

Введение

Нельзя сказать, что данные — это что-то новое. Люди занимаются количественными измерениями и составлением таблиц уже не одно столетие. Тем не менее в последние годы — с тех пор как я начал писать для FlowingData, моего сайта по дизайну, визуализации и статистике, — я наблюдаю настоящий бум в этой области, и конца-края ему не видно. Совершенствование технологий сделало сбор и хранение данных задачей чрезвычайно простой, а сеть позволяет получать к ним доступ в любой момент, когда захочется. Это изобилие данных, окажись они в хороших руках, способно стать кладезем информации, помогающей принимать более дальновидные решения, излагать свои идеи более убедительно и создавать более объективное представление о том, как люди смотрят на мир и на самих себя.

Значительные сдвиги в публикации правительственных данных произошли в середине 2009 года, когда Соединенные Штаты запустили портал Data.gov. Он представляет собой всеобъемлющий каталог информации, предоставляемой федеральными министерствами и ведомствами, и демонстрирует прозрачность и подотчетность всех этих организаций и должностных лиц. Сайт задумывался для того, чтобы предоставить гражданам страны возможность ознакомиться с тем, на что власти тратят налоговые поступления. А ведь до этого правительство больше походило на черный ящик. Значительная часть данных на Data.gov и раньше лежала в открытом доступе — она находилась на сайтах ведомств, но те были разбросаны по всей Сети. Теперь же все сведения собраны в одном месте и отформатированы так, что их анализ и визуализация стали намного проще. И у Организации Объединенных Наций есть подобный портал — UNdata. Прошло совсем немного времени, и этому примеру последовало Соединенное Королевство, организовав свой Data.gov.uk. Крупные города мира, такие как Нью-Йорк, Сан-Франциско и Лондон, также являются поставщиками большого количества данных.

Открытию коллективной сети также во многом поспособствовало появление тысяч интерфейсов программирования приложений (application programming interfaces, API). Задача была — приобщить разработчиков и соблазнить их сделать что-то со всеми этими данными. Такие приложения, как Twitter и Flickr, имеют API с широким функционалом, что делает возможным создание пользовательского интерфейса, совершенно отличного от того, который можно видеть на самих сайтах. На ProgrammableWeb — ресурсе, занимающемся каталогизацией API, — их представлено более двух тысяч. Не так давно появились и новые приложения, такие как Infochimps и Factual, разработанные специально для предоставления структурированных данных.

У себя, на индивидуальном уровне, вы можете обновить друзей в Facebook, поделиться сведениями о своем местонахождении через Foursquare или пошебетать о том, чем занимаетесь, в Twitter — все это можно сделать парой кликов мышью или нажатием нескольких клавиш на клавиатуре. Более специализированные приложения дают вам возможность вести учет того, что вы едите, сколько вы

[Купить книгу на сайте kniga.biz.ua >>>](http://kniga.biz.ua)

весите, каково ваше настроение и многого другого. Какую бы информацию о себе вы ни захотели собрать, наверняка найдется приложение, которое поможет вам это сделать.

Если вокруг собрано столько данных — в магазинах, на складах и в базах, — значит, ситуация созрела для появления людей, которые способны их осмыслить. Сами по себе данные не так уж интересны (по крайней мере, для большей части человечества). Интересна информация, которую можно извлечь из этих данных. Люди хотят знать, о чем говорят их данные, и если вы способны помочь им в этом, то будете весьма востребованы. Вот почему Хол Вариан (Hal Varian), главный экономист Google, говорит, что статистик — самая секси работа грядущего десятилетия, и отнюдь не потому, что статистики — такие красавчики. (Хотя если взглянуть на нас сквозь призму гиковского* шика, то мы очень даже ничего.)

Визуализация

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.

Джон Тьюки (John Tukey), мой любимый статистик и отец разведочного анализа данных, разбирался в статистических методах и свойствах, как мало кто другой, и верил, что графические методы также имеют законное право на существование. Он был глубоко убежден, что картинки способны открывать нам неожиданное. Вы можете очень многое узнать из данных, просто визуализируя их, а в ряде случаев это и есть все, что вам нужно сделать, чтобы принять информированное решение или рассказать историю.

Например, в 2009 году в Соединенных Штатах произошел значительный рост уровня безработицы. В 2007 году он составлял в среднем 4,6 процента, в 2008 году поднялся до 5,8 процента, а к сентябрю 2009 года дошел уже до 9,8 процента. Однако такие средние по стране величины способны поведать лишь часть истории. Это обобщенные данные по США в целом. Но, может, были какие-то регионы, в которых уровень безработицы оказался выше, чем в других местах? Может, были регионы, которые эта беда вообще обошла стороной?

Карты, представленные на рис. 0.1, рассказывают более полную версию этой истории, вам достаточно лишь взглянуть на них — и вы сможете ответить на вопросы из предыдущего абзаца. Округа, окрашенные в более темный цвет, — это области, в которых уровень безработицы был сравнительно высоким, в то время как в округах, окрашенных светлым, ее уровень оказался относительно низким. В 2009 году вы уже видите множество регионов на западе, где уровень безработицы стал выше 10 процентов. Такая же ситуация сложилась и в большинстве регионов на востоке. А вот регионы Среднего Запада пострадали не так сильно (см. рис. 0.2).

