COMPÉTITIONS D'ANALYSE DES DONNÉES À L'UNIVERSITÉ GRENOBLE ALPES : MOTIVATIONS, ORGANISATION ET RETOURS D'EXPÉRIENCE

Jean-Baptiste Durand
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP , LJK, 38000 Grenoble, France
Jean-Baptiste.Durand@univ-grenoble-alpes.fr

Résumé. Dans cet exposé, nous proposons un retour d'expérience sur divers modules d'enseignement à l'Université Grenoble Alpes mettant en œuvre des compétitions d'analyse de données, entre 2017 et 2019. Ces enseignements se voulaient transversaux à plusieurs formations ou disciplines, et visaient à renforcer ou compléter des méthodes enseignées via des approches traditionnelles, tout en favorisant le travail en groupe et la communication. Nous présentons l'organisation de ces compétitions, y compris les infrastructures et projets institutionnels sur lesquelles elles s'appuient, ainsi que les difficultés rencontrées; s'ensuivent un bilan et quelques perspectives.

Mots-clés. Data challenges, compétitions, analyse des données, science des données, apprentissage statistique.

Abstract. In this presentation, some feedback is provided on various teaching units taking place at Université Grenoble Alpes that involve competitions on data analysis, between 2017 and 2019. These teaching units claimed to be transversal to serveral fields and training programmes. They aimed at renforcing or completing methods taught by classical approaches, while promoting teamworking and communication. The organization of these competititions is presented, including some infrastructures and institutional projects on which they relied, as well as encountered difficulties. This is followed by a short, global assessment and by perspectives.

Keywords. Data challenges, data competitions, data analysis, data science, machine learning.

1 Motivations et contexte

Depuis ces dernières années, la ComUE UGA (Université Grenoble Alpes) fait face à une demande croissante de formations faisant intervenir la science des données à différents niveaux (IUT, master, doctorat et recherche), et dans des contextes disciplinaires variés (statistique, mathématiques appliquées, informatique, traitement du signal, physique, biologie ou autre). L'accroissement constant au cours des ans des performances des méthodes d'apprentissage automatique et de fouille de données, combiné à leur accessibilité favorisée par la disponibilité de bibliothèques dans divers langages (R et python notamment) se sont traduits par une forte augmentation de la demande de scientifiques des données dans les établissements publics et dans l'industrie (Agarwal *et al.*, 2014 ; Davenport et Patil, 2012). Suivant les disciplines, les compétences à acquérir vont de l'utilisation plus ou moins directe d'outils, jusqu'au développement et à l'étude de nouvelles méthodes et, parfois même, à la mise en place d'infrastructures spécifiques de calcul haute performance. La formation d'un nombre adéquat d'étudiants avec des compétences adaptées à la fois en informatique, statistique, avec des expériences de mise en œuvre dans des problèmes réels traitant des données de complexité significative, est devenu un enjeu fort de l'université (Donoho, 2017).

Les difficultés pédagogiques, au regard de ces enjeux, sont principalement liées à l'hétérogénéité des centres d'intérêt, des compétences des étudiants, et des carrières visées (par exemple,

1 Institute of Engineering Univ. Grenoble Alpes

recherche ou industrie). En particulier, il existe pour une partie d'entre eux une relative défiance vis-à-vis soit des mathématiques, de la statistique ou de l'informatique.

Répondre à une telle demande de formation est un défi pour l'équipe pédagogique de science des données à l'UGA, qui a souhaité proposer des solutions permettant, d'une part, de mieux impliquer les étudiants dans leur formation et d'autre part, de mutualiser les moyens humains et matériels mobilisés. Notre réponse vise à aider ces étudiants à compléter leur formation à la fois par la théorie et la pratique. Elle est basée sur quatre axes, qui font l'objet d'un projet « Transformations pédagogiques et Plateformes *Learning-by-doing* » et d'un projet Cross-Disciplinary Program (CDP) « DATA@UGA (Sciences des données) », financés par l'Idex « Université Grenoble Alpes : Université de l'innovation ». L'axe présenté ici est celui des compétitions d'analyse de données. Cette approche nous a paru à même d'impliquer plus fortement les étudiants dans leurs études et de les plonger dans un contexte proche de celui de l'entreprise ou de la recherche. Un objectif supplémentaire était de les confronter à des travaux multidisciplinaires, ce qui apparaît comme une composante importante de la science des données (Donoho, 2017).

