

Базовые понятия интеллектуального анализа данных



*Только кончая задуманное сочинение,
мы уясняем себе, с чего нам следовало
его начать.*

Блез Паскаль

Содержание

- Феномен Больших данных
- Понятие интеллектуального анализа данных
- Технологический цикл анализа данных
- Основные задачи анализа данных
 - поиск шаблонов
 - классификация
 - кластеризация
 - поиск аномалий

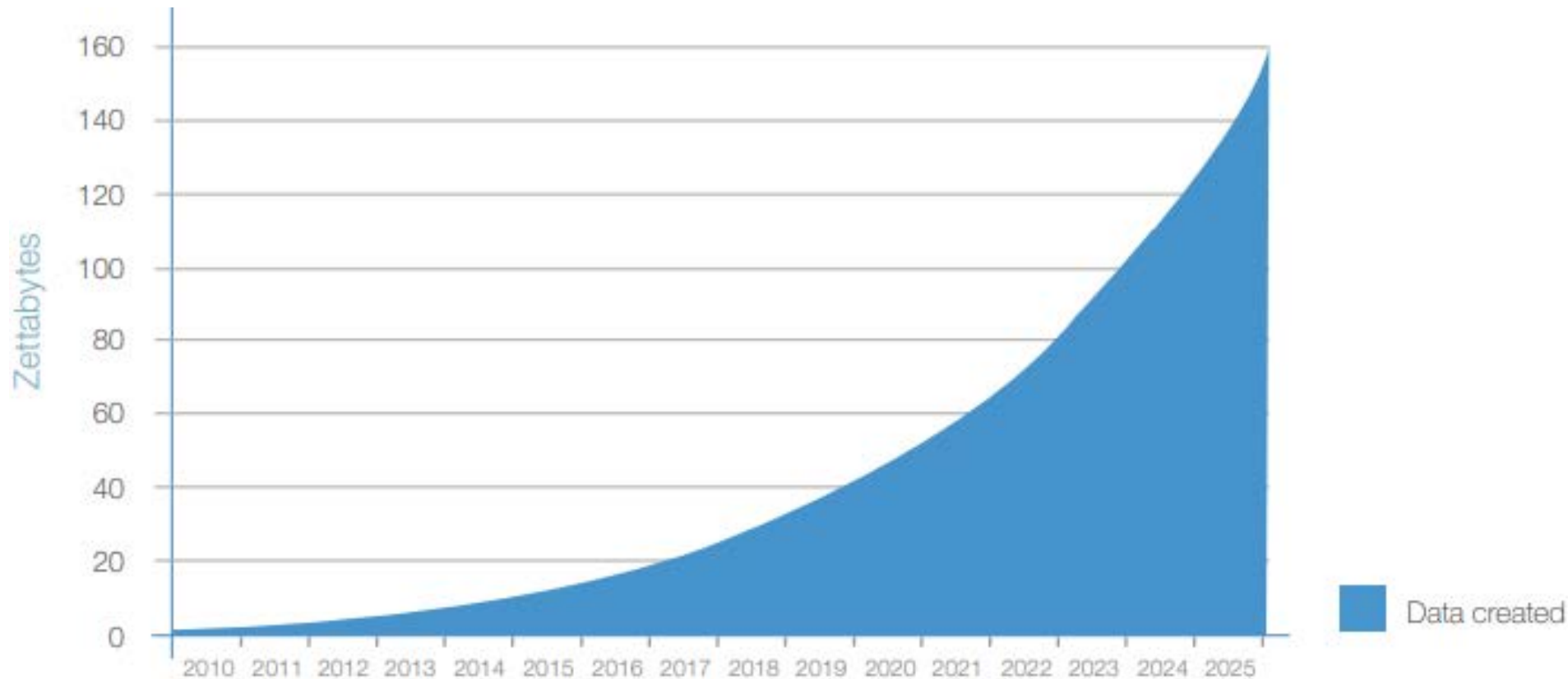
Прирост данных в современном мире



Turner V., Gantz J., Reinsel D., *et al.* The Digital Universe of opportunities: rich data and the increasing value of the Internet of Things. 2014.

URL: <https://www.iotjournaal.nl/wp-content/uploads/2017/01/idc-digital-universe-2014.pdf>.

Прирост данных в современном мире

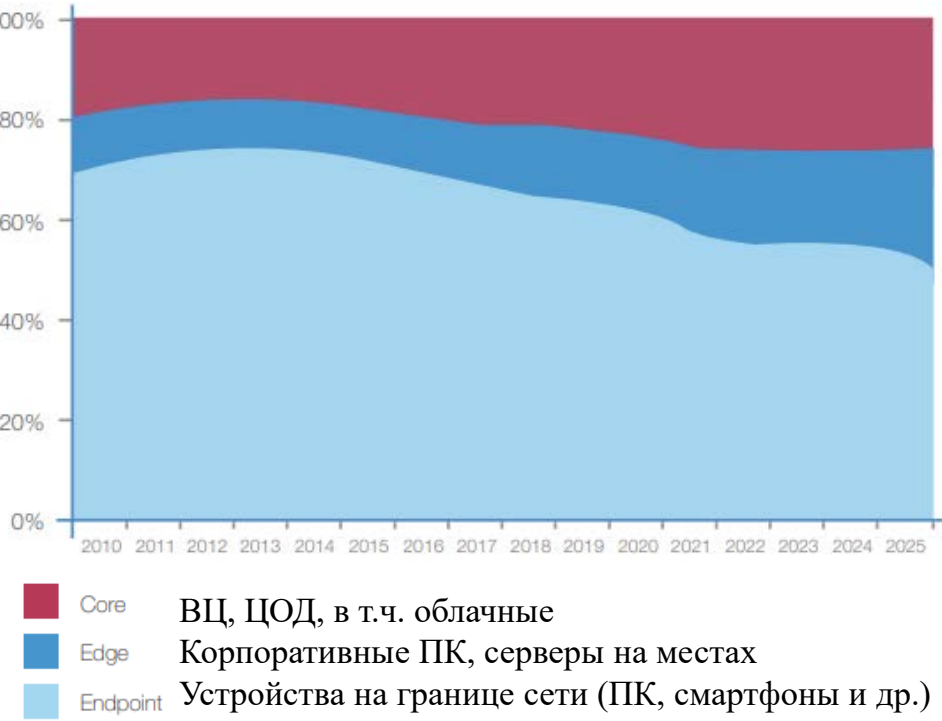


Reinsel D., Gantz J., Rydning J. Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big. 2017.

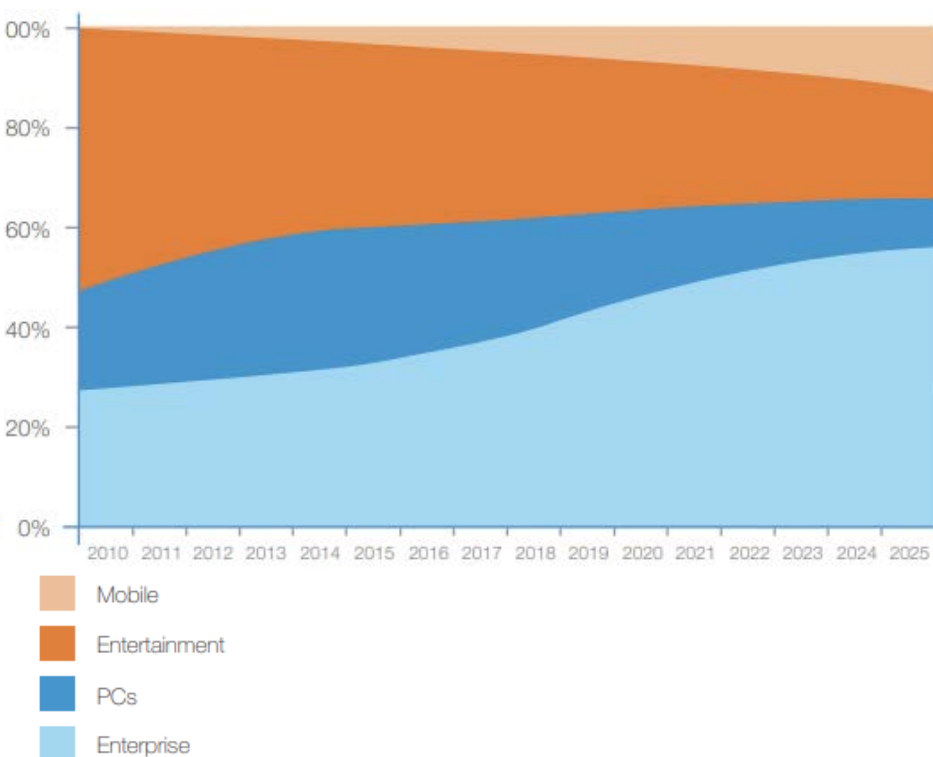
URL: <https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>.

Прирост данных в современном мире

Где создаются данные



Где хранятся данные



Reinsel D., Gantz J., Rydning J. Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big. 2017.

URL: <https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>.

Понятие Больших данных

- Большие данные (Big Data) – социально-экономический феномен, связанный с появлением технологий, позволяющих анализировать огромные массивы данных (вплоть до общемирового объема)
- Хронология основных событий
 - **2008:** спец. выпуск журнала Nature
 - **2009-11:** распространение термина в прессе и у разработчиков
 - **2011:** мировые тренды ИТ – виртуализация, Big Data, энергосбережение (Gartner)
 - **2013:** магистерская программа “Data Science” в университетах Данди (Шотландия) и Гарварде (США)
 - **2015:** Big Data прошел стадию хайпа и вступил в стадию практического применения (Gartner)



Новая идеология аккумуляирования данных

- Собирать любые возможные данные, когда и где это ВОЗМОЖНО
- Собранные данные будут иметь ценность либо для первоначальной, либо для непредвиденной цели



С. Плюшкин
1800(?) – ?

Характеристики Больших данных

- Основные
 - Volume: размер
 - Velocity: скорость прироста
 - Variety: разнообразие
- Дополнительные
 - Veracity: достоверность
 - Value: ценность результатов
 - Viability: жизнеспособность
 - Variability: изменчивость
 - Visualization: наглядное представление результатов

Интеллектуальный анализ данных (Data Mining)

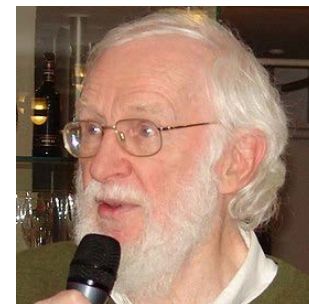
Алгоритмы и ПО для обнаружения в данных знаний, которые

- *ранее неизвестны*:
новы и не подтверждают уже имеющиеся сведения
- *нетривиальны*:
незаметны при визуальном анализе или вычислении простых статистических характеристик
- *практически полезны*:
имеют ценность для исследователя/потребителя
- *доступны для интерпретации*:
представимы в наглядной для пользователя форме и объяснимы в терминах предметной области

Frawley W.J., Piatetsky-Shapiro G., Matheus C.J. Knowledge Discovery in databases: an overview // Knowledge Discovery in Databases. AAAI/MIT Press, 1991. P. 1–30.

