ETF Working together Learning for life

Большие данные для аналитики рынка труда

День 3, занятие 7

К «умной аналитике рынка труда»
Задачи, лучшие практики, полученный опыт

Алессандро Ваккарино — Мауро Пелуччи

Ноябрь 2021

Emsi | burningglass

Темы

1. Цель и контекст
2. Задачи
    1. Анализ стороны спроса и предложения
    2. Визуализация
    3. Готовые решения на базе ИИ

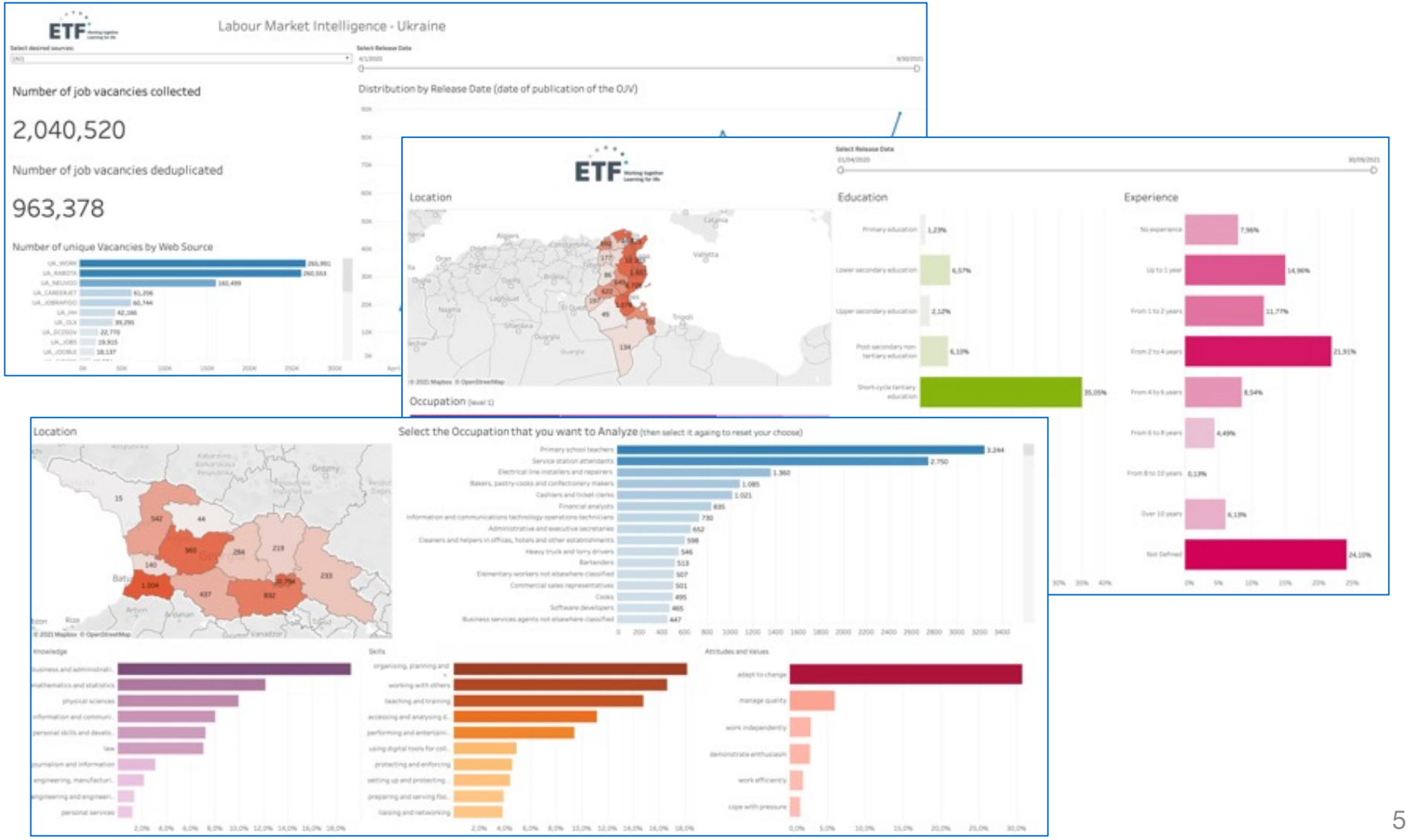
Темы

1. **Цель и контекст**
2. Задачи
    1. Анализ стороны спроса и предложения
    2. Визуализация
    3. Готовые решения на базе ИИ

Наша отправная точка

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

Наша отправная точка



...что теперь?

