Оглавление

От научного редактора русского издания 10

Предисловие 11

Мнение автора о ПО и процессах для построения графиков 13

Условные обозначения 14

Использование примеров кода 15

Благодарности 15

Введение 17

Некрасивые, плохие и ложные изображения 18

ЧАСТЬ I
ОТ ДАННЫХ
ДО ВИЗУАЛИЗАЦИИ

Глава 1. Визуализация данных: соответствие данных и эстетики 22

Эстетика и типы данных 22

Использование шкал для отображения данных на эстетические элементы 25

Глава 2. Оси и системы координат 29

Прямоугольная (декартова) система координат 29

Нелинейные оси 32

Системы координат с изогнутыми осями 38

Глава 3. Цветовые шкалы 41

Цвет как средство различения 41

Цвет как средство представления значений данных 43

Цвет как средство выделения данных 46

Глава 4. Каталог визуализаций 49

Количественные диаграммы 49

Диаграммы распределения 50

Пропорциональные диаграммы 51

Диаграммы двух переменных 52

Геопространственные диаграммы 54

Неопределенность на диаграммах 54

Глава 5. Визуализация количественных данных 56
Столбчатые диаграммы 56
Столбчатые диаграммы с группировкой и накоплением 61
Точечные графики и тепловые карты 64

Глава 6. Визуализация распределений: гистограммы и графики плотности 69
Визуализация одного распределения (Single Distribution) 69
Визуализация нескольких распределений одновременно 75

Глава 7. Визуализация распределений: функции распределения и графики «квантиль-квантиль» 79
Функции распределения 79
Сильно искаженные распределения 82
Графики «квантиль-квантиль» 86

Глава 8. Одновременная визуализация множества распределений 88
Визуализация распределений вдоль вертикальной оси 88
Визуализация распределений на горизонтальной оси 95

Глава 9. Визуализация пропорций 99
Время круговых диаграмм! 99
Пример в пользу столбчатых диаграмм 102
Пример в пользу столбчатых диаграмм и графиков плотности с наложением 104
Визуализация пропорций по отдельности как частей целого 106

Глава 10. Визуализация пропорций на нескольких уровнях 109
Как не надо строить многоуровневые пропорции 109
Мозаичные графики и древовидные карты 111
Многоуровневые круговые диаграммы 115
Диаграммы в параллельных координатах 117

Глава 11. Визуализация связей между двумя и более количественными переменными 120
Диаграммы рассеяния 120
Коррелограммы 124
Снижение размерности 127
Парные выборки 130

Глава 12. Визуализация временных рядов и других функций независимой переменной 134
Самостоятельные временные ряды 134
Множественные временные ряды и кривые «доза — эффект» 137
Временной ряд двух или более объясняемых переменных 140
Глава 13. Визуализация трендов 146
Сглаживание 146
Подгонка трендов при помощи заданных функциональных форм 152
Удаление трендов и декомпозиция временных рядов 156
Глава 14. Визуализация геопространственных данных 162
Проекции 162
Слои 169
Фоновые картограммы 172
Картограммы 176
Глава 15. Визуализация неопределенности 179
«Кадрирование» вероятностей в виде частот 179
Визуализация неопределенности точечной оценки 184
Визуализация неопределенности подгонки кривых 197
Диаграммы гипотетических исходов 200

ЧАСТЬ II
ПРИНЦИПЫ ДИЗАЙНА ВИЗУАЛИЗАЦИЙ

Глава 16. Принцип пропорциональной заливки 204
Визуализации на линейных шкалах 204
Визуализации на логарифмических шкалах 209
Прямая визуализация площадей 212
Глава 17. Обработка накладывающихся точек 215
Частичная прозрачность и джиттеринг (jittering) 215
Двухмерные гистограммы 219
Изолинии 221

Глава 18. Распространенные ошибки при использовании цвета 227
Отображаем слишком много или ненужную информацию 227
Использование немонотонных цветовых шкал для передачи значений данных 231
Игнорирование потребностей людей с нарушениями цветового зрения 232

