

Big Data for Labour Market Intelligence

Day 3, Session 7

Toward «Smart LMI»

Challenges, best practices and lessons learned

Alessandro Vaccarino – Mauro Pelucchi

November 2021

Topics

1. Goal & context
2. Challenges
 1. Demand and Supply side analysis
 2. Visualization
 3. Off the shelf AI solutions

Topics

1. Goal & context

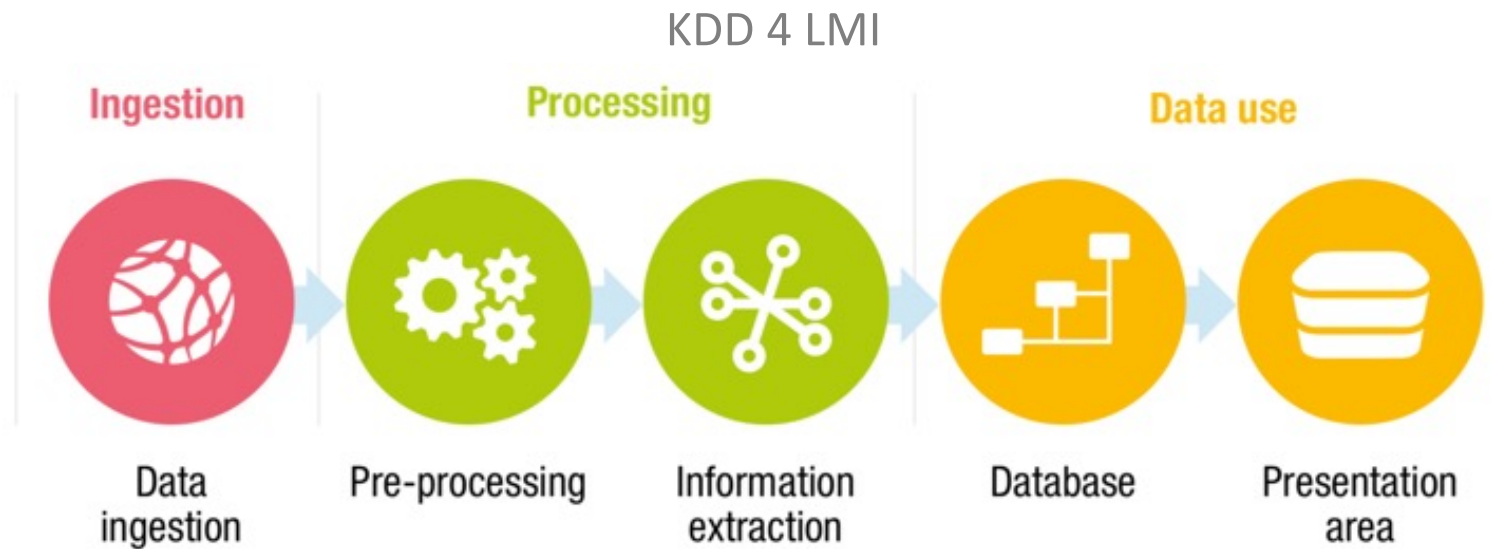
2. Challenges

1. Demand and Supply side analysis

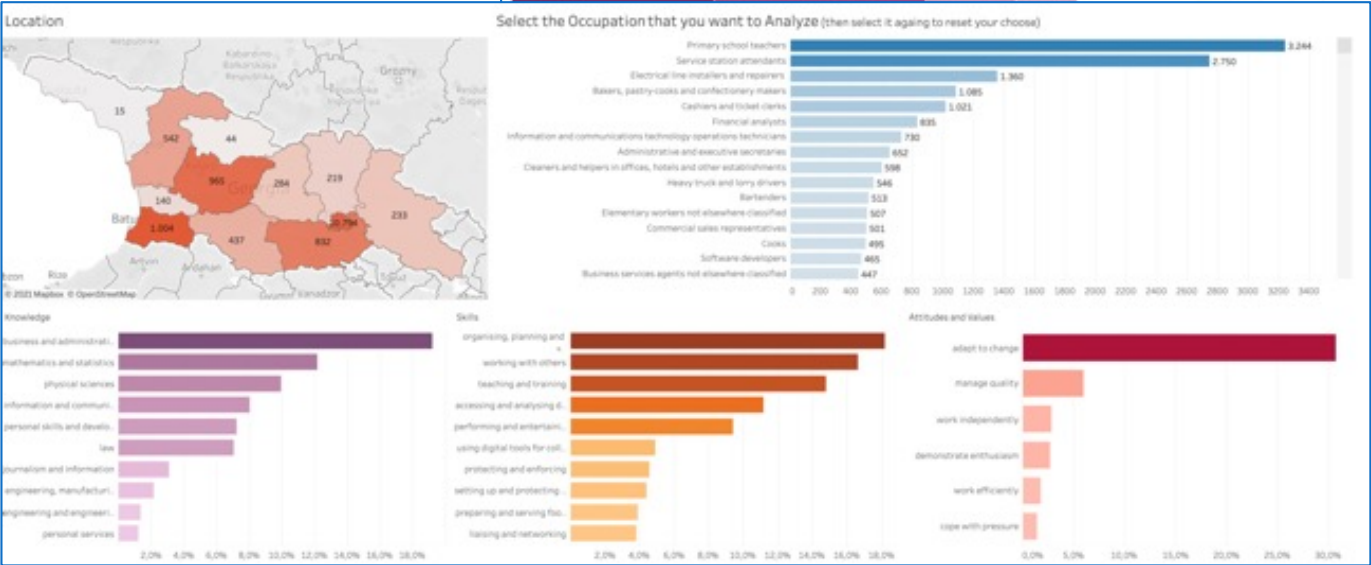
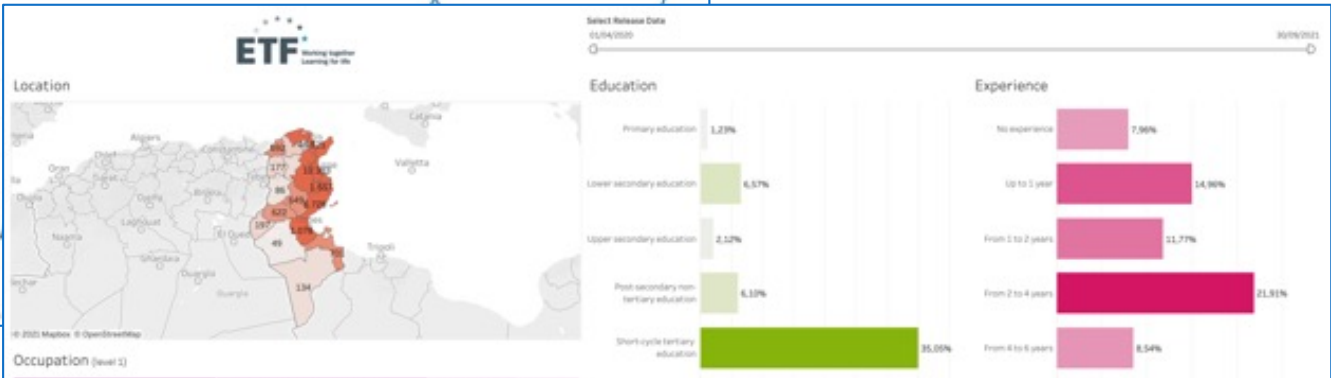
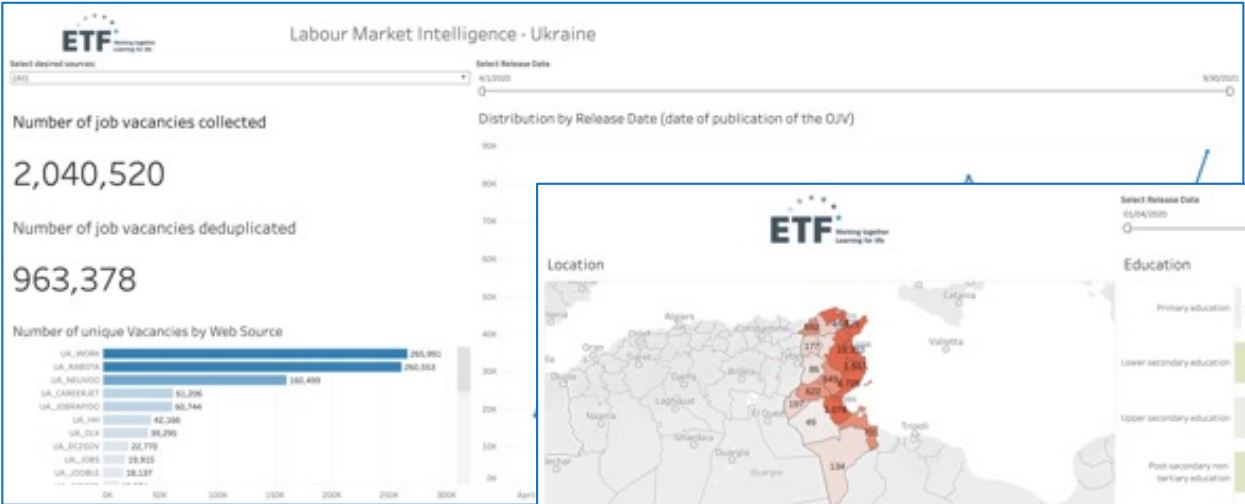
2. Visualization