- Это отправная точка
  - Большие данные могут стать «золотой жилой» информации.
  - То, что мы с вами видели, лишь отправная точка, описание того, что происходит.
  - Определив методологию и систему, способную собирать и классифицировать большие данные, мы можем увидеть больше и попробовать совершенно новые виды анализа.

Новые виды анализа?

- Что мы можем получить?
  - Интеграция новых данных
  - Распространение информации, адаптированной под потребности разных заинтересованных лиц
  - Повторное использование существующих компонентов и знаний
  - Определение нового *ракурса анализа*

Рассмотрим некоторые примеры

Темы

1. Цель и контекст
2. Задачи
    1. **Анализ стороны спроса и предложения**
    2. Визуализация
    3. Готовые решения на базе ИИ

Зачем?

- Анализ потребностей — один из углов рассмотрения рынка труда
- Анализируя сторону предложения, мы можем получить дополнительные и добавочные сведения
  - Предлагаемые профессиональные умения
  - Сопоставление спроса и предложения
  - Эволюция профессиональных профилей
  - …

Как?

- Нам нужно найти дополнительный источник информации, который сможет помочь понять, как развивается сторона предложения
- Этот источник данных должен охватывать:
  - профессиональные профили
  - профессии
  - профессиональные умения
  - …

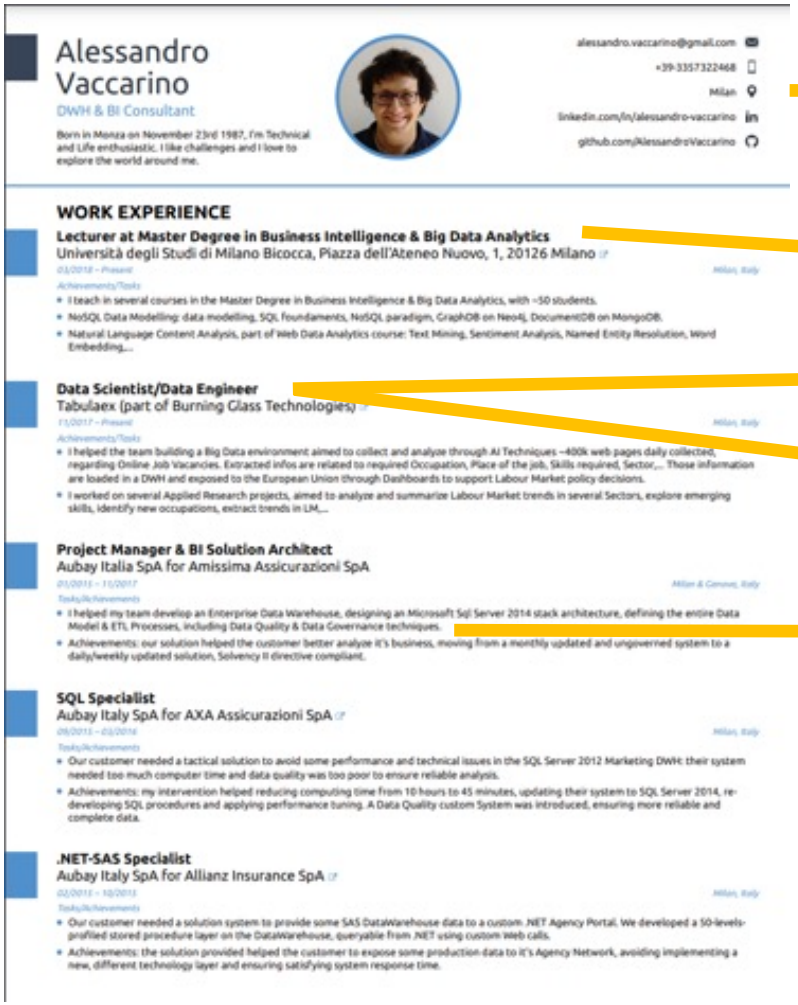
Какие источники могут предоставить подобную информацию?