Глава 19. Избыточная передача данных 237
Проектирование легенд с применением принципа избыточной передачи данных 237
Проектирование визуализаций без легенды 242

Глава 20. Многопанельные визуализации 247
Малые панельные визуализации 247
Составные визуализации 252

Глава 21. Заголовки, подписи и таблицы 258
Заголовки и подписи к рисункам 258
Названия осей и легенд 260
Таблицы 264

Глава 22. Баланс данных и контекста 267
Предоставление подобающего объема контекста 267
Фоновые сетки 272
Парные данные 277
Вывод 279

Глава 23. Подписи осей должны быть крупными 281

Глава 24. Избегайте лишних линий 286

Глава 25. Не используйте 3D 293
Избегайте неоправданного применения 3D 293
Не используйте трехмерную систему координат 295
Когда трехмерные визуализации уместны 301

ЧАСТЬ III
РАЗНОЕ

Глава 26. Наиболее распространенные форматы файлов изображений 304
Растровая и векторная графика 304
Сжатие растровой графики с потерями и без 306
Преобразования между форматами изображений 309

Глава 27. Как выбрать подходящее программное обеспечение для визуализации ... 310
Воспроизводимость и повторимость ... 311
Исследование данных и представление данных ... 313
Разделение содержания и дизайна ... 315

Глава 28. Как рассказать историю и донести свою мысль ... 318
Что такое история? ... 319
Создавайте визуализации «для генералов» ... 322
Постепенный переход к сложным визуализациям ... 326
Визуализации должны быть запоминающимися ... 328
Будьте последовательны, но не повторяйтесь ... 330

Аннотированный список литературы ... 335
Размышления о данных и их визуализации ... 335
Книги по программированию ... 336
Тексты по статистике ... 336
Исторические тексты ... 337
Книги по смежной тематике ... 338

Технические примечания ... 339

Примечания ... 341

Предметный указатель ... 344

Об авторе ... 349

Об изображении на обложке ... 350

От научного редактора русского издания

Перед вами прекрасная книга, которая в сжатой и очень доступной форме рассказывает о том, как можно подавать и визуализировать самые разноплановые данные. Книга хороша тем, что написана профессионалом своего дела и человеком, который не просто собрал несколько статей, а по-настоящему уже много лет использует визуализацию данных в своей работе.

Оригинал книги написан на английском языке и содержит очень большое количество названий различных диаграмм, графиков и других форм визуализации, у части которых нет эквивалентов в русском языке, а часть может в различных изданиях и приложениях быть переведена по-другому.

Мы старались пользоваться как можно более широкой базой для того, чтобы максимально качественно перевести все названия, но заранее приносим извинения, если вы привыкли к другим названиям форм представления и визуализации данных, нежели к тому, как они переведены в этой книге.

Андрей Бояршинов,
Director of Production, General Arcade

Предисловие

Если вы ученый, аналитик, консультант или любой другой специалист, чьи обязанности включают в себя подготовку технических документов или отчетов, то вы не понаслышке знаете, как важно уметь убедительно визуализировать данные в формате изображений. Чем нагляднее графики, тем весомее смотрятся аргументы. Изображения должны быть ясными, привлекательными и убедительными. Разница между хорошим и плохим графиком подобна разнице между влиятельной и малоизвестной газетой, выигранными и упущенными грантами или контрактами, удачным и провальным собеседованием. При этом, несмотря на, казалось бы, очевидную востребованность методических материалов по этой теме, существует на удивление мало ресурсов, которые рассказывают о том, как качественно и красиво иллюстрировать данные. Высших учебных заведений, которые предлагают курсы по этой теме, совсем немного, да и специализированную литературу придется поискать. (Но что-то, разумеется, есть.) Учебные материалы, посвященные программным средствам построения графиков, обычно рассказывают лишь о создании некоторых визуальных эффектов, однако объяснений, почему следует выбрать тот или иной вариант, не дается. При этом в повседневной рабочей рутине предполагается, что вы знаете, как создавать хорошие графики. А если вам повезет, у вас появится терпеливый научный руководитель, который научит нескольким приемам визуализации при написании ваших первых научных статей.

