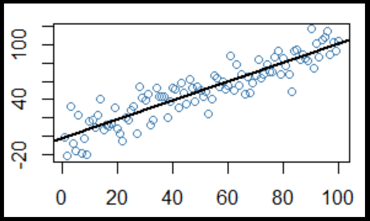
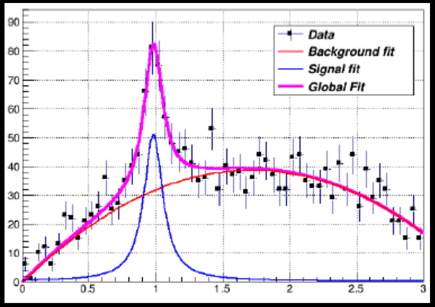
Temas avanzados en física computacional
Análisis de datos





Semestre

2016-I

Clase-1

José Bazo

jbazo@pucp.edu.pe

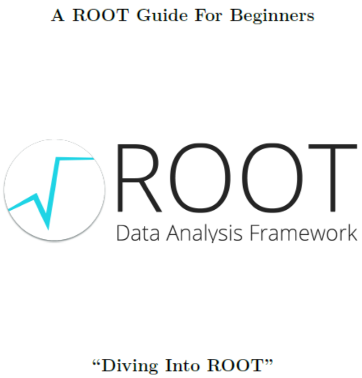
Contenidos del curso

1. Introducción al análisis de datos y data science
2. Lenguaje de programación R
3. ROOT Data Analysis Framework
4. Manipulación y visualización de datos
5. Modelamiento estadístico
6. Machine Learning
7. TMVA (Toolkit for Multivariate Data Analysis)

Bibliografía

Libros:

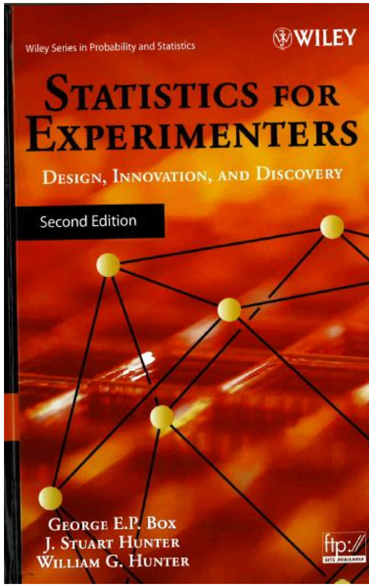
- Venables, Smith, et al. An Introduction to R. 2015
- Peng. R Programming for Data Science. 2015
- A ROOT Guide for Beginners, 2015.
- Hoecker et al. Toolkit for Multivariate Data Analysis with ROOT User´s Guide. 2007
- Box et al. Statistics for Experimenters. 2005
- Hastie et al. The Elements of Statistical Learning. 2008
- Witten, Frank & Hall. Data mining: practical machine learning tools and techniques. 2011.



A ROOT Guide For Beginners

ROOT
Data Analysis Framework

"Diving Into ROOT"

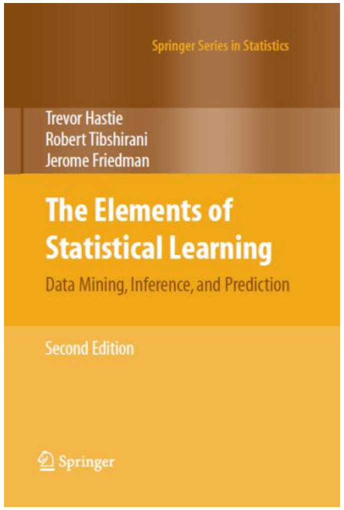


Wiley Series in Probability and Statistics
WILEY

STATISTICS FOR EXPERIMENTERS
DESIGN, INNOVATION, AND DISCOVERY

Second Edition

GEORGE E.P. BOX
J. STUART HUNTER
WILLIAM G. HUNTER



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning
Data Mining, Inference, and Prediction

Second Edition

Springer

R Programming for Data Science

Roger D. Peng

Bibliografía

Internet:

Coursera: Data Science at Johns Hopkins University:
https://www.coursera.org/specializations/jhu-data-science

The Comprehensive R Archive Network
https://cran.r-project.org/

ROOT
https://root.cern.ch/

Stack Overflow
http://stackoverflow.com/

Kaggle
https://inclass.kaggle.com/

Material en Intranet

Los archivos de la clase se encontrarán en Intranet del curso

Sistema de Evaluación

En este curso se aplica la modalidad de nota única.

