Introducción a la Ciencia de Datos

Maestría en Ciencias de la Computación

Dr. Irvin Hussein López Nava

ANIVERSARIO CICESE

CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

Presentación del curso

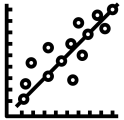
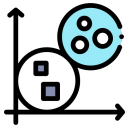
Objetivo del curso

El **curso** busca proporcionar al estudiante **herramientas** que le permitan iniciar en el área de **ciencias de datos**, en particular, generar **modelos de aprendizaje** basados en grandes volúmenes de datos.

Pretende también que el estudiante entienda y maneje los métodos de **estadística** e **inteligencia artificial** más empleados.

Temario

Unidad 1	2	3	4	5	6	7
Introducción	Procesamiento de datos	Regresión	Aprendizaje no supervisado	Aprendizaje supervisado	Meta aprendizaje	Aprendizaje profundo

Evaluación

20%

Lecturas y tareas

+

50%

Prácticas

+

30%

Proyecto final

- Las lecturas, tareas y prácticas son obligatorias y deben ser completadas en tiempo y forma.
- Por cada lectura deberá crearse un reporte de máximo una cuartilla.
- Para cada práctica debe crearse un Notebook y deberá agregarse al repositorio personal.
- El reporte del proyecto deberá describir el trabajo de investigación en formato de artículo.

Timeline del curso

66% inscritos

2020

Satellite Imagery Classification Using Shallow and Deep Learning Approaches

Michelle Saisse-Vinett[SM] and Irvin Hussein Lopez-Nava

Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), 22860 Ensenada, Baja California, Mexico
mzinn@cicese.edu.mx, hussein@cicese.mx
https://www.cicese.edu.mx/

Abstract. Recent advances in remote sensing technology and high-resolution satellite imagery offer great possibilities for understanding the earth's surfaces. However, satellite image classification is a challenging problem that involves the analysis of complex patterns in spatial and spectral space, two learning approaches are proposed and compared for classifying a large-scale dataset including different types of land-use and land-cover surfaces (Eurosat). Traditional (shallow) machine learning models and deep learning models are built by using a set of features extracted from the satellite images for both approaches and using the RGB images for deep models. The best F1-score obtained by the shallow approach was 0.87, while for the deep approach it was 0.91. No significant difference was found in these results; however, significant improvements can be made by exploring the deep approach in greater depth.

Keywords: Satellite images · Satellite imagery classification · Image classification · Land use classification · Land cover classification · Remote sensing · Deep learning

2021

60% inscritos

Fingerspelling recognition in Mexican Sign Language (LSM) using Machine Learning*

Abstract. Sign languages allow deaf people to express their thoughts, feelings, and ideas through a combination of hand shapes, movements, and languages. Each sign language is unique and has its own grammar, syntax, and vocabulary. Mexican Sign Language (LSM) is characterized by rich gestural and facial expression that gives it a great communicative and linguistic capacity. In the study of LSM, two main components have been identified: (i) fingerspelling, and (ii) ideograms. The first is similar to spelling in oral languages, and is used to communicate proper nouns, technical terms, and words for which there is no ideogram, which are little known to the deaf community. In this paper we propose a method for recognizing the LSM alphabet by using machine learning-based techniques capable of classifying the signs made by 10 test subjects. 21 keypoints of the hands were extracted from the MediaPipe library, in order to have a better representation to feed the classification models. The results when classifying the 21 letters ascended an F1-score of 0.08 when using a Random Forest classifier, and when classifying only the first 3 letters. Tools such as those proposed in this work can facilitate seamless communication by translating Spanish into LSM and vice versa, allowing both communities to engage effectively in various settings.

Keywords: LSM · Sign language recognition · Sign language classification · Mexican Sign Language · Fingerspelling · Dactylology

No se ofertó

2022

Towards Detection of Depression and Anxiety in Adolescents through Social Media Texts

Abstract. Anxiety and depression represent widespread mental health challenges encountered by a substantial proportion of young individuals, carrying profound implications for their overall well-being. Social media has become an integral part of the daily lives of young people, providing a unique opportunity to identify potential indicators of anxiety and depression. This study focuses on exploring the presence of both anxiety and depression in adolescents through the analysis of data collected from social media comments. To achieve this, a series of classification models have been trained and evaluated, combining different n-grams representations, feature selection techniques, and machine learning algorithms. Bigrams emerged as the most effective, enabling the identification of word pairs such as 'feel like', 'feel restless', and 'every day', commonly indicative of anxiety and depression. The Random Forest model with bigrams achieved an F1-score of 0.836 using feature selection. Consequently, these results shed light on significant findings regarding the connection between sleep problems, anxiety, and depression. The present study is oriented towards the comprehensive understanding of the prevalence, contributing factors, and potential manifestation of anxiety and depression in the youth population.

Keywords: Anxiety · Depression · Text classification · Social media.

