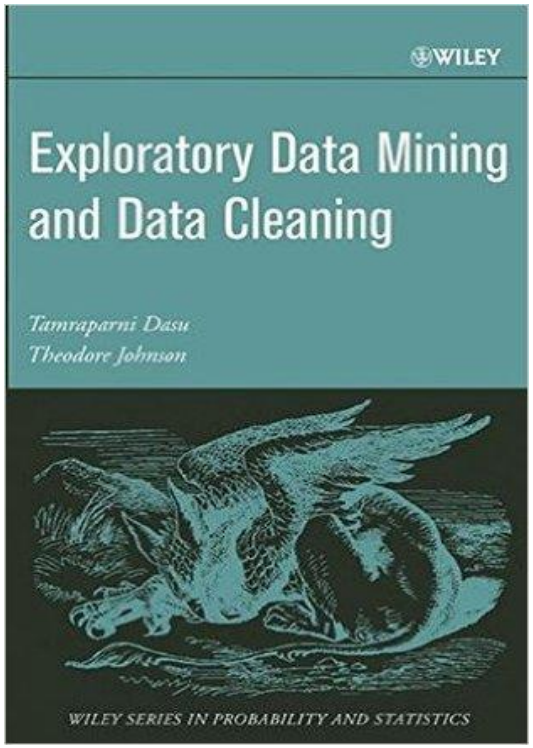
Data Wrangling en R para Programadores SQL

Bienvenidos a la conversación:
#DataWranglers #SGVirtual

Presenta:
Marciano A. Moreno D.C.
@marciano_moreno

El problema…

"Se dice que frecuentemente el 80% del análisis de datos se dedica al proceso de limpiar y prepararlos." [1]



Parte del problema radica en la amplitud de actividades asociadas, incluyendo detección de outliers, reconocimiento de fechas, imputación de valores faltantes, por mencionar algunas.

[1] Dasu, T; Johnson, T. Exploratory Data Mining and Data Cleaning. New York, NY, USA: John Wiley & Sons, Inc., 2003.

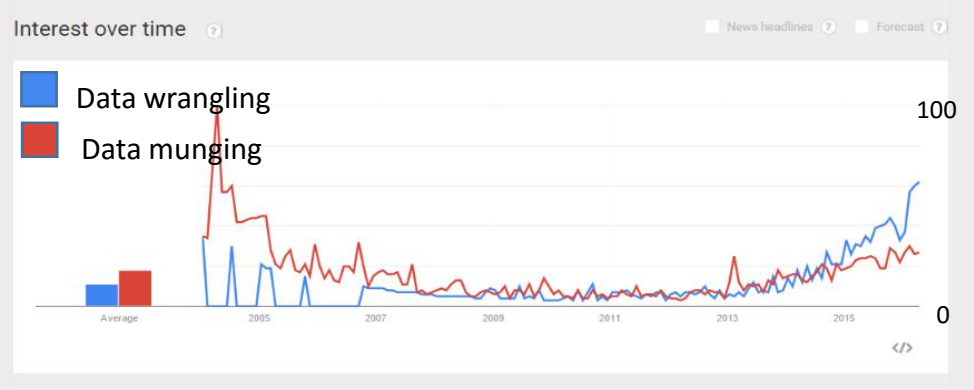
¿Qué es Data Wrangling?

- *"Un proceso de exploración y transformación iterativa de datos que habilita análisis." [1]*
- *"Cualquier transformación de datos requerida para preparar el dataset para análisis posterior, visualización o consumo operativo." [2]*
- *"El proceso de manualmente convertir o mapear datos de forma "cruda" a otro formato que permita un consumo más conveniente de los datos con la ayuda de herramientas semi-automatizadas." [3]*

[1] Kandel Sean, et. Al. *Research Directions in Data Wrangling: Visualizations and transformations for usable and credible data*. United Kingdom: SAGE, 2011.
[2] Rattenbury, T, el. al. Data Wrangling: Techniques and concepts for agile analytics. Sebastopol, CA, USA : O'Reilly, 2015 (preview).
[3] Wikipedia: Autores diversos. Data wrangling. http://bit.ly/1KslZb7, 2016

