Cultura de Datos: Pensar, Visualizar, Experimentar
Clase 1

Diplomado IIEG 2026
Dr. E. Ulises Moya Sánchez
ulises.moya@iieg.gob.mx



Índice

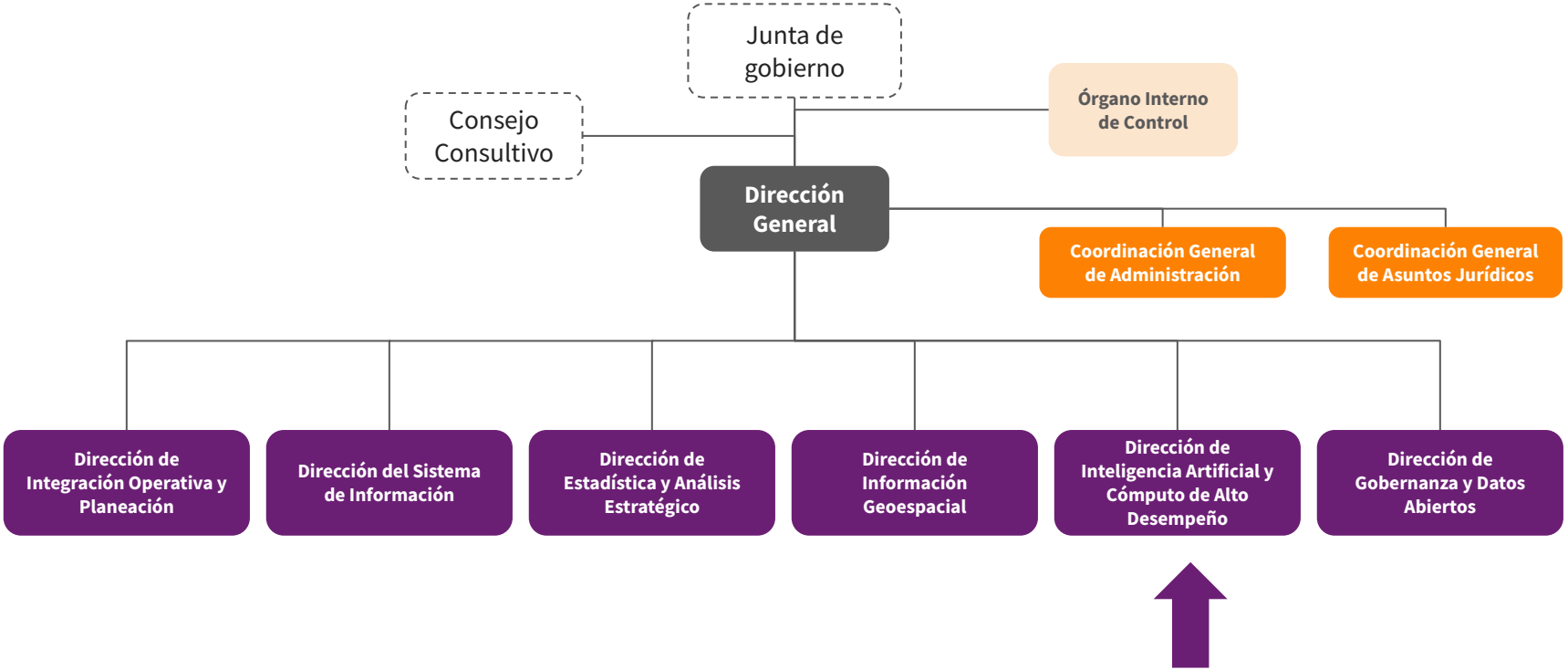
- Introducción: Profesor, alumnos y curso
- Motivación y cultura de datos
- ¿Qué son los datos?
- Componentes de los datos
- Tipo de datos
- Ciclo de vida los datos

IIEG
Instituto de Información Estadistica y Geográfica de Jalisco

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Introducción

Estructura propuesta (simplificada)



Ecosistema de Inteligencia Artificial en Jalisco.

Startups, Spin-offs y pymes.
Financieras, educación, servicios, robótica.

Empresas líderes
Diseño electrónico, hardware, análisis de datos, aprendizaje automático.

Centros de análisis de datos y supercómputo

Centros de investigación and Clusters empresariales

Gobierno e instituciones

Socios

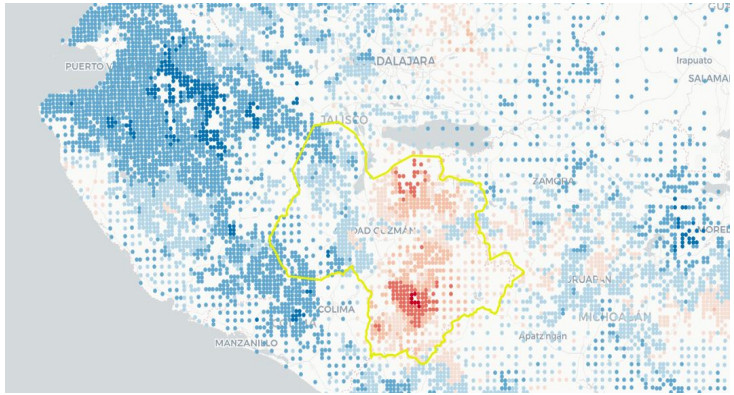
Conferencias internacionales Inteligencia artificial

Universidades

Eventos internacionales y Comunidades



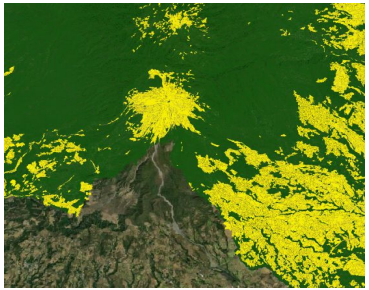
Alertas tempranas de deforestación con IA



Mapa original



Predicción con IA



El proyecto es una colaboración entre CONAFOR, SEMADET, USAID, y US Forest de desarrollo de modelos de inteligencia artificial (IA) que **detectan la deforestación ilegal en una región del cinturón de producción de aguacate**. El objetivo es **permitir a las autoridades identificar áreas críticas que requieren restauración y conservación.** El proyecto fue seleccionado entre 23 propuestas a nivel global por el GPAI.

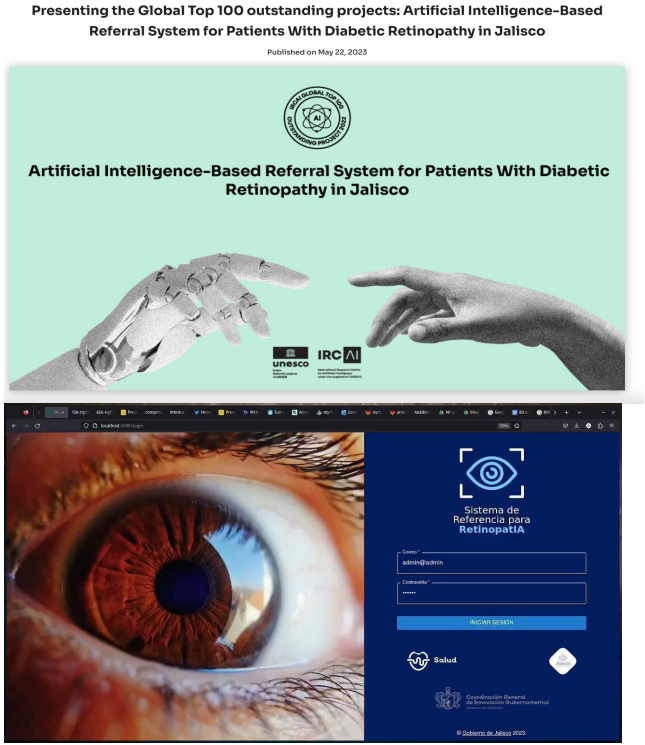
1
Premio internacional OCDE

1
Colaboración internacional

14
Municipios en Jalisco

[GPAI- Scale up solutions 2023]

Sistema de referencia de retinopatía diabética con IA



Este sistema de referencia con IA, desarrollado en la Coordinación, **posibilita tamizajes masivos para retinopatía diabética en la población**. El sistema de referencia **permite priorizar la atención a pacientes con grados de retinopatía más avanzados, evitando que pierdan la visión**. Este proyecto ha recibido tres premios internacionales, fue elegido en el 2023 dentro del Top 10 a nivel mundial por la UNESCO. También tenemos cerca de 6 artículos internacionales sobre este trabajo.

3	2	1074
Premios internacionales	Colaboraciones internacionales	Pacientes atendidos en Piloto

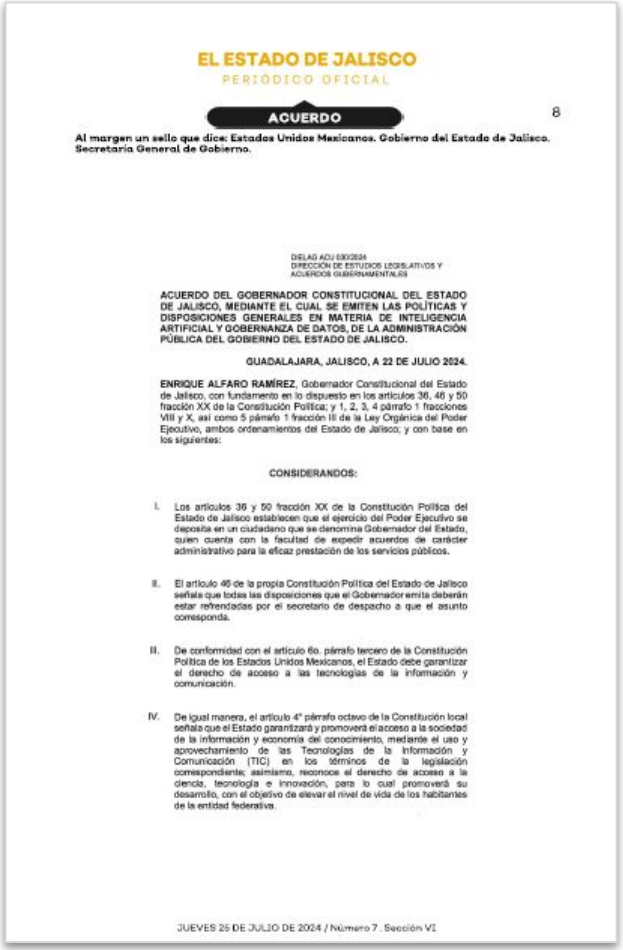
https://retina.jalisco.gob.mx/

Gracias a BID Fair-LAC, SSJ, Hosp Civil, Eyepacs, Mesidor II y IDRID

Política IA y Gobernanza de Datos
Acuerdo del Gobernador

El Gobierno de **Jalisco ha tomado la delantera en la implementación de políticas y disposiciones en materia de Inteligencia Artificial (IA) y Gobernanza de Datos**, estableciendo un marco normativo que regula su uso ético y responsable **en la Administración Pública Estatal**.

De este modo, Jalisco **se convierte en pionero en México al establecer las bases para la regulación de la IA en Jalisco** y se posiciona como un **referente en la creación de un entorno propicio para el avance tecnológico en el sector público** al tiempo que se asegura el respeto a los derechos de las y los ciudadanos y se fomenta una cultura de confianza en el uso de datos.



EL ESTADO DE JALISCO
PERIÓDICO OFICIAL

ACUERDO

8

Al margen un sello que dice: Estados Unidos Mexicanos. Gobierno del Estado de Jalisco. Secretaría General de Gobierno.

DIELAG ACU 030/2024
DIRECCIÓN DE ESTUDIOS LEGISLATIVOS Y ACUERDOS GUBERNAMENTALES

ACUERDO DEL GOBERNADOR CONSTITUCIONAL DEL ESTADO DE JALISCO, MEDIANTE EL CUAL SE EMITEN LAS POLÍTICAS Y DISPOSICIONES GENERALES EN MATERIA DE INTELIGENCIA ARTIFICIAL Y GOBERNANZA DE DATOS, DE LA ADMINISTRACIÓN PÚBLICA DEL GOBIERNO DEL ESTADO DE JALISCO.

GUADALAJARA, JALISCO, A 22 DE JULIO 2024.

ENRIQUE ALFARO RAMÍREZ, Gobernador Constitucional del Estado de Jalisco, con fundamento en lo dispuesto en los artículos 36, 46 y 50 fracción XX de la Constitución Política; y 1, 2, 3, 4 párrafo 1 fracciones VIII y X, así como 5 párrafo 1 fracción III de la Ley Orgánica del Poder Ejecutivo, ambos ordenamientos del Estado de Jalisco; y con base en los siguientes:

CONSIDERANDOS:

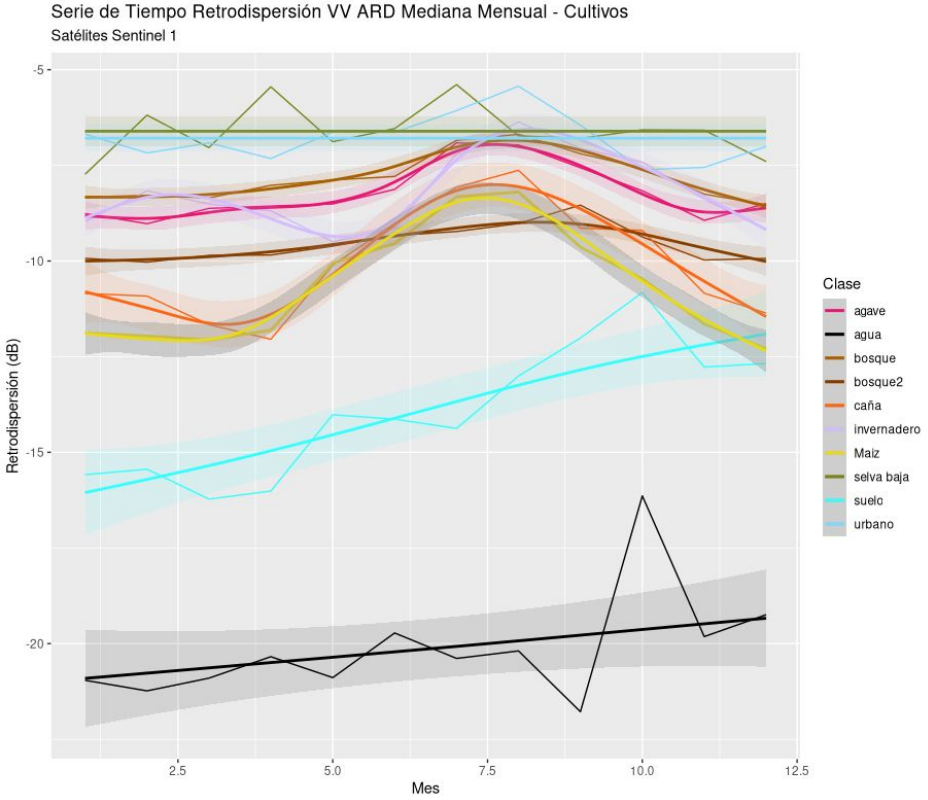
I. Los artículos 36 y 50 fracción XX de la Constitución Política del Estado de Jalisco establecen que el ejercicio del Poder Ejecutivo se deposita en un ciudadano que se denomina Gobernador del Estado, quien cuenta con la facultad de expedir acuerdos de carácter administrativo para la eficaz prestación de los servicios públicos.

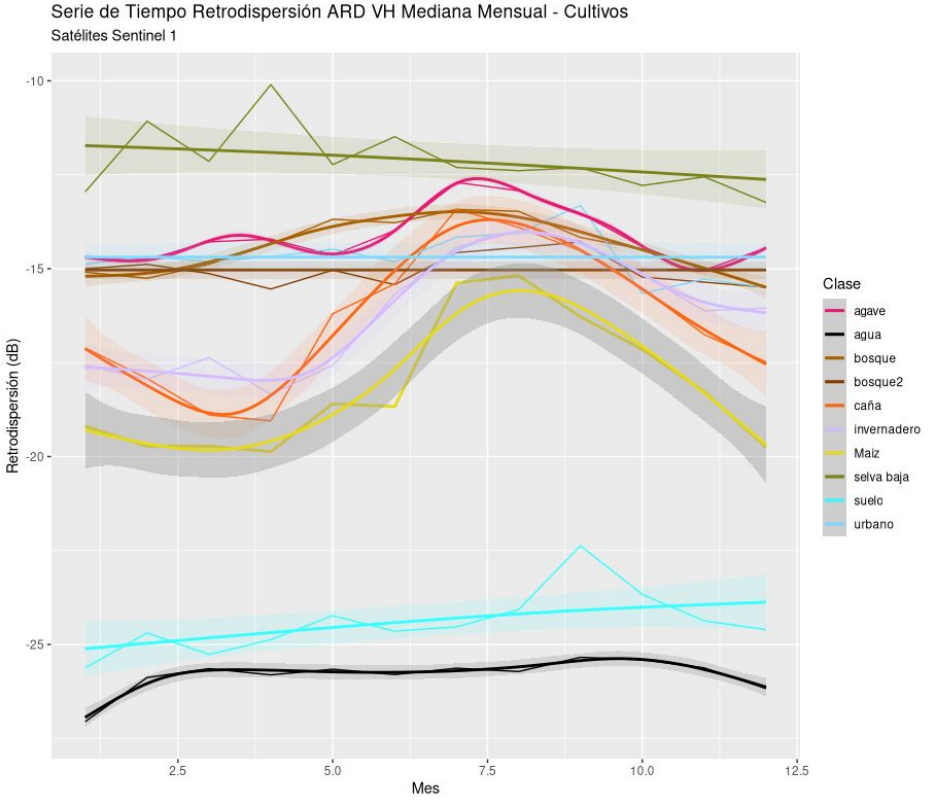
II. El artículo 46 de la propia Constitución Política del Estado de Jalisco señala que todas las disposiciones que el Gobernador emita deberán estar refrendadas por el secretario de despacho a que el asunto corresponda.

III. De conformidad con el artículo 6o. párrafo tercero de la Constitución Política de los Estados Unidos Mexicanos, el Estado debe garantizar el derecho de acceso a las tecnologías de la información y comunicación.

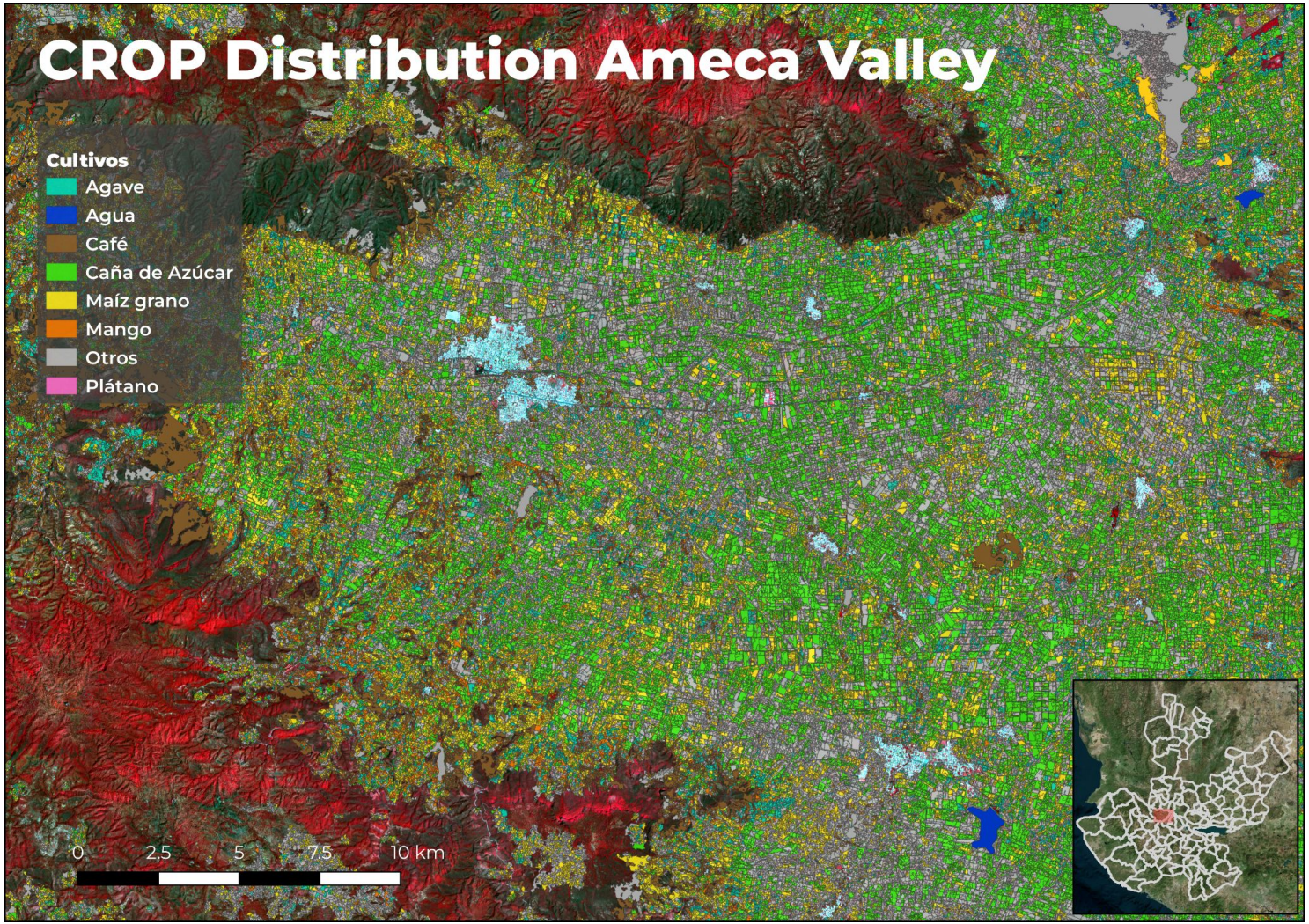
IV. De igual manera, el artículo 4° párrafo octavo de la Constitución local señala que el Estado garantizará y promoverá el acceso a la sociedad de la información y economía del conocimiento, mediante el uso y aprovechamiento de las Tecnologías de la Información y Comunicación (TIC) en los términos de la legislación correspondiente; asimismo, reconoce el derecho de acceso a la ciencia, tecnología e innovación, para lo cual promoverá su desarrollo, con el objetivo de elevar el nivel de vida de los habitantes de la entidad federativa.

JUEVES 25 DE JULIO DE 2024 / Número 7 . Sección VI

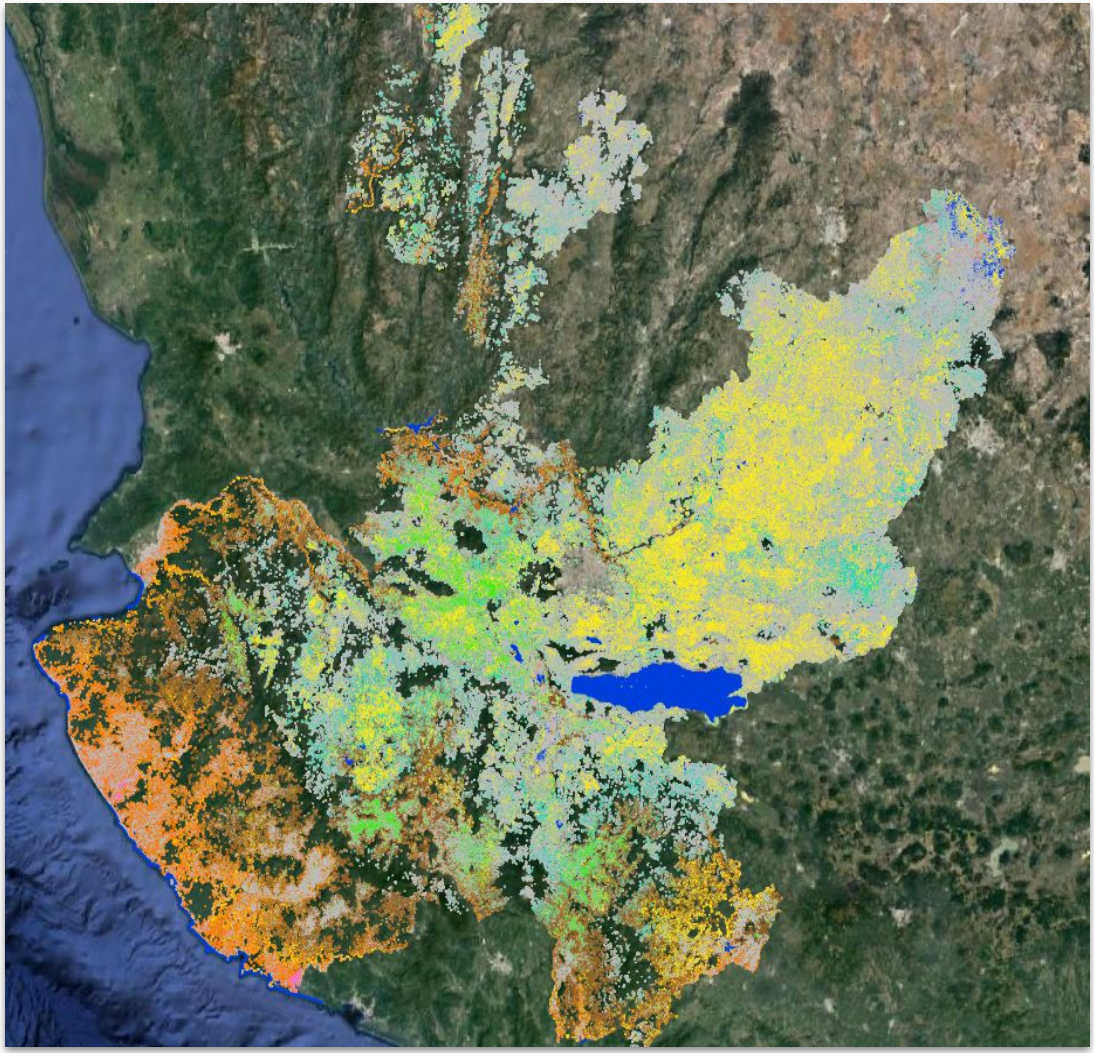

Serie de Tiempo Retrodispersión VV ARD Mediana Mensual - Cultivos
Satélites Sentinel 1
Retrodispersión (dB)
-5
-10
-15
-20
2.5
5.0
7.5
10.0
12.5
Mes
Clase
agave
agua
bosque
bosque2
caña
invernadero
Maiz
selva baja
suelo
urbano


Serie de Tiempo Retrodispersión ARD VH Mediana Mensual - Cultivos
Satélites Sentinel 1
Retrodispersión (dB)
-10
-15
-20
-25
2.5
5.0
7.5
10.0
12.5
Mes
Clase
agave
agua
bosque
bosque2
caña
invernadero
Maiz
selva baja
suelo
urbano