Что касается писательской работы, опытные редакторы говорят о так называемом «слухе» — способности слышать (внутренне, когда вы читаете отрывок прозы), хорош ли этот текст. Думаю, что для графиков и других визуализаций нам тоже нужно «зрение», то есть способность смотреть на картинку и видеть, является ли этот график сбалансированным, ясным и убедительным. Как и в случае с текстом, умению отличать эффективный график от нерабочего можно научиться. Наличие «зрения» — это прежде всего знакомство с более широким спектром простых законов и принципов хорошей визуализации, а также внимание к мелким деталям, зачастую упускаемым другими людьми.

По своему опыту я знаю, что, как и в случае с текстом, невозможно развить «зрение», всего лишь прочитав одну книжку на выходных. Это длительный процесс, и он будет с вами всю вашу жизнь, и может случиться так, что концепции, которые сейчас вам кажутся слишком сложными или, наоборот, малозначимыми, через пять лет станут для вас куда более существенными. Что касается меня, то я и сегодня продолжаю совершенствоваться в искусстве создания графиков. Я постоянно ищу новые подходы и обращаю внимание на визуальные и дизайнерские решения других людей. Не исключаю, что буду менять свое мнение, если появятся убедительные аргументы в пользу иной точки зрения. Сегодня я могу считать один график отличным, а уже через месяц у меня появится повод его покритиковать. Помните об этом, читая книгу, и не принимайте все мои слова за истину в последней инстанции. Будьте критичны к моим аргументам в пользу определенных решений и выбирайте сами, принимать их или нет.

Материал книги выстроен в логической последовательности, однако большинство глав можно читать как самостоятельный текст: штудировать книгу от корки до корки вовсе не обязательно. Не стесняйтесь пропускать шаги, чтобы выбрать наиболее интересный для вас в данный момент раздел или тот, который посвящен тому типу дизайна, над которым вы сейчас размышляете. Более того, думаю, что вы извлечете из этой книги максимум пользы только в том случае, если будете читать ее не всю целиком за раз, а по частям и в течение длительного периода времени. Попробуйте применить парочку-другую концепций из книги, а потом вернитесь к ней, чтобы узнать о других принципах или освежить в памяти уже изученные разделы. Вполне возможно, что одна и та же глава предстанет перед вами совсем в другом свете, если вы вновь вернетесь к ней через несколько месяцев.

Несмотря на то, что почти все рисунки в этой книге сделаны с помощью R и ggplot2, я не считаю эту книгу справочником по созданию графиков при помощи R. Моя цель — рассказать об общих принципах создания графиков. Выбор программного обеспечения, используемого для создания изображений, зависит от ваших конкретных потребностей. Если вы захотите воспроизвести визуализации, приведенные в этой книге, то можете использовать любое ПО для построения графиков. Тем не менее хочу отметить, что многое из того, что я демонстрирую, сделать посредством ggplot2 и аналогичных пакетов будет гораздо проще, чем с помощью других библиотек для построения графиков. Важно отметить, что, поскольку данная книга не справочник по созданию графиков при помощи R, вы не найдете здесь обсуждений кода или методов программирования. Я хочу, чтобы вы сосредоточились на концепциях и графиках, а не на их реализациях. Если вам интересно, как были сделаны те или иные рисунки, обратитесь к исходному коду по ссылке на с. 15.

Мнение автора о ПО и процессах для построения графиков

За моими плечами больше двух десятилетий опыта подготовки графиков для научных публикаций, я являюсь автором тысяч рисунков. Если в течение этого времени что-то и оставалось незыблемым, так это постоянные изменения рабочего процесса их подготовки. Каждые несколько лет появляется новая библиотека для построения графиков или даже новая парадигма, после чего огромное количество ученых переключается на более актуальный инструментарий. Я делал рисунки с помощью gnuplot, Xfig, Mathematica, Matlab, matplotlib в Python, base R, ggplot2 в R и, возможно, других инструментов, названия которых я уже и не вспомню. В настоящее время я предпочитаю подход ggplot2 в R, но не ожидаю, что он дотянет до моей пенсии.

