Place de R et Python dans les formations en Data Science

Ricco Rakotomalala

Université Lumière Lyon 2 – Data, informatique et statistique

http://dis.univ-lyon2.fr/

Curriculum vitae

- Formation en économétrie (statistique, économie mathématique)
- Thèse de doctorat en Machine Learning (Apprentissage statistique)

- Enseignant chercheur, en poste à l'Université Lumière Lyon 2
- Spécialité : statistique, data mining et ses applications, informatique - Data Science

- « Père » des logiciels gratuits SIPINA v.3 et TANAGRA (open source)
- Auteur d'une dizaine d'ouvrages libres
- Auteur de plus de 500 supports de cours et tutoriels en français et en anglais

- 644 visites par jour depuis 10 ans (depuis le 1er février 2008, compteur Google Analytics)

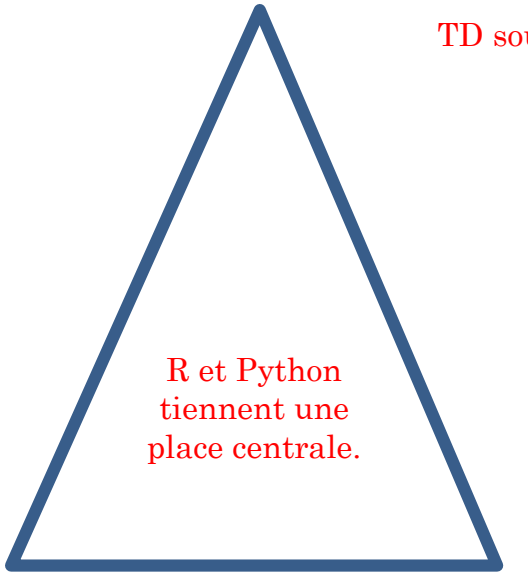
Master SISE (Statistique et Informatique pour la Science des donnéEs)

Existe depuis 1986 – Y intervient depuis 1995 – Responsable depuis 2013

Aujourd'hui

STATISTIQUE DATA MINING

Méthodes statistiques : séries temporelles, analyse de variance, biostat – données catégorielles
Méthodes de data mining / machine learning : méthodes supervisées, non supervisées, adaptées aux grandes dimensions, techniques de réduction de dimension, réseaux de neurones, deep learning.
TD sous R et Python.



Programmation R et **Python**.
Bases de données, **entrepôts de données, BI**.
Reporting - **Dataviz** (Qlik, Tableau)
Technologies big data (hadoop, spark, mapreduce, …)

INFORMATIQUE

Traitement des données non structurées : **Text Mining**, Image Mining, **Web mining**, analyse des réseaux sociaux.
Analyse des **données de sécurité**.
Marketing. Professionnalisation, projets transversaux. TD sous R et Python.

APPLICATIONS (Connaissances métiers)

Un point de vue personnel sur le choix et l'intérêt pédagogique des logiciels libres (gratuits) de data science dans les enseignements.

Plan

1. Les prémices (dont SIPINA v3)
2. Tanagra
3. R
4. Python
5. Etude des offres d'emploi
6. Conclusion

Les prémices

Logiciels (amateurs) pour les statistiques et le data mining

Regress

https://eric.univ-lyon2.fr/~ricco/regress.html

- Projet étudiant (Maîtrise d'Econométrie – Université Lyon 2 – 1992)
- Régression linéaire simple et multiple – Diagnostic de la régression
- Pas de réelles spécifications (devait permettre de réaliser le projet)
- 1994 : Utilisation par P. Sylvestre-Baron pour ses enseignements. Affinage des spécifications, extension des fonctionnalités, plus d'interactivité. Piloté par menu.
- 1997 : Amélioration de la gestion de données avec les structures de SIPINA v3
- 2004 : Intégration dans un pool incluant SIPINA v3 (StatPackage)

Un peu quand-même : chargement des données, lancement des traitements, affichage des résultats, graphiques.

Peu documenté (un peu quand même, PSB), peu valorisé,… peu utilisé.

Les principales fonctionnalités ont été reprises dans Tanagra.

Démonstration : infidélité.xls

Sipina v2… v2.5

http://tutoriels-data-mining.blogspot.fr/2010/05/sipina-version-25-presentation.html

- Projet du Laboratoire ERIC, initié par des étudiants du DESS SISE (1994 - 1995)
- Stage de DESS au Laboratoire ERIC (Gestion des règles)
- Maintenance et évolution durant thèse de doctorat (1995 – 1997)
- Pas de réelles spécifications, ni de cible (recherche ? enseignement ?)
- Limité aux arbres de décision, graphes d'induction
- Gestion déficiente des données et faiblesses structurelles ont empêché toutes possibilité d'évolution

Visibilité Internet – Promotion du Laboratoire ERIC

Un des très rares outils gratuits de data mining diffusé à l'époque

Arbres interactifs exclusivité des logiciels payants

Démonstration : iris

Sipina v3…

http://sipina-arbres-de-decision.blogspot.fr/

- Projet personnel (1997) nourri par les (dé)boires de la version 2.5
- Axes de réflexion : gestion performante des données, simplification des architectures internes, extension aux méthodes supervisées
- Tourné essentiellement vers la recherche, support des publications (à partir de 1998)
- Fantasmes de l'industrialisation – Pas de diffusion réelle au départ, statut incertain

Mis en ligne et documenté à partir de 2004

Reste le seul gratuit avec des arbres interactifs à ce jour

Association avec d'autres outils (Regress, Règles d'association)

Désactivation des modules purement recherche

Très peu d'évolutions depuis 2004

Démonstration : arbres interactifs, multithreading, swap

Au-delà de Sipina v3…

Restreint aux méthodes supervisées

Tentative de création d'un pool de logiciels (classification automatique, méthodes factorielles) basés sur le même gestionnaire de données, sur le modèle de STATISTICA, mais peu concluant

Impossibilité de garder une mémoire des enchaînements des traitements

Pilotage par menu = définition manuelle des traitements, aucune mémoire des enchaînements, obligation de refaire les mêmes clics à la réouverture du logiciel

Complexité des structures internes

Pour chaque nouvelle méthode : programmation de l'algorithme, mais aussi programmation de la fenêtre d'affichage des résultats. Rapport (code utile / code d'interface graphique) pas favorable du tout.

