

Fundamentos de Ciencias de Datos

Ciencia de Datos ¿Una moda?



© CanStockPhoto.com - csp45445583

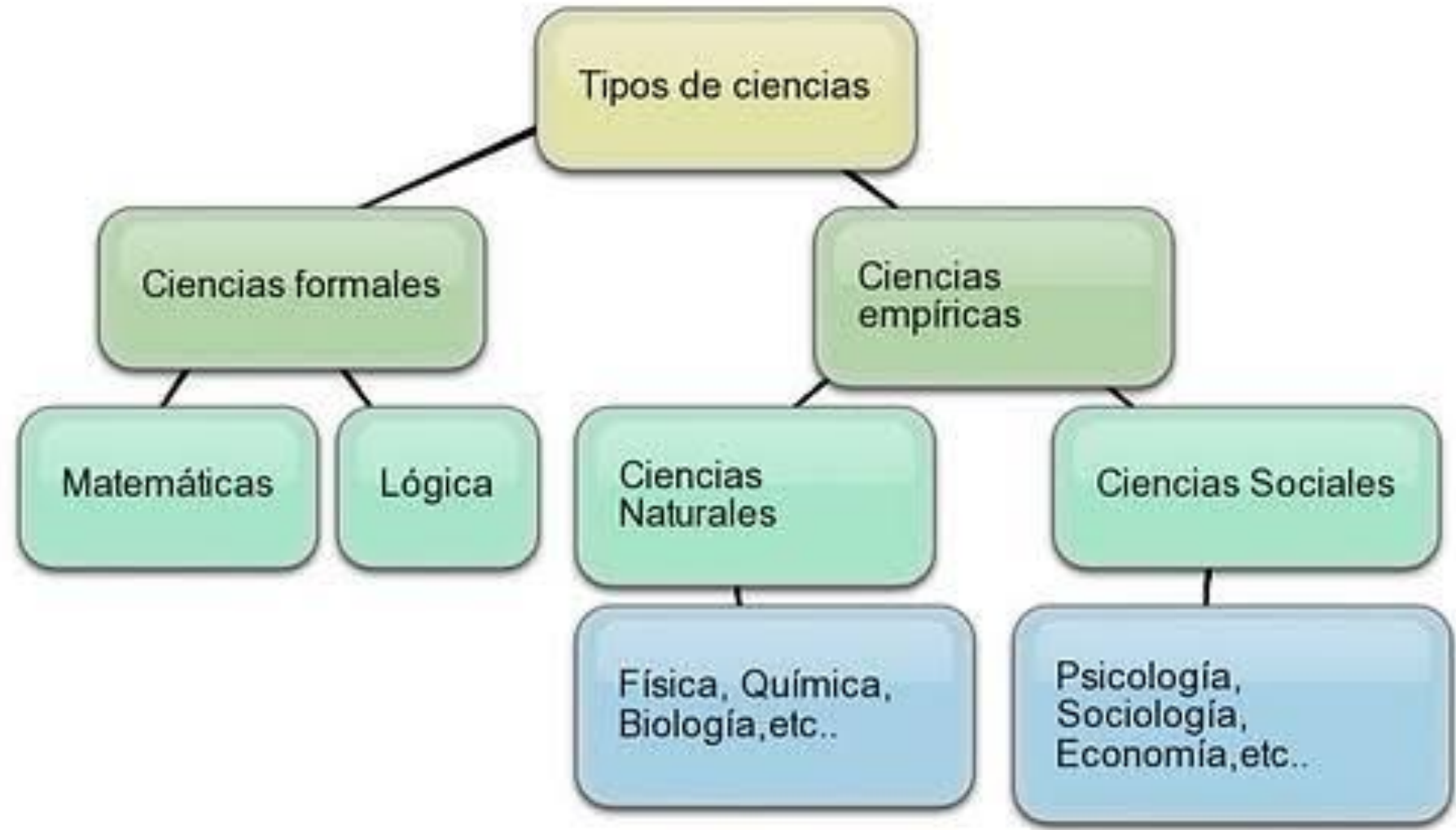
¿Qué es ciencia de Datos?

