Визуализация больших данных

Максим Губин

Томск

econophysica

На этой лекции будут рассмотрены

- Введение в визуализацию данных;
- Особенности визуализации больших данных;
- Kibana;
- Matplotlib.

Вступление

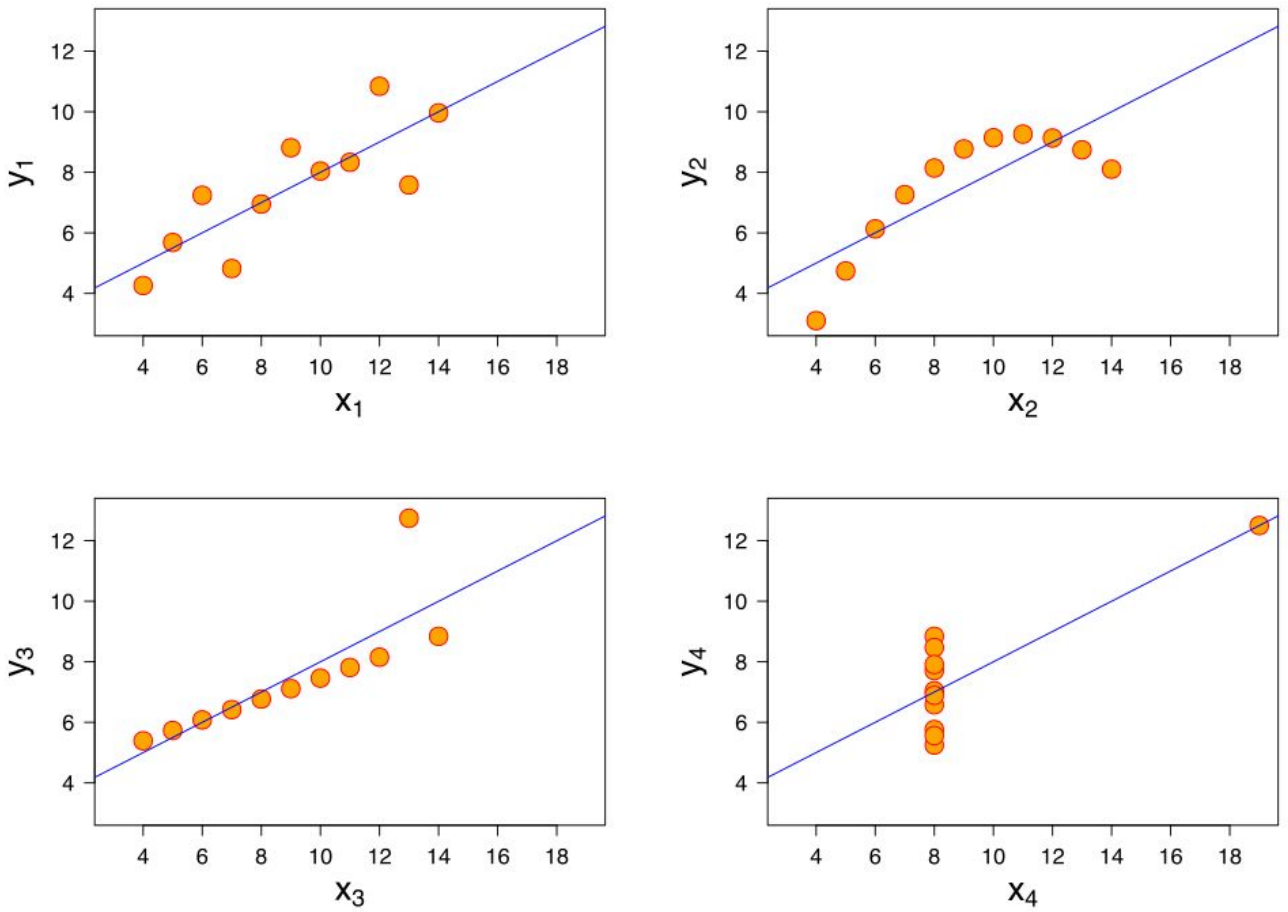
- Визуализация данных является ключевой частью любого рабочего процесса в науке о данных, но ее часто рассматривают как неудобный дополнительный шаг в отчете о результатах анализа.
- Принимать такую позицию – ошибочно: картинка стоит тысячи слов.

Зачем нужно визуализировать данные?

- Чтобы четко и эффективно донести информацию до пользователей;
- Чтобы помочь пользователям анализировать и рассуждать о данных и знаниях;
- Чтобы сделать сложные данные более доступными, понятными и удобными для использования;
- Чтобы помочь выявить закономерности, понять идеи, изучить источники данных.

Да, это все о пользователях. Но дело не в подаче менеджерам. Речь идет о повседневном анализе и принятии решений.

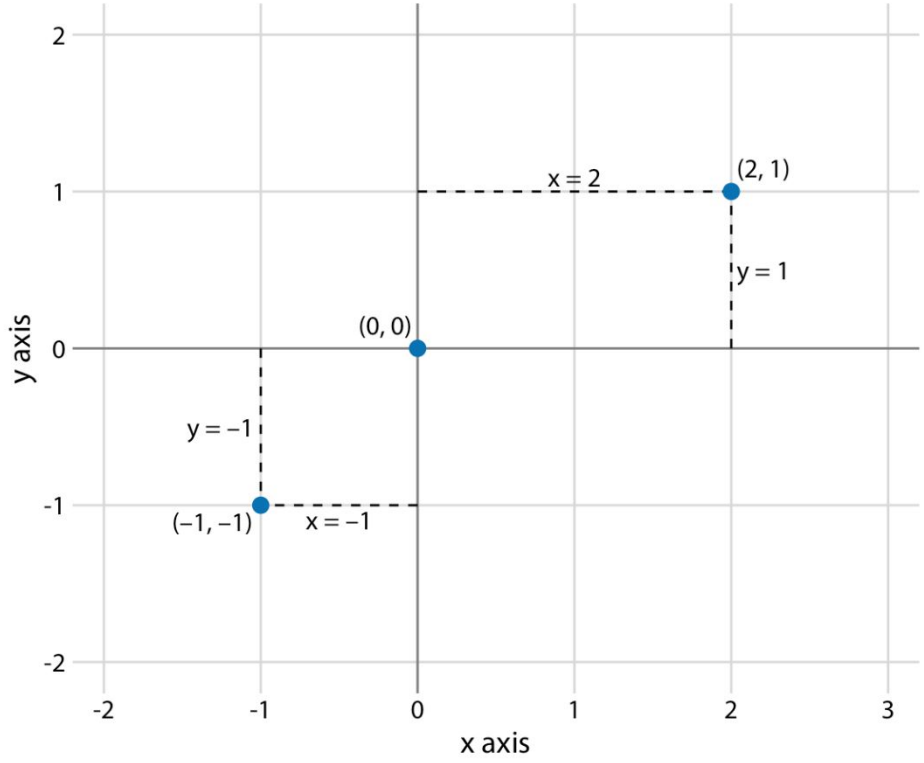
Квартет Энскомба



Параметр	Значение	Точность
Среднее x	9	точно
Дисперсия x	11	точно
Среднее y	7.50	до 2 знаков после запятой
Дисперсия y	4.125	+/- 0.003
Корреляция между x и y	0.816	до 3 знаков после запятой
Формула линейной регрессии	y = 3.00 + 0.500x	до 2 и 3 знаков после запятой соответственно
Коэффициент детерминации линейной регрессии	0.67	до 2 знаков после запятой

econophysica

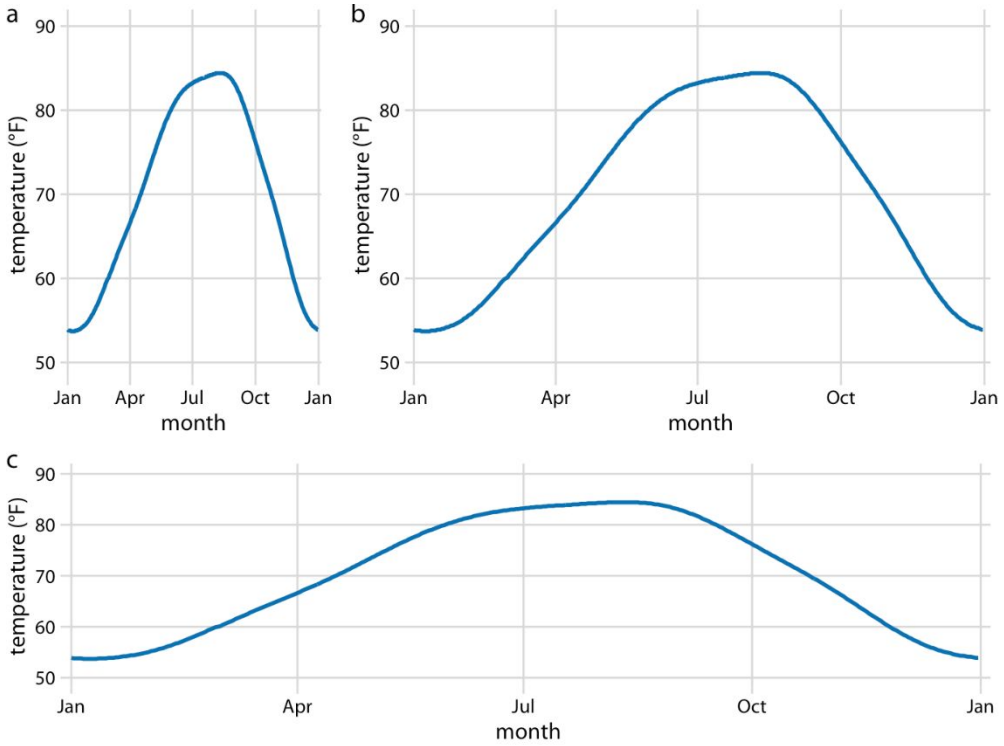
Декартовы координаты

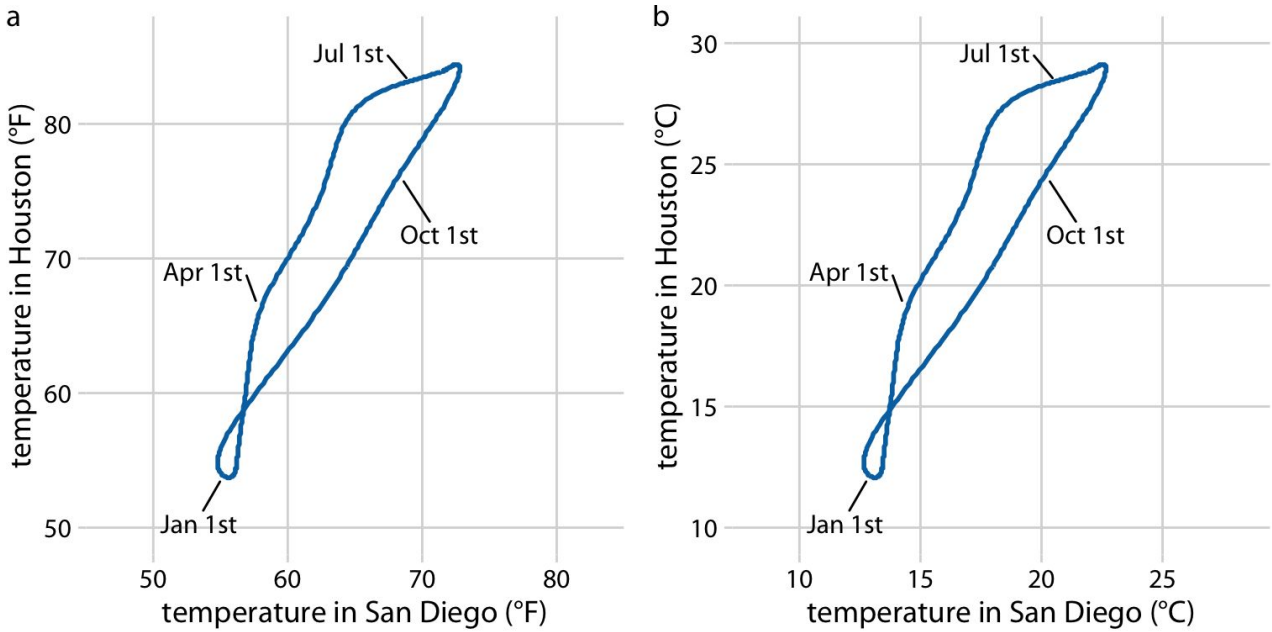


- 2 оси
- Единицы измерения для каждой оси
- Координатная сетка
- Легенда

econophysica

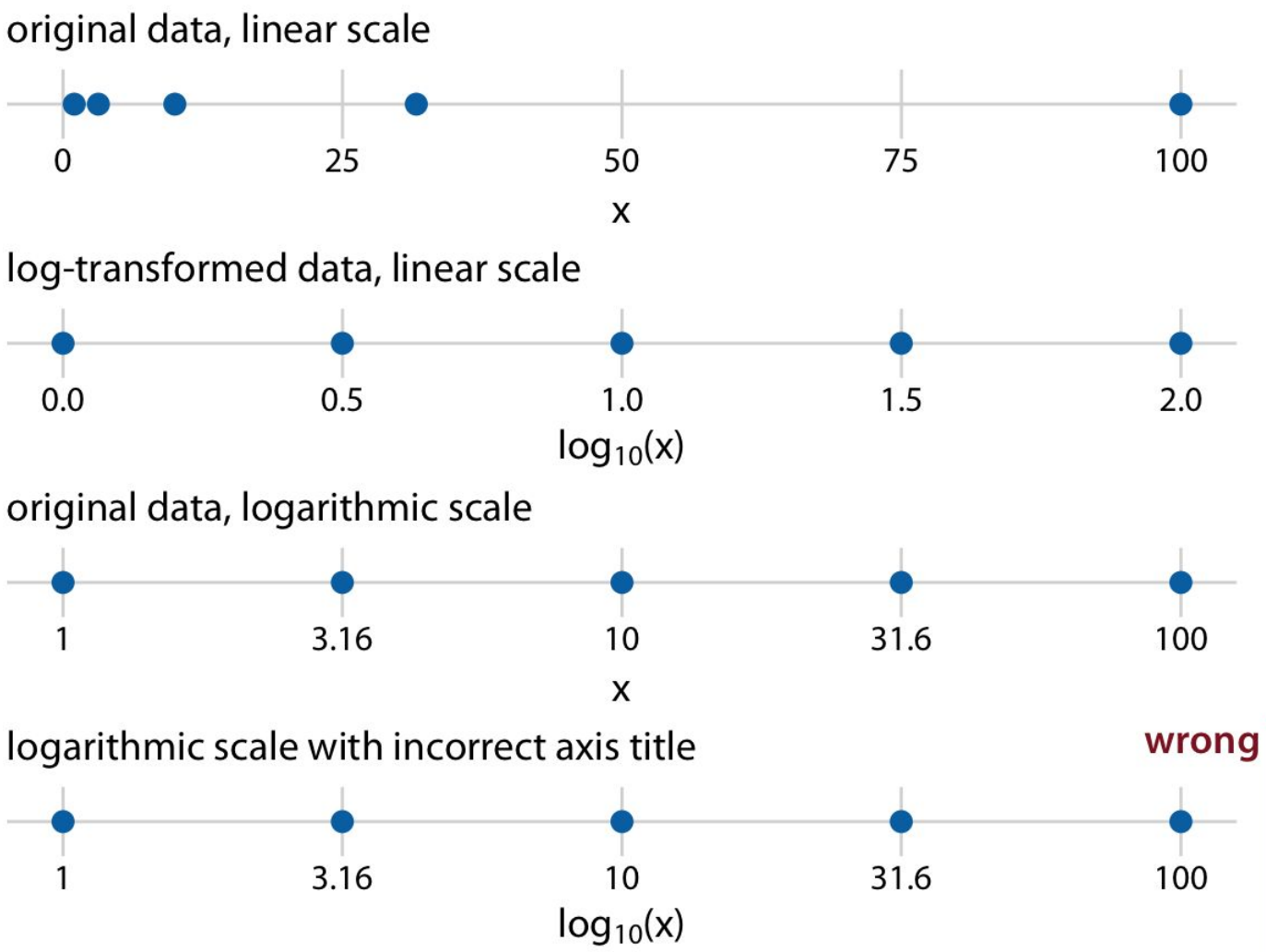
Декартовы координаты





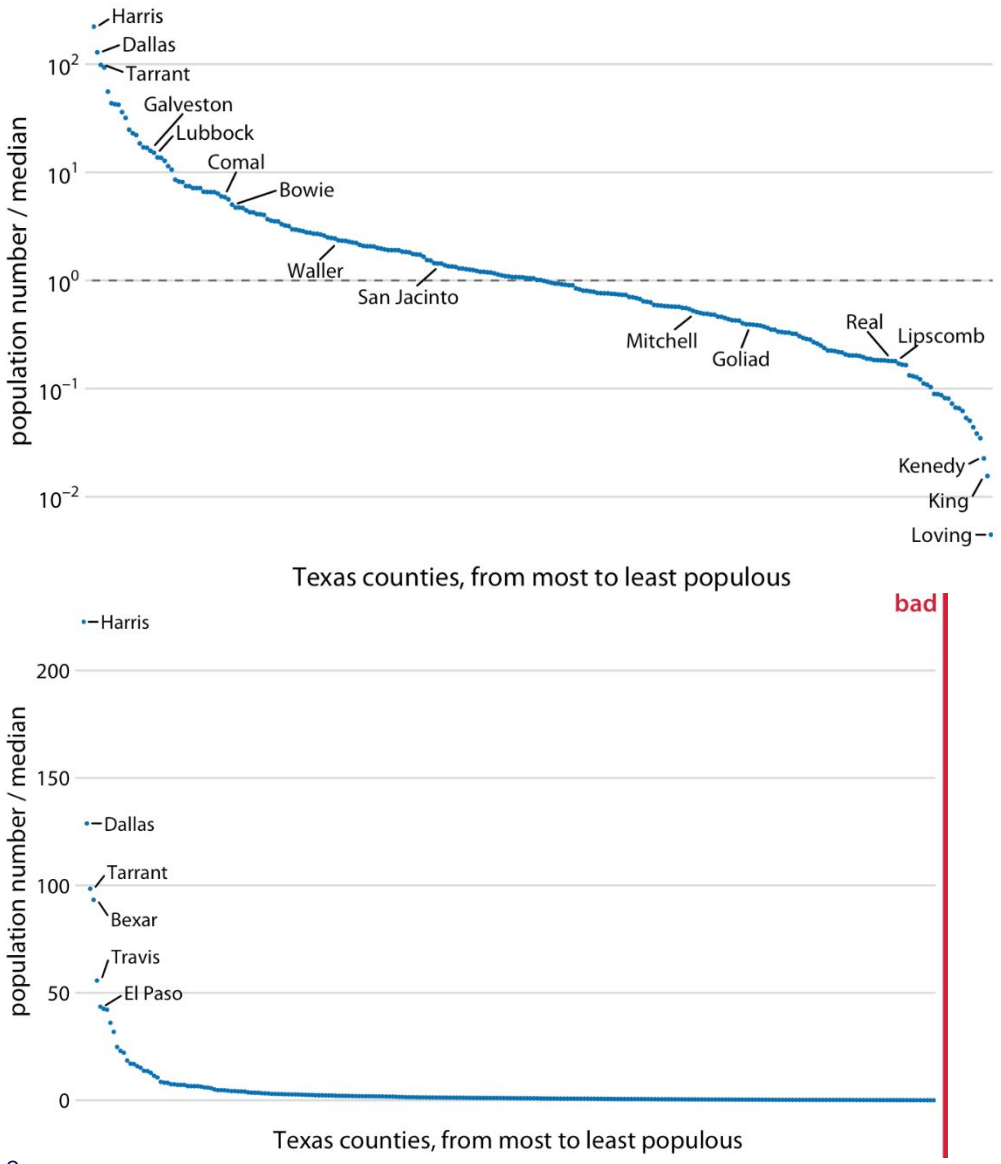
econophysica

Декартовы координаты



Обращайте внимание на единицы изменения данных, особенно при использовании логарифмической шкалы.

Декартовы координаты



Логарифмические шкалы часто используются, когда набор данных содержит числа очень разных величин. В округах Техаса, показанных на рисунках, в самом густонаселенном округе (Харрис) согласно переписи 2010 года было 4092459 человек, а в наименее густонаселенном (Лавинг) - 82.

econophysica

Цветовые шкалы: качественные

- Мы часто используем цвет как средство для различения отдельных предметов или групп, которые не имеют внутреннего порядка, таких как разные страны на карте или разные производители определенного продукта. В этом случае мы используем качественную цветовую шкалу.
- Такая шкала содержит конечный набор определенных цветов, которые выбираются так, чтобы они выглядели четко отличными друг от друга, а также были эквивалентны друг другу.
- Второе условие требует, чтобы ни один цвет не выделялся относительно других.
- И цвета не должны создавать впечатление порядка, как в случае с последовательностью цветов, которые становятся последовательно све

econophysica

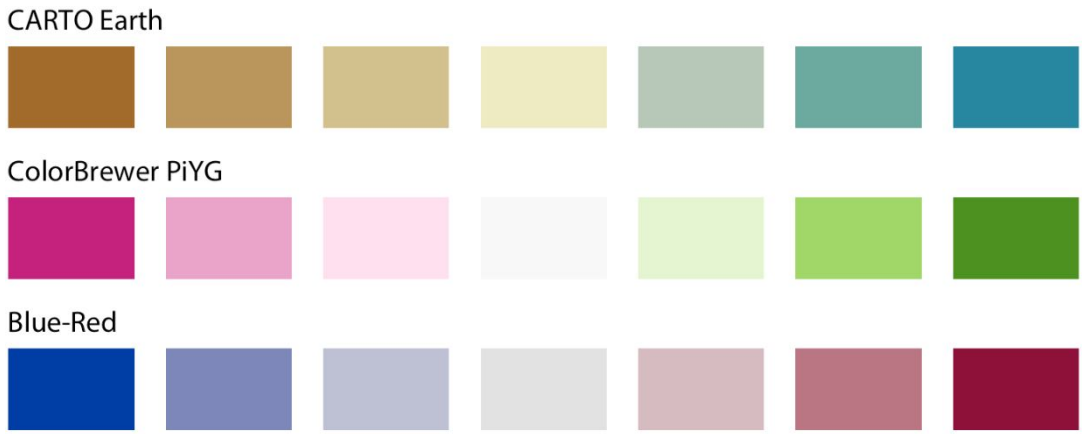
Цветовые шкалы: последовательные

- Цвет также можно использовать для представления численных данных, таких как доход, температура или скорость.
- В этом случае мы используем последовательную цветовую шкалу.
- Такая шкала содержит последовательность цветов, которые четко указывают (i), какие значения больше или меньше, чем другие, и (ii) насколько два конкретных значения удалены друг от друга.
- Второй момент подразумевает, что цветовая гамма должна восприниматься как равномерно изменяющаяся во всем диапазоне.

econophysica

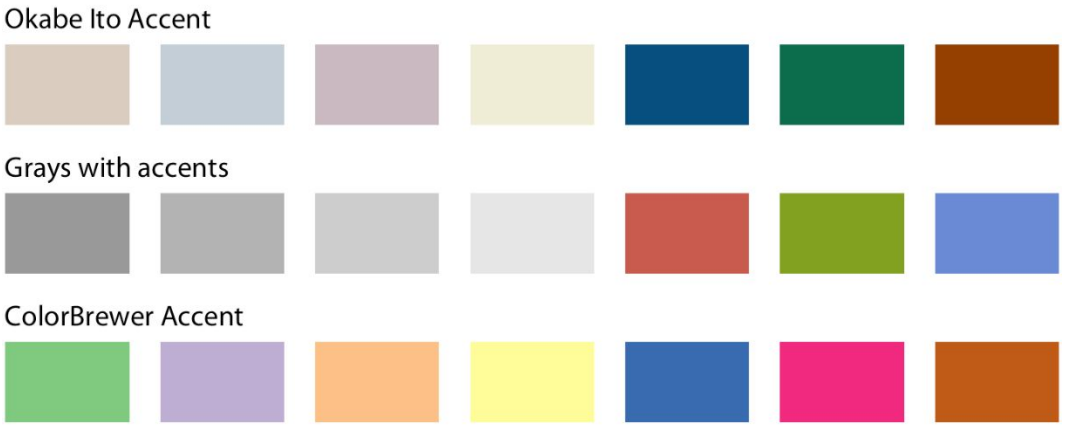
Цветовые шкалы: расходящиеся

- В некоторых случаях нам нужно визуализировать отклонение значений данных в одном из двух направлений относительно нейтральной средней точки.
- Одним простым примером является набор данных, содержащий как положительные, так и отрицательные числа. Мы можем захотеть показать те, которые имеют разные цвета, чтобы сразу стало ясно, является ли значение положительным или отрицательным, а также то, как далеко в любом направлении оно отклоняется от нуля.
- Подходящей цветовой шкалой в этой ситуации является расходящаяся цветовая гамма.
- Мы можем думать о расходящейся шкале как о двух последовательных шкалах, соединенных вместе в общей средней точке, которая обычно представлена светлым цветом.