* Гик (англ. geek) — человек, чрезмерно увлеченный какой-либо темой и по этой причине несколько выпадающий из реальности. В русском языке чаще применяется к людям, одержимым (компьютерными) технологиями. *Прим. пер.*

Безработица, 2004–2009 гг.

2004 — средний уровень по стране 5,5%

2005 — 5,1%

2006 — 4,6%

2007 — 4,6%

2008 — 5,8%

2009 — 9,8%



С 2000 до 2004 г. безработица росла стабильно, достигнув максимума в 6,3% в июне 2003 г. В последующие четыре года уровень безработицы стабильно снижался.



Небольшой спад в 0,4 процентных пункта по сравнению с 2004 г.



Второй год подряд уровень безработицы снижается, на этот раз всего чуть менее чем на 10%.



Уровень безработицы по стране в целом остается на уровне 2003 г., но мы видим, что на Западном побережье и в Мичигане он повышается.



В 2008 г. безработица поднимается более чем на 1 процентный пункт — безработных становится 8,9 млн людей.



Средний по стране уровень безработицы поднимается до наивысшего уровня с июня 1983 г., когда он составлял 10,1%. С апреля безработица росла ежемесячно, за исключением одного месяца, когда она снизилась на 0,1%.

УРОВЕНЬ БЕЗРАБОТИЦЫ (%)

0 2 4 6 8 10+

Источник: Бюро трудовой статистики. Разработано: FlowingData, <http://flowingdata.com>

Рис. 0.1. Карты, демонстрирующие уровень безработицы в США с 2004 по 2009 гг.

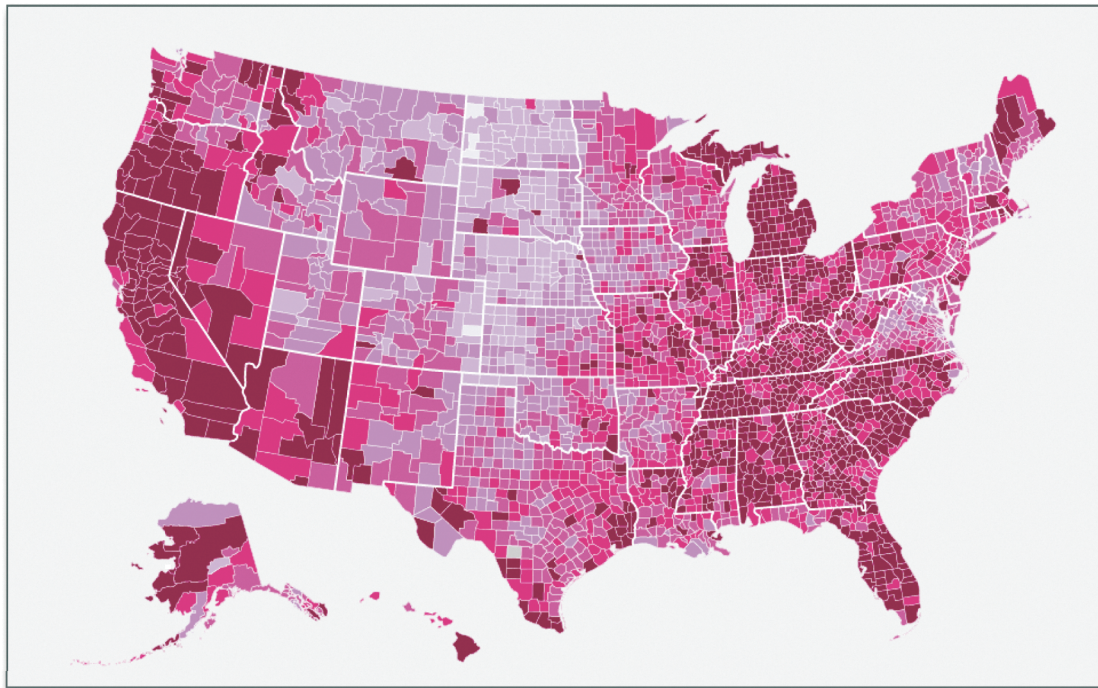


Рис. 0.2. Карта, демонстрирующая уровень безработицы в 2009 г.

Таблица 126. Браки и разводы — общее количество и в расчете на 1000 человек, по штатам, с 1990 по 2007 гг.

(Значение 2443,5 представляет собой 2 443 500 человек. По месту проживания.)

