



DISCLAIMER: The information provided in this document is limited and will be complemented by additional materials in SP8's SGA1 Deliverables

Grant Agreement:	604102	Project Title:	Human Brain Project
Document Title:	Medical Informatics Platform v3		
Document Filename:	SP8_D8.6.4_Resubmission_Final_v1.2		
Deliverable Number:	D8.6.4		
Deliverable Type:	Prototype		
Work Package(s):	WPs 8.1, 8.3, 8.4, 8.5, 8.6		
Dissemination Level:	PU		
Planned Delivery Date:	M30 / 31 Mar 2016		
Actual Delivery Date:	M30 / 31 Mar 2016		
Resubmission Date:	13 Oct 2016		
Authors:	Ferath KHERIF, CHUV (P23), T8.6.1		
Compiling Editors:	Nathanaëlle MINARD, CHUV (P23), T8.6.1 Tea DANELUTTI, CHUV (P23), T8.6.1		
Contributors:	Anastasia AILAMAKI, EPFL (P1), T8.1.1 John ASHBURNER, UCL (P70), T8.3.2 Yoav BENJAMINI, TAU (P55), T8.3.1 Martin BRESKVAR, JSI (P87), T8.3.4 Ludovic CLAUDE, CHUV (P23), T8.4.1 Mihaela DAMIAN, CHUV (P23), T8.4.2 Harry DIMITROPOULOS, UoA (P37), T8.1.3. Tal GALILI, TAU (P55), T8.3.1 Thomas HEINIS, EPFL (P1), T8.1.1 Giannis KAZADEIS, AUEB (P3), T8.1.2 Dragi KOCEV, JSI (P87), T8.3.4 Tal KOZLOVSKI, TAU (P55), T8.3.1 Jan KRALJ, JSI (P87), T8.3.4 Boudewijn LELIEVELT, LUMC (P81), T8.3.3 Thanh LUU-THO, CHUV (P23), T8.4.1 Mira MARCUS-KALISH, TAU (P55), T8.3.1 Cesar MATOS, EPFL (P1), T8.1.1 Alexis MITELPUNKT, TAU (P55), T8.3.1 Mirco NASUTI, CHUV (P23), T8.4.1 Alexandros PAPADOPOULOS, UoA (P37), T8.1.3		



	Alberto REDOLFI, HUG (P113), T8.5.3 Görkem SAYGILI, LUMC (P81), T8.3.3 Elia SBEITI, HUG (P113), T8.5.3 Darius SIDLAUSKAS, EPFL (P1), T8.1.1 Lefteris STAMATOGIANNAKIS, UoA (P37), T8.1.3 Tassos VENETIS, AUEB (P3), T8.1.2 Vasilis VASSALOS, AUEB (P3), T8.1.2 Anže VAVPETIC, JSI (P87), T8.3.4 Eleni ZACHARIA, UoA (P37), T8.1.3 Bernard ZENKO, JSI (P87), T8.3.4
Coordinator Review:	EPFL (P1): Jeff MULLER, Martin TELEFONT UHEI (P45): Sabine SCHNEIDER, Martina SCHMALHOLZ
Editorial Review:	EPFL (P1): Guy WILLIS, Lauren ORWIN, Colin McKINNON
Abstract:	<p>The Medical Informatics Platform (MIP) was released to the public on 30 March 2016. The MIP is directly accessible online at https://mip.humanbrainproject.eu and via the HBP Collaboratory.</p> <p>The MIP is intended for end users. Therefore, users are key not only in defining the functional requirements, but also in approving the MIP end product and its quality. In addition to involving internal users actively and intensively in the specification, design and testing of the platform, the MIP was presented to external audiences. The involvement of external users will be fostered after the public release on 30 March 2016.</p> <p>The Platform is composed of two main components. The first component is the Web Portal for research services (Epidemiological Exploration, Interactive Analysis, and Biological Signature of Diseases). The Web Portal provides the connection to analytical services (data mining servers) via a microservice architecture. It also provides access to data hosted in hospitals (via the Hospital Bundle) and data from biobanks, public databases and research databases. The second component is the Hospital Bundle, a software stack that will run at every participating hospital or medical centre of the Federated Network of Hospitals and Centers (FNHC) of the MIP.</p> <p>The detail of the technology used for these two components is described in the Annex B.</p>
Keywords:	Medical Informatics Platform, architecture, user documentation, user testing, software and services, use cases, medical data
Available at:	www.humanbrainproject.eu/ec-deliverables



Table of Contents

1.	The Aim of this Document	7
2.	How to Access the Medical Informatics Platform	7
3.	Data Providers	9
4.	Platform User Instructions	9
4.1	The MIP Knowledge Base (KB)	10
4.2	General User Guidelines	10
4.3	Introduction	11
4.4	Scientific Description: Use case UC-W0010	12
	Scientific goal:	12
	Objectives:	12
	Study title:	12
4.5	Implementation	13
4.6	User Guidelines for Data Mining Method Developers	16
4.7	Software Catalogue	16
5.	Platform Testing and Quality Strategy	17
5.1	Summary of Testing Strategy	20
5.2	Quality Management Strategy	21
5.2.1	Quality Planning:	21
5.2.2	Quality Control:	22
5.2.3	MIP Project Management Tools:	23
6.	Platform User Adoption Strategy	24
6.1	Until M30	24
6.2	Beyond M30	25
7.	Help and User Feedback	26
7.1	User Feedback Received Month 18 - Month 30	28
Annex A: Platform Architectural Diagram.....		29
A.1	The Medical Informatics Platform	29
A.2	The Medical Informatics Web Portal and Research Services	31
	Release Version 1	35
7.1.1	Release Version 2	36
7.1.2	Release Version 3	37
A.3	Hospital Bundle	39
A.3.1	Hospital bundle deployment	43
Annex B: Software and Services Included in this Platform Release		45
B.1	Presentation Layer	45
	Product/Software Package/Service name: Web Portal	45
	Product/Software Package/Service name: MIP User Knowledge Base	47
B.2	Application & Computation Service	48
	Product/Software Package/Service name: Research and Modelling services: EE/IA/BSD	48
	Product/Software Package/Service name: Algorithm Factory	49
B.3	Management services	51
	Product/Software Package/Service name: Microservice architecture	51
B.4	Hospital Bundle	54
	Product/Software Package/Service name: Hospital Bundle	54
	Product/Software Package/Service name: Exareme	56
	Product/Software Package/Service name: Schema Mapping and Data Exchange (MIPMap)	58
	Product/Software Package/Service name: WebMIPMap	60
	Product/Software Package/Service name: MIPMapRew	63
	Product/Software Package/Service name: NoDB/RAW	64



Product/Software Package/Service name: Anonymization Module.....	66
Product/Software Package/Service name: Administration User Interface.....	68
B.5 Algorithms.....	71
Product/Software Package/Service name: MIP Function — Semi-supervised rule based clustering algorithm.....	71
Product/Software Package/Service name: MIP Function —Informatics-based Model: Enriched Automated Diagnostic Tools	72
Product/Software Package/Service name: MIP Function — Informatics-based Model: Deep Learning for Automated Features Extraction	73
Product/Software Package/Service name: MIP Function — Informatics-based Model: Rasch model and factor analysis for learning disease severity	75
Product/Software Package/Service name: MIP Function Informatics-based Model: Bi-clustering applied to gene expression and brain volumetric data	77
Product/Software Package/Service name: MIP Function Informatics-based Model: Bayesian Causal Model	77
Product/Software Package/Service name: MIP Function — Disease subtypes signatures - Big medical data strategy (3-C)	80
Product/Software Package/Service name: MIP Function —Label Propagation Framework.....	81
Product/Software Package/Service name: MIP Function — Multi-Target Regression on Data Streams	84
Product/Software Package/Service name: MIP Function — Predictive Clustering Trees	85
Product/Software Package/Service name: MIP Function — Rule Ensembles.....	87
Product/Software Package/Service name: MIP Function — Feature Ranking for Structured Targets	89
Product/Software Package/Service name: MIP Function — Subgroup Discovery from Multi-Resolution Data	91
Product/Software Package/Service name: MIP Function — Subgroup Discovery from Heterogeneous Data.....	93
Product/Software Package/Service name: MIP Function — Visual Performance Evaluation	95
Product/Software Package/Service name: MIP Function — Brainspan co-expression clustering	97
Product/Software Package/Service name: MIP Function — BH-tSNE	98
Annex C: Summary - Platform Use Case Status	100
Annex D: Summary - Service IT Resource Planning	110
Annex E: Summary - Service Technology Readiness Levels (TRLs) Metrics	112
Performance indicators and benchmark data	119
EXAREME	119
MIPMap	119
RAW	121
Annex F: Backlog (Remaining bugs and new features to be added).....	128
Product/Software Package/Exareme	128
Product/Software Package/MIPMap	129
Product/Software Package: MIP Computation Services.....	130
Product/Software Package: Information & Scientific Reference Services	132
Product/Software Package: Web UI.....	133
Product/Software Package: MIP User Knowledge Base	136
Annex G: IPR Status, Ownership and Innovation Potential	137
Annex H: Medical Informatics Platform Data	138
Annex J: Note on Data Standardisation.....	139
Annex K: Hospital Bundle - Deployment Experience at CHUV.....	140
Deployment Steps	140
Deployment Experience (WP8.1).....	143
Lessons Learned.....	144
Implications for the future.....	144



List of Figures and Tables

Figure 1: The Medical Informatics Platform: Model Building.....	7
Figure 2: Users of the platform can explore variables, test models and explore the biological signature of brain diseases.....	8
Figure 3: Process, Roadmap and Current Status of Data Providers.....	9
Figure 4: MIP Use Cases process, from the exploration of variables to the design and the estimation of models	11
Figure 5: SUV frontal values.....	16
Figure 6: Software catalogue	17
Figure 7: Development lifecycle at CHUV, including testing (in orange - user involvement)	19
Figure 8: Examples of functional UI testing.....	20
Figure 9: Project roles and organisation (CHUV).....	21
Figure 10: SP8 Communication Board in Trello.....	24
Table 1: Example of Platform user activities coordinated by WP8.5 and WP8.6.....	24
Figure 11: Process of managing and incorporating feedback beyond M30	27
Figure 12: Architectural Diagram of the Medical Informatics Platform.....	30
Figure 13: MIP Web Portal Architecture Diagram	31
Figure 14 Logical and physical architecture at M30	32
Table 2: Summary of MIP first release.....	35
Table 3: Summary of MIP second release	36
Table 4: Summary of MIP third release.....	37
Figure 15: Hospital bundle architecture overview.....	41
Figure 16: Exareme architecture overview	42
Figure 17: Architecture of a cluster in the microservice infrastructure	52
Figure 18: WebMIPMap User Interface	62
Figure 19: Administration UI displaying log information of a server	69
Figure 20: Administration UI showing summary of diagnosis	70
Figure 21: Administration UI showing the system load	70
Figure 22: Results for Classification Based on Random Patches Used as Input Features....	74
Figure 23: The probability density distributions of the 3 clinical groups.....	76
Figure 24: Bayesian Linear Regression.....	79
Figure 25: An original T1-weighted MRO scan (left), along with automatically extracted grey (middle) and white matter (right) tissue maps; the tissue maps encode the probability of each tissue type (given the model and data).	82
Figure 26: Grey and white matter from the original tissue atlases (left), together with registered versions (middle and right).	83

Table 5: Platform use case status	100
Table 6: Service IT resource planning	110
Table 7: Service Technology Readiness Levels Metrics.....	113
Table 8: Task description.....	120
Table 9: Evaluation times	121
Figure 27: Projection-intensive queries over JSON data	123
Figure 28: Projection-intensive queries over binary relational data.....	124
Figure 29: Selection queries over JSON data	126
Figure 30: Selection queries over binary relational data	127

1. The Aim of this Document

This document provides access to the Medical Informatics Platform (MIP) v1 and related information.

2. How to Access the Medical Informatics Platform

The MIP is one of six ITC Platforms that comprise the HBP Scientific Research Infrastructure. All these Platforms can be accessed via the HBP Collaboratory web interface:

<https://collab.humanbrainproject.eu/#/collab/19/nav/403>

Direct link to the MIP on the Collaboratory:

<https://collab.humanbrainproject.eu/#/collab/50/nav/242>

Direct link to the MIP:

<https://mip.humanbrainproject.eu>

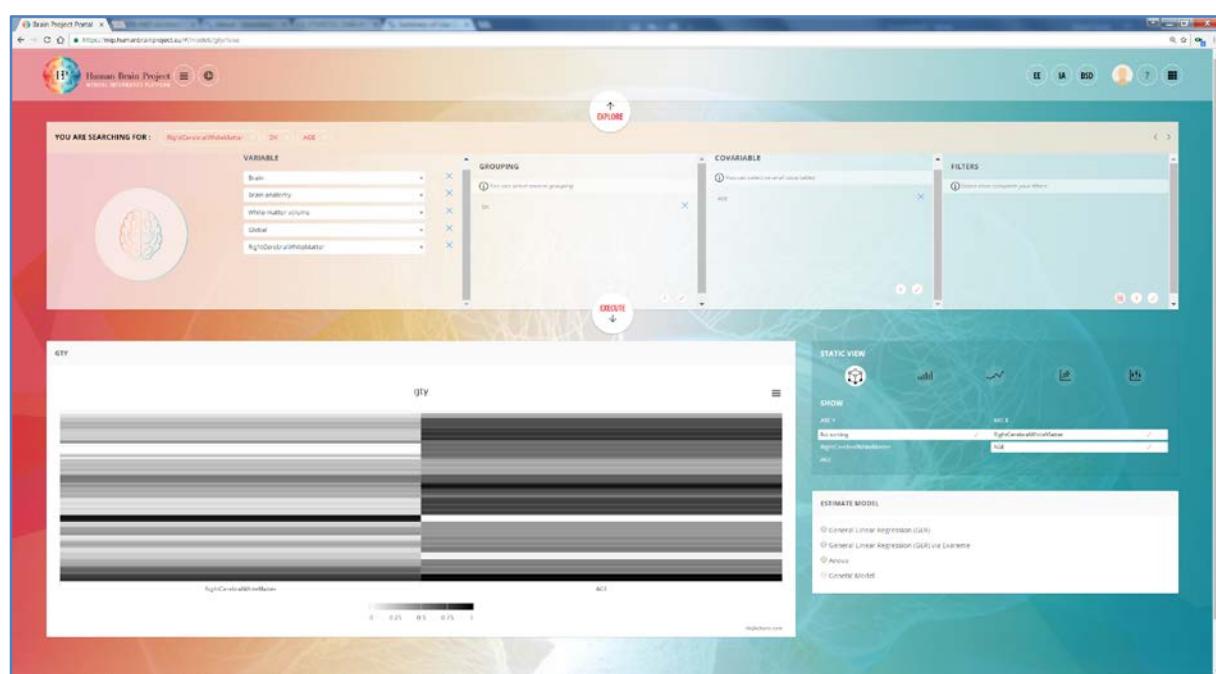


Figure 1: The Medical Informatics Platform: Model Building.



Figure 2: Users of the platform can explore variables, test models and explore the biological signature of brain diseases.



3. Data Providers

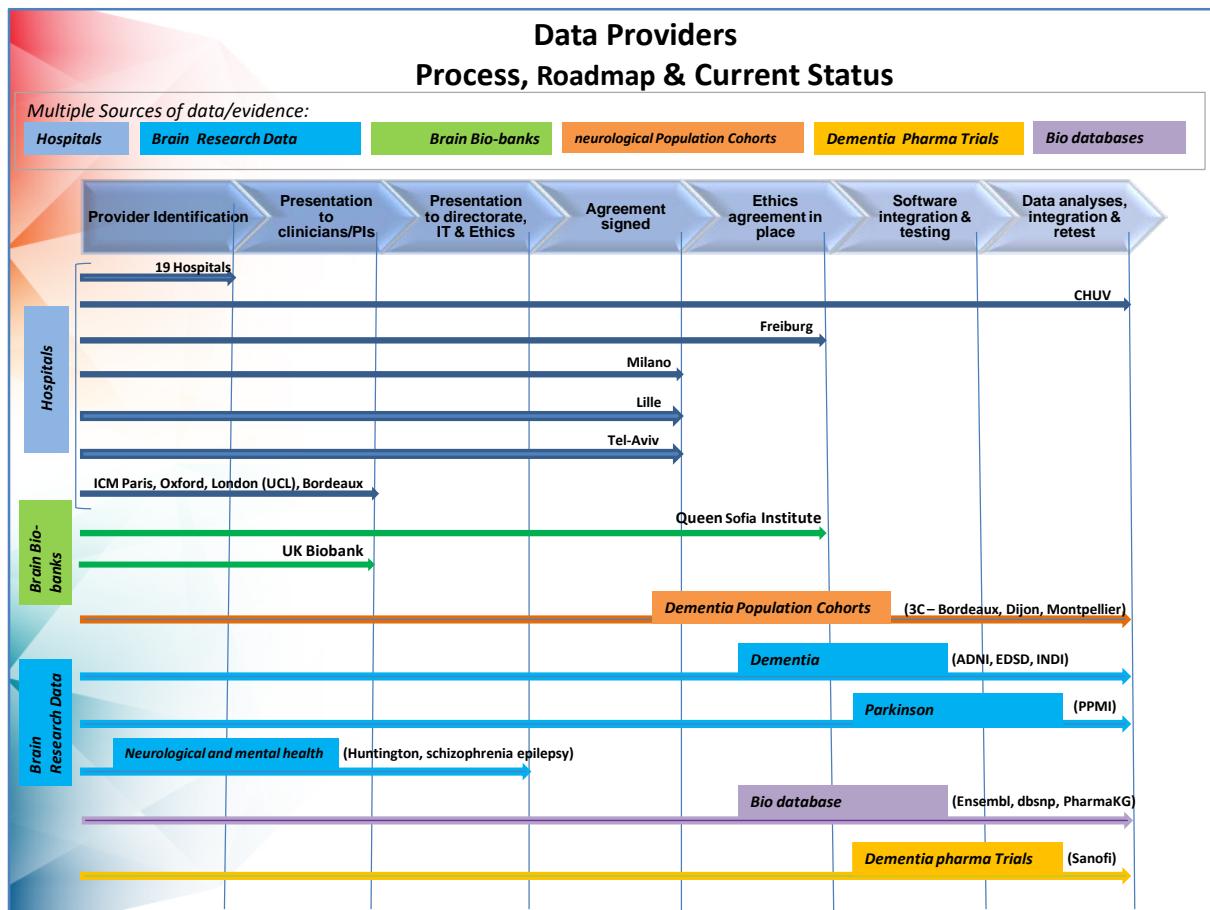


Figure 3: Process, Roadmap and Current Status of Data Providers

4. Platform User Instructions

The Platform Documentation constitutes a separate Deliverable (D8.6.5 - Medical Informatics Platform v1 – Documentation). The document includes direct links to technical and user documentation (e.g. user guidelines), software catalog and knowledge base.

D8.6.5 was scheduled to include a roadmap describing plans for future Platform development but this topic is covered in the present document (D8.6.4) - see Annex F: Backlog (remaining bugs and new features to be added).

The functionality delivered at M30 is targeted towards the majority of envisaged users of the MIP - clinicians, neuroscientists, statisticians, pharmaceutical and biotech companies, health managers and even the general public. This covers exploration of brain variables, descriptive statistical analyses on the variables/data available in the system, design of models, estimation of models, saving models/results and sharing them with the community, as well as writing descriptions/articles on the models. The models are designed and estimated by the users interactively, by selecting variables, covariates, groupings and desired methods from the MIP library. Users may also apply filters to narrow the data pool to be considered.

It should be noted that as the complexity of the functionality increases progressively in the Platform, so do the skills required. For example, the Interactive Analysis (IA) requires more advanced skills in statistics than the Epidemiological Exploration (EE).

An early version of the Biological Signature of Diseases (BSD) service is also available in the MIP. Currently, pre-computed results of unsupervised clustering algorithms permit users to browse brain disease groups and their relationships to variables (e.g. proteins, genes), and so identify subgroups with distinct biological signatures. BSD functionality will be improved and better integrated with the MIP in SGA1. New functionalities will also be added.

Information on data and data mining tools, models, the MIP library and options to explore signatures of diseases are part of the documentation available in the MIP Knowledge Base.

4.1 The MIP Knowledge Base (KB)

The MIP Knowledge Base (KB) platform is an online Content Management System (based on Liferay technology). The platform has been especially customised by T8.5.3 to empower the user by enabling him/her to participate directly to future development of the MIP and so increase the MIP community. At M30, the functionality of the KB is still limited (display of information related to the MIP - guidelines, project and tools description). Beyond M30 functionality will be extended to offer a more interactive browsing environment to users, including latest news on development of features, a calendar of focus group meetings and other events of interest, MIP roadmap, feedback forms, forum, etc. See Annex B for future development plans.

The KB platform is well integrated into the MIP through a seamless HBP OpenID authentication. It is accessible from the MIP itself (see D8.6.5 for detailed instructions).

4.2 General User Guidelines

Direct link to the User Guidelines of the MIP (regularly updated):

https://hbppsp8repo.github.io/documentation/HBP_SP8_UserGuide_latest.pdf

The figure below describes the logic of the user interface functionality flow in the MIP, based on the use cases described in the specification document of the Ramp-Up Phase (RUP; Section 3.3 in D8.6.1: Medical Informatics Platform v1 - specification document, available at <https://www.humanbrainproject.eu/ec-deliverables>).

The functionality of the process is described in detail in the MIP User Guidelines.

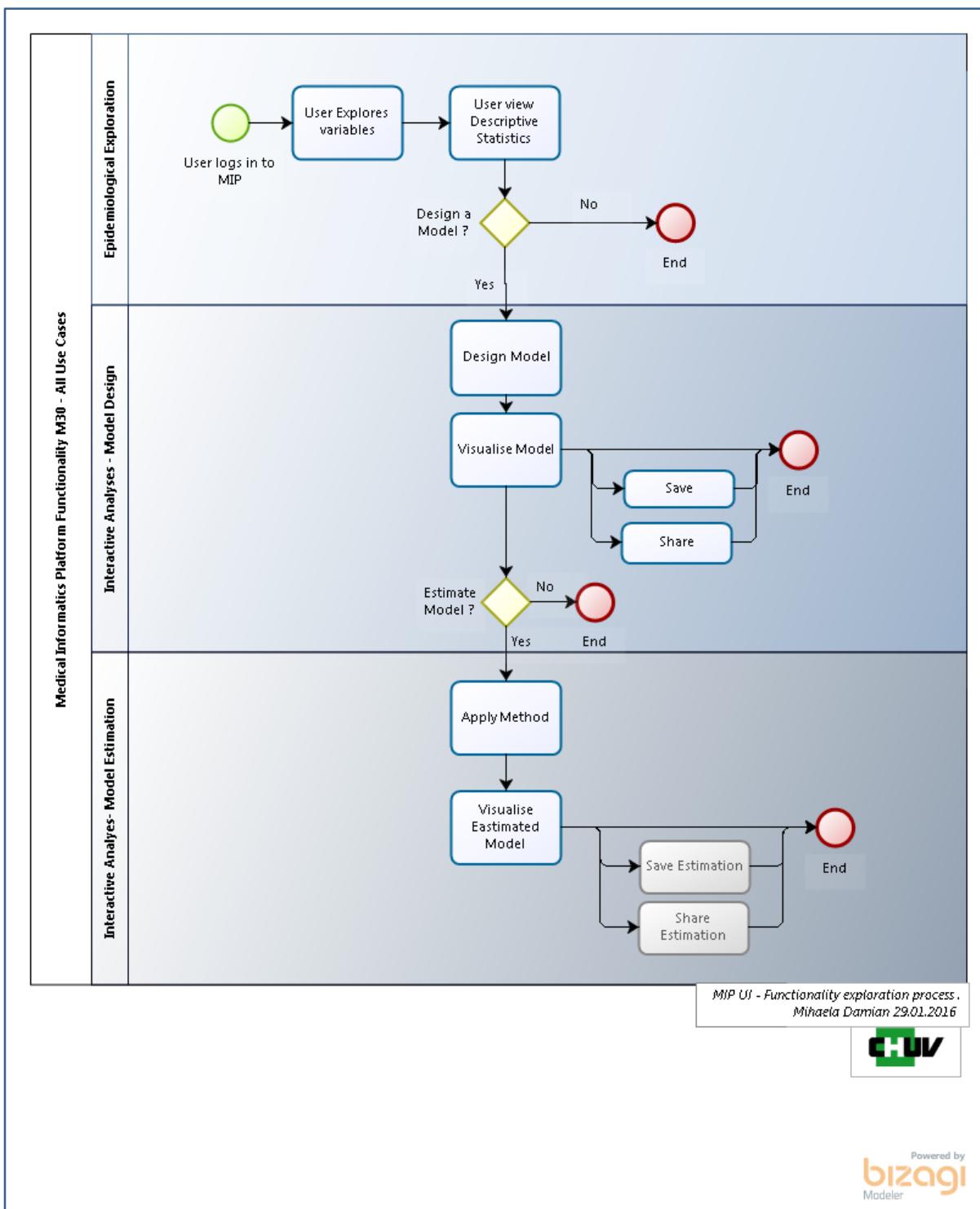


Figure 4: MIP Use Cases process, from the exploration of variables to the design and the estimation of models

Below is an example of use: Study Reproduction (UC_W0010) - use case developed for M24 Influence of biological, genetic and demographic factors on development of Alzheimer's (Dr Ferath Kherif, Gretel Sanabria-Diaz, Mihaela Damian, CHUV 09.09.2015)

4.3 Introduction

To validate the first prototype of the MIP (M24 - end Sept 2015), the team at WP8.2 CHUV decided to illustrate the use case by reproducing the results of published paper. The paper below was chosen to ensure that the existing functionality of the Platform can be realistically tested and that dataset used was also available in the MIP (i.e. ADNI data). The method used, visualisation method expected, and the desired effect were the same. Below is a summary of steps performed in "implementing" the paper:

Step 1: Identify the paper to match the functionality in MIP to be tested.

Step 2: Understand the paper.

Step 3: Identify the input parameters (e.g. dataset), the constraints (e.g. elements of the paper that cannot be reproduced, but which will not affect the quality of the tests) and the output desired.

Step 4: Define the use case, by "translating" the paper into the technical needs.

Step 5: Plan the use case implementation in detail .

Step 6: Implement.

Step 7: Test and compare results.

4.4 Scientific Description: Use case UC-W0010

This use case exemplifies the use cases UC-SP8-UC-001 to SP8-UC-005.

Scientific goal:

Examine the relative contribution of various factors (genetic, biological and demographic) and the effect of their interactions on the accumulation of regional cortical amyloid (characteristic of Alzheimer's disease(AD)).

Objectives:

For this use case, will apply a published scientific study and analyse the results using the MIP platform.

Study title:

Mapping the effects of ApoE4, age and cognitive status on 18F-florbetapir PET measured regional cortical patterns of beta-amyloid density and growth. (*NeuroImage* 78 (2013) 474-480) see <http://www.ncbi.nlm.nih.gov/pubmed/23624169>)

Necessary data: ADNI (Alzheimer's Disease Neuroimaging Initiative, ADNI) (Landau et al., 2012; Weiner et al., 2010) with florbetapir (AV45) PET scans and of various stages of cognitive impairment - early mild cognitive impairment (EMCI), late MCI (LMCI), mild AD (AD), normal controls (CN) .

Subjects with values for AV45 PET measures of Standard Uptake Value Ratios (SUVRs) for cortical regions (frontal, anterior/posterior cingulate, lateral parietal, and lateral temporal) will be selected so that the mean florbetapir uptake from gray matter relative to uptake in the whole cerebellum (white and gray matter) can be extracted. The ratio of ROI-to-whole cerebellum generates standard uptake value ratios (SUVRs) for each subject in the four florbetapir cortical regions. An average will be calculated (for each region respectively - frontal, parietal, temporal and cingulate) compared to the whole cerebellum to account for the composite/global ratios for each subject.

Demographic data considered: Age at the scan date, Education score, Gender



Genetic data to be considered: APOE status. *Genotyping methods are described www.ADNI.org*

Computation method used: Linear regression

The model: $Y = (\text{Diagnosis}) \times (\text{APOE status}) + (\text{Age}) + (\text{Education}) + (\text{Gender})$

The model will return the mean values for each subgroups according to their APOE status: CN (control), AD, EMCI, LMCI.

4.5 Implementation

1) Actors

Alan is a researcher in neurology.

2) Description

Alan wants to assess how the APOE4 gene influences the beta-amyloid markers in humans, and hence inform about AD diagnosis, taking into account age, gender and education. He is interested in subjects with AV45 PET scans taken between May 2010 and March 2012 for the frontal lobe brain region. He wants to visualise the results grouped by the different stages of cognitive impairment: CN, EMCI, LMCI and AD, and APOE4 status absent (=0) or present (>0).

3) Preconditions/Prerequisites

- PET imaging data has been pre-processed and extracted features are available (details regarding processing described at <http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/> and <http://adni.loni.usc.edu/methods/documents/>).
- PET images available online at <http://adni.loni.usc.edu/data-samples/pet/>.
- APOE status has been determined and exists in the database. Genotyping methods used to determine APOE status are described at <http://adni.loni.usc.edu/>.
- The subjects have been diagnosed before the scan (inclusion and exclusion ADNI criteria in www.adni-info.org) and data for the different disease stages (EMCI, MCI, mild AD) is available.
- The SUVRs for the frontal region are available (pre-calculated and in the API).
- Linear regression computations are ready and integrated within MIP.
- The interface between the Web UI and the MIP back-end (and computation scripts - Linear regression) are completed.
- The age of the subject at the time the PET was taken is available.
- The web interface of the MIP is available and permits the selection of all the needed criteria.

