```
library(dplyr)

r-ladies_global %>%
  filter(city = 'Dublin')
```

R-Ladies

TOUR OF R

SONYA ABBAS
Data Scientist - AdaptiveMobile
sonya.abs@gmail.com

1.
Introduction to R

What - Why – Example: R for Data Science

What is R

- Language and environment for statistical computing and graphics
- Created by Ross Ihaka and Robert Gentleman - University of Auckland, New Zealand (conceived 1992 - released 1995 - stable beta version 2000)
- Similar to the S language and environment - Bell Laboratories by John Chambers and colleagues (Old S: ~1975, New S: ~1988)
- Open source project

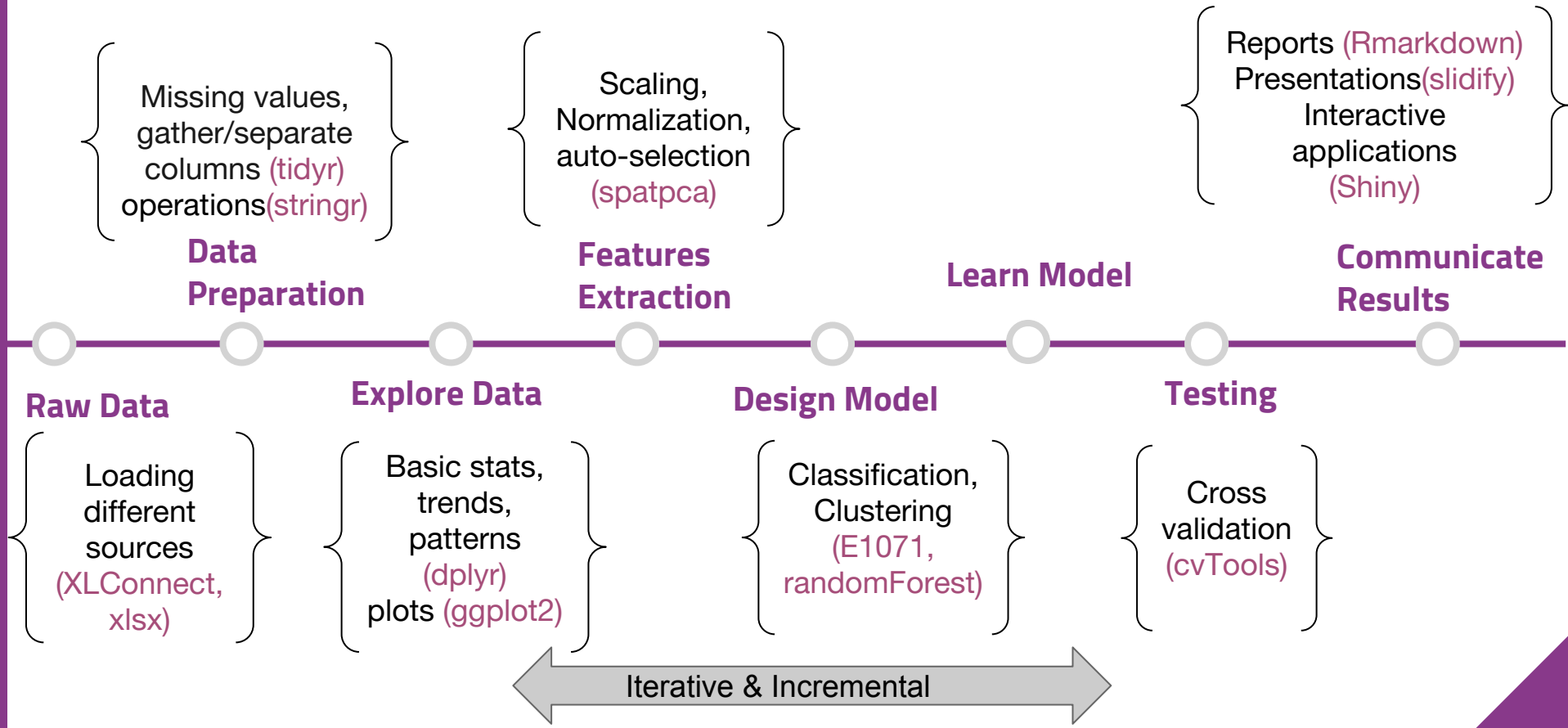
Statistical computing in R

- Linear and nonlinear modelling
- Classical statistical tests
- Time-series analysis
- Classification
- Clustering
- Graphical techniques

