==========

Date de rédaction : 13/11/2015
Nom du rédacteur du document : Matthieu Cisel
Spécialités : Analyse de données
==========

Cas d'étude Hubble: MOOCAZ

Scénario hubble : Scénarios 2 et 4
Personnes impliquées pour la collecte et l'analyse : Matthieu Cisel (collecte), Tony Doat (Traitement), Matthieu Cisel - Mattias Mano - Sébastien Iksal - Serge Garlatti (Analyse)
Période de la collecte : Printemps 2015
Periode de l'analyse : Septembre - Novembre 2015 ?

Dispositif d'apprentissage (Etude de cas de Hubble)

Type de dispositif : MOOC
Finalité de l'apprentissage : Apprendre à monter un MOOC
Utilisation du dispositif et fonctionnalités : Consommation de vidéo, réalisation de quizz, de devoirs, forums de discussion

Contexte de production de données : MOOC organisé sur FUN (Exceptionnellement, les images présentées sont celles issues d'un MOOC de Coursera, mais le principe reste le même)

Au besoin indiquer les différentes moments de la production (savoir si des données ont été produites sur plusieurs années)

La problématique posée pour l'analyse :

Scénario 2 V1 : On veut suivre les évolutions de types d'apprenants définis à l'avance. (8 comportements définis sur une semaine donnée)

Scénario 4 : On veut essayer de suivre des parcours d'apprenants et les changements d'état d'un format d'engagement à l'autre au fil des semaines

Objectifs de l'analyse : Reproduire des analyses réalisées dans le cadre d'un autre MOOC

Description du stockage des données:

Plateformes/outils utilisés: Les données sont stockées sur les serveurs de STEF en premier lieu, puis partagés sur Assembla. Le tableau se rapporte à cette dernière plate-forme.

Points forts de ces plateformes	Points faibles

Accessibilité relativement facile de la part de tous les acteurs du projet.	Pas de points faibles pour le moment pour les usages actuels

Production des données avant le traitement :

Décrire le processus de production des données brutes : Les serveurs du CINES stockent en théorie tous les logs de la plate-forme. Une extraction est requise, les logs extraits par l'équipe du CINES sont envoyés à FUN qui nous les retransmet.

Liste des variables initiales : L'ensemble des variables du jeu de données sont disponibles sur la documentation d'edX
http://edx.readthedocs.org/en/latest/internal_data_formats/tracking_logs.html
Parmi les nombreuses variables qui nous intéressent, nous allons pour le moment nous intéresser aux événements de déclenchement de vidéo, de téléchargement de vidéo, et de réalisation de quizz

Plateformes/outils utilisés:

Points forts	Points faibles
	FUN avait jusqu'à peu des problèmes de stockage des logs (des pertes importantes ont été constatées -30- 80%)

Description des pré-traitements: Transformer un fichier de logs qui se présente comme un mélange de JSON pas très propre en une matrice CSV plus facilement analysable

Objectifs des pré-traitements : Rendre un fichier de JSON plus digeste pour une analyse sur R.

Décrire le processus de pré-traitement: Une moulinette en Python mise en place par Tony décompose l'ensemble des événements de la partie GET ou POST en une série de colonne. Tony ne garantit pas le contrôle qualité du pré-traitement (mais en première instance c'est correct).

Outils utilisés: Moulinette Python Tony

Points forts	Points faibles
Rapide, efficace	

Description des analyses :

Description Analyse Scénario 2 V1

Objectif de la création de ces nouvelles données : On ne s'intéresse pas tant à aux données brutes qu'aux indicateurs construits à partir de ces données et censés décrire le comportement de participants.

Script R : Disponible sur demande (beaucoup trop large pour être collé dans un fichier comme celui-ci)

Mode de calcul des variables : Nous souhaitons étudier le comportement des individus tout le long du MOOC. Nous définissons, à partir des trois types de ressources pédagogiques, quatre actions possibles : avoir vu une vidéo, avoir téléchargé une vidéo, avoir répondu à un quiz, avoir effectué un devoir. Nous allons donc étudier la séquence d'actions des individus par semaine. Nous considérons que la même activité au sein d'une même journée (voir deux fois la même vidéo par exemple) est un doublon et ne considérons donc que les actions uniques de chaque jour. En revanche, durant une semaine, un individu peut visionner plusieurs fois la même vidéo. A partir de ces séquences, nous avons déterminé des états par semaine pour chaque individu :

