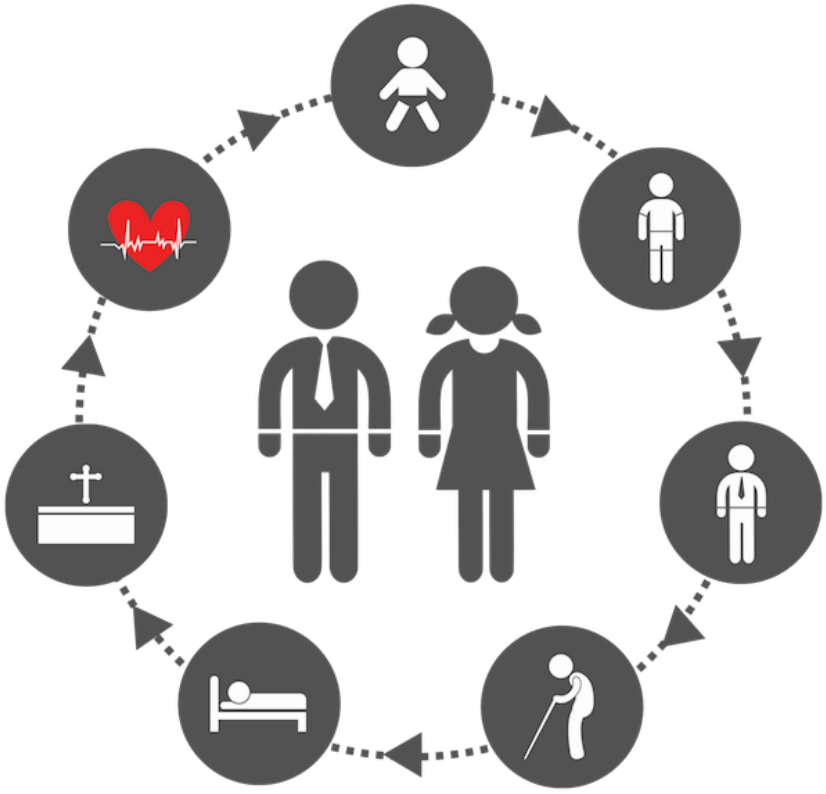
Технологический цикл аналитической обработки данных



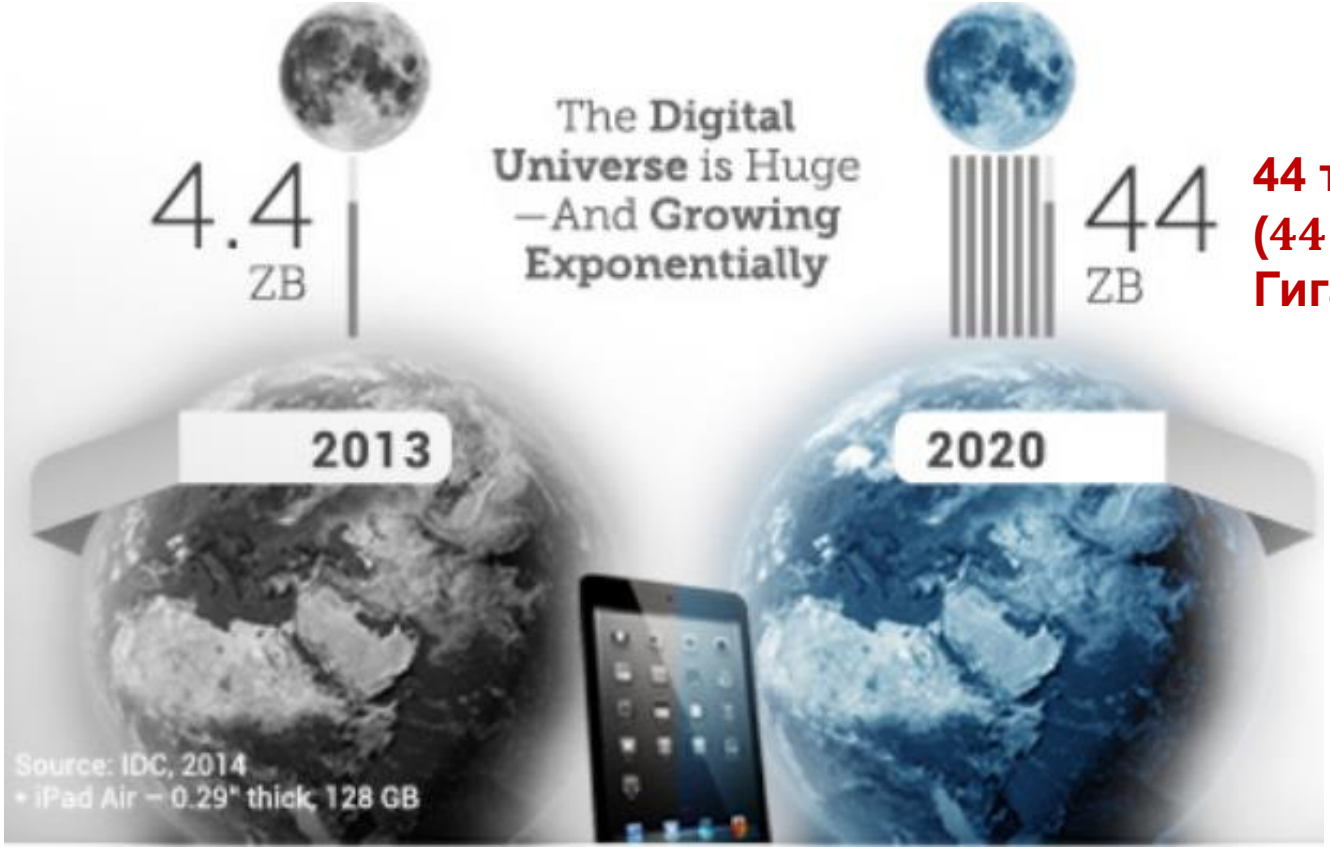
Анализ – это умение делать обоснованные выводы, а не придумывать их.

Н. Пащенко

Содержание

- Понятие Больших данных
- Технологический цикл аналитической обработки информации
  - Хранилище данных
  - Оперативная обработка информации (OLAP)
  - Интеллектуальный анализ данных (Data Mining)
- Основные задачи интеллектуального анализа данных
  - поиск шаблонов
  - классификация
  - кластеризация

Прирост данных в современном мире



44 триллиона ($44 \cdot 10^{12}$) Гигабайт

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: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