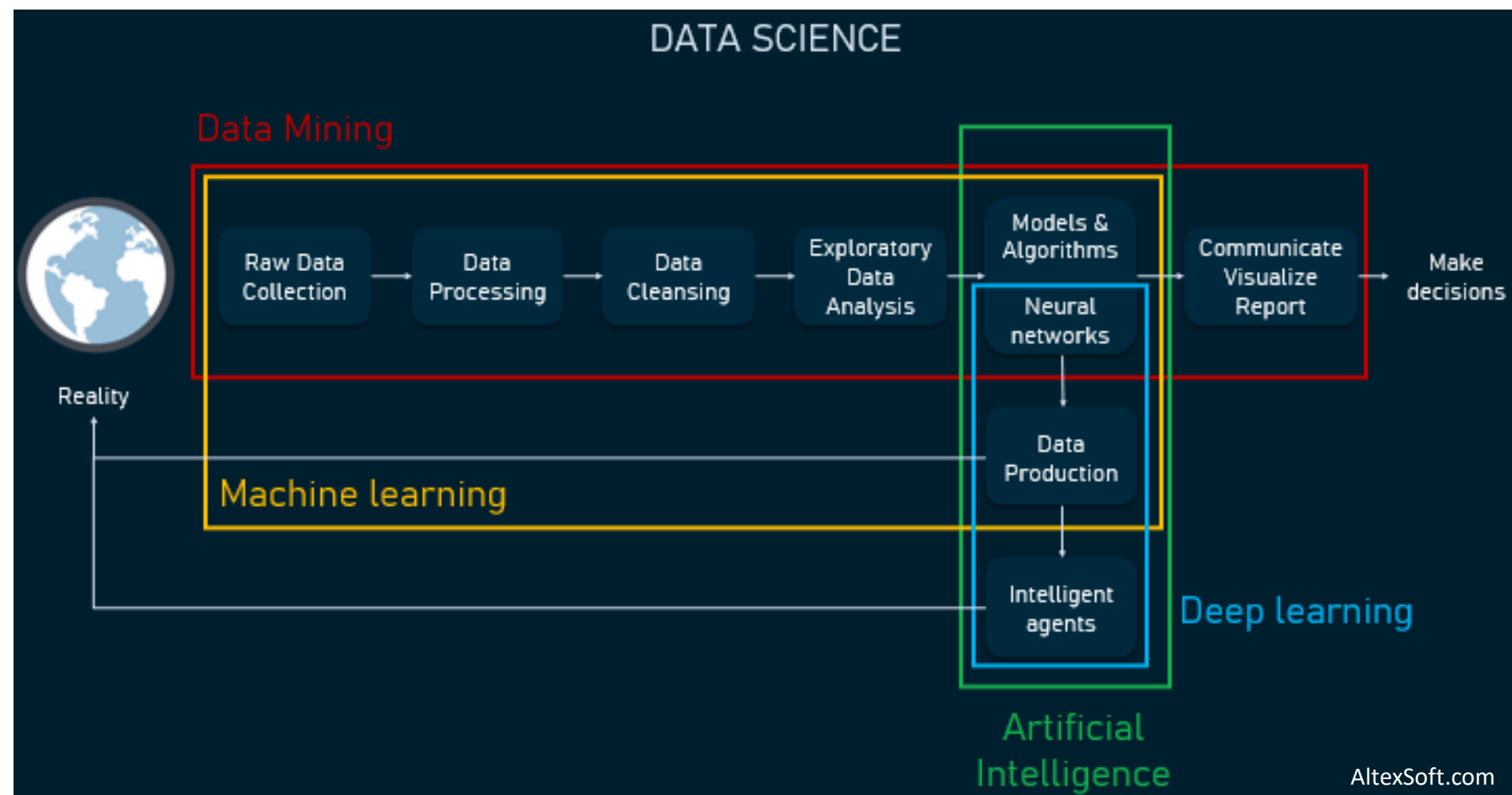
Наука о данных (даталогия, Data Science)

- Раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме
- Объединяет методы проектирования и обработки данных (в т.ч. большого объема и на основе параллелизма), статистику, интеллектуальный анализ данных, машинное и глубокое обучение, ИИ
- Хронология основных событий
 - **1966:** Учреждение CODATA (Committee on Data for Science and Technology)
 - **1974:** Naur P. A Basic Principle of Data Science // Concise Survey of Computer Methods. Lund, 1974
 - **2000:** Cleveland W.S. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics // Int. Stat. Review J. 2001. Vol. 69, no. 1. P. 21–26.
 - **2002-03:** CODATA Data Science Journal, The Journal of Data Science at Columbia University
 - **2013:** магистерская программа “Data Science” в университетах Данди (Шотландия) и Гарварде (США)

