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

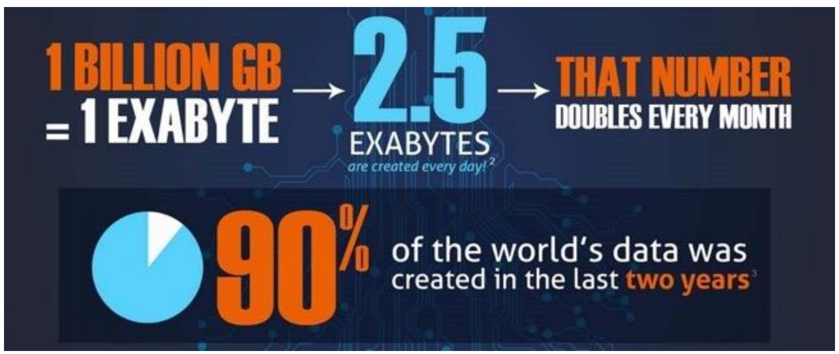
Big Data at a Glance

”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

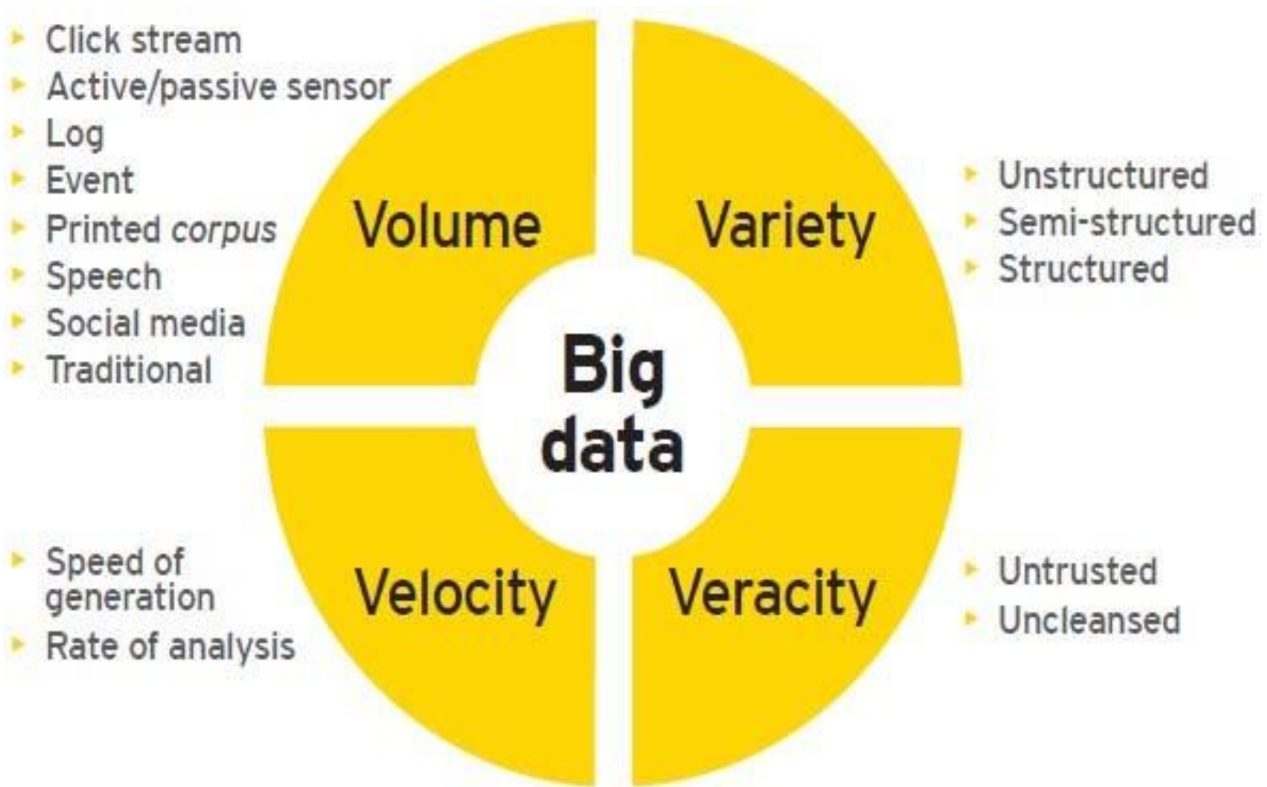
Making these data useful for stakeholders requires to turn these **data** into knowledge, **as the knowledge is the end product of a data-driven discovery**

The 4V's Big Data model

Data grows fast



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

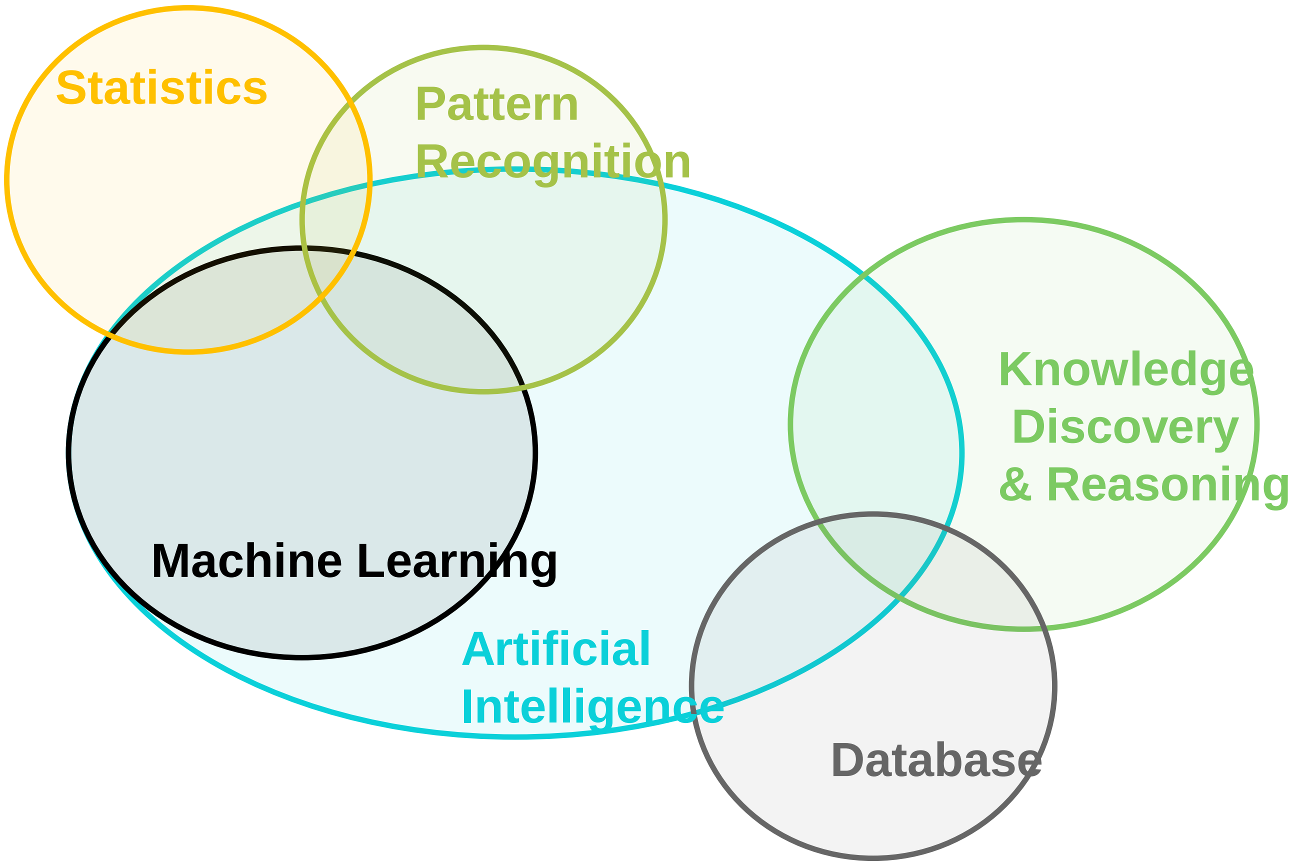
Schalkoff (1990)

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

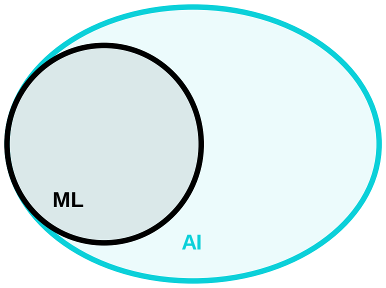
EU - AI for Europe (2018)