Постоянное изменение программных платформ — одна из основных причин того, что данная книга не является учебником по программированию и что в ней нет ни единого примера кода. Я хочу, чтобы эта книга была полезной вне зависимости от того, какое ПО вы используете, и хочу, чтобы она оставалась таковой даже после того, как все уйдут с ggplot2 и перейдут к следующему поколению программ визуализации данных. Я понимаю, что выбранный мною подход наверняка расстроит некоторых пользователей ggplot2, которые хотели бы знать, как я сделал тот или иной рисунок, и поэтому те, кому интересно узнать о моих методах программирования, могут обратиться к исходному коду книги: он находится в открытом доступе. Кроме того, вероятно, в будущем появится дополнительный материал, посвященный исключительно вопросам программирования.

За прошедшие годы я понял одну вещь: автоматизация — ваш друг. Мое мнение — графики должны генерироваться автоматически как часть процесса анализа данных (который также должен быть автоматизирован), и они должны в результате создаваться уже готовыми к отправке на принтер, без необходимости ручной доработки. Я часто вижу, как стажеры сначала делают грубый набросок будущего графика, а затем импортируют его в Illustrator, чтобы привести в порядок. Эта идея плоха в силу нескольких причин. Во-первых, если вы вручную редактируете график, конечный результат становится невоспроизводимым. Никто другой не сможет сгенерировать график, идентичный созданному вами. Даже если вы всего лишь изменили шрифт подписей к засечкам осей, линии могут получиться размытыми, и уже одно это может нанести ущерб информативности изображения. Например, вы решили вручную заменить некоторые непонятные метки на более читаемые — другой человек вряд ли сможет проверить правильность этой замены. Во-вторых, если

вы добавите много ручной постобработки в процесс подготовки графиков, вы будете менее охотно вносить какие-либо изменения в свою работу или переделывать ее. В результате может случиться так, что вы просто проигнорируете разумные предложения ваших соавторов или коллег об изменении графиков или у вас может возникнуть соблазн повторно использовать старую визуализацию, даже если данные уже обновились. В-третьих, вы можете попросту забыть, что именно вы делали, и не сможете создать аналогичный график, но с другим наполнением. Все эти примеры взяты из жизни: я лично видел, как такое происходило с реальными людьми и настоящими публикациями.

Поэтому использование интерактивных программ по созданию графиков — плохая идея. По сути, они заставляют вас создавать графики вручную. На самом деле, вероятно, лучше автоматически создать эскиз фигуры и украсить его в Illustrator, чем создавать всю фигуру вручную в какой-нибудь интерактивной программе. Имейте в виду, что Excel тоже является интерактивной программой построения графиков и не рекомендуется для подготовки рисунков (или анализа данных).

Одним из важнейших компонентов книги по визуализации данных является возможность воспроизведения предлагаемых графиков. Приятно изобрести какой-нибудь новый изящный тип визуализации, но, если воссоздать график, используя ваш метод, невозможно, пользы от вашей идеи будет немного. Например, когда Эдвард Тафти предложил так называемые спарклайны, то первое время никто толком не понимал, как их делать. И хотя нам безусловно нужны гении, которые заставляют мир двигаться вперед, расширяя границы возможного, эта книга посвящена практике и может использоваться в повседневной работе специалистов по данным, готовящих графики для своих публикаций. Поэтому предлагаемые мной визуализации могут быть созданы с помощью нескольких строк кода R, ggplot2 и легкодоступных пакетов расширений. Почти все изображения в этой книге, за исключением тех, что находятся в главах 26–28, были автоматически сгенерированы именно в том виде, в каком они здесь представлены.

Условные обозначения

В этой книге используются следующие условные обозначения.

Курсивный шрифт

Обозначает новые термины, имена файлов и их расширения.

`Моноширинный шрифт`

Используется для обозначения элементов кода, таких как имена переменных или функций, операторов и ключевых слов.

