预测，估计和归因*

原作者：Bradley Efron[1]
译者：史宏炜[2]，郭旭[2]

[1] *Department of Statistics, Stanford University*
[2] 北京师范大学统计学院

摘要

20 世纪在科学需求和计算方面的限制塑造了经典的统计学方法论，但在 21 世纪，需求和限制都有所改变，因此方法论也相应地发生了变化。大规模的预测算法——神经网络、深度学习、boosting、支持向量机、随机森林等已经在大众传媒上获得了“明星地位”。它们被视作在巨大规模的数据集上运行的回归传统的继承者。这些算法与标准回归方法（例如最小二乘或逻辑回归）相比有什么区别？本文对几个关键的差异进行了探究，主要集中在“预测”和“估计”之间或者“预测”和“归因（显著性检验）”之间的差异。本文大部分的讨论是通过一些小的数值实例展示的。

关键词：黑箱，短期预测变量，随机森林，表面加噪声

1 引言

统计回归方法要追溯到 19 世纪初的 Gauss 和 Legendre 时期，特别是 1877 年的 Galton 时期。在 20 世纪期间，回归的思想被应用于众多重要的统计任务：新案例的预测、回归表面的估计以及对单个预测变量的显著性进行赋值（在文章的标题中称作“归因”）。在 20 世纪，大量强有力的统计思想都和回归有着密切的联系：最小二乘拟合、逻辑回归、广义线性模型、ANOVA、预测变量显著性检验以及均值回归。

21 世纪见证了一种新型的“纯预测算法”的崛起——神经网络、深度学习、boosting、支持向量机和随机森林，这些算法继承了 Guass-Galton 的传统，同时它们可以在规模巨大的数据集

*统计学大师 Bradley Efron 在统计学顶级期刊 Journal of the American Statistical Association 发表了题为 Prediction, Estimation, and Attribution (Efron et al., 2020) 的文章，并有 Friedman et al. (2020); Candès and Sabatti (2020); Xie and Zheng (2020) 等统计学大家的讨论评价，后文章进一步被转载在期刊 International Statistical Review。本文已得到原作者 Bradley Efron 的中文翻译许可。

上运行。这类数据集往往拥有数百万个数据点，甚至包含着比数据点还多的预测变量。这些算法（尤其是深度学习）在自动化任务方面非常成功，如在线购物、机器翻译和航线信息。它们凭借自身的优势成为了媒体的宠儿，引起了商业界极大的兴趣。最近这股热潮已经扩散到了科学领域，在浏览器中马上就可以检索到“生物学中的深度学习”、“计算语言学与深度学习”以及“调控基因组学的深度学习”等相关词条。

纯预测算法如何与传统的回归方法相关联呢？这是本文要探讨的中心问题。本文将探究一系列突出的差异，包括假定方面、科学理念以及目标方面的差异。讲清楚这件事是很复杂的，两者也并没有清晰的胜负之分；但是至少在我看来，可以给出一个粗略的总结：纯预测算法为统计学家的“武器库”增添了强大的力量，但为了它们常规的科学适用性，还需要开展进一步的实质性发展。这样的发展已经在统计界展开，并为我们的学科提供了可喜的活力。

本文最初是一个演讲，写作风格很宽泛，旨在描述当前的实践而不是强调事情必须是怎样的。本文假设读者没有关于各种预测算法的预先知识，虽然这将严重低估许多读者。

本文不是一篇研究性的论文，大部分的论证都是通过数值例子展开的。这些例子的规模很小，以现行的预测标准来衡量甚至是不值一提的。某种“巨型主义”已经占据了预测类的文献，会使用一些夸张的前缀，如 *tera-*，*peta-*和 *exa*。但是小的数据集可以更好地暴露出新方法的局限性。

Hastie et al. (2009) 是关于经典和现代预测方法的一个非常好的参考文献。本文关于纯预测算法机制的讨论很少：我希望这样就足以让读者了解它们与传统算法的根本不同之处。

2 表面加噪声模型

对于预测算法和传统的回归方法，本文假设统计学家可获取的数据集 $\mathbf{d}$ 具有以下结构：

$$\mathbf{d} = \{(x_i, y_i), i = 1, 2, \ldots, n\}; \tag{1}$$

其中 x_i 是 p 维的预测向量，x_i 来自一个已知空间 $\mathcal{X} \subset \mathbb{R}^p$，$y_i$ 是一个实值响应变量。假设 n 对数据是相互独立的，更简化的形式为

$$\mathbf{d} = \{\mathbf{x}, \mathbf{y}\}, \tag{2}$$

其中 $\mathbf{x}$ 是 $n \times p$ 的矩阵，x_i^t 是 $\mathbf{x}$ 的第 i 行，$\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$。假定最传统回归模型为“具有正态误差的最小二乘”，表示如下

$$y_i = x_i^t \beta + \epsilon_i \quad (i = 1, 2, \ldots, n), \tag{3}$$

其中 $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$，$\beta$ 是一个未知的 p 维参数向量。用矩阵记号写为

$$\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\epsilon}. \tag{4}$$

对任意的 $x \in \mathcal{X}$，模型 (3) 中响应变量 y 具有期望 $\mu = x^t\beta$，因此 $y \sim N(\mu, \sigma^2)$。线性表面 $\mathcal{S}_\beta$ 可以表示为

$$\mathcal{S}_\beta = \left\{\mu = x^t\beta, x \in \mathcal{X}\right\}, \tag{5}$$

其包含着所有真实的期望，但是真实情况会被噪声项 ϵ_i 所模糊。

更一般的是，本研究将模型 (3) 扩展为

$$y_i = s(x_i, \beta) + \epsilon_i \quad (i = 1, 2, \ldots, n), \tag{6}$$

其中 $s(x, \beta)$ 是某个已知的函数形式，对于任意固定的 β，所得的期望 $\mu = s(x, \beta)$ 是一个关于 $x \in \mathcal{X}$ 的函数。那么真实的期望表面，也就是*回归表面*可以表示为

$$\mathcal{S}_\beta = \{\mu = s(x, \beta), x \in \mathcal{X}\}. \tag{7}$$

大多数传统的回归模型依赖于某种*表面加噪声*的形式（尽管“加”可能指的是二项可变性）。表面描述的是我们希望学习到的科学真理，但是我们只能观察到表面上被噪声模糊的点。统计学家的传统估计任务就是是从数据 **d** 中尽可能多地了解表面的信息。

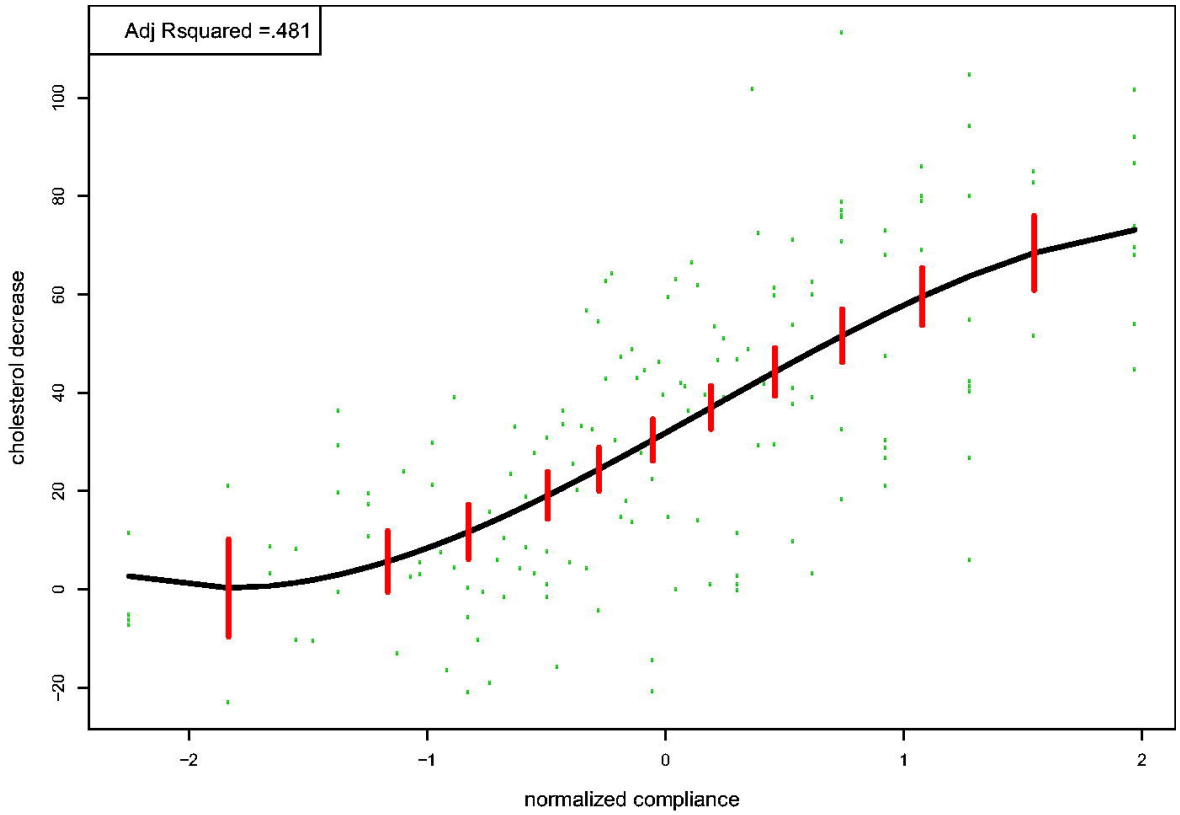

图 1: 黑色的曲线是对胆甾烯胺数据的 OLS 拟合回归；垂直的竖线表示 ± 一个标准误的估计。

图 1 展示了一个小例子, 数据集是从 Efron and Feldman (1991) 的大数据集中提取的：$n = 164$ 名男性医生志愿服用降胆固醇药物胆甾烯胺。对于每个医生记录以下两组数据

$$\begin{aligned} x_i &= \text{正态化后的依从性,} \\ y_i &= \text{观察到的胆固醇降低量.} \end{aligned} \tag{8}$$

依从性是指某个受试者实际服用的剂量占计划剂量的比例，范围从 0% 到 100%，正态化之后是从-2.25 到 1.97。本研究希望看到依从性越好，胆固醇降低的越多。

本研究利用这一数据拟合得到一个正态回归模型 (6)，其中 $s\left(x_i, \beta\right)$ 具有以下形式

$$s\left(x_i, \beta\right) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3. \tag{9}$$

也就是说这是一个三次回归模型，图 1 中黑色的曲线表示一个估计的表面，表示为

$$\widehat{\mathcal{S}} = \{s(x, \hat{\beta}), x \in \mathcal{X}\}, \tag{10}$$

这是通过极大似然的方法拟合得到的，也等价于通过最小二乘（OLS）方法进行拟合。垂直的竖线表示在选取的 11 个 x 点上，估计值 $s(x, \hat{\beta})$ 加减一个标准误，展示出使用 $\widehat{\mathcal{S}}$ 来估计真实的 $\mathcal{S}$ 的不准确程度。

这一例子是估计方面的内容。对于归因而言，只有 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是显著非零时才会被考虑。调整的 R^2 为 0.482，R^2 是模型预测能力的一种传统度量。

传统研究方法的另一个支柱是逻辑回归, 表 1 考虑了一组新生儿数据 (Mediratta et al., 2020)：在非洲的工厂中，$n = 812$ 个患病的新生儿经历了长达一年的观察，其中存活 605 个，死亡 207 个。新生儿数据集中记录了 11 个协变量，包括胎龄、体重和 Apgar 评分等，因此数据集 (1) 中的 x_i 是 11 维的，同时 y_i 等于 0 或 1 用来表示新生儿是存活还是死亡。这是一个具有线性对数表面以及伯努利噪声的表面加噪声模型。

对 11 个预测变量进行标准化，使其均值为 0，方差为 1，然后进行逻辑回归分析。表 1 展示了一些输出结果。第 1 列和第 2 列给出了回归系数的估计和标准误（这是对线性逻辑表面 $\widehat{\mathcal{S}}$ 的估计及其准确度的描述）。

第 3 列展示了 11 个变量的标准双边 p 值，有 6 个变量显著非零，其中有 5 个非常明显。这就是分析的归因部分。就预测而言，拟合逻辑回归模型给出了每个婴儿的死亡概率 p_i 的估计值。预测规则为

$$\begin{aligned} &\text{如果} \quad p_i > 0.25 \qquad \text{预测为死亡,} \\ &\text{如果} \quad p_i \leq 0.25 \qquad \text{预测为存活,} \end{aligned} \tag{11}$$

表 1: 新生儿数据的逻辑回归分析

	估计	SE	p 值
Intercept	–1.549	0.457	0.001***
gest	–0.474	0.163	0.004**
ap	–0.583	0.110	0.000***
bwei	–0.488	0.163	0.003**
resp	0.784	0.140	0.000***
cpap	0.271	0.122	0.027*
ment	1.105	0.271	0.000***
rate	–0.089	0.176	0.612
hr	0.013	0.108	0.905
head	0.103	0.111	0.355
gen	–0.001	0.109	0.994
temp	0.015	0.124	0.905

11 个预测变量中的 6 个具有显著的双边 p 值；逻辑回归的估计结果产生了 18% 的预测误差。

得到的经验误差率为 18%。（选择阈值 0.25 是为了补偿较小的死亡比例。）

所有这些都是熟悉的东西，用来提醒读者传统的回归分析是如何开始的：设定一个对潜在科学真相（“表面”）的描述，连同一个关于随机误差的模型。纯预测算法则遵循不同的路径，如第 3 节所述。

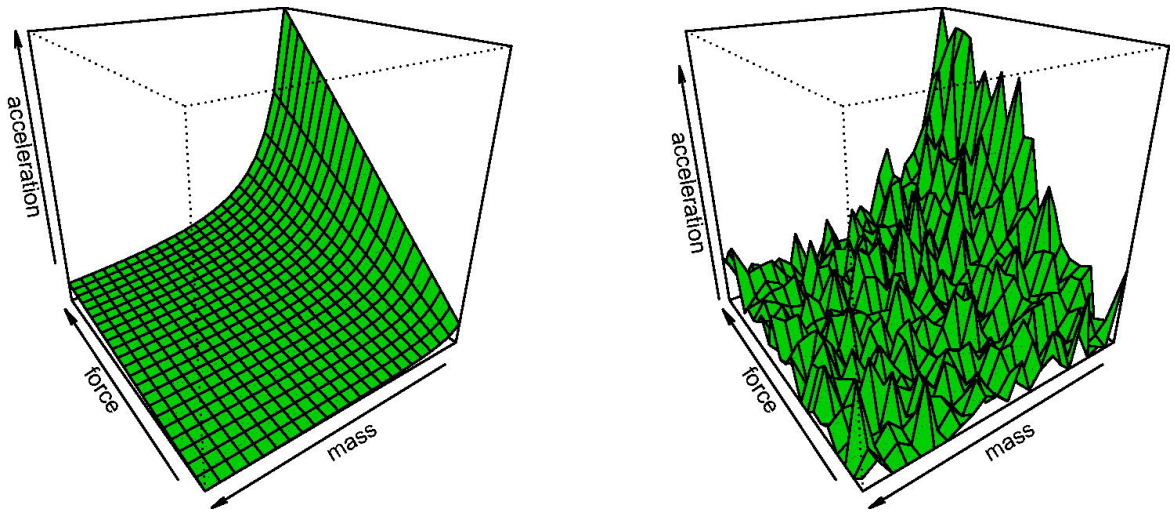

图 2: 左边是描述牛顿第二运动定律的曲面，加速度 = 力/质量；右边是一个添加噪声的版本。

图 2 的左边展示了牛顿第二定律的表面，

$$\text{加速度} = \frac{\text{力}}{\text{质量}}. \tag{12}$$

想象第二定律产生自牛顿的大脑是令人高兴的，但其实牛顿是一位实验大师。右边是一幅想象出来的实验数据可能的样子[1]。

在缺乏天才般的洞察力的情况下，统计的估计理论旨在作为一种工具，从嘈杂的数据中窥视并识别出平滑的潜在真相。胆甾烯胺和新生儿的例子都不像牛顿第二定律那么根本，但他们的共同目标都是在嘈杂的环境中提取到可靠的科学结构。噪声是短暂的，但结构是永恒的，或至少持久的（见第 8 节）。

