

资讯

指数

活动

创投罗盘

综合 | 专家 | 专访 | 学堂 | 报告 | 案例 | **大数据+** | 能源 | 环保 | 营销 | 金融 | 征信 | 医疗 | 零售 | 交通 | 通信 | 互娱 | 农业 | 政府

## 大数据如何保持正轨？揭秘10个大数据神话

36大数据 | 2017-08-31 11:38

大数据 正轨 神话

【数据猿导读】也许对大数据更好的一个类比是它就像一匹意气风发的冠军赛马: 通过适当的训练和天赋的骑师, 良种赛马可以创造马场记录-但没有训练和骑手, 这个强大的动物根本连起跑门都进不了。



如果数据有一点点就不错了, 那么数据是海量的话就一定棒极了, 对不对 这就好比说, 如果一个炎日夏日里的微风让你感觉凉爽, 那么你会为一阵一阵的凉风感到欣喜若狂。以下为译文:

也许对大数据更好的一个类比是它就像一匹意气风发的冠军赛马: 通过适当的训练和天赋的骑师, 良种赛马可以创造马场记录-但没有训练和骑手, 这个强大的动物根本连起跑门都进不了。

为了确保你组织的大数据计划保持正轨, 你需要消除以下10种常见的误解。

### 1. 大数据就是‘很多数据’

大数据从其核心来讲, 它描述了结构化或非结构化数据如何结合社交媒体分析, 物联网的数据和其他外部来源, 来讲述一个“更大的故事”。该故事可能是一个组织运营的宏观描述, 或者是无法用传统的分析方法捕获的大局观。从情报收集的角度来看, 其所涉及的数据的大小是微不足道的。

### 2. 大数据必须非常干净

在商业分析的世界里, 没有“太快”之类的东西。相反, 在IT世界里, 没有“进垃圾, 出金子”这样的东西, 你的数据有多干净 一种方法是运行你的分析应用程序, 它可以识别数据集中的弱点。一旦这些弱点得到解决, 再次运行分析以突出“清理过的”区域。

### 3. 所有人类分析人员会被机器算法取代

数据科学家的建议并不总是被前线的业务经理们执行。行业高管Arijit Sengupta在 TechRepublic 的一篇文章中指出, 这些建议往往比科学项目更难实施。然而, 过分依赖机器学习算法也同样具有挑战性。Sengupta说, 机器算法告诉你该怎么做, 但它们没有解释你为什么这么做。这使得很难将数据分析与公司战略规划的其余部分结合起来。

## 精品栏目

- [2017/07/27] 大数据24小时 [More>](#)
- [2017/07/24-28] 大数据周周看 [More>](#)
- [2017/07/24-28] 大数据投融资 [More>](#)
- [2017/07/24-28] 大咖周语录 [More>](#)
- [2017/07/24-28] 大数据周聘汇 [More>](#)
- [2017/07/24-28] 每周一本书 [More>](#)
- [2016/08-10] 大数据活动公告 [More>](#)

## 专家推荐

[More >](#)



涂子沛



傅志华



马亮



苏萌



崔晓波



韩涵



车品觉



刘雷鸣



董飞



郭伟



张涵诚



陈运文

## 人物专访

[More >](#)




曾被167家VC拒绝, 如今公司估值百

## 活动推荐

[More >](#)

- ACS 2017中国汽车CIO峰会10 2017-10-25
- 2017金融科技价值—数据驱动 2017-10-19

 TOP PREDICTION ALGORITHMS				
TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	The "best fit" line through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
	Logistic regression	The adaptation of linear regression to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a branching method to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but by combining them we get better overall performance.	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
	Gradient Boosting	Uses even weaker decision trees, that are increasingly focused on "hard" examples.	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Neural networks	Neural networks	Mimics the behavior of the brain. Neural networks are interconnected neurons that pass messages to each other. Deep learning uses several layers of neural networks put one after the other.	Can handle extremely complex tasks - no other algorithm comes close in image recognition.	X Very, very slow to train, because they have so many layers. Require a lot of power. X Almost impossible to understand predictions.

