

ЗМІСТ

<i>Передмова</i>	13
Розділ 1. Вступ: дата-аналітичне мислення	22
Всюдисутність можливостей даних	22
Приклад: ураган Френсіс	24
Приклад: передбачення плинності клієнтів	25
Data science, програмування і ухвалення рішень на основі даних	26
Обробка даних та «біг-дата»	29
Від біг-дати 1.0 до біг-дати 2.0	30
Вміння працювати з даними і data science як стратегічне надбання	31
Дата-аналітичне мислення	34
Ця книжка	36
Майнинг даних і data science, новий погляд	36
Пробірки — не суть хімії: data science і робота дата-спеціаліста	38
Підсумки	39
Розділ 2. Задачі бізнесу і рішення data science	41
Від завдань бізнесу до завдань майнингу даних	42
Контрольовані і неконтрольовані методи	47
Майнинг даних і його результати	48
Процес майнингу даних	50
Розуміння бізнесу	51
Розуміння даних	52
Підготовка даних	53
Моделювання	54
Оцінка	55
Запуск	56
Складнощі управління командою data science	58
Інші техніки й технології аналітики	59
Статистика	59
Постановка запиту бази даних	61
Організація сховища даних	63
Регресійний аналіз	63

Машинне навчання і майнинг даних	64
Як відповідати на питання бізнесу за допомогою цих технік	65
Підсумки	66
Розділ 3. Вступ у прогностичне моделювання:	
від кореляції до контрольованої сегментації	67
Моделі, індукція і прогнози	69
Направлена сегментація	72
Вибір інформативних атрибутів	74
Приклад: вибір атрибутів з приростом інформації	81
Направлена сегментація з моделями з деревовидною структурою	86
Візуалізуємо сегментації	92
Дерева як набори правил	94
Підрахунок вірогідності	96
Приклад: розв'язуємо проблему з плинністю за допомогою індукції дерева рішень	98
Підсумки	103
Розділ 4. Як навчити модель під дані	
Класифікація проти математичних функцій	107
Лінійні дискримінантні функції	109
Оптимізуємо цільову функцію	112
Приклад майнингу лінійного дискримінанта з даних	112
Лінійні дискримінантні функції для призначення оцінок зразкам і їхнього ранжування	114
Машини опорних векторів, коротко	115
Регресія і математичні функції	118
Визначення вірогідності належності до класу і логістична «регресія»	121
* Логістична регресія: деякі технічні деталі	124
Приклад: логістична регресія проти індукції дерева рішень	127
Нелінійні функції, машини опорних векторів і нейронні мережі	132
Підсумки	134
Розділ 5. Перенавчання і як його уникнути	
Генералізація	136
Перенавчання	138
Досліджуємо перенавчання	138
Контрольні дані і графік навчання	139
Перенавчання в індукції дерева рішень	141
Перенавчання в математичних функціях	143
Приклад: перенавчання лінійних функцій	145
* Приклад: чому перенавчання — це погано?	149
Від оцінки контрольних даних до перехресної перевірки	151
Повертаємося до набору даних про плинність	155

Криві навчання	156
Як уникати перенавчання і контролювати складність	158
Як уникнути перенавчання в індукції дерева рішень	159
Загальний метод, як уникати перенавчання	160
* Як уникнути перенавчання при оптимізації параметрів	162
Підсумки	166
Розділ 6. Подібність, сусіди й кластери	168
Подібність і відстань	169
Логіка «найближчого сусіда»	172
Приклад: аналітика по віскі	172
Найближчі сусіди у предиктивному моделюванні	175
Скільки сусідів і як впливає?	178
Геометрична інтерпретація, перенавчання і контроль складності	180
Проблеми методу найближчих сусідів	183
Деякі важливі технічні деталі, пов'язані з подібностями й сусідами	186
Гетерогенні атрибути	186
* Інші функції відстані	187
* Об'єднувальні функції: рахуємо оцінки від сусідів	191
Кластеринг	193
Приклад: повернімося до аналітики віскі	193
Ієрархічний кластеринг	194
Повертаємося до найближчих сусідів: кластеринг навколо центроїда	199
Приклад: кластеризуємо статті про новини бізнесу	203
Як зрозуміти результати кластерингу	207
* Як використовувати прогнозування залежної змінної, щоб генерувати описи кластерів	209
Крок назад: вирішення задач бізнесу і дослідження даних	212
Підсумки	214
Розділ 7. Аналітичне дизайн-мислення I:	
що таке хороша модель?	216
Як оцінювати класифікатори	217
Проста точність і проблеми з нею	218
Матриця невідповідностей	219
Задачі з незбалансованими класами	219
Проблема неоднакової ціни й переваг	221
Генералізація після класифікації	223
Ключовий сценарій аналітики: очікуване значення	224
Як створити шаблон використання класифікатора за допомогою очікуваного значення	225
Використання очікуваного значення для оцінки класифікатора	226
Оцінка, базова ефективність та інвестиції в дані	235
Підсумки	238