Nota Final = (1 Par + 2 Ta + 3.5 Ex + 3.5 Pre) / 10

Par=Participación en clase

Ta=Promedio de 3 tareas

Ex= Examen

Pre=Presentación del trabajo

Cronograma

Tarea 1: 8 abril
Tarea 2: 22 abril
Tarea 3: 13 mayo

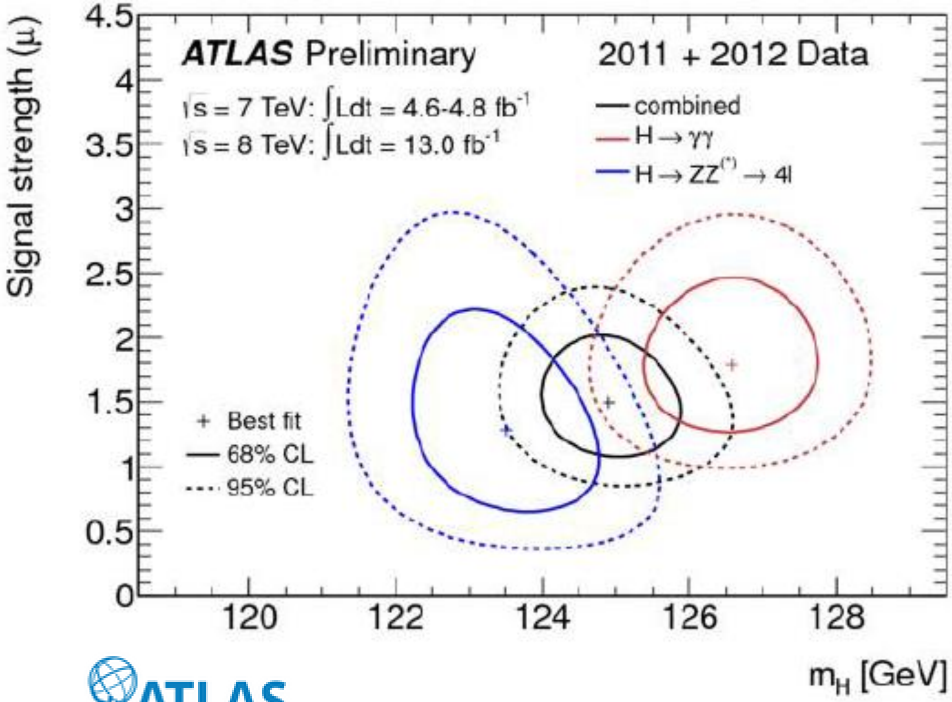
Examen: 27 mayo

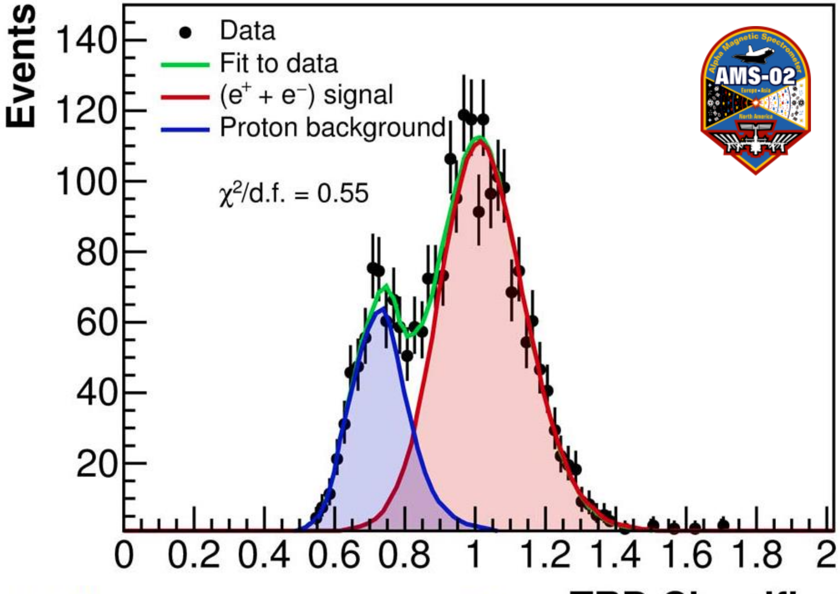
Presentación Trabajo: 1 julio

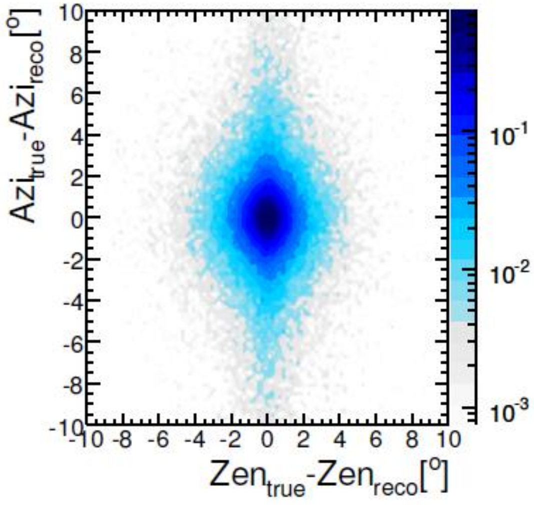
Asesorías: Oficina 305, tercer piso Edificio Física

1. Introducción al análisis de datos y data science

Análisis de datos con ROOT





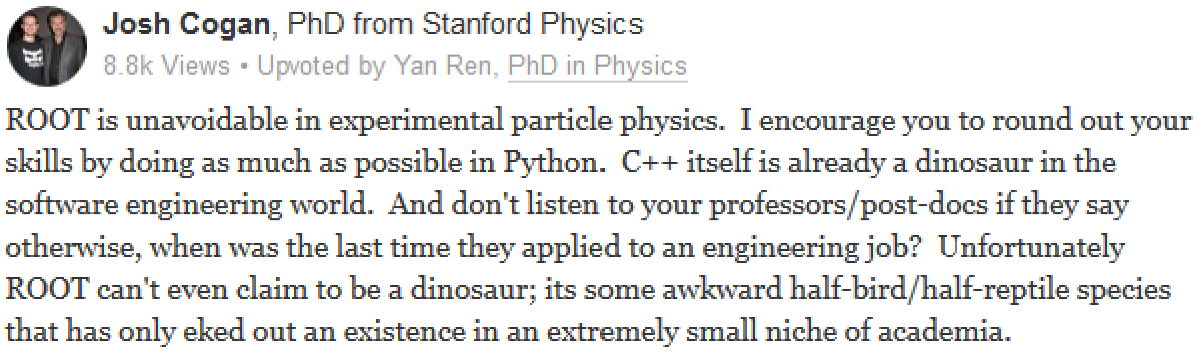


(b) Point Spread Function

ROOT does what physicists do:

It makes plots.

Análisis de datos con ROOT

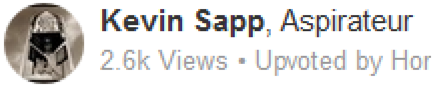


Josh Cogan, PhD from Stanford Physics
8.8k Views • Upvoted by Yan Ren, PhD in Physics

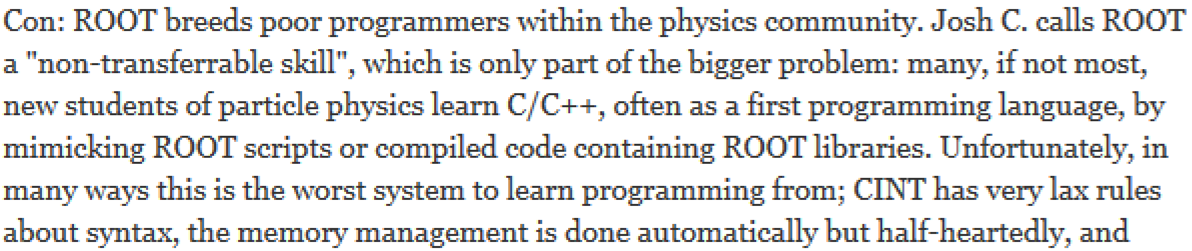
ROOT is unavoidable in experimental particle physics. I encourage you to round out your skills by doing as much as possible in Python. C++ itself is already a dinosaur in the software engineering world. And don't listen to your professors/post-docs if they say otherwise, when was the last time they applied to an engineering job? Unfortunately ROOT can't even claim to be a dinosaur; its some awkward half-bird/half-reptile species that has only eked out an existence in an extremely small niche of academia.

Jay Wacker, Researcher in particle physics.
5.6k Views • Upvoted by Hongwan Liu, Graduate
student in Theoretical Cosmology, Andy Buckley

Almost no one likes ROOT. The only advantage it has going for it is that there is a huge code base built up and that starting over from scratch is nearly impossible. The good news is that there are things like PyROOT that is a Python module that allows you to interact with ROOT. ROOT was a big step forward from FORTRAN based PAW which lingered until the late 90s. I think no one would think about using ROOT if they were starting again, but it is impossible to change out in either experiment at the LHC because they are running and no other experiment is big enough to devote resources to creating a new



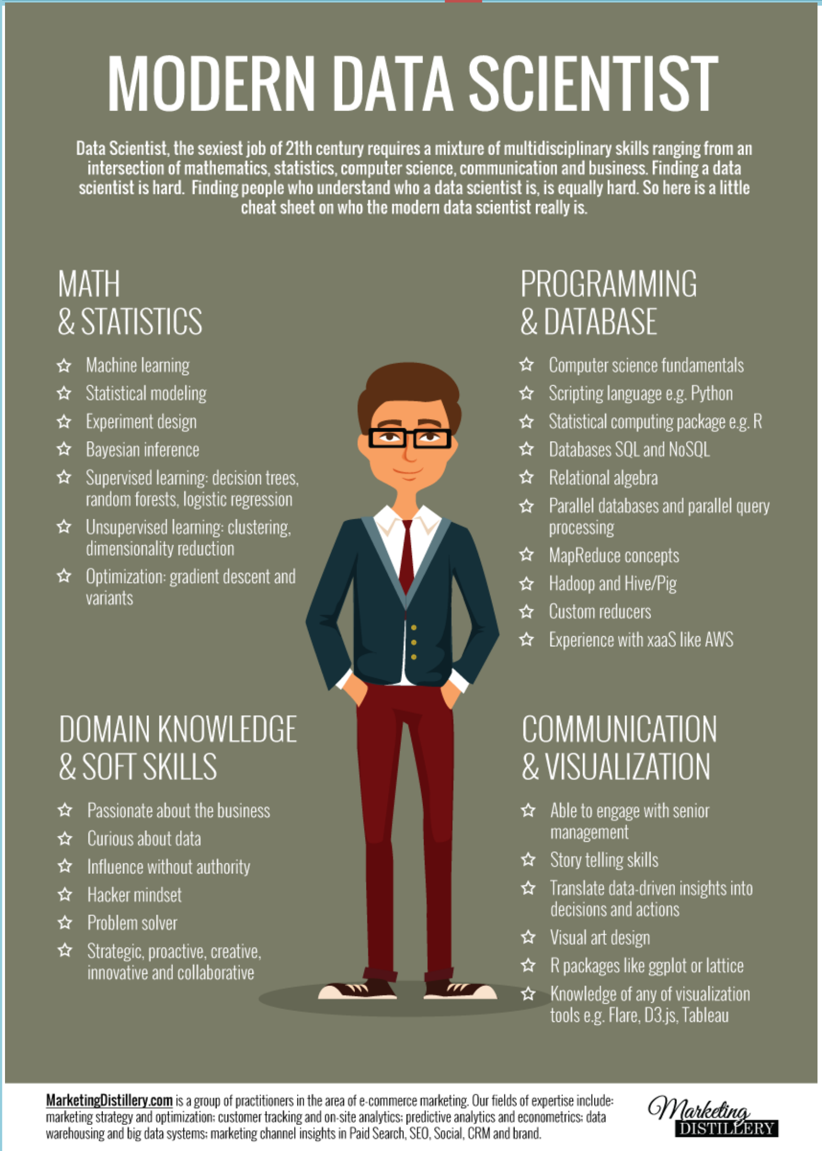
Kevin Sapp, Aspirateur
2.6k Views • Upvoted by Hor



