KSCHOOL

Máster Data Science

Barcelona, 2019.

Henry Navarro

KSCHOOL

About me

- Lead Data Scientist: Altran Technologies.
- Data Scientist: Equifax – Solutio – Enefgy.
- Profesor Universidad Carlos III de Madrid – Grupo ML4DS & GISC. (https://goo.gl/9eeHAz)
- Profesor Escuela de Organización Industrial.
- Data Analyst – Policía Ven.
- Máster en Ingeniería Matemática – UC3M.
- Lienciado en Matemática – UCV.

KSCHOOL

Intro Data Engineering

- El rol del data engineer ha ido avanzando gradualmente a medida que se han creado más herramientas.
- Es necesario en el desarrollo de procesos big data.
- A diferencia de los data scientists, existen metodologías para la ingeniería de datos.

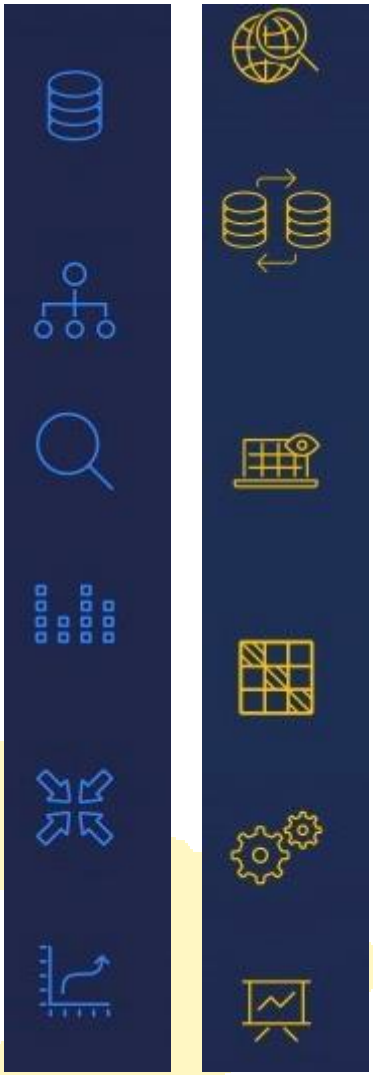
KSCHOOL

¿Hablamos de lo mismo?

¿Hablamos de lo mismo?

