Reconhecimentos

"Um recurso de leitura obrigatória para qualquer pessoa decidida a aproveitar a oportunidade do Big Data."
— *Craig Vaughan*
Vice-presidente global da SAP

"Um livro oportuno que informa, em alto e bom som, aquilo que finalmente se tornou evidente: no mundo moderno, Dados são Negócios e você não pode mais pensar em negócios sem *pensar em dados*. Leia este livro e você compreenderá a ciência por trás dos dados."
— *Ron Bekkerman*
Diretor de dados da Carmel Ventures

"Um ótimo livro para gestores de negócios que lideram ou interagem com cientistas de dados e que desejam compreender melhor os princípios e algoritmos disponíveis, sem os detalhes técnicos dos livros sobre um assunto específico."
— *Ronny Kohavi*
Arquiteto parceiro da Microsoft Online Services Division

"Provost e Fawcett reuniram toda sua maestria na arte e na ciência da análise de dados do mundo real em uma incomparável introdução ao assunto."
— *Geoff Webb*
Editor-chefe do *Data Mining and Knowledge Discovery Journal*

"Eu adoraria que todos com quem eu trabalhei tivessem lido este livro."
— *Claudia Perlich*
Cientista-chefe da Dstillery e Grande Vencedora do Prêmio Advertising Research Foundation Innovation (2013)

"Uma peça fundamental no desenvolvimento acelerado do mundo de Data Science. Uma leitura obrigatória para todos os interessados na revolução do Big Data."
— *Justin Gapper*
Gerente de análise da unidade de negócios da Teledyne Scientific and Imaging

"Os autores, ambos renomados especialistas em Data Science mesmo antes do tema receber esse nome, escolheram um tópico complexo e o tornaram acessível a todos os níveis, mas, principalmente, muito útil para os novatos. Até onde eu sei, este é o primeiro livro do tipo — com foco em conceitos de Data Science aplicados a problemas práticos de negócios. Está generosamente recheado de exemplos do mundo real que definem problemas familiares e acessíveis no mundo dos negócios: rotatividade de clientes, marketing direcionado, até mesmo análise de uísque!

A obra é única, no sentido de que não é um livro de receita de algoritmos, ao contrário, ajuda o leitor a compreender os conceitos subjacentes por trás do Data Science e, mais importante, como abordar e ser bem-sucedido na resolução de problemas. Se você está procurando uma visão geral sobre Data Science ou se você é novato no assunto e precisa conhecer o básico, esta é uma leitura obrigatória."

— Chris Volinsky
Diretor de Pesquisas Estatísticas na AT&T Labs e
Membro da Equipe Vencedora do Desafio Netflix de US$ 1 milhão

"Este livro vai além da análise de dados para principiantes. É o guia essencial para aqueles (ou todos?) cujas empresas são construídas sobre a onipresença de oportunidades envolvendo dados e a nova ordem de tomada de decisão baseada em dados."

— Tom Phillips
CEO da Dstillery e ex-diretor do Google Search e Analytics

"O uso inteligente de dados se tornou uma força que impulsiona os negócios para novos níveis de competitividade. Para prosperar neste ecossistema orientado por dados, engenheiros, analistas e gerentes devem compreender suas opções, escolhas projetadas e implicações. Com exemplos motivadores, exposição clara e uma grande variedade de detalhes que abrange não só o "como", mas os "porquês", Data Science para Negócios é fundamental para aqueles que desejam se engajar no desenvolvimento e na aplicação de sistemas orientados por dados."

— Josh Attenberg
Chefe em Data Science do Etsy

"Os dados são o alicerce de novas ondas de crescimento de produtividade, inovação e uma maior percepção do cliente. Apenas recentemente o tópico passou a ser visto como uma fonte de vantagem competitiva. Lidar bem com os dados está rapidamente se tornando um requisito mínimo para entrar no jogo. A profunda experiência aplicada dos autores faz com que esta seja uma leitura obrigatória — uma janela para a estratégia de seu concorrente."

— Alan Murray
Empreendedor Serial; Parceiro da Coriolis Ventures

"Um dos melhores livros sobre mineração de dados e que me ajudou a ter várias ideias sobre análise de liquidez no negócio FX. Os exemplos são excelentes e ajudam a dar um mergulho profundo no assunto! Este livro ficará na minha estante para sempre!"

— Nidhi Kathuria
Vice-presidente de FX do Royal Bank of Scotland

"Um livro excelente e acessível para ajudar as pessoas de negócio a apreciarem melhor os conceitos, ferramentas e técnicas utilizadas pelos cientistas de dados. E para quem trabalha com Data Science apreciar melhor o contexto empresarial em que suas soluções são implantadas."
— Joe McCarthy
Diretor de análise e Data Science da Atigeo

"Na minha opinião, é o melhor livro sobre Data Science e Big Data para uma compreensão profissional de analistas de negócios e gerentes que devem aplicar essas técnicas no mundo real."
— *Ira Laefsky*
MS em Engenharia (Ciência da Computação) /MBA em Tecnologia da Informação e Pesquisador da Interação Humana e Computador anteriormente na Equipe de Consultoria Sênior de Arthur D. Little, Inc. and Digital Equipment Corporation

"Com exemplos motivadores, exposição clara e uma grande variedade de detalhes que abrangem não só o "como", mas os "porquês", Data Science para Negócios é fundamental para aqueles que desejam se envolver no desenvolvimento e na aplicação de sistemas orientados por dados."
— *Ted O'Brien*
Cofundador/ Diretor de Aquisição de Talentos da Starbridge Partners e Editor da *Data Science Report*

Data Science para Negócios

Foster Provost e Tom Fawcett

ALTA BOOKS
EDITORA
Rio de Janeiro, 2016

Para nossos pais.

