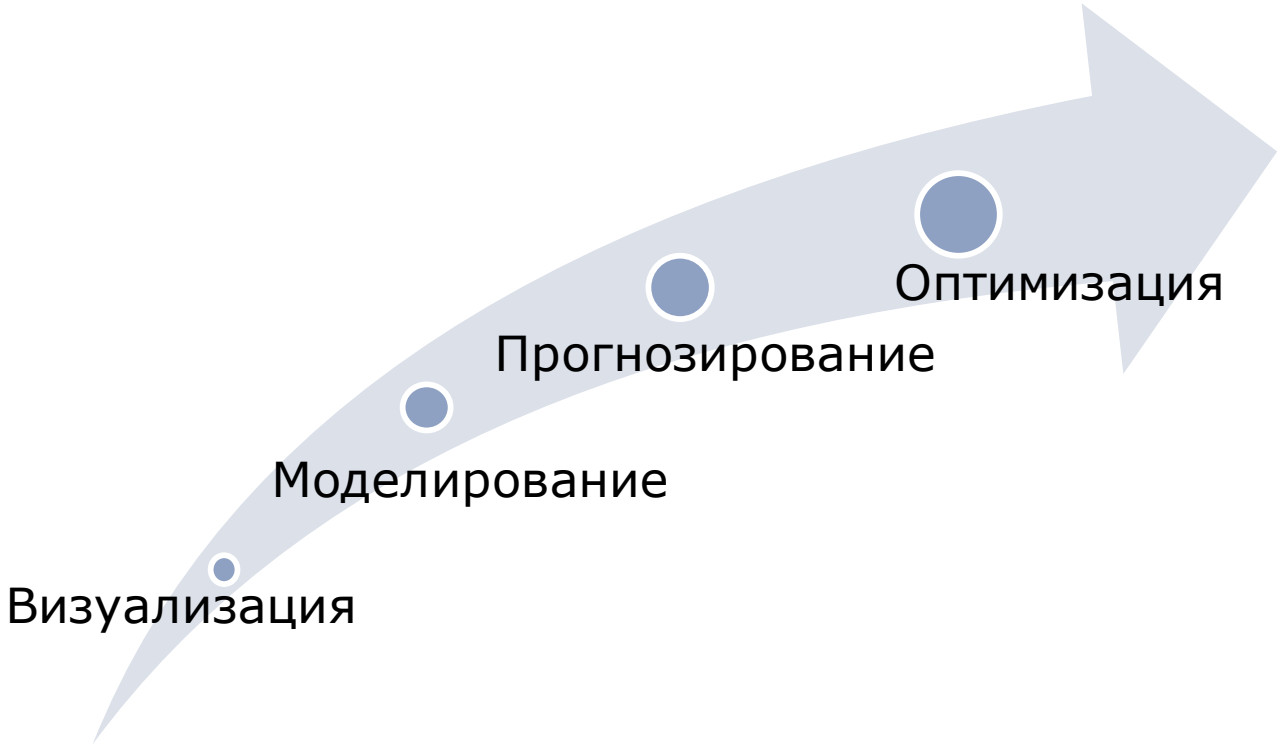
Процесс анализа данных

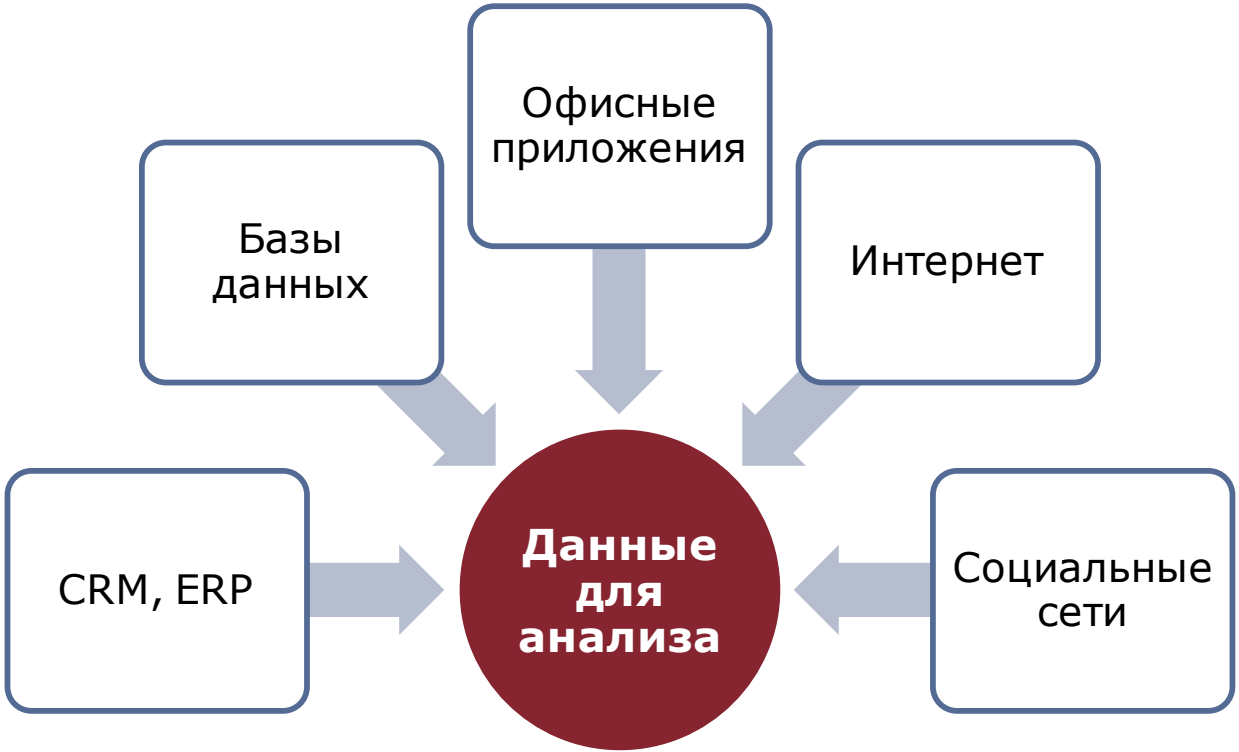
Основные подходы

Уровни анализа



Процесс анализа

Выборка данных

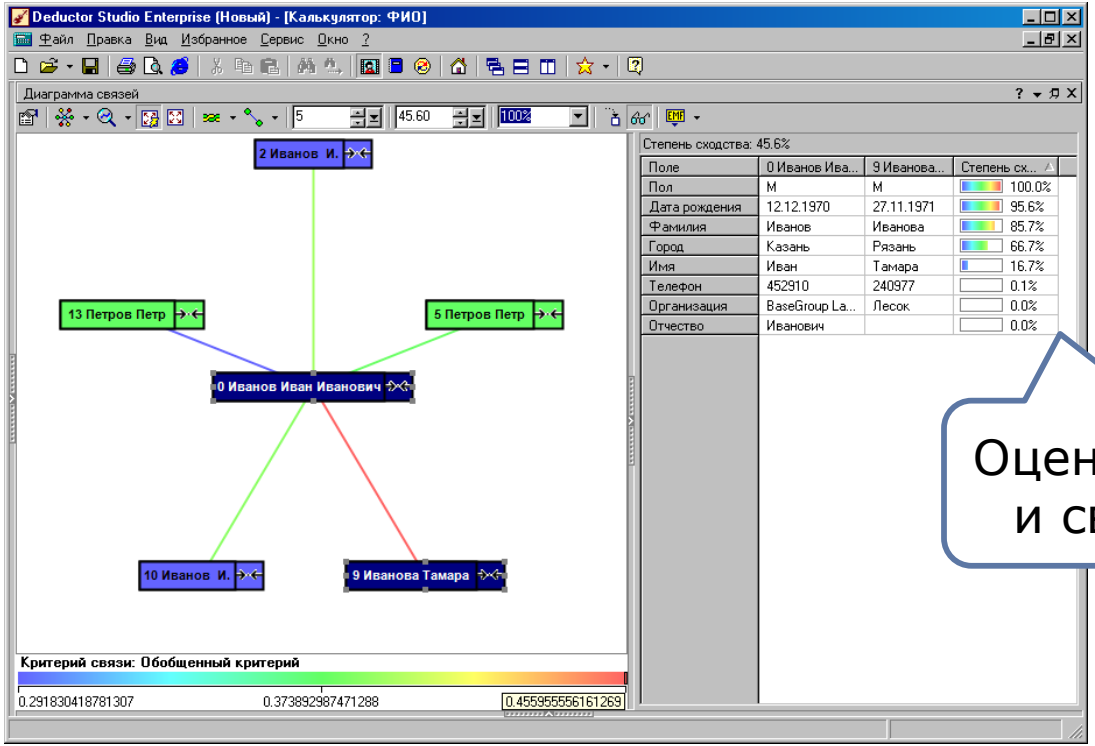


Выборка данных: проблема

Атрибут	Персона 1	Персона 2
ФИО	Иванов Иван Иванович	Иван Иванович
Адрес		г. Рязань ул. Новая 53в
Телефон	+7 (4912) 24-09-77	
