Introducción © EDICIONES ROBLE, S.L.

campusformacion.imf.com © EDICIONES ROBLE, S.L.
ALEXIS GASTON VILLAGRA RAMIREZ

Indice

campusformacion.imf.com © EDICIONES ROBLE, S.L.
ALEXIS GASTON VILLAGRA RAMIREZ

I. Introducción

Introducción. Antonio Sarasa Cabezuelo

Las tecnologías Big Data habilitan la extracción de valor de los datos que generan las empresas. No obstante, ese valor solo puede obtenerse a través de la aplicación de un conjunto de técnicas y métodos que permiten procesarlos, transformarlos, estudiarlos y, finalmente, construir modelos sobre los mismos. Todas estas tareas se recogen en la figura profesional del científico de datos *(data scientist).*

El conjunto de habilidades y técnicas se ha reflejado en diferentes "diagramas de Venn" que han proliferado en presentaciones, charlas y documentos de Internet en distintas versiones. Reproducimos aquí la versión original publicada por Drew Conway con licencia abierta CC-BY-NC.

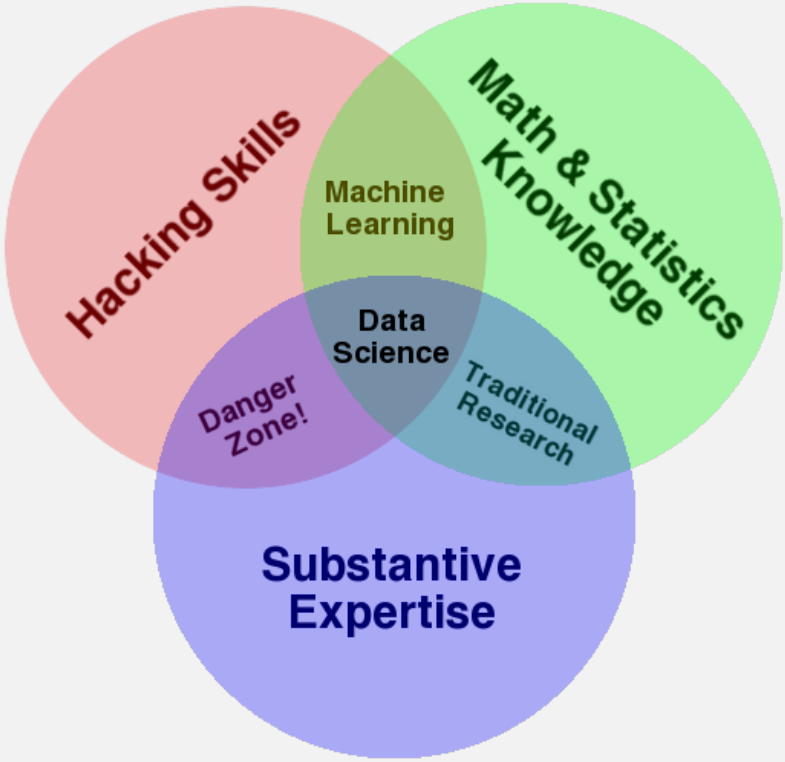

Imagen 1. Diagrama de Venn.

La idea del diagrama es la de mostrar la interdisciplinariedad inherente al trabajo del *data scientist.* Los tres grupos de habilidades pueden definirse sucintamente como sigue:

Conocimiento de matemática y estadística

La interpretación de los datos requiere conocimientos estadísticos y de modelos de aprendizaje estadístico, así como entender las distribuciones de los datos y las relaciones entre ellos. Estos conocimientos, combinados con conocimientos algorítmicos, dan como intersección el aprendizaje automático *(machine learning).*

Conocimientos de tecnologías de la información *(hacking skills)*

Si bien no es necesario hacer una carrera de informática *(computer science),* sí es necesaria una habilidad y soltura con el uso de diferentes tecnologías, incluyendo las bases de datos, las tecnologías de Internet —como fuente de datos— y la programación —no de manera profesional pero sí suficiente para la adquisición, transformación y tratamiento de datos—. Los especialistas en tecnología que también tienen conocimientos de un área del negocio carecen de los fundamentos estadísticos y matemáticos para obtener soluciones, corriendo por tanto el riesgo de llegar a soluciones simplistas o mal justificadas (de ahí el *Danger Zone!*).

Conocimientos específicos de áreas de negocio o dominios particulares

La aplicación de las técnicas analíticas sin un conocimiento del dominio no permite valorar los hallazgos, formular las preguntas correctas o evaluar los resultados de forma significativa para el negocio. Tradicionalmente se han utilizado métodos de investigación en muchas de esas áreas, pero el *data scientist* lo combina con automatización en el tratamiento de datos y técnicas algorítmicas como el *machine learning.*

En este módulo, se estudian los fundamentos mínimos necesarios en el conjunto de habilidades tecnológicas que se recogen en el diagrama de Venn anterior.