Sumário

Prefácio

Data Science para Negócios destina-se a diversos tipos de leitores:

- Pessoas de negócios que trabalham com cientistas de dados, gerenciando projetos orientados para Data Science ou investindo em empreendimentos de Data Science,
- Desenvolvedores que implementam soluções de Data Science, e
- Aspirantes a cientistas de dados.

Este não é um livro sobre algoritmos, nem é um substituto para o mesmo. Evitamos, deliberadamente, uma abordagem centrada em algoritmos. Acreditamos que existe um conjunto relativamente pequeno de conceitos ou princípios fundamentais que norteiam as técnicas para extrair conhecimento útil a partir dos dados. Esses conceitos servem como *base* para muitos algoritmos bem conhecidos de mineração de dados. Além disso, esses conceitos são a base da análise de problemas de negócios centrados em dados, da criação e da avaliação de soluções de Data Science, e da avaliação de estratégias e propostas gerais de Data Science. Por conseguinte, organizamos a exposição em torno desses princípios gerais e não de algoritmos específicos. Onde foi necessário descrever detalhes processuais, usamos uma combinação de texto e diagramas que consideramos mais acessíveis do que uma listagem de etapas algorítmicas detalhadas.

O livro não presume um conhecimento matemático sofisticado. No entanto, por sua própria natureza, o material é um pouco técnico — o objetivo é transmitir uma compreensão significativa de Data Science, não apenas uma visão geral de alto nível. De modo geral, tentamos minimizar a matemática e tornar a exposição o mais "conceitual" possível.

Colegas da indústria comentam que o livro é inestimável para alinhar a compreensão das equipes de negócios, técnica/de desenvolvimento e de Data Science. Essa observação se baseia em uma pequena amostra, por isso, estamos curiosos para ver o quão geral ela realmente é (veja o Capítulo 5!). Idealmente, vislumbramos um livro que qualquer cientista de dados daria aos seus colaboradores das equipes de desenvolvimento ou de negócios dizendo:

se você realmente deseja projetar/implementar soluções de primeira linha em Data Science para problemas de negócios, precisamos ter um conhecimento comum sobre este material.

Os colegas também nos dizem que o livro foi muito útil de uma maneira inusitada: na preparação para entrevistar candidatos a uma vaga em Data Science. A demanda das empresas pela contratação de cientistas de dados é forte e crescente. Em resposta, mais e mais candidatos se apresentam como cientistas de dados. Cada candidato à vaga em Data Science deve compreender os fundamentos apresentados neste livro. Nossos colegas do setor dizem que ficam surpresos com o fato de que muitos não compreendem. Discutimos, com alguma seriedade, um panfleto de acompanhamento "Cliff's Notes to Interviewing for Data Science Jobs" (conteúdo em inglês).

Nossa Abordagem Conceitual ao Data Science

Neste livro, apresentamos uma coleção dos conceitos mais importantes e fundamentais em Data Science. Alguns desses conceitos são "destaques" nos capítulos e outros são introduzidos mais naturalmente através de debates (e, portanto, não são necessariamente identificados como conceitos fundamentais). Os conceitos abrangem desde o processo de vislumbrar o problema, aplicar as técnicas de Data Science, até implantar os resultados para melhorar a tomada de decisão. Eles também embasam uma grande variedade de métodos e técnicas de análise de negócios.

Os conceitos se encaixam em três categorias gerais:

1. Conceitos sobre como a ciência de dados (Data Science) se encaixa na organização e no cenário competitivo, incluindo formas de atrair, estruturar e nutrir equipes de Data Science; maneiras de pensar sobre como Data Science leva a uma vantagem competitiva; e conceitos táticos para se sair bem com projetos de Data Science.

2. Formas gerais de pensar em dados de maneira analítica. Isso ajuda a identificar os dados apropriados e a considerar métodos adequados. Os conceitos incluem o *processo de mineração de dados*, bem como o acúmulo de diferentes *tarefas de alto nível de mineração de dados*.

3. Conceitos gerais para realmente extrair conhecimento a partir de dados, que sustentam a vasta gama de atividades de Data Science e seus algoritmos.

Por exemplo, um conceito fundamental é o de determinar a similaridade de duas entidades descritas pelos dados. Essa capacidade forma a base de várias tarefas específicas. Ela pode ser usada diretamente para *encontrar* clientes semelhantes em uma base de dados. Ela forma o núcleo de vários algoritmos de *previsão* que estimam um valor alvo, como o uso esperado de recursos de um cliente ou a probabilidade de ele responder a uma oferta. É também a base para técnicas de *agrupamento*, que reúne entidades por suas características compartilhadas, sem um objetivo focado. A similaridade forma a base da *recuperação de informação*, na qual documentos ou páginas da web, pertinentes a uma consulta de pesquisa, são recuperados.

