



FORSIDES

EXPERTS – FLEXIBLES – PROCHES – ENGAGÉS

GROUPE DE TRAVAIL - DATA SCIENCE

19 JUILLET 2017

ORDRE DU JOUR

- Le GT Data Science
- La Data Science, à quoi ça sert ?
- Classer les problèmes de Data Science en catégories
- Pourquoi Python ?

LE GT DATA SCIENCE

OBJECTIFS DU GT DATA SCIENCE

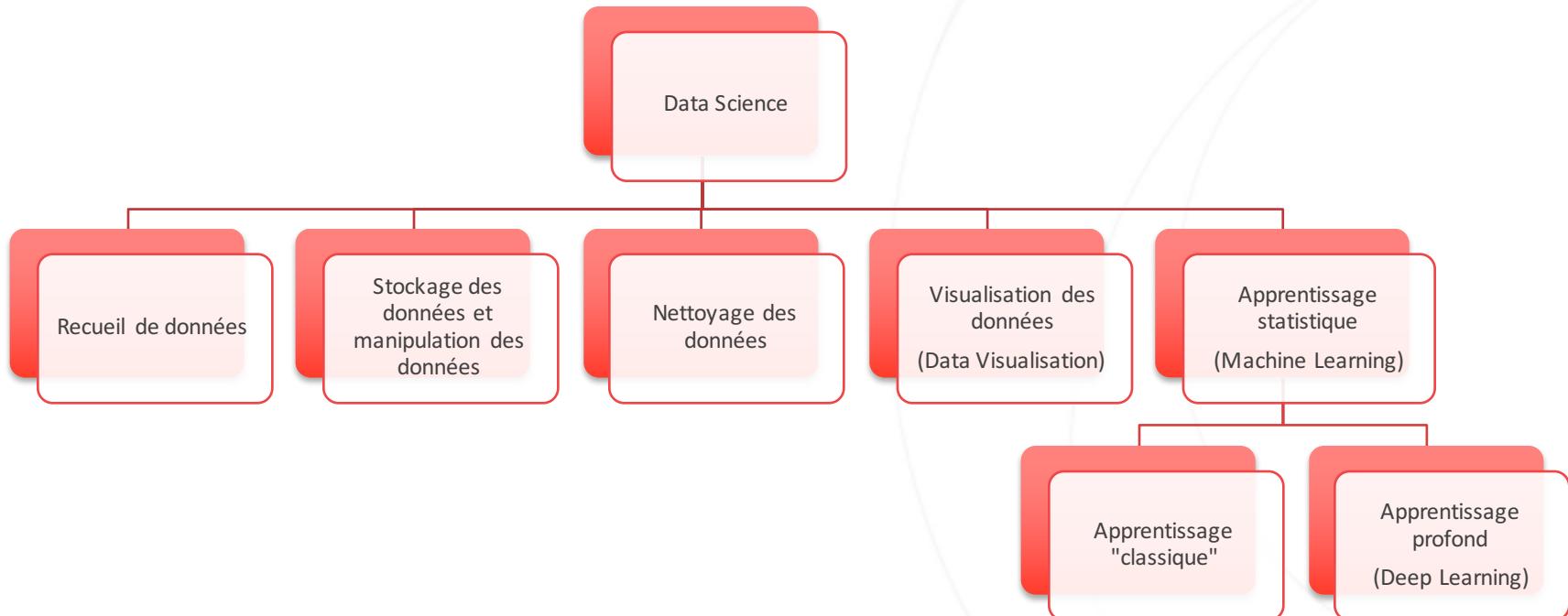
- Connaître les principales techniques de data science, et savoir les utiliser à travers le langage Python
- Réfléchir à l'impact de la data science sur la société, et en particulier sur le secteur de l'assurance
- Organiser des rendez-vous réguliers pour échanger sur la data science

ORGANISATION

- Une réunion tous les 15 jours environ
- Partage de codes/exemples pour faire les premiers pas avec Python
- Organisation de concours/challenge dans le domaine des data science, et interne à Forsides

LA DATA SCIENCE, À QUOI ÇA SERT ?

