EDM Webinar

Data Lake or Data Swamp?

Pandemic-accelerated issue

Mirek Sopek

The founder and CTO of MakoLab, CEO of LEI.INFO

Robert Sendacki

The founder and CEO of MakoLab Consulting

Tomasz Stepien

DCAM and BCM Manager, MakoLab Consulting

Jans Aasman

Psychologist and Cognitive Science Expert. CEO of Franz Inc.

Richard Wallis

The founder of Data Liberate. Contributor to the development of Schema.org

Andrzej Grochowalski

CIO of InPost

MakoLab | MakoLab CONSULTING

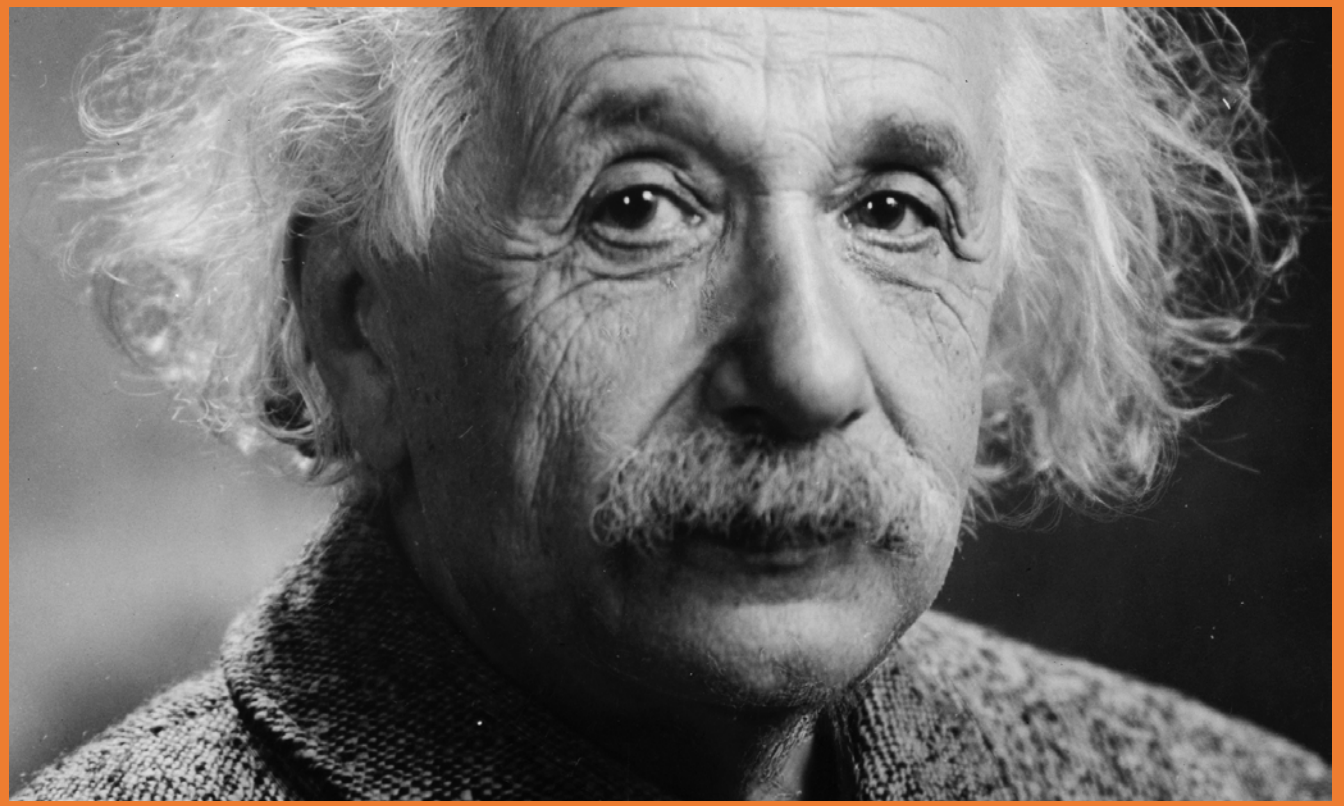
EDM Council

Moderated by **Mark Zill**
Senior Advisor, EDM Council

- Mark joined the Council in 2017
- Mark works with Council members, staff and partners to apply innovative technologies in support of the Council's focus areas
- Prior to joining the Council, Mark was the Chief Operating Officer of the GoldenSource Corporation for 10 years
- He has also held numerous senior management positions in technology, operations and general management spanning both private and public companies

EDMCouncil

How to bring sanity to data projects?

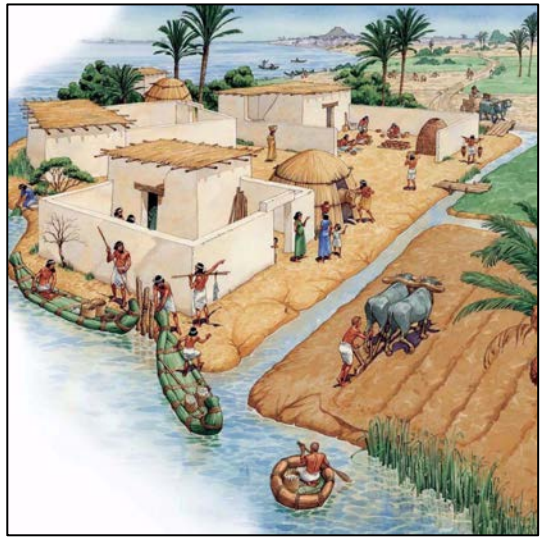


“Doing the same thing over and over again and expecting a different result – is the definition of insanity.”