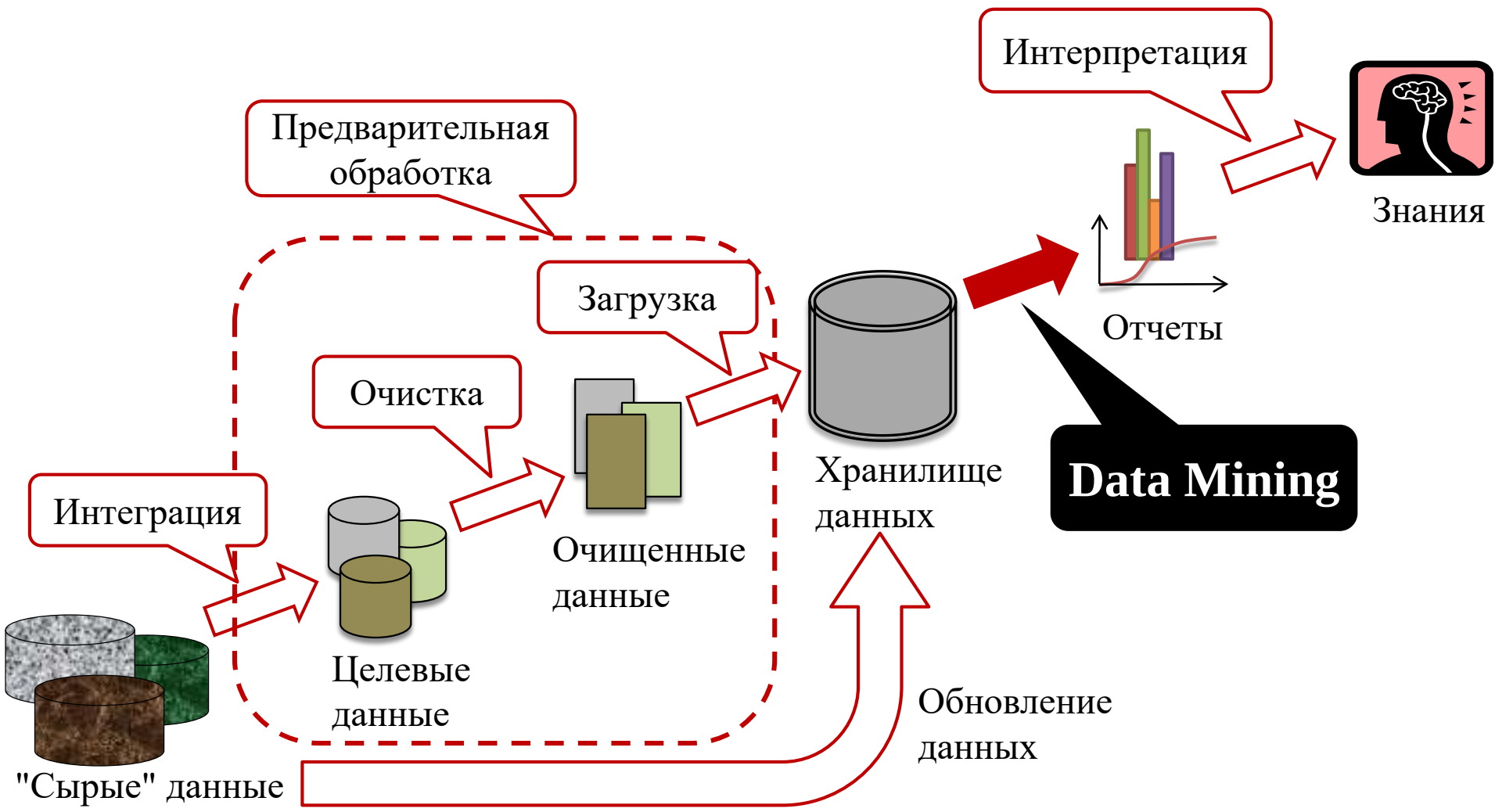


Петер Наур
1928–2016

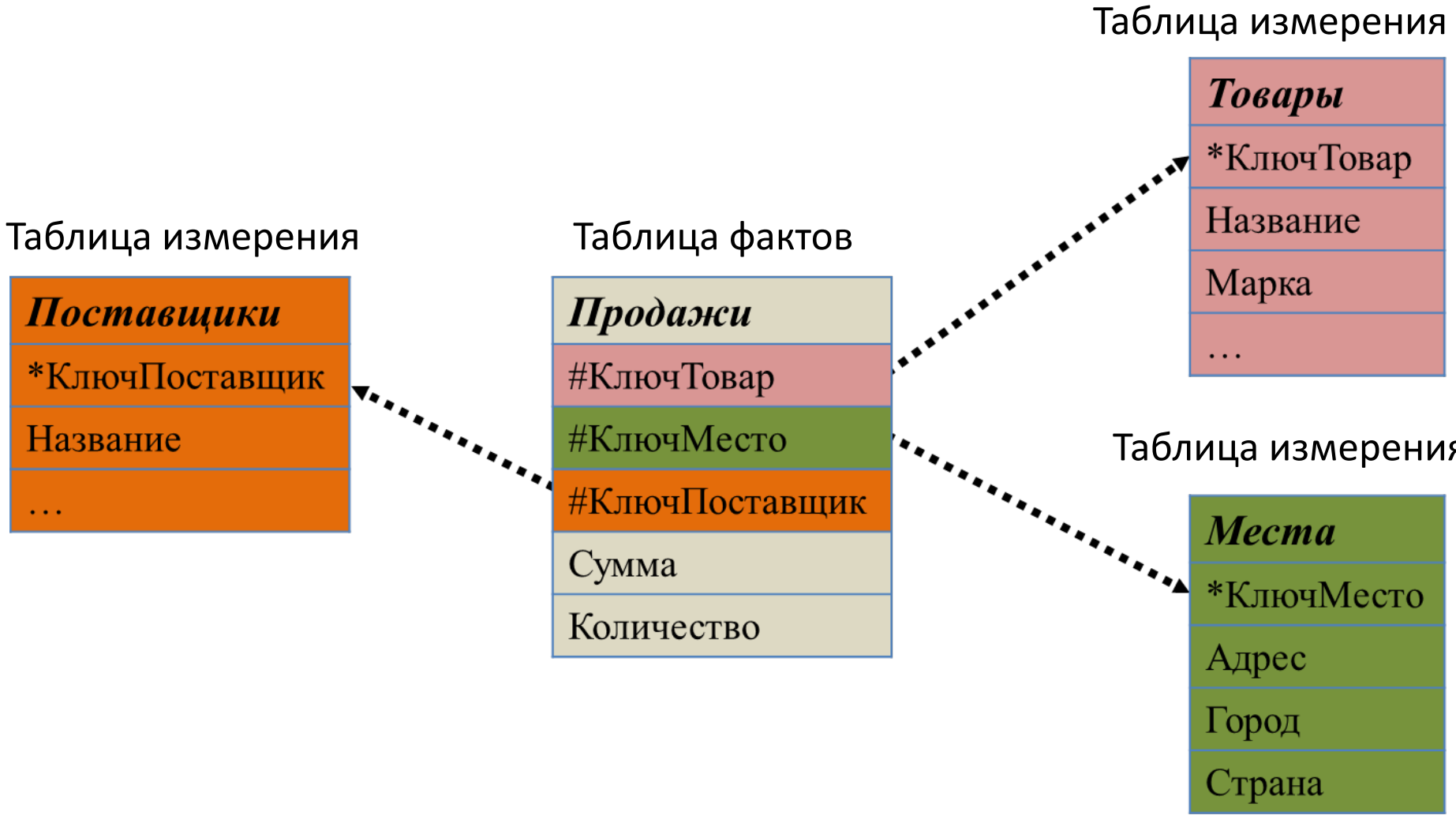
Data Science и другие термины



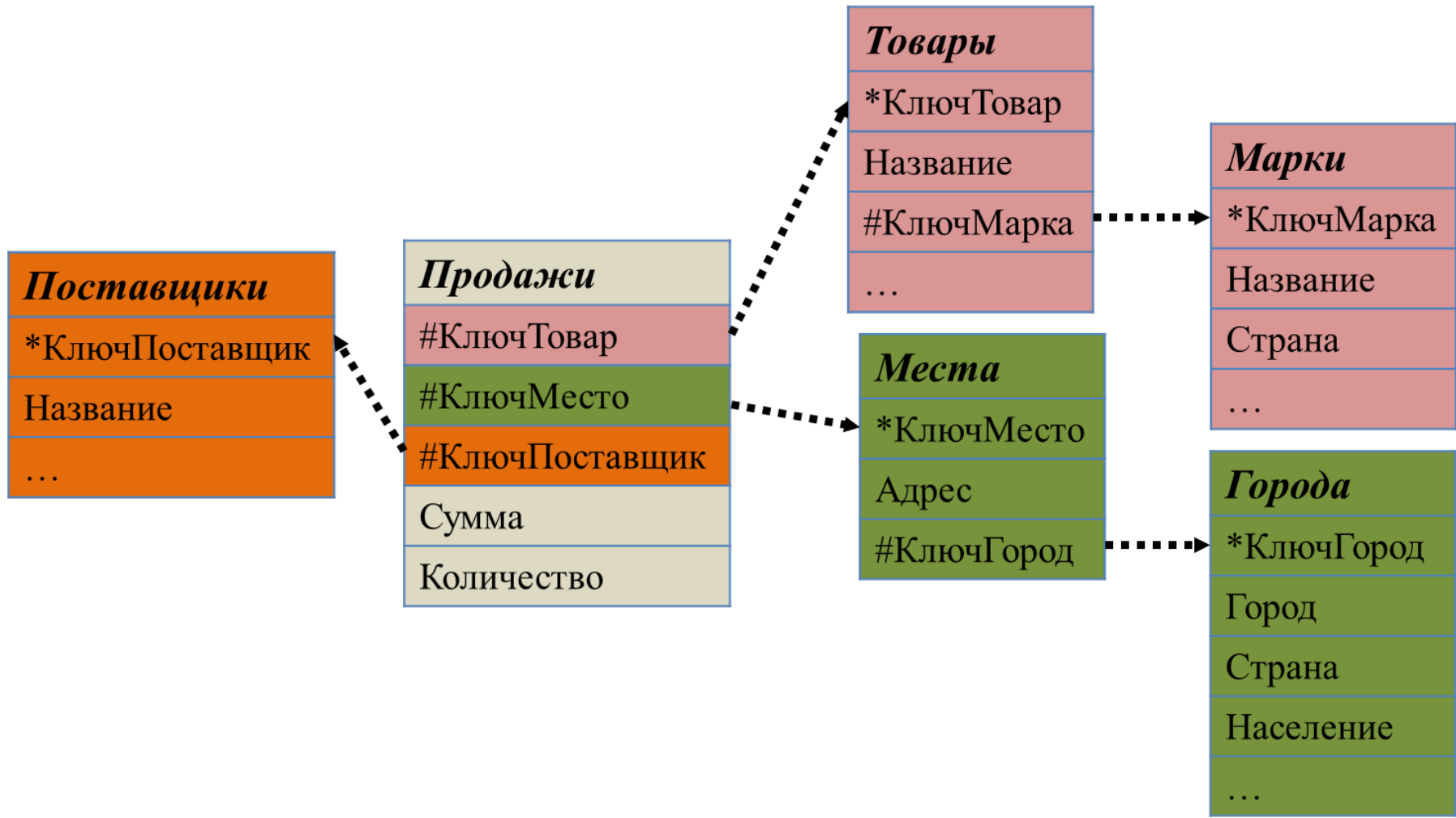
Процесс извлечения знаний (сообщество баз данных)



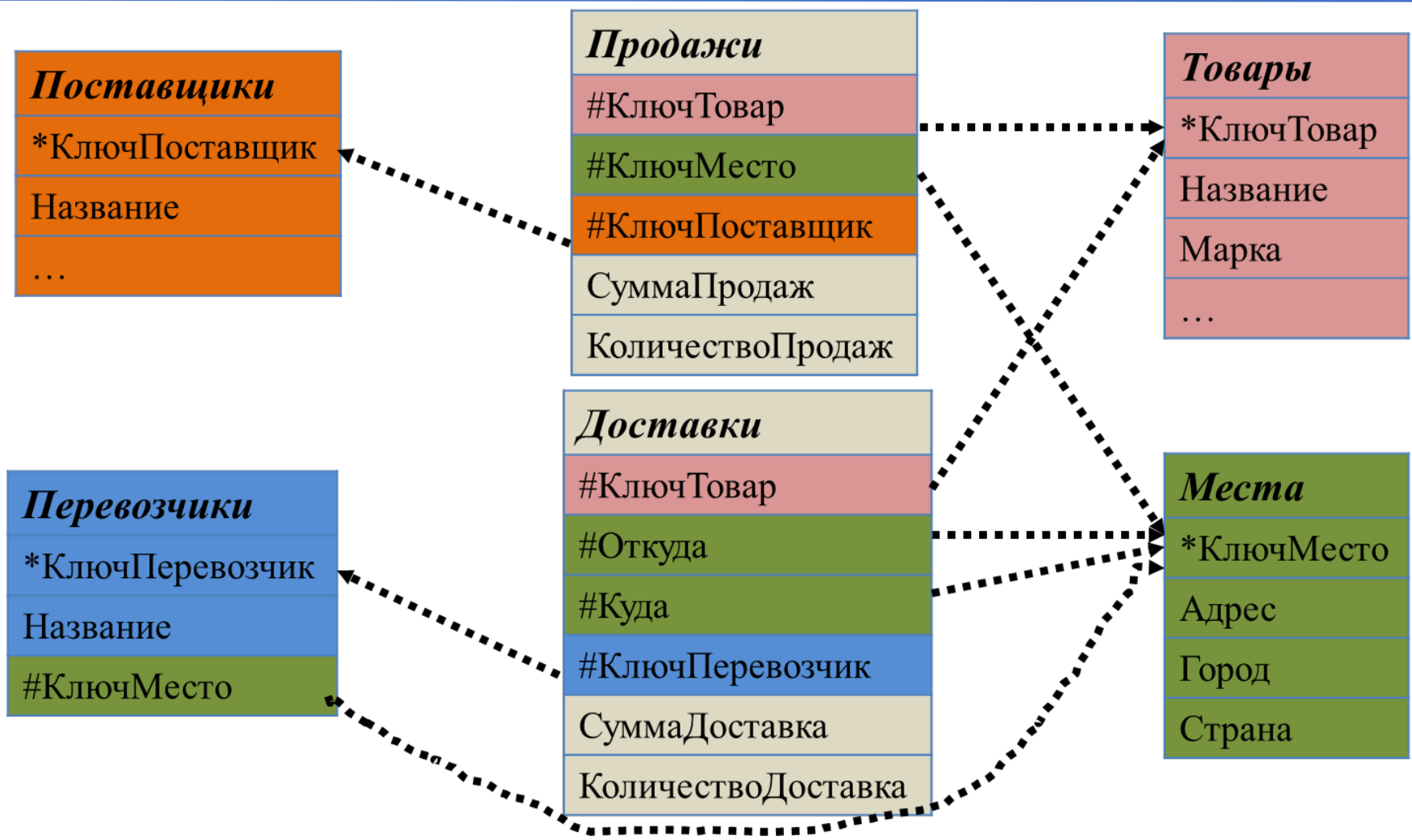
Хранилище данных: «звезда»



Хранилище данных: «снежинка»



