Naissance d'un projet :

Convaincu depuis longtemps que les sciences de la donnée, des méga-données, sont un enjeu majeur de nos sociétés, ParisTech et ses écoles ont mis en œuvre un grand nombre d'initiatives dans le domaine : masters, mastères spécialisés, formation continue, chaires d'enseignement et de recherche, etc.

Courant 2014, dans la continuité de ces actions, ParisTech, en partenariat avec trois de ses écoles (ENSAE ParisTech, Télécom ParisTech et ENSTA ParisTech), a décidé de lancer pour 2015 un projet de challenge international étudiant autour **des Big Data**[1] (ou méga-données) : Le Data Science Game.

Data Science Game

Une initiative de ParisTech et ses écoles

ENSAE ParisTech — ENSTA ParisTech université PARIS-SACLAY — TELECOM ParisTech

Objectifs :

- Encourager les étudiants à mesurer leurs capacités ;
- Démontrer l'importance de ce domaine pour ParisTech et ses écoles sur la scène internationale.

Il aura fallu un an pour monter de toute pièce un challenge ambitieux, pertinent et fédérateur autour des méga-données, sur le modèle des challenges d'économétrie qui existent depuis de nombreuses années et font référence dans le monde. Un an pour monter une équipe d'organisation, trouver des partenaires, structurer la compétition, rassembler des participants partout à travers le monde et mener la compétition à son terme. À la clé, le sacre de la meilleure équipe étudiante en Big Data.

[1] Les big data, littéralement les « grosses données », ou méga-données, parfois appelées données massives, désignent des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information. L'on parle aussi de datamasse en français par similitude avec la biomasse (définition Wikipedia : ici).

Construire une nouvelle compétition sur la scène internationale n'est pas chose aisée. Pour ce faire, et parce que ParisTech et ses écoles croient fortement à la force de leurs étudiants, une décision collective a été prise de monter un projet à structure mixte :

- Un pilotage par **ParisTech et ses écoles**[(2)] [2]mené par Guillaume Gaudron (ENSAE ParisTech), chef de projet ;
- **Une équipe d'organisation**[(3)] [3]entièrement composée d'étudiants de ParisTech, porté par Benjamin Donnot (3ème année à l'ENSAE ParisTech).

Fin 2014, l'équipe s'est attelée à mettre en œuvre ce projet ambitieux, dans un calendrier contraint. Trois axes clés ont permis de concrétiser ce projet :

- Concevoir un challenge pertinent, fondé sur des jeux de données réels ;
- Nouer des partenariats structurants pour sponsoriser et accompagner la compétition et lui permettre d'exister ;
- Bénéficier d'une couverture internationale.

Fort du travail collectif de l'équipe de pilotage et d'organisation, des partenariats ont vite été noués pour sécuriser ces axes, avec deux figures majeures des enjeux numériques et digitaux dans leurs domaines :

Google

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

- Google France pour la fourniture des données et la construction du challenge. Avec un impératif : conserver le secret sur la teneur des sujets et des données ;
- Capgemini, pour l'accompagnement logistique avec notamment la mise à disposition dans le cadre du partenariat, du Château des fontaines, situé près de Chantilly, pour accueillir les participants.

[2] Equipe de Pilotage : Guillaume Gaudron (ENSAE ParisTech), Mathieu Trystram (ParisTech), Alexander Geperth (ENSTA ParisTech), Yves Grenier (Télécom ParisTech), Alain Bamberger (ParisTech), Guillaume Ravel (ParisTech).

[3] Equipe d'organisation (ENSAE ParisTech pour l'édition 2015) : Benjamin Donnot, Antoine Ly, Thibaud Laugel, Loïc Michel, Peter Naylor, Raphaël Denis, Antoine Guillot, Audrey Ribeiro, Lucile Beaune.

D'autres partenariats ont vite suivi, permettant de donner à la compétition l'envergure nécessaire pour lancer sa première édition : Sanofi-Aventis, Ekimetrics, Datascience.net, la fondation ParisTech et Capgemini Consulting.

Un site internet a également été mis en place autour de l'évènement : www.datasciencegame.com.

Les équipes :

Le Data Science Game a été conçu pour les étudiants de niveau Master qualifiés en « data sciences », réunis par équipes de 4.

Pour cette première édition, les réseaux de ParisTech ont permis la bonne diffusion de l'information sur cette compétition dans les Grandes Écoles et Universités en France et à l'étranger.

