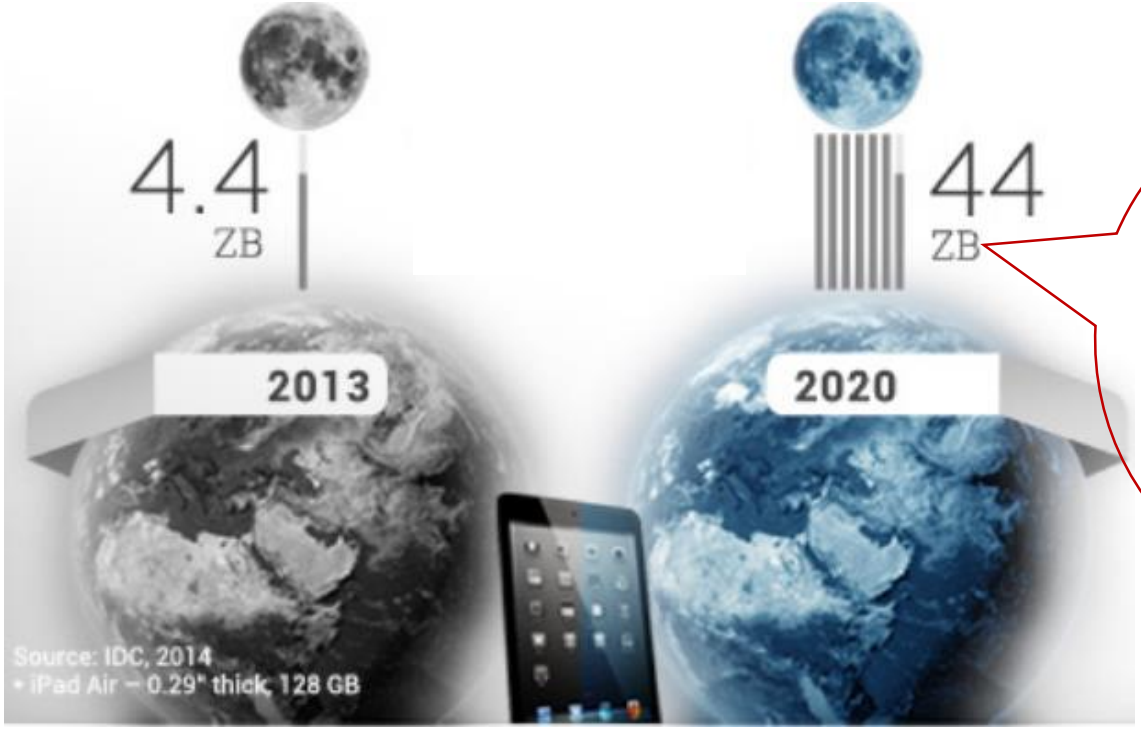
МАШИННОЕ ОБУЧЕНИЕ:

КАК НАЙТИ ТО, ЧТО СКРЫТО В ДАННЫХ?

Михаил Леонидович Цымблер,

доктор физ.-мат. наук,
профессор кафедры системного программирования,
Высшая школа электроники и компьютерных наук ЮУрГУ

Пока Вы читали этот текст, объем мировых данных увеличился на 10^5 Гигабайт



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

Новая реальность Больших данных

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

С. Плюшкин
1800(?) – ?

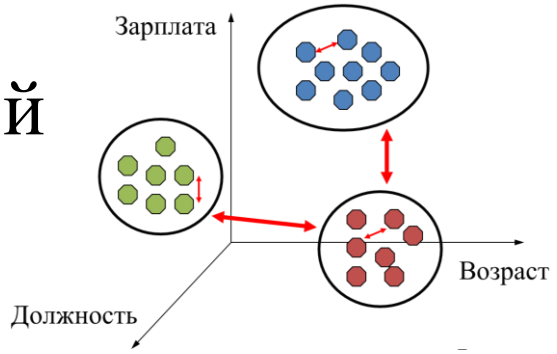
Машинное обучение (Machine Learning)

технология решения задач, в которой **компьютер обучается решать задачу за счет выявления закономерностей в данных**, а не выполняет прямое решение задачи по заданному алгоритму

Основные задачи машинного обучения

- ***Классификация***
  Распределение объектов с одинаковой структурой по разным группам, смысл которых известен заранее

- ***Кластеризация***
  Распределение объектов с одинаковой структурой по разным группам, смысл которых не известен заранее

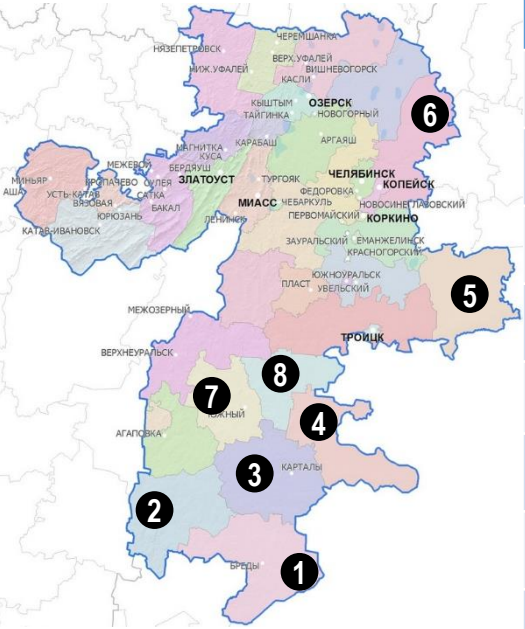


- ***Поиск шаблонов***
  Нахождение часто встречающихся зависимостей между объектами

