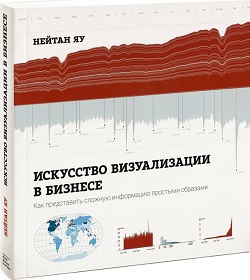
**Нейтан Яу. Искусство визуализации в бизнесе**

С тех пор, как я перешел из издательского бизнеса в торговую компанию (а было это в 2005 г.), Excel – является моим основным инструментом работы. Конечно, я далек от мысли, что владею им в совершенстве, но всё же думаю, что могу себя назвать экспертом. Так вот, эта книга не об Excel'е… Автор вскользь упоминает о нем, но, скорее, в уничижительном контексте, как о программе для начинающих. Серьезная инфографика и иная визуализация данных выполняются в других программах. Автор начинает с легкого 10-страничного знакомства с тем, как извлекать данные из Интернета на основе кода, написанного на языке Python. В умелых руках обилие информации – не проблема, а настоящий кладезь. Ведь интересны не сами цифры, а та информация, которую можно из них извлечь. Люди хотят знать, о чем именно говорят все эти таблицы и графики. Книга покажет, с помощью какого ПО удобнее обрабатывать информацию в каждом из случаев. Рассмотрены примеры визуализации в программах R и Adobe Illustrator, HTML, CSS и JavaScript. Также в книге вы познакомитесь с применением Flash-графики для нужд визуализации, научитесь создавать карты с помощью Python и SVG и анимированные карты в ActionScript и Flash. Книга – реальный учебник для тех, кто никогда не работал в упомянутых программах.

Нейтан Яу. Искусство визуализации в бизнесе. – М.: Манн, Иванов и Фербер, 2013. — 352 с.



На момент публикации заметки доступна только цифровая версия книги; купить в [ЛитРес](http://www.litres.ru/neytan-yau/iskusstvo-vizualizacii-v-biznese-kak-predstavit-slozhnuu-informaciu-prostymi-obrazami-2/?lfrom=13042861)

**Введение.** Один из лучших способов исследовать крупную базу данных и попытаться разобраться в ней — это визуализация. Поместите числа в видимое пространство и предоставьте мозгу — своему или ваших читателей — выявить паттерны. В этом деле мы все мастера. Вы сможете разглядеть истории, которые, возможно, никогда бы не увидели, применяя лишь формальные статистические методы. Если вы за всю жизнь не написали ни строчки кода, тогда R — вычислительная среда, которой отдают предпочтение многие статистики, — может показаться вам пугающей. Но после того как вы изучите несколько примеров, вы быстро набьете руку. Вы можете также скачать коды и файлы данных и опробовать работающие демоверсии на сайте [издательства](http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470944889.html) и [блоге](http://book.flowingdata.com/) автора (на момент публикации заметки файлы в блоге не доступны).

### Глава 1. Как рассказать историю с помощью данных

Великий смысл статистики и визуализации — в том, что они помогают увидеть, что именно стоит за данными. Помните: данные есть отражение реальной жизни. Это не просто груда каких-то чисел. В этой груде содержится множество историй. В полной мере это дошло до меня во время стажировки в газете New York Times. Длилась она всего три летних месяца после второго курса магистратуры, но изменила мой подход к данным навсегда. Я не только научился создавать графику для новостей. Я научился преподносить данные как новость, а вместе с тем узнал многое о дизайне, организации и проверке фактов, а также об их поиске и изучении. Некоторые из лучших диаграмм и графиков New York Times можно посмотреть на [здесь](http://flowingdata.com/?s=nytimes).

Когда материалы собраны, подходите к визуализации так, как будто вы собираетесь рассказать историю. Какого рода историю вы хотите поведать? Это будет репортаж? Или рассказ о жизни? Вы хотите убедить людей в необходимости каких-то действий? Чему бы ни была посвящена ваша графика, вам следует всегда искать **паттерны** и **зависимости**. В качестве простого примера можно привести данные о подключении к Интернету, взятую из [Доклада о мировом прогрессе](http://flowingprints.com/print4.php) (рис. 1).

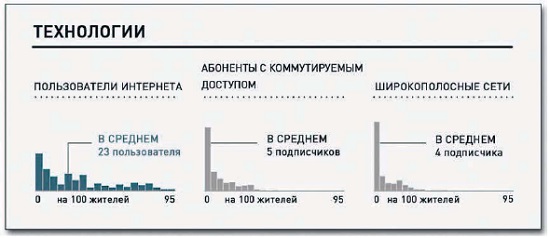


Рис. 1. Подключение к Интернету

Проще и быстрее всего — предоставить компьютерной программе, которая у вас установлена, самой решить, какой диапазон использовать для каждой диаграммы. Однако все диаграммы здесь выстроены на базе одного и того же диапазона, несмотря на то что нет ни одной страны, у которой имелось бы по 95% абонентов с коммутируемым доступом или пользователей широкополосных сетей. Но единый масштаб дает вам возможность легко сравнивать распределения между группами. А потому, когда вам приходится работать с множеством различных наборов данных, постарайтесь думать о них как о группе, а не как об отдельных, никак не связанных между собой массивах. Так вы сможете получить гораздо более интересные результаты.

*Сомнительные данные.* Пытаясь разглядеть истории, скрывающиеся за вашими данными, стоит всегда подвергать сомнению то, что вы видите. Помните: то, что это — числа, еще не означает, что они верны. Проверка и верификация данных — одна из важнейших, если не самая важная, часть графического дизайна.

*Дизайн.* При разработке дизайна вам соблюдать несколько основных принципов:

* объясните кодировки,
* дайте осям названия (а также укажите, какая именно шкала применяется в каждом конкретном случае: логарифмическая, дифференциальная, экспоненциальная…),
* укажите источники,
* учитывайте аудиторию (диаграмма, создаваемая для слайда презентации, должна быть простой; вы, конечно, можете включить в нее массу деталей, но их увидят только те, кто сидит в первом ряду; а вот если вы делаете постер, который люди будут пристально разглядывать и изучать, тогда вы можете включить в диаграмму гораздо больше подробностей),
* держите геометрию под контролем.

Распространенная ошибка обычно происходит тогда, когда дизайнеры используют для обозначения величин двухмерные формы, но измеряют их так, как если бы они использовали только одно измерение. На рис. 2а показаны два круга, размер которых задавался по площади. Это правильный подход. На рис. 2б представлена пара кругов, чьи размеры определялись по диаметру. Первый круг диаметром в два раза меньше второго, но по площади он меньше в четыре раза.

