# FRONTIERS IN MASSIVE DATA ANALYSIS

# 海量数据分析前沿

National Research Council of the National Academies
美国国家学术院国家研究委员会 组编
华东师范大学数据科学与工程研究院 译

清华大学出版社

# FRONTIERS IN MASSIVE DATA ANALYSIS

# 海量数据分析前沿

美国国家学术院国家研究委员会 组编

华东师范大学数据科学与工程研究院 译



清华大学出版社

## 内 容 简 介

近年来，大数据成为学术界和工业界的热点，其本质就是海量数据分析。海量数据的来源包括互联网、传感器、生产生活、科学观测、科学实验等。海量数据分析不仅可以帮助人们取得新的科学发现，也可以推动技术的适应性、个性化和健壮性方面的进步。海量数据分析是一个跨学科的研究领域，理解本书的内容需要具备计算机科学、统计学和优化理论的基础知识。本书从计算和推理的角度分析了与海量数据分析相关的前沿问题，重点介绍海量数据挖掘分析以及流数据挖掘的进展，讨论了并行和分布式系统架构方面的最新发展，具体内容包括数据建模、任务建模、计算复杂性问题分析、数据采样以及人工参与的数据分析方法等。

本书对于指导我国从事数据科学研究的科技人员制订计划和开展研究具有参考价值。



北京市版权局著作权合同登记号　图字：01-2014-1704



**图书在版编目(CIP)数据**

海量数据分析前沿/美国国家学术院国家研究委员会组编；华东师范大学数据科学与工程研究院译. --北京：清华大学出版社，2015

书名原文：Frontiers in massive data analysis

ISBN 978-7-302-39547-8

Ⅰ.①海…　Ⅱ.①美…②华…　Ⅲ.①数据处理　Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 041384 号

**责任编辑**：薛　慧
**封面设计**：何凤霞
**责任校对**：刘玉霞
**责任印制**：宋　林
**出版发行**：清华大学出版社
**网　址**：http://www.tup.com.cn，http://www.wqbook.com
**地　址**：北京清华大学学研大厦 A 座　**邮　编**：100084
**社总机**：010-62770175　**邮　购**：010-62786544
**投稿与读者服务**：010-62776969，c-service@tup.tsinghua.edu.cn
**质　量　反　馈**：010-62772015，zhiliang@tup.tsinghua.edu.cn
**印 刷 者**：北京富博印刷有限公司
**装 订 者**：北京市密云县京文制本装订厂
**经　销**：全国新华书店
**开　本**：175mm×245mm　**印　张**：13.5　**字　数**：169 千字
**版　次**：2015 年 5 月第 1 版　**印　次**：2015 年 5 月第1 次印刷
**印　数**：1～2500
**定　价**：39.00 元

产品编号：059326-01

# 译 者 序

这是一本值得一读的书，我第一次读到它的时候就想把它介绍给大家。由于各种原因，这本书的中文译本的出版比我们期望的晚了一年多，但我相信它仍不过时，还是值得认真拜读。我是在 2013 年 11 月 13 日下午陪同国家自然科学基金委员会信息学部大数据考察团访问加州大学伯克利分校时获赠这本书的，伯克利 AMP 实验室主任 Michael J. Franklin 在介绍美国大数据研究计划以及他们实验室的工作之后把这本他参与撰写的刚刚出版的书送给了我们。我们考察团一行四人，包括基金委信息学部常务副主任秦玉文教授、计算机处处长刘克教授、华东师范大学何晓丰教授。考察的首站是硅谷，我们拜访了位于 Mountain View 的微软硅谷研究院搜索实验室的 Rakesh Agrawal 博士和位于 Palo Alto 的 SAP 美国总部的 Dina Bitton 博士和 Ming-Chien Shan 博士。访问伯克利是我们大数据考察的重要一站，2013 年 11 月 13 日上午，我们一行受到劳伦斯伯克利国家实验室（LBNL）常务副主任 Horst Simon 的热情接待。LBNL 不仅是世界上第一个加速器的诞生地，也是科学数据管理的发祥地。来自 LNBL 的科学家们向我们介绍了他们在科学计算、科学数据管理、可视化和可视分析等方面的工作，让我们领略了他们在科学数据管理和分析方面源远流长的历史和做出的卓越贡献。加州大学伯克利分校 AMP 实验室是受美国大数据研究计划资助成立的。AMP 实验室主任 Michael J. Franklin 教授 2013 年上半年受邀在华东师范大学进行学术休假访问，访问期间他两次返回美国华盛顿就是为了讨论本书的撰写和定稿，他在学术报告和学术交流中，多次提到这本书，很令我们期待。

