

УДК 004.8+336  
ББК 32.973.26-018.1+65.26  
Я65

## Янсен С.

Я65 Машинное обучение для алгоритмической торговли на финансовых рынках. Практикум: Пер. с англ. — СПб.: БХВ-Петербург, 2020. — 560 с.: ил.  
ISBN 978-5-9775-6595-0

Книга посвящена практике применения машинного обучения с целью создания мощных алгоритмических стратегий для успешной торговли на финансовых рынках. Изложены базовые принципы работы с данными: оценивание наборов данных, доступ к данным через API на языке Python, доступ к финансовым данным на платформе Quandl и управление ошибками предсказания. Рассмотрены построение и тренировка алгоритмических моделей с помощью Python-библиотек pandas, Seaborn, StatsModels и sklearn и построение, оценка и интерпретация моделей AR( $p$ ), MA( $q$ ) и ARIMA( $p, d, q$ ) с использованием библиотеки StatsModels. Описано применение библиотеки PyMC3 для байесового машинного обучения, библиотек NLTK, sklearn (Scikit-learn) и spaCy для назначения отметок финансовым новостям и классифицирования документов, библиотеки Keras для создания, настройки и оценки нейронных сетей прямого распространения, рекуррентных и сверточных сетей. Показано, как применять трансферное обучение к данным спутниковых снимков для предсказания экономической активности и как эффективно использовать подкрепляемое обучение для достижения оптимальных результатов торговли.

*Для финансовых аналитиков и программистов на языке Python*

УДК 004.8+336  
ББК 32.973.26-018.1+65.26

### Группа подготовки издания:

|                       |                          |
|-----------------------|--------------------------|
| Руководитель проекта  | <i>Евгений Рыбаков</i>   |
| Зав. редакцией        | <i>Екатерина Сависте</i> |
| Перевод с английского | <i>Андрея Логунова</i>   |
| Компьютерная верстка  | <i>Ольги Сергиенко</i>   |
| Оформление обложки    | <i>Карины Соловьевой</i> |

© Packt Publishing 2019. First published in the English language under the title 'Hands-On Machine Learning for Algorithmic Trading – (9781789346411)'

Впервые опубликовано на английском языке под названием 'Hands-On Machine Learning for Algorithmic Trading – (9781789346411)'

"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

ISBN 978-1-78934-641-1 (англ.)  
ISBN 978-5-9775-6595-0 (рус.)

© Packt Publishing 2019  
© Перевод на русский язык, оформление.  
ООО "БХВ-Петербург", ООО "БХВ", 2020

# Оглавление

|   |           |
|---|-----------|
| <b>Об авторе</b> .....  | <b>23</b> |
| <b>О рецензентах</b> .....  | <b>24</b> |
| <b>Комментарии переводчика</b> .....                                      | <b>25</b> |
| О терминологии .....  | 26        |
| <b>Предисловие</b> .....  | <b>28</b> |
| Для кого эта книга предназначена .....                                    | 29        |
| Что эта книга охватывает.....   | 29        |
| Получение максимальной отдачи от этой книги .....                         | 32        |
| Скачивание файлов с примерами исходного кода.....                         | 32        |
| Скачивание цветных изображений .....                                      | 33        |
| Принятые в книге условные обозначения .....                               | 33        |
| <b>Глава 1. Машинное обучение для торговли на финансовых рынках</b> ..... | <b>34</b> |
| Как читать эту книгу .....  | 35        |
| Чего ожидать .....  | 36        |
| Кто должен прочесть эту книгу.....  | 37        |
| Как эта книга организована .....  | 37        |
| Каркас — от данных к конструированию стратегии.....                       | 37        |
| Основы машинного обучения.....  | 38        |
| Обработка естественного языка .....                                       | 39        |
| Глубокое и подкрепляемое обучение .....                                   | 40        |
| Что вам нужно для успеха .....  | 40        |
| Источники данных.....   | 41        |
| Репозиторий GitHub .....  | 41        |
| Библиотеки Python.....  | 42        |
| Рост популярности машинного обучения в инвестиционной индустрии .....     | 42        |
| От торговли электронной к торговле высокочастотной .....                  | 43        |
| Факторное инвестирование и умные бета-фонды .....                         | 46        |
| Алгоритмические первопроходцы превосходят людей в широком масштабе .....  | 51        |
| Фонды, ведомые машинным обучением, привлекают 1 трлн долларов AUM.....    | 52        |

|  |            |
|--|------------|
| Возникновение квантоментальных фондов .....  | 53         |
| Инвестиции в стратегический потенциал.....   | 54         |
| Машинное обучение и альтернативные данные .....                                    | 54         |
| Привлечение "толпы" в разработку алгоритмов торговли.....                          | 56         |
| Конструирование и исполнение стратегии торговли .....                              | 57         |
| Привлечение источников данных и управление данными .....                           | 58         |
| Исследование и оценивание альфа-факторов .....                                     | 58         |
| Оптимизация портфеля и риск-менеджмент .....                                       | 60         |
| Бэктестирование стратегии.....   | 60         |
| Машинное обучение и стратегии алгоритмической торговли .....                       | 61         |
| Примеры использования машинного обучения для торговли<br>на финансовых рынках..... | 62         |
| Глубинный анализ данных для извлечения признаков .....                             | 62         |
| Контролируемое обучение для создания и агрегирования альфа-факторов .....          | 63         |
| Размещение финансовых средств среди портфельных активов .....                      | 64         |
| Тестирование торговых идей.....  | 64         |
| Подкрепляемое обучение.....  | 65         |
| Резюме .....   | 65         |
| <b>Глава 2. Рыночные и фундаментальные данные.....</b>                             | <b>66</b>  |
| Как работать с рыночными данными.....  | 67         |
| Микроструктура рынка .....   | 68         |
| Торговые площадки.....   | 68         |
| Типы ордеров .....   | 71         |
| Работа с данными ордерной книги.....   | 72         |
| Протокол FIX .....   | 72         |
| Данные ордерной книги TotalView-ITCH биржи Nasdaq.....                             | 73         |
| Регуляризация тиковых данных .....   | 83         |
| Доступ к рыночным данным через API .....   | 87         |
| Дистанционный доступ к данным с помощью библиотеки pandas .....                    | 87         |
| Платформа Quantopian .....   | 90         |
| Библиотека zipline.....  | 91         |
| Платформа Quandl .....   | 93         |
| Другие поставщики рыночных данных .....  | 93         |
| Как работать с фундаментальными данными .....                                      | 94         |
| Данные финансовой отчетности.....  | 95         |
| Автоматизированная обработка — XBRL.....   | 95         |
| Построение временного ряда с фундаментальными данными .....                        | 96         |
| Другие источники фундаментальных данных .....                                      | 101        |
| Макроэкономические и отраслевые данные с помощью pandas-datareader .....           | 101        |
| Эффективное хранение данных с помощью библиотеки pandas.....                       | 102        |
| Резюме .....   | 103        |
| <b>Глава 3. Альтернативные данные для финансов.....</b>                            | <b>104</b> |
| Революция альтернативных данных .....  | 105        |
| Источники альтернативных данных .....  | 107        |
| Физические лица.....   | 107        |
| Производственные процессы предприятия .....  | 108        |
| Датчики.....   | 109        |

|   |            |
|---|------------|
| Оценивание альтернативных наборов данных .....                                      | 110        |
| Критерии оценивания.....  | 111        |
| Качество сигнального содержимого .....  | 111        |
| Качество данных.....  | 113        |
| Технические аспекты.....  | 115        |
| Рынок альтернативных данных .....   | 115        |
| Поставщики данных и примеры использования .....                                     | 116        |
| Данные социальных настроений .....  | 117        |
| Спутниковые данные.....   | 118        |
| Геолокационные данные .....   | 119        |
| Данные электронных почтовых квитанций.....  | 119        |
| Работа с альтернативными данными .....  | 119        |
| Выскабливание данных службы OpenTable .....   | 119        |
| Извлечение данных из HTML с помощью библиотек requests и BeautifulSoup .....        | 120        |
| Знакомство с библиотекой Selenium — использование браузерной автоматизации .....    | 121        |
| Построение набора данных бронирований ресторанов .....                              | 122        |
| Следующий шаг — библиотеки Scrapy и splash .....                                    | 124        |
| Стенограммы телеконференций о корпоративных зарплатах .....                         | 125        |
| Разбор HTML с помощью регулярных выражений .....                                    | 126        |
| Резюме .....  | 128        |
| <b>Глава 4. Исследование альфа-факторов .....</b>                                   | <b>129</b> |
| Выработка альфа-факторов .....  | 130        |
| Важные категории факторов.....  | 131        |
| Импульсные и сентиментные факторы .....   | 132        |
| Стоимостные факторы .....   | 136        |
| Волатильностные и размерные факторы .....   | 140        |
| Качественные факторы .....  | 141        |
| Как преобразовывать данные в факторы.....   | 143        |
| Полезные методы библиотек pandas и NumPy.....                                       | 144        |
| Факторы, встроенные в платформу Quantopian .....                                    | 147        |
| Библиотека TA-Lib .....   | 147        |
| Поиски сигналов — как использовать библиотеку zipline .....                         | 148        |
| Архитектура — событийная симуляция торговли.....                                    | 148        |
| Единственный альфа-фактор из рыночных данных .....                                  | 149        |
| Сочетание факторов из разнообразных источников данных .....                         | 152        |
| Разделение сигнала и шума — как использовать библиотеку alphalens.....              | 154        |
| Создание форвардных финансовых возвратов и факторных квантилей .....                | 155        |
| Предсказательная результативность по факторным квантилям .....                      | 156        |
| Информационный коэффициент .....  | 159        |
| Оборачиваемость фактора .....   | 161        |
| Ресурсы альфа-факторов.....   | 162        |
| Альтернативные библиотеки алгоритмической торговли .....                            | 162        |
| Резюме .....  | 163        |
| <b>Глава 5. Оценивание стратегии.....</b>   | <b>164</b> |
| Как строить и тестировать инвестиционный портфель с помощью библиотеки zipline..... | 165        |
| Торговля по плану и перебалансировка портфеля.....                                  | 165        |

|   |            |
|---|------------|
| Как измерять результативность с помощью библиотеки <code>pyfolio</code> .....                     | 167        |
| Коэффициент Шарпа .....   | 167        |
| Фундаментальный закон активного менеджмента .....   | 169        |
| Внутривыборочная и вневыборочная результативность с помощью библиотеки <code>pyfolio</code> ..... | 170        |
| Получение входа в библиотеку <code>pyfolio</code> из библиотеки <code>alphalens</code> .....      | 170        |
| Получение входа в библиотеку <code>pyfolio</code> из бэктеста <code>zipline</code> .....          | 171        |
| Форвардное тестирование вневыборочных возвратов .....   | 171        |
| Сводная статистика результативности .....   | 172        |
| Периоды просадки и влияние факторов .....   | 173        |
| Моделирование событийного риска .....   | 174        |
| Как избегать ловушек бэктестирования .....  | 175        |
| Сложности данных .....  | 176        |
| Систематическое смещение из-за забегания вперед .....   | 176        |
| Систематическое смещение из-за выживших .....   | 176        |
| Управление выбросами .....  | 176        |
| Нерепрезентативный период .....   | 177        |
| Вопросы реализации .....  | 177        |
| Результативность на основе рыночной цены .....  | 177        |
| Торговые издержки .....   | 177        |
| Определение времени сделок .....  | 178        |
| Прочесывание данных и переподгонка при бэктестировании .....                                      | 178        |
| Минимальная длина бэктеста и дефлированный коэффициент Шарпа .....                                | 179        |
| Оптимальное прекращение бэктестирования .....   | 179        |
| Как управлять портфельным риском и возвратностью .....  | 180        |
| Среднедисперсная оптимизация .....  | 182        |
| Как это работает .....  | 182        |
| Эффективная граница на Python .....   | 182        |
| Сложности и недостатки .....  | 185        |
| Альтернативы среднедисперсной оптимизации .....   | 186        |
| Портфель $1/n$ .....  | 186        |
| Минимально-дисперсный портфель .....  | 186        |
| Глобальная портфельная оптимизация — подход Блэка — Литгермана .....                              | 187        |
| Как определять размер ставок — правило Келли .....  | 187        |
| Паритет риска .....   | 190        |
| Риск-факторное инвестирование .....   | 191        |
| Иерархический паритет риска .....   | 191        |
| Резюме .....  | 192        |
| <b>Глава 6. Процесс машинного обучения .....</b>  | <b>194</b> |
| Усвоение регулярностей из данных .....  | 195        |
| Контролируемое обучение .....   | 197        |
| Неконтролируемое обучение .....   | 198        |
| Применения .....  | 198        |
| Кластерные алгоритмы .....  | 198        |
| Снижение размерности .....  | 199        |
| Подкрепляемое обучение .....  | 200        |

|  |            |
|--|------------|
| Рабочий поток машинного обучения.....  | 201        |
| Простое пошаговое руководство — $k$ ближайших соседей.....                     | 201        |
| Очертить рамки задачи — цели и метрики.....                                    | 202        |
| Предсказание против статистического вывода.....                                | 202        |
| Регрессионные задачи.....  | 204        |
| Классификационные задачи.....  | 205        |
| Собрать и подготовить данные.....  | 209        |
| Разведать, извлечь и выработать признаки.....                                  | 209        |
| Использование теории информации для оценивания признаков.....                  | 210        |
| Отобрать автоматически обучающийся алгоритм.....                               | 211        |
| Сконструировать и настроить модели.....  | 211        |
| Компромисс между смещением и дисперсией.....                                   | 211        |
| Недоподгонка против переподгонки.....  | 212        |
| Управление компромиссом.....   | 213        |
| Кривые усвоения.....   | 214        |
| Как использовать перекрестный контроль для отбора модели.....                  | 215        |
| Как реализовать перекрестный контроль на языке Python.....                     | 215        |
| Перекрестный контроль.....   | 216        |
| Настройка параметров с помощью библиотеки <i>sklean</i> .....                  | 219        |
| Кривые перекрестного контроля с помощью библиотеки <i>yellowbricks</i> .....   | 220        |
| Кривые усвоения.....   | 220        |
| Настройка параметров с помощью интерфейса <i>GridSearchCV</i> и конвейера..... | 221        |
| Сложности перекрестного контроля в финансах.....                               | 221        |
| Перекрестный контроль временного ряда в библиотеке <i>sklearn</i> .....        | 222        |
| Прочистка, наложение эмбарго и комбинаторный перекрестный контроль.....        | 222        |
| Резюме.....  | 223        |
| <b>Глава 7. Линейные модели.....</b>   | <b>224</b> |
| Линейная регрессия для статистического вывода и предсказания.....              | 225        |
| Множественная линейная регрессионная модель.....                               | 226        |
| Как формулировать модель.....  | 227        |
| Как тренировать модель.....  | 228        |
| Наименьшие квадраты.....   | 228        |
| Оценивание максимального правдоподобия.....                                    | 229        |
| Градиентный спуск.....   | 230        |
| Теорема Гаусса — Маркова.....  | 230        |
| Как проводить статистический вывод.....  | 232        |
| Как диагностировать и устранять проблемы.....                                  | 233        |
| Качество подгонки.....   | 234        |
| Гетероскедастичность.....  | 235        |
| Внутрирядовая корреляция.....  | 236        |
| Мультиколлинеарность.....  | 236        |
| Как выполнять линейную регрессию на практике.....                              | 237        |
| Обычные наименьшие квадраты из библиотеки <i>StatsModels</i> .....             | 237        |
| Стохастический градиентный спуск с помощью библиотеки <i>sklearn</i> .....     | 239        |
| Как строить линейную факторную модель.....                                     | 240        |
| От модели <i>SARIM</i> к пятифакторной модели Фама — Френча.....               | 241        |
| Получение рисков факторов.....   | 242        |
| Регрессия Фама — Макбета.....  | 244        |