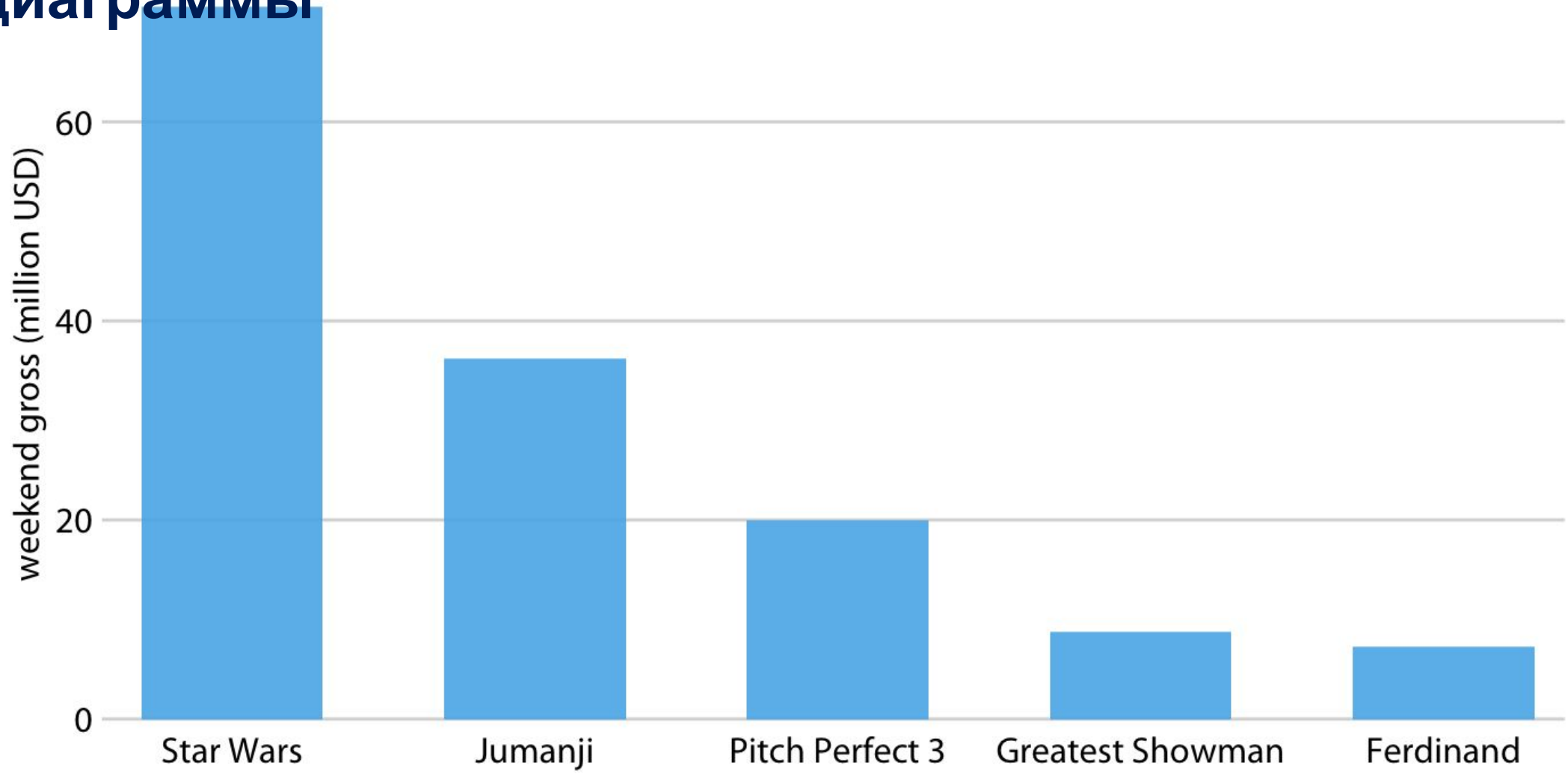
Цветовые шкалы: акцент

- Цвет также может быть эффективным инструментом для выделения определенных элементов в данных.
- В наборе данных могут быть определенные категории или значения, которые содержат ключевую информацию об истории, которую мы хотим рассказать, и мы можем усилить историю, выделив читателю соответствующие элементы фигуры.
- Простой способ добиться этого - раскрасить элементы фигуры цветом или набором цветов, которые ярко выделяются на фоне остальной части фигуры. Этот эффект может быть достигнут с помощью **акцентирующих** цветовых шкал, которые представляют собой цветовые шкалы, которые содержат как набор приглушенных цветов, так и соответствующий набор более сильных, темных и / или более насыщенных цветов.

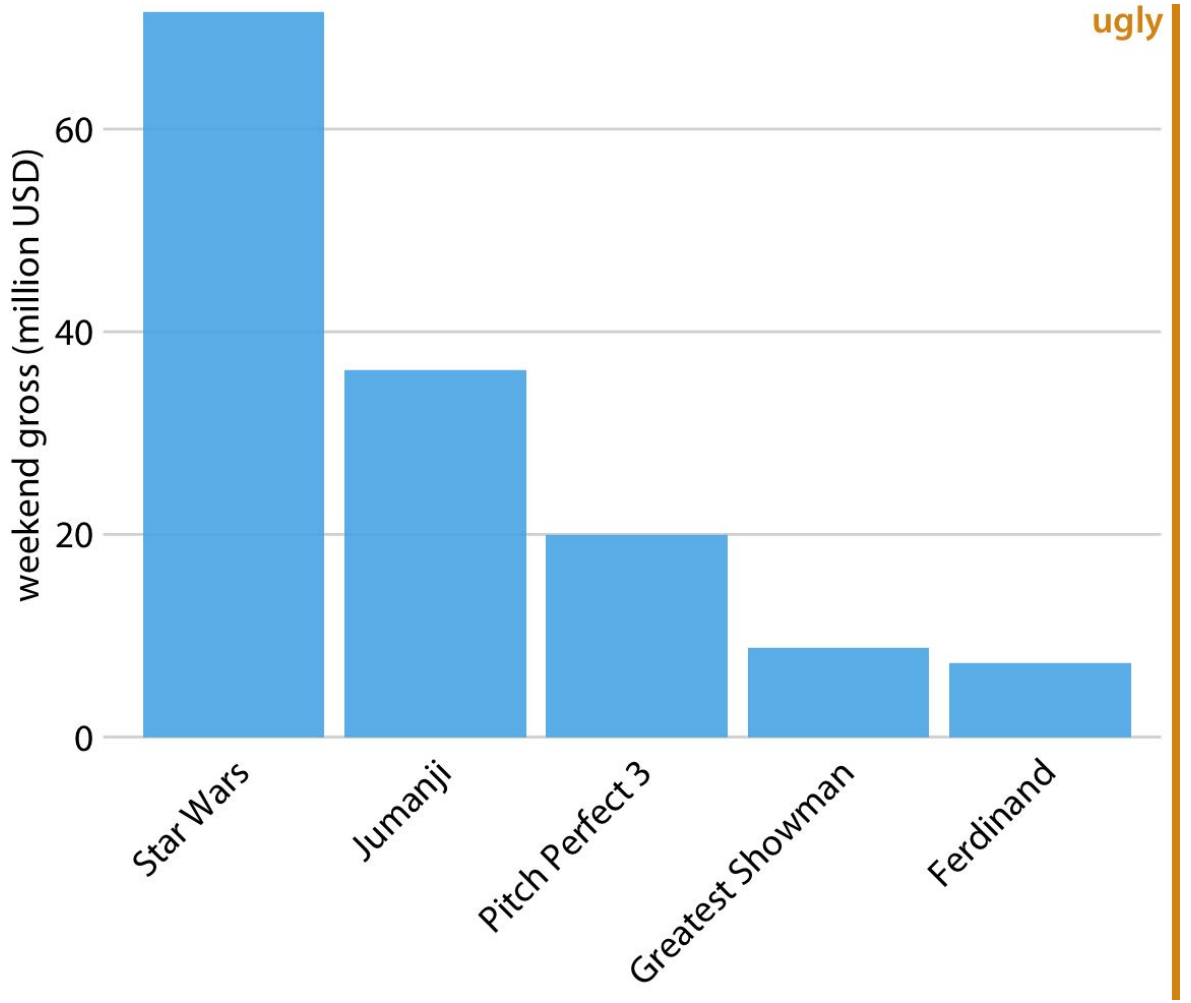


econophysica

Визуализация численных величин: столбчатые диаграммы

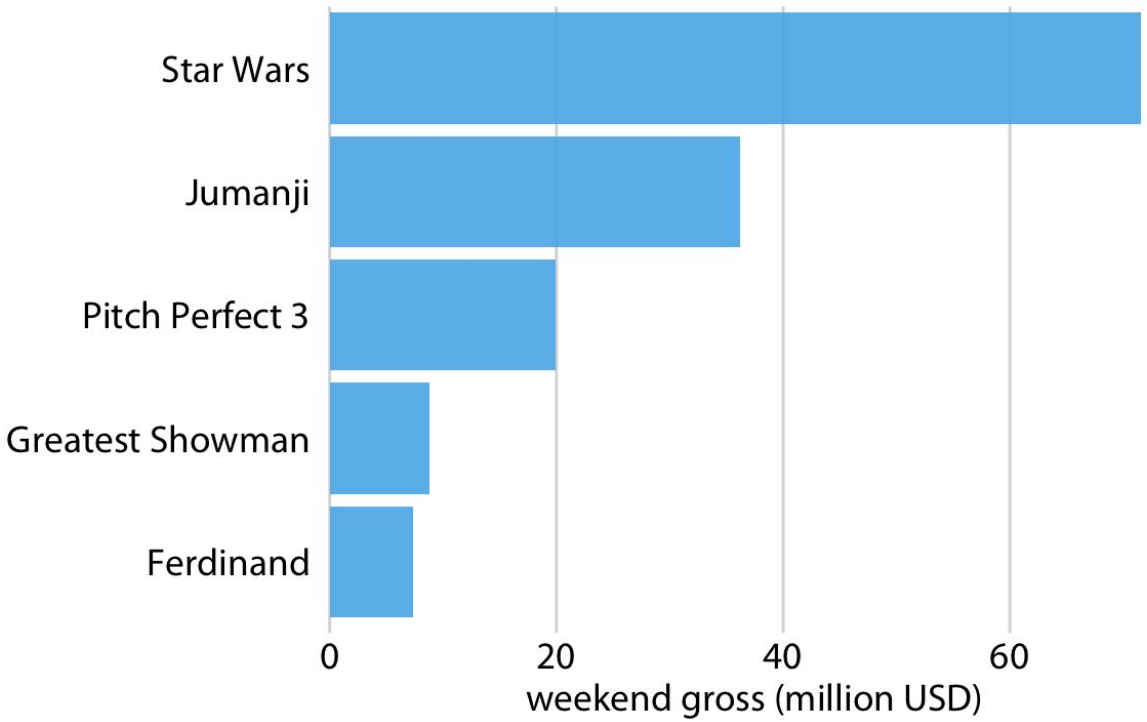


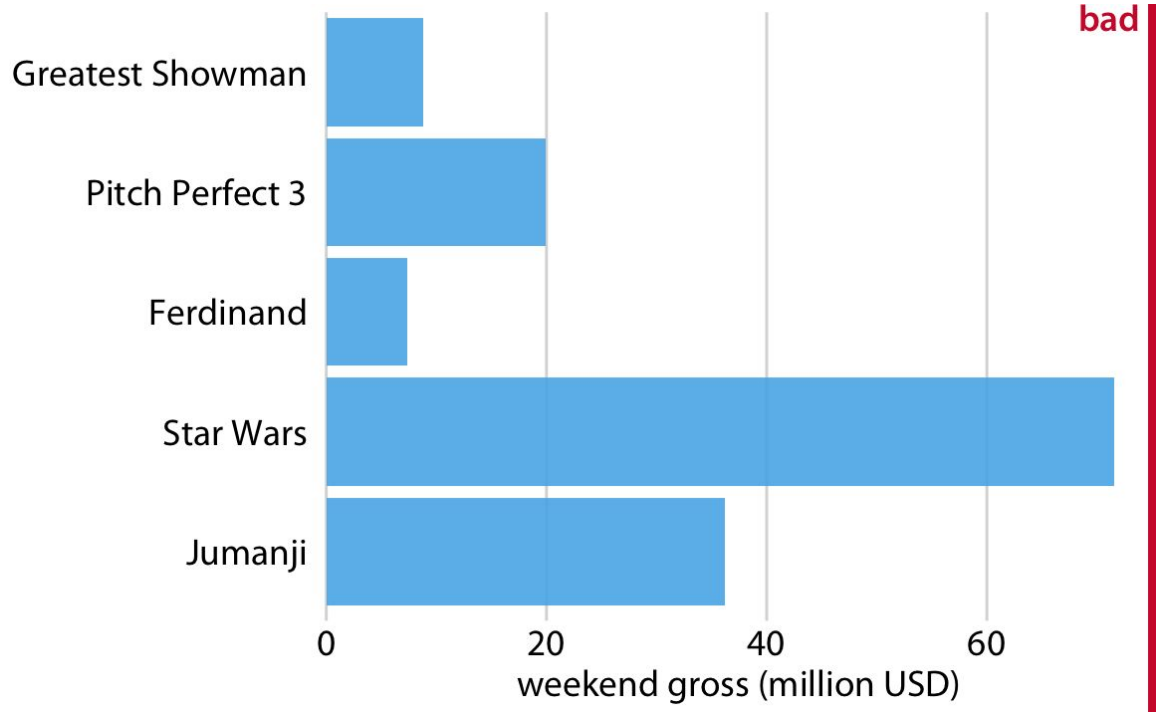
Визуализация численных величин: столбчатые диаграммы



econophysica

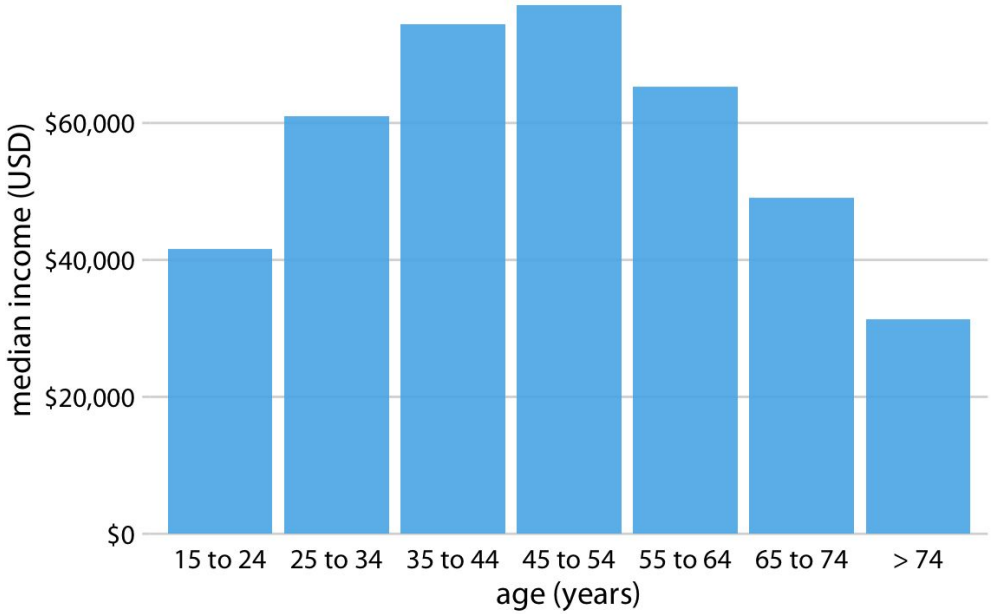
Визуализация численных величин: столбчатые диаграммы

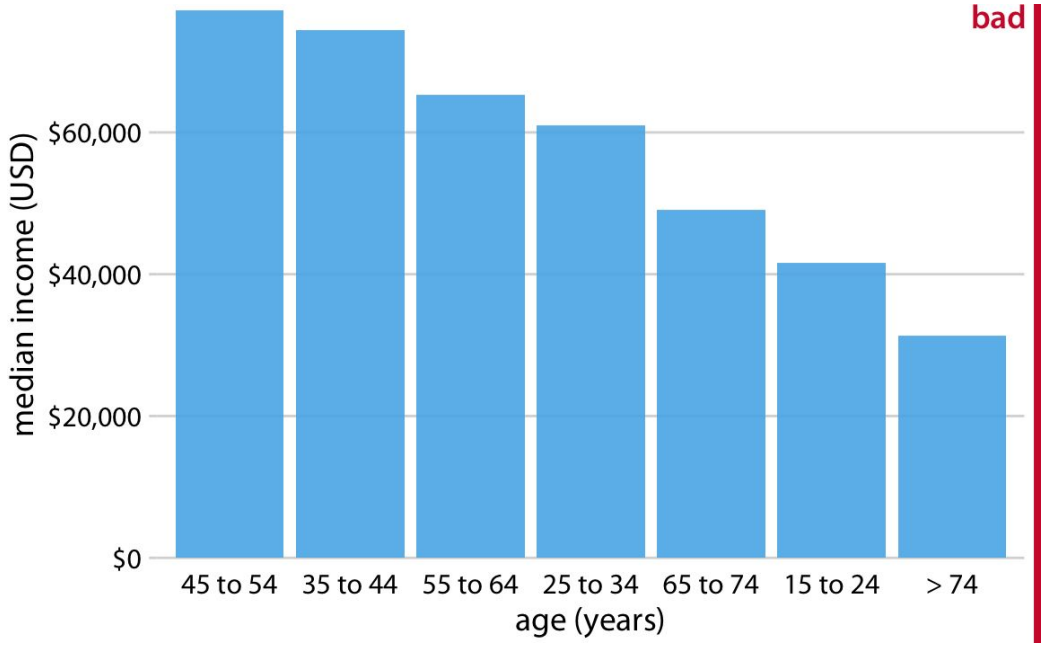




econophysica

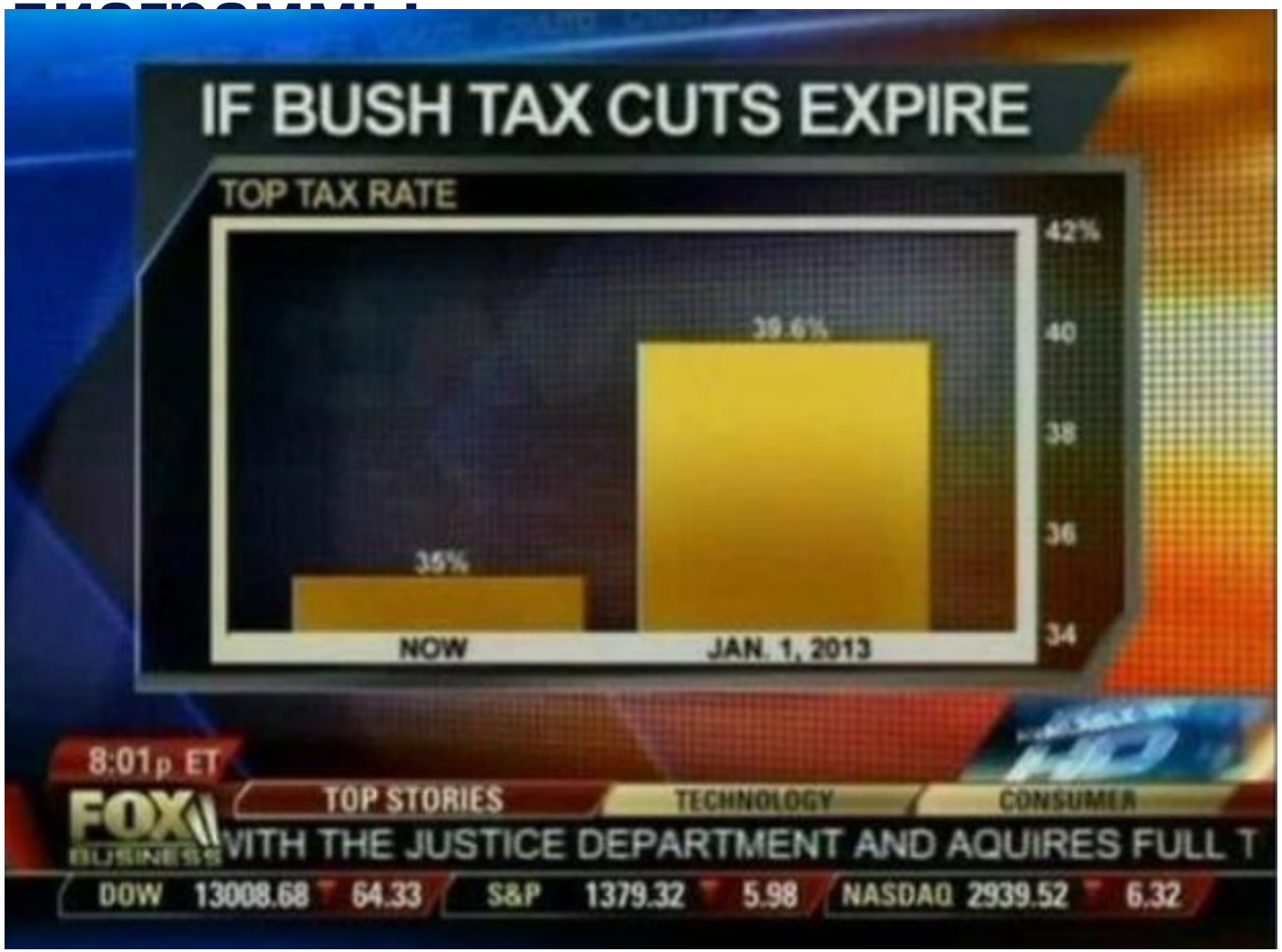
Визуализация численных величин: столбчатые диаграммы

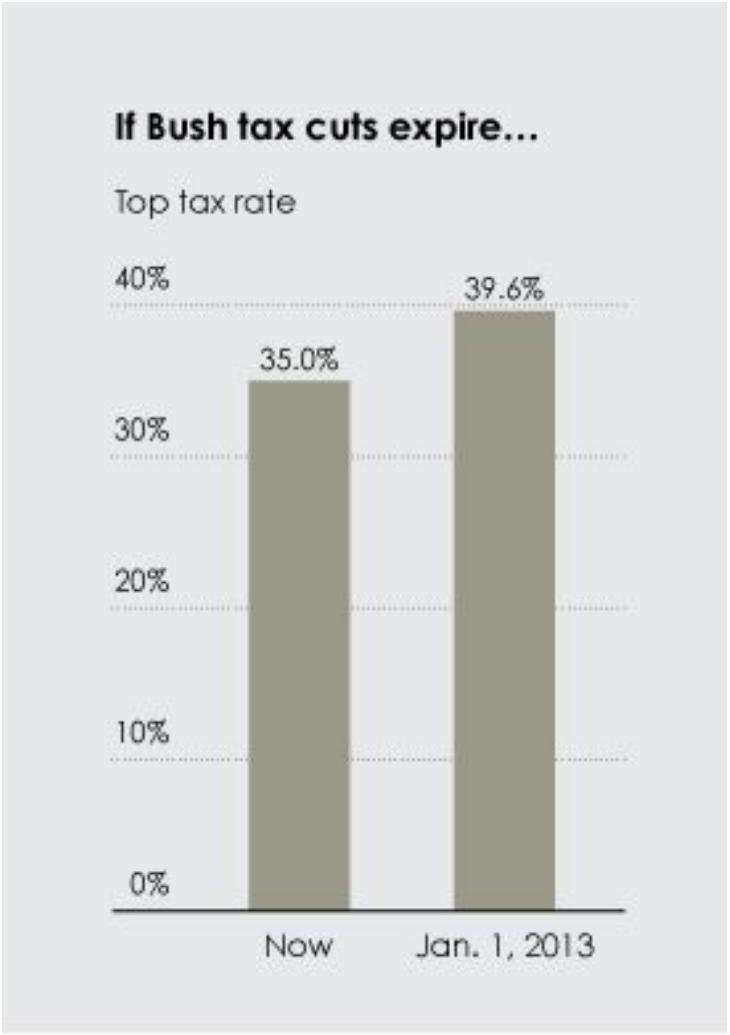




econophysica

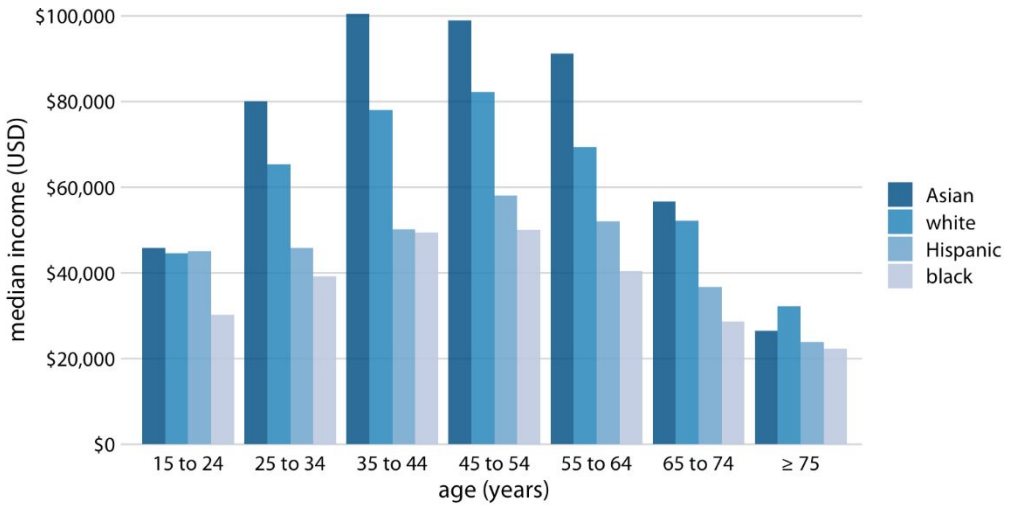
Визуализация численных величин: столбчатые

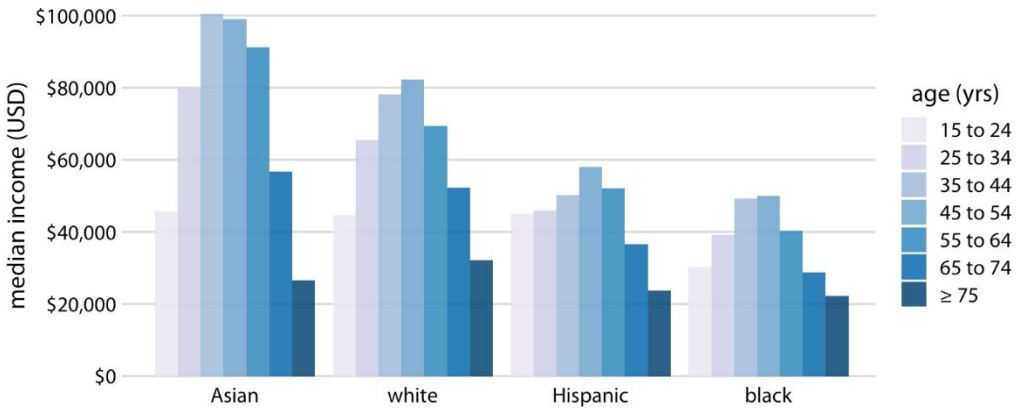




econophysica

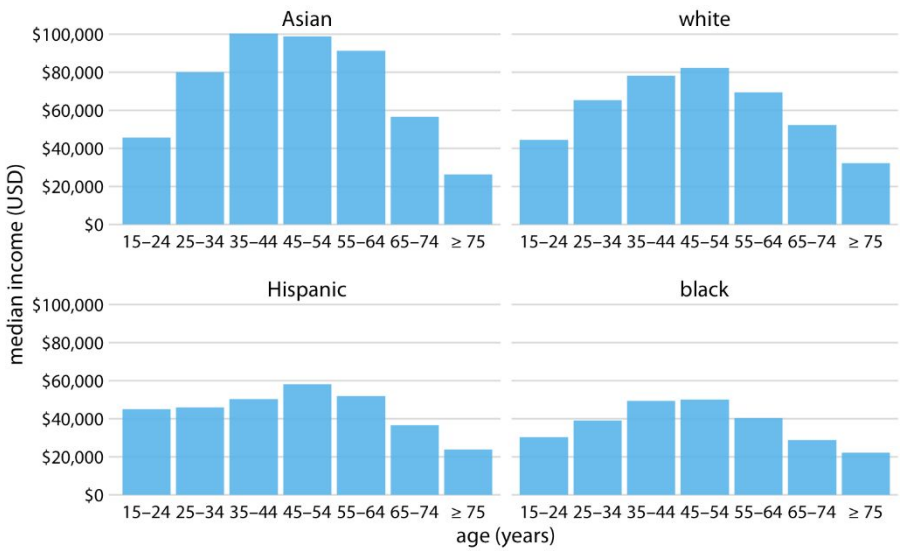
Комбинации столбчатых диаграмм

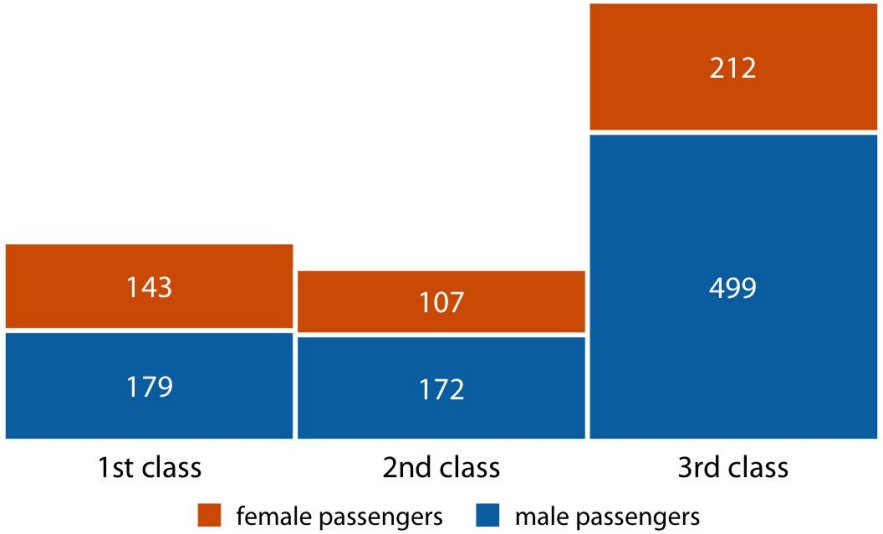




- Столбчатые диаграммы с группировкой позволяют показать множество данных за раз, но их труднее воспринимать.
- Всегда выбирайте условие группировки, связанное с историей, которую вы хотите показать вашей визуализацией.

