ETF Working together Learning for life

# Большие данные для аналитики рынка труда

## День 1, занятие 2

Роль ИИ в системе сбора и анализа данных

Алессандро Ваккарино — Мауро Пелуччи

22 ноября 2021

burningglass TECHNOLOGIES

# Темы

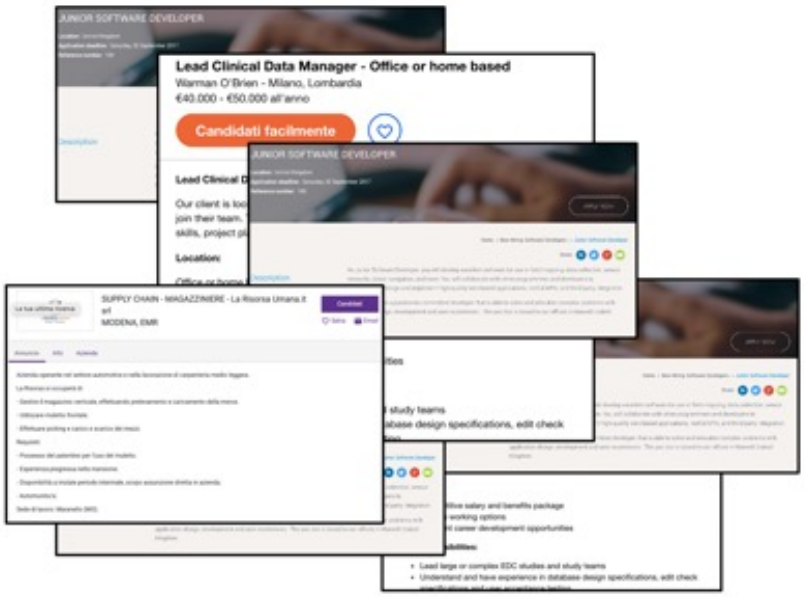
1. Краткое повторение
2. Система сбора и анализа данных
    1. Функциональная архитектура
    2. Методы приема данных
    3. Конвейер обработки данных
    4. Методы классификации

# Темы

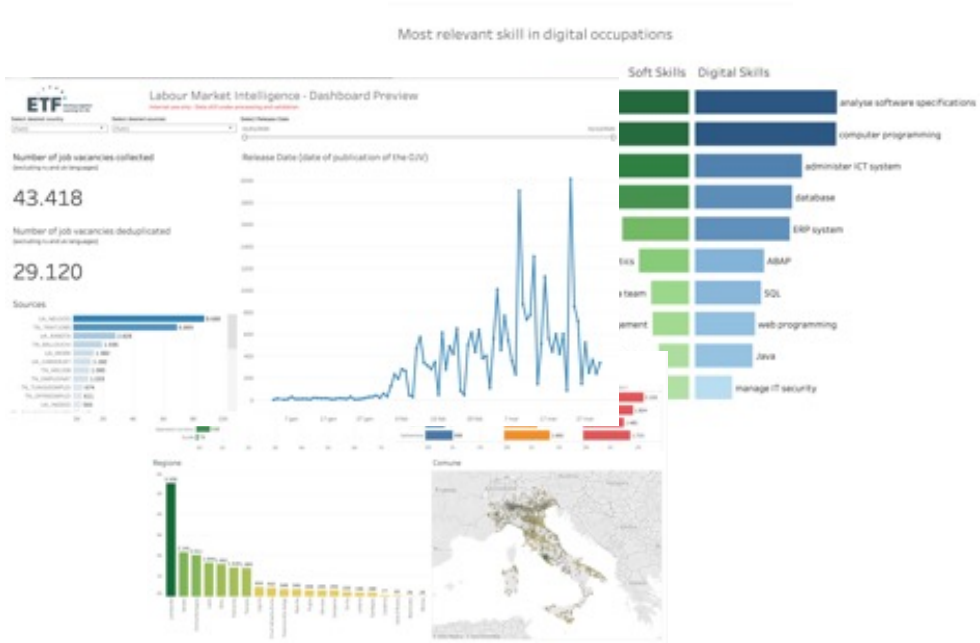
1. **Краткое повторение**
2. Система сбора и анализа данных
    1. Функциональная архитектура
    2. Методы приема данных
    3. Конвейер обработки данных
    4. Методы классификации

# Наша цель

Превратить онлайн-объявления о работе…




…в статистику и аналитику



# Задачи

- Обработка громадного **объема** данных, поступающих практически в режиме реального времени
- Данные из Интернета → Необходимость выявлять **шум** и уменьшать его количество
- **Многоязычная** среда
- Необходимость привязки к **стандартам классификации**
- Найти способ **сформулировать и представить** широкий и комплексный сценарий

# Наш подход

Обнаружение знаний в базах данных (ОЗБД) для аналитики рынка труда (АРТ)

