

AMAT- Introducción a Ciencia de Datos y Machine Learning

Karina Lizette Gamboa Puente Oscar Arturo Bringas López

Contents

Chapter 1

BIENVENIDA

1.1 Objetivo

La primera parte del curso tiene como finalidad que el alumno tenga un entendimiento general de conceptos, técnicas, algoritmos y del proceso de desarrollo de proyectos de Ciencia de Datos. Entenderá la diferencia entre Big Data, Machine Learning, Business Intelligence y Ciencia de Datos. Todo lo anterior será cumplido mientras el alumno aprende las paqueterías y funciones más novedosas que se usan en R para Ciencia de Datos y las tecnologías que dan soporte a este software.

Se asume que el alumno tiene conocimientos generales de estadística, bases matemáticas y de programación básica en R.

1.2 ¿Quienes somos?

ACT. ARTURO BRINGAS

LinkedIn: arturo-bringas **Email:** act.arturo.b@ciencias.unam.mx

Actuario, egresado de la Facultad de Ciencias y Maestría en Ciencia de Datos, ITAM. Experiencia en modelos predictivos y de clasificación de machine learning aplicado a seguros, deportes y movilidad internacional. Es jefe de departamento en Investigación Aplicada y Opinión de la UNAM, donde realiza estudios estadísticos de impacto social. Es consultor para empresas y organizaciones como GNP, El Universal, UNAM, Simmia, la Organización de las Naciones Unidas Contra la Droga y el Delito (UNODC), entre otros. Actualmente es profesor de machine learning y programación en R en AMAT y se desempeña como consultor independiente en diferentes proyectos contribuyendo a empresas en temas de

machine learning, estadística, series de tiempo, visualización de datos y análisis geoespacial.

**ACT. KARINA LIZETTE GAMBOA**

LinkedIn: KaLizzyGam **Email:** lizzygamboa@ciencias.unam.mx

Actuaria, egresada de la Facultad de Ciencias, UNAM, candidata a Maestra en Ciencia de Datos por el ITAM.

Experiencia en áreas de analítica predictiva e inteligencia del negocio. Lead y Senior Data Scientist en consultoría en diferentes sectores como tecnología, asegurador, financiero y bancario. Experta en entendimiento de negocio para la correcta implementación de algoritmos de inteligencia y explotación de datos. Actualmente se desarrolla como Arquitecta de Soluciones Analíticas en Merama, startup mexicana clasificada como uno de los nuevos unicornios de Latinoamérica. Senior Data Science en CLOSTER y como profesora del diplomado de Metodología de la Investigación Social por la UNAM así como instructora de cursos de Ciencia de Datos en AMAT.

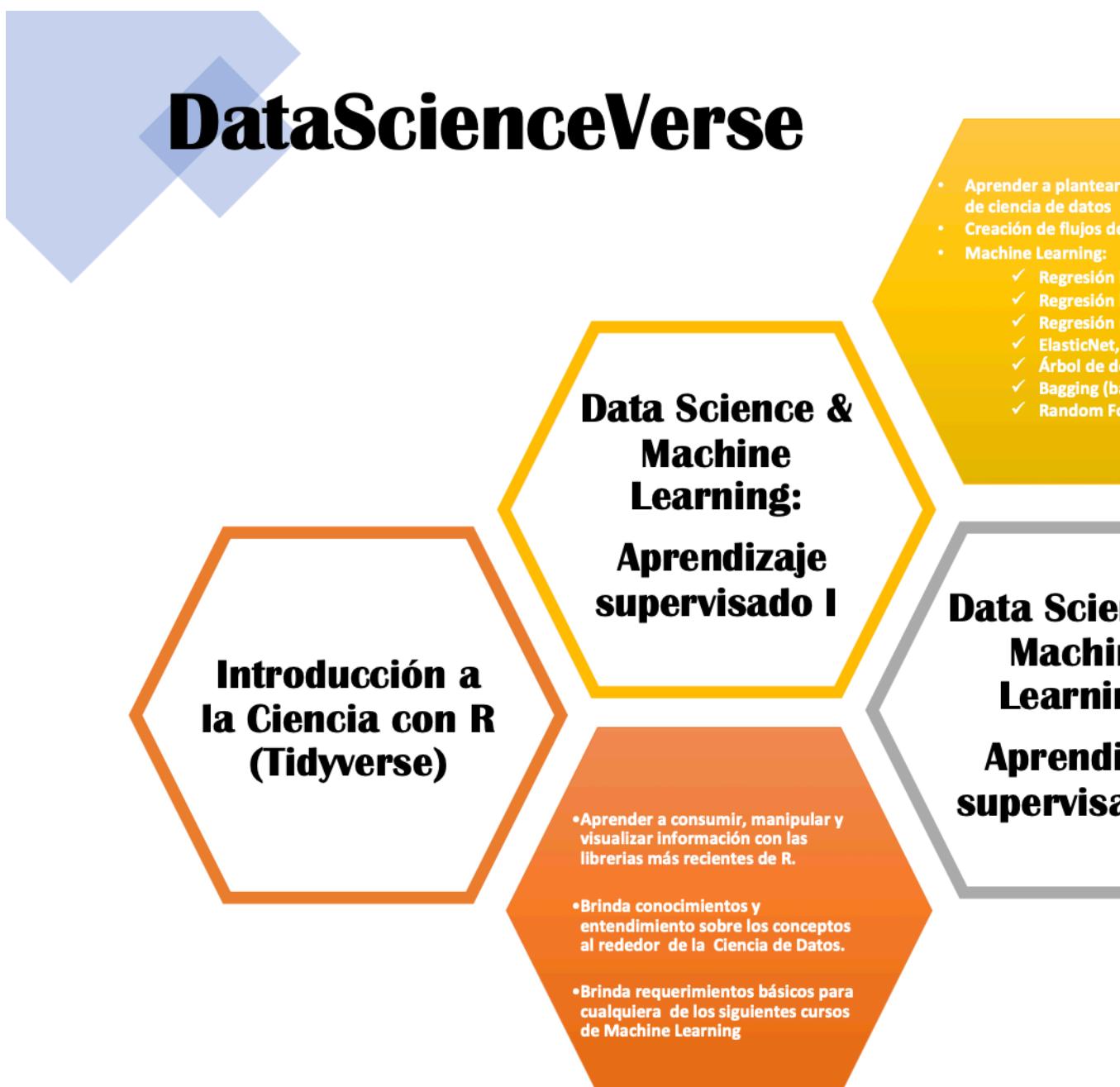
1.2. ¿QUIENES SOMOS?

