



**Thématique :**

**EVADATA**

**DEPP**



**CONTACT POUR LE CHALLENGE**

**Nom, Prénom : ROCHER, Thierry**

---

**Fonction : Adjoint au sous-directeur des évaluations et de la performance scolaire, DEPP**

---

**Numéro de téléphone : 060744044**

**RESUME DU CHALLENGE**

Le challenge consiste à établir une méthodologie permettant d'exploiter les traces numériques des élèves recueillies dans des situations d'évaluations interactives (*process data*). Il s'agit de définir un cadre du processus de traitement et d'analyse de ces données, qui soit pertinent d'un point de vue statistique et pédagogique et qui pourra être transférable à de nombreuses situations d'évaluations, qu'elles soient à visée diagnostique ou formative.

# Sommaire

1. CONTEXTE ET OBJECTIFS DU CHALLENGE .....	3
2. PRESENTATION DU SPONSOR ET EXPRESSION DU BESOIN.....	3
3. ÉTAT DE L'ART ET CHALLENGE A RESOUDRE .....	3
4. RESULTATS ATTENDUS.....	3
5. ENGAGEMENT DU SPONSOR, ÉQUIPE MOBILISEE ET MODALITES OPERATIONNELLES .....	3
6. RESSOURCES ET/OU JEUX DE DONNEES PROPOSES PAR LE SPONSOR.....	4
7. DEROULEMENT DU CHALLENGE - METHODOLOGIE.....	4
8. CRITERES DE SELECTION .....	4
9. RESTITUTION DU CHALLENGE .....	4
10. PERSPECTIVES ET RETOMBEES POSSIBLES DU CHALLENGE POUR LE LAUREAT .....	4

## 1. CONTEXTE ET OBJECTIFS DU CHALLENGE

Depuis 2015, les évaluations menées par la DEPP (service statistique du ministère de l'éducation nationale) ont opéré une transition vers le tout numérique. Ceci a donné lieu au développement de situations interactives qui permettent de recueillir des démarches de résolution de tâches et non plus seulement des réponses à des questions.

Ainsi, l'objectif principal est de pouvoir rendre compte des compétences des élèves de manière beaucoup plus fine et également de manière à ce que les enseignants puissent agir efficacement d'un point de vue pédagogique, en fonction de profils précis.

Le challenge a pour but de définir une méthodologie d'analyse des données recueillies dans le cadre de ces items interactifs (*process data*) qui pourra être transférable à de nombreuses situations d'évaluations, qu'elles soient à visée diagnostique ou formative.

Ce challenge s'inscrit dans les problématiques plus générales de l'application des méthodologies du domaine des Big Data au champ de l'éducation, ici plus particulièrement à celui de l'évaluation des élèves.

## 2. PRESENTATION DU SPONSOR ET EXPRESSION DU BESOIN

La **Direction de l'évaluation, de la prospective et de la performance (DEPP)** est à la fois une direction du ministère en charge de l'Éducation nationale et le service statistique ministériel en charge des statistiques sur l'éducation : <https://www.education.gouv.fr/direction-de-l-evaluation-de-la-prospective-et-de-la-performance-depp-12389>

Concernant les **évaluations standardisées des élèves**, la DEPP conduit de nombreux programmes :

- évaluation de l'ensemble des élèves de CP, CE1, 6<sup>ème</sup>, 2<sup>nde</sup>, JDC
- enquêtes nationales sur de larges échantillons (CEDRE, Socle, Panels, Évaluation de politiques publiques)
- enquêtes internationales sur de larges échantillons (PISA, PIRLS, TIMSS, ICILS)

L'ensemble de ces programmes d'évaluations standardisées concernent au total environ 4 millions d'élèves chaque année. L'équipe chargée du développement de ces programmes est composée d'environ 50 personnes, dont 15 statisticiens/psychométriciens.

Depuis 2015, ces évaluations ont opéré une **transition vers le tout numérique**. Cette transformation pose de nouveaux défis mais fait naître de nouvelles opportunités, notamment en matière d'**automatisation de corrections**.

Dans ce cadre, la DEPP a développé de nombreuses situations innovantes d'évaluations, aussi appelées TEI (*Technology-Enhanced Items*). Ces **items interactifs** permettent de recueillir des données sur la démarche de l'élève et non plus seulement sur sa simple bonne ou mauvaise réponse. Ces données, aussi appelées **process data**, sont nombreuses et semi-structurées.

