Session 2.

A Project perspective on Big Data architectural pipelines and benchmarks

EUROPEAN BIG DATA VALUE FORUM 2020

DataBench

Panelists

Arne Berre
Chief Scientist, SINTEF

Leonidas Kallipolitis
Aegis - I-BiDaaS Projects

Brian Elvesæter
Research scientist, SINTEF
TheyBuyForYou Project

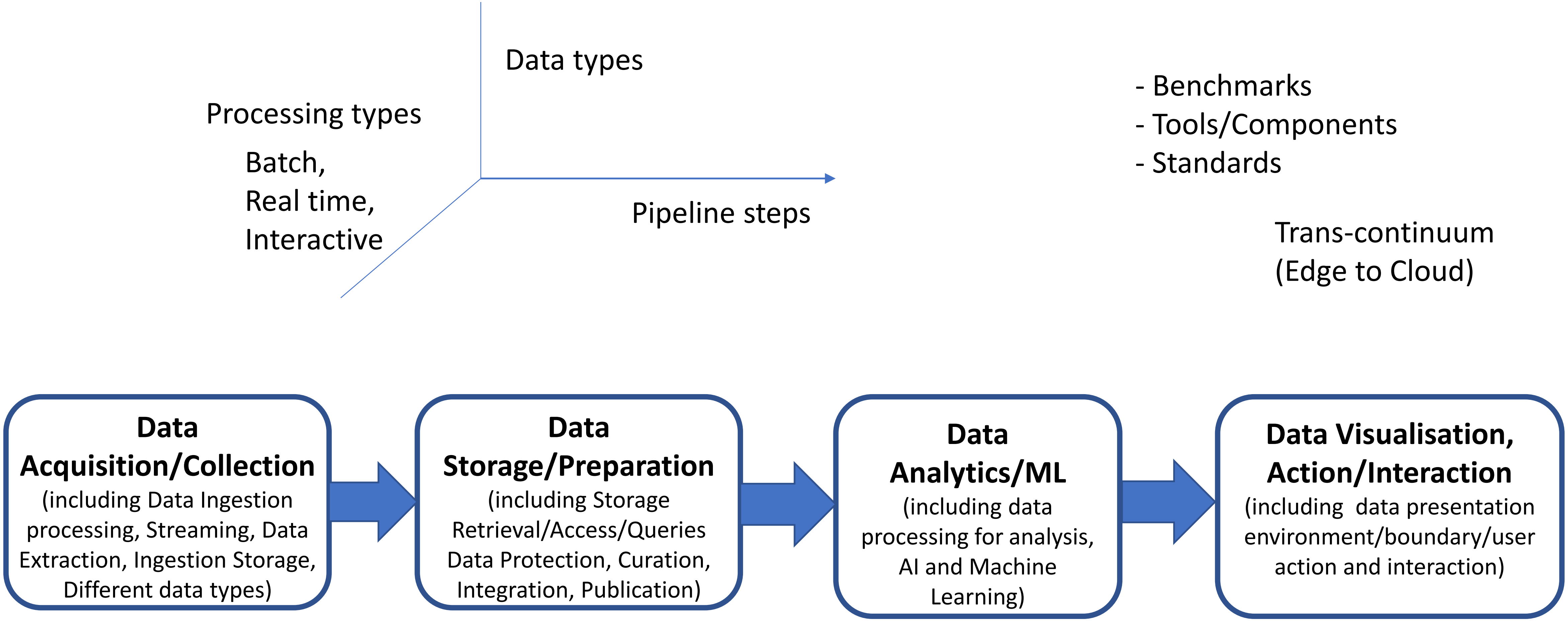
Athanasios Koumparos
Senior Software Engineer,
Vodafone Innovus
Track&Know Project

Jon Ander Gómez Adrián
Universitat Politecnica de Valencia
DeepHealth Project

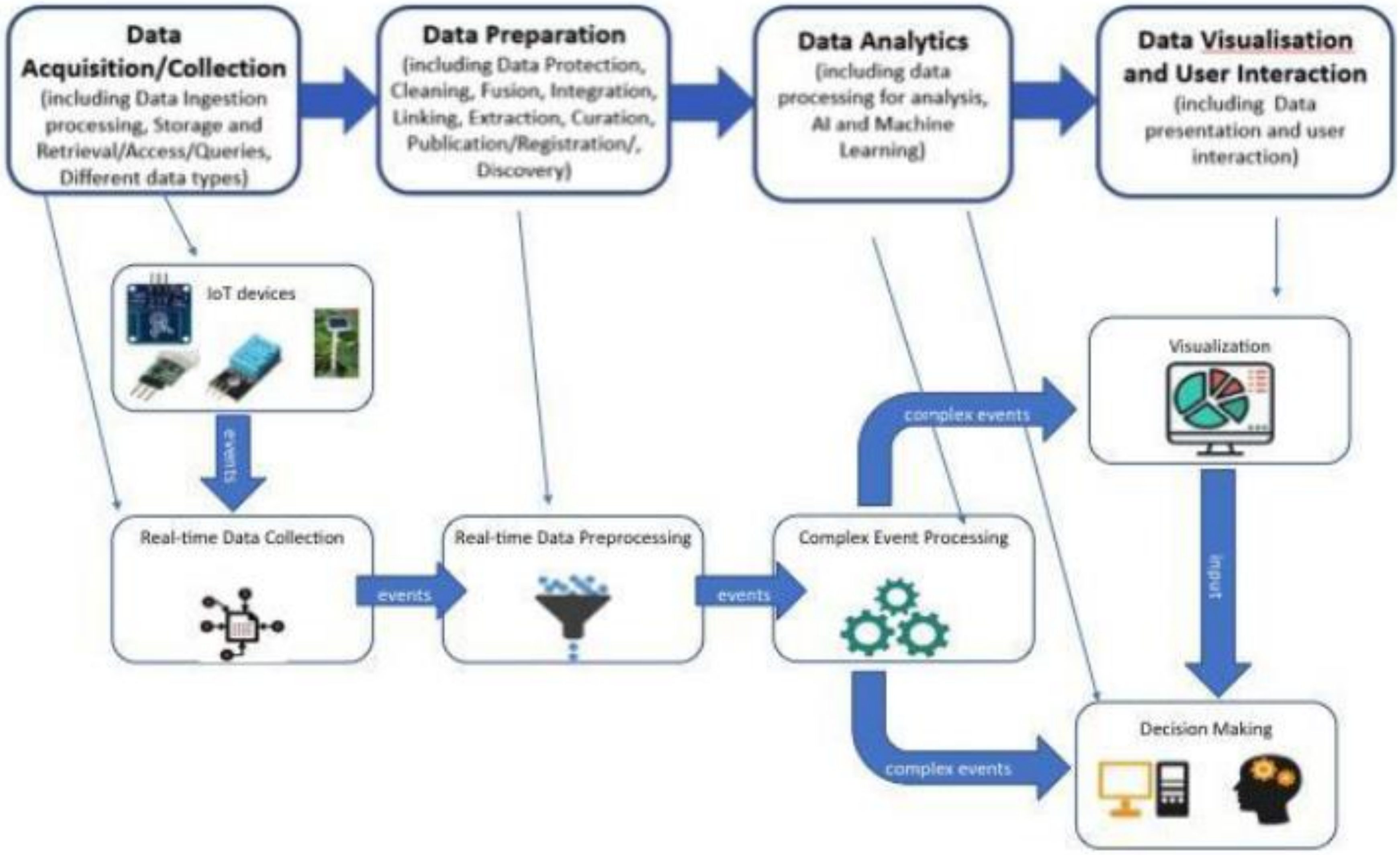
Caj SÖDERGÅRD
Professor, VTT Data-driven
Solutions
DataBio Project

Introduction to Architectural pipelines

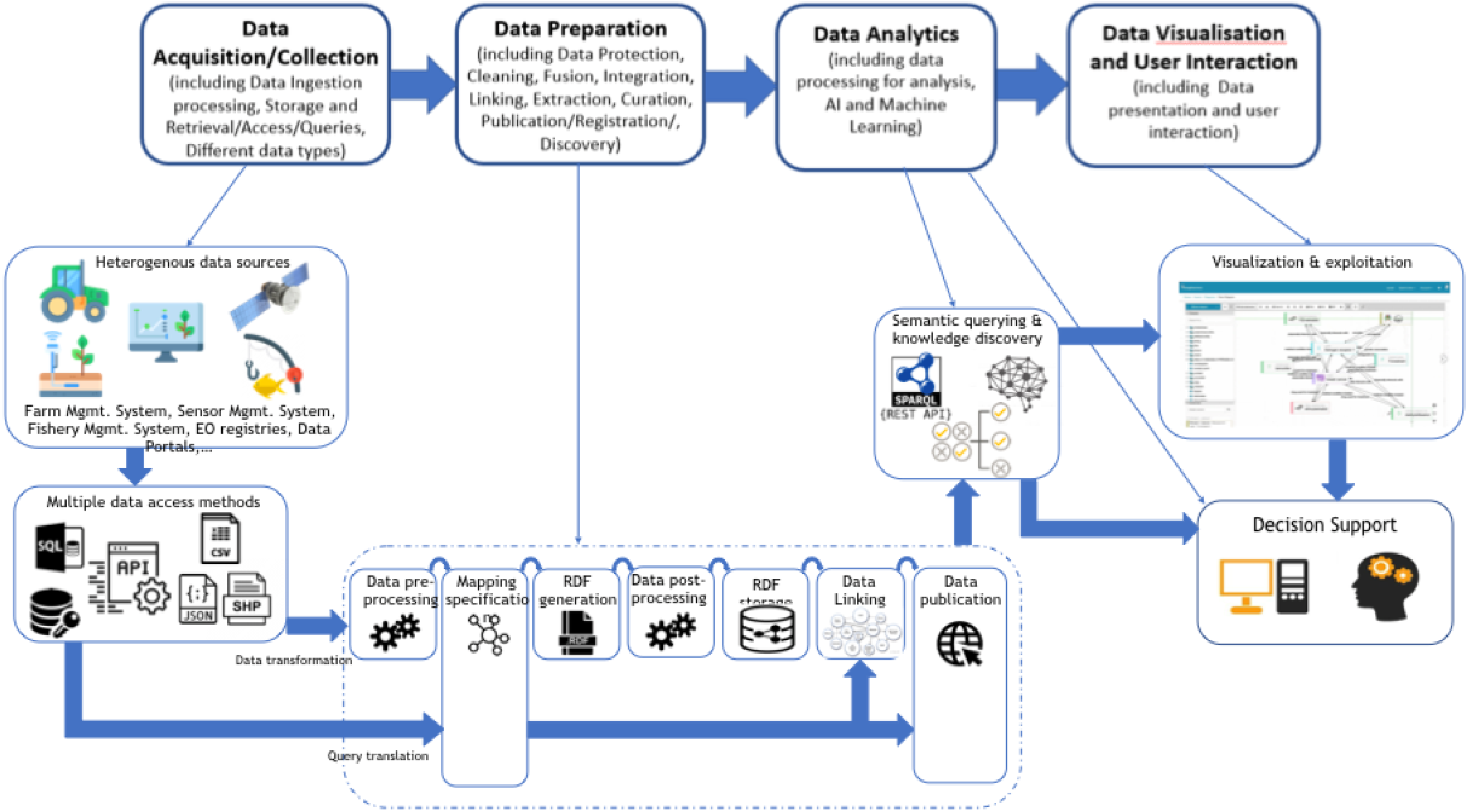
- I-BiDaaS
- TBFY
- Track&Know
- DataBio
- DeepHealth

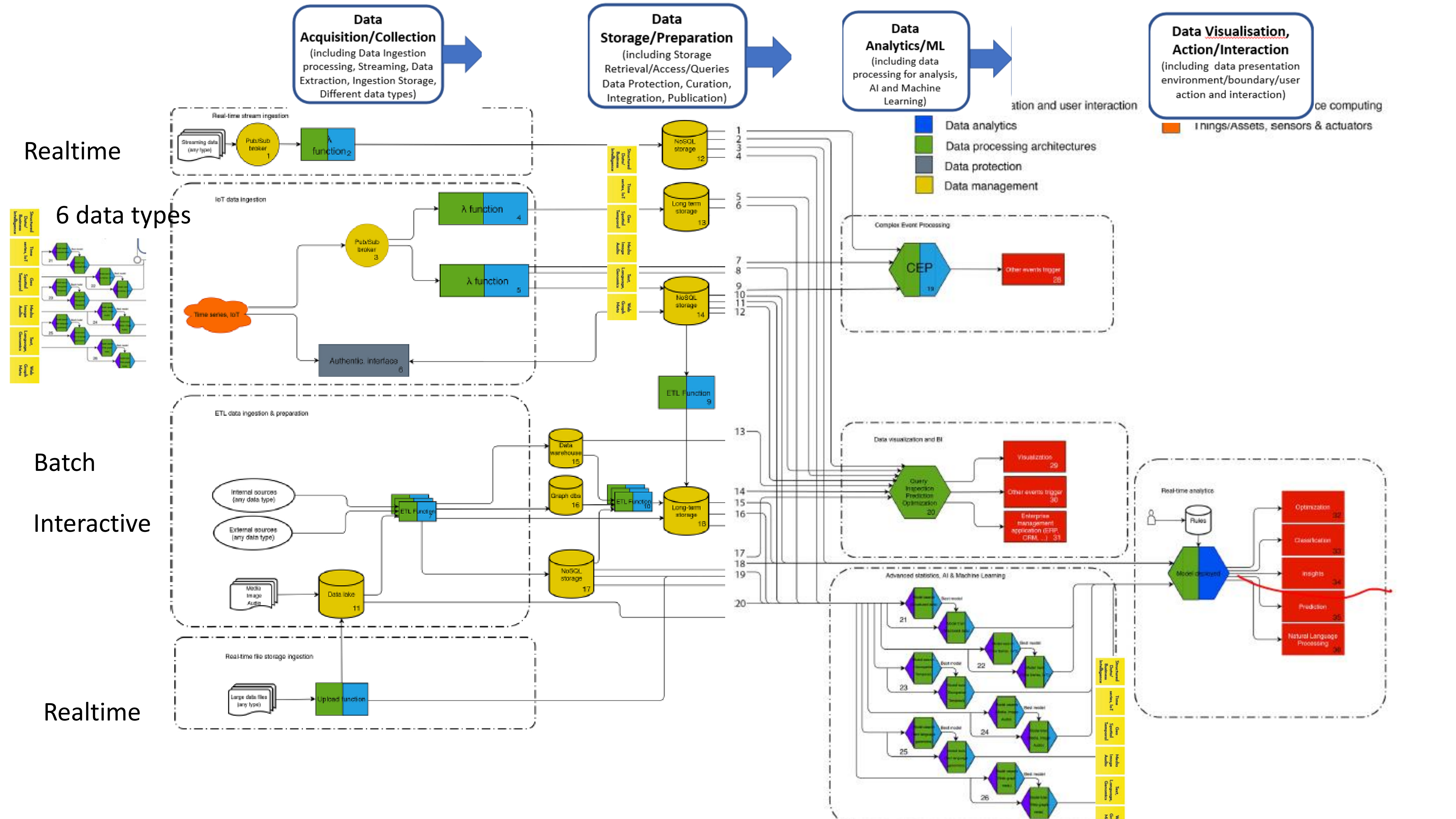

Data types
Processing types
Batch,
Real time,
Interactive
Pipeline steps
- Benchmarks
- Tools/Components
- Standards
Trans-continuum
(Edge to Cloud)
Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)
Data Visualisation, Action/Interaction
(including data presentation environment/boundary/user action and interaction)

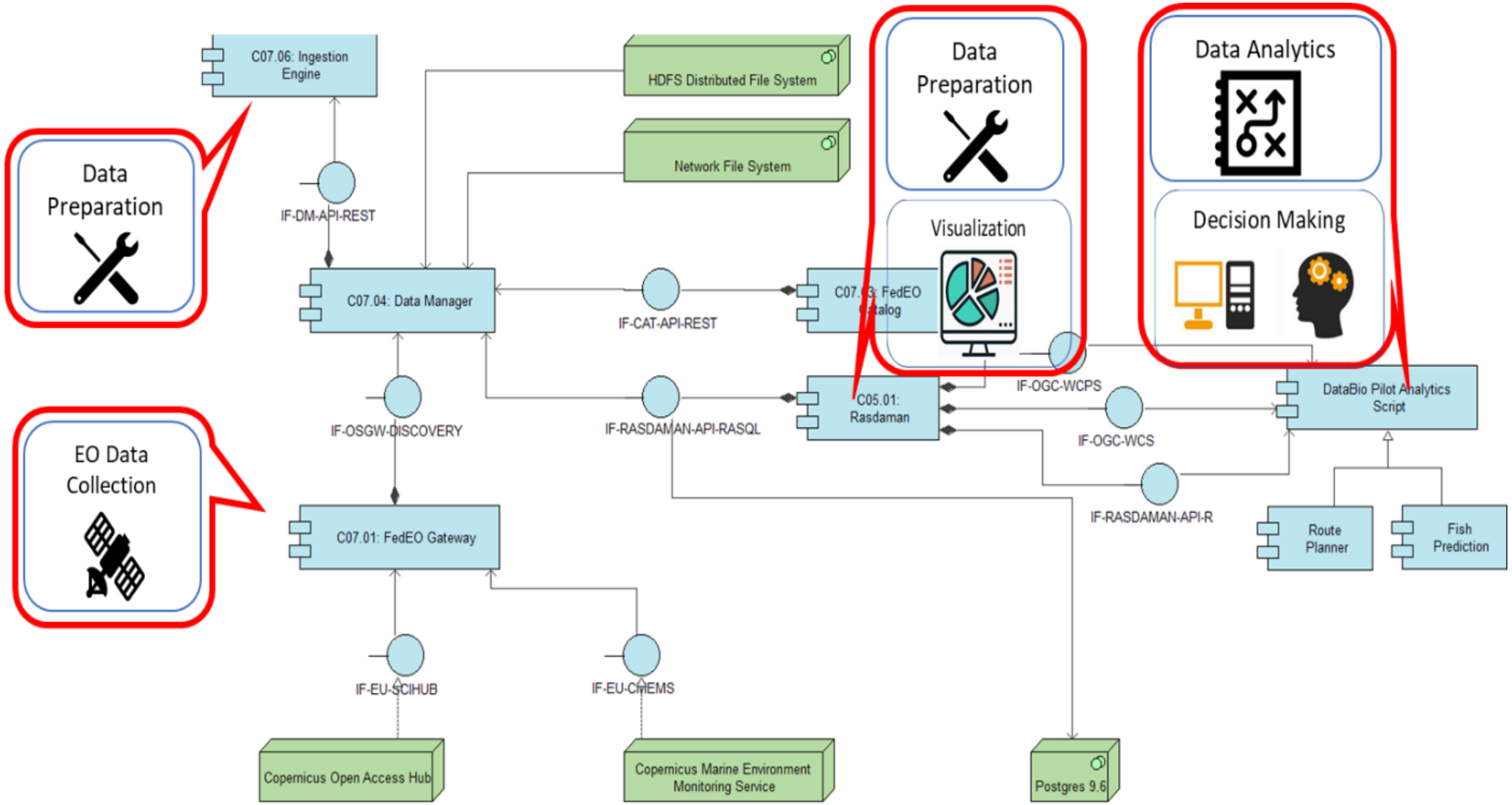
IoT – Real time Data Pipeline



Knowledge Graph/Ontology/Linked Data pipeline




Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)
Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)
Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)
Data Visualisation, Action/Interaction
(including data presentation environment/boundary/user action and interaction)
Realtime
6 data types
Batch
Interactive
Realtime
Real-time stream ingestion
Streaming data (any type)
Pub/Sub broker 1
λ function 2
IoT data ingestion
λ function 4
Pub/Sub broker 3
λ function 5
Time series, IoT
Authentic. interface 6
ETL data ingestion & preparation
Internal sources (any data type)
External sources (any data type)
ETL Function 7
Media Image Audio
Data lake 11
Real-time file storage ingestion
Large data files (any type)
Upload function
NoSQL storage 12
Long term storage 13
NoSQL storage 14
ETL Function 9
Data warehouse 15
Graph dbs 16
ETL Function 10
Long-term storage 18
NoSQL storage 17
ation and user interaction
Data analytics
Data processing architectures
Data protection
Data management
ce computing
Things/Assets, sensors & actuators
Complex Event Processing
CEP 19
Other events trigger 28
Data visualization and BI
Query Inspection Prediction Optimization 20
Visualization 29
Other events trigger 30
Enterprise management application (ERP, CRM, ...) 31
Advanced statistics, AI & Machine Learning
Real-time analytics
Rules
Model deployed
Optimization 32
Classification 33
Insights 34
Prediction 35
Natural Language Processing 36