Дата рождения	1971 г.	15 декабря
E-mail	ivanov@mail.ru	ivanoff@gmail.com
Место работы	BaseGroup Labs	BGL
Источник	CRM-система	Facebook

Это один человек?

Выборка данных: решение

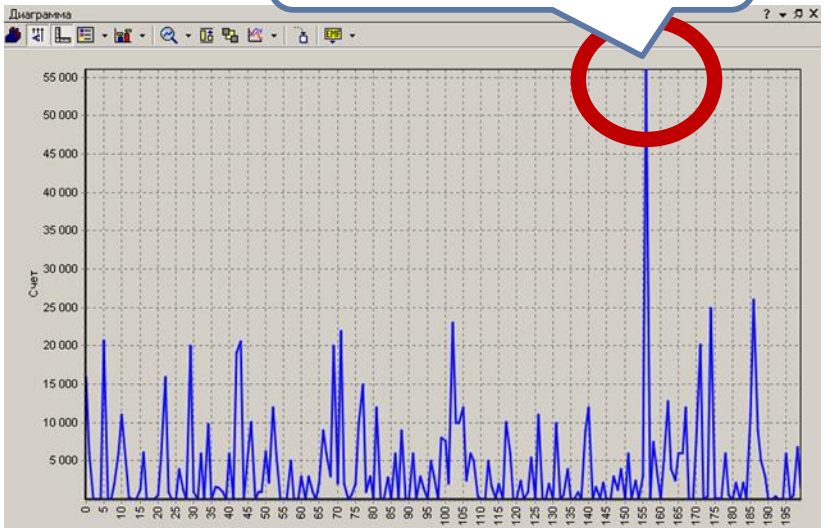


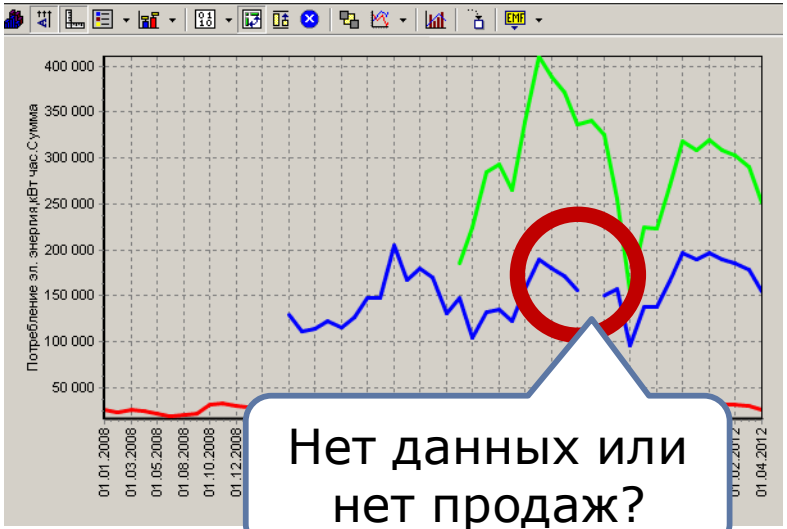
Оценка близости и связывание

Очистка данных

Очистка данных: пример

Аномалия или норма?