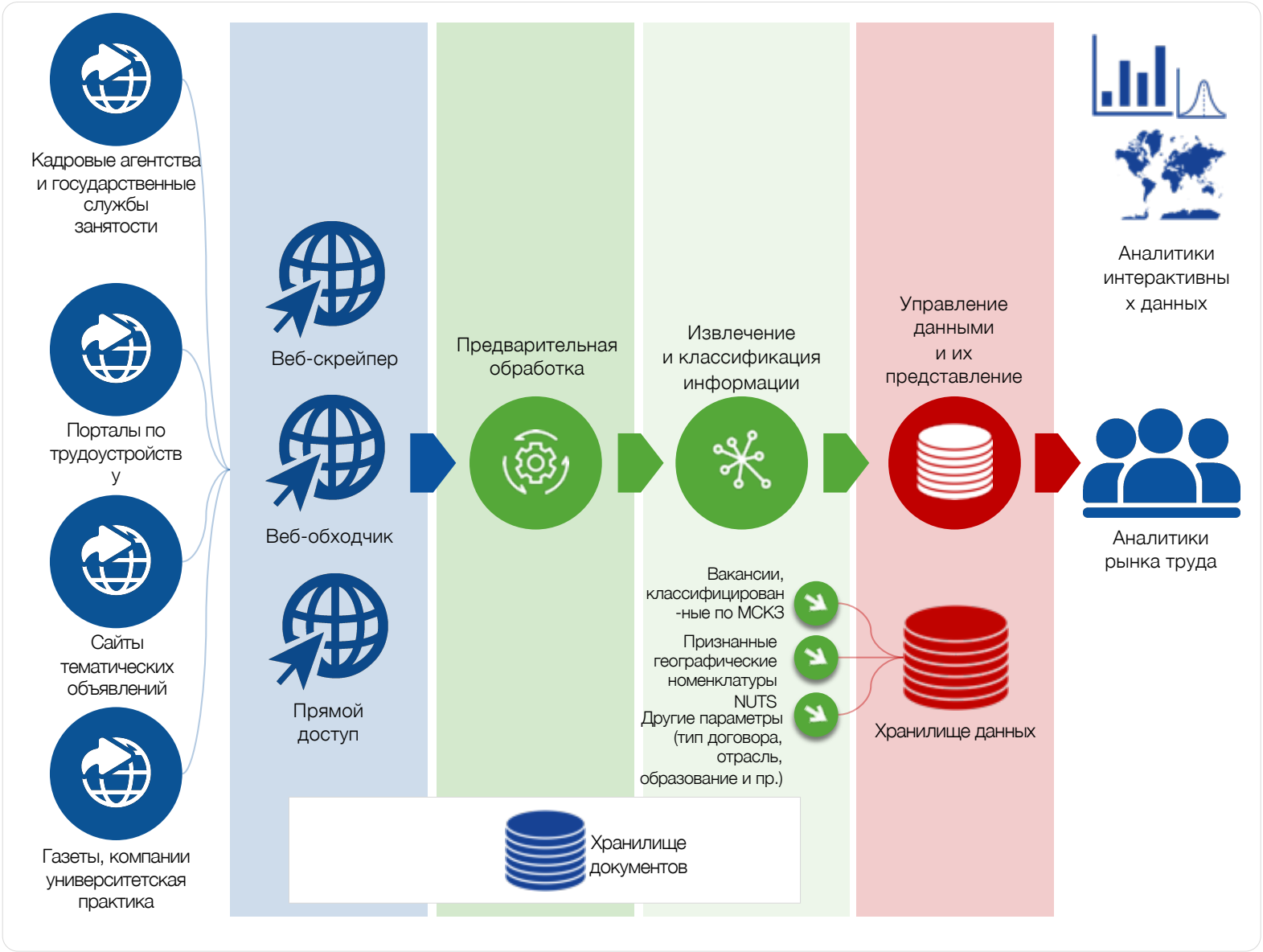
# Темы

1. Краткое повторение
2. Система сбора и анализа данных
    1. **Функциональная архитектура**
    2. Методы приема данных
    3. Конвейер обработки данных
    4. Методы классификации

