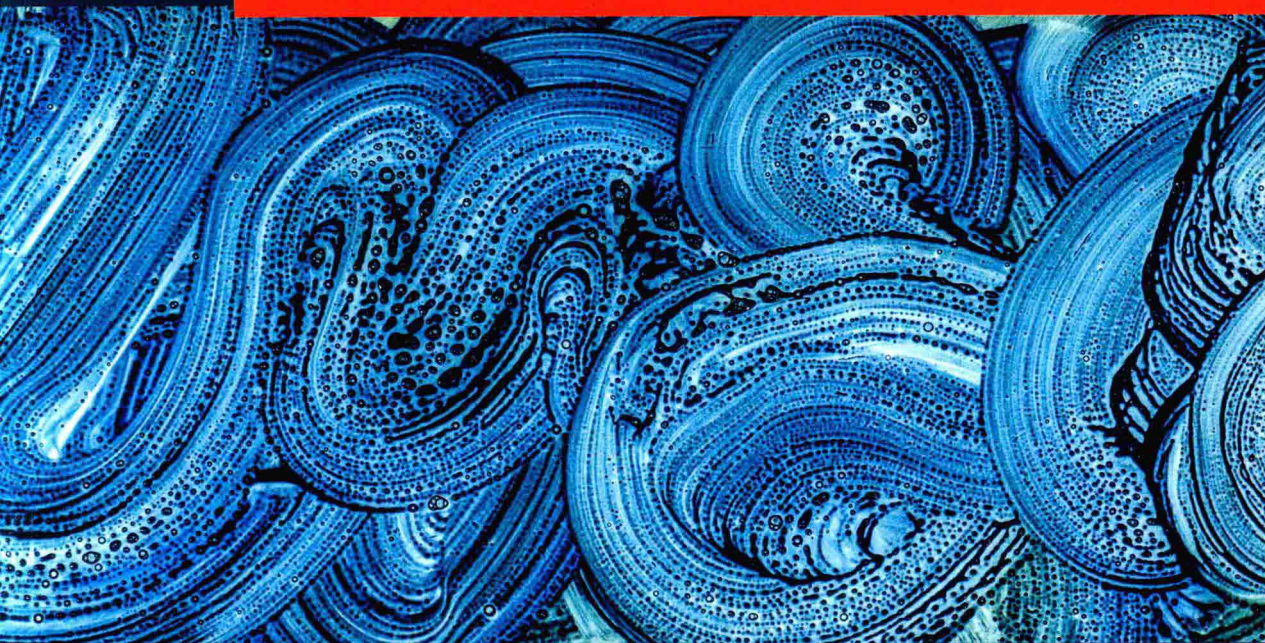


大数据时代的软件工程

软件科学家与数据科学家
的思维碰撞

Perspectives on Data Science for
Software Engineering



蒂姆·孟席斯 (Tim Menzies)

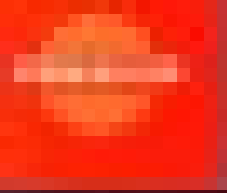
[美] 劳里·威廉姆斯 (Laurie Williams) 编著

托马斯·齐默尔曼 (Thomas Zimmermann)

王永吉 陈力 吕荫润 等译



机械工业出版社
China Machine Press



中国科学院
中国科学院



大数据时代的软件工程

中国科学院软件研究所
软件研究所

Forecasting and Data Science: The
Software Engineering



中国科学院软件研究所

中国科学院软件研究所

中国科学院软件研究所

中国科学院软件研究所

中国科学院软件研究所

华章程序员书库



Perspectives on Data Science for
Software Engineering

大数据时代的软件工程

软件科学家与数据科学家
的思维碰撞

蒂姆·孟席斯 (Tim Menzies)

[美]

劳里·威廉姆斯 (Laurie Williams)

编著

托马斯·齐默尔曼 (Thomas Zimmermann)

王永吉 陈力 吕荫润 王培霞 王翀 余海 游逸 张艺品 译



机械工业出版社
China Machine Press

图书在版编目(CIP)数据

大数据时代的软件工程：软件科学家与数据科学家的思维碰撞 / (美)蒂姆·孟席斯 (Tim Menzies) 等编著；王永吉等译. —北京：机械工业出版社，2018.1

(华章程序员书库)

书名原文：Perspectives on Data Science for Software Engineering

ISBN 978-7-111-58835-1

I. 大… II. ①蒂… ②王… III. 软件工程 - 研究 IV. TP311.5

中国版本图书馆 CIP 数据核字 (2017) 第 330930 号

本书版权登记号：图字 01-2017-2752

ELSEVIER

Elsevier (Singapore) Pte Ltd.

