

# Connectome Stability Analysis using OpenWorm

## The AWESOME Connectome Proposal of Awesomeness

Proposal for a joint FWF/ARRS Stand-Alone Project

by

Andreas HOLZINGER

Holzinger Group, HCI-KDD, Institute for Medical Informatics, Statistics and  
Documentation, Medical University Graz, Austria

Graz , March, 15, 2016

### **Requested Funding:**

349,450 EUR (1 Postdoc and 1 PhD for the duration of 36 months)

**OEFOS 2012 Discipline:** 102 Computer Science - 033 Data Mining

**Keywords:** heterogeneous data, machine learning, visual data mining, information networks,  
ontologies, semantic data mining, redescription mining, feature selection, multi-view learning

**Note:** This proposal is original and has not been submitted to any other grant authority.

**Ethical Declaration:** This proposal does not raise any ethical issues.

**Remark:** This type of proposal is limited to a total number of 26 pages and the project  
duration is limited to a max. duration of 36 Months.

Find an  
actual  
title

Accurate?



Actual  
Date

Get ac-  
tual  
amounts

Confirm?

UPDATE!

**Abstract:**

This is a simple paragraph I inserted to preserve the formatting of the original abstract. It serves no purpose other than being a placeholder for when an actual abstract is written.  

## Contents

<b>1</b>	<b>Scientific Aspects</b>	<b>4</b>
1.1	Motivation for our Research . . . . .	4
1.2	Scientific Questions, Hypotheses and Goals . . . . .	5
1.3	Scientific Relevance and Innovative Aspects . . . . .	5
1.4	Importance of the expected results . . . . .	6
<b>2</b>	<b>Work plan</b>	<b>6</b>
2.1	WP 1: Enriched Heterogeneous information networks . . . . .	7
2.2	WP 2: Multi-view learning and feature selection . . . . .	10
2.3	WP 3: Workflows with parallel processing and meta learning . . . . .	13
2.4	WP 4: Data use case - Parkinson's disease . . . . .	15
<b>3</b>	<b>Organizational Aspects</b>	<b>17</b>
3.1	Work Organization, Supervision and Risk Management . . . . .	17
3.2	Strategies for Dissemination of Results . . . . .	18
3.3	Economic, Social and Practical Impact . . . . .	18
3.4	Career Benefits for those Involved . . . . .	19
3.5	Infrastructure . . . . .	19
3.6	National and International Cooperation . . . . .	20
3.7	Project Team . . . . .	21
<b>4</b>	<b>Financial Aspects</b>	<b>23</b>
4.1	In-kind contribution of partners . . . . .	23
4.2	Grant applications of partners . . . . .	23
<b>5</b>	<b>References</b>	<b>24</b>

# 1 Scientific Aspects

## 1.1 Motivation for our Research

The field of Connectomics is very promising in terms of furthering our understanding of nervous systems. While it has already very successfully increased our understanding of biological neural networks, connectomics still has several large challenges to overcome, to the point where over 20 years after the connectome of *c.elegans* has been found as the first connectome of a multi-cellular organism, we are still not able to fully accurately simulate its nervous system, with progress for more complex neural networks lagging even further behind. There are several distinct obstacles that hamper progress within this field:

1. The data collection itself has several technical limitations imposed on it. In order to accurately not only map a nervous system anatomically in its entirety, but also find the appropriate parameters that govern the operation of every single cell and synapse within it, a large variety of different techniques need to be utilized, and some of those may prove impossible to perform on a given organism.
2. As connectomics aims to analyse ever more complex organism, these technical difficulties get exacerbated by ethical considerations, which severely limit the possible techniques that can be used.
3. Finally, all these problems get severely inflated by the sheer amount of data required.

All these problems mean that finding the full connectome of any organism is a very lengthy task, albeit one that produces partial results at a constant rate.

Some techniques have in the past been used to work around this problem. From a neurophysiological point of view, some measurements that could not be conducted on e.g. *c. elegans* have been conducted on closely related species, with the results being extrapolated. More recently, evolutionary algorithms have been used to find missing parameters in order to be able to simulate the connectome accurately. This however will always raise the issue of whether the simulation can actually be used to infer explanations on the connectome underlying them, since it is ultimately possible that the simulation works entirely differently, only accidentally producing similar resulting behaviour.

There should be something more here...

I do not  
like this  
paragraph  
at all  
text

## 1.2 Scientific Questions, Hypotheses and Goals

The main scientific questions in our project are aimed at *feasibility* and *scientific value* of the use of simulations in *connectomics*. Having extensive experience in we will evaluate the current state of such simulations and their applicability to future studies on more complex organism than *c. elegans*.

something  
relevant

It is therefore our goal to explore and justify the following statements in detail. They are the key hypotheses of our proposal:

1. While current connectome simulations are not yet entirely accurate to every nuance of neuron operations, they provide highly accurate and useful data to be used to better evaluate possible future studies for their research value. \_\_\_\_\_
2. As macro-scale connectomes have been used to gain valuable insights into the inner workings of human brains despite only providing a quite abstracted and simplified, we will show that the same principle can be applied to micro-scale connectomes: That even simplified and incomplete models and simulations based on these can yield valuable insights and should be considered regardless of known or unknown inaccuracies to the actual known details of neuron operation. \_\_\_\_\_
3. The current simulations can be used to gain insight into the stability of the *c. elegans* nervous system, which will give useful data as to how optimized biologically evolved networks actually are. \_\_\_\_\_

Confirm?