2023

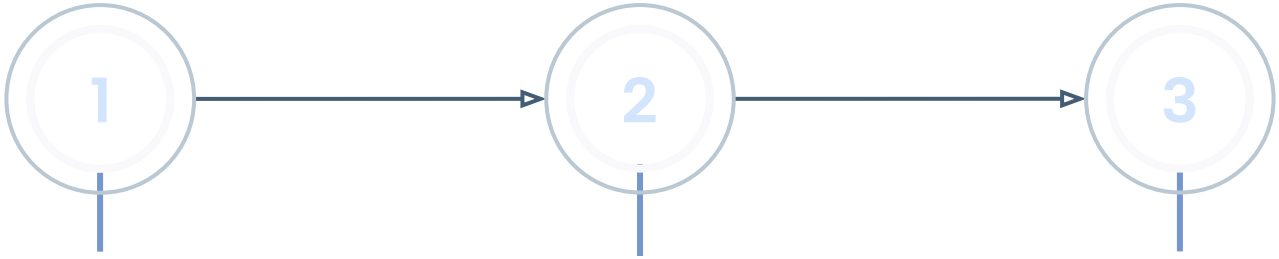
73% inscritos

Antecedentes 2023

Notebooks

Videos

Secuencia de cursos



Introducción a la Ciencia de Datos

Ciencia de Datos para Sensado Inteligente

Reconocimiento de Patrones

Recursos

3sspvfl

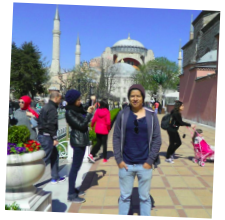
GitHub

jupyter

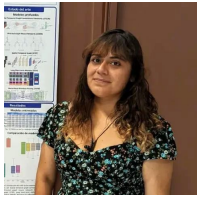
scikit learn

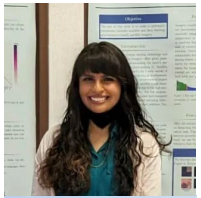
Acerca de mí

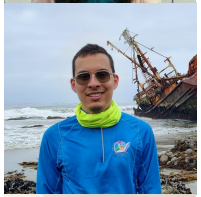
data
gait
activity
sensors
analysis
classification
human
recognition
features
activities
participants
time
sensor
acceleration
wearable
system
signals
subjects
parameters
model
algorithm
accelerometer
device
people
dataset
systems
devices
models
inertial
movements
motion
actions
research
training
detection
dementia
learning
upper
processing
joints
technology
algorithms
peaks
instances
physical
class
action
health
according
classification
exergames
intervention
evaluation
dynamic
measure
estimation
rate
images
mean
place
classifier
older
set
hand
annotation
position
best
variability
methods
sound
function
monitoring
associate
social
anxiety
behavior
method
test
adult
living
signal
values
process
groups
tower
joint
high
care
respect
balance
wii
age
vector
sessions
application
provide
driving
orientation
movement
accuracy
walking
caregivers
cognitive
daily
subject
places
network
participant
labeling
environment
environments
right
characteristic
review
classes
feature
forearm
ambient
case
group
dementia
function
sound
methods
social
monitoring
associate
anxiety
hand
annotation
position
best
set
place
rate
images
mean
classifier
older
measure
dynamic
estimation
variability
technology
processing
joints
symptoms
algorithms
peaks
instances
level
classifier
physical
arm
class
action
health
according
classification
exergames
important
intervention
evaluation
dataset
algorithm
model
people
signal
behavior
method
test
adult
living
values
process
groups
tower
joint
high
care
respect
balance
wii
age
vector
sessions
application
provide
driving
orientation
movement
accuracy
walking
caregivers
cognitive
daily
subject
places
network
participant
labeling
environment
environments
right
characteristic
review
classes
feature
forearm
ambient
case
group
dementia
function
sound
methods
social
monitoring
associate
anxiety
hand
annotation
position
best
set
place
rate
images
mean
classifier
older
measure
dynamic
estimation
variability
technology
processing
joints
symptoms
algorithms
peaks
instances
level
classifier
physical
arm
class
action
health
according
classification
exergames
important
intervention
evaluation

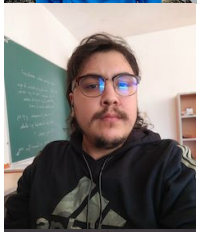


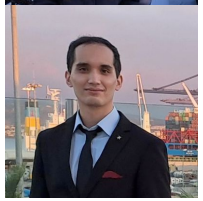
My team

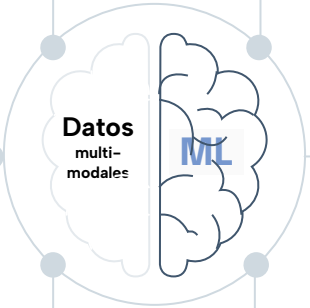












Marcha humana

Lengua de Señas

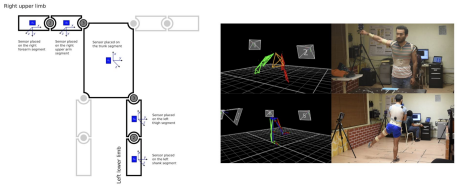
Somnolencia

Astronomía

Redes sociales

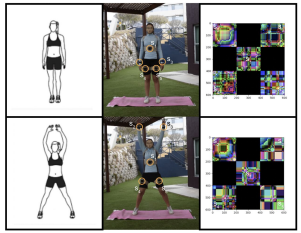
Perros + emociones

Interacción humano-robot

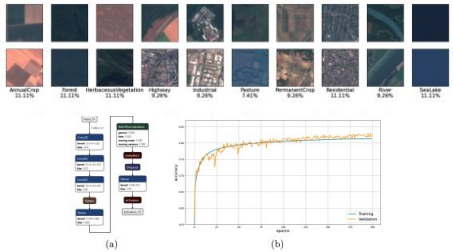


Seguimiento del movimiento humano con sensores inerciales vestibles

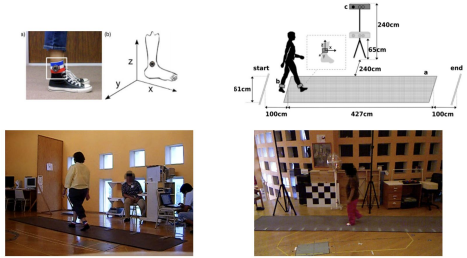
Reconocimiento de ADL's con sensores vestibles en escenario naturalista



