Big Data for Labour Market Information

Сессия 3

Обнаружение знаний в базах данных (ОЗБД) для АРТ

Алессандро Ваккарино – Фабио Меркорио

Большие данные в контексте информации рынка труда – акцент на данных онлайн-вакансий. Учебный семинар

Милан, 21–22 ноября 2019 года

ETF Working together Learning for life

European Commission

EaP Eastern Partnership

Темы

1. Этапы ОЗБД для прорыва в АРТ
2. Как можно использовать обработанную ИРТ? [примеры]
3. Проблемы, ограничения и вызовы

Знаменуют ли большие данные прорыв в сфере рынка труда?

Аналитике рынка труда
могут содействовать три главных
источника данных:

(1) статистические источники
(2) административные источники
(3) интернет-источники (большие
данные для АРТ)

К чему идет рынок труда?

ВЫЗОВЫ РТ

1. Эволюция профессиональных умений
2. Новые нарождающиеся профессии
3. Автоматизация работы и замена человека
4. Мобильность

Потребности РТ

1. Обновление информации (почти в реальном времени)
2. Решения на основе данных (пусть говорят данные)
3. Возможность предугадывать тенденции

Знания приобретают критическое значение для разных игроков и разработчиков политики рынка труда, давая возможность понимать его динамику и тенденции

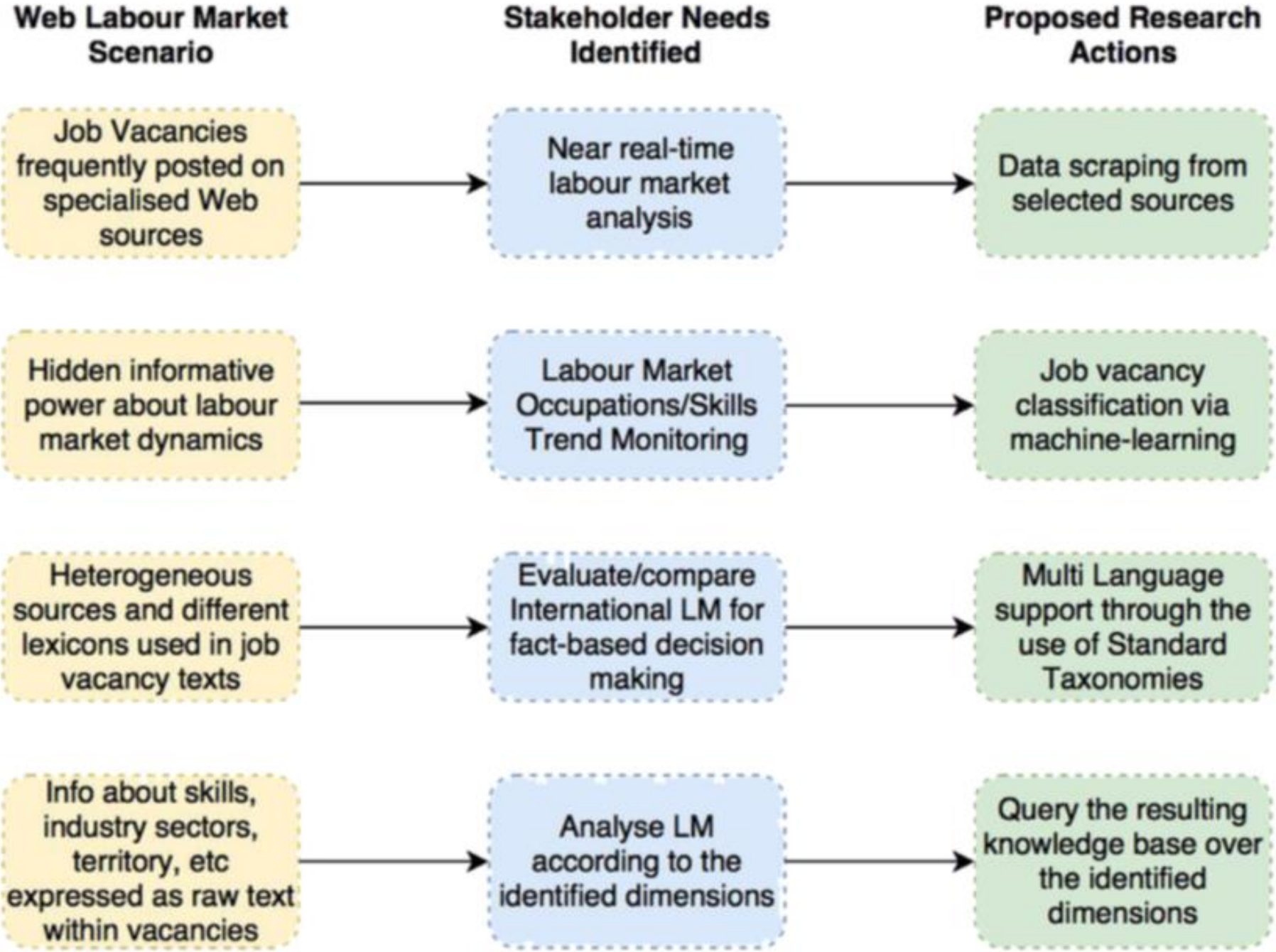

Web Labour Market Scenario
Stakeholder Needs Identified
Proposed Research Actions
Job Vacancies frequently posted on specialised Web sources
Near real-time labour market analysis
Data scraping from selected sources
Hidden informative power about labour market dynamics
Labour Market Occupations/Skills Trend Monitoring
Job vacancy classification via machine-learning
Heterogeneous sources and different lexicons used in job vacancy texts
Evaluate/compare International LM for fact-based decision making
Multi Language support through the use of Standard Taxonomies
Info about skills, industry sectors, territory, etc expressed as raw text within vacancies
Analyse LM according to the identified dimensions
Query the resulting knowledge base over the identified dimensions

Table 1 Main characteristics for LM Data Sources

LM Source Type	*Data Type*[2]	*Generation Rate*	*Data Model Paradigm*	*Quality*	*Coverage*	*Analysis Paradigm*	*Believability*	*Value*
Statistical	Structured	Periodically	Relational	Owner's responsibility	Owner's responsibility	Top Down & Model Based	Owner's responsibility	intrinsic
Administrative	Structured or Semi-structured	Periodically	Relational	Owner's responsibility	Owner's responsibility & User's responsibility	Top Down & Model Based	Owner's responsibility & User's responsibility	intrinsic
Web	Structured, Semi-structured or Unstructured	Near-real-time or real-time	Relational and Non Relational (NoSQL)	User's responsibility	User's responsibility	Bottom up & Data Driven	User's responsibility	extrinsic

Table 2 Most significant limitations of Big Data architectures

Issue (most significant)	Caused by	Conceptual Blocks of Big Data Architectures
Schema-free data are out: only structured data sources can be manipulated. Roughly, this means that only data that obey a rigid, well-defined data model can be handled, to the exclusion of all "unstructured" data, such as free text, comments and Web content in general.	Variety	Data ingestion; NoSQL models;
No adaptability to change: the addition of a new source requires the whole process to change, and this makes it difficult to scale the architecture over multiple (albeit structured) sources.	Variety, Velocity	Data lake
Rigid ETL: the procedures that transform content from source formats to target formats have to be precisely written to fit the desired data structure (e.g., data warehouse).	Variety	Schema free; data-driven approach (bottom-up rather than top-down)
Time consuming: the larger the volume of data to be processed, the greater the time needed to complete the process. ETL procedures are usually high time and memory consumers, as they need to "scan" all the data sources at any time to transform source data.	Volume, Variety, Velocity	Scale-out rather than scale-up

Как работать с онлайн-вакансиями (ОВ) в широких масштабах?

Пример размещённой в Интернете вакансии

Должность: специалист по обработке и анализу данных

Описание: Растущая компания-разработчик ищет талантливого специалиста в области вычислительных технологий. Ваши знания в области работы с данными помогут нам выйти на новый уровень. В ваши обязанности будет входить выявление возможностей для дальнейшего улучшения наших механизмов поиска вакансий и соискателей, а также разработка и внедрение решений. […] Требуемые умения и знания:

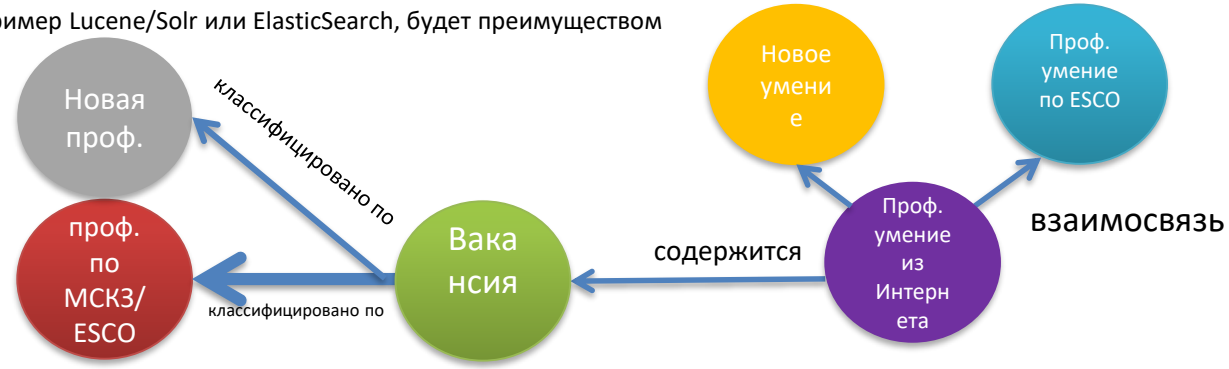
- SQL и реляционные базы данных;
- Анализ данных в R (или Matlab);
- Обработка крупных массивов данных в MapReduce и Hadoop;
- Аналитика в реальном времени при помощи Spark, Storm или похожих решений;
- Машинное обучение;
- Обработка естественного языка (NLP) и интеллектуальный анализ текста;
- Разработка на языках C++, Python, Perl;
- Опыт работы с поисковыми машинами, например Lucene/Solr или ElasticSearch, будет преимуществом