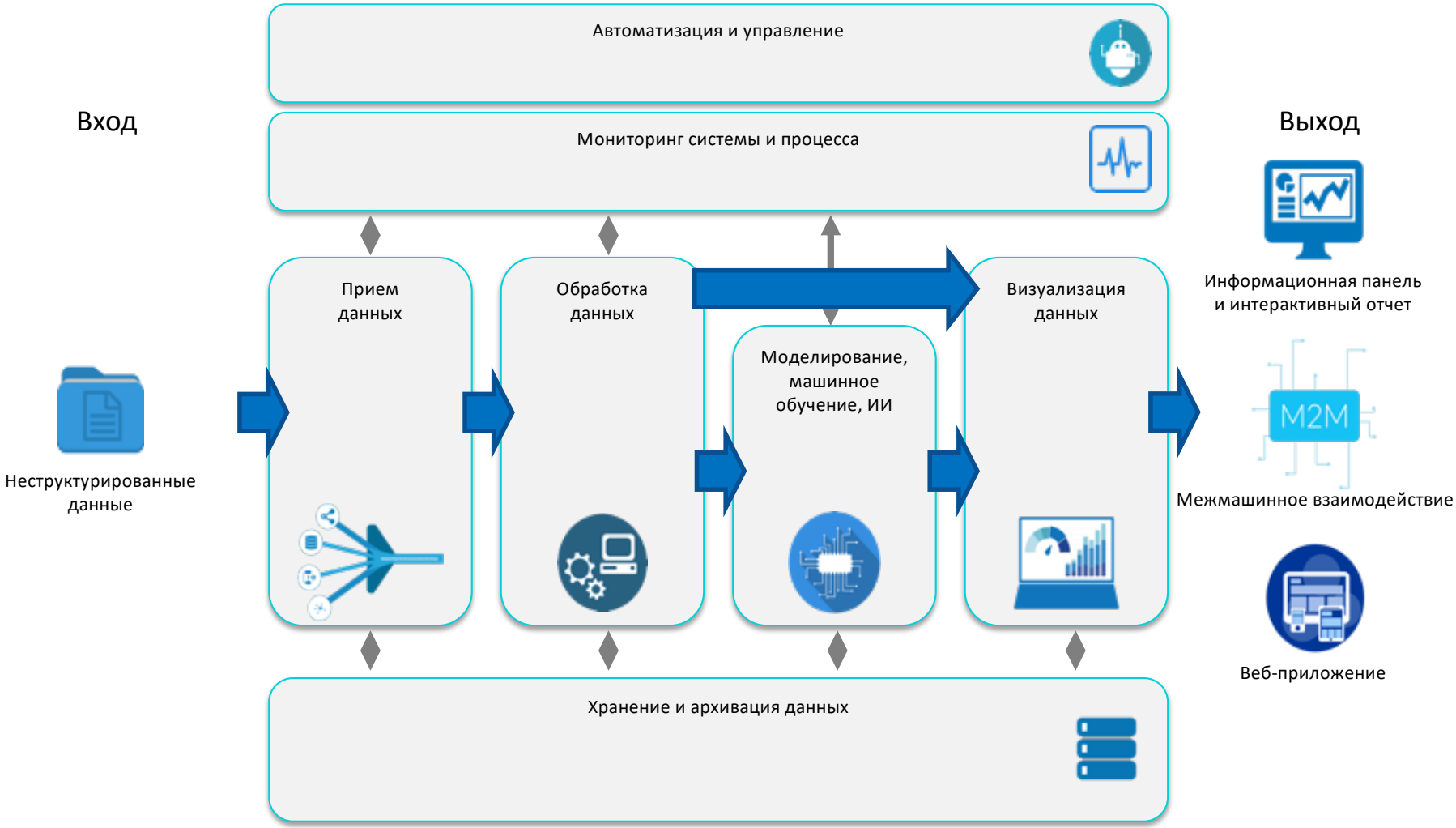
# Общий поток данных

# Концептуальная архитектура

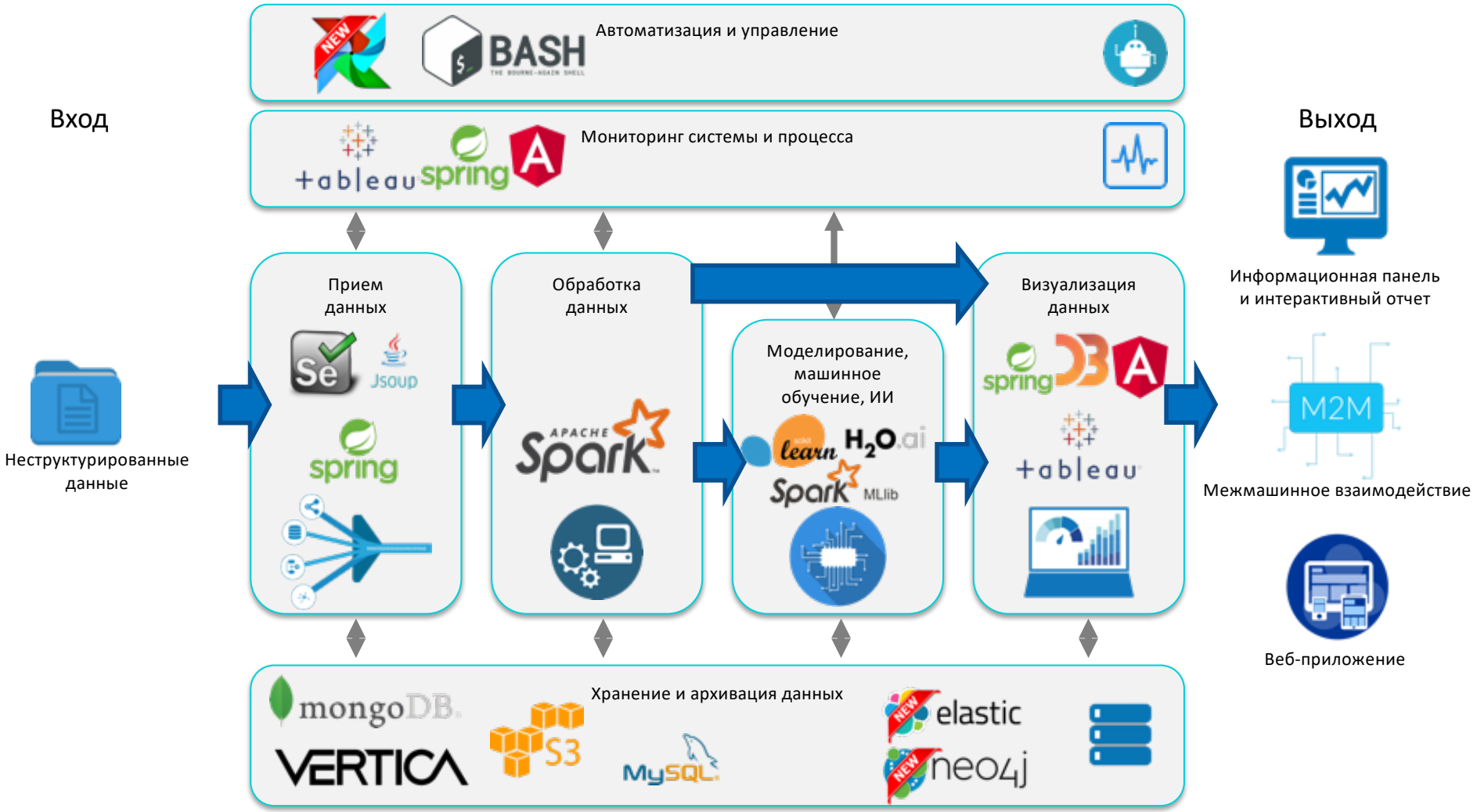
# Логическое представление



# Физическое представление



# Технологическое представление



# Темы

1. Краткое повторение
2. Система сбора и анализа данных
    1. Функциональная архитектура
    2. **Методы приема данных**
    3. Конвейер обработки данных
    4. Методы классификации

# Фаза приема данных

Процесс получения и импорта данных с веб-порталов

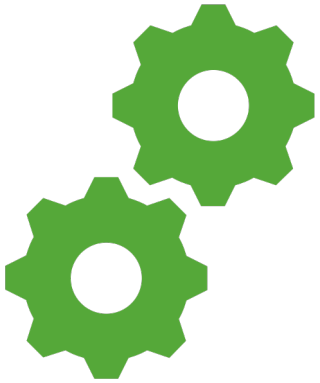
и их помещение в базу данных

Акцент на
объемах

Расширение и
и максимизация
охвата

Прямые соглашения
с наиболее
значимыми
источниками

# Задачи в связи с приемом данных



Надежность процесса

Качество собранных данных

Масштабируемость и управление

# Задачи в связи с приемом данных

## 1. Надежность

Проблема: потенциальные технические проблемы при сборе данных из источника (недоступность, блокирование, изменения в структуре данных)

Риск: потеря данных

Решение: избыточность

- Прием данных с самых важных сайтов (с точки зрения объема и/или охвата) минимум из двух источников
- Избегание потерь данных в случае проблем с источником
- Сбор данных из первичных и вторичных источников

# Задачи в связи с приемом данных

## 2. Качество

Проблема: необходимость получения как можно более чистых данных и выявления структурированных данных (при наличии)

Риск: потеря качества

Решение: прием с учетом особенностей источника. Мы собираем данные, используя индивидуальный подход к каждому конкретному источнику:

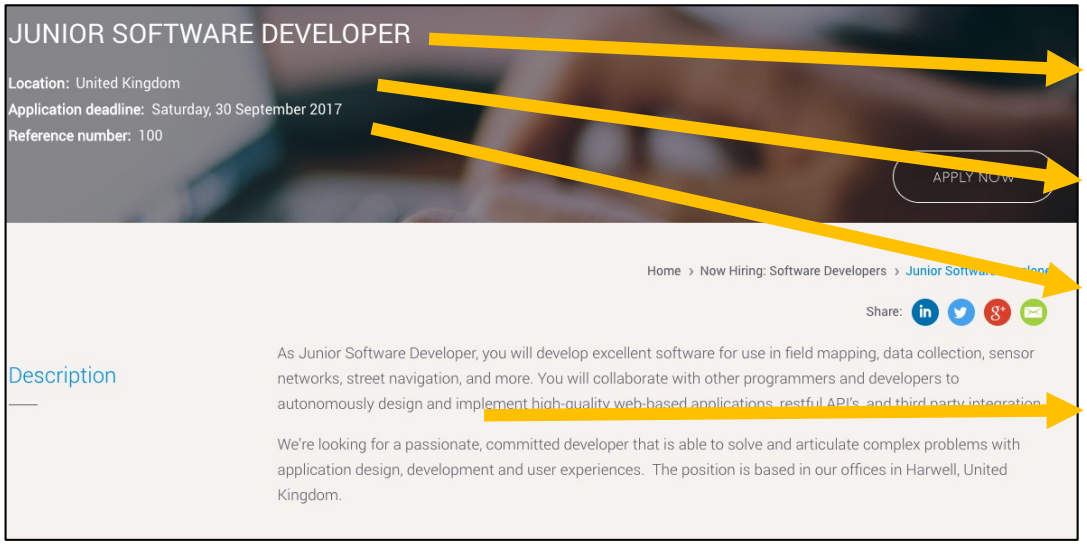
- интерфейс прикладного программирования (API)
- скрейпинг
- обход контента

# Задачи в связи с приемом данных — качество

- API: когда возможно (соглашения), мы собираем с веб-порталов преимущественно структурированные данные.
  - Преимущества: очень высокое качество (большинство полей структурированы)
  - Недостатки: требуются соглашения, которые не всегда есть
- Скрейпинг: если API использовать нецелесообразно, а структура веб-портала единообразна, мы разрабатываем специальный скрейпер, извлекающий со страниц структурированные и неструктурированные данные
  - Преимущества: высокое качество (много структурированных полей)
  - Недостатки: требуется разработка под определенный веб-портал
- Обход контента: если структура страниц веб-портала не единообразна, мы осуществляем прием данных с использованием многоцелевого обхода
  - Преимущества: более низкое качество (нет структурированных полей)
  - Недостатки: быстрый и универсальный подход

# Скрейпинг — пример

Веб-скрейпинг — это агрегация данных для извлечения их с веб-сайтов в структурированной форме

