AI 时代的 R 语言

李舰

2017 中国 R 语言会议（上海）

华东师范大学

2017 年 12 月 02 日

CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY

IT大咖说 知识共享平台

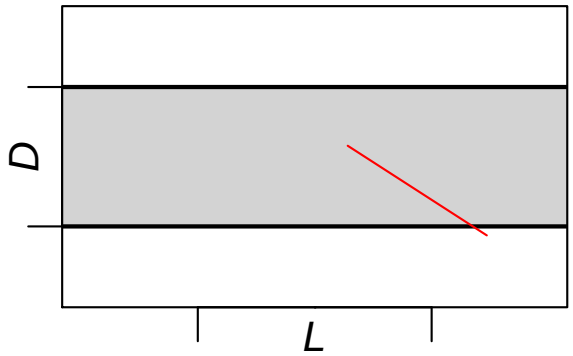
目 录

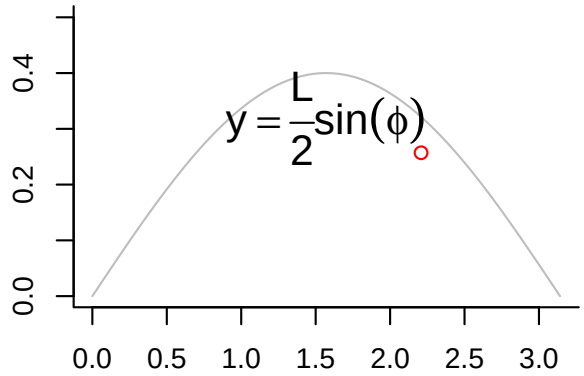
目 录

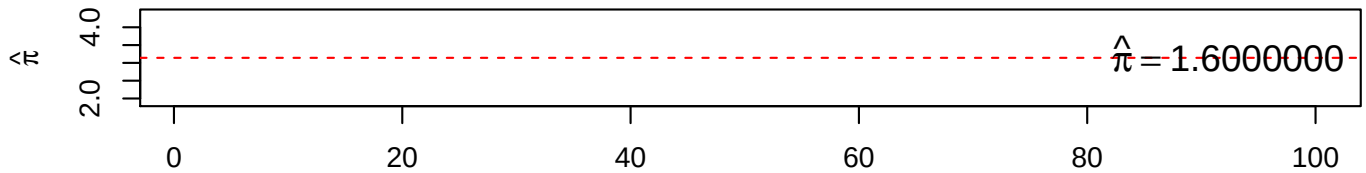
概率论的发展

- 1494 年，现代会计学之父帕西奥利提出奖金分配问题
  - 假设两个人 A 和 B 在玩一种游戏，胜者得 10 分，负者得 0 分，先得 60 分者获胜。如果突然游戏终止，而此时 A 的得分是 50 分，B 的得分是 30 分，奖金应该如何分配给 A 和 B 才算公平？
- 1654 年，帕斯卡正式创立概率论
  - 32 岁的帕斯卡和 54 岁的费马通信讨论奖金分配的问题，得到了正确的答案 7:1。
  - 在和费马的通信中，帕斯卡针对很多概率问题都提出了很多清晰而全面的解决方案，因此后世人们认为是帕斯卡创立了概率论，费马是重要贡献者。
- 1812 年，拉普拉斯完善了古典概率论
  - 拉普拉斯的著作《概率的解析理论》是古典概率完善的标志，明确给出了概率的古典定义，并建立了较为严密的体系。
  - 古典概率论也称为传统概率论，以概率的计算和大数定律为基础，偏重于解决实际问题。

