Визуализация

- Наглядное представление результатов анализа данных
  - Имеет принципиальное значение для их интерпретации
  - Причина: восприятие человека ограничено
- Ученые продолжают вести исследования в области совершенствования современных методов представления данных в виде
  - изображений,
  - диаграмм
  - Анимаций
- Ничего нового здесь придумать уже невозможно, но на самом деле это не так

Почему важна визуализация

Визуализация позволяет видеть то, что иначе сложно заметить. Да, информация присутствует в данных, но без визуализации вы не замечаете зависимостей.

- Дает ответы на многие вопросы **быстрее**. В простейшем случае гораздо проще посмотреть на график и увидеть, чем на колонку цифр.
- Хорошая визуализация позволяет “исследовать данные”, поиграть с ними, выявляя интересные вещи, что особенно важно в исследованиях.
- В наше время объемы данных растут с невероятной скоростью. Визуализация помогает справиться с возрастающей сложностью и разнообразием данных.
- Все любят смотреть на интригующие цветные картинки, но практически никто не любит скучные таблицы с цифрами. Субъективное восприятие информации, доверие к информации выше, когда она подана визуально.

Области использования визуализации

- **Статистика и отчеты**. Данные за некий период времени показываются вместе. Например, статической картинкой в приложении к отчету или настраиваемым графиком в сервисе статистики, с возможностью изменения параметров его отображения.
- **Справочная информация**. Дополнение к основному тексту, наглядно иллюстрирующее его упоминаемыми данными. Например, дать общее представление о динамике одного из показателей, либо отобразить какой-то процесс и его этапы; может быть — показать структуру некоего явления.
- **Интерактивные сервисы**. Продукты и проекты, в которых инфографика является частью функциональности. Так, в качестве средства навигации по сервисам может являться диаграмма процесса. Почти все, что связано с работой с картами специализированных системах вроде диспетчерских и большей части компьютерных игр.
- **Иллюстрации**. Красивое отображение данных для создания самостоятельных иллюстраций.
- **Чертежи и схемы**. Специализированные документы, показывающие структуру и процесс работы сложных инженерных и природных систем
- **Эксперименты и искусство**. Визуализация данных в виде сложных и громоздких изображений, которые сложно "прочитать" бегло — объем данных и взаимосвязей между ними таков, что нужно разбираться с картинкой по частям; либо просто абстрактные изображения, автоматически сгенерированные. В последнее время направление все более популярно и периодически выходит за рамки компьютерной графики — например, в виде графиков-скульптур.

Визуализация больших данных

- Визуализация «больших данных»
  - Big Data Visualization
  - Large Data Visualization
- Области
  - научная визуализация
    - «Большие данные» возникают в результате сложного компьютерного моделирования различных объектов и процессов
  - информационная визуализация
    - Визуальное описание и представление абстрактной информации, получаемой в результате процесса сбора и обработки многокатегориальных данных, для анализа которых необходимо применение нескольких количественных и качественных мер оценки.
  - визуализация программного обеспечения

Задачи визуализации “больших данных”

- визуализация потоков данных;
- визуальный интеллектуальный анализ данных;
- визуальный поиск и рекомендации;
- описание ситуаций на основе больших данных с использованием визуализации;
- **масштабируемые методы параллельной визуализации**;
- современные аппаратные средства и архитектуры для анализа и визуализации данных;
- человеко-компьютерный интерфейс и визуализация больших данных;
- приложения визуализации больших данных, включая кибер разведку и контрразведку, бизнес-анализ (бизнес разведку), электронную коммерцию, анализ научной информации, образование

Требования

- оценка пригодности (адекватности в визуализации) видов отображения естественность (привычность для пользователей)
- устойчивость к масштабированию
- возможность вывода сверхбольших объемов данных
- возможности для представления сложных структур, а
- также объектов особого интереса, особых точек, аттракторов, сингулярностей

Сингулярность (особенность) — точка, в которой математическая функция стремится к бесконечности или имеет какие-либо иные нерегулярности поведения.