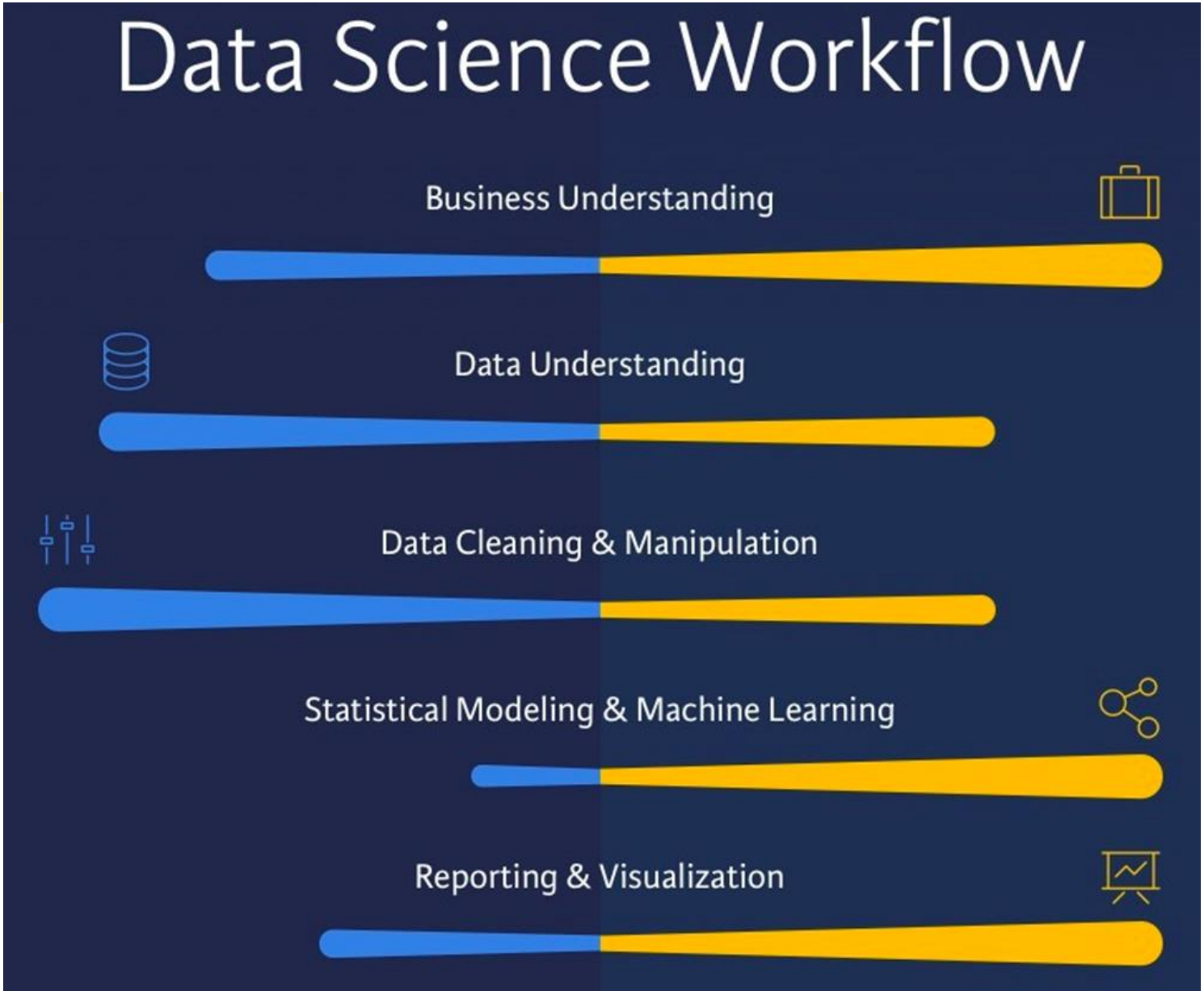
Data Engineer

- Desarrollar, construir, probar y mantener arquitecturas (como bases de datos y sistemas de procesamiento a gran escala).
- Implementar una arquitectura útil para los casos de negocios
- Desarrollo de oportunidades para la adquisición de datos.
- Establecer procesos de conjuntos de datos para modelización, minería y producción.
- Emplear una variedad de lenguajes y herramientas para conectar sistemas.
- Recomendar maneras de mejorar la fiabilidad, eficiencia y calidad de los datos.