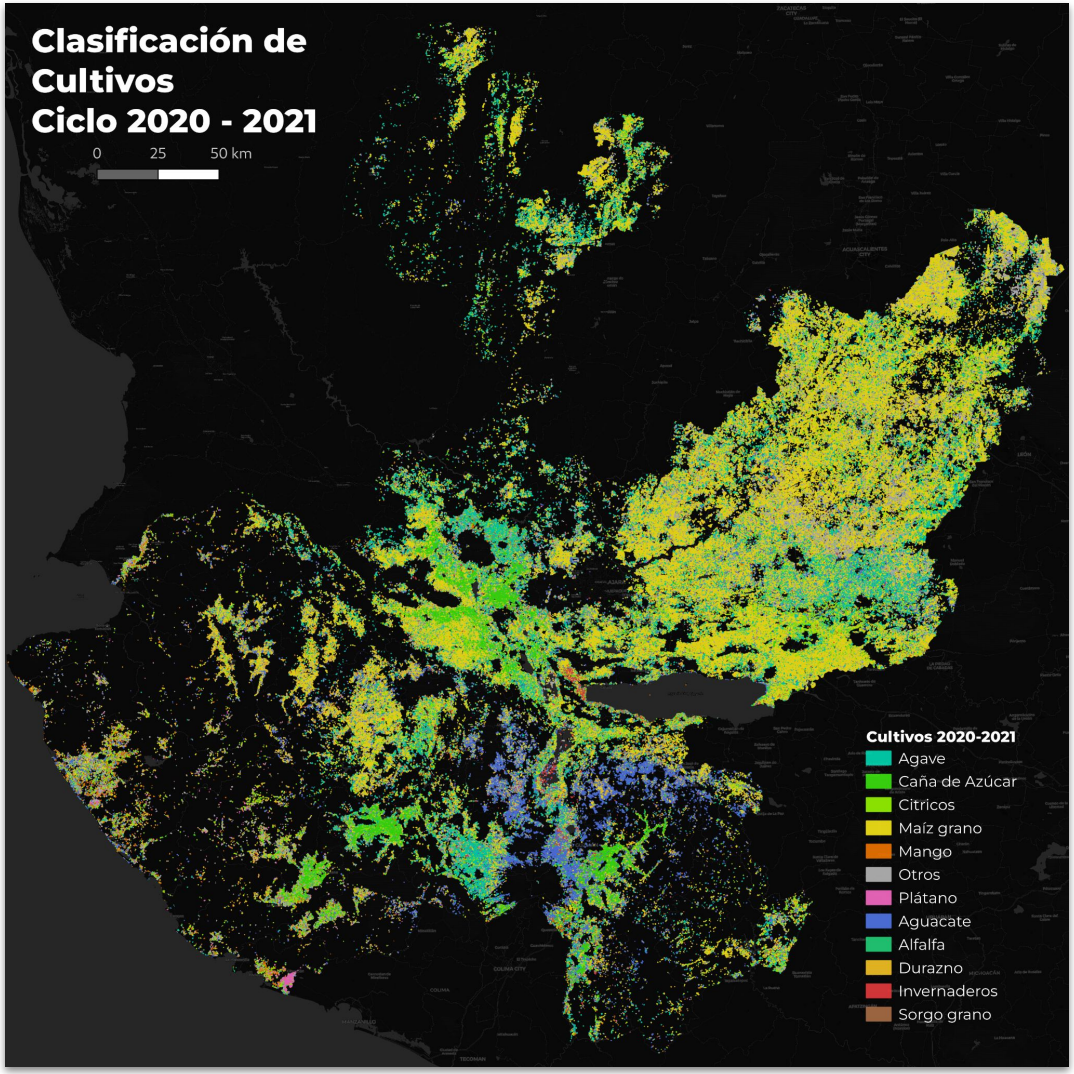
Resultados de clasificación de cultivos, con métodos de segmentación automática de imágenes tradicionales



Resultados de clasificación de cultivos, con métodos de **segmentación automática** de imágenes tradicionales

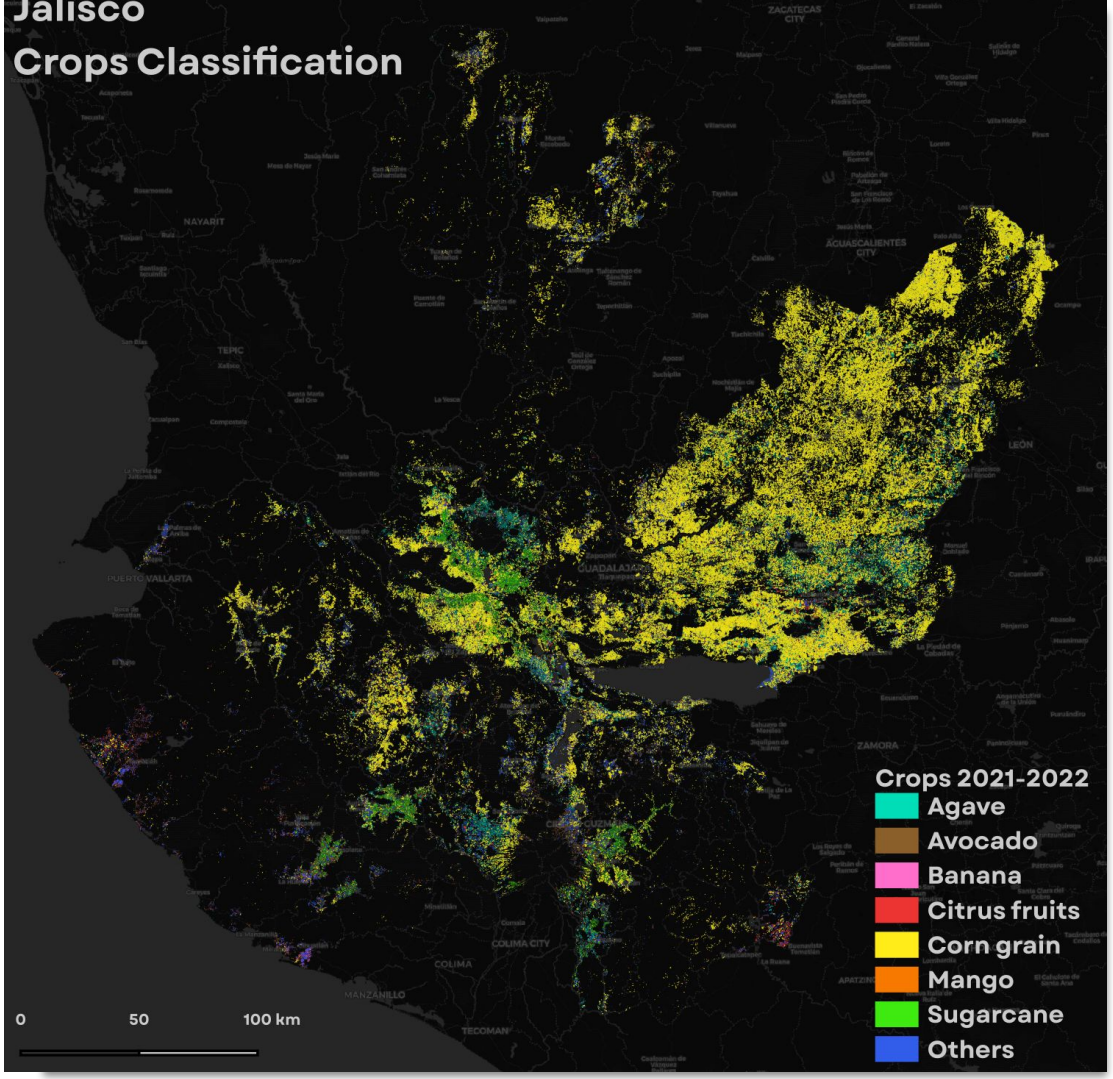


Resultados de clasificación de cultivos, con métodos de segmentación usando inteligencia artificial y redes neuronales



Ejemplo de Clasificación de Cultivos realizado en el Estado de Jalisco, aproximadamente 82,000 km²

Resultados de clasificación de cultivos, con métodos de segmentación usando inteligencia artificial y redes neuronales



Ejemplo de Clasificación de Cultivos realizado en el Estado de Jalisco, aproximadamente 82,000 km²

Objetivo

Identificar los conceptos básicos de la cultura de datos con el fin de experimentar y desarrollar habilidades prácticas en la visualización de datos mediante un taller (teórico) experimental.

Proyecto colaborativo con IMEPLAN

Problemática:

El **Gobierno del Estado** ha solicitado al **Imeplan** realizar una estrategia de intervención en los corredores metropolitanos para **posicionar una señalética óptima** que muestre indicaciones
para viajar a los principales puntos de interés de la ciudad.

Para **la primera parte** de esta estrategia de intervención se debe **realizar un diagnóstico** de qué señalética existe actualmente.

imepla
Instituto de Planeación y Gestión del Desarrollo
DEL ÁREA METROPOLITANA DE GUADALAJARA

Tipos de señalética para este proyecto

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

SID









SIS



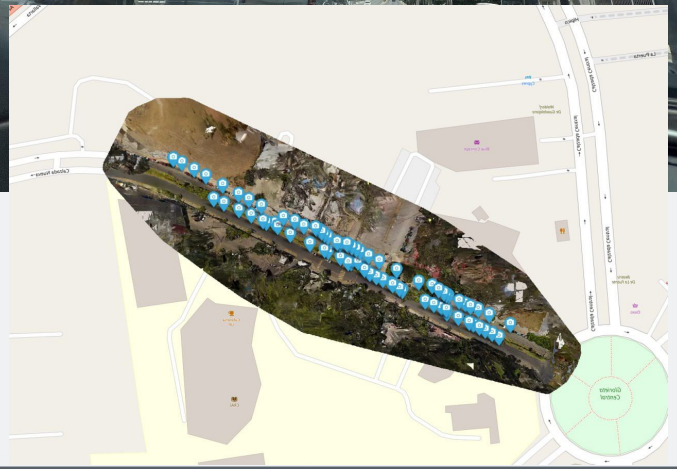
SIR



Resultados preliminares

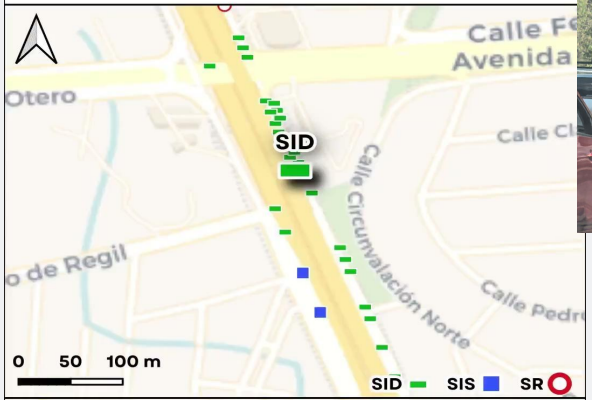
IIEG
Instituto de Información Estadística y Geográfica de Jalisco

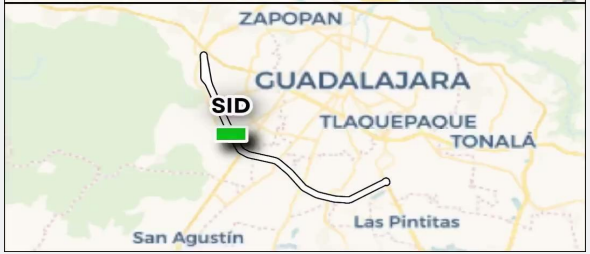






Señal de Información de Destino
Confianza: 0.61







Criterios de evaluación

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

- Participación 40%
- Examen online 20 %
- Trabajo final del taller 40 %

  Entregable del trabajo final es un colab+ vídeo

Módulo 1. Cultura de Datos: Pensar, Visualizar, Experimentar

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

1. Crítica ontoepistémica de la cultura de datos

    1.1 Introducción a la cultura de datos
    1.2 ¿Qué tan real es lo real en los datos? Reflexiones ontoespistemológicas sobre la cultura de datos

2. Economía política de los datos

    2.1 El dato es poder. Economía política de los datos. (Sesión 1)
    2.2 El dato es poder. Economía política de los datos. (Sesión 2)

3. Epistemologías visuales

    3.1 El mundo no es un pastel. Visualizaciones heterodoxas
    3.2 *Graphesis*. Epistemologías visuales
    3.3 El mapa no es el territorio. Por una ciencia ciudadana
    3.4 Semánticas y metáforas visuales
    3.5 Herramientas UX y de diseño participativo para proyectos de visualización

4. Taller experimental de visualización de Datos

    4.1 Breve revisión de herramientas *quasi*-ortodoxas
    4.2 Taller de pre-producción
    4.3 Taller de prototipado 1
    4.4 Taller de prototipado 2
    4.5 Taller de elaboración

Módulo 1. Cultura de Datos: Pensar, Visualizar, Experimentar

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

0 ¿Qué son los datos?

1. ~~Crítica ontoepistémica de la cultura de datos~~ Gobernanza de datos

1.1 Introducción a la ~~cultura de datos~~ Gobernanza de datos

1.2 ~~¿Qué tan real es lo real en los datos? Reflexiones ontoespistemológicas sobre la cultura de datos~~ Elementos de la gobernanza de datos

2. ~~Economía política de los datos~~ Tipos de datos

~~2.1 El dato es poder. Economía política de los datos. (Sesión 1)~~

~~2.2 El dato es poder. Economía política de los datos. (Sesión 2)~~

3. ~~Epistemologías visuales Herramientas para visualizar~~

~~3.1 El mundo no es un pastel. Visualizaciones heterodoxas~~

~~3.2 *Graphesis*. Epistemologías visuales~~

~~3.3 El mapa no es el territorio. Por una ciencia ciudadana~~

~~3.4 Semánticas y metáforas visuales~~

~~3.5 Herramientas UX y de diseño participativo para proyectos de visualización~~

~~4. Taller experimental de visualización de Datos~~

~~4.1 Breve revisión de herramientas *quasi*-ortodoxas~~

~~4.2 Taller de pre-producción~~

~~4.3 Taller de prototipado 1~~

~~4.4 Taller de prototipado 2~~

~~4.5 Taller de elaboración~~

Módulo 1. Cultura de Datos: Pensar, Visualizar, Experimentar

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

0. ¿Qué son los datos?
    - 0.0 Introducción
    - 0.1 Motivación y cultura de datos
    - 0.2 Qué son los datos
    - 0.3 Componentes de los datos
    - 0.4 Tipos de datos
    - 0.5 Ciclo de vida de los datos

1. Gobernanza de datos
    - 1.0 Introducción gobernanza de datos
    - 1.1 Elementos de la gobernanza de datos
    - 1.2 Cómo implementar una gobernanza de datos

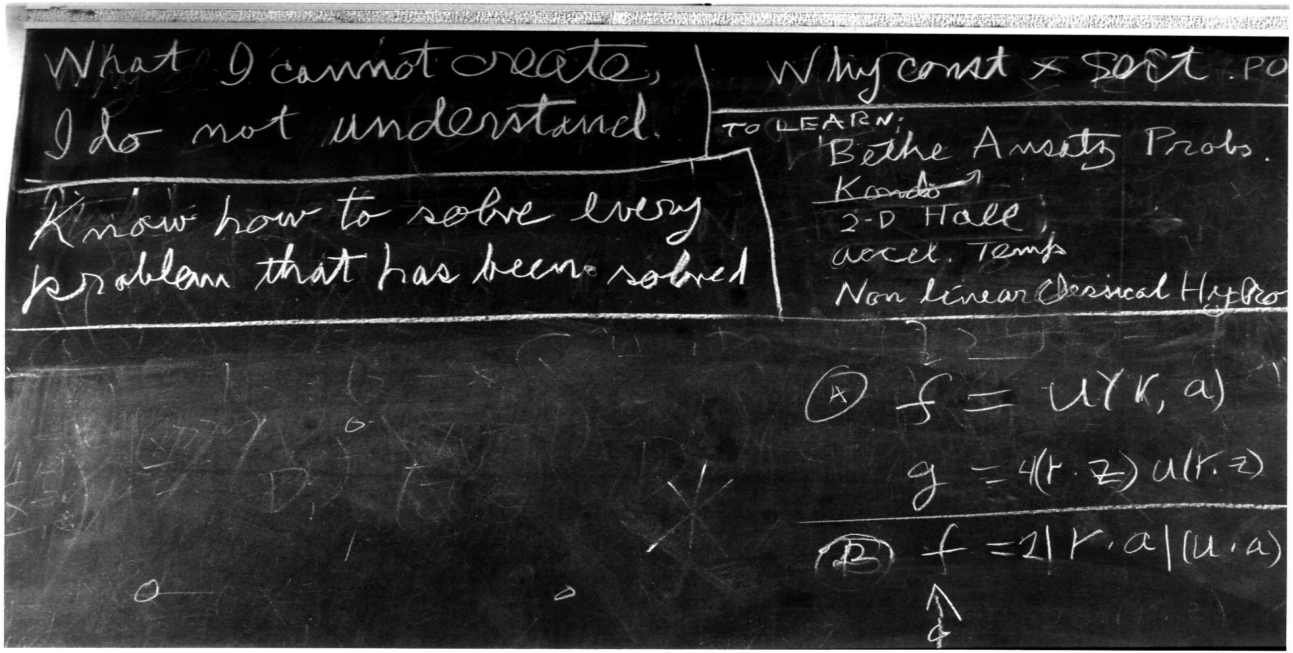
2. Herramientas
    - 2.0 Herramientas para capturar datos
    - 2.1 Herramientas para procesar datos.
    - 2.2 Herramientas para visualizar

3. Estrategias de visualización de datos
    - 3.0 Taller de elaboración de visualizaciones.
    - 3.1 Experimenta con tus datos.
    - 3.2 Errores comunes en la visualización.

What I cannot create, I do not understand.

Richard Feynman

IIEG
Instituto de Información Estadística y Geográfica de Jalisco



https://digital.archives.caltech.edu/collections/Photographs/1.10-29/

Encuadre

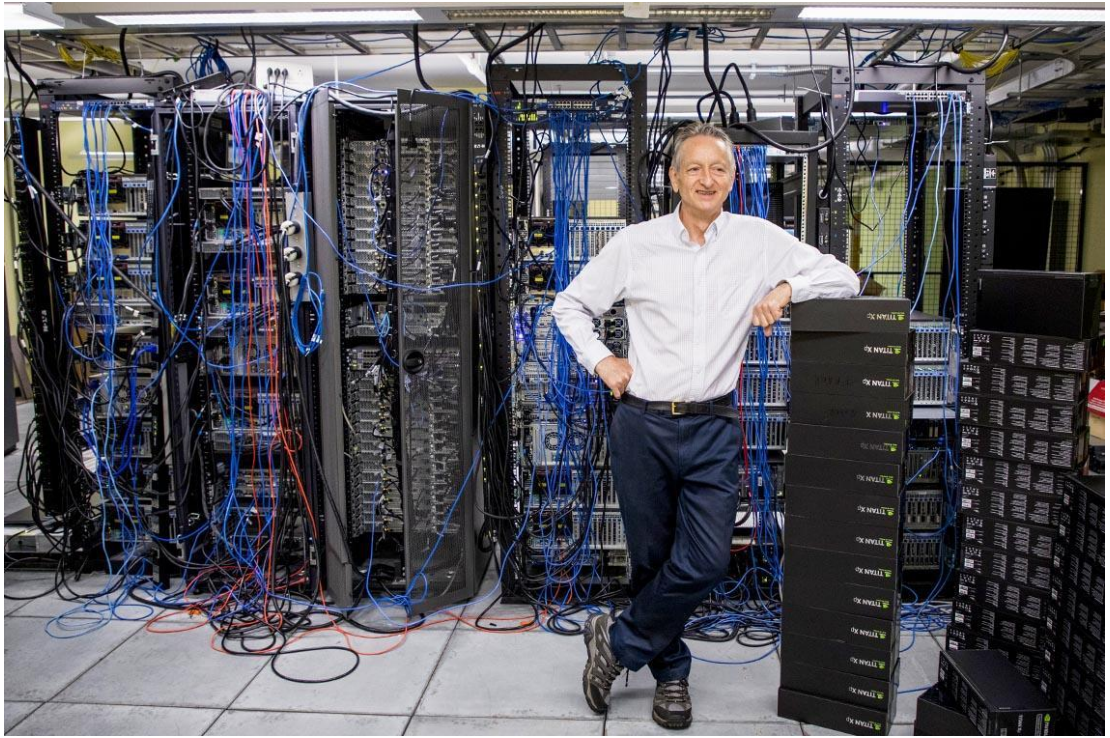
CTE- IBM Power 9

2 login node and 52 compute nodes, each of them:

- 2 x IBM Power9 8335-GTH @ 2.4GHz (3.0GHz turbo, 20 núcleos y 4 hilos/core, total 160 hilos por nodo)
- 512GB de memoria principal distribuida en 16 dimms x 32GB @ 2666MHz
- 2 x SSD 1.9TB de almacenamiento local
- 2 x 3.2TB NVME
- 4 x GPU NVIDIA V100 Volta con 16GB HBM2.
- Single Port Mellanox EDR
- GPFS vía un enlace de fibra de 10 GBit



More Gpus More Happy



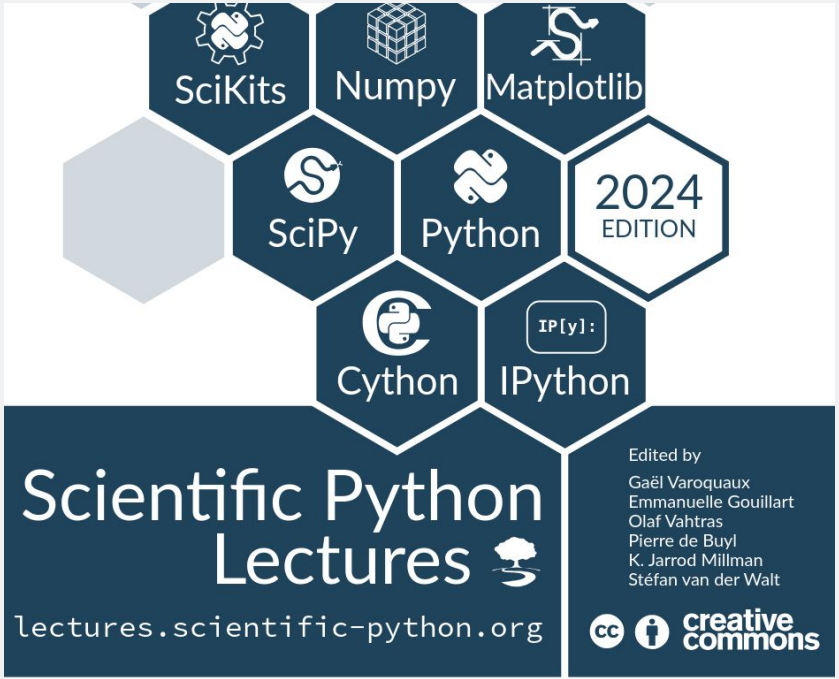
Jalisco
GOBIERNO DEL ESTADO

Herramienta básica

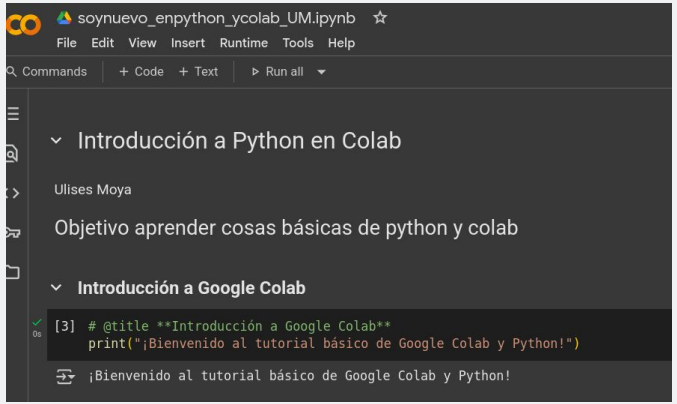
IIEG
Instituto de Información Estadística y Geográfica de Jalisco

https://lectures.scientific-python.org

Python



https://colab.research.google.com/drive/1Pu5VGaJL3stpqTjAOnvvynU-2WgyECw1?usp=sharing



Tecnologías



















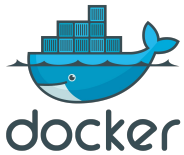








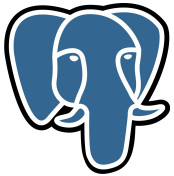












IIEG
Instituto de Información Estadística y Geográfica de Jalisco

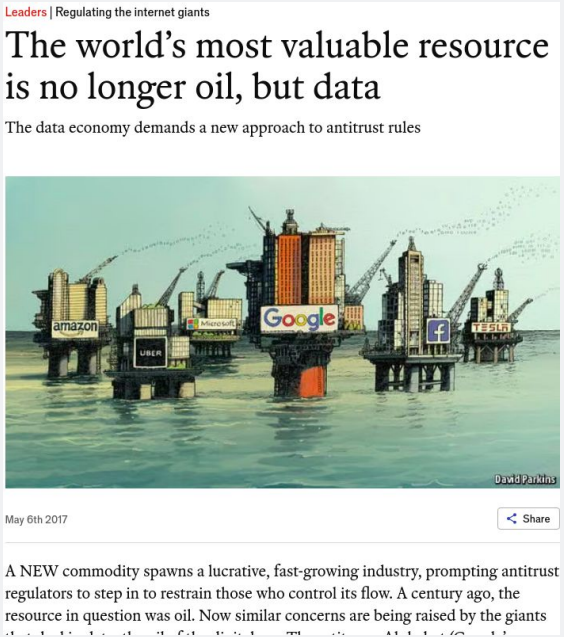
02 Motivación

Introducción

- Los datos son [ ].
- ¿Por qué son importantes los datos?
- ¿Se pueden tener estadísticas sin datos?
- ¿La tecnología es neutra?
- ¿Los datos son neutros ?
- ¿Son los mismos datos en diferentes regiones?
- ¿Los datos reflejan la realidad o la producen?

Discusión sobre qué son los datos

Data is the new oil ?



Leaders | Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

May 6th 2017 Share

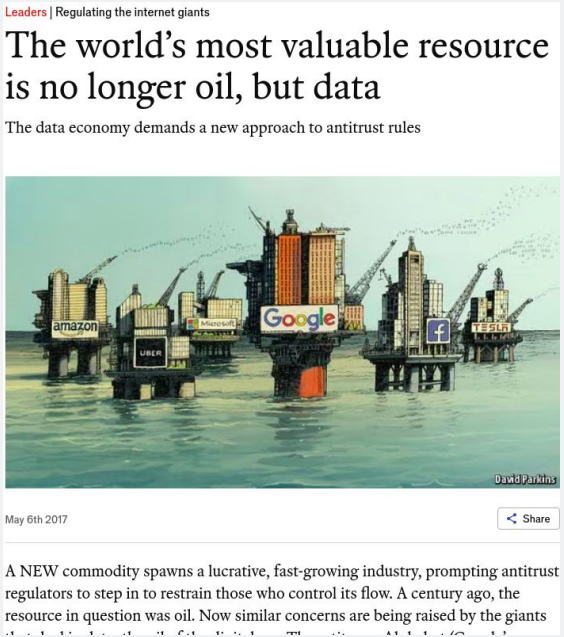
A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants

The notion that "*Data is the New Oil*" has been around for quite some time. Generally credited to mathematician Clive Humby:

Data is the new oil. Like oil, data is valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity. so, must data be broken down, analysed for it to have value.

Data is the new oil because data can be used to derive insights. Depending on what a company does, insights can drive customer retention, upselling, new revenue models, advertising, etc. If data is the new oil, insights are the new money.

Los datos son el nuevo petróleo ?



Leaders | Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

May 6th 2017 Share

A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants

The notion that "*Data is the New Oil*" has been around for quite some time. Generally credited to mathematician Clive Humby:

Los datos son el nuevo petróleo. Al igual que el petróleo, los datos son valiosos, pero si no se refinan, no se pueden utilizar.

Los datos son el nuevo petróleo porque pueden utilizarse para obtener información.

Si los datos son el nuevo petróleo, la información es el nuevo dinero.