Albert Einstein

Data in Business is like Water on Earth: there would be no life without it …

Three major events in the history of mankind:



First cities

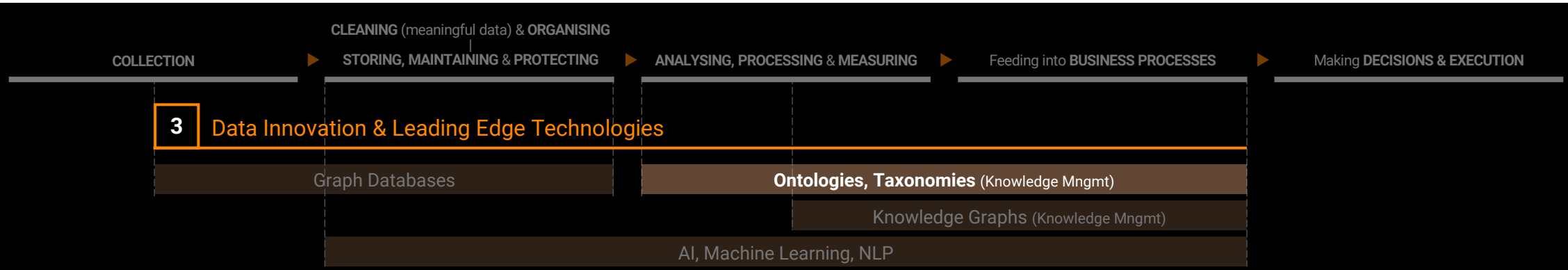
Mediaeval cities

Modern civilization

Tech Thread 1:
Attempt organizing your Knowledge Accumen through Taxonomies & Ontologies

- The ontology definition (slightly indirect) we like most:
  „Knowledge engineering is the application of logic and ontology to the task of building computable models of some domain for some purpose.” (Sowa, 1999)”.
- The taxonomy can be thought of as a simpler ontology, limited to categorization (classification) of concepts into hierarchical structures.
- Taxonomies and Ontologies are of highest importance for any organization – from the perspective „organizing” knowledge about any business domain.
- Some notable industrial examples of ontologies: **FIBO** (Financial Industry Business Ontology), The **ICARUS** Ontology (representation of the knowledge for the aviation sector), **Industry Ontology Foundry (IOF)** - a family of interoperable ontologies (in the making).

… Where we are in the „DATA Process”:



Data in Business is like Water on Earth: there would be no life without it ...

Just like our civilization depends on water - we, as individuals, organizations and societies depend on data.

At first, we don't have much data

Then we learned to collect more data

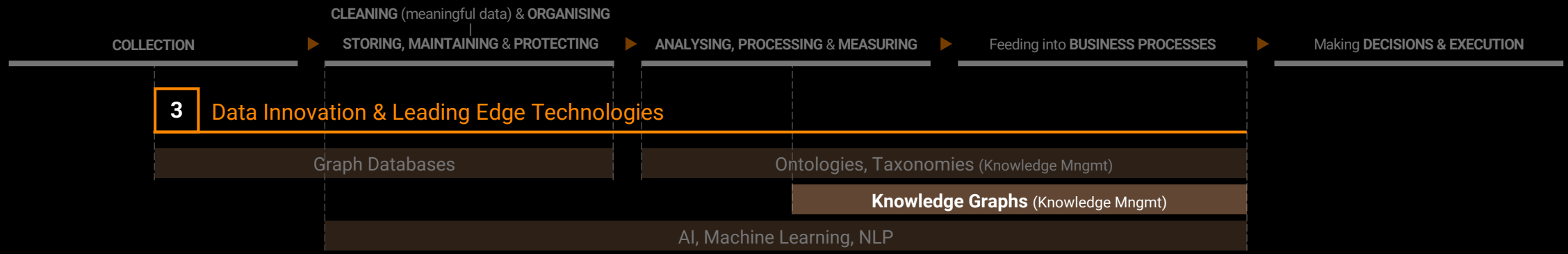
And now the amount is overwhelming

DATA for Business in Information Era = Water for all the life on Earth

Tech Thread 2:
Build Knowledge Graphs – You will need them…

- Two Knowledge Graph definitions we like most:
  1. **„The knowledge graph (KG) represents a collection of interlinked descriptions of entities: real-world objects and events, or abstract concepts ” (by Ontotext)**
  2. **"A Knowledge Graph is a model of a knowledge domain created by subject-matter experts with the help of intelligent machine learning algorithms" (by Poolparty)**
- KGs superseded the older "semantic systems": they are focused on creation of a common knowledge interface for ALL of your data and structural representation for potentially all of data maintained by your organization.
- Knowledge Graphs comply with the data FAIR Guiding Principles (Findability, Accessibility, Interoperability & Reusability).
- Knowledge Graphs help you to build **Data Fabric** on top of your **Data Lakes** and **Data Warehouses.**

… Where we are in the „DATA Process”:



However, most of the rain falling on us goes down the drain

However, most of the rain falling on us goes down the drain

What data do we collect?

Why do we collect certain data?