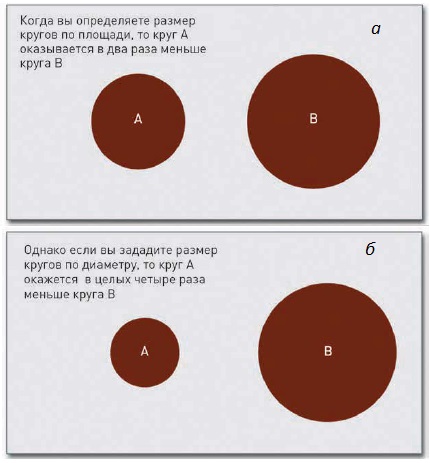


Рис. 2. Определение размеров кругов в диаграммах: а) правильный подход, б) неправильный подход

### Глава 2. Как обращаться с данными

Вам нужно научиться добывать и перерабатывать данные, и тогда ваши способности в области визуализации существенно возрастут. Получив данные, вам необходимо изучить контекст. Не нужно становиться экспертом в той области, из которой были почерпнуты данные, однако вам следует знать, из какого именно первоисточника они взяты и о чем говорят. Это поможет вам создать более качественную графику и поведать через нее более цельную историю.

При поиске данных, помимо Google, я время от времени обращаюсь и к [WolframAlpha](http://www.wolframalpha.com/), поисковой машине с вычислительными возможностями. Эта поисковая машина может оказаться особенно полезной при поиске базовых статистических данных по разным темам. Данные также можно поискать на сайтах университетов, например, [Калифорнийского университета](http://statistics.ucla.edu/) в Лос-Анджелесе (данные, которые кафедра статистики университета использует при выполнении лабораторных работ). Если вас интересует география, обратитесь на [Flickr ShapefiLes](https://www.flickr.com/services/api/) (здесь вы найдете географические границы), спорт – [Basketball Reference](http://www.basketball-reference.com/) (источник исчерпывающих данных и сведений вплоть до прямых репортажей с игр НБА), мир в целом – [UNdata](http://data.un.org/) (агрегатор данных из разных источников со всего мира), государственные и политические организации США – [Data.gov](http://www.data.gov/) (каталог данных, поставляемых органами правительства). Вы также можете научиться самостоятельно извлекать данные с web-сайтов, например, с помощью кода на языке Python (см. [Извлечение данных с web-страниц с помощью кода на языке Python](http://baguzin.ru/wp/?p=14390)).

*Форматы данных.* Большинство людей привыкли работать с данными в Excel. И в этом нет ничего плохого, если вы будете делать в данной программе все — от анализа до визуализации. Но если вам понадобится выйти за эти рамки, тогда вам придется ознакомиться и с другими форматами данных: текст с разделителями, объектная нотация JavaScript или расширяемый язык разметки.

Текст с разделителями-запятыми имеет очень широкое применение и читается большинством программ с таблицами, таких как Excel или Google Documents. Электронные таблицы вы также можете экспортировать в виде текста с разделителями.

Объектная нотация JavaScript ([JSON](http://json.org/)) – это один из самых распространенных форматов, который предлагают веб-API. JSON работает с ключевыми словами и значениями и относится к ним как к объектам. Если вы конвертируете данные в формате JSON в значения, разделенные запятыми (CVS), каждый объект станет строчкой. Если вы собираетесь создавать диаграммы и графики для Сети, вы наверняка столкнетесь с этим форматом.

Расширяемый язык разметки (XML) применяют для передачи данных через различные API. Его можно охарактеризовать как формат текстового документа со значениями, заключенными в теги.

*Инструменты форматирования.* Google Refine (он же [OpenRefine](http://openrefine.org/)) запускается через браузер, а вся обработка данных происходит у вас на компьютере. Когда вы откроете Refine, вы увидите знакомый табличный интерфейс со строчками и столбцами. В нем можете сортировать данные по полям и искать определенные величины. А еще можно относительно быстро находить ошибки.

Опробуйте также [Mr. Data Converter](https://shancarter.github.io/mr-data-converter/), чтобы конвертировать свои таблицы Excel в формат, удобный для размещения в Сети. Скопируйте данные из Excel и вставить их в область ввода в верхней части экрана, а внизу укажите, в каком формате вы хотите получить их (рис. 3).

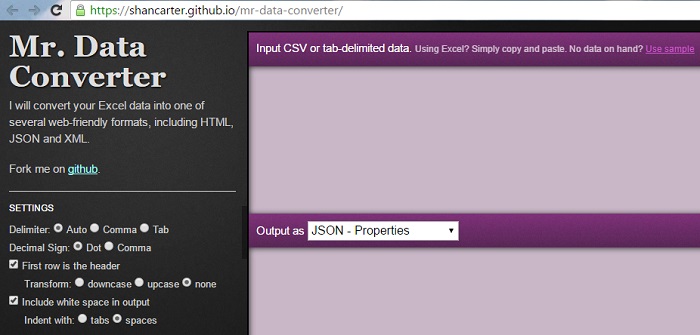


Рис. 3. Интерфейс Mr. Data Converter

А еще вы можете отформатировать ваши данные с помощью кода (см. [Форматирование данных с помощью кода на языке Python](http://baguzin.ru/wp/?p=14445)).

### Глава 3. Выбор инструментов для визуализации данных

Настоящая глава посвящена теме выбора программного обеспечения для визуализации данных. Готовые решения для визуализации – это самый простой вариант, с которого могут начать новички. Некоторые из этих инструментов, такие как Microsoft Excel или Google Documents, предназначены для управления данными и графикой на базисном уровне, в то время как другие сконструированы для более углубленного анализа и визуального исследования.

[Google Spreadsheets](https://www.google.ru/intl/ru/sheets/about/) — это, по сути, облачная версия Microsoft Excel. Однако у Google Spreadsheets есть и некоторые преимущества перед Microsoft Excel. Во-первых, стандартный набор диаграмм, несколько иной (рис. 4).

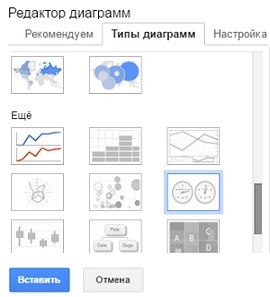


Рис. 4. Набор диаграмм у Google Spreadsheets богаче, чем в Excel

Many Eyes — это действующий исследовательский проект Лаборатории визуальной коммуникации IBM. Представляет собой онлайн-приложение, дающее возможность загружать свои данные в виде текстового файла с разделителями и изучать этот файл с помощью набора интерактивных инструментов визуализации. В нем есть почти все традиционные типы визуализации, такие как линейный график и диаграмма рассеяния, но, помимо этого, Many Eyes предлагает целый набор различных, более продвинутых и экспериментальных типов визуализации, а также ряд базовых инструментов для составления карт, например, дерево слов или облако слов (рис. 5).



