J'ai testé pour vous... un MOOC

Nathalie Villa-Vialaneix

J'AI TESTE POUR VOUS... UN MOOC

Nathalie VILLA-VIALANEIX[1]

I HAVE TESTED FOR YOU... A MOOC

RÉSUMÉ

Le terme « MOOC » désigne aussi bien des plateformes d'enseignement en ligne dont la particularité est de proposer des cours ouverts que les cours eux-mêmes. Depuis les premiers MOOCs, qui ont vu le jour en 2012, ceux-ci ont connu un développement important et une attention forte des gouvernements et des universités, partout dans le monde. Le but du présent article n'est pas de proposer un diagnostic de ces cours en ligne, ni même un diagnostic des plateformes en général, mais de me focaliser sur un cours de statistique proposé sur un des plus grands MOOCs mondiaux pour montrer ce que les méthodes pédagogiques proposées dans celui-ci ont d'innovantes, d'en expliquer l'intérêt et les éventuelles limites.

Mots-clés : *MOOC, analyse de données, cours en ligne, CLOM.*

ABSTRACT

The word « MOOC » is used for on-line teaching platforms as well as for the courses available on these platforms. Since the first MOOC, released in 2012, a huge number of new courses have been proposed on these platforms and MOOCs have gained a large amount of attention from governments and universities all around the world. The purpose of the present article is not to provide a diagnosis for all MOOCs but to focus on a specific course, in the area of data analysis, which is offered by one of the biggest on-line platform. I will try to explain which innovative and interesting practices are used in these courses and also to underline their main drawbacks.

Keywords : *MOOC, data analysis, on-line course.*

1 Les MOOC, mon MOOC

1.1 Bref panorama du phénomène « MOOC »

MOOC (Massive Online Open Course[2]) est un acronyme qui désigne, à l'origine, des plateformes d'enseignement en ligne dont la particularité est de proposer des cours ouverts (à la différence des plateformes d'enseignement en ligne que beaucoup d'universités françaises ont ouvertes mais qui sont restreintes aux étudiants légalement inscrits dans un cursus de l'université). Le principe des MOOCs est de rendre le savoir accessible à tous et, pour ce faire, d'utiliser les supports numériques dans leurs diversités : les premières et principales plateformes en ligne ne se contentent pas de mettre à disposition des supports classiques (polycopiés de cours, énoncés d'exercices et correction...) mais utilisent les possibilités

[1] SAMM, Université Paris 1, 90 rue de Tolbiac, 75013 Paris ; nathalie.villa@univ-paris1.fr

[2] Cours en Ligne Ouverts et Massifs, parfois aussi appelés « CLOM » (Cours en Ligne Ouverts et Massifs) en français.

d'interactivité offertes par les technologies web avec des vidéos, quiz, forum, wiki, ... Les plateformes proposent aux étudiants de suivre les cours à leur rythme, de manière autonome et avec la progression qui leur convient, ou bien elles synchronisent l'avancement de tous les étudiants inscrits à un cours sur une période donnée de quelques semaines. Ainsi, sur les plus grosses plateformes américaines, certains cours sont suivis simultanément par plusieurs dizaines de milliers d'étudiants, localisés dans le monde entier. Enfin, certaines plateformes proposent de sanctionner la réussite à un cours par un certificat de réussite, qui certifie que le cours a été suivi et qu'une proportion minimale d'exercices réussis a été atteinte.

Les premières plateformes de cours en ligne ont vu le jour en 2012 avec Udacity[3], fondée en février 2012 par Sebastien Thrun (professor à l'université de Stanford et employé chez Google) et Mike Sokolsky (ancien personnel des universités de Stanford et Alberta), suivi de peu par Coursera[4], fondée en avril par Daphne Koller (professeur à l'université de Stanford) et Andrew Ng (également professeur à l'université de Stanford) et par edX[5] fondée par le MIT et l'université de Harvard. Depuis, elles se sont développées, soit par l'intégration d'un nombre important d'universités et de cours dans les plateformes déjà existantes[6], soit par la création de nouvelles plateformes qui restent malgré tout des acteurs de moindre importance face aux géants américains (voir Brafman, 2013, pour un petit bilan des MOOCs publié dans Le Monde, le 29/05/2013). L'engouement massif que ces plateformes ont suscité ont fait des MOOCs un objet d'attention important des institutions nationales et internationales et même de certaines grandes entreprises[7] : ainsi, le groupe d'experts de haut niveau « Former les professeurs à l'enseignement » de la commission européenne annonçait, dans son communiqué du 18 juin 2013 :

> *« Le groupe de haut niveau va s'atteler à la seconde partie de sa mission, centrée sur la manière de donner un effet maximal aux nouvelles méthodes permettant de produire un enseignement supérieur de qualité, comme les cours en ligne ouverts et massifs (Massive Open Online Course, ou « MOOC »), qui permettent aux personnes d'accéder à l'enseignement supérieur chez elles. Des partenaires dans 11 pays ont récemment lancé les premiers MOOC paneuropéens avec l'aide de la Commission européenne (IP/13/349). »*

Le premier MOOC paneuropéen a vu le jour en avril 2013 et se nomme OpenupEd[8] et compte actuellement 11 partenaires dont le ministère français de l'enseignement supérieur et de la recherche. La France a également lancé son premier MOOC national, FUN (France Université Numérique[9]), à l'initiative du ministère de l'enseignement supérieur et de la recherche. Ce MOOC est porté par trois acteurs clés (INRIA pour le déploiement de la

---

[3] https://www.udacity.com

[4] https://www.Coursera.org

[5] https://www.edx.org

[6] Coursera a actuellement plus de 70 institutions partenaires dans le monde : https://www.Coursera.org/partners ; edX près de 30 https://www.edx.org/schools

[7] Dans le domaine de la statistique, SAS© s'est, par exemple, investi dans la création d'un cours sur Coursera en partenariat avec l'université américaine de Wesleyan (voir Finkel, 2013).

[8] http://www.openuped.eu

[9] https://www.france-universite-numerique-mooc.fr

plateforme qui s'appuie sur la technologie d'edX[10], CINES pour l'hébergement et RENATER pour les infrastructures réseaux).

Au-delà des objectifs enthousiastes mis en avant par les plateformes américaines,

> *« We believe in connecting people to a great education so that anyone around the world can learn without limits. »*[11]

les MOOCs intéressent gouvernements, universités et organismes de formation car ils permettent de toucher un large public, en formation initiale ou continue, de mutualiser les ressources pédagogiques, de modifier les rapports à l'enseignement (en diminuant le nombre d'heures de cours mais en augmentant le suivi individuel). Toutefois, le modèle économique sur lequel repose les MOOCs n'est pas clair : si certaines plateformes proposent des certifications payantes à l'issue d'un MOOC, aucune université ne s'est encore, semble-t-il, avancée à valider officiellement un cours proposé sur une plateforme. Si l'ambition d'intéresser des recruteurs au phénomène est affichée – via les statistiques de participation et de réussite des apprenants – de telles pratiques sont encore à un stade très préliminaire (voir Finkel, 2013).

