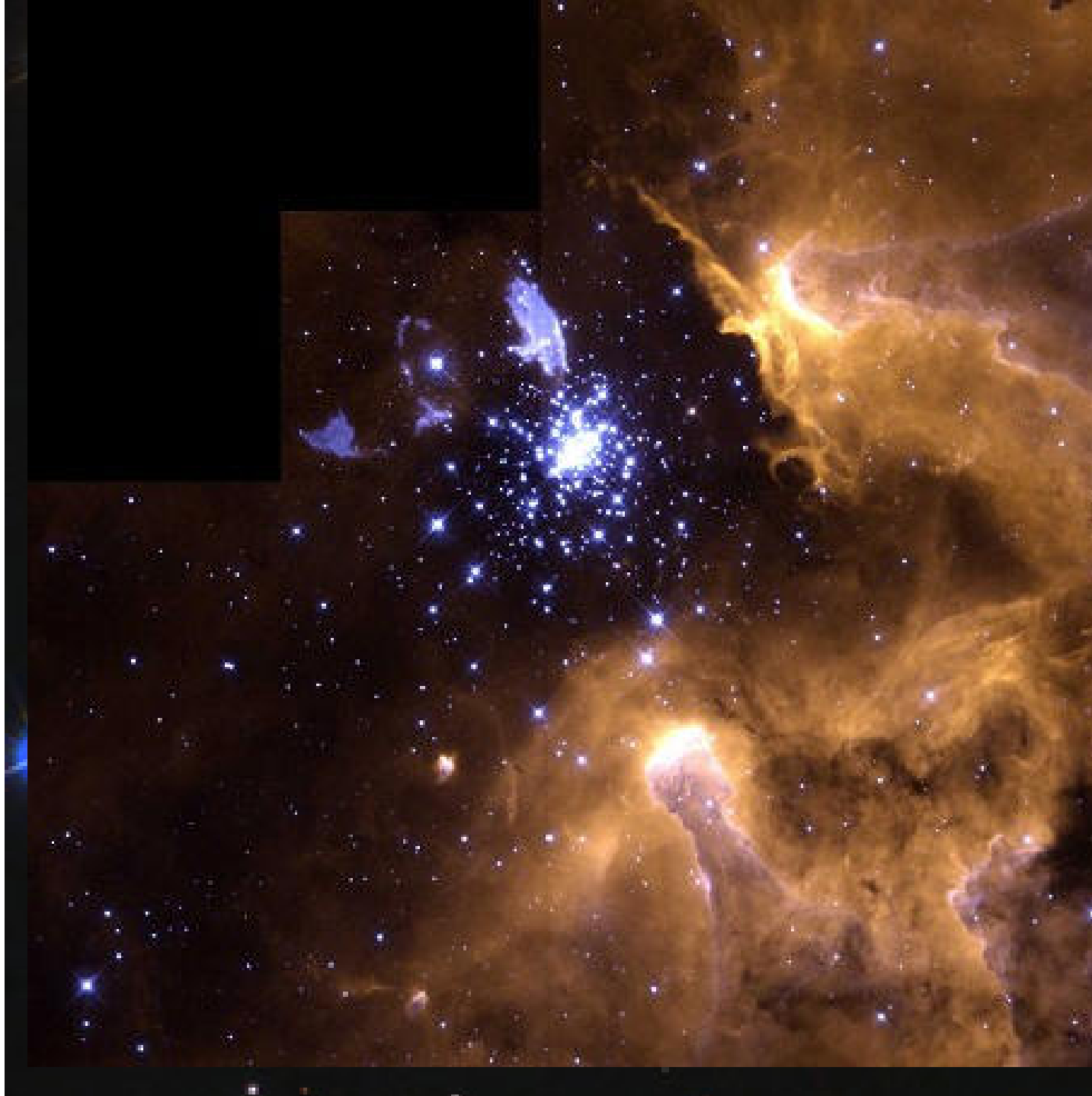
统计面临的挑战

吴喜之



科学与统计

统计的地位

- 统计在人类生活的各个方面所起的**重大作用**无可置疑
- 当然，很多人不知道这一点。
- 还有一个问题:什么是真正的统计?

