KEYRUS

insight into value

CONTEXTE

- Keyrus Régions **présent aux rencontres R depuis la création** de la practice Big Data & Analytics en juin 2014
  2014 Montpellier : 1 participant – 1 consultant
  2015 Grenoble : 1 participant – 7 consultants
  2016 Toulouse : 3 participants – 18 consultants

- Keyrus Régions **sponsorise** les 5èmes rencontres R 2016.

- R est un **MUST HAVE** pour tous les consultants de la practice.

- Présents aux rencontres R Toulouse 2016 :

Sylvain Coppéré | Consultant Manager
Practice Big Data & Analytics
+33 (0)6 63 85 86 82
sylvain.coppere@keyrus.com

Victoria Granger | Consultant Senior / Docteur
Practice Big Data & Analytics
+33 (0)6 32 98 83 41
victoria.granger@keyrus.com

Gabriel Julien | Responsable Sud-Ouest
Commerce
+33 (0)6 15 34 50 54
gabriel.julien@keyrus.com

© Copyright 2015 – Keyrus

AGENDA

R DANS LE DÉVELOPPEMENT DE NOTRE PRACTICE BIG DATA & ANALYTICS EN RÉGIONS

BIG DATA & ANALYTICS chez KEYRUS REGION

POSITIONNEMENT DE R DANS NOS MISSIONS

TYPOLOGIES DE MISSION

CONCLUSIONS

© Copyright 2015 – Keyrus

AGENDA

BIG DATA & ANALYTICS chez KEYRUS REGION

POSITIONNEMENT DE R DANS NOS MISSIONS

TYPOLOGIES DE MISSION

CONCLUSIONS

© Copyright 2015 – Keyrus

KEYRUS REGIONS

3 UNIVERS D'EXPERTISE

Management & Transformation

Gestion du Changement
Finance
Organisation RH
Stratégie IT
Marketing et Communication
Achats
Pilotage des Services

KEYRUS management

Data Intelligence

Business Intelligence
Gestion des données et des référentiels
Performance de l'administration
Big Data / Analytics
Open Data

KEYRUS data

Digital Experience

Stratégies Digitales
CRM, gestion de l'administré / usager
Transformation Digitale
Portail 360° citoyen / 360° Patient
Applications mobiles

KEYRUS digital

© Copyright 2015 – Keyrus

DATA INTELLIGENCE

BIG DATA & ANALYTICS INTÉGRÉE À NOTRE OFFRE DATA INTELLIGENCE

Data Intelligence Région Centre-Est

***100 consultants**, experts, managers, consultants*

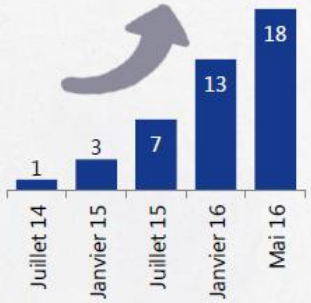
PRACTICES DATA INTELLIGENCE

BIG DATA & ANALYTICS	BI & EPM	SMART BI	DATA MANAGEMENT
18	55	13	7

Direction des Technologies

POLES D'EXPERTISES

... suite à un fort développement



18 consultants

... 4,5 ans d'expérience moyenne ...

... et une diversité accrue des profils

© Copyright 2015 – Keyrus

BIG DATA & ANALYTICS

UNE PRACTICE EN FORT DÉVELOPPEMENT

DATA ANALYST / STATISTICIEN

- # Datamining et DataViz
- # Forte connaissances « métier »
- # Outils : SAS, SPSS, SQL, R Shiny, Tableau

DATA SCIENTIST

- # Techniques de machine Learning
- # Notion en distribution des calculs
- # Outils : Python, R, Matlab, Spark, Hive
- # Ecosystème Big Data

DOCTEUR EN SCIENCES

- # Modélisation complexe
- # Gestion de données très variées
- # Veille sur les techniques d'analyse
- # Outils : R, Matlab, RShiny

BIG DATA

- # Architecture Big Data
- # Datalake et des connecteurs
- # Ecosystème Hadoop

DATA & ANALYTICS PROJECT MANAGER

- # Profils expérimentés
- # Gestion de projet et d'équipes de Data Analysts & Scientists
- # Accompagnement méthodologique et technologique

© Copyright 2015 – Keyrus

AGENDA

BIG DATA & ANALYTICS chez KEYRUS REGION

POSITIONNEMENT DE R DANS NOS MISSIONS

TYPOLOGIES DE MISSION

CONCLUSIONS

© Copyright 2015 – Keyrus

PANORAMA « ANALYTICS »

POSITIONNEMENT R

~50% des missions utilisent R
~25% des missions sont exclusivement R

MUST HAVE dans les recrutements

R (l'open-source) accède au **monde privé,** ce n'était pas le cas il y a 5-10 ans