|  |            |
|--|------------|
| Усадочные методы: регуляризация для линейной регрессии.....                          | 248        |
| Как хеджироваться от переподгонки .....  | 248        |
| Как работает гребневая регрессия.....  | 249        |
| Как работает регрессия лассо .....   | 251        |
| Как применять линейную регрессию для предсказания финансовых возвратов.....          | 251        |
| Подготовка данных.....   | 251        |
| Создание универсума и временной горизонт.....  | 251        |
| Расчет целевого финансового возврата .....   | 252        |
| Отбор и преобразование альфа-факторов .....  | 253        |
| Очистка данных — пропущенные данные .....  | 253        |
| Разведывательный анализ данных .....   | 254        |
| Кодирование категориальных переменных с помощью фиктивных значений .....             | 254        |
| Создание форвардных возвратов .....  | 255        |
| Линейная регрессия методом OLS с использованием библиотеки StatsModels .....         | 256        |
| Диагностическая статистика.....  | 256        |
| Линейная регрессия по методу OLS с использованием библиотеки sklearn.....            | 257        |
| Кастомизированный перекрестный контроль временного ряда.....                         | 257        |
| Отбор признаков и цели.....  | 258        |
| Перекрестный контроль модели.....  | 258        |
| Тестовые результаты — информационный коэффициент и RMSE .....                        | 259        |
| Гребневая регрессия с использованием библиотеки sklearn.....                         | 261        |
| Настройка регуляризационных параметров с помощью перекрестного контроля .....        | 261        |
| Результаты перекрестного контроля и траектории гребневых коэффициентов.....          | 262        |
| Десятка ведущих коэффициентов .....  | 262        |
| Лассо-регрессия с использованием библиотеки sklearn.....                             | 263        |
| Перекрестно-контрольный информационный коэффициент и траектория лассо-регрессии..... | 264        |
| Линейная классификация.....  | 265        |
| Логистическая регрессионная модель .....   | 265        |
| Целевая функция .....  | 266        |
| Логистическая функция .....  | 266        |
| Оценивание максимального правдоподобия.....  | 267        |
| Как проводить статистический вывод с помощью библиотеки StatsModels.....             | 268        |
| Как применять логистическую регрессию для предсказания.....                          | 270        |
| Как предсказывать ценовые движения с помощью библиотеки sklearn.....                 | 270        |
| Резюме .....   | 272        |
| <b>Глава 8. Модели временных рядов .....</b>   | <b>273</b> |
| Аналитические инструменты для диагностики и извлечения признаков.....                | 274        |
| Как разложить временной ряд на регулярности .....                                    | 275        |
| Как вычислять статистику скользящего окна .....                                      | 276        |
| Скользящие средние и экспоненциальное сглаживание .....                              | 277        |
| Как измерять автокорреляцию .....  | 277        |
| Как диагностировать и достигать стационарности.....                                  | 278        |
| Преобразования временного ряда .....   | 279        |
| Как диагностировать и обращаться с единичными корнями.....                           | 280        |
| Единично-корневые тесты .....  | 282        |
| Как применять преобразования временных рядов .....                                   | 283        |

|  |            |
|--|------------|
| Модели одномерных временных рядов .....  | 285        |
| Как строить авторегрессионные модели .....   | 285        |
| Как выявлять число сдвигов .....   | 286        |
| Как диагностировать подгонку модели .....  | 286        |
| Как строить модели скользящего среднего .....  | 287        |
| Как выявлять число сдвигов .....   | 287        |
| Связь между моделями AR и MA .....   | 287        |
| Как строить модели ARIMA и их расширения .....   | 287        |
| Как выявлять число членов моделей AR и MA .....  | 288        |
| Добавление признаков — ARMAX .....   | 289        |
| Добавление сезонного исчисления последовательных разностей — SARIMAX .....                     | 289        |
| Как прогнозировать макроэкономические фундаментальные показатели .....                         | 290        |
| Как использовать модели временных рядов для прогнозирования волатильности .....                | 292        |
| Модель авторегрессионной условной гетероскедастичности (ARCH) .....                            | 293        |
| Обобщение модели ARCH — модель GARCH .....   | 294        |
| Как строить модель прогнозирования волатильности .....   | 294        |
| Модели многомерных временных рядов .....   | 298        |
| Система уравнений .....  | 298        |
| Векторная авторегрессионная модель (VAR) .....   | 299        |
| Как использовать модель VAR для прогнозов макроэкономических фундаментальных показателей ..... | 300        |
| Коинтеграция — временные ряды с общим трендом .....  | 303        |
| Тестирование на коинтеграцию .....   | 304        |
| Как использовать коинтеграцию для стратегии парной торговли .....                              | 305        |
| Резюме .....   | 305        |
| <b>Глава 9. Байесово машинное обучение .....</b>   | <b>307</b> |
| Как работает байесово машинное обучение .....  | 308        |
| Как обновлять допущения на основе эмпирического наблюдения .....                               | 309        |
| Точный вывод: оценивание апостериорного максимума .....  | 310        |
| Как отбирать априорные распределения .....   | 311        |
| Как не усложнять вывод — сопряженные априорные распределения .....                             | 312        |
| Как динамически оценивать вероятности движения цены актива .....                               | 312        |
| Приближенный вывод: стохастический и детерминированный подходы .....                           | 314        |
| Стохастический вывод на основе выборок .....   | 315        |
| Генерирование выборок методами Монте-Карло марковских цепей .....                              | 315        |
| Вариационный вывод .....   | 318        |
| Вероятностное программирование с помощью библиотеки PyMC3 .....                                | 319        |
| Байесово машинное обучение с помощью библиотеки Theano .....                                   | 319        |
| Рабочий поток библиотеки PyMC3 .....   | 320        |
| Определение модели — байесова логистическая регрессия .....                                    | 320        |
| Приближенный вывод — метод Монте-Карло марковской цепи .....                                   | 323        |
| Приближенный вывод — вариационный Байес .....  | 324        |
| Диагностика модели .....   | 324        |
| Предсказание .....   | 327        |
| Практические применения .....  | 327        |
| Байесов коэффициент Шарпа и сравнение результативности .....                                   | 328        |

|  |            |
|--|------------|
| Байесовы модели временных рядов .....  | 330        |
| Стохастические модели волатильности .....                                      | 330        |
| Резюме .....   | 330        |
| <b>Глава 10. Деревья решений и случайные леса .....</b>                        | <b>331</b> |
| Деревья решений .....  | 332        |
| Как деревья усваивают и применяют правила принятия решения .....               | 332        |
| Как использовать деревья решений на практике .....                             | 334        |
| Как готовить данные .....  | 334        |
| Как кодировать собственный класс перекрестного контроля .....                  | 335        |
| Как строить регрессионное дерево .....   | 336        |
| Как строить классификационное дерево .....                                     | 338        |
| Как визуализировать дерево решений .....                                       | 340        |
| Как оценивать предсказания дерева решений .....                                | 340        |
| Свойство важности признаков .....  | 342        |
| Переподгонка и регуляризация .....   | 342        |
| Как регуляризовать дерево решений .....  | 343        |
| Обрезка дерева решений .....   | 345        |
| Как настраивать гиперпараметры .....   | 345        |
| Класс <i>GridSearchCV</i> для деревьев решений .....                           | 346        |
| Как обследовать древесную структуру .....                                      | 347        |
| Кривые усвоения .....  | 348        |
| Сильные и слабые стороны деревьев решений .....                                | 349        |
| Случайные леса .....   | 350        |
| Ансамблевые модели .....   | 350        |
| Как бэггинг снижает дисперсию модели .....                                     | 351        |
| Бутстрап-агрегированные деревья решений .....                                  | 353        |
| Как строить случайный лес .....  | 355        |
| Как тренировать и настраивать случайный лес .....                              | 356        |
| Свойство важности признаков для случайных лесов .....                          | 358        |
| Внепакетное тестирование .....   | 359        |
| Сильные и слабые стороны случайных лесов .....                                 | 360        |
| Резюме .....   | 360        |
| <b>Глава 11. Градиентно-бустинговые машины .....</b>                           | <b>362</b> |
| Адаптивный бустинг .....   | 363        |
| Алгоритм AdaBoost .....  | 364        |
| AdaBoost в библиотеке <i>sklearn</i> .....                                     | 366        |
| Градиентно-бустинговые машины .....  | 368        |
| Как тренировать и настраивать модели на основе GBM .....                       | 370        |
| Размер ансамбля и досрочная остановка .....                                    | 371        |
| Усадка и темп усвоения .....   | 371        |
| Подвыборка и стохастический градиентный бустинг .....                          | 372        |
| Как использовать градиентный бустинг с помощью библиотеки <i>sklearn</i> ..... | 372        |
| Как настраивать параметры с помощью класса <i>GridSearchCV</i> .....           | 374        |
| Влияние параметров на тестовые отметки .....                                   | 375        |
| Как тестировать на отложенном наборе данных .....                              | 377        |

|  |            |
|--|------------|
| Быстро масштабируемые реализации градиентно-бустинговых машин..... | 377        |
| Как алгоритмические инновации стимулируют результативность.....    | 378        |
| Аппроксимация второпорядковой функции потерь.....                  | 378        |
| Упрощенные алгоритмы поиска разбинок.....                          | 379        |
| Поуровневый рост против полистового роста.....                     | 380        |
| Тренировка на основе GPU.....                                      | 381        |
| Отсеивающая регуляризация DART — отсев для деревьев.....           | 381        |
| Трактовка категориальных признаков.....                            | 382        |
| Дополнительные признаки и оптимизации.....                         | 382        |
| Как использовать библиотеки XGBoost, LightGBM и CatBoost.....      | 383        |
| Как создавать двоичные форматы данных.....                         | 383        |
| Как оценивать результаты.....                                      | 388        |
| Результаты перекрестного контроля между моделями.....              | 388        |
| Как интерпретировать результаты GBM.....                           | 391        |
| Свойство важности признаков.....                                   | 392        |
| Графики частичной зависимости.....                                 | 393        |
| Аддитивные объяснения Шепли (SHapley).....                         | 395        |
| Как резюмировать значения SHAP по признакам.....                   | 396        |
| Как использовать графики силы для объяснения предсказания.....     | 397        |
| Как анализировать взаимодействие признаков.....                    | 398        |
| Резюме.....  | 399        |
| <b>Глава 12. Неконтролируемое обучение.....</b>                    | <b>401</b> |
| Снижение размерности.....  | 402        |
| Линейные и нелинейные алгоритмы.....                               | 403        |
| Проклятие размерности.....   | 405        |
| Линейное снижение размерности.....                                 | 407        |
| Анализ главных компонент.....                                      | 407        |
| Анализ независимых компонент.....                                  | 413        |
| Анализ главных компонент для алгоритмической торговли.....         | 415        |
| Рисковые факторы, ведомые данными.....                             | 415        |
| Характеристические портфели.....                                   | 418        |
| Усвоение проекций в топологическом многообразии.....               | 421        |
| Алгоритм t-SNE.....  | 423        |
| Алгоритм UMAP.....   | 424        |
| Кластеризация.....   | 425        |
| Кластеризация на основе $k$ средних.....                           | 426        |
| Оценивание качества кластеров.....                                 | 427        |
| Иерархическая кластеризация.....                                   | 429        |
| Визуализация — дендрограммы.....                                   | 430        |
| Плотностная кластеризация.....                                     | 432        |
| Алгоритм DBSCAN.....   | 432        |
| Иерархический алгоритм DBSCAN.....                                 | 432        |
| Модели гауссовых смесей.....                                       | 433        |
| Алгоритм максимизации ожиданий.....                                | 433        |
| Иерархический паритет риска.....                                   | 434        |
| Резюме.....  | 436        |

|  |            |
|--|------------|
| <b>Глава 13. Работа с текстовыми данными</b> .....                                 | <b>438</b> |
| Как извлекать признаки из текстовых данных.....                                    | 439        |
| Сложности обработки естественного языка.....                                       | 439        |
| Рабочий поток обработки естественного языка.....                                   | 440        |
| Разбор и лексемизация текстовых данных.....  | 441        |
| Лингвистическая аннотация.....   | 442        |
| Семантическая аннотация.....   | 442        |
| Закрепление меток.....   | 442        |
| Примеры использования.....   | 443        |
| От текста к лексемам — конвейер обработки естественного языка.....                 | 443        |
| Конвейер по обработке естественного языка с помощью библиотек spaCy и textacy..... | 443        |
| Разбор, лексемизирование и аннотирование предложения.....                          | 444        |
| Пакетная обработка документов.....   | 446        |
| Определение границ предложений.....  | 446        |
| Распознавание именованных сущностей.....   | 447        |
| <i>N</i> -граммы.....  | 447        |
| Потоковый API библиотеки spaCy.....  | 448        |
| Многоязычная обработка естественного языка.....                                    | 448        |
| Обработка естественного языка с помощью библиотеки TextBlob.....                   | 450        |
| Выделение основ слов и лемматизация.....   | 450        |
| Полярность и субъективность настроений.....  | 451        |
| От лексем к числам — терм-документная матрица.....                                 | 451        |
| Модель мешка слов.....   | 452        |
| Измерение сходства документов.....   | 452        |
| Терм-документная матрица с помощью библиотеки sklearn.....                         | 453        |
| Применение класса <i>CountVectorizer</i> .....                                     | 454        |
| Классы <i>TfidfTransformer</i> и <i>TfidfVectorizer</i> .....                      | 456        |
| Предобработка текста — краткий обзор.....  | 458        |
| Классификация текста и sentimentный анализ.....                                    | 458        |
| Наивный байесов классификатор.....   | 459        |
| Памятка по теореме Байеса.....   | 459        |
| Допущение об условной независимости.....   | 460        |
| Классифицирование новостных статей.....  | 461        |
| Тренировка и оценивание мультиномиального наивного байесова классификатора.....    | 461        |
| Sentimentный анализ.....   | 462        |
| Данные социальной сети Twitter.....  | 462        |
| Отзывы о деятельности предприятий — конкурсные данные веб-сайта Yelp.....          | 463        |
| Резюме.....  | 466        |
| <b>Глава 14. Тематическое моделирование</b> .....                                  | <b>467</b> |
| Усвоение скрытых тем: цели и подходы.....  | 468        |
| От линейной алгебры к иерархическим вероятностным моделям.....                     | 469        |
| Латентно-семантическое индексирование.....   | 469        |
| Как реализовать LSI с помощью библиотеки sklearn.....                              | 471        |
| Сильные и слабые стороны.....  | 472        |

|  |            |
|--|------------|
| Вероятностный латентный семантический анализ.....  | 474        |
| Как реализовать rLSA с помощью библиотеки sklearn .....  | 475        |
| Латентное размещение Дирихле .....   | 476        |
| Как работает LDA.....  | 476        |
| Распределение Дирихле .....  | 477        |
| Генеративная модель.....   | 477        |
| Реконструирование процесса.....  | 478        |
| Как оценивать темы LDA.....  | 479        |
| Перплексивность .....  | 479        |
| Тематическая когерентность .....   | 479        |
| Как реализовать LDA с помощью библиотеки sklearn.....  | 480        |
| Как визуализировать результаты LDA с помощью библиотеки pyLDAvis.....                                    | 481        |
| Как реализовать LDA с помощью библиотеки gensim.....   | 482        |
| Тематическое моделирование применительно к телеконференциям о<br>корпоративных зарплатах .....           | 484        |
| Предобработка данных.....  | 485        |
| Тренировка и оценивание модели.....  | 485        |
| Проведение экспериментов .....   | 487        |
| Тематическое моделирование применительно к отзывам о деятельности<br>предприятий на веб-сайте Yelp ..... | 488        |
| Резюме .....   | 490        |
| <b>Глава 15. Векторное вложение слов.....</b>  | <b>491</b> |
| Как векторные вложения слов кодируют семантику .....   | 492        |
| Как нейронно-языковые модели усваивают словопотребление в контексте .....                                | 492        |
| Модель Word2vec — усвоение векторных вложений в широком масштабе.....                                    | 493        |
| Модельная цель — упрощенная активационная функция softmax .....  | 494        |
| Автоматическое обнаружение фраз .....  | 495        |
| Как оценивать векторные вложения — векторная арифметика и аналогии .....                                 | 496        |
| Как использовать предварительно натренированные словарные векторы .....                                  | 497        |
| GloVe — глобальные векторы для представления слов.....   | 498        |
| Как тренировать собственные векторные вложения слов .....  | 499        |
| Архитектура модели Skip-Gram в библиотеке Keras.....   | 499        |
| Оценивание контрастивное к шуму .....  | 499        |
| Компоненты модели.....   | 500        |
| Словарные векторы из финансовой отчетности SEC с использованием<br>библиотеки gensim.....                | 500        |
| Предобработка .....  | 500        |
| Автоматическое обнаружение фраз .....  | 501        |
| Тренировка модели.....   | 502        |
| Оценивание модели.....   | 502        |
| Влияние параметрических настроек на результативность .....   | 503        |
| Сентиментный анализ с помощью модели Doc2vec.....  | 504        |
| Тренировка модели Doc2vec на сентиментных данных Yelp.....   | 504        |
| Создание входных данных.....   | 505        |
| Бонус — модель Word2vec для машинного перевода .....   | 508        |
| Резюме .....   | 508        |

|  |            |
|--|------------|
| <b>Глава 16. Дальнейшие действия</b> .....   | <b>509</b> |
| Ключевые итоги и извлеченные уроки.....  | 510        |
| Данные — единственный самый важный ингредиент .....                                    | 510        |
| Контроль качества .....  | 510        |
| Интеграция данных .....  | 511        |
| Компетентное знание предметной области помогает разблокировать ценность данных .....   | 511        |
| Выработка признаков и исследование альфа-факторов .....                                | 512        |
| Машинное обучение — это комплект инструментов для решения задач с помощью данных ..... | 512        |
| Диагностика модели помогает ускорить оптимизацию .....                                 | 513        |
| Принятие решений без бесплатного обеда .....   | 513        |
| Управление компромиссом между смещением и дисперсией .....                             | 514        |
| Определение адресных модельных целей .....   | 514        |
| Верификационная проверка оптимизации .....   | 515        |
| Остерегайтесь переподгонки к историческим данным .....                                 | 515        |
| Как проникать в сутьность черно-ящичных моделей .....                                  | 515        |
| Машинное обучение для торговли на финансовых рынках на практике .....                  | 516        |
| Технологии управления данными .....  | 516        |
| Системы управления базами данных .....   | 517        |
| Технологии больших данных — Hadoop и Spark .....                                       | 517        |
| Инструменты машинного обучения.....  | 518        |
| Онлайновые торговые платформы .....  | 519        |
| Платформа Quantopian .....   | 519        |
| Платформа QuantConnect .....   | 519        |
| Платформа QuantRocket.....   | 520        |
| <b>Заключение</b> .....  | <b>521</b> |
| <b>Глоссарий</b> .....   | <b>522</b> |
| <b>Предметный указатель</b> .....  | <b>546</b> |