3 纯预测算法

21 世纪[2]见证了一系列不同寻常的预测算法的发展：随机森林、梯度提升、支持向量机、神经网络（包括深度学习）以及一些其他算法。为了区别于上一节介绍的传统预测方法，我将把这些统称为“纯预测算法”。纯预测算法已经在多个方面取得了惊人的成绩，并引起了公众极大的兴趣，如：机器翻译、iPhone 的 Siri、面部识别、国际象棋锦标赛和围棋等多个项目。如果媒体关注是一种合适的度量标准，那么纯预测算法就是我们这个时代的统计明星。

之所以使用形容词“纯”，是因为预测算法的目标是预测，基本上不考虑估计和归因。他们的基本策略很简单：获得很高的预测准确度是其最直接的目的，而不考虑表面加噪声模型。因此预测算法具有一些突出的优势，也有一些缺点。优点和缺点将在文章之后的内容中说明。

一个预测算法的常规程序是：输入一个数据集 $\mathbf{d} = \{(x_i, y_i), i = 1, 2, \ldots, n\}$ (1)，输出一个准则 $f(x, \mathbf{d})$，对于任意的预测变量 x，基于这一准则可以得到一个预测值

$$\widehat{y} = f(x, \mathbf{d}). \tag{13}$$

我们希望该准则的明显错误率很小。对于分类问题而言，错误率为 $\hat{y}_i \neq y_i$ 的占比为

$$\widehat{\text{err}} = \#\left\{f\left(x_i, \mathbf{d}\right) \neq y_i\right\}/n. \tag{14}$$

更重要的是，我们希望真正错误率

$$\text{Err} = E\{f(X, \mathbf{d}) \neq Y\} \tag{15}$$

也很小，其中 (X, Y) 是从给定数据集 $\mathbf{d}$ 中的 (x_i, y_i) 所服从的概率分布中随机提取得到的；见第 6 节。随机森林、boosting、深度学习等算法都以能在复杂情况下得到较小的 Err 而得名。

纯预测算法除了和传统的预测方法有很大的不同以外，这些算法之间也存在着很大的差异。最容易去描述的算法是随机森林 (Breiman et al., 2001)。对于二分类预测问题，如新生儿数据，随机森林的预测结果依赖于分类树的总效果。

[1]早在半个世纪前，伽利略就因为使用斜面和水钟来估计落体加速度而闻名。
[2]实际上，在“漫长的 21 世纪”，大部分活动开始于 20 世纪 90 年代。

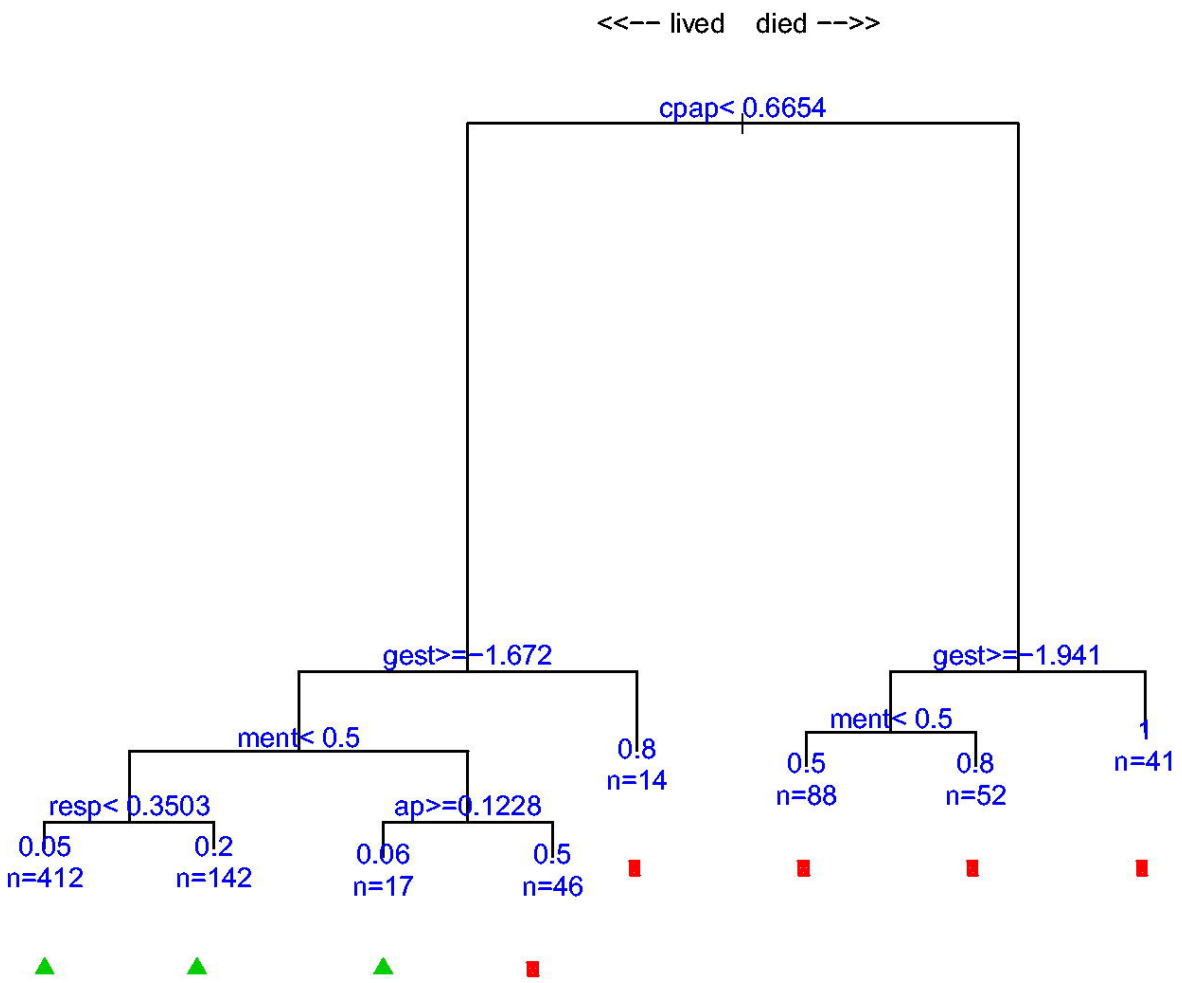

图 3: 新生儿数据的分类树。三角形终端节点预测婴儿为存活，方块预测婴儿为死亡；该规则的明显预测错误率为 17%，交叉验证错误率为 18%。

图 3 展示了将 R 程序 `Rpart3`[3]应用于新生儿数据获得的单个分类树。在树的顶端，812 个新生儿被分为两组：预测变量 cpap（气道阻塞标志物）小于阈值 0.6654 被置于左侧预断病情较好的组；cpap ≤ 0.6654 的新生儿患者被分配到右侧预断病情较差的组。选择预测变量 cpap 以及阈值 0.6654 是为了在所有可能的选择（预测变量、阈值）中最大限度地提高两个组中的死亡率差异[4]。接着按照相同的基尼系数准则，把两组各自再分为两组。分割过程将一直进行，直到触发了某种停止规则。

在图 3 的底端可以看到分割过程以 8 个*终端结点*结束：最左边的结点包含了最初 812 个新生儿中的 412 个，只有 5% 死亡；最右侧结点有 41 名新生儿，全部死亡。三角形表示死亡比

[3]一个实现 CART (Breiman et al., 1984) 的 R 语言程序包。

[4]更准确地来说：如果 n_{L} 和 n_{R} 表示左右两个组的样本大小，$\hat{p}_{\mathrm{L}}$ 和 $\hat{p}_{\mathrm{R}}$ 表示死亡占比，那么算法最小化了基尼准则 $n_{\mathrm{L}}\hat{p}_{\mathrm{L}}\left(1-\hat{p}_{\mathrm{L}}\right)+n_{\mathrm{R}}\hat{p}_{\mathrm{R}}\left(1-\hat{p}_{\mathrm{R}}\right)$。这个式子等价于 $n\hat{p}(1-\hat{p})-\left(n_{\mathrm{L}}n_{\mathrm{R}}/n\right)\left(\hat{p}_{\mathrm{L}}-\hat{p}_{\mathrm{R}}\right)^2$，其中 $n=n_{\mathrm{L}}+n_{\mathrm{R}}$ 以及 $\hat{p}=\left(n_{\mathrm{L}}\hat{p}_{\mathrm{L}}+n_{\mathrm{R}}\hat{p}_{\mathrm{R}}\right)/n$，因此对于所有给定的 n_{L} 和 n_{R}，最小化基尼准则等价于最大化 $\left(\hat{p}_{\mathrm{L}}-\hat{p}_{\mathrm{R}}\right)^2$。

例小于初始比例 25.5% 的三个终端结点，方块表示死亡比例超过 25.5% 的终端结点。预测规则是：三角形处代表存活，方块处代表死亡。如果一个新生儿带着 11 个测量值的向量 x 来到医院，医生可以根据 x 的取值沿着分类树从上到下来预测新生儿的生死。

如果将节点上观测到的比例，如 0.05 等视作是正确的，那么此分类准则得到的明显错误率为 17%。分类树通常被认为是贪婪的过拟合者，然而对于新生儿案例，10 折交叉验证分析给出的错误率为 18%，两者几乎相同。Mediratta et al. (2020) 对新生儿数据进行了传统分析，得到了 20% 的交叉验证错误率。值得一提的是，图 3 中的分割变量和表 1 中的显著性变量非常吻合。

到目前为止，回归树的表现还不错，但在一些规模更大的例子中，回归树的预测表现很差，见 Breiman et al. (2001) 的 9.2 节。作为一种改进，Breiman 提出的随机森林算法依赖于对大量 bootstrap 树进行平均，每棵树生成的方法如下：

1. 从原始数据 **d** 中提取一个非参数 bootstrap 样本 $\mathbf{d}^*$，即从 **d** 中有放回地抽取 n 对 (x_i, y_i) 的随机样本。

2. 像之前一样，利用样本 $\mathbf{d}^*$ 构建一个分类树，但仅使用从原始 p 个变量中独立随机抽取的 p^* 个预测变量来选择每次分割（$p^* \doteq \sqrt{p}$）。

生成 B 个这样的分类树之后，一个新的观测 x 在每棵树上得到一个分类结果；最终 $\widehat{y} = f(x, \mathbf{d})$ 是由 B 中的大多数分类结果决定的。通常 B 取值范围为 100-1000。

本研究使用 R 程序 `randomForest`，将随机森林算法应用到了新生儿的预测问题中，结果如图 4 所示。预测错误率[5]被表示为关于 bootstrap 树的个数的函数。总的来说，本研究共使用了 $B = 501$ 棵树，但在 200 棵树之后预测错误率的变化不大。总体预测错误率在 17% 左右波动，与图 3 中 18% 的交叉验证错误率相比只有一点点改进。在第 4 节的微阵列例子中，随机森林算法则显示出了更好的优势。

随机森林以 x 的 p 列作为预测变量，但随后通过分割过程生成一系列新的预测变量（如，“cpap 小于或大于 0.6654”）。新的变量给分析带来了高度的交互作用，例如，在图 3 中的 cpap 和 gest 之间就存在着高交互性。尽管不同的纯预测算法实现的方式不同，但高交互性以及可以产出丰富的预测变量是所有纯预测算法的特征。

[5]这些是预测误差的“出袋”估计值，是交叉验证的一种形式，在附录 A 中有解释。

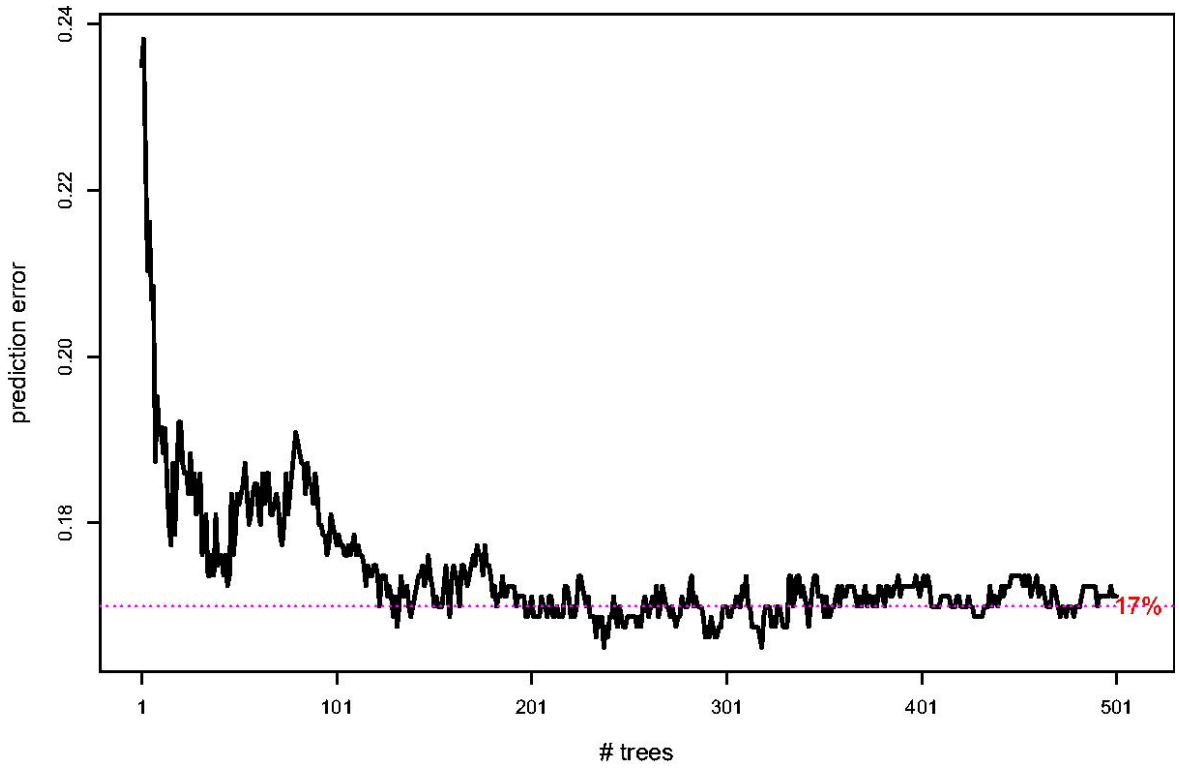

图 4: 新生儿数据的随机森林预测错误率，是一个关于 bootstrap 树的个数的函数；交叉验证错误率 17%。

4 一个微阵列的预测问题

纯预测算法的突破性进展涉及到了真正庞大的数据集。例如，谷歌上最初的英法翻译工具，就是用从加拿大和欧盟的立法记录中获得的数百万个英法平行翻译片段中进行训练的。本文目前没有提供这种规模的研究，但作为新生儿数据的一个小进步，我们将考虑前列腺癌的微阵列研究。