1.2 MOOC et statistique

Les MOOCs proposent un panel très large de cours, en différentes langues bien que l'anglais soit la langue principalement utilisée. La statistique fait évidemment partie de ce catalogue. Par exemple,

- pour Udacity, 5 cours sont répertoriés dans la catégorie « Mathematics » dont deux sont liés à la statistique « Introduction to statistics » (Sebastian Thrun) et « Statistics » (Sean Laraway et Ronald Rogers, San José University, US) et un cours d'un sujet connexe à la statistique, référencé dans « Computer Science » sous le titre « Introduction to Artificial Intelligence » (Peter Norvig et Sebastian Thrun) qui contient des éléments sur la statistique, l'apprentissage, les processus de Markov...
- pour Coursera, 33 cours sont répertoriés dans la catégorie « Statistics and Data Analysis », ces cours sont parfois très élémentaires, du « Statistics One » (Andrew Conway, Princeton University, US) ou au contraire très ciblés, « Case-Based Introduction to Biostatistics » (Scott L. Zger, Johns Hopkins University, US), et quelques-uns de ces 30 cours ne sont pas directement, à proprement parler, liés à la statistique elle-même, comme « Analyse Numérique pour les Ingénieurs » (Marco Picasso, École Polytechnique Fédérale de Lausanne, Suisse). Les cours proposés dans cette catégorie sont donnés en 3 langues : 31 cours en anglais, 2 en chinois et 1 en français ;
- pour edX, 15 cours sont proposés dans la catégorie « Statistics and Data Analysis », là encore certains très généraux comme « Introduction to Statistics: Inference » (Ani

[10] La solution technique proposée par edX, openedX est soutenue et financée par Google et a été utilisée et adaptée pour porter FUN, voir Dupont-Calbo (2013).

[11] About Coursera® : https://www.Coursera.org/about : « Nous croyons au fait de proposer aux personnes un bon enseignement de manière à ce que n'importe qui dans le monde ait la possibilité d'apprendre sans limite. »

Adhikari, UC Berkeley University, US) et d'autres plus spécialisés comme « Health in Numbers: Quantitative Methods in Clinical and Public Health Research » (Earl Francis Cook et Marcello Pagano, Harvard University, US) ;

- la plateforme française FUN propose un cours en statistique, « Fondamentaux en statistique ».

Par extension, le terme « MOOC » est actuellement fréquemment utilisé pour désigner les cours eux-mêmes, disponibles sur ces plateforme. C'est dans ce sens que nous l'emploierons dans la suite de cet article. Le but du présent article n'est pas de proposer un diagnostic de tous ces cours, ni même un diagnostic des MOOCs en général, mais de me focaliser sur l'un d'eux pour montrer ce que les méthodes pédagogiques proposées dans celui-ci ont d'innovantes, d'en expliquer l'intérêt et les limites. Ma motivation est d'illustrer que, contrairement à certaines idées reçues, les MOOCs ne sont pas un simple enchaînement de vidéo et de quiz (voir aussi à ce propos le post « MOOC et pédagogie par projet » du blog de Mathieu Cisel sur les MOOCs[12]). En effet, si une proportion croissante de personnes ont entendu parler du terme, peu semble en connaître la signification exacte et encore moins les ont utilisés[13]. En avant-propos, je précise que j'ai testé plusieurs de ces cours en ligne et que si mes propos n'ont pas de portée générale, je souhaite illustrer ici la particularité des cours en ligne au travers d'un exemple choisi et les différences que j'ai pu observer par rapport, à la fois aux cours en présentiel mais aussi par rapport aux supports multimédias que l'on trouve habituellement sur les sites web des universités ou des collègues.

Pour ce faire, je me focaliserai sur l'exemple d'un cours donné sur la plateforme Coursera, celui de « Data Analysis » de Jeff Leek (Johns Hopkins University, US). La présentation de ce cours est disponible à l'URL https://www.Coursera.org/course/dataanalysis et il est référencé dans les catégories « Health and Society » et « Statistics and Data Analysis ». C'est un cours qui a déjà été proposé deux fois sur la plateforme Coursera, la première session ayant débuté en mars 2013 et la deuxième en octobre 2013. La section 2 présentera les objectifs généraux du cours et le programme, la section 3 aura pour objectif de décrire l'organisation pratique du cours, son avancement et le matériel pédagogique mis à la disposition de l'apprenant, la section 4 fera le bilan de mon ressenti comme apprenante.

2 Présentation générale du cours

Le cours « Data Analysis » est présenté en deux paragraphes sur la page web qui lui est consacrée. Le cours est motivé par une référence aux « Big Data » et a pour objectif, d'une part de donner des bases pour apprendre à extraire des informations importantes d'un fichier de données (ce que l'on appelle généralement la « fouille de données ») et, d'autre part, d'apprendre à communiquer sur les résultats obtenus. Le cours se définit comme un « applied statistics course focusing on data analysis » (cours de statistique appliqué principalement orienté sur l'analyse des données). Le cours a une durée de 8 semaines et couvre des sujets

[12] http://blog.educpros.fr/matthieu-cisel et le post http://blog.educpros.fr/matthieu-cisel/2013/12/23/mooc-et-pedagogie-par-projet/ ; Matthieu Cisel est doctorant à l'ENS Cachan et effectue actuellement une thèse sur les MOOCs. Il consigne une partie de ses réflexions sur le sujet sur son blog.

[13] Voir les statistiques tirées d'une enquête du New York Times et relayée dans l'article (Finkel, 2013).

comme la régression linéaire, l'ACP, la validation croisée, la notion de p-value, … sans toutefois rentrer dans les fondements mathématiques de ces méthodes : le point de vue du cours est, en effet, de faire mettre en œuvre et expliquer comment utiliser les méthodes et concepts présentés. Aussi, le cours est basé sur application à des données réelles, via le logiciel statistique libre **R**[14] que tout apprenant peut donc installer librement sur son propre ordinateur. Peu de prérequis sont recommandés à part une connaissance préalable (au moins minimale) du langage de programmation **R** (des cours supplémentaires pour apprendre à utiliser le logiciel sont toutefois proposés à ceux qui ne le connaîtraient pas) et une certaine aisance à la rédaction en anglais (les évaluations étant essentiellement basées sur des rapports d'analyse de données devant être rédigés en anglais).