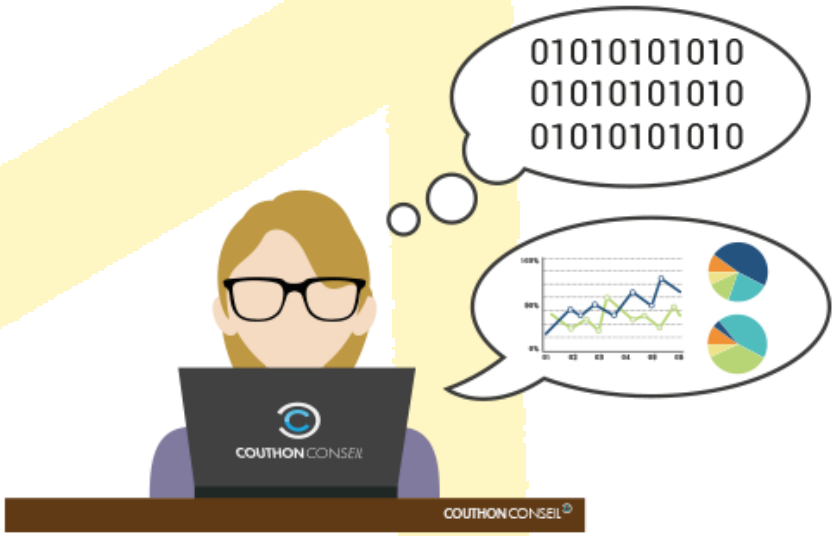


Data Scientist

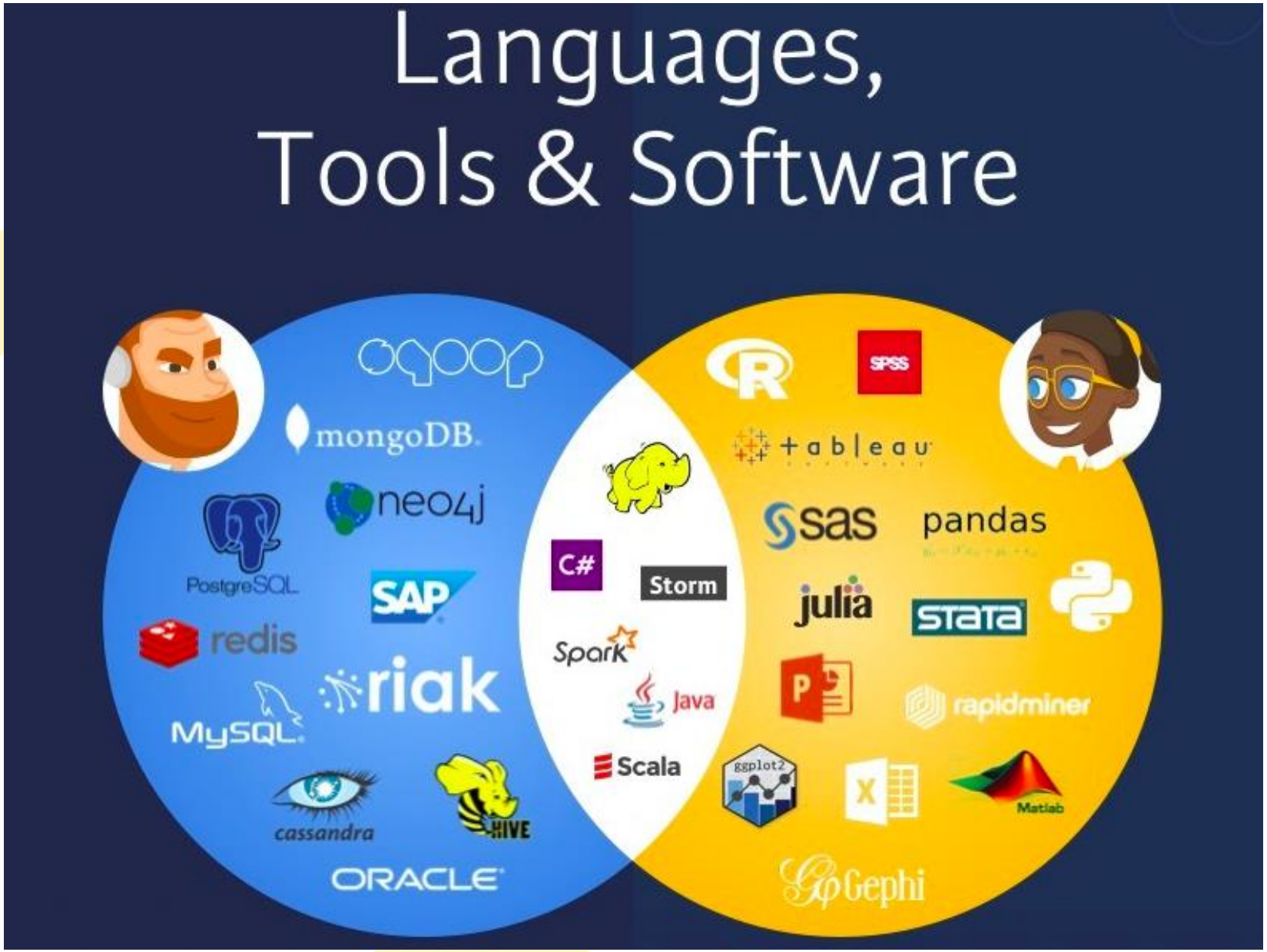
- Llevar a cabo desarrollos para responder a preguntas de la industria y los negocios.
- Aprovechar grandes volúmenes de datos de fuentes internas y externas para responder a ese negocio.
- Implementar softwares de análisis, aprendizaje automático y métodos estadísticos para preparar los datos para su uso en modelos predictivos y explicativos.
- Explorar y examinar los datos para encontrar patrones.


Data Science Workflow
Business Understanding
Data Understanding
Data Cleaning & Manipulation
Statistical Modeling & Machine Learning
Reporting & Visualization

Data Analyst



- Los Data Analysts desempeñan un papel importante al solicitar información de una base de datos o realizar consultas.
- Pueden procesar y aprovechar los conjuntos de datos para proporcionar informes resumidos y visuales.
- La función principal de un analista de datos es inferir datos sin procesar para explorar patrones y sacar conclusiones.
- Utilizan procesos metodológicos y aplican algoritmos para llegar a estas conclusiones.
- Si bien un analista de datos puede usar algoritmos para respaldar su función, no se espera que los desarrollen ni suelen tener una sólida formación matemática o de investigación, su función requiere una comprensión básica de estadística, recopilación de datos y muy fuerte en visualización y análisis de datos exploratorios.
- Los analistas de datos ayudan a simplificar los datos complejos a informes y cuadros ad hoc, lo que permite a las empresas explotar al máximo sus datos.