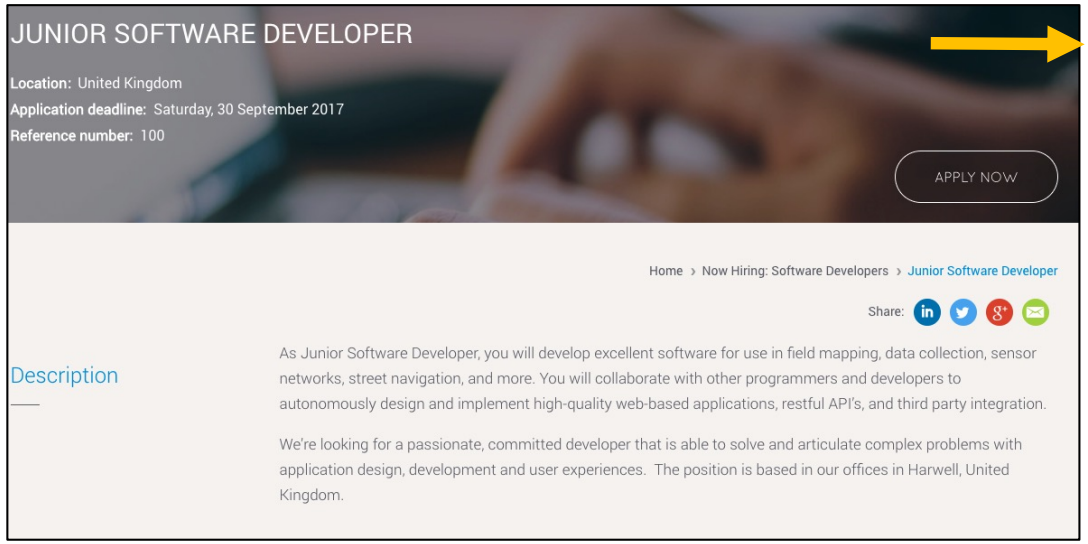



| | |
|---|---|
| Должность: | Младший разработчик ПО |
| Регион: | Великобритания |
| Время: | Суббота, 30 сентября 2017 г. |
| Описание: | В качестве младшего разработчика ПО вы будете разрабатывать прекрасное ПО для использования… |

# Обход контента (краулинг) — пример

Веб-обходчик (краулер) — это бот, который систематически просматривает веб-порталы с целью загрузки всех их страниц.

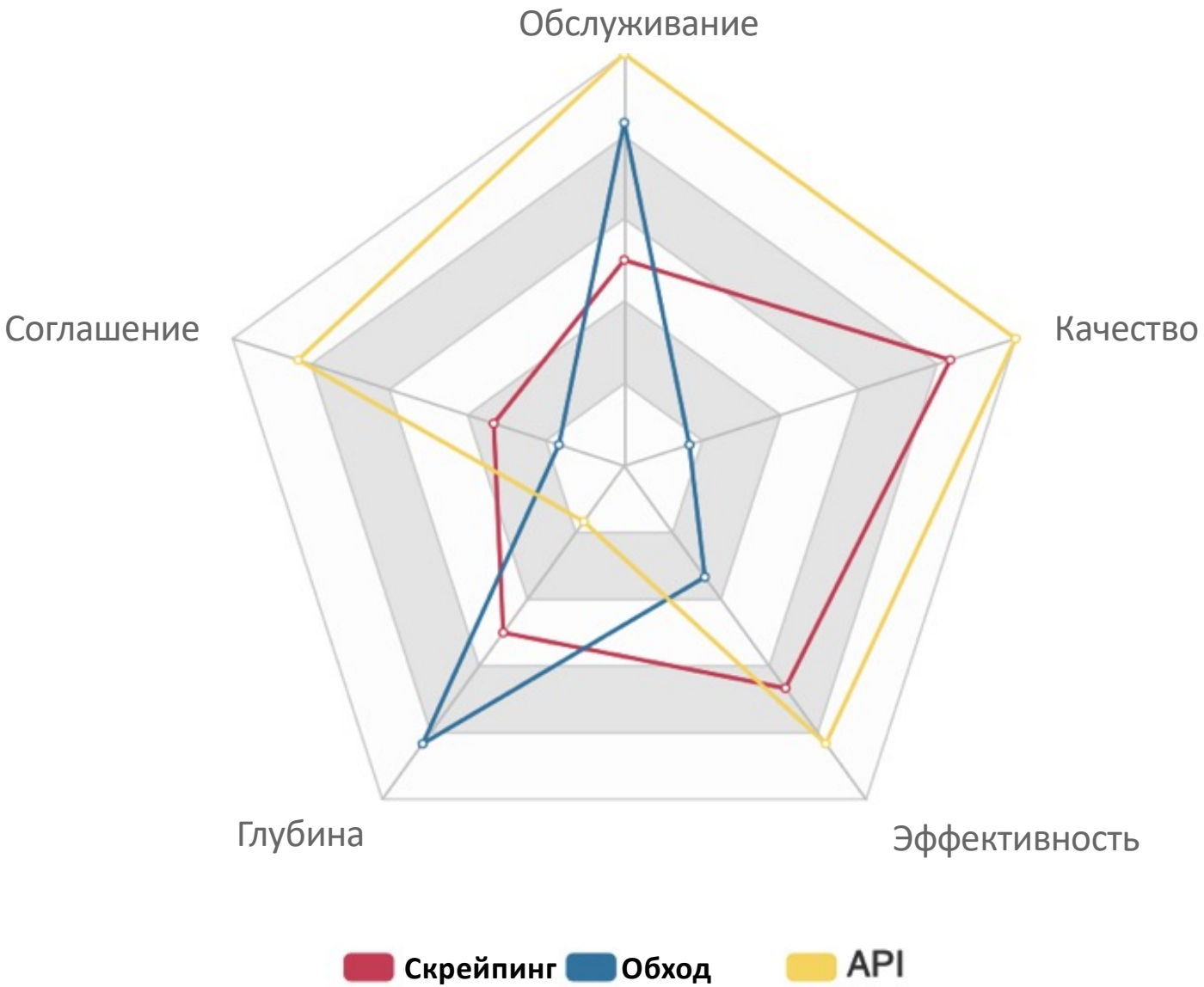
Обход контента (краулинг) — это самый распространенный способ получения больших массивов информации из Интернета с помощью поисковых роботов (например, GoogleBot)




Веб-страница:

```
<!DOCTYPE html>
  <head>
    <meta name="должность" content="Младший
разработчик ПО" />
  </head>
  <body>
    <header>
      <h2>Младший разработчик ПО</h2>

<div><div>Регион</div>Великобритания</div>
    …
    </header>
    <div><div>Описание</div>
    <span>В качестве младшего разработчика ПО
вы будете разрабатывать прекрасное ПО для
использования…
```


Обслуживание
Качество
Эффективность
Глубина
Соглашение
Скрейпинг
Обход
API

# Задачи в связи с приемом данных

## 3. Масштабируемость и управление

Проблема: необходимость работать в реальной и сложной среде больших данных с одновременным подключением к тысячам веб-сайтов

Риск: потеря контроля над процессом и потеря данных онлайн-вакансий из-за его медлительности

Решение:

- масштабируемая инфраструктура
- специальный инструмент для мониторинга и управления

# Задачи в связи с приемом данных — масштабирование

Мы разработали решение на основе микросервисов, с помощью которого можно при необходимости создавать и удалять «виртуальные поисковые компьютеры» . Каждый компьютер имеет множество браузеров, способных имитировать навигацию человека на веб-сайтах.