统计的地位

- 在美国统计早已经取代计算机，成为**最容易**找工作的专业
- 美国普通公众对统计有着**过分**的崇拜
- 而中国数学类学生赴美留学的**首选专业**也是统计
- 在美国，大量学物理、计算机、电子等专业的人**改行学统计**

那么，什么是统计呢？

STATISTICS

– **the science of collecting, analyzing, presenting, and interpreting data.**

统计

- 统计方法就是科学的方法.
- 什么是科学和科学的方法呢？
- （面对需要、收集数据、根据数据建立模型、利用模型做预测或得到其它结论、模型则根据新的信息进行更新）

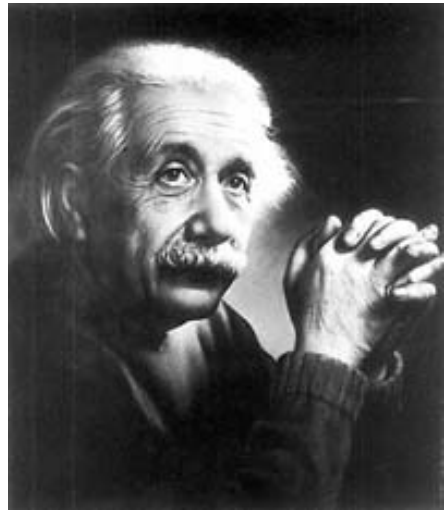
科学的方法

- 对世界的认识源于获得的信息或数据。
- 总结信息时会形成模型(假说或理论)
- 模型会指导进一步的探索……，直到遇到这些模型无法解释的现象。这就导致对这些模型的更新和替代。
- 这就是科学的方法。用科学方法进行的探索才叫科学。

例：天文学

- 公元2世纪托勒玫宇宙地心说
- 1543年哥白尼阐明了日心说
- 开普勒发现行星运动原理→伽利略把望远镜用于天文观测→牛顿又建立了运动和万有引力定律→赖特在1750年提出宇宙是由众多星系构成→18世纪末，赫歇尔首先进行了巡天观测，奠定了现代恒星天文学的基础。

例:牛顿→爱因斯坦



- 牛顿建立了运动定律和万有引力定律，可解释相当大部分人们周围所观测到的现象。
- 后来在亚原子尺度上、在行星观测中出现牛顿的惯性定律或万有引力定律无法解释的现象。这就导致了爱因斯坦狭义和广义相对论的产生。
- 又出现和相对论矛盾的现象，将会促进对相对论的修正。

科学方法的步骤

- 科学方法是目前已知的筛去谎言和错觉的最好方式。科学方法的步骤可做如下大致的描述：

1. 观测宇宙的某些方面。
2. 发明或提出可以解释这些观测的假说或假设，它必须和观测结果是相容的。
3. 利用该假说进行预测。
4. 用实验来检验这些预测（证伪），或者做进一步观测并根据结果修正假说。
5. 重复第3、4步直到在理论和实验或观测中没有矛盾为止。

理论

- 能够说明很多现象的假说可称为理论。
- 但任何理论都不能达到绝对的真理。
- 科学理论都应该是可证伪的(falsifiable)。应该存在某种实验或可能的发现可能证明理论是不对的。
- 科学是在证伪中发展的
- 基于不能重复观测或重复实验的现象而产生的许多说法，都不是科学，最多是信仰。
- 神的存在是无法证伪的．宗教不是科学，而是信仰。

科学是靠证据说话

- 理论适用与否靠实验或观测，不能靠辩论
- 古希腊的伟大哲学家亚里士多德用各种理由辩论说男人和女人的牙齿数目不同。
- 基于含糊不清或者不适当的前提的逻辑推理是没有多大意义的。

科学研究必需是毫无偏见的

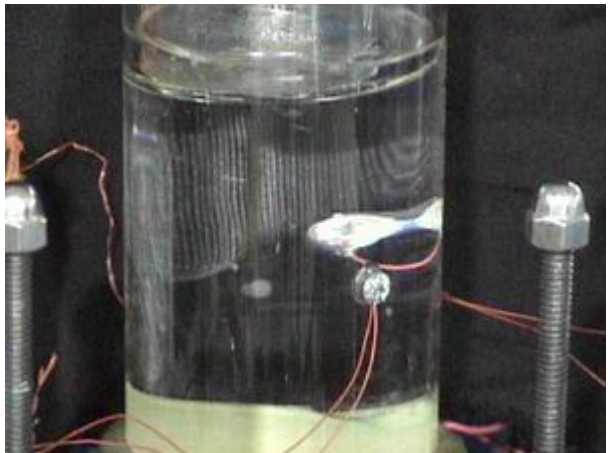
- 科学的结论应该独立于研究人员的文化背景、社会背景、种族、习惯、宗教和政治信仰等因素。

