



Введение в анализ данных

Лекция 1.1

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

План дисциплины

■ 1. Введение, основные понятия анализа данных

- Введение в анализ данных. Анализ данных в различных прикладных областях. Основные определения. Этапы анализа данных. Примеры прикладных задач и их типы: классификация, регрессия, ранжирование, кластеризация, поиск структуры в данных.

■ 2. Применение встроенных функций Excel

- Различные типы ссылок. Связывание листов и рабочих книг. Применение различных типов встроенных функций. Математические функции. Статистические функции.
- Функции ссылок и подстановки. Логические функции. Текстовые функции. Функции для работы с датами.

План дисциплины

- **3. Форматы. Условное форматирование. Работа с большими табличными массивами**
 - Форматы. Создание пользовательских форматов. Числовые форматы. Форматы даты и времени. Группы пользовательских форматов. Редактирование, применение и удаление форматов.
 - Правила выделения ячеек. Гистограммы. Цветовые шкалы. Создание правила форматирования. Сортировка данных. Сортировка по одному критерию. Многоуровневая сортировка. Сортировка по форматированию. Спарклайны
 - Фильтрация данных. Срезы. Расширенный фильтр. Подведение промежуточных итогов. Консолидация данных
 - Функции работы с данными. Особенности совместной работы.
- **4. Сводные таблицы.**
 - Создание сводных таблиц. Преобразование сводных таблиц. Фильтрация данных: фильтры, срезы, временная шкала. Настройка полей сводной таблицы.
 - Добавление вычисляемых полей в сводную таблицу. Группировка полей в сводных таблицах. Сводные диаграммы. Обновление сводных таблиц и диаграмм

План дисциплины

■ 5. Формулы массивов

- Адресация. Функции. Формулы массивов.
- Решение задач по извлечению данных из массива данных.
- Поиск по нескольким критериям. Использование именованных диапазонов в расчетах. Обработка данных с одного или нескольких листов.

■ 6. Визуализация данных.

- Диаграммы. Комбинированные диаграммы. Гистограмма с отображением итогов.
- Проектная диаграмма Ганта. Диаграмма сравнений Торнадо. Каскадная диаграмма (диаграмма отклонений Водопад).
- Иерархические диаграммы. Статистические диаграммы. Диаграммы с пользовательскими элементами управления.

План дисциплины

- **7. Прогнозирование данных. Вариативный анализ "Что Если" и Оптимизация.**
 - Выделение тренда: скользящее среднее, функции регрессионного анализа: ПРЕДСКАЗ, ТЕНДЕНЦИЯ, РОСТ. Построение линий тренда.
 - Использование инструмента Таблица данных для анализа развития ситуации при 2-х переменных. Оценка развития ситуации и выбор оптимальной стратегии с помощью Сценариев.
 - Решение однокритериальной задачи оптимизации с помощью Подбора параметра.
 - Решение многокритериальных задач оптимизации с использованием надстройки Поиск решения
- **8. Обработка внешних баз данных**
 - Импорт внешних данных: Web, Access, Text.
 - Запросы (Microsoft Query) к внешним базам данных: Access, Excel.

Организация работы

- Лекция
- Практическая работа в компьютерном классе
- Домашнее задание

Система оценок

	количество	баллы	всего
занятие	20	4	80
аттестация	1	20	20
Итого			100

Оценка	Незачет		Зачет				
Оценка ECTS	F (2)	FX (2+)	E (3)	D (3+)	C (4)	B (5)	A (5+)
Максимальная сумма баллов	0–35	36–50	51–60	61–68	69–83	84–92	93–100

Введение, основные понятия анализа данных

Введение в анализ данных. Анализ данных в различных прикладных областях. Основные определения. Этапы анализа данных. Примеры прикладных задач и их типы: классификация, регрессия, ранжирование, кластеризация, поиск структуры в данных.

Источники данных

- **Наблюдение**

- Воздействие на объект исследования минимально

- **Эксперимент**

- На объект исследования оказывается заранее рассчитанное воздействие

- **Генеральная совокупность и выборка**

- Простой отбор – случайное извлечение объектов из генеральной совокупности с возвратом или без возврата.
- Типический отбор, когда объекты отбираются не из всей генеральной совокупности, а из ее «типической» части.
- Серийный отбор – объекты отбираются из генеральной совокупности не по одному, а сериями.
- Механический отбор - генеральная совокупность «механически» делится на столько частей, сколько объектов должно войти в выборку и из каждой части выбирается один объект.

- **Исследования**

- Сплошные
- Выборочные

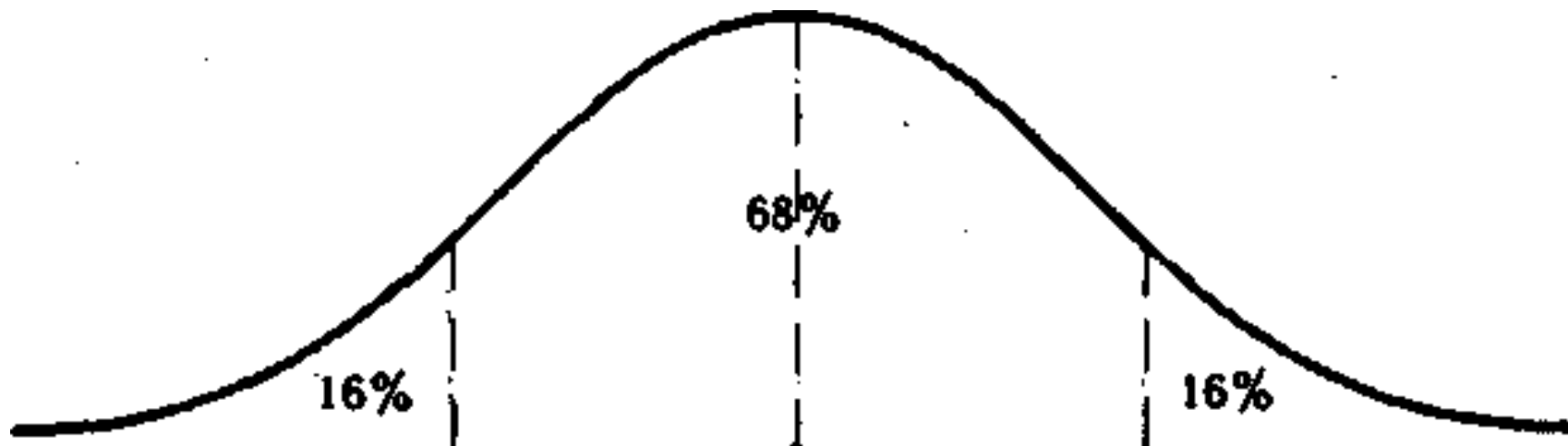
Пример сплошной выборки – перепись 1897г

