



Overview of the Mobility program

DESTINI
*Smart Data Processing and Systems of Deep
Insight*



DESTINI



HORIZON 2020

Table of Contents

03 About Mobility **04** About DESTINI **27** Conclusion

Satellite Event

- 05** Day 1
- 06** Day 2
- 07** Day 3
- 08** Day 4 (1)
- 09** Day 4 (2)
- 10** Day 5

Execution of the mobility program

- 11-12** Dario Benvenuti
- 13-14** Silvestro Veneruso
- 15-16** Flavia Monti
- 17-18** Jerin George Mathew
- 19-20** Alberto Morvillo
- 21-22** Francesca De Luzi
- 23-24** Mirella Sangiovanni
- 25-26** Stefan Driessen



About Mobility

The mobility program aimed at bringing together Early Stage Researchers (ESRs) to collaborate and/or be recruited by CUT for a specific period of time, to conduct research with CUT's team in areas revolving around smart data processing and systems of deep insight.

Due to the traveling restrictions posed by the COVID-19 virus outbreak, the mobility program was delayed and moved towards the end of the project.

The corresponding activities were executed in two stages:

1. Organization of Satellite Event in the context of SummerSoc

During this event, lectures, presentations, trainings, and brainstorming sessions were conducted. ESRs presented ongoing research work that align with the areas of interest to DESTINI and new ideas were proposed and discussed for future research. The best ideas were selected for further elaboration and the ESRs were invited to continue taking part to the mobility program through site visits and short relocation to Cyprus.

2. Execution of the mobility program in Cyprus

ESRs that were invited to Cyprus to investigate further their ideas, comprised PhD candidates from Sapienza and JADS. During their period of stay, they worked closely with researchers at CUT, both within the Cypriots partner's group and beyond that, to explore further subjects identified and to suggest new research topics of mutual interest. In this context, new methods, techniques, algorithms, models and approaches on areas such as Digital Twins, Data Lakes, Data Meshes, Blockchain, Predictive Maintenance and Analytics, AI-supported smart data processing were proposed, discussed, elaborated and analyzed. The ultimate goal of all these activities will be to produce new research papers for publication and establishing further research collaboration. Finally, this exercise led to investigating the potential of writing and submitting new proposals for attracting external funding at National or European level.



About DESTINI

TWINNING PROJECT - HORIZON2020

Smart Data ProcESSing and SysTEms of Deep INsight (DESTINI) is a H2020 Twinning Project that proposes a series of coordination and support actions for promoting research in the area of Smart Data. The Cyprus University of Technology (CUT) joins forces with two internationally recognized scientific groups from the Netherlands (Tilburg University and Jheronimus Academy of Data Science) and Italy (Sapienza Università di Roma), aiming to strengthen its research and scientific profile in the relevant area. Specifically, DESTINI's activities revolve around exchanging scientific knowledge and transferring best research practices amongst its partners in the field of Smart Data Processing and Systems of Deep Insight.



04, July, 2022

Day 1



Day 1 included an introductory presentation from Professor Andreas Andreou, explaining the context of the mobility program, the activities to be conducted and the agenda of the event. In addition, brainstorming sessions were held throughout the day about research-oriented topics.



05, July, 2022

Day 2



Day 2 began with presentations from the Dr Andreas Christoforou and Panayiotis Christodoulou, so as to give ideas in sub-areas of DESTINI and establish discussions and ideas for brainstorming. In addition, Flavia Monti, ESR from Sapienza University presented her research work on Industry 4.0.



06, July, 2022

Day 3



Day 3 included several presentations from ESRs. Stelios Mappouras, a researcher of DESTINI initiated the day by presenting his research work on Blockchain and NFTs. Then, ESRs from JADS presented their research work. Stefan Driessen presented his work on Data Product marketplace, Stefano Dalla Palma on MLOps, testing, Mirella Sangiovanni on Sustainable cities and Smart Healthcare Intelligence and lastly Nemanja Borovits on Privacy Preservation. The day concluded with brainstorming on these ideas, and discussing about coupling them with DESTINI's research activities.



