



Ciencia de datos para la producción de estadísticas

Área de Ciencia de Datos – Unidad
de Gobierno de Datos

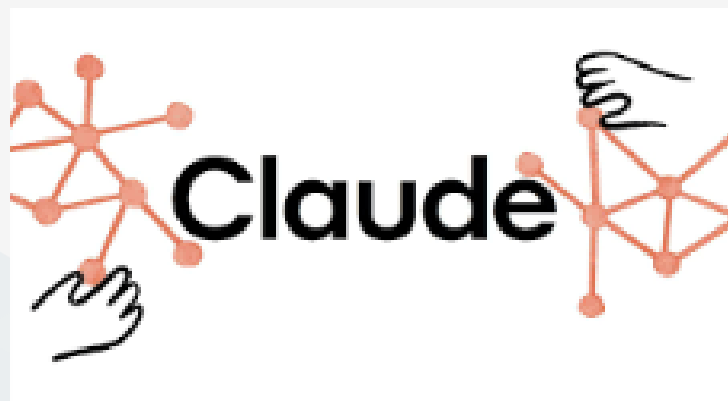
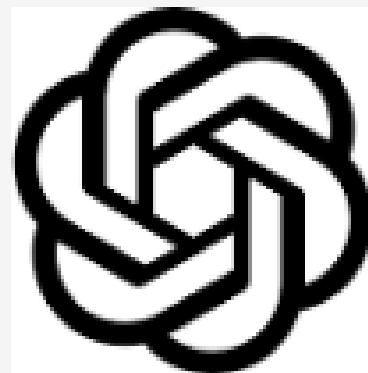
INSTITUTO NACIONAL DE ESTADÍSTICAS

Junio de 2025

El contexto: datos en todas partes





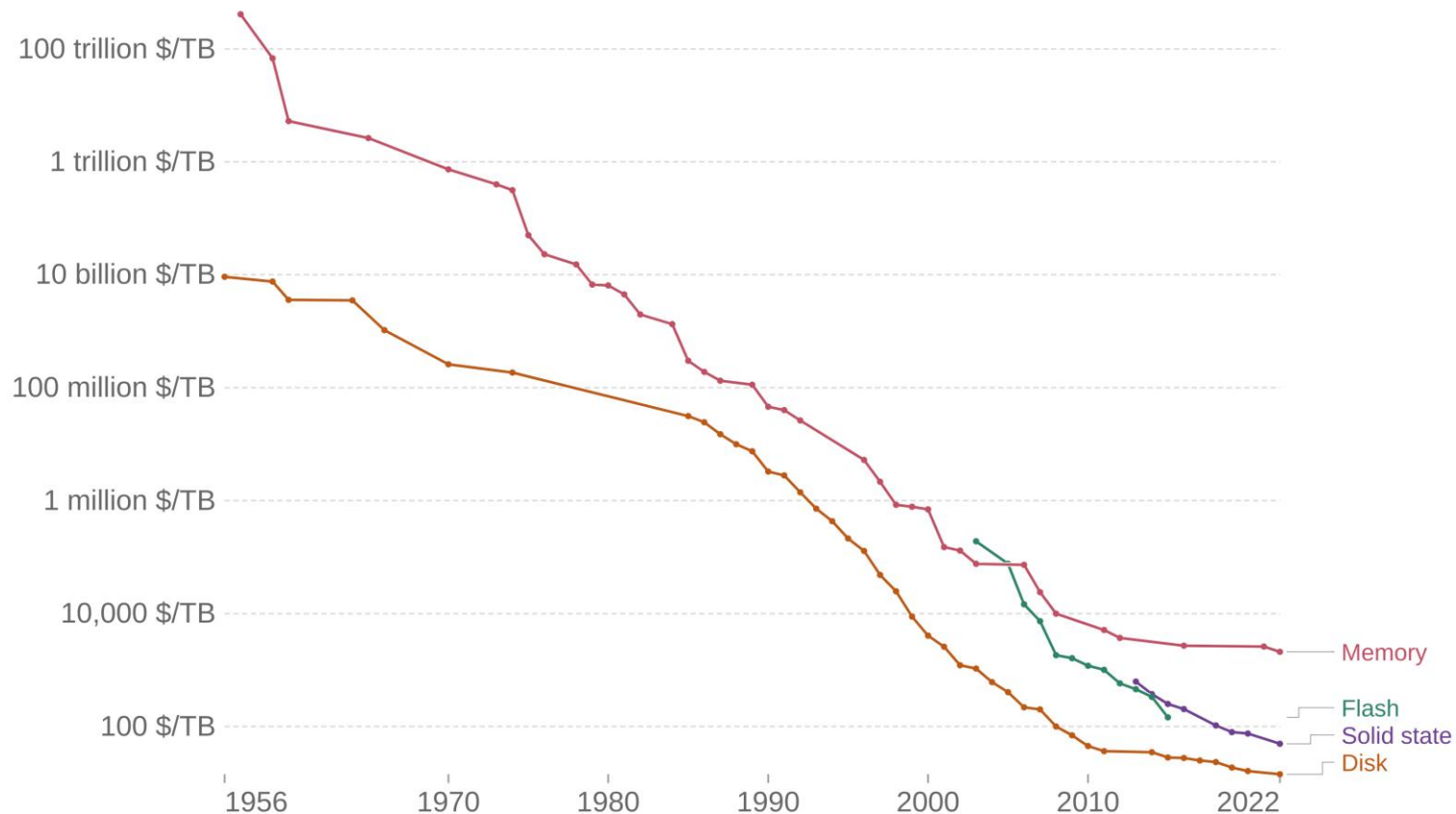


El contexto: el costo computacional

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.