Needs  
rephras-  
ing

Not  
happy  
with this.  
This has  
been  
done.  
Further  
study?  
Confir-  
mation  
study?

## 1.3 Scientific Relevance and Innovative Aspects

Connectomics are a promising field of study to further our understanding of nervous systems. However, progress in this field is slow due to the sheer amount of data required and the difficulty in measuring that data. This problem has been worked around by using macro-scale connectomes, which do appear to provide useful data while still presenting a significant level of abstraction from the actual workings of the brain.

Conversely, the field of micro-scale connectomics has focussed on providing detailed data on the exact workings of neurons and their interconnections into a nervous system to the point of best possible match to known parameters from in vivo measurements. While this discipline focusses on accuracy, the amount of data to be measured is quite staggering. Also, as the example of *c. elegans* shows, when the expected data set is completed, it often leads to important distinctions in hitherto disregarded details that now also need to be captured.

Ref

Given this reliance on a set of data that can take enormous amounts of time to compile, the ability to work with the incomplete datasets would allow further research, basing their studies on such simulations, to get a head start, allowing researchers to get preliminary findings before the dataset is entirely completed.

#### 1.4 Importance of the expected results

The main areas of contribution will correspond with our key hypotheses as well as the work package structure that follows; therefore we content ourselves with a very concise list at this point: /todoStill a copy of hedatbio

- Browser-based, visually composable, immediately executable and shareable workflows will empower researchers in fields such as biomedicine without requiring them to have specialist data mining skills.
- The developed data preprocessing techniques, including efficient network preprocessing approaches and unsupervised feature selection, will enable heterogeneous information fusion and handling of large data sets.
- The development of new data analytic techniques, including new semantic pattern mining, redescription mining, and heterogeneous network mining algorithms will allow enrichment of existing data sets with new and rich information sources that are so far poorly exploited.
- Application of the proposed methods on challenging biomedical data sets (Parkinson disease) will make progress beyond current state-of-the-art in diagnostics/prognostics.

## 2 Work plan

Our project goals can be achieved by two Postdoc researchers and two PhDs students (one pair from each side) over the project duration of three years supported by the senior researchers and regular staff from both the Austrian and Slovenian research groups. The expertise of the two groups are complementary, with the Austrian side specializing in human-computer interaction of data mining process (human-in-the-loop) while the Slovenian side is focused on developing new data mining algorithms and platforms. Both groups apply their data mining expertise to the biomedical domain. We believe that this intertwining and cross-disciplinary experience of our personnel will ensure the full success of HEDATBIO.

As a guiding use case we will strive to design and implement our experiments according to the following processing pipeline:

Figure 1: In order to support human comprehension of complex, heterogeneous data we 1) merge information from diverse sources into 2) a heterogeneous network enriched with structural information 3) utilizing advanced network analysis, feature subset selection and redescription methods we create comprehensible 4) classification and / or clustering results.

We split our research into four work packages (WP), where WP 1 develops methods for enriched heterogeneous information networks, and WP 2 deals with multi-view data (obtained also from vectorizations in WP 1) by enabling large data sets (feature selection) and human-comprehensive methods (redescription mining). WP 3 provides human friendly visual integration of all developed algorithms. WP 4 validates the work by applying it in the study of Parkinson's disease, comparing it against existing state-of-the-art methods.

## 2.1 WP 1: Enriched Heterogeneous information networks

**Motivation:** Very large databases of documents, semantic triplets, ontologies and relationship networks exist which are currently poorly exploited. Such heterogeneous data are frequently collected from multiple diverse domains or obtained from various sources and exhibit heterogeneous properties, because variables of each data sample can be naturally partitioned into groups. In contrast to standard (homogeneous) information networks, heterogeneous networks describe heterogeneous types of entities and different types of relations. Our approach enriches heterogeneous information networks, where nodes of certain types contain additional information, for example in the form of experimental results or documents. To improve on the state-of-the-art it is necessary to harness additional background knowledge and exploit it in the form of structural links among different features describing the instances as well as semantic relations between features and objects.

**Selected Related Work:** As shown by D. Page and his collaborators (??), relational data mining, which we address through efficient decomposition of heterogeneous information networks, can substantially improve the results of classical machine learning approaches. In network data analysis, instances are linked in a web of connections. ? introduced the concept of authority ranking for heterogeneous networks, where the impact is transferred along edges to simultaneously rank nodes of different types. Network propositionalization (?) decomposes heterogeneous network into a set of informative homogeneous networks which are used to create feature vectors corresponding to nodes in the network. Vectors are obtained by Personal

PageRank (PPR) algorithm which computes local impact of neighborhood for each network node, thereby vectorizing the network structure. The feature vectors are classified to predict the class values of these nodes. Class labels can also be propagated through the network (?). Our extension to vectorization will a) improve its efficiency by taking the sparse nature of PPR into account and b) by weighting the components of the vector with their information content. The latter idea stems from information retrieval, where it is important to correctly set weights of terms in documents (just as not all words in a document are equally informative about its contents, so in a network not all neighboring nodes transfer equal amount of information). Although the term-frequency inverse-document-frequency (tf-idf) weighting is currently the most popular method, recent studies have shown that these weights can be improved with learning (?).

