DE
DS

ULB

Data Engineering for Data Science

AALBORG UNIVERSITET

ATHENA
Research & Innovation
Information Technologies

UPC

Program information: **https://deds.ulb.ac.be**

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

DTIM

Outline

- Joint doctorate
- Data Engineering for Data Science
- Research Projects
- Industrial Partners

Erasmus Mundus

Marie Sklodowska-Curie Innovative Training Network

JOINT DOCTORATE

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

DTIM

Former Joint PhD programme

8 Years of Experience

IT4BI DC

Information Technologies for Business Intelligence

Doctoral College

Université Libre de Bruxelles (ULB)
Aalborg Universitet (AAU)
Universitat Politècnica de Catalunya (UPC)
Technische Universität Dresden (TUD)
Poznań University of Technology (PUT)

https://it4bi-dc.ulb.ac.be

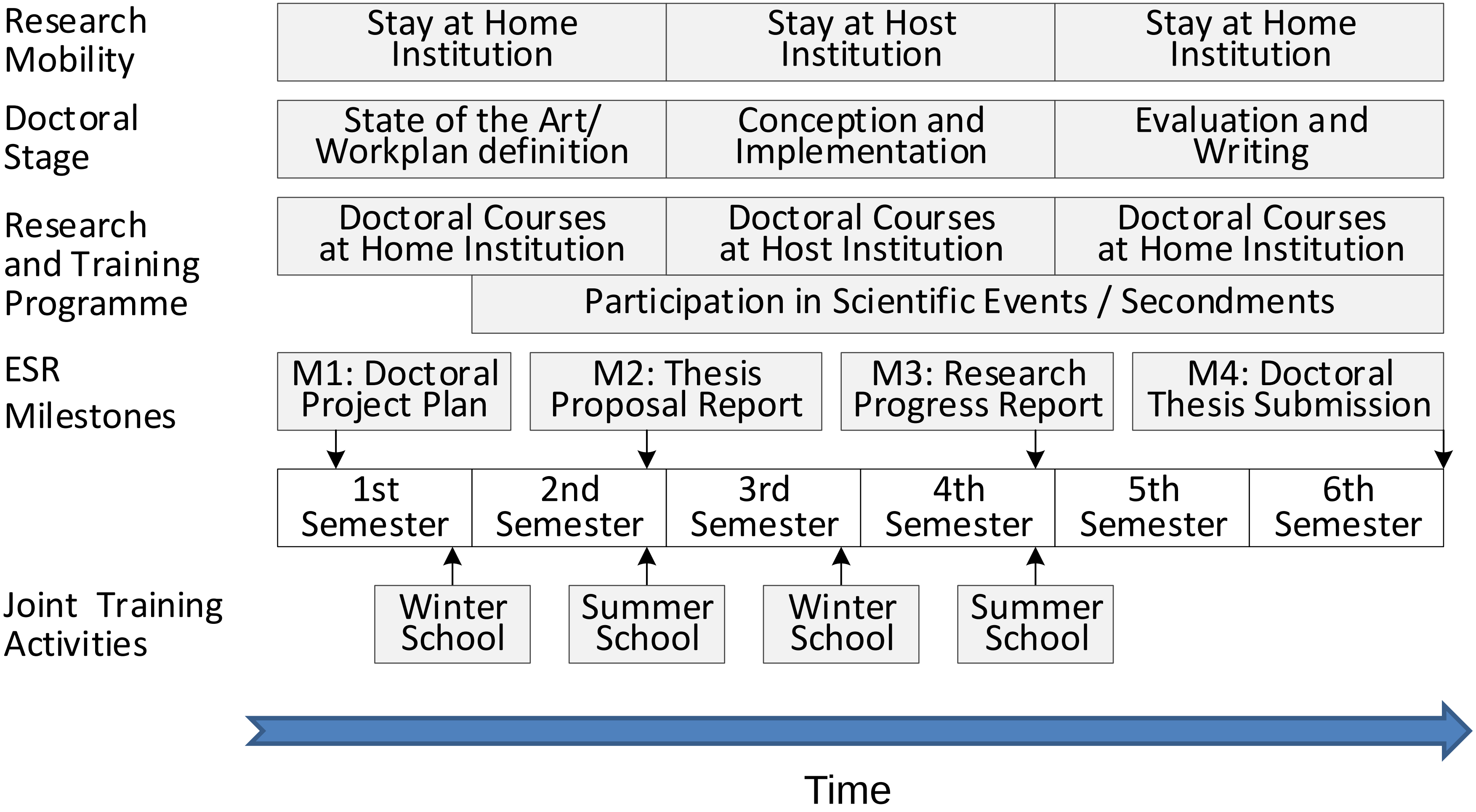
Joint PhD programme

147 / 1503 Granted

Horizon 2020
Programme

Université Libre de Bruxelles (ULB)
Aalborg Universitet (AAU)
Universitat Politècnica de Catalunya (UPC)
Athena Research Center (ARC)

