Как визуализировать данные с помощью лиц Чернова

<https://docs.google.com/presentation/d/1MzO5TtLgBMnzAO_nckMGdFBW1L93bCV4fp2L0i-jasY/edit#slide=id.g1100790d91_0_50>

http://www.cs.middlebury.edu/~jwood/chernoff.html

<https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D1%86%D0%B0_%D0%A7%D0%B5%D1%80%D0%BD%D0%BE%D0%B2%D0%B0>

<https://searchengineland.com/my-data-have-feelings-49512>

https://www.improbable.com/airchives/paperair/volume16/v16i4/about%20Chernoff%20faces%20AIR%2016-4.pdf

слайд про визуализацию

**Методы визуализации**

Методы визуализации, в зависимости от количества используемых измерений, принято классифицировать на две группы [22]:

* представление данных в одном, двух и трех измерениях;
* представление данных в четырех и более измерениях.

**Представление данных в одном, двух и трех измерениях**

К этой группе методов относятся хорошо известные способы отображения информации, которые доступны для восприятия человеческим воображением. Практически любой современный инструмент *Data Mining* включает способы визуального представления из этой группы.

В соответствии с количеством измерений представления это могут быть следующие способы:

* одномерное (univariate) измерение, или ***1-D*** ;
* двумерное (bivariate) измерение, или ***2-D*** ;
* трехмерное или проекционное (projection) измерение, или ***3-D***.

Следует заметить, что наиболее естественно человеческий глаз воспринимает двухмерные представления информации.

При использовании двух- и трехмерного представления информации *пользователь* имеет возможность увидеть закономерности набора данных:

* его кластерную структуру и распределение объектов на классы (например, на диаграмме рассеивания);
* топологические особенности;
* наличие трендов;
* информацию о взаимном расположении данных;
* существование других зависимостей, присущих исследуемому набору данных.

Если набор данных имеет более трех измерений, то возможны такие варианты:

* использование многомерных методов представления информации (они рассмотрены ниже);
* снижение размерности до одно-, двух- или трехмерного представления. Существуют различные способы снижения размерности, один из них - факторный анализ - был рассмотрен в одной из предыдущих лекций. Для снижения размерности и одновременного визуального представления информации на двумерной карте используются *самоорганизующиеся карты* Кохонена.

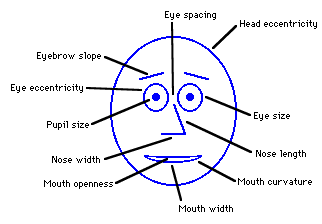
**Представление данных в 4 + измерениях**

Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации.

Наиболее известные способы многомерного представления информации:

* *параллельные координаты* ;
* " *лица Чернова* ";
* лепестковые диаграммы.

Основная идея представления информации в " ***лицах Чернова*** " состоит в кодировании значений различных переменных в характеристиках или чертах человеческого лица. Пример такого "лица" приведен на рис.



Для каждого наблюдения рисуется отдельное "лицо". На каждом "лице" относительные значения переменных представлены как формы и размеры отдельных черт лица (например, *длина* и ширина носа, размер глаз, размер зрачка, угол между бровями).

*Анализ* информации при помощи такого способа отображения основан на способности человека интуитивно находить сходства и различия в чертах лица.

В 1981 году Бернард Флури и Ганс Ридвил (Bernhard Flury and Hans Riedwyl) улучшили концепцию и добавили лицам Чернова асимметрию. Таким образом количество переменных увеличилось вдвое - до 36.

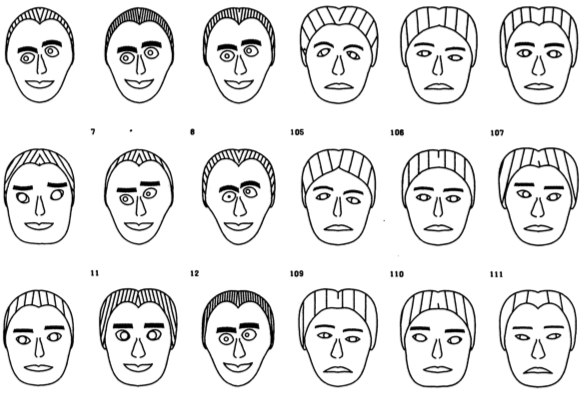
Итак, каждое лицо - это массив из 18 элементов, каждый из которых принимает значение от 0 до 1. Значению соответствует внешний вид соответствующей части лица. Параметры исследуемых объектов приводятся к этим значениям. Экстремумы реальных данных будут приняты как 0 и 1. Все остальное - лежащим в этом промежутке. По полученному массиву конструируется лицо.