Основные отличия от настоящих компьютеров:

1. Отсутствие мониторов, сохранение страниц в нашем озере данных
2. Масштабируемость в обоих направлениях по мере необходимости

# Резюме и ключевые слова

- Изучение источников, их выбор и увеличение охвата
- Оптимизированный подход
  - Компоненты API, скрейпинга, обхода контента
- Акцент на объемах
  - Масштабирование и сбор данных в режиме реального времени
- Мониторинг собранных данных в режиме реального времени

# Темы

1. Краткое повторение
2. Система сбора и анализа данных
    1. Функциональная архитектура
    2. Методы приема данных
    3. **Конвейер обработки данных**
    4. Методы классификации

# Предварительная обработка данных — задачи и определения

- Цель:
  - Обеспечить ввод подходящих данных на этапе извлечения информации
- Задачи:
  - Измерение, мониторинг и повышение качества данных для повышения полноты, единообразия, комплексности, своевременности и периодичности
- Подход:
  - Разработка многофазного конвейера, нацеленного на:
    - выявление вакансий: анализировать страницу веб-сайта, чтобы выделить только то содержимое, которое относится к вакансиям
    - дедупликация: выявлять дублирующиеся публикации вакансий, чтобы получить единую запись вакансии
    - обнаружение даты: определять даты выпуска вакансий и истечения их срока посредством анализа описания вакансии
    - длительность вакансии: метод определения даты истечения срока, если она не указано прямо
- Особенности:
  - Гарантированное качество данных на всех фазах обработки

# Предварительная обработка данных — задачи и определения

Процесс очистки принимаемых данных и дедупликация онлайн-вакансий, чтобы во время аналитической фазы обрабатывались данные как можно более высокого качества

**Определение** языка

Снижение шума

**Дедупликация** онлайн-вакансий

# Этапы предварительной обработки

Объединение

Очистка

Обработка и резюмирование текста

# Предварительная обработка данных
# Определение языка

- Зачем:
  - В каждом языке свои ключевые слова, стоп-слова…
  - Он может отражать разные культуры и сценарии рынка труда…
  - ...Поэтому крайне важно определить язык онлайн-вакансии, чтобы использовать наиболее подходящий конвейер классификации
- Как:
  - Для каждого языка (свыше 60) мы обучили отдельный классификатор на основе корпуса Википедии
  - Полученные модели отличаются высокой точностью (~99 % точности) и возможностью быстрого внедрения в конвейер обработки данных
- Что мы получаем:
  - Быструю и надежную классификацию языка, используемого в каждой онлайн-вакансии
  - Способ архивировать те онлайн-вакансии, для которых у нас нет конвейера классификации

# Предварительная обработка данных
# Как справляться с шумом?

- В среде больших данных приходится иметь дело с шумом
  - Почему? Потому что информация собирается из Интернета, одного из самых зашумленных мест
- Во-первых, нам нужно понять, с каким типом шума нам придется столкнуться...:
  - Веб-страницы, не имеющие прямого отношения к онлайн-вакансиям:
    - страницы социальных сетей
    - страницы новостей
    - страницы с политикой конфиденциальности
    - ...
  - Веб-страницы, которые выглядят, как онлайн-вакансии:
    - обучающие курсы
    - резюме
    - консультационные услуги
    - ...

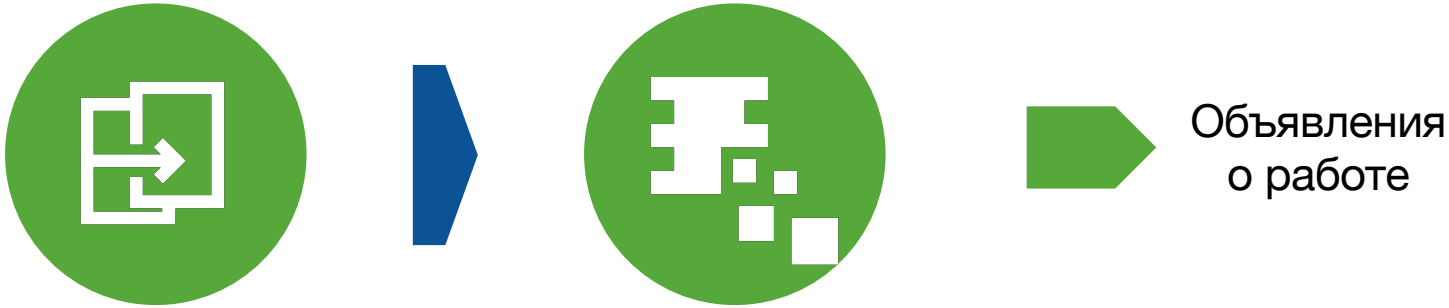

- ...Затем нам нужно выявить и обработать дублирующиеся онлайн-вакансии:
  - Обычно одну вакансию размещают на нескольких порталах
  - Если мы рассмотрим их как отдельные, то переоценим спрос на рынке труда
  - Поэтому нам нужно выявлять дублирующиеся онлайн-вакансии и объединять информацию из них в одну запись

# Предварительная обработка данных Выявление шума — как?

- Подход в 2 этапа:
  - Метод машинного обучения
    - Для каждого языка мы подготовили наивный байесовский классификатор, обучив его на более чем 20 тыс. веб-страниц:
      - 10 тыс. страниц, связанных с реальными онлайн-вакансиями
      - 10 тыс. страниц, не связанных с онлайн-вакансиями
    - Точность ~99 %
    - Быстро обучается и используется
    - Подход похож на систему обнаружения спама в электронной почте
  - Метод нечетких соответствий
    - Используется для выявления веб-страниц, похожих на онлайн-вакансии, но относящихся к предложениям обучающих курсов, консультационных услуг и т. д.
    - Сканируется заголовок и основной текст страницы с объявлением на предмет ключевых слов (в зависимости от языка), благодаря чему можно пометить страницу как «не имеющую отношения к онлайн-вакансии»

Но до начала фазы дедупликации онлайн-вакансий нам нужно очистить текст, чтобы упростить и консолидировать его…

# Предварительная обработка данных
# Фаза дедупликации




**Физическая дедупликация или определение нечетких соответствий**

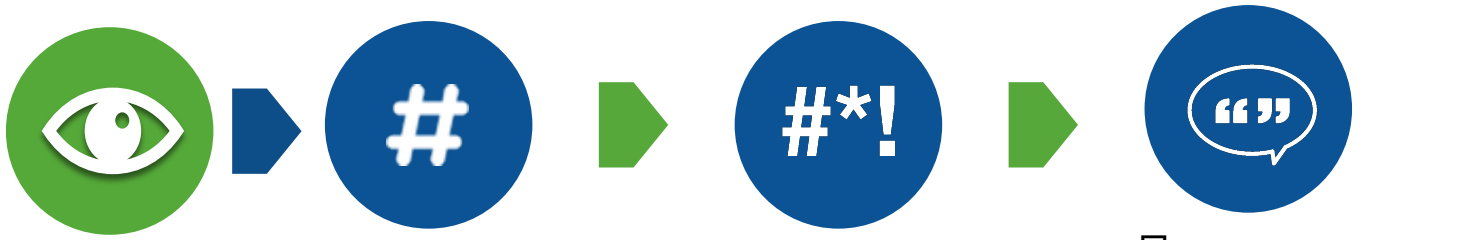
Проводится на основе той части вакансии, в которой содержится ее описание (или содержание).

