I. Data Science & Big Data: Basics

Introduction to Data Science

Data Everywhere

Contenidos I

DataOwl

1. ¿Qué es el Data Science?

1.1 Conceptos básicos

DataOwl

Big Data

Concepto general para cualquier colección de datos tan grande o compleja que resulta difícil (si no imposible) de procesar utilizando herramientas tradicionales.

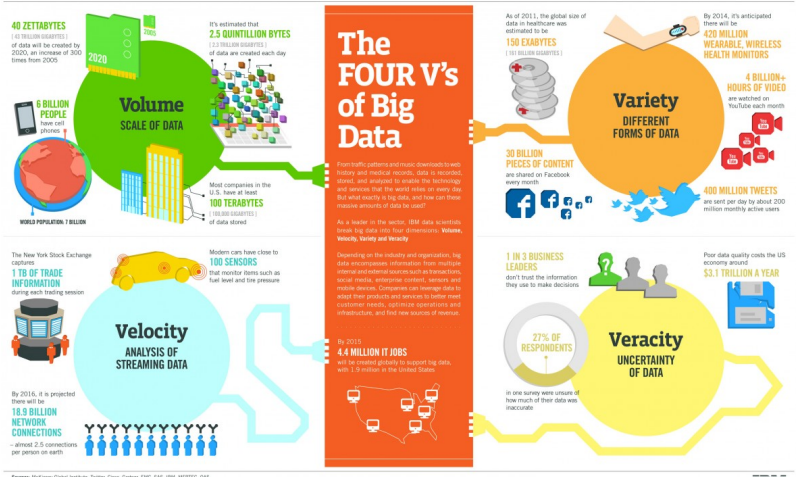
Data Science

El Data Science es el campo en el que se utilizan diversos métodos para analisar cantidades masivas de datos y extraer el conocimiento que esta pueda contener.

Desafíos: Busqueda, extracción, limpieza, almacenamiento, envío, transferencia y visualización de datos.

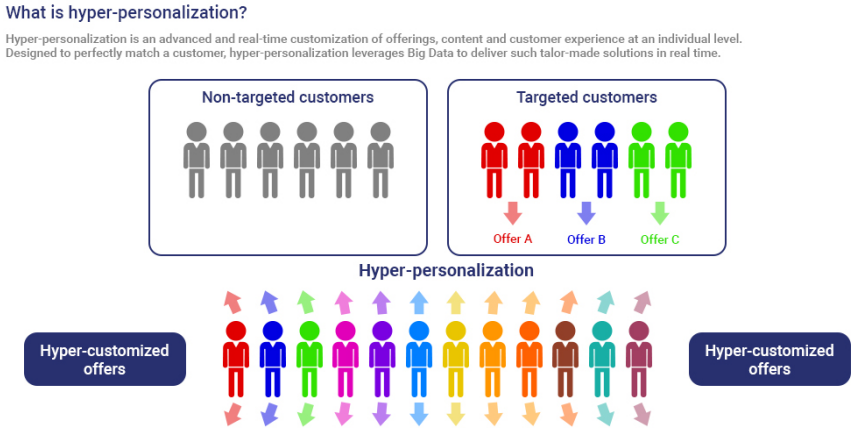
Skills de un DS: Buena base estadística, entendimiento del Big Data, modelos de Machine Learning, construcción de algoritmos (Hadoop, Pig, Spark, R, Python, Java, etc.), bases de datos (SQL and NoSQL).

1.1 Basics


The FOUR V's of Big Data
Volume
SCALE OF DATA
40 ZETTABYTES
[40 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[150 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States.
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