C'est l'une des raisons pour laquelle les compétitions de données ont récemment connu un fort développement, qui s'est accompagné de la création de plateformes adaptées pour les héberger (kaggle.com, dextra.sg, drivendata.org entre autres). Pour les entreprises ou les instituts publics, ils sont une manière de sous-traiter un problème dont la résolution ne peut être menée en interne (l'exemple le plus connu étant la compétition Netflix de 2009 pour la prédiction de scores d'appétence de consommateurs pour la location de DVD). Pour les compétiteurs, la participation est un moyen de mettre en valeur leurs compétences auprès de recruteurs, une activité ludique… ou une opportunité de faire leurs premières armes en analyse de données. Ce dernier aspect a conduit les universités à créer leurs propres sites de challenges (challengedata.ens.fr), ou les sites existants à proposer une version éducative (Kaggle in Class, inclass.kaggle.com, utilisé notamment par Molla, 2013).

Le principe d'utiliser ces pratiques comme outils de formation a fait l'objet de quelques publications (D'Aquin *et al.*, 2014 ; Drachsler, 2014 ; Durand, 2017), assez peu au regard de l'expansion des compétitions à vocation pédagogique sur Kaggle. Les retours d'expérience sont restés rares jusqu'en 2017, à l'exception notable de Molla (2013), puis ont commencé à être publiés (voir Besse, 2017, puis le Dossier spécial « Concours et challenges en statistique » de Statistique et Enseignement **8**(2), 2017). Dans ce qui suit, nous présentons l'organisation de quelques unes de ces compétitions et des infrastructures utilisées.

2 Compétitions mises en œuvre

Nous présentons, dans un premier temps, les infrastructures et les pratiques organisationnelles ayant permis la réalisation des compétitions. Dans un deuxième temps, nous évoquons le contenu de quelques compétitions organisées, en détaillant celles qui présentent un caractère interdisciplinaire. Pour ces dernières, un bilan est proposé.

Organisation et infrastructure. Dans le cadre du projet Idex Data Challenges, nous avons obtenu un financement pour l'aménagement d'une salle multimodale et pour le recrutement d'un ingénieur chargé de développer une plateforme de compétitions et plus généralement, de venir en appui de leur mise en œuvre. Ce projet est le volet « formation » d'un projet plus vaste, épaulé par le Grenoble Alpes Data Institute[2] qui comprend également un volet « recherche ». Dans ce cadre sont régulièrement organisées des compétitions à destination des chercheurs, qui bénéficient des mêmes infrastructures et de l'appui de l'ingénieur associé au projet. Rappelons brièvement le principe usuel de ces compétitions : on dispose de deux jeux de données, l'un d'entraînement, l'autre de test. Conceptuellement chacun de ces jeux est divisé entre deux catégories de données : des prédicteurs x et des quantités à prédire y, le but étant de construire une fonction $f(x)$ qui approxime au mieux y au sens d'une métrique m fixée par les organisateurs. Les valeurs de y ne sont connues des participants que pour le jeu d'entraînement et utilisées pour construire f ; celles de x sont disponibles dans les deux jeux de données. Les participants sont classés sur la base de la valeur de $m(y,f(x))$ sur le jeu de

2 https://data-institute.univ-grenoble-alpes.fr/

test. Il existe des variantes où y n'est jamais requis : il s'agit alors d'optimiser une certaine fonction $m(x)$.