Вот какие параметры задаются у лица:

1. Размер глаза
2. Размер зрачка
3. Позиция зрачка
4. Наклон глаза
5. Горизонтальная позиция глаза
6. Вертикальная позиция глаза
7. Изгиб брови
8. Плотность брови
9. Горизонтальная позиция брови
10. Вертикальная позиция брови
11. Верхняя граница волос
12. Нижняя граница волос
13. Обвод лица
14. Темнота волос
15. Наклон штриховки волос
16. Нос
17. Размер рта
18. Изгиб рта

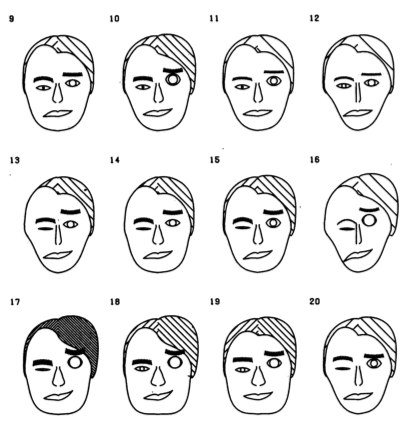
Сложность заключается в правильном сопоставлении исследуемых переменных с частями лица. При ошибке важные закономерности могут остаться незамеченными.

Флури приводит пример удачного многофакторного анализа с помощью лиц. Он проанализировал 100 реальных и 100 поддельных банкнот по параметрам размера границ, отступов и диагоналей. Вот что получилось:



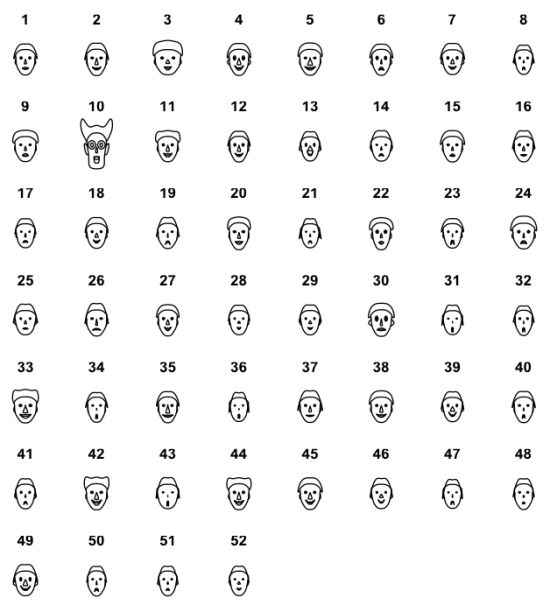
Поддельные банкноты четко выделились в отдельную группу. Таким образом анализ позволил выявить различающиеся группы объектов.

Асимметрия позволяет рассматривать объекты в прогрессе. Второй пример показывает различные факторы у пациентов, к которым применялось лечение. Левая сторона лица показывает значения параметров до, а правая - после лечения.



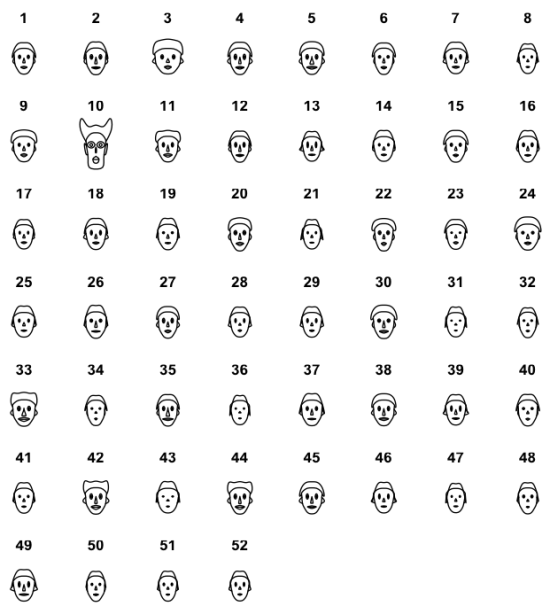
Посмотрите как изменилось состояние параметров. Легко можно понять кому и насколько стало лучше, даже не вникая в сущность исследуемых параметров.

Для начала следует установить пакет  [aplpack](http://cran.r-project.org/web/packages/aplpack/index.html)(Another Plot Package). После чего нужно загрузить данные (будем использовать выборку «уровень преступности по штатам»). В выборке 8 столбцов. Первый столбец - это имя штата, за исключением строки для среднего значения для США и округа Колумбия. Остальные столбцы - семь категорий преступлений. После чего просто вызываем функцию **faces(crime[,2:8])**с нужными параметрами. Получим следующее:

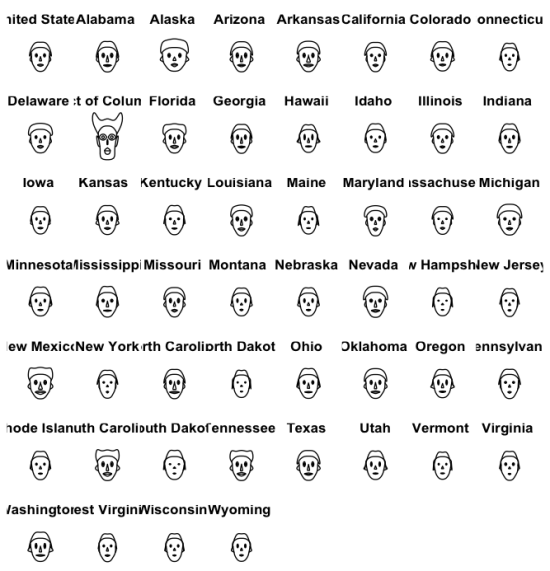


Можно увидеть, что некоторые лица улыбаются. Для более нейтральных наборов данных, таких как качество жизни или бейсбольная статистика, это имело бы смысл. Чем выше значение, тем лучше. Однако для нашей выборки, чем выше значение, тем хуже. Улыбки за уровень краж и воровства не кажутся совершенно правильными.