Pour sa première édition, le Data Science Game 2015 a vu s'affronter vingt équipes, onze issues d'écoles d'ingénieurs ou **d'universités françaises**[4], dont 5 écoles de ParisTech**,** et neuf issues d'établissements basés en Europe (Angleterre, Allemagne, Italie, Pays-Bas…) mais aussi **en Inde et en Russie**[5]**.**

[4] **Etablissements français** participant à la compétition : École Polytechnique, ENS Cachan, ENSAE ParisTech, Grenoble INP, INSA Toulouse, MINES Paristech, Télécom ParisTech, ENSTA ParisTech, Toulouse School of Economics, Université Paul Sabatier - Toulouse 3

[5] **Etablissements étrangers** participant à la compétition : Universitat Mannheim (Allemagne), University of Padova (Italie), Indian Institute of Technology Delhi (Inde), University College Dublin (Irlande), Universiti degli Studi di Milano (Italie), Sapienza University of Rome (Italie), University of Amsterdam (Pays-Bas), Moscow State University (Russie), Imperial College London (Angleterre).

Objectif des participants : Trouver des solutions à un problème complexe en s'appuyant sur l'analyse de données massives, à travers l'élaboration d'algorithmes de traitement capables de gérer et de comprendre ces données.

Compétences en Science de la donnée ou « Data Sciences » attendues :

- Extraire, nettoyer et manipuler de grands ensembles de données structurés ou non (web, texte, etc.) ;
- Construire des modèles prédictifs fondés sur des méthodes de statistiques prédictives et de techniques d'apprentissage (« machine learning »).

Écoles / Université	Nom d'équipe
École Polytechnique	Poly Unicorns
ENS Cachan	Ayrbus-Fishmanati
ENSAE ParisTech	Full Metal Scientists
ENSTA ParisTech	eNStA
Grenoble INP	Breaking Data
Imperial College London	Designated Neurons
Indian Institute of Technology Delhi .	CTech IITD
INSA Toulouse	GMM
MINES Paristech	Cocotte Data
Moscow State University	MSU
Sapienza University of Rome	Sapienza
Télécom ParisTech	Telecominers
Toulouse School of Economics	Toulouse School of Economics
Università degli Studi di Milano	MInd overflow
UniversitÃ© Paul Sabatier - Toulouse 3	SID
Universität Mannheim	UniMAnalytics
University College Dublin	Insight UCD
University of Amsterdam	Nedap Hinton Hyenas
University of Padua	A Giant Mind
UPMC	MADatascience

Fig 1. Les équipes en compétition

La compétition :

La compétition, entièrement en anglais, s'est disputée en deux temps :

1. **Une première phase, non éliminatoire, entre le 15 mai et le 15 juin 2015**, pour permettre aux étudiants de se familiariser avec les données et les outils ;
2. **Puis une compétition intensive le week-end des 20 et 21 juin** à l'issue de laquelle les équipes ont été classées et récompensées. Les épreuves ont eu lieu au sein du centre international de formation du Groupe Capgemini, Les Fontaines, situé à Gouvieux près de Chantilly.

Chaque phase de la compétition comportait un sujet spécifique avec son propre jeu de données, construit et fournit en collaboration avec Google France.

1. Cette première phase, amicale, s'est déroulée en ligne, depuis le site datascience.com qui permet la mise à disposition sécurisée de jeux de données et leur traitement.

Pour cette première épreuve, les étudiants ont reçu un jeu de données composé d'extraits de livres. Ce jeu de données, ou jeu d'apprentissage, était scindé en 2 colonnes avec, d'un côté, les textes et de l'autre, les auteurs correspondant. Les équipes devaient construire un modèle et des algorithmes pour identifier les auteurs de ces textes parmi William Shakespeare, Mark Twain, Oscar Wilde, Edgar Allan Poe, Jane Austen et Arthur Conan Doyle.

L'efficacité de leurs modèles était ensuite évaluée sur datascience.net avec un autre jeu, un jeu de test. Le score obtenu suite à la soumission correspondait au taux de bonnes réponses du modèle.

En dépit de son caractère amical, il a été demandé à chaque équipe de faire au moins trois soumissions durant la phase amicale afin de pouvoir évaluer les modèles et leur pertinence. Un classement amical permettait de suivre l'évolution des équipes et de la pertinence de leurs modèles, avec un score médian autour de 85% de bonnes réponses.

2. La deuxième phase constituait le cœur de la compétition et l'attrait principal du challenge. Elle s'est déroulée les 20 et 21 juin 2015, au château des Fontaines.

Les fontaines – © les fontaines

Les étudiants sont arrivés sur le site le vendredi dans la journée par leurs propres moyens ou ceux mis en place par l'organisation.