Academic processes



Research Mobility

Stay at Home Institution | Stay at Host Institution | Stay at Home Institution

Doctoral Stage

State of the Art/ Workplan definition | Conception and Implementation | Evaluation and Writing

Research and Training Programme

Doctoral Courses at Home Institution | Doctoral Courses at Host Institution | Doctoral Courses at Home Institution

Participation in Scientific Events / Secondments

ESR Milestones

M1: Doctoral Project Plan | M2: Thesis Proposal Report | M3: Research Progress Report | M4: Doctoral Thesis Submission

1st Semester | 2nd Semester | 3rd Semester | 4th Semester | 5th Semester | 6th Semester

Joint Training Activities

Winter School | Summer School | Winter School | Summer School

Time

Economics per ESR

- **Gross amount: 3,270€/month (39,240€/year)**
  - Correction coefficient per country (BE:100.0%; DK:135.0%; EL:88.7%; ES:95.4%)
  - The net salary results from deducting all compulsory (employer /employee) social security contributions as well as direct taxes (e.g., income tax)
- **Mobility allowance: 600€/month**
- **Family allowance: 500€/month**
  - Should the person be linked to the researcher by
    - (i) marriage,
    - (ii) a relationship with equivalent status to a marriage, or
    - (iii) dependent children
  - The status of the researcher will not evolve over the lifetime of the action

Elegibility

- Be **Early-Stage Researchers** (ESRs)
  - Having 300ECTS (typically 180BSc+120MSc)
- Be of **any nationality**
  - Nationality is therefore not a criterion
- Undertake **transnational mobility**
  - Must not have resided or carried out their main activity (work, studies, etc.) in the country of the recruiting beneficiary for more than 12 months in the 3 years immediately before the recruitment date
  - The eligibility of the researcher will be determined at the date of their **first recruitment** in the action

