Data Science The Mit Press Essential Knowledge Series

Raw Data Is an OxymoronChild Data CitizenAll Data Are LocalData ActionBig Data, Little Data, No DataOptimization for Machine LearningMachine Learning for Data StreamsPredicting Structured DataComputational ThinkingSpatial ComputingData ScienceCultural AnalyticsIndexing It AllArtificial UnintelligenceToo SmartMathematics of Big DataFundamentals of Machine Learning for Predictive Data AnalyticsIntroduction to Natural Language ProcessingOpen SpaceTrusted DataAI EthicsAn Introduction to Statistical Genetic Data AnalysisIntroduction to Machine LearningCloud Computing for Machine Learning and Cognitive ApplicationsDeep LearningMetadataThe BookA Hands-On Introduction to Data ScienceBig Data Is Not a MonolithDevelopment of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language SciencesThe InfographicIntroduction to Computation and Programming Using PythonAnalyzing Neural Time Series DataDeep LearningData FeminismDecoding the Social WorldA Vast MachineInformation and the Modern CorporationMachine LearnersElements of Causal Inference

Raw Data Is an Oxymoron

An accessible synthesis of ethical issues raised by artificial intelligence that moves beyond hype and nightmare scenarios to address concrete questions. Artificial intelligence powers Google's search engine, enables Facebook to target advertising, and allows Alexa and Siri to do their jobs. AI is also behind self-driving cars, predictive policing, and autonomous weapons that can kill without human intervention. These and other AI applications raise complex ethical issues that are the subject of ongoing debate. This volume in the MIT Press Essential Knowledge series offers an accessible synthesis of these issues. Written by a philosopher of technology, AI Ethics goes beyond the usual hype and nightmare scenarios to address concrete questions. Mark Coeckelbergh describes influential AI narratives, ranging from Frankenstein's monster to transhumanism and the technological singularity. He surveys relevant philosophical discussions: questions about the fundamental differences between humans and machines and debates over the moral status of AI. He explains the technology of AI, describing different approaches and focusing on machine learning and data science. He offers an overview of important ethical issues, including privacy concerns, responsibility and the delegation of decision making, transparency,

and bias as it arises at all stages of data science processes. He also considers the future of work in an AI economy. Finally, he analyzes a range of policy proposals and discusses challenges for policymakers. He argues for ethical practices that embed values in design, translate democratic values into practices and include a vision of the good life and the good society.

Child Data Citizen

An up-to-date account of the interplay between optimization and machine learning, accessible to students and researchers in both communities. The interplay between optimization and machine learning is one of the most important developments in modern computational science. Optimization formulations and methods are proving to be vital in designing algorithms to extract essential knowledge from huge volumes of data. Machine learning, however, is not simply a consumer of optimization technology but a rapidly evolving field that is itself generating new optimization ideas. This book captures the state of the art of the interaction between optimization and machine learning in a way that is accessible to researchers in both fields. Optimization approaches have enjoyed prominence in machine learning because of their wide applicability and attractive theoretical properties. The

increasing complexity, size, and variety of today's machine learning models call for the reassessment of existing assumptions. This book starts the process of reassessment. It describes the resurgence in novel contexts of established frameworks such as first-order methods, stochastic approximations, convex relaxations, interior-point methods, and proximal methods. It also devotes attention to newer themes such as regularized optimization, robust optimization, gradient and subgradient methods, splitting techniques, and second-order methods. Many of these techniques draw inspiration from other fields, including operations research, theoretical computer science, and subfields of optimization. The book will enrich the ongoing cross-fertilization between the machine learning community and these other fields, and within the broader optimization community.

All Data Are Local

A comprehensive introduction to modern applied statistical genetic data analysis, accessible to those without a background in molecular biology or genetics. Human genetic research is now relevant beyond biology, epidemiology, and the medical sciences, with applications in such fields as psychology, psychiatry, statistics, demography, sociology, and economics. With advances in computing power, the

availability of data, and new techniques, it is now possible to integrate large-scale molecular genetic information into research across a broad range of topics. This book offers the first comprehensive introduction to modern applied statistical genetic data analysis that covers theory, data preparation, and analysis of molecular genetic data, with hands-on computer exercises. It is accessible to students and researchers in any empirically oriented medical, biological, or social science discipline; a background in molecular biology or genetics is not required. The book first provides foundations for statistical genetic data analysis, including a survey of fundamental concepts, primers on statistics and human evolution, and an introduction to polygenic scores. It then covers the practicalities of working with genetic data, discussing such topics as analytical challenges and data management. Finally, the book presents applications and advanced topics, including polygenic score and gene-environment interaction applications, Mendelian Randomization and instrumental variables, and ethical issues. The software and data used in the book are freely available and can be found on the book's website.

Data Action

Introduction -- Definitions -- Descriptive metadata -- Administrative metadata -- Use metadata -- Enabling technologies for metadata -- The Semantic Web -- The future of metadata

Big Data, Little Data, No Data

Perspectives on the varied challenges posed by big data for health, science, law, commerce, and politics.