Возраст	Мужчины	Женщины	Оба пола		Возраст	Мужчины	Женщины	Оба пола
20	1 203 695	1 601 899	2 805 594		40	1 256 887	1 589 111	2 845 998
21	1 027 337	771 514	1 798 851		41	378 000	305 971	683 971
22	1 158 645	1 099 310	2 257 955		42	621 999	542 531	1 164 530
23	1 053 373	998 226	2 051 599		43	512 332	451 597	963 929
24	929 468	885 048	1 814 516		44	424 793	393 336	818 129
25	1 188 690	1 480 517	2 669 207		45	924 863	979 753	1 904 616
26	958 475	907 086	1 865 561		46	474 362	409 166	883 528
27	985 563	922 210	1 907 773		47	438 236	392 561	830 797
28	1 014 745	1 019 581	2 034 326		48	542 175	503 505	1 045 680
29	625 075	530 513	1 155 588		49	299 949	265 337	565 286
30	1 494 510	1 889 008	3 383 518		50	1 077 316	1 386 835	2 464 151
31	508 760	421 847	930 607		51	230 845	205 362	436 207
32	780 758	730 968	1 511 726		52	394 749	357 427	752 176
33	721 160	643 810	1 364 970		53	322 410	301 518	623 928
34	549 678	536 352	1 086 030		54	268 326	269 386	537 712
35	1 109 835	1 223 840	2 333 675		55	680 823	700 127	1 380 950
36	724 769	675 392	1 400 161		56	370 174	318 765	688 939
37	724 478	649 983	1 374 461		57	285 355	245 565	530 920
38	796 176	739 313	1 535 489		58	310 510	278 453	588 963
39	483 817	401 594	885 411		59	170 292	147 219	317 511
					60	899 491	1 142 709	2 042 200

Примеры данных с неопределенной достоверностью

- Оценки в школьном аттестате
- Результаты ЕГЭ
- Результаты работы в группе (коллективе)

Оценка знаний путем тестирования



Хороший нормативно-ориентированный тест обеспечивает нормальное распределение индивидуальных баллов репрезентативной выборки учеников, когда среднее значение баллов находится в центре распределения, а остальные значения концентрируются вокруг среднего по нормальному закону, т.е. примерно 70% значений в центре, а остальные сходят на нет к краям распределения

Как получать данные

Два принципа составления выборки

■ Принцип повторностей

- предполагает, что один и тот же эффект будет исследован несколько раз
- повторности должны быть независимы друг от друга

Все объекты отличаются.

Если брать объекты поодиночке, то можно учитывать незначительные характеристики, не влияющие на сущность исследования

Исследование нескольких объектов позволят выявить характеристики, влияние которых на объект незначительное

■ Рандомизация

- каждый объект должен иметь разные шансы попасть в выборку

Что дает анализ данных

- **Общие характеристики для больших выборок**
 - центральная тенденция
 - разброс, насколько сильно разбросаны данные
- **Сравнение между разными выборками**
 - насколько велика вероятность, что различия вызваны случайными причинами
- **Сведения о взаимосвязях**
 - соответствия
 - корреляции
 - зависимости
 - предсказания
- **Структура**
 - классификация объектов

Основные этапы анализа данных

- **4 основных этапа анализа:**
 - Описание совокупности данных
 - Уплотнение исходной информации.
 - Углубление интерпретации и переход к объяснению
 - Прогноз развития явлений.

Описание совокупности данных

■ Чистка массива.

- выявление ошибок и пропусков, допущенных в ходе сбора и ввода информации.
- Здесь задача - поиск "выбросов" (неправильно забитых ответов респондента) и логических нарушений в ходе интервью (например, не сделанный переход).
- коррекция выборки.
- Наиболее распространенным методом коррекции выборки является перевзвешивание. При его использовании, ответы более представленной категории респондентов учитываются с определенным коэффициентом (например, 0,9).

■ Описание

- описание распределения данных по существенным с точки зрения целей и задачи признакам.

Уплотнение исходной информации

- **Цель данного этапа – сокращение числа признаков, необходимых для анализа**
 - укрупнение шкал (например, группировка возраста)
 - расчет индексов и агрегированных показателей.

Углубление интерпретации

- и переход к объяснению путём выявления возможных прямых и косвенных влияний по полученным агрегированным показателям.
- Цель данного этапа – поиск статистических закономерностей в распределении данных.
- Здесь же проверяются основные гипотезы, строятся выводы. На данном этапе основной применяемый метод – корреляционный анализ.

Прогноз развития явлений

- и процессов при определенных условиях.
- Этот этап имеет место лишь в аналитических исследованиях. Происходит построение содержательных представлений об основе процесса.
- Методы – регрессионный анализ

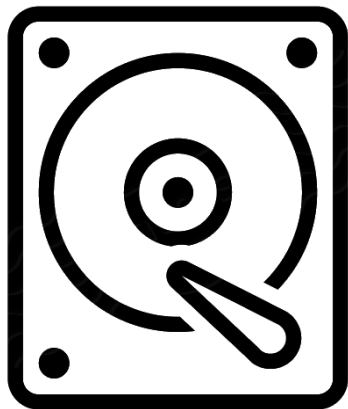
BigData

- Большие данные - обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами
- В широком смысле о «больших данных» говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных, в некоторых проблемных областях — весь мировой объём данных, и вытекающих из этого трансформационных последствий.
- С точки зрения информационных технологий в совокупность подходов и инструментов изначально включались средства массово-параллельной обработки неопределённо структурированных данных, прежде всего, системами управления базами данных категории NoSQL
- NoSQL → no relational → No SQL → **Not Only SQL**

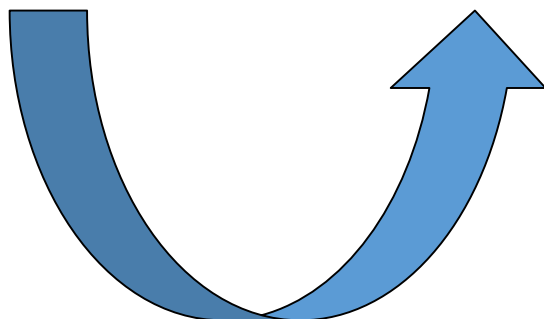
BigData

- **Совокупность данных, методов их обработки, технологий хранения и доступа, применения**
 - большие объемы метаданных
 - количество записей остается, меняется объем самой записи
 - географически распределенные данные
 - системно распределенные данные
 - специализированные аппаратные комплексы
 - дисковые подсистемы выделяются в отдельный класс технологических решений
 - узким местом становятся системы коммутации и передачи данных
 - вычислительные узлы строятся на новых принципах
 - развитие и распространение языков программирования
 - S, R, Python
- **IP4 → IP6**
- **3G → 4G → 5G**