DESTINI

07, July, 2022

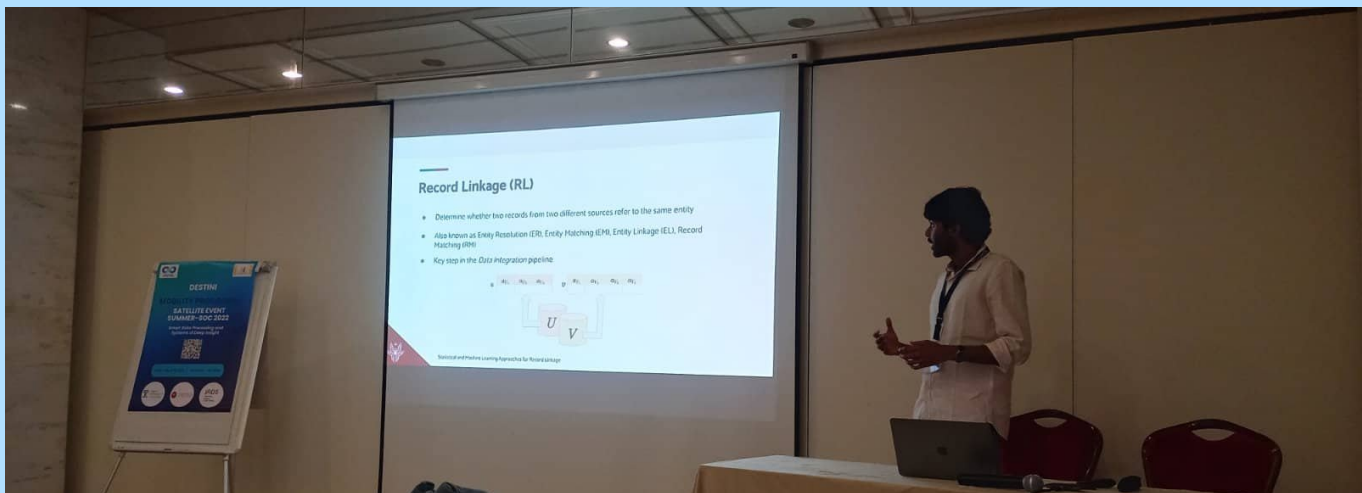
Day 4



During Day 4, Presentations from ESRs continued with 2 ESRs of CUT, Michalis Pingos and Spyros Loizou presented their PhD work on Data Lakes and semantic enrichment mechanism and Digital Twins for visual querying and process mining. In addition, 3 ESRs from Sapienza University followed, Alberto Morvillo presented Geo-based visual exploration of digital document collections, Silvestro Veneruso presented Applying Process Mining to Human Daily Activities and Jerin George Mathew presented Statistical and Machine Learning Approaches to Record Linkage. Following the presentations, the day concluded with brainstorming sessions.

07, July, 2022

Day 4

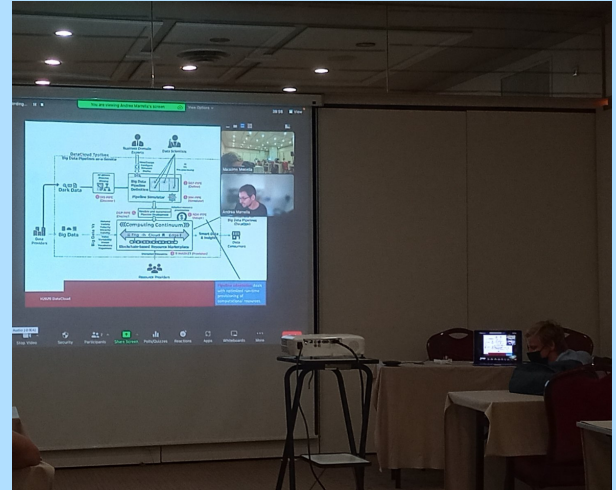
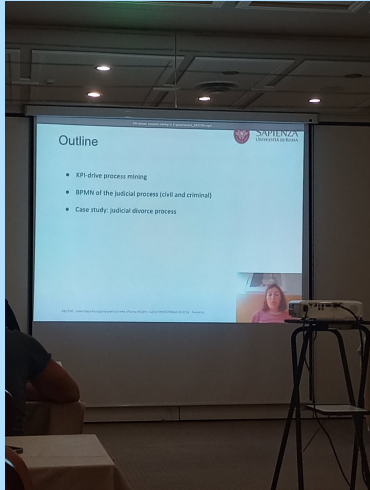




DESTINI

08, July, 2022

Day 5

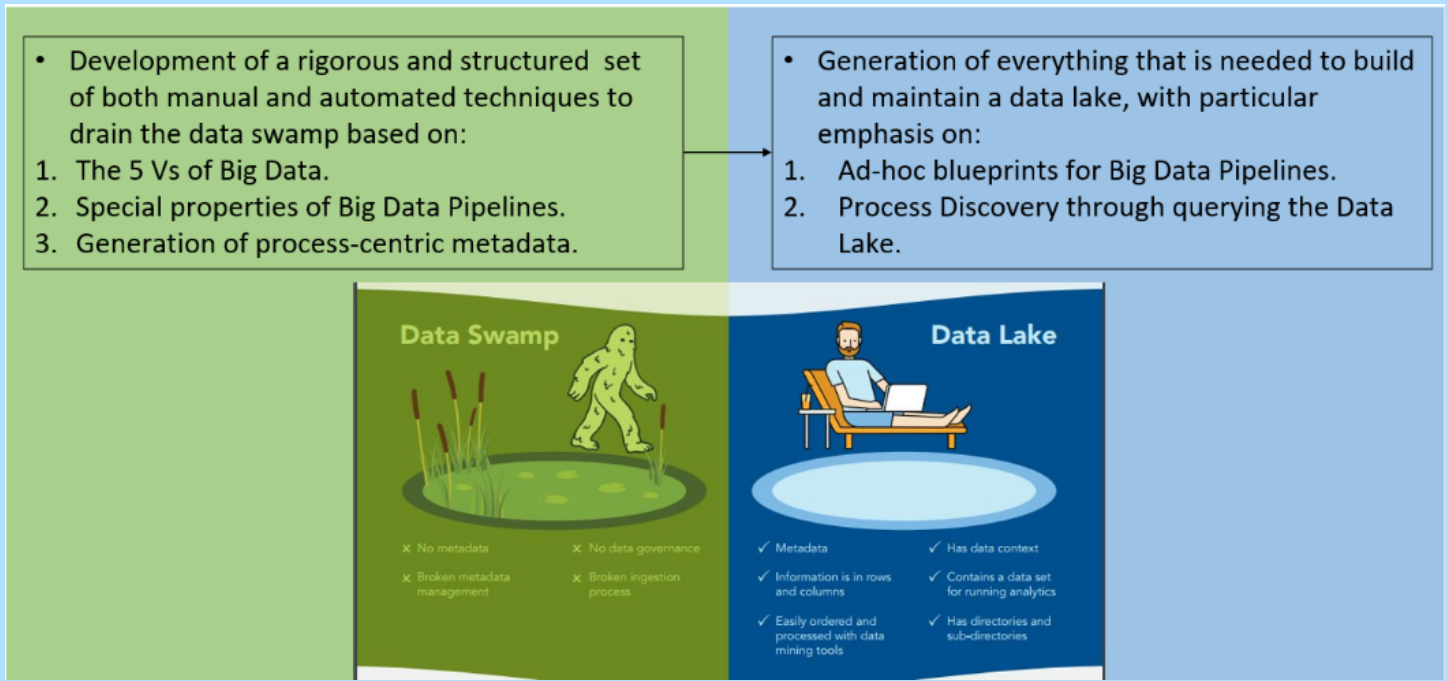


Day 5, the final day of the event, included presentations from the Professor Andrea Marella on Big Data Pipelines Discovery from Dark Data: Challenges and Some Solutions. Also, the remaining ESRs from Sapienza presented their research work: Dario Benvenuti on Event Log Extraction and Generation in the context of Big Data Pipeline and Francesca De Luzi on KPI-driven process mining in E-government. After the conclusion of the presentations, The Professors Andreas Andreou, Massimo Mecella and Willem Jan Van Den Heuvel discussed, and selected the best ideas that align with the project's JRAs.