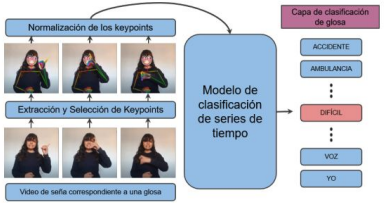
Reconocimiento de actividades físicas con sensores vestibles y video



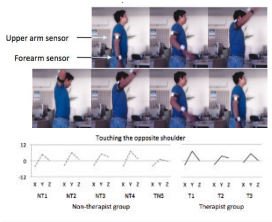
Clasificación de imágenes satelitales utilizando aprendizaje automático



Estimación de parámetros espacio-temporales de la marcha humano con sensores y video



Reconocimiento de glosas de la Lengua de Señas Mexicana

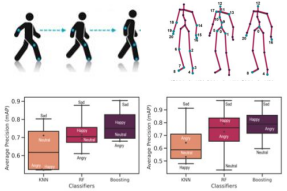


Variabilidad en movimientos terapéuticos con sensores inerciales

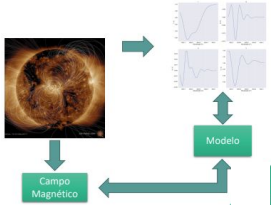
Detección de eventos violentos en publicaciones de redes sociales

Detección de obstáculos con sensores inerciales vestibles

Detección de somnolencia durante la conducción en un escenario simulado



Reconocimiento de emociones a partir de patrones en la marcha humana



Medición del campo magnético estelar con aprendizaje profundo

Acerca de ustedes

- Formación académica y procedencia
- Intereses de investigación
- Interés en el curso
- Algo extra-académico

01

Introducción

1.1 La Ciencia de Datos

DATA SCIENCE
DETECTION
BIG DATA
COMPUTING
INFORMATION
TECHNOLOGY
SOCIAL MEDIA
PROMOTION
MODELS
E-MARKETING
COMMUNICATION
WWW
SERVICES
CONTENT
PROCESSING
BRANDING
COMPUTER
MULTIMEDIA
NETWORK
PROJECTS
CONSUMER
CONSUMER DEMAND
MARKETS
ORGANIZATION
SOFTWARE
PREDICTIVE
PLANNING
WEB MARKETING
PROGRAM
ANALYTICS
EVENTS
PROGRAMMING
DATA MINING
MACHINE LEARNING
VISION
ENGINEERING
RESEARCH
INFORMATION
WEB SERVICES
BIG
KDD
PROBABILITY
WEB DEV
COMPUTING
MATHS PATTERN
PROJECTS
DATA
STRATEGY
WORLDWIDE
SERVICE
VISUALIZATION
BIG
DATA
ENGINEERING
SOLUTIONS
PLANNING
MEDIA
MOBILE
PRO
PRICING
CODING
SEGMENTATION
SOCIAL
NETWORK
STATISTICS
INFORMATION
SOCIAL NETWORKS
TARGET
DIGITAL

¿Qué es
la Ciencia
de Datos?

Definiciones

ChatGPT

Campo interdisciplinario que involucra métodos, procesos, algoritmos y sistemas para obtener conocimiento y perspicacia a partir de datos en diversas formas

https://chat.openai.com/

Wikipedia

Campo académico interdisciplinario que utiliza estadística, computación científica, métodos, procesos, algoritmos y sistemas científicos para obtener, tratar, analizar y presentar informes a partir de datos ruidosos, estructurados y no estructurados.

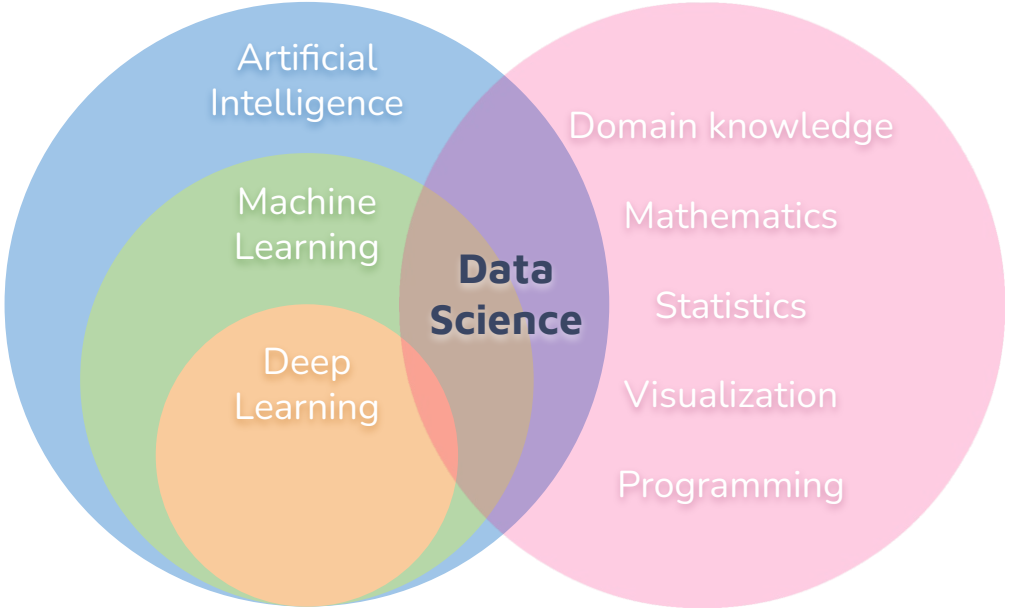
https://es.wikipedia.org/wiki/Ciencia_de_datos

Science Direct

La Ciencia de Datos proporciona procesos y sistemas para extraer y gestionar conocimientos a partir de fuentes de datos estructuradas y no estructuradas.

https://www.sciencedirect.com/topics/social-sciences/data-science

Campos relacionados



El término “Ciencia de Datos”

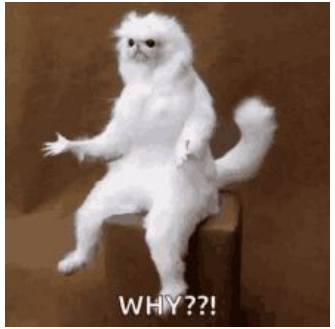
- El término “**ciencia**” implica **conocimiento** adquirido a través del **estudio sistemático**.
- En una definición: trabajo **sistemático** que construye y organiza el **conocimiento** en forma de explicaciones y predicciones.