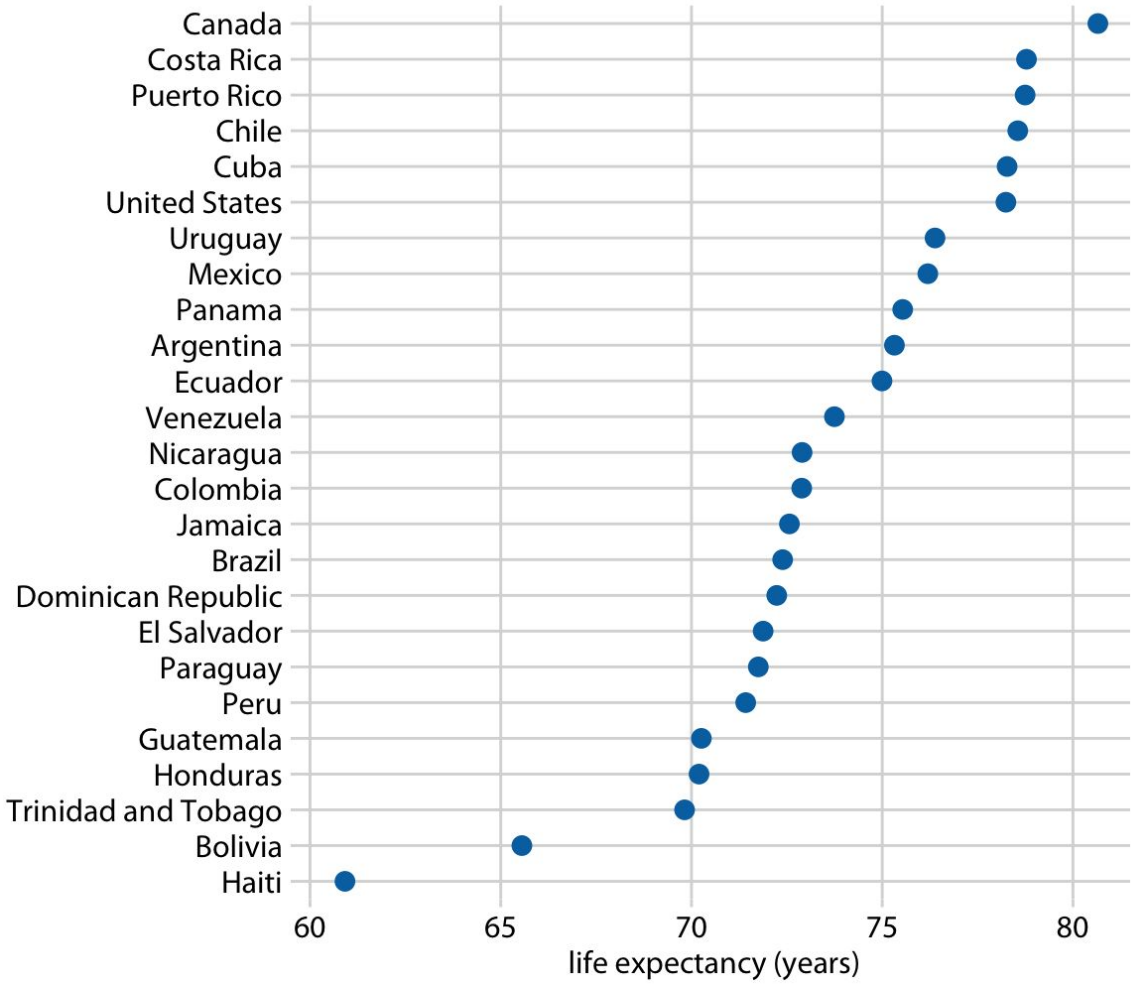
Комбинации столбчатых диаграмм

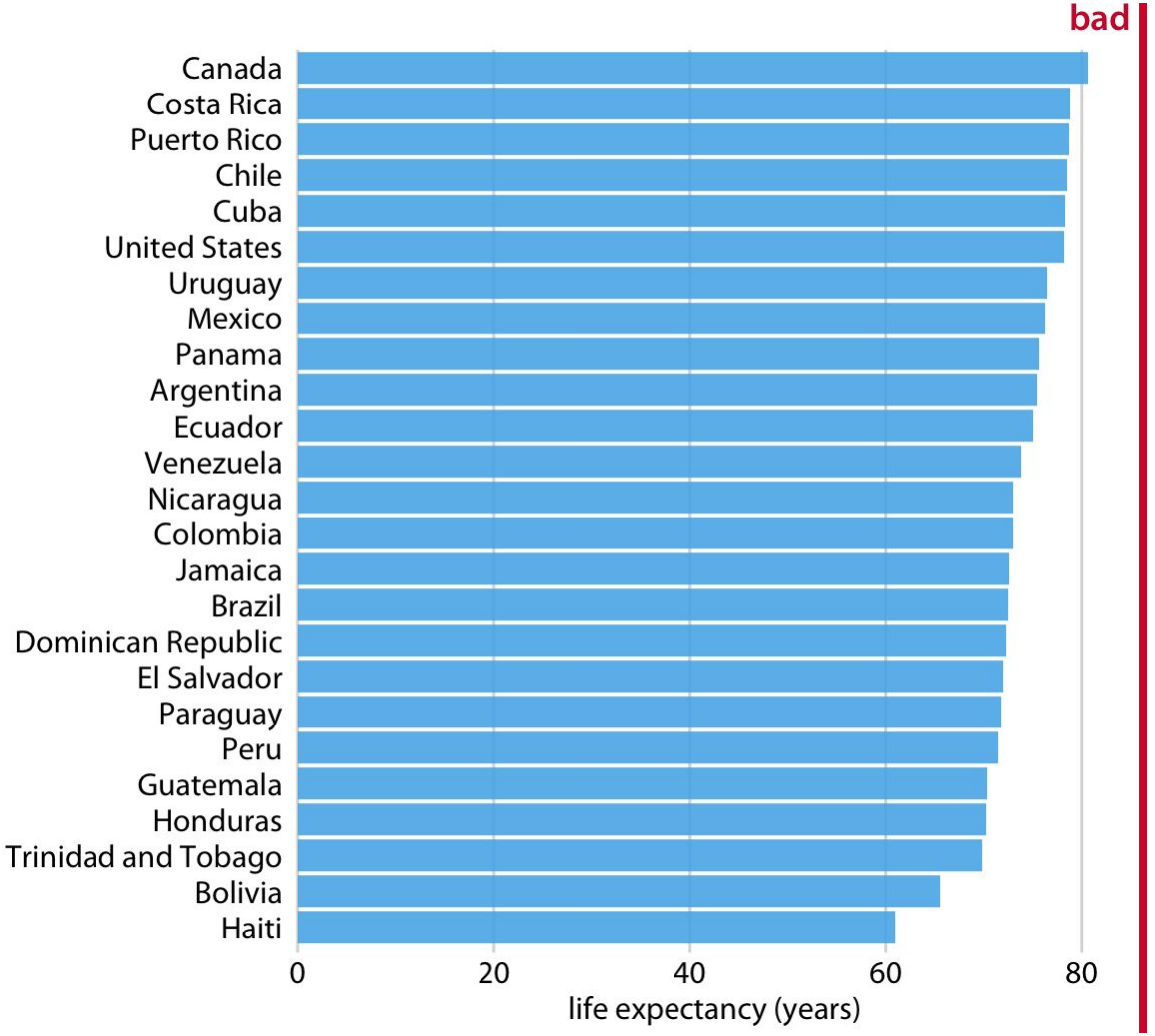




- Иногда проще воспринять несколько столбчатых диаграмм, чем одну сгруппированную.
- Вместо того, чтобы рисовать группы столбцов рядом, иногда предпочтительнее складывать столбцы друг на друга. Такое объединение полезно, когда сумма величин, представленных отдельными столбцами, сама по себе является значимой.

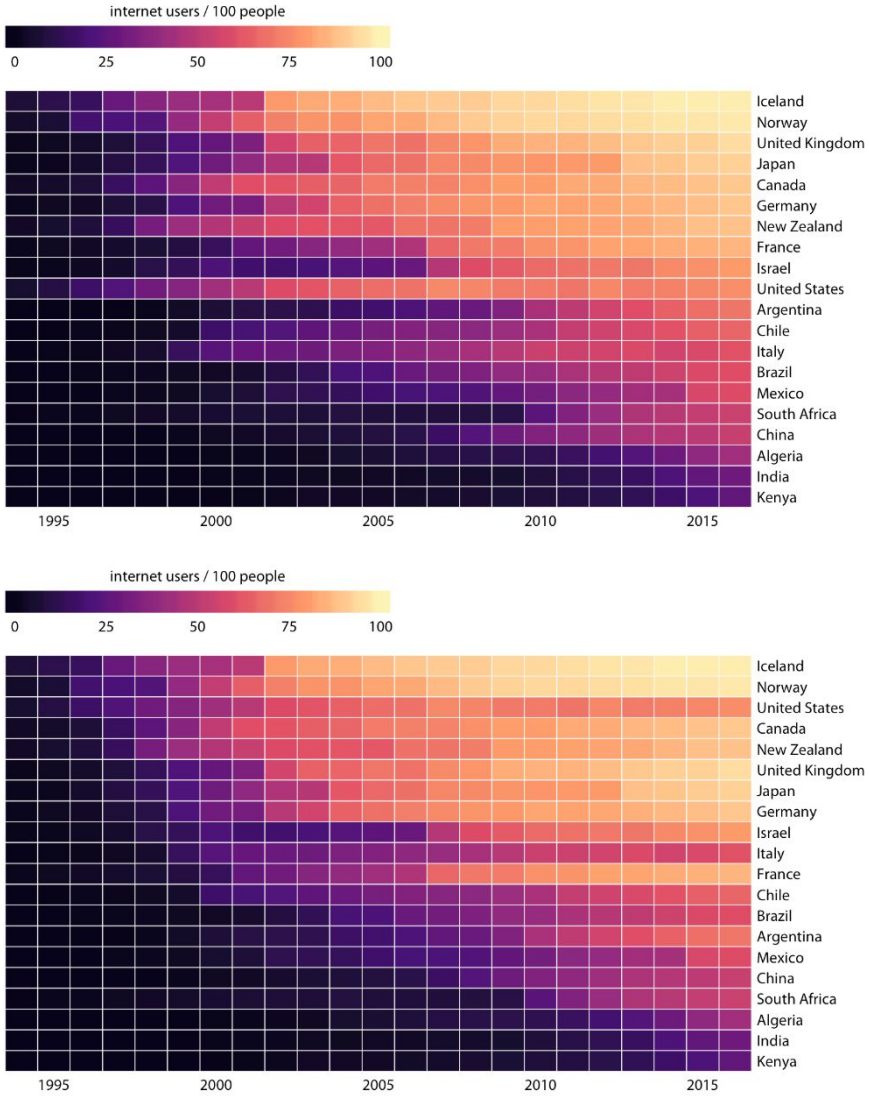
Точечные графики





econophysica

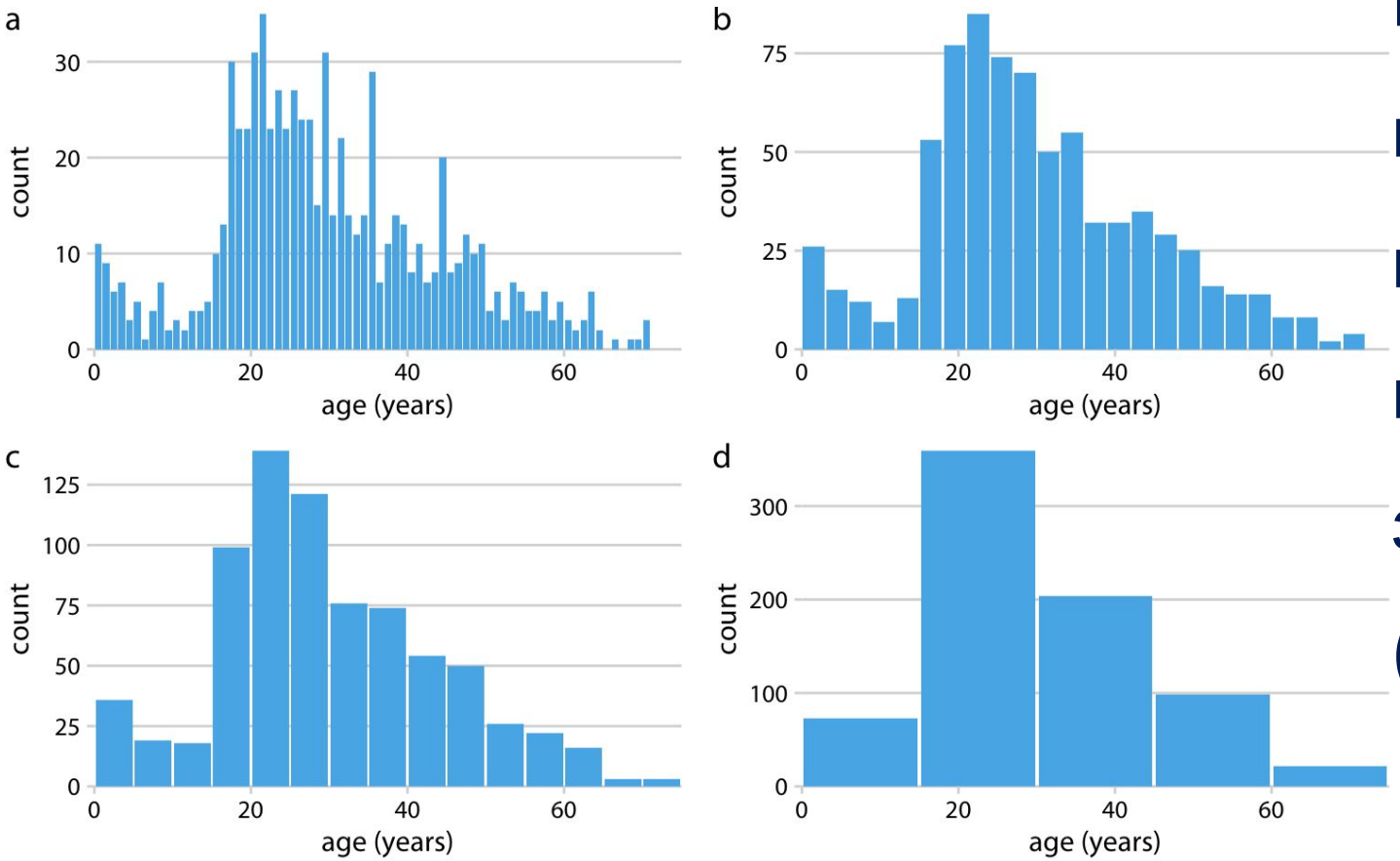
Тепловые карты



- Тепловые карты позволяют отобразить динамику развития данных, подобно множеству графиков, но с сохранением читаемости.
- Для тепловых карт также критически важно выбрать условие упорядочивания строк, связанное с теми особенностями данных, на которые вы хотели бы обратить внимание.

econophysica

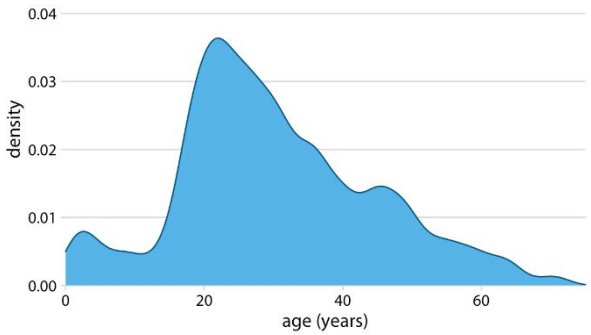
Гистограммы



При создании гистограммы всегда нужно попробовать несколько возможных значений ширины столбца (bin).

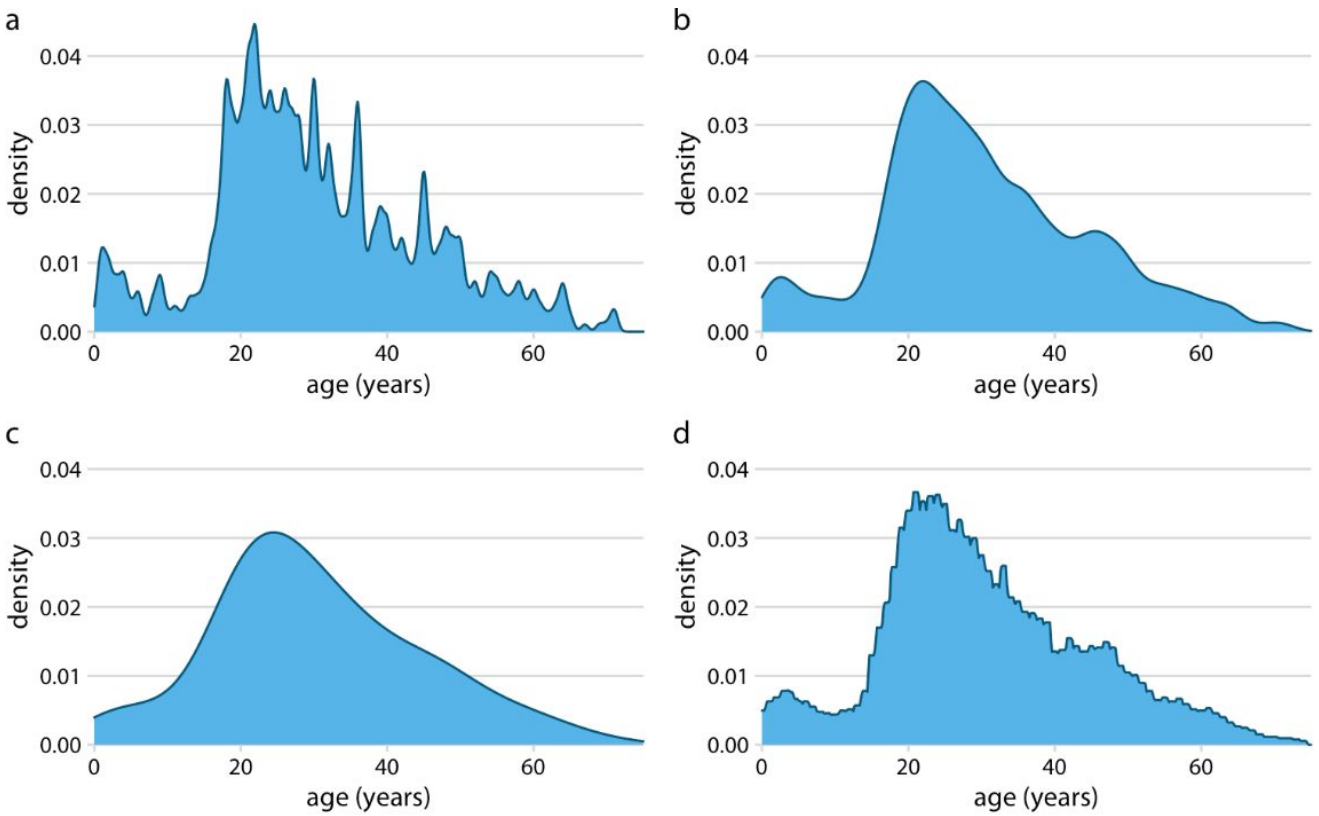
econophysica

Диаграмма плотности



- На графике плотности мы пытаемся визуализировать базовое распределение вероятности данных, рисуя соответствующую непрерывную кривую.
- Эта кривая должна быть построена из данных, и наиболее часто используемый метод для этой процедуры оценки называется оценкой плотности ядра.
- При оценке плотности ядра мы рисуем непрерывную кривую (ядро) с небольшой шириной (контролируемой параметром, называемым полосой пропускания) в местоположении каждой точки данных, а затем складываем все эти кривые, чтобы получить окончательную оценку плотности.
- Наиболее широко используемым ядром является ядро, но есть много других вариантов.

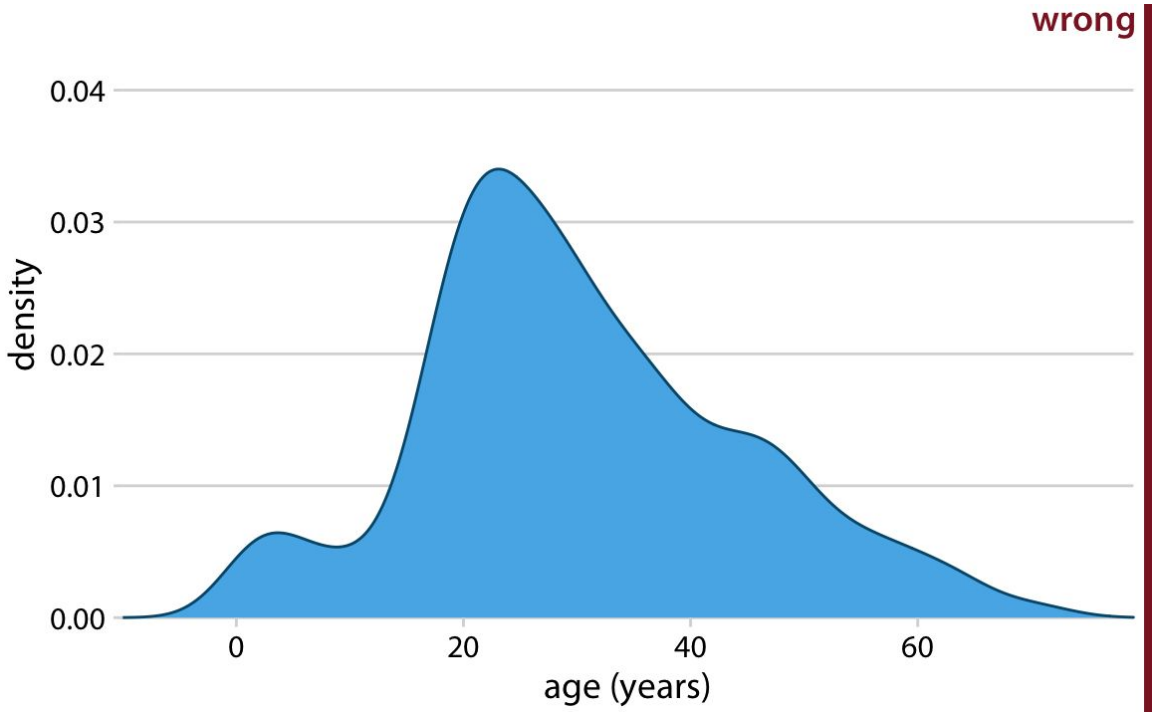
Диаграмма плотности



При создании диаграмм плотности всегда нужно попробовать несколько возможных значений полосы пропускания (bandwith).

econophysica

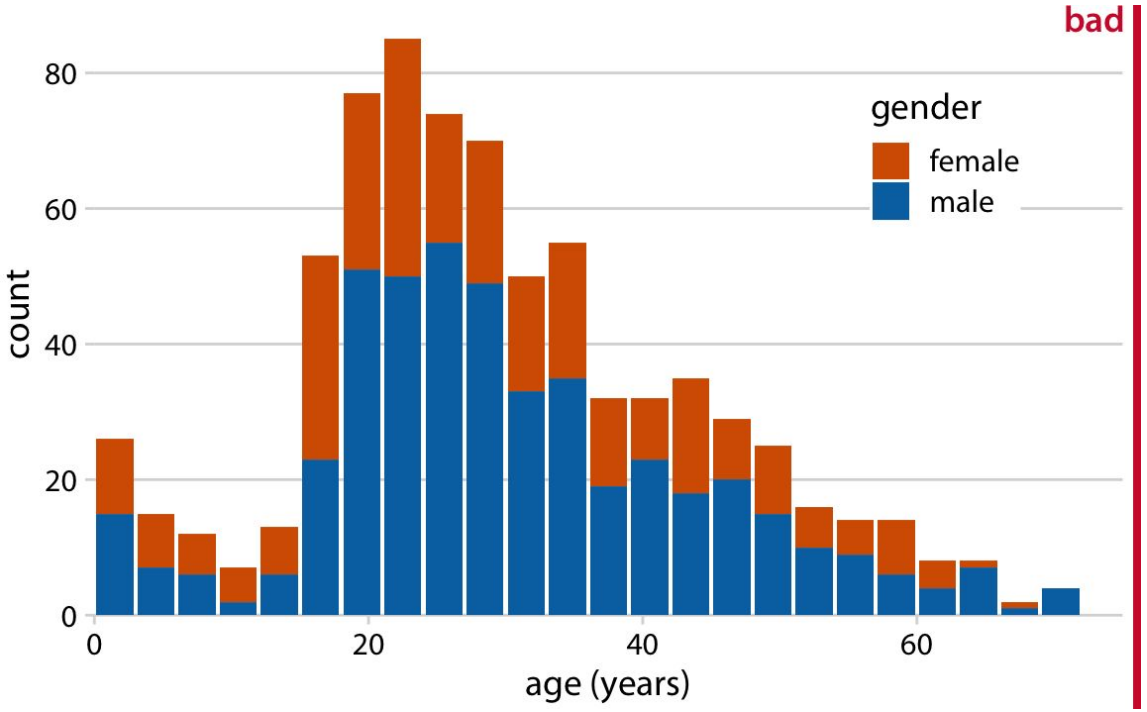
Диаграмма плотности



Оценки плотности ядра имеют тенденцию приводить к появлению данных там, где их нет, особенно в хвостах. Как следствие, небрежное использование оценок плотности может легко привести к явно неверным визуализациям. Например, визуализацию распределения по возрасту, которая включает отрицательный возраст.