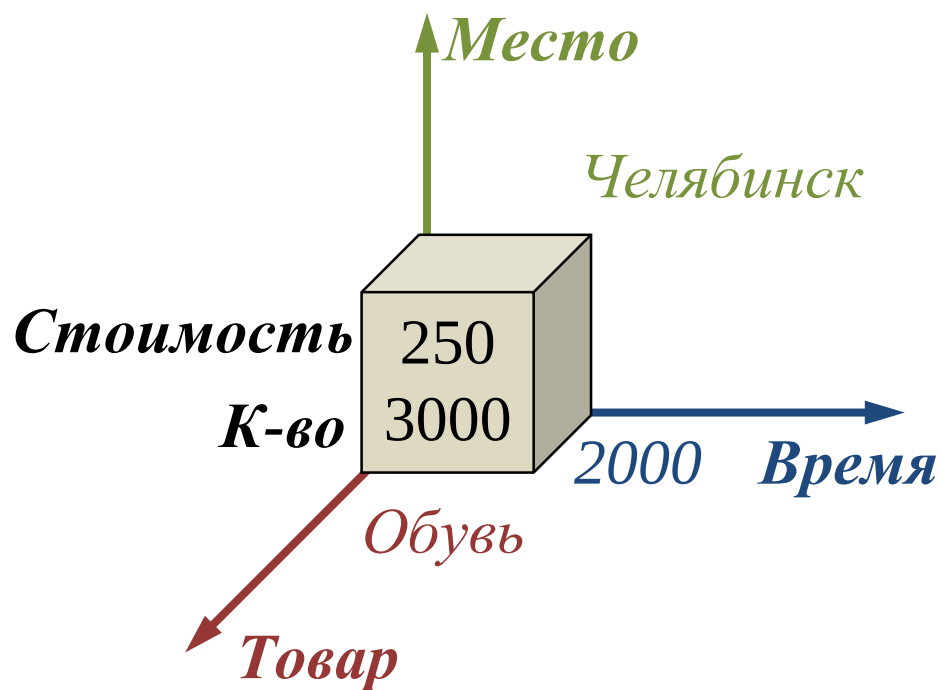
Хранилище данных: «созвездие»



НЕ интеллектуальный анализ данных

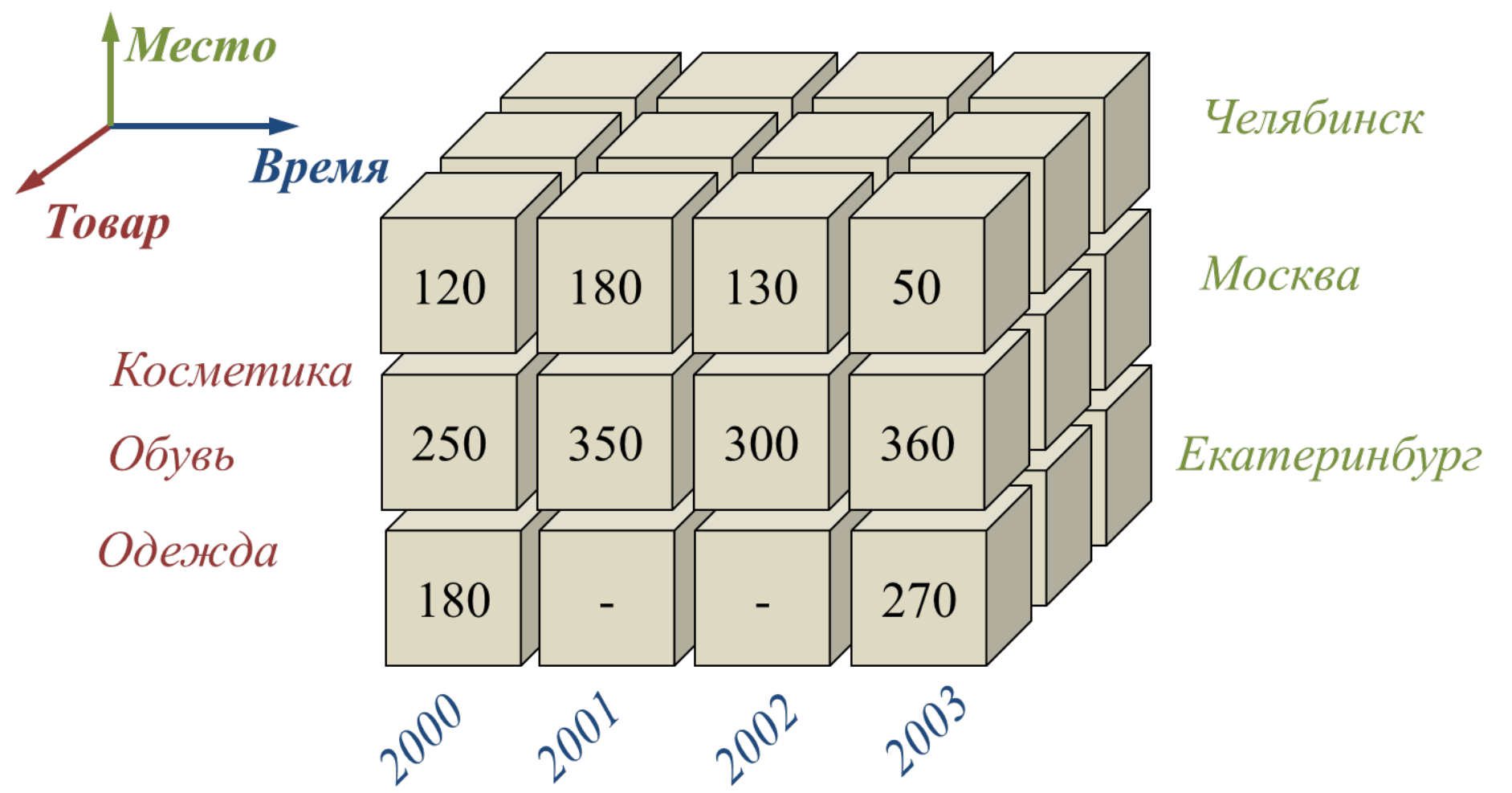
- Online Analytical Processing
 - Агрегирующие многокритериальные выборки из БД
- Online Transaction Processing
 - INSERT, DELETE, UPDATE
 - SELECT something FROM tables [WHERE condition]
 - SELECT count(*), min(*), max(*), avg(*) FROM tables [WHERE condition]

OLAP: многомерная модель данных

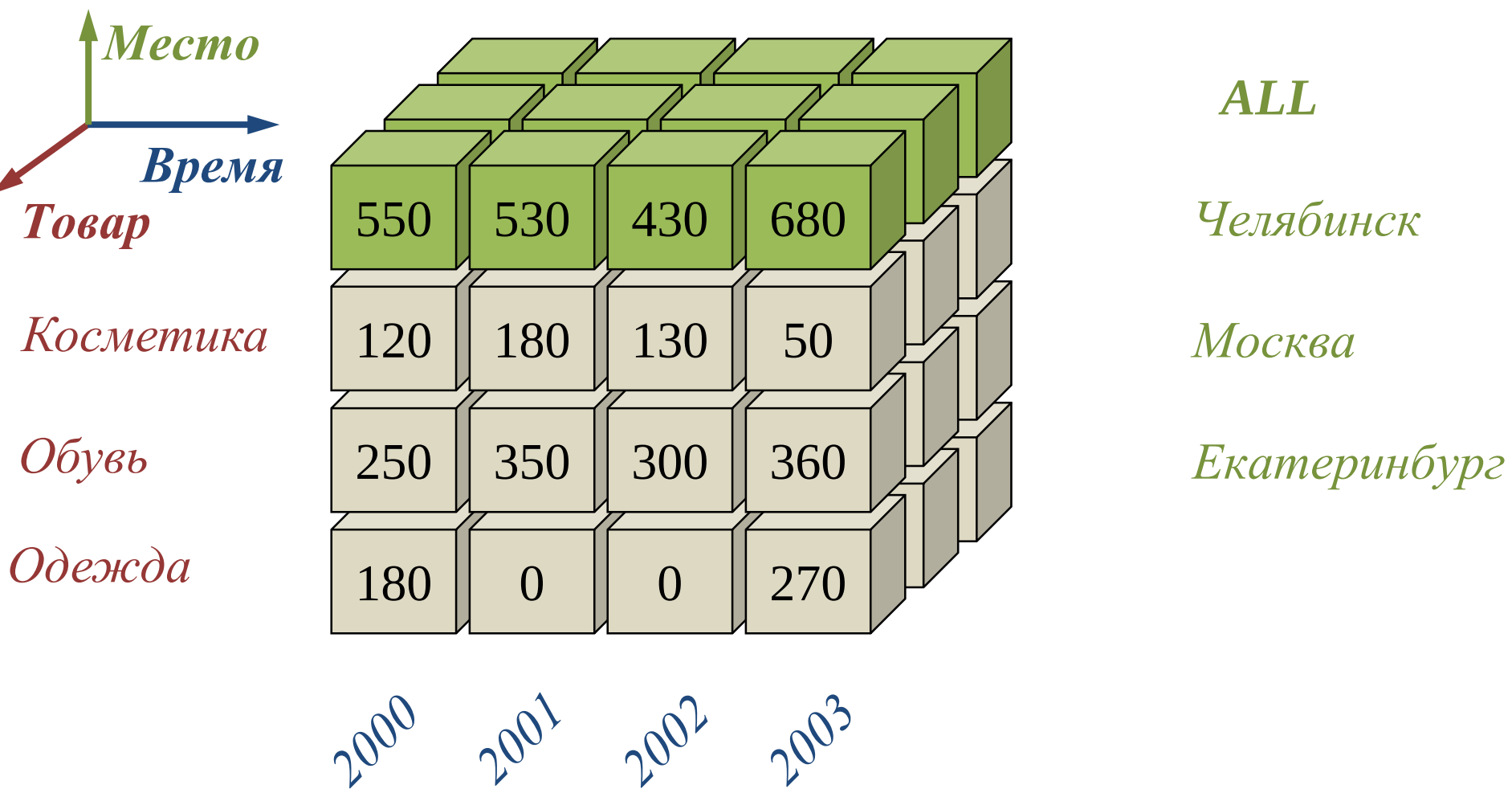


- Элемент данных – сведения о событии в предметной области
 - Измерения – набор категориальных атрибутов, характеризующих событие
 - Меры – вещественные функции от измерений