Por lo tanto, la **ciencia de datos** podría implicar:

un enfoque que involucre **datos** y, por extensión, **estadística;**

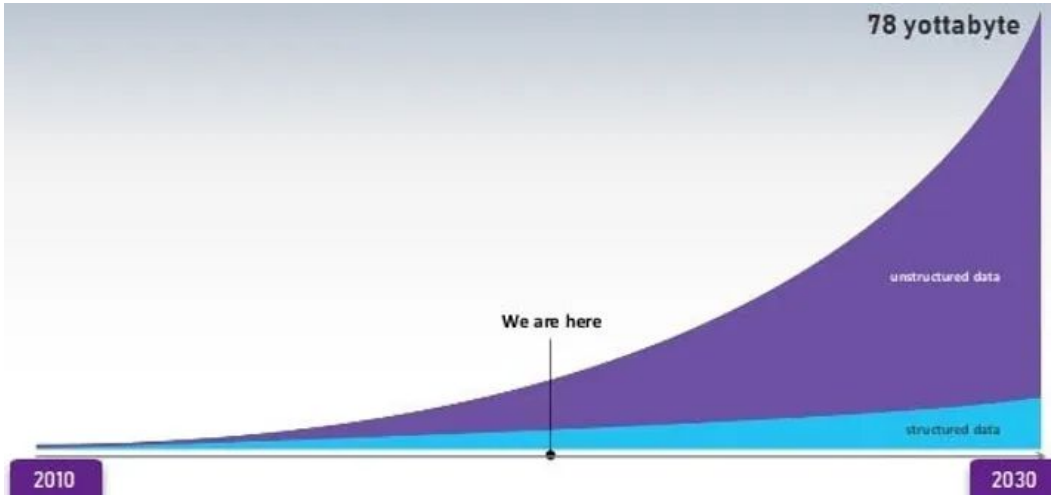
o el **estudio sistemático** de la organización, propiedades, y análisis de datos y su papel en la **inferencia**, incluida la confianza misma en la inferencia

¿Por qué el término “**ciencia de datos**” cuando se ha tenido la “**estadística**” durante siglos?



Necesidad de este término

Primero: la materia prima, la parte de "**datos**", es cada vez más heterogénea y desestructurada: texto, imágenes, video...

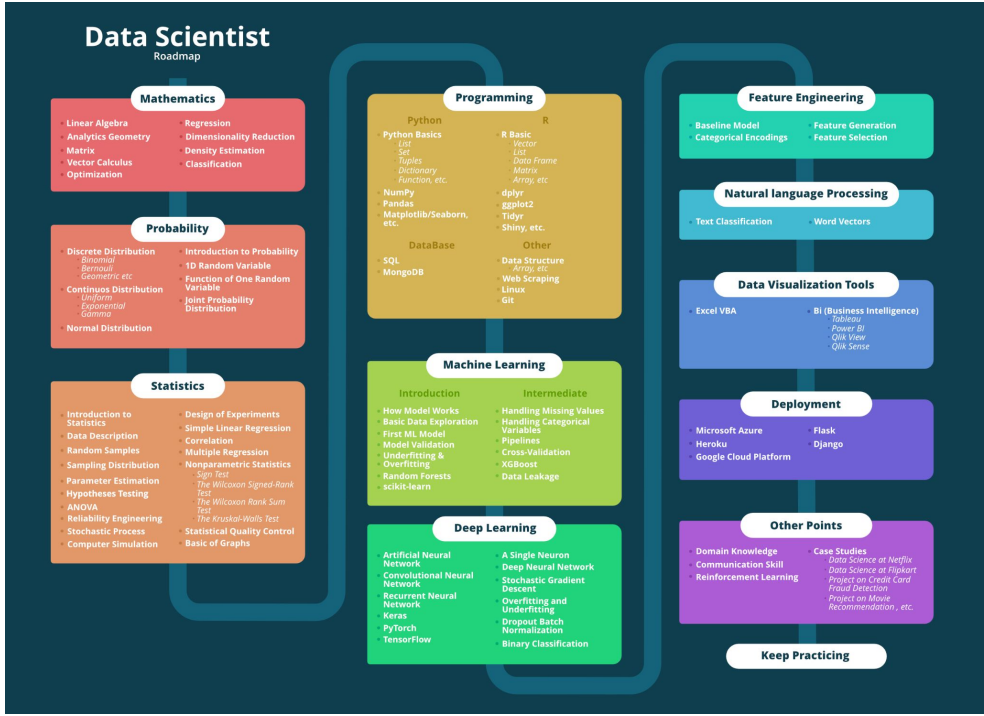


https://www.d2ktech.com/big-data

Necesidad de este término

Segundo: la proliferación de **lenguajes** diseñados para que las máquinas interpreten los **datos** automáticamente.

En otras palabras, las computadoras están haciendo cada vez más el trabajo de fondo.