Новое отношение к данным

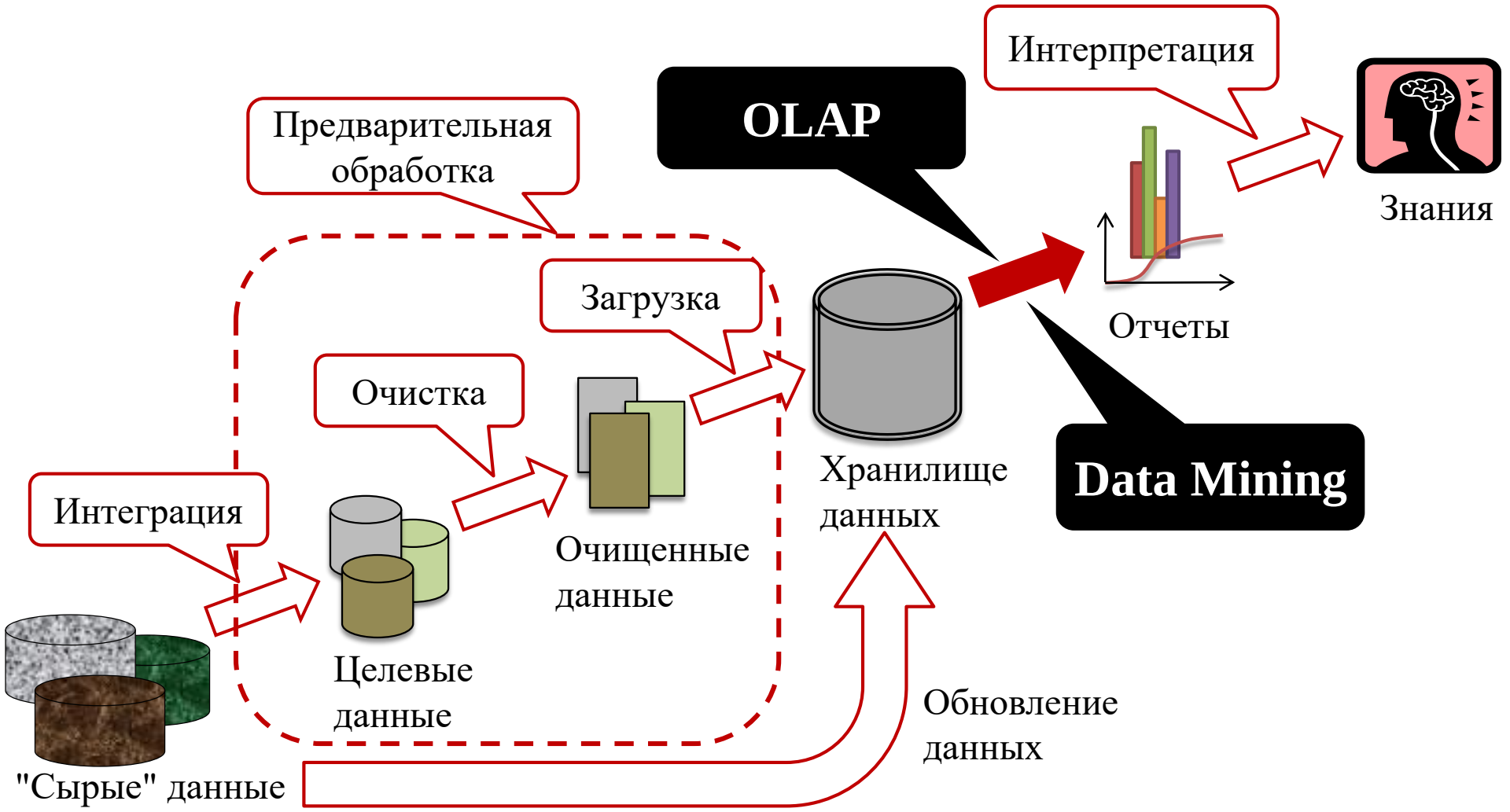
- Собирать любые возможные данные, когда и где это возможно
- Собранные данные будут иметь ценность либо для первоначальной, либо для непредвиденной цели

- Интеграция
  - комбинирование разл. источников данных
  - отбор релевантных источников данных
  - агрегация, конвертация в нужный формат
- Очистка
  - исправление ошибок, удаление шумов

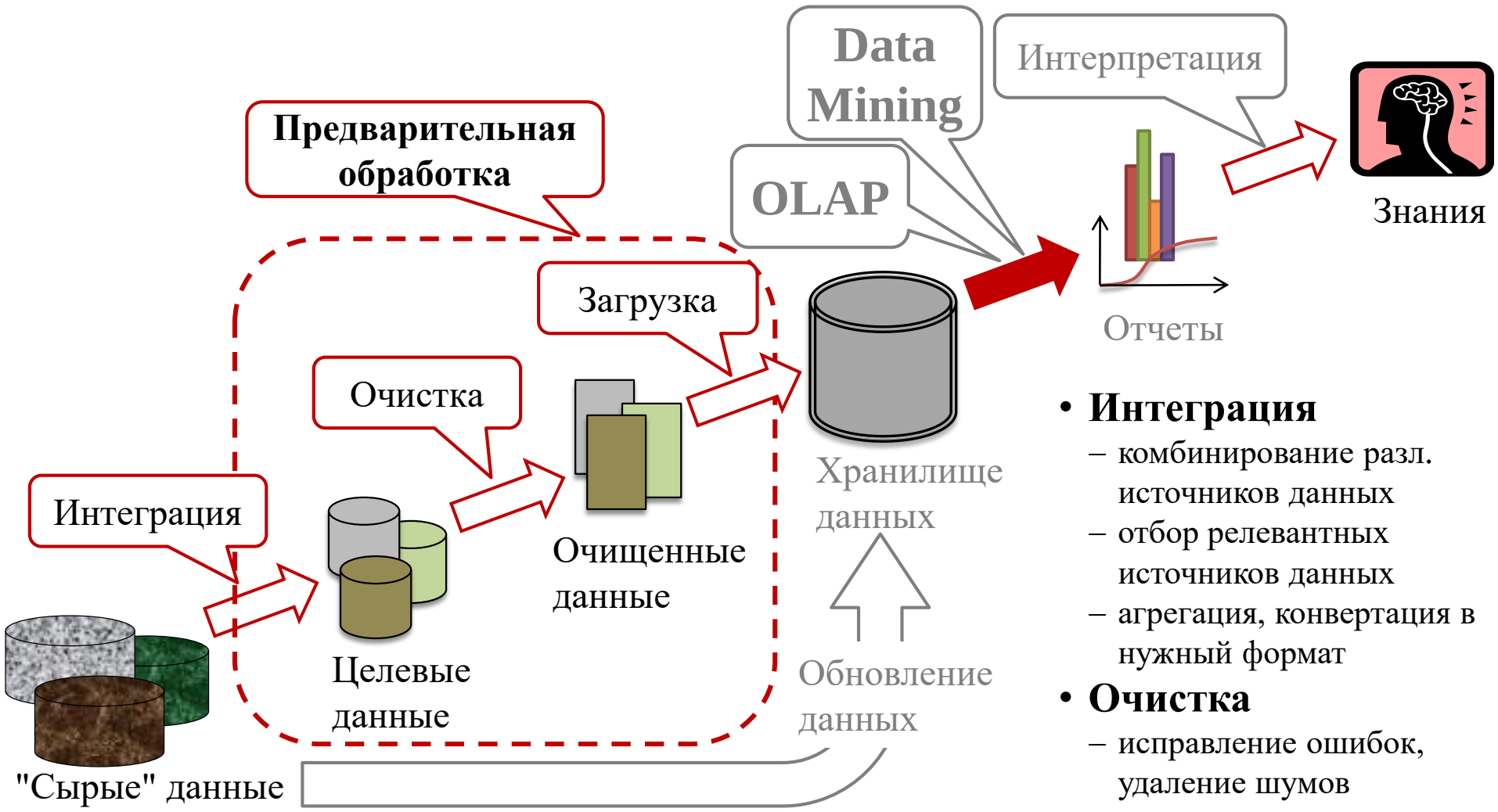
Большие данные (Big Data)

- Социально-экономический феномен, связанный с появлением технологий, позволяющих анализировать огромные массивы данных (вплоть до общемирового объема)
- Основные характеристики
  - **V**olume (размер): Тб, Пб,
  - **V**elocity (скорость прироста): онлайн
  - **V**ariety (разнообразие): аудио, видео, изображения текст, HTML, таблицы БД и др.