Our World
in Data



Source: John C. McCallum (2022)

OurWorldInData.org/technological-change • CC BY

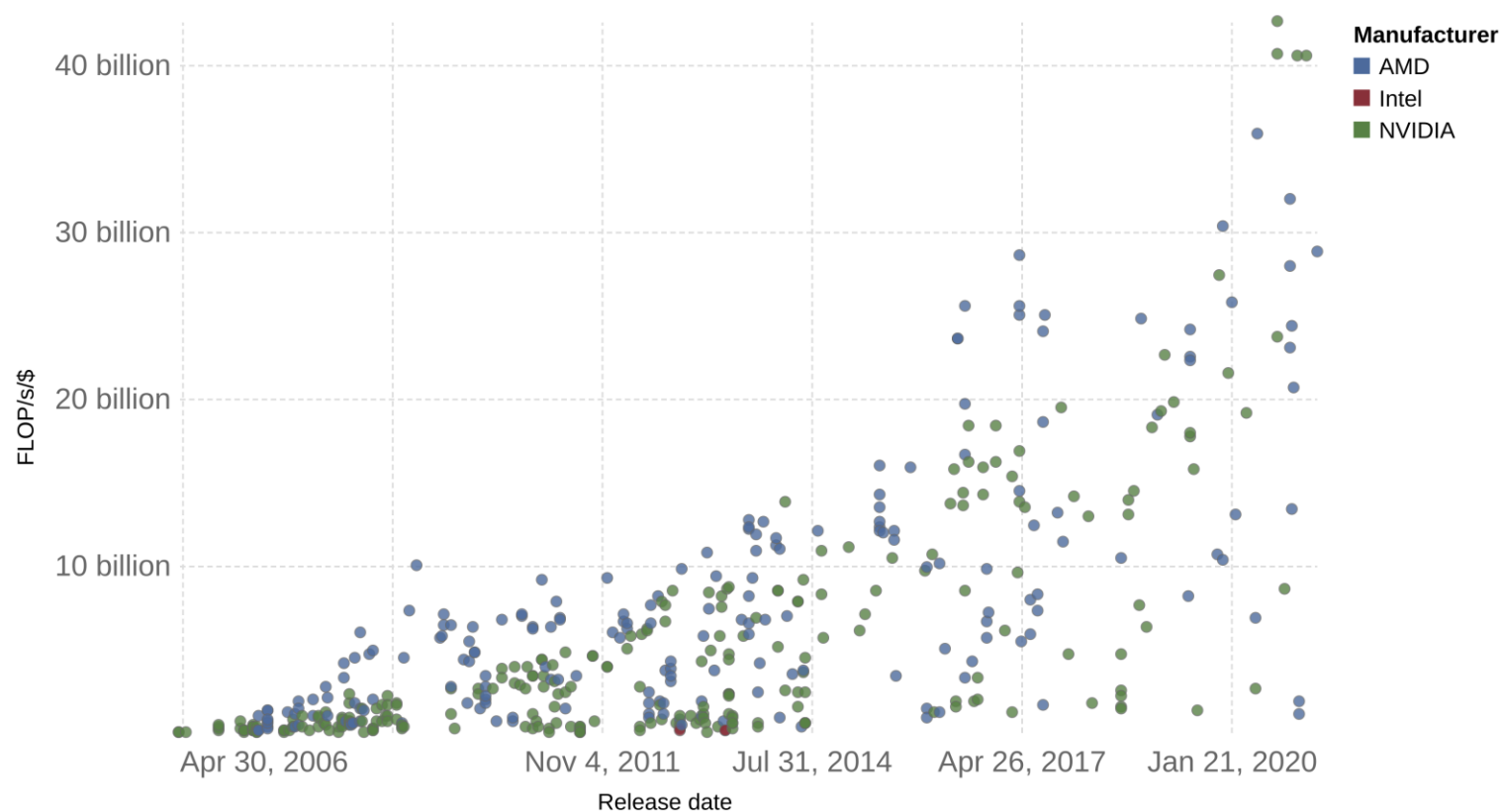
Note: For each year, the time series shows the cheapest historical price recorded until that year.

El contexto: el costo computacional

GPU computational performance per dollar

Graphics processing units (GPUs) are the dominant computing hardware for artificial intelligence systems. GPU performance is shown in floating-point operations¹/second (FLOP/s) per US dollar, adjusted for inflation.

Our World
in Data