Nécessité d'un outil à la fois plus générique et plus simple (à utiliser, à programmer).

Tanagra

Un logiciel gratuit pour l'enseignement et la recherche

http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html

Contexte (2002)

- Vogue du data mining – Deux ouvrages font date : Lefébure & Venturi (1998), Tufféry (2002)
- Weka (ouvrage de référence en 2000) était le seul outil réellement populaire, mais trop tourné Machine Learning, pas de culture statistique (ex. régression logistique, méthodes factorielles, etc.)
- Expériences pédagogiques désastreuses avec deux éditeurs de solutions de Data Mining
- Nécessité de développer un outil dédié à l'enseignement

Réflexion entamée à l'été 2002 (veille techno 1, 2, 3, 4 + maquettes)

Début du développement au printemps 2003

Diffusion en janvier 2004

Eléments de réflexion

Evaluer les logiciels gratuits pour l'enseignement du data Mining (Séminaire déc. 2005)

1. S'attacher au fond et non à la forme

L'étudiant ne doit pas être dépendant de l'outil

2. Former des étudiants qui vont sur le marché du travail

Respecter les standards du domaine, conforme à la pratique du data mining (contre-ex. WinIDAMS)

3. L'apprentissage de l'outil ne doit pas requérir des compétences additionnelles

Avoir à apprendre un langage de script spécifique par ex. rogne sur le temps du data mining

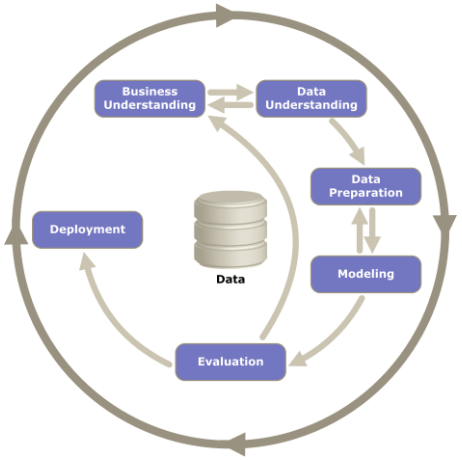
4. (*A posteriori*) Interfaçage fort avec Excel, outil privilégié du chargé d'études

Excel est systématiquement demandé pour les postes de chargé d'étude (à l' époque, mais encore aujourd'hui).

Voir aussi les Sondages KDnuggets 2005 et 2017. Le maîtriser est un atout.

Spécifications

Calé sur le processus Data Mining (CRISP-DM)



Critères pour les logiciels de Data Mining

1. Architecture : stand-alone, client-serveur, via un navigateur, …
2. Mode opératoire : diagramme de traitements, langage de script, pilotage par menu,…
3. Performances, capacités de traitement, temps de calcul
4. Accès aux données : fichiers textes, Excel, accès aux bases de données,…
5. Manipulation des données : transformations, recodage,…
6. Exploration graphique : représentations, visualisations, interactions,…
7. Bibliothèques de techniques de machine learning : supervisées, non-supervisées, …
8. Evaluation et comparaisons : comparaison des approches, benchmarking…
9. Reporting et solutions pour le déploiement (PMML,…)

Cahier des charges de Tanagra

Utilisateur

1. Gratuité totale – Non commercial – Choix d'une licence privative (Œuvre de l'esprit)
2. Installation simplifiée – Pas de serveurs ou librairies à installer
3. Gestion simplifiée des données – Connexion avec les tableurs (Excel, Libre/Open Office)
4. Fonctionnement par diagramme de traitements
5. Couvrir les statistiques, l'analyse de données et le data mining dans un cadre unifié
6. Résultats lisibles. Possibilité de les reprendre dans un tableur
7. Mettre de côté les aspects opérationnels (interfaçage BD, reporting dynamique, déploiement)

Développeur

A. Minimiser le code dédié à la gestion des données et à l'interface (sorties HTML)
B. Gestion simplifiée des méthodes (avec fichier de configuration)
C. Structures permettant l'ajout de toutes méthodes traitant des tableaux « individus x variables »

Tanagra – Outil pour la recherche

Accès libre au logiciel = Reproductibilité des expérimentations

Toute publication doit être reproductible en l'état par tout chercheur. Cela n'est possible que si l'outil est accessible librement (les données aussi d'ailleurs)

Accès libre au code source = Validation des implémentations

Comparer les implémentations permet de les valider, de les améliorer, de les optimiser. Ex. différence de temps de traitement entre Tanagra et Weka, sur les arbres de décision, sur les SVM, …

Accès libre au code source = Outil vivant (ex. Gong Yu – version 1.4.20 – Oct. 2007)

Possibilité pour les autres chercheurs d'introduire leurs propres algorithmes pour mener des expérimentations. Un peu au début, mais très peu l'ont fait finalement…

Promotion - Documentation

Ecrire un article de référence

Marque le coup en annonçant le logiciel. Sera la référence citée par les utilisateurs. Deux articles pour Tanagra : EGC 2005, Revue MODULAD (2005). Participation à quelques conférences, séminaires et ateliers même.

Documenter le logiciel

Sous la forme de tutoriels sur les méthodes plutôt que de rédiger un « manuel de référence » toujours en retard d'une version.

Supports de cours et tutoriels pour d'autres outils (R et Python notamment) ont pris une place importante (prépondérante) !

Monter un site web

La visibilité internet est une promotion à moindre coût. Un forum a été envisagé puis abandonné, trop énergivore.

Démonstration : infidélité.xls

R

Pourquoi R alors qu'il y a déjà Tanagra ?