表 2: 在前列腺数据的 100 次训练/测试的随机分割中，随机森林测试集中的错误个数。

错误个数	0	1	2	3	4	5	7
频数	3	26	39	12	5	4	1

本研究中有 $n = 102$ 个男性，其中 52 个是癌症患者，50 个是正常控制组。本研究使用一组 $p = 6033$ 个基因来衡量每个男性的基因表达水平，

$$x_{ij} = \text{第 } i \text{ 个男性的第 } j \text{ 个基因的活性}, \tag{16}$$

其中 $i = 1, 2, \ldots, 102$ 以及 $j = 1, 2, \ldots, 6033$。在这种情况下，$n \times p$ 的矩阵 x 的宽度要比高度

大的多，学者通常将这种 $p \gg n$ 的情形称作“宽数据”，这与传统的 $p \ll n$ 的“高数据”形成了对比。

接下来利用随机森林算法对男性微阵列数据进行预测，判断是正常还是患癌。按照标准程序，102 名男性被随机分为大小均为 51 的训练集和测试集[6]，每个数据集中有 25 名正常控制组和 26 名癌症患者。

测试集 $\mathbf{d}_{\text{train}}$ 中包含了 51 组 (x, y) 数据对，其中 x 是一个包含 $p = 6033$ 个基因活性测量值的向量，y 等于 0 或 1 分别代表正常控制组或癌症患者。本研究利用 R 程序 `randomForest` 得到了预测准则 $f(x, \mathbf{d}_{\text{train}})$。然后将这一准则应用到测试集中，得到了 51 个测试对象的预测结果 $\hat{y}_i = f(x_i, \mathbf{d}_{\text{train}})$。

图 5 绘制了测试集错误率随着随机森林树的个数增加的情况。在 100 棵树之后，测试集的错误率为 2%。也就是说，$\hat{y}_i$ 和 y_i 的结果基本一致，实际结果表现为 51 个测试对象中有 50 个被预测正确：按照任何人的标准这都是一个出色的表现！但这并不是一个特别幸运的结果。随后进行了 100 次训练/测试的随机分割，每次都重复图 5 中的随机森林预测，并计算出测试集中预测错误的数量。如表 2 所示，预测错误个数的众数为 2，“1 个预测错误” 的情形经常发生。

一个分类树可以被认为是一个函数 $f(x)$，对于任意的 $x \in \mathcal{X}$，$f(x)$ 取值为 0 或 1。图 3 中的树将 11 维空间 $\mathcal{X}$ 划分为 8 个矩形区域，其中 3 个为 $y = 0$，5 个为 $y = 1$。如果在第一次分割之后就停止的话可以得到一个更简单的函数，这种情况下 $\mathcal{X}$ 会被划分为两个区域，cpap $<$ 0.6654 和 cpap $\geq$ 0.6654。这种简单的树被称为“树桩”。

由此引出了另一种著名的纯预测方法，*boosting*。图 6 展示了 R 语言程序 `gbm`（gradient boosting model）算法应用于前列腺癌症数据中的预测结果[7]。Gbm 连续地拟合了分类树的加权和，

$$\sum_{k=1}^{K} w_k f_k(x), \tag{17}$$

在第 $k+1$ 步中，选择树 $f_{k+1}(x)$ 使得可以最大程度地提升拟合效果。本研究需要保持较小的权重 w_k，以避免陷入一个错误的序列。在 400 步之后，图 6 的结果显示测试集上的错误率为 4%，即 51 个测试对象中有两个被错误预测，这也是一个出色的预测结果。(Hastie et al. (2009) 中的例子显示，gbm 通常比随机森林表现得更好。)

以 boosting 这种引人共鸣的语言来说，进入图 6 结构的树桩被认为是“弱学习者”：他们中的任何一个都能勉强将预测错误率降低到 50% 以下。无数的弱学习者能够如此有效地结合在一起，这是一个惊喜，也是纯预测算法的一大进步。相比之下，传统的方法更侧重于信号强度

[6]更常见的做法是选择 81 个数据进行训练以及 21 个进行测试，但为了后续的比较，更大的测试集是有利的。
[7]在应用时 $d = 1$，即拟合“树桩”，收缩系数设为 0.1

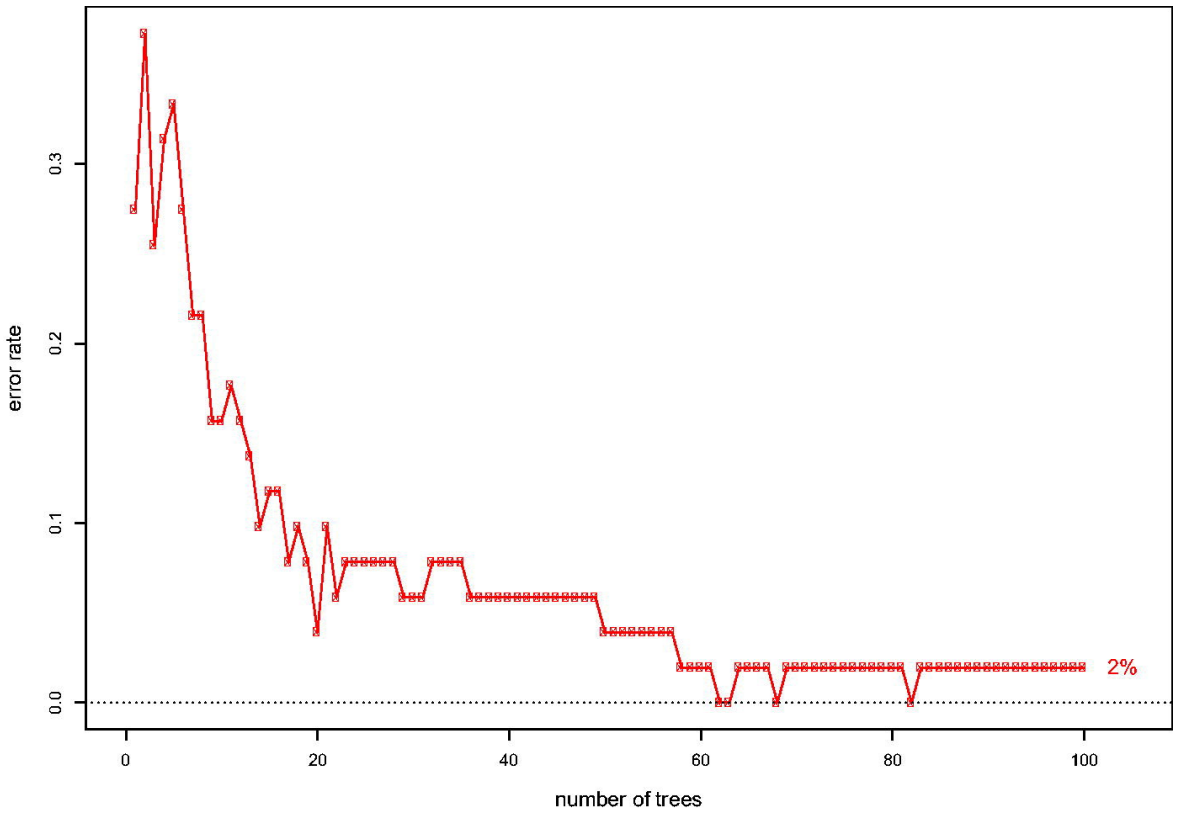

图 5: 将随机森林算法应用到前列腺癌微阵列研究的测试集错误率，是关于 bootstrap 树的个数的函数。

高的单个预测变量，如表 1 中带星号的变量，这是传统统计方法与纯预测算法之前的一个关键区别，将在后面的章节中讨论。

在图 6 中比较细的那条曲线表示在训练集上 gbm 准则的预测错误率。在第 86 步时训练集的预测错误率变为 0，但此时模型训练仍在继续，测试集的预测错误率有所改善。交叉验证计算给出了一些何时停止拟合的提示——这个例子中最好在第 200 步停止，但这不是一个确定性的问题。

本研究使用程序包 `keras` 将神经网络/深度学习算法应用于前列腺数据中。结果显示深度学习算法的预测效果要比随机森林或着 gbm 差：根据精确的停止规则，测试集中有 7 个或者 8 个预测错误。支持向量机算法表现地更差，在测试集上有 11 个预测错误。

深度学习算法相较其他预测算法是非常错综复杂的，报告显示“使用了 780,736 个参数”，这些参数是通过交叉验证设置的内部调节参数。这一切的实现要归功于现代强大的计算能力，它是纯预测算法潜在的推动者。

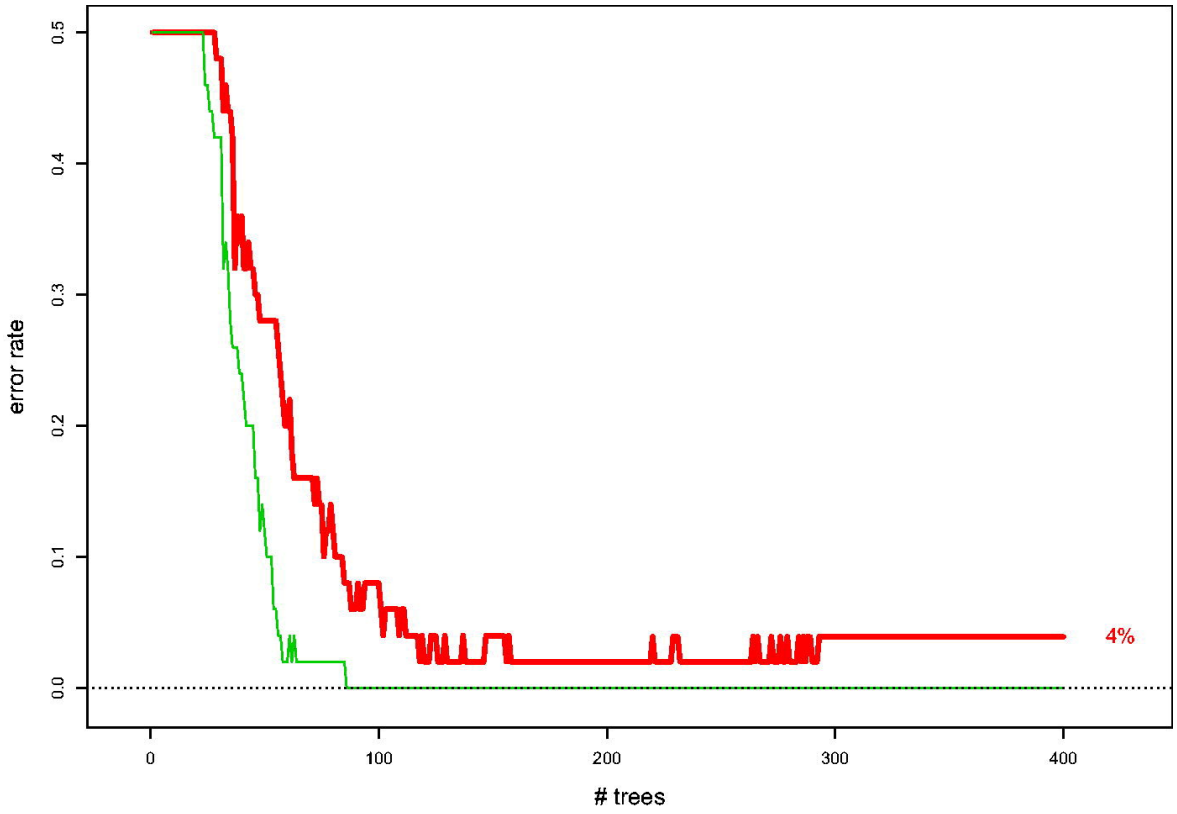

图 6: 将 boosting 算法 gbm 应用到前列腺癌微阵列研究的测试集错误率。细曲线是训练集错误率，在第 86 步时达到零。

5 预测的优点和缺点

对于我们这些努力在微阵列研究[8]中寻找“重要”基因的人来说，随机森林和 gbm 在前列腺癌症预测中得到的几乎完美的结果是一个令人不安的惊喜。在不忽视预测算法的惊喜或独创性的情况下，一个促成因素可能是预测比归因或估计更容易实现。总的来说，这是一个难以支持的猜想，但有几个例子有助于说明这一点。

对于估计而言，假设我们有 25 个独立同分布的观测，他们来自于期望 μ 未知的正态分布，

$$x_1, x_2, \ldots, x_{25} \overset{\text{ind}}{\sim} \mathcal{N}(\mu, 1), \tag{18}$$

并且进一步考察使用样本均值 $\bar{x}$ 还是样本中位数 $\breve{x}$ 来估计 μ。就均方误差而言，样本均值以压倒性的优势胜出，效果提高了一半以上，

$$E\left\{(\breve{x}-\mu)^2\right\} / E\left\{(\bar{x}-\mu)^2\right\} \doteq 1.57. \tag{19}$$

[8] 见 Efron and Hastie (2016) 中的图 15.5。

假设任务换成了对新数据 $X \sim N(\mu, 1)$ 做预测，样本均值仍然胜出，但效果仅仅提高了2%，

$$E\left\{(\breve{x} - X)^2\right\} / E\left\{(\bar{x} - X)^2\right\} \doteq 1.02. \tag{20}$$

造成这种情况的原因是：大部分的预测误差来自于 X 的多变性，这是 $\bar{x}$ 和 $\breve{x}$ 都无法解决的[9]。

预测比估计更容易，至少在更宽容的意义上是这样的。预测允许使用效率低的估计量，如 gbm 的“树桩”，同时这对于大规模部署是很方便的。纯预测算法的操作是非参数的，因此带来的好处是不太需要担心估计效率。

为了对比预测和归因，本研究考虑了微阵列研究的一个理想化版本，有 n 个受试者，其中 $n/2$ 个为健康的控制组，$n/2$ 个为生病的患者：每个受试者都提供了一个由 N 个基因测量值组成的向量，$\mathbf{X} = (X_1, X_2, \ldots, X_N)^t$，其中

$$X_j \overset{\text{ind}}{\sim} \mathcal{N}\left(\pm\delta_j/2c, 1\right) \quad (c = \sqrt{n/4}), \tag{21}$$

$j = 1, 2, \ldots, N$，“+”表示患病，“-”表示健康；δ_j 表示基因 j 的效应大小。大多数基因都是 $\delta_j = 0$，个数记为 N_0，但少数的 N_1 个基因的 δ_j 具有正值 Δ，记为

$$N_0 : \delta_j = 0 \quad \text{以及} \quad N_1 : \delta_j = \Delta. \tag{22}$$

当一个新的人到来并且产生了满足条件 (21) 的微阵列测量值 $\mathbf{X} = (X_1, X_2, \ldots, X_N)^t$ 时，没有人知道他是健康的还是患病的，即不知道 $\pm$ 值。由此引出一个问题：在预测变得不可能之前 N_1/N_0 可以多小？附录 A 中给出的答案是当 $N_0 \to \infty$ 时，如果

$$N_1 = O\left(N_0^{1/2}\right), \tag{23}$$

但不能低于这个值时，准确预测是有可能的。

相比之下，附录 A 展示了实现有效归因的要求是

$$N_1 = O\left(N_0\right). \tag{24}$$

就像“大海捞针”(Johnstone and Silverman, 2004) 一样，归因需要的“针”比预测多一个数量级。结合弱学习者的预测策略并不适用于归因，归因从定义上看基本上是在寻找较强的单个预测变量。至少在这个例子中我们似乎可以公平地说：预测比归因容易得多。

[9]这里可以想象成本研究有一个新的观测要预测。相反，假设本研究要预测 m 个新的观测 $X_1, X_2, \ldots, X_m \overset{\text{ind}}{\sim} \mathcal{N}(\mu, 1)$，并且本研究希望去预测他们的均值 $\bar{X}$。当 $m = 10$ 时，效率比为 $E\left\{(\bar{X} - \breve{x})^2\right\} / E\left\{(\bar{X} - \bar{x})^2\right\} = 1.16$；当 $m = 100$ 时，效率比则为 1.46；并且当 $m = \infty$ 时，效率比为 1.57。因此我们可以把估计看作是对未来平均值的预测。

三种主要的回归类别可以按以下顺序排列，

$$\text{预测} \quad \cdots \quad \text{估计} \quad \cdots \quad \text{归因}, \tag{25}$$

估计处于中心位置，而预测和归因彼此之间的距离更远。在传统统计学中，估计通过 p 值以及置信区间与归因联系在一起，如表 1 所示。从另一个角度来看，好的估计量通常也是好的预测量。预测和估计都将他们的输出集中在 $n \times p$ 的矩阵 $\mathbf{x}$ 的 n 侧，而归因则集中在 p 侧。在 (25) 中，估计面对两个方向。

`randomForest` 算法也做了一些将预测和归因联系在一起的尝试。在进行预测的同时，还计算了 p 个预测变量的*重要性度量*[10]。图 7 展示了图 5 的前列腺癌症应用中重要性得分从高到低的排序结果。在 $p = 6033$ 个基因中，348 个得分为正，这些基因曾被选择为分割变量。图中可以看出基因 1031 最重要，另外有 25 个基因在重要性曲线的急弯上。我们是否可以用重要性得分作为归因，就像表 1 中的星号那样?