Data Science flows

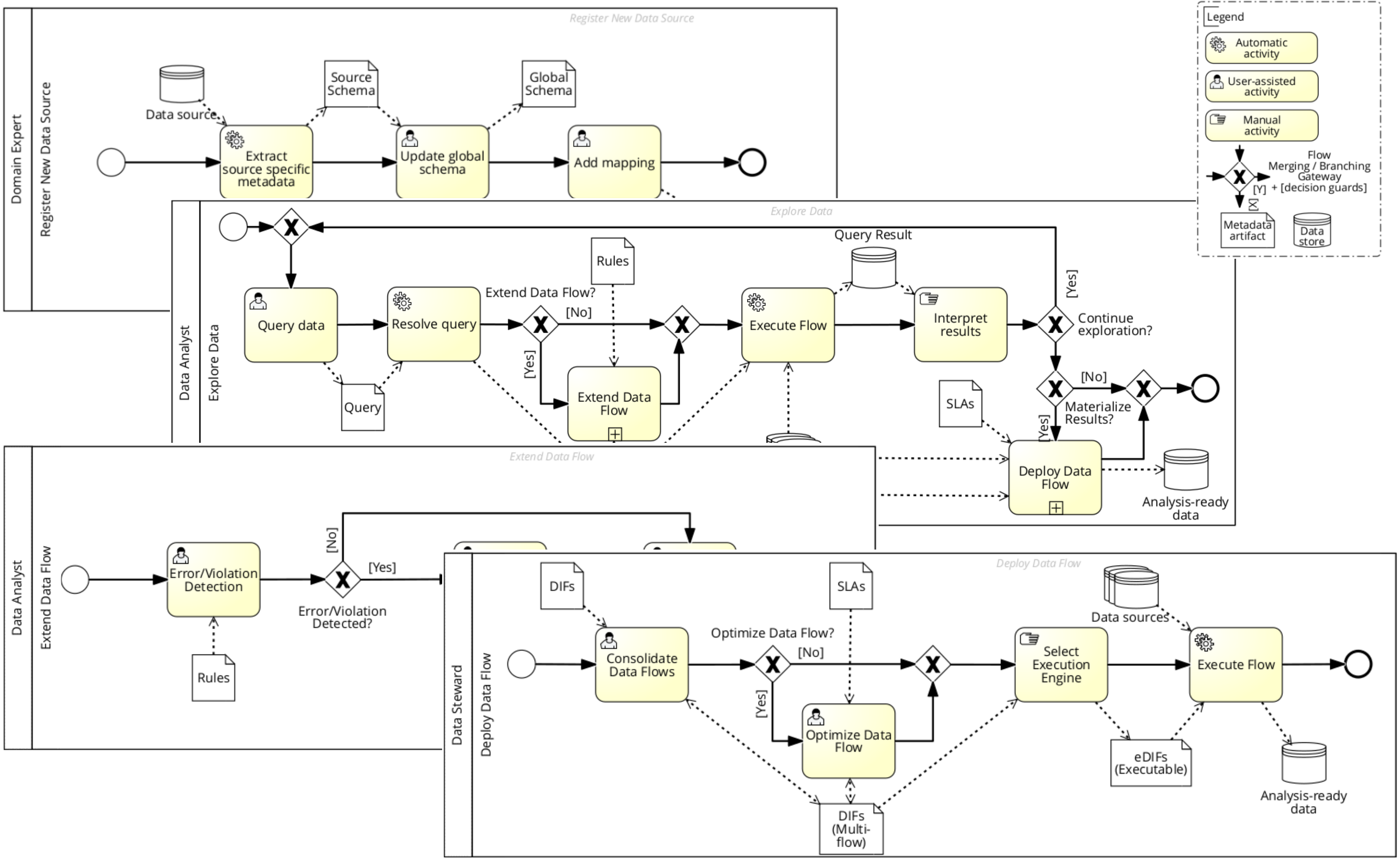
Big Data Management System

DATA ENGINEERING FOR DATA SCIENCE

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

DTIM

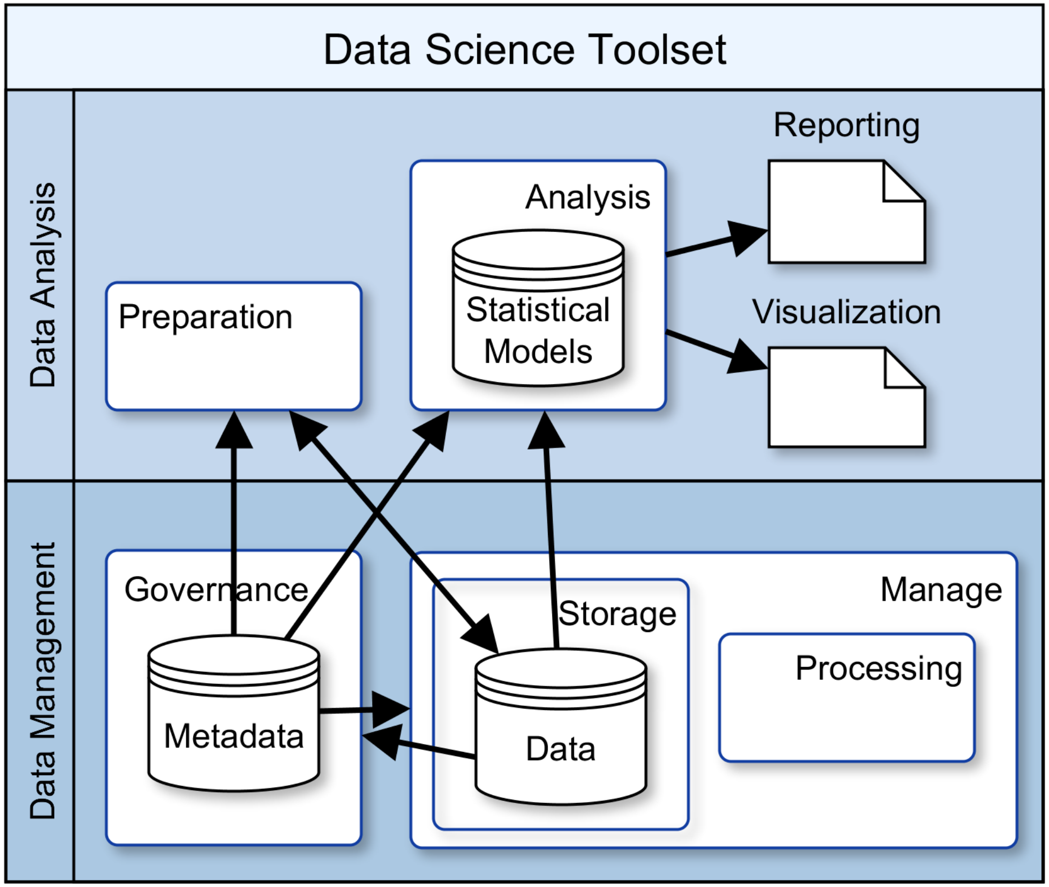
Data Science processes



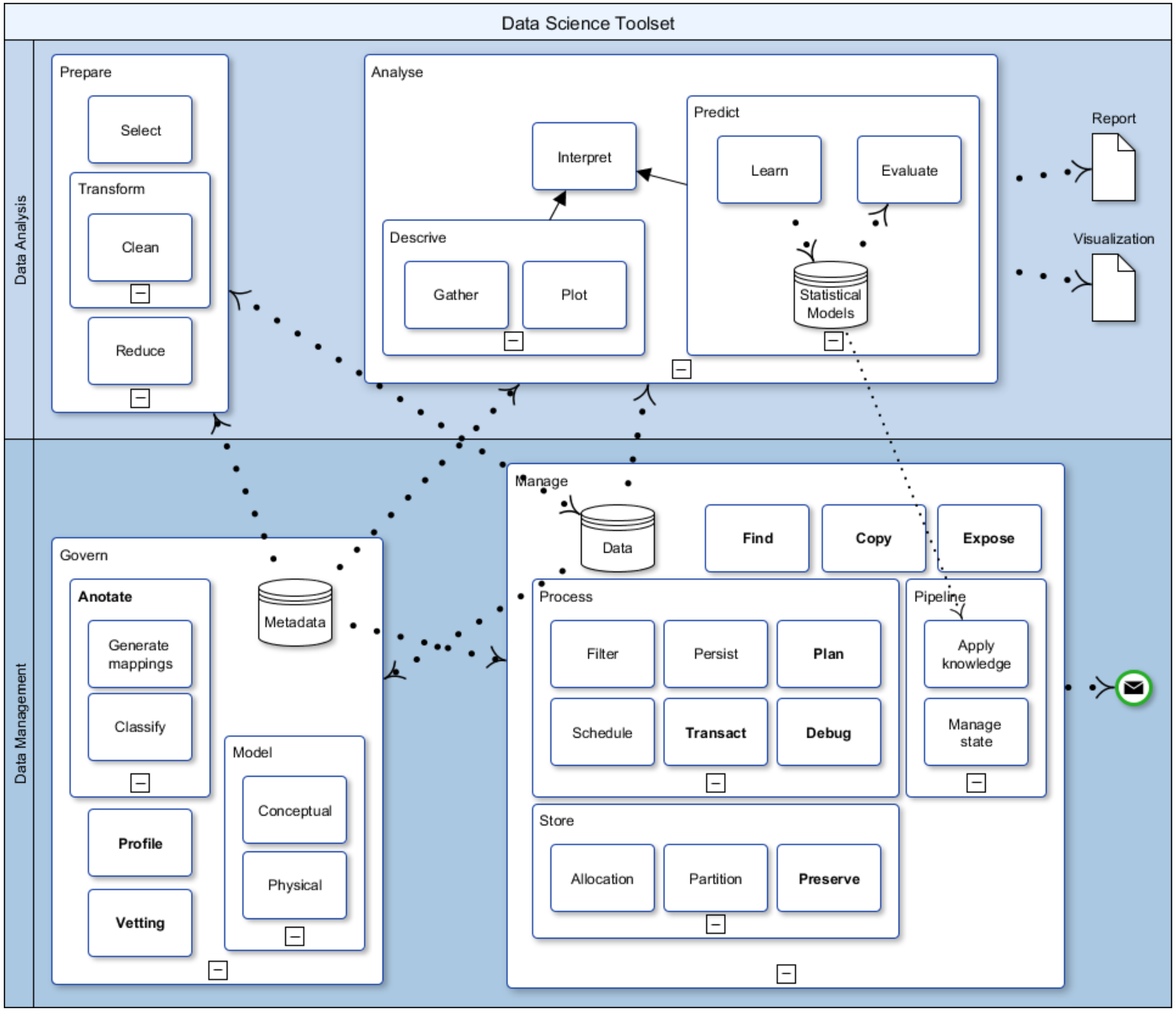
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

DTIM

Big Data Management System



Submodules



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

DTIM

Work packages

Early Stage Researchers

Challenges

Approaches

RESEACH PROJECTS

Early Stage Researchers

Research Projects		ULB	UPC	AAU	ARC	Discipline	Functional Modules
ESR1.1	Semantic-aware heterogeneity management	□	■			Customer	WP1 Governance
ESR1.2	Traceability in big data processing		■	□		Health	
ESR1.3	Privacy-aware data integration			■	□	Health	
ESR2.1	Transparent in-situ data processing		■		□	Transport	WP2 Storage and Processing
ESR2.2	Distribution and replication for feature selection	□	■			Customer	
ESR2.3	Model-based storage for time series	□		■		Energy	
ESR2.4	Analytic operators for trajectories	■		□		Transport	
ESR2.5	End-to-end optimisation for data science in the wild	■	□			Energy	
ESR2.6	Physical optimisation for large scale, DS workloads		□		■	Transport	
ESR3.1	Spatio-temporal data integration & analysis	□		■		Transport	WP3 Preparation
ESR3.2	Synopses-driven data integration & federated learning		□		■	Finance	
ESR3.3	Unified information extraction for data preparation	■		□		Customer	
ESR4.1	Interactive exploration & analytics on complex big data			□	■	Customer	WP4 Analysis
ESR4.2	Scalable model selection in stream settings	■			□	Finance	
ESR4.3	A platform for prescriptive analytics			■	□	Energy	