1.2 Uses of Data Science

(i) Commercial Companies: Hyper Personalization

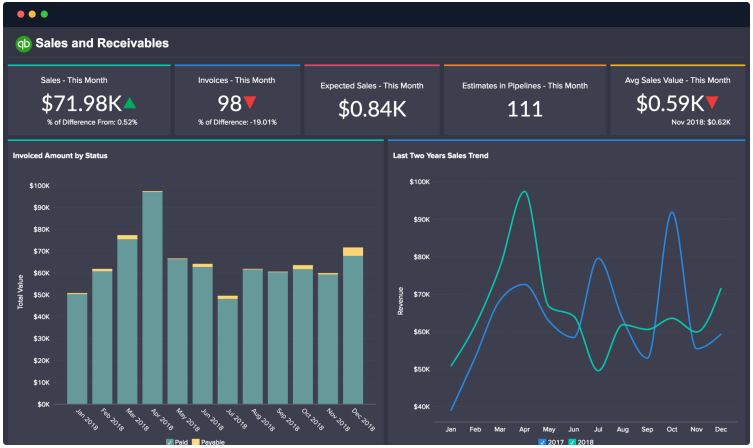


1.2 Uses of Data Science

(ii) Commercial Companies: People Analytics

1.2 Uses of Data Science

(iii) Commercial Companies: Financial Analytics

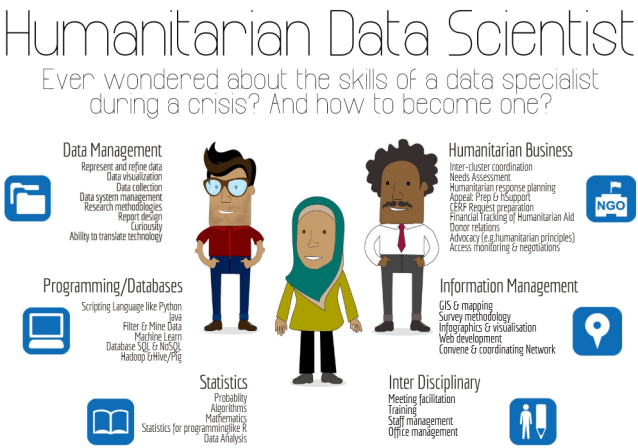


1.2 Uses of Data Science

(iv) Governmental Organizations

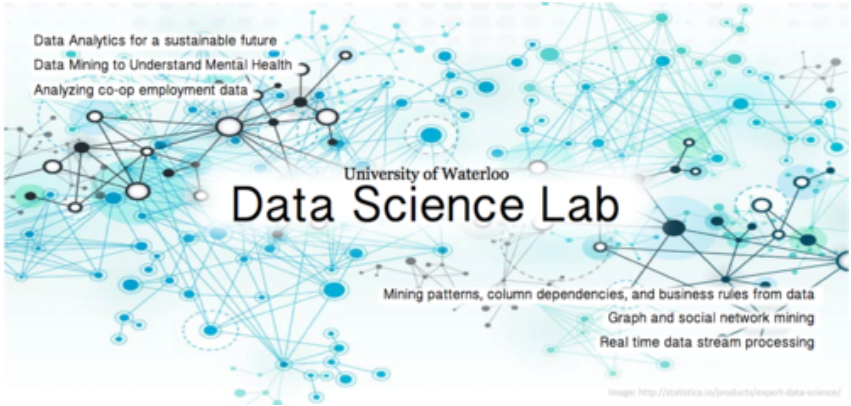
1.2 Uses of Data Science

(v) Nongonvernmental Organizations



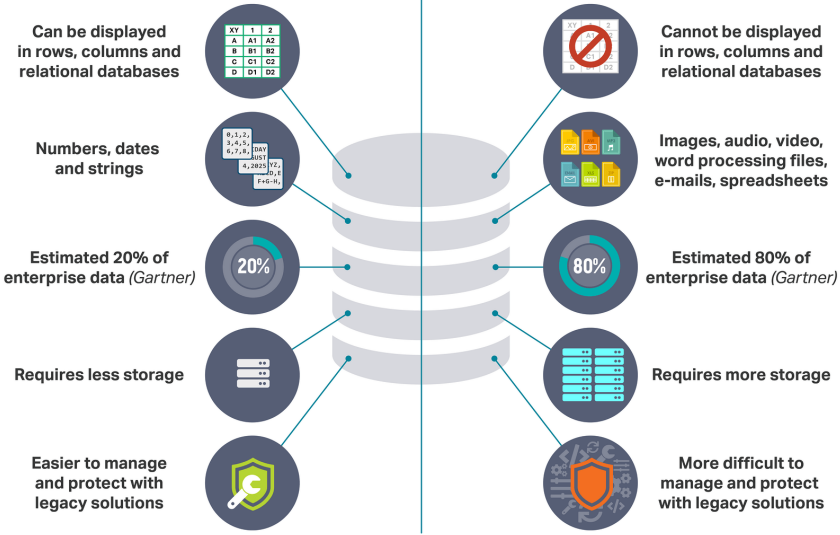
1.2 Uses of Data Science

(vi) Universities



2. Tipos de datos

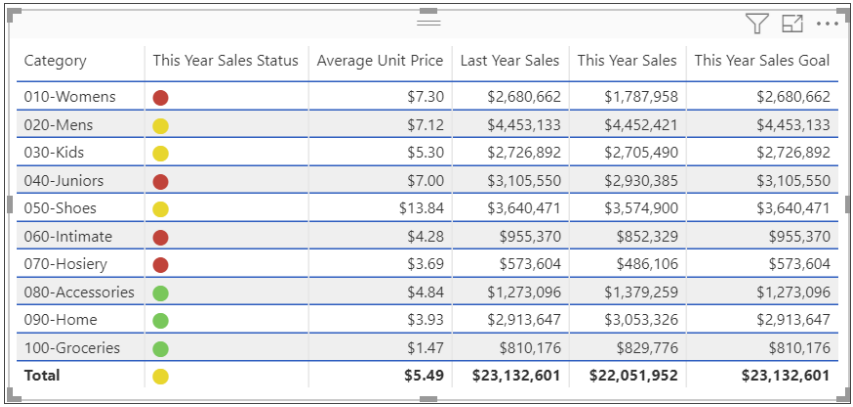
Structured Data vs Unstructured Data



Data Everywhere

DataOwl

(i) Datos estructurados: Tablas



Category	This Year Sales Status	Average Unit Price	Last Year Sales	This Year Sales	This Year Sales Goal
010-Womens	🔴	$7.30	$2,680,662	$1,787,958	$2,680,662
020-Mens	🟡	$7.12	$4,453,133	$4,452,421	$4,453,133
030-Kids	🟡	$5.30	$2,726,892	$2,705,490	$2,726,892
040-Juniors	🔴	$7.00	$3,105,550	$2,930,385	$3,105,550
050-Shoes	🟡	$13.84	$3,640,471	$3,574,900	$3,640,471
060-Intimate	🔴	$4.28	$955,370	$852,329	$955,370