Optimization for Machine Learning

If machine learning transforms the nature of knowledge, does it also transform the practice of critical thought? Machine learning—programming computers to learn from data—has spread across scientific disciplines, media, entertainment, and government. Medical research, autonomous vehicles, credit transaction processing, computer gaming, recommendation systems, finance, surveillance, and robotics use machine learning. Machine learning devices (sometimes understood as scientific models, sometimes as operational algorithms) anchor the field of data science. They have also become mundane mechanisms deeply embedded in a variety of systems and gadgets. In contexts from the

everyday to the esoteric, machine learning is said to transform the nature of knowledge. In this book, Adrian Mackenzie investigates whether machine learning also transforms the practice of critical thinking. Mackenzie focuses on machine learners—either humans and machines or human-machine relations—situated among settings, data, and devices. The settings range from fMRI to Facebook; the data anything from cat images to DNA sequences; the devices include neural networks, support vector machines, and decision trees. He examines specific learning algorithms—writing code and writing about code—and develops an archaeology of operations that, following Foucault, views machine learning as a form of knowledge production and a strategy of power. Exploring layers of abstraction, data infrastructures, coding practices, diagrams, mathematical formalisms, and the social organization of machine learning, Mackenzie traces the mostly invisible architecture of one of the central zones of contemporary technological cultures. Mackenzie's account of machine learning locates places in which a sense of agency can take root. His archaeology of the operational formation of machine learning does not unearth the footprint of a strategic monolith but reveals the local tributaries of force that feed into the generalization and plurality of the field.

Machine Learning for Data Streams

The first textbook to teach students how to build data analytic solutions on large data sets using cloud-based technologies.

Predicting Structured Data

A comprehensive guide to the conceptual, mathematical, and implementational aspects of analyzing electrical brain signals, including data from MEG, EEG, and LFP recordings.

Computational Thinking

An introduction to a broad range of topics in deep learning, covering mathematical and conceptual background, deep learning techniques used in industry, and research perspectives. “Written by three experts in the field, Deep Learning is the only comprehensive book on the subject.” —Elon Musk, cochair of OpenAI; cofounder and CEO of Tesla and SpaceX Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Because the computer gathers knowledge

from experience, there is no need for a human computer operator to formally specify all the knowledge that the computer needs. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones; a graph of these hierarchies would be many layers deep. This book introduces a broad range of topics in deep learning. The text offers mathematical and conceptual background, covering relevant concepts in linear algebra, probability theory and information theory, numerical computation, and machine learning. It describes deep learning techniques used by practitioners in industry, including deep feedforward networks, regularization, optimization algorithms, convolutional networks, sequence modeling, and practical methodology; and it surveys such applications as natural language processing, speech recognition, computer vision, online recommendation systems, bioinformatics, and videogames. Finally, the book offers research perspectives, covering such theoretical topics as linear factor models, autoencoders, representation learning, structured probabilistic models, Monte Carlo methods, the partition function, approximate inference, and deep generative models. Deep Learning can be used by undergraduate or graduate students planning careers in either industry or research, and by software engineers who want to begin using deep learning in their products or platforms. A website offers supplementary material for

both readers and instructors.

Spatial Computing

A guide to information as the transformative tool of modern business. While we have been preoccupied with the latest i-gadget from Apple and with Google's ongoing expansion, we may have missed something: the fundamental transformation of whole firms and industries into giant information-processing machines. Today, more than eighty percent of workers collect and analyze information (often in digital form) in the course of doing their jobs. This book offers a guide to the role of information in modern business, mapping the use of information within work processes and tracing flows of information across supply-chain management, product development, customer relations, and sales. The emphasis is on information itself, not on information technology. Information, overshadowed for a while by the glamour and novelty of IT, is the fundamental component of the modern corporation. In Information and the Modern Corporation, longtime IBM manager and consultant James Cortada clarifies the differences among data, facts, information, and knowledge and describes how the art of analytics has all but eliminated decision making based on gut feeling, replacing it with fact-based decisions. He describes the working style of “road

warriors," whose offices are anywhere their laptops and cell phones are and whose deep knowledge of a given topic becomes their medium of exchange. Information is the core of the modern enterprise, and the use of information defines the activities of a firm. This essential guide shows managers and employees better ways to leverage information—by design and not by accident.

Data Science

Making diverse data in linguistics and the language sciences open, distributed, and accessible: perspectives from language/language acquistiion researchers and technical LOD (linked open data) researchers. This volume examines the challenges inherent in making diverse data in linguistics and the language sciences open, distributed, integrated, and accessible, thus fostering wide data sharing and collaboration. It is unique in integrating the perspectives of language researchers and technical LOD (linked open data) researchers. Reporting on both active research needs in the field of language acquisition and technical advances in the development of data interoperability, the book demonstrates the advantages of an international infrastructure for scholarship in the field of language sciences. With contributions by researchers who

produce complex data content and scholars involved in both the technology and the conceptual foundations of LLOD (linguistics linked open data), the book focuses on the area of language acquisition because it involves complex and diverse data sets, cross-linguistic analyses, and urgent collaborative research. The contributors discuss a variety of research methods, resources, and infrastructures. Contributors Isabelle Barrière, Nan Bernstein Ratner, Steven Bird, Maria Blume, Ted Caldwell, Christian Chiarcos, Cristina Dye, Suzanne Flynn, Claire Foley, Nancy Ide, Carissa Kang, D. Terence Langendoen, Barbara Lust, Brian MacWhinney, Jonathan Masci, Steven Moran, Antonio Pareja-Lora, Jim Reidy, Oya Y. Rieger, Gary F. Simons, Thorsten Trippel, Kara Warburton, Sue Ellen Wright, Claus Zinn