Штат	Браки ¹						Разводы ³					
	Количество (в тыс.)			На 1000 человек ²			Количество (в тыс.)			На 1000 человек ²		
	1990	2000	2007	1990	2000	2007	1990	2000	2007	1990	2000	2007
США*	2443,5	2329,0	2204,6	9,8	8,3	7,3	1182,0	НД	НД	4,7	4,1	3,6
Алабама	43,1	45,0	42,4	10,6	10,3	9,2	25,3	23,5	19,8	6,1	5,4	4,3
Аляска	5,7	5,6	5,8	10,2	8,9	8,4	2,9	2,7	3,0	5,5	4,4	4,3
Аризона ⁵	36,8	38,7	39,5	10,0	7,9	6,2	25,1	21,6	21,5	6,9	4,4	3,9
Арканзас	36,0	41,1	33,7	15,3	16,0	11,9	16,8	17,9	16,8	6,9	6,9	5,9
Калифорния	237,1	196,9	225,8	7,9	5,9	6,2	128,0	НД	НД	4,3	НД	НД
Колорадо	32,4	35,6	29,2	9,8	8,6	6,0	18,4	НД	21,2	5,5	НД	4,4
Коннектикут	26,0	19,4	17,3	7,9	5,9	4,9	10,3	6,5	10,7	3,2	2,0	3,1
Делавэр	5,6	5,1	4,7	8,4	6,7	5,5	3,0	3,2	3,9	4,4	4,2	4,5
Округ Колумбия	5,0	2,8	2,1	8,2	5,4	3,6	2,7	1,5	1,0	4,5	3,0	1,6
Флорида	141,8	141,9	157,6	10,9	9,3	8,6	81,7	81,9	86,4	6,3	5,3	4,7
Джорджия	66,8	56,0	64,0	10,3	7,1	6,7	35,7	30,7	НД	5,5	3,9	НД
Гавайи	18,3	25,0	27,3	16,4	21,2	21,3	5,2	4,6	НД	4,6	3,9	НД
Айдахо	14,1	14,0	15,4	13,9	11,0	10,3	6,6	6,9	7,4	6,5	5,4	4,9
Иллинойс	100,6	85,5	75,3	8,8	7,0	5,9	44,3	39,1	32,8	3,8	3,2	2,6
Индиана	53,2	334,5	51,2	9,6	5,8	8,1	НД	НД	НД	НД	НД	НД
Айова	24,9	20,3	20,1	9,0	7,0	6,7	11,1	9,4	7,8	3,9	3,3	2,6
Канзас	22,7	22,2	18,6	9,2	8,3	6,7	12,6	10,6	9,2	5,0	4,0	3,3
Кентукки	49,8	39,7	33,6	14,5	10,0	7,9	21,8	21,6	19,7	5,8	5,4	4,6
Луизиана	40,4	40,5	32,8	9,6	9,3	7,6	НД	НД	НД	НД	НД	НД
Мэн	11,9	10,5	10,1	9,7	8,3	7,7	5,3	5,8	5,9	4,3	4,6	4,5
Мэриленд	46,3	40,0	35,5	9,7	7,7	6,3	16,1	17,0	17,4	3,4	3,3	3,1
Массачусетс	47,7	37,0	38,4	7,9	6,0	6,0	16,8	18,6	14,5	2,8	3,0	2,2
Мичиган	76,1	66,4	59,1	8,2	6,7	5,9	40,2	39,4	35,5	4,3	4,0	3,5
Миннесота	33,7	33,4	29,8	7,7	6,9	5,7	15,4	14,8	НД	3,5	3,1	НД
Миссисипи	24,3	19,7	15,7	9,4	7,1	5,4	14,4	14,4	14,2	5,5	5,2	4,9
Миссури	49,1	43,7	39,4	9,6	7,9	6,7	26,4	26,5	22,4	5,1	4,8	3,8
Монтана	6,9	6,6	7,1	8,6	7,4	7,4	4,1	2,1	3,6	5,1	2,4	3,7
Небраска	12,6	13,0	12,4	8,0	7,8	7,0	6,5	6,4	5,5	4,0	3,8	3,1
Невада	120,6	144,3	126,4	99,0	76,7	49,3	13,3	18,1	16,6	11,4	9,6	6,5
Нью-Гэмпшир	10,5	11,6	9,4	9,5	9,5	7,1	5,3	7,1	5,1	4,7	5,8	3,9
Нью-Джерси	58,7	50,4	45,4	7,6	6,1	5,2	23,6	25,6	25,7	3,0	3,1	3,0
Нью-Мексико ⁵	13,3	14,5	11,2	8,8	8,3	5,7	7,7	9,2	8,4	4,9	5,3	4,3
Нью-Йорк ⁵	154,8	162,0	130,6	8,6	8,9	6,8	57,9	62,8	55,9	3,2	3,4	2,9
Северная Каролина	51,9	65,6	68,1	7,8	8,5	7,5	34,0	36,9	37,4	5,1	4,8	4,1
Северная Дакота	4,8	4,6	4,2	7,5	7,3	6,6	2,3	2,0	1,5	3,6	3,2	2,4
Огайо	98,1	88,5	70,9	9,0	7,9	6,2	51,0	49,3	37,9	4,7	4,4	3,3
Оклахома	33,2	15,6	26,2	10,6	4,6	7,5	24,9	12,4	18,8	7,7	3,7	5,2
Орегон	25,3	26,0	29,4	8,9	7,8	7,8	15,9	16,7	14,8	5,5	5,0	4,0
Пенсильвания	84,9	73,2	71,1	7,1	6,1	5,7	40,1	37,9	35,3	3,3	3,2	2,8
Род-Айленд	8,1	8,0	6,8	8,1	8,0	6,4	3,8	3,1	3,0	3,7	3,1	2,8
Южная Каролина	55,8	42,7	31,4	15,9	10,9	7,1	16,1	14,4	14,4	4,5	3,7	3,3
Южная Дакота	7,7	7,1	6,2	11,1	9,6	7,7	2,6	2,7	2,4	3,7	3,6	3,1
Теннесси	68,0	88,2	65,6	13,9	15,9	10,6	32,3	33,8	29,9	6,5	6,1	4,9
Техас	178,6	196,4	179,9	10,5	9,6	7,5	94,0	85,2	79,5	5,5	4,2	3,3
Юта	19,4	24,1	22,6	11,2	11,1	8,6	8,8	9,7	8,9	5,1	4,5	3,4
Вермонт	6,1	6,1	5,3	10,9	10,2	8,6	2,6	5,1	2,4	4,5	8,6	3,8
Вирджиния	71,0	62,4	58,0	11,4	9,0	7,5	27,3	30,2	29,5	4,4	4,3	3,8
Вашингтон	46,6	40,9	41,8	9,5	7,0	6,5	28,8	27,2	28,9	5,9	4,7	4,5
Западная Вирджиния	13,0	15,7	13,0	7,2	8,7	7,2	9,7	9,3	9,0	5,3	5,2	5,0
Висконсин	38,9	36,1	32,2	7,9	6,8	5,8	17,8	17,6	16,1	3,6	3,3	2,9
Вайоминг	4,9	4,9	4,8	10,7	10,3	9,3	3,1	2,8	2,9	6,6	5,9	5,5

