Organización de Datos

Martes teoria

Jueves práctica

100% virtual

**Método de aprobación**:

* 1 tp durante la cursada (sin reentrega)
* 1 tp durante la cursada (1 reentrega)
* 1 parcial (2 recuperatorios)
* 1 tp final (2 reentregas en cualquier fecha de coloquio). Defensa oral.

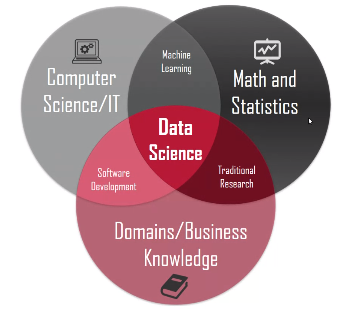
[pagina de la materia](https://datos7506-fiuba.github.io/)

Clases Teóricas

# Clase 1 Intro a Data Science (23/08)

**Data Scientist**: Alguien capaz de obtener, interpretar, procesar y filtrar los datos. Llega a conclusiones en base a esto y construye soluciones.

**!= Data engineer**: Solapamiento entre habilidades y responsabilidades, cambia el enfoque. Este se enfoca en la creación de infraestructura y arquitectura para la generación, soporte y extracción de los datos.



**Habilidades técnicas**:

* Data mining: Extracción/minería de datos. Convierte el conjunto de datos en información encontrando patrones en la misma.

Ejemplo: Pañales y cerveza.

* Visualización de datos
* Matematica y estadistica
* Machine Learning: Redes neuronales, árboles de decisión, métodos de ensamble. Permite que la pc realice tareas complejas (ej: detectar personas en una foto).

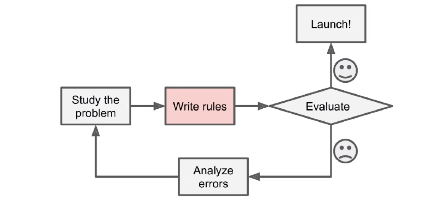
**Machine Learning**: Aprendizaje automático. Es la ciencia y el arte de programar computadoras para que aprendan a partir de datos.

Se dice que un programa de computadora se aprende de la experiencia respecto de una tarea y una medida de rendimiento, si su rendimiento medido mejora la experiencia.

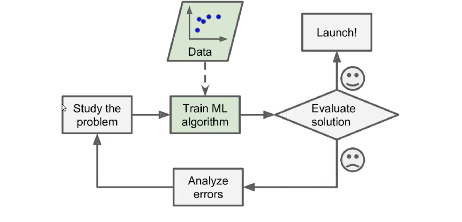
Es el campo de estudio que da a las pcs la capacidad de aprender sin ser programada de manera explícita.

**Ejemplo: Filtro de SPAM**

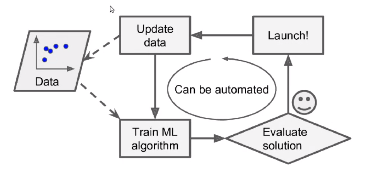
* Enfoque tradicional



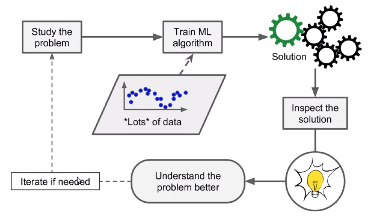
* Enfoque ML



* Adaptación automática al cambio



* Pueden ayudarnos a aprender



**Áreas de aplicación**

*Medicina*: Test de coronavirus a través de la voz - Generación de nuevas drogas - Pre-diagnósticos en base a historial clinico - Epidemiologia

*Gaming*: Monetización - Aumentar el engagement - Mejorar la calidad visual - Hacer jugadores virtuales de distintas dificultades.

*Energía*: Minimizar riesgos - Costos e inversiones óptimos - Optimización y eficiencia de abastecimiento - Maximizar extracción/generación.

*Turismo*: Perfiles de viajantes - Optimización de rutas.

*Seguridad*: Análisis de imágenes.

**Metodología**



**Visualización de Datos**

*Por qué es necesario graficar?* Comunicar de forma clara y concisa, entender para qué sirve cada tipo de gráfico, entender de forma eficiente los datos, encontrar patrones/relaciones.

En los histogramas, el eje x tiene que ser continuo, no discreto.

**Números Útiles**

*Media*: Promedio

*Mediana*: Valor que está en la mitad de la población

*Cuatril*: Valores límite que dejan al 25% de la población entre ellos (se puede con nunpy, ver diapos).

*Rango intercuartílico*: El rango entre el cuartil 1 y el cuartil 3.

El **box plot** ordena todos los valores que tenemos en una recta numérica. Dibuja donde caen estos valores.

Más plots: <https://www.python-graph-gallery.com/>

Clases Prácticas

# Clase 1 Intro a Data Science (25/08)