Clase 1: Introducción a la ciencia de datos.

Diplomado de Ciencia de Datos
Notas por José Daniel Conejeros - jdconejeros@uc.cl

Junio del 2020

I. Consideraciones generales

El siguiente documento es un registro de los principales temas abordados en las sesiones del Diplomado de Ciencia de Datos de la UC. Las notas e información de este documento no buscan ser exhaustivos, sino que presentar de manera acotada los temas tratados clase a clase. El diplomando, y las notas, van a estar organizadas de la siguiente manera:

Tema 1: Fundamento de la ciencia de datos

- Aspectos principales de la ciencia de datos.
- Metodologías para proyectos.
- Organizaciones y aspectos comerciales.
- Consideraciones normativas y éticas al operar con datos.

Tema 2: Herramientas estadística para la ciencia de datos.

- Estadística descriptiva en R.
- Forescasting, modelos predictivos, etc.

Tema 3: Machine Learning

- Aprendizaje supervisado/no supervisado.
- Herramientas computacionales para desarrollar.

Tema 4: Visualización de datos y aplicaciones

- Visualización.
- Aplicaciones al Marketing y las Finanzas.
- Aplicaciones al análisis de texto.

Cada uno de los temas será resumido en estos documentos incluído las imágenes ilustrativas y ejemplos prácticos tanto de análisis estadístico y programación. A continuación se procede con las notas sobre los principales aspectos de la ciencia de datos.

II. Introducción a la ciencia de datos

Profesor: Alexis Alvear - aalvear@mat.uc.cl

El Big Data implica el procesamiento y manejo de grandes volúmenes de datos (Eric Schmidt). Lo cual no es sencillo, pues se requieren de herramientas y habilidades para lograr esto, pero la pregunta general es:

a. ¿Por qué creamos y generamos tantos datos?

Creamos información (datos) para dar explicación de los fenómenos que ocurren en el mundo con el fin de lograr nuestra supervivencia. Así, por ejemplo, vivir en un país sísmico requiere de mayor información (mayor data) para tomar mejores decisiones. Hay una orientación a buscar información sobre fenómenos extraños de los cuales no tenemos mayores antecedentes y que generan un alto nivel de incertidumbre. En ese sentido:

Fenómenos del mundo

Fenómeno en el mundo →Incertidumbre →Riesgo

La construcción de un desarrollo científico ayuda a dar explicación a los fenómenos que ocurren y, a partir de eso, reducir la incertidumbre a través de las certezas científicas. No obstante, dicho desarrollo siempre va a estar enmarcado en un paradigma, en otras palabras, el conjunto de valores y creencias compartidos por una comunidad y que definen un marco de conducta. En este caso, dicho paradigma tiene su origen práctico en el método científico:

Figura 1: Método Científico

En sí la ciencia para Popper no descubre realidades, sino que descubre falsedades y esto nos permite obtener certezas sobre un mundo que es complejo y cambiante. Siguiendo esa línea, Thomas Khun dirá que estos paradigmas se irán sustituyendo por nuevos, lo cual describe como revoluciones científicas. En ese sentido, es de pensar que el paradigma actual intenta responder a los desafíos e interrogantes de su época.

Paradigma científico

Método Ciéntico →Certezas →(-)Intertidumbre →(-)Riesgo

Paradigma sistémico-complejo

Con el objetivo de tratar la realidad de forma más adecuada se necesita un nuevo tipo de pensamiento que sea **complejo, sistémico, holístico, multidimensional, ecológico y global**. En otras palabras, que pueda entregar una compresión integral, holística y multidimensional del contexto, las interconexiones, las estructuras y los procesos. En definitiva, comprender la dinámica del todo es el *desafío de la complejidad*.

Sistemas complejos

- La información que contiene
- La información que produce
- La información que se interelaciona

El **modelamiento matemático** abre la posibilidad para comprender y trabajar con la complejidad. Esto se reduce a escribir de forma simple un problema complejo, en otras palabras, simplificar el todo. Nos permite simular un determinado sistema (comportamiento de las interacciones y relaciones) y predecir su comportamiento. El desarrollo científico y matemático nos ha permitido comprender de mejor forma el mundo complejo en el que vivimos (múltiples variables influyen en el comportamiento de los fenómenos). Una de las maneras en que podemos comprender el mundo es a través del registro del conocimiento a través del **dato**.

b. ¿Dónde registramos y guardamos tantos datos?