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



Due macro tipologie di AI

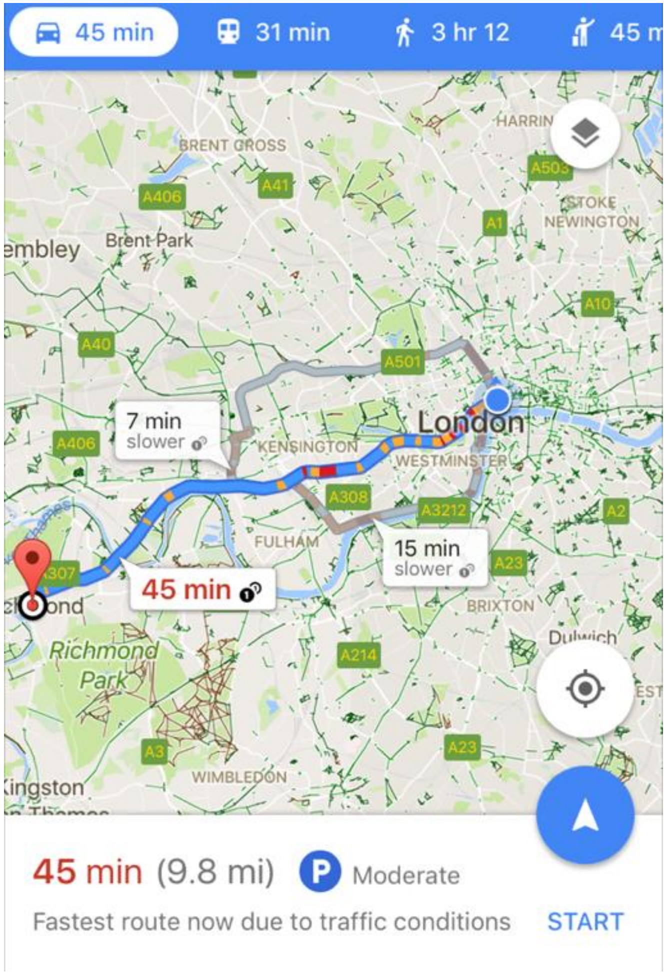


Narrow (weak) AI: able to perform single tasks (play chess, recommend products, forecast, etc.). The context and tasks are defined.

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



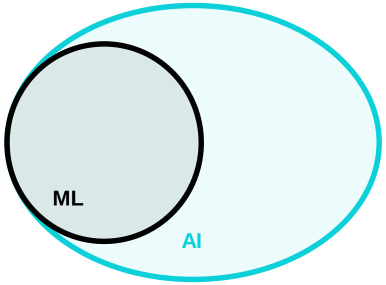
INPUT:

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

OUTPUT:

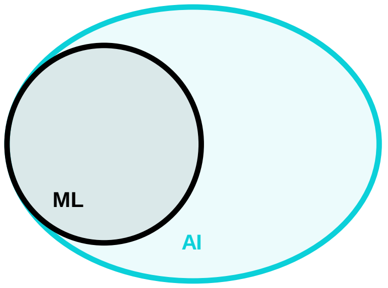
-Plan (min time/km/etc)

