USP-RUG Workshop "Big Data in Science"
Sao Paulo, September 26/27, 2017

1. Introduction

Aim: The workshop should bring together scientists from Brazil and The Netherlands, coordinated by Universidade de São Paulo (USP) and University of Groningen/University Medical Center Groningen (RUG/UMCG), as well as other public and private partners to explore collaborative projects on the theme "Extracting Science from Big Data". The idea is to identify common challenges across the scientific disciplines in the handling, mining, and exploring of Big Data for scientific purposes. The workshop focusses on methodologies for turning raw Big Data into large sets of information and subsequently extracting scientific insights from that. **In other words: From Big Data to Big Information to Big Science**.

Potential follow up funding: FAPESP and NWO consider opening one or more calls in 2018 for collaborative projects between São Paulo and Dutch partners. The exact format for the call is not yet known, but the theme will likely be Healthy Ageing related, and both funding agencies have indicated that they share an interest also in the topic of Big Data. The workshop may provide us with an opportunity to connect in the best possible way to any future call for applications (NWO-FAPESP or other).

Rationale: Big data related research has become pervasive in many branches and specializations of the current research directions active at both USP and RUG. On the one hand, increasingly many such disciplines are able to generate and process big data, due to developments in their fields. On the other hand, successfully setting up and completing research involving big data becomes increasingly hard, due to the sheer size of the data, complexity of the involved methods, and variety of the types of knowledge one needs to possess to manage such an endeavor. As such, there is strong consensus that *big data related research has to be inter- and transdisciplinary*. This consensus is also shared by various other actors in the field, such as stakeholders at the Dutch and Brazilian funding agencies. In their words, "1+1=3": the combination of expertise from multiple domains and locations is required to set up, obtain financing for, and execute successful big data related projects.

The aim of this workshop is to support the creation of synergies aiming at setting up such multidisciplinary and multisite collaboration teams aiming at big data research proposals. Particular sub-aims are:

- getting the teams and researchers at RUG and USP know each other's interests and assets related to big data;
- identifying the concerns, aims, and assets that have a large impact;
- forming multidisciplinary subteams interested in setting up joint proposals, based on the identified commonalities of interests.

Anticipated outcome: This workshop should yield *roughly three* collaborative interdisciplinary research *proposal teams* (including potential industrial collaborators) that should be a perfect fit for the upcoming FAPESP-NWO calls. The elaboration of the proposals is to be done after the workshop, and based on the concrete details of the upcoming call (to be known at a later date).

2. Structuring the discussion

Having only plenary discussions between all the involved participants is very likely not to lead to the desired outcome, as the variety of interests is too large to be handled

effectively in the limited available time. Moreover, this can dilute the focus of the planned proposal teams. As such, structuring the workshop into subgroups is needed.

We propose two structuring axes, as follows.

2.1. Themes (TH)

Every participant belongs to one of the three main themes identified as relevant, at a general level, for the big data research at USP and RUG. The Themes are divided into sub-themes, each of which has a coordinator at USP and RUG ("Sub-theme Coordinators"). These themes are as follows:

TH1: Health sciences (Medical Sciences)

This theme relates to research in medicine and life sciences. Sub-themes of interest are neurosciences, oncology, cardiology, virology, radiology and nuclear medicine, epidemiology, microbiology, and public health.

TH2: Life Sciences

This theme relates to research in pharmacy, molecular biology and biotechnology, including molecular modelling and genomics.

TH3: Science and Technology

This theme relates to research in computing science (informatics), astronomy and data-intensive technology. As an important note, research in this theme is closer to generating technological *solutions* for big data related problems than to generating domain-specific *problems* that require big data related solutions.

Every theme TH1..TH3 will naturally involve domain experts who contribute in generating *problems* (and related big data) in that theme. However, such problems require big-data-related *solutions*. Such solutions are typically created by big data experts, whose work is (usually) not strictly related to a predefined theme. For instance, a computer scientist expert in pattern-finding in large networks can collaborate with problem owners generating such networks in health, live, or engineering sciences.

To foster the match-making between problem owners and solution providers, every theme TH1..TH3 will be allocated a number of big data technical experts from RUG and USP. The workshop coordinators will try to optimally pre-assign the technical experts to the themes based on the questionnaires filled in by the problem owners of the three themes.

2.2. Topics (TO)

While the themes listed above can exist and evolve independently, there are a number of topics which cut across all research (and solutions) that are related to big data. These topics are of a technological nature: They relate to the challenges anyone that needs to use big data faces; and also to the technical big data solutions which, once constructed, can be deployed to support any application field. We have identified four main technological topics related to big data, as follows.

TO1: Data generation and federation

Collaborative projects between USP and RUG will be geographically distributed enterprises by definition. This holds for data generation: experiments generating large data are likely geographically distributed across hospitals and research centers. It holds

for data curation and operations: data storage and processing relies on distributed hardware. It holds for scientific exploitation: the research teams are geographically distributed. This leads to the concept of data federations per discipline. This subtheme aims to identify the commonality of outstanding challenges (and solutions) for data federations in medical and the physical sciences. It starts from an inventory of current approaches at USP and RUG.

TO2: Data integration

In science, current breakthroughs typically come from a joint analysis of Big Data sets coming from diverse instruments / experiments on a common population of persons / objects of study. Advanced data warehousing offers methods to cope with the challenging complexity. This subtheme aims to identify the commonality of outstanding challenges for such complex data warehousing in the medical and the physical sciences; and it looks for technological solutions for such challenges which exist among the participants. It starts from an inventory of current approaches at USP and RUG.