Le programme du cours est découpé de la manière suivante :

- la première semaine est consacrée à des présentations générales sur l'analyse de données avec une introduction basique à la manière dont les fichiers de données sont gérés par **R** et dont on peut construire des graphiques ;
- la deuxième semaine est consacrée à la gestion des données et à l'organisation d'une analyse statistique ainsi qu'à montrer des exemples de nettoyage (importation, homogénéisation des formats, …) et de résumés élémentaires des données ;
- la troisième semaine approfondit la représentation graphique des données et aborde les notions de classification (non supervisée) et d'ACP ;
- la quatrième semaine se focalise sur des notions de statistique plus avancées avec l'introduction de l'inférence statistique et de la notion de p-value et sur la régression linéaire (simple, multivariée, sur une variable qualitative...) ;
- la cinquième semaine approfondit encore les notions de modèle linéaire avec l'ANOVA, le modèle linéaire généralisé et des notions sur la validation et la sélection de modèle ;
- la sixième semaine aborde la notion d'apprentissage et de prévision et présente la validation croisée et les arbres de régression ;
- la septième semaine aborde des notions d'apprentissage plus avancées comme le lissage, le bootstrap ou la combinaison de méthodes de prévision ;
- la huitième semaine résume les notions et principes vus durant le cours et aborde les notions de correction des tests multiples et de validation des modèles par simulation.

Le programme du cours est donc assez copieux puisqu'il part de notions très générales sur l'analyse de données pour aller jusqu'à des notions assez fines de sélection de modèles et de combinaison de modèles. À ma connaissance, aucun cours d'un cycle universitaire français n'a un programme aussi ambitieux. En comparaison, les étudiants de DUT STID[15] abordent les notions des cinq premières semaines durant leurs 2 années d'études post-baccalauréat. Également, il faut souligner que ces notions sont abordées sans pratiquement s'appuyer sur des concepts mathématiques. La durée estimée (sur le descriptif du cours) de travail

[14] http://www.r-project.org
[15] http://www.stid-france.com/

hebdomadaire est de 3 à 5 heures. Lors de sa première édition, ce cours a été suivi par plus de 130 000 étudiants, ce qui en fait un des cours les plus suivis de la plateforme Coursera[16].

D'un point de vue personnel, la raison pour laquelle j'ai choisi ce cours est qu'il se situe dans mon domaine d'expertise pour que je puisse le suivre sans investissement de travail important (du moins, a priori) et avec suffisamment de recul pour pouvoir en apprécier les choix pédagogiques.

3 Matériel pédagogique

L'inscription au cours est simple et rapide : après avoir créé son compte sur la plateforme en ligne (uniquement un e-mail et un mot de passe sont nécessaires pour cela, bien que l'on puisse choisir d'enrichir son profil pour tenter de se faire connaître et de valoriser, après d'éventuels recruteurs, les connaissances validées au travers des cours proposés).

Sur Coursera, l'interface en ligne est conviviale et la navigation facile : le style de page web utilisé pour la mise en ligne du cours est aéré avec des gros boutons, divers outils de navigation interactifs, des polices aérées et assez peu de surcharge graphique. Un panneau à gauche du l'écran permet d'accéder aux différentes rubriques du cours. Ces rubriques dépendent en partie du cours lui-même mais sont relativement semblables. Sur la page d'accueil, la partie droite contient les informations sur les derniers contenus publiés, les dates limites de rendu des devoirs ou les discussions récentes des forums de discussion et la partie centrale répertorie les derniers messages de l'enseignant (également reçus sur sa boîte e-mail). Pour le cours de « Data Analysis », la page d'accueil la première semaine de cours se présente comme dans la figure 1.

[16] http://simplystatistics.org/2013/09/16/data-analysis-in-the-top-9-courses-in-lifetime-enrollment-at-Coursera/ ; voir également http://www.r-bloggers.com/complete-list-of-Coursera-courses-using-r-ranked-by-popularity/ qui liste les mentions « j'aime » obtenus par les différents cours de statistique de Coursera sur facebook® .

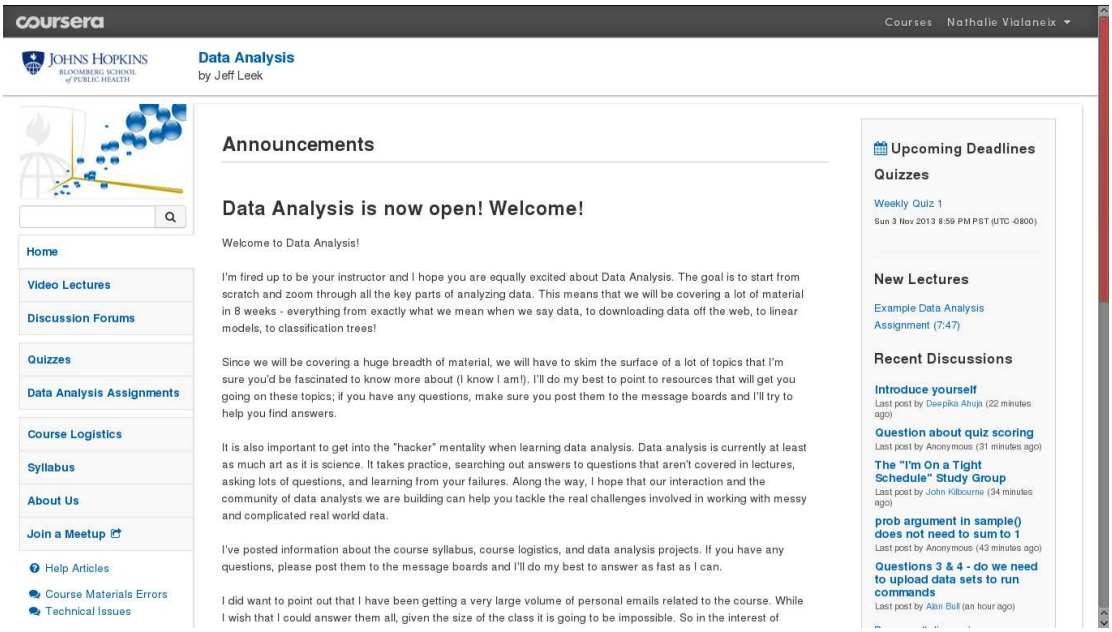

FIGURE 1 : Page d'accueil du cours « Data Analysis » lors de la première semaine de cours

Pour ce cours, on peut accéder, sur la gauche, à une page répertoriant les vidéos elles-mêmes, à des forums de discussion entre apprenants (dans lesquels interviennent parfois quelques assistants d'enseignement), à une page contenant des quiz, à une page contenant les devoirs à rendre, à une page récapitulant les objectifs et échéances du cours et à un lien vers un wiki extérieur contenant du matériel supplémentaire et que les apprenants peuvent également enrichir eux-mêmes. La première découverte d'un cours en ligne peut donc désarçonner car il faut se familiariser avec les divers items et leur utilité spécifique (prendre donc quelques dizaines de minutes pour parcourir le site). Les cours suivants sont plus faciles à aborder, la structuration étant très similaire.