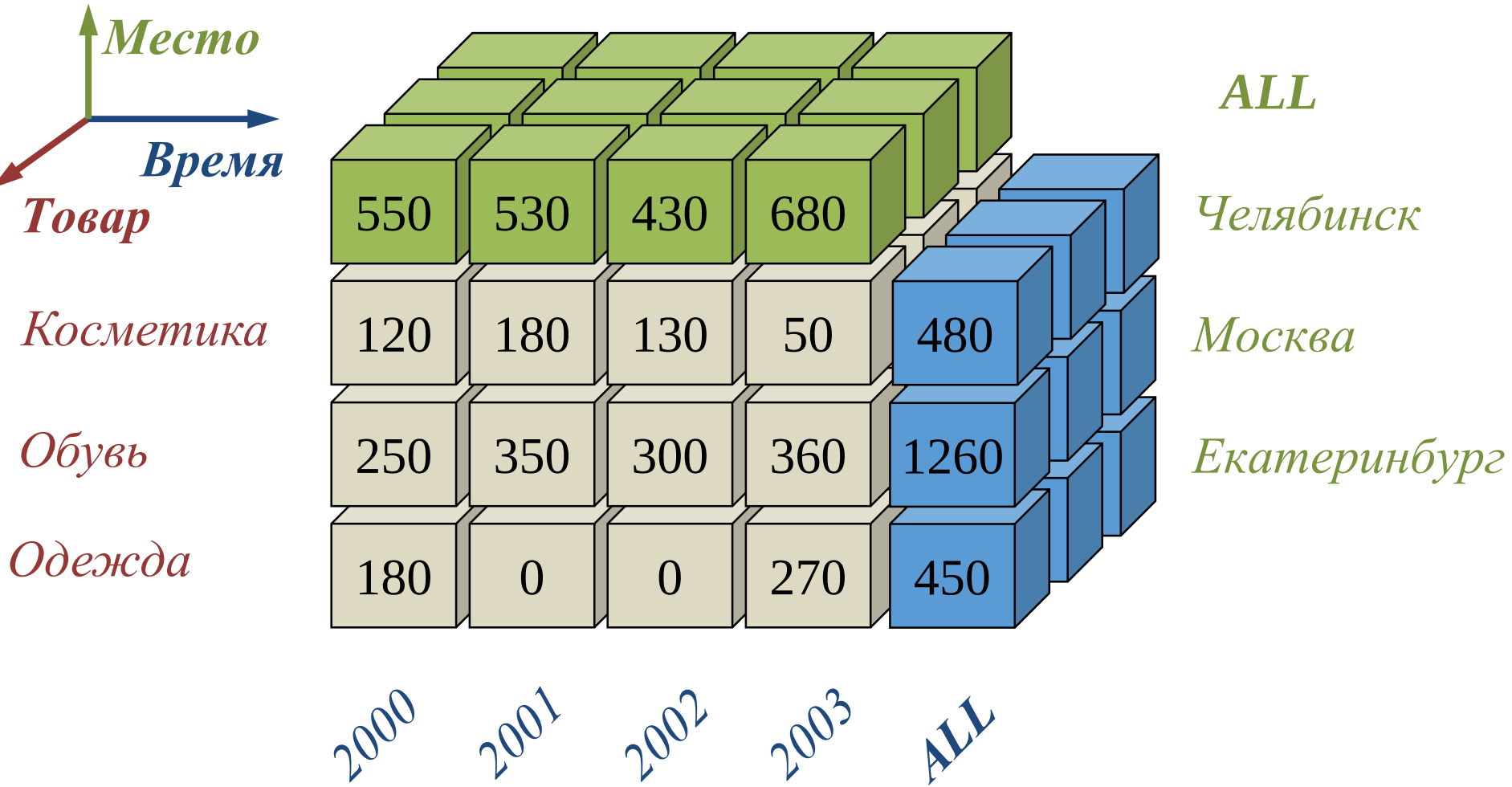
OLAP: многомерная модель данных



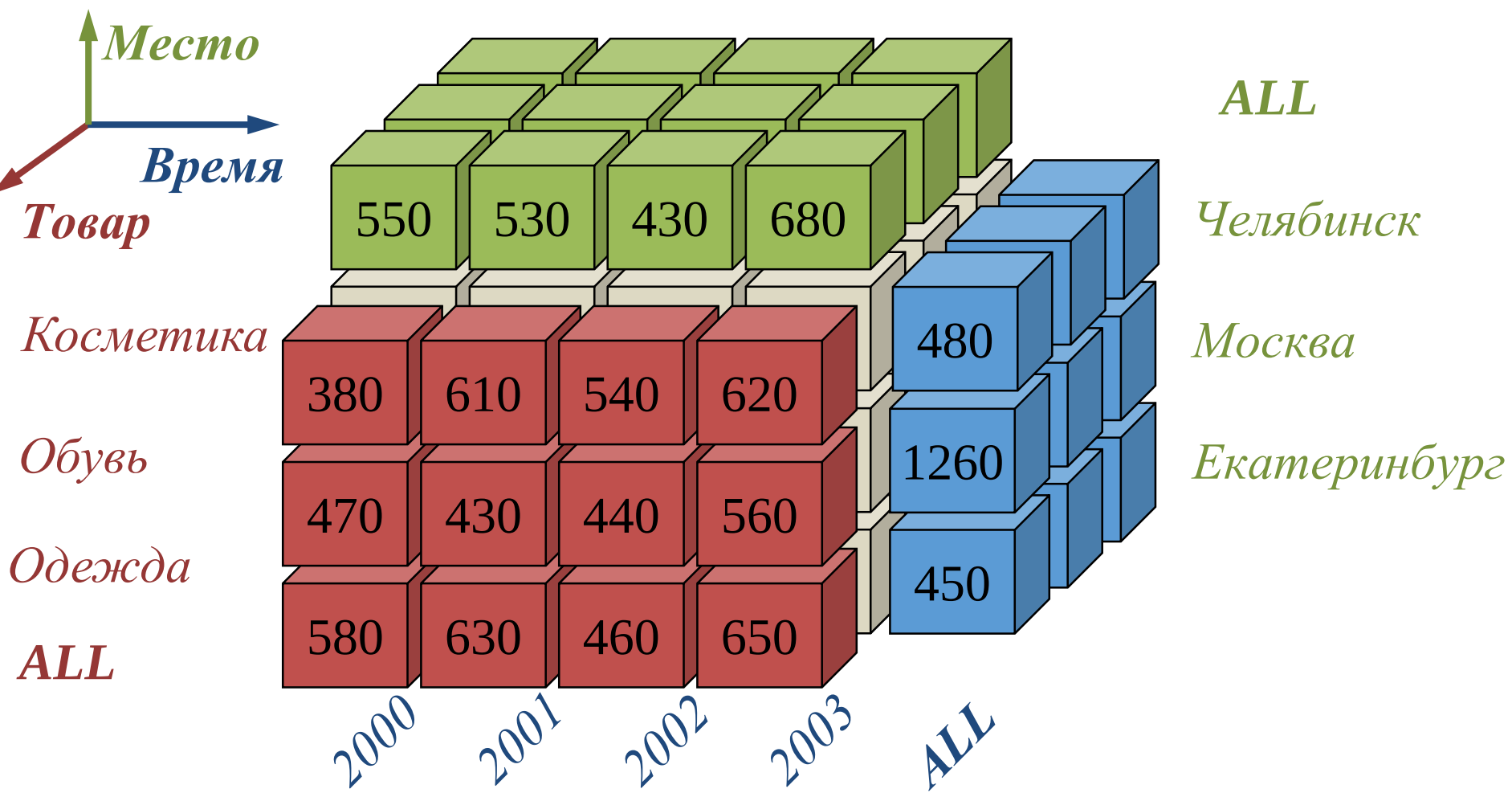
OLAP-куб



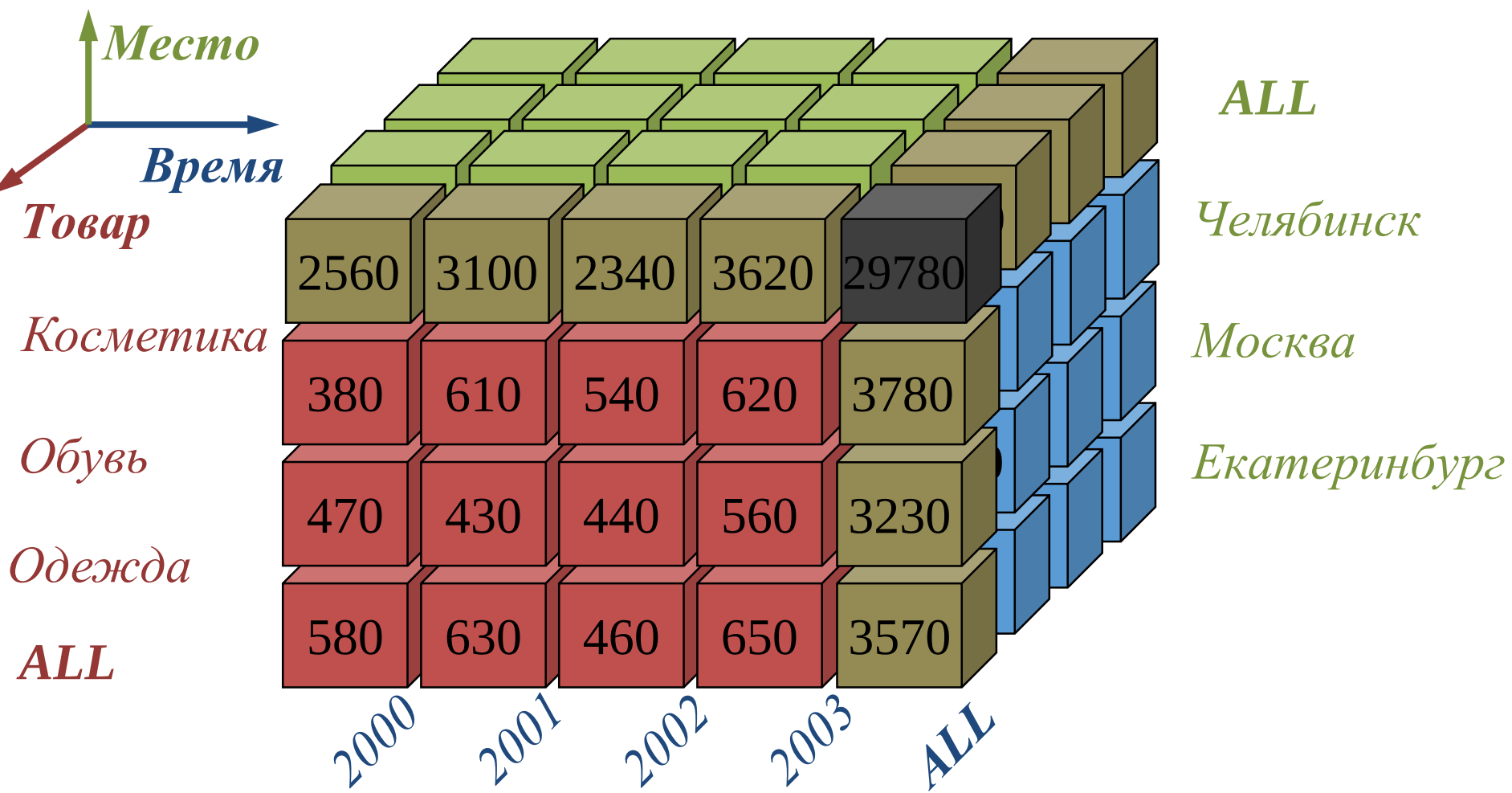
OLAP-куб



OLAP-куб



OLAP-куб



OLAP: многомерный запрос

Время	Место	Товар	Сумма
2000	Челябинск	Одежда	100
2000	Челябинск	Косметика	120
2000	Москва	Одежда	250
2000	Москва	Косметика	75
2001	Челябинск	Одежда	230
2001	Челябинск	Косметика	310
2001	Москва	Одежда	170
2001	Москва	Косметика	350

```
select
  Время, Место, Товар,
  sum(Сумма) as Прибыль
from Продажи
cube by (Время,
        Место, Товар)
```