Всегда проверяйте, что ваша оценка плотности не приводит к появлению бессмысленных значений.

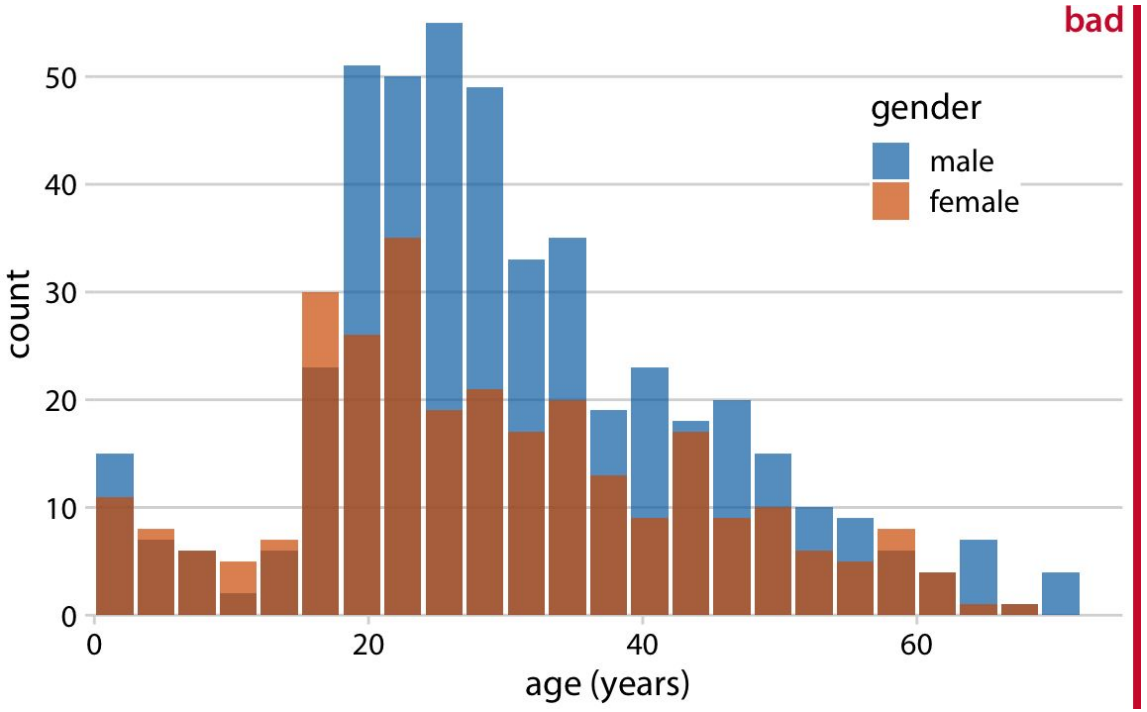
Несколько распределений на одной визуализации



- Гистограмма возрастов пассажиров Титаника с разделением по полу.
- Гистограммы с накоплением легко спутать с перекрывающимися гистограммами.
- Кроме того, высоты столбцов, представляющих пассажиров женского пола, нельзя легко сравнить друг с другом.

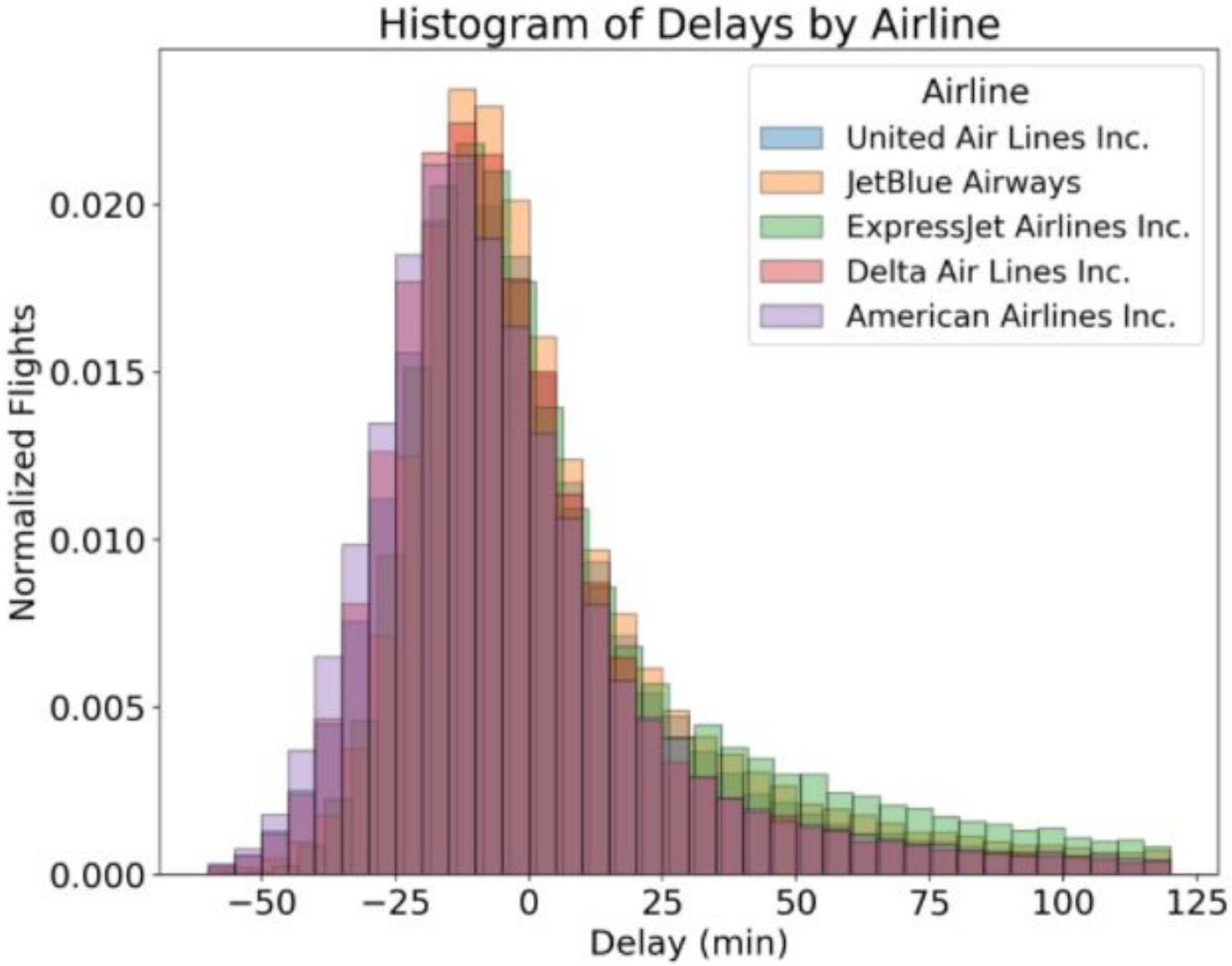
econophysica

Несколько распределений на одной визуализации



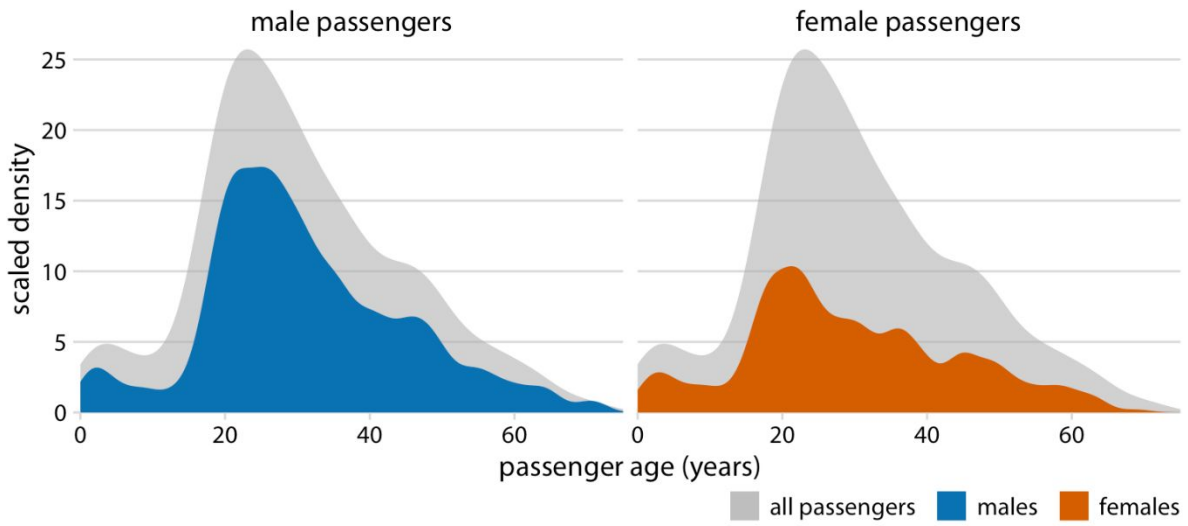
Теперь кажется, что на самом деле есть три разные группы, а не две, и мы до сих пор не до конца уверены, где начинается и заканчивается каждая полоса. Перекрывающиеся гистограммы не работают должным образом, потому что полупрозрачная полоса, нарисованная поверх другой, имеет тенденцию не выглядеть как полупрозрачная полоса, а вместо этого похожа на полосу, нарисованную другим цветом.

Несколько распределений на одной визуализации



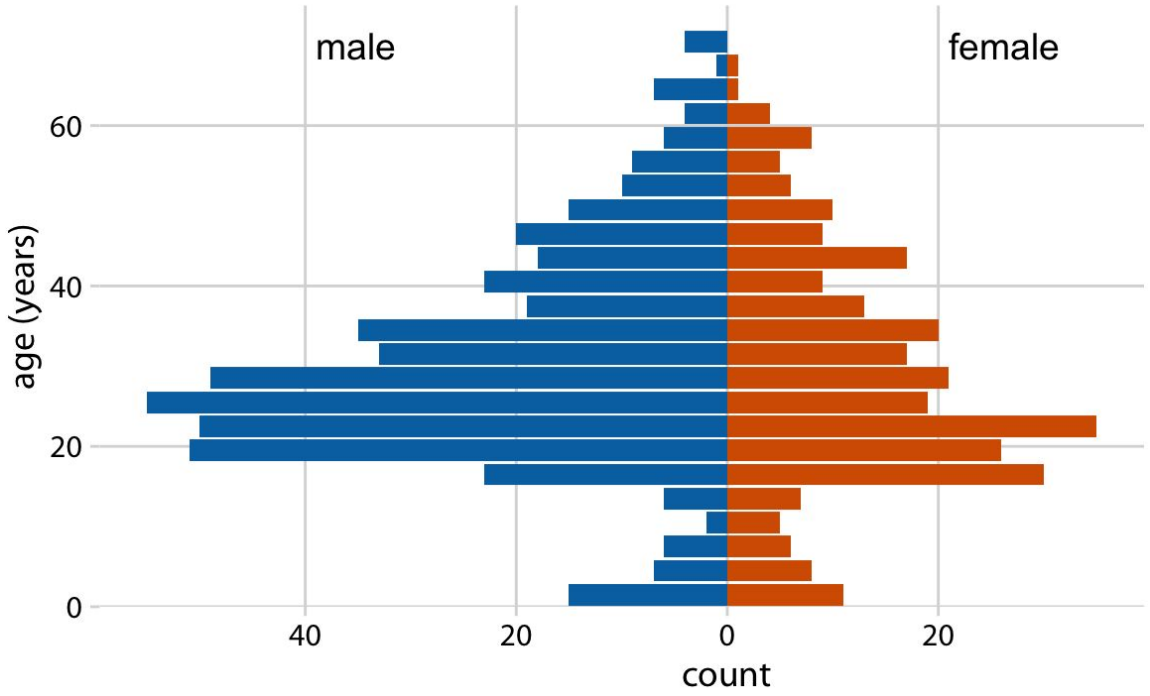
econophysica

Несколько распределений на одной визуализации



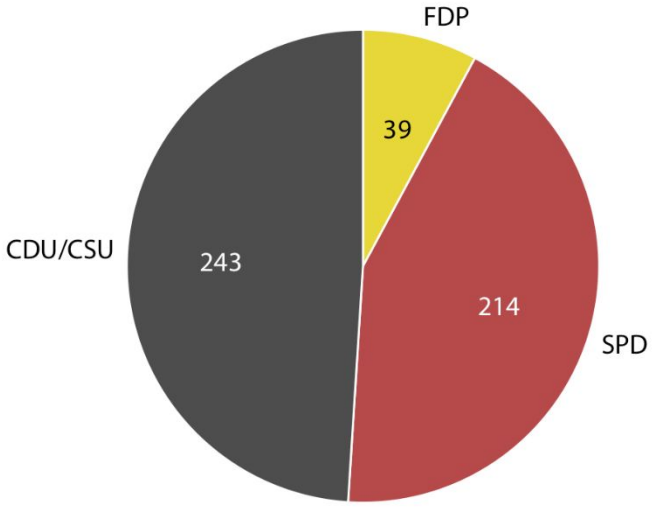
Решение, которое хорошо работает для этого набора данных, состоит в том, чтобы показать распределение по возрасту пассажиров мужского и женского пола по отдельности, каждое из которых как часть общего распределения по возрасту.

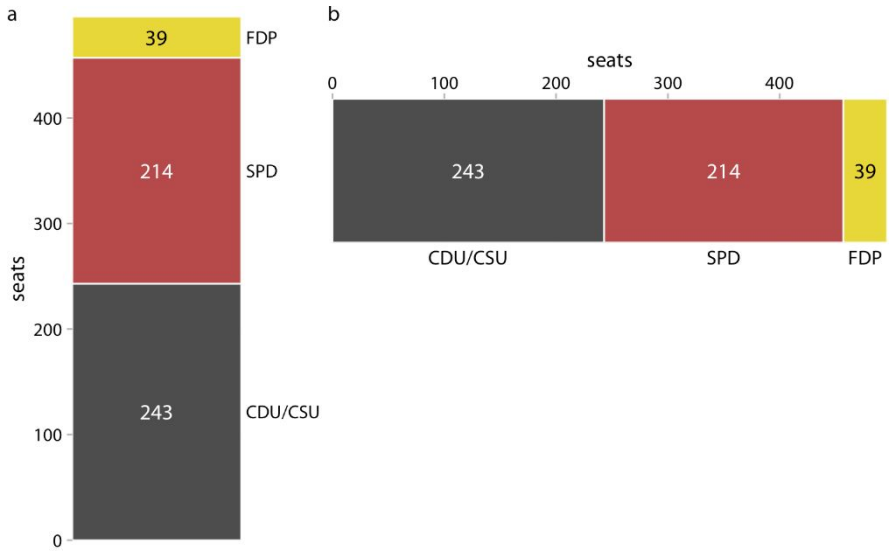
Несколько распределений на одной визуализации



Наконец, когда мы хотим визуализировать ровно два распределения, мы также можем создать две отдельные гистограммы, повернуть их на 90 градусов и сделать так, чтобы столбцы в одной гистограмме указывали в противоположном направлении от другого. Этот трюк обычно используется при визуализации возрастных распределений, а получающийся график обычно называется *возрастной пирамидой*

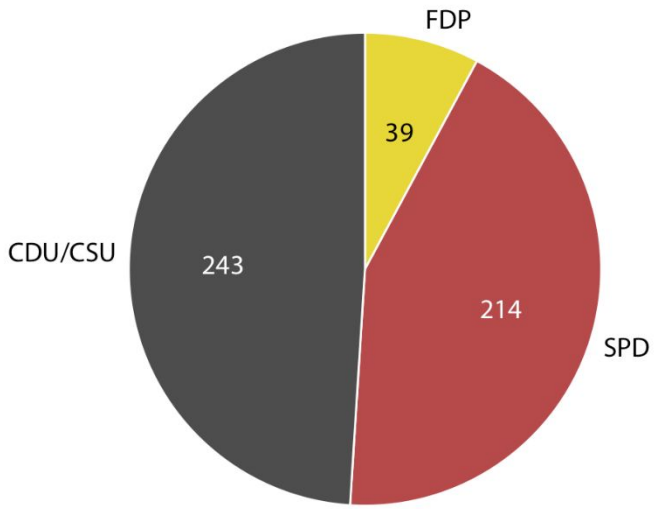
Круговые диаграммы

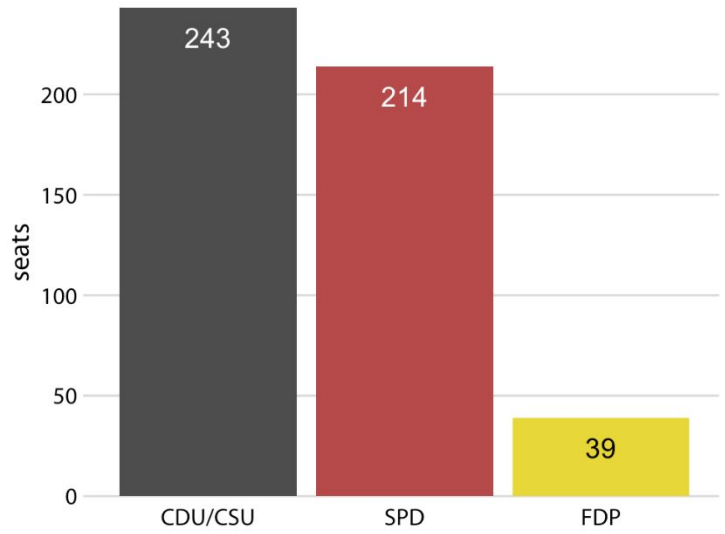




Круговая диаграмма разбивает круг на срезы, так что площадь каждого среза пропорциональна доле итога, которую он представляет. Эту же процедуру можно выполнить для прямоугольника, и в результате получается столбчатая диаграмма с суммированием.

Круговые диаграммы

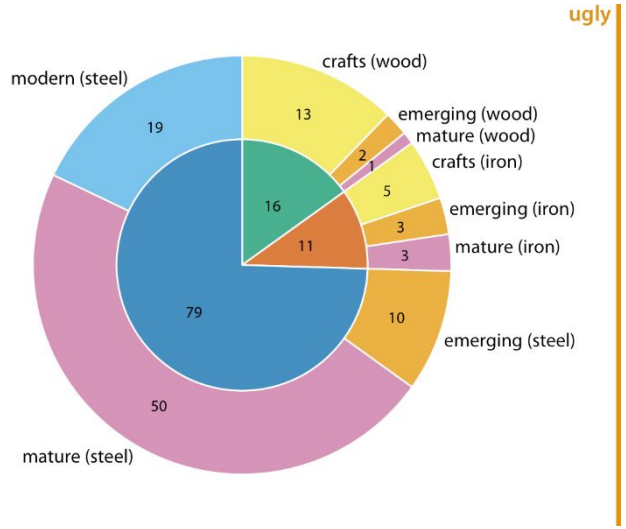


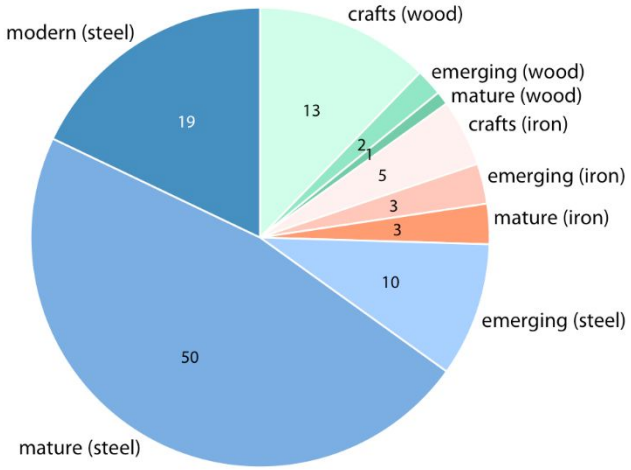


- Для сравнения, столбчатая диаграмма облегчает прямое сравнение трех групп, хотя и затеняет другие аспекты. Так. отношение каждого столбца к общему числу не является визуально очевидным.

- Помните: вам всегда нужно выбирать визуализацию, которая наилучшим образом соответствует вашему конкретному набору данных и выделяет ключевые особенности данных, которые вы хотите подчеркнуть.

Круговые диаграммы с подкатегориями

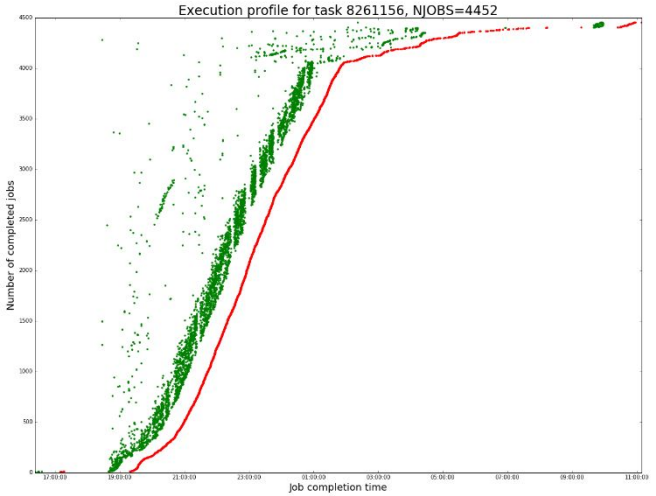


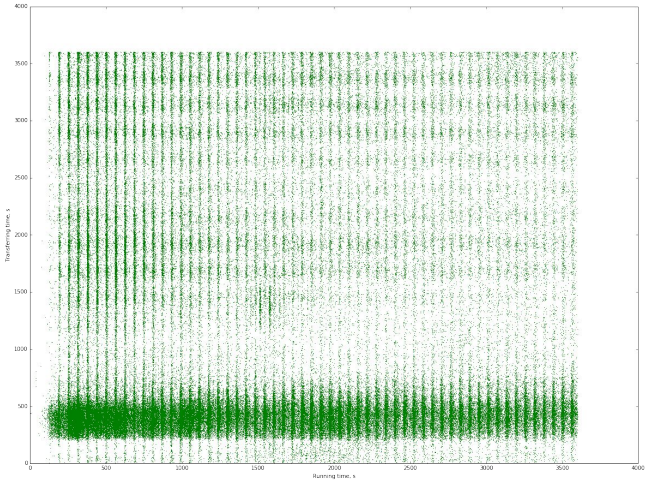


Круговые диаграммы можно делать с разбиением частей на ещё более мелкие части, но такие диаграммы трудно воспринимаются.

Диаграмма, где каждый подэлемент – наивысшего уровня, и группировка показана цветом, воспринимается лучше.

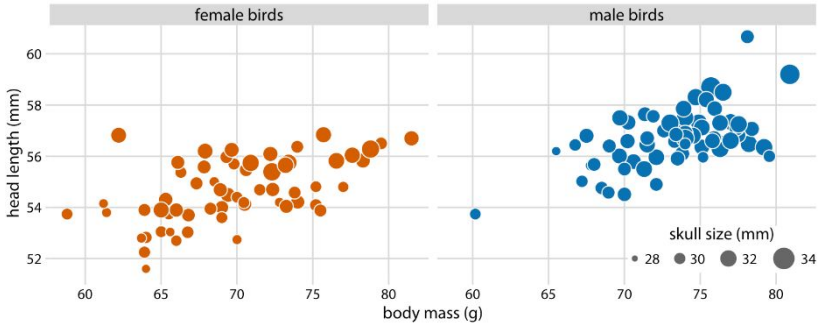
Диаграммы рассеяния





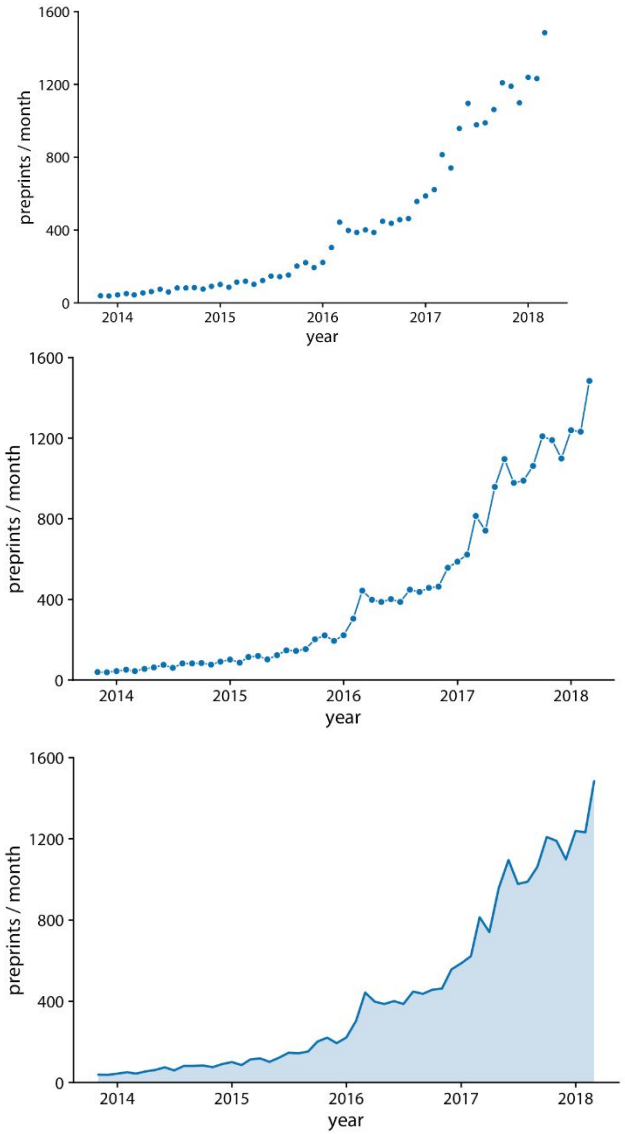
- Многие наборы данных содержат две или более количественных переменных, и нас может интересовать, как эти переменные связаны друг с другом.
- Чтобы построить отношения только двух таких переменных, например, рост и вес, обычно применяют диаграмму рассеяния.
- Если нужно показать более двух переменных разом, можно использовать пузырьковую диаграмму.

Пузырьковые диаграммы



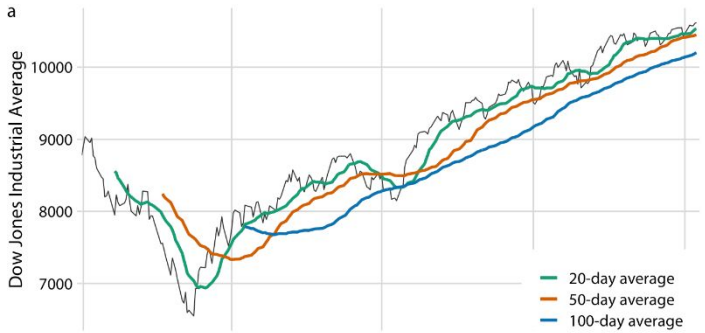
- Главный недостаток пузырьковых диаграмм в том, что они показывают одинаковые типы переменных, количественные переменные, с двумя различными типами шкал, положением и размером.
- Это затрудняет визуальное выявление сильных связей между различными переменными.
- Более того, различия между значениями данных, закодированных как размер пузырьков, сложнее воспринимать, чем различия между значениями данных, закодированными как положение.
- Поскольку даже самые большие пузырьки должны быть маленькими по сравнению с общим размером фигуры, различия в размерах даже между самыми большими и самыми маленькими пузырьками обязательно будут небольшими.
- Следовательно, меньшие различия в значениях данных будут соответствовать очень маленьким различиям в размерах, которые практически невозможно увидеть.