Cours de programmation R en Master de Statistique (2006)

Cours de programmation en Master SISE

Il y a toujours eu un cours de programmation en Master SISE . Compétence indispensable pour un statisticien (ex. Estimations des régionales pour les soirées électorales de France 3 – Estimations à 20h)

Quel langage enseigner ?

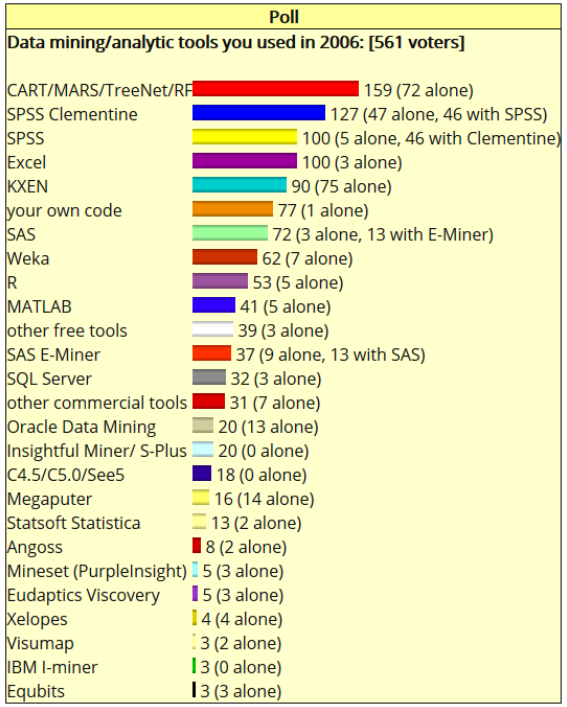
Turbo Pascal, Turbo Pascal pour Windows, Delphi, Java, C++…

R devenait un acteur important du Data Mining / Data Science

Les sondages des KDnuggets (pourtant ce n'était pas totalement évident en 2006)

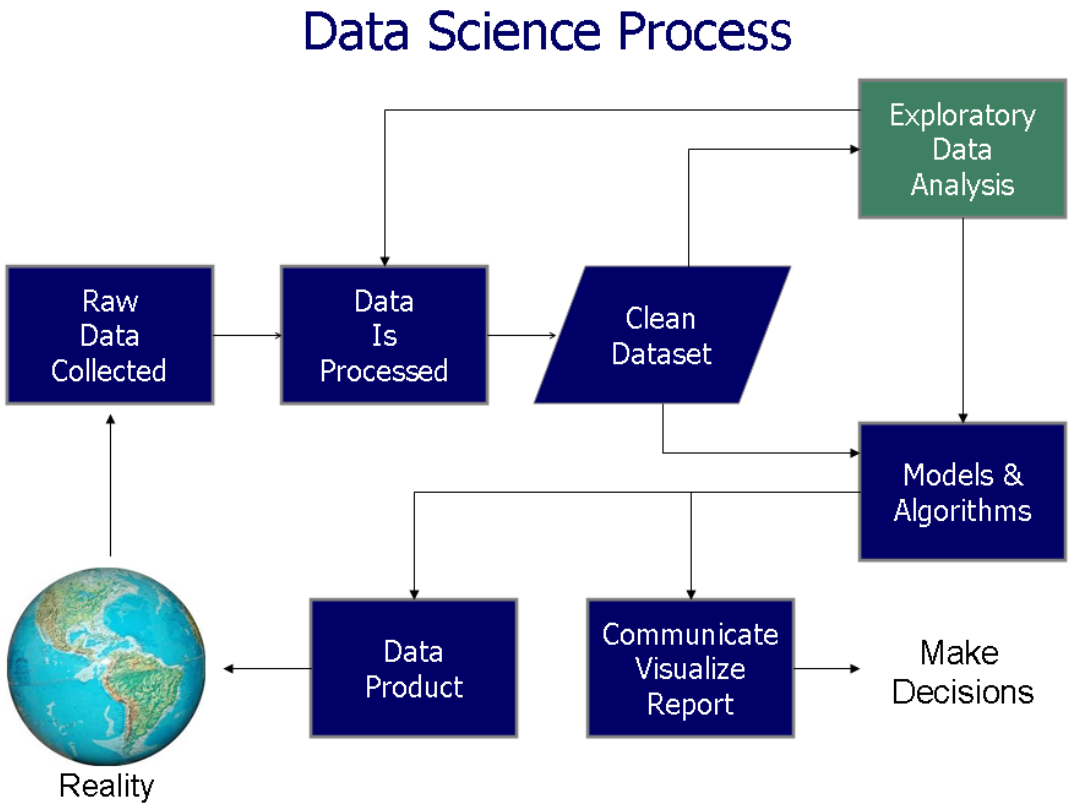
Programmation R + Statistiques sous R

Cumuler les avantages : savoir réaliser des traitements sous R (peu l'avaient manipulé avant le M2 à l'époque) et monter en compétence en programmation



Introduction du cours en M2 à la rentrée de septembre 2007
https://eric.univ-lyon2.fr/~ricco/cours/cours_programmation_R.html

R, encore plus indispensable dans le contexte de la Data Science



(https://en.wikipedia.org/wiki/Data_science)

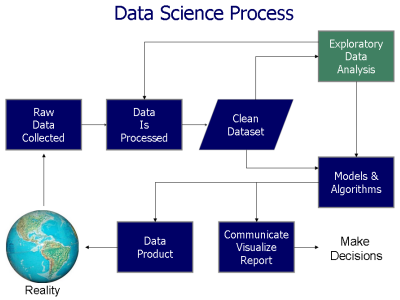
VARIETE

VOLUMETRIE

VELOCITE

A chaque étape sont associées des tâches spécifiques que doivent assurer les outils / logiciels de data mining. Préparation des données, modélisation et présentation sont au cœur du métier de statisticien.