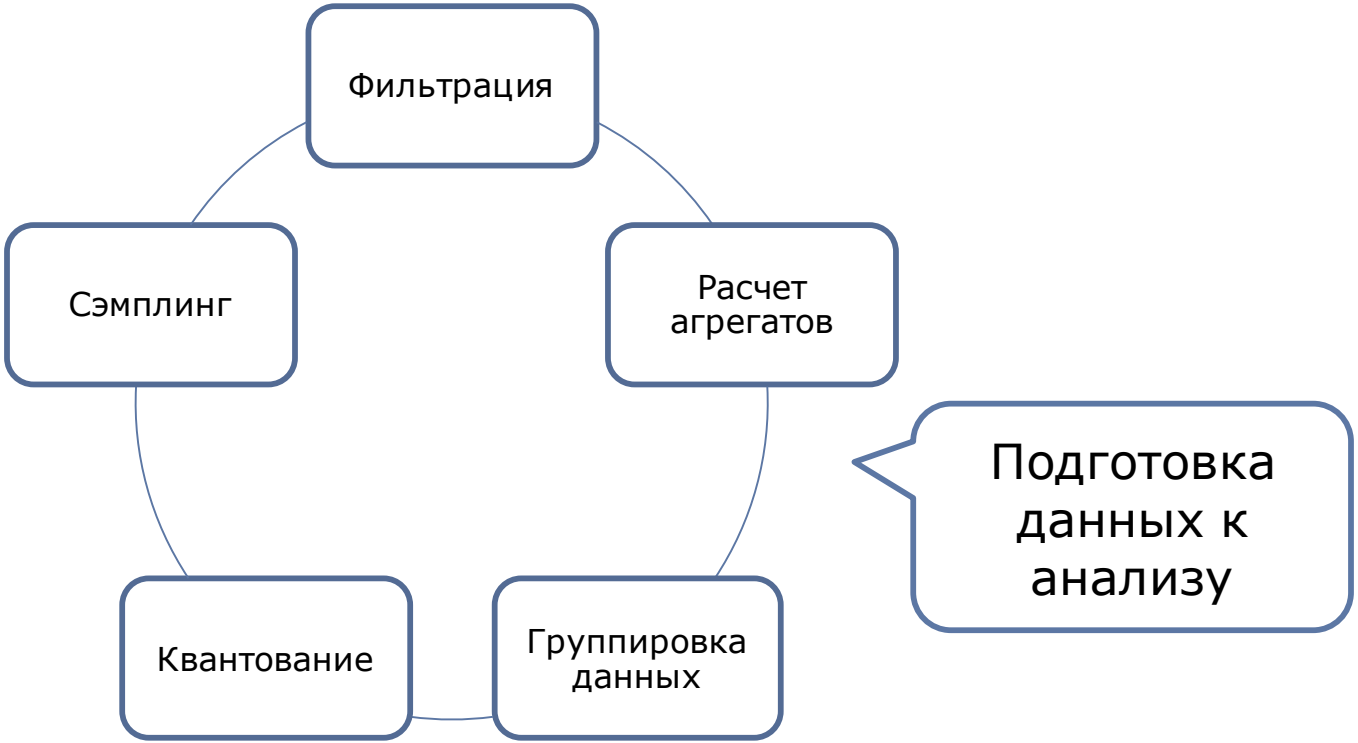


Нет данных или нет продаж?

Очистка данных: решение

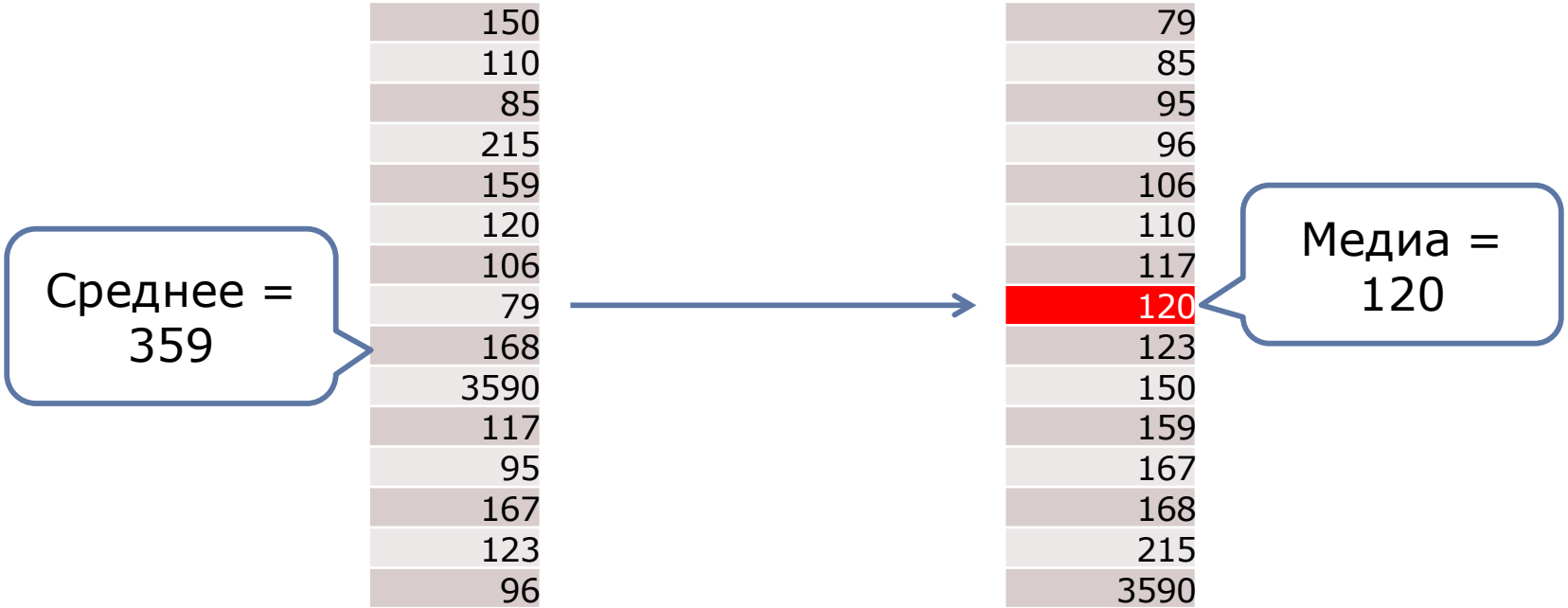
Проблема	Вариант решения
Ошибки ввода	Проверить по справочникам
Пропуски	Интерполировать
Аномалии	Срезать выбросы
Дубли	Оставить одну запись
Противоречия	Удалить записи