7

Empresas anteriores: GNP, Activer Banco y Casa de Bolsa, PlayCity Casinos, RakenDataGroup Consulting, entre otros.



1.3 Ciencia de Datos en R



1.4 Estructura del curso actual

1.4.1 Alcances del curso

Al finalizar el módulo, el participante sabrá plantear un proyecto de ciencia de datos, desde sus requerimientos hasta sus alcances. Sabrá crear flujos de trabajo limpios y ordenados para crear poderosos modelos de Machine Learning. Este curso brindará las bases para introducirse al módulo intermedio de Ciencia de datos que se imparte en AMAT:

1. Data Science & Machine Learning (Aprendizaje Supervisado II)

Requisitos:

Computadora con al menos 4Gb Ram.

Instalación de R con al menos versión 4.1.0

Instalación de Rstudio con al menos versión 1.4

Kit básico para Ciencia de Datos con R (Tidyverse) ó

Dominio de las funciones de manipulación y visualización de datos con Tidyverse en R

Temario:

1.- Introducción a Ciencia de Datos (8 HRS)

- ¿Qué es Ciencia de Datos?
- Objetivo de la ciencia de datos
- ¿Qué se requiere para hacer ciencia de datos?
- Tipos de problemas que se pueden resolver
- Tipos de algoritmos y aprendizaje
- Ciclo de vida de un proyecto de Ciencia de Datos
- Taller de Scoping

2. Machine Learning: conceptos básicos (4 HRS)

- ML y algoritmos
- Análisis supervisado vs no supervisado
- Sesgo y varianza

- Pre-procesamiento e ingeniería de datos
- Partición de datos: test & train

3. Machine Learning: Modelos de aprendizaje supervisado (20 HRS)

- Regresión lineal
- Regresión logística
- Regresión lasso
- Regresión ridge
- ElasticNet
- KNN
- Árbol de decisión
- Bagging (básico)
- Random Forest
- Comparación de modelos

1.5 Duración y evaluación del curso

- El programa tiene una duración de 32 hrs.
- Las clases serán impartidas los días sábado, de 9:00 am a 1:00 pm
- Serán asignados ejercicios que el participante deberá resolver entre una semana y otra.
- Al final del curso se solicitará un proyecto final, el cual **deberá ser entregado para ser acreedor a la constancia de participación.**

1.6 Recursos y dinámica de clase

En esta clase estaremos usando:

- R da click aquí si aún no lo descargas
- RStudio da click aquí también
- Miro úsame
- Zoom Clases
 - Pulgar arriba: Voy bien, estoy entendiendo!
 - Pulgar abajo: Eso no quedó muy claro
 - Mano arriba: Quiero participar/preguntar ó Ya estoy listo para iniciar
- Grupo de WhatsApp El chismecito está aquí
- Google Drive
- Notas de clase Revisame si quieres aprender
- Documento del taller de Scoping.

Chapter 2

INTRODUCCIÓN A CIENCIA DE DATOS

2.1 ¿Qué es Ciencia de Datos?

2.1.1 Definiendo conceptos:

Estadística Disciplina que recolecta, organiza, analiza e interpreta datos. Lo hace a través de una población muestral generando estadística descriptiva y estadística inferencial.

- La estadística descriptiva, como su nombre lo indica, se encarga de describir datos y obtener conclusiones.
- La estadística inferencial argumenta sus resultados a partir de las muestras de una población.
- En la estadística descriptiva se utilizan números, como medidas, para analizar datos y llegar a conclusiones de acuerdo a ellos. Con la estadística inferencial se intenta conseguir información al utilizar un procedimiento ordenado en el manejo de los datos de la muestra.
- Algunos dicen que La estadística inferencial se encarga de realizar el cálculo de la probabilidad de que algo ocurra en el futuro



