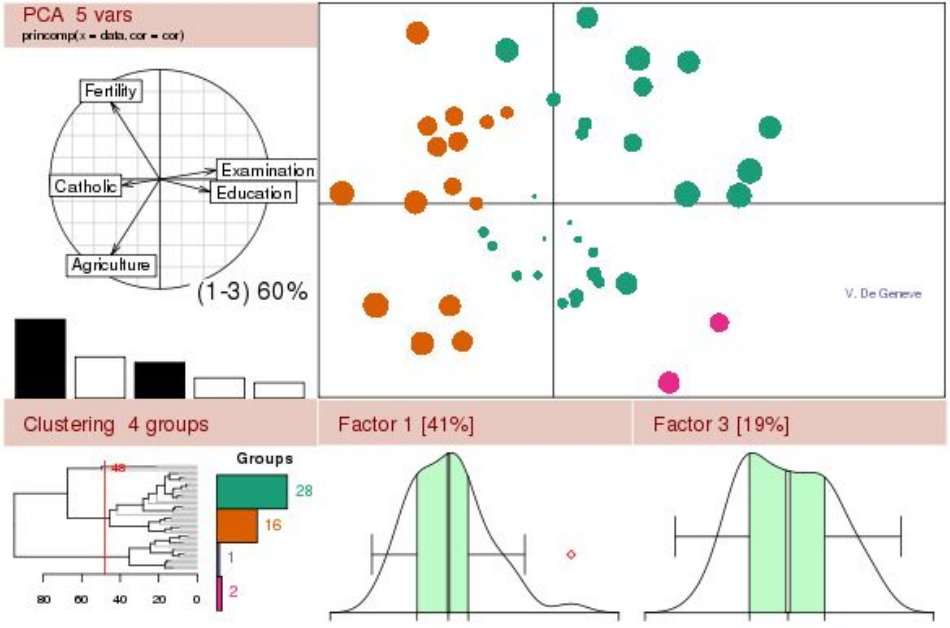
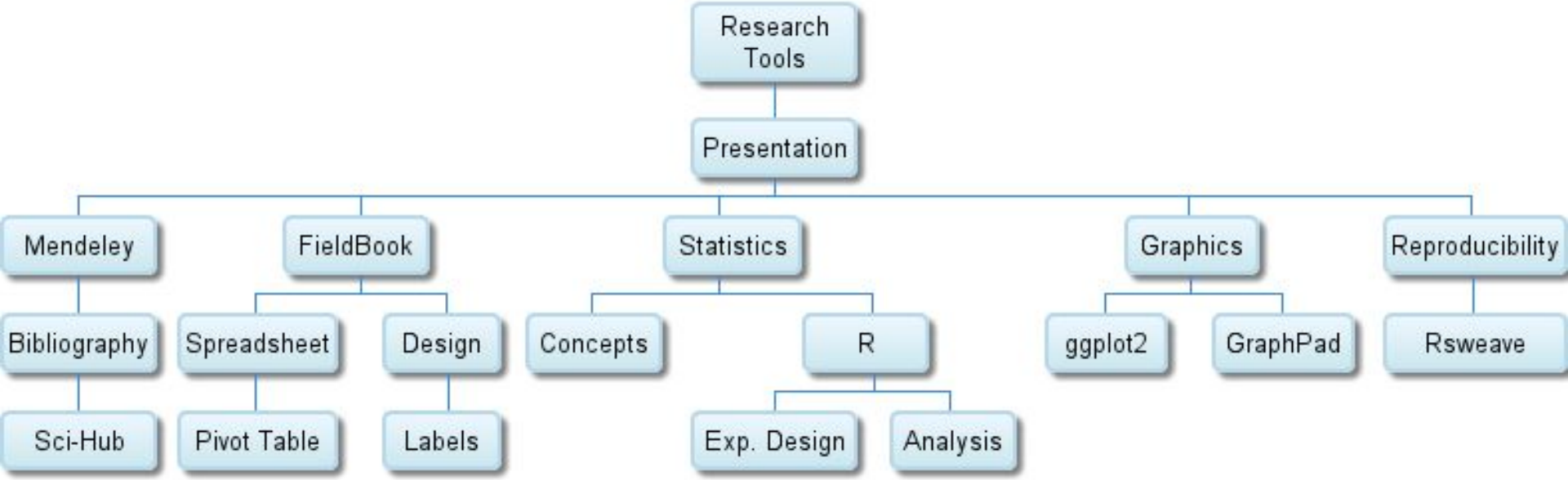