Initialement le cahier des charges de la plateforme comprenait des fonctionnalités classiques sur ce type de plateforme : l'inscription des participants en équipes, le dépôt du règlement et du calendrier, la soumission de solutions sous forme de tableaux de donnés par les compétiteurs et leur évaluation. Il comprenait aussi les fonctionnalités suivantes, qui ne sont pas proposées sur la plupart des plateformes :

a) création de compétitions où les fichiers soumis par les participants ne sont pas des tableaux numériques mais des fonctions, ou plus généralement du code (représentant par exemple des lois de probabilité)

b) accès à des outils de développement collaboratif pour les participants

c) exécution du code sur la plateforme, non seulement pour établir le classement mais aussi en phase de développement

d) octroi de droits différenciés aux équipes suivant leur provenance – internes à notre université (accès aux ressources de calcul intégrées, voir c) ou externes (utilisation de ressources propres pour l'exécution des tests)

e) forum pour les organisateurs permettant d'échanger autour des compétitions en cours et futures, avec archivage des compétitions passées avec divers niveaux de visibilité (thème, données, solutions).

Étant donné que ce projet était trop ambitieux, vu nos ressources, et que ces fonctionnalités sont partiellement offertes par diverses plateformes existantes, nous nous sommes orientés vers la réutilisation et la mutualisation de ces dernières : Codalab[3] pour les point a) et c) – évaluation des solutions ; la plateforme gitlab de l'Université Grenoble Alpes pour le point b) ; Slack[4] pour le point e).

Codalab est une plateforme co-développée par Stanford University, l'Université de Paris-Saclay et l'Université Autonome de Barcelone. L'une de ses composantes est dédiée aux compétitions. Elle fonctionne sur le principe de dépôt d'une archive qui comprend des données d'entraînement et de test, une image docker utilisée pour exécuter les solutions soumises, soit sur la plateforme, soit sur un serveur mis à disposition par les organisateurs d'une compétition donnée, et des scripts python définissant essentiellement la métrique m et le format attendu pour la fonction f. Ce format comprend les sorties y et les entrées x au sens large, y compris des contraintes éventuelles sur les données d'entraînement (par exemple pour des séries temporelles, on n'a pas le droit d'utiliser les valeurs futures pour prédire la valeur courante de y).

La salle multimodale a été aménagée en 4 îlots de tables partiellement mobiles et détachables organisées autour de colonnes avec des prises électriques et réseaux comprenant 4 PC fixes et des emplacements pour portables. Il y a un grand écran par îlot, partageable avec les 4 PC fixes et les éventuels portables, plus un grand écran commun, partageable par n'importe quel portable ou PC fixe parmi les 16 plus l'ordinateur de l'enseignant. Ces écrans sont destinés à faciliter la communication dans les équipes, ou la diffusion d'information et d'exemples par les enseignants.

Exemples de compétitions. À notre connaissance, dans le cadre du volet formation du projet Idex, en deux ans 11 compétitions ont été organisées, pour les publics suivants : IUT STID de 1re et 2e année, licence professionnelle, M1 et M2 en science des données, informatique, mathématiques appliquées et traitement du signal.

Nous présentons de manière plus détaillée les compétitions à destination des étudiants issus de trois M2 internationaux (informatique, mathématiques appliquées et traitement du signal) dans la mesure où elles sont emblématiques des différents types de problèmes à surmonter :

- Interdisciplinarité. Les problèmes choisis requièrent des compétences en informatique, mathématiques appliquées et traitement du signal et aucun étudiant n'est spécialiste des trois à la fois. Nous avons donc imposé des équipes devant mélanger les trois profils.

3 https://competitions.Codalab.org/

4 https://slack.com/

- Caractère inter-universitaire. Les étudiants proviennent d'universités différentes et de composantes différentes d'une même université. Ceci complique l'accès aux ressources informatiques, salles, emplois du temps, les inscriptions, la diffusion des informations et des calendriers.
- Données non-standard. Les données à analyser étaient des séries chronologiques, possiblement multivariées, avec des métriques complexes.

Les compétitions comportent deux phases. La première a lieu d'octobre à janvier et comprend une présentation de la problématique scientifique, des jeux de données, règlement et métriques, puis le développement, sur le temps libre des étudiants, d'une solution simple de référence qui donne lieu à une présentation orale et un mini-rapport. Les équipes peuvent évaluer et exprimer leurs besoins spécifiques en logiciel et matériel pour la 2e phase. Celle-ci se déroule à temps plein sur trois jours. Lors de la dernière journée, un nouveau jeu de test est publié, les équipes sont classées en fin de matinée et font une présentation orale devant les autres équipes dans l'après-midi. La note d'équipe comprend trois parties : classement et rapport finaux et présentation orale.