DESTINI

Dario Benvenuti



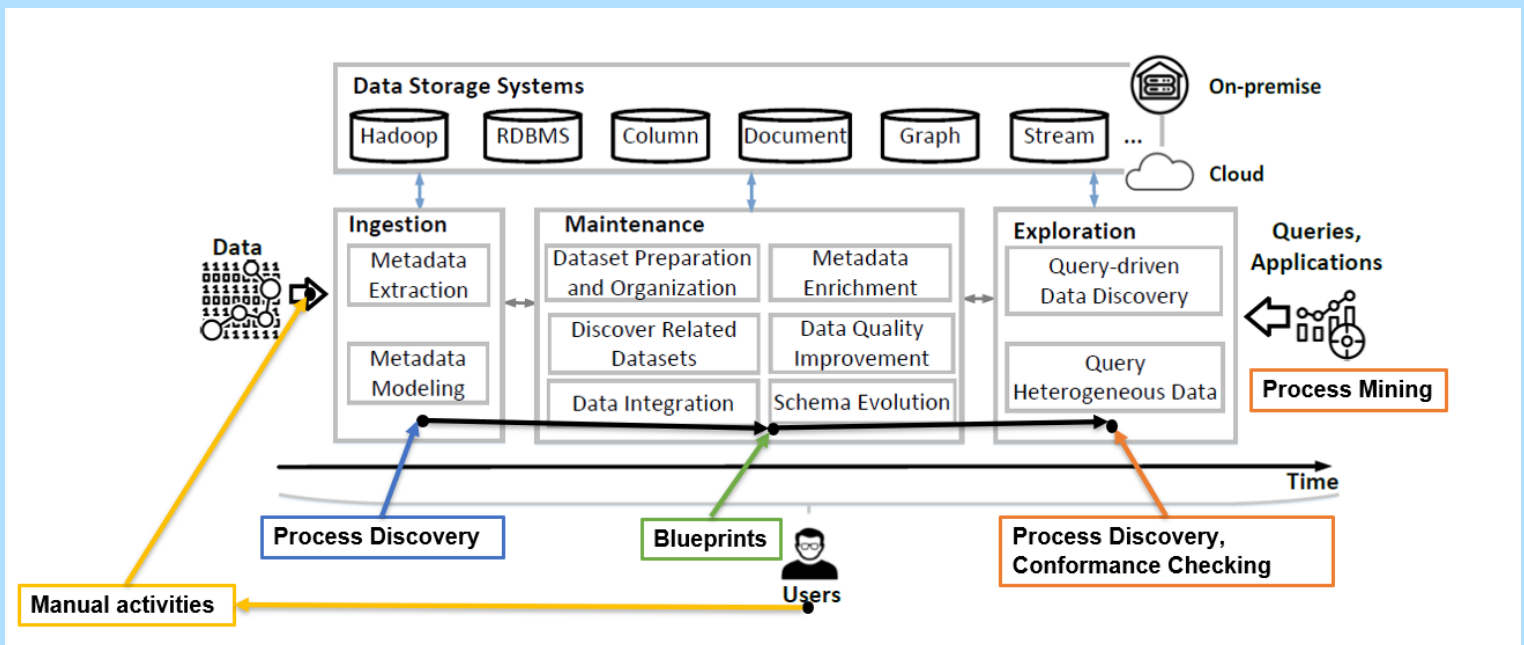
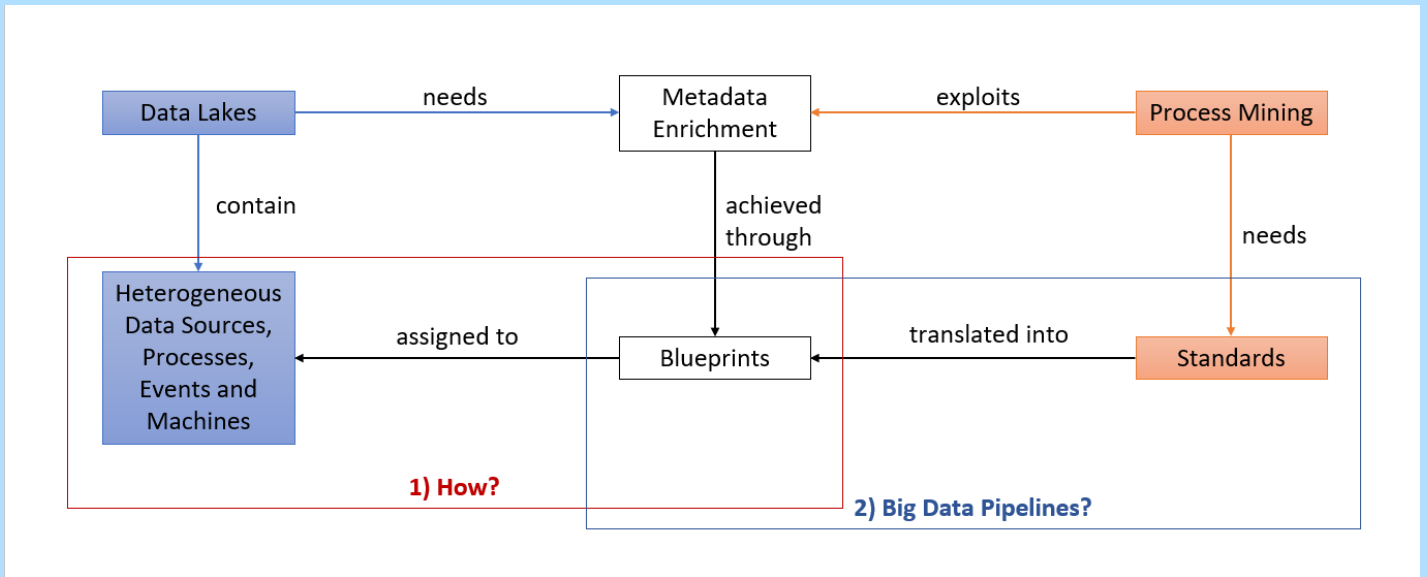
In practice, this research direction can be split in six phases:

1. Collection of the most popular manual and automated (or semi-automated) techniques to drain DSs (e.g., data cleaning, data integration, metadata management/enrichment).
2. Analysis of the collected techniques to make a clear map of what can be automated and under which conditions.
3. Development of the required ad-hoc blueprints to work in the Big Data Pipeline context and to assign process-centric parameters in a strictly data-centric environment.
4. Development of the draining and maintenance techniques, and analysis of the role of PM as an automated technique to build blueprints and to exploit the DL itself to extract knowledge from every kind of data (e.g., even from unstructured data with no timestamp).
5. Validation with a real organization.
6. Writing of the paper.



DESTINI

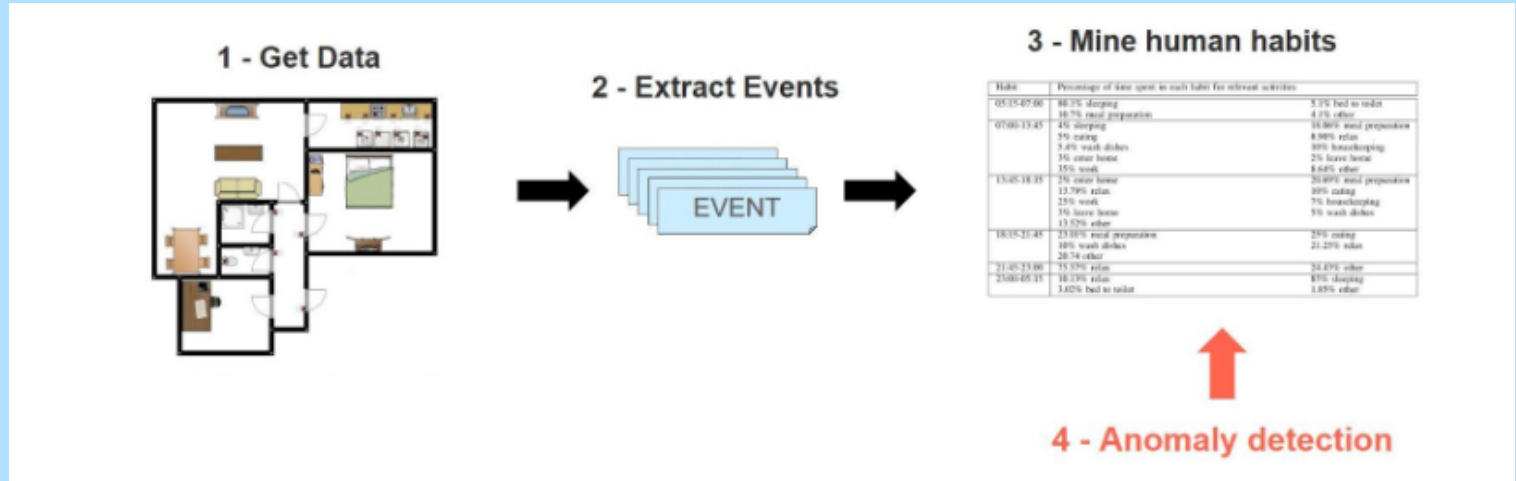
Dario Benvenuti





DESTINI

Silvestro Veneruso



Research Proposal #1

TARGET: Elderly people living alone, people with some mental diseases (e.g., dementia) or who are rehabilitating from some trauma (e.g., a stroke).

SETTING: Smart environments. E.g., elderly homes, nursing homes, rehabilitation's centers.

PROBLEM: How can we ensure that their daily routines go the right way?

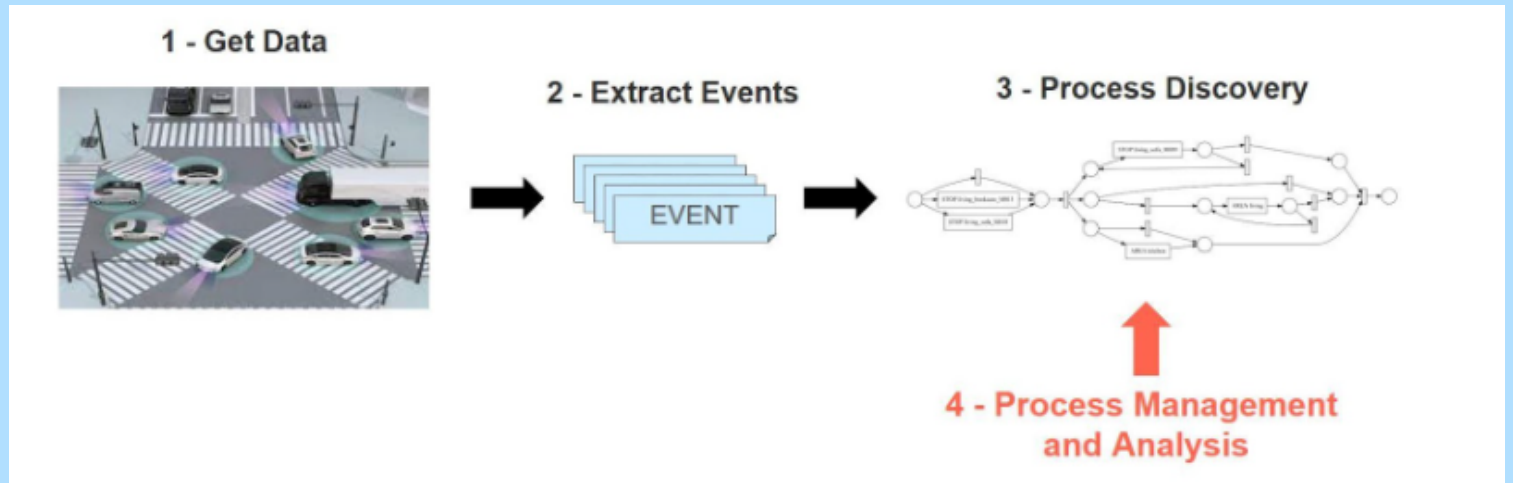
GOAL: Starting from a sensor log, we propose an approach to automatically mine human habits on a daily basis and then perform anomaly detection to capture any possible wrong behavior from the observed inhabitants.