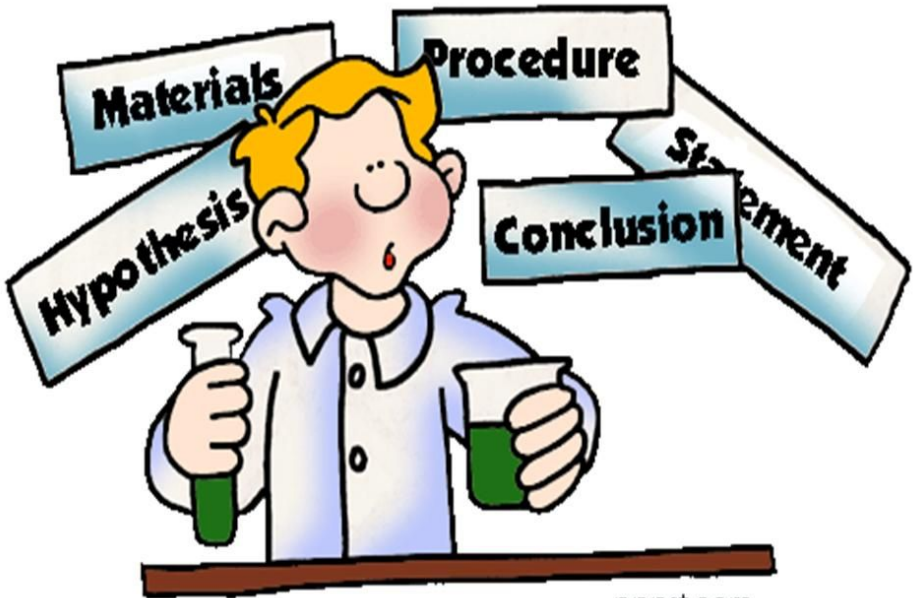
Herramientas para la Investigación Científica

Flavio Lozano Isla



Course Summary

Objetivos del curso



El método científico

Comunicación

UN ANÁLISIS BRILLANTE RESULTA INÚTIL SI LOS RESULTADOS NO SON COMUNICADOS CON UN LENGUAJE SENCILLO Y CLARO, INCLUIDO EL GRADO DE INCERTIDUMBRE ESTADÍSTICA DE LAS CONCLUSIONES. POR EJEMPLO, EN LA ACTUALIDAD, LOS MEDIOS DE COMUNICACIÓN PUBLICAN MÁS A MENUDO LOS MÁRGENES DE ERROR DE LOS RESULTADOS DE LAS ENCUESTAS QUE REALIZAN.

Statistical Language Wars

SAS vs R vs SPSS

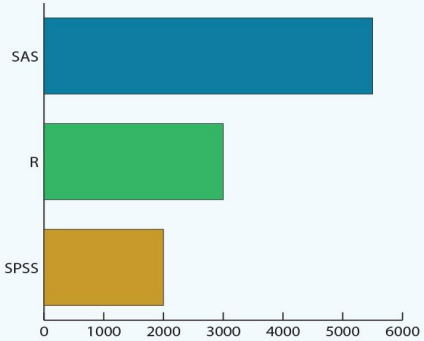
To the amusement of many, programmers are continuously debating fiercely about which programming language is better. Until recently, these "rivalries" were a specific characteristic to the computer science industry, but today the analytics industry community is catching-up. Huge debates on what should be the statistical programming language of choice are occupying forums and meet-ups. Here we compare **SAS**, **R** and **SPSS** to see how they stack up.

