数据科学时代的数据智慧和十大原则

王小宁

中国传媒大学数据科学与智能媒体学院

2021 年 5 月 23 日

目录

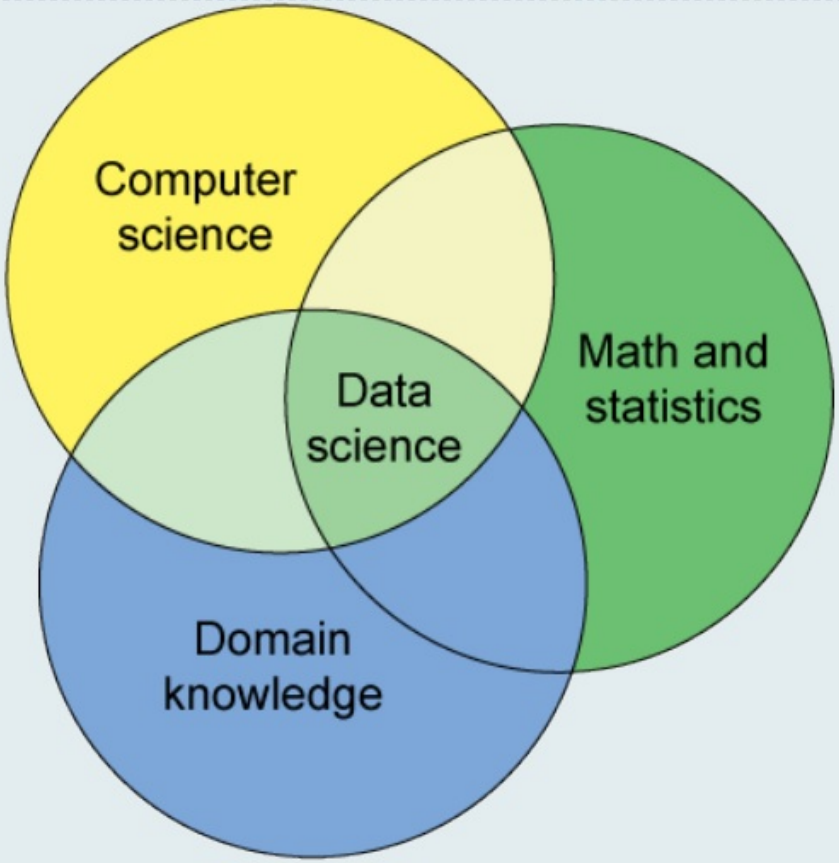
简述

三篇文章

- 让我们拥抱数据科学English,中文,幻灯片
- 数据科学中的 “数据智慧”English,中文
- 有效统计实践的十项简明原则,English,中文

让我们拥抱数据科学

Data science is all the rage



http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png

SSNSW First President: Helen Newton Turner

A statistician and a geneticist

(1908-1995)

Helen Newton Turner, the animal geneticist who spent most of her professional life working for the improvement of the wool industry, died this week [late November1995-Ed.]. Not many scientists in Australia have contributed as much and as directly to the growth and well-being of a major Australian industry as she did. Not that she would have agreed with this description; she was a very modest woman and would have said, 'It wasn't me, it was my team'.

http://sydney.edu.au/senate/documents/Students/Turner_obituary.pdf

Annals First Editor: Harry C. Carver

A mathematical statistician and an aerial navigation expert

(1890-1977)

Decoration for Exceptional Service by US Air Force

When he was ready to retire at 70, he set up criteria for average temperature, total rainfall, number of days of sunshine, etc., and then conducted a systematic search of the U.S. weather records to find the winning location.

His decision was to choose Santa Barbara, California, and he moved.

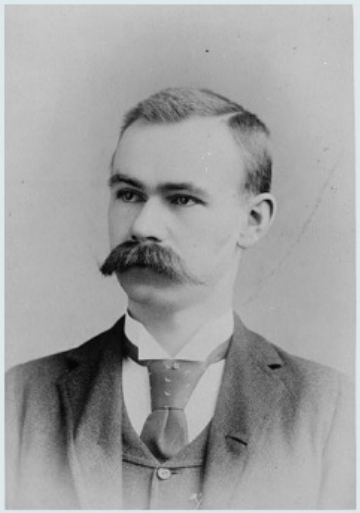
Carver: early "machine learner"

Air navigation is the process of planning, recording, and controlling the movement of a craft from one place to another.