Por fim, sustenta vários algoritmos comuns para *recomendação*. Um livro tradicional orientado para algoritmos pode apresentar cada uma dessas tarefas em um capítulo diferente, sob nomes diferentes, com aspectos comuns encobertos por detalhes de algoritmos ou proposições matemáticas. Neste livro, em vez disso, focamos em conceitos unificadores, apresentando tarefas e algoritmos específicos como manifestações naturais deles.

Como outro exemplo, ao avaliar a utilidade de um padrão, vemos uma noção de *elevação* — quão mais prevalente é um padrão do que o esperado de modo aleatório — amplamente recorrente em Data Science. É muito utilizado para avaliar diferentes tipos de padrões em diferentes contextos. Algoritmos de anúncios direcionados são avaliados pelo cálculo da elevação que se obtém com a população alvo. A elevação é utilizada para julgar o peso da evidência a favor ou contra uma conclusão. A elevação ajuda a determinar se uma coocorrência (uma associação) nos dados é interessante, em vez de simplesmente ser uma consequência natural da popularidade.

Acreditamos que explicar Data Science em torno de conceitos tão fundamentais não só ajuda o leitor, mas também facilita a comunicação entre os executivos e os cientistas de dados. Fornece um vocabulário comum e permite que ambas as partes compreendam melhor uma à outra. Os conceitos compartilhados levam a discussões mais profundas que podem revelar problemas críticos que passariam despercebidos.

Para o instrutor

Este livro tem sido utilizado com sucesso como livro-texto para uma grande variedade de cursos de Data Science. Historicamente, o livro surgiu a partir do desenvolvimento das aulas multidisciplinares de Data Science de Foster na Escola Stern, da NYU, no outono de 2005.[1] A aula original foi destinada para alunos de MBA e alunos MSIS, mas atraiu alunos de toda a universidade. O aspecto mais interessante da aula não foi a inesperada atratividade para alunos fora do MBA e do MSIS, para quem se destinava. O mais interessante foi que também se mostrou muito valiosa para estudantes com sólido conhecimento em aprendizado de máquina e outras disciplinas técnicas. Parte da explicação parecia ser a falta de um enfoque em princípios fundamentais e outras questões, além de algoritmos, em seus currículos.

Agora, na NYU, usamos o livro como complemento para uma variedade de programas relacionados ao Data Science: os programas originais de MBA e MSIS, graduação em análise de negócios, novo MS da NYU/Sterns no programa de Análise de Negócios e como Introdução ao Data Science para novos MS da NYU em Data Science. Além disso, (antes da publicação) o livro foi adotado por mais de vinte outras universidades para programas em nove países (e aumentando), em escolas de negócios, em programas de ciência da computação e para introduções mais gerais ao Data Science.

1 É claro que cada autor tem a nítida impressão de que fez a maior parte do trabalho do livro.

Fique atento para os sites dos livros (veja abaixo) para obter informações sobre como conseguir material instrucional útil, incluindo slides de aulas, amostras de questões e problemas, exemplo de instruções de projetos com base na estrutura do livro, perguntas de provas e muito mais.

Mantemos uma lista atualizada de adoções conhecidas no site do livro (*http://www.data-science-for-biz.com/* — conteúdo em inglês). Clique em *Who's Using It* no topo.

Outras Habilidades e Conceitos

Há muitos outros conceitos e habilidades que um cientista de dados precisa saber além dos princípios fundamentais de Data Science. Essas habilidades e conceitos serão discutidos nos Capítulo 1 e 2. Incentivamos o leitor interessado para a visitar o site dos livros para indicação de material para aprender essas habilidades e conceitos adicionais (por exemplo, criação de scripts em Python, processamento de linha de comando Unix, arquivos de dados, formatos comuns de dados, bases de dados e consultas, arquiteturas de Big Data e sistemas como MapReduce e Hadoop, visualização de dados e outros tópicos relacionados).

Seções e Notação

Além de notas de rodapé ocasionais, o livro contém "quadros". Eles são, essencialmente, notas de rodapé estendidas. Nós os reservamos para materiais que consideramos interessantes e válidos, porém longo demais para uma nota de rodapé e uma digressão ao texto principal.

Detalhes Técnicos à Frente — Uma Nota Sobre as Seções Marcadas indicadas por ícones

Os detalhes matemáticos ocasionais são relegados para seções opcionais indicadas por um ícone. Esses títulos de seção trazem um ícone e contém um parágrafo como este. Tais seções "indicadas por ícones" contêm uma matemática mais detalhada e/ou detalhes mais técnicos do que os outros locais, e o parágrafo introdutório explica seu propósito. O livro é escrito de modo que essas seções possam ser puladas sem perda de continuidade, embora, em alguns lugares, lembramos os leitores de que os detalhes aparecem lá.

Construções no texto como (Smith e Jones, 2003) indicam uma referência a uma entrada na bibliografia (neste caso, o artigo ou livro de 2003 de Smith e Jones); "Smith e Jones (2003)" é uma referência semelhante. Uma única bibliografia para o livro inteiro aparece no final.