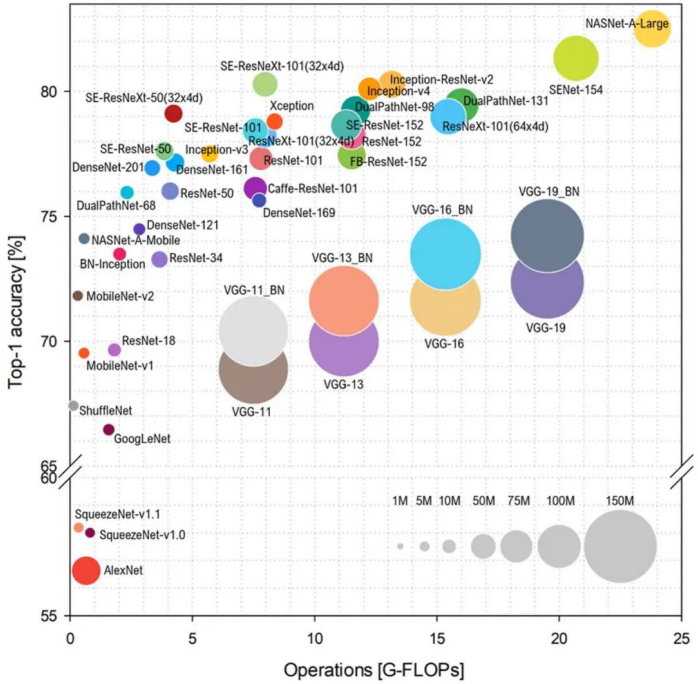


https://www.d2ktech.com/big-data

Necesidad de este término

Tercero: el hardware de última generación que permite procesar grandes cantidades de información.

Algoritmos computacionales que fueron propuestos décadas atrás y que hoy en día pueden ser utilizados.

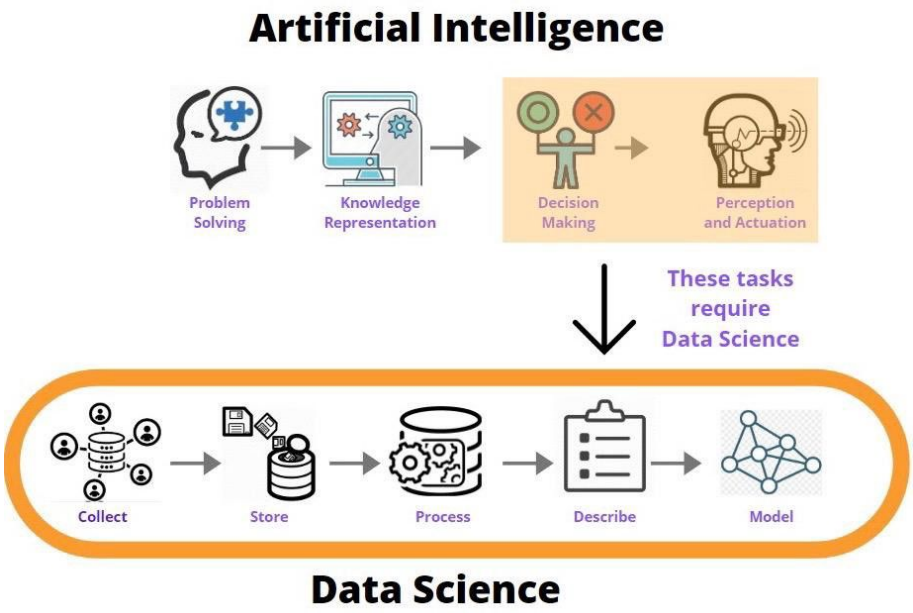


https://deci.ai/blog/close-gap-cpu-performance-gpu-deep-learning-models/

Necesidad de este término

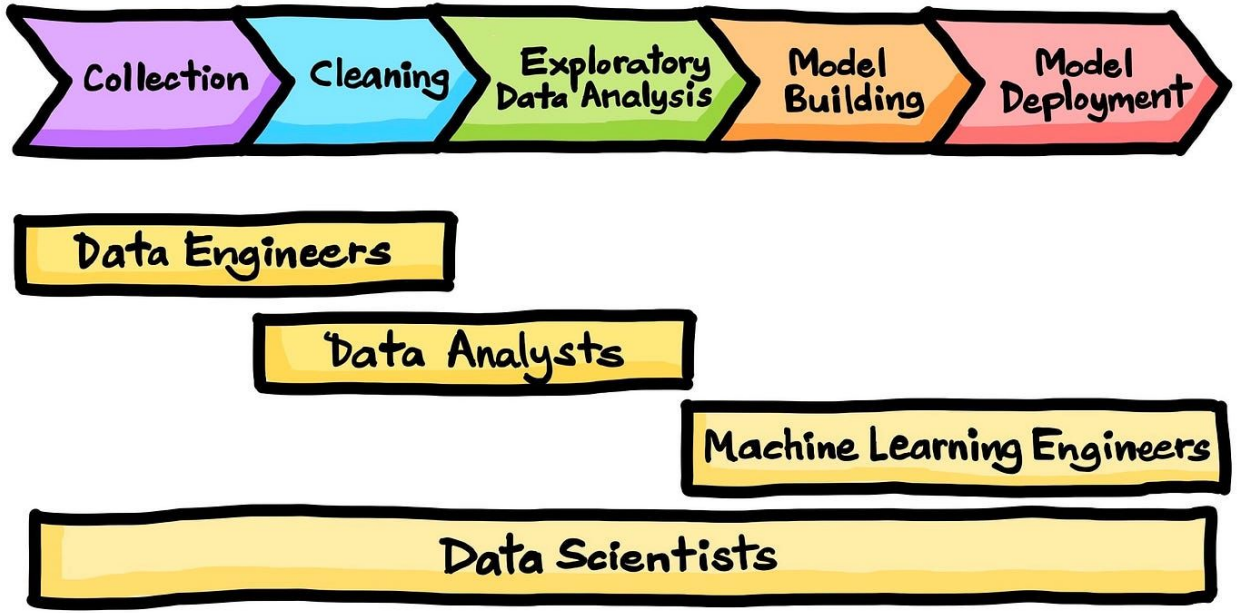
Por último, la **toma de decisiones**: Cada vez es más común que la computadora tome decisiones, sin la ayuda de humanos.