Machine Learning



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

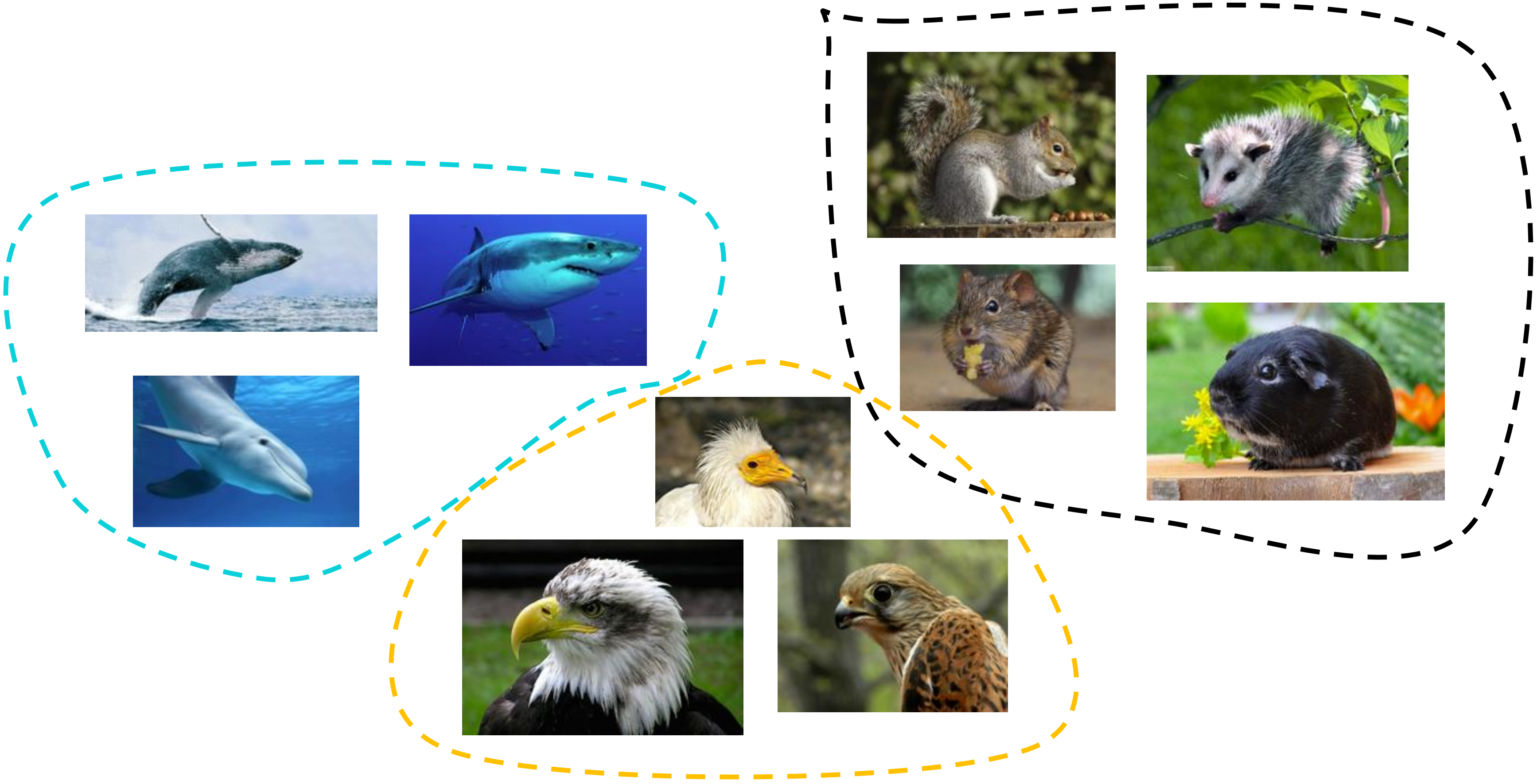
Machine Learning: Due categorie



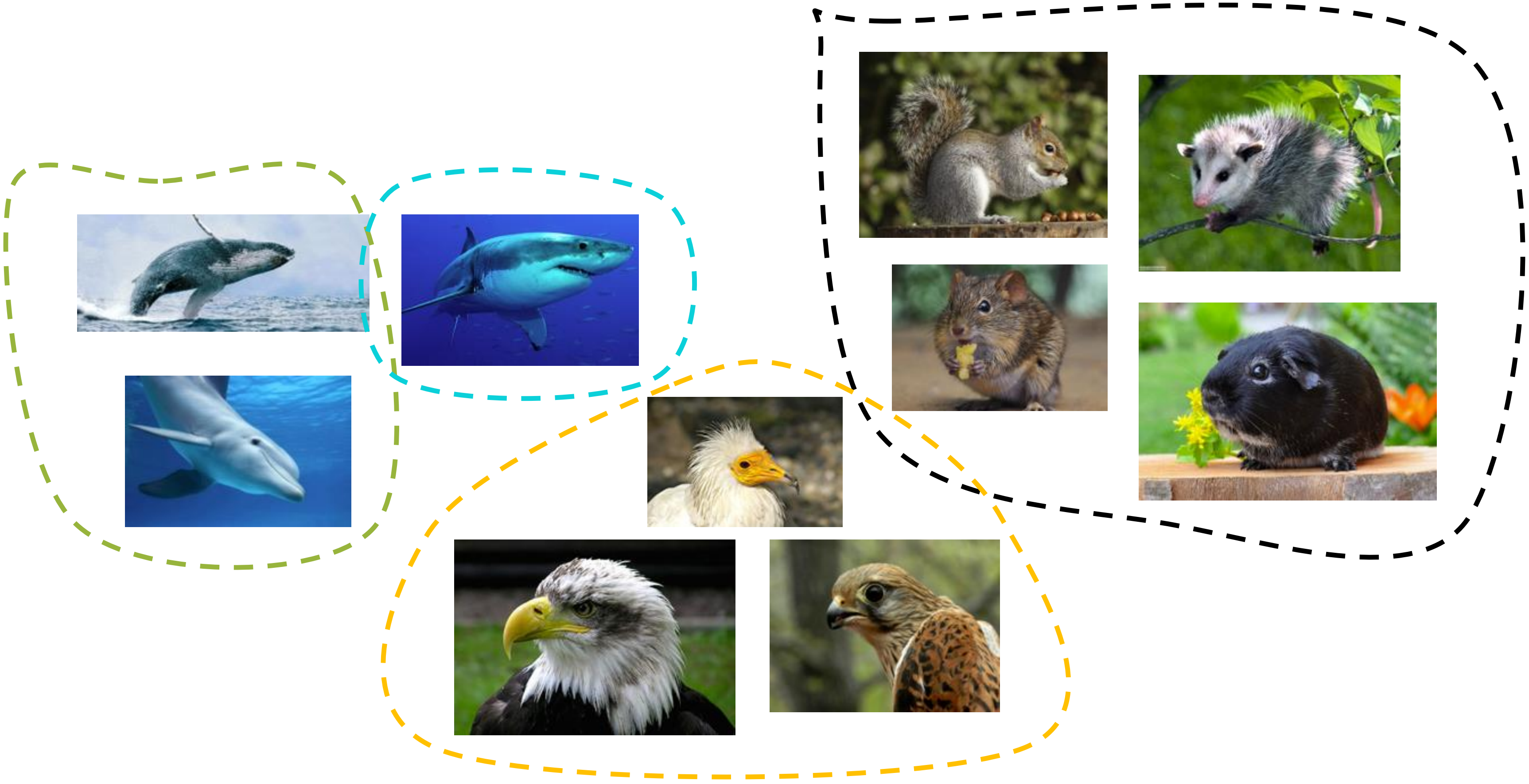
Unsupervised

Supervised

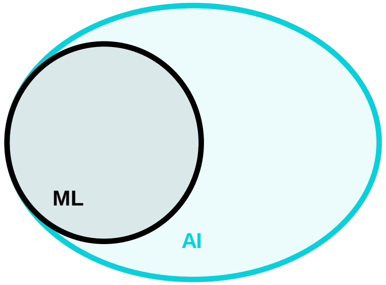
Unsupervised Learning



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

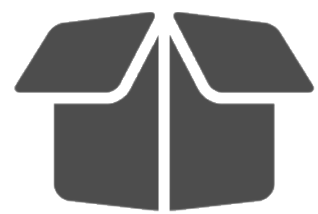
Supervised Learning (Learning Phase)

Squalo Bianco

Balena

Training Set

(the bigger, the better)



Machine Learning Algorithm

Supervised Learning (Evaluation Phase)

Squalo Bianco (V)

Balena (X)

Test Set

Machine Learning Algorithm

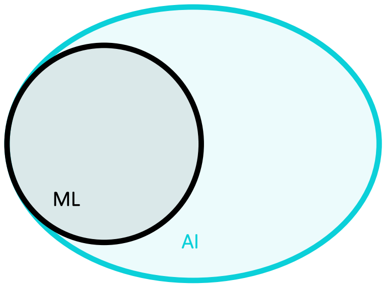
Score: 92% accuracy

Supervised Learning (in production)

Shark or Whale?

Machine Learning Algorithm

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**:

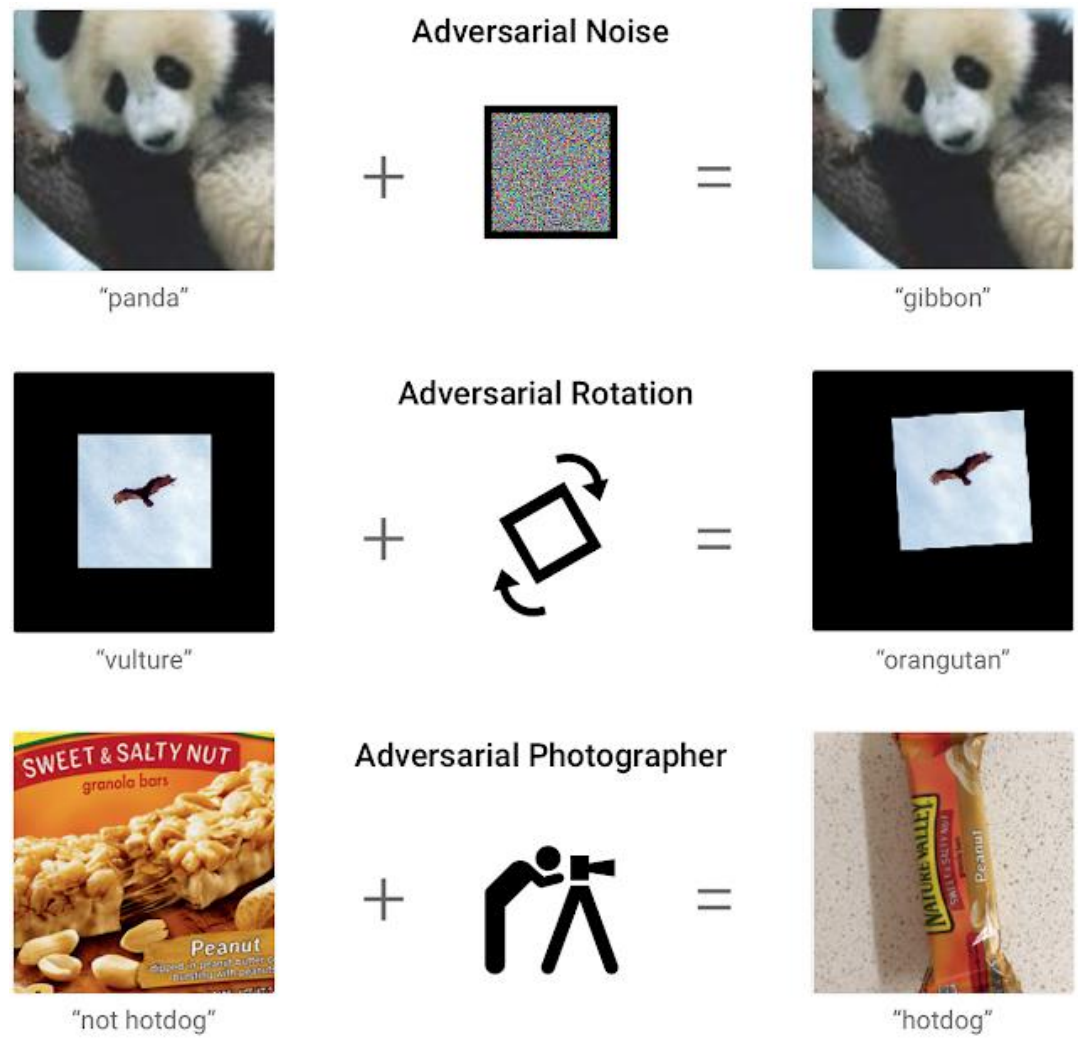
- Big
- Labeled (ground truth) by domain experts

Pros/Cons

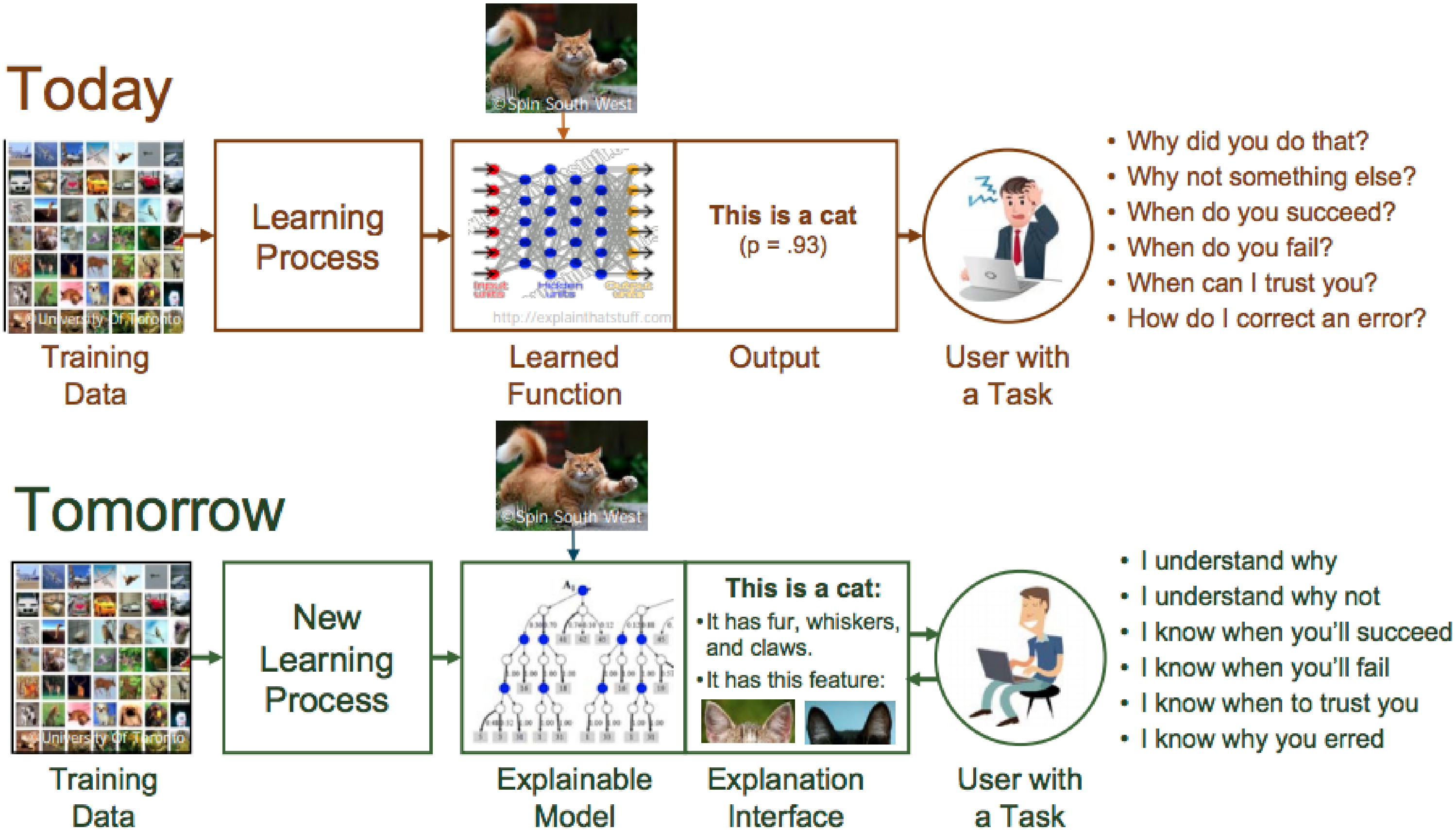
- ML good in **NON mission critical** applications as they fall in explaining to humans the rationale behing their behaviours… **eXplainable AI**