■■■■ Main supervisor from the Home institution ▨ Use case disciplines

□□□□ Co-supervisor from the Host Institution

WP1
Governance

ESR1.1 (UPC/ULB)

Semantic-aware heterogeneity management

The **variety** of data models and technologies among different systems requires that metadata are flexibly represented and governed. While current approaches introduce vocabularies to represent basic metadata artifacts (e.g., data schema, user queries), most of the processing still relies on the user explicitly querying the original artifacts and not the vocabulary. This project will **use semantic graphs** to develop models, techniques, and methods that **automatically select and combine** basic **metadata artifacts** into more complex ones, supporting advanced user assistance data governance tasks such as integration of new instances, traceability or privacy-aware processing, considering semantics.

WP1
Governance

ESR1.2 (UPC/AAU)
Traceability in big data processing

It is common for analysts to work in collaborative environments relying on **partial results** produced by others, having only a limited view of the whole data flow. Thus, it is needed to systematically **keep track of all transformations** of data from the source to the result of the analysis, and evaluate how the original characteristics of the data are affected, to **guarantee traceability and improve trustability**. In this project, we will cope with problems ranging from the amount of annotations produced (bigger than the data themselves), to the impact of data transformations in the **quality metrics**, going through the process distribution. We will first collect metadata at different granularities without overloading the processing engine, then link them properly, and **estimate the effect of transformations**.

WP1
Governance

ESR1.3 (AAU/ARC)
Privacy-aware data integration

Basically all applications in Data Science need to **combine/integrate data** originating from multiple **heterogeneous datasets** (XML, Excel, RDF, etc.). This challenging task becomes even more challenging if we consider that some parts of the data might have to be protected to **ensure privacy**. This is particularly true for applications in **health care**, where some data requires protection while still allowing for efficient access and analysis (e.g., patient data or results of certain studies). This project will focus on the **integration of private personal data** from multiple sources, by carefully ensuring the level of exposition of every instance.

WP2
Storage and
Processing

ESR2.1 (UPC/ARC)
Transparent in-situ data processing

Bringing computation closer to data is a must when dealing with large datasets in highly distributed systems with very **specialised hardware** components. Some relevant examples are supporting in-situ data transformations (e.g., decrypt on read) and tier crossings (e.g., select the form of compression based on physical medium to be moved to). The goal of this project is to **improve the performance** of the data processing engine, such as Spark, by leveraging hardware specificities, without affecting the interface to applications. The approach consists in improving scalability by decoupling the engine primitives from the underlying data store platform, in such a way that tasks can be delegated to **avoid data transfers** through the different levels of the stack, and benefit from specific hardware.

WP2
Storage and
Processing

ESR2.2 (UPC/ULB)

Distribution and replication for feature selection

The **curse of dimensionality** is more and more relevant as the number of features to be tackled keeps on increasing in many areas, such as bioinformatics, predictive maintenance and IoT. The solution is to perform a careful and efficient **feature selection** before analysing the dataset. Thus, the purpose of this project is to propose the appropriate **data distribution and replication policies**, taking into account their specific semantics rather than using a generic approach independent of the application, to optimise filter-based feature selection methods, such as mRMR, and their generalisation to other kinds of algorithms and their associated data structures.

ESR2.3 (AAU/ULB)

Model-based storage for time series

WP2
Storage and
Processing

Industrial sensors, like those in wind turbines, generate large amounts of **never ending time series**, but only small parts of them (e.g., averages over 15 minute windows) can be handled and stored for analysis. Provided that many TBs or even PBs must be handled, it is vital to develop new methods for **incremental storage and fast retrieval**, that avoid accessing raw data to produce results. In this project, we focus on how to store such time series data by means of models and investigate how these models can be incrementally maintained and used to access both past data as well as **predict future data**.

ESR2.4 (ULB/AAU)
Analytic operators for trajectories

WP2
Storage and Processing

Similarity joins, i.e., the matching of similar pairs of objects, is a fundamental operation in data management. In the **spatial setting**, the problem can be stated as follows: Given the sets P and Q of trajectories and a similarity threshold, return all **pairs of trajectories** with a similarity that exceeds . In the maritime domain, identifying similarities between trajectories of vessels is crucial for classifying vessel activities (e.g., fishing). In particular, **(near) real-time** similarity search is necessary to successfully identify events related to navigational safety (e.g., piracy). Real-time similarity search is a computationally challenging problem, however. As such, the objective in this project is to develop techniques that allow efficient (near) real-time trajectory similarity search.