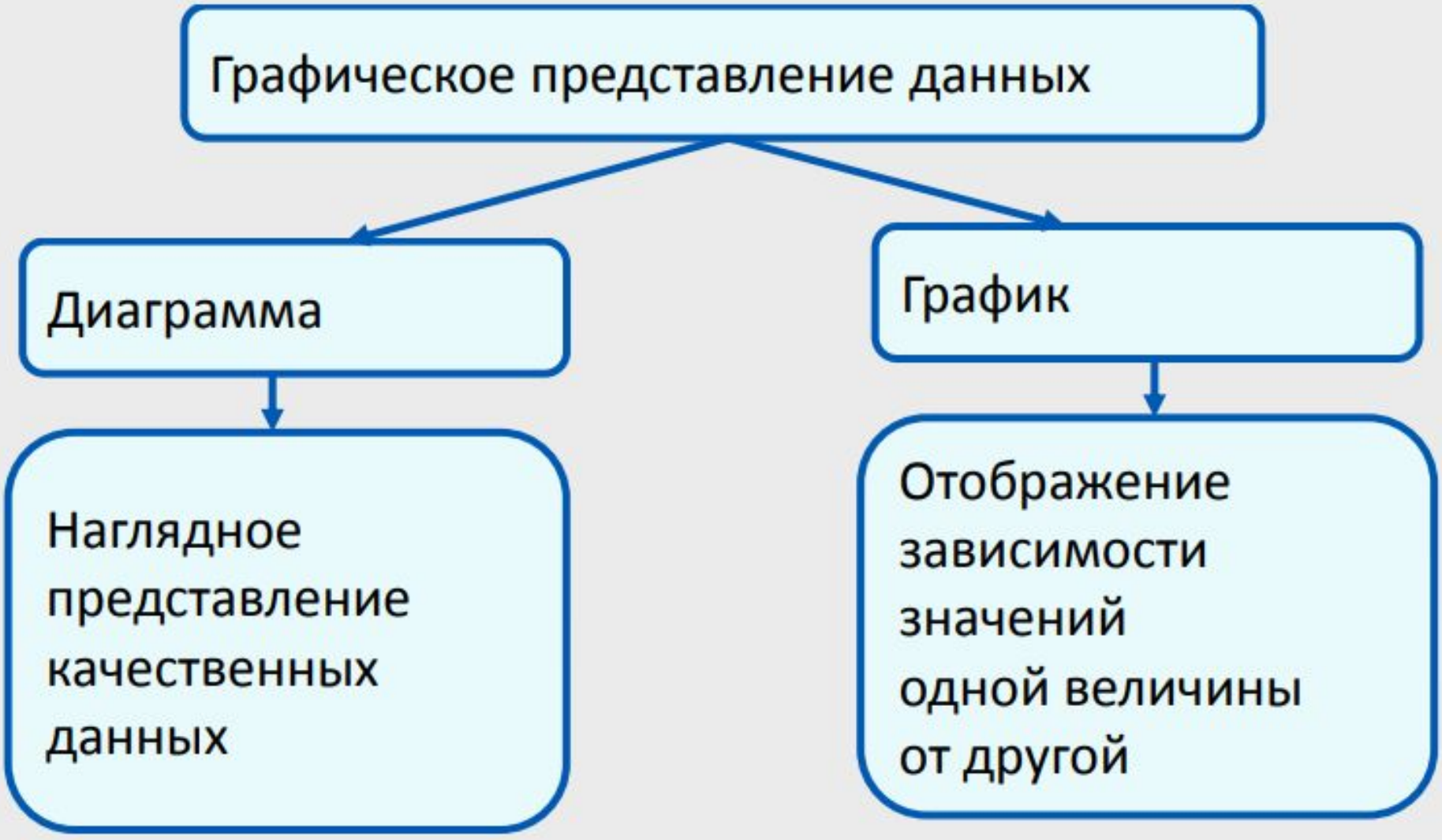
Решение проблемы визуализации БД

- Использование достаточно простых, но четко интерпретируемых методов представления
- Формализация описания объектов программного обеспечения
- Верификация и валидация визуализации

Простые способы визуализации

Выполнить анализ и визуализацию данных в таблице позволяют сортировка, поиск и графическое отображение.

Простые способы визуализации



Традиционные виды визуализации

- Графики
- Диаграммы сравнения
- Деревья и структурные диаграммы
- Карты
- Диаграммы связей
- Деревья и структурные диаграммы
- Диаграммы визуализации процесса
- Матрицы
- Диаграммы времени
- Карты
- Диаграммы связей
- Иллюстрации

Еще одна классификация

- Графики и диаграммы
- Инфорграфика и схемы
- Презентация и анализ данных
- Интерактивный сторителлинг
- Бизнес аналитика и дашборды
- Научная и медицинская визуализация
- Карты и картограммы

Графики и диаграммы

- Наверное самый привычный для нас вид визуализации данных. Используется как для презентации данных, так и для анализа. Встретить их можно и на работе, и в журнале и в научном отчете. Обычно знания о существующих типах диаграмм и графиков мы получаем из школы или из стандартного набора в экселе. Однако, мало кто знает, что мир графиков и диаграмм не ограничивается точечным графиком, столбиковой и круговой диаграммой. Существуют порядка 15 общеизвестных типов диаграмм, а всего их более 60, при этом их количество увеличивается с каждым днём — люди придумывают новые типы для визуализации сложных и необычных данных.