## ГЛАВЫ, ПОМЕЩЕННЫЕ В ЭЛЕКТРОННЫЙ АРХИВ

|  |          |
|--|----------|
| <b>Глава 17. Глубокое обучение</b> .....   | <b>1</b> |
| Глубокое обучение и искусственный интеллект .....  | 2        |
| Сложности высокоразмерных данных .....   | 3        |
| Глубокое обучение как усвоение представлений .....                                       | 4        |
| Как глубокое обучение извлекает иерархические признаки из данных .....                   | 5        |
| Аппроксимация универсальных функций.....   | 6        |
| Глубокое обучение и усвоение проекций в топологическом многообразии.....                 | 6        |
| Как глубокое обучение соотносится с машинным обучением и искусственным интеллектом ..... | 7        |
| Как конструировать нейронную сеть.....   | 8        |
| Как работают нейронные сети.....   | 9        |
| Простая архитектура сети прямого распространения .....                                   | 10       |

|   |           |
|---|-----------|
| Ключевые варианты выбора конструкции .....  | 11        |
| Стоимостные функции .....   | 12        |
| Выходные элементы .....   | 12        |
| Скрытые элементы .....  | 13        |
| Как регуляризовать глубокие нейронные сети .....  | 14        |
| Штрафы по норме параметров .....  | 14        |
| Досрочная остановка .....   | 15        |
| Отсев .....   | 15        |
| Оптимизация для глубокого обучения .....  | 15        |
| Стохастический градиентный спуск .....  | 16        |
| Импульс .....   | 17        |
| Адаптивные темпы усвоения .....   | 17        |
| Как строить нейронную сеть на языке Python .....  | 18        |
| Входной слой .....  | 18        |
| Скрытый слой .....  | 19        |
| Выходной слой .....   | 20        |
| Прямое распространение .....  | 21        |
| Перекрестно-энтропийная стоимостная функция .....   | 21        |
| Как тренировать нейронную сеть .....  | 22        |
| Как реализовать обратное распространение с помощью Python .....   | 22        |
| Как вычислять градиент .....  | 23        |
| Градиент функции потерь .....   | 23        |
| Градиенты выходного слоя .....  | 24        |
| Градиенты скрытого слоя .....   | 24        |
| Собирая все вместе .....  | 25        |
| Тренировка сети .....   | 26        |
| Как применять библиотеки глубокого обучения .....   | 28        |
| Как применять библиотеку Keras .....  | 28        |
| Как применять инструмент визуализации TensorBoard .....   | 30        |
| Как применять библиотеку PyTorch 1.0 .....  | 31        |
| Как создать загрузчик данных в библиотеке PyTorch .....   | 32        |
| Как формировать нейросетевую архитектуру .....  | 33        |
| Как тренировать модель .....  | 34        |
| Как оценивать модельные предсказания .....  | 35        |
| Как применять библиотеку TensorFlow 2.0 .....   | 35        |
| Как оптимизировать нейросетевые архитектуры .....   | 35        |
| Создание временного ряда финансовых возвратов от акции для предсказания<br>ценового движения актива ..... | 36        |
| Определение нейросетевой архитектуры с заполнителями .....  | 37        |
| Определение кастомизированной метрики потерь для досрочной остановки .....                                | 38        |
| Выполнение GridSearchCV для настройки нейросетевой архитектуры .....                                      | 38        |
| Как еще больше улучшить результаты .....  | 40        |
| Резюме .....  | 40        |
| <b>Глава 18. Сверточные нейронные сети .....</b>  | <b>41</b> |
| Как работают сети ConvNet .....   | 42        |
| Как работает сверточный слой .....  | 43        |
| Сверточный этап — обнаружение локальных признаков .....   | 45        |

|  |           |
|--|-----------|
| Детекторный этап — добавление нелинейности .....                                   | 47        |
| Сводный этап — понижающая дискретизация признакововой карты .....                  | 47        |
| Источники вдохновения из нейробиологии .....                                       | 48        |
| Опорные архитектуры сети ConvNet .....   | 49        |
| Сеть LeNet5 — первая современная CNN-сеть (1998) .....                             | 49        |
| Сеть AlexNet — закрепление CNN-сетей на технологической карте (2012) .....         | 50        |
| Сеть VGGNet — ориентация на фильтры меньших размеров .....                         | 51        |
| Сеть GoogLeNet — меньше параметров посредством модуля Inception .....              | 52        |
| Сеть ResNet — современное состояние дел .....                                      | 53        |
| Эталоны .....  | 53        |
| Извлеченные уроки .....  | 54        |
| Компьютерное зрение за пределами классификации — обнаружение и сегментация .....   | 54        |
| Как конструировать и тренировать CNN-сеть с помощью языка Python .....             | 55        |
| Сеть LeNet5 и набор данных MNIST с библиотекой Keras .....                         | 55        |
| Как готовить данные .....  | 55        |
| Как определить архитектуру .....   | 56        |
| Сеть AlexNet и набор данных CIFAR10 с библиотекой Keras .....                      | 58        |
| Как готовить данные с помощью аугментации снимка .....                             | 58        |
| Как формировать архитектуру модели .....   | 59        |
| Как использовать CNN-сеть с данными временного ряда .....                          | 60        |
| Трансферное обучение — более быстрая тренировка с меньшим количеством данных ..... | 62        |
| Как надстраивать над предварительно натренированной CNN-сети .....                 | 62        |
| Как извлекать бутылочные признаки .....  | 63        |
| Как продолжить дальнейшую тренировку предварительно натренированной модели .....   | 64        |
| Как обнаруживать объекты .....   | 66        |
| Набор данных номеров домов Google Street View .....                                | 66        |
| Как определять CNN-сеть с многочисленными выходами .....                           | 67        |
| Новейшие разработки .....  | 68        |
| Быстрое обнаружение объектов на спутниковых снимках .....                          | 68        |
| Как капсульные сети улавливают позу .....  | 68        |
| Резюме .....   | 68        |
| <b>Глава 19. Рекуррентные нейронные сети .....</b>                                 | <b>70</b> |
| Как работают RNN-сети .....  | 71        |
| Развертывание вычислительного графа с циклами .....                                | 73        |
| Обратное распространение во времени .....  | 74        |
| Альтернативные архитектуры RRN-сетей .....   | 74        |
| Выходная рекуррентия и вмешательство учителя .....                                 | 74        |
| Двунаправленные RNN-сети .....   | 75        |
| Кодировочно-декодировочные архитектуры и механизм внимания .....                   | 75        |
| Как конструировать глубокие RNN-сети .....   | 76        |
| Сложности усвоения долгосрочных зависимостей .....                                 | 77        |
| Элементы с долгой краткосрочной памятью .....                                      | 77        |
| Вентильные рекуррентные элементы .....   | 79        |
| Как строить и тренировать RNN-сети с помощью языка Python .....                    | 79        |
| Регрессия одномерного временного ряда .....  | 80        |
| Как привести данные временного ряда в форму для RNN-сети .....                     | 80        |
| Как определять двухслойную RNN-сеть, используя единственный слой LSTM .....        | 82        |

|   |           |
|---|-----------|
| Уложенные друг на друга элементы LSTM для классификации временного ряда .....     | 84        |
| Как готовить данные .....   | 84        |
| Как формировать архитектуру .....   | 86        |
| Регрессия многомерного временного ряда .....                                      | 88        |
| Загрузка данных .....   | 88        |
| Подготовка данных .....   | 88        |
| Определение и тренировка модели .....   | 89        |
| Элементы LSTM и векторные вложения слов для классификации сентимента .....        | 91        |
| Загрузка данных IMDB с отзывами о кинофильмах .....                               | 91        |
| Определение архитектур вложения и RNN-сети .....                                  | 92        |
| Сентиментный анализ с предварительно натренированными словарными векторами .....  | 93        |
| Предобработка текстовых данных .....  | 94        |
| Загрузка предварительно натренированных вложений GloVe .....                      | 94        |
| Резюме .....  | 95        |
| <b>Глава 20. Автокодировщики и генеративные состязательные сети .....</b>         | <b>97</b> |
| Как работают автокодировщики .....  | 98        |
| Нелинейное снижение размерности .....   | 99        |
| Сверточные автокодировщики .....  | 100       |
| Ограничения по разреженности с регуляризованными автокодировщиками .....          | 100       |
| Исправление поврежденных данных с помощью шумоподавляющих автокодировщиков .....  | 101       |
| Автокодировщики с отображением одной последовательности в другую .....            | 101       |
| Вариационные автокодировщики .....  | 102       |
| Конструирование и тренировка автокодировщиков с использованием языка Python ..... | 102       |
| Подготовка данных .....   | 103       |
| Однослойный автокодировщик прямого распространения .....                          | 104       |
| Определение кодировщика .....   | 105       |
| Определение декодировщика .....   | 105       |
| Тренировка модели .....   | 106       |
| Оценивание результатов .....  | 106       |
| Автокодировщик прямого распространения с ограничениями по разреженности .....     | 107       |
| Автокодировщик глубокого прямого распространения .....                            | 107       |
| Визуализация кодирования .....  | 108       |
| Сверточные автокодировщики .....  | 109       |
| Шумоподавляющие автокодировщики .....   | 110       |
| Как работают GAN-сети .....   | 111       |
| В чем различие между генеративными и дискриминативными моделями .....             | 111       |
| Как работает состязательная тренировка .....                                      | 112       |
| Как эволюционируют архитектуры GAN-сетей .....                                    | 113       |
| Глубокая сверточная GAN-сеть (DCGAN) .....  | 113       |
| Условные GAN-сети .....   | 114       |
| Успешные и быстро развивающиеся приложения GAN-сетей .....                        | 114       |
| Сеть CycleGAN — неспаренная трансляция из изображения в изображение .....         | 114       |
| Сеть StackGAN — синтез изображений "текст в фото" .....                           | 114       |
| Сверхразрешающая способность фотореалистичных изображений .....                   | 115       |
| Синтетические временные ряды с помощью рекуррентных cGAN-сетей .....              | 115       |

|  |            |
|--|------------|
| Построение GAN-сетей с использованием языка Python.....                                      | 115        |
| Определение дискриминаторной сети .....  | 116        |
| Определение генераторной сети .....  | 116        |
| Совмещение обеих сетей для определения GAN-сети .....  | 117        |
| Состязательная тренировка.....   | 117        |
| Оценивание результатов .....   | 118        |
| Резюме .....   | 118        |
| <b>Глава 21. Подкрепляемое обучение .....</b>  | <b>120</b> |
| Ключевые элементы подкрепляемого обучения.....   | 121        |
| Компоненты интерактивной системы подкрепляемого обучения .....                               | 122        |
| Политика — от состояний к действиям .....  | 122        |
| Вознаграждения — самообучение на основе действий .....                                       | 123        |
| Ценностная функция — хорошие решения в долгосрочной перспективе .....                        | 123        |
| Модельные агенты против безмодельных.....  | 124        |
| Как решать задачи подкрепляемого обучения.....   | 124        |
| Ключевые сложности в решении задач подкрепляемого обучения .....                             | 125        |
| Учет заслуг .....  | 125        |
| Разведывание против эксплуатации.....  | 125        |
| Фундаментальные подходы к решению задач подкрепляемого обучения .....                        | 125        |
| Динамическое программирование — цикл по ценностным значениям и цикл по политике .....        | 127        |
| Конечные задачи марковского процесса принятия решений .....                                  | 127        |
| Последовательности состояний, действий и вознаграждений .....                                | 127        |
| Ценностные функции — как оценивать долгосрочное вознаграждение .....                         | 128        |
| Уравнение Беллмана.....  | 129        |
| От ценностной функции к оптимальной политике.....  | 129        |
| Цикл по политике .....   | 130        |
| Цикл по ценностным значениям .....   | 131        |
| Обобщенный цикл по политике .....  | 132        |
| Динамическое программирование на языке Python .....  | 132        |
| Настройка решетчатого мира GridWorld.....  | 133        |
| Вычисление матрицы переходов.....  | 135        |
| Цикл по ценностным значениям .....   | 136        |
| Цикл по политике .....   | 137        |
| Решение задач марковского процесса принятия решений с помощью библиотеки rumpdttoolbox ..... | 137        |
| Выводы .....   | 138        |
| <i>Q</i> -обучение.....  | 138        |
| Компромисс между разведыванием и эксплуатацией — $\epsilon$ -жадная политика .....           | 139        |
| Алгоритм <i>Q</i> -обучения .....  | 139        |
| Тренировка агента <i>Q</i> -обучения с помощью языка Python .....                            | 139        |
| Глубокое подкрепляемое обучение .....  | 140        |
| Аппроксимация ценностной функции с помощью нейронной сети .....                              | 141        |
| Алгоритм глубокого <i>Q</i> -обучения и его расширения .....                                 | 141        |
| Воспроизведение опыта .....  | 142        |
| Медленно изменяющаяся целевая сеть.....  | 142        |
| Двойное глубокое <i>Q</i> -обучение .....  | 143        |

|  |            |
|--|------------|
| Платформа OpenAI Gym — лунный посадочный модуль .....                      | 143        |
| Сеть DDQN с использованием библиотеки TensorFlow .....                     | 144        |
| Архитектура сети DQN .....   | 144        |
| Настройка среды платформы OpenAI.....                                      | 145        |
| Гиперпараметры .....   | 145        |
| Вычислительный граф сети DDQN.....   | 146        |
| Результативность .....   | 148        |
| Подкрепляемое обучение для торговли на финансовых рынках .....             | 148        |
| Конструирование торговой среды в платформе OpenAI.....                     | 148        |
| Элементарная торговая игра.....  | 149        |
| Результативность глубокого $Q$ -обучения на фондовом рынке.....            | 150        |
| Резюме .....   | 151        |
| <b>Приложение 1. Настройка среды программирования на языке Python.....</b> | <b>152</b> |
| Настройка дистрибутива Anaconda.....                                       | 152        |
| Создание среды.....  | 153        |
| Установка программных библиотек .....                                      | 155        |
| Установка библиотеки TensorFlow .....                                      | 158        |
| Блокноты Jupyter.....  | 159        |
| <b>Приложение 2. Скачивание и подготовка данных .....</b>                  | <b>162</b> |
| Цены Quandl Wiki.....  | 162        |
| Метаданные цен Wiki.....   | 163        |
| Цены фондового индекса S&P 500 .....                                       | 163        |
| Компонентные ценные бумаги фондового индекса S&P 500.....                  | 164        |
| Метаданные по торгуемым компаниям США .....                                | 164        |
| Индексы вексельных цен .....   | 165        |



# Об авторе

**Стефан Янсен** (Stefan Jansen), дипломированный финансовый аналитик, является основателем и ведущим исследователем данных в компании Applied AI (<http://www.applied-ai.solutions/>), где он консультирует компании и стартапы списка Fortune 500 из разных отраслей промышленности по переводу деловых целей в стратегию данных и искусственного интеллекта, комплектует команды исследователей данных и разрабатывает решения на основе автоматизированного (машинного) обучения. До своей нынешней должности он был управляющим партнером и ведущим исследователем данных в международной инвестиционной фирме, в которой организовал и внедрил применение прогнозной аналитики и инвестиционных исследований в практику.