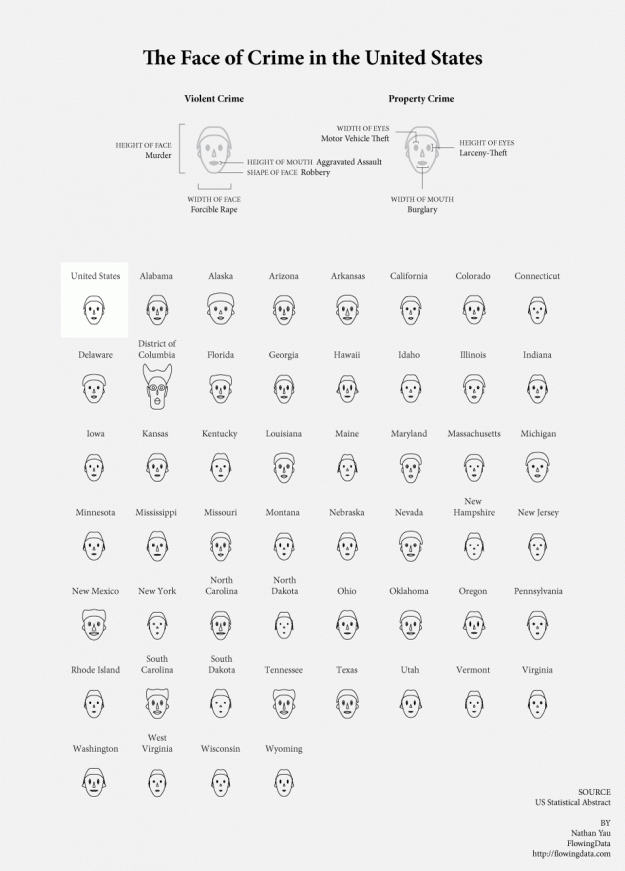
Функция faces () не позволяет нам выбирать, какие элементы ассоциировать с каждым параметром, поэтому нам нужно найти другое решение. Согласно документации кривая улыбки применяется к шестому столбцу входной матрицы. Заполним шестой столбец в наших данных одинаковым значением. Таким образом, все кривые улыбки будут нейтральными. Получим лица без улыбок



Добавим метки классов



Теперь можно ассоциировать каждое лицо со штатом.



жестокое преступление:

* Высота лица (убийство)
* высота рта (нападение при отягчающих обстоятельствах)
* форма лица (грабеж)

имущественное преступление

* ширина глаз (кража автомобилей)
* высота глаз (воровство)
* ширина рта (кража со взломом)

Второй пример:

Проанализируем данные по игре в крикет, и попытаемся сравнить 5 подающих игроков. Можно посмотреть на несколько показателей, таких как скорость удара, средний уровень подачи и т. д. Но глядя на эти показатели в таблице, сделать вывод сложно. Можно создать лицо Чернова и сравнить множество показателей среди этих игроков.

Игроки сравнивались по следующим показателям:

1. Средний уровень подачи

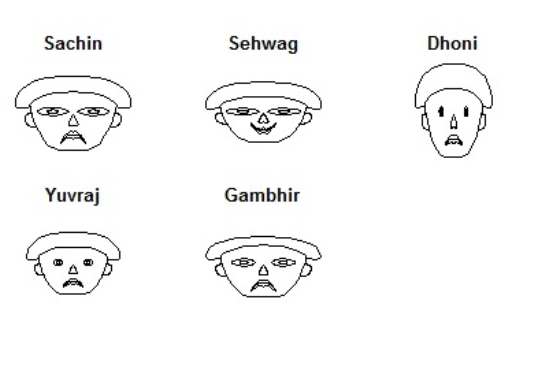
2. Скорость удара

3. Количество четырех очков в матче

4. Количество очков, равных шести, за матч

5. Соотношение подач мяча к общему количеству сыгранных матчей

Чтобы создать лица, эти метрики были сопоставлены с некоторыми чертами лица: 1. Высота лица, 2. Изгиб улыбки, 3. Ширина глаз 4. Высота глаз 5. Ширина лица



Как видно, самое счастливое лицо у Sehwag, так как он имеет наибольшую скорость удара, переменную, отображаемую изгибом улыбки. Кроме того, обратите внимание, что у Dhoni очень вытянутое лицо, это опять-таки связано с тем, что среднее значение удара отображается высотой лица, а Dhoni имеет очень хорошее среднее значение.

Еще одно замечание - это ширина глаз как для Dhoni, так и для Yuvraj, очень маленькая, это говорит о том, что оба игрока**, хотя и очень хорошие мастера удара, сделали много пробегов, пробегая между калитками.**