Most data mining algorithms only work with tabular data; however, relational data mining approaches such as inductive logic programming and statistical relational learning approaches (?) can also directly incorporate domain knowledge (background knowledge) without vectorization in order to build richer and more accurate predictive models. However, the knowledge contained in ontologies is rarely used, since it usually cannot be directly employed. With the emergence of Linked Data (?) the data mining community is faced with a new challenge of exploiting this vast resource for knowledge discovery. Semantic data mining is a promising data mining approach to this challenge, however the search space explored by semantic data mining methods has a size that is exponential in terms of the amount of data and annotations being analyzed, and can thus not be exhaustively searched. General methods such as Hedwig (?) use a search heuristic that limits the search space, however, current methods are still unable to handle the amount of data that are now routinely available.

To make it practicable to do effective semantic mining of large networks we will use advanced network ranking and community detection. These are currently highly active research fields, so we list only representatives of relevant research directions. Network ranking is the task of ranking network nodes by either global or local importance (?) and network community detection is the task of discovering communities (sets of nodes with several connections between them and few connections pointing out of the set) in networks (?). Both network community detection and network ranking can be used to answer the question most relevant to us: ‘*For my experimental data, what are the most important parts of the knowledge graph?*’. In our work, we will analyze and compare both classes of methods.

## Objectives:

**O1: Mining enriched heterogeneous information networks.** Within this package we



will extend the three step methodology to mining enriched heterogeneous information networks by  $\mathcal{H}$ , where the input to a data mining algorithm is a set of instance descriptions of a single data type, enriched by contextual knowledge - structural links among different features describing the instances, e.g., documents. In the first step of the methodology, data is preprocessed and transformed into a heterogeneous information network. In the second step the heterogeneous network is decomposed into a set of homogeneous networks, containing only the target nodes of the original network. In the third step, the homogeneous networks are used to predict the labels of target nodes. We propose three extensions to this methodology: a) efficient decomposition of the networks by taking sparse nature of PPR process into account, b) weighting the components of the vectors with their information content, and c) prediction of labels using the label propagation algorithms as well as specialized variants of SVM (support vector machine) classifiers. The most appropriate classification method will be determined by empirical evaluation.

The overall objective is therefore to get a viable methodology for mining large heterogeneous networks where local information content is taken into account and the best classification methods are employed based on the type of problem.

**O2: Mining semantic data.** Network analysis is ideally suited to deal with large amounts of data while semantic data mining is able to garner deep insights from the data. Our objective is to bring these two approaches together, so that semantic data mining will be feasible for large heterogeneous network data. Doing this directly addresses several currently open questions. New semantic data mining algorithms will be developed based on different network analysis approaches, network ranking and network community detection.

Network ranking and community detection are both primarily defined in homogeneous information networks, not taking different node types into account. Based on the approach of (?) and work in this project, we will test several approaches to deal with the heterogeneous nature of networks. The first and simplest way is to ignore the node types altogether and view the network as a homogeneous network. A second, more sophisticated method is to apply different weights to different edge types. The third option is to focus on a single node type and use network propositionalization techniques to construct homogeneous networks that contain nodes of just one type. We will compare these different methods of dealing with heterogeneous networks. In the second step we will use recently developed efficient state-of-the-art network ranking and community detection approaches (???) coupled with semantic data mining (?).

#### Tasks:

**T1: Design enriched heterogeneous network mining methods.** This involves three

different tasks: 1) being able to enrich heterogeneous networks with textual information, 2) taking both local and global network information into account, and 3) testing the different design choices to make the approach efficient on large data sets.

**T2: Design semantic data mining approach.** This involves two tasks: 1) fusing semantic data mining and network analysis to exploit the plethora of LinkedData becoming available, and 2) testing different design choices to make the approach efficient on large data sets.

**T3: Implement and test a visual data mining support for heterogeneous data.** This involves extending Clowdflows, adding visual widgets, feedback and user interaction through a user interface based on graphical representation of algorithmic building blocks. Testing will conduct a series of evaluations to demonstrate the utility of the approaches and of the visual components, using open-access data sets and data sets available from project partners.

#### **Deliverables:**

**D1:** An implemented use case demonstrating the visual data mining of enriched heterogeneous data.

**D2:** An implemented use case demonstrating the semantic data mining combined with Linked-Data.

**D3:** Analyzing the performance and usability of use cases D1 and D2. We plan to present the results to the international scientific community, e.g. at the KDD conference 2018 or similar.

**D4:** Extending D1, D2, and D3 to a practically useful big data real world use case; a consolidated output is planned to be provided to an international SCI indexed journal.