Он также был руководителем глобального финансово-технологического стартапа, работающего на 15 рынках, работал во Всемирном банке, консультировал центральные банки на развивающихся рынках и вел проекты на шести языках на четырех континентах. Стефан имеет степень магистра Гарвардского и Берлинского университетов и преподает науку о данных в частной школе General Assembly (<https://generalassemb.ly/>) и на платформе онлайн-обучения Datacamp (<https://www.datacamp.com/>).

"Благодарю Packt за эту возможность и команду, которая воплотила ее в конечный результат, в особенности хочу поблагодарить Снегил Колте (Snehal Kolte) за оказанное содействие в процессе редактирования. Многие мои сторонники заслуживают упоминания, но профессор Цекхаузер (Zeckhauser) из Гарварда выделяется вдохновляющим интересом к творческому использованию количественных методов для решения задач. Я в долгу перед родителями за то, что они поощряли мое любопытство и поддерживали меня. Но больше всего я благодарен Мариане за то, что она делает все это стоящим того".

*Стефан Янсен*

# О рецензентах

**Дуг Ортис** (Doug Ortiz) является опытным архитектором корпоративного облака, больших данных, анализа данных и решений, который проектировал, конструировал, разрабатывал, реконструировал и интегрировал корпоративные решения. Его знания также охватывают облачные платформы Amazon Web Services, Azure, Google Cloud, деловую аналитику, базы данных Hadoop, Spark, NoSQL и SharePoint. Он является основателем проекта астрофизического симулирования Illustris.

"Огромное спасибо моей замечательной жене Милле, а также Марии, Николаю и нашим детям за их поддержку".

*Дуг Ортис*

**Сандипан Дей** (Sandipan Dey) является исследователем данных с широким спектром интересов, включая такие темы, как машинное обучение, глубокое обучение, обработка изображений и компьютерное зрение. Он работал во многих областях науки о данных, включая рекомендательные системы, предсказательные модели для индустрии событий, модели датчиковой локализации, сентиментный анализ и машинную прогностику. Он получил степень магистра в области компьютерных наук в Университете Мэриленда, округ Балтимор, его статьи были опубликованы в бюллетенях нескольких конференций и в журналах IEEE по глубинному анализу данных.

Он имеет сертификаты более 100 массовых открытых курсов дистанционного обучения (МООС) по науке о данных, машинному обучению, глубокому обучению, обработке изображений и родственным курсам/специализациям. Регулярно обновляет свой блог (sandipanweb) и является энтузиастом машинного обучения.

# Комментарии переводчика

В эпоху всеобщей цифровизации и сетевого взаимодействия сфера биржевой торговли становится все более демократичной, позволяя любому попробовать в ней свои силы, используя многочисленные открытые источники данных и торговые платформы. Эта книга является одной из первых на русском языке, полностью посвященной алгоритмической торговле на финансовых рынках с использованием технологии машинного обучения, и она безусловно вызовет интерес как у профессионалов в области биржевой торговли, так и у специалистов по машинному обучению и, возможно, станет настольной у каждого кванта и разработчика автоматически обучающихся систем.

Книга написана высококлассным профессионалом и для профессионалов. Автор подробно разбирает все магистральные и новейшие технические решения машинного обучения применительно к торговле на финансовых рынках, предоставляя обширный справочный материал и примеры использования в своем репозитории на GitHub. Постоянно обновляемая кодовая база книги изначально была создана с использованием Ubuntu 18.04. Ее репозиторий содержит файлы операционных сред Ubuntu и Mac. Дальнейшие обновления репозитория будут содержать адаптированные версии для Windows. При переводе книги исходный код был частично протестирован в среде Windows 10. При тестировании исходного кода за основу взята Python версии 3.7.2. Время перевода — апрель-май 2019 г.

С целью расширить аудиторию книги за счет разработчиков обучающихся систем и, наоборот, облегчить биржевым торговцам работу с технологией машинного обучения настоящий перевод снабжен сносками в виде комментариев и определений терминов из финансов, машинного обучения, теории вероятностей и статистики, которые облегчат понимание материала и помогут "зацепиться" за тему. Все терминологические сноски в конце книги сведены в глоссарий основных терминов. Почти для всех терминов приведены ссылки на источник. Перевод книги содержит два приложения, посвященные настройке среды программирования на языке Python и подготовке данных, используемых на протяжении всей книги (см. электронный архив к книге по адресу <ftp://ftp.bhv.ru/9785977565950.zip> или на странице книги на сайте издательства [www.bhv.ru](http://www.bhv.ru)).

## О терминологии

**Финансовый возврат и возвратность.** Одним из ключевых понятий биржевой торговли является *доходность*, или, если точнее, *финансовый возврат* (return) на инвестированный капитал. С точки зрения биржевого торговца (трейдера), если представить ценовую информацию в виде барного временного графика, то финансовый возврат — это не что иное, как разница между ценой актива в момент времени  $t-1$  и ценой в момент времени  $t$  ( $\text{return}[t] - \text{return}[t-1]$ ). Если торговец сделал ставку в момент времени  $t-1$ , то он получает положительный, нулевой или отрицательный финансовый возврат в момент времени  $t$ . С точки зрения портфельного менеджера, это возврат на портфельные инвестиции в течение любого периода оценивания, включающего изменение рыночной стоимости портфеля. Процентное соотношение этого показателя называется *возвратностью* (rate of return,  $(\text{return}[t] - \text{return}[t-1]) / \text{return}[t-1]$ ). Цель машинного обучения в финансах состоит в том, чтобы натренировать автоматически обучающуюся систему предсказывать состояние рынка и цену актива/стоимость портфеля в момент времени  $t+1$  с целью достижения максимальной возвратности.

**Результативность.** В англоязычной литературе термин "результативность" (performance) широко применяется как в финансах, так и в машинном обучении, в особенности в финансах, где самое главное — это финансовый результат, получаемый от актива или инвестиционного портфеля. Данный термин ассоциируется не с созданием чего-либо за единицу времени (производительность, productivity) или эффекта (эффективность, effectiveness), а с созданием результата, который затем сравнивается с неким эталоном. Например, результативность актива (asset performance) означает способность предприятия брать операционные ресурсы, управлять ими и в результате получать прибыльную возвратность, превышающую индекс или кредитную ставку. Результативность в МО означает некое достижение по исполнению задачи, измеряемое общепринятыми стандартными мерами точности, полноты, стоимости и скорости. Хотя этот термин синонимичен термину "эффективность", однако эффективность имеет собственное слово и не раз встречается в этой книге отдельно, а термин "результативность" контрастнее подчеркивает нацеленность на результат и поэтому предпочтителен.

**Терминология машинного обучения.** В центре внимания машинного обучения и его подобласти, глубокого обучения, находится автоматически обучающаяся система, т. е. система, способная с течением времени приобретать новые знания и улучшать свою работу, используя поступающую информацию. В зарубежной специализированной литературе для *передачи* знаний ученику и *получения* знаний учеником существуют отдельные термины — train (*натренировать*) и learn (*выучить*, усвоить), где тренировка — это работа, которую выполняет исследователь-проектировщик для получения обучившейся модели, в основе которой лежит обучающийся алгоритм, по сути искатель минимумов (или максимумов) для надлежащим образом сформулированных функций, а усвоение, или самообучение, — это работа, которую выполняет алгоритм-ученик по заучиванию связей и регулярно-

стей в данных или изменению и закреплению своего поведения. Когда же для обоих английских терминов в русской спецлитературе используется один-единственный термин "обучение", то он несет в себе двусмысленность, потому что под ним может подразумеваться и передача знаний, и получение знаний одновременно, как, например, в случае с термином "машинное обучение", который может означать и тренировку алгоритмических машин, и способность таких машин автоматически обучаться, что нередко вносит путаницу и терминологический разброд в переводной литературе при решении дилеммы "training-learning" в то время, как появление в зарубежной технической литературе термина learning в любом виде подразумевает исключительно второе — *самообучение, заучивание, усвоение* алгоритмом регулярностей или параметров. Отсюда вытекает одно важное следствие: английский термин machine learning обозначает *усвоение знаний алгоритмической машиной*, а, следовательно, более соответствовать оригиналу будет термин "*машинное самообучение*" или "*автоматическое обучение*". Весомым аргументом в пользу этих вариантов термина является и то, что с начала 1960-х и до середины 1980-х годов в ходу был похожий термин — "обучающиеся машины" (см. работы А. Тьюринга, К. Шеннона, Н. Винера, Н. Нильсона, Я. З. Цыпкина и др.). В настоящем переводе, следуя принципам здравого смысла и бритвы Оккама, за основу принят зарубежный подход.

<https://ru.tradingview.com/u/capissimo/#published-scripts>

# Предисловие

Наличие и доступность разнообразных данных повысила спрос на компетентные знания в области стратегий алгоритмической торговли. Благодаря этой книге вы выберете машинное обучение (МО), будете его применять к широкому спектру источников данных и создавать мощные алгоритмические стратегии.

Книга начинается с введения существенных элементов, таких как оценивание наборов данных, доступ к данным через API с помощью Python, использование платформы Quandl для доступа к финансовым данным и управление ошибками предсказания. Далее мы рассмотрим различные технические решения машинного обучения и автоматически обучающиеся алгоритмы, которые могут использоваться для построения и тренировки алгоритмических моделей с помощью программных Python-библиотек pandas, Seaborn, StatsModels и sklearn. Потом мы построим, оценим и дадим интерпретацию моделей  $AR(p)$ ,  $MA(q)$  и  $ARIMA(p, d, q)$  с использованием библиотеки StatsModels. Вы примените байесовы понятия "априорное распределение", "наблюдение" и "апостериорное распределение" для того, чтобы различать понятие неопределенности с помощью библиотеки PyMC3. Затем мы задействуем библиотеки NLTK, sklearn (Scikit-learn) и spaCy для назначения sentimentных отметок финансовым новостям и классифицирования документов для извлечения торговых сигналов. Мы научимся конструировать, строить, настраивать и оценивать нейронные сети прямого распространения, рекуррентные нейронные сети (RNN-сети) и сверточные нейронные сети (CNN-сети), используя библиотеку Keras для разработки изощренных алгоритмов. Вы примените трансферное обучение к данным спутниковых снимков для предсказания экономической активности. Наконец, мы применим подкрепляемое обучение для достижения оптимальных результатов торговли.

Прочитав книгу до конца, вы будете способны применять алгоритмическую торговлю, реализуя умные инвестиционные стратегии.

## Для кого эта книга предназначена

Книга предназначена для аналитиков и исследователей данных и разработчиков на языке Python, а также инвестиционных аналитиков и портфельных менеджеров, работающих внутри финансово-инвестиционной индустрии. Если вы хотите реализовать эффективную алгоритмическую торговлю, разрабатывая интеллектуальные разведывающие стратегии с использованием автоматически обучающихся алгоритмов, то настоящая книга — именно то, что вам нужно! Некоторое понимание приемов языка Python и технических решений МО является обязательным.

## Что эта книга охватывает

*Глава 1 "Машинное обучение для торговли на финансовых рынках"* отождествляет центральную тему данной книги, обрисовывая роль машинного обучения в формировании и оценивании сигналов для разработки и исполнения торговой стратегии. В ней описывается стратегический процесс от генерирования и моделирования гипотез, отбора данных и бэктестирования до оценивания и исполнения стратегий в портфельном контексте, включая риск-менеджмент.

*Глава 2 "Рыночные и фундаментальные данные"* охватывает источники данных и работу с исходными тиковыми данными, поступающими с бирж, и данными финансовой отчетности, а также способы доступа к многочисленным поставщикам общедоступных данных, на которые мы будем опираться на протяжении всего повествования в этой книге.

*Глава 3 "Альтернативные данные для финансов"* предоставляет категории и критерии для оценивания качества стремительно растущего числа источников и поставщиков. В ней также демонстрируются способы создания альтернативных наборов данных путем выскабливания веб-сайтов, например для сбора стенограмм телеконференций о корпоративных заработках для использования в алгоритмах естественно-языковой обработки и сентиментного анализа во второй условной части книги.

*Глава 4 "Исследование альфа-факторов"* обеспечивает каркас для понимания того, как работают факторы и как измерять их эффективность, например, используя информационный коэффициент (information coefficient, IC). В ней демонстрируется способ конструирования альфа-факторов из данных с помощью библиотек Python в режиме офлайн и на веб-платформе Quantopian. В ней также вводятся библиотека zipline для бэктестирования факторов и библиотека alphalens для оценивания их предсказательной мощности.

*Глава 5 "Оценивание стратегии"* знакомит с тем, как строить, тестировать и оценивать стратегии торговли, используя исторические данные с помощью библиотеки zipline в режиме офлайн и на веб-платформе Quantopian. Она представляет и демонстрирует способы вычисления результативности портфеля и метрических показателей риска с помощью библиотеки pyfolio. В ней также рассматриваются спосо-

бы управления методологическими сложностями бэктестирования стратегий и вводятся методы оптимизации стратегии с точки зрения портфельного риска.

*Глава 6 "Процесс машинного обучения"* закладывает фундамент, давая описания того, как формулировать, тренировать и настраивать автоматически обучающиеся модели и оценивать их предсказательную результативность как систематический рабочий процесс.

*Глава 7 "Линейные модели"* показывает, как использовать линейную и логистическую регрессию для выведения заключения и предсказания и применять регуляризацию для управления риском переподгонки. Она знакомит с торговой платформой Quantopian и демонстрирует способы построения факторных моделей и предсказания цен на финансовые активы.

*Глава 8 "Модели временных рядов"* посвящена одномерным и многомерным временным рядам, включая векторные авторегрессионные модели и коинтеграционные тесты, а также способы их применения к стратегиям парной торговли.

*Глава 9 "Байесово машинное обучение"* знакомит с тем, как формулировать вероятностные модели и как извлечение выборок из вероятностных распределений методами Монте-Карло марковских цепей (Markov Chain Monte Carlo, MCMC) и вариационный Байес обеспечивают аппроксимирование вывода. В ней также иллюстрируются способы применения библиотеки PyMC3 для вероятностного программирования с целью понимания сущности параметрической и модельной неопределенности.

*Глава 10 "Деревья решений и случайные леса"* показывает, как строить, тренировать и настраивать нелинейные древесные модели для проникновения в сущность данных и предсказания. В ней представлены древесные ансамблевые модели и иллюстрируется то, как случайные леса используют агрегирование бутстраповских выборок для преодоления некоторых недостатков деревьев решений.

*Глава 11 "Градиентно-бустинговые машины"* демонстрирует использование библиотек XGBoost, LightGBM и CatBoost для выполнения высокорезультативной тренировки моделей и предсказания. Кроме того, в ней подробно анализируется настройка многочисленных гиперпараметров.

*Глава 12 "Неконтролируемое обучение"* посвящена применению задач снижения размерности и кластеризации в алгоритмической торговле. В ней используется анализ главных и независимых компонент для извлечения ведомых данными рисков факторов. В ней также представлено несколько технических решений для кластеризации и показано использование иерархической кластеризации для размещения финансовых средств среди портфельных активов.

*Глава 13 "Работа с текстовыми данными"* демонстрирует способы конвертирования текстовых данных в числовой формат. В ней также показывается применение классификационных алгоритмов из второй условной части, связанных с сентиментным анализом, к крупным наборам данных.

*Глава 14 "Тематическое моделирование"* посвящена применению байесова неконтролируемого обучения с целью извлечения латентных тем, которые способны ре-

зюмировать большое число документов и предлагать более эффективные способы разведывания текстовых данных либо использования тем в качестве признаков в классификационной модели. В ней демонстрируется способ применения этого технического решения к стенограммам телеконференций о корпоративных заработках, источники которых привлечены в *главе 3*, и к годовым отчетам, предъявляемым в Комиссию по ценным бумагам и биржам (Securities and Exchange Commission, SEC).

*Глава 15 "Векторное вложение слов"* посвящена использованию нейронных сетей для усвоения передовых языковых признаков в форме словарных векторов, которые улавливают семантический контекст намного лучше, чем традиционные текстовые признаки, и представляют собой очень перспективное направление для извлечения торговых сигналов из текстовых данных.