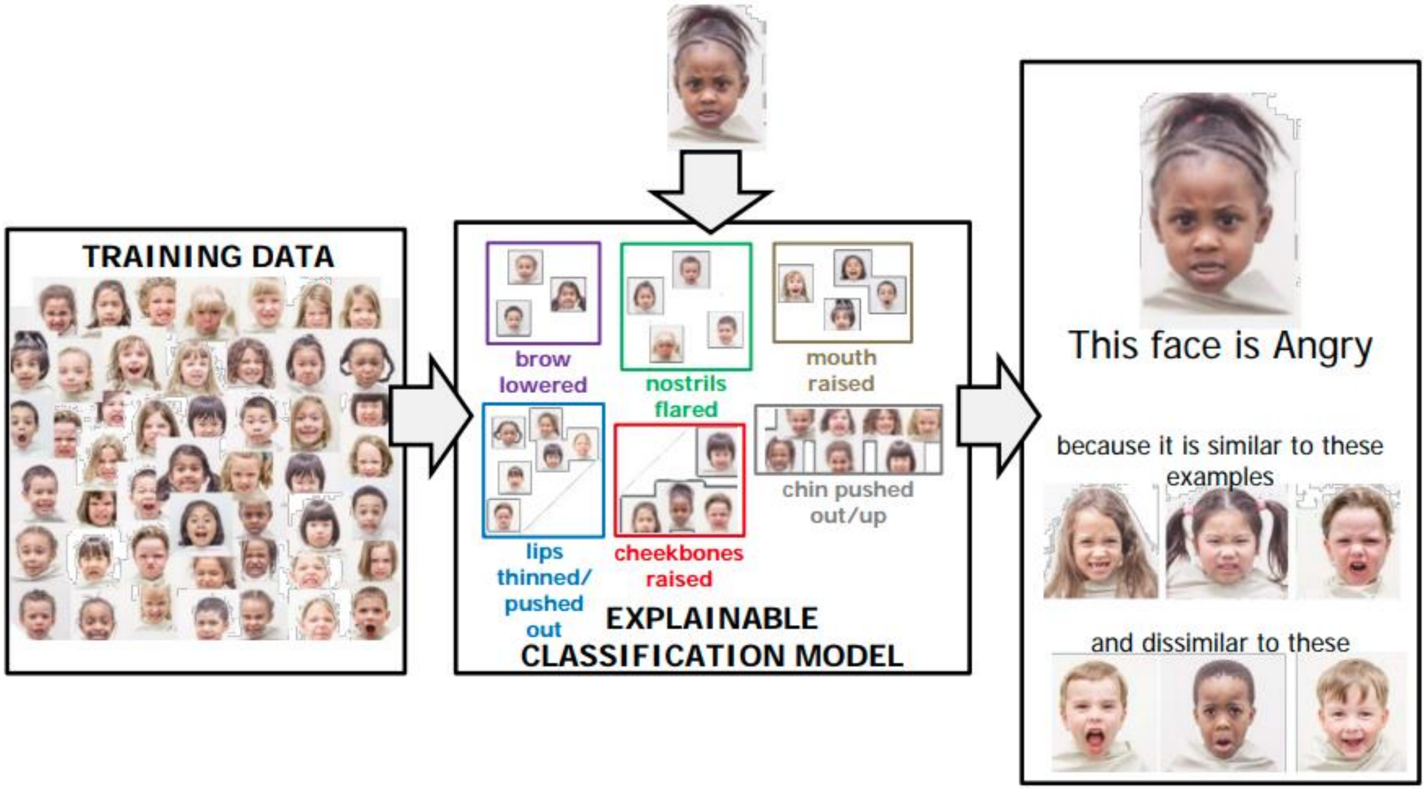
Deep Neural Network

So strong…
so weak



Explainable AI – l’AI explains itself to humans

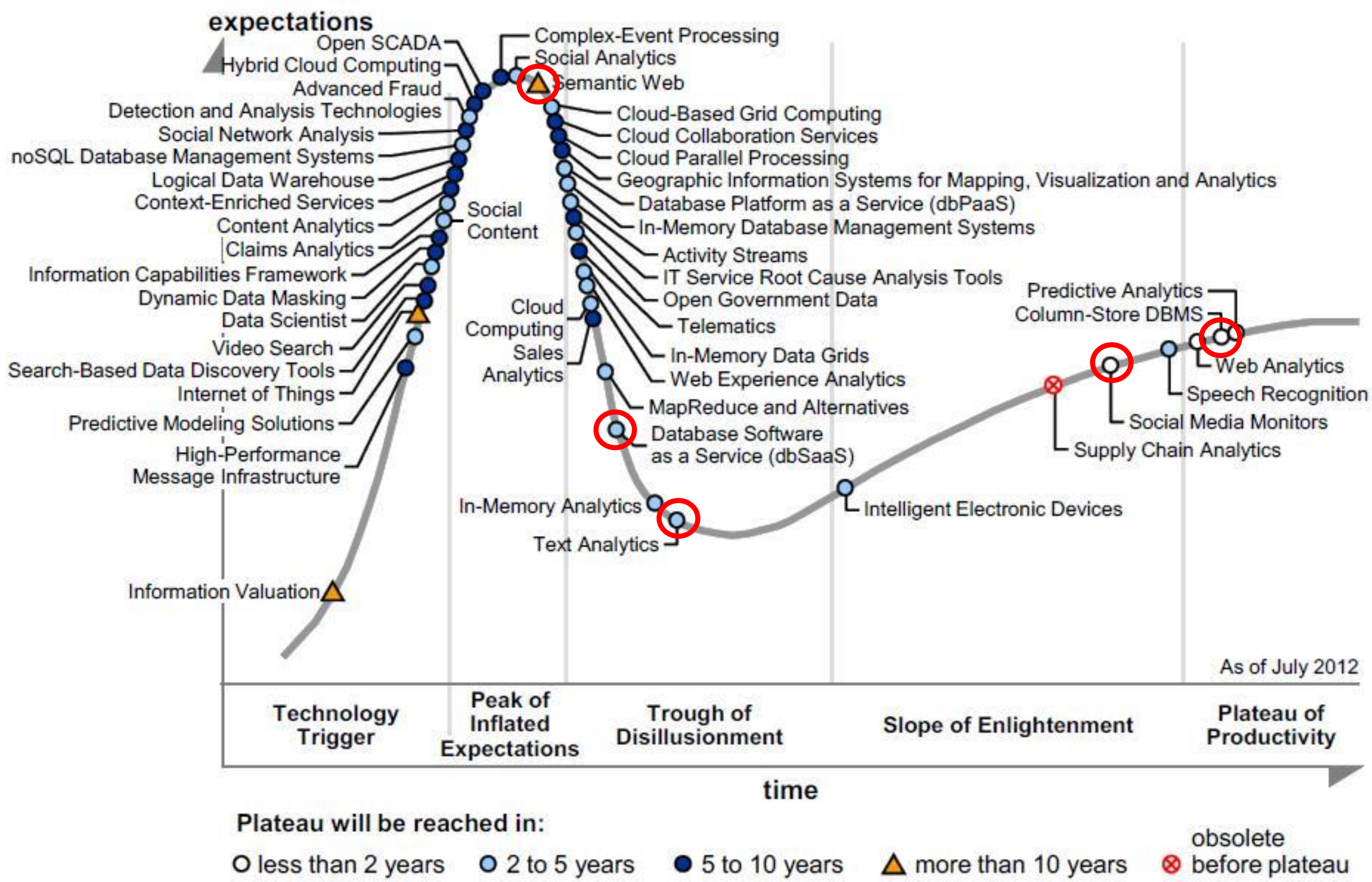




Domini che richiedono explainability:

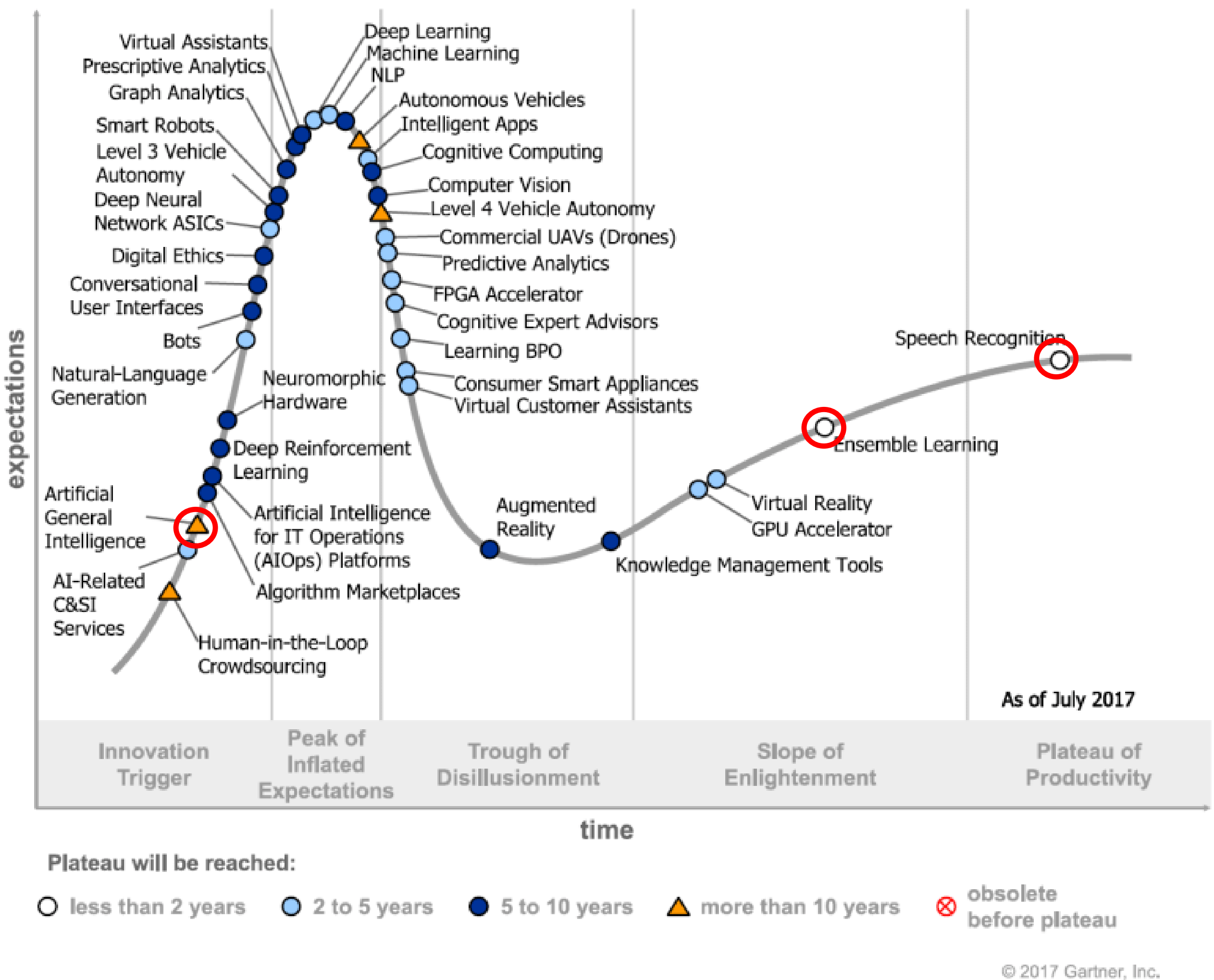
-- Medicina
-- Trasporti
-- Militare
-- Finanza
-- Legale
…

Figure 1. Hype Cycle for Big Data, 2012



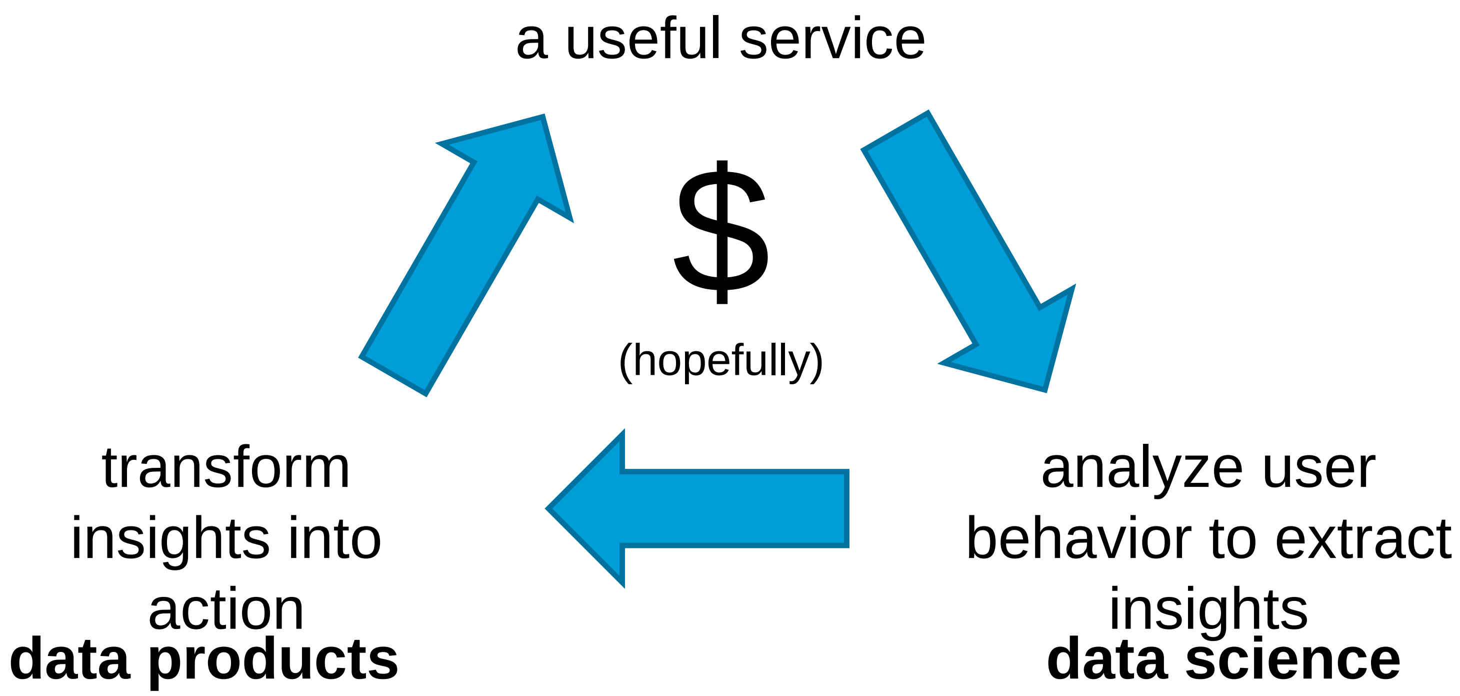
Source: Gartner (July 2012)

Figure 1. Hype Cycle for Artificial Intelligence, 2017

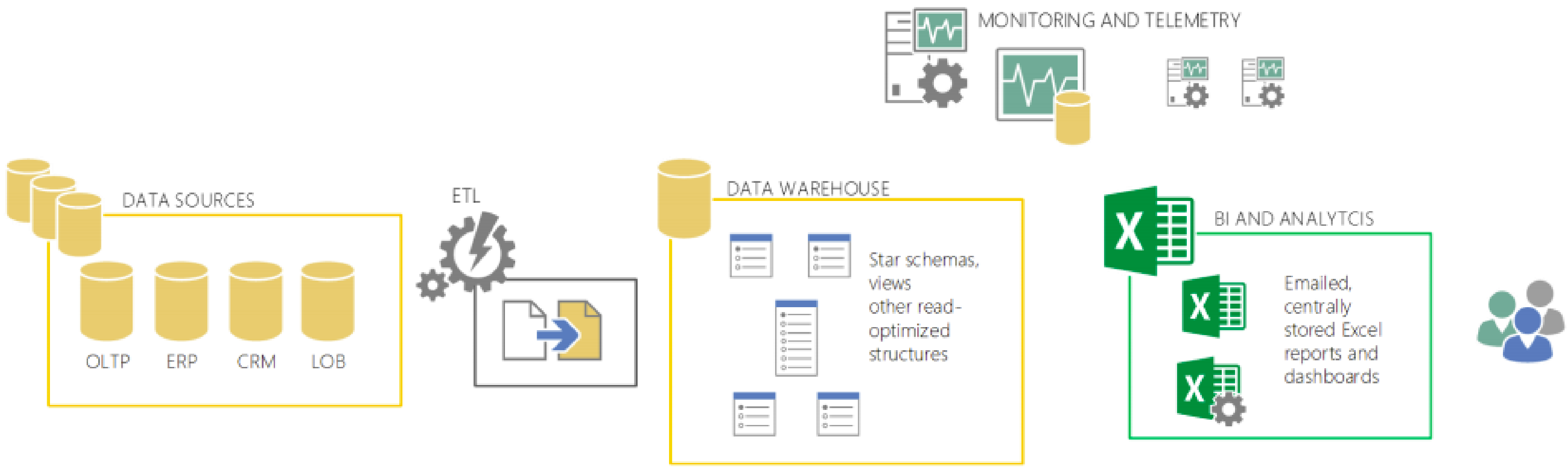


So, how Big Data and AI can interact to derive knowledge from data?