Neste livro, tentamos manter a matemática ao mínimo, e quando ela aparece está simplificada ao máximo possível sem causar confusão. Para os nossos leitores com formação técnica, alguns comentários podem ser feitos a respeito de nossas escolhas simplificadas.

1. Evitamos a notação Sigma (Σ) e Pi (Π), comumente usadas em livros didáticos para indicar somas e produtos, respectivamente. Em vez disso, simplesmente usamos equações com elipses como esta:

   $$f(x) = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

   Nas seções técnicas "indicadas por ícones", algumas vezes adotamos a notação Sigma e Pi quando essa abordagem de elipse é muito complicada. Supomos que as pessoas que leem essas seções estão um pouco mais acostumadas com a notação matemática e não ficarão confusas.

2. Livros de estatística costumam ser cuidadosos em distinguir entre um valor e sua estimativa, colocando um "chapéu" em variáveis que são estimativas, por isso, em tais livros, você verá uma probabilidade verdadeira denotada p e sua estimativa denotada $\hat{p}$. Neste livro, quase sempre falamos de estimativas de dados, e colocar o circunflexo em tudo torna as equações prolixas e feias. Tudo deve ser considerado como uma estimativa de dados, a menos que se diga o contrário.

3. Simplificamos a notação e removemos variáveis externas onde acreditamos que estejam claras a partir do contexto. Por exemplo, quando discutimos classificadores matematicamente, estamos, tecnicamente, lidando com decisões predicadas sobre vetores de característica. Expressar isso formalmente levaria a equações como:

   $$\hat{f}_R(\mathbf{x}) = x_{\text{Age}} \times \ -1 + 0.7 \times x_{\text{Balance}} + 60$$

   Em vez disso, optamos por algo mais legível, como:

   $$f(\mathbf{x}) = Age \times \ -1 + 0.7 \times \text{Balance} + 60$$

   com o entendimento de que $\boldsymbol{x}$ é um vetor e *Idade* e *Equilíbrio* são componentes dele.

Tentamos ser coerentes com a tipografia, reservando fontes tipográficas de largura fixa como Sepal Width para indicar atributos ou palavras-chave em dados. Por exemplo, no capítulo sobre exploração de texto, uma palavra como "*discutir*" designa uma palavra em um documento, enquanto discutir poderia ser um indício resultante dos dados.

As seguintes convenções tipográficas são usadas neste livro:

Itálico

Indica novos termos, URLs, endereços de e-mail, nomes de arquivos e extensões de arquivos.

`Monoespaçada`

Usada para listagem de programas, bem como dentro de parágrafos para se referir a elementos como variáveis ou nomes de funções, bancos de dados, tipos de dados, variáveis de ambiente, declarações e palavras-chave.

Monoespaçada com itálico

Mostra texto que deve ser substituído por valores fornecidos pelo usuário ou por valores determinados pelo contexto.

Ao longo do livro, colocamos dicas e avisos especiais pertinentes ao material. Eles serão processados de forma diferente, dependendo se você está lendo em papel, PDF ou e-book, da seguinte forma:

Uma frase ou parágrafo composto como este significa uma dica ou sugestão.

Este texto e elemento significa uma nota geral.

Texto reproduzido desta forma significa uma advertência ou precaução. Estes são mais importantes do que as dicas e são usados com moderação.

Usando Exemplos

Além de ser uma introdução ao Data Science, este livro pretende ser útil em discussões e trabalhos do dia a dia na área. Não é preciso permissão para responder a uma pergunta citando este livro e seus exemplos. Agradecemos, mas não exigimos atribuição. A atribuição formal geralmente inclui título, autor, editora e ISBN. Por exemplo: "*Data Science para Negócios* de Foster Provost e Tom Fawcett (Altabooks). Copyright 2013 Foster Provost e Tom Fawcett, 978-1-449-36132-7.

Agradecimentos

Agradecemos aos muitos colaboradores, colegas de área ou não, que prestaram inestimável *feedback*, críticas, sugestões e incentivo com base em muitos manuscritos anteriores. Correndo o risco de deixar de citar alguém, deixe-nos agradecer em particular: Panos Adamopoulos, Manuel Arriaga, Josh Attenberg, Solon Barocas, Ron Bekkerman, Josh Blumenstock, Ohad Brazilay, Aaron Brick, Jessica Clark, Nitesh Chawla, Peter Devito, Vasant Dhar, Jan Ehmke, Theos Evgeniou, Justin Gapper, Tomer Geva, Daniel Gillick, Shawndra Hill, Nidhi Kathuria, Ronny Kohavi, Marios Kokkodis, Tom Lee, Philipp Marek, David Martens, Sophie Mohin, Lauren Moores, Alan Murray, Nick Nishimura, Balaji Padmanabhan, Jason Pan, Claudia Perlich, Gregory Piatetsky-Shapiro, Tom Phillips, Kevin Reilly, Maytal Saar-Tsechansky, Evan Sadler, Galit Shmueli, Roger Stein, Nick Street, Kiril Tsemekhman, Craig Vaughan, Chris Volinsky, Wally Wang, Geoff Webb, Debbie Yuster e Rong Zheng. Também gostaríamos de agradecer, de forma mais geral, aos alunos das aulas de Foster, Mineração de Dados para Análises de Negócios, Data Science na Prática, Introdução ao Data Science e o Seminário de Pesquisa em Data Science. As questões e os problemas que surgiram durante os primeiros rascunhos deste livro forneceram *feedback* substancial para seu aprimoramento.

