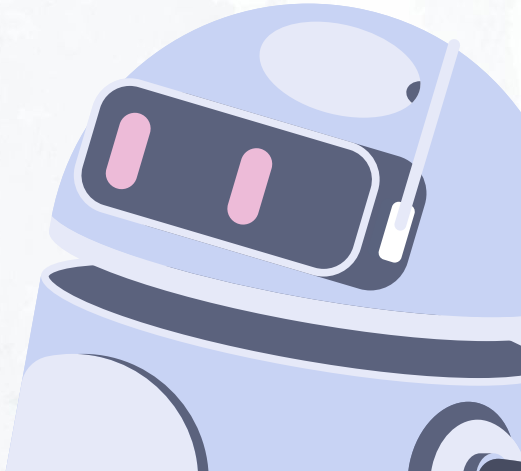
Analítica de Datos (Aprendizaje de máquina)



UNAL

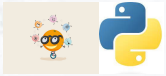
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Contenido

01 → **Introducción**

Introducción al análisis de datos
Bases de datos y tipo de variables
Taller introducción a Orange Data Mining



02 → **Preproceso y transformación de variables**

Taller en Orange Data Mining

03 → **Selección y extracción de características**

Taller en Orange Data Mining

04 → **Distancia y similitud**

Clustering y análisis de correspondencia
Taller en Orange Data Mining

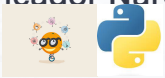
Contenido

05 → Taller guiado Orange Data Mining 

06 → Aprendizaje supervisado

Aprendizaje supervisado, básicos de ML, métricas de evaluación, técnicas de validación, clasificador KNN
Taller en Orange Data Mining 

07 → Naive Bayes

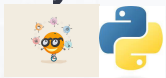
Repaso de probabilidad y clasificador Naive Bayes
Taller en Orange Data Mining 

08 → Árboles de decisión y SVMs

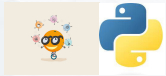
Taller en Orange Data Mining 

Contenido

09 → Regresión logística y redes neuronales

Taller en Orange Data Mining 

10 → Random Forest, ensemble learning (boosting y bagging)

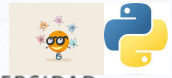
Taller en Orange Data Mining 

11 → Regresión lineal, métricas de evaluación

Taller en Orange Data Mining 

12 → Análisis de series de tiempo

Conceptos básicos de series de tiempo, manejo de la variable tiempo, series de tiempo, modelos auto-regresivos

Taller en Orange Data Mining 

Contenido

13 → **Manejo de datos como redes y grafos**

Taller en Orange Data Mining 

14 → **Deep Learning**

Redes neuronales convolucionales y transfer learning

Taller en Python 

15 → **Herramientas de análisis de datos**

Aplicativos web y dashboards 

Metodología de la asignatura

- **Componentes mixtas de aprendizaje:**
  60% de clases de contenido teórico
  40% de clases de contenido práctico

- **Horarios de clases:**
  Lunes de 9:00am a 11:00am
  Miércoles de 2:00 pm a 4:00pm

- **Asistencia:** Se debe cumplir con una asistencia mínima al 80% del curso y la presentación de los resultados de un proyecto integrador.
- Porcentajes:
  - Proyecto integrador: 40%
  - Talleres: 45 %
  - Asistencia: 10%
  - Laboratorio: 5 %

Metodología de la asignatura

- La **evaluación del curso** se hará mediante el desarrollo de un proyecto integrador
    - Se espera la aplicación de técnicas y conocimientos en la solución de un problema real de análisis de datos.
    - Los **datos** de origen pueden ser reales (un problema real de sus empresas, negocios o proyectos) o artificiales. Se firmarán acuerdos de confidencialidad de los datos en caso de ser necesario.

