Лекция 6

Анализ информации, данных

Данные

В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.

Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.

Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки.

Иными словами, **данные** - это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

Данные

Объект описывается как набор атрибутов.

Атрибут - свойство, характеризующее объект.

Атрибут также на
измерением, характе

Переменная (variab
для всех изучаемь
изменяться от объек

Значение (value) пер

	Атрибуты				
Объекты	Код клиента	Возраст	Семейное положение	Доход	Класс
	1	18	Single	125	1
	2	22	Married	100	1
	3	30	Single	70	1
	4	32	Married	120	1
	5	24	Divorced	95	2
	6	25	Married	60	1
	7	32	Divorced	220	1
	8	19	Single	85	2
	9	22	Married	75	1
	10	40	Single	90	2

Данные

Генеральная совокупность (population) - вся совокупность изучаемых объектов, интересующая исследователя.

Выборка (sample) - часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

Параметры - числовые характеристики генеральной совокупности.

Статистики - числовые характеристики выборки.

Гипотеза - частично обоснованная закономерность знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

Измерения

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

Дискретные данные являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.

Непрерывные данные - данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность.

Шкалы

Существует пять типов шкал измерений: *номинальная, порядковая, интервальная, относительная и дихотомическая.*

Номинальная шкала (nominal scale) - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Шкалы

Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.

Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории.

Шкалы

Номер объекта	Профессия (номинальная шкала)	Средний балл (интервальная шкала)	Образование (порядковая шкала)
1	слесарь	22	среднее
2	ученый	55	высшее
3	учитель	47	высшее

Дата измерения	Облачность (номинальная шкала)	Температура в 8 часов утра (интервальная шкала)	Сила ветра (порядковая шкала)
1 сентября	облачно	22 deg C	Ветер сильный
2 сентября	пасмурно	17 deg C	Ветер слабый
3 сентября	ясно	23 deg C	Ветер очень сильный

Типы наборов данных

Данные, состоящие из записей

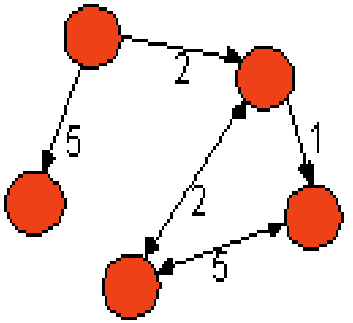
Табличные данные - данные, состоящие из записей, каждая из которых состоит из фиксированного набора атрибутов.

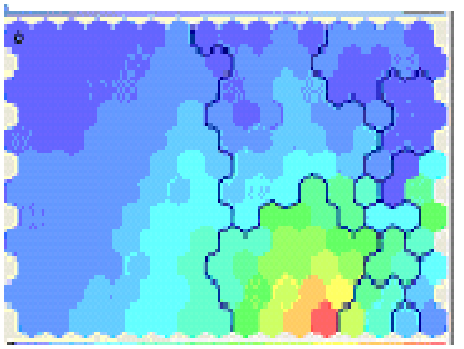
Транзакционные данные представляют собой особый тип данных, где каждая запись, являющаяся транзакцией, включает набор значений.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