Compétition de 2017-2018. Le thème était l'identification de locuteurs dans des vidéos. Le but est d'une part d'attribuer une étiquette aux personnes présentes dans la vidéo afin de permettre leur identification d'une image à l'autre. D'autre part, pour chacune des personnes identifiées, il faut détecter à chaque instant si elle est en train de parler ou non. Une métrique permet de quantifier différents types d'erreurs : faux positifs et faux négatifs dans la détection des personnes, erreurs d'étiquetage et de détection de la parole.

Compétition de 2018-2019. Il s'agissait de prédire la concentration de 4 types de polluants avec un pas de temps de 10 min sur les 540 pas de temps suivants, pour chaque jour, à l'aide de l'historique des mesures effectuées sur plusieurs stations localisées géographiquement et des prévisions d'un modèle d'émission et de diffusion des polluants accompagnées de données météorologiques localisées. La métrique donnait un poids variable aux différents polluants et aux divers horizons de prédiction ; elle était basée sur une hypothèse gaussienne multivariée et nécessitait donc la prédiction d'espérances et matrices de covariance d'assez grande taille, variables temporellement.

Pour la 1re compétition, aucune plateforme n'était disponible. Les inscriptions ont été faites par mail ; le respect des contraintes de mixité a occasion l'échange de nombreux messages. Les informations sur le règlement et le calendrier ont été diffusées via un site web ad-hoc. Les sites web se sont multipliés pour diffuser des jeux de données et des exemples de codes. Dans la phase finale les données de tests ont été diffusées sur un site web et contenaient par erreur les vraies valeurs de y à prédire. Les solutions ont été évaluées par les organisateurs en lançant manuellement des scripts sur leurs propres PC. Les présentations orales et les rapports des étudiants ont montré que malgré des résultats parfois bons voire très bons, de grandes difficultés subsistaient pour valider les modèles et leurs hypothèses ainsi que pour justifier les méthodes utilisées.

Pour la 2e compétition, l'usage de la plateforme Codalab a facilité les inscriptions, la diffusion centralisée des informations sur le règlement et le calendrier, des données d'entraînement et de test, l'évaluation des solutions et la publication des rangs en temps réel. Les consignes ont insisté fortement sur la nécessité de justifier et valider soigneusement les approches, mais l'attrait du classement est resté le plus fort pour les étudiants et l'aspect technique des solutions a primé sur la qualité des restitutions. Les organisateurs avaient sous-estimé le temps d'évaluation des solutions définitives du fait d'une part de leur temps d'exécution, les scripts pouvant ne jamais se terminer d'eux-mêmes ou bien produisant des erreurs après des temps très longs et d'autre part, du caractère séquentiel de leur exécution (par opposition à une parallélisation des évaluations, non prévue par la plateforme).

Dans les deux cas, les étudiants en traitement du signal ont eu le ressenti – et se sont parfois plaints – que la compétition était avant tout une affaire de programmation. Les étudiants en informatique ont assumé la programmation sans toujours bien assumer le contenu épistémologique. Les autres étudiants ont manifestement essayé d'établir une dynamique d'équipe et de maintenir une cohérence dans le projet, avec parfois des tensions à résoudre. Ceci témoigne de la réalité du caractère interdisciplinaire du projet, mais aussi des difficultés qui en découlent et de ce que les organisateurs les ont peu anticipées et en ont minimisé l'ampleur.

3 Bilan et perspectives

Les objectifs initiaux du projet Idex « Data Challenges » comprenaient l'organisation de compétitions, la promotion de ces pratiques pédagogiques dans les formations en statistique et au-delà, le développement d'une plateforme et l'aménagement d'une salle multimodale. A posteriori ces objectifs paraissent un peu trop vastes pour un projet de deux ans finançant une personne à 30 % sur le projet, mais l'essentiel a été réalisé : plateforme, salle et organisation de compétitions.