El **dato** es una representación simbólica que es generada por una acción (información cualitativa y cuantitativa). Algunos ejemplos:

- Primer censo en chile: Censo 1813 con el objetivo de contabilizar la cantidad de estudiantes que había en el país.
- Registro de nacimientos por parte del registro civil.
- Años 80: se masifica la digitalización
- 2000: registros digitalizados, se masifica el excel y antes el lottus123.

Actualmente estamos en la **era del big-data** en que todo se almacena en una nube y podemos acceder a ella desde cualquier parte del mundo. La materialización de esta nueva era ha sido gracias a un proceso de construcción científico - tecnológica.

- Alan Turing: formalizó el concepto de los algoritmos a través de la máquina de turing (1936).
- Norbert Wiener: fundador de la cibernética. Lenguaje y técnicas que permite abordar el problema del control y la comunicación en general.
- Claude Shannon: Matemático e ingeniero eléctrico, padre de la teoría de la información y creador del bit (1948) como unidad de medida.

Base del Big Data

Ciencia ←→Teconología

El big data es un fenómeno científico tecnológico que gracias a la integración de estas dos disciplinas nos permiten **transformar la complejidad en simplicidad con ayuda de las tecnologías de la información**. Nos permite mirar los fenómenos desde una perspectiva más integral. En definitiva nos permite **simplificar complejidad**:

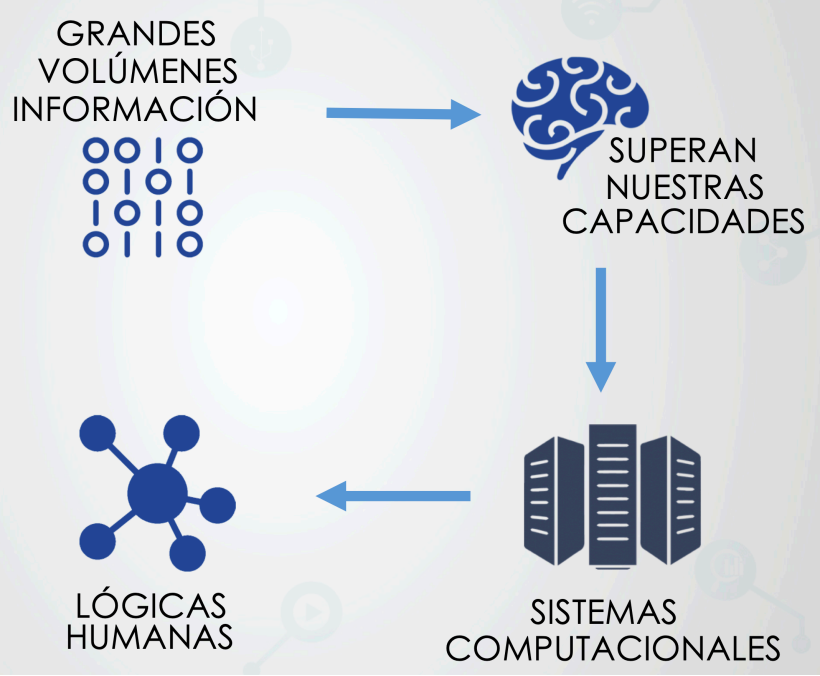

Figura 2: Proceso de Big Data

Sin embargo, no es solo un tema de cantidad, la complejidad está representada por:

1. Volumen: el desarrollo tecnológico nos permite capturar, procesar y analizar miles de datos minuto a minuto. El desafío es desarrollar capacidades técnicas para el procesamiento y análisis de datos masivos.

2. Velocidad: al trabajar con datos masivos es necesario contar con capacidades robustas que hagan frente a la volatilidad de los datos. Esto dado que muchos de los datos tienen una corta vida útil y es necesario capturarlos y analizarlos en el modelo oportuno para que no pierdan valor.

- **Batch:** procesamiento de datos es programado para determinados periodos una vez recolectados y almacenados en una cantidad suficiente de información acumulada. Se requiere de tiempo para procesar los datos.