What about data we do not collect, like:

Webpages' logs

Production sensors' data

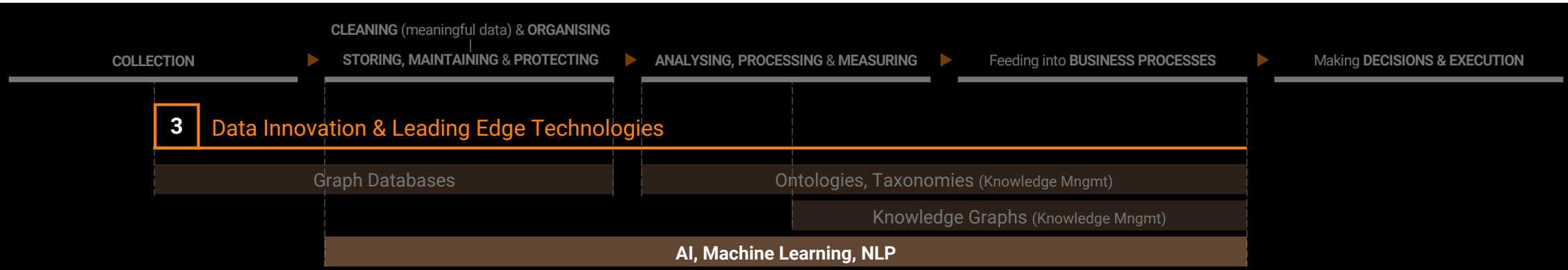
IoT sensors

Customers' behaviours in physical shops

Tech Thread 3:
AI & NLP

- There is no doubt that AI technologies like Machine Learning will form an important part of any modern data handling strategy.
- However, we must listen carefully to warnings coming from practitioners of AI. Gary Marcus and Ernest Davis (in **„Rebooting AI: Building Artificial Intelligence We Can Trust**") say: "*What we have for now are basically digital idiot savants*", "*What's missing from AI today—and likely to stay missing, until and unless the field takes a fresh approach—is* ***broad (or "general") intelligence***. (…)
- We need to adopt newer strategies: XAI (**Explainable AI**), **Semantic AI** and adhere to **HITL principles** (Human-In-The-Loop) principles.
- Makolab's KnowML is one of examples of approaching the problem. We have recently applied KnowML tools for Covid related data.

… Where we are in the „DATA Process":



Survey summary

As we have not been able to collect enough data (statistically relevant) from the surveys, we shall proceed to analyze what we observe to be the most common business reality.

Why are our 'buckets' leaking?

- **Trouble to show ROI of data management:**
  - WWW logs
  - Production sensors
  - IoT sensors
  - Physical shops

- **Very often we live in „fake data bubbles”, due to:**
  - No „data work” culture
  - Input data quality
  - Data processing mistakes/errors

Real case examples:

Over the years telecoms have neither collected nor analysed data from BTS (base transceiver station) about customers as costs were significant and business could not figure out what could be the purpose.

The manufacturing company’s Overall Equipment Efficency (OEE), measured’ between 75 and 80%, so almost World Class. Really?

Once you look into the process you could see the data was based on paper forms filled out by line operators. When given the choice to tune the line, or fill out the report for the previous activity - they alway selected tuning instead of paper work … So in reality the OEE was ~48% !!!

The most popular steps companies are taking to improve 'handling water'

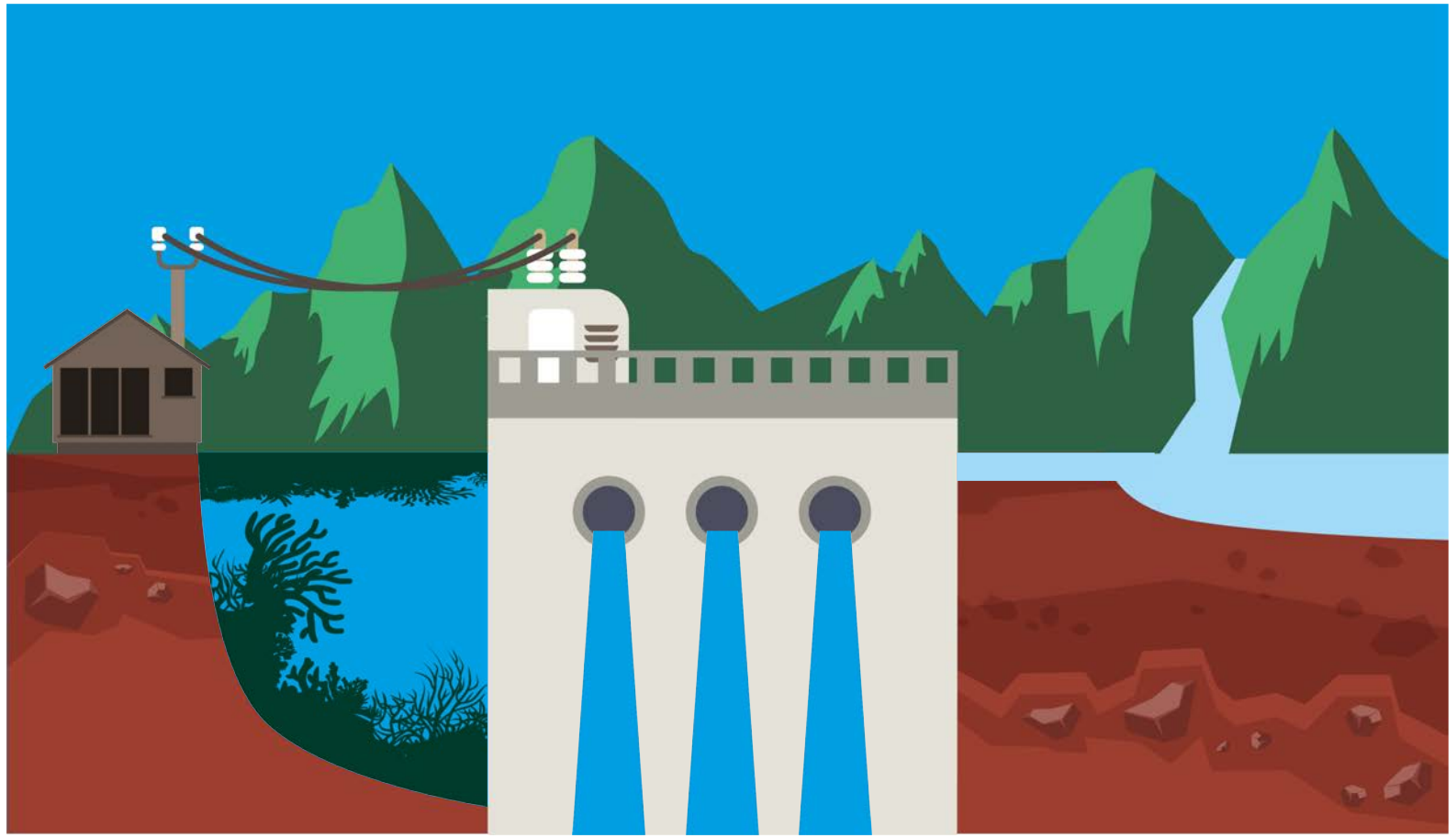
'Traditional' On-Premises Data Warehouse