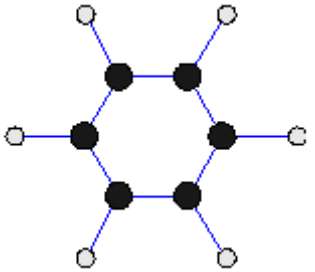
Типы наборов данных

Графические данные
Граф, карты





Химические данные

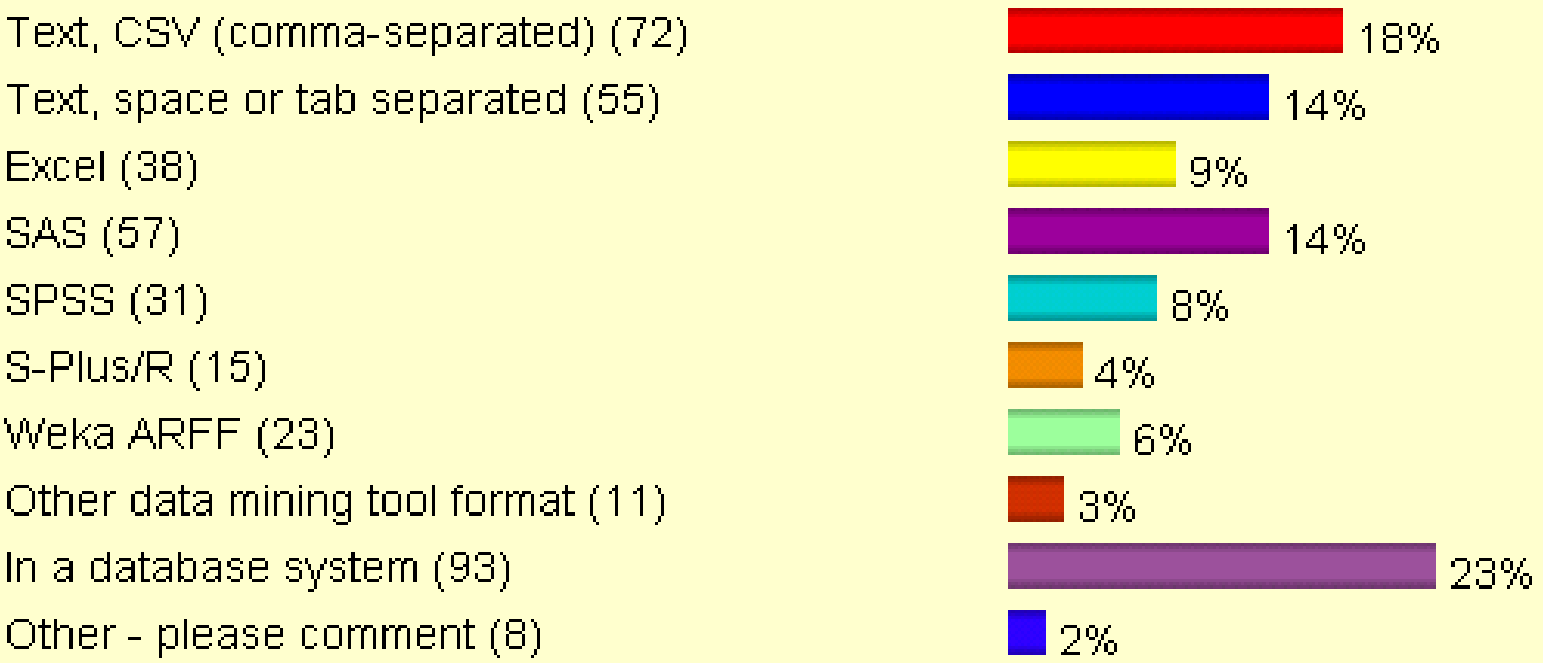


Форматы хранения данных

KDnuggets : **Polls** : Data Storage Formats (June 2005)

Poll

What are your preferred methods for storing data for data mining? [403 votes total]