Cultural Analytics

A comprehensive introduction to the most important machine learning approaches used in predictive data analytics, covering both theoretical concepts and practical applications. Machine learning is often used to build predictive models by extracting patterns from large datasets. These models are used in predictive data analytics applications including price prediction, risk assessment, predicting customer behavior, and document classification. This introductory

textbook offers a detailed and focused treatment of the most important machine learning approaches used in predictive data analytics, covering both theoretical concepts and practical applications. Technical and mathematical material is augmented with explanatory worked examples, and case studies illustrate the application of these models in the broader business context. After discussing the trajectory from data to insight to decision, the book describes four approaches to machine learning: information-based learning, similarity-based learning, probability-based learning, and error-based learning. Each of these approaches is introduced by a nontechnical explanation of the underlying concept, followed by mathematical models and algorithms illustrated by detailed worked examples. Finally, the book considers techniques for evaluating prediction models and offers two case studies that describe specific data analytics projects through each phase of development, from formulating the business problem to implementation of the analytics solution. The book, informed by the authors' many years of teaching machine learning, and working on predictive data analytics projects, is suitable for use by undergraduates in computer science, engineering, mathematics, or statistics; by graduate students in disciplines with applications for predictive data analytics; and as a reference for professionals.

Indexing It All

How to analyze data settings rather than data sets, acknowledging the meaning-making power of the local. In our data-driven society, it is too easy to assume the transparency of data. Instead, Yanni Loukissas argues in All Data Are Local, we should approach data sets with an awareness that data are created by humans and their dutiful machines, at a time, in a place, with the instruments at hand, for audiences that are conditioned to receive them. The term data set implies something discrete, complete, and portable, but it is none of those things. Examining a series of data sources important for understanding the state of public life in the United States—Harvard's Arnold Arboretum, the Digital Public Library of America, UCLA's Television News Archive, and the real estate marketplace Zillow—Loukissas shows us how to analyze data settings rather than data sets. Loukissas sets out six principles: all data are local; data have complex attachments to place; data are collected from heterogeneous sources; data and algorithms are inextricably entangled; interfaces recontextualize data; and data are indexes to local knowledge. He then provides a set of practical guidelines to follow. To make his argument, Loukissas employs a combination of qualitative research on data cultures and exploratory data visualizations. Rebutting the “myth of digital

universalism," Loukissas reminds us of the meaning-making power of the local.

Artificial Unintelligence

A hands-on approach to tasks and techniques in data stream mining and real-time analytics, with examples in MOA, a popular freely available open-source software framework. Today many information sources—including sensor networks, financial markets, social networks, and healthcare monitoring—are so-called data streams, arriving sequentially and at high speed. Analysis must take place in real time, with partial data and without the capacity to store the entire data set. This book presents algorithms and techniques used in data stream mining and real-time analytics. Taking a hands-on approach, the book demonstrates the techniques using MOA (Massive Online Analysis), a popular, freely available open-source software framework, allowing readers to try out the techniques after reading the explanations. The book first offers a brief introduction to the topic, covering big data mining, basic methodologies for mining data streams, and a simple example of MOA. More detailed discussions follow, with chapters on sketching techniques, change, classification, ensemble methods, regression, clustering, and frequent pattern mining. Most of these

chapters include exercises, an MOA-based lab session, or both. Finally, the book discusses the MOA software, covering the MOA graphical user interface, the command line, use of its API, and the development of new methods within MOA. The book will be an essential reference for readers who want to use data stream mining as a tool, researchers in innovation or data stream mining, and programmers who want to create new algorithms for MOA.

Too Smart

The mathematization of causality is a relatively recent development, and has become increasingly important in data science and machine learning. This book offers a self-contained and concise introduction to causal models and how to learn them from data. After explaining the need for causal models and discussing some of the principles underlying causal inference, the book teaches readers how to use causal models: how to compute intervention distributions, how to infer causal models from observational and interventional data, and how causal ideas could be exploited for classical machine learning problems. All of these topics are discussed first in terms of two variables and then in the more general multivariate case. The bivariate case turns out to be a particularly hard problem for causal

learning because there are no conditional independences as used by classical methods for solving multivariate cases. The authors consider analyzing statistical asymmetries between cause and effect to be highly instructive, and they report on their decade of intensive research into this problem. The book is accessible to readers with a background in machine learning or statistics, and can be used in graduate courses or as a reference for researchers. The text includes code snippets that can be copied and pasted, exercises, and an appendix with a summary of the most important technical concepts.