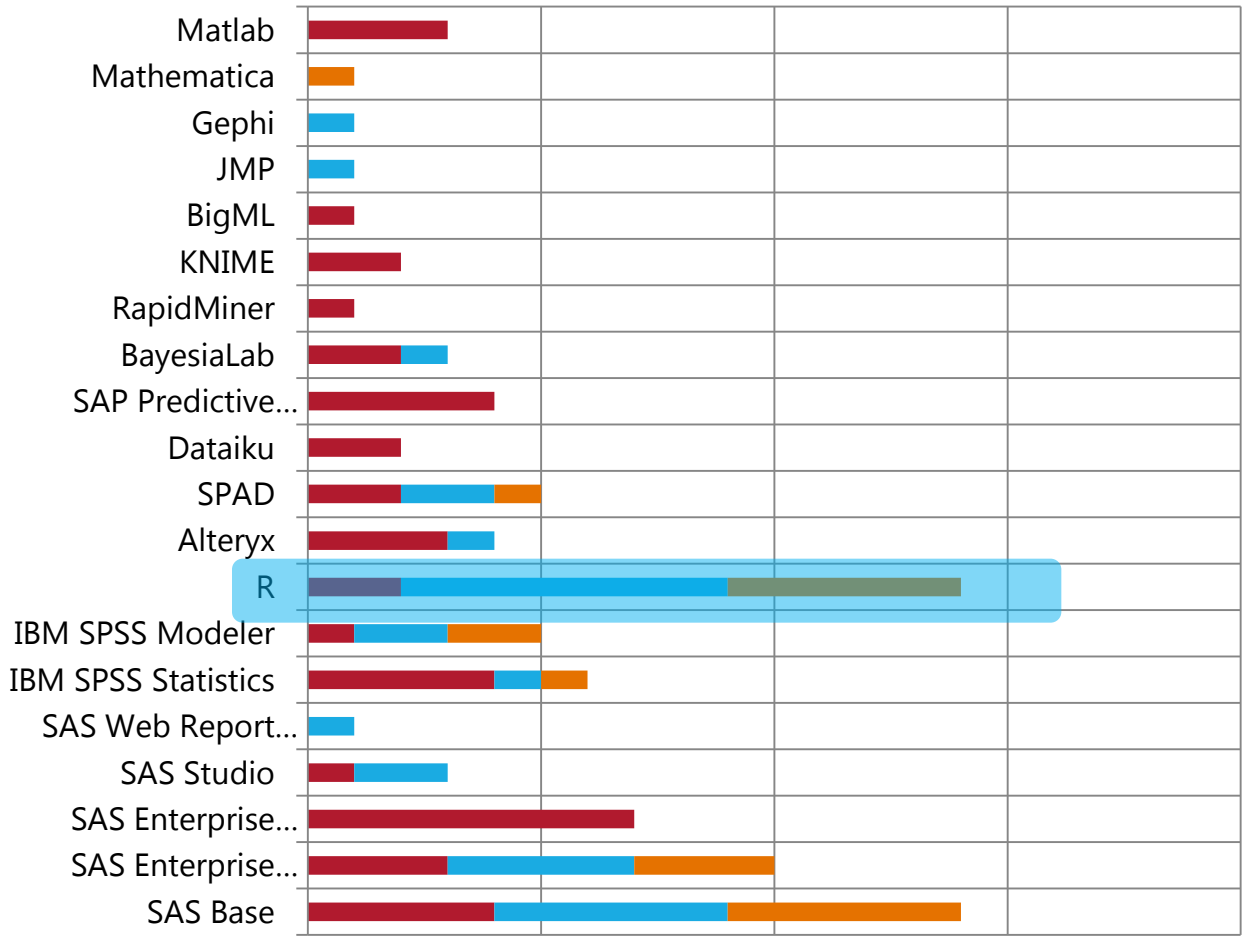
Notoriété en hausse **chez les non-initiés**

Tous les éditeurs embarquent R (lecture script, définition de nœuds spécifiques…)

Recherche pour coupler la force de R (langage de statisticiens) avec la **puissance informatique** (Revolution Analytics, R++…)

Des grands groupes qui se questionnent **sur des migrations** SAS → R / serveur R

Des **écosystèmes R / serveur R déjà éprouvés**



© Copyright 2015 – Keyrus

KEYRUS

ENVIRONNEMENTS DE TRAVAIL

R Studio Desktop seul

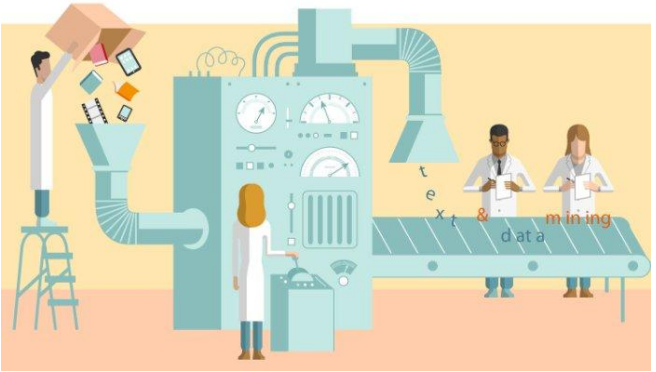
R Studio Desktop ; données dans un cluster MapR
-Connexion et récupération des données via un package ODBC pour interfaçage R-Hive

R Studio Server ; données dans un cluster Hortonworks
-Connexion et récupération des données via un package JDBC pour interfaçage R-Hive

R Studio Server ; Calculs et Sessions RStudio simultanées
-Données dans bases de données MongoDB, Oracle, NoSQL Hive, HDFS (hadoop distributed file system) permettant des récupérer des fichiers et des archives des fichiers EN MASSE
-Partage de packages / Utilisation des applications Shiny et partage avec utilisateurs métiers / Connexion à des outils DataViz (Tableau)

© Copyright 2015 – Keyrus

PRINCIPAUX PACKAGES UTILISES PAR NOS CONSULTANTS



Data collection : rodbc, rjdbc, xml, rjson

Data preparation : datatable, dplyr, doBy, sqldf

Data vizualization : ggplot2, lattice, plotly, rgl (3D), shiny, RmarkDown, igraph

Data modeling : Hmisc (clustering), randomForest, glm, ade4 (acp), lda, kknn

Data performance : gbm (boosting), boot (bootstrapping)