Розділ 8. Візуалізуємо ефективність моделі	240
Ранжування замість класифікацій	241
Криві прибутку	243
Графіки й криві помилок	246
Ділянка під кривою ROC (AUC)	251
Сумарна відповідь і підйомні криві	251
Приклад: аналітика ефективності моделювання	
в задачі з плінністю	254
Підсумки	263
Розділ 9. Докази й вірогідності	265
Приклад: рекламне таргетування онлайн-користувачів	265
Імовірнісне поєднання доказів	267
Сумарна ймовірність та незалежність	268
Правило Баєса	270
Застосування правила Баєса в data science	271
Умовна незалежність і Наївний Баєс	273
Плюси й мінуси Наївного Баєса	275
Модель доказу «Підйом»	277
Приклад: підйоми зразків з фейсбучних лайків	279
Докази в дії: таргетуємо користувачів	280
Підсумки	281
Розділ 10. Репрезентація і майнинг тексту	283
Чому текст важливий	284
Чому текст — це складно	285
Репрезентація	285
Мультимножина слів	286
Частота термів	287
Вимірювання розрідженості:	
зворотна частота документа	289
Поєднуємо їх: TFIDF	290
Приклад: джазові музиканти	291
* Як пов'язані IDF і ентропія	295
Способи без використання мультимножини	297
N-грамні послідовності	297
Добування іменованих сутностей	298
Тематичні моделі	299
Приклад: майнинг новин для прогнозу	
змін вартості акцій	300
Задача	300
Дані	303
Перед-обробка даних	305
Результати	306
Підсумки	310

Розділ II. Аналітичне мислення для рішень II:	
аналітична інженерія	311
Таргетування на найкращих потенційних	
благодійників через розсилку	312
Принцип очікуваного значення: розбиваємо	
бізнес-задачу і збираємо рішення по шматках	312
Короткий відступ про помилку вибірки	315
Наш приклад із плинністю — розглядаємо ще детальніше	316
Принцип очікуваного значення:	
структуруємо складнішу задачу	316
Оцінка впливу стимуляції	317
Від розкладання очікуваного значення	
до рішення data science	319
Підсумки	322
Розділ 12. Інші задачі й техніки data science	323
Взаємозв'язки й збіги: пошук об'єднаних одиниць	324
Вимірювання неочікуваності: підйом і балансування	325
Приклад: пиво і лотерейні квитки	326
Асоціації у фейсбучних лайках	327
Профілювання: пошук типової поведінки	331
Передбачення зв'язків і соціальні рекомендації	336
Зменшення кількості даних, латентна інформація	
і рекомендації фільмів	338
Ангажованість, варіативність і ансамблеві методи	341
Базоване на даних пояснення причин і приклад	
з вірусного маркетингу	344
Підсумки	346
Розділ 13. Data science і бізнес-стратегія	347
Дата-аналітичне мислення, повернення	347
Як досягти переваги над конкурентами	
за допомогою data science	349
Як втримати конкурентну перевагу	
за допомогою data science	351
Дуже суттєва історична перевага	351
Унікальна інтелектуальна власність	352
Унікальні неявні додаткові активи	352
Найкращі дата-саєнтисти	353
Найкращий менеджмент data science	354
Як знаходити й утримувати дата-саєнтистів	
та їхні команди	356
Дослідження прикладів із практики data science	358
Будьте готові почути креативні ідеї від будь-кого	359
Будьте готові оцінювати пропозиції	
за проектами data science	359
Приклад пропозиції майнінгу даних	360
Недоліки пропозиції Big Red	361
Повнота потенціалу data science у фірмі	362