©2017 Dataiku, Inc. | www.dataiku.com | contact@dataiku.com | @dataiku

预测算法的范围从相对简单的线性算法到更复杂的基于树的算法，最后是极其复杂的神经网络。

来源：dataiku，dataconomy。

#### 4.数据湖是必须的

据丰田研究所数据科学家Jim Adler说，巨量存储库，一些IT经理们设想用它来存储大量结构化和非结构化数据，根本就不存在。企业机构不会不加区分地将所有数据存放到一个共享池中。Adler说，这些数据是“精心规划”的，存储于独立的部门数据库中，鼓励“专注”的专业知识”。这是实现合规和其他治理要求所需的透明度和问责制的唯一途径。

#### 5.算法是万无一失的预言家

不久前，谷歌流感趋势项目被大肆炒作，声称比美国疾病控制中心和其他健康信息服务机构更快、更准确地预测流感疫情的发生地。正如《纽约客》的Michele Nijhuis在2017年6月3日的文章中所写的那样，人们认为与流感有关词语的搜索会准确地预测疫情即将爆发的地区。事实上，简单地绘制本地温度是一个更准确的预测方法。

谷歌的流感预测算法陷入了一个常见的大数据陷阱——它产生了无意义的相关性，比如将高中篮球比赛和流感爆发联系起来，因为两者都发生在冬季。当数据挖掘在一组海量数据上运行时，它更可能发现具有统计意义而非实际意义的信息之间的关系。一个例子是将缅因州的离婚率与美国人均人造黄油的消费量挂钩：尽管没有任何现实意义，但这两个数字之间确实存在“统计上显著”的关系。

#### 6.你不能在虚拟化基础架构上运行大数据应用

2017第二届中国国际大数据产	2017-08-17
GIEC2017全球互联网经济大会	2017-08-08
2017年第二届上海大数据与分	2017-08-01

### 不容错过的资讯

- 1 大数据24小时：Facebook“神童”跳槽谷歌
- 2 金融科技&大数据产品推荐：神策分析—
- 3 四部委审查微信淘宝隐私条款，互联网企
- 4 分享：解析6个公司的大数据岗位的面试:
- 5 小白做数据分析的一点感悟
- 6 如何将数据可视化技术应用于广告投放？
- 7 机器人即将抢走你的工作？数据表明你可
- 8 如何让大数据分析更有效？这里有5种技:
- 9 大数据是什么？一文秒读懂大数据
- 10 走进大数据院：当大数据成为思维习惯时

### 大数据学堂

More >



### 大数据企业推荐

More >

	九次方   贡献中国数据智慧
	星图数据   Data turn biz
	晶赞科技   数据推动产业智能化
	TalkingData   移动·数据·价
	百分点   大数据践行者

### 热门职位

More >

大约10年前，当“大数据”首次出现在人们眼前时，它就是Apache hadoop的代名词。就像VMware的Justin Murray在 2017年5月12日的文章 中所写的，大数据这一术语现在包括一系列技术，从NoSQL(MongoDB，Apache Cassandra)到Apache Spark。

此前，批评者们质疑Hadoop在虚拟机上的性能，但Murray指出，Hadoop在虚拟机上的性能与物理机相当，而且它能更有效地利用集群资源。Murray还炮轰了一种误解，即认为虚拟机的基本特性需要存储区域网络(SAN)。实际上，供应商们经常推荐直接连接存储，这提供了更好的性能和更低的成本。

7.机器学习是人工智能的同义词

一个识别大量数据中模式的算法和一个能够根据数据模式得出逻辑结论的方法之间的差距更像是一个鸿沟。ITProPortal 的Vineet Jain在 2017年5月26日的文章 中写道，机器学习使用统计解释来生成预测模型。这是算法背后的技术，它可以根据一个人过去的购买记录来预测他可能购买什么，或者根据他们的听歌历史来预测他们喜欢的音乐。

