



Data Science

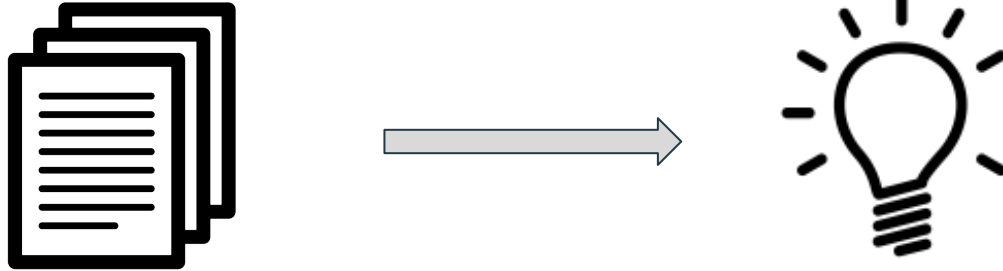
Antonio Jesús Gil

Contexto

1. **Ciencia de datos**
2. **Datos**
3. **Científico de datos**
4. **Proceso de ciencia de datos**

[illegible]

Objetivo común:



Partir de datos y obtener conocimiento

Ejemplo: Análisis estadístico de datos

- Explicar los datos
- Vocación descriptiva
- Se cumplen las hipótesis
- Inferencia estadística



Ejemplo: Aprendizaje automático

Data

[illegible]

Algorithm

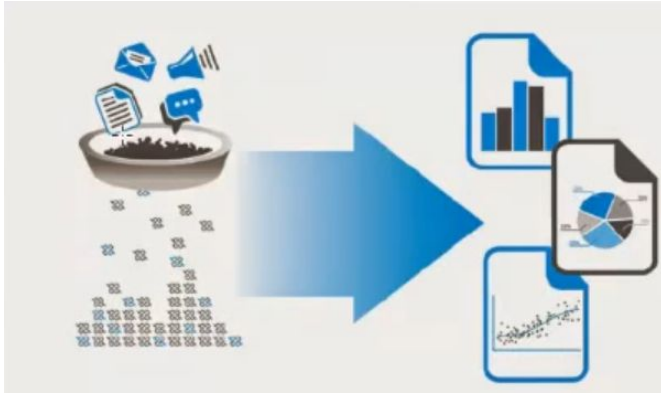


Model

$$f(x)$$

- Obtener **buenos modelos** a partir de los datos
- Selección y validación de modelos
- Vocación “**predictiva**”
- Más ciencia, menos datos

Ejemplo: Ciencia de datos



Ejecución de los modelos

- **Valorización** de los datos
- Más datos, menos ciencia
- Objetivo: valor a partir de datos



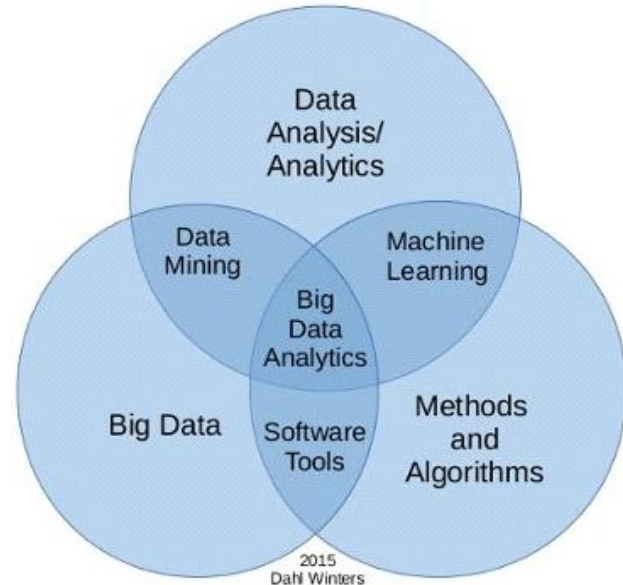
Ejemplo: Big Data



- **Big data** requiere de Infraestructura
- Necesita de los algoritmos ML
- Grandes volúmenes de datos
- Plataformas elásticas, cloud