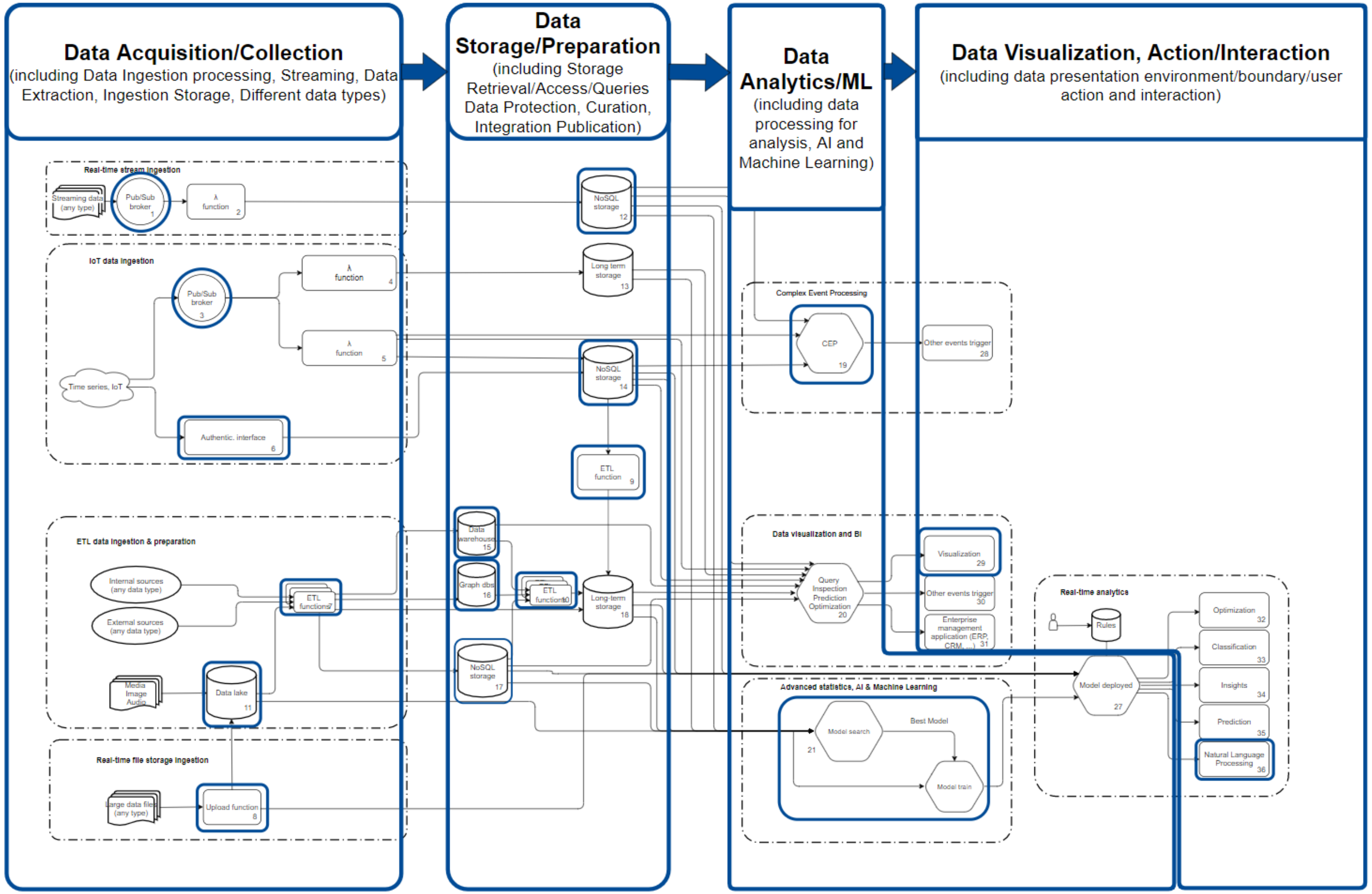
Data Preparation
EO Data Collection
C07.06: Ingestion Engine
IF-DM-API-REST
HDFS Distributed File System
Network File System
C07.04: Data Manager
IF-CAT-API-REST
C07.03: FedEO Catalog
Data Preparation
Visualization
Data Analytics
Decision Making
IF-OSGW-DISCOVERY
IF-RASDAMAN-API-RASQL
C05.01: Rasdaman
IF-OGC-WCPS
IF-OGC-WCS
IF-RASDAMAN-API-R
DataBio Pilot Analytics Script
Route Planner
Fish Prediction
C07.01: FedEO Gateway
IF-EU-SCIHUB
IF-EU-CMEMS
Copernicus Open Access Hub
Copernicus Marine Environment Monitoring Service
Postgres 9.6

Search Search by Blueprint/Pipeline

Search by step of the Data Value Chain or for the specific components of the generic big data pipeline generated by DataBench

DataBench has devised a generic architectural blueprint mapped to a top level pipeline along the Data Vale Chain covering the steps of **Data Acquisition/Collection** (including data ingestion, processing, streaming, extraction and ingestion storage), **Data Preparation/Storage** (including storage retrieval/access/queries, data protection, curation, integration and publication), **Data Analytics** (including data processing for analysis, AI and Machine Learning) and **Data Visualization/Interaction** (including data presentation, environment/boundary/user action and interaction).

By clicking on the image below on one of the four steps of the pipeline, or in one of the specific elements from the generic blueprint, this search interface will help you discovery benchmarks and associated knowledge (nuggets)



Conclusion on Pipelines and related benchmarks,
Arne J. Berre, SINTEF

- I-BiDaaS
- TBFY
- Track&Know
- DataBio
- DeepHealth

EUROPEAN BIG DATA VALUE FORUM

3 - 5 NOVEMBER . 2020 - BERLIN + VIRTUAL

I-BiDaaS

Industrial-Driven Big Data as a Self-Service Solution

Industrial-Driven Big Data as a Self-Service Solution

Leonidas Kallipolitis, AEGIS

European Commission

German Research Center for Artificial Intelligence

BDV BIG DATA VALUE ASSOCIATION

BERLIN

BERLIN PARTNER for Business and Technology

Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

PLATTFORM INDUSTRIE 4.0

I-BiDaaS
Industrial-Driven Big Data as a Self-Service Solution

Project Consortium
13 partners

TOTAL BUDGET / TOTAL EC FUNDING
€ 4 997 035

START DATE
1 January 2018

PROJECT NAME
Industrial-Driven Big Data as a Self-Service Solution

PROJECT TYPE
RIA

DURATION
36 months

Identity Card

http://www.ibidaas.eu/ @Ibidaas https://www.linkedin.com/in/i-bidaas/

Consortium

1. FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS (**FORTH**)
2. BARCELONA SUPERCOMPUTING CENTER - CENTRO NACIONAL DE SUPERCOMPUTACION (**BSC**)
3. IBM ISRAEL - SCIENCE AND TECHNOLOGY LTD (**IBM**)
4. CENTRO RICERCHE FIAT SCPA (**CRF**)
5. SOFTWARE AG (**SAG**)
6. CAIXABANK, S.A (**CAIXA**)
7. THE UNIVERSITY OF MANCHESTER (**UNIMAN**)
8. ECOLE NATIONALE DES PONTS ET CHAUSSEES (**ENPC**)
9. ATOS SPAIN SA (**ATOS**)
10. AEGIS IT RESEARCH LTD (**AEGIS**)
11. INFORMATION TECHNOLOGY FOR MARKET LEADERSHIP (**ITML**)
12. UNIVERSITY OF NOVI SAD FACULTY OF SCIENCES SERBIA (**UNSPMF**)
13. TELEFONICA INVESTIGACION Y DESARROLLO SA (**TID**)

I-BiDaaS
Industrial-Driven Big Data as a Self-Service Solution

Key messages

A **complete** and **safe environment** for methodological **big data experimentation**

Tool and services to **increase the quality** of data analytics

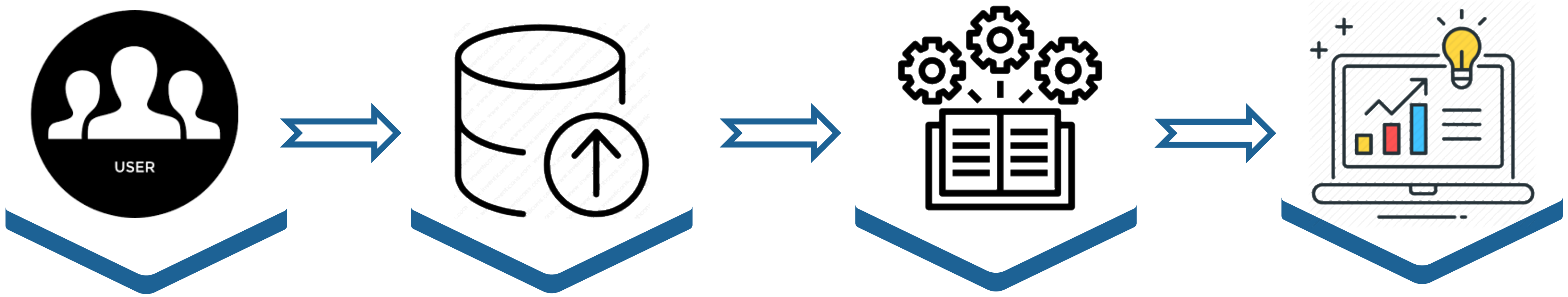
A Big Data as a **Self-Service solution** that helps in **breaking silos** and boosts EU's data-driven economy

Tools and services for **fast ingestion and consolidation** of both realistic and fabricated data

Increases impact in research community and contributes to industrial innovation capacity

Tools and services for the management of **heterogeneous infrastructures**

I-BiDaaS pipeline



Users

- Expert mode
- Self-service mode
- Co-develop mode

Data

- Import your data
- Fabricate Data
- Tokenize data

Analyze your Data

- Stream & Batch Analytics
- Expert: Upload your code
- Self-service: Select an algorithm from the pool
- Co-develop: custom end-to-end application

Results

- Visualize the results
- Share models

Benefits of using I-BiDaaS

Do it yourself In a flexible manner

Break data silos

Safe environment

Interact with Big Data technologies

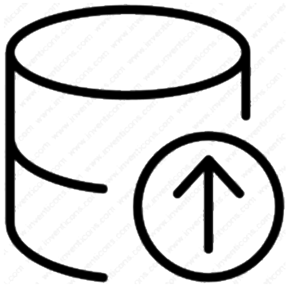
Increase speed of data analysis

Intra- and inter-domain data-flow

Cope with the rate of data asset growth

I-BiDaaS pipeline

Industrial-Driven Big Data as a Self-Service Solution



Users

- Expert mode
- Self-service mode
- Co-develop mode

Data

- Import your data

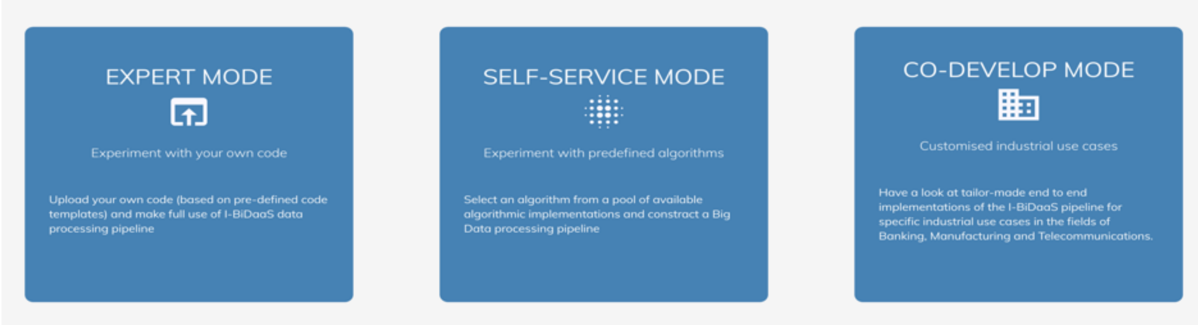
Analyze your Data

- Stream & Batch Analytics
- Expert: Upload your code

Results

- Visualize the results
- Share models

Flexible solution



Benefits of using I-BiDaaS

Do it yourself
In a flexible manner

Break data silos

Safe environment

Interact with Big Data technologies

Increase speed of data analysis

Intra- and inter-domain data-flow

Cope with the rate of data asset growth

I-BiDaaS pipeline

I-BiDaaS
Industrial-Driven Big Data as a Self-Service Solution

Users

- Expert mode
- Self-service mode
- Co-develop mode

Data

- Import your data
- **Fabricate Data**
- **Tokenize data**

Analyze your Data

- Stream & Batch Analytics
- Expert: Upload your code
- Self-service: Select an
  e pool
  om end-
  n

Results

- Visualize the results
- Share models

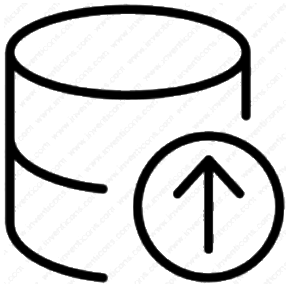
Data sharing & breaking silos

Benefits of using I-BiDaaS

- Do it yourself In a flexible manner
- Break data silos
- Safe environment
- Interact with Big Data technologies
- Increase speed of data analysis
- Intra- and inter-domain data-flow
- Cope with the rate of data asset growth

I-BiDaaS pipeline

I-BiDaaS
Industrial-Driven Big Data as a Self-Service Solution



Users

- Expert mode
- Self-service mode
- Co-develop mode

Data

- Import your data
- Fabricate Data
- Tokenize data

Analyze your Data

- <u>Stream & Batch Analytics</u>
- Expert: Upload your code
- Self-service: Select an algorithm from the pool
- Co-develop: custom end-to-end application

Results

- Visualize the results
- Share models

Benefits of using I-BiDaaS

Do it yourself In a flexible manner

Break data silos

Safe environment

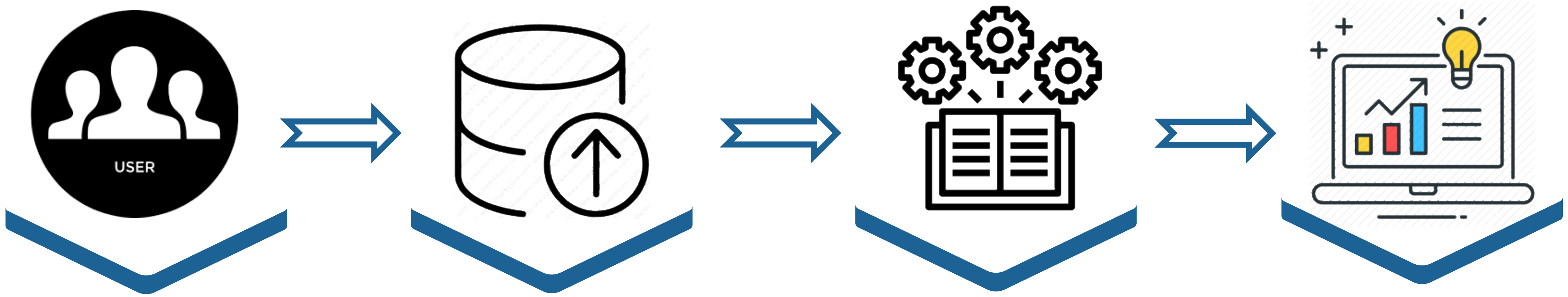
Interact with Big Data technologies

Increase speed of data analysis

Intra- and inter-domain data-flow

Cope with the rate of data asset growth

I-BiDaaS pipeline



Users

- Expert mode
- Self-service mode
- Co-develop mode

Data

- Import your data
- Fabricate Data
- Tokenize data

Analyze your Data

- Stream & Batch Analytics
- Expert: Upload your code
- Self-service: Select an algorithm from the pool
- Co-develop: custom end-to-end application

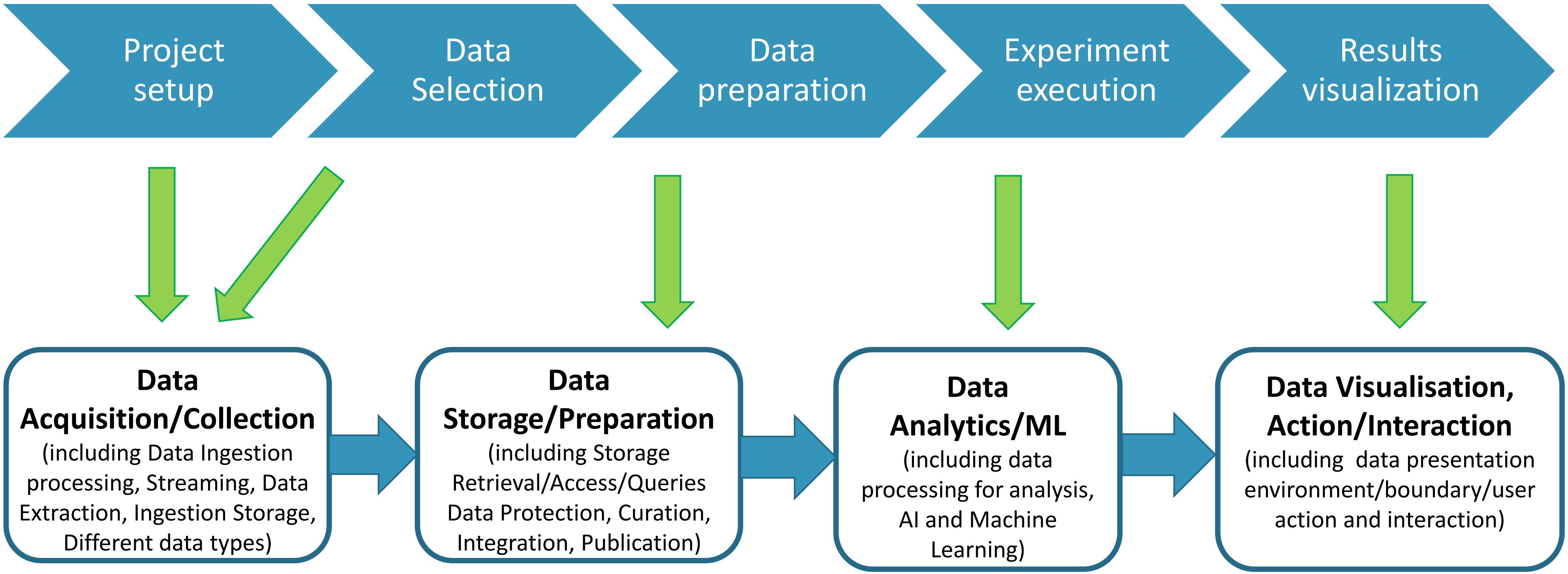
Results

- Visualize the results
- Share models

Benefits of using I-BiDaaS

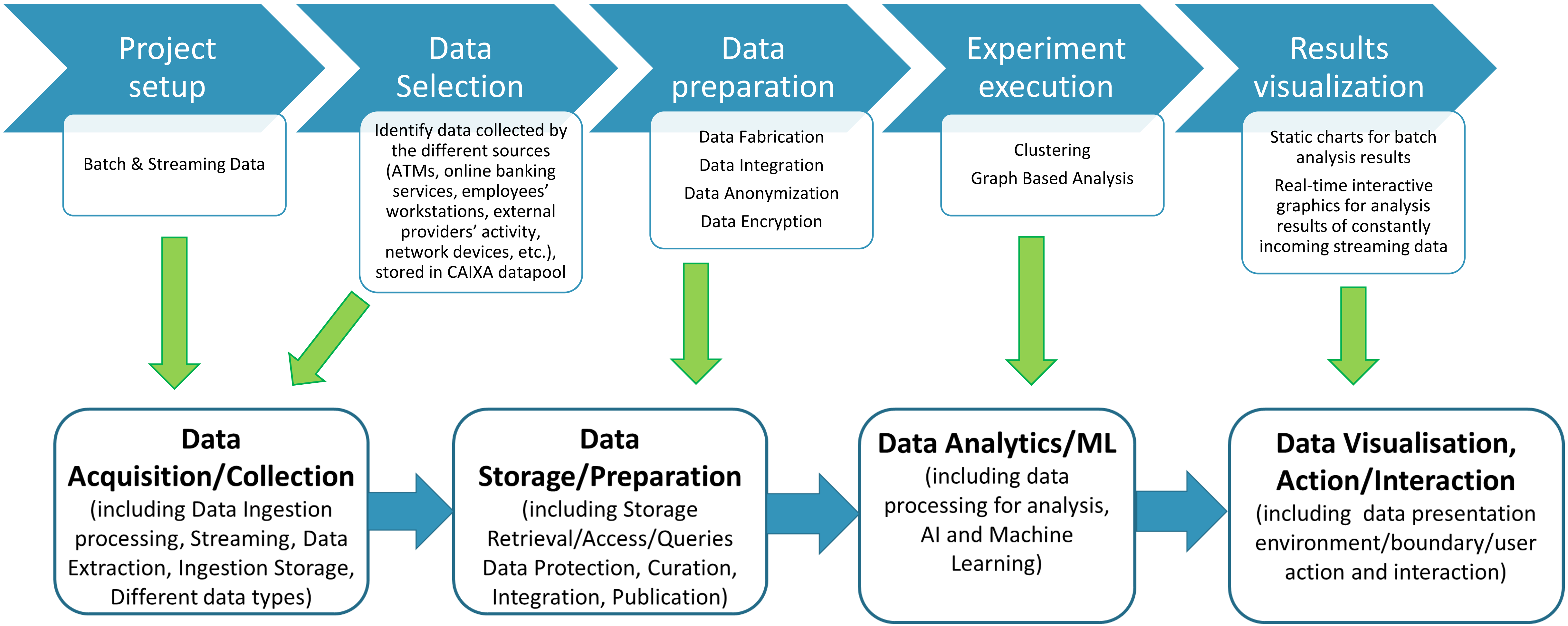
- Do it yourself In a flexible manner
- Break data silos
- Safe environment
- Interact with Big Data technologies
- Increase speed of data analysis
- Intra- and inter-domain data-flow
- Cope with the rate of data asset growth