## **2.2 WP 2: Multi-view learning and feature selection**

#### **Motivation:**

Heterogeneous data are frequently collected from multiple diverse domains or obtained from various sources and exhibit heterogeneous properties, because variables of each data sample can be naturally partitioned into groups. Each variable group is referred to as a particular view, and the multiple views for a particular problem can take different forms, e.g. a) clinical measurements, genetic data, and images, b) words in documents, meta-information describing documents (e.g. title, author and journal) and the co-citation network graph for scientific papers. Conventional machine learning algorithms (e.g., support vector machines or random forests) have to merge all multiple views into one single view. This may cause overfitting and is in many cases not even meaningful, because each view has a specific statistical property. In contrast to single-view learning, multi-view learning uses a separate function to model

each particular view and jointly optimizes all the functions simultaneously. While several multi-view approaches exist (?) they are mostly not concerned with the interpretability and comprehensibility of the trained models, even though these aspects are of utmost importance in the medical domain.

Redescription mining is aimed at generating interpretable models for cases when there are two or more distinct attribute sets (multivariate descriptors) describing the same set of samples. Such techniques are highly applicable in biology, economy, pharmacy, ecology and many other fields, where it is important to understand connections between different descriptors and to find regularities that are valid for different element subsets. Redescriptions are represented in a form of rules and the aim is to make these rules understandable, interpretable and relevant across views.

To manage large data sets, distributed computing and specialized feature subset selection approaches are needed which can exploit multi-view heterogeneous data. Distributed computing is pervasively used in our target implementation environment ClowdFlows, and we will design the algorithms to exploit this feature of our system. The aim of feature subset selection is to reduce the computational load of processing high dimensional data by selecting only features relevant for a given task. While unlabeled data is abundant in many application areas, most existing feature selection techniques require single-view labeled data. The development of unlabeled multi-view feature selection methods will be important for efficient redescription mining scenarios.

**Selected Related Work:** ? provides an overview of multi-view learning approaches and lists the multi-views methods in the following learning scenarios: dimensionality reduction, semi-supervised and supervised learning, active learning, ensemble learning, transfer learning, and clustering. Redescription mining is a human-comprehensible aspect of multi-view learning. Originally starting with clustering, it was recently extended to frequent itemset mining ?, relational learning (?) and interactive redescriptions (?). So far, redescription learning lacks comprehensible supervised methods, such as rule learning and subgroup discovery.

In unsupervised learning, feature subset selection aims to express high-dimensional data with low-dimensional representations to reveal significant latent information. It can be used to compress, visualize or re-organize data, and as a preprocessing step for other machine learning tasks. Recently, several approaches to feature selection in this setting were developed, e.g., (??). Our algorithm ReliefF (?) is well-known and is an efficient feature subset selection algorithm for classification and regression problems. It runs in low-order polynomial time, is noise-tolerant and can detect strong feature interactions. It has been extensively adapted to

different scenarios, and is frequently applied in life-sciences, e.g., (??). So far, this approach has not been used in unsupervised or multi-view context.

### Objectives:

**O1: Redescription rule learning.** We will develop new redescription mining algorithms for rule learning. We will adapt the rule selection heuristics to incorporate information on rules already generated in other views. This will provide interpretable models yet exploit knowledge from several views.

**O2: Clustering ensemble for multi-view learning.** Clustering is of great importance in multi-view learning as it can group similar instances in each of the views and thus provide sets of related meaningful features for each of the views. We plan to develop similarity-based clustering using rule ensembles by extending our own approach to cases when examples are described with two or more different attribute sets (multi-view clustering). The aim of the extension is to use concurrent clustering across distinct sets of attributes to increase the stability/reliability of final groupings of examples.

**O3: Efficient feature subset selection for redescription mining.** Extending the ideas of ReliefF, we will develop an efficient similarity-based iterative algorithm which will be able to exploit the multi-view nature of data by taking similarity from other views into account when computing the relevance of features in the selected view. The same principle will be applied in unsupervised setting via virtual labels.

### Tasks:

**T1: Develop classification rule learning for redescription mining.** This involves:

- the development of several heuristic rule-quality estimators that will bias the learned rules in one view to more likely cover instances already covered by rules learned in other views and thereby increase the probability of overlap and consequently the quality of redescriptions,
- conducting a series of evaluations to demonstrate the utility of the algorithm using open-access data sets and data sets available from project partners.

**T2: Develop ensemble clustering for multi-view learning.** This will extend our rule based clustering approach to rule ensembles, by:

- the development of ensemble clustering using rule based clusterings that can run concurrently across distinct sets of attributes.

- conducting a series of evaluations to demonstrate the utility of the algorithm using open-access data sets and data sets available from project partners.

**T3: Develop feature subset selection for redescription mining.** We will generalize the efficient ReliefF filter algorithm and adapt it to the multi-view and unsupervised case. The key idea is sharing of similarity structures between different views in multi-view learning.

**T4: Implementation in Clowdflows.** Implement and test a visual data mining support for redescription mining and multi-view feature subset selection in Clowdflows, adding visual widgets, feedback and user interaction through a UI based on graphical representation of algorithmic building blocks. The testing phase will outline and conduct a series of evaluations to demonstrate the utility of the approaches and the visual components, using open-access data sets and data sets available from project partners.

#### **Deliverables:**

**D1:** A case study on the use of redescription rule learning on neurological data, which will be presented at an international conference (e.g. ECML 2017 or similar).