Nom	Description
Viewer	Un apprenant ne regardant que les vidéos
Collector	L'apprenant a uniquement téléchargé les vidéos
Quizzer	L'apprenant se sera contenté d'effectuer les quiz
Active viewer	Un apprenant qui mêle actions de vidéos et de quiz
Completer	Un apprenant qui fait toutes les activités possible au sein de la semaine
Low completer	individus effectuant les devoirs couplés avec un autre type d'actions
Solver	Un individu ayant effectué uniquement les devoirs
Inactive	n'a effectué aucune action cette semaine

Cette classification nous permet d'observer l'évolution des individus durant les semaines de cours, nous excluons les activités précédant la première semaine et ayant lieu après la dernière semaine.

Liste des méthodes mise en œuvre : Construction de variables nouvelles (classification de comportements)

Mode opératoire technique, logiciels utilisés : R

Résultats : On réalise deux types d'analyse : la première est une analyse temporelle (on étudie le comportement semaine après semaine). La seconde est une analyse par section. On regarde le comportement sur chacun des modules du cours.

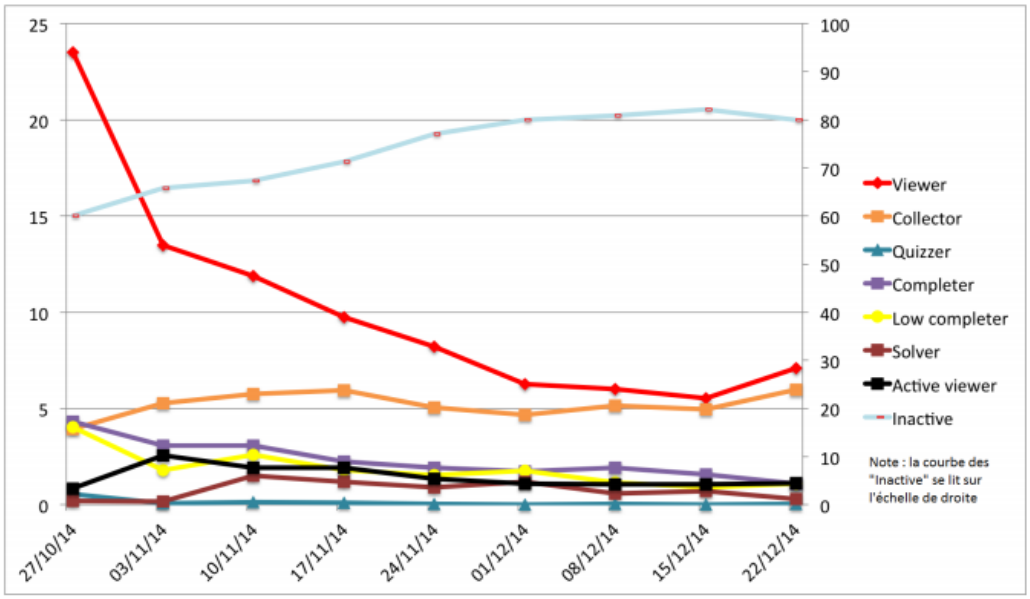


FIGURE 9 – Proportion d'états par semaine (%)