	SAS	R	SPSS
HISTORY	**Creator**: Jim Goodnight and Jim Barr, North Carolina State University **Year Released**: mass distributed since 1972 **Must Knows**: • SAS started because of a need for a computerized statistics program to analyze vast amounts of agricultural data • The SAS institute was founded in 1976 and currently has 13,733 employees • In 2013, SAS invested 25% of revenue in R&D	**Creator**: Ross Ihaka and Robert Gentleman, University of Auckland, New Zealand and the R foundation **Year Released:** 1995 **Must Knows**: • R is an implementation of the S programming language created at Bell Labs • The design and evolution of R is controlled by the R-core group and R foundation • The source code for the R software environment is written primarily in C, Fortran and R.	**Creator**: Norman H. Nie, Dale H. Bent, and Hadlai "Tex" Hull **Year Released**: 1968 **Must Knows**: • In 1976 SPSS jeopardized the University of Chicago's status as a tax-exempt organization • SPSS was acquired by IBM in 2009 for US$1.2 billion • In 1993 SPSS was taken public on the NASDAQ exchange

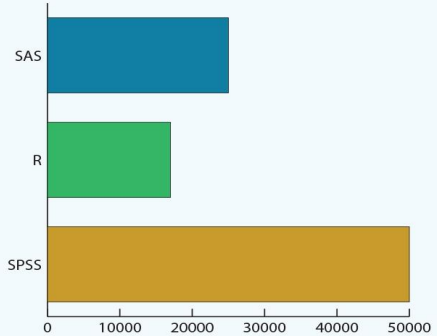
	SAS	R	SPSS
PURPOSE AND USABILITY	• SAS accumulated since the 1970s a large amount of high-quality production code for multiple purposes • SAS has a strong leading position in the commercial analytics space. Code legacy plays an important role here • SAS has strong data handling capabilities. Furthermore, it releases its software updates in a controlled environment, which make them well tested. Nevertheless, SAS is an expensive solution.	• R has been used in academics and research for a long time. Today, its finding its way into commercial applications as well. See R as the open-source counterpart of SAS. • R has advanced graphical capabilities thanks to for example packages like ggplot2, googleVis and rCharts. • Due to its open-source nature, R has a large and supportive community. The latest techniques are developed and released quickly.	• SPSS is a great tool for non-statisticians since it has a user-friendly Interface and easy-to-use drop down menus. • Just like SAS, SPSS has a rather hefty price tag. • SPSS has applications in many fields, but mainly plays a leading role in social sciences.
COMPANIES USING IT	hp, Scotiabank, AIG, STAPLES	Google, FDA, The New York Times	ABN·AMRO, Canon, BT, Credit Suisse

	SAS	R	SPSS
EASE OF LEARNING	Although it's not like learning Microsoft Word, getting a basic understanding of how to work with SAS **shouldn't take you to long**. However, to become really good you will need to work through a lot of specifics. There are **many official and unofficial tutorials** available, and official certifications can be obtained via SAS training institutes.	R has a reputation for being hard to learn. Instead of setting up a complete analysis at once, R users need to **learn how to analyze data interactively**. For most data analysts, this is a mind shift they first need to undergo. The open-source community of R is rapidly lowering this learning curve by creating **high-quality introductory tutorials** and interactive coding tutorials.	SPSS is by far the **easiest to learn** among the 3 languages listed here. So if you only open a statistical program twice a month SPSS is the way to go. One of the biggest advantages in terms of learning is its **similarities with Excel**, something most of us are familiar with.

MARKETABILITY



Number of analytics jobs on Indeed.com 2/2014



Use of analytic software in academia 05/2013.
Based on number of google scholar hits