Data requires transformation.
Expensive to store large volumes.
Schema-on-write.
Tightly coupled storage & compute.

Data Lake

Data stored in native format.
Can store unlimited data forever.
Schema-on-read.
Decoupled storage & compute.

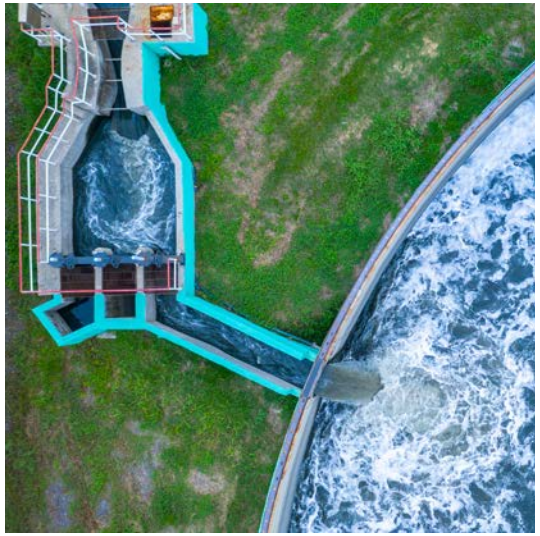
Target: to build an effective metaphorical water-electric plant
Sad reality: a swamp not capable of generating electricity



Commonly observed 'Remedies' for 'dead-ends' in DWH / Data Lakes

Invest in cloud / software extension

Taking shortcuts



1:1 digitisation

Common pitfalls

- Lack of data management basics
- Lack of structured
- No audit trails how data are collected

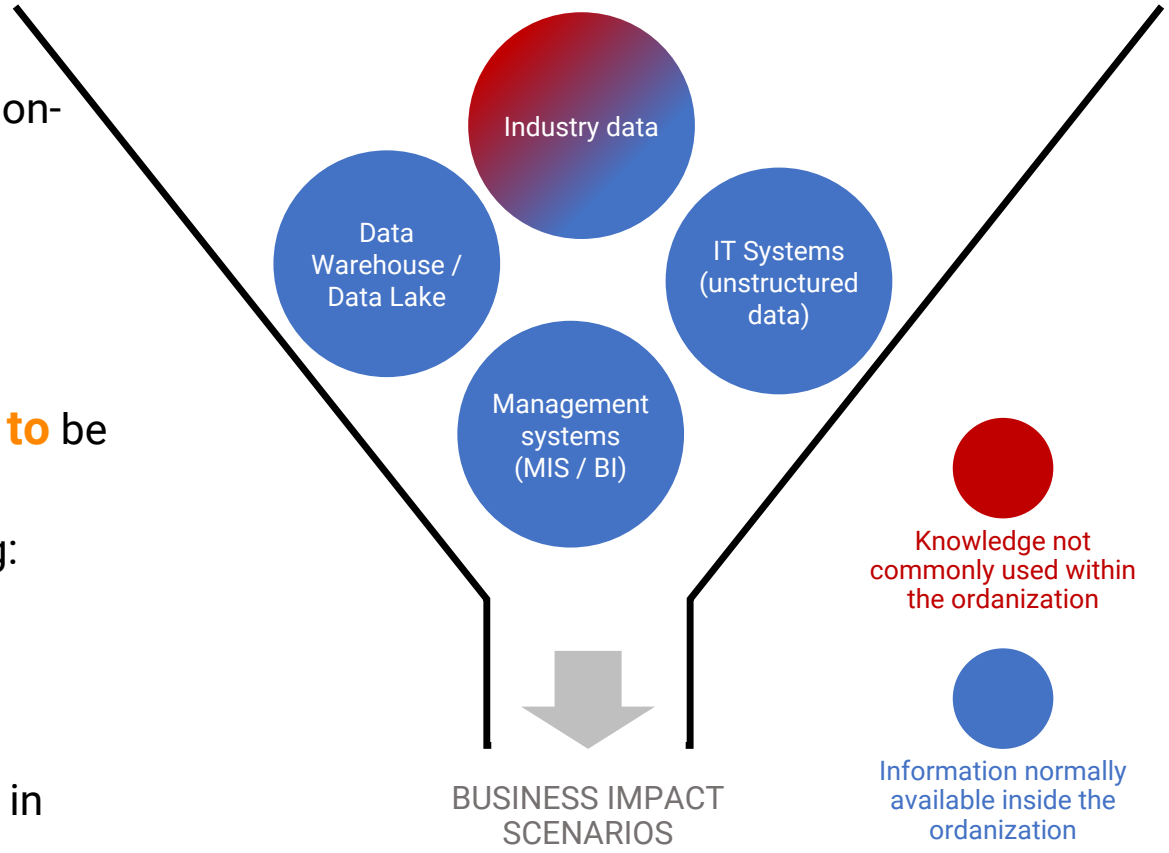
How has the CoViD pandemic exposed some key issues?

