Big Data for Labour Market Information

Session 1
General overview of Big Data for LMI

Alessandro Vaccarino – Fabio Mercorio

Big Data for Labour Market Information – focus on data from online job vacancies – training workshop
Milan, 21-22 November 2019
Topics

1. Big Data at a Glance
2. So what’s AI? [by examples]
3. Big Data for LMI
Big Data at a Glance

“Big Data” usually refers to large amounts of different types of data produced with high velocity from a high number of various types of sources.

Making these data useful for stakeholders requires to turn these data into knowledge, as the knowledge is the end product of a data-driven discovery.
The 4V’s Big Data model

- **Volume**: Data size
- **Velocity**: Speed of change
- **Variety**: Different forms of data sources
- **Veracity**: Uncertainty of data
Data grows fast

1 billion GB → 2.5 exabytes → that number doubles every month

90% of the world’s data was created in the last two years

More iPhones are sold than babies born
Not just “a lot of data”
Big Data are nothing without «Artificial Intelligence» that derive knowledge from them
Artificial Intelligence: A changing definition

Haugeland (1985)
The exciting new effort to make computers think ...
*machines with minds*, in the full and literal sense

Rich & Knight (1991)
The study of how to make computers do things at which, at the moment, people are better

Schalkoff (1990)
A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes

EU - AI for Europe (2018)
systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals
AI: A multidisciplinary approach

Big Data: the fuel of AI

- Statistics
- Pattern Recognition
- Machine Learning
- Artificial Intelligence
- Knowledge Discovery & Reasoning
- Database
Due macro tipologie di AI

**Narrow (weak) AI:** able to perform single tasks (play chess, recommend products, forecast, etc.). The context and tasks are defined.

**General (strong) AI:** able to reason, take decisions autonomously, and perform an undefined number of tasks as a human. The context and tasks are not defined (reality).
Videos will be showed
AI Planning

INPUT:
- Maps
- Initial Condition (traffic, GPS, etc..)
- Goal (destination)

OUTPUT:
- Plan (min time/km/etc)
A software that learns to perform a task using its experience, and it increases its experience by improving its ability to perform the task for which has been designed over time.
Machine Learning: Due categorie

Unsupervised

Supervised
Unsupervised Learning
Unsupervised Learning
Machine Learning: Due categorie

Unsupervised:

The system classifies items having similar and common characteristics (feature) on the basis of a similarity criterion. The results vary as the classification criterion varies.
Supervised Learning (Learning Phase)

Training Set
(the bigger, the better)

Machine Learning Algorithm
Supervised Learning (Evaluation Phase)

Score: 92% accuracy
Supervised Learning (in production)
Machine Learning: Due categorie

**Supervised**
The system classifyies items having similar features on the basis of the characteristics found during the training phase. The test phase just allows one to know how good the system performed the training phase. There is no way to know how good the system will be in production phase (i.e., working on novel items never seen)
Supervised Learning: Issue

Dataset **must be:**
- Big
- Labeled (ground truth) by domain experts

Pros/Cons
- ML good in **NON mission critical** applications as they fall in explaining to humans the rationale behind their behaviours... **eXplainable AI**
Deep Neural Network

So strong... so weak
Explainable AI – l’AI explains itself to humans

Today

- Training Data
- Learning Process
- Learned Function
- Output
- User with a Task

Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow

- Training Data
- New Learning Process
- Explainable Model
- Explanation Interface
- User with a Task

I understand why
- I understand why not
- I know when you’ll succeed
- I know when you’ll fail
- I know when to trust you
- I know why you erred
Domini che richiedono explainability:

-- Medicina
-- Trasporti
-- Militare
-- Finanza
-- Legale
...
Figure 1. Hype Cycle for Big Data, 2012

Source: Gartner (July 2012)
Figure 1. Hype Cycle for Artificial Intelligence, 2017
So, how Big Data and AI can interact to derive knowledge from data?

...towards «Data Science»
a useful service

analyse user behavior to extract insights
data science

(hopefully)

transform insights into action
data products
Before Big Data

DATA SOURCES
OLTP, ERP, CRM, LOB

ETL

DATA WAREHOUSE
Star schemas, views, other read-optimized structures

BI AND ANALYTICS
Emailed, centrally stored Excel reports and dashboards

MONITORING AND TELEMETRY

Well manicured, often relational sources
Known and expected data volume and formats
Little to no change

Complex, rigid transformations
Required extensive monitoring
Transformed historical into read structures

Flat, canned or multi-dimensional access to historical data
Many reports, multiple versions of the truth
24 to 48h delay
Top Down approach
After Big Data

All data sources are considered
Leverages the power of on-prem technologies and the cloud for storage and capture
Native formats, streaming data, big data

Refineries transform data on read
Produce curated data sets to integrate with traditional warehouses
Users discover published data sets/services using familiar tools
Bottom Up Approach

- Ingest all data regardless of requirements
- Store all data in native format without schema definition
- Do analysis using analytic engines like Hadoop

Batch queries
Interactive queries
Real-time analytics
Machine Learning
Data warehouse
How to put all those data?

It is a “lake” of data where:

- Incoming flows are input data that can have many form/structure
- Outcoming flows are output data, that are the analysed data
What is Data Lake?

How do Data Lakes Work?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

1. Structured Data
   - Information in rows and columns
   - Easily ordered and processed with data mining tools

2. The reservoir of water is a dataset, where you run analytics on all the data.

3. The outflow of water is the analyzed data.

4. Through this process, you are able to “sift” through all the data quickly to gain key business insights.

Unstructured Data

- Raw, unorganized data
- Emails
- PDF files
- Images, video and audio
- Social media tools
How to allow machines to process those data such that the data Volume does not affect performances?
Scale up
Scale out
How Big Data changed the way of doing LMI?