科学领域的造假

存在制造假的研究结果的现象。

但除非造假者的结论没有多大意义，总是会被人发现的。

如1989美国犹他大学的彭斯和英国南安普敦大学的弗莱什曼冷核聚变以及韩国科学家黄禹锡克隆胚胎干细胞的例子。

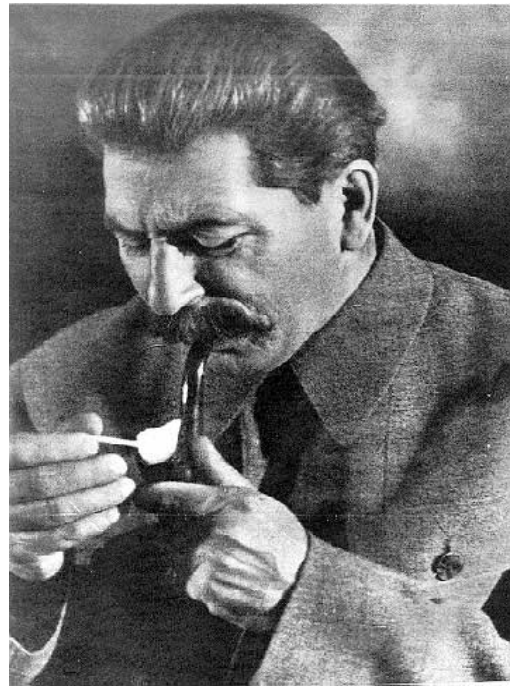


权力、宗教和意识形态对科学造成严重干扰

- 拥护哥白尼的“天体运行论”的布鲁诺被罗马教廷以“异端分子和异端分子的老师”的罪名，于1600年2月17日被烧死在罗马鲜花广场。
- 加利略由支持日心说于1633年被罗马天主教廷判决软禁，他在软禁中度过余生；结果使得地中海地区的科学传统完全停止了。

权力、宗教和意识形态科学造成严重干扰

- 在1930－60年代，苏联的全苏列宁农业科学院院长李森科把孟德尔和摩尔根遗传学斥为资产阶级的异端邪说，并在斯大林的支持下对苏联的研究基因的学者实行人身迫害。此事也对中国遗传学界产生了恶劣影响。



统计学是所有学科的工具
统计学方法是科学的方法

统计应该是一门科学

由于历史和国情，在很长一段时期中，这里所说的统计学在苏联和在我国被官方认为是资本主义的，同时我国一些与“官方观点”不一致的统计学家受到持续的批判。比如中国人民大学留美教授戴世光。

统计的应用

统计学与各个学科的数据都打交道；统计学实际上已经应用于所有领域。作为例子，它们包括：精算，农业，动物学，人类学，考古学，审计学，晶体学，人口统计学，牙医学，生态学，经济计量学，教育学，选举预测和策划，工程，流行病学，金融，水产渔业研究，遗传学，地理学，地质学，历史研究，人类遗传学，水文学，工业，法律，语言学，文学，劳动力计划，管理科学，市场营销学，医学诊断，气象学，军事科学，核材料安全管理，眼科学，制药学，物理学，政治学，心理学，心理物理学，质量控制，宗教研究，社会学，调查抽样，分类学，气象改善，遥感，搏采，等等。

当今，任何领域的研究结果，如果没有根据数据所作出的结论，很难被认可的。

中国统计中的伪科学

- 中国统计过去(现在?)分为“统计学”(文科的“列宁主义”统计，即现在所谓“社会经济统计学”)和“数理统计”(国际意义上的统计)

- 由于国情，国人对统计的尊重远远不如任何其他国家的人(可能北朝鲜除外)，往往误解统计学；
- 根据前苏联传统，国内一些学者把统计称为是经济学科的一部分，
- 这种经济学中的苏联式统计学的数学水平低于小学数学水平。
- 与现代经济学所需的大量的统计和数学形成鲜明对照.