*Глава 16 "Последующие шаги"* представляет собой резюме всех предыдущих глав.

*Глава 17 "Глубокое обучение"* знакомит с библиотеками Keras, TensorFlow и PyTorch, наиболее популярными каркасами глубокого обучения, которые мы будем использовать на протяжении всей четвертой условной части. В ней также представлены технические решения для тренировки и настройки, включая регуляризацию, и дан краткий обзор общепринятых архитектур.

*Глава 18 "Сверточные нейронные сети"* охватывает CNN-сети, очень мощные сетевые архитектуры, предназначенные для классификационных задач с широкомасштабными неструктурированными данными. Мы введем успешные архитектурные проекты, натренируем CNN-сеть на спутниковых данных, например для предсказания экономической активности, и применим трансферное обучение для ускорения тренировочного процесса.

*Глава 19 "Рекуррентные нейронные сети"* показывает, как RNN-сети могут быть полезными для моделирования отображений последовательности в последовательность, в том числе для временных рядов. В ней демонстрируется, как RNN-сеть улавливает нелинейные регулярности за более продолжительные периоды.

*Глава 20 "Автокодировщики и генеративные состязательные сети"* обращается к вопросам неконтролируемого глубокого обучения, включая автокодировщики для нелинейного сжатия многомерных данных и генеративные состязательные сети (Generative Adversarial Networks, GAN) — одно из самых важных последних нововведений, которое применяется для генерирования синтетических данных.

*Глава 21 "Подкрепляемое обучение"* знакомит с автоматическим обучением на основе максимизации вознаграждения за принимаемое решение, позволяющее конструировать и тренировать агентов, которые учатся оптимизировать решения с течением времени в ответ на окружающую их среду. Вы увидите процесс построения агента, который откликается на рыночные сигналы, с использованием библиотеки Open AI Gym (<https://gym.openai.com>), так называемого "тренажерного зала искусственного интеллекта".



Главы 17–21 (а также приложения и цветные иллюстрации к главам 1–15) представлены в электронном архиве данной книги, размещенном на сайте издательства. Электронный архив можно скачать по ссылке <ftp://ftp.bhv.ru/9785977565950.zip> и со страницы книги на веб-сайте издательства по адресу [www.bhv.ru](http://www.bhv.ru).

## Получение максимальной отдачи от этой книги

Для чтения данной книги требуется только базовое знание языка Python и элементарное понимание технических решений машинного обучения.

## Скачивание файлов с примерами исходного кода<sup>1</sup>

Файлы с примерами можно скачать с вашего аккаунта по адресу <http://www.packtpub.com/> для всех книг издательства Packt Publishing, которые вы приобрели. Если вы купили эту книгу в другом месте, то можно посетить <http://www.packtpub.com/support> и зарегистрироваться там, чтобы получить файлы прямо по электронной почте.

Скачать файлы с примерами можно, выполнив следующие шаги:

1. Войдите на наш веб-сайт или зарегистрируйтесь там, используя ваш адрес электронной почты и пароль.
2. Наведите указатель мыши на вкладку **SUPPORT** вверху страницы.
3. Щелкните по разделу **Code Downloads & Errata**, посвященному примерам программного кода и опечаткам.
4. Введите название книги в поле поиска.

Скачав файл, пожалуйста, убедитесь, что вы распаковали или извлекли папку, воспользовавшись последней версией указанных ниже архиваторов:

- ◆ WinRAR/7-Zip для Windows;
- ◆ Zipeg/iZip/UnRarX для Mac OS;
- ◆ 7-Zip/PeaZip для Linux.

Помимо этого, комплект примеров программного кода, прилагаемый к данной книге, размещен на GitHub в разделе **Packt** по адресу <https://github.com/PacktPublishing/Hands-On-Machine-Learning-for-Algorithmic-Trading>. В случае обновления программного кода он будет обновлен в существующем репозитории GitHub.

---

<sup>1</sup> Данный раздел относится к электронным архивам оригинального издания книги Stefan Jansen "Hands-On Machine Learning for Algorithmic Trading". Описание электронного архива к русскому изданию см. в конце книги. — *Прим. ред.*

Мы также располагаем другими комплектами примеров программного кода, которые можно выбрать из нашего богатого каталога книг и видеороликов, предлагаемого на странице <https://github.com/PacktPublishing/>. Можете убедиться сами!

## Скачивание цветных изображений

Мы также предоставляем PDF-файл с цветными изображениями скриншотов/диаграмм, используемых в этой книге. Вы можете скачать его здесь: [https://static.packt-cdn.com/downloads/9781789346411\\_ColorImages.pdf](https://static.packt-cdn.com/downloads/9781789346411_ColorImages.pdf).<sup>2</sup>

## Принятые в книге условные обозначения

В этой книге используется ряд текстовых условных обозначений.

- ◆ Код в тексте обозначает кодовые слова в тексте, имена таблиц базы данных, ввод данных пользователем и дескрипторы Twitter. Приведем пример: "Алгоритм продолжает исполняться после вызова функции `run_algorithm()` и возвращает тот же кадр данных с бэктестовой результативностью".
- ◆ Блок кода выглядит следующим образом:
 

```
interesting_times = extract_interesting_date_ranges(returns=returns)
(interesting_times['Fall2015'].to_frame('pf')
 .join(benchmark_rets)
 .add(1).cumprod().sub(1)
 .plot(lw=2, figsize=(14, 6),
       title='Паника после голосования по Брекситу'))
```
- ◆ **Полужирный шрифт** обозначает элемент интерфейса — команду меню, кнопку и др.
- ◆ *Курсивом* выделены новые термины, важные слова или слова, отображаемые на экране.
- ◆ Узким шрифтом выделены имена файлов, папок.



Данный элемент обозначает подсказку или совет.



Данный элемент обозначает предупреждение или предостережение.

<sup>2</sup> PDF-файл с цветными изображениями к русскому изданию книги представлен в электронном архиве, который доступен по ссылке <ftp://ftp.bhv.ru/9785977565950.zip> и со страницы книги на веб-сайте издательства по адресу [www.bhv.ru](http://www.bhv.ru). — *Прим. ред.*

# 1

## Машинное обучение для торговли на финансовых рынках

Алгоритмическая торговля на финансовых рынках опирается на компьютерные программы, которые исполняют алгоритмы, автоматизирующие некоторые или все элементы торговой стратегии. Алгоритмы представляют собой последовательность шагов или правил по достижению цели и могут принимать множество форм. В случае *машинного обучения* (МО — machine learning, ML) эти алгоритмы ставят своей задачей усвоить другие алгоритмы, а именно правила по достижению цели на основе данных, такой как минимизация ошибки предсказания.

Указанные алгоритмы кодируют различные действия портфельного менеджера, который наблюдает за рыночными транзакциями и анализирует соответствующие данные, принимая решения о размещении ордеров на покупку или продажу. Последовательность ордеров определяет элементы содержимого инвестиционного портфеля, которые со временем призваны произвести финансовые возвраты на инвестированный капитал, привлекательные для поставщиков капитала, с учетом их аппетита к риску.

В конечном счете цель активного инвестиционного менеджмента заключается в достижении альфы, т. е. финансовых возвратов, превышающих эталонный показатель, используемый для оценивания. Фундаментальный закон активного инвестиционного менеджмента применяет *информационное соотношение* (information ratio, IR), которым выражается стоимость активного менеджмента как отношение портфельных финансовых возвратов сверх возвратов эталонного показателя, обычно индекса, к волатильности этих возвратов. Он аппроксимирует это информационное соотношение как произведение *информационного коэффициента* (information coefficient, IC), которым измеряется качество прогноза, как корреляция возвратов с результатами, и широты стратегии, выраженной как квадратный корень из числа ставок.

Следовательно, ключом к генерированию альфы является прогнозирование. Успешные предсказания, в свою очередь, требуют опережающей информации или превосходящей способности обрабатывать публичную информацию. Алгоритмы обеспечивают оптимизацию на протяжении всего инвестиционного процесса, от размещения финансовых средств среди портфельных активов до генерирования идей,

исполнения сделок и риск-менеджмента. Использование МО для алгоритмической торговли, в частности, нацелено на более эффективное использование обычных и альтернативных данных с целью порождения как более качественных, так и более действенных прогнозов, тем самым улучшая добавочную стоимость активного менеджмента.

Исторически алгоритмическая торговля на финансовых рынках обычно более узко определялась как автоматизация исполнения торговли с целью минимизации затрат, предлагаемых стороной продажи, но мы рассмотрим более полную перспективу, поскольку использование алгоритмов и, в частности, машинного обучения повлияло на более широкий спектр деятельности от генерирования идей и конструирования альфа-факторов до размещения финансовых средств среди портфельных активов, калибровки позиций и тестирования и оценивания стратегий.

В этой главе рассматривается более широкая картина того, как использование МО стало критически важным источником конкурентного преимущества в инвестиционной индустрии и где именно оно вписывается в инвестиционный процесс, обеспечивая условия для стратегий алгоритмической торговли.

В этой главе мы рассмотрим следующие темы:

- ◆ как организована эта книга и кто должен ее читать;
- ◆ как МО стало играть стратегическую роль в алгоритмической торговле;
- ◆ как конструировать и исполнять торговую стратегию;
- ◆ как МО наращивает добавочную стоимость в стратегии алгоритмической торговли.

## **Как читать эту книгу**

Если вы читаете эти строки, то, вероятно, знаете, что МО стало стратегическим потенциалом во многих отраслях промышленности, включая инвестиционную индустрию. Взрывной рост цифровых данных, который стимулирует значительную долю роста МО, оказывает особенно сильное влияние на инвестиции, которые уже имеют долгую историю использования сложных моделей по обработке информации. Из размаха торговли по всем классам финансовых активов следует, что в дополнение к рыночным и фундаментальным данным, которые раньше были в центре внимания аналитических усилий, становится актуальным широкий спектр новых альтернативных данных.

Возможно, вы также столкнулись с пониманием того, что успешное применение МО или науки о данных требует интеграции статистических знаний, вычислительных навыков и компетентного знания предметной области на индивидуальном или командном уровне. Другими словами, важно задавать правильные вопросы, выявлять и понимать данные, которые способны обеспечивать ответы, разворачивать широкий спектр инструментов для получения результатов и интерпретировать их таким образом, который приводит к правильным решениям.

Следовательно, в этой книге рассматривается комплексный подход к применению МО в области инвестиций и торговли на финансовых рынках. В этом разделе мы расскажем о том, чего ожидать, как достигать своих целей и что вам нужно для достижения своих целей и получения удовлетворения от всего этого процесса.

## Чего ожидать

Цель этой книги — ознакомить вас со стратегической перспективой, концептуальным пониманием и практическими инструментами, позволяющими наращивать добавочную стоимость от применения МО в торговом и инвестиционном процессах. В этой связи она охватывает МО не как отдельное мероприятие, а как важный элемент данного процесса.

Прежде всего, она охватывает широкий спектр алгоритмов контролируемого, неконтролируемого и подкрепляемого обучения, широко используемых для извлечения сигналов из разнообразных источников данных, относящихся к различным классам финансовых активов. Она знакомит с рабочим потоком МО и концентрирует внимание читателя на примерах практического его применения с соответствующими данными и многочисленными примерами исходного кода. Однако она также развивает математический и статистический фундамент с целью обеспечения настройки алгоритма или интерпретации результатов.

В этой книге признается, что инвесторы могут извлекать добавочную стоимость из сторонних данных больше, чем другие отрасли промышленности. Как следствие, в ней охвачены не только способы работы с рыночными и фундаментальными данными, но и способы привлечения, оценивания, обработки и моделирования альтернативных источников данных, таких как неструктурированные текстовые и графические данные.

Настоящая книга увязывает использование МО с исследованием и оцениванием альфа-факторов, с квантитативными и факторно обусловленными стратегиями и вводит портфельный менеджмент в качестве контекста для развертывания стратегий, объединяющих несколько альфа-факторов. В ней также подчеркивается, что МО может наращивать добавочную стоимость, выходя за пределы предсказаний, относящихся к ценам на индивидуальные финансовые активы, например к размещению финансовых средств среди портфельных активов, и обращается к проблеме рисков ложных обнаружений, возникающих в результате использования МО с крупными наборами данных для разработки торговой стратегии.

Не следует удивляться тому, что эта книга не содержит рекомендаций по инвестициям или готовых алгоритмов торговли. Вместо этого она знакомит со строительными блоками, необходимыми для выявления, оценивания и комбинирования наборов данных, подходящих для той или иной инвестиционной цели, отбирает и применяет автоматически обучающиеся алгоритмы к этим данным, а также разрабатывает и тестирует стратегии алгоритмической торговли, основываясь на полученных результатах.

## Кто должен прочесть эту книгу

Вы найдете книгу информативной, если вы аналитик, исследователь данных или инженер в области МО с пониманием финансовых рынков и интересом к стратегиям торговли. Вы также должным образом оцените ее как инвестиционный профессионал, который стремится эффективно задействовать МО для принятия более оптимальных решений.

Если ваша квалификация — программное обеспечение и МО, вы можете бегло просмотреть или просто пропустить вводный материал по МО. Если ваши компетентные знания принадлежат области инвестиций, то вы, вероятно, будете знакомы с некоторым или всем финансовым контекстом. Вы, пожалуй, найдете книгу наиболее полезной в качестве краткого обзора ключевых алгоритмов, строительных блоков и примеров использования, чем в качестве источника специализированного анализа конкретного алгоритма или стратегии. Однако в этой книге мы исходим из того, что вы заинтересованы в продолжении изучения этой очень динамичной области. С этой целью в ней приводятся справочные материалы со ссылками на многочисленные ресурсы с целью поддержания вашего движения вперед к индивидуальным стратегиям торговли, в которых эффективно задействуются и развиваются охватываемые ею фундаментальные методы и инструменты.

Вы не должны испытывать трудностей в использовании языка Python 3 и различных научно-вычислительных библиотек, таких как NumPy, pandas или SciPy, и должны быть заинтересованы подхватить на этом пути целый ряд других. Некоторый опыт работы с МО и библиотекой sklearn был бы полезен, однако мы кратко рассмотрим базовый рабочий поток и будем давать различные справочные материалы со ссылками на всевозможные ресурсы с целью заполнения пробелов или более глубокого погружения в рассматриваемую тему.

## Как эта книга организована

В книге представлено всестороннее введение в то, как МО может наращивать добавочную стоимость в конструировании и исполнении стратегий торговли. Книга состоит из четырех условных частей, которые охватывают различные аспекты процесса привлечения источников данных и разработки стратегии, а также разные решения всевозможных задач МО.

## Каркас — от данных к конструированию стратегии

Первая часть обеспечивает каркас для разработки стратегий алгоритмической торговли. Она сосредоточена на данных, которые питают энергией обсуждаемые в этой книге автоматически обучающиеся (МО) алгоритмы и стратегии, описывает, как МО может использоваться для выведения торговых сигналов, и как развертывать и оценивать стратегии в качестве компонента инвестиционного портфеля.

В оставшейся части этой главы резюмируется то, как и почему МО заняло центральное место в инвестициях, описывается процесс торговли и обрисовывается то, как МО может наращивать добавочную стоимость.

*Глава 2 "Рыночные и фундаментальные данные"* охватывает источники и работу с исходными тиковыми данными, поступающими с бирж, и данными финансовой отчетности, а также способы доступа к многочисленным поставщикам общедоступных данных, на которые мы будем опираться на протяжении всей этой книги.

*Глава 3 "Альтернативные данные для финансов"* предоставляет категории и критерии для оценивания качества стремительно растущего числа источников и поставщиков. В ней также демонстрируются способы создания альтернативных наборов данных путем выскабливания веб-сайтов, например для сбора стенограмм телеконференций о корпоративных заработках для использования в алгоритмах естественно-языковой обработки и сентиментного анализа во второй части книги.

*Глава 4 "Исследование альфа-факторов"* обеспечивает каркас для понимания того, как работают факторы и как измерять их эффективность, например, используя информационный коэффициент (IC). В ней демонстрируется способ конструирования альфа-факторов из данных с помощью библиотек Python в режиме офлайн и на веб-платформе Quantopian. В ней также вводятся библиотека zipline для бэкестирования факторов и библиотека alphalens для оценивания их предсказательной мощности.

*Глава 5 "Оценивание стратегии"* знакомит с тем, как строить, тестировать и оценивать стратегии торговли, используя исторические данные с помощью библиотеки zipline в режиме офлайн и на веб-платформе Quantopian. Она представляет и демонстрирует способы вычисления результативности портфеля и метрических показателей риска с помощью библиотеки pyfolio. В ней также рассматриваются способы управления методологическими сложностями бэкестирования стратегий и вводятся методы оптимизации стратегии с точки зрения портфельного риска.