4) Input parameters

Variables:

Brain Metabolism > PET > AV45 > FRONTAL	<i>Note: this will be on the y axis</i>
---	---

Grouping:

APOE4	<i>Note: APOE4 status values are 0 or >0.</i>
DX_bI	<i>Note: DX_bI contains the various cognitive impairment stages (CN, EMCI, LMCI, AD)</i>

Covariates:



AGE
PTGENDER (aka gender)
PTEDUCAT (aka education)

Filters:

For this particular use case, Alan decides to select the following filters:

Filter description	Filter Details
Previously selected variables	Brain Metabolism > PET > AV45> Frontal
Only PET scans taken between May 2010 & March 2012	May 2010 < EXAMDATE < March 2012
Only DX_b1 stages with certain MMSE and CDR scores :	<p>For DX_b1= « EMCI », only consider subjects with: $24 \leq \text{MMSE} \leq 30$ and $\text{CDR} = 0.5$</p> <p>OR</p> <p>For DX_b1= « LMCI », only consider subjects with: $24 \leq \text{MMSE scores} \geq 30$ and $\text{CDR} = 0.5$</p> <p>OR</p> <p>For DX_b1= « AD », only consider subjects with: $20 \leq \text{MMSE scores} \geq 26$ and $\text{CDR} = 0.5 \text{ or } \text{CDR} = 1.0$</p> <p>OR</p> <p>For DX_b1= « CN », only consider subjects with: $24 \leq \text{MMSE scores} \geq 30$ and $\text{CDR} = 0$</p>
Do not include DX_b1 = « SMC »	$\text{DX_bl} < > \text{« SMC »}$

5) Basic flow of events (in gray - related but outside the scope of Web UI)

- Alan wants to visualise the Standard Uptake Value Ratios (SUVRS) in a plot for the cortical FRONTAL region with respect to APEO4 gene and cortical impairment stages, considering age, gender and education.
- Alan connects to MIP and accesses the user interface.
- On the web interface, Alan selects the criteria and applies the filters (see Input Parameters).
- Alan launches the criteria.
- Criteria are being sent to the MIP computation layer where linear regression method is applied.
- MIP computation layer returns back to the Web UI the result of the linear regression (vectors).



g) The Web UI displays the plot.

6) Alternative flows

A1:

1.A1	Alan wants to run the same analysis for the CINGULATE, PARIETAL, and TEMPORAL lobes respectively, so he will change the variable accordingly, e.g. Brain Metabolism > PET > AV45 > CINGULATE
2.A1	<i>Alan launches the criteria.</i>
3.A1	<i>Criteria is being sent to the MIP computation layer where linear regression method is applied.</i>
4.A1	<i>MIP computation layer returns back to the Web UI the result of the linear regression (vectors).</i>
5.A1	The Web UI displays the plot.

A2:

1.A2	Alan decides to change the default type of the graph. He selects the new type of graph and applies it to the result before.
2.A2	The Web UI displays the plot.

A3:

1.A3	Alan decides to save the graph and share it within his group of interest.
2.A3	The graph is available in Alan's interest group.

7) Exceptions (if any)

8) Post-conditions

The results are successfully displayed on the Web UI.

Basic Flow:

X axis: 1) the different stages of disease (the groups in DX_bl), split by the APOE4 status (0 or !0).

Y axis: returned by the Linear Regression (sent to the Web UI via the API): the values and the average of the SUV Frontal for each record (see picture below as example).

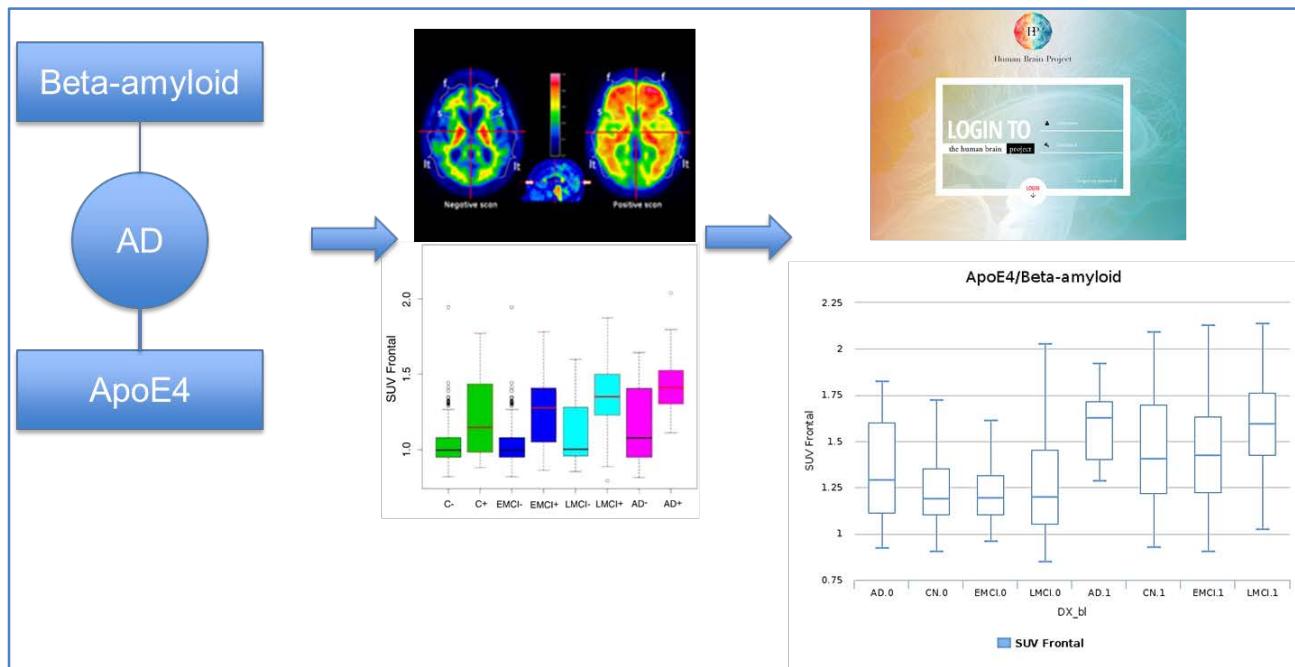


Figure 5: SUV frontal values

4.6 User Guidelines for Data Mining Method Developers

At M30, developer users are internal to SP8. They are part of W8.2, WP8.3 and WP11.2. Their role is to develop and provide methods/algorithms for the Platform.

Guidelines for submitting algorithms into the catalogue pipeline are available at <https://github.com/LREN-CHUV/functions-repository/blob/master/Guidelines.md>.

Beyond M30, the MIP roadmap plans to open the development of methods to the external community of developers. Once this functionality is added, developer users will be able to create their own KB items on the methods created (description and guidelines), and to open discussion threads.

4.7 Software Catalogue

The software catalogue is accessible online (<https://hbpsp8repo.github.io/>). The web page lists all the software developed in the RUP with additional information such as type of licence, authors, access to the code source, and bug tracking, where available.



Software Catalog

[Twitter](#) [GitHub](#)

HBP SP8 repository

- [HBP SP8](#)
- [Documentation](#)
- [Software Catalog](#)
- [Open source software](#)
- [Docker images](#)
- [The author](#)
- [@HBPmedical on Twitter](#)
- [@HBPSP8Repo on GitHub](#)
- [Contact via email](#)

Open source software

Deployment	Organization	License	Management	Continuous Integration
mip-microservices-infrastructure	CHUV_LREN	license Apache2.0	trellis infrastructure	
dev-setup	CHUV_LREN	license Apache2.0	trellis development-tools	
RAW-deploy	EPFL_DIAS			
ansible-airflow	CHUV_LREN	license MIT		
docker-flyway	CHUV_LREN	license Apache2.0		
xnat-docker	CHUV_LREN	license Apache-2.0		
labkey-docker	CHUV_LREN	license Apache-2.0		

Portal	Organization	License	Management	Continuous Integration
portal-specs	CHUV_LREN			
portal-backend	CHUV_LREN	license AGPL-3.0		
portal-frontend	CHUV_LREN	license AGPL-3.0	trellis web-frontend	
bootstrap-mip-app	CHUV_LREN	license MIT		
sunbeam	CHUV_LREN	license CERN-YU		

Figure 6: Software catalogue

5. Platform Testing and Quality Strategy

The figure below represents the development lifecycle of the Platform, including testing.

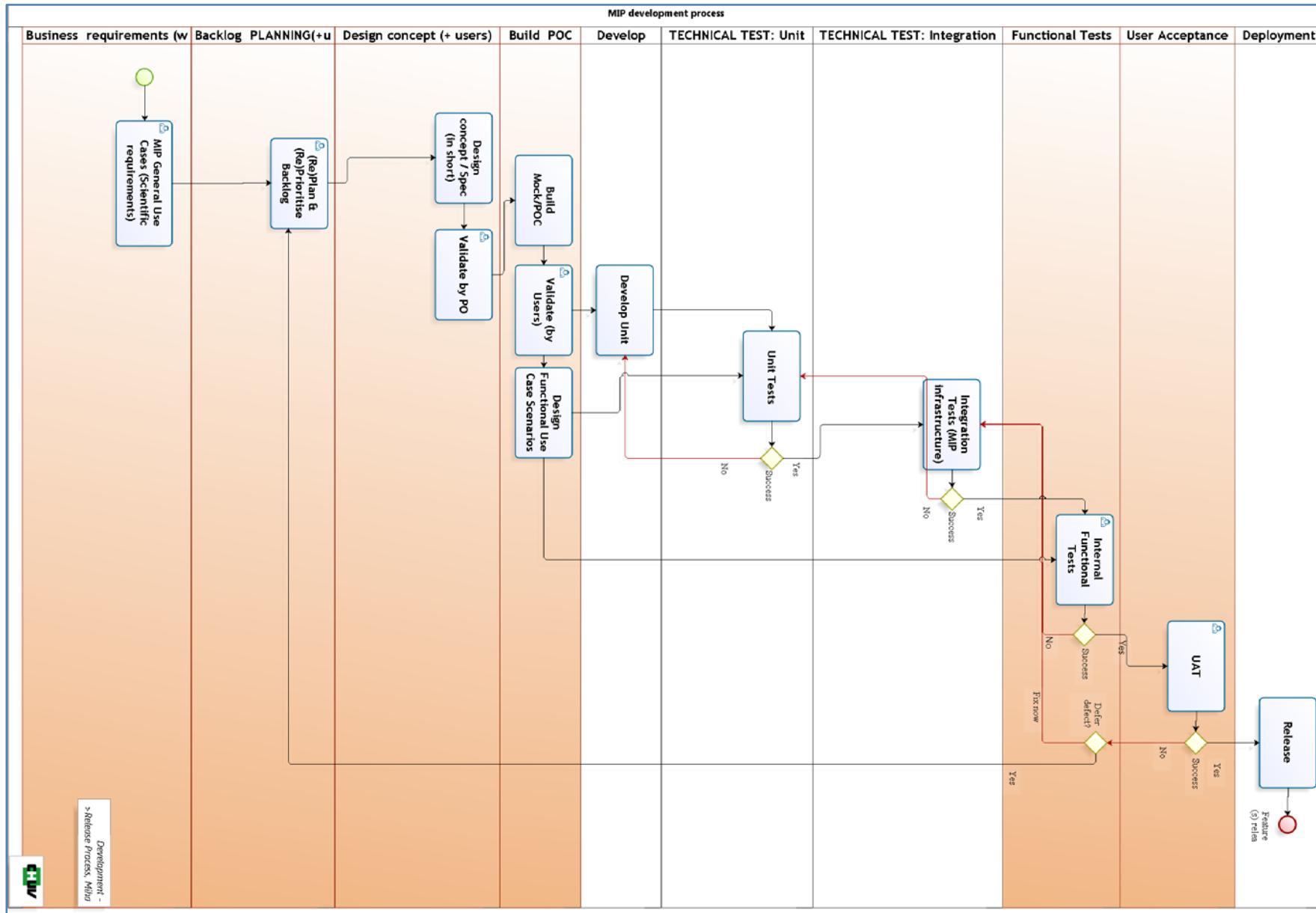




Figure 7: Development lifecycle at CHUV, including testing (in orange - user involvement)



The full Testing and Quality Strategy defined and implemented by WP8.4 (CHUV) for the Platform is available online, in the MIP Knowledge Base:

<https://mip.humanbrainproject.eu/help/testing-quality>.

The Strategy includes the development lifecycle process, tools used in testing, sample testing results (at M24), and specifications following users' feedback received at M18 and M24.

A short summary of the Testing Strategy is available in the first section of this chapter. The second section gives a detailed description of the Quality Management Strategy.

Sample of the Functional UI tests at M24 (before UAT)								
Category / Area	Type	ID	Action	Description	Expected result	Validation	Validation CHUV - MD - 4.12.2015	
Login page	Functional	R-01	Login	The user can connect himself to the platform	The user accesses the whole platform. Only users into the platform database can login	Success	Success (all 3 test users)	
Widget (global)	Functional	R-02	Widget reload	The widget can be reloaded	The displayed data are updated according to the platform DB (not the API)	Abandonné	Why? Explanation useful.	
Widget (global)	Ergonomic	R-03	Drag'n'drop	The widget can be moved with a drag'n'drop on the widget header	The widget stays to its new place according to the page grid. The other widgets moved if necessary, following the grid	Success	Success	
Widget - List of models	Functional	R-04	On load	Models displayed in the widget	Only validated models can be listed	Success	Success if Details section true. See ->	
Widget - Single model	Functional	R-05	On load	Model displayed in the widget	Only validated models can be displayed on the Community page and the List of models on the dashboard	Success	Success, but see above	
Widget - List of articles	Functional	R-06	On load	Articles displayed in the widget	Only validated articles can be listed	Success	Success, but see above	
Widget - List of articles	Functional	R-07	Edit	Access to the article edition	Only accessible if the current user is the article's author	Success	Success	
Widget - List of articles	Functional	R-08	Download	The user download the article	The article is downloaded in PDF	Success	Success	
Widget - List of articles	Functional	R-09	Preview	The user preview the article	The article is displayed in its final format	Success	Success	
Widget - Single article	Functional	R-10	On load	Article displayed in the widget	Only validated article can be displayed	Success	Success (i.e. articles displayed), but validated = ? (as R04-R06)	
Widget - Single article	Functional	R-11	Edit	Access to the article edition	Only accessible if the current user is the article's author	Success	Success	
Widget - Single article	Functional	R-12	Download	The user download the article	The article is downloaded in PDF	Success	Success	

Figure 8: Examples of functional UI testing

5.1 Summary of Testing Strategy

In addition to involving internal users actively and intensively in the specification, design and testing of the platform, the MIP was presented to external audiences. The involvement of external users will be fostered after the public release on 30 March 2016.

The MIP is intended for end users. Therefore, users are key not only in defining the functional requirements, but also in approving the MIP end product and its quality. The latter refers to the User Acceptance Tests (UAT). These are carried out per release, after all the technical tests (unit, integration and full-system) and internal functional tests have been carried out and passed successfully.

Up to M30, internal users have been continuously and intensively involved in all steps along the project lifecycle. After M30, the MIP will start involving external users (see section on [Platform User Adoption Strategy](#)).

- **Users involved before M30 (internal):**



- neuroscientists, statisticians, clinical psychologists, neurologists
- **Users envisaged to be involved after M30 (external):**
 - clinicians (epidemiologists, neuroclinicians), researchers (in neurology, neuroimaging, pharmacology), statisticians (scientific developers, method and algorithm developers), platform developers (to deploy tools and algorithms), pharmaceutical industry, R&D departments (for cooperation in clinical study designs, personalised medicine therapy studies), medical & research writers, the general public.

Documentation on the different types of users can be found online in the RUP Deliverable D8.6.1: Medical Informatics Platform v1 at <https://www.humanbrainproject.eu/ec-deliverables>.

The tools used in testing are JIRA, Excel, SOAP UI, Karma, Jenkins, Selenium, Matlab, R. More details are available at <https://mip.humanbrainproject.eu/help/testing-quality>.

5.2 Quality Management Strategy

For the MIP, SP8 decided to make use of the best practices of the two best methodologies currently used in industry. PRINCE2 for high-level project governance, project organisation, controls, processes and themes (including Quality Management) and Agile for the implementation phases.

This has proved to be extremely effective in such a complex scientific project where flexibility and incremental delivery is best carried out within organised, well established and controlled boundaries.

To achieve and maintain a high project and product Quality, we start with a well organised team, with clear roles and responsibilities. Our *Product Owner* represents the users' interests and ensures their needs are reflected correctly in the product. The *Project Scrum/Manager* works closely with the Product Owner and ensures the *Team* understands the scientific needs and the product is delivered according to the specification, within deadline and quality tolerances.

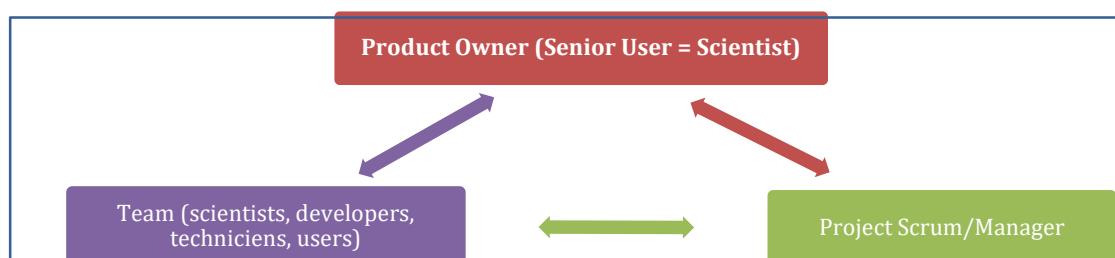


Figure 9: Project roles and organisation (CHUV)

The MIP coordination team at CHUV considers quality management to be a continuous process, running along the whole MIP product lifecycle. To manage the project and product quality, we consider quality planning, quality control and MIP project management tools.

5.2.1 Quality Planning:

- End-user quality expectations (from Project's Brief).
- Defining and applying best practice in daily project implementation (i.e. PRINCE2 and Agile for implementation phases, as described above).
- Quality tolerances (at each important milestone).

- Acceptance criteria (defined with the end user, product tested against it).
- Standards (i.e. most appropriate for software, project methodology and implementation approach).
- Management of change - analyse impact of new proposed functionality (onto unit, whole system and users), prioritisation, plan of deployment and communication to users.
- Deployment and user engagement - user engagement will be tailored to fit the released features at each milestone and the long-term aim (see section on [Platform User Adoption Strategy](#)).

5.2.2 *Quality Control:*

We ensure quality through:

- Rigorous, incremental and responsive planning (development and release dependencies are highly important).
 - Automation of software processes (to eliminate human error, increase performance quality and ensure consistency).
 - Automation of test cases: currently semi-automated, automation being mainly done at the low unit level. Post M30, when the Platform will be more stable and framework in place, more testing processes can be automated.
 - Definition and implementation of business and operational processes on a daily basis (e.g. clear processes).
 - Risk and Issue management (ensuring quality by overcoming risks before becoming issues and risking jeopardising the quality).
 - Risks and issues within SP8 are considered, (logged), dealt with and escalated from the individual units/Tasks up - i.e. within a Task, escalated as necessary to the WP, to SP level and furthermore to the HBP where, in the near future, we envisage to integrate with the HBP central risk register.
 - The coordination team at CHUV is an active member of the HBP Risk Management group, where it contributes with solutions to the overall programme and project.
 - End Stage Report to Product Owner (Product Owner = Senior Scientist/Senior User).
 - Create Lessons Learnt Log, and apply them along the project.
 - Quality Review Technique (organised procedures assessing the fitness of the product or particular functions).
 - *Participants:* users and stakeholders
 - *When:*
 - to validate functional specification (with direct impact on user experience, especially UI)
 - to validate implemented functionality (as UAT)
 - at any point in the project, as required
- Benefits:* early identification of defects, proved ideal in Agile methodology
- early involvement of the users and stakeholders, and hence willingness to commit to the product
- objectivity offered to management in assessing the project progress



5.2.3 MIP Project Management Tools:

A variety of Project Management tools is used by CHUV team (WP8.4, WP8.6). Their use depends on the purpose and the skills of collaborators involved:

- **MindJet/MindMap** - Good for drawing an overall visual picture of the product to build (product based diagram). Can add dependencies, deadlines, resources. We (the Project Manager (PM) mostly) use it to create the Product Breakdown Structure (PBS) of the whole MIP. It is also used for individual phases.
- **MS Project Management** - Good for creating Gantt Charts and reports to the management, creating dependencies between tasks and phases (more reliable than MindJet), graphical interface of task tracking, highlight critical paths, etc. Disadvantage: the audience must be used to understanding Gantt Charts. We mostly use it for the big-picture plan in conjunction or addition to PBS.
- **JIRA** - We use this for issue and task tracking internally at CHUV, but also in our common tasks with the other teams within or outside SP8 (e.g. the Collaboratory, SP5). Advantages: working with remote and disperse teams; all communication on a task or issue is kept together and visible; can have watchers just following the status of the tasks. Disadvantages: takes time to configure, create and link tasks; not easy to see all tasks.
- **Excel** - The easiest for very small tasks at very short deadlines when under pressure; easy for everybody to understand and use.
- **Trello** - WP8.6 team created a “SP8 Communication Board” where important project messages can be shared amongst SP8 teams - sprints and their status, backlog, business processes. Access possible by request <https://trello.com/hbpsp8>. Some members of the team use it to organise own workloads.
- **Bizagi** - Business Process Modeller, excellent for modelling, simulating and running processes. We use it intensively. Examples include: modelling the user request workflow in the MIP; visualising the different use cases in the Web UI and their interaction; for user support and feedback mechanism (operational phase); for developer user method deployment mechanism, etc.

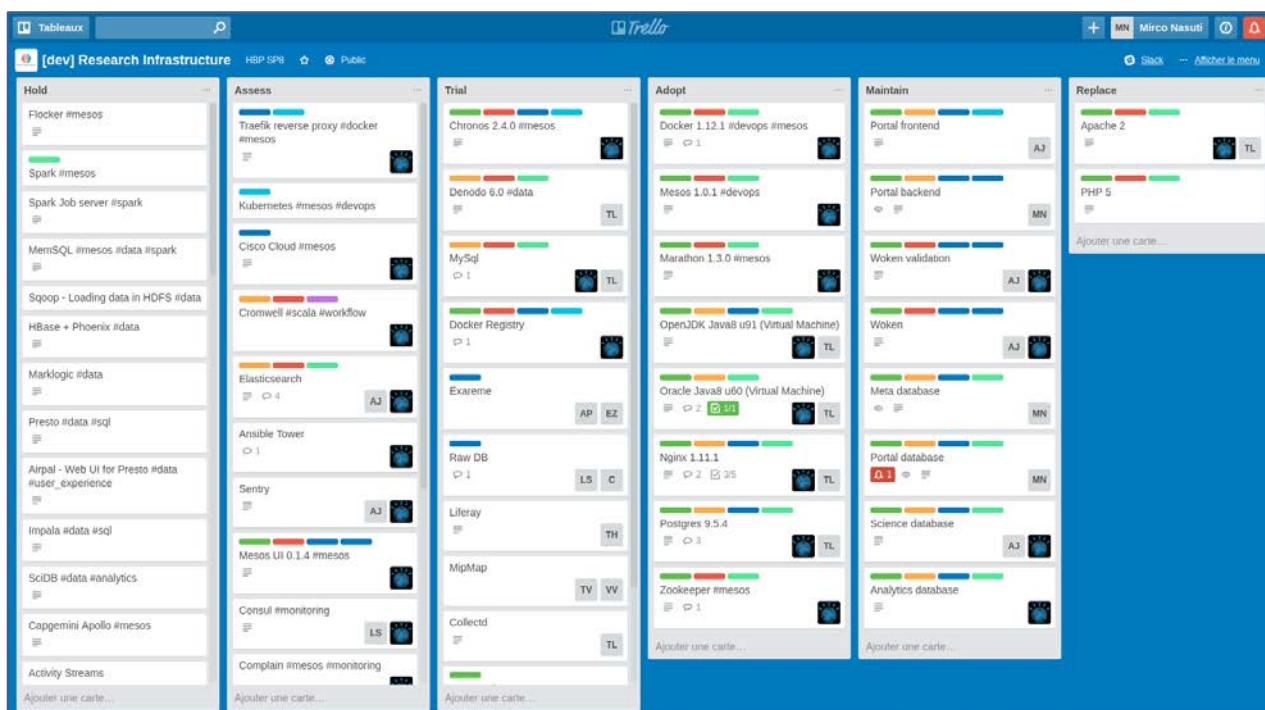


Figure 10: SP8 Communication Board in Trello

6. Platform User Adoption Strategy

6.1 Until M30

In addition to involving internal users actively and intensively in the specification, design and testing of the Platform, the MIP has been presented several times to external audiences. Their feedback was used for validating ideas and direction:

Table 1: Example of Platform user activities coordinated by WP8.5 and WP8.6

Date	Activity	Attendees (approx.)
End Feb 2016	Annual presentation to representatives of the IT at CHUV.	5
End Oct 2015	Presentation to Neuroscience and Medicine working group in Lausanne. The group included directors of hospitals and clinics, heads of clinical departments, and heads of Life Science departments of the University of Lausanne, the EPFL and the University of Geneva.	12
End Sept 2015	Presented solution to HBP members - during the HBP Summit 2015 in Madrid.	300



July 2015	Presented solution to Brain/MINDS Project (Tokyo and Kyoto Universities).	5
15-18 March 2015	Second HBP Education Workshop: Future Medicine - Medical intelligence for Brain Diseases (CHUV, Lausanne).	100
Jan 2015	Presentation to the IMI initiative.	20
Nov 2015	Web Summit Dublin - "HealthTex".	42K >3K talk attendance

6.2 Beyond M30

After M30 the MIP will be slowly opened to external collaborators:

- External method developers, which will be able to include their algorithms into the Platform.
 - At that time, the MIP Knowledge Base will also be open to external users for editing
 - Users will be able to describe the methods developed, create guidelines or open forum discussions with the community.
- All users:
 - Invite 100 external users (before the end of June 2016) to participate in organised Focus Groups and online chats where functionality will be discussed, feedback received, impact analyses on current version made and approved changes prioritised.
 - Users will be able to vote online on a list of suggested new features or improvements ("e-news").
 - Most voted features, together with mandatory system upgrades (related to technical performances, improvement of processes etc.), will be discussed in the Focus Groups.
 - Approved/agreed improvements will be displayed in a "News" section on the MIP, so that users will always be informed about upcoming new features or changes.

The following activities will also be organised to encourage the use of the Platform and the involvement in its future development:

- Widely broadcast videos and tutorials explaining not only the existing functionality but also the vision, and short- and long-term benefits. This includes the progression to the discovery of Biological Signature of Diseases, opening the Platform usage to disease spaces other than brain, and the discovery of intercorrelations between diseases and their treatments, wanted and unwanted effects of treatments, etc. We believe that end users need to understand the large goal and long-term desired benefits of the Platform in order to be motivated to contribute effectively to its development.
- Recording and publishing Focus Groups meetings.
- Workshops and lectures at HBP and non-HBP education events.
- Presentations and information exchange with similar projects worldwide, as well as with the European Commission.

- Ideally, “roadshows” at neurology units of big hospitals - to specifically attract expert users and additional contributing hospitals.

7. Help and User Feedback

To obtain help in using the Platform, users can start by checking the online user guidelines at https://hbpsp8repo.github.io/documentation/HBP_SP8_UserGuide_latest.pdf.

They can also browse the Knowledge Base, which is available directly in the MIP.
<http://193.204.145.212:8080/liferay/>

If personal assistance is needed, or if users want to provide feedback or contribute to the on-going development of the Platform, they can write an email to: hbp_medicalinformatics@chuv.ch.

The next section describes how user feedback has been incorporated into MIP developments up to M30.

The diagram below represents the process of managing and incorporating feedback after M30.

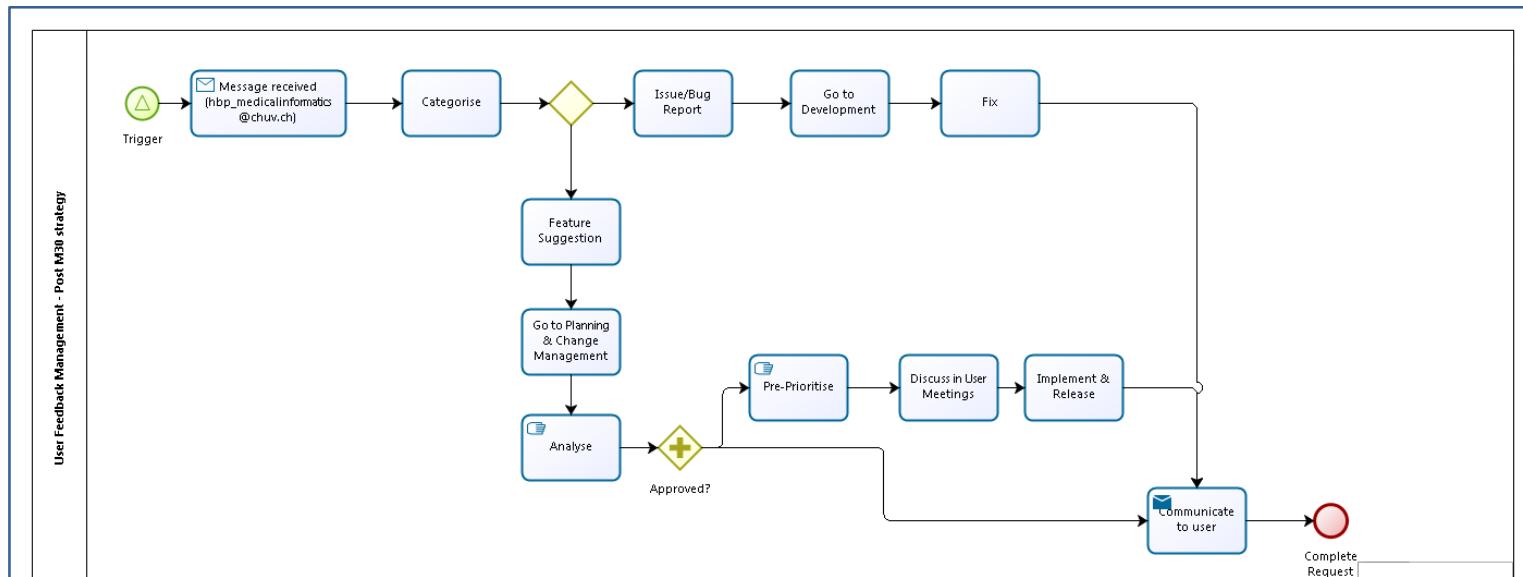


Figure 11: Process of managing and incorporating feedback beyond M30



7.1 User Feedback Received Month 18 - Month 30

The MIP is a sophisticated ICT system that consists of various conceptual layers, and the Web UI is just one small part. Therefore, although releases happen very often, they do not always affect the end user directly (e.g. automation of an infrastructure back-end software process). The tables in the next section describe only user feedback received on the relevant releases on the Web UI.