Коллизия развития дисковых подсистем для обработки данных

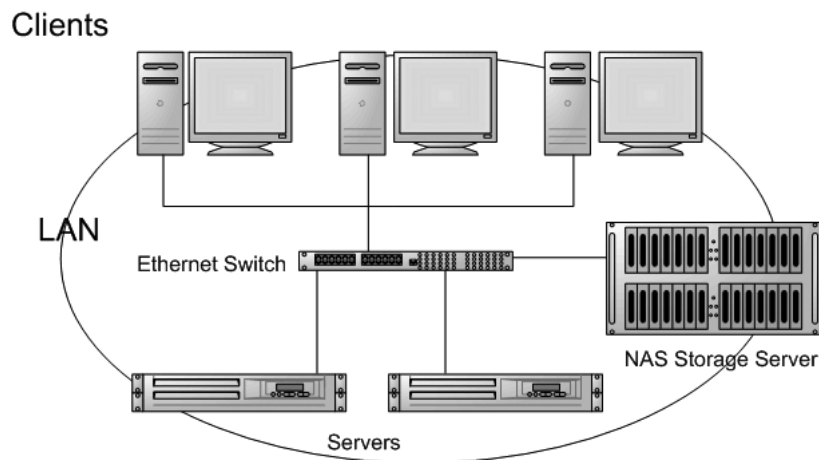


HDD
RAID

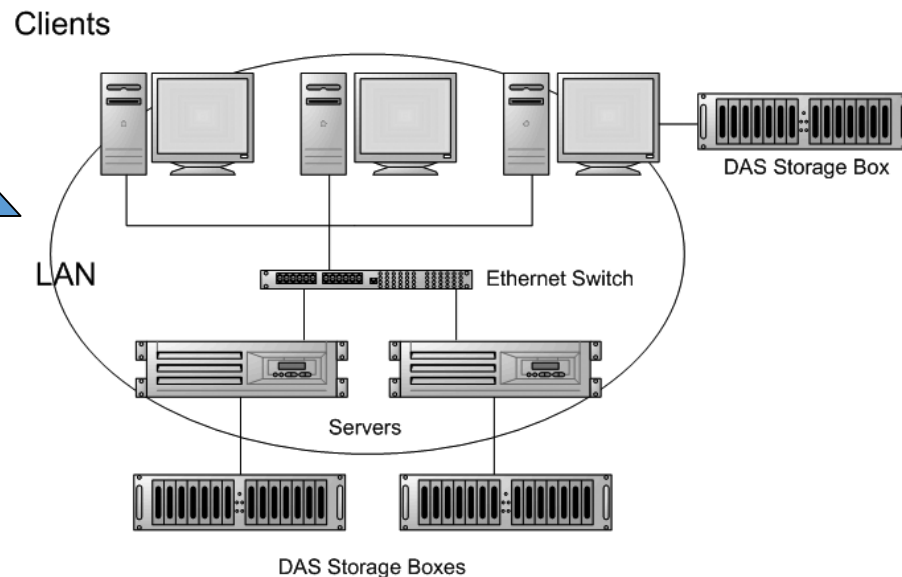


NAS, SAN

Network Attached Storage



Direct Attached Storage



DAS

система хранения
данных с прямым
подключением

Техники и методы анализа, применимые к BigData

- Data Mining;
- Краудсорсинг;
- Смешение и интеграция данных;
- Машинное обучение;
- Искусственные нейронные сети;
- Распознавание образов;
- Прогнозная аналитика;
- Имитационное моделирование;
- Пространственный анализ;
- Статистический анализ;
- Визуализация аналитических данных.

Горизонтальная масштабируемость, которая обеспечивает обработку данных - базовый принцип обработки больших данных.

Данные распределены на вычислительные узлы, а обработка происходит без деградации производительности.

Применимость реляционных систем управления и Business Intelligence.

Программные средства обработки данных

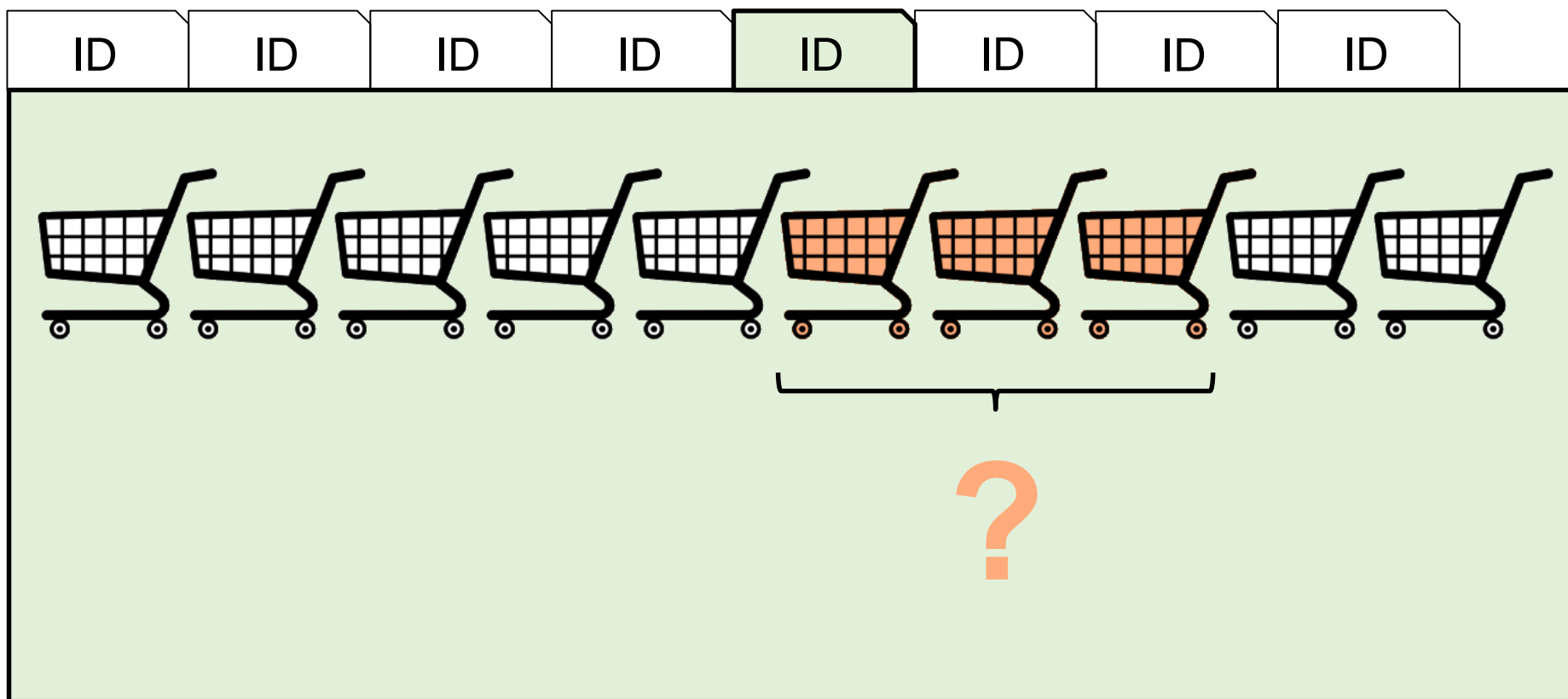
- **Калькулятор**
- **Электронные таблицы**
- **СУБД**
 - MS Access
 - MS SQL Server
- **Специализированные статистические программы**
 - оконно-кнопочные системы
 - STATISTICA
 - STADIA
 - IBM SPSS Statistics Base
 - статистические среды
 - SAS
 - R