070-Hosiery	🔴	$3.69	$573,604	$486,106	$573,604
080-Accessories	🟢	$4.84	$1,273,096	$1,379,259	$1,273,096
090-Home	🟢	$3.93	$2,913,647	$3,053,326	$2,913,647
100-Groceries	🟢	$1.47	$810,176	$829,776	$810,176
Total	🟡	**$5.49**	**$23,132,601**	**$22,051,952**	**$23,132,601**

(ii) Datos no estructurados: Lenguaje Natural

DataOwl

Capítulo 14

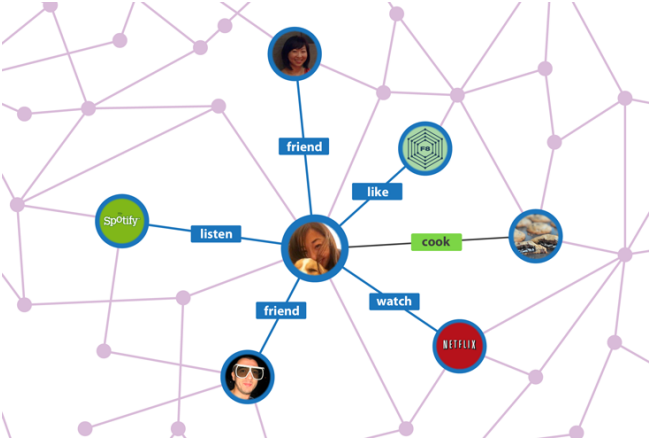
Crecimiento económico con ahorro óptimo*

En los capítulos anteriores hemos analizado el crecimiento asumiendo que la tasa de ahorro es constante e igual a s. Aunque en una primera aproximación esta es una buena idea, tiene también algunas limitaciones. La primera es que el crecimiento al final depende de lo que pase con el crecimiento de la productividad y otros factores, todo lo cual debiera incidir en la tasa de ahorro. Solo podemos especular acerca de cómo cambia la tasa de ahorro sin mayores fundamentos. Y en segundo lugar, desde el punto de vista de tener una buena teoría de crecimiento que nos permita analizar el bienestar, se debe tener un modelo bien especificado, que incluya la utilidad de los hogares.