Классификация: отток клиентов

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

Классификация: поиск полезных ископаемых







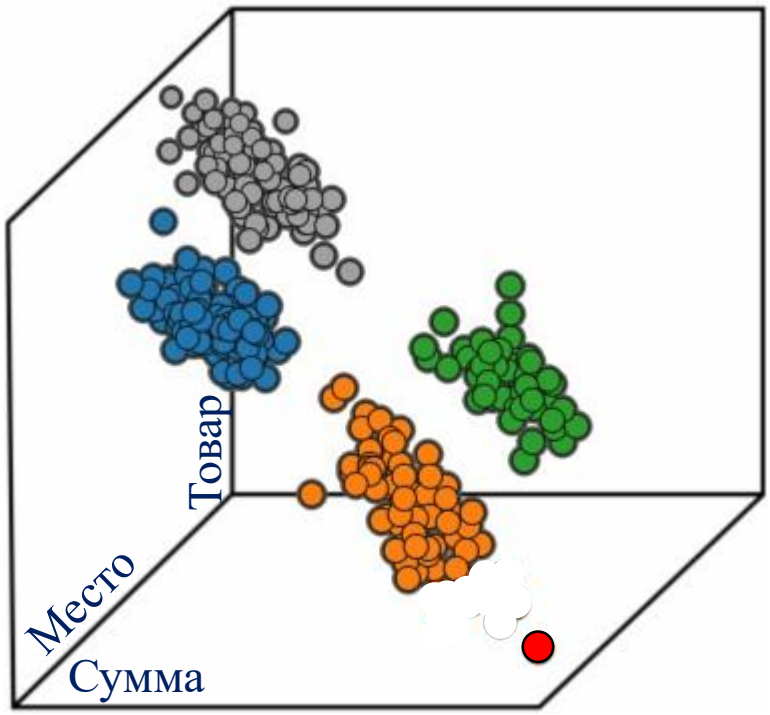
#	$FeSO_4 \bullet 7H_2O$		$NH_4H_2PO_4$		$(Na,Ca)(Si,Al)_4O_8$		$SiO_2 \bullet nH_2O$		$LiAlSi_4O_{10}$		…	Ископаемое
	Наличие	Плотность	Наличие	Плотность	Наличие	Плотность	Наличие	Плотность	Наличие	Плотность		
1	Да	3.40	Нет	-	Да	7.80	Нет	-	Да	23.92		Железо
2	Нет	-	Да	7.22	Да	2.97	Да	5.97	Да	16.54		Медь
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		НЕТ
7	Да	0.75	Нет	-	Нет	-	Да	3.10	Да	2.99		НЕТ
8	Нет	-	Да	0.36	Да	2.08	Нет	-	Нет	-		НЕТ

Кластеризация: выделение целевых групп

№	Пол	Возраст	Кол-во звонков	Длит. звонков	Интернет трафик	Тариф	Сумма платежа	…	Группа
1	М	53	78	279	1.2	В	80	…	
2	М	28	217	307	2.7	А	60	…	
3	Ж	58	65	201	0.9	А	100	…	
4	Ж	29	308	965	3.5	В	30	…	
5	М	33	92	180	3.1	Б	100	…	
6	М	56	89	223	0.3	В	35	…	
7	М	37	144	268	1.4	Б	40	…	
8	М	24	103	146	1.1	А	50	…	
9	М	53	48	158	0.6	А	60	…	
10	Ж	68	19	117	0	В	25	…	

Кластеризация: обнаружение мошенничества

ИД	Время	Место	Товар	Сумма	…
123	2016-04-30	Челябинск	Пиджак	2 499.99	
124	2016-05-01	Челябинск	Брюки	1 499.99	
125	2016-05-02	Коркино	Носки	129.99	
	…				
131	2016-05-02	Владивосток	Toyota LC	99 999.99	

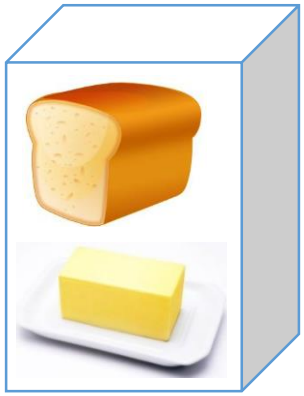
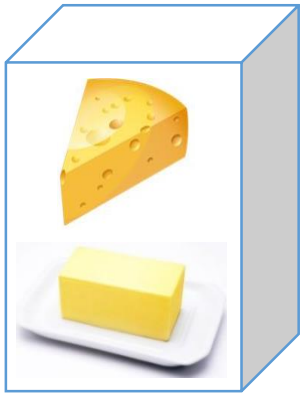


- Распознавание мошеннического использования как выбросов (точек, существенно далеких от остальных)

Поиск шаблонов: анализ рыночных корзин

- Какие наборы товаров в супермаркете часто покупают совместно?

Пример: анализ рыночных корзин

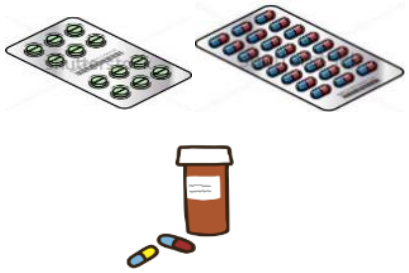

- Увеличение товарооборота и прибыли
  - Близкое расположение товаров в торговом зале
  - Скидка при совместной покупке товаров
  - Ход мыслей покупателя: "**если** санитайзер **то** гречка"

Пример: поиск побочных эффектов лекарств

- Какие симптомы у пациентов часто встречаются совместно с принимаемыми лекарствами?