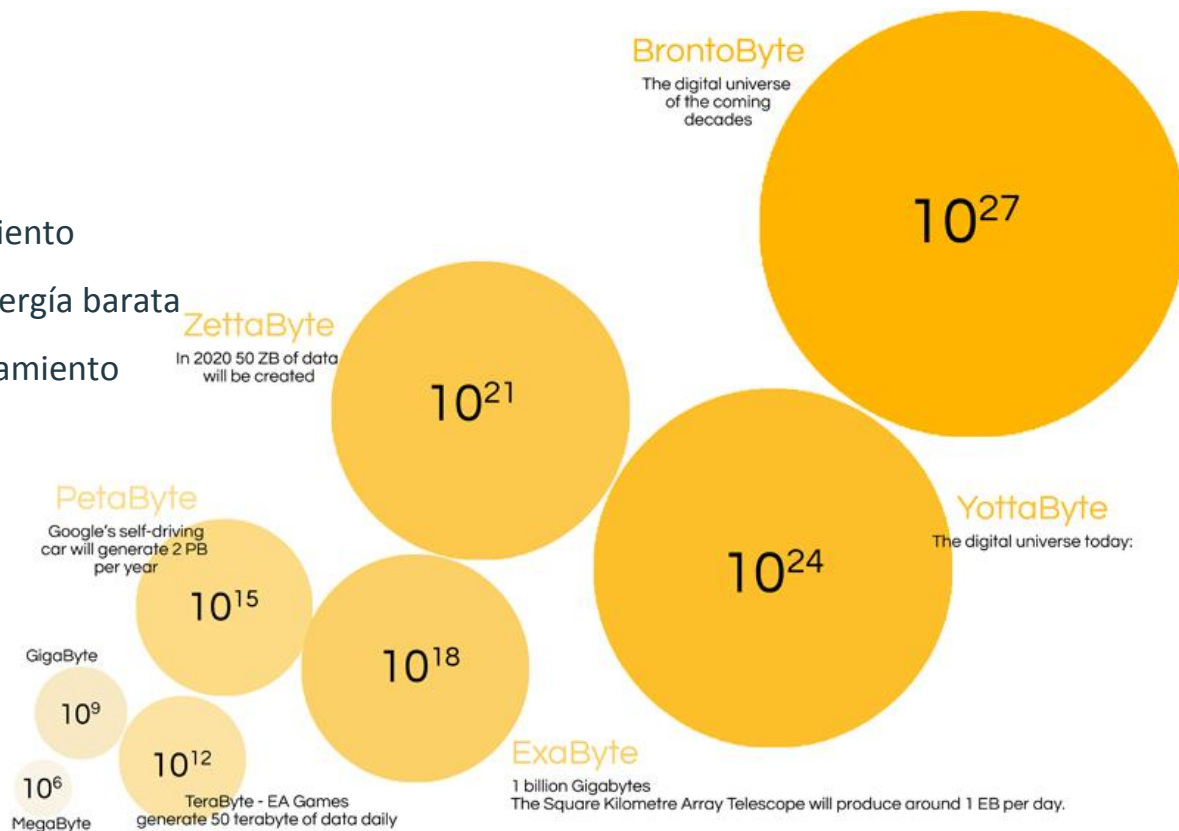
Big Data & Data Science

- Big data también son ficheros y bases de datos, configuración
- La analítica también es estadística y conocimiento del dato



Datos

- Coste de almacenamiento
- Aparecen granjas, energía barata
- Capacidad de procesamiento



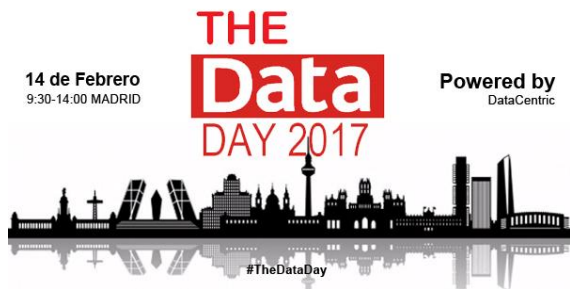
IS DATA REALLY THE NEW OIL?



Datos II

Adquieren **valor estratégico** para las empresas

Se convierten en una nueva clase de **activo**



... Las **compañías que no son capaces de entender** la necesidad de plantear su negocio en torno a la **importancia del dato**, de su procesamiento, de la obtención de los mejores algoritmos y de la mejora continua de sus sistemas de Machine Learning están **destinadas a ser cada vez menos competitivas y desaparecer.**

Enrique Dans (14/02/2017)

Científico de Datos

Depende quien te lo diga:

- Un estadístico que vive en Silicon Valley
- Un informático que hace estadística



Josh Wills
@josh_wills



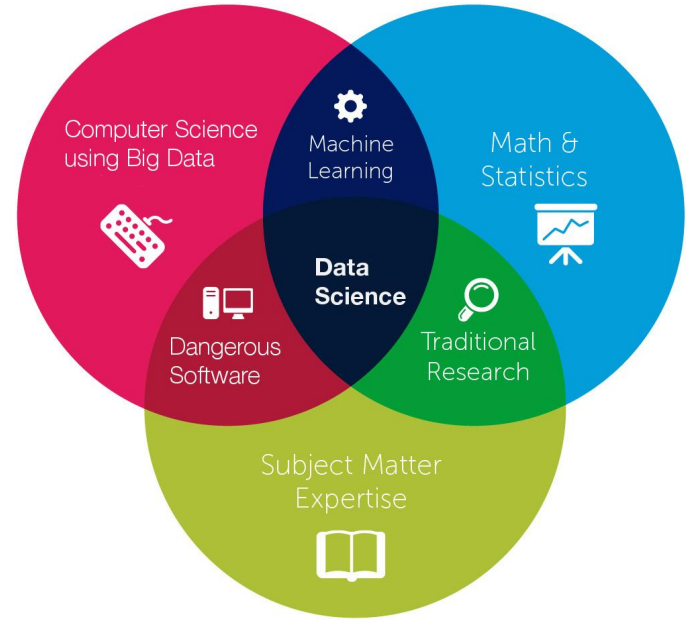
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS
1,326

LIKES
789



9:55 AM - 3 May 2012



“Data scientist is
the sexiest job
of the 21st century.”

Harvard Business Review



Ciencia de datos:

“La **ciencia de datos** es un campo interdisciplinar que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados”

La ciencia de datos cierra el círculo, desde la recolección de datos reales a su procesamiento y análisis y a influenciar en el mundo real otra vez

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st 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

DOMAIN KNOWLEDGE & SOFT SKILLS

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

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, 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

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.