Población estadística





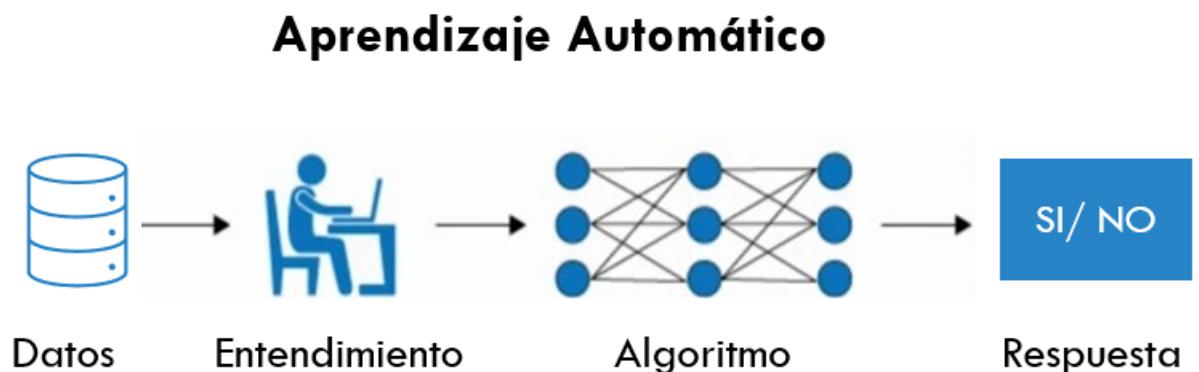
- **Business Intelligence:** BI aprovecha el software y los servicios para transformar los datos en conocimientos prácticos que informan las decisiones empresariales estratégicas y tácticas de una organización. Las herramientas de BI acceden y analizan conjuntos de datos y presentan hallazgos analíticos en informes, resúmenes, tableros, gráficos, cuadros, -indicadores- o KPI's y mapas para proporcionar a los usuarios inteligencia detallada sobre el estado del negocio. (BI está enfocado en analizar la historia pasada)

¿Qué características tiene un KPI?

- Específicos
- Continuos y periódicos
- Objetivos
- Cuantificables
- Medibles
- Realistas
- Concisos
- Coherentes
- Relevantes



- **Machine Learning:** El ‘machine learning’ –aprendizaje automático– es una rama de la inteligencia artificial que permite que las máquinas aprendan de los patrones existentes en los datos. Se usan métodos computacionales para aprender de datos con el fin de producir reglas para mejorar el desempeño en alguna tarea o toma de decisión. (Está enfocado en la programación de máquinas para aprender de los patrones existentes en datos principalmente estructurados y anticiparse al futuro)