DESTINI

Execution of the mobility program in Cyprus

Silvestro Veneruso



Research Proposal #2

TARGET: Civil Protection's institution.

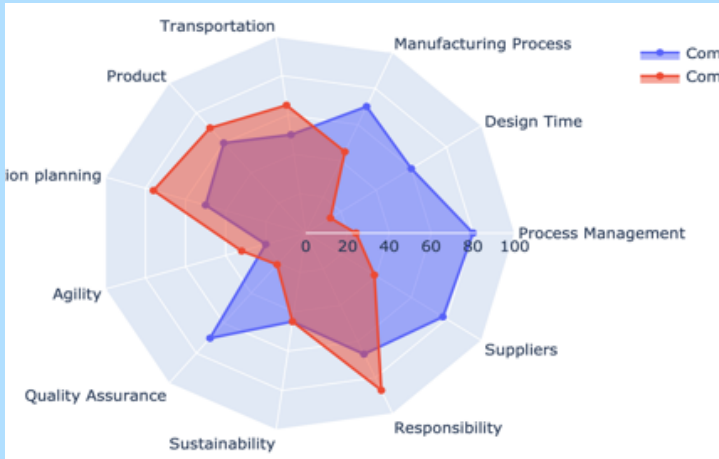
SETTING: Public environments. E.g., roads.

PROBLEM: How is a car crash usually managed? Who intervenes first: public officers or the ambulance? Which is the timing?

GOAL: We want to mine models describing how car crashes are handled in order to perform some analysis over this information.



Flavia Monti



Research Proposal #1

- Construct the Maturity Model FCM
- Execute the model to:
 - Study dependencies between dimensions
 - Find roadmaps towards Industry 4.0 readiness

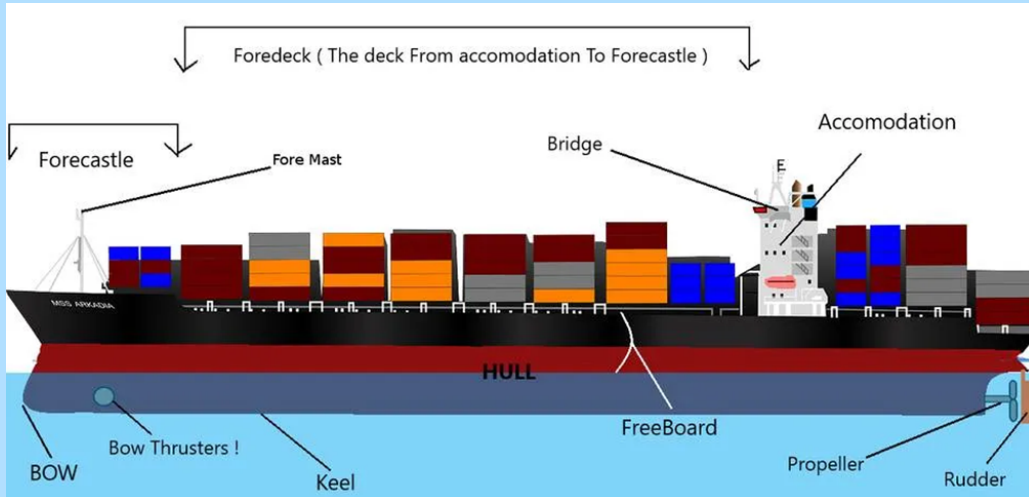
Methodology:

- Define concepts
 - Dimensions and levels of the maturity model
- Validate the concepts through questionnaire to the experts
 - Industry and academia
- Define concept causality through questionnaire to the experts
- Execute the map for static and dynamic analyses



DESTINI

Flavia Monti



Predictive Maintenance (PdM) Strategies in Maritime using Digital Twins

- Maritime domain is undergoing a digital transformation
 - Promotes automation and digitization of processes and control operations
- Maintenance in the maritime domain is crucial
 - Failures in ship components may have significant consequences

Goal: run ships avoiding extraordinary problems

- Detecting anomalies and possible defects and fix them before they run into failure

Short-term approach

- Identification of critical components
 - Availability of data
 - FMEA, FMECA, 4-quadrants selection methods
- Definition of Digital Twin architecture
 - Service-based digital twin
 - Maritime Digital Twin Architecture (MDTA)



Jerin George Mathew

Schema Alignment

Source 1

Studio Album	Artist name	Genre	Release date
Stadium Arcadium	RHCP	Alternative Rock	February 24, 2004



Album name	Artist name	Genre	Release date
Stadium Arcadium	RHCP	Alternative Rock	February 24, 2004

Source 2

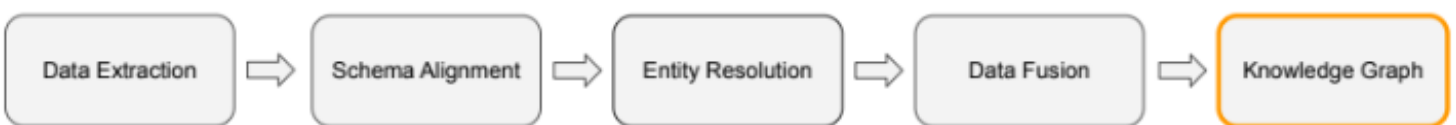
Album name	Artist	Release month	Release year
Stadium Arcadium	Red Hot Chili Peppers	02	2006