Critères pour les logiciels de data science



1. Architecture : stand-alone, client-serveur, via un navigateur, …
2. Mode opératoire : diagramme de traitements, langage de script, pilotage par menu,…
3. Performances, capacités de traitement, temps de calcul
4. Accès aux données : fichiers textes, Excel, accès aux bases de données,…
5. Solutions pour la volumétrie, technologies big data
6. Accès aux données non structurées et primitives de traitements (texte, image, …)
7. Interfaçage avec les API du web (ex. Twitter, Google+, Google Analytics, …)
8. Manipulation des données : transformations, recodage,…
9. Exploration graphique : représentations, visualisations, interactions,…
10. Bibliothèques de techniques de machine learning : supervisées, non-supervisées, …
11. Evaluation et comparaisons : comparaison des approches, benchmarking…
12. Reporting et solutions pour le déploiement (PMML,…)

Atouts pour l'enseignement

Richesse des bibliothèques de data mining (machine learning, statistique, …)

Grâce au système des packages, extensible à l'infini. Attention à la validité des packages proposés (The R Journal). Exploration facilitée des nouveaux domaines (ex. Analyse des données spatiales)

Programmation statistique (ex. nouvelles méthodes de data mining, etc.)

Le langage le permet facilement. Production des solutions « clés en main » avec le développement des packages. Ce savoir faire est un véritable atout pour les étudiants. De plus en plus d'offres de stage !

Programmation Big Data

Programmation avancée MapReduce (Hadoop). Programmation parallèle. Une vraie compétence additionnelle qui fait le lien avec l'analyse fine des algorithmes de data mining.

Accès aux données exotiques – Utilisation des API

La profusion des packages ouvre démultiplie les possibilités d'accès aux données (ex. Google Analytics). Packages pour le traitement des données non-structurées (ex. Text Mining)

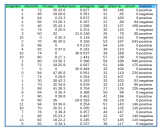
L'éclectisme de R facile l'acquisition des compétences additionnelles !

Démo – Analyse prédictive

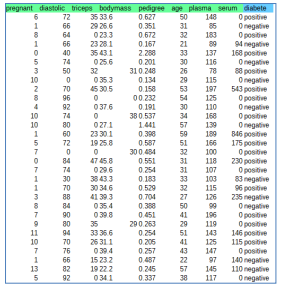
Prédiction de l'infidélité

Ensemble d'apprentissage

Construction de la fonction f(.) à partir des données d'apprentissage



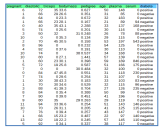
$Y = f(X1,X2,\ldots) + \varepsilon$



pregnant	diastolic	triceps	bodymass	pedigree	age	plasma	serum	diabete
6	72	35	33.6	0.627	50	148	0	positive
1	66	29	26.6	0.351	31	85	0	negative
8	64	0	23.3	0.672	32	183	0	positive
1	66	23	28.1	0.167	21	89	94	negative
0	40	35	43.1	2.288	33	137	168	positive
5	74	0	25.6	0.201	30	116	0	negative
3	50	32	31.0	0.248	26	78	88	positive
10	0	0	35.3	0.134	29	115	0	negative
2	70	45	30.5	0.158	53	197	543	positive
8	96	0	0.0	0.232	54	125	0	positive
4	92	0	37.6	0.191	30	110	0	negative
10	74	0	38.0	0.537	34	168	0	positive
10	80	0	27.1	1.441	57	139	0	negative
1	60	23	30.1	0.398	59	189	846	positive
5	72	19	25.8	0.587	51	166	175	positive
7	0	0	30.0	0.484	32	100	0	positive
0	84	47	45.8	0.551	31	118	230	positive
7	74	0	29.6	0.254	31	107	0	positive
1	30	38	43.3	0.183	33	103	83	negative
1	70	30	34.6	0.529	32	115	96	positive
3	88	41	39.3	0.704	27	126	235	negative
8	84	0	35.4	0.388	50	99	0	negative
7	90	0	39.8	0.451	41	196	0	positive
9	80	35	29.0	0.263	29	119	0	positive
11	94	33	36.6	0.254	51	143	146	positive
10	70	26	31.1	0.205	41	125	115	positive
7	76	0	39.4	0.257	43	147	0	positive
1	66	15	23.2	0.487	22	97	140	negative
13	82	19	22.2	0.245	57	145	110	negative
5	92	0	34.1	0.337	38	117	0	negative

Ensemble de données (dataset)