LES DIFFÉRENTES ÉTAPES DE LA DATA SCIENCE



- Prévoir, pour optimiser :
 - Publicité
 - Facebook, Google, Critéo, ...
 - Recommandation
 - Youtube, Deazer, Booking, Air France
 - Ciblage client
 - Optimisation de process : EDF (Prédition de consommation d'énergie, maintenance prédictive), Trading algorithmique, Prédiction de la probabilité de défaut (Crédit), estimation de valeur client
- Analyser, pour créer du savoir
 - Tests médicaux, Sciences sociales, Economie, Résultats d'une entreprise
- Système intelligent
 - Véhicule autonome, Robot



matching : Utilisateur <-> Contenu

APPRENTISSAGE « CLASSIQUE »

- Les données ont une structure « classique » : Base de données relationnelle

Observations

Variables

Total rows: 19 Total columns: 10

Rows 1-19

	Name	Sex	Age	Height	Weight
1	Joyce	F	11	51.3	50.5
2	Louise	F	12	56.3	77
3	Alice	F	13	56.5	84
4	James	M	12	57.3	83
5	Thomas	M	11	57.5	85
6	John	M	12	59	99.5
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84
10	Carol	F	14	62.8	102.5
11	Henry	M	14	63.5	102.5
12	Judy	F	14	64.3	90
13	Robert	M	12	64.8	128
14	Barbara	F	13	65.3	98
15	Mary	F	15	66.5	112
16	William	M	15	66.5	112
17	Ronald	M	15	67	133
18	Alfred	M	14	69	112.5
19	Philip	M	16	72	150

Valeurs

Identifiants

APPRENTISSAGE PROFOND

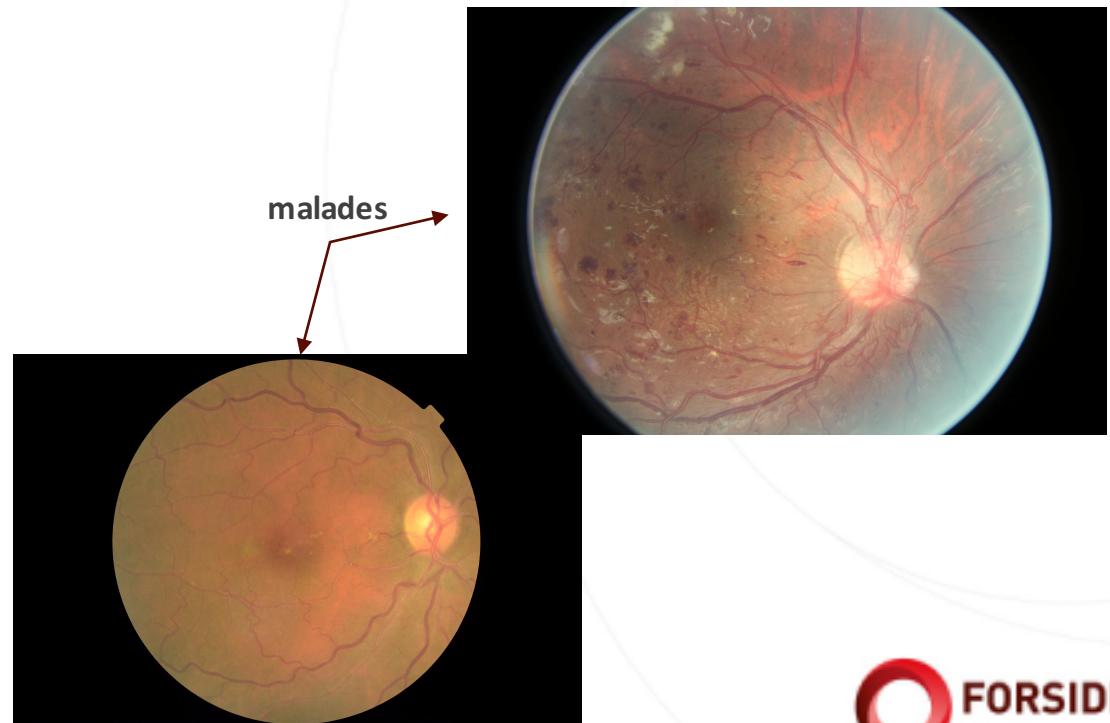
- Traitement du langage naturel (Natural Language Processing, NLP)
 - Facile : Combien y-a-t-il d'étoiles sur le drapeau européen ?
 - Difficile : La prochaine phrase est fausse. Vladimir a plus de trente ans. Est-il possible que Vladimir ait 25 ans ?
- Traitement d'image :
 - Description d'une image
 - Lecture de formulaires
 - Reconnaissance faciale
- Traitement du son
 - Reconnaître un style de musique, un artiste
 - Siri
- Traitement de données « non structurées »
(ou structurées mais pas sous forme de base de données relationnelles)



- Combien y-a-t-il de personnes sur cette photo ?
- De quelle couleur est le t-shirt de l'homme à droite ?

