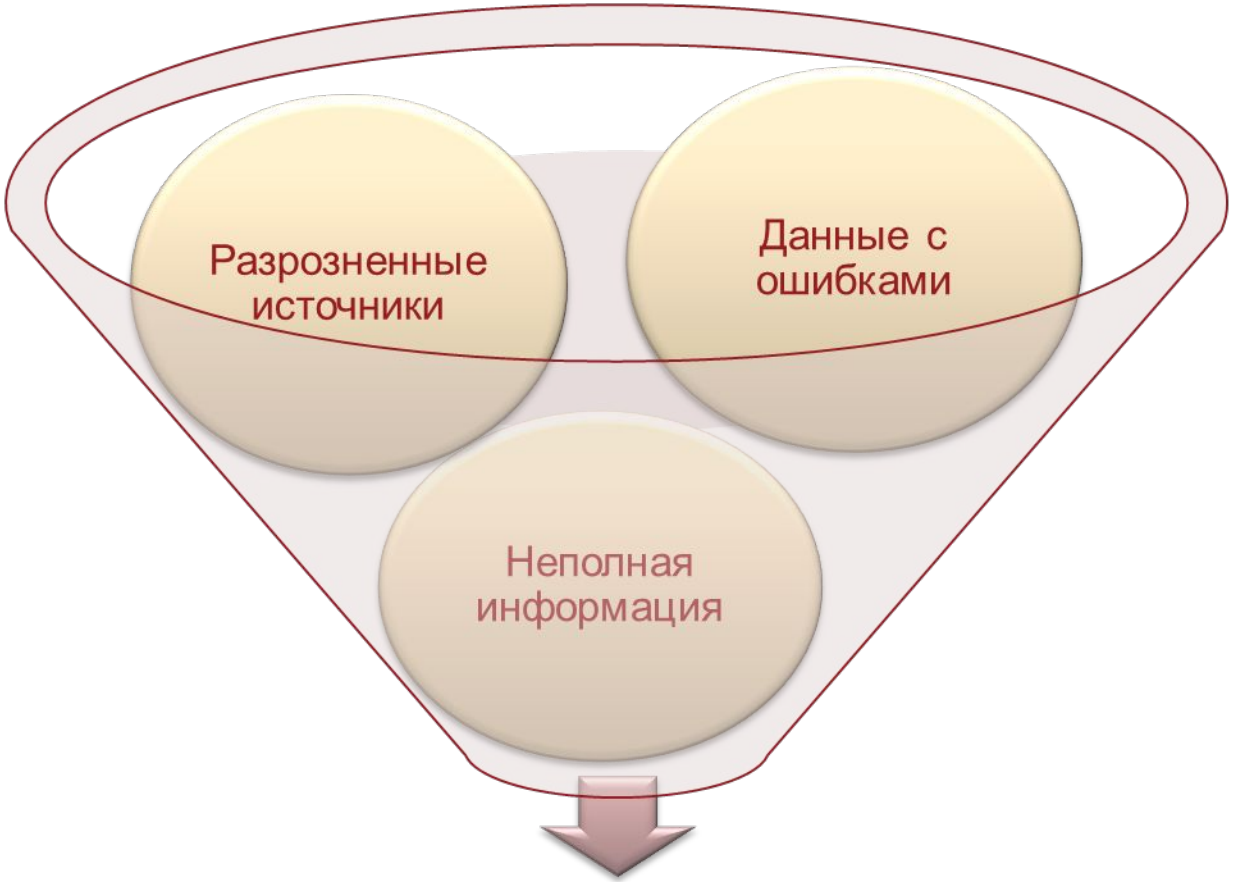
Методы стандартизации, очистки и обогащения данных

Проблема


Разрозненные источники
Данные с ошибками
Неполная информация
Некачественный анализ
Невозможность проведения анализа!

Решение

«Грязные» данные

Стандартизация

Очистка и контроль

Обогащение

Комплекс мер по улучшению качества данных

Качественные данные

Стандартизация: понятие

Стандартизация – это унификация представления и приведение к единому формату данных.

Задачи:

- Нормализация баз данных с целью удаления избыточности: разбиение на несколько таблиц, выделение первичных ключей…
- Разбор строк на атомарные объекты: разделение поля «ФИО» на значения «Фамилия», «Имя», «Отчество», разбор адреса по КЛАДР…
- Унификация представления: преобразование номеров телефонов к стандартному виду +7 (XXX) XXXXXXX…

Стандартизация: парсинг

Парсинг – грамматический или лексический анализ текста.