**D2:** A case study on the use of ensemble clustering for redescription mining, which will be presented at an international conference (e.g. Discovery Science 2017 or similar).

**D3:** Analyzing D1 and D2 in a solid research paper submitted to a SCI indexed international journal.

**D4:** A case study on the use of feature subset selection for multi-view supervised and unsupervised setting coupled with redescription rule mining on neurological data, which will be presented at an international conference, e.g. BIH 2017 or similar.

**D5:** Extending D1 and D4 and combining them in a case study on biological data, to form a solid research paper submitted to a SCI indexed international journal.

## **2.3 WP 3: Workflows with parallel processing and meta learning**

#### **Motivation:**

As data analysis pipelines become more complex as well as ubiquitous, the need for standardization and community data platforms emerge. We therefore propose automated workflow components for ClowdFlows, an open access Web-based data mining platform, allowing experts and users alike to visually compose state-of-the-art processing pipelines. Where applicable, components of possible new workflows will be processed in parallel using ClowdFlows inherent infrastructure capabilities. Furthermore, meta data about each experiment will be stored; this will enable heuristic recommendation engine which will be able to propose optimal

processing paths (algorithmic sequence) if presented with input data plus problem specification.

### **Selected Related Work:**

As (?) perfectly states: *“Using self-contained solutions often results in a glue code system design pattern, in which a massive amount of supporting code is written to get data into and out of general-purpose packages.”* This signifies to us the need for a workflow system easily usable by domain experts without programming. ? defines an ontology of cloud computing that encompasses five layers: the hardware; the software kernel; cloud infrastructure (communication); cloud software environment (the platform); cloud applications. The implementation of our algorithms should fit into the cloud software environment and application layers of this description. Concerning the parallel execution of components in specific data mining workflows, the methodology chosen may be specific to the underlying problem domain, as described in the “asynchronous, dynamic, graph-parallel” model used for Belief Propagation in (?). It will therefore be necessary to corroborate that our use case presented in WP 4 is amenable to such an approach.

Meta-learning applies learning algorithms on data collected about machine learning experiments. In our scenario it concerns selection of workflow components and their parameters based on tackled problems, their features, available learning algorithms, preprocessing methods, parameters and performance measures. Several approaches and successful applications of this principle exist and we will follow best practices (?).

### **Objectives:**

**O1: Meta-learning process.** In order to perform meta-learning a suitable data model has to be developed, which supports collection of data from experiments in the ClowdFlows environment. We aim to build several description formats for data, algorithms and results. We will use them to collect data and perform meta-learning.

**O2: Parallelization.** Multi-view data typically allow parallelization and thereby speed-up of several components. We aim to demonstrate how to apply parallelization inside workflows and how to construct parallel workflows for complex heterogeneous-data multi-view learning tasks. We will use our Parkinson’s disease use case described in WP 4.

### **Tasks:**

**T1: Workflow components with meta-learning.** We will first develop a data format for representing meta-learning in all the developed workflow components, including their input data, problem features, preprocessing options, algorithms and their parameters, output,

dependencies between individual stages etc. We will implement methods that collect data in this format. In order to collect sufficient quantities of data for effective meta-learning, we will need to analyze behavior on several data sets and thus identify structural properties and dependencies in the analyzed problems.

**T2: Parallel processing.** We will parallelize the constructed workflow components and demonstrate their efficiency in the cloud-based ClowdFlows environment. This means that input data has to be packaged into convenient units that correspond to individual tasks, then use the ClowdFlows platform to split them up into separate jobs, collect their results and assemble final results. The whole process will be demonstrated on our use case from WP 4.

#### **Deliverables:**

**D1: ClowdFlows workflow components.** These will support collection of data for meta-learning. Will be presented at an international workshop.

**D2: An implemented use case.** This will demonstrate the feasibility of our parallel model on a specific, chosen problem within our use case domain (Parkinson’s disease).

**D3: A survey paper on meta-learning formats.** This will analyze achieved performance on selected data sets within our use case. An outlook on the possibility of implementing an algorithmic recommender system for similar problems will also be provided.

## **2.4 WP 4: Data use case - Parkinson’s disease**

#### **Motivation:**

In order to demonstrate the feasibility of our methodologies, we need a real-world problem domain providing us with suitable data sets that are diverse enough to show the value of our research, yet specific enough for our results to be clearly interpretable by domain experts. The use case of Parkinson’s disease (PD) comprises several sub-problems (tremor classification and prediction, gait discrimination models, etc.) which are amenable to different machine learning approaches and feature their own, distinctive input data sets. Coming from EEG, EMG, implanted body sensors and force detectors, these data sets have distinct attribute domains, which are - via their time dimension, and probably via many other biological attributes too - interlinkable with one another. As such, they will be very suitable to demonstrate our approach as outlined in WP 1, WP 2 & WP 3 and also fit well with our cross-institutional approach.