Carver got the first Hollerith tabulating machine at U-M (later IBM).

His approach to air navigation could be viewed with a modern lens as on-line "machine learning" -- fast estimation based on instantly measured data through optimization (with possibly uncertainty measure)

Inventor of Hollerith Machine: Herman Hollerith



A statistician and an inventor

(1860-1929)

Founder of the Tabulating Machine Company

that later merged to become IBM

Hollerith is widely regarded as the father of modern machine data processing. With his invention of the punched card evaluating machine the beginning of the era of automatic data processing systems was marked. His draft of this concept dominated the computing landscape for nearly a century.

-- Wikipedia

Turner + Carver = "Data Scientist"

Putting all the traits of Turner and Carver together, we get a good portrait of data scientist:

1. Statistics (S)
2. Domain (science) knowledge (D)
3. Computing (C)
4. Collaboration ("team work") (C)
5. Communication (to outsiders) (C)

Data Science = SDC^3

W. Cochran (1953): Sampling Techniques

A statistician
(1909 – 1980)

S. S. Wilks Medal of ASA

Set up (bio)stat depts of
Iowa-State, NC-State, Johns Hopkins, and Harvard

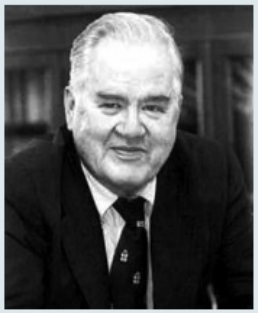
http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Cochran.html

http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Cochran.html

"Our knowledge, our attitudes, and our actions are based to a very large extent upon samples. This is equally true in everyday life and in scientific research."

"But when the material is far from uniform, as is often the case, the method by which the sample is obtained is critical, and the study of techniques that ensure a trustworthy sample becomes important." -- Introduction in Sampling Techniques

J.W. Tukey (1962): Future of data analysis



A mathematician
(1915 – 2000)

U.S. Medal of Science
IEEE Medal for co-invention of FFT

It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out "how things really are" by investigation and the confrontation of insights with experience.

Tukey's definition of "data science"?

To fortify our position in DS, we should focus on

Critical thinking enables Statistics + Domain knowledge

Computing

(parallel computation, memory and communication dominate scalability)

Leadership, interpersonal, and communication

abilities enable collaboration + communication with outside

Claiming "data science" as our own in 1998 by Jeff Wu

Quotes from Jeff Wu's inaugural lecture of his Carver (!) Chair Professorship at Univ of Michigan in 1998.

Statistics = Data Science ?

C. F. Jeff Wu

University of Michigan, Ann Arbor

A proposal:

"Statistics" $\longrightarrow$ "Data Science"

"Statisticians" $\longrightarrow$ "Data Scientists"

Quotes from Wu's slides (cont)

- Several good names have been taken up: computer science, information science, material science, cognitive science
- "Data Science" is likely the remaining good name reserved for us

L. Breiman (2001): Statistical modeling: the two cultures

A probabilist, and statistician, machine learner
(1928 – 2005)

CART, Bagging, Random Forests

"If our goal as field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a diverse set of tools."

让我们拥抱数据科学-总结

- 拥抱数据科学
- 思考数据与现实业务的结合
- 计算科学，适当学习计算机原理，硬件和存储（并行计算和分布式计算）
- 结果的可重复性

数据智慧

数据科学中的“数据智慧”

“数据智慧”-定义

“数据智慧“是将**领域知识、数学和方法论与经验、理解、常识、洞察力以及良好的判断力相结合，思辨性地理解数据和依据数据做决策的一种能力。**

“数据智慧“是数学、自然科学和人文主义这三方面能力的融合，是科学和艺术的结合。在缺乏有实践经验者的指导下，个人很难仅仅靠从读书中获得“数据智慧”，想要学习它的最好方法就是和拥有它的人**一起共事。**

要回答的问题

- 数据科学的问题最开始往往来自于统计学或者数据科学以外的学科。
- 如大脑是如何工作的？解决这个问题需要与专业领域的人合作。
- 由于误差的存在，我们谨慎的避免对于数据中出现的模式进行过度拟合，要提前做探索性数据分析（EDA）。

1. 数据收集

- 数据是如何收集的?
- 在哪些地点?
- 在什么时间段?
- 谁收集的?
- 用什么设备收集的?
- 中途操作人员和设备被更换过吗?