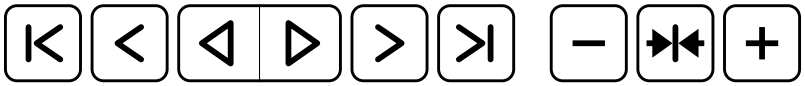
蒲丰投针（1777）



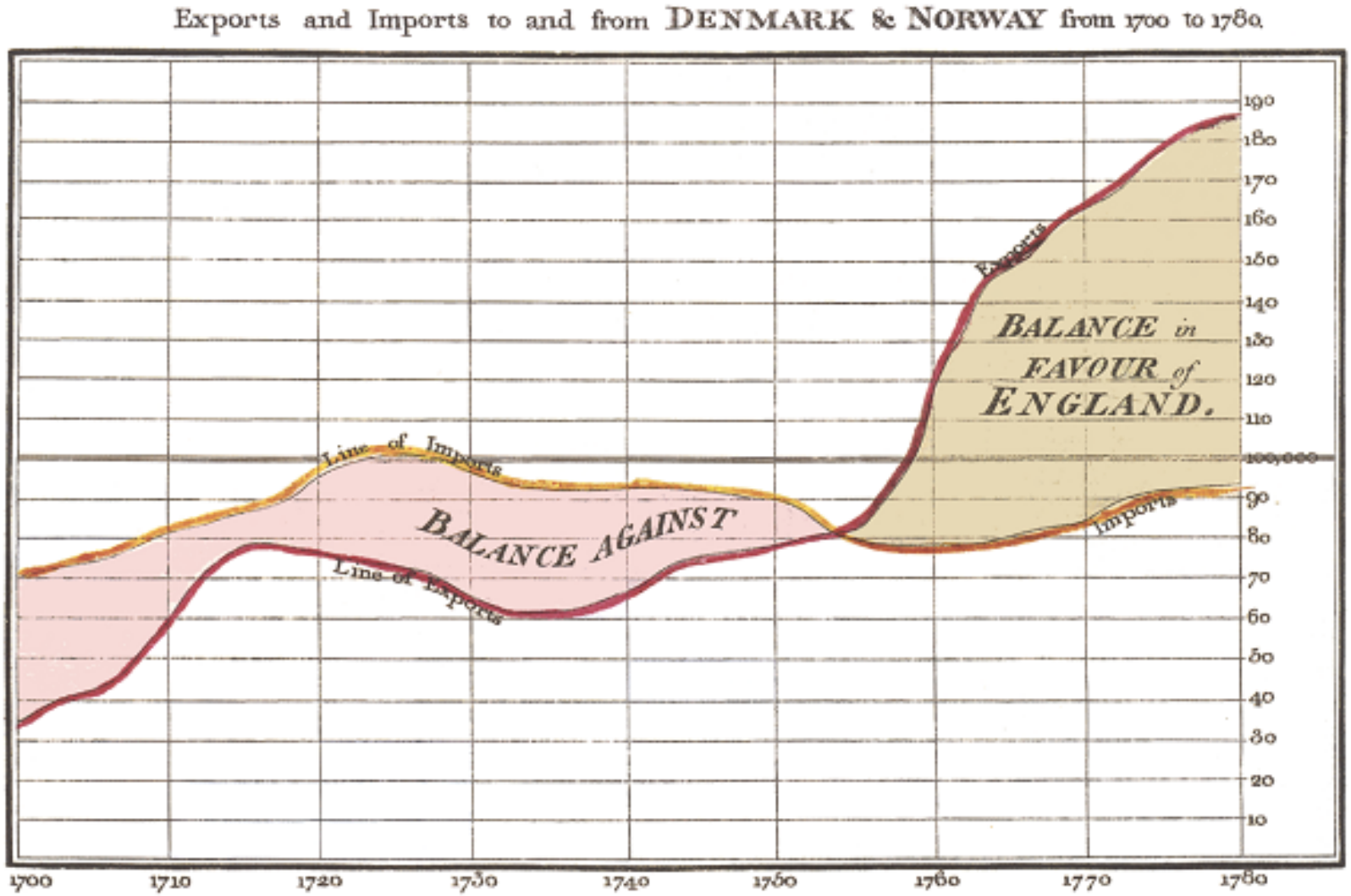




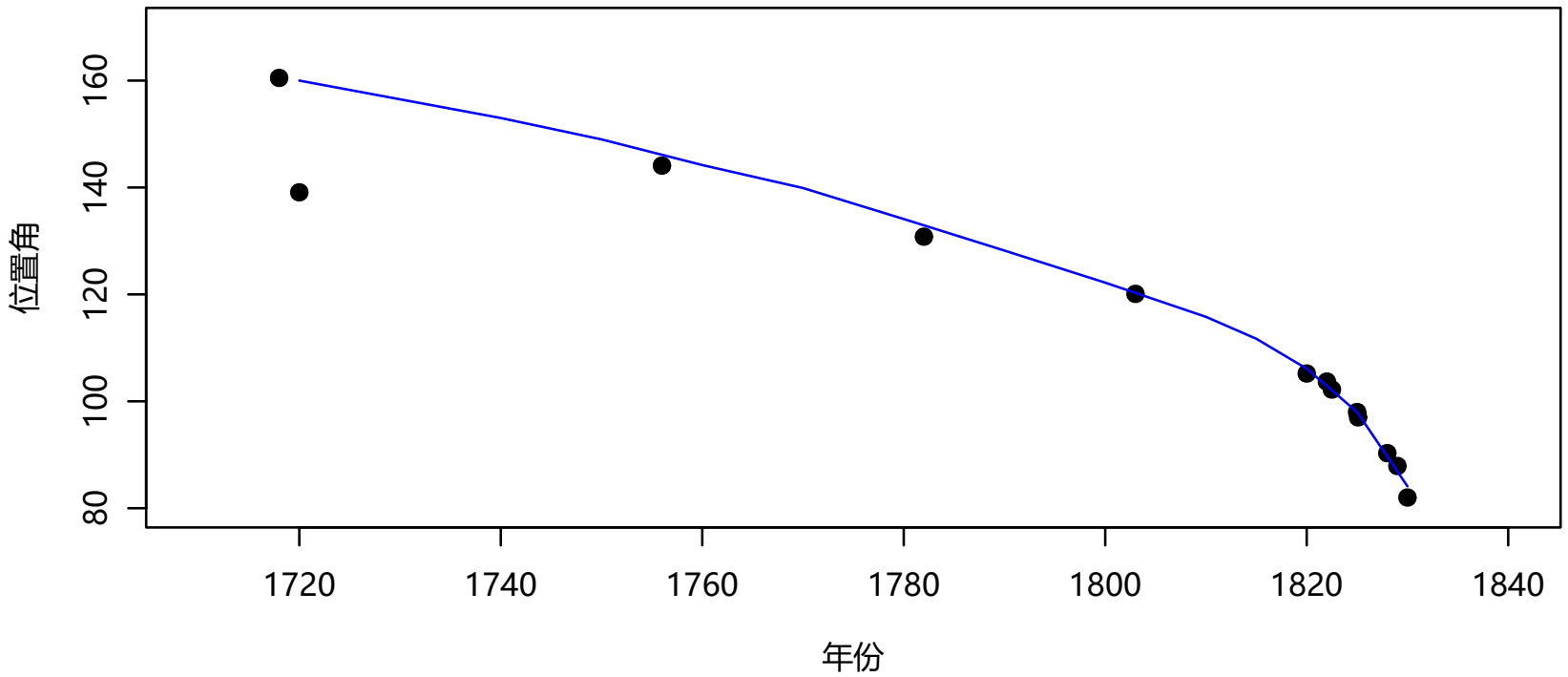
Times of Dropping: 1



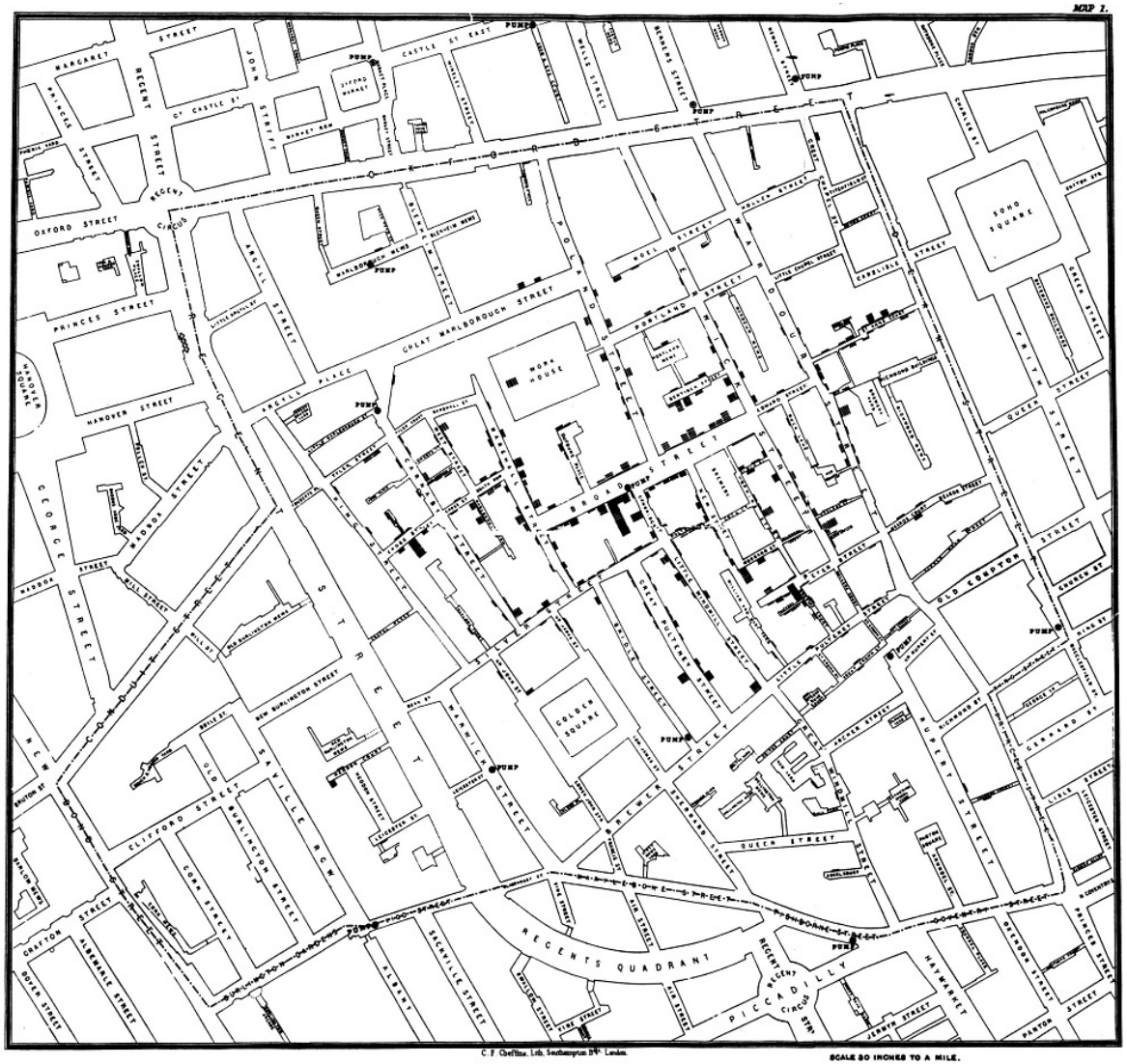
Playfair 的线图（1786）



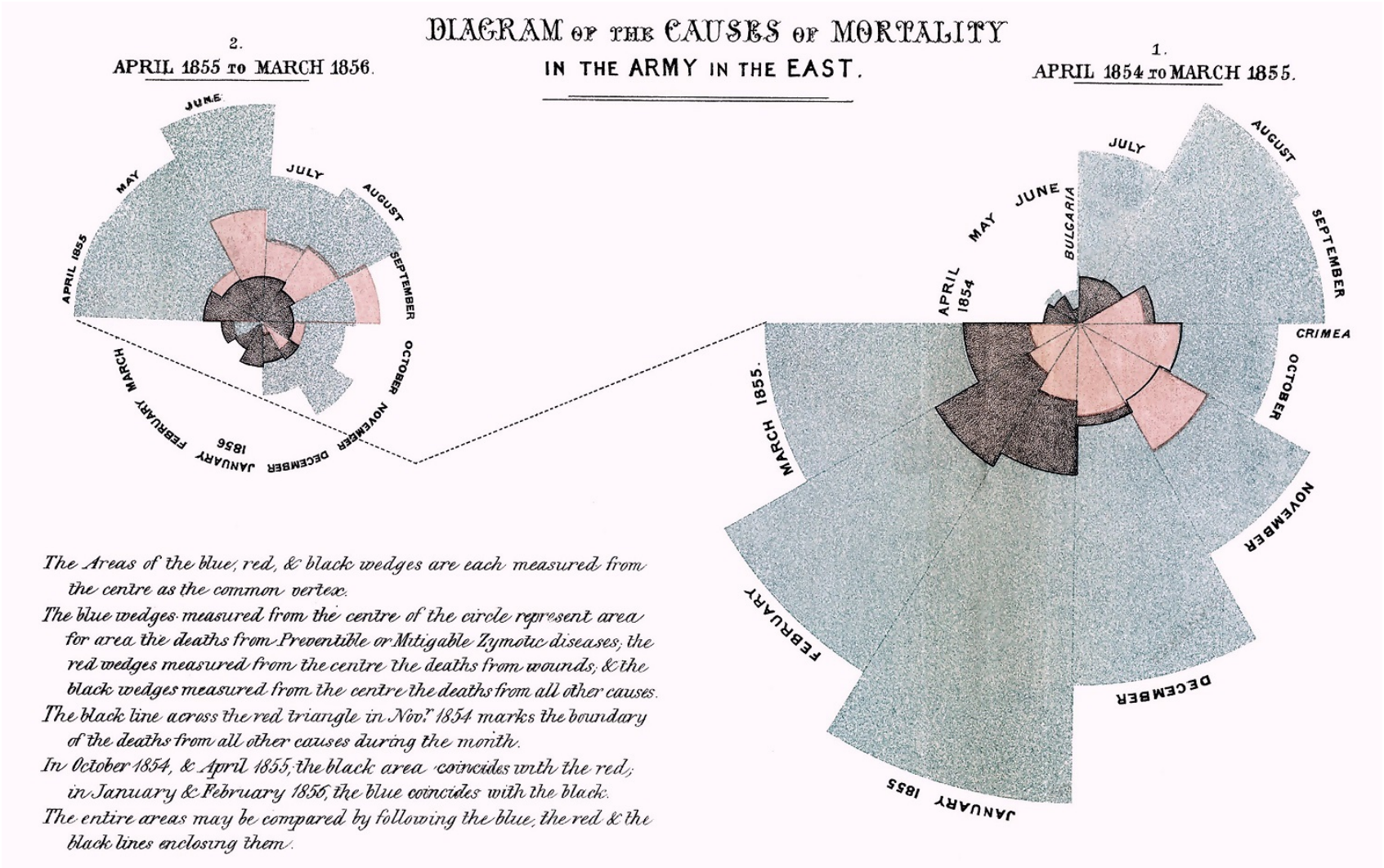
赫歇尔的散点图（1833）



伦敦霍乱防治图（1855）



南丁格尔的玫瑰图（1858）



目 录

统计学的起源

- 凯特勒，近代统计学之父
  - 比利时统计学家凯特勒被誉为是“近代统计学之父”，也被称为是“统计学的鼻祖”。