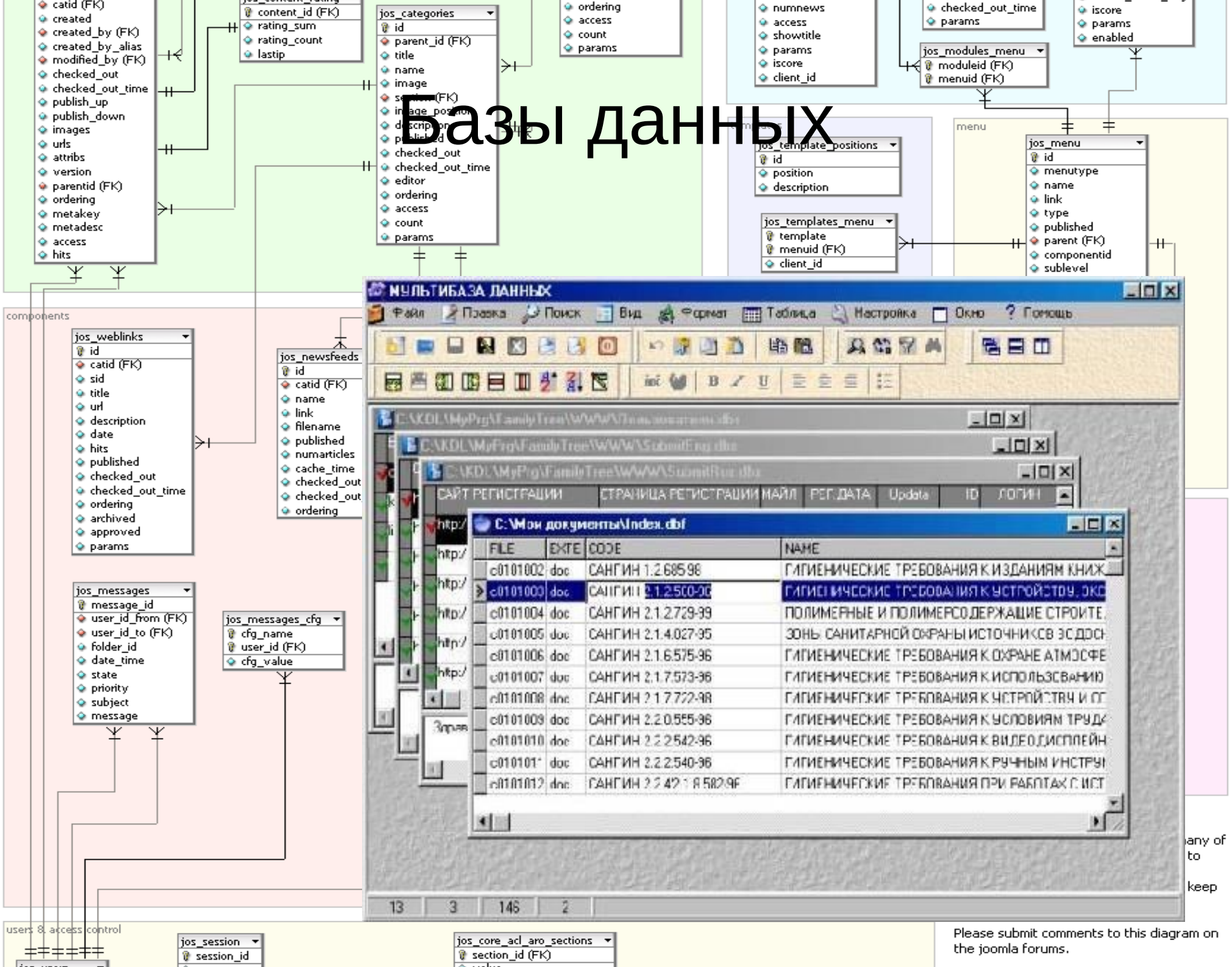
Базы данных

Совокупность связанных данных, организованных по определенным правилам, предусматривающим общие принципы описания, хранения и манипулирования, независимая от прикладных программ. База данных является информационной моделью предметной области. Обращение к базам данных осуществляется с помощью системы управления базами данных.

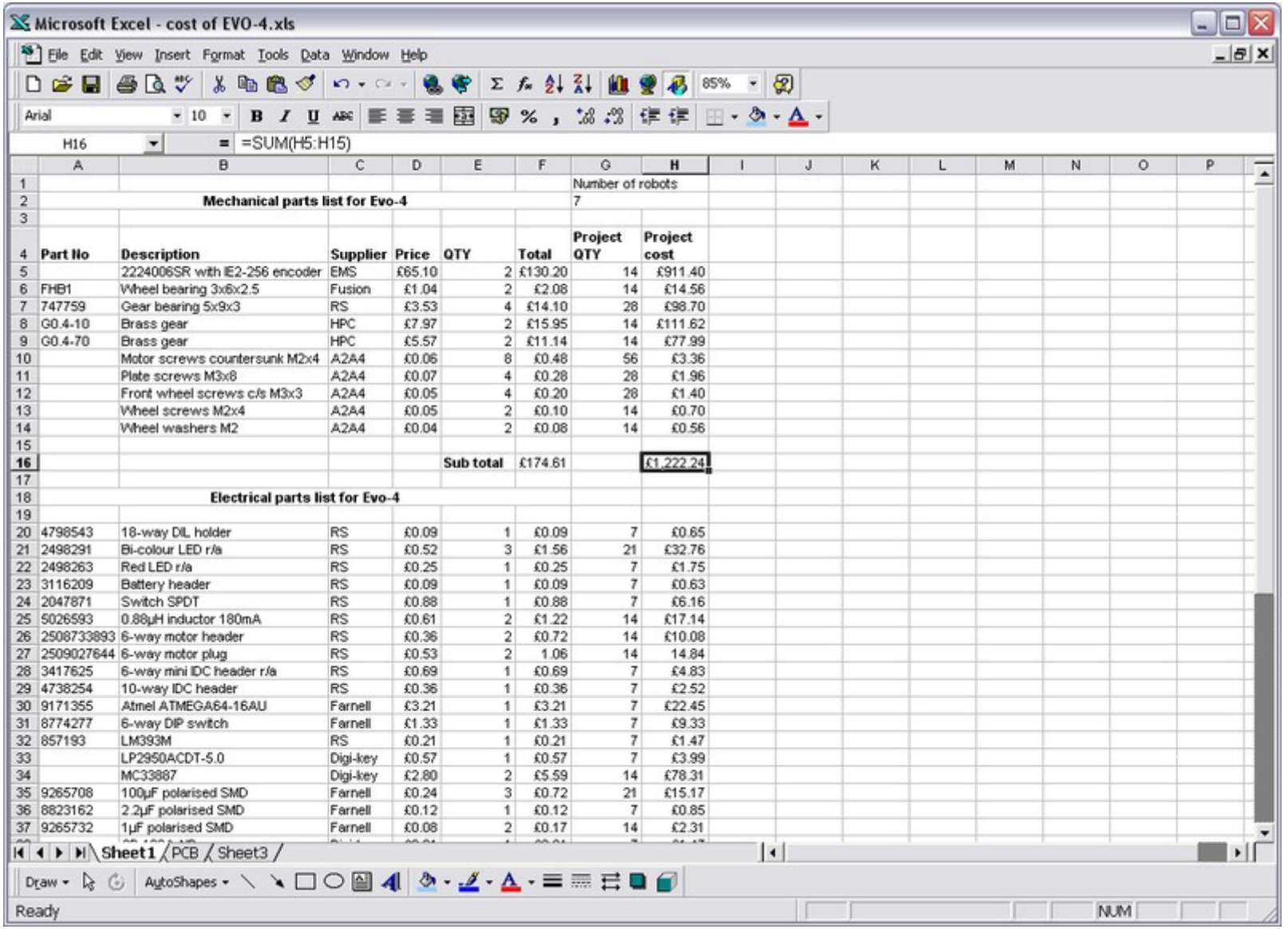
по законодательству РФ

Объективная форма представления и организации совокупности данных, систематизированных таким образом, чтобы эти данные могли быть найдены и обработаны с помощью ЭВМ