Цикл аналитической обработки информации



Цикл аналитической обработки информации

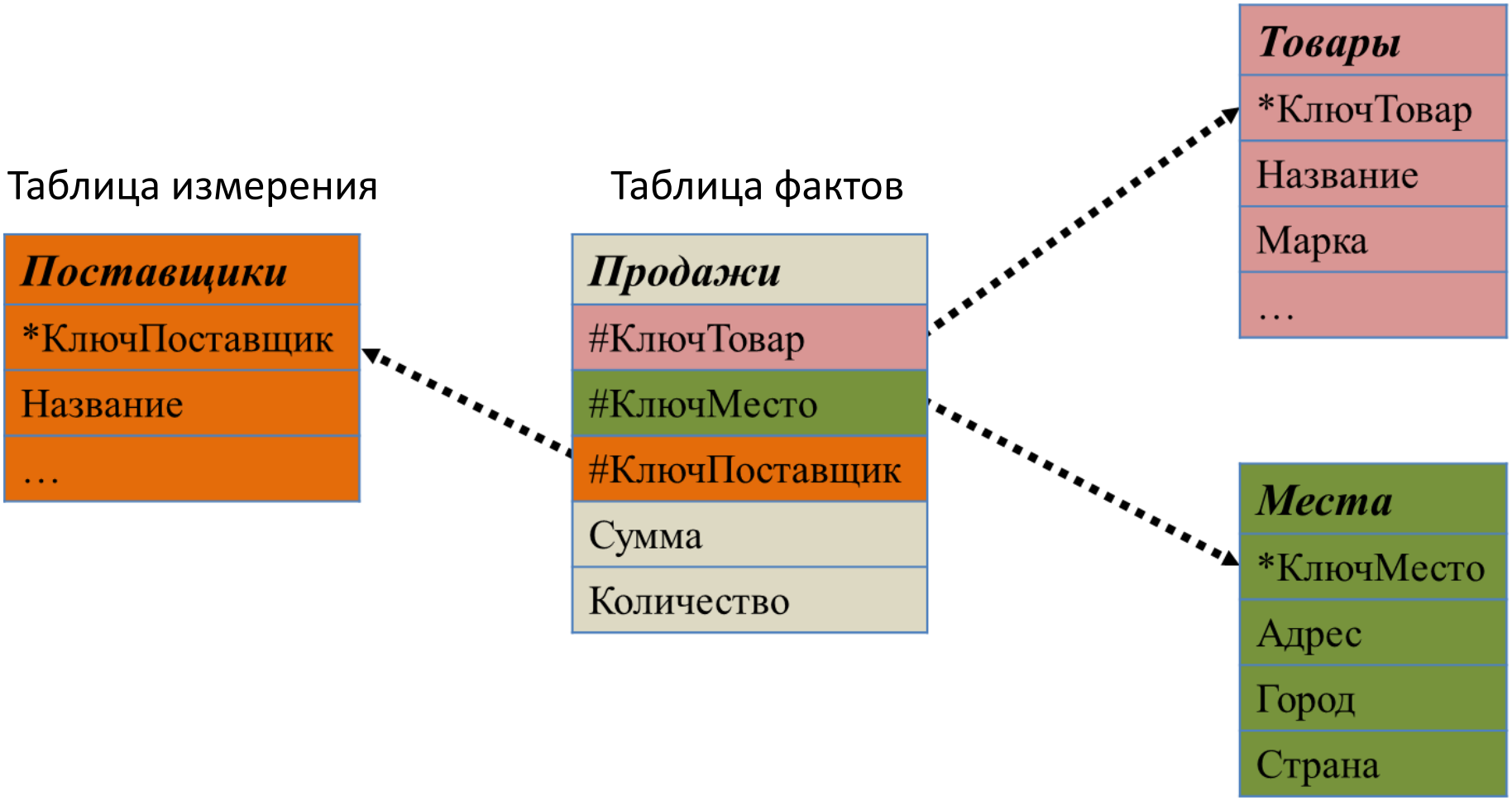


- **Интеграция**
  - комбинирование разл. источников данных
  - отбор релевантных источников данных
  - агрегация, конвертация в нужный формат
- **Очистка**
  - исправление ошибок, удаление шумов

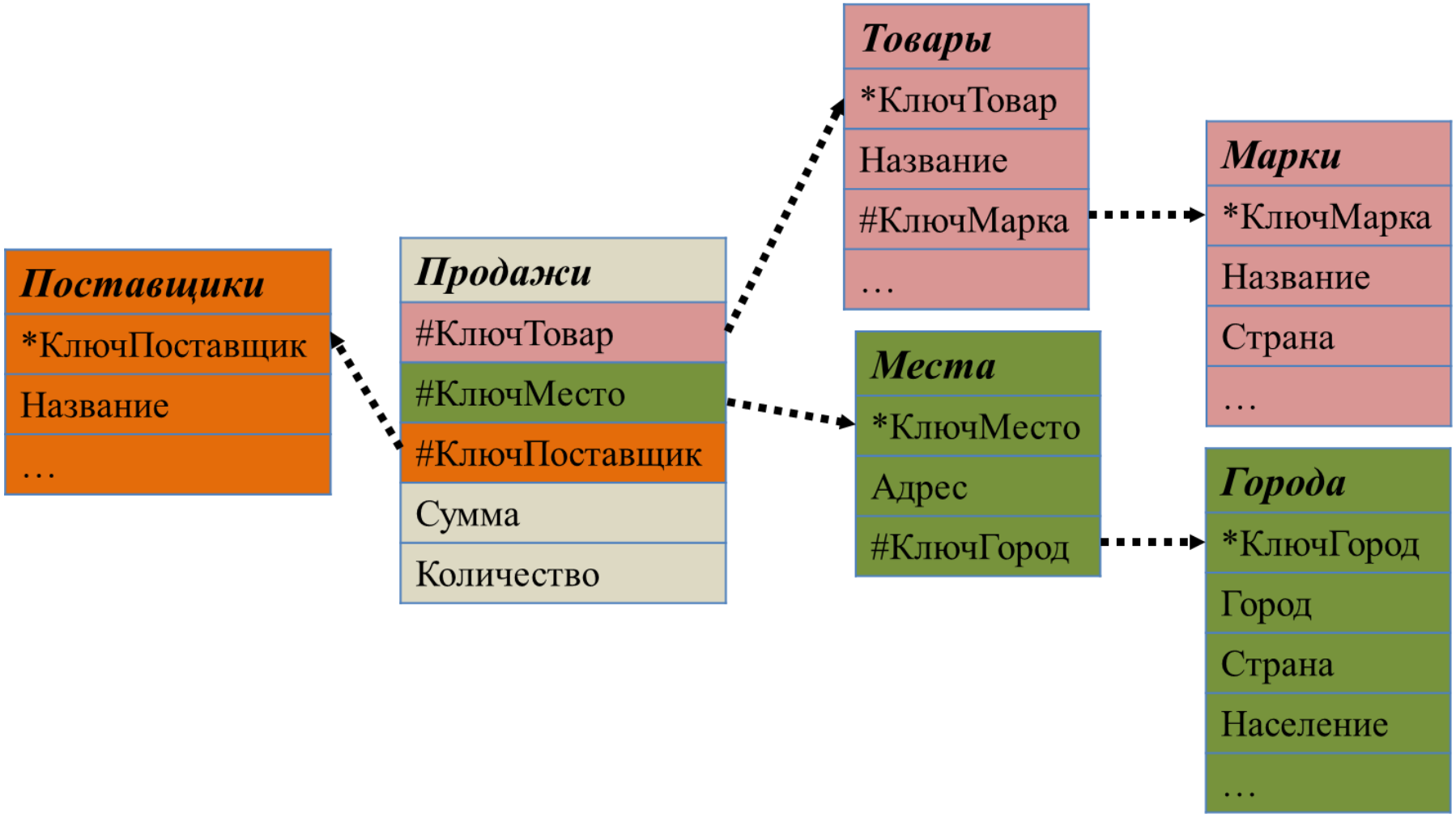
Хранилище данных

- Предметно-ориентированное хранилище для целей аналитики
  - физическое отделение от источников (баз) данных
  - интеграция информации из источников (баз) данных
  - поддержка хронологии в данных
  - неизменчивость (только загрузка новых) данных

