Paris-Saclay Center for Data Science

Compte rendu du COPIL, 26/03/2014

Membres présents ou représentés

Gilles Faÿ	MAS/Centrale
Balázs Kégl	LAL/UPSud
Michèle Sebag	LRI/UPSud
Arnak Dalalyan	Stats/ENSAE/CREST
Patricio Leboeuf	FCS
Erwan Le Pennec	CMAP/Polytechnique
Emmanuel Vazquez	E3S/Supelec
Florence d'Alche Buc	IBISC/Uevry
Cécile Germain-Renaud	LRI/UPSud
Francois Yvon	LIMSI/UPSud
Laurent Barthès	LATMOS/UVSQ
Cedric Gouy-Pailler	LIST/CEA

Sur les 16 membres du comité de pilotage, 12 membres sont présents ou représentés; le quorum est en conséquence atteint.

1 Informations générales

Le projet Paris-Saclay Center for Data Science (CDS@PS) a été retenu par l'IDEX Paris Saclay et va bénéficier d'un financement de 1.200K€ pour deux ans (01.03.2014-29.02.2016). Un renouvellement d'un an sera possible si l'IDEX est reconduit. L'ambition du projet est de former une vraie communauté scientifique autour de la science des données (data science) dans le périmètre de l'Université Paris Saclay.

Le financement accordé par l'IDEX est une aide au démarrage qui doit servir, en particulier, pour initier des projets de recherche pouvant être subventionnés par d'autres sources (nationales, européennes, etc.). La participation active de tous les membres du projet est fortement encouragée pour faire avancer le projet.

2 Data challenges

Une des actions initiallement prévues dans le projet de CDS@PS est l'aide à l'organisation des challenges sur des données scientifiques. Dans ce cadre, deux projets de challenges ont été présentés.

1. Le challenge sur le boson de Higgs et est piloté par Balázs Kégl (LAL). Le challenge sera lancé en mai-août 2014 sur Kaggle (https://www.kaggle.com) et est co-financé par Google. Il est prévu d'organiser un workshop à la conférence NIPS'14 pour présenter les résultats du challenge. Il y aura également une présentation faite à la journée thématique sur la physique des particules. **Le COPIL a validé le financement du challenge sur le boson de Higgs pour un montant de 20K€ (14K€ pour les frais de Kaggle et pour le prix et 6K€ pour les missions reliées à l'organisation du workshop NIPS'14.)**

2. Le deuxième challenge est porté par Claire Nédellec (MIG/INRA) et concerne des données bio-informatiques. L'aide demandés servira à effectuer l'annotation des données utilisées pour le challenge. Les résultats du challenge seront présentés à la journée thématique sur le NLP (Natural Language Processing) ou Bio-informatique. **Le COPIL a validé le financement du challenge bioinformatique pour un montant de 18K€ pour un CDD ingénieur de six mois.**

3 Organisation d'école d'été

IN2P3 organise une école bi-annuelle en statistiques pour les physiciens (School of Statistics). L'école aura lieu entre du 26 au 30 mai 2014 (https://indico.in2p3.fr/conferenceOtherViews.py?view=standard&confId=9742). Le COPIL a validé l'attribution d'un montant de 4.8K€ pour participer au financement de cette école.

Balázs Kégl informe que le LAL organisera une jornée de formation sur le "cloud computing" d'ici printemps 2015. Ce cours sera ouvert à tous les membres du CDS@PS et ce dernier en fera la publicité.

Pour renforcer l'action du CDS@PS, il serait souhaitable que les membres se portent bénévoles pour faire des cours doctoraux sur les divers aspect de la data science. L'information sur ces cours sera diffusée par le biais du CDS@PS.

Il serait également souhaitable d'organiser une école d'été sur nos thèmes en 2015. Le CDS peut participer à hauteur de 20K€ au financement d'une telle école. Les propositions doivent être faites avant la fin du mois de juin 2014.

4 Réunions, séminaires, brainstormings, reading club

Le succès de l'initiative lancée par le CDS@PS dépend fortement de implication des membres du projet dans les différentes manifestations organisées, que ce soit en tant qu'organisateur ou en tant que participant. Pour les projets soumis en réponse à un appel à projet du CDS, l'implication des auteurs du projet dans l'animation sera un des critères d'évaluation.

Trois types de manifestations sont envisagées:

Brainstormings des demi-journées organisées à la demande dont l'objectif est de se connaître, initier une discussion, tisser des liens et monter des projets (ANR, européen, etc.). Deux demi-journées de ce type sont déjà programmées:

- chimie analytique à la faculté de pharmacie, 01/04, inscription à: https://indico.lal.in2p3.fr/conferenceDisplay.py?confId=2414
- économie et gestion (RITM) au LRI, 28/04, inscription à: https://indico.lal.in2p3.fr/conferenceDisplay.py?confId=2413

L'inscription sur le site indico est très fortement recommandée pour pouvoir organiser le déjeuner-buffet, et pour avoir une trace.

Journées thématiques Il est envisagé d'organiser entre 8 et 10 journées sur la période de 2 ans, avec 5-6 exposés longs ou 10-12 exposés courts. Il est prévu de financer 1-2 invités externes, ainsi que les déjeuners et les pauses cafés. Ces journées peuvent être accompagnées par des sessions de posters présentant les travaux des étudiants/postdocs. Chaque journée aura un responsable et une équipe d'organisation. Le support technique peut être centralisé au LAL.

Le COPIL a décidé d'accorder 30K€ pour l'organisation de ces journées (20K€ pour les frais de mission des conférenciers invitées et 10K€ pour d'autres frais (déjeuner buffet, pause café, communication).

Un calendrier provisoire pour ces journées avec des responsables devrait être proposé avant le 30 juin 2014.