Базы данных


jos_categories
id
parent_id (FK)
title
name
image
checked_out
checked_out_time
editor
ordering
access
count
params
jos_template_positions
id
position
description
jos_templates_menu
template
menuid (FK)
client_id
jos_menu
id
menutype
name
link
type
published
parent (FK)
componentid
sublevel
jos_modules_menu
moduleid (FK)
menuid (FK)
jos_weblinks
id
catid (FK)
sid
title
url
description
date
hits
published
checked_out
checked_out_time
ordering
archived
approved
params
jos_newsfeeds
id
catid (FK)
name
link
filename
published
numarticles
cache_time
checked_out
checked_out
ordering
jos_messages
message_id
user_id_from (FK)
user_id_to (FK)
folder_id
date_time
state
priority
subject
message
jos_messages_cfg
cfg_name
user_id (FK)
cfg_value
jos_session
session_id
jos_core_acl_aro_sections
section_id (FK)
components
menu
users & access control
МУЛЬТИБАЗА ДАННЫХ
Файл
Правка
Поиск
Вид
Формат
Таблица
Настройка
Окно
Помощь
САЙТ РЕГИСТРАЦИИ
СТРАНИЦА РЕГИСТРАЦИИ
МАЙЛ
РЕГДАТА
Update
ID
ЛОГИН
C:\Мои документы\Index.dbf
FILE
EXTE
CODE
NAME
c0101002 doc САНПИН 1.2.685-98 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К ИЗДАНИЯМ КНИЖ
c0101003 doc САНПИН 2.1.2.500-00 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К УСТРОЙСТВУ, ОКО
c0101004 doc САНПИН 2.1.2.729-99 ПОЛИМЕРНЫЕ И ПОЛИМЕРСОДЕРЖАЩИЕ СТРОИТЕ
c0101005 doc САНПИН 2.1.4.027-95 ЗОНЫ САНИТАРНОЙ ОХРАНЫ ИСТОЧНИКОВ ВОДОСН
c0101006 doc САНПИН 2.1.6.575-96 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К ОХРАНЕ АТМОСФЕ
c0101007 doc САНПИН 2.1.7.573-96 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К ИСПОЛЬЗОВАНИЮ
c0101008 doc САНПИН 2.1.7.722-98 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К УСТРОЙСТВУ И СО
c0101009 doc САНПИН 2.2.0.555-96 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К УСЛОВИЯМ ТРУДА
c0101010 doc САНПИН 2.2.2.542-96 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К ВИДЕОДИСПЛЕЙН
c0101011 doc САНПИН 2.2.2.540-96 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ К РУЧНЫМ ИНСТРУ
c0101012 doc САНПИН 2.2.4/2.1.8.582-96 ГИГИЕНИЧЕСКИЕ ТРЕБОВАНИЯ ПРИ РАБОТАХ С ИСТ
13
3
146
2
Please submit comments to this diagram on the joomla forums.

Excel (xls, xlsx)



Microsoft Excel - cost of EVO-4.xls

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U

H16 =SUM(H5:H15)