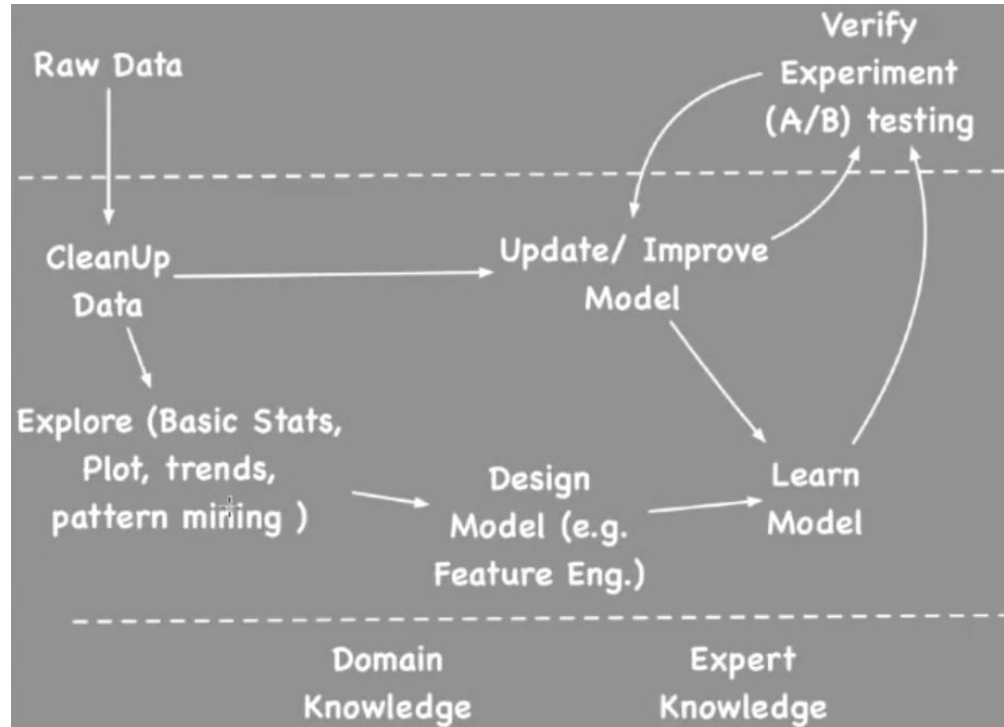
Marketing
DISTILLERY
©2 Krzysztof Zawacki

si si, pero ... ¿Qué es ciencia de datos?

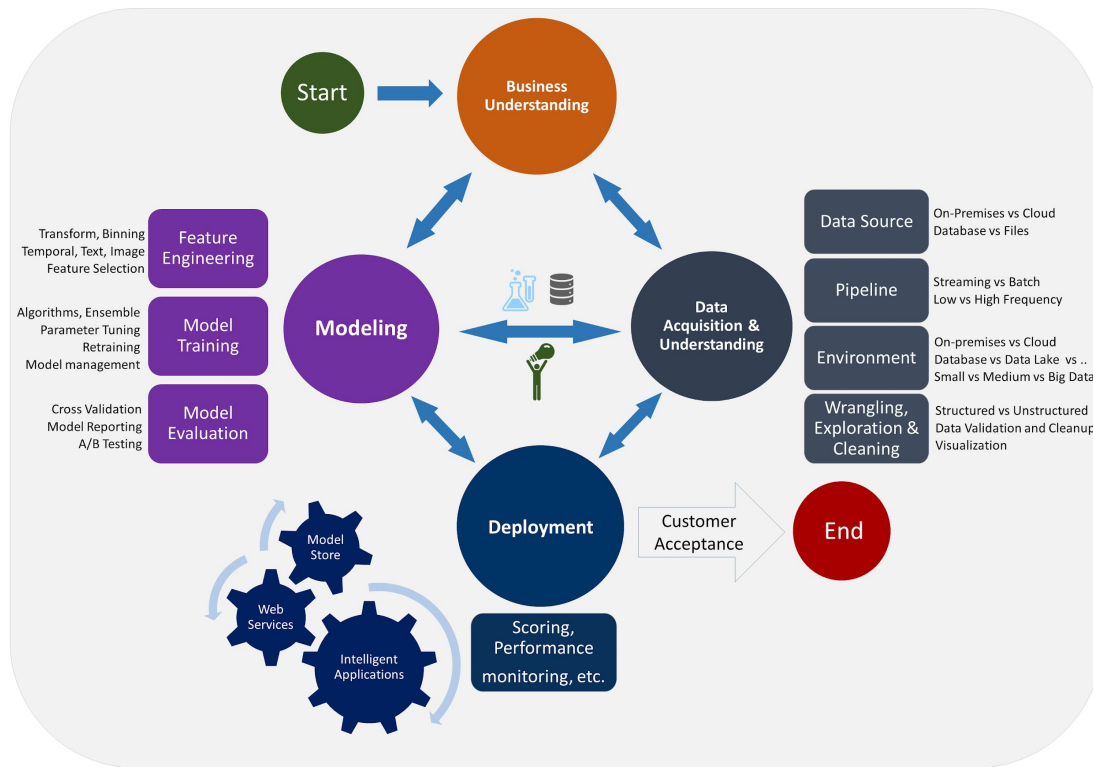
“la capacidad de **tomar** datos y ser capaz de comprenderlos, procesarlos, extraer valor de ellos, visualizarlos y comunicarlos”, con el objetivo de **crear productos**

- Datos → Productos basados en datos
- Transacciones tarjetas de crédito → detección del fraude
- Datos de sensores → smart cities / home
- Texto, datos en rrss → Satisfacción consumidores
- Logs → Diagnóstico automático
- Datos genómicos → Tratamientos personalizados

Proceso de ciencia de datos



Proceso de ciencia de datos II



Caso de estudio: Netflix prize

video → <https://youtu.be/ImpV70uLxyw>





















- En 2006 Netflix una competición:
- \$1 Million por mejorar **cinematch** en al menos un 10% (Reducción del error)
- Conjunto de datos basado en usuario, película y rating
- Sistema de recomendaciones