Источник данных со стороны предложения

- Резюме
  - подробные
  - связаны с действительностью
  - обновленные
  - **не**структурированные
- Профили в социальных сетях (например, LinkedIn)
  - подробные
  - связаны с действительностью
  - обновленные
  - **частично** структурированные
  - частный источник (требуется прием данных)

Схема обработки данных на стороне предложения

Извлечение информации



Alessandro Vaccarino

DWH & BI Consultant

Born in Monza on November 23rd 1987, I'm Technical and Life enthusiastic. I like challenges and I love to explore the world around me.

alessandro.vaccarino@gmail.com
+39-3357322468
Milan
linkedin.com/in/alessandro-vaccarino
github.com/AlessandroVaccarino

WORK EXPERIENCE

Lecturer at Master Degree in Business Intelligence & Big Data Analytics
Università degli Studi di Milano Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano

- I teach in several courses in the Master Degree in Business Intelligence & Big Data Analytics, with ~50 students.
- NoSQL Data Modelling: data modelling, SQL foundaments, NoSQL paradigm, GraphDB on Neo4j, DocumentDB on MongoDB.
- Natural Language Content Analysis, part of Web Data Analytics course: Text Mining, Sentiment Analysis, Named Entity Resolution, Word Embedding,...

Data Scientist/Data Engineer
Tabulaex (part of Burning Glass Technologies)

- I helped the team building a Big Data environment aimed to collect and analyze through AI Techniques ~400k web pages daily collected, regarding Online Job Vacancies. Extracted infos are related to required Occupation, Place of the job, Skills required, Sector,... Those information are loaded in a DWH and exposed to the European Union through Dashboards to support Labour Market policy decisions.
- I worked on several Applied Research projects, aimed to analyze and summarize Labour Market trends in several Sectors, explore emerging skills, identify new occupations, extract trends in LM,...

Project Manager & BI Solution Architect
Aubay Italia SpA for Amissima Assicurazioni SpA

- I helped my team develop an Enterprise Data Warehouse, designing an Microsoft Sql Server 2014 stack architecture, defining the entire Data Model & ETL Processes, including Data Quality & Data Governance techniques.
- Achievements: our solution helped the customer better analyze it's business, moving from a monthly updated and ungoverned system to a daily/weekly updated solution, Solvency II directive compliant.

SQL Specialist
Aubay Italy SpA for AXA Assicurazioni SpA

- Our customer needed a tactical solution to avoid some performance and technical issues in the SQL Server 2012 Marketing DWH: their system needed too much computer time and data quality was too poor to ensure reliable analysis.
- Achievements: my intervention helped reducing computing time from 10 hours to 45 minutes, updating their system to SQL Server 2014, re-developing SQL procedures and applying performance tuning. A Data Quality custom System was introduced, ensuring more reliable and complete data.

.NET-SAS Specialist
Aubay Italy SpA for Allianz Insurance SpA

- Our customer needed a solution system to provide some SAS DataWarehouse data to a custom .NET Agency Portal. We developed a 50-levels-profiled stored procedure layer on the DataWarehouse, queryable from .NET using custom Web calls.
- Achievements: the solution provided helped the customer to expose some production data to it's Agency Network, avoiding implementing a new, different technology layer and ensuring satisfying system response time.