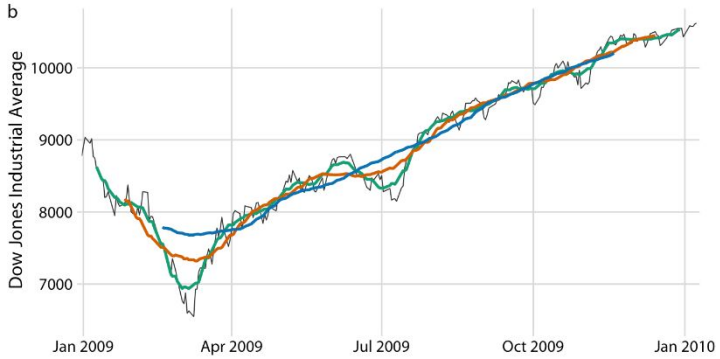
Временные ряды



- Важная особенность временных рядов: у каждой точки есть строго заданные левый и правый соседи, строго по одному, и интервалы по оси X между точками одинаковы.
- Мы можем визуально подчеркнуть это, соединяя соседние точки линиями.
- Некоторые люди возражают против рисования линий между точками, потому что линии не представляют наблюдаемые данные. В частности, если бы только несколько наблюдений были расположены далеко друг от друга, если бы наблюдения проводились в промежуточное время, они, вероятно, не попали бы точно в показанные линии.
- Таким образом, в некотором смысле, линии не соответствуют реальным данным.
- Тем не менее, они могут помочь с восприятием, когда точки разнесены далеко или неравномерно.
- Мы можем в некоторой степени решить эту дилемму, указав ее в подписи к рисунку, например, написав «линии приведены для иллюстрации».

Тренды

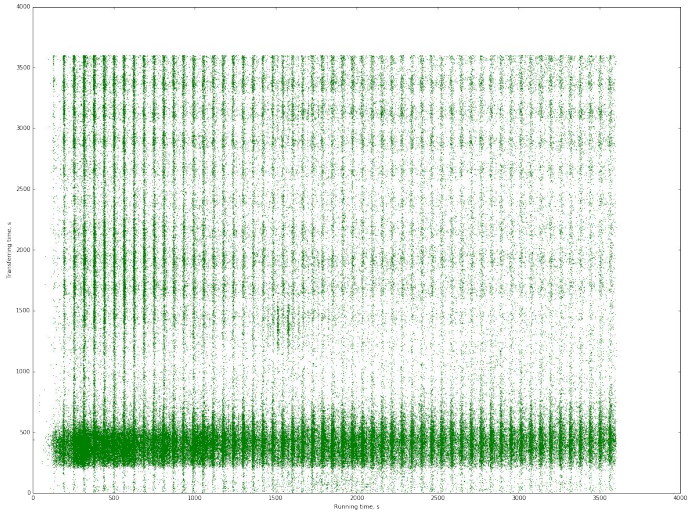


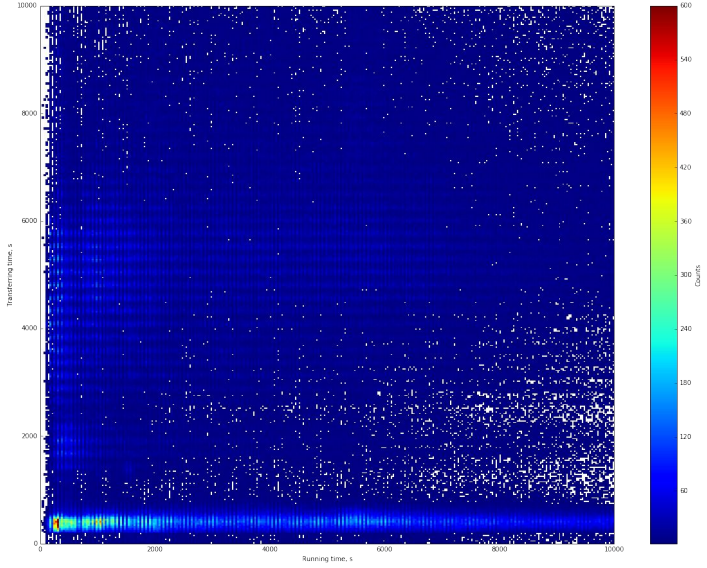


- Акт сглаживания создает функцию, которая сохраняет важные закономерности в данных, удаляя ненужные незначительные детали или шумы.
- Финансовые аналитики обычно сглаживают данные фондового рынка, вычисляя скользящие средние.
- Чтобы сгенерировать скользящее среднее, мы берем временное окно, скажем, первые 20 дней во временном ряду, вычисляем среднюю цену за эти 20 дней, затем перемещаем временное окно на один день, так что теперь оно охватывает 2-й и 21-й день, вычислите среднее значение за эти 20 дней, снова переместите временное окно и т. д. Результатом является новый временной ряд, состоящий из последовательности усредненных цен.
- Чтобы построить эту последовательность скользящих средних, нам нужно решить, какой конкретный момент времени связать со средним для каждого временного окна.
- Финансовые аналитики часто отображают каждое среднее значение в конце соответствующего временного окна. Этот выбор приводит к кривым, которые отстают от исходных данных (рисунок a), с более серьезными задержками, соответствующими большим временным окнам усреднения.
- Статистики, с другой стороны, строят среднее значение в центре временного окна, в результате чего получается кривая, идеально наложенная на исходные данные.

econophysica

2D-гистограммы





- 2D-гистограмма концептуально похожа на одномерную гистограмму, как обсуждалось в главе 7, но теперь мы объединяем данные в двух измерениях.
- Мы подразделяем всю плоскость x – y на маленькие прямоугольники, подсчитываем, сколько наблюдений приходится на каждый из них, и затем окрашиваем прямоугольники по этому количеству.

econophysica

Визуализация должна содержать историю

Большая часть визуализации данных делается с целью общения.

У нас есть представление о наборе данных, и у нас есть потенциальная аудитория, и мы хотели бы донести нашу информацию до нашей аудитории.

Чтобы успешно донести наше понимание, нам нужно будет представить зрителям ясную и захватывающую **историю**.

Потребность в истории может раздражать ученых и инженеров, которые могут приравнять ее к придумыванию, продаже результатов.

Тем не менее, эта точка зрения упускает важную роль, которую истории играют в рассуждениях и памяти.

Любое общение создает историю в умах аудитории.

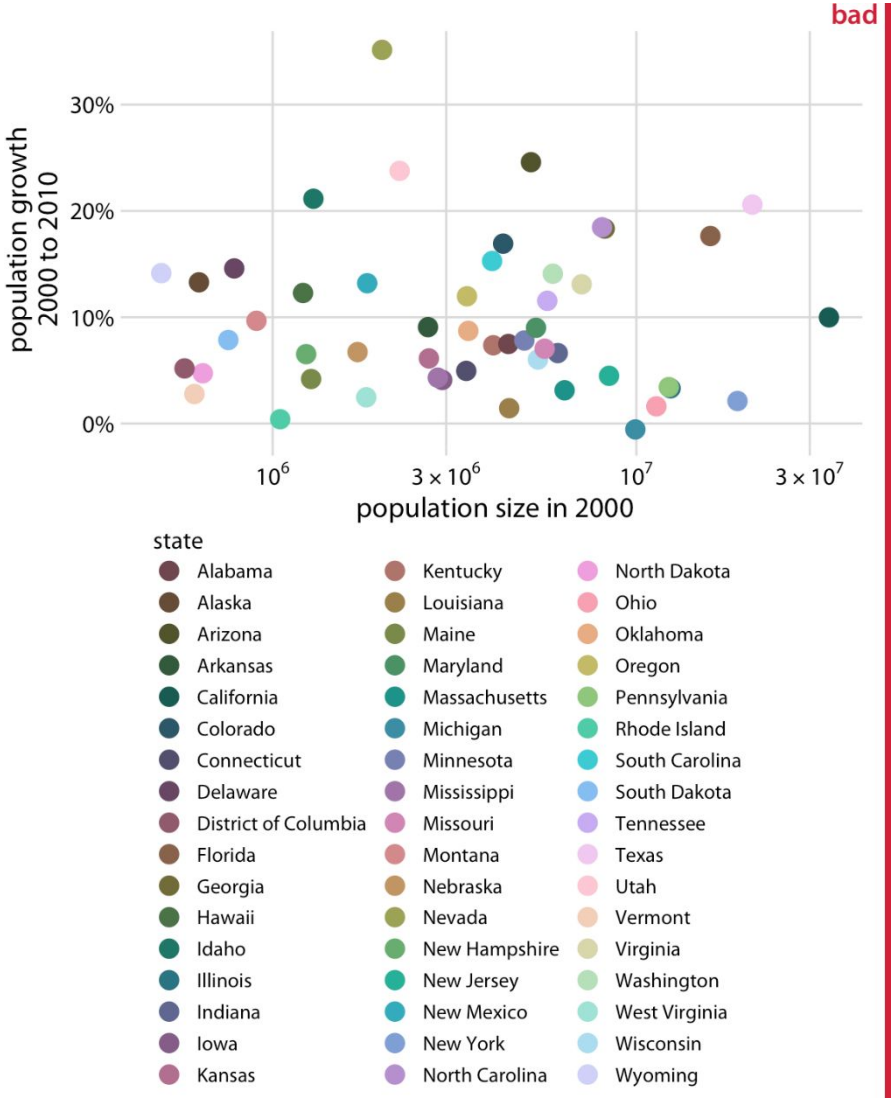
Если мы сами не представим ясную историю, то наша аудитория создаст ее за нас.

В лучшем случае эта историябудет достаточно близка нашему взгляду на представленный материал.

Тем не менее, это может быть и часто намного хуже. Придуманная история может быть «это скучно», «автор неправ» или «автор некомпетентен».

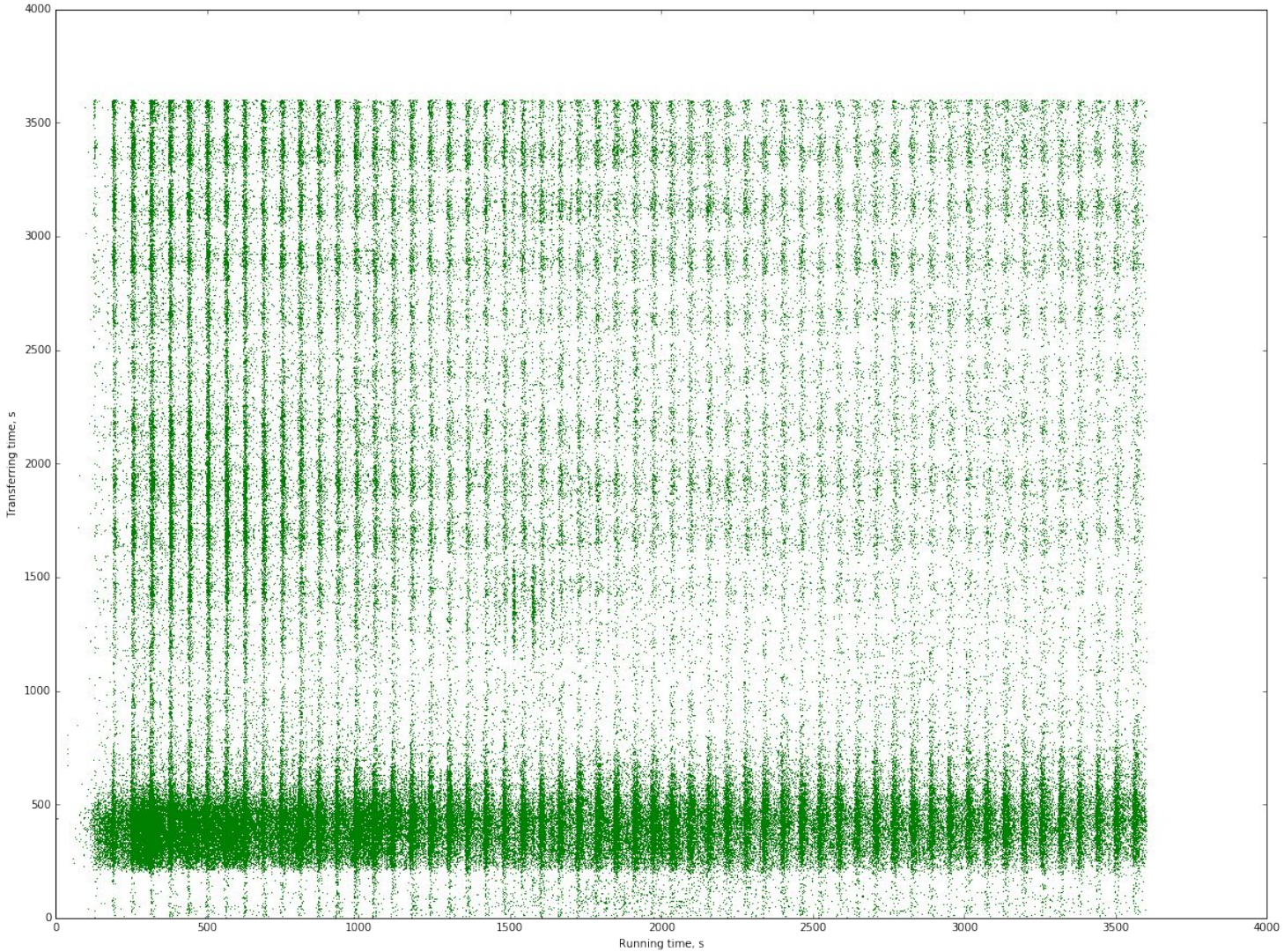
Ваша цель в рассказывании истории должна состоять в том, чтобы использовать факты и логические рассуждения, чтобы заинтересовать аудиторию.

Частые ошибки использования цвета



Используйте прямую маркировку вместо цветов, когда нужно различать более восьми категорий.

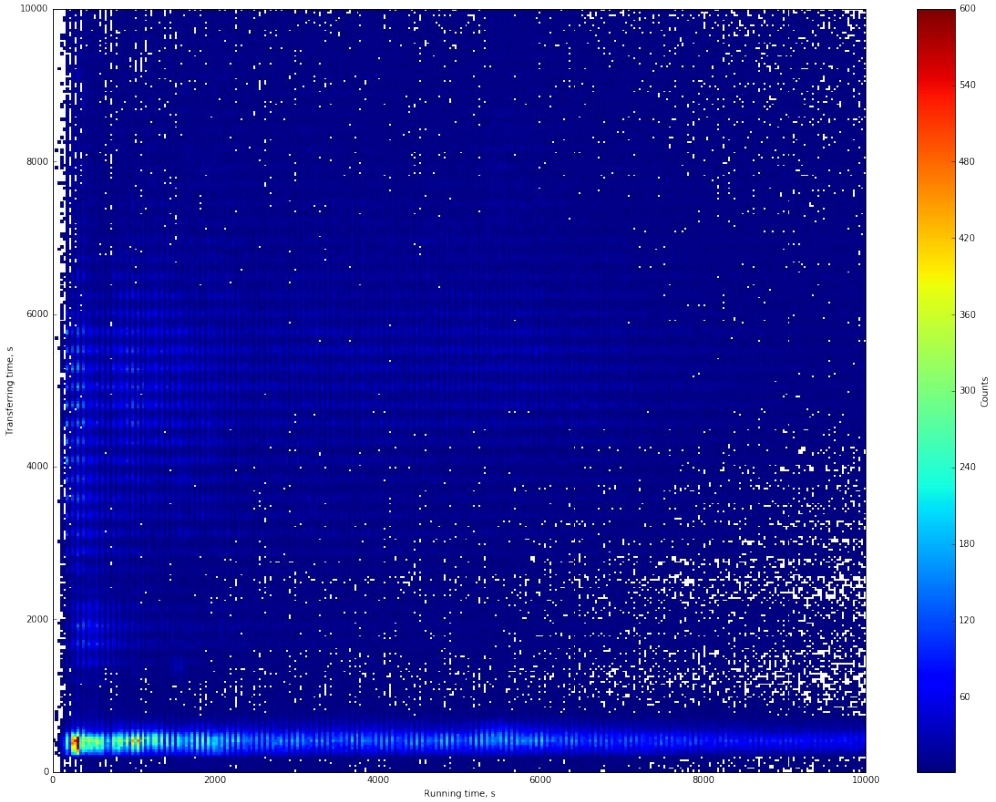
Частые ошибки использования цвета



Избегайте больших заполненных областей чрезмерно насыщенных цветов. Они мешают вашему читателю воспринимать ваши данные.

econophysica

Частые ошибки использования цвета



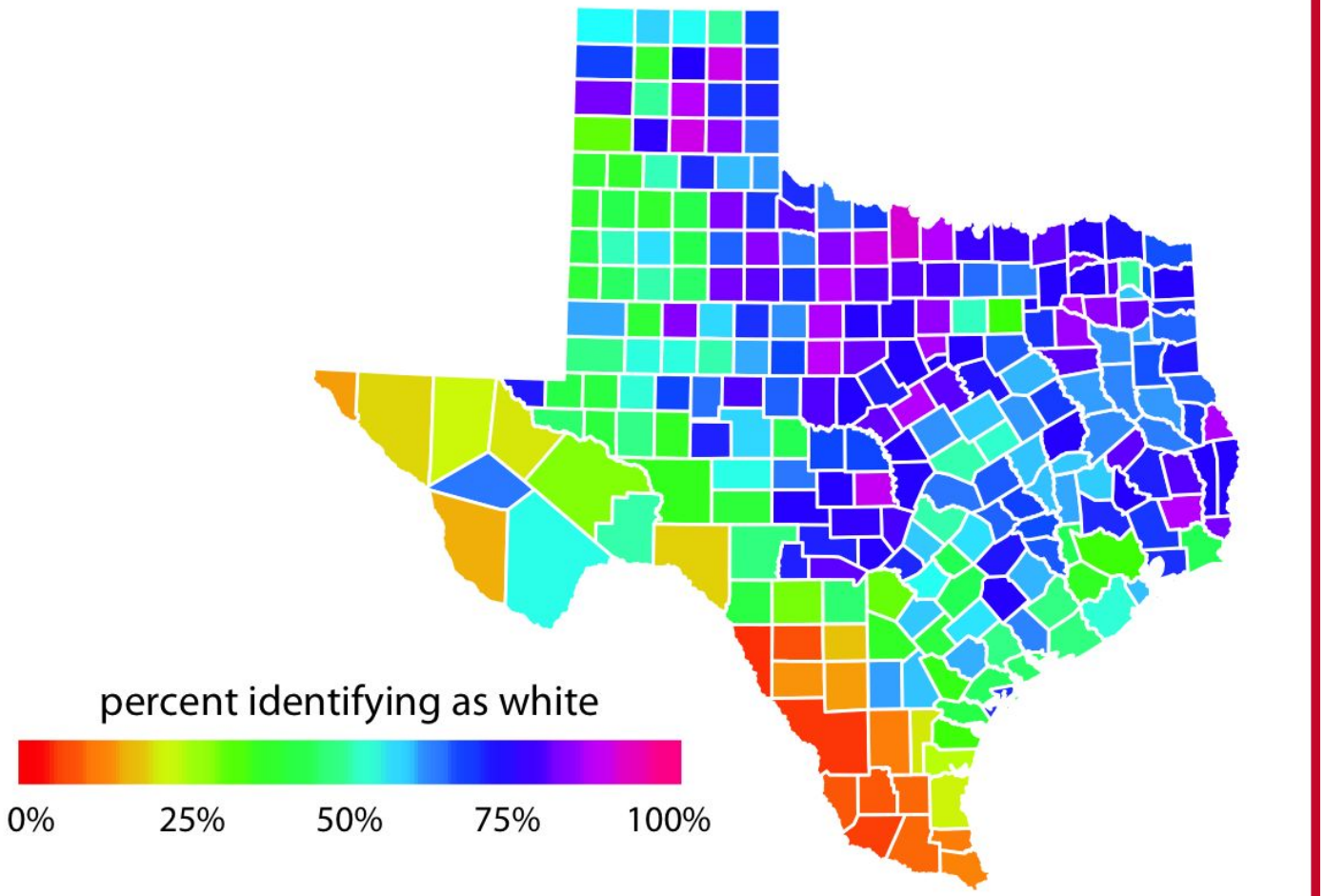
Цветовая гамма радуги очень немонотонна. Это становится ясно видимым путем преобразования цветов в значения серого.

Слева направо шкала меняется от умеренно темной до светлой до очень темной и обратно до умеренно темной.

econophysica

Частые ошибки использования цвета

bad

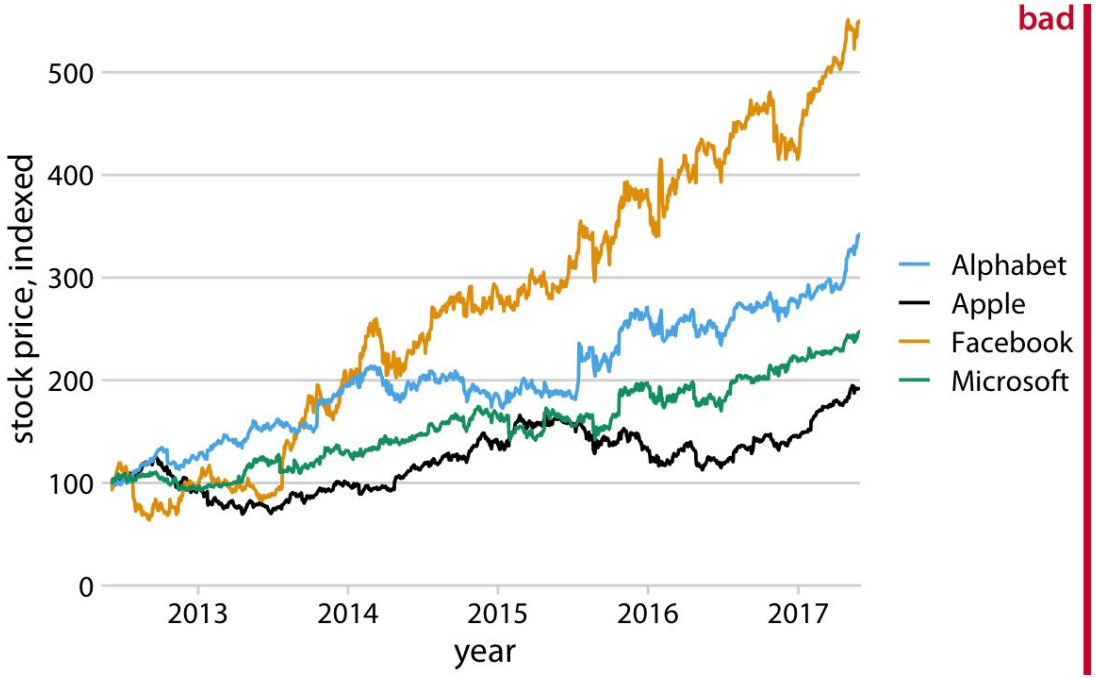


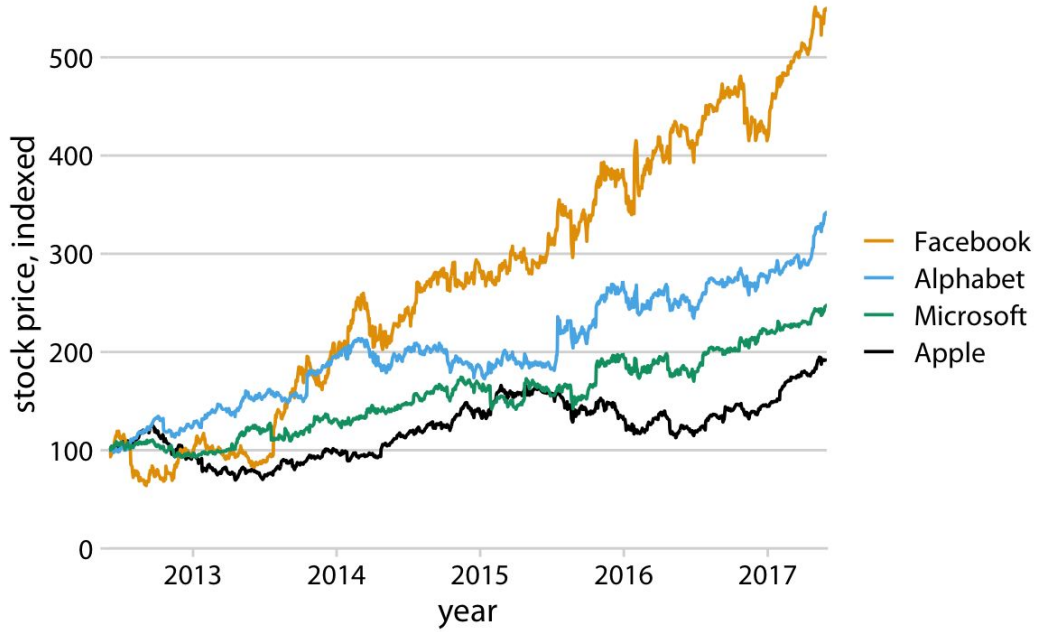
econophysica

Частые ошибки использования цвета: дальтонизм

- По первым двум примерам может показаться, что почти невозможно найти два контрастных цвета, которые безопасны при всех формах нарушений цветовосприятия.
- Однако ситуация не так страшна. Часто можно внести небольшие изменения в цвета, чтобы они имели желаемый вид, а также были различимы людьми с ограниченными возможностями.
- Например, шкала ColorBrewer PiYG (от розового до желто-зеленого) выглядит красно-зеленой для людей с нормальным цветовым зрением, но остается различимой для людей с проблемами цветовосприятия.