Рис. 5. Текстовое облако из слов Конституции США

В настоящий момент – это пока что самый богатый по возможностям бесплатный набор инструментов для исследования данных.

Несмотря на то что все эти инструменты просты в применении, у них есть и некоторые недостатки. Чтобы не лишать себя удобной возможности работать методом перетаскивания (drag and drop), вам придется несколько ограничить себя в гибкости относительно того, что можно сделать с данными. По мере того, как вы начнете работать со все большими объемами информации, чаще станут возникать ситуации, в которых существующего программного обеспечения вам будет недостаточно. И когда наступит этот момент, вы сможете обратиться к программированию.

**Программирование.** Умение писать программы позволяет вам поступать более гибко и лучше адаптироваться к различным типам данных. Вся классная инфографика создается с помощью специально разработанного кода. Я предпочитаю использовать *Python* (см., например, [Майк МакГрат. Программирование на Python для начинающих](http://baguzin.ru/wp/?p=14338)). *PHP* был первым языком, который я выучил, когда начал заниматься созданием программ для Сети. Существует множество PHP-библиотек для построения диаграмм. Самая популярная из них — это библиотека для создания спарклайнов, позволяющая вставлять маленькие, размером со слово, графики в текст или добавлять визуальные компоненты в числовые таблицы (рис. 6; о спарклайнах в Excel см. [Спарклайны или микрографики в ячейках Excel](http://baguzin.ru/wp/?p=9029)).

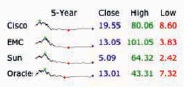


Рис. 6. Инфографики, созданные с применением графической библиотеки РНР

*Processing* — это язык программирования с открытым исходным кодом, предназначенный для дизайнеров и художников. Для интерактивных и динамических инфографиков в Сети часто используют *Flash* и *ActionScript*. Графику можно создавать и только во Flash, тем более что интерфейс программы работает по принципу перетаскивания (drag and drop), однако ActionScript позволяет лучше контролировать процесс взаимодействия пользователей с интерактивными объектами. Многие приложения написаны целиком на ActionScript без использования среды Flash. Тем не менее код компилируется как Flash-приложение. В последнее время произошел заметный сдвиг в сторону визуализации, которая легко запускается без «посторонней помощи» в браузере через *HTML, JavaScript и CSS*. Вы можете использовать плагин Sparkline для jQuery, чтобы делать маленькие диаграммы (рис. 7).



Рис. 7. Инфографики, созданные с применением плагина jQuery Sparklines

R — мой излюбленный инструмент для создания инфографики. Большинство статистиков предпочитают работать именно с R.

Существуют также и платные альтернативы, такие как S-plus и SAS, но им тяжело тягаться с бесплатным вариантом и с активным сообществом его сторонников. Когда вы доберетесь до R, не думайте, что увидите интерфейс, где все можно делать мышью. С помощью разных пакетов R доступны практически все виды диаграмм. К сожалению, R мало подходит для интерактивной графики и анимации. Для анализа «сырой» продукт R подходит отлично, но для презентаций и сторителлинга лучше будет немного поработать над эстетикой. Я делаю основу графического объекта в R, а затем редактирую его и совершенствую в каком-нибудь приложении для подготовки макетов документов, например, в Adobe Illustrator.

Любая графика, построенная на статистических данных, которая выглядит как сделанная на заказ или появляется в том или ином крупном издании, скорее всего, на одном из этапов прошла обработку в Adobe Illustrator. Эта программа — отраслевой стандарт. Каждый графический объект, который печатается на страницах New York Times, был или создан, или отредактирован в Illustrator. [Inkscape](https://inkscape.org/ru/) — это бесплатная альтернатива Adobe Illustrator. Также доступно [обучающие руководства](https://inkscapetutorials.org/) по Inkscape.

В последние годы объемы географических данных значительно возросли, а вместе с ними увеличилось и количество способов, которыми можно пользоваться для создания карт. Услуги мобильного позиционирования находятся на подъеме, все больше становятся массивы данных с привязанными к ним широтой и долготой. Помимо прочего, карты — невероятно интуитивный способ визуализации данных. Карты Google, Yahoo и Microsoft – самое простое онлайновое решение. Чем лучше вы умеете писать коды, тем большего вы сможете добиться с помощью API для создания карт.

### Глава 4. Визуализация паттернов во времени

Столбчатые диаграммы — один из самых популярных типов диаграмм (рис. 8). Всегда начинайте ось значений с нуля, иначе соотношение величин на вашей диаграмме будет отображаться некорректно.



Рис. 8. Структура столбчатой диаграммы

Пример диаграммы для визуализации темпоральных (временн*ы*х) данных, и работа с ними в R более подробно рассмотрена в отдельной заметке – [Создание столбчатой диаграммы в R](http://baguzin.ru/wp/?p=14499). В заметке также рассмотрены основы работы в Adobe Illustrator.

Штабельные столбчатые диаграммы помимо временн*о*й динамики позволяют анализировать подкатегории (рис. 9). В R для построения такой диаграммы используется функция barplot(), первым аргументом которой должна быть матрица (в обычной столбчатой диаграмме используется та же функция, но аргументом выступает один или два столбца).

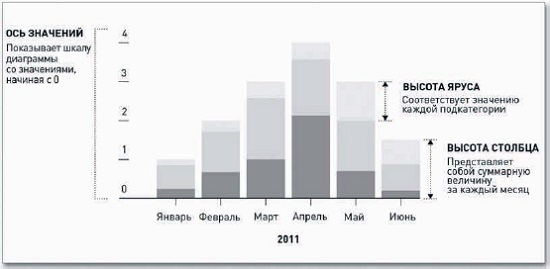


Рис. 9. Структура штабельной столбчатой диаграммы

Иногда лучше использовать не столбцы, а точки. Они занимают меньше места и, поскольку нет брусков, более удачно создают ощущение непрерывного потока, переходящего от одной позиции к другой (рис. 10). В R для построения такой диаграммы используется функция plot() с типом type="p" ("p" – points/точки, "l" – lines/линии, "b" – both, обе).