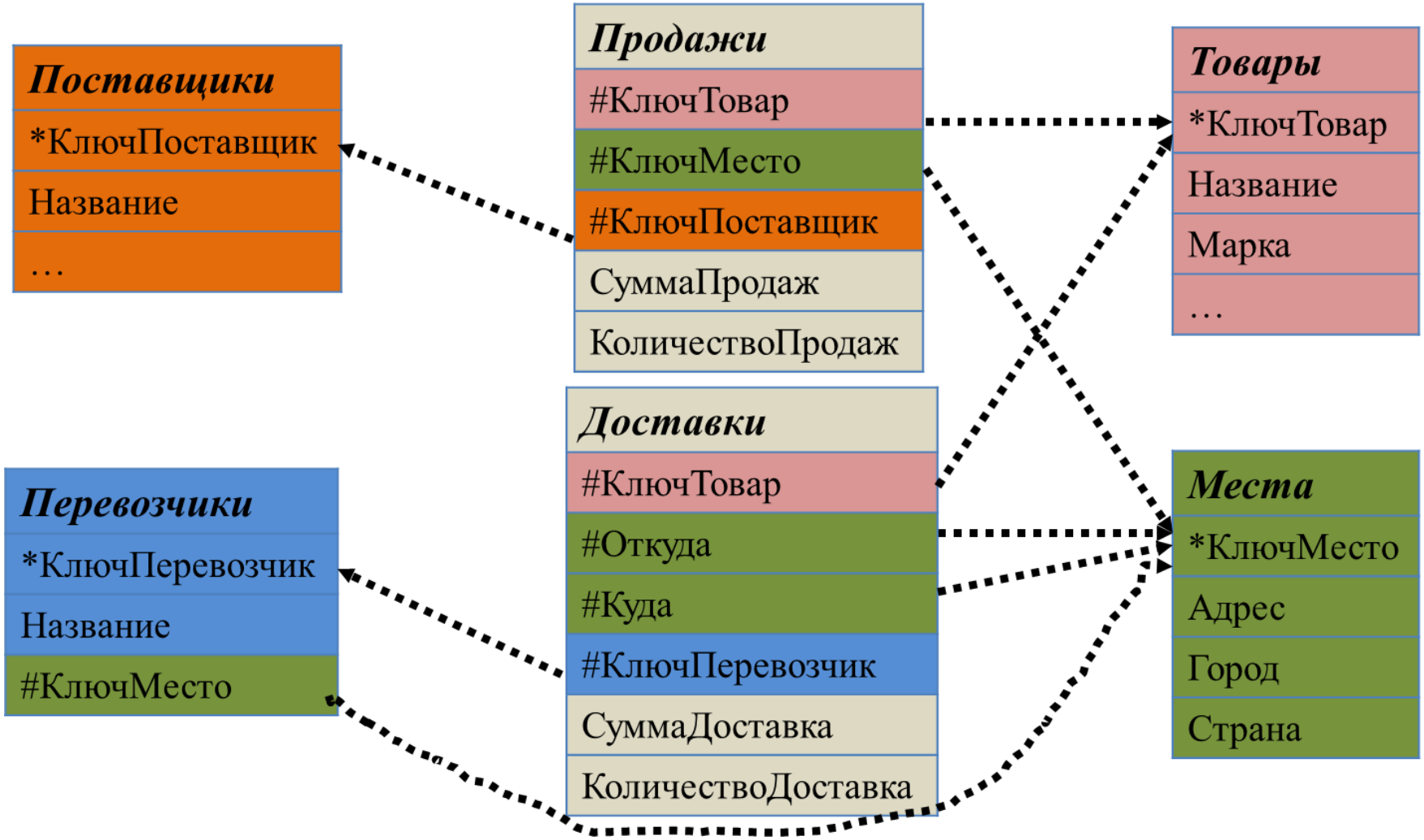
Хранилище данных: «звезда»



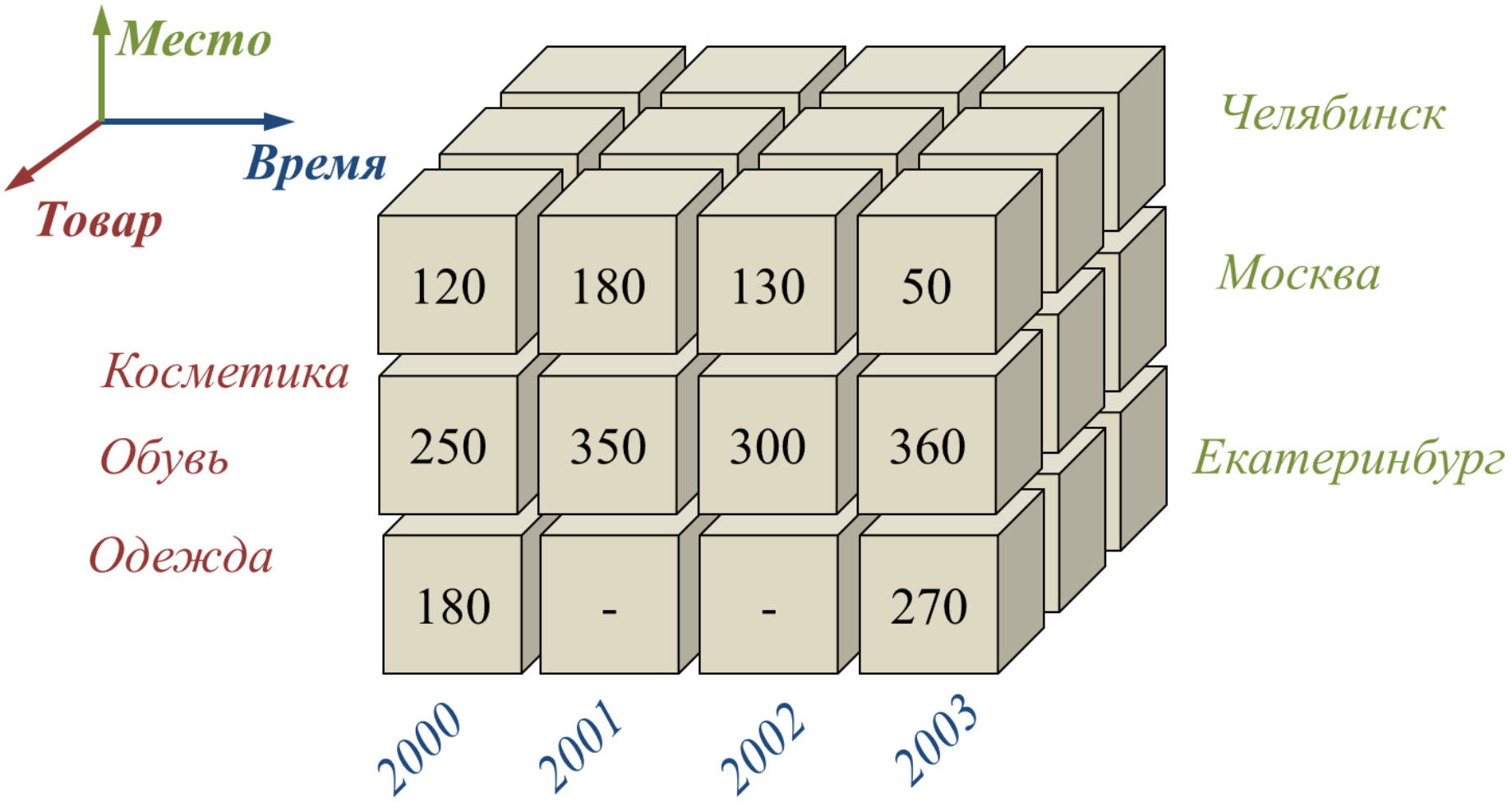
Хранилище данных: «снежинка»



