
UNIVERSIDAD CATÓLICA
de Colombia
Vigilada Mineducación

# Uso de Herramientas Informáticas para proyectos de Big Data

Diego Alberto Rincón Yáñez
darincon@ucatolica.edu.co

UNIVERSIDAD CATÓLICA de Colombia
Vigilada Mineducación

# ¡Ejercicio No.3!

- Problema: Extraer la mayor cantidad de meta-datos del Libro el quijote.
- URL: http://www.elmundo.es/quijote/index.html

- Grupos de 4 Personas
  - Ningún compañero cercano!!

# Extracción de Datos

- Descargar Ejemplo Hecho
  - Java
  - VisualBasic 6.0
  - Python

# Buenas Practicas para la presentación de datos sin tratamiento

# ¿Problema?

- Generalmente las habilidades del equipo de Data Science no son las mismas (diferencias de nivel de código, matemáticas entendimiento del problema).

- Crear un protocolo de publicación del tratamiento/recolección y catalogación se vuelve importante.

# Pasos Generales


1. Los datos en bruto (raw data).
2. Los datos ordenados
3. Libro especificando los valores y tipos (Diccionario de Datos).
4. Una receta explicita del proceso.

# 1. Datos en Bruto - ¿Qué son?

- Archivos Binarios?
- Archivo no formateado de Excel
- Archivo JSON del API de Twitter.
- Datos introducidos a mano.

# 1. Datos en Bruto - ¿Están correctos?

- Ningún software modifico los datos.
- Nadie ha manipulado los números en los datos.
- Nadie ha quitado pedazos de datos.
- Nadie a factorizado los datos de ninguna manera.

# 1. Datos en Bruto – ¿Por qué no manipular los datos?

- Demorar el proceso de análisis
- Leyes de tratamiento de la información.
- Métodos Forenses de tratamiento
- Cadena de Custodia.

# 2. Los datos ordenados

- Cada variable debe estar en una columna.
- Cada observación diferente debe estar en una fila.
- Debe haber una tabla por cada variable/entidad.

# 2. Los datos ordenados

- Cada columna debe tener un nombre.
- Cada nombre debe ser fácil de leer.
- Presentaciones en Excel a través de tablas, archivos CSV o delimitados por TAB.

# 3. Diccionario (Code book)

- Información de las Variables (incluyendo unidades), esto no esta contenido en los datos ordenados.

- Información de las agrupaciones realizadas.

- Generalmente hay una sección “Diseño de Estudio”, ¿como recolecto los datos?.

# 3. Diccionario (Code book)

- Como  se definieron las variables, https://en.wikipedia.org/wiki/Statistical_data_type.

- Ejemplos de una muestra.

- Tratar de ser lo más especifico posible.

¡¡¡NO TODO EL MUNDO PIENSA IGUAL!!!!

¡¡¡PARA TODOS NO ES OBVIO LO MISMO!!!!

# 4.Una receta explicita del proceso

- Un paso a paso de lo que realizo metodológica y tecnológicamente.
  **Reproducible Research in Computational Science**
  **http://www.sciencemag.org/content/334/6060/1226**

- Listado de pasos para recrear el proceso.
  1. Tome el archivo y paselo por el scriptX su salida sera llamada 1.out
  2. Ubique el archivo 1.out en este directorio y abralo con X software
  3. Etc…..

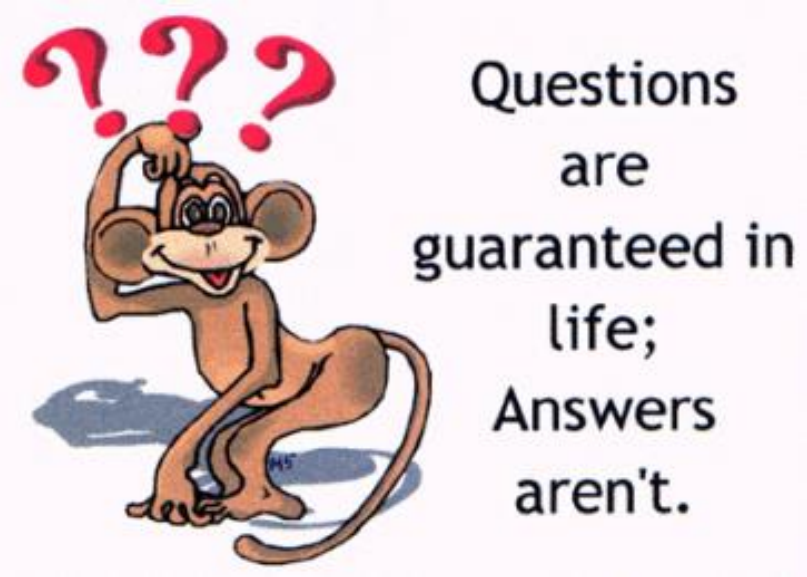
- Incluir información de la tecnología, versiones, arquitecturas, sistemas operativos, etc.

# ¡Taller!

1. Datos a Catalogar?
2. Información a extraer?
3. Como extraer los datos de la fuente?
4. Code Book (ver Formato)
5. ¿Receta del Proceso?

Producto: ¡Documento!

Entrega: 1 semana




# ¿Preguntas?

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.