жар

тошнота
рвота



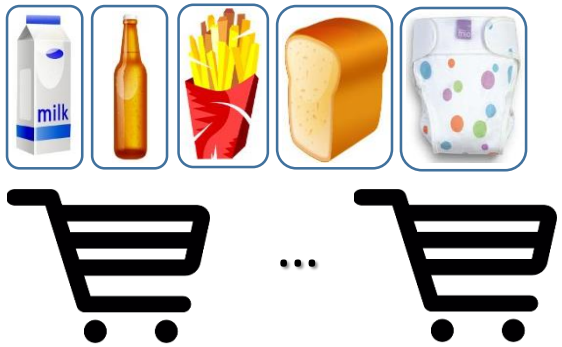
тошнота
жар

тошнота
жар
сыпь

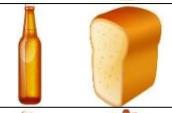
...

Анализ рыночных корзин: постановка задачи

- Дано:
  - товары
  - корзины



- Найти:
  наборы, товары в которых часто покупают совместно («частые» наборы)

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

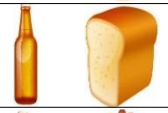
Поддержка набора

- *Поддержка набора (support)* – доля корзин с этим набором

1-наборы	*Support*
	70%
	70%
	20%

2-наборы	*Support*
	40%
	10%
	20%

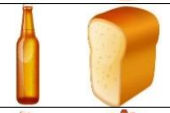
3-наборы	*Support*
	20%
	10%
	20%

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Частый набор

- *minsup – минимальная поддержка* частого набора

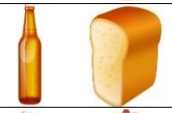
	$\boldsymbol{minsup} = 70\%$
Частые 1-наборы	
Частые 2-наборы	нет
Частые 3-наборы	нет
Частые 4-наборы	нет

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Частый набор

- *minsup – минимальная поддержка* частого набора

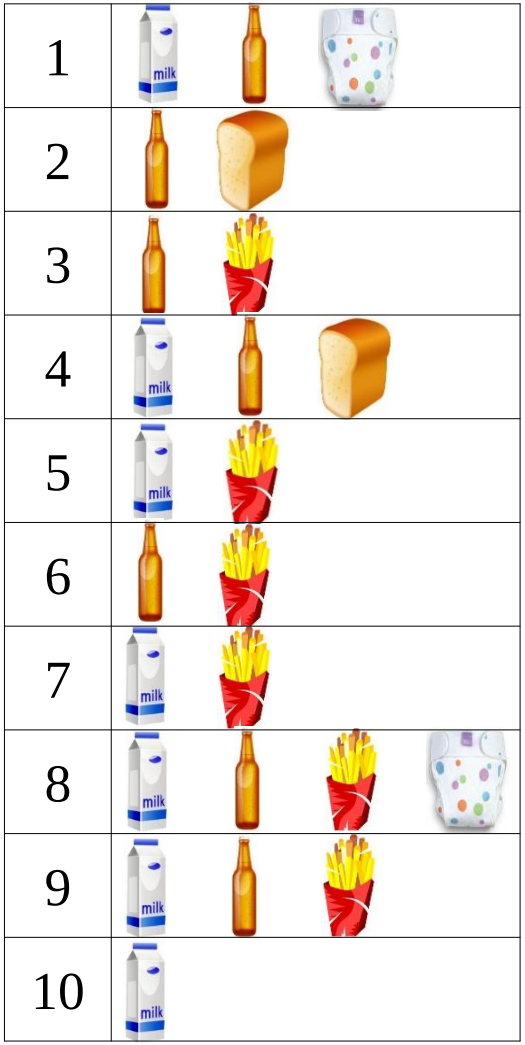
	minsup = 70%	***minsup*** = 40%
Частые 1-наборы		
Частые 2-наборы	нет	
Частые 3-наборы	нет	нет
Частые 4-наборы	нет	нет

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Частый набор

- *minsup – минимальная поддержка* частого набора

	minsup = 70%	***minsup*** = 40%	***minsup*** = 10%
Частые 1-наборы			
Частые 2-наборы	нет		
Частые 3-наборы	нет	нет	
Частые 4-наборы	нет	нет	