- **Es el estudio de los datos** que pueden provenir de diferentes fuentes, tales como:
 - teléfonos celulares, expedientes de hospitales, registros satelitales, mercados financieros, redes sociales y otros.
- Los datos pueden ser:
 - **estructurados**: de bases de datos y sistemas de almacenamiento de datos debidamente formalizado, incluso hojas de cálculo
 - **no-estructurados**: textos libres como noticias periodísticas, tuits, forma de audio, video, imágenes digitalizadas
- Se pretende obtener con el estudio, su entendimiento.
- Para el estudio, se apoya en:
 - la estadística, la ciencia de la computación, la minería de datos, el aprendizaje automático y el análisis predictivo.

Lectura recomendada: “A very short history of Data Science”.

<https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#1bb5557555cf>

¿Qué tipo de ciencia es?



¡ Discusión !

¿Ciencia transversal?

¿Por qué es importante la ciencia de datos?

- **Al entender los datos** se pueden identificar situaciones problemáticas o de interés particular, en diferentes rangos de aplicación.
- Los problemas pueden ser **pasados, presentes o futuros**, es decir:
 - ¿Que paso, Que esta pasando o Que va a pasar?
- Las **aplicaciones** pueden ser en:
 - Mercadotecnia, Gobierno, Economía Administración, Medicina, Comportamiento humano, Finanzas,....
- En el rango de la aplicación, la ciencia de datos **identifica problemas**.
- La ciencia de datos por sí misma **no genera soluciones**.

Lectura recomendada: “¿Qué diablos es Ciencia de Datos?”.

<https://medium.com/datos-y-ciencia/qu%C3%A9-diablos-es-ciencia-de-datos-f1c8c7add107>



¿Por qué es importante la ciencia de datos?

- La ciencia de datos ha tenido un crecimiento rápido en los últimos años
- Las industrias y áreas de estudio han encontrado un gran *potencial* en la Ciencia de Datos
- Principales factores que han influenciado el crecimiento son:
 - La reducción de los **costos** de la computación,
 - La facilidad de almacenamiento y procesamiento en la **nube**,
 - La **conectividad y accesibilidad** para llegar a millones de usuarios con productos e información
- La capacidad competitiva de una organización ahora se mide en la manera en que aplica el análisis de sus datos, ***para impulsar la innovación***



¿Por qué es importante la ciencia de datos?

Son cuatro elementos en los cuales centrarse, referido a las organizaciones:

- **Evaluar la salud de la organización** ... generar metas numéricas y/o métricas, monitorearla y saber como se ha desarrollado
- **Construir los productos o servicios adecuados** ... No cualquier producto o servicio que se fabrique, será exitoso para la organización. El proceso de fabricación, distribución y venta debe ser monitoreado con datos y el análisis de éstos, dirán que tan exitoso es el producto
- **Obtención de pronósticos y tendencias futuras** ...
 - que pasaría si ...?
 - Como se comportaría esto... ?
 - Hagamos un prototipo y...
- **Definir la ruta de la organización y su estrategia** ... la importancia de saber hacia donde dirigir las organización y como, es uno de los elementos mas importantes del análisis de datos