QUELQUES APPLICATIONS DE L'APPRENTISSAGE PROFOND

- Jouer à un jeu : casse brique, jeu de go
(vidéo casse brique : <https://www.youtube.com/watch?v=V1eYniJ0Rnk>)
- Véhicule à conduite autonome
- Identifier une pathologie sur une image médicale (kaggle rétinopathie : <https://www.kaggle.com/c/diabetic-retinopathy-detection>) :



- Tarification
 - Assurance de biens : auto, habitation
 - Assurance santé
- Ciblage de client
- Valeur client
- Provisionnement
- Modèles de durée
 - Tables de mortalité/maintien, durée de résiliation
- Open Data
 - Santé, auto
- Assurance automobile télématique
 - Pay as/how you drive
 - Italie : 15% du marché (4,5 M polices)
 - UK : 450K polices
 - France : Allianz, Axa Direct Assurance (YouDrive) -> peu de polices
- Objets connectés (santé, habitation, auto)
 - USA : John Hancock (modulation prime assurance)
 - France : AXA
- Lutte contre la fraude
 - Blanchiment d'argent/financement du terrorisme
 - Escroquerie à l'assurance

EXEMPLES DE PROBLÈMES STATISTIQUES ÉTUDIÉS EN ASSURANCE

1) Scoring de prospect :

- Comment construire un score permettant de prédire la valeur rapportée par un client (en fonctions des caractéristiques socio-professionnelles du client : propriétaire ou nom, CSP, etc.) ? Et ainsi optimiser un processus commercial.

2) Déterminer si le détenteur d'un contrat d'assurance vie va racheter son contrat avant 8 ans

- Construire un modèle en utilisant les informations (âge, sexe, historique des opérations sur le contrat (ou sur les comptes du même client), etc ...)

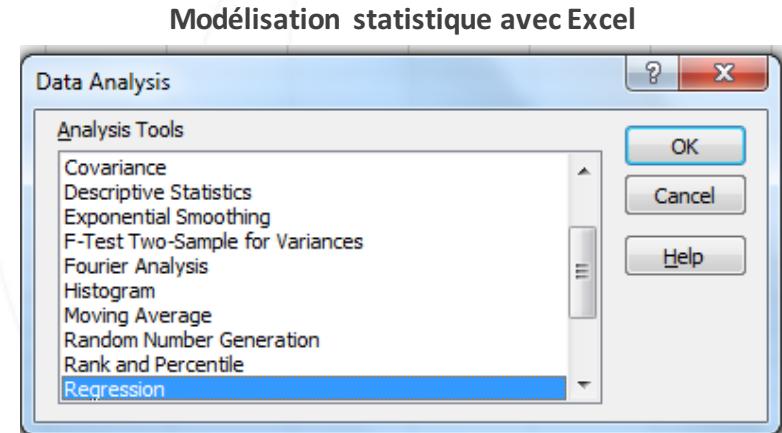
3) Classification des contrats d'un portefeuille d'assurance santé en groupes homogènes selon un ensemble de caractéristiques

- Âge, catégorie socio-professionnelle, lieu de résidence, etc...

4) Déterminer si un automobiliste a une conduite à risque à partir de relevés GPS de ses trajets au volant

LES AVANTAGES DE MAITRISER UN LANGAGE D'ANALYSE DE DONNÉES DANS LE QUOTIDIEN

- Traitement des données
 - Plus volumineuses qu'avec Excel
 - Opérations spécifiques : « merge » de bases de données, concaténation de base, ...
 - Automatiser des traitements que l'on aurait fait à la main sous Excel
 - Moins de risque d'erreur due à une fausse manipulation
 - Accès à des opérations plus complexes
- Analyser des données
 - Excel limité dans son offre de méthodes statistiques →
 - Python dispose de bibliothèques (Scikit Learn, Xgboost, ...) où sont programmées beaucoup de méthodes statistiques
- Construire des systèmes dont le moteur est codé en Python (ou R)



IMPORTANCE DES DONNÉES

- Nature de l'information :
 - Dans l'exemple 4), l'information sur les relevés GPS de trajets est beaucoup plus pertinente que l'information sur le sexe/âge etc... pour répondre au problème
- Qualité de l'information (Qualité des données) :
 - L'information doit être vraie (ou pour le moins contenir une faible proportion d'informations erronées) et structurée
- Recueil de l'information souvent prépondérant devant le modèle statistique utilisé
 - Le Big Data tend à fournir plus de données exploitables (exemples : données GPS, données objets connectés, données génétiques etc...)