© Copyright 2015 – Keyrus

AGENDA

BIG DATA & ANALYTICS chez KEYRUS REGION

POSITIONNEMENT DE R DANS NOS MISSIONS

TYPOLOGIES DE MISSION

CONCLUSIONS

© Copyright 2015 – Keyrus

MISSIONS

MODES D'INTERVENTION / PROBLÉMATIQUES

	DATA PREPARATION	DATA ANALYSIS	DATA MODELING	DATA VIZUALISATION
ACCOMPAGNEMENT	API access	Segmentation		Cartographie SIG
ETUDES	Open data	Datamining	Scoring	
PROGRAMMES	Parsing	Graph mining		d3js
PROOF OF VALUE / MAQUETTE	Fuzzy matching	Web Analytics	Machine Learning	Shiny app
FORMATION				

© Copyright 2015 – Keyrus

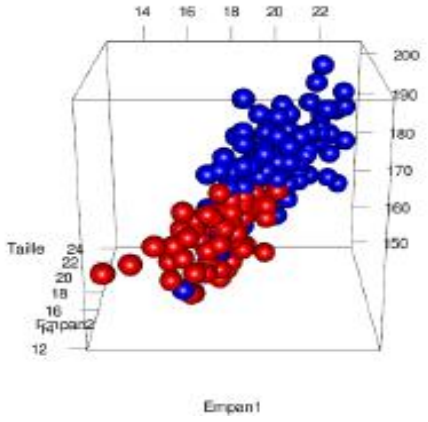
MISSIONS
SEGMENTATION

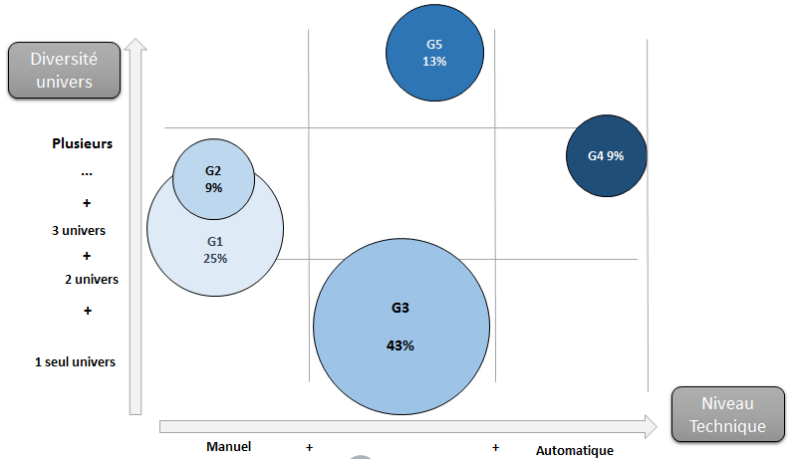
Contexte de la mission

Dans le cadre de la refonte d'un portefeuille produits ou d'une meilleure compréhension des habitudes clients, il s'agit de travailler sur les transactions brutes pour définir des **groupes de clients** homogènes en terme de comportement ou d'attentes.

Solution apportée

Calcul d'**indicateurs** de profiling pertinents en fonction des besoins métiers.
Analyse de corrélations et classifications des individus.
Comparaison des groupes obtenus et interprétation métier.