“大数据”无疑是近几年最热的一个科技术语。据2012年12月4日美国《时代》周刊网站报道，在美国的2012年十大流行词评比中，“大数据”名列第二，排第一的是美国人当年最为关心的政治事件“财政悬崖”。在IT领域，“大数据”是继高性能计算机、互联网、网格计算、云计算之后的又一被大众所关注的技术术语。从某种意义上讲，“大数据”已经远远超出了技术范畴，变成一个被赋予各种解读的流行词。“大数据”在我国的热度还在持续上升，只是从今年两会以后稍稍让位于“互联网＋”。正是因为“大数据”这个词的含义太过宽泛，各人可以有自己的一套解读方式。在不少场合听到过各种有关“大数据”的报告，一个普遍的情况是：报告的大数据应用大多不是报告人熟悉的领域。似乎印证了那句话“互联网企业做大数据，做的不说，说的不做”。

实际上，虽然互联网是推动大数据热的始作俑者，但广泛来说，大数据不仅仅局限于互联网数据。要讨论这林林总总的数据，从认识论的观点来看，首先就是要对大数据进行分类，这非常必要，是确保大家在同一论域进行讨论的前提。按照我的理解，大数据大致可以分为Web数据、决策数据、科学数据三大类。顾名思义，Web数据是与Web相关的数据，包括网页、链接、日志等具体类型，门户网站、搜索引擎、社交网络、电子商务等以Web形式呈现或以Web为载体的新型信息服务系统产生的数据大多可以归纳为此类型。决策数据主要指以前由传统的数据库和数据仓库管理的，在生产过程中产生的数据，是用于决策的，也可称为商务智能(BI)数据。科学数据实际上是最早的一类大数据，包括科学实验数据、科学观测数据、科学文献数据、设计数据等，这类数据与科学领域密切相关，品种最多，研究最难，没有领域专家的参与IT专家难以胜任科学数据的管理和分析任务。

以上是大数据类型的一个划分，关于大数据研究的认识，我也有一个分三个层次的观点。大数据的研究全景可以看作是一个倒

立的三角形。这个倒立三角形分为三层，最上面一层，也就是最宽的那一层，代表形形色色的各种应用，这些应用是数据的来源也是数据的应用场所；最底下的一层，也就是那个小三角形，就代表 IT 计算系统或平台，这是传统信息技术行业关心和擅长的领域；中间那一层代表模型和算法，指的就是对应用进行理解、抽象、建模，然后在底层的计算平台上予以实现。我读这本书，就是按照这三个层次来理解的。这也是我喜欢这本书的一个原因。这三个层次中，应用这一层，每一类应用有各自对应的学科去深入研究；计算平台那一层对应的学科就是我们计算机或 IT 学科。关于这两层，本书的第二、第三章以及其他部分章节有所涉及。本书的主要章节讨论的内容都是和第二层模型和算法相关的。

按照本书的观点，大数据的本质就是海量数据分析。海量数据的来源包括互联网、传感器、生产生活、科学观测、科学实验等。海量数据分析不仅可以帮助人们获得新的科学发现，也可以推动技术在适应性、个性化和健壮性方面的进步。海量数据分析是个跨学科的研究领域，理解本书的内容需要具备计算机科学、统计学和优化理论的基础知识。本书从计算和推理的角度分析了与海量数据分析相关的前沿问题，重点介绍海量数据挖掘分析以及流数据挖掘的进展，讨论了并行和分布式系统架构方面最新发展，具体内容包括数据建模、任务建模、计算复杂性问题分析、数据采样以及人工参与的数据分析方法等。