Why R

- R is free
- Access to cutting-edge technologies
- Useful skill
- Well-developed
- Can be extended via packages
- Effective data handling and storage facility
- Graphical facilities for data analysis

Data Science Pipeline



2.
Closer Look into R

R foundation - CRAN – RStudio - Conferences – Useful materials

R Foundation - CRAN - RStudio

R Foundation

- Non profit organization that support R project, provide reference point for all community members

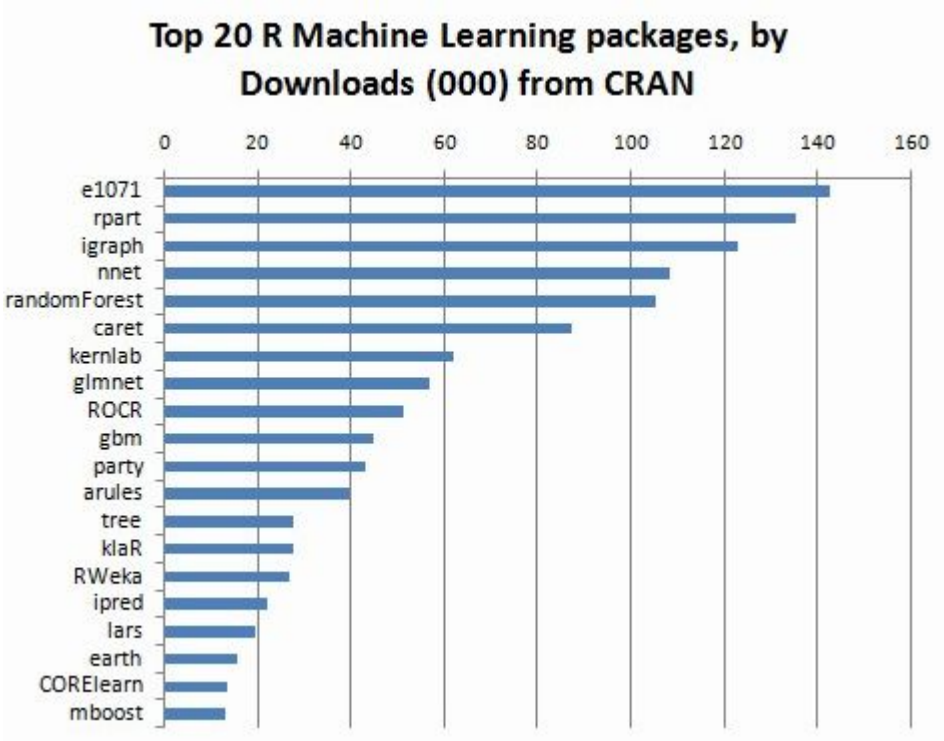
CRAN

- Comprehensive R archive network to store identical, up-to-date, versions of code and documentation for R
- Download and Install R (Linux, Mac and Windows)
- CRAN package repository features 9,952 available packages

RStudio

- IDE for R, code editor, debugging and visualisation tools

R Packages



Geethika Bhavya Peddibhotla blog, Kdnuggets 2015

To load data
(RODBC, RMySQL, XLConnect)

To manipulate data
(dplyr, tidyr, stringr)

To visualize data
(ggplot2, ggvis)

To model data
(car, randomForest, caret)

To report results
(shiny, R markdown, xtable)

For Spatial data
(sp, maptools)

For Time Series and Financial data
(zoo, xts, quantmod)

Garrett Grolemund, quick list of useful R packages 2017

R Conferences - useR!