Беременная девочка и магазин Target

Discount Card



Беременная девочка и магазин Target



Машинное обучение

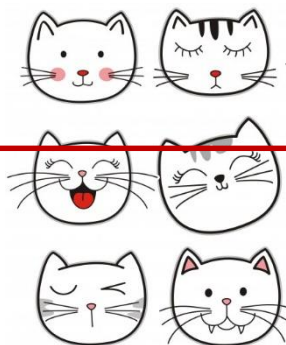
Машинное обучение (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Имеется множество объектов (ситуаций) и множество возможных ответов (откликов, реакций). Существует некоторая зависимость между ответами и объектами, но она неизвестна. Известна только конечная совокупность прецедентов — пар «объект, ответ», называемая обучающей выборкой. На основе этих данных требуется восстановить неявную зависимость, то есть построить алгоритм, способный для любого возможного входного объекта выдать достаточно точный классифицирующий ответ.

Важной особенностью при этом является способность обучаемой системы к обобщению, то есть к адекватному отклику на данные, выходящие за пределы имеющейся обучающей выборки. Для измерения точности ответов вводится оценочный функционал качества.

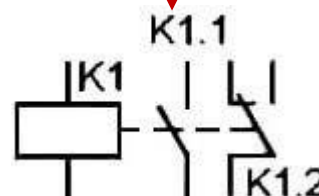
Лужайка и нехорошие коты

IP-камера
Foscam FI9800P



Видеокарта Nvidia Jetson TX1
+
Нейронная сеть глубокого
обучения Caffe

Particle Photon



Искусственная нейронная сеть

Искусственная нейронная сеть (ИНС) — математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма.

После разработки алгоритмов обучения получаемые модели стали использовать в практических целях: в задачах прогнозирования, для распознавания образов, в задачах управления и др.

ИНС представляет собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты (особенно в сравнении с процессорами, используемыми в персональных компьютерах). Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И, тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие по отдельности простые процессоры вместе способны выполнять довольно сложные задачи.

Искусственная нейронная сеть

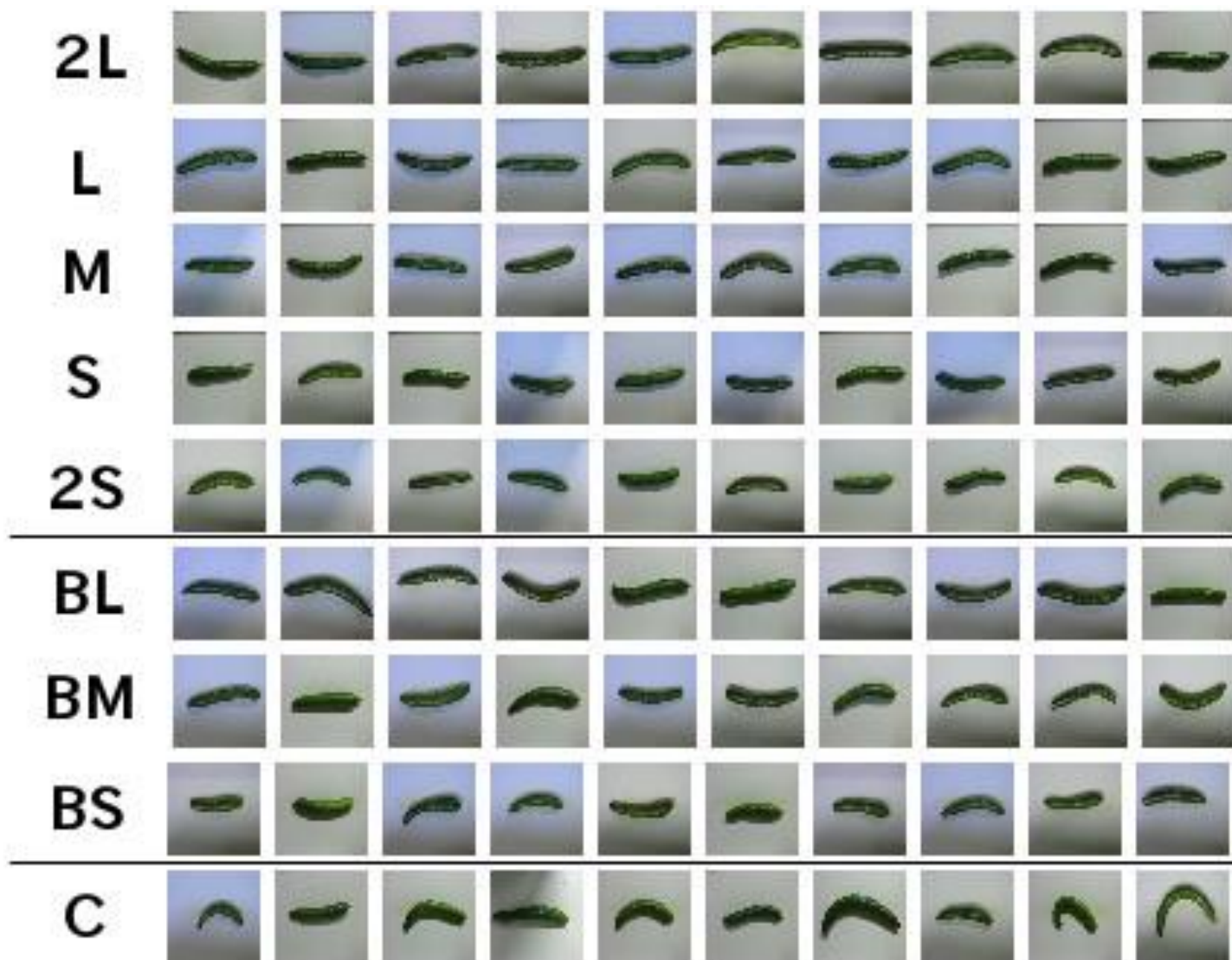
Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения — одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных и/или «зашумленных», частично искажённых данных.