Contenidos

- **Módulo 1: Jupyter y Pandas**
- **Módulo 2: Recolección, preparación y almacenamiento de datos**
 - **JSON, Expresiones Regulares, Web Scraping, APIs REST**
- **Módulo 3: Explorar y visualizar, Matplotlib, seaborn, bokeh**

Módulo 1: ¿Python? ... why not

- Es un potente lenguaje de programación
- Es interactivo, resultados visibles al momento
- Dispone de una gran comunidad, solucionan problemas
- Fácil aprendizaje
- <https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2018>

Language Rank	Types	Spectrum Ranking
1. Python	  	100.0
2. C++	  	99.7
3. Java	  	97.5
4. C	  	96.7
5. C#	  	89.4
6. PHP		84.9
7. R		82.9
8. JavaScript	 	82.6
9. Go	 	76.4
10. Assembly		74.1

Python

- **Scripting:** No necesita compilar. Permite ejecutar código directamente en un intérprete
- **POO:** Útil para describir modelos, datasets ...
- ¿Qué es un intérprete?
- **Entorno de trabajo** que contiene la información de lo que se está ejecutando (variables, funciones, paquetes importados...)
- Aislados: dos entornos distintos **no comparten** nada

Entorno de desarrollo

- Docker
- Jupyter Notebook, Orígenes:
 - iPython → shell interactivo de Python con soporte para gráficos e interfaz
 - Jupyter → Viene de Julia, Python y R y soporta más de 40 lenguajes (instalando su kernel)
 - Interactivo → proporciona una web interactiva. Cliente / Servidor
 - Gestión → Como un paquete más de Python

Tratamiento de datos

- Tidy data: “Wickham, H (2014). Tidy data. Journal of Statistical Software”
- Filosofía de organización y almacenaje de los datos para su posterior tratamiento y filtrado de los mismos, Python dispone:
 - Pandas: Representación y tratamiento de datos con filosofía Dataframe de R
 - Matplotlib y Seaborn: análogos a ggplot2 de R o Matlab
 - ScikitLearn: Machine Learn aplicado a problemas tradicionales
 - PySpark: Machine Learn aplicado a problemas Big data

Jupyter!

```
docker run -p 8888:8888 -v "$PWD":/home/jovyan ajgil/ds-notebook
```

La pantalla principal es **Home**. Se diferencian dos estados o modos de trabajo:

- Modo de comandos: No aparece el cursor, cambia a este estado pulsando **Esc**
- Modo de edición: Aparece el cursor en la celda, cambia a este cursor haciendo click en la celda o pulsando **Enter**

Shortcuts: Consulta en modo comandos pulsando **H** o en la barra de menú Help / Keyboard Shortcuts

Lab1 - Jupyter

Pandas

Basado en estructuras de datos denominadas Dataframes → Bidimensionales
indexado implícito por filas y columnas (Excel)

Series

- Estructuras de nivel inferior
- Indexado explícito
- Se indexan mediante tags
- Implementan muchas operaciones
- Es la forma natural de operar con un Dataframe



Series

- Partiendo de una **colección** de elementos
- Desde una colección y un índice debe ser “**hashable**”
- Desde un **diccionario**
- Podemos dar un nombre tanto a la **serie** como al **índice**
- Pueden ser **accedidas** mediante su **índice** o **posición**: Funciones `loc[]` y `iloc[]`

Pandas II

Datos categóricos

- Son tipo de datos específico para datos categóricos (número limitado de valores)
- Se visualizan generalmente como Strings
- Internamente se representan como enteros
- Tratamiento más eficiente
- Permite operaciones

Pandas III

Dataframes

- Estructura de datos bidimensional
- Se indexa por filas
- Las columnas son accedidas individualmente
- Cada columna puede almacenar datos diferentes
- Son tratadas como objetos de tipo `Series`

Columns

	<i>Name</i>	<i>Team</i>	<i>Number</i>	<i>Position</i>	<i>Age</i>
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

Rows

Data

Pandas III

Dataframes

Se construyen desde:

- Diccionarios

```
df_titanic = pd.read_csv('datos/Titanic.csv',  
                           names=['ID', 'Nombre']  
df_titanic.head()
```

