Introduction aux Data Sciences

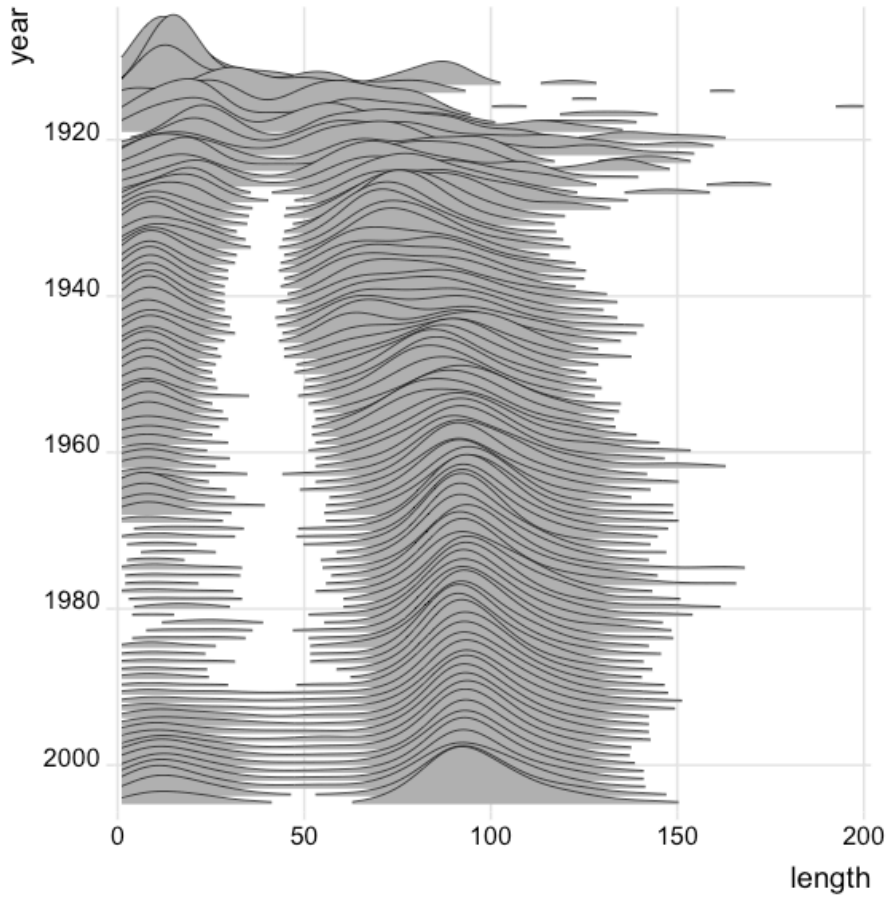
Christophe Benavent - Université Paris Dauphine

2022-09-25

Contents

Chapter 1

Avant propos



Ce bookdown présent les éléments d'un cours de data science avec r. Il est reproductible, on peut en cloner les élement à partir du repository. Le texte est encore hasardeux, les codes sont vérifiés.

Il sera dynamique, modifié à mesure de nos cours, séminaires et ateliers.

L'illustration de couverture représente l'évolution de la longueur des films de la base Imbd et raconte en chiffres un aspect de l'histoire du cinema. Jusqu'aux années 30, la longueur est hétérogène ensuite elle se se stabilise : les courts-métrages ont une durée de l'ordre de 15mn qui se raccourcit avec les décennies, ce genre menace de disparaitre dans les années 80 et reprend du poil de la bête dans les années 2000. Les films longs voient leur longueur s'accroitre et se stabiliser autour d'un peu moins de 100 mn, soit une heure et quarantes minutes. On observera enfin qu'au cours des années 1990 les films de taille intermédiaires réapparaissent. On devinera dans cette évolution l'émergence de standard, ou de convention.Les faits viennent au secours des théories…

Dans ce graphique il y a tous les éléments des data sciences contemporaine : un jeu de donnée riche et systématique, un modèle statistique fondamental avec la notion de densité de probabilité, une mesure, un critère de comparaison.

Les diagrammes ridges, ce sont leurs dénominations, sont inspirés de la pochette de l'album Unknown Pleasures de Joy division sorti en pleine New Wave , en 1979. Un article de Vice en rappele l'origine et le destin du graphisme qu'on connait mieux imprimé sur des t-shirt que dans les cours de statistiques.

1.1 Plan du manuel

C'est un projet en cours, Les chapitres projetés sont les suivants. certains sont dans les limbes, d'autres ont pris consistances

- 1 - L'environnement r x
- 2 - Installation et prise en main x
- 3 - Usage de ggplot - uni et bivarié x
- 4 - Usage de ggplot - multivarié x
- 6 - Analyse de variance et régression linéraire x
- 5 - Tables
- 6 - Modèles factoriels (Psych) x
- 7 - AFC x
- 8 - MDS

- 9 - Clustering x
- 10 - Analyse de réseaux
- 10 - Modèle d'équations structurelles (Lavaan)
- 11 - Modèle linéaire généralisé
- 12 - Modèles à décomposition d'erreur

- 13 - Times series
- 14 - Analyse geospatiale
- 15 - Machine learning x

1.2 Les jeux de données

Au cours du développement, plusieurs cas pratiques - souvent réduit en volume pour rester exemplaire, seront employés. Les données seront partagées.