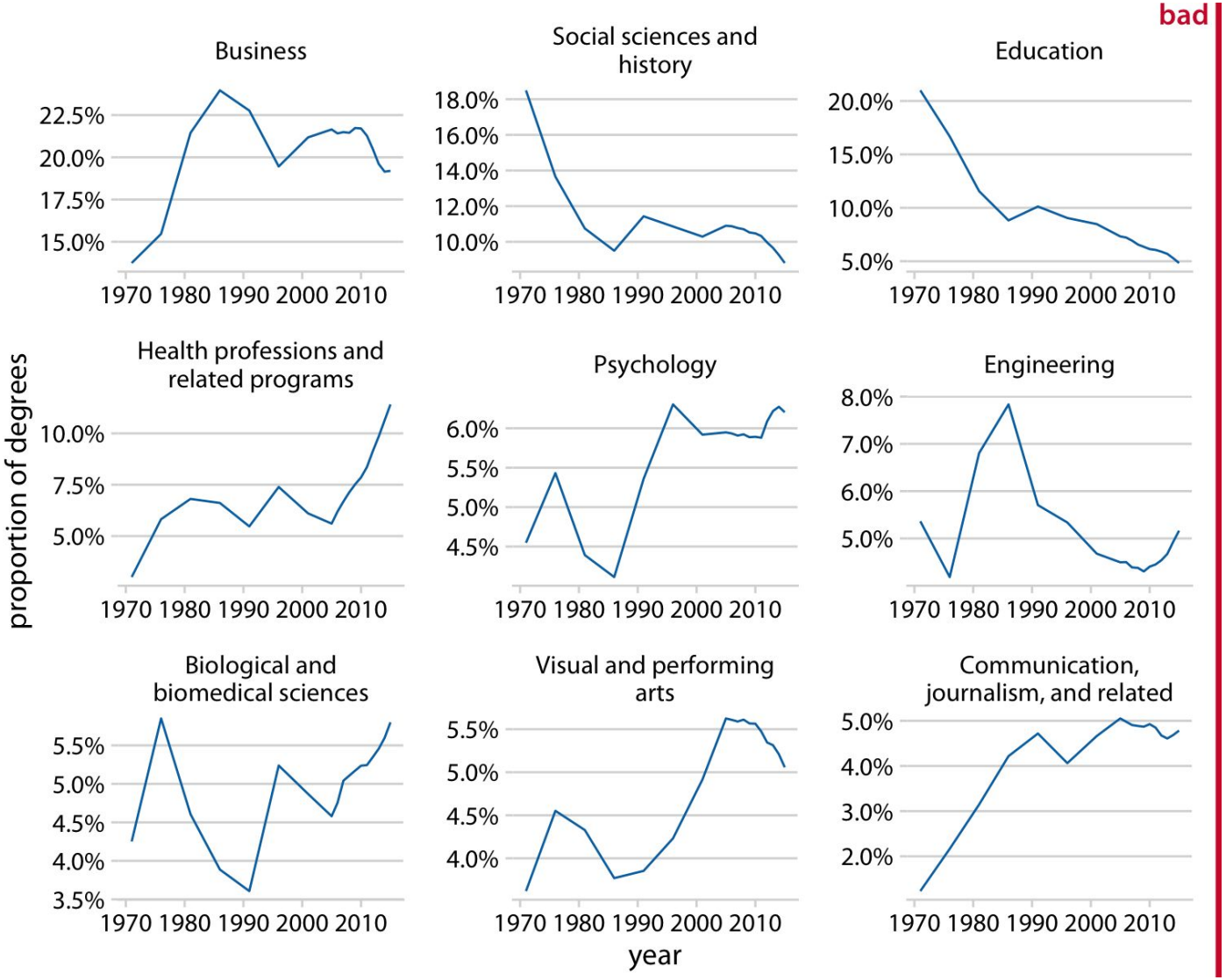
Частые ошибки композиции: легенда





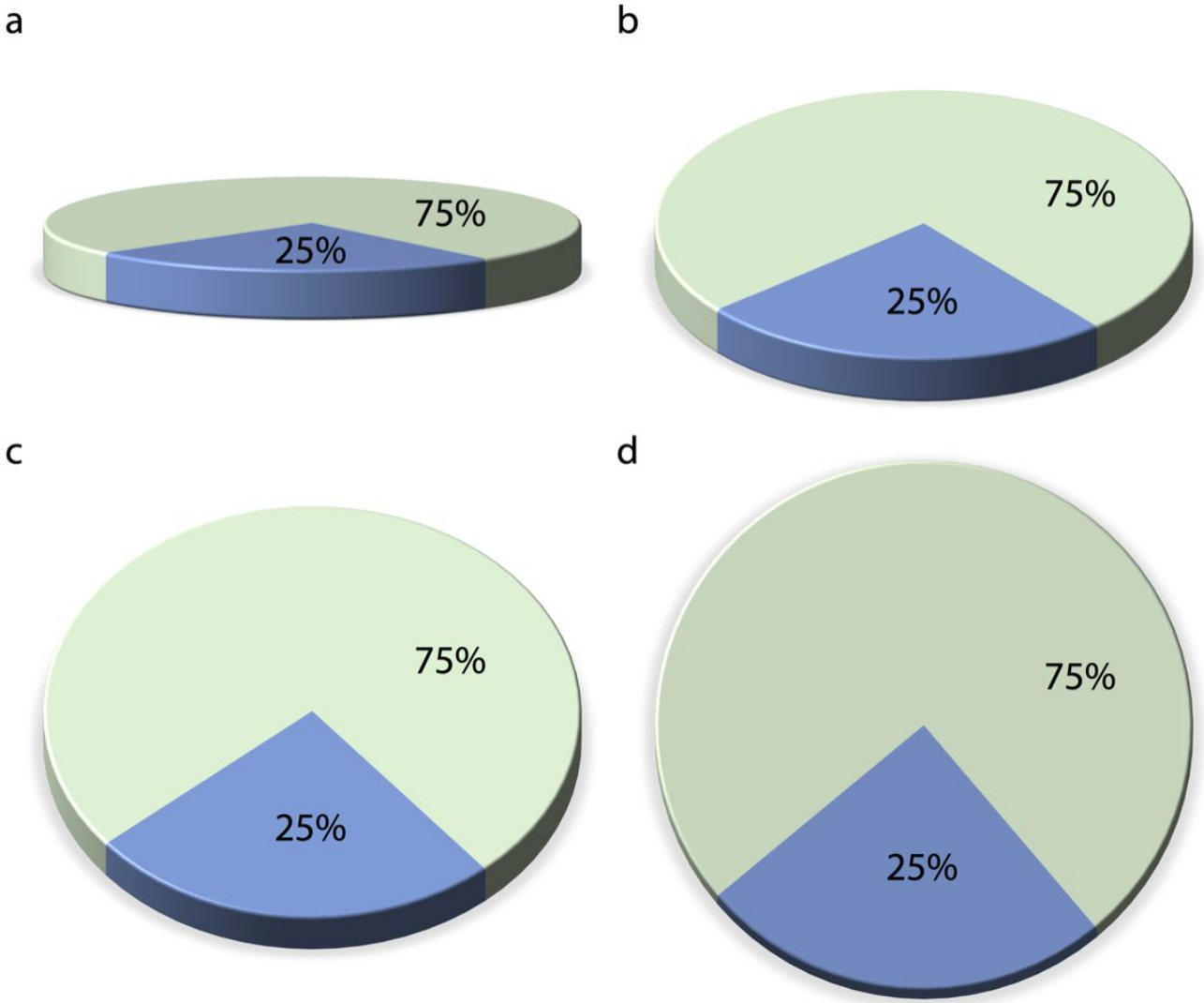
econophysica

Частые ошибки композиции: шкалы

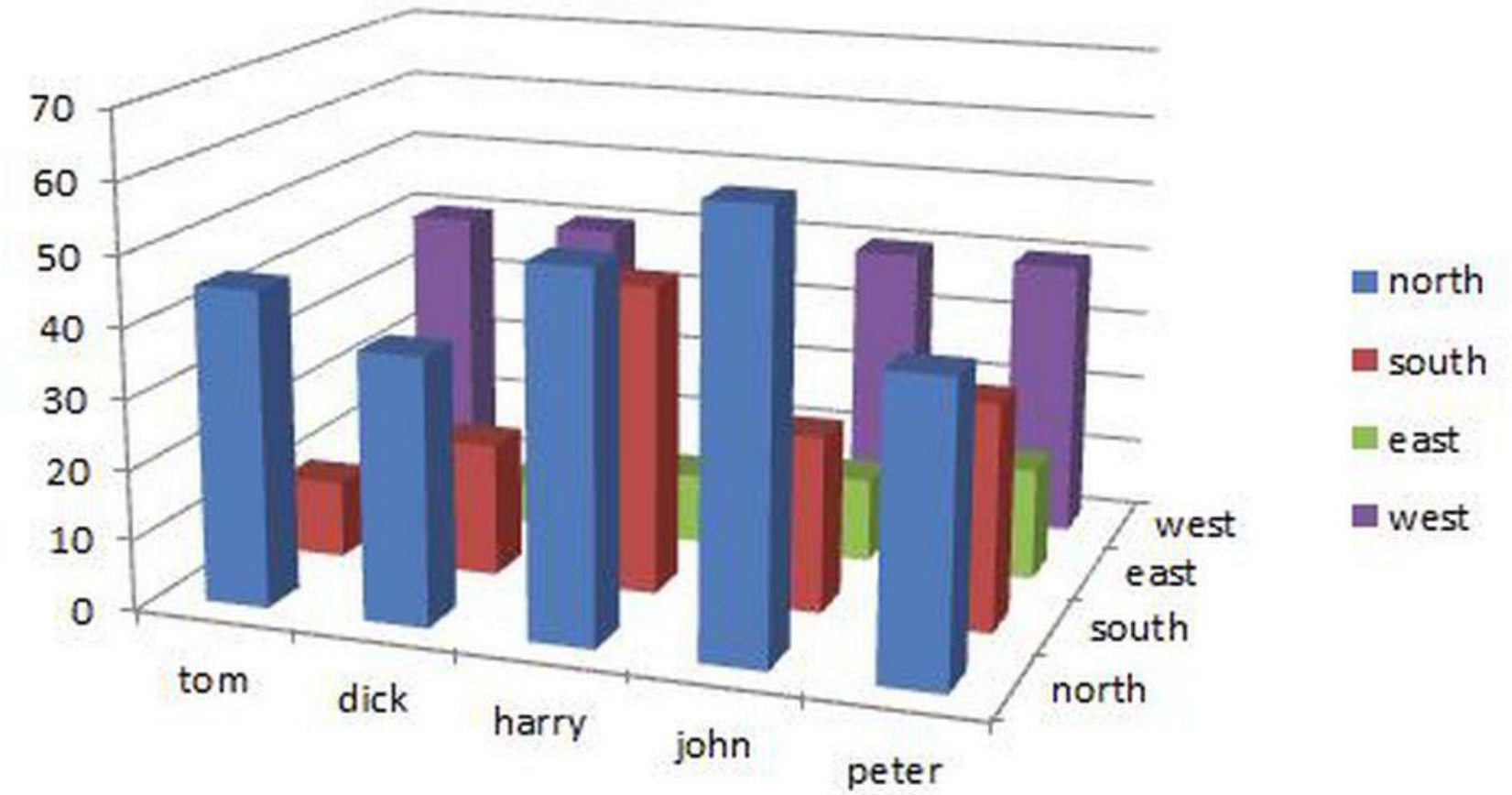


econophysica

Частые ошибки композиции: 3D

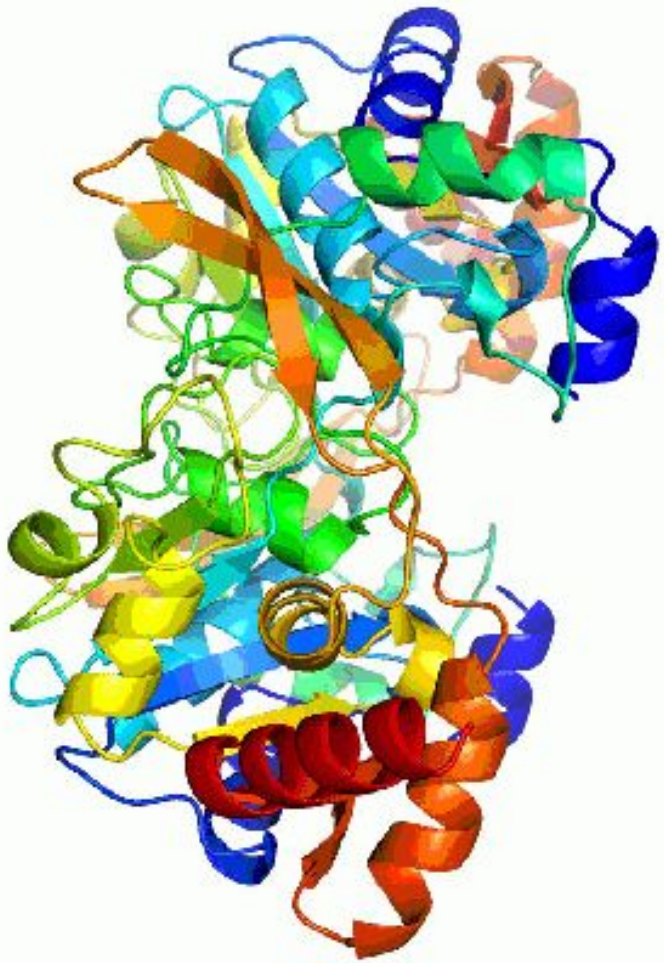


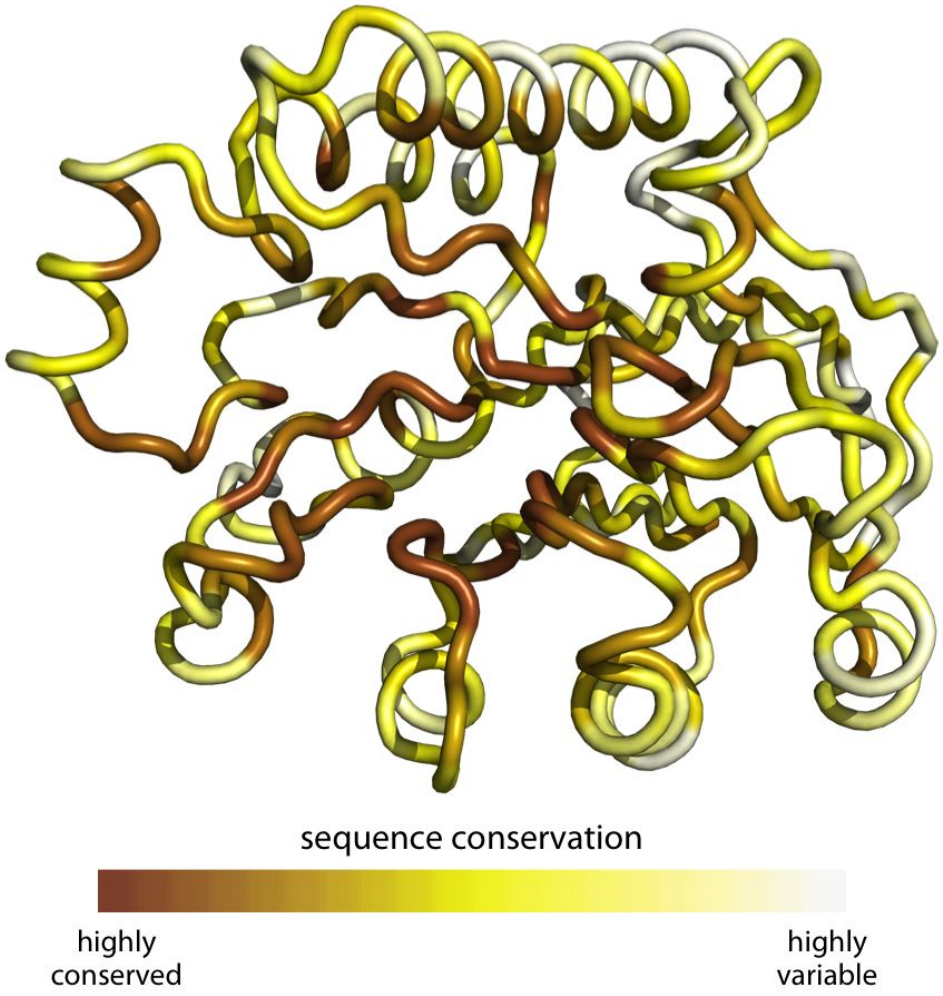
Частые ошибки композиции: 3D



econophysica

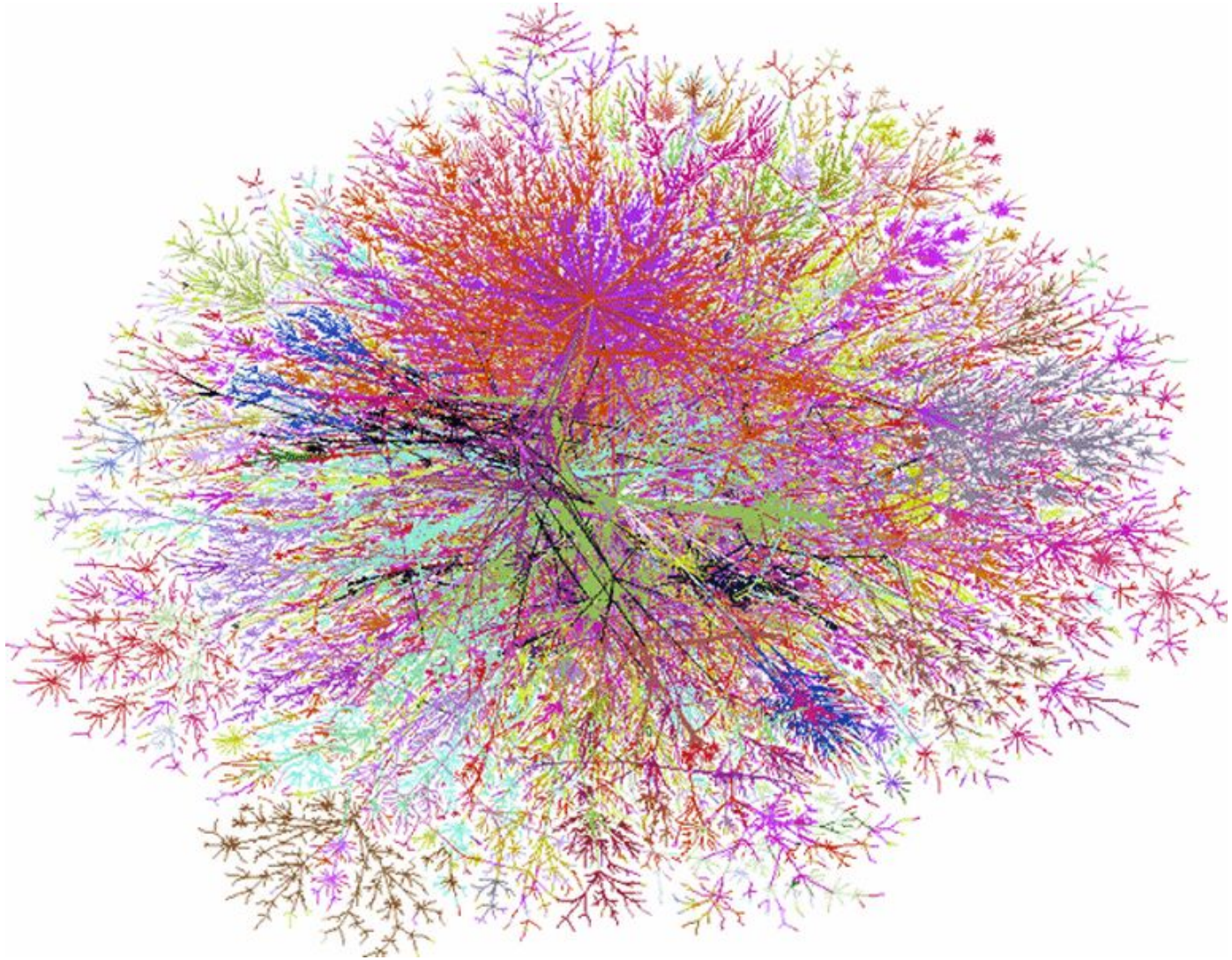
Частые ошибки композиции: 3D



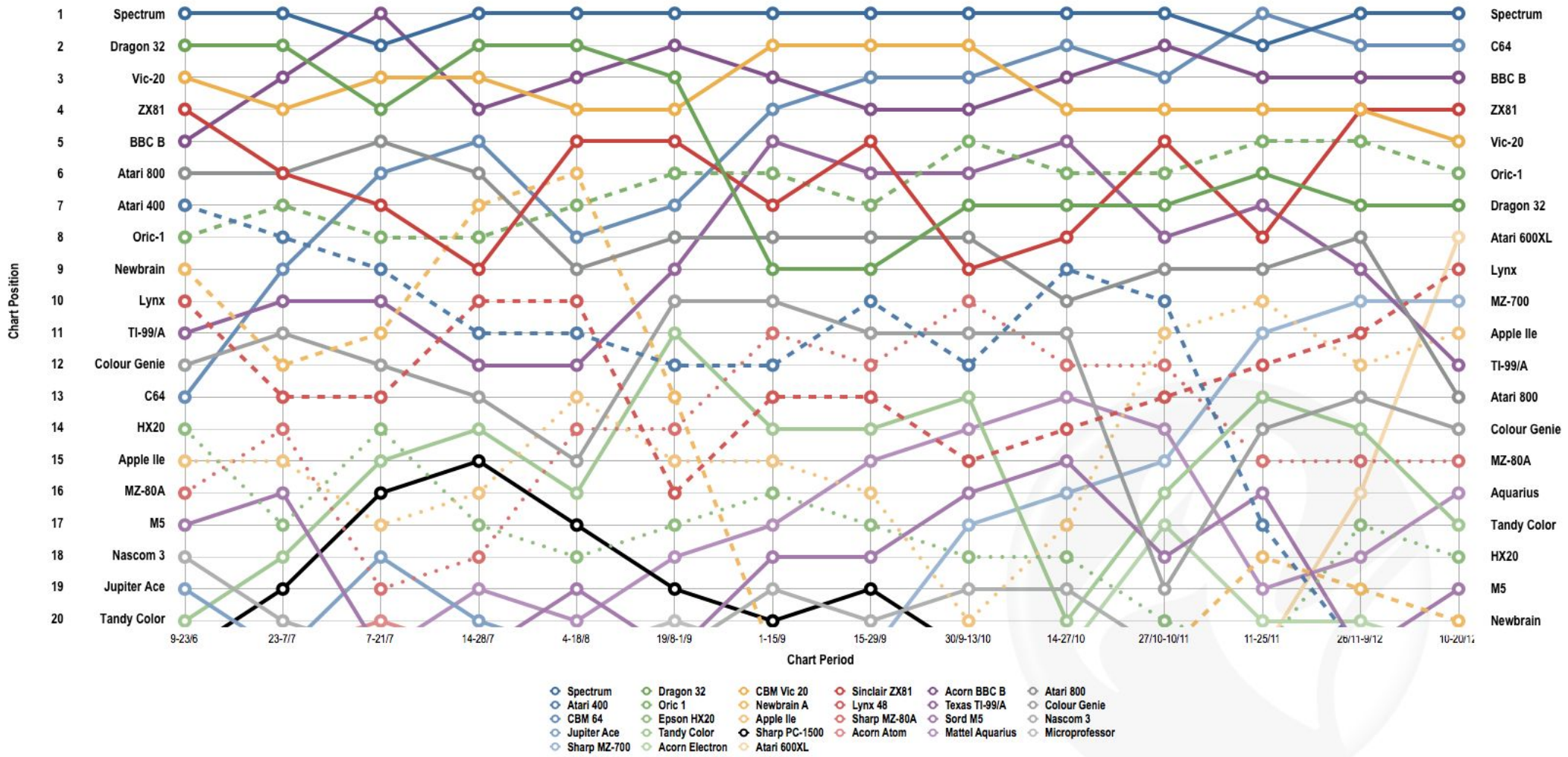


econophysica

Частые ошибки композиции: нагроможденность

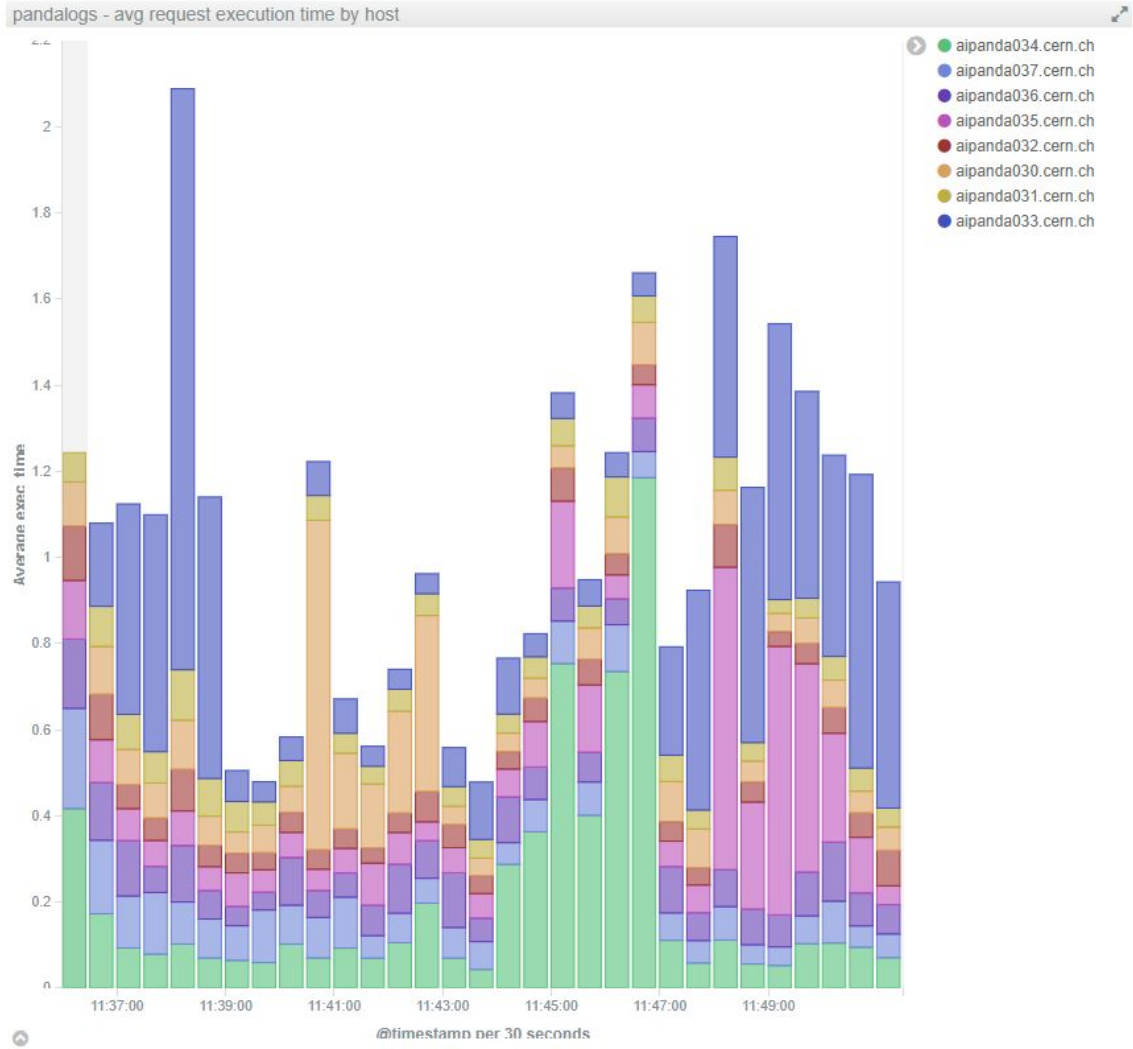


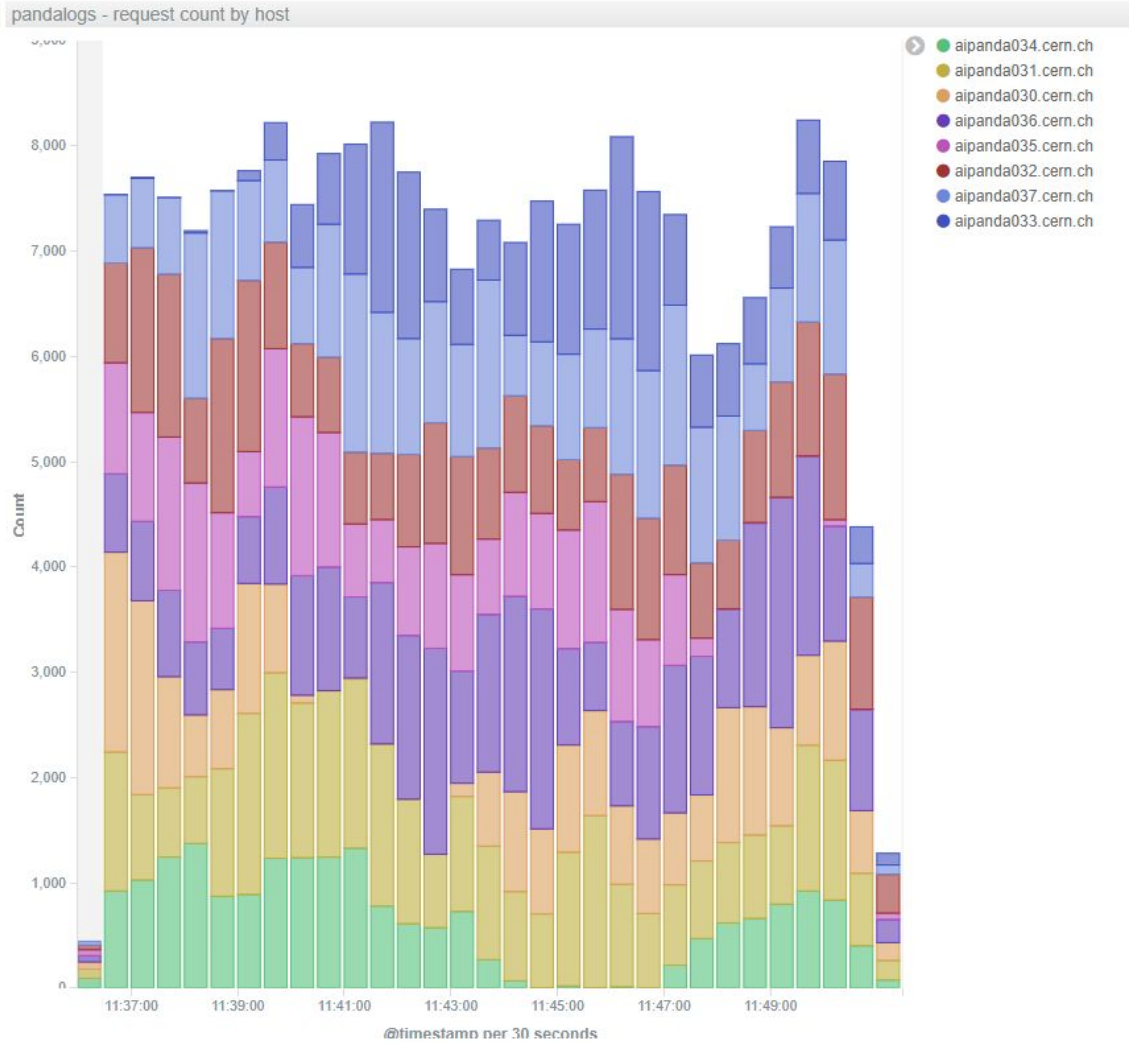
Частые ошибки композиции: нагроможденность