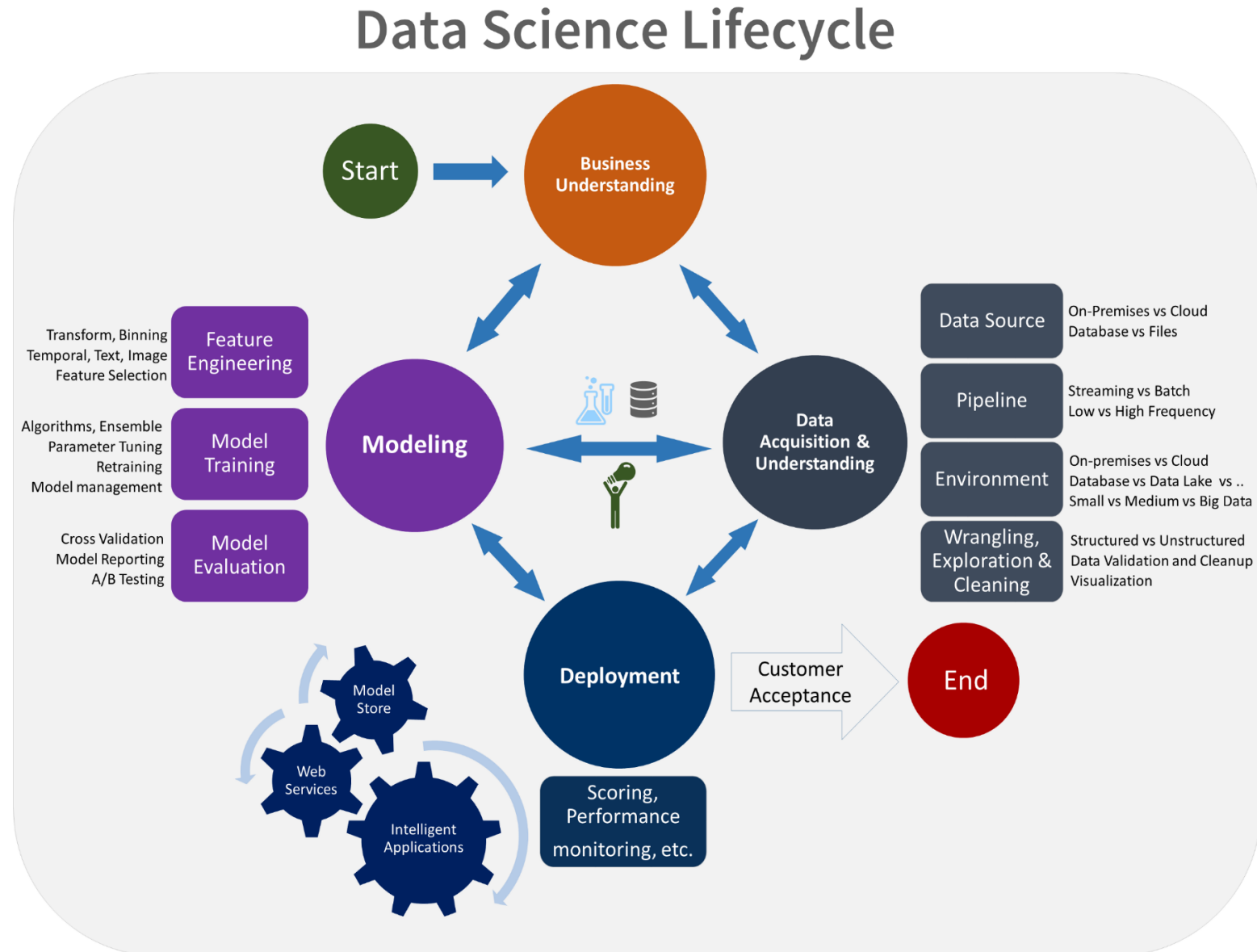
Ciclo de vida de un proyecto de Ciencia de Datos

- Hay muchas formas de representar el **ciclo de vida de un proyecto**, por lo que tomaremos la siguiente:

De..
<https://towardsdatascience.com/data-science-life-cycle-101-for-dummies-like-me-e66b47ad8d8f>