1.3 Les supports de cours

Les supports de cours sont constitués principalement de vidéos : chaque semaine, de 5 à 10 vidéos, d'une dizaine de minutes chacune, sont publiées sur la plateforme du cours. Elles sont ordonnées par ordre de progression dans le cours et organisées en semaine. Lorsque l'apprenant a regardé une vidéo, celle-ci apparaît comme visionnée, comme montré dans la figure 2, ce qui permet de bien visualiser l'avancement du cours de l'apprenant.

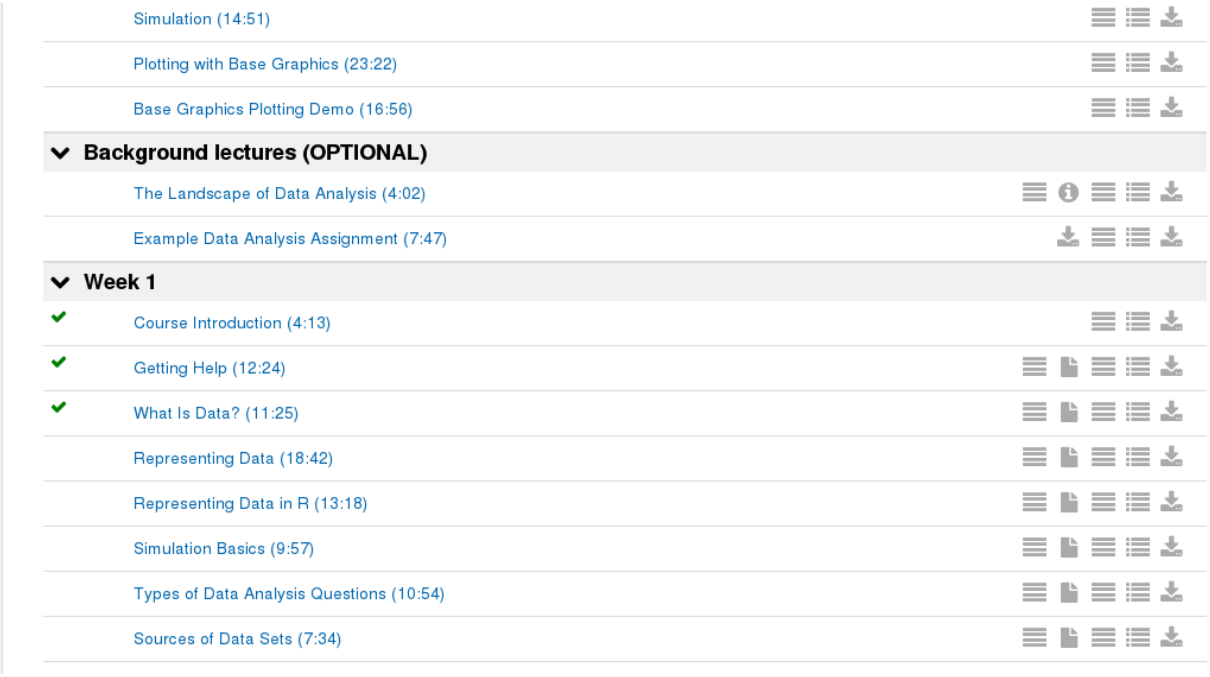

FIGURE 2 : Écran de présentation des vidéos. Les vidéos sont listées par semaine ; à gauche une marque verte indique que la vidéo a été visionnée ; à droite divers boutons permettent de télécharger la vidéo (au format MP4), les sous-titres et parfois un diaporama (formats PDF ou HTML) ou un fichier supplémentaire ou deux (un exemple d'analyse de données au format ZIP, avec plusieurs fichiers à l'intérieur : données, script R et rapport, par exemple).

Tous les supports fournis sont téléchargeables : les vidéos (au format MP4), les sous-titres en anglais (la synchronisation vidéo/sous-titre est prise en charge par la plupart des lecteurs de vidéos modernes), le PDF du diaporama de présentation quand la vidéo consiste en le commentaire d'un diaporama.

Le contenu même des vidéos est divers : soit un diaporama de présentation commenté par l'enseignant et sur lequel des actions sont réalisées (principalement des annotations manuelles ou bien des pointages sur un endroit précis du transparent par un pointeur de souris), soit des films, montrant éventuellement l'instructeur en train de parler, soit des vidéos d'un écran d'ordinateur sur lequel des actions sont réalisées... Toutes les vidéos sont en anglais, fournies presque immédiatement avec les sous-titres en anglais (parfois un délai de 1 à 2 jours est nécessaire entre la publication de la vidéo et la publication des sous-titres en anglais), et elles sont simultanément publiées sur la plateforme de sous-titrage amara[17]. J'ai moi-même sous-titré en français quelques-unes des vidéos du cours et ai pu constater que, dès la finalisation d'un sous-titrage, celui-ci est automatiquement proposé à l'utilisateur sur Coursera. Les vidéos peuvent directement être visionnées en ligne (c'est-à-dire, sans téléchargement préalable, en

[17] http://www.amara.org : Amara est un site web de mise en ligne de vidéos avec une interface conviviale et simple qui permet leur sous-titrage. Muni d'un compte sur ce site web, vous pouvez sous-titrer (ou corriger les sous-titres) de n'importe laquelle des vidéos mises à disposition sur ce site web dans la langue de votre choix.

« streaming ») et, bien que cela ne soit que très rarement le cas dans le cours « Data Analysis », le fait de visionner en ligne permet à certains enseignants d'insérer des quiz basiques sur la compréhension directe d'une notion à l'intérieur du cours même : la vidéo est alors stoppée jusqu'à ce que l'apprenant donne sa réponse à un QCM à réponse unique ou multiple, comme le montre la figure 3.