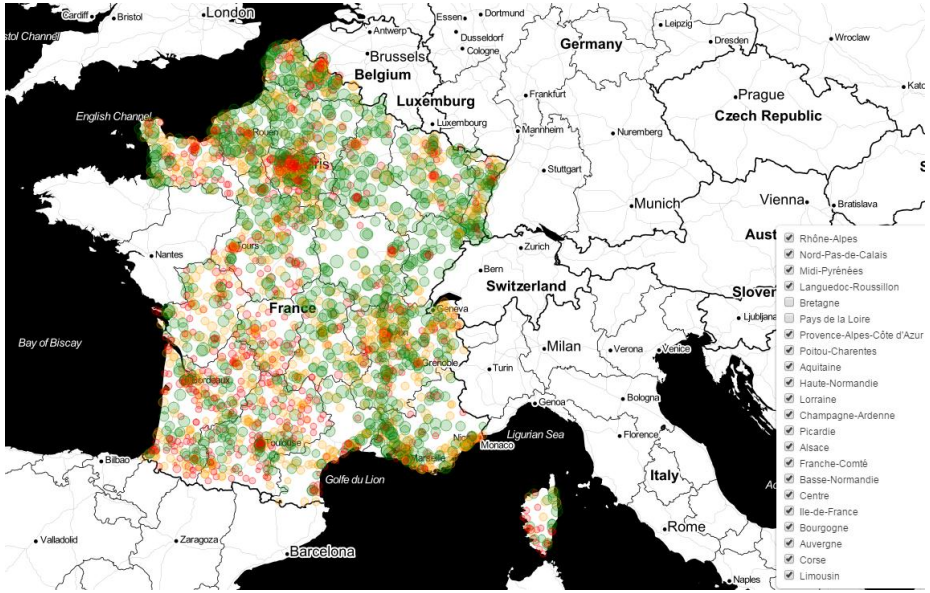
© Copyright 2015 – Keyrus

KEYRUS

MISSIONS
DATAVIZ

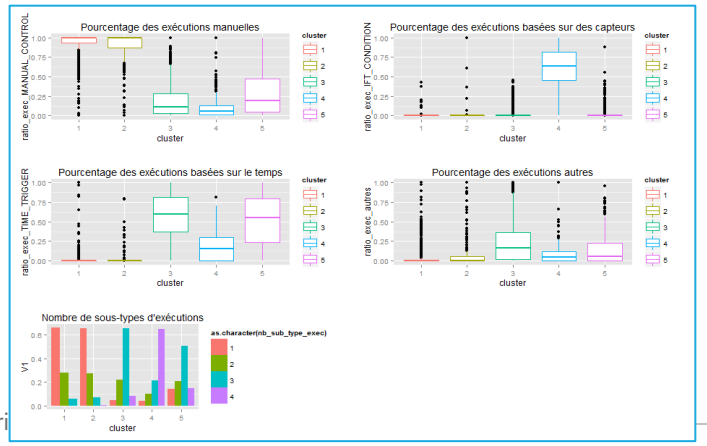
Contexte de la mission

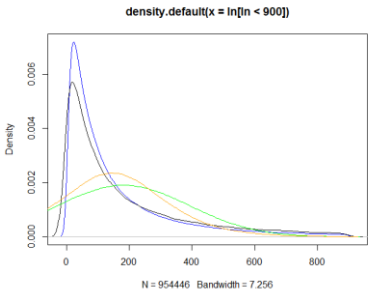
Pour **appréhender une nouvelle source de données complexes et variées**, il est nécessaire de visualiser les données individuellement, en les croisant afin de détecter des tendances, des valeurs aberrantes…

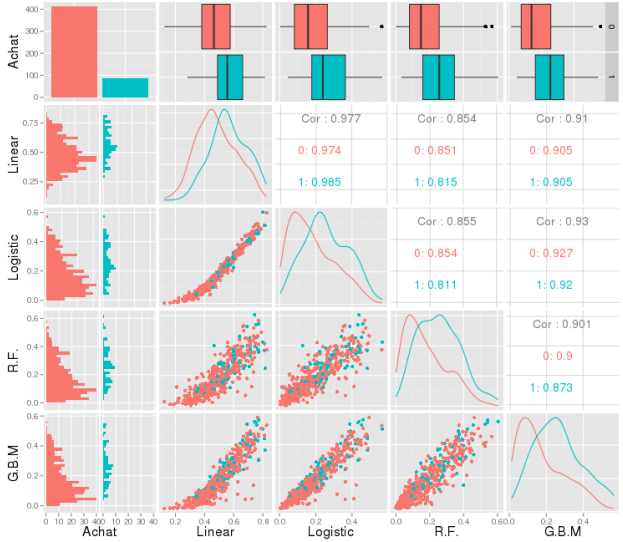


Solution apportée

Nous utilisons des fonctions de dataviz pour **représenter graphiquement** les données afin de comprendre les proximités entre les données (histogrammes, boxplot, cartographie, scatterplot…)







© Copyri

KEYRUS

MISSIONS

MAQUETTE - SHINY

Contexte de la mission

Dans le cadre du **processus de changement de pièces** d'un objet de haute technologie, l'objectif était d'extraire des **associations de pièces fréquentes** afin d'aider les acteurs du processus et accélérer le temps de changement.

Solution apportée

Calcul d'**indicateurs** basé sur les **modifications passées** associé à une interface R-Shiny simple et ergonomique pour les utilisateurs métier.
A partir d'une (ou plusieurs) première(s) pièce(s) à modifier, l'application suggère d'autres éléments ayant une forte probabilité d'être modifiés avec des indicateurs de **probabilité**, des options de **filtres**, de **description** et de **représentation graphique** (notamment interactive à partir d'éléments vis.js).