  - 凯特勒 1834 年成为英国皇家统计协会的创始成员之一，把概率论引入了统计学，取得了很多开创性的成就。
- 高尔顿，早期的统计大家
  - 1855 年发现了父子的遗传身高向平均值回归的现象。1969 年在表哥达尔文《物种起源》的激发下研究了遗传的统计规律，发表了专著《遗传天才》。1892 年发表了专著《指纹学》。
  - 1901 年，高尔顿资助并与其学生皮尔逊等人联合创办了科学期刊《生物统计》。
- 提勒，数理统计的先驱
  - 19 世纪时就发现了很多现代统计学中的成果，但是由于论文都是丹麦语，很多先驱性的工作当时没被重视，直到 1980 年后才广为人知。
  - 也是精算领域的先驱，在最初的精算师国际组织中担任要职。

推断统计学

- 卡尔 · 皮尔逊
  - 1857 年出生于英国，被誉为是“数理统计的创始人”。
  - 1895 年提出皮尔逊分布族，1900 年提出卡方检验。
  - 皮尔逊以倾斜分布的方式提出了革命性的思想，对 19 世纪主流的决定论科学思想进行了沉重打击。
- 费希尔
  - 1890 年出生于英国，被誉为是“推断统计之父”。
  - 1912 年提出了最大似然估计。1925 年出版的《研究者用的统计方法》是第一本推断统计学的教科书，对统计方法的数学化、统计理论的实用化做了突出的贡献，开创了方差分析、统计检验、实验设计等诸多统计学领域。
- 奈曼
  - 1894 年出生于俄国，区间估计和假设检验理论的创始人。
  - 和艾贡 · 皮尔逊（卡尔 · 皮尔逊之子）共同做了很多伟大的研究，1928 年提出了区间估计的理论，1933 年完善了假设检验的理论。

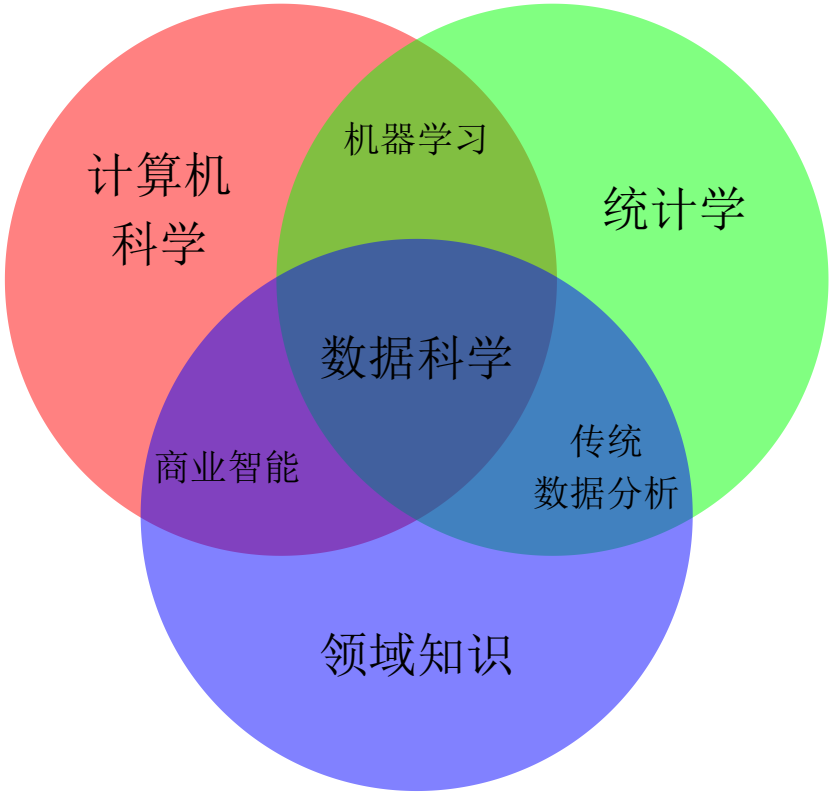
目 录

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？

- 数据科学（Data Science），也称为资料科学 [a]