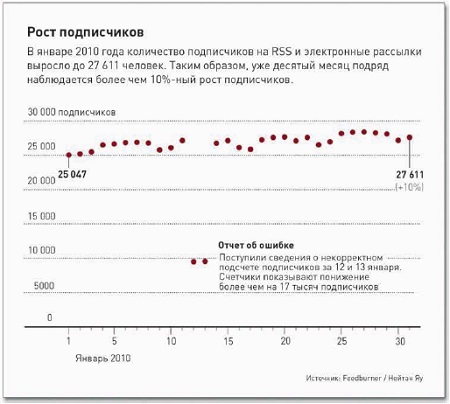


Рис. 10. Точечная диаграмма, созданная в R и обработанная в Illustrator

Несложную диаграмму можно сразу построить в Adobe Illustrator. Для начала в панели *Инструменты* выберите *Вертикальные полосы*. Кликните правой кнопкой мыши на треугольнике в правом нижнем углу пиктограммы и выберите *Линейная диаграмма* (рис. 11).

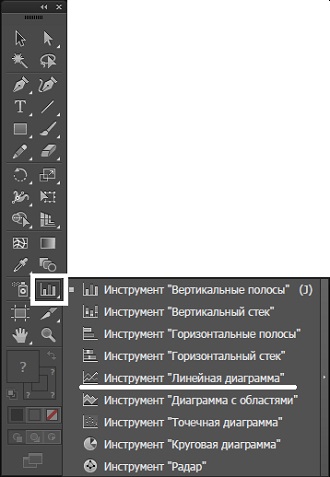


Рис. 11. Инструмент *Линейная диаграмма* в Adobe Illustrator

Откройте CSV-файл world-population.csv в Excel или Google Documents (рис. 12а), скопируйте все строчки за исключением шапки таблицы (то есть ее верхней строки). Вернитесь обратно в Illustrator. Используя инструмент *Линейная диаграмма*, в новом файле щелкните мышью и перетащите курсор по диагонали, очертив прямоугольную область, примерно равную по размеру тому графику, который вы хотите получить в конечном итоге. Появится таблица. Вставьте в нее данные из буфера памяти (рис. 12б), а затем щелкните по кнопке с галочкой в верхнем правом углу. Вы построили линейный график по умолчанию (рис. 13). Теперь доработайте его, и получите небольшой дизайнерский шедевр (рис. 14).

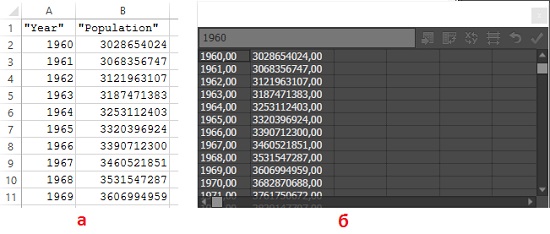


Рис. 12. Перенос данных через буфер из Excel (а) в Illustrator (б)

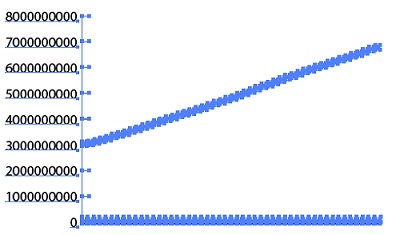


Рис. 13. Линейный график по в Illustrator умолчанию



Рис. 14. Изменение численности населения мира за последние пять десятилетий

Один из недостатков стандартного линейного графика состоит в том, что он внушает мысль о стабильном движении от точки А к точке В. Иногда все так и обстоит, как в примере выше. Но в других случаях после длительного времени пребывания в неизменном состоянии вдруг происходит скачок. Так ведут себя тарифы, ставки и т.п. Для такого типа данных существует ступенчатый график (рис. 15). В R для такого типа графиков используется функция plot() с типом type="s".



Рис. 15. Изменение почтовых ставок

### Глава 5. Визуализация пропорций

Временные ряды самым естественным образом группируются по... времени. Данные о пропорциях также группируются, но по категориям, подкатегориям и совокупностям. В случае с пропорциями обычно ищут три параметра: минимум, максимум и распределение в целом. Для демонстрации минимума и максимума вам не нужна диаграмма. Самое интересное — это распределение пропорций. Как соотносится группа людей, сделавшая тот или иной выбор, с остальными? Как распределяются калории между жирами, белками и углеводами — поровну, или одна из этих групп доминирует?

Части целого — это самый простой вид пропорций, а круговые диаграммы — старый испытанный вариант их представления (рис. 16а). Вы можете создать круговую диаграмму сразу же в Illustrator, а можете начать с R с помощью функции pie() (подробнее см. [Базовые графические возможности R: круговые диаграммы](http://r-analytics.blogspot.ru/2012/01/r.html#.V0rFspGLSUk)). У круговой диаграммы есть младшая кузина: кольцевая диаграмма (рис. 16б).



Рис. 16. Интересы посетителей сайта FlowingData: (а) круговая диаграмма, (б) кольцевая диаграмма

Создание кольцевой диаграммы с помощью программы Protovis рассмотрено в [отдельной заметке](http://baguzin.ru/wp/?p=14537). Protovis — бесплатный инструмент для визуализации с открытым исходным кодом. Protovis — это JavaScript-библиотека, позволяющая использовать возможности современных браузеров для работы с масштабируемой векторной графикой. Поскольку графические объекты генерируются динамически, это дает возможность делать их анимированными и интерактивными. А потому Protovis — отличный выбор для создания онлайн-графики.

В 1990 году Бен Шнайдерман из Университета Мэриленда придумал тримап. Этот способ визуализации основан на областях, при котором размер каждого прямоугольника выражает собой количественный показатель. Внешние прямоугольники представляют родительские категории, а прямоугольники внутри родительских категорий — это своего рода подкатегории. Вы можете использовать тримап для визуализации прямых пропорций, но, чтобы опробовать технологию в полном объеме, ее лучше применять с иерархическими или, точнее, с древовидно структурированными данными (рис. 17). В R тримап можно построить с помощью функции map.market(), содержащейся в пакете Portfolio.



Рис. 17. Тримап, созданный в R и отредактированный в Illustrator

Иногда приходится работать с пропорциями во времени. Такой тип диаграмм можно рассматривать как серию штабельных столбчатых диаграмм, столбцы которых соединены во времени (рис. 18). В Illustrator для этих целей используется *Диаграмма с областями* (аналог есть и в Excel).