Sin embargo, hay una multitud de **problemas** que van desde los costos de decisiones incorrectas hasta problemas éticos y de privacidad.



https://towardsdatascience.com/data-science-vs-the-world-of-ai-793f6da8cf86

El proceso de la Ciencia de Datos



Secuencia de cursos

1 Introducción a la Ciencia de Datos

2 Ciencia de Datos para Sensado Inteligente

3 Reconocimiento de Patrones

¿Para qué se utiliza?

Análisis **descriptivo**

Examina los datos para obtener información sobre **lo que ha** ocurrido u ocurre en el entorno de los datos.

Análisis de **diagnóstico**

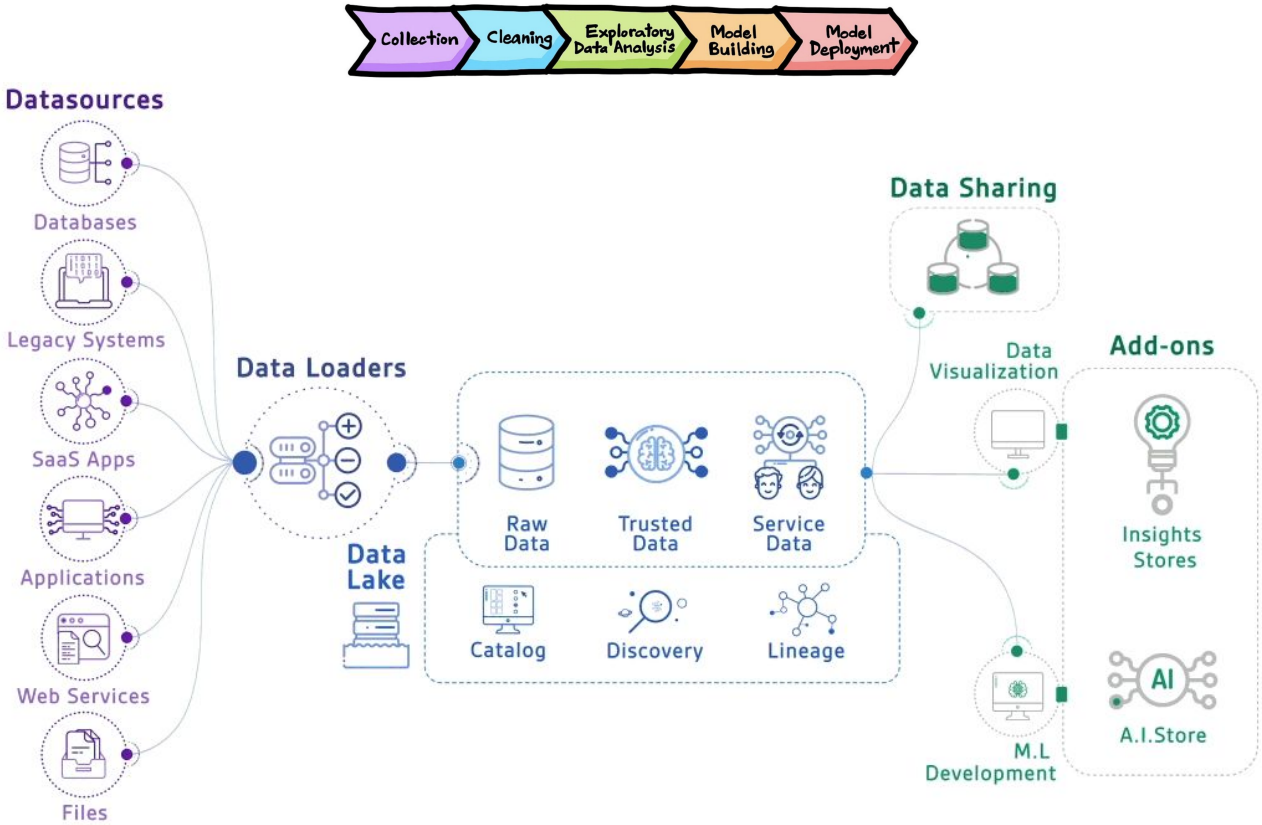
Es un análisis detallado de los datos para entender **por qué ha** ocurrido algo.

Análisis **predictivo**

Utiliza datos históricos para hacer **previsiones** sobre los patrones de datos que pueden producirse en el futuro.