En voici la présentation des sets de données utilisées dans le syllabus. Elle sont disponible dans le répertoire "./data/"

- ESS : c'est une très belle base de données de sociologie.
- happydemics : observatoire de la présidentielle2022
- Arpur

1.3 Le cadre technique et les packages utilisés

Ce *syllabus* est écrit en **Markdown** (**?**) et avec le package **Bookdown** (**?**)

Le code s'appuie très largement sur `tidyverse` et emploie largement les ressources de `ggplot`. Les packages seront introduits au fur et à mesure. En voici la liste complète.

```
options(tinytex.verbose = TRUE)

knitr::opts_chunk$set(echo = TRUE, include=TRUE, cache=TRUE, message=FALSE, warning=FALSE)

#boite à outils et dataviz
library(tidyverse) # inclut ggplot pour la viz, readr et
library(cowplot) #pour créer des graphiques composés
library(ggridges) # le joy division touch
library(ggmosaic)

#networks
library(igraph)
library(ggraph)

# Accéder aux données
library(rtweet)  # une interface efficace pour interroger l'api de Twitter

# NLP
library(tokenizers)
```

```
library(quanteda)
library(quanteda.textstats)
library(udpipe) #annotation syntaxique
library(tidytext)
library(cleanNLP) #annotation syntaxique

#sentiment
library(syuzhet)              #analyse du sentimeent



#mise en page des tableaux
library(flextable)


#statistiques et modèles
library(lme4) #pour des modèles plus complexe que les mco
library(jtools) #une série d'utiltaire pour bien représenter les résultats
library(interactions) #traitement des interactions

library(corrplot)
library(psych)


library("FactoMineR")
library("factoextra")


#ML
library(caret)

#utilitaires
library(rcompanion)

#graphismes
library(ggthemes)
theme_set(theme_bw())


#palettes
library(colorspace) #pour les couleurs
library(wesanderson)
```

```
# Utilitaires

library(citr) #pour insérer des références dans le markdown
```

L'ensemble du code est disponible sur github. A ce stade c'est encore embryonnaire. Les proches et nos étudiants pourrons cependant y voir l'évolution du projet et de la progression

Quelques conventions d'écriture du code r

- On appele les dataframes de manière générale `df`, les tableaux intermédiaires sont appelé systématiquement `foo`
- Gestion des palettes de couleurs ** une couleur :" royalblue" ** deux couleurs ** 3 à 7 couleurs
- On emploie autant que possible le dialecte tidy.
- Les chunks sont notés en 4 chiffre : 2 pour le chapitre et deux pour le chunck. 0502 est le second chunk du chapitre 5.
- On commente au maximum les lignes de code pour épargner le corps du texte et le rendre lisible

1.4 A faire

todo list :

- insérer un compteur google analytics (voir https://stackoverflow.com/questions/41376989/how-to-include-google-analytics-in-an-rmarkdown-generated-github-page)
- modifier le titre en haut à gauche
- vérifier le système de références voir (https://doc.isara.fr/tuto-zothero-5-bibtex-rmarkdown-zotero/)
- Vérifier la publication en pdf

Chapter 2

Introduction aux data sciences

2.1 Objectif et sommaire

L'objet du manuel est de donner un aperçu général des méthodes d'analyses de données et de data science.

2.2 Science ou technique ?

Plûtôt que le terme consacré de Data sciences, il vaudrait mieux parler de data ingiénérie dans la mesure où le data scientiste participe à un processus de production qui va de l'acquisition des donnée à leur propagation dans l'organisation ou la société. La technique domine sur la science et l'unité se trouve dans l'intégration de ce processus. La révolution des données vient de l'interopérabilité croissante de ces techniques et d'une intégration qui fluidifie le passage d'une étape à une autre. Standards et langages en sont les éléments clés.

Du côté des sciences, ce dont bénéficie l'univers des data sciences, c'est l'héritage de cultures statistiques foisonnantes qui après s'être développées dans leur cocon displinaire, se retrouvent désormais rassemblées dans un même langage. Bien sur il y a de manière sous-jascente à ces cultures les mathématiques et les statistiques mathématiques qui construisent les fondements des modèles et des techniques. Mais le développement s'est fait souvent quand le scientifique se retrouve face à un problème où une observation.

Prenons le cas des psychologues qui ont inventé l'analyse factorielle dans le but de pouvoir tester certains de leurs concepts : un degré d'intelligence, une

personnalité, des attitudes.