I-BiDaaS Experimental Workflow



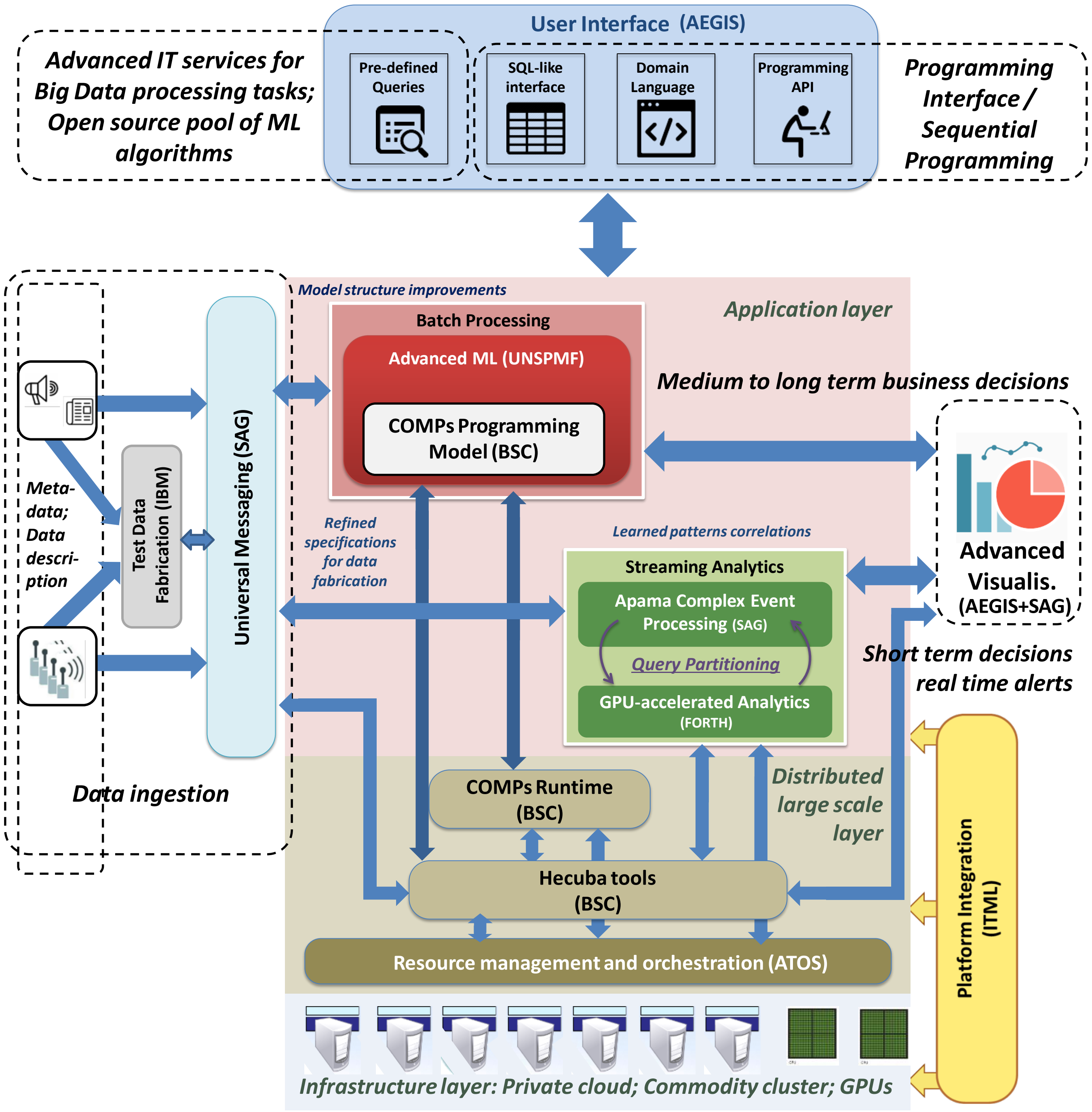
DataBench Pipeline

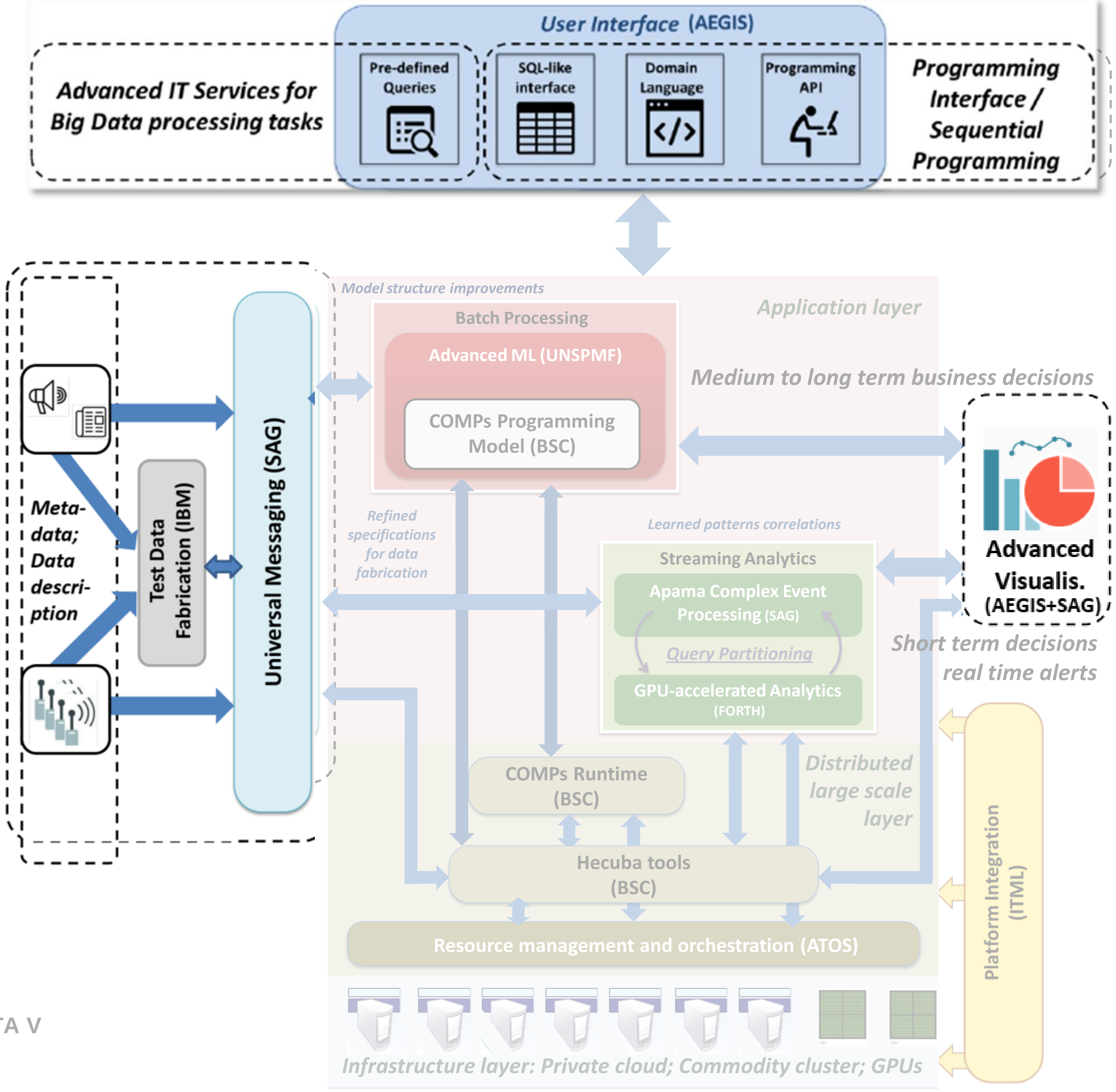
Example: Banking Experiments Workflow



DataBench Pipeline

The I-BiDaaS solution: Architecture & technologie





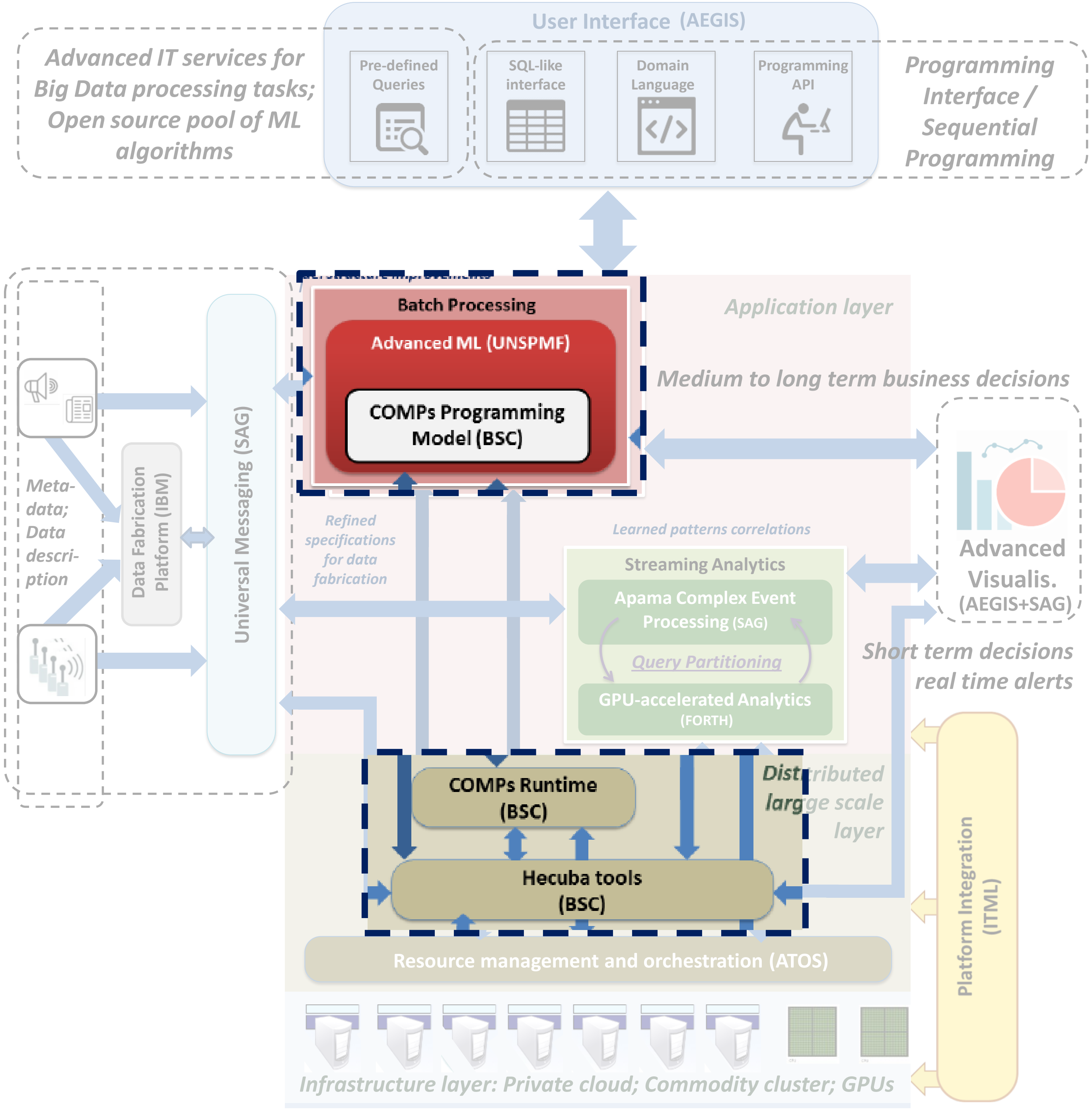
EUROPEAN BIG DATA V

WP2
Data, user interface, visualization

Technologies:

- *IBM TDF*
- *SAG UM*
- *AEGIS AVT*

http://ibidaas.eu/tools



EUROPEAN BIG DATA V

WP3
Batch Analytics

Technologies:

- *BSC COMPSs*
- *BSC Hecuba*
- *BSC Qbeast*
- *Advanced ML (UNSPMF)*

http://ibidaas.eu/tools

Benchmarking: Technology level

I-BiDaaS partner	Technology name	Platform role	DataBench pipeline	Current benchmarks
FORTH	GPU accelerator technology	Data pre-processing, Streaming Analytics	Step 3	Custom benchmark (throughput, latency)
BSC	COMPSs	Sequential programming model for distributed architectures	Step 3	Applications (Own use cases)
BSC	Hecuba	Data management framework with easy interface	Step 2	Applications (Own use cases)
BSC	**Qbeast**	**Multidimensional indexing and storage**	**Step 2**	**TPC-H**
IBM	Test Data Fabrication	Synthetic test data fabrication	Step 1	Several open source + commercial products (e.g., Grid tools of CA) / No known benchmarks yet
SAG	Apama Streaming Analytics Platform	Streaming Analytics	Step 3	Custom benchmark (throughput)
SAG	Universal Messaging	Message Broker	Step 1	Custom benchmark (throughput)
AEGIS	Advanced visualization and monitoring	Visualization and interface	Step 4	N/A
UNSPMF	**Pool of ML algorithms in COMPSs/Python**	**Batch analytics**	**Step 3**	**Respective MPI implementation; Sklearn; HiBench**
ATOS	Resource management and orchestration module	Resource management	Step 2	N/A

Benchmarking: Business, data & analytics

	Business Objectives	Data Sets	Data Size	Processing Type	Type of Analysis
Telecoms	– improve and optimize current operations	– Anonymized mobility data (structured) – Anonymized call center data (unstructured)	TB	– batch & streaming	– predictive – descriptive / diagnostic
Finance	– improve decision making – improve efficiency of Big Data solutions	– Tokenized online banking control data (structured) – Tokenized bank transfer data (structured) – Tokenized IP address data (structured)	PB	– batch – batch & streaming	– descriptive / diagnostic
Manufacturing	– improve and optimise current operations – improve the quality of the process and product	– Anonymized SCADA/MES data (structured) – Anonymized Aluminum Die-casting (structured)	GB	– batch – batch & streaming	– predictive – diagnostic

Benchmarking: Business level

<table>
<tr><th>I-BiDaaS Partner</th><th>Use Case</th><th>Most relevant business KPIs</th></tr>
<tr><td>TID</td><td>Accurate location prediction with high traffic and visibility</td><td rowspan="3">- Acquisition of insights on the dynamics of cellular sectors
- Processing costs (cost reduction)
- Customer satisfaction</td></tr>
<tr><td>TID</td><td>Optimization of placement of telecommunication equipment</td></tr>
<tr><td>TID</td><td>Quality of service in Call Centers</td></tr>
<tr><td>CAIXA</td><td>Enhanced control on online banking</td><td rowspan="3">- Cost reduction
- Data accessibility
- Time efficiency
- End-to-end execution time (from data request to data provision)</td></tr>
<tr><td>CAIXA</td><td>Advanced analysis of bank transfer payment in financial terminal</td></tr>
<tr><td>CAIXA</td><td>Analysis of relationships through IP addresses</td></tr>
<tr><td>CRF</td><td>Production process of aluminium die-casting</td><td rowspan="2">- 3 Product quality levels (High, Medium, Low)
- Overall Equipment Effectiveness (OEE),
- Maintenance cost
- Cost reduction</td></tr>
<tr><td>CRF</td><td>Maintenance and monitoring of production assets</td></tr>
</table>

Thank you!

Questions?

Enabling procurement data value chains for economic development, demand management, competitive markets and vendor intelligence

Brian Elvesæter, SINTEF

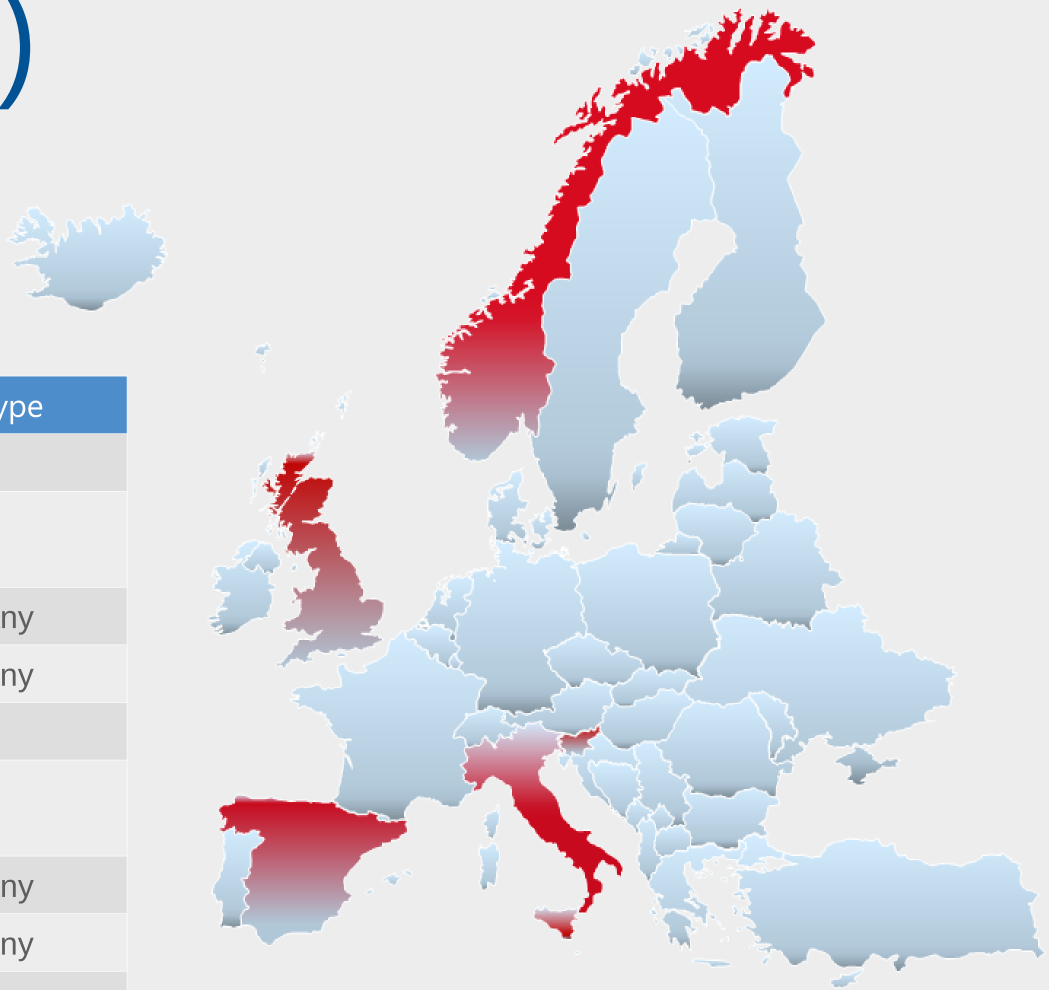
brian.elvesater@sintef.no

https://theybuyforyou.eu

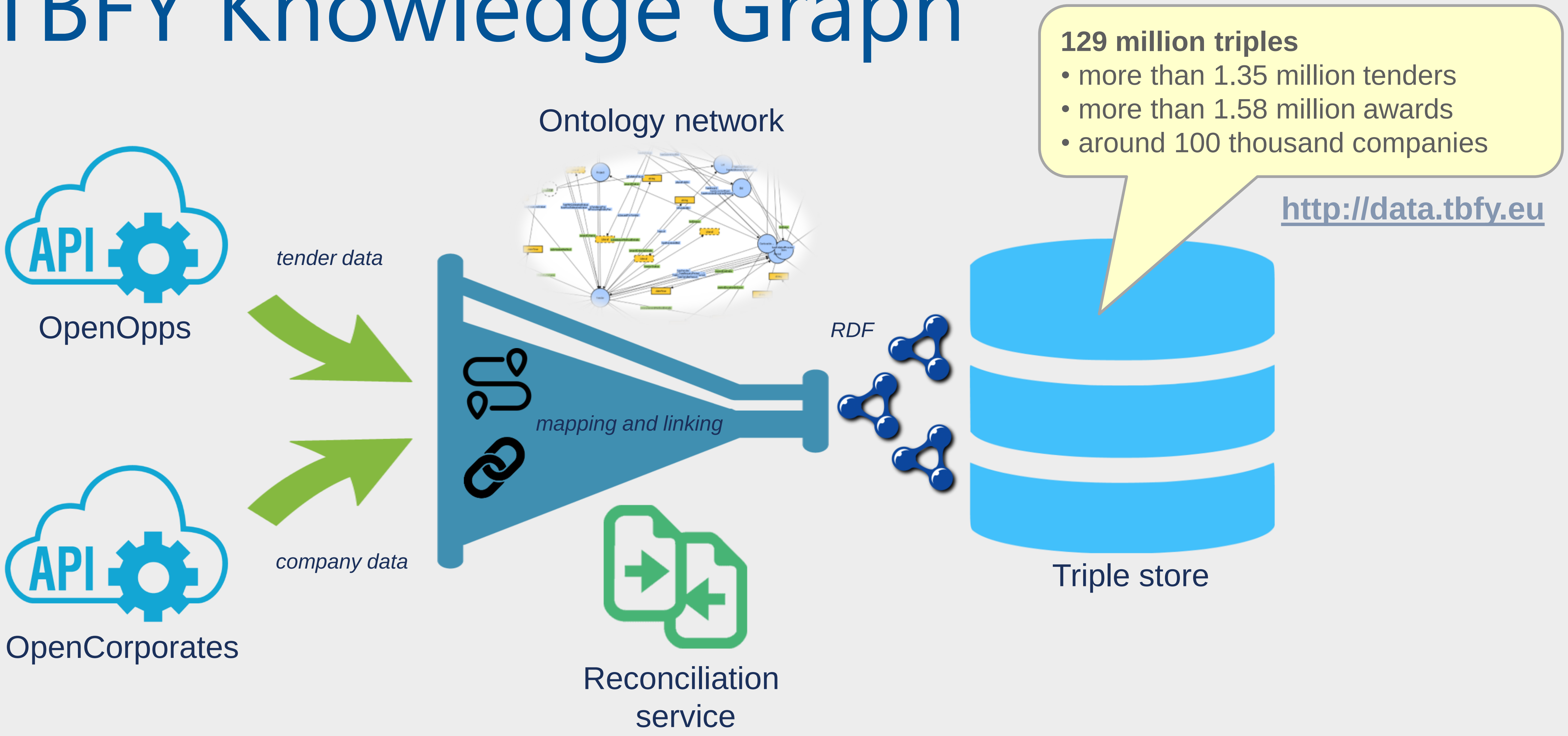
TheyBuyForYou (TBFY)