**Сопоставление метаданных**

Использование метаданных с порталов по трудоустройству для удаления дубликатов вакансий на сайтах-агрегаторах (например, контрольный идентификатор, URL-адрес страницы)

# Обработка и резюмирование текста

Фаза обработки и резюмирования текста нацелена на его сокращение с целью улучшения процесса классификации вакансий по европейским стандартам.




Определитель языка

Текст объявления о работе

Устранение шума и обработка

Представление векторно-пространственной модели

JUNIOR SOFTWARE DEVELOPER

**Location:** United Kingdom
**Application deadline:** Saturday, 30 September 2017
**Reference number:** 100

Description

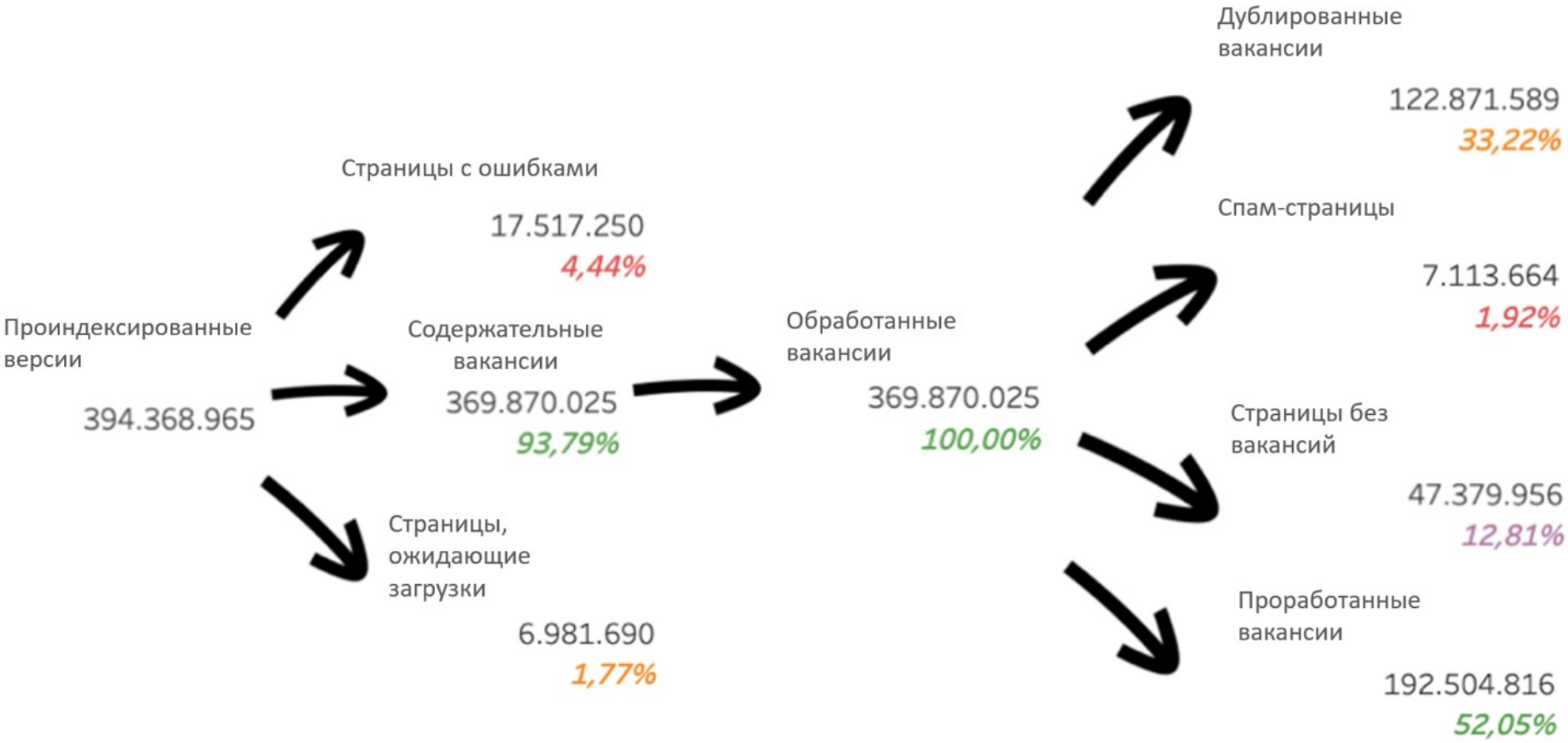
As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

В качестве младшего ⟨разработчика ПО⟩, вы будете разрабатывать прекрасное ⟨программное обеспечение⟩ для использования в сфере ⟨сопоставления полей⟩, ⟨сбора данных⟩, ⟨сенсорных сетей⟩, ⟨уличной навигации⟩, и многих других. Вы будете ⟨сотрудничать⟩ с другими ⟨программистами⟩ и ⟨разработчиками⟩ в целях ⟨самостоятельного⟩ проектирования и внедрения высококачественных ⟨веб-приложений⟩, ⟨API⟩, соответствующий ограничениям REST, и обеспечения ⟨интеграции⟩ сторонних решений.

Мы ищем увлеченного, преданного делу ⟨разработчика⟩, умеющего ⟨решать⟩ и формулировать ⟨сложные задачи⟩ в сфере ⟨проектирования приложений⟩, ⟨разработки⟩ и ⟨взаимодействия с пользователем⟩. Работа в нашем офисе в ⟨Харвелле⟩, ⟨Великобритания⟩.

# Предварительная обработка данных — Результаты



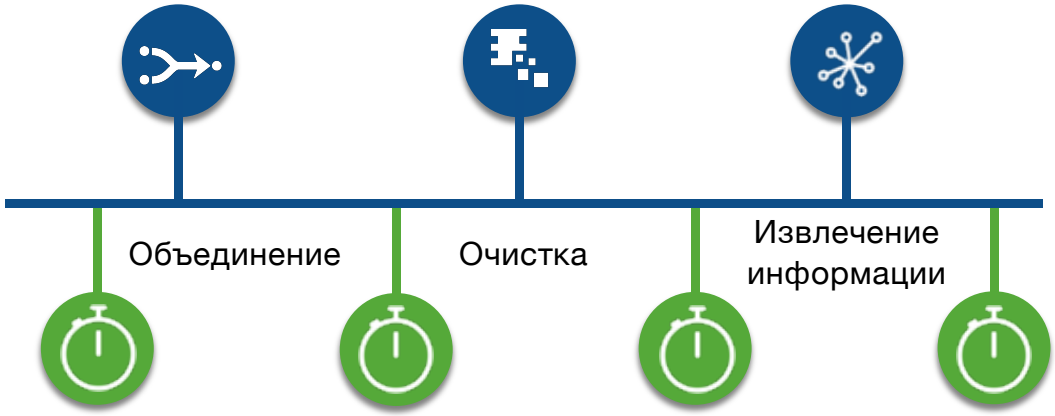
# Предварительная обработка данных
# Что делать с шумом?

## Мы не удаляем шум физически

Мы собираем его, чтобы отслеживать процесс в целом, и фиксируем:

- Тип шума → Чтобы определить необходимость разработки процесса более глубокой проверки качества
- Тенденции шума → Для выявления источников, повышающих или понижающих количество шума, и работы с ними
- Аналитические цели → Анализ культурных сред конкретных стран, например использование портала по трудоустройству для рекламы обучающих курсов
- Мониторинг → Для отслеживания процесса в целом

Как мы отслеживаем качество данных?