Mathematics of Big Data

How to create an Internet of Trusted Data in which insights from data can be extracted without collecting, holding, or revealing the underlying data. Trusted Data describes a data architecture that places humans and their societal values at the center of the discussion. By involving people from all parts of the ecosystem of information, this new approach allows us to realize the benefits of data-driven algorithmic decision making while minimizing the risks and unintended consequences. It proposes a software architecture and legal framework for an Internet of Trusted Data that provides safe, secure access for everyone and protects against bias, unfairness, and other

unintended effects. This approach addresses issues of data privacy, security, ownership, and trust by allowing insights to be extracted from data held by different people, companies, or governments without collecting, holding, or revealing the underlying data. The software architecture, called Open Algorithms, or OPAL, sends algorithms to databases rather than copying or sharing data. The data is protected by existing firewalls; only encrypted results are shared. Data never leaves its repository. A higher security architecture, ENIGMA, built on OPAL, is fully encrypted. Contributors Michiel Bakker, Yves-Alexandre de Montjoye, Daniel Greenwood, Thomas Hardjoni, Jake Kendall, Cameron Kerry, Bruno Lepri, Alexander Lipton, Takeo Nishikata, Alejandro Noriega-Campero, Nuria Oliver, Alex Pentland, David L. Shrier, Jacopo Staiano, Guy Zyskind An MIT Connection Science and Engineering Book

Fundamentals of Machine Learning for Predictive Data Analytics

How data science and the analysis of networks help us solve the puzzle of unintended consequences. Social life is full of paradoxes. Our intentional actions often trigger outcomes that we did not intend or

even envision. How do we explain those unintended effects and what can we do to regulate them? In Decoding the Social World, Sandra González-Bailón explains how data science and digital traces help us solve the puzzle of unintended consequences—offering the solution to a social paradox that has intrigued thinkers for centuries. Communication has always been the force that makes a collection of people more than the sum of individuals, but only now can we explain why: digital technologies have made it possible to parse the information we generate by being social in new, imaginative ways. And yet we must look at that data, González-Bailón argues, through the lens of theories that capture the nature of social life. The technologies we use, in the end, are also a manifestation of the social world we inhabit. González-Bailón discusses how the unpredictability of social life relates to communication networks, social influence, and the unintended effects that derive from individual decisions. She describes how communication generates social dynamics in aggregate (leading to episodes of “collective effervescence”) and discusses the mechanisms that underlie large-scale diffusion, when information and behavior spread “like wildfire.” She applies the theory of networks to illuminate why collective outcomes can differ drastically even when they arise from the same individual actions. By opening the black box of unintended effects, González-Bailón identifies strategies for

social intervention and discusses the policy implications—and how data science and evidence-based research embolden critical thinking in a world that is constantly changing.

Introduction to Natural Language Processing

An exploration of infographics and data visualization as a cultural phenomenon, from eighteenth-century print culture to today's data journalism. Infographics and data visualization are ubiquitous in our everyday media diet, particularly in news—in print newspapers, on television news, and online. It has been argued that infographics are changing what it means to be literate in the twenty-first century—and even that they harmonize uniquely with human cognition. In this first serious exploration of the subject, Murray Dick traces the cultural evolution of the infographic, examining its use in news—and resistance to its use—from eighteenth-century print culture to today's data journalism. He identifies six historical phases of infographics in popular culture: the proto-infographic, the classical, the improving, the commercial, the ideological, and the professional. Dick describes the emergence of infographic forms within a wider history of journalism, culture, and communications, focusing his analysis on the UK. He considers their use in the partisan British journalism of late

eighteenth and early nineteenth-century print media; their later deployment as a vehicle for reform and improvement; their mass-market debut in the twentieth century as a means of explanation (and sometimes propaganda); and their use for both ideological and professional purposes in the post-World War II marketized newspaper culture. Finally, he proposes best practices for news infographics and defends infographics and data visualization against a range of criticism. Dick offers not only a history of how the public has experienced and understood the infographic, but also an account of what data visualization can tell us about the past.

Open Space

A new way of thinking about data science and data ethics that is informed by the ideas of intersectional feminism. Today, data science is a form of power. It has been used to expose injustice, improve health outcomes, and topple governments. But it has also been used to discriminate, police, and surveil. This potential for good, on the one hand, and harm, on the other, makes it essential to ask: Data science by whom? Data science for whom? Data science with whose interests in mind? The narratives around big data and data science are overwhelmingly white, male, and techno-heroic. In Data Feminism,

Catherine D'Ignazio and Lauren Klein present a new way of thinking about data science and data ethics—one that is informed by intersectional feminist thought. Illustrating data feminism in action, D'Ignazio and Klein show how challenges to the male/female binary can help challenge other hierarchical (and empirically wrong) classification systems. They explain how, for example, an understanding of emotion can expand our ideas about effective data visualization, and how the concept of invisible labor can expose the significant human efforts required by our automated systems. And they show why the data never, ever "speak for themselves." Data Feminism offers strategies for data scientists seeking to learn how feminism can help them work toward justice, and for feminists who want to focus their efforts on the growing field of data science. But Data Feminism is about much more than gender. It is about power, about who has it and who doesn't, and about how those differentials of power can be challenged and changed.