- **Fechas importantes:**
    - Presentación de idea del proyecto integrador – 18 de diciembre de 2024
    - Presentación proyecto integrador – 5 de marzo de 2025

Proyecto integrador

El **objetivo** de este proyecto es la integración final de todas las técnicas y conocimientos adquiridos en este curso en la solución de un problema real de análisis de datos.
Este proyecto puede ser desarrollado en **grupos de 4 personas**.

Respecto al proyecto:

- Debe partir de datos confiables y de calidad.
- Trabajar en una definición clara de un problema.
- Considerar un **alcance medible y coherente**.

UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

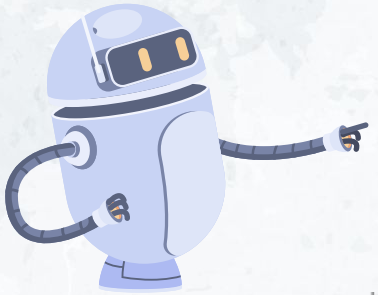
Pregunta →

¿Cuentan con experiencia en programación - Python?

Con poca, moderada o mucha experiencia: XX estudiantes
Ninguna experiencia: XX estudiantes

Tener datos NO es tener información

- Los datos se pueden encontrar fácilmente en cualquier lado
- La información hay que saber cómo y dónde buscarla.
- Normalmente subyace escondida detrás de los datos.
- Obtenerla, requiere no solo del procesamiento y del análisis de los datos.



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Tener datos y saber analizar datos NO es tener información

(+) Para poder analizar los datos de manera efectiva, es esencial entender el problema subyacente que se está estudiando.

(+) Al tener una buena comprensión del problema, los analistas de datos pueden entender mejor los datos, refinar sus preguntas de investigación y desarrollar mejores insights y soluciones.

(+) Entender el problema también permite a los analistas de datos asegurarse de que los datos recopilados sean válidos y precisos para el análisis.

UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

(AD) = Análisis de datos

- Recopilar, procesar, almacenar y analizar datos para extraer información relevante.
- Utilizar herramientas de análisis estadístico para identificar tendencias, patrones y relaciones entre los datos.
- Elaborar modelos de inferencia para la predicción y clasificación de patrones.
- Ayudar a otras dependencias en la institución a tomar decisiones basadas en los resultados de los análisis de datos.

UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

¿Cómo relacionamos datos con información?

Raw Data (datos en crudo): Tener las cifras de un determinado fenómeno.

Information (información): Poder analizar los datos y extraer de esas cifras relaciones, dependencias,influencias, causas y posibles consecuencias.

Knowledge (conocimiento): Saber cómo hacer frente a la información obtenida.

Wisdom (sabiduría): Tener el poder para hacer frente y tomar decisiones

Veamos un ejemplo: el mito de los pañales y la cerveza

El mito cuenta que en una cadena de almacenes de los Estados Unidos (Walmart o Costco) analizaron los datos de compras de sus clientes

***Raw Data*:** Los registros de los artículos que los clientes habían comprado, junto con datos relativos a la hora, el género del comprador y la edad.

Information*:** Se descubrió una alta correlación entre: ***compradores hombres, ***compras entre 5pm y 7pm***, ***pañales*** y ***cervezas***.

***Knowledge*:** Saber que los padres, después de salir del trabajo, suelen comprar pañales y también cervezas.

***Wisdom*:** Implementar nuevas estrategias de publicidad y mercadeo.

La realidad de los datos hoy

¿Cuál es el origen de los datos?

Los datos, hoy en día, tienen un origen multimodal.

Las personas nos encargamos de generar datos mediante:

- El uso de aplicaciones en nuestros dispositivos móviles (interacciones en redes sociales, geolocalización, monitoreo de la actividad física).
- El empleo de medios de pago electrónico.
- El uso de las tarjetas de fidelización (en almacenes de cadena, supermercados).
- Dispositivos conectados a la nube (asistentes en casa, internet de las cosas, domótica, wearables).