Розділ 14. Висновки	365
Фундаментальні концепції data science	365
Застосуємо фундаментальні концепції до нової задачі:	
майнинг даних з мобільних пристроїв	368
Як інакше підійти до рішення задач	371
Чого не можуть дані: обчислення з оператором	
у контурі управління	372
Приватність, етика і майнинг даних про людей	375
Що ще можна сказати про data science?	376
Останній приклад: від краудсорсингу до клаудсорсингу	377
Наостанок	378
 Додаток А. Інструкція до оцінки пропозиції	 380
Розуміння бізнесу й даних	380
Підготовка даних	381
Моделювання	381
Оцінка й запуск	382
 Додаток Б. Ще один зразок пропозиції	 383
Сценарій і пропозиція	383
Недоліки пропозиції CGC	384
 Глосарій	387
Бібліографія	392

ПЕРЕДМОВА

- «**D**ata Science для бізнесу» призначена для декількох типів читачів:
- Людей бізнесу, які працюватимуть зі спеціалістами з аналізу даних, управлятимуть проектами, які орієнтовані на дані, або робитимуть венчурні інвестиції в data science компанії;
 - Розробників, які будуть впроваджувати пов'язані з data science рішення в життя, і
 - Людей, які хочуть у майбутньому професійно працювати з даними.

Це не книжка про алгоритми, і вона не може замінити книжку про алгоритми. Ми свідомо вирішили не зосереджуватись на алгоритмах. Ми вважаємо, що в основі того, як діставати корисну інформацію з даних, лежить відносно невелика кількість фундаментальних концепцій або принципів. Ці концепції і стали *основою* для багатьох відомих алгоритмів добування даних. Навіть більше: ці концепції лежать в основі аналізу датоцентричних управлінських проблем, створення та оцінки рішень із використанням data science і оцінки загальних стратегій та пропозицій data science. Відповідно й ми організували цю книжку довкола цих загальних принципів, а не конкретних алгоритмів. Там, де обов'язково потрібно пояснювати деталі з процедури, ми поєднували текст і діаграми — вважаємо, що так зрозуміліше, ніж просто крок за кроком прописати алгоритм.

Серйозний математичний бекграунд для цієї книжки не потрібен. Але матеріал сам по собі дещо технічний — у нас було на меті розповісти найважливіше з того, що потрібно знати про data science, а не просто наговорити великих незрозумілих загальних слів. Загалом, ми постаралися мінімізувати математику й пояснити все якомога більш «схематично».

Колеги з галузі кажуть, що книжка просто безцінна для тих, кому треба в бізнесі постійно працювати і зі стандартними процесами, і з технічно-

виробничою стороною, і з людьми, що займаються data science. Але група людей, про яку вони говорять, насправді дуже маленька — і нам було цікаво, скільком людям це насправді може бути важливо (див. розділ 5!). Ідеальну картину ми бачимо так: коли дата-науковцю потрібно почати працювати з новою командою розробників чи менеджерів, він дає їм цю книжку і каже: якщо ви справді хочете придумати/застосувати в бізнесі data science рішення екстра-класу, нам усім потрібно знати з цієї теми одне й те саме.

Колеги також кажуть, що книжка згодилася їм так, як ми того не очікували: щоб готуватися до співбесід із кандидатами на посади з data science. У бізнесі дуже сильний попит на спеціалістів з data science, і він щоразу зростає. Як результат, дедалі більше людей, які шукають роботу, починають називати себе дата-спеціалістами. Кожен кандидат на посаду такого спеціаліста має розуміти фундаментальні засади професії, які описані в цій книжці. (Колеги з галузі розповідають, що вони дуже здивовані тим, скільки людей насправді цих засад не розуміють. Ми напівсерйозно пишемо про це в «Записках Кліффа до співбесіди з кандидатом по data science»).

НАШ КОНЦЕПТУАЛЬНИЙ ПІДХІД ДО DATA SCIENCE

У цій книжці ми зібрали колекцію найфундаментальніших принципів data science. Деякі з них винесено в заголовки розділів, про інші просто згадано у тексті (і тому вони не обов'язково виокремлені як фундаментальні принципи). Цими принципами описаний весь процес, від передбачення того, яка виникне проблема, до застосування технік data science і використання результатів на те, щоб обрати кращий із варіантів рішення. Ті самі принципи лежать в основі великої кількості аналітичних методів і технік, які застосовуються в бізнесі.