Application du modèle (prédiction) sur l'ensemble de test



Ensemble de test

$(Y, \hat{Y})$

Y : valeurs observées

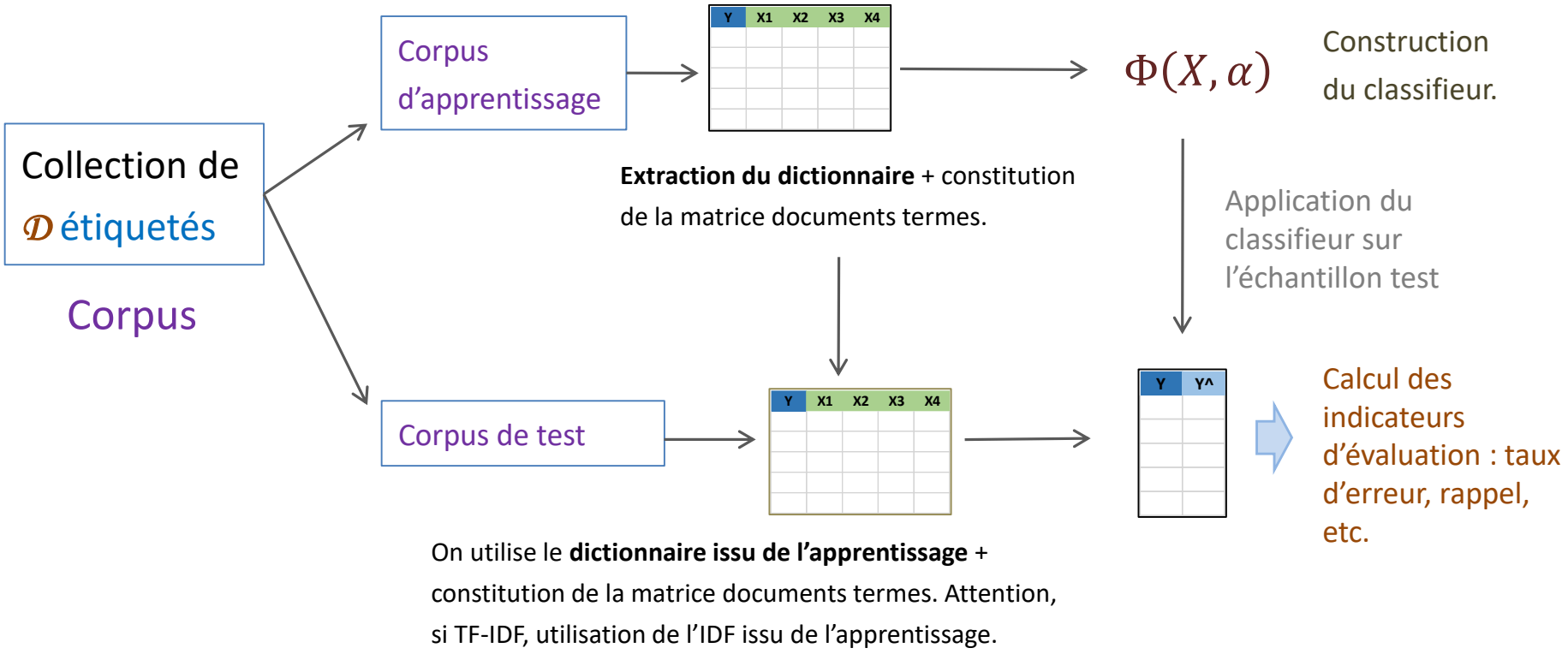
Y^ : valeurs prédites par f(.)

Mesures de **performances** par confrontation entre Y et Y^ : matrice de confusion + indicateurs

Démo – Text Mining – Catégorisation de documents

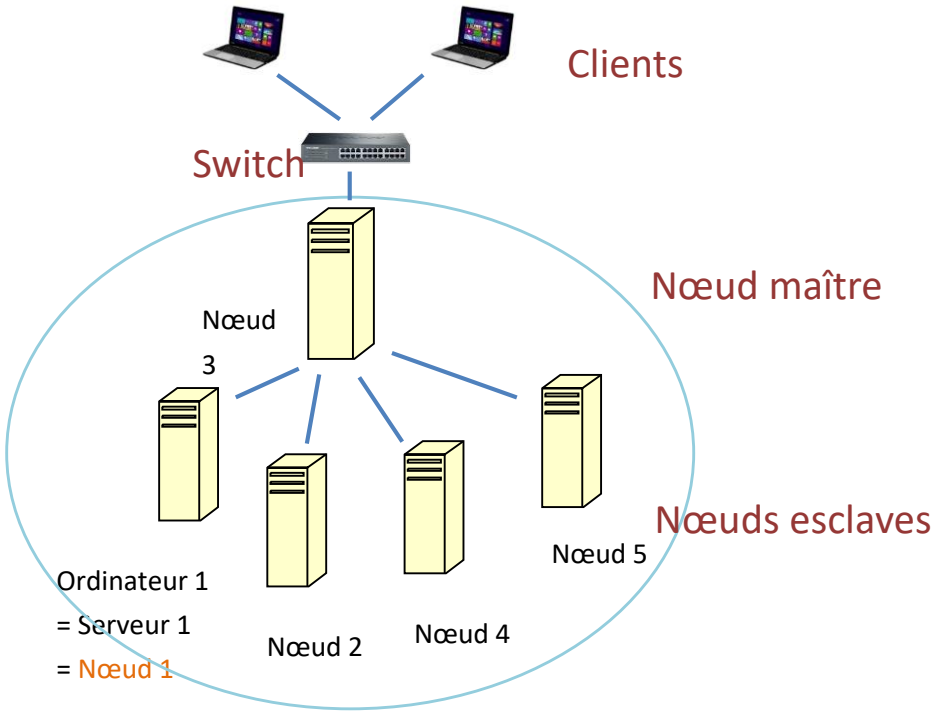
```
...
<document>
<sujet>acq</sujet>
<texte>
Resdel Industries Inc said it has agreed to acquire San/Bar Corp in a share-for-share exchange, after San/Bar distributes all shgares of its Break-Free Corp subsidiary to San/Bar shareholders on a share-for-share basis.
</texte>
</document>

<document>
<sujet>acq</sujet>
<texte>
Warburg, Pincus Capital Co L.P., an
investment partnership, said it told representatives of Symbion Inc it would not increase the 3.50-dlr-per-share cash price it has offered for the company. In a filing with the Securities and Exchange Commission, Warburg Pincus said one of its top executives, Rodman Moorhead, who is also a Symbion director, met April 1 with Symbion's financial advisor, L.F. Rothschild, Unterberg, Towbin Inc.
</texte>
</document>
...
```

Etiquetage automatique des nouvelles de « Reuters »