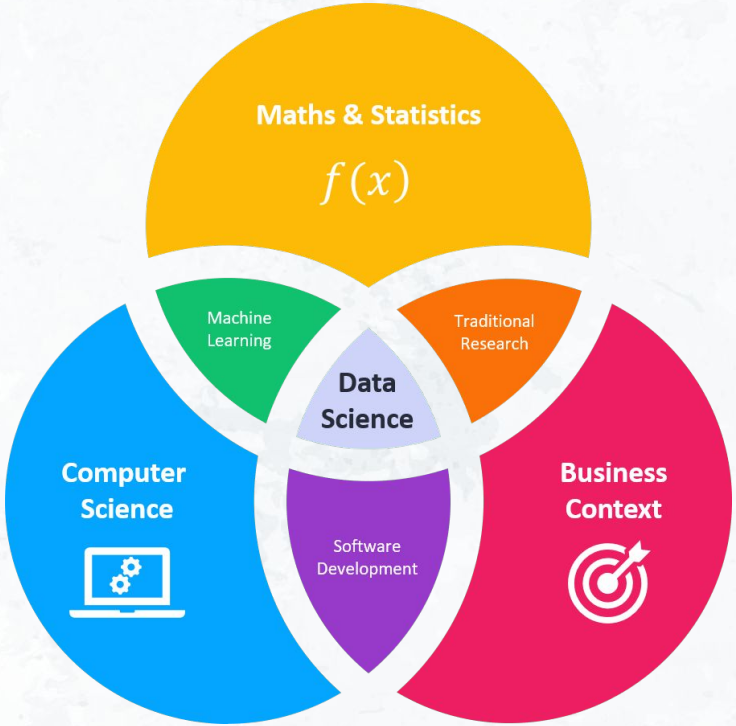
... y en muchas otras formas.

UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Alrededor de la ciencia de datos

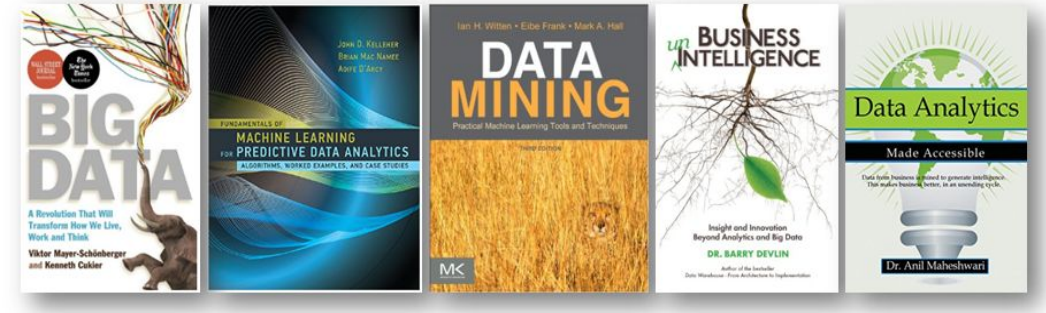


La ciencia de datos está relacionada con diversas áreas.

Alrededor de la ciencia de datos

La ciencia de datos está relacionada también con diversas disciplinas.

En donde a veces encontramos mucho ruido.



Data science

From Wikipedia, the free encyclopedia

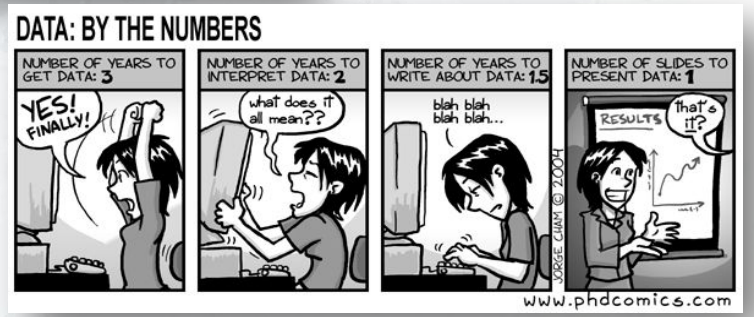
Not to be confused with information science.