Распознавание образов В качестве образов могут выступать различные по своей природе объекты: символы текста, изображения, образцы звуков и т. д. При обучении сети предлагаются различные образцы образов с указанием того, к какому классу они относятся. Образец, как правило, представляется как вектор значений признаков. При этом совокупность всех признаков должна однозначно определять класс, к которому относится образец. В случае, если признаков недостаточно, сеть может соотнести один и тот же образец с несколькими классами, что неверно. По окончании обучения сети ей можно предъявлять неизвестные ранее образы и получать ответ о принадлежности к определённому классу.

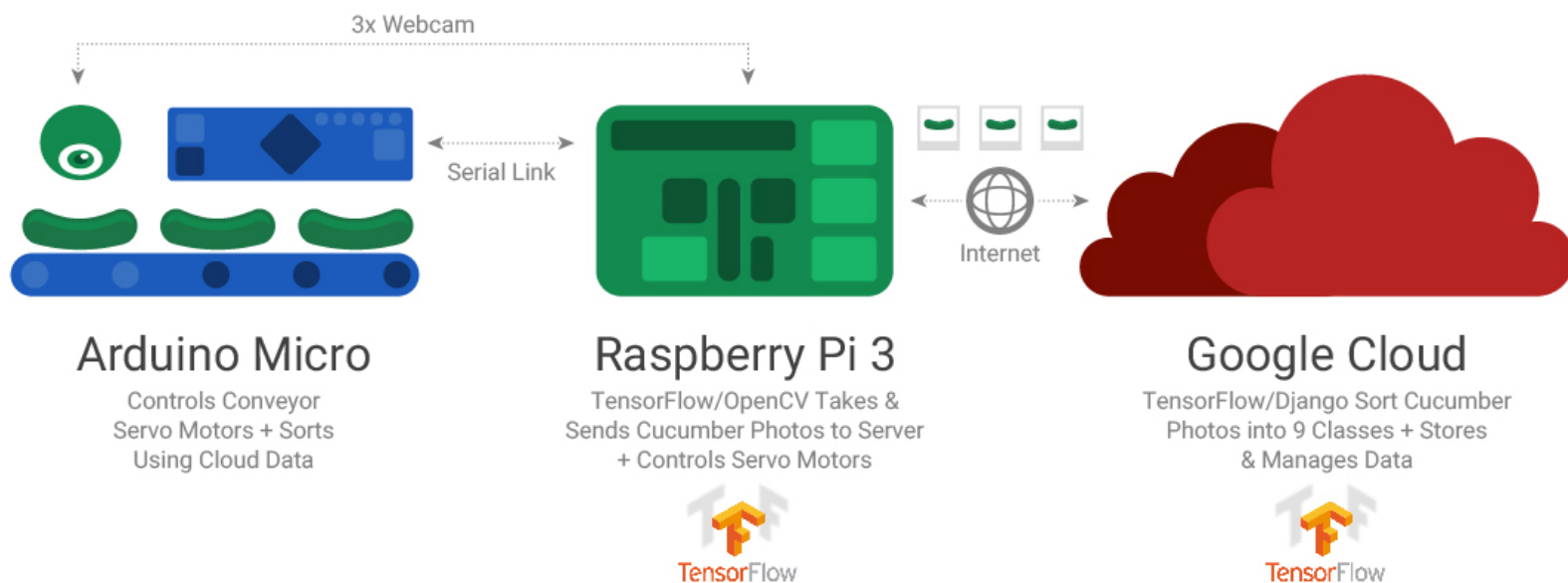
Лужайка и нехорошие коты



Сортировка огурцов



Сортировка огурцов



Одномерные типы анализа данных

Дескриптивные (или описательные) статистики являются базовым и наиболее общим методом анализа данных. Задача - провести опрос с целью составления портрета потребителя товара. Респонденты указывают свой пол, возраст, семейное и профессиональное положение, потребительские предпочтения и т.д., а описательные статистики позволяют получить информацию, на основе которой будет строиться весь портрет.

Потенциальный спрос на товар

Стоимость товара, руб.	Абсолютная частота, чел.	Относительная частота, %	Кумулятивная частота, %
5000	23	19,2%	19,2%
4500	41	34,2%	53,4%
4399	56	46,6%	100%

Абсолютная частота показывает, сколько раз тот или иной ответ повторяется в выборке. Например, 23 человека купили бы предложенный товар стоимостью 5000руб., 41 человек – стоимостью 4500руб. и 56 человек – 4399руб. Относительная частота показывает, какую долю данное значение составляет от всего объема выборки (23 человека – 19,2%, 41 – 34,2%, 56 – 46,6%). Кумулятивная или накопленная частота показывает долю элементов выборки, не превышающих определенное значение. Например, изменение процента респондентов, готовых приобрести тот или иной товар при уменьшении цены на него (19,2% респондентов готовы купить товар за 5000руб., 53,4% — от 4500 до 5000руб., и 100% — от 4399 до 5000руб.).

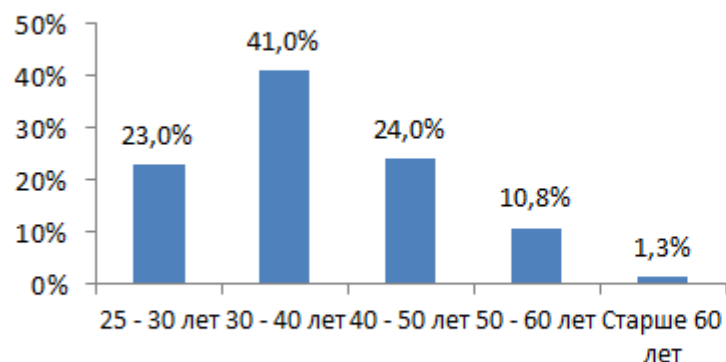
Дескриптивный анализ

Наряду с частотами, дескриптивный анализ предполагает расчет различных описательных статистик. Соответствуя своему названию, они предоставляют основную информацию о полученных данных. Уточним, использование конкретной статистики зависит от того, в каких шкалах представлена исходная информация. Номинальная шкала используется для фиксации объектов, не имеющих ранжированного порядка (пол, место жительства, предпочитаемая марка и т.д.). Для подобного рода массива данных нельзя рассчитать каких-либо значимых статистических показателей, кроме моды — наиболее часто встречающегося значения переменной. Несколько лучше в плане анализа ситуация обстоит с порядковой шкалой. Здесь становится возможным, наряду с модой, расчет медианы – значения, разбивающего выборку на две равные части. Например, при наличии нескольких ценовых интервалов на товар (500-700 руб. руб., 700-900, 900-1100 руб.) медиана позволяет установить точную стоимость, дороже или дешевле которой потребители готовы приобретать или, наоборот, отказаться от покупки. Наиболее богатыми на все возможные статистики являются количественные шкалы, которые представляют собой ряды числовых значений, имеющих равные интервалы между собой и поддающихся измерению. Примерами подобных шкал могут служить уровень дохода, возраст, время, отводимое на покупки и т.д. В данном случае становятся доступными следующие информационные меры: среднее, размах, стандартное отклонение, стандартная ошибка среднего.