3. 数据含义

- 数据中的某个数值代表了什么含义?
- 它测量了什么?
- 它是否测量要测量的?
- 哪些环节可能会出差错?
- 在哪些统计假设下可以认为数据收集没有问题?

4. 相关性

- 收集来的数据能完全或部分地回答要研究的问题吗?
- 如果不能，还需要收集什么其他数据?
- 第 2 个问题中提到的要点在此处同样受用。

5. 问题转化

- 如何将 1. 数据收集中的问题转化成一个数据相关的统计问题，使之能够很好回答与原始问题呢?
- 有多种转换方式吗？比如，我们可以把问题转换成一个与统计模型有关的预测问题或者统计推断问题吗?

❸ 在选择模型前，列出将每一种能解决与实质性问题的转化方式的优点和缺点。

6. 可比性

- 各数据单元是否是可比的，或经过标准化处理而可视为可交换的?
- 苹果和橘子是否被组合在一起了?
- 数据单元是否相互独立?
- 两列数据是不是同一个变量的副本?

7. 可视化

- 询问数据范围是什么?
- 数据正常吗？是否有缺失值?
- 多使用颜色和动态图，注意有意料之外的情况记住，我们大脑皮层的 30% 都是用来处理图像的，所以可视化在挖掘数据模式和特殊情况时非常有效。
- 通常情况，为了找到大数据的模式，可视化在建立某些模型之后使用最有用，比如，计算残差并进行可视化展示。

8. 随机性

- 统计推断的概念，比如 p 值和置信区间，都依赖于随机性。那数据中的随机性是什么含义呢?
- 我们要对统计模型的随机性尽量明确地定义。
- 哪些所研究的领域中知识支持所用统计模型中的随机性的描述?
- 一个表现统计模型中随机性的最好例子，就是因果关系分析中 Neyman-Rubin 的随机分组原理（在 AB 检验中也有使用)。

9. 稳定性

- 你会使用哪些现有的方法?
- 不同的方法会得出同一个定性的结论吗?
- 可重复性研究最近在科学界中吸引了很多注意,《Science》的主编 Marcia McNutt 指出“**实验再现是科学家用以增加结论信度的一种重要方法**”。

10. 结果验证

- 人们怎样能知道数据分析是不是做的好呢?
- 衡量标准是什么?
- 可以考虑用其他类型的数据或者先验知识来衡量有效性，不过可能需要收集新的数据以确认结果的有效程度。

有效统计实践的十项简明原则

“十项简则”系列

- Ten Simple Rules” series 的创始人和长期作者 Phil Bourne 建议一些统计学家写一篇关于统计学“十项简则“的文章。
- 目标群体：**一定统计知识，并且有可能得到周围统计学家的帮助，或者有亲力亲为的态度并在电脑里已经安装了一些统计软件。**
- 提及的原则是作者从合作研究与教学经验，以及不止一次的令人沮丧的求助：**“麻烦看一下我学生的毕业论文/我的基金申请/审稿人的意见：这需要再加点统计内容，但还要看上去简洁明了。”**中总结出来的。

原则 1：统计方法应使得数据能够解决科学问题}

- 初级统计使用者和统计专家之间的巨大差别在他们思考如何利用手头的数据时就显露出来。
- 初级使用者往往潜意识里就已经默认数据和所研究的科学问题之间存在联系，然后直接考虑该用哪种方法对数据进行操作，而不是思考研究目标。
- 在充分了解这些问题后，统计专家会和他们的合作者讨论**数据如何能解决问题以及哪种方法是最有效的。**

原则 2：信号与噪音共存

- 刻画变异性是统计学的重要课题之一。变异性以各种形式存在。
- 比如当我们三次测量同一物体却取得三种不同数值时。这种变异性往往被叫做“噪音”，因为它既不能被解释也被认为与研究无关。
- 统计分析的目标就是在存在噪音和无关的变异性的情况下评估数据中的信号以及研究者感兴趣的变异性。

原则 3：提前计划，越早越好

- 类似“实验预期的理想结果是什么？你该如何解释它？”这样的问题作为出发点
- 在设计阶段提出问题可以减轻在分析阶段的头痛程度：谨慎的数据收集能简化分析并使之更加严谨。
- Ronald Fisher 爵士所说：“在实验结束后去咨询统计学家就好比让他去做尸检一样，他可能只能告诉你实验是因为什么原因而失败的”