Трансформация

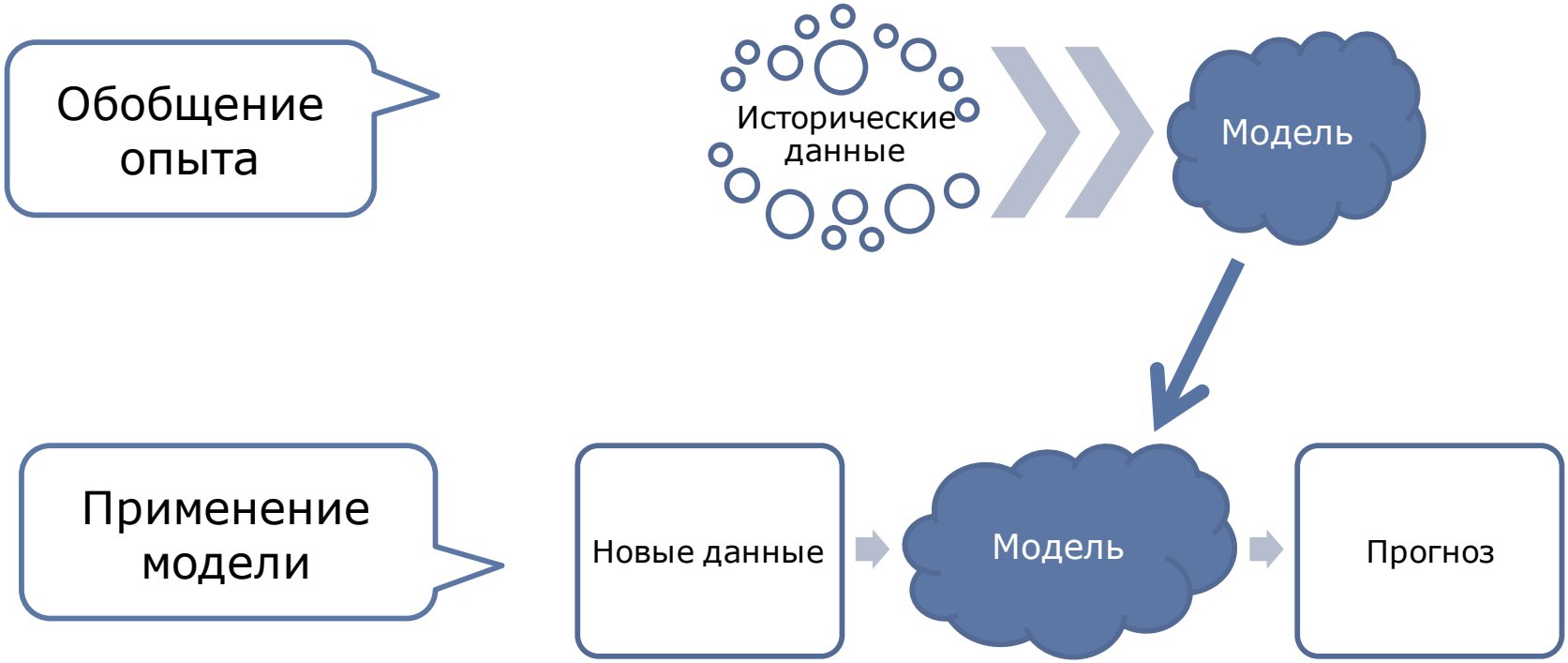


Трансформация: проблема

Трансформация: решение



Data Mining


Обобщение опыта
Исторические данные
Модель
Применение модели
Новые данные
Модель
Прогноз

