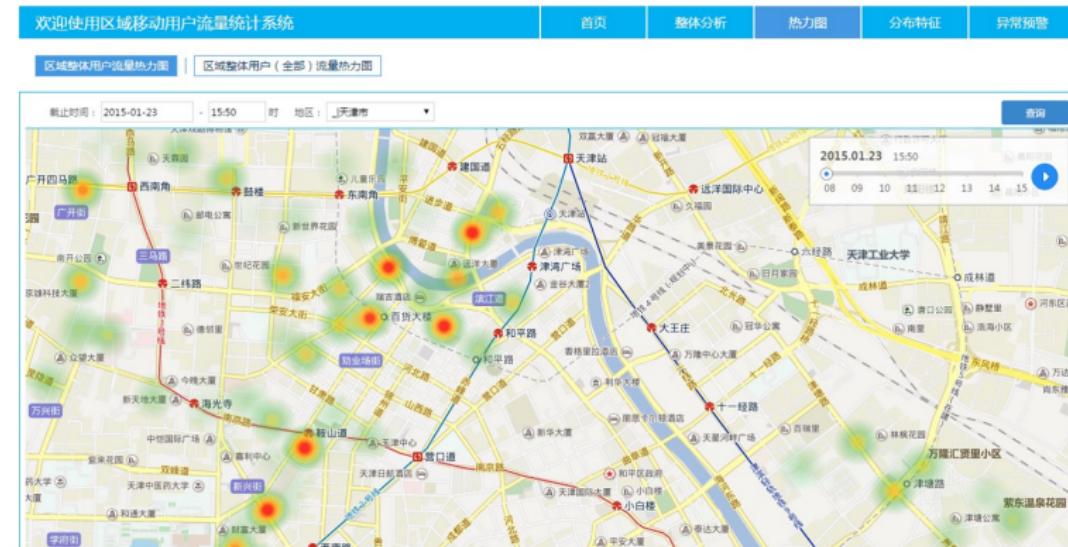
4 案例介绍

- 制药
- 食品
- 零售和快消
- 其他

GIS 与新店选址



移动数据与人流量分析



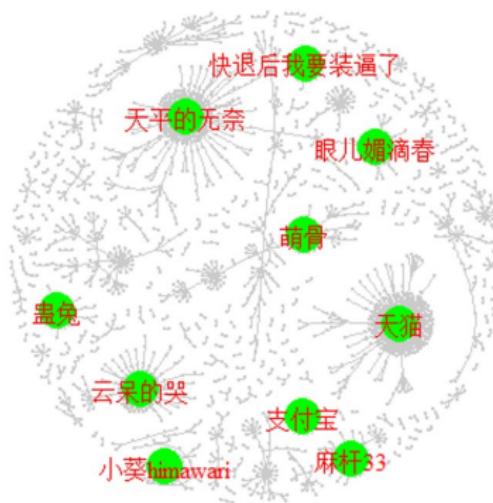
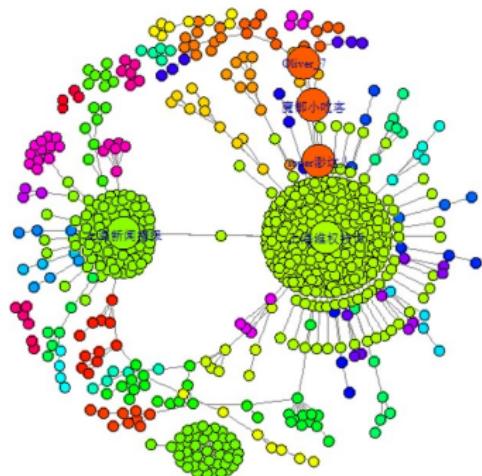
SFA 数据与铺货成功率



产品代言人的选择



口碑监控



目 录

1 数据科学简介

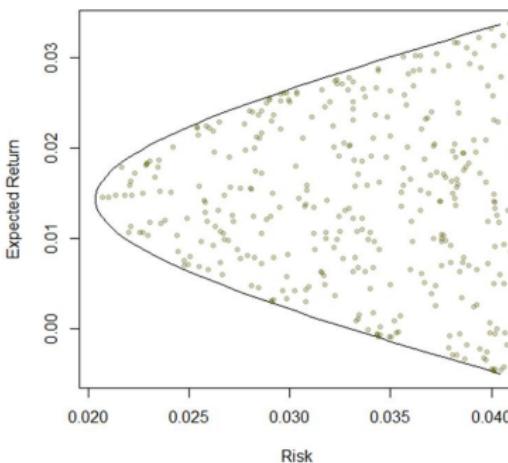
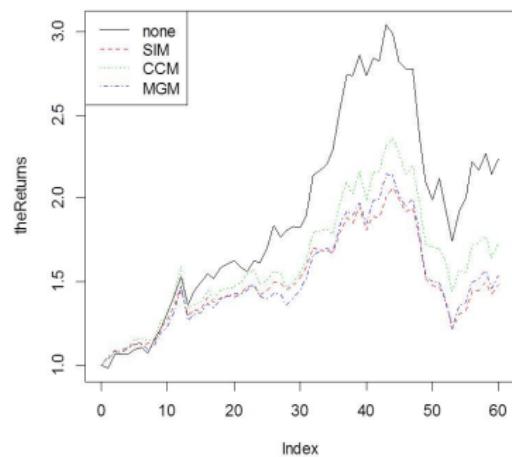
2 数据科学与 R