Technologie

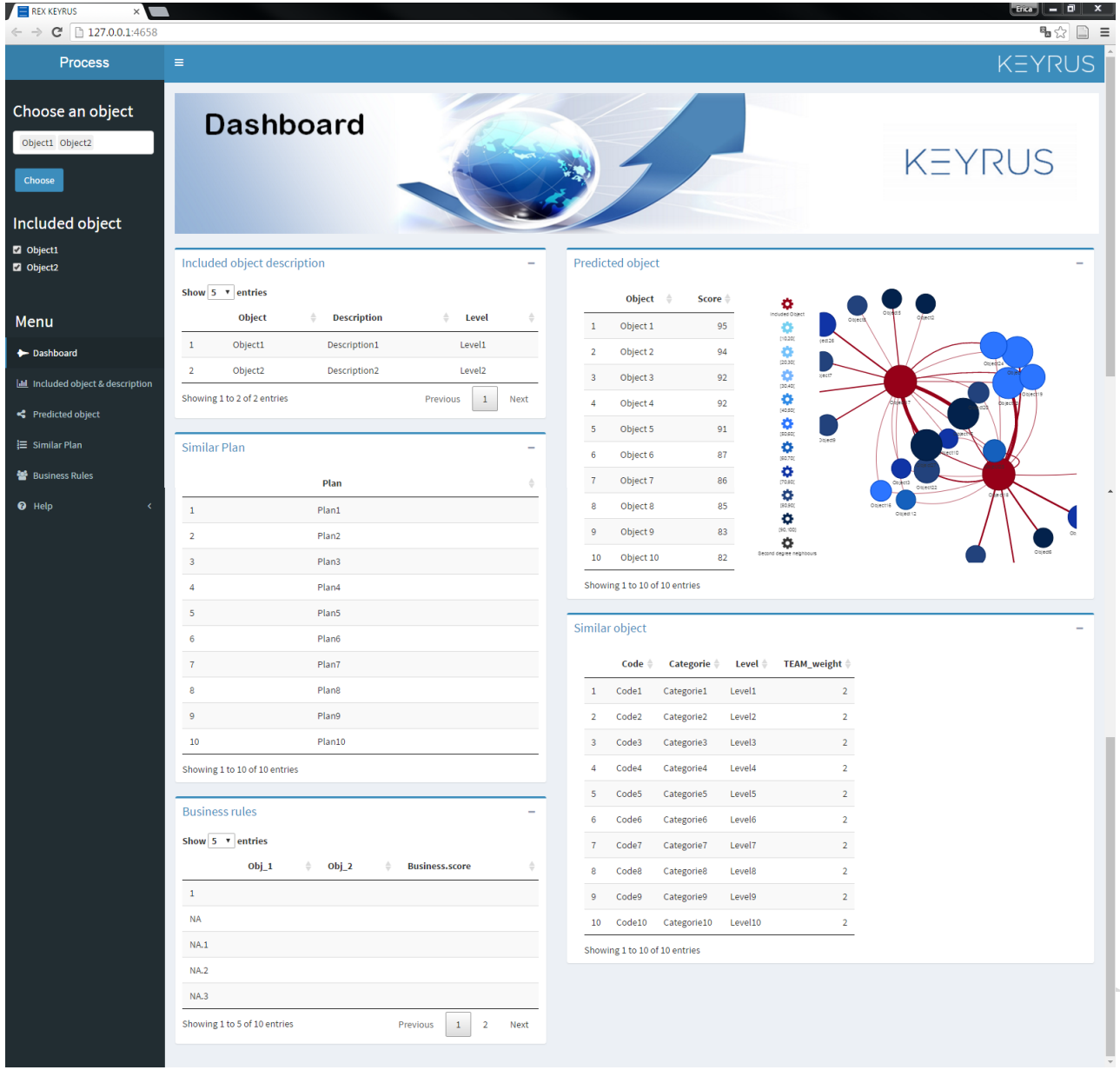
Langage

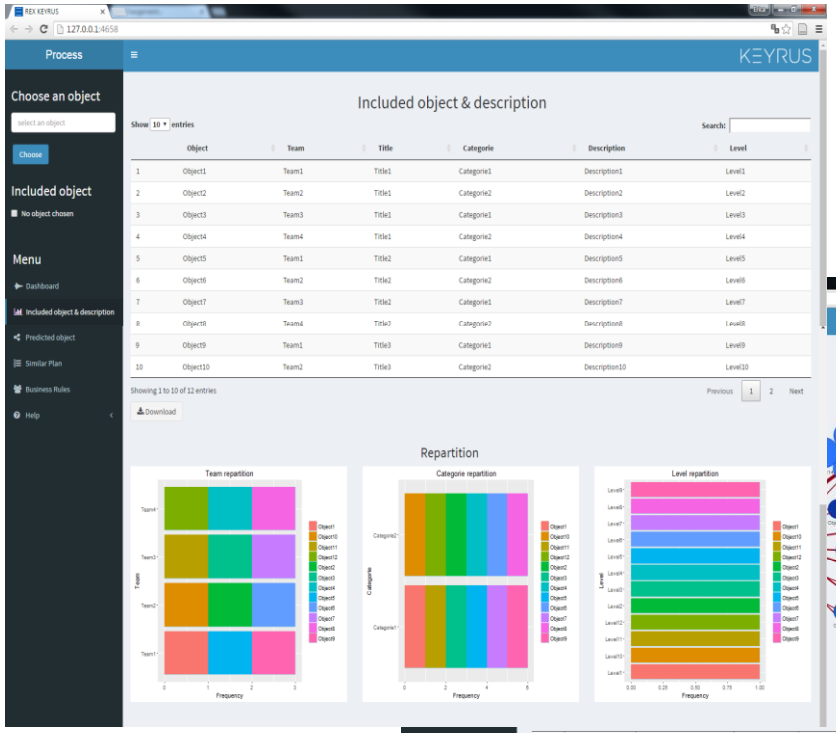
Ide

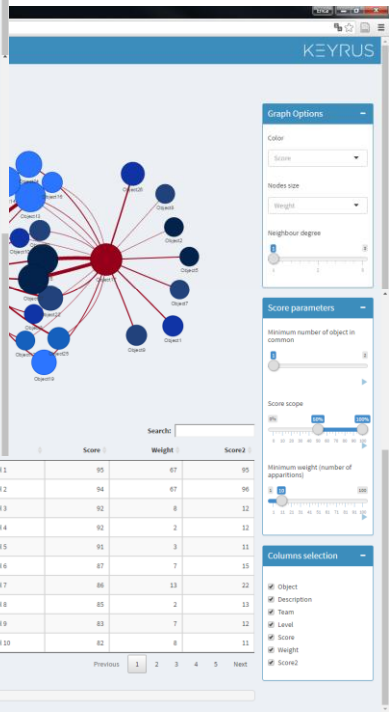
Principaux packages

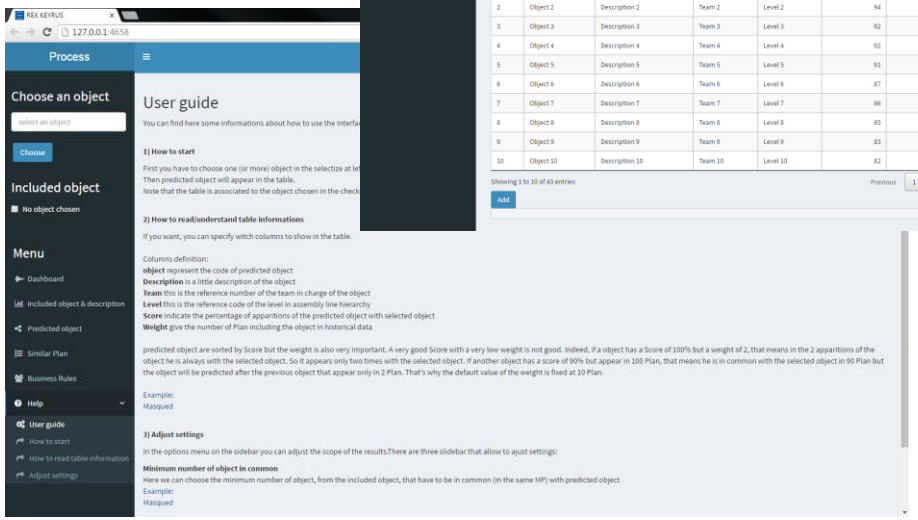
MISSIONS

MAQUETTE - SHINY









MISSIONS

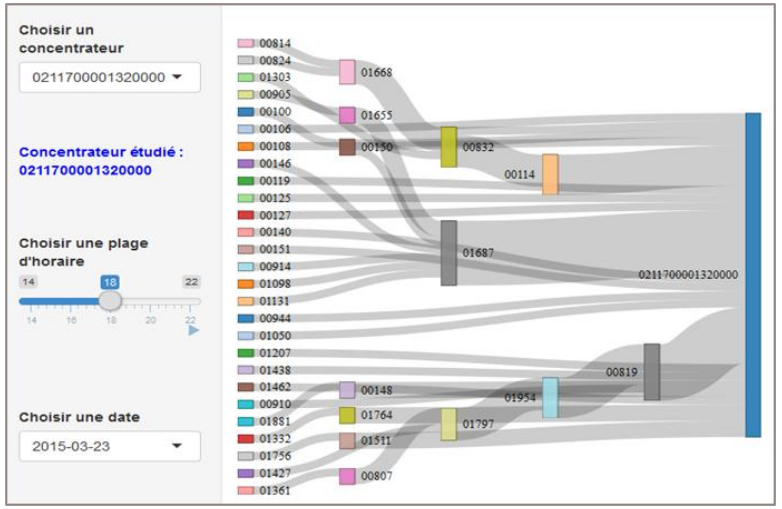
D3JS

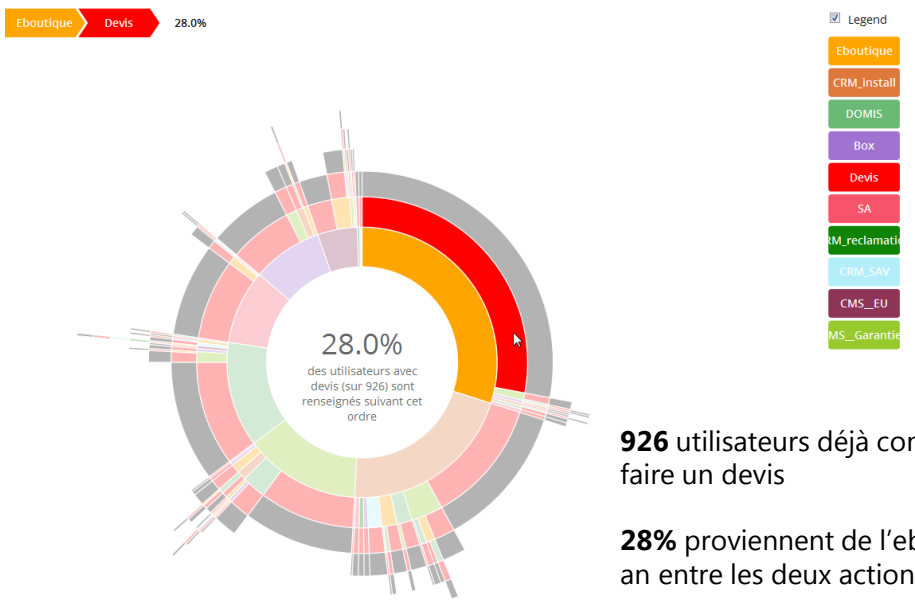