Languages,
Tools & Software
hadoop
mongoDB.
neo4j
PostgreSQL
SAP
redis
riak
MySQL
cassandra
HIVE
ORACLE
C#
Storm
Spark
Java
Scala
R
SPSS
tableau
sas
pandas
julia
STATA
rapidminer
ggplot2
Matlab
Gephi

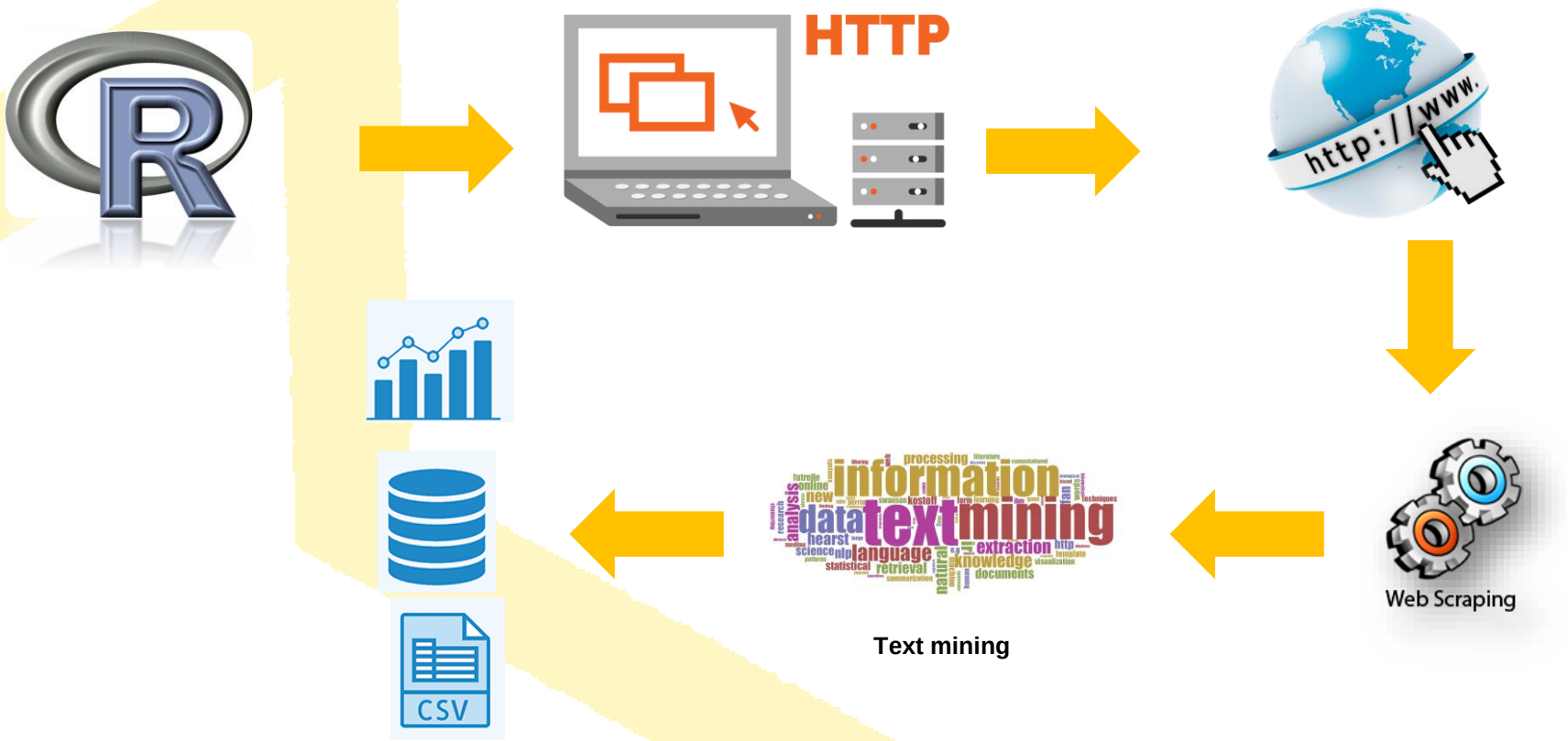
Index

- Extracción de datos mediante bots (Web scraping).
- Conexión a API's o Interfaz de programación de aplicaciones (Application Programming Inteface).
- Extracción de datos desde API's.
- Conexión desde R a Bases de Datos.
- Almacenamiento desde R a Bases de Datos.
- Introducción a procesos ETL (Extract, Transform and Load)

Intro: Web scraping

- ¿Cuál es la mayor “base de datos” del mundo?
- ¿Copiar y pegar datos podría tener sentido?
- Los mayores casos de uso de negocio son datos en directorios de páginas amarillas, sitios de bienes raíces, redes sociales, sitios de compras en línea, etc.
- Por un lugar debemos conocer un poco de html a nivel conceptual y por el otro debemos disponer de los conocimientos técnicos necesarios para lograr extraer datos
- Gestión de grandes cantidades de datos (big data).
- Plantear la estrategia de extracción debemos saber cómo y cuáles serán los datos que a extraer con la finalidad de poder dar un sentido informativo para un caso de uso de negocio.