НД — нет данных. 1. Данные на основе подсчета заключенных браков, за исключением отмеченных. 2. На основании общего количества населения, проживающего на территории; по данным переписи от 1 апреля 1990 и 2000 гг. и по оценкам к 1 июля за все остальные годы. 3. Включая аннулированные браки. 4. Данные о количестве браков и разводов по США в целом оценочные и включают также штаты, по которым нет информации. Начиная с 2000 г. данные об уровне разводов основываются на комбинированной системе подсчета населения в отчетных штатах и в округе Колумбия. Сбор детализированных данных о браках и разводах был отменен в январе 1996 г. 5. Некоторые данные основаны на количестве выданных свидетельств о браке.

Источник: Национальный центр статистики в области здравоохранения США. "Births, Marriages, Divorces and Deaths: Provisional Data for 2007, Vol. 56, №21. July 14, 2008" и предыдущие отчеты

Вы бы не смогли выявить эти географические и временные закономерности так быстро, если бы перед вами была только сводная таблица, и уж точно не сумели бы это сделать, располагая лишь средними по стране величинами. И хотя при наличии данных на уровне округов картина становится более сложной, большинство людей тем не менее способны интерпретировать эти карты. Такие карты помогают политикам решать, куда направлять финансовую помощь или другие формы поддержки.

А самое замечательное — то, что все данные, использованные для создания представленных выше карт, абсолютно бесплатны и доступны широкой общественности на сайте Бюро трудовой статистики. И хотя «нарыть» их было не то чтобы очень просто с такой устаревшей системой представления данных, как у них, так или иначе все тамошние цифры в вашем распоряжении: они сидят и ждут, чтобы кто-нибудь их малость обработал визуально.

«Статистический ежегодник Соединенных Штатов», например, существует в виде сотен таблиц данных (рис. 0.3), но в нем нет ни одного графика. Это неплохая возможность представить всеобъемлющую картину страны. Вот уж действительно интересная штука. Некоторое время назад я перевел часть таблиц в диаграммы просто так, ради идеи. На рис. 0.4 вы можете увидеть динамику браков и разводов, почтовых тарифов, потребления электроэнергии и кое-чего еще. В первом варианте,

Рис. 0.3. Таблица из «Статистического ежегодника Соединенных Штатов»

то есть в таблице, разобраться в данных непросто, и все, что у вас получится извлечь из них, — это отдельные величины. А вот в графическом варианте вы легко можете заметить тенденции и паттерны и с одного взгляда провести сравнение.

Такие поставщики новостей, как New York Times и Washington Post, прекрасно справляются с задачей делать данные более доступными и наглядными. Они, наверное, лучше многих других пользуются всей этой открытой информацией, каждый день рассказывая читателям все новые и новые истории. Иногда графики с данными применяются для обогащения истории иной точкой зрения, а в других случаях графики, собственно, и излагают всю историю.

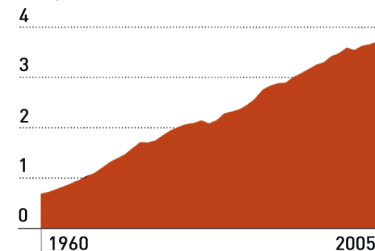
С развитием онлайн-медиа диаграммы и графики получили еще большее распространение. Сегодня в новостных компаниях существуют специальные службы, которые занимаются только интерактивами, или только диаграммами, или только картами. У New York Times, например, есть даже отдел новостей, созданный исключительно для работы с тем, что они называют «компьютеризированными репортажами». Тамошние журналисты специализируются на новостях с цифрами. И отделу графики в New York Times не привыкать к работе с огромными массивами данных.