[a] 图形摘自《数据科学中的 R 语言》

计算机的发明

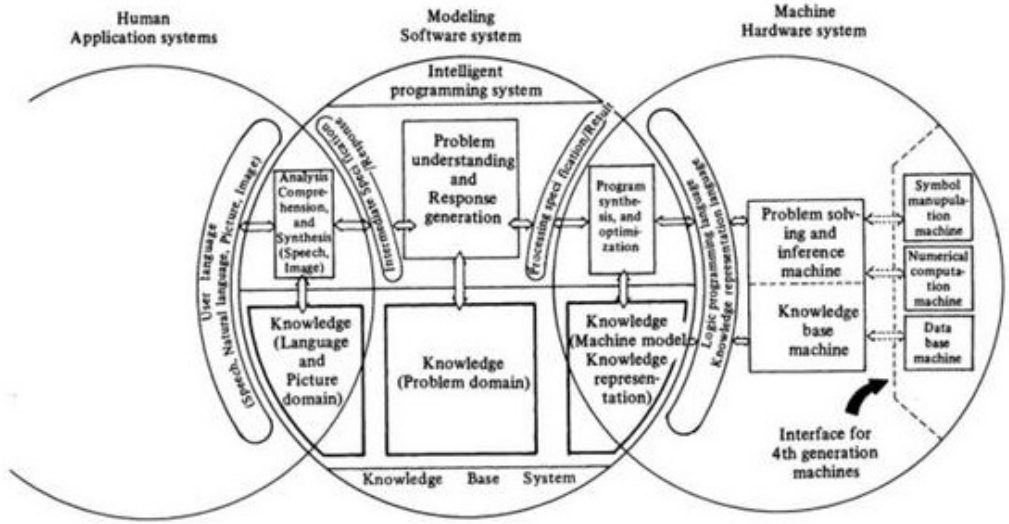
- ENIAC
  - 世界上第一台通用计算机，1946 年在美国的宾夕法尼亚大学诞生。
  - 美国国防部用它来进行弹道计算，占地 170 平方米，重达 30 吨，每秒钟可进行 5000 次运算。
  - ENIAC 以电子管作为元器件（一共用了 18000 个电子管），所以又被称为电子管计算机，也称为第一代电子计算机。
- UNIVAC I
  - 世界上第一台商用计算机，1951 年研制成功并交付给美国人口统计局用于人口普查。
  - 采用晶体管作为元器件，晶体管不仅能实现电子管的功能，又具有尺寸小、重量轻、寿命长、效率高、发热少、功耗低等优点。
  - 是第二代计算机的代表。

早期的 AI 时代

- 人工智能
  - 1940 年，控制论之父维纳研究计算机如何像大脑一样工作。
  - 1950 年，人工智能之父的图灵提出了著名的“图灵测试”。
  - 1956 年，达特茅斯大学的会议上正式使用了“人工智能”这个术语，宣告了这个学科的诞生。
- 神经网络
  - 1943 年，心理学家 Warren McCulloch 和数理逻辑学家 Walter Pitts 提出神经元的数学模型。
  - 1957 年，康内尔大学教授 Frank Rosenblatt 提出的“感知机”模型，第一次用算法来精确定义神经网络，其通过训练学习逻辑运算的成功引起了轰动。
  - 1969 年，Marvin Minsky 和 Seymour Papert 出版了《感知机：计算几何简介》，该书指出了神经网络技术的局限性。
  - 1986 年，Hinton 和 David Rumelhart 发表了 BP 算法。

“第五代计算机”时代

- 日本第五代计算机计划
  - 1978 年，日本通产省委托东京大学计算机中心主任的 TohruMoto-Oka 研究下一代计算机系统。
  - 1981 年，TohruMoto-Oka 为首的委员会提交了报告《知识信息处理系统的挑战：第五代计算机系统初步报告》。
  - 日本人选择了逻辑程序语言 Prolog，走的是规则和逻辑路线。
  - 1992 年，日本政府宣布第五代计算机研制失败。



数据挖掘时代

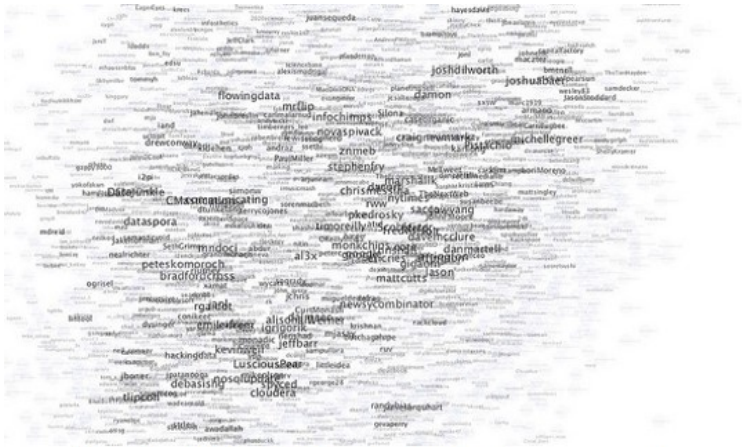
- Data Mining
  - 上个世界 90 年代开始流行，世纪之交时跟随人们对知识爆炸的预期变得很火。
  - 一般是指从大量的数据中通过算法挖掘出隐藏于其中信息的过程。
  - 目前已经融入到机器学习的范畴中，一般认为使用机器学习方法、遵循数据挖掘流程来进行数据分析。

商业智能时代

- Business Intelligence (BI)
  - 上个世界 90 年代末开始在业界出现，本世纪初非常火热，主要针对当时流行的“海量数据”进行存储和分析。
  - 通常指用数据仓库、OLAP、数据挖掘和数据可视化技术进行数据分析以实现商业价值。

MeasuresLevel	大类	上海	云南	内蒙古	北京	台湾	吉林	四川	天津	宁夏	安徽	山东	山西	广东
环境分						5.487								
	丽人	7.077	7.157	7.05	7.165		6.908	7.059	7.164	7.188	7.054	7.152	7.051	7.079