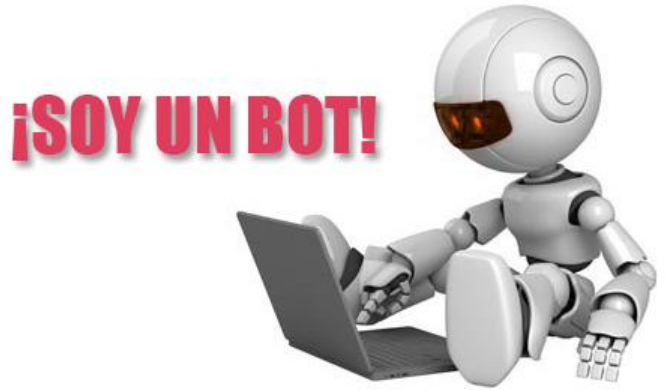
Web scraping



Text mining

Web scraping

- El web scraping es la técnica de extracción automática de datos de sitios web mediante software / script.
- Se puede extraer: imágenes, videos, texto, información de contacto como emails, números de teléfonos, etc.
- No es robusto, depende de la estructura de la web.
- Legalidad en la extracción de datos.
- Leyes de protección de datos.



Bibliografía

- Simon Munzert, Christian Rubba, Peter Meiner, and Dominic Nyhuis. 2014. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining (1st ed.). Wiley Publishing.
- Aydın, Olgun. (2018). R Web Scraping Quick Start Guide.

KSCHOOL

Let's do it...

API INTERFACE

KSCHOOL

API: Application Programming Interface

- Es una manera segura de extraer datos, tanto para la empresa como para el usuario.
- Requieren de una autenticación, en algunos casos complicada.
- Suelen devolver los datos en formato JSON.

Google Places API

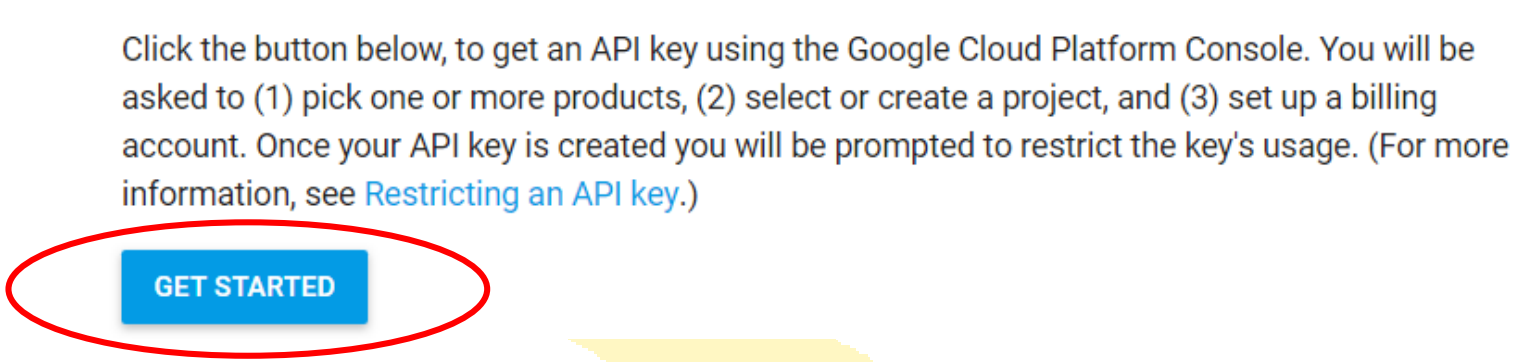
- Funcionalidades exactas a las de Google Maps.
- Todos los datos de Google Maps a disposición de una query.
- Documentación disponible: https://developers.google.com/places/web-service/intro?hl=es
- Soporte: https://stackoverflow.com/questions/tagged/google-maps. Con los tags Google-places, Google-maps, otras.

Google Places API

- Requiere de un API key que se obtiene con una cuenta Google desde el enlace https://developers.google.com/places/web-service/get-api-key.
- Fácil de utilizar.
- Documentación muy bien explicada.
- Bug reporting.
- Múltiples ejemplos.

¿Cómo obtener el API key?