WP8.4 was responsible for managing the releases & user feedback mentioned below (feedback below also described in the Testing and Quality Strategy document available at <https://mip.humanbrainproject.eu/help/testing-quality>).



Annex A: Platform Architectural Diagram

A.1 The Medical Informatics Platform

Overall the Platform is composed of two main building blocks:

- The first is the Web Portal for research services (Epidemiological Exploration, Interactive Analysis and Biological Signature of Diseases). It provides access to data hosted in hospitals (via the Hospital Bundle) and data from biobanks, public databases and research databases. The Web Portal provides the connection to analytical services (data mining servers) via a microservice architecture.
- The second is the Hospital Bundle, a software stack that will run at every participating hospital or medical centre of the Federated Network of Hospitals and Centers (FNHC) of the MIP.

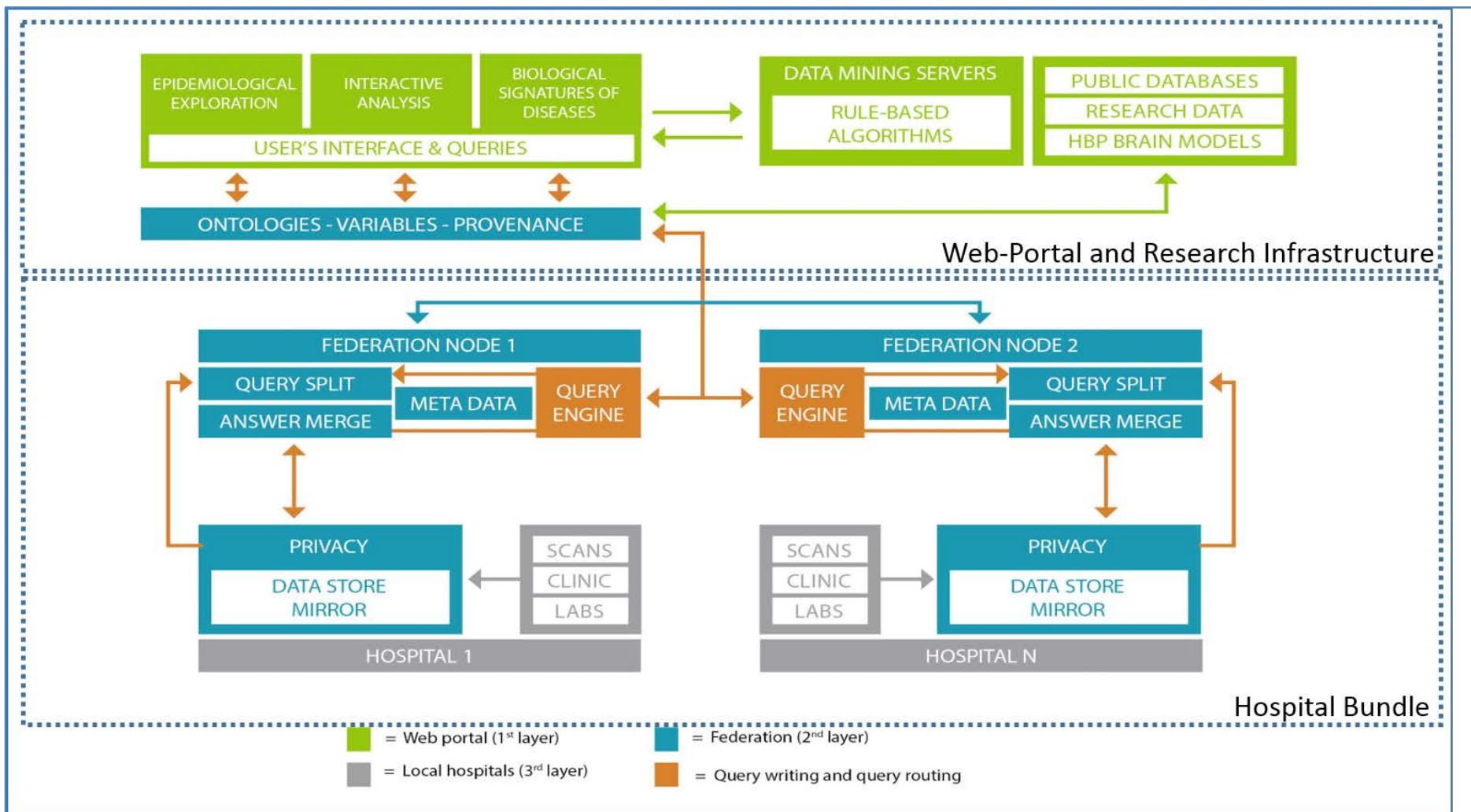


Figure 12: Architectural Diagram of the Medical Informatics Platform



A.2 The Medical Informatics Web Portal and Research Services

The Medical Informatics Web Portal and Research Services have been developed by WP8.2, WP8.4 and WP8.5. They allow the users to connect to the MIP and access the main user services. The web portal is the main user interface for the MIP. It provides an easy-to-use interface that allows researchers to explore the data coming from research sources and hospitals, build models, perform online analytics on the data and to publish and share results.

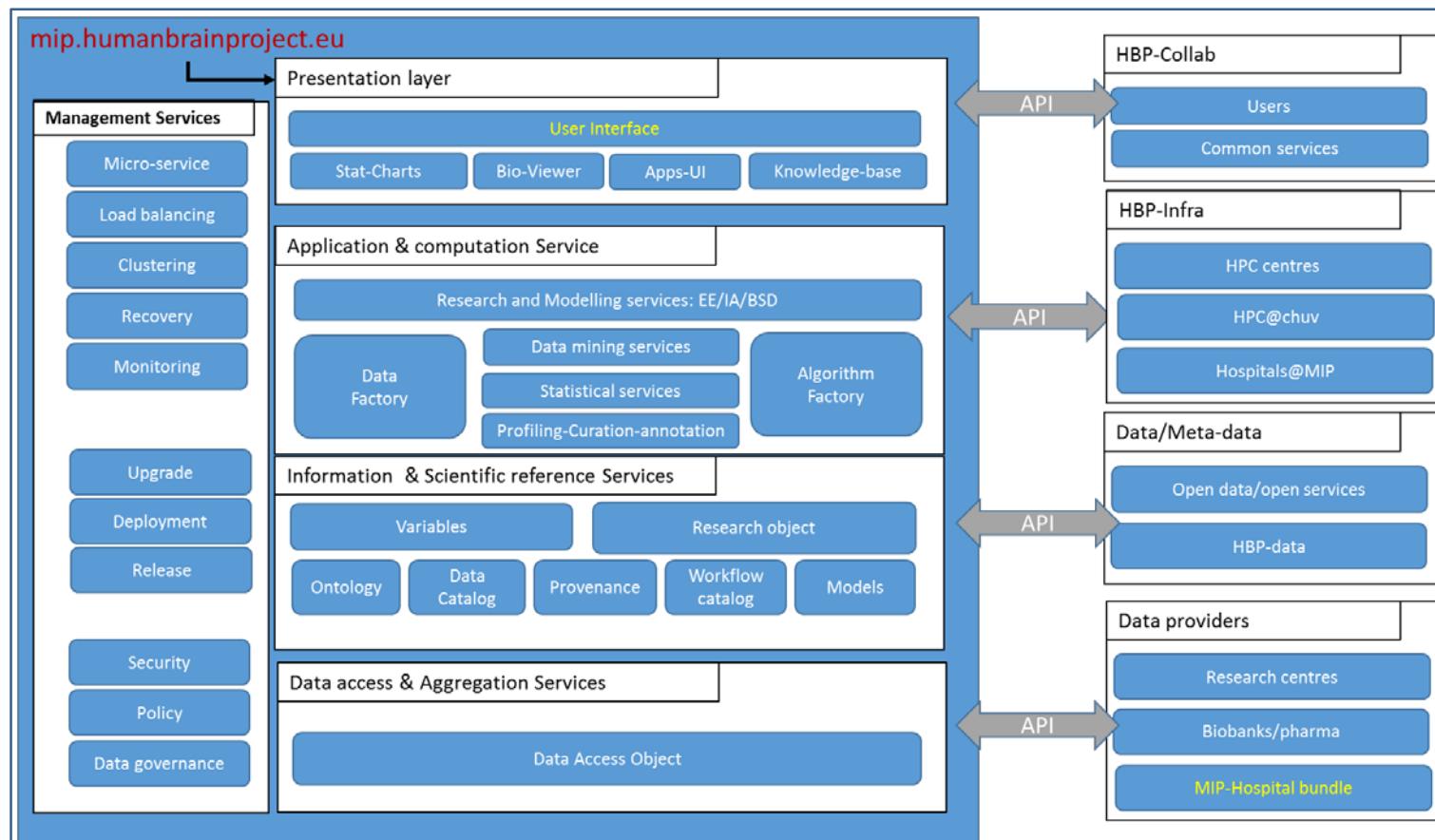


Figure 13: MIP Web Portal Architecture Diagram

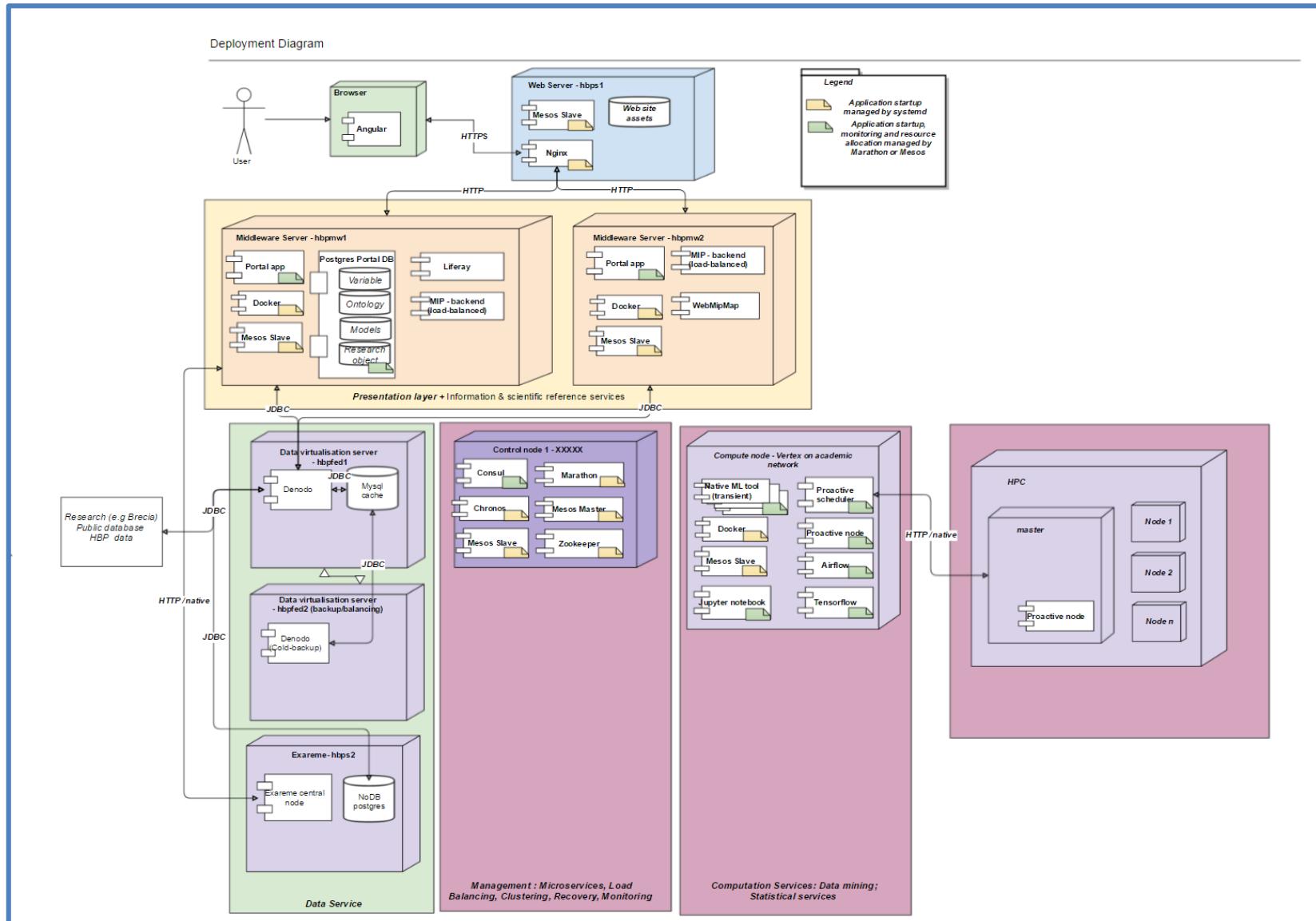


Figure 14 Logical and physical architecture at M30



The following is a list of the components of the back-end services developed by WP8.2, WP8.4 and WP8.5.

Research services:

Access to Epidemiological Exploration, Interactive Analysis and Biological Signature of Diseases, and connection to analytical services and data mining servers via a microservice architecture.

It also provides access to data hosted in hospitals (via the Hospital Bundle) and data from biobanks, public databases and research databases. The detail of the technology used is described in [Annex B](#). The portal back-end provides the services consumed by the front-end.

The MIP User Knowledge Base (KB) is a platform especially designed for the users of the MIP, to:

- provide users with an easy way of accessing the information about the MIP (guidelines),
- provide feedback and interact with MIP administrators, developers and team.

The KB has followed the functionality/releases of the MIP in terms of content and functions (i.e. the external users will be given access and have an environment set up for recording own guidelines, once the MIP is open to external developer users).

- In the long run, the objective of the KB is to provide a flexible tool, fully integrated in the MIP, allowing remote attendance of courses, tutorials, hand-on-sessions, chat, forum, discussion boards, and ad-hoc quizzes to test the increase of knowledge of MIP end-users (i.e. neuroscientists, physicians, researchers, as well as general public).

Management Services: The Microservice architecture provides support for :

- user management and login
- reference data
- controlled access to the various back-end systems, deployment of microservices in a cluster
- monitoring
- recovery.

Application and computation services:

- The data factory includes tools to
 - start, execute and monitor the feature extraction algorithms from MRI data (spatial registration and atlasing)
 - start, execute and monitor the feature extraction algorithms from genetic data (plink tools).
- The algorithm factory provides tools to



- execute the self and custom data analytics algorithms written in a variety of languages and platforms.
- Algorithms are exposed as on-demand web services.
- Algorithms can be executed anywhere in the HPC and in the hospitals.

The Information & Scientific reference Services include:

- Services related to the meta-data
- Services related to the Variable (ontology)
- The scientific workflows: the data used, the algorithms, the models created and the results saved as Research Objects (RO).
- The semantic language for the description of the meta-data is based on RDF and Json-ld.
- The RO contains all the provenance information needed to reproduce the results. RO (<http://www.researchobject.org/>) are saved as a publication and shared by default.
- The standard for the description of results is based on PFA (PFA - Portable Format for Analytics - Data Mining Group dmg.org/pfa/).
- Ontology services will be connected to the knowledge graph of SP5 and other external resources using open-data and open-services.

Data access & Aggregation Services: Data access layer that provide access to:

- the data from hospitals
- the data from biobanks
- the data from research cohorts
- the data from epidemiological cohorts


Release Version 1

Table 2: Summary of MIP first release

Month/version	M18/v0.1 (POC)	
URL version tested	http://hbps1.chuv.ch/webgraph/#/	
Summary of functionality	Basic functionality and UI design. Purpose: to get 1 st feedback from users. <ul style="list-style-type: none"> • View Manager (create « views », share « views ») • Basic plots • Saved Filters (create and share filters) • Variables and Data - minimal (flat files, not distributed) 	
Method of validation	Quality Review Technique (see Quality Strategy for details) - early validation of incipient functionality	
Feedback received		
Users	Functionality Feedback (main)	Usability & Ease of use
Internal: neuroscientist, statistician, clinical psychologist, neurologist	(+) liked the concept of creating "views" and filters, and sharing them with community (-) suggested to be able to search easier for variable/ontology; also to have a larger library (-) suggested to have a library of methods, to run them as see the detailed statistics; also a library of plot types	(-) basic design, fields need described better, as well as the flow in the system; not very clear what the steps are
Incorporated Results		
1) Feedback incorporated and new specification (see new spec at Knowledge Base > Project Documentation > Testing & Strategy section at https://mip.humanbrainproject.eu/help/)		
2) New upgraded UI release v0.3		



<https://hbps1.chuv.ch/mip/#/intro>

7.1.1 Release Version 2

Table 3: Summary of MIP second release

Month/version	M24/v0.3
URL version tested	https://hbps1.chuv.ch/mip/#/intro
Summary of functionality	<p>Software Framework</p> <ul style="list-style-type: none"> Connection to 5 data sources (CHUV, 3C, INDI, EDSD, ADNI) & distributed computation using R and User Defined Functions (UDFs). This release was the 1st step (POC) towards the federated infrastructure, as the 5 data sources were located on 3 different servers, in 2 different physical locations (Switzerland - Lausanne (CHUV) and Italy - Brescia). Reproduction of a published scientific study (<i>NeuroImage</i> 78 (2013) 474-480) see http://www.ncbi.nlm.nih.gov/pubmed/23624169 (same ADNI dataset used, same method applied (General Linear Regression), same display method - boxplot obtained, and resulting effect compared - see Annex I). Apps integration: Genexpression, "Sunburst" graph on distributed nodes <p>Web UI:</p> <ul style="list-style-type: none"> Reviewed flow and design of Web UI Interactive Analyses service User and Community dashboards (aka “views” in previous release)
Method of validation	<ol style="list-style-type: none"> Quality Review Technique - direct involvement in functionality specification, early validation of incipient functionality UAT
Feedback received	
Users	Functionality Feedback (main)
	Usability & Ease of use



Internal: neuroscientist, statistician, clinical psychologist, neurologist	(-) inability to vary the use case scenarios too much (hence get more variables and data in) (-) plots not displaying data correctly (+) much improved dashboards (+) much improved designs	(-) difficult to select variables from drop-down menus; also not always understanding the variables' name (hence description to be added) (+) much improved navigation flow
Incorporated Results		
3) Feedback incorporated and new specification (see new spec at Knowledge Base > Project Documentation section at https://mip.humanbrainproject.eu/help/)		
4) New upgraded UI release v1.0 https://mip.humanbrainproject.eu/		
Project Lessons Learnt		
- Each connectivity might have particularities, connection problems might arise, which may impact time and quality. - Clear process and connection requirements should be known in advance, discussed with the local data provider. This will also help to plan and estimate the work.		

7.1.2 Release Version 3

Table 4: Summary of MIP third release

Month/version	M30/v1.0 (public release)
URL version tested	https://mip.humanbrainproject.eu/
Summary of functionality	Software Framework: <ul style="list-style-type: none"> • Computational Infrastructure (Algorithm Factory) completed, packaged and integrated (See Annex B for full technical details).



	<ul style="list-style-type: none"> The algorithm factory provides the tooling to execute of the self and custom data analytics algorithms written in a variety of languages and platforms. Algorithms are exposed as on-demand web services. Integration of R algorithms: Creation of machine and human readable standard formats for description of statistical results of algorithms (to be further used in general UDFs). Data Factory framework iteration #1 (preprocessing). All functionality generalised (all mocks replaced) and packaged. <p>Integrated into MIP product:</p> <p>Web UI (web front- and back-end):</p> <ul style="list-style-type: none"> Much improved usability and UI flow, inc. reorganisation of APPS section, addition of Epidemiological Exploration (EE) Variable Dictionary Added Descriptive Statistics in the Epidemiological Exploration Improved Interactive Analysis (IA) - Model Design, Model Exploration, improved Visualisation of results (Design Matrix, tables stats) Corrected Display of information in plots Added Design Matrix as visualisation option All functionality generalised (all mocks replaced) and packaged Corrected bugs from previous versions Created and incorporated Terms of Services and Disclaimer Designed and added onboarding pages Integrated with User Knowledge Base (with single-sign-on) <p>Additional unit developments completed (to be integrated into MIP):</p> <p>Hospital Bundle:</p> <ul style="list-style-type: none"> WebMIPMap, MipMapRew, MIPMap Hospital Bundle integration (Anonymization, MIPMap, Raw, Exareme)
Method of validation	<ol style="list-style-type: none"> Quality Review Technique - direct involvement in functionality specification, early validation of incipient functionality UAT



Feedback received		
Users	Functionality Feedback	Usability & Ease of use
Internal: neuroscientist, paediatrician/data mining scientist	<ul style="list-style-type: none"> (+) much improved functionality (+) generic results of applied methods (+) fixed Apps and added description (e.g. for 3C) (-)₁ add description on variables in EE (bubbles), also on mouse-over (-)₂ ideally explain what My Data and My Community contain (-)₃ EE: to improve the grouping of variables in EE (explorative variables in "bubbles") (-)₄ Apps: Brain 3D viewer: measurements are outside the 3D brain (+) able to see the huge future potential 	(+) Excellent, much improved usability, easy to explore, a lot to explore
Incorporated Results		
<ul style="list-style-type: none"> (-)₁, (-)₂, (-)₃ fixed/improved and included into release (-)₄ to be included post M30 		

A.3 Hospital Bundle

The MIP-Hospital Bundle includes software for schema and data integration, *in situ* querying, federated querying, and dataflow processing. Each hospital will be a node of the MIP, and will communicate through appropriate modules in the hospital bundle with all other hospital nodes. The described development is done in part by T8.4.1 and integrates all components from the Work Package WP8.1 into a fully-distributed MI Hospital Bundle. The bundle consists of the following major components.

The ***in situ* query engine, RAW**. The engine serves as a database back-end at each participating hospital and is responsible for executing queries on raw hospital data.

The **schema mapping and data exchange tool, MIPMap**. MIPMap is used in the Hospital Bundle as a declarative ETL tool (Extract, Transform, Load) that translates data provided by participating hospitals to the MIP schema, thus populating the Local Data Store Mirror of each hospital.



Additionally, it is utilized in the translation of the research data, used in the MIP, to also integrate them to the MIP schema, making them interoperable with other MIP data, hence also used in the Web Portal. The tool offers an easy-to-use graphical interface where the user, given a source and target schema, can define correspondences/mappings by simply drawing arrow lines between the elements of different schemas. These mappings are executed using a data exchange engine and the data produced are the hospital data that conform to the MIP schema. The AUEB team is responsible for the tool and deploying the data warehousing preparation workflow. Since the last reporting period (M18), the prototype MIPMap has been developed to be a stable software application, integrated into the hospital bundle. It has been utilized in collaboration with WP8.2 to translate the available MIP data (CHUV, ADNI, INDI, and ESDS) to the MIP schema, thus creating respective LDSMs. A detailed evaluation of the performance of MIPMap can be seen in Annex E - Performance indicators.

Supporting MIPMap's functionality, the AUEB team has developed WebMIPMap (See Annex B), a web interface that can create mapping tasks which can later be downloaded and run on the desktop version, i.e. MIPMap. The use of WebMIPMap, however, is not restricted to the bundle. It is provided as a service in the SP8 Web Portal to allow users to create mappings between the MIP schema and hospital research data schemata and/or ontologies. These mappings are used to translate terms of the MIP schema to hospital and research data schema terms in case their Local Data Store Mirror does not fully comply with the MIP schema. To achieve this the AUEB team has also developed MIPMapRew (see Annex B), a service for rewriting queries posed at the Web Portal so that the nomenclature used by its predicates conforms to the schema of the hospital/research centres. Finally, WebMIPMap can be used to map MIP schema to existing ontologies.

The distributed and privacy preserving processing system, EXAREME. The system offers a declarative language which is based on SQL with user-defined functions (UDFs) extended with parallelism and data pipeline primitives. The UoA team implements this federated querying system and is responsible for the distributed workflow component of the bundle. Currently, Exareme is integrated with RAW and executes a variety of distributed algorithms in a privacy preserving manner. The integration with RAW and with the web portal have been the major developments since M18. Depending on changes that may happen with RAW or the web portal, further work might be needed on finalizing and testing the integration.

The anonymization and query filter module, Anonymizer. The goal of the anonymizer module deployed at a hospital is to (a) remove all data exported from the hospital's systems from personal identifiers, (b) check that incoming queries conform to privacy standards, i.e., the columns they read are limited, and (c) strip all query results from personal identifiers as well.

Based on the above, the subcontractor (Gnubila) has developed the anonymization module including the following features:

- Anonymizer: blacklist/whitelist manager, DICOM anonymizer, full text anonymizer.
- Query Filter: Query Parser, Black Listed Fields exclusion, Aggregation Field check.
- Response Cleaner: Privacy Information Webservice, Cleaner Process.

Federated Workflow: Coming from the web portal, queries and analyses will first run through an Exareme instance (at any hospital), which distributes them to other Exareme instances (at other hospitals). At every hospital, the queries will run through the query filter to ensure that only allowed queries are passed on. If allowed, the queries are forwarded to RAW that is used to access the hospital data *in situ*. Once RAW returns the aggregated results, they are sent back to Exareme, which collects all results from different hospitals.

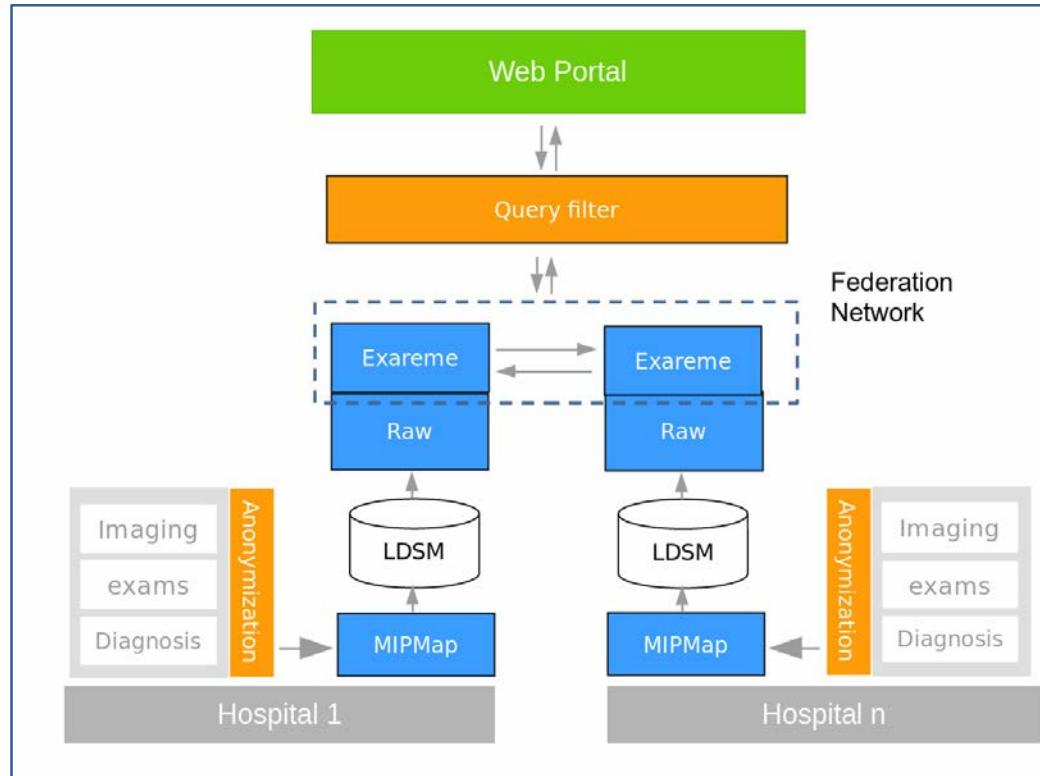


Figure 15: Hospital bundle architecture overview

The distributed privacy preserving processing engine (EXAREME¹) is an open source project supported by the Management of Data, Information & Knowledge Group (MADgIK group²) at University of Athens (UoA). The system offers a declarative language which is based on SQL with user-defined functions (UDFs) extended with parallelism and data pipeline primitives. The system architecture is shown below.