#### **Selected Related Work:**

Data mining methods are routinely used for the computational assessment of neurological

disorders including PD (??). Much of the recent work is based on body -fixed sensors (BFS) for long-term monitoring of patients (?). ? study the effects of deep brain stimulation (DBS) on ground reaction force (GRF) during gait and try to discriminate between normal and PD subjects. Multi-view learning is highly competitive in this area as evidenced in (?). This and other recent state-of-the-art methods will be used for comparison with our approaches.

### Objectives:

**O1: Challenging open access data sets.** We aim to select several challenging problems described with rich, multi-view data, heterogeneous data. The data sets selected will allow use of existing linked data, textual data and biological attributes related to PD.

**O2: Evaluation of the proposed approaches.** The selected data set will allow application of our approaches (heterogeneous networks, relational learning, multi-view learning, feature selection, redescription mining) to the Parkinson’s disease problem domain using the ClowdFlows data mining platform. We aim to demonstrate the advantages of our approaches and our computational environment as well as to find their strengths and weaknesses compared to other state of the art approaches.

### Tasks:

**T1: Survey of the data sets.** We will analyze existing open access data sets and sub-problems in the field of PD with a focus on data sets and their properties. We will select a subset of challenging data sets suitable for our approaches.

**T2: Extensive evaluation.** We will use the distinct properties of diverse data sets to arrive at a heterogeneous graphs and multi-view representations suitable to our approaches. We will encapsulate the developed workflows into ClowdFlows built-in evaluation workflows to evaluate them and compare them with existing state-of-the-art approaches concerning their performance and computational characteristics. Biomedical experts (from Medical University Graz) will check that results are meaningful and correct from the biomedical perspective.

### Deliverables:

**D1: A survey paper** describing the state of the art in data sets used for Parkinson’s research to be presented at an international conference and extended into an archival journal contribution, e.g. in ACM Computing Surveys.

**D2: A set of association rules** amenable to our objectives; the results will be presented at a suitable international conference and extended into a SCI indexed international journal.

**D3: A research paper** about fusing diverse multi-view data sets with rich, heterogeneous graphs and advantages our ClowdFlows based data mining environment offers for such tasks,



published e.g. in Springer Knowledge and Information Systems journal.

### 3 Organizational Aspects

#### 3.1 Work Organization, Supervision and Risk Management

The HEDATBIO lead applicant, Andreas Holzinger, will act as project leader and will work directly on this project and allocate a significant amount of time to the HEDATBIO project. He is associate professor for Computer Science and a member of the doctoral school for Computer Science at Graz University of Technology, hence he is in the position to supervise the involved PhD students and to bring in new or additional work-force on demand. He has extensive project management experience and know-how in software development and his diverse scientific background will be a further success factor for this project. Moreover, this project has the full commitment of the Holzinger Group, the applicant's institute and the University, so all technical facilities and organizational support is ensured.

The principal investigator of the Slovenian partner, Nada Lavrač and two senior researchers from her group (Marko Robnik-Šikonja and Dragan Gamberger) will also work directly on this project. Marko Robnik-Šikonja, who is associate professor of computer science and informatics at University of Ljubljana, Faculty of Computer and Information Science will supervise the involved PhD student and allocate a significant amount of time to the HEDATBIO project. The PhDs employed by this project will also be supported by the PostDocs, who will only be hired after the proposal is granted in order to find the best possible international candidates for ensuring the full success of HEDATBIO.

We take care on risks at three levels (scientific risks, management risks, technical risks):

**Scientific risks** lie in the uncertainty of research. We follow the approach that key elements in managing uncertainty are reflective learning and sense-making as well as a good communication strategy and a well-balanced atmosphere to stimulate thinking and problem solving. We take measures to control progress of work, ensure detailed and clear definition of architecture/interfaces and focus on implementing key features in sane iterations. To further reduce risks we have incorporated an international Scientific Advisory Board and will ensure regular communication/feedback.

**Managerial risks** include lack of resources and/or staff changes forced upon the project by

one or more collaborators, therefore the quality of the outcome might decrease as it depends on having access to high-quality resources and staff. We will ensure a good level of communication to talk about any problems that arise and will strive to quickly bring in new scientific staff if necessary. For this we are well prepared as we can allocate additional human resources from permanent staff of both sides, at short notice.

**Technical risks** may include difficulties in data acquisition. This risk is minimal, since we are planning to use open source data sets as well as locally available data from both sides. All equipment is at our disposal and this project has full commitment by both Institutes and both Universities.

### 3.2 Strategies for Dissemination of Results

**1. Science-to-Science:** We will produce internal progress reports. The most promising ones will be extended to peer-reviewed conferences papers, symposia and workshop contributions and the most valuable results will be developed into solid international journal publications. The project team also plans to organize workshops/special sessions at international conferences to disseminate the gained knowledge, and to make algorithms and tools accessible to the international scientific community. The lead applicant has established an HCI-KDD expert network since 2011, which will also serve as an international dissemination platform.

**2. Science-to-Business:** As our algorithmic libraries will be open source and available online, businesses might find it interesting to utilize or embed our software in their products. While we have no co-operation with existing commercial vendors of KDD software at present, good business cases might open up in the future as more and more algorithms become available on the platform. As ClowdFlows is web based and follows the ideas of 'executable paper', it could for example benefit editors of journals vetting submissions, students trying to learn from real world examples, or engineers collaborating on prototyping new algorithms.