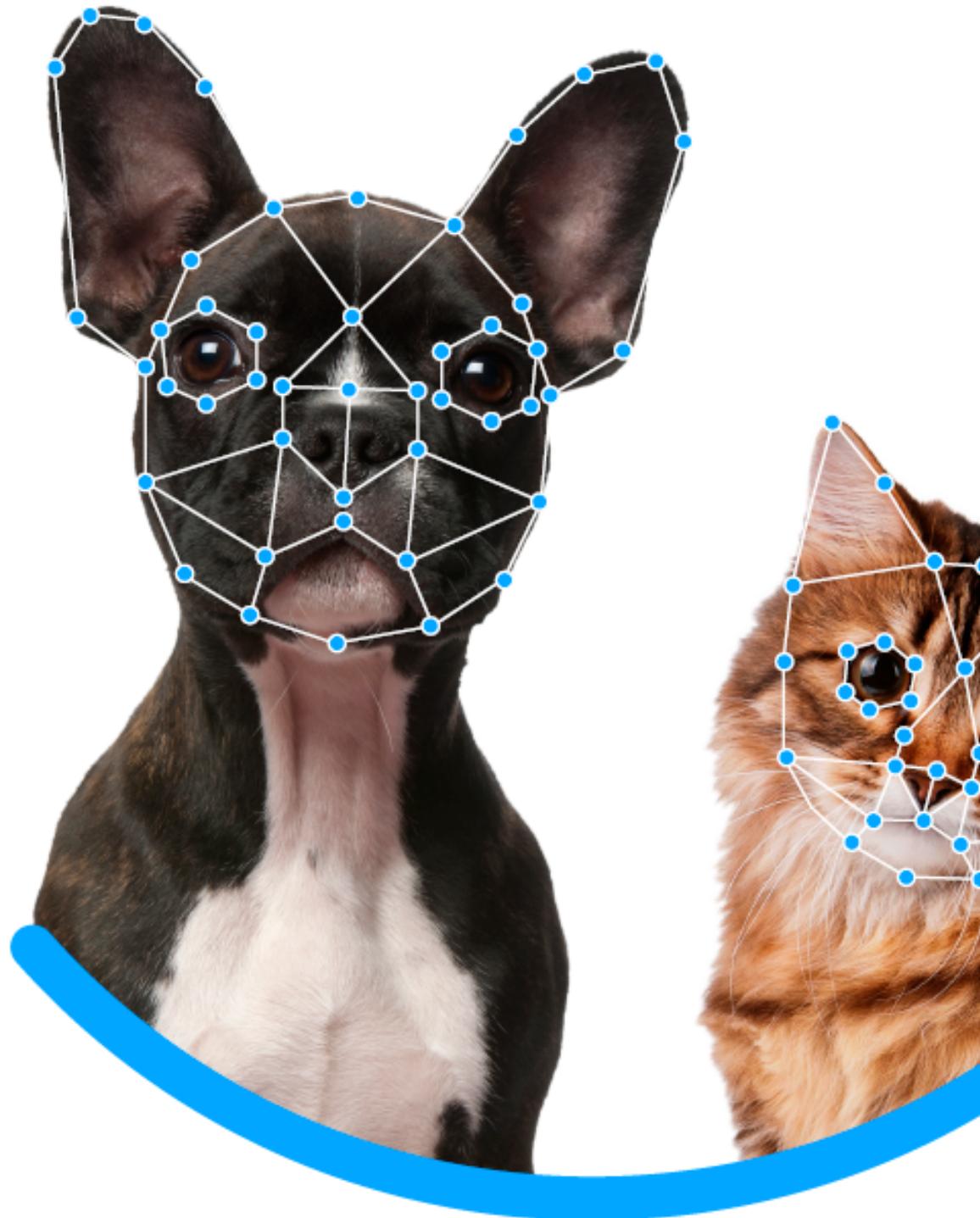


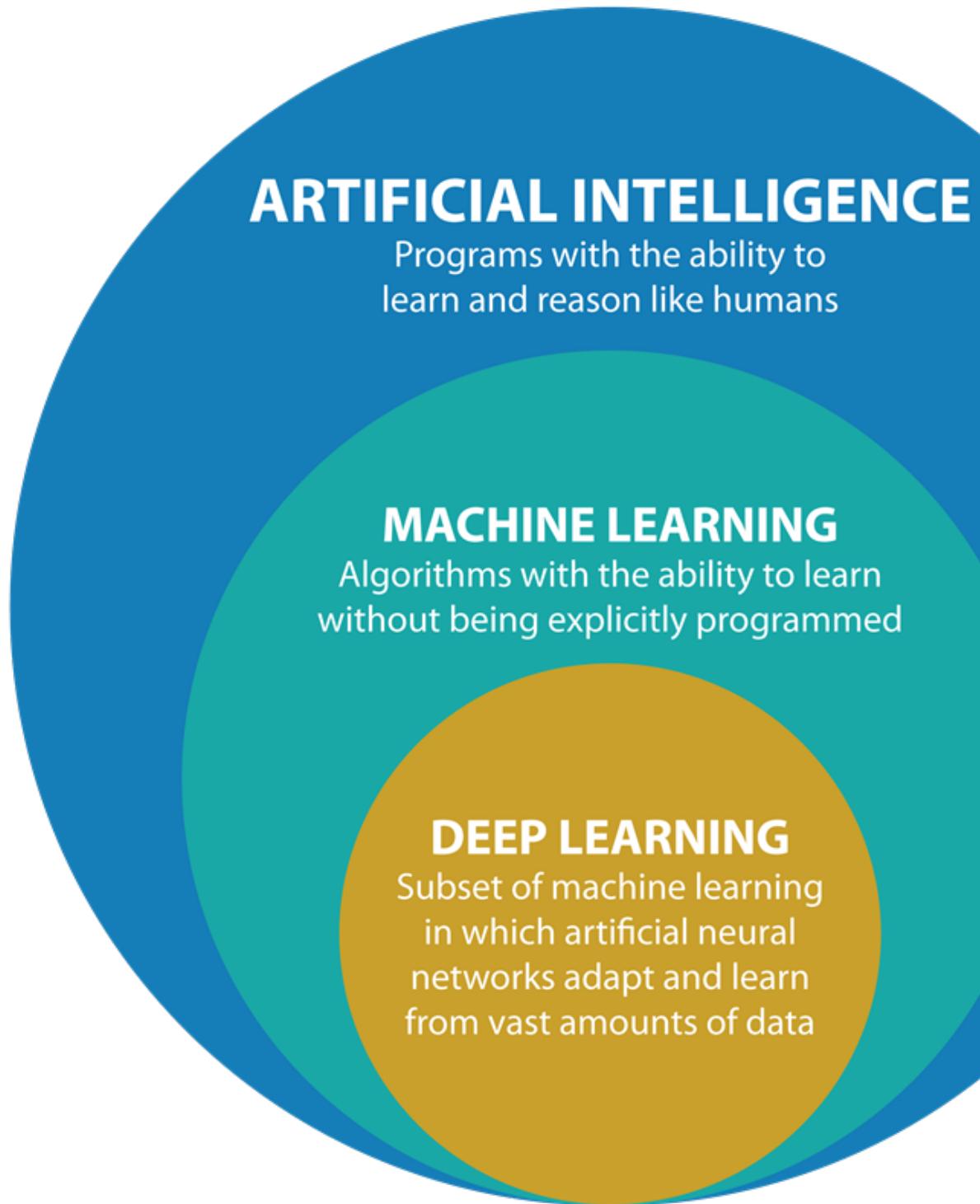


- **Deep Learning:** El aprendizaje profundo es un subcampo del aprendizaje automático que se ocupa de los algoritmos inspirados en la estructura y función del cerebro llamados redes neuronales artificiales.

En *Deep Learning*, un modelo de computadora aprende a realizar tareas de clasificación directamente a partir de imágenes, texto o sonido. Los modelos de aprendizaje profundo pueden lograr una precisión de vanguardia, a veces superando el rendimiento a nivel humano. Los modelos se entran mediante el uso de un gran conjunto de datos etiquetados y arquitecturas