¹ <http://www.exareme.org/>

² <http://www.madgik.di.uoa.gr/>

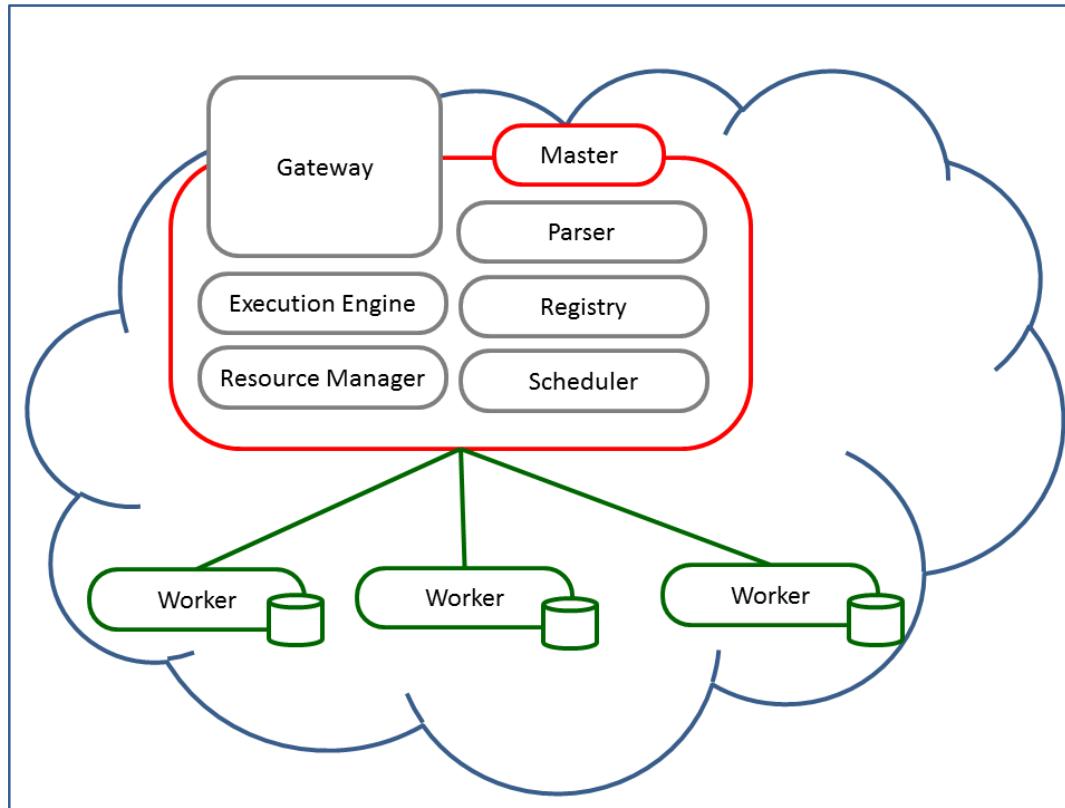


Figure 16: Exareme architecture overview

Exareme is separated into the following components: The Master is elected from the worker pool and is the main entry point, through the gateway, to the system. The Master is responsible for the orchestration of all the components. The Execution Engine communicates with the resource manager and schedules the operators of the query respecting their dependencies in the dataflow graph and the available resources. It also monitors the dataflow execution and handles failures. All the information related to the data and the allocated resources is stored in the Registry. The Resource Manager is responsible for the allocation and deallocation of resources on each node. The Optimizer/Scheduler engine translates a high level query into the distributed machine code of the system and creates the final execution plan by assigning operators to workers. Finally, the Worker executes operators (relational operators and UDFs) and transfers intermediate results to the master. Madis³ is

³ <https://github.com/madgik/madis>



the core engine of the Worker, and is a wrapper of SQLite based on the python APSW. Madis processes the data in a streaming fashion and performs pipelining when possible, even for UDFs. The UDFs are executed inside the database along with the relational operators to push them as close to the data as possible.

In the context of MIP bundle, Exareme acts as the federation layer. This layer is responsible for the communication of each hospital and web portal. It does not allow communication amongst hospitals. Worker components are deployed in the local layer on each hospital node and act as connectors with the RAW query engines. The Master component acts as the global layer and can be deployed in one or more hospital nodes.

Exareme currently supports the following functionalities:

- Get list of the available algorithms.
- Submit any of the available algorithms for execution.
- Get the execution status of a submitted algorithm.
- Get the execution results of a completed algorithm.

We reviewed other similar systems before choosing Exareme and RAW. All of these systems have extensive complexity and different trade-offs. DataShield offers a variety of modelling and statistics functions implemented in a distributed non-disclosive summary statistics way. Exareme/Raw follow the same concept. The main difference is that instead of using R with C/C++ extensions on top of a database like DataShield does, Exareme/Raw uses the database (extended with Python UDFs) for all computational needs. Developing new functionality on Exareme only requires a few SQL queries and small Python functions. In such a manner and in a short time, we have implemented a number of complex statistical/modelling functions which follow the distributed non-disclosive summary statistics concept and preserve privacy.

A.3.1 Hospital bundle deployment

The process to prepare and set up a running LDSM at a new hospital is as follows:

- 1) Ensuring all the confidentiality and ethical documentation is in place and, when required, to perform testing on non-anonymised data samples.
- 2) Discussing data selection (focus on neurodegenerative diseases) and collecting as many variables as possible from each subject (with focal point on brain imaging data).
- 3) Performing an anonymization test (whether through Gubila software provided by SP8 or a software chosen by the hospital).
- 4) Mapping tasks from hospital data to MIP schema data (performed by a clinician who understands the semantics of the data together with the HBP engineer).
- 5) Coordination with the IT department:



- a) To get the right authorization to connect to the relevant systems
 - b) To define the data import policy (frequency, if mirrors are needed)
 - c) To install the anonymization server and the LDSM
- 6) After all of this, the last tests are performed, data is imported from the hospital and the LDSM is connected to the federation network.

The **hardware requirements** for the installation of the federation software bundle in a member hospital are as follows:

- 1) At least two (2) servers with 32G RAM, four (4) cores
 - a) Note: The number of servers actually needed will depend on the currently available infrastructure and on the amount of data.
 - b) Note: If there is need for *in situ* feature extraction, the corresponding software will need at least 15 GB RAM (recommended: 64GB) and four (4) cores.
- 2) Computer network access with three (3) computer ports available, two (2) for internal communication in private network between hospitals (RMI, HTTP) and one (1) for external communication with the web portal (HTTP).
 - a) Note: A fourth port can be added for remote administration, if requested.



Annex B: Software and Services Included in this Platform Release

This annex gives details of each Software Package and Service of the MIP.

B.1 Presentation Layer

Product/Software Package/Service name: Web Portal

Metadata

Category	Application
Maintainers	CHUV
Homepage	https://github.com/LREN-CHUV/portal-frontend
License	Apache 2.0 License
Current Version	0.1
All Versions	

Description

The web portal is the main user interface for the MIP.

It provides an easy-to-use interface that allows researchers to explore the data coming from research sources and the hospitals, to perform online analytics on that data and to write and share papers.



Technologies used

- 1) User interface: angular, bootstrap, Nginx
 - Stats-Charts: highCharts d3.js
 - Density distribution
 - Time-series
- 2) Bio-viewer:
 - MRI based Atlas 2D and 3D (three.js version 0.7) for
 - Gene expression atlas, Gene Browser, GWAS Manhattan plot
 - Maximum intensity projection Map
- 3) Apps UI:
 - Nodejs, Grunt

**Product/Software Package/Service name: MIP User Knowledge Base****Metadata**

Category	User Guidelines
Maintainers	WP8.4 - WP8.5
Homepage	https://mip.humanbrainproject.eu/help/
Documentation	https://www.liferay.com/
Support	hbp_medicalinformatics@chuv.ch
Source Code	Liferay
License	No-IPR
Current Version	V1.0

[All Versions](#)**Description**

The MIP User Knowledge Base (KB) is a platform especially designed for the users of the MIP, to provide them with an easy way to access information about the MIP (guidelines), provide feedback, and interact with the MIP administrators, developers and team. The KB has followed the functionality/releases of the MIP in terms of content and in terms of functions (i.e. the external users will be given access and have an environment set up for recording own guidelines, once MIP will be open to external developer users).

In the long run, the objective of the KB is to provide a flexible tool, fully integrated in the HBP MIP, allowing remote attendance of courses, tutorials, hand-on-sessions, chat, forum, discussion boards, and ad-hoc quizzes to test the increase of knowledge of the MIP end-users (i.e. neuroscientists, physicians, researchers, as well as general public).



Technologies

The KB is based on Liferay technology. It is an open-source web standard mainly written in Java. It includes built-in web Content Management System and it has been highly configurable for building customized websites/portals as well as for assembly of themes, html pages, web contents, portlets and widgets. The KB is supported as live interactive content where the end-user has an active role.

B.2 Application & Computation Service

Product/Software Package/Service name: Research and Modelling services: EE/IA/BSD

Metadata

Category	Application
Maintainers	CHUV
Source Code	https://github.com/LREN-CHUV/portal-backend
License	Apache 2.0 License
Current Version	0.1

All Versions

Description

The portal back-end provides the services consumed by the front-end. This includes user management and login, reference data, controlled access to the various back-end systems.

Technologies used

- 1) Custom Java application (portal-back-end)



- a) REST API
 - b) Swagger documentation of the web services
- 2) Postgres Database

Product/Software Package/Service name: Algorithm Factory

Metadata

Category	Application
Maintainers	CHUV
Homepage	https://github.com/LREN-CHUV/woken
Source Code	https://github.com/LREN-CHUV/woken
License	Apache 2.0 License
Current Version	0.1
All Versions	

Description

The algorithm factory provides tools to execute the self and custom data analytics algorithms written in a variety of languages and platforms. Algorithms are exposed as on-demand web services. At its core, the application today is composed of woken, an engine for on-demand analytics



that launches Docker containers and collects their results, a few scripts in R encapsulated in Docker images and producing PFA models, and a web portal allowing the user to select data, build analytical model from a limited set of algorithms and view the results.

Technologies used

- 1) Custom Scala application ([woken](#))
 - REST API
 - Swagger documentation of the web services
 - Possible future merge with [Cromwell](#) from the Broad Institute
 - Using the [PFA](#) standard to represent, exchange and store models
- 2) [Chronos](#) - externally maintained application
- 3) Docker-based framework for the inclusion of algorithms implemented in various technologies ([mip-docker-images](#))
- 4) Docker images embedding the algorithms ([functions-repository](#)), published on public [Docker Hub repository](#) and on a on-premise Docker registry.
- 5) Algorithms coming from standard R distribution, Weka, our partners (CCC strategy from TAU team, rule based algorithms from JSI, BH-tSNE algorithm from LUMC)
- 6) Postgres Database
- 7) Data Mining services
 - Currently using the workflow application for on-demand calculations
 - Future: Apache Spark + Jupyter notebook for more advanced use
 - Future: HPC analytics services
- 8) Statistical services
 - Currently using stats extracted by Rapid Miner
 - Integration with Exareme
- 9) Profiling, curation, annotation
 - Currently: ad-hoc R scripts, Spring batch
- 10) Data Factory:
 - Custom Spring Batch application



- Feature extraction: Matlab + SPM + [HPC@chuv](#) + Activeeon Proactive scheduler
- Future: XNAT

B.3 Management services

Product/Software Package/Service name: Microservice architecture

Metadata

Category	Application
Maintainers	CHUV
Homepage	https://github.com/LREN-CHUV/mip-microservices-infrastructure
Source Code	https://github.com/LREN-CHUV/mip-microservices-infrastructure
License	Apache 2.0 License
Current Version	0.1
All Versions	

Description

The Microservice architecture provides support for deploying microservices in a cluster, monitoring and recovery. It is a custom version of

Cisco's [Mantl](#) running on Ubuntu servers and includes Apache Mesos, Marathon, Zookeeper, Docker. In the future, we plan to use Consul, dns-masq, Vault, Logstash, and Traefik. It allows the control of a cluster of machines as a single machine and simplifies operations. Cisco is able to use this architecture over several data centres and this could be used as the technical backbone to connect the data centre at CHUV with other data centres.

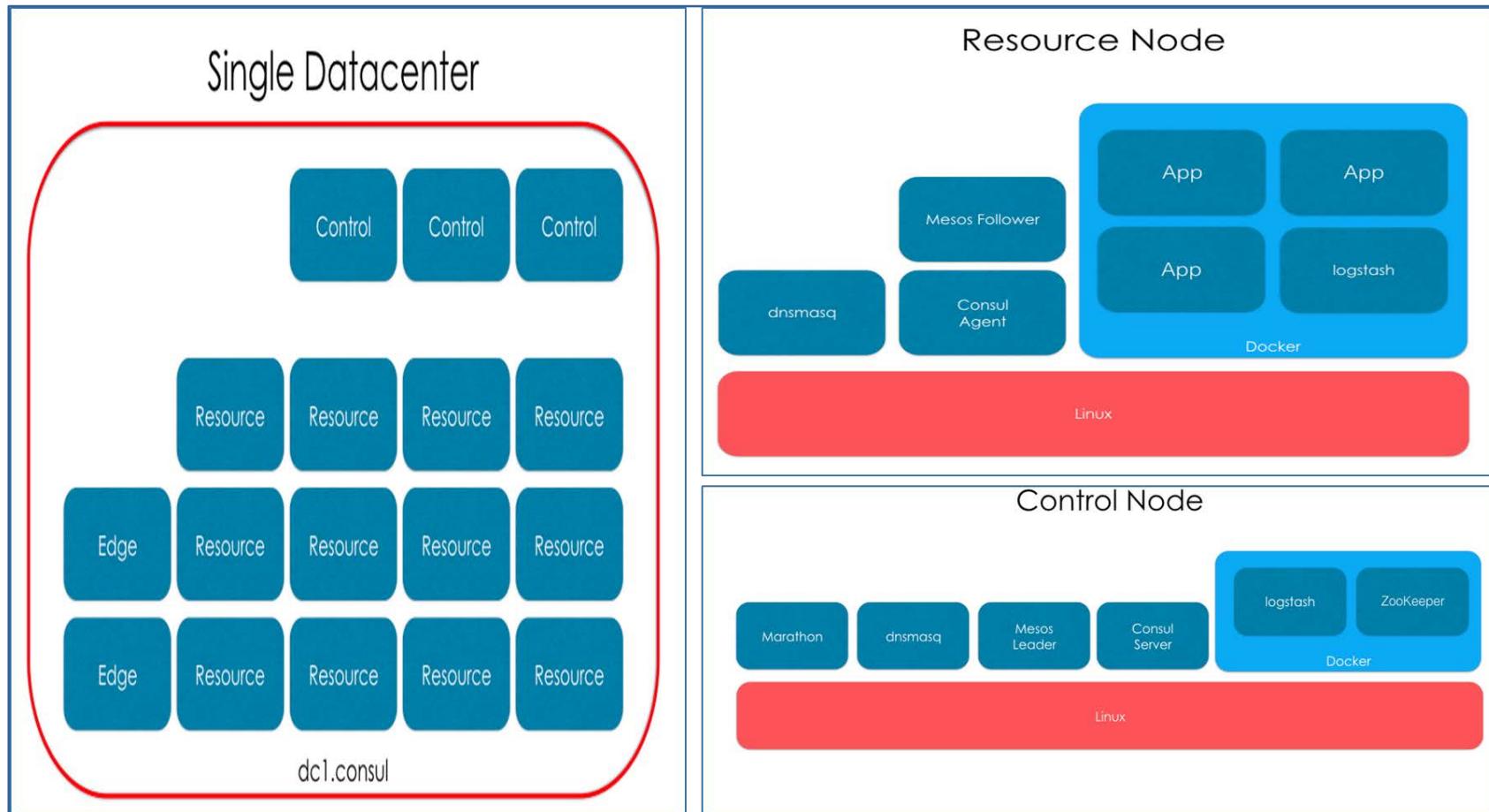


Figure 17: Architecture of a cluster in the microservice infrastructure

In a data centre, we organize a pool of computers to work together and form a cluster.



Each machine is given a role: control node, resource node or edge node.

A control node contains only the software necessary to manage a data centre: Mesos to distribute the work between machines; Zookeeper to keep track of what is happening in the cluster; Marathon to control long running processes such as databases, application servers and web services. Resource nodes are where all work gets executed, e.g. algorithms running inside Docker containers and launched by the Algorithm Factory, long running services such as databases and micro services.

1) Monitoring

- Marathon
- collectd
- Future: Consul

2) Deployment, upgrade, release

- Jenkins
- Ansible + environment specific scripts derived from mip-microservice-infrastructure (prod-infrastructure, qa-infrastructure, dev-infrastructure)
- Docker

3) Security

- Network security provided by IT team at CHUV
- Hardening of servers provided by Ansible scripts (first-five-minutes, unattended-upgrades, firewall) installing and configuring Logwatch, Fail2ban, UFW firewall
- HTTPS certificate provided by Letsencrypt.org+ automatic certificate renewal scripts installed on Nginx servers
- Anonymised users on production servers
- Future: adoption of [CIS best practices](#)

4) Policy, data governance



B.4 Hospital Bundle

Product/Software Package/Service name: Hospital Bundle

Application version

Metadata

Category	Application
Maintainers	Cesar MATOS , Alexandros PAPADOPOULOS , Eleni ZACHARIA , Tassos VENETIS , Giannis KAZADEIS
Homepage	https://github.com/HBSP8Repo
Documentation	https://github.com/HBSP8Repo
Support	Cesar MATOS , Alexandros PAPADOPOULOS , Eleni ZACHARIA
Source Code	https://github.com/HBSP8Repo
License	Mixed
Current Version	0.1
All Versions	0.1



Description

The Hospital Bundle consists of the following major components: the *in situ* query engine, RAW, the schema mapping and data exchange tool, MIPMap, the distributed and privacy preserving processing system, EXAREME, and the anonymization and query filter module, Anonymizer, presented below.

**Product/Software Package/Service name: Exareme**

Federation engine

Application version 0.1

Metadata

Category	Application
Maintainers	MADgIK group, University of Athens
Homepage	http://www.exareme.org
Documentation	http://madgik.github.io/exareme/ https://github.com/madgik/exareme/wiki/Running-mip-algorithms-on-exareme
Support	exareme-support@googlegroups.com
Source Code	https://github.com/madgik/exareme/
License	MIT License
Current Version	0.1
All Versions	application



Description

EXAREME is a distributed privacy preserving processing engine. The system offers a declarative language based on SQL with user-defined functions (UDFs) extended with parallelism and data pipeline primitives. In the context of the MIP bundle, Exareme acts as the federation layer. This layer is responsible for the communication of each hospital and web portal. It does not allow communication amongst hospitals. Worker components are deployed on each hospital node and act as connectors with the RAW query engines. The Master component merges the partial hospital results and can be deployed in one or more hospital nodes.

Technologies used

Exareme uses JDK 1.7, Python 2.7 and SQLite 3.9.2.

**Product/Software Package/Service name: Schema Mapping and Data Exchange (MIPMap)**

MIPMap and WebMIPMap provide interfaces for their users. MIPMap is tool that provides a user interface to data providers in order to allow them to translate their data to the MIP schema, while WebMIPMap provides a user interface to the MIP Web Portal users, allowing them to create mappings following the use cases described in the WebMIPMap description section.

Schema mapping and data exchange tool

Tool version 1.0

Metadata

Category	Tool
Maintainers	Giannis KAZADEIS (AUEB, P3) Tassos VENETIS (AUEB, P3)
Documentation	https://github.com/aueb-wim/MIPMap
Support	g.kazadeis@gmail.com
Source Code	https://github.com/aueb-wim/MIPMap
License	GNU GENERAL PUBLIC LICENSE Version 3
Current Version	1.0
All Versions	1.0



Description

MIPMap is a schema mapping and data exchange tool specifically tailored for the needs of the HBP. It offers an easy-to-use graphical interface where the user, given a source and target schema, can define correspondences/mappings by simply drawing arrow lines between the elements of the two tree-form representations. In order to specify these mappings we use declarative representations, under the formalism of Tuple-generating dependencies (TGD), that are executed using an advanced, highly scalable and efficient mapping execution engine designed to deal with the complexity and size of HBP data transformations.

MIPMap is used in the Hospital Bundle as a declarative ETL tool (Extract, Transform, Load) that translates data provided by participating hospitals to the MIP schema, thus populating the Local Data Store Mirror of each hospital. Additionally, it is utilized in the translation of the research data used in the MIP, to also integrate them to the MIP schema, making them interoperable to other MIP data.

MIPMap's requirements are:

Operating System: Windows 7 (or newer) 64-bit, Linux, Mac OS

RAM: 4GB or higher

Java version: 7.0 or higher

MIPMap also requires the installation of a Postgres database server (version 9.2+) with a user with administration rights. Access to the database can be configured through a properties file after the installation of MIPMap.

Technologies used

Java 8

**Product/Software Package/Service name: WebMIPMap**

Web interface of the MIPMap tool

Service version 1.0

Metadata

Category	service
Maintainers	Giannis KAZADEIS Tassos VENETIS
Documentation	https://github.com/aueb-wim/WebMIPMap
Support	g.kazadeis@gmail.com
Source Code	https://github.com/aueb-wim/WebMIPMap
License	GNU GENERAL PUBLIC LICENSE Version 3
Current Version	1.0
All Versions	1.0

Description

WebMIPMap is a web interface of the MIPMap tool. It is provided as a service in the SP8 Web Portal to allow users to create mappings between the MIP schema and hospital research data schemata and/or ontologies. These mappings are used to translate terms of the MIP schema to hospital and research data schema terms in case their Local Data Store Mirror does not fully comply with the MIP schema. Additionally, it can



be used to map MIP schema to existing ontologies. Moreover, WebMIPMap can be utilized as a web interface to create mapping tasks that can run on the desktop version (MIPMap).

It is obvious that the use of WebMIPMap is similar to the use of MIPMap; however, WebMIPMap, due to the no move no copy policy imposed by SP8, does not support data translation. Nonetheless, it can be used to create a mapping task that can be downloaded and executed on MIPMap.

The figure below shows a screenshot of the user interface of the tool where the mapping task presented in the MIPMap tutorial is loaded.

Technologies used

- Javascrip
- Java Servlets
- Java.

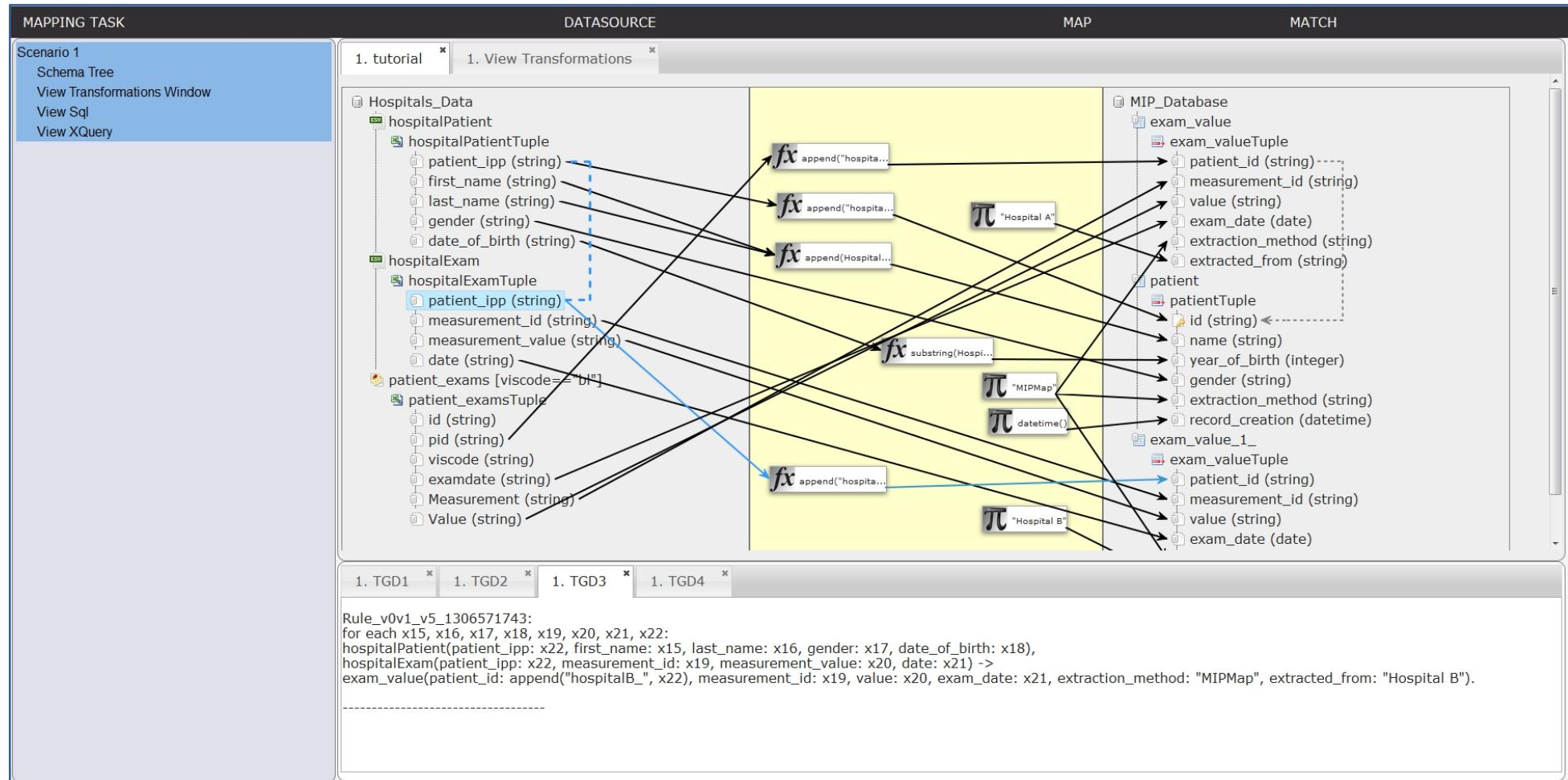


Figure 18: WebMIPMap User Interface

**Product/Software Package/Service name: MIPMapRew**

Complementary service to MIPMap tool to rewrite queries

Service version 0.7

Metadata

Category	Service
Maintainers	Giannis KAZADEIS Tassos VENETIS
Support	g.kazadeis@gmail.com
Source Code	https://github.com/aeub-wim/MIPMapRew
License	GNU GENERAL PUBLIC LICENSE Version 3
Current Version	0.7
All Versions	0.7

Description

MIPMapRew is a service for rewriting queries posed at the Web Portal so that the nomenclature used by its predicates conforms to the schema of the hospital/research centres. As stated in the WebMIPMap description, it is highly likely that some hospitals/research centres might not fully comply with the MIP schema. In such cases WebMIPMap will be used to create the mappings between the two schemata and MIPMapRew will be responsible to rewrite the original query (created at the Web Portal), with the terms and nomenclature followed by the participating hospital so that the appropriate results can be collected. Note, however, that such a scenario will not be very common at the beginning of the Project since, according to the specifications of the MIP, all the hospitals and research centres that are going to contribute data to the Platform will use the common MIP schema in order to make their data available.



Technologies used

- SQL
- Java.

Product/Software Package/Service name: NoDB/RAW

First version of the query engine deployed at a local data store mirror (CHUV).

Metadata

Category	Application
Maintainers	<u>Cesar MATOS</u> , <u>Manos KARPATHIOTAKIS</u>
Homepage	http://dias.epfl.ch/RAW
Documentation	https://github.com/HBPSP8Repo/SP8LocalMirrorBundle/blob/master/RAW_README.md
Support	<u>Cesar MATOS</u> , <u>Manos KARPATHIOTAKIS</u>
Source Code	https://github.com/HBPSP8Repo/NoDB
License	Missing
Current Version	0.1



Category	Application
All Versions	0.1

Description

The Local data store query engine is a database system specially tailored for the needs of the HBP. Its main purpose is to offer efficient querying services directly on files inside the hospital. The type of research involved, biological signatures of diseases, requires more advanced data structures and more expressive operations than those provided by traditional databases.

The query language enables users to apply powerful transformations over the output of a query. RAW uses a query language, similar to SQL (Structured Query Language) that provides support for a multitude of data models: collection types, hierarchies and multi-dimensional arrays. This flexibility enables queries to transparently access a great variety of datasets (e.g., relational tables, CSV and JSON files, array image data, etc.).

Furthermore, to be able to query from heterogeneous data formats, RAW utilizes code generation techniques. When a query is posed code generation plug-ins are invoked to produce code specific to both the file format and the query posed. Code generation acts as an enabler for queries targeting multiple data formats. In addition, it improves performance, as each query leads to execution of very specific code, avoiding generic methods that would take place (e.g., parsing the data files using a generic parser).

RAW currently supports the following functionality:

- Get list of the available schemas/ registered files patients, exams values, brain features.
- Submit queries and retrieve results.
- Submit paginated queries (streaming).
- Register new files to be queried.

Technologies

The query engine system requirements are:

- Operating system: Linux
- RAM: 8GB or higher
- Java version: 8.0 or higher



- Scala version: 2.11 or higher

More information is available at https://github.com/HBPSP8Repo/SP8LocalMirrorBundle/blob/master/RAW_README.md

Product/Software Package/Service name: Anonymization Module

Metadata

Category	Application
Maintainers	Gnubila
Homepage	https://gnubila.fr
Documentation	https://gnubila.fr
Support	Gnubila
Source Code	Closed
License	By Gnubila
Current Version	0.1
All Versions	0.1



Description

Data to be used in the context of the MIP needs to be anonymized twice, firstly when exported from the hospital systems and secondly when queried.

Data is first anonymized when exported from hospital information systems before it is integrated into the MIP. On export from the hospital systems all personal identifiers are stripped from it, i.e. identifiers (such as name, social security numbers etc.) that allow one to directly infer the identity of the patient are removed. Once this is accomplished, the MIP integrates the data and can answer queries on it.

Any personal identifiers also need to be stripped from query results before they are returned to the Platform (and therewith before sending the results back to the user).

The data to be anonymized is multi-modal, i.e. encompasses MRI data (based on the DICOM⁴ standard), potentially PET, full text (patient files, notes etc.), genetic, proteomic, etc.

Anonymization of the data is to be verifiable, i.e. after anonymization the data needs to be tested to ensure that there are no personal identifiers left.

Technologies used

The anonymization module requires the following:

- a Java 7 JRE/JDK (or greater) is required to run the Anonymizers
- a valid license

The language for the configuration files is YAML ("YAML Ain't Markup Language"). YAML is a human readable data serialization language. Due to the YAML Parser used, some syntax constraints have to be respected:

- Indent using the space character (always use the same number of space characters)
- Do not indent using the tabulation character

More information is available at https://github.com/HBPSP8Repo/SP8LocalMirrorBundle/blob/master/Anonymizer_README.md.