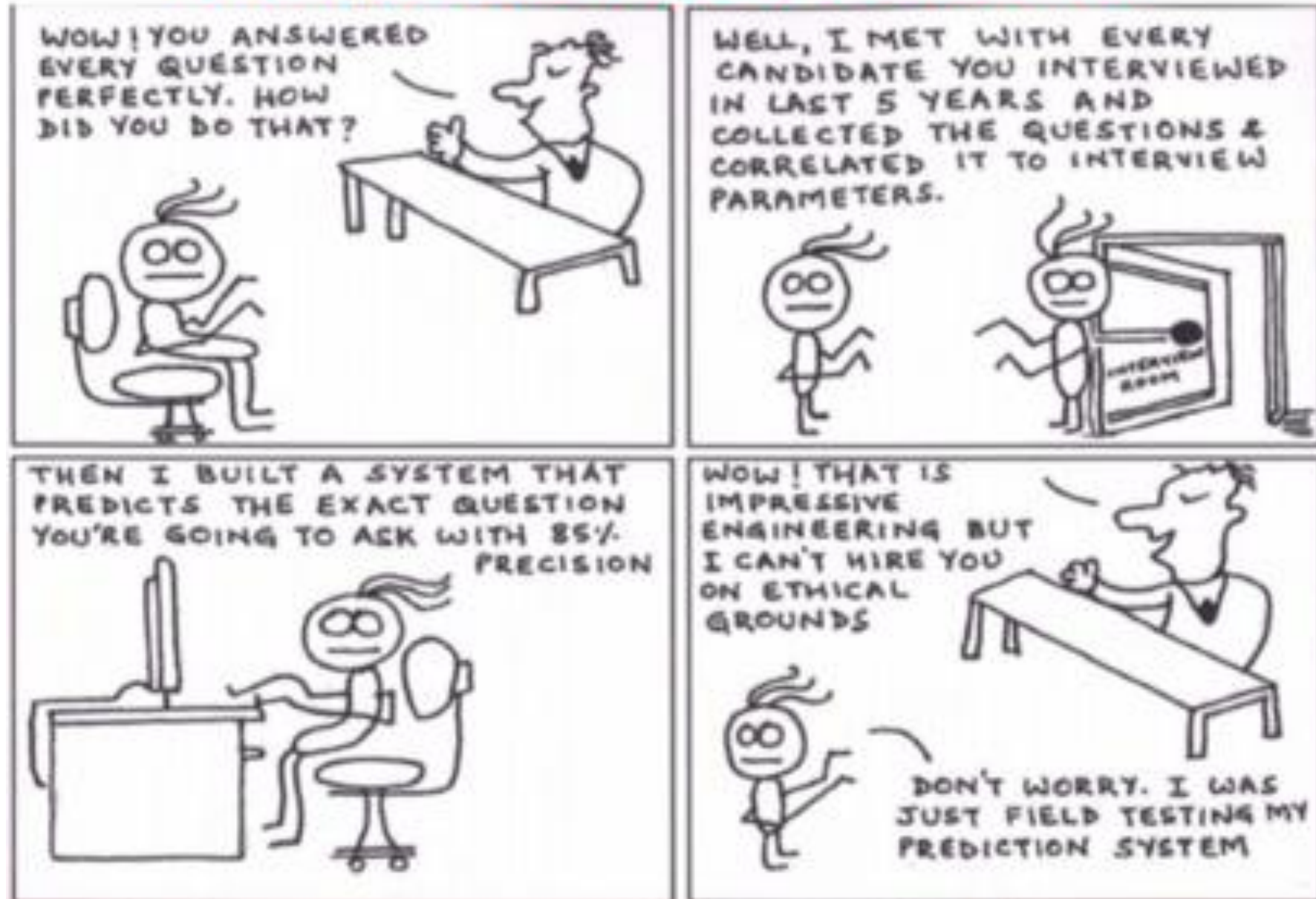
Esta vista se puede resumir en 7 pasos:

1. Entender el negocio
2. Extraer los datos
3. Limpiar los datos
4. Explorar los datos
5. Generar un modelo de características
6. Elaborar un modelo de predicción
7. Compartir/mostrar los resultados



Entender el negocio

When you interview a data scientist...



(Del Blog Microsoft Azure...)

- Hay cinco **preguntas** que se deben responder:
 - ¿cuánto o cuántos? (regresión)
 - ¿de que categoría? (clasificación)
 - ¿de que grupo? (clustering/agrupamiento)
 - ¿está raro? (detección de anomalías)
 - ¿que opción tomar? (recomendación)
- Estas preguntas nos ayudarán a definir el **objetivo** de nuestro proyecto de Ciencia de Datos