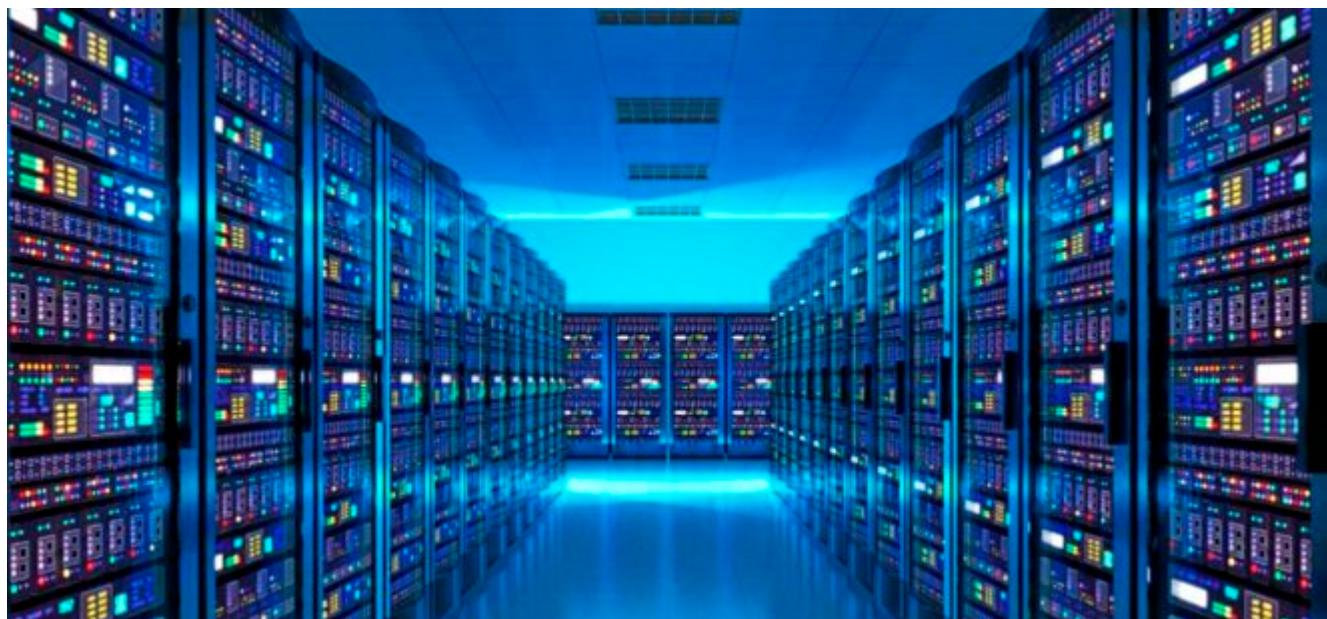
de redes neuronales que contienen muchas capas. (Está enfocado en la programación de máquinas para el reconocimiento de imágenes y audio (datos no estructurados))





- **Big data** se refiere a los grandes y diversos conjuntos de información que crecen a un ritmo cada vez mayor. Abarca el volumen de información, la velocidad a la que se crea y recopila, y la variedad o alcance de los puntos de datos que se cubren. Los macrodatos a menudo provienen de la minería de datos y llegan en múltiples formatos.

Es común que se confunda los conceptos de *Big Data* y *Big Compute*, como habíamos mencionado *Big Data* se refiere al procesamiento de conjuntos de datos que son más voluminosos y complejos que los tradicionales y *Big Compute* a herramientas y enfoques que utilizan una gran cantidad de recursos de CPU y memoria de forma coordinada para resolver problemas que usan algoritmos muy complejos.



Curiosidad: Servidores en líquido para ser enfriados

Curiosidad 2: Centro de datos en el océano

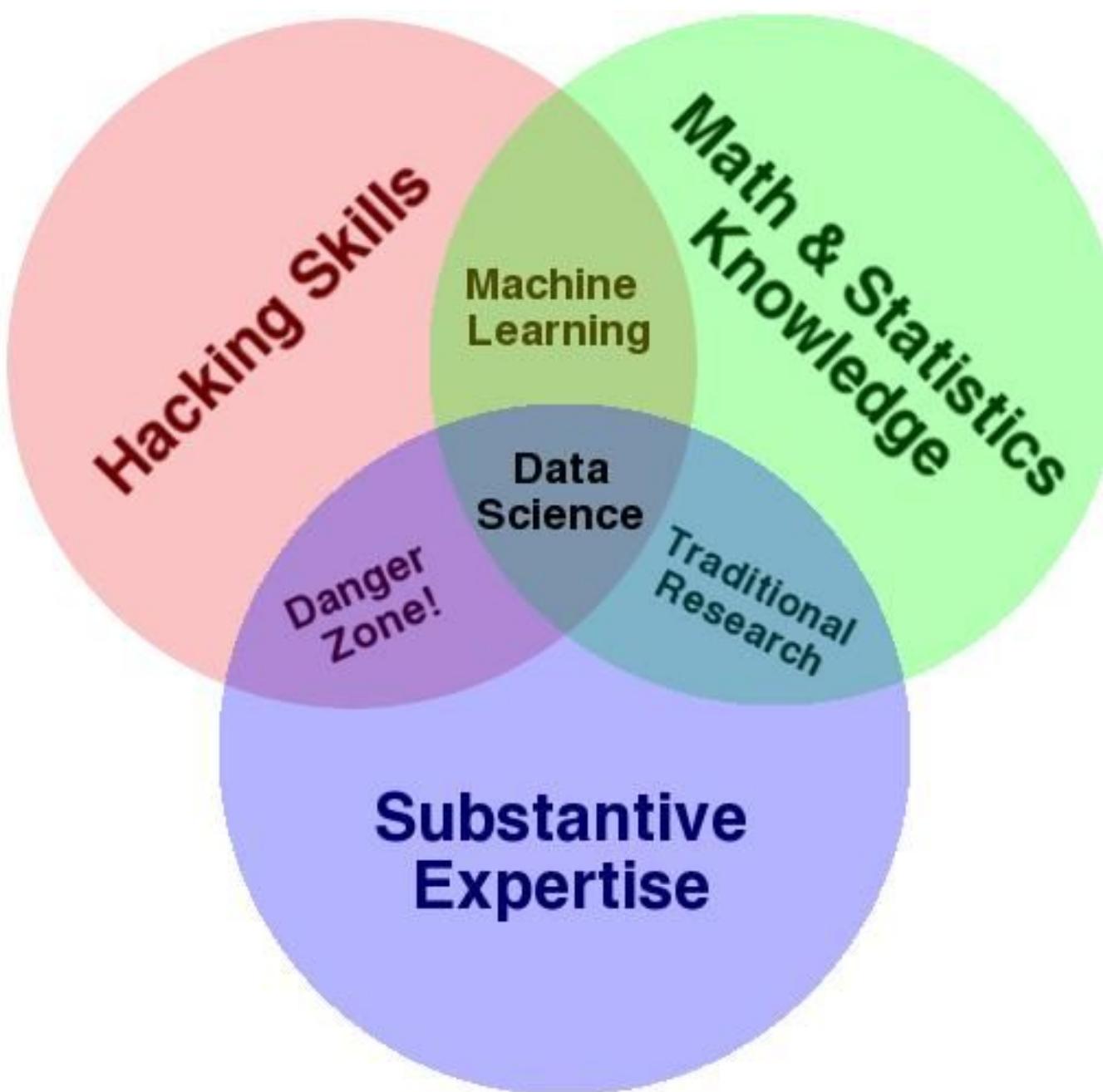
Entonces, ¿qué NO es ciencia de datos?

