# Data science : Savoir collecter, décrypter, analyser et prédire à partir de mégadonnées

*CATEGORIE : C*

## Vue d'ensemble

Domaine(s) d'activité professionnel dans lequel(s) est utilisé la certification :

Transverse :

- **Bancaire**
- **Assurance**
- **Energie**
- **Transport**
- **Industrie**
- **Conseil**
- **Agroalimentaire**
- **Recherche**
- **Médical**

Les compétences en data science s'exercent dans de nombreux domaines à enjeux économiques et sociétaux (marketing ; industrie ; agroalimentaire ; conseil ; recherche ; médical ; emploi....) et dans toute activité utilisant un volume important de données

Code(s) NAF : **—**
Code(s) NSF : **326**, **114**
Code(s) ROME : **M1806**, **M1805**, **M1803**, **M1801**
Formacode : **31054**

Date de création de la certification : **04/11/2014**
Mots clés : **Modelisation**, **algorithmes**, **Big data**, **Data science**

## Identification

Identifiant : **2862**
Version du : **06/06/2017**

## Références

Consensus, reconnaissance ou recommandation :
Formalisé :

- **Requête gouvernementale formalisée dans les 34 plans de la nouvelle France industrielle (p.51)**

Non formalisé :

- **Les technologies de traitement des données massives (big data) sont depuis peu disponibles pour toutes les entreprises, que ce soit sous forme de services Cloud, de plates-formes open source ou de solutions proposées par les éditeurs de logiciels propriétaires. Mais les compétences pour valoriser cette manne de données manquent. La demande de data scientists est croissante et l'écart entre la demande et l'offre se creuse tant au niveau national qu'international. Compte tenu de l'extrême spécialisation nécessaire pour exercer la profession de data scientist, les opportunités d'embauche sont**

**extrêmement nombreuses et largement supérieures à la quantité de profils qualifiés. Ainsi, face aux difficultés des entreprises pour recruter des spécialistes rompus aux avancées techniques, statistiques et informatiques, dans l'exploitation des données massives, plusieurs écoles de management et d'ingénieurs ainsi que des universités ont développé des formations pour répondre à la demande importante des établissements publics et privés en data scientists**

## Descriptif

### *Objectifs de l'habilitation/certification*

Le certificat doit permettre de connaitre les outils algorithmiques et techniques liés à la data science et de piloter un projet en data science.  L'objectif de la certification est de fournir l'expertise nécessaire pour la gestion et l'analyse pointues des données massives. Le data scientist certifié pourra alors déterminer les indicateurs permettant de mettre en place une stratégie répondant à une problématique de l'entreprise.

### *Lien avec les certifications professionnelles ou les CQP enregistrés au RNCP*

- Aucun

### *Descriptif général des compétences constituant la certification*

Le Data Scientist est un expert de l'analyse de données massives ("big data"). Il a de bonnes connaissance de la gestion des bases de données,  il récupère à partir de sources de données multiples et dispersées, structurées ou non, appartenant à l'entreprise ou en open data, les données dont il a besoin pour traiter le problème posé. Ses connaissances métier lui permettent de bien cibler la méthodologie adaptée au problème. Il est spécialisé en statistique, informatique et connait parfaitement le secteur ou la fonction d'application des données analysées.

A ce titre, il est chargé de :

1) poser de façon claire et précise le problème à résoudre

2) récupérer les données permettant de résoudre ce problème. Les données peuvent venir des différents entrepôts de données de l'entreprise (les types de bases de données ne sont pas un problème

## Public visé par la certification

- Tout public disposant de compétences en statistique et en informatique. Public travaillant de manière préférentielle à un poste d'analyse de données (Data Analyst ou similaire) ou dans un environnement datacenter (au contact de grandes structures de données).

pour lui). En fonction de la problématique, il peut aussi récupérer des données externes à l'entreprise (opendata, site web, Api dédiées, données INSEE...).

3) mettre en forme toutes les données de manière optimale en fonction des algorithmes utilisés

4) choisir les différentes méthodes potentielles pour apporter une solution. comparer les différentes méthodes en utilisant des données d'apprentissage (estimer les paramètres des méthodes) et des données de test (pour effectuer des prévisions) et avoir des indicateurs fiables de comparaison de méthodes en utilisant souvent des logiciels de traitements comme R ou Python.

5) rédiger des codes propres documentés et réutilisables par ses collègues dans un souci de répétabilité des analyses.

6) présenter de manière convaincante au donneur d'ordre les résultats lorsque la meilleure méthode est sélectionnée

### *Modalités générales*

L'inscription à la formation fait l'objet d'une sélection sur dossier. Chaque dossier comprend un cv détaillé une lettre de motivation et un test de positionnement relatif au logiciel R effectué par le candidat.