Des premières analyses encourageantes de ces données, en mathématiques, ont permis de conclure qu'il était possible d'établir des **profils cognitifs** des élèves, appuyés sur une théorie didactique et qui permettent de se renseigner sur le niveau mais aussi sur le type de compréhension des apprentissages par les élèves. L'objectif est de pouvoir rendre compte des compétences des élèves de manière beaucoup plus fine mais également de fournir des informations en fonction de profils précis des élèves de manière à ce que les enseignants puissent agir efficacement d'un point de vue pédagogique.

### 3. ÉTAT DE L'ART ET CHALLENGE A RESOUDRE

Encore trop peu de travaux existent dans le domaine de l'analyse des *process data*. Et parmi ces travaux, la prise en compte de fondements théoriques concernant les apprentissages est relativement peu fréquente. Or, cet aspect est essentiel si l'on souhaite que les analyses statistiques aboutissent à identifier des profils pertinents d'un point de vue pédagogique.

Ainsi, la DEPP a conduit de premiers travaux sur ces questions, qui ont donné lieu à une publication : Salles et al. (2020)<sup>1</sup>. L'approche retenue est de situer l'analyse au croisement de la théorie didactique et de l'apport des Data Science.

Les enjeux renvoient maintenant à l'« industrialisation » des processus d'analyse des *process data* issus de ces items. D'une part, le nombre de ces items ne cesse de croître (environ 50 à ce jour). D'autre part, le nombre d'élèves concernés est amené à considérablement augmenter, passant de quelques milliers d'élèves d'échantillons représentatifs à des générations entières (800 000 élèves par niveau scolaire). Par conséquent, il est primordial de pouvoir établir une méthodologie robuste permettant de traiter efficacement des données et d'en rendre compte de façon adaptée pour les élèves et les enseignants.

Le challenge est de **construire une méthodologie permettant d'industrialiser cette approche**, grâce aux méthodologies employées dans le champ des Big Data. En effet, ces items interactifs se multiplient et ont vocation à être passés par le plus grand nombre d'élèves dans le cadre d'**évaluations nationales** (une cohorte représente 800 000 élèves). En outre, dès lors qu'ils sont validés par ces approches au croisement de la didactique et de la statistique, ils constituent des outils potentiellement très précieux, **pour les enseignants et leur action pédagogique** au quotidien.

Trois aspects sont à considérer :

- Industrialisation des processus de nettoyage, de structuration et d'analyse des données
- Définition d'un cadre méthodologique de conception et d'analyse des *process data*, s'appuyant sur des technologies Big Data et en lien avec des théories didactiques approtées par la DEPP, dans le domaine des mathématiques
- Application à des items existants

Un exemple d'item interactif de mathématiques est donné ci-dessous. Pour résoudre la tâche, l'élève utilise un tableur-grapheur. Il peut également utiliser des outils (crayon, calculatrice, etc.). Toutes les actions de l'élève sont enregistrées et donnent lieu à la production de données massives et semi-structurées.

---

<sup>1</sup> <https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-020-00085-y>

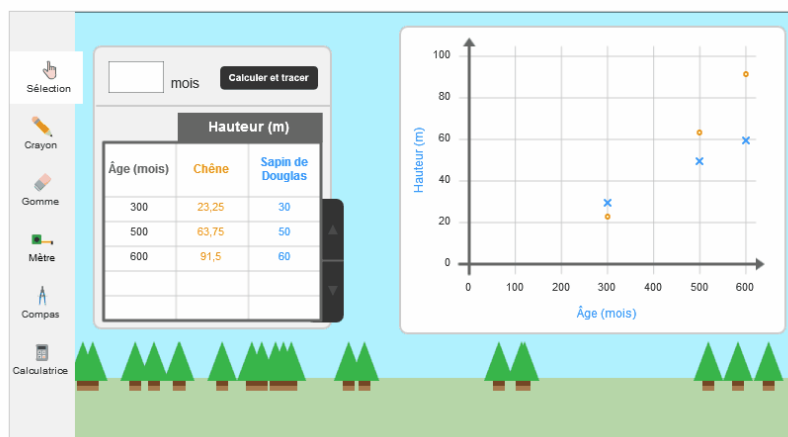
Deux graines d'arbres sont plantées au même moment : un chêne et un sapin de Douglas.

En entrant dans la première colonne, l'âge (en mois) des arbres, on obtient leur hauteur (en mètre) dans les deuxième et troisième colonnes.

Les points correspondants s'affichent sur le graphique : en orange le chêne, en bleu le sapin.

**A quel âge (autre que 0 mois) ont-ils la même hauteur ?**

L'âge est de  mois.