3 如何成为数据科学家

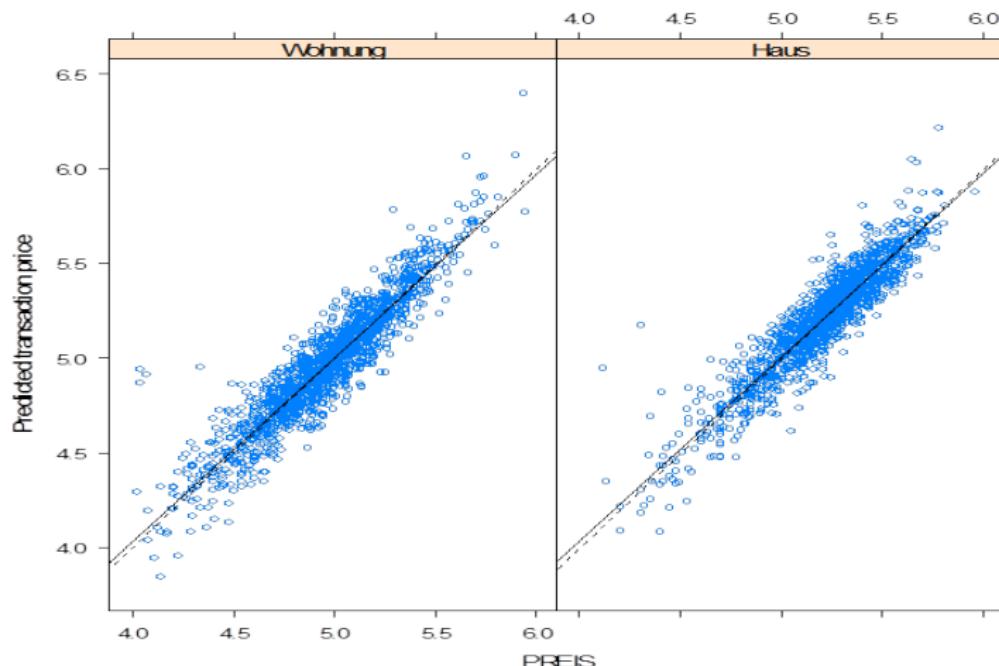
4 案例介绍

- 制药
- 食品
- 零售和快消
- 其他

金融投资

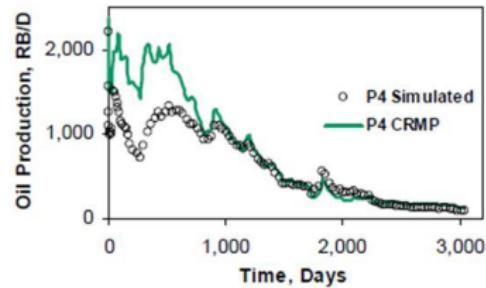
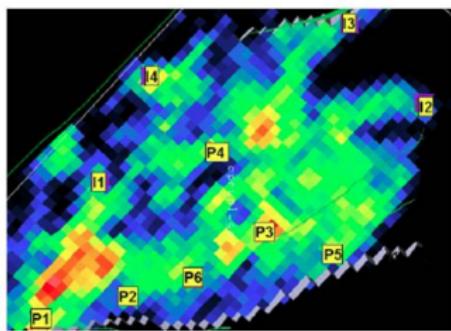
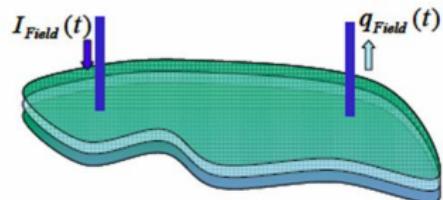
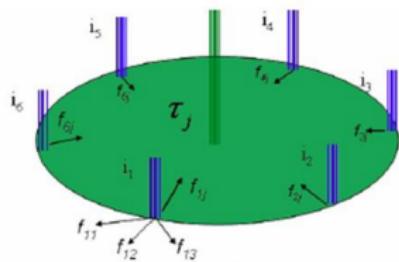


房地产



石油

● 容抗模型



模拟和仿真

图像分析

● 图像识别与分析



ID	总面积(cm ²)	梗面积(cm ²)	最大长度(cm)	R值	G值	B值	描述	区域	等级	年份	报告时间
1 IMG_0008.jpg	711.38	3.10	58.15	0.76	0.57	0.15	IMG_0008			2013-01-17 11:58:42	
2 IMG_0009.jpg	719.20	0.02	48.03	0.77	0.60	0.24	IMG_0009			2013-01-17 11:58:54	
3 IMG_0010.jpg	532.30	9.44	60.84	0.74	0.51	0.06	IMG_0010			2013-01-17 11:59:28	

Thank you!