Интерпретация результатов

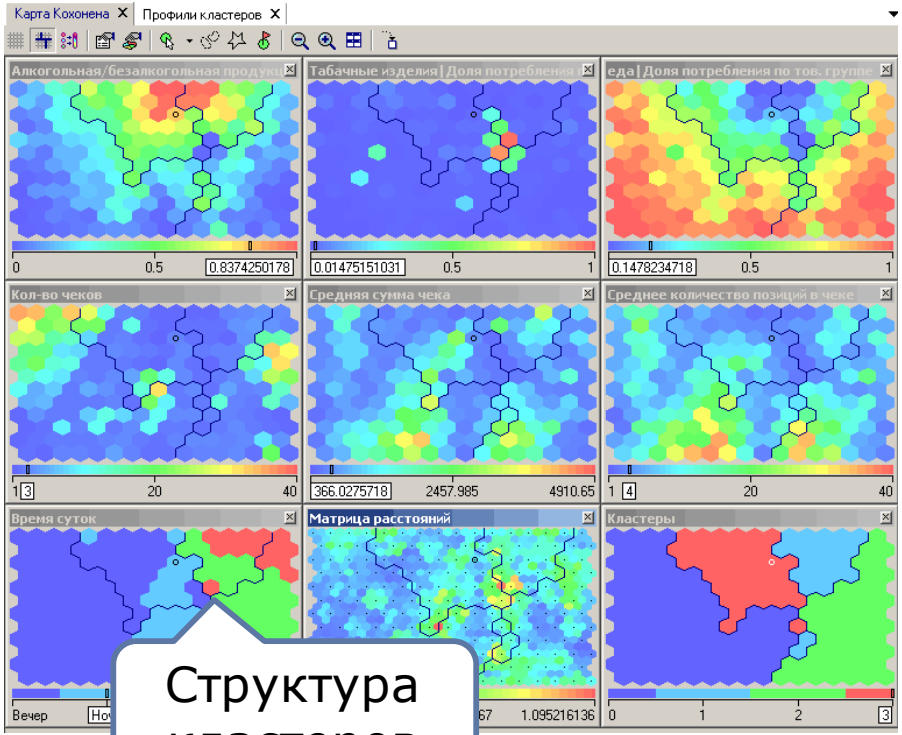
Трудно понять модель

Нет доверия к результатам

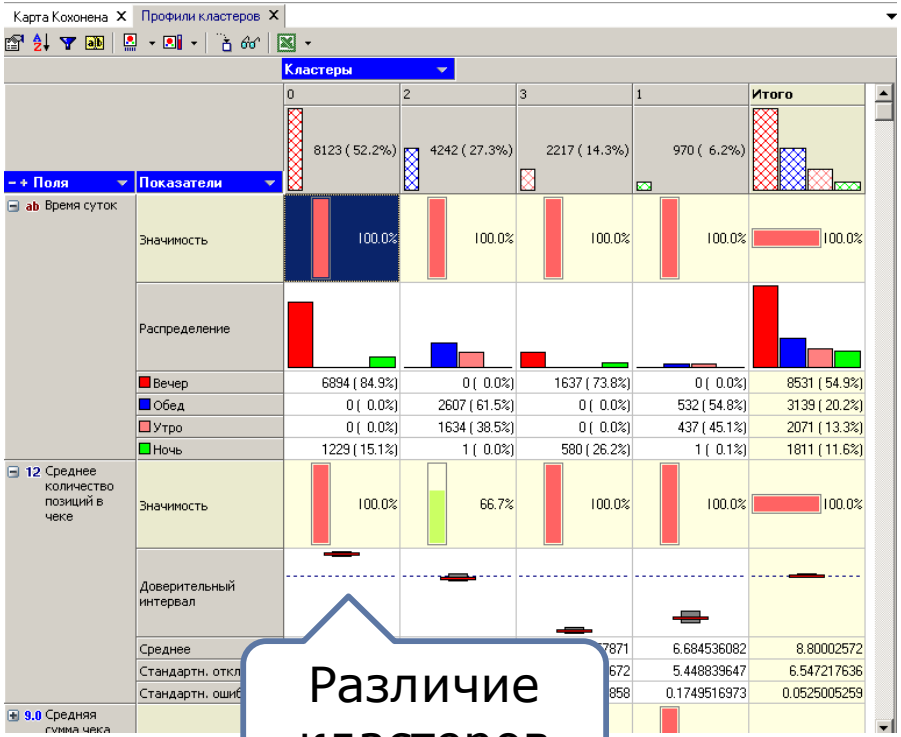
Отказ в применении модели

BaseGroup Labs
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

Визуализация – способ понять

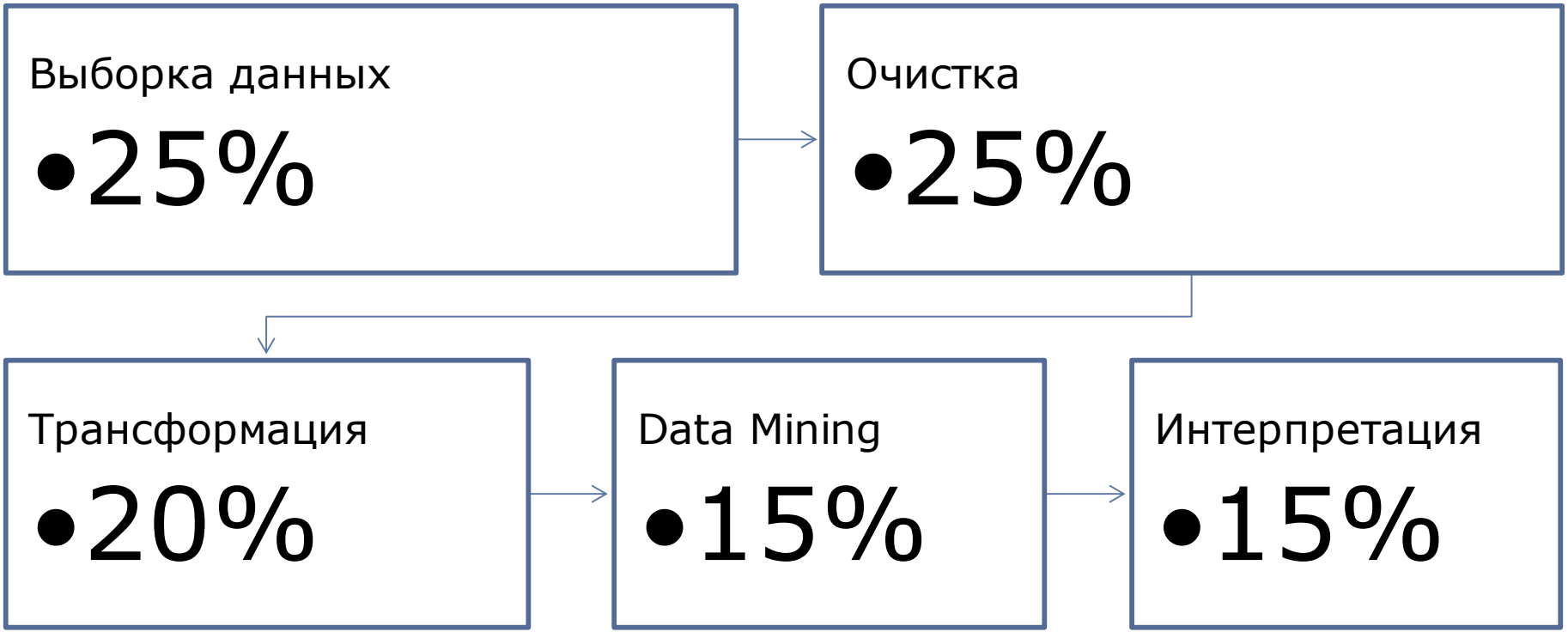


Структура кластеров



Различие кластеров

Трудоемкость этапов



Пример: прогнозирование

Пример: отток клиентов

Data Mining

Data Mining

Data Mining – это процесс обнаружения

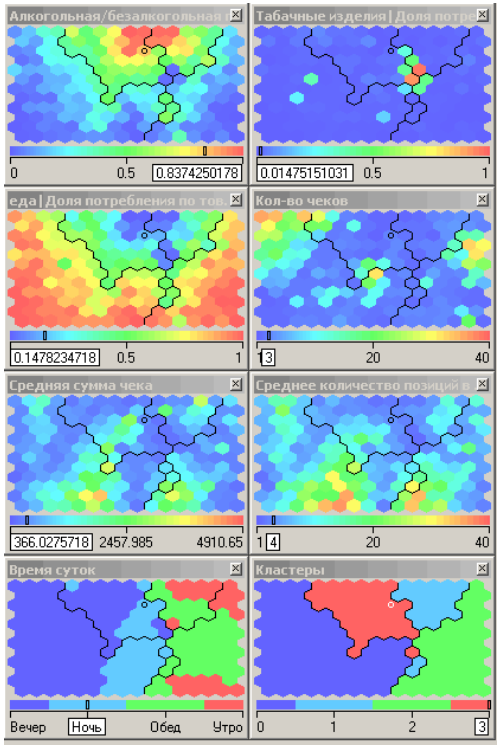
- в **'сырых'** данных
- ранее неизвестных **нетривиальных**
- практически **полезных** и
- доступных **интерпретации** знаний,
- необходимых для принятия **решений**

Классы задач Data Mining

- Кластеризация
- Регрессия
- Классификация
- Ассоциативные правила
- Последовательные шаблоны
- Анализ временных рядов
- Анализ связей
- Анализ отклонений

Кластеризация

Объединение «похожих» объектов в сравнительно **однородные** группы, существенно отличающихся от других групп

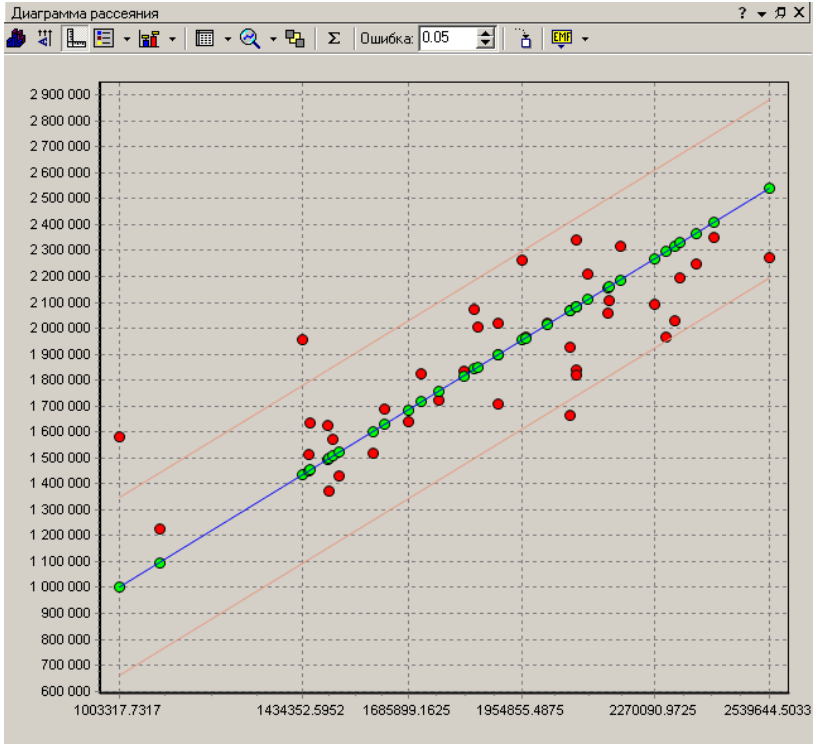


Кластеризация: задачи

- Сегментация клиентов
- Выявление целевой аудитории
- Анализ миграции клиентов
- Канибализация товаров

Регрессия

Предсказание значения **непрерывной** зависимой переменной с помощью независимых переменных