Регион: → Милан, Италия

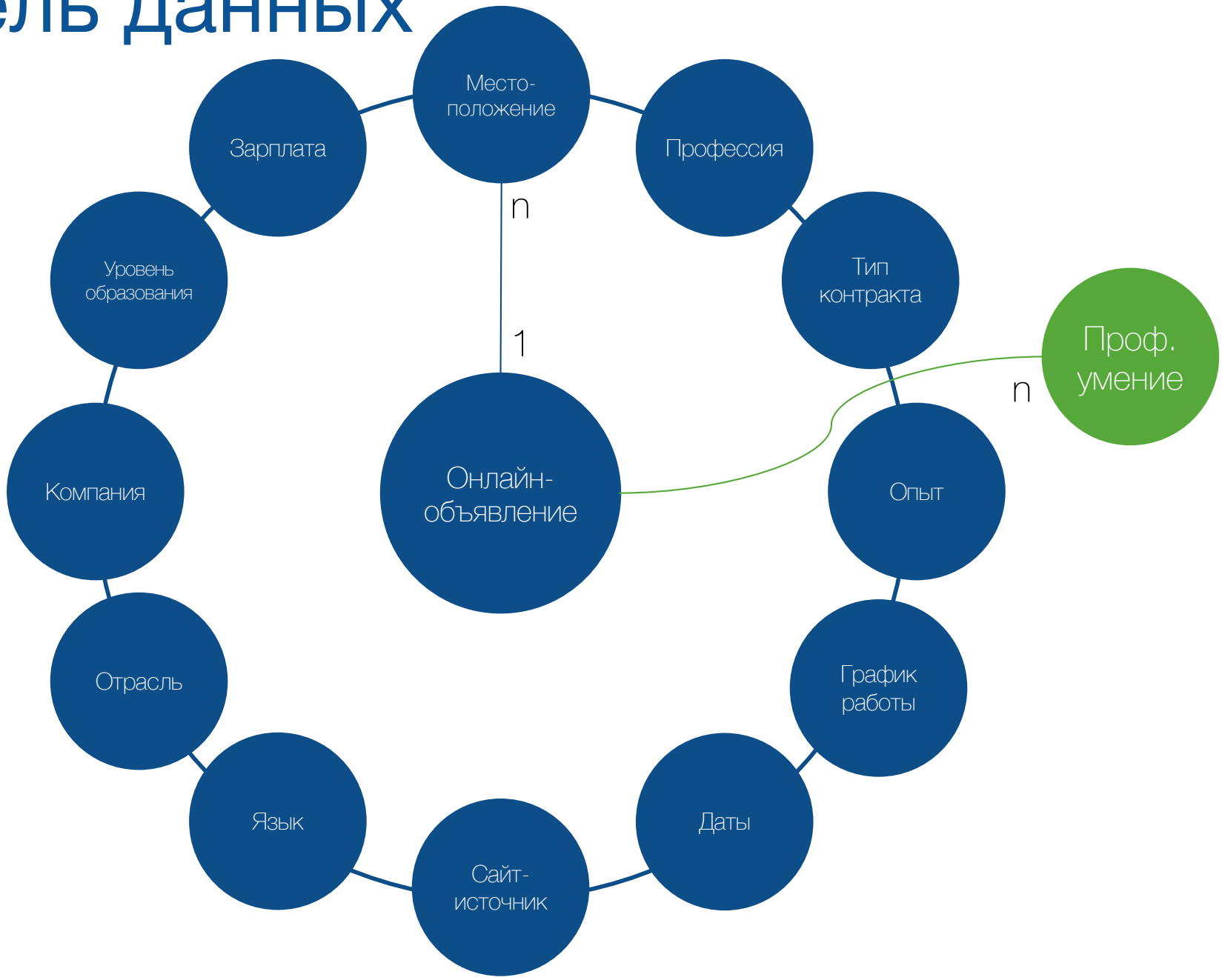
Должность: → Преподаватель

Должность: → Специалист по обработке данных

Должность: → Инженер по работе с данными

Проф. умения: → Управление данными, качество данных, процессы ETL, управление проектами, …

Модель данных



Возможные пути анализа данных — 1

- Анализ стороны предложения
  - наиболее предлагаемые профессии
  - самые популярные проф. умения
  - географические местоположения с самым высоким уровнем доступности специалистов
  - …

См. пример итальянского агентства по трудоустройству, чтобы увидеть, как это работает

Возможные пути анализа данных — 2

- Сопоставление спроса и предложения
  - Сравнение наборов профессиональных умений, предлагаемых специалистами и требуемых рынком
  - Интеграция с дополнительными источниками, поступающими от образовательных учреждений (например, от университетов)
  - Выявление дисбаланса между профессиональными умениями и траекториями обучения с целью внедрения правильных стратегий по:
    - повышению квалификации
    - переподготовке

См. пример итальянского университета, чтобы увидеть, как это работает

Возможности и задачи

- Профили в социальных сетях и резюме, а также размещенные объявления о работе, дают:
  - актуальные профессиональные профили и умения
  - глубокую информацию
  - актуальные сведения

- Кроме того, лексикон можно связать с теми же классификациями, что приняты для анализа потребностей, что обеспечит согласованность

Возможности и задачи

- С другой стороны, по сравнению со стороной спроса, здесь труднее обеспечить доступность данных
  - Объявления о работе, по своей природе, должны размещаться в Интернете в открытом доступе
  - А резюме, как правило, являются личными и не размещаются на публичных сайтах и страницах
  - Профессиональные социальные сети, такие как LinkedIn, заботятся о правах собственности на свои данные и не предоставляют доступ к ним сторонним игрокам
  - Резюме содержат персональную информацию, что порождает дополнительные задачи в отношении обеспечения конфиденциальности на этапе получения и обработки данных