3 Killiney Road, #08-01 Winsland House I, Singapore 239519

Tel: (65) 6349-0200; Fax: (65) 6733-1817

Perspectives on Data Science for Software Engineering

Tim Menzies, Laurie Williams, Thomas Zimmermann

Copyright © 2016 by Elsevier Inc. All rights reserved.

ISBN-13: 978-0-12-804206-9

This translation of Perspectives on Data Science for Software Engineering by Tim Menzies, Laurie Williams, Thomas Zimmermann was undertaken by China Machine Press and is published by arrangement with Elsevier (Singapore) Pte Ltd.

Perspectives on Data Science for Software Engineering by Tim Menzies, Laurie Williams, Thomas Zimmermann 由机械工业出版社进行翻译，并根据机械工业出版社与爱思唯尔（新加坡）私人有限公司的协议约定出版。

《大数据时代的软件工程：软件科学家与数据科学家的思维碰撞》(王永吉 陈力 吕荫润 等译)

ISBN: 978-7-111-58835-1

Copyright © 2018 by Elsevier (Singapore) Pte Ltd.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from Elsevier (Singapore) Pte Ltd. Details on how to seek permission, further information about the Elsevier's permissions policies and arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by Elsevier (Singapore) Pte Ltd. and China Machine Press (other than as may be noted herein).

注意

本译本由 Elsevier (Singapore) Pte Ltd. 和机械工业出版社完成。相关从业及研究人员必须凭借其自身经验和知识对文中描述的信息数据、方法策略、搭配组合、实验操作进行评估和使用。由于医学科学发展迅速，临床诊断和给药剂量尤其需要经过独立验证。在法律允许的最大范围内，爱思唯尔、译文的原文作者、原文编辑及原文内容提供者均不对译文或因产品责任、疏忽或其他操作造成的人身及/或财产伤害及/或损失承担责任，亦不对由于使用文中提到的方法、产品、说明或思想而导致的人身及/或财产伤害及/或损失承担责任。

Printed in China by China Machine Press under special arrangement with Elsevier (Singapore) Pte Ltd. This edition is authorized for sale in the People's Republic of China only, excluding Hong Kong SAR, Macau SAR and Taiwan. Unauthorized export of this edition is a violation of the contract.

本书封底贴有 Elsevier 防伪标签，无标签者不得销售。

大数据时代的软件工程：软件科学家与数据科学家的思维碰撞

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：张志铭

责任校对：殷虹

印刷：北京兆成印刷有限公司

版次：2018 年 1 月第 1 版第 1 次印刷

开本：186mm × 240mm 1/16

印张：15.5

书号：ISBN 978-7-111-58835-1

定价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010) 88379426 88361066

投稿热线：(010) 88379604

购书热线：(010) 68326294 88379649 68995259

读者信箱：hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问：北京大成律师事务所 韩光/邹晓东

The Translator's Words 译者序

本书是 2014 年在德国 Dagstuhl 举行的“软件开发分析”研讨会文章合集。在该研讨会上，来自工业界以及学术界的软件开发领域的众多优秀专家齐聚一堂，围绕软件开发与数据科学的议题进行了充分的讨论。本书作为载体，记录了广泛的软件分析工作，同时也传达了与会者“时间有限，数据无限”的观点。

随着软件行业以及数据科学的发展，如何利用收集的数据来进行软件分析，成为软件工程行业的一大重要议题。近些年来，数据呈现爆炸式增长，我们已进入“大数据”时代，同时，基于人工智能、可视化技术等的数据分析工具也变得触手可及，因此现在软件分析领域面临的不是缺乏数据或者缺少数据分析工具和方法的问题，而是如何将这些工具作用于对应数据之上来进行软件分析的问题。

因此，至关重要的是从数据科学的视角来洞悉软件工程领域，更好地收集、整理软件开发管理过程中产生的软代码、Git 提交信息等数据，以及从现有种类繁多的数据分析工具和方法中挑选适合软件分析这一特定领域的工具和方法。

本书内容大致可以归为两类：软件工程的实战经验，这部分提供软件分析过程中的宝贵经验和具体数据科学方法在软件工程实践中存在的问题；数据分析的实战经验，这部分讨论了从日常实践数据中总结的结论，其中涉及的案例研究材料来自软件工程领域，这些材料也为其他领域的科学家提供了很多数据。