OLAP: многомерный запрос

Время	Место	Товар	Сумма
2000	Челябинск	Одежда	100
2000	Челябинск	Косметика	120
2000	Москва	Одежда	250
2000	Москва	Косметика	75
2001	Челябинск	Одежда	230
2001	Челябинск	Косметика	310
2001	Москва	Одежда	170
2001	Москва	Косметика	350

select
 Время, Место, Товар,
 sum(Сумма) as Прибыль
from Продажи
cube by (Время,
 Место, Товар)

Время	Место	Товар	Прибыль
2000	Челябинск	Одежда	100
2000	Челябинск	Косметика	120
2000	Челябинск	[NULL]	220
2000	Москва	Одежда	250
2000	Москва	Косметика	75
2000	Москва	[NULL]	325
2000	[NULL]	Одежда	350
2000	[NULL]	Косметика	195
2000	[NULL]	[NULL]	545
2001	Челябинск	Одежда	230
2001	Челябинск	Косметика	310
2001	Челябинск	[NULL]	540
2001	Москва	Одежда	170
2001	Москва	Косметика	350
2001	Москва	[NULL]	520

OLAP: многомерный запрос

Время	Место	Товар	Сумма
2000	Челябинск	Одежда	100
2000	Челябинск	Косметика	120
2000	Москва	Одежда	250
2000	Москва	Косметика	75
2001	Челябинск	Одежда	230
2001	Челябинск	Косметика	310
2001	Москва	Одежда	170
2001	Москва	Косметика	350

select
 Время, Место, Товар,
 sum(Сумма) as Прибыль
from Продажи
cube by (Время,
 Место, Товар)

Время	Место	Товар	Прибыль
[NULL]	Челябинск	Одежда	330
[NULL]	Челябинск	Косметика	430
[NULL]	Челябинск	[NULL]	760
[NULL]	Москва	Одежда	420
[NULL]	Москва	Косметика	425
[NULL]	Москва	[NULL]	845
[NULL]	[NULL]	Одежда	750
[NULL]	[NULL]	Косметика	855
[NULL]	[NULL]	[NULL]	1 605

OLTP

Accounts

AcctNo	Balance	...
000374	25000	
...		
123456	10000	
...		

Перевод
средств

Accounts

AcctNo	Balance	...
000374	20000	
...		
123456	15000	
...		

BEGIN TRANSACTION

```
update Accounts
set Balance=Balance-5000
where AccNo='000374';

if SQLState<>'00000' goto UNDO;

update Accounts
set Balance=Balance+5000
where AccNo='123456';

if SQLState<>'00000' goto UNDO;
```

COMMIT;

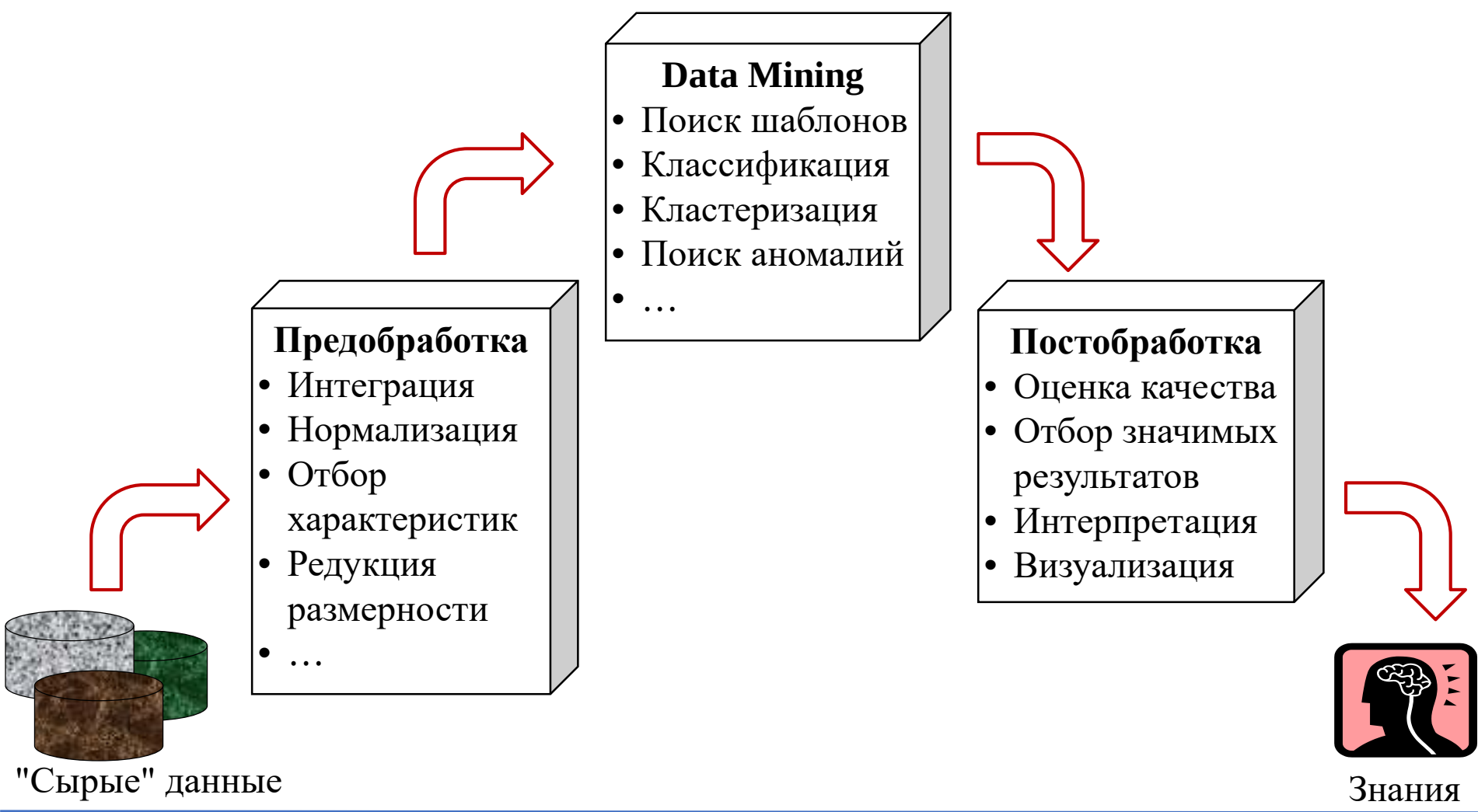
goto Finish;

Undo: **ROLLBACK;**

Finish: ;

END TRANSACTION

Процесс извлечения знаний (сообщество машинного обучения)



Базовые задачи анализа данных

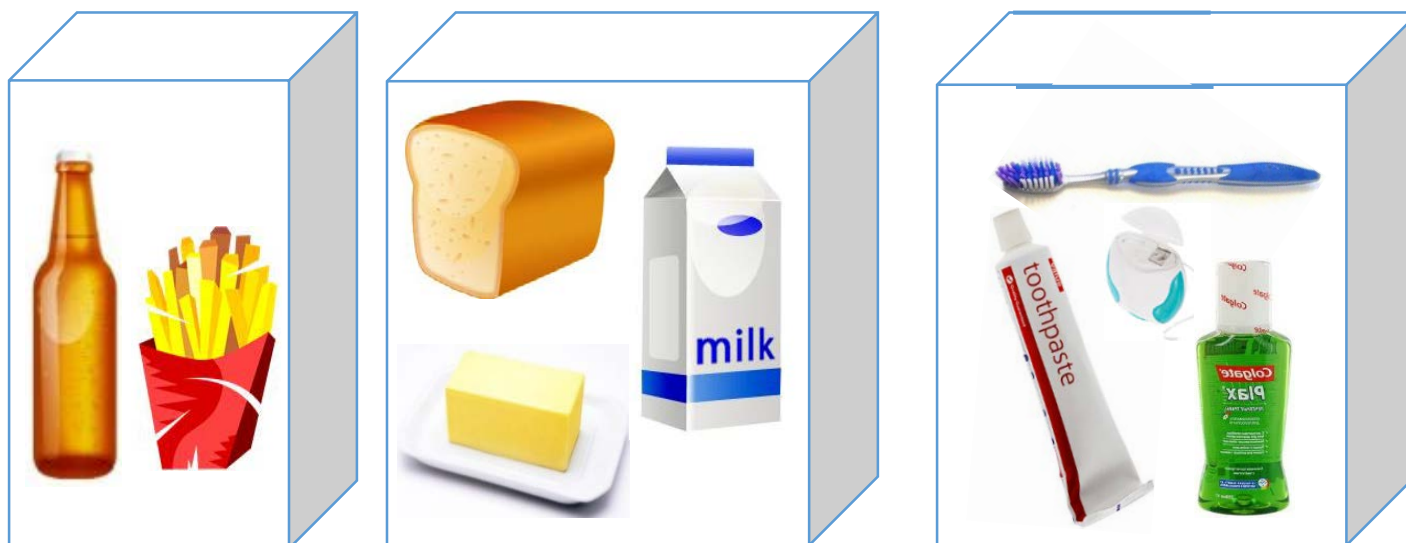
- Поиск шаблонов
 - Нахождение в данных о множестве объектов часто повторяющихся закономерностей
- Классификация
 - Распределение множества объектов одинаковой структуры по заранее известным группам (классам) в зависимости от похожести свойств объектов
- Кластеризация
 - Распределение множества объектов одинаковой структуры по заранее неизвестным группам (кластерам) в зависимости от похожести свойств объектов
- Поиск аномалий
 - Нахождение объектов, наиболее непохожих на другие объекты множества