Темы

1. Цель и контекст
2. Задачи
    1. Анализ стороны спроса и предложения
    2. **Визуализация**
    3. Готовые решения на базе ИИ

Визуализация данных

- Цель:
  - Представить информацию, преобразовав ее в полезную для принятия решений
- Задачи:
  - Удовлетворить потребности нескольких типов заинтересованных лиц, обладающих разными целями и наборами умений
- Подход:
  - Разработать адаптивную систему, состоящую из разных специальных подходов, которая сможет обеспечить:
    - поиск **правильного решения для каждой потребности**
    - **единообразие** данных, даже при использовании разных аналитических платформ
    - предложить **масштабируемое** решение, которое сможет обеспечить как потребности в визуализации данных, так и в углубленном анализе данных

Что мы имеем

2 таблицы

Документ FT

1 строка для каждого элемента:

- Основной идентификационный номер → Ключ
- Онлайн-объявление
- Источник
- Местоположение

Почему? Потому что для каждой онлайн-вакансии
мы можем обнаружить указание нескольких местоположений,

например «Разработчик ПО в Лондоне / Ливерпуле»

Анализ проф. умений FT

1 строка для каждого элемента:

- Основной идентификационный номер → Ключ
- Онлайн-объявление
- Источник
- Местоположение
- Профессиональное умение

Почему? Потому что для каждой онлайн-вакансии мы, очевидно, можем выявить
множество профессиональных умений,
например «Разработчик ПО в Лондоне / Ливерпуле,
способный ориентироваться на клиентов, владеющий английским языком и обладающий стрессоустойчивостью».

Что нам потребуется

Руководитель проекта

Ключевые пользователи

Эксперты отрасли

Конечные пользователи

Граждане

Организации

Ответственные за принятие решений

Аналитики

Что нам потребуется

Граждане

Где востребован мой профессиональный профиль?

Какова тенденция спроса на профессии в сфере информационных технологий?

Организации

Ответственные за принятие решений

Соответствуют ли курсы моего университета реальным потребностям рынка?

Как в моей стране развивается зеленая экономика?

Аналитики

Что нам потребуется

Низкий уровень навыков в предметной области
Низкий уровень аналитических навыков
Малая глубина информации
Высокий уровень стандартизации сведений

Граждане

Высокий уровень навыков в предметной области
Низкий уровень аналитических навыков
Средняя глубина информации
Высокий уровень стандартизации сведений

Организации

Высокий уровень навыков в предметной области
Средний уровень аналитических навыков
Средняя глубина информации
Средний уровень стандартизации сведений

Ответственные за принятие решений

Высокий уровень навыков в предметной области
Высокий уровень аналитических навыков
Большая глубина информации
Низкий уровень стандартизации сведений

Аналитики

Что нам нужно обеспечить

Граждане

Что-то **легко** доступное, с **небольшой** глубиной информации, которую **легко понять**

Ключевые показатели результативности (KPIs)

Что-то **легко** доступное, со **средней** глубиной информации, которую **легко понять**

Сторителлинг

Организации

Ответственные за принятие решений

Что-то, что сможет дать **больше информации** людям, которые знают, **как ее читать и интерпретировать**

Информационные панели

Что-то, что сможет дать людям, обладающим **высоким уровнем** знаний в **предметной области**, **техническими** и **аналитическими** навыками, **полный доступ** к данным.

Лаборатория данных

Аналитики

Визуализация данных. Просто график?

- Цель:
  - Предоставить нужной заинтересованной стороне нужный инструмент
- Задачи:
  - Найти способ удовлетворить разные потребности с учетом такого значимого объема информации
- Подход:
  - Определить несколько подходов к визуализации и анализу данных:
    - **Инфографика**: привлекательная, статическая и понятная для широкого круга пользователей
    - **Общедоступные порталы для граждан**: просто, быстро и высокоинформативно
    - **Информационная панель**: повышенная информативность, доступна в Интернете, для лиц, отвечающих за принятие решений
    - **Лаборатория для самостоятельного анализа**: доступ к данным, самый высокий уровень информативности, требует особого набора предметных, технических и аналитических навыков