Album name	Artist name	Genre	Release date
Stadium Arcadium	Red Hot Chili Peppers	-	02-2006

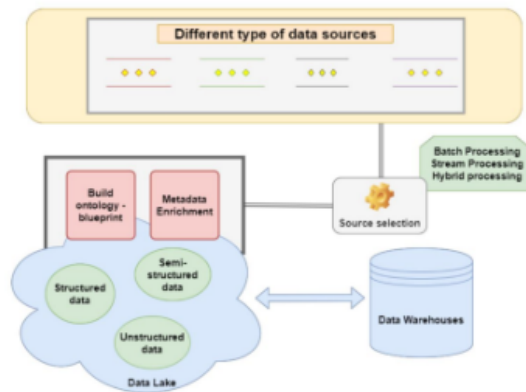
Statistical and ML approaches: a comparison

	Statistical approaches	ML approaches
Statistical guarantees	Yes	No
Scalability (blocking)	somewhat	Yes
Data types	categorical and numerical	textual data
Supervised	no	yes
Core techniques	missing data estimation techniques, MLE	Transformer based nets
Popular techniques	Fellegi-Sunter[2], Blink [3], DBlink [4]	DeepMatcher [5], DeepER [6], Ditto [7]



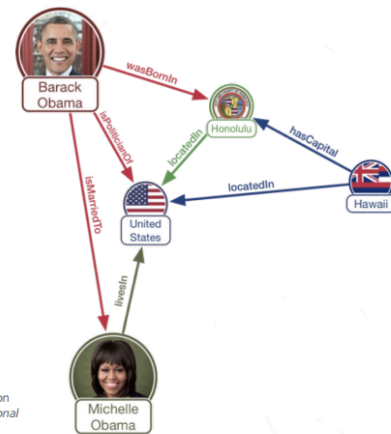
Jerin George Mathew

Proposed framework



Knowledge graph (KG)

- Graph-based model used to represent *knowledge* about real-world objects
- Nodes represent real-world entities (e.g. "Barack Obama")
- Arcs represent relationships between nodes (e.g. "was born in")
- We continuously interact with a KG on a daily basis: Google Knowledge Graph



[10] Rossi, Andrea, et al. "Explaining link prediction systems based on knowledge graph embeddings." *Proceedings of the 2022 International Conference on Management of Data*. 2022.

Paper extension proposal: key ideas

- Initial setting: Data Lake of healthcare-related data sources
- Main idea: extend the framework by extracting more fine-grained details from the data lake
 - Retrieve the set of medical examination of a patient across different hospitals
 - Find patients having a similar medical treatment history
- Such details can be arranged as a KG
- Such KG provides an additional view of the Data Lake that can coexist with the blueprint-based view

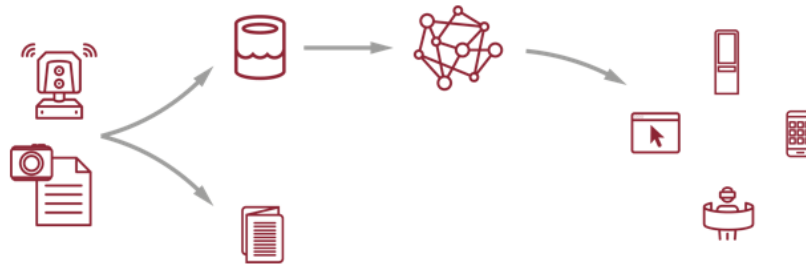
High-level overview of the methodology

1. Finding record pairs in the Data Lake that refers to the same patient:
 - a. Temporal record linkage
2. Data cleaning and data fusion
3. Knowledge graph building
4. Knowledge graph update (when the content of the Data Lake changes)



DESTINI

Alberto Morvillo



Research Proposal #1 - Digitalization of a place of interest

Target: Tourists, visitors and researchers of all ages and gender.

Setting: Any physical place that represents a place of interest (e.g.: monuments, small villages, museums) and does not already allow a digital fruition of his content.

Goal: Facilitate a place fruition in order to

- Increase tourism by providing easy-to-access content, which will also increase the accessibility of the place (e.g., people with disabilities, people who cannot reach the physical place, etc.)
- Facilitate the research by acquiring and managing detailed data of the place, which the researcher can explore using search engines or data navigation techniques.
- Preserve the physical place, by applying analysis and preservation techniques based on the data collected for digitization.

The proposal is to first use data acquisition technologies, like digital 3D scanners (LIDAR) for artifacts, monuments and place, text recognition (OCR) for manuscripts and manual acquisition for other contents.

The collected data is then inserted in a Data Lake and, with the use of knowledge graphs, they will be organized in a concept-oriented way in order to facilitate searches and navigation.

This approach will require a standardized metadata and data parsing.

Following a user-oriented approach, the data could be presented using different media and interfaces specifically designed for a designed target that could be Brochures, multimedia kiosks, website, mobile apps, VR devices or Braille stands.



DESTINI

Execution of the mobility program in Cyprus

Alberto Morvillo

Research Proposal #2 - Tracking of an artifact transaction using blockchains

Target: Entities or companies that must manage artifact transitions.

Goal: Allow to track artifact transactions easily and securely.

Blockchains are a decentralized approach to store information and are widely used in financial and business fields to avoid the use of a central Certification Authority to validate the transactions. In respect to a centralized approach, which a central entity manages all the transactions and stores all the information in single place; the blockchains uses a peer-to-peer approaches and each party involved in a transaction have a copy of the data, while the validation of them is demanded to other parties.