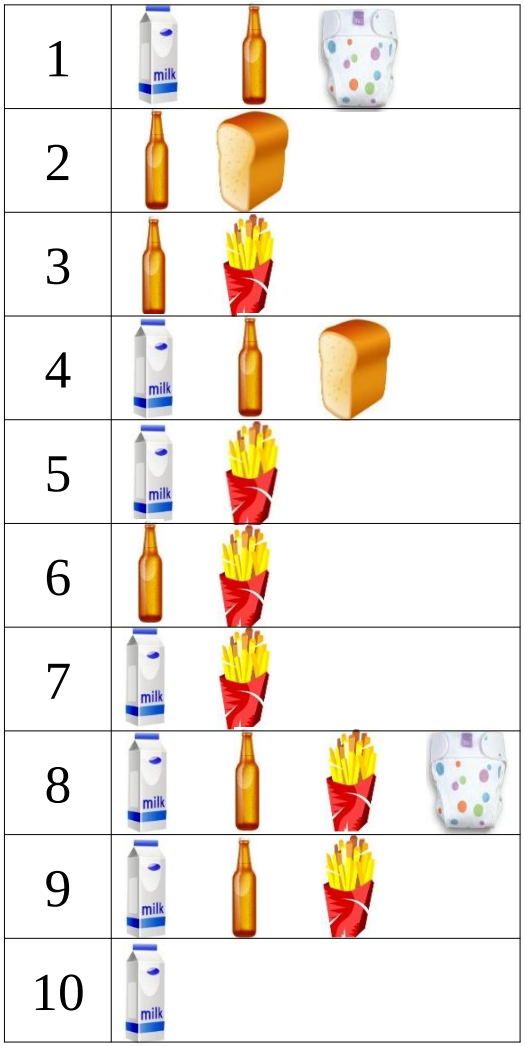


Частый набор

- *minsup – минимальная поддержка* частого набора

	minsup = 70%	***minsup*** = 40%	***minsup*** = 10%
Частые 1-наборы			
Частые 2-наборы	нет		
Частые 3-наборы	нет	нет	
Частые 4-наборы	нет	нет	

- Чем меньше *minsup*, тем больше частых наборов



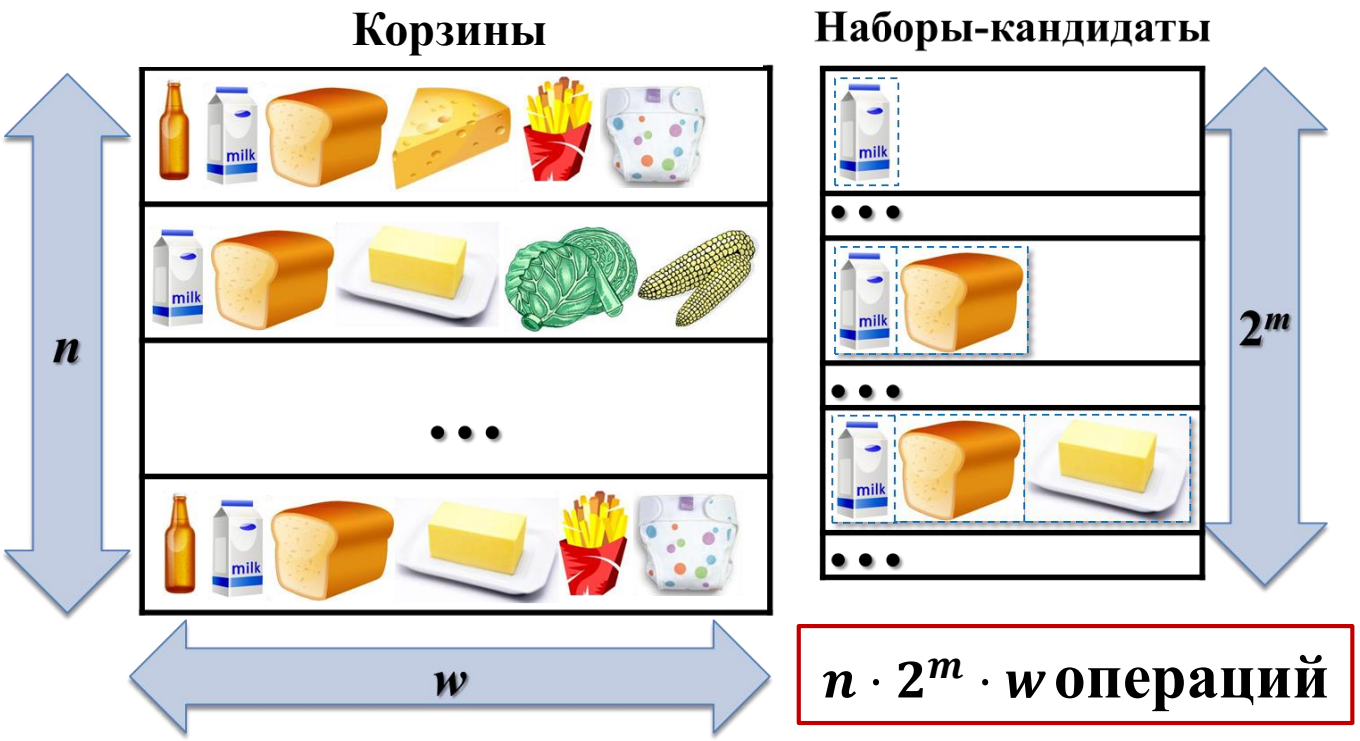
Поиск частых наборов: полный перебор

1. **Сгенерируем всех кандидатов в частые наборы**
2. Вычислим их поддержку и наборы с поддержкой ниже *minsup*