econophysica

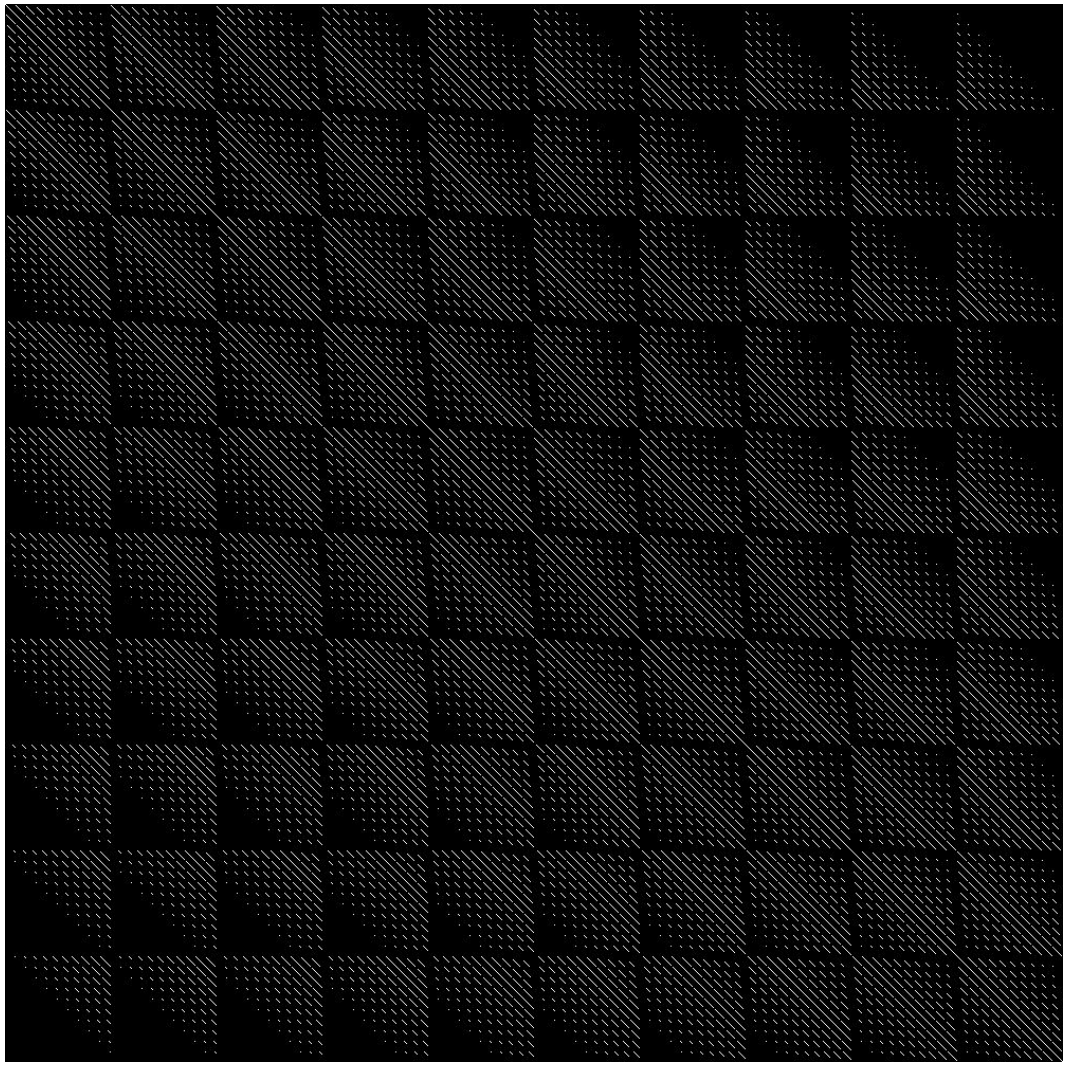
Примеры реальных диаграмм





econophysica

Визуализация больших данных



- Отображение миллионов точек данных в одной визуализации почти всегда делает визуализацию загроможденной до степени нечитаемости.
- Решение этой проблемы требует группировки, агрегирования данных или построения репрезентативных выборок.
- Круговые, линейные, гистограммы и диаграммы рассеяния могут быть очень полезны для отображения агрегированных данных.

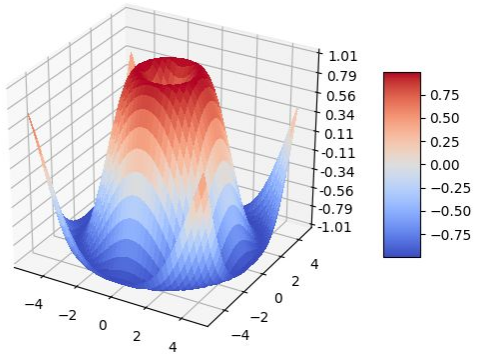
Matplotlib

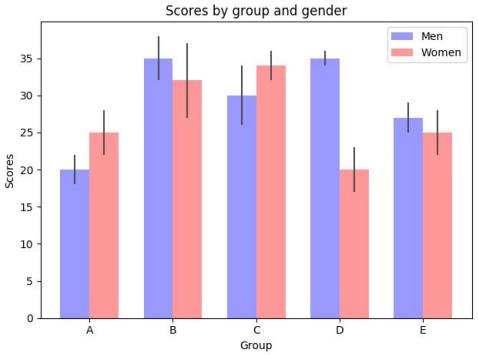
- Matplotlib это самая большая и мощная из библиотек визуализации, доступных в Python.
- Она мощная, гибкая, и имеет огромный выбор визуализаций.
- Для начинающего пользователя, она может показаться слишком сложной.
- Вы можете потратить много времени на работу со всеми доступными опциями, даже если все, что вам нужно, - это создать простой точечный график.

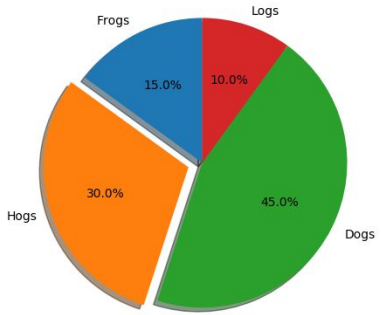
Matplotlib - самый популярный инструмент визуализации для Python, поэтому вы никогда не будете одиноки в случае возникновения проблем.

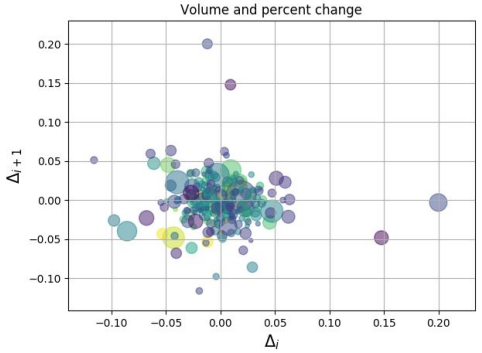
Matplotlib

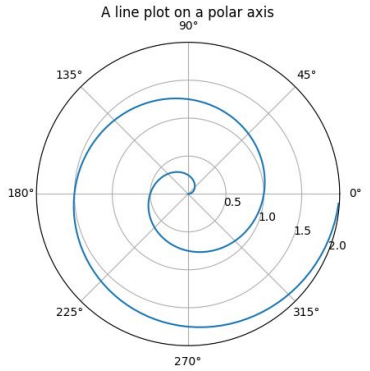
matplotlib

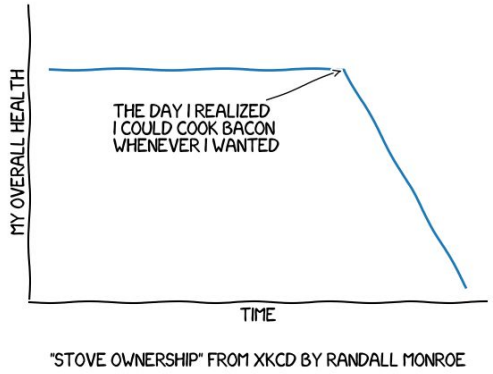










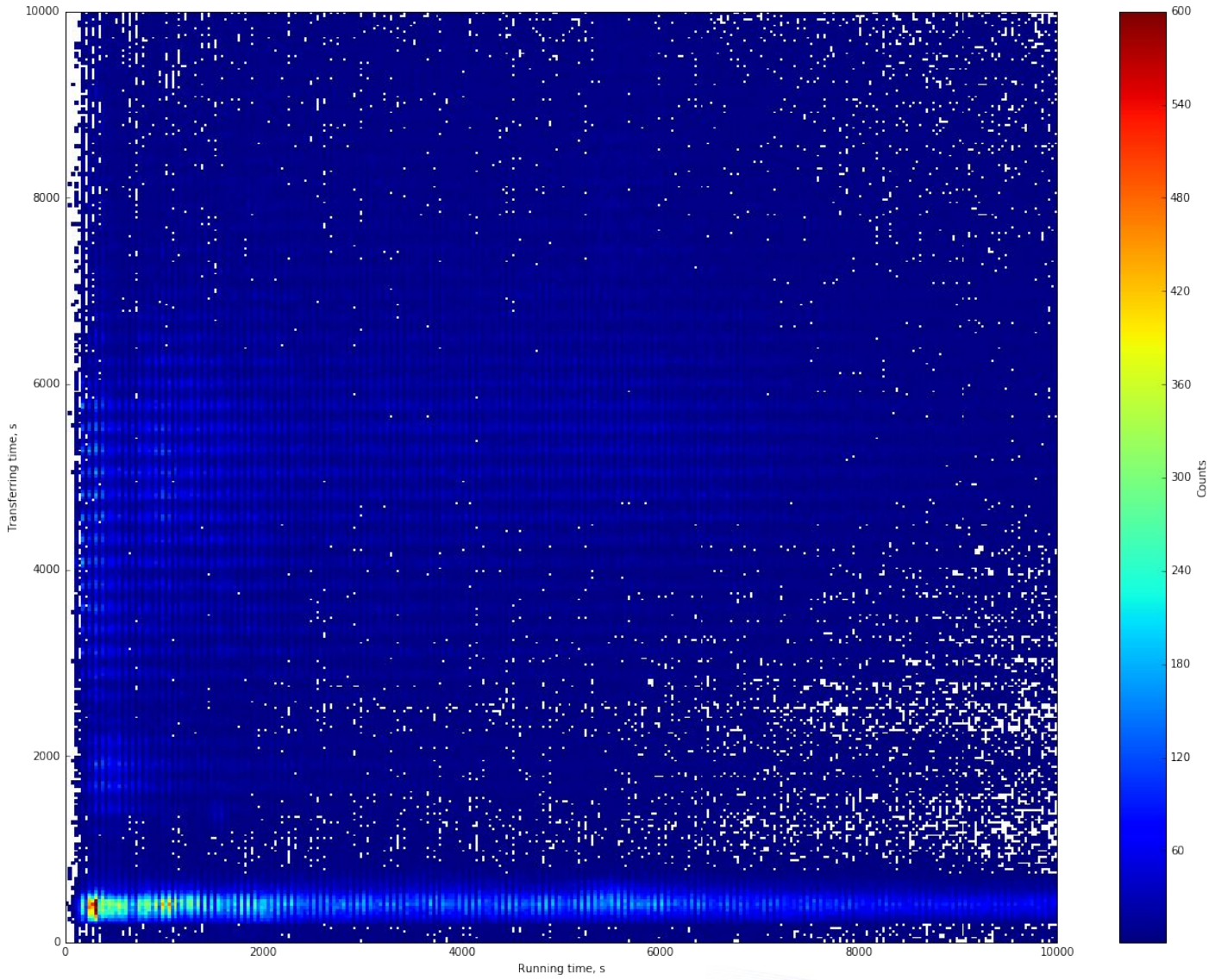


"STOVE OWNERSHIP" FROM XKCD BY RANDALL MONROE

econophysica

Matplotlib: 2-мерная гистограмма

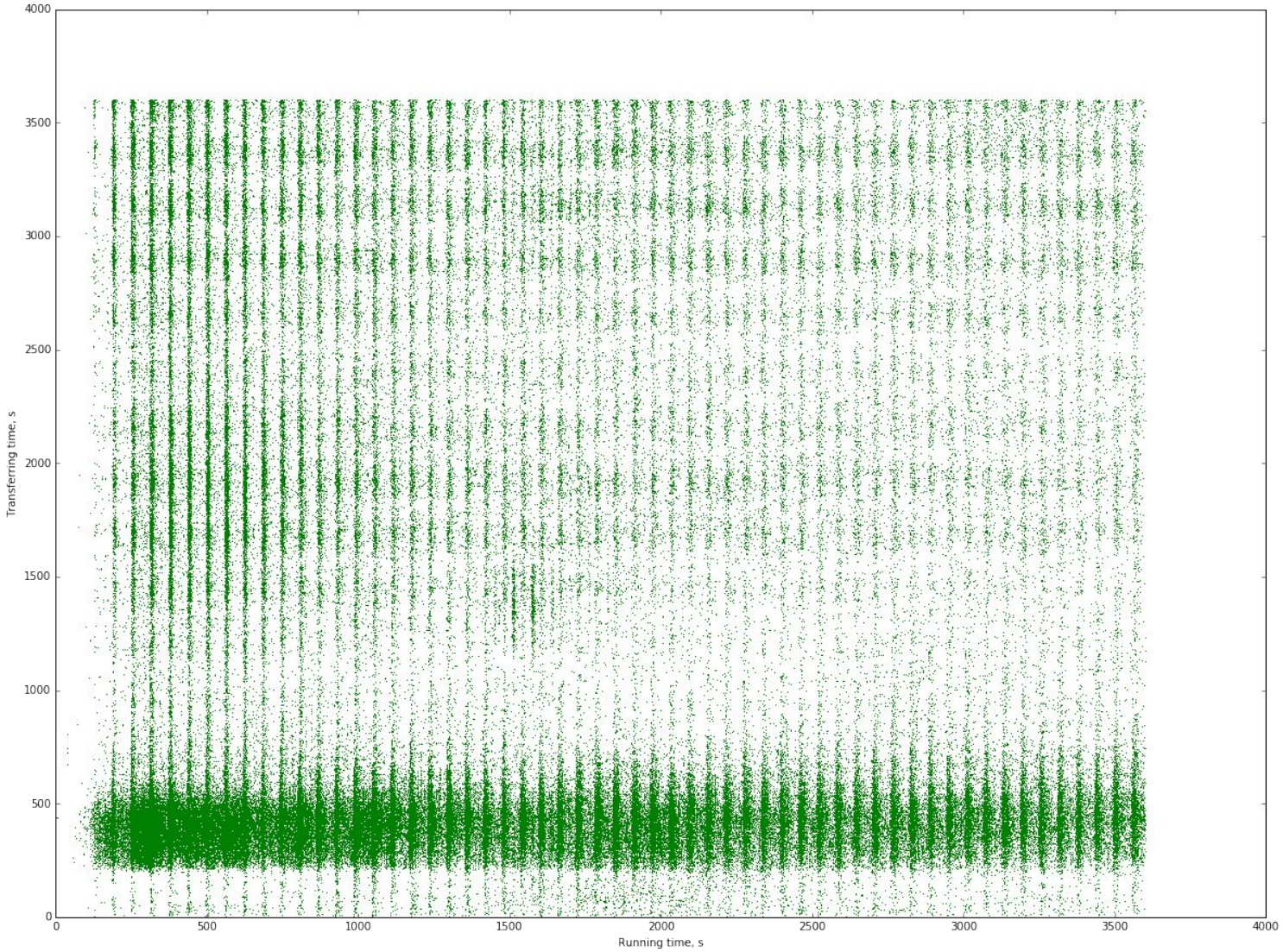
matplotlib



econophysica

Matplotlib: диаграмма рассеяния

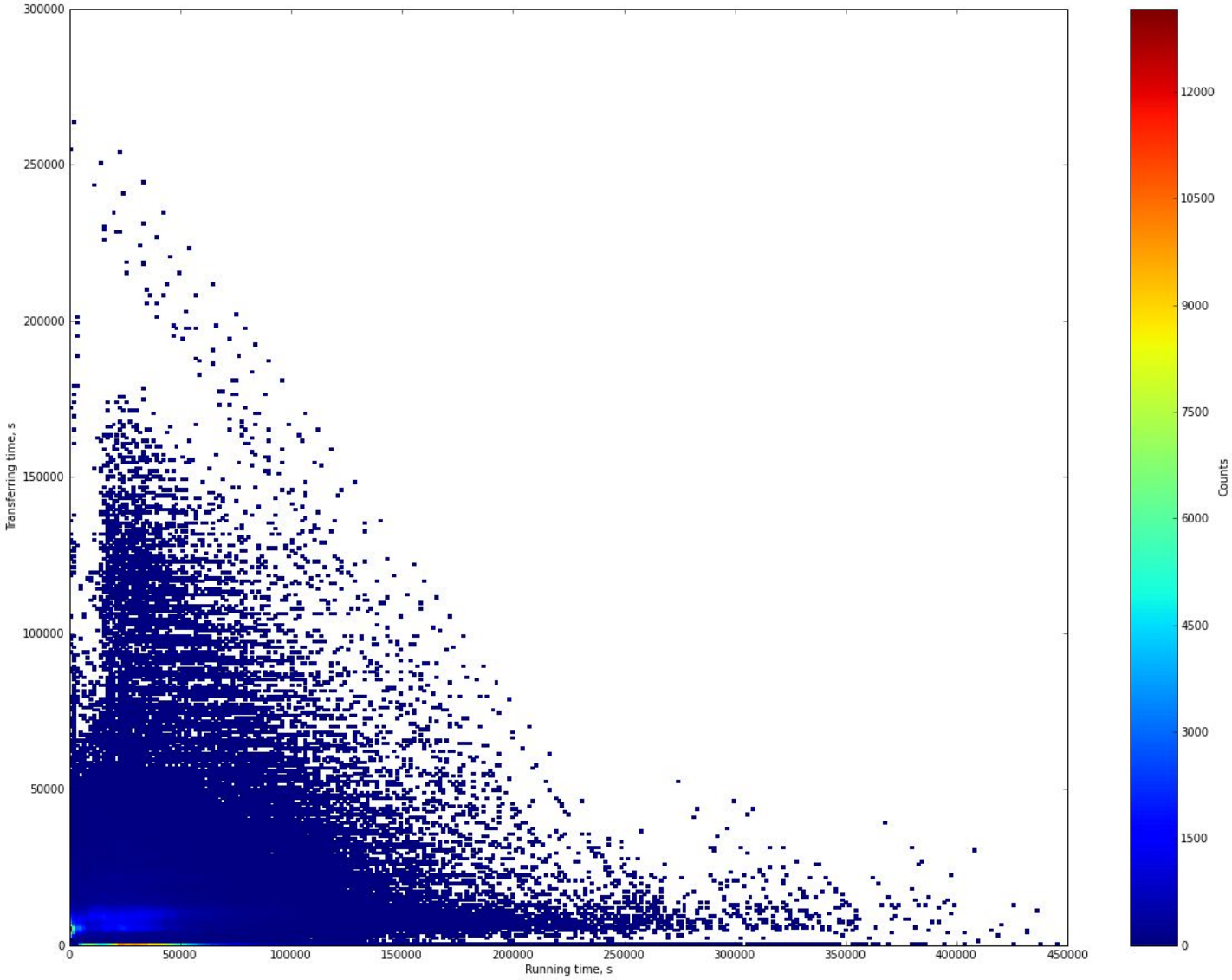
matplotlib



econophysica

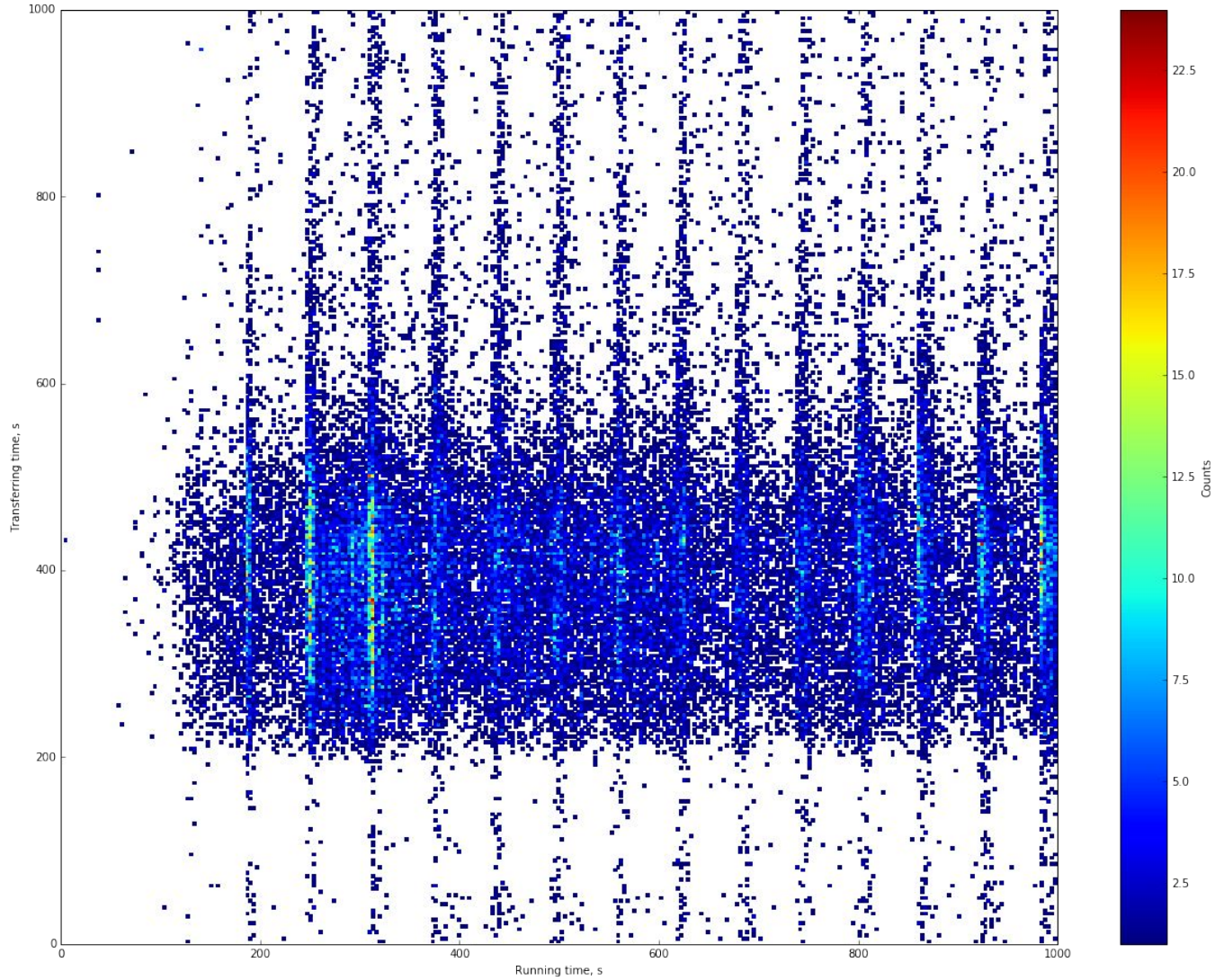
Matplotlib: 2 миллиона элементов на 2d-гистограмме