La plateforme n'a pas été développée par notre université mais l'appropriation de son fonctionnement par notre ingénieur a donné lieu à un rapprochement remarquable avec les développeurs de Codalab. Nous avons pu exprimer nos besoins et une dynamique de mise en commun des ressources s'est mise en place. L'enjeu principal, pour les prochaines années, est d'arriver à la concrétiser, notamment par la mutualisation des ressources de calcul. Ces ressources sont nécessaires à deux étapes des compétitions : la phase d'entraînement et la phase de test. Classiquement le coût de calcul est à la charge des participants pour l'entraînement et à celle de la plateforme pour le test. Ce paradigme peut être remis en question dans deux types de contextes : 1) la quantité de mémoire ou la puissance de calcul requises ne sont pas facilement accessibles aux étudiants ; 2) elles sont rendues accessibles aux étudiants par des accords spécifiques avec des centres de calcul mais ne sont pas accessibles aux développeurs de la plateforme. A priori des participants venus du monde entier ont des ressources qui se limitent à leur ordinateur personnel, et les universités ne souhaiteront sans doute pas ouvrir leurs moyens à n'importe qui, d'autant qu'il est difficile de vérifier que ces moyens sont réellement utilisés uniquement pour la compétition. Il y a donc des accords à trouver pour qu'à la fois les étudiants et la plateforme Codalab puisse bénéficier des ressources de calcul combinées de plusieurs universités. L'une des compétitions organisées en 2019-2020 portera sur des données massives, et ce sera un challenge pour les organisateurs d'arriver à la faire fonctionner via Codalab, ne serait-ce que la partie de test.

La salle multimodale est à présent utilisée pour les compétitions du volet « formation » du projet, mais aussi celles liées à son volet « recherche » et pour divers TP classiques ayant lieu dans l'UFR hébergeant la salle. Ceci manifeste une mutualisation réussie de ce type de ressource, d'autant que les étudiants proviennent parfois d'une autre université.

L'Idex demande régulièrement des indicateurs concernant le nombre d'étudiants bénéficiant du projet, le nombre de formations, d'heures, le type de pédagogie pratiquée (formation par la pratique, via une plateforme ou non, pédagogie active, interdisciplinarité). Si ces indicateurs sont satisfaisants, l'essaimage n'est pas allé aussi loin que nous l'aurions souhaité. Certes, les problèmes traités sont issus des domaines de la santé, la biologie ou la physique. Il y a donc une forme faible d'interdisciplinarité dans la mesure où les étudiants sont amenés à côtoyer des données – mais pas d'étudiants – d'autres disciplines : ils sont eux-mêmes tous issus du domaine de la science des données au sens large. Afin d'arriver à une forme plus complète d'interdisciplinarité, nous souhaiterions associer des étudiants à des projets issus non seulement de la science des données, mais aussi des sciences socio-économiques, la communication, le management, le journalisme, la cognition, la psychologie expérimentale, ou autres. Un autre projet concerne les étudiants de M2 de l'École supérieure du professorat et de l'éducation (ESPE) et du master de statistique. Les étudiants de M2 de l'ESPE sont amenés à mettre en place une expérimentation pédagogique, à recueillir des données pendant leur stage et à les analyser. Chaque groupe d'étudiants de l'ESPE serait binômé avec un étudiant en statistique pour l'aide à la mise en place de l'expérience et de l'analyse. L'ouverture devra également se faire du côté d'étudiants et de formations extérieurs à Grenoble. Ainsi, les compétitions pourraient contribuer à la visibilité nationale et internationale de l'Idex et de la ComUE UGA. Cette ouverture pose la question de l'ouverture des données également. Par exemple dans la compétition « polluants atmosphériques » de 2018-2019, les données météorologiques ne peuvent être rendues publiques. Le projet pourrait s'efforcer d'utiliser des données libres et d'interagir avec la plateforme de stockage et de partage de données « PersciDo » du Labex Persyval à Grenoble.

