UAX Universidad Alfonso X el Sabio

CIENCIA DE DATOS

GRADO EN BIOMEDICINA

DR. ARIEL CARIAGA-MARTÍNEZ

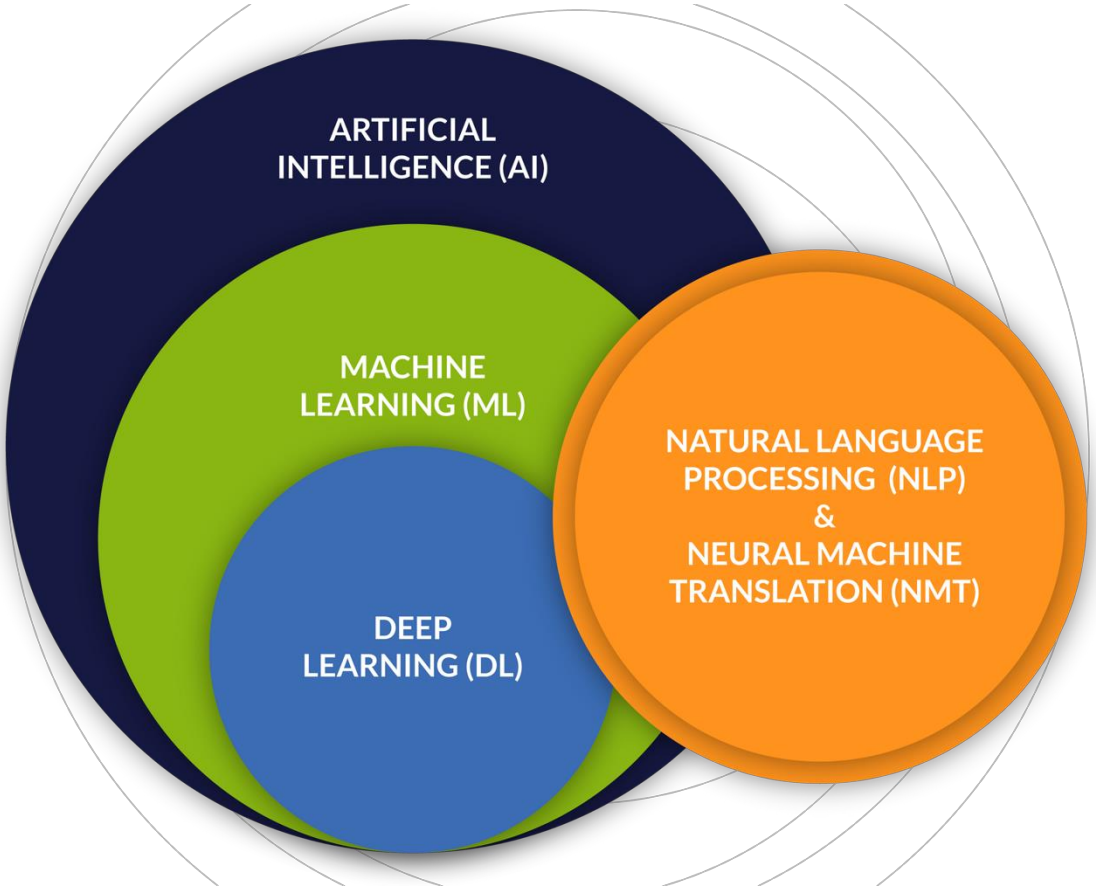
CIENCIA DE DATOS
Recordatorio

UAX

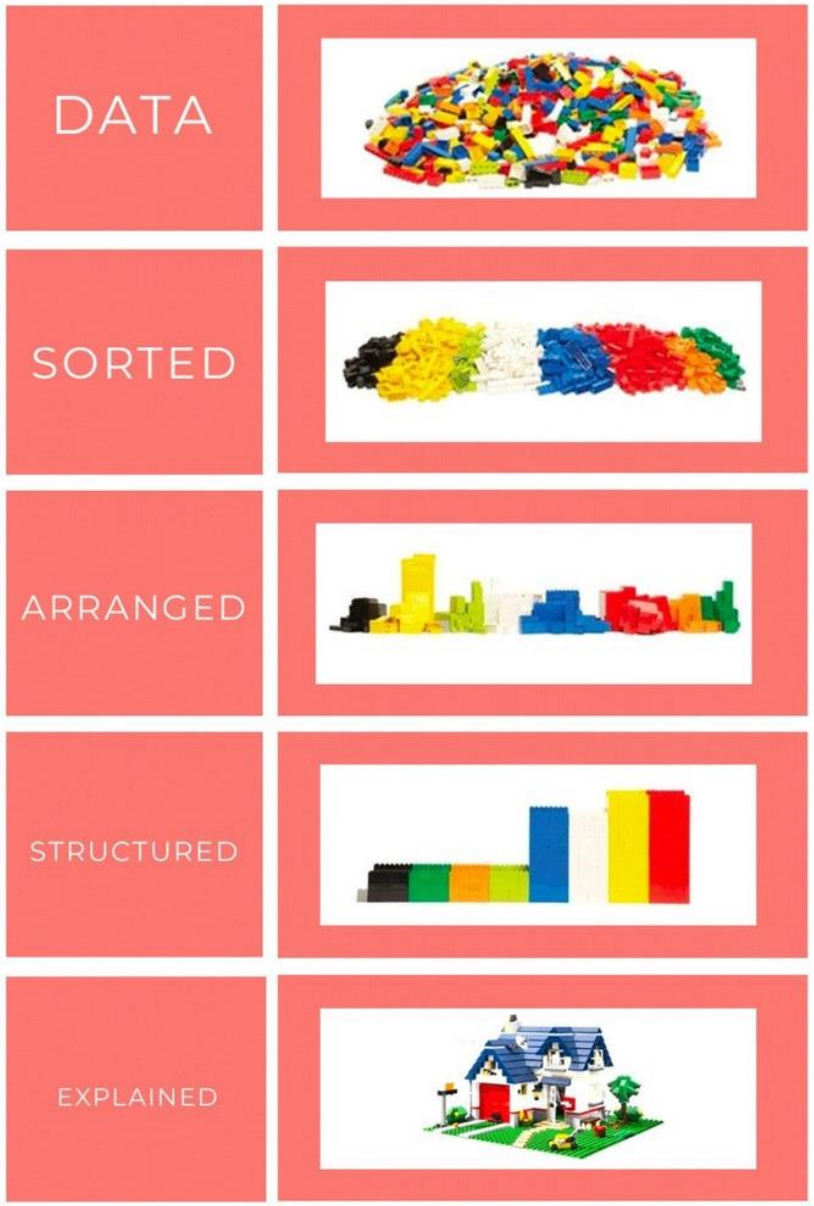
Inteligencia artificial = campo científico multidisciplinar completo.

ML = algoritmos de "aprendizaje" (sin ser programados para ello específicamente).

DL = (parte del ML) funciones matemáticas que colaboran entre sí para generar un "output".



PUNTOS PARA VERIFICAR



¿De qué forma recogemos los datos?

¿Cómo los podríamos organizar/ordenar?

¿Cómo los podríamos transformar?

¿Cómo los podríamos modelar?

¿QUÉ HISTORIA (CONVINCENTE) PODEMOS CONTAR?

Conceptos para repasar…

Probabilidad → Variables aleatorias (distribuciones, etc.).

Estadística descriptiva → Análisis exploratorio de los datos (EDA)

Resumir, entender, detectar patrones, anomalías, suposiciones…

Las fases de un proyecto de Ciencia de datos

1. Definición del Problema
   - Identificación de la pregunta de investigación o el problema empresarial.
   - Establecimiento de objetivos claros y de métricas para medir el éxito.