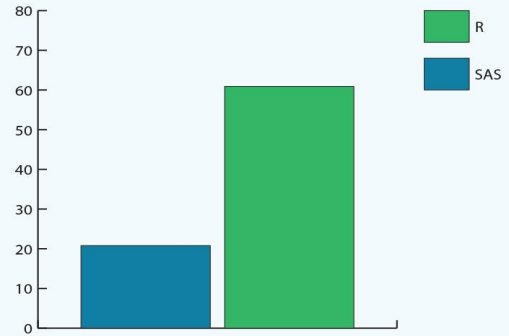
POPULARITY

kaggle

50% of Kaggle winners use R

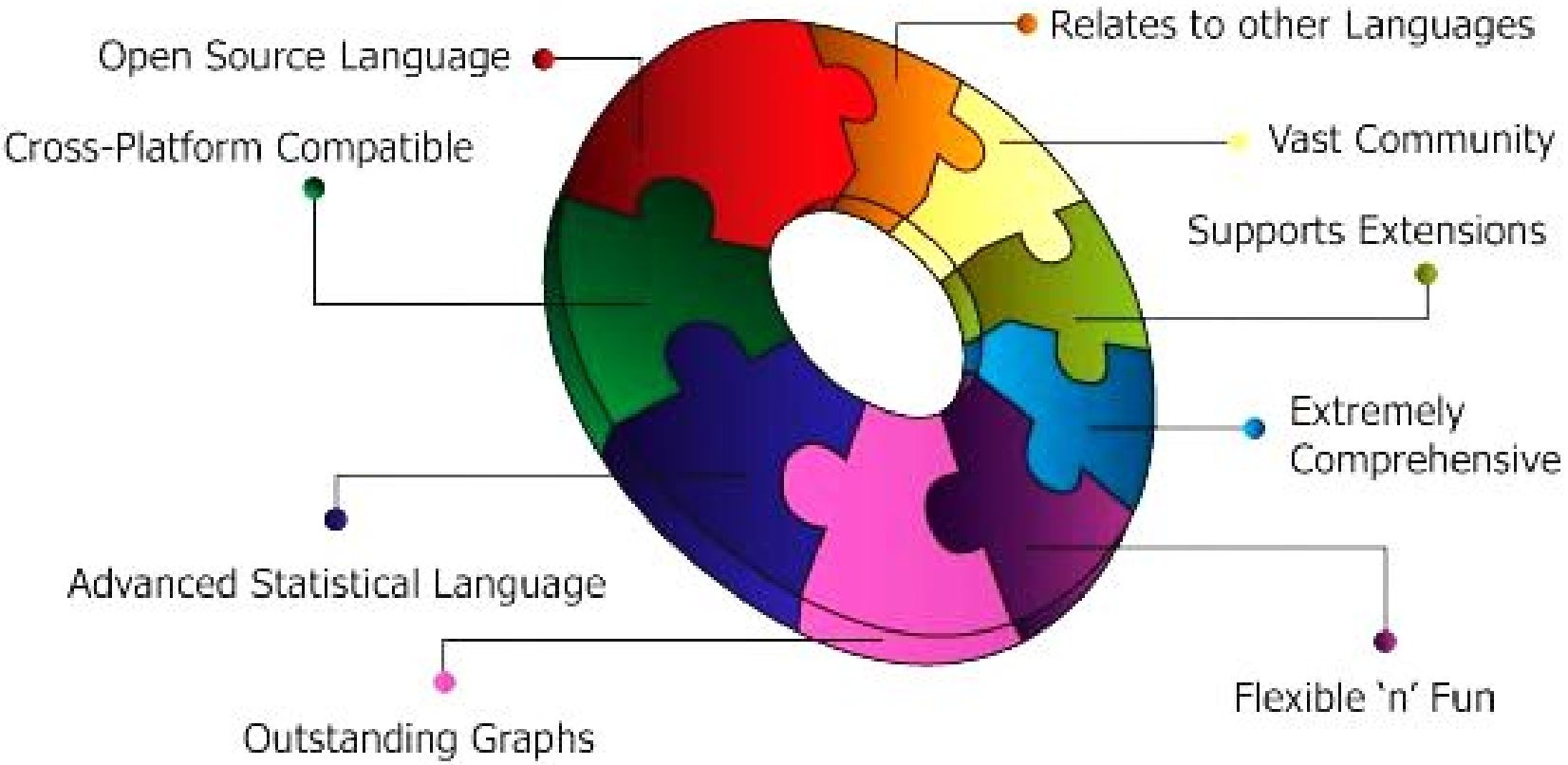
stackoverflow

The number of R related posts on Stack Overflow is more than 7-fold the number of posts on SAS



Percentage of "What programming/statistics languages you used for an analytics / data mining / data science work in 2013?" (KD nuggets)

Why Learn R?


Open Source Language
Relates to other Languages
Cross-Platform Compatible
Vast Community
Supports Extensions
Extremely Comprehensive
Advanced Statistical Language
Flexible 'n' Fun
Outstanding Graphs

R is FREE.

Software	Cost
STATA	$1,140 - $4,370 + maintenance
SAS	$8,700 - $140,000 / year
SPSS	$2,390 - $40,600 / year
MATLAB	$2,150 + $1,000s for modules
R	**$0**

Razones para usara R!

Thuner (2014): Introduction to R