…towards «Data Science»


a useful service
$
(hopefully)
transform insights into action
data products
analyze user behavior to extract insights
data science

Before Big Data



Well manicured, often relational sources

Known and expected data volume and formats

Little to no change

Complex, rigid transformations

Required extensive monitoring

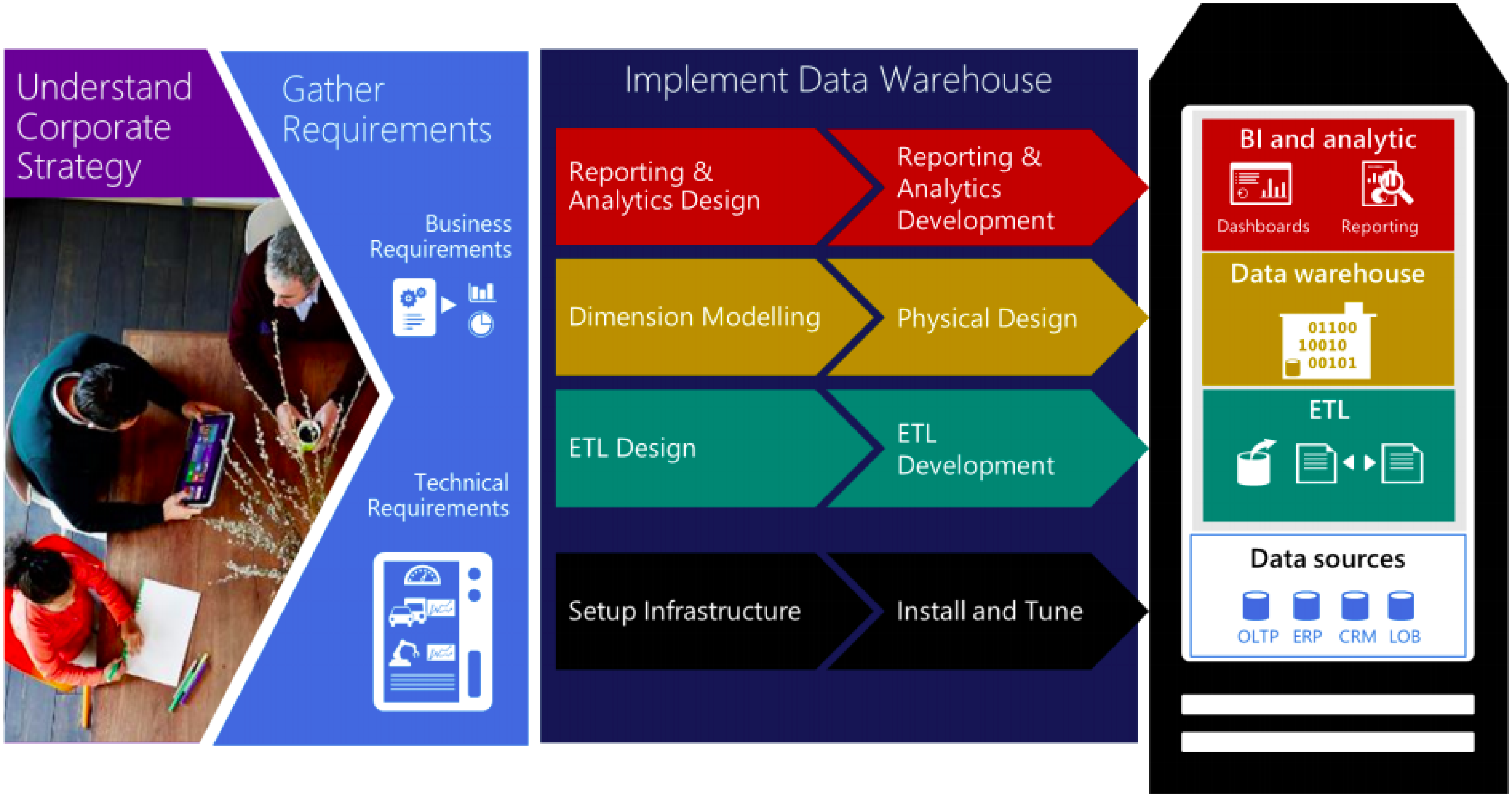
Transformed historical into read structures

Flat, canned or multi-dimensional access to historical data

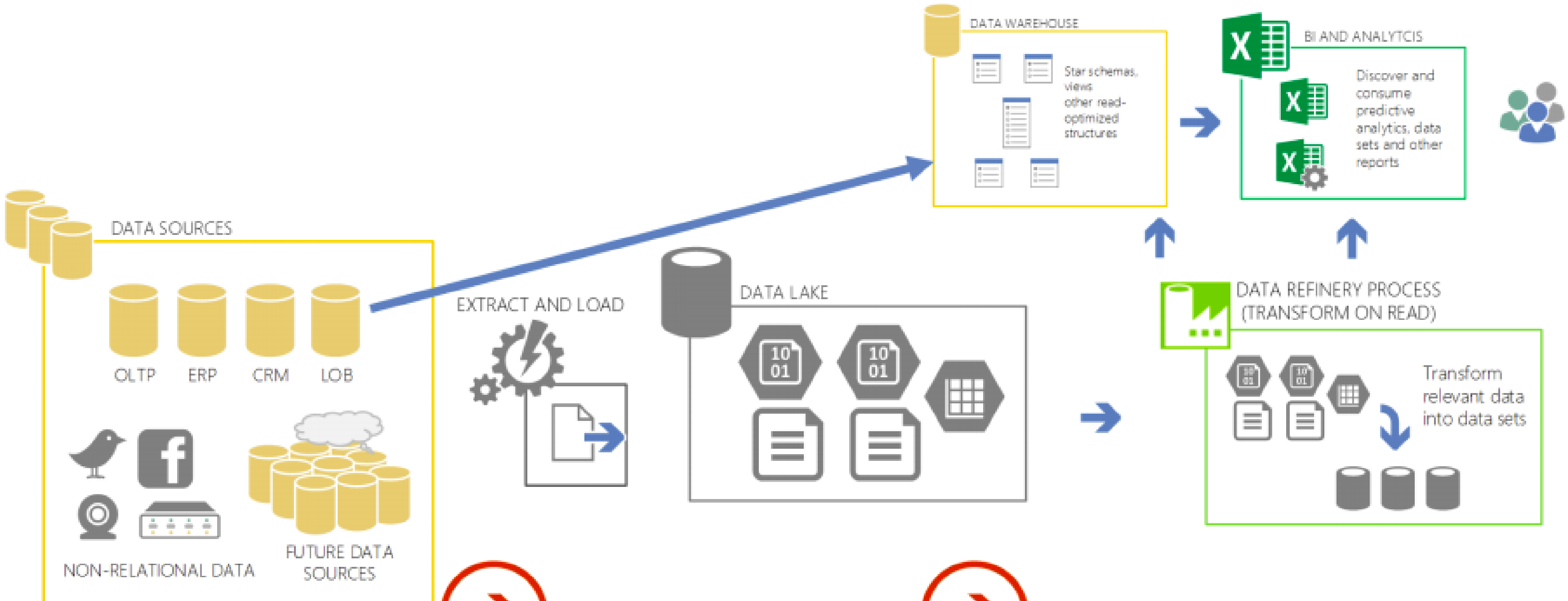
Many reports, multiple versions of the truth