Lectura de archivos → csv, json, excel

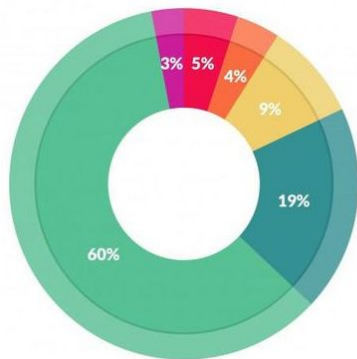
- Series
- Listas
- Colecciones

Lab2 - Pandas

Módulo 2: Recolectar, preparar, almacenar

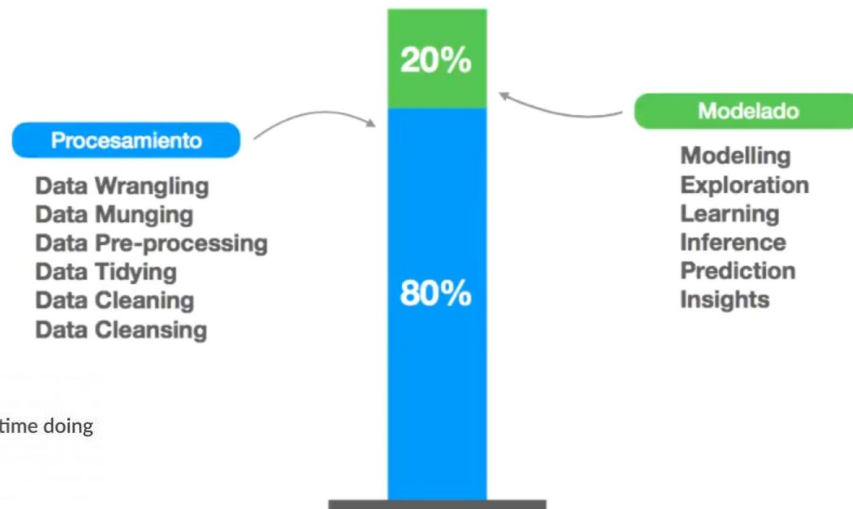
- Resolver : ¿De dónde vienen? ¿Cómo están almacenados? ¿Qué forma tienen?
- Fuentes y tecnologías de recolección y acceso a datos
- Formato de los datos y herramientas
- Tecnologías de recolección y procesamiento de datos
- Almacenamiento

Data scientist time



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



Habilidades + Objetivos

- Muy tecnológico con orientación a desarrollo de software
- Análisis descriptivo / exploratorio
- Modelado / Inferencia
- Machine Learning
- Producto (Servicio, Visual, ...)

Orígenes de los datos

Transacciones

Instituciones
(open data)

Sensores IoT

Procesos
Industriales

Ciencia

Medicina

“Data is the new Oil”



Formato de los Datos

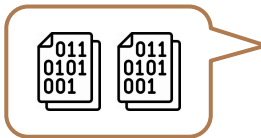
Los datos originales no vienen en un formato propicio para su análisis

Necesitan de ser procesados

Ficheros sin estructurar / desorganizados



Número indeterminado de ficheros excel
con una o varias hojas en su interior



Ficheros binarios o con formato privativo,
mediciones de sensores o equipo industrial

Semi-estructurados: Extracto salida API

```
    "result_type": "Recent"
  },
  "source": "<a href='\"https://mobile.twitter.com\"' rel='\"nofollow\"'>Twitter Web App</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 2244994945,
    "id_str": "2244994945",
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "description": "Your official source for Twitter Platform news, updates & events. Need",
    "url": "https://t.co/FGl7V0ULyL",
    "entities": {
      "url": {
        "urls": [
          {
            "url": "https://t.co/FGl7V0ULyL",
            "expanded_url": "https://developer.twitter.com/",
            "display_url": "developer.twitter.com",
            "indices": [
              0,
              23
```

Semi-estructurados: Extracto HTML

```
<h1 id="firstHeading" class="firstHeading" lang="en">Web scraping</h1>

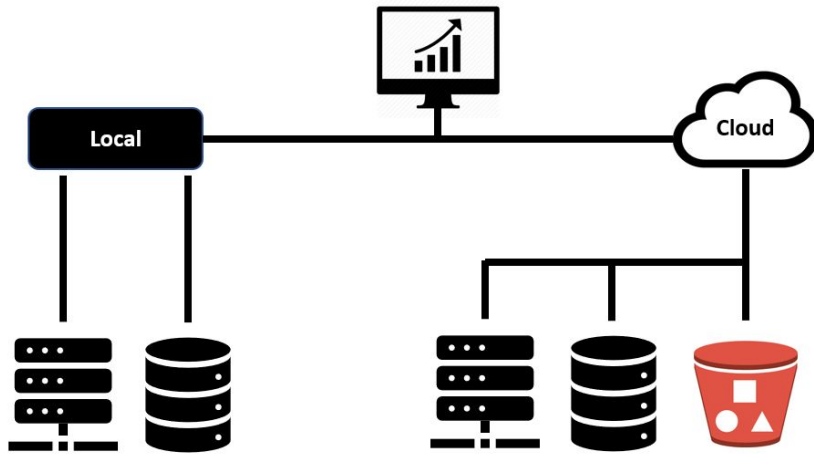
<div id="bodyContent" class="mw-body-content">
  <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia</div>
  <div id="contentSub"></div>

  <div id="jump-to-nav"></div>
  <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
  <a class="mw-jump-link" href="#p-search">Jump to search</a>
  <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr"><div class="mw-parser-output">
    <table class="box-More_citations_needed plainlinks metadata ambox ambox-content ambox-Refimprove" ro
      <tbody><tr><td class="mbox-image"><div style="width:52px"><a href="/wiki/File:Question_book-new.
        
        <div role="note" class="hatnote navigation-not-searchable">For broader coverage of this topic, see <a href="/wiki
        <p><b>Web scraping</b>, <b>web harvesting</b>, or <b>web data extraction</b> is <a href="/wiki/Data_scraping" ti
        </p><p>Web scraping a web page involves fetching it and extracting from it.<sup id="cite_ref-Boeing2016JPER_1-1"
        </p><p>Web scraping is used for <a href="/wiki/Contact_scraping" title="Contact scraping">contact scraping</a>,</p>
      </td></tr></tbody></table>
    </div>
  </div>
```