**3. Science-to-Public:** We will inform the public about our research and aim to disseminate our knowledge broadly on a regional level, for example in public exhibits, at public Open-lab days ("FWF Lange Nacht der Forschung" and "ARRS Night of researchers"), in regional newspapers and local German and Slovenian speaking workshops, etc. A project web-site HEDATBIO will be available and provide a showcase.

### 3.3 Economic, Social and Practical Impact

entire section copy of hedat-bio

entire section copy of hedat-bio

The main goal of this project is on *enabling basic scientific research via methods of Machine Learning by non-programmers*, but in order to achieve this we will have to devise new tools of practical importance. There is excellent potential for mid- and long-term industrial co-operations and for bringing benefits to the Machine Learning / Data Science community. Moreover, enabling faster research cycles by eliminating the need to manually configure processing pipelines could potentially yield higher scientific output in many areas that utilise computational learning methods. A successful HEDATBIO project could have significant potential to further both national and international developments and enormous impact on supporting international life sciences research (e.g. cancer and neurological disorders research).

### 3.4 Career Benefits for those Involved

The Postdocs and PhD students involved will have an excellent opportunity to achieve knowledge in an extremely interesting and stimulating setting within a realistic time horizon. There are several promising research directions worth being pursued as a PhD within this project. The PostDocs will have an enormous opportunity to demonstrate and further develop their expertise in cutting edge questions of scientific research at the intersection between knowledge discovery and bioinformatics as well as to establish an international track record in the field. The cooperation in an ambitious international team is a further advantage. All HEDATBIO members have the opportunity to achieve expert know-how in an area that is becoming ever more important in the future.

### 3.5 Infrastructure

---

The Institute for Medical Informatics, Statistics and Documentation, Medical University Graz is working on biomedical information systems, with emphasis on making clinical data usable. This includes the development and evaluation of algorithms, software and statistical methods, from data acquisition to data analytics. The Institute offers statistical expertise for biomedical research projects, data management for clinical trials and data extraction and reporting services for medical research. The dedication is to deliver high-quality, methodical contributions and to develop software systems for the support of clinical researchers with a focus on information quality (?). The Institute maintains a Quality Management System, has long-standing software engineering expertise, and is ISO 9001 (certification number: Q-11627/0) certified in both project management and software development. All necessary equipment is available and HEDATBIO has the full support of both the Institute and the University.

entire  
section  
copy of  
hedatbio  
with SLO  
taken out

### 3.6 National and International Cooperation

The lead applicant is employed at the Medical University Graz, located at Graz University Hospital, where excellent local co-operations to relevant clinicians and biological domain experts are established. There are well-established cooperations with Graz University of Technology, where the project applicant teaches the main required lecture Biomedical Informatics at the Faculty of Computer Science and supervises engineering students. On an international level the Group is well connected to the international HCI-KDD network, which the project applicant established. To guarantee that HEDATBIO will be successful, we will collaborate with international partners whose use cases connect to ours.

The Slovenian group is coordinating the FP7 FET project MAESTRA (Learning from Massive, Incompletely Annotated, and Structured Data), which addresses tasks of analyzing complex data, with focus on structured output prediction in the context of massive, networked and incompletely labeled data. Within MAESTRA the group developed many new methods for learning decision trees and ensembles for structured output prediction (multi-target classification, regression and (hierarchical) multi-label classification). The group is a partner in the Human Brain Project (FET Flagship), where it develops a platform, enabling the analysis of large quantities of data routinely collected in hospitals for diagnostic purposes. The group develops methods for rule-based clustering, based on subgroup discovery and predictive clustering, which produce understandable cluster descriptions and coordinates H2020 project PD\_manager, which is a health platform for Parkinson's disease management. The goal of the project is to build and evaluate an innovative, health, patient-centric ecosystem for Parkinsons disease. Moreover, Prof Lavrač coordinates several national projects from the area of life science (in cooperation with National Institute of Biology), the most recent ones being analysis of heterogeneous information networks for knowledge discovery in life sciences, development and applications of new semantic data mining methods in life sciences, and semantic rule discovery in the context of Web services. Prof. Lavrač is also the Principal Investigator of the national research program Knowledge Technologies which is the best Slovenian program in terms of international research activities and collaborations. All this will significantly contribute to the full success of HEDATBIO.

We will involve three international experts as our **scientific advisory board** who will also help in dissemination of our work: Prof. Dr. Ning ZHONG from Japan, and Prof. Dr. Nitesh CHAWALA from the USA, and Prof. Dr. Igor JURISICA from Canada.

### 3.7 Project Team

Note: Short CVs of team members are attached in separate documents

The Austrian group is approaching the problem of knowledge discovery from complex data from the view of domain experts and decision makers. They work on a synergistic combination of methods from two areas, offering ideal conditions to unravel problems with complex data sets: Human-Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with machine learning (human-in-the-loop). This HCI-KDD approach is of importance for solving problems in the health informatics domain generally, and an important step towards personalized medicine.

The Slovenian group specializes in intelligent data analysis of complex data, e.g. prediction in the context of massive, networked, incomplete, structured, multi-view, and/or heterogeneous data. Life sciences present a common field of interest for both groups and require their complementary expertise. For example problems arising in medicine require on one hand, analysis of highly complex data and on the other hand, human comprehensive results.