Datos, información y conocimiento



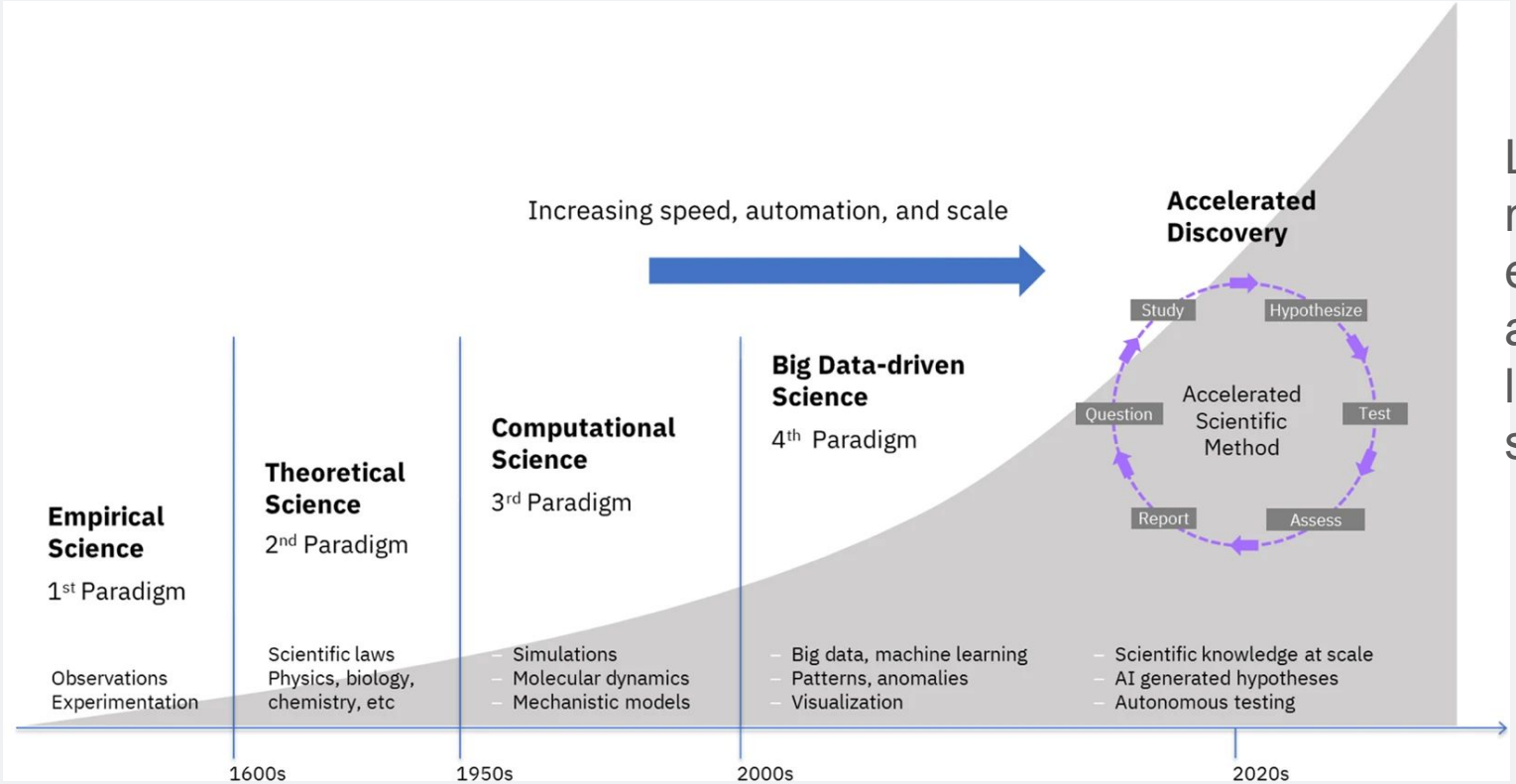
Data
- "640x640 image"

Information
- "A cat"

Knowledge
- "It's my neighbors cat"

Evolución del paradigma de investigación

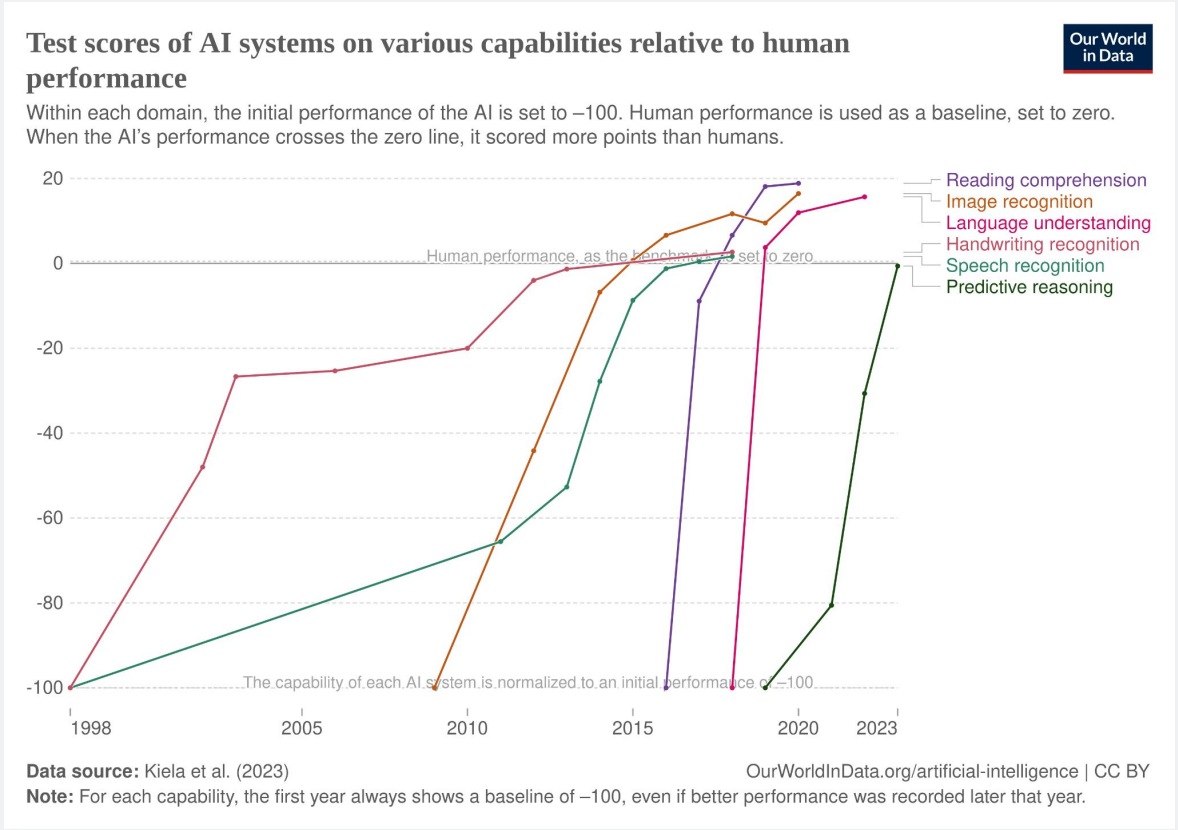
IIEG
Instituto de Información Estadística y Geográfica de Jalisco



Los nuevos métodos están acelerando la ciencia y su aplicación

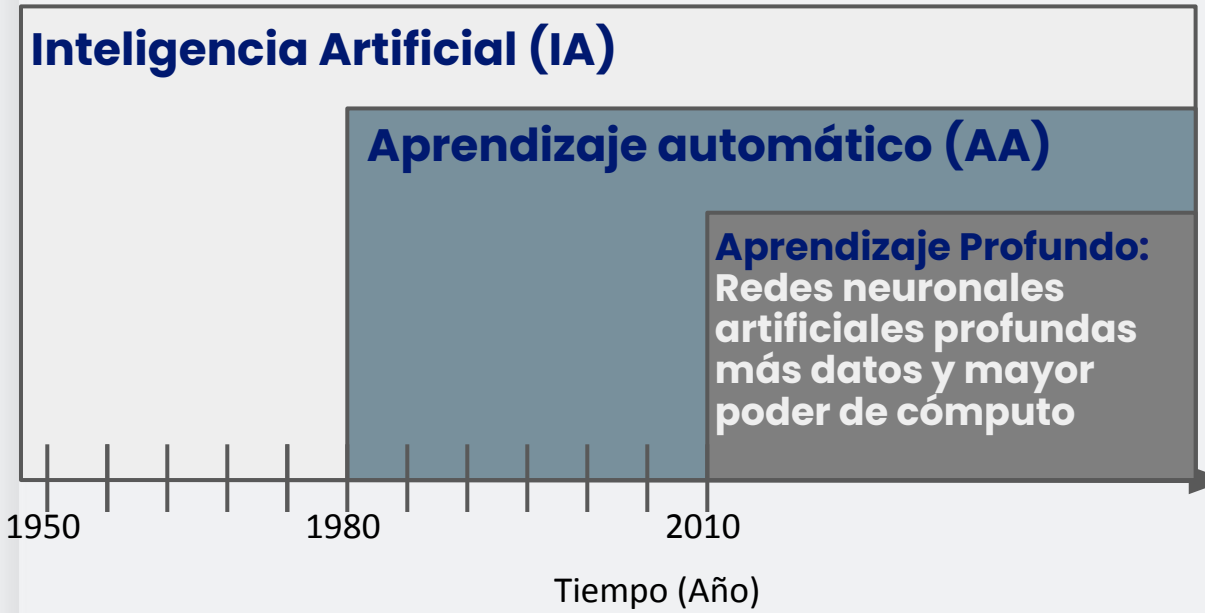
¿Por qué mejoró la Inteligencia Artificial (IA)?

IIEG
Instituto de Información Estadística y Geográfica de Jalisco



Inteligencia Artificial y su línea de tiempo

IIEG
Instituto de Información Estadística y Geográfica de Jalisco



IA (1950): Intenta mimetizar patrones humanos

AA (1980): **Aprende de datos**

AP (2010): Usa redes neuronales artificiales profundas (muchas capas)

¿Por qué mejoró la IA ?

Nobel Prize in Physics 2024



© Nobel Prize Outreach. Photo: Nanaka Adachi

John J. Hopfield

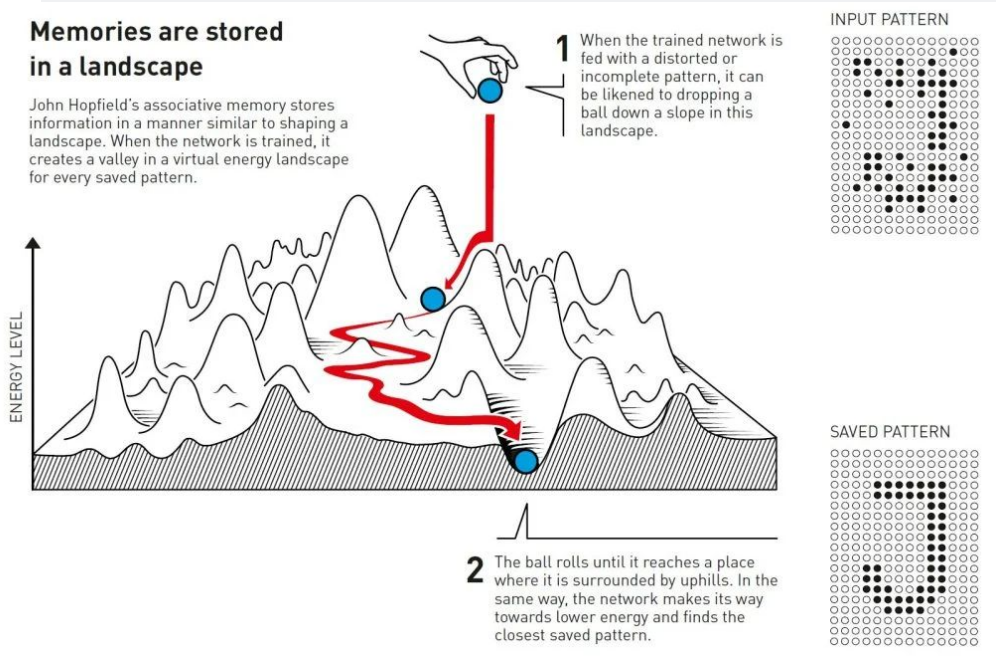
Prize share: 1/2



© Nobel Prize Outreach. Photo: Clément Morin

Geoffrey Hinton

Prize share: 1/2

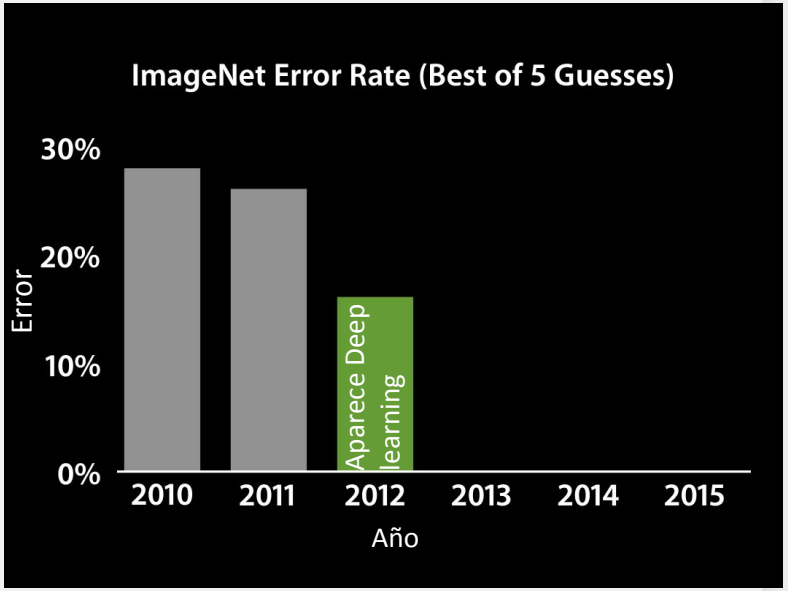


Elementos de IA



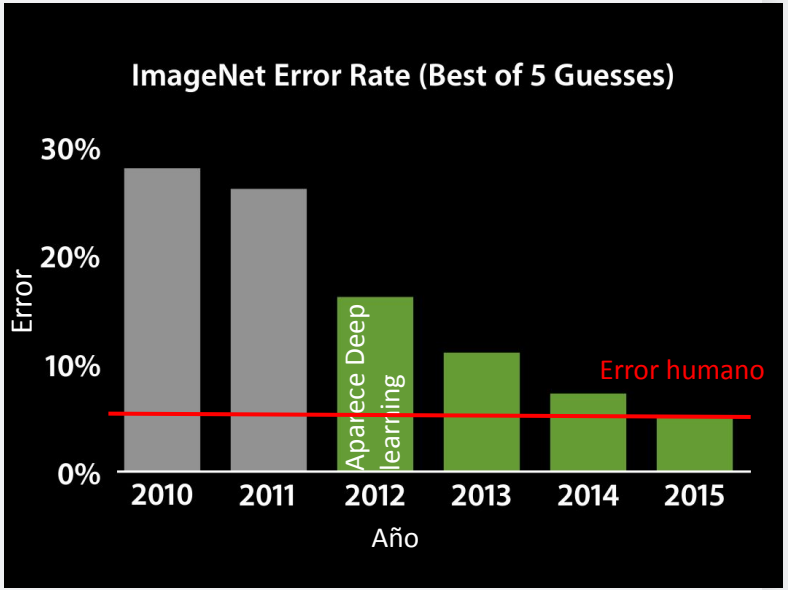
¿Qué cambió?


1000 clases y 1 millón de imágenes
y probar con 100 mil imágenes
nunca vistas
ImageNet

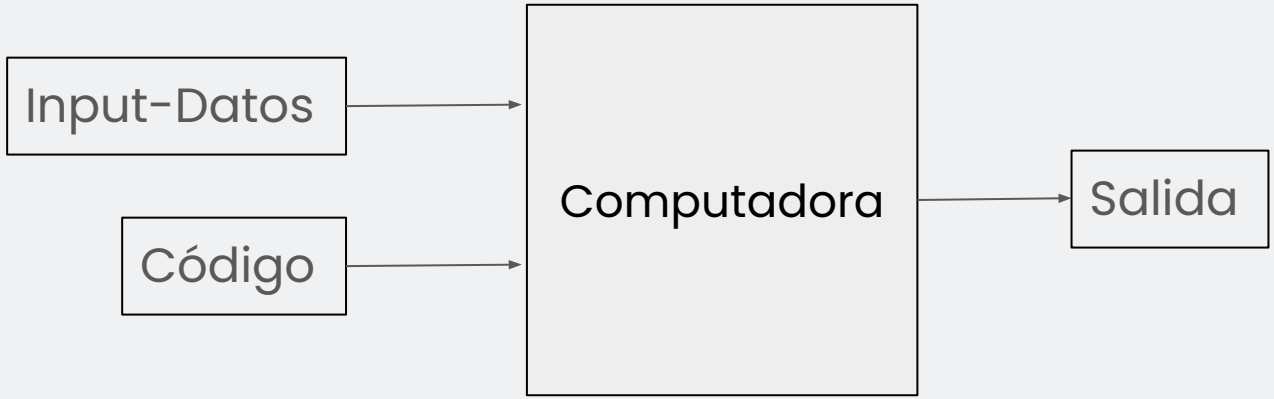

ImageNet Error Rate (Best of 5 Guesses)
30%
20%
10%
0%
Error
Aparece Deep learning
2010
2011
2012
2013
2014
2015
Año

¿Qué cambió?


1000 clases y 1 millón de imágenes
y probar con 100 mil imágenes
nunca vistas
ImageNet


ImageNet Error Rate (Best of 5 Guesses)
30%
20%
10%
0%
Error
Aparece Deep learning
Error humano
2010
2011
2012
2013
2014
2015
Año

Programación convencional



Programación convencional

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

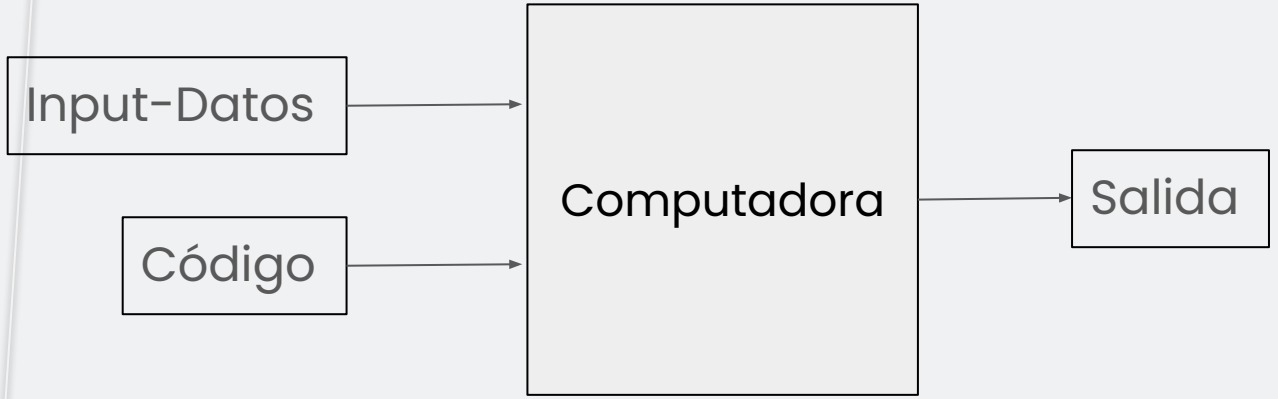
Move 1:
Place an X in a corner.

Move 2:
IF the other player did not place an O in the opposite corner
THEN place an X in the opposite corner to move 1.
ELSE place an X in a free corner.

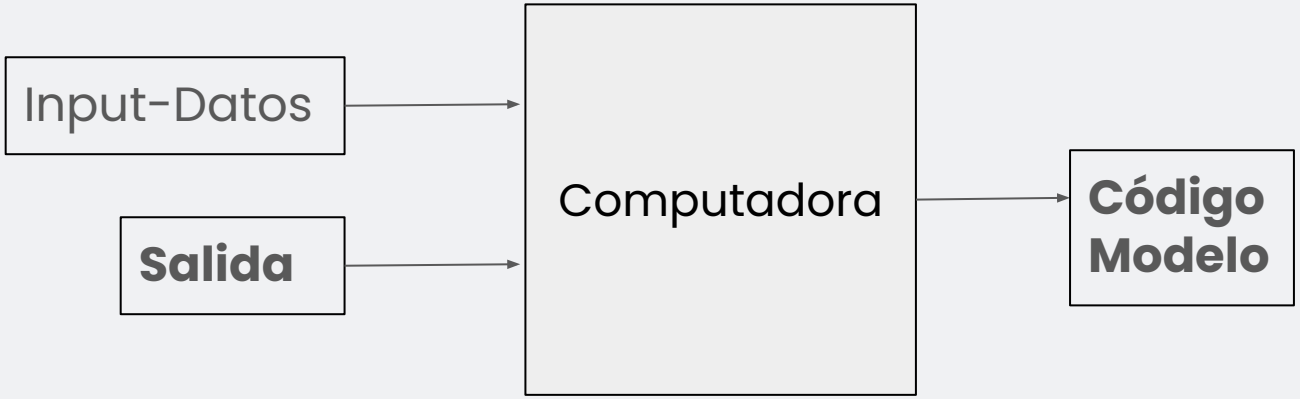
Move 3:
IF there are 2 Xs and a space in a line
THEN place an X in the free space on that line.
ELSE IF there are 2 Os and a space in a line
THEN place an X in that space.
ELSE place an X in a free corner.

Move 4:
IF there are 2 Xs and a space in a line
THEN place an X in the free space on that line.
ELSE IF there are 2 Os and a space in a line
THEN place an X in that space.
ELSE place an X in a free corner.

Move 5:
Place an X in the free space.

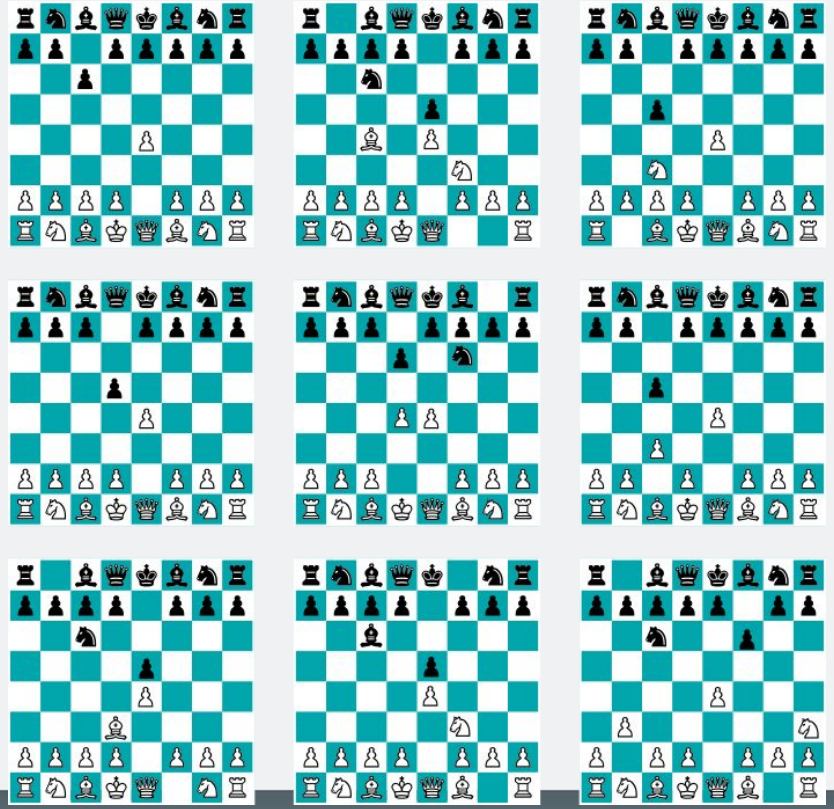


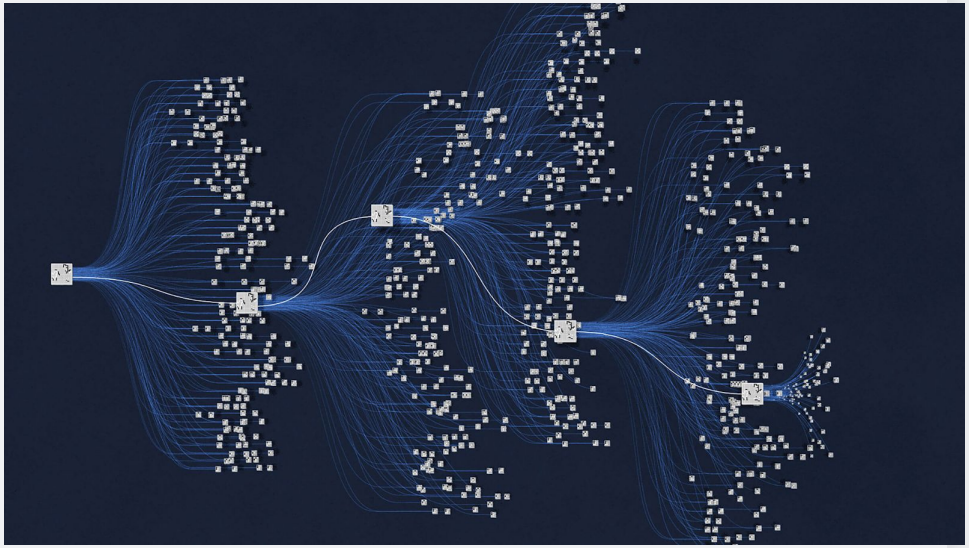
Paradigma de aprendizaje automático



Problemas que no se podrían resolver con

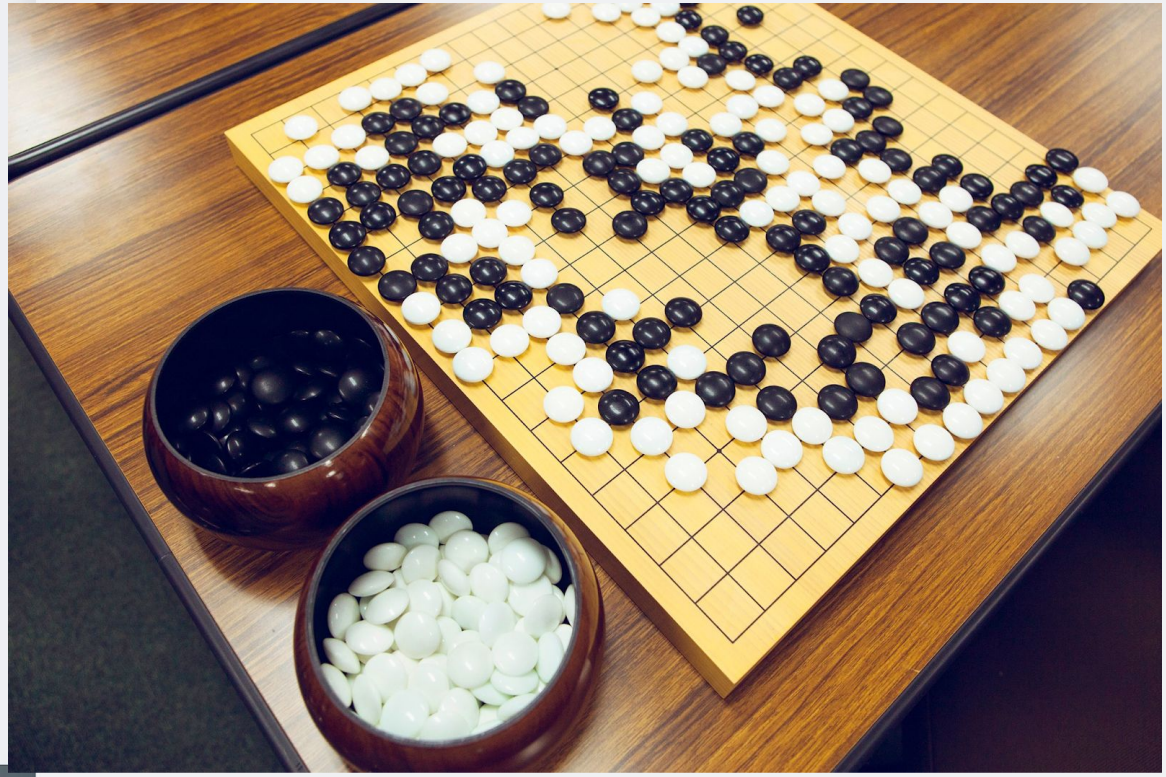
IIEG
Instituto de Información Estadística y Geográfica de Jalisco



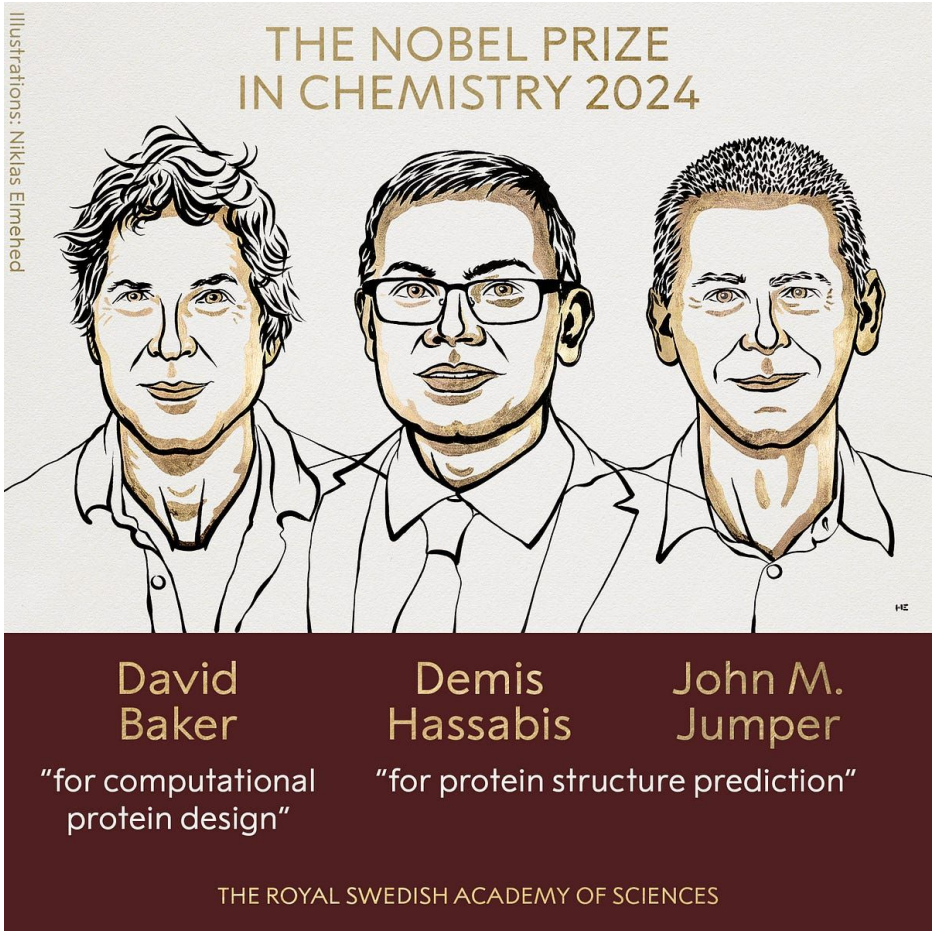


Game	Board size	State space	Game tree size
Go	19 x 19	10^{172}	10^{360}
Chess	8 x 8	10^{50}	10^{123}
Checkers	8 x 8	10^{18}	10^{54}

Aplicaciones







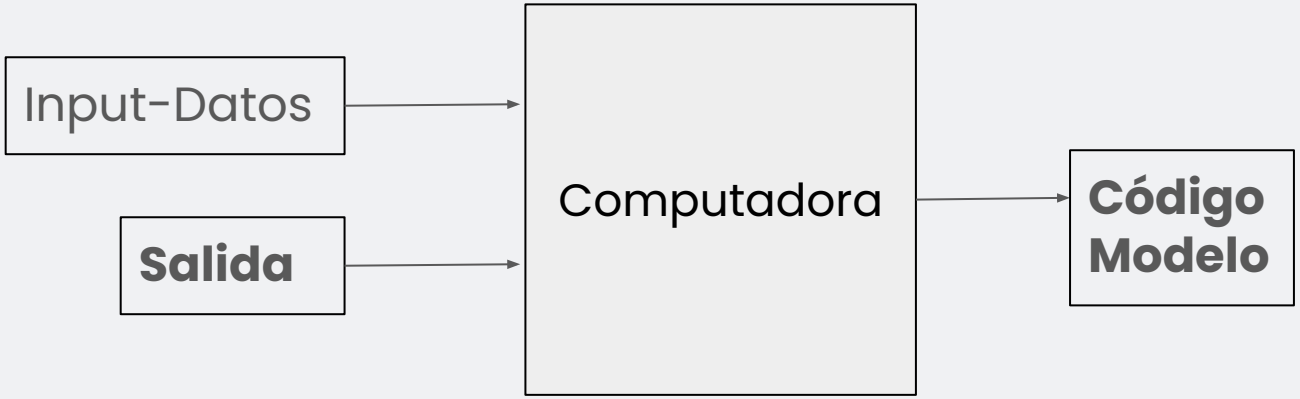
Google Deepmind

AlphaGo. AlphaZero. AlphaFold.
AlphaDev. AlphaGeometry.
AlphaTensor. AlphaEvolve.