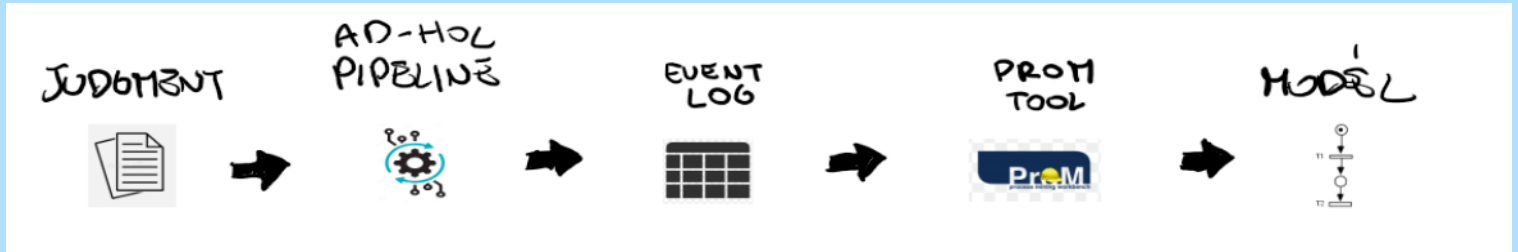
As logistic and supply chains can take advantage of a distributed approach to store the product history and in the food industry there is already some use cases, the proposal is to apply the same methodology to physical artifact, like archaeological or fine arts, to track their transactions or logistic.



DESTINI

Execution of the mobility program in Cyprus

Francesca De Luzi



Research Proposal #1

TARGET: Judge and other persons in the domain of justice (e.g., administrative and technical employee)

SETTING: E-justice

PROBLEM: Management of the backlog in justice processes

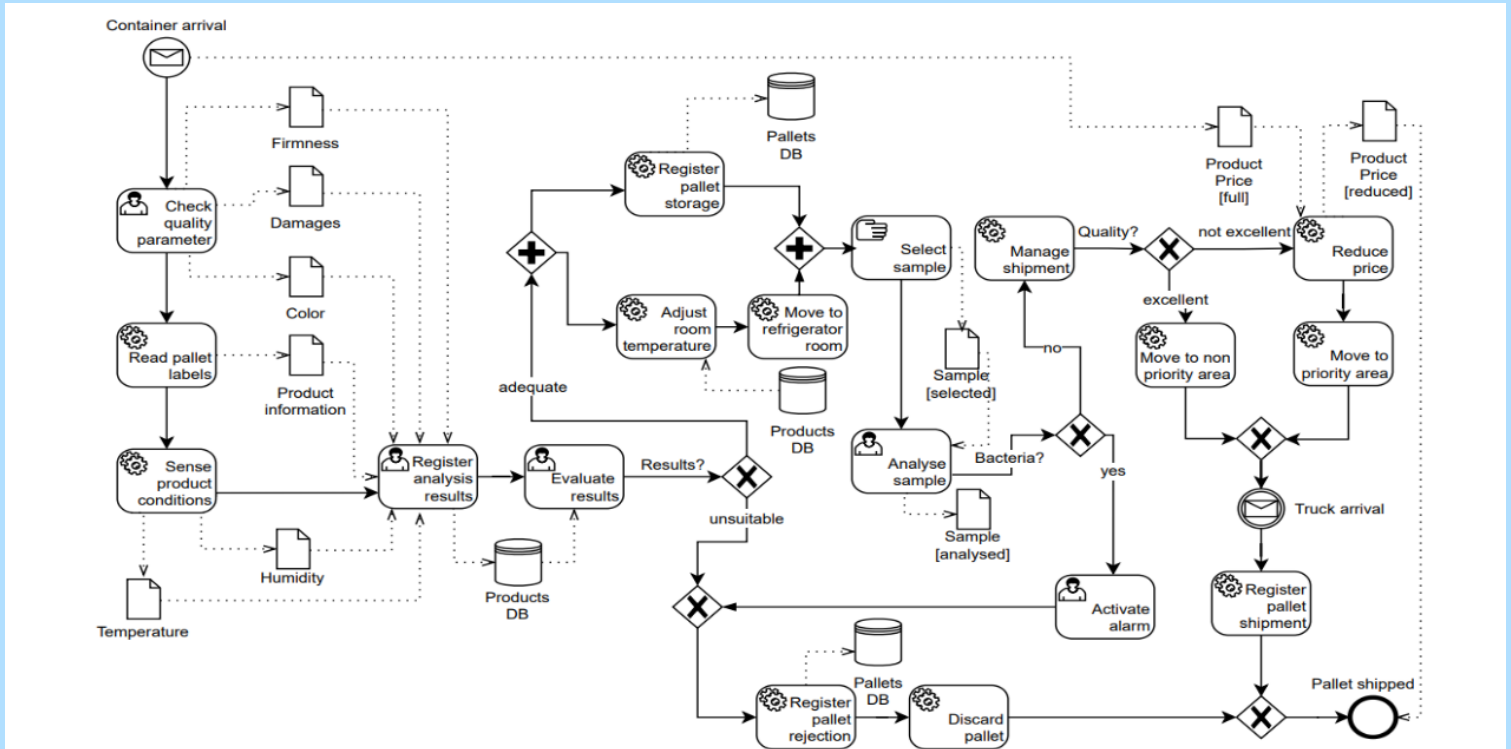
oFor example, the judge often fails to read all the papers carefully, or it is often necessary to differentiate the work immediately

GOAL: combine Digital Twins and process mining in a unified graphical and interactive dashboard to deliver support to the judge to write judicial documents and other tasks such as calendaring through a visual representation of business data and logs

ADVANTAGES: flexible and user-friendly dashboard, generic framework that can be able applied in every business and data context



Francesca De Luzi



Research Proposal #2

TARGET: BP designer – domain expert

SETTING: IoT-aware BPs (e.g., smart manufacturing, logistics, smart health or whatever application domain)

PROBLEM: IoT data quality - due to their limitations in terms of computational power and energy autonomy, they often compromise the quality characteristics with the risk of preventing the correct execution of the entire BP.