Поиск шаблонов: анализ рыночной корзины



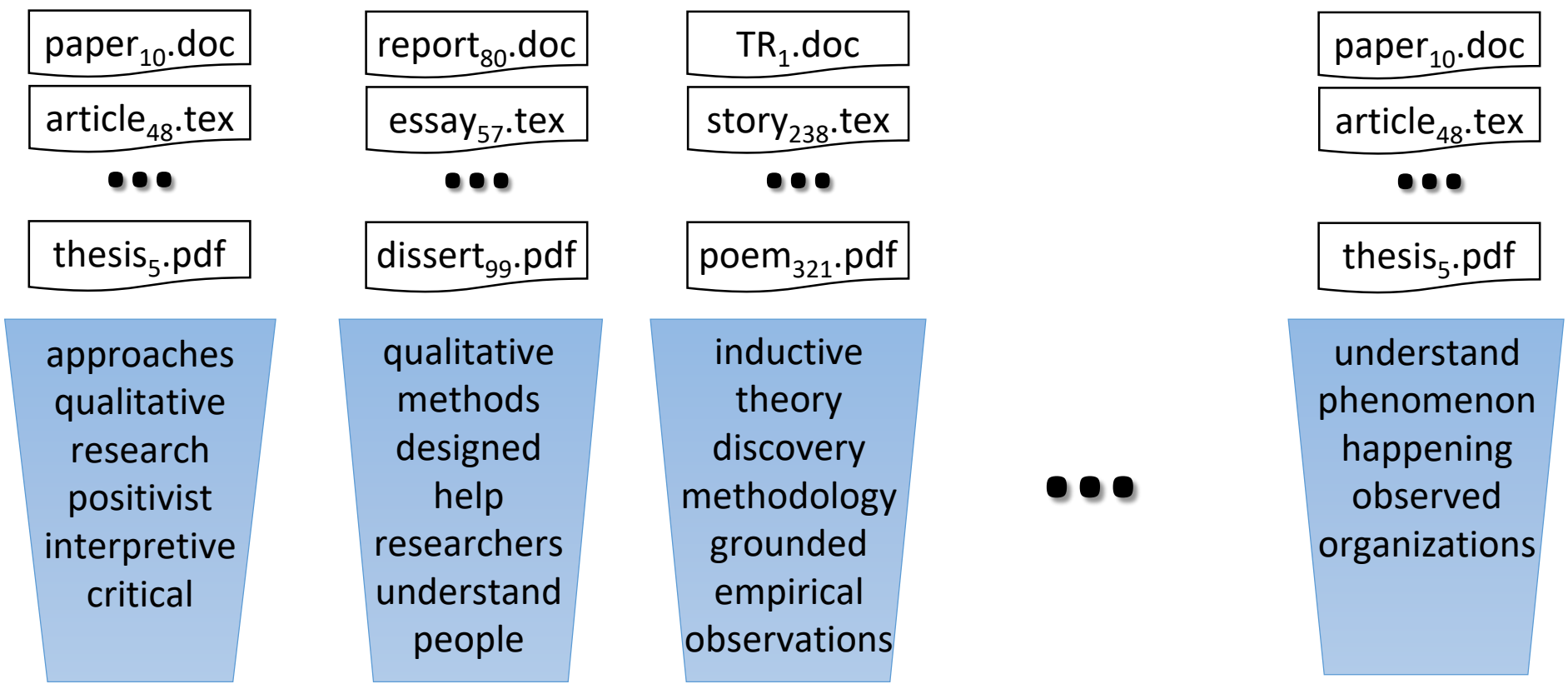
Какие наборы товаров в супермаркете
часто покупают совместно?

Поиск шаблонов: анализ рыночной корзины



- Мерчендайзинг, программы лояльности
- Ассоциативные правила
 - антисептик → {гречка, туалетная бумага}

Поиск шаблонов: плагиат



Документы, часто возникающие совместно,
вероятный плагиат

Поиск шаблонов: побочные эффекты лекарств



жар

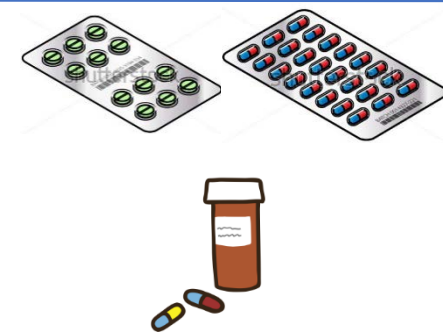




тошнота
рвота

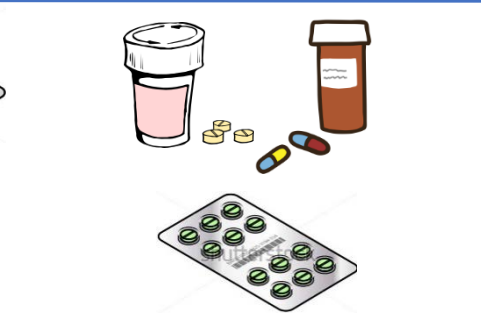


...



тошнота
жар





тошнота
жар
сыпь



Найти сочетания лекарств,
дающие заданный побочный эффект

Примеры приложений поиска шаблонов

- Анализ рыночной корзины
 - правила для продвижения, расстановки и хранения товаров
- Диагностика телекоммуникационного оборудования
 - правила для поиска комбинаций сигналов тревоги, которые возникают часто вместе в один и тот же период времени
- Медицинская информатика
 - правила для поиска сочетания симптомов пациента и результатов тестов, связанных с определенными заболеваниями

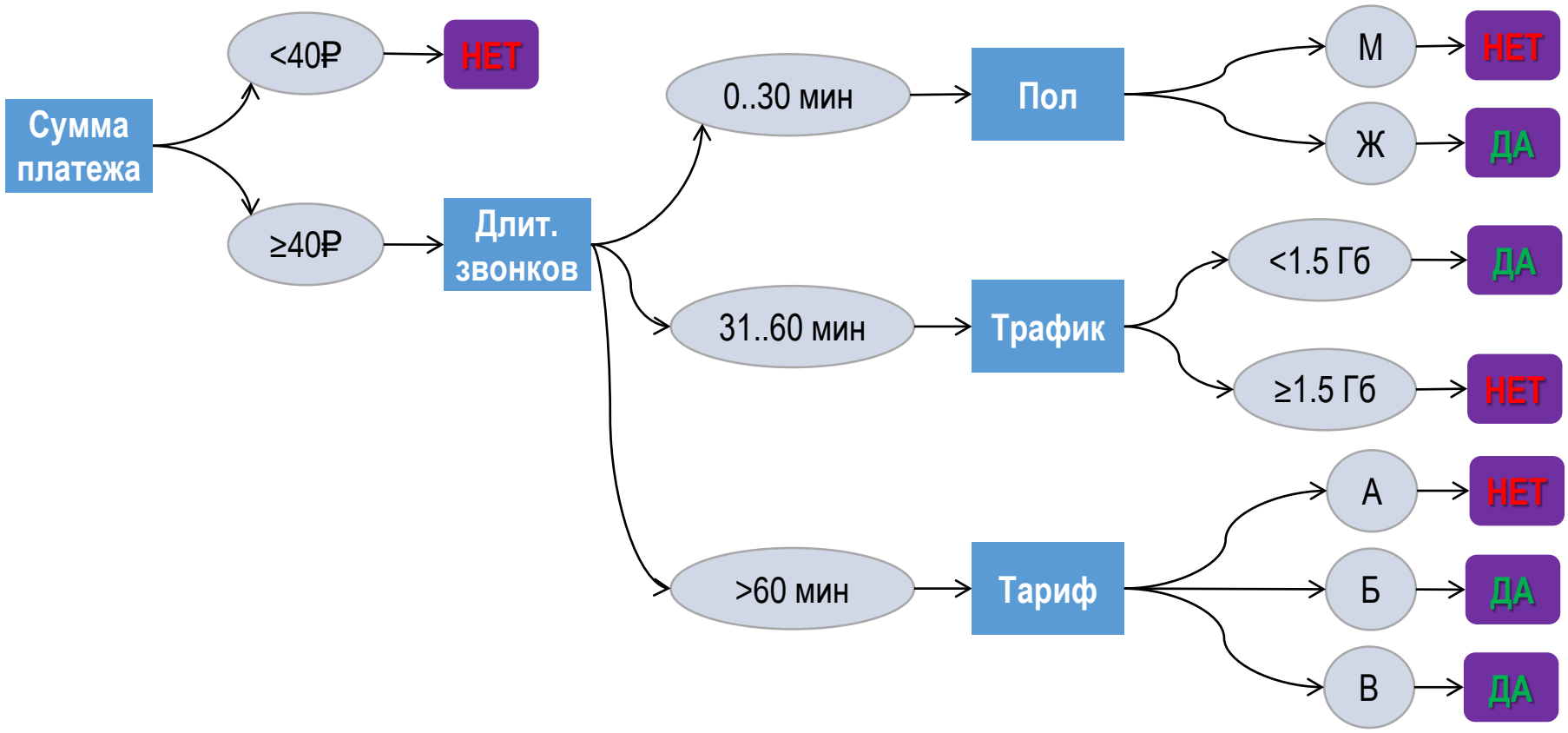
Классификация: анализ оттока клиентов



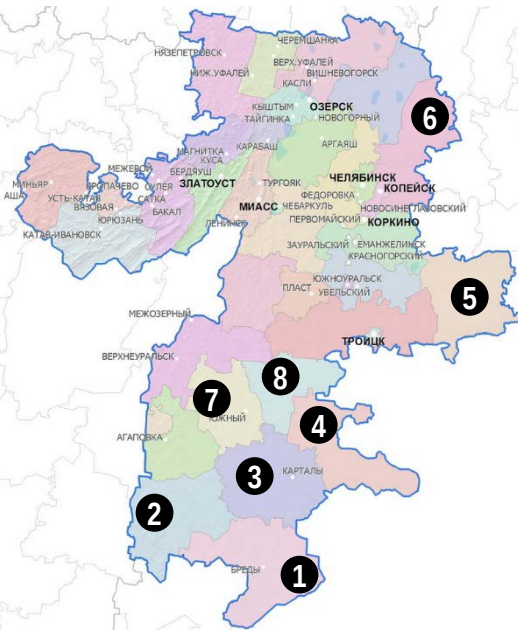
№	Пол	Возраст	К-во звонков	Длит. звонков	Интернет трафик	Тариф	Сумма платежа	...	ЛОЯЛЬНОСТЬ
1	М	24	103	146	1.1	А	50	...	НЕТ
2	М	28	217	307	2.7	А	60	...	ДА
3	Ж	29	308	965	3.5	В	30	...	НЕТ
4	Ж	33	92	180	3.1	Б	100	...	НЕТ
5	М	53	78	279	1.2	В	80	...	ДА
6	М	37	144	268	1.4	Б	40	...	НЕТ
7	Ж	38	65	201	0.9	А	100	...	ДА
8	М	43	48	158	0.6	А	60	...	ДА
9	М	46	89	223	0.3	В	35	...	ДА
10	Ж	68	19	117	0	В	25	...	НЕТ

Определить, уйдет ли клиент к другому оператору

Классификация: анализ оттока клиентов



Классификация: поиск полезных ископаемых



#	FeSO ₄ • 7H ₂ O		NH ₄ H ₂ PO ₄		(Na,Ca) (Si,Al) ₄ O ₈		SiO ₂ •nH ₂ O		LiAl Si ₄ O ₁₀		...	Мине- рал
	Нали чие	Плот ность	Нали чие	Плот ность	Нали чие	Плот ность	Нали чие	Плот ность	Нали чие	Плот ность		
1	Да	3.40	Нет	-	Да	7.80	Нет	-	Да	23.92		Железо
2	Нет	-	Да	7.22	Да	2.97	Да	5.97	Да	16.54		Медь
3	Да	4.67	Да	5.45	Да	5.43	Да	8.95	Да	28.49		Серебро
4	Нет	-	Да	3.12	Нет	-	Да	9.12	Нет	-		Цинк
5	Да	2.78	Да	0.18	Нет	-	Нет	-	Да	25.02		Железо
6	Да	1.02	Нет	-	Нет	-	Да	1.23	Да	2.12		ПУСТО
7	Да	0.75	Нет	-	Нет	-	Да	3.10	Да	2.99		ПУСТО
8	Нет	-	Да	0.36	Да	2.08	Нет	-	Нет	-		ПУСТО
	...											