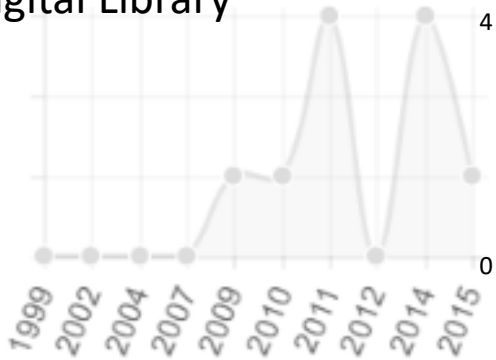
Tendencias en Data Wrangling

Google Trends

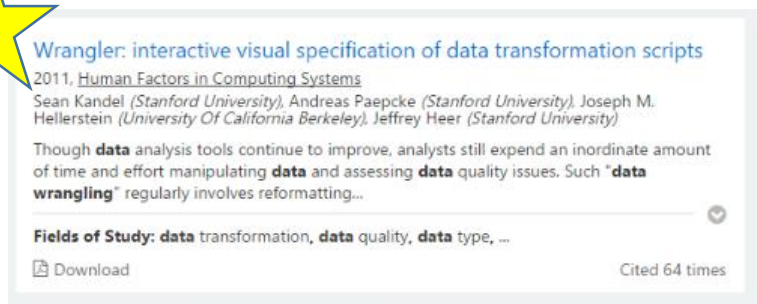


https://www.google.com/trends/explore#q=data%20wrangling%2C%20data%20munging&cmpt=q&tz=Etc%2FGMT%2B5

ACM Digital Library

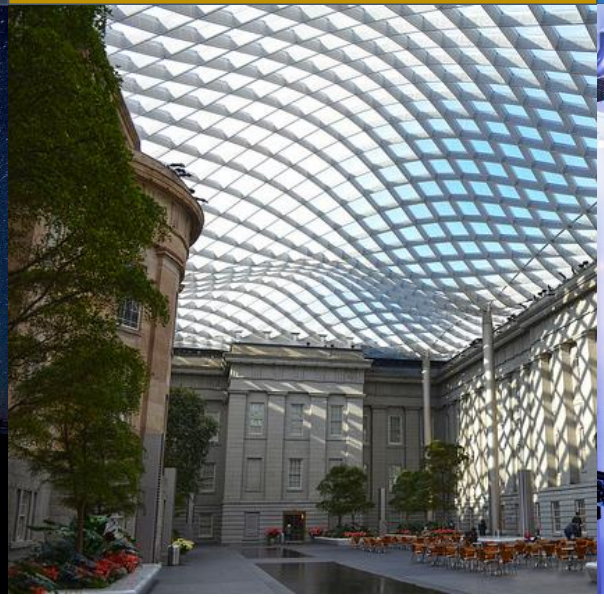


Microsoft Academic Search (top)