L'ambition du projet est de pouvoir généraliser l'emploi de ces situations interactives auprès de millions d'élèves, lors des évaluations nationales mais également à terme pour une mise à disposition d'outils d'évaluations interactifs à l'ensemble des enseignants, dans le cadre d'évaluations formatives.

Pour cela, le projet se déroulera selon deux axes, dénommés et décrits ci-dessous:

- **Data engineering :**
  - garantir la qualité des données, à travers une procédure de nettoyage, de consolidation et de stockage ou ETL (*Extract Transform and Load*)
  - assurer la maintenance de l'accessibilité aux données
  - assurer l'automatisation du pipeline de données
- **Data analytics :**
  - assurer l'analyse des données, à travers la construction de variables secondaires (*feature engineering*) et la modélisation (*machine learning, deep learning, ..*)
  - garantir la reproductibilité de l'étude, et assurer la documentation des scripts réalisés
  - garantir une maîtrise du champ étudié suffisant pour passer de la théorie à son application pratique à travers les données

## 4. RESULTATS ATTENDUS

Le livrable attendu est un cadre méthodologique général permettant une industrialisation des processus d'analyse des *process data*, ainsi que les codes et la documentation associée ayant servi au traitement des données des items interactifs de mathématiques.

**Data engineering :** méthodologie et codes employés pour industrialiser le nettoyage, processus, organisation, stockage (ETL)

- préconisation d'optimisation des log data issus des TEIs (format, structure, volume...)
- scripts paramétrables de transformation des données : passer du JSON à des données structurées

- *pipeline* : proposer un *workflow* automatisé complet, du stockage à l'extraction des données structurées, voire à l'output du modèle dans le cas où l'item a déjà été analysé

*Data analytics* : cadre méthodologique de recommandations de modèles (IA, *Machine Learning*, NLP...), choix de modélisations en fonction de critères

- quelle méthode selon quelles hypothèses didactiques et modèles cognitifs d'élèves définis par la DEPP *a priori* ? (neuf familles didactiques ont été identifiées parmi l'ensemble de 50 items)
- *feature engineering* / *feature selection*
- si nécessaire, préconisation de modifications des *log data* capturées par l'item

## 5. ENGAGEMENT DU SPONSOR, ÉQUIPE MOBILISÉE ET MODALITÉS OPERATIONNELLES

La DEPP investit beaucoup dans des solutions technologiques innovantes en matière d'évaluation (jusqu'à un engagement fort au niveau international à travers l'association FLIP+). Dans ce secteur, la correction automatisée de production d'élèves occupe une place importante et un secteur stratégique pour la DEPP en tant que service statistique.

L'équipe mobilisée pour ce projet sera composée de :

- Une cheffe de projets, en charge des programmes innovants
- Une *data scientist*
- Un professeur de mathématiques, spécialiste de l'évaluation numérique
- Un psychométricien

En outre, le service informatique de la DEPP assistera l'équipe sur les aspects liés aux données.

## 6. RESSOURCES ET/OU JEUX DE DONNÉES PROPOSÉS PAR LE SPONSOR

Les données test dont dispose la DEPP sont constituées de 50 items interactifs, avec pour chacun environ les réponses de 3000 élèves. A terme, le volume de données est voué à être beaucoup plus important quand ces items seront déployés sur des cohortes entières (jusqu'à 800 000 élèves par cohorte). On estime qu'un item représentera alors 200 Go. Il est donc important que les procédures puissent être mises à l'échelle (scalables) sur un environnement distribué.

Les données sont anonymisées, au format JSON.

Il est possible d'organiser un travail sur le site de la DEPP sur un environnement Spark.

## 7. DÉROULEMENT DU CHALLENGE - MÉTHODOLOGIE

Les passages surlignés en vert correspondent à l'axe *Data Engineering* et ceux surlignés en jaune correspondent à l'axe *Data Analytics*.

T0 : Mise à disposition des données, des scripts et de la documentation associée.

T0 + 1 mois : Point de régulation. *Lancement du nettoyage et traitement des données existantes.*

T0 + 6 mois : *Résultats et identification des points forts et points faibles des données existantes en vue de la standardisation et industrialisation des processus. Proposition de nouveaux modèles et standards pour la structuration des *log data* issues des TEIs. Scripts d'automatisation du nettoyage et traitement des données existantes (ETL).*

T0 + 18 mois : Cadre méthodologique de conception et d'analyse des *process data* (service obligatoire), *workflow* automatisé allant du nettoyage et traitement (ETL) au scoring automatique de données (service possible).