⁴ <http://medical.nema.org/standard.html>



*Product/Software Package/Service name: Administration User Interface***Metadata**

Category	Application
Maintainers	Cesar MATOS
Homepage	https://github.com/HBPSP8Repo
Documentation	https://github.com/HBPSP8Repo
Support	Cesar MATOS
Source Code	https://github.com/HBPSP8Repo
License	The MIT License (MIT)
Current Version	0.1
All Versions	0.1

Description

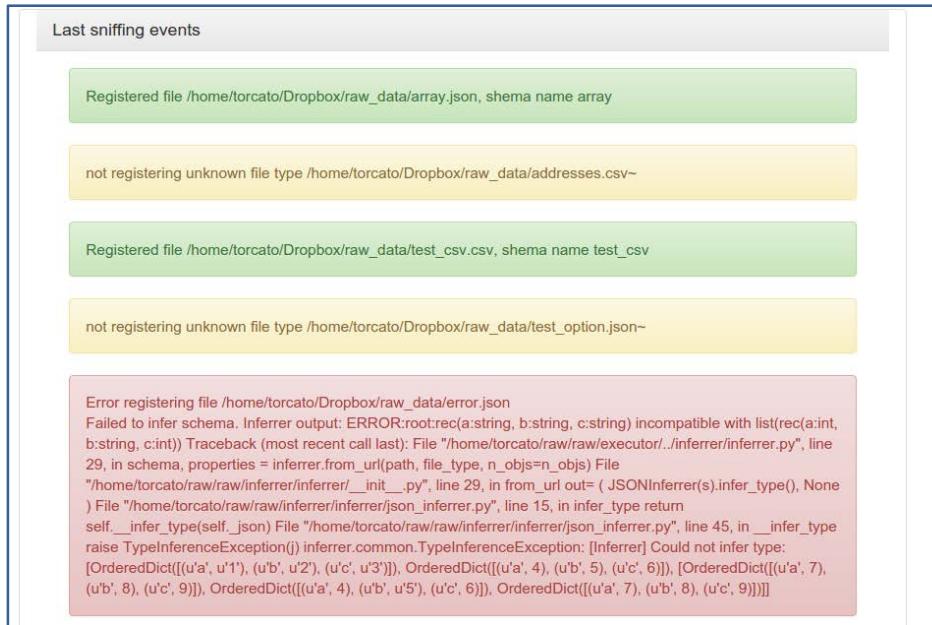
An administration UI is available to manage/configure Hospital Bundle. The UI is available only locally (within each hospital) and only to privileged users. The goal of this software is to give system administrators a more user-friendly way of handling the system.

Using the software, a system administrator can:

- Check configuration options



- Change user permissions
- Add/remove data
 - Applies only for local data in the hospital which has been already pre-processed (anonymized and mapped to MIP schema)
- Perform some basic data exploration
- View system status information (e.g., current load of servers and available resources)



```
Last sniffing events

Registered file /home/torcato/Dropbox/raw_data/array.json, shema name array
not registering unknown file type /home/torcato/Dropbox/raw_data/addresses.csv~
Registered file /home/torcato/Dropbox/raw_data/test_csv.csv, shema name test_csv
not registering unknown file type /home/torcato/Dropbox/raw_data/test_option.json~

Error registering file /home/torcato/Dropbox/raw_data/error.json
Failed to infer schema. Inferrer output: ERROR:root:rec(a:string, b:string, c:string) incompatible with list(rec(a:int, b:string, c:int)) Traceback (most recent call last): File "/home/torcato/raw/raw/executor/../inferrer/inferrer.py", line 29, in schema, properties = inferrer.from_url(path, file_type, n_objs=n_objs) File "/home/torcato/raw/raw/inferrer/inferrer/_init_.py", line 29, in from_url out= ( JSONInferrer(s).infer_type(), None ) File "/home/torcato/raw/raw/inferrer/inferrer/json_inferrer.py", line 15, in infer_type return self._infer_type(self._json) File "/home/torcato/raw/raw/inferrer/inferrer/json_inferrer.py", line 45, in __infer_type raise TypeInferenceException(j) inferrer.common.TypeInferenceException: [Inferrer] Could not infer type: [OrderedDict([(u'a', u'1'), (u'b', u'2'), (u'c', u'3')]), OrderedDict([(u'a', 4), (u'b', 5), (u'c', 6)]), [OrderedDict([(u'a', 7), (u'b', 8), (u'c', 9)])], OrderedDict([(u'a', 4), (u'b', u'5'), (u'c', 6)]), OrderedDict([(u'a', 7), (u'b', 8), (u'c', 9)])]]
```

Figure 19: Administration UI displaying log information of a server

To aid the user to check for errors in the data, this UI will have some data visualization capabilities, e.g. display histograms, bar charts and other graphs.

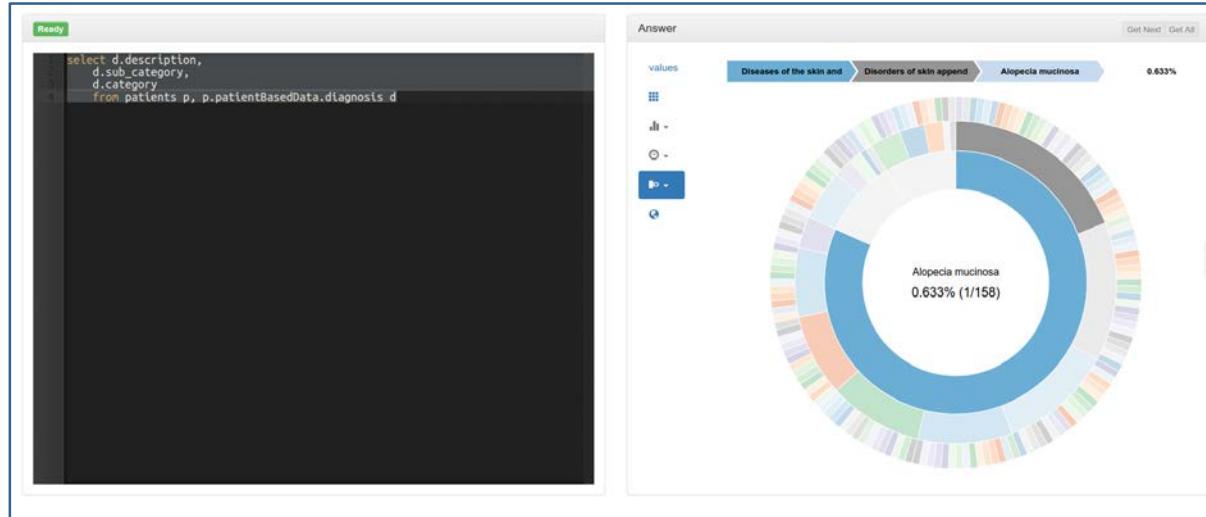


Figure 20: Administration UI showing summary of diagnosis

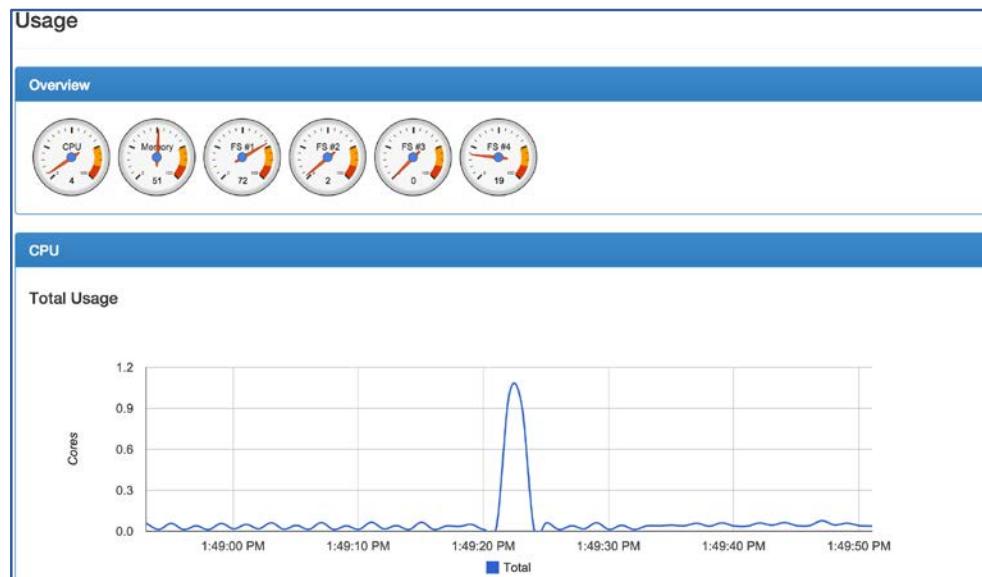


Figure 21: Administration UI showing the system load



B.5 Algorithms

Product/Software Package/Service name: MIP Function – Semi-supervised rule based clustering algorithm

Application version 1

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Ferath KHERIF
Documentation	https://github.com/HBPSP8Repo/woken
Source Code	https://github.com/HBPSP8Repo/woken
License	Apache 2.0 License
Current Version	0.1

Description

The rule-based algorithm aims to explain the variability between individuals, and describes a population by a group of “local over-densities”. These are defined as subspaces over combinations of variables. The algorithm performs an exhaustive search of the data space to predict the outcome variables;

Typical use case, the health status of each subject in terms of the presence or absence of AD. In our experiment, the predictive variables are the 90 brain region volumes, age, gender, and individual subject global volumes.



Product/Software Package/Service name: MIP Function –Informatics-based Model: Enriched Automated Diagnostic Tools

Application version 1

Metadata

Category MIP function, data analysis algorithm

Maintainers Ferath KHERIF and Jing CUI

Source Code
<https://github.com/HBPSP8Repo/woken>

License Apache 2.0 License

Current Version 0.1

Description

This is an automated classifier from a set of MRI scans from deceased, pathologically diagnosed individuals. This classifier provides prognostic value on clinically categorised living people.



Product/Software Package/Service name: MIP Function –Informatics-based Model: Deep Learning for Automated Features Extraction

Application version 1

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Ferath KHERIF and Bart VANDAME
Source Code	https://github.com/HBSP8Repo/woken
License	Apache 2.0 License
Current Version	0.1

Description

The increasing calculation power of computers has led to a rising interest in complex machine learning methods. In particular, the investigation of artificial neural networks with many hidden layers continuously results in promising new applications. These include image and face recognition (1, 2, 3), speech recognition (1, 2) and signal processing (1). Very recently, these deep learning networks have also been used in the classification of AD patients versus healthy control subjects, resulting in accuracies of up to 95% (Suk, Heung-II; Shen, Dinggang; Deep learning-based feature representation for AD/MCI classification Medical Image Computing and Computer-Assisted Intervention-MICCAI 2013 583-590,2013).

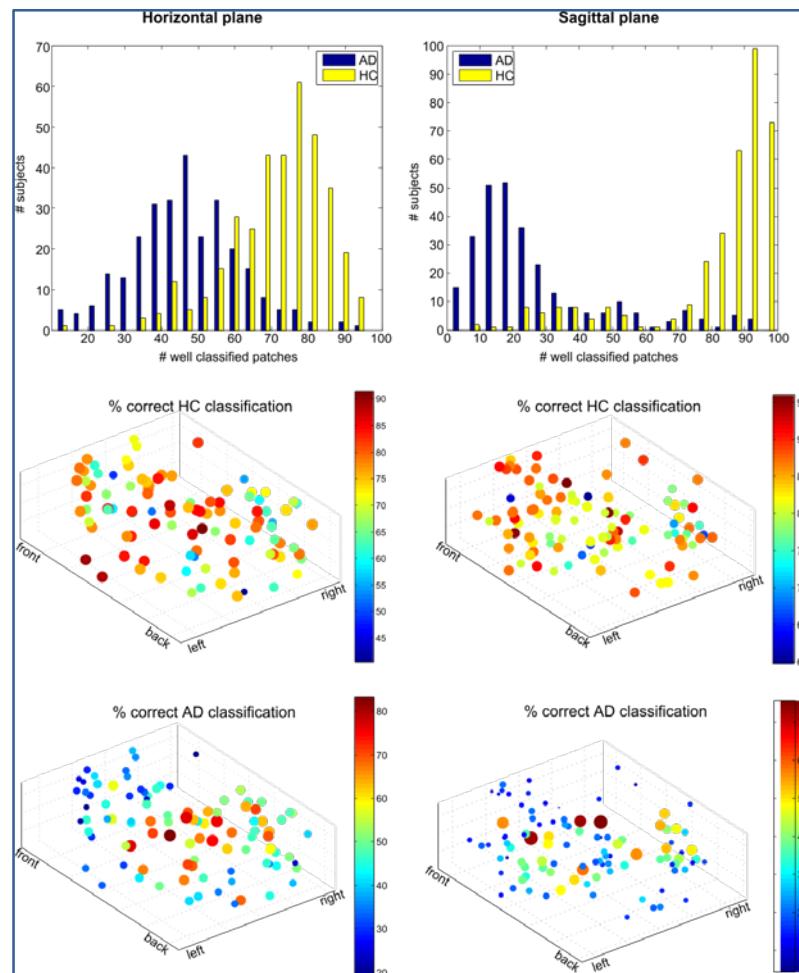


Figure 22: Results for Classification Based on Random Patches Used as Input Features.

Histograms of well classified patches (top) per group. Location of patches showing the percentage of correct classifications for HC (middle) and AD (bottom). The size and colour of the dots refer to the percentage of correct classifications.



Product/Software Package/Service name: MIP Function –Informatics-based Model: Rasch model and factor analysis for learning disease severity

Application version 1

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Ferath KHERIF and Bart VANDAME
Source Code	https://github.com/HBPS8Repo/woken
License	Apache 2.0 License
Current Version	0.1

Description

We propose to extract an index or a latent variable from the neuroimaging data to quantify the disease severity for each subject and regional vulnerability by applying factor analysis. Since the atrophy pattern correlates with the loss of neurons, this severity has a biological meaning and is independent of symptoms. We aim to test if the estimated severity significantly associates with clinical diagnosis and identify the regions highly weighted in calculating the severity.

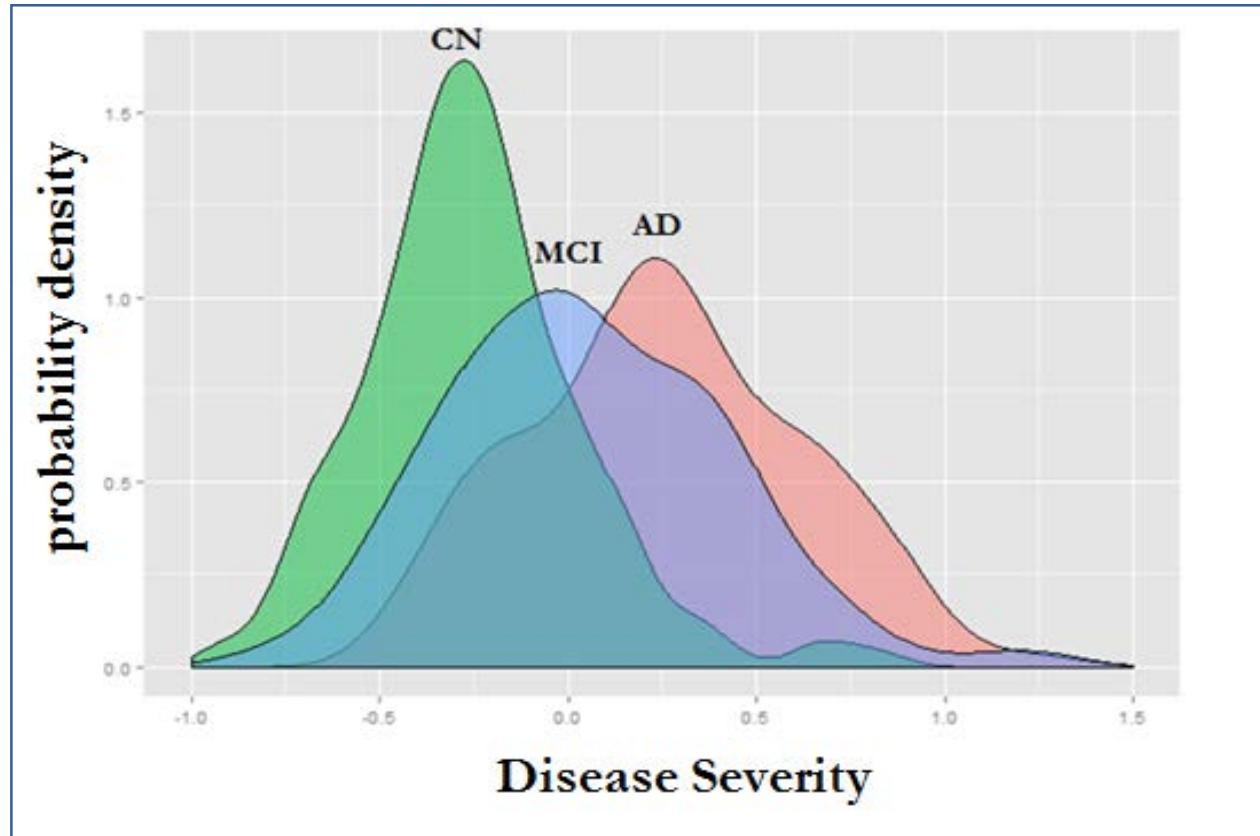


Figure 23: The probability density distributions of the 3 clinical groups.

In the post-hoc analysis, we applied two sample T-tests to compare each pair of groups: cognitive normals (CN) vs mild cognitive impairments (MCI), CN vs Alzheimer's disease (AD) and MCI vs AD. Three pairs of groups showed significant different distributions with p-value < 0.001.



Product/Software Package/Service name: MIP Function Informatics-based Model: Bi-clustering applied to gene expression and brain volumetric data

Application version 1

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Ferath KHERIF and Jonathan SULC
Source Code	https://github.com/HBPSp8Repo/woken
License	Apache 2.0 License
Current Version	0.1

Description

The Ping-Pong Algorithm (PPA) is a bi-clustering method used to compare two datasets with one common dimension. The algorithm uses a random weighted set of genes as a starting point (called a seed). It then selects subjects in which these genes deviate from the mean across subjects. Using these subjects, it then selects brain regions whose volumes deviate from the mean in these subjects. Using these regions, it selects a second vector of subjects for which the volumes of these regions deviate from the mean across subjects. Finally, it selects a new set of genes whose expression in these subjects deviates from the norm. This process is repeated until convergence is reached (i.e. the gene, subject, and region sets do not change from one iteration to the next).

Product/Software Package/Service name: MIP Function Informatics-based Model: Bayesian Causal Model

Application version 1



Metadata

Category	MIP function, data analysis algorithm
Maintainers	Ferath KHERIF and Lester MELI-GARCIA
Source Code	https://github.com/HBPSP8Repo/woken
License	Apache 2.0 License
Current Version	0.1

Description

We aimed at designing, deriving equations and implementing the causal Bayesian model similar to the General Linear Model (GLM) for distributed Data. GLM is one of the most used models to estimate dependencies between clinical, neuropsychological and neuroimaging variables. In our case, the data is distributed in different hospitals and it is not possible to move them to a unique Federation node where the GLM could be computed in a classical way. Therefore, special equations should be developed that allow us to have reliable GLM estimations under this condition. The Bayesian Formalism provides us the necessary armamentarium to deal with it and offers general sophisticated ways to extend to other models and managing high dimensional and Multimodal Data.



Divide and Recombine Bayesian paradigm (Parallel)

Example: Bayesian Linear Regression

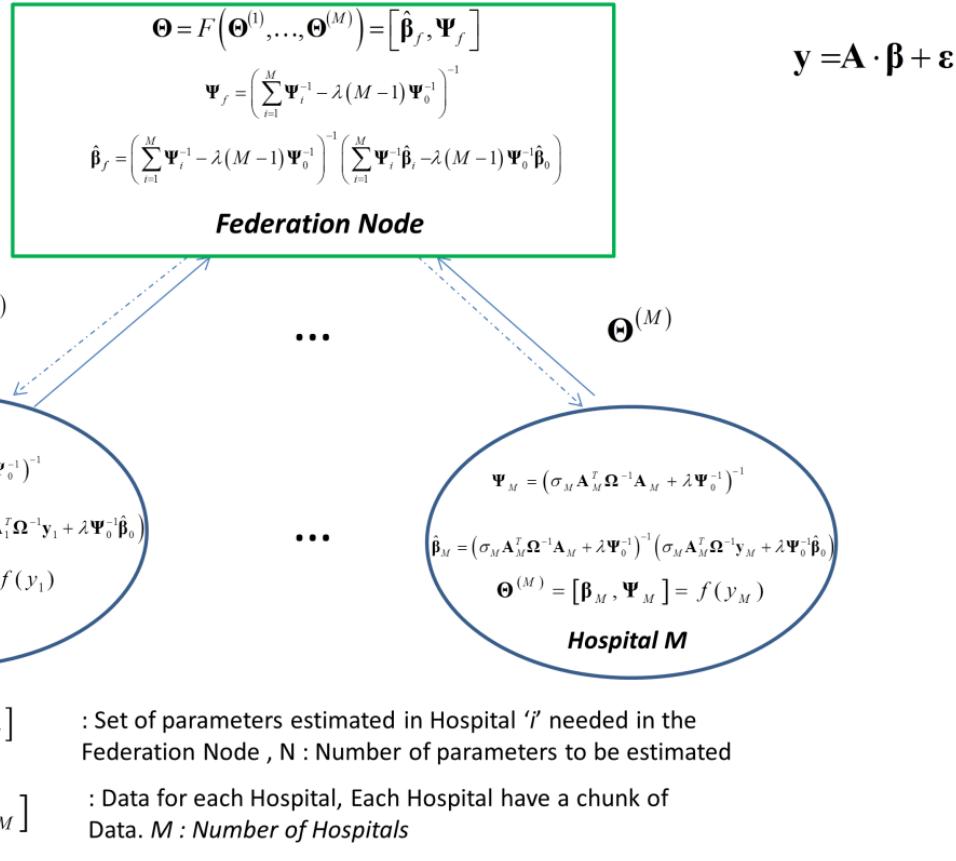


Figure 24: Bayesian Linear Regression

**Product/Software Package/Service name: MIP Function – Disease subtypes signatures - Big medical data strategy (3-C)**

Categorize, Cluster & Classify (3-C) - Disease subtype signatures - Big medical data strategy.

Application version 1.0

Metadata

Category MIP function, data analysis algorithm, disease signatures

Maintainers Alexis Mitelpunkt
Tal Galili

Documentation https://github.com/HBPPSP8Repo/CCC/blob/master/docs/Knit_Package.md

Current Version 0.2.0

All Versions 0.2.0

Description

Health informatics is facing many challenges these days, in analyzing current medical data and especially hospital data towards understanding disease mechanisms, predicting the course of a disease or assisting in targeting potential therapeutic options. Alongside the promises, many challenges emerge. Among the major ones we identified are: current diagnosis criteria that are too vague to capture disease manifestation; the irrelevance of personalized medicine when only heterogeneous classes of patients are available; and, how to properly process big data to avoid false claims. We offer a 3C strategy that starts from the medical knowledge, categorizing the available set of features into three types: the patients' assigned disease diagnosis, clinical measurements and potential biological markers, proceeds to an unsupervised learning process targeted to create new disease diagnosis classes, and finally, classifying the newly proposed diagnosis classes utilizing the potential biological markers. Our strategy, developed as part of the Medical Informatics Work Package at the EU Human Brain flagship Project strives to connect between potential biomarkers, and more homogeneous classes of disease manifestation that are expressed by meaningful features.

The development of the algorithm and its implementation was partly paid by the HBP.

**Product/Software Package/Service name: MIP Function –Label Propagation Framework**

Feature Extraction Framework. Data mining is to be based on a number of brain structure volume features, which are automatically extracted from patient MRI scans.

Application version 1.0

Metadata

Category	MIP function, data mining algorithm
Maintainers	John ASHBURNER
Homepage	http://www.fil.ion.ucl.ac.uk/~john/LabelProp/
Documentation	http://www.fil.ion.ucl.ac.uk/~john/LabelProp/Label%20Propagation%20Framework.pdf
Support	j.ashburner@ucl.ac.uk
Source Code	http://www.fil.ion.ucl.ac.uk/~john/LabelProp/
License	GNU General Public License
Current Version	1.0

Description

Single NIfTI volumes of the brain are first partitioned into three classes: grey matter, white matter and background. This procedure also incorporates an approximate image alignment step and a correction for image intensity non-uniformities. This procedure is done using the Segment⁵ tool from within the SPM12⁶ software, which runs within the MATLAB⁷ programming language.

⁵ Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005 Jul 1;26(3):839-51.

⁶ <http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>

⁷ <http://uk.mathworks.com/products/matlab/>

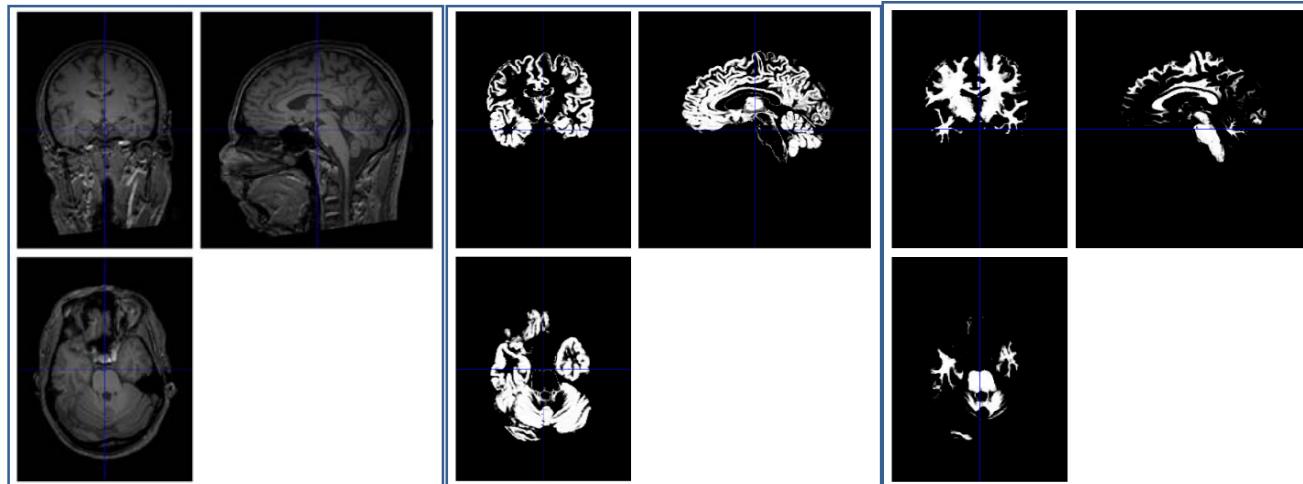


Figure 25: An original T1-weighted MRO scan (left), along with automatically extracted grey (middle) and white matter (right) tissue maps; the tissue maps encode the probability of each tissue type (given the model and data).

Tissue atlases, precomputed from the training data (see later), are then spatially registered with the extracted grey and white matter maps, using the *Shoot*⁸ tool from *SPM12*. The warps estimated from this registration step are then used to project other pre-computed image data in to alignment with the original scans (and their grey and white matter maps).

The rules of probability⁹ are then used to combine the various images to give a probabilistic label map for each brain structure. These probabilities are summed for each structure, to give probabilistic volume estimates. These estimates serve as features for data mining. Optionally, the method also allows maximum probability label maps to be saved.

⁸ Ashburner, John, and Karl J. Friston. "Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation." *NeuroImage* 55.3 (2011): 954-967.

⁹For example, for labelling the hippocampus, we have: $P(\text{structure}=\text{hippocampus}) = P(\text{structure}=\text{hippocampus} \mid \text{tissue}=\text{grey}) \times P(\text{tissue}=\text{grey}) + P(\text{structure}=\text{hippocampus} \mid \text{tissue}=\text{white}) \times P(\text{tissue}=\text{white}) + P(\text{structure}=\text{hippocampus} \mid \text{tissue}=\text{other}) \times P(\text{tissue}=\text{other})$.

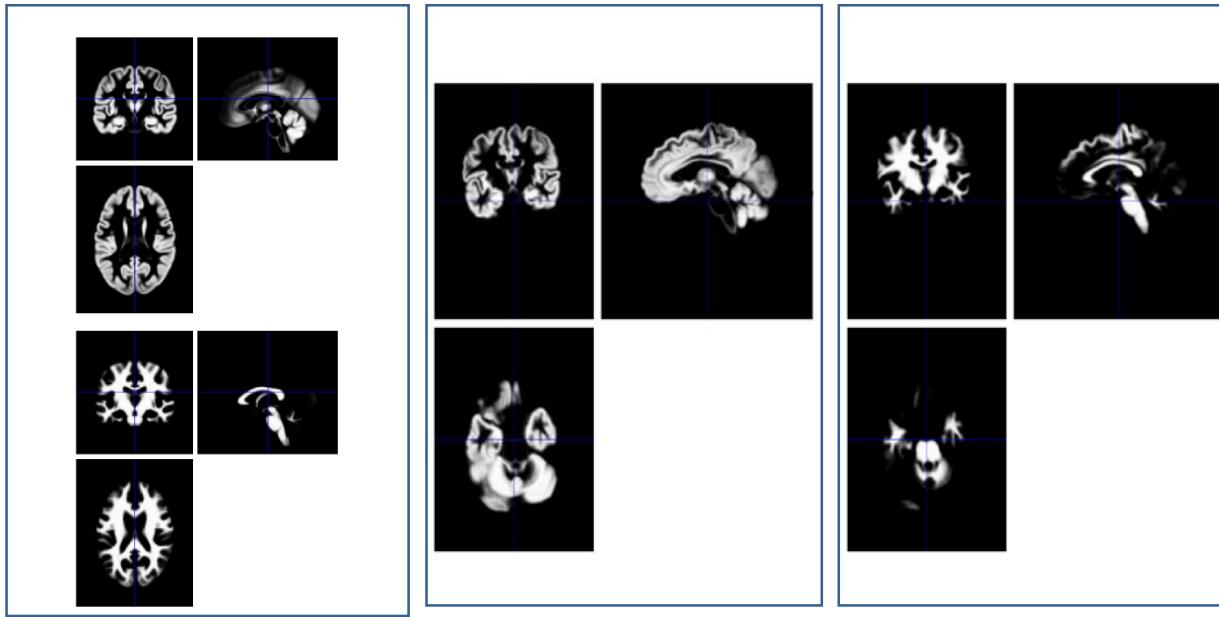


Figure 26: Grey and white matter from the original tissue atlases (left), together with registered versions (middle and right).