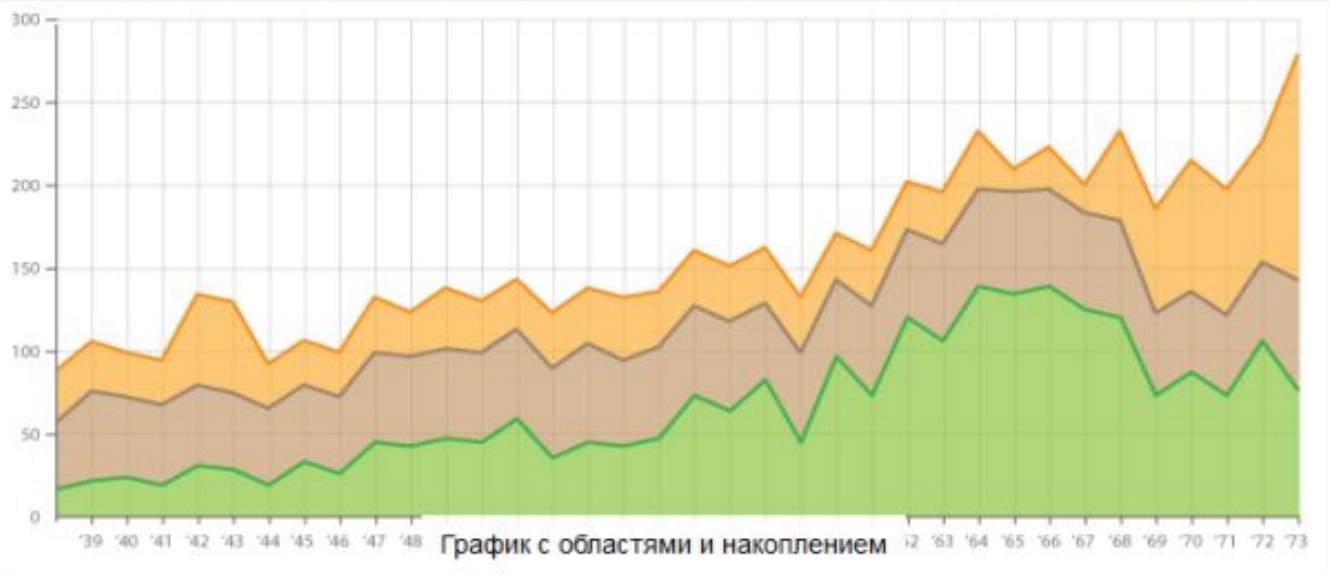
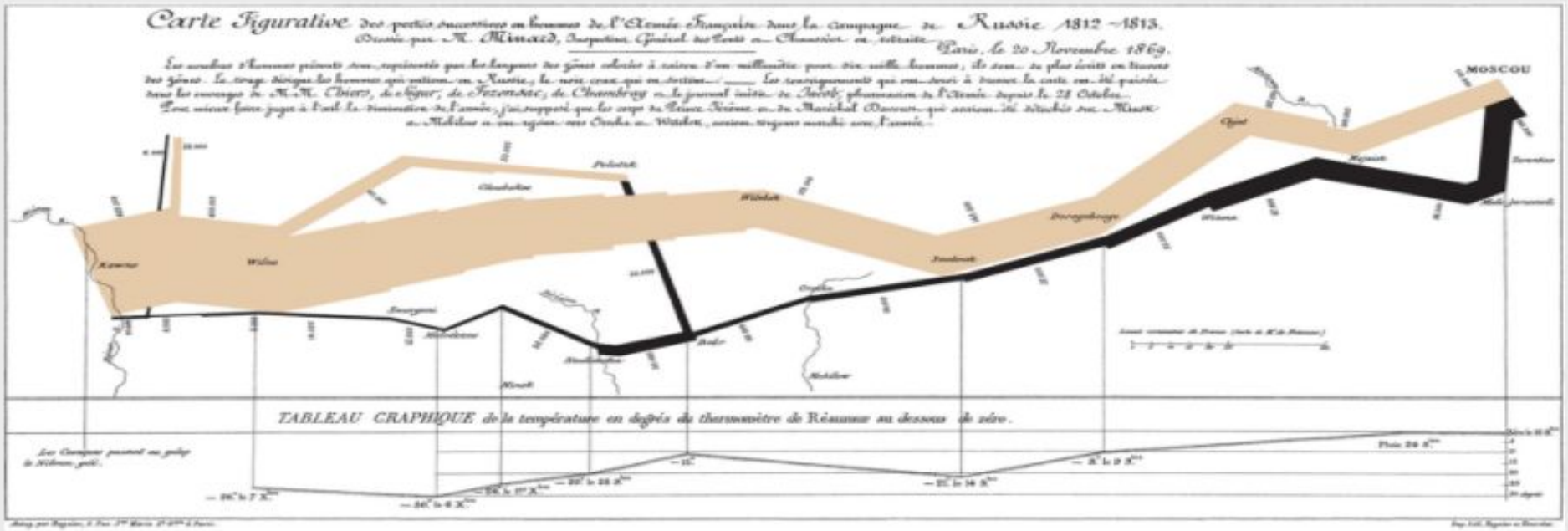


График с областями и накоплением

Инфографика

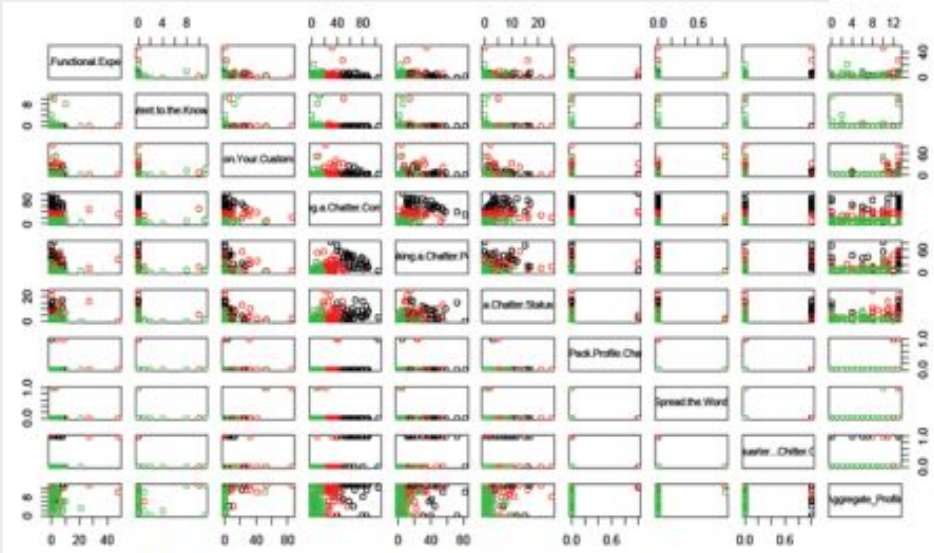
- Инфографика стала очень популярна в последние годы, хотя существуют уже давно. Инфографика относиться к журналистике данных, где графики и схемы объясняют какие-либо факты по выбранной теме. Обычно инфографика статична и представляет собой длинную «простыню» с картинками и текстом. Отличительной особенностью инфографики является то, что в ней приводятся уже готовые выводы, то есть читателя проводят за руку по выбранной теме и при этом приправляют это все цифрами и картинками. Часто используется рисованный или мультяшный стиль. Часто используется не к месту или «для красоты», хотя конечно же есть замечательные и интересные примеры.



Презентация и анализ данных

- Один самых привычных способов использования визуализации данных — презентация информации в виде диаграмм или инфографики. И если с этим все понятно, то использование визуализации для анализа информации, в основном, используется только бизнес-аналитиками и учеными. В чем заключается отличие?
- При анализе данных с помощью визуализации используют так называемое быстрое прототипирование — то есть создание **большого количества различных визуальных представлений одних и тех же данных.** Делается это для возможности нахождения скрытых, на первый взгляд, взаимосвязей и зависимостей, а также первичной оценки набора данных для возможности применения в дальнейшем более сложных инструментов анализа. Этот подход называется Exploratory data analysis (EDA), что на русский можно перевести как разведочный анализ данных. Основное отличие от презентации данных — визуализация здесь может быть «черновой» и некрасивой, но выполняется быстро и одним человеком или небольшой рабочей группой. Для этого чаще всего используют эксель, R или матлаб