Совет или подсказка

Общее замечание

Предостережение

Использование примеров кода

Дополнительные материалы можно скачать по следующему адресу: *https://addons.eksmo.ru/it/Data_Visualization.zip*.

Цель этого руководства — помочь вам в решении ваших задач. В работе над своими программами или документацией вы можете пользоваться фрагментами кода из данной книги.

Благодарности

Появление этого проекта на свет было бы невозможно без той фантастической работы, которую проделала команда RStudio, превратив вселенную R в первоклассную платформу для подготовки оригинал-макетов. В частности, я должен поблагодарить Хэдли Уикхэма за создание ggplot2 — программы для построения графиков, — которая использовалась для создания всех изображений в этой книге. Я также хотел бы поблагодарить Се Ихуэй за создание R Markdown и за написание пакетов `knitr` и `bookdown`. Не думаю, что я решился бы взяться за настоящий проект, не будь у меня под рукой этих инструментов. Писать R Markdown-файлы — одно удовольствие, собирать материал и наращивать темп работы — легче легкого. Особую благодарность я хочу выразить Ахиму Зейлеису и Рето Штауфферу за `colorspace`, Томасу Лину Педерсену за `ggforce` и `gganimate`, Камилу Словиковски за `ggrepel`, Эдзеру Пебесме за `sf` и Клэр Маквайт за ее работу над пакетами `colorspace` и `colorblindr` для имитации дальтонизма при просмотре готовых иллюстраций.

Отдельно хочу поблагодарить людей, которые предоставили полезные отзывы о черновых версиях этой книги. Наибольший вклад внесли Майк Лукидес, мой редактор в O'Reilly, и Стив Хароз — они прочитали и прокомментировали каждую главу. Я также получил полезные комментарии от Карла Бергстрома, Джессики Халлман, Мэтью Кея, Тристана Мара, Эдзера Пебесмы, Джона Швабиша и Хэдли Уикхэма. Блог Лена Кифера, а также книги и сообщения Кирана Хили послужили источником вдохновения для создания графиков и наборов данных для использования. Ряд людей указали на незначительные проблемы или опечатки: Тьяго Аррайс, Малкольм Барретт, Джессика Бернетт, Джон Колдер, Антонио Педро Камарго, Дарен Кард, Ким Крессман, Акос Хайду, Томас Йохманн, Эндрю Кинсман, Уилл Керсен, Алекс Лаледжини, Джон Лидли, Катрин Лайнвебер, Микель Мадина, Клэр Маквайт, С'бусисо Мхондване, Хосе Назарио, Стив Путман, Маэль Салмон, Кристиан Шудома, Джеймс Скотт-Браун, Энрико Спиниелли, Воутер ван дер Бейл и Рон Юрко.

Я также хотел бы поблагодарить всех остальных участников tidyverse и R-сообщества в целом. Действительно, для любой задачи по визуализации, с которой вы можете столкнуться, существует R-пакет. Все эти приложения были разработаны силами большого сообщества, состоящего из тысяч специалистов по обработке данных и статистике, и многие из них в той или иной форме внесли свой вклад в создание этой книги.

Наконец, я хотел бы поблагодарить мою жену Стефанию за ее терпение в течение несметного количества вечеров и выходных, когда я часами сидел перед компьютером, писал код `ggplot2` и был полностью погружен в мельчайшие детали графиков и проработку деталей глав.

Введение

Визуализация данных — это отчасти наука, а отчасти искусство. Самое сложное в этом деле — сделать так, чтобы искусство получилось хорошим, при этом не переврав науку, и наоборот. Визуализация данных — это прежде всего точная передача информации. Недопустимо даже малейшее искажение данных. Если вы заметили, что одно число в два раза больше другого, но при этом на схеме соответствующие этим числам элементы имеют одинаковый размер, знайте: вся работа по визуализации пошла под откос. В то же время нельзя забывать, что визуализация должна быть приятна глазу. Качественное визуальное преподнесение данных обычно повышает их информативность. Если у зрителя рябит в глазах от ярких цветов, элементы графика несбалансированы или на нем множество отвлекающих внимание объектов, зрителю будет сложнее рассмотреть изображение и верно считать его смысл.