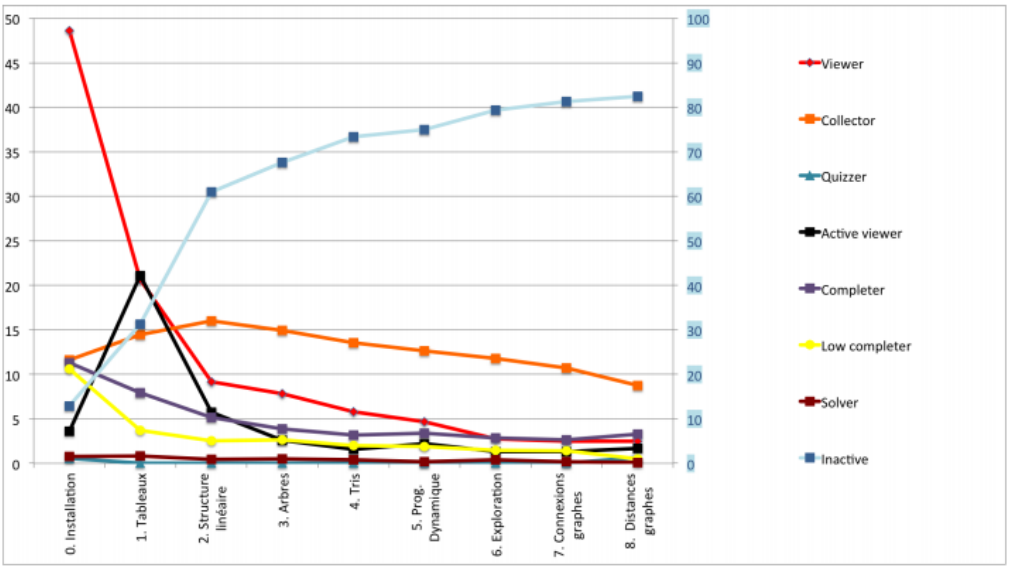


FIGURE 13 – Proportion d'états par section (%)

Scripts R : Disponibles sur demande (très long)

Points forts des analyses	**Points faibles des analyses**
Approche efficace pour capturer la dynamique d'un MOOC	Résultats très dépendants de la manière de construire les variables. Exemple : forte influence de la manière de prendre en compte de doublons, etc

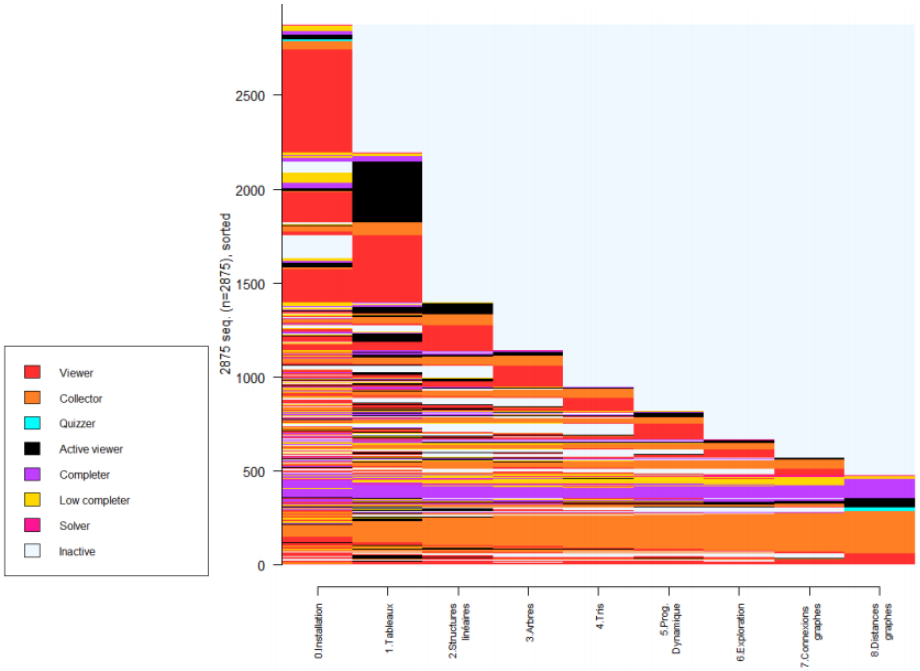
On fait ensuite les analyses correspondant au scénario 4 : analyse de parcours d'apprenants au sein de la plate-forme. L'analyse de séquence sous R. Cf. ci-dessous une analyse modulaire.