Contexte de la mission

Dans le cadre de l'exploration de nouvelles données ou de la présentation de résultats, nos clients demandent de plus en plus **un accès à la donnée travaillée et des visualisations interactives.**

Solution apportée

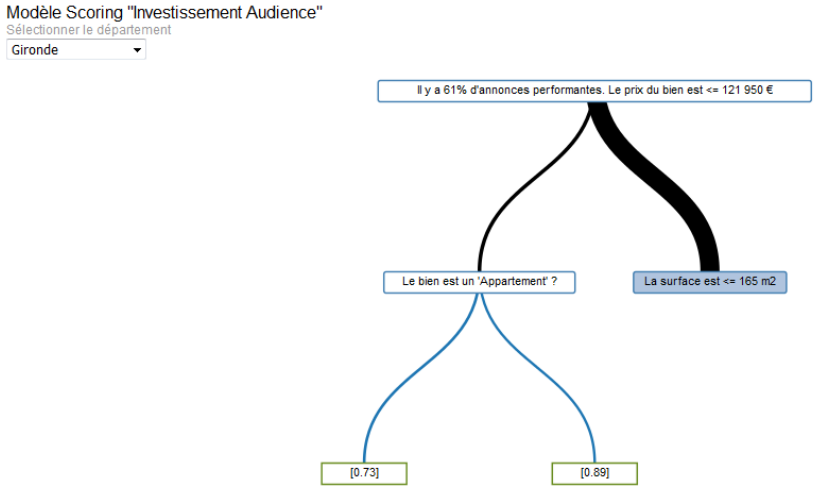
Préparation de fonctions R générant des visualisations d3js.