在这种情况下，答案似乎是否定的。本研究从数据集中删除了基因 1031，将数据矩阵 $\mathbf{x}$ 减少到 102×6032，并重新运行 `randomForest` 预测算法。现在测试集预测错误的个数为零。当移除最重要的 5 个基因，最重要的 10 个，直到最重要的 348 个基因时，对测试集预测错误的个数产生的影响同样都很小，如表 3 所示。

表 3: 当移除图 7 所示的顶部预测变量时，对前列腺癌数据应用随机森林算法所得的测试集预测错误的个数。

# 移除	0	1	5	10	20	40	80	160	348
# 错误	1	0	3	1	1	2	2	2	0

最后，图 5 中参与构建初始预测准则的所有基因都已被移除。现在 $\mathbf{x}$ 是 102×5685，但是基于删减后的数据集 $\mathbf{d} = \{\mathbf{x}, \mathbf{y}\}$ 的随机森林准则仍给出了很好的预测结果。表 3 所示实际结果为测试集错误个数是零。最后一步的预测准则产生了 364 个“重要”基因，与最初的 348 个基因是不相交的。当从预测集中去除所有 712=348+364 个基因时，$\mathbf{x}$ 是 102×5321，仍然给出了一个随机森林预测准则，并且此准则只产生了一个测试集错误。

在这个例子中，预测的“弱学习者”模型似乎是占优势的。很明显大部分基因与前列腺癌症的相关性很弱，但是预测算法可以用这些弱学习者的不同组合来给出近乎完美的预测结果。如果预测是唯一的目标，这确实属于一个优势，但从归因的角度来看这是一个劣势。如表 1 所示，传统的归因是在努力识别一小部分有因果关系的协变量（即使不能推断出严格的因果关系）。

[10]有几种这样的度量值。第 3 节的图 7 中的度量值与基尼准则有关。在算法结束时，本研究得到了一个很长的列表，列出了所有 bootstrap 树的所有分割情况；单个预测变量的重要性得分是在所有分割中基尼系数下降的总和，其中该预测变量是指分割变量。

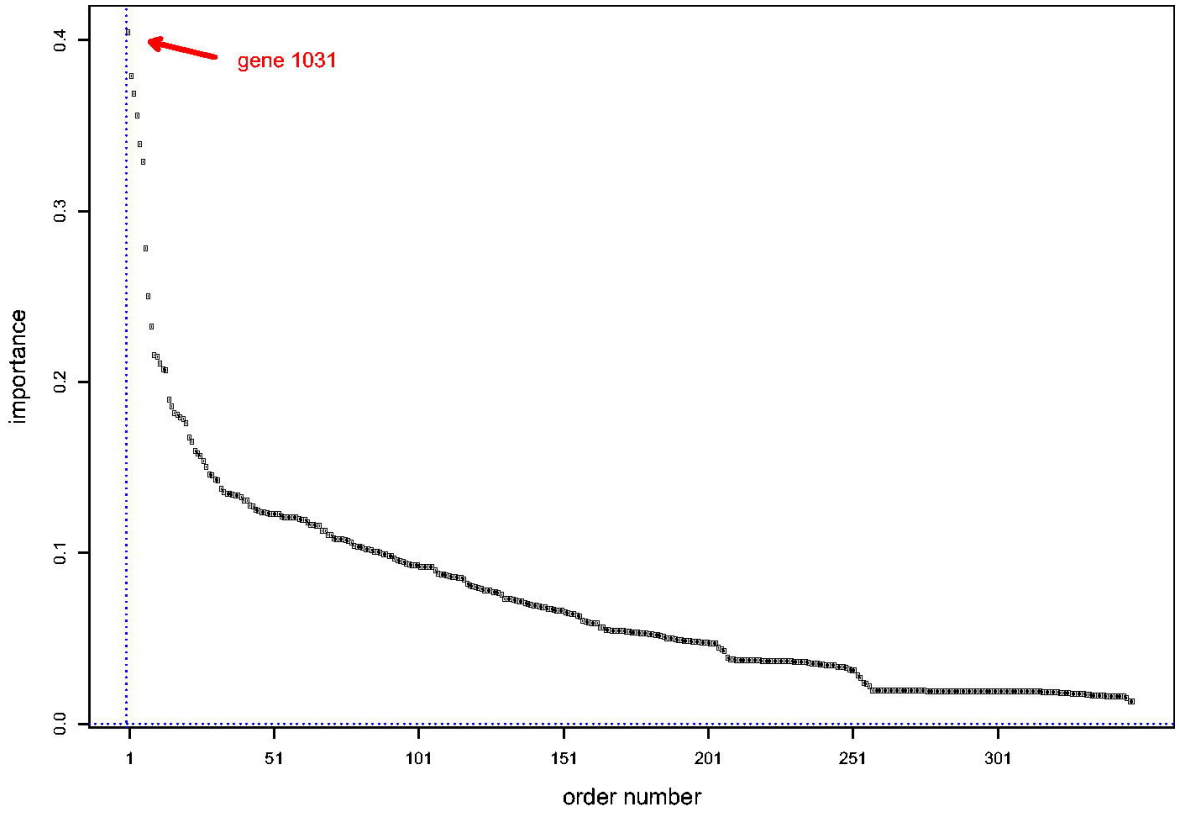

图 7: 图 5 中基于随机森林算法得到的前列腺癌症预测准则的重要性度量，按重要性递减的顺序绘制。

纯预测算法倾向于产生相关性较弱的预测变量，这使他们与归因背道而驰。第 9 节中讨论了*稀疏性*，指只存在少数重要预测变量的工作假设，而这些预测变量根本不是表 3 所传递的信息。

6 训练/测试集范式

现代预测方法的一个重要组成部分就是训练/测试集范式：数据集 $\mathbf{d}$ (1) 被划分为一个训练集 $\mathbf{d}_{\text{train}}$ 以及一个测试集 $\mathbf{d}_{\text{test}}$；然后仅使用训练集 $\mathbf{d}_{\text{train}}$ 计算得到一个预测准则 $f(x, \mathbf{d}_{\text{train}})$；最后利用 $f(x, \mathbf{d}_{\text{train}})$ 对测试集 $\mathbf{d}_{\text{test}}$ 进行预测，由此得到了此准则预测错误率的一个可靠估计，但可靠并不意味着完美。

在第 4 节的前列腺癌症微阵列研究中就采用了这种范式，应用随机森林算法得到了一个非常小的预测错误率 2%[11]。这样的预测结果对于我来说是非同寻常的。那么为什么不利用这一预测准则，基于一组新的 6033 个基因表达水平去诊断某人是否患有前列腺癌症呢？下面的例子说明了这可能会出错。

[11] 考虑到表 2 中的信息，更准确的错误率估计为 3.7%。

在第 4 节中，我们将 102 名男性随机地分为两个 51 人的小组，每组 25 名正常控制组和 26 名癌症患者，从而得到了前列腺癌症数据的训练集与测试集。文献中强调随机化是为了防止产生偏差。基于此，我再次对前列腺癌症数据进行了研究，但这次选择了 ID 号码较低的 25 名正常控制组和 26 名癌症患者作为训练集，测试集则是剩下的 51 名 ID 较高的受试者，同样包括 25 名正常控制组和 26 名癌症患者，这其实违背了随机化。

在重新分析中，`randomForest` 的表现不如图 5 的结果：$f(x, \mathbf{d}_{\text{train}})$ 在 $\mathbf{d}_{\text{test}}$ 上产生了 12 个错误预测，预测错误率高达 24% 而不是之前的 2%，如图 8 所示。boosting 算法 `gbm` 的表现也是一样糟糕，产生的预测错误率为 28%（14 个错误预测），如图 9 所示。

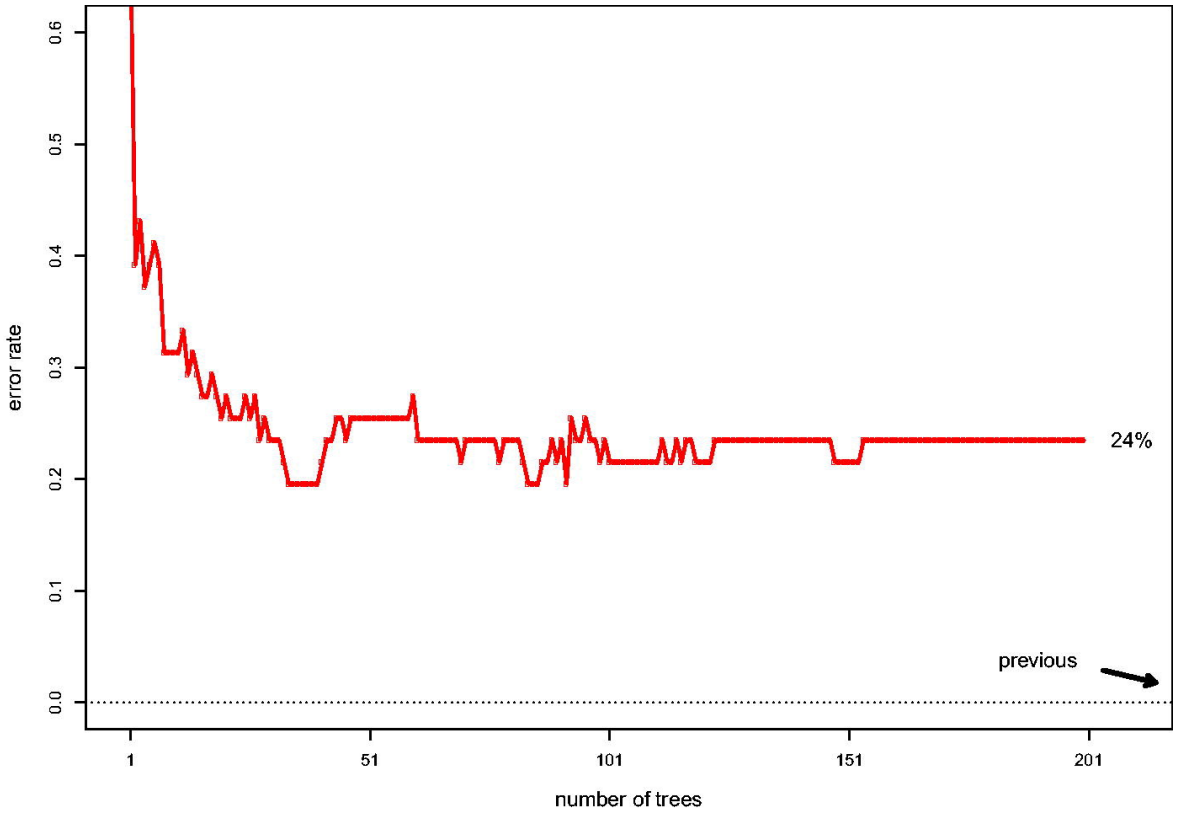

图 8: 基于 `randomForest` 得到的前列腺癌症微阵列研究的测试集误差，训练/测试集由 ID 号码早/晚决定。结果相较图 5 要差很多。

为什么此时的预测效果这么差呢？从检查结果来看并不是很明显，但前列腺研究的对象可能是按照所列的顺序收集的[12]，并且随着时间的推移会产生一些小的方法上的差异。也许所有被纳入 `randomForest` 和 `gbm` 的弱学习者都非常容易受到这种差异的影响。有关预测的文献用概念漂移作为这类难题的标签，一个臭名昭著的例子是 Google 的流感预测器，它在打败了

[12]在正常受试者数据的奇异值分解结果中，第二主向量与 ID 号之间有一个向上的倾斜关系，但对于癌症患者数据并非如此。

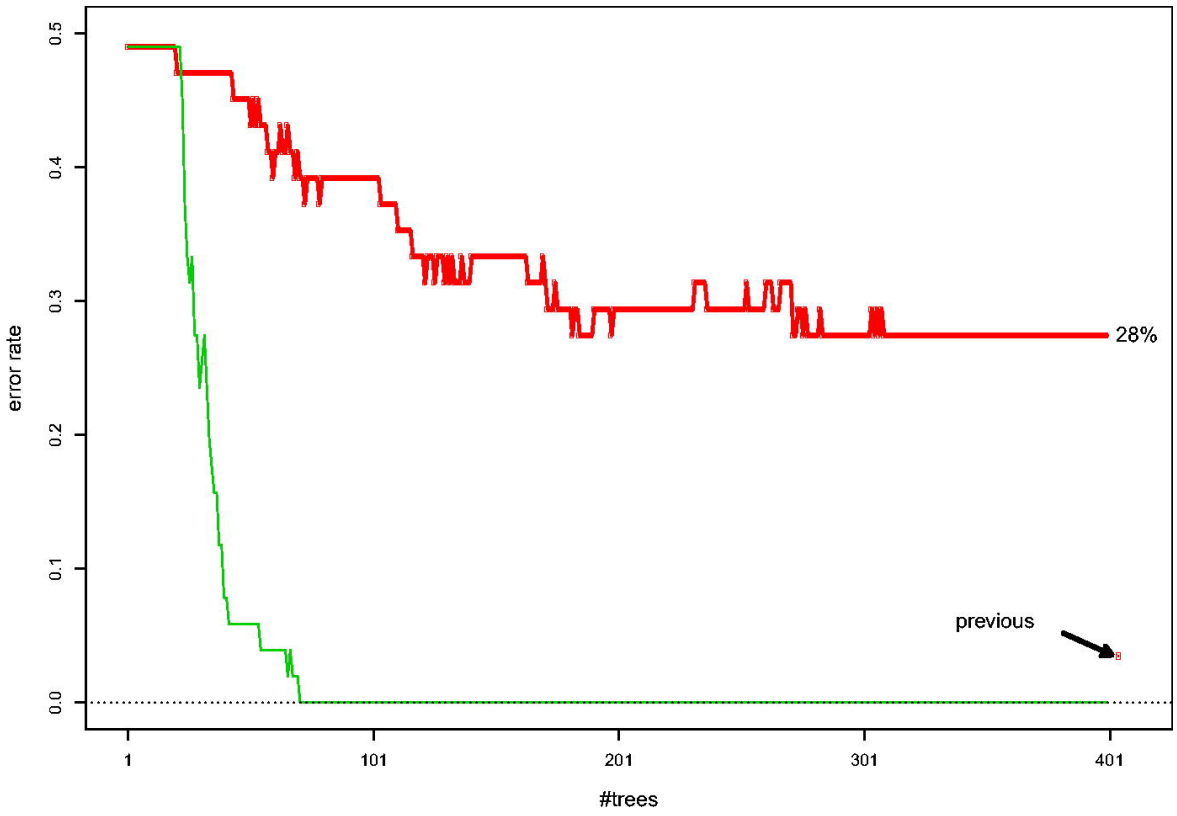

图 9: 基于 gbm 算法得到的测试集误差，利用 ID 的早/晚划分训练/测试集，并且与图 6 进行比较。继续使用 800 棵树，误差估计值会降低到 26%。细曲线表示训练集的错误率，在第 70 步后为零，然而测试集的错误率继续下降。见第 8 节的标准 5 中的简要讨论。