Ou celui des écologues qui souhatent estimer une population de poisson dans une rivière, problème qui a donné naissance aux modèles de capture recapture. On pourrait ajouter les géographes avec les modèles d'analyse spatiale, les financiers face à la variabilité des cours des places boursières, etc. Celui des économètres est peut-être le plus évident. Les biostatisticiens sont des contributeurs importants.

Ce que la technique apporte c'est l'intégration par un langage et donc un ensemble de conventions, incarnées par r et python, d'algoritmes, et de programmes qui ne sont plus spécifique à un domaine, mais peuvent circuler de l'un à l'autre. C'est ainsi que le catalogues de toutes les techniques psychométriques devient accessible aux autres disciplines par le biais d'un pachage en particulier , psych. De la même manière l'outillage des linguiste devient accessible aux autres disciplines, pensons aux économiste qui intéègre dans le indicateurs des sources textuelle telle que l'analyse du sentiment.

L'interopérabilité apportée par ces langages ne se définit pas que par l'algorithme qui aurait été porté d'un autre langage vers celui-ci (des cas de réécriture ?) mais aussi par des programme passerelle qui à partir de r permettent d'activité des algorithme écrit en C, en javascript ou tout autre langage "plus informatiques" et souvent plus éfficace.

2.3 histoires des logiciels statistiques

Et c'est ce qu'on observe dans l'évolution des logiciels

- 1980 : statitcf
- 1980 : SAS comme accès à r
- 1990 : SPSS

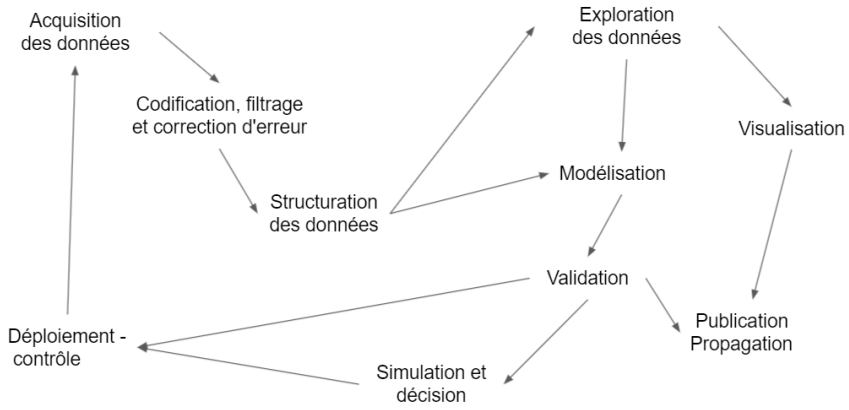
http://www.deenov.com/blog-deenov/histoire-du-logiciel-spad.aspx

des système portable

intégration graphique

la modularisation : base /fonction/ packages

2.4 Le processus de traitement des données



- Acquisition
- Codification , filtrage et correction d'erreur
- Structuration des données : api, open data
- Exploration
- Modélisation :
- validation : tests versus AB testing
- Simulation et décision
- Vizualisation et sensemaking
- Déploiement :
- Contrôle :
- Publication : dash board, pdf , slide etc, webb site

2.5 Les facteurs de développement des data-sciences

Ces développements sont favorisés par un environnement fertile dont trois facteurs se renforcent mutuellement.

2.5.1 Une lingua franca

histoire de r histoire de python

2.5.2 Une communauté

Le second facteur , intimement lié au premier, est la constitution d'une large communauté de développeurs et d'utilisateurs qui se retrouvent aujourd'hui

dans des plateformes de dépots (Github, Gitlab), de plateformes de type quora (StalkOverFlow), de tutoriaux, de blogs (BloggeR), de journaux (Journal of Statistical Software) et de bookdown.

Des ressources abondantes sont ainsi disponibles et facilitent la formation des chercheurs et des data scientists. Toutes les conditions sont réunies pour engendrer une effervescence créative.

2.5.3 La multiplication des sources de données.

Le troisième est la multiplication des sources de données et leur facilité d'accès. Les données privées, et en particulier celles des réseaux sociaux, même si un péage doit être payé pour accéder aux APIs, popularisent le traitement de données massives. Le mouvement des données ouvertes (open data) proposent et facilitent l'accès à des milliers de corps de données : retards de la SNCF, grand débat, le formidable travail de l'Insee, european survey etc.

2.5.4 du ML à l'IA

Le retour au boites noires dans les années 2000. Ce qui distingue les statistiques traditionnelles de l'approche machine learning réside d'abord par une approche de la modèlisation différente. Les modèles statistiques et économétriques considèrent non seulement une structure (moddèle linéaires par ex), la spécification du modèle, mais aussi des modèles de distribution qui définissent le cadre d'estimation. L'évaluation passe par le test du respects des hypothèses de constructions (distribution des erreurs), et de la qualité d'ajustement. Le machine learning, se concentre sur la valeur prédictive, et considère n'importe quelle spécification même si elle est peu intelligible et comprend de grandes quantité de paramètres.

KNN, SVM, rf et le retour des réseaux de neurones.

La révolution des convolutions et la multiplication des architectures