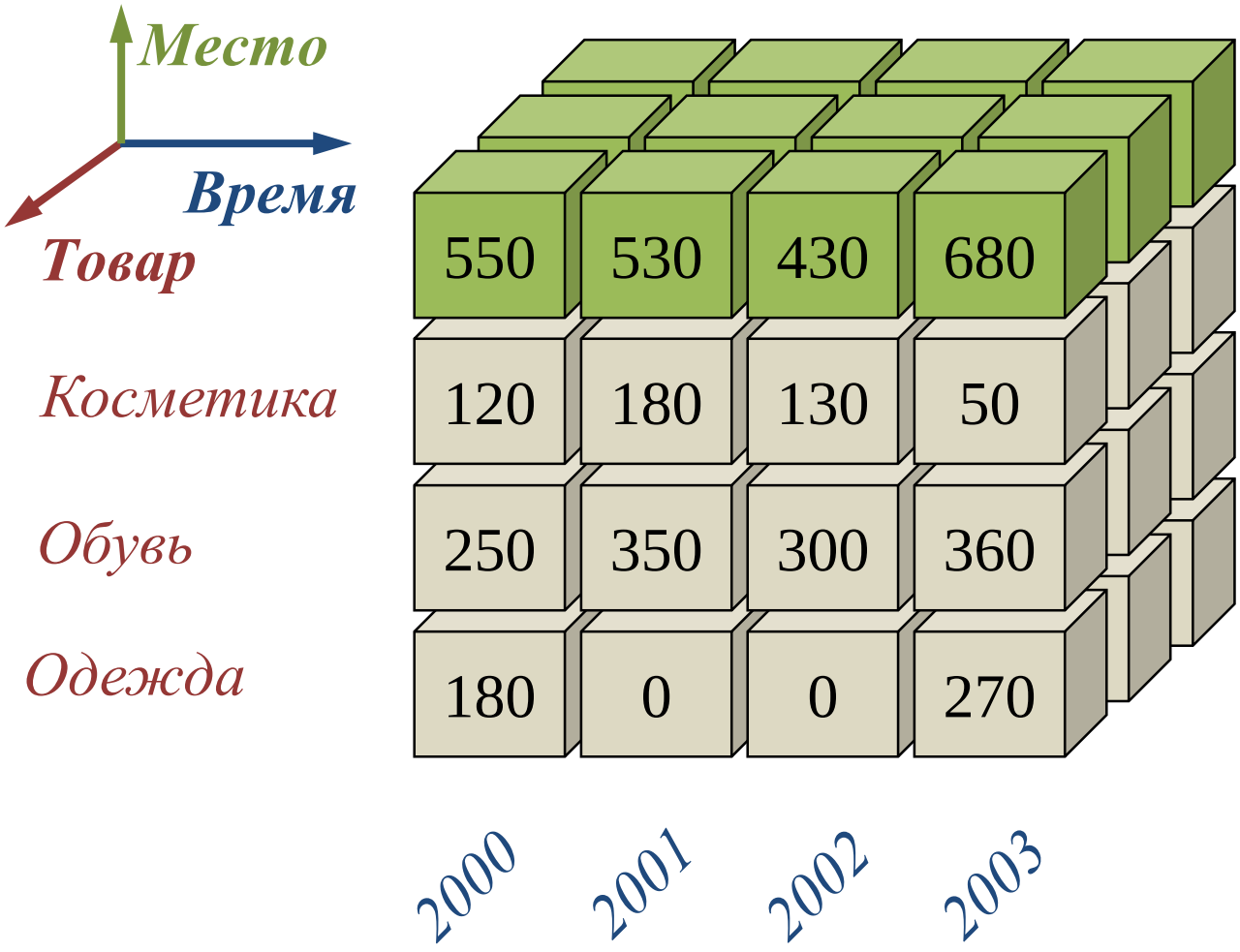
Хранилище данных: «созвездие»