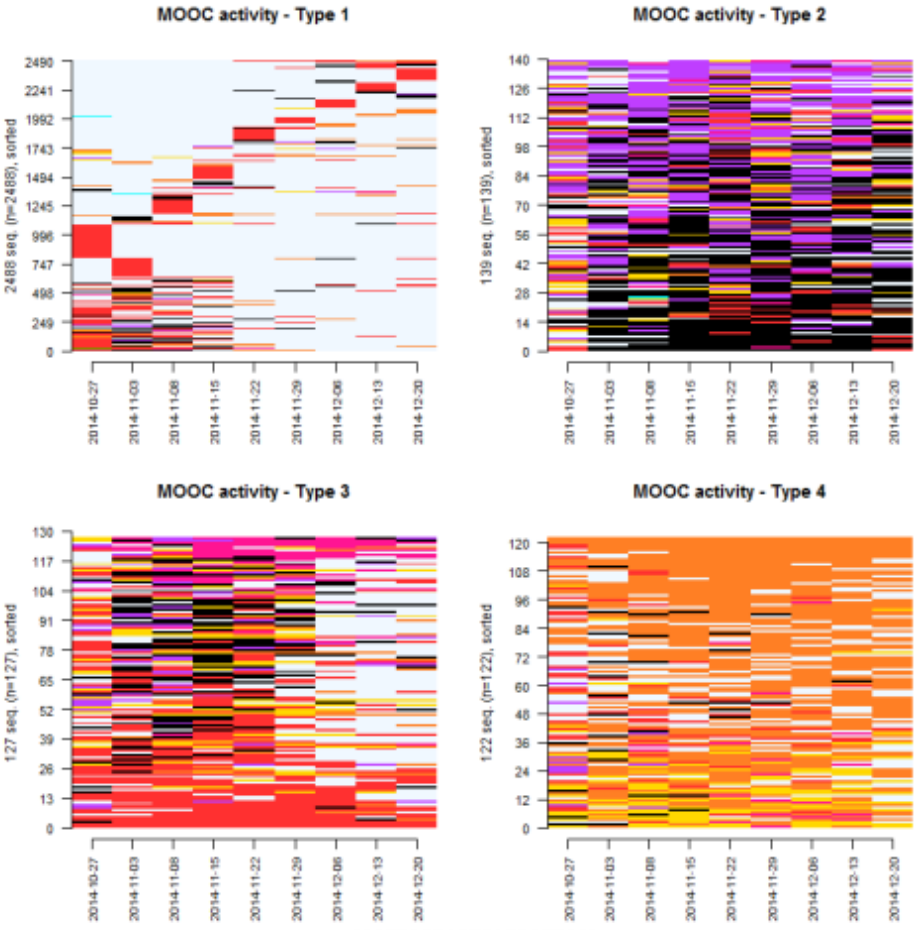
Analyse Scénario 4 : Parcours de participants au sein de la plate-forme

Liste des méthodes mise en œuvre : Construction de variables nouvelles (classification de comportements), analyse de séquence

Mode opératoire technique, logiciels utilisés : R

Résultats : Ce graphique permet de visualiser l'évolution d'état pour chacun des individus ayant été actif au moins une fois durant les neuf semaines de cours. Chaque ligne correspond à un participant dont nous pouvons suivre l'évolution de la première semaine à gauche à la dernière à droite. Nous constatons visuellement que l'activité constante réside dans les catégories Viewer , Collector et Active Viewer (respectivement en rouge, orange et noir).
Cependant, le nombre d'individus actifs toutes les semaines est faible. Chaque semaine semble apporter son lot de nouveaux inscrits qui vont majoritairement regarder ou télécharger les vidéos, voir faire les quiz. Enfin, pour valider notre classification, nous appliquons un algorithme de clustering prenant en compte la probabilité de passage d'un état à un autre sur deux semaines consécutives. L'algorithme détermine que deux découpages sont pertinents : le premier en quatre classes, le second en six. Étudions ces différentes répartitions. La classification en quatre classes nous permet d'avoir une première idée de division. Le type 1 correspond aux individus parcourant peu le cours : ils viennent rarement plus de la moitié du cours et effectuent principalement du visionnage de vidéos. Les types 2 et 3 représentent les individus les plus actifs du MOOC, regardant les vidéos mais répondant aussi aux quiz et rendant les devoirs. Cependant, le type 2 réunit clairement les Completers avec les Active Viewers, alors que les individus du type 3 sont un peu moins actifs (Viewers et Solvers principalement). Enfin, le type 4 regroupe principalement ceux qui vont venir chaque semaine pour télécharger les vidéos : les Collectors.





Script R : Disponible sur demande (vraiment très long)

Points forts des analyses	**Points faibles des analyses**
Permet de réaliser des classifications sur des trajectoires et non pas sur des comportements “moyens”. Analyse beaucoup plus fine de ce qui se passe au sein du MOOC	Il y a une part d’arbitraire dans la manière de construire la matrice de distances, i.e., la façon de pondérer les différentes actions et de décider quelles sont les plus similaires les unes des autres.

Description des Itérations

Pourquoi le processus d’analyse a été reproduit ?

Points forts des itérations	Points faibles des itérations