Осуществляет деление поля на атомарные значения

Значение
Иван Петрович Сидоров, студент 5 курса РГРТУ, кафедра «ЭВМ»

Название	Значение
Имя	Иван
Отчество	Петрович
Фамилия	Сидоров
Вуз	РГРТУ
Кафедра	ЭВМ
Курс	5

Стандартизация: словари

Использование **машинных словарей** (справочников имен, телефонных кодов, КЛАДР, БИК…) позволяет стандартизировать представление данных.

Исходный адрес
Пос. Кустаревка, Ул. Кооперативная

Информация из КЛАДР	
Пос. Кустаревка	06201800005100
Ул. Кооперативная	0620180000510000700

Стандартизированный адрес		
Индекс		391450
Область	062	Рязанская область
Район	018	Сасовский район
Код населённого пункта	051	п. Кустаревка
Код улицы	0700	ул. Кооперативная

Стандартизация: регулярные выражения

Регулярные выражения позволяют производить манипуляции с данными, используя шаблоны:

- находить в строке подстроки, удовлетворяющие заданному шаблону: поиск жителей, прописанных в Москве…
- извлекать из строки фрагменты, с заданным стандартом написания: выделение почтового индекса или года рождения…
- изменять в строке подстроки, соответствующие шаблону: удаление нечисловых символов из паспортных данных или телефона…
- проверять, соответствует ли строка заданному шаблону: проверка корректности e-mail…

Очистка данных: понятие

Очистка данных – процесс выявления и исправления ошибок, позволяющий обеспечить качественный анализ.

Задачи:

- Оценка достоверности информации
- Выявление ошибочных и подозрительных данных: аномалий, дубликатов, противоречий…
- Исправление выявленных ошибок

Очистка: частотный анализ

Метод основывается на анализе **частоты появления** определенного значения или комбинаций таких значений во всей совокупности данных

Имя	Количество человек	
	Жен	Муж
Александр	20	**80**
Жанна	**95**	5
Наргиз	**92**	8
Хамзат-оглы	3	**97**
Юлия	**99**	1

Имя	Пол
Александр	Мужской
Жанна	Женский
Наргиз	Женский
Хамзат-оглы	Мужской
Юлия	Женский

Очистка: контрольные числа

В основе алгоритма **контрольных чисел** лежит расчет определенных функций, которые применяются для проверки правильности номеров банковских карт, ИНН, СНИЛС, ОКПО, ОКАТО, ОГРН