Дескриптивный анализ

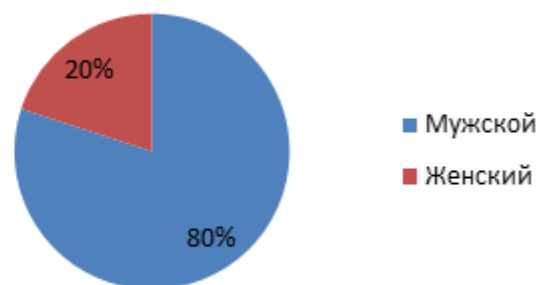
Конечно, язык цифр является довольно «сухим» и для многих весьма непонятным. По этой причине дескриптивный анализ дополняется визуализацией данных путем построения различных диаграмм и графиков, как, например: гистограммы, линейные, круговые или точечные диаграммы.

Возраст респондентов



Гистограмма

Пол респондентов



Круговая диаграмма

Таблицы сопряженности и корреляции

Таблицы сопряженности – это средство представления распределения двух переменных, предназначенное для исследования связи между ними. Таблицы сопряженности можно рассматривать как частный тип дескриптивного анализа. В них также является возможным представление информации в виде абсолютных и относительных частот, графическая визуализация в виде гистограмм или точечных диаграмм. Наиболее эффективно таблицы сопряженности проявляют себя в определении наличия взаимосвязи между номинальными переменными (например, между полом и фактом потребления какого-либо продукта). В общем виде таблица сопряженности выглядит так. Зависимость между полом и использованием страховыми услугами

Зависимость между полом и использованием страховыми услугами

Пол	Пользуетесь ли Вы услугами страхования жизни?	
	Да	Нет
Мужской	39%	54%
Женский	61%	46%
Итого по столбцу	100%	100%

Таблицы сопряженности и корреляции

На основе представленных в таблице данных и можно делать выводы о наличии/отсутствии взаимосвязи между исследуемыми переменными.

Для более точного выявления наличия связи между переменными используют разные статистические критерии.

Наиболее часто применяются такие, как: критерий Хи-квадрат (χ^2); коэффициент сопряженности; критерий лямбда; коэффициент R Спирмена; критерий корреляции Пирсона и др.

Правильный выбор критерия является решающим шагом для получения корректных результатов.

Понятие «корреляции»

Корреляция предназначена для выражения силы взаимосвязи по безразмерной шкале от -1 до + 1.

Положительная корреляция означает сильную положительную взаимосвязь, т.е. увеличение одной переменной вызывает увеличение другой переменной. Например, такая корреляция наблюдается между ростом и весом человека.

Отрицательная корреляция означает сильную отрицательную взаимосвязь, т.е. увеличение одной переменной вызывает уменьшение другой переменной. Например, увеличение цены товара может сопровождаться уменьшением объема продаж.

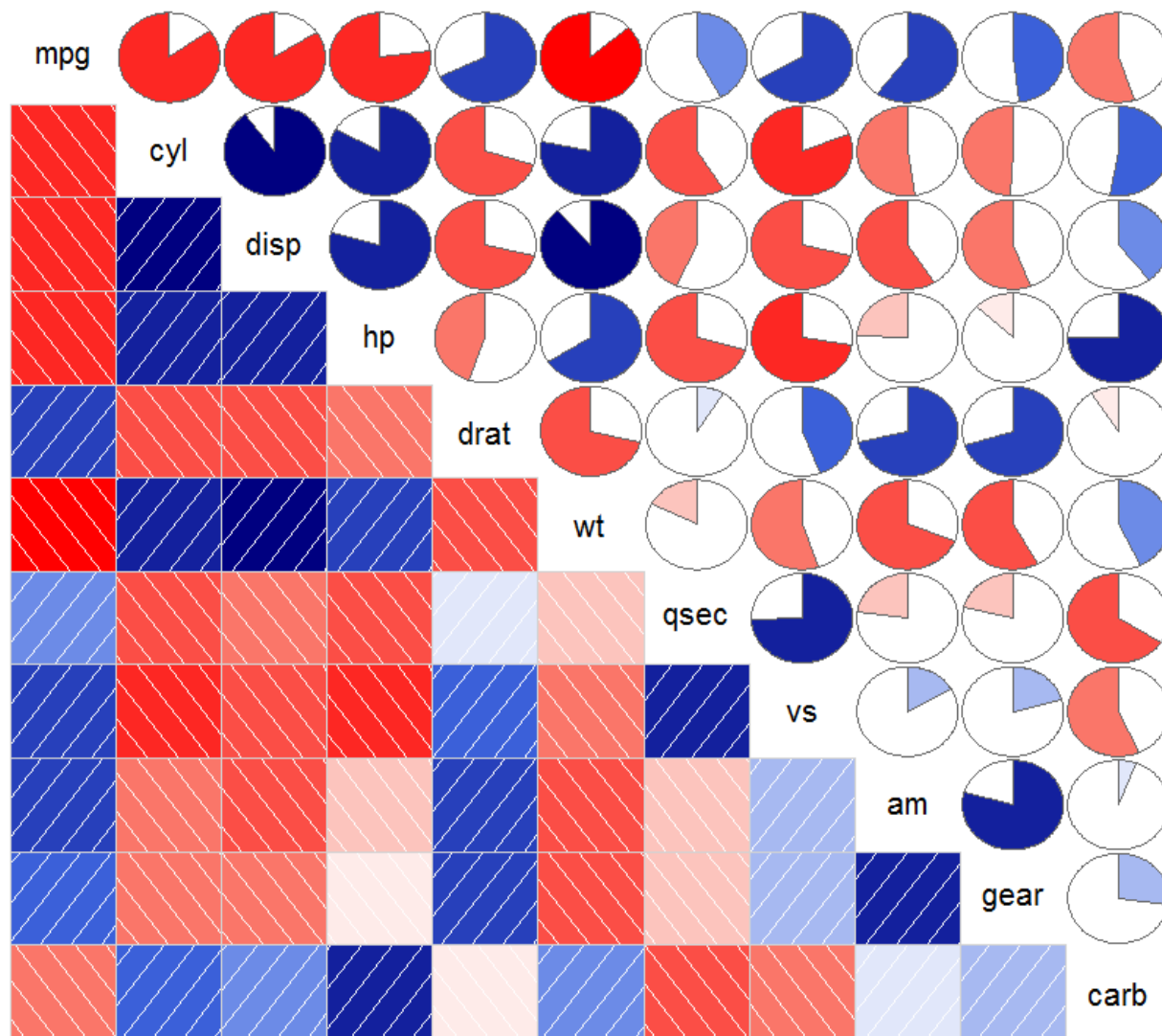
Близкая к нулю корреляция означает, что между двумя переменными нет никакой взаимосвязи. Кроме того, между переменными может существовать нелинейная взаимосвязь, которая характеризуется нулевой корреляцией.