- International R User Conference where R users and developers meet – June
- Invited talks: technical, computing issues, topics of current interest
- User-contributed presentations and posters: R-related topics

R Conferences - DSC

- Directions in Statistical Computing where developers and researchers in statistical software and computing meet
- Started in1999 to 2009 - open registration, calls for papers and peer-reviewed conference proceedings
- Restarted in 2014 as an annual conference coinciding with the General Assembly of the R Foundation but by invitation only
- Topics: big-data extensions, database interfaces, graphical subsystems, and user interfaces and scientific computing

R Useful Materials

- **Manuals:** an introduction to R, R Data Import/Export, Writing R Extensions, etc.
- **R Journal:** open access, short to medium length articles covering interesting topics
- **R news:** changes in R, CRAN, coming conferences and conference reports
- **Books:** Learning Base R, an Introduction to R for Quantitative Economics: Graphing, Simulating and Computing, empirical Research in Economics: Growing up with R, introduction to data science, elements of statistical learning, etc.
- **Other:** online courses (coursera), swirl package for interactive course, RStudio, Kaggle, UCI machine learning repository (360 data sets) – dataCamp, etc.

3.
My Experience using R

Contribution to UseR! 2015 – recommendation systems – auto completion text writing

E-governments Action Plans Clustering

Centre for Data Analytics

Insight

Sonya Abbas and Adegboyega Ojo

Problem

- The pressure of evaluating and improving the government's actions plans
- The need to evaluate the progress of governments as basis for assistance from organizations such as world bank
- The difficulties of learning experience from other countries
- Challenges of discovering similarities between countries action plans

The World Bank

Open Government Partnership

Methodology

- **Getting Data:**
  - Text documents descriptions of action plans from OGP.
  - Different languages
  - Different formats
- **Preparing Data:**
  - Unify documents format to txt
  - Translate documents to english
- **Analysing Data:**
  - Remove spaces, punctuations, numbers and countries names
  - Create similarity matrix(40 000 features) using tri gram tokenizer
  - Optimize features by removing ones appears in one document only: output is optimized matrix with 647 features
  - Hierarchical clustering: construct distance matrix (hclust func)
  - Kmeans Clustering: contains matrix normalization, clustering with 5 number of clusters fixed (kmeans func),
- **Visualize Data:**
  - Kmeans Clustering output as plots
- **Interpret data:**
  - Egove experts interpret the relations between the countries action plans and the categoreis of the countries.
  - Compare the results to the clusters we get from OGP indicators

Approach

Step 1:

- Hierarchical clustering based on OGP indicators
- For each category, we apply latent dirichlet allocation (LDA) for topic modeling: 1) topic extraction for action plans docs, 2) topic extraction for challenges, 3) topic extraction for commitments.
- Adjust parameters and Analyse results by eGov experts
- In parallel, we do qualitative data analysis for topic extraction over the action plans using Nvivo.

Step 2:

- Hierarchical clustering for all countries based on similarities – distance matrix
- K-means clustering
- Compare the results – clusters from both methods
- Visulize the results and show relations between terms , countries and terms - countries

Implementation

- We use R in order to implement the work
- We visualize using different packages such as ade4, ggplot2, ellipse, HSAUR and flexclust

Results

- Countries clusters

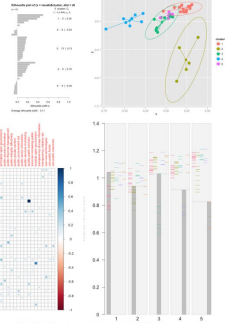
Cluster 0: Canada, Denmark, Finland, Israel, Netherlands, Norway South Korea, Sweden, UK, US.

Cluster 1: Albania, Armenia, Azerbaijan, Dominican Republic, Guatemala, Honduras, Indonesia, Kenya, Moldova, Paraguay, Philippines, Tanzania, Ukraine.

Cluster 2: Chile, Czech Republic, Estonia, Lithuania, Malta, Slovak Republic, Spain, Uruguay.

Cluster 3: Bulgaria, Croatia, Greece, Italy, Latvia, Montenegro, Romania, South Africa.