2. Recopilación de Datos
   - Recolección de datos relevantes, que pueden provenir de bases de datos, APIs, archivos, u otros.
   - Consideración de la calidad, la disponibilidad y la relevancia de los datos.

3. Preparación de los Datos
   - Limpieza de datos: manejo de valores faltantes, duplicados y datos inconsistentes.
   - Transformación: normalización, codificación y escalado de datos.
   - Ingeniería de características para mejorar el rendimiento del modelo.

4. Análisis Exploratorio de Datos (EDA)
   - Análisis inicial para comprender distribuciones, correlaciones y patrones.
   - Visualización de datos para detectar tendencias y relaciones clave.

5. Modelado
   - Selección de algoritmos y técnicas de modelado adecuados para el problema.
   - Entrenamiento y ajuste de modelos con los datos de entrenamiento.
   - Validación y selección del modelo óptimo basado en métricas de desempeño.

6. Evaluación del Modelo
   - Prueba del modelo en datos de prueba para evaluar su rendimiento.
   - Cálculo de métricas de error y ajuste para refinar el modelo, si es necesario.

7. Implementación
   - Despliegue del modelo en un entorno de producción para su uso.
   - Configuración de un sistema de monitoreo para supervisar el rendimiento.

8. Mantenimiento y Actualización
   - Monitoreo del modelo en producción para asegurar su precisión y relevancia.
   - Realización de actualizaciones periódicas o reentrenamiento del modelo.

UAX

Fase 3-4) EDA → ¿Para qué?

¿Qué?

- Estadística descriptiva → Mínimo
- Descripción de tipos de datos.
- Presencia de NAs / Evaluación de la normalidad → ¿Por qué?
- (Análisis univariante)
  - Presencia de outliers
  - Distribución aproximada de los datos → ¿Por qué?
- Análisis bivariante → ¿Por qué?