Agradecemos a todos os colegas que nos ensinaram sobre Data Science e sobre como ensiná-lo ao longo dos anos. Agradecemos especialmente a Maytal Saar-Tsechansky e Claudia Perlich. Maytal graciosamente compartilhou com Foster suas anotações sobre sua aula de mineração de dados muitos anos atrás. O exemplo de classificação com árvore de decisão no Capítulo 3 (agradecimento especial à visualização de "corpos") baseia-se, principalmente, na ideia e no exemplo dela; suas ideias e exemplo foram a origem da visualização comparando a divisão do espaço instância com árvores de decisão e as funções discriminantes lineares no Capítulo 4, o exemplo "David responderá?", no Capítulo 6, baseia-se no exemplo dela, e, provavelmente, outras coisas esquecidas. Claudia lecionou sessões conjuntas de Mineração de Dados para Análise de Negócios/Introdução ao Data Science com Foster durante os últimos anos e lhe ensinou muito sobre Data Science no processo (e além).

Agradecemos a David Stillwell, Thore Graepel e Michal Kosinski por fornecer os dados de curtidas do Facebook para alguns dos exemplos. Agradecemos a Nick Street por fornecer os dados de núcleos celulares e por permitir que usássemos a imagem de tais dados no Capítulo 4. Agradecemos a David Martens por sua ajuda com a visualização dos locais de terminais móveis. Agradecemos a Chris Volinsky por fornecer os dados de seu trabalho no Desafio Netflix. Agradecemos a Sonny Tambe pelo acesso antecipado aos seus resultados sobre tecnologias e produtividade em Big Data. Agradecemos a Patrick Perry por nos indicar o exemplo do *call center* do banco utilizado no Capítulo 12. Agradecemos a Geoff Webb pelo uso do sistema de mineração da associação Magnum Opus.

Acima de tudo, agradecemos às nossas famílias por seu amor, paciência e incentivo.

Uma grande quantidade de software de código-fonte aberto foi utilizado na preparação deste livro e de seus exemplos. Os autores gostariam de agradecer aos desenvolvedores e contribuidores de:

- Python e Perl
- Scipy, Numpy, Matplotlib e Scikit-Learn
- Weka
- O Repositório de Aprendizado de Máquina da Universidade da Califórnia, em Irvine (Bache & Lichman, 2013).

Por fim, gostaríamos de incentivar os leitores a visitar nosso site (*http://www.data-science-for-biz.com* — conteúdo em inglês) para atualizações deste material, novos capítulos, erratas, adendos e conjuntos de slides complementares.

— Foster Provost e Tom Fawcett

CAPÍTULO 1

Introdução: Pensamento Analítico de Dados

Não sonhe pequeno, pois esses sonhos não têm poder
para mover os corações dos homens.

— Johann Wolfgang von Goethe

Os últimos quinze anos testemunharam grandes investimentos em infraestrutura de negócios que têm melhorado a capacidade de coletar dados em toda a empresa. Agora, praticamente todos os aspectos dos negócios estão abertos para a coleta de dados e, muitas vezes, até instrumentados para isso: operações, manufatura, gestão da cadeia de fornecimento, comportamento do cliente, desempenho de campanha de marketing, procedimentos de fluxo de trabalho e assim por diante. Ao mesmo tempo, atualmente, a informação está amplamente disponível em eventos externos, como tendências de mercado, notícias industriais e os movimentos dos concorrentes. Essa ampla disponibilidade de dados levou ao aumento do interesse em métodos para extrair informações úteis e conhecimento a partir de dados — o domínio de data science.

A Onipresença das Oportunidades de Dados

Agora, com grandes quantidades de dados disponíveis, as empresas em quase todos os setores estão focadas em explorá-los para obter vantagem competitiva. No passado, as empresas podiam contratar equipes de estatísticos, modeladores e analistas para explorar manualmente os conjuntos de dados, mas seu volume e variedade superaram muito a capacidade da análise manual. Ao mesmo tempo, os computadores se tornaram muito mais poderosos, a comunicação em rede é onipresente, e foram desenvolvidos algoritmos que podem conectar conjuntos de dados para permitir análises muito mais amplas e profundas do que antes. A convergência desses fenômenos deu origem à aplicação, cada vez mais difundida, de princípios de data science e de técnicas de mineração de dados nos negócios.

Provavelmente, a maior aplicação de técnicas de mineração de dados está no marketing, para tarefas como marketing direcionado, publicidade online e recomendações para venda cruzada. A mineração de dados é usada para gestão de relacionamento com o cliente para analisar seu comportamento a fim de gerenciar o desgaste e maximizar o valor esperado do cliente. A indústria financeira utiliza a mineração de dados para classificação e negociação de crédito e em operações via detecção de fraude e gerenciamento de força de trabalho. Os principais varejistas, do Walmart à Amazon, aplicam a mineração de dados em seus negócios, do marketing ao gerenciamento da cadeia de fornecimento. Muitas empresas têm se diferenciado estrategicamente com data science, às vezes, ao ponto de evoluírem para empresas de mineração de dados.