Постоянные точки проверки качества

# Резюме и ключевые слова

- Акцент на качество
  - Как устранить шум?
  - Действия по дедупликации
- Задачи, связанные с языками
  - Специально подобранный компонент для каждого языка
- Отслеживание качества данных
  - Постоянная проверка качества и точки проверки

# Темы

1. Краткое повторение
2. Система сбора и анализа данных
    1. Функциональная архитектура
    2. Методы приема данных
    3. Конвейер обработки данных
    4. **Методы классификации**

# Классификация данных

- Цель:
  - Извлечь и структурировать информацию из данных для передачи на уровень представления
- Задачи:
  - Обработка огромного массива неоднородных данных на разных языках
- Подход:
  - Разработать адаптируемую схему с учетом языка, подстроенную к различным особенностям информации. Некоторые актуальные задачи:
    - Классификация по **профессии**: комбинированные методы, такие как машинное обучение, тематическое моделирование, обучение без учителя
    - Классификация по **профессиональным умениям**: другие различные комбинированные методы, такие как анализ текста с учетом сходства на основе корпуса или знаний
- Особенности:
  - Гарантировать извлечение объяснимой информации, регистрацию методов классификации и соответствующие особенности.

# Классификация данных — пример


Младший разработчик ПО

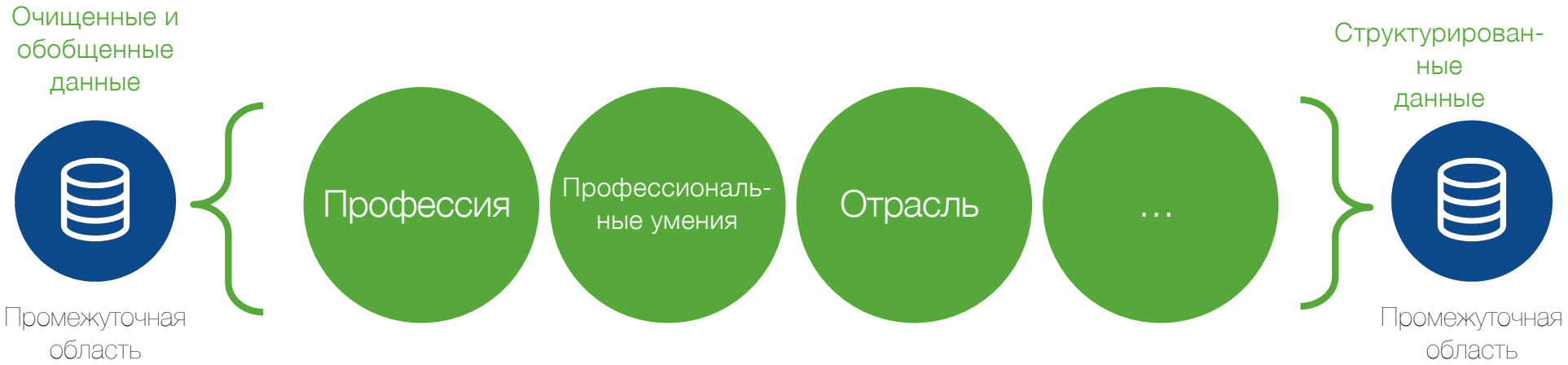
В качестве младшего разработчика ПО, вы будете разрабатывать прекрасное программное обеспечение для использования в сфере сопоставления полей, сбора данных, сенсорных сетей, уличной навигации, и многих других. Вы будете сотрудничать с другими программистами и разработчиками в целях самостоятельного проектирования и внедрения высококачественных веб-приложений, API, соответствующих ограничениям REST, и обеспечения интеграции сторонних решений.

Мы ищем увлеченного, преданного делу разработчика, умеющего решать и формулировать сложные задачи в сфере проектирования приложений, разработки и взаимодействия с пользователем.  Работа в нашем офисе в Харвелле, Великобритания.

# Извлечение и классификация информации Аналитика рынка труда в реальном времени

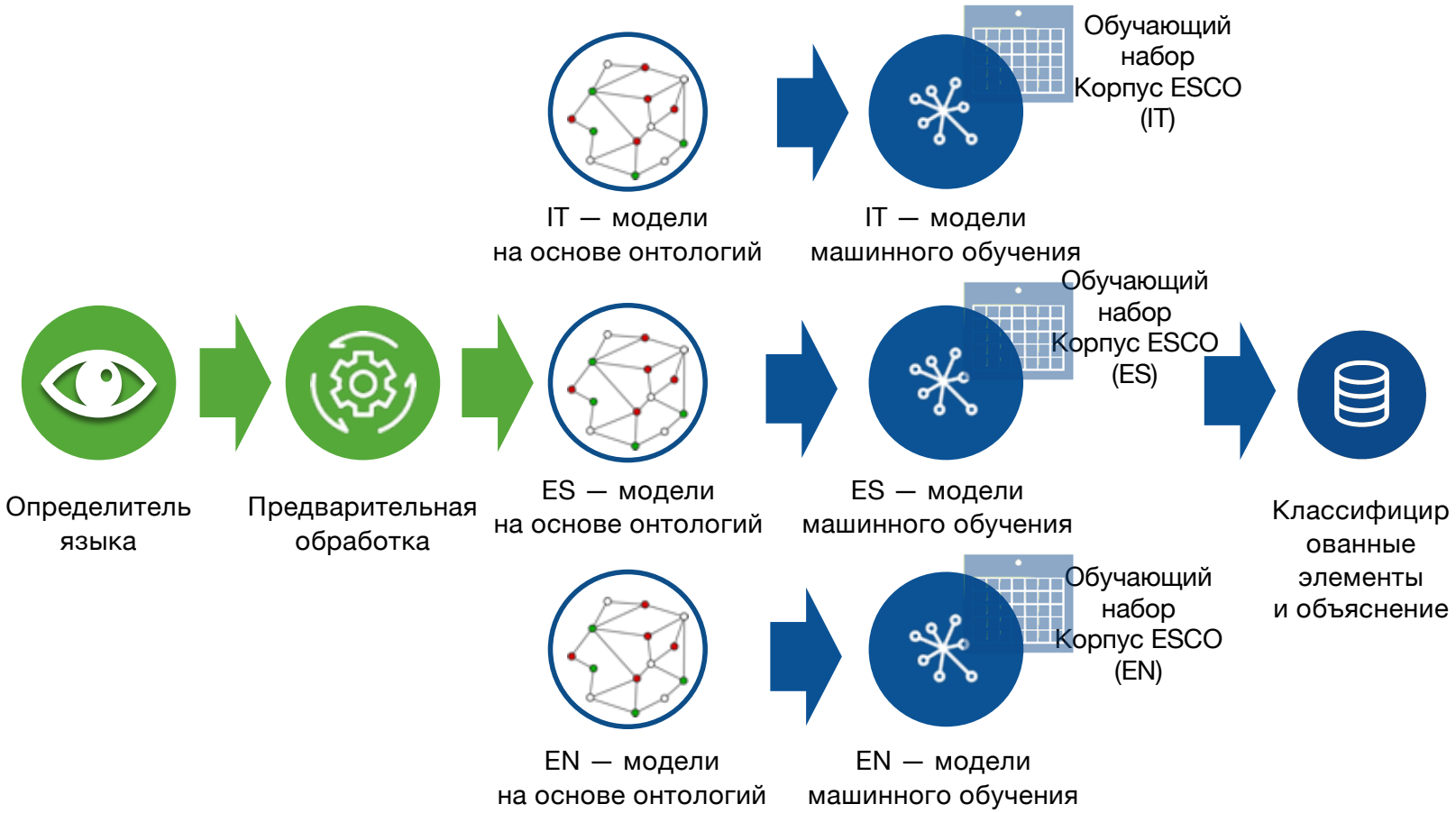
Извлечение информации — это сфера обработки естественного языка, которая связана с поиском фактической информации в произвольном тексте.