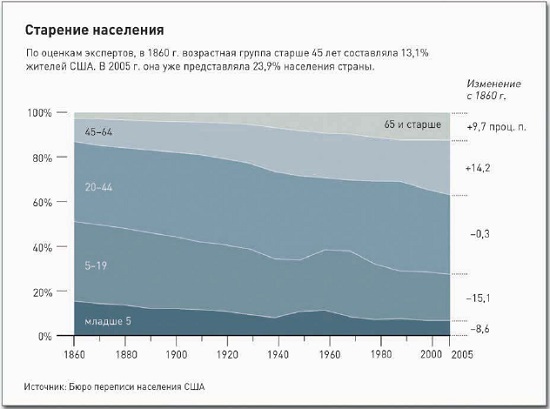


Рис. 18. Штабельная диаграмма с областями, сопровождаемая подписями

Один из недостатков использования штабельных диаграмм с областями состоит в том, что они становятся нечитабельными и бесполезными, если у вас много категорий и много точек данных. В случае с разбивкой на возрастные группы диаграмма работает, так как категорий всего пять. Но если добавлять их еще и еще, ярусы в конце концов начнут выглядеть как узенькие полоски. Ситуация усложняется и тогда, когда у вас есть одна категория с относительно небольшой долей в общей сумме — она легко может затеряться между более внушительными категориями. Однако если сделать штабельную диаграмму с областями интерактивной, это проблема становится решаемой.

Интерактивные штабельные диаграммы с областями сделал популярными Мартин Ваттенберг с его [NameVoyager](http://www.babynamewizard.com/voyager). Данный инструмент используется для визуализации выбора имен для новорожденных во времени. Когда вы вводите имя в окошко поиска, диаграмма автоматически обновляется.

Один из недостатков штабельных диаграмм с областями состоит в том, что в них тяжело разглядеть тенденции в той или иной отдельной группе, потому что местоположение каждой точки зависит от точек, которые ниже нее. А потому иногда бывает лучше представить пропорции как простые временные ряды (рис. 19).

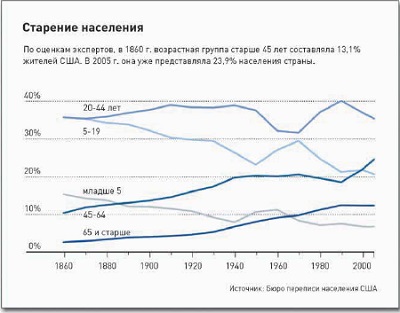


Рис. 19. Линейная диаграмма иногда будет хорошей альтернативой штабельной диаграмме с областями

### Глава 6. Визуализация зависимостей

Корреляция — наверное, это первое, о чем вы подумали, когда услышали про поиск зависимостей в данных. Второе — скорее всего, это причинно-следственная связь. А теперь вы наверняка вспомнили мантру о том, что корреляция и причинно-следственная связь — это «две большие разницы». Корреляцию легко выявить и отобразить графически. У ученых на доказательство причинно-следственной связи порой уходят годы. В качестве примера рассмотрим уровень преступности в США по штатам за 2005 год в пересчете на 100 000 человек. Всего мы рассмотрим семь типов преступлений, а начнем с пары: кражи со взломов / убийства (рис. 20).

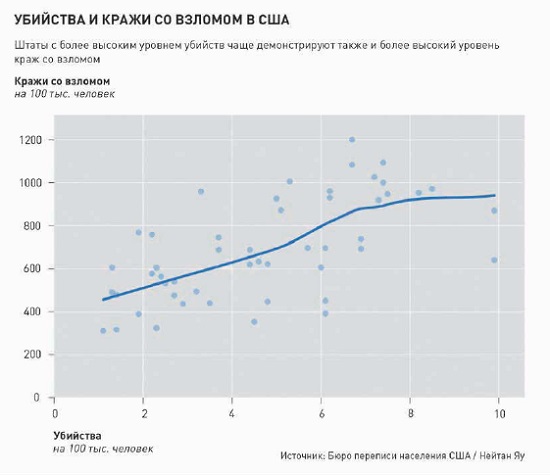


Рис. 20. Диаграмма рассеяния, сопоставляющая данные по убийствам и по кражам со взломом

Чтобы рассказать полную историю, вы должны понимать данные. Чем больше вы о них будете знать, тем лучше получится история, которую сможете поведать. Вот почему, чтобы получить все возможные парные сравнения, полезно воспользоваться матрицей диаграмм рассеяния (рис. 21).

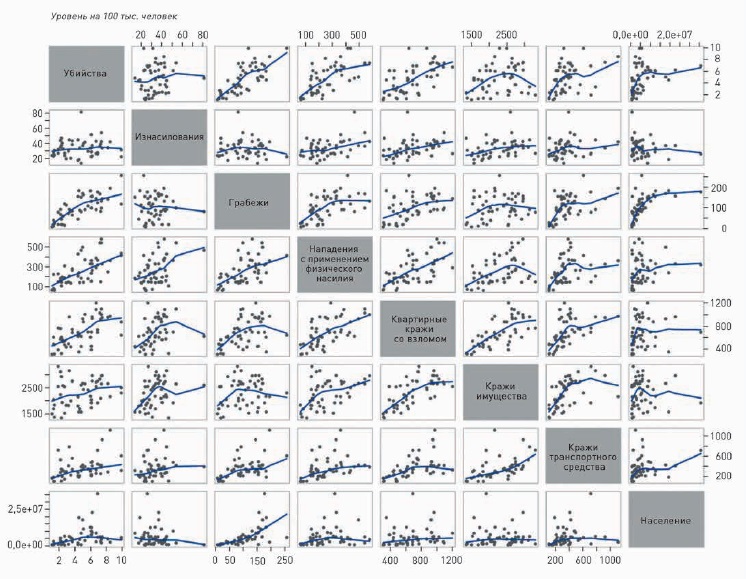


Рис. 21. Матрица диаграмм рассеяния для сопоставления уровней различных типов преступности

Преимущество пузырьковой диаграммы состоит в том, что они дают возможность сравнивать три переменные одновременно (рис. 22). Одна переменная располагается на оси Х, другая — на оси Y, а выражением третьей служит размер площади пузырьков (именно площади, а не диаметра; подробнее см. [Пузырьковая диаграмма](http://baguzin.ru/wp/?p=1077)).