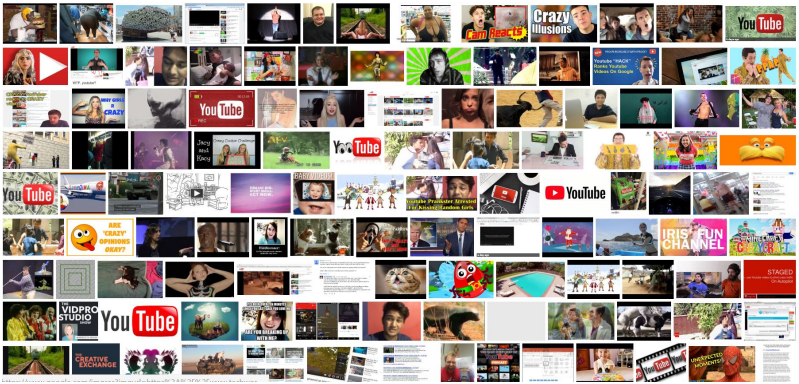
(iii) Datos estructurados/no estructurados: Datos generados por máquinas

```
[08/04/00 14:28:27](0)    Server: PSNT sleeping for 15 seconds
[08/04/00 14:28:41](0)    Server: PSNT looking for work
[08/04/00 14:28:41](5)    Server: PSNT checking status...
[08/04/00 14:28:41](5)    Server action mode: Ok (looking for requests)
[08/04/00 14:28:41](5)    Checking Process cancels...
[08/04/00 14:28:41](4)    Checking status of active processes...
[08/04/00 14:28:41](5)    Process 9836 is still running as Session ID 711
[08/04/00 14:28:41](5)    Process 9837 is still running as Session ID 634
[08/04/00 14:28:41](5)    Process 9838 is still running as Session ID 703
[08/04/00 14:28:41](5)    Info for array of Request(s) associated with a Job slated to be submitted
[08/04/00 14:28:41](5)       Size of array: 1
[08/04/00 14:28:41](5)    Info for array of Active Processes
[08/04/00 14:28:41](5)       Size of array: 3
[08/04/00 14:28:41](5)       Crystal                 :  Active: 3   Max: 3
[08/04/00 14:28:41](5)    Server: PSNT checking status...
[08/04/00 14:28:41](5)    Server action mode: Submitting request
[08/04/00 14:28:41](5)    Number of New Process Request(s) To Start: 1
[08/04/00 14:28:41](1)    Process Instance: 9843 started (PID: 645)
[08/04/00 14:28:41](4)    Starting process:: 9843
[08/04/00 14:28:41](4)        Command Line: Y:\BIN\CLIENT\WINX86\PSSQR.EXE
[08/04/00 14:28:41](4)        Parm List: -CT ORACLE -CS  -CD E800R21B -CA %ACCESSID% -CAP %ACCESSPSWD%
[08/04/00 14:28:41](4)        Working Dir: c:\apps\db\oracle81\bin
[08/04/00 14:28:41](4)          Session Id: 645
[08/04/00 14:28:41](0)    Server: PSNT sleeping for 14 seconds
[08/04/00 14:28:55](0)    Server: PSNT looking for work
[08/04/00 14:28:55](5)    Server: PSNT checking status...
[08/04/00 14:28:55](5)    Server action mode: Ok (looking for requests)
[08/04/00 14:28:55](5)    Checking Process cancels...
[08/04/00 14:28:55](4)    Checking status of active processes...
[08/04/00 14:28:55](5)    Process 9836 is still running as Session ID 711
[08/04/00 14:28:55](5)    Process 9837 is still running as Session ID 634
[08/04/00 14:28:55](5)    Process 9838 is still running as Session ID 703
[08/04/00 14:28:55](5)    Info for array of Request(s) associated with a Job slated to be submitted
[08/04/00 14:28:55](5)       Size of array: 0
[08/04/00 14:28:55](5)    Info for array of Active Processes
[08/04/00 14:28:55](5)       Size of array: 3
[08/04/00 14:28:55](5)       Crystal                 :  Active: 3   Max: 3
[08/04/00 14:28:55](5)    Info for array of Queued Request(s) found in Process Request table
[08/04/00 14:28:55](5)       Size of array: 16
```