本书是由美国国家科学院、美国国家工程院和医学科学研究院的运营机构——美国国家研究委员会下属的海量数据分析委员会、应用和理论统计委员会、数学科学及其应用委员会、工程和物理科学部组织编写的。项目得到了美国国家安全局的支持，全美多个领域七八十位国际顶级专家参与了本书的撰写或评审工作。本书的中文翻译得到了清华大学出版社的大力支持，华东师范大学数据科学与工程研究院的周傲英教授、何晓丰教授、周敏奇副教授、金澈清

教授、王晓玲教授、王长波教授、钱卫宁教授、宫学庆教授、张蓉副教授、张召副教授、高明副教授，以及云南大学的岳昆教授和复旦大学的沙朝锋副教授参与了本书的翻译和校对。由于本书涉及的学科领域广泛，参与翻译的人员较多，再加上译者水平有限，如有翻译不准确甚至错误之处，敬请读者谅解并给予指正。

华东师范大学数据科学与工程研究院
周傲英
2015 年 4 月 13 日

# 美国国家学术院

## ——全国科学、工程和医学咨询机构

美国国家科学院由从事科学和工程研究的杰出学者组成，是投身于科学和技术发展、致力于造福人类的民间、非营利性、自组织的团体。该组织于1863年由国会特许授权组建而成，其职责是在科学和技术方面为联邦政府建言献策。Ralph J. Cicerone博士是美国国家科学院现任主席。

美国国家工程院是于1964年在美国国家科学院特许下成立的由杰出工程师代表组成的并行组织，日常管理和院士选举都自主进行，与美国国家科学院一起为联邦政府建言献策。美国国家工程院的职责还包括：资助满足国家需求的工程项目、鼓励教育和研究以及发现具有杰出成就的工程师。现任主席是C. D. Mote博士。

美国国家医学院由美国国家科学院于1970年组建，致力于在与公共卫生有关的政策审查方面提供相关专业专家的服务。该组织的职责由美国国家科学院宪章规定，并向联邦政府建言献策，主动确定医疗保健、医学研究和教育中存在的问题。现任主席是Harvey V. Fineberg博士。

美国国家研究委员会由美国国家科学院于1916年组建，其宗旨是联合广泛的科学和技术团体和美国国家科学院一起推动知识发现的进程，并为联邦政府提供咨询服务。根据美国国家学术院的政策，该委员会已成为美国国家科学院和美国国家工程院的主要运行机构，为政府、公众、科学和技术团体提供服务。委员会由美国国家科学院、美国国家工程院和美国国家医学院共同管理。Ralph J. Cicerone博士和C. D. Mote博士分别是委员会现任主席和副主席。

——引自 www.national-academies.org

# 海量数据分析委员会成员

MICHAEL I. JORDAN，加利福尼亚大学（伯克利）（主席）
KATHLEEN M. CARLEY，卡内基梅隆大学
RONALD R. COIFMAN，耶鲁大学
DANIEL J. CRICHTON，喷气推进实验室
MICHAEL J. FRANKLIN，加利福尼亚大学（伯克利）
ANNA C. GILBERT，密歇根大学
ALEX G. GRAY，佐治亚理工学院
TREVOR J. HASTIE，斯坦福大学
PIOTR INDYK，麻省理工学院
THEODORE JOHNSON，AT&T 实验室
DIANE LAMBERT，谷歌公司
DAVID MADIGAN，哥伦比亚大学
MICHAEL W. MAHONEY，斯坦福大学
F. MILLER MALEY，国防分析研究所
CHRISTOPHER OLSTON，谷歌公司
YORAM SINGER，谷歌公司
ALEXANDER SANDOR SZALAY，约翰·霍普金斯大学
TONG ZHANG，罗格斯，新泽西州立大学

**工作人员**

SUBHASH KUVELKER，研究中心主任（2011 年 10 月 17 日前）
SCOTT WEIDMAN，研究中心主任（2011 年 10 月 17 日后）
BARBARA WRIGHT，行政助理