Con: ROOT breeds poor programmers within the physics community. Josh C. calls ROOT a "non-transferrable skill", which is only part of the bigger problem: many, if not most, new students of particle physics learn C/C++, often as a first programming language, by mimicking ROOT scripts or compiled code containing ROOT libraries. Unfortunately, in many ways this is the worst system to learn programming from; CINT has very lax rules about syntax, the memory management is done automatically but half-heartedly, and

link

Data Science



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

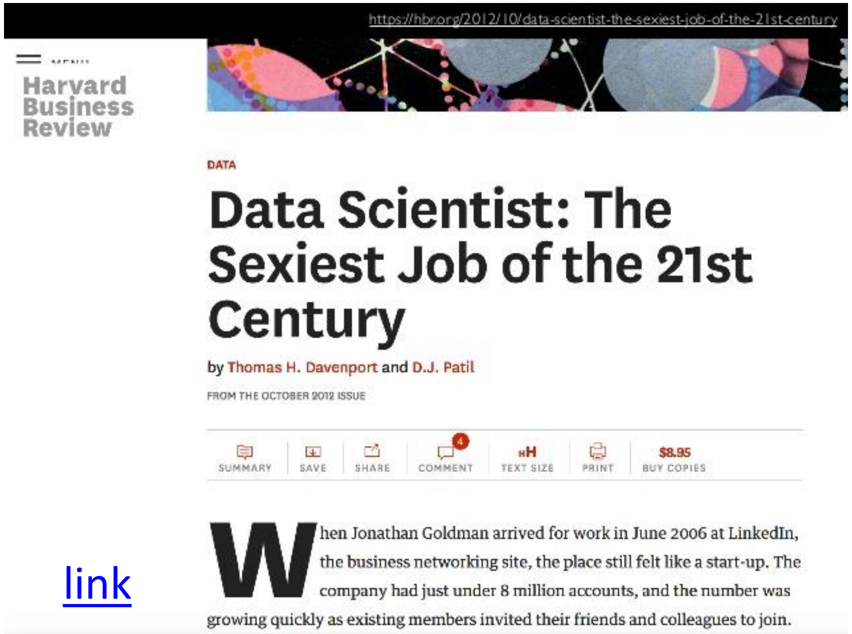
COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

link



https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

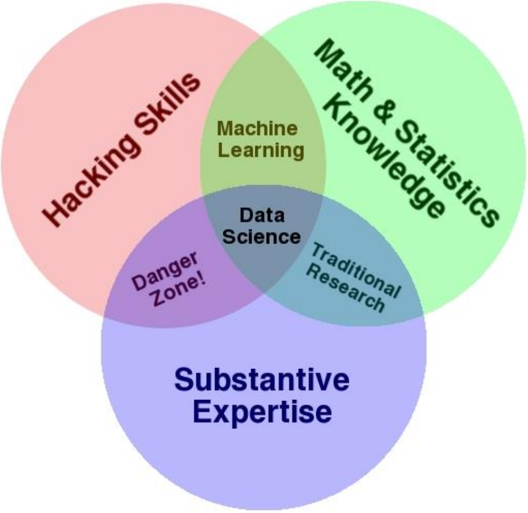
DATA

Data Scientist: The Sexiest Job of the 21st Century

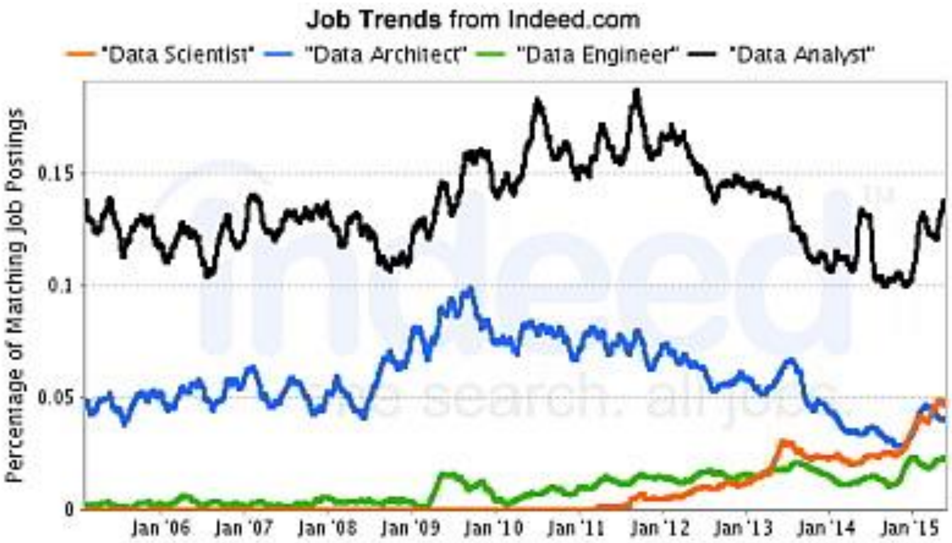
by Thomas H. Davenport and D.J. Patil

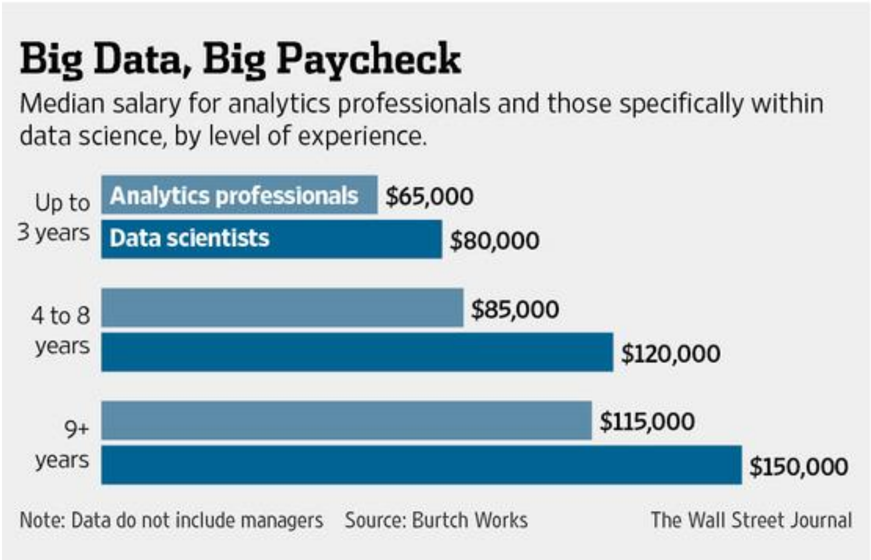
FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.