Презентация и анализ



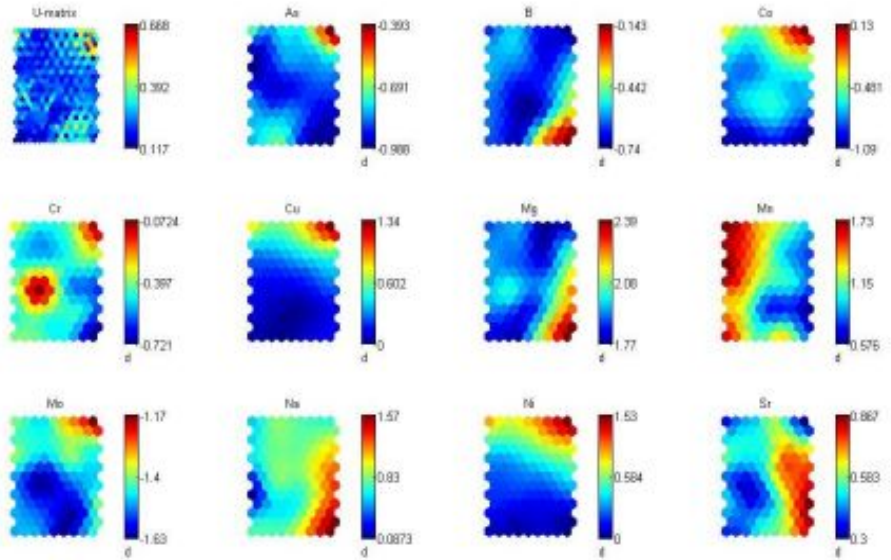
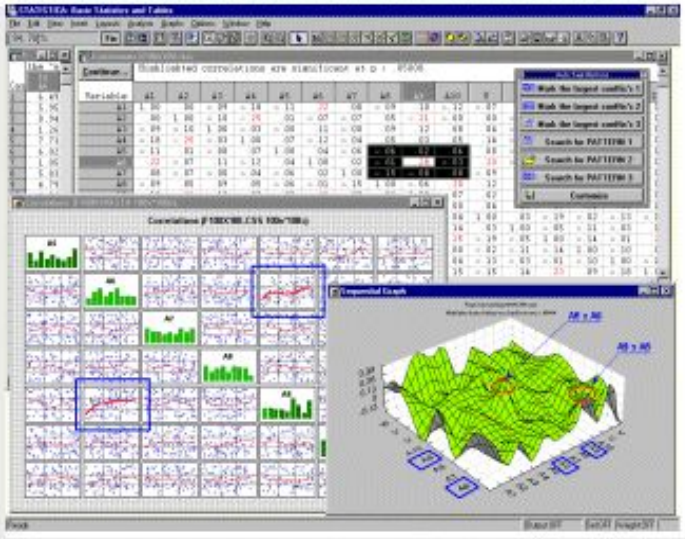
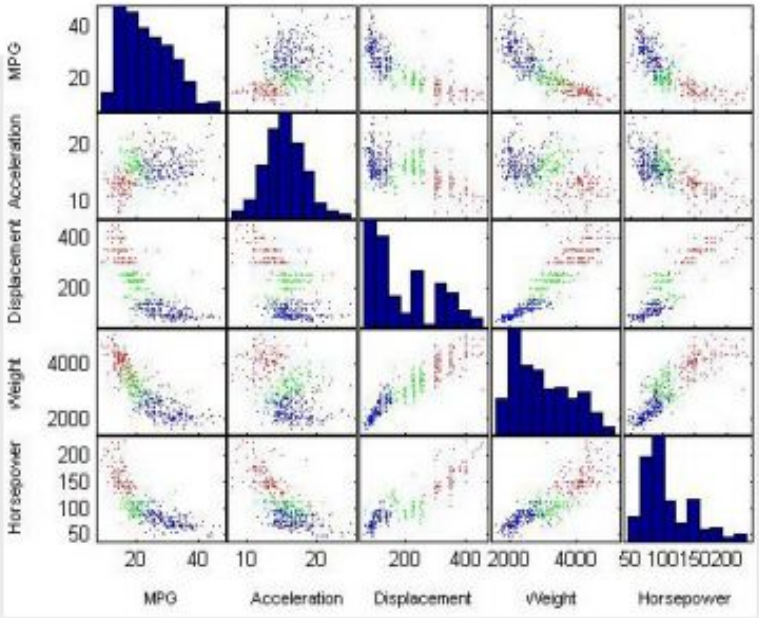