Trusted Data

A book at the intersection of data science and media studies, presenting concepts and methods for computational analysis of cultural data. How can we see a billion images? What analytical methods can we

bring to bear on the astonishing scale of digital culture--the billions of photographs shared on social media every day, the hundreds of millions of songs created by twenty million musicians on Soundcloud, the content of four billion Pinterest boards? In Cultural Analytics, Lev Manovich presents concepts and methods for computational analysis of cultural data. Drawing on more than a decade of research and projects from his own lab, Manovich offers a gentle, nontechnical introduction to the core ideas of data analytics and discusses the ways that our society uses data and algorithms.

AI Ethics

An introduction to computational thinking that traces a genealogy beginning centuries before the digital computer. A few decades into the digital era, scientists discovered that thinking in terms of computation made possible an entirely new way of organizing scientific investigation; eventually, every field had a computational branch: computational physics, computational biology, computational sociology. More recently, “computational thinking” has become part of the K-12 curriculum. But what is computational thinking? This volume in the MIT Press Essential Knowledge series offers an accessible overview, tracing a genealogy that begins centuries before digital computers and

portraying computational thinking as pioneers of computing have described it. The authors explain that computational thinking (CT) is not a set of concepts for programming; it is a way of thinking that is honed through practice: the mental skills for designing computations to do jobs for us, and for explaining and interpreting the world as a complex of information processes. Mathematically trained experts (known as “computers”) who performed complex calculations as teams engaged in CT long before electronic computers. The authors identify six dimensions of today's highly developed CT—methods, machines, computing education, software engineering, computational science, and design—and cover each in a chapter. Along the way, they debunk inflated claims for CT and computation while making clear the power of CT in all its complexity and multiplicity.

An Introduction to Statistical Genetic Data Analysis

An accessible guide to the ideas and technologies underlying such applications as GPS, Google Maps, Pokémon Go, ride-sharing, driverless cars, and drone surveillance. Billions of people around the globe use various applications of spatial computing daily—by using a ride-sharing app, GPS, the e911 system, social media check-ins, even Pokémon Go. Scientists and researchers use spatial computing to track

diseases, map the bottom of the oceans, chart the behavior of endangered species, and create election maps in real time. Drones and driverless cars use a variety of spatial computing technologies. Spatial computing works by understanding the physical world, knowing and communicating our relation to places in that world, and navigating through those places. It has changed our lives and infrastructures profoundly, marking a significant shift in how we make our way in the world. This volume in the MIT Essential Knowledge series explains the technologies and ideas behind spatial computing. The book offers accessible descriptions of GPS and location-based services, including the use of Wi-Fi, Bluetooth, and RFID for position determination out of satellite range; remote sensing, which uses satellite and aerial platforms to monitor such varied phenomena as global food production, the effects of climate change, and subsurface natural resources on other planets; geographic information systems (GIS), which store, analyze, and visualize spatial data; spatial databases, which store multiple forms of spatial data; and spatial statistics and spatial data science, used to analyze location-related data.

Introduction to Machine Learning

The book as object, as content, as idea, as interface. What is the

book in a digital age? Is it a physical object containing pages encased in covers? Is it a portable device that gives us access to entire libraries? The codex, the book as bound paper sheets, emerged around 150 CE. It was preceded by clay tablets and papyrus scrolls. Are those books? In this volume in the MIT Press Essential Knowledge series, Amaranth Borsuk considers the history of the book, the future of the book, and the idea of the book. Tracing the interrelationship of form and content in the book's development, she bridges book history, book arts, and electronic literature to expand our definition of an object we thought we knew intimately. Contrary to the many reports of its death (which has been blamed at various times on newspapers, television, and e-readers), the book is alive. Despite nostalgic paeans to the codex and its printed pages, Borsuk reminds us, the term “book” commonly refers to both medium and content. And the medium has proved to be malleable. Rather than pinning our notion of the book to a single form, Borsuk argues, we should remember its long history of transformation. Considering the book as object, content, idea, and interface, she shows that the physical form of the book has always been the site of experimentation and play. Rather than creating a false dichotomy between print and digital media, we should appreciate their continuities.

Cloud Computing for Machine Learning and Cognitive Applications

The new edition of an introductory text that teaches students the art of computational problem solving, covering topics ranging from simple algorithms to information visualization.