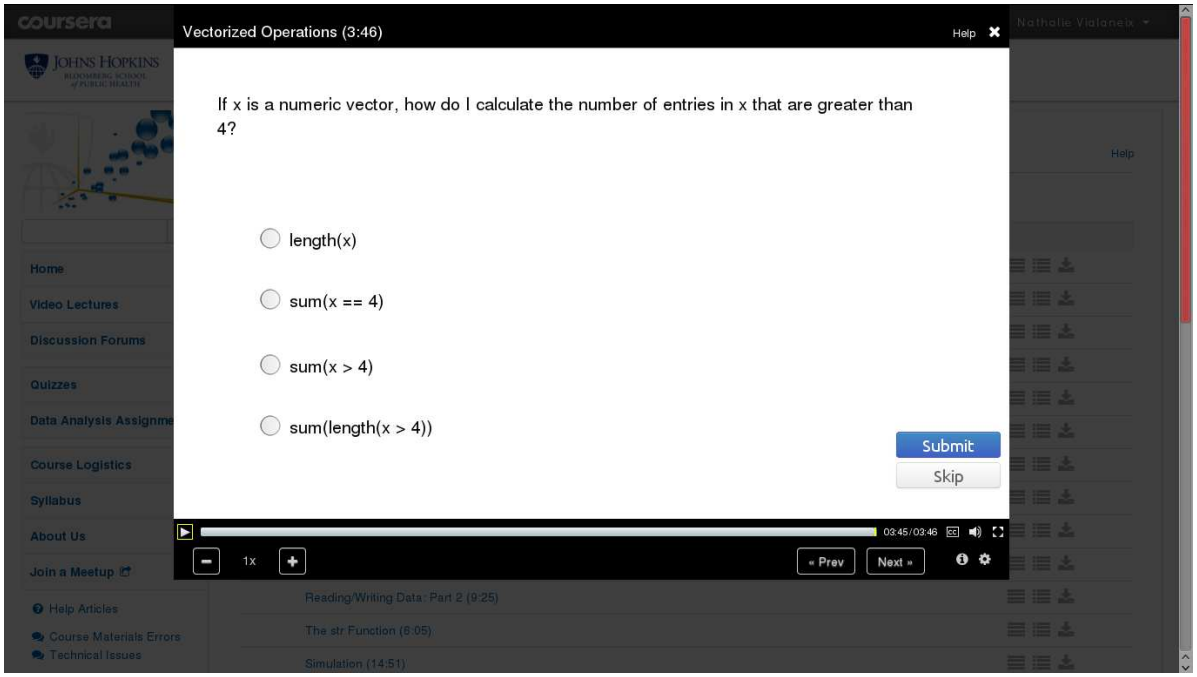

FIGURE 3 : Exemple d'une vidéo visionnée en streaming avec un quiz à l'intérieur : la vidéo est stoppée jusqu'à ce que l'apprenant réponde. L'apprenant a droit éventuellement à plusieurs essais puis la réponse est donnée et commentée.

Cette réponse est ensuite corrigée, la réponse correcte éventuellement expliquée avant que la vidéo ne passe au sujet suivant. Sur les cours utilisant cette modalité, 2 à 4 quiz sont ainsi proposés par vidéo, ce qui permet de conserver une attention soutenue durant la courte durée de la vidéo. Cette modalité est sans nul doute excellente pour conserver l'attention de l'apprenant qui se voit obligé d'écouter de manière active afin de comprendre de manière suffisante pour répondre à la question posée.

D'un point de vue pédagogique toutes les notions sont abordées sans recours à des connaissances en probabilités. Le principe général est de partir d'exemples, réels ou bien de petits exemples simulés, de faire des graphiques et de décrire les phénomènes observés pour illustrer et commenter la notion. Pour la notion de p-value, par exemple, celle-ci est introduite, dans une première approche « intuitive », par une description sommaire

> « *Idea : Suppose nothing is going on – how unusual is it to see the estimate we got ? »*

Il faut noter que la notion même d'estimateur n'est pas formellement définie dans les cours précédents (simplement illustrée sur un exemple en régression linéaire simple). Ensuite, une sorte de « mode opératoire » est donné :

« *Approach :*

1. Define the hypothetical distribution of data summary (statistic) when "nothing is going on" (null hypothesis)

2. Calculate the summary/statistics with the data we have (test statistic)

3. *Compare what we calculated to our hypothetical distribution and see if the value is " extreme" (p-value) »*

La notion de distribution a été introduite dans les cours précédents et illustrée, de manière empirique, par des histogrammes et principalement par des simulations selon diverses lois usuelles à l'aide du logiciel **R**. Cette notion est illustrée dans les diapositives suivantes par une application réalisée sur le logiciel **R** : une régression linéaire est réalisée sur un jeu de données public (le jeu de données « galton » du package `UsingR`) et la loi de l'estimateur du coefficient est donnée, de manière générale et sous l'hypothèse nulle

$$\frac{\ddot{b}_1 - b_1}{S.E.\left(\hat{b}_1\right)} \approx t_{n-2} \ .$$

Les notations standards et résultats classiques sont donc bien énoncés de manière mathématique mais tout cela est commenté de manière à expliquer intuitivement la notion qui doit donc pouvoir être comprise par un apprenant qui ne lit pas la notation mathématique. Le diaporama montre, toujours grâce au logiciel R, la densité de la distribution sous hypothèse nulle et place la valeur observée de l'estimateur sur le graphique pour illustrer ce que cette valeur a d'exceptionnel. Sur cet exemple, on dégage les principales caractéristiques du mode opératoire du cours :

- les notions statistiques sont introduites en limitant les références mathématiques mais celles-ci sont néanmoins présentes ;
- l'apprentissage est basé sur de nombreux exemples et graphiques issus de données réelles ou de simulations. Le code **R** qui permet de générer les exemples est toujours donné, ce qui est une aide supplémentaire pour quelqu'un qui connaît le langage de programmation mais peut probablement troubler quelqu'un qui ne le connaît pas.

1.4 Les exercices

Une fois que l'apprenant a regardé toutes les vidéos de la semaine, un quiz lui est proposé. L'apprenant doit répondre à ce quiz dans un délai d'environ une semaine après publication des vidéos et celui-ci est noté (voir la section « L'évaluation » ci-dessous). Il est composé d'une dizaine de questions à choix multiples qui sont des applications assez directes du cours, soit sous la forme de questions de cours, soit sous la forme de petits exercices très courts à réaliser avec **R**. Dans le cours de « Data Analysis », les réponses aux QCM sont d'une seule réponse exacte sur 4 ou 5 proposées mais dans d'autres cours les QCM sont parfois plus complexes avec possibilités de réponses multiples ou davantage de réponses possibles. Les QCM proposés sont généralement pertinents pour la compréhension du cours et l'assimilation par l'apprenant des notions proposées et de leur mise en œuvre. Compte tenu du code d'honneur

que l'apprenant s'engage à respecter en début de cours, il ne m'est pas possible de reproduire dans cet article un des quiz, puisque ceux-ci sont utilisés pour l'évaluation de l'apprenant.