CURRENT SITUATION UNDER CoViD

- **Managers are put in a situation of extreme unpredictability**
- **Existing time-series have been proven** useless for forecating the future. Most of the methods that help decision-making processes can lead to completely incorrect decisions.
- **Sudden need for different, richer and better data;** risk for even more shortcuts and „temporary patchworks".

CHALLENGE

- In an unpredictable environment, **business scenarios need to** be planned that **take into account different factors.**
- **Flexibility is required** to make strategic decisions, including:
  - undefined timetable for removing restrictions,
  - changing scope of government support,
  - behaviour of business entities in new circumstances,
  - signal information from the market (statistical indicators in micro - mezzo - macro and cross sections).



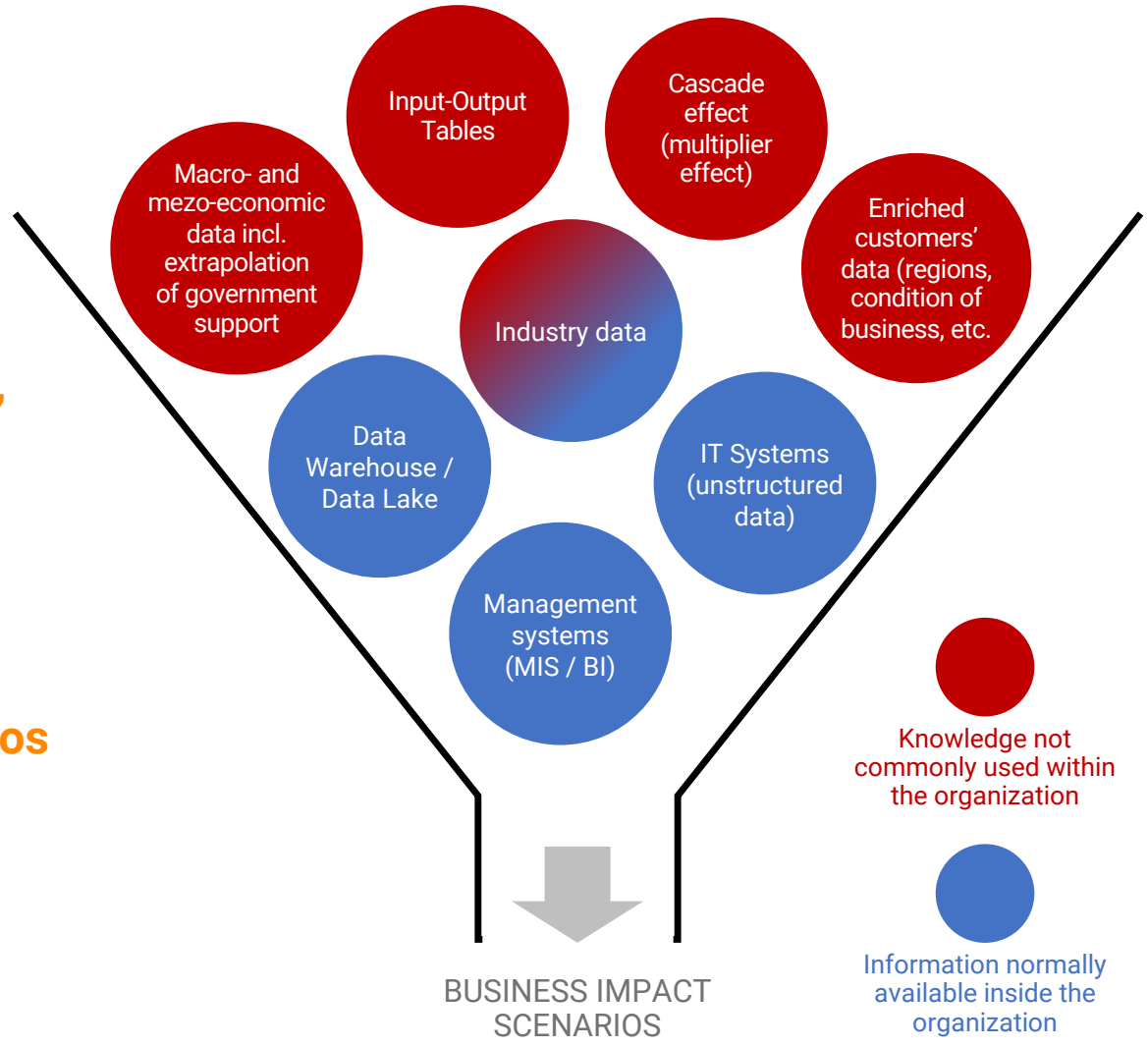
Agile analytics to address the business need