大数据时代



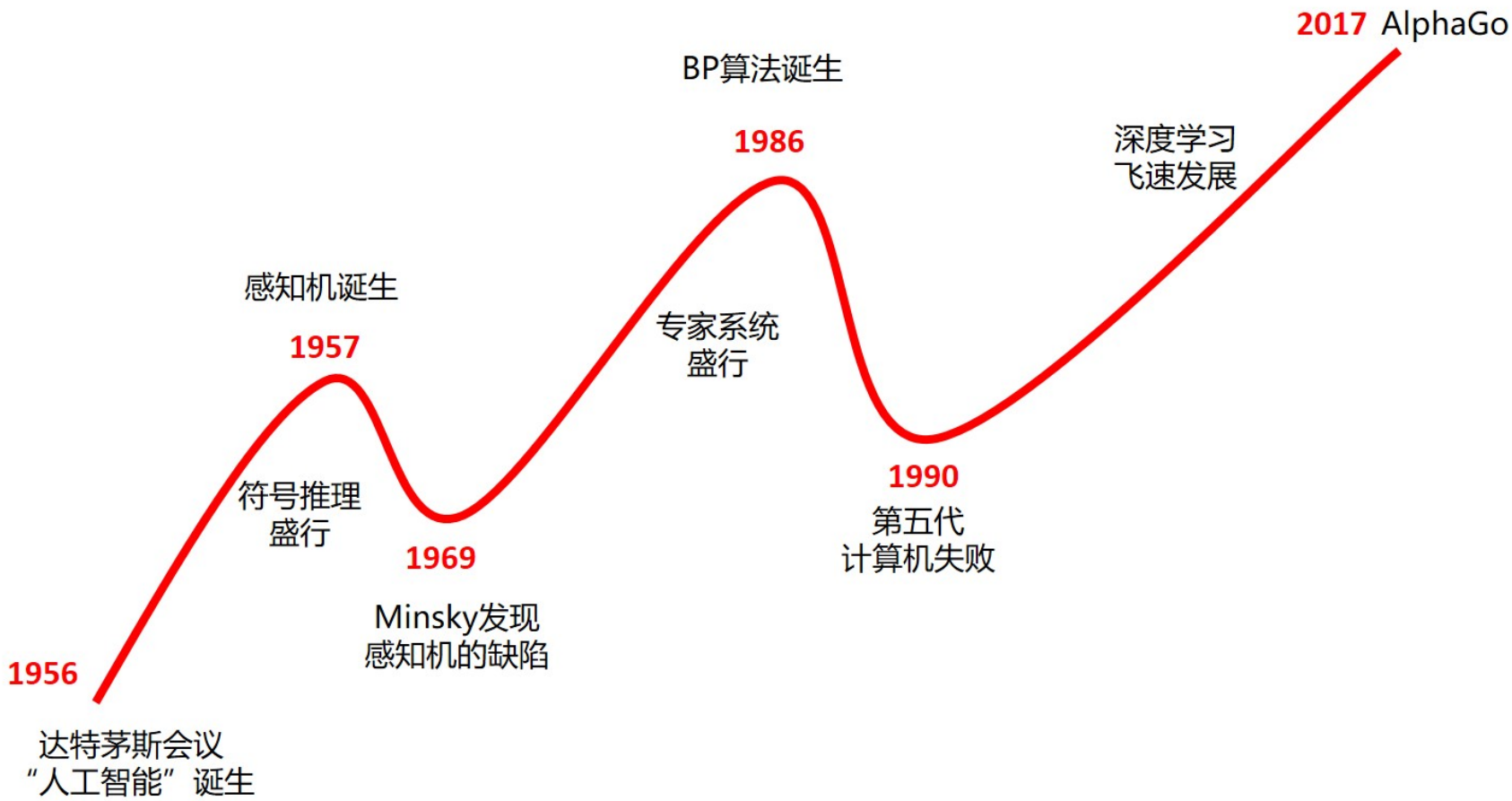
很大的数据量（Volume）

快速地响应（Velocity）

巨大的价值（Value）

多样的形式（Variety）

AI 时代



目 录

目 录

R 的诞生（I）

- S 语言是 R 语言的前身
  - S 语言诞生于 20 世纪 70 年代由 John M. Chambers 领导的贝尔实验室统计研究部。
  - 1998 年美国计算机学会（ACM）授予了 S 语言的主要设计者 John M. Chambers “软件系统奖”。
  - 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。
  - 2008 年，TIBCO 收购了已改名的 Insightful 公司。
- R 语言吸收了很多 Scheme 语言的特性
  - Scheme 语言诞生于 1975 年的 MIT，是 LISP 语言的一个方言。
  - 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 的书，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 的诞生（II）

- 1993 年，R 语言诞生
  - 1992 年 Ross Ihaka 和 Robert Gentleman 在奥克兰大学成为同事。后来为了方便教授初等统计课程，二人开发了一种语言；而他们名字的首字母都是 R，于是 R 便成为这门语言的名称。
  - 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。
  - 苏黎世理工学院的 MartinM 极力劝说两位原作者公布源代码，让 R 成为自由软件。于是 1995 年 6 月 R 的源代码正式发布到了自由软件协会的 FTP 上。
- 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 年开始，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，“万物皆对象”）；
  - modular，built from standardized pieces（模块化，由标准化块构建）；
  - collaborative，aworld-wide，open-source effort（协作性，全球范围的开源力量）。