ИНН
12345678 9046

Контрольные числа	
4	7

Контрольные числа не совпадают

ИНН введен с ошибкой

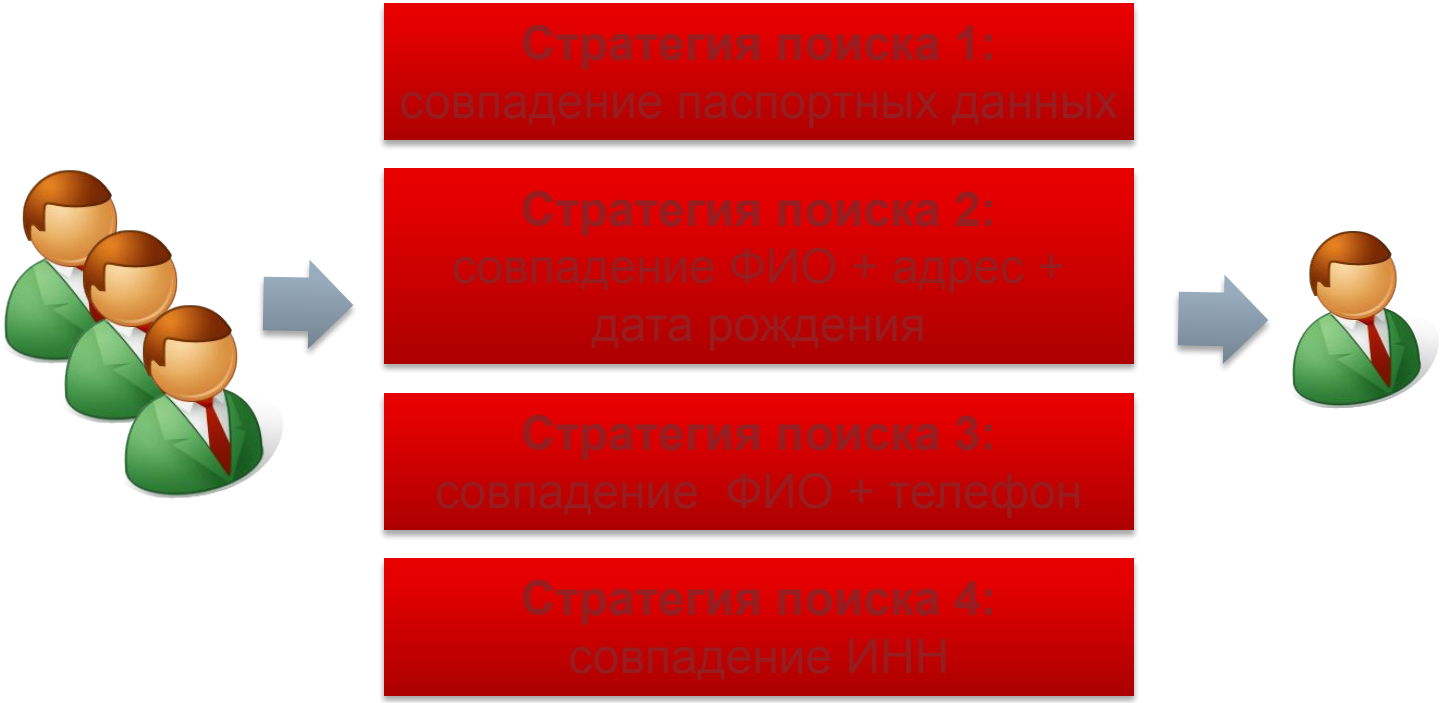
Очистка: схожесть строк

Анализ строк выявляет «похожесть» записей с помощью алгоритмов сравнения значений: метода Левенштейна, Soundex, нахождения максимальной общей подпоследовательности…

Имена из словаря	Расстояние Левенштейна
Игнатий	5
Игорь	2
Измаил	5
Изот	3
…	…

Очистка: дедубликация

Дедубликация основывается на поиске совпадающих и похожих объектов по определенным стратегиям с целью устранения повторов.



Очистка: другие методы

Для очистки данных используются и другие методы:

- Формализованные правила: накладывание заранее определенных правил очистки на контролируемые поля
- Способы замены: индексирование слов по их звучанию, кодирование…
- Проверка по статистическим значениям: по доверительному интервалу, средним значениям…
- Кластерный анализ: проверка написания значения с учетом попадания его в кластер…

Обогащение – процесс насыщения данных новой информацией, которая позволяет сделать их более ценными, значимыми и информативными с точки зрения решения той или иной аналитической задачи.

Задачи:

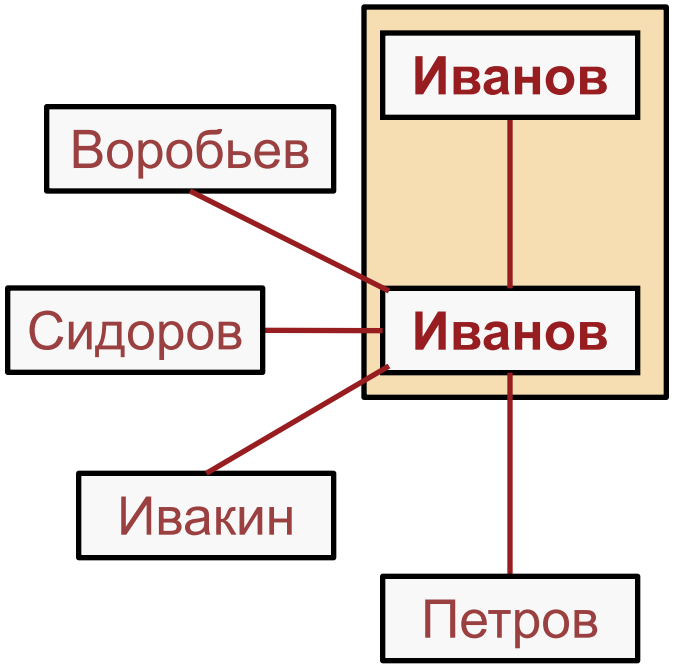
- Интеграция данных из множества источников
- Выявление связей между объектами
- Заполнение пропусков

Обогащение: анализ связей

Анализ связей исследует взаимосвязанные объекты и определяет закономерности между ними.

Обогащение: поиск близких объектов

Поиск близких объектов основывается на «схожести» значений признаков объектов.



Признак	Объект 1	Объект 2
Фамилия	Иванов	Иванов
Город	Рязань	г. Рязань
E-mail	i@mail.ru	i@mail.ru
Место работы	ООО «Русь» →	
Должность	Директор →	

Обогащение: другие методы

Обогащение данных предполагает применение и комбинирование множества методов:

- Реорганизация самих данных: введение кодировок, признаков состояний объектов, подразделение их на категории…
- Нечеткий поиск: восстановление пропусков с помощью нечетких запросов…
- Анализ источников данных: рейтингование источников данных по достоверности…

Результат


Повышение информационной насыщенности
Отсутствие ошибок, опечаток, дублей…
Качественные данные
Стандартизованное представление данных
Пригодность для дальнейшего анализа