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

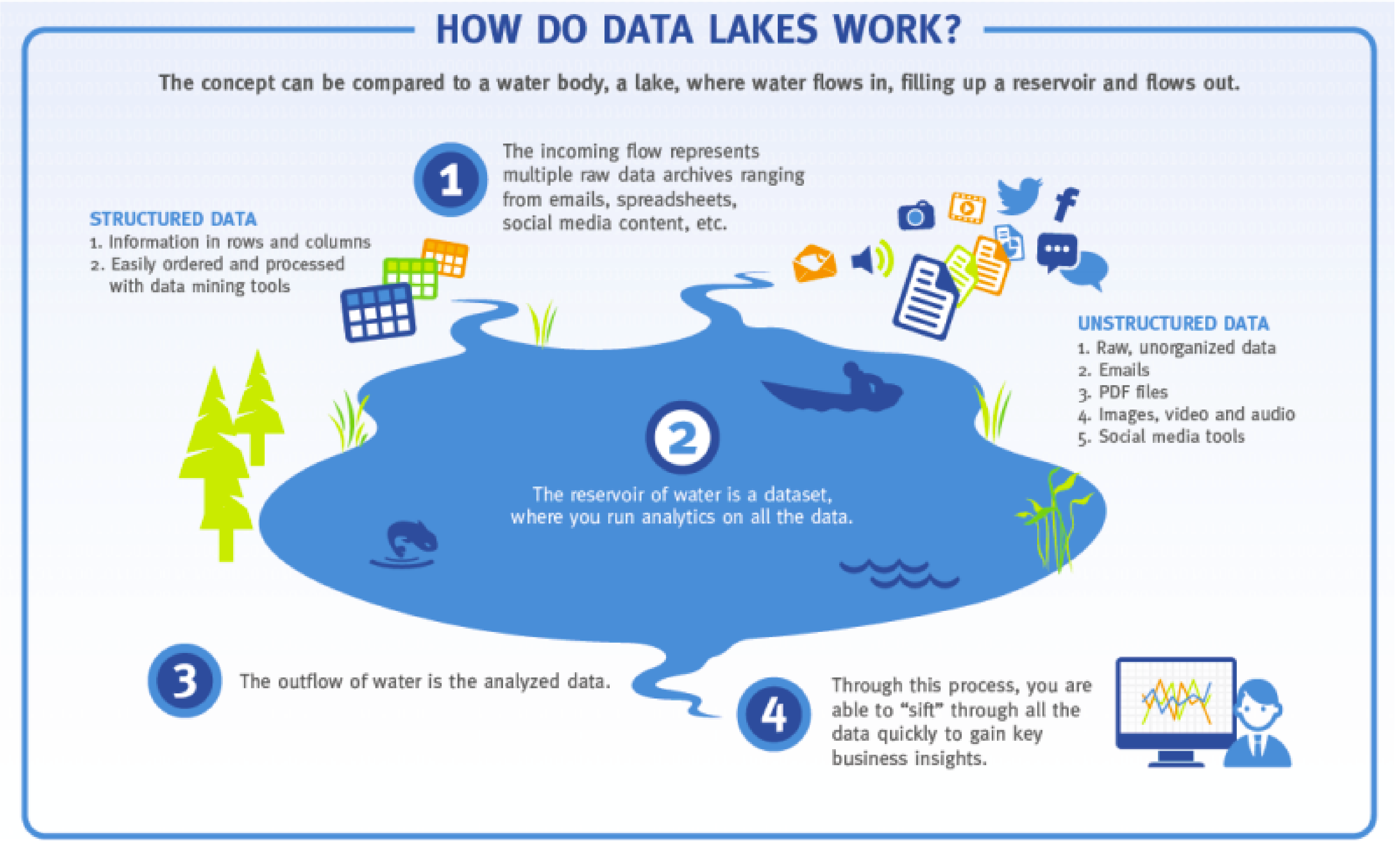
Bottom Up Approach

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?

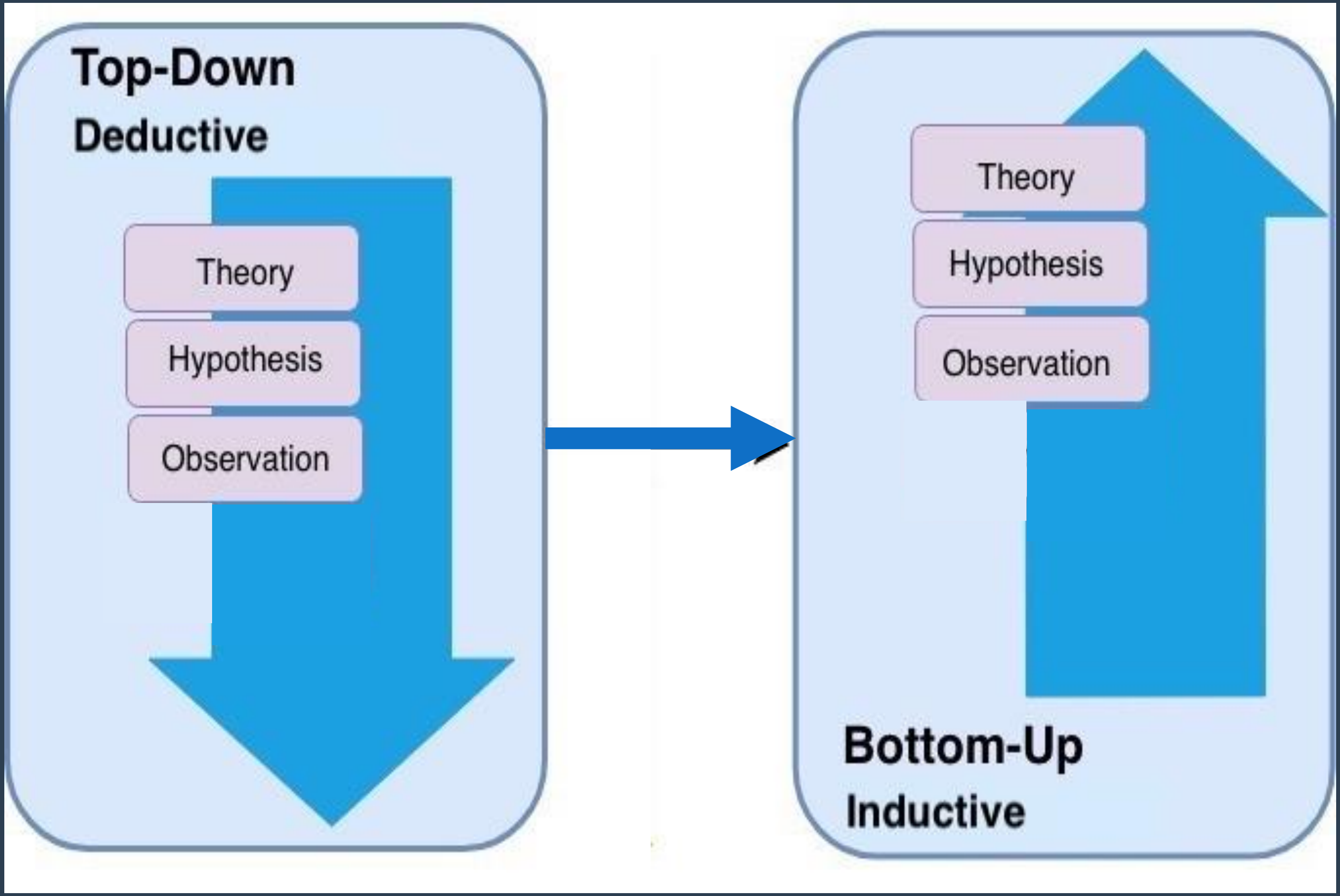


How to allow machines to process those data such that the data Volume does not affect performances?

Scale up

Scale out

How Big Data changed the way of doing LMI?