- Horizon 2020 Innovation Action
  - Grant agreement No 780247
- Duration: 36 months
  - Jan 2018 – Dec 2020
- Overall budget
  - € 3 274 440
- Developing Big Data tools for Public Procurement
  - data access
  - data analytics
  - data interaction
  - data visualization
- 10 Partners from 5 countries

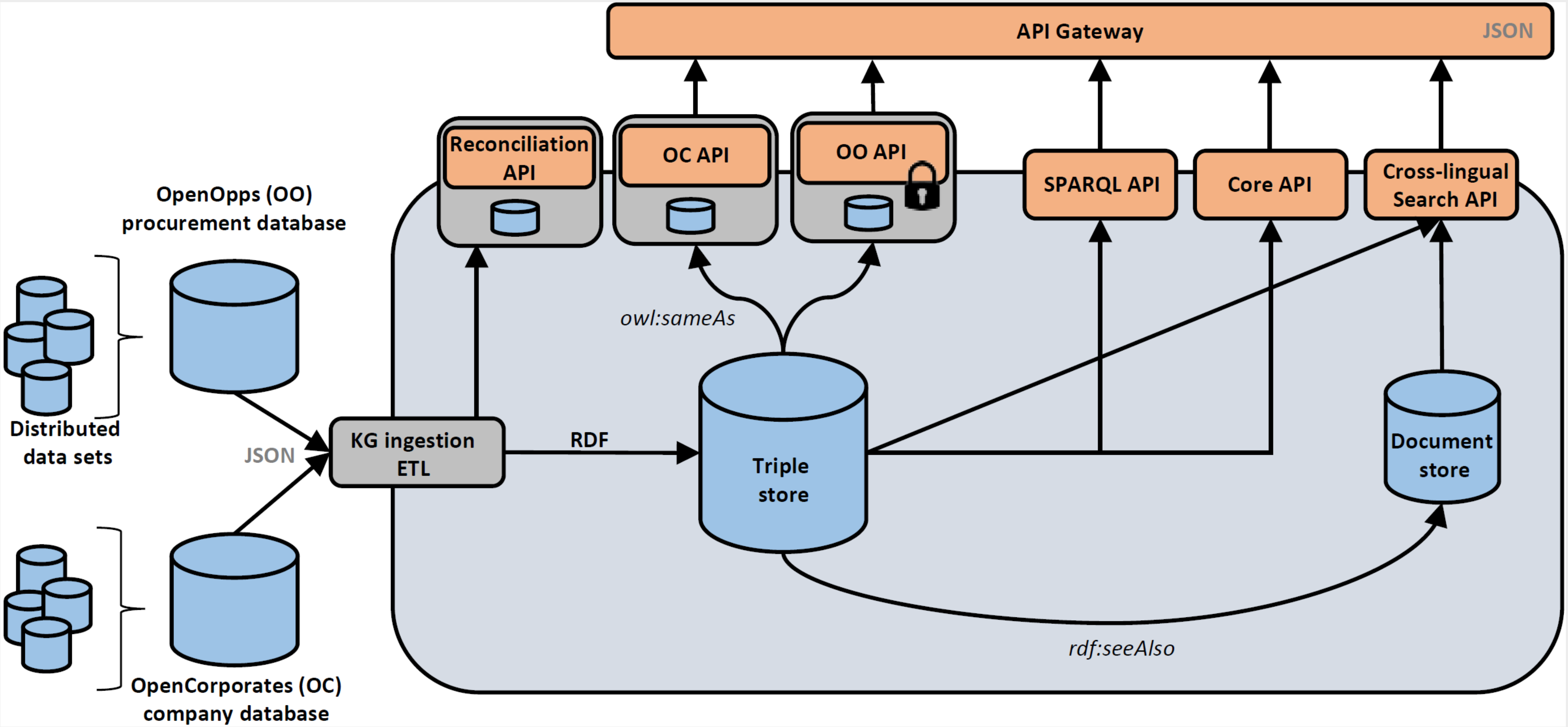
Partner	Country	Organisation Type
SINTEF AS	NO	RTO
Municipality of Zaragoza	ES	Public sector
CERVED	IT	Private Company
OpenCorporates	UK	Private Company
Josef Stefan Institute	SI	RTO
Ministry of Public Sector Innovation	SI	Public Sector
OESÍA Networks	ES	Private Company
OpenOpps	UK	Private Company
Universidad Politécnica de Madrid	ES	University
King's College London	UK	University



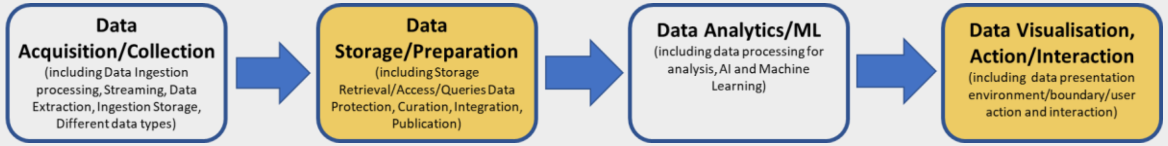
TBFY Knowledge Graph



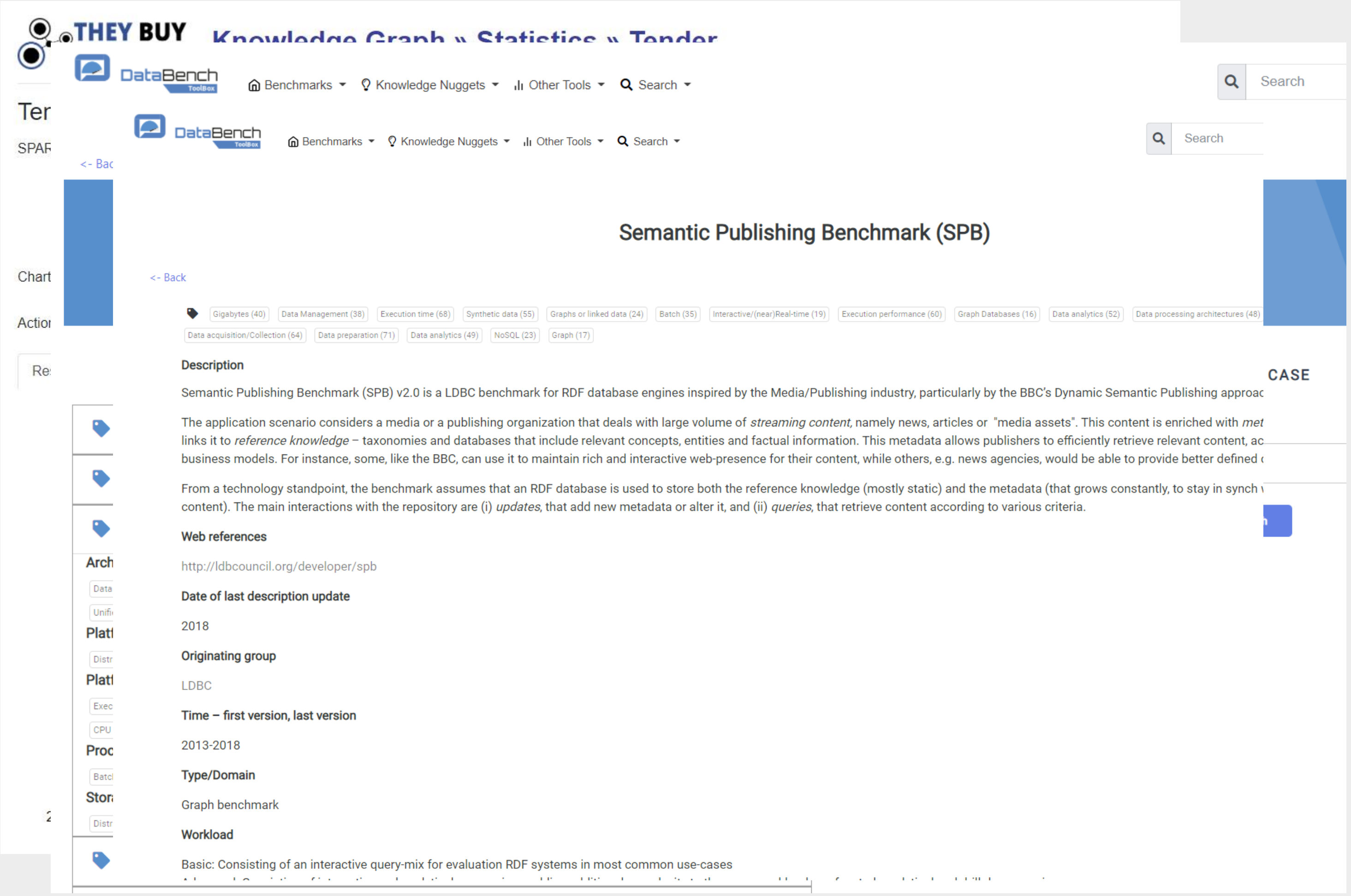
TBFY Platform



Benchmark case #2:
Query/visualization



- SPARQL Query Performance
  - KG API (RESTful API wrapping pre-defined SPARQL queries)
  - Visualization
- Benchmark questions?
  - Which triple store (RDF) database to choose?
  - Which cloud computing (resource) plan to choose?
    - https://aws.amazon.com/pricing/
  - Database stability issues
    - JVM heap size
- DataBench ToolBox
  - LDBC Semantic Publishing Benchmark (SPB) to benchmark RDF triple stores
    - GraphDB
    - Apache Jena Fuseki & TDB
    - ...



Questions?

Brian Elvesæter, SINTEF

brian.elvesater@sintef.no

https://theybuyforyou.eu

Vodafone Innovus
Fleet Management solutions

Athanasios Koumparos
Vodafone Innovus

European Big Data Value Forum, 2020.11.04

Track & Know

Getting to know Vodafone Innovus

Who we are

a 100% Vodafone company located in Athens, Greece, est. 2004

Operating in 7 Vodafone OpCos

What we do

We design and implement innovative IoT solutions based on VF **Group strategy & customer needs**

What we deliver

We have built Vodafone **Group products** as well as the **Global Device Management solution**

Our advantages

- ✓ World class Platform
- ✓ Global & local expertise
- ✓ Operations experience
- ✓ Agile way of working

World class capabilities build a world class eco-system

Enterprise Fleet | Overview

Logistics Vertical

Soho / SME / Corporate segments

Developed by VF Innovus

Refreshed platform, GDSP connected

A well-established solution, applicable to customers with
advanced needs for location tracking, sensor monitoring & in-depth reports / analytics !

Markets deployed

VF Greece & VF Albania

>13k active assets
>6k assets in cold chain logistics

11 years of experience

providing an excellent solution since 2009

A fresh, award winning approach

new platform with re-designed UI, advanced Driver Safety & enhanced capabilities

MOBILITY AWARDS 2019 GOLD

Innovation

We participate in EU programs (Track & Know, Trustonomy) and collaborate with the Hellenic Organization of Intelligent Transport Systems to enhance the solution capabilities

VFI IoT Platform

Enterprise Fleet

API, GUI

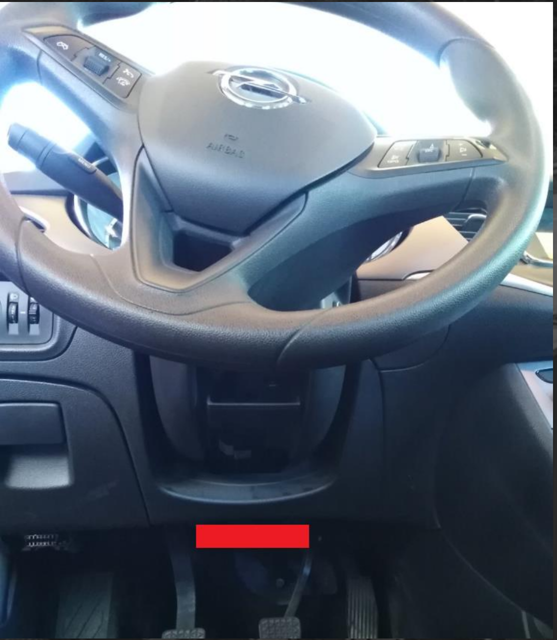
Analytics – Offline Processing

Online Data Processing

Low Level Device – Platform Communication

GSM Network

Devices on vehicles



Device in vehicles

- Devices are installed under the hood (or bonnet)
- Vehicles vary in type, age and electronics

Data

- More than 3M data packets received daily by the platform
- GPS data (coordinates, angle, speed, date)

Enterprise Fleet++

Challenges

Reception problems

- Low signal strength due to installation (under the bonnet)
- Urban areas (bridges, tall buildings etc.)
- GPS signal reflections

Vehicle

- Poor quality of power supply
- Vehicle aging – noisy electricals

Issues caused by problems

- Incorrect display position on map

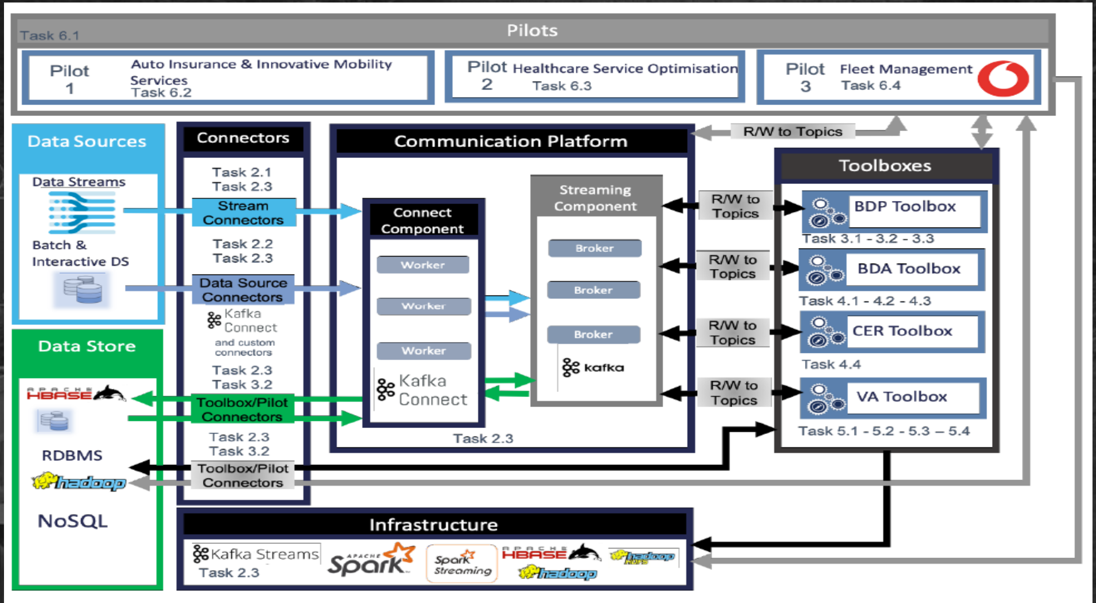
(on top of buildings, fields, wrong road)

- False positive alerts (e.g. speed violation)
- Location gaps during route display

Track & Know

Enterprise Fleet++

- Big Data Analytics
- ML based components
- Scalable
- Online Processing
- Offline Processing
- Versatile
- Easy to integrate (Kafka Topics)



Data Collection / Collection

Live data streaming from devices (GPS/Sensors)

Data Storage / Preparation

Cleansing, enrichment, storage in NoSQL stores

Data Analytics / ML

Pattern recognition
Location forecasting
Clustering
Mobility networks

Data Visualization Action / Interaction

Visual analytics
End user GUI

Integration VFI IoT Platform to Track & Know

Enterprise Fleet++

Online Processing Flows:

- Data cleansing
- Data enrichment
- Data pattern recognition
- Future location prediction
- Driver behavior analysis

Offline Processing Flows

- Individual mobility networks, predict next service period
- Hot spot analysis
- Trajectory matching
- Visualization