本书的每一章节都不是长篇大论，而是通过言简意赅的语言向读者普及软硬件技术。作者们通过成功案例研究、应用场景分析、技术介绍、经验教训总结等方式来呈现自己的见解。每一章节的作者并没有试图事无巨细地向读者介绍一项很具体的技术或者方法，而是通过简洁的语言向读者传达自己的见解。细心的读者可能会注意到，整本书几乎没有专业术语、公

式，甚至参考文献也很少。总之，本书不会给读者一种在研读科技论文的感觉，读者可以将本书当作一本科普读物进行阅读。

感谢中国科学院大学的李孟岳、李奉治、鲍鼎之、胡靖宇为本书的校对提供的帮助。感谢翻译过程中机械工业出版社的曲熠等人给予的帮助。感谢家人以及所有朋友一如既往的支持和帮助，正是由于你们的支持与帮助，本书的翻译工作才得以顺利进行。

由于译者水平有限，书中难免会有疏漏，还望读者不吝提出意见和建议。

About the Authors 本书作者

Bram Adams

Polytechnique Montreal, Canada

A. Bacchelli

Delft University of Technology, Delft, The Netherlands

T. Barik

North Carolina State University

E. T. Barr

University College London, London, United Kingdom

O. Baysal

Carleton University, Ottawa, ON, Canada

A. Bener

Ryerson University, Toronto ON, Canada

G. R. Bergersen

University of Oslo, Norway

C. Bird

Microsoft Research, Redmond, WA, United States

D. Budgen

Durham University, Durham, United Kingdom

B. Caglayan

Lero, University of Limerick, Ireland

T. Carnahan

Microsoft, Redmond, WA, United States

J. Czerwinka

Principal Architect Microsoft Corp., Redmond, WA, United States

P. Devanbu

UC Davis, Davis, CA, United States

M. Di Penta

University of Sannio, Benevento, Italy

S. Diehl

University of Trier, Trier, Germany

T. Dybå

SINTEF ICT, Trondheim, Norway

T. Fritz

University of Zurich, Zurich, CHE

M. W. Godfrey

University of Waterloo, Waterloo, ON, Canada

G. Gousios

Radboud University Nijmegen, Nijmegen, The Netherlands

P. Guo

University of Rochester, Rochester, NY, United States

K. Herzig

Software Development Engineer, Microsoft Corporation, Redmond, United States

A. Hindle

University of Alberta, Edmonton, AB, Canada

R. Holmes

University of British Columbia, Vancouver, Canada

Zhitao Hou

Microsoft Research, Beijing, China

J. Huang

Brown University, Providence, RI, United States

Andrew J. Ko

University of Washington, Seattle, WA, United States

N. Juristo

Universidad Politecnica de Madrid, Madrid, Spain

S. Just

Researcher, Software Engineering Chair & Center for IT-Security, Privacy and Accountability,
Saarland University, Germany

M. Kim

University of California, Los Angeles

E. Kocaguneli

Microsoft, Seattle, WA, United States

K. Kuutti

University of Oulu, Oulu, Finland

Qingwei Lin

Microsoft Research, Beijing, China

Jian-Guang Lou

Microsoft Research, Beijing, China

N. Medvidovic

University of Southern California, Los Angeles, CA, United States

A. Meneely

Rochester Institute of Technology, Rochester, NY, United States

T. Menzies

North Carolina State University, Raleigh, NC, United States

L. L. Minku

University of Leicester, Leicester, United Kingdom

A. Mockus

The University Of Tennessee, Knoxville, TN, United States

J. Münch

Reutlingen University, Reutlingen, Germany; University of Helsinki, Helsinki, Finland; Herman
Hollerith Center, Böblingen, Germany

G. C. Murphy

University of British Columbia, Vancouver, BC, Canada

B. Murphy

Microsoft Research, Cambridge, United Kingdom

E. Murphy-Hill

North Carolina State University

M. Nagappan

Rochester Institute of Technology, Rochester, NY, United States

M. Nayebi

University of Calgary, Calgary, AB, Canada

M. Oivo

University of Oulu, Oulu, Finland

A. Orso

Georgia Institute of Technology, Atlanta, GA, United States

T. Ostrand

Mälardalen University, Västerås, Sweden

F. Peters

Lero - The Irish Software Research Centre, University of Limerick, Limerick, Ireland