La formation a lieu en présentiel et comprend 18 jours (ou 15 jours pour les sessions intensives) d'enseignements théorique et pratique. Les stagiaires mènent également un projet cas d'école pendant leur formation. Ils disposent d'un accès à la plateforme Teralab pendant toute la durée de la formation.

Trois types de session sont proposés aux candidats:

Une session intensive de 15 jours (120 heures) soit 3 fois 5 jours
Une session de 18 jours (126 heures) soit 3 jours par mois.
Une session de 14 jours (112 heures) en partenariat avec l'université de Berkeley (intervention de deux formateurs de l'université de Berkeley)
Les sessions se déroulent sur une durée maximale de 6 mois.

La validation du certificat est soumise à la réussite de l'examen, à la présentation d'un projet et à la présence obligatoire aux cours.

Un examen de rattrapage est proposé aux stagiaires échouant à la certification.

### *Liens avec le développement durable*

Aucun

## Valeur ajoutée pour la mobilité professionnelle et l'emploi

### *Pour l'individu*

Le certificat permet aux stagiaires d'acquérir des compétences multiples à savoir la maîtrise des techniques statistiques et de l'analyse des données, une connaissance des technologies et des outils informatiques des bases de données et éventuellement un savoir-faire métier dans le domaine d'application des données étudiées.

Le certifié possède des compétences transversales et est en capacité de gérer un projet en data

### *Pour l'entité utilisatrice*

La certification valide les compétences d'une personne à gérer un projet de data science. Le certifié est capable de déterminer le problème à résoudre dans l'entreprise, de le formuler de manière mathématique, algorithmique puis de combiner les données nécessaires entre elles pour y répondre.

Le Data Scientist joue un grand rôle dans la création d'indicateurs précieux à tous les niveaux de

science.

En relation avec l'émergence du big data, ces spécialistes sont recrutés dans tous les secteurs : transports, énergie, assurance, banque, industrie, commerce, santé ..... Ils peuvent avoir un large choix de domaines d'application.

l'entreprise. Il s'investit dans l'amélioration de l'activité globale grâce à la précision de l'analyse et en mettant sur pied des modèles de prédiction. L'analyse et l'exploitation des données massives permettent aux entreprises de façonner leur marché,d'accroitre leur efficacité et de rester compétitives.

## Evaluation / certification

### *Pré-requis*

Diplôme (ou niveau) Bac + 4 ou 5 ou expérience professionnelle équivalente.

Prérequis : maitriser les méthodes de régression et le logiciel R

### *Compétences évaluées*

Le titulaire est capable de :

1) Requêter dans le système informatique de l'entreprise pour récupérer les données pertinentes

2) De récupérer les données adaptées à l'extérieur de l'entreprise en utilisant, si le besoin s'en fait sentir, des techniques d'acquisition automatisées de données externes à l'entreprises

3) D'écrire un programme pour analyser automatiquement de larges volumes de données et en extraire les informations pertinentes

4/ Mettre en œuvre les fonctions de R adaptées pour traiter et visualiser un jeu de données.

5) D'utiliser les algorithmes de machine learning implémentés dans R ou Python

6) De comparer différents algorithmes en utilisant des techniques d'apprentissage/validation

7) Selon le besoin, utiliser un serveur de calculs pour accélérer les temps de traitement.

8) Utiliser, pour un volume de données important, la bonne architecture de répartition des données.

9/ Evaluer la qualité du modèle selon sa finalité (prédictive ou explicative)

10) Présenter aux donneurs d'ordre son travail et ses choix. Utiliser avec intelligence des outils de visualisation ou créer un serveur web permettant de rejouer les méthodes.

11) Proposer aux services informatiques les modèles à mettre en production

### *Niveaux délivrés le cas échéant (hors nomenclature des niveaux de formation de 1969)*

Sans objet

## Centre(s) de passage/certification

- Ensae Ensai Formation Continue (le Cepe) 60 rue Etienne Dolet 92240 Malaloff

La validité est Permanente

**Possibilité de certification partielle :** non
Matérialisation officielle de la certification :
Certificat de compétences délivré par le GENES ou Certificat de compétences délivré par le GENES et l'Université de Berkeley

## Plus d'informations

### *Statistiques*

11 promotions et une promotion spécifique à la Banque de France soit un total de 143 stagiaires inscrits depuis la création du certificat

96 certifiés et 40 en cours de certification

### *Autres sources d'information*

http://www.lecepe.fr/certificats/data-scientist/

http://www.actu-cci.com/emploi/13695-l-ensae-ensai-formation-continue-le-choix-de-la-performance