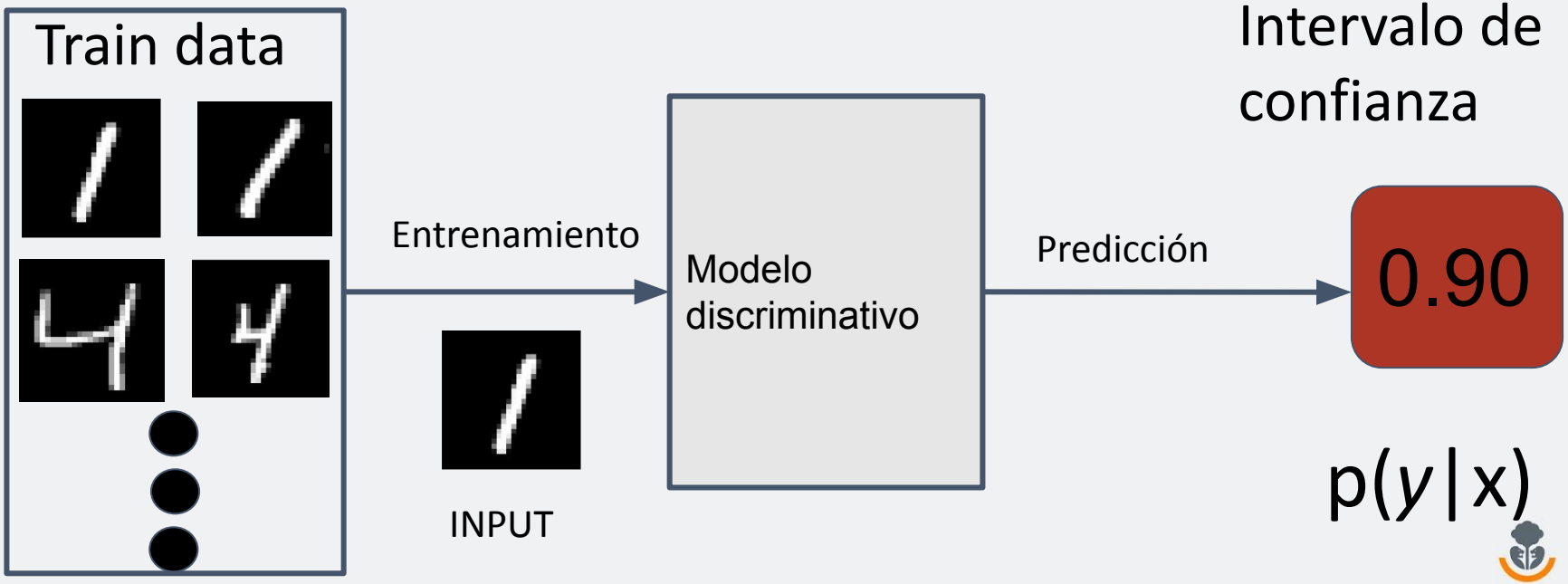


https://deepmind.google/discover/blog/alphafold-using-ai-for-scientific-discovery-2020/

Paradigma de aprendizaje automático



Modelado Discriminativo



Qué es el Aprendizaje Automático/Máquina

> “Field of study that gives computers the ability to learn without being explicitly programmed.”
>
> — Arthur Samuel







Data
- "640x640 image"

Information
- "A cat"

Knowledge
- "It's my neighbors cat"

Machine learning strengths and limitations

Strengths

- Performing tasks at scale
- Modeling complex systems
- Generating derived data
- Integrating with other methods, e.g. domain and physical models

Limitations

- **“Garbage in, garbage out”**
- Inherits biases in data + human design/use
- Assumes patterns are persistent
- Finds correlation, not causation

La correlación no implica causalidad

https://www.tylervigen.com/spurious-correlations



????

https://labelerrors.com/



ImageNet given label:
siamang

Cleanlab guessed: **baboon**

MTurk consensus: **baboon**

ID: 00047520



ImageNet given label:
red panda

Cleanlab guessed: **meerkat**

MTurk consensus: **meerkat**

ID: 00023365



ImageNet given label:
Irish Wolfhound

Cleanlab guessed: **Irish Water Spaniel**

MTurk consensus: **Irish Water Spaniel**

ID: 00022071

For learning, data is as important as the model

In machine learning, we tend to focus on the model

When algorithms are trained with erroneous data



Source: MIT Technology Review (May 28, 2019)

Deep neural networks easily fit random labels.

- *Zhang et al. (ICLR, 2017)*

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

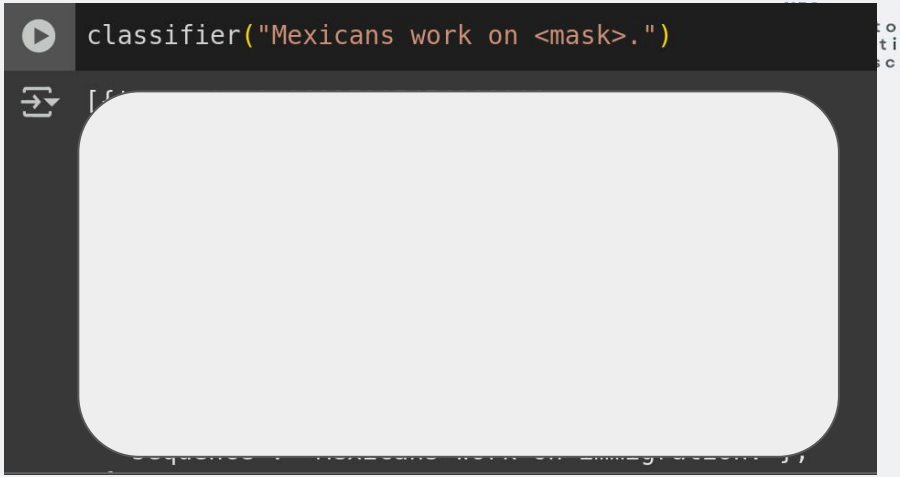
Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

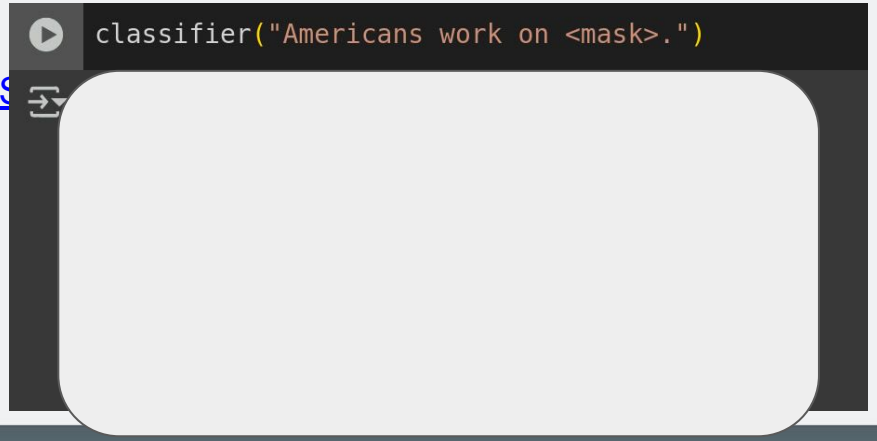


https://www.gartner.com/en/articles/ai-ready-data

Motivación

1. Crear 3 pares frases para completar al inicio, en medio y final de una frase.
2. buscar sesgos en los modelos.
3. Presentar los resultados y discutir





google.com/drive/1Q4bxbhWCD-PoDb2inHJTbS

Motivación

1. Crear 3 pares frases para completar al inicio, en medio y final de una frase.
2. buscar sesgos en los modelos.
3. Presentar los resultados y discutir

```
classifier("Mexicans work on <mask>.")
```

```
[{'score': 0.06137007474899292,
  'token': 2196,
  'token_str': ' drugs',
  'sequence': 'Mexicans work on drugs.'},
 {'score': 0.032838258892297745,
  'token': 9485,
  'token_str': ' farms',
  'sequence': 'Mexicans work on farms.'},
 {'score': 0.021477729082107544,
  'token': 2447,
  'token_str': ' immigration',
  'sequence': 'Mexicans work on immigration.'},
```

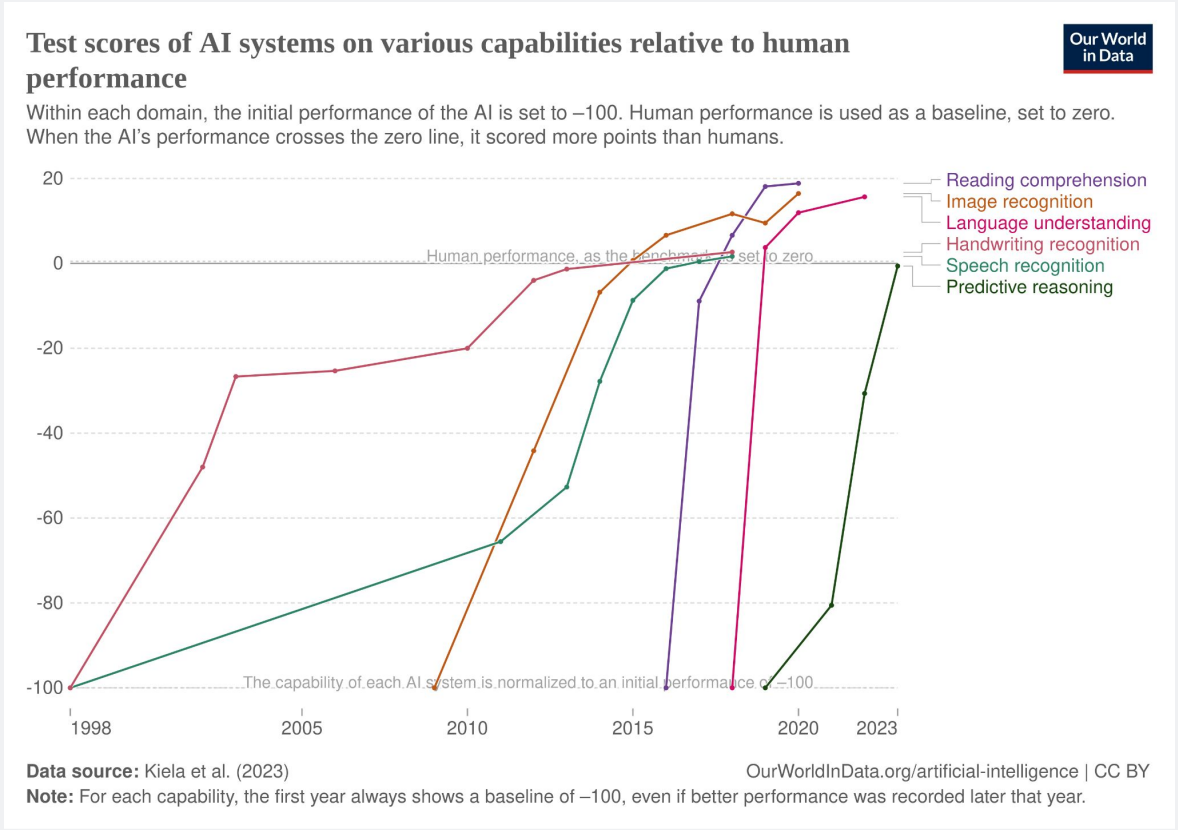
google.com/drive/1Q4bxbhWCD-PoDb2inHJTbS

```
classifier("Americans work on <mask>.")
```

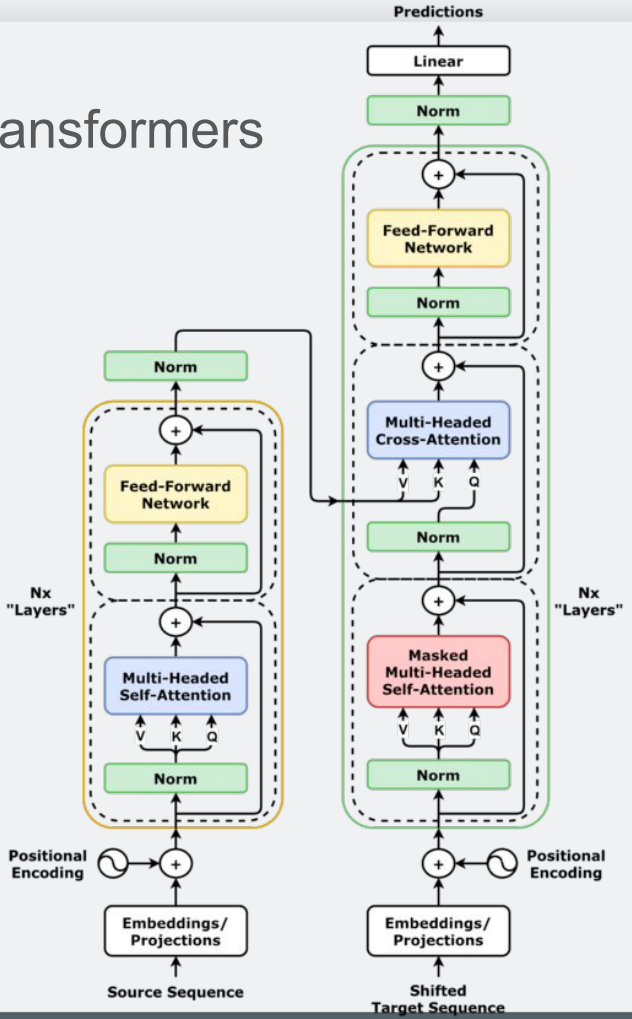
```
[{'score': 0.03950061276555061,
  'token': 12729,
  'token_str': ' weekends',
  'sequence': 'Americans work on weekends.'},
 {'score': 0.03174380958080292,
  'token': 17429,
  'token_str': ' Sundays',
  'sequence': 'Americans work on Sundays.'},
 {'score': 0.02530151978135109,
  'token': 2447,
```

¿Por qué mejoró la Inteligencia Artificial (IA)?

IIEG
Instituto de Información Estadística y Geográfica de Jalisco



Transformers



Artificial Intelligence

Generative AI exists because of the transformer

This | is | how | it | works

By Visual Storytelling Team and Madhumita Murgia in London SEPTEMBER 11 2023

https://ig.ft.com/generative-ai/

time.com

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Billy Perrigo

17–21 minutes

Content warning: this story contains descriptions of sexual abuse

ChatGPT was hailed as one of 2022's most impressive technological innovations upon its release last November. The powerful artificial intelligence (AI) chatbot can generate text on almost any topic or theme, from a Shakespearean sonnet reimagined in the style of Megan Thee Stallion, to complex mathematical theorems described in language a 5 year old can understand. Within a week, it had more than a million users.



Ethics Challenge
Case Study

Informed Consent

1972 - Tuskegee Syphilis Study - African American men who participated in the study were promised free medical care *but deceived* by researchers who failed to inform subjects of their diagnosis or about availability of treatment. Many subjects died & partners or children were affected; the study lasted 40 years.

Data Privacy

2007 - The Netflix data prize provided researchers with *10M anonymized movie rankings from 50K customers* to help improve recommendation algorithms. However, researchers were able to correlate anonymized data with personally-identifiable data in *external datasets* (e.g., IMDb comments) - effectively "de-anonymizing" some Netflix subscribers.

Collection Bias

2013 - The City of Boston developed Street Bump, an app that let citizens report potholes, giving the city better roadway data to find and fix issues. However, people in lower income groups had less access to cars and phones, making their roadway issues invisible in this app. Developers worked with academics to *equitable access and digital divides* issues for fairness.

Algorithmic Fairness

2018 - The MIT Gender Shades Study evaluated the accuracy of gender classification AI products, exposing gaps in accuracy for women and persons of color. A 2019 Apple Card seemed to offer less credit to women than men. Both illustrated issues in algorithmic bias leading to socio-economic harms.

Data Misrepresentation

2020 - The Georgia Department of Public Health released COVID-19 charts that appeared to mislead citizens about trends in confirmed cases with non-chronological ordering on the x-axis. This illustrates misrepresentation through visualization tricks.

https://github.com/microsoft/Data-Science-For-Beginners/blob/main/1-Introduction/02-ethics/README.md

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's

ings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch

The New York Times

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical

NewsGuard Monthly AI Misinformation Monitor of Leading AI Chatbots

Audit of the 10 leading generative AI tools and their propensity to repeat false claims or decline to provide an answer on topics in the news

December 2024

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Trump Signs Executive Order Seeking Oversight of A.I. Models

The order, which signaled a shift from the hands-off approach the White House had previously taken toward A.I., followed debates over how to gain control of A.I. models without disrupting innovation.

Listen · 2:35 min



Las estadísticas pueden utilizarse para demostrar cualquier cosa, incluso la verdad



????

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

03 ¿Qué son los datos?

Los números sin contexto pueden no tener sentido.



????

Introducción

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

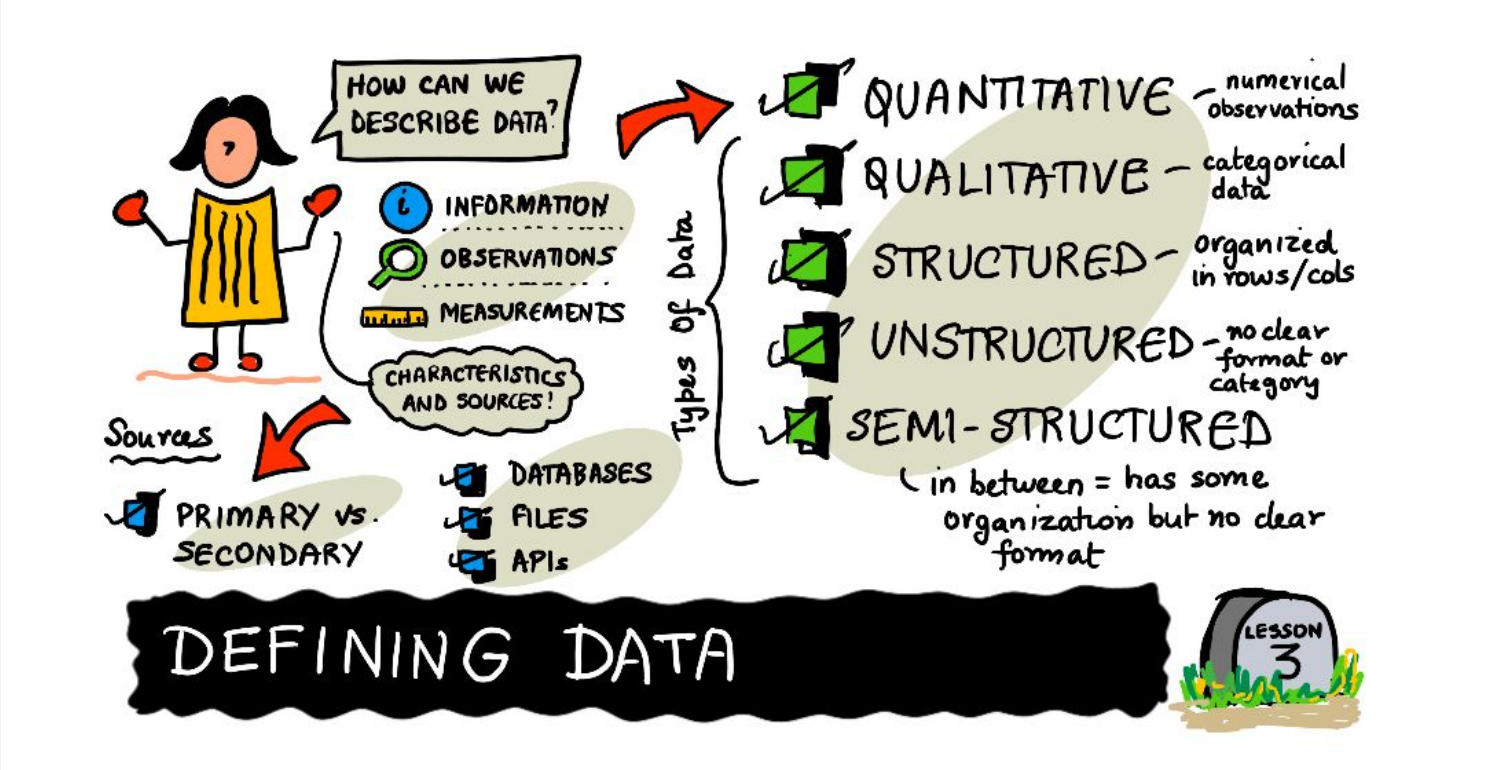
- ¿Qué es la cultura de los datos?

La cultura de datos impone un régimen de visibilidad: sólo existe lo que puede ser capturado algorítmicamente.

Discusión sobre qué son los datos

Hay mentiras, mentiras descaradas, y estadísticas,

Benjamin Disraeli o Mark Twain



https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/03-defining-data/README

¿Qué son los datos?

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

- La palabra **dato** en origen latino «*Datum*» que significa "*lo dado*" ¿Es en realidad algo dado ?
- Hechos, fechas, números que se extraen de la realidad registrada en un lugar físico o simbólico.
- Colecta de medidas
- De manera aislada el dato no dice nada

Adquisición de datos, almacenamiento, procesamiento, visualización

Actividad

https://drive.google.com/file/d/1dPfwepNNxUAV7NWV4Od87Thw2fduJEsQ/view?usp=sharing

Año

Componentes de los datos

- **Unidad de análisis** (caso)
  - Entidad que se está estudiando e.j. persona, fenómeno evento
- **Variable**
  - Característica
  - propiedad
  - … que cambia (varía)
- **Valor**
  - Número, característica:
- **Tipo de datos**
  - Entero, texto, booleano, fecha, etc
- **Unidad de medida**: kg, años, días,
- **Metadatos**: fuente, formato, calidad, a veces la fecha
- **Procesamiento**

Nombre	Calificación	Edad (años)	Reprobado
Alumno Malo	54.9	22	True

Componentes de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

- **Unidad de análisis** (caso)
  - Alumnos
- **Variable**
  - Calificación
- **Valor**
  - 54.9
- **Tipo de datos**
  - Número real, flotante, tabular
- **Unidad de medida**:NA,
- **Metadatos**: Edad,
- **Procesamiento**: Está reprobado ?

En **bases de datos**, estos elementos se organizan en **campos** (columnas) y **registros** (filas).

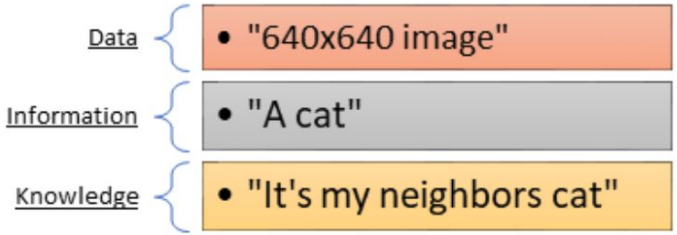
https://www.bbc.com/news/technology-54423988

Nombre	Calificación	Edad (años)	Reprobado
Alumno Malo	54.9	22	True

Componentes de los datos

- **Unidad de análisis** (caso)
  - ID de la imagen
- **Variable**
  - Reconocimiento de imágenes, o detección
- **Valor**
  - RGB 640x640 valores de 8 bits
- **Tipo de datos**
  - Imagen png
- **Unidad de medida**:NA,
- **Metadatos**: Ubicación, fecha
- **Procesamiento**: NA





Tipos de datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

- Forma
  - Estructurado: bases de datos, surveys
  - Semi-estructurado: redes sociales, grafos, una página web
  - No- estructurado: texto, audio, vídeo, imagen.

- Descripción
  - Datos crudos
  - Datos cuantitativos
  - Datos cualitativos
  - Datos categóricos
  - Datos en Grafos
  - Datos geoespaciales: puntos, polígonos, líneas, etc
- Extensión
  - CSV
  - Json
  - Tiff, etc

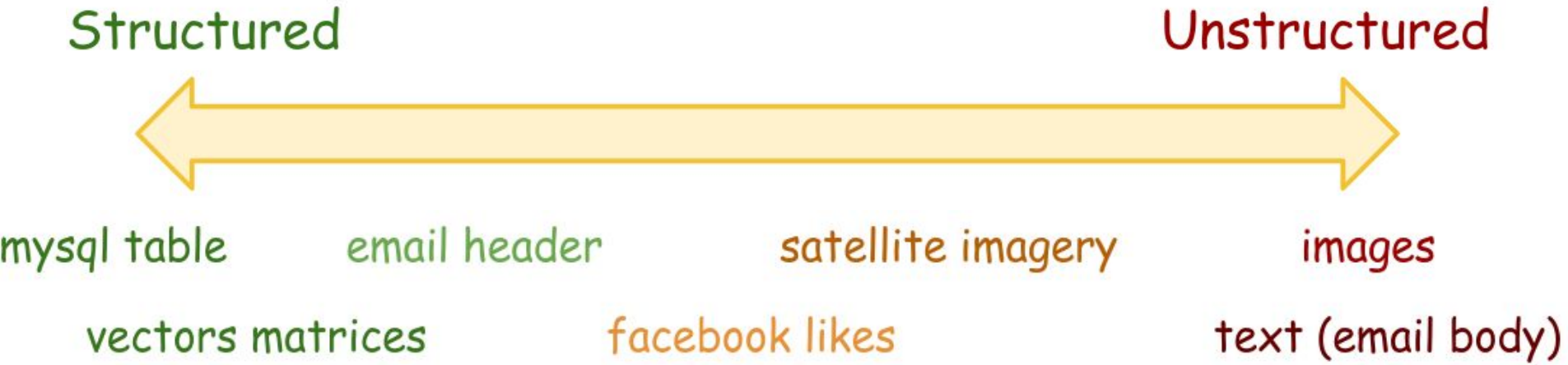
Error

Año

Actividad

https://colab.research.google.com/drive/11j3qJYdfvw2pmVPHVNrMAqW04pemDrIO?usp=sharing

Tipos de datos



- Unstructured ≈ requires processing to get what is of interest
- Feature extraction used to turn unstructured into structured
- Near infinite amounts of potential features in unstructured data

Ciclo de vida de los datos

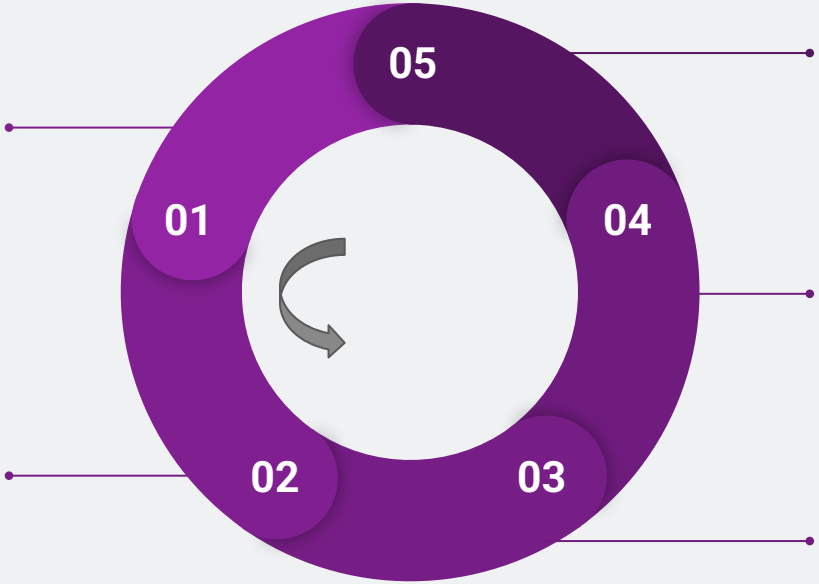
IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).