D. Posnett

University of California, Davis, CA, United States

L. Prechelt

Freie Universität Berlin, Berlin, Germany

Venkatesh-Prasad Ranganath

Kansas State University, Manhattan, KS, United States

B. Ray

University of Virginia, Charlottesville, VA, United States

R. Robbes

University of Chile, Santiago, Chile

P. Rotella

Cisco Systems, Inc., Raleigh, NC, United States

G. Ruhe

University of Calgary, Calgary, AB, Canada

P. Runeson

Lund University, Lund, Sweden

B. Russo

Software Engineering Research Group, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy

M. Shepperd

Brunel University London, Uxbridge, United Kingdom

E. Shihab

Concordia University, Montreal, QC, Canada

D. I. K. Sjøberg

University of Oslo; SINTEF ICT, Trondheim, Norway

D. Spinellis

Athens University of Economics and Business, Athens, Greece

M.-A. Storey

University of Victoria, Victoria, BC, Canada

C. Theisen

North Carolina State University, Raleigh, NC, United States

A. Tosun

Istanbul Technical University, Maslak Istanbul, Turkey

B. Turhan

University of Oulu, Oulu, Finland

H. Valdivia-Garcia

Rochester Institute of Technology, Rochester, NY, United States

S. Vegas

Universidad Politecnica de Madrid, Madrid, Spain

S. Wagner

University of Stuttgart, Stuttgart, Germany

E. Weyuker

Mälardalen University, Västerås, Sweden

J. Whitehead

University of California, Santa Cruz, CA, United States

L. Williams

North Carolina State University, Raleigh, NC, United States

Tao Xie

University of Illinois at Urbana-Champaign, Urbana, IL, United States

A. Zeller

Saarland University, Saarbrücken, Germany

Dongmei Zhang

Microsoft Research, Beijing, China

Hongyu Zhang

Microsoft Research, Beijing, China

Haidong Zhang

Microsoft Research, Beijing, China

T. Zimmermann

Microsoft Research, Redmond, WA, United States

Contents 目 录

译者序

本书作者

第1章 简介 1

- 1.1 软件工程的数据科学视角 1
- 1.2 软件分析与实际应用 4
- 1.3 归纳式软件工程的七个原则：
我们做的是不同的 7
- 1.4 软件工程中数据分析模式的需求 11
- 1.5 从软件数据到软件理论：到达的
捷径 14
- 1.6 为何理论是重要的 17

第2章 成功案例与应用 21

- 2.1 挖掘应用中的异常 21
- 2.2 拥抱动态工件 25
- 2.3 移动应用商店分析 28
- 2.4 软件的自然性 30
- 2.5 版本发布就绪的研究进展 34
- 2.6 如何征服你的在线服务 37

- 2.7 评价个人生产率 40
- 2.8 基于堆栈跟踪的攻击面检测 43
- 2.9 软件工程数据的视觉分析 46
- 2.10 游戏数据分组后效果更佳 49
- 2.11 实践中应用数据科学的成功
故事 52
- 2.12 从来没有足够的时间做所有
需要的测试 56
- 2.13 能源开采的危险：多次衡量，
一次比较 59
- 2.14 大规模商业软件中的错误文件
定位 63
- 2.15 定制组件：个性化问题跟踪的
机遇 66
- 2.16 至关重要的是决策而非数字：
分析设计表 68
- 2.17 编程语言对代码质量影响的
系统研究 72
- 2.18 代码审查不是为了寻找缺陷：
即使已建立的工具也需要偶尔
进行评估 74

第3章 技术 78

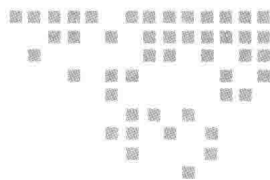
- 3.1 会谈 78
- 3.2 查找暂存数据中的状态转换 83
- 3.3 卡片分类：从文本到主题 85
- 3.4 工具！工具！我们需要工具 88
- 3.5 基于证据的软件工程 93
- 3.6 你需要哪种机器学习方法 96
- 3.7 首先结构化你的非结构化数据！用标签云概括非结构化数据的案例 99
- 3.8 解析数据！准备原始数据的实用技巧 105
- 3.9 自然语言处理不是免费的午餐 109
- 3.10 聚集经验证据进行更可信的决策 112
- 3.11 如果是软件工程，它可能是一个贝叶斯因子 117
- 3.12 隐私和数据共享都要有恰如其分的条件 120
- 3.13 软件工程可预测模型中的群体智慧 124
- 3.14 挖掘软件数据时兼顾定量方法和定性方法 128
- 3.15 为存活而调查设计的过程和为远航而调查部署的过程 132