Cluster 4: Brazil, Colombia, El Salvador, Georgia, Jordan, Macedonia, Mexico, Peru, Turkey



Emails: sonya.abbas@insight-centre.org adegboyega.ojo@insight-centre.org

DCU NUI Galway OÉ Gaillimh UCC UCD sfi

Data

E-government action plans
World bank and OGP indicators

Analysis

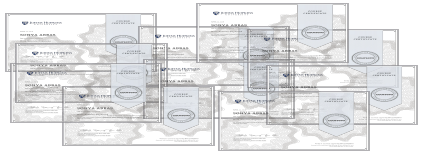
Topic modeling
Hierarchical clustering - K-means clustering
Regression models

Results

Give recommendation to governments based on experiences from other governments belong to same group.

R packages used

ade4, ggplot2, hsaur, flexclust, topicmodels



Auto Completion Text Writing

Enter Your Text:

hi, how are

Go!

Lazy to continue writing! no worries, Click this button to auto complete your text.

Hello, Here is the results of our prediction:

hi, how are you

Data

Datasets from SwiftKey- covers blogs, news and tweets

Analysis

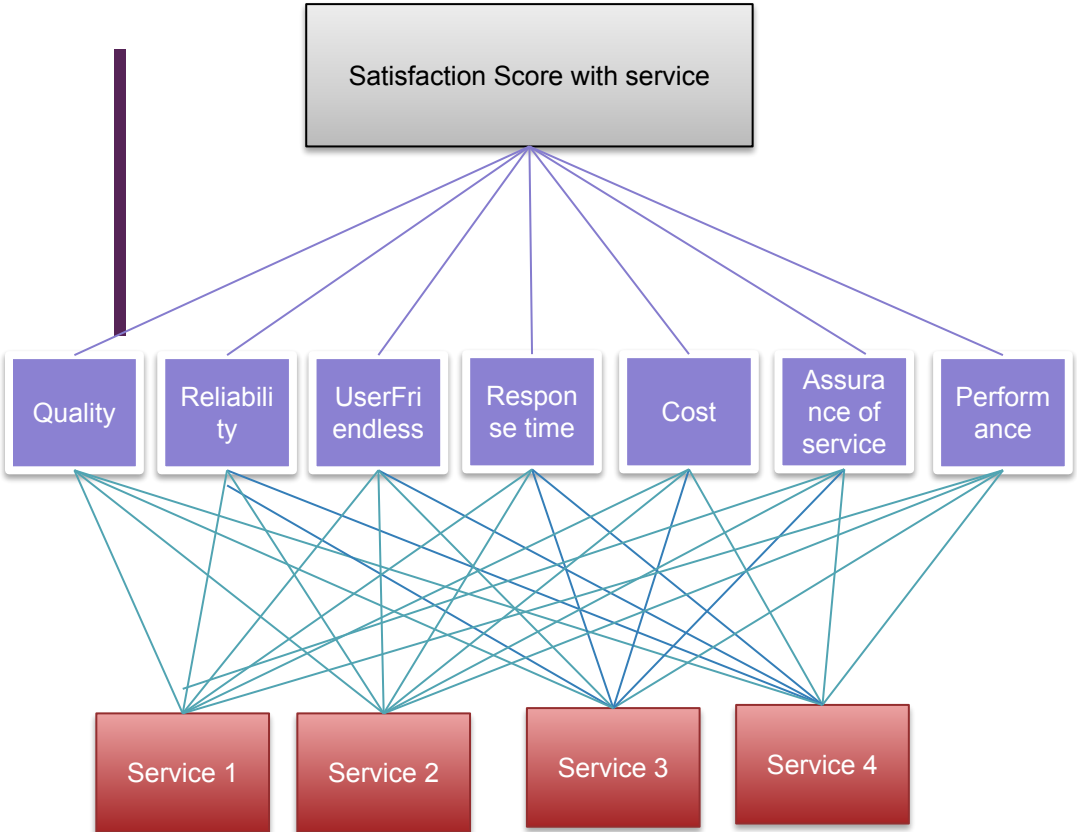
Probabilistic language model

Results

Auto completion text writing shiny app

R packages used

Stringr- topicmodels



Data

Synthetic data

Analysis

Multicriteria decision making approach

Results

Recommendations for users based on their preferences

R packages used

Pmr package (AHP), open cpu API for data analysis based on R

Thank You

sonya.abs@gmail.com