Визуализация пробилась также в поп-культуру. Фирма Stamen Design, занимающаяся визуализацией и широко известная своими онлайн-интерактивами, в последние несколько лет готовила для ежегодной церемонии вручения премий MTV Video Music Awards трекееры твитов.

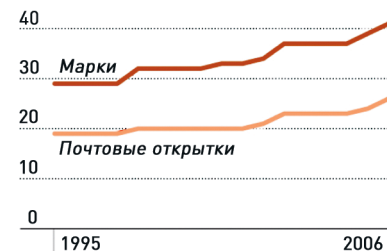
Листая «Национальный статистический справочник»

Недавно Бюро переписи населения США выпустило краткий статистический отчет. Он охватывает сферы искусства, образования, выборов, коммуникаций и многое другое. Ниже представлена часть доступных данных.

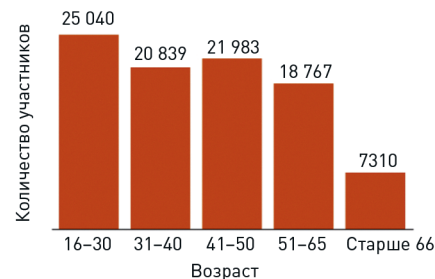
Потребление электроэнергии, 1960–2005 гг.
в млрд киловатт-часов



Тарифы почтовых услуг, 1995–2006 гг.
в центах



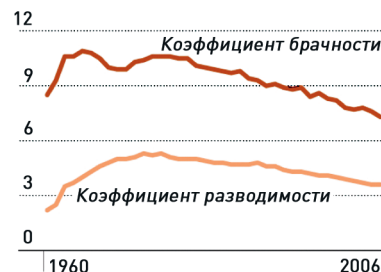
Участники образовательных программ
для взрослых, 2005 г.



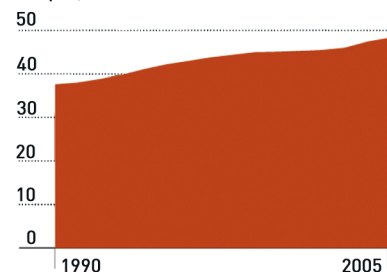
Домохозяйства, имеющие проблемы
с нехваткой питания, 2000–2005 гг.
в процентах



Браки и разводы, 1960–2006 гг.
на 1000 человек



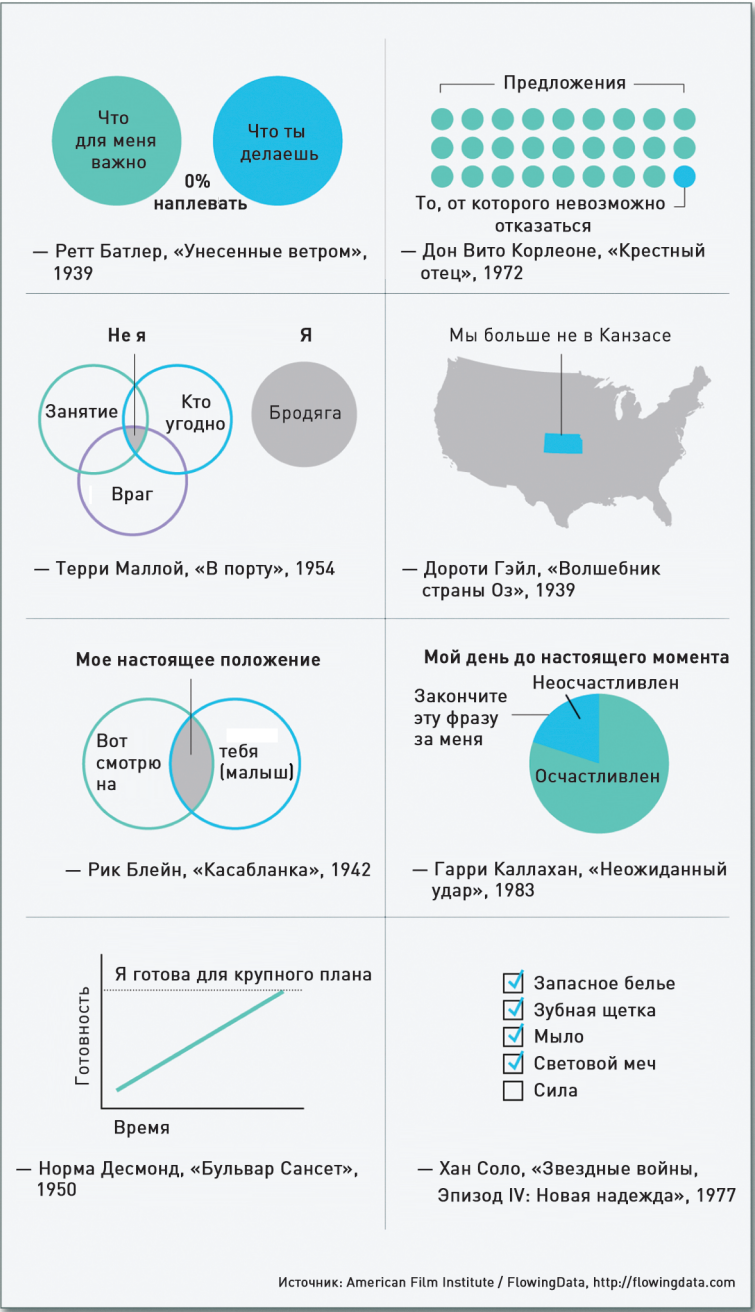
Процент женщин-аспирантов в области
естественных и технических наук,
в процентах