第4章 智慧与技巧 138

- 4.1 记录一切？ 138
- 4.2 起源的重要性 140

- 4.3 从一开始就保持开放 144
- 4.4 减少洞察时间 147
- 4.5 获取成功的五步：如何在组织中部署数据科学 150
- 4.6 发布过程如何影响软件分析 153
- 4.7 安全无价 156
- 4.8 挖掘错误报告中的陷阱 159
- 4.9 使可视化成为分析过程的一部分 163
- 4.10 不要忘记开发者（小心你的假设） 165
- 4.11 研究的局限性和语境 168
- 4.12 可执行的指标才是更好的指标 171
- 4.13 可复制的结果更可靠 175
- 4.14 软件工程研究的多样性 178
- 4.15 一次还不够：为什么需要重复 181
- 4.16 不仅仅是数字：可视化的需求 184
- 4.17 不要使自己尴尬：小心数据中的偏差 187
- 4.18 操作数据丢失、错误和脱离语境 193
- 4.19 数据科学在改进和评估过程中演化？ 197
- 4.20 相关性不是因果关系（不要大叫“找到了！”） 199
- 4.21 小软件公司的软件分析：问题比答案更多 202

- 4.22 路灯下的软件分析（《星际迷航》教会我们提出正确问题的重要性）..... 205
- 4.23 软件工程实验中会出现哪些错误..... 208
- 4.24 同样的大小并不适合所有情况..... 211
- 4.25 对于好的模型，简单的解释更好..... 212
- 4.26 白衬衫效应：从失败的期望中学习..... 215
- 4.27 简单的问题能导致更好的见解..... 218
- 4.28 连续实验以尽早评估价值..... 222
- 4.29 谎言、可恶的谎言和分析：为什么大数据需要厚数据..... 225
- 4.30 整个世界都是你的测试套件..... 229



第 1 章 Chapter 1

简介

1.1 软件工程的数据科学视角

T. Menzies^{*}, L. Williams^{*}, T. Zimmermann[†]

North Carolina State University, Raleigh, NC, United States^{}*

Microsoft Research, Redmond, WA, United States[†]

为什么选择本书

本书的内容来自为期一周的德国 Dagstuhl 研讨会^[1]。该会议的目标是记录广泛的软件分析工作。该会议以“时间有限，数据无限”为前提。

也就是说，鉴于近期我们收集数据的增加，并且缺乏可以评估数据的熟练分析师^[2]，现在的数据远超人们所能研究的。因此，至少对此而言，现实世界数据的分析（使用半自动或全自动方法）是一个爆炸性的领域。这个问题由于两个因素而变得紧迫：

- **许多有用的方法：**人工智能、社会科学方法、可视化以及统计等长达数十年的研究已经产生了大量从数据中学习的有效方法。
- **这些方法得到大量支持：**标准教科书和教育课程探讨了这些方法。广泛使用的工具包也支持这些方法（有时甚至通过免费下载的方式）。此外，考虑到“大数据”革命，现在可以获得必要的硬件，即使是长期使用这些工具。所以现在问题不是“如何获取这些工具”，而是如何“使用”这些工具。

如果通用分析是一个活跃的领域，那么软件分析就是双重的。想想我们对软件项目的

认识：

- 源代码；
- 关于该代码的电子邮件；
- 签到；
- 工作项目；
- 错误报告；
- 测试套件；
- 测试执行；
- 甚至有关开发者的一些背景信息。

所有信息都记录在软件仓库中，如 CVS、Subversion、GIT、GITHUB 和 Bugzilla。遥测数据、运行时跟踪和日志文件可以在这些存储库中找到，这反映了客户如何体验软件，应用程序和功能使用情况，性能和可靠性记录等。

我们一直缺少有关人们如何合作的细节信息：

- 借助别人的见解和软件工具；
- 产生和分享新的见解和软件工具；
- 维护和更新现有的见解和软件工具。

在这里，借助“工具”，是指从浏览网页时触发的四行 SQL 语句到只有上百行代码的脚本或更大的开源及专有系统的所有内容。此外，我们使用的“工具”包括构建新工具以及持续的维护工作、硬件和软件系统的组合。

因此，考虑到你的需求，本书探讨了从软件开发应用程序数据分析到见解产生的过程。各章节由 Dagstuhl 研讨会的参与者（图 1-1）以及工业和学术数据挖掘领域的众多专家共同撰写。我们的目标是总结和分享他们的经验，并结合自身智慧表达对数据分析过程的理解。



图 1-1 Dagstuhl 研讨会 14261 上“软件开发分析”的参与者（2014 年 6 月 22 日至 27 日）