BUSINESS PERSPECTIVE

- **Input data**
  - **macro- and mezo-economic forecasts** on the specifics of markets and consumers,
  - **IT / DWH / Data Lake / BI structures.**
- Collect available data **(e.g. percentage of unemployment, sectors, etc.),**
- Enrich **(e.g. customers' regions, general condition of business from external databases, etc.),**
- Supplement available data **(e.g. economy input-output tables, cascade effect, etc.),**
- **Construct and validate matrixes** based on which **scenarios** are generated **to estimate the impact on a company.**

RESULT

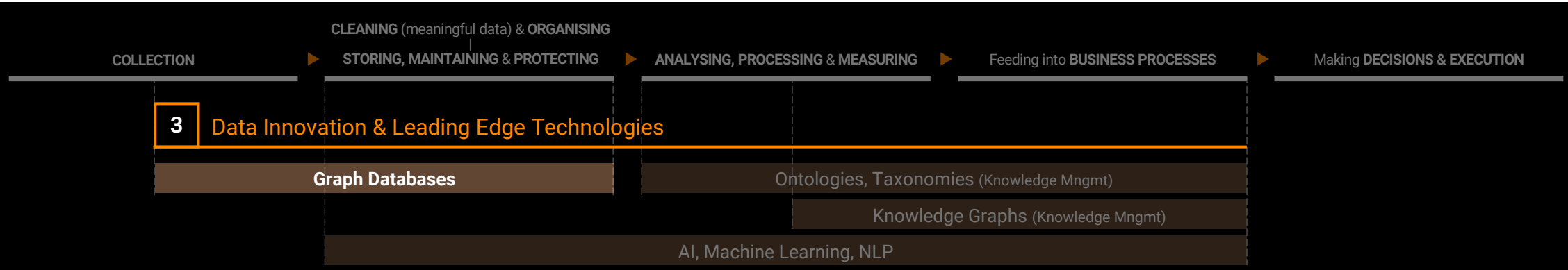
- **Flexible plan to respond to changing circumstances**



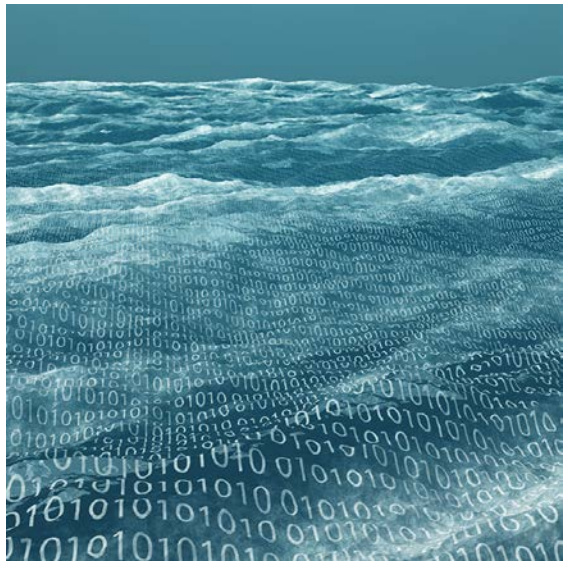
Tech Thread 4:
Organize your data & store it in the way optimal for searching & extracting knowledge.

- Graph Databases exploit graph structures (nodes, edges and their properties) to represent and store data.
- Semantic databases ("triplestores"), based on RDF data model belong to the family of graph databases.
- However, latest **(L)PGs** – (Labelled) Property Graphs offer **much more efficient alternatives to RDF graphs**.
- Property Graphs systems have also proven to offer better analytical tools for heterogenous data.
- The classical Semantic Web community invented the new RDF* approach to compete with PGs --> RDFs are still important.
- The newly proposed ISO project - GQL (Graph Query Language) - slated for becoming an international standard in 2022 - will change the fundamental paradigm of data use – first time since SQL was standardized in 1986.

… Where we are in the „DATA Process":



Data Fabric to the rescue



Data fabric allows you to:

- Links sometimes disparate data points
- Allows for a temporary structure to be created, deeply analyzed, used and dissolved when no longer needed
- In the structured state – increased speed of analytical operations

Data Fabric – structured approach to combine advantages of Data Lakes and Data Warehouses

DATA WAREHOUSE
A central repository of data from disparate sources.

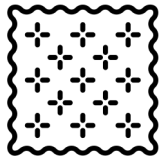
→ Structured
→ Integrated
→ Searchable
→ Order imposed by ETL procedures

→ **Ready for Analytics and BI**

DATA LAKE
A repository of data from disparate sources in original/raw formats.

→ Usually unstructured
→ Highly scalable (bigdata)
→ Heterogeneous
→ Explorable

→ **Ready for Data Mining and Science**



DATA FABRIC
Distributed system with a top layer acting like a living tissue stretched over disparate sources of data from data warehouses and data lakes.