Top-Down Deductive
- Theory
- Hypothesis
- Observation

Bottom-Up Inductive
- Theory
- Hypothesis
- Observation
How Big Data and AI are related to Labour Market?

Labour Market Intelligence

LMI
LMI at a glance

- Labor market intelligence (LMI) is a term that is emerging in the whole labor market community, especially in the European Union.
- There is no unified definition of what LMI is, it can be referred to the design and use of AI algorithms and frameworks to analyze data related to labor market (aka labor market information) for supporting policy and decision-making.
Needs: new tools for LMI

• Famous study of Frey and Osborne (THE FUTURE OF EMPLOYMENT, Oxford)
  • 47% of Jobs will disappear in the next 25 years.
• 65% of children entering primary school today (2017) will ultimately end up working in completely new job types that don’t yet exist.
• Huge implications in terms of skill requirements
  • Numbers are worrying but are they really true?
• We need to implement several complementary tools for investigating these changes
Why Big Data Analytics for LMI?

- Lacking data on skill demands by employers
- Conventional methods are
  - expensive
  - suffer from time lags
  - focus on specific types of skills
  - Surveys are rigid and lengthy tools
- Forecasting tools to identify the most relevant trends
  - But forecasting tools are necessarily imprecise about the features and skill requirements of the jobs of the future
- Skills anticipation
- Useful
  - Understand the real market demands
  - Inform career mobility and training choices
  - Fine-tune training offer
5 Vs of Big Data in the LMI context

Data Availability
Growing of Computational power
Advances in AI

The Rise of Big Data

VOLUME
- Num of Records
- Num of Sources
- Num of Countries

VELOCITY
- Real/Near - time by:
  - Crawling
  - Scraping
  - API access

VARIETY
- Structured
- Semi-Structured
- Unstructured

VERACITY
- Data Quality:
  - Consistency
  - Deduplication
  - Availability
  - Credibility

VALUE
- Understand LM dynamics for decision making according to Stakeholder needs
How Big Data changed the way of doing LMI?
Is Big Data a game changer in the field of labour market?
Three main Labour Market Sources can support LM Intelligence

1. Statistical sources
2. Administrative sources
3. Web Sources (Big Data 4 LMI)
Quo vadis Labour Market?

**LM CHALLENGING FACTORS**
1. Skills Evolution
2. New Emerging Occupations
3. Job Automatisation/Replacement
4. Mobility

**LM NEEDS**
1. Updated information (near-real-time)
2. Data driven decisions (let data speak)
3. Prediction can be done to anticipate trends

Knowledge becomes crucial to support different LM actors and policy makers in understanding LM dynamics and trends.
Web Labour Market Scenario

1. Job Vacancies frequently posted on specialised Web sources
   - Stakeholder Needs Identified: Near real-time labour market analysis
   - Proposed Research Actions: Data scraping from selected sources

2. Hidden informative power about labour market dynamics
   - Stakeholder Needs Identified: Labour Market Occupations/Skills Trend Monitoring
   - Proposed Research Actions: Job vacancy classification via machine-learning

3. Heterogeneous sources and different lexicons used in job vacancy texts
   - Stakeholder Needs Identified: Evaluate/compare International LM for fact-based decision making
   - Proposed Research Actions: Multi Language support through the use of Standard Taxonomies

4. Info about skills, industry sectors, territory, etc expressed as raw text within vacancies
   - Stakeholder Needs Identified: Analyse LM according to the identified dimensions
   - Proposed Research Actions: Query the resulting knowledge base over the identified dimensions
<table>
<thead>
<tr>
<th>LM Source Type</th>
<th>Data Type</th>
<th>Generation Rate</th>
<th>Data Model Paradigm</th>
<th>Quality</th>
<th>Coverage</th>
<th>Analysis Paradigm</th>
<th>Believability</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>Structured</td>
<td>Periodically</td>
<td>Relational</td>
<td>Owner’s responsibility</td>
<td>Owner’s responsibility</td>
<td>Top Down &amp; Model Based</td>
<td>Owner’s responsibility</td>
<td>intrinsic</td>
</tr>
<tr>
<td>Administrative</td>
<td>Structured or Semi-structured</td>
<td>Periodically</td>
<td>Relational</td>
<td>Owner’s responsibility</td>
<td>Owner’s responsibility &amp; User’s responsibility</td>
<td>Top Down &amp; Model Based</td>
<td>Owner’s responsibility &amp; User’s responsibility</td>
<td>intrinsic</td>
</tr>
<tr>
<td>Web</td>
<td>Structured, Semi-structured or Unstructured</td>
<td>Near-real-time or real-time</td>
<td>Relational and Non Relational (NoSQL)</td>
<td>User’s responsibility</td>
<td>User’s responsibility</td>
<td>Bottom up &amp; Data Driven</td>
<td>User’s responsibility</td>
<td>extrinsic</td>
</tr>
</tbody>
</table>
# Table 2 Most significant limitations of Big Data architectures

<table>
<thead>
<tr>
<th>Issue (most significant)</th>
<th>Caused by</th>
<th>Conceptual Blocks of Big Data Architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>Variety</td>
<td>Data ingestion; NoSQL models;</td>
</tr>
<tr>
<td>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.</td>
<td>Variety, Velocity</td>
<td>Data lake</td>
</tr>
<tr>
<td>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).</td>
<td>Variety</td>
<td>Schema free; data-driven approach (bottom-up rather than top-down)</td>
</tr>
<tr>
<td>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.</td>
<td>Volume, Variety, Velocity</td>
<td>Scale-out rather than scale-up</td>
</tr>
</tbody>
</table>