1			**2**
2			**4**
3			**8**
…		…	…
m			2^m

Поиск частых наборов: полный перебор

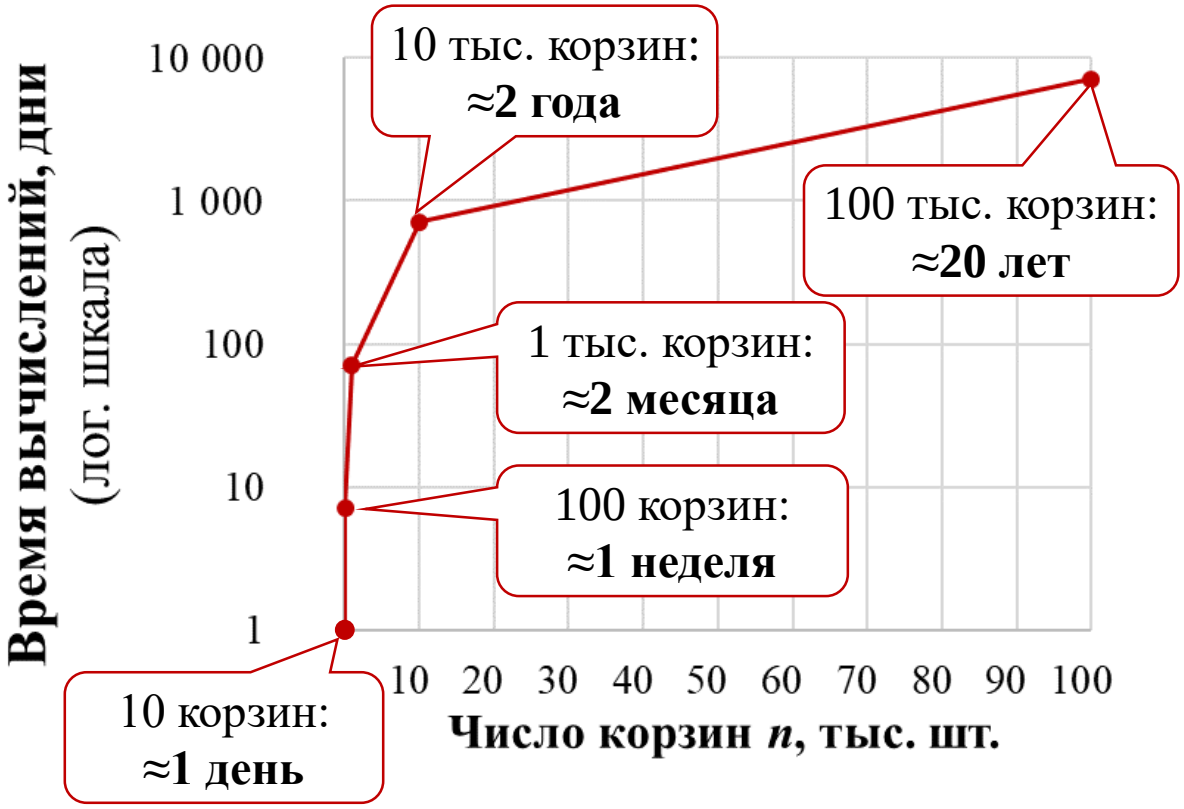
1. Сгенерируем всех кандидатов в частые наборы
2. **Вычислим их поддержку и наборы с поддержкой ниже $minsup$**



Поиск частых наборов: полный перебор

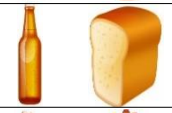
1. Сгенерируем всех кандидатов в частые наборы
2. **Вычислим их поддержку и наборы с поддержкой ниже *minsup***

$n \cdot 2^m \cdot w$ **операций**
(одна операция в секунду)
$m = 10$
$w = 6$



Сокращение перебора: принцип Априори

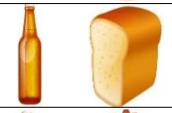
- Наборы и их поддержка
  - sup()=70%
  - sup()=40%
  - sup()=20%
  - sup()=10%

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Сокращение перебора: принцип Априори

- Наборы и их поддержка
  - sup()=70%
  - sup()=40%
  - sup()=20%
  - sup()=10%

- **Поддержка набора не больше поддержки любого его поднабора**

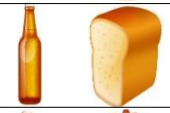
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Сокращение перебора: принцип Априори

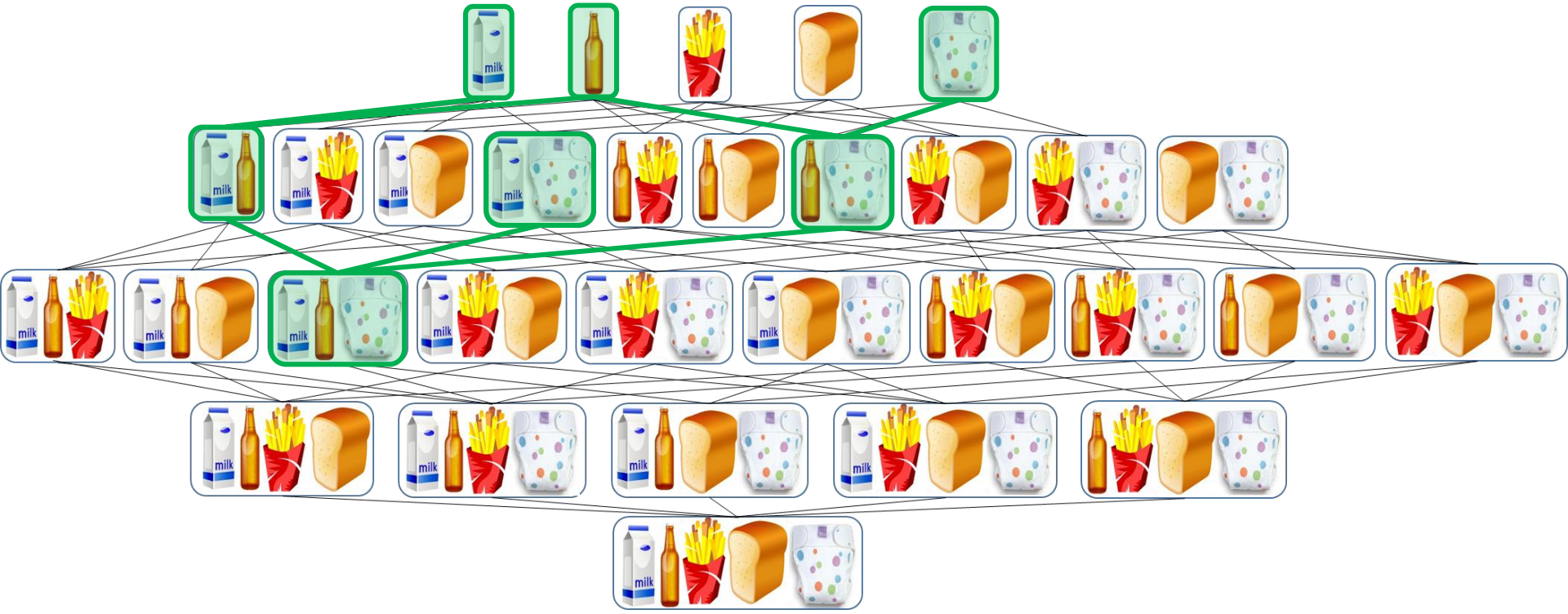
- **Если набор частый,**
  то все его *под*наборы частые
  - sup()=20%
  - sup()=40%, sup()=40%, sup()=40%
  - sup()=60%, sup()=70%, sup()=60%