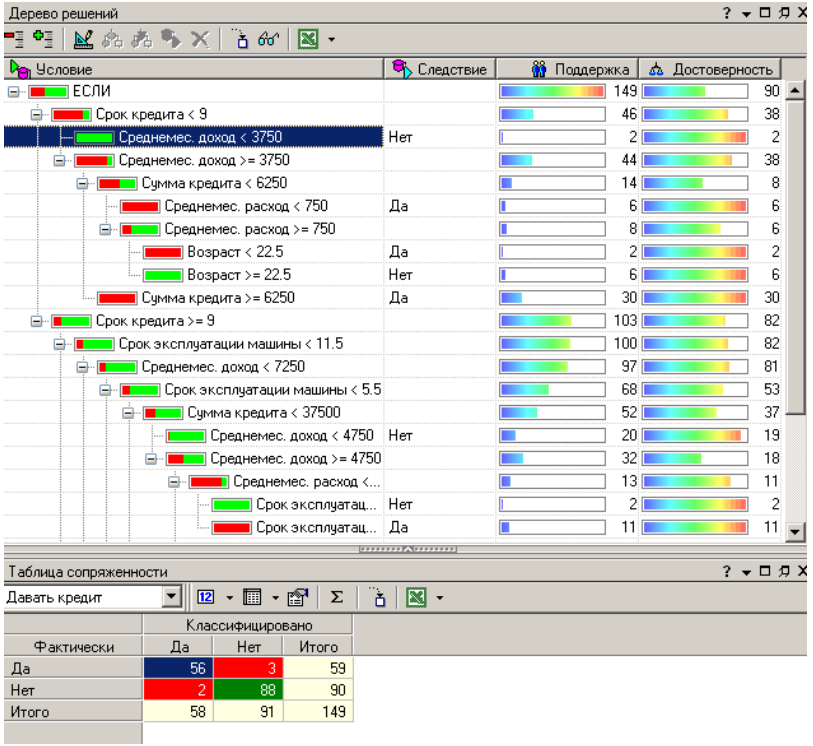


Регрессия: задачи

- Прогнозирование спроса
- Вероятность отклика на предложение
- Оценка эластичности цен
- Кредитный скоринг

Классификация

Отнесение объектов к одному из **известных классов** с помощью независимых переменных



Классификация: задачи

- Оценка перспективности клиента
- Предсказание мошенничества
- Прогнозирование оттока
- Анализ рисков

Ассоциация

Обнаружение в транзакциях зависимостей, что из события X с определенной вероятностью следует событие Y

Правила

Правил: 65 из 65 | Фильтр: Без фильтрации

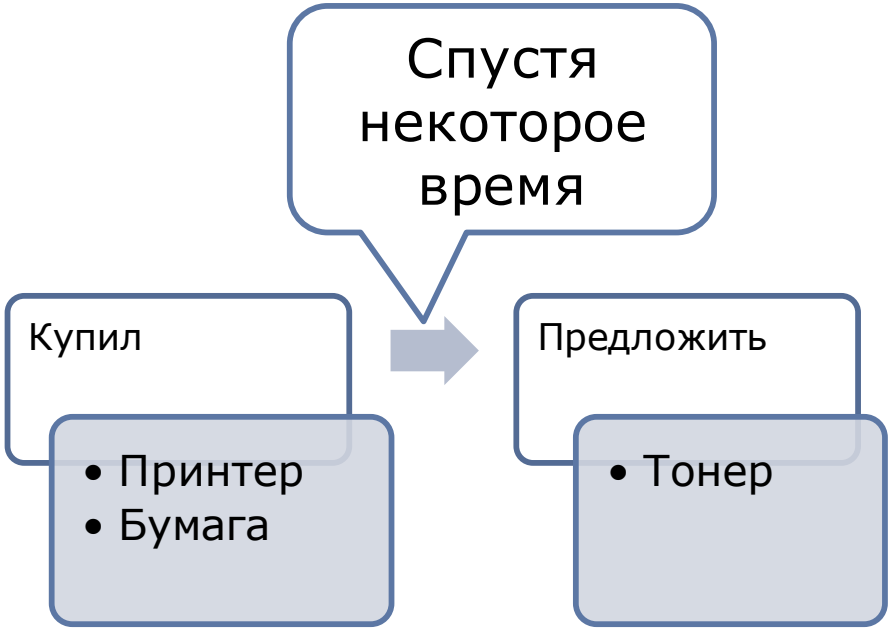
№	Условие	Следствие	Поддержка		Достоверност	Лифт
			Кол-во	%		
52	Срочный вклад		126	1.58	60.58	3.913
53	Кредитные карты	Депозитные сертификаты	126	1.58	28.25	1.152
	Срочный вклад					
54	Депозитные сертификаты	Экспресс-кредит	92	1.15	27.96	1.698
	Кредитные карты					
55	Депозитные сертификаты	Кредитные карты	92	1.15	39.48	2.551
	Экспресс-кредит					
56	Кредитные карты	Депозитные сертификаты	92	1.15	24.86	1.014
	Экспресс-кредит					
57	Кобрендинговые карты	Срочный вклад	142	1.78	51.08	4.520
	Кредитные карты					
58	Кобрендинговые карты	Кредитные карты	142	1.78	67.62	4.368
	Срочный вклад					
59	Кредитные карты	Кобрендинговые карты	142	1.78	31.84	1.825
	Срочный вклад					
60	Кобрендинговые карты	Экспресс-кредит	84	1.05	30.22	1.835
	Кредитные карты					
61	Кобрендинговые карты	Кредитные карты	84	1.05	37.33	2.412
	Экспресс-кредит					
62	Кредитные карты	Кобрендинговые карты	84	1.05	22.70	1.301
	Экспресс-кредит					
63	Кредитные карты	Экспресс-кредит	160	2.00	35.87	2.178
	Срочный вклад					

Ассоциация: задачи

- Анализ рыночной корзины
- Кросс-продажи (Cross-sale)
- Повышение доходности (Up-sale)
- Лучшее товарное предложение (Next Best Offer)

Последовательность

Выявление зависимости, что **после** события X, с определенной вероятностью наступит событие Y

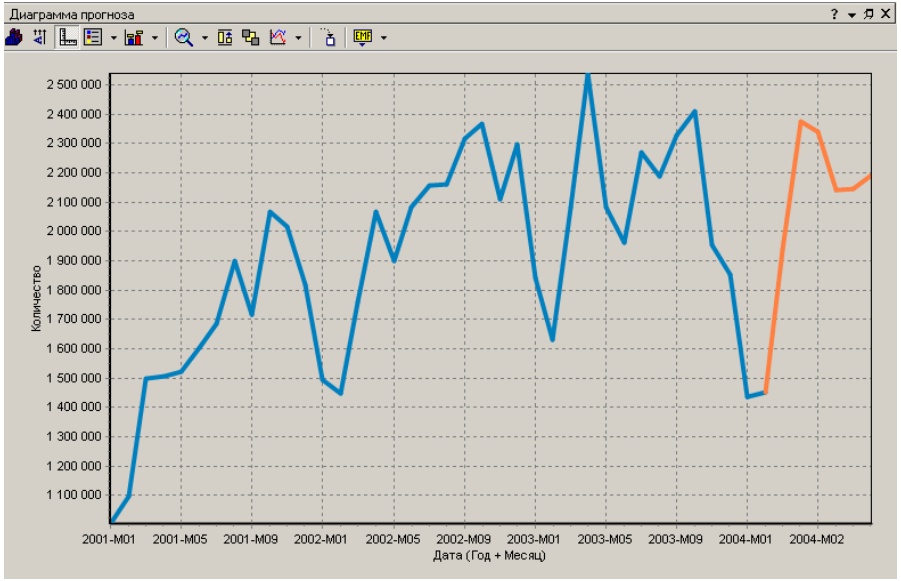


Последовательность: задачи

- Предсказание переходов по сайту
- Анализ отложенного спроса
- Оптимизация работы службы технической поддержки