Datos no digitales

Día	Precipitación mm	Meteoros observados			Viento Km/h e		Temperatura °C		Humedad %		Presión hPa	
		M.	T.	N.	Máx.	Dir.	Máx.	Mín.	Máx.	Mín.	Máx.	Mín.
1	0.4	☉			37	E	20.0	17.1	83	56	1021.5	1018.8
2	2.0	☉			26	ES	21.8	15.8	90	66	1020.9	1015.0
3	10.0	☉ ☉			25	W	21.1	18.7	90	73	1016.7	1014.8
4	0.0				35	NW	20.6	17.6	82	63	1019.6	1016.5
5	0.0				35	NW	24.8	19.3	90	63	1016.6	1010.0
6	0.0				54	NW	28.7	20.6	85	32	1016.6	1006.5
7	0.0				13	W	28.3	18.7	80	39	1017.7	1016.3
8	3.2		☉	☉	25	W	27.1	16.2	86	42	1019	1017.7
9	4.3	☉			44	E	22.2	15.1	88	55	1023	1029
10	0.0				80	SW	24.3	16.6	87	59	1019	1009
11	0.0				31	SW	25.4	15.3	88	71	1009	1008
12	0.0				37	SW	25.2	15.4	90	50	1015	1008
13	0.0				25	W	25.1	15.8	81	48	1014	1013
14	0.0				28	NW	24.3	15.1	75	44	1016	1013
15	0.0				26	N	24.5	15.0	79	45	1018	1016
16	0.0				33	SE	24.2	14.3	85	49	1019	1018
17	0.0				32	E	24.7	15.4	90	53	1022	1018
18	0.2	A			15	W	23.3	15.8	79	54	1023	1020
19	0.0				12	SE	23.2	16.3	89	48	1025	1023
20	0.3	A			10	SE	22.7	15.3	88	59	1028	1025
21	0.0				20	SW	22.8	14.2	79	58	1027	1023
22	0.0				28	S	22.9	14.8	80	59	1023	1026
23	0.0				31	S	24.5	14.3	80	60	1016	1005
24	7.3		☉		35	E	21.2	13.2	77	63	1005	1003
25	0.0				30	NE	20.5	14.3	78	62	1003	1003
26	15.4		☉ R ☉ B		44	NE	20.1	15.1	90	79	1005	1003
27	17.7	☉ ☉	☉		48	E	20.0	14.1	100	80	1005	1002
28	25.3	☉ ☉	☉ ☉		56	SE	18.7	14.3	100	75	1007	1003
29	0.0	≡			31	SW	16.8	12.1	100	65	1011	1003
30	0.0				25	SW	17.9	12.7	77	60	1011	1005
31	2.1			☉	31	E	17.5	12.6	89	54	1018	10011

Acceso y recolección de los datos



- El acceso a los datos en la nube se hace generalmente a través de interfaces programáticas
- Los detalles de las conexiones acaban formando parte de las estrategias de procesamiento de los datos (API - SDK)

Propiedades de los datos y tecnologías de recolección

¿Es la tecnología?

- Escalable:
 - Velocidad acceso
 - Capacidad almacenamiento
- Segura

¿Son los datos?

- Completos
- Veraces
- Estáticos, dinámicos, temporales (streams)
- Legales

Formato de los datos

Podemos clasificar tres formatos básicos de los datos conforme a la estructuración de la información

- **Datos estructurados:** Toda la información está identificada conforme a un patrón común y definido que representa un modelo. Óptima para ser analizada Ej. Modelo relacional, base de datos SQL
- **Datos semi-estructurados:** La estructura está definida y es conocida, pero no responde a un modelo estricto. Óptima para ser procesada, Ej. JSON, XML, MD
- **Datos no estructurados:** No representan un modelo y carece de estructura conocida. Requiere ser transformada para su análisis, Ej, multimedia, texto libre

No estructurados:

- **Documentos** de texto, imágenes, video, audio y señales
- **Diversidad** de formatos; texto plano, binarios, formato privado, codificados, encriptados.
- En ocasiones nos encontramos datos **híbridos** ya que contienen **metadatos**

Data Wrangling: Transformando los datos

- Objetivo : transformar datos en bruto (**raw data**) en un formato apropiado y valioso desde el punto de vista del análisis.
- El proceso debe realizarse a partir de la forma más pura posible de los datos en bruto y debe documentarse para ser reproducible.
- No es suficiente con obtener un conjunto de datos estructurado ya que no siempre es óptimo para el análisis.

DATOS ESTRUCTURADOS \neq BUENOS DATOS

- Generalmente buscamos que los datos mantengan unas propiedades deseables, conocidas como: **Tidy Data**

Tidy Data

“Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types”

Hadley Wickham 2014 (Tidy data - Journal of Statistical Software)



Un dataset tidy mantiene las siguientes **propiedades**:

- Cada variable representa una columna
- Cada observación representa una fila
- Cada unidad observacional representa una tabla

Permite definir objetivos, estrategias y herramientas **estandarizadas** para la limpieza y transformación de datos.

Permite definir un vocabulario y operadores de transformación desde un punto de vista **agnóstico** a cualquier lenguaje

Data Wrangling: De “raw data” a “tidy data”

- Transformación **reproducible** entre formatos cuyo resultado general es un “**dataset**” formado por una o mas tablas con información semántica asociada:
 - Fenómenos observacionales: Que representan las observaciones de cada tabla y su contexto experimental.
 - Diccionario de variables (origen, tipo, dominio, restricciones)
 - Índices y relaciones que relacionen observaciones entre tablas
- Diferentes tareas: (nomenclatura no **estándar**)
 - Data **scraping**: extracción de fragmentos relevantes en un corpus
 - Data **wrangling / munging**: Operaciones destructivas entre formatos
 - Data **cleaning**: Detección y limpieza de errores, valores perdidos
 - Data **tyding**: Adaptar datos estructurados al formato tidy

Recolección y procesamiento de los datos

Toda información a ser recolectada se encuentra en un servidor

Fuera del entorno corporativo la mayor fuente de datos se encuentra en **Internet**

Nos centramos en peticiones a **servicios web**

HTTP 101

Stateless: La comunicación se realiza en secuencias de pregunta / respuesta independientes entre sí, sin asumir detalles ni coordinación entre las partes.