- **Если набор редкий,**
  то все его *над*наборы редкие
  - sup()=10%

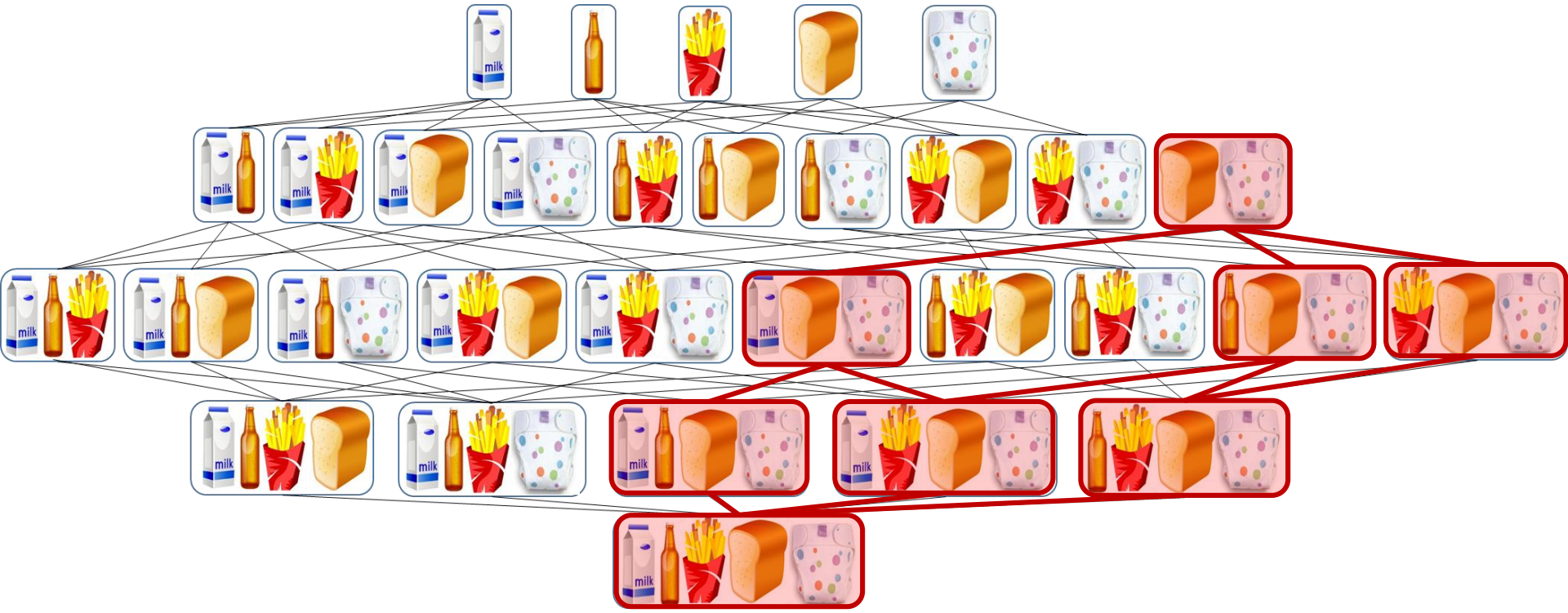
***minsup*=20%**

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Сокращение перебора: принцип Априори