TO3: Data exploration

With access to diverse and complex Big Data sets in place the challenge becomes to mine this data effectively and efficiently. Artificial neural networks and other machine learning approaches can assist the human researcher in (i) finding the critical pieces of information in the ocean of data and (ii) determining subtle statistical aspects of the data ocean itself. A separate challenge is to allow visual exploration by humans to derive scientific insights from large sets of information. This involves smart and human-guided mapping of high-dimensional spaces to lower dimensions; cluster analysis and outlier detection; jointly depicting data attributes having many different types and semantics; and depicting time-dependent (dynamic) data. This subtheme aims to cross fertilize the medical and physical sciences by looking for opportunities for joint research on data-exploration techniques, based on technological solutions developed (or to be developed) by the science and technology theme.

TO4: Data application

The focus of this topic is on methods, methodologies, and techniques for using big data to design and execute experimental strategies.

Topics will serve as a 'common language' that will bind the problem owners with solution providers by offering them a framework in which they can express their big-data-related needs, respectively their big-data-related expertise.

3. Workshop outline

We propose a two-day workshop in São Paulo to be organized 26-27 September 2017 with at least 60-80 participants (at least 30 or so from Groningen). To streamline discussions for an effective outcome, participants will enroll in a theme (based on their interests) and also express their interest in one or more topics. Details on the procedure are given below. The sub-theme coordinators will provide input to draft the list of invitees, after which the invitees may confirm their participation via their enrolment in a theme.

3.1. Theme enrolment

We propose a *theme-based* structure of the workshop, along the three themes TH1..TH3 mentioned in section 2.1. This way, researchers who share interests in the same theme can discuss closest with their peers.

To gather information on the affinities of participants to themes and topics, all participants who belong, in terms of problem ownership, to a theme, are asked, before the workshop, to complete a questionnaire in which they have to specify

- their main theme of activity (TH1..TH3). Only one such theme is permitted. This input is mandatory;
- their interests in the technological topics (TO1..TO4). Several topics are allowed. This is an optional input, which, if present, will help the match-making of problem owners with big-data solution providers.

3.2. Schedule generation

Based on the response of the participants, the workshop coordinators will associate big-data experts from USP/RUG to the themes. We aim at having a good preliminary match of theme-specific problems and big-data-specific solutions, but also a good balance of different kinds of big data expertise across all three themes.

Based on the final compositions of the themes (domain experts and big-data experts), the three theme coordinators will draft a preliminary program for their respective themes. This can involve talks from the participants, round tables, or (optimally) a mix of the two. Details are given below.

3.3. Workshop preliminary schedule

Day 0
Arrival at USP.

Day 1
9:00-9:30 - *Opening* (Carsten Wrenger (USP), Alex Telea (RUG))
9:30-10:00 - *Introduction into aims, scope, and set-up of the workshop*

The workshop starts with a plenary session in which we discuss the overall aims, scope, and anticipated outcome of the workshop. The theme coordinators are also introduced, as well as the per-theme participants are listed. The way of working (this document) and its rationale is explained.

10:00-10:30 – *Introduction of the themes*

We propose three plenary keynote lectures (~10 minutes) on each of the themes. The lectures are based on the collected input from theme participants. In particular, the challenges (and/or solutions) related to the four technological topics that are important for the theme should be outlined. The aim of these lectures is to give everybody a good insight in what 'lives' within each theme.

10:30-10:45 – *Coffee break*

11:00-13:00 - *Theme sessions*

We then split up in 3 per-theme groups. Within each group, presentations are held by the individual participants, to favor knowing one another and one's research interests. The presentations should focus, among others, on the big-data-related challenges of the current research of the participants; and also on highlighting existing RUG/USP collaborations which can be strengthened by additional participants. The exact format of the theme sessions is to be decided by the theme coordinators, based also on the final number of participants. For instance, a presentation in Health (TH1) can introduce some very specific problem in medicine, which in turn raises several technological challenges related to, say, storage and analysis of multidimensional hybrid big data. The conclusion

of this presentation is that joint work between the medical problem owner and machine learning and/or database experts is needed.

13:00-14:00 *Lunch*

14:00-17:00 - *Theme sessions (continued)*

We continue the per-theme work initiated in the morning.

17:00-19:00 *Social event*

20:00 *Dinner*

Day 2

9:00 – 11:00 *Presentations of Big Data applications by Big data experts in relation to the themes*

In this session, we switch the focus for problem domains to generic big data solutions. The floor is given to 5..6 presentations from the big data experts participating to the workshop to describe their successes and specialized research in big data related work along the four technological topics TO1..TO4. The speakers are pre-selected by the workshop general organizers. The aim of this session is to expose all participants to *solutions* and state-of-the-art related to big data, so as to further foster collaborations.

11:00-11:15 – *Coffee break*

11:15-13:00 – *Theme based round table discussion between "solution seekers" and Big Data experts*

In this session, discussions continue towards the match-making of solution seekers with big data experts, based on the insights accumulated so far.

13:00-14:00 *Lunch*

14:00 – 16:00: *Plenary summaries and discussions*

In this session, all three themes present the summary of their findings, along the lines presented above in this document, in plenary. In particular, interesting matches discovered during the workshop, including details over the involved themes and technological topics, are presented plenarily. Time is split equally for all the three involved themes. The individual participants in each theme have now the opportunity of making their concluding remarks and outline future collaboration thoughts.

16:00-16:45 *Closing remarks* (USP/RUG)

19:00 *Dinner*

Day 3

Individual meetings between members of the project teams
This day is reserved for elaboration of per-project-team specific activities, based on the already formed project teams.

Departure (end of day)