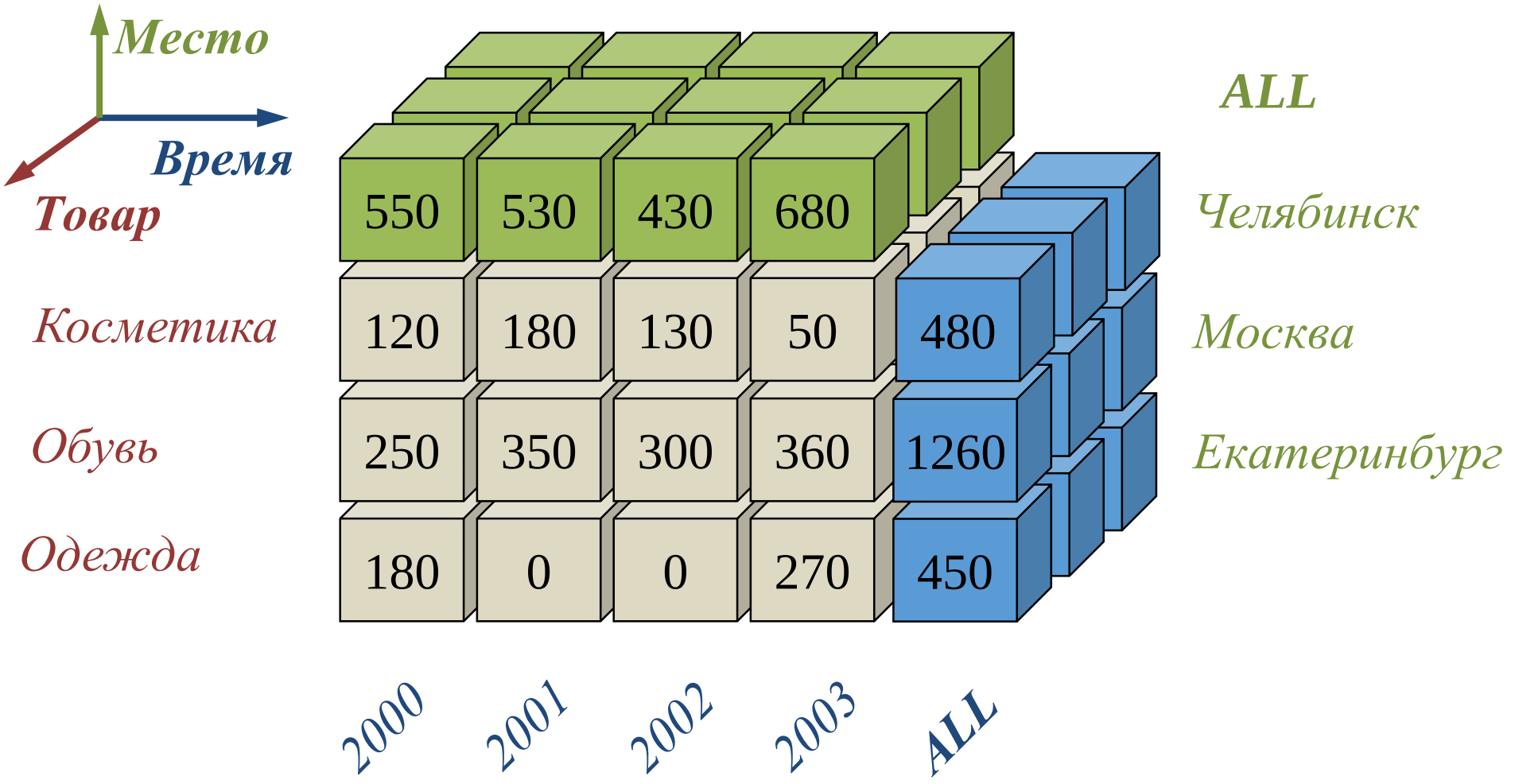
OLAP: многомерная модель данных



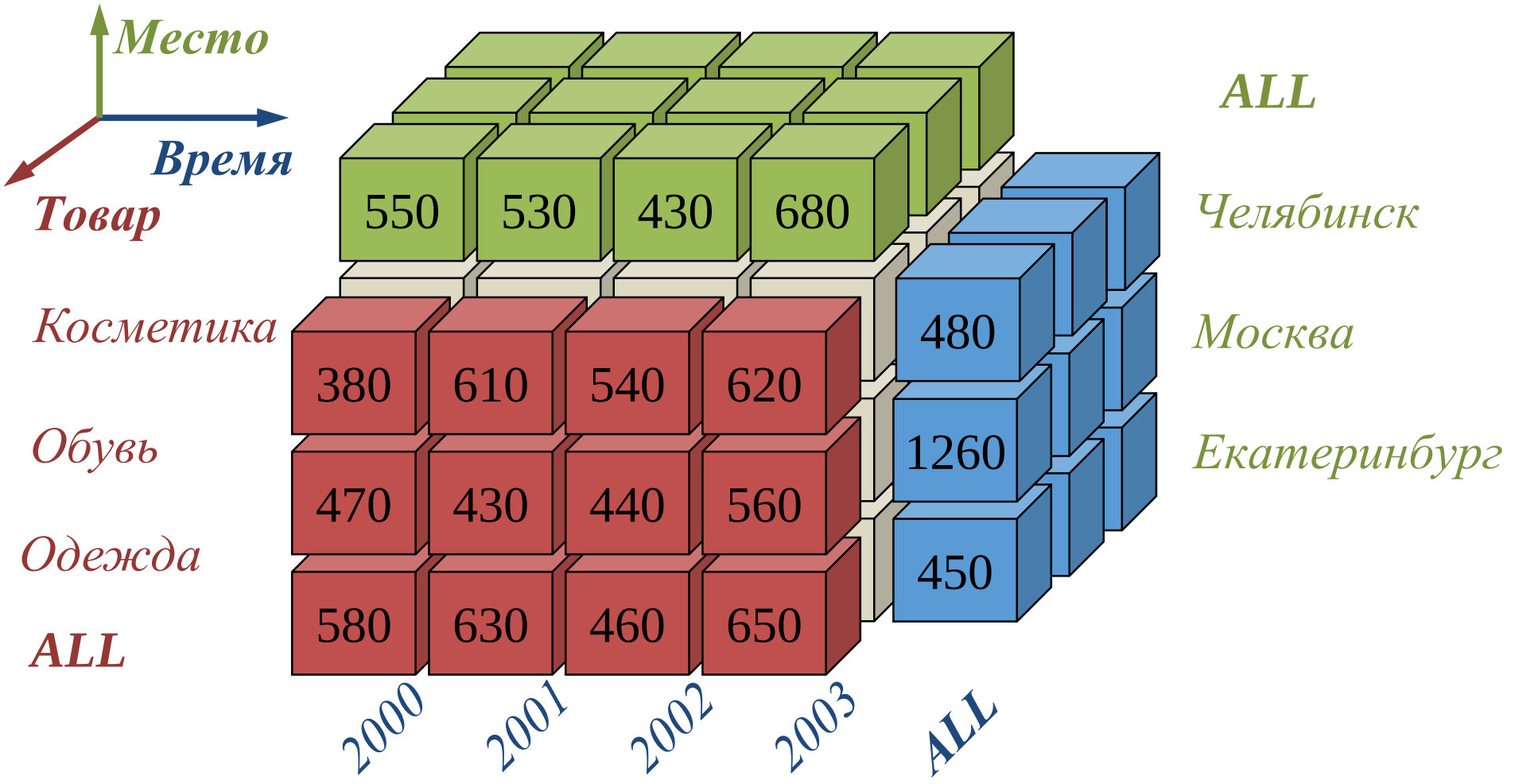
OLAP-куб


Место
Время
Товар
Косметика
Обувь
Одежда
550
530
430
680
120
180
130
50
250
350
300
360
180
0
0
270
2000
2001
2002
2003
ALL
Челябинск
Москва
Екатеринбург

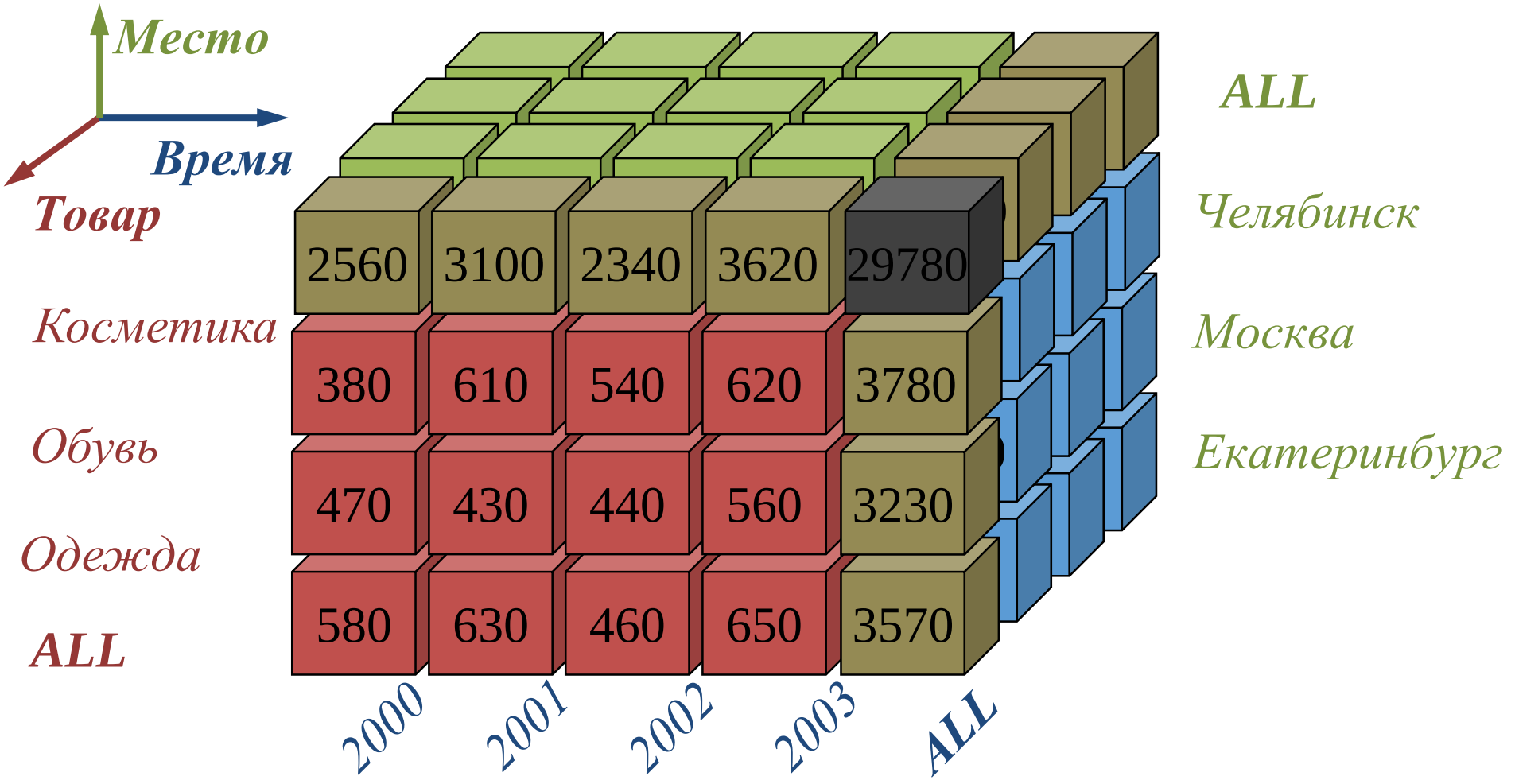
OLAP-куб


Место
Время
Товар
Косметика
Обувь
Одежда
550
530
430
680
120
180
130
50
480
250
350
300
360
1260
180
0
0
270
450
ALL
Челябинск
Москва
Екатеринбург
2000
2001
2002
2003
ALL