# 应用和理论统计委员会成员

CONSTANTINE GATSONIS，布朗大学（主席）
MONTSERRAT FUENTES，北卡罗莱纳州立大学
ALFRED O. HERO Ⅲ，密歇根大学
DAVID M. HIGDON，洛斯阿拉莫斯国家实验室
IAIN JOHNSTONE，斯坦福大学
ROBERT E. KASS，卡内基梅隆大学
JOHN LAFFERTY，芝加哥大学
XIHONG LIN，哈佛大学
SHARON-LISE T. NORMAND，哈佛医学院
GIOVANNI PARMIGIANI，达纳法伯癌症研究所
RAGHU RAMAKRISHNAN，微软公司
ERNEST SEGLIE，国防部长办公室（已退休）
LANCE WALLER，埃默里大学
EUGENE WONG，加利福尼亚大学（伯克利）

**工作人员**

MICHELLE SCHWALBE，主任
BARBARA WRIGHT，行政助理

# 数学科学及其应用委员会成员



**工作人员**

# 致　谢

经美国国家研究委员会报告审议小组同意，已经邀请来自不同领域的技术专家对报告草案进行评审。该次独立评审的目的在于提供坦率和批判性的意见以帮助委员会发布尽可能准确的报告，确保报告符合委员会客观性、实证性和研究导向的标准。为维护审议过程的完整性，评阅意见和草案依然处于保密状态。我们由衷地感谢以下专家对本报告的审阅：

Amy Braverman，喷气推进实验室

John Bruning，康宁-特罗佩尔公司（已退休）

Jeffrey Hammerbacher，Cloudera 公司

Iain Johnstone，斯坦福大学

Larry Lake，得克萨斯大学

Richard Sites，谷歌公司

Hal Stern，加利福尼亚大学（尔湾）

虽然上述评审人提供了许多建设性的意见和建议，但并不要求他们同意报告的结论或建议。而且在报告发布前，他们也没有看到本报告的最终稿。本报告由加利福尼亚州圣巴巴拉大学的 Michael Goodchild 监督评阅，他（她）由美国国家研究委员会任命，负责按照委员会规定的程序组织对本报告的独立评审，并合理考虑所有评审人的意见。因本报告最终内容所引发的责任由写作委员会和本机构承担。

委员会还要感谢在会议或其他交流过程中提出宝贵建议的以下专家：

Léon Bottou，NEC 实验室

Jeffrey Dean，谷歌公司

John Gilbert，加利福尼亚大学（圣巴巴拉）

Jeffrey Hammerbacher，Cloudera 公司

Patrick Hanrahan，斯坦福大学

S. Muthu Muthukrishnan，罗格斯新泽西州立大学

Ben Shneiderman，马里兰大学

Michael Stonebraker，麻省理工学院

J. Anthony Tyson，加利福尼亚大学（戴维斯）

# 目　录

# 概　要

## 海量数据的机遇与挑战

在科学和商业的诸多领域中，实验、观测和数值仿真可产生 TB 量级的数据，在某些领域达到 PB 量级甚至更多。基于此种数据规模的信息分析促进了从基因组学、天文学、高能物理，到新兴信息产业等领域的重大突破。传统的分析方法假设分析人员可以在自己的计算环境中处理数据，但大数据的出现改变了这一范式，特别是当海量数据分布在不同的物理位置上时。

虽然科学界和国防工业长期以来一直主导着海量数据集的生成和使用，但电子商务和大规模搜索引擎的兴起使得其他行业也同样面临着海量数据的挑战。例如，谷歌、雅虎、微软和其他互联网公司拥有 EB 量级（$10^{18}$ 字节）的数据，像 Facebook、YouTube 和 Twitter 等社交媒体数据的爆炸性增长超出了人们的想象，部分社交媒体已经拥有数亿用户。对这些海量数据的挖掘正在改变人们对危机公关、市场营销、娱乐、网络安全和国家情报等的处理方式。它也正在改变人们对信息存储和检索的认知。文件、图片、视频和网络等数据集合不仅仅是以二进制串的形式存储、索引和检索，它们也是科学发现和知识的潜在源泉，需要远比传统的索引和关键词计数复杂得多的分析技术来对数据中的现象进行关联和语义上