Os principais objetivos deste livro são ajudá-lo a visualizar problemas de negócios a partir da perspectiva de dados, e a entender os princípios da extração de conhecimento útil a partir deles. Existe uma estrutura fundamental para o pensamento analítico de dados e princípios básicos que devem ser compreendidos. Há também áreas específicas onde intuição, criatividade, bom senso e conhecimento de domínio devem ser exercidos. Uma perspectiva de dados fornecerá estrutura e princípios, e isso lhe dará uma base para analisar sistematicamente tais problemas. Conforme você se aprimora no pensamento analítico de dados, você desenvolve intuição sobre como e onde aplicar a criatividade e o conhecimento de domínio.

Ao longo dos dois primeiros capítulos deste livro, discutimos em detalhes vários temas e técnicas relacionados ao data science e à mineração de dados. Os termos "Data Science" e "Data Mining" são, muitas vezes, utilizados de forma intercambiável, e este último tem desenvolvido vida própria, uma vez que vários indivíduos e organizações tentam tirar proveito do atual alarde que o cerca. Em um nível mais elevado, *data science* é um conjunto de princípios fundamentais que norteiam a extração de conhecimento a partir de dados. *Data mining* é a extração de conhecimento a partir deles, por meio de tecnologias que incorporam esses princípios. Como termo, "data science" muitas vezes é aplicado mais amplamente do que o uso tradicional de "data mining", mas as técnicas de mineração de dados fornecem alguns dos mais claros exemplos de princípios de data science.

É importante compreender data science, mesmo que você nunca vá aplicá-lo. O pensamento analítico de dados permite avaliar propostas para projetos de mineração de dados. Por exemplo, se um funcionário, um consultor ou um potencial alvo de investimento propõe melhorar determinada aplicação de negócios apartir da obtenção de conhecimento de dados, você deve ser capaz de avaliar a proposta de maneira sistemática e decidir se ela é boa ou ruim. Isso não significa que será capaz de dizer se será bem-sucedido — pois projetos de mineração de dados, muitas vezes, exigem experimentação —, mas você deve conseguir identificar falhas óbvias, hipóteses fantasiosas e partes faltando.

Neste livro, descrevemos uma série de princípios fundamentais de data science e ilustramos cada um com, pelo menos, uma técnica de mineração de dados que os incorpore. Para cada princípio, normalmente, há muitas técnicas específicas que o envolvem, assim, neste livro,

escolhemos enfatizar os princípios básicos em vez das técnicas específicas. Dito isso, não daremos muita importância à diferença entre data science e mineração de dados, exceto nos casos em que isso terá um efeito substancial na compreensão dos conceitos efetivos.

Vamos analisar dois breves estudos de caso de análise de dados para extrair padrões preditivos.

Exemplo: O Furacão Frances

Considere um exemplo de uma história do New York Times de 2004:

> O furacão Frances estava a caminho, avançando pelo Caribe, ameaçando atingir a costa atlântica da Flórida. Os residentes se mudaram para terrenos mais elevados, porém distantes, em Bentonville, Arkansas. Executivos das lojas Walmart decidiram que a situação oferecia uma grande oportunidade para uma de suas mais recentes armas orientadas em dados: a tecnologia preditiva.
>
> Uma semana antes de a tempestade atingir a costa, Linda M. Dillman, diretora executiva de informação, pressionou sua equipe para trabalhar em previsões baseadas no que havia acontecido quando o furacão Charley apareceu, várias semanas antes. Com o apoio dos trilhões de bytes de histórico de compras contidos no banco de dados do Walmart, ela sentiu que a empresa poderia "começar a prever o que aconteceria, em vez de esperar que acontecesse". (Hays, 2004)

Pense *porque* previsões orientadas em dados podem ser úteis neste cenário. Elas podem ser úteis para prever que as pessoas na trilha do furacão comprariam mais garrafas de água. Talvez, mas isso parece um pouco óbvio, e por que precisaríamos de data science para descobrir isso? Pode ser útil para projetar o *aumento* nas vendas devido ao furacão, assegurando que os Walmarts locais estejam bem abastecidos. Talvez a mineração de dados possa revelar que determinado DVD esgotou na trilha do furacão — mas, talvez, isso tenha acontecido naquela semana em Walmarts de todo o país, e não apenas onde o furacão era iminente. A previsão pode, de certa forma, ser útil, mas provavelmente é mais genérica do que a Sra. Dillman pretendia.

Seria mais valioso descobrir padrões não tão óbvios causados pelo furacão. Para fazer isso, os analistas podem examinar o grande volume de dados do Walmart a partir de situações prévias semelhantes (como o furacão Charley) para identificar demanda local *incomum* de produtos. A partir desses padrões, a empresa pode ser capaz de antecipar a demanda incomum de produtos e correr para abastecer as lojas antes da chegada do furacão.