虽然这些算法很聪明，但它们远远不能达到人工智能的目的，即复制人类的决策过程。基于统计的预测缺乏人类的推理、判断和想象力。从这个意义上说，机器学习可能被认为是真正AI的必要先导。即使是迄今为止最复杂的AI 系统，比如 IBM沃森，也无法提供人类数据科学家所提供的大数据的洞察力。

8.大多数大数据项目至少实现了一半的目标

IT经理们知道没有数据分析项目是100%成功的。当这些项目涉及大数据时，成功率就会直线下降，NewVantage Partners最近的调查结果显示了这一点。在过去的五年中，95%的企业领导人表示，他们的公司参与了一个大数据项目，但只有48.4%的项目取得了“可衡量的结果”。

NewVantage Partners的大数据执行调查显示，只有不到一半的大数据项目实现了目标，而“文化”变化是最难实现的。资料来源: Data Informed 。

事实上，根据2016年10月发布的 Gartner的研究结果，大数据项目很少能跨过试验阶段。Gartner的调查发现，只有15%的大数据实现被部署到生产中，与去年调查报告的14%的成功率相对持平。

9.大数据的增长将减少对数据工程师的需求

如果你公司大数据计划的目标是尽量减少对数据科学家的需求，你可能会得到令人不快的惊喜。2017 Robert Half 技术薪资指南 指出，数据工程师的年薪平均跃升到13万美元和19.6万美元之间，而数据科学家的薪资目前平均在11.6万美元和16.3万美元之间，而商业情报分析员的薪资目前平均在11.8万美元到13.875万美元之间。

10.员工和一线经理将张开双臂拥抱大数据

NewVantage Partners的调查发现，85.5%的公司都致力于创建一个“数据驱动的文化”。然而，新的数据计划的整体成功率仅为37.1%。这些公司最常提到的三个障碍是缺乏组织一致性(42.6%)，缺乏中层管理人员的采纳和理解(41%)，以及业务阻力或缺乏理解(41%)。

未来可能属于大数据，但获得这一技术的好处需要大量的针对多样人性的老式辛勤工作。

来源：36大数据

收藏

分享











**声明：**数据猿尊重媒体行业规范，相关内容都会注明来源与作者；转载我们原创内容时，也请务必注明“来源：数据猿”与作者名称，否则将会受到数据猿追责。

北京 | 数据堂 大数据架构师&大数据分析

北京 | 聚合数据 业务拓展经理&JAVA工程

北京 | 慧米数据 三个职位

湖南 | 银杏数据科技有限公司 数据工程师

北京 | 中献电子技术开发中心 大数据分析

大家都在搜

百度	融资	漏洞	互联网
追随	机器人	大数据	
云计算	北京	小米	创业
春节	互联网+	物联网	
大数据应用	陕西	阿里巴巴	
机器学习	大数据	中国	
python	营销	数据分析	
医疗	科技部	投资	金融
数据挖掘	人工智能	电商	



走进大数据院：当大数据成为思维习惯时 产业发展才算成熟



金融科技&大数据产品推荐：神策分析——可私有化部署的用户行为...



神策分析  
驱动企业决策和产品智能  
可以私有化部署的用户行为分析平台

我要评论

我想要评论.....

提交评论

ok

热点导航

- |         |         |
|---------|---------|
| 大数据人物专访 | 大数据活动推荐 |
| 大数据学堂   | 商业智能    |
| 互联网广告   | 央行征信    |
| 检察系统    | 人工智能    |
| 内容为王    | 宽带资本    |
| 硅谷大数据   | 通联数据    |
| 京东金融    | 二次元大数据  |
| 信息安全    | 大数据风控   |
| 大数据研究基地 | 原生数据    |
| 大数据地形图  | 位置大数据   |
| 互联网     | 人工智能    |
| 大数据技术   | 快递      |
| 大数据入门   | 网站地图    |
| 大数据物流   |         |

关于数据猿  
成为专栏专家  
好文投递&寻求报道  
广告推广与活动合作  
数据支持&合作



数据合作伙伴：