R 在数据分析应用领域的发展

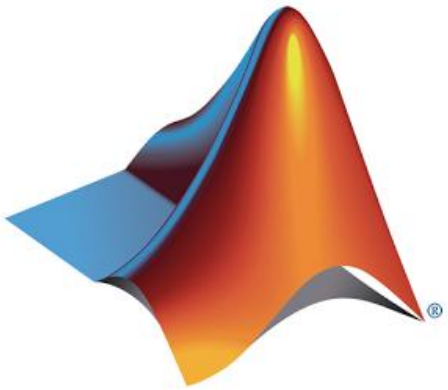
- KDNuggets 关于“数据分析/数据挖掘/数据科学”的调查
  - R 从 2011 年 KDNuggets 调查数据分析类编程语言开始就排名第一，从 2012 年开始，在关于“数据挖掘和数据分析”工具的调查中，也超过了 Excel 和 Rapidminer 成为第一。
  - 2017 年 8 月的“数据科学和机器学习平台”的调查中，Python 超过 R 成为第一。
- IEEE 热门编程语言排行榜
  - 2017 年 8 月发布的排行版中，R 排名第 6，前 10 名的编程语言为 Python、C、Java、C ++、C#、R、JavaScript、PHP、Go、Swift。
- TIOBE 编程语言排行榜
  - 2017 年 11 月发布的排行版中，R 语言排名第 11，前 10 名的编程语言为 Java、C、C ++、Python、C#、JavaScript、VB.NET、PHP、Delphi、Assembly Language。

目 录

统计计算

- 简介
  - R 的设计目地之一是成为一个统计计算的编程环境。
  - 最早版本的 R 用 Fortran 编写，当前版本主要是 C/C++ 。R 包可以很方便地支持 C/C++ 和 Fortran 的开发。
- R 的优劣
  - R 语言编程容易，代码易读。
  - 性能相对较差，不过可以通过集成 C/C++ 或 Fortran 的库来解决。

矩阵式编程



GAUSS
Expanding Probabilities.

- 简介
  - R 最早出现在公众的视线是作为矩阵式编程语言而著称的，易于满足统计建模的需求。
  - 早期的 R 常被拿来和 Matlab、GAUSS 进行比较。
- R 的优劣
  - 开源免费，第三方资源丰富。
  - 性能相对较差，受制于开源的 BLAS/LAPACK。

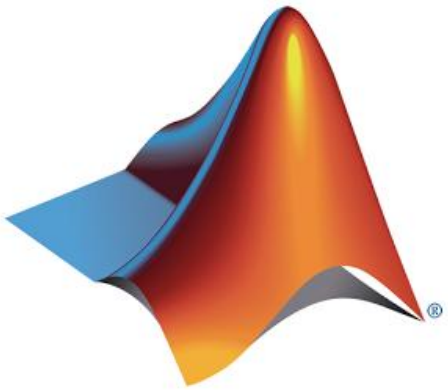
数据可视化

- 简介
  - 早期的 R 受欢迎的重要原因是可以生成高质量的矢量图形。
  - 编程灵活，是自定义统计图形的首选。
- R 的优劣
  - 可视化资源非常丰富，其中 ggplot2 包实现了“The Grammar of Graphics”中的语法，几乎成了统计图形作图的事实标准。
  - 动态可视化的能力比较弱，但是和目前主流的 JavaScript 有很好的结合，社区中存在很多像 recharts 这样的优秀第三方包。

统计学方法

- 简介
  - R 语言崛起之初常被拿来和 SAS 比较。
  - 业界中 R 和 SAS 可以进行很好的配合。
- R 的优劣
  - 分析方法非常丰富，编程轻松简洁，代码开源而透明。
  - 性能相对较差，直接处理大量数据的能力弱，之前在业界缺少大公司的背书，微软加入后有了很大改观。

蒙特卡洛方法



Arena®

- 简介
  - R 内置的统计分布非常丰富，随机数机制也非常灵活。
  - 编程灵活而简洁，在蒙特卡洛领域有很广泛的应用
- R 的优劣
  - 开源免费，编程容易。
  - 仿真的图形界面比较弱。

最优化方法

AMPL

STREAMLINED MODELING
FOR REAL OPTIMIZATION

LINDO SYSTEMS INC.

- 简介
  - 早期 R 的优化功能很弱，常被用来和一些商业软件进行对比。
  - 目前具备了比较完善而强大的最优化能力，在开源软件中处于领先地位。
- R 的优劣
  - 开源免费，资源丰富，编程灵活。
  - 相比商业软件，性能较差且缺少一些复杂的方法，但是可以结合 COIN-OR 进行扩展。

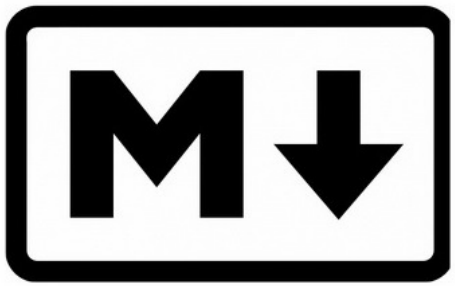
机器学习

- 简介
  - 早期机器学习资源不如 Python 丰富，导致很多 R 用户投入了 Python 阵营。
  - 近来机器学习的功能比较完善，一些主流工具的作者直接参与了 R 包的开发，比如 xgboost。
  - 最近得到了一些商业公司很好的支持，有成为商业大数据机器学习主流框架的趋势。
- R 的优劣
  - 学习门槛低，代码易读。
  - 直接使用的运算性能相对较差。

深度学习

- 简介
  - 早期神经网络功能弱，并且主流深度学习框架很少直接提供 R 的支持。
  - MXNet 对 R 提供了完美的原生支持，在 RStudio 的贡献下，目前也有了 Tensorflow 和 Keras 的 R 包。
- R 的优劣
  - 在 R 的框架下可以完美地融合到分析流程中，Windows 下的安装和使用非常容易。
  - 案例和文档相对较少，相比 Python，在深度学习领域不是很主流。