原则 4：关心数据质量

- 在数据分析时，“无用的输入导致无用的输出”。
- 然而现代数据收集的复杂性导致其需要许多关于技术功能的假设，通常包括数据预处理技术。
- 即使数据已经过预处理，但是在分析前往往还需要做很多努力；这些通常被叫做“数据清洗”，“数据修整”，“数据刨削”。
- 测量单位也应该被充分理解并被一致记录。缺失数据值能被相关软件正确识别是非常必要的。
- 数据被转换为便利的格式后，我们应该先好好地观察一番。这也被叫做探索性数据分析 (EDA)，它通常是分析中提供信息最多的环节。

原则 5：统计分析并不只是一系列数字计算}

- 统计软件提供了帮助分析的工具，而不是定义了分析。
- 问题的科学背景是至关重要的，同时分析的关键就是把分析方法和科学问题紧密联系起来。
- 将分析中的某些步骤转换为结构化的算法，这样会对你自己和有相似或相同数据的其他研究者今后重复这种分析很有帮助。

原则 6：保持简约

- 在其他条件相同时，简约远胜于复杂，它有着不同的称谓：“奥卡姆剃刀”，”KISS 原则”（Keep it simple, Stupid），“少即是多”和“至繁归于至简”。
- 简单的模型能帮助我们从复杂的现象中建立规律，并方便我们与同行以及更广泛的世界的交流。

原则 7：对变异性进行评估

- 几乎所有的生物学**测量**在重复时都会展现明显的差异，造成基于数据计算出的所有结果都具有**变异性**。
- 统计分析的一个基本目标就是帮助研究者**评估变异性**，通常以**标准差或置信区间**的形式体现。
- 统计建模和推断的一个最大的成功就是用**估计参数的数据来估计误差**。
- 当报告结果时，提供一些变异性的说明是非常必要的。一个常见的错误就是计算标准误差时忘记考虑数据或变量间的相关性，这样通常会低估真实的变异性。

原则 8：检查你的假设

- 任何统计推断都需要假设。这些假设是基于已有知识和对数据变异性的概率表示—后者我们叫做统计模型—而提出的。
- 常见的统计模型都会用到线性关系的假设。然而在给定数据时，线性模型的合适程度是一个实际中经常碰到的问题，因此需要被仔细研究。
- 统计中令人头痛且很常见的假设就是数据中的不同观测结果是**统计独立**的。
- 充分理解你所采用的方法中蕴含的假设和尽力去评估这些假设是至关重要的。
- 至少你会想要检查统计模型与数据的拟合程度。可视化展示和拟合残差图对评价假设的成立和模型的拟合是非常有益的。

原则 9：尽一切可能的去重复

- 一个好的分析师会仔细地检查数据，寻找各种类型的模式并搜寻可预测与不可预测的结果。
- 当统计推断太关注于数据时，例如 p 值，他们就会失去通常的解释性。忽略这个事实是不诚实的：就好像是把靶的红心画在你射的箭的落点附近。
- 近来有大量的批评是关于在科学研究中 p 值的使用，其主要是针对 “如果 p 值小于 0.05 则结果不足以发表” 这一误解。
- 统计学家往往知道最明显的数据窥探的方法，例如选择特定的变量来做分析报告，以及一系列在该情况下调整结果的方法。
- 理想情况下，重复实验应该由独立的研究员进行。

原则 10：使你的分析可再现

- 可再现性（reproducibility）：在给定相同数据和对分析细节的完整描述，再现那些结果中的表格，图像和统计推断应该是可能的。
- 有一些方法可以极大的提高结果的可再现能力：将一些步骤系统化（原则 5），将用于生成结果的数据和代码共享。

总结

- 马克·吐温有句广为流传的话："**有三种谎言：谎言，可恶的谎言，以及统计。**"
- 研究者所面临的一个核心且常见的课题就是揭示数据所能告诉我们关于要研究的科学问题的信息。**统计学**可以看作为了帮助这个过程而**构建的语言**，**概率论**则是其**语法**。
- 在众多报告美国统计协会批判 p 值的文章中，尤其喜欢生物统计学家 Andrew Vickers 的一句话："**将统计学看做一门科学，而不是一份菜谱。**"这是 0 号原则的最佳候选。

- 在终极的分析中，一切知识都是历史；
- 在抽象的意义下，一切科学都是数学；
- 在理性的世界里，所有的判断都是统计学
- 摘自 C.R.Rao《统计与真理》

谢　谢!