Deep Learning

A guide to understanding the inner workings and outer limits of technology and why we should never assume that computers always get it right. In Artificial Unintelligence, Meredith Broussard argues that our collective enthusiasm for applying computer technology to every aspect of life has resulted in a tremendous amount of poorly designed systems. We are so eager to do everything digitally--hiring, driving, paying bills, even choosing romantic partners--that we have stopped demanding that our technology actually work. Broussard, a software developer and journalist, reminds us that there are fundamental limits to what we can (and should) do with technology. With this book, she offers a guide to understanding the inner workings and outer limits of technology--and issues a warning that we should never assume that

computers always get things right. Making a case against technochauvinism--the belief that technology is always the solution--Broussard argues that it's just not true that social problems would inevitably retreat before a digitally enabled Utopia. To prove her point, she undertakes a series of adventures in computer programming. She goes for an alarming ride in a driverless car, concluding "the cyborg future is not coming any time soon"; uses artificial intelligence to investigate why students can't pass standardized tests; deploys machine learning to predict which passengers survived the Titanic disaster; and attempts to repair the U.S. campaign finance system by building AI software. If we understand the limits of what we can do with technology, Broussard tells us, we can make better choices about what we should do with it to make the world better for everyone.

Metadata

A survey of computational methods for understanding, generating, and manipulating human language, which offers a synthesis of classical representations and algorithms with contemporary machine learning techniques. This textbook provides a technical perspective on natural language processing—methods for building computer software that

understands, generates, and manipulates human language. It emphasizes contemporary data-driven approaches, focusing on techniques from supervised and unsupervised machine learning. The first section establishes a foundation in machine learning by building a set of tools that will be used throughout the book and applying them to word-based textual analysis. The second section introduces structured representations of language, including sequences, trees, and graphs. The third section explores different approaches to the representation and analysis of linguistic meaning, ranging from formal logic to neural word embeddings. The final section offers chapter-length treatments of three transformative applications of natural language processing: information extraction, machine translation, and text generation. End-of-chapter exercises include both paper-and-pencil analysis and software implementation. The text synthesizes and distills a broad and diverse research literature, linking contemporary machine learning techniques with the field's linguistic and computational foundations. It is suitable for use in advanced undergraduate and graduate-level courses and as a reference for software engineers and data scientists. Readers should have a background in computer programming and college-level mathematics. After mastering the material presented, students will have the technical skill to build and analyze novel natural language processing

systems and to understand the latest research in the field.

The Book

A critical history of the modern tradition of documentation, tracing the representation of individuals and groups in the form of documents, information, and data. In this book, Ronald Day offers a critical history of the modern tradition of documentation. Focusing on the documentary index (understood as a mode of social positioning), and drawing on the work of the French documentalist Suzanne Briet, Day explores the understanding and uses of indexicality. He examines the transition as indexes went from being explicit professional structures that mediated users and documents to being implicit infrastructural devices used in everyday information and communication acts. Doing so, he also traces three epistemic eras in the representation of individuals and groups, first in the forms of documents, then information, then data. Day investigates five cases from the modern tradition of documentation. He considers the socio-technical instrumentalism of Paul Otlet, “the father of European documentation” (contrasting it to the hermeneutic perspective of Martin Heidegger); the shift from documentation to information science and the accompanying transformation of persons and texts into users and

information; social media's use of algorithms, further subsuming persons and texts; attempts to build android robots—to embody human agency within an information system that resembles a human being; and social “big data” as a technique of neoliberal governance that employs indexing and analytics for purposes of surveillance. Finally, Day considers the status of critique and judgment at a time when people and their rights of judgment are increasingly mediated, displaced, and replaced by modern documentary techniques.

A Hands-On Introduction to Data Science

State-of-the-art algorithms and theory in a novel domain of machine learning,prediction when the output has structure.

Big Data Is Not a Monolith

An examination of the uses of data within a changing knowledge infrastructure, offering analysis and case studies from the sciences, social sciences, and humanities. “Big Data” is on the covers of Science, Nature, the Economist, and Wired magazines, on the front pages of the Wall Street Journal and the New York Times. But despite

the media hyperbole, as Christine Borgman points out in this examination of data and scholarly research, having the right data is usually better than having more data; little data can be just as valuable as big data. In many cases, there are no data—because relevant data don't exist, cannot be found, or are not available. Moreover, data sharing is difficult, incentives to do so are minimal, and data practices vary widely across disciplines. Borgman, an often-cited authority on scholarly communication, argues that data have no value or meaning in isolation; they exist within a knowledge infrastructure—an ecology of people, practices, technologies, institutions, material objects, and relationships. After laying out the premises of her investigation—six "provocations" meant to inspire discussion about the uses of data in scholarship—Borgman offers case studies of data practices in the sciences, the social sciences, and the humanities, and then considers the implications of her findings for scholarly practice and research policy. To manage and exploit data over the long term, Borgman argues, requires massive investment in knowledge infrastructures; at stake is the future of scholarship.

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language

Sciences

An introductory textbook offering a low barrier entry to data science; the hands-on approach will appeal to students from a range of disciplines.