Request: Petición al servicio

- **URL** (Unified Resource Locator) Descripción del recurso a solicitar

protocolo://host:puerto/recurso/query

- **Verbo:** Acción a realizar sobre el recurso solicitado, GET, POST, PUT, DELETE

HTTP 101

- **Response:** El mensaje de respuesta enviado desde el servidor
 - **Code:** Código de estado indicando el resultado de la operación
 - 200 -> Éxito
 - 400 -> Error en la petición
 - 500 -> Error interno del servidor
 - **Body:** Contenido del mensaje con el recurso solicitado, mensajes de información o errores.
- **Headers:** Presentes tanto en Request como Response para indicar metadatos y propiedades de la petición tales como tipo de contenido, autenticación y detalles de la conexión.

APIs - Application Programming Interface

- Conjunto de **subrutinas, protocolos y herramientas** que definen la comunicación entre distintos componentes software
- En el contexto **web**, un API permite **encapsular** un **servicio** detrás del protocolo HTTP, definiendo una serie de recursos y acciones sobre los mismos
- Puede utilizarse para definir **consultas y mutaciones** sobre una base de datos o sobre otra serie de recursos accesibles, partiendo desde un servidor web
- Necesario consensuar y documentar el esquema y el modelo que se publica para aportar significado a las operaciones y datos obtenidos
- El estándar de facto es **RESTfull**
- **GET** nos permite consultar el modelo subyacente en la API,
una salida en formato **JSON**

RESTful API
GET PUT POST DELETE

Disponibilidad de datos en servicios web y APIs

- Podemos encontrar una gran cantidad de APIs y otros servicios de datos **públicos**
- Algunas de estas APIs pueden establecer **límites de acceso** o métodos de **autenticación**
- Ejemplos de APIs:
 - Twitter permite el acceso a tweets, hashtags y otros datos de interacción social
 - Google Books permite acceder a un catálogo inmenso de libros, incluyendo abstracts y portadas
 - SWAPI es un recurso educativo que publica un API proporcionando un extenso índice relacional de elementos del universo Star Wars.

Postman

<https://www.googleapis.com/books/v1/volumes?q=rings>

The screenshot shows the Postman interface with a GET request to `https://www.googleapis.com/books/v1/volumes?q=rings`. The request is successful, returning a 200 OK status. The response body is displayed in JSON format, showing details for the book "The Lord of the Rings" by J.R.R. Tolkien, published by Houghton Mifflin Harcourt in 2012.

```
{
  "id": "yl4dILkcqm4C",
  "etag": "08mYk2ibtJE",
  "selfLink": "https://www.googleapis.com/books/v1/volumes/yl4dILkcqm4C",
  "volumeInfo": {
    "title": "The Lord of the Rings",
    "subtitle": "One Volume",
    "authors": [
      "J.R.R. Tolkien"
    ],
    "publisher": "Houghton Mifflin Harcourt",
    "publishedDate": "2012-02-15",
    "description": "A PBS Great American Read Top 100 Pick One Ring to rule them all, One Ring to find them, One Ring to bring them all and in the darkness bind them In ancient times the Rings of Power were crafted by the Elven-smiths, and Sauron, the Dark Lord, forged the One Ring, filling it with his own power so that he could rule all others. But the One Ring was taken from him, and though he sought it throughout Middle-earth, it remained lost to him. After many ages it fell by chance into the hands of the hobbit Bilbo Baggins. From Sauron's fastness in the Dark Tower of Mordor, his power spread far and wide. Sauron gathered all the Great Rings to him, but always he searched for the One Ring that would complete his dominion. When Bilbo reached his eleventy-first birthday he disappeared, bequeathing to his young cousin Frodo the Ruling Ring and a perilous quest: to journey across Middle-earth, deep into the shadow of the Dark Lord, and destroy the Ring by casting it into the Cracks of Doom. The Lord of the Rings tells of the great quest undertaken by Frodo and the
```

Tecnologías para el procesamiento de datos

- Las técnicas aplicadas dependen de la **naturaleza** de los datos y del **objetivo**
 - **Texto plano, lenguaje natural: Técnicas NLP**
 - **Imágenes, audio y señales, anotación automática. Deep Learning**
 - **Datos no estructurados, HTML, logs, expresiones regulares**
 - **datos estructurados y semi-estructurados: Pandas, R**