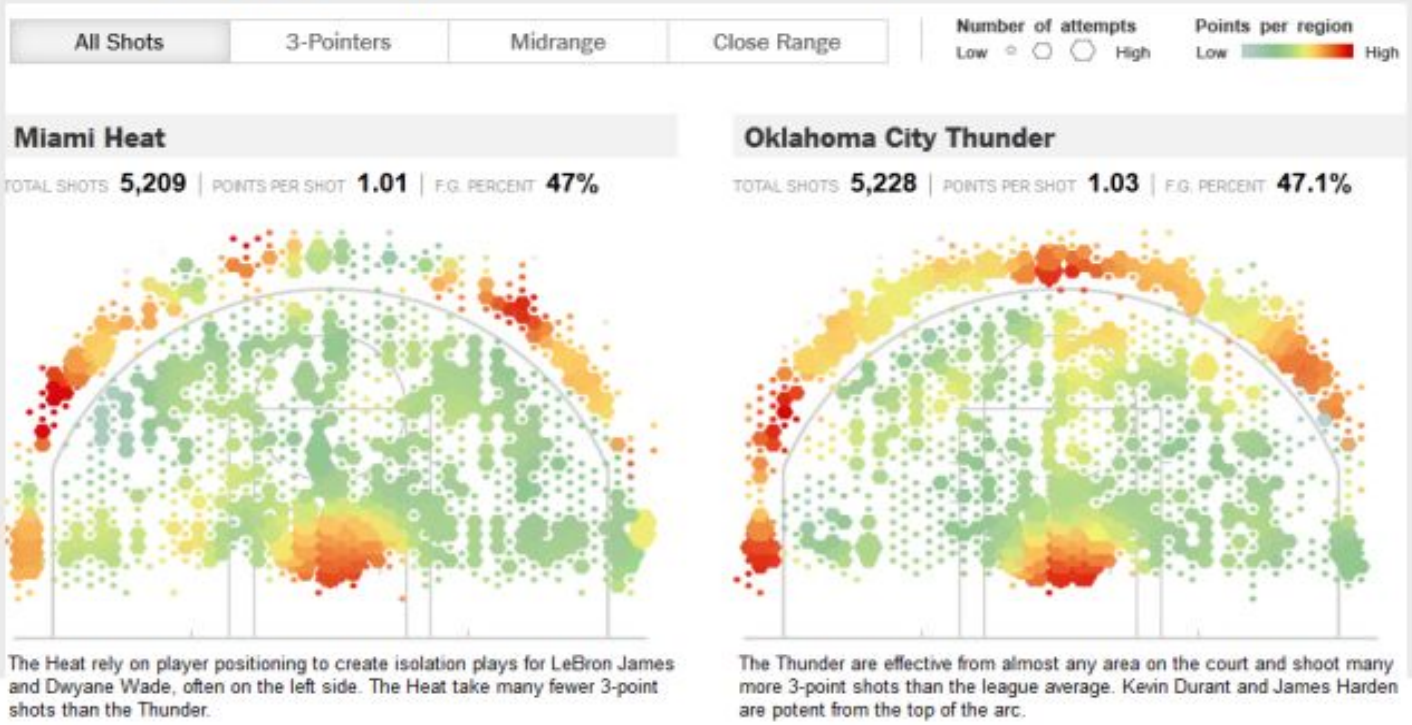
Figure 4 Clustering Analysis of Salesforce usage for Bunchball client





Интерактивный сторителлинг

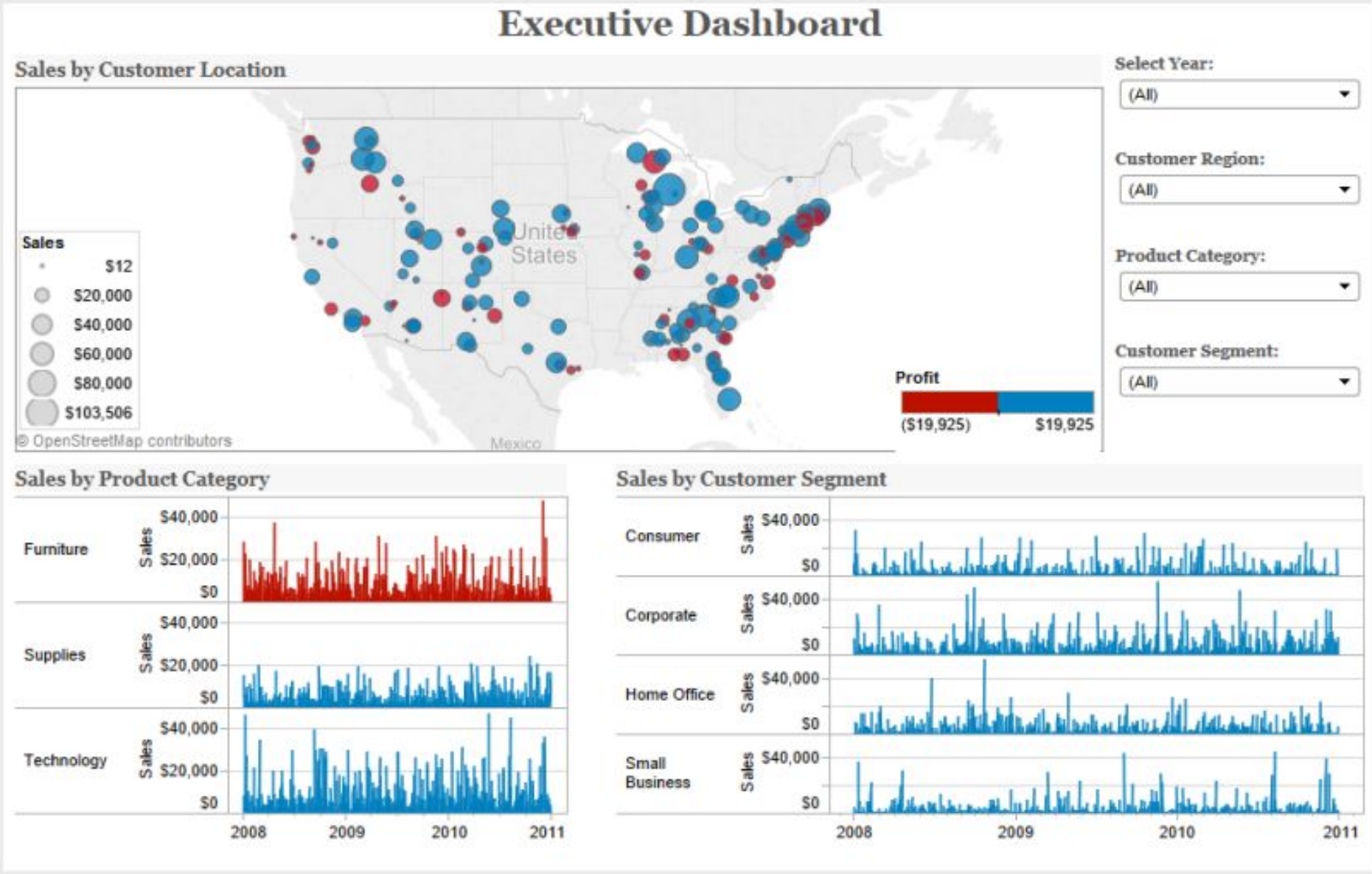
- Сторителлинг — это преподнесение какой-либо полезной информации в форме интересного рассказа. Интерактивный сторителлинг — рассказ с которым слушатель может взаимодейтсовать. Пользователь может управлять отображением информации и находить те зависимости, которые не нашёл автор. В этом смысле он близок к разведочному анализу данных, но отличается тем, что данные заранее обработаны и представлены в удобном для анализа виде, а также имеются подсказки или заранее прописанные сценарии использования. Поэтому, чаще всего интерактивный сторителлинг называют интерактивной инфографикой, но для того чтобы ей стать не достаточно просто к статичной инфографике добавить всплывающие окошки.