Источник: U.S. Census Bureau

FLOWINGDATA

Рис. 0.4. Графический вид данных из «Статистического ежегодника Соединенных Штатов»



Каждый год Stamen делала что-то новое, но суть была одна: показать, о чем говорят люди в Twitter в реальном времени. Когда в 2009 году во время речи получившей награду Тейлор Свифт (Taylor Swift) произошла известная неприятность с Канье Уэстом (Kanye West)*, через трекер можно было сразу увидеть, что люди об этом думают.

На данном этапе вы подходите к визуализации не столько аналитически, сколько на уровне ощущений. Определение визуализации кажется несколько туманным. На протяжении длительного времени визуализация была чем-то связанным с количественными показателями. От вас требуется своими инструментами выявлять паттерны, а от паттернов — каким-то образом помогать вам в процессе анализа. Визуализация — это не просто получение объективных голых фактов. Как и в случае с трекером Stamen, она почти всегда связана в первую очередь с фактором развлечения. Это способ дать зрителям возможность смотреть шоу с присуждением призов и в процессе взаимодействовать с остальными людьми. Другим прекрасным тому примером может служить творчество Джонатана Харриса (Jonathan Harris). Харрис выстраивает свои работы, такие как «Все у нас хорошо» и «Охота на китов», скорее, вокруг историй, нежели вокруг аналитических прозрений, и эти истории базируются на человеческих эмоциях, вызванных цифрами и аналитикой.

Схемы и диаграммы также переросли состояние просто инструментов и служат теперь средствами коммуникации идей — они способны даже на шутки. Такие сайты, как GraphJam и Indexed, используют диаграммы Венна, секторные диаграммы и пр. для того, чтобы представлять популярные песни или показывать, что комбинирование

Рис. 0.5. Цитаты из кинофильмов в графической форме

* Во время выступления Тейлор Свифт, победившей в номинации за лучшее женское видео, Канье Уэст поднялся на сцену, отобрал у нее микрофон и заявил, что лучшим он считает ролик Бейонсе. Прим. пер.

черного, белого и красного равносильно выпуску коммунистической газеты или убийству панды. «Недозагрузка данных» (Data Underload) — серия юмористических постов, которые я публикую на FlowingData, — это моя собственная «проба пера» в данном жанре. Каждый день я веду наблюдения и выкладываю их в форме диаграмм. На рис. 0.5 таким образом проиллюстрированы цитаты из известных фильмов, вошедших в рейтинги Американского института киноискусства. Получилось нечто абсолютно нелепое, но забавное (по крайней мере, для меня).

Итак, что же такое визуализация? Ну, все зависит от того, с кем вы разговариваете. Некоторые люди утверждают, что визуализация — это традиционные схемы и графики. Другие придерживаются более либеральных взглядов. Для них все, что способно иллюстрировать данные, и есть визуализация, и не имеет значения, что это: произведение data-арта или таблица в Microsoft Excel. Я лично больше склоняюсь к последним, хотя, бывает, иногда обнаруживаю себя среди членов первой группы. В конце концов, все это не так уж и важно. Просто делайте то, что подходит для ваших целей.

На каком бы определении визуализации вы ни остановились, когда вы станете создавать диаграммы для своей презентации, анализировать большой массив цифровой информации или готовить новостной репортаж, содержащий некие данные, вы в конечном счете будете искать правды. В какой-то момент ложь и статистика стали почти синонимами, но на самом деле лгут не числа. Лгут люди, использующие числа. Иногда они делают это намеренно, реализуя какой-то план, но в большинстве случаев подобное происходит неумышленно. Когда вы не знаете, как правильно составить диаграмму или как подать данные беспристрастно, есть большая вероятность, что у вас появится некая информационная свалка, дающая совершенно искривленное представление о реальности. Однако если вы усвоите годные приемы визуализации и научитесь работать с данными, вы сможете уверенно излагать свои идеи и радоваться своим открытиям.

► Другие публикации из серии «Недозагрузка данных» («Data Underload») можно найти на сайте FlowingData по адресу: <http://dataf1.ws/underload>

Как научиться работать с данными

Я начал изучать статистику на первом курсе колледжа. Это был обязательный учебный предмет для получения никак не связанной со статистикой ученой степени по электротехнике. Я слышал много ужасных историй, но, в отличие от описываемых в этих байках преподавателей, мой профессор оказался большим энтузиастом своего дела и занимался им с явным удовольствием. Читая свои лекции, он быстро передвигался вверх-вниз по ступеням зала и размашисто жестикулировал, задевая студентов, мимо которых проходил. До того дня, как мне кажется, у меня не было более вдохновленного учителя, и это, несомненно, повлияло на меня — я увлекся миром данных. В итоге четыре года спустя я закончил колледж специалистом в области статистики.