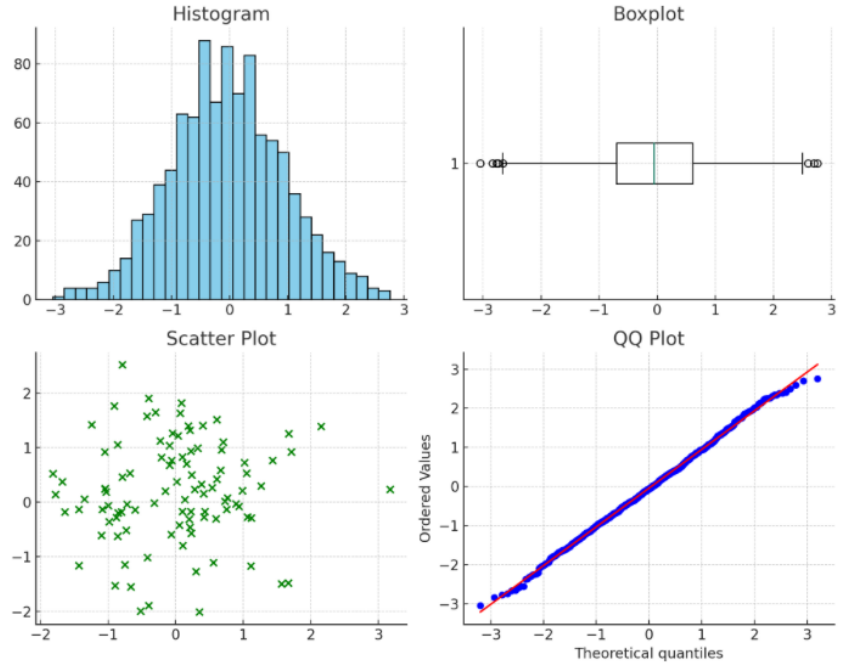
EDA

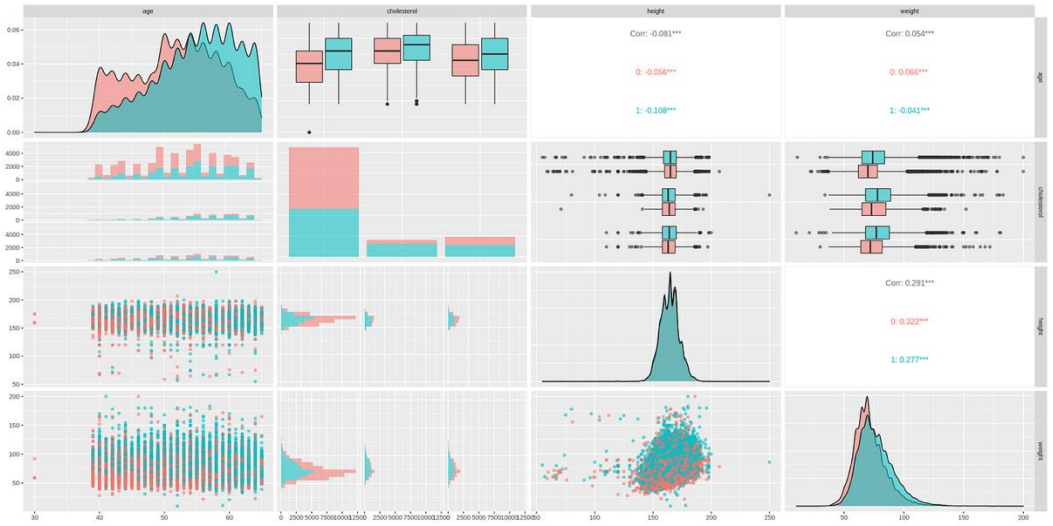
Intro-to-R.R | new_metadata

Filter

	genotype	celltype	replicate	samplemeans	age_in_days
sample1	Wt	typeA	1	10.266102	40
sample2	Wt	typeA	2	10.849759	32
sample3	Wt	typeA	3	9.452517	38
sample4	KO	typeA	1	15.833872	35
sample5	KO	typeA	2	15.590184	41
sample6	KO	typeA	3	15.551529	32
sample7	Wt	typeB	1	15.522219	34
sample8	Wt	typeB	2	13.808281	26
sample9	Wt	typeB	3	14.108399	28
sample10	KO	typeB	1	10.743292	28
sample11	KO	typeB	2	10.778318	30
sample12	KO	typeB	3	9.754733	32

Showing 1 to 12 of 12 entries, 5 total columns





Algunas preguntas respondidas tras un EDA adecuado.

¿Hay valores faltantes? ¿Cómo se distribuyen los datos?

¿Hay normalidad?
¿Hay alguna otra distribución de interés?
¿Hay correlación de variables?

¿Necesito normalizar/escalar los datos?
¿Necesito cambiar tipos de variables?
¿Qué puedo hacer con los outliers?
¿Qué puedo hacer con los NAs?

¿QUÉ QUIERO (PUEDO) MODELAR?