Пример размещённой в Интернете вакансии

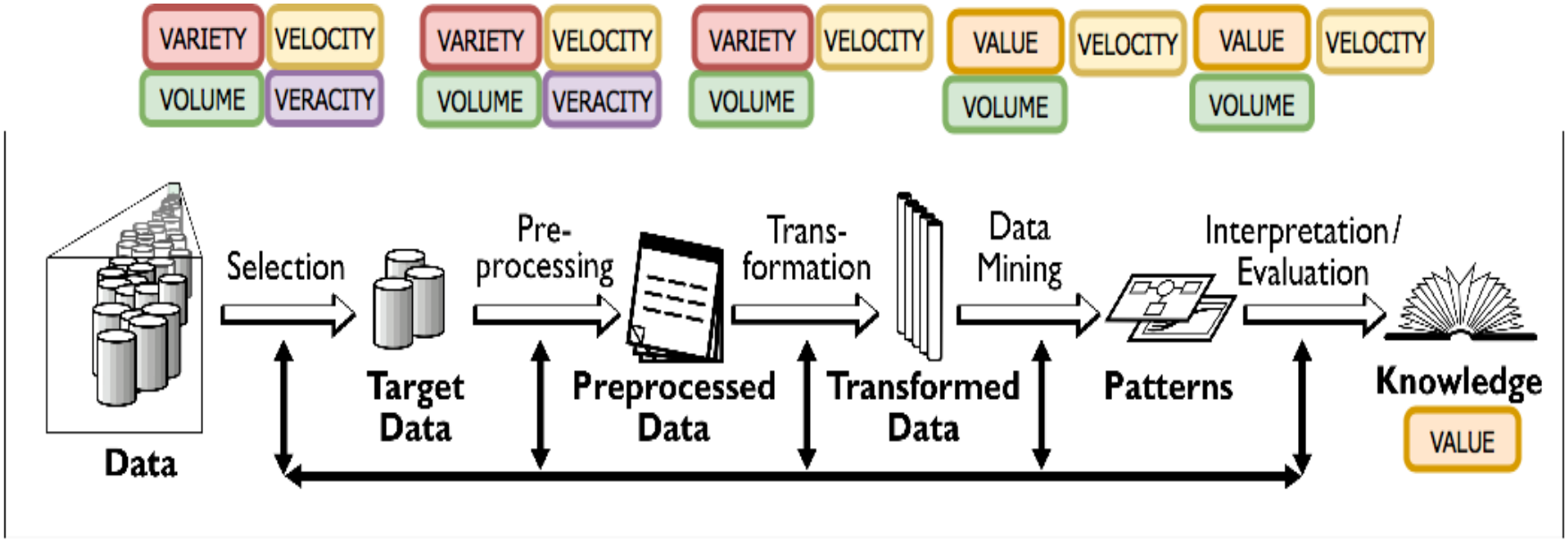
Должность: специалист по обработке и анализу данных

Описание: Растущая компания-разработчик ищет талантливого специалиста в области вычислительных технологий. Ваши знания в области работы с данными помогут нам выйти на новый уровень. В ваши обязанности будет входить выявление возможностей для дальнейшего улучшения наших механизмов поиска вакансий и соискателей, а также разработка и внедрение решений. […] Требуемые умения и знания:

- SQL и реляционные базы данных;
- Анализ данных в R (или Matlab);
- Обработка крупных массивов данных в MapReduce и Hadoop;
- Аналитика в реальном времени при помощи Spark, Storm или похожих решений;
- Машинное обучение;
- Обработка естественного языка (NLP) и интеллектуальный анализ текста;
- Разработка на языках C++, Python, Perl;
- Опыт работы с поисковыми машинами, например Lucene/Solr или ElasticSearch, будет преимуществом



Процесс (ОЗБД, Файяд 1997)



Процесс (ОЗБД для АРТ)

Какие нужны знания и умения, чтобы построить систему АРТ?

Пять шагов для превращения больших данных в АРТ

Профессионалы, вовлекаемые на разных этапах проекта АРТ:

- Статистики
- Инженеры и специалисты по обработке и анализу данных
- Эксперты в области РТ – экономисты и социологи

Такой набор знаний и умений этих специалистов говорит о необходимости междисциплинарного подхода к разработке проекта АРТ

Пять шагов для превращения больших данных в АРТ

Эти шаги включают в себя обработку текста, NLP, очистку данных, устранение шума и дедупликацию данных

Статистики: определяют меры по обеспечению качества данных, их распределению и значимости

Специалист в области вычислительных технологий: гарантирует предварительную обработку данных в масштабе свыше миллиона объектов

Эксперты в области РТ: как определить синонимы сферы РТ, которые помогут повысить точность данных? Как определить критерии, характеризующие недостающие значения и дубликаты?

Пять шагов для превращения больших данных в АРТ

Включает сокращение и проекцию данных с целью определения единой модели

Статистики: измеряют полноту единой модели данных

Специалист в области вычислительных технологий: гарантирует масштабное преобразование сырых данных в целевые данные

Эксперты в области РТ: как нам определить формат и таксономию целевых данных?

Пять шагов для превращения больших данных в АРТ

Определение подходящих алгоритмов ИИ (например, классификация, прогнозирование, регрессия, кластеризация, фильтрация информации) путём поиска нужных моделей в определённой репрезентативной форме, исходя из цели анализа

Статистики и специалисты в области вычислительных технологий: определение алгоритма, корректировка параметров, внедрение

Эксперты в области РТ: Какие умения отбирать, а какие следует отбросить? Оценка результата

Пять шагов для превращения больших данных в АРТ

Внедрение визуальной парадигмы для визуального представления полученных знаний в зависимости от целей пользователя. В контексте АРТ это означает учёт способности пользователя понять данные и его главной цели

Статистики и специалисты в области вычислительных технологий: определяют подход к визуализации и изложению и реализуют его

Эксперты РТ: Как передать нужные знания в соответствии с потребностями заинтересованного лица?

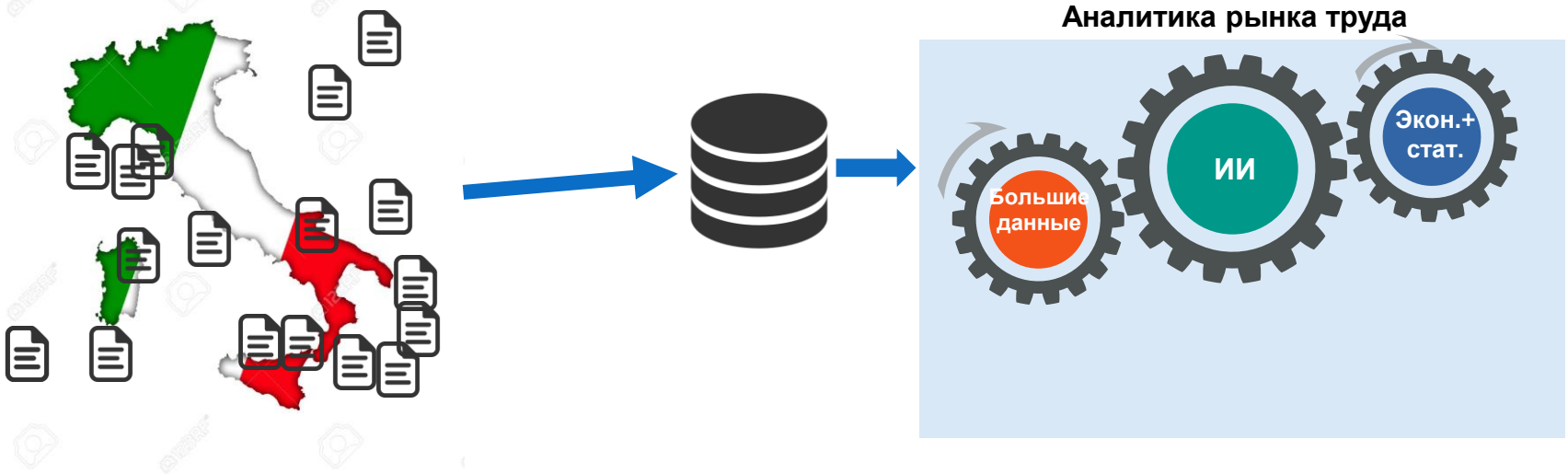
Как определить траектории визуальной навигации для каждого заинтересованного лица?

Как получить обратную связь (если она есть) он пользователей РТ?

Как использовать знания о РТ в бизнес-целях?