Бизнес аналитика и дашборды

- Визуализация активно используется в бизнесе. Принцип «говорите с данными» помогает компаниям зарабатывать больше, а клиентам получать лучший сервис. Для разового анализа обычно используется эксель или R. Однако это не удобно если необходимо следит за какими-то показателями (KPI) на постоянной основе. Для отслеживания рутинных KPI используют дашборды — дисплеи на которых выведены все необходимые показатели в одном месте в виде графиков, диаграмм и таблиц.
- Проектирование эффективных дашбордов — сложная и неординарная задача. Зачастую их перегружают ненужной информацией или стараются использовать все возможные типы шаблонных графиков. Часто для того чтобы спроектировать хороший дашборд необходимо создание новых типов визуализации информации. Тематика активно развивается за счет все большего применения аналитики в бизнесе. Также дашборды применяются и для личного использования (фитнес трекеры, анализ личных расходов и т. п.)

Пример Dashboard



Научная и медицинская визуализация

- Специфический вид визуализации, который используется как следует из названия в медицине и науке. Его целью обычно является выделение закономерностей или аномалий. От обычной визуализации данных отличается тем, что часто бывает трёхмерной и требует специальной подготовки для интерпретации.

Карты и картограммы

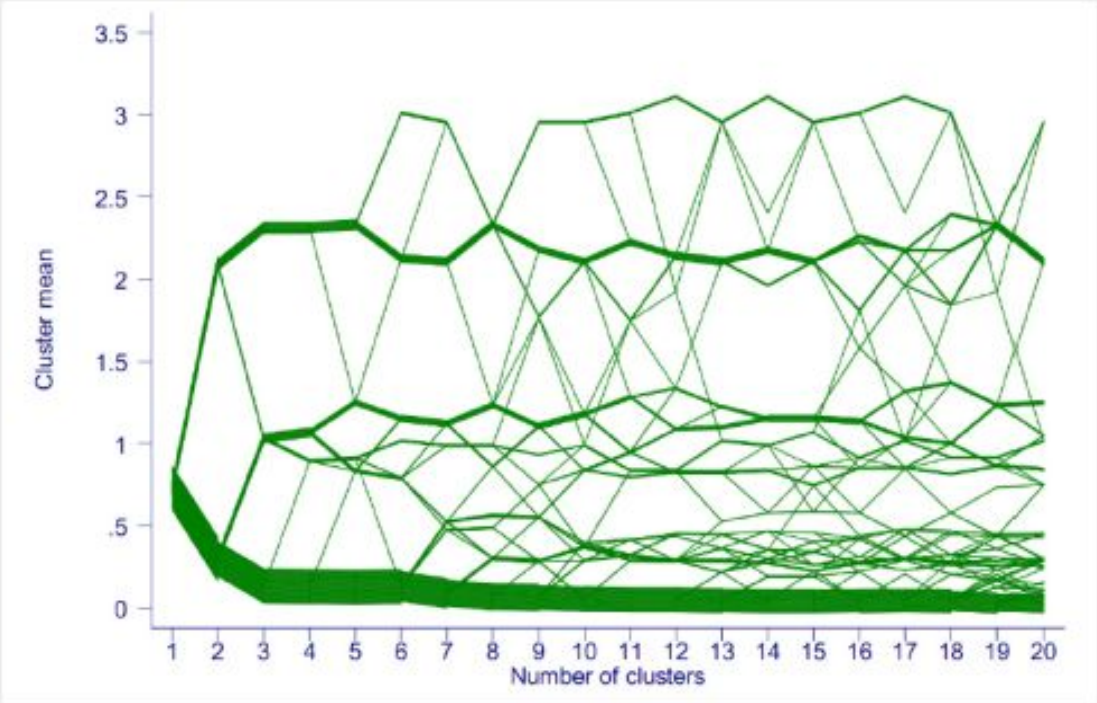
- Карты — одни из древнейших способов визуализации, отображающих окружающую реальность. Картограмма — карта с нанесенной на неё информацией в виде цвета или других способов. Картограммы могут быть использованы для отображения любой информации — от плотности населения, до частоты использования мобильных телефонов в каждом районе страны.

Облако тегов

- Каждому элементу в облаке тега присваивается определенный весовой коэффициент, который коррелирует с размером шрифта. В случае анализа текста величина весового коэффициента напрямую зависит от частоты употребления (цитирования) определенного слова или словосочетания.
- Позволяет читателю в сжатые сроки получить представление о ключевых моментах сколько угодно большого текста или набора текстов.