L'apprenant n'est pas limité dans le temps pour donner ses réponses : il peut répondre aux questions qui lui paraissent les plus simples, les enregistrer puis revenir consulter des parties du cours et reprendre la réalisation de ses exercices. Lorsque l'apprenant est satisfait de ses réponses, il soumet son travail qui est immédiatement corrigé : il reçoit une note (généralement sur 10) et, si son score est maximal, il peut également consulter la correction des questions avec des commentaires justifiant les réponses. Si son score est inférieur au maximum, l'apprenant reçoit une correction partielle : cette correction lui indique quelles étaient les réponses correctes et les questions incorrectement traitées (mais sans toutefois lui dévoiler la réponse exacte aux questions mal traitées). Il peut alors choisir de refaire le quiz et cela jusqu'à 3 fois. Dans les essais ultérieurs, l'ordre des réponses est parfois modifié aléatoirement et certaines réponses légèrement modifiées.

La seule date limite imposée est celle du délai pour la notation : celui-ci est d'une semaine après la publication des vidéos. Après cette date, l'apprenant peut toujours répondre aux quiz mais ceux-ci ne feront pas partie de sa note d'évaluation finale. Le délai pour assimiler est donc relativement court puisqu'il y a une incitation forte à avoir regardé toutes les vidéos dans la semaine et à les avoir suffisamment bien comprises pour pouvoir répondre à tous les exercices. En revanche, un apprenant qui réussit à investir le temps nécessaire pour soutenir ce rythme a de fortes chances d'obtenir un très fort score aux quiz, du fait de la possibilité de répéter les essais. Dans d'autres cours, cette possibilité est plus restreinte : outre le plus grand nombre de réponses possibles et la possibilité d'avoir des réponses multiples, les réponses et/ou les questions varient légèrement (tout en ciblant le même type de compétences) et une pénalité est appliquée à la note à partir du 2ème ou 3ème essai (la note retenue pour le deuxième essai correspondant, par exemple, à 80 % de la note initiale), voire un seul essai est autorisé. Au final, si plusieurs essais sont autorisés, le meilleur score sur l'ensemble des essais est conservé.

1.5 L'évaluation

Le point difficile, dans un cours en ligne, est l'évaluation. La plupart des plateformes existantes proposent divers types de certificat pour valider les connaissances d'un apprenant à un cours :

- un certificat de réussite simple « certificate of accomplishement », gratuit, que l'on obtient si l'on a atteint une proportion d'exercices réussis minimale, définie à l'avance ;
- un certificat de réussite dit « vérifié », certifié par la plateforme de cours en ligne, et payant. Ce certificat peut être de nature diverse, parfois certifié par une grande université américaine (comme sur la plateforme edX) mais ne semble jamais donner droit à la validation d'une UE équivalente dans l'université correspondante. Le coût d'un certificat « vérifié » dépend de la plateforme et du cours mais se situe généralement aux environs de 40/60 US$. Pour le cours de « Data Analysis », la possibilité d'obtenir un tel certificat n'était pas offerte aux apprenants.

Dans la mesure où un certificat, éventuellement authentifié, voire validé par une grande université, est délivré, l'évaluation doit être conçue de manière sérieuse. Différentes modalités d'évaluations sont prévues :

- les quiz (voir la section « Les exercices ») : dans le cours de « Data Analysis », chacun des 8 quiz donnait une note sur 10 ; le total des points accordés pour la réalisation des quiz était donc de 80 ;
- des devoirs à effectuer à la maison qui peuvent prendre diverses formes : certains cours demandent de réaliser un travail de programmation en relation avec les notions du cours (qui ne relève pas réellement d'un cours d'informatique mais s'appuie sur un langage de programmation donné pour illustrer les connaissances acquises : l'évaluation est alors effectuée par l'exécution du programme par le serveur qui en vérifie les sorties). Pour le cours de « Data Analysis » la notion de « présentation des résultats de l'analyse » était assez centrale et les devoirs ont donc consisté en deux analyses de données à effectuer, chacune constituant en une brève description d'un jeu de données public avec sa problématique sous-jacente (deux petits paragraphes de quelques lignes). Le rendu était effectué en deux fichiers: le premier était un texte principal (2000 mots), incluant une introduction, une analyse structurée répondant à la problématique posée, une conclusion et des références. Le deuxième était une figure (formats PNG, JPG ou PDF) avec sa légende (500 mots). Les devoirs à rendre ont été fournis les semaines 3 et 6 avec un délai de deux semaines entre la date de publication du sujet et la date de remise du travail, qui est à télécharger sur le site web au format PDF ou bien à copier/coller directement au format texte. Chacun de ces devoirs était noté sur 40 points, soit 80 points au total également.

L'apprenant obtient un certificat de réussite s'il obtient au moins 100 points. Un certificat « avec mention » lui est décerné à partir de 144 points. Si le seuil de validation est relativement bas (les quiz sont assez faciles et rapides à réaliser et suffisent presque, à eux seuls, à obtenir le minimum nécessaire), la complexité du deuxième type de devoir est, en revanche, à souligner : non seulement l'apprenant doit savoir, parmi les méthodes abordées dans le cours, et qui, a priori, sont nouvelles pour lui, lesquelles mettre en œuvre. Il doit ensuite réaliser la mise en œuvre pratique avec un logiciel qui est un vrai langage de programmation et non un logiciel utilisable simplement de manière intuitive et sur des données qui ne sont pas de simples cas d'école mais des cas d'études réels, même si ceux-ci sont passés dans le domaine public (le premier fichier de données contenait 2 500 observations et 12 variables, à la fois numériques et qualitatives ; le second fichier de données contenait plus de 7 000 observations et plus de 500 variables, toutes sauf une, numériques). L'analyse des données pour répondre à la question posée demande donc des allers/retours multiples entre analyse, interprétation, affinement de l'analyse, nouvelle interprétation, etc. Enfin, le travail doit être synthétisé et structuré pour produire un compte-rendu et une figure (à rédiger en anglais) : les exigences pour ce rapport sont importantes puisqu'il doit être correctement rédigé et structuré, contenir une description des données assez complète, une justification des choix méthodologiques, une analyse des résultats et aussi une critique de ceux-ci et des possibilités d'ouverture. Le résultat final est évalué sur des aspects rédactionnels, des aspects de mise en œuvre pratique de méthodologies adaptées mais aussi des aspects de capacité d'analyse et de recul sur les résultats. En remarque personnelle, c'est le

type de rendu (en plus court) que l'on peut attendre d'un projet étudiant (DUT STID, cursus d'école d'ingénieurs) qui est le fruit de l'accumulation d'une ou plusieurs années d'études en statistique, accompagnées de cours de communication et de rédaction.