## **Основы машинного обучения**

Вторая условная часть книги охватывает фундаментальные контролируемые и неконтролируемые обучающиеся алгоритмы и иллюстрирует их применение к стратегиям торговли на финансовых рынках. Она также знакомит с веб-платформой Quantopian, где можно эффективно задействовать и комбинировать данные и методы МО, разработанные в этой книге, для реализации алгоритмических стратегий, которые исполняют сделки на живых рынках.

*Глава 6 "Процесс машинного обучения"* закладывает фундамент, давая описания того, как формулировать, тренировать и настраивать автоматически обучающиеся модели и оценивать их предсказательную результативность в качестве систематического рабочего процесса.

*Глава 7 "Линейные модели"* показывает, как использовать линейную и логистическую регрессию для выведения заключения и предсказания и применять регуляризацию для управления риском перепогонки. Она знакомит с торговой платформой Quantopian и демонстрирует способы построения факторных моделей и предсказания цен на финансовые активы.

*Глава 8 "Модели временных рядов"* посвящена одномерным и многомерным временным рядам, включая векторные авторегрессионные модели и коинтеграционные тесты, а также способы их применения к стратегиям парной торговли.

*Глава 9 "Байесово машинное обучение"* знакомит с тем, как формулировать вероятностные модели и как извлечение выборок из вероятностных распределений методами Монте-Карло марковских цепей (Markov Chain Monte Carlo, MCMC) и вариационные байесовы методы обеспечивают аппроксимирование вывода. В ней также иллюстрируются способы применения библиотеки PyMC3 для вероятностного программирования с целью понимания сущности параметрической и модельной неопределенности.

*Глава 10 "Деревья решений и случайные леса"* показывает, как строить, тренировать и настраивать нелинейные древесные модели для проникновения в сущность данных и предсказания. В ней представлены древесные ансамблевые модели и иллюстрируется то, как случайные леса используют бутстраповское агрегирование для преодоления некоторых недостатков деревьев решений.

*Глава 11 "Градиентно-бустинговые машины"* демонстрирует, как использовать библиотеки xgboost, lightgbm и catboost для выполнения высокорезультативной тренировки и предсказания. Кроме того, в ней подробно анализируется настройка многочисленных гиперпараметров.

*Глава 12 "Неконтролируемое обучение"* посвящена применению задач снижения размерности и кластеризации в алгоритмической торговле. В ней используется анализ главных и независимых компонент для извлечения ведомых данными рисков факторов. В ней также представлено несколько технических решений для кластеризации и показано использование иерархической кластеризации для размещения финансовых средств среди портфельных активов.

## **Обработка естественного языка**

В третьей условной части основное внимание уделяется текстовым данным и вводятся современные методы неконтролируемого обучения для извлечения высококачественных сигналов из этого ключевого источника альтернативных данных.

*Глава 13 "Работа с текстовыми данными"* демонстрирует способы конвертирования текстовых данных в числовой формат. В ней также показано применение классификационных алгоритмов из второй условной части, связанных с сентиментным анализом, к большим наборам данных.

*Глава 14 "Тематическое моделирование"* посвящена применению байесова неконтролируемого обучения с целью извлечения латентных тем, которое способно резюмировать большое число документов и предлагать более эффективные способы разведывания текстовых данных либо использования тем в качестве признаков в классификационной модели. В ней демонстрируется способ применения этого технического решения к стенограммам телеконференций о корпоративных заработках, источники которых будут привлечены в *главе 3*, и к годовым отчетам, по-

даваемым в Комиссию по ценным бумагам и биржам (Securities and Exchange Commission, SEC).

*Глава 15 "Векторное вложение слов"* посвящена использованию нейронных сетей для заучивания ультрасовременных языковых признаков в форме словарных векторов, которые улавливают семантический контекст намного лучше, чем традиционные текстовые признаки, и представляют собой очень перспективное направление для извлечения торговых сигналов из текстовых данных.

## Глубокое и подкрепляемое обучение

Четвертая условная часть знакомит с глубоким обучением и подкрепляемым обучением.

*Глава 17 "Глубокое обучение"* знакомит с библиотеками Keras, TensorFlow и PyTorch — наиболее популярными каркасами глубокого обучения — и иллюстрирует способы тренировки и настройки различных архитектур.

*Глава 18 "Сверточные нейронные сети"* иллюстрирует применение CNN-сетей с графическими и текстовыми данными.

*Глава 19 "Рекуррентные нейронные сети"* знакомит с RNN-сетями для данных временных рядов.

*Глава 20 "Автокодировщики и генеративные состязательные сети"* показывает способы применения глубоких нейронных сетей для неконтролируемого автоматического обучения с автокодировщиками и знакомит с GAN-сетями, которые производят синтетические данные.

*Глава 21 "Подкрепляемое обучение"* демонстрирует использование автоматического обучения с максимизацией подкрепления для создания динамических агентов, которые заучивают функцию линии поведения, основываясь на вознаграждениях, с использованием программного инструмента Open AI gym.



*Главы 17–21* (а также приложения и цветные иллюстрации к *главам 1–15*) представлены в электронном архиве данной книги, размещенном на сайте издательства. Электронный архив можно скачать по ссылке <ftp://ftp.bhv.ru/9785977565950.zip> и со страницы книги на веб-сайте издательства по адресу [www.bhv.ru](http://www.bhv.ru).

## Что вам нужно для успеха

Изложение в книге вращается вокруг применения автоматически обучающихся алгоритмов к разным наборам данных. Значительный дополнительный материал размещен в репозитории GitHub, который обеспечивает повторение изученных тем и эксперименты с примерами, обсуждаемыми в книге. В нем представлены дополнительные сведения и инструкции, а также многочисленные справочные материалы.

## Источники данных

Мы будем использовать свободно доступные исторические данные из рыночных, фундаментальных и альтернативных источников. Главы 2 и 3 охватывают характеристики и обращаются к этим источникам данных и знакомят с ключевыми поставщиками, которых мы будем использовать на протяжении всей книги. Сопутствующий репозиторий GitHub (см. далее) содержит инструкции о том, как получать или создавать некоторые наборы данных, которые мы будем использовать повсюду, и включает некоторые меньшие наборы данных.

Несколько примеров источников данных, которые мы будем привлекать и с которыми будем работать, включают следующие, но не ограничиваются только ими:

- ◆ данные ордерной книги протокола передачи данных ИТЧ биржи NASDAQ;
- ◆ документы официальной отчетности в системе электронного сбора, анализа и извлечения данных (Electronic Data Gathering, Analysis and Retrieval, EDGAR) Комиссии по ценным бумагам и биржам США (U.S. Securities and Exchange Commission, SEC);
- ◆ стенограммы телеконференций о корпоративных заработках из портала Seeking Alpha (<https://seekingalpha.com/>);
- ◆ дневные цены рыночной площадки Quandl и другие точки данных более чем по 3000 акций США;
- ◆ различные макроэкономические фундаментальные данные Федеральной резервной системы и др.;
- ◆ крупный набор данных отзывов о деятельности предприятий веб-сайта Yelp и наборы данных социальной сети Twitter;
- ◆ снимковые данные по нефтяным танкерам.

Некоторые данные имеют размер в несколько гигабайт (например, данные биржи NASDAQ и документы финансовой отчетности, подаваемые в комиссию SEC). Такие моменты будут оговариваться в блокнотах Jupyter.

## Репозиторий GitHub

Репозиторий GitHub содержит блокноты Jupyter, которые подробнее иллюстрируют многие понятия и модели. Отсылки к блокнотам рассеяны по всей книге, где они используются. Каждая глава имеет собственный каталог с отдельными инструкциями, где это необходимо, а также ссылку на содержимое конкретной главы.

Блокноты Jupyter — отличный инструмент для создания воспроизводимых вычислительных повествований. Он позволяет пользователям создавать и обмениваться документами, которые соединяют в себе "живой" код с повествовательным текстом, математическими уравнениями, визуализациями, интерактивными элементами управления и другим богатым выводимым материалом. Они также предостав-

ляют строительные блоки для интерактивных вычислений с данными, такие как файловый браузер, консоли и текстовый редактор<sup>1</sup>.



Файлы исходного кода размещены по адресу <https://github.com/PacktPublishing/Hands-On-Machine-Learning-for-Algorithmic-Trading>.

## Библиотеки Python

В данной книге используется Python 3.7, и рекомендуется применение мини-версии дистрибутива Anaconda — `miniconda` — для установки менеджера пакетов `conda` и создания среды `conda` с целью установки реквизитных библиотек. В связи с этим репозиторий GitHub содержит манифестный файл `environment.yml`. Пожалуйста, обратитесь к инструкциям по установке, указанным в файле `README` в репозитории GitHub<sup>2</sup>.

## Рост популярности машинного обучения в инвестиционной индустрии

За последние несколько десятилетий инвестиционная индустрия значительно эволюционировала и продолжает это делать в условиях возросшей конкуренции, технического прогресса и сложных экономических условий. В этом разделе мы рассмотрим несколько ключевых трендов, которые сформировали инвестиционную среду в целом, а также контекст алгоритмической торговли на финансовых рынках, в частности, и связанные с этим темы, которые будут повторяться на протяжении всей книги.

Тренды, которые привели алгоритмическую торговлю и МО к нынешнему видному положению, включают:

- ◆ изменения в микроструктуре рынка, такие как распространение электронной торговли и интеграция рынков по всем классам финансовых активов и географиям;
- ◆ разработку инвестиционных стратегий с учетом влияния рисков факторов в отличие от классов финансовых активов;
- ◆ революции в вычислительной мощности, генерировании и управлении данными, а также аналитических методах;

---

<sup>1</sup> Если по каким-то причинам файл блокнота из репозитория книги открыть не удастся, то имеется обходной путь — перейти на веб-сайт обозревателя блокнотов `nbviewer` по адресу <https://nbviewer.jupyter.org/>, вставить в поисковое поле адрес нужного блокнота и нажать кнопку **Go!**. — *Прим. перев.*

<sup>2</sup> См. подробное описание организации работы с блокнотами Jupyter и настройку дистрибутива Anaconda в *приложении 1*. — *Прим. перев.*

- ◆ повышенную результативность первопроходцев в алгоритмической торговле по сравнению с человеческими, дискреционными инвесторами.

В дополнение к этому финансовые кризисы 2001 и 2008 гг. повлияли на подход инвесторов к диверсификации и риск-менеджменту и привели к появлению недорогих пассивных инвестиционных механизмов в форме *биржевых фондов*<sup>3</sup> (ETF). На фоне низкой доходности и низкой волатильности после кризиса 2008 г. экономные инвесторы перевели 2 трлн долларов из активно управляемых взаимных фондов<sup>4</sup> в пассивно управляемые биржевые фонды ETF. Конкурентное давление также отражается в более низких хедж-фондовых комиссионных, которые к 2017 г. упали с традиционных 2%-х комиссионных за годовой менеджмент и 20%-й доли в прибыли в среднем соответственно до 1,48 и 17,4%.

## От торговли электронной к торговле высокочастотной

Электронная торговля радикально продвинулась в плане возможностей, объема, охвата классов финансовых активов и географии с тех пор, как в 1960-х годах сети начали маршрутизировать цены на компьютерные терминалы.

Фондовые рынки, т. е. рынки долевых ценных бумаг<sup>5</sup> частных компаний, возглавили этот тренд во всем мире. Правила обработки ордеров комиссией SEC от 1997 г. внесли в биржи конкуренцию через *электронные коммуникационные сети*<sup>6</sup> (ECN). Сети ECN представляют собой автоматизированные *альтернативные системы торговли* (Alternative Trading Systems, ATS), которые сочетают ордера на покупку и продажу по конкретным ценам, прежде всего на долевые ценные бумаги и валюты, и зарегистрированы как брокеры-дилеры. Она позволяет крупным брокерам и индивидуальным биржевым торговцам в различных географических точках торговать напрямую без посредников как на биржах, так и в нерабочее время. Темные пулы

<sup>3</sup> Биржевой фонд (exchange-traded fund, ETF), или торгуемый на бирже фонд, — это ликвидная ценная бумага, которая отслеживает фондовый индекс, товар, облигации или корзину финансовых активов. См. <https://www.investopedia.com/terms/e/etf.asp>. — *Прим. перев.*

<sup>4</sup> Взаимный фонд (mutual fund), или фонд взаимных инвестиций, — это финансовый механизм, состоящий из денежных средств, собранных от многочисленных инвесторов с целью инвестирования в ценные бумаги, такие как акции, облигации, инструменты денежного рынка и другие финансовые активы. Представляет собой портфель активов, тщательно отобранных и приобретенных профессиональными финансистами на вложения многих тысяч мелких вкладчиков. См. <https://www.investopedia.com/terms/m/mutualfund.asp>. — *Прим. перев.*

<sup>5</sup> Долевая ценная бумага (equity) — это ценная бумага, обеспечивающая право собственности в компании через покупку обыкновенных либо привилегированных акций (долей). Термин equity также означает нетто-стоимость или чистую стоимость компании. См. [http://www.morningstar.com/InvGlossary/equity\\_definition\\_what\\_is.aspx](http://www.morningstar.com/InvGlossary/equity_definition_what_is.aspx). — *Прим. перев.*

Рынок долевых ценных бумаг (equity market), или фондовый рынок, — это рынок, где выпускаются в обращение и торгуются акции, т. е. доли собственности в компании, через биржу или внебиржевые рынки. См. <https://www.investopedia.com/terms/s/shareholdersequity.asp>. — *Прим. перев.*

<sup>6</sup> Электронные коммуникационные сети ECN (electronic communication networks) — это электронная система осуществления сделок купли-продажи биржевых товаров. — *Прим. перев.*

ликвидности — это еще один тип альтернативных систем торговли, которые позволяют инвесторам размещать ордера и торговать, не раскрывая свою информацию публично, как в ордерной книге, поддерживаемой биржей. Темные пулы, выросшие из постановления комиссии SEC 2007 г., часто размещаются в крупных банках и подлежат регламентированию со стороны комиссии SEC.

С ростом влияния электронной торговли алгоритмы затратно эффективного исполнения прошли ускоренную эволюцию, и их принятие быстро распространилось со стороны продажи на сторону покупки<sup>7</sup> и по всем классам активов. Автоматизированная торговля появилась примерно в 2000 г. как инструмент на стороне продажи, направленный на затратно эффективное исполнение сделок, который распределяет ордера во времени, ограничивая их влияние на финансовый рынок<sup>8</sup>. Эти инструменты распространились на сторону покупки и становились все более изощренными, принимая в расчет, например, транзакционные издержки и ликвидность, а также краткосрочные ценовые и объемные прогнозы.

*Прямой доступ к рынку* (Direct Market Access, DMA) дает биржевому торговцу более высокий контроль над исполнением, позволяя ему отправлять ордера непосредственно на биржу, используя инфраструктуру и идентификацию участника рынка у брокера, который является членом биржи. Спонсируемый доступ снимает с брокеров обязанность осуществлять контроль предсделочных рисков и формирует основу для *высокочастотной торговли* (high-frequency trading, HFT).

Под термином "высокочастотная торговля" имеются в виду автоматизированные сделки с финансовыми инструментами, которые исполняются с чрезвычайно низкой временной задержкой в микросекундном интервале, и где участники занимают позиции в течение очень коротких периодов. Ее цель заключается в выявлении и использовании неэффективности микроструктуры рынка, институциональной инфраструктуры торговых площадок. За последние 10 лет высокочастотная торговля существенно выросла и по оценкам составляет примерно 55% объема торговли на фондовых рынках США и около 40% на фондовых рынках Европы. То же самое произошло на фьючерсных рынках, где высокочастотная торговля выросла примерно до 80% от объемов фьючерсов на иностранную валюту и двух третей от объемов фьючерсов на процентные ставки и 10-летние казначейские векселя (согласно ежегодному "Обзору доступа к финансовым службам FAS 2016")<sup>9</sup>.