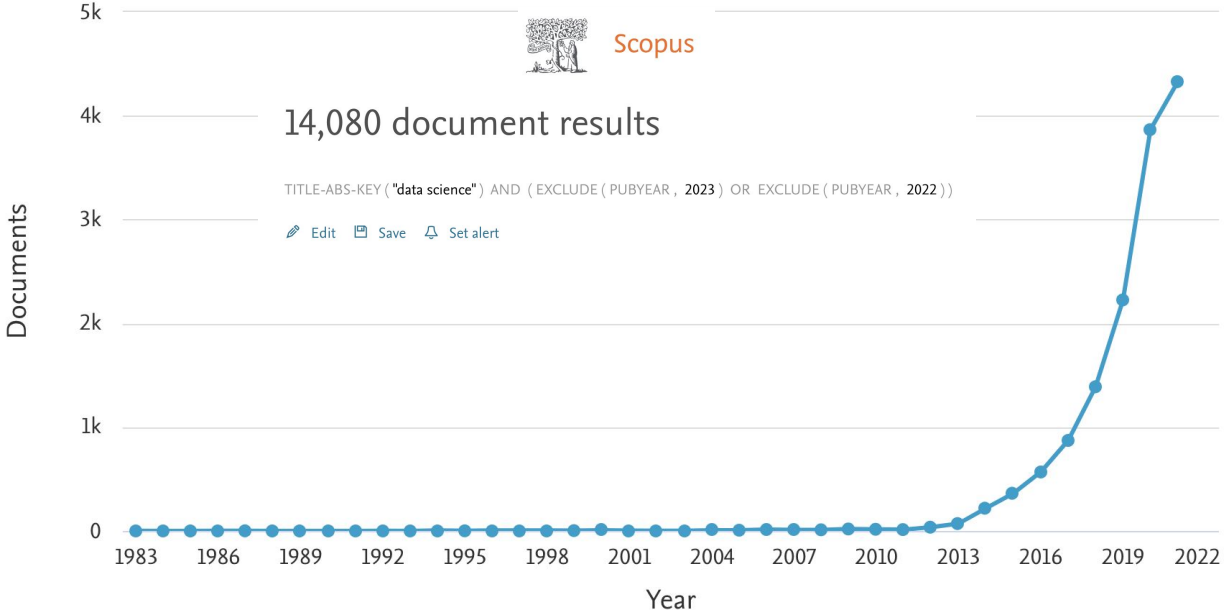
Análisis **prescriptivo**

No solo predice lo que es probable que ocurra, sino que sugiere una respuesta **óptima** para ese resultado.