	A	B	C	D	E	F	G	H
1							Number of robots	
2		**Mechanical parts list for Evo-4**					7	
3								
4	**Part No**	**Description**	**Supplier**	**Price**	**QTY**	**Total**	**Project QTY**	**Project cost**
5		2224006SR with IE2-256 encoder	EMS	£65.10	2	£130.20	14	£911.40
6	FHB1	Wheel bearing 3x6x2.5	Fusion	£1.04	2	£2.08	14	£14.56
7	747759	Gear bearing 5x9x3	RS	£3.53	4	£14.10	28	£98.70
8	G0.4-10	Brass gear	HPC	£7.97	2	£15.95	14	£111.62
9	G0.4-70	Brass gear	HPC	£5.57	2	£11.14	14	£77.99
10		Motor screws countersunk M2x4	A2A4	£0.06	8	£0.48	56	£3.36
11		Plate screws M3x8	A2A4	£0.07	4	£0.28	28	£1.96
12		Front wheel screws c/s M3x3	A2A4	£0.05	4	£0.20	28	£1.40
13		Wheel screws M2x4	A2A4	£0.05	2	£0.10	14	£0.70
14		Wheel washers M2	A2A4	£0.04	2	£0.08	14	£0.56
15								
16					**Sub total**	£174.61		£1,222.24
17								
18		**Electrical parts list for Evo-4**						
19								
20	4798543	18-way DIL holder	RS	£0.09	1	£0.09	7	£0.65
21	2498291	Bi-colour LED r/a	RS	£0.52	3	£1.56	21	£32.76
22	2498263	Red LED r/a	RS	£0.25	1	£0.25	7	£1.75
23	3116209	Battery header	RS	£0.09	1	£0.09	7	£0.63
24	2047871	Switch SPDT	RS	£0.88	1	£0.88	7	£6.16
25	5026593	0.88µH inductor 180mA	RS	£0.61	2	£1.22	14	£17.14
26	2508733893	6-way motor header	RS	£0.36	2	£0.72	14	£10.08
27	2509027644	6-way motor plug	RS	£0.53	2	1.06	14	14.84
28	3417625	6-way mini IDC header r/a	RS	£0.69	1	£0.69	7	£4.83
29	4738254	10-way IDC header	RS	£0.36	1	£0.36	7	£2.52
30	9171355	Atmel ATMEGA64-16AU	Farnell	£3.21	1	£3.21	7	£22.45
31	8774277	6-way DIP switch	Farnell	£1.33	1	£1.33	7	£9.33
32	857193	LM393M	RS	£0.21	1	£0.21	7	£1.47
33		LP2950ACDT-5.0	Digi-key	£0.57	1	£0.57	7	£3.99
34		MC33887	Digi-key	£2.80	2	£5.59	14	£78.31
35	9265708	100µF polarised SMD	Farnell	£0.24	3	£0.72	21	£15.17
36	8823162	2.2µF polarised SMD	Farnell	£0.12	1	£0.12	7	£0.85
37	9265732	1µF polarised SMD	Farnell	£0.08	2	£0.17	14	£2.31

Sheet1 PCB Sheet3

Draw AutoShapes

Ready NUM

CSV

CSV (от англ. **Comma-Separated Values** — значения, разделённые запятыми) — текстовый формат, предназначенный для представления табличных данных. Каждая строка файла — это одна строка таблицы. Значения отдельных колонок разделяются разделительным символом (*delimiter*) — запятой (,). Однако, большинство программ вольно трактует стандарт CSV и допускают использование иных символы в качестве разделителя. В частности в локалях, где десятичным разделителем является запятая, в качестве табличного разделителя, как правило, используется точка с запятой. Значения, содержащие зарезервированные символы (пробел, запятая, точка с запятой, новая строка) обрамляются двойными кавычками ("); если в значении встречаются кавычки — они представляются в файле в виде двух кавычек подряд.

CSV

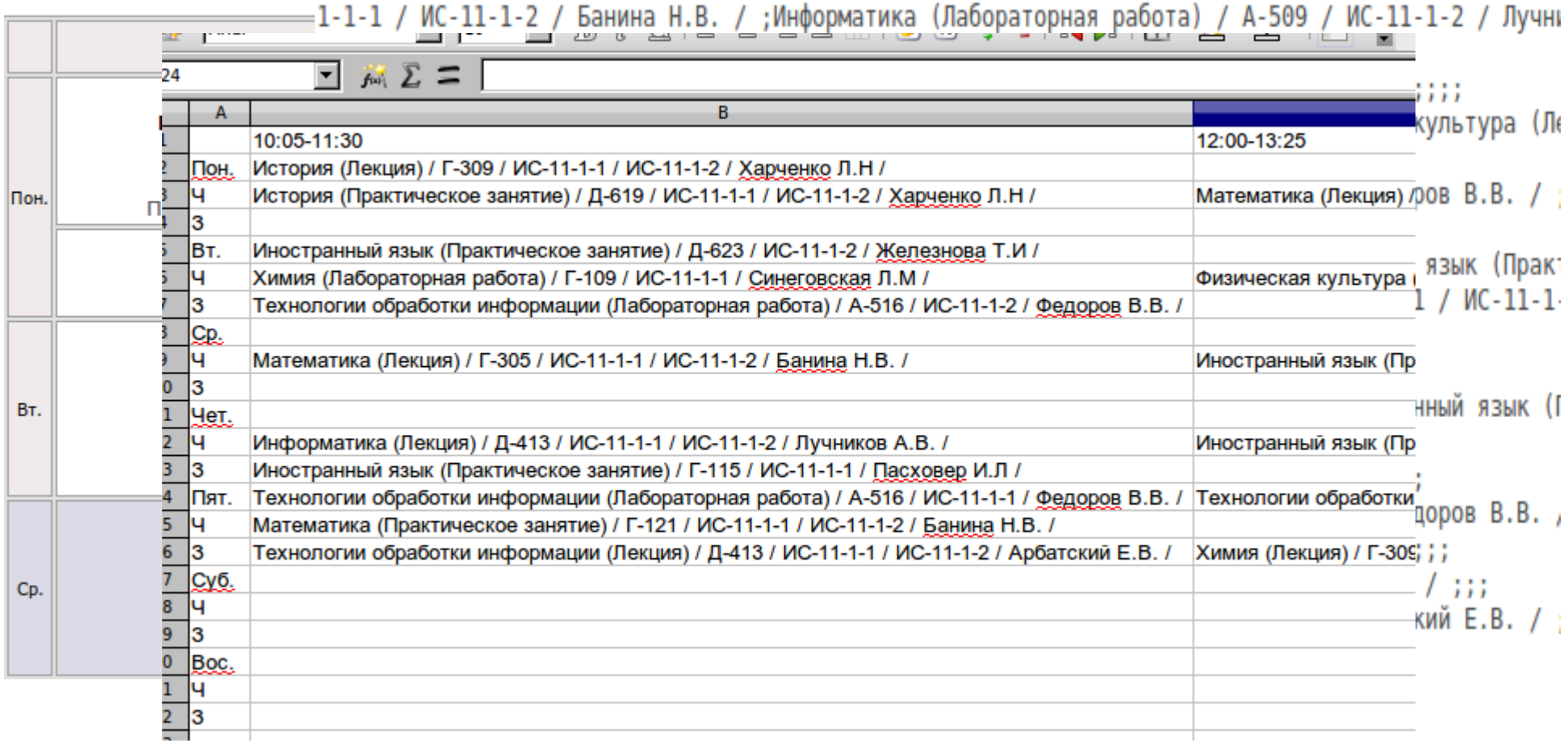
```
1965;Пиксел;E240 — формальдегид (опасный консервант)!;"красный, зелёный, битый";3000,00
1965;Мышка;"А правильней ""Использовать Ёлочки""";;4900,00
"Н/д";Кнопка;Сочетания клавиш;"MUST USE! Ctrl, Alt, Shift";4799,00
```