- 前苏联式的“统计学”目前即使在俄国也无人问津
- 但其八股形式在中国仍然流行；而且存在于在官方的统一考试中☹

什么是有用的统计？

有用：在市场经济下找得到工作

美国统计学及其相关专业学位的学校研究方向与学历分布

层次 / 专业	学校数量	本科	硕士	博士
统计学	156	99	146	97
生物统计学	53	2	51	46
商业统计学	15	7	9	9
其他	60	20	43	36

数学的重要性

- 真正严格的逻辑仅存在于数学之中，只能够从学习数学中获得。
- 数学的逻辑服务于现代理性社会的所有方面。

统计和数学的思维方式差异

- 数学思维是以演绎为主
- 统计思维是以归纳为主，兼有演绎

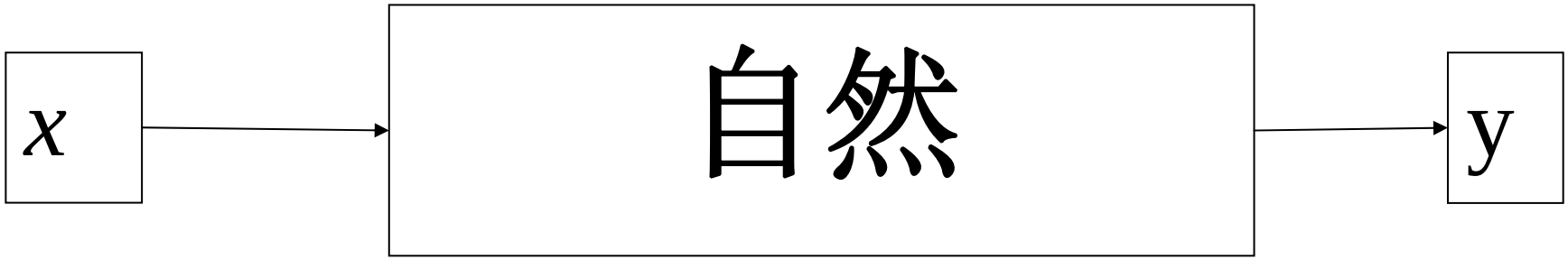
统计主要需要……

- 数学、计算机及研究对象领域的知识
- 加上想象力、通常的逻辑推理和常识判断的能力。

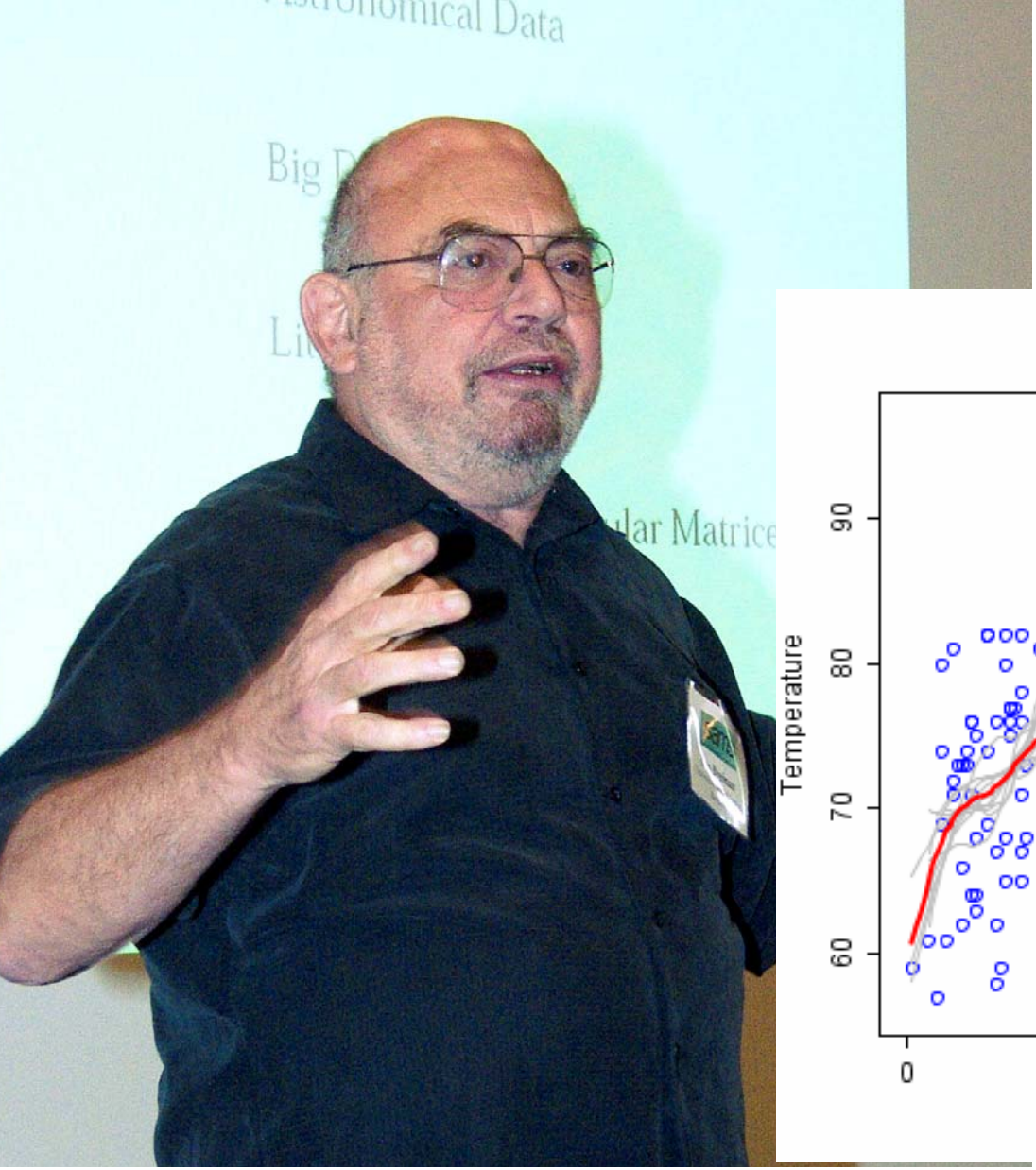
统计面对的挑战

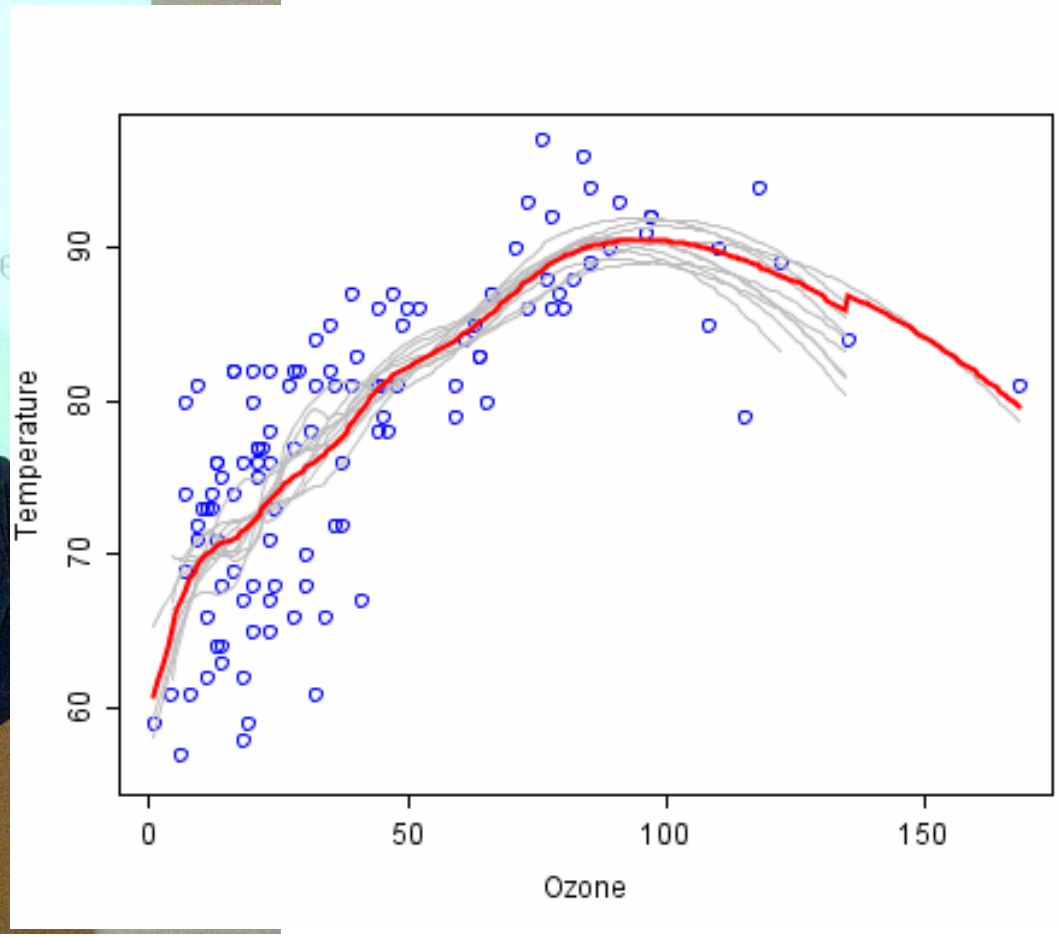
统计所研究的对象中的许多关系，很难被诸如物理定律那样的理论明确描述，被认为具有某种随机性，类似于黑匣子