Исходя из своего опыта, хочу отметить, что в большинстве случаев (но не всегда!) ученые умеют визуализировать данные, не вводя читателей в заблуждение. К сожалению, они не всегда обладают развитым художественным вкусом и периодически используют вещи, сильно отвлекающие внимание от передаваемой информации. Дизайнеры, в свою очередь, делают все невероятно красиво, но при этом крайне небрежно относятся к самим данным. Посыл моей книги состоит в том, чтобы донести до каждой из этих групп полезную для них информацию.

В этой книге я концентрируюсь на базовых принципах, методах и концепциях, которые применяются при иллюстрации публикаций, докладов или презентаций. Поскольку визуализация данных — обширное поле, которое в широчайшем толковании может включать в себя такие разнообразные вещи, как схемотехника, 3D-анимации и пользовательские интерфейсы, я вынужден сузить спектр. Данная книга посвящена исключительно способам визуализации, представленным в печатном виде, онлайн или в виде слайдов. Здесь вы не встретите информацию об интерактивных визуальных элементах или видео, кроме как в небольшом разделе главы 25. Поэтому в рамках данной книги я буду достаточно вольно использовать слова «визуализация» и «изображение», имея в виду одно и то же. Помимо перечисленного, в этой книге не идет речь о том, *как* создавать изображения при помощи существующих

программ для визуализации и программных библиотек. Библиография в конце книги содержит ссылки на подходящие для изучения этих тем статьи.

Данная книга состоит из трех частей. Первая — «От данных до визуализации» — описывает различные виды графиков и диаграмм, такие как гистограммы, диаграммы рассеяния и круговые диаграммы. Основной упор в этой части делается на научный аспект визуализации. Однако вместо того, чтобы писать обширную энциклопедию со статьями, посвященными каждому мыслимому подходу к визуализации, я расскажу вам о тех способах визуализации, которые вы наверняка встретите в публикациях или будете использовать в своих работах. При написании данной части я постарался сгруппировать подходы к визуализации по тому, какой посыл они несут зрителям, а не по типам используемых данных. Учебники статистики обычно описывают анализ и визуализацию данных в привязке к типам данных и организуют материал по количеству и типу переменных (одна непрерывная переменная, дискретная переменная, две непрерывные переменные, одна непрерывная и одна дискретная переменная и т. д.). Мое мнение на этот счет состоит в том, что найти в таких текстах что-то полезное смогут только ученые-статистики. Большинство людей воспринимает данные через призму информативной составляющей сообщения: например, насколько что-то велико или мало, каковы его составные части, как оно соотносится с чем-то другим и т. д.

Вторая часть, «Принципы дизайна визуализаций», посвящена вопросам дизайна, которые зачастую возникают в процессе создания схем или диаграмм. Основная, но не единственная тема, поднимаемая в данной главе, — эстетический аспект визуализации данных. После того как мы выбрали подходящий для наших данных график или диаграмму, перед нами встает проблема выбора визуальных элементов: цвет, символы, размеры и гарнитуры шрифтов. От выбора этих элементов зависит то, насколько наш способ визуализации будет понятен и эстетичен. Из второй части вы узнаете о наиболее частых проблемах, с которыми мне приходится сталкиваться во время непосредственной работы над визуализацией.

Третья часть, «Разное», посвящена темам, которые не вошли ни в первую, ни во вторую часть. Сюда относятся, например, форматы файлов, в которых обычно хранятся рисунки и графики, а также подсказки по выбору ПО. Кроме того, эта глава рассказывает о том, как правильно встраивать диаграммы в большие документы с учетом контекста.

Некрасивые, плохие и ложные изображения

На протяжении этой книги вам будут встречаться разные варианты одних и тех же изображений. Часть из них будет использоваться в качестве примеров того, как надо и не надо делать. Чтобы вам было понятнее, какие примеры