***Product/Software Package/Service name: MIP Function – Multi-Target Regression on Data Streams***

The data stream mining algorithm for predicting structured target variables.

Application version 15.10

Metadata

Category MIP function, data analysis algorithm

Maintainers [Aljaž OSOJNIK](#)

Homepage <http://moa.cms.waikato.ac.nz/>

Documentation <http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-moa-trees.pdf>

Support <http://source.ijs.si/hbp/mipfunctions/issues>

Source Code <http://source.ijs.si/hbp/moa.git>
<http://source.ijs.si/hbp/mipfunctions/tree/master/r-clus-trees>

License GNU GPL

Current Version 15.10

All Versions 15.10



Description

Methods for data stream mining are often used in Big Data problems, as they offer the ability to quickly process large amounts of data. However, the models obtained using this paradigm are often not interpretable for humans. We use trees learned in an online manner, which allows us to produce accurate and highly interpretable models, at high speeds. Through the use of the Hoeffding bound, the model can infer statistically supported hypotheses and use them to construct a decision tree.

This MIP function implements the FIMT-DD and iSOUP-Tree algorithms for learning decision trees from data streams, the former for single-target prediction and the latter for multi-target prediction. Both of these algorithms produce models in the form of a model tree, i.e. a decision tree, which uses linear functions in the leaves to achieve better performance.

The development of the algorithm and its implementation was partly paid by the HBP.

Product/Software Package/Service name: MIP Function – Predictive Clustering Trees

The predictive clustering tree algorithm for predicting structured target variables.

Application version 2.12

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Martin BRESKVAR Bernard ŽENKO
Homepage	http://source.ijs.si/hbp/clus/wikis/home
Documentation	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-clus-trees.pdf http://source.ijs.si/hbp/clus/wikis/documentation
Support	http://source.ijs.si/hbp/clus/issues



Category	MIP function, data analysis algorithm
Source Code	http://source.ijs.si/hbp/clus.git http://source.ijs.si/hbp/mipfunctions/tree/master/r-clus-trees
License	GNU GPL
Current Version	2.12
All Versions	2.12

Description

Predictive clustering combines aspects from both predictive modelling and clustering. Predictive clustering trees (PCTs) partition the set of examples into subsets in which examples have similar values of the target variable, while clustering produces subsets in which examples have similar values of the descriptive variables. The task of predictive clustering is to find clusters of examples that have similar values of both the target and the descriptive variables.

While most decision tree learners induce classification or regression trees, PCTs generalize this approach and represent trees that are interpreted as cluster hierarchies. Depending on the learning task at hand, different goal criteria are to be optimized while creating the clusters, and different heuristics will be suitable to achieve this. Classification and regression trees are special cases of PCTs, and by choosing the right parameter settings PCTs can closely mimic the behaviour of tree learners such as CART or C4.5. However, its applicability goes well beyond classical classification or regression tasks: PCTs have been successfully applied to many different tasks including multi-task learning (multi-target classification and regression), structured output learning, multi-label classification, hierarchical classification, and time series prediction. Next to these supervised learning tasks, PCTs are also applicable to semi-supervised learning, subgroup discovery, and clustering.

This MIP function implements the PCT algorithm.

The development of the algorithm and its implementation was not paid by the HBP.

***Product/Software Package/Service name: MIP Function – Rule Ensembles***

The rule ensemble algorithm for predicting structured target variables.

Application version 2.12

Metadata

Category MIP function, data analysis algorithm

Maintainers [Martin BRESKVAR](#)

[Bernard ŽENKO](#)

Homepage <http://source.ijs.si/hbp/clus/wikis/home>

Documentation <http://source.ijs.si/hbp/clus/wikis/documentation>

Support <http://source.ijs.si/hbp/clus/issues>

Source Code <http://source.ijs.si/hbp/clus.git>

License GNU GPL

Current Version 2.12

All Versions 2.12



Description

Methods for learning decision rules are being successfully applied to many problem domains, particularly when understanding and interpretation of the learned model is necessary. In many real life problems, we would like to predict structured target variables, e.g. multiple related numeric variables, or time series. While several methods for learning rules that predict multiple targets at once exist, they are all based on the covering algorithm, which does not work well for regression problems. A better solution for regression is the rule ensemble approach that transcribes an ensemble of decision trees into a large collection of rules. An optimization procedure is then used to select the best (and much smaller) subset of these rules and to determine their respective weights.

This MIP function implements the FIRE algorithm, which employs the rule ensemble approach for solving multi-target regression and time series problems. We can improve the accuracy of the rule model by adding simple linear functions to the ensemble, which results in a model that is a combination of global linear functions and rules.

The development of the algorithm and its implementation was partly paid by the HBP.

***Product/Software Package/Service name: MIP Function – Feature Ranking for Structured Targets***

The feature ranking algorithm for structured target variables.

Application version 2.12

Metadata

Category MIP function, data analysis algorithm

Maintainers [Dragi KOCEV](#)

Homepage <http://source.ijs.si/hbp/clus/wikis/home>

Documentation <http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-clus-franking.pdf>
<http://source.ijs.si/hbp/clus/wikis/documentation>

Support <http://source.ijs.si/hbp/clus/issues>

Source Code <http://source.ijs.si/hbp/clus.git>
<http://source.ijs.si/hbp/mipfunctions/tree/master/r-clus-franking>

License GNU GPL

Current Version 2.12

All Versions 2.12



Description

Methods for feature ranking are used in many domains with many descriptive variables, i.e. high-dimensional problems. The obtained rankings provide an additional insight about the importance of the variables for the target and/or reduce the dimensionality of the problem. Many real-life problems have structured targets that need to be predicted. However, the task of feature ranking in the context of predicting structured target variables is more complex than the same task for simple classification or regression. Typical approaches for this task decompose the output to primitive components, perform feature ranking on these smaller problems, and then aggregate the resulting rankings into a single ranking. They are computationally intractable for large output spaces (e.g., genetic or imaging data) and ignore the dependencies between components of the output. We have developed efficient feature ranking methods in the context of predicting structured targets. The developed methods are based on the ensemble learning paradigm.

This MIP function implements two algorithms for feature ranking for structured targets: (1) RF-RANK exploits the random forests mechanism, and (2) GENIE3 exploits the variance reduction at each tree node from the ensemble. For the latter method, the ensemble could be random forest or an ensemble of extra trees. Both methods use predictive clustering trees as base predictive models.

The development of the algorithm and its implementation was partly paid by the HBP.

***Product/Software Package/Service name: MIP Function – Subgroup Discovery from Multi-Resolution Data***

A subgroup discovery tool that can use ontological domain knowledge (RDF graphs) in the learning process. Subgroup descriptions contain terms from the given domain knowledge and enable potentially better generalizations.

Application version 0.3.1 (hbp branch)

Metadata

Category MIP function, data analysis algorithm

Maintainers [Anže VAVPETIČ](#)

Homepage <https://github.com/anzev/hedwig/tree/hbp>

Documentation <http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-hedwig.pdf>
<https://github.com/anzev/hedwig/tree/hbp>

Support <https://github.com/anzev/hedwig/issues>

Source Code <https://github.com/anzev/hedwig/tree/hbp>

License MIT

Current Version 0.3.1

All Versions 0.3.1



Description

This MIP function implements the Hedwig algorithm [1, 2, 3]. Given a data set consisting of examples (e.g. patients) described in terms of several descriptive binary attributes (e.g. clinical variables) labelled with a single target attribute (e.g. a potential biological marker) the algorithm produces a set of rules, i.e. subgroup descriptions or subgroups. In contrast to standard subgroup discovery, Hedwig can also exploit additional domain knowledge encoded as RDF graphs: these can be simple facts or ontologies like the Gene Ontology or a combination of several types of domain knowledge. These RDF graphs represent additional relationships of the attributes describing the examples and can be used to automatically generate generalizations not possible only with the “bottom level” descriptive attributes. Currently it is unclear if such hierarchical information will be available within the MIP, so the Hedwig algorithm (as implemented in the MIP function) at the moment generates subgroup descriptions only from bottom level attributes.

The development of the algorithm and its implementation was partly paid by the HBP.

***Product/Software Package/Service name: MIP Function – Subgroup Discovery from Heterogeneous Data***

The subgroup discovery algorithm for analysis of heterogeneous data.

Application version 1.0

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Jan KRALJ
Homepage	http://source.ijs.si/hbp/tehin/wikis/home
Documentation	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-tehin.pdf
Support	http://source.ijs.si/hbp/tehin/issues
Source Code	http://source.ijs.si/hbp/tehin.git http://source.ijs.si/hbp/mipfunctions/tree/master/r-tehin
License	GNU GPL
Current Version	1.0
All Versions	1.0



Description

Network analysis is an ever-growing field of research capable of reasoning with data in a network setting. An important part of network analysis is so called network propositionalisation, which allows us to construct feature vectors for each node in a network. An example of network propositionalisation is a specific application of the personalized PageRank algorithm.

In a heterogeneous network, network propositionalisation becomes less obvious. In a network with several different types of nodes, it does not make sense to construct feature vectors for all nodes because the nodes may be entirely non-comparable. In our approach, we therefore take a heterogeneous network and deconstruct it into several homogeneous networks, each containing nodes of the same (so called target) type. Network propositionalisation can then be applied to the homogeneous networks and the resulting vectors can be concatenated to construct a single feature vector for each node of the target type.

This function performs network propositionalisation on a heterogeneous network by first deconstructing the network into several homogeneous networks. The homogeneous networks are constructed using user-supplied meta-paths in the heterogeneous network (for example, in a network consisting of papers and their authors, we can construct a homogeneous network of papers where two papers are connected if they share an author). The result of this function is a set of feature vectors, one for each node of the target type. Recently, the function was updated so that it can accept not only heterogeneous networks, but also standard data instances (with feature vectors) as input. In that case, the function constructs a proximity network of instances and performs network propositionalisation on the resulting network.

The development of the algorithm and its implementation was partly paid by the HBP.

**Product/Software Package/Service name: MIP Function – Visual Performance Evaluation**

The ViperCharts web-based platform for visual performance evaluation of data analysis algorithms.

Application version 1.0

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Jan KRALJ
Homepage	http://viper.ijs.si/
Documentation	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-viper.pdf http://viper.ijs.si/about/
Support	http://source.ijs.si/hbp/mipfunctions/issues
Source Code	http://source.ijs.si/hbp/mipfunctions/tree/master/r-viper
License	GNU GPL
Current Version	1.0
All Versions	1.0

Description

ViperCharts is a web-based platform for visual performance evaluation of classification, prediction, and information retrieval algorithms.



The platform enables users to create interactive charts for easy and intuitive evaluation of performance results. It includes standard performance visualizations used in machine learning, data mining, information retrieval, etc., and extends them by offering alternative evaluation methods like F-isolines, and by establishing relations between corresponding presentations like ROC, Precision-Recall and Lift curves, or ROC Hull and Cost curves.

Additionally, the interactive performance charts can be saved, exported to several formats, and shared via unique web addresses. A web API to the service is also available.

ViperCharts support the following charts for visual performance evaluation:

- Scatter charts: PR space charts and ROC space charts.
- Curve charts: Lift curves, ROC curves, PR curves, Cost curves, Kendall curves and Rate-driven curves.
- Column charts: General column charts for visualizing multiple performance measures for a set of algorithms.

The implementation of the algorithm was partly paid by the HBP.

**Product/Software Package/Service name: MIP Function – Brainspan co-expression clustering****Metadata**

Category MIP function, Brainspan co-expression clustering

Maintainers LUMC

Documentation <http://lvdmaaten.github.io/tsne/>

Description

We developed a methodology based on Matlab technology to extract exploit disease patterns using co-expression network analysis in Human Brain Transcriptome. In <http://www.sciencedirect.com/science/article/pii/S1046202314003211>, we present a multi-dimensional co-expression analysis method for extracting disease signatures. We used BrainSpan human transcriptome database for our experiments. The results reveal that the Autism Spectral Disorder candidate genes share transcriptional networks related to synapse formation and elimination, protein turnover and mitochondrial function. We believe similar analysis can also be used in the SP8 MIP to find different disease signatures through lists of disease-implicated genes. Additionally, as an ongoing research, we aim to classify Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) patients accurately using both imaging and genetic features in the ADNI data and BH-tSNE.

An example validation / application use case of how this gene co-expression clustering in normal brains can be used to uncover human brain disease signatures from GWAS data of ~ 20000 Migraine patients and ~100000 controls is worked out in the paper "Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas" (<https://www.broadinstitute.org/publications/broad7938> - DOI 10.1007/s00439-016-1638-x). This type of analysis reveals which regions, cell types and biological processes may be involved in particular brain diseases, given a large GWAS cohort once it becomes available from the hospitals participating in the MIP.

**Product/Software Package/Service name: MIP Function – BH-tSNE****Metadata**

Category MIP function, Co-expression Network Analysis

Maintainers LUMC

Homepage <https://surfdrive.surf.nl/files/public.php?service=files&t=1e4242bba90e33ab0cf1f4154ab5ce44>
<https://github.com/lvdmaaten/bhtsne/>

Documentation http://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf
http://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

Description

tSNE has been developed by Laurens van der Maaten as a non-linear dimensionality reduction (DR) algorithm to visualize high-dimensional data. The main advantage of non-linear DR algorithms in comparison to linear DR algorithms such as PCA is that non-linear DR algorithms can represent neighbouring samples in the high-dimensional space better than linear DR algorithms in the lower dimensional space such as 2D or 3D. This advantage is crucial to visualize similarities between the features of samples of the data (such as gene expressions, disease phenotypes), hence to observe possible correlations between the samples and classify them in the same cluster.

tSNE has been successfully applied to analyse a broad variety of different data from hand-written digits to gene expression. As a main drawback, tSNE is computationally expensive and it is not feasible to apply it to large data. tSNE with Barnes-Hut approximation (BH-tSNE) has recently been developed to overcome the sample size limitation and computational overhead of tSNE. With BH-tSNE, it is possible to create an embedding of high-dimensional data with millions of samples. Furthermore, BH-tSNE requires $O(N\log N)$ computations compared to tSNE's computational complexity of $O(N^2)$. Hence, Barnes Hut approximation lowers computational time of tSNE substantially.

BH-tSNE can be successfully used to find disease patterns on disease phenotypes and gene expression data. As an example, we have recently used it to classify Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) patients.

The software tool for BH-tSNE is available in C++. Matlab and Python wrappers are also available. Recently, we also implemented it in R language and share it on GitHub repository.



Technologies

<https://github.com/lvdmaaten/bhtsne/>

<https://github.com/jkrijthe/Rtsne>



Annex C: Summary - Platform Use Case Status

The table below gives the release status in the MIP for each Use Case described in the RUP Deliverable D8.6.1 (specification of the MIP).

Table 5: Platform use case status

Use Case ID	Success Scenario	Status	ID of Related Functional Requirement	Related Product/Software Package/Service	TRL at end of RUP	Contributors	
SP8-UC-001	<p><u>Primary actor:</u> Alan is a GU. He is an epidemiologist.</p> <p><u>Description:</u> Alan wants to write a report for the World Health Organisation on Alzheimer's disease. He is interested in estimating the incidence rate (i.e. the number of cases) of the disease in a sampled population in Europe for demographic factors.</p> <p><u>Preconditions:</u></p> <ul style="list-style-type: none"> • The clinical disease classification Variables (e.g. ICD-10) have been released and are available at local data sources. • Metadata describing the local data-source Variables are available at the MIP Web Portal. <p><u>Success scenario:</u></p> <ol style="list-style-type: none"> 1) Alan logs into the Collaboratory. 2) He selects the Epidemiological Exploration service and uses the interface. 3) He browses the Variable space of the ICD-10 classification and selects "Alzheimer's disease" 4) Alan narrows his research to the specific geographical region and to demographic factors. 5) The interface displays a graph/table output with the number of cases of Alzheimer's disease patients in Europe for the demographic factors. 	<p>Released</p> <p>This use case is supported by the EE service on the website.</p>	SP8-FR-001-G SP8-FR-003-G SP8-FR-005-G SP8-FR-007-G SP8-FR-009-G SP8-FR-015-WeP SP8-FR-026-WeP SP8-FR-028-WeP SP8-FR-029-DF SP8-FR-031-DF SP8-FR-033-DF SP8-FR-035-DF SP8-FR-037-DF	SP8-FR-002-G SP8-FR-004-G SP8-FR-006-G SP8-FR-008-G SP8-FR-010-G SP8-FR-025-WeP SP8-FR-027-WeP SP8-FR-030-DF SP8-FR-032-DF SP8-FR-034-DF SP8-FR-036-DF SP8-FR-038-DF	*Elastic Search *Interactive variable browser *Interactive graph/charts display for stats	5-6	WP8.4



			SP8-FR-039-DF SP8-FR-047-LD SP8-FR-049-LD SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD	SP8-FR-040-DF SP8-FR-048-LD SP8-FR-050-LD SP8-FR-052-LD SP8-FR-054-LD				
SP8-UC-002	<p>Primary actor: Beth is a GU. She is a clinician in neurology.</p> <p>Description:</p> <p>Beth is interested in carrying out a study on dementia. First, she wants to know the number of cases for different types of dementia in Europe: dementia in Alzheimer's disease with early and late onset, vascular dementia and dementia in Pick's disease.</p> <p>Preconditions:</p> <ul style="list-style-type: none"> The clinical disease classification Variables (e.g. ICD-10 World Health Organisation International Statistical Classification of Diseases and Related Health Problems 10th Revision; code adopted by hospitals) have been released and are available at local data sources. Metadata describing the local data source Variables are available at the MIP Web Portal. <p>Success scenario:</p> <ol style="list-style-type: none"> 1) Beth logs into the Collaboratory. 2) She selects the Epidemiological Exploration service and uses the interface. 3) She uses the multiple-choice drop-down menu to select the different types of dementia from the ICD-10 classification. 4) Beth narrows her research to the geographical region and age ranges. 5) The interface displays a graph/table output with the number of cases for each type of dementia and age range. 	Released	<p>Epidemiological Exploration</p> <p>This use case is supported by the EE service on the website.</p>	SP8-FR-001-G SP8-FR-003-G SP8-FR-005-G SP8-FR-007-G SP8-FR-009-G	SP8-FR-002-G SP8-FR-004-G SP8-FR-006-G SP8-FR-008-G SP8-FR-010-G	*Elastic Search *Interactive variable browser *Interactive graph/charts display for stats	5-6	WP8.4



			SP8-FR-039-DF SP8-FR-047-LD SP8-FR-049-LD SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD	SP8-FR-040-DF SP8-FR-048-LD SP8-FR-050-LD SP8-FR-052-LD SP8-FR-054-LD			
SP8-UC-003	<p>Primary actor: Charlotte is a GU. She is a neuroscientist.</p> <p>Description:</p> <p>Charlotte wants to study the relation between the mean volume of grey matter of the hippocampi and the types of dementia.</p> <p>Preconditions:</p> <ul style="list-style-type: none"> The brain "grey matter volume." Variables have been produced at the local level, released and are available through the MIP Web Portal. Metadata describing the local data source Variables are available at the MIP Web Portal. <p>Success scenario:</p> <ol style="list-style-type: none"> 1) Charlotte logs into the Collaboratory. 2) She accesses the Epidemiological Exploration service (as per Use Cases 1&2) where she selects the Variable "Volume of grey matter of hippocampi" in the cases of dementia. 3) The system retrieves the summary counts, i.e. the number of records for the selected Variables, and displays it. 4) Charlotte views the results, and she decides that the sample size is sufficient to examine further the values for a particular type of population. 5) She selects the Interactive Analysis service and chooses the statistical analysis for comparing differences in grey matter volume for the sub-types of dementia. 6) She retrieves the mean volume of grey matter for populations of interest and their standard deviations. 	Released	<p>Interactive Analysis</p> <p>The users can now run similar use cases via the IA service on the Platform.</p>	<p>SP8-FR-001-G SP8-FR-003-G SP8-FR-005-G SP8-FR-007-G SP8-FR-009-G</p> <p>SP8-FR-016-WeP SP8-FR-018-WeP SP8-FR-020-WeP SP8-FR-025-WeP SP8-FR-027-WeP</p> <p>SP8-FR-029-DF SP8-FR-031-DF SP8-FR-033-DF SP8-FR-035-DF</p>	<p>SP8-FR-002-G SP8-FR-004-G SP8-FR-006-G SP8-FR-008-G SP8-FR-010-G</p> <p>SP8-FR-017-WeP SP8-FR-019-WeP SP8-FR-021-WeP SP8-FR-026-WeP SP8-FR-028-WeP</p> <p>SP8-FR-030-DF SP8-FR-032-DF SP8-FR-034-DF SP8-FR-036-DF</p>	<p>*Interactive graph/charts display for stats</p> <p>*Data filtering</p> <p>*Configuration panel for methods</p> <p>*Link to online help & tutorial for method configuration (MIP Knowledge Base)</p>	WP8.4 WP8.5 WP8.2 WP8.3 WP11.2



	7) Charlotte can then use these measures and compare them to the values that she observed in her clinic, and evaluate whether they are typical or unusually high/low.		SP8-FR-037-DF SP8-FR-039-DF SP8-FR-041-LD SP8-FR-043-LD SP8-FR-045-LD SP8-FR-047-LD SP8-FR-049-LD SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD SP8-FR-057-LD SP8-FR-059-LD	SP8-FR-038-DF SP8-FR-040-DF SP8-FR-042-LD SP8-FR-044-LD SP8-FR-046-LD SP8-FR-048-LD SP8-FR-050-LD SP8-FR-052-LD SP8-FR-054-LD SP8-FR-056-LD SP8-FR-058-LD SP8-FR-060-LD		
SP8-UC-004	<p><u>Primary actors:</u></p> <ul style="list-style-type: none"> • Charlotte is a GU. She is a neuroscientist expert in neuroimaging. • Mike is a GU. He is a neuroscientist expert in genomics. <p><u>Description:</u></p> <p>Charlotte and Mike want to collaborate on a project for understanding the link between genetics and brain features in dementia.</p> <p><u>Preconditions:</u></p> <ul style="list-style-type: none"> • The brain "grey matter volume" Variables have been released and are available at the local data source. • The "APOE gene" Variables have been released and are available from the local data source. • Metadata describing the local data source Variables are available at the MIP Web Portal. 	Released	<p>Interactive Analysis</p> <p>The users can now run similar use cases via the IA service on the Platform.</p>	<p>SP8-FR-001-G SP8-FR-003-G SP8-FR-005-G SP8-FR-007-G SP8-FR-009-G SP8-FR-016-WeP SP8-FR-018-WeP SP8-FR-020-WeP</p> <p>SP8-FR-002-G SP8-FR-004-G SP8-FR-006-G SP8-FR-008-G SP8-FR-010-G SP8-FR-017-WeP SP8-FR-019-WeP SP8-FR-021-WeP</p>	<p>*Interactive graph/charts display for stats</p> <p>*Data filtering</p> <p>*Configuration panel for methods</p> <p>*Link to online help & tutorial for method configuration (MIP)</p>	WP8.4 WP8.5



	<u>Success scenario:</u> 1) Charlotte and Mike log into the Collaboratory. 2) Charlotte accesses the Epidemiological Exploration service (as per Use Cases 1&2) where she selects the Variable "Volume of grey matter of hippocampi" in the cases of dementia. With Mike's expertise in genetics, she is now able to also select the "ApoE genetic" phenotype. 3) The system retrieves the summary counts, i.e. the number of records for the selected Variables, and displays it. 4) Charlotte and Mike view the results and they decide to examine further the values for a particular type of population. 5) They select the Interactive Analysis service and choose the statistical analysis for comparing differences in grey matter volume for the sub-types of dementia and the "ApoE genetic" phenotype. They can then assess the replicability of the results on another subset. 6) Charlotte and Mike save their results in a common project folder in the UP.		SP8-FR-025-WeP SP8-FR-027-WeP SP8-FR-029-DF SP8-FR-031-DF SP8-FR-033-DF SP8-FR-035-DF SP8-FR-037-DF SP8-FR-039-DF SP8-FR-041-LD SP8-FR-043-LD SP8-FR-045-LD SP8-FR-047-LD SP8-FR-049-LD SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD SP8-FR-057-LD SP8-FR-059-LD	SP8-FR-026-WeP SP8-FR-028-WeP SP8-FR-030-DF SP8-FR-032-DF SP8-FR-034-DF SP8-FR-036-DF SP8-FR-038-DF SP8-FR-040-DF SP8-FR-042-LD SP8-FR-044-LD SP8-FR-046-LD SP8-FR-048-LD SP8-FR-050-LD SP8-FR-052-LD SP8-FR-054-LD SP8-FR-056-LD SP8-FR-058-LD SP8-FR-060-LD	Knowledge Base)	
SP8-UC-005	Primary actor: John is a DU (Developer User). He is a scientific developer. Description:	Released A developer user guide has been produced, which shows	Interactive Analysis	SP8-FR-001-G SP8-FR-003-G	SP8-FR-002-G SP8-FR-004-G	WP8.4 WP8.5 WP8.2 WP8.3



	<p>John has developed new tools for combining genetic information and brain volume, and wants to share them with the MIP community to be used in data analyses.</p> <p>Success scenario:</p> <ol style="list-style-type: none"> 1) John logs into the Collaboratory. 2) John uploads his script and tests the script in the sandbox. 3) The script is retested and validated by the Admin User's group. The script is then made available to the GU community. 4) The script is made available to all users in the Interactive Analysis service. 	<p>how users can add apps to the Platform.</p>	SP8-FR-005-G SP8-FR-007-G SP8-FR-009-G SP8-FR-025-WeP SP8-FR-027-WeP SP8-FR-029-DF SP8-FR-031-DF SP8-FR-033-DF SP8-FR-035-DF SP8-FR-037-DF SP8-FR-039-DF SP8-FR-046-LD SP8-FR-048-LD SP8-FR-050-LD SP8-FR-052-LD SP8-FR-054-LD	SP8-FR-006-G SP8-FR-008-G SP8-FR-010-G SP8-FR-026-WeP SP8-FR-028-WeP SP8-FR-030-DF SP8-FR-032-DF SP8-FR-034-DF SP8-FR-036-DF SP8-FR-038-DF SP8-FR-040-DF SP8-FR-047-LD SP8-FR-049-LD SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD		WP11.2
SP8-UC-006	<p>Primary actor: Beth is a GU. She is a clinician in neurology.</p> <p>Description: Beth is interested in taking forward personalised diagnostics using the biological signatures of the disease.</p> <p>Preconditions:</p>	<p>Released Exploration of the biological signature of diseases is implemented as an</p>	<p>Biological Signatures of Diseases</p>		4	WP8.4 WP8.5 WP8.2 WP8.3 WP11.2



<ul style="list-style-type: none"> The biological signatures of diseases produced by the data mining algorithms are available at the MIP Web Portal. The Variables that describe each disease signature cluster have been released and are available at the MIP Web Portal. <p>Success scenario:</p> <ol style="list-style-type: none"> 1) Beth logs into the Collaboratory. 2) She selects the Biological Signatures of Diseases service and uses the interface to classify her own patient by comparing his clinical and biological characteristics with the whole range of provided biological signatures of diseases using an optimal matching algorithm. 3) She does this by selecting Variables of interest - e.g. demographic data, blood cholesterol, neuropsychological scores, genetic burden, etc. 4) She enters values for those Variables. 5) She retrieves a list of disease signatures ordered according to the best match. The distribution of values of the other unselected Variables is also displayed along with their uncertainty - e.g. genotype, clinical scores and cardiovascular risk factors. 6) She also retrieves a 3D brain map with highlighted anatomical regions affected by the particular disease corresponding to the optimally matched disease signature. She can compare the map with the anatomy pattern of her own patients. 7) Depending on how well the disease signature cluster matches her criteria, Beth can add new Variables to determine the stability of her classification in relation to the number of criteria or Variables used. 8) She can compare the derived disease signature cluster to conventional clinical classification - e.g. ICD-10, DSM V classification. 9) If needed, she can review her patients (data) to verify the derived disease signature cluster by similarity and by differences with other patients 	interactive viewer.	SP8-FR-009-G SP8-FR-022-WeP SP8-FR-024-WeP SP8-FR-026-WeP SP8-FR-028-WeP SP8-FR-029-DF SP8-FR-031-DF SP8-FR-033-DF SP8-FR-035-DF SP8-FR-037-DF SP8-FR-039-DF SP8-FR-041-LD SP8-FR-043-LD SP8-FR-045-LD SP8-FR-047-LD SP8-FR-049-LD SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD SP8-FR-057-LD SP8-FR-059-LD	SP8-FR-010-G SP8-FR-023-WeP SP8-FR-025-WeP SP8-FR-027-WeP SP8-FR-030-DF SP8-FR-032-DF SP8-FR-034-DF SP8-FR-036-DF SP8-FR-038-DF SP8-FR-040-DF SP8-FR-042-LD SP8-FR-044-LD SP8-FR-046-LD SP8-FR-048-LD SP8-FR-050-LD SP8-FR-052-LD SP8-FR-054-LD SP8-FR-056-LD SP8-FR-058-LD SP8-FR-060-LD				
---	---------------------	--	--	--	--	--	--