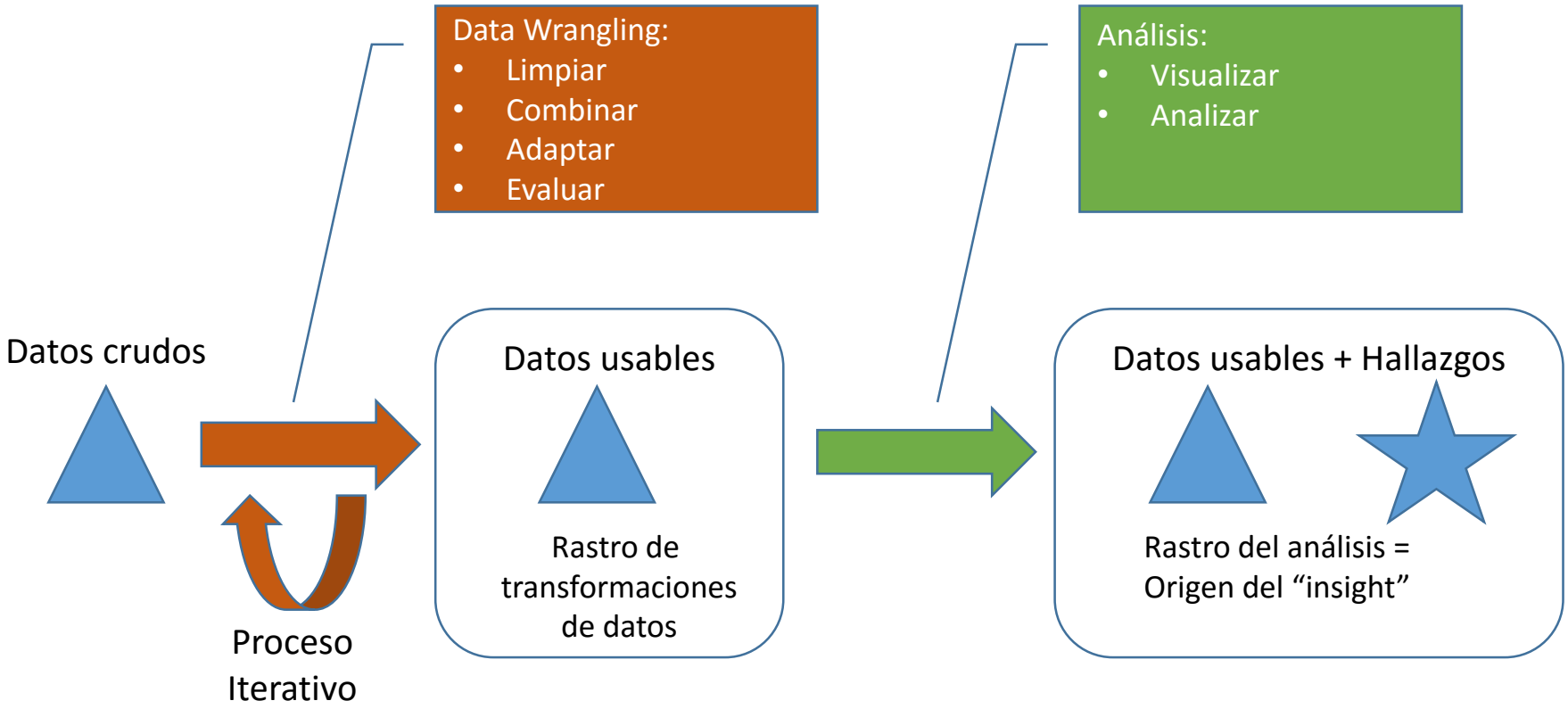


http://dl.acm.org/results.cfm?within=owners.owner%3DHOSTED&srt=publicationDate&query=wrangling+munging&Go.x=0&Go.y=0

Contextos de Data Wrangling

Exploración	Curación	Producción
		
© ESO/S. Brunier	© Tim Evanson	© Visual Factories

Data Wrangling:
El elefante en el cuarto[1]



[1] Kandel Sean, et. Al. *Research Directions in Data Wrangling: Visualizations and transformations for usable and credible data*. United Kingdom: SAGE, 2011.

Herramientas de Data Wrangling: propiedades clave

- Escala.
- Poder de expresión.
- Asistencia en especificación de transformaciones.
- Perfilamiento integrado.
- Usuario.
- Caso de uso.

El lenguaje de programación R

- Basado en el lenguaje de programación S que fue desarrollado por John Chambers, et. al. en Laboratorios Bell, E.U.A. en 1976.
- R fue desarrollado a principios de 1990 por Ross Ihaka y Robert Gentleman (Universidad de Auckland, Nueva Zelanda).
- Desde 1997 fue desarrollado por el R Development Core Team. Siendo parte del proyecto GNU de la Free Software Foundation
- La Fundación R es una asociación sin fin de lucro establecida por los integrantes del R Development Core Team.
- En 2015 la Linux Foundation anunció el Consorcio R como un proyecto colaborativo para fortalecer las comunidades técnicas y de usuarios. El consorcio no interfiere con el desarrollo y el lenguaje R en sí.