Unidad 1

El módulo comienza con el concepto de máquina virtual, cómo pueden crearse y para qué sirven. Esto tiene el propósito práctico de que el alumno sepa utilizarlas durante su proceso de aprendizaje y también pueda hacer uso de ellas profesionalmente, dado que la externalización de la computación en la nube, en muchos casos, se basa en tecnología de máquinas virtuales. En esta primera unidad se introducen también los fundamentos del uso de la Shell de comandos en Linux y de la creación de scripts, dado que en la configuración, instalación y despliegue de soluciones de Big Data es muy habitual trabajar con la línea de comandos y se requiere una cierta familiaridad con ella.

Unidad 2

La segunda unidad proporciona los conocimientos básicos del lenguaje de programación Python, orientados a su ampliación posterior en el tratamiento de datos. Python es uno de los lenguajes más populares utilizados por los *data scientists* y tiene además la ventaja de que se puede utilizar para cualquier otro propósito, ya que cuenta con un conjunto de bibliotecas muy amplio. El lenguaje R, especializado en el tratamiento estadístico, se estudia en el segundo módulo.

Unidad 3

La tercera unidad trata de la tecnología de bases de datos relacional, que es la base de muchos de los sistemas de información actuales. Conocer las bases de datos relacionales es fundamental por dos motivos. El primero, porque el *data scientist* tiene que ser capaz de obtener datos de esas bases de datos utilizando el lenguaje de consulta SQL. Por otro lado, porque las nuevas tecnologías de base de datos, que se denominan habitualmente NoSQL, también soportan en muchos casos el lenguaje de consulta SQL, aunque internamente no sean bases de datos relacionales. También sucede esto con otros sistemas, como por ejemplo en Apache Hive, que permite el uso de SQL para ejecutar procesamiento de datos paralelo sobre almacenamiento distribuido que utilice Apache Hadoop.

Unidad 4

La cuarta unidad introduce las tecnologías de Internet y sus principales lenguajes. Esto es importante para la adquisición de datos de la web en muchas aplicaciones: es necesario un conocimiento y habilidad para tratar con estos datos, al menos, en dos tipos de fuentes. La primera son los datos que muchos sitios web como Facebook, Twitter o StackExchange proporcionan a través de interfaces de programación de aplicaciones o APIs —usualmente siguiendo la arquitectura REST y ofreciendo datos en formatos como JSON o XML—. La segunda es la recogida de páginas web directamente, que requiere comprender y saber aplicar herramientas de crawling —recoger conjuntos de páginas siguiendo enlaces— y de scraping —procesar el contenido de esas páginas HTML para extraer la información relevante—.

Unidad 5

La quinta unidad trata de preparar al estudiante en la forma de compartir ficheros, código o datos en repositorios compartidos (abiertos o corporativos). Concretamente, se introduce el control de versiones y las herramientas basadas en Git, que es el sistema utilizado por repositorios on-line populares, como GitHub o BitBucket.

Unidad 6

Finalmente, la sexta unidad introduce los conceptos básicos de las bibliotecas fundamentales del stack científico de Python: NumPy, Pandas y Matplotlib. Estos bloques básicos proporcionan las bases para después profundizar en otras bibliotecas de ese stack o para aprender por analogía otros lenguajes, como R, que se basan en estructuras de datos similares.

campusformacion.imf.com © EDICIONES ROBLE, S.L. ALEXIS GASTON VILLAGRA RAMIREZ

II. Objetivos

campusformacion... ALEXIS GASTO...

Los objetivos generales del módulo que los alumnos alcanzarán pueden resumirse en los siguientes:

- Comprender y saber utilizar tecnologías de virtualización para el prototipado, desarrollo y despliegue de sistemas.
- Adquirir competencias de fundamentos de programación especialmente orientadas al tratamiento de datos.
- Conocer y saber interpretar y diseñar bases de datos relacionales, así como ser capaz de utilizar el lenguaje de consulta SQL para el acceso a fuentes de datos relacionales o no.
- Comprender las tecnologías básicas de Internet y de la web, así como sus lenguajes fundamentales, para ser capaz de obtener datos de fuentes en Internet.
- Saber utilizar herramientas de colaboración en grupos de trabajo para compartir código, datos y otros recursos.
- Entender los fundamentos de la programación con estructuras vectoriales y matriciales, que son la base de los lenguajes que utilizan los *data scientists*.

... ROBLE...