Data Science





De un mail de Hans Beck (The ALICE juniors) sobre "Qualities for next spokesperson": …

8. Support careers within and outside of academia

The job market is tough and quite some ALICE people experience unemployment or uncertain futures. How can we make sure the hard work everybody puts in their project is being rewarded? …. A career outside academia is a viable option; but for many of us it is uncharted territory. Can we prepare the next generation better to be fit also for industry? We miss a regular space to share experiences in the job search and highlight opportunities…

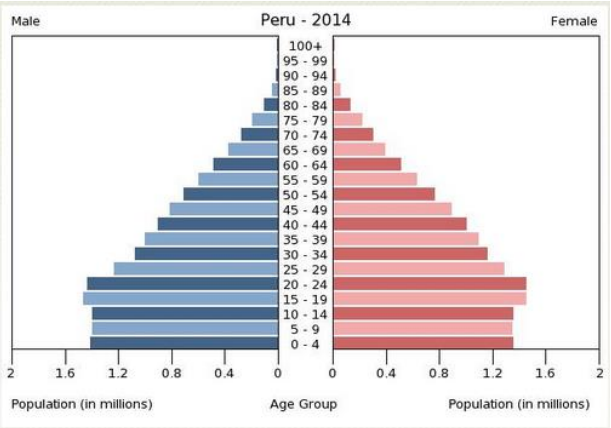
@PUCP : maestría en Informática con mención en Ciencias de la Computación

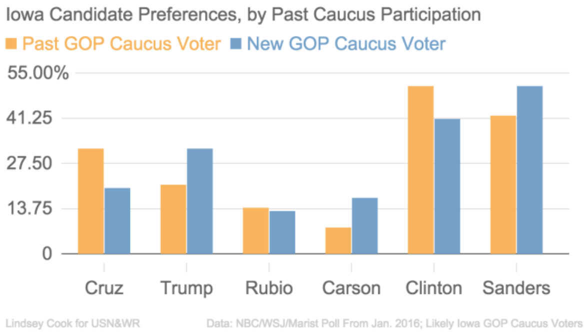
Proceso del análisis de datos

Paso	Ejemplo
Definir pregunta	Correlación temporal entre flujo de fotones y neutrinos astrofísicos
Definir juego de datos ideal	Datos de Fermi y IceCube
Obtener datos accesibles	Fermi públicos formato FITS, IceCube (en txt) no todo
Limpiar datos	Eliminar períodos off-line, calibración, primer fondo
Análisis exploratorio	Graficar número de eventos versus tiempo en intervalos de minutos
Modelamiento estadístico	Modelar señal y fondo, método likelihood, cortes con BDT
Graficar resultados	Flujo versus tiempo
Interpretar resultados	Datos compatibles con ruido de fondo (sin correlación)
Crear código reproducible	Macro/programa con ROOT
Distribuir resultados	Publicación

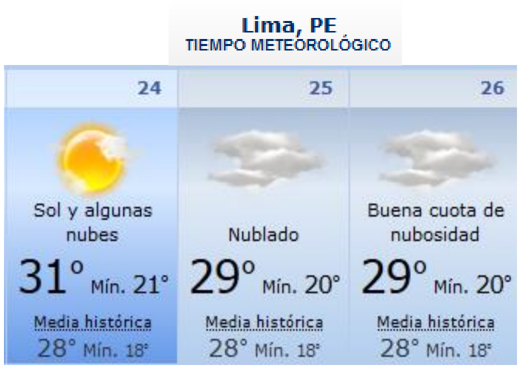
Tipos de análisis

- Descriptivo
- Exploratorio
- Inferencial
- Predictivo
- Causal
- Mecanista



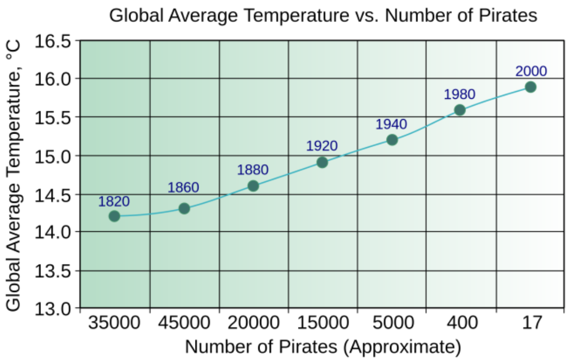


"Prediction is not inference"



Lima, PE
TIEMPO METEOROLÓGICO

24	25	26
Sol y algunas nubes	Nublado	Buena cuota de nubosidad
31° Mín. 21°	29° Mín. 20°	29° Mín. 20°
Media histórica 28° Mín. 18°	Media histórica 28° Mín. 18°	Media histórica 28° Mín. 18°

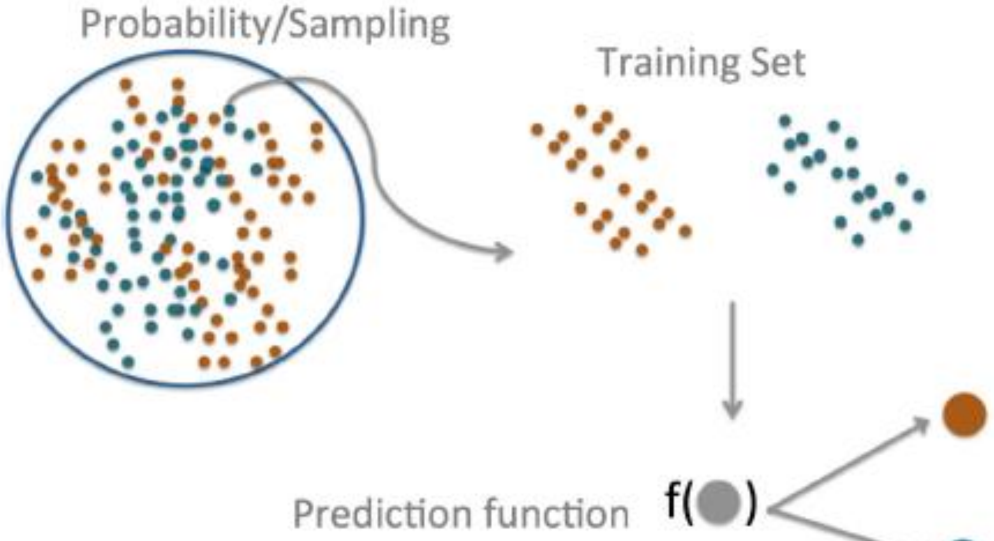


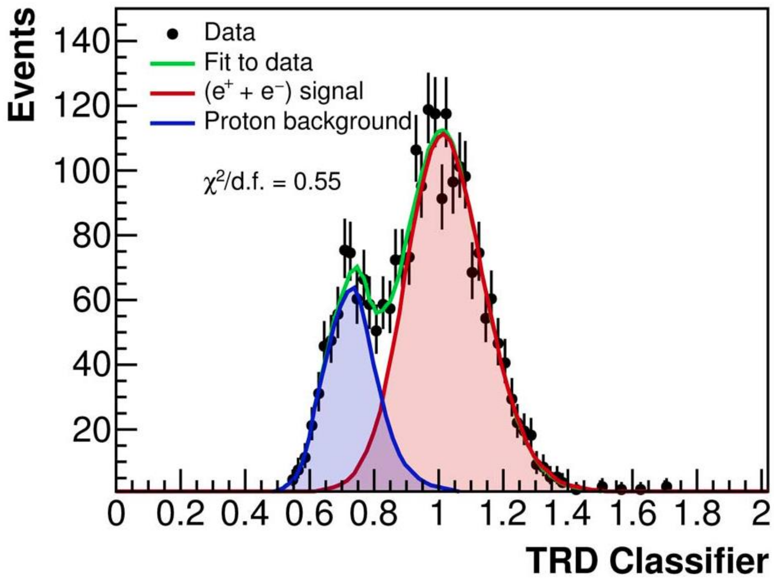
$$F_g = G\frac{m_1 m_2}{r^2}$$

"Correlation does not imply causation"

Categorías tomadas de J. Leek (JHU)

Predicción


Probability/Sampling
Training Set
Prediction function
f( )


Events
140
120
100
80
60
40
20
● Data
Fit to data
(e⁺ + e⁻) signal
Proton background
χ²/d.f. = 0.55
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
TRD Classifier

Big Data

Raw data per event ~1 Mb, with rate of 6x10^8 events/s

IceCube data volumes are 1TB per day raw data and 135GB per day of filtered data.

AMS-02 data sample: raw frames, reconstructed and simulated data ~150 TB/year.

Human genome data size ~770 Mb

As of June 2015, English wikipedia size ~10 TB uncompressed.