CDC（疾控中心）数年之后惨败[13]。通过随机的方式来选取测试集听起来很谨慎，但也肯定会隐藏一些漂移效应。

漂移这一概念让我们陷入这样一个问题：我们的各种新旧回归方法，应该告诉我们什么？从历史上看，科学一直在寻找控制宇宙的潜在真理：真理被认为是永恒的，如牛顿定律。在物理学和天文学中，永恒表现得十分清晰，如 $E = mc^2$，Hubble 定律；或许在医学和生物学中也是如此，例如 DNA 与血液循环。但是，现代科学已经进入了真理可能更具偶然性的领域，如经济学、社会学以及生态学。

在不使用牛顿的标准的情况下，传统的估计和归因旨在获得超越即时数据集的长期结果。在第 2 节的表面加噪声模型中，表面扮演着真理的角色——至少永恒到足以证明其在为最接近的估计而努力着。

在表 1 的新生儿例子中，我们希望像 gest 和 ap 这种被标星的预测变量在未来的研究中继续发挥重要作用。在第 2 年的数据中，只有 $n = 246$ 个新生儿。对第 2 年的数据进行了同样的

[13]疾控中心本身赞助了每年度基于互联网的流感预测挑战 (Schmidt, 2019)；可以在 *predict.cdc.gov* 查看他们过去的结果。

逻辑回归模型，得出的系数估计值与第 1 年的值非常相似，见表 4。牛顿不会羡慕这样的结果，但似乎已经发现了比眼前利益更重要的东西。

表 4: 比较第 1 年（如表 1）和第 2 年新生儿数据的逻辑回归系数；相关系数为 0.79。

	gest	ap	bwei	resp	cpap	ment	rate	hr	head	gen	temp
第 1 年	–0.47	–0.58	–0.49	0.78	0.27	1.10	–0.09	0.01	0.1	0.00	0.02
第 2 年	–0.65	–0.27	–0.19	1.13	0.15	0.41	–0.47	–0.02	–0.2	–0.04	0.16

没有什么可以排除纯预测算法寻求永恒真理的可能性，但他们最著名的用途还是应用在更短暂的现象中：信用评分、Netflix 电影推荐、面部识别、*Jeopardy!* 竞赛。预测算法的一大优势是能够从大量混杂的数据集中提取信息，即使只是为了短期使用。只有满足估计的错误率在当前数据集的有限范围之内，或者不会在有限范围之外太远的设定，对测试集进行随机选取才是有意义的。

接下来是一个人为设计的微阵列实例，所有的预测变量都是短暂的：$n = 400$ 名受试者参与研究，每天在治疗组和控制组之间交替到达一个，每个受试者被测量一个 $p = 200$ 的基因微阵列。400×200 数据矩阵 $\mathbf{x}$ 的每条数据是相互独立的，其中

$$x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ij}, 1) \quad \text{对于} \quad i = 1, 2, \ldots, 400 \quad \text{以及} \quad j = 1, 2, \ldots, 200. \tag{26}$$

大多数的 μ_{ij} 为 0，但是偶尔会有一个基因有 30 天的活跃片段，在这期间

$$\mu_{ij} = 2 \quad \text{对于治疗组} \quad \text{以及} \quad -2 \quad \text{对于对照组}, \tag{27}$$

或

$$\mu_{ij} = 2 \quad \text{对于对照组} \quad \text{以及} \quad -2 \quad \text{对于治疗组}, \tag{28}$$

在 (27) 和 (28) 之间的选择是随机的，每一个活跃片段的开始日期也是随机的。每个基因的平均活跃次数为 1 次。图 10 中黑色的线段代表所有的活跃时期。

400 名假设的受试者被随机分为 320 人的训练集和 80 人的测试集。`randomForest` 分析的结果展示在图 11 的左边，测试集错误率为 19%。接着本研究进行了第 2 次的 `randomForest` 分析，使用第 1 天到 320 天的受试者作为训练集，以及第 321 天到 400 天的受试者作为测试集。结果展示在图 11 的右边，测试错误率约为 45%。

在这种情况下，很容易看出问题出在哪里。从任何一天的测量结果中，预测准则都有可能从附近几天的活跃片段来预测是控制组还是治疗组（即使不知道图 10 中的活跃图）。这样的情况适用于随机分割法，因为其大部分的测试日期与训练日期是混在一起的。但对于先后次

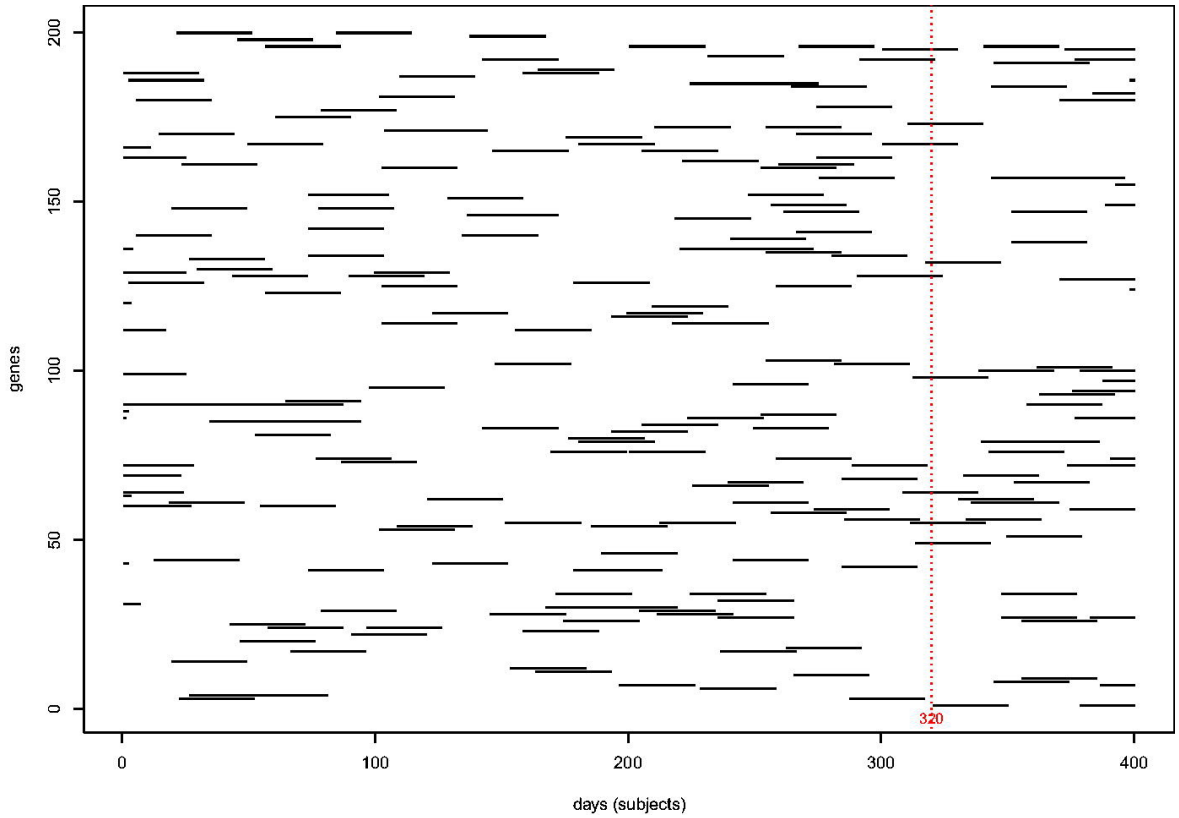

图 10: 黑色线段表示在假设的微阵列研究中基因的活跃期。(矩阵转置是为了便于排版。)

序分割法就不是这样了，因为大部分的测试日期都与训练日期相差很远。换言之，插值预测比外推预测更容易[14]。

通常我们可以从训练/测试集误差估计中学到什么？回到公式 (1)，一般的假设是数据对 (x_i, y_i) 是独立同分布于某个 $(p+1)$ 维空间上的概率分布 F，即

$$(x_i, y_i) \overset{\text{iid}}{\sim} F \quad \text{对于} \quad i = 1, 2, \ldots, n. \tag{29}$$

我们选择了大小为 n_0 的训练集 $\mathbf{d}_0$ 以及大小为 $n_1 = n - n_0$ 的测试集 $\mathbf{d}_1$ （在模型 (29) 下怎样才是不相关的），进一步计算得到预测准则 $f(x, \mathbf{d}_0)$ 并应用到 $\mathbf{d}_1$ 上，从而得到了一个误差估计

$$\widehat{\text{Err}}_{n_0} = \frac{1}{n_1} \sum_{d_1} L\left(y_i, f\left(x_i, \mathbf{d}_0\right)\right), \tag{30}$$

其中 L 是某种损失函数，如均方误差或计数误差。然后基于模型 (29)，$\widehat{\text{Err}}_{n_0}$ 的无偏估计为

$$\text{Err}_{n_0}(F) = E_F\left\{\widehat{\text{Err}}_{n_0}\right\}, \tag{31}$$

[14]Yu and Kumbier (2019) 提出了“内部检验”与“外部检验”的主要区别。

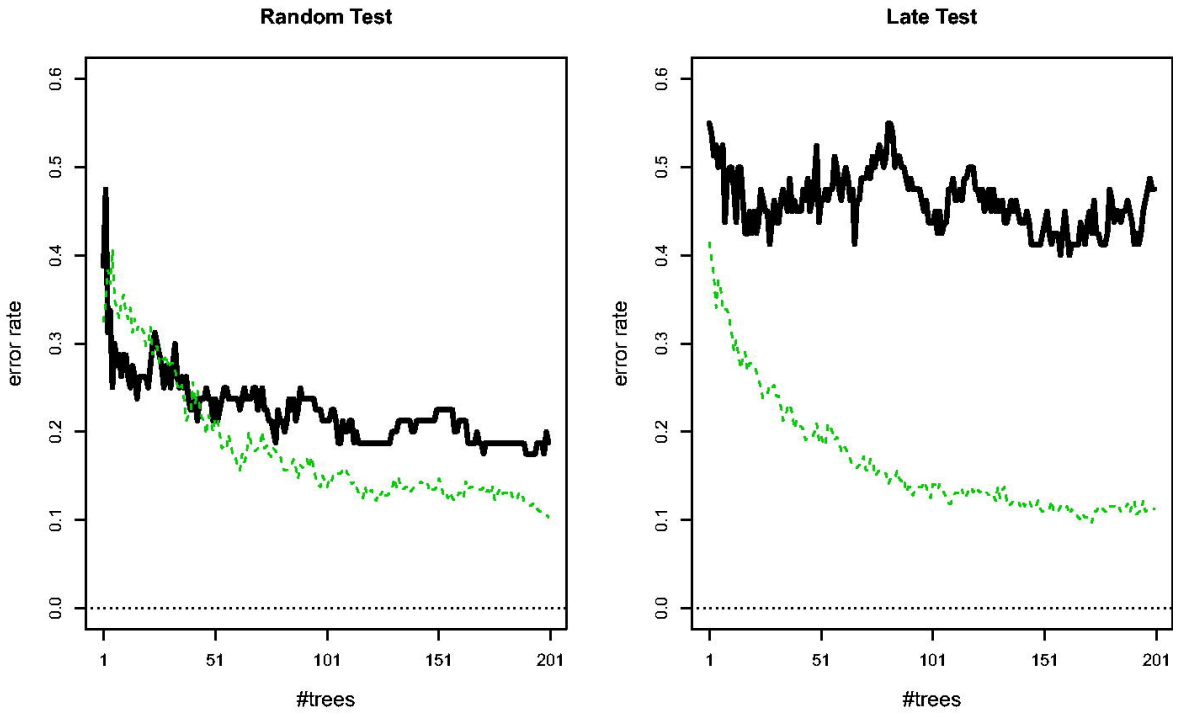

图 11: 对图 10 中的人造微阵列研究进行 randomForest 预测。左面板：从 400 天中随机选取了大小为 80 的测试集；粗黑色曲线显示出最终估计的测试错误率为 19%。右面板：测试集为 321-400 天；现在错误率估计是 45%。两幅图中的浅虚线曲线均为训练集误差。

由 n_0 形成的准则[15] $f(x, \mathbf{d}_0)$ 的平均预测误差取自 F。

漂移这一概念可以被解释为数据生成机制 (29) 的变化，比如分布 F 变化为一些新的分布 $\tilde{F}$，这可能就是图 8 和 9 中前列腺癌症预测效果很差的罪魁祸首[16]。传统的预测方法也容易受到这种变化的影响。在新生儿研究中，基于第 1 年数据的逻辑回归预测模型所得的交叉验证错误率为 20%，而应用于第 2 年数据时，交叉验证错误率增加至 22%。

对于图 10 和图 11 中人为构造的例子来说情况是更加复杂的，因为模型 (29) 并不严格适用。有效的预测变量是短暂的，其只在短时间内活跃。一个合理的推测是（但也仅此而已），纯预测算法的弱学习者容易出现短暂性，或者至少比传统方法中青睐的主要预测变量更容易出现短暂性。不管这是不是真的，我觉得通过随机选择来构建训练/测试集是有一些危险的，他们的误差在统计上一定要有所保留。从实际操作上来讲，向一个关心前列腺癌症的朋友推荐图 5 中的随机森林的预测准则另我很惶恐。

这不仅仅是一种假想出来的担忧。在 2019 年的文章 “Deep Neural Networks are Superior to

[15]重要的一点是，“一个准则”是指根据感兴趣的算法和数据生成机制形成的规则，而不是手头的特定规则 $f(x, \mathbf{d}_0)$；见 Efron and Hastie (2016) 的图 12.3。