Примеры приложений классификации

- Выявление спама в электронной почте
- Выявление мошеннических транзакций кредитной карты
- Рубрикация текста в новостной ленте (финансы, погода, развлечения, спорт и др.)
- Прогнозирование оттока клиентов провайдера услуг
- Определение вида поверхности по спутниковому снимку (водоем, городское поселение, лес и др.)
- Определение вида опухолевых клеток (доброкачественные или злокачественные)
- Определение вида вторичных структур белка (альфа-спираль, бета-лист или случайная спираль)
- Предсказание вида небесного тела (звезда или галактика) по снимкам телескопа

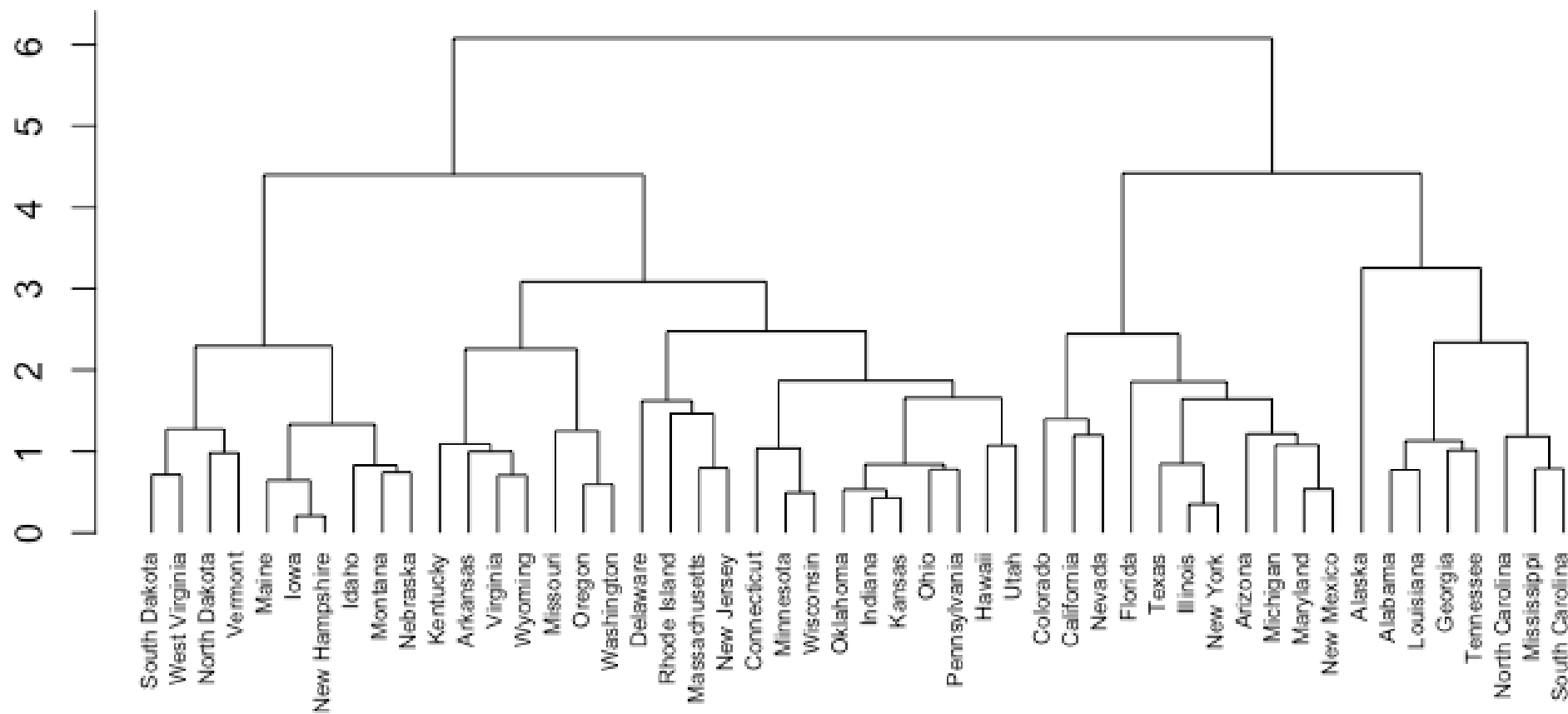
Кластеризация: разбиение



#	Sex	Age	Day calls	Day charge	Eve calls	Eve charge	Night calls	Night charge	Intl calls	Intl charge	Plan	...	GROUP
1	M	24	110	45.07	99	16.78	91	11.01	3	2.7	A		
2	M	28	123	27.47	103	16.62	103	11.45	3	3.7	B		
3	F	29	114	41.38	110	10.3	104	7.32	5	3.29	C		
4	M	33	71	50.9	88	5.26	89	8.86	7	1.78	A		
5	M	53	113	28.34	122	12.61	121	8.41	3	2.73	C		
6	M	37	98	37.98	101	18.75	118	9.18	6	1.7	A		
7	F	78	88	37.09	108	29.62	118	9.57	7	2.03	B		
8	M	63	79	26.69	94	8.76	96	9.53	6	1.92	B		
9	F	46	97	31.37	80	29.89	90	9.71	4	2.35	A		

- Определение групп клиентов, имеющих сходные характеристики

Кластеризация: иерархии



Сходство 50 штатов США относительно данных по арестам за различные правонарушения на 100 тыс. жителей в 1973 г.

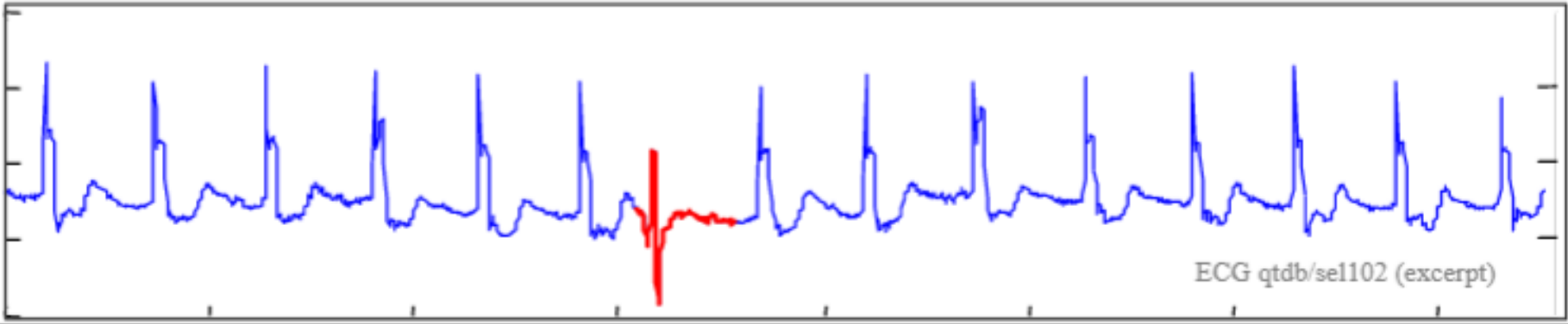
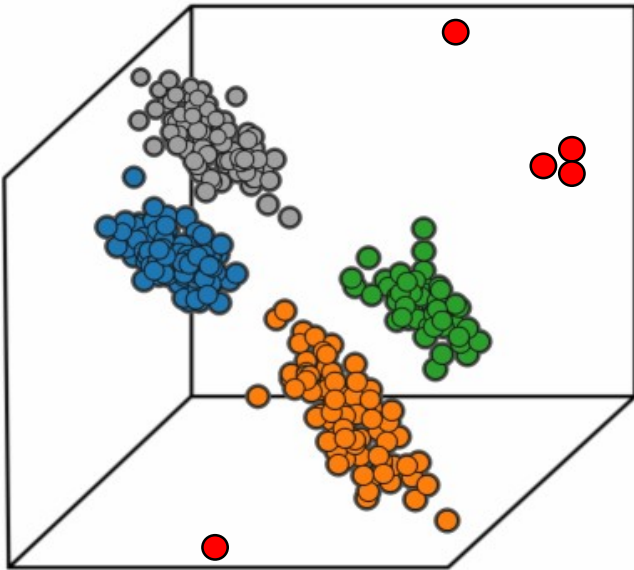
Примеры приложений кластеризации

- Биология: иерархическая систематика живых существ (царство, тип, класс, порядок, семейство, род и виды)
- Информационный поиск: группировка документов сходного содержания
- Землепользование: идентификация территорий с аналогичным землепользованием в базе данных наблюдения Земли
- Маркетинг: выявление групп клиентов для разработки целевых маркетинговых программ
- Градостроительство: группировка строений в соответствии с их типом, стоимостью и географическим положением
- Исследования землетрясений: группировка наблюдаемых эпицентров землетрясений вдоль разломов континентов
- Климат: понимание климата Земли, поиск закономерностей атмосферы и океана
- Предобработка анализируемых данных: редукция данных (центр кластера вместо объектов кластера), поиск ближайших соседей, поиск выбросов и др.

Поиск аномалий (выбросов)



ID	Time	Amount	Location	Item	...
123	2016-04-30	44.99	Moscow	Jacket	
124	2016-05-01	39.99	Moscow	Shirt	
...					
131	2016-05-02	49,999.99	Tokyo	Toyota LC	



Примеры приложений поиска аномалий (выбросов)

- Обнаружение мошенничества с кредитными картами
- Обнаружение сетевых вторжений
- Выявление аномального поведения сенсорных сетей для мониторинга и наблюдения
- Выявление изменений глобального лесного покрова

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2012. 740 p. ISBN 978-0123814791
 - 1. Introduction, pp. 1-38
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN-13: 978-0-13-312890-1
 - 1. Introduction, pp. 1-21