- 记输入的数据为*x*，而输出为*y*；那么根据*x*产生出*y*的过程则可以用如下图形描述。



- 一般来说统计数据分析有两个目的。
- 一个是能够由输入数据*x*来预测*y*。
- 而另一个为解释这个联系输入变量和输出变量的“自然”部分，即所谓的“黑匣子”。





eo Breiman (January 27, 1928 – July 7, 2005) was a distinguished statistician at the niversity of California, Berkeley. He was the recipient of numerous honors and wards, and was a member of the United States National Academy of Science.

- 按照Breiman(2001)[1]的说法，统计有两个文化。一个是**数据建模**文化（data modeling culture），它在黑匣子中假定一个随机产生数据的模型（最典型的包括线性回归模型、logistic回归模型和Cox模型等等）。
- 这里对模型是否适当采用诸如拟合优度检验和残差分析等方法来确定。而模型通常为下面的函数形式：
  响应变量=*f*(预测变量，参数，随机噪声)或
  Y=f(X, θ, ε)

- 而Breiman所说的另一种为**算法建模**文化(algorithmic modeling culture)。
- 它也是找一个函数$f(x)$来预测y。
- 只不过这里的函数不局限于一些明确表达的数学公式，而是一个算法。
- 这里主要关心的是预测；而黑匣子到底是什么，能够解释就解释，但并不强求。

- 典型的算法包含决策树、关联规则、随机森林、支持向量机等等。
- 这里对模型是否适当，则采用预测精度来衡量。
- Breiman认为，专注于数据模型会产生无关的理论以及有问题的结论，使得统计学家远离适当的算法模型，不去研究崭新的实际问题。

- 多数专业统计学家属于数学出身。
- 他们认为“数理统计学只是从数量表现的层面上来分析问题，完全不触及问题的专业内涵。”
- 在这个意义上，“数理统计方法是一个中立性的工具。这‘中立’的含义是。它既不在任何问题上有何主张，也不维护任何利益或在任何学科中坚持任何学理。
- 作为一个工具，谁都可以使用。如果谁不同意这种方法，可以不使用[1]。”

- 对于统计方法或统计模型本身的这种在各学科中的“中立性”是普遍同意的。
- 但是，任何统计方法的发展、任何模型的建立都有其应用背景。
- 统计学家的研究，就其本质来说，是不可能独立于这些领域的具体目标，除非他们所做的工作是统计推断中间的一个局部数学环节的演绎式推导。

- 按照Breiman，数据建模文化包含了目前统计课程所涉及的大部份统计模型。
- 建立这些模型需要一些在实际中不一定能够满足的数学假定；
- 在模型选择、对结果的解释和预测等方面有很多不明确或不清楚的地方。
- 这些模型的使用对于非统计领域的人员来说并不方便。

- 而算法建模文化，则针对实际课题的问题，选择一些方法，利用计算机来根据训练样本建模。
- 人们用对测试样本的预测精度来判断这些模型是否适用。
- 由于没有多少中间的人为干预，Breiman觉得，这种文化是其他领域的工作者容易掌握的。

如果脱离应用背景而把统计作为纯粹数学的一部分，统计学没有存在的必要。

- 第一，统计学的方法都是在应用的推动下产生的，如果没有应用，它们不会出现。
- 其次，如果以应用为目的而产生的统计方法不能满足应用的要求，再漂亮的数学表达也不能保证其存在；
- 第三，统计中的数学本身不能形成一个完整的逻辑体系（贝叶斯统计可能被认为是例外），其中有大量的人为或主观因素在起作用；这是不符合纯粹数学的本质的。