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics,[3] similar to Knowledge Discovery in Databases (KDD).

Overview [edit]

Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, operations research,[4] information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing. Methods that scale to big data are of particular interest in data science, although the

El flujo de trabajo en la ciencia de datos



¿Qué podemos hacer con los datos?

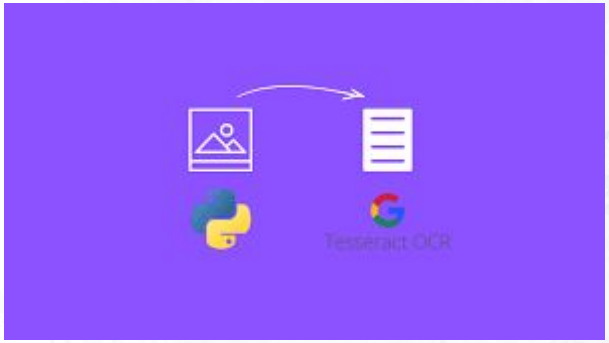
Análisis

- Describir el estado actual de una organización o de un proceso.
- Detectar eventos anómalos o poco frecuentes.
- Diagnosticar la causa de eventos o comportamientos.

Modelamiento

- Predecir futuros eventos.
- Pronosticar cambios.

Análisis = extraer información



Cuando analizamos datos, estos:

- Se inspeccionan.
- Se limpian.
- Se transforman.

Con el fin de encontrar información útil que lleve a concluir sobre un proceso o fenómeno y apoye la toma de decisiones.

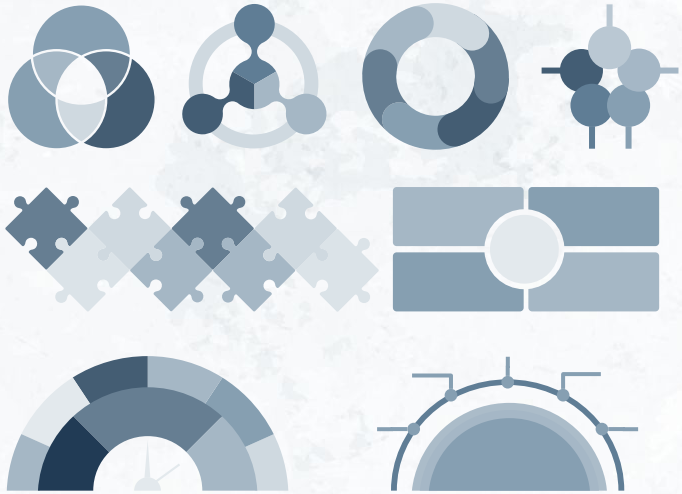
Tipos de análisis

De acuerdo con la naturaleza de la información

- Cualitativo.
- Cuantitativo.

De acuerdo con el objetivo

- Análisis Exploratorio (ADE)
- Análisis Confirmatorio (ADC)



Modelamiento = entender y predecir

Cuando modelamos datos:

- Buscamos **extrapolar el comportamiento** de un fenómeno ante otras condiciones de entrada.
- **Clasificamos** distintos conjuntos de datos en clases que pueden ser:
  - Conocidas (aprendizaje supervisado)
  - Desconocidas (aprendizaje no-supervisado)
- Predecimos comportamientos de personas o variables.