1. Ir al siguiente enlace: https://developers.google.com/places/web-service/get-api-key#quick-guide



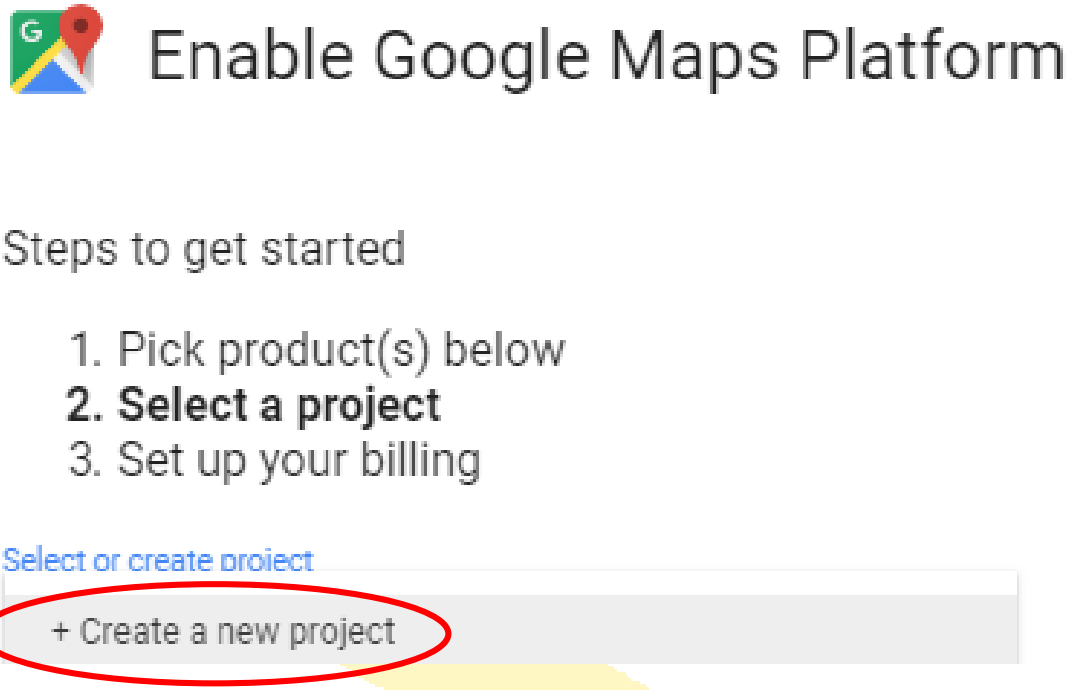
KSCHOOL

¿Cómo obtener el API key?

2. Seleccionar todos los en los servicios

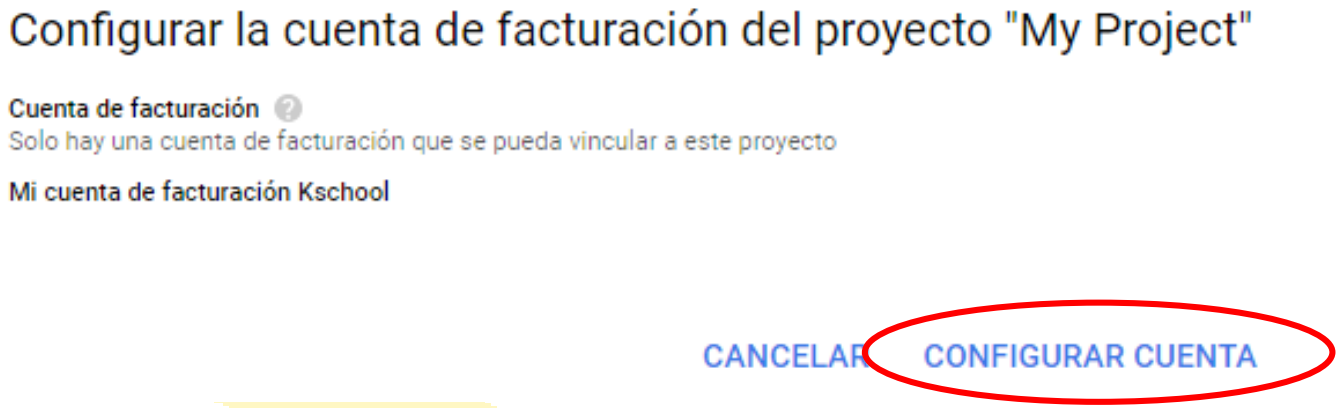
¿Cómo obtener el API key?