Сокращение перебора: принцип Априори

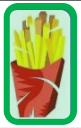


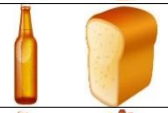
Алгоритм Априори, шаг 1

***minsup*=20%**

- Кандидаты в частые 1-наборы

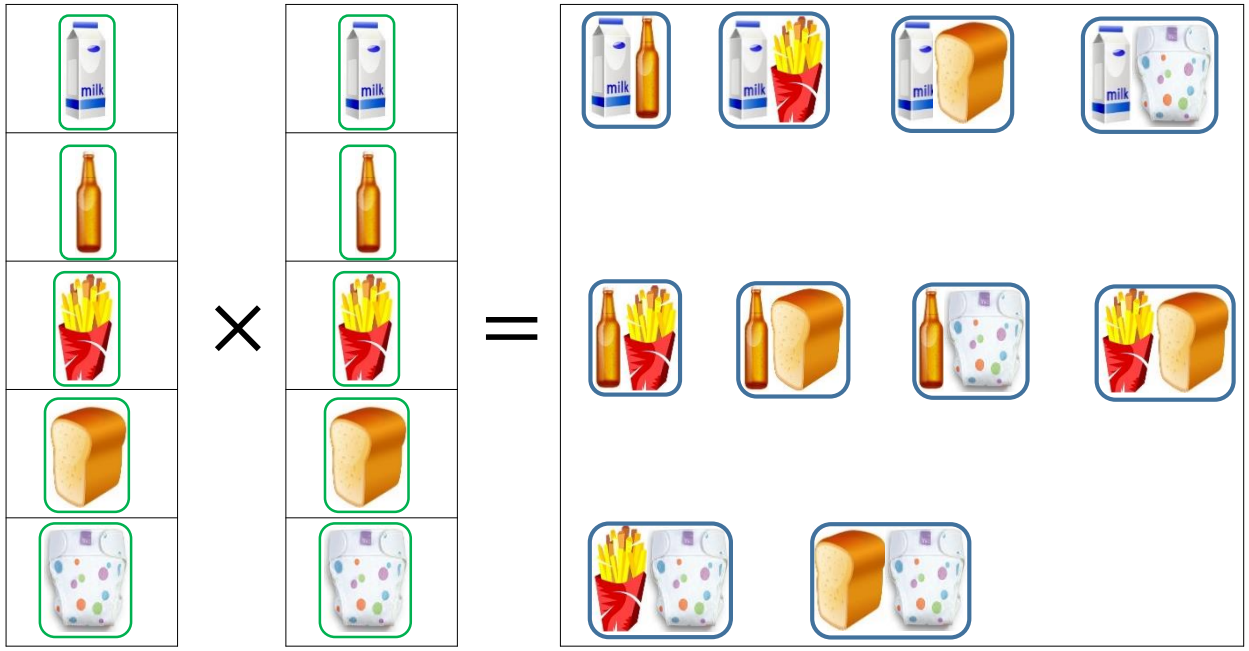
- Подсчет поддержки, частые 1-наборы

1		70%
2		70%
3		60%
4		20%
5		20%

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Алгоритм Априори, шаг 2

- Кандидаты в частые 2-наборы:
  прямое произведение частых 1-наборов на себя



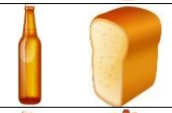
Алгоритм Априори, шаг 2

***minsup*=20%**

- Подсчет поддержки, частые 2-наборы

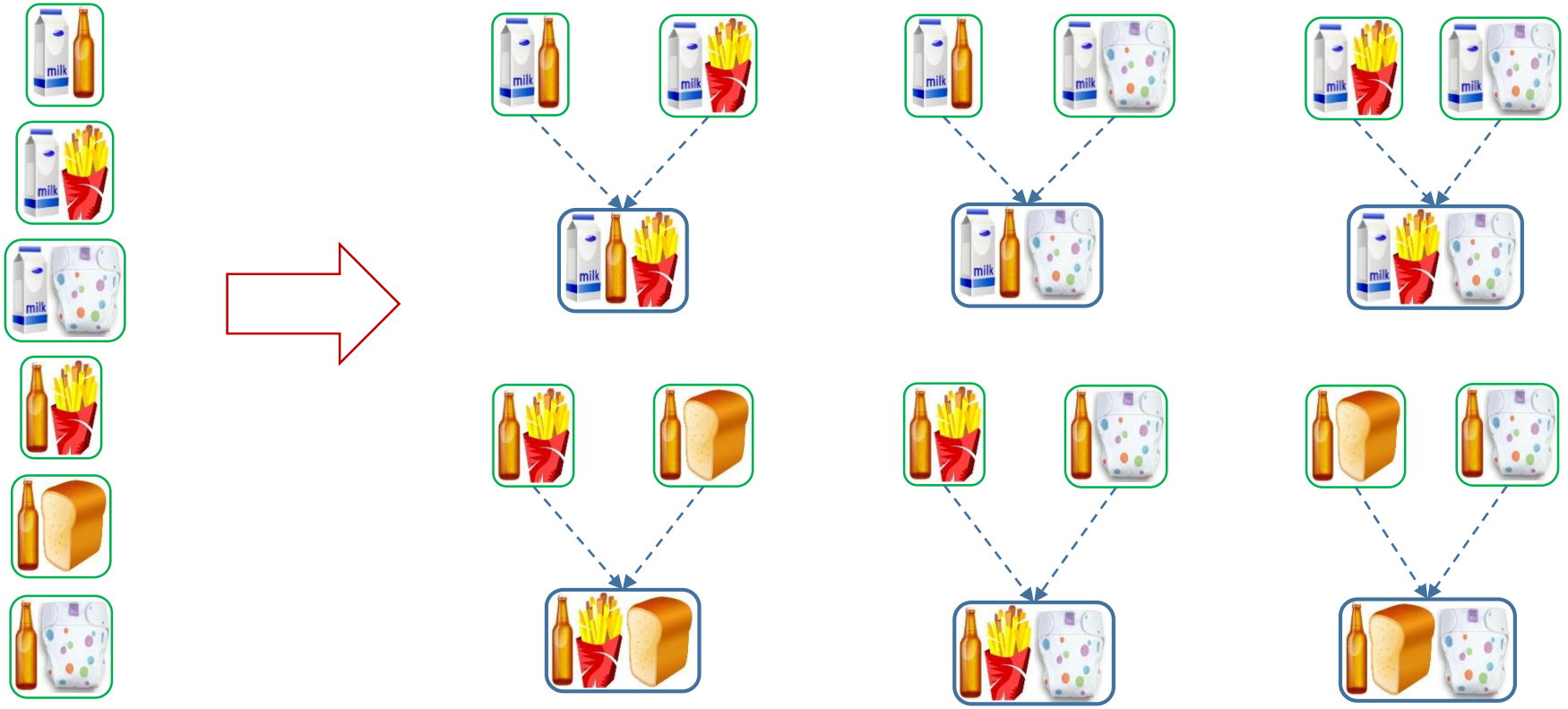
1		40%
2		40%
3		10%
4		20%
5		40%

6		20%
7		20%
8		0
9		10%
10		0

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Алгоритм Априори, шаг 3

- Кандидаты в частые 3-наборы: самосоединение частых 2-наборов



Алгоритм Априори, шаг 3

- Отбрасывание редких 3-наборов кандидатов

Принцип Априори

Если набор частый, то все его поднаборы частые

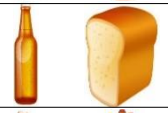
1		
2		
3		
4		
5		
6		

Алгоритм Априори, шаг 3

***minsup*=20%**

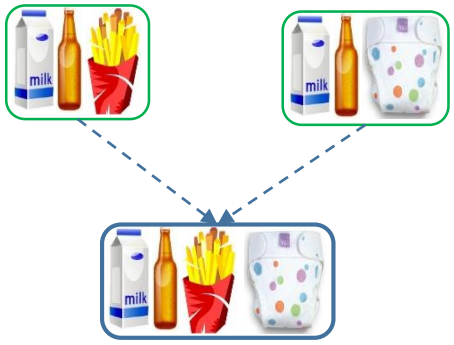
- Подсчет поддержки, частые 3-наборы

1		20%
2		20%

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Алгоритм Априори, шаг 4

- Кандидаты в частые 4-наборы



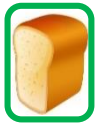
- Отбрасывание редких 4-наборов кандидатов

1		

- Конец обработки: нет наборов-кандидатов

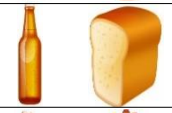
Алгоритм Априори, результаты

***minsup*=20%**

1		**70%**
2		**70%**
3		**60%**
4		**20%**
5		**20%**

6		**40%**
7		**40%**
8		**20%**
9		**40%**
10		**20%**
11		**20%**

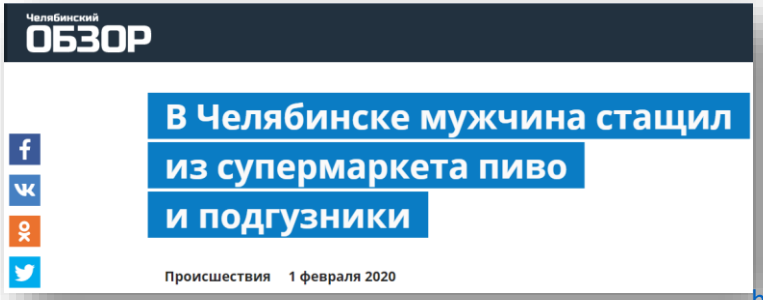
12		**20%**
13		**20%**