[16]Cox 在对 Breiman et al. (2001) 的讨论中谈到模型 (29) 的适用性时说：“然而，很多预测都不是这样的。通常情况下，预测基于的是某些完全不同的条件。例如，如果医用 x 射线减少 10% 会对美国每年的癌症发病率产生什么影响？等等。”

Dermatologists in Melanoma Image Classification" (深度神经网络在黑色素瘤图像分类方面优于皮肤科医生) 中，Brinker et al. (2019) 证明了标题中所说的结论。但作者非常谨慎，建议要在未来进行研究加以验证。此外，他们承认了随机选取测试集的局限性，以及算法中某些预测变量可能是短暂的。对于任何这种诊断算法的严格使用，以及在表明某些亚群未被误诊的研究中，频繁的更新都是十分必要的[17]。

7 平滑性

牛顿的微积分与牛顿的运动定律相吻合并不仅仅是一个愉快的巧合。如拉普拉斯所描述的那样，牛顿的世界是一个无限平滑的世界，在这个世界里，微小的因果变化会产生微小的结果变化，并且许多阶导数具有物理意义。传统统计方法的参数化模型强化了平滑世界的范式。回头看第 2 节的图 1，我们可能不同意胆甾烯胺三次回归曲线的确切形状，但响应变量的平滑性无可争议：依从性从 1 到 1.01 的变化不会对预测的胆固醇降低量产生太大影响。

纯预测算法并没有构建响应变量的平滑性。图 12 的左侧展示了 `randomForest` 对胆固醇降低量的估计结果，是关于正态化依从性的函数。我们可以看到，左侧的结果大致符合 OLS 三次曲线，但呈现锯齿状并且绝对是不平滑的。右侧的图是由 `gbm` 算法得到的一条锯齿较少的"曲线"，但仍然有大量的局部不连续。

图 1 中的三次曲线是通过比较 1 到 8 阶多项式回归对应的 Cp 值来选择的，三次对应的 Cp 值表现最优。图 12 中的 `randomForest` 和 `gbm` 都首先将 **x** 设为 164×8 的矩阵 `poly(c,8)`（这是在 R 语言中的记号），其中 **c** 是由调整的依从性所构成的向量，它是一个 8 次多项式的基。图 12 中右侧的虚线曲线为 8 阶多项式的 OLS 拟合，令人惊喜的是，`gbm` 预测结果在依从性的大部分范围内都与 8 阶拟合结果相符。也许这是预测算法能够被用作非参数回归估计的一个有希望的先兆，但在更高的维度上问题会变得更加困难。

接下来考虑超新星数据：记录了 $n = 75$ 个超新星的绝对亮度 y_i，以及在 $p = 25$ 个不同波长下的光谱能量测量值 x_i，因此数据集是

$$\mathbf{d} = \{\mathbf{x}, \mathbf{y}\}, \tag{32}$$

其中 **x** 为 75×25，**y** 是一个 75 维向量。经过一些预处理，一个合理的模型是

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1). \tag{33}$$

本研究希望通过 x_i 去预测 μ_i。

[17]面部识别算法已被证明具有性别、年龄和种族偏见。

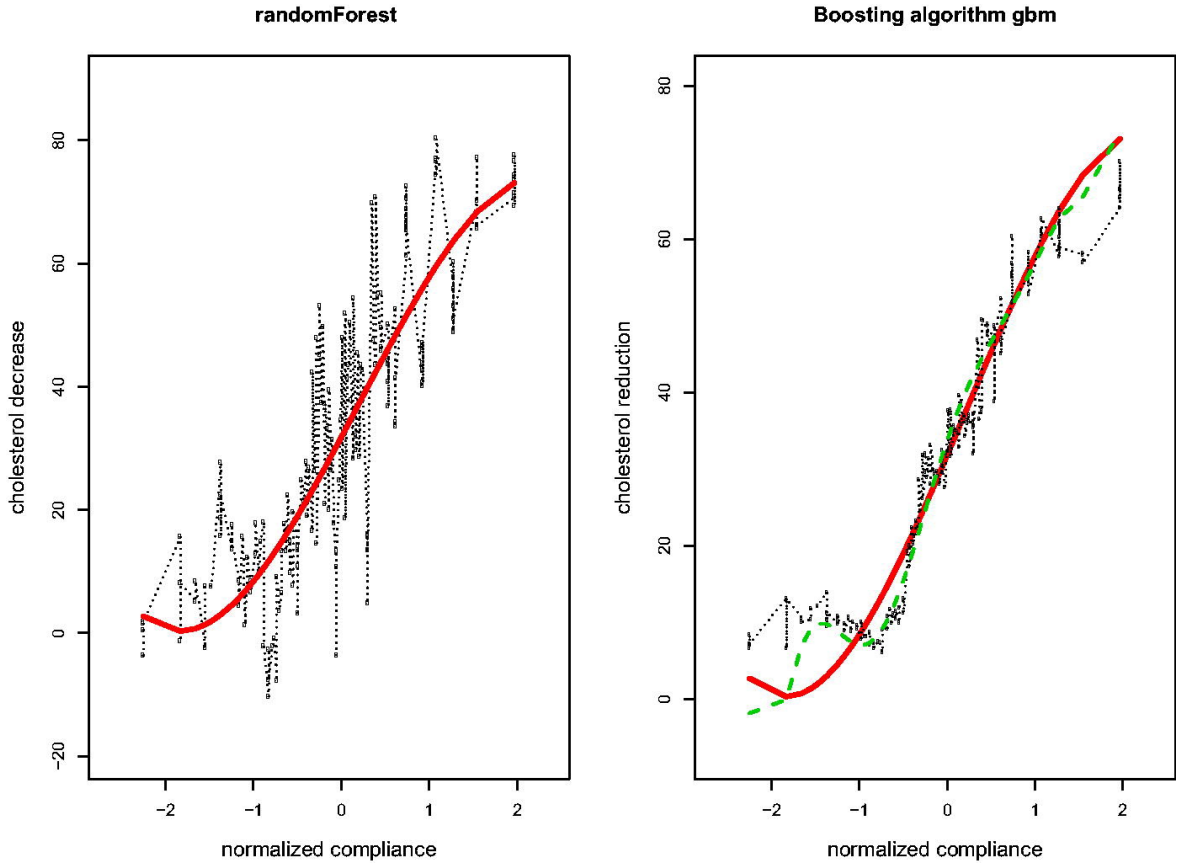

图 12: 对第 2 节中图 1 的胆甾烯胺数据进行 randomForest 以及 gbm 拟合的结果。粗曲线为立方 OLS；右侧面板中的虚线曲线为 8 阶 OLS 拟合。

我们的数据 **d** 是非常有利的，因为这 75 颗超新星在离地球足够近的地方，不需要使用 x_i 数据就可以直接确定 y_i 的值。然而这种测定通常是不可用的，并且 x_i 总是可观察到，因此准确的预测准则为

$$\hat{y}_i = f\left(x_i, \mathbf{d}\right), \tag{34}$$

这可以让天文学家更好地利用 1a 型超新星作为“标准烛光”来确定到遥远星系的距离[18]。

接下来利用 randomForest 以及 gbm 算法去拟合超新星数据 (32)。得到的结果有多光滑或者参差不齐呢？将 75 个观测中的任意两个记为 i_1 和 i_2，令 $\{x_\alpha\} \in R^{25}$ 是连接 x_{i1} 和 x_{i2} 的直线，

$$\{x_\alpha = \alpha x_{i_1} + (1-\alpha)x_{i_2} \quad \text{对于} \quad \alpha \in [0,1]\}, \tag{35}$$

用 $\{\hat{y}_\alpha\}$ 表示对应的预测值。基于一个线性模型可以得到线性插值 $y_\alpha = \alpha y_{i_1} + (1-\alpha)y_{i_2}$。

[18]暗能量的发现和宇宙膨胀需要将 1a 型超新星视为始终具有相同的绝对亮度，也就是说，视为完美的标准烛光。这并不是完全正确的。这一分析的目的是通过回归方法使蜡烛更标准，从而改进宇宙膨胀下的距离测量。Efron and Hastie (2016) 的第 12 章中讨论了这个数据的一个子集。

图 13 展示了 $\{\hat{y}_\alpha\}$ 在三种情形下的结果：$i_1 = 1$ 和 $i_2 = 3$，$i_1 = 1$ 和 $i_2 = 39$ 以及 $i_1 = 39$ 和 $i_2 = 65$。randomForest 的预测轨迹在局部和全局范围内都非常曲折离奇，gbm 没有那么明显，但离达到光滑仍然很远[19]。

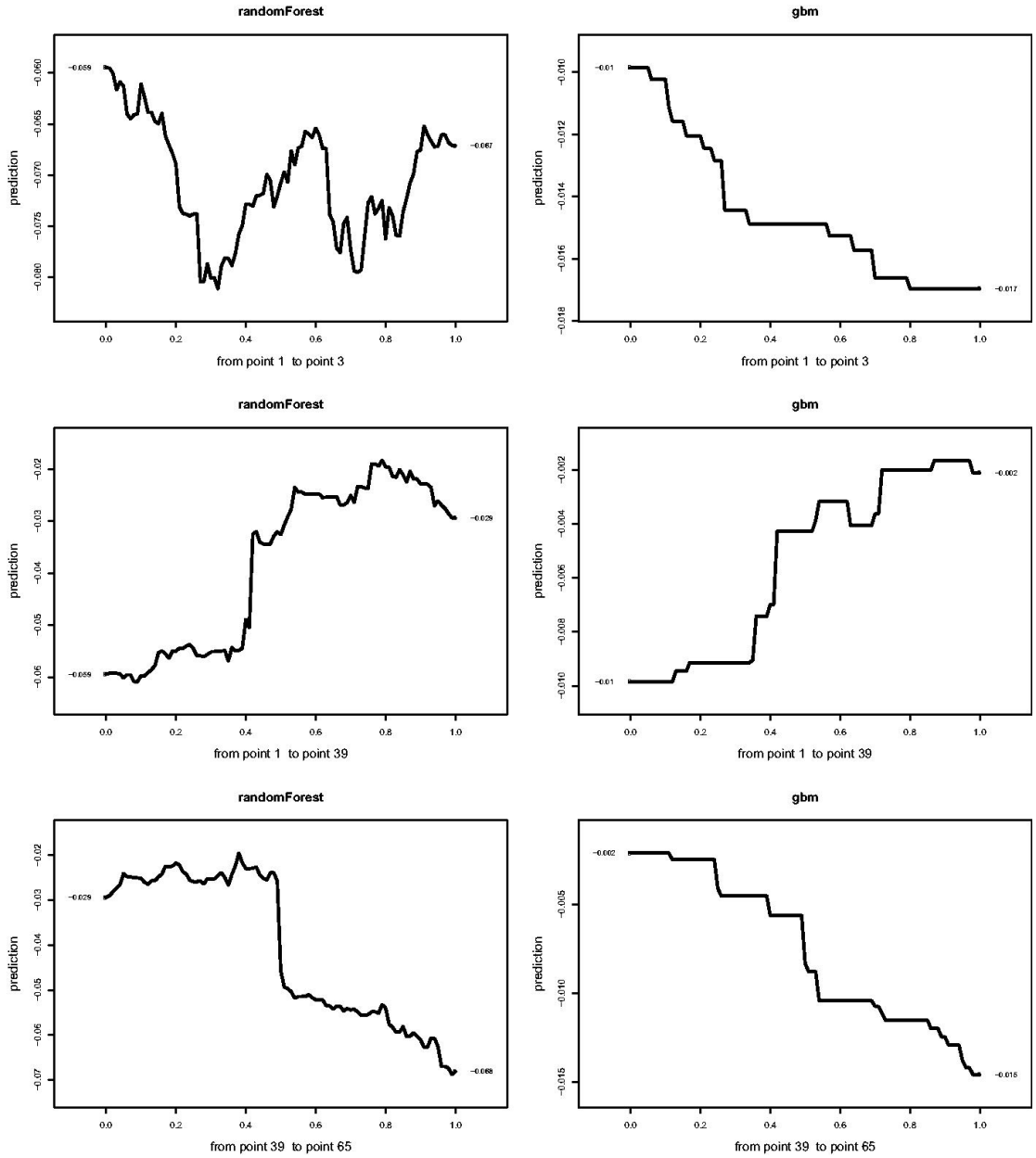

图 13: 超新星数据中点对之间的插值。左边是 randomForest，右边是 gbm。

如果目标对象本身就是离散的则不需要模型的平滑性：比如说电影推荐、信用评分以及国际象棋走法等。而对于一些科学应用来说，平滑性对于模型的可信度非常重要。据我所知，纯

[19]由 gbm 得到的相对平滑的结果必须与它对超新星数据给出更糟糕的预测这一事实进行权衡，这很大程度上缩小了 $\hat{y}_i$ 向零的方向。

预测算法得到参差不齐的结果并没有内在原因。神经网络本质上是一种复杂的逻辑回归程序，因此可能有望产生更平滑的输出。

8 一个比较清单

尽管预测和估计经常被混为一谈，但是他们并不一样。这篇文章的大部分是在关注差异。作为对过去的总结以及进行更广泛讨论的跳板，本节列出了一些重要的区别以及他们在统计实践方面的意义。

Breiman et al. (2001) 在 *Statistical Science* 期刊中发表的文章 "Statistical Modeling: The Two Cultures"（统计建模：两种文化）为新千年在预测的优点上开了一个好头。在这项工作中，Leo 对 "算法文化"（我一直称之为纯预测算法）进行了充满活力和热情的论证，同时批评了 "数据建模文化"（即传统方法）在大数据时代的局限性。首席讨论者 David Cox 教授以其特有的平衡态度为主流统计学辩护，他没有拒绝预测算法，而是指出了他们的局限性。我是第二个讨论者，对 Leo 的观点持怀疑态度（当时他对随机森林很热衷，这在当时还很新鲜），但也有些印象深刻。

事实证明，Breiman 比我更有先见之明：在 21 世纪，纯预测算法已经占据了统计领域的风头，在很大程度上沿着 Leo 提出的思路不断发展。本文可以被认为是我为回答预测算法如何与传统回归推断相关联的问题而做出的持续努力。

表 5 列出了区分传统回归方法和纯预测算法的 6 个标准。我需要再次提出之前的 "宽泛" 警告：我确信所有这六个区别都可以找到例外，列出的属性也不是一成不变的，唯一的含义是他们反映了当前的用法。

表 5: 关于传统回归方法和纯预测算法的差异的比较清单。

	传统回归方法	纯预测算法
1.	表面加噪声模型（连续的、光滑的）	直接预测（可能是离散的，锯齿状的）
2.	科学真理（长期）	经验预测准确性（可能是短期的）
3.	参数化建模（因果关系）	非参数化建模（黑箱）
4.	简约建模（研究者选择协变量）	反简约建模（算法选择预测变量）
5.	$\mathbf{x}_{p\times n}$：$p \ll n$ （同质数据）	$p \gg n$，两者都可能是巨大的（混合数据）
6.	最优推断理论（mle，Neyman-Pearson）	训练/测试范式（通用任务框架）

准则 *1*：表面加噪声模型在传统回归方法中无处不在，以至于在纯预测领域中他们的缺失是令人不安的。`randomForest`，`gbm` 或他们的同类算法都不需要表面或者噪声这样的输入，这在易于使用方面是一个巨大的优势。另外一个好处是：如果没有模型也就无法使用错误的模