На протяжении всего моего обучения в бакалавриате статистика сводилась для меня к анализу данных, распределению и проверке гипотез, и мне это нравилось. Было забавно разглядывать

наборы данных и выявлять в них различные тенденции, паттерны и корреляции. Но когда я перешел в магистратуру, мои взгляды изменились, и все стало еще интереснее.

Теперь заниматься статистикой не означало лишь выявлять паттерны и проверять гипотезы (что, как оказалось, во многих случаях не так уж и полезно). Стоп, нет, последние слова я беру назад. Статистика по-прежнему состояла из этого, но тем не менее все воспринималось по-другому. Статистика, как я понял, — это рассказывание историй посредством данных. Вы берете кучку данных, описывающих ваш материальный мир, и анализируете их не только ради того, чтобы найти некие корреляции, но и чтобы выяснить, что же происходит вокруг вас. И когда вы это выясняете, полученные истории оказываются способны помочь вам решить некие проблемы и задачи реального мира — такие как снижение уровня преступности, улучшение здравоохранения и облегчение движения автотранспорта — или хотя бы оставаться человеком, информированным по всем этим вопросам.

Многие люди не видят связи между данными и реальной жизнью. Именно поэтому, когда я сообщал, что учусь в магистратуре на статистика, многие ребята говорили мне, что в колледже ненавидели курс статистики. Я знаю, вы не допустите подобной ошибки, верно? Вы ведь уже начали читать эту книгу.

Как обрести нужные умения, чтобы научиться извлекать пользу из данных? Для этого вы можете прослушать курс лекций, как я, но можете учиться и самостоятельно, опытным путем. Это ведь именно то, чем люди занимаются большую часть времени в магистратуре.

Точно так же обстоят дела и с визуализацией и информационной графикой. Вам не нужно быть великим дизайнером, чтобы делать классную графику. Равно как не нужно быть кандидатом наук в области статистики. Вам достаточно иметь страстное желание учиться и — как это бывает почти с каждым делом в жизни — практиковаться, чтобы совершенствоваться.

Первую диаграмму с данными я создал классе в четвертом. Сделал я ее, участвуя в школьной ярмарке научных проектов. Мой партнер по проекту и я исследовали (очень углубленно, можете не сомневаться), по какой поверхности улитки передвигаются быстрее всего. Мы ставили улиток на гладкие и шершавые поверхности и засекали время, чтобы посмотреть, сколько минут им понадобится, чтобы пройти определенное расстояние. Таким образом мы получили данные о времени для различных поверхностей. На их основе я создал столбцовую диаграмму. Я не помню, догадался ли я выстроить столбцы по величине от наименьшего к наибольшему, но зато я хорошо помню, как боролся с программой Excel. Однако на следующий год, когда мы выяснили, что именно предпочитают есть мучные каштановые хрущаки, диаграмма получилась просто блестящей. Когда вы освоите основной набор функциональных возможностей и научитесь работать с программным обеспечением, разобраться с остальным не составит трудности. И если это не отличный пример обучения на собственном опыте, тогда я не знаю, что еще может быть примером. Кстати, быстрее всего улитки передвигались по стеклу, а каштановые хрущаки предпочитали сухой завтрак Grape Nuts — на тот случай, если вам это интересно.

Здесь мы будем говорить о самых основополагающих моментах, но по сути своей процесс выглядит аналогично, какую бы программу или язык программирования вы ни взяли изучать.

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. Данная книга поможет вам в этом.

Я говорю вам это, потому что именно так я и учился. Помню, как я впервые углубился в дизайнерские аспекты визуализации. Было это летом после второго курса магистратуры, когда я получил потрясающее известие: меня взяли на стажировку в должности редактора графики в New York Times. До того момента графика для меня всегда была лишь инструментом анализа (в том числе и столбцовые диаграммы для школьной ярмарки научных проектов), а эстетика и дизайн не имели особо большого значения — если они вообще имели для меня хоть какое-нибудь значение. Мне и в голову не приходило, сколь велика роль данных в журналистике.

Чтобы подготовиться к стажировке, я прочитал все книги по дизайну, которые смог достать, и проштудировал руководство пользователя Adobe Illustrator, потому что знал: в New York Times работают именно с этой программой. Но только когда я действительно начал делать диаграммы и графики, я стал по-настоящему чему-то учиться. Обучаясь в процессе работы, вы окажетесь вынуждены очень быстро набираться знаний и навыков, необходимых в деле. И по мере того, как вы будете работать со все большими массивами данных и создавать все больше диаграмм, ваши умения станут развиваться активнее.