文学化编程



- 简介
  - 早期的 R 就引入了文学化编程的思想，Sweave 时代就和 LaTeX 结合得很好。
  - 谢益辉的 knitr 发布后颠覆了这个领域，后续的 bookdown 有可能会改变科技类书籍出版的模式。
- R 的优劣
  - 资源非常丰富，功能非常强大。
  - 学习 LaTeX 的门槛相对较高，但基于 Markdown 框架的话可以比较容易。

系统架构

- 简介
  - 早期的 R 并未向后台服务器方向发展，在业界通常是和 Java 进行整合。
  - Shiny 问世后拥有了完美的服务器框架，此外 fiery 包也可以提供强大的后台服务能力。
- R 的优劣
  - Shiny 框架非常简洁，可能是门槛最低的网站系统构建工具，与 JavaScript 可视化库的结合也非常容易。
  - 性能较差，一般不用作生产系统。

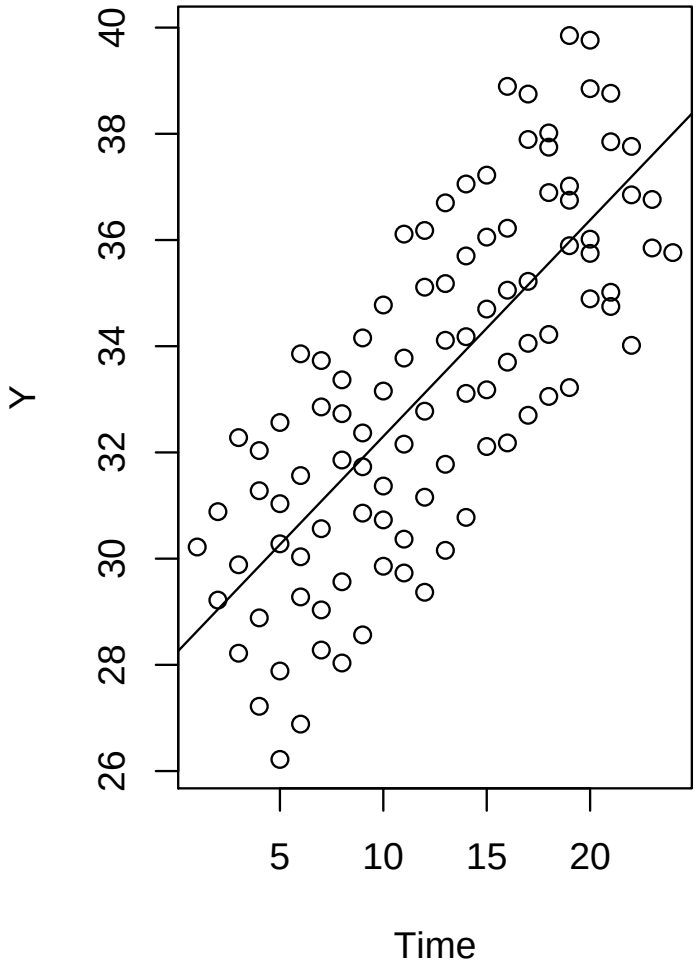
目 录

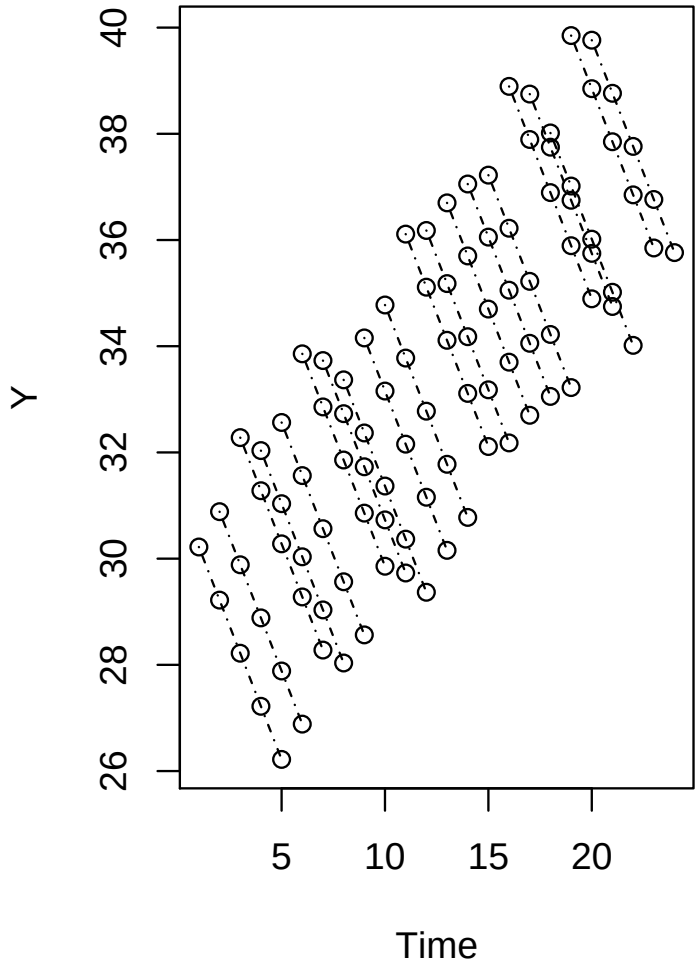
针对使用需求找准定位

- R User，使用者
  - 用 R 来进行数据分析。
  - 把 R 当作一种统计分析及作图软件，只是通过程序和函数调用而已。
  - 自如地通过 R 语言来操作数据和分析建模。
  - 熟练地编写函数和控制语句简化操作。
- R Developer，开发者
  - 基于 R 开发供他人使用的函数及工具。
  - 把 R 当作一种计算机编程语言。
  - 除了用 R 进行分析，还基于 R 为他人开发工具。
- R Hacker，高手
  - 优化和改变 R。
  - 由于 R 专注于统计分析而且很开放，非常欢迎各种外部工具（例如 C、Fortran、Hadoop、Spark），统计和编程高手可以很方便地扩展 R。

要有“神仙意识”

专注建模而不是套用工具





Thank you!