“Data Science”: Añadiendo valor a los datos

MARÍA-EGLÉE PÉREZ

UPR RÍO PIEDRAS

¿Qué se conoce como ”Big Data”?

“Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

McKinsey Global Institute “Big data: The next frontier for innovation, competition,and productivity” Junio 2011

Las cuatro V’s de “Big Data”

Volumen

Velocidad

Variedad

Veracidad

“Even though we now can generate a lot of data, there’s no substitute for a good biological question.”

Dr Heather Allen (USDA)
EDAMAME 2017

¿Por qué recolectamos datos?

Para comprender procesos

Para predecir situaciones futuras

Para tomar decisiones

…..

Recolectamos datos para convertirlos en información que permita generar respuestas.

Ciencia de Datos (“Data Science”)

Si bien inicialmente la discusión se centró sobre el uso de grandes conjuntos de datos, se ha evolucionado hacia el concepto de Ciencia de Datos (”Data Science”)

Ciencia de Datos es la rama del conocimiento que permite aprender de los datos para obtener conocimiento y predicciones útiles (R. Irizarry, U. of Harvard)

Ciencia de Datos (“Data Science”)

Concepto aún en evolución.

Ligado a la disponibilidad de grandes conjuntos de datos (“Big Data”), pero no limitado por ella.

Agencias gubernamentales y empresas están dedicando grandes cantidades de dinero al establecimiento de equipos de trabajo y al entrenamiento de personal. (“Big Data to Knowledge BD2K”, NIH)

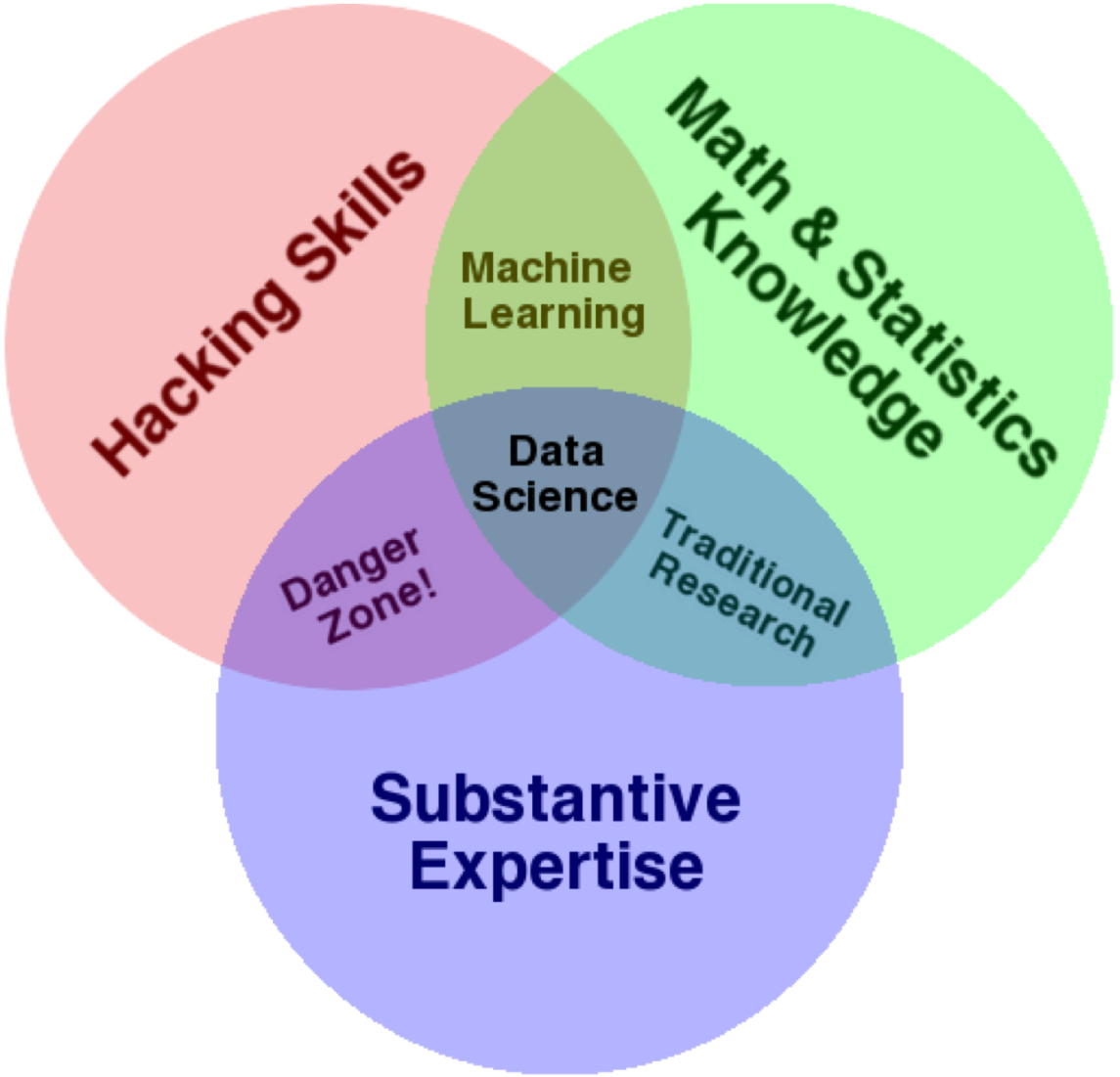
En 2011, el McKinsey Global Institute (MGI) advertía que en el futuro la competitividad de una empresa dependerá de su capacidad para generar valor a partir de los datos disponibles, y que la demanda de recursos humanos capaces de enfrentar estas tareas será superior a la oferta.