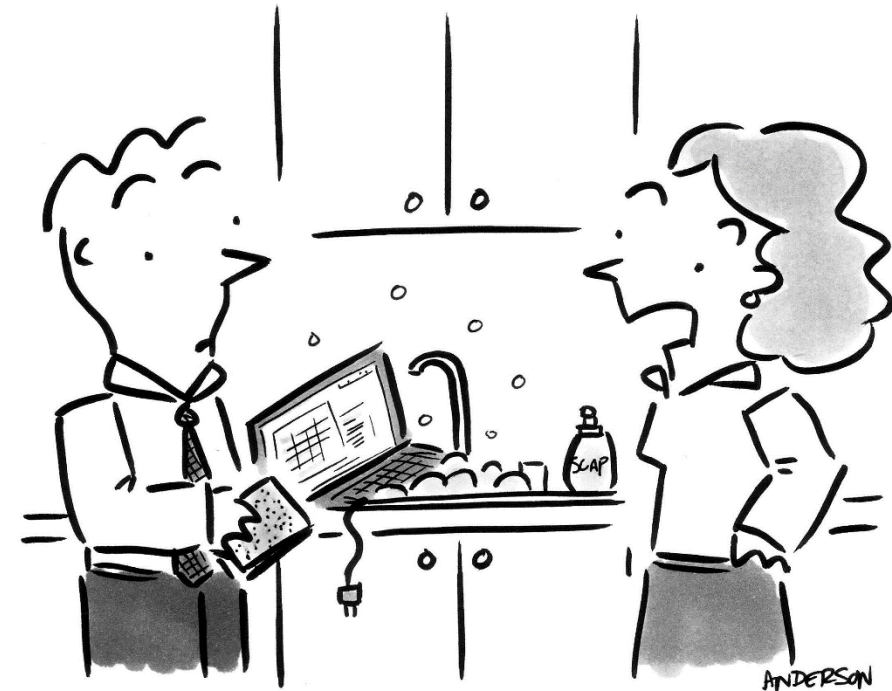
Extracción de datos

- Una vez que se ha definido el objetivo del proyecto, las preguntas son:
- ¿Qué datos necesito?
- ¿De donde saco los datos?
- ¿Cómo los saco?
- ¿Con que herramientas?
- ¿Qué me voy a encontrar?
- ¿Dónde o como los guardo?

Limpieza de los datos

Esta es una de las actividades que mas tiempo nos consumen

- Se buscan y arreglan inconsistencias:
- 1, 0 o True, False o Si, No
- Masculino o Femenino: M/F o H/M
- 31/12/2019 o 12/31/2019 o 2019/12/31
- Pesos o Dólares o Euros o Pesetas o Francos...



"This is not what I meant when I said 'we need better data cleansing!'"

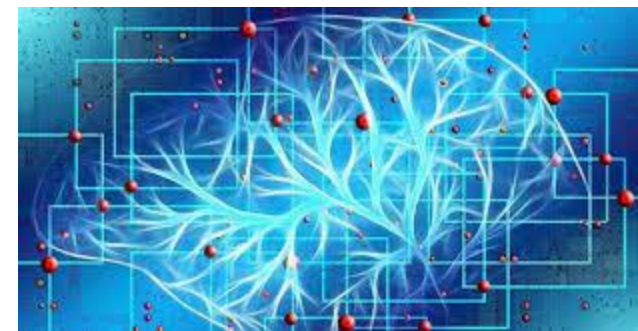
Exploración de datos

- Hay que conocer los datos, buscar algunos patrones, tendencias o sesgos, repeticiones, máximos y mínimos
- Es conveniente obtener algunas gráficas como histogramas o curvas de distribución, de tal forma que se pueda observa la tendencia general
- En este punto se puede uno cuestionar si esta muestra de datos va a ser útil para cumplir con los objetivos de nuestro proyecto