Comenzando a desarrollar con R

- Instalar R -> https://cran.rstudio.com/
- Instalar un GUI de R (RStudio, hay otros) -> https://www.rstudio.com/
- Introducción a R (Torfs, Brauer) -> http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Demostración

Operaciones de Data Wrangling

- Selección de columnas
- Filtrado de reglones
- Agrupación y cálculo de agregados
- Ordenamiento
- Joins
- Unión
- Transformación

Tecnologías de Data Wrangling en R que revisaremos

- Data Frame
- Data Table
- Dplyr
- Tidyr

Data Frame

https://cran.rstudio.com/doc/manuals/r-release/R-intro.html#Data-frames

- Clase que forma parte de la base de código de R (no se requiere de un paquete externo).
- Puede ser conceptualizada como una matriz con columnas de distintos modos y atributos.
- Soporta las convenciones de indexado de las matrices.

```
df <- data.frame(…)
is.data.frame(df)
df[i, j]
df$col
```

Data Table

https://cran.rstudio.com/web/packages/data.table/index.html

- Paquete externo, extensión del data frame.
- Desarrollado por Dowle, et. al.
- Agregación rápida de datos
- Sintaxis concisa y consistente.
- Diseñada para reducir tiempo de programación y cómputo.
- Puede ser conceptualizada como una matriz con columnas de distintos modos y atributos.
- Soporta las convenciones de indexado de las matrices.

```
install.packages("data.table")
library(data.table)
dt <- data.table(…)
is.data.table(dt)
dt[i, j, k]
dt$col
```

Dplyr – Gramática para manipulación de datos

https://cran.rstudio.com/web/packages/dplyr/index.html

- Herramienta rápida y consistente para trabajar con objetos tipo data frame, in-memory y out of memory.
- Desarrollada por Hadley Wickham, Romain Fracois y RStudio.
- Facilita el uso de herramientas de manipulación de datos para análisis en R.
- Alto rendimiento, piezas escritas en código de C++.
- Emplea la misma interfaz para trabajar con datos sin importar dónde se encuentren almacenados: data frame, data table o base de datos.

```
install.packages("dplyr")
library(dplyr)

flights %>%
        select(origin, dest)
```

TidyR – “Tidy Data”

https://cran.rstudio.com/web/packages/tidyr/index.html

- Estructurar datasets para facilitar el análisis.
- Funciones para expandir y recolectar información de renglones y columnas.
- Evolucion de funciones reshape().
- Funciona bien con dplyr.
- Tidy Data = 3FN
  - Cada variable forma una columna.
  - Cada observación forma un renglón.
  - Cada tipo de unidad observacional forma una tabla.

```
install.packages(“tidyr”)
install.packages(“dplyr”)
library(tidyr)
library(dplyr)


preg2 <- preg %>%
  gather(treatment, n, treatmenta:treatmentb) %>%
  mutate(treatment = gsub("treatment", "", treatment)) %>%
  arrange(name, treatment)
preg2
```

Continuando tu camino en Data Wrangling

- Instala R.
  - https://cran.rstudio.com/
- Consulta las referencias, whitepapers y viñetas oficiales en CRAN.
- Descarga la presentación y el código de esta sesión:
  - https://github.com/marcianomoreno/datawrangling
- Continuemos la conversación sobre #DataWranglers #SGVirtual en redes sociales.

SG VIRTUAL CONFERENCE

Marciano A. Moreno Díaz C.

@marciano_moreno

marciano_moreno@acm.org

marciano_moreno

bit.ly/marmo-linkedin

#DataWranglers #SGVirtual