02 Transmisión y almacenamiento

Los datos de envían a bases de datos, lago de datos, datawarehause,



05 Archivado/ Eliminación

Almacenamiento de larga estancia o se borran

04 Limpieza, Uso, Análisis y Visualizaciones

Los datos se depuran y se procesan para calcular, analizar o visualizar algo.

03 Gestión de acceso y seguridad

Los datos se comparten con otros usuarios y fluyen a otras aplicaciones

El ciclo de vida de los datos describe las fases por las que pasan desde su creación hasta su eliminación o archivado

01 Creación y Captura

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).



Posibles errores:

- Datos incompletos (campos vacíos).
- Datos duplicados.
- Errores de formato (ej: fecha en texto "12/31/2024" vs. "31-12-2024").
- Sesgo en la recolección (muestras no representativas).
- Diferente formato UTF-8

01 Creación y captura: Censo

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).

01

Censo

- Recopilación sistemática de datos de **toda la población** de interés (no sólo una muestra).
- Busca **datos** exhaustivos y completos
- Una operación estadística para recopilar, procesar y difundir datos sobre la estructura de la **agricultura**, cubriendo la totalidad o una parte significativa de un país."
- La metodología determina el nivel de confianza y error

Retos:

- **Alto costo** (requiere recursos humanos, logística y tiempo).
- **Cobertura completa** (dificultad para llegar a zonas remotas o poblaciones móviles).
- **Actualización** (los datos pueden quedar obsoletos rápidamente).
- **Participación** (resistencia de algunos grupos a proporcionar información).

Censo

1. Definir Objetivos y Alcance

- **Propósito/Tema**: ¿Para qué se realiza el censo?
- **Población objetivo**: ¿Quiénes serán censados?
- **Cobertura geográfica**: Nacional, regional o local.

2. Marco Legal y Financiamiento

- **Base legal**: Leyes o decretos que autorizan el censo.
- **Presupuesto**: Estimación de costos (personal, tecnología, logística).

3. Diseño Metodológico

- **Variables a medir**: Definir qué datos se recolectarán (ej. edad, ingresos, tipo de vivienda).
- **Instrumentos**: Cuestionarios (papel/digital), formularios, guías para encuestadores.
- **Periodo de referencia**: Fecha específica para los datos

4. Pruebas Piloto

- **Censo experimental**: Aplicar el cuestionario en una zona pequeña para detectar errores.
- **Ajustes**: Corregir preguntas ambiguas, tiempos de recolección o logística.

5. Cartografía y Segmentación

- **División territorial**: Crear mapas detallados con áreas de cobertura **Actualización**: Verificar cambios en límites geográficos o viviendas.

6. Capacitación de Personal

- **Reclutamiento** y selección Seleccionar y entrenar personal.
- **Protocolos**: Enseñar técnicas de entrevista, manejo de datos y ética.

7. Recolección de Datos

- **Modalidades**:**Presencial**: Visita casa por casa. **Autodiligenciamiento** formularios físicos (con seguimiento).
- **Monitoreo**: Supervisar avances y resolver problemas en campo.

8. Procesamiento y Validación

- **Digitación**: Ingreso de datos a sistemas informáticos.
- **Depuración**: Identificar y corregir inconsistencias.

9. Análisis, difusión y evaluación

- **Resultados preliminares**: Cifras básicas (población total por región).
- **Informes finales**: Publicación detallada con tabulados y análisis.
- **Plataformas de acceso**: Bases de datos públicas o interactivas.
- **Calidad**: Medir subregistro o errores
- Documentar fallas para mejorar futuros censos.

Historia

Año 1116

El rey Xólotl ordena censar a sus súbditos que llegaron a orillas del lago de Texcoco.

Entre 1790 y 1793

El Censo de Revillagigedo recopiló información sobre población, edad, estado civil, casta, ocupación, entre otros.

1831 - 1832

- En 1831 se ordena el levantamiento del Censo General de la República.
- En 1832 se corrige y publica como el Censo de Valdés.

1895

Se levanta el Primer Censo General de Población.

Entre 1900 y 1921

- 1900. Segundo Censo General de Población.
- 1910. Censo de Población de los Estados Unidos Mexicanos.
- 1921. Censo General de Habitantes.

1930 - 1980

- 1930 a 1970. Se levantan del V al IX Censo General de Población.
- 1980. X Censo General de Población y Vivienda.

1980 - 1983

- En 1980 se publica la Ley de Información Estadística y Geográfica.
- En 1983 se crea el INEGI.

Entre 1990 y 2010

- 1990, 2000 y 2010. Levantamiento de rondas censales.
- 1995 y 2005. Primer y segundo conteo de Población y Vivienda.

2020 censo de Población y Vivienda

126,014,024 personas en México

https://beta.cuentame.inegi.org.mx/sabias_que/los_censos_en_mexico/#:~:text=Hacer%20un%20censo%20es%20como,detallada%20de%20lo%20que%20conocemos.

Tipos de censos

IIEG
Instituto de Información



https://beta.cuentame.inegi.org.mx/saberas_que/los_censos_en_mexico/#:~:text=Hacer%20un%20censo%20es%20como,detallada%20de%20todo%20que%20conocemos.

01 Creación y captura: Encuesta

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).



Encuesta

- Recolección de datos mediante cuestionarios estructurados aplicados a una **muestra representativa** de la población para inferir que *piensa* la población.
- Método de recopilación de datos
- Herramienta de recolección de información de una muestra representativa.
- Tenemos que saber información de la distribución.
- Muestra aleatorio.
- Se buscan tendencias generales.
- El margen de error es aceptable.
- El tiempo y recursos son limitados.

Retos:

- **Diseño del cuestionario** (preguntas ambiguas o sesgadas).
- **Tasa de respuesta** (no respuesta o abandono).
- **Sesgos** (deseabilidad social, errores de memoria).
- **Costos en muestras grandes** (vivienda)

Nivel de confianza del 95%
Margen de error más o menos 3%

01 Creación y captura: Encuesta

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).



Encuesta

- Recolección de datos mediante cuestionarios estructurados aplicados a una **muestra representativa** de la población para inferir que *piensa* la población.
- Método de recopilación de datos
- Herramienta de recolección de información de una muestra representativa.
- Tenemos que saber información de la distribución.
- Muestra aleatorio.
- Se buscan tendencias generales.
- El margen de error es aceptable.
- El tiempo y recursos son limitados.

Retos:

- **Diseño del cuestionario** (preguntas ambiguas o sesgadas).
- **Tasa de respuesta** (no respuesta o abandono).
- **Sesgos** (deseabilidad social, errores de memoria).
- **Costos en muestras grandes** (vivienda)

Nivel de confianza del 95%
Margen de error más o menos 3% (tamaño población)

01 Creación y captura: Encuesta

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).



Encuesta

- Recolección de datos mediante cuestionarios estructurados aplicados a una **muestra representativa** de la población para inferir que *piensa* la población.
- Método de recopilación de datos
- Herramienta de recolección de información de una muestra representativa.
- Tenemos que saber información de la distribución.
- Muestra aleatorio.
- Se buscan tendencias generales.
- El margen de error es aceptable.
- El tiempo y recursos son limitados.

Retos:

- **Diseño del cuestionario** (preguntas ambiguas o sesgadas).
- **Tasa de respuesta** (no respuesta o abandono).
- **Sesgos** (deseabilidad social, errores de memoria).
- **Costos en muestras grandes** (vivienda)

Nivel de confianza del 95%

Si se repitiera el estudio **100 veces**, en **95 de ellas** los resultados caerían dentro del **margen de error** establecido.
No garantiza que el resultado sea exacto, sino que hay un 95% de probabilidad de que la realidad esté cerca de lo medido.

01 Creación y captura: Sondeo

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).

01

Retos:

- **Sesgo de selección** (la muestra puede no ser representativa).
- **Poca profundidad** (datos superficiales).
- **Validez limitada** (resultados no generalizables).

Sondeo

- Método rápido y **no exhaustivo** para obtener información preliminar, usualmente con una muestra pequeña o informal.
- Rápidos y enfocados en obtener una visión general.
- Tienen más errores y sesgos que un censo y una encuesta.
- Los usuarios deciden si participar.
- Sus resultados no son generalizables.

01 Creación y captura: Datos administrativos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).

01

Datos administrativos

- Información recolectada para fines no estadísticos
- Registros salud, impuestos
- Procesos internos.

Retos:

- **Calidad inconsistente** (errores en registros).
- **Acceso restringido** (privacidad, burocracia).

01 Creación y captura: Sensores

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).

01

Sensores

- Datos medidos de forma directa e indirecta
- Monitoreo

Retos:

- **Costo**
- **Acceso restringido** (privacidad,

01 Creación y captura: Levantamiento en campo

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).



Levantamiento en campo

- Recolección de datos **in situ**, mediante observación directa, mediciones o entrevistas en terreno

Retos:

- **Accesibilidad** (terrenos difíciles o peligrosos).
- **Precisión** (errores humanos en medición o registro).
- **Condiciones ambientales** (clima, fauna).
- **Logística compleja** (equipos, permisos).

01 Creación y captura: Minería de datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

01 Creación y captura

Generación u obtención de datos (sensores, formularios, transacciones, APIs, etc.).

01

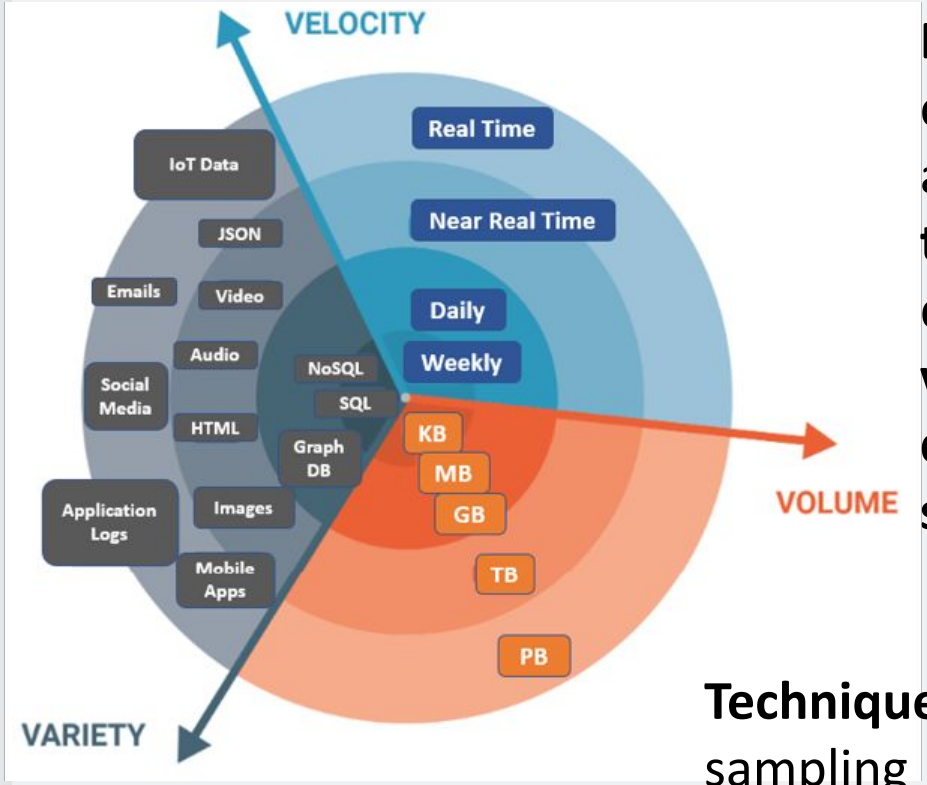
Minería de datos

- Análisis de grandes volúmenes de datos generados digitalmente (redes sociales, transacciones).

Retos:

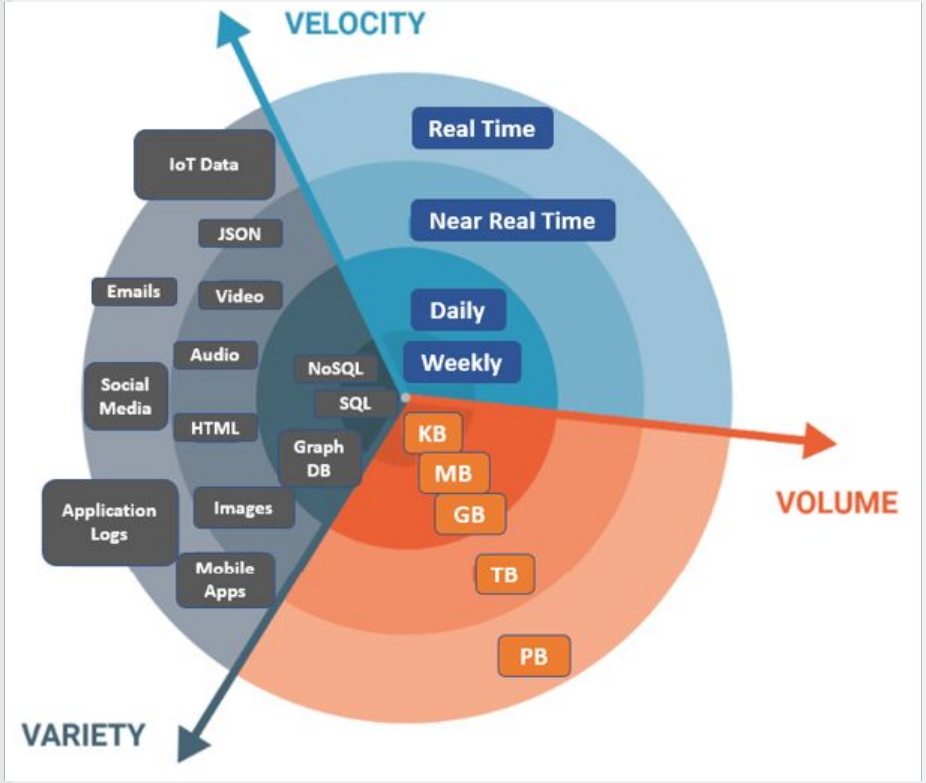
- **Privacidad**
- Ruido de datos.

data that will not fit in main memory.



Big data: extremely **large** and **diverse** collections of structured, unstructured, and semi-structured data that **continues to grow** exponentially over time. These datasets are complex in **volume, velocity, and variety,** that **traditional data management systems cannot store, process, and analyze** them.

Techniques: Mapreduce, parallel computing and sampling



Data that will not fit in main memory.

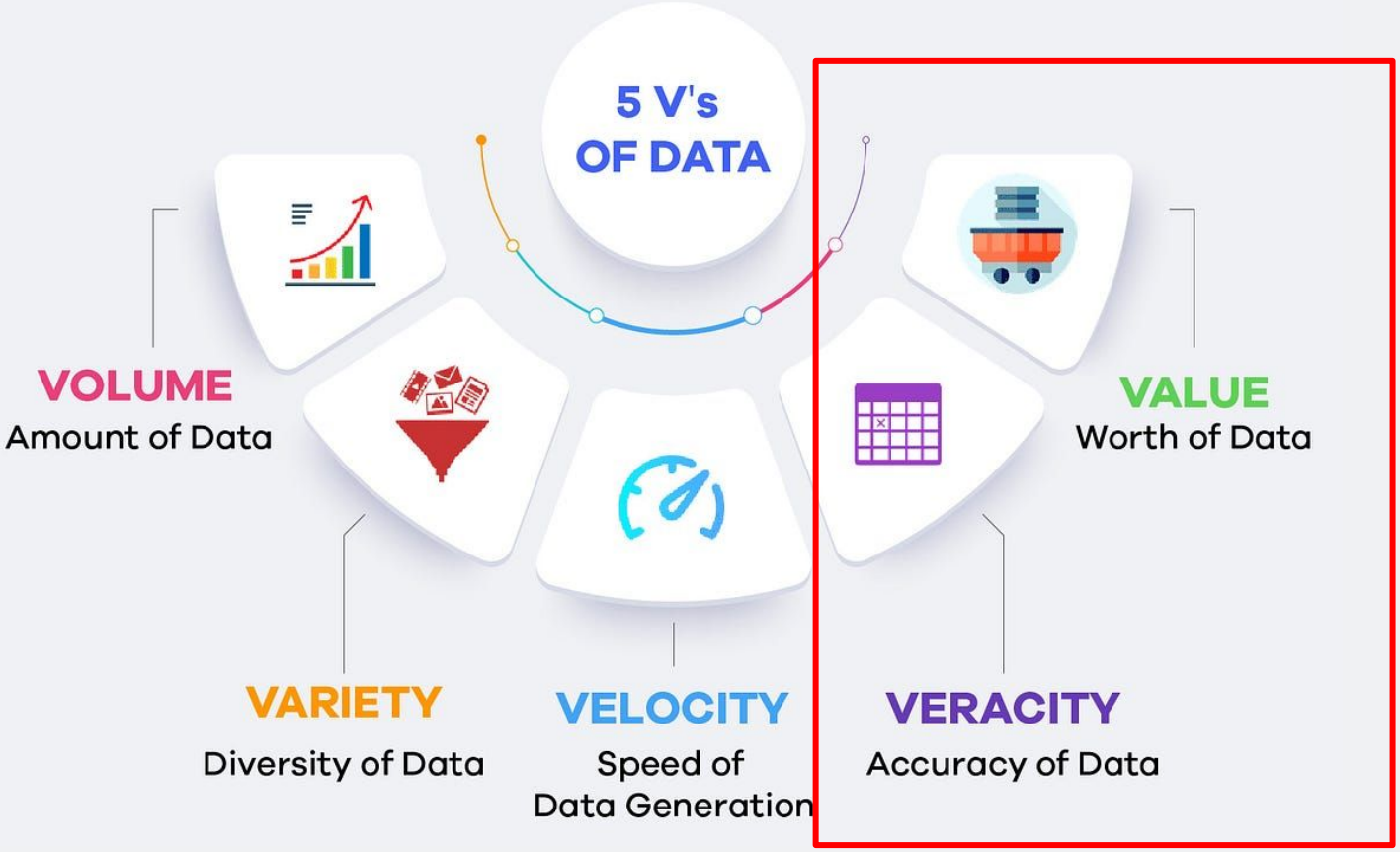
Data with a large number of observations and/or features (statistics)

non-traditional sample size

One of the main challenges of big data is the ability to store, process, and analyze it effectively

Goal: Generalizations
A model or summarization of the data.

Focus on scalability and algorithms/analyses not possible without large data.

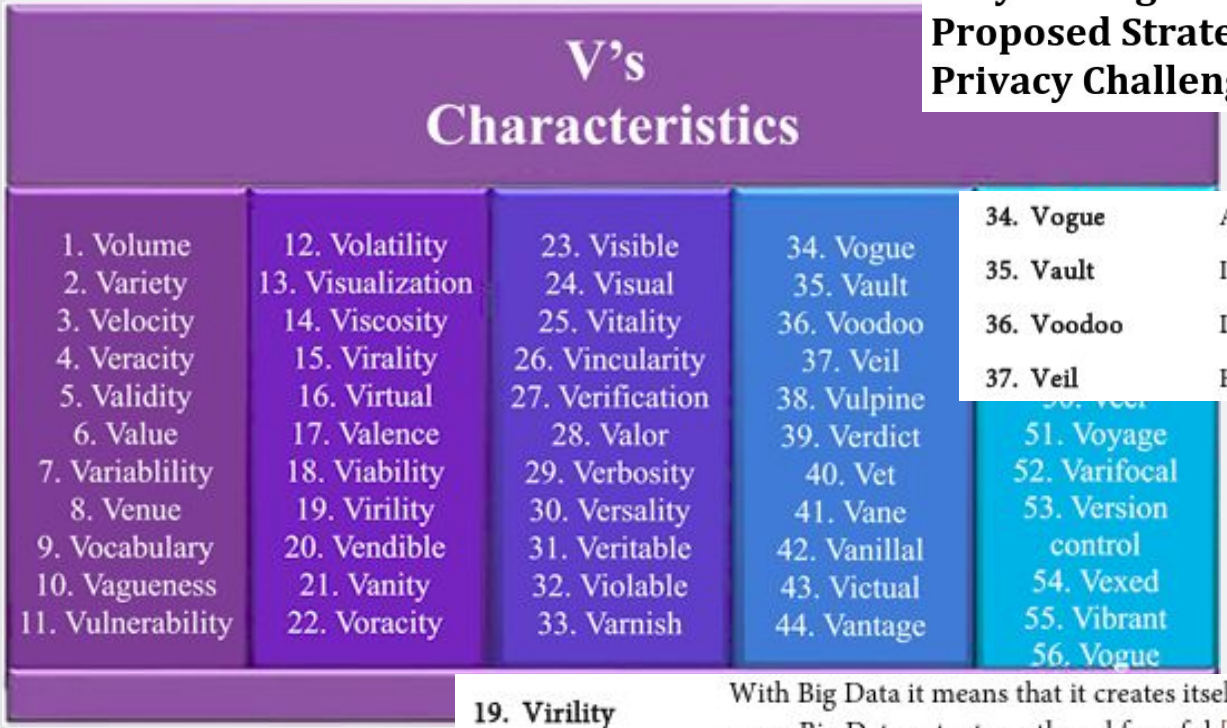

5 V's
OF DATA
VOLUME
Amount of Data
VARIETY
Diversity of Data
VELOCITY
Speed of
Data Generation
VERACITY
Accuracy of Data
VALUE
Worth of Data

56V of Big data

Scientific Research Publishing

Journal of Information Security, 2020, 11, 304-328
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2)



34. Vogue	Artificial intelligence are become? [1] [16].
35. Vault	Importance of data security [1] [16].
36. Voodoo	Deliver results with real-world impact [1] [16].
37. Veil	Examine latent variables from behind the curtain [1] [16].

19. Virility	With Big Data it means that it creates itself. The more Big Data you have, the more Big Data gets strength and forceful [1] [7] [8] [9] [10].



2008



2011



2012



2010



2011



2018

Estados Unidos: Data.gov (clima, salud, educación, transporte, etc.).

Unión Europea: European Data Portal.

Francia: data.gouv.fr (transportes, medio ambiente, economía, entre otros).

India: data.gov.in.

Reino Unido: data.gov.uk.

NASA: Open Data Portal (datos espaciales, imágenes satelitales, etc.).

NOAA: National Centers for Environmental Information (clima, océanos, atmósfera).

European Space Agency: Copernicus Open Access Hub (imágenes satelitales).

Kaggle Datasets: Kaggle

DrivenData: drivendata.org

https://datos.gob.mx/ iiieg.gob.mx

Laboratorios con Colab

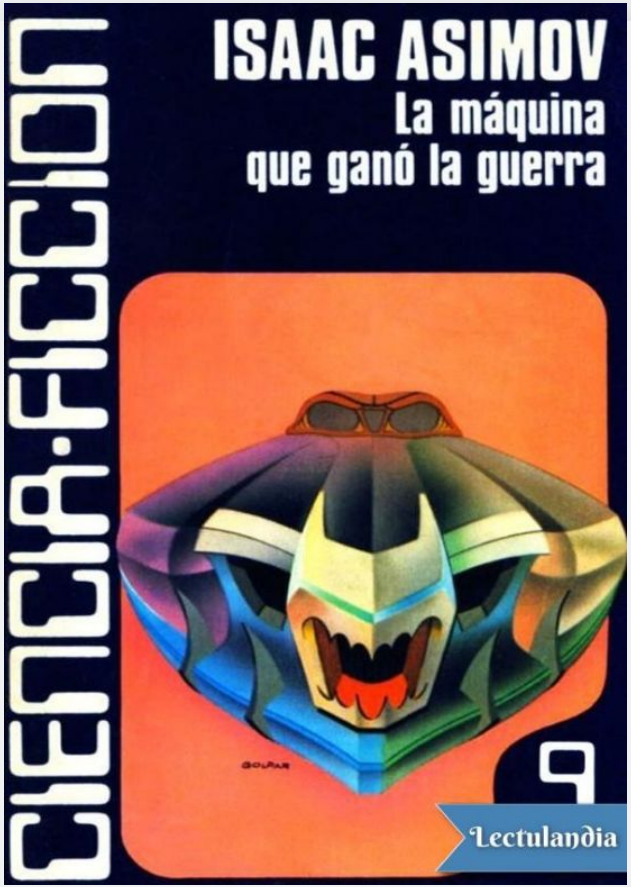
IIEG
Instituto de Información
Estadística y Geográfica
de Jalisco

Titanic Pandas Profile

EDA datos autos

Iris cleanlab





https://pdfcoffee.com/la-maquina-que-gano-la-guerra-isaac-asimov-pdf-free.html

Datos y su importancia II

—Claro —dijo Swift con ánimo consolador—. Comprendo que lo intentaras.

En esta ocasión Jablonsky decidió encender su cigarrillo.

—No obstante, supongo que suministraste datos a los programas de Multivac. No nos dijiste nada sobre esa inconfiabilidad.

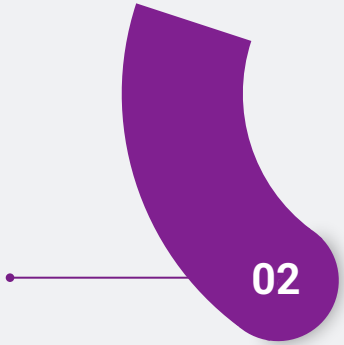
—¿Cómo podía decíroslo? Y si lo hubiera hecho, ¿cómo asegurarme de que ibais a creerme? —preguntó Henderson salvajemente—. El esfuerzo total de la guerra recaía sobre Multivac. Era nuestra única gran arma, pues los denebianos no tenían nada igual. ¿Qué otra cosa podría elevar más la moral que la seguridad de saber que Multivac estaría siempre prediciendo y cercando cualquier movimiento denebiano y que dirigiría siempre y siempre prevería la infalibilidad de nuestras propias maniobras? Por el Gran Espacio, desde que nuestra red de espionaje fue arrojada del hiperespacio no pudimos ya obtener ningún dato fiable acerca de los denebianos para seguir alimentando Multivac, y eso era algo que no podíamos arriesgarnos a hacer público.