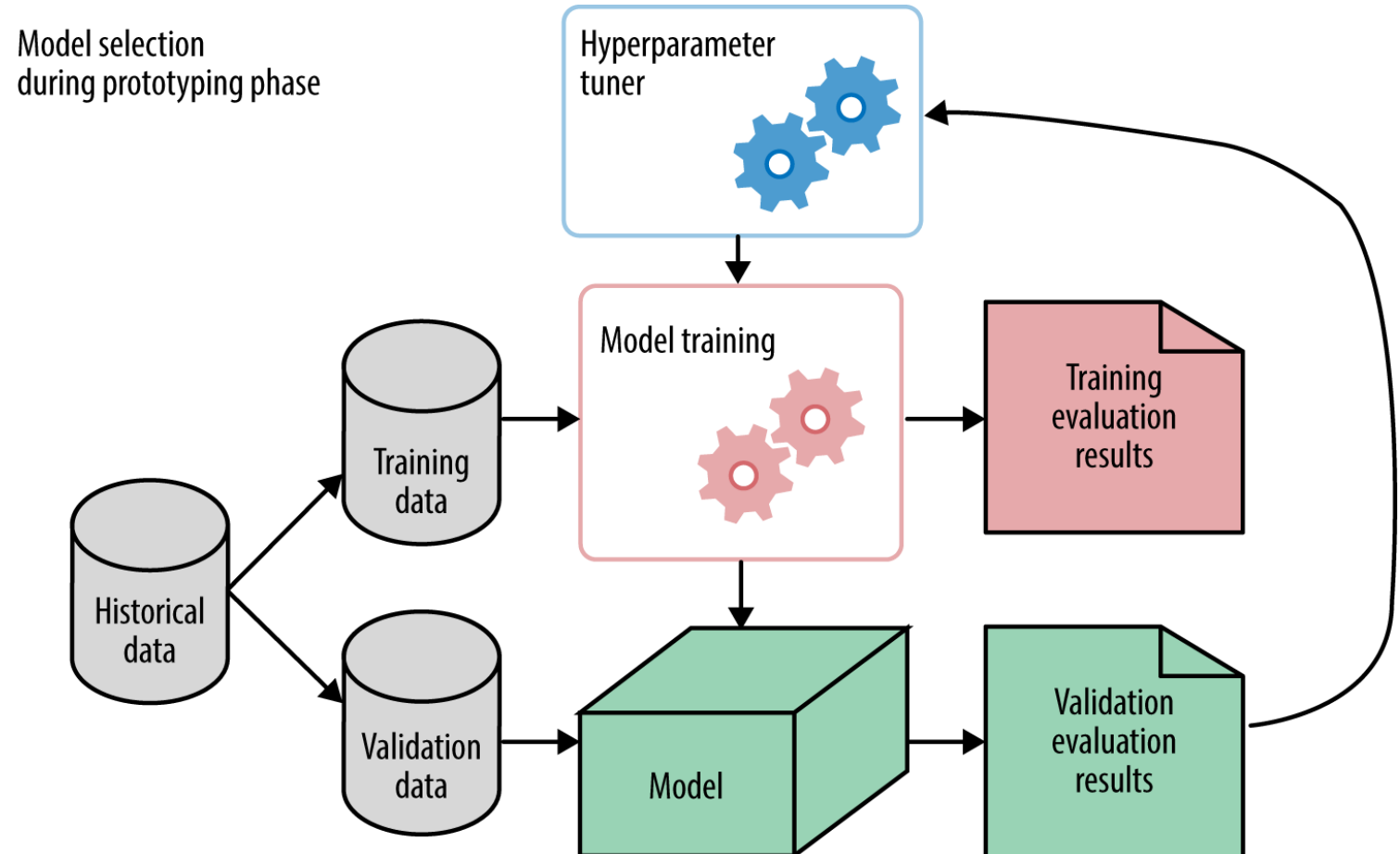
Modelo de características

- Además de los datos de interés principal para el objetivo del proyecto, podría haber otros datos de referencia que influyen en el fenómeno que se desea estudiar
- En este modelo podría generarse algo sobre *Machine Learning*
- Las características pueden ser seleccionadas o construidas
 - *Seleccionadas*, de la misma muestra de datos
 - *Construidas*, a partir de los datos de la muestra y ciertas formulas



Modelo de predicción

- Típicamente vamos a aplicar el modelo de *Machine Learning*
- Puede haber diferentes algoritmos de *Machine Learning*; deben ser evaluados
- .. De Microsoft Azure:



Mostrar los resultados

- La importancia de esta etapa es que hay muchas formas de expresar lo que se ha encontrado, pero cual es la óptima, dependiendo de las persona que lo van a ver y usar?
- Hay muchas herramientas para visualizar datos; hay que dominar las técnicas de éstas para estar listos ante cualquier cuestionamiento y generar nuevas formas de visualización

Lectura: “Así que Ciencia de Datos .. eh ? Juan Carlos Vázquez”

<https://medium.com/datos-y-ciencia/as%C3%AD-que-ciencia-de-datos-eh-e05b49549dc6>