**The complementary expertise and cross-domain integration will provide an atmosphere to foster different perspectives and opinions; it will offer the opportunity to find truly novel ideas and a fresh look on the methodologies to put these ideas into practice and enables jointly what neither group might do on their own.**

**Assoc.Prof. Dr. Andreas HOLZINGER, PhD** is the project applicant and principal investigator at the Austrian side, head of the Research Unit HCI-KDD at the Medical University Graz, and currently Visiting Professor for machine learning in health informatics at Vienna University of Technology. His interdisciplinary experience in management of several national and EU Projects (e.g. REACTION Remote Accessibility to Diabetes Management and Therapy in Operational Healthcare Network; EMERGE Emergency Monitoring and Prevention) will be very beneficial in HEDATBIO. He will support the PostDoc and supervise the PhD on the Austrian side and foster the exchange of researchers and students between the Slovenian and Austrian side, hence help in supervising students in Slovenia.

**Prof. Nada Lavrač** is the principal investigator of the Slovenian group, head of the Department of Knowledge Technologies in JSI and has extensive experience in management of national and EU projects (e.g. FP7 FET project MAESTRA (Learning from Massive, Incompletely Annotated, and Structured Data), EU FET Flagship Human Brain Project, H2020 project PD\_manager (a health platform for Parkinson's disease management), and nation-

entire section copy of hedat-bio

ally funded Analysis of heterogeneous information networks for knowledge discovery in life sciences, and Development and applications of new semantic data mining methods in life sciences). Her main interests are machine learning and data mining, in particular inductive logic programming and intelligent data analysis in medicine. She closely cooperates with other members of the Slovene group and has recently established a collaboration with Holzinger Group.

**Assoc.Prof. Marko Robnik-Šikonja, PhD.** His main research interests are artificial intelligence and machine learning. Relevant to the HEDATBIO is his work on ensemble learning, feature evaluation, feature subset selection, and rule-based learning. He is (co)author and maintainer of three open-source, R-based software packages for data analytics and data mining, containing state-of-the-art analytical methods. He works with Prof. Lavrč on the problems of rule induction for redescription mining and analysis of heterogeneous networks. They jointly supervise two PhD students working on these topics. He is strongly involved with doctoral study at University of Ljubljana and will advise and supervise one PhD student working on the project.

**Dragan Gamberger, PhD,** is a senior scientist working at Rudjer Boskovic Institute, Zagreb and part time at International Graduate School at JSI, Ljubljana. He works with Prof. Lavrč on the problems of knowledge representation by ontologies, reasoning for decision support, and applications of these techniques in medicine. He is co-author of the public service Data Mining Server which is available at <http://dms.irb.hr/>.

**Bernd Malle, PhD student** Bernd has finished his Master studies in Software Development at Graz University of Technology, supervised by Prof. Dr. A.Holzinger. He is interested in bringing Machine Learning to the browser, and has initiated the project 'Graphinius' connecting a JavaScript graph library to an in-browser code editor & visualization module. The combination of theoretical interests combined with practical ambition makes him an ideal candidate for HEDATBIO.

**2 x N.N. PostDoc, and 1 x N.N. PhD** To ensure a theoretically well-founded progress of our research, international PostDocs with expertise in the relevant fields *will be hired once the project has been granted*. They will extend existing algorithms in rule ensembles for clustering, incorporate network structure into re-description mining and extend mining of heterogeneous information networks, adapt semantic data mining algorithms etc. One of them will take care of SE aspects of the developed methods and act as an entry point for integration of research software into the ClowdFlows environment. The second PhD will work on: a) feature

subset selection for redescription mining and unsupervised learning and b) rule learning in redescription mining for comprehensibility and sharing of information between views.

## 4 Financial Aspects

### 4.1 In-kind contribution of partners

**Austrian side in-kind contribution:** 94,000 EUR (senior staff costs, organization of international workshops), all necessary equipment will be provided; open access costs will be covered; invitations for the international scientific advisory board will be covered;

**Slovenian side in-kind contribution:** 74,000 EUR (senior staff costs, organization of international workshops), all necessary equipment will be provided; open access costs will be covered; invitations for the international scientific advisory board will be covered;

### 4.2 Grant applications of partners

**Austrian side grant application:** in order to successfully carry out the HEDATBIO project one full time PostDoc and one full time PhD over the whole project duration of 36 months is needed; travel costs of 2,000 EUR in the first year, 3,000 EUR in the second year, and 6,000 EUR in the third year, shall be exclusively used for traveling to conferences for the PhD and the PostDoc. This results in 332,810 EUR, to which the obligatory overhead must be added, which results into the total grant application of **349,450 EUR** from the Austrian side.

**Slovenian side grant application:** in order to successfully carry out the HEDATBIO project one full time PostDoc and one full time PhD over the whole project duration of 36 months is needed, which results into a total grant application of **300,000 EUR**.

Entire  
section  
still a  
copy from  
hedatbio

## **5 References**

Note: Due to the page limit, this list contains only limited related work.