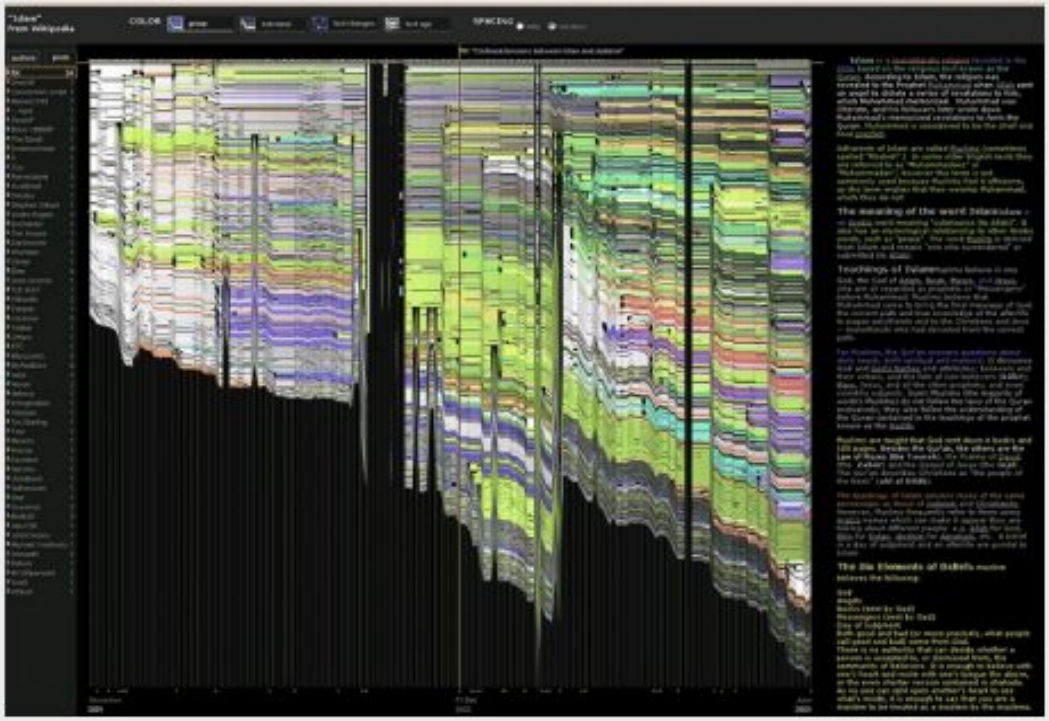
Кластергармма

- Метод визуализации, использующийся при кластерном анализе
- Показывает как отдельные элементы множества данных соотносятся с кластерами по мере изменения их количества
- Выбор оптимального количества кластеров – важная составляющая кластерного анализа



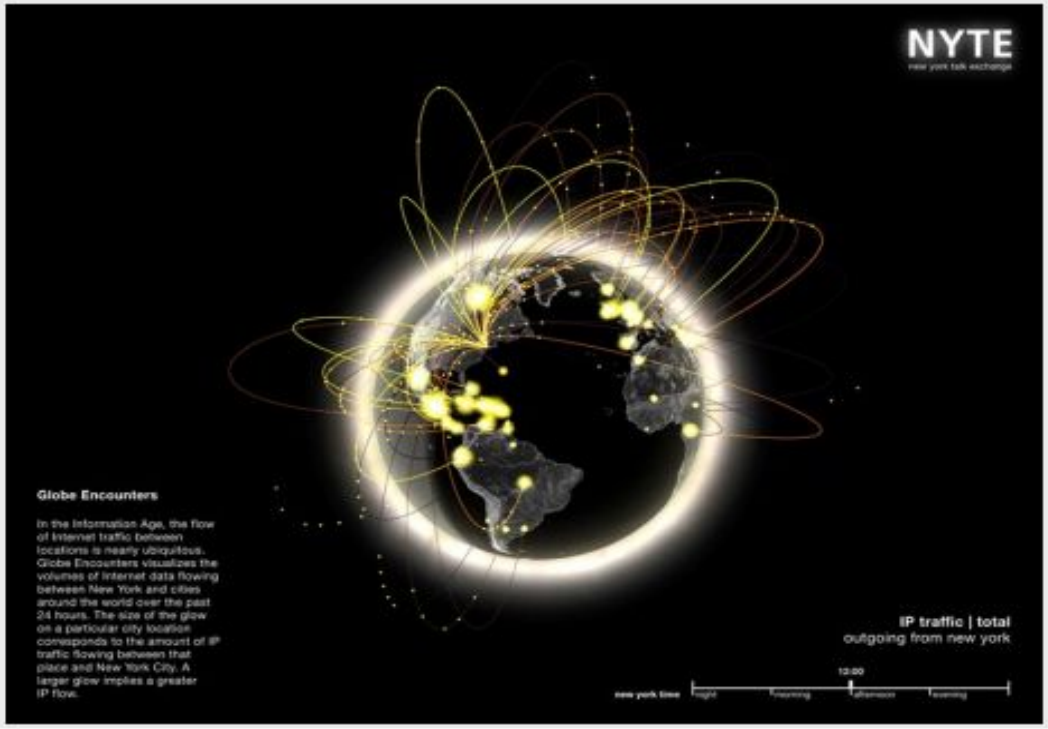
Исторический поток

- Помогает следить за эволюцией документа, над созданием которого работает одновременно большое количество авторов. В частности, это типичная ситуация для сервисов wiki в том числе.
- По горизонтальной оси откладывается время,
- По вертикальной – вклад каждого из соавторов, т.е. объем введенного текста.
- Каждому уникальному автору присваивается определенный цвет на диаграмме. Приведенная диаграмма – результат анализа для слова «ислам» в Википедии. Хорошо видно, как возрастала активность авторов с течением времени.



Пространственный поток

- Эта диаграмма позволяет отслеживать пространственное распределение информации. Приведенная в качестве примера диаграмма построена с помощью сервиса New York Talk Exchange. Она визуализирует интенсивность обмена IP-трафиком между Нью-Йорком и другими городами мира. Чем ярче линия – тем больше данных передается за единицу времени. Не составляет труда выделить регионы, наиболее близкие к Нью-Йорку в контексте информационного обмена.