3. Crea un proyecto si aún no lo tienes creado



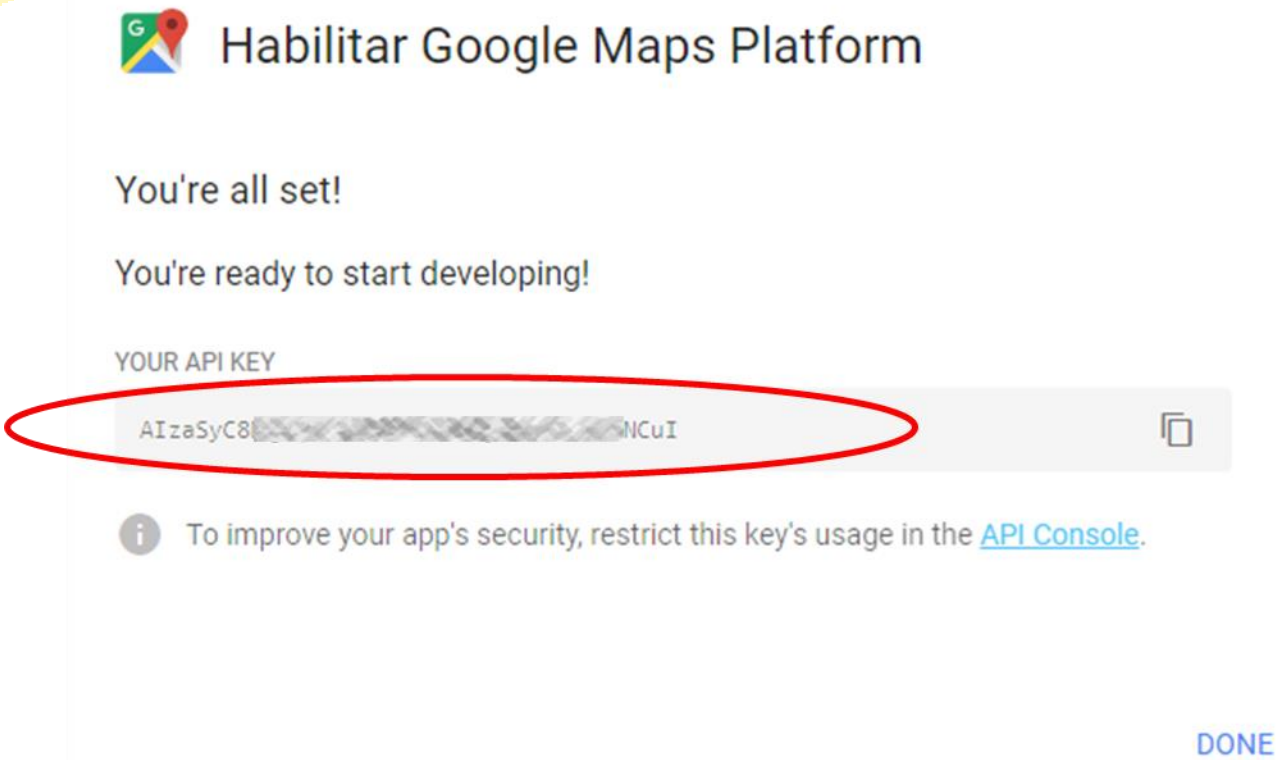
¿Cómo obtener el API key?

4. Se debe habilitar la facturación. Esto no implica ningún cobro, se tienen un número limitado de solicitudes.



¿Cómo obtener el API key?

5. ¡Tu API key ya está listo para usarse! (si tienes la tarjeta de crédito añadida)



KSCHOOL

Let's do it...

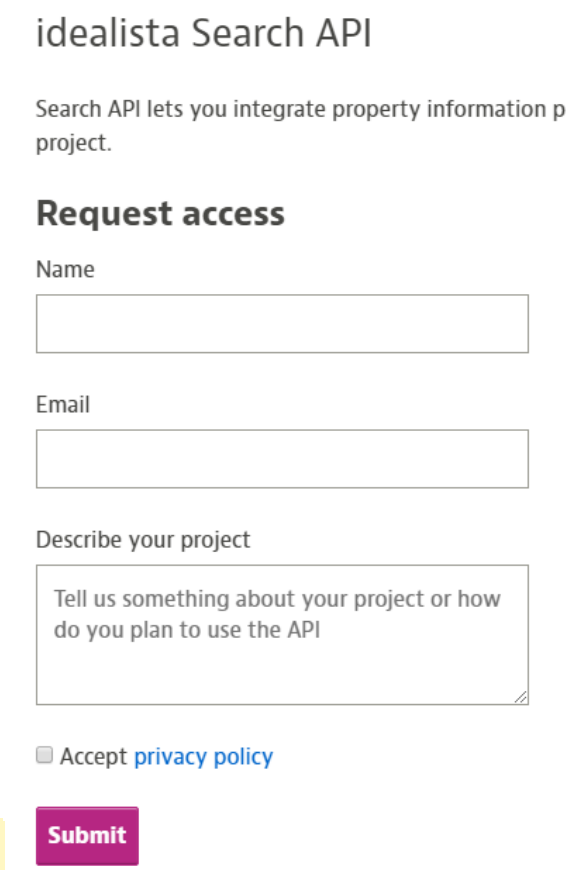
API Idealista

- Se extraen datos de inmuebles que se encuentran publicados en la web.
- Difícil de conectar, necesita una autenticación Oauth 2.0.
- No es fácil obtener un API key.
- Devuelve solamente 50 inmuebles en un radio.
- Documentación solo disponible luego de la solicitud del API key.
- Es la web de inmuebles más completa de España.
- Precios elevados.