Стратегии высокочастотной торговли ориентированы на получение малой прибыли в расчете на сделку с использованием пассивных или агрессивных стратегий. *Пас-*

---

<sup>7</sup> Сторона покупки (buy side) и сторона продажи (sell side) относятся к фирмам, которые соответственно приобретают и торгуют товарами и услугами. Применительно к финансам, сторона покупки относится к покупке ценных бумаг и включает инвестиционных менеджеров, пенсионные фонды и хедж-фонды, а сторона продажи относится к фирмам, которые выпускают, продают или торгуют ценными бумагами, и включает инвестиционные банки, консультативные фирмы и корпорации. — *Прим. перев.*

<sup>8</sup> Влияние на финансовый рынок (market impact) означает влияние, которое участник рынка оказывает на цену финансового актива, когда он покупает или продает его. — *Прим. перев.*

<sup>9</sup> См. <http://data.imf.org/?sk=E5DCAB7E-A5CA-4892-A6EA-598B5463A34C>. — *Прим. перев.*

сивные стратегии включают арбитражную торговлю<sup>10</sup>, ориентированную на получение прибыли от очень небольших ценовых разниц для одного и того же финансового актива или его производных инструментов, торгуемых на разных площадках. Агрессивные стратегии включают предвосхищение заявок или инициацию импульса. Предвосхищение заявок, также именуемое *обнаружением ликвидности*, охватывает алгоритмы, которые размещают небольшие разведывательные ордера с целью обнаружения скрытой ликвидности у крупных институциональных инвесторов и торгуют перед крупным ордером с целью извлечь выгоду из последующих ценовых движений. Инициация импульса подразумевает алгоритм, исполняющий и аннулирующий серию ордеров с целью заманить другие алгоритмы высокочастотной торговли агрессивнее покупать (или продавать), и извлекает выгоду из результирующих ценовых изменений.

В связи с этим регуляторы выражали озабоченность по поводу потенциальной взаимосвязи между некоторыми агрессивными стратегиями высокочастотной торговли и повышенной хрупкостью и волатильностью рынка, например, во время молниеносного обвала в мае 2010 г., волатильности рынка казначейских обязательств в октябре 2014 г. и молниеносного обвала промышленного индекса Доу-Джонса в августе 2015 г. более чем на 1000 пунктов. Вместе с тем из-за присутствия высокочастотной торговли рыночная ликвидность увеличивалась вместе с объемами торгов, что снижало совокупные транзакционные издержки<sup>11</sup>.

Сочетание сниженных торговых объемов на фоне снижения волатильности и растущей стоимости технологии и доступа как к данным, так и к торговым площадкам привело к финансовому давлению. По оценкам, совокупные поступления высокочастотной торговли от американских акций впервые с 2008 г. упали ниже 1 млрд долларов по сравнению с 7,9 млрд долларов в 2009 г.

Этот тренд привел к отраслевой консолидации за счет различных приобретений, например, крупнейшей зарегистрированной проприетарной торговой фирмой Virtu Financial и долевыми инфраструктурными инвестициями, такими как новый маршрут с ультранизкой временной задержкой Go West между Чикаго и Токио. Одновременно с этим такие стартапы, как Alpha Trading Lab, обеспечивают наличие торговой инфраструктуры и данных высокочастотной торговли, демократизируя высокочастотную торговлю за счет привлечения разработчиков алгоритмов через Интернет в обмен на долю прибыли.

<sup>10</sup> Арбитраж (arbitrage) — это практика использования преимуществ разницы цен между двумя или более рынками и заключения комбинации совпадающих сделок, которые извлекают выгоду из таких дисбалансов. Например, под валютным арбитражем понимается одновременная покупка и продажа валюты на разных валютных рынках с целью получения прибыли от разницы обменных курсов в двух местах. См. <https://www.thefreelibrary.com/Currency+arbitrage+as+a+tool+of+corporate+financial+management-a0469641512>. — *Прим. перев.*

<sup>11</sup> Транзакционные издержки (transaction cost) — это стоимость осуществления любой экономической сделки, сопровождающая участие в рынке. В инвестировании это затраты, понесенные при покупке или продаже активов, такие как комиссии и спред. См. [https://en.wikipedia.org/wiki/Transaction\\_cost](https://en.wikipedia.org/wiki/Transaction_cost). — *Прим. перев.*

## Факторное инвестирование и умные бета-фонды

Финансовый возврат<sup>12</sup>, обеспечиваемый активом, является функцией неопределенности или риска, связанного с финансовой инвестицией. Инвестирование в долевые ценные бумаги подразумевает, например, принятие корпоративного делового риска, а инвестиция в облигации — принятие риска дефолта.

Поскольку конкретные характеристики рисков несут в себе предсказания финансовых возвратов, выявление и прогнозирование поведения этих рисковых факторов<sup>13</sup> становится первостепенным направлением при конструировании инвестиционной стратегии. Это позволяет добиться ценных торговых сигналов и является ключом к превосходящим результатам активного менеджмента. Понимание рисковых факторов в финансовой индустрии со временем существенно эволюционировало и повлияло на то, как МО используется для алгоритмической торговли.

*Теория современного инвестиционного портфеля* (МРТ)<sup>14</sup> ввела разграничение между источниками идиосинкратического и систематического риска для определенного актива. Идиосинкратический риск<sup>15</sup> может быть устранен путем диверсификации, но систематический риск — нет. В начале 1960-х годов *модель ценообразования капитальных активов* (САРМ)<sup>16</sup> выявила один-единственный фактор, сти-

<sup>12</sup> Финансовый возврат (return), или возврат на инвестицию или финансовая отдача, — деньги, сделанные или потерянные на инвестициях. Возврат выражается номинально как изменение денежной стоимости инвестиции с течением времени, либо в процентах из соотношения прибыли к инвестициям. В этом случае он именуется возвратностью.

См. <https://www.investopedia.com/terms/r/return.asp>. — Прим. перев.

<sup>13</sup> Рисковый фактор (risk factor) — это измеримая характеристика или элемент, изменение которого может повлиять на стоимость актива, например, обменный курс, процентная ставка и рыночная цена. См. <http://www.businessdictionary.com/definition/risk-factor.html>. — Прим. перев.

<sup>14</sup> Теория современного инвестиционного портфеля (modern portfolio theory, МРТ), или портфельная теория Марковица, — это теория, которая описывает то, как инвесторы, не склонные к риску, могут создавать портфели для оптимизации или максимизации ожидаемой возвратности на основе заданного уровня рыночного риска, подчеркивая, что риск является неотъемлемой частью более высокого вознаграждения. См. <https://www.investopedia.com/terms/m/modernportfoliotheory.asp>. — Прим. перев.

Портфель (portfolio), как правило, представляет собой группу финансовых активов, таких как акции, облигации, товары, валюты и эквиваленты денежных средств, а также их фондовые аналоги, включая взаимные, биржевые и закрытые фонды. См. <https://www.investopedia.com/terms/p/portfolio.asp>. — Прим. перев.

<sup>15</sup> Идиосинкратический риск (idiosyncratic risk) — это риск, который относится к факторам, влияющим на отдельный актив, по сравнению с систематическим риском, который относится к более широким трендам, сказывающимся на всем рынке в целом.

См. <https://www.investopedia.com/search?q=idiosyncratic+asset>. — Прим. перев.

<sup>16</sup> Модель ценообразования капитальных активов (Capital Asset Pricing Model, САРМ) описывает связь между систематическим риском и ожидаемой возвратностью от активов, в особенности от акций. Указанная модель является однофакторной, и единственным фактором служит бета-фактор, который, согласно модели, показывает, насколько хорошо акция движется относительно рынка. Акции, которые двигались выше рынка, имели более высокий бета-фактор и, следовательно, более высокий риск и возврат.

См. <https://www.investopedia.com/terms/c/capm.asp>. — Прим. перев.

мулирующий финансовые возвраты от всех активов: финансовый возврат, приносимый рыночным портфелем<sup>17</sup> сверх казначейских векселей. Рыночный портфель состоял из всех торгуемых ценных бумаг, взвешенных по их рыночной стоимости. Систематическое влияние на актив рынка измеряется бетой, которая представляет собой корреляцию между возвратами, приносимыми активом, и рыночным портфелем.

Признание, что риск того или иного актива зависит не от него в отдельности, а от того, как он движется по отношению к другим активам и рынку в целом, стало крупным концептуальным прорывом. Другими словами, активы получают рисковую премию не из-за их идиосинкратических особенностей, а из-за влияния на них базовых факторных рисков.

Вместе с тем крупный пласт академической литературы и многолетний опыт инвестирования опровергли предсказание CAPM-модели о том, что рисковые премии финансовых активов зависят от влияния на них лишь одного-единственного фактора, измеряемого бетой финансового актива. Напротив, с тех пор были обнаружены многочисленные дополнительные рисковые факторы. Фактор — это квантифицируемый сигнал, атрибут или любая величина, которая исторически коррелировала с будущими возвратами от акций и, как ожидается, останется коррелированной в будущем.

Эти рисковые факторы были объявлены аномалиями, поскольку они противоречили *гипотезе об эффективном рынке* (Efficient Market Hypothesis, EMH), которая доказывала, что рыночное равновесие всегда будет формировать цену ценных бумаг в соответствии с CAPM-моделью, вследствие чего никакие другие факторы не должны иметь предсказательной мощности. Экономическая теория в основе факторов может быть либо рациональной, когда премии за факторные риски компенсируют низкие финансовые возвраты в плохие времена, либо поведенческой, когда агенты не в состоянии задействовать избыточные возвраты за счет арбитража.

Хорошо известные аномалии включают стоимостные, размерные и импульсные эффекты, которые помогают предсказывать возвраты, учитывая при этом рыночный фактор CAPM-модели. Размерный эффект зиждется на том факте, что малые фирмы систематически превосходят по результативности крупные, как было обнаружено Банцем (Banz, 1981) и Рейнганумом (Reinganum, 1981). Стоимостной эффект (Basu, 1982) утверждает, что фирмы с низкими метриками стоимости демонстрируют повышенную результативность. Это говорит о том, что фирмы с низкими ценовыми мультипликаторами, такими как соотношение цены к корпоративным заработкам или цены к балансовой стоимости, показывают более высокую результативность, чем их более дорогие сверстники (как было предложено изобретателями инвестиций в стоимость Бенджамином Грэмом (Benjamin Graham) и Дэвидом Доддом (David Dodd) и популяризировано Уорреном Баффетом (Warren Buffet)).

---

<sup>17</sup> Рыночный портфель (market portfolio) — это теоретическая диверсифицированная группа инвестиций, каждый актив которой взвешен пропорционально его суммарному присутствию на рынке. См. <https://www.investopedia.com/terms/m/market-portfolio.asp>. — Прим. перев.



Корпоративные заработки (earnings) обычно относятся к чистому доходу после уплаты налогов, иногда именуемому нижней строкой отчета о прибылях и убытках, или прибылями компании, согласно формуле:

Валовая выручка

*минус* Стоимость реализованной продукции/услуг

= Чистая выручка

*плюс* Другие доходы

*минус* Операционные расходы

= Валовая прибыль (ЕВITDA, корпоративный заработок)

*минус* Налоги, проценты, износ и амортизация

= Чистая прибыль (нераспределенная прибыль, это тоже корпоративный заработок).

Корпоративные заработки являются основным определяющим фактором долевой цены компании, поскольку заработки и связанные с ними обстоятельства могут указывать на прибыльность и успешность предприятия в долгосрочной перспективе. Они являются, пожалуй, самым важным и наиболее изученным показателем в финансовой отчетности компании, поскольку показывают прибыльность по сравнению с оценками со стороны аналитиков и руководства компании (см. <https://www.investopedia.com/terms/e/earnings.asp>. — *Прим. перев.*)

Импульсный эффект, обнаруженный в конце 1980-х годов, в частности, Клиффордом Эссесом (Clifford Asness), партнером-основателем глобальной компании по инвестиционному менеджменту AQR, констатирует, что акции с хорошим импульсом, с точки зрения недавних 6–12-месячных возвратов, имеют более высокие возвраты в перспективе, чем слабо-импульсные акции с аналогичным рыночным риском. Исследователи еще обнаружили, что стоимостные и импульсные факторы объясняют возвраты от акций за пределами США, а также от других классов финансовых активов, таких как облигации, валюты и товары, и дополнительные рисковые факторы.

Применительно к ценным бумагам с фиксированной доходностью стоимостная стратегия называется *выездом на кривой отдачи*<sup>18</sup> и является формой премии за продолжительность. Применительно к товарам она называется *скользящей возвратностью*<sup>19</sup> с положительной возвратностью для фьючерсов с наклоном кривой

<sup>18</sup> Выезд на кривой отдачи (riding the yield curve) — это торговая стратегия, включающая покупку долгосрочной облигации и ее продажу до ее созревания с целью получения прибыли от снижения доходности, которое происходит в течение срока действия облигации.

См. <https://www.investopedia.com/terms/r/ridingtheyieldcurve.asp>. — *Прим. перев.*

<sup>19</sup> Скользящая возвратность (roll return, rolling return) представляет собой среднегодовую возвратность за период, заканчивающийся указанным годом.

См. <https://www.investopedia.com/terms/r/rollingreturns.asp>. — *Прим. перев.*

вверх и отрицательной возвратностью в противном случае. Применительно к иностранным валютам стоимостная стратегия называется *кэрри-трейдом*<sup>20</sup>.

Существует также премия за неликвидность. Ценные бумаги, которые являются более неликвидными, торгуются по низким ценам и имеют высокие средние избыточные возвраты по сравнению с их более ликвидными аналогами. Облигации с более высоким риском дефолта, как правило, в среднем имеют более высокие возвраты, отражая премию за кредитный риск. Поскольку инвесторы готовы платить за страхование от высокой волатильности, когда возвраты тяготеют к обрушению, продавцы защиты от волатильности на опционных рынках чаще всего получают высокие возвраты.

Многофакторные модели определяют риски в более широких и разнообразных терминах, чем просто рыночный портфель. В 1976 г. Стивен Росс (Stephen Ross) предложил теорию арбитражного ценообразования, которая утверждала, что инвесторы получают компенсацию за многочисленные систематические источники риска, которые нельзя диверсифицировать. Три наиболее важными макроэкономическими факторами являются рост, инфляция и волатильность наряду с производительностью, демографическим и политическим рисками. В 1992 г. Юджин Фама (Eugene Fama) и Кеннет Френч (Kenneth French) объединили рискованные факторы долевых ценных бумаг<sup>21</sup> — размер и стоимость — с рыночным фактором в единую модель, которая лучше объясняла срезовые (кросс-секционные) финансовые возвраты от акций. Позже они добавили модель, которая также включала рискованные факторы облигаций, одновременно объясняя возвраты для обоих классов активов.

Особо привлекательным аспектом рискованных факторов является их низкая или отрицательная корреляция. Например, стоимостные и импульсные рискованные факторы имеют отрицательную корреляцию, снижая риск и повышая скорректированные на риск возвраты за пределами выгоды, вытекающей из рискованных факторов. Более того, за счет кредитного плеча<sup>22</sup> и длинно-коротких инвестиционных стратегий<sup>23</sup> фак-

---

<sup>20</sup> Кэрри-трейд (carry-trade, carry) — это стратегия, в соответствии с которой высокодоходная валюта финансирует сделку с низкодоходной валютой. Иными словами, это получение прибыли на валютном рынке за счет разницы величин процентных ставок.

См. <https://www.investopedia.com/terms/c/currencycarrytrade.asp>. — Прим. перев.

<sup>21</sup> Фактор (factor) — это характеристика, поддающаяся квантификации особенность актива с существенной информацией о риске и возвратности. В случае долевых ценных бумаг наиболее известные и наиболее документированные факторы включают стоимость, размер, импульс, низкую волатильность и качество. — Прим. перев.

<sup>22</sup> Кредитное плечо (leverage), или финансовый рычаг, — это дополнительная покупательная способность, создаваемая маржинальной торговлей, позволяющая эффективно платить меньше полной цены за актив, используя заемные средства. Кредитное плечо обычно представлено в виде соотношения: например, если на торговом счете имеется 10 тыс. долларов и вы занимаете еще 10 тыс. долларов, то ваше кредитное плечо равно 2:1. — Прим. перев.

<sup>23</sup> Длинно-короткая инвестиционная стратегия (long-short strategy) — это инвестиционная стратегия, обычно связанная с хедж-фондами и с некоторыми прогрессивными традиционными менеджерами активами, которая заключается в покупке длинных акций, т. е. с ожидаемым ростом цены, и продаже

торные стратегии могут быть объединены в рыночно-нейтральные подходы. Сочетание длинных позиций в ценных бумагах, находящихся под влиянием положительных рисков, с короткими позициями в ценных бумагах, находящихся под влиянием отрицательных рисков, позволяет собирать динамические рисковые премии.