- No es una tecnología
- No es una herramienta
- No es desarrollo de software
- No es Business Intelligence*
- No es Big Data*
- No es Inteligencia Artificial*

- No es (solo) machine learning
- No es (solo) deep learning
- No es (solo) visualización
- No es (solo) hacer modelos

2.2 Objetivo de la Ciencia de Datos

- Los científicos de datos analizan qué preguntas necesitan respuesta y dónde encontrar los datos relacionados. Tienen conocimiento de negocio y habilidades analíticas, así como la capacidad de extraer, limpiar y presentar datos. Las empresas utilizan científicos de datos para obtener, administrar y analizar grandes cantidades de datos no estructurados. Luego, los resultados se sintetizan y comunican a las partes interesadas clave para impulsar la toma de decisiones estratégicas en la organización.



Fuente: Blog post de Drew Conway

Más sobre Conway: Forbes 2016

2.3 ¿Qué se requiere para hacer Ciencia de Datos?

- Background científico
 - Conocimientos generales de probabilidad, estadística, álgebra lineal, cálculo, geometría analítica, programación, conocimientos computacionales... etc
- Datos
 - Relevancia y suficiencia

Es indispensable saber si los datos con los que se trabajará son relevantes y suficientes, debemos evaluar qué preguntas podemos responder con los datos con los que contamos.

- Suficiencia: Los datos con los que trabajamos tienen que ser representativos de la población en general, necesitamos que las características representadas en la información sean suficientes para aproximar a la población objetivo.
- Relevancia: De igual manera los datos tienen que tener relevancia para la tarea que queremos resolver, por ejemplo, es probable que información sobre gusto en alimentos sea irrelevante para predecir número de hijos.

Relevance and

Irrelevant and Insufficient



Relevant but Insufficient



Relevant and Sufficient



- Etiquetas

- Se necesita la intervención humana para etiquetar, clasificar e introducir los datos en el algoritmo.



- Software

- Existen distintos lenguajes de programación para realizar ciencia de datos:



2.4 Aplicaciones de Ciencia de Datos

Dependiendo de la industria en la que se quiera aplicar Machine Learning, podemos pensar en distintos enfoques, en la siguiente imagen se muestran algunos ejemplos:

MACHINE LEARNING APLICACIONES

ENERGÍA

1



- Predecir fallas en refinerías
- Localizar nuevas fuentes de energía
- Analizar minerales

2

GOBIERNO

3



- Elevar eficiencia y ahorros
- Minimizar el robo de identidad
- Prevenir la corrupción

5



MINORISTAS

- Mejorar campañas de mercadotecnia
- Personalizar la oferta
- Reducir la pérdida de clientes durante el proceso de compra
- Mejorar la experiencia de compra

6



HOSPITALES

- Incrementar el éxito de una operación
- Predecir tiempos de espera en

Podemos pensar en una infinidad de aplicaciones comerciales basadas en el análisis de datos. Con la intención de estructurar las posibles aplicaciones, se ofrece a continuación una categorización que, aunque no es suficiente para englobar todos los posibles casos de uso, sí es sorprendente la cantidad de aplicaciones que abarca.

1. Aplicaciones centradas en los clientes

- Incrementar beneficio al mejorar recomendaciones de productos
- Upselling
- Cross-selling
- Reducir tasas de cancelación y mejorar tasas de retención
- Personalizar experiencia de usuario
- Mejorar el marketing dirigido
- Análisis de sentimientos
- Personalización de productos o servicios

2. Optimización de problemas

- Optimización de precios
- Ubicación de nuevas sucursales
- Maximización de ganancias mediante producción de materias primas
- Construcción de portafolios de inversión

3. Predicción de demanda

- Número futuro de clientes
- Número esperado de viajes en avión / camión / bicis
- Número de contagios por un virus (demanda médica / medicamentos / etc)
- Predicción de uso de recursos (luz / agua / gas)

4. Análisis de detección de fraudes

- Detección de robo de identidad
- Detección de transacciones ilícitas
- Detección de servicios fraudulentos
- Detección de zonas geográficas con actividades ilícitas

2.5 Tipos de aprendizaje

La diferencia entre el análisis supervisado y el no supervisado es la etiqueta, es decir, en el análisis supervisado tenemos una etiqueta “correcta” y el objetivo de los algoritmos es predecir esta etiqueta.

2.5.1 Aprendizaje supervisado

- Conocemos la respuesta correcta de antemano.
- Esta respuesta correcta fue “etiquetada” por un humano (la mayoría de las veces, en algunas circunstancias puede ser generada por otro algoritmo).
- Debido a que conocemos la respuesta correcta, existen muchas métricas de desempeño del modelo para verificar que nuestro algoritmo está haciendo las cosas “bien”.