KSCHOOL

¿Cómo obtener el API key?

- Rellena el siguiente formulario: http://developers.idealista.com/access-request
- Cruza los dedos para que te lo aprueben. 🤞
- Tendrás solo 150 solicitudes mensuales.



KSCHOOL

Let's do it...

Otras API's interesantes

- Airquality: http://aqicn.org/here/es/
- Twitter: https://developer.twitter.com/en/docs/api-reference-index.html
- Facebook: https://developers.facebook.com/docs/apis-and-sdks?locale=es_ES
- Spotify: https://developer.spotify.com/documentation/web-api/
- Youtube: https://developers.google.com/youtube/v3/

KSCHOOL

Let's do it...

KSCHOOL

Bases de datos y R, la combinación (casi) perfecta.

KSCHOOL

Intro: R y Bases de datos

- R tiene problemas con la manipulación de grandes datasets.
- Con R, los datos se almacenan en la memoria.
- SQL puede ser usado para análisis de datos con R.
- No es recomendable escribir datos desde R a cualquier base de datos.

KSCHOOL

Intro: ¿Qué es una base de datos relacional?

Columna

Tabla

	A	B	C	D	
1	**Last Name**	**Sales**	**Product Type**	**Company**	**Contact**
2	Smith	$1,675.00	EEE-312	Wok N Roll	Adams
3	Johnson	$1,480.00	DC-1	Wok N Roll	Rogers
4	Williams	$1,064.00	EE-2	Peace A Pizza	Evans
5	Jones	$1,390.00	DF-3	Kung Food	Webb
6	Brown	$4,865.00	EEE-45	Peace A Pizza	Fields
7	Williams	$1,243.00	FD-2	Kung Food	Mccoy
8	Johnson	$9,339.00	DC-1	Kung Food	Hansen

Fila

Conceptos R vs. SQL

R	SQL
NA	Database
dataframe	tabla
variable	columna
observación	fila
`subset(), data[cond], order`	`select, join, order by`

Tipos de conectividades

ODBC (Open Database Connectivity)

- RODBC (Open Database Connectivity usando R)

DBI-Based (Database Interface-Based)

- Database specific (Based on DBI)
  - RMySQL, RPostgreSQL, ROracle, RSQLite, etc…
- JDBC (Based on DBI)
  - RJDBC

El paquete RODBC

- El más popular
- Usa “Open Database Connetivity” (ODBC)
- Originalmente desarrollado por Microsoft
- Permite la conexión con muchísimas bases de datos
- Configuración requerida ODBC Driver Manager
- Contiene útiles funciones de R

Paquetes para DBI's

- Interfaz virtual
- Se basan en controladores DBI individuales para proporcionar métodos para interfaces definidas
- Contienen diferentes controladores basados en DBI para MySQL, PostgreSQL y Oracle.

El paquete RJDBC

- Utiliza JDBC drivers para proporcionar la conectividad
- Es necesario Java Runtime Environment

KSCHOOL

Let's do it...

Conclusiones

- Un Data Scientist es un profesional que tiene conocimientos de diversos softwares de análisis y visualización de datos, sin embargo, su fuerte siempre es la modelización de algoritmos de machine learning.
- La mejor y más completa base de datos del mundo es el internet, una técnica como el web scraping permite no solo extraer datos de manera automática sino también la más barata.
- Deben tenerse en cuenta para el web scraping los avisos de legales de cada web.
- A diferencia del web scraping, las API's son una fuente de acceso seguro y completamente autorizado a las bases de datos de las empresas a las que se consulten.
- Las API's son limitadas en cuanto a solicitudes, suelen ser de pago y el coste puede llegar a ser elevado.
- Como buenos Data Scientist que somos, es indispensable un conocimiento al menos medio de bases de datos, especialmente del tipo SQL.
- Tener conocimientos para la gestión de datos desde R hacia SQL es indispensable para el ahorro de tiempo, flujo de trabajo e inserción segura de datos, la base de todo el ciclo que comprende la ciencia de datos.

KSCHOOL

Fin...