Accueillis au sein du château, ils ont pu profiter des installations et du cadre pour souffler quelques heures avant le début de la compétition. Après un dîner au château, une présentation du déroulé du week-end leur a été faite, précisant l'organisation de la compétition et ses règles essentielles, parmi lesquelles la stricte interdiction de faire appel à une aide extérieure. En revanche, les participants étaient libres d'utiliser des bases de données externes, voire des modèles pré-entraînés, pour renforcer leurs algorithmes.

La compétition s'est déroulée en 2 étapes :

1. **La première**, sans interruption, du samedi matin jusqu'au dimanche en fin d'après-midi, visant à construire un modèle sur un jeu de données différent de celui de la première phase.
2. **La seconde**, de 17h à 18h le dimanche, pour soumettre l'un des modèles élaborés à un nouveau jeu de données test et établir le classement final.

Cette fois-ci, il a été demandé aux participants de classer des vidéos Youtube en 15 catégories, chaque vidéo étant associée à une seule catégorie. Leur but : utiliser l'ensemble des données à leur disposition pour classer ces vidéos : titres, descriptions, durées, commentaires, dates, formats, etc. La difficulté ? Des problèmes de syntaxe, d'orthographe (ex : dans les commentaires), de langue, etc.

Pour ce faire, les équipes avaient à leur disposition un jeu d'entrainement de 240 000 vidéos. Chaque vidéo étant décrite par une quinzaine de variables et les catégories associées sur le modèle des textes et des auteurs de la première phase. Comme lors de la phase amicale, les équipes ont développé des modèles à partir de ce jeu d'apprentissage puis soumis ceux-ci sur le site datascience.net pour vérifier leur justesse à l'aide d'un jeu de données test.

La compétition a eu lieu en continu du samedi à 7h du matin (!) au dimanche à 17h, y compris durant la nuit. Les équipes étaient libres de s'organiser comme elles le souhaitaient. Leur seule contrainte : travailler et laisser leurs ordinateurs dans la salle de compétition mise à leur disposition. Les nombreux repas et buffets réapprovisionnés en permanence fournissaient les forces nécessaires à leur concentration.

Durant cette phase, il a été demandé aux équipes quatre soumissions obligatoires (deux par jour à des horaires donnés), afin de réaliser des classements intermédiaires. Le site internet affichait leurs meilleurs résultats parmi toutes les soumissions faites.

Dès 16h le samedi, les Russes de l'Université de Moscou (MSU) ont pris la tête de la compétition avec un score de 70%, talonnés par L'Institut de Technologie de Delhi, et les équipes de l'École Polytechnique et de l'ENSAE ParisTech.

1.	MSU	40 contributions	20/06/15 15:17	Score 73,36108%
2.	ctech IITD	4 contributions	20/06/15 15:54	Score 69,57502%
3.	Poly Unicorns	10 contributions	20/06/15 13:16	Score 66,24062%
4.	FullMetalScientists	2 contributions	20/06/15 14:22	Score 65,91936%
5.	MADatascience !	4 contributions	20/06/15 12:46	Score 65,50655%

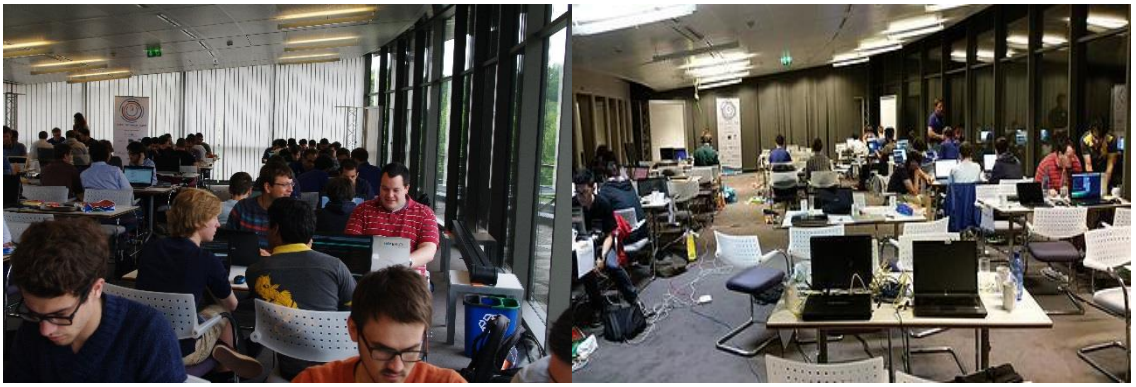
Fig 2. 1er classement intermédiaire

La compétition a continué ainsi jusqu'au dimanche 17h. A cette heure-là, le classement intermédiaire montrait les équipes au coude-à-coude : l'équipe de Russie, MSU, toujours en tête, suivi par Télécom ParisTech et Sapienza de l'Université de Rome.