API, GUI

Analytics – Offline Processing

Online Track & Know

Online Processing

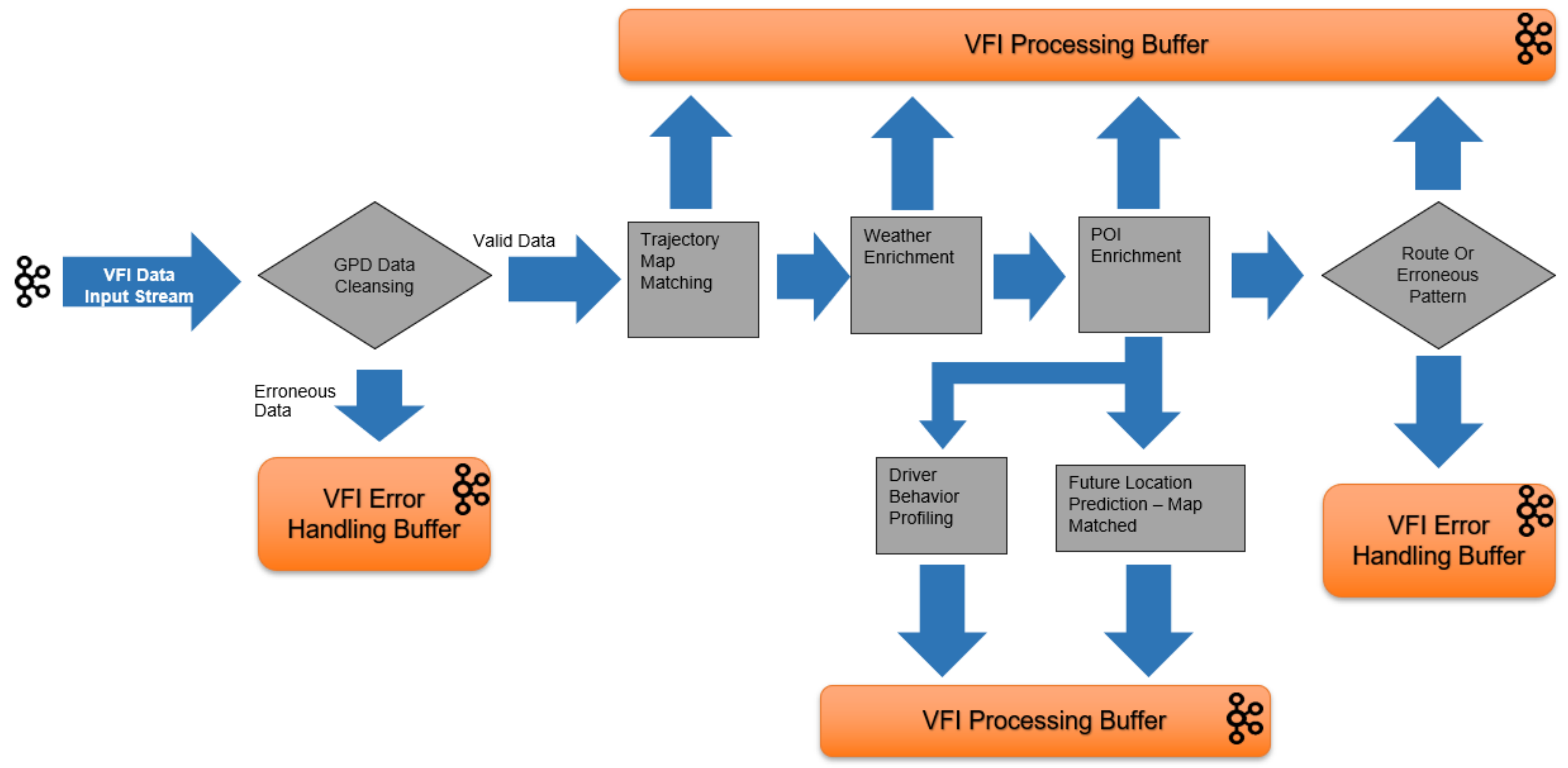
Offline Track & Know

Low Level Device – Platform Communication

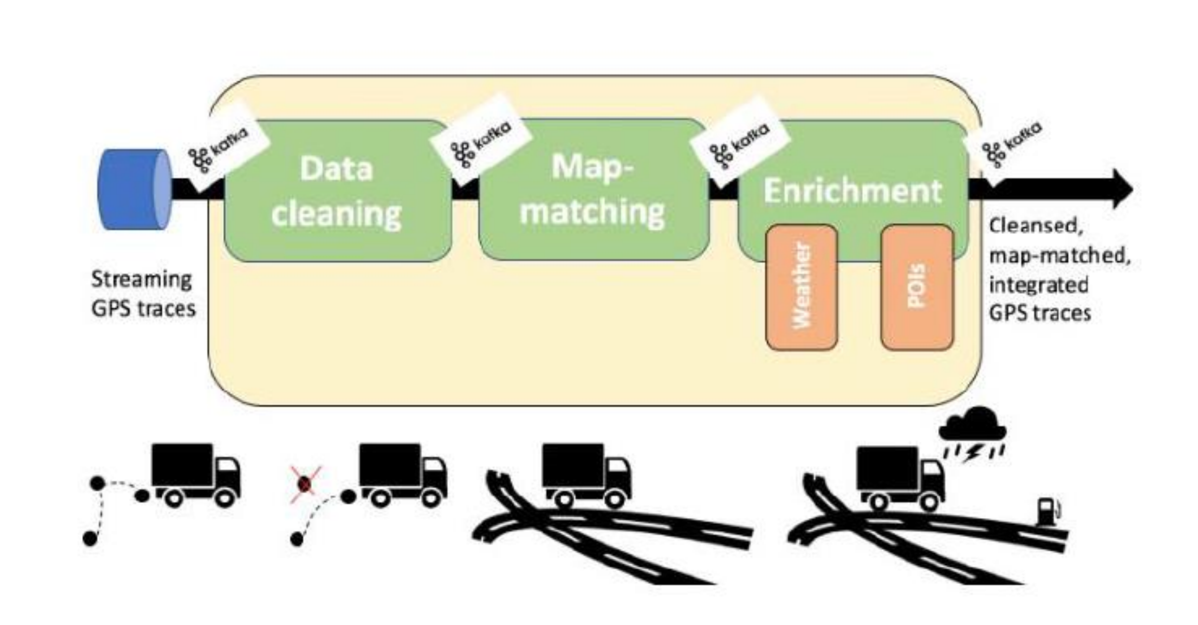
GSM Network

Data Flows T&K Platform

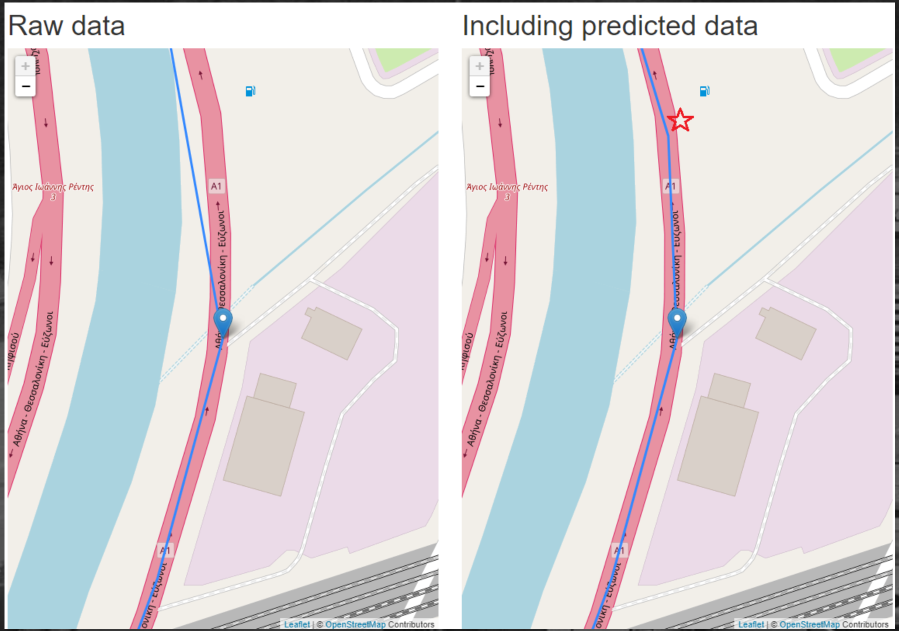
Enterprise Fleet++



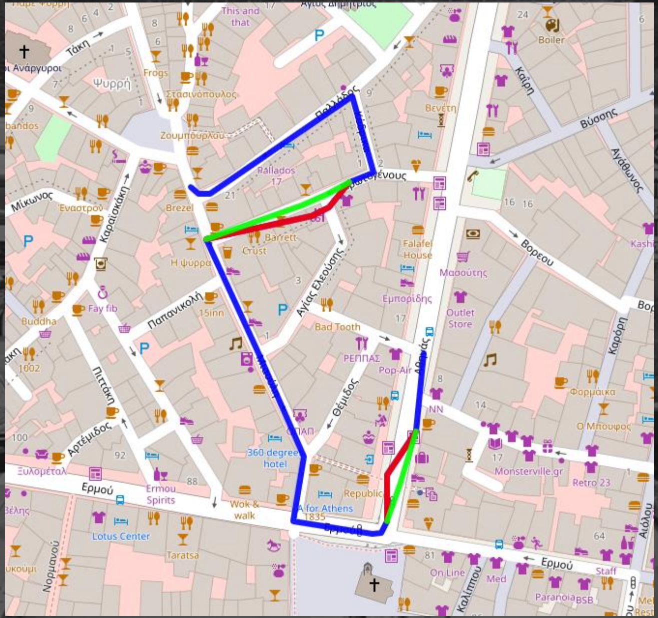
Data cleansing tool - Map matching tool



GPS post enhancements



- Vehicle moving on a highway
- Sampling rate at 30 seconds



- Vehicle moving in dense urban area
- GPS reception not always good

Thank you

Data-driven Bioeconomy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732064

This project is part of BDV PPP

DataBio pipelines

Prof. Dr. Caj Södergård,
Technical Manager of DataBio

VTT Technical Research Centre of Finland

European Big Data Forum 2020
DataBench session
4.11.2020

Data-driven Bioeconomy - DataBio

DataBio aim: Develop big data tools for enhancing production of raw materials for food, energy and biomaterials industries

- A EU-funded project with 48 partners
- 27 bioeconomy pilots in 17 countries
- Duration 2017-2020

Responsible and sustainable production of food, energy and biomaterials

Better raw material utilisation from agriculture, forestry and fishery sources

New business opportunities through market-ready big data technologies

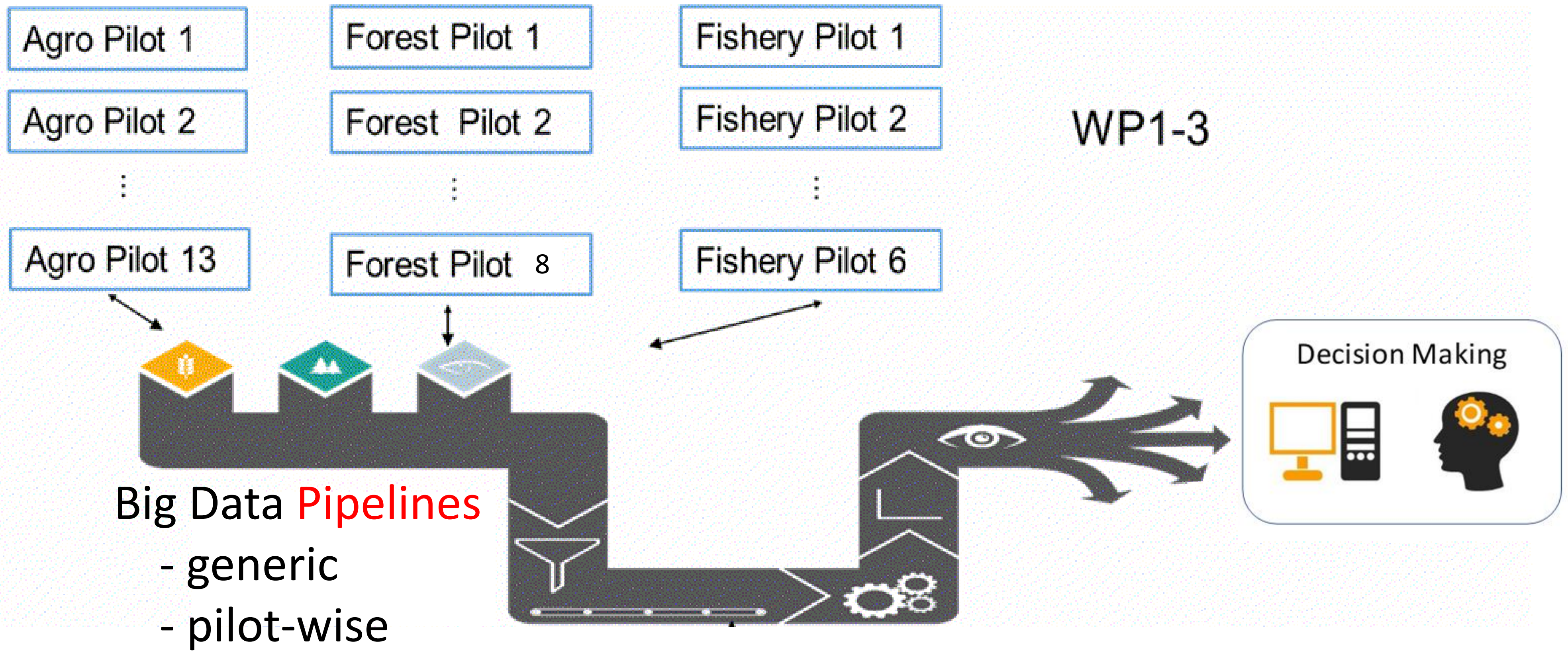
Motivation

Population growth and urbanisation are increasing the demand for natural resources, which is putting a strain on the Earth's carrying capacity.

We aim to develop new sustainable ways to use forest, farm and fishery resources and to communicate real-time information to decision-makers and producers.

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

DataBio platform serves the 27 pilots

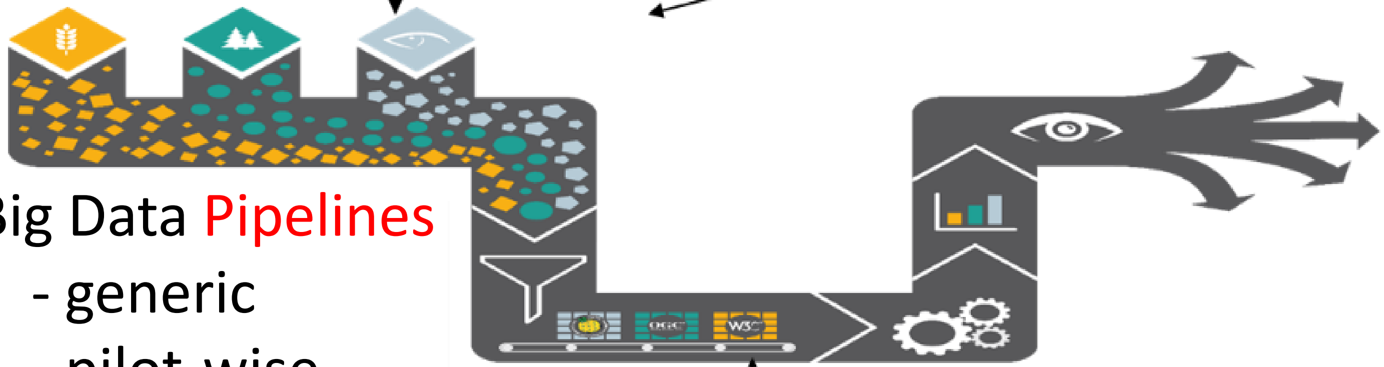


DataBio platform serves the 27 pilots

Agro Pilot 1
Agro Pilot 2
⋮
Agro Pilot 13

Forest Pilot 1
Forest Pilot 2
⋮
Forest Pilot 8

Fishery Pilot 1
Fishery Pilot 2
⋮
Fishery Pilot 6



Big Data Pipelines
- generic
- pilot-wise

DataBio platform
with big data components
and datasets

WP4 Components & IoT datasets

This document is part of a project that has received funding
from the European Union's Horizon 2020 research and innovation programme
under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or
reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

DataBio platform serves the 27 pilots

Agro Pilot 1

Agro Pilot 2

⋮

Agro Pilot 13

Forest Pilot 1

Forest Pilot 2

⋮

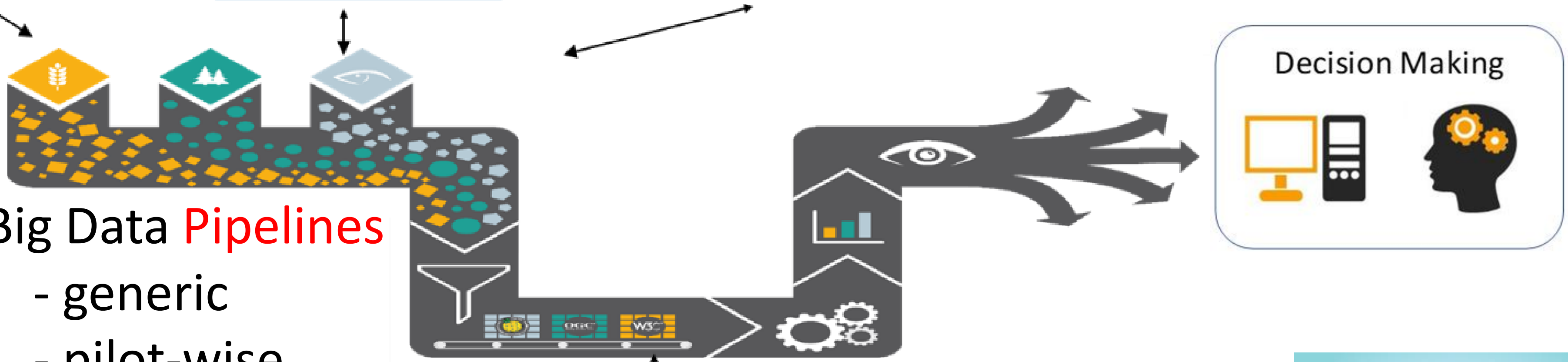
Forest Pilot 8

Fishery Pilot 1

Fishery Pilot 2

⋮

Fishery Pilot 6



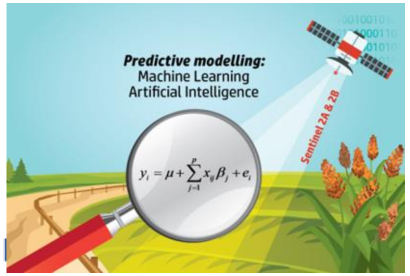
Big Data Pipelines

- generic
- pilot-wise

DataBio platform
with big data components
and datasets

WP4 Components & IoT datasets

WP5 Components & Earth Observation datasets



This document is part of a project that has received funding
from the European Union's Horizon 2020 research and innovation programme
under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or
reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

Pipelines vs. Services

Pipeline

- A chain of real-time processing components:
  - Acqusition/Collection,
  - Preparation
  - Analytics
  - Visualisation, User interaction
- Clear interfaces between components and to outside
- A "*white box*" showing internal wiring for developers

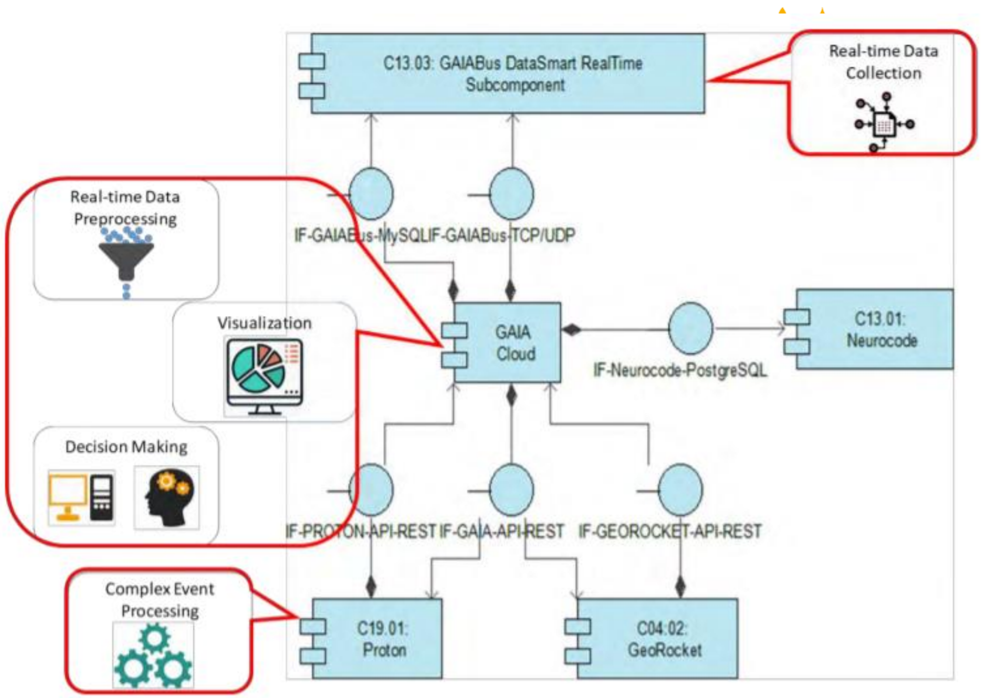
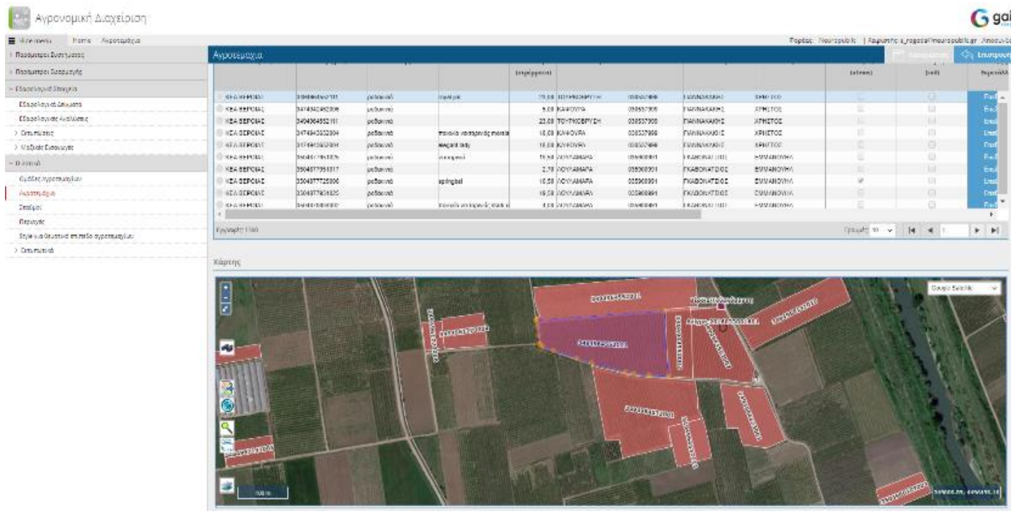


Figure 17: Mapping of generic components into pilot A1.1 component view

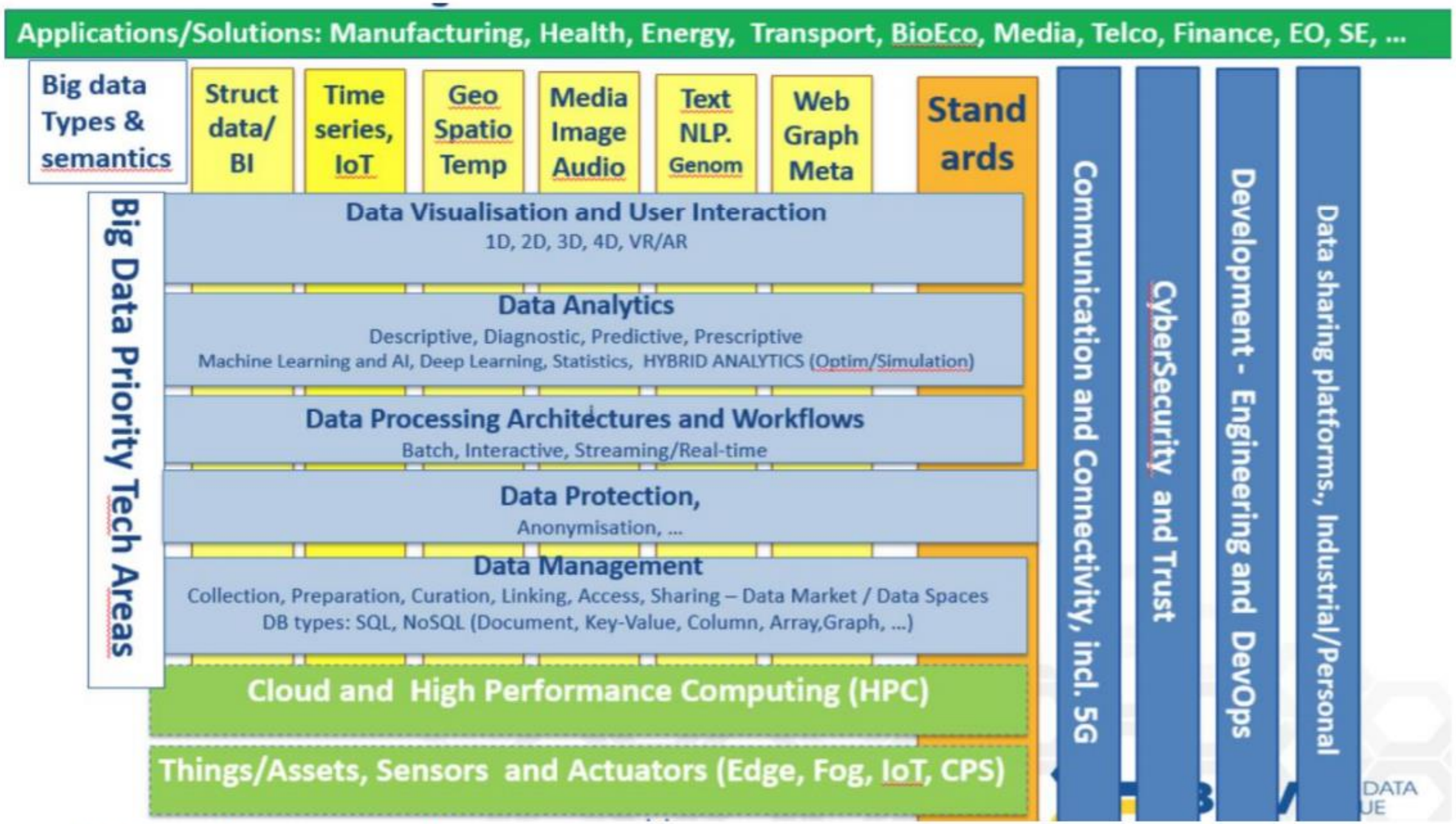
Service

- Provides usability to end users
- No display of internal wirings of components
- Accessed through API:s (web services, remote calls)
- Activated remotely through database queries ("end points") and executed in the cloud.
- Represents a "*black box*".