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

А при чем здесь  и  ?

- В 1992 г. в США был проведен анализ 1.2 млн. рыночных корзин в 25 магазинах формата «у дома» компании Osco, который выявил частый набор  для покупок в рабочие дни с 5 до 7 час. вечера

  - Руководство Osco не стало ставить эти товары рядом на полках (им были неясны причины такого частого набора)
  - *Объяснение*: молодая семья приходит с работы домой, жена отправляет мужа в ближайший магазин купить для ребенка  , а муж дополнительно покупает себе 

- 

Сотрудники Росгвардии в Челябинске задержали мужчину, подозреваемого в краже из супермаркета. «Уловом» грабителя стали три упаковки подгузников и три банки пива.

https://obzor174.ru/v-chelyabinske-muzhchina-stashchil-iz-supermarketa-pivo-i-podguzniki

Что в итоге?

- Большие данные – феномен современного информационного общества
- Машинное обучение – технология решения задач обработки данных, в которой компьютер обучается решать задачу за счет выявления закономерностей в данных
- Задача поиска шаблонов – нахождение часто встречающихся зависимостей между объектами
- Алгоритм Априори – поиск частых наборов на основе отбрасывания заведомо редких наборов

Обучение машинному обучению в ЮУрГУ

Высшая школа электроники и компьютерных наук
https://eecs.susu.ru/

Группа ВКонтакте: https://vk.com/susu_eecs

Кафедра системного программирования
https://sp.susu.ru/

Шифр	Название	Бюджетные места в 2023 году	ЕГЭ	Проходной балл в 2022 г.
Бакалавриат				
02.03.02	Фундаментальная информатика и информационные технологии	55	Математика, Русский язык, Информатика или Физика	240 (из 300)
09.03.04	Программная инженерия	50	Математика, Русский язык, Информатика или Физика	225 (из 300)