2.5.1.1 Tipos de aprendizaje supervisado (Regresión vs clasificación)

Existen dos tipos principales de aprendizaje supervisado, esto depende del tipo de la variable respuesta:

- Los algoritmos de **clasificación** se usan cuando el resultado deseado es una etiqueta discreta, es decir, clasifican un elemento dentro de diversas clases.
- En un problema de **regresión**, la variable target o variable a predecir es un valor numérico.



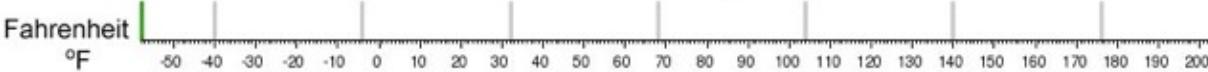
Regression

What is the temperature going to be tomorrow?

PREDICTION

84°

Fahrenheit
°F



Classification

Will it be Cold or Hot tomorrow?

PREDICTION

HOT

COLD

Fahrenheit
°F



2.5.2 Aprendizaje no supervisado

- Aquí no tenemos la respuesta correcta de antemano ¿cómo podemos saber que el algoritmo está bien o mal?
- Estadísticamente podemos verificar que el algoritmo está bien

- Siempre tenemos que verificar con el cliente si los resultados que estamos obteniendo tienen sentido de negocio. Por ejemplo, número de grupos y características



Chapter 3

CICLO DE VIDA

3.1 Ciclo de un proyecto de Ciencia de Datos

Data Science Project Lifecycle

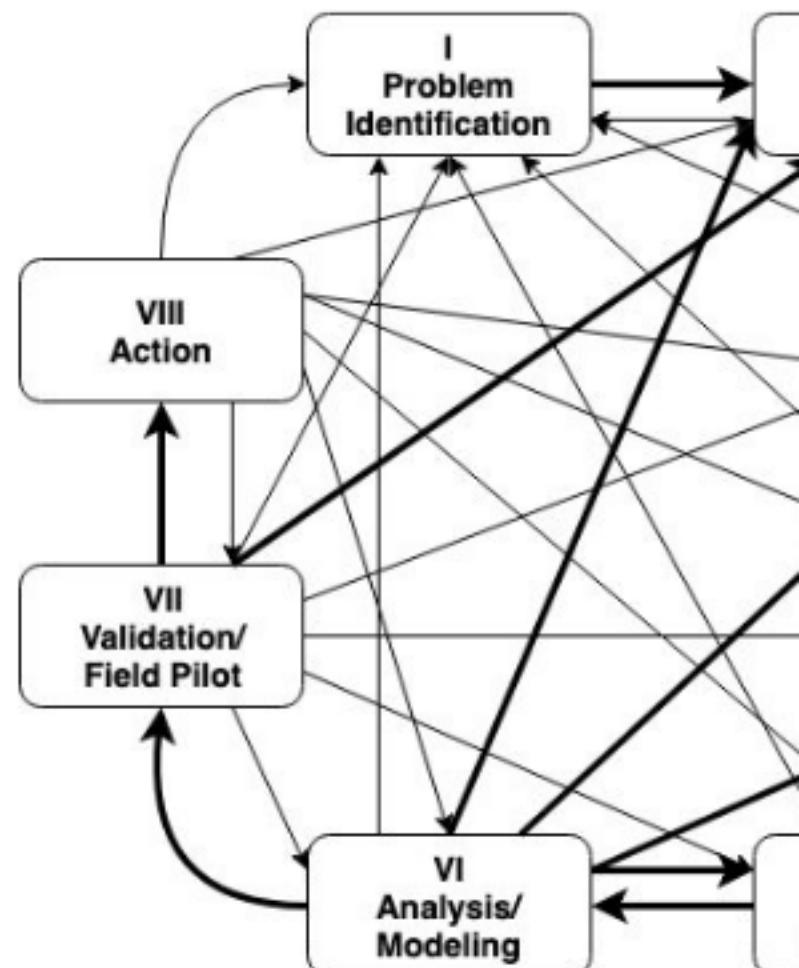


Figure 1: Lifecycle of a data science project. The phas...

1. Identificación del problema

- Debemos conocer si el problema es significativo, si el problema se puede resolver con ciencia de datos, y si habrá un compromiso real del lado de cliente/usuario/partner para implementar la solución con todas sus implicaciones: recursos físicos y humanos.



2. Scoping

- El objetivo es definir el alcance del proyecto y por lo tanto definir claramente los objetivos.

- Conocer las acciones que se llevarán a cabo para cada objetivo. Estas definirán las soluciones analíticas a hacer.
- Queremos saber si los datos con los que contamos son relevantes y suficientes.
- Hacer visible los posibles conflictos éticos que se pueden tener en esta fase.
- Debemos definir el cómo evaluaremos que el análisis de esos datos será balanceada entre eficiencia, efectividad y equidad.



3. Adquisición de datos

- Adquisición, almacenamiento, entendimiento y preparación de los datos para después poder hacer analítica sobre ellos.

- Asegurar que en la transferencia estamos cumpliendo con el manejo adecuado de datos sensibles y privados.



4. EDA

- El objetivo en esta fase es conocer los datos con los que contamos y contexto de negocio explicado a través de los mismos.
- Identificamos datos faltantes, sugerimos cómo imputarlos.
- Altamente apoyado de visualización y procesos de adquisición y limpieza de datos.