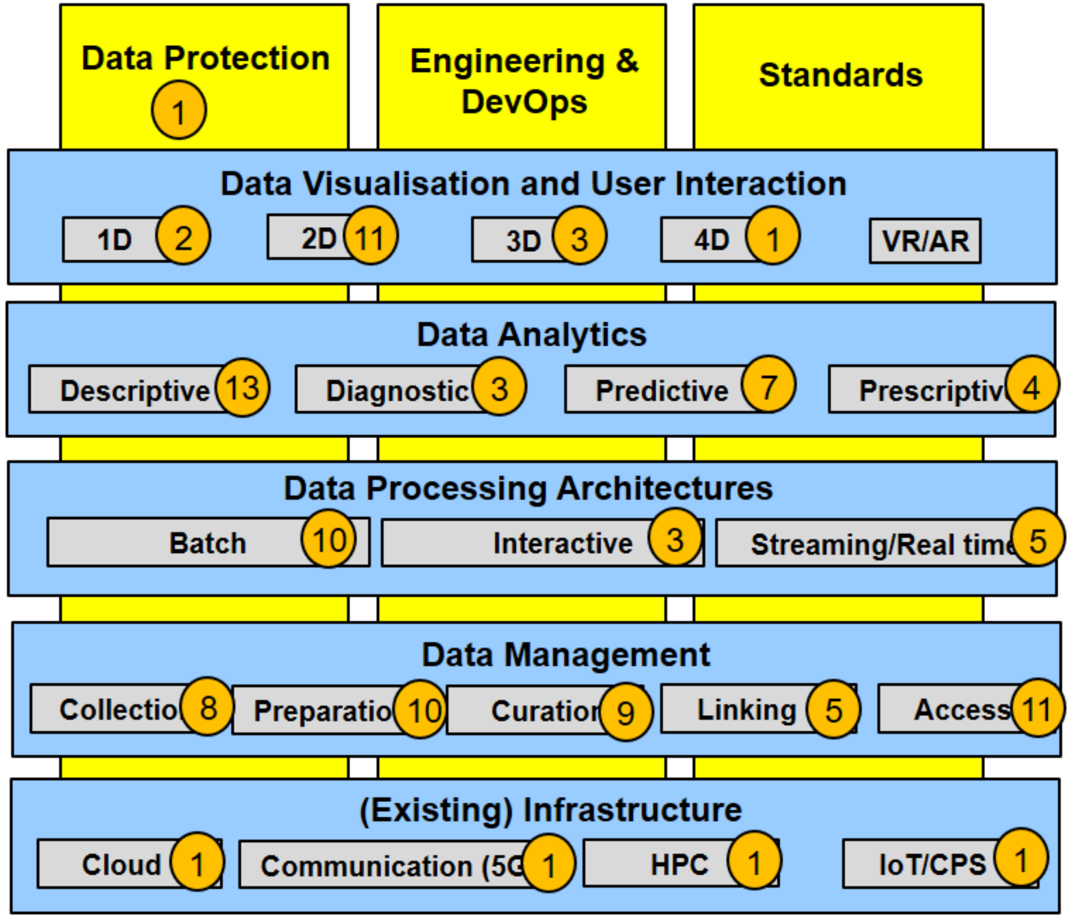
This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

We used the BDVA Reference Model



This document is part of a project that has received funding
from the European Union's Horizon 2020 research and innovation programme
under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or
reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

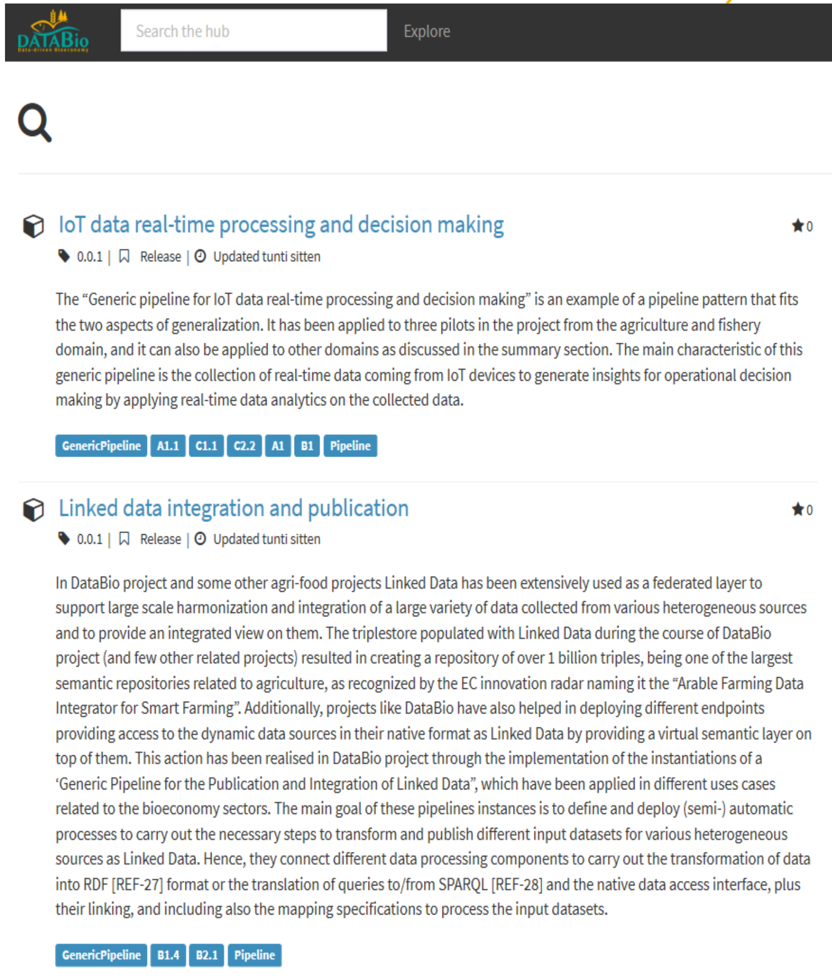
Platform development in DataBio in numbers



- 62 components from 28 partners in two trial rounds
- 1-6 components per pilot (average 2)
- 14 new **user interfaces**
- 59 new **APIs**
- +2,7 in **Technology Readiness Level** (1-9)

Generic pipelines in DataBio

Generic pipeline	Applied in pilots
Earth Observation and Geospatial data processing	**A1.1** Precision agri **B1.2** Precision agri , **C1.1** Insurance, **C2.2** CAP Support Fishery: **A1** Tuna fishing operations , **B1** Tuna fishing planning,
IoT data real-time processing and decision making	Agri: **A1.1** Precision agri, **B1.1** Cereals & biomass crops Fishery: **A1**
Linked Data Integration and publication	Agri: **B1.4** Precision agri **B2.1** Machinery mgmt, **Other** open farm and geospatial datasets Fishery: **Virtual pilot**
Privacy-aware analytics	Fishery (Norway): **Virtual pilot**
Genomics	Agri: **A2.1** Greenhouse , **B1.3** Biomass sorghum
Forestry Data management	Forestry (Finland): **2.2.1** Data Sharing , **2.2.2** Monitoring, **2.4.2** Shared Multiuser environment
Fisheries decision support in catch planning	Fishery: **B2, C1, C2, Virtual pilot**

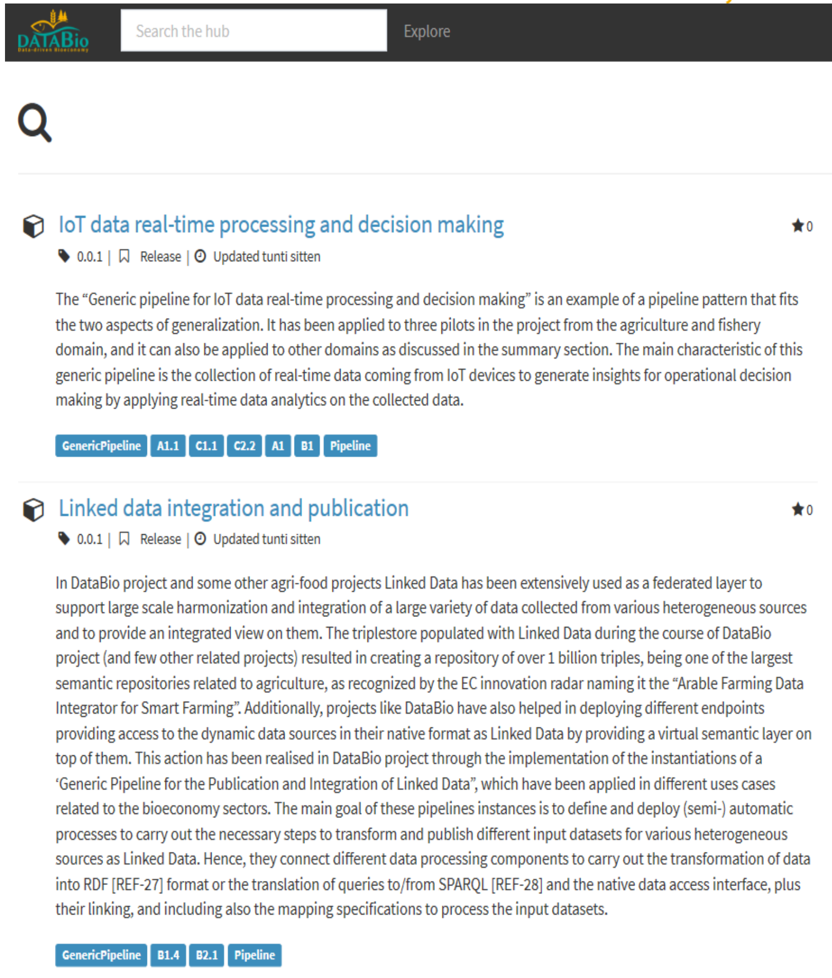


Source: Deliverable D4.4 Service documentation

VTT

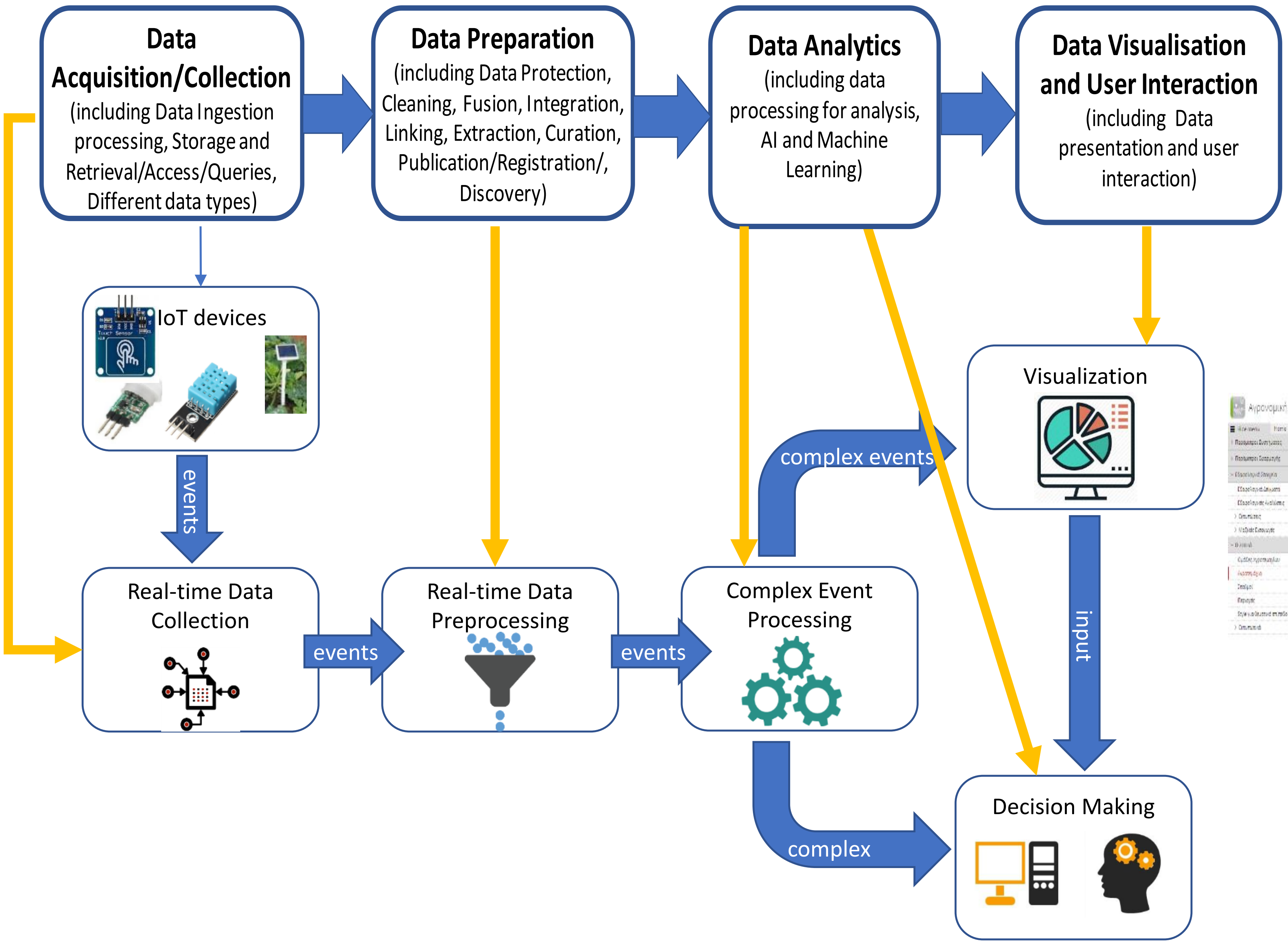
Generic pipelines in DataBio

Generic pipeline	Applied in pilots
Earth Observation and Geospatial data processing	**A1.1** Precision agri **B1.2** Precision agri , **C1.1** Insurance, **C2.2** CAP Support Fishery: **A1** Tuna fishing operations , **B1** Tuna fishing planning,
IoT data real-time processing and decision making	Agri: **A1.1** Precision agri, **B1.1** Cereals & biomass crops Fishery: **A1**
Linked Data Integration and publication	Agri: **B1.4** Precision agri **B2.1** Machinery mgmt, **Other** open farm and geospatial datasets Fishery: **Virtual pilot**
Privacy-aware analytics	Fishery (Norway): **Virtual pilot**
Genomics	Agri: **A2.1** Greenhouse , **B1.3** Biomass sorghum
Forestry Data management	Forestry (Finland): **2.2.1** Data Sharing **, 2.2.2** Monitoring**, 2.4.2** Shared Multiuser environment
Fisheries decision support in catch planning	Fishery: **B2, C1, C2, Virtual pilot**

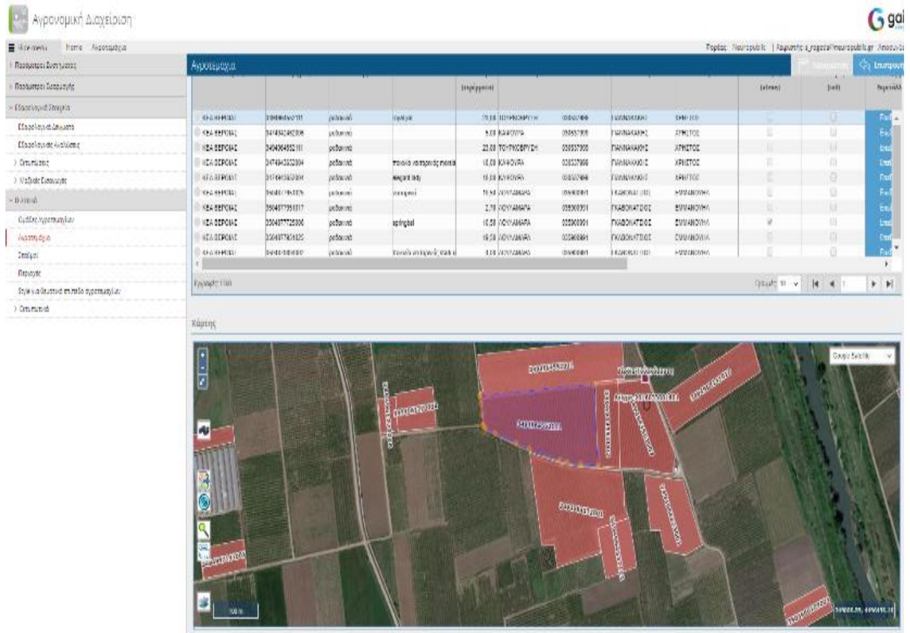


Source: Deliverable D4.4 Service documentation

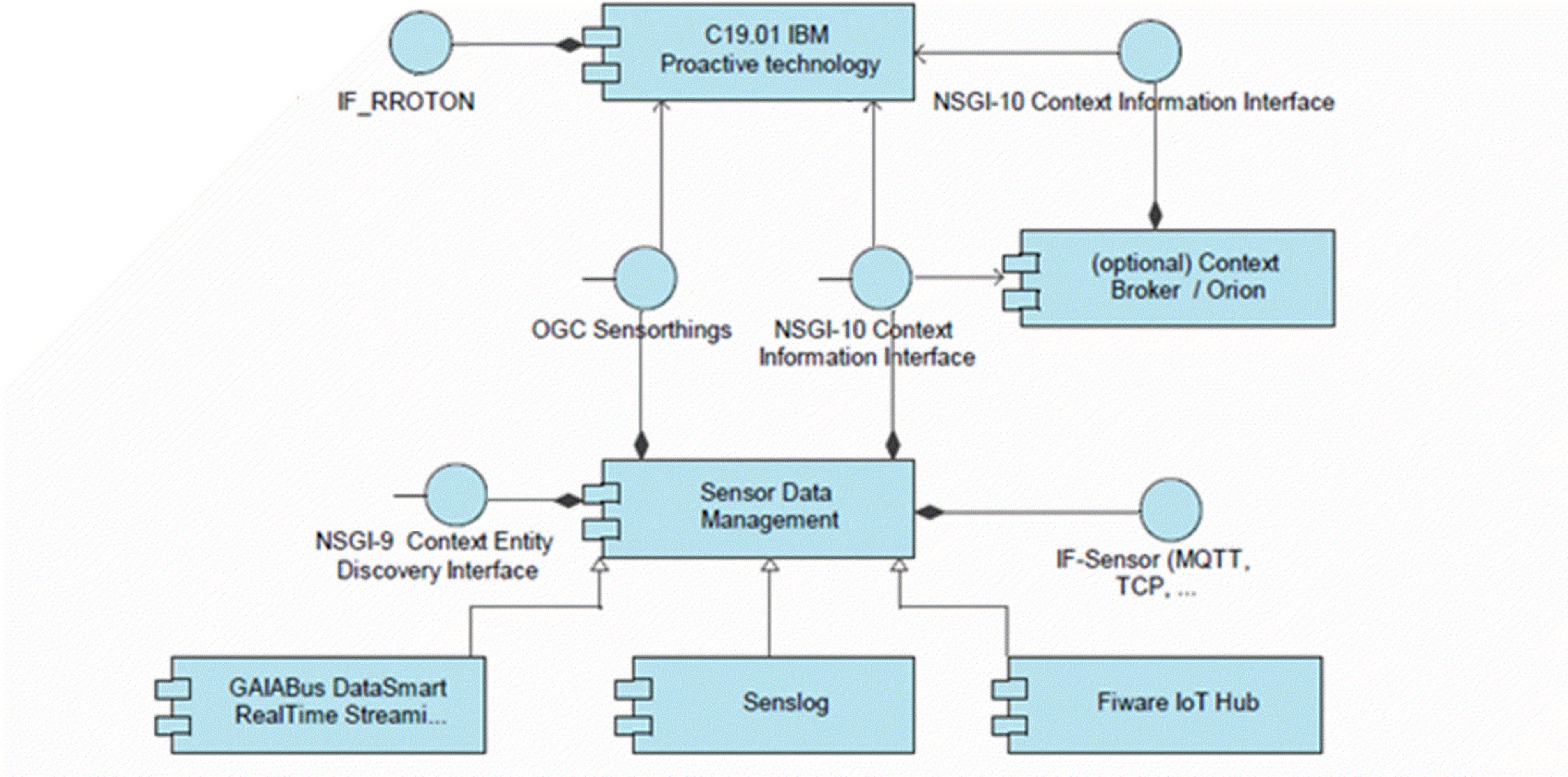
Generic Pipeline – IoT data realtime processing