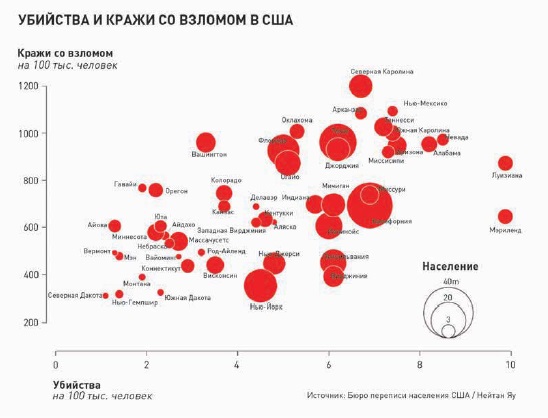


Рис. 22. Пузырьковая диаграмма, демонстрирующая уровень преступности в США; аналогична рис. 20, но также добавлено население штатов в качестве третьей величины

**Диаграммы распределения**

В 1970-х годах, когда компьютеры не были так широко распространены, графики и диаграммы чертились в основном от руки. Часть советов знаменитого статистика Джона Тьюки, которые он давал в своей книге Анализ результатов наблюдений. Разведочный анализ (М.: Мир, 1981) касались использования ручек и карандашей для достижения плотности линий и оттенков. Тогда же возникла и диаграмма типа «стебель с листьями» (или «опора и консоль»). Все, что вам нужно сделать, — это написать числа в определенном порядке, и в результате вы получите общее представление о распределении. Данный метод пользовался особой популярностью в 1980-х годах (когда применение статистических графиков для анализа данных только набирало обороты), потому что включить такой график в документ было легко, даже если текст набирался на печатной машинке (подробнее см. [Как упорядочить массив данных](http://baguzin.ru/wp/?p=5296)). Однако данный метод считается уже устаревшим, так что лучше, наверное, будет воспользоваться гистограммой (рис. 23). По сути, гистограмма является более наглядной версией диаграммы «стебель с листьями».



Рис. 23. Распределение стран по уровню рождаемости

Хотя ось значений непрерывная, распределение все равно разбивается на дискретное число столбцов. Каждый столбец представляет собой некую совокупность объектов, в случае с предыдущим примером — совокупность стран. Но что происходит внутри каждого такого деления, каждой «корзины»? На рис. 24 показано использование кривой вместо столбцов. Общая площадь под кривой равняется единице, а вертикальная ось представляет вероятность или пропорцию данного значения в этой самой совокупности.

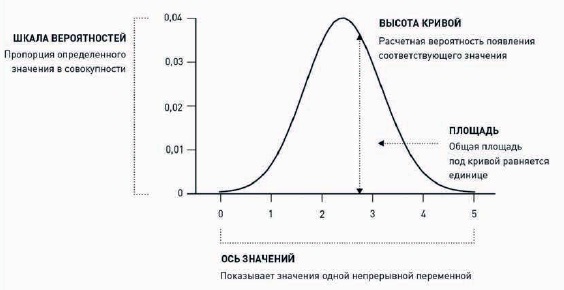


Рис. 24. Структура графика плотности

Очень часто бывает полезно сравнивать несколько распределений. Это может быть динамика во времени, или по регионам. В Excel для этих целей могут использоваться спарклайны (подробнее см. [Спарклайны или микрографики в ячейках Excel](http://baguzin.ru/wp/?p=9029)). Small multiples впервые использовал Эдвард Тафти для описания множества небольших по размеру, однотипных по структуре, но многовариантных (в плане данных) диаграмм или графиков, которые легко сравнивать. Как и большинство терминов Тафти, small multiples тяжело поддается переводу. На русском встречаются различные варианты: «типовые множества», «маленькие однотипности», «размноженные миниатюры», «мультиграфики» и даже «раскадровки». Вот вам две истории, рассказанные с помощью миниграфиков (рис. 25 и 26; о Тафти см., например, [Принцип Эдварда Тафти минимизации количества элементов диаграммы](http://baguzin.ru/wp/?p=2071)).

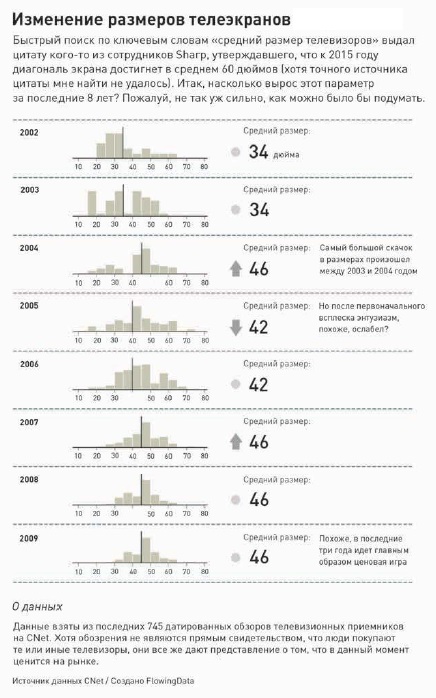


Рис. 25. Распределение размеров телевизионных экранов во времени

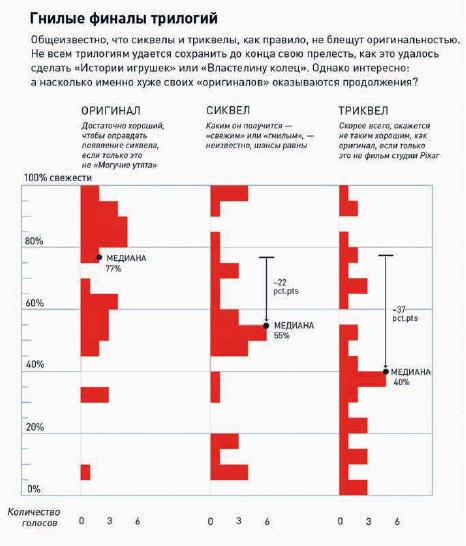


Рис. 26. Рейтинги трилогий от оригинала до финала

### Глава 7. Выявление различий

Эта глава посвящена тому, как обособлять группы в рамках совокупности, причем делать это по большому числу критериев, а также как выявлять выбросы, применяя здравый смысл. Один из самых прямолинейных способов визуализации таблицы данных — показать ее одновременно всю. Но при этом вместо чисел в качестве индикатора значений вы можете использовать цвета. Такое представление называется тепловой картой.[[1]](#footnote-1) Для создания тепловой карты в R используется функция heatmap().