Même si la proportion des 130 000 apprenants qui rendent effectivement les devoirs est faible, il est exclu que l'équipe pédagogique (composé de l'enseignant et de quelques assistants d'enseignement[18]) puisse corriger les 2 ou 3 pages de rapport fourni. La modalité de notation choisie est donc la notation par les pairs. D'un point de vue pratique, la notation est organisée comme suit : après la date limite de rendu du devoir, chaque apprenant doit noter le travail de quatre autres apprenants, anonymisé : la notation est à effectuer par le biais de réponses à une vingtaine de questions, décrivant une compétence donnée qui est évaluée sur une échelle de 0 à 5, comme par exemple :

- « Does the analysis have an introduction, methods, results, and conclusions section? » (est-ce que l'analyse contient une partie introduction, une partie méthodes, une partie résultats et une partie conclusion ?)
- « Is the analysis written in clear and understandable English? » (est-ce que l'analyse est écrite dans un anglais clair et compréhensible ?)
- Are the statistical models correctly applied? (est-ce que les modèles statistiques sont utilisés correctement ?)

Une fois ces quatre évaluations effectuées, l'apprenant doit reprendre son travail et l'auto-évaluer : le résultat des évaluations des pairs donne la note finale de l'apprenant (la médiane par item a été utilisée). Quand le processus d'évaluation est terminé, on peut consulter le détail par item de sa note, et donc voir quels points sont à améliorer.

Comme le souligne Jeffrey Leek, l'enseignant de « Data Analysis » sur son blog « Simply Statistics »[19] :

> « *I knew in advance that it was likely to be the most controversial component of the class.* »[20]

Il justifie son choix par des considérations de faisabilité et économiques. Toutefois, ayant participé au processus de l'intérieur, je dois ajouter une justification complémentaire : sur les deux fois quatre travaux que j'ai eu à évaluer, une bonne moitié étaient réellement d'un bon niveau et intéressants. Sur un sujet aussi vaste qu'une problématique générale d'analyse de données, les réponses que l'on peut apporter sont très diverses et la lecture d'autres travaux, où des choix méthodologiques différents ont été faits, est extrêmement intéressante et participe en elle-même au processus d'apprentissage. On y lit des points de vue originaux, des vérifications que l'on n'avait pas pensé à effectuer, etc. J'ai été positivement surprise de la

---

[18] Je n'ai pas trouvé le nombre d'assistants d'enseignement du cours Data Analysis mais le premier e-mail de présentation du cours précise qu'il est réalisé par un co-instructeur, Roger Peng (qui s'occupe de la partie de formation à **R**) et de « teaching assistants » (en nombre indéfini). Sur un autre cours, celui de Roger Peng, « Computing Data Analysis » (d'une durée deux fois moins importante en nombre de semaines), le nombre d'assistants d'éducation était égal à 5.

[19] http://simplystatistics.org/2013/03/26/an-instructors-thoughts-on-peer-review-for-data-analysis-in-Coursera/

[20] « Je savais par avance que ce point serait le plus controversé du cours. »

qualité de plusieurs des travaux que j'ai eu à évaluer et ma propre évaluation de mon travail a été influencée par ceux-ci. Au final, mon auto-évaluation était toujours très proche de la note que j'ai obtenue. Il serait sans doute intéressant de quantifier cette question mais mon ressenti final était que la notation obtenue de cette manière-là était finalement de bonne qualité.

1.6 La pédagogie active

Bien que j'y ai peu participé moi-même, le cours proposait un forum et un wiki : il était possible pour les apprenants de poser leurs questions sur les forums (à l'exception de demander ou donner les réponses des devoirs et quiz) et de fournir du matériel additionnel personnel sur le wiki. L'animation sur les forums et le wiki a été importante et plusieurs apprenants ont proposé leur propre matériel pédagogique supplémentaire. Le nombre très important d'apprenants assure, à mon avis, la pertinence de ce type de modalité pédagogique : même si une faible proportion d'apprenants sont réellement actifs sur ces outils, ils suffisent à les rendre très vivants et la plupart des questions posées obtiennent une réponse correcte en un temps très court, sans une participation importante de l'enseignant (mais des assistants d'enseignement sont néanmoins indispensables pour surveiller et animer les forums). Ce type de cours s'appuie sur un apprentissage et une évolution au travers de projets personnels, faits à la maison, ce qui est assez éloigné des méthodes d'apprentissage les plus courantes que nous pratiquons en France où l'apprentissage repose très souvent sur des exercices plus courts et plus ciblés dont l'enseignant encadre la recherche en classe et pour lesquels il fournit souvent une correction commune au tableau. Ces projets demandent un investissement important de l'apprenant mais comportent aussi des aspects pédagogiques intéressants en eux-mêmes (aller chercher l'information, s'organiser, s'autonomiser par rapport à l'apprentissage), si l'apprenant a la maturité (cognitive et personnelle) nécessaire pour les aborder.

4 Commentaires et conclusion

Dans cette dernière partie, je voudrais faire le point sur mon ressenti comme apprenante. En avant-propos, je rappelle que le cours ne présentait aucune difficulté cognitive pour moi. Le temps hebdomadaire que j'y ai consacré a pourtant été, fréquemment, supérieur au temps nécessaire annoncé. En particulier, la réalisation des deux projets m'a demandé l'équivalent d'une journée de travail environ (le second projet a été particulièrement chronophage) et pourtant, ceux-ci ne présentaient pas, pour moi, de difficultés méthodologiques ni rédactionnelles. Le temps de notation des devoirs des autres apprenants était aussi important (bien que moindre comparé à celui de réalisation des devoirs) et pourtant, là encore, je n'avais aucune difficulté de compréhension des devoirs des autres apprenants et suis, de plus, rompue à la correction de copies. Ma conclusion à la fin des 8 semaines de cours a été de me demander comment quelqu'un qui partirait réellement avec un niveau basique en statistique et analyse de données pourrait supporter le rythme soutenu du cours. Pour quelqu'un qui voudrait réellement se former à l'analyse de données avec un niveau de départ faible, j'évalue le temps d'apprentissage hebdomadaire à pratiquement un mi-temps, entre l'assimilation des données et la réalisation des exercices. Ce temps n'est pas excessif puisque, comme je le soulignais dans la partie « Présentation générale du cours », le cours correspond probablement

à plusieurs cours d'un cursus universitaire français classique. Cela est cependant très loin de l'investissement annoncé qui serait plutôt celui nécessaire à quelqu'un qui aurait déjà des notions en analyse de données et souhaiterait avoir une vue globale du sujet et des approfondissements.