De acuerdo con una encuesta reportada en el boletín de enero 2017 de la American Statistical Association, durante los pasados 4 años

- 65% de las organizaciones incrementaron el número de posiciones requiriendo habilidades en análisis de datos.
- 59% esperan incrementarlos durante los próximos 5 años.
- 78% de las organizaciones buscando empleados con habilidades para el análisis de datos reportaron haber tenido problemas para encontrar candidatos calificados.

Diagrama de Venn de la Ciencia de Datos

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



Aspectos fundamentales de la investigación con datos

Selección y limpieza de un conjunto de datos para adecuarlos al objetivo del estudio,

Organización de los datos para obtener un acceso rápido y eficiente a grandes volúmenes de datos.

Análisis descriptivo de los datos para obtener intuiciones y establecer hipótesis iniciales

Obtención de predicciones basadas en métodos y modelos estadísticos

Comunicación de resultados a través de visualizaciones y resúmenes cuantitativos y cualitativos que sean interpretables y relevantes para el campo de aplicación

(R. Irizarry, Harvard)

Habilidades fundamentales para un científico de datos

Habilidades técnicas: Ciencias de cómputo

- Bases de datos, almacenamiento de grandes volúmenes de datos
- Procesamiento de grandes volúmenes de datos (Hadoop, Spark,…)
- Programación: SQL, Java, Python, R …

Habilidades fundamentales para un científico de datos

Habilidades técnicas: Matemática y Estadística

- Álgebra lineal: Muchos métodos de Machine Learning y Estadística Multivariada están basados en métodos de descomposición de matrices (QR, SVD)
- Análisis exploratorio de datos. Visualización

Habilidades fundamentales para un científico de datos

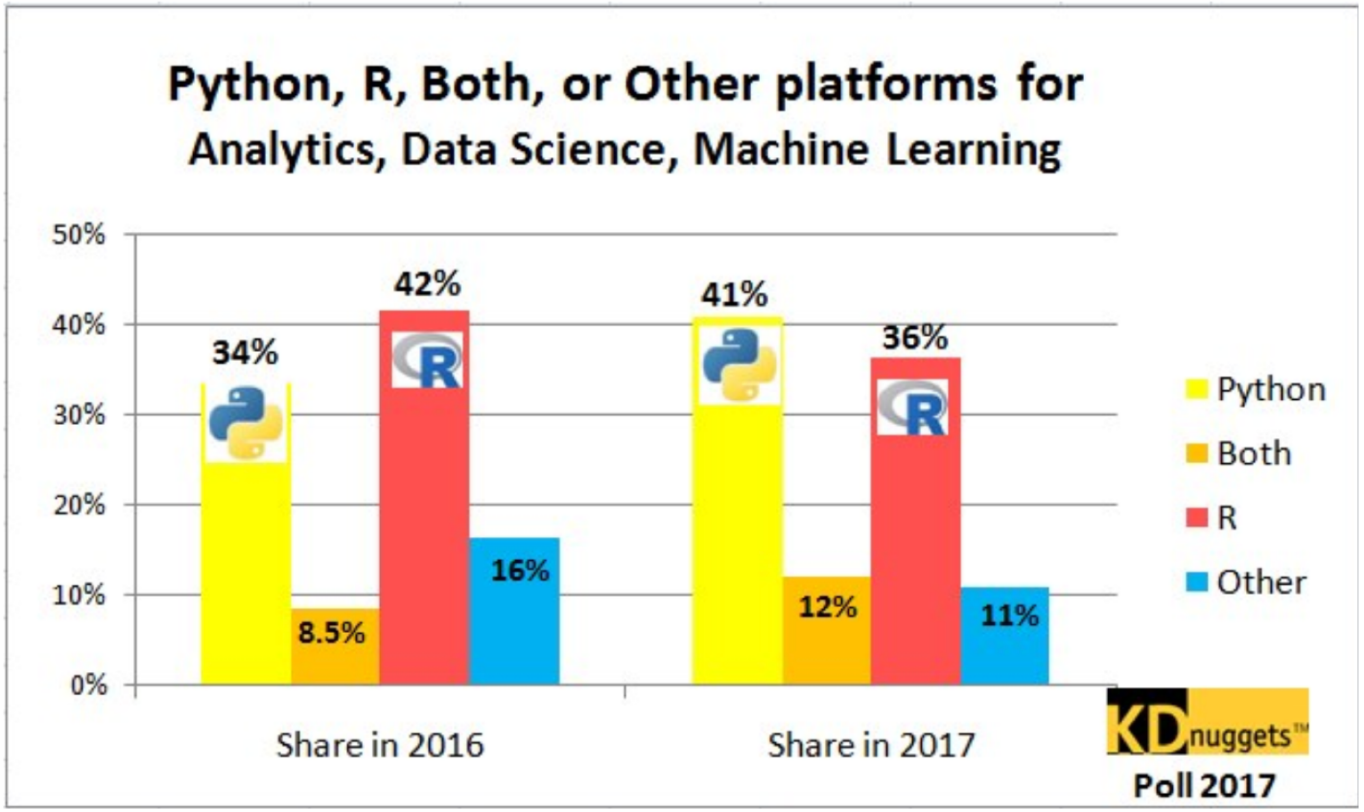
Habilidades técnicas: Matemática y Estadística (cont.)