By 2016, global IP traffic will reach 1.1 zettabytes (10^{21}b) per year

shell / terminal

Command Line interface:

- Explorar directorios
- Crear y editar archivos, directorios, programas
- Ejecutar programas

- Git Bash Cygwin: Windows
- Terminal: Linux/Mac

Path:

Directorio superior: root = “/”

Home = “~”

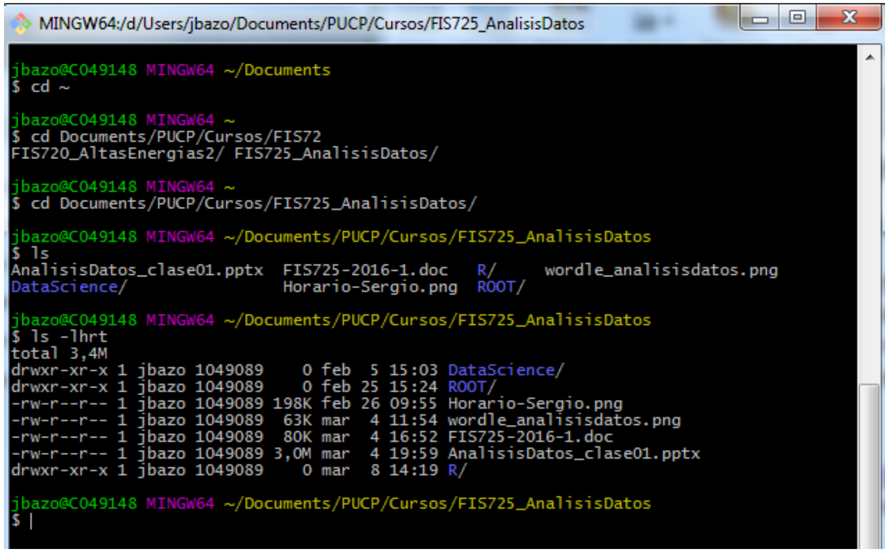
Comandos: command -flags arguments

ej. ls –lhrt ~

chmod –R 777 ./

Básicos: pwd, clear, ls, cd, mkdir, touch, cp, rm, mv, date, echo

Para Windows: Bitwise SSH client
Notepad++


MINGW64:/d/Users/jbazo/Documents/PUCP/Cursos/FIS725_AnalisisDatos

```
jbazo@C049148 MINGW64 ~/Documents
$ cd ~

jbazo@C049148 MINGW64 ~
$ cd Documents/PUCP/Cursos/FIS72
FIS720_AltasEnergias2/ FIS725_AnalisisDatos/

jbazo@C049148 MINGW64 ~
$ cd Documents/PUCP/Cursos/FIS725_AnalisisDatos/

jbazo@C049148 MINGW64 ~/Documents/PUCP/Cursos/FIS725_AnalisisDatos
$ ls
AnalisisDatos_clase01.pptx  FIS725-2016-1.doc   R/     wordle_analisisdatos.png
DataScience/                Horario-Sergio.png  ROOT/

jbazo@C049148 MINGW64 ~/Documents/PUCP/Cursos/FIS725_AnalisisDatos
$ ls -lhrt
total 3,4M
drwxr-xr-x 1 jbazo 1049089    0 feb  5 15:03 DataScience/
drwxr-xr-x 1 jbazo 1049089    0 feb 25 15:24 ROOT/
-rw-r--r-- 1 jbazo 1049089 198K feb 26 09:55 Horario-Sergio.png
-rw-r--r-- 1 jbazo 1049089  63K mar  4 11:54 wordle_analisisdatos.png
-rw-r--r-- 1 jbazo 1049089  80K mar  4 16:52 FIS725-2016-1.doc
-rw-r--r-- 1 jbazo 1049089 3,0M mar  4 19:59 AnalisisDatos_clase01.pptx
drwxr-xr-x 1 jbazo 1049089    0 mar  8 14:19 R/

jbazo@C049148 MINGW64 ~/Documents/PUCP/Cursos/FIS725_AnalisisDatos
$ |
```

Linux Bash Shell Cheatsheet .bashrc

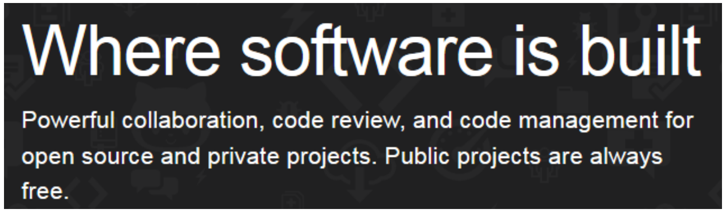
Version Control

Para compartir, almacenar y trabajar en grupo en código cambiante:
(Sub)Version Control Software: **Repositorio central**

Crear cuenta (free) en **GitHub**

https://github.com/

GitHub



Where software is built

Powerful collaboration, code review, and code management for open source and private projects. Public projects are always free.

Otros sistemas:

CVS: Concurrent Versions System (GNU license)

SVN: Subversion (Apache license)

versus

Otro lugar donde guardar código libre:

sourceforge

ejemplo

https://sourceforge.net/

Version Control

Sistema que guarda cambios hechos archivos en el tiempo y permite acceder a versiones anteriores

Git Cheatsheet

Instalar Git (Bash o GUI):
guardar en **repositorio local**

Configuración:

```
$ git config --global user.name jlbazo
$ git config --global user.email jbazo@pucp.edu.pe
$ git config --list
$ exit
```

Usar mismo username y correo de GitHub

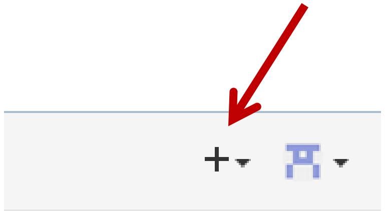
GitHub: web-based hosting service
repositorio remoto

GitHub: online

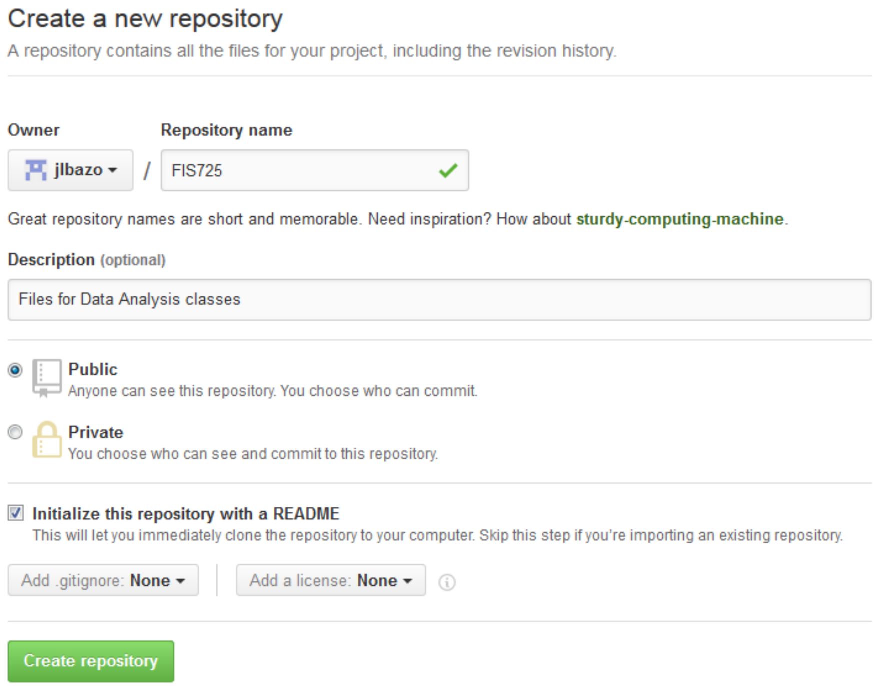
Permite subir/bajar **repositorios locales** hacia/de **remotos** (web)
Homepage de repositorios públicos: compartir, exposición

¿Cómo crear un repositorio online?

Profile page -> click create a new repo



Escribir nombre y descripción
Seleccionar “Public”
Marcar “Initialize …”



Create a new repository

A repository contains all the files for your project, including the revision history.

Owner: jlbazo ▾ / Repository name: FIS725 ✓

Great repository names are short and memorable. Need inspiration? How about **sturdy-computing-machine**.

Description (optional): Files for Data Analysis classes

● **Public**
Anyone can see this repository. You choose who can commit.

○ **Private**
You choose who can see and commit to this repository.

☑ **Initialize this repository with a README**
This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.

Add .gitignore: **None** ▾ | Add a license: **None** ▾ ⓘ

Git: copia local

¿Cómo copiar repositorio creado en disco local?