Результирующая таблица:

1965	Пиксел	E240 — формальдегид (опасный консервант)!	красный, зелёный, битый	3000
1965	Мышка	А правильней "Использовать Ёлочки"		4900
Н/д	Кнопка	Сочетания клавиш	MUST USE! Ctrl, Alt, Shift	4799

CSV

```
eugine@eugene-945P-S3:~/Downloads$ iconv -f cp1251 IS-11-1.csv
;10:05-11:30;12:00-13:25;13:35-15:00;15:10-16:35;16:45-18:10;18:20-19:45;19:55-21:20
Пон.;;История (Лекция) / Г-309 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н / ;;;;;
Ч;;;История (Практическое занятие) / Д-619 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н / ;Математика
1-1-1 / ИС-11-1-2 / Банина Н.В. / ;Информатика (Лабораторная работа) / А-509 / ИС-11-1-2 / Лучни
```



	A	B	
1		10:05-11:30	12:00-13:25
2	Пон.	История (Лекция) / Г-309 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н /	
3	Ч	История (Практическое занятие) / Д-619 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н /	Математика (Лекция) /
4	З		
5	Вт.	Иностранный язык (Практическое занятие) / Д-623 / ИС-11-1-2 / Железнова Т.И /	
6	Ч	Химия (Лабораторная работа) / Г-109 / ИС-11-1-1 / Синеговская Л.М /	Физическая культура (
7	З	Технологии обработки информации (Лабораторная работа) / А-516 / ИС-11-1-2 / Федоров В.В. /	
8	Ср.		
9	Ч	Математика (Лекция) / Г-305 / ИС-11-1-1 / ИС-11-1-2 / Банина Н.В. /	Иностранный язык (Пр
10	З		
11	Чет.		
12	Ч	Информатика (Лекция) / Д-413 / ИС-11-1-1 / ИС-11-1-2 / Лучников А.В. /	Иностранный язык (Пр
13	З	Иностранный язык (Практическое занятие) / Г-115 / ИС-11-1-1 / Пасховер И.Л /	
14	Пят.	Технологии обработки информации (Лабораторная работа) / А-516 / ИС-11-1-1 / Федоров В.В. /	Технологии обработки
15	Ч	Математика (Практическое занятие) / Г-121 / ИС-11-1-1 / ИС-11-1-2 / Банина Н.В. /	
16	З	Технологии обработки информации (Лекция) / Д-413 / ИС-11-1-1 / ИС-11-1-2 / Арбатский Е.В. /	Химия (Лекция) / Г-309
17	Суб.		
18	Ч		
19	З		
20	Вос.		
21	Ч		
22	З		

Подготовка данных

1. Выделение анализируемых объектов
2. Определение анализируемых параметров
3. Сбор измерений
4. Приведение данных к табличному виду
5. Фильтрация / сортировка / группировка данных

Сортировка

	A
1	**Время наблюдения**
2	10:00:00 10.11.2011
3	11:00:00 10.11.2011
4	12:00:00 10.11.2011
5	11:00:00 11.11.2011
6	12:30:00 11.11.2011
7	12:00:00 11.11.2011
8	13:30:00 11.11.2011
9	13:00:00 11.11.2011
10	09:12:00 12.11.2011
11	09:34:00 12.11.2011
12	09:37:00 12.11.2011
13	10:00:00 12.11.2011
14	11:00:00 12.11.2011