WP2
Storage and
Processing

ESR2.5 (ULB/UPC)

End-to-end optimisation for data science in the wild

Contemporary analytical pipelines **interconnect a myriad of scripts**, programs and tools, often spanning a **multitude of programming languages** (e.g., R, Python) and associated libraries (e.g., Spark, Tensorflow). While this re-use of existing frameworks reduces development time, it leads to sub-optimal performance due to the lack of **end-to-end optimisations** such as merge of loops or removal of materialisation points. The objective of this project is to propose techniques that allow such optimisation for analytic pipelines expressed in multiple languages and spanning multiple existing libraries, by **using a common intermediate representation** in which known DBMS-style optimisations can be expressed.

WP2
Storage and
Processing

ESR2.6 (ARC/UPC)
Physical optimisation for large scale, DS workloads

Modern Data Science processing workloads typically involve computations of extreme-scale analytics that can be encoded in various forms (e.g., queries, workflows, programs) and executed on **more than one platform**. Parts of the processing could be pushed to the edge level (e.g., input sensors), while other more computationally intensive parts (e.g., stock correlation in finance, gene simulations in life sciences) could be executed on one or more, potentially distributed, Big Data platforms or clusters (e.g., GPUs) of a supercomputer or in the cloud. The decision on what is **the right platform and timing to execute** a Data Science workload is based on a **multitude of criteria and optimization objectives**, including hardware and processing capabilities, scheduled and running workloads, available resources and pricing, and so on. In this project, we develop tools and techniques for optimizing (e.g., in terms of runtime, throughput, latency, scheduling, system resources, monetary resources) the execution of Data Science workloads across different computing platforms.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

DTIM

WP3
Preparation

ESR3.1 (AAU/ULB)
Spatio-temporal data integration & analysis

Many companies collect very large quantities of **spatio-temporal data**, e.g., transport companies collect telemetric data that are currently used for scheduling, billing, and other administrative tasks. **Eficiently managing, integrating, and analyzing** such spatio-temporal data together with other types of data is a very challenging task. Based on multiple, large datasets, the project **annotates a digital map** with novel information to be able to accurately quantify the driving, e.g., in terms of fuel consumption that is used for both analytic and predictive analysis. This project will therefore develop **scalable processing** techniques for integrating heterogeneous data with spatio-temporal datasets.

ESR3.2 (ARC/UPC)
Synopses-driven data integration & federated learning

WP3
Preparation

Data preparation is among the most time-demanding parts of analytics. The aim of this project is to investigate techniques, including federated learning, to **scale** data preparation, enabling **low-latency responses and automated proposals** for data integration actions, without compromising data security and data privacy. In one embodiment, the data will be pre-processed for **constructing compact synopses** (i.e., summaries), which will enable the user to quickly navigate through different actions and interactively observe the effect on the data. The project will thus get the set of synopses that are best suited for the problem at hand (e.g., the best instance sample), build these synopses, use them to **visualize the data and their relationships**, and to **propose actions**.

WP3
Preparation

ESR3.3 (ULB/AAU)

Unified information extraction for data preparation

Information extraction, the activity of **extracting structured information from unstructured text**, is a core data preparation step. Systems for information extraction fall into two main categories: **machine-learning-based (ML-based), and rule-based**. In the former, models for extraction tasks are automatically obtained by ML methods, which however requires large amounts of training data. In the latter, extraction rules are manually written by human experts, which obviates training data and has the benefit of the rules being explainable, but is labor intensive. Despite advances in ML, rule-based systems are still widely used in practice. The objective in this project is to design and implement an information extraction system that **combines the benefits of both categories**.

WP4
Analysis

ESR4.1 (ARC/AAU)
Interactive exploration & analytics on complex big data

The large scale and unstructured nature of complex Big Data makes **interactive data exploration and analytics** a non-trivial, expensive task for expert and non-expert data scientists. Traditional exploration and analysis methods often cannot cope with the generation pace and versatility of modern heterogeneous data sources. Supporting interactive data exploration and analytics, without sacrificing **performance** and overall user experience, requires rethinking and co-designing of the data management and analysis layers. In this project, we revisit the interface between data management and data exploration and analysis focusing on performance and expressivity. We identify new primitive **operations for exploration and analytics**, form workflows comprising primitive operations, and design effective solutions for optimizing such workflows. We develop techniques for enabling an advanced, user-friendly experience through enhanced **exploration in natural language** and by **providing recommendations** for spot on data analysis based on the characteristics of the data and the user requirements.