3. Off the shelf AI solutions

Our starting point



Our starting point



...and now?

- This is a starting point
 - Big Data can provide us a gold mine of information
 - What we've seen together is just a starting point, the description of what's happening
 - One defined a methodology and a system able to collect and classify Big Data, we can extend our view and approach novel analyses

Novel analyses?

- What can we obtain?
 - Integration of new data
 - Tailored dissemination to different stakeholders
 - Re-use of existing components and knowledge
 - Definition of new *angle of analyses*

Let's see some examples

Topics

1. Goal & context
2. Challenges
 - 1. Demand and Supply side analysis**
 2. Visualization
 3. Off the shelf AI solutions

Why?

- Demand analysis is one of the points of view of Labour Market
- We can obtain additional and complementary information analyzing supply side
 - Offered skills
 - Matching demand-supply
 - Evolution of professional profiles
 - ...

How?

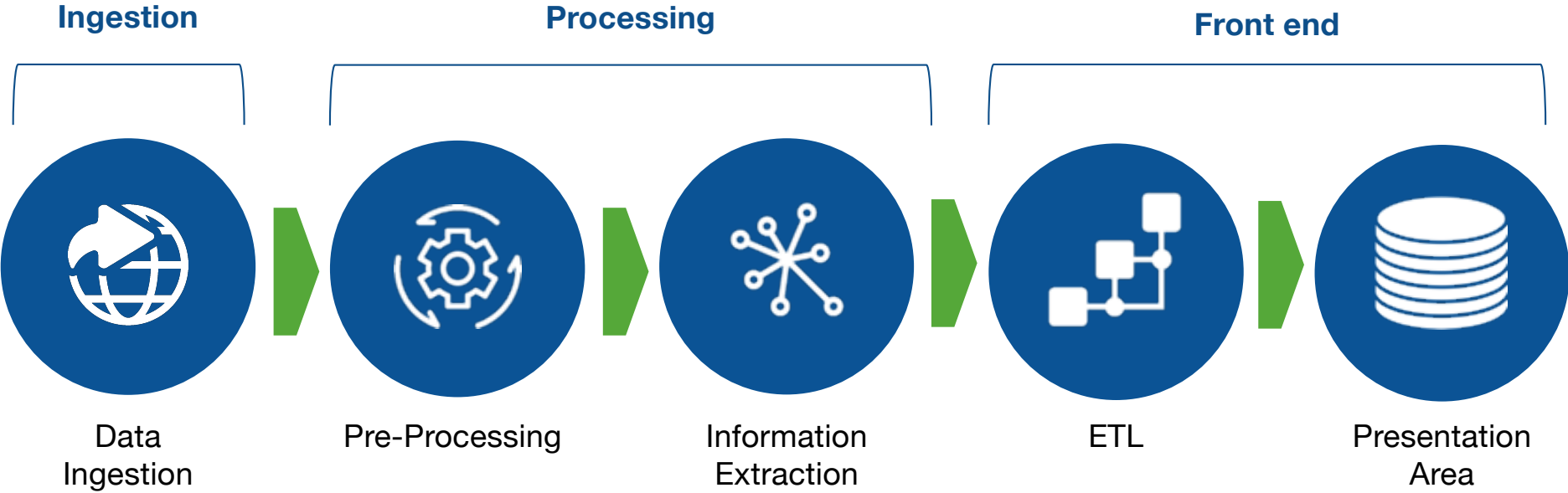
- We need to find additional information source that can help us understanding how supply side is evolving
- Data source must cover:
 - Professional profiles
 - Occupations
 - Skills
 - ...