Ці принципи можна розбити на три великі групи:

1. Принципи, які стосуються того, яке взагалі у data science місце в компаніях і конкуренції, включно з тим, яким способом можна зацікавлювати й структурувати команди по data science і робити з них успішну частину компанії, яким способом перетворити data science на конкурентну перевагу і які можна використовувати тактики, щоб добре давати собі раду із проектами, пов'язаними з data science.
2. Загальні принципи аналітичного мислення з використанням даних. Вони допоможуть визначити прийнятні дані й застосувати відповідні методи. Серед цих принципів — *процес майнінгу даних*, а також підбірка різних завдань на майнінг даних високого рівня.
3. Загальні принципи того, яким чином отримувати із даних потрібну інформацію — те, для чого власне й існує велика кількість завдань з data science та алгоритми їхнього виконання.

ВСТУП: ДАТА-АНАЛІТИЧНЕ МИСЛЕННЯ

Не мрійте маленькими мріями; вони не здатні рухати серцями мужів.

Йоганн Вольфганг Гете

За останні 15 років у інфраструктуру бізнесу прийшли величезні інвестиції, від чого можливостей збирати дані по всьому підприємству стало куди легше. Майже кожен аспект бізнесу тепер відкритий до збору даних і часто навіть під це підлаштований: операції, виробництво, управління логістикою, поведінка користувачів, ефективність маркетингової кампанії, організація робочого процесу тощо. Водночас інформація широко доступна і поза межами компаній: тренди ринку, новини галузі, дії конкурентів. Через загальну доступність даних зростає цікавість до методів, якими із даних можна дістати корисну інформацію та знання, — царини data science.

Всюдисутність можливостей даних

Тепер, коли стала доступною безмежна кількість даних, компанії в майже кожній галузі зосереджено намагаються експлуатувати її так, щоб отримати конкурентну перевагу. У минулому фірми наймали команди статистиків, розробників моделей і аналітиків, щоб вони вручну досліджували отримані дані, але тепер обсяги й різноманіття цих даних вже вийшло далеко за межі, де їх можна опрацювати вручну. Водночас комп'ютери стали набагато потужнішими, мережева кооперація — поширеним явищем, а алгоритми розвинулися так, що можуть поєднувати бази даних між собою, і так проводити ширший і глибший аналіз, ніж той, що взагалі був раніше можливий. Збіглися два феномени, і відтак у бізнесі почали дедалі активніше використовувати техніки дата-майнінгу і принципи науки про дані.

Найширше, мабуть, техніки дата-майнингу використовують у маркетингу — для таргетування, онлайн-реклами і рекомендацій для крос-продажу. Дата-майнинг використовують для управління стосунками з умовним клієнтом, аналізу поведінки клієнта, щоб мати змогу контролювати тертя і зробити споживчу цінність максимально високою. У фінансовій галузі дата-майнингом користуються для того, щоб створювати кредитні рейтинги й торгувати в кредит, а також щоб визначати шахраїв і управляти персоналом. Великі ритейлери, наприклад, Walmart чи Amazon, використовують дата-майнинг у своєму бізнесі всюди: і в маркетингу, і в управлінні логістикою. Багато компаній стратегічно диференціювалися за допомогою data science, деякі аж так сильно, що перетворилися на компанії з дата-майнингу.

Головна мета цієї книжки — допомогти вам побачити бізнес-проблеми з погляду даних і зрозуміти, за яким принципом із даних можна видобувати корисну інформацію. У дата-аналітичного мислення є фундаментальна структура і базові принципи, які потрібно розуміти. Подекуди тут також потрібно застосовувати інтуїцію, креативність, здоровий глузд і знання проблемної ділянки. Якщо ви дивитиметеся на проблеми з погляду даних, у вас буде структура і принципи, а відтак ви отримаєте шаблон, за яким можна буде систематично аналізувати подібні проблеми. Коли ви краще освоїте дата-аналітичне мислення, то вже почнете інтуїтивно розуміти, де тут можна придумати креативне рішення, а де потрібні знання проблемної ділянки.

У перших двох розділах цієї книжки ми детально обговорюватимемо різноманітні теми й техніки, які стосуються data science і дата-майнингу. Терміни «data science» і «дата-майнинг» часто вважають взаємозамінними, але коли різні компанії та індивідууми почали намагатися заробляти на дата-хайпі, перший термін зажив власним життям. Якщо не вдаватися в подробиці, data science — це набір принципів, якими потрібно керуватися, щоб з даних отримати інформацію. Дата-майнинг — це вилучення інформації з даних, за допомогою технологій, які створені за цими принципами. Сам термін «data science» використовують частіше, ніж «дата-майнинг», але техніки дата-майнингу — це нерідко найкращі ілюстрації принципів data science.