Онлайн сервисы (1)

- Lovely Charts Сервис позволяет создавать диаграммы всех видов, таких как блок-схемы, карты сайта, бизнес-процессов, организационных диаграмм, каркасы и многое другое.
- Gliffy. Сервис позволяет легко создавать блок-схемы, диаграммы, поэтажные планы, чертежи и многое другое.
- Google Chart Сервис позволяет создавать линейные графики, различные диаграммыдиаграммы
- Cacoo создание различных диаграмм, графиков, карт и др
- ChartGo быстрое создание диаграмм
- Create a Graph создание диаграмм
- Diagramly создание схем, диаграмм, информационных карт и др
- Google Ngram Viewer визуализация частоты упоминания
- Mapwing создание интерактивных туров с использованием различных объектов
- Mind42 Создание ментальных карт
- Mindomo создание ментальных карт

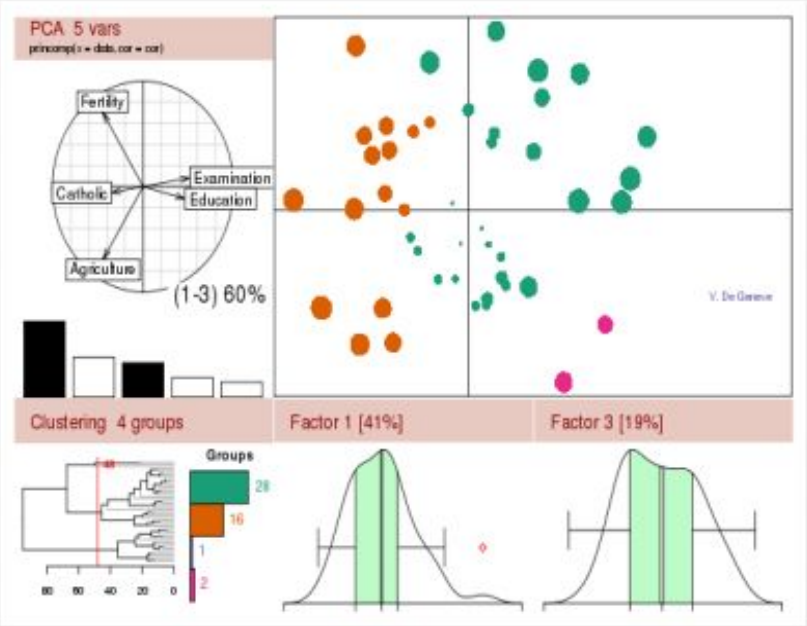
Онлайн сервисы (2)

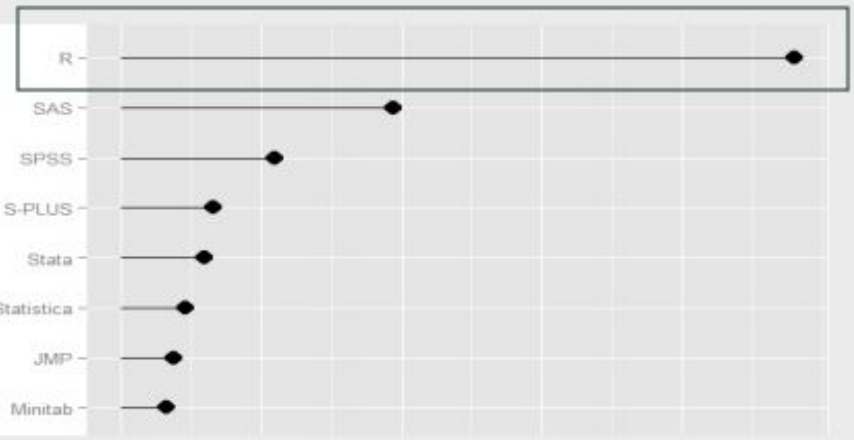
- Piecolor создание диаграмм
- Rich Chart Live создание "живых" диаграмм
- Spiderscribe_NET сервис для создания когнитивных карт
  Tagxedo Генерация облака слов с действующими ссылками поиска
- Ultimate FlashFace создание фотороботов
- WordCloud генерация облака слов сайта/блога по ссылке
  Word It Out генерация облака слов
  Wordle-net генерация облака ключевых слов
- Google карты используем и редактируем карты
  Quickmaps быстрое редактирование карт
- PinPoint привязка тегов к карте, информация, Карты от Google и Яндекса
  Wikimapia редактируем карты, объекты добавляем фото и комментарии

Визуальный анализ с использованием R

- R - интуитивно понятный язык сценариев
- Позволяет импортировать и использовать большое количество пакетов **для анализа и визуализации**.
- Для подобного основанного на модели анализа можно также использовать другие научные инструменты визуализации, такие как MATLAB или даже Gnuplot, но R содержит различные пакеты, хорошо выполняющие многомерный анализ наборов данных, не являющихся научными по своей природе (визуализация бизнес-аналитики)
- Инсталяция
  - http://www.r-project.org/ - интерпретатор языка R
  - http://www.rstudio.com/ - Rstudio, IDE для работы с R
  - igraph — пакет для анализа и визуализации данных

Проект R для статистических вычислений





- Язык для статистических исследован ий и работы с графикой (Росс Айхэк, Роберт Джентельмен, Оклендский ун-т, 1997)
- **Open source проект, R Foundation**
- Широкий спектр различных функций (временные ряды, прогнозирование, классификация, кластеризация и др.)
- Важное отличительное преимущество – простые средства построения самых сложных графиков и диаграмм
- Возможность расширения, технология разработки дополнительных пакетов участниками проекта

Отличие R

- Инструменты, подобные MATLAB, предоставляют среду интерактивного научного и инженерного анализа для исследования моделей и данных инженерам и ученым;
- R предоставляет те же средства бизнес-аналитикам и аналитикам больших данных всех типов.
- Возможность интерактивного исследования больших данных при помощи таких инструментов, как R и BigQuery, отличает анализ больших данных от пакетного и глубинного анализа данных, которые часто выполняются с помощью MapReduce.
- В любом случае целью является формирование новых моделей и поддержка принятия решений с использованием больших данных.

Основные возможности

- Можно вводить команды одну за другой в командной строке (>) или запустить последовательность команд из файла-источника
- Язык R содержит огромное количество различных типов данных, включая векторы (числовые, строковые, логические), матрицы, блоки данных и списки (vector, matrix, data.frame, list)
- Для выхода из R достаточно ввести команду >q()
- Гибкость языка R обеспечивается посредством встроенных и пользовательских функций. Во время работы с R все пользовательские данные хранятся в памяти программы
- Базовые функции доступны по умолчанию. Другие функции содержатся в особых статистических пакетах и могут быть загружены во время работы
- R содержит встроенные базы данных, которые можно использовать для обучения
  - > data( ) # Загрузка всех доступных пакетов
  - > help(datasetname)