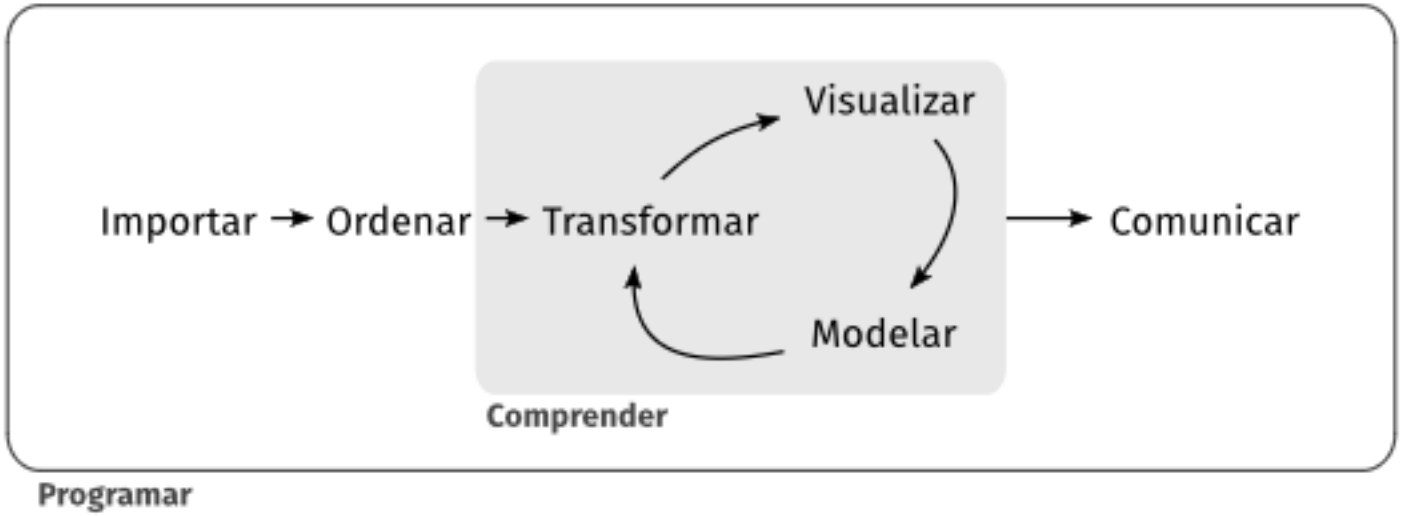
UAX

Siguiente paso: “Transformar/Visualizar/Modelar”

SQL

Tableu,
PowerBI…

Python, R…



Tu propio
conocimiento

https://es.r4ds.hadley.nz

OPTIMIZAR LAS MÉTRICAS + PRESENTAR

UAX

¿Cómo? (HOW?) → "Automágicamente"

ERRORES / ENTRENAR

ML: ¿Cómo "aprendizaje" sin ser programado para ello específicamente?

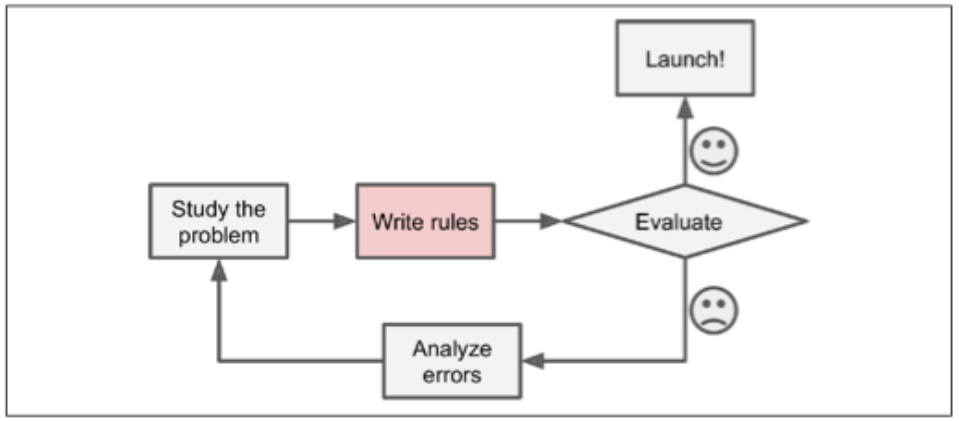


Figure 1-1. The traditional approach

Descomponer el problema + "escribir las reglas" (= proponer hipótesis + modelos estadísticos)