The Infographic

"Data Action will offer a model for reading, collecting, visualizing, and putting data to work on civic change. Using arresting graphics and influential case studies, as well as incorporating cultural and historical context, Data Action presents a helpful corrective to standard practice. Historically, data has been used and manipulated to make policy decisions without input from the general public. Data Action asks advocates of big data to rethink how they work by laying out a methodology for more transparent and accountable data analysis. The tools outlined in this book will help anyone, not just government officials, but data scientists, civic activists and hackers, as well as all citizens reaching for more effective civic debates and policy reforms, to shape our environment, economy, public health, and quality of life, with greater transparency and democratic participation"--

Introduction to Computation and Programming Using Python

The first book to present the common mathematical foundations of big data analysis across a range of applications and technologies. Today, the volume, velocity, and variety of data are increasing rapidly across a range of fields, including Internet search, healthcare, finance, social media, wireless devices, and cybersecurity. Indeed, these data are growing at a rate beyond our capacity to analyze them. The tools—including spreadsheets, databases, matrices, and graphs—developed to address this challenge all reflect the need to store and operate on data as whole sets rather than as individual elements. This book presents the common mathematical foundations of these data sets that apply across many applications and technologies. Associative arrays unify and simplify data, allowing readers to look past the differences among the various tools and leverage their mathematical similarities in order to solve the hardest big data challenges. The book first introduces the concept of the associative array in practical terms, presents the associative array manipulation system D4M (Dynamic Distributed Dimensional Data Model), and describes the application of associative arrays to graph analysis and machine learning. It provides a mathematically rigorous definition of associative arrays and describes the properties of associative arrays

that arise from this definition. Finally, the book shows how concepts of linearity can be extended to encompass associative arrays. Mathematics of Big Data can be used as a textbook or reference by engineers, scientists, mathematicians, computer scientists, and software engineers who analyze big data.

Analyzing Neural Time Series Data

Episodes in the history of data, from early modern math problems to today's inescapable “dataveillance,” that demonstrate the dependence of data on culture. We live in the era of Big Data, with storage and transmission capacity measured not just in terabytes but in petabytes (where peta- denotes a quadrillion, or a thousand trillion). Data collection is constant and even insidious, with every click and every “like” stored somewhere for something. This book reminds us that data is anything but “raw,” that we shouldn't think of data as a natural resource but as a cultural one that needs to be generated, protected, and interpreted. The book's essays describe eight episodes in the history of data from the predigital to the digital. Together they address such issues as the ways that different kinds of data and different domains of inquiry are mutually defining; how data are variously “cooked” in the processes of their collection and use; and

conflicts over what can—or can't—be “reduced” to data. Contributors discuss the intellectual history of data as a concept; describe early financial modeling and some unusual sources for astronomical data; discover the prehistory of the database in newspaper clippings and index cards; and consider contemporary “dataveillance” of our online habits as well as the complexity of scientific data curation. Essay Authors Geoffrey C. Bowker, Kevin R. Brine, Ellen Gruber Garvey, Lisa Gitelman, Steven J. Jackson, Virginia Jackson, Markus Krajewski, Mary Poovey, Rita Raley, David Ribes, Daniel Rosenberg, Matthew Stanley, Travis D. Williams

Deep Learning

The goal of machine learning is to program computers to use example data or past experience to solve a given problem. Many successful applications of machine learning exist already, including systems that analyze past sales data to predict customer behavior, optimize robot behavior so that a task can be completed using minimum resources, and extract knowledge from bioinformatics data. Introduction to Machine Learning is a comprehensive textbook on the subject, covering a broad array of topics not usually included in introductory machine learning texts. Subjects include supervised learning; Bayesian decision theory;

parametric, semi-parametric, and nonparametric methods; multivariate analysis; hidden Markov models; reinforcement learning; kernel machines; graphical models; Bayesian estimation; and statistical testing.Machine learning is rapidly becoming a skill that computer science students must master before graduation. The third edition of Introduction to Machine Learning reflects this shift, with added support for beginners, including selected solutions for exercises and additional example data sets (with code available online). Other substantial changes include discussions of outlier detection; ranking algorithms for perceptrons and support vector machines; matrix decomposition and spectral methods; distance estimation; new kernel algorithms; deep learning in multilayered perceptrons; and the nonparametric approach to Bayesian methods. All learning algorithms are explained so that students can easily move from the equations in the book to a computer program. The book can be used by both advanced undergraduates and graduate students. It will also be of interest to professionals who are concerned with the application of machine learning methods.

Data Feminism

An accessible introduction to the artificial intelligence technology

that enables computer vision, speech recognition, machine translation, and driverless cars. Deep learning is an artificial intelligence technology that enables computer vision, speech recognition in mobile phones, machine translation, AI games, driverless cars, and other applications. When we use consumer products from Google, Microsoft, Facebook, Apple, or Baidu, we are often interacting with a deep learning system. In this volume in the MIT Press Essential Knowledge series, computer scientist John Kelleher offers an accessible and concise but comprehensive introduction to the fundamental technology at the heart of the artificial intelligence revolution. Kelleher explains that deep learning enables data-driven decisions by identifying and extracting patterns from large datasets; its ability to learn from complex data makes deep learning ideally suited to take advantage of the rapid growth in big data and computational power. Kelleher also explains some of the basic concepts in deep learning, presents a history of advances in the field, and discusses the current state of the art. He describes the most important deep learning architectures, including autoencoders, recurrent neural networks, and long short-term networks, as well as such recent developments as Generative Adversarial Networks and capsule networks. He also provides a comprehensive (and comprehensible) introduction to the two fundamental algorithms in deep learning: gradient descent and

backpropagation. Finally, Kelleher considers the future of deep learning—major trends, possible developments, and significant challenges.