	C
	Время наблюдения
	09:12:00 12.11.2011
	09:34:00 12.11.2011
	09:37:00 12.11.2011
	10:00:00 10.11.2011
	10:00:00 12.11.2011
	11:00:00 10.11.2011
	11:00:00 11.11.2011
	11:00:00 12.11.2011
	12:00:00 10.11.2011
	12:00:00 11.11.2011
	12:30:00 11.11.2011
	13:00:00 11.11.2011
	13:30:00 11.11.2011

Группировка

Под группировкой понимают расчленение единиц статистической совокупности на группы, однородные в каком-либо существенном отношении, и характеристику таких групп системой показателей в целях выделения типов явлений, изучения структуры и взаимосвязей. С помощью группировок решаются три задачи:

• разделение всей совокупности на качественно однородные группы. Эти группировки называются типологическими.

• характеристика структуры явления и структурных сдвигов. Эти группировки называются структурными.

• изучение взаимосвязей между отдельными признаками изучаемого явления. Такие группировки называются аналитическими.

Статистические расчеты

- **Количество рядов значений**;
- **Минимальные значения**;
- **Максимальные значения**;
- **Размах** - разница между наибольшим и наименьшим значениями выборки;
- Средние значения, **оценка математического ожидания**;
- **Дисперсия** - среднее арифметическое квадратов отклонений значений от их среднего;
- **Стандартное отклонение** - квадратный корень из дисперсии выборки - мера того, насколько широко разбросаны точки данных относительно их среднего.
- **Медиана**;
- **Выбросы** (outliers) - данные, резко отличающиеся от основного числа данных.

Расчеты

$$M = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$D = \frac{\sum_{i=1}^{n} (x_i - M)^2}{n}$$

$$\sigma = \sqrt{D}$$

Расчеты

• **Медиана** - *точная середина выборки, которая делит ее на две равные части по числу наблюдений.* Обязательным условием нахождения медианы является упорядоченность выборки. Таким образом, для нечетного количества наблюдений медианой выступает наблюдение с номером (n+1)/2, где n - количество наблюдений в выборке. Для четного числа наблюдений медианой является среднее значение наблюдений n/2 и (n+2)/2.

Сравнение результатов

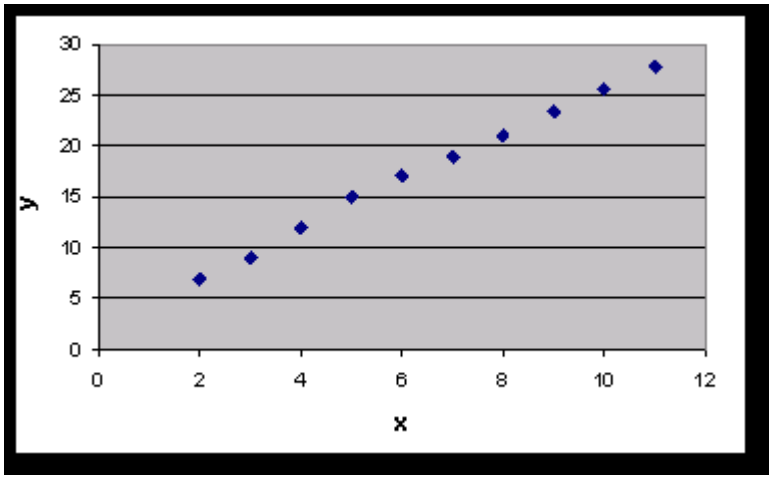
Абсолютные значения

Относительные значения

Расчеты

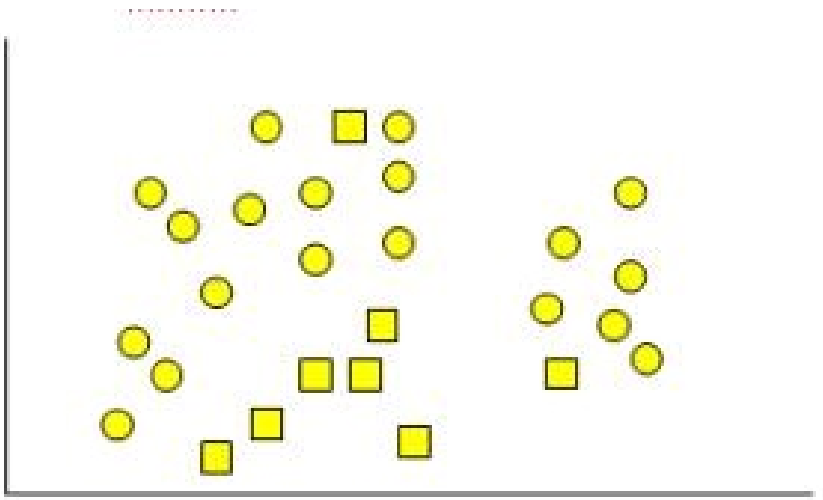
- Корреляционный анализ — обнаружение зависимости;
- Регрессионный анализ — анализ зависимостей;
- Прогнозирование;
- Тренды, сезонность;
- Кластеризация и классификация;
- Нейронные сети.