WP4
Analysis

ESR4.2 (ULB/ARC)
Scalable model selection in stream settings

Accurate machine learning requires the execution of a model selection phase to perform a number of design choices (e.g., learner), and to calibrate a number of parameters and **hyperparameters** (e.g., number of layers in a neural network). While model selection is conventionally performed off-line and in main memory, the characteristics of contemporary Big Data scenarios require novel methodologies for performing **accurate model selection in streaming and distributed settings**. So far, the continuous incremental arrival of data has limited the range of considered model selection strategies to simple tuning and **drift detection**. The ambition in this thesis is to provide an extended library of resource-aware model selection techniques that manage the **trade-off between reaction time and accuracy**.

WP4
Analysis

ESR4.3 (AAU/ARC)
A platform for prescriptive analytics

Prescriptive analytics does not only predict the future but also suggests (prescribes) the best course of action to take. It entails information collection, extraction, consolidation, visualisation, forecast, optimisation, and what-if analysis. Thus, as a range of non-integrated and specialized tools have to be glued together in an ad-hoc fashion, building prescriptive analytics solutions is labor-intensive, error-prone, and inefficient. This project will **build a generic platform** for prescriptive analytics, which tightly integrates scalable data storage with **declarative specification of queries**, constraints, objectives, requirements, and mathematical optimisations in a unified framework.

Challenges and approaches

		Solution approach	
Big Data Charac.	Research challenge	Data simplif.	Processing improv.
Volume	Scalability	ESR3.2	ESR2.1 ESR3.1 ESR4.3
	Curse of dimensionality	ESR2.2	
	Curse of modularity		ESR2.4
Velocity	Incremental arrival		ESR2.3 ESR4.2
	Real-time processing		ESR2.6 ESR4.1
Variety	Heterogeneity	ESR1.1	
	Minimise data movements		ESR2.5
	Dirty and noisy data	ESR3.3	
Veracity	Traceability		ESR1.2
Privacy		ESR1.3	