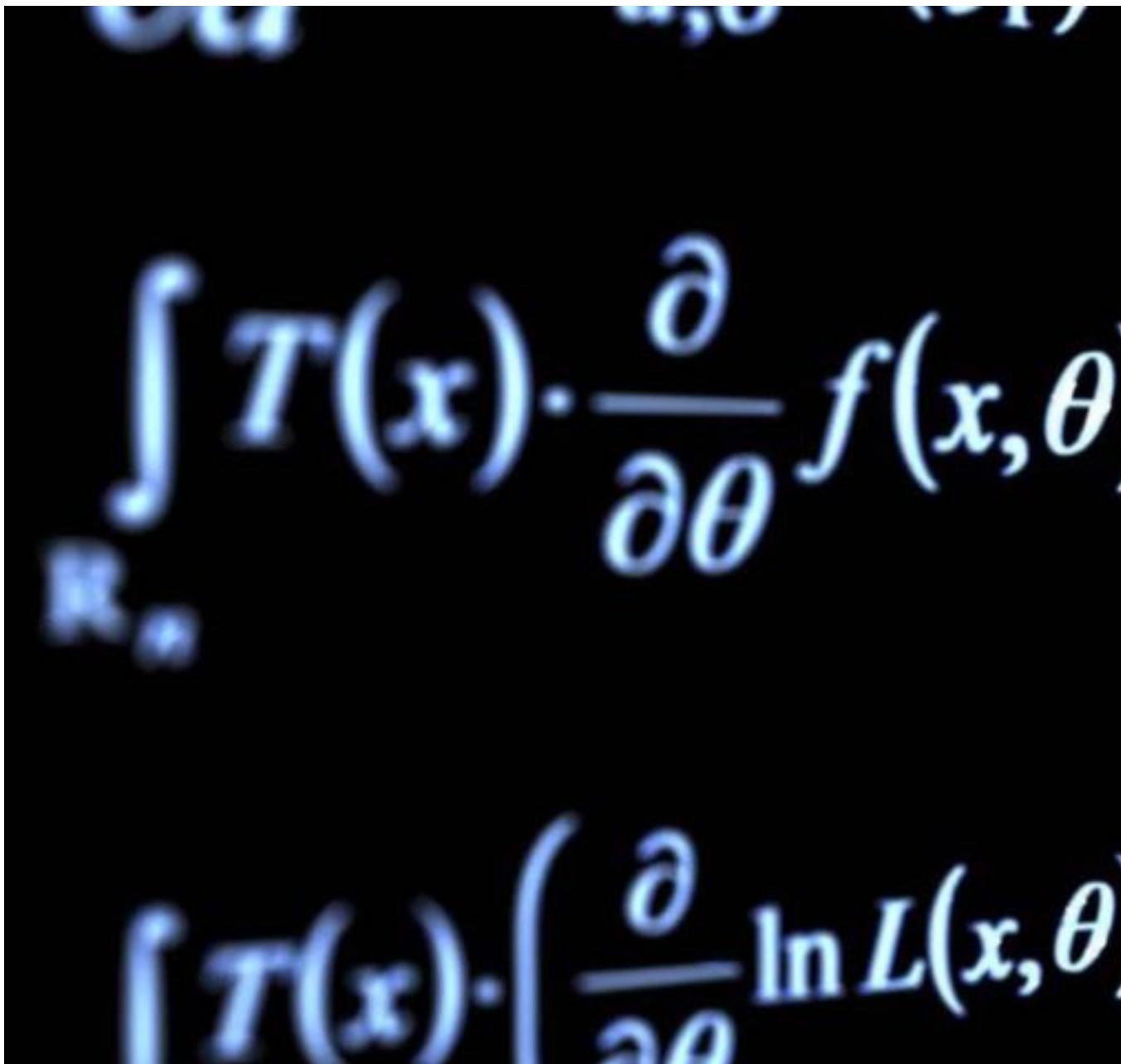
5. Formulación analítica

- Esta fase incluye empezar a formular nuestro problema como uno de ciencia de datos, el conocimiento adquirido en la fase de exploración

nos permite conocer a mayor detalle del problema y por lo tanto de la solución adecuada.

6. Modelado

- Proceso iterativo para desarrollar diferentes “experimentos”.
 - Mismo algoritmo/método diferentes hiperparámetros (grid search).
 - Diferentes algoritmos.
- Selección de un muy pequeño conjunto de modelos tomando en cuenta un balance entre interpretabilidad, complejidad, desempeño, fairness.
- Correcta interpretación de los resultados de desempeño de cada modelo.


$$\int_{\mathbb{R}_n} T(x) \cdot \frac{\partial}{\partial \theta} f(x, \theta)$$
$$\int_{\mathbb{R}_n} T(x) \cdot \left(\frac{\partial}{\partial \theta} \ln L(x, \theta) \right)$$

7. Validación

- Es muy importante poner a prueba el/los modelo/modelos seleccionados en la fase anterior. Esta prueba es en campo con datos reales, le llamamos prueba piloto.

- Debemos medir el impacto causal que nuestro modelo tuvo en un ambiente real.



8. Acciones a realizar

- Finalmente esta etapa corresponde a compartir con los tomadores de decisiones/stakeholders/creadores de política pública los resultados obtenidos y la recomendación de acciones a llevar a cabo -menú de opciones-.
- Las implicaciones éticas de esta fase consisten en hacer consciente el impacto social de nuestro trabajo.



3.2 Data Science *scoping*

El **scoping** es uno de los pasos más importante en los proyectos de ciencia de datos, es ideal realizarlo con ayuda del cliente, tiene como objetivo **definir el alcance del proyecto**, definir los objetivos, conocer las acciones que se llevaran