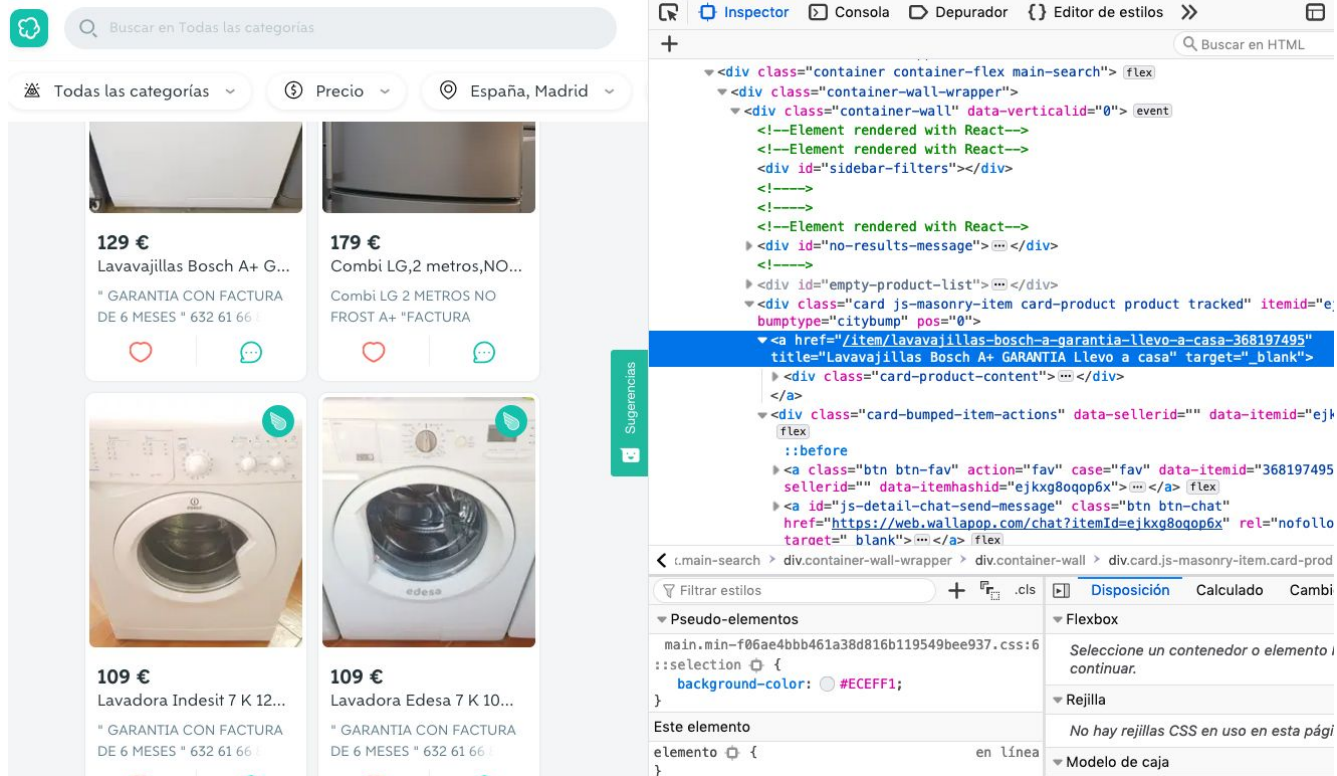
Expresiones Regulares

- Son la herramienta básica de análisis de texto no estructurado
- Permiten definir un **patrón de búsqueda** mediante un **lenguaje formal** que es contrastado con el cuerpo del mensaje si existe coincidencia
- Las implementaciones del lenguaje formal pueden variar, pero los operadores y la estrategia de aplicación es siempre la misma
- Podemos emplear expresiones regulares para **analizar cadenas** de texto e identificar los datos que necesitamos.

Web Scraping

- Solo una pequeña parte de los datos se ofrece de forma estructurada o semiestructurada
- La mayor parte, el **80% solo están** disponibles como no-estructurados
- Web scraping consiste en extraer información de la web, combinando el uso del HTTP para obtener los documentos e interactuar con las páginas de manera automática, utilizando parsers y expresiones regulares sobre los mismos
- De este modo un bot o crawler puede analizar automáticamente un conjunto de páginas y procesar toda la información necesaria para devolverla en formatos con estructura para su análisis

Web Scraping II



Datos semiestructurados en Python

- Una de las ventajas del formato JSON es su similaridad con las **estructuras de datos** usadas en muchos lenguajes de programación, equivalentes a la notación de objetos en Javascript y muy similar a los **diccionarios de Python**.
- Estas propiedades son extensivas a otros lenguajes de serialización con **XML** o **YAML**
- El paquete *json* permite codificar y decodificar JSON

```
student1 = Student(first_name="Jake", last_name="Doyle")
student2 = Student(first_name="Jason", last_name="Durkin")
team = Team(students=[student1, student2])
```

```
// Serialization
json_data = json.dumps(team, default=lambda o: o.__dict__, indent=4)
print(json_data)
```

```
// Deserialization
decoded_team = Team(**json.loads(json_data))
print(decoded_team)
```

Datos estructurados en Python

- Datos en formato tabular se suelen relacionar con bases de datos relacionales y SQL
- Usar bases de datos para análisis exploratorio es costoso e introduce un grado de **complejidad adicional** no deseado.
- Las aplicaciones de ciencia de datos suelen trabajar con representaciones tabulares de los datos en memoria (sino caben, Apache Spark)
- Pandas es el candidato ideal ya que no solo no permite obtener una representación a bajo nivel (series) de los datos, sino que provee una serie de operadores de alto nivel equivalentes a cualquier implementación SQL
- https://pandas.pydata.org/pandas-docs/stable/getting_started/comparison/comparison_with_sql.html

Datos estructurados en Python II

- Pandas no es único en su naturaleza ya que en concepto Dataframe está arraigado en otros lenguajes, como R
- Es importante dominar la sintaxis y fundamental conocer las operaciones y el formato de los datos a un **mayor nivel de abstracción**
- Se aprende un vocabulario de manera **agnóstica al lenguaje** y podremos definir operaciones realizadas sobre los datos para garantizar su **reproducibilidad y portabilidad**
- Usar el estándar Tidy data garantiza un conjunto de operaciones bien definidas además de optimizar el **cómputo** y aprovechar **operaciones vectorizadas**
- **Data wrangling with Pandas cheat sheet**

Lab3 - Expresiones Regulares

Lab4 - Web Scraping

Lab5 - Consumo APIs

Lab6 - Datos estructurados Pandas (tidy)