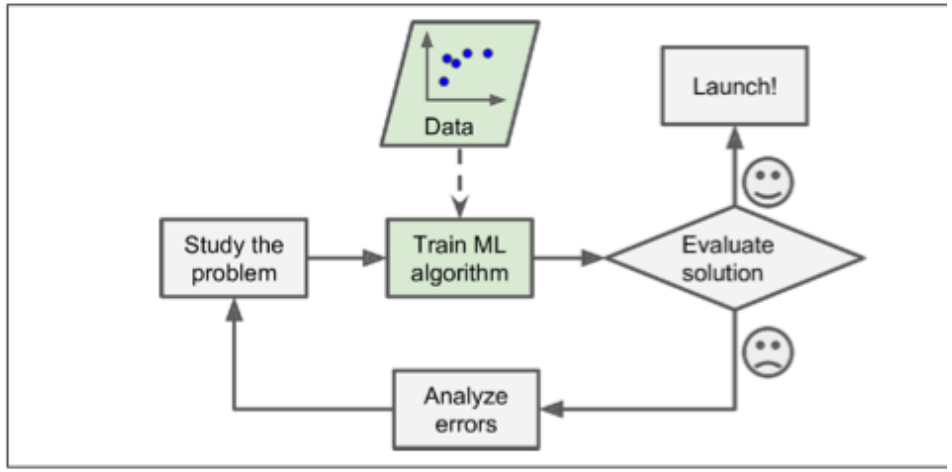
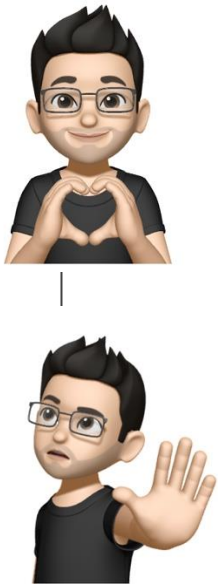


Figure 1-2. The Machine Learning approach



El algoritmo (función matemática/estadística + evaluación del "error") recibe datos y va ofreciendo mejores soluciones…

¡YA TENGO EL DATASET LISTO!

ESCOJO LA VARIABLE RESPUESTA → SELECCIONO LAS PREDICTORAS (INGENIERÍA DE VARIABLES) → VERIFICO LOS TIPOS → PROPONGO EL MODELO A UTILIZAR (SEGÚN EL TIPO DE VARIABLE → VISUALIZO RESULTADOS / MÉTRICAS DEL MODELO → REPITO