型。

前列腺癌症的临床医生肯定会对有效的预测感兴趣，但研究型科学家更感兴趣的是疾病的病因，这也是传统统计方法发挥作用的地方。如果随机森林算法自 1908 年就存在了，而有人刚刚发明了回归模型显著性检验，那么新闻媒体现在可能正在宣告一个“锋利数据”时代的到来。

从推断中消除表面构建会产生一系列的不好的后果，如下所述。其中一个损失是平滑性。为了产生轰动效应，预测算法的应用大多集中在离散的目标空间中，如亚马逊推荐、翻译程序、驾驶方向等，这些都和平滑性无关。将他们用于科学研究的愿望可能会加速开发出更平滑、更合理的算法。

准则 *2*：表 5 的两边使用了类似的拟合标准——对于定量的响应变量使用某种版本的最小二乘，但他们使用的是不同的思维模式。遵循着 200 年的科学道路，传统的回归方法旨在从噪声数据中提取潜在的真理：虽然不是永恒的真理，但至少是一些超越当前经验的启示。

如果不需要对表面加噪声机制进行建模，科学真理在预测方面的重要性就会褪色。可能不存在任何潜在的真相。预测方法可以适应短暂的关系，这种关系只需要下一次更新之前保持有效即可。引用 Breiman 的话，“该领域的理论将重点从数据模型转移到了算法的性质上”，也就是从物理世界转移到了计算机。预测届的研究是一个巨大的事业，它的研究确实主要集中在算法的计算特性上，特别是关注当 n 和 p 变得巨大时他们表现如何，而较少关注他们与数据生成模型之间的关系。

准则 *3*：参数化建模在传统推断方法中扮演着中心角色，而预测算法是非参数化的，如 (29) 所示。(“非参数”可能涉及大量的调整参数，在深度学习中有数百万个这样的参数，所有这些参数都与算法有关而与数据的生成无关。）参数模型的背后通常隐藏着一些因果关系。在第 2 节图 1 的胆甾烯胺例子中，即使严格的因果关系是我们难以摸透的 [20]，我们也可能认为增加药物胆甾烯胺的摄入会导致胆固醇呈 S 型下降。

放弃数学模型就等于放弃理解自然这一历史性的科学目标。Breiman 直言不讳地陈述道：“数据模型在算法文化这个领域中很少被使用。这种方法是指，大自然在一个黑箱中产生数据，黑箱的内部是复杂的，神秘的，至少部分是不可知的 [21]。”

黑箱方法在科学上有一种反智力的感觉，但从另一方面看，如果预测是唯一的目标，那么科学理解可能就无关紧要了。机器翻译提供了一个有用的研究案例，其中基于语言结构的语言学分析与或多或少的纯预测方法之间存在着几十年的冲突。在统计机器翻译（SMT）的总称下，后一种方法已经席卷了整个领域，例如，谷歌翻译目前使用的是深度学习预测算法。

传统的统计教育包括大量的概率论课程。概率在非参数纯预测的视角下所占比例较小，

[20] Efron and Feldman (1991) 努力对因果关系进行了论证，这个论证没有被后来的作者不加批判地接受

[21] Cox 反驳道：形式化模型是有用的，而且对于深刻的思考来说几乎是必不可少的。

而交叉验证以及 bootstrap 等概率上的简单技术则肩负了方法论的重担。Mosteller and Tukey (1977) 的著作，*Data Analysis and Regression: A Second Course in Statistics* 倾向于一种非概率的推断方法，而这与机现代机器学习课程相吻合。

准则 4：表 1 中的 11 个新生儿预测变量是从最初的 81 个预测变量中筛选出来的，这是经过初步测试并与医学科学家进行讨论得到的。简约建模是传统方法的一个特征。这对于估计尤其是归因是至关重要的，因为在通常情况下发现的能力会随着预测变量的增加而减弱。

纯预测世界是反简约的。对于预测变量集合的控制，或所谓的“特征”的控制，在由统计学家传递到算法的过程中可以创造出高度交互的新特征，例如随机森林的树变量。Breiman 说道：“预测变量越多，信息就越多”，这是一个对深度学习时代特别准确的预见。

但是我是持怀疑态度的。我对 Breiman 的文章的评论是这样开头的：“乍一看，Leo Breiman 的这篇令人振奋的文章似乎是在反对简约和科学洞察力，并支持使用带有许多调整参数的黑箱。再一看，它仍然是这样，但这篇文章是令人兴奋的 ...”。如在第 4 节的图 5 和图 6 中，对前列腺癌症数据进行 `randomForest` 以及 `gbm` 预测得到了不错的结果，这肯定支持了 Leo 的说法。但仍可能有所保留。这里创造出的特征似乎属于弱学习者类型，也许本质上比表 1 中假定的强学习者更短暂。

这是第 6 节中提出的建议。如果预测算法是通过一群弱学习者的巧妙组合而起作用的，那么这些弱学习者对于预测会比估计，特别是比归因更有用（如第 5 节所述）。“短期科学”是一种矛盾的说辞。预测算法用于科学发现将取决于其能够提供长期有效性的证明。

准则 5：传统的应用要求 $n \times p$ 数据矩阵 $\mathbf{x}$（n 个受试者，p 个预测变量）的 n 充分大于 p，比如说 $n > 5 \cdot p$，这样的数据被称为“高数据”。新生儿数据中 $n = 812$、$p = 12$（将截距项算在内）完全符合高数据的特点。第 7 节的超新星数据中 $n = 75$ 以及 $p = 25$ 就没有完美符合高数据。在表 5 的右边，纯预测算法则允许甚至鼓励 $p \gg n$ 的“宽数据”。前列腺癌症微阵列的研究数据明显属于宽数据，其中 $n = 102$ 以及 $p = 6033$。即使我们一开始是对高数据进行分析，如胆甾烯胺的例子，预测算法也会创造出一些新的特征来扩大原始数据的变量。

预测算法如何在 $p \gg n$ 的情况下避免过拟合？这一问题有各种各样的答案，但没有一个是完全令人信服的：首先，使用测试集可以保证对误差进行可靠评估（见准则 6 的讨论）。第二，大多数算法在训练阶段会使用交叉验证进行检验。最后，有一个活跃的研究领域，它们声称能展示算法的“自正则化”特性，因此，即使运行其中一个算法的时间远远超过了训练数据完美拟合的点，如第 6 节的图 9 所示，仍然会产生合理的预测[22]。

估计特别是归因在同质数据集中的效果很好，其中 (x, y) 来自一个狭义的总体。对于一项从特定的疾病种类中选择受试者的随机临床试验，其所收集的数据具有严格的同质性。而预

[22]例如，在 $p > n$ 的 OLS 拟合问题中，一般的估计 $\hat{\beta} = (\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{y}$ 是得不到的，算法应该收敛到可以完美拟合数据 $\mathbf{y} = \mathbf{x}\hat{\beta}$ 的 $\hat{\beta}$，并且具有最小范数 $\|\hat{\beta}\|$；见 Hastie et al. (2019)。

测算法不要求同质性，这一特点可以使其具有更广泛的应用，在结果的普遍适用性方面是一个优点，但在可解释性方面是一个缺陷。

纯预测算法的大规模处理特性使得对于很大的 n 和 p 也可以产生预测结果，但这是一个危险的优点。这导致了对更大规模训练集的渴望。巨大的训练集对预测有很好的影响，使任务更具插值性且更少外推性（即，更像图 5 和图 6，而不是图 8 和图 9），但在归因方面会造成混淆视听[23]。

传统的回归方法将预测矩阵 **x** 视作一个固定的辅助统计量。这大大简化了参数回归模型的理论；在纯预测的领域中，**x** 和 **y** 都是随机的，唯一的概率模型是：$(x, y) \overset{\text{iid}}{\sim} F$。此时理论是更加困难的，鼓励在准则 6 中讨论的经验强调。贝叶斯统计在概率预测世界中被削弱了，留下一个默认的频率基础作为理论基础。

准则 6：传统的统计实践是以一个世纪的理论发展为基础的。极大似然估计和 Neyman-Pearson 引理是指导应用方法学的最优准则。在表 5 的预测方面，理论的有效性被经验方法所代替，尤其是对训练/测试集的误差估计方面。

预测的优点是无需进行理论建模，但由于缺乏坚实的理论结构，导致了“百花齐放”：流行的纯预测算法之间是完全不同的。在过去的四分之一个世纪里，先是神经网络，然后是支持向量机，boosting，随机森林，以及以深度学习形式再现的神经网络都处在焦点位置。在缺乏理论指导的情况下，我们或许可以期待更多。

在所谓的“通用任务框架”中，各种预测类比赛被用来为算法评分以取代理论的标准。常见的任务是围绕着一些众所周知的数据集进行预测分析，其中最有名的是 Netflix 电影推荐数据。这些都不能很好的替代最优预测理论，虽然最优预测理论目前还不存在[24]。

测试集是估计预测误差的可靠工具，但从完整的数据集 **d** (1) 中随机选取测试集可能会削弱推断效果。即使是适度的概念上的转移也会大大增加实际的预测误差，如第 6 节的前列腺癌症微阵列研究的例子。在某些情况下，除了随机选择法还有一些其他替代方法，例如通过收集日期的早/晚选择训练/测试集，如图 8 和 9 所示。第 7 节中超新星数据的研究目标是将预测准则应用于距离地球更远的超新星，因此选择距离地球更远的超新星作为测试集是相对谨慎的做法。

在 1914 年，著名天文学家、统计学家 Arthur Eddington[25]提出，从正态分布的数据中估算标准误差时，平均绝对偏差比均方根更有效。Fisher 在 1920 年做出了回应，他证明了均方根不仅

[23] 如果一篇文章以“超过 100 万人被问及 ...”为开头，那么一个有经验的统计学家会停止阅读，因为他知道随机抽取 1000 个样本会更可取。在大数据时代，这种统计上的民间智慧可能会消失。在一本名为 Big Data（大数据）的畅销书中，作者在样本量问题上失去了所有平衡，其主张 $n =$ 所有：全国所有的流感病例，亚马逊网站上所有的书，所有可能的狗/猫图片。“在大数据时代寻找随机样本，就像在汽车时代紧紧抓住马鞭。”公平地说，这本书中“$n =$ 所有”的例子实际上是狭义定义的，例如曼哈顿的所有下水道。

[24] 贝叶斯准则提供了这样一种理论，但代价是其假设远远超出了当前预测环境的限制。

[25] 他后来以在天文学上证实了爱因斯坦的相对论而闻名。

优于平均绝对偏差，而且优于任何其他可能的估计量，这是他的充分性理论的一个早期例子。

传统方法建立在这些参数化见解的基础上。表 5 的两边都遵循着不同的准则：左边相对有序，布局良好，像瑞士，而右边是狂野的西部繁荣。双方都可以从相互交往中获益匪浅。在 20 世纪 20 年代之前，统计学家并没有真正理解估计，而在 Fisher 的工作之后，我们理解了。我们在大规模预测算法方面也面临同样的情况：有很多好的想法和令人兴奋的东西，没有从原则上理解他们，但进展可能就在眼前。

9 宽数据时代的传统方法

纯预测算法的成功对传统的理论和实践都产生了激励作用。传统统计理论在 20 世纪上半叶形成，主要面向高数据：n 很小且 p 更小，通常情况下 $p = 1$ 或 $p = 2$。不管人们是否喜欢预测算法，部分现代科学已经进入了宽数据时代。以下有三个例子。

大数据并没有被预测算法所独占。计算遗传学的研究也是基于大数据进行的，特别是 GWAS 的形式，即全基因组关联研究。Ikram et al. (2010) 在一项关于眼睛血管收缩的研究[26]中给出了一个令人印象深刻的例子。该例测量了 $n = 15,358$ 个人的收缩量；对每个人的基因组进行了评估，大约 $p = 10^6$ 个 SNPs（single-nucleotide polymorphisms，单核苷酸多态性），一个典型的 SNP 具有一定的 ATCG 值，该值出现在大多数人群中，或出现在一个较小的、不太普遍的替代值中。这一研究的目的就是发现与血管收缩相关的 SNPs。

对于 $\mathbf{x} = 15,356 \times 10^6$，我们无疑处在大数据和宽数据领域。此时表面加噪声模型将不再适用。取而代之的是，对每个 SNP 进行单独研究：进行线性回归，预测变量为该位置染色体对中微小多态性的数量，对于每个个体来说取值为 0、1 或 2，以及响应变量为他或她的收缩测量值。接着给出了一个针对原假设的 p 值 p_i，原假设为位置 i 处的多态性对收缩没有影响，$i = 1, 2, \ldots, 10^6$。显著性水平为 0.05 的 Bonferroni 阈值为：

$$p_i \leq 0.05/10^6. \tag{36}$$

Ikram et al. (2010) 在“曼哈顿图”中展示了他们的结果，纵坐标为 $z_i = -\log_{10}(p_i)$，横坐标对应的是在基因组中的位置。阈值 (36) 对应的 $z_i \geq 7.3$；10^6 个 SNPs 中 179 个为 $z_i > 7.3$，拒绝了基因无效的原假设。这些 SNPs 被聚集在基因组的五个位置上，其中一个位置几乎不起作用。作者声称发现了四个新的基因位点。这些可能代表了与血管收缩有关的 4 个基因（尽管 12 号染色体上的一个突峰被发现分布在一些相邻的基因上）。

GWAS 程序没有使用 $p = 10^6$ 个预测变量进行传统的归因分析，而是令 $p = 1$，进行了 10^6 次归因分析，然后使用第二层的推断来解释第一层的结果。本文下一个例子是两层策略更详

[26]微血管收缩被认为是导致心脏病发作的原因，但很难观察到心脏的情况；在眼睛中观察要容易得多。

细的实现。

虽然不是 10^6，但第 4 节中前列腺癌症微阵列研究的 $p = 6033$ 个特征足以限制了整体无法采用表面加噪声模型。相反，我们首先对每个基因进行单独的 $p = 1$ 分析，如 GWAS 例子中所示。第 j 个基因的数据 (16) 可以表示为

$$\mathbf{d}_j = \{x_{ij} : i = 1, 2, \ldots, 102\}, \tag{37}$$

其中 $i = 1, 2, \ldots, 50$ 为正常控制者，$i = 51, 52, \ldots, 102$ 为癌症患者。

在正态性假设下，我们可以计算出统计量 z_j，将患者与对照组进行比较，以达到良好的近似值[27]，

$$z_j \sim \mathcal{N}(\delta_j, 1), \tag{38}$$

其中 δ_j 表示第 j 个基因的*效用大小*：δ_j 等于 0 表示“无效基因”，即在患者与对照组中表现出相同遗传活性的基因，而对于正在寻找的基因种类的 $|\delta_j|$ 很大，即那些对比患者与对照组有不同反应的基因。

对单个基因本身的推断是直接的。例如，

$$p_j = 2\Phi(-z_j) \tag{39}$$

是用来检验 $\delta_j = 0$ 的双边 p 值。然而这忽略了对 6033 个 p 值的同时解释。因此与 GWAS 一样，此研究也需要第二层的推断。

贝叶斯分析给效应大小假定了一个先验的“密度” $g(\delta)$，在 $\delta = 0$ 处的取值为概率 π_0，用来表示无效基因。实际上大多数的基因对前列腺癌症症都是不起作用的，因此 π_0 可以看作接近于 1。*局部错误发现率* $\text{fdr}(z)$，表示给定 z 值 z 的情况下基因是无效基因的概率，根据贝叶斯准则有

$$\text{fdr}(z) = \pi_0\phi(z - \delta)/f(z) \doteq \phi(z - \delta)/f(z), \tag{40}$$