统计从数学继承了什么？

- 统计应用最初是由政府的需要而产生的；但目前统计的方法和理论基础是由一批数学家奠定的。
- 很多人认为统计学是“数学的一个分支”。
- 这当然不仅涉及统计和数学的定义，而且涉及统计的性质和应用背景。

- 由于统计发展历史中的数学背景，上个世纪中期基本定型的数理统计教科书充满了数学味极强的定义、引理、定理、推论，以及贯串其中的纯粹数学推导和证明。
- 数学是一个“是非明确”的理想世界，它自我形成严格的封闭逻辑体系；只要逻辑正确，数学研究最多得不出结果，但不会犯错误。
- 这也是以演绎为主的数学魅力之所在。数学教科书没有负面的内容。数学的逻辑完全是客观的。

- 但以归纳为主要思维方式的统计是描述现实世界的，是为各领域服务的。
- 统计需要建立各种数学模型来近似现实世界；但任何数学模型都不可能精确地描述现实世界或自然；正如没有科学理论能够等于真理一样。
- 数学是不能证伪的；而统计和其他科学的理论一样，必须是可以证伪的。

- 基本上由数学老师教授的数理统计课程多是按照纯粹数学的模式设计的，
- 对于背后的基于数据的统计思想介绍得不很充分，也不强调这些充满假定的数学模型都是对现实世界的不同程度的简化。很少教科书指出违背这些假定的后果。
- 几乎没有人告诉学生，所有统计教科书中对数据（或其总体）的数学假定都是无法用数据验证的。
- 数学化的统计教科书极少提到统计应用中一系列决策的主观性和任意性。

传统的数据建模在应用中所遇到的问题

- 所有模型都仅仅是对现实世界的某种近似。
- 模型存在的一个必要条件是它们必须能够被人们解出来，无论是近似的，或者是精确的。
- 任何可得到的结论由于模型的近似性而必然是近似的。
- 而这些结果到底和现实世界有多么近似，可能永远不清楚。

- 衡量模型是否合适或者统计结果是否合理的传统方法包括各种拟合优度检验、准则，以及残差分析等等，当然还采用无偏性等大样本或总体概念。
- 正如Efron(2001)[1]指出的，二十世纪的统计可标以“100年的无偏性”，“大多数我们的统计理论和实践是围着无偏或几乎无偏估计（特别是MLES）和基于这样估计的检验转的。”

- 然而，要使用这些判别方法，必须对模型和产生数据的总体做出一些假定，诸如模型的数学形式、误差的结构和分布的假定。这些假定是基于经验、数据的特征，或数学上的方便。
- 然而，Bickel *et al* (2001)[2]表明除非备选假设有明确的方向，拟合优度检验的效率很低。
- 而残差分析也是不可靠的；它在变量数目多的时候无法揭示欠缺的拟合。不同的残差分析方法会导致不同的结论。

- 虽然拟合优度检验和残差分析可能会误导，但是正如Breiman（2001）[3]所说，近年来在*JASA*发表的关于数据的应用文章连这些方法也很少利用。
- 似乎和独创性的统计模型相比，模型拟合好坏是次要的。
- 只欣赏模型本身，而忽略实际应用背景是危险的。当结论仅仅描述模型的机制而不反映模型应该反映的现实世界时，结论必然是错误的。

- Mostelling & Tukey(1977)[4]在讨论回归的谬误时说：整个按部就班的回归领域充满着智力的、统计的、计算的和主题的困难。
- 很难想象我们面对着从包含未知的物理、化学、生物或社会机制的复杂系统中产生的未受控制的观测数据背后的机制能够被一些统计学家主观选择的参数模型来充分解释。而从这样模型得到的结论不能由拟合优度检验和残差分析来证实。

- 传统统计方法的另一个问题是数据建模的结果的多重性。也就是说，若干模型都显著，但它们对现实世界有不同的描述。
- 这些不同、但又都“显著”的模型对黑匣子的解释各异。
- Mountain & Hsiao (1989)[1]表明，很难构造一个能够包含所有竞争模型的复杂模型。而且，鉴于利用有限的样本所建立的依赖于渐近理论的各种检验的合法性和效率，所导致的结论是靠不住的。