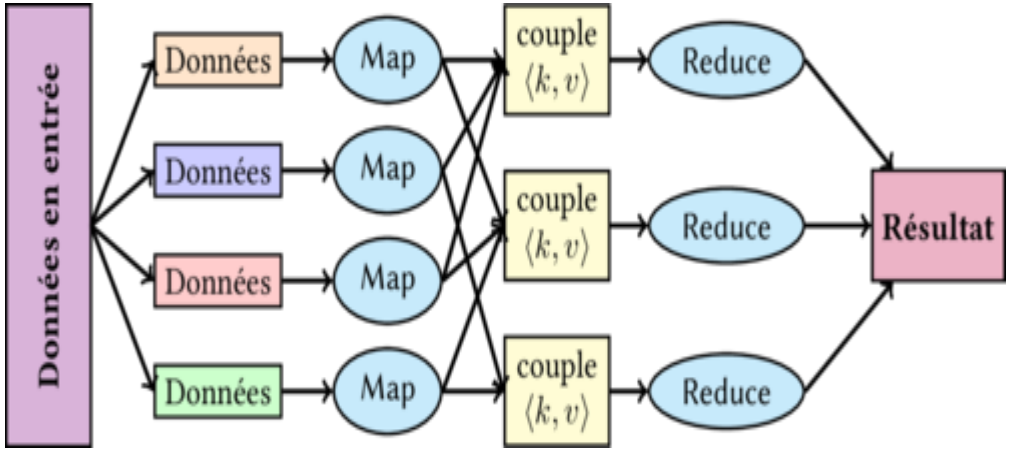


Démo – Programmation MapReduce

Une interprétation du calcul distribué / parallélisation des calculs
Calcul de la somme d'un vecteur



Grappe de machines



https://fr.wikipedia.org/wiki/MapReduce

Python

Pourquoi Python alors qu'il y a déjà R ?

Cours de programmation en L3 IDS (2014)

Cours de programmation en L3 IDS (Informatique et Data Science)

Couplage avec un cours d'algorithmie. « Initiation » à la programmation. Était basé sur Delphi/Lazarus (langage Pascal).

Avantages pédagogiques : bases de la programmation + prog. évènementielle + interfaces graphiques

Faire évoluer l'enseignement sans verser dans le phénomène de mode

De plus en plus d'articles en parlent en tant qu'outil pédagogique (ex. Python et enseignement, Juillet 2014) et outil professionnel dans notre domaine (Python et Big Data, Mars 2015). On est au-delà de l'épiphénomène.

Etre en phase avec le marché du travail

Au-delà des aspects pédagogiques, les offres d'emploi pour Delphi ne sont pas légions (APEC.FR, 69 offres au 17/01/2018).

Elles sont un « tout petit peu » plus nombreuses pour Python (APEC.FR, 1145 offres au 17/01/2018).

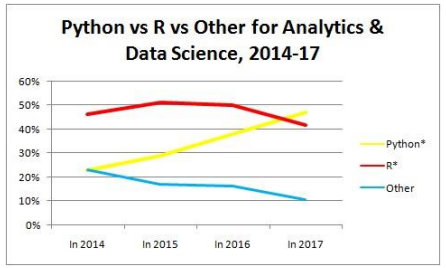
Introduction du cours en L3 à la rentrée de septembre 2015

http://tutoriels-data-mining.blogspot.fr/search/label/Python

Python en Master Data Science

Python un acteur incontournable de la Data Science

Progression forte ces dernières années (Sondage KDnuggets, Mai 2015), au point de dépasser (supplanter ?) R aujourd'hui (Août 2017).



Python présent dans les offres d'emploi Data Science / Big Data en France

Dans les offres d'emploi estampillées « data science », Python est incontournable (ex. Data Science Job Report, 2017). Voir l'étude des étudiants du Master SISE (recensement et qualification d'offres d'emploi statistique, data mining, data science, BI,…).

Richesse de la bibliothèque Big Data / Data Science

Elles complètent celles de R. Certaines sont plus compétitives (ex. Text mining – NLTK – semble plus complet) ou plus novatrices (ex. Deep Learning).

La difficulté (pédagogique) additionnelle par rapport à R est faible

!

Python est plus généraliste que R. Mais, dans nos domaines, le passage de l'un à l'autre est relativement facile au final. Voir le projet de « Reconnaissance faciale » et son contexte. On peut maîtriser facilement l'un ET l'autre.

En réalité, Python et R se complètent. Il nous appartient de déterminer le plus adapté selon les circonstances et les objectifs. !

Démo – Analyse prédictive

Détection de spams

Ensemble d'apprentissage

Construction de la fonction f(.) à partir des données d'apprentissage

$Y = f(X1,X2,\ldots) + \varepsilon$

Application du modèle (prédiction) sur l'ensemble de test

$(Y, \hat{Y})$

Y : valeurs observées
Y^ : valeurs prédites par f(.)

Mesures de **performances** par confrontation entre Y et Y^ : matrice de confusion + indicateurs

Ensemble de test

Ensemble de données (dataset)

pregnant	diastolic	triceps	bodymass	pedigree	age	plasma	serum	diabete
6	72	35	33.6	0.627	50	148	0	positive
1	66	29	26.6	0.351	31	85	0	negative
8	64	0	23.3	0.672	32	183	0	positive
1	66	23	28.1	0.167	21	89	94	negative
0	40	35	43.1	2.288	33	137	168	positive
5	74	0	25.6	0.201	30	116	0	negative
3	50	32		31 0.248	26	78	88	positive
10	0	0	35.3	0.134	29	115	0	negative
2	70	45	30.5	0.158	53	197	543	positive
8	96	0		0 0.232	54	125	0	positive
4	92	0	37.6	0.191	30	110	0	negative
10	74	0		38 0.537	34	168	0	positive
10	80	0	27.1	1.441	57	139	0	negative
1	60	23	30.1	0.398	59	189	846	positive
5	72	19	25.8	0.587	51	166	175	positive
7	0	0		30 0.484	32	100	0	positive
0	84	47	45.8	0.551	31	118	230	positive
7	74	0	29.6	0.254	31	107	0	positive
1	30	38	43.3	0.183	33	103	83	negative
1	70	30	34.6	0.529	32	115	96	positive
3	88	41	39.3	0.704	27	126	235	negative
8	84	0	35.4	0.388	50	99	0	negative
7	90	0	39.8	0.451	41	196	0	positive
9	80	35		29 0.263	29	119	0	positive
11	94	33	36.6	0.254	51	143	146	positive
10	70	26	31.1	0.205	41	125	115	positive
7	76	0	39.4	0.257	43	147	0	positive
1	66	15	23.2	0.487	22	97	140	negative
13	82	19	22.2	0.245	57	145	110	negative
5	92	0	34.1	0.337	38	117	0	negative

Démo – Reconnaissance faciale

Démarche de recherche d'information par le contenu. Projet en Python.