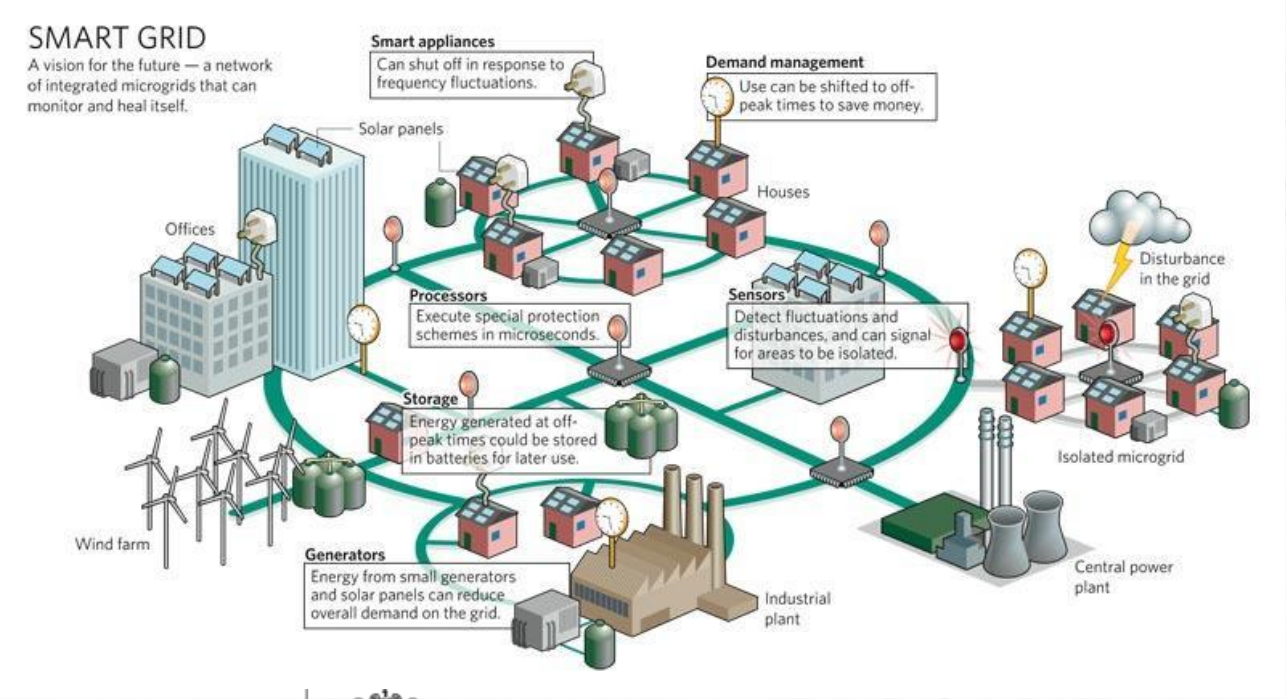
Todo con el fin de adquirir un entendimiento más amplio de los fenómenos bajo estudio.

Aplicación de la Analítica de Datos →

Algunos ejemplos:

Smart Grids

Redes eléctricas inteligentes



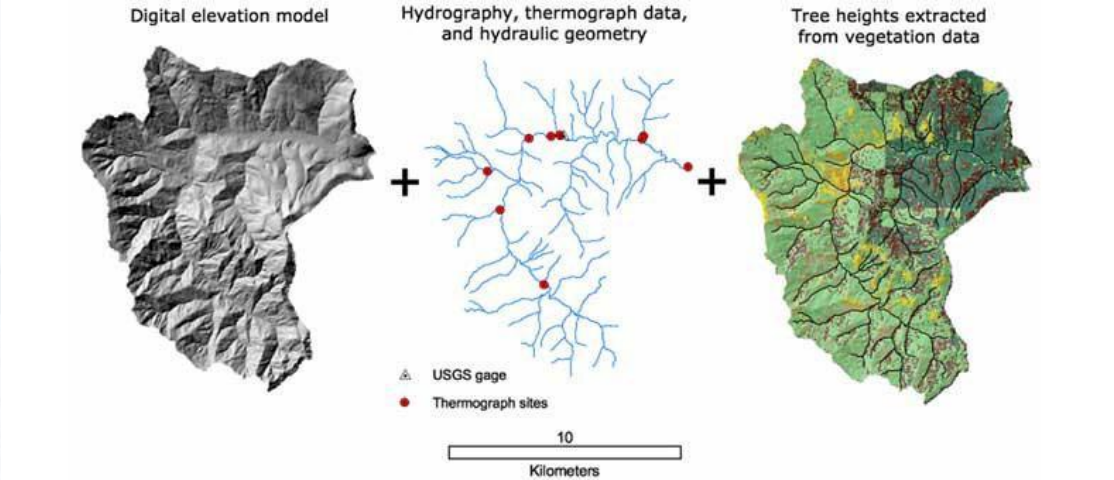
Sector Agro

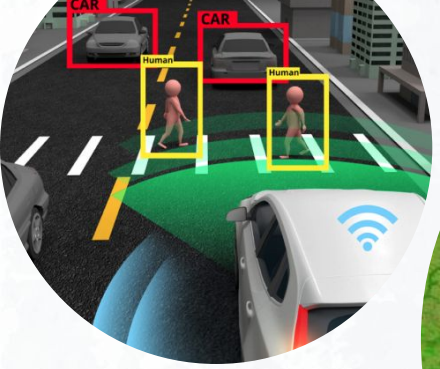
Agricultura de precisión

- Nutrientes en el suelo.
- Contenido de agua.
- Imágenes de espectroscopia.

Impacto en ciencias ambientales

- Planes de reparación y mitigación del medio ambiente.
- Economía medioambiental.
- Apoyar políticas gubernamentales.





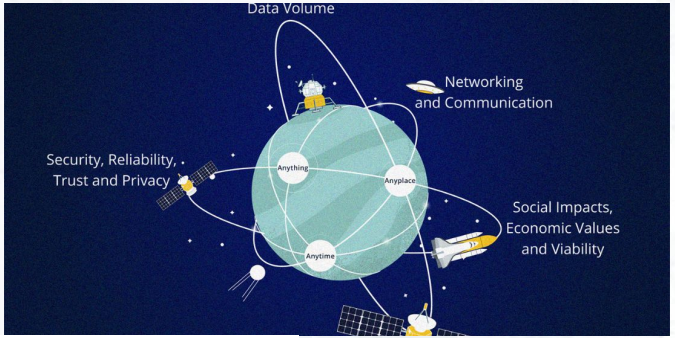
En reconocimiento y clasificación de imágenes

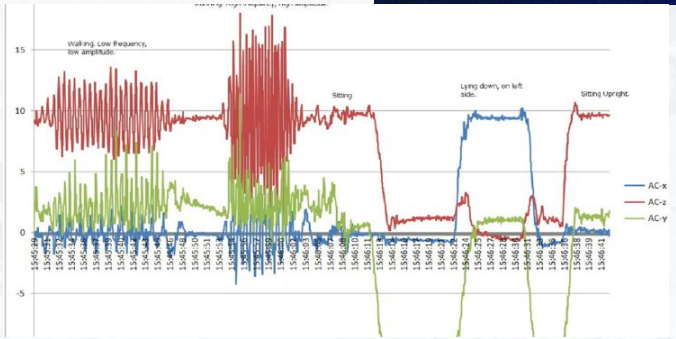
- Imágenes satelitales en estudios ambientales.
- Imágenes multiespectrales en detección de patrones.

Internet de las cosas (Internet of Things – IoT)

El Internet de las cosas involucra:

- Hardware, programación, seguridad, redes de trabajo e integración en la nube, análisis de datos y predicción, aprendizaje de máquina e inteligencia artificial.





UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Inteligencia de negocios

La inteligencia de negocios se ha abierto campo como la disciplina encargada de involucrar el análisis cuantitativo de datos en la toma de decisiones. Algunos ejemplos:

- Tarjetas de fidelización de clientes.
- Segmentación de mercados regionales.
- Mejorar la logística y los canales de distribución de bienes y servicios.