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**

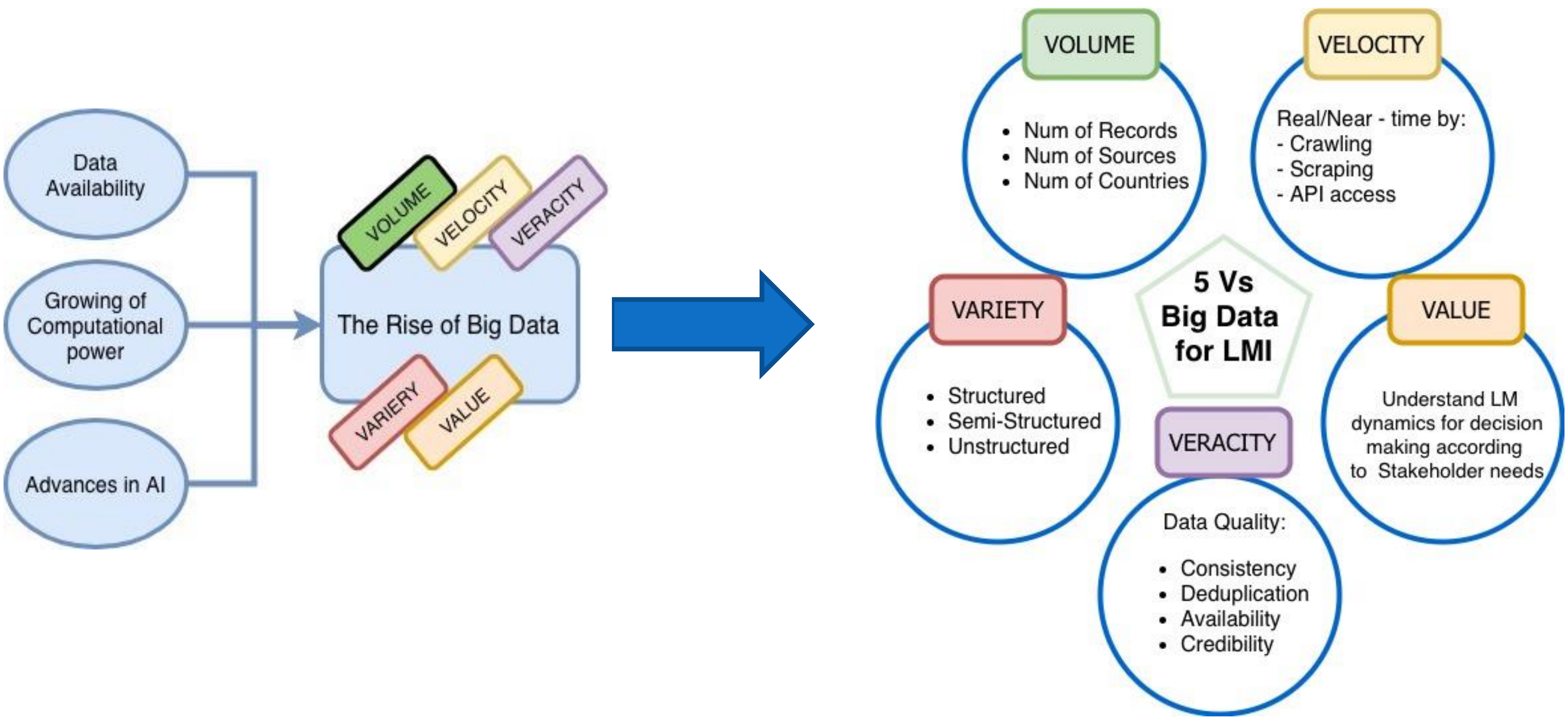
Needs: new tools for LMI

- 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

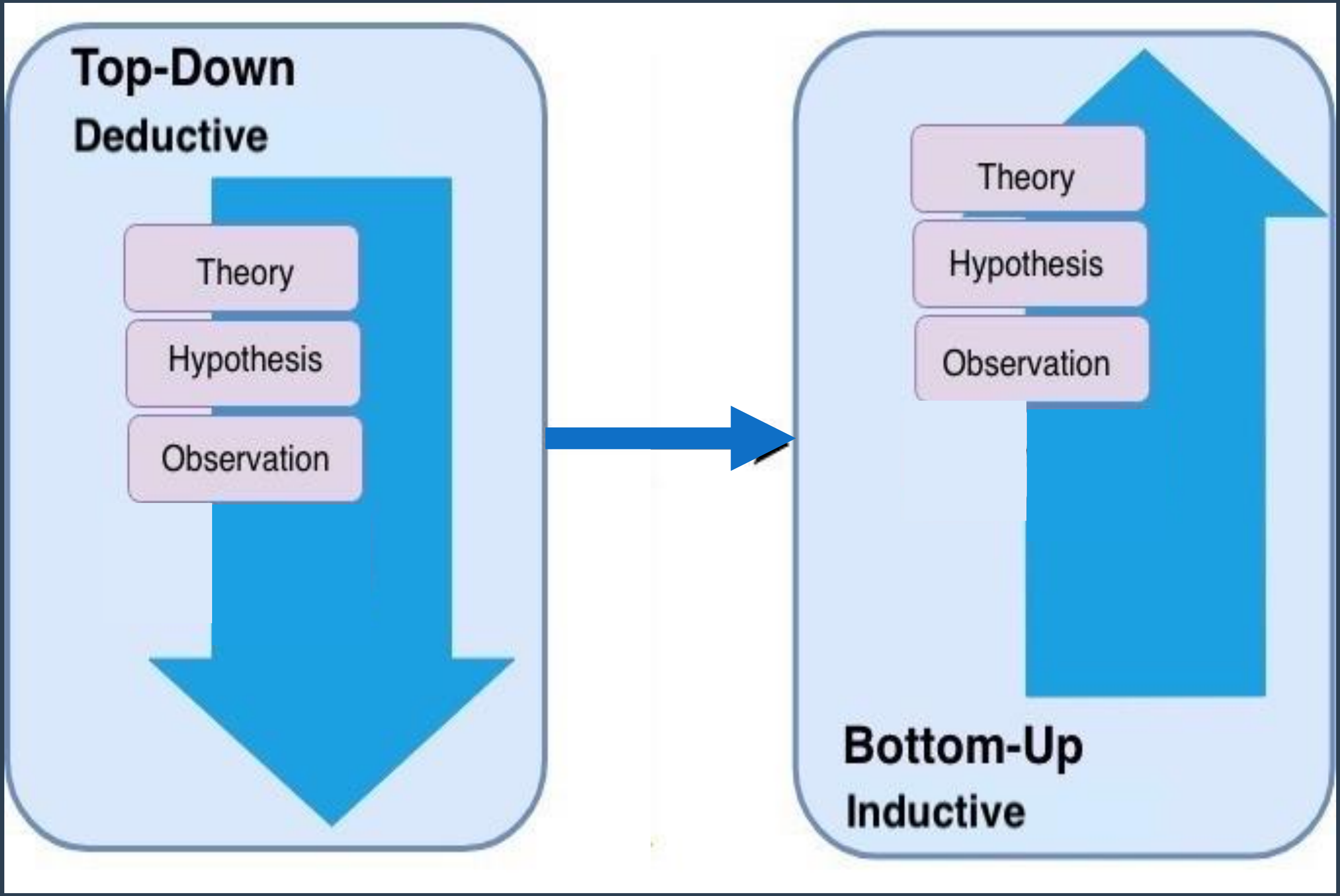
Why Big Data Analytics for LMI?

- 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



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

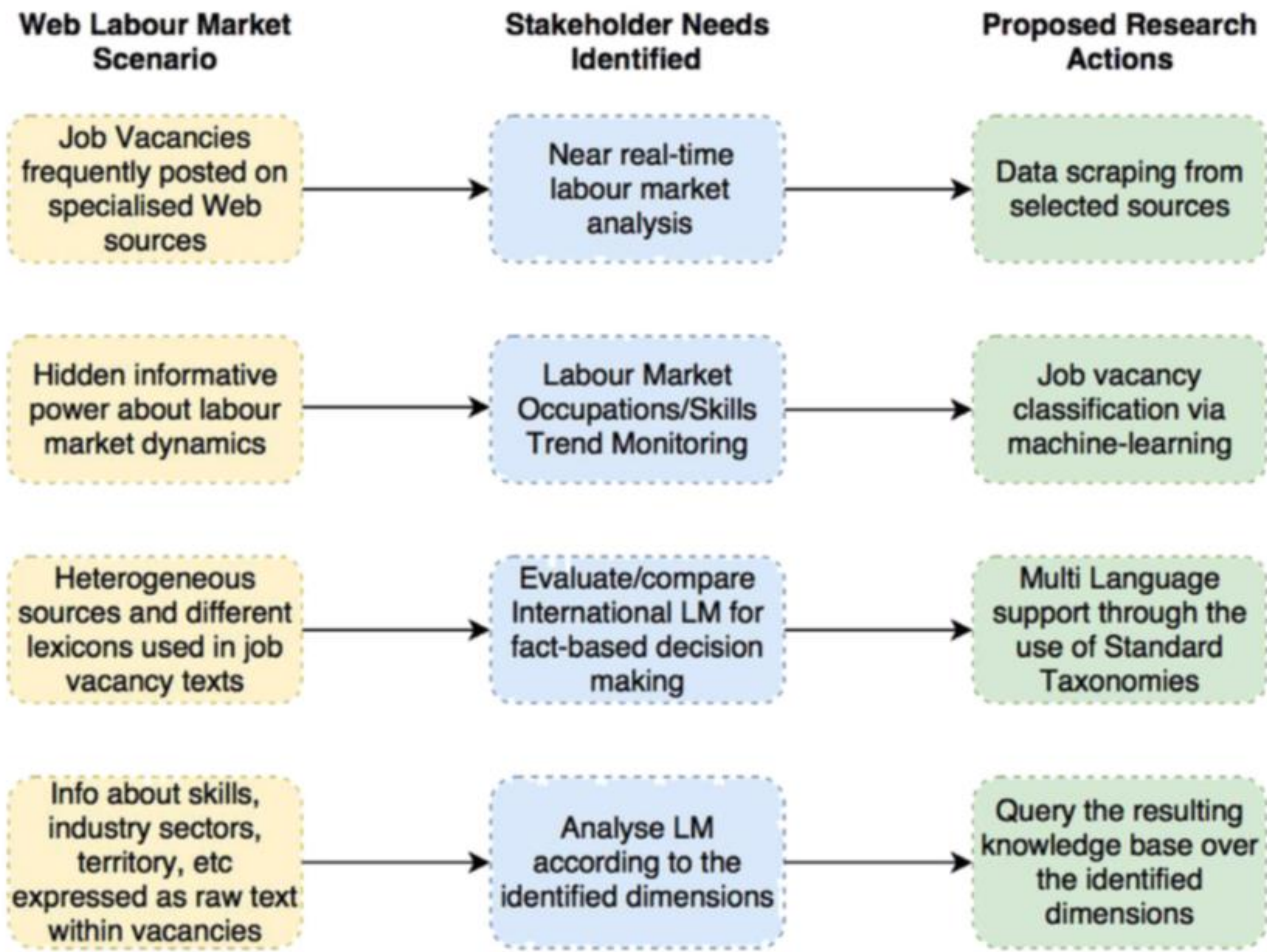

Web Labour Market Scenario
Stakeholder Needs Identified
Proposed Research Actions
Job Vacancies frequently posted on specialised Web sources
Near real-time labour market analysis
Data scraping from selected sources
Hidden informative power about labour market dynamics
Labour Market Occupations/Skills 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