其中 $\phi(z) = \exp\{-z^2/2\}/\sqrt{2\pi}$，$f(z)$ 是 z 的边际密度函数，

$$f(z) = \int_{-\infty}^{\infty} \phi(z - \delta)g(\delta)d\delta. \tag{41}$$

先验 $g(\delta)$ 大多情况下是未知的。*经验贝叶斯分析*假设 $f(z)$ 属于某个参数族 $f_\beta(z)$；通过在所有 6033 个 z 值的观测集合 $\{z_1, z_2, \ldots, z_p\}$ 上拟合 $f_\beta(\cdot)$ 计算出 β 的极大似估计 $\hat{\beta}$，从而得到

[27]如果 t_j 是一个对比患者与控制组的两样本 t 统计量，我们取 $z_j = \Phi^{-1}F_{100}(t_j)$，其中 F_{100} 是自由度为 100 的 t 统计量的累积密度函数，Φ 是标准正态分布的累积密度函数。效用大小 δ_j 是患者与对照组之间期望差异的单调函数；见 Efron (2012) 的 7.4 节。

错误发现率的估计值为

$$\widehat{\text{fdr}}(z) = \phi(z-\delta)/f_{\hat{\beta}}(z). \tag{42}$$

图 14 中的虚线展示了 $\widehat{\text{fdr}}(z)$ 的结果，其中 f_β 是通过 5 阶对数多项式模型得到的，

$$\log\{f_\beta(z)\} = \beta_0 + \sum_{k=1}^{5}\beta_k z^k; \tag{43}$$

$\widehat{\text{fdr}}(z)$ 在 $|z|<2$ 时接近 1（即基因几乎可以肯定是无效的），随着 $|z|$ 的增大 $\widehat{\text{fdr}}(z)$ 逐渐降为 0，例如在 $z=4$ 是等于 0.129。显著性检验的传统阈值为 $\widehat{\text{fdr}}(z)\le 0.2$；如图 14 中的三角形所示，有 29 个基因满足这个条件。

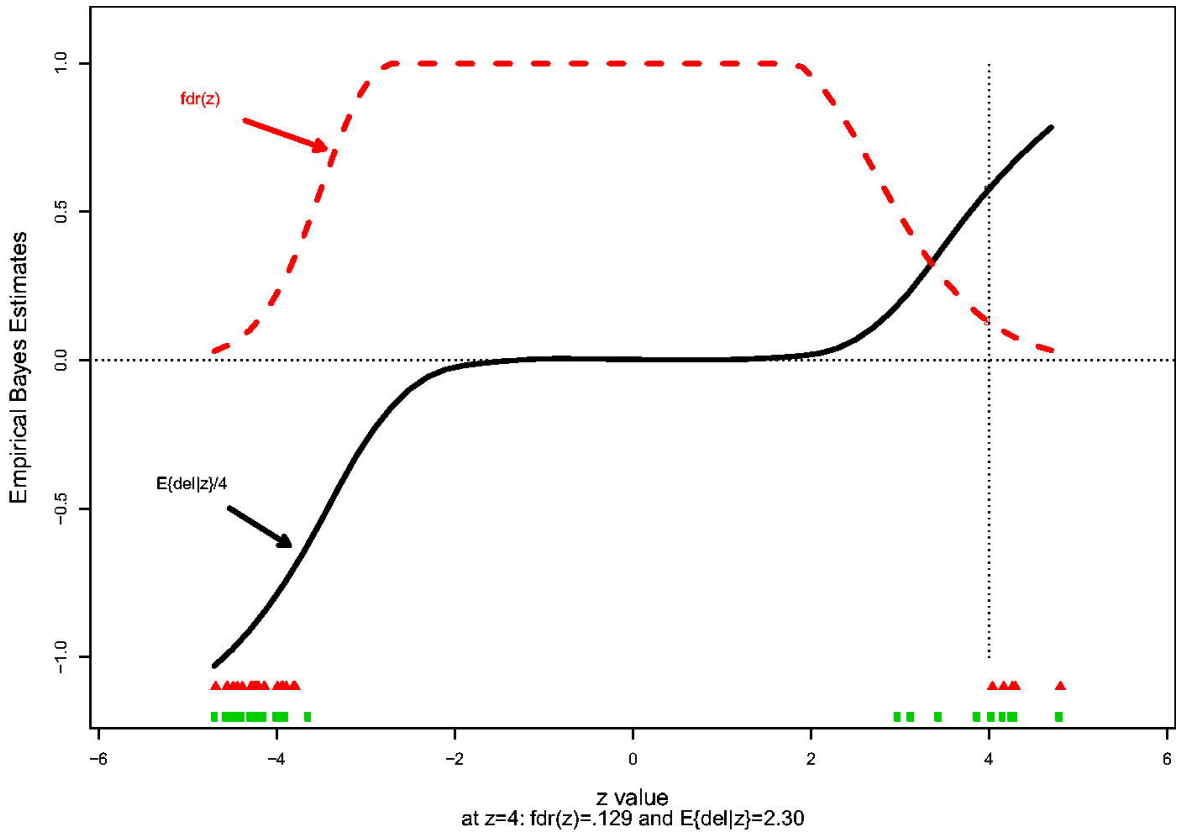

图 14: 对前列腺癌症数据进行验贝叶斯分析，得到局部错误发现曲线的估计 $\widehat{\text{fdr}}(z)$ 以及后验效用大小的估计 $\hat{E}\{\delta|z\}$（为了展示的目的，对后者除以 4。）三角形表示 29 个基因具有 $\widehat{\text{fdr}}(z)\le 0.2$；方块是由 `glmnet` 分析得到的 29 个最重要的基因。

我们也可以估计预期的效用大小。*Tweedie* 公式 (Efron, 2011) 给出了后验期望的一个简单表达式

$$E\{\delta \mid z\} = z + \frac{d}{dz}\log f(z), \tag{44}$$

$f(z)$ 为边际密度 (41)。在图 14 中，用 $f_{\hat{\beta}}$ 代替 f 来计算 $E\{\delta \mid z\}$。当 $|z| \le 2$ 时 $E\{\delta \mid z\}$ 接近为 0，在 $z = 4$ 时上升为 2.30。

通过使用一个两层次模型，经验贝叶斯分析将 $p = 6033$ 减少到 $p = 5$。这样就又回到了传统方法的舒适区，从而可以很好地进行估计和归因分析。图 14 对此均有说明。

*稀疏性*为宽数据的估计和归因提供了另一种方法：我们假设 p 个预测变量中的大多数是不起作用的，并集中精力寻找为数不多的重要变量。*lasso* (Tibshirani, 1996) 提供了一个关键的研究方法。在一个 OLS 类型的问题中，我们通过最小化以下式子来估计 p 维回归系数 β，

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^t\beta\right)^2 + \lambda\|\beta\|_1, \tag{45}$$

其中 $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$。

这里 λ 是一个固定的调整参数：$\lambda = 0$ 对应了 β 的普通最小二乘解（如果 $p \le n$）而 $\lambda = \infty$ 使得 $\hat{\beta} = 0$。λ 取值很大时，只有少数的协变量系数 $\hat{\beta}_j$ 是非零的。该算法从 $\lambda = \infty$ 开始然后减小 λ，每次引入一个新的非零坐标 β_j。即使在 $p > n$ 的情形下该方法也是可行的。

将 lasso 方法应用到第 7 节的超新星数据中，其中 $\mathbf{x}$ 的 $n = 75$ 以及 $p = 25$。图 15 展示了在前六步中非零系数 $\hat{\beta}_j$ 被逐步添加进来的过程。其中预测变量 15 首先被选取，接下来是变量 16、18、22、8 和 6，在第 25 步得到完整的 OLS 解 $\hat{\beta}$。基于某个精度公式表明第 4 步为最优的拟合结果，此时只有系数 15，16，18 和 22 是非零的。（这些对应于光谱中铁元素部分的能量测量值。）

稀疏性和 lasso 将我们带向与纯预测算法相反的方向。lasso 的推断不是结合无数的弱预测变量，而是基于几个最强的解释变量。这很适合归因但不太适合预测。

接下来将实现 lasso 的 R 程序 `glmnet` 应用到第 4 节的前列腺癌症预测问题中，使用与图 5 相同的训练/测试分割方法。它的表现比 `randomForest` 差得多，在测试集上有 13 个预测错误。然而，当应用到 102 名男性的整个数据集上时，`glmnet` 给出了重要基因的有用指示：图 14 中的方块展示了它认为最有影响力的 29 个基因的 z 值。这些基因具有较大的 $|z_i|$ 值，即使算法事并先不“知道”在癌症组和对照组之间得到的 t 统计量。

lasso 产生了有偏差的 β 估计，因为坐标值 $\hat{\beta}_j$ 向零收缩。对预测方法的批评也适用于此：有偏差的估计还没有坚实的理论基础。

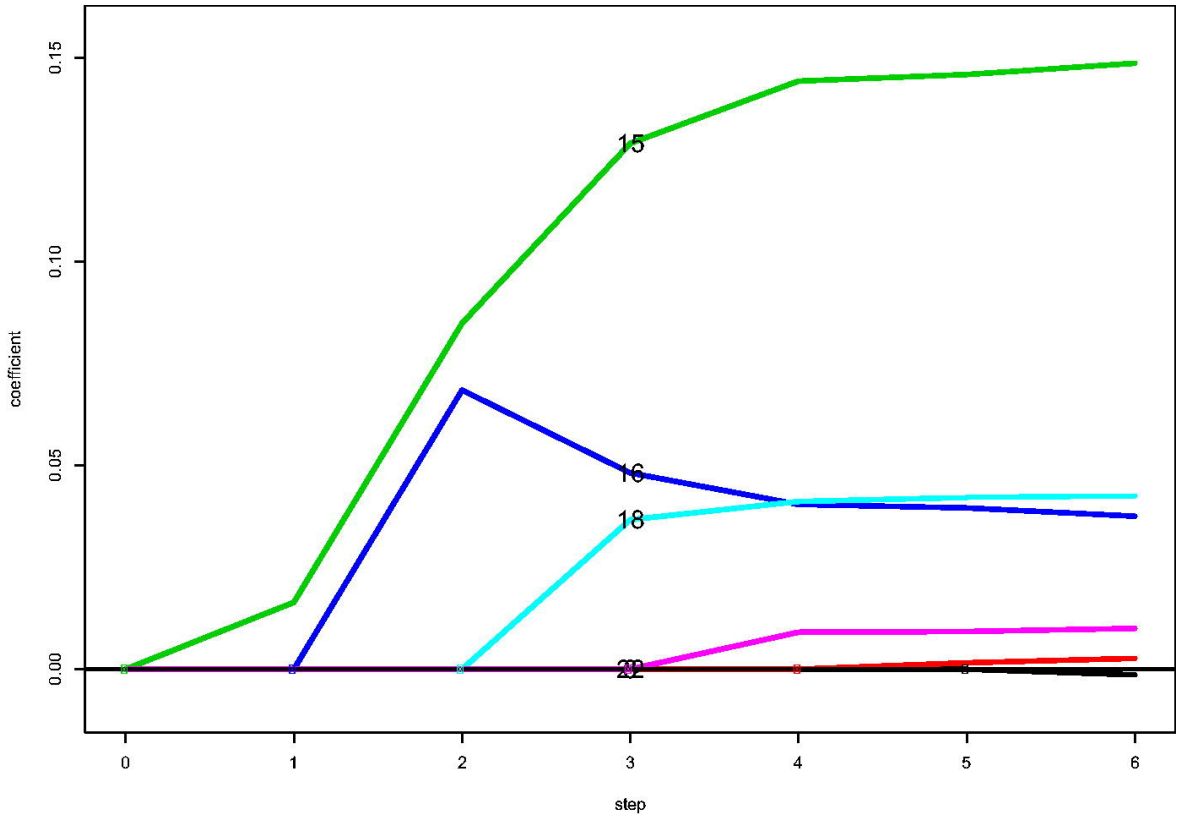

图 15: lasso 算法应用于超新星数据的前 6 步结果；各种预测变量的系数被绘制成步长的函数。预测变量 15 首先被选择，随后是 16、18、22、8 和 6。在第 4 步停止可得到估计误差的最低 Cp 估计值。

10 两个充满希望的趋势

这篇文章并不是在讲一个“皇帝没有衣服”的故事，而是说“皇帝有很好的衣服，但并不是适合所有场合。”在合适的情况下，纯预测算法可以取得惊人的成功。当人们在媒体上读到一篇关于人工智能充满热情的报道时，通常会有这样一个算法，以巨大的规模运行，做着繁重的工作。自高斯时代以来，回归方法已取得了长足的进展。

本文的大部分内容都是关于预测算法不能做到的事情，至少在他们目前的形式下是无法做到的。他们复杂的“黑箱”特性使得算法很难被评判。本文尝试使用相对较小的数据集（根据预测文献的标准）来说明他们与传统的估计和归因方法的差异。对于预测界来说大多数的批评都不足为奇，本文将这些批评归纳为第 8 节中表 5 的六个准则。

大约在 2000 年的某个时候，统计学界出现了分歧[28]。这里的讨论中，我们可以将这两个分支称为“纯预测”和“GWAS”：两者都适用于巨大的数据集，但前者已经完全算法化而后者则停留在更传统的数学建模路径上。本节标题中的“两个充满希望的趋势”是指统一的尝试，

[28]参见 Efron and Hastie (2016) 后记中的三角图。

诚然，统一的尝试并没有走得很远。

趋势 1 的目标是使预测算法的输出更具可解释性，即更像传统统计方法的输出。可解释的表面特别是线性模型表面，是实现这一成就的理想之选。尽管通常不是在统计显著性的特定意义上，归因之类的目标也是需要的。

一种策略是使用传统方法来分析预测算法的输出；见 Hara and Hayashi (2016) 以及 Efron and Hastie (2016) 的 346 页。Wager et al. (2014) 使用 bootstrap 和 jackknife 的思想提出了随机森林预测的标准误差的计算方法。Murdoch et al. (2019) 以及 Vellido et al. (2012) 对可解释性进行了概述，尽管两者都没有重点介绍纯预测算法。Achille and Soatto (2018) 利用信息论的思想，讨论了预测算法在统计上的充分性度量。

从另一个方向来看，趋势 2 为从表 5 的左边移动到右边，希望能够在传统框架内至少实现一些预测算法的优势。一个明显的目标是大规模运算。Qian et al. (2019) 提供了一个 $n = 500,000$ 以及 $p = 800,000$ 的 `glmnet` 示例。Hastie et al. (2009) 成功地将 boosting 与逻辑回归联系在一起。逻辑回归模型是预测界最常用的传统参数模型，因此有理由期待在该领域取得统一进展。

对于本节的标题来说，“有雄心的”可能比“充满希望的”更准确。表 5 中所看到的鸿沟很大，而如果有统一的计划也只是在进行中。在我看来，这些障碍都是理论上的。极大似然理论提供了一个估计精度的下界，以及接近实现它的实用方法。关于预测，我们能说些什么呢？“通用任务框架”通常只能展示出竞争者之间错误率的微小差异，但无法知道是否有其他算法会做得更好。简而言之，我们没有预测的最优理论。

这些天来，无论是关于统计学还是生物统计学，我听到的演讲都充满了对预测算法的活力和兴趣。目前算法的发展大多都来自统计学科之外，但我相信未来的进步特别是在科学应用方面的进步，将在很大程度上取决于我们。