Le niveau de difficulté du cours était tout à fait raisonnable vu l'amplitude des sujets abordés : les notions présentées l'étaient de manière intuitive mais toujours parfaitement correcte. En illustrant des notions difficiles comme celles d' « effets confondants » ou de « sur-apprentissage » sur des données de petite taille, réelles ou simulées et en ayant recours aux graphiques, l'enseignant arrive à faire passer des notions fondamentales en analyse de données. Il ne me semble toutefois pas qu'un débutant puisse réellement, à l'issue du cours maîtriser le programme de celui-ci ; cependant, il aura sans doute gagné une bonne vision d'ensemble du sujet et aura un point de départ pour commencer à approfondir s'il le souhaite... à condition qu'il ait réussi à suivre le cours en entier. Je ne sais pas combien d'apprenants ont rendu les devoirs ni combien ont abandonné en cours de route mais, si j'ai parlé de la surprenante qualité de la moitié des devoirs que j'ai eu à évaluer, je n'ai pas encore parlé des autres : pour la plupart, ils étaient très en dessous d'un niveau de compréhension même basique du cours (erreur de compréhension des données, de la méthodologie, voire absence totale de tout traitement statistique). Les causes de ces échecs (manque de temps ou mauvaise assimilation du cours) sont malheureusement impossibles à analyser de l'extérieur. Pour un autre cours (d'informatique celui-ci), « Initiation à la programmation en C++ », de Vincent Lepetit, Jean-Cédric Chappelier et Jamila Sam de l'École Polytechnique de Lausanne, les statistiques d'abandon ont été données : sur un peu plus de 17 000 inscrits, 11 000 environ ont été actifs et un peu plus de 2 000 ont rendu au moins un devoir sur 4 demandés. 836 ont finalement obtenu un certificat de réussite, ce qui représente un peu plus de 5 % des inscrits : la proportion peut paraître faible mais le nombre en lui-même est déjà important (surtout pour un cours donné en français). Aucune information n'est donnée sur le profil des apprenants donc il est difficile de voir combien ont réellement acquis une connaissance valide par ce biais.

Par rapport aux cours « traditionnels », la pédagogie en ligne m'a séduite sur plusieurs aspects : d'un point de vue format, la durée (courte) des vidéos et le fait qu'elles soient entrecoupées parfois de QCM permet de maintenir l'attention de l'apprenant sans le lasser. L'interactivité, le fait de répondre à des quiz en ligne et d'avoir immédiatement le résultat de son travail et de pouvoir le corriger, est également un aspect très agréable et motivant. Enfin, une très grande force du type d'enseignement proposé par certaines plateformes en ligne comme Coursera est finalement la synchronisation de l'apprentissage de plusieurs dizaines de milliers de participants : il est indéniable que, d'une part le nombre important de participants et, d'autre part, les délais d'assimilation très resserrés créent une émulation qui est très bénéfique. J'ai eu l'occasion de suivre d'autres cours en ligne, dans lesquels l'apprentissage était désynchronisé et la vitesse d'évolution laissée à la libre appréciation de l'apprenant : il y est beaucoup plus difficile de s'imposer une discipline personnelle pour progresser dans l'apprentissage. L'impression de participer à une entreprise collective est un facteur de motivation supplémentaire et encourage effectivement une forme de pédagogie active. Cependant, certaines critiques soulignent que tous les cours en ligne ne sont pas de cette qualité : sur son blog « More or Less Bunk », très critique à l'égard des MOOCs, Jonathan

Reeds décrit un cours sur l'holocauste qu'il a suivi et dont il explique qu'il n'était constitué que de vidéos sans originalité qui existaient apparemment avant la publication du cours sur la plateforme[21]. Derrière un cours, plus que le support lui-même, l'investissement de l'enseignant et ses capacités pédagogiques personnelles restent la principale source de qualité du cours. D'autres critiques soulignent aussi que la pédagogie en ligne ne peut remplacer le contact direct : de ce point de vue, il peut naître chez l'enseignant une certaine frustration à faire un cours sans auditoire et de ne pas avoir les retours et réactions des apprenants pour orienter son propos[22]. À l'inverse, une fois le cours publié, les demandes des étudiants (même si elles ne proviennent que d'une petite proportion des inscrits) peuvent aussi être « massives ». Jeffrey Leek, précise en effet, lors du premier message de la deuxième session de son cours :

> *« I did want to point out that I have been getting a very large volume of personal emails related to the course. While I wish that I could answer them all, given the size of the class it is going to be impossible. So in the interest of fairness I won't be responding to any personal emails about the class. »*[23]

Ainsi, c'est l'intégralité de l'organisation des enseignements qui est à revoir, avec le soutien indispensable de « tuteurs » qui animent le cours et corrigent les principaux problèmes des apprenants.

En conclusion, l'expérience pédagogique m'a plu et je la recommande. Toutefois, elle a aussi ses limites : bien que le format des cours soit pensé pour être attractif et accessible, il n'en demeure pas moins que l'investissement de travail pour acquérir les connaissances est considérable et que ce travail est majoritairement un travail individuel, voire solitaire, malgré la présence de forums. Aussi, si je renouvelle régulièrement l'expérience « MOOC », j'ai tendance à picorer plutôt qu'à suivre un cours dans son entier et à me plier à l'exercice du rendu des devoirs, ce qui limite, de fait, ma compréhension des sujets abordés. Pour l'instant, les certifications obtenues via ce type de plateforme ne semblent pas avoir de valeur même symbolique sur le marché de l'emploi et on peut imaginer le public comme majoritairement intéressé par de l'auto-formation complémentaire pour une activité professionnelle déjà établie ou bien constitué de curieux souhaitant apprendre, sur un sujet donné, pour leurs loisirs.

Remerciements

Je remercie Jean-Michel Poggi pour les nombreuses références qu'il m'a envoyées, pour les discussions sur le sujet et ses idées toujours intéressantes. Je remercie également Jean Villa-Vialaneix pour sa revue de presse attentive.

[21] http://moreorlessbunk.wordpress.com/2013/09/06/why-Coursera-and-udacity-are-the-worst-things-that-ever-happened-to-moocs/

[22] Voir, à ce propos, le retour d'expérience de Sylvie Méléard, durant son cours « Aléatoire » (plateforme Coursera), donné à l'occasion des 30 ans de la SMAI et visualisable à http://smai.emath.fr/smai2013/smai30ans/

[23] « Je voulais souligner que j'ai reçu une quantité très importante d'e-mails personnels relatifs à ce cours. J'aimerais pouvoir répondre à tous mais, étant donné le nombre d'étudiants, cela m'est impossible. Ainsi, par soucis d'égalité, je ne répondrai à aucun e-mail personnel durant le cours. »

Références

[1] Brafman, N. (2013), Tous diplômés d'Harvard, le fantasme des MOOC, *Le Monde*, **29 mai 2013**.

[2] Dupont-Calbo, J. (2013), Derrière le MOOC à la française : Google, *Le Monde*, **16 octobre 2013**.

[3] Finkel, E. (2013), Data mining the MOOCs, *University Business*, **October 2013**. http://www.universitybusiness.com/article/data-mining-moocs

[4] Groupe d'experts de haut niveau : former les professeurs à l'enseignement (2013), *Communiqué de presse de la Commission européenne du 18 juin 2013*, **IP/13/554**. http://europa.eu/rapid/press-release_IP-13-554_fr.htm