CLASSEZ LES MÉTHODES DE DATA SCIENCE EN CATÉGORIES

APPRENTISSAGE SUPERVISÉ OU APPRENTISSAGE NON SUPERVISÉ

On classe les méthodes utilisées en Data Sciences en 2 catégories :

- Apprentissage supervisé :
 - **On explique une variable Y à partir de variables explicatives, notées X** (exemples : Rétinopathie, Scoring prospects, Rachat assurance vie)
 - Le plus souvent Y est une variable à une dimension
- Apprentissage non supervisé :
 - On cherche à classer les individus en classes homogènes à partir de caractéristiques X
 - **Aucune variable ne joue le rôle de variable à expliquer.** (exemple : Classification des contrats assurance santé)
 - (Plus généralement, on cherche à extraire de l'information de données sans qu'aucune variable ne joue le rôle de cible)

APPRENTISSAGE SUPERVISÉ – CLASSEMENT OU RÉGRESSION

- Il existe 2 types de problème en apprentissage supervisé :
 - La **régression** : on cherche à prédire une variable Y **quantitative** à partir de variables explicatives X (scoring, tarification)
 - Le **classement** : on cherche à prédire une variable Y **qualitative** à partir de variables explicatives X . Y peut compter 2 modalités (**classement binaire** : Rétinopathie, Rachat assurance vie) ou plus de 2 (**classement multiple**)

POURQUOI PYTHON ?

PYTHON

- L'un des 2 langages de référence pour la data science, avec R
- Moins de bibliothèques statistiques que sur R, mais généralement les bibliothèques disponibles sont de meilleure qualité
- Python n'est pas spécifique à la statistique, contrairement à R.
- Des environnements de développement (IDE) très pratiques
- Quelques bibliothèques clefs pour l'étude de données avec Python
 - Lecture/Manipulation de données : pandas, numpy
 - Visualisation de données : matplotlib
 - Modélisation : scikit-learn, Xgboost, (StatsModels)



DES RESSOURCES POUR APPRENDRE LA DATA SCIENCE AVEC PYTHON

- Python :
 - Pour débutant
 - [Cours interactif sur Codecademy](#)
 - Livre : Swinnen G., [Apprendre à programmer avec Python3](#)
- Livres
 - Débutant et en français : Biernat E. & Lutz M., *Data science : fondamentaux et études de cas*, 2015
 - Plus difficile et en anglais : VanderPlas J., *Python Data Science Handbook*, 2016
 - Plus didactique et en anglais : Coelho & Richert, *Building machine learning systems with Python*, 2015
- Contenu que l'on utilisera dans les premières séances du GT
 - Fortement inspiré du cours de Ricco Rakotomalala ([lien](#)) et de Xavier Dupré ([lien](#))
- Challenge de Data Science
 - [Kaggle](#), [Kaggle in class](#), <https://www.datascience.net/>
- MOOC
 - [Coursera](#), [EdX](#), [Udacity](#), [FUN](#) (en français)
- Une question ? Un problème ?
 - Google -> [Stack overflow](#) / [Cross validated](#)



Forsides

52 Rue de la Victoire, 75009 Paris

T : +33 (0)1 42 97 91 70

Nos associés

Arnaud Cohen

T : 06 78 47 25 39 | 01 42 97 91 73
arnaud.cohen@forsides.fr

Marc Raymond

T : 06 42 18 94 08 | 01 42 97 91 69
marc.raymond@forsides.fr

Véronique Mattei

T : 06 83 25 42 83 | 01 42 97 91 72
veronique.mattei@forsides.fr

Valéry Jost

T : 06 83 30 64 88 | 01 42 97 91 76
valery.jost@forsides.fr

Avertissement : les informations, données et analyses présentées ne peuvent en aucun cas être assimilées à des prestations de services ou de conseil rendues par leurs auteurs ou par Forsides. Aussi, elles ne peuvent être utilisées comme un substitut à une consultation rendue par une personne professionnellement compétente. En tout état de cause, la responsabilité des auteurs de Forsides ne pourra en aucun cas être engagée du fait ou à la suite d'une décision prise sur la base des informations, données et analyses présentées.