Анализ зависимостей

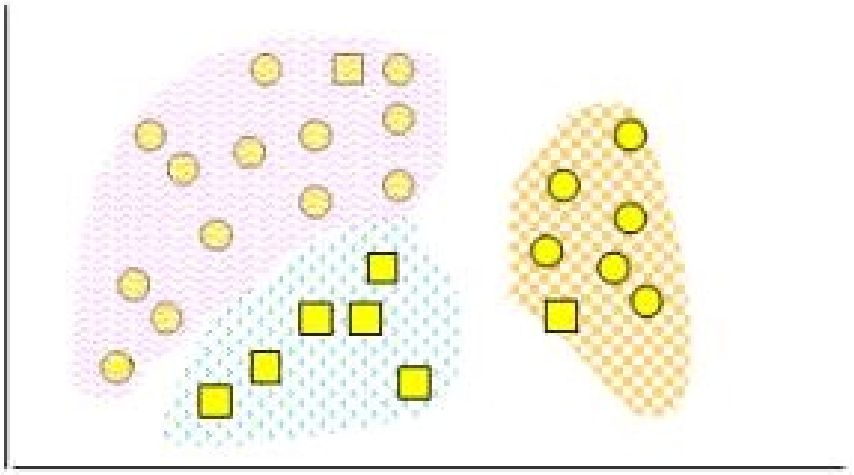


Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая
		средняя		сильная	

Классификация и кластеризация



Классификация: классы предопределены изначально



Кластеризация: классы не предопределены, осуществляется поиск наиболее похожих, однородных групп

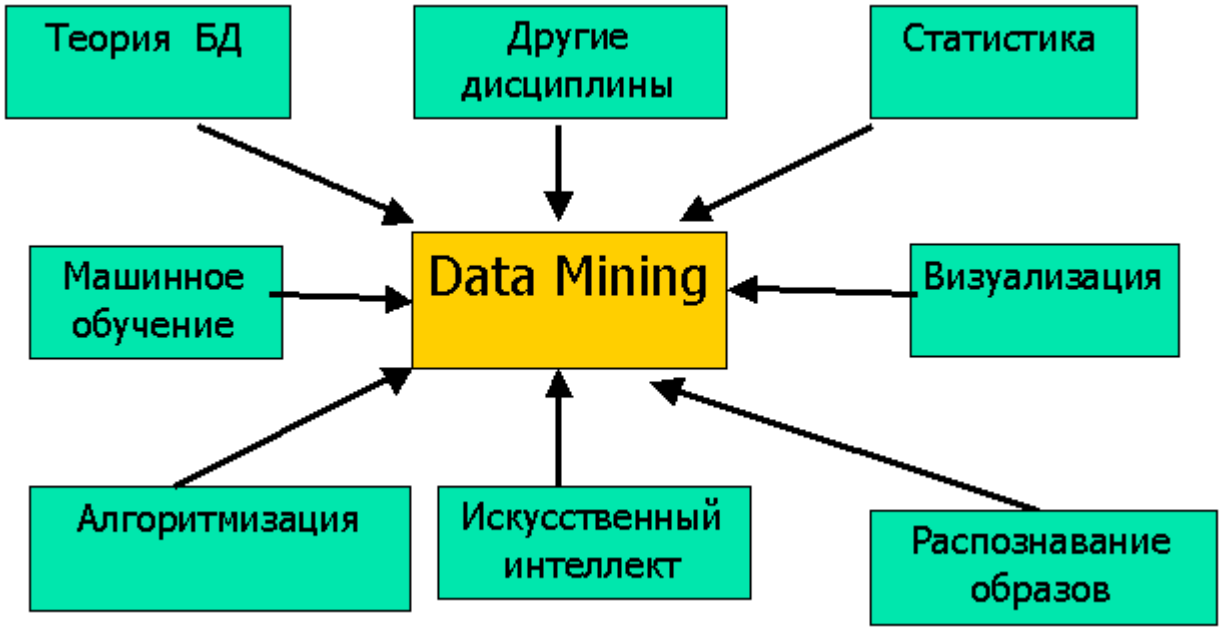
Data mining

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Data Mining (рус. *добыча данных, интеллектуальный анализ данных, глубинный анализ данных*) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

1) Термин введён Григорием Пиатецким-Шапиро в 1989 году.
2) Понятие Data Mining, появившееся в 1978 году, приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялся в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных.

Data mining


Теория БД
Другие дисциплины
Статистика
Машинное обучение
Data Mining
Визуализация
Алгоритмизация
Искусственный интеллект
Распознавание образов