Which sources can provide similar information?

Supply side data source

- Resumes
 - Detailed
 - Adherent to reality
 - Up-to-date
 - **Unstructured**
- Social profiles (e.g. LinkedIn)
 - Detailed
 - Adherent to reality
 - Up-to-date
 - **Semi-structured**
 - Private source (data ingestion needed)

Supply side processing flow



Information extraction

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.

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

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
03/2018 - Present

Data Scientist/Data Engineer
TabulaeX (part of Burning Glass Technologies)
11/2017 - Present

Project Manager & BI Solution Architect
Aubay Italia SpA for Amissima Assicurazioni SpA
05/2013 - 11/2017

SQL Specialist
Aubay Italy SpA for AXA Assicurazioni SpA
06/2013 - 03/2016

.NET-SAS Specialist
Aubay Italy SpA for Allianz Insurance SpA
02/2013 - 10/2013

Area:

Milan, Italy

Title:

Lecturer

Title:

Data Scientist

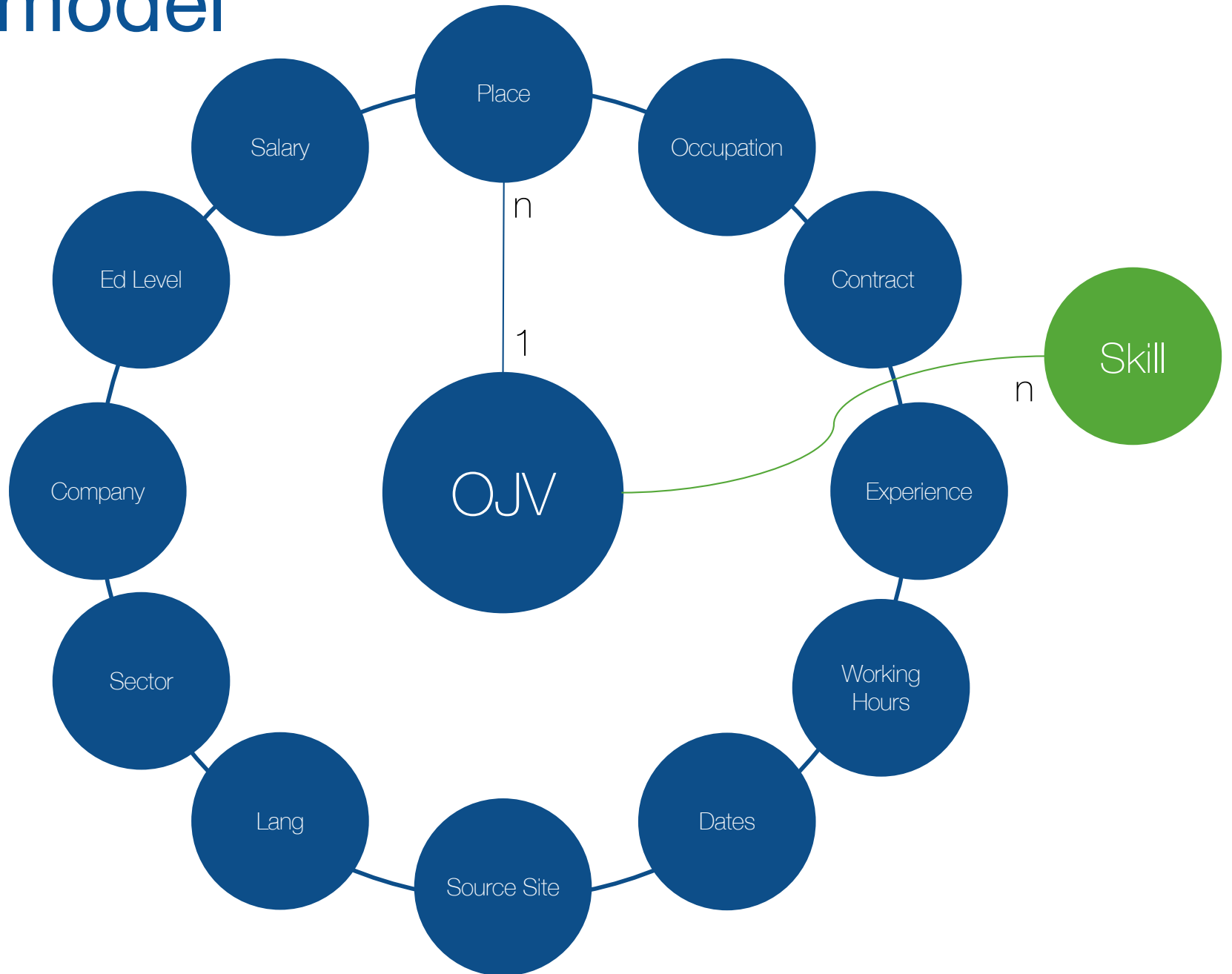
Title:

Data Engineer

Skills:

Data Governance, Data Quality, ETL Processes, Project Management, ...