Usando **Git Bash**:

```
$ mkdir FIS725
$ cd FIS725
```

Crear directorio donde copiar repositorio

```
$ git init
```

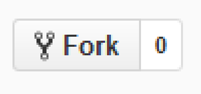
inicializar repositorio local

```
$ git remote add origin https://github.com/jlbazo/FIS725.git
```

dirección externa

```
$ git pull origin master
```

También se pueden copiar repositorios de otros usuarios (**fork**) (hacerlo primero en GitHub)

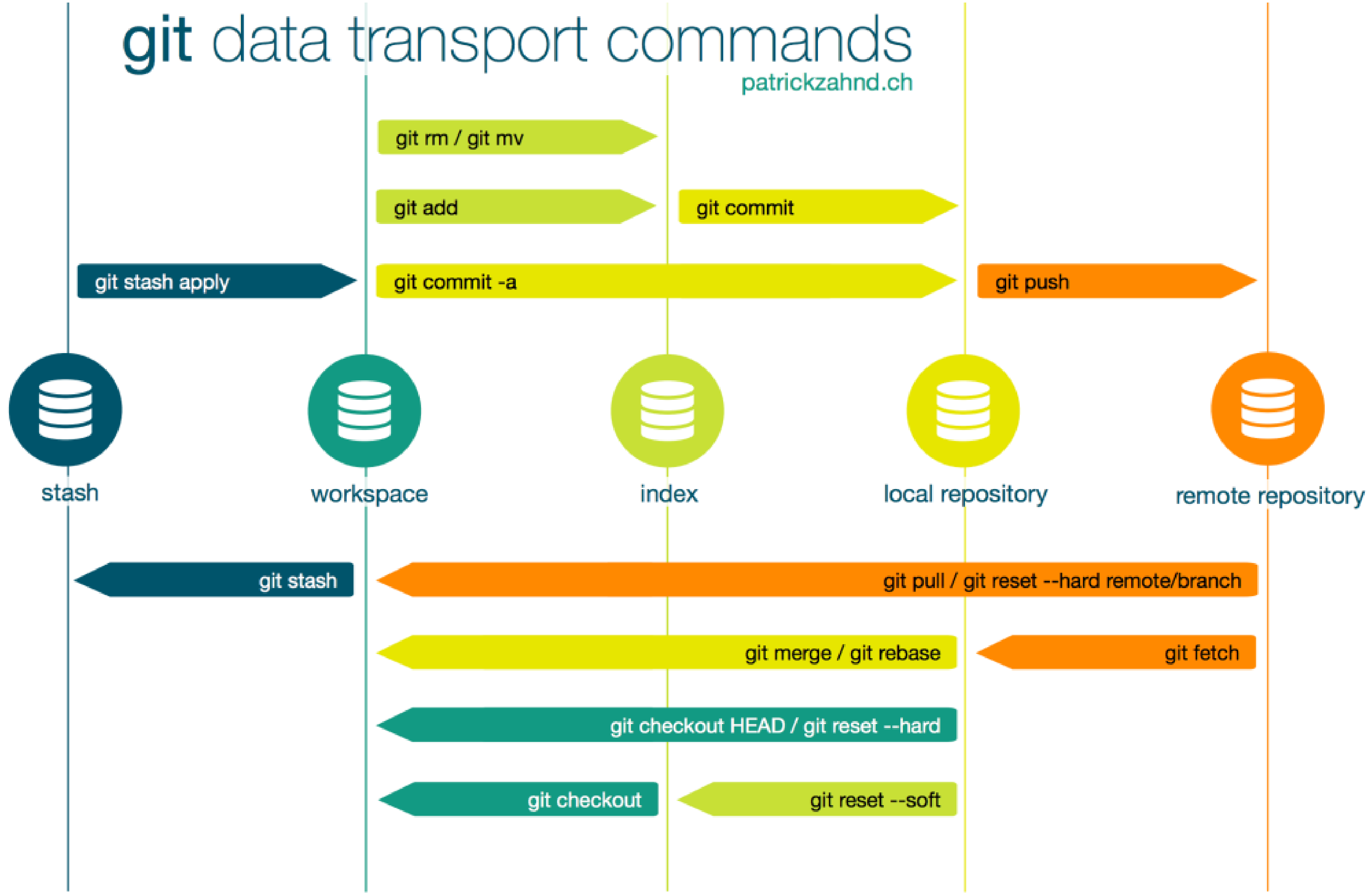


```
$ git clone https://github.com/jlbazo/forked_repo.git
```

Copia repositorio

Para incluir cambios de forked repo pedir "Pull request" en GitHub

Git: estructura



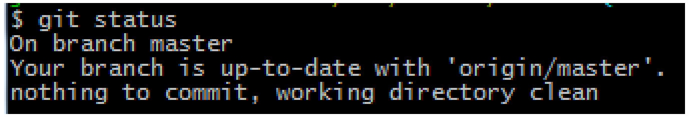
Git: comandos

$ git add -A ← añadir (y actualizar) nuevos archivos al index, antes de *commit*

$ git commit –m “message” ← Guardar cambios en repositorio local

$ git push ← Guardar cambios en repositorio remoto (GitHub)

$ git status



```
$ git status
On branch master
Your branch is up-to-date with 'origin/master'.
nothing to commit, working directory clean
```

Tutorial

Git: branches

Crear nueva versión (branch) del proyecto para editar sin interferir con otros:

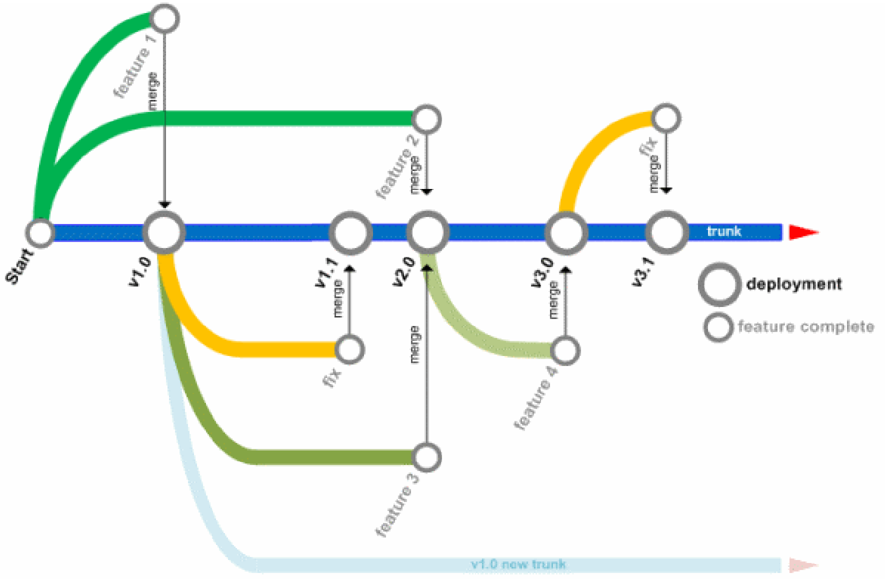
$ git checkout –b *branchname* ← crear nueva rama

$ git branch ← ver rama actual

$ git checkout master ← Regresar a rama principal

$ git pull origin *branchname*

$ git push origin *branchname*



Preparándonos para R

Instalar R desde CRAN Comprehensive R Archive Network

Instalar R Studio mejor interfaz gráfica

Instalar **paquetes adicionales** (más de 8000) cuando sean necesarios:

```
>  available.packages()

> install.packages(“ggplot2”)

>  library(ggplot2)
```

← Cargar paquete

```
>  search()
```

← Ver paquetes cargados

```
>  find.package(“ggplot2”)
```

← comprobar si paquete está instalado

Instalando paquetes externos a CRAN (ej. BioConductor) →

```
>  source(“http://bioconductor.org/biocLite.R”)

> biocLite(“rhdf5”)
```

Preparándonos para R

Para Windows

Instalar Rtools   Necesarias para los paquetes de R

Seleccionar versión de R correspondiente

Dejar al instalador seleccionar el *path* (marcar casilla correspondiente)

```
> install.packages(“devtools”)

> library(devtools)

> find_rtools()
```

← Si regresa TRUE, instalación exitosa

Encontrando respuestas

"Try first thyself, and after call in God; for to
the worker God himself lends aid.“

Εὐριπίδης

Google

Siempre buscar en inglés para obtener más resultados

Cross Validated

stackoverflow

RootTalk
ROOT Discussion Forums

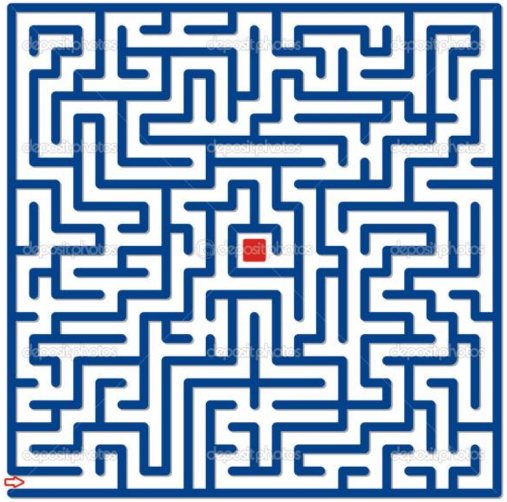
No hay necesidad de reinventar la rueda.
Ya existe una solución para tu problema.