Как читать эту книгу

Эта книга построена на примерах и написана с целью дать вам знания, необходимые для создания различных объектов информационной графики от начала и до конца. Вы можете прочитать ее от корки до корки, а можете выхватить только нужные вам идеи, если у вас уже есть какие-то данные или если вы владеете какими-то способами визуализации. Главы организованы так, что примеры в них независимы и самодостаточны. Если вы новичок в этой области, первые главы будут вам особенно полезны. В них рассказывается о том, как подходить к имеющимся данным, что в них искать и какими инструментами пользоваться. Вы узнаете, где именно находить данные и как их форматировать и готовить для визуализации. Далее представлены приемы визуализации, структурированные по типу данных и по типу историй, которые с их помощью вы можете рассказать. Но помните: говорить всегда должны сами данные.

Какой бы подход к чтению этой книги вы ни выбрали, я горячо рекомендую вам читать ее, сидя за работающим компьютером, чтобы прорабатывать примеры шаг за шагом и просматривать все ресурсы, упомянутые в примечаниях и ссылках. Вы можете также скачать коды и файлы данных и опробовать работающие демоверсии на сайтах <http://www.wiley.com/go/visualizethis> и <http://book.flowingdata.com>.

Чтобы сделать сказанное выше предельно ясным, на рис. 0.6 я представляю вам схему, которая поможет понять, что именно вам необходимо. Развлекайтесь в свое удовольствие!

* R — язык программирования для статистической обработки данных и работы с графикой; свободная программная среда с открытым исходным кодом. *Прим. пер.*

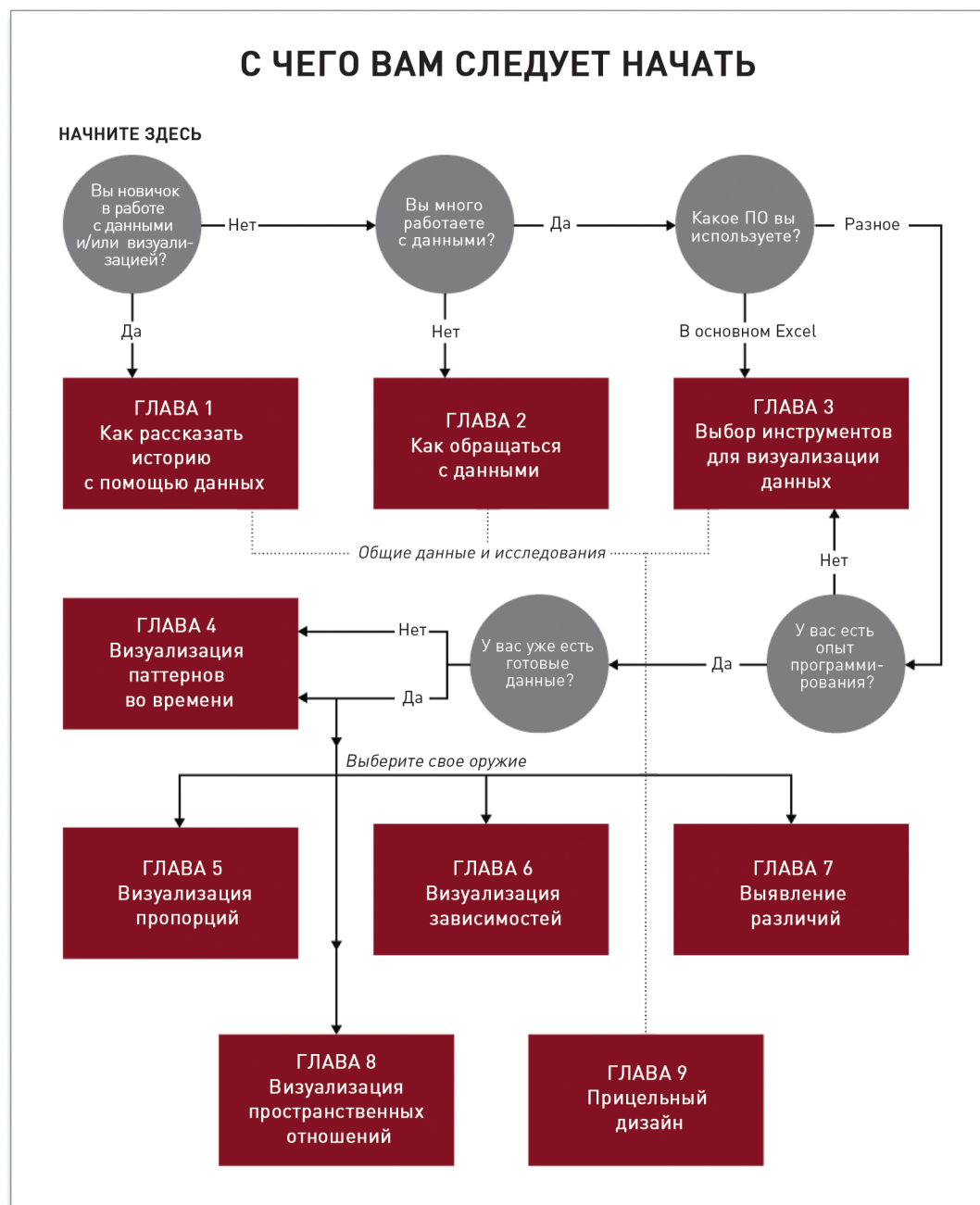


Рис. 0.6. С какого места следует начать читать эту книгу