Decoding the Social World

A concise introduction to the emerging field of data science, explaining its evolution, relation to machine learning, current uses, data infrastructure issues, and ethical challenges. The goal of data science is to improve decision making through the analysis of data. Today data science determines the ads we see online, the books and movies that are recommended to us online, which emails are filtered into our spam folders, and even how much we pay for health insurance. This volume in the MIT Press Essential Knowledge series offers a concise introduction to the emerging field of data science, explaining its evolution, current uses, data infrastructure issues, and ethical challenges. It has never been easier for organizations to gather, store, and process data. Use of data science is driven by the rise of big data and social media, the development of high-performance computing, and the emergence of such powerful methods for data analysis and modeling as deep learning. Data science encompasses a set of principles, problem definitions, algorithms, and processes for

extracting non-obvious and useful patterns from large datasets. It is closely related to the fields of data mining and machine learning, but broader in scope. This book offers a brief history of the field, introduces fundamental data concepts, and describes the stages in a data science project. It considers data infrastructure and the challenges posed by integrating data from multiple sources, introduces the basics of machine learning, and discusses how to link machine learning expertise with real-world problems. The book also reviews ethical and legal issues, developments in data regulation, and computational approaches to preserving privacy. Finally, it considers the future impact of data science and offers principles for success in data science projects.

A Vast Machine

The science behind global warming, and its history: how scientists learned to understand the atmosphere, to measure it, to trace its past, and to model its future. Global warming skeptics often fall back on the argument that the scientific case for global warming is all model predictions, nothing but simulation; they warn us that we need to wait for real data, “sound science.” In A Vast Machine Paul Edwards has news for these skeptics: without models, there are no data. Today,

no collection of signals or observations—even from satellites, which can “see” the whole planet with a single instrument—becomes global in time and space without passing through a series of data models. Everything we know about the world's climate we know through models. Edwards offers an engaging and innovative history of how scientists learned to understand the atmosphere—to measure it, trace its past, and model its future.

Information and the Modern Corporation

Who benefits from smart technology? Whose interests are served when we trade our personal data for convenience and connectivity? Smart technology is everywhere: smart umbrellas that light up when rain is in the forecast; smart cars that relieve drivers of the drudgery of driving; smart toothbrushes that send your dental hygiene details to the cloud. Nothing is safe from smartification. In Too Smart, Jathan Sadowski looks at the proliferation of smart stuff in our lives and asks whether the tradeoff—exchanging our personal data for convenience and connectivity—is worth it. Who benefits from smart technology? Sadowski explains how data, once the purview of researchers and policy wonks, has become a form of capital. Smart technology, he argues, is driven by the dual imperatives of digital capitalism: extracting data

from, and expanding control over, everything and everybody. He looks at three domains colonized by smart technologies' collection and control systems: the smart self, the smart home, and the smart city. The smart self involves more than self-tracking of steps walked and calories burned; it raises questions about what others do with our data and how they direct our behavior—whether or not we want them to. The smart home collects data about our habits that offer business a window into our domestic spaces. And the smart city, where these systems have space to grow, offers military-grade surveillance capabilities to local authorities. Technology gets smart from our data. We may enjoy the conveniences we get in return (the refrigerator says we're out of milk!), but, Sadowski argues, smart technology advances the interests of corporate technocratic power—and will continue to do so unless we demand oversight and ownership of our data.

Machine Learners

Key to understanding and addressing climate change is continuous and precise monitoring of environmental conditions. Satellites play an important role in collecting climate data, offering comprehensive global coverage that can't be matched by in situ observation. And yet,

as Mariel Borowitz shows in this book, much satellite data is not freely available but restricted; this remains true despite the data-sharing advocacy of international organizations and a global open data movement. Borowitz examines policies governing the sharing of environmental satellite data, offering a model of data-sharing policy development and applying it in case studies from the United States, Europe, and Japan -- countries responsible for nearly half of the unclassified government Earth observation satellites. Borowitz develops a model that centers on the government agency as the primary actor while taking into account the roles of such outside actors as other government officials and non-governmental actors, as well as the economic, security, and normative attributes of the data itself. The case studies include the U.S. National Aeronautics and Space Administration (NASA) and the U.S. National Oceanographic and Atmospheric Association (NOAA), and the United States Geological Survey (USGS); the European Space Agency (ESA) and the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT); and the Japanese Aerospace Exploration Agency (JAXA) and the Japanese Meteorological Agency (JMA). Finally, she considers the policy implications of her findings for the future and provides recommendations on how to increase global sharing of satellite data.

Elements of Causal Inference

"Explores the collection of children's biometric, educational, and social media data and its immediate and downstream effects for individuals and families"--