→ Empowered by data semantics
→ Highly connected through Linked Data principles
→ Integrated on the metadata and access levels
→ Searchable and explorable on the knowledge level

→ **Ready for Analytics, BI, Data Science and AI**

METHODS:

Existing Data ▸ **Semantic Knowledge Models** *Taxonomies, Ontologies* ▸ **Superordinate Database with meaningful and dynamic linking of business objects** *Knowledge Graphs* enhance **Data Analytics**, **Business Intelligence**, **Visualization**, **Artificial Intelligence**

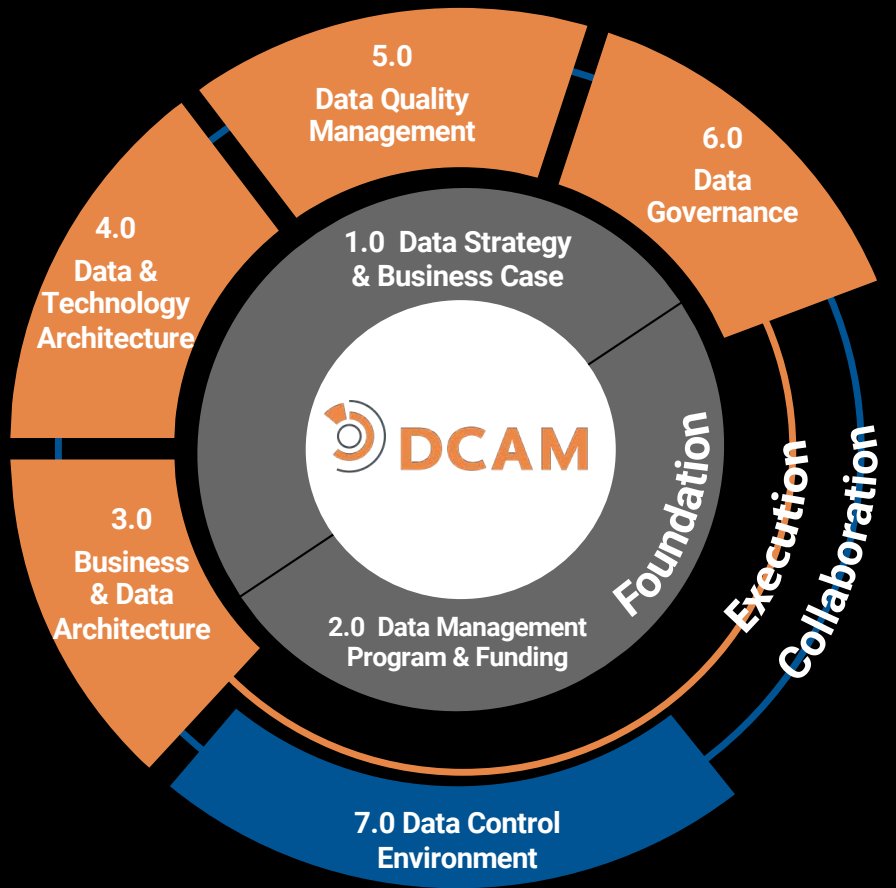
Every journey starts the same way, with THE FIRST STEP...

Data Capability Assessment Model (DCAM)
foundation tenets:

1. **MANAGING CONTENT** (Identify, Define, Locate).
2. **ENSURE DATA QUALITY** (Data must be Fit-For-Purpose).
3. **BUILD A SUSTAINABLE** PROGRAM (Skills, Governance, Culture).
4. **ENABLE CROSS ORGANIZATIONAL COLLABORATION.**

MOST ORGANISATIONS DO NOT TRULY AND DEEPLY UNDERSTAND THEIR DATA BLOODSTREAM.

MISSING DATA, INCONSISTENCIES, ERRORS, ETC. ARE ALMOST STANDARD.



We are EDMC's authorised DCAM partner

WE ARE A MEMBER OF EDMCouncil AND ENGAGE WITH DATA COALITION

DCAM

The **Data Management challenges facing businesses** across the globe are both varied and multi-dimensional as they are complex and difficult in terms of resolution:

DISTRIBUTED ARCHITECTURE OF ENTERPRISES

Not only on the side of IT, but rather (and more importantly) on the business side – caused to a large extent by the global COVID 19 pandemic. **Workers and employees have to deal with much less direct interactions while relying on more and more remote data/information access**, where such data and information have to be of high quality and instant accessibility (no delays) to a much higher degree than earlier.

INCREASED REGULATION

Companies operating in specific industries (such as – for example - finance industry) are subject to increased and intensifying regulation (e.g. from ECB, DG FISMA, KNF in Poland, and others). This increasing (width and depth) regulation results in the necessity (and formal requirements) to design and implement strictly controlled, predictable processes which ensures full trustworthiness of information and data.
On top of this **- business organisations distributed and operating across different geographies and local legislations/regulatory regimes face truly complex issues while striving to achieve data consistency, integrity and trustworthiness.**

DATA SENSITIVITY

Majority of enterprises but specifically financial institutions have access to vast amounts of their customers' personal and sensitive data. Such information resources require stringent and specific means of protection – especially given the (mentioned above) incerased degree of regularion and reporting requirements, as well as distributed and remote work model, **the latter nececcitated and resulting from the current pandemic/post-pandemic work environment.**

Hence, it is necessary to adopt and use Information & Data Management methodology:

DATA CAPABILITY ASSESSMENT MODEL (DCAM)

DCAM is recognised as the **industry standard** (across industries) and allows enterprises to achieve full control of the quality of their data.

COMMENTS FROM OUR PANELISTS:

Jans Aasman

Dr. Jans Aasman is a Ph.D. psychologist and expert in Cognitive Science, as well as CEO of Franz Inc., an early innovator in Artificial Intelligence and provider of Knowledge Graph Solutions based on AllegroGraph.

As both a scientist and CEO, Dr. Aasman continues to break ground in the areas of Artificial Intelligence and Knowledge Graphs, as he works hand-in-hand with numerous Fortune 500 organizations, as well as government organizations worldwide.