De fato, foi o que aconteceu. O *New York Times* (Hays, 2004) relatou que: "... especialistas exploraram os dados e descobriram que as lojas realmente precisariam de certos produtos — e não apenas das habituais lanternas. 'Não sabíamos, no passado, que havia tido um aumento nas vendas de Pop-Tarts de morango, sete vezes acima do normal, antes de um furacão', disse a Sra. Dillman em uma entrevista recente. 'E o principal produto pré-furacão mais vendido era a cerveja.[1]

1 É claro! O que combina melhor com Pop-Tarts de morango do que uma boa cerveja gelada?

Exemplo: Prevendo a Rotatividade de Cliente

Como são realizadas essas análises de dados? Considere um segundo e mais típico cenário de negócios e como ele pode ser tratado a partir de uma perspectiva de dados. Este problema servirá como um exemplo recorrente que iluminará muitas das questões levantadas neste livro e fornecerá um quadro de referência comum.

Vamos supor que você acabou de ingressar em um ótimo trabalho analítico na MegaTelCo, uma das maiores empresas de telecomunicação nos Estados Unidos. Eles estão tendo um grande problema com a retenção de clientes no negócio de produtos e serviços sem fio. Na região do Médio Atlântico, 20% dos clientes de telefonia celular abandonam o serviço quando seus contratos vencem, e está ficando cada vez mais difícil adquirir novos clientes. Como agora o mercado dos telefones celulares está saturado, o enorme crescimento do mercado sem fio diminuiu. Agora, as empresas de comunicação estão engajadas em batalhas para atrair os clientes da concorrência, ao mesmo tempo que mantêm seus próprios. A transferência de clientes de uma empresa para outra é chamada de *rotatividade*, e é algo dispendioso em todos os sentidos: uma empresa precisa gastar em incentivos para atrair um cliente, enquanto outra empresa perde rendimento quando o cliente vai embora.

Você foi chamado para ajudar a entender o problema e encontrar uma solução. Atrair novos clientes é muito mais caro do que manter os que já existem, por isso, uma boa verba de marketing é alocada para evitar a rotatividade. O marketing já projetou uma oferta especial de retenção. Sua tarefa é elaborar um plano preciso, passo a passo, para saber como a equipe de data science deve usar os vastos recursos de dados da MegaTelCo para decidir quais clientes devem receber uma oferta especial de retenção antes do término de seus contratos.

Pense cuidadosamente sobre quais dados você pode usar e como serão usados. Pense, especificamente, como a MegaTelCo deve escolher um conjunto de clientes para receber sua oferta a fim de melhor reduzir a rotatividade para uma verba de incentivo em particular? Responder a essa pergunta é muito mais complicado do que pode parecer inicialmente. Voltaremos a este problema várias vezes, acrescentando sofisticação a nossa solução conforme desenvolvemos uma compreensão dos conceitos fundamentais de data science.

Na verdade, a retenção de clientes tem sido uma das grandes utilizações para tecnologias de mineração de dados — especialmente nos setores de telecomunicação e finanças. Esses, de forma mais geral, foram alguns dos primeiros e mais amplos adotantes das tecnologias de mineração de dados, por motivos que serão discutidos mais adiante.

Data Science, Engenharia e Tomada de Decisão Orientada em Dados

Data science envolve princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados. Neste livro, analisaremos o objetivo primordial de

data science, que é o aprimoramento da tomada de decisão, uma vez que isso geralmente é de interesse direto para os negócios.

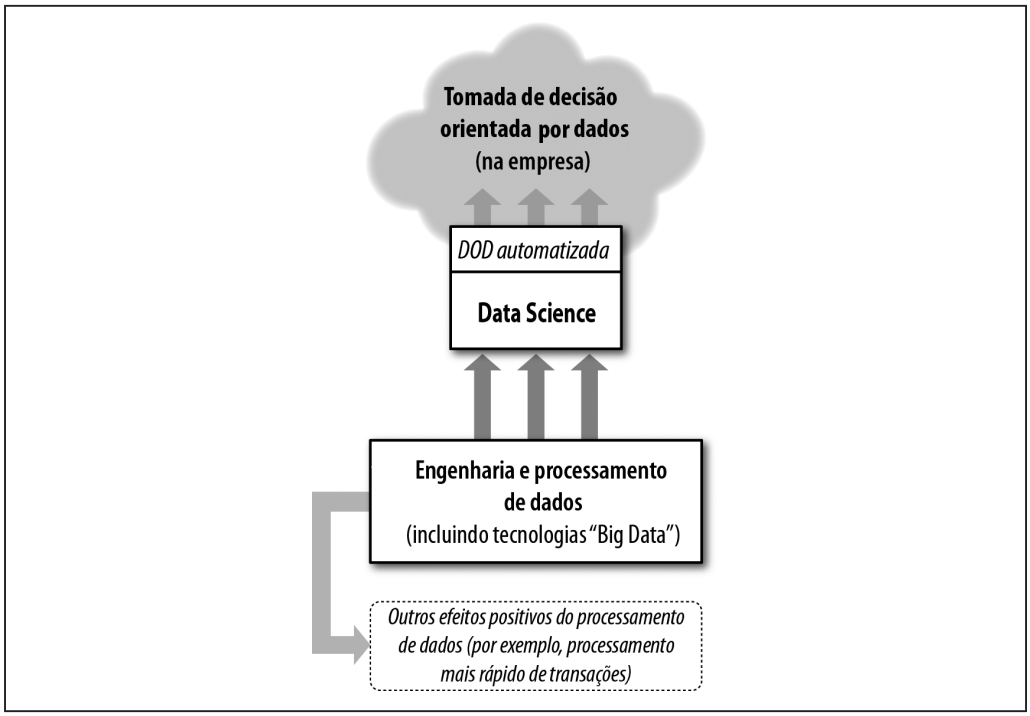

Figura 1-1. Data science no contexto dos diversos processos relacionados a dados na organização.