Des réunions de suivi auront lieu toutes les 2 semaines entre le lauréat et la DEPP.

Deux processus selon la méthode Agile :

#### **Data engineering :**

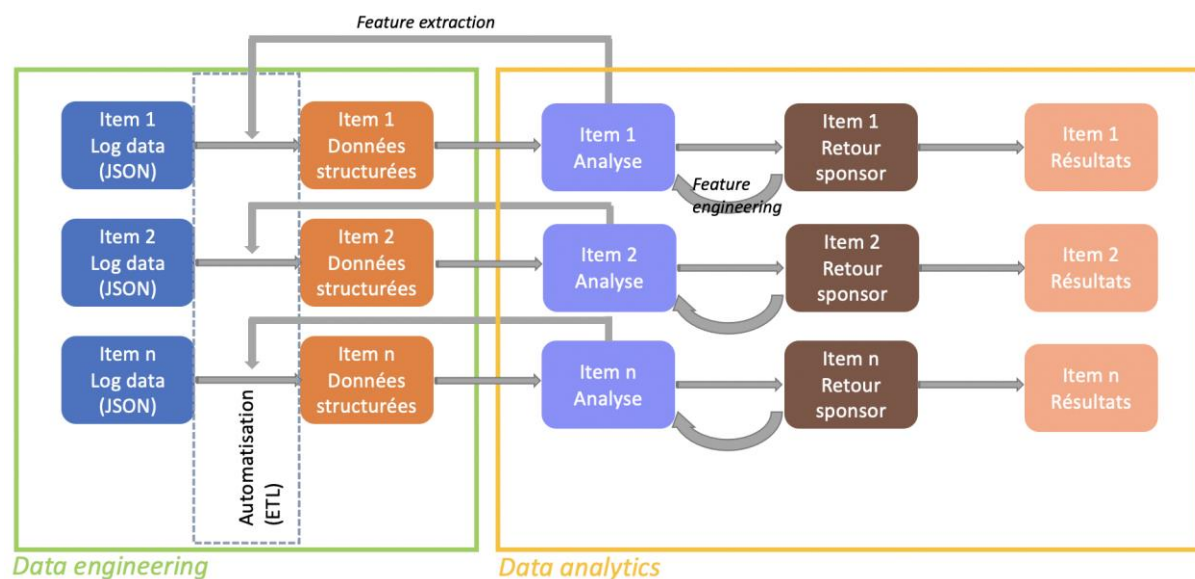
Les *process data* seront extraites des données JSON, et transformées afin d'obtenir des données structurées, utilisables pour l'analyse. Cette étape pourra être répétée suite au *feedback* du *data analyst*, si nécessaire.

En parallèle, les scripts d'extraction des données seront construits de façon à pouvoir systématiser cette étape pour l'ensemble des items, à travers un code paramétrable. Les différents items serviront de points de versioning dans ce but. Cette automatisation sera suivie d'une analyse architecturale SI (système d'information), dans l'optique d'une intégration directe dans le *workflow* ETL du sponsor.

#### **Data analytics :**

Les données structurées seront analysées, afin d'en obtenir une interprétation en phase avec les hypothèses didactiques. Ces analyses feront l'objet d'un point d'étape avec le sponsor, afin de réorienter les travaux si besoin est. Cette étape aboutira à la rédaction d'un rapport d'étude. L'étude de chaque item est estimée à 2 mois.

En parallèle, les différents items étudiés permettront de faire émerger un cadre méthodologique approprié à la généralisation des études aux autres items reposant sur les mêmes hypothèses didactiques. Ce cadre méthodologique s'enrichira au fur et à mesure de l'avancée dans les travaux.





## 8. CRITERES DE SELECTION

### 1- Pertinence de la proposition de réponse (pondération : 30)

- **Innovation** : caractère innovant des travaux réalisés/proposés par le lauréat (pondération : 10)
- **Cohérence et réalisme** de la proposition du calendrier des travaux (pondération : 10)
- **Prise en compte de la problématique** posée par le sponsor : la solution proposée par le lauréat répond au challenge du sponsor et à ses critères spécifiques (pondération : 10)

#### Critères de sélection spécifiques :

- Méthodologie proposée (Agile), description de l'organisation générale des prestations et processus.
- L'utilisation de langages et de technologies : Python, R, JSON, Hadoop, HDFS, Spark, Hive, HBASE, MongoDB
- Facilité d'accès aux solutions et qualité de la documentation dans l'objectif de futures évolutions et utilisations par le sponsor.
- Utilisation de technologies *open source*.