算法建模

- 和传统的所谓数据建模文化不同，Breiman所定义的算法建模文化则多数由没有传统统计背景的研究人员所发展。
- 早在1980年代，算法建模在心理计量学、社会科学、医学中就有不同程度的应用。但最有影响的是80年代中期出现的神经网络和决策树。

- 这些方法的目的是提高预测的精度。最初的研究人员由年轻的计算机科学家、物理学家、工程师和少数统计学家。
- 他们在数据模型无法使用的复杂预测问题上试验他们的新的方法。
- 这些问题包括语言识别、图象识别、非线性时间序列预测、笔迹识别、以及金融市场的预测。

- 算法建模的势力迅速扩展，并且产生了数千篇文章。
- 最初的算法建模的研究人员多数没有传统统计训练，或者不受传统统计的约束；现在也有一些著名的统计学家加入了他们的行列。
- 他们的问题除了传统统计无法用武的领域，比如处理由遥感卫星、互联网、光学和射电天文望远镜、基因研究等产生的海量数据之外，也进入了传统的数据建模的领地。

- 目前的算法建模方法对于模型的评价主要是预测精度，比如利用试验数据集来对训练数据集所建立的模型进行交叉验证。
- 他们的方法也逐步改进，比如支持向量机就比早期的神经网络更有效，助推法（boosting）或其改进型进行分类和回归的方法也在不断改进。
- 这些方法许多在机器学习、人工智能或数据挖掘等各种名称下产生和发展。

- 算法建模和传统统计不仅仅区别于前面所说的着重于预测精度和适用于海量数据，它还有其他一些优点。
- 比如在基因数据中，变量个数可以达到4682个，而样本量仅有81个(参见Dudoit et al., 2000[1])。
- 这样巨大的变量和观测值数目的比例是传统统计不可想象的。比如，Diaconis & Efron(1983)[2]年曾经说过，“统计经验表明，基于19个变量和仅仅155个数据点来拟合模型是不明智的。”

- 它不仅不畏惧巨大的维数，而且认为变量越多，包含的信息越多。实际上，有大量的信息在各种预测变量的组合之中。
- 算法建模文化不仅不减少维数，而且在预测变量中增加许多变量。
- 此外，前面所说的数据建模文化所无法解决的模型多重性问题在算法建模文化中也是有利的，它可以把大量的竞争模型整合起来增加预测的精度。

- 高精度总是与数据背后的机制更可靠的信息相关联的。因此，算法模型比数据模型提供更好的预测精度，也提供了关于数据背后机制的较好的信息。
- 此外，应用算法模型需要较少的专业知识和专家干预，对于各领域的工作者来说，更易于掌握和理解。

- 尽管算法建模有如此广泛的应用和优势，但是，由于算法建模文化的研究成果基本上没有传统统计所固有的总体分布假定、假设检验、参数估计等标志性因素。
- 这些成果多数发表在工程、计算机及其他非统计应用领域的期刊上。
- 人们可能会问，按照（比如不列颠百科全书的）统计为“收集、分析、展示和解释数据的科学”的定义，难道这些算法建模不属于统计吗？

- 实际上，在统计学的社区中，统计的定义是由各个统计系研究生课程的内容来确定的，是由统计杂志的文章范围来确定的。
- 当然，统计系的课程目录是由受过传统统计训练的教授确定的；而统计杂志的内容是被该杂志的主编和编辑来决定的。

- 在这种自我贴标签和自我约束下，统计界过去更多地聚焦在模型形式本身，而不是作为建模目的之所在的实际问题上。
- 统计也因此失去了大量活力、创造力和领地。上面所提到的统计的形式定义从来没有被完全当真。
- 坚持数据模型的损害在于：统计学家把自己排除于有些最有意义和挑战性的统计问题之外，而许多有意义的结论最终由非统计学家来找到。

回到统计的最初宗旨

- 统计最初是为解决实际问题而产生的，现在，统计学必须重新回到它针对实际问题而与数据打交道，并且创造有关理论的传统。
- 为了解决实际问题，必须毫无偏见地接受任何有效的建模方法。无论是数据模型，还是算法模型，还是它们的结合都可能很好地解决面对的问题。
- 统计学家还需要与其他领域的科学家合作，共同工作。只有这样，我们才能应对新时代中不断产生的问题所带来的挑战。


谢谢大家