926 utilisateurs déjà connus avant de faire un devis

28% proviennent de l'eboutique (plus d'1 an entre les deux actions)



© Copyright 2015 – Keyrus

KEYRUS

MISSIONS

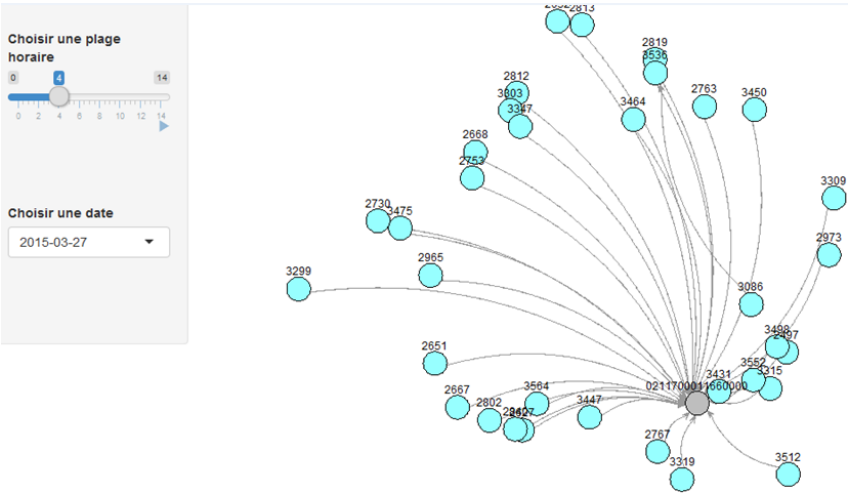
GRAPH MINING

Contexte de la mission

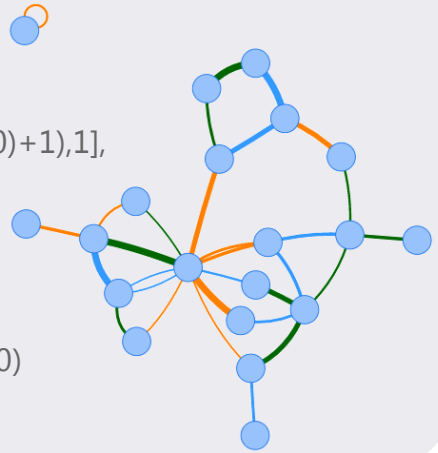
Dans des systèmes complexes (réseaux, interactions…), il s'agit de représenter les forces des relations afin de pouvoir avancer dans la **recherche d'une panne ou d'une interaction forte.** Les interactions sont modélisées selon un modèle mathématique (calcul des centralités…)

Solution apportée

Utilisation des packages VisNetwork et rpart




```
edges=data.frame(from=nodes[floor((runif(30)*20)+1),1],
    to=nodes[floor((runif(30)*20)+1),1],
    title=c(1:30),
    width=rep(c(3,4,5,6,9,12),each=5),
    length=rep(c(50,40,30),each=10),
    color=rep(c("#FF8000","#3399FF","#006600"),10)
)
```



© Copyright 2015 – Keyrus

MISSIONS

PARSING

Contexte de la mission

Le client possède des fichiers avec des formats hétérogènes et a **besoin de les uniformiser**. Il s'git alors de définir le format de restitution et de modifier les sources en conséquence.

Solution apportée

Il a été défini une **batterie de fonctions** « type » (mapping, transposition, transformation) permettant de modifier un nouveau fichier afin d'automatiser au maximum les modifications.

Utilisation des packages stringr

Transposition

	Label 1	Données 1	Label 2	Donnée 2
Ref 1	longeur	3,5	rawWeight kg	0,03
Ref 2	taille du corps	21-janv	rawWeight kg	1,20

	longeur	rawWeight kg	taille du corps
Ref 1	3,5	0,03	
Ref 2		1,2	21,00

Mapping

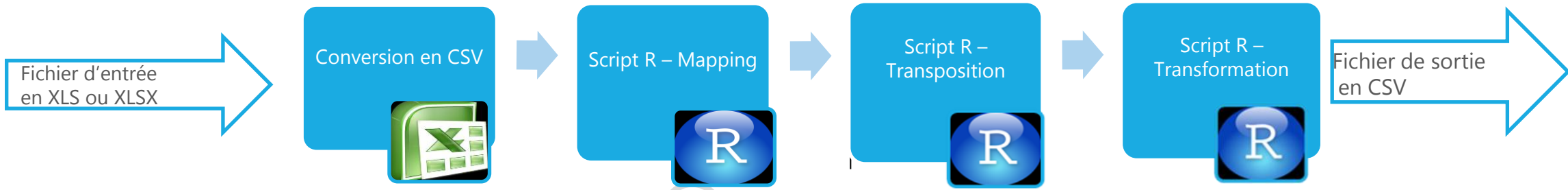
	longeur	rawWeight kg	taille du corps
Ref 1	3,5	0,03	
Ref 2		1,2	21,00

	longeur	Poids brut	Taille
Ref 1	3,5	0,03	
Ref 2		1,2	21,00

Transformation

	longeur	Poids brut	Taille
Ref 1	3,5	0,03	
Ref 2		1,2	21,00

	longeur	Poids brut (kg)	Taille
Ref 1	3,5	30	
Ref 2		1200	21,00



© Copyright 2015 – Keyrus

MISSIONS

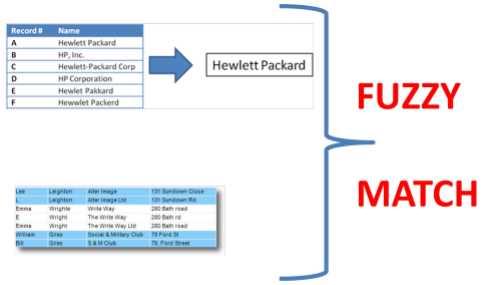
FUZZY MATCHING

Contexte de la mission

Certains clients ont besoin de **réconcilier différentes sources de données avec des nomenclatures différentes**. Il s'agit alors de reconnaître des proximités entre les bases à l'aide de distances à concevoir.