Для выполнения этой задаче используются методы машинного обучения (обучение на основе онтологий, обучение с учителем и без учителя), чтобы сопоставлять объявления о работе со стандартными классификациями.




Машинное обучение → Обучение на основе онтологий, обучение с учителем и без учителя, и др.

# Классификация




Что означает «модели на основе онтологий»?

Как мы можем использовать онтологии для классификации?

# Конвейер профессий

# Вопросы касательно классификатора профессий

- Обучение на основе онтологий + обучение с учителем
  - онтология ESCO
  - новые метки из тематического моделирования
- Одна модель для каждого языка
- Данные, помеченные экспертом из каждой страны
  - около 100 тыс. объявлений о работе (очищенный обучающий набор с использованием нашей онтологии)
  - 436 возможных целей
- Набор для оценивания — 20 % «золотого» набора данных объявлений о работе
  - взвешенная точность ~86 %
  - около 430 обнаруженных профессий

# Подходы на основе схожести текста


Измерение схожести строк оперирует последовательностями строк и комбинациями символов.

Сходство Джаро — Винклера, коэффициент Жаккара, косинусное сходство


Сопоставление по корпусу текстов — это мера семантической близости, по которой определяется сходство между словами в соответствии с информацией, полученной из большого корпуса текстов.

Латентно-семантический анализ, явный семантический анализ, дистрибуционно связанные слова с использованием совместного появления (инструмент DISCO)


Сопоставление на основе знаний заключается в определении степени схожести между словами с использованием информации, полученной из семантических сетей

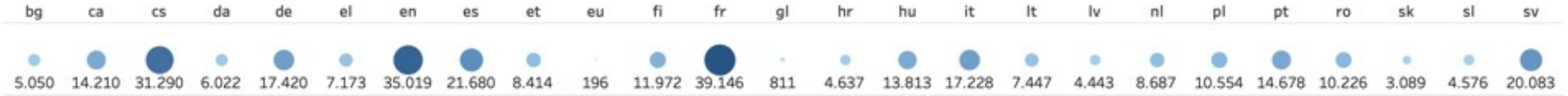
## Precision of occupation (overall)

86,66%

## Validation Set (overall)

317.864

## Validation Set by language



| bg | ca | cs | da | de | el | en | es | et | eu | fi | fr | gl | hr | hu | it | lt | lv | nl | pl | pt | ro | sk | sl | sv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5.050 | 14.210 | 31.290 | 6.022 | 17.420 | 7.173 | 35.019 | 21.680 | 8.414 | 196 | 11.972 | 39.146 | 811 | 4.637 | 13.813 | 17.228 | 7.447 | 4.443 | 8.687 | 10.554 | 14.678 | 10.226 | 3.089 | 4.576 | 20.083 |

## Precision of occupation by language

| bg | ca | cs | da | de | el | en | es | et | eu | fi | fr | gl | hr | hu | it | lt | lv | nl | pl | pt | ro | sk | sl | sv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 85,96% | 86,45% | 88,96% | 96,89% | 83,17% | 85,02% | 82,83% | 87,39% | 83,63% | 68,59% | 98,80% | 83,89% | 83,07% | 89,45% | 97,40% | 81,30% | 92,86% | 96,95% | 93,31% | 80,02% | 83,01% | 95,12% | 98,50% | 87,78% | 78,43% |

## Precision of occupation (lv1)

| | |
|---|---|
| Clerical support workers | 85,77% |
| Craft and related trades .. | 86,10% |
| Elementary occupations | 86,19% |
| Managers | 86,32% |
| Plant and machine operat.. | 86,29% |
| Professionals | 86,61% |
| Service and sales workers | 89,38% |
| Skilled agricultural, fores.. | 88,79% |
| Technicians and associate.. | 85,54% |

## Precision of occupation (lv2)

| | |
|---|---|
| Administrative and comm.. | 85,06% |
| Agricultural, forestry and .. | 80,82% |
| Assemblers | 84,87% |
| Building and related trad.. | 92,30% |
| Business and administrati.. | 85,66% |
| Business and administrati.. | 80,06% |
| Chief executives, senior o.. | 91,36% |
| Cleaners and helpers | 85,11% |
| Customer services clerks | 82,21% |
| Drivers and mobile plant .. | 86,49% |
| Electrical and electronic t.. | 74,60% |
| Food preparation assista.. | 89,08% |
| Food processing, wood w.. | 82,61% |
| General and keyboard cler.. | 97,20% |
| Handicraft and printing w.. | 89,65% |

## Precision of occupation (lv3)

| | |
|---|---|
| Administration professio.. | 86,21% |
| Administrative and specia.. | 84,92% |
| Agricultural, forestry and .. | 80,82% |
| Animal producers | 83,13% |
| Architects, planners, surv.. | 87,56% |
| Artistic, cultural and culin.. | 91,74% |
| Assemblers | 84,87% |
| Authors, journalists and li.. | 90,72% |
| Blacksmiths, toolmakers .. | 86,70% |
| Building and housekeepin.. | 90,33% |
| Building finishers and rel.. | 95,47% |
| Building frame and relate.. | 90,00% |
| Business services agents | 89,57% |
| Business services and ad.. | 79,10% |
| Car, van and motorcycle d.. | 90,40% |

## Precision of occupation (lv4)

| | |
|---|---|
| Accountants | 83,60% |
| Accounting and bookkeepi.. | 58,14% |
| Accounting associate prof.. | 85,65% |
| Actors | 93,41% |
| Administrative and execu.. | 84,32% |
| Advertising and marketin.. | 65,30% |
| Advertising and public rel.. | 71,63% |
| Aged care services manag.. | 78,81% |
| Agricultural and forestry .. | 94,55% |
| Agricultural and industria.. | 76,49% |
| Agricultural technicians | 81,32% |
| Air conditioning and refri.. | 85,95% |
| Air traffic controllers | 84,43% |
| Air traffic safety electroni.. | 95,52% |
| Aircraft engine mechanics.. | 79,61% |

# Резюме и ключевые слова

- Акцент на резюмировании
  - Как резюмировать данные и улучшить результаты наших аналитиков данных?
- Связь со стандартными классификациями
  - сравнение данных онлайн-вакансий с другими источниками
- Задачи, связанные с «золотыми» наборами данных (мощность множества, качество и разнообразие)
- Смешанные подходы
  - машинное обучение
  - обучение на основе онтологий
  - методы сопоставления текста и извлечения информации
- Жизненный цикл модели