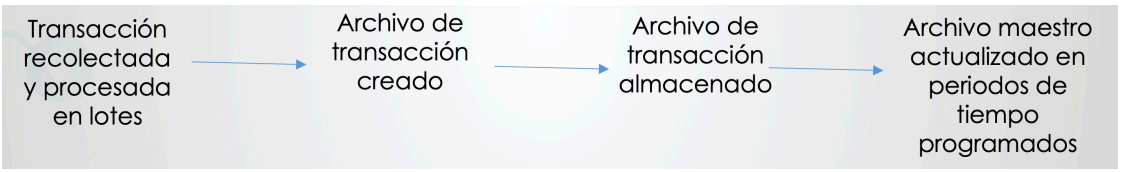

Figura 3: Proceso Batch

- **Streaming:** El procesamiento de datos es inmediato una vez ocurrida la transacción. La base de datos se actualiza una vez ocurrido el evento (proceso repetitivo e inmediato)

Figura 4: Proceso Streaming

3. Variedad: hay que considerar los tipos de datos, sean estos estructurados, semiestructurados o no estructurados. El desafío es procesar datos que provienen de distintas fuentes y formatos.

- **Estructurados:** aquellos que tienen un modelo definido y que provienen de un campo determinado o regitro (fichas de clientes).
- **Semiestructurados:** cuenta sin una estructura fija, pero contienen atributos o etiquetas (correos).
- **No estructurado:** datos que no tienen un modelo predefinido o no tiene una forma específica (videos, audios, fotografías).

4. Veracidad: Resguardar la calidad de los datos. Al trabajar con grandes volumenes de datos se debe considerar registros erroneos, incompletos, discrepantes o faltantes.

- **Datos erróneos:** campos o atributos mal consignados por problemas de lectura o tipeo.
- **Datos faltantes:** información incompleta de dimensiones considerables que puede impactar en los resultados esperados.
- **Fuentes discrepantes:** info proveniente de más de una fuente de información que presenta antecedentes previos.

5. Valor: el valor está dado por la capacidad analítica. Obtengo mayor valor a medida que los análisis se complejizan, en ese sentido existe una linealidad positiva entre valor y dificultad de análisis:

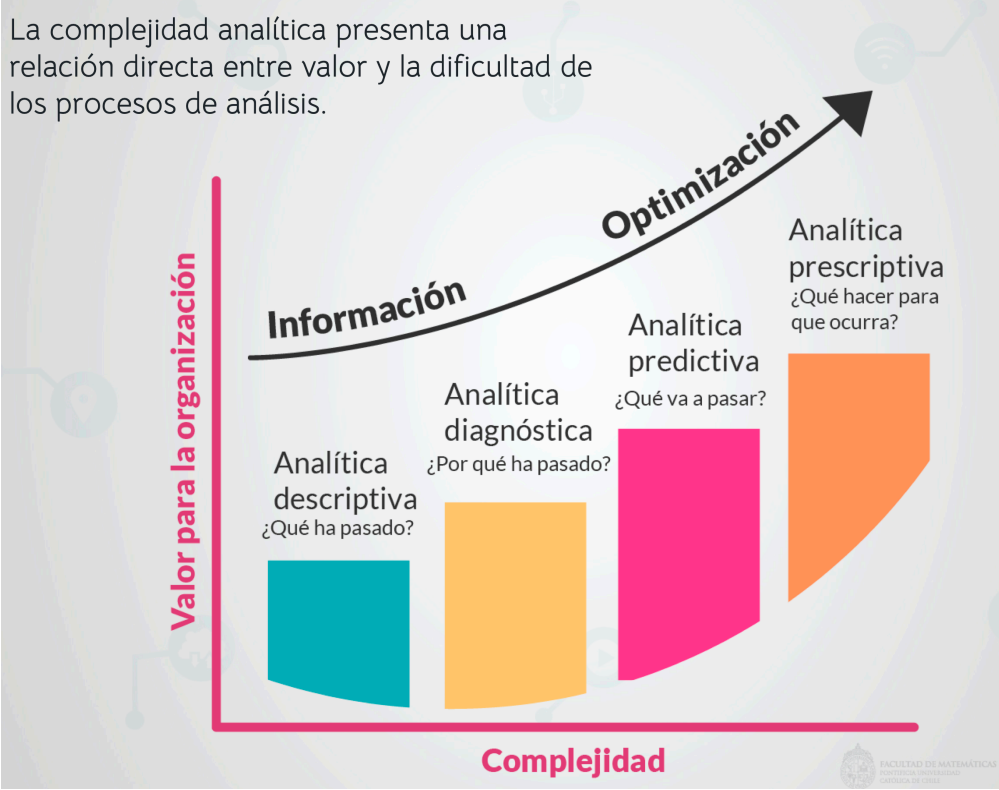

Figura 5: Capacidad Analítica

Estos pasos se sustentan en la capacidad analítica con los distintos modelos técnicos y estadísticos. Lo predictivo no implica lo prescriptivo necesariamente, en otras palabras, *correlación no implica causalidad*. En este sentido, la ciencia de datos genera valor.