Ciclo de vida de los datos:02

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

02 Transmisión y almacenamiento

Los datos de envían a bases de datos, lago de datos, datawarehause,



Posibles errores:

- Pérdida de datos por fallos de hardware.
- Almacenamiento redundante (duplicidad).
- Duplicidad
- Estructura inadecuada (ej: tablas mal normalizadas en SQL).
- **Sin respaldo Regla 3-2-1 o sin respaldos periódicos.**
- **Corrupción de datos.**
- **Errores en el ETL**
- **Formatos incompatibles**

3-2-1: mantener **tres** copias de los datos, en **dos** tipos diferentes de medios de almacenamiento, y **una** copia fuera del sitio

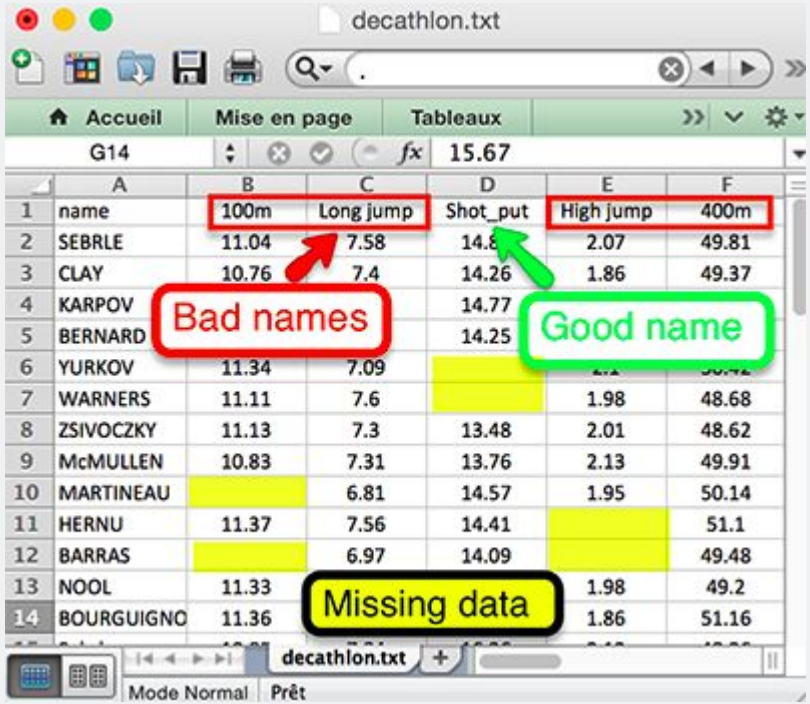
Hola soy _____ y tengo un excel

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

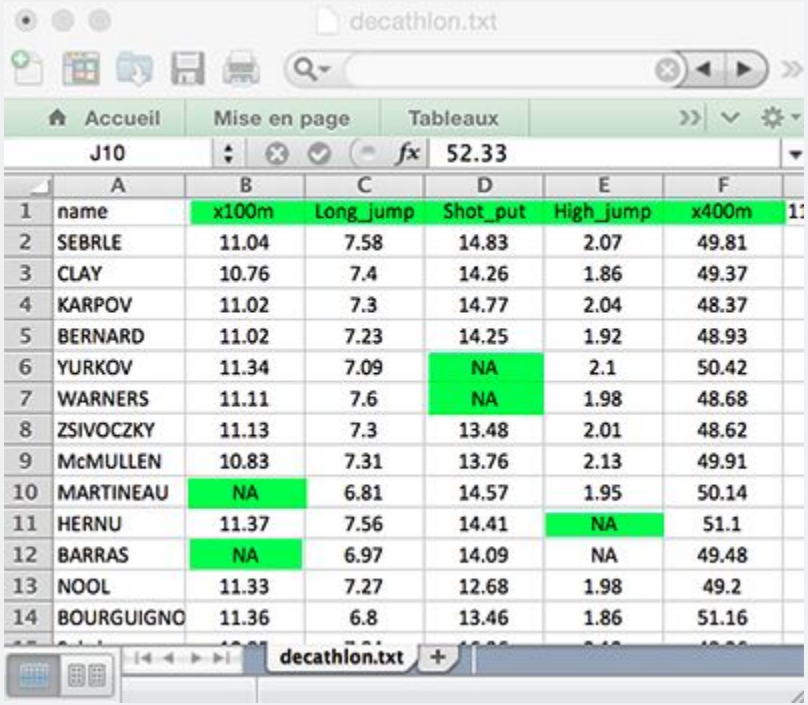
- Me van a regañar?
- Mis datos no sirven ?
- Soy un ser malo ?
- Tengo veneno en la sangre?
- Hay algo peor que un excel?

Tengo un excel

IIEG
Instituto de Información Estadística y Geográfica de Jalisco



	A	B	C	D	E	F
1	name	100m	Long jump	Shot_put	High jump	400m
2	SEBRLE	11.04	7.58	14.8	2.07	49.81
3	CLAY	10.76	7.4	14.26	1.86	49.37
4	KARPOV			14.77		
5	BERNARD			14.25		
6	YURKOV	11.34	7.09			
7	WARNERS	11.11	7.6		1.98	48.68
8	ZSIVOCZKY	11.13	7.3	13.48	2.01	48.62
9	McMULLEN	10.83	7.31	13.76	2.13	49.91
10	MARTINEAU		6.81	14.57	1.95	50.14
11	HERNU	11.37	7.56	14.41		51.1
12	BARRAS		6.97	14.09		49.48
13	NOOL	11.33			1.98	49.2
14	BOURGUIGNO	11.36			1.86	51.16



	A	B	C	D	E	F
1	name	x100m	Long_jump	Shot_put	High_jump	x400m
2	SEBRLE	11.04	7.58	14.83	2.07	49.81
3	CLAY	10.76	7.4	14.26	1.86	49.37
4	KARPOV	11.02	7.3	14.77	2.04	48.37
5	BERNARD	11.02	7.23	14.25	1.92	48.93
6	YURKOV	11.34	7.09	NA	2.1	50.42
7	WARNERS	11.11	7.6	NA	1.98	48.68
8	ZSIVOCZKY	11.13	7.3	13.48	2.01	48.62
9	McMULLEN	10.83	7.31	13.76	2.13	49.91
10	MARTINEAU	NA	6.81	14.57	1.95	50.14
11	HERNU	11.37	7.56	14.41	NA	51.1
12	BARRAS	NA	6.97	14.09	NA	49.48
13	NOOL	11.33	7.27	12.68	1.98	49.2
14	BOURGUIGNO	11.36	6.8	13.46	1.86	51.16

Tengo un excel

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

[Inversión extranjera directa]

[Inversión por entidad federativa]

flujosporentidadfederativa_actu__5_1 .XLSX
Archivo Editar Ver Insertar Formato Datos Herramientas Ayuda

Nota: Nota: Datos acumulados para cada año..

Inversión Extranjera Directa en México por entidad federativa y tipo de inversión
CIFRAS ACTUALIZADAS, por trimestre, en *millones de dólares corrientes*

	2006				2007	
Entidad Federativa/Tipo de Inversión	**1**	**2**	**3**	**4**	**1**	**2**
Total general	**7,437.8**	**14,072.1**	**16,418.9**	**21,236.8**	**10,815.8**	**16,953.4**
Aguascalientes	**49.9**	**90.2**	**114.1**	**140.6**	**51.9**	**142.5**
Cuentas entre compañías	14.7	30.7	40.9	57.5	-2.2	14.7
Nuevas inversiones	10.9	22.6	29.3	38.0	24.8	89.6
Reinversión de utilidades	24.4	37.0	43.9	45.1	29.2	38.2
Baja California	**305.0**	**619.5**	**947.5**	**1,340.7**	**488.4**	**749.4**
Cuentas entre compañías	182.4	410.0	687.2	952.4	222.6	383.8
Nuevas inversiones	65.8	115.6	162.6	281.9	171.4	231.2
Reinversión de utilidades	56.8	93.9	97.7	106.4	94.4	134.5
Baja California Sur	**266.4**	**411.9**	**534.6**	**592.2**	**213.3**	**408.2**
Cuentas entre compañías	190.7	217.6	249.3	272.3	44.2	83.0
Nuevas inversiones	64.1	177.0	267.4	301.5	153.6	304.9
Reinversión de utilidades	11.6	17.3	17.9	18.4	15.5	20.3
Campeche	**15.2**	**25.1**	**34.3**	**30.4**	**12.4**	**73.1**
Cuentas entre compañías	2.2	3.2	10.4	-3.4	-4.7	-8.1

Tipo de inversión | País de origen | Actividad económica

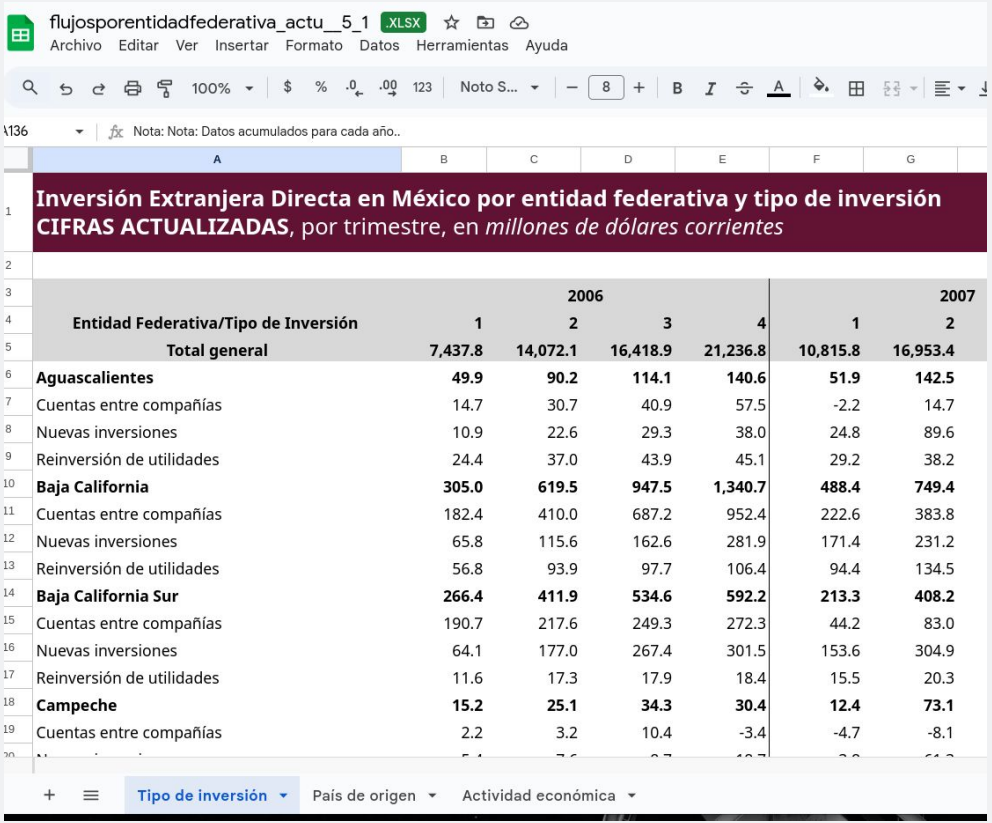
Tengo un excel

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

[Inversión extranjera directa]

[Inversión por entidad federativa]

Periodo	Aguascalientes	... (otras entidades)
q1-2006	49.9	...
q2-2006	90.2	...



flujosporentidadfederativa_actu__5_1 .XLSX
Archivo Editar Ver Insertar Formato Datos Herramientas Ayuda

Nota: Nota: Datos acumulados para cada año..

Inversión Extranjera Directa en México por entidad federativa y tipo de inversión
CIFRAS ACTUALIZADAS, por trimestre, en *millones de dólares corrientes*

	2006				2007	
Entidad Federativa/Tipo de Inversión	**1**	**2**	**3**	**4**	**1**	**2**
Total general	**7,437.8**	**14,072.1**	**16,418.9**	**21,236.8**	**10,815.8**	**16,953.4**
Aguascalientes	**49.9**	**90.2**	**114.1**	**140.6**	**51.9**	**142.5**
Cuentas entre compañías	14.7	30.7	40.9	57.5	-2.2	14.7
Nuevas inversiones	10.9	22.6	29.3	38.0	24.8	89.6
Reinversión de utilidades	24.4	37.0	43.9	45.1	29.2	38.2
Baja California	**305.0**	**619.5**	**947.5**	**1,340.7**	**488.4**	**749.4**
Cuentas entre compañías	182.4	410.0	687.2	952.4	222.6	383.8
Nuevas inversiones	65.8	115.6	162.6	281.9	171.4	231.2
Reinversión de utilidades	56.8	93.9	97.7	106.4	94.4	134.5
Baja California Sur	**266.4**	**411.9**	**534.6**	**592.2**	**213.3**	**408.2**
Cuentas entre compañías	190.7	217.6	249.3	272.3	44.2	83.0
Nuevas inversiones	64.1	177.0	267.4	301.5	153.6	304.9
Reinversión de utilidades	11.6	17.3	17.9	18.4	15.5	20.3
Campeche	**15.2**	**25.1**	**34.3**	**30.4**	**12.4**	**73.1**
Cuentas entre compañías	2.2	3.2	10.4	-3.4	-4.7	-8.1

Tipo de inversión | País de origen | Actividad económica

Característica	Bases de Datos (SQL/NoSQL)	Excel (XLSX)	CSV
Estructura	Esquema definido (tablas, documentos).	Hojas con celdas (semiestructurado).	Texto plano, sin esquema (solo filas/columnas).
Volumen de datos	✅ Alto (TB o más). Optimizado para grandes datasets.	❌ Limitado (hasta ~1M filas). Lento con datos grandes.	⚠️ Moderado (depende del sistema de archivos).
Consultas	Lenguaje SQL (ej: `JOIN`, `GROUP BY`).	Fórmulas básicas (VLOOKUP, tablas dinámicas).	❌ Sin capacidades de consulta nativas.
Concurrencia	Múltiples usuarios acceden simultáneamente con transacciones.	❌ Bloqueo al editar (un usuario a la vez).	❌ Solo lectura concurrente (sin edición simultánea).
Seguridad	Permisos por usuario/rol (ej: `GRANT SELECT`).	🔒 Contraseñas de archivo (menos seguro).	❌ Sin seguridad integrada.
Integridad	Validaciones, claves foráneas, ACID (SQL).	Validaciones manuales (pocas garantías).	❌ Sin validaciones.
Formato	Binario o gestionado por el motor (ej: `.db`, `.mdf`).	Binario (`.xlsx`).	Texto plano (`.csv`).
Velocidad	Optimizado para búsquedas y operaciones complejas.	Lento con muchos datos o fórmulas.	Rápido para lectura/escritura simple.
Escalabilidad	✅ Horizontal (sharding) o vertical.	❌ No escalable.	❌ No escalable.
Uso típico	Apps web, análisis empresarial, transacciones.	Reportes pequeños, cálculos manuales.	Intercambio de datos, migraciones simples.

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

SQL

Standard language for querying and manipulating data

Structured Query Language

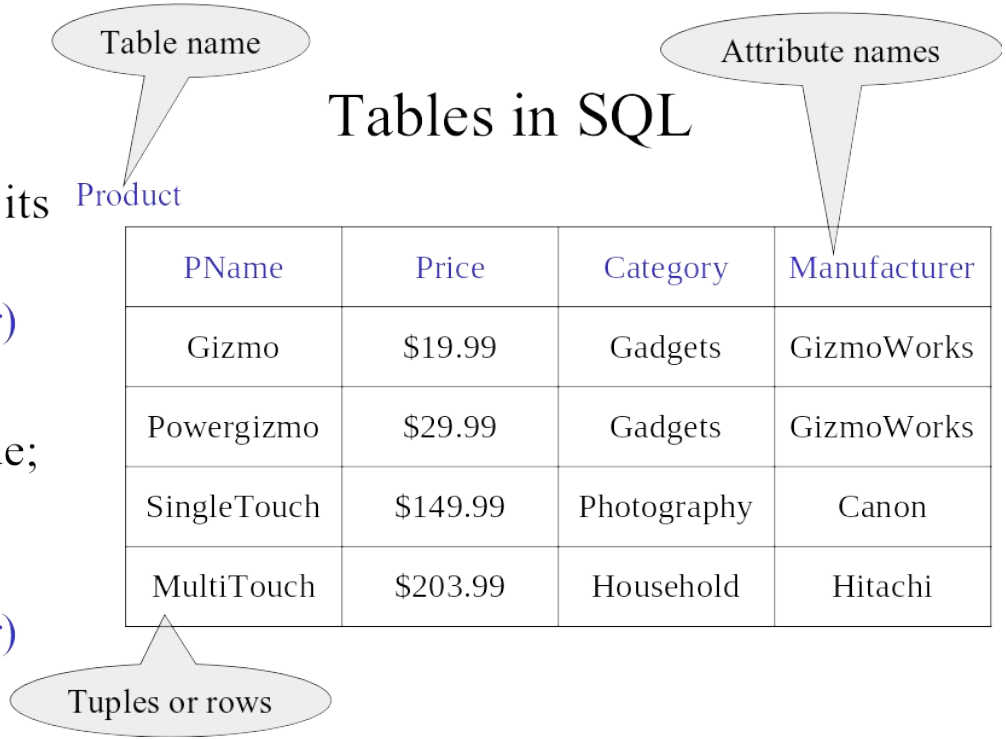
Tables Explained

- The *schema* of a table is the table name and its attributes:

Product(PName, Price, Category, Manfacturer)

- A *key* is an attribute whose values are unique; we underline a key

Product(PName, Price, Category, Manfacturer)



Standard language for querying and manipulating data

Structured Query Language

Basic form: (plus many many more bells and whistles)

```
SELECT  <attributes>
FROM    <one or more relations>
WHERE   <conditions>
```

SQL

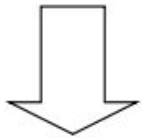
Standard language for querying and manipulating data

Structured Query Language

Product

PName	Price	Category	Manufacturer
Gizmo	$19.99	Gadgets	GizmoWorks
Powergizmo	$29.99	Gadgets	GizmoWorks
SingleTouch	$149.99	Photography	Canon
MultiTouch	$203.99	Household	Hitachi

```
SELECT *
FROM Product
WHERE category='Gadgets'
```



SQL

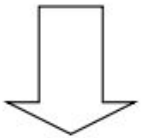
Standard language for querying and manipulating data

Structured Query Language

Product

PName	Price	Category	Manufacturer
Gizmo	$19.99	Gadgets	GizmoWorks
Powergizmo	$29.99	Gadgets	GizmoWorks
SingleTouch	$149.99	Photography	Canon
MultiTouch	$203.99	Household	Hitachi

```
SELECT  PName, Price, Manufacturer
FROM    Product
WHERE   Price > 100
```



SQL

Standard language for querying and manipulating data

Structured Query Language

Data Types in SQL

- Atomic types:
  - Characters: CHAR(20), VARCHAR(50)
  - Numbers: INT, BIGINT, SMALLINT, FLOAT
  - Others: MONEY, DATETIME, …

- Every attribute must have an atomic type
  - Hence tables are flat
  - Why ?

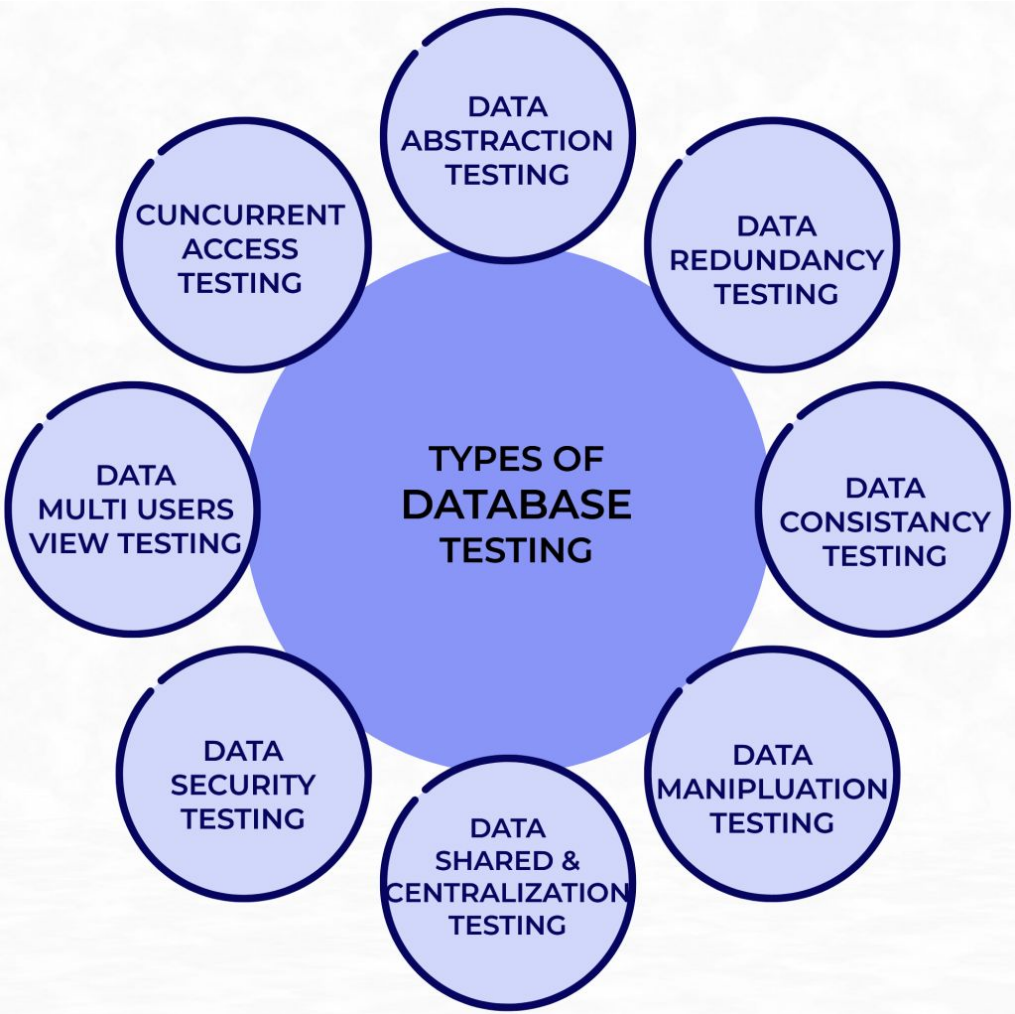
SQL

SQLite: Base de datos ligera, sin servidor, almacenada en un solo archivo.

Ventajas de SQLite

✅ No requiere configuración.
✅ Ideal para proyectos pequeños y aprendizaje.
✅ Soporta SQL estándar.

Colab Sqlite



Database testing is the process to **ensure the quality, accuracy, and security** of the **data** stored in a database.

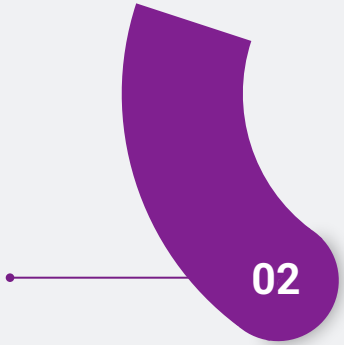
it involves verifying the data within a database by assessing the components of the software applications or systems that influence the data and its associated functionalities.

Ciclo de vida de los datos:02

IIEG
Instituto de Información
fica

02 Transmisión y almacenamiento

Los datos de envían a bases de datos, lago de datos, datawarehause,



ETL

- **Extracción-Transformación-Carga**
- ETL = Extract -> Transform -> Load

– **Extract**:Obtener los los datos tan eficientemente como sea posible

– **Transform**: hacer algunos cálculos, transforman los datos, limpiar datos

– **Load**: cargar los datos en la siguiente aplicación

Plomero de los datos: Los datos se mueven desde las fuentes a los siguientes procesos.

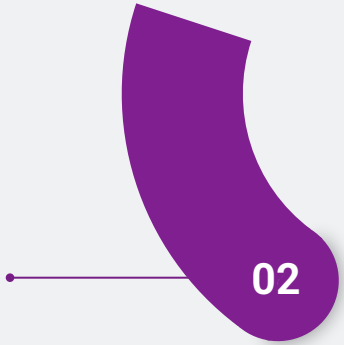
https://indepth-network.org/sites/indepth.mrmdev.co.uk/files/2017/workshops/ishare/presentations/Introduction_ETL.pdf

Ciclo de vida de los datos:02

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

02 Transmisión y almacenamiento

Los datos de envían a bases de datos, lago de datos, datawarehause,



Data Lake

Un **Data Lake** es un repositorio que almacena datos **en bruto** (estructurados, semiestructurados y no estructurados) sin procesar inicialmente.

Datos en crudo: Sin transformar (JSON, CSV, imágenes, logs, etc.).

Esquema flexible: No requiere estructura definida al ingresar.

Big Data: Usado para análisis masivo con operaciones simples por ejemplo map reduce.

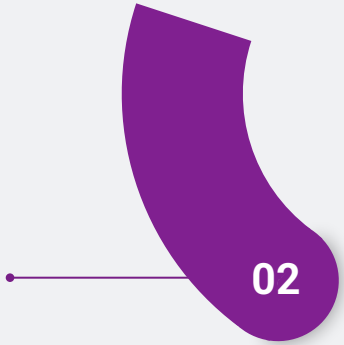
Escalable: Diseñado para grandes volúmenes (Hadoop, AWS S3).

Ciclo de vida de los datos:02

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

02 Transmisión y almacenamiento

Los datos de envían a bases de datos, lago de datos, datawarehause,



Datawarehouse (Almacén de datos)

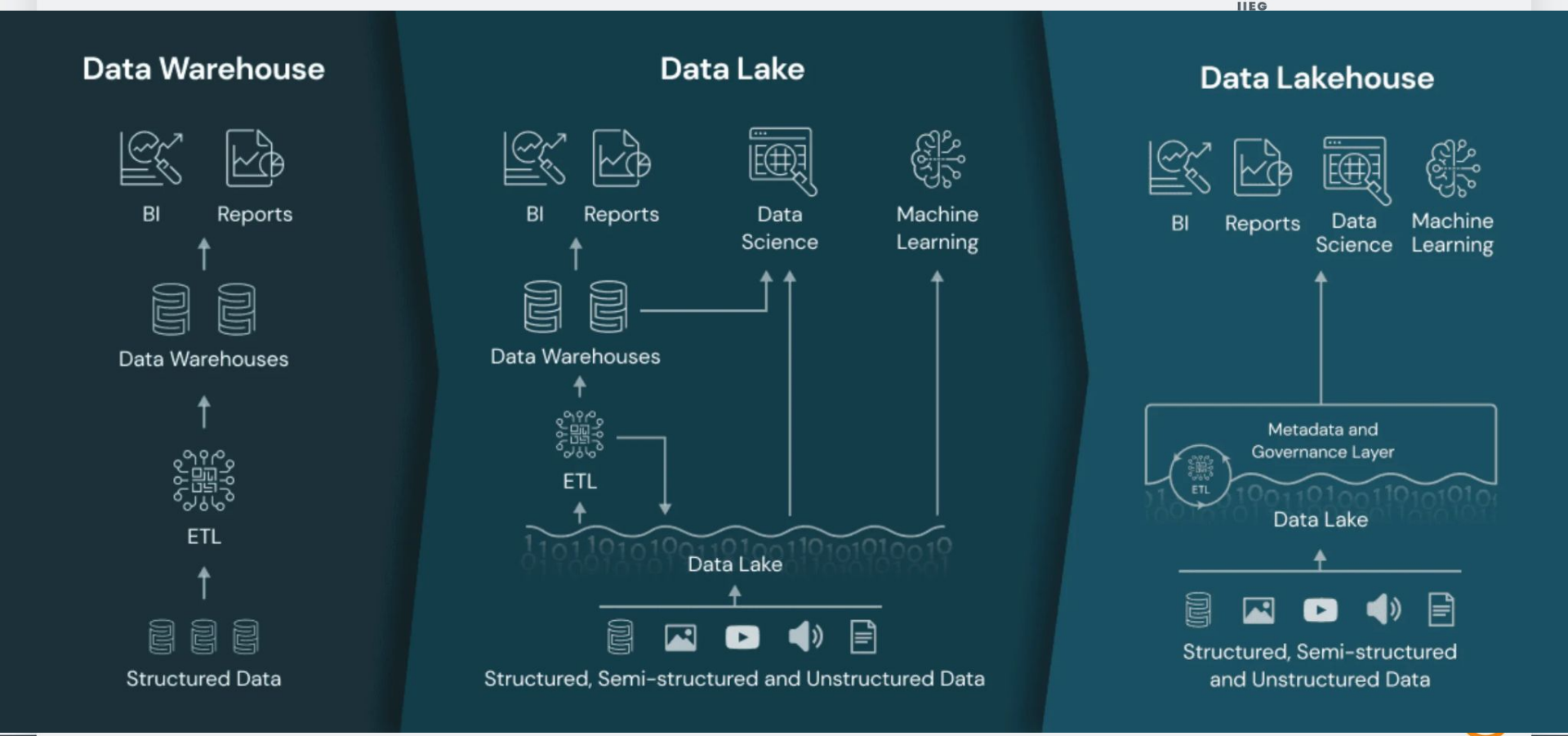
Es un repositorio centralizado de datos estructurados y procesados, diseñado para **consultas y análisis empresarial**

Características

Datos estructurados: Organizados en tablas (esquema relacional).
Procesados (ETL): Los datos se limpian, transforman y cargan antes de almacenarse.
Orientado a consultas: Optimizado para reportes y análisis

Input datos estructurados

Ciclo de vida de los datos:02



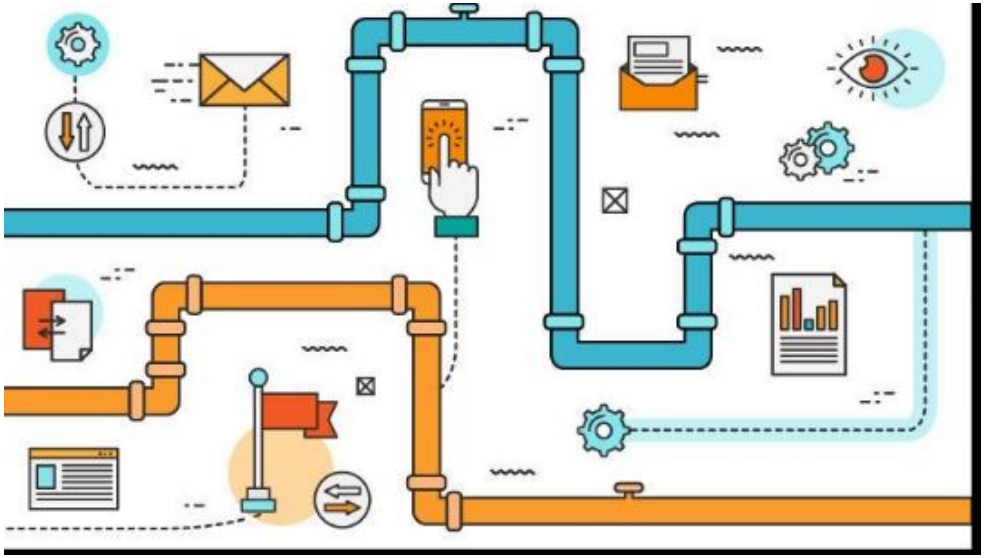
Data Warehouse Vs Data Lake

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Aspecto	Data Warehouse	Data Lake
Tipo de datos	Sólo estructurados (tablas SQL).	Cualquier tipo (JSON, imágenes, logs).
Procesamiento	Datos transformados antes de almacenar (ETL).	Datos en crudo (se procesan al leer).
Esquema	Definido antes de cargar (schema-on-write).	Definido al consultar (schema-on-read).
Usuarios	Analistas de BI.	Científicos de datos/Ingenieros.
Herramientas	Snowflake, Redshift, BigQuery.	Hadoop, AWS S3, Azure Data Lake.