Retraite Afin de favoriser les échanges entre les différentes équipes, il est souhaitable d'organiser une retraite de 3 à 5 jours au printemps/été 2015. Le coût approximatif d'un tel événement serait de 20K€.

La page web du CDS va centraliser les annonces de ces différents événements ainsi que des séminaires et groupes de travail organisés par les équipes membres. Il est également envisagé de mettre en place un groupe de lecture pour les thésards.

Pour gagner en visibilité, le CDS peut également sponsoriser des conférences et workshops externes. Dans ce cadre, la demande de participation financière au symposium sur la visualisation et Data Science lors de la conférence VIS 2014
http://ieeevis.org/year/2014/info/exhibition/supporters-and-exhibition
a été faite par Jean-Daniel Fekete. Le COPIL a attribué 3K€ à cette demande, et garde la décision ouverte sur une subvention de 6K€ conditionnée sur une présentation de Jean-Daniel Fekete au prochain COPIL.

Suite à la discussion engendrée par cette demande, Emmanuel Vazquez a proposé de piloter et élaborer une politique de communication générale pour guider nos futures décisions sur le sponsoring des conférences et workshops externes.

5 Site WEB

Le site provisoire du CDS se trouve à http://cds.lal.in2p3.fr/ et est édité sous Wordpress. Pour le moment, seuls B. Kégl et A. Dalalyan peuvent modifier le site, mais toute proposition d'aide serait la bienvenue. Il est envisagé que l'Université Paris-Saclay héberge notre site, auquel cas la migrations peut être faite relativement facilement.

6 Coding sprints et Open Software Initiative

Dans le cadre des *Coding sprints* il y a eu deux demandes.

Bertrand Thirion demande le financement d'un coding sprint pour scikit-learn en 2015 à hauteur de 5k€. Il nous fournira plus de détails sur cette demande ultérieurement.

Alexandre Gramfort (Telecom Paris) demande le financement de coding sprint pour les logiciels MNE (http://martinos.org/mne/) et NiLearn (http://nilearn.github.io/) à hauteur de 5k€ que le COPIL lui a attribué.

Alexandre Gramfort rédigera un document sur le "Open Software Initiative" (date limite pour le texte 04/04/2014), dont le but est d'inciter les étudiants (Master et Doctorat, voir post-doc) à participer au développement des logiciels open source. Le COPIL se prononcera sur ce projet par le biais d'un vote par mail.

7 Gestion

Plusieures options sont envisagées. La demandes d'un CDD de 18 mois coûterait environ 50-60K€. Balázs Kégl nous informe des discussions en cours avec la direction du LAL pour étudier la possibilité de mise à disposition du CDS de certains services administratifs du LAL.

8 Appel d'offre

Il y aura un appel d'offre pour les actions suivantes prévues dans le projet initial du CDS@PS:

- Des postdocs pour une durée d'un an maximum sur un sujet interdisciplinaire. Financement prévu: 50K€ par projet. Le COPIL va évaluer la qualité scientifique des projets et leur adéquation avec la politique du CDS@PS. Il s'agit d'évaluer le projet, pas les candidats. Une fois les projets sélectionnés, la date limite d'embauche est le 01/01/2015.
- Des "code consolidators" pour une durée allant de 3 à 6 mois. Cela s'adresse typiquement aux étudiants juste après leur thèse afin qu'ils puissent "professionnaliser" les outils informatiques développés pendant la thèse. Cela devrait se concrétiser par la publication d'un logiciel et/ou un package sous une licence open source et/ou l'intégration dans un toolbox. La somme prévue est de 12.5 - 25k€par projet (à peu près 4150€ par mois).
- Des semestres sabbatiques intra-Saclay d'une durée allant de 6 mois à un an. L'idée principale est d'encourager les "data scientists" (statisticien, machine learner, optimiseurs, etc.) de passer un semestre dans un laboratoire de "domaine science" (physique, biologie, chimie, économie, etc.) qui possède des données, ou inversement, un "domaine scientist" de passer du temps dans un laboratoire de "data science" pour développer des techniques d'analyse. Le budget du CDS@PS pourrait alors être utilisé, par exemple, pour le rachat des heures d'enseignements. Chaque candidat intéressé par cette action est invité à se renseigner auprès de son établissement pour les aspects administratifs et comptables d'une telle action.

Le calendrier proposé est le suivant. L'appel d'offre sera approuvé au prochain COPIL et publié début mai 2014 avec une date limite de soumission fixée au 31/05/2014. Au cours du mois de juin le COPIL examinera les projets reçus et rendra sa décision au plus tard au début du mois de juillet 2014. Le texte de l'appel sera rédigé par Balázs Kégl, Arnak Dalalyan et Alexandre Gramfort. L'aide de toute autre personne intéressée est la bienvenue.

Concernant les financements des thèses de doctorat, il est décidé de repousser l'appel d'offre à l'automne 2014 afin de laisser le temps aux membres du CDS d'initier des collaborations et laisser mûrir les projets qui en résultent. Il s'agira alors de faire un appel à projet sans candidat, pour des thèse qui pourraient débuter à partir de janvier ou septembre 2015. Tous les membres du CDS ayant des projets de thèse pour la rentrée 2014-2015 sont invités à répondre à l'appel d'offre IDI (Initiative Doctorale Interdisciplinaire) lancé par l'IDEX. Le CDS@PS pourra fournir des lettres de soutien au projets qui portent sur la Data Science.

9 WG open data/data centers

Cécile Germain a rapidement présenté le role que le CDS pourrait jouer dans le recensement des besoins et l'organisation de l'infrastructure "data center". Pour le bon fonctionnement de ce WG, il serait souhaitable d'identifier un interlocuteur dans chaque groupe/établissement concerné.