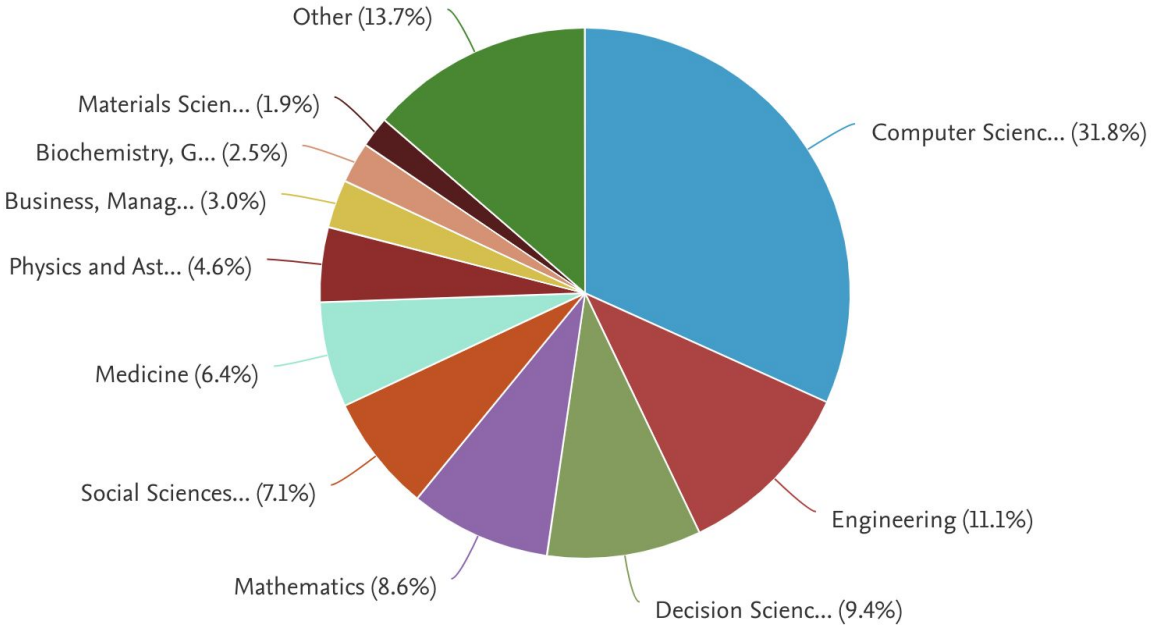


https://www.datasciencecentral.com/data-management/

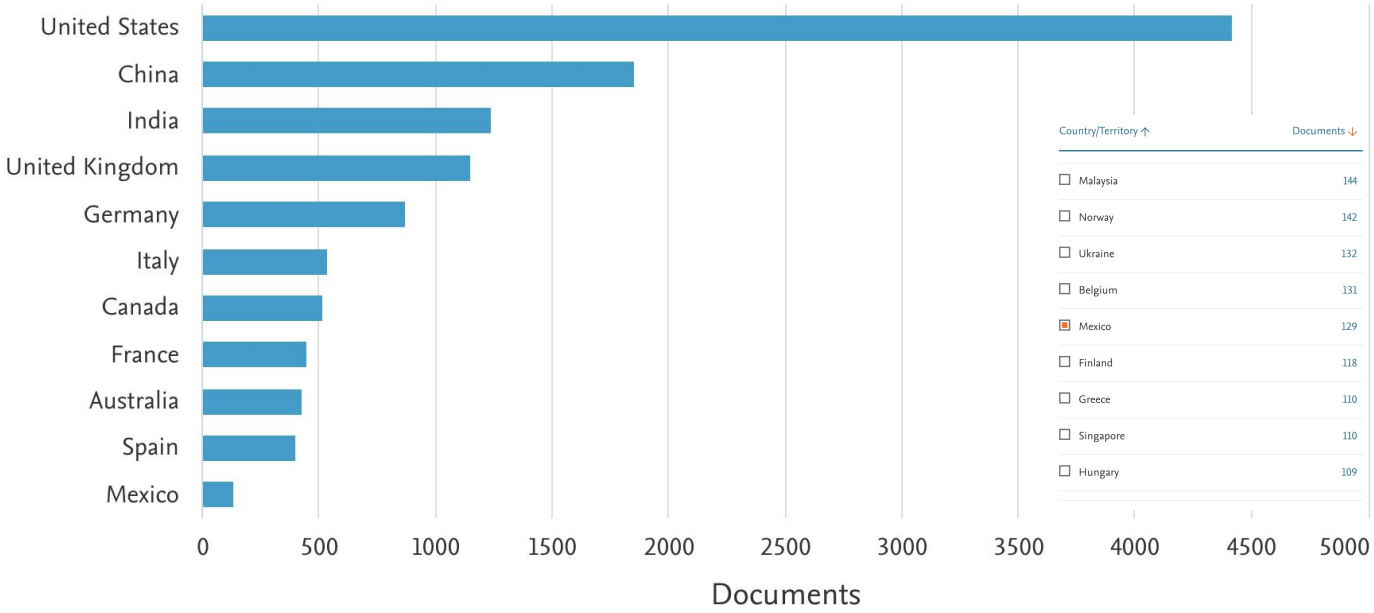
Publicaciones científicas



Publicaciones por área



Publicaciones por país



Country/Territory	Documents
Malaysia	144
Norway	142
Ukraine	132
Belgium	131
Mexico	129
Finland	118
Greece	110
Singapore	110
Hungary	109

data
big
analytics
science
machine
learning
research
network
algorithm
management
information
engineering
classification
case
method
intelligent
workshop
models
networks
approach
applications
knowledge
visualization
computing
development
technology
health
medicine
software
discovery
monitoring
quality
healthcare
challenges
detection
model
digital
data-driven
sciences
covid-19
statistics
evaluation
predictive
education
systems
system
process
design
framework
deep
neural
prediction
review
mining
artificial
cloud
clinical
support
new
medical
smart
clustering
social
time
computer
cancer
future
study
intelligence
application
techniques
modeling
algorithms
industry
business
assessment
platform
training
graph
recognition
novel
students
tool
human
statistical
implementation
building
dynamic
precision
scientists
skills
automated
global
security
era
decision
technologies
approaches
text
power
natural
impact
materials
tools
diagnosis
making
convolutional
regression
fuzzy
field
generation
driven
educational
toward
role
curriculum
overview
service
perspectives
supervised
reinforcement
reproducible
bioinformatics
blockchain
decision-making
cardiovascular
behavioral
utility
probabilistic
uncertainty
correction
implementing
reconsider
recommender
death
cancer
cardiac
graphs
help
modulating
methodology
operations
states
structure
python
water
traffic
mobile
energy
environment
user
exploring
innovation
evidence
workflows
multiple
food
distributed
industrial
datasets
communication
lot
can
comprehensive
ensemble
learning
semantic
sensor
understanding
modelling
things
academic
potential
finding
experience
factors
computational
patients
spatial
online
optimization
estimation
applied
extraction
ethics
context
drug
supply
patterns
visual
control
segmentation
transformation
simulation
exploration
scalable
learned
function
influence
bayesian
competition
advances
trends
informatics
urban
processing
policy
privacy
cities
dynamics
parallel
university
effective
disease
outcomes
series
via
strategies
problems
predicting
forecasting
opportunities
improving
services
interactive
efficient
advanced
methods
risk
care
studies
feature
automatic
distributed
multiple
food
large
sustainable
construction
environmental
distribution
customer
production
program
costs
issues
association
solving
sharing
mapping
integration
large-scale
designing
enhancing
ethical
projects
recent
winning
imaging
strategy
concepts
vehicle
quantitative
developing
planning
thinking
hierarchical
mathematical
investigating
quantum
analyzing
validation
solution
exploratory
multidimensional
relationship
government
analytical
people
operational
machines
values
computation
discussion
qualitative
crowdsourcing
recommendation
scientific
collaborative
project
industrial
communication
datasets
cities
dynamics
parallel
university
effective
disease
outcomes
series
via
strategies
problems
predicting
forecasting
opportunities
improving
services
interactive
efficient
advanced
methods
risk
care
studies
feature
automatic
distributed
multiple
food
large
sustainable
construction
environmental
distribution
customer
production
program
costs
issues
association
solving
sharing
mapping
integration
large-scale
design

What we are told…

DATA

SORTED

ARRANGED

PRESENTED VISUALLY

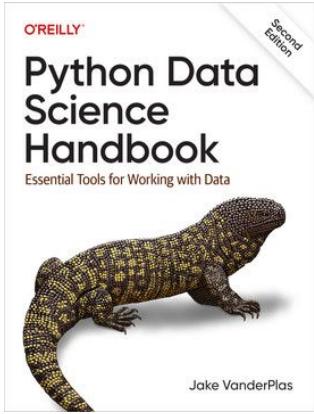
EXPLAINED WITH A STORY

ACTIONABLE (USEFUL)

What we actually have…

©2021 Dean of Big Data, LLC - Bill Schmarzo. All rights reserved.

Extra Libro



GitHub

- 00.00-Preface.ipynb
- 01.00-IPython-Beyond-Normal-Python.ipynb
- 01.01-Help-And-Documentation.ipynb
- 01.02-Shell-Keyboard-Shortcuts.ipynb
- 01.03-Magic-Commands.ipynb
- 01.04-Input-Output-History.ipynb
- 01.05-IPython-And-Shell-Commands.ipynb
- 01.06-Errors-and-Debugging.ipynb
- 01.07-Timing-and-Profiling.ipynb
- 01.08-More-IPython-Resources.ipynb

Gracias!

¿Alguna pregunta?

hussein@cicese.mx

https://sites.google.com/view/husseinlopeznava

SCAN ME

CREDITS: This presentation was based on a template by **Slidesgo**, and includes icons by **Flaticon**.