A Figura 1-1 coloca data science no contexto de diversos outros processos intimamente associados e relacionados com dados na organização. Ela distingue data science de outros aspectos do processamento de dados que estão ganhando cada vez mais atenção nos negócios. Vamos começar pelo topo.

Tomada de decisão orientada por dados (DOD) refere-se à prática de basear as decisões na análise dos dados, em vez de apenas na intuição. Por exemplo, um negociante poderá selecionar anúncios baseados puramente em sua longa experiência na área e em sua intuição de que funcionará. Ou, pode basear sua escolha na análise dos dados sobre a forma como os consumidores reagem a diferentes anúncios. Ele também poderia utilizar uma combinação dessas abordagens. A DOD não é uma prática do tipo "tudo ou nada", e diversas empresas a adotam em maior ou menor grau.

Os benefícios da tomada de decisão orientada por dados têm sido demonstrados conclusivamente. O economista Erik Brynjolfsson e seus colegas do MIT e da Penn's Wharton School realizaram um estudo de como DOD afeta o desempenho das empresas (Brynjolfsson, Hitt & Kim, 2011). Eles desenvolveram uma medida de DOD que classifica as empresas quanto ao uso de dados para tomar decisões. Eles mostram que, estatisticamente, quanto mais orientada por dados, mais produtiva uma empresa é — mesmo controlando uma vasta gama

de possíveis fatores de confusão. E as diferenças não são pequenas. Um desvio padrão a mais na escala de DOD está associado com um aumento de 4%–6% na produtividade. A DOD também está correlacionada com maior retorno sobre ativos, retorno sobre o patrimônio líquido, utilização de ativos e valor de mercado e a relação parece ser causal.

O tipo de decisões que interessam neste livro se enquadram, principalmente, em dois tipos: (1) decisões para as quais "descobertas" precisam ser feitas nos dados e (2) decisões que se repetem, principalmente em grande escala, e, assim, a tomada de decisão pode se beneficiar até mesmo de pequenos aumentos na precisão deste processo com base em análise de dados. O exemplo do Walmart, acima, ilustra um problema tipo 1: Linda Dillman gostaria de descobrir "fatos" que ajudariam o Walmart a se preparar para a chegada iminente do furacão Frances.

Em 2012, o competidor do Walmart, Target, virou notícia por um caso próprio de tomada de decisão orientada por dados, também um problema tipo 1 (Duhigg, 2012). Como a maioria dos varejistas, a Target se preocupa com os hábitos de compra dos consumidores, o que os motiva e o que pode influenciá-los. Os consumidores tendem a permanecer inertes em seus hábitos e fazê-los mudar é difícil. Quem tomava as decisões na Target sabia, no entanto, que a chegada de um novo bebê na família é um momento em que as pessoas mudam significativamente seus hábitos de compras. Nas palavras do analista da Target, "assim que percebemos que estão comprando nossas fraldas, eles comprarão todo o resto também." A maioria dos varejistas sabe disso e, portanto, competem entre si tentando vender produtos de bebês para novos pais. Como a maior parte dos registros de nascimento é pública, os varejistas obtêm informações sobre nascimentos e enviam ofertas especiais para os novos pais.

No entanto, a Target desejava sair na frente da concorrência. Eles estavam interessados em saber se conseguiriam *prever* se as pessoas *estavam esperando* um bebê. Se pudessem, ganhariam uma vantagem ao fazer ofertas antes de seus concorrentes. Usando técnicas de data science, a Target analisou dados históricos sobre os clientes que souberam *posteriormente* que estavam grávidas, e foi capaz de obter informações que poderiam predizer quais consumidores estavam esperando um bebê. Por exemplo, mulheres grávidas costumam mudar a dieta, o guarda-roupa, as vitaminas e assim por diante. Esses indicadores podem ser extraídos dos dados históricos, montados em modelos preditivos e, em seguida, implantados em campanhas de marketing. Discutiremos modelos preditivos de forma mais detalhada conforme avançarmos no livro. No momento, é suficiente entender que um modelo preditivo abstrai a maior parte da complexidade do mundo, concentrando-se em um conjunto específico de indicadores que se correlacionam, de algum modo, com uma quantidade de interesses (quem apresentará rotatividade ou quem comprará, quem está grávida, etc). O mais importante, nos exemplos Walmart e Target, é que a análise dos dados não estava testando uma simples hipótese. Ao invés disso, os dados foram explorados com a esperança de que algo útil pudesse ser descoberto.[2]

2 A Target foi tão bem-sucedida que este caso levantou questões éticas sobre a implantação de tais técnicas. Preocupações relacionadas à ética e à privacidade são interessantes e muito importantes, mas deixaremos essa discussão para outro momento.