R Mailing Lists

Encontrando respuestas

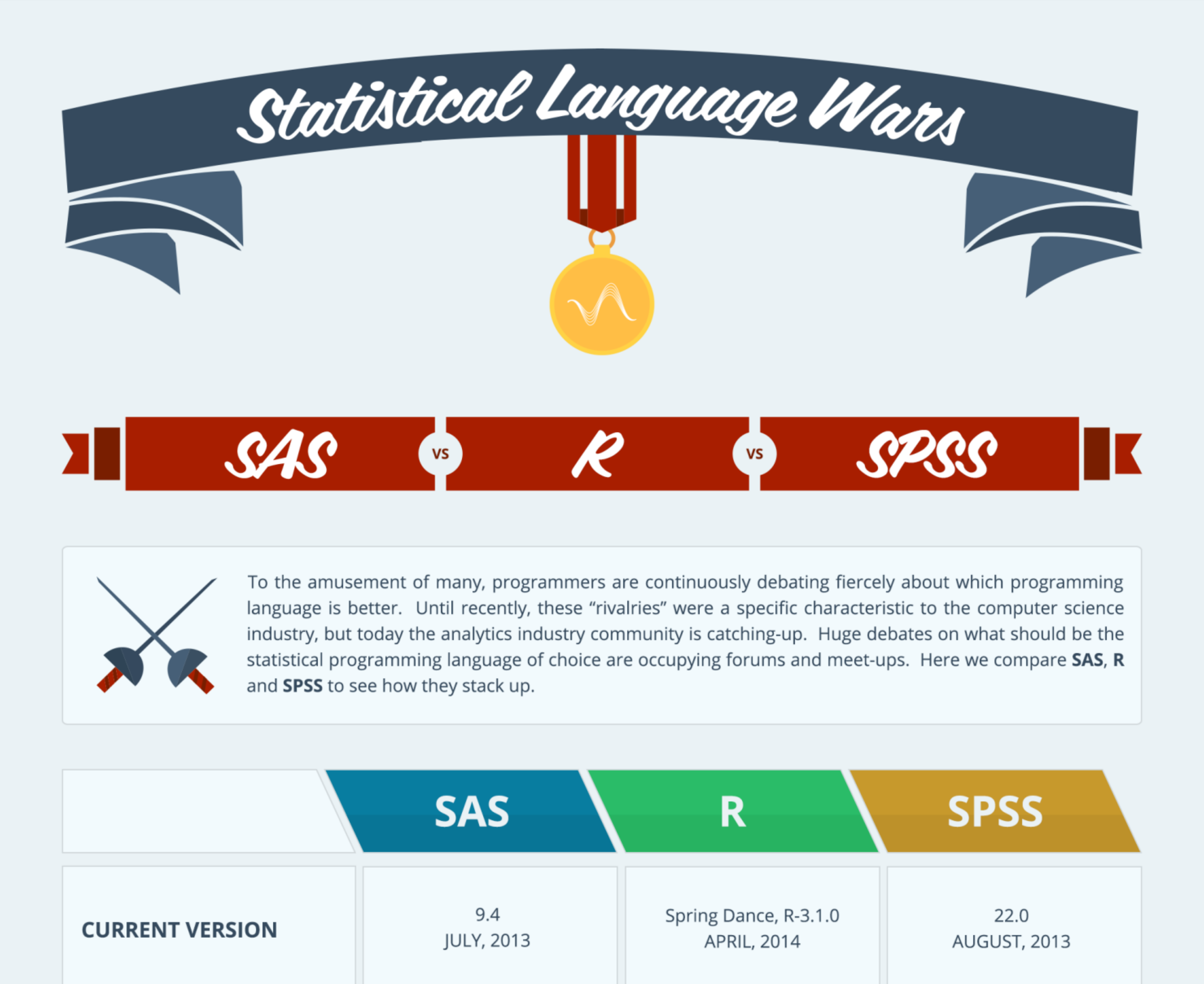
1. Buscar en internet
2. Leer manuales y archivos de ayuda
3. Preguntar a colegas expertos
4. Preguntar en un foro (antes agotar otros medios)

Escribir un título descriptivo, adjuntar código reproducible y datos del sistema usado.



Tener paciencia e insistir.

To R or not to R?



Statistical Language Wars

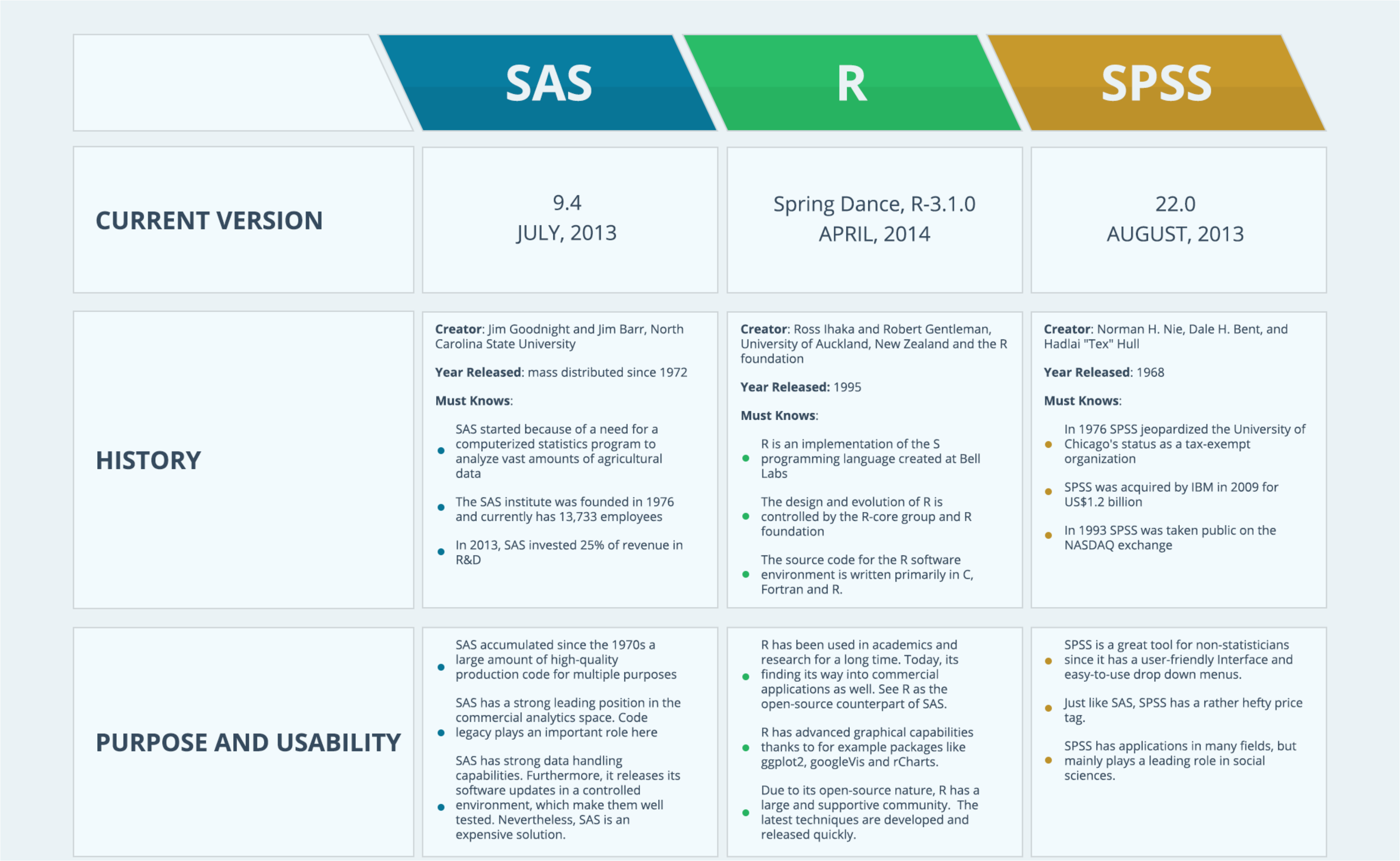
SAS vs R vs SPSS

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
CURRENT VERSION	9.4 JULY, 2013	Spring Dance, R-3.1.0 APRIL, 2014	22.0 AUGUST, 2013