CADA MODELO TIENE MÉTRICAS DIFERENTES → SE PUEDEN OPTIMIZAR (FINE-TUNING) SEGÚN LA PREGUNTA

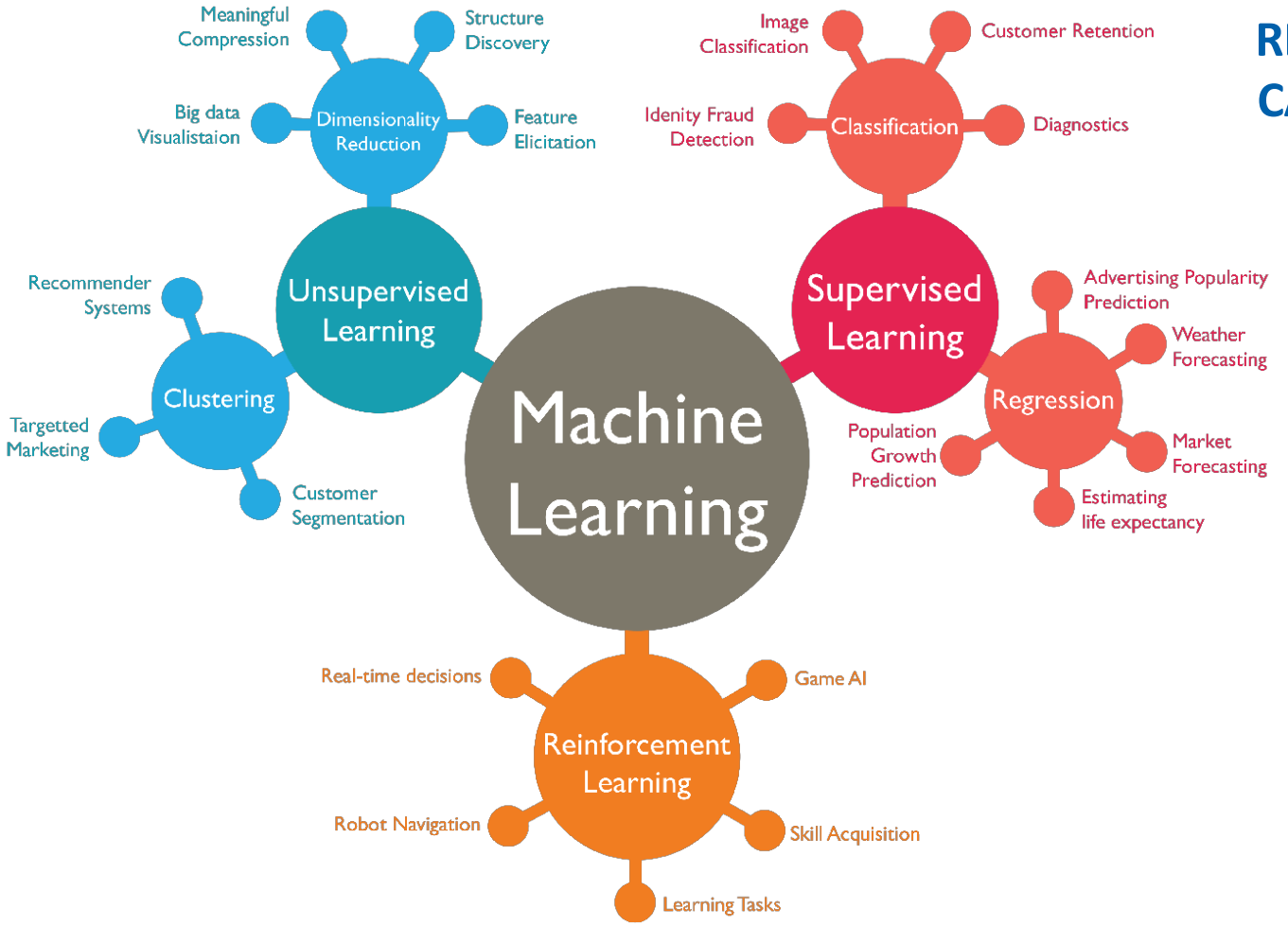
PRESENTA TUS DATOS → DATA STORY TELLING → (EN ANEXOS PUEDEN IR LOS SCRIPTS) SOLO IRÁN LOS RESULTADOS CON SUS MÉTRICAS Y BREVES PINCELADAS DEL MODELADO (WEBS DE REPOSITORIOS : GITHUB / GITLAB)

UAX

DATOS TABULARES

NO HAY VARIABLE RESPUESTA CONOCIDA → NUEVA PERSPECTIVA

VARIABLE RESPUESTA = CATEGÓRICA



VARIABLE RESPUESTA = CONTINUA

ML: Los datos NO necesariamente deben seguir ningún "patrón" (organización/"leyes")

UAX

GOBERNANZA DEL DATO.

- Políticas y procedimientos: no podemos hacer ”lo que queremos como queremos”. Documentación legal vinculante.
- Roles y responsabilidades: *data owners, data stewards y data users* (entre otros).
- Calidad del dato → POLÍTICAS
- Seguridad/privacidad, accesiblidad/disponibilidad.
- Trazabilidad y auditoría. → RESPONSABLES
- CONSIDERACIONES ÉTICAS.

EN DATA SCIENCE TODO POR ESCRITO

UAX Universidad Alfonso X el Sabio

GRACIAS

DR. ARIEL CARIAGA-MARTINEZ

CIENCIA DE DATOS

ACARIMAR@UAX.ES