Как следствие, факторы, объяснявшие возвраты выше и за пределами однофакторной CAPM-модели, встраивались в так называемые инвестиционные стили, которые накренивают портфели в пользу одного или нескольких факторов, и в результате активы начали мигрировать в факторно-ориентированные портфели. Финансовый кризис 2008 г. подчеркнул, насколько метки классов финансовых активов могут вводить в заблуждение и создавать ложное ощущение диверсифицированности, когда инвесторы не смотрят на лежащие в основе факторные риски, поскольку все классы активов обрушились вместе.

За последние несколько десятилетий квантитативное факторное инвестирование<sup>24</sup> эволюционировало от простого подхода, основанного на двух или трех стилях, до многофакторных умных или экзотических бета-продуктов. В 2017 г. умные бета-фонды<sup>25</sup> пересекли отметку 1 трлн долларов активов под управлением, что свидетельствует о популярности гибридной инвестиционной стратегии, сочетающей активный и пассивный менеджмент. Умные бета-фонды используют пассивную стратегию, но модифицируют ее в соответствии с одним или несколькими факторами, такими как более дешевые акции или их фильтрация в соответствии с выплатами дивидендов, генерируя более высокие финансовые возвраты. Этот рост совпал с усилением критики по поводу высоких комиссионных, взимаемых традиционными активными менеджерами, а также усилением контроля за их работой.

Продолжающееся обнаружение и успешное прогнозирование рисков факторов, которые индивидуально либо в сочетании с другими рисковыми факторами значительно влияют на будущие возвраты от активов по всем классам активов, является ключевой движущей силой всплеска интереса к МО в инвестиционной индустрии и будет ключевой темой на протяжении всей этой книги.

---

коротких акций, т. е. с ожидаемым снижением цены. Прилагательные "длинный" и "короткий" обозначают направление движения цены соответственно вверх и вниз, не обязательно обозначая продолжительность. — *Прим. перев.*

<sup>24</sup> Факторное инвестирование (factor investing) — это стратегия, которая выбирает ценные бумаги по атрибутам, связанным с более высокой возвратностью. Существует два основных типа факторов, определяющих возвратность активов: макроэкономические факторы и стилевые факторы. Первые охватывают широкие риски по всем классам активов, в то время как вторые призваны объяснять возвратности и риски в рамках классов активов. Некоторые общие макроэкономические факторы включают кредит, инфляцию и ликвидность, в то время как стилевые факторы среди прочих охватывают стиль, стоимость и импульс. См. <https://www.investopedia.com/terms/f/factor-investing.asp>. — *Прим. перев.*

<sup>25</sup> Умное бета (smart beta) инвестирование сочетает в себе преимущества пассивного инвестирования и преимущества активных инвестиционных стратегий. Цель умного бета-инвестирования — получить альфа, снизить риск или увеличить диверсификацию за стоимость ниже, чем традиционный активный менеджмент, и слегка выше, чем прямое индексное инвестирование. Оно стремится отыскать наилучшую структуру оптимально диверсифицированного портфеля.

См. <https://www.investopedia.com/terms/s/smart-beta.asp>. — *Прим. перев.*

## Алгоритмические первопроходцы превосходят людей в широком масштабе

Репутация и рост *активов под управлением* (Assets Under Management, AUM) фирм<sup>26</sup>, которые возглавили алгоритмическую торговлю, сыграли ключевую роль в формировании интереса инвесторов и последующих отраслевых усилий по воспроизведению их успеха. Систематические фонды отличаются от высокочастотной торговли тем, что в отличие от преимуществ от чистой скорости сделки могут проводиться значительно дольше, стремясь использовать возможности арбитража.

Систематические стратегии, которые в основном или исключительно полагаются на алгоритмическое принятие решений, приобрели большую известность благодаря внедрившему их математику Джеймсу Саймонсу (James Simons), который в 1982 г. основал хедж-фонд Renaissance Technologies и построил ее в первую квантовую фирму. Ее скрытный фонд Medallion, который закрыт для посторонних, по оценкам, с 1982 г. зарабатывал финансовый возврат в размере 35% в годовом исчислении.

Три самых известных квантитативных хедж-фонда, D.E. Shaw, Citadel и Two Sigma, которые используют систематические стратегии, основанные на алгоритмах, в 2017 г. впервые добрались до 20 лучших исполнителей на рынке с точки зрения суммарного долларового дохода для инвесторов после сборов и с момента создания.

Квантитативный хедж-фонд D.E. Shaw, основанный в 1988 г., с 47 млрд долларов AUM в 2018 г. присоединился к списку под номером 3. Квантитативный хедж-фонд Citadel, запущенный в 1990 г. Кеннетом Грифффином (Kenneth Griffin), управляет 29 млрд долларов и занимает 5 место, и квантитативный хедж-фонд Two Sigma, запущенный лишь в 2001 г. бывшими сотрудниками D.E. Shaw Джоном Овердеком (John Overdeck) и Дэвидом Сигелом (David Siegel), вырос с 8 млрд долларов AUM в 2011 г. до 52 млрд долларов в 2018 г. Хедж-фонд Bridgewater, запущенный в 1975 г., с более чем 150 млрд долларов AUM продолжает лидировать благодаря своему фонду Pure Alpha Fund, который также встраивает систематические стратегии.

Аналогичным образом, в списке 100 лучших хедж-фондов журнала Institutional Investor за 2017 г. (<https://www.institutionalinvestor.com>) пять из шести ведущих фирм при принятии инвестиционных решений в значительной степени или полностью опираются на компьютеры и торговые алгоритмы — и все они увеличивают свои финансовые активы в условиях, которые в иных отношениях были бы сложными. Несколько квантитативно-ориентированных фирм поднялись на несколько ступеней и в некоторых случаях увеличили свои активы на двузначные проценты. Номер 2 в списке, глобальная компания по управлению инвестициями *Applied*

<sup>26</sup> Активы под управлением (assets under management, AUM) служат мерой общей рыночной стоимости всех финансовых активов, которыми финансовое учреждение, такое как взаимный фонд, венчурная компания или брокерский дом, управляет от имени своих клиентов и себя. — *Прим. перев.*

*Quantitative Research* (AQR) в 2017 г. увеличила свои хедж-фондовые активы на 48% до 69,7 млрд долларов и управляла 187,6 млрд долларов в масштабах всей фирмы.

Среди всех хедж-фондов, ранжированных по совокупной результативности за последние 3 года, квантовые фонды, управляемые компанией Renaissance Technologies, достигли позиций 6 и 24, хедж-фонд Two Sigma достиг позиции 11, хедж-фонд D. E. Shaw — позиций 18 и 32, хедж-фонд Citadel — позиций 30 и 37. Помимо лучших исполнителей на рынке, за последние несколько лет хорошо зарекомендовали себя алгоритмические стратегии. За последние 5 лет квантовые хедж-фонды росли примерно на 5,1% в год, в то время как за тот же период средний хедж-фонд вырастал на 4,3% в год.

## **Фонды, ведомые машинным обучением, привлекают 1 трлн долларов AUM**

Три общеизвестные революции в вычислительной мощности, данных и методах МО сделали принятие систематических, ведомых данными стратегий не только убедительнее и экономически эффективнее, но и ключевым источником конкурентного преимущества.

Вследствие этого алгоритмические подходы находят более широкое применение не только в хедж-фондовой индустрии, которая была первопроходцем этих стратегий, но и в более широком круге менеджеров финансовыми активами и даже в пассивно управляемых механизмах, таких как биржевые фонды ETF. В частности, предсказательная аналитика с использованием машинного обучения и алгоритмической автоматизации играет все более заметную роль на всех этапах инвестиционного процесса по всем классам финансовых активов, от генерирования идей и исследований до формулирования стратегии и конструирования портфеля, исполнения сделок и риск-менеджмента.

Оценочные размеры отрасли различаются по той причине, что нет объективного определения квантитативного или алгоритмического фонда, и многие традиционные хедж-фонды или даже взаимные фонды и биржевые фонды ETF внедряют компьютерные стратегии или интегрируют их в дискреционную среду в подходе "человек плюс машина".

По оценкам инвестиционного банка Morgan Stanley за 2017 г., за последние 6 лет алгоритмические стратегии росли на 15% в год и контролируют около 1,5 трлн долларов между хедж-фондами, взаимными фондами и умными биржевыми бета-фондами. Другие отчеты предполагают, что индустрия квантитативных хедж-фондов уже практически превысила 1 трлн долларов AUM, почти удвоив свой размер с 2010 г. на фоне оттока из традиционных хедж-фондов. В отличие от этого, согласно последнему глобальному докладу поставщика хедж-фондовых данных Hedge Fund Research (<https://www.hedgefundresearch.com/>), суммарный капитал хедж-фондовой индустрии достиг 3,21 трлн долларов.

По оценкам исследовательской фирмы Preqin, почти 1500 хедж-фондов совершают большинство сделок с помощью компьютерных моделей. Квантитативные хедж-фонды теперь ответственны за 27% всех сделок инвесторами с акциями США по сравнению с 14% в 2013 г. Но многие используют исследователей данных — или квантов<sup>27</sup> — которые, в свою очередь, используют машины для построения крупных статистических моделей<sup>28</sup>.

Вместе с тем в последние годы фонды перешли к истинному МО, где искусственно-интеллектуальные системы могут быстро анализировать большие объемы данных и самосовершенствоваться посредством такого анализа. Недавние примеры включают фирмы по инвестиционному менеджменту Rebellion Research, Sentient и Aidiya, которые опираются на эволюционные алгоритмы и глубокое обучение при разработке полноавтоматических инвестиционных платформ, приводимых в действие *искусственным интеллектом* (ИИ).

Из стержневой хедж-фондовой индустрии принятие алгоритмических стратегий распространилось на взаимные фонды и даже на пассивно управляемые биржевые фонды в форме умных бета-фондов и на дискреционные фонды в форме квантоментальных<sup>29</sup> подходов.

## Возникновение квантоментальных фондов

В активном инвестиционном менеджменте сформировалось два четких подхода: систематическое (или квантовое) инвестирование и дискреционное инвестирование. Систематические подходы опираются на алгоритмы повторяющегося и ведомого данными подхода по выявлению инвестиционных возможностей по многочисленным ценным бумагам; в отличие от этого дискреционный подход предусматривает углубленный анализ меньшего числа ценных бумаг. Эти два подхода становятся все более похожими, поскольку фундаментальные менеджеры используют все больше подходов, приводимых в действие наукой о данных.

Теперь даже фундаментальные торговцы вооружаются квантитативными методами, что, согласно инвестиционного банка Barclays, составляет 55 млрд долларов систематических финансовых активов. Нейтральные к конкретным компаниям, квантитативные фонды торгуют регулярностями и динамикой по широкому спектру ценных бумаг. Как показывают данные, собранные инвестиционным банком Barclays, на квантов теперь приходится около 17% от суммарного объема хедж-фондовых активов.

---

<sup>27</sup> Квант (quant) — это биржевой специалист по квантитативным методам.

См. <https://habr.com/company/iticapital/blog/307854/>. — *Прим. перев.*

<sup>28</sup> См. статью "The Quants Run Wall Street Now" ("Теперь на Уолл-стрит заправляют кванты") в Wall Street Journal. — <https://www.wsj.com/articles/the-quants-run-wall-street-now-1495389108>.

<sup>29</sup> Квантоментальный (quantamental) — неологизм, который образован из двух терминов "квантитативный подход" и "фундаментальный подход" и означает сочетание новейших квантитативных методов, в том числе на основе машинного обучения и ИИ, с классическими методами на основе фундаментальных величин. — *Прим. перев.*

Американский хедж-фонд Point72 Asset Management с финансовыми активами в размере 12 млрд долларов переводит около половины своих портфельных менеджеров на подход "человек плюс машина". Хедж-фонд Point72 также инвестирует десятки миллионов долларов в группу, которая анализирует большие объемы альтернативных данных и передает результаты биржевым торговцам.

## **Инвестиции в стратегический потенциал**

Рост инвестиций во взаимосвязанный потенциал — технологии, данные и, самое главное, квалифицированных людей — подчеркивает, насколько важной стала алгоритмическая торговля с использованием МО для конкурентного преимущества, в особенности в свете растущей популярности пассивных индексированных инвестиционных механизмов, таких как биржевые фонды ETF, после финансового кризиса 2008 г.

Инвестиционный банк Morgan Stanley отметил, что только 23% его квантовых клиентов говорят, что они не рассматривают возможность использования или уже не используют МО, по сравнению с 44% в 2016 г.

Например, компания Guggenheim Partners LLC построила в Национальной лаборатории Лоуренса Беркли в Калифорнии так называемый суперкомпьютерный кластер за 1 млн долларов с целью помочь ее инвестиционным фондам перемалывать числа. Электричество для компьютеров стоит еще 1 млн долларов в год.

Квантитативная инвестиционная группа AQR опирается на научные исследования с целью выявления и систематического учета торговых факторов, которые со временем доказали свое превосходство над более широким рынком. Указанная фирма привыкла избегать ведомых чисто компьютером стратегий квантовых коллег, таких как Renaissance Technologies или DE Shaw. Однако в последнее время AQR начала искать прибыльные регулярности на рынках, используя МО, занимаясь разбором и анализом новых наборов данных, таких как спутниковые снимки теней, отбрасываемых нефтяными скважинами и танкерами.

Ведущая фирма BlackRock, с более чем 5 трлн долларов AUM, также делает ставки на алгоритмы с целью победить менеджеров дискреционных фондов, обильно инвестируя в систематическую трейдинговую фирму SAE, которую она приобрела во время финансового кризиса. Американская компания по менеджменту финансовыми активами Franklin Templeton приобрела Random Forest Capital — инвестиционную компанию, ориентированную на долги и управляемую данными — за нераскрываемую сумму, надеясь, что ее технология сможет поддержать более широкого менеджера финансовыми активами.

## **Машинное обучение и альтернативные данные**

Хедж-фонды давно занимались поисками альфа через информационное преимущество и способностей обнаруживать новые некоррелированные сигналы. Исторически сюда входили такие вещи, как проприетарные опросы покупателей или избира-

телей перед выборами или референдумами. Время от времени использование корпоративных инсайдеров, диагностов и экспертных сетей с целью расширения знаний об отраслевых трендах или компаниях пересекает правовые грани: после 2010 г. индустрию потрясла серия судебных преследований биржевых торговцев, портфельных менеджеров и аналитиков за использование инсайдерской информации.

В отличие от этого, информационное преимущество от использования традиционных и альтернативных источников данных с помощью МО связано не с экспертными и отраслевыми сетями или доступом к корпоративному управлению, а скорее с возможностью сбора крупных объемов данных и их реально-временного анализа.

Революция в использовании данных в стратегиях алгоритмической торговли была совершена тремя трендами, которые способны еще больше сдвинуть инвестиционную отрасль от дискреционных стилей в сторону количественных:

- ◆ экспоненциальное увеличение объема цифровых данных;
- ◆ увеличение вычислительной мощности и емкости хранения данных по более низкой цене;
- ◆ достижения в методах МО для анализа многосложных наборов данных.

Обычные данные охватывают экономическую статистику, торговые данные или корпоративные отчеты. Альтернативные данные намного шире и охватывают такие источники, как спутниковые снимки, кредитно-карточные продажи, сентиментный анализ, данные мобильной геолокации и выискивание веб-сайтов, а также преобразование данных, генерируемых в результате обычного протекания бизнеса, в ценную развединформацию. Они охватывают, в принципе, любой источник данных, содержащий торговые сигналы, которые можно извлечь с помощью МО.

Например, данные страховой компании о продажах новых полисов автострахования не только служат индикатором объемов продаж новых автомобилей, но и могут быть разбиты на бренды или географии. Многие поставщики выискивают веб-сайты в поисках ценных данных, начиная со скачивания приложений и отзывов пользователей до бронирования билетов авиакомпаний и номеров в гостиницах. Социально-медийные сайты также могут выискиваться в поисках подсказок о мнениях потребителей и трендах.

Как правило, наборы данных являются крупными и требуют хранения, доступа и анализа с использованием масштабируемых решений для параллельной обработки данных, таких как Hadoop и Spark; согласно финансовому конгломерату Германии Deutsche Bank, существует более 1 млрд веб-сайтов с более чем 10 трлн разных веб-страниц с 500 экзабайтами (или 500 млрд гигабайт) данных. И более 100 млн веб-сайтов добавляется в Интернет с каждым годом.

Проникновение в суть перспектив компании в режиме реального времени, задолго до того, как ее результаты будут опубликованы, может быть получено из снижения списков вакансий на ее веб-сайте, внутреннего рейтинга ее руководителя сотрудниками на сайте найма Glassdoor или падения средней цены одежды на ее веб-сайте.