Richard Wallis

An independent Linked Data, Knowledge Graph, and Schema.org consultant.

Trading as Data Liberate, Richard has been an active contributor to the adoption and distribution of each of these technologies from their introduction.

Working with Google in establishing and expanding the Schema.org RDF based vocabulary on the web whilst assisting clients in its pragmatic use and adoption.

Andrzej Grochowalski

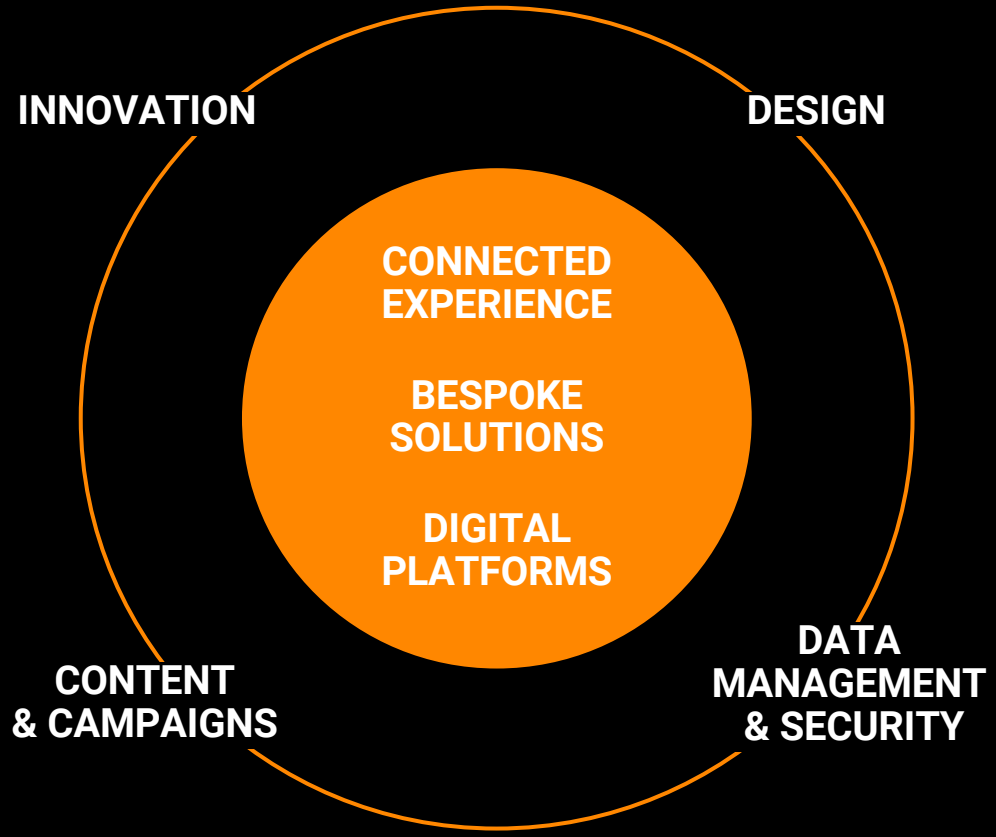
CIO of InPost,
the most successful operator of automated parcel lockers in Europe.

Andrzej works with IT and modern technologies for over 18 years. In 2017–2019, he was the CIO and vice president of the management board at the Tauron Group (large energy supplier), and previously for 10 years he worked at BNP Paribas as IT director, where he also coordinated the work of the innovation center.

He has successfully completed several major IT transformations and implementation projects, including three bank mergers.

MakoLab (established 1989)

A GLOBALLY OPERATING **DIGITAL PROJECT HOUSE** AND A TEAM OF TECHNOLOGY EXPERTS THAT FUSE IT ENGINEERING AND CREATIVITY **TO BUILD USER-INSPIRED SOLUTIONS**.



MakoLab MakoLab CONSULTING

MakoLab Consulting

- **BOUTIQUE CONSULTING** EXTENDED BY **DIGITAL PROJECT HOUSE** FROM **MAKOLAB S.A.**
  CONSULTING PROJECTS FOLLOWED BY DIGITAL TRANSFORMATION DELIVERY.
- **AGILITY – WE EMBRACE CHANGE**
  AND WE DO NOT WASTE YOUR TIME. WE FOCUS ON YOUR REAL PROBLEMS AND WE TEACH YOUR PEOPLE A „CAN DO!" ATTITUDE.
- **ENABLING ENGAGEMENT & LASTING CHANGE**
  WE WORK CLOSELY WITH PEOPLE THROUGH STRUCTURED WORKSHOPS, BREAKING INTERNAL BARRIERS, ORGANISING PEOPLE AROUND FACTS (NOT MYTHS) AND DEPLOYING ORGANISED FREEDOM OF ACTION.
- **STRONG R&D COMPONENT**
  RESEARCH & SCIENCE APPLIED.
- **TRUE COOPERATIVE CULTURE**
  HISTORY OF SUCCESFUL DELIVERY & PROVEN RELIABILITY.
  WILLING TO SHARE RISKS & REWARDS.

Questions?

MakoLab | MakoLab CONSULTING

EDM Webinar










FOR MORE INFORMATION PLEASE CONTACT:

Robert Sendacki

The founder and CEO
of MakoLab Consulting

robert.sendacki@makolab.com

www.makolab.consulting

MakoLab | MakoLab CONSULTING

EDMCouncil