Розуміти data science важливо, навіть якщо ви взагалі не збираєтеся її застосовувати. Із дата-аналітичним мисленням ви зможете оцінювати пропозиції щодо проектів із дата-майнингу. Наприклад, якщо співробітник, консультант або потенційна інвестиційна мета запропонують краще застосування для конкретного бізнесу із використанням інформації, яку можуть дати дані, ви зможете системно оцінити пропозиції і вирішити, чи розважна вона, а чи в ній є недоліки. Це не означає, що ви знатимете, чи спрацює запропоноване — в проектах із дата-майнингом для цього найчастіше потрібно спробувати — але ви зможете помітити явні недоліки, нереалістичні припущення й частини, яких не вистачатиме.

У цій книжці ми опишемо низку фундаментальних принципів data science, і проілюструємо кожен із них щонайменше однією технікою дата-майнінгу, в якій буде використано цей принцип. Зазвичай кожен із принципів застосовується в багатьох техніках, тож у цій книжці ми вирішили скоріше говорити про базові принципи, аніж про конкретні техніки. А отже, ми не будемо сильно наполягати на різниці між data science і дата-майнінгом, якщо тільки це не буде напряму впливати на пояснення принципів.

Розгляньмо два невеликих кейси з аналізу даних, де потрібно отримати прогнозовані сценарії.

ПРИКЛАД: УРАГАН ФРЕНСІС

Розгляньмо приклад зі статті 2004 року в New York Times.

Ураган Френсіс на повній швидкості мчав через Карибське море і загрожував ударити просто по атлантичному узбережжю Флориди. Мешканці узбережжя втекли туди, де було повище, але управління магазинів Wal-Mart у Бентонвілі вирішили, що в цій ситуації буде дуже доречно використати їхню найновішу зброю... технології передбачення.

За тиждень до того, як ураган мав дістатися землі, керівниця інформаційного управління Wal-Mart Лінда Ділман наказала своїм працівникам створити прогнози погоди на основі того, що сталося, коли за кілька тижнів до того налетів ураган Чарлі. На основі трильйонів байтів історії покупок, які були у сховищі даних Wal-Mart, вона вирішила, що компанія може «почати передбачати, що станеться, а не чекати, коли це станеться», сказала вона. (Гейс, 2004)

Подумайте, чому передбачення на основі даних в такому сценарії може бути корисним. Воно може бути корисним, тому що так можна передбачити, що люди, які тікатимуть від урагану, будуть купувати більше води у пляшках. Можливо, але це трішки очевидно, та й хіба потрібні дані, щоб це зрозуміти? Воно може бути корисним, щоб спрогнозувати, *на скільки саме піднімуться продажі* через ураган, щоб закупити в місцеві Wal-Mart потрібну кількість товару. Можливо, дата-майнінг покаже, що через ураган зростають продажі певних DVD — але можливо, вони того тижня розпродалися у Wal-Mart по всій країні, не тільки там, куди сунув ураган. Передбачення могло бути корисним, але, мабуть, загальнішим, аніж те, яке хотіла отримати міс Ділман.

Було би куди цінніше пошукати пов'язані з ураганом патерни, які були б не такі очевидні. Щоб це зробити, аналітики могли би оцінити величезний обсяг даних Wal-Mart із попередніх, подібних ситуацій (наприклад, коли був

мога більше скоротити плинність з урахуванням бюджету? На це питання відповісти куди складніше, ніж може здатися спочатку. У книжці ми неодноразово повертатимемося до цього завдання, і в силу того, як ми розумітимемо принципи data science, наше рішення ставатиме дедалі елегантнішим.



У реальності технології дата-майнінгу переважно і використовували, коли розбиралися з відтоком клієнтів — особливо в телекомунікаційному та фінансовому бізнесах. Вони одними з найперших почали широко використовувати технології дата-майнінгу, з причин, які ми обговорюватимемо пізніше.