OLAP-куб


Место
Время
Товар
Косметика
Обувь
Одежда
ALL
550
530
430
680
380
610
540
620
470
430
440
560
580
630
460
650
480
1260
450
2000
2001
2002
2003
ALL
ALL
Челябинск
Москва
Екатеринбург

OLAP-куб


Место
Время
Товар
Косметика
Обувь
Одежда
ALL
2560
3100
2340
3620
29780
380
610
540
620
3780
470
430
440
560
3230
580
630
460
650
3570
2000
2001
2002
2003
ALL
ALL
Челябинск
Москва
Екатеринбург

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**

Интеллектуальный анализ данных (Data Mining)

Совокупность алгоритмов, методов и ПО для обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия стратегически важных решений в различных сферах человеческой деятельности

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.

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

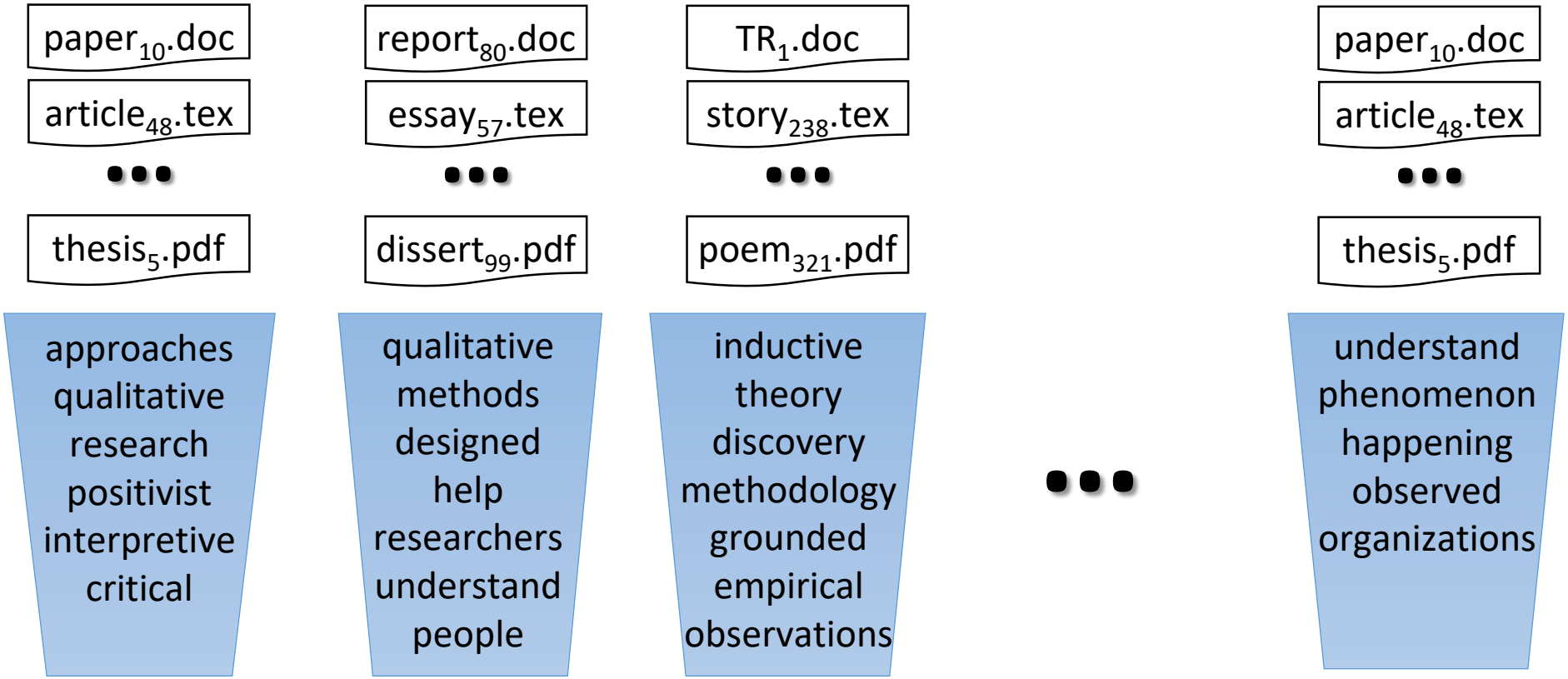
Какие наборы товаров в супермаркете часто покупают совместно?

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

Классификация: общий подход

Name	Income	Age	Credit rating
Peter Parker	low	youth	risky
Anakin Skywalker	low	youth	risky
Tony Stark	high	middle	safe
Han Solo	low	middle	risky
Clark Kent	low	senior	safe
James Bond	medium	senior	safe
Bruce Wayne	high	middle	safe

IF *age=youth*
THEN *rating:=risky*
IF *income=high*
THEN *rating:=safe*
IF *age=middle* AND
income=low
THEN *rating:=risky* ...

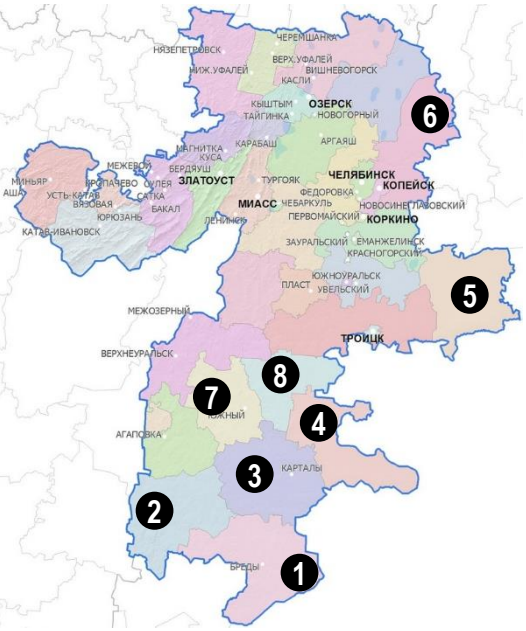
Name	Income	Age	Credit rating
John Doe	low	senior	?
Jane Doe	high	youth	?

Классификация: анализ оттока клиентов

#	Sex	Age	Day calls	Day charge	Eve calls	Eve charge	Night calls	Night charge	Intl calls	Intl charge	Plan	…	CHURN
1	M	24	110	45.07	99	16.78	91	11.01	3	2.7	A		YES
2	M	28	123	27.47	103	16.62	103	11.45	3	3.7	B		NO
3	F	29	114	41.38	110	10.3	104	7.32	5	3.29	C		YES
4	M	33	71	50.9	88	5.26	89	8.86	7	1.78	A		YES
5	M	53	113	28.34	122	12.61	121	8.41	3	2.73	C		NO
6	M	37	98	37.98	101	18.75	118	9.18	6	1.7	A		YES
7	F	78	88	37.09	108	29.62	118	9.57	7	2.03	B		NO
8	M	63	79	26.69	94	8.76	96	9.53	6	1.92	B		NO
9	F	46	97	31.37	80	29.89	90	9.71	4	2.35	A		NO
					…								