Solution apportée

Nous utilisons des approches de « fuzzy matching » (package **stringdist)** pour rapprocher différentes sources (nom, prenom, identifiant, adresse…)



Fuzzy Match Graphic by Megter.com

```
# Method 2: applying different string matching methods #osa Optimal string aligment, (restricted Damerau-Levenshtein distance). #lv Levenshtein distance (as in R's native adist). #dl Full Damerau-Levenshtein distance. #hamming Hamming distance (a and b must have same nr of characters). #lcs Longest common substring distance. #qgram q-gram distance. #cosine cosine distance between q-gram profiles #jaccard Jaccard distance between q-gram profiles #jw Jaro, or Jaro-Winker distance.   #install.packages('stringdist')
library(stringdist)
 distance.methods<-c('osa','lv','dl','hamming','lcs','qgram','cosine','jaccard','jw')
 dist.methods<-list() for(m in 1:length(distance.methods)) { dist.name.enh<-matrix(NA, ncol = length(source2.devices$name),nrow = length(source1.devices$name)) for(i in 1:length(source2.devices$name)) { for(j in 1:length(source1.devices$name)) { dist.name.enh[j,i]<-stringdist(tolower(source2.devices[i,]$name),tolower(source1.devices[j,]$name),method = distance.methods[m]) #adist.enhance(source2.devices[i,]$name,source1.devices[j,]$name) } }
dist.methods[[distance.methods[m]]]<-dist.name.enh }
```

© Copyright 2015 – Keyrus

MISSIONS

API

Contexte de la mission

En amont de missions « analytics », il est parfois utile **d'enrichir certaines données** par de l'open-data : météo, web analytics…

Solution apportée

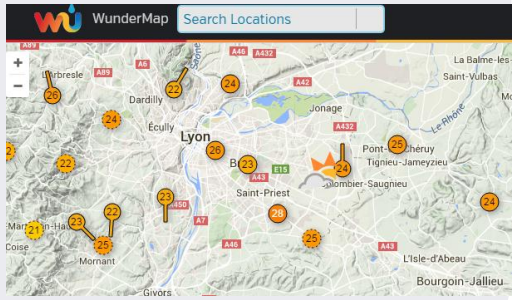
Utilisation de **packages/fonctions pour se connecter aux API**

Ex: RSiteCatalyst, WeatherUnderground

```
In [2]: import pandas as pd

        import rpy2.robjects.packages as rpackages
        from rpy2.robjects import pandas2ri

        # Activate ability to translate R objects to pandas data frames
        pandas2ri.activate()

        # Load RSiteCatalyst into Python
        rsc = rpackages.importr('RSiteCatalyst')

In [5]: # Set authentication credentials
        rsc.SCAuth('username', 'secret');

        [1] "Credentials Saved in RSiteCatalyst Namespace."

In [6]: # Call GetReportSuites to confim it works
        grs = rsc.GetReportSuites()
        pandas2ri.ri2py_dataframe(grs)

Out[6]:
```

	rsid	site_title
1	zwitchdev	ZwitchCorp Dev
2	zwitchjulia	Julia Bloggers

```
getHistoricalWeather <- function(airport.code="SFO", date="Sys.Date()")
{
  base.url <- 'http://api.wunderground.com/api/{your key here}/'
  # compose final url
  final.url <- paste(base.url, 'history_', date, '/q/', airport.code, '.json', sep='')

  # reading in as raw lines from the web service
  conn <- url(final.url)
  raw.data <- readLines(conn, n=-1L, ok=TRUE)
 # Convert to a JSON
  weather.data <- fromJSON(paste(raw.data,
collapse=""))
  close(conn)
  return(weather.data)
}
```



	feed_id	rsid	name	interval	timezone	status	ftp_host	ftp_path	fields
1	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	accept_language
2	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	browser
3	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	connection_type
4	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	country
5	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	date_time
6	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	domain
7	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar1
8	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar2
9	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar3
10	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar4
11	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar5
12	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar6
13	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar7
14	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar8
15	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar9
16	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar10
17	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar11
18	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar12
19	1109801	zwitchdev	RSiteCatalyst Data Feed	daily	US/Eastern	Active	ftp.juliabloggers.com	/	evar13

© Copyright 2015 – Keyrus

AGENDA

BIG DATA & ANALYTICS chez KEYRUS REGION

POSITIONNEMENT DE R DANS NOS MISSIONS

TYPOLOGIES DE MISSION

CONCLUSIONS

© Copyright 2015 – Keyrus

PANORAMA « ANALYTICS »

POSITIONNEMENT R POUR NOS CLIENTS

FORCES

Accès total aux algorithmes et à leurs implémentations

Liberté de redistribuer des copies de programme

Communauté : fournit un forum permettant aux chercheurs d'explorer et de développer des méthodes

Ce sont le plus souvent **des experts** de leur domaine qui proposent des packages

FAIBLESSES

Interaction graphique limitée avec l'interface par défaut

Pas de presse-bouton (il faut tout coder), mais il est possible d'utiliser l'interface graphique « Rcommander »

Pour traiter de gros volumes de données, il est nécessaire de mettre en place **une architecture adaptée et d'optimiser son code**
➔ In-Memory / Nécessité de réaliser un dimensionnement précis des ressources en fonction de la volumétrie et des besoins

Langage interprété, peut être lent sur les instructions itératives (boucle for)

En vue d'une industrialisation, l'absence d'un **support commercial** (garantie) et de **support technique direct** (type hotline)

© Copyright 2015 – Keyrus

PANORAMA « ANALYTICS »

POSITIONNEMENT R

Etude analytique avec livraison de rapports / modélisations optimales **=> OK**

Réalisation de maquette / POC / démonstration **=> OK**

Mise en production

=> OK si couplé à des serveurs dimensionnés

=> KO si pas de serveur adapté => Python / Hadoop / Dev

Passation auprès du client métier

=> OK si ressources « data »

=> KO si peu d'appétence => outils plus ergonomiques avec flux d'analyse

© Copyright 2015 – Keyrus