Instance: Prediction and real-time alerts of diseases and pests breakouts in crops



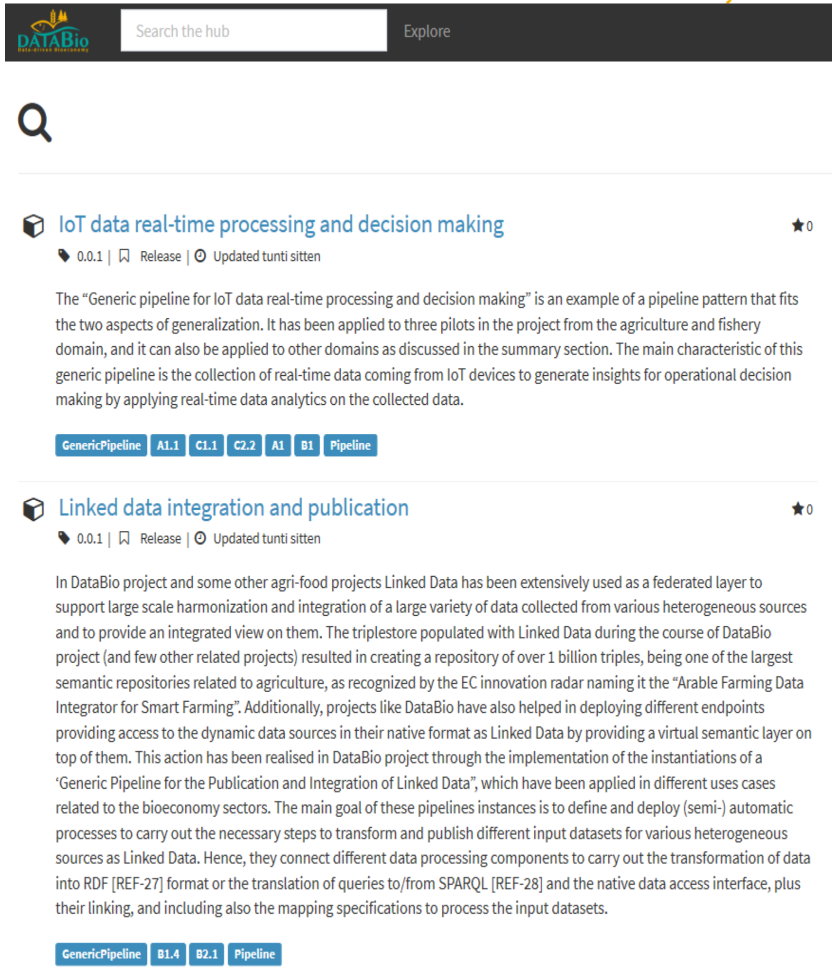
Crop monitoring using ”IoT Generic Pipeline



This document is part of a project that has received funding
from the European Union’s Horizon 2020 research and innovation programme
under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or
reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

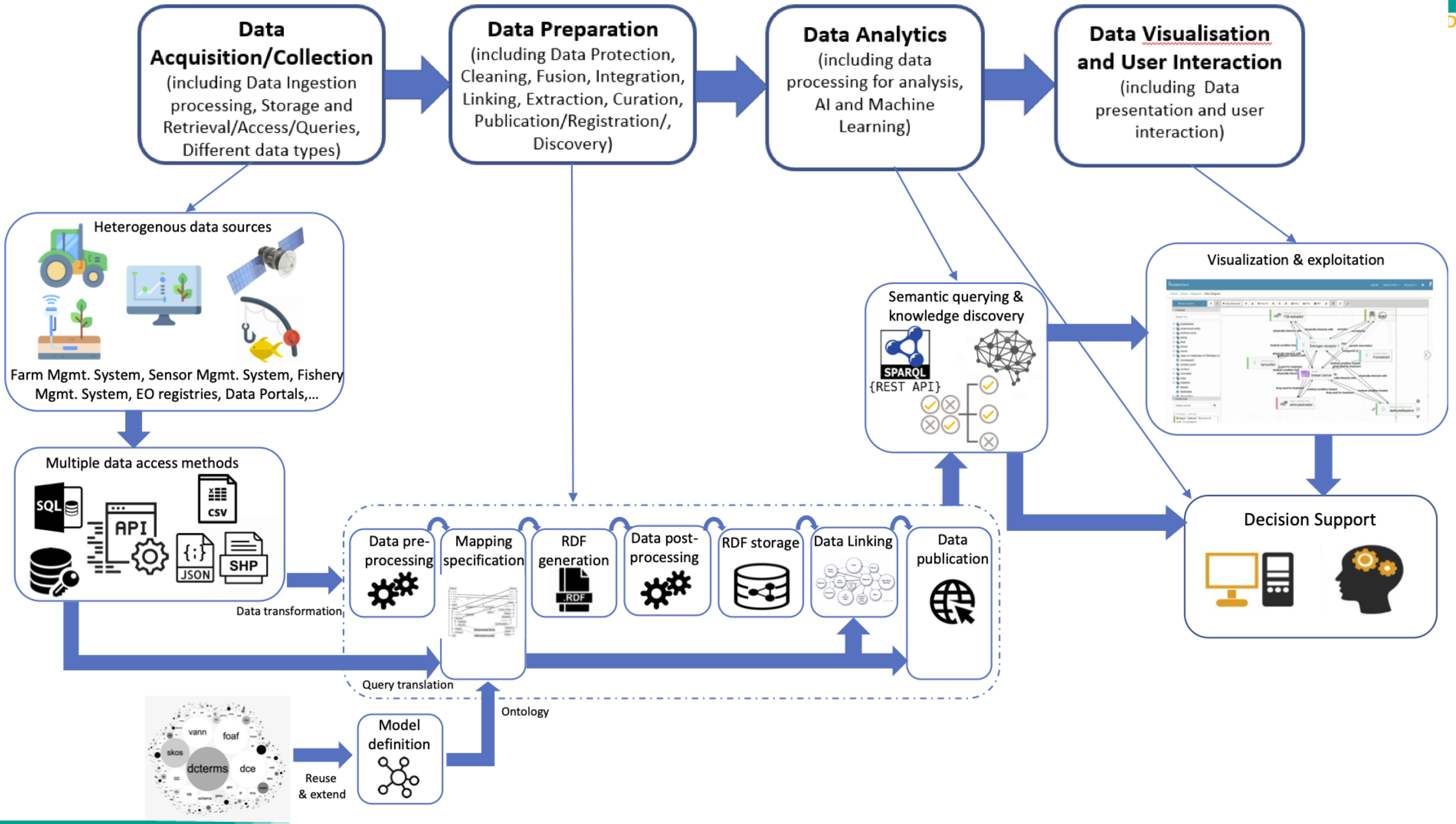
Generic pipelines in DataBio

Generic pipeline	Applied in pilots
Earth Observation and Geospatial data processing	**A1.1** Precision agri **B1.2** Precision agri , **C1.1** Insurance, **C2.2** CAP Support Fishery: **A1** Tuna fishing operations , **B1** Tuna fishing planning,
IoT data real-time processing and decision making	Agri: **A1.1** Precision agri, **B1.1** Cereals & biomass crops Fishery: **A1**
Linked Data Integration and publication	Agri: **B1.4** Precision agri **B2.1** Machinery mgmt, **Other** open farm and geospatial datasets Fishery: **Virtual pilot**
Privacy-aware analytics	Fishery (Norway): **Virtual pilot**
Genomics	Agri: **A2.1** Greenhouse , **B1.3** Biomass sorghum
Forestry Data management	Forestry (Finland): **2.2.1** Data Sharing **, 2.2.2** Monitoring**, 2.4.2** Shared Multiuser environment
Fisheries decision support in catch planning	Fishery: **B2, C1, C2, Virtual pilot**



Source: Deliverable D4.4 Service documentation

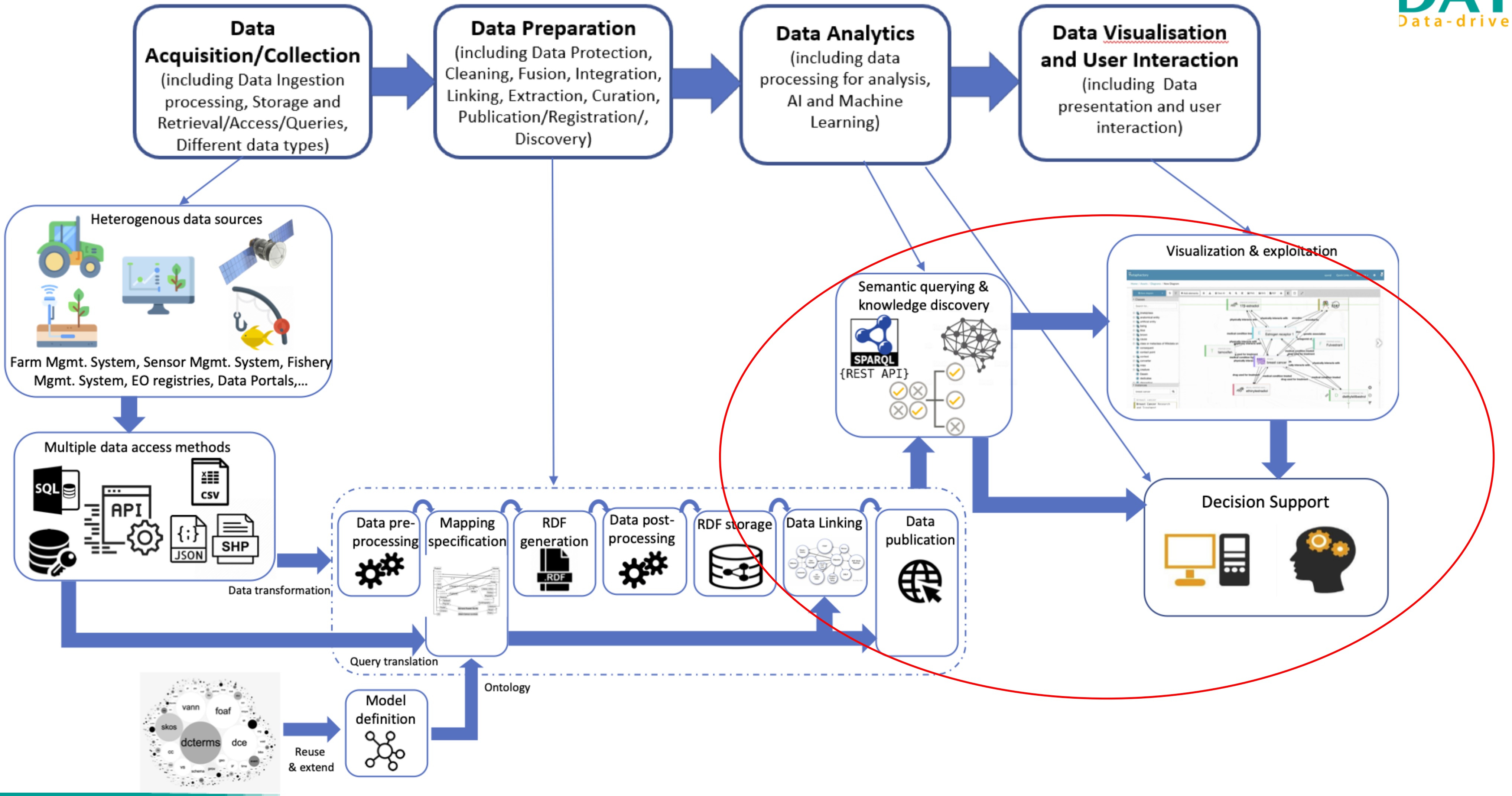
Linked Data Integration and Publication Pipeline



Source: Raul Palma

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

Case: Link discovery from knowledge graphs



Source: Raul Palma

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

Case: Link discovery from knowledge graphs

We applied discovery of RDF spatial links based on topology (Geo-L):

• Identifying fields from *Czech LPIS* data with specific soil type, from *Czech open data*

• Identifying all fields in a specific region which grow the same type of crops like the one grown in a specific field over a given period of time

• Identifying ”risky” plots from *Czech LPIS* data which intersect with *buffer zones* around water bodies.

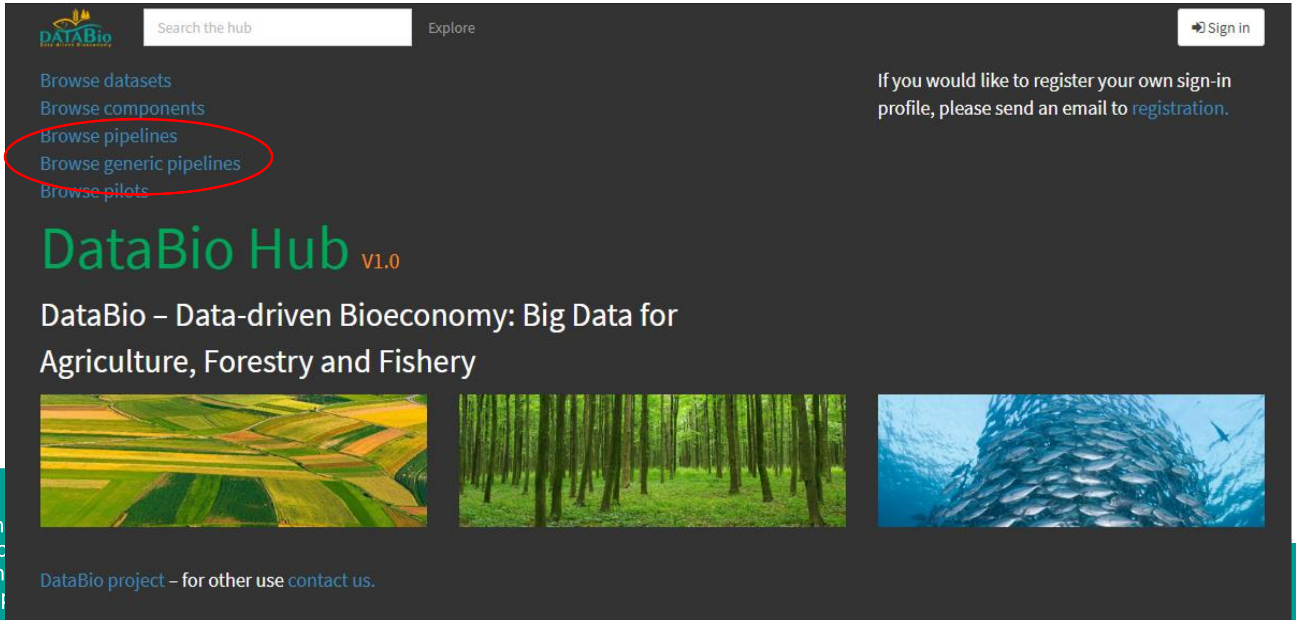
Figure: Risky overlap area between a crop field and a buffer zone of a water

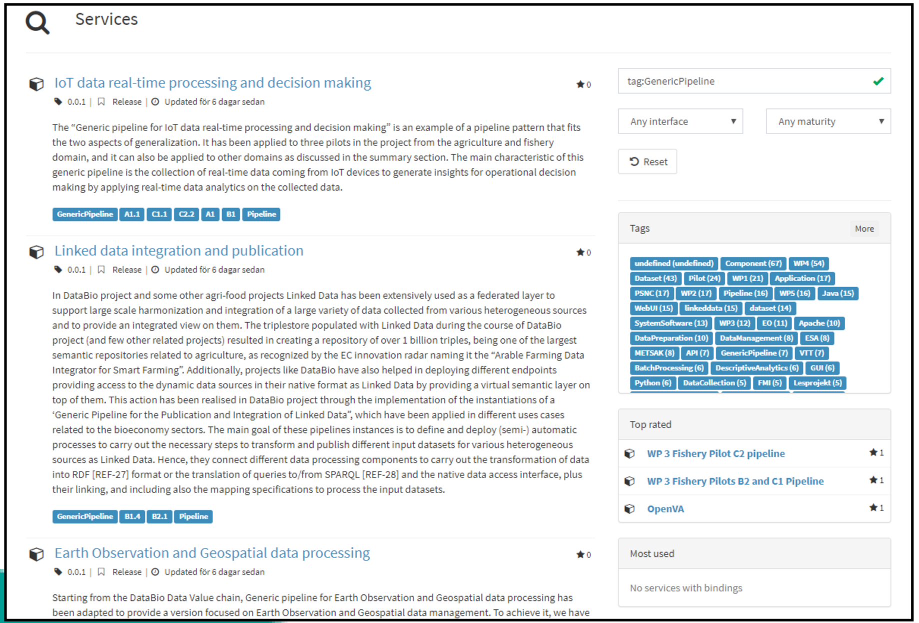
Source: Raul Palma

This document is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation programme under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

Managing project assets: components, pipelines, datasets and reports

- **DataBio Hub** (DataBioHub.eu) is central in the development platform
  - Provides a catalogue of public (and private) digital assets of DataBio
  - Links resources together (project reports, models, docker modules etc)
  - Describes currently 101 **components**, 65 **datasets**, **25 pipelines** (7 generic), 27 **pilots**
- Provides links to Deliverables, Interfaces (SPARQL endpoints, REST…)





Conclusions

- DataBio designed a *common* development platform for 27 pilots in agriculture, forestry and fishery
- Big Data *pipelines* were central in this platform as a tool for developers
- We showed that a few *generic* pipelines can be applied in numereous diverse applications
- *Instantations* of the generic pipelines were used in the pilots
- All Dtabio results are available in the searchable and cross-linked *DataBio Hub*.

This document is part of a project that has received funding
from the European Union's Horizon 2020 research and innovation programme
under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or
reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

Thank you for your attention!

This document is part of a project that has received funding
from the European Union's Horizon 2020 research and innovation programme
under agreement No 732064. It is the property of the DataBio consortium and shall not be distributed or
reproduced without the formal approval of the DataBio Management Committee. Find us at www.databio.eu.

DEEPHEALTH

Deep-Learning and HPC to Boost Biomedical Applications for Health

Pipelines for Medical Imaging Use Cases & Requirements for Benchmarking

Federico **Bolelli**, Jon Ander **Gómez**, Costantino **Grana**, Roberto **Paredes**

Evaluation schemes for Big data and AI Performance of high Business impact

EBDVF 2020, November 3-5, 2020

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

About DeepHealth

DEEPHEALTH

Aim & Goals

- **Put HPC computing power at the service of biomedical applications** with DL needs and apply DL techniques on large and complex image biomedical datasets to support **new and more efficient ways of diagnosis, monitoring and treatment of diseases.**
- Facilitate the daily work and **increase the productivity of medical personnel and IT professionals** in terms of image processing and the **use and training of predictive models** without the need of combining numerous tools.
- Offer a **unified framework** adapted to exploit underlying heterogeneous **HPC and Cloud architectures** supporting state-of-the-art and next-generation **Deep Learning (AI) and Computer Vision algorithms** to enhance European-based medical software platforms.