Data model



Possible data analysis paths - 1

- Supply analysis
 - Most offered professions
 - Top trending skills
 - Geographies with the highest professional availability
 - ...

See it in action, with an Italian Job Agency example

Possible data analysis paths - 2

- Demand and Supply match
 - Compare skillsets offered by professionals and requested by market
 - Integrate with additional sources coming from training providers (e.g. universities)
 - Detect skill mismatching and training path, aimed to ensure proper strategies on:
 - Upskilling
 - Reskilling

See it in action, with an Italian University example

Opportunities and challenges

- Social profiles and resumes, as well as job postings, provides:
 - Actual professional profiles and skills
 - Deep information
 - Up-to-date insights
- In addition, lexicon can be linked to the same taxonomies adopted for demand analysis, ensuring that data comparison will be consistent and straightforward

Opportunities and challenges

- On the other hand, data availability is more challenging than demand side
 - By their own nature, postings must be public on the internet
 - Resumes are generally personal and not shared on public websites/pages
 - Professional Social Networks, such as LinkedIn, pay attention to the ownership of their data, denying access to external players
 - Resumes contain personal information, increasing privacy related challenges in data acquisition and processing

Topics

1. Goal & context
2. Challenges
 1. Demand and Supply side analysis
 - 2. Visualization**
 3. Off the shelf AI solutions

Data Visualization

- **Goal:**
 - Present information transforming it in actionable insights
- **Challenges:**
 - Handle several types of stakeholders, with different goals and skillsets
- **Approach:**
 - Develop an adaptable framework, composed by different tailored approach, that can ensure:
 - Find the **right solution for the right need**
 - Ensure **consistency** of data, even through different data analysis platforms
 - Propose a **scalable** solution, that can support from data visualization to advanced data analytics needs

What we have

2 Tables

FT Document

- 1 Row for each:
- General_ID → Key
 - OJV
 - Source
 - Place

Why? Because, for each OJV, we can detect a multi-place Vacancy
e.g. «Software Developer in London / Liverpool»

FT Skill Analysis

- 1 Row for each:
- General_ID → Key
 - OJV
 - Source
 - Place
 - Skill

Why? Because, for each OJV, we can obviously detect multiple skills
e.g. «Software Developer in London / Liverpool, with customer orientation culture, that speaks english and tolerates stress»

What we need



Project
Leader



Key
Users



Domain
Experts



End
Users



Citizens



Organizations



Decision
Makers



Analysts

What we need



Citizens

Where is my professional profile required?



Organizations

What's the demand trend for ICT related occupations?



Decision Makers

Are my University courses aligned with actual market needs?



Analysts

How green economy is evolving in my country?

What we need



Citizens

Low domain skills
Low analytical skills
Low depth of information
High standardization of insights



Decision
Makers

High domain skills
Medium analytical skills
Medium depth of information
Medium standardization of insights

High domain skills
Low analytical skills
Medium depth of information
High standardization of insights



Organizations

High domain skills
High analytical skills
High depth of information
Low standardization of insights



Analysts

What do we need to provide



Citizens

Something **easy** to be accessed,
with **low** but **ready-to-be-
understood** information

KPIs

Something **easy** to be accessed,
with **medium** but **ready-to-be-
understood** information

Storytelling



Organizations



Decision
Makers

Something that can provide
more information to people
that knows **how to read and
interpret** it

Dashboarding

Something that can provide **full
access** to data to people with a
**strong domain/technical/analytical
background**

Data Lab



Analysts

Data Visualization. Just graphic?