### 2- Compétences et engagements (pondération : 30)

- **Niveau d'engagement du candidat** : identification des personnes clés (pondération : 10) (nombre et qualité, fourniture des CV) et quotité de disponibilité dédiée au challenge, mise à disposition d'un interlocuteur identifié et unique pour favoriser les interactions avec le sponsor

Deux types de profils sont attendus :

#### Profil data engineer

- ☐ Connaissance approfondie de SQL et autres langages de base de données
  - ☐ spécifiquement, la maîtrise des bases NoSQL est nécessaire
- ☐ Stockage de données et outils ETL
- ☐ Connaissance architectures Big Data
  - ☐ spécifiquement, la maîtrise experte de Hadoop et Spark est demandée
- ☐ Usage de langages maîtrisés en interne par le sponsor
  - ☐ R, Python, JavaScript sont particulièrement attendus

#### Profil data analyst / data scientist :

- ☐ Usage expert des environnements Big Data
  - ☐ Apache Spark est particulièrement attendu
- ☐ Maîtrise de la palette d'algorithmes de modélisation, prédiction et catégorisation
  - ☐ machine learning, deep learning
  - ☐ connaissance des algorithmes de NLP serait appréciée
- ☐ Usage de langages maîtrisés en interne par le sponsor
  - ☐ R, Python, SQL
- ☐ Qualités pédagogiques, capacité de communication avec les experts du champ étudié
  - ☐ spécifiquement, appétence pour le champ de la didactique et des sciences de l'éducation

- **Maturité de la réflexion** (pondération : 10) et des connaissances du candidat sur le monde de l'éducation et ses enjeux ;

Cadres règlementaires et juridiques (code de l'éducation, RGPD, handiaccessibilité...) et cadre de confiance.

Connaissance des programmes scolaires.

- **Ressources en démonstration** (pondération : 10) : indication de réalisations numériques déjà produites (dans l'éducation ou autre, produits industriels ou prototypes).



### 3 - Prix. (pondération : 40)

## 9. RESTITUTION DU CHALLENGE

Services obligatoires :

- Identification des points forts et points faibles des données existantes en vue de la standardisation et industrialisation des processus de traitement et d'analyse. Proposition de nouveaux modèles et standards pour la structuration des log data issues des TEIs. Scripts d'automatisation du nettoyage et traitement des données existantes (ETL)
- Cadre méthodologique de conception et d'analyse des *process data*.

Services possibles :

- Conception d'un *workflow* automatisé et industrialisable de nettoyage et traitement de données (ETL) : des log data (JSON) aux données structurées.
- Conception d'un *workflow* d'analyse des données item par item : outil de *scoring* automatique

Tous les codes ainsi que la documentation associée seront demandés au lauréat à l'issue du challenge.

Les solutions développées doivent utiliser des technologies open source.

La propriété intellectuelle des solutions développées doit appartenir au sponsor afin de garantir l'autonomie du sponsor au regard de leur implémentation sur des données futures.

## 10. PERSPECTIVES ET RETOMBÉES POSSIBLES DU CHALLENGE POUR LE LAURÉAT

Le lauréat s'inscrit dans un axe parmi les plus innovants au croisement du numérique et de l'éducation, de la recherche et de l'action pédagogique.

Le lauréat aura accès à un corpus unique de données (traces numériques de milliers d'élèves dans des situations d'évaluation complexes). Le lauréat bénéficiera en outre de l'expertise de la DEPP en matière d'évaluation (didactique, psychométrie en particulier), sur laquelle il pourra capitaliser pour d'autres activités, dans un domaine en plein essor.

L'expérience sera valorisable au niveau national – à travers l'élaboration de solutions logicielles par exemple – ainsi qu'au niveau international, à travers les services étendus à des institutions partenaires de la DEPP à l'étranger (notamment via l'association FLIP+ : <https://flip-plus.org/>)

Pour le lauréat, cette expérience lui permettra de se positionner de façon optimale dans le secteur des Big Data en éducation, suivant en cela l'essor du *learning analytics*.

En outre, la DEPP a une gamme de publications très variées : notes, articles, dossiers, rapports. Par ailleurs, les équipes de la DEPP interviennent fréquemment dans des conférences scientifiques et dans des conférences réunissant des professionnels de l'évaluation et publient régulièrement dans des revues scientifiques. En outre, la DEPP est très active dans les réseaux internationaux sur l'évaluation.

Tous ces outils seront mobilisés pour valoriser les travaux réalisés avec le lauréat.