DATA SCIENCE, ПРОГРАМУВАННЯ І УХВАЛЕННЯ РІШЕНЬ НА ОСНОВІ ДАНИХ

Data science — це принципи, процеси й техніки, потрібні для розуміння феноменів через (автоматичний) аналіз даних. У цій книжці ми говоритимемо про прийняття рішень як про головну мету data science, оскільки зазвичай саме це напряму цікавить бізнес.

На рис. 1.1 data science поставлена в контекст різноманітних пов'язаних із даними та близьких до неї процесів у організації. За ним можна розі-

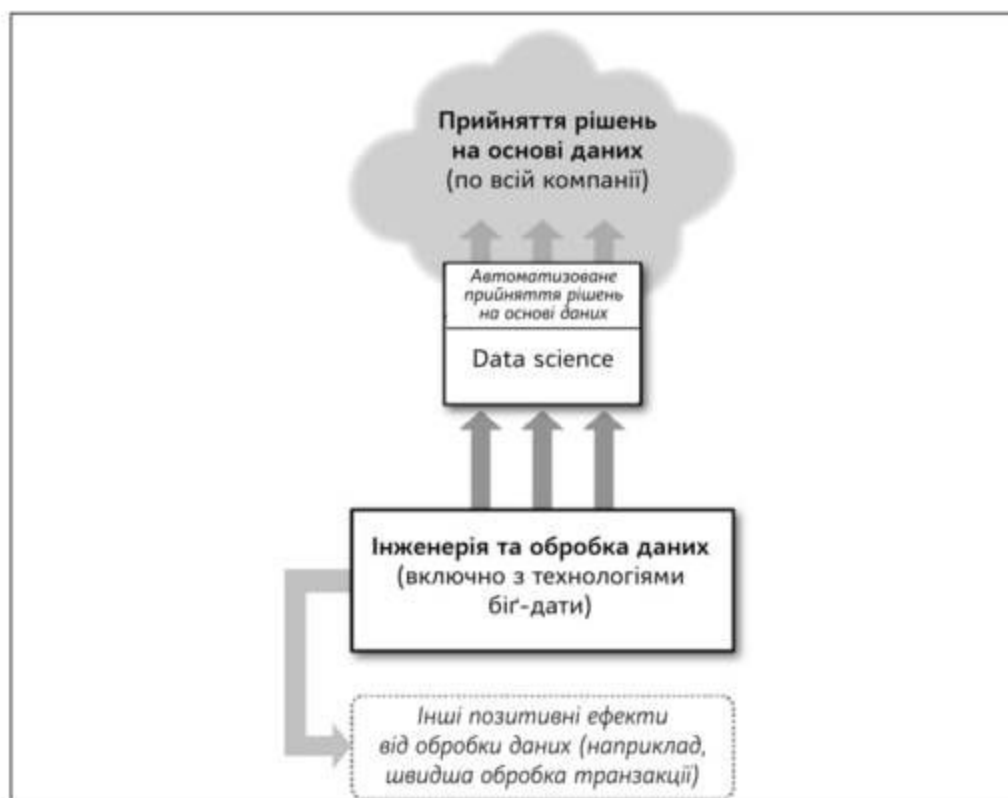


Рис. 1.1. Data science у контексті різних пов'язаних із даними процесів у організації

братися, чим data science відрізняється від обробки даних, до якої в бізнес-середовищі зараз зростає інтерес. Почнімо згори.

Ухвалювання рішення на основі даних — це ґрунтувати рішення на аналізі даних, а не лише на інтуїції. Наприклад, маркетолог може вибрати рекламу, засновуючись суто на своєму великому досвіді роботи в галузі та інтуїції до прийомів, які працюють. Або ж на аналізі даних про те, як споживачі реагують на різні реклами. Можна також використати обидва підходи. Ухвалення рішень на основі даних — не такий підхід, де або все, або нічого, і в різних компаніях до цього вдаються де більше, а де й менше.

Переваги ухвалення рішень на основі даних підтверджені беззаперечно. Економіст Ерік Брінюлфссон та його колеги з MIT та Ліги Плюща дослідили, як прийняття рішень на основі даних впливає на продуктивність компаній (Brynjolfsson, Hitt, & Kim, 2011). Вони розробили оцінку такого підходу, який ранжував компанії за тим, наскільки активно вони використовують дані, коли приймають рішення в компанії. Дослідження показало, що за статистикою, що більше в компанії орієнтуються на дані, то продуктивніша вона — вона навіть може контролювати велику кількість розбіжних факторів. І різниця немаленька. Одне стандартне відхилення вгору по шкалі прийняття рішень на основі даних — це зростання продуктивності на 4–6 %. Ухвалення рішень на основі даних корелюється також із вищим прибутком на активи, рентабельністю капіталу, використанням наявних ресурсів і ринковою цінністю, і схоже, що ці фактори між собою пов'язані.