Анализ временных рядов

Предсказание будущих значений временного ряда по настоящим и прошлым значениям

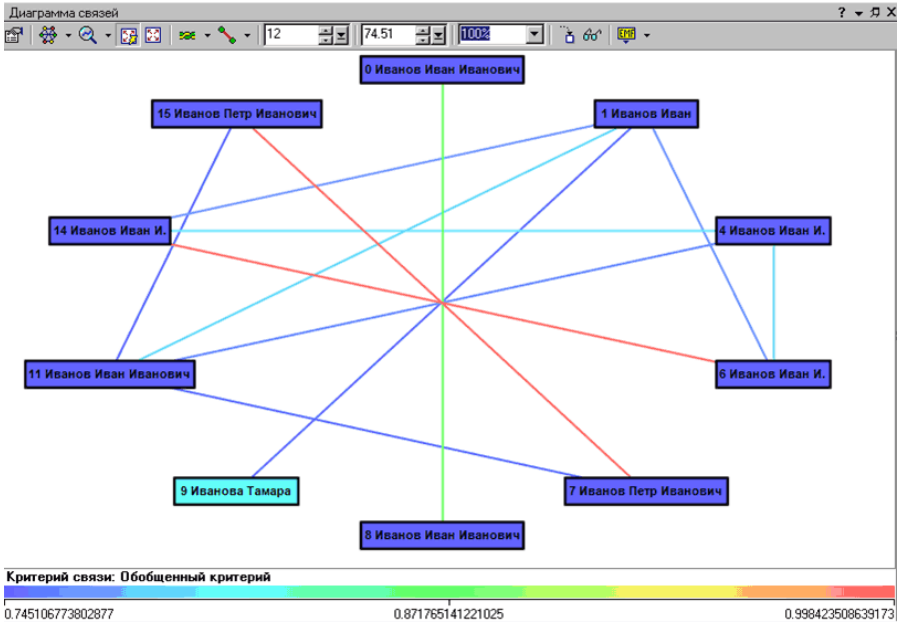


Анализ временных рядов: задачи

- Прогнозирование спроса
- Оптимизация складских запасов
- Прогнозирование финансовых потоков
- Прогнозирование потребности в ресурсах

Анализ связей

Выявление отношений между объектами сети для определения ранее неизвестных характеристик объектов

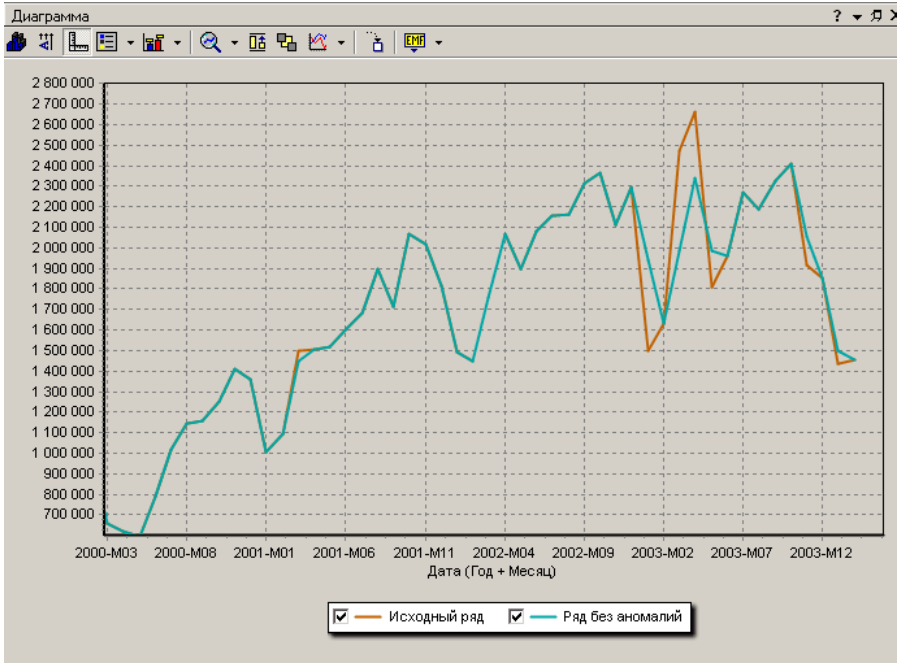


Анализ связей: задачи

- Противодействие мошенничеству
- Защита конфиденциальных данных
- Построение профилей клиентов
- Выбор каналов воздействия

Анализ отклонений

Обнаружение наиболее нехарактерных случаев, выбивающихся из общих закономерностей



Анализ отклонений: задачи

- Выявление подозрительной активности
- Анализ влияния маркетинговых акций
- Автоматический контроль выполнения KPI

Применение в бизнесе

Решение большинства задач бизнес-аналитики сводятся к описанным классам задач Data Mining или их комбинациям.

Кейс: мошенничество

Мошенничество в рознице

До 70% потерь происходит по вине персонала. Проблемная зона – касса:

- Воровство и обман покупателей
- Неправомерное использование скидок
- Начисление баллов на бонусные карты

Что такое мошенничество

Мошенничество – не только воровство, но и **осознанное** нарушение корпоративных правил работы:

- Начисление баллов на свою карту
- Использование служебной карты для родственников и знакомых

Выявление мошенничества

Противодействия мошенничеству базируются на выявлении последовательности подозрительных действий, оценке вероятности обмана и расчете финансовых потерь:

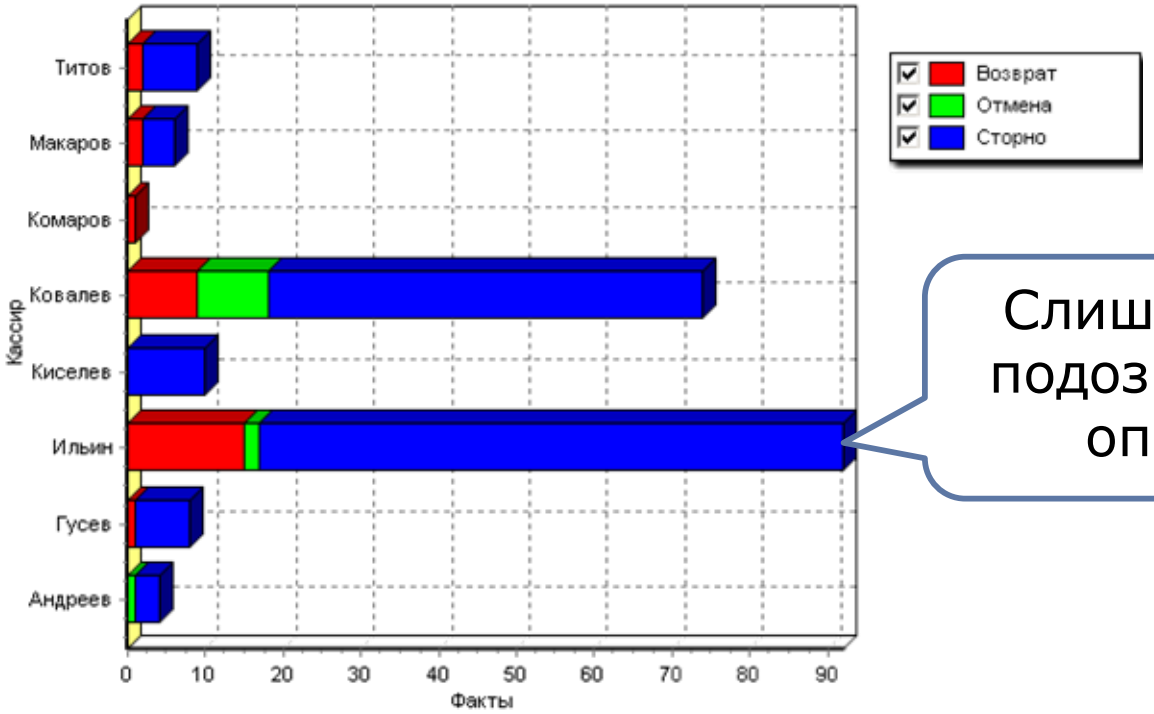
- Жесткие правила известных схем обмана
- Частотный анализ действий
- Аномальные выбросы во временных рядах
- Анализ последовательности действий
- Поиск подозрительных сочетаний
- Кластеризация транзакций

Много бонусов

Магазин	Месяц	Вариант
	# Количество	
Кассир	Σ Значение	%↕ Процент
Ильин	4 877	34,70%
Гусев	1 796	12,78%
Ковалев	1 720	12,24%
Петров	1 306	9,29%
Киселев	1 148	8,17%
Лебедев	741	5,27%
Кузнецов	653	4,65%
Андреев	521	3,71%
Соколов	452	3,22%
Голубев	358	2,55%
Новиков	153	1,09%
Козлов	147	1,05%
Павлов	105	0,75%
Морозов	33	0,23%
Иванов	30	0,21%
Комаров	16	0,11%
Итого:	14 056	100,00%

Кассир начислил подозрительно много бонусов

Аномальное сторно



Слишком много подозрительных операций

BaseGroup Labs
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

Странный возврат

Создан чек

№ транзакции	Дата транзакции	Время транзакции	Тип транзакции	№ ККМ	№ чека	№ возвр. чека	Код кассира	Код товара	Цена	Количество	Сумма
82721	40319	18:19:15	11 – регистрация товара	2	7095		8	1583	56.4	3	169.2
82722	40319	18:19:20	11 – регистрация товара	2	7095		8	249875	15.5	12	186
82723	40319	18:19:27	11 – регистрация товара	2	7095		8	13752	35	8	280
82724	40319	18:19:51	40 – оплата	2	7095		8				635.2
82725	40319	18:19:51	55 – закрытие чека	2	7095		8				635.2

№ транзакции	Дата транзакции	Время транзакции	Тип транзакции	№ ККМ	№ чека	№ возвр. чека	Код кассира	Код тов.	Цена	Кол-во	Сумма
83326	40319	19:53:33	80 – возврат по номеру чека	2	7175	7095	8				
83327	40319	19:53:50	13– возврат	2	7175	7095	8	13752	35	-2	-70
83328	40319	19:53:53	40 – оплата	2	7175	7095	8				-70
83329	40319	19:53:53	55 – закрытие чека	2	7175	7095	8				-70

Отмена чека через час

Плохие сочетания

Кластеров: 2 из 2 | Фильтр: Без фильтрации

№	Номер кластера	Количество транзакций	Ширина кластера	Мощность кластера
0	0	12	3	36
1	1	16	10	78

Кластер 0

Элементы
регистрация товара
оплата
закрытие чека

Профиль нормального чека

Кластер 1

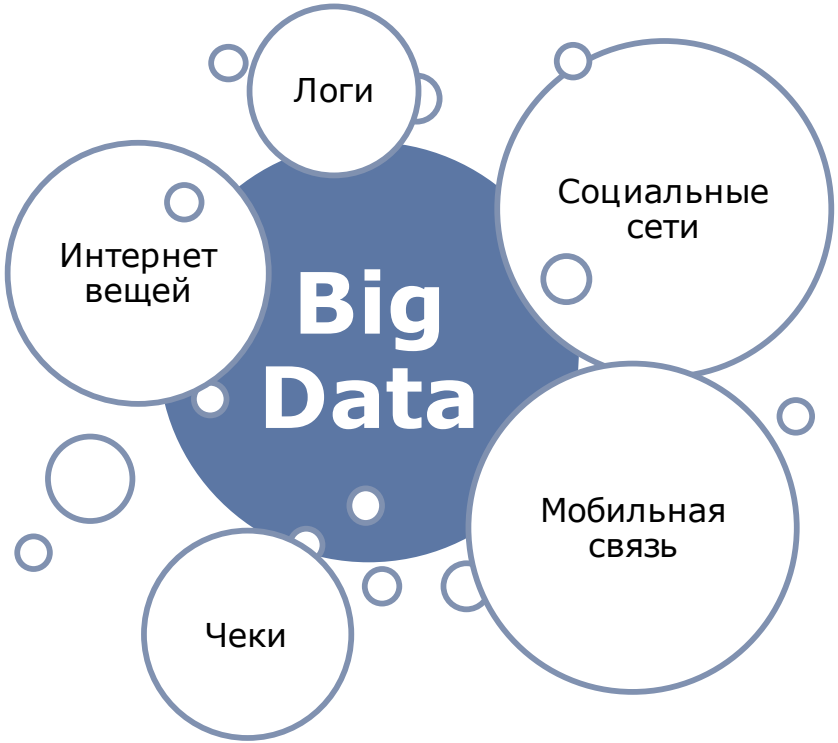
Элементы
возврат
возврат по номеру чека
детализация скидки на чек
закрытие чека
оплата
отложенный чек
отмена чека
регистрация товара
скидка на чек, %
сторно

Профиль «плохого» чека

Big Data

Обвал данных

- Каждый день в мире генерируется 10^{18} байт информации
- 90% всех существующих данных созданы за последние 2 года
- Каждый час Wal-Mart генерирует данных в 170 раз больше объема данных Библиотеки Конгресса США



Big Data

Проблемы больших данных:

- Volume – огромные объема данных
- Velocity – высокая скорость генерации новых данных
- Variety – многообразие структурированных и неструктурированных источников данных

Потенциал Big Data

- Мнение клиентов
- Рекомендательные системы
- Массовая кастомизация услуг
- Противодействие оттоку
- Борьба с мошшеничеством
- Построение профилей клиентов

Знания из данных

Ручная обработка огромных потоков данных практически бесполезна.

Технологии Data Mining – реальный способ извлечь **ценные знания** из Big Data, превратив умение анализировать данные в конкурентное преимущество.

basegrpoup.ru