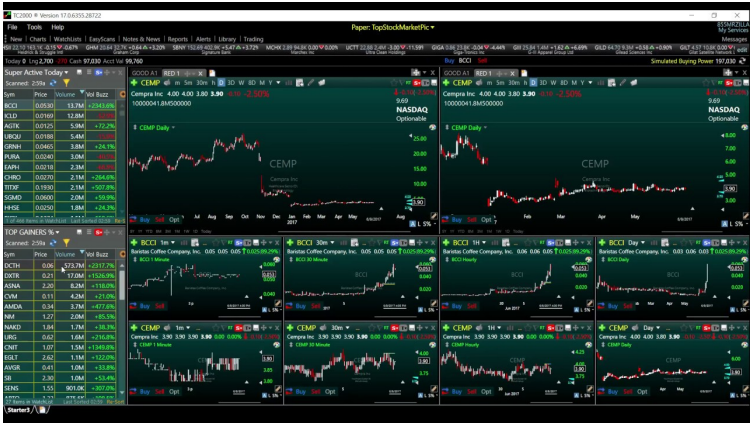
(iv) Datos no estructurados: Grafos



(v) Datos no estructurados: Audio, Imagen y Video

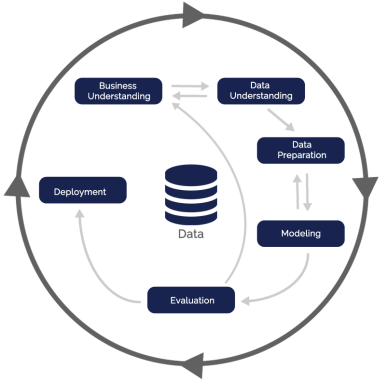


(vi) Datos estructurados/ no estructurados: Datos en Streaming



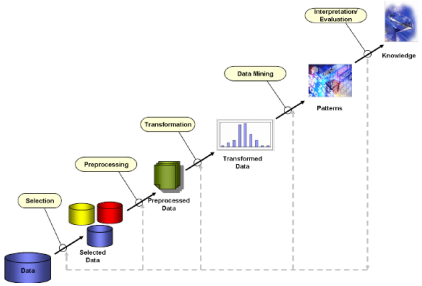
3. El proceso de la ciencia de datos

3. El proceso de la ciencia de datos: Un acercamiento



La metodología **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*) es un proceso de ciclo de vida para un proyecto de datos, en donde subdivide el proyecto en 6 fases. No obstante, la secuencia no es estricta, de hecho, la mayoría de los proyectos avanzan y retroceden entre fases según sea necesario.

3. The Data Science Process: An approaching



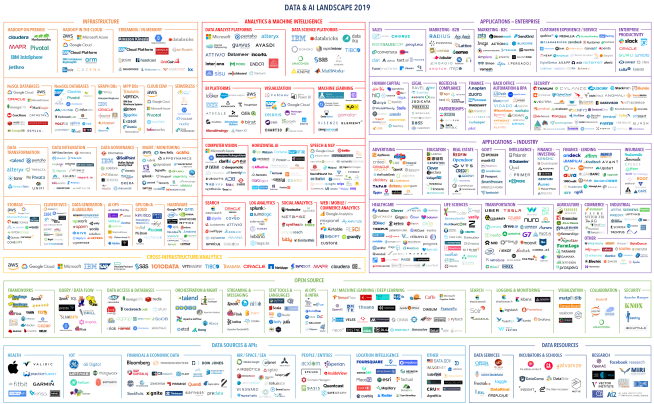
KDD (Knowledge Discovery in Databases) is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.

3. The Data Science Process: An approaching

SAS Institute defines data mining as the process of Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA) large amounts of data to uncover previously unknown patterns which can be utilized as a business advantage.

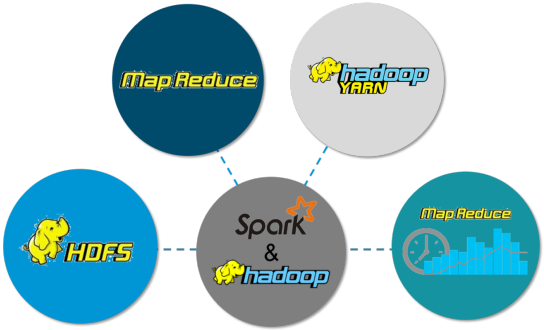
4. Ecosistema de los datos: Frameworks y Herramientas

4. Ecosistema de los datos: Frameworks y Herramientas



Data Everywhere

4.1 Sistemas de archivos distribuídos



4.2 Lenguajes de programación

4.3 Frameworks de Machine Learning

4.4 Bases de datos SQL/NoSQL

A. Herramientas para el curso

A.1 Anaconda

Anaconda es una distribución de Python (y R). Es gratuito y de código abierto, y simplifica la administración y la implementación de paquetes. Es la plataforma estándar para la ciencia de datos de Python y el aprendizaje automático de código abierto.

A.2 PostgreSQL

PostgreSQL, también conocido como Postgres, es un sistema de administración de bases de datos relacionales (RDBMS) gratuito y de código abierto que enfatiza la extensibilidad y el cumplimiento de SQL.

I. Data Science & Big Data: Basics

Introduction to Data Science

Data Everywhere