D'un point de vue plus pédagogique et moins stratégique, nous avons identifié un intérêt à mieux expliciter aux étudiants la nature interdisciplinaire du projet et à les aider à mieux prendre en main leur rôle. Dans cette perspective, l'année prochaine, le déroulement de la compétition sera un peu modifié : dans un 1er temps, les équipes ne seront pas interdisciplinaires mais au contraire, spécialisées. Chaque équipe aura un aspect spécifique à prendre en main (par exemple pour le M2 informatique : prise en main du cluster de calcul). Puis dans un 2e temps, les équipes seront dissoutes et des équipes pluridisciplinaires seront constituées, de manière à favoriser l'explicitation par chaque membre aux coéquipiers, des compétences spécifiques acquises en vue de la réalisation de la solution. La compétition envisagée est relative à la prédiction de l'épaisseur de couches de neige / glace en montagne, l'identification de sous-couches et de leur densité, à partir de séries temporelles d'images multispectrales avec des problèmes d'interpolation spatiale. Ce sujet va dans le sens d'une progression dans la technicité des infrastructures utilisées au fil des ans, pour vérifier que nous arrivons à couvrir une palette suffisante de situations réalistes possiblement rencontrées en entreprise : données temporelles la 1re année, prédictions probabilistes la 2e année, gros jeu de données nécessitant des données et calculs distribués la 3e année.

Du point de vue des étudiants, les compétitions organisées dans le cadre de la formation sont une opportunité de valoriser le travail en équipe et l'analyse de données pour des problèmes réels, mais aussi de s'entraîner à d'autres compétitions plus visibles. Ainsi, suite à l'un des challenges, des étudiants ont de nouveau fait équipe pour participer aux Data Science Games[5] 2018 et se sont qualifiés pour la finale au rang de 10e / 340, pour être classés en finale avec le rang de 6e /20.

Remerciements

Cette communication bénéficie d'un financement dans le cadre des investissements d'avenir par le projet Idex Université Grenoble Alpes porté par la Communauté Université Grenoble Alpes (ComUE).

Bibliographie

[1] Agarwal, R., Bapna, R., Yong Goh, K., Ghose, A., Shmueli, G. et Slaughter, S. (2014). Does Growing Demand for Data Science Create New Opportunities for Information Systems? *In :*Karahanna, E., Srinivasan, A. et Tan, B. (ed.), *Thirty Fifth International Conference on Information Systems (ICIS 2014): Building a Better World through Information Systems, 14-17 décembre 2017, Auckland, New Zealand*. Publ. Association for Information Systems (AIS).
[2] Besse, P. (2017). Enseigner la Science des grosses Données à l'Université de Toulouse-INSA. *In* : *Colloque Francophone International sur l'Enseignement de la Statistique – CFIES2017, 6-8 septembre 2017, Grenoble*.
[3] D'Aquin, M., Dietze, S., Drachsler, H., Guy, M., Herder, E. et Parodi, E. (2014). Building the Open Elements of an Open Data Competition. *D-Lib Magazine* 20 (5/6). http://www.dlib.org/dlib/may14/daquin/05daquin.html.
[4] Davenport, T. H. et Patil, D. J. (2012). Data scientist. *Harvard business review*, 90(5), 70-76.
[5] Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.
[6] Drachsler, H., Stoyanov, S., d'Aquin, M., Herder, E., Guy, M. et Dietze, S. (2014). An evaluation framework for data competitions in TEL. *In* : Rensing, C. *et al.* (ed.), *Ninth European Conference on Technology Enhanced Learning (EC-TEL 2014), Graz, Austria*. Springer, Cham, p. 70-83.
[7] Durand, J.-B. (2017). Challenges d'analyse de données : une formation par la pratique transversale et multidisciplinaire en science des données. *In* : *Colloque Francophone International sur l'Enseignement de la Statistique – CFIES2017, 6-8 septembre 2017, Grenoble*.
[8] Molla, D. (2013). Overview of the 2013 ALTA Shared Task. *In* : Karimi, S. et Verspoor, K. (ed.), *Proceedings of the Australasian Language Technology Association Workshop, 4-6 décembre 2013, Brisbane, Australia*. Queensland University of Technology, Brisbane, Australia, p. 132-136.

5 https://datasciencegame.com/2018/