matplotlib

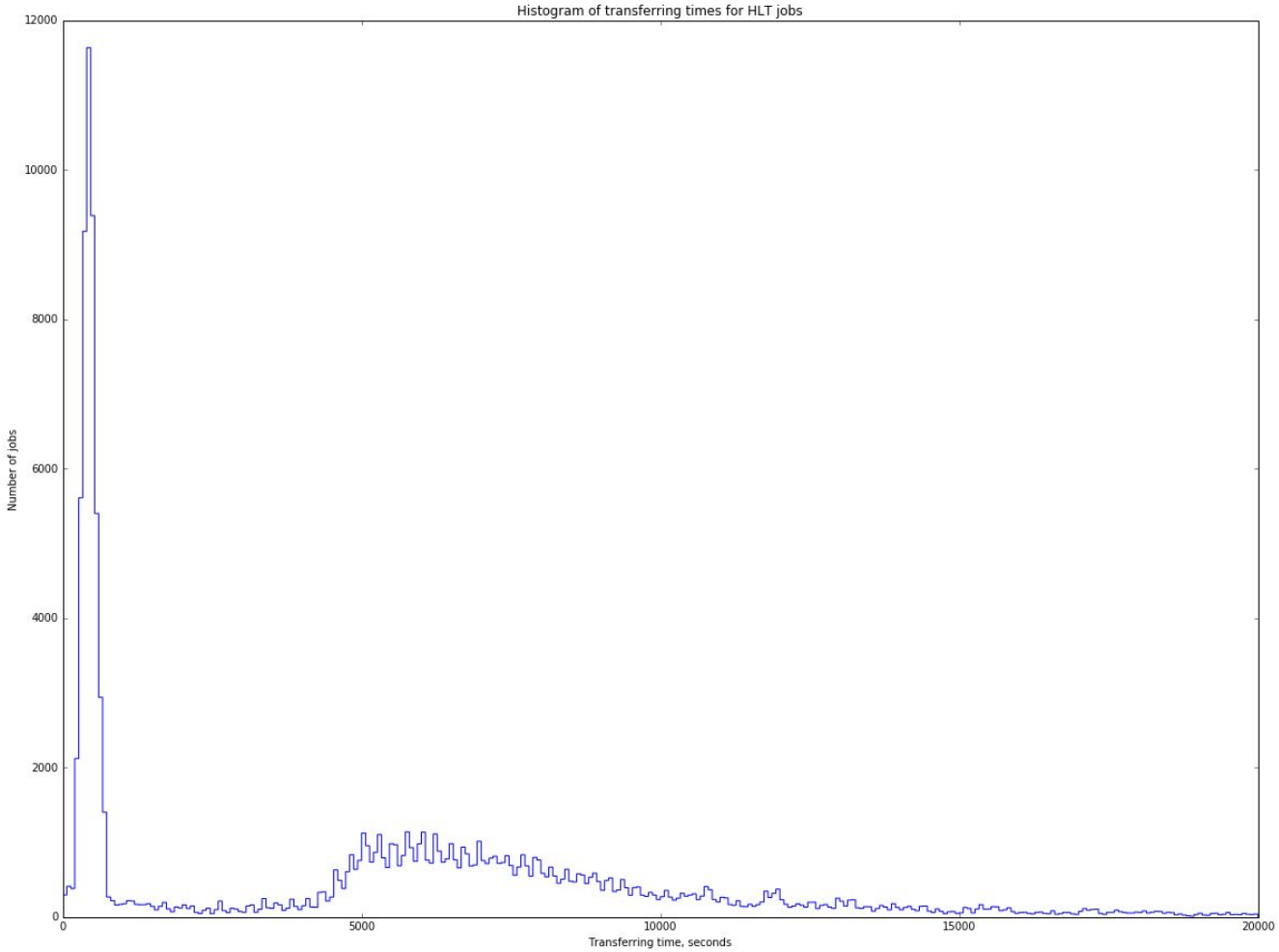


econophysica

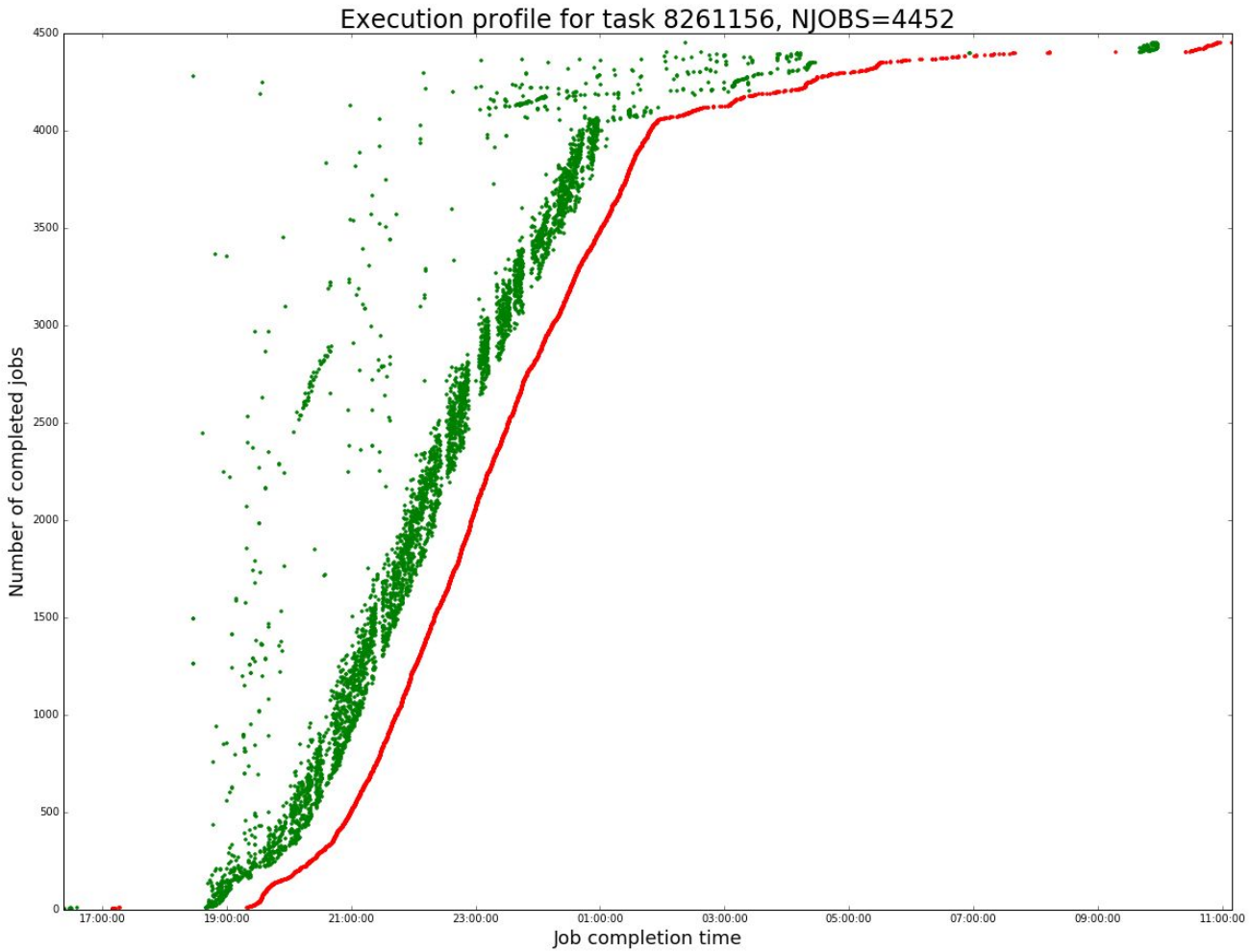
Matplotlib: интересные закономерности поведения задач WLCG



Matplotlib: гистограмма



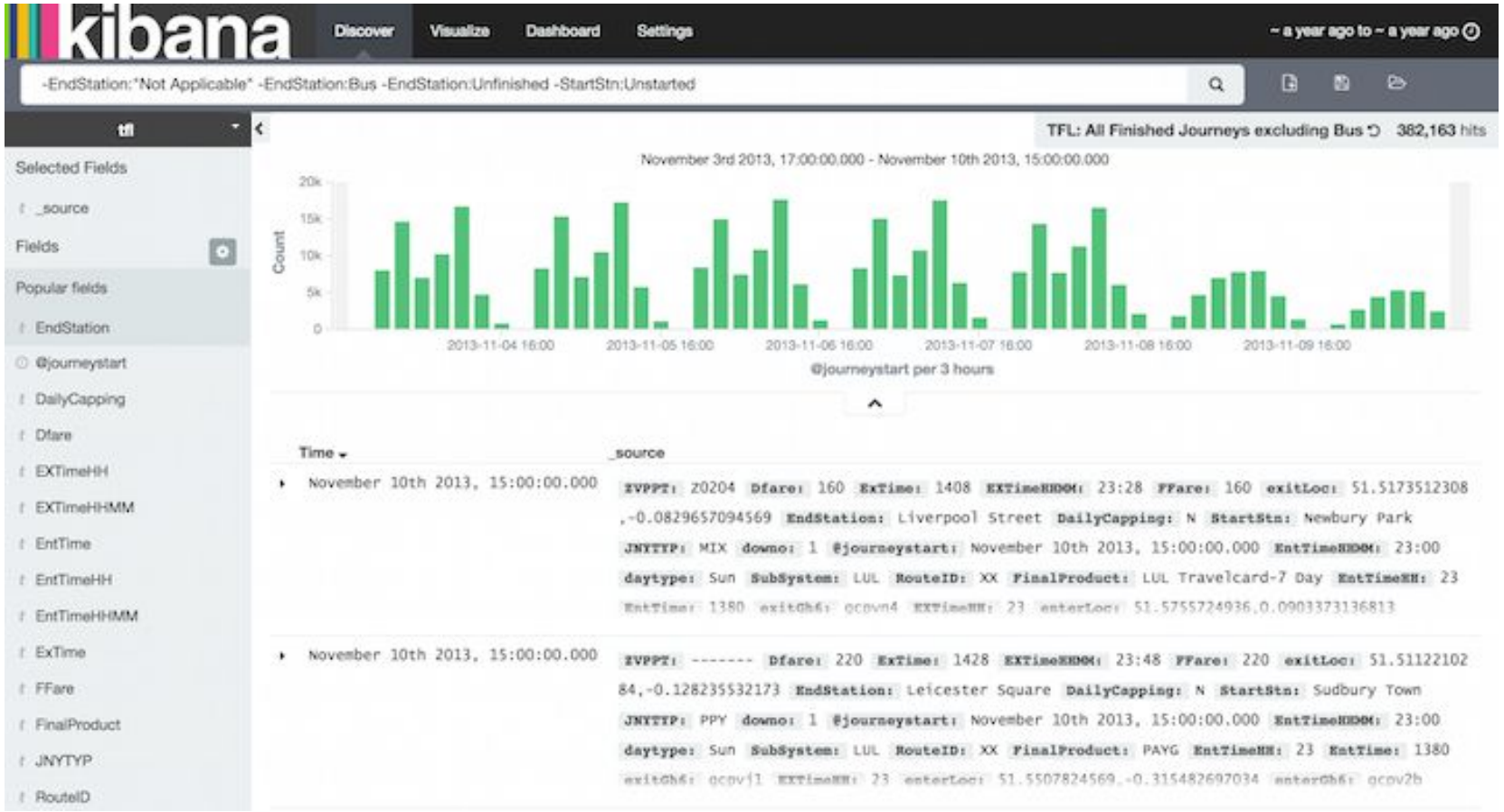
Matplotlib: диаграмма рассеяния



Kibana

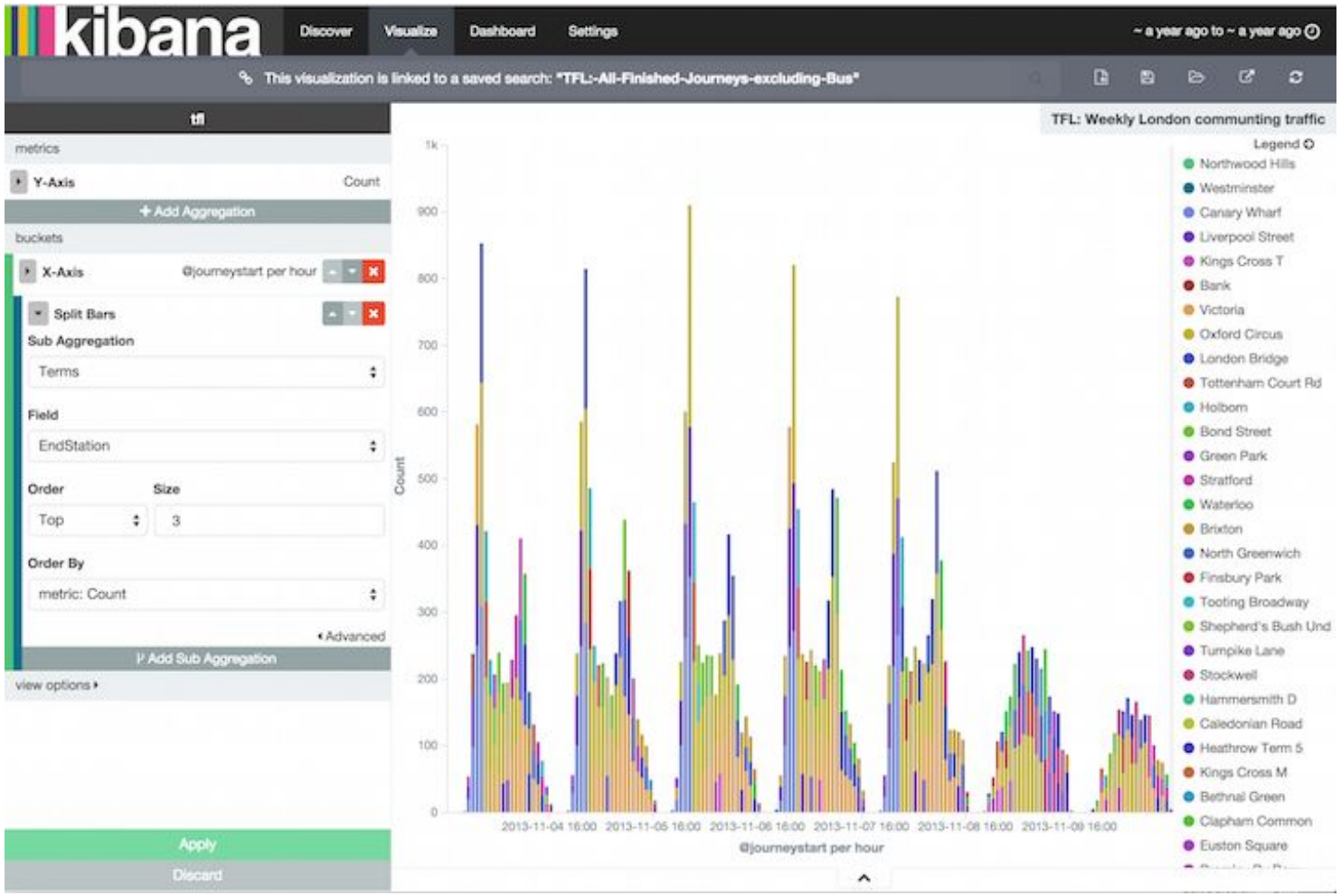
- Веб-приложение для исследования и визуализации данных
- Современный браузерный интерфейс (HTML5 + JavaScript)
- Поставляется с собственным веб-сервером для простой настройки
- Полная интеграция с Elasticsearch

Kibana: обнаружение данных



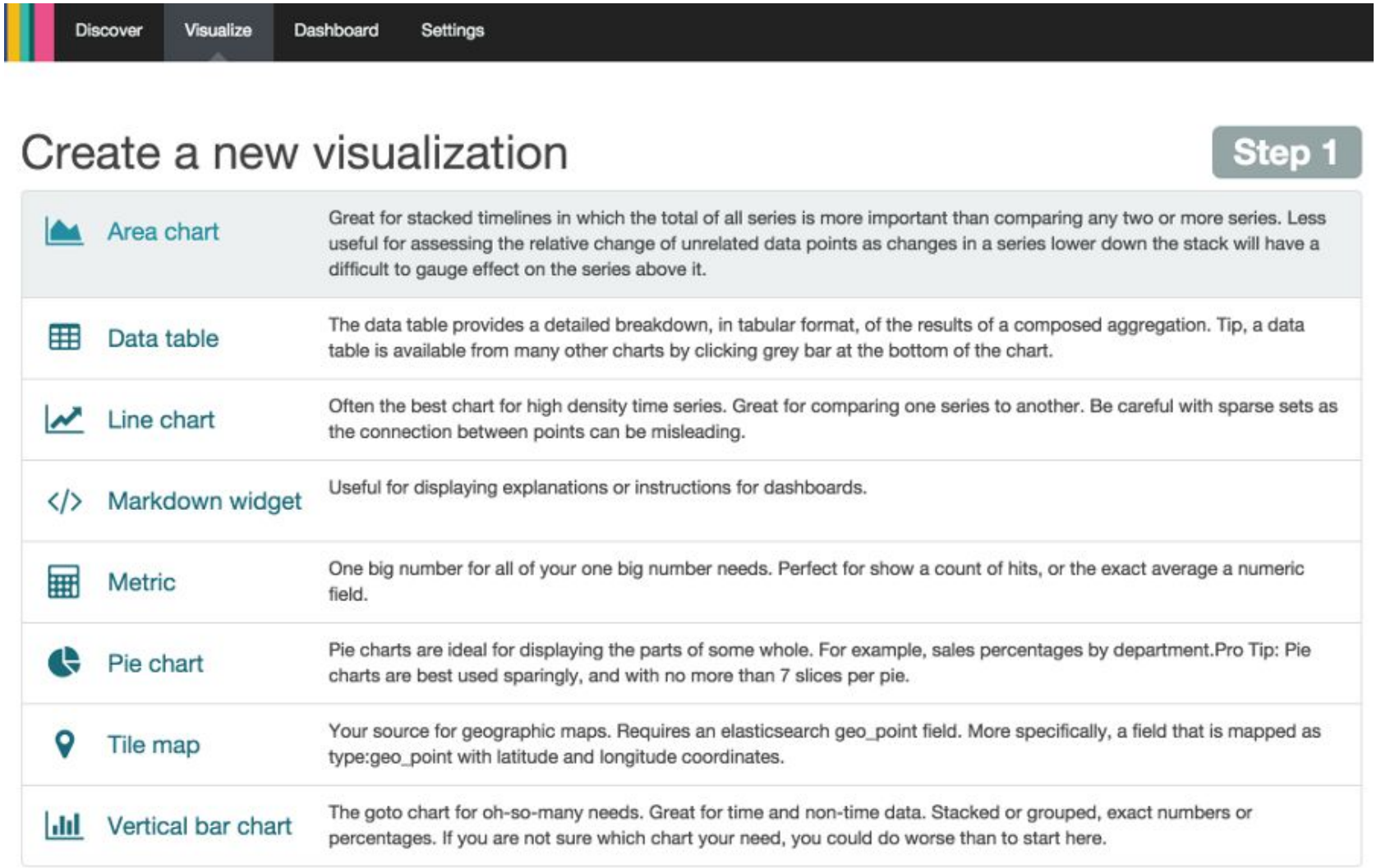
econophysica

Kibana: создание визуализаций

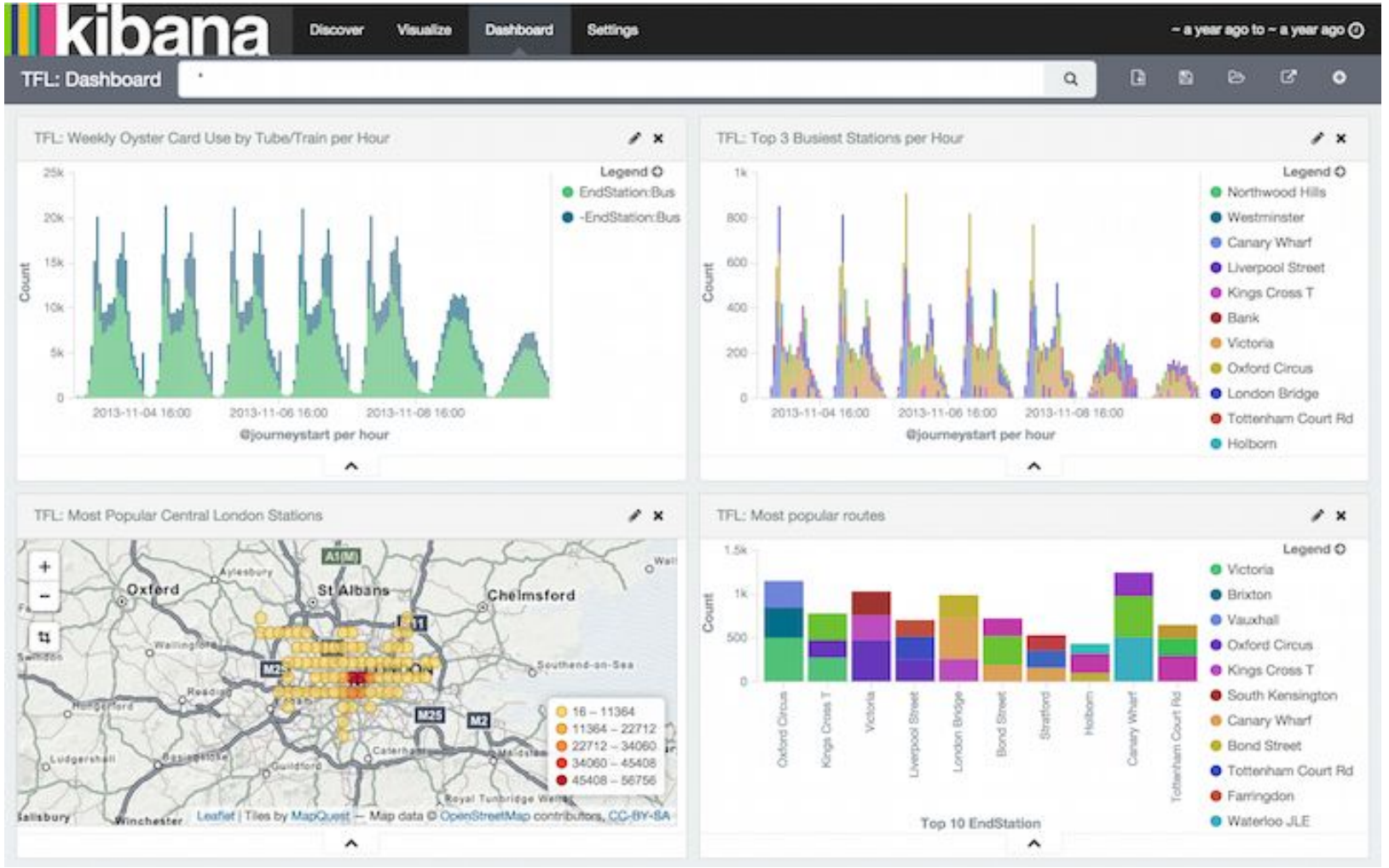


econophysica

Kibana: разные типы визуализаций

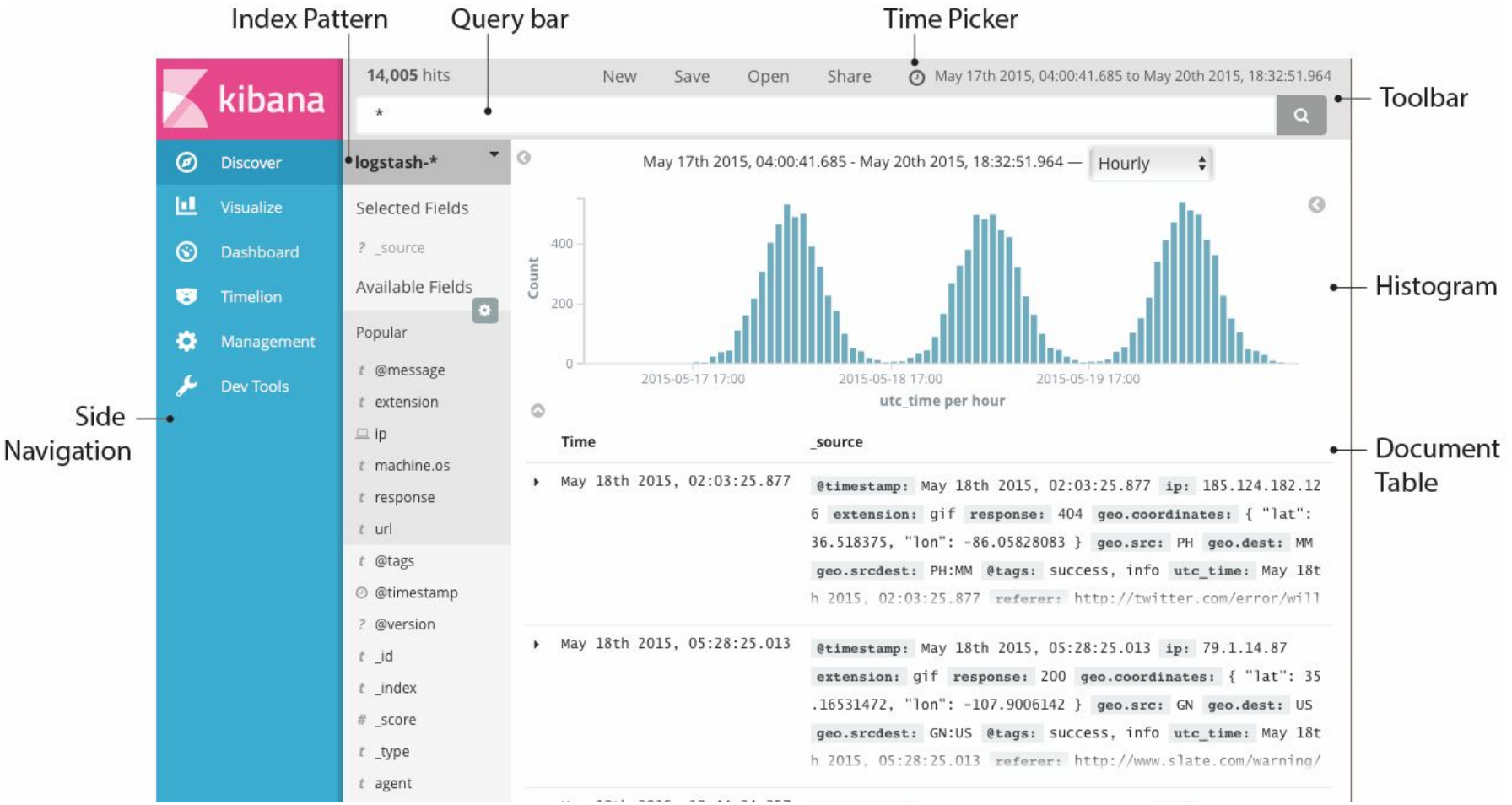


Kibana Dashboard



econophysica

Kibana: основные элементы интерфейса



econophysica

Заключение

- Помните: большинство всех возможных визуализаций больших данных на самом деле не делаются с самими большими данными, они обычно основаны на предварительно обработанных выборках данных.
- Вы не обязаны использовать специализированные инструменты для визуализации, если вам не нужны ваши графики **быстро**.
- Таким образом, вы можете свободно использовать любые инструменты визуализации, которые вы предпочитаете.
- Кибана не обязательна, но удобна.

Спасибо за внимание!

mgubin@tpu.ru

econophysica