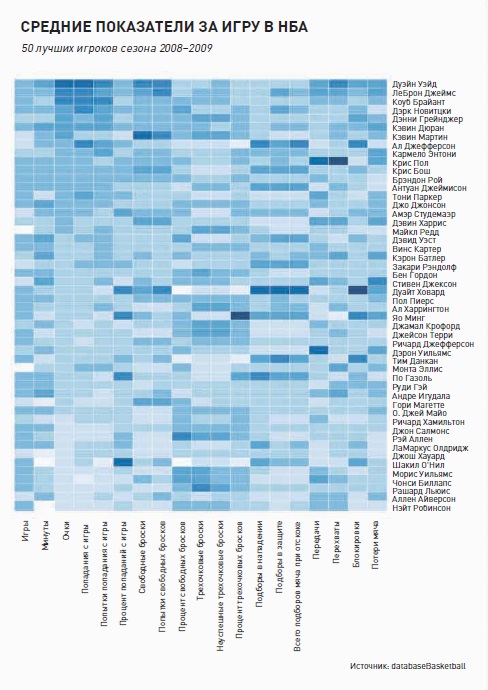


Рис. 27. Тепловая карта, демонстрирующая средние показатели за игру 50 лучших игроков НБА сезона 2008-2009 гг. Более темный цвет соответствует лучшим показателям. Игроки отсортированы по колонке *Очки*

Для наглядного представления многомерных данных хорошо подходят диаграммы-радары, также известные как паутинные, или лепестковые, диаграммы (а еще их иногда называют диаграммами-звездами; рис. 28). Центр — это минимальное значение каждой из переменных, а конец оси — ее максимум. На одной диаграмме можно представить несколько элементов. Однако в спешке легко напортачить и сделать графику бесполезной, иными словами, плохо рассказать историю. Так что лучше придерживаться принципа «один элемент — одна диаграмма» и сравнивать их между собой. В R для создания таких диаграмм используется функция stars().

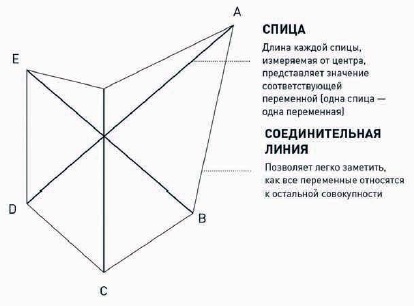


Рис. 28. Структура диаграммы-радара

Еще один вид диаграмм для выявления общности и различий – параллельные координаты, изобретенные в 1885 году Морисом д'Оканем (Maurice d'Ocagne). В этом типе графики вы размещаете множество осей параллельно друг другу (рис. 29). Верхушка каждой оси представляет максимальное значение переменной, а ее нижняя точка — минимальное. Для каждого элемента слева направо вычерчивается линия, которая двигается вверх и вниз в зависимости от значений конкретного элемента.

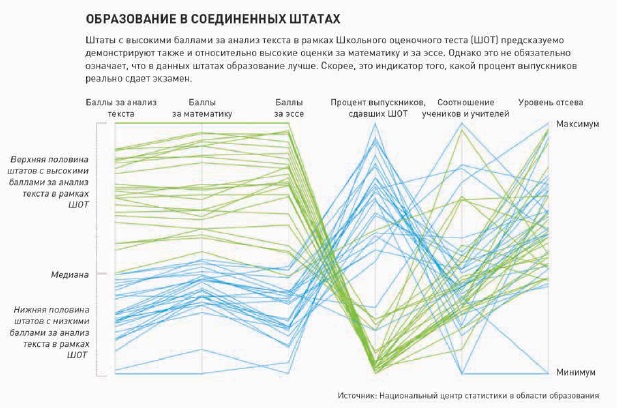


Рис. 29. Диаграмма с параллельными координатами

Следующий метод выделения групп – кластерный анализ, и его разновидность – многомерное шкалирование (подробнее см. [Кластерный анализ: сетевые графы и определение сообществ](http://baguzin.ru/wp/?p=13926)). Представьте себе, что вы находитесь в пустой комнате и что помимо вас в ней есть еще два человека. Ваша задача — определить, где именно им следует располагаться согласно их росту. Чем более они похожи по этому показателю, тем ближе должны стоять, и, соответственно, чем существенней разница в их росте, тем дальше они должны оказаться друг от друга. Но представьте, что у вас не двое, а 50 человек и вам необходимо их расставить по местам в комнате на основе, скажем, пяти критериев. Вот для этого и существует многомерное шкалирование. Если на основании данных о прохождении теста (см. рис. 29) расположить штаты на 2-мерной оси координат, то они сгруппируются в соответствии с полученными баллами (рис. 30).

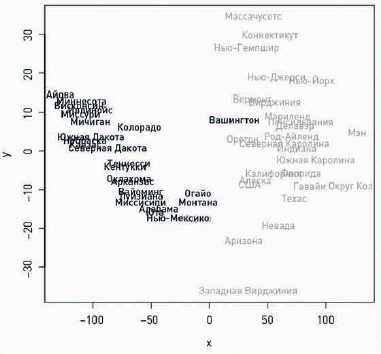


Рис. 30. Штаты окрашены в соответствии с баллами за анализ текста

Помимо того, как группируются отдельные элементы данных, вас должно интересовать и то, как они не группируются. Иными словами, хорошо бы изучить выбросы. Иногда они могут оказаться самой интересной частью вашей истории, в других случаях — просто досадными опечатками.

### Глава 8. Визуализация пространственных отношений

Карты — это форма визуализации, которая обладает дополнительным преимуществом: она невероятно интуитивна. В Сети самый распространенный способ определить на карте некую точку — это воспользоваться Google или Bing Maps. Через их API вы получаете интерактивную карту, которую можете изменять в масштабе и прокручивать с помощью всего нескольких строк JavaScript.

Поскольку исходные данные, как правило, включают перечень адресов, вам придется обратиться к геокодированию для извлечения широты и долготы точек. Если вам необходимо геокодировать несколько адресов, проще всего зайти на соответствующий сайт и ввести адреса вручную. Однако, если вам необходимо геокодировать множество пространственных точек, тогда вам лучше сделать это программными средствами, например, с помощью пакета [Geopy](https://pypi.python.org/pypi/geopy) языка Python. Хотя функционал R в области маппинга довольно ограничен, тем не менее он существенно облегчает процесс размещения точек на карте. Почти всю работу за вас выполнит пакет maps.

В реальной жизни данные обычно содержат не только координаты неких местоположений. Как правило, к ним привязана и еще какая-нибудь величина, такая как объем продаж компании или население города. В этом случае вы можете наложить на карту пузырьковую диаграмму (рис. 31).