SP8-UC-007	<p>Primary actor: Nathalie is a GU. She is a researcher in pharmaceutical R&D.</p> <p>Preconditions:</p> <ul style="list-style-type: none"> The biological signatures of diseases produced by the data mining algorithms are available at the MIP Web Portal. The Variables that describe each disease signature cluster have been released and are available at the MIP Web Portal. <p>Description: Nathalie is interested in defining inclusion criteria and a set of non-invasive biomarkers for a clinical trial on a new drug for Alzheimer's disease.</p> <p>Success scenario:</p> <ol style="list-style-type: none"> 1) Nathalie logs into the Collaboratory. 2) She selects the Biological Signatures of Diseases service and uses the interface to retrieve the set of features of interest according to the provided disease signatures for dementia of the Alzheimer type. 3) She identifies the features leading to the creation of homogeneously stratified set of rules she would apply to create the cohorts undergoing pharmacological intervention. 4) She uses the provided predictive tools to infer potential therapeutic targets and positive as well as adverse effects based on multi-scale information - e.g. molecular pathways, proteomics interactions, genetic profiles, etc. up to the system/behavioural level. 5) She is now in a position to specify trials using well defined homogeneous, and therefore small cohorts to test the effects of a drug or cocktail of drugs that modulates the targets suggested by the rules that define her disease signature of interest. 	<p>Released</p> <p>Exploration of the biological signature of diseases is implemented as an interactive viewer.</p>	<p>Biological Signatures of Diseases</p> <table border="0"> <tr> <td>SP8-FR-001-G</td> <td>SP8-FR-002-G</td> </tr> <tr> <td>SP8-FR-003-G</td> <td>SP8-FR-004-G</td> </tr> <tr> <td>SP8-FR-005-G</td> <td>SP8-FR-006-G</td> </tr> <tr> <td>SP8-FR-007-G</td> <td>SP8-FR-008-G</td> </tr> <tr> <td>SP8-FR-009-G</td> <td>SP8-FR-010-G</td> </tr> </table> <table border="0"> <tr> <td>SP8-FR-022-WeP</td> <td>SP8-FR-023-WeP</td> </tr> <tr> <td>SP8-FR-024-WeP</td> <td>SP8-FR-025-WeP</td> </tr> <tr> <td>SP8-FR-026-WeP</td> <td>SP8-FR-027-WeP</td> </tr> <tr> <td>SP8-FR-028-WeP</td> <td></td> </tr> </table> <table border="0"> <tr> <td>SP8-FR-029-DF</td> <td>SP8-FR-030-DF</td> </tr> <tr> <td>SP8-FR-031-DF</td> <td>SP8-FR-032-DF</td> </tr> <tr> <td>SP8-FR-033-DF</td> <td>SP8-FR-034-DF</td> </tr> <tr> <td>SP8-FR-035-DF</td> <td>SP8-FR-036-DF</td> </tr> <tr> <td>SP8-FR-037-DF</td> <td>SP8-FR-038-DF</td> </tr> <tr> <td>SP8-FR-039-DF</td> <td>SP8-FR-040-DF</td> </tr> <tr> <td>SP8-FR-041-LD</td> <td>SP8-FR-042-LD</td> </tr> <tr> <td>SP8-FR-043-LD</td> <td>SP8-FR-044-LD</td> </tr> <tr> <td>SP8-FR-045-LD</td> <td>SP8-FR-046-LD</td> </tr> <tr> <td>SP8-FR-047-LD</td> <td>SP8-FR-048-LD</td> </tr> <tr> <td>SP8-FR-049-LD</td> <td>SP8-FR-050-LD</td> </tr> </table>	SP8-FR-001-G	SP8-FR-002-G	SP8-FR-003-G	SP8-FR-004-G	SP8-FR-005-G	SP8-FR-006-G	SP8-FR-007-G	SP8-FR-008-G	SP8-FR-009-G	SP8-FR-010-G	SP8-FR-022-WeP	SP8-FR-023-WeP	SP8-FR-024-WeP	SP8-FR-025-WeP	SP8-FR-026-WeP	SP8-FR-027-WeP	SP8-FR-028-WeP		SP8-FR-029-DF	SP8-FR-030-DF	SP8-FR-031-DF	SP8-FR-032-DF	SP8-FR-033-DF	SP8-FR-034-DF	SP8-FR-035-DF	SP8-FR-036-DF	SP8-FR-037-DF	SP8-FR-038-DF	SP8-FR-039-DF	SP8-FR-040-DF	SP8-FR-041-LD	SP8-FR-042-LD	SP8-FR-043-LD	SP8-FR-044-LD	SP8-FR-045-LD	SP8-FR-046-LD	SP8-FR-047-LD	SP8-FR-048-LD	SP8-FR-049-LD	SP8-FR-050-LD		
SP8-FR-001-G	SP8-FR-002-G																																												
SP8-FR-003-G	SP8-FR-004-G																																												
SP8-FR-005-G	SP8-FR-006-G																																												
SP8-FR-007-G	SP8-FR-008-G																																												
SP8-FR-009-G	SP8-FR-010-G																																												
SP8-FR-022-WeP	SP8-FR-023-WeP																																												
SP8-FR-024-WeP	SP8-FR-025-WeP																																												
SP8-FR-026-WeP	SP8-FR-027-WeP																																												
SP8-FR-028-WeP																																													
SP8-FR-029-DF	SP8-FR-030-DF																																												
SP8-FR-031-DF	SP8-FR-032-DF																																												
SP8-FR-033-DF	SP8-FR-034-DF																																												
SP8-FR-035-DF	SP8-FR-036-DF																																												
SP8-FR-037-DF	SP8-FR-038-DF																																												
SP8-FR-039-DF	SP8-FR-040-DF																																												
SP8-FR-041-LD	SP8-FR-042-LD																																												
SP8-FR-043-LD	SP8-FR-044-LD																																												
SP8-FR-045-LD	SP8-FR-046-LD																																												
SP8-FR-047-LD	SP8-FR-048-LD																																												
SP8-FR-049-LD	SP8-FR-050-LD																																												
	WP8.4 WP8.5 WP8.2 WP8.3 WP11.2																																												



			SP8-FR-051-LD SP8-FR-053-LD SP8-FR-055-LD SP8-FR-057-LD SP8-FR-059-LD	SP8-FR-052-LD SP8-FR-054-LD SP8-FR-056-LD SP8-FR-058-LD SP8-FR-060-LD		
SP8-UC-008	<p>Primary actor: Paul is an AU. He is a platform developer.</p> <p>Preconditions:</p> <ul style="list-style-type: none"> • Paul developed, tested and validated a new functionality of the rule-based clustering algorithm on a small dataset. • The algorithm has been deployed onto the data mining servers. <p>Description:</p> <p>Paul wants to run the new functionality of the data mining algorithm on the data available through the MIP. The results - the Biological Signature of Disease - will then be available for other users.</p> <p>Success scenario:</p> <ol style="list-style-type: none"> 1) Paul identifies the input data for the algorithm. He sets research criteria using the Ontology, Variables and Provenance descriptions. 2) Paul sets up the parameters of the data mining servers to execute the algorithm on the local data. 3) Paul runs the algorithm. 4) The results - the Biological Signature of Disease - are stored in the system. 5) Paul explores the rules on the variables that the algorithm has returned as biological signatures. 6) Paul registers the rules of the Biological Signature of Disease as new Variables in the Variables descriptions. He also describes the method used to obtain them in the Provenance database. 7) Paul validates the Biological Signature of Disease and releases it. The Biological Signature of Disease can now be used by other users. 	Not released fully. User-guide is available on how to import algorithms. Registration of the results is manual.	Data mining SP8-FR-001-G SP8-FR-003-G SP8-FR-005-G SP8-FR-007-G SP8-FR-009-G	SP8-FR-002-G SP8-FR-004-G SP8-FR-006-G SP8-FR-008-G SP8-FR-010-G		WP8.4 WP8.5 WP8.2 WP8.3 WP11.2



			SP8-FR-041-LD	SP8-FR-042-LD			
			SP8-FR-043-LD	SP8-FR-044-LD			
			SP8-FR-045-LD	SP8-FR-046-LD			
			SP8-FR-047-LD	SP8-FR-048-LD			
			SP8-FR-049-LD	SP8-FR-050-LD			
			SP8-FR-051-LD	SP8-FR-052-LD			
			SP8-FR-053-LD	SP8-FR-054-LD			
			SP8-FR-055-LD				



Annex D: Summary - Service IT Resource Planning

Table 6: Service IT resource planning

Product/Software Package/Service	TRL	Data Storage Capacity used by this Product	Data Storage Capacity Allocated for this Product	Location(s) of Data Storage	Data Access Protocol(s)*	Compute Resource(s) Allocated	Location(s) of Compute Resource(s) Allocated	Compute Access Protocol(s)**
Central research services - production	5	100 GB	32 GB	CHUV Academic network	Microservice architecture	2 Cores/16 GB Ram/4 GB OS 3 hbps2, hbps3, hbps4 4 Cores/32 GB Ram/8 GB OS 2 hpfed1, hpfed2 8 Cores/64 GB Ram/16 GB OS 2 hpmdw1, hpmdw	CHUV Academic network	API (java)
Central research services - QA and DEV	5	100 GB	1 6GB	CHUV Academic network	Microservice architecture	1 Core/8 GB Ram/2 GB OS 1 hpgate1 2 Cores/16 GB Ram/4 GB OS 1 hbps1,hpfedqa2,hpfedqa3 4 Cores/32 GB Ram/8 GB OS 2 hpfedqa1	CHUV Academic network	API (java)



EXAREME	5	10 GB	10 GB	Local Data Store Mirrors	REST API, JDBC	RAM: 4GB or more	CHUV	REST API
EXAREME algorithms	4	1 GB	1 GB	Local Data Store Mirrors	REST API	RAM: 4GB or more	CHUV	REST API
MIPMap	5	100 MB	100 MB	Local Data Store Mirrors		RAM: 4GB or more	CHUV	
WebMIPMap	5	40 MB	40 MB	Web Portal		RAM: 2GB or more	CHUV	
MIPMapRew	4	5 MB	5 MB	Web Portal		RAM: 2GB or more		
RAW	5	300 MB	300 MB	Local Data Store Mirrors	Files	RAM: 16GB or more	CHUV	REST API

* Data Access Protocols such as GPFS, N.FS, S3, Collab storage, etc.

** Compute Access Protocols such as EC2, Task Framework, Unicore, OCCI, Slurm, ssh, gLite, Condor, etc.



Annex E: Summary - Service Technology Readiness Levels (TRLs) Metrics

Documentation URL - User, Developer and/or Administrator documentation is available at this URL. Strong preference should be given for publicly available documentation services.

Target User Count (TRL6+) - Target user counts (concurrent service users).

SLA Defined - The software documentation defines some Quality of Service metrics in the service documentation. These metrics may or may not be enforced by the service itself. The service has not been tested to adhere to the documented QoS metrics.

SLA Monitored - The Quality of Service metrics are monitored by a monitoring service.

SLA Enforced - The Quality of Service metrics are enforced by implementing service. If the SLA Definition indicates on 3 API/request/sec/user, there are suitable mechanisms implemented in the service to ensure these limits are not exceeded.



Table 7: Service Technology Readiness Levels Metrics

Product/Software Package/Service	TRL	Documentation URL	Target User Count (TRL6+)	SLA Defined (TRL7+)	SLA Monitored (TRL7+)	SLA Enforced (TRL7+)	Comments
EXAREME	5	http://madgik.github.io/exareme/	500 registered users of which 30 users are executing simple queries and 15 users are executing complex queries concurrently, depending on the hardware capabilities.				Deployed for several data sources. Monitoring services in place.
EXAREME algorithms	5	https://github.com/madgik/mip-algorithms	1 user / process.				Deployed for one hospital (CHUV). Monitoring services in place.
MIPMap	4	https://github.com/aeub-wim/MIPMap	1 user/process (desktop application).	NA	NA	NA	Deployed for one hospital (CHUV). Monitoring services in place.
WebMIPMap	4	https://github.com/aeub-wim/WebMIPMap	As many users as web portal supports (1 instance/web portal user).	NA	NA	NA	Deployed for one hospital (CHUV).



							Monitoring services in place.
MIPMapRew	4	https://github.com/aueb-wim/MIPMapRew	As many users as web portal supports (1 instance/web portal user).	NA	NA	NA	Deployed for one hospital (CHUV). Monitoring services in place.
SP8 bundle query engine	5	https://github.com/HBPSP8Repo/SP8LocalMirrorBundle/blob/master/RAW_README.md	As many users as web portal supports (1 instance / web portal user).	NA	NA	NA	
Central research services- microservice architecture	4	https://github.com/LREN-CHUV/portal-backend https://github.com/LREN-CHUV/portal-frontend https://github.com/LREN-CHUV/portal-specs https://github.com/LREN-CHUV/bootstrap-mip-app https://github.com/LREN-CHUV/mip-apps-manager	The main user interface for the MIP.	4			
MIP User Knowledge Base	6	https://www.liferay.com/	An easy way of accessing the information about the MIP (guidelines), provide feedback and interact with the MIP administrators, developers and team.				TRL6, only because we don't have the proper SLA monitoring and rule enforcement in place.



Research and services: Modelling EE/IA/BSD	4	https://github.com/LREN-CHUV/portal-backend https://github.com/LREN-CHUV/portal-frontend	The portal back-end provides the services consumed by the front-end. This includes user management and login, reference data, controlled access to the various back-end systems.				
Algorithm Factory	4	https://github.com/LREN-CHUV/woken	Executes the self and custom data analytics algorithms written in a variety of languages.				TRL4 but backed by mature technologies (Docker, Mesos, etc.).
Microservice architecture		https://github.com/LREN-CHUV/mip-microservices-infrastructure	Provides support for deploying microservices in a cluster, monitoring and recovery.				Deployed at CHUV and monitoring services in place.
SP8 bundle		https://github.com/HBPSP8Repo	Software stack that will run at every participating hospital or medical centre of the FNHC of the MIP of the HBP.	NA	NA	NA	Deployed for one hospital (CHUV).



SP8 bundle query engine		https://github.com/HBPSP8Repo/SP8LocalIMirrorBundle/blob/master/RAW_README.md	First version of the query engine deployed at a local data store mirror (CHUV).				
Anonymization Module	6	http://gnubila.fr/	Data to be used in the context of the MIP needs to be anonymized twice: first when exported from the hospital systems and second when queried.				Deployed for one hospital (CHUV).
Administration User Interface	4	https://github.com/HBPSP8Repo	Available to manage/configure Hospital Bundle.				Part of Raw db component.
MIP Application: GeneExpression	3	https://github.com/LREN-CHUV/gene-expression	Find correlations between gene expression and brain area.				Results are viewable in the biological signature viewer.
MIP Application: Biological Rules viewer	3	https://github.com/LREN-CHUV/portal-frontend/tree/master/app/scripts/external/ViewerProject	3D viewer for the biological signatures of disease.				Results are viewable in the biological signature viewer.
MIP Application: Graph Bargraph	2	JavaScript code	Bar graph viewer for the biological signatures of disease.				
MIP Application: Graph Sunburst	3	https://github.com/LREN-CHUV/sunburst					
MIP Application: Brain 3D viewer	3	Javascript code available					
MIP Application: Brain 2D atlas	3	Matlab code available					



MIP Function – Semi-supervised rule based clustering algorithm	3	https://github.com/LREN-CHUV/woken	Aims to explain the variability between individuals, and describes a population by a group of “local over-densities”.				
MIP Function – Informatics-based Model: Enriched Automated Diagnostic Tools	3	https://github.com/LREN-CHUV/woken	This is an automated classifier from a set of MRI scans that came from deceased, pathologically diagnosed individuals. This classifier provides prognostic value on clinically categorised living people.				
MIP Function – Informatics-based Model: Deep Learning for Automated Features Extraction	3	Matlab code available	Include image and face recognition (1, 2, 3), speech recognition (1, 2) and signal processing (1).				
MIP Function – Informatics-based Model: Rasch model and factor analysis for learning disease severity	3	Matlab code available	Extract an index or a latent variable from the neuroimaging data to quantify the disease severity for each subject and regional vulnerability by applying factor analysis				
MIP Function Informatics-based Model: Bi-clustering applied to gene expression and brain volumetric data	3	Matlab code available	Bi-clustering method used to compare two datasets with one common dimension				
MIP Function Informatics-based Model: Bayesian Causal Model	3	https://github.com/LREN-CHUV/woken https://github.com/LREN-CHUV/hbplregress	Designing, equations derivation and implementing the causal Bayesian model similar to the General Linear Model (GLM) for distributed Data				



MIP Function – Disease subtypes signatures - Big medical data strategy (3-C)	3	https://github.com/HBPSP8Repo/CCC	Categorize, Cluster & Classify (3-C) - Disease subtype signatures - Big medical data strategy				
MIP Function – Label Propagation Framework	3	http://www.fil.ion.ucl.ac.uk/~john/LabelProp/Label%20Propagation%20Framework.pdf	To be based on a number of brain structure volume features, which are automatically extracted from patient MRI scans.				
MIP Function – Multi-Target Regression on Data Streams	3	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-moa-trees.pdf	The data stream mining algorithm for predicting structured target variables.				
MIP Function – Predictive Clustering Trees	3	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-clus-trees.pdf http://source.ijs.si/hbp/clus/wikis/documentation	The predictive clustering tree algorithm for predicting structured target variables.				
MIP Function – Rule Ensembles	3	http://source.ijs.si/hbp/clus/wikis/documentation	The rule ensemble algorithm for predicting structured target variables.				
MIP Function – Feature Ranking for Structured Targets	3	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-clus-franking.pdf http://source.ijs.si/hbp/clus/wikis/documentation	The feature ranking algorithm for structured target variables.				
MIP Function – Subgroup Discovery from Multi-Resolution Data	3	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-hedwig.pdf https://github.com/anzev/hedwig/tree/hbp	A subgroup discovery tool that can use ontological domain knowledge (RDF graphs) in the learning process. Subgroup descriptions contain terms from the given domain knowledge and enable potentially better generalizations.				
MIP Function –	3	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-tehin.pdf	The subgroup discovery algorithm for analysis of heterogeneous data.				



Subgroup Discovery from Heterogeneous Data							
MIP Function – Visual Performance Evaluation	3	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-viper.pdf http://viper.ijs.si/about/	The ViperCharts web-based platform for visual performance evaluation of data analysis algorithms.				
MIP Function – Brainspan co-expression clustering	3	http://lvdmaaten.github.io/tsne/	To extract exploit disease patterns using co-expression network analysis in Human Brain Transcriptome.				
MIP Function – BH-tSNE	3	http://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf http://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf	tSNE has been developed by Laurens van der Maaten as a non-linear dimensionality reduction (DR) algorithm to visualize high-dimensional data.				

Performance indicators and benchmark data

EXAREME

We examine the scaling characteristics of response times of the EXAREME federation engine and RAW query engine based on the first four use cases (ref: Deliverable D8.6.1). The two important factors that affect the system performance are the number of participating hospital and the number of simultaneous user requests.

The setup simulated three hospital nodes loaded with the ADNI dataset. The workload simulated 1-10 simultaneous users sending queries to the system. The response time, of each individual user's query, increased by 20% when going from 1 to 10 simultaneous users. The response time remained the same when going from 1 to 3 hospitals and increasing the dataset by 3 times. The above experiments can be reproduced using the Exareme/Raw (<https://github.com/madgik/exareme/tree/mip>) along with the test script (mining-api-test.sh) in the Exareme source code.

MIPMap

The response times of the user interface of MIPMap and WebMIPMap are characterized as interactive since their order of magnitude is in the sub-second scale.



However, in order to measure the loading and execution time needed for MIPMap to perform data translation (WebMIPMap does not support this functionality), we ran the following experiments. The experiments were performed on a desktop computer running Windows 7, with intel zeon 3.20GHz processor, 2 cores and 6GB of RAM.

We have created five mapping tasks of different sizes and complexity shown in Table 8. The nomenclature used in the table is defined below:

- #S: number of source tuples
- #T: number of target tuples created
- S: source size
- T: target size
- #J: number of joins in the source schema
- #Sel: number of selection conditions
- #F: number of functions in the mapping task
- #FK: number of foreign keys in the target schema
- #TGDs: number of TGDs in the mapping task

Table 8: Task description

	#S	#T	S	T	#J	#Sel	#F	#FK	#TGDs
Task1	~150	~70	5 KB	4,50 KB	1	1	6	1	4
Task2	~2500	~350	222 KB	26 KB	2	1	7	2	4
Task3	~54000	~54000	3 MB	4.70 MB	2	0	10	0	10
Task4	~93000	~93000	8.10 MB	8.60 MB	1	0	12	0	8
Task5	~1850000	~1236000	200 MB	152 MB	2	1	22	12	12

For these mapping tasks we have created the TGDs needed in order to translate them to the MIP schema, and in the following executed each mapping task. Note that when a mapping task is created or loaded in MIPMap, the source data are not loaded until the first time the mapping



task is executed. However, after the first execution the data are not loaded again. The results are shown in Table 9 and are measured in seconds, while the meaning of each column is defined below:

- t_{TGD} : TGD generation time
- t_{Trl} : Translation time taking into consideration the time needed to load the data (i.e. first execution of the mapping task)
- t_{Tr} : Translation time without taking into consideration the time needed to load the data ()
- t_{Totl} : Overall time taking into consideration the time needed to load the data ($t_{TGD} + t_{Trl}$)
- t_{Tot} : Overall time without taking into consideration the time needed to load the data ($t_{TGD} + t_{Tr}$)
- t_{Exp} : Time needed to export the created data to CSV files

Table 9: Evaluation times

	t_{TGD}	t_{Trl}	t_{Tr}	t_{Totl}	t_{Tot}	t_{Exp}
Task 1	0.552	1.925	1.531	2.477	2.083	0.123
Task 2	0.186	2.733	1.781	2.919	1.967	0.377
Task 3	1.605	8.340	7.197	9.945	8.802	0.350
Task 4	0.499	12.365	7.284	12.864	7.783	0.714
Task 5	1.241	324.300	79.109	325.541	80.350	10.142

RAW

We compared RAW with a range of systems and configurations to see how well it performs i) versus systems that at some point were extended to support richer data models, and ii) versus systems specialized for a specific scenario by design. We consider both relational and hierarchical data. For the latter, we choose the JSON data format because of its popularity as a data exchange format, which also explains why multiple relational DBMS have been adding support for it as a datatype.

We compare RAW against i) PostgreSQL 9.4.1, ii) DBMS X (a commercial DBMS), iii) MonetDB 11.19.9, iv) DBMS C (a commercial column store), and iv) MongoDB 3.0.3. We include PostgreSQL and DBMS X to our comparison because they support both relational and JSON data, to observe how a generic system performs in the two diverse cases. The MonetDB and DBMS C column stores are designed to efficiently support relational



analytical queries, and have also recently added JSON support. Finally, we use MongoDB as a specialized system for JSON data. MongoDB uses a binary serialization (BSON) for JSON data. PostgreSQL supports both a binary (jsonb) and a character-based serialization; we use jsonb because of its efficiency. The other systems essentially treat JSON as a subtype of VARCHAR. Neither the systems we compared against, nor RAW, make any assumption about the field order in the JSON files.

All experiments are run on a dual socket Xeon Haswell CPU E5-2650L (12 cores per socket @ 1.80 GHz), equipped with 64KB L1 cache and 256 KB L2 cache per core, 30 MB L3 cache shared, 256 GB RAM, and 2TB 7200 RPM SATA 3 disk storage. The operating system is Red Hat Enterprise Linux 7.1 (Maipo). RAW uses LLVM 3.4 to produce custom code for each query, the compilation time being a few milliseconds per query. MonetDB and PostgreSQL were built using gcc 4.8.3, with optimization flags on. We run all systems in single threaded mode.

To generate input datasets for the experiment, we use the data generator of TPC-H for the lineitem and order tables using scale factors 10 (SF10 - 60 million lineitem tuples, 15 million order tuples) and 100 (SF100 - 600 million lineitem tuples, 150 million order tuples). We shuffle each file's contents to avoid any potential optimizations that might exploit interesting orders, which could introduce noise to our experiments. To test performance over JSON data, we convert the TPCH-SF10 version to JSON, resulting in a 20GB file, and load it in PostgreSQL, MonetDB and MongoDB. RAW natively operates over the JSON file, making no assumptions about field order, and building a positional index during the first data access.

Projections

For queries projecting a varying number of fields, we use three variations of the following template:

```
SELECT AGG(val1), ..., AGG(valN) FROM lineitem WHERE l_orderkey < [X].
```

The first two variations compute one aggregate, COUNT and MAX respectively. The third variation computes four aggregations (COUNT or MAX). We examine the difference in performance for queries with selectivity 10%, 20%, 50%, and 100%.

The figure below presents the results for the JSON version (SF10). The support for JSON is not yet mature in MonetDB, which results in having the worst performance in all cases. Similarly, DBMS-C underperforms in all the experiments we conducted over JSON data. As for the other systems, JSON access is also expensive for DBMS-X because it uses a character-based encoding. MongoDB is competitive with PostgreSQL only for the COUNT() query. In the rest of the cases, and as the amount of aggregates to compute increases, PostgreSQL outperforms MongoDB. RAW has the best performance; its lightweight generated code path makes it more efficient for the CPU-intensive task of processing JSON entries. In addition, contrary to PostgreSQL, RAW does not treat entire JSON objects as bulky BLOB data; RAW uses the positional index to retrieve the information it needs from each object, which it then feeds in the query pipeline without "polluting" the CPU caches with the verbose JSON object any further.

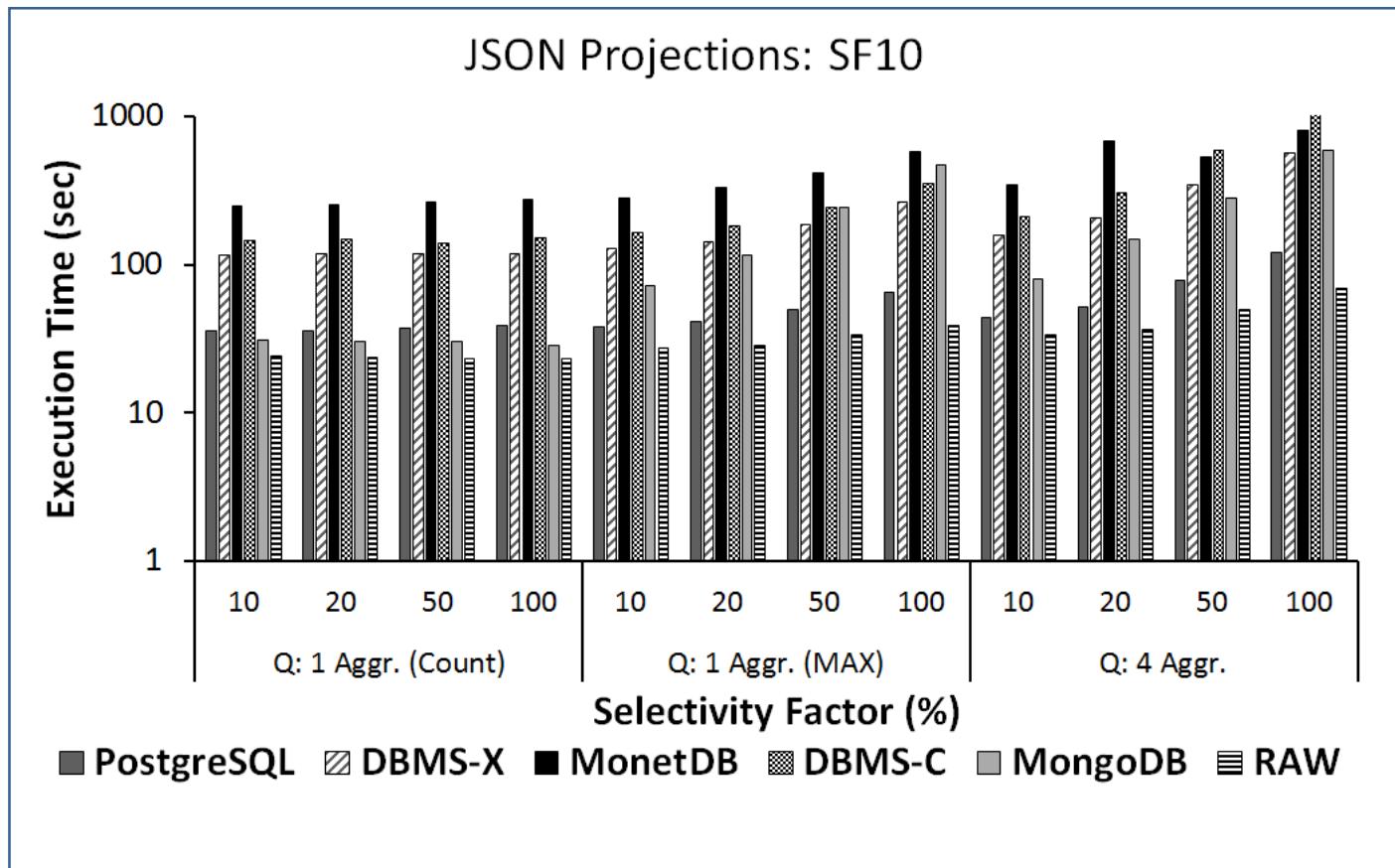


Figure 27: Projection-intensive queries over JSON data

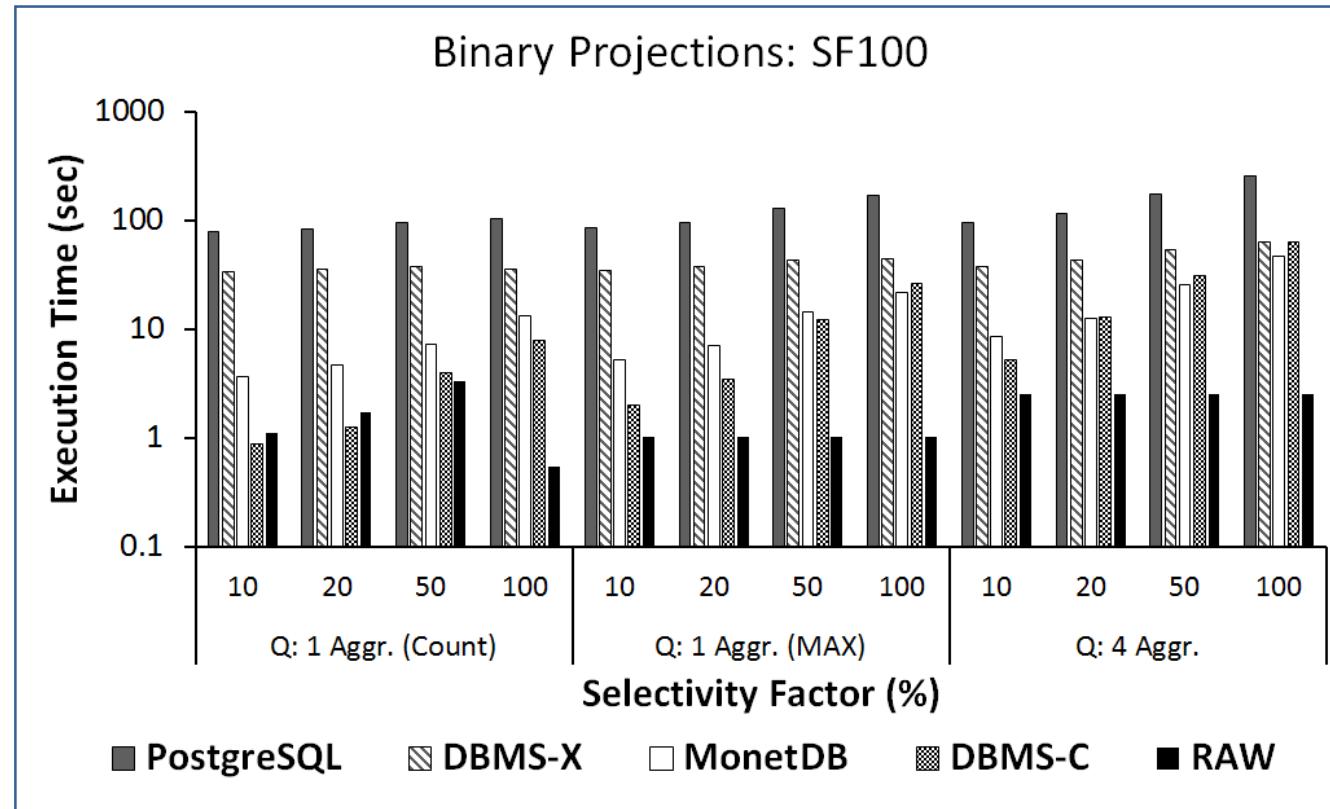


Figure 28: Projection-intensive queries over binary relational data

The figure above presents results for the queries over binary data (SF100). In this case, MonetDB and DBMS-C outperform PostgreSQL and DBMS-X because the analytical query template we study is more suitable for column-oriented engines (i.e., a small subset of the relation is accessed). For the simplest COUNT query template, DBMS-C is the fastest for selective queries because although we shuffled the original input data, DBMS-C sorts it at loading time by design. Given that the query has a predicate on the sorting key, DBMS-C skips many data entries while answering the query. In addition, this query does not project any attributes, therefore DBMS-C does not incur any tuple reconstruction cost. For less selective instances of this query and for the other queries that are more complex, RAW outperforms both DBMS-C and MonetDB; their columnar operators produce intermediate results (i.e. they fully materialize their output), thus paying a materialization cost for the columns involved. This cost increases as queries become less selective; RAW pipelines data instead. In addition, the resulting code of RAW is a tight, minimal for-loop which only contains an if block evaluating the selection condition. Hence, it is very branch- and cache-friendly. The importance of generating minimal code is highlighted in the calculation of the COUNT query (left side of the figure above). The generated



code is minimal enough for the effect of the branch predictor to be visible. When selectivity reaches 100%, very few mispredictions occur, therefore the query becomes much faster for RAW, although intuitively RAW has to do more work to calculate the aggregate value.