1.	MSU	112 contributions	21/06/15 14:25	Score 75,45016%
2.	Telecominers -	21 contributions	21/06/15 16:02	Score 73,91034%
3.	Sapienza	71 contributions	21/06/15 16:46	Score 73,72380%
4.	Designated Neurons	31 contributions	21/06/15 16:28	Score 73,60549%
5.	MADatascience BRO	93 contributions	21/06/15 16:48	Score 73,43104%
6.	ctech IITD	46 contributions	21/06/15 13:41	Score 73,34295%
7.	Full Metal	27 contributions	21/06/15 15:36	Score 73,30322%
8.	Nedap Hinton Hyenas	70 contributions	21/06/15 16:48	Score 73,25313%
9.	Poly Unicorns	81 contributions	21/06/15 13:40	Score 72,97160%
10.	Breaking Data	25 contributions	21/06/15 15:45	Score 71,50000%

Fig 3. 4ème et dernier classement intermédiaire

Tout s'est très bien déroulé durant ces 34h non-stop de compétition, émaillées par quelques soucis techniques inhérents à ce type de compétition mais vite pris en charge par les équipes.



Les écoles de ParisTech : ENSAE ParisTech, Télécom ParisTech, Ecole Polytechnique, MINES ParisTech et ENSTA ParisTech

À 17h le dimanche, les travaux et le classement intermédiaire ont été arrêtés et il a été demandé aux équipes de procéder à une ultime soumission, celle déterminant le classement final. Les équipes ont choisi l'un de leur modèle qui a été ensuite utilisé avec un nouveau jeu de données test.

La remise des prix :

La cérémonie de remise des prix de la première édition du Data Science Game a eu lieu le dimanche 21 juin à 18h.

Pour l'occasion, l'ensemble des équipes participantes ainsi que les sponsors de la compétition étaient réunis aux Fontaines. Orchestré par Benjamin Donnot (ENSAE ParisTech), responsable de l'équipe étudiante d'organisation, la cérémonie visait a récompensé les 5 premières places de la compétition par un prix high-tech remis par chacun des sponsors (drones, tablettes, téléphones dernier cri, casques audio haute performance, serveurs personnels).

Après un discours d'introduction par Julien Pouget, directeur de l'ENSAE ParisTech, pour le compte de ParisTech et de ses écoles et présentant les convictions des écoles dans le domaine des Big Data, la cérémonie s'est poursuivie avec une intervention de Paul Séguineau, d'Ekimetrics venu remettre le 5ème prix. C'est l'Imperial College de Londres qui a été récompensé.

Puis, Anne Olivier de Sanofi-Aventis a remis le prix à la quatrième équipe, les Hinton Hyenas de l'Université d'Amsterdam, qui a remonté de manière fulgurante lors du classement final.

Le 3ème prix a été remis par Guillaume Ravel, de la Fondation ParisTech, à l'équipe des Telecominers de Télécom ParisTech.

Capgemini est venu remettre le prix de la seconde place lors d'une intervention de Marc Chemin. Ce 2ème prix a été décerné à l'équipe italienne de Rome, de l'Université Sapienza.

Enfin, le 1er prix a été décerné à l'équipe de l'université de Moscou, qui n'a pas quitté une seule fois la tête du classement pendant toute la durée de la compétition. C'est Anne-Claire Haury de Google France, ayant participé à l'élaboration du contenu du challenge, qui leur a remis ce prix.

ParisTech est fier de ce palmarès international, 5 places, 5 pays, et félicite toutes les équipes lauréates.

1. **Russie** : Université de Moscou
2. **Italie** : Université de Rome
3. **France** : Télécom ParisTech
4. **Pays-Bas** : Université d'Amsterdam
5. **Royaume-Uni** : Imperial College

Une fois les prix remis, direction le château pour une photo de groupe et un cocktail de célébration bien mérité où participants et sponsors ont pu faire plus ample connaissance et relâcher un peu la pression.

Les équipes ont ensuite quitté le site dans la soirée ou le lundi matin. ParisTech et les équipes d'organisation espèrent qu'elles garderont un bon souvenir de cette 1ère édition du Data Science Game, qui, pour ParisTech a été un succès et à pleinement rempli les objectifs fixés au départ.

ParisTech remercie l'ensemble des équipes pour leur participation en espérant les revoir pour la prochaine édition. ParisTech vous donne rendez-vous l'année prochaine !

Plus d'informations sur : www.paristech.fr et www.datasciencegame.com