ИТАЛЬЯНСКИЙ монитор рынка труда в реальном времени

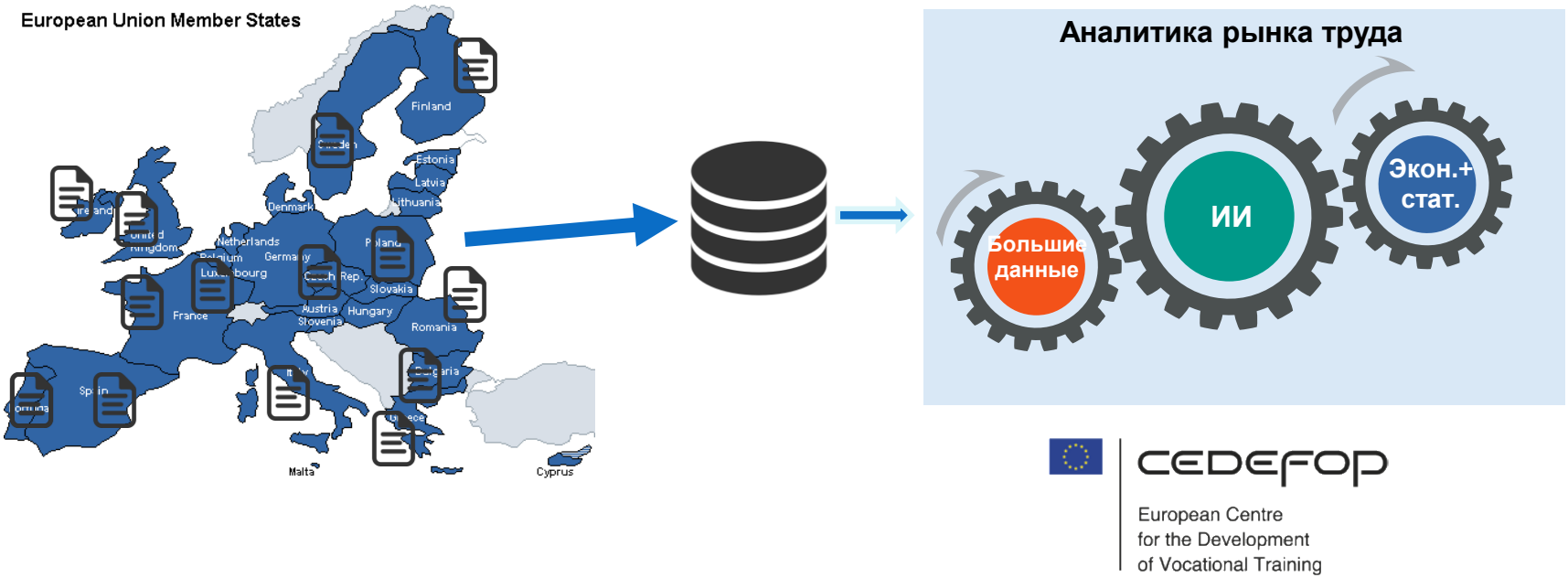
WOLLYBI
powered by Tabulaex



CRISP interuniversity research centre for public services

ОВ с 2013 года – свыше 4 млн. уникальных вакансий

ЕВРОПЕЙСКИЙ монитор рынка труда в реальном времени



28 стран ЕС – 32 языка – более 6 млн уникальных вакансий в месяц

Как можно использовать обработанную ИРТ?

1. Обнаружение профессий и проф. умений
2. Уровни межличностных, цифровых и специальных умений
3. Новые нарождающиеся профессии
4. Расширение таксономии
5. Объяснить людям механизм работы машинного обучения

Обнаружение профессий и проф. умений

Акцент на профессиях и умениях, на которые есть спрос на онлайновом РТ

Наглядная демонстрация: территориальное измерение

ВИДЕО

Уровни межличностных, цифровых и специальных умений

Как оценить влияние цифровизации на профессии?

Вычислить уровни умений

Цель: Оценить распространённость ИКТ в профессиях, связанных и не связанных с ИКТ

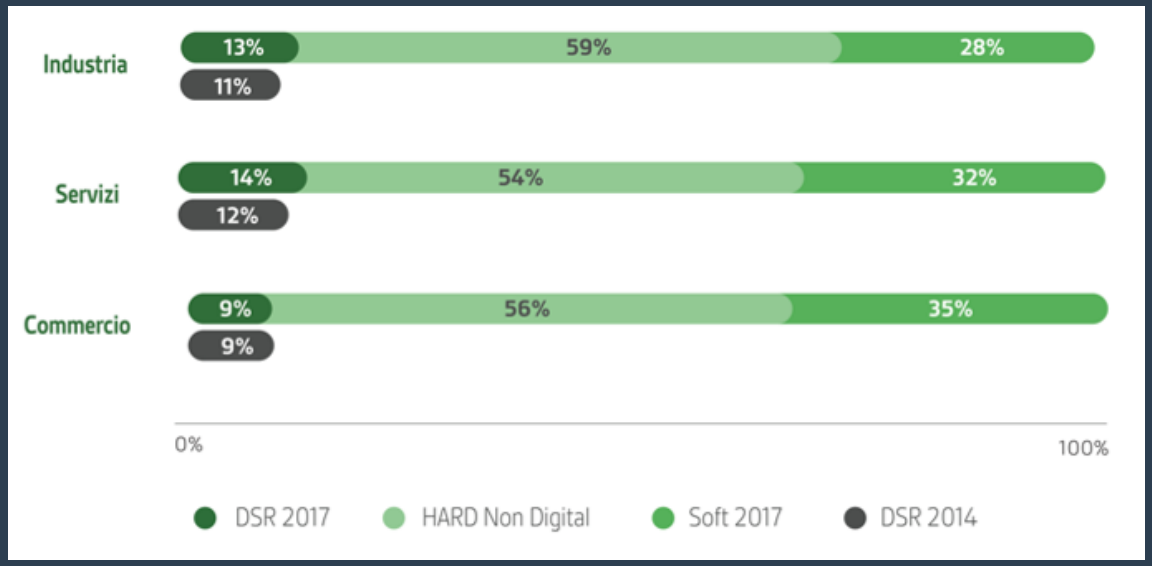
Идея: Использовать информативность систематизированных ОВ для вычисления уровня цифровых навыков (УЦН), межличностных и узкоспециальных нецифровых умений

УЦН измеряет уровень проявления цифровых навыков в отдельно взятой профессии, исходит из наблюдений за распространённостью цифровых навыков во всех профессиях, независимо от того, связаны ли они с миром ИКТ.

Вычисление уровней умений и навыков

Источник: WollyBI

Спрос на цифровые, узкоспециальные и межличностные умения по секторам



Специальный анализ на разных уровнях детализации. В данном случае – с акцентом на «сектор»...

Источник – WollyBI, торговая марка TabulaeX

Вычисление уровней умений и навыков

Источник: WollyBI

- **Прикладные и управленческие навыки** = умение использовать инструменты и программное обеспечение для управления как операционными процессами, так и принятием решений
- **Специальные умения ИКТ** = для работы с узкоспециализированными решениями, платформами и программными языками
- **Базовые умения** = для повседневного использования основных цифровых инструментов
- **Навык информационного посредничества** = для использования цифровых инструментов, предназначенных для корпоративной коммуникации

... и другие, в зависимости от особенностей каждой профессии...

Источник – WollyBI, торговая марка TabulaeX

Источник: WollyBI

Вычисление уровней умений и навыков [умения ESCO + новые]

	Прикладные и управленческие навыки			Специальные ИКТ		Информационное посредничество	
Профессия	Использование баз данных	Программы для проектировщиков	3D-моделирование	Создание интерфейсных веб-сайтов	Веб-программирование	Использование графических программ	Использование программной разметки
Граф. и мультимедийные дизайнеры	2.5	3.5	3	4.5	2	5	2

	Прикладные и управленческие навыки		Информационное посредничество		
Профессия	Использование баз данных	ERP	Управление цифровыми данными	Оптимизация поисковых систем (SEO)	Использование социальных сетей
Специалист по обучению сотрудников	4.5	4	4.5	4	2.5

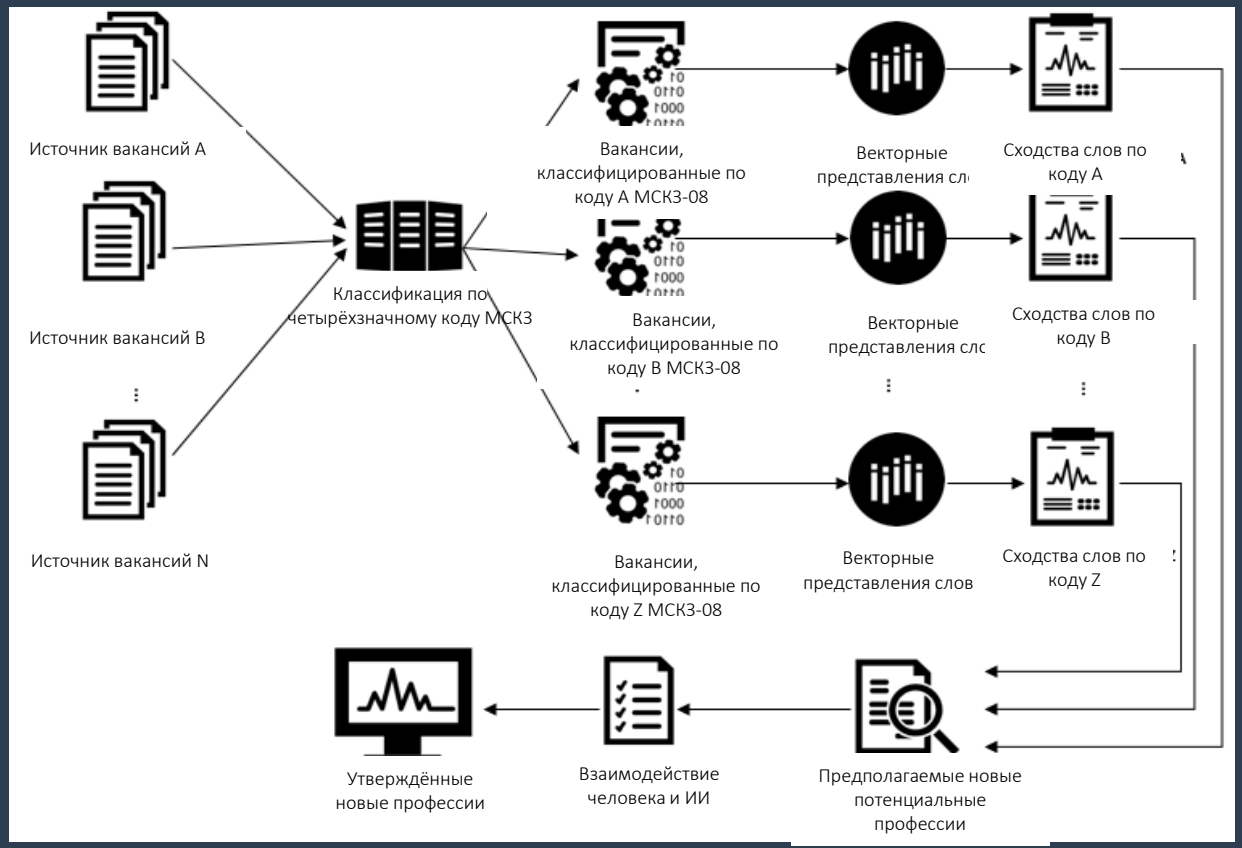
... и другие, в зависимости от элементарных умений