Selections

To test queries with multiple selection predicates, we use three variations of the following template:

```
SELECT COUNT(*) FROM lineitem WHERE val1 < [X] AND ... AND valN < [Z].
```

The examined queries include one, three, and four predicates in the WHERE clause respectively. We examine the performance difference for queries with 10%, 20%, 50%, and 100% selectivity.

The figure below presents the results over JSON data (SF10). We observe that RAW outperforms the other systems across the whole experiment. DBMS-X is the slowest because of its character-based JSON encoding. Compared to Figure 1, MongoDB closes the gap from

PostgreSQL and RAW. The reason is that the current query template projects out a count instead of more complex aggregates which MongoDB does not compute as efficiently. RAW converts the values it needs on the fly, whereas PostgreSQL and MongoDB operate over a more efficient binary serialization. Still, RAW outperforms the other systems because it reduces the rest of the CPU overheads significantly. Besides pipelining, it consults its positional index to pinpoint needed values, thus reducing navigational cost in the file. These benefits become more apparent as queries for less selective queries. In the case of binary data presented in Figure 30, the outcome is similar to the one for projection queries. RAW is faster in the majority of cases because it pipelines data through all operators. MonetDB and DBMS~C operators materialize their output, which becomes more expensive as selectivity moves towards 100%.

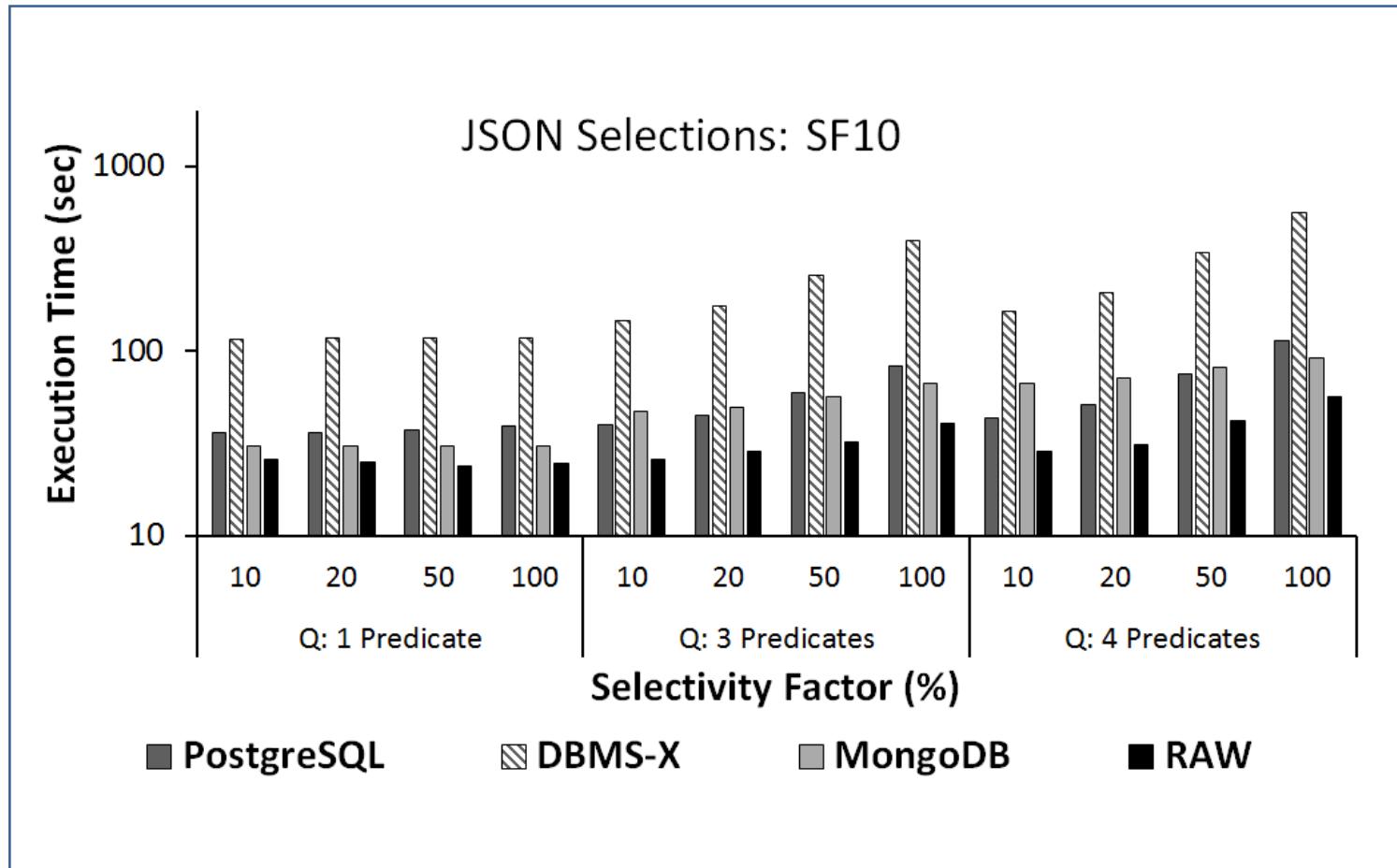


Figure 29: Selection queries over JSON data

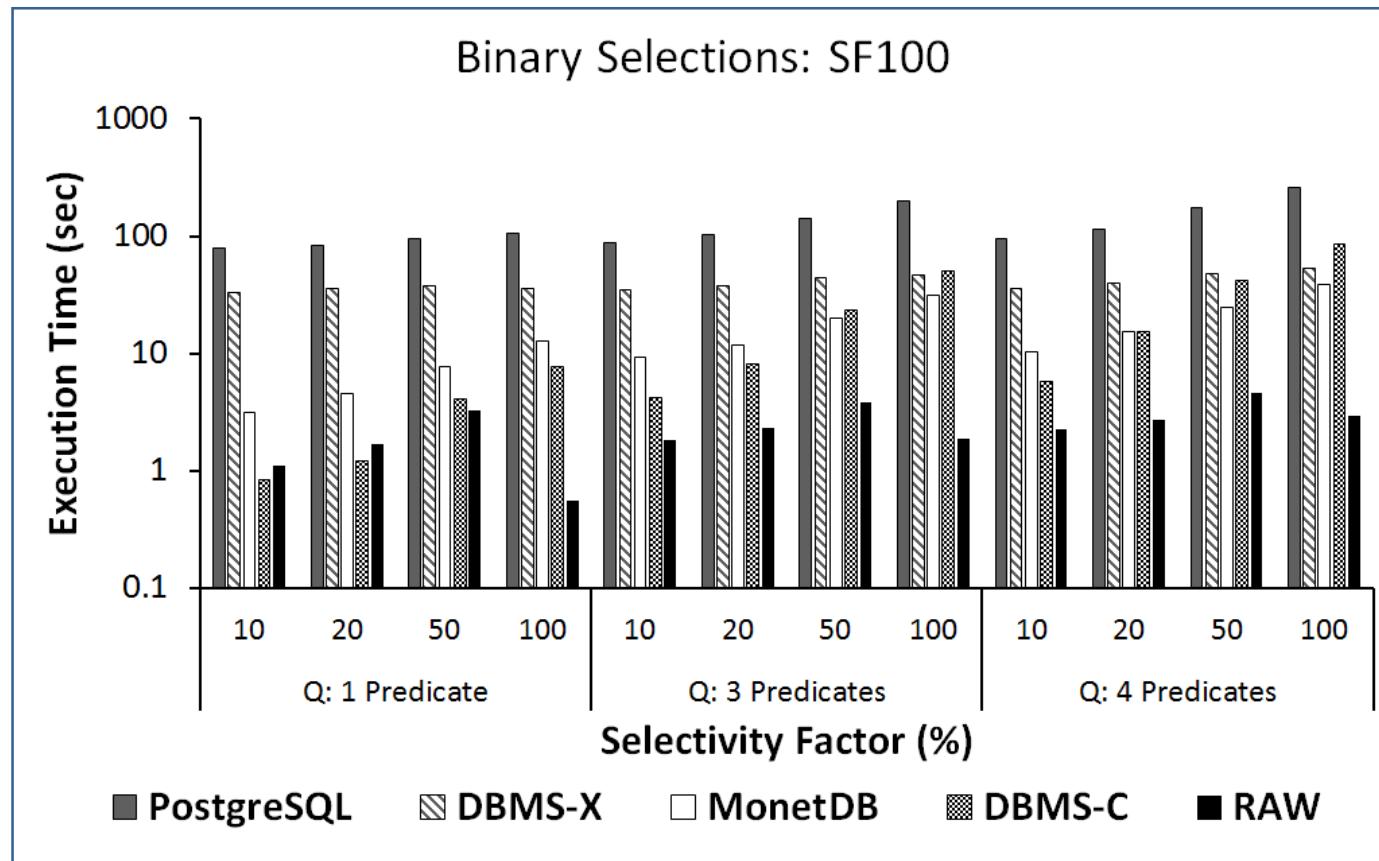


Figure 30: Selection queries over binary relational data

Our evaluation demonstrates how RAW serves a challenging real-world workload efficiently: it outperforms state-of-the-art open-source and commercial approaches without being tied to a single data model or format, all while operating transparently across heterogeneous data. Its ability to morph based on the query requirements opens multiple opportunities for further optimizations.



Annex F: Backlog (Remaining bugs and new features to be added)

For each Product/Software Package/Service, please find complete two tables, one for bugs and one for new features to be added).

The MIP is currently under testing and the list of bugs will be provided in the final release of the present report (<https://mip.humanbrainproject.eu/help/testing-quality>)

Product/Software Package/Exareme

Remaining Bugs

Bug ID	Related Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments

Features

Feature ID	Necessary for Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
	All use cases.	Iterative workflow at global level (between the central node and hospital nodes)		Although Exareme supports iterative workflow computations at the local level (iterations inside the hospital nodes or inside the central node), it does not support iterative workflows at the global level (iterations between the central node and hospital nodes). This missing functionality is critical for certain data mining algorithm' classes.



Product/Software Package/MIPMap

Remaining Bugs

Bug ID	Related Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
No limitations, problems or bugs exist.				

Features

Feature ID	Necessary for Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
	NA	Integration of MIPMapRew into WebMIPMap and the Web portal		Testing dataset need to be created.



Product/Software Package: MIP Computation Services

Remaining Bugs

Bug ID	Related Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments

Features

Feature ID	Necessary for Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
1	Y (for developer users - development. For general users - running integrated algorithms)	Integrate off-the-shelf machine learning algorithms	Jira	<p>Technology: use current dockerised back-end infrastructure and develop the service as a new docker image that runs a JAVA program, calling the RapidMiner framework.</p> <p>The service will Train the 2 classifiers Naive Bayes and k-nearest neighbours and will validate them using standard 10-fold cross-validation.</p> <p>It will also write validation results consisting in the accuracy measure with confidence intervals as well as the confusion matrix into the result database.</p> <p>Benefits: allow users, via the existing Web UI, to select any combination of variables for which models can be trained and validated using generic data-driven machine learning algorithms. The results of the validation (e.g. accuracy, confusion matrix, any other relevant metrics) and properties of the trained models (e.g. relevance of the features) would be then displayed to the user.</p> <p>This service will allow any other algorithms developed in JAVA to be integrated in MIP.</p>
2	Y	Integrate Matlab technology	Jira	<p>So that developer users can integrate Matlab code directly, without translation in other languages.</p> <p>R is already integrated, and feature 1 above will allow to integrate JAVA.</p>
3	N	Full automation of Data Factory	Jira	Data Factory = all tools, processes and activities carried out to "produce data" for MIP use. Ex: curation and import of data, production of features etc.



				<p>Output: data catalogue describing the variables in MIP ontology, summary stats about variables (counts), provenance.</p> <p>Process is in place, semi-automated, but we want to fully automate it.</p>
--	--	--	--	---



Product/Software Package: Information & Scientific Reference Services

Remaining Bugs

Bug ID	Related Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments

Features

Feature ID	Necessary for Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
1	Y	Elastic Search to store and manage Web UI information such as ontologies and variable dictionary, statistical descriptive statistics, catalogue of live algorithms etc. This is an important crosslink with the other platforms, i.e. SP5 ontology.		



Product/Software Package: Web UI

Remaining Bugs

Bug ID	Related Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments

Features

Feature ID	Necessary for Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
1.	Y	Apply filters to the dataset of interest when designing and visualising models	Jira	Currently users can design models by categorising the variables of interests (Epidemiological Exploration). They can also design a filter (e.g. to only consider subjects of particular age range), BUT the filter is currently not applied to the dataset. This feature will apply the filter to the dataset selected.
2.	Y	Create "Bins" of variables	Jira	Permit users to interactively and easily define up new variables based on existing ones, including simple mathematical operations (e.g. group two types of variables, the product of two variables etc.).
3.	Y	Save estimated model and results into the Research Object (RO) of the user.	Jira	Currently, users can estimate models by applying methods. The new feature will allow the estimated model and results to be saved to the user's Research Object.
4.	N	Graphic visualisation of estimated models (not mandatory, but ideal for an effective user experience)	Jira	Currently only displayed as table of statistics, but ideally should be also visually displayed. Challenge: output is dependent on the method selected - tree, graph, plot etc.
5.	Y	Review models and estimations	Jira	Linked to (3), allow user to return to any of the steps in the flow - model design, model estimation, apply changes and save as new model or estimation.



6.	Y	Model comparison	Jira	Linked to (3) , to allow users to compare own estimations to of other users. By doing so, users can see for example risk factors for different demographic groups (females vs males, between BMI values etc.).
7.	N	Permit users create their own apps	Jira	Apps section already exists on MIP UI, but the process of allowing external users to directly contribute is under development, including the guidelines.
8.	N	Create "News" section	Jira	To alert users of functionality to be released, scheduled down times etc. (Might be done in the Knowledge Base, TBC).
9.	N	TBC: open vote for proposed functionality (as part of Use Engagement programme)	Jira	So that all users have the chance to express their weight on a particular proposed new function (proposed by MIP team or similar users).
10.	Y	Expand Biological Signature of Diseases service	Jira	
11.	N	User Profile - Implement the "Settings" functionality	Jira	
12.	N	Search across MIP	Jira	Permit user to search for terms across the whole platform - articles, models etc.
13.	N	(Virtual) Interest Groups	Jira	Allow users to belong to certain interest groups and so build virtual interest groups (aimed all to be dynamic, based on most used words in own profile - TBC).
14.	N	Algorithm Factory Community Platform (long term)	Jira	Interactive platform to allow the developer community build methods together - build, test, share them, vote, package and deploy them into the MIP pipeline. (may only be integration of existing platforms with MIP - TBC).



15.	N	Business KPIs (aka platform usability stats)	Jira	For development strategy: online reports will be designed and created to timely report on the usage of the platform - what is being viewed, what not (methods, data, variables, functionality), most used searches etc. This will inform about what new data or functionality that might be required, functionality that is not used etc, and it will help to define the MIP product roadmap.
16.	Y	Enrich visualisation types	Jira	e.g. tree structures (TBC).
17.	Y	Enrich the catalogue of methods and Data Mining algorithms available on Web UI	Jira	Dependent on the Algorithm Factory.
18.	Y	Allow selection of group of variables	Jira	
19.	N	User Role management	Jira	To be reviewed, especially when external method developers will be able to deploy own methods.

Note: the above Web UI features have been proposed by the internal users. Since after M30 we will start involving external users, the list is subject to change.



Product/Software Package: MIP User Knowledge Base

Feature ID	Necessary for Use Case(s) or FR(s)	Short Description	Bug tracker URL (if available)	Comments
1	N/A	Working with media	N/A	KB will deploy a “built-in media” to support and enable the MIP course organizers to easily search for and insert video / audio files in their courses.
2	N/A	Lessons, courses, curriculum and categories management.	N/A	Create ad-hoc personalized lessons for MIP end-users. Every file (.ppt; .pdf, etc.) could be moved / merged / split dynamically to re-arrange new lessons and courses.
3	N/A	Certification of the MIP users in the correct usage of the MIP tools	N/A	The KB will qualify with specific certificates users before using MIP tools. This process should be done iteratively once new MIP features are added.
4	N/A	Learning Transcript and Progress tracking	N/A	The KB will show all user's learning history and statistics (i.e. list of courses, lessons attended for each course, how many attempts a user do to accomplish a lesson and get a certificate, list of user's valid and expired KB certificates) to track their progress. For the benefit of end-users, KB can also print the transcript to a PDF allowing users to download it.
5	N/A	Customizable notification system through email and social tools (e.g.: Facebook)	N/A	Users, registered to the MIP, will be informed as soon as new features will be deployed.



Annex G: IPR Status, Ownership and Innovation Potential

Product/Software Package/Service	IPR Status*	Owner(s)	Non-HBP users**	Innovation Potential***
EXAREME	Open Source	MADgIK group, University of Athens	OpenAIRE project, Optique project, MD-Paedigree project	Distributed processing
MIPMap				
RAW				
Anonymizer	Licenced by Gnúbila	Gnúbila	Licenced by Gnúbila	Commercial
EXAREME algorithms	Open Source	MADgIK group, University of Athens		Distributed Privacy Preserving Data Mining Algorithms
Woken	Open source	CHUV-LREN	In TBR	Web analytics platform

* IPR Status: Open Source, Copyright, Patent, Trade Secret, pre-IPR (i.e. you intend to obtain some form of IPR in the future)

** If this product/software package/service is currently being used outside HBP (e.g. donated, loaned, licensed, sold), please specify by whom.

*** Innovation Potential: Potential practical applications beyond HBP, commercial and/or non-commercial.



Annex H: Medical Informatics Platform Data

The table below gives an overview of the data available at M30 in the MIP.

Data Provenance	Data Type*	Features Captured**	Number of Data Sets	Clinical Data Standard Conformity***
Alzheimer's Disease Neuroimaging Initiative (ADNI) ADNIMERGE R package (version 0.0.1)	Research: - MRI results - FDG-PET imaging - CSF biomarkers - Neuropsychological tests - Proteins	200 relevant features (cognitive domains, demographic data, neuropsychological exams, etc.)	1500 Available in the platform	Neuromorphometrics
3-C	Population cohorts : - MRI results - Neuropsychological tests	Brain features Genetics SNP	3000 Available in the platform	Neuromorphometrics
CHUV	Diagnostic, blood test	Diagnostic and risk factors	8600 Processing and curation stage: 2000 available in the platform	Modified version of ICD10
INDI	- MRI results - Neuropsychological tests	Brain features Diagnostic and risk factors	463 Processing and curation stage : <i>Used for data mining</i>	
EDSD	- MRI results - Neuropsychological tests	Brain features Diagnostic and risk factors Demographics	934	Neuromorphometrics
Biobank	- MRI results and pathology	Brain features Diagnostic and risk factors Demographics	33	Neuromorphometrics Braak stages

* Patient, research, etc.

** Physiology, images, expression, cognitive, etc.

*** CDISC, C-FAST, etc.



The recruitment strategy of institutional providers of data to the MIP is shown in the diagram below.

Annex J: Note on Data Standardisation

The functionality of the MIP is largely based on the use of a federation that provides uniform access to data distributed in hospitals. In order for these hospitals to be interoperable and hence for the users to have a centralized experience of the available data, all the hospital LDSMs (that contain the data that each hospital contributes to the MIP) follow a common schema, the MIP schema that is an internal standard for the MIP. This schema is a generic (star) schema that captures information regarding demographics, examinations, brain features and genetic information. The design of the schema is largely based on the CDISC standard, in the sense that a large number of attributes of the schema are also present in the CDISC STDM terminology. However, the whole spectrum of CDISC STDM has not been covered, since this does not follow from the scientific users requirements so far; additionally, extra fields have been used to capture the available information. It is worth noting that a team, named Data Governance and Data Selection Workgroup, has been formed in SP8 for SGA1 that is dedicated to further address the issue of MIP schema standardization as the needs of the project evolve over time.

The generality of the MIP schema is perceived such that in order for the represented data to be interoperable, it stores information regarding the standard or ontology to which each (represented) variable conforms. The standards that are used to represent variables are given by the data provider, except for cases where there is straight-forward mapping of the variable to a standard. In case the standards are given by the data provider, in the form of mapping between the provided variables and the standard nomenclature or coding, MIPMap, a visual data exchange tool that performs the data translation process (i.e. transforms the data of the participating hospitals to populate each hospital's LDSM) utilizes this information to populate the LDSM with standardized data.

Taking into account the data that have been available so far in the MIP, the standards supported at the moment are the Cortical Labeling Protocol atlas for brain regions, ICD-10 for diagnoses, LOINC for clinical measurements (that is also the standard used by CDISC for clinical data), and dbSNP for genetic data.



Annex K: Hospital Bundle - Deployment Experience at CHUV

Deployment Steps

The steps were defined by a Project Manager, also acting as Deployment Manager. The steps were between CHUV-IT, CHUV-LREN (HBP: 8.2, 8.4, 8.5).

Theme	Action	Main responsibility to make sure the action is done			oct 2013-march 2014	april 2014-sept 2014	oct 2014-march 2015	april 2015-sept 2015	oct 2015-march 2016
		chuv itrc et dsi	hbp chuv (WP8.2, WP8.4, WP8.5)	hbp epfl					
Project Management	Define, monitor and adjust the steps for the deployment at CHUV								
data store miror	Define list of data to be extracted from DSI CHUV : HBP specification								
data store miror	Batch extraction script for images from PACS								
data store miror	Budget allocation to purchase Server and Network and storage								
data store miror	Acquire and set up Server and Network and storage								
data store miror	Create the data store mirror for RAW								
query engine	Define specifications of the needed queries								
query engine	Define specifications for the query engine RAW depending on the queries that need to be answer and the data source								
data store miror	Choose standard ontologies as models create of meta data dictionary of variables (provenance, definition, quality control, type) : ontology								



data store mirror	Map the CHUV data into a pilot-model (link it to a standard ontology) : it will serve as model for the 4 other hospitals)								
query engine	Build and implement the query engine RAW locally at EPFL : dev, prod, access conditions								
query engine	Test the query engine RAW at CHUV with the Data store mirror								
query engine	Deploy the query engine RAW at CHUV								
federation infrastructure	Define provenance model for the federated infrastructure for the data and the querying (for each data item, attach origin)								
query engine	Implement provenance capabilities for RAW (bring at same time data and origin information)								
pre-processing	Implement provenance capabilities for workflow (link info on pre-processing of data with data)								
pre-processing	Implement automatic annotation of data (which hospital)								
anonymisation	Define strategy for data privacy : define the field in the images and other data that need to be hidden for privacy								
anonymisation	Define strategy for data privacy : decide if a coding-decoding module is necessary								
anonymisation	Depending on global agreement already signed check if CHUV ethic committee validation necessary with Mr Savary								
anonymisation	Extract and anonymized patient data from CHUV EMR based upon the HBP specification (molis, axya, soarian, pacs)								
data store mirror	Populate the data store mirror with anonymized patient data : one shot								
data store mirror	Refresh the data store mirror every 6 months : extract								
data store mirror	Refresh the data store mirror every 6 months : populate								
administration HBP CHUV	Integrate the 1 new HBP CHUV collaborators : 1 Financial Administrative officer								



administration HBP CHUV	Recruit and integrate the 3 new HBP CHUV collaborators : 1 IT engineers, 1 post doc, 1 Project manager							
administration HBP EPFL	Recruit and integrate the 1 new HBP EPFL collaborator : 1 software engineer							
pre-processing	Curation quality data check of the data store mirror, for images and biologic data							
pre-processing	Implement and test of image pre-processing software and fine tuning on matlab, for images							
pre-processing	Implement on top of matlab scripts able to pre-process data : transforming images into variables							
pre-processing	Curation quality data check of the data store mirror, for genomic data							
pre-processing	Pre-process data : transforming genomic data into variables							
pre-processing	Implement and test of pre-processing software and fine tuning on R for genomic data							
pre-processing	Analyse variables from processed data based on the users query (User Define Functions)							
pre-processing	Create workflow to wrap the operation sequence that will extract and process automatically the data (transforming images and genomic data)							
federation infrastructure	Define the specification of software for the federation infrastructure							
federation infrastructure	Choose a subcontractor to build the federation infrastructure in 2 steps : qualification round, tender process							
federation infrastructure	Manage the subcontractor that will build the federation infrastructure							
federation infrastructure	Test of a versions of the federation software within the CHUV IT environment							
federation infrastructure	Integrate of the productive version of the federation software within the CHUV IT environment							
anonymisation	Choose a subcontractor to implement the anonymization software : 3 companies asked because <230 kchf							



anonymisation	Manage the subcontractor that will implement the anonymization software								
anonymisation	Check with the DSI security officer RSSI if this anonymization software is ok								
federation infrastructure	Check with the DSI security officer RSSI if this federation infrastructure is ok								
federation infrastructure	Integrate the query engine RAW of CHUV into federation architecture								
Web Portal	Create a web interface for end users								
Web Portal	Implement statistical models								
MIP platform	Create the DEV, QA and production environment								
MIP platform	Implement the MIP research services for supporting the use case								
MIP platform	Connect the MIP to the collab, MIP KB								

Deployment Experience (WP8.1)

Deployment was based on the docker (<http://docker.com>) container technology.

Preparation and packaging phase:

- 1) Prepare docker images per service.
- 2) Upload the images to docker hub.
- 3) Write shell scripts to start the docker containers in the correct order.

Deployment phase:

- 1) Clone deployment repo.
- 2) Use shell scripts to start/stop the services.



Lessons Learned

- 1) The process is greatly simplified by the use of docker & docker hub.
- 2) Still requires the administrator to login on the server, and execute the required scripts, which is fine for a low number of deployments (less than 10).
- 3) Not all software is configured to run into docker containers, which has proven to be less robust and requires more work for deployment and maintenance.
- 4) Due to extremely low resources at hospitals, personnel have limited availability, which requires a lot more resources and time than initially expected by SP8, for communication and administration (meetings, getting access, etc.)

Implications for the future

- 1) To enable better scalability of the federation to multiple hospitals, we are looking into distributed frameworks to deploy and manage remotely all the MIP services, in order to reduce management overheads.
- 2) We are looking into dockerizing the remaining services, as this has been proven to be a very effective way of managing the service.
- 3) Complete and improve the Platform with more advanced remote status monitoring.



Annex L: SP8 Hospital Bundle specifications

This document gives the specification and description of the Hospital Bundle functionality. It also provides information on how anonymous data is obtained, processed, and stored locally.