Orquestación de datos

La orquestación de datos es el proceso de extraer datos aislados de múltiples ubicaciones de almacenamiento, combinarlos, organizarlos y ponerlos a disposición de las herramientas de análisis de datos. También puede incluir tareas como el aprovisionamiento de recursos y la monitorización.



Apache Airflow

Luigi

Azkaban

Oozie

and more..

Quartz

https://airflowsummit.org/slides/2022/b1-choosing-airflow-v2.pdf

History

Apache Airflow

- Started in October 2014 by Maxime Beauchemin at Airbnb.
- Open source from the very first commit and officially brought under the Airbnb GitHub and announced in June 2015
- Joined the Apache Software Foundation's Incubator program in March 2016

Broadcom AutoSys

- First developed by William Arntz and Walte r Goodwin who created AutoSystems Ltd.
- Sold to Platinum Technology International in 1995
- Bought by Computer Associates in 1999
- Broadcom acquired CA in 2018 and known today as "AutoSys Workload Automation"

bmc Control-M

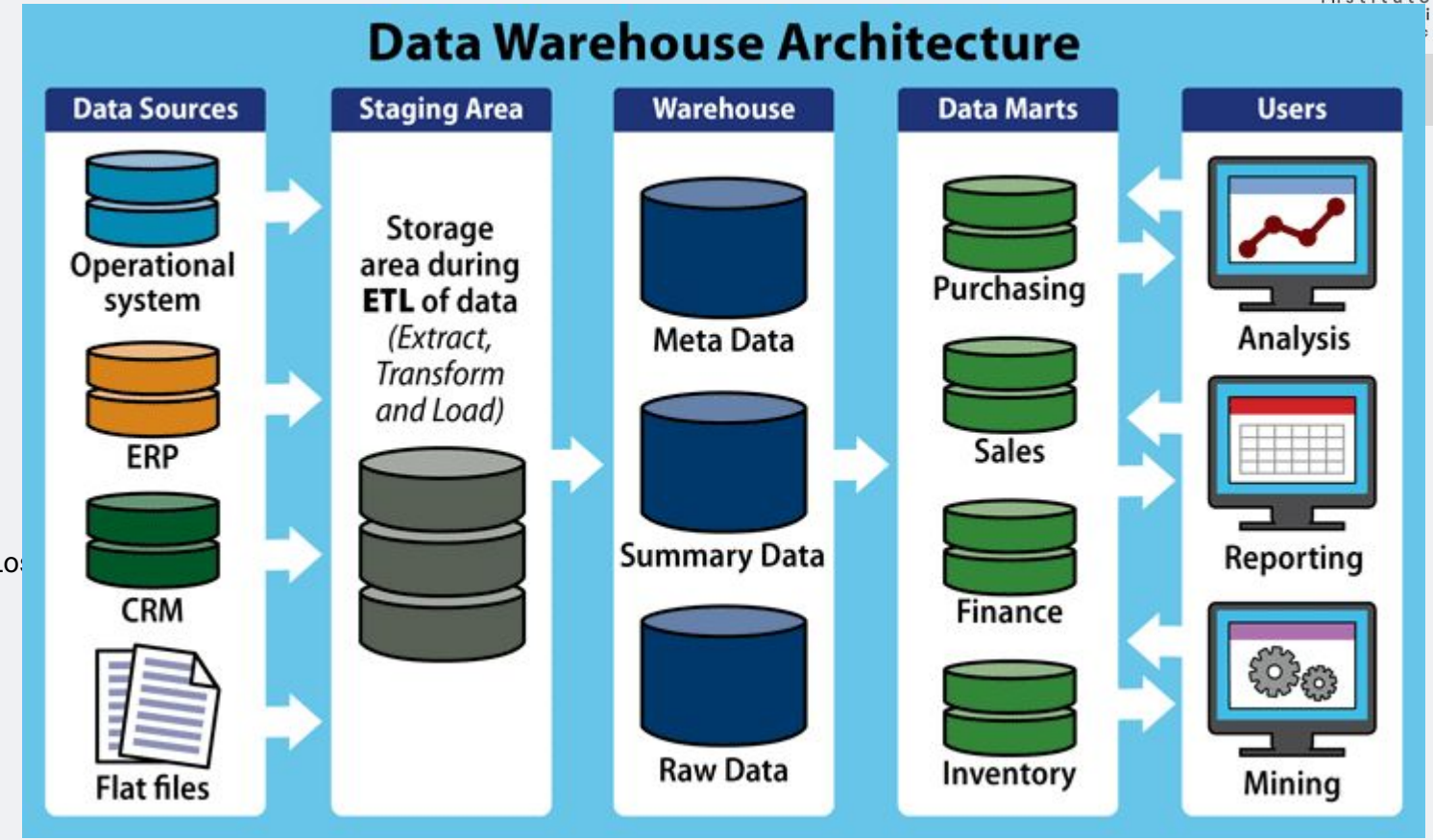
- Originally developed for scheduling jobs on mainframe computer systems, by an Israel-based company - New Dimension Software
- BMC software acquired New Dimension Software in **1999**
- "BMC Helix Control-M" - SaaS solution was released in December - **2020**

Spotify Luigi

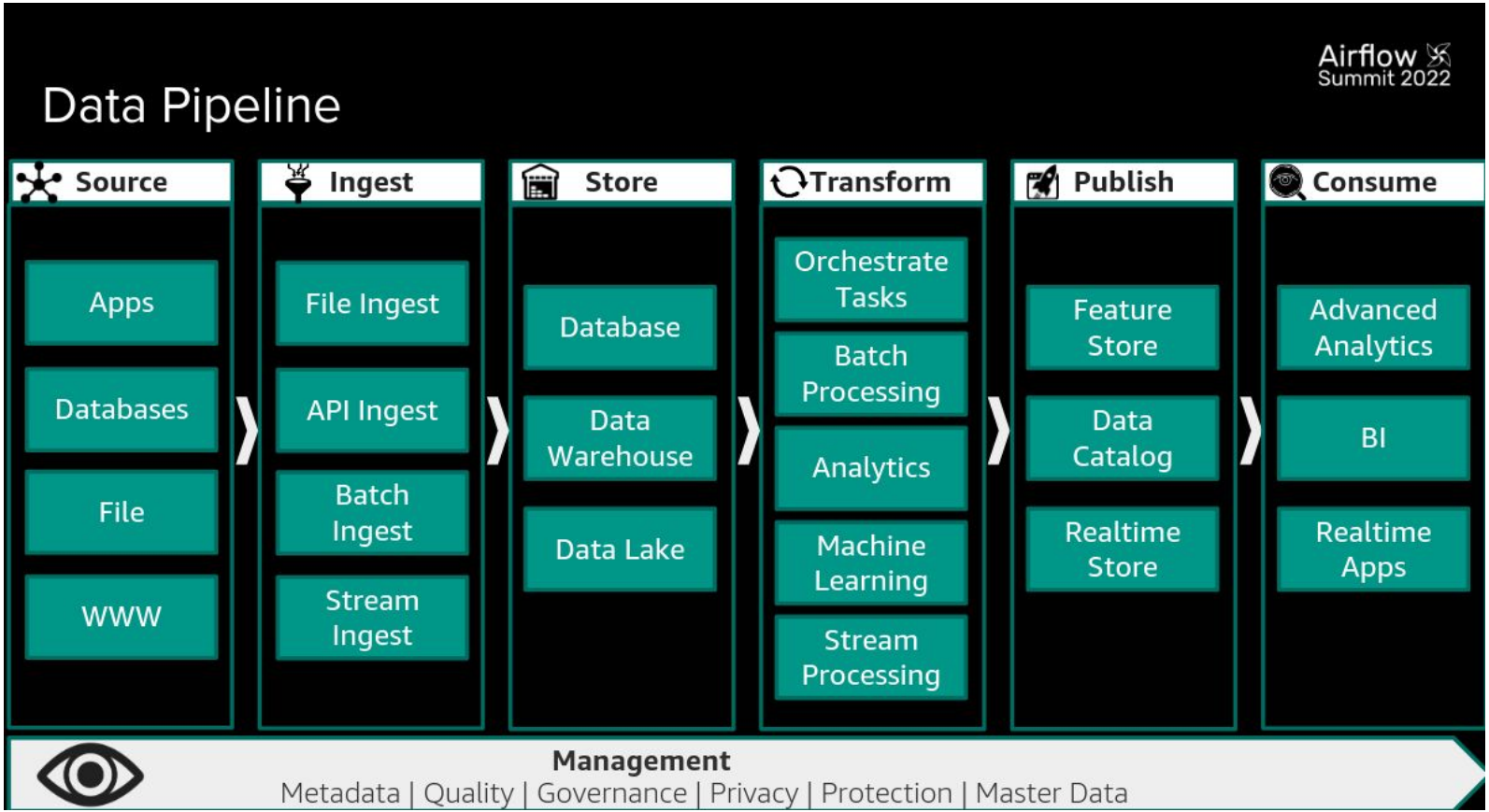
- Created by Spotify mainly by Erik Bernhardsson and Elias Freider
- Initial commit on github/spotify/luigi on Nov 17, **2011**
- Open sourced in **2012**
- Spotify's Data Team maintains Luigi

https://airflowsummit.org/slides/2022/b1-choosing-airflow-v2.pdf

Ciclo de vida de los datos:02



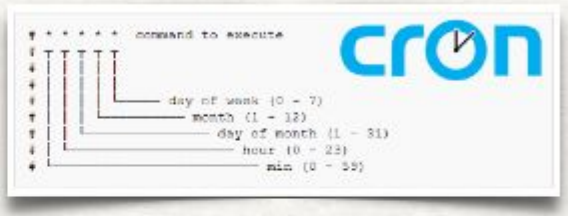
Los



https://airflowsummit.org/slides/2022/b1-choosing-airflow-v2.pdf



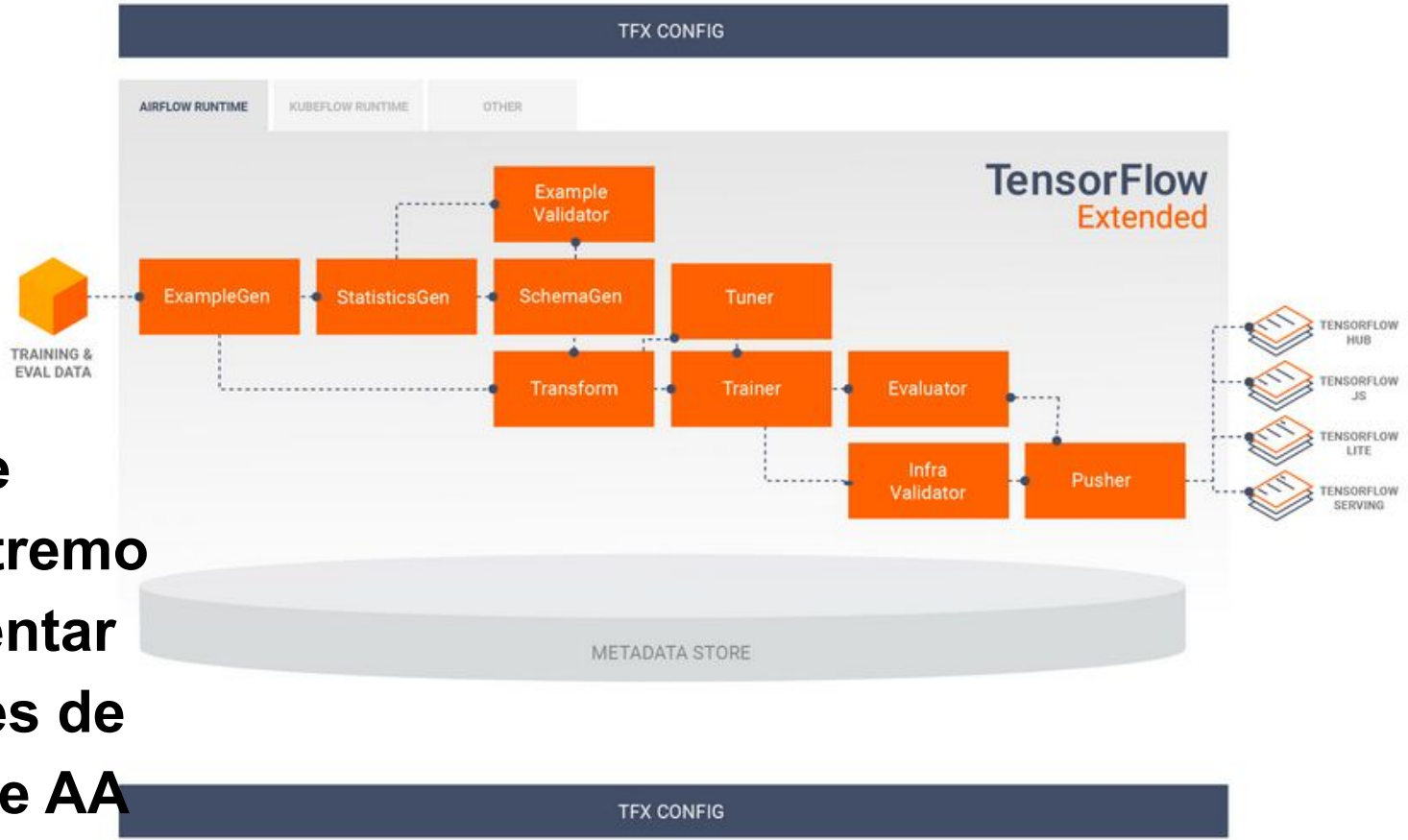


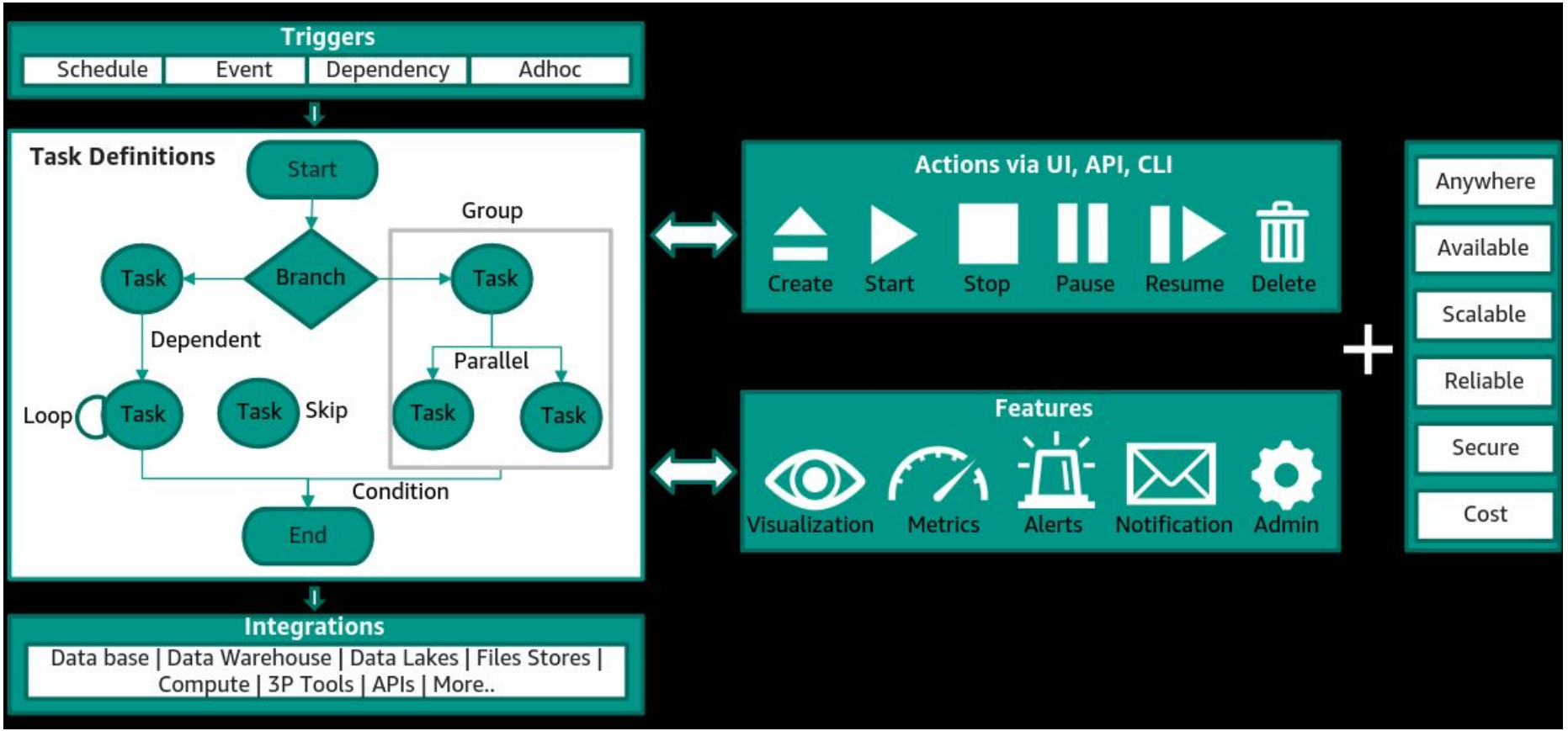


"Scripting + Cron would do!"

TFX

TFX es una plataforma de extremo a extremo para implementar canalizaciones de producción de AA





https://airflowsummit.org/slides/2022/b1-choosing-airflow-v2.pdf

Definiciones básicas

- **Apache Airflow:**
  - Plataforma de **orquestación de workflows** programables.
  - Open-source, escrito en Python.
- Características clave:
  - Escalable, dinámico, extensible.
  - Basado en **DAGs** (*Directed Acyclic Graphs*).
- Ejemplos comunes:
  - ETL/ELT (Data Pipelines).
  - Entrenamiento de modelos de ML.

Estructura del Airflow

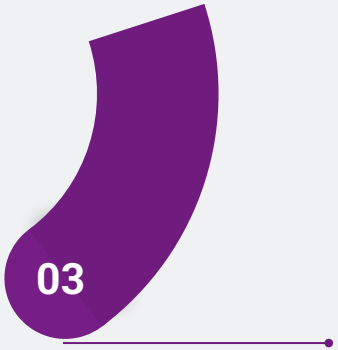
- **DAG**: Grafo acíclico dirigido (conjunto de tareas ordenadas).
- **Operators**: Bloques de construcción (PythonOperator, BashOperator, etc.).
- **Tasks**: Instancias de un Operator.
- **Scheduler**: Ejecuta tareas según dependencias.
- **Executor**: Define cómo se ejecutan las tareas (LocalExecutor, CeleryExecutor).

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Posibles errores:

- Robo/secuestro de datos
- Datos inaccesibles
- Formatos obsoletos
- Riesgos legales
- Falta de encriptación o anonimización
- No cumplir con estándares.



03 Gestión de acceso y seguridad

Los datos se comparten con otros usuarios y fluyen a otras aplicaciones

Ciberseguridad

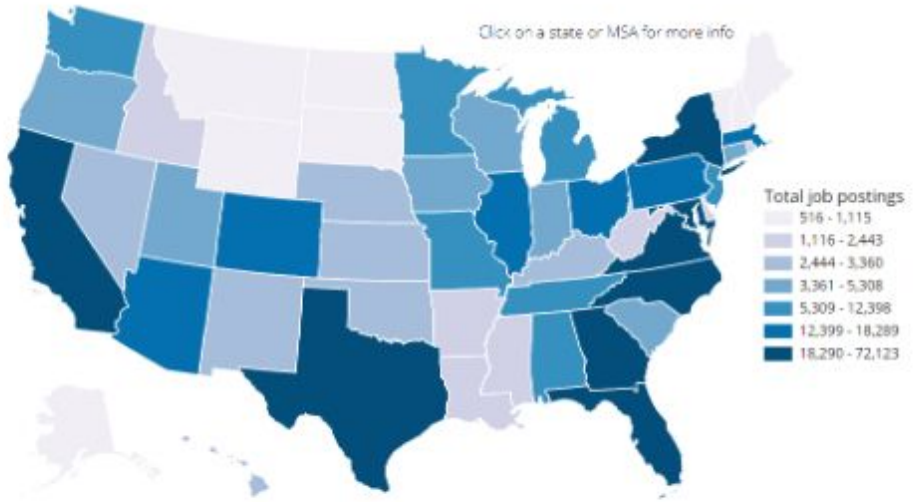
- Protección y seguridad de sistemas y datos contra ataques digitales.
- Son un conjunto de prácticas y principios diseñadas para salvaguardar la infraestructura e información en contras de ataques.

THE WHY
Critical Cyber Workforce Shortage

Our national security and economic stability depends on us solving it.

- Employers are desperate for cyber degree program graduates.[3]
- **504,316 cybersecurity job vacancies** (Oct. 2018- Sep. 2019)[1], and the shortage continues to grow[2]
- Cybercriminals now responsible for **billions in losses per year** and **state-sponsored hacking groups** posing an ever-greater threat.[3]

Cybersecurity Supply/Demand Heat Map
Total Job Openings



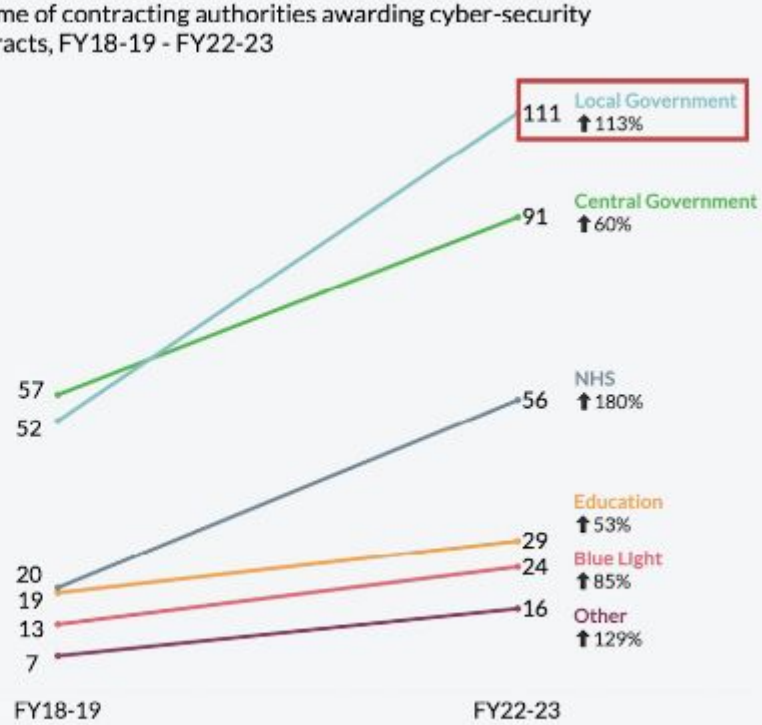
1 CyberSeek.org
2 Supporting the Growth and Sustainment of the Nation's Cybersecurity Workforce. https://www.nist.gov/ sites/default/files/documents/2018/07/24/eo_wf_report_ to_potus.pdf
3 The Cybersecurity Workforce Gap. William Crumpler & James A. Lewis. https://csis-prod.s3.amazonaws.com/s3fs-public/publication/190129_Crumpler_Cybersecurity_FINAL.pdf

CYBER.ORG

The number of cyber security buyers has doubled since FY18-19

In 4 years, the number of contracting authorities awarding cyber-security contracts almost **doubled from 168 to 327**

There is still plenty of room for growth – in FY22-23 this was **only 16%** of all UK public sector contracting authorities



AI IN CYBERSECURITY



1

Social engineering schemes:

These schemes rely on psychological manipulation to trick individuals into revealing sensitive information or making other security mistakes. They include a broad range of fraudulent activity categories, including phishing, vishing and business email compromise scams.

2

Password hacking:

Cybercriminals exploit AI to improve the algorithms they use for deciphering passwords. The enhanced algorithms provide quicker and more accurate password guessing, which allows hackers to become more efficient and profitable. This may lead to an even greater emphasis on password hacking by cybercriminals.

3

Deepfakes:

This type of deception leverages AI’s ability to easily manipulate visual or audio content and make it seem legitimate. This includes using phony audio and video to impersonate another individual. The doctored content can then be broadly distributed online in seconds—including on influential social media platforms—to create stress, fear or confusion among those who consume it.

4

Data poisoning:

Hackers “poison” or alter the training data used by an AI algorithm to influence the decisions it ultimately makes. In short, the algorithm is being fed with deceptive information, and bad input leads to bad output.

Inyección SQL

- **Qué es**: Ataque donde se inserta código malicioso en consultas SQL
- **Impacto**: Robo, modificación o eliminación de datos.
- **Mitigación**:
  - Usar consultas parametrizadas (prepared statements).
  - Validar y sanitizar inputs.

Exposición de Datos Sensibles

- **Qué es**: Acceso no autorizado a datos por falta de encriptación o permisos excesivos.
- **Mitigación**:
  - Encriptar datos en reposo (AES-256) y en tránsito (TLS).
  - Implementar **mínimos privilegios** (RBAC).

RBAC (*Role-Based Access Control)*

Configuraciones Inseguras

- **Ejemplos**: Credenciales por defecto, puertos abiertos (ej: 3306 para MySQL).
- **Solución**:
  - Hardening de la base de datos (deshabilitar servicios innecesarios).

4. Falta de Backup Seguro

- **Riesgo**: Pérdida de datos por ransomware o fallos.
- **Mitigación**:
  - Backups automáticos + encriptados.
  - Probarlos regularmente.



Riesgo	Bases de Datos	Procesos ETL
Inyección	SQL Injection	Inyección en transformaciones
Exposición	Datos en reposo	Datos en tránsito
Autenticación	Privilegios excesivos	Credenciales en código
Integridad	Corrupción de datos	Datos no validados

Taller de datos

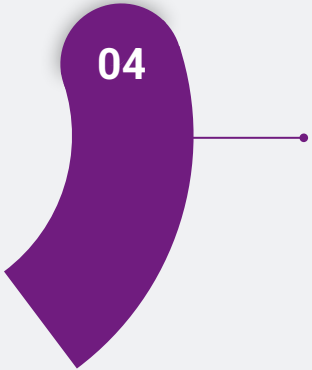
- Acomodar los datos, nombres de columna eliminar encabezados, imágenes(Casi dejar como CSV)
- Explorar:
  - Buscar los datos faltantes (reemplazarlos, eliminarlos),
  - Describir los tipos de datos y en su caso transformarlos por ejemplo enteros, o fechas.
  - Visualizar los datos.
  - Resolver problemas.