- Inferencia estadística básica: probabilidades, estimación, pruebas de hipótesis
- Modelos estadísticos: Modelos de regresión lineal simple y múltiple. Modelo logístico. Modelo log-lineal. Métodos de predicción en series cronológicas. Estadística Bayesiana.

Habilidades fundamentales para un científico de datos

- **Habilidades técnicas: En la frontera**
  - Machine Learning: Algoritmos de clasificación (clustering jerárquico, K-medias, K vecinos más cercanos, ”random forest”), algoritmos de reducción de dimensionalidad (componentes principales, análisis de factores, escalamiento multidimensional,etc.)
  - Uso de software para análisis de datos.

R vs Python



https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

Habilidades fundamentales para un científico de datos

Habilidades no técnicas

- Curiosidad intelectual
- Capacidades de comunicación
  - Para contribuir a la formulación de preguntas
  - Para comunicar resultados en forma accesible a los usuarios.
- Conocimiento del área específica de aplicación.

Buscar al “unicornio” *vs* crear equipos multidisciplinarios.

“Data Science” es una actividad interdisciplinaria que requiere conocimientos en diversas áreas.

No es fácil encontrar personas con todas las habilidades requeridas.

En mi opinión, es necesario crear equipos multidisciplinarios y favorecer su integración.

Buscar al “unicornio” *vs* crear equipos multidisciplinarios.

La implantación de una gerencia basada en datos requiere un compromiso consciente por parte de la organización y en particular de sus líderes.

Ejemplos

- Sistemas de recomendación
- Selección de anuncios
- Traductores
- Correctores de escritura
- Predicción de resultados electorales
- Estudio de redes sociales
- Diagnóstico médico
- Medicina personalizada
- “Sabermetrics”
- Detección de fraude

Retos del trabajo con grandes volúmenes de datos

Estar siempre consciente de la existencia de incertidumbre.

Los resultados de los análisis estadísticos son válidos en una ventana de tiempo y espacio definida.

¿Son nuestras conclusiones realmente objetivas?

Tener *más* datos no siempre significa tener *mejores* datos

¿Es ético usar cierto tipo de datos?

Efecto de las desigualdades sobre los datos (¡y de los datos sobre las desigualdades!)

¿Por dónde empezar?

Cursos en línea

Existen numerosos recursos en línea, muchos de ellos gratuitos.

- Serie de cursos “Data Analysis for Life Sciences” R. Irizarry en edX.com
- *Data Science (Professional Certificate Program)* R Irizarry en edX.com.
- *Especialización en Ciencia de Datos*, John Hopkins (Coursera)
- *Especialización en Ciencia de Datos Aplicada con Python*, University of Michigan (Coursera)
- Otros cursos en Coursera, edX, Datacamp, Udacity, Udemy, etc.

¿Por dónde empezar?

Programas graduados en Ciencia de Datos (o áreas afines)

http://www.mastersindatascience.org

Consideraciones finales

La disponibilidad de grandes volúmenes de datos genera grandes posibilidades en Ciencia, Gerencia y otras muchas áreas.

Los datos por sí mismos no generan conocimiento. El conocimiento es generado al hacer las preguntas correctas y analizar los datos relevantes para llegar a conclusiones.

El reto:

Añadir valor a los datos existentes