Источник – WollyBI, торговая марка TabulaeX

Новые нарождающиеся профессии на основе схожести и различий умений

Выявление новых нарождающихся профессий при помощи ИИ

1. Классификация ОВ по четырёхзначному коду МСКЗ
2. Построение нескольких векторно-пространственных представлений слов (профессии и проф. умения) для выявления сходства лексики в разных ОВ
3. Вычисление сходств между известными (профессии и умения) и новыми терминами
4. Предложение новых потенциальных профессий



(Некоторые) новые нарождающиеся профессии

Специалист по обработке и анализу данных

Облачные вычисления

Эксперт по кибербезопасности

Бизнес-аналитик

Аналитик больших данных

Маркетинг в социальных сетях

В период с января 2014 по август 2019 гг. собрано 9000 онлайн-вакансий, связанных с (некоторыми) новыми нарождающимися профессиями (см. выше)

СПЕЦИАЛИСТ ПО ОБРАБОТКЕ И АНАЛИЗУ ДАННЫХ – 1,7 тыс. вакансий (2014–2019)

ПРОФ. УМЕНИЯ

48 %	39 %	13 %
МАТЕМАТИКА И СТАТИСТИКА	ВЫЧИСЛЕНИЯ	ДЕЛ. АДМ.
• Анализ данных, статистическое обучение • SAS, R • SAP & SPSS*	• SQL, Python, Hadoop • Бизнес-аналитика, машинное обучение • Интеграция данных	• Связи с общественность ю • Управление • Управление взаимоотношени ями с клиентами

УМСТВЕННЫЕ И МЕЖЛИЧНОСТ НЫЕ УМЕНИЯ

38 %	29%	13 %	7 %	5 %
ПОВЕДЕНИЕ	ИНОСТРАННЫЕ ЯЗЫКИ	РЕШЕНИЕ ПРОБЛЕМ	СОТРУДНИЧЕСТВО	ЛИДЕРСТВО

Перемена между 2019 и 2018 гг.: +31 %
Перемена между 2019 и 2017 гг.: +149 %

СПЕЦИАЛИСТ ПО СОЦИАЛЬНЫМ СЕТЯМ – 0,4 тыс. вакансий (2014–2019)

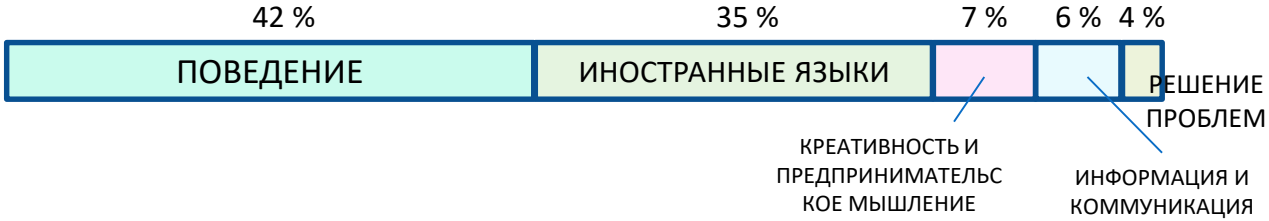
ПРОФ. УМЕНИЯ

- Управление
- Связи с общественностью
- Маркетинг

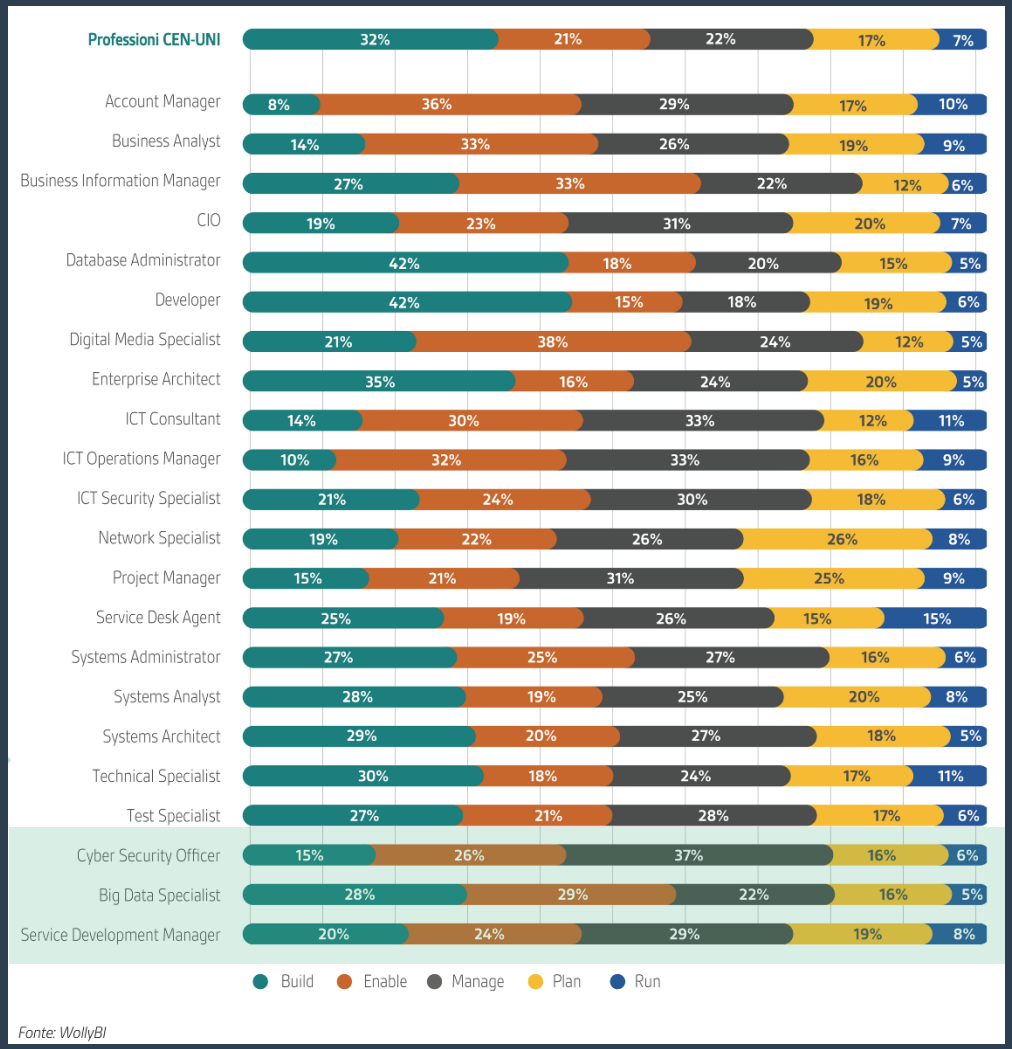
- Adobe Photoshop и HTML5
- Google Analytics и AdWords*
- CMS (система управления контентом)*