Рішення, які цікавитимуть нас у цій книжці, можна поділити на два типи: 1) рішення, для яких у межах даних потрібно робити «відкриття», 2) рішення, які повторюються, особливо коли масштаб величезний, і навіть якщо прийняття рішень на основі аналізу даних стане трішечки точнішим, це може позитивно вплинути на процес прийняття рішень загалом. Наведений вище приклад із Wal-Mart — ілюстрація проблеми першого типу: Лінді Ділман потрібна була інформація, як допомогти Walmart підготуватися до неминучого урагану Френсис.

У 2012 році магазин Target, конкурент Walmart, потрапив у новини зі своєю історією про прийняття рішень на основі даних. Це теж була проблема першого типу (Duhigg, 2012). Target, як і переважна більшість роздрібних торговців, переймається купівельними звичками людей, тим, що їх підштовхує купувати і як на них можна вплинути. Зазвичай у покупців звички досить-таки усталені й змінити їх дуже складно. Однак люди, які приймали рішення в Target, знали, що один момент, коли купівельні звички суттєво змінюються, існує. Це момент, коли в людей з'являється дитина. «Якщо ми зробимо так, що вони купуватимуть у нас підгузки — вони купуватимуть у нас і все інше». Більшість роздрібних магазинів про це знають і конкурують

У різних галузях автоматичне ухвалення рішень набуло різних масштабів. Галузі фінансів і телекомунікацій почали застосовувати це раніше, переважно тому, що мережі даних у них розвивалися не за роками, і вони проводили обрахунки величезного масштабу. Вони могли об'єднувати дані й створювати великомасштабні моделі, а також застосовувати отримані моделі у прийнятті рішень.

У 1990-х автоматичне ухвалення рішень до невпізнаного змінило банківські системи й споживчі кредити. Банки й телекомунікаційні компанії в 1990-х також застосовували системи великих масштабів, щоб мати змогу приймати рішення стосовно контролю за шахраями на основі даних. Роздрібні системи комп'ютеризувалися дедалі більше, рішення щодо мерчандайзингу вже були повністю автоматизовані. Відомі приклади — програми винагород у казино Harrah's і автоматичні рекомендації Amazon і Netflix. Зараз ми з вами спостерігаємо революцію в рекламі, оскільки тепер користувачі величезну частку часу проводять в інтернеті, і в онлайні рішення щодо реклами можна ухвалювати за (буквально) долю секунди.

ОБРОБКА ДАНИХ ТА «БІГ-ДАТА»

Зараз важливо відхилитися й обговорити ще один момент. В обробці даних є багато такого, що не стосується data science — хоча, якщо судити по медіа, то можна подумати інакше. Інженерія й обробка даних критично важливі для існування data science, але це більш загальні поняття. Наприклад, сьогодні багато вмінь, систем і технологій з обробки даних помилково називають data science. Але щоб розуміти data science і компанії, які ухвалюють рішення на основі даних, важливо розуміти цю відмінність. Для data science потрібен доступ до даних, і правильна інженерія може тільки піти на користь, але ці технології — не технології власне data science. Як показано на рис. 1.1, вони підтримують data science, але можуть бути корисні ще багато для чого. Технології обробки даних дуже важливі для багатьох задач у бізнесі, для яких потрібні дані, але де не потрібно вміти діставати з даних корисну інформацію або ухвалювати на їх основі рішення. Це, наприклад, ефективна обробка транзакції, підтримування роботи сучасної веб-системи і рекламна онлайн-кампанія.

Останнім часом досить багато уваги медіа привернули технології біг-дати (Hadoop, Hbase, і MongoDB). По суті, «біг-дата» означає, що для традиційних технологій обробки даних ці набори даних завеликі, і потрібні нові технології. Технології біг-дати використовують для багатьох задач, включно з інженерією даних — як і традиційні технології. Інколи технології біг-дати застосовують, щоб імплементувати техніки майнінгу даних. Однак набагато частіше так широко відомі технології біг-дати використо-