Minería de datos

Portales de noticias como

https://news.google.com

agregan en tiempo real noticias de distintas fuentes.

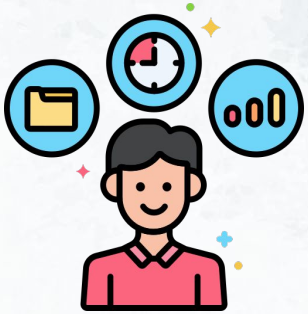
Spotify

amazon prime video

NETFLIX

Algoritmos de recomendación de plataformas de Streaming (de video y música)

La realidad de los datos hoy



(+) ¿Cuántos datos tenemos?

El 90% de los datos existen hoy en el mundo fueron creados en los últimos dos años.

De hecho, en los últimos dos años se han creado más datos que en toda la historia de la raza humana.

Internet usage worldwide – statistics & facts | Statista

Solo el 1% de estos datos es aprovechado para extraer información y ser aprovechada en los negocios.

La realidad de los datos hoy

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Data Scientist: The Sexiest Job of the 21st Century (hbr.org)

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022

Is Data Scientist Still the Sexiest Job of the 21st Century? (hbr.org)

What Do We Do About the Biases in AI?

by James Manyika, Jake Silberg, and Brittany Presten

October 25, 2019

What Do We Do About the Biases in AI? (hbr.org)

¿Qué debe motivarnos?

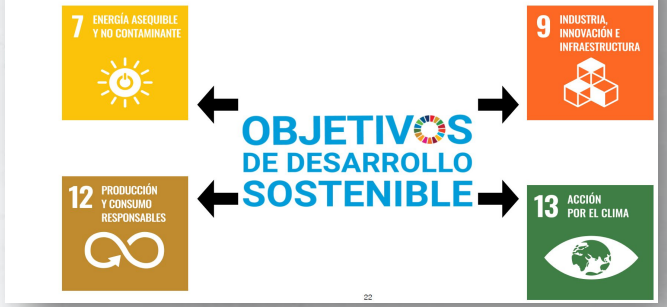
unesco

Ética de la inteligencia artificial

Ética de la inteligencia artificial | UNESCO

Desarrollo sostenible

Industrias 4.0 apoyando la sostenibilidad



¿Qué recursos usaremos?

https://www.python.org

https://colab.research.google.com/

https://orangedatamining.com

¿Qué recursos usaremos?

Algunos repositorios para bases de datos

Repositorios comunes (populares)

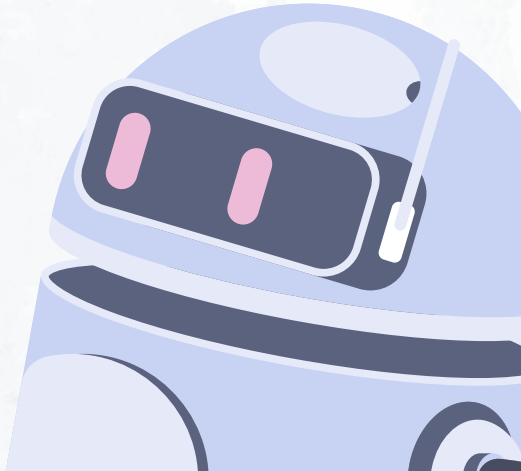
https://www.kaggle.com/datasets
https://github.com/awesomedata/awesome-public-datasets
https://archive.ics.uci.edu/ml/datasets.php
https://paperswithcode.com/datasets
https://www.openml.org/
https://datasetsearch.research.google.com

Repositorios de google

https://ai.google.com/research/ConceptualCaptions/
http://hdrplusdata.org/dataset.html
https://research.google.com/youtube8m/
http://hdrplusdata.org/dataset.html (un challenge en kaggle del anterior dataset)
https://google.github.io/realestate10k/
https://research.google.com/ava/

Gracias !

dfcollazosh@unal.edu.co



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD
NACIONAL
DE COLOMBIA