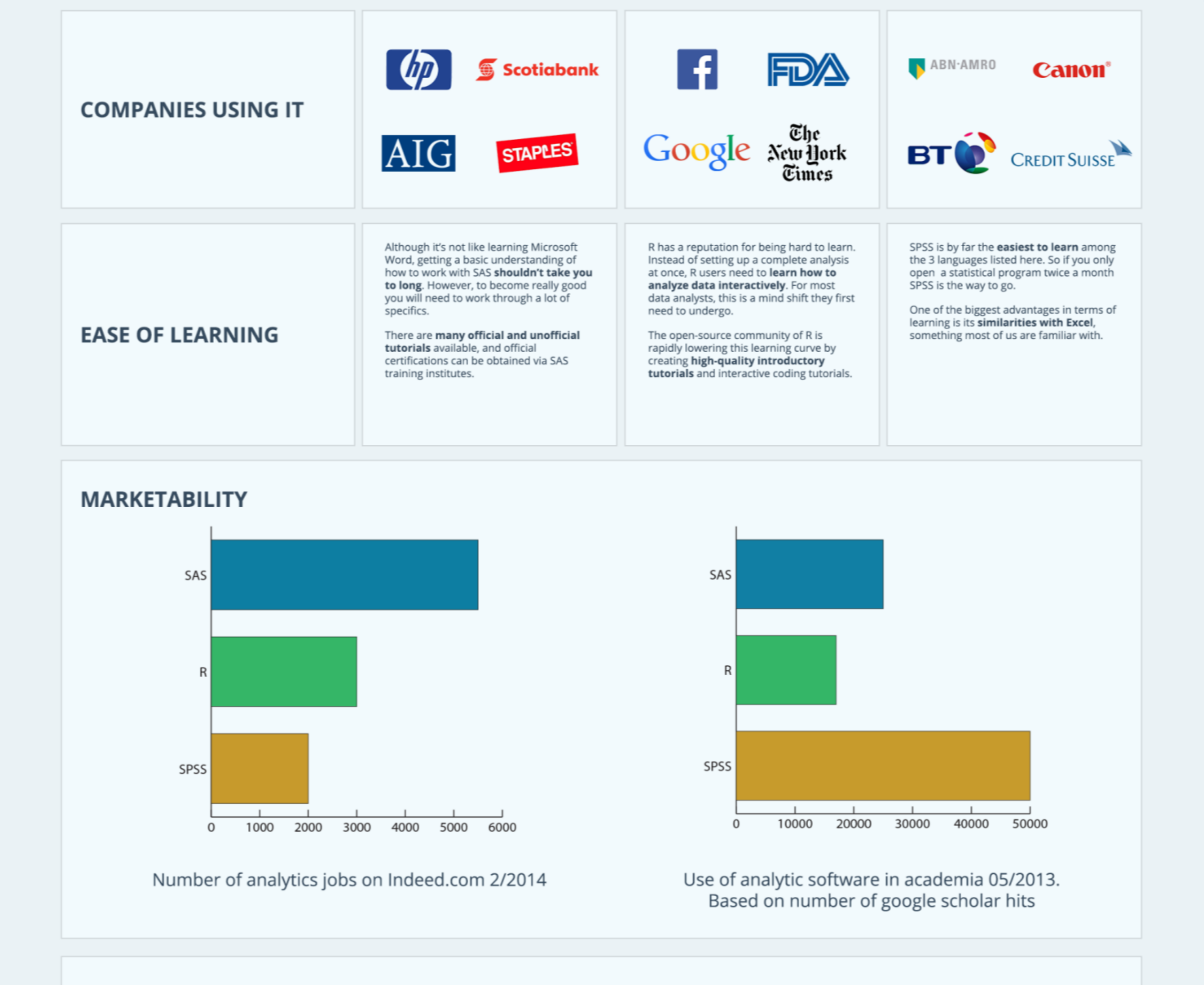
link

To R or not to R?



	SAS	R	SPSS
CURRENT VERSION	9.4 JULY, 2013	Spring Dance, R-3.1.0 APRIL, 2014	22.0 AUGUST, 2013
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
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.

To R or not to R?



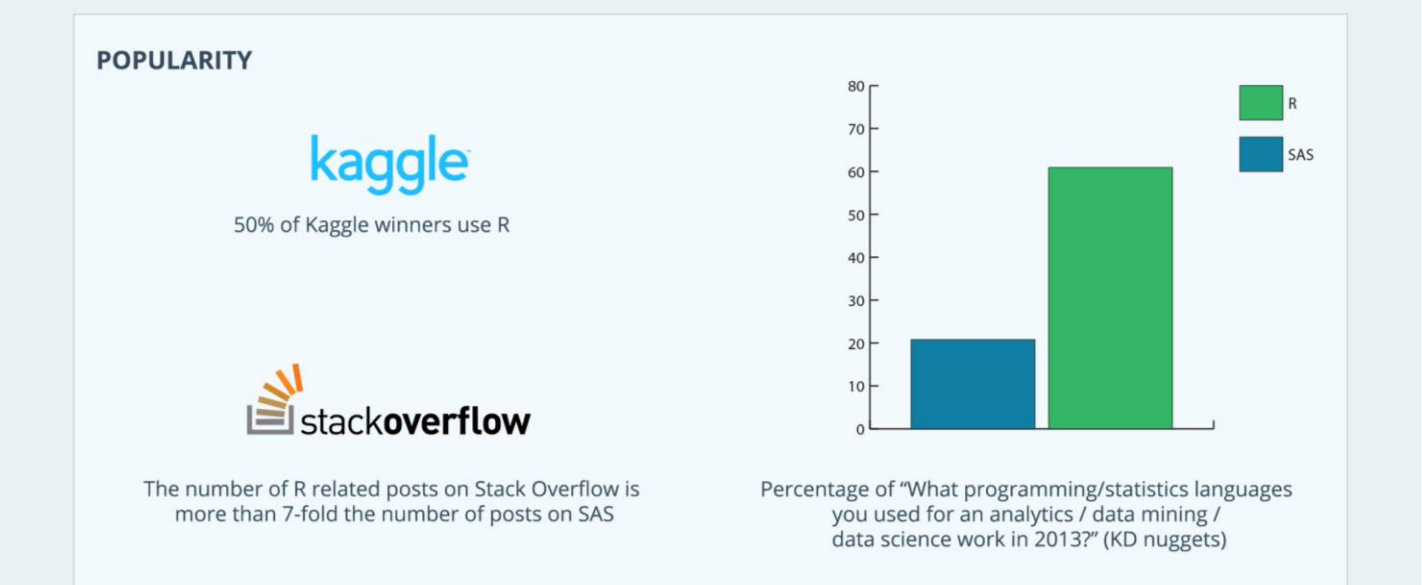
COMPANIES USING IT	hp Scotiabank AIG STAPLES	f FDA Google The New York Times	ABN·AMRO Canon BT CREDIT SUISSE
EASE OF LEARNING	Although it's not like learning Microsoft Word, getting a basic understanding of how to work with SAS **shouldn't take you to long**. However, to become really good you will need 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

To R or not to R?



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)