c. La ciencia de datos

La ciencia de datos es una disciplina que se encarga de transformar datos en información para generar conocimiento que tenga valor y relevancia a la hora de tomar decisiones. El cientista de datos es un profesional que desarrolla procesos de análisis de alta complejidad mediante el diseño y desarrollo de algoritmos que permitan encontrar hallazgos relevantes en la información, interpretar resultados y obtener conclusiones relevantes.

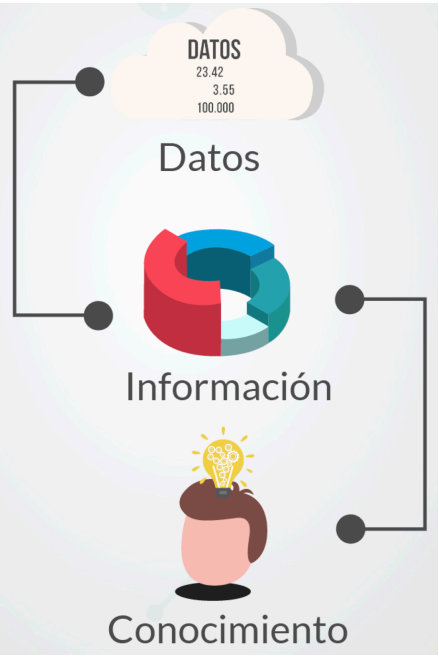

Figura 6: Ciencia de Datos

Elementos fundamentales para el desarrollo de la ciencia de datos:

- **Modelamiento matemáticos y estadístico:** caja de herramientas para diseñar e implementar procesos de análisis a distintos niveles de complejidad. Esto le permite reconocer patrones de comportamiento en los datos, estimar probabilidades de ocurrencia de diferentes fenómenos, realizar pronósticos, simular escenarios futuros, etc.
- **Programación computacional:** manejar un lenguaje de programación nos dota de capacidades para automatizar procesos de análisis. Además puede ser capaz de diseñar interfaces de aprendizaje automático entre humanos y computadoras, estableciendo las bases para soluciones de inteligencia artificial. `R` / `Python` son lenguajes para sistematizar procesos (códigos que un computador comprende y ejecuta).

Machine Learning

Disciplina que permite a las computadores aprender de forma autónoma (sin intervención humana). En otras palabras, podemos programar a las computadoras para que aprendan a través de imitación, observación o instrucción con algoritmos que identifiquen de forma iterativa patrones y tendencias en un conjunto de datos masivos.
Algoritmos de aprendizaje:

- **Supervisados:** Descubrir la relación existente entre una variable de entrada y otra de salida (el resultado buscado). Aquí hay métodos de clasificación como árboles de decisión, random forest, naive bayes, support vector machine y redes neuronales artificiales.
- **No supervisados:** Generan conocimiento a partir de los datos de entrada sin necesidad de supervisar un resultado esperado. Aquí podemos encontrar el clustering, análisis de componentes principales, k-means y reglas de asociación.

- **Visualización de datos:** la capacidad de contar historia con los datos se mezcla con criterios para seleccionar técnicas de visualización y gráficos adecuados para quienes consuman la información e interpreten adecuadamente las conclusiones obtenidas. Podemos usar herramientas como `tableau`, `Power Bi`. `Shiny`, `Rmarkdown`.
- **Comunicación de datos:** podemos utilizar enfoques estructurados sobre cómo comunicar nuestros hallazgos relevantes a partir de los datos e involucra tres elementos (datos, visualización y narrativa) esto se conoce como **Storytelling**. Por otro lado, los **dashboard** o paneles de datos consisten en la consolidación de la representación gráfica de lo indicadores claves y está orientado a la toma de decisiones estratégicas (transforma los datos en información relevante).

Documento elaborado con las herramientas de Rmarkdown