Расчет коэффициентов корреляции

```
> t1 <- mtcars[,c(1,2,9)]
> cor(t1,method="pearson")
           mpg          cyl          am
mpg  1.0000000 -0.852162  0.5998324
cyl -0.8521620  1.000000 -0.5226070
am   0.5998324 -0.522607  1.0000000
> cor(t1,method="spearman")
           mpg          cyl          am
mpg  1.0000000 -0.9108013  0.5620057
cyl -0.9108013  1.0000000 -0.5220712
am   0.5620057 -0.5220712  1.0000000
> cor(t1,method="kendall")
           mpg          cyl          am
mpg  1.0000000 -0.7953134  0.4690128
cyl -0.7953134  1.0000000 -0.4946212
am   0.4690128 -0.4946212  1.0000000
```


Визуализация корреляции

mpg	расход топлива (количества миль на галлон топлива)
cyl	кол-во цилиндров
disp	объем двигателя
hp	мощность двигателя (лошадиные силы)
drat	передаточное число заднего моста
wt	вес
qsec	значение времени разгона
vs	тип двигателя (v- образный, рядный)
am	тип коробки передач
gear	кол-во передач
carb	число карбюраторов



Многомерные типы анализа данных

Многомерный анализ данных позволяет одновременно исследовать взаимоотношения двух и более переменных и проверять гипотезы о причинных связях между ними.

Техники многомерного анализа разнообразны.

Самые распространенные:

- Факторный анализ
- Кластерный анализ

Факторный анализ

Суть факторного анализа, состоит в том, чтобы имея большое число параметров, выделить малое число макропараметров, которыми и будут определяться различия между измеряемыми параметрами.

Это позволит оптимизировать структуру анализируемых данных.

Применение факторного анализа преследует две цели:

- сокращение числа переменных;
- классификация данных.

Факторный анализ. Пример по исследованию имиджа компании

Клиенту предлагается оценить данную компанию по целому ряду критериев, общее число которых может превышать несколько десятков.

Применение факторного анализа в данном случае позволяет снизить общее количество переменных путем распределения их в обобщенные пучки факторов, например,

- «материальные условия компании»,
- «взаимодействие с персоналом»,
- «удобство обслуживания».

Факторный анализ. Пример в составление социально-психологических портретов потребителей

Респонденту необходимо выразить степень своего согласия/несогласия с перечнем высказываний о стиле жизни.

В итоге, можно выделить, например, целевые группы потребителей:

- «новаторы»;
- «прогрессисты»;
- «консерваторы».

Факторный анализ. Пример в сфере банковского дела

Актуальным примером исследования в сфере банковского дела, может послужить, изучение уровня доверия клиента к банку, которое можно описать следующими факторами:

- надежность сделок (включающий такие параметры, как сохранность средств, возможность беспрепятственного их перевода);
- обслуживание клиентов (профессионализм сотрудников, их благожелательность) и
- качество обслуживания (точность выполнение операций, отсутствие ошибок) и др.

Кластерный анализ

Кластерный анализ (от англ. cluster – сгусток, пучок, гроздь) – это один из способов классификации объектов.

Он позволяет рассматривать достаточно большой объем информации, сжимая его и делая компактными и наглядными.

Термин «кластерный анализ» был введен в 1939 году английским ученым Р. Трионом, предложившим соответствующий метод, который сводился к поиску групп с тесно коррелирующим признаком в каждой из них.

Целью кластерного анализа является выделение сравнительно небольшого числа групп объектов, как можно более схожих между собой внутри группы, и как можно более отличающихся в разных группах. В настоящее время разработано достаточно большое число алгоритмов кластерного анализа.

Кластерный анализ

Планируется провести опрос потребителей, соответственно, для разных потребителей необходимы различные стратегии для их привлечения. Для решения данной задачи мы предлагается сегментировать клиентов, прибегнув к методу кластеризации. Для этого выполняются следующие шаги:

- формируется выборка и проводится опрос клиентов,
- определяются переменные (характеристики), по которым будут оцениваться респонденты в выборке,
- вычисляются значения меры сходства и различия между ответами респондентов,
- выбираются метод кластеризации (т.е. правила объединения респондентов в группы),
- определяется оптимальное число кластеров (групп).

Кластерный анализ

	Уровень дохода			Возраст			Ежемесячные затраты на покупку одежды		
	15-25 тыс.	25-35 тыс.	35-45 тыс.	25-30 лет	30-35 лет	35-40 лет	До 10 тыс.	До 20 тыс.	До 40 тыс.
Кластер 1	72%	18%	10%	73%	20%	7%	75%	15%	10%%
Кластер 2	20%	64%	16%	17%	67%	16%	15%	73%	12%
Кластер 3	9%	14%	77%	9%	11%	80%	8%	10%	82%

Информация, представленная в таблице, позволяет составить портрет клиентов каждого кластера, которые впоследствии необходимо учитывать при составлении стратегии успешного продвижения продукта на рынке.

Кластерный анализ. Применение

- В социологии: разделение респондентов на различные социально-демографические группы.
- В маркетинге: сегментация рынка по группам потребителей, группировка конкурентов по факторам конкурентоспособности.
- В менеджменте: выделение групп сотрудников с разным уровнем мотивации, выявление мотивирующих/демотивирующих факторов в организации, классификация конкурентоспособных отраслей и поставщиков, и др.
- В медицине — классификация симптомов, признаков заболеваний, пациентов, препаратов для успешной терапии.
- А также психиатрии, биологии, экологии, информатике и т.д.

Социометрия / Социометрический анализ

Метод социометрии (от лат. socius – товарищ, компаньон, соучастник; metrim — измерение) как метод анализа данных был разработан в США психологом Дж. Морено в 1934 году.

Социометрия относится к социально-психологическим тестам и позволяет измерять межличностные отношения, внутригрупповые связи и иерархии в малых социальных группах, то есть реально существующих образованиях, где люди собраны вместе и объединены каким-либо общим признаком (например, рабочий коллектив, студенческая группа, и др.). Социометрия(sociometry) позволяет выявить лидеров и аутсайдеров, измерить авторитет формального и неформального лидеров, определить социально-психологический климат в группе, выявить наличие конфликта, а также ценностные ориентации.