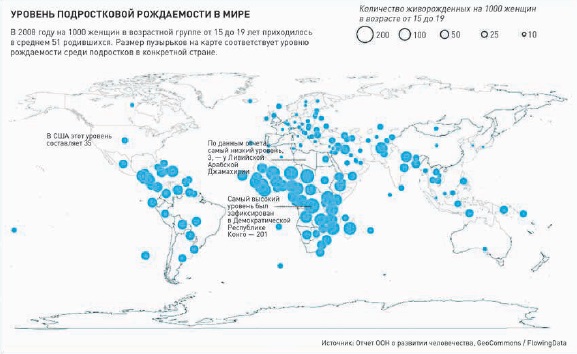


Рис. 31. Данные об уровне подростковой рождаемости

Однако чаще данные соответствуют не точкам, а регионам или странам. А самый распространенный способ нанесения на карту региональных данных — это картограмма. В ней регионы окрашиваются в различные цвета на базе неких количественных показателей и в соответствии с разработанной вами цветовой шкалой (рис. 32).

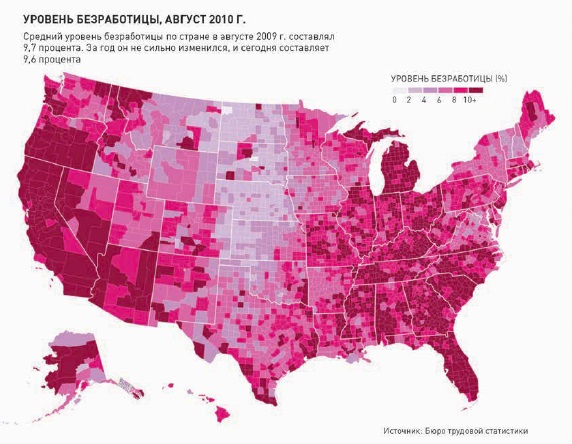


Рис. 32. Уровень безработицы по округам США

Если вы добавите еще одно измерение данных, вы сможете рассмотреть изменения одновременно и во времени, и в пространстве. В главе 4 вы визуализировали время с помощью линий и графиков, но, когда к подобного рода данным добавляется еще и местоположение, может оказаться, что карты — это более интуитивно понятный способ представить паттерны и изменения. Так бывает легче разглядеть кластеры или группы регионов, которые находятся близко друг к другу в физическом плане. Для этих целей отлично подойдут маленькие панели (рис. 33).

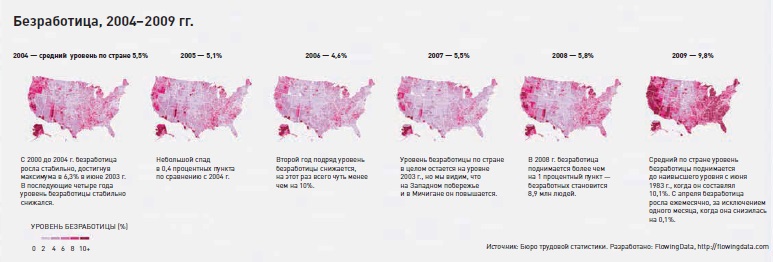


Рис. 33. Уровень безработицы с 2004 по 2009 гг.

Один из наиболее очевидных способов визуализации изменений во времени и пространстве — это анимация данных. Вместо того чтобы показывать отдельные карты для различных срезов времени, вы можете продемонстрировать изменения в динамике на одной-единственной интерактивной карте. Несколько лет назад я разработал карту, показывающую распространение магазинов сети Walmart на территории [Соединенных Штатов](http://projects.flowingdata.com/walmart/) (подробнее см. Сэм Уолтон. Сделано в Америке. Как я создал Wal-Mart).

### Глава 9. Прицельный дизайн

Когда вы занимаетесь самостоятельным изучением данных, вам не нужно думать о сторителлинге. В конце концов, вы и есть автор истории. Однако в тот момент, когда вы решите использовать эти данные, чтобы донести некую информацию до других голой диаграммы окажется уже недостаточно.

Для начала, чтобы рассказать хорошую историю с данными, вам необходимо знать их источник. Именно этот начальный процесс сбора данных и делает графику New York Times настолько качественной. Сбор и приведение данных в порядок — процедура намного более трудоемкая, нежели собственно создание графики.

Ваша работа в качестве дизайнера инфографики состоит в донесении того, что вы знаете, до вашей аудитории, которая, скорее всего, данных не знает, а потому, если не получит каких-нибудь объяснений или установок, может и не разглядеть в графике то, что видите вы. Вам не нужно писать целые эссе, чтобы сопровождать ими каждую вашу работу, но заголовок и кое-какие пояснения во вводном абзаце всегда бывают полезны. Не принимайте по умолчанию, что ваши читатели в курсе всего или что они и сами разберутся в вашей графике. Такого не будет, особенно в Сети, где все привыкли, чуть что, сразу щелкать по следующей ссылке и переходить далее.

В главе 1 вы увидели, как работает кодировка. По сути, у вас есть данные, и вы кодируете их с помощью геометрии, цвета или анимации. А затем читатели декодируют эти фигуры, цвета и движения, представляя себе их снова в виде чисел. Это и есть основа визуализации. Кодирование означает перевод на визуальный язык. Декодирование помогает взглянуть на данные под другим углом и разглядеть паттерны, которые вы бы не заметили, если б смотрели на таблицу.

На примерах, приведенных выше, вы могли убедиться, что хороший дизайн работает не только на эстетику. Он делает графику более понятной и способен менять отношение читателей к данным или к истории, которую вы пытаетесь им рассказать. Графики, создаваемые в R или Excel с настройками по умолчанию, смотрятся сыровато и похожи на штамповку. В инфографике выбор цвета играет важную роль. Цвет способен вызывать (или не вызывать) эмоции, он помогает доносить информацию. И это ваша ответственность — выбрать цвета, которые правильно передадут ваше послание. Цвета должны соответствовать истории, которую вы пытаетесь рассказать.

По теме см. также

[Левин. Статистика для менеджеров с использованием Microsoft Excel](http://baguzin.ru/wp/?p=5285)

[Идеи Байеса для менеджеров](http://baguzin.ru/wp/?p=6355)

[Как с помощью диаграммы приукрасить действительность? или о факторе лжи Эдварда Тафти](http://baguzin.ru/wp/?p=2086)

[Дарелл Хафф. Как лгать при помощи статистики](http://baguzin.ru/wp/?p=12682).

1. Любопытно, что я использовал тепловую карту, еще не зная ее названия, при анализе частоты обращения к ячейкам хранения на складе; см. [Совершенствование складской логистики на основе методов менеджмента качества](http://baguzin.ru/wp/?p=3806), рис. 11. [↑](#footnote-ref-1)