Правильное решение для каждой потребности

- Потребность:
  - обобщенные показатели
  - не требуется знаний предметной области
  - без риска неверной интерпретации
- Подход:
  - Ключевые показатели результативности (KPIs)
  - Метод блога: мало цифр, подробная интерпретация

Граждане

Смотрите в действии

Правильное решение для каждой потребности

- **Потребность:**
  - показатели средней детализации
  - требуются знания предметной области
  - без риска неверной интерпретации
- **Подход:**
  - Сторителлинг
  - *Информационная панель с пошаговыми инструкциями*: стандартная аналитическая модель, которая знакомит пользователя с разными сведениями

Организации

Смотрите в действии

Правильное решение для каждой потребности

- Потребность:
  - подробные показатели
  - требуются знания предметной области и аналитические навыки
  - без риска неверной интерпретации
- Подход:
  - Информационная панель
  - Свободный анализ в заранее определенной среде

Ответственные за принятие решений

Смотрите в действии

Правильное решение для каждой потребности

- Потребность:
  - без показателей
  - доступ к необработанным и очищенным данным
  - высокий риск неверной интерпретации
  - требуется высокий уровень знаний:
    - предметной области
    - модели данных
    - методологии
    - технологий
- Подход:
  - Лаборатория данных
  - Место, где можно запрашивать данные, применяя аналитические решения по своему выбору

Аналитики

Смотрите в действии

Темы

1. Цель и контекст
2. Задачи
    1. Анализ стороны спроса и предложения
    2. Визуализация
    3. **Готовые решения на базе ИИ**

Почему?

- Мы увидели большое количество компонентов ИИ, которые могут помочь нам анализировать данные:
  - определение языка
  - спам-фильтры для онлайн-объявлений о работе
  - фильтр дедупликации
  - классификаторы профессий по ESCO
  - классификатор на основе пользовательских таксономий
  - …

Есть ли у нас возможность использовать их снова?

Повторное использование ИИ

Конечно. Повторное использование модели в ИИ — это общепринятая и передовая практика:

- снижается **стоимость** разработки
- **повышается** эффективность
- централизованное **обслуживание**
- **единообразие** классификации в разных проектах

Повторное использование ИИ присутствует даже в представленном примере анализа спроса и предложения: мы повторно использовали и адаптировали большую часть конвейера классификации, чтобы обеспечить согласованность данных из онлайн-объявлений и резюме

Платформы для совместной работы

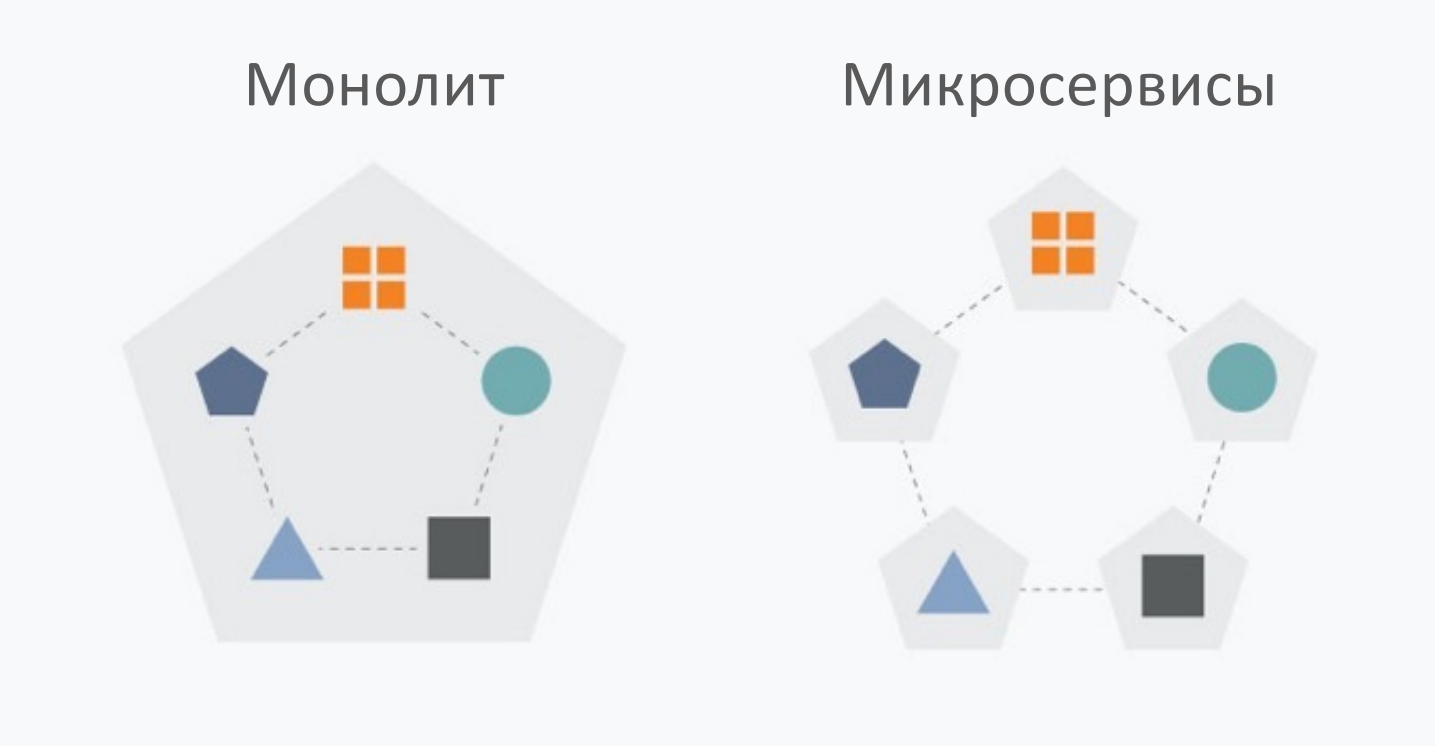
На рынке представлено несколько образцов платформ для совместной работы и обмена данными:

- Databricks
- AWS SageMaker
- Anaconda
- H2O

Большинство из них предназначено для обмена информацией и совместной работы в рамках единого конвейера анализа данных, включая модель машинного обучения

Микросервисы

Современные технологии позволяют повторно использовать компоненты программного обеспечения как отдельные кирпичики, из которых можно составить более сложную систему.



Микросервисы

Эта архитектура может обеспечить эффективное повторное использование компонентов, которые будут взаимодействовать между собой, обмениваясь информацией.

Нет необходимости знать «что лежит в основе сервиса» (например, классификация ESCO). Нужно всего лишь знать, «как с ним взаимодействовать» (как сделать правильный запрос и как обработать ответ).

Этот похоже на одного из наших интеллектуальных помощников (Alexa, Siri,…)

Задачи

То, что мы только что рассмотрели, дает возможность существенно повысить возможности масштабирования и повторного использования системы.

НО

Нам нужно помнить о том, что мы делаем, что мы хотим получить и как была построена используемая нами система

Риски готовых решений ИИ

Использование готовой модели ИИ — это отличная идея: нам не нужно заниматься ее проектированием, разработкой и обслуживанием. Мы можем просто *пользоваться* ею.

Но если нам не известно, как эта система была спроектирована, как ее тренировали, строили и как она обслуживается, мы можем получить неожиданные результаты.

Модель ИИ (как и любая другая аналитическая система) — это не оракул.

Передовые практики работы с готовыми системами ИИ

В рамках сотрудничества с Евростатом мы работаем над созданием Центра веб-аналитики (Web Intelligence Hub), который должен служить репозитарием компонентов ИИ с тем, чтобы позволить и максимизировать их повторное использование и обмен знаниями.

Чрезвычайно важно обеспечить, чтобы поведение каждого компонента, а также лежащей в его основе методологии, было должным образом задокументировано. Это позволит избежать неожиданных результатов и неверной интерпретации полученных сведений.

Давайте поразмышляем о дедупликации размещаемых объявлений и отраслевой классификации

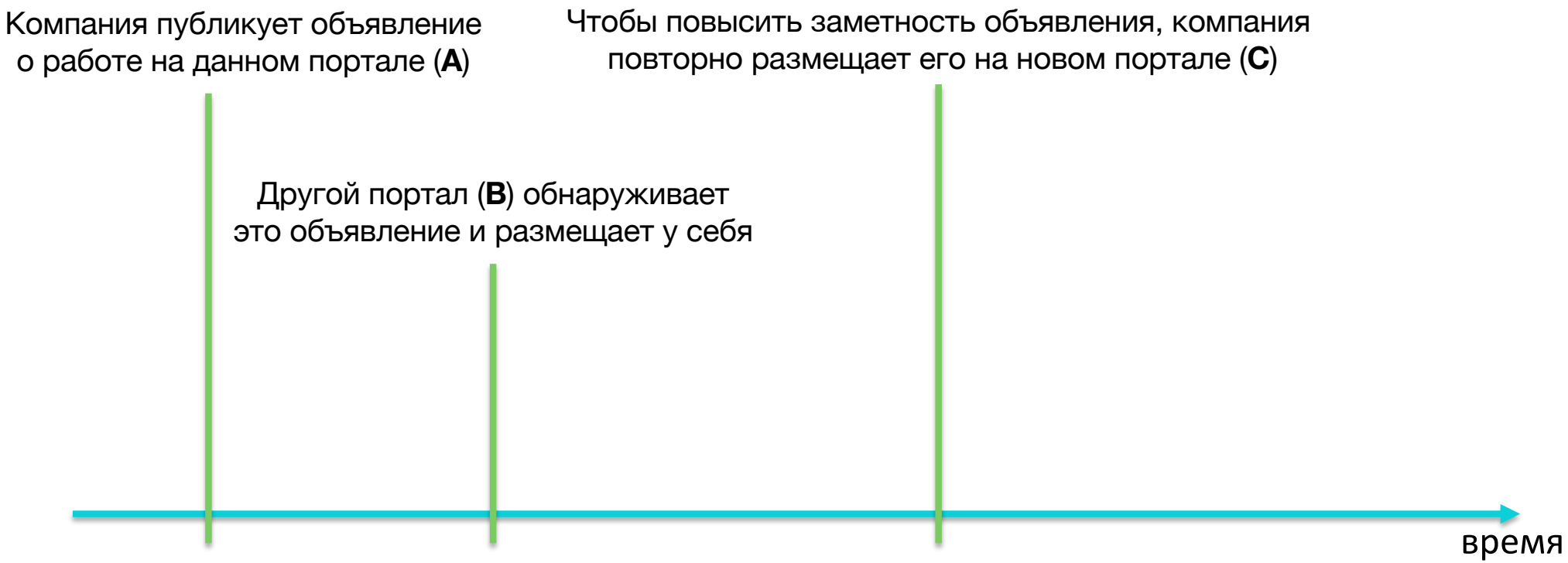
Проблема дедупликации публикаций

Я знаю, что на данном портале размещено 100 публикаций. Почему я нахожу только 10? Система делает что-то неправильно?

Это типичный вопрос, который возникает при анализе данных онлайн-объявлений о работе.

Это не система выдает неверный результат, а просто нам нужно знать, **что мы анализируем**

Проблема дедупликации публикаций



Проблема дедупликации публикаций

Вакансия	Наименование	Описание	Отрасль	Дата публикации	Портал
0001	Менеджер по продажам	Наша компания…	Производство	01.01.2021	A
0002	Менеджер по продажам	Наша компания…	Производство	15.01.2021	B
0003	Менеджер по продажам	Нашей компании…	Производство	01.02.2021	C

Высокий уровень сходства

Связанный временной период

Три онлайн-объявления определяются как дубликаты.
Первое из них (портал A) выбирается как дедуплицированное объявление.

Проблема отраслевой классификации

У нас есть официальная статистика, согласно которой на сектор образования приходится 5 % рынка. Почему я нахожу в системе только 0,2? Она делает что-то неправильно?

Это еще один типичный вопрос, который возникает при анализе данных онлайн-объявлений о работе.

Опять-таки, нам нужно глубоко понимать, **что мы анализируем**

Проблема отраслевой классификации

Сектор здравоохранения большинства стран нанимает работников через публичные тендерные процедуры.

В таком случае этот сектор не отражается в онлайн-объявлениях о работе.

Это проблема не классификации, но *искаженного отбора*: если мы будем принимать это искажение во внимание, мы будем способны лучше понять полученные результаты.

Передовые практики

Мы не защищаем систему или полученные результаты: большие данные, как мы увидели, могут привнести существенные трудности, с которыми придется справляться.

С другой стороны, если мы способны справиться с существующими рисками, эти данные дадут возможность осуществлять анализ новыми и прорывными методами.

Не верьте этой системе, но понимайте ее и верьте в нее.

Спасибо большое

Алессандро Ваккарино