WP1 WP2 WP3 WP4 Functional Modules

Secondments

Tools

INDUSTRIAL PARTNERS PARTICIPATION

Disciplines

ESR2.4 MarineTraffic

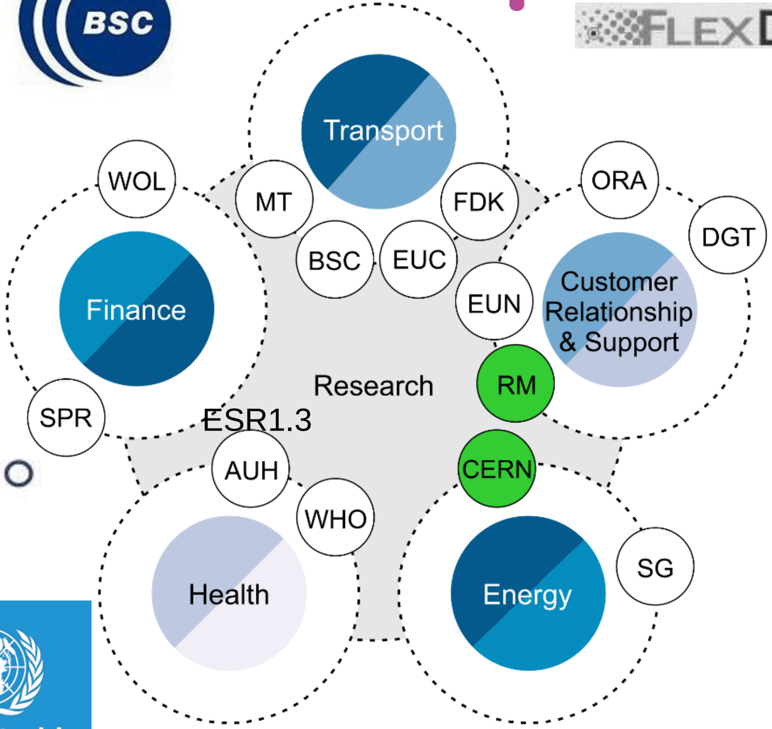
ESR2.1 BSC

ESR2.6 eurecat

ESR3.1 FLEXDANMARK

ESR2.2 orange™

Worldline ESR4.2



Generalitat de Catalunya ESR1.1

spring techno ESR3.2

EURANOVA ESR3.3

World Health Organization ESR1.2

CERN ESR2.5

SIEMENS Gamesa RENEWABLE ENERGY ESR2.3 ESR4.3

rapidminer ESR4.1

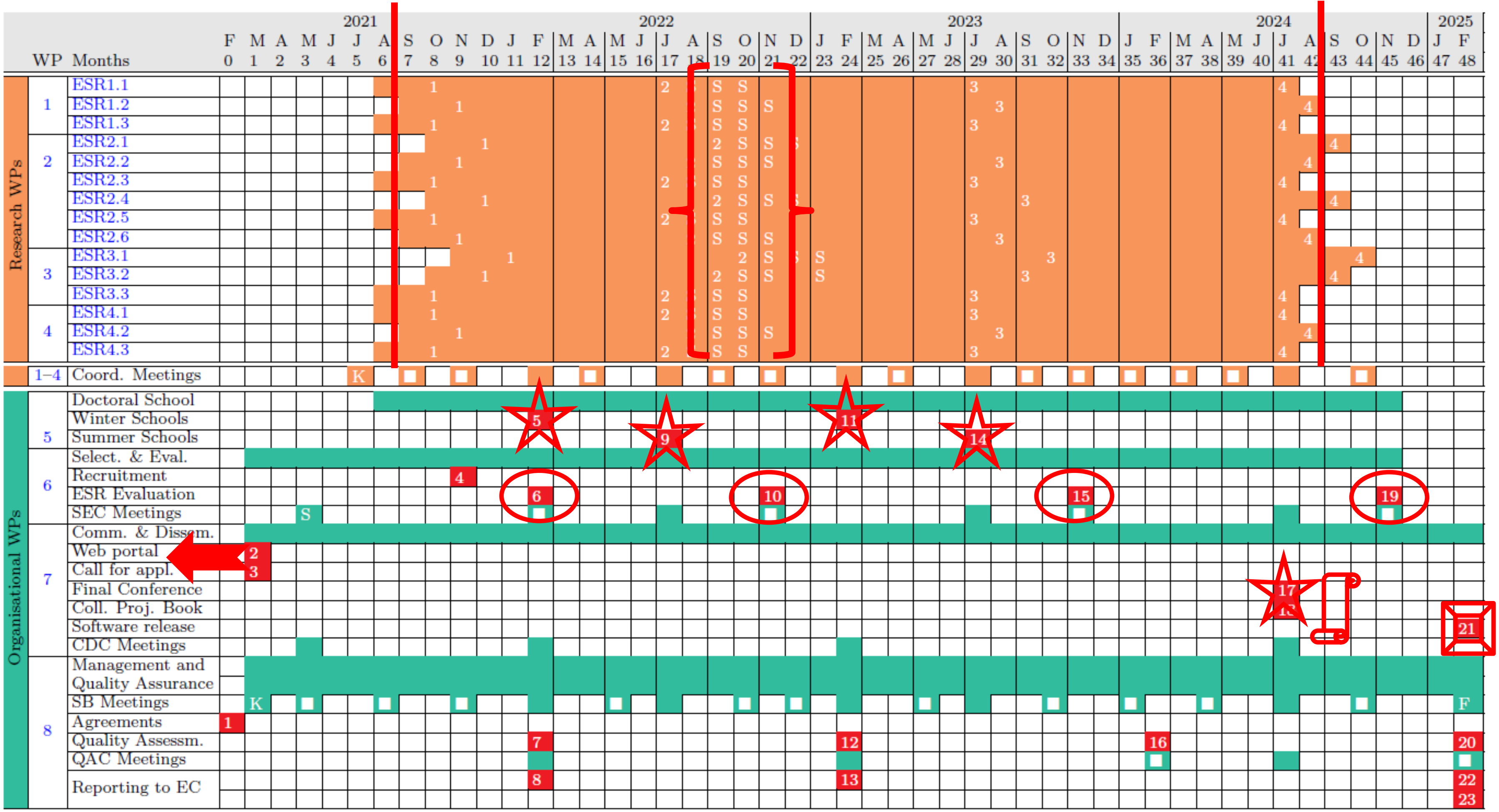
Open source tools

Plan

Contact

CONCLUSIONS

Gantt



S = Secondment, 1 = ESR Milestone 1 (DPP), 2 = ESR Milestone 2 (TPR), 3 = ESR Milestone 3 (RPR), 4 = ESR Milestone 4 (Thesis Submission), n = Project Milestone,
K K = Kick-off meeting, S = Selection meeting, 🟧 🟩 = Face-to-face meeting, □ □ = Teleconference, F = Final meeting

N.B. For maximising the availability of potential candidates, the project is planned to start in March 2021 so that the hiring of the ESRs is synchronised with the start of the academic year. The initial month of the ESRs may vary by up to 3 months depending on various factors (Sect. 3.2.6); the Gantt chart reflects this but does not commit any ESR to a particular starting month. Since 3 months are typically required to organise a thesis defence after the thesis submission, all defences may take place within the scope of the project. The months for the secondments will be adapted to the specificities of the candidate, the topic, and the hosting institution; the Gantt chart depicts a typical situation for the secondments. The leaders of the research WPs will meet with the ESRs of their WP every 2.5 months to coordinate their work in producing the deliverables and the collective project book (Annex M).

DE
DS

Database Technologies and Information Management

http://www.essi.upc.edu/dtim

Contact (at UPC): {aabello|oromero}@essi.upc.edu

@romero_m_oscar

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

DTIM