Disposer d'une banque d'images

Extraction de caractéristiques

Matrice de description, ligne : individus, colonnes : caractéristiques.

X1	X2	X3	X4	X5

Recherche de similarités.

Extraction de caractéristiques

X1	X2	X3	X4	X5

Image « requête »

Vecteur de description de l'individu « requête »

Identification avec degré de fiabilité.

Démo - Reconnaissance musicale

Démarche de recherche d'information par le contenu. Projet en Python.

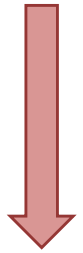
Disposer d'une banque de musiques

Extraction de caractéristiques

Matrice de description, ligne : chansons, colonnes : caractéristiques.

X1	X2	X3	X4	X5

Matching + recherche de similarités.



Extraction de caractéristiques

X1	X2	X3	X4	X5

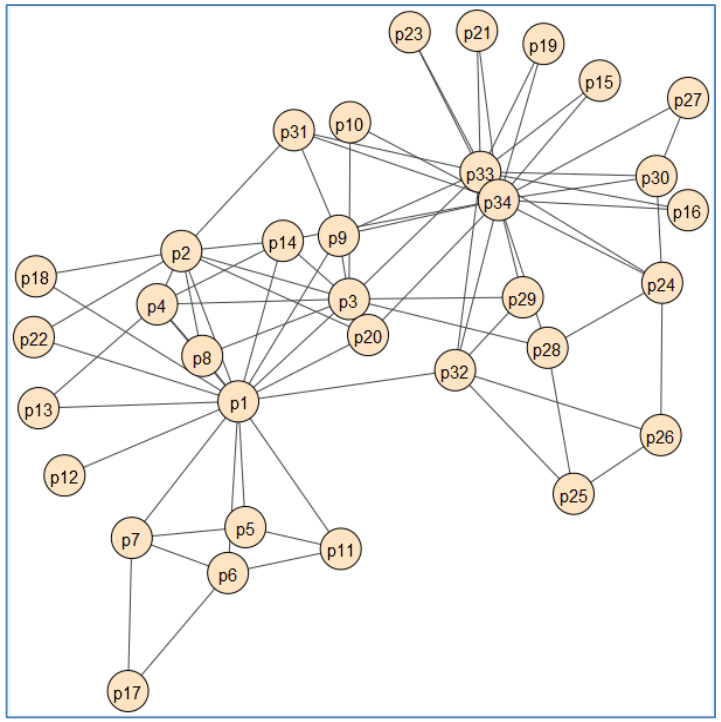
Chanson « requête »

Vecteur de description de la chanson « requête »

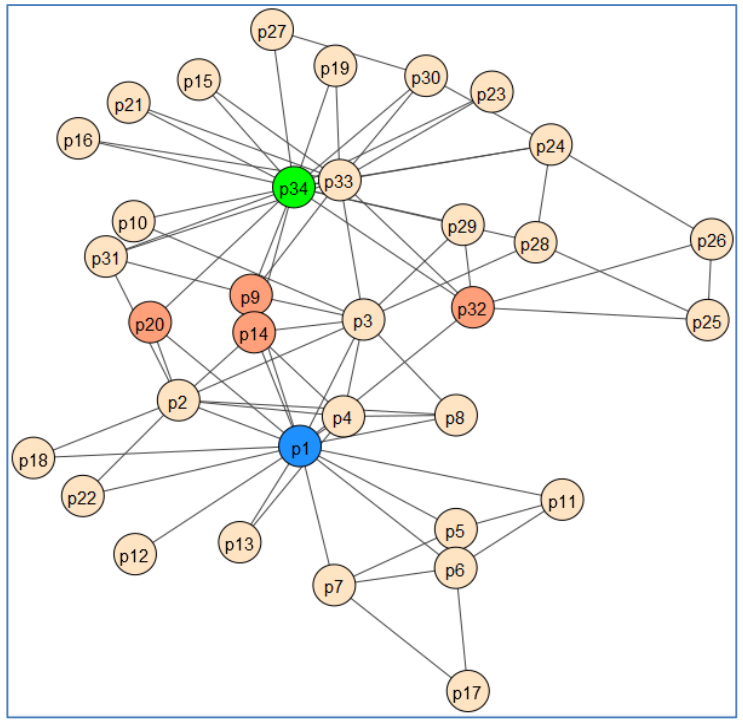
Identification et recommandation

Démo – Détection de communautés

Code comparé Python et R pour la détection des communautés dans les réseaux sociaux



Point de départ : des individus plus ou moins connectés entre eux.



Découvrir des individus centraux (centralité) autour desquels s'agglomèrent des communautés. Identifier les individus qui peuvent jouer le rôle de relais.