Key facts

Duration: 36 months
Starting date: Jan 2019

Budget 14.642.366 €
EU funding 12.774.824 €

22 partners from **9 countries:**
Research centers, Health organizations, large industries and SMEs

Research Organisations

Universitat Politècnica de València; Barcelona Supercomputing Center – Centro Nacional de Supercomputación; list; UNIMORE Università degli Studi di Modena e Reggio Emilia; Karolinska Institutet; CRS4; Università degli Studi di Torino; Otto von Guericke Universität Magdeburg

Health Organisations

Azienda Ospedaliero-Universitaria Città della Salute e della Scienza di Torino; Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana; CHUV; Spitalul Clinic Prof. Dr. Theodor Burghele București; Stockholms läns landsting

Large Industries

everis an NTT DATA Company; PHILIPS; THALES; SIMAVI Software Imagination & Vision

SMEs

WINGS ICT Solutions; TREE Technology; prodesign electronic gmbh; Stelar

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Developments & Expected Results

DEEPHEALTH

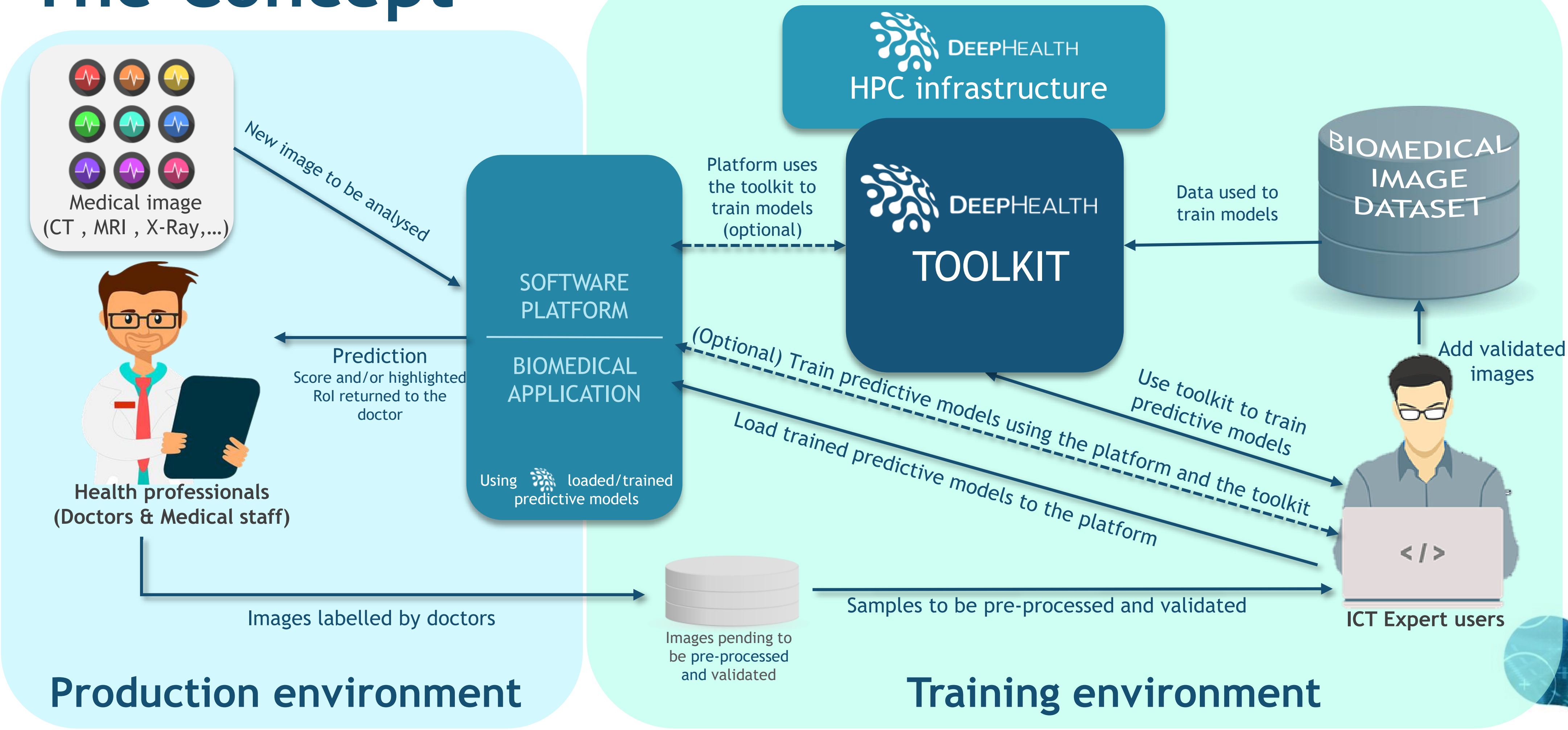
- **The DeepHealth toolkit**
  - Free and open-source software: 2 libraries + front-end.
    - ***EDDLL:*** *The European Distributed Deep Learning Library*
    - ***ECVL:*** *the European Computer Vision Library*

  EU libraries

  - Ready to run algorithms on Hybrid HPC + Cloud architectures with heterogeneous hardware (Distributed versions of the training algorithms)
  - Ready to be integrated into end-user software platforms or applications
- **HPC infrastructure** for an efficient execution of the training algorithms which are computationally intensive by making use of heterogeneous hardware in a transparent way
- Seven enhanced **biomedical and AI software platforms** provided by EVERIS, PHILIPS, THALES, UNITO, WINGS, CRS4 and CEA that integrate the DeepHealth libraries to improve their potential
- Proposal for a structure for anonymised and pseudonymised data lakes
- **Validation** in 14 use cases (*neurological diseases, tumor detection and early cancer prediction, digital pathology and automated image annotation*).

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

The Concept

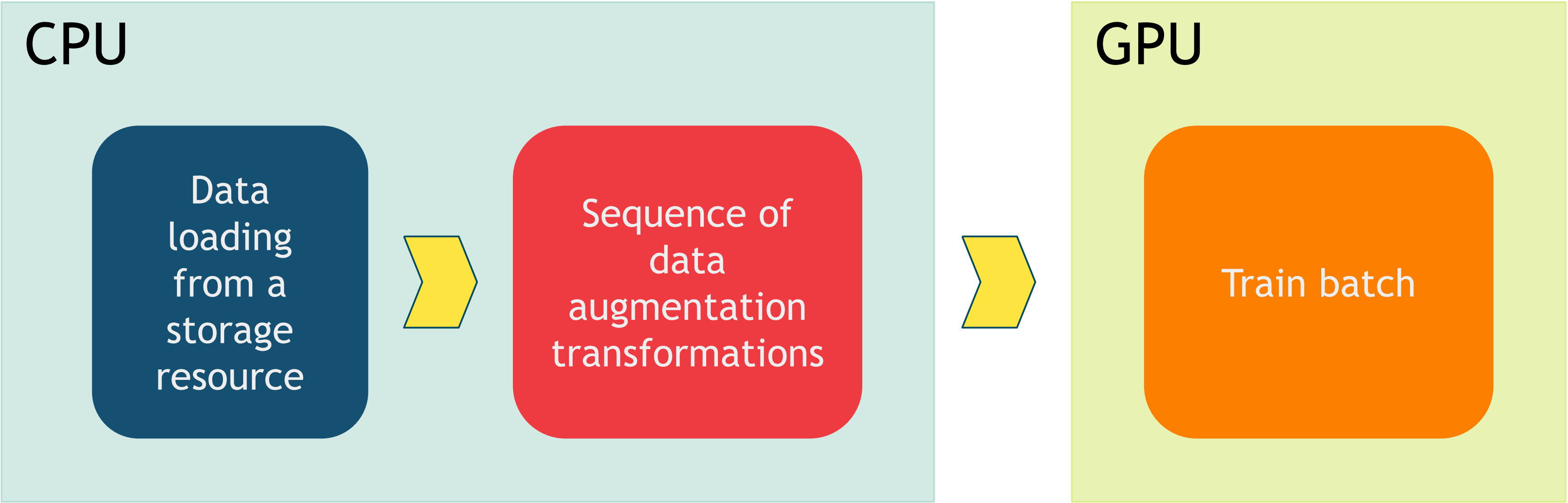


The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Pipeline Definition

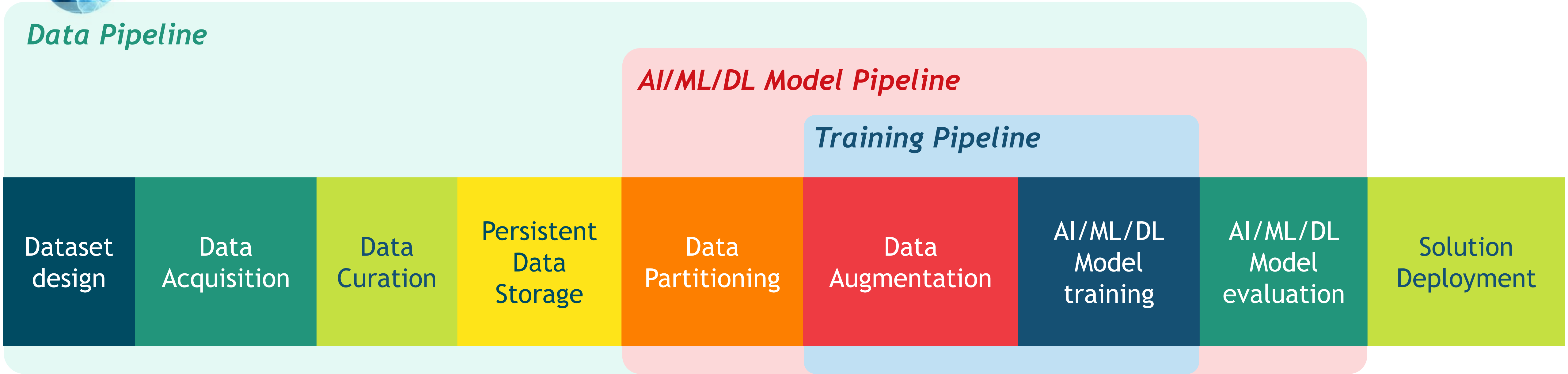
What do we understand by pipeline in this context?

A **pipeline** is a set of operations **sequentially applied** to a data block (subset of samples)



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Data & Model Pipelines



Training Pipeline *is at the core of the* ***Model Pipeline*** *which in turn* is considered part of the ***Data Pipeline***

Both pipelines are suitable for business and research applications

The whole ***Data Pipeline*** is applicable to any sector. Our project is focused on the Health sector

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Data Acquisition / Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)

Data Storage / Preparation
(including Storage Retrieval / Access / Queries Data Protection, Curation, Integration, Publication)

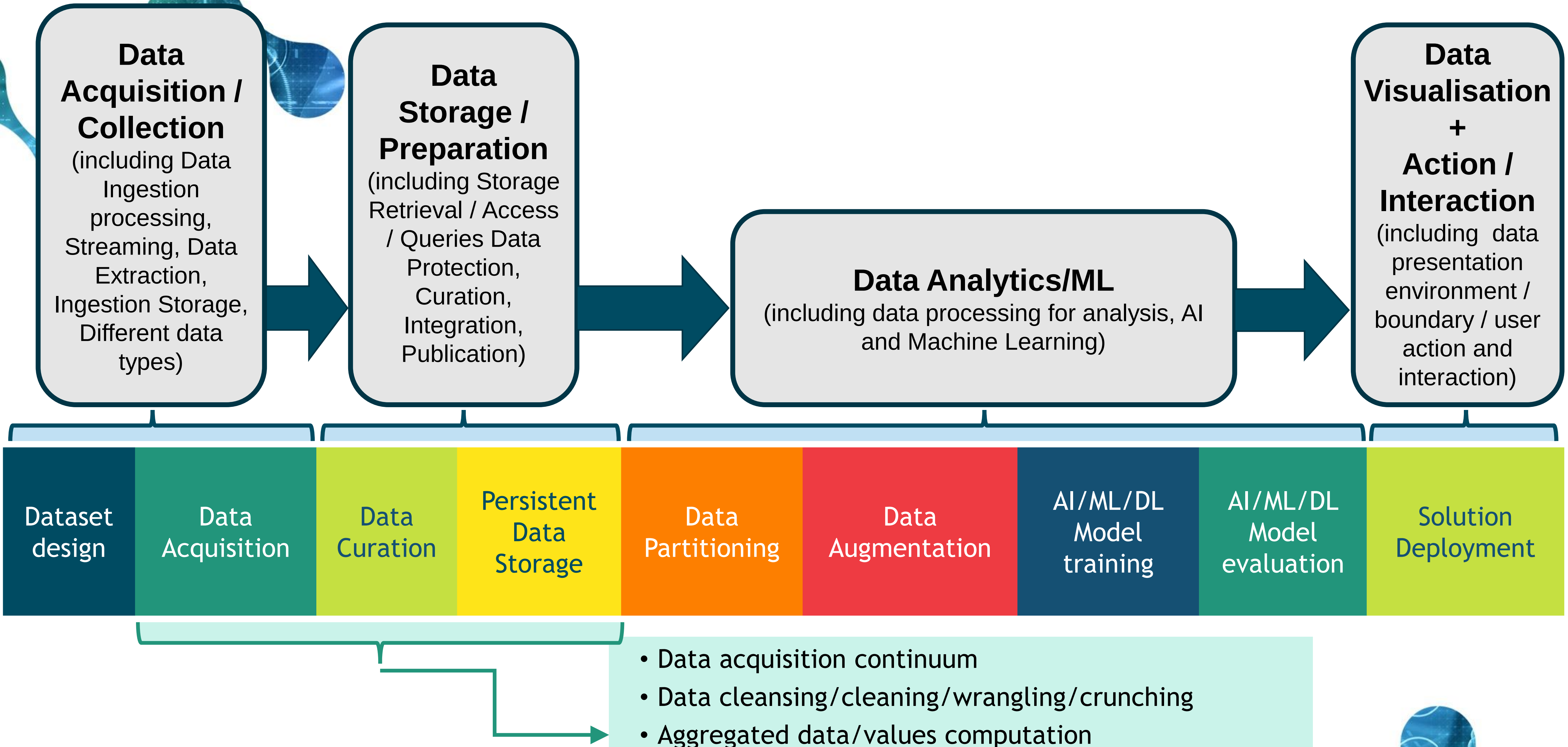
Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)

Data Visualisation + Action / Interaction
(including data presentation environment / boundary / user action and interaction)

Dataset design	Data Acquisition	Data Curation	Persistent Data Storage	Data Partitioning	Data Augmentation	AI/ML/DL Model training	AI/ML/DL Model evaluation	Solution Deployment

- Data types and formats –identification and definition
- Metadata definition
- Data Lake structure definition
- Guidelines / HOW-TOs
- Etc.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.



- Data acquisition continuum
- Data cleansing/cleaning/wrangling/crunching
- Aggregated data/values computation
- Implementation of the data-lake definition
- Creation of users and permissions or make data public

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Data Acquisition / Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)

Data Storage / Preparation
(including Storage Retrieval / Access / Queries Data Protection, Curation, Integration, Publication)

Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)

Data Visualisation + Action / Interaction
(including data presentation environment / boundary / user action and interaction)

Dataset design	Data Acquisition	Data Curation	Persistent Data Storage	Data Partitioning	Data Augmentation	AI/ML/DL Model training	AI/ML/DL Model evaluation	Solution Deployment

Model training loop:

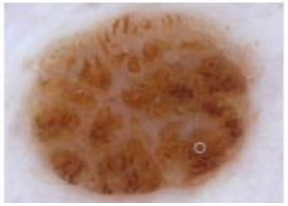
- Partitioning in training/validation/test subsets
- Data augmentation on-the-fly
- Training and evaluation of models
- Cloud & High Performance Computing requirements

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

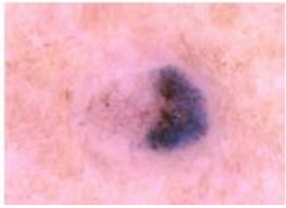
Skin Lesion Detection and Classification

- Use case nº 12 of the DeepHealth project is based on the International Skin Imaging Collaboration dataset
- Aims: identification (**segmentation**) and diagnosis (**classification**) of skin lesion images among different classes

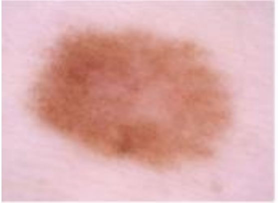
Actinic Keratosis



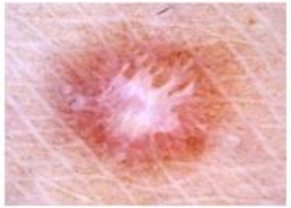
Benign Keratoses



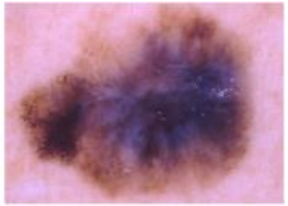
Basal Cell Carcinoma



Nevus



Dermatofibroma



Melanoma

Dataset design

Data Acquisition

Data Curation

Persistent Data Storage

- Retrospective acquisition
- 23.906 annotated images
- Publicly available on the ISIC archive website
- jpeg data format

Data Partitioning

Data Augmentation

Model training

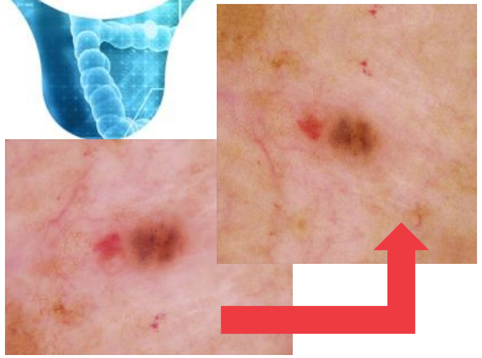
Model evaluation

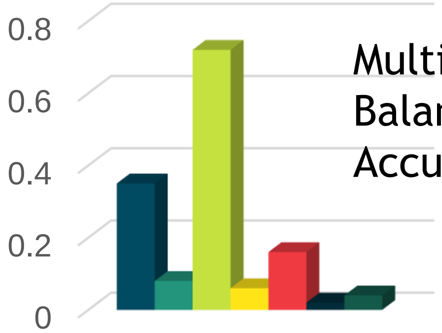
- Training 19.000
- Validation 906
- Test 4.000

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Skin Lesion Detection and Classification

Performed using the DeepHealth toolkit. Models are already available in the front-end.



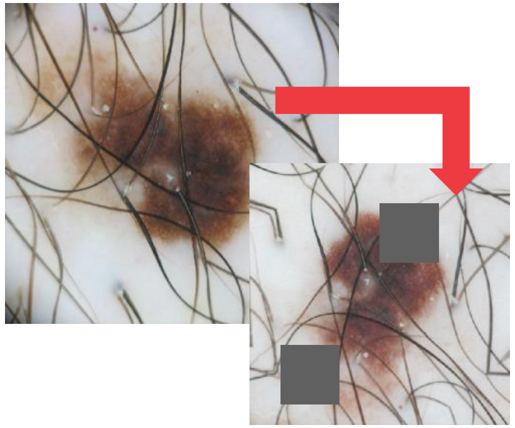


Multiclass Balanced Accuracy

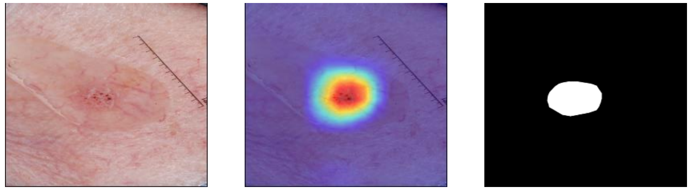
Data Augmentation

AI/ML/DL Model training

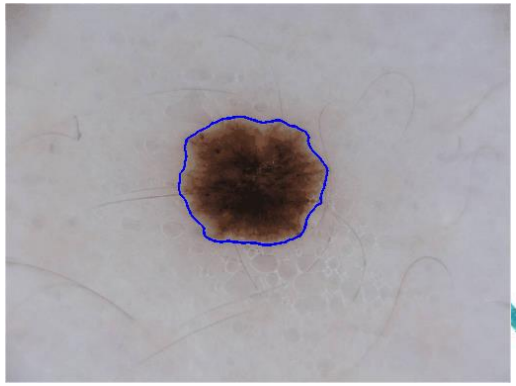
AI/ML/DL Model evaluation



Explainability plays a fundamental role in this context. Ensuring **Confidence Calibration** and providing a **Visual Explanation** of the models is essential to support clinicians.



Jaccard Index (Intersection over Union)



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Needs & Requirements

Evaluate datasets in terms of

1. **Findability** – where should a data scientist search for the dataset?
2. **Availability** – how long does a data scientist need to start the initial exploratory data analysis?
3. **Interoperability** – how long does a data scientist need to start training AI/ML/DL models with a dataset?
4. **Reusability** – are previously obtained results with a dataset public and available to other researchers / data scientists?
5. **Privacy / Anonymisation** – can the dataset be made public without compromising the identity of individuals?
6. **Quality** – is the dataset biased or unbalanced? What procedure has been followed to validate annotations?

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Needs & Requirements

Evaluate Deep Learning libraries in terms of

1. **Speed-up** – is distributed learning really efficient?
2. **Convergence** – does the distributed learning reach the same model accuracy in less time?
3. **Usability** – how long does a developer need to use the libraries effectively?
4. **Integrability** – how difficult is it to integrate the libraries as part of solutions to deploy?
5. **KPIs**: time-of-training-models (**totm**), performance/power/accuracy trade-off, etc.
6. **Others** – can you help us to evaluate other aspects?

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Needs & Requirements

Evaluate Software Platforms in terms of

1. **Usability** – how long does a domain application expert need to manage the software tool effectively?
2. **Completeness** – does the application platform provide all the algorithms/procedures/functions to allow domain application experts to easily define the sequences of steps to implement the data and/or model pipelines?
3. **Compatibility** – how many data formats does the platform admits to import/export data and models from/to other frameworks?
4. **KPIs**: time-to-model-in-production (**ttmip**), time-of-pre-processing-images (**toppi**), etc.
5. **Others** – can you help us to evaluate other aspects?

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Questions?

Federico Bolelli federico.bolelli@unimore.it
Jon Ander Gómez jon@prhlt.upv.es
Costantino Grana costantino.grana@unimore.it
Roberto Paredes rparedes@prhlt.upv.es

https://deephealth-project.eu

@DeepHealthEU

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Contacts

www.databench.eu

info@databench.eu

@DataBench_eu

DataBench

DataBench Project

DataBench

DataBench Project

DataBench

EUROPEAN BIG DATA VALUE FORUM

BERLIN + VIRTUAL
3-5 NOVEMBER 2020

This project has received funding from the European Horizon 2020 Programme for research, technological development and demonstration under grant agreement n° 780966