Социометрия / Социометрический анализ

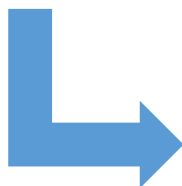
В основе процедуры рассматриваемого метода лежит так называемый социометрический опрос, который состоит из следующих этапов:

1. Подготовка: исследователь знакомится с различными характеристиками коллектива и определяет, какие данные он хочет получить.
2. Исследователь устанавливает содержание и количество социометрических критериев. Критерий выглядит как вопрос, который задается группе. Члены группы делают положительный либо отрицательный выбор в отношении других членов группы, таким образом, обнаруживая свои симпатии и антипатии. Критерий формулируется таким образом, чтобы побуждать человека отдавать предпочтение по какому-либо вопросу людям своей группы. Например: «С кем из твоей группы ты бы стал готовить совместный проект?», «Кого из коллег ты бы пригласил на день рождения в первую очередь?» и т.д.
3. Проведение опроса. Опрос лучше проводить в коллективах, которые имеют опыт совместной деятельности, в результате которой уже возникли определенные устойчивые взаимоотношения между его членами.
4. Социометрия и матрица. Обработка информации и интерпретация данных: на основании ответов участников заполняется социометрическая матрица – таблица, в которой представлено распределение индивидуальных выборов.

Социометрическая матрица → социограмма

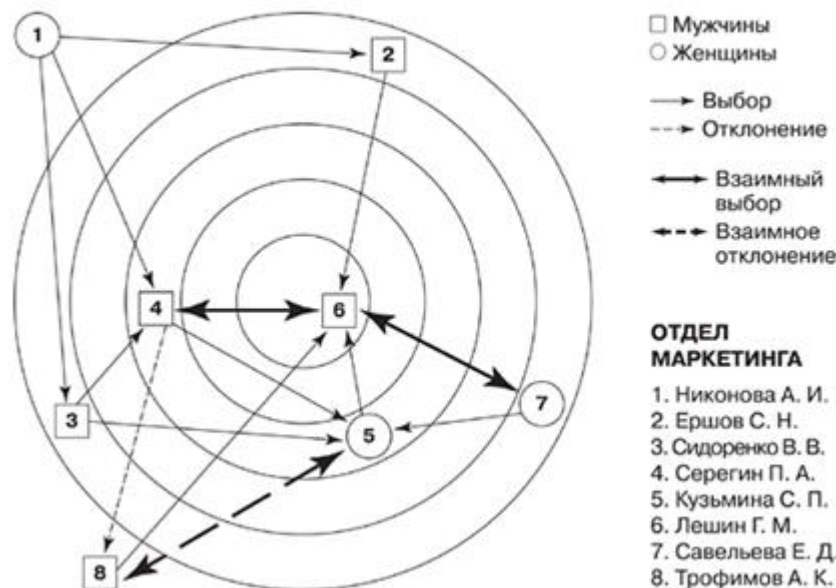
Социометрическая матрица

Фамилия выбирающего члена группы	Фамилия выбираемого члена группы	1. Егоров	2. Кузьмин	3. Федоров	4. Алешин	...	Характеристики (качества) по каждому выбору
1. Егоров			3		2		Энергичен, лентяй и зануда, нормальный парень
2. Кузьмин		1		2	3		Мой друг, любитель подставлять
3. Федоров		-3	-2				Консервативен
4. Алешин							Хороший человек, всегда помогает
...							
Кол-во полученных выборов		1	1	1	2		
Σ полученных выборов		1	3	2	5		
Кол-во полученных отклонений		1	1	-	-		
Σ полученных отклонений		3	2	-	-		



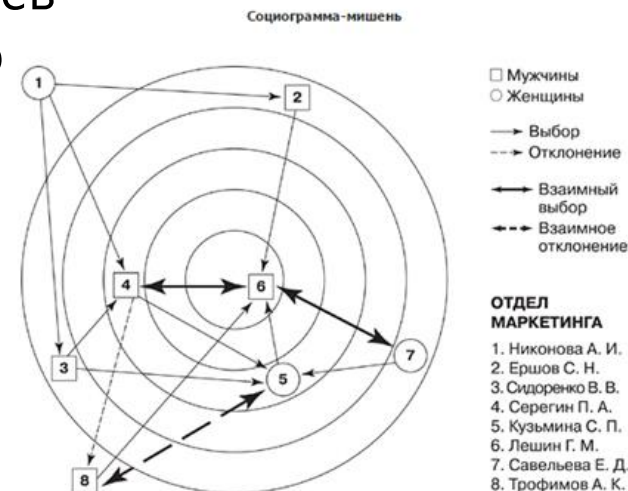
На основе заполненной социометрической матрицы строится социограмма

Социограмма-мишень





Социограмма

В центр помещается наиболее популярный член группы (или несколько членов), далее — менее популярные, по мере убывания, вплоть до изгоев (внешний круг). Стрелки обозначают взаимную либо одностороннюю симпатию/антипатию. Социграмма позволяет визуализировать результаты, наглядно увидеть картину сложившихся взаимоотношений в группе. На основании полученной информации можно строить «рейтинги популярности», определять позиции участников в структуре межличностных отношений, выделять подгруппы и т. п. На сегодняшний день социометрический анализ широко применяется в психологии (социальная психология, психотерапия), педагогике, социологии, а также менеджменте и управлении персоналом.



Категориальный анализ/ Метод категоризации

Категориальный анализ или метод категоризации применяется для анализа качественных данных, представленных текстом. Например, ответов на свободные вопросы, где респондент отвечает самостоятельно, формулируя ответ своими словами. В результате использования данного метода можно получить новый массив данных для последующей статистической обработки. Для качественного выполнения подобного рода анализа необходимо глубоко понимать и разбираться в исследуемой проблеме и ее понятийном аппарате, так как он основан на непосредственной работе с семантикой ответов, то есть значениями слов и словосочетаний. Важным замечанием является необходимость сохранения баланса. Нельзя сильно упрощать понятия, приводя их к наиболее общим категориям, потому что при этом теряется часть смысла, вложенного в ответ респондентов. Также и наоборот, нельзя составлять слишком узкие категории, содержащие 1-2 понятия, т.к. структура данных остается громоздкой, а ее анализ затруднительным. Поэтому рассматриваемый метод предполагает в большей степени творческий подход, а не механическую обработку.



Категориальный анализ/ Метод категоризации

Собраны ответы о причинах неудовлетворенности работой сотрудников какой-либо организации

Причины неудовлетворенности работой (на основе свободных ответов)

Маленькая зарплата
Жара в кабинетах
Неудобные стулья
Зарплата не соответствует объему работы, который я выполняю
В кабинете постоянно жарко, невозможно работать
Плохая уборка помещений, везде пыль
Премии весьма символические
Зарплата никуда не годится



Причины неудовлетворенности работой (на основе свободных ответов)

Финансовые причины:
Размер заработной платы
Размер премиальных выплат
Санитарно-гигиенические условия:
Температура в рабочем кабинете
Неподходящая мебель
Некачественная уборка помещений

Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57