GOAL: show how a Digital Twin of data can be used to validate its quality in terms of completeness, accuracy, timeliness, consistency, uniqueness, integrity, compliance

ADVANTAGES: automatically detection of faulty data with the possibility to correct measurement errors



DESTINI

Mirella Sangiovanni

During visiting week in Cyprus

Health & Well-being



Crime, Security & Safety



Nature & AgriFood



My domains as Biologist and future Data Scientist

Talk with
dr. Andreas Christoforou

Sustainability &
Social Entrepreneurship

[6]

JADS



Defining the future steps:

1. Exploring the Forensics Data of NFI
2. Make more clear the Research Problems
3. Starting with simple algorithms to extract informations
4. Look forward to define an Architecture and possibilities to how proceed in this project
5. Looking to find correlations with FCM

Discussion of Fuzzy Cognitive Maps (FCMs) model

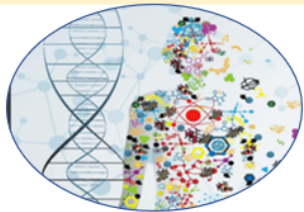


DESTINI

Mirella Sangiovanni

During visiting week in Cyprus

Health & Well-being



Crime, Security & Safety



Nature & AgriFood



Talk with Prof. George Botsaris

My domains as Biologist and future Data Scientist



Sustainability & Social Entrepreneurship

[12]

JADS

JADS Jheronimus Academy of Data Science

TILBURG UNIVERSITY

Cyprus University of Technology



Problems: contaminations and diseases of cow (Mastitis) -> effects on people

1. some ideas came up to write a proposal to support and alert the farms in time
2. Evaluate monitoring systems to get results faster and predict "types of contamination"
3. Data from labs (input)

Department of Agricultural Sciences, Biotechnology and Food Science

14



DESTINI

Execution of the mobility program in Cyprus

Stefan Driessen

Proposal 1: Extending Metadata Blueprints for Data Meshes

Stefan Driessen¹ Michalis Pingos²

¹ *Jheronimus Academy of Data Science, Tilburg University, 's-Hertogenbosch, the Netherlands.*

E-mail: s.w.driessen@jads.nl

² *SEiS Research Lab, Cyprus University of Technology, Limassol, Cyprus.*

E-mail: mfpingos@gmail.com

This document briefly introduces the first of two possible research collaborations based on the research areas of Stefan and Michalis. The ideas discussed herein are intended to serve as high-level early-stage talking points that can be communicated between interested parties.

Short Overview

The proposal focuses on extending the current blueprint for a metadata ontology based on the 5 V's for big data. The blueprints could include new entities that are explicitly relevant for data products on data meshes or data markets.

1 Background and Problems Addressed

Data Lakes were proposed in the 2010's as architectures suitable for dealing with "Big Data" and for assisting organisations towards being *data-driven*. Current literature shows a trend towards more decentral data exchange architectures such as data markets and data meshes [1]. Data Meshes, in particular, attempt to address some of the shortcomings of monolithic data platforms such as data lakes [2, 3]. Creating proper data products puts requirements on metadata templates that are not yet addressed by existing approaches.

2 Proposed Work and Links to Undergoing Research

Currently, Michalis has proposed some blueprint for a metadata ontology based on the 5 V's that drive data lake development [4]. Since data meshes are also designed for handling big data, these metadata aspects should still be relevant. However, as indicated by Stefan's undergoing research collaborations with companies (e.g., Automotive and Telecom), data mesh imposes new requirements and restrictions on the metadata of their data products. One potentially interesting approach is therefore to extend Michalis' templates to include new metadata that is required for data products such as *data product owner*, *usage policy*, *Domain*, etc. This ties into Stefan's current work of setting up a standardised metadata template for data products.



DESTINI

Execution of the mobility program in Cyprus

Stefan Driessen

Proposal 2: Discovering Data Products and Domains

Stefan Driessen¹ Michalis Pingos²

¹ *Jheronimus Academy of Data Science, Tilburg University, 's-Hertogenbosch, the Netherlands.*

E-mail: s.w.driessen@jads.nl

² *SEiIS Research Lab, Cyprus University of Technology, Limassol, Cyprus.*

E-mail: mfpingos@gmail.com

This document briefly introduces the second of two possible research collaborations based on the research areas of Stefan and Michalis. The ideas discussed herein are intended to serve as high-level early-stage talking points that can be communicated between interested parties. Both Michalis and Stefan currently believe this to be the more interesting approach.

Short Overview

The proposal focuses on addressing one of the largest challenges in transitioning towards a data mesh architecture: how to create domains and prioritise data products.

1 Background and Problems Addressed

Many (large) companies and organisations have initiatives to transition from their existing, monolithic, data platforms towards more decentral data exchanges, such as internal data markets or data meshes [1, 2]. One of the main challenges for this transition, in addition to the novelty of the concepts, is how to divide up the data landscape into domains and identifying data assets that should be turned into data products [3]. These organisational challenges are in fact often perceived to be more daunting than the technical challenges associated with data mesh design [4]. Methods for creating domains can (presumably) be found in domain-driven design literature? Methods for prioritising data products are supposedly related to data product valuation techniques, which are currently lacking.

2 Proposed Work and Links to Undergoing Research

Based on Michalis' existing metadata framework it could be possible to establish a hierarchical structure within the domain assets. On the one hand primarily static metadata attributes such as origin and type of data can be used to establish domain boundaries that are neither too large nor too small. On the other hand, dynamic metadata on usage can be used to help organisations and companies prioritise which data assets to turn into data products first.



DESTINI

Smart Data ProcESSing and SysTems of Deep INsight

Conlusion

TWINNING PROJECT - HORIZON2020

The mobility program of DESTINI executed succesfully, and resulted in interesting research outcomes, such as exploration of new research fields, establishment of future collaboration between partners and ESRs.

Contact us



destini2020.eu



facebook.com/destini2020eu



twitter.com/destini2020eu