Определить, уйдет ли клиент к другому оператору

Классификация: поиск полезных ископаемых







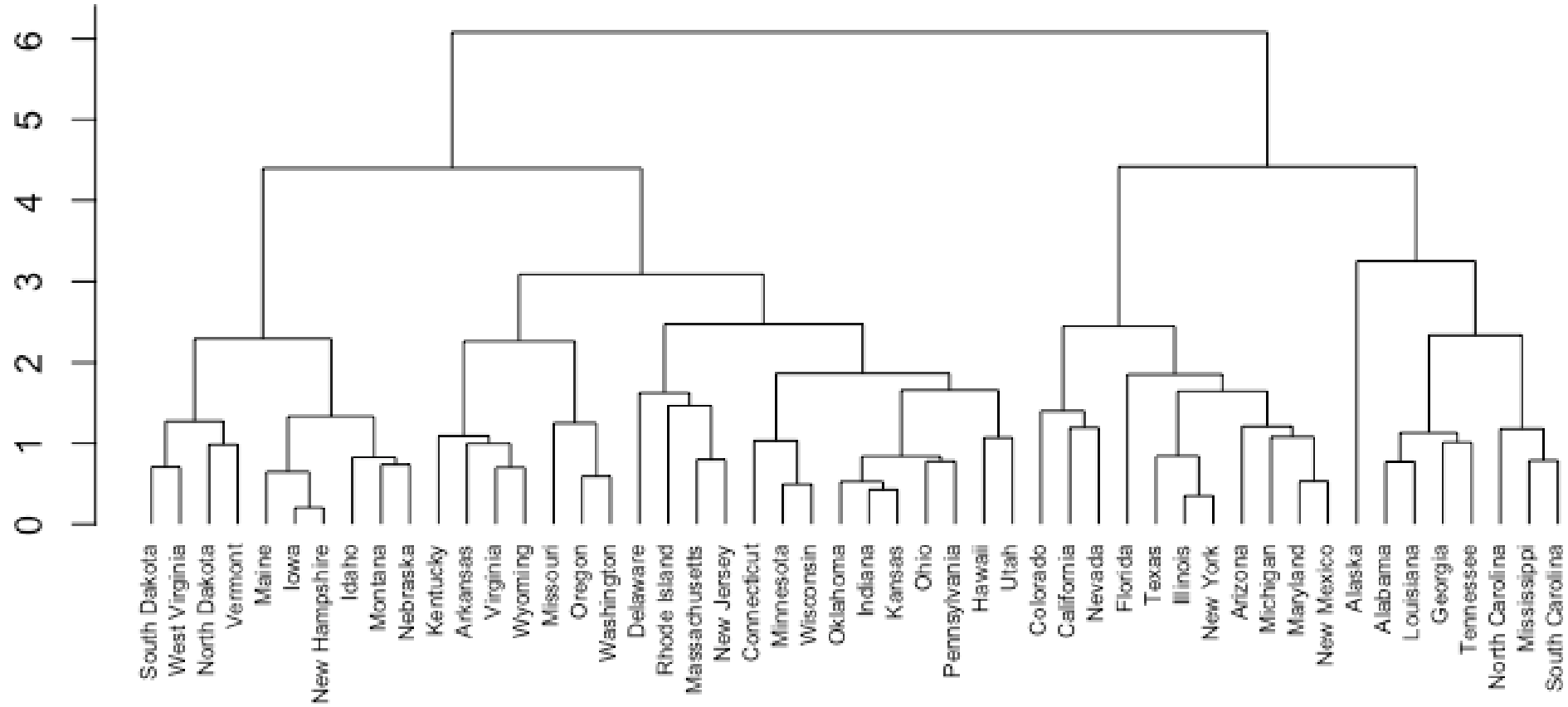
#	$FeSO_4 \bullet 7H_2O$		$NH_4H_2PO_4$		$(Na,Ca)(Si,Al)_4O_8$		$SiO_2 \bullet nH_2O$		$LiAlSi_4O_{10}$		…	Mineral
	Наличие	Плотность	Наличие	Плотность	Наличие	Плотность	Наличие	Плотность	Наличие	Плотность		
1	Yes	3.40	No	-	Yes	7.80	No	-	Yes	23.92		Iron
2	No	-	Yes	7.22	Yes	2.97	Yes	5.97	Yes	16.54		Copper
3	Yes	4.67	Yes	5.45	Yes	5.43	Yes	8.95	Yes	28.49		Silver
4	No	-	Yes	3.12	No	-	Yes	9.12	No	-		Zinc
5	Yes	2.78	Yes	0.18	No	-	No	-	Yes	25.02		Iron
6	Yes	1.02	No	-	No	-	Yes	1.23	Yes	2.12		NONE
7	Yes	0.75	No	-	No	-	Yes	3.10	Yes	2.99		NONE
8	No	-	Yes	0.36	Yes	2.08	No	-	No	-		NONE

Кластеризация: таргетирование

#	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 смысловых групп клиентов

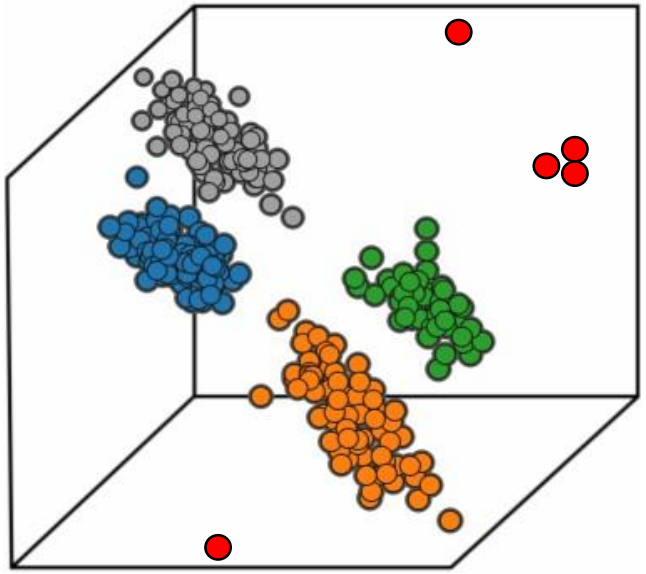
Кластеризация: иерархии

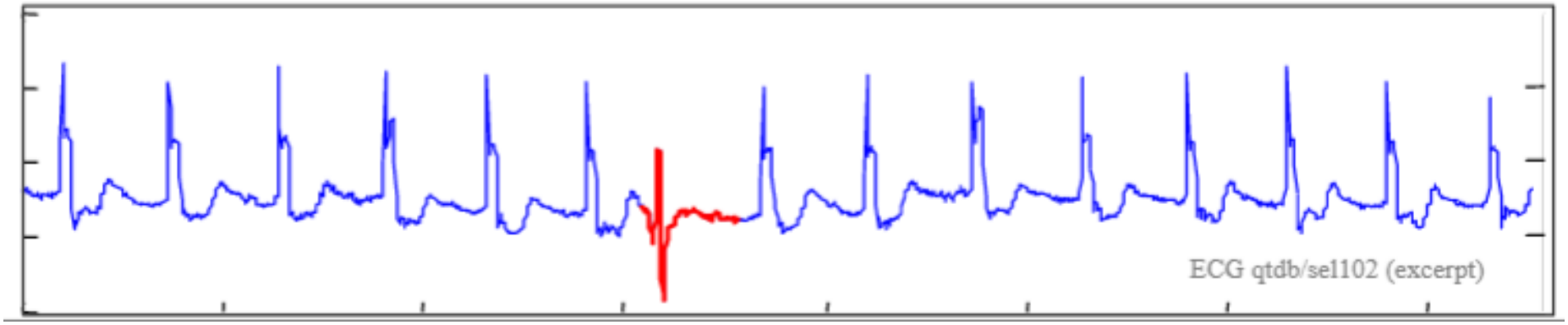


Сходство 50 штатов США относительно данных по арестам за различные правонарушения на 100 тыс. жителей в 1973 г.

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	





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