- **Goal:**
 - Provide the right tool for the right stakeholder
- **Challenges:**
 - Find a way to support different needs on such a relevant amount of information
- **Approach:**
 - Define several data visualization and analysis approaches:
 - **Infographics:** appealing, static and widely understandable
 - **Public portals for citizens:** easy, fast and high level informative
 - **Dashboard:** deeper informative, web based, for decision makers
 - **Self-service analysis labs:** access data, highest informative opportunity, requires domain/technical/analytical skillset

The right solution for the right need

- **Need:**
 - High level indicators
 - No domain background needed
 - No risk of misinterpretation
- **Approach:**
 - KPIs
 - Blog approach: few numbers, verbose interpretation



Citizens

See it in action

The right solution for the right need

- **Need:**
 - Medium level indicators
 - Domain background needed
 - No risk of misinterpretation
- **Approach:**
 - Storytelling
 - *Guided dashboard*: standard analytical pattern, that guide the user across different insights



Organizations

See it in action

The right solution for the right need

- **Need:**
 - Detailed indicators
 - Domain and analytical background needed
 - No risk of misinterpretation
- **Approach:**
 - Dashboard
 - Free analysis in a pre-defined environment



Decision
Makers

See it in action

The right solution for the right need

- **Need:**

- No indicator
- Access to raw and cleaned data
- High risk of misinterpretation
- Need of a strong knowledge of:
 - Domain
 - Data Model
 - Methodology
 - Technology

- **Approach:**

- Data Lab
- A place where to query data using the preferred analytical solutions



Analysts

See it in action

Topics

1. Goal & context
2. Challenges
 1. Demand and Supply side analysis
 2. Visualization
 3. **Off the shelf AI solutions**

Why?

- We've seen a lot of AI components that can help us analyzing data:
 - Language Detection
 - Spam filters for OJAs
 - Deduplication filter
 - ESCO Occupation classifiers
 - Custom taxonomy classifiers
 - ...

Do we have a way to reuse them?

AI re-usage

Of course, model reuse in AI is a common and best practice:

- Reduced **cost** of development
- **Enhancement** efficiency
- Centralized **maintenance**
- Classification **consistency** across projects

Examples of AI re-usage are present even on the Demand-Supply example provided: we reused and adapted most of the classification pipeline, to ensure coherent data between OJAs' and Resumes' information

Collaboration platforms

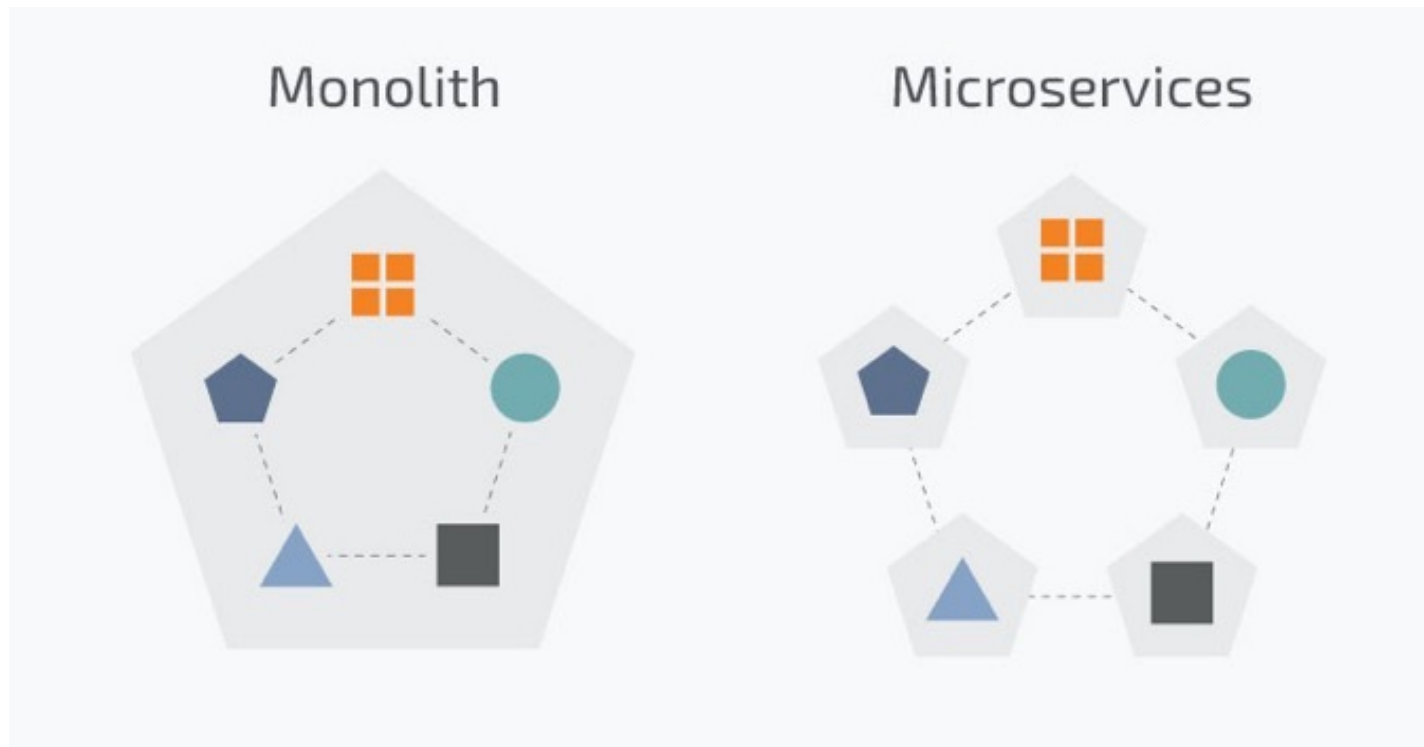
Several examples of collaboration and sharing platforms are available on the market:

- Databricks
- AWS SageMaker
- Anaconda
- H2O

Most of them aim to support community sharing and collaboration on the same data analysis pipeline, including machine learning model

Microservices

Modern technology enables us re-usage of software components as independent bricks that can be connected to build a more complex system.



Microservices

This architecture can ensure an efficient reuse of components, that will interact together sharing information.

It's not needed to know "*what's behind a service*" (e.g. an ESCO Occupation classifier) but will be just needed to know how "*how to interact with it*" (how to make a proper request and how to handle the response).

It's like one of our smart assistant (Alexa, Siri,...)

Challenges

What we've just discussed provides of course a great improvement to system scalability and reusage.

But

We need to keep attention to what we're doing, what we want to obtain and how the system we're using was built

Risks of off-the-shelf AI

Using a pre-build AI model is a great deal: we don't need to take care of the design, development and maintenance of it. We just have to *use* it.

But if we don't know how the system was designed, trained, built and how it's maintained, it could give us unexpected results.

An AI model (as well as any other Analytical system) is
not an oracle

Best practices for off-the-shelf AI

Within EUROSTAT collaboration, we're working on a Web Intelligence Hub that has the goal to provide a repository of AI components to ensure and maximize reuse and knowledge sharing.

It's crucial to ensure that the behavior of each component, as well as the methodology behind, is well documented. This will avoid unexpected results and misinterpretations of obtain results

*Let's think about posting deduplication
and Industry classification*

Posting deduplication issue

I know that a given portal contains 100 postings. Why do I find just 10? Is the system doing something wrong?

This is a typical question while analyzing OJA data.

The system is not giving a wrong result, we just need to know **what we're analyzing**

Posting deduplication issue

The company publishes a OJVs on a given portal **(A)**


To increase visibility, the company republish the OJV to a new portal **(C)**

The same OJV is detected By another portal **(B)** and republished



time

Posting deduplication issue



Vacancy	Title	Description	Sector	Publish date	Portal
0001	Sales Manager	Our company...	Manufacturing	01/01/2021	A
0002	Sales Manager	Our company...	Manufacturing	15/01/2021	B
0003	Sales Manager	For our company...	Manufacturing	01/02/2021	C

High similarity

Coherent timeframe

The three OJVs are detected as duplicates and the first one published (portal A) is elected as deduplicated OJV

Industry classification issue

We've official statistics saying that Education industry has a market share of 5%. Why do I find just 0.2 on the system? Is it doing something wrong?

This is another typical question while analyzing OJA data.

Again, we need to deeply understand **what we're analyzing**

Industry classification issue

Most countries have a public healthcare sector, that's hiring through public tenders.

In this case, the sector is not expressed through OJAs.

This is not a classification issue, but a *selection bias*: if we keep in mind this bias, we'll be able to better understand results provided

Best practices

We're not defending the system or the provided results: Big Data, as we've seen, can introduce relevant critical topics to be handled.

On the other hand, if we're able to deal with the risk that are present, that data will enable novel and disruptive analytics.

Don't trust this system, but understand and believe it

Thank you very much

Alessandro Vaccarino