Analyse des offres d’emploi

Positionnement des outils dans les offres d’emploi

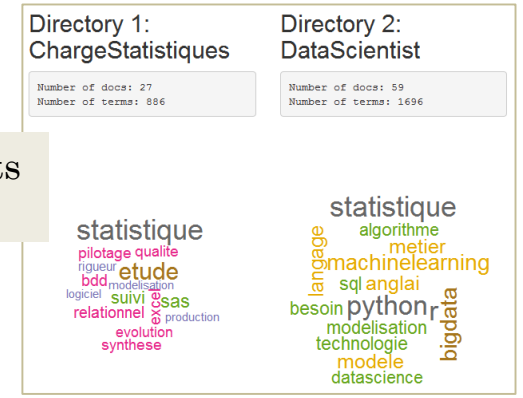
Analyse des offres d'emploi

Analyse de documents textuels (text mining) et classement / classification. Projet sous R (Shiny)

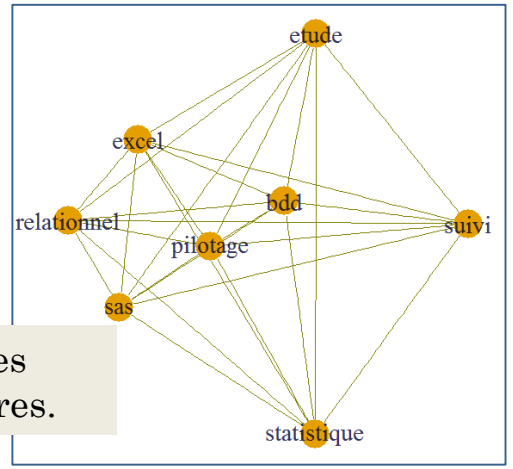
Offres d'emploi qui ont été étiquetées manuellement.

Analyse et développement d'un application Shiny

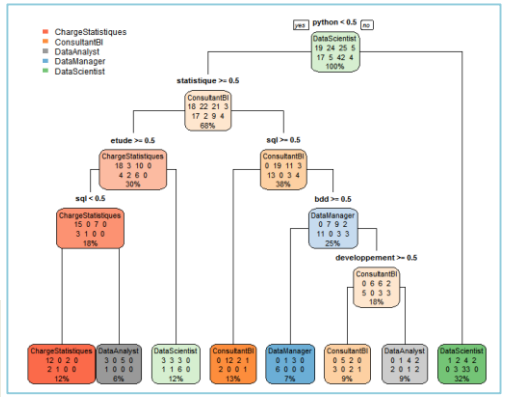
Mots clés fréquents selon les métiers



Association entre les termes dans les offres.



Identification des métiers selon les termes de l'offre.



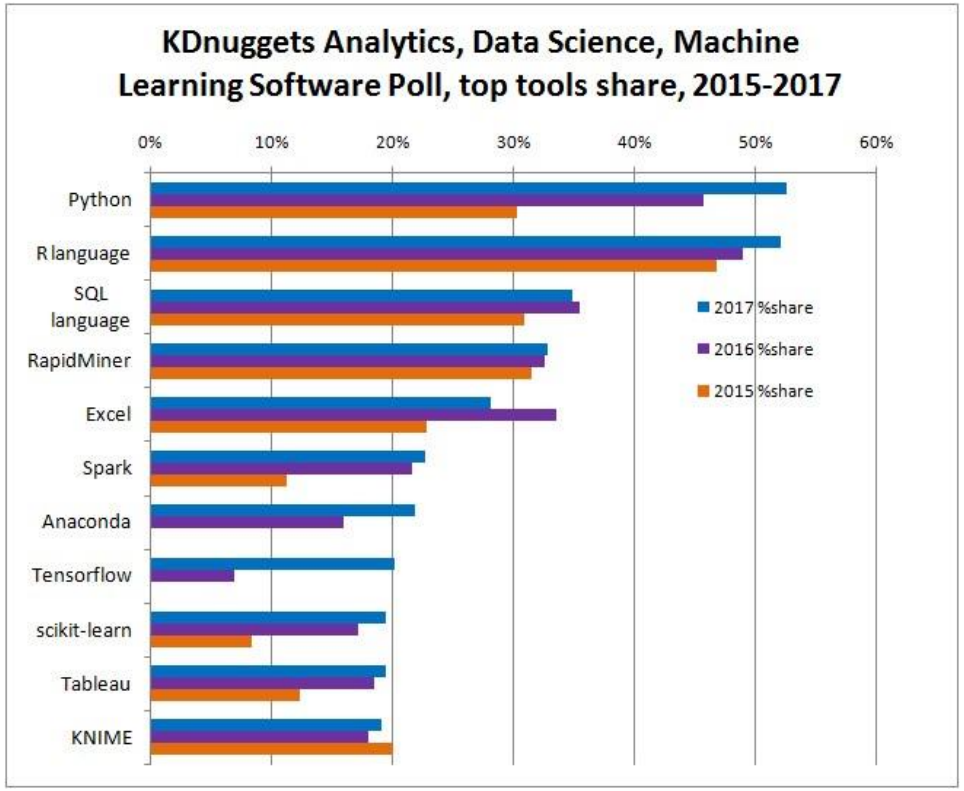
Métiers : Chargés d'études statistique, consultant BI, data analyst, data engineer, data manager, data miner, data scientist, data visualisation

Conclusion

R et Python réellement incontournables ?

Enquête Kdnuggets + Etude Gartner

On demande aux internautes d'indiquer les logiciels qu'ils utilisent (Mai 2017, 2900 votants) (2017 Software Polls Results) + Analyse détaillée (Juin 2017), notamment des associations entre outils.



Evaluation de 16 outils analytiques commerciaux (*pas tous*) sur la base de 15 critères. (Février 2017, Data Science Platforms: gainers and losers).

Codes de lecture (Gartner.com) :

Leaders execute well against their current vision and are well positioned for tomorrow.

Visionaries understand where the market is going or have a vision for changing market rules, but do not yet execute well.

Niche Players focus successfully on a small segment, or are unfocused and do not out-innovate or outperform others.

Challengers execute well today or may dominate a large segment, but do not demonstrate an understanding of market direction.

Conclusion

R et **Python** proposent des fonctionnalités et des performances opérationnelles. Ils sont incontournables aujourd'hui dans la Data Science. Les éditeurs de logiciels l'ont compris et les intègrent dans leurs solutions.

- Il faut une formation dédiée et de la pratique pour savoir réellement tirer parti de ces outils (d'où la profusion de MOOC…).
- Pas R ou Python mais plutôt R et Python. Et bien maîtriser les deux n'est pas un problème.
- Le choix dans la pratique dépend du contexte, des objectifs de l'étude, des packages disponibles et de leur qualité, etc.