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

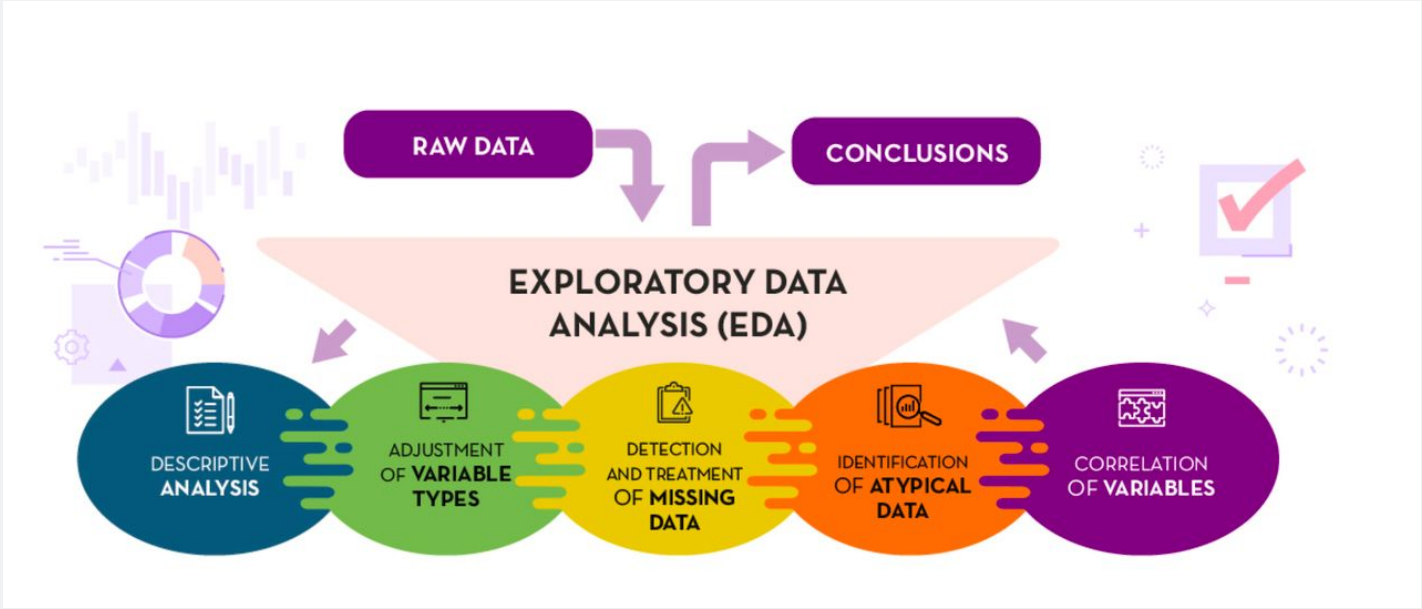
Posibles errores:

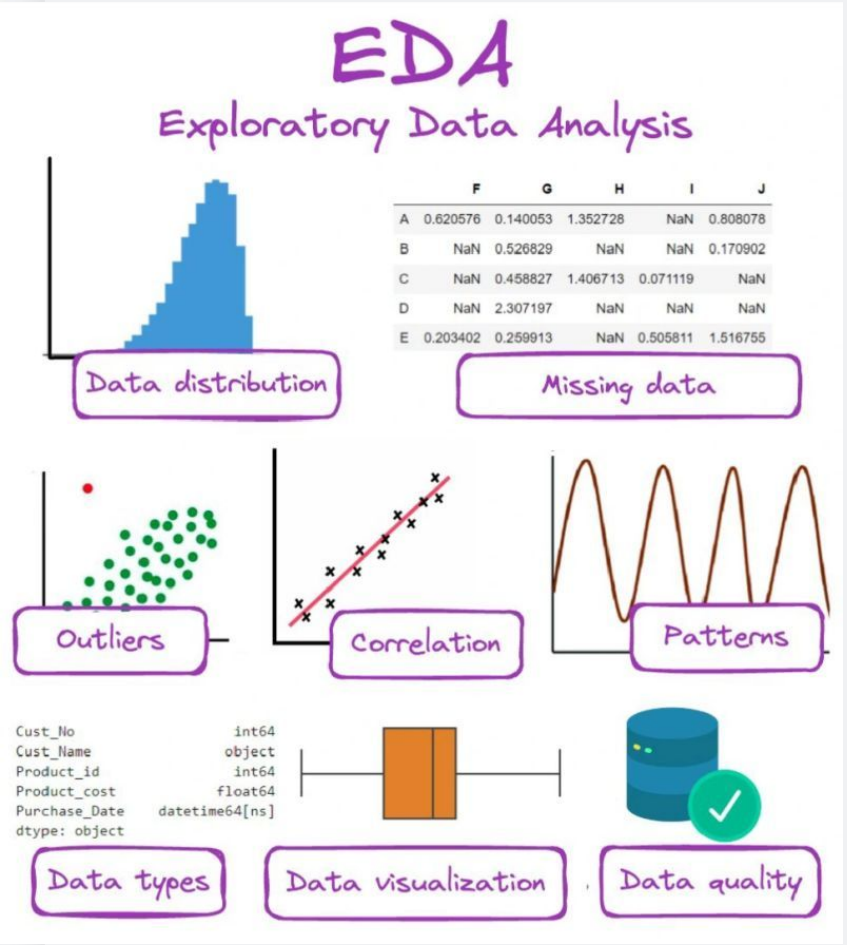
- Uso de datos desactualizados
- Interpretación incorrecta
- Falta de revisión de los procedimientos



04 Limpieza, Uso, Análisis y Visualizaciones

Los datos se depuran y se procesan para calcular, analizar o visualizar algo.


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES



1. Distribution Analysis

Distribution analysis is a fundamental technique used in EDA. It helps us understand how the data points in our dataset are distributed across its range. It also helps in identifying any outliers and is often represented using histograms, density plots, and box plots.

2. Correlation Analysis

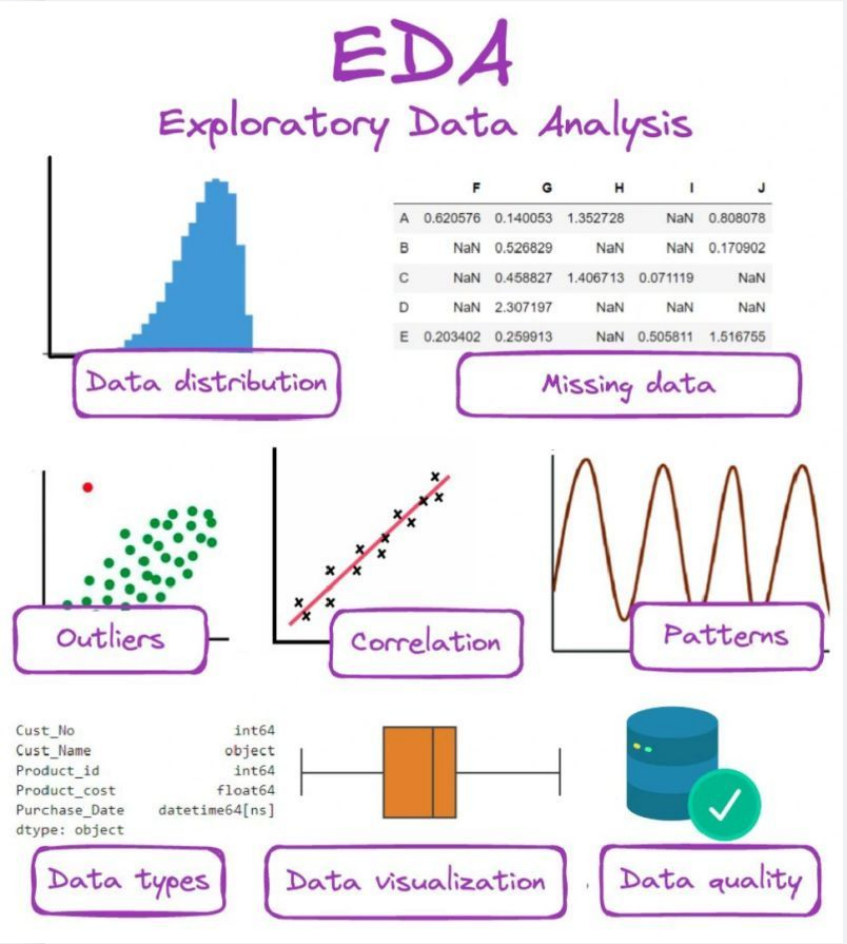
Correlation analysis helps to identify the relationships between various variables in our dataset. Understanding these dependencies is an important concept in machine learning, as it allows the selection of relevant predictors for your model and avoids multicollinearity.

3. Outlier Detection

Effective outlier detection is a crucial phase in exploratory data analysis. Using tools such as box plots or scatter plots can help recognize anomalies in your data. Predictive modeling greatly benefits from a clean dataset, free of outliers. Specialized algorithms for outlier detection can be incorporated into your analysis process for further refinement and optimization.

4. Multivariate Analysis

Multivariate data analysis is a crucial aspect of exploratory data analysis. It utilizes statistical methods to help decipher patterns or relationships between multiple data



	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

```
Cust_No              int64
Cust_Name           object
Product_id           int64
Product_cost       float64
Purchase_Date  datetime64[ns]
dtype: object
```

4. Multivariate Analysis

Multivariate data analysis is a crucial aspect of exploratory data analysis. It utilizes statistical methods to help decipher patterns or relationships between multiple data variables simultaneously. This technique offers a profound insight into the dataset, helping knowledge workers make informed decisions.

5. Time Series Visualization

Decoding patterns and seasonal trends through time series visualization is a powerful component. With interactive tools, knowledge workers can visually examine fluctuations over a period, facilitating the recognition of anomalies and forecasting.

6. Dimensionality Reduction

To handle complex datasets with a large number of features, efficient dimensionality reduction techniques are critical. Techniques such as PCA and t-SNE offer effective ways to reduce data dimensionality, speeding up the learning process without losing important information.

Laboratorios con Colab

IIEG
Instituto de Información
Estadística y Geográfica
de Jalisco

Titanic Pandas Profile

EDA datos autos

Iris cleanlab



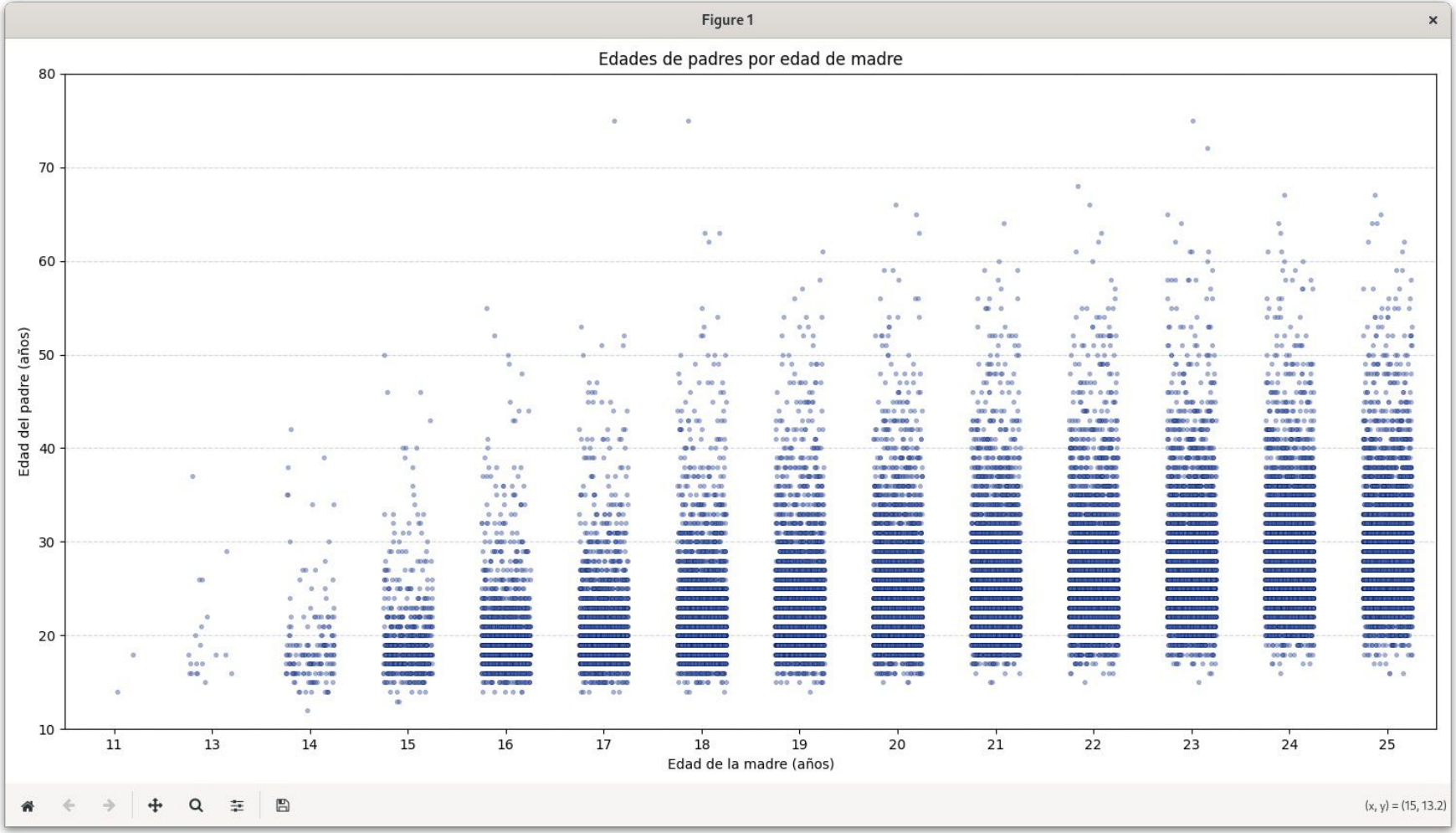
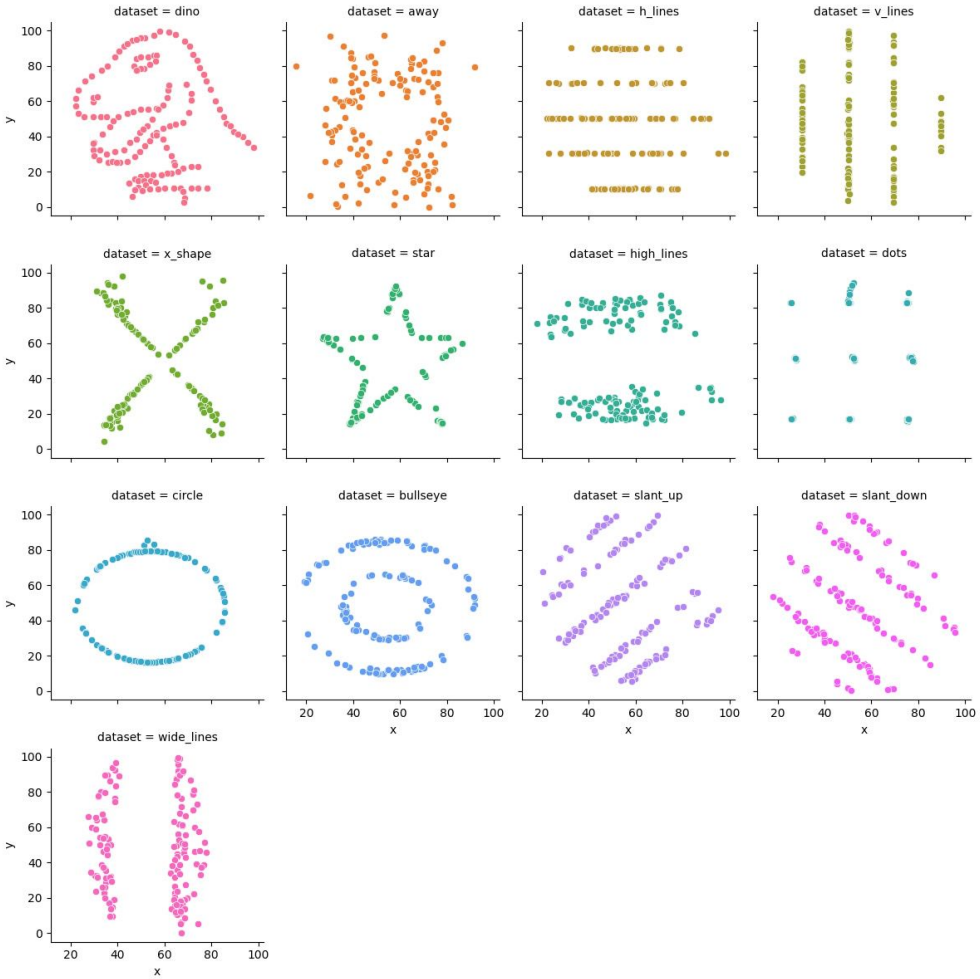
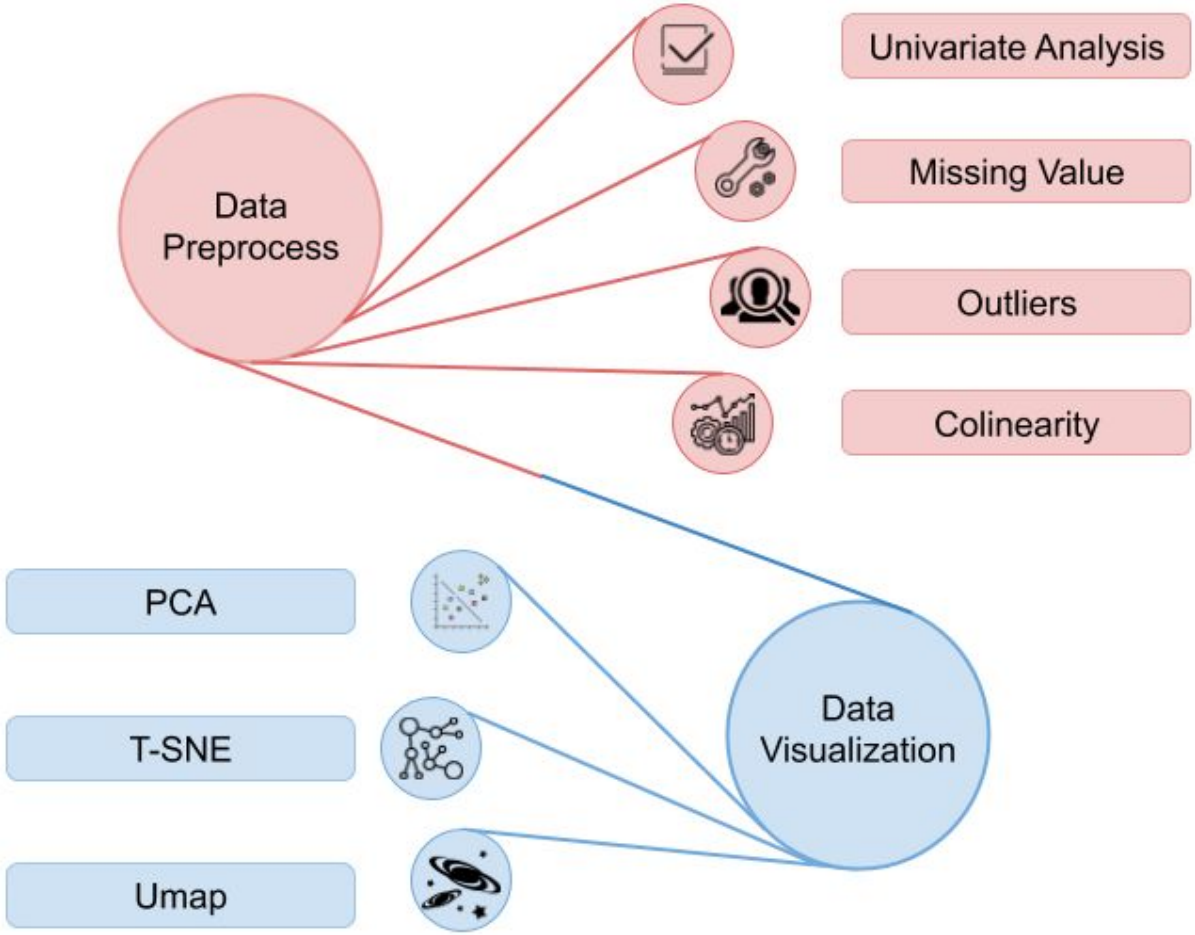


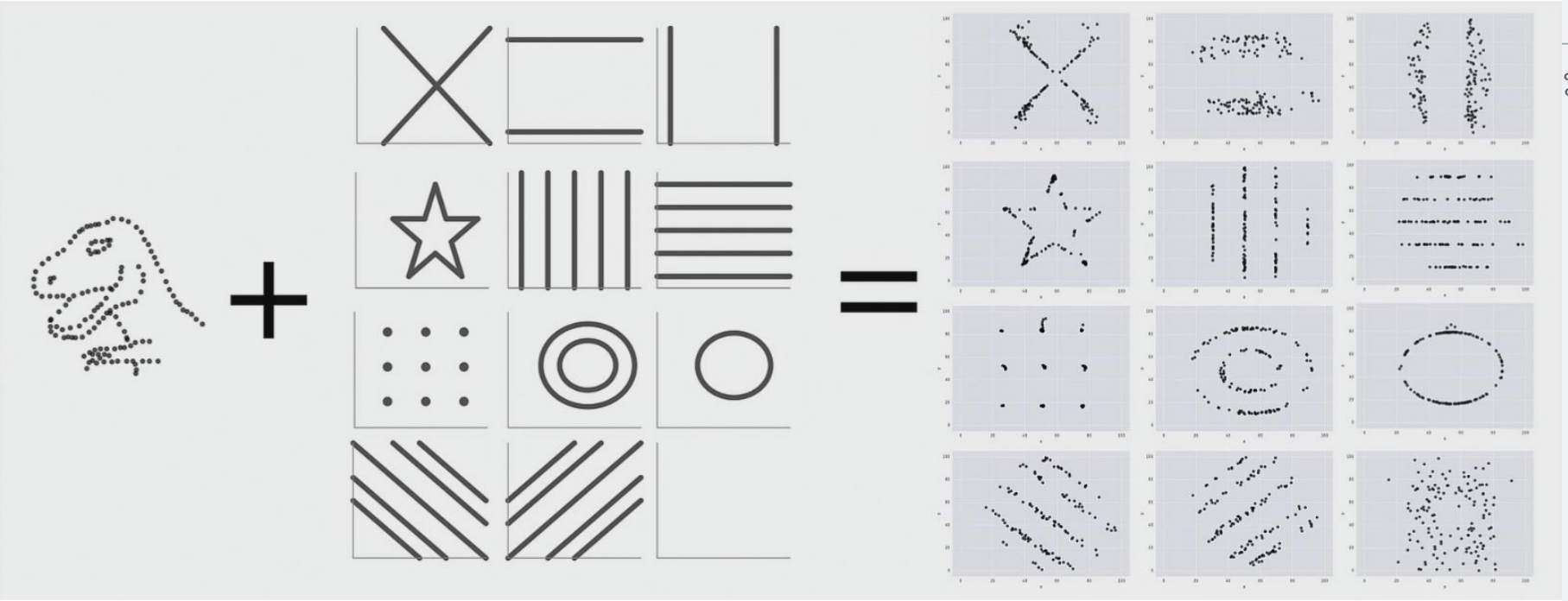
Figure 1
Edades de padres por edad de madre
Edad del padre (años)
Edad de la madre (años)
80
70
60
50
40
30
20
10
11
13
14
15
16
17
18
19
20
21
22
23
24
25
(x, y) = (15, 13.2)

Laboratorio Colab

https://colab.research.google.com/drive/1MKcGODLRthj9Hximu6W7VEiWWBVz91TU?usp=sharing




Data Preprocess
Univariate Analysis
Missing Value
Outliers
Colinearity
Data Visualization
PCA
T-SNE
Umap



https://www.research.autodesk.com/publications/same-stats-different-graphs/

Taller de datos

- Acomodar los datos, nombres de columna eliminar encabezados, imágenes(Casi dejar como CSV)
- Explorar:
  - Buscar los datos faltantes (reemplazarlos, eliminarlos),
  - Describir los tipos de datos y en su caso transformarlos por ejemplo enteros, o fechas.
  - Visualizar los datos.
  - Resolver problemas.

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Posibles errores:

- Archivado inaccesible
- Eliminación de datos por error
- Incumplimiento de regulaciones



05 Archivado/ Eliminación

Almacenamiento de larga estancia o se borran

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Clasificación

Clasificación de Datos

- **Categorizar por**:
  - **Sensibilidad** (datos personales, financieros, de salud).
  - **Frecuencia de uso** (activos, inactivos, históricos).
  - **Requisitos legales** (tiempos de retención obligatorios).
- **Herramientas**: Usar etiquetado automático



05 Archivado/ Eliminación

Almacenamiento de larga estancia o se borran

Apache Ranger

AWS Macie

opendlp

https://colab.research.google.com/drive/1EJVRgHrCqQCVVIT4099WDx0Mg1V1Ktw-?usp=sharing

https://github.com/microsoft/presidio

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Políticas de Archivado

- **¿Cuándo archivar?** Cuando los datos son inactivos pero aún necesarios para cumplimiento o análisis histórico.
- **Mejores prácticas**:
  - **Mover a almacenamiento frío** (ej: Amazon Glacier, Azure Blob Archive).
  - **Comprimir y cifrar** antes de archivar.
  - **Documentar metadatos** (fecha, responsable, razón del archivado).
  - **Automatizar procesos** con herramientas



05 Archivado/ Eliminación

Almacenamiento de larga estancia o se borran

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Herramienta	Tipo	Costo	Caso de Uso Típico
AWS Glacier	Cloud	Bajo (pago por uso)	Archivado masivo en la nube.
MinIO	On-Prem/Cloud	Open-Source	Reemplazo de S3 Glacier on-premise.
Veritas Enterprise Vault	On-Prem	Comercial	Archivado corporativo con cumplimiento.
Archivematica	On-Prem	Open-Source	Preservación digital a largo plazo.
PostgreSQL Partitioning	On-Prem	Open-Source	Archivado de tablas grandes.

Ciclo de vida de los datos

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

Cumplimiento Legal y Auditoría

- **Regulaciones clave**:
  - **GDPR**: Derecho al olvido (Art. 17).
  - **CCPA**: Eliminación de datos personales bajo solicitud.
  - **HIPAA**: Retención de registros médicos por 6 años.
  - **ARCO**
- **Documentar todo**: Mantener registros de cuándo, cómo y por qué se archivaron/eliminaron datos.
  -



05 Archivado/ Eliminación

Almacenamiento de larga estancia o se borran

Ciclo de vida de los datos

1. **A – Acceso**
   - Derecho a conocer qué datos personales están siendo procesados, para qué fines, y con quién se comparten.
   - Ejemplo: Solicitar a un banco una copia de tus datos registrados.
2. **R – Rectificación**
   - Derecho a corregir datos personales inexactos, incompletos o desactualizados.
   - Ejemplo: Actualizar tu dirección en una plataforma de e-commerce.
3. **C – Cancelación** (o **Supresión**)
   - Derecho a eliminar datos personales cuando ya no sean necesarios para los fines para los que fueron recopilados o cuando retires tu consentimiento.
   - Ejemplo: Solicitar a una red social borrar tu cuenta y datos asociados.
4. **O – Oposición**
   - Derecho a oponerse al tratamiento de tus datos para fines específicos (ej: publicidad, prospección comercial).
   - Ejemplo: Negarse a recibir correos promocionales de una tienda.

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

05 Archivado/ Eliminación

Almacenamiento de larga estancia o se borran

IIEG
Instituto de Información Estadística y Geográfica de Jalisco

04 Referencias

Referencias y recomendaciones

https://github.com/GeostatsGuy/DataScienceInteractivePython

https://dcai.csail.mit.edu/

https://www.datascience-pm.com/

https://microsoft.github.io/Data-Science-For-Beginners/#/

https://cognitiveclass.ai/

https://www.bbc.com/news/technology-54423988

https://sheetcast.com/articles/ten-memorable-excel-disasters

Gracias



IIEG
Instituto de Información Estadística y Geográfica de Jalisco