Анализ данных

УМСТВЕННЫЕ И МЕЖЛИЧНОСТНЫЕ УМЕНИЯ



Перемена между 2019 и 2018 гг.: +105 %
Перемена между 2019 и 2017 гг.: +123 %

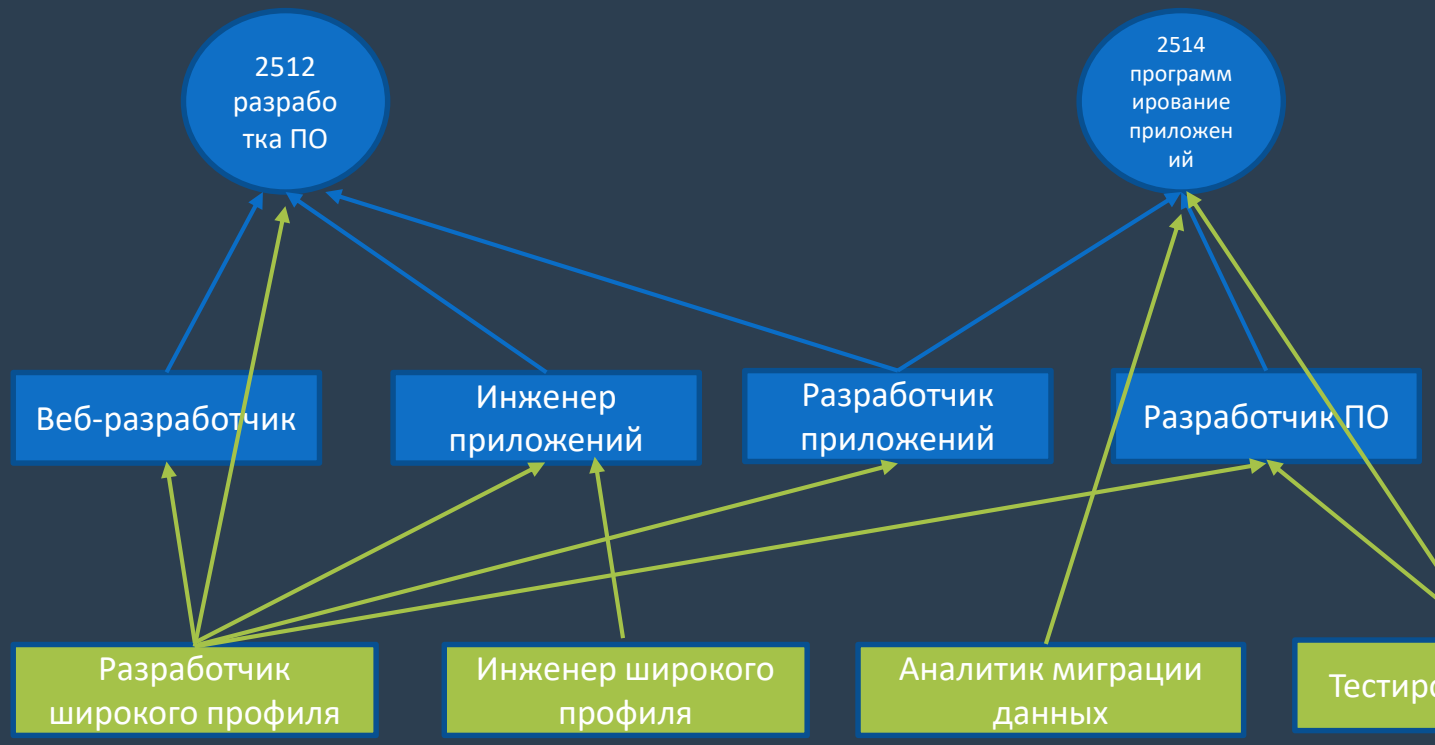


Fonte: WollyBI

Новые профессии можно сравнивать с традиционными при помощи разных таксономий (в этом случае – рамки компетентности eCF)

Расширение таксономии: как улучшить таксономии проф. умений и профессий при помощи семантических сходств ОВ?

Основная идея



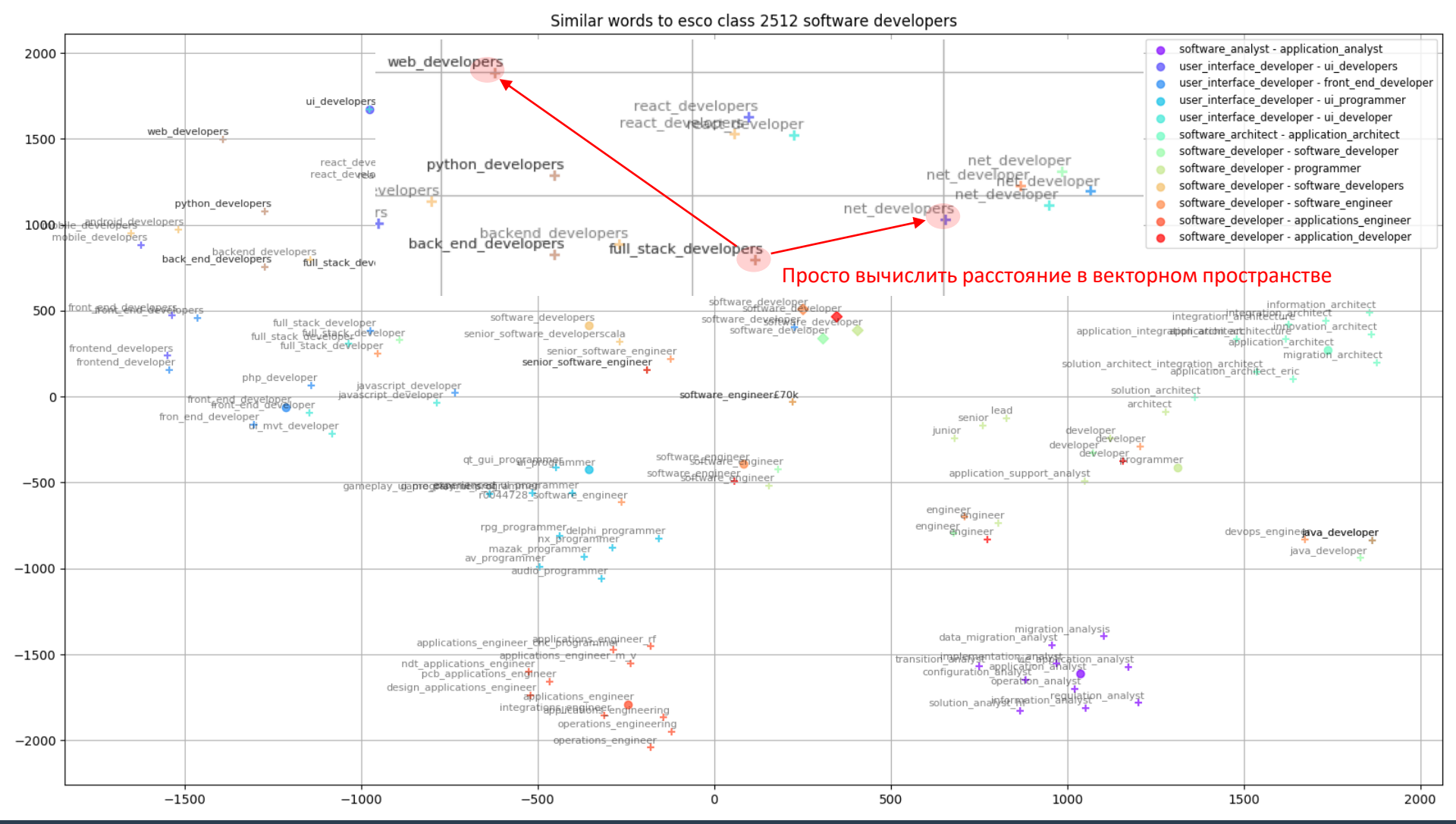
Четырёхзначный код МСКЗ

Пятизначный код ESCO и альтернативные метки

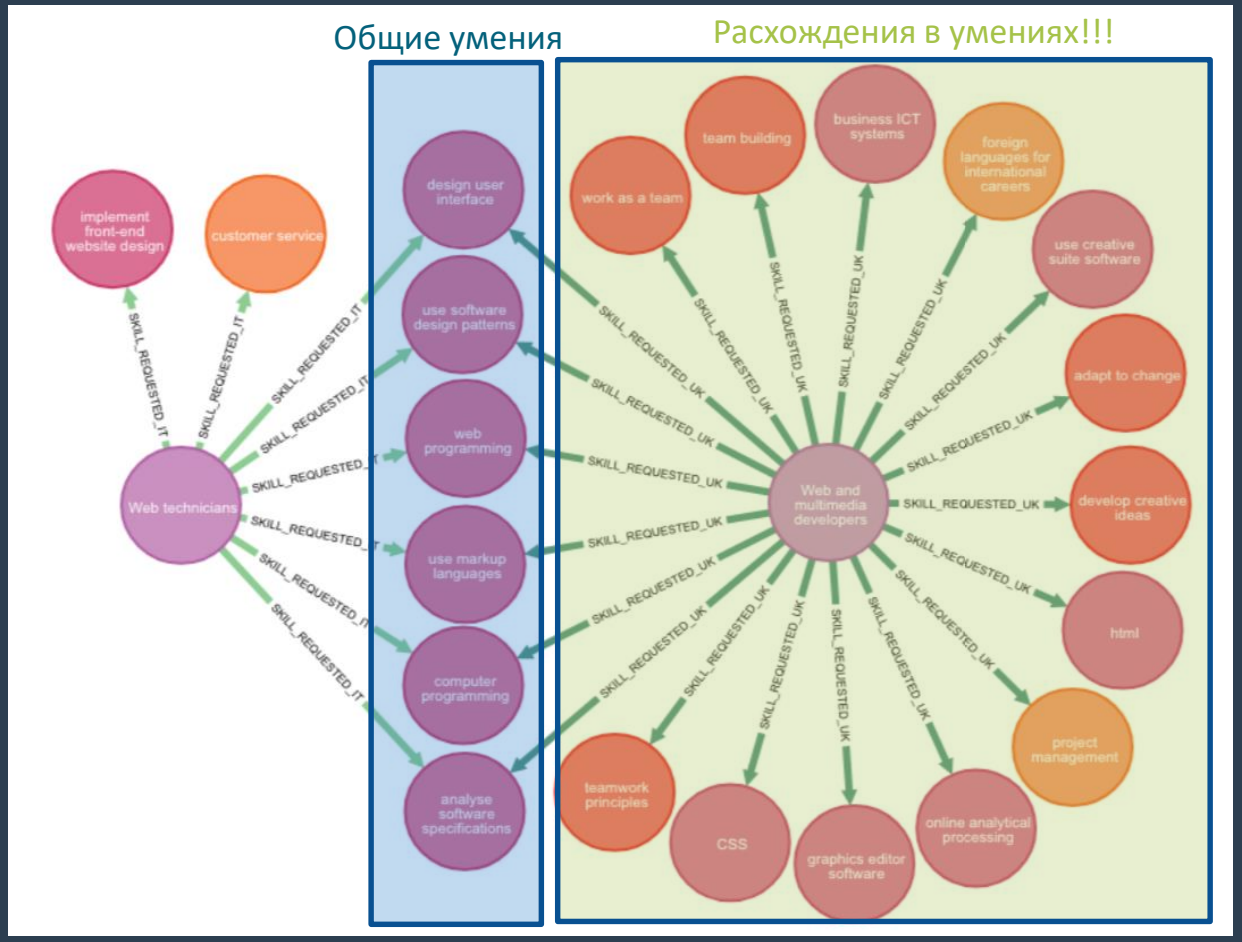
Упоминания в ОВ

Основные вопросы: Как...

1. *поддерживать актуальность таксономии в соответствии с ожиданиями и лексикой рынка труда?*
2. *вносить эти упоминания в соответствующие записи таксономии?*
3. *оценить схожесть между всеми записями таксономии?*
4. *оценить релевантность записей таксономии?*


Similar words to esco class 2512 software developers
software_analyst - application_analyst
user_interface_developer - ui_developers
user_interface_developer - front_end_developer
user_interface_developer - ui_programmer
user_interface_developer - ui_developer
software_architect - application_architect
software_developer - software_developer
software_developer - programmer
software_developer - software_developers
software_developer - software_engineer
software_developer - applications_engineer
software_developer - application_developer
web_developers
ui_developers
python_developers
backend_developers
back_end_developers
full_stack_developers
net_developers
net_developer
react_developers
software_developers
senior_software_engineer
software_engineer£70k
java_developer
Просто вычислить расстояние в векторном пространстве

Расширение таксономии профессиональных умений



Сравнение разных рынков труда веб-отрасли [в данном случае – Италия, Великобритания и Германия] для осуществления анализа расхождений в проф. умениях, взяв за основу таксономию.

Вопрос: *«если взять за основу определённую профессию в таксономии МСКЗ для итальянского рынка труда, какая профессия на британском рынке труда наиболее соответствует ей с точки зрения требуемых профессиональных умений?»*

Умения, ассоциируемые с профессией «веб-техник» в Италии, больше всего похожи на профессию «Веб- и мультимедиа разработчик» в Великобритании

Построение онтологии АРТ

ВИДЕО

Сравнение разных стран (веб-техник)

Nazione	Top 10 skill	*tf-idf* IT	UK	DE
IT	web programming	**100.0**	38.0	100.0
IT	analyse software specifications	**95.18**	36.84	21.72
IT	computer programming	**89.17**	48.82	59.35
UK	Marketing	39.69	**100.0**	56.56
UK	online analytical processing	0	**94.39**	0
UK	adapt to change	69.21	**87.21**	79.47
DE	web programming	100.0	38.0	**100.0**
DE	use markup languages	80.47	34.03	**82.97**
DE	adapt to change	69.21	87.21	**79.47**

Совершенствование классификации при помощи визуальных пояснений: объяснить и проверить

Объяснимая аналитика рынка труда (OAPT)

Цель. **Повысить достоверность анализа и результатов, предоставляемых конечным пользователям, объяснив им поведение алгоритмов ИИ, которые используются для получения этих результатов.**

Идея. Алгоритмы машинного обучения действуют, как черный ящик, и угадать причину того или иного решения невозможно. Объяснимый ИИ (ОИИ, запущен управлением перспективных исследований Министерства обороны США в 2016 году) призван создать новое поколение систем машинного обучения, способных объяснять свои решения, как человек.

Преимущества. Улучшение способности понимать и отслеживать процессы классификации (даже в случае сложных последовательностей алгоритмов) и выявлять случаи ошибочной классификации с целью повышения точности и понимания системы.

Зачем нам объяснения?

Нет возможности гарантировать (или хотя бы аргументировать) пользователям, что система выучила *«правильную модель»*. Можно лишь говорить о том, что она выучила *«модель правильно»*.

Возможность объяснить (**в понятном человеку формате**), почему система отнесла тот или иной элемент именно к этой категории, сделала бы систему более надёжной и достойной доверия в глазах пользователя.

Способность понять и выделить факторы, которые заставили систему предугадать определённый результат, можно использовать в процессе проверки результатов с участием человека, что позволит улучшить:

- **прозрачность** для конечных пользователей, которые принимают решения на основе выданных вами результатов;
- **достоверность и полезность** системы;
- **точность** классификации;
- и многое другое (смотрите далее)

RA2: Совершенствование классификации при помощи визуальных пояснений: объяснить и проверить

Предложение. Использовать данные ОВ для текстового выражения алгоритмов ОАРТ

Потенциальные заинтересованные лица. Национальные ассоциации компании сферы информационных технологий

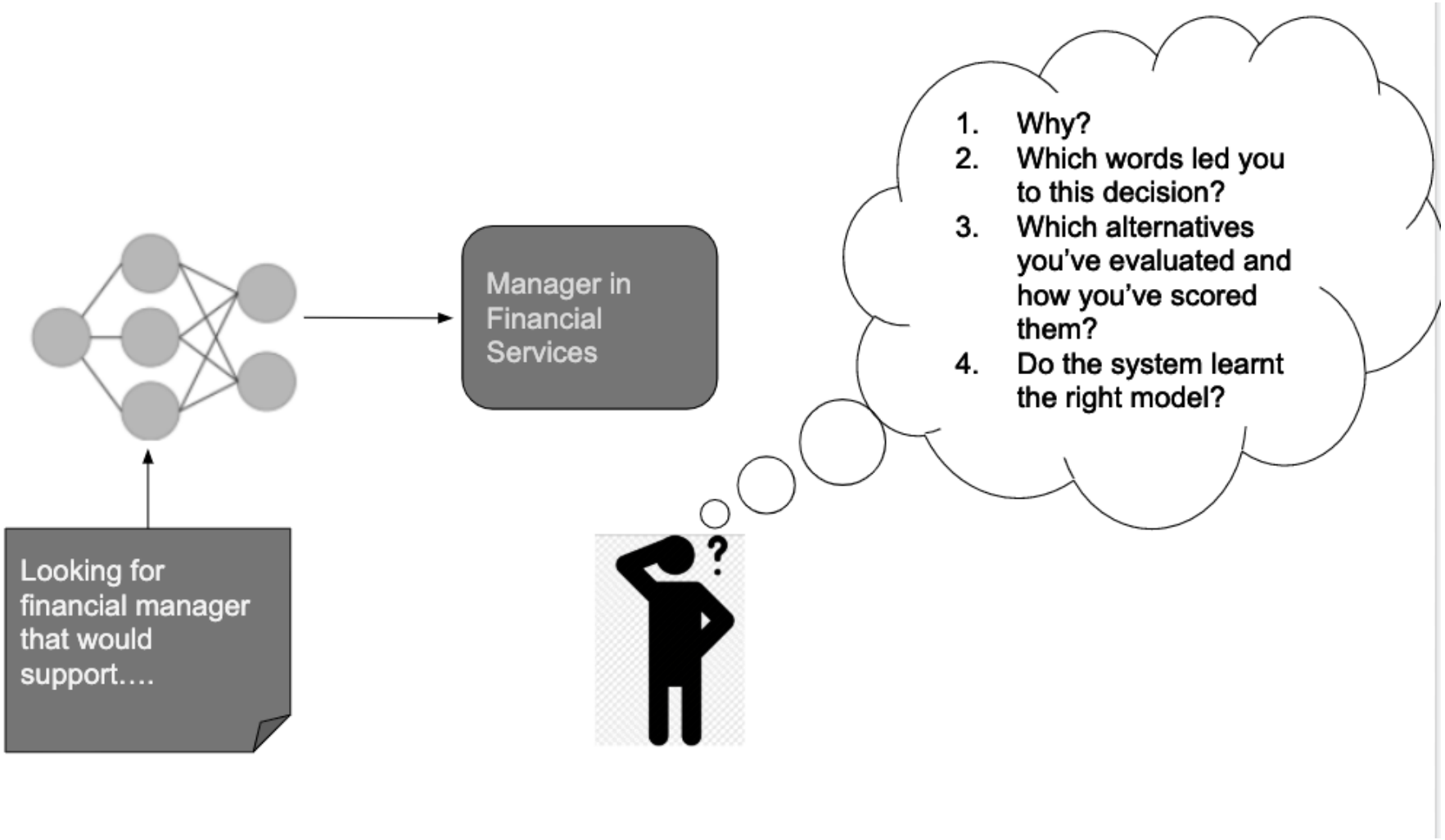
Фазы исследования

- Определение глобальной и местной модели объяснимости
- Оценка алгоритмов ОИИ
- Внедрение алгоритма ОАРТ
- Текстовое выражение определённого алгоритма ОАРТ на основе данных ОВ

Результат исследования.

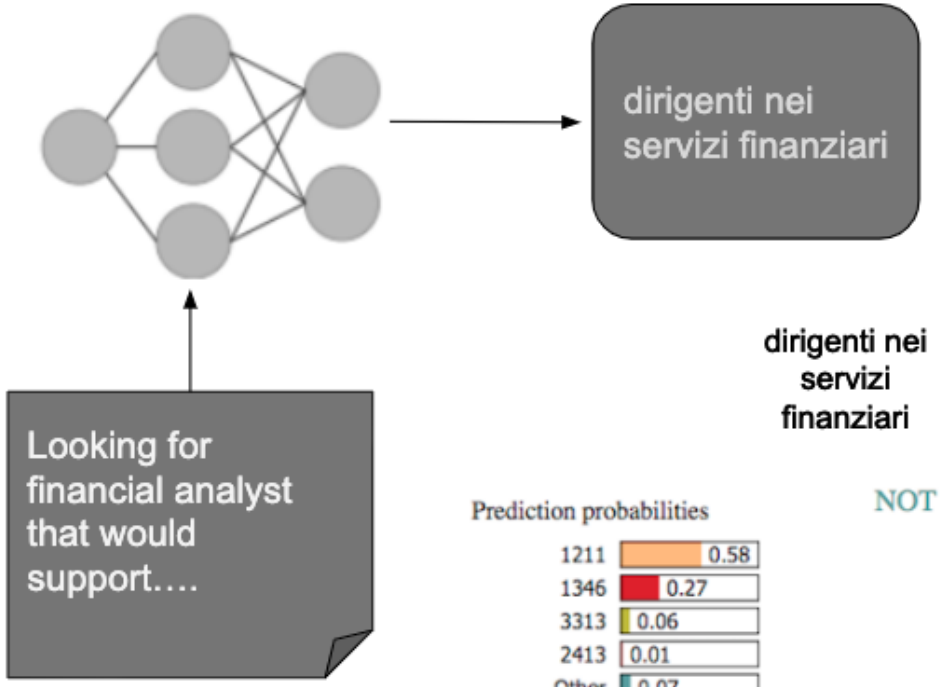
- Рабочие документы;
- Предложение исследования для объяснения характеристик классификации WolliBI [в работе].

Основное внимание: объяснить и проверить

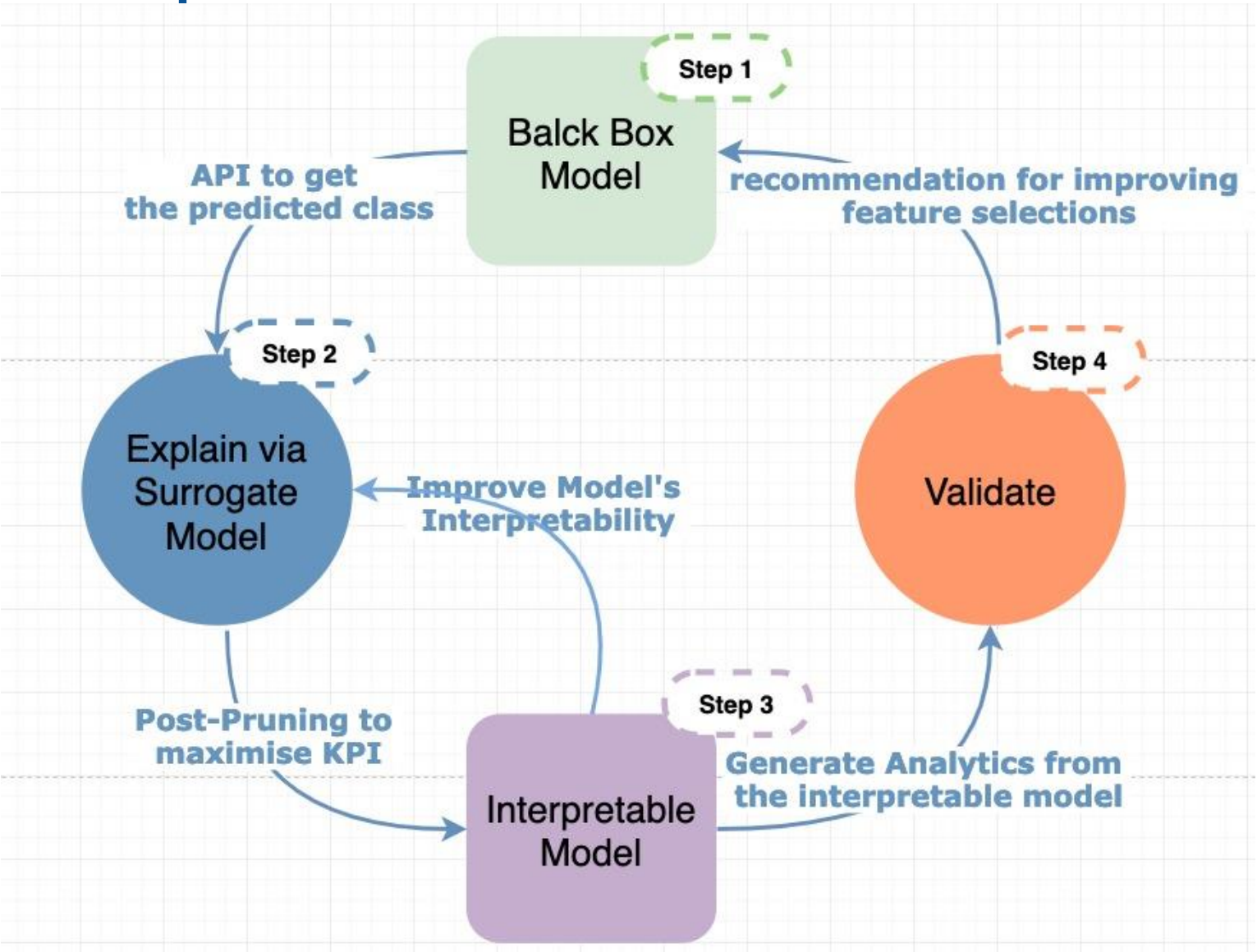


Основное внимание: объяснить и проверить

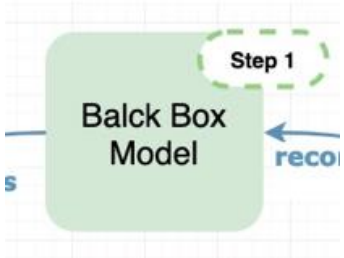
Might this help?



Основное внимание: объяснить и проверить



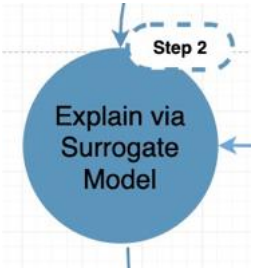
Основное внимание: объяснить и проверить



Классифицировать каждую вакансию по критериям классификации системы

Примечание: последовательность предварительной обработки представляет собой уникальное вычисление по модулю языка, т.е. некоторые термины, которые не учитываются в задаче классификации (например, «старший», «Милан» и т.п.) не отбрасываются, поскольку они полезны для выявления секторов, опыта и прочего.

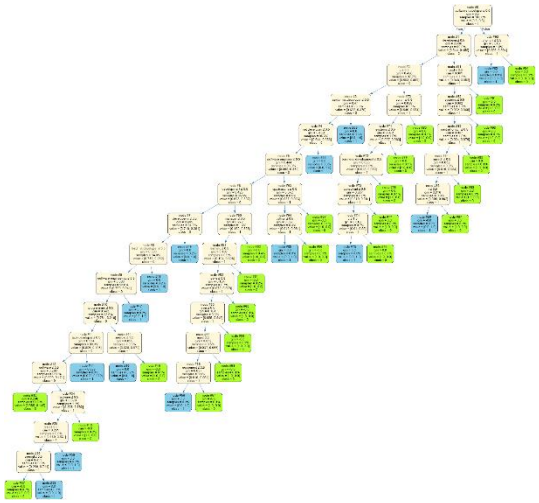
Основное внимание: объяснить и проверить



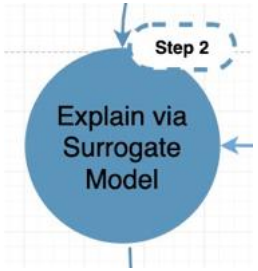
Использование глобальных интерпретируемых моделей для объяснения каждого класса (т.е. одной из более 400 профессий ESCO)

В результате получаем дерево, иллюстрирующее поведение системы при классификации вакансий, относящихся к такой профессии ESCO

Разработчик ПО→



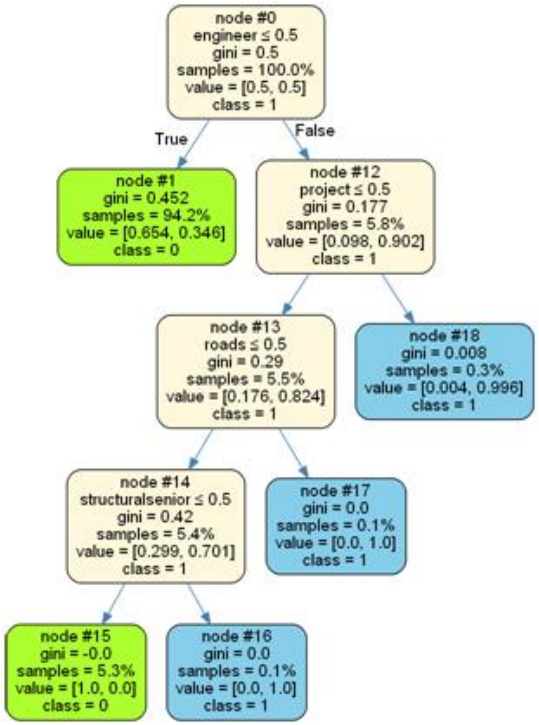
Основное внимание: объяснить и проверить



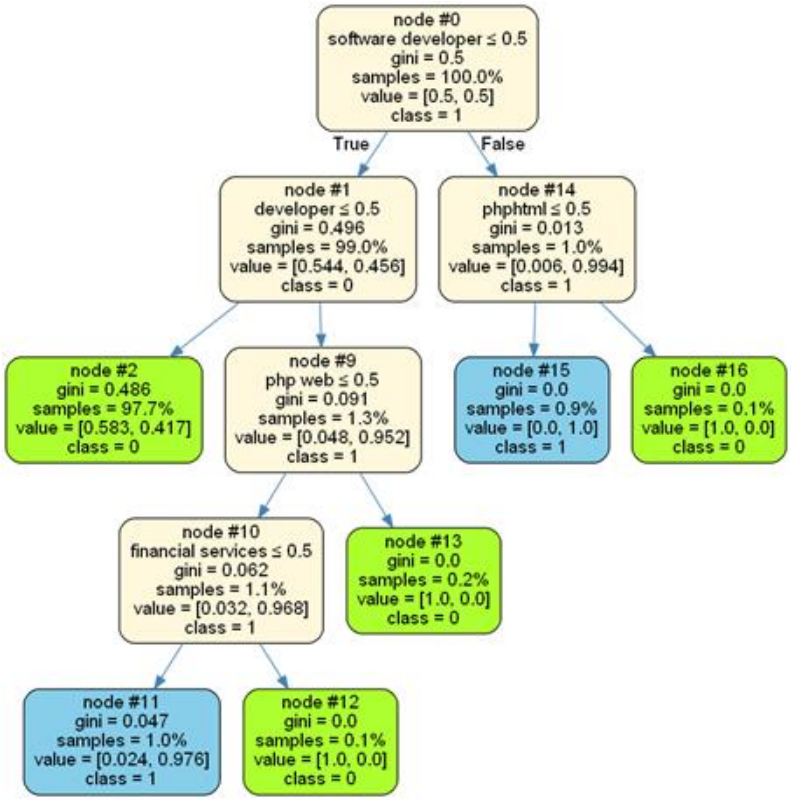
По сути, это позволяет заглянуть внутрь системы классификации

Её можно увидеть в форме набора вложенных операторов if-then-else, которые раскрывают критерии системы

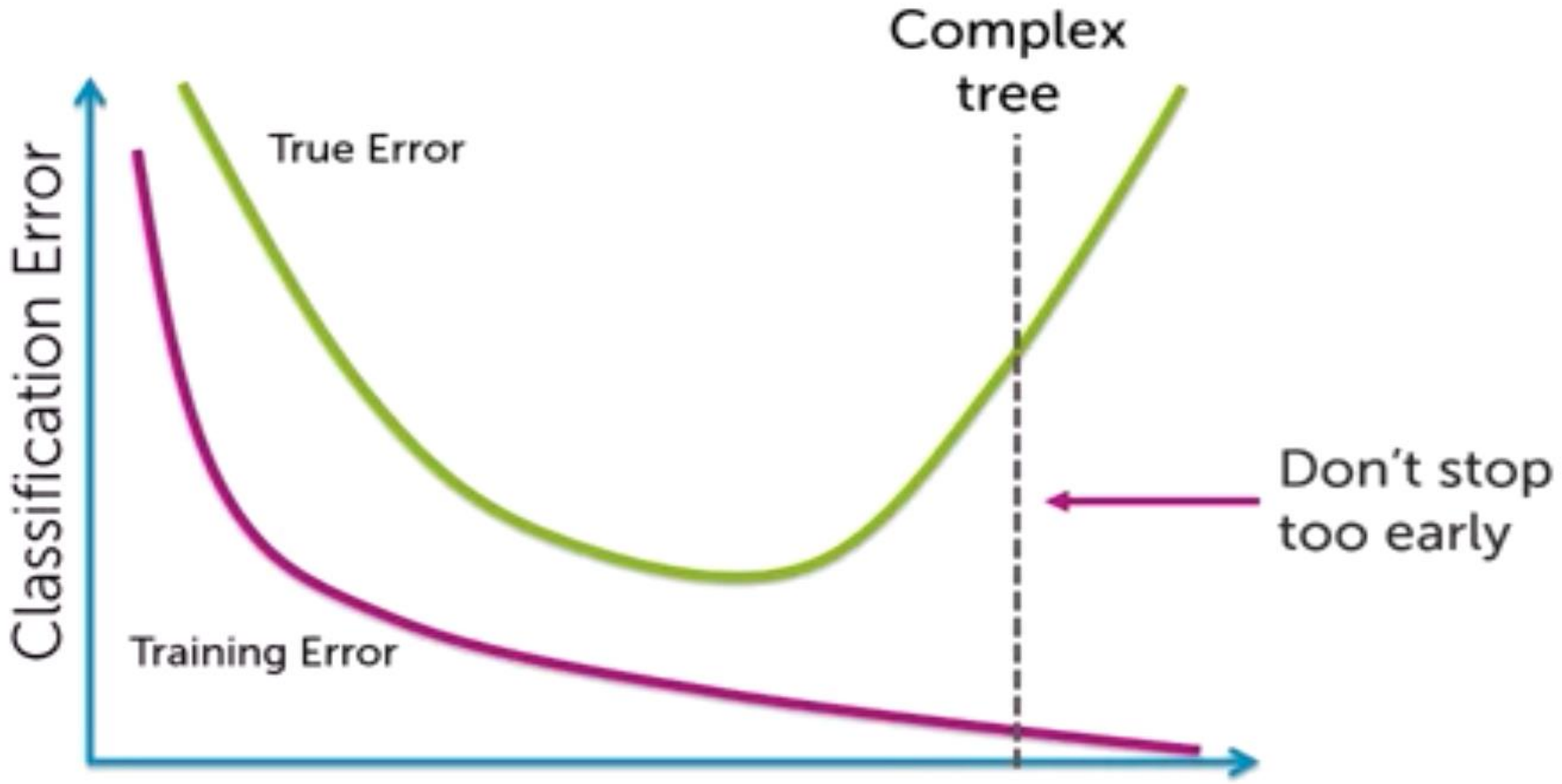
Инженер-строитель



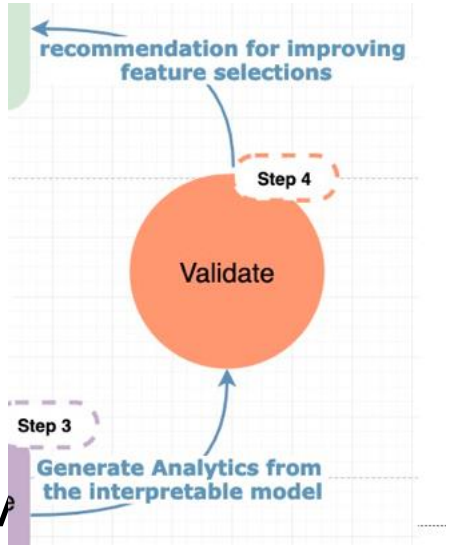
Разработчик ПО



Основное внимание: объяснить и проверить



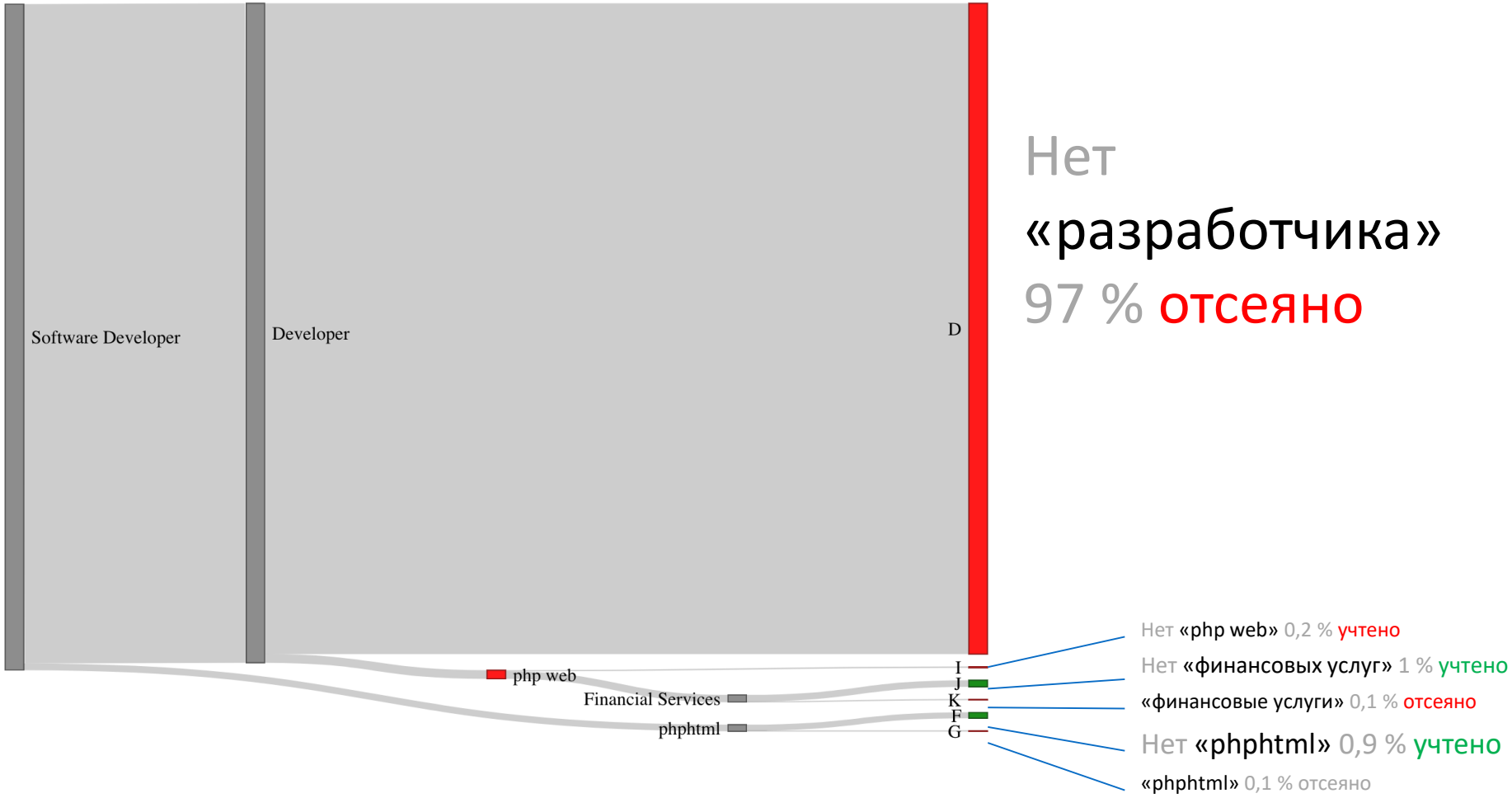
Основное внимание: объяснить и проверить



Переход от **интерпретаций** к **объяснениям**

После того, как система проинтерпретирована, мы используем визуализацию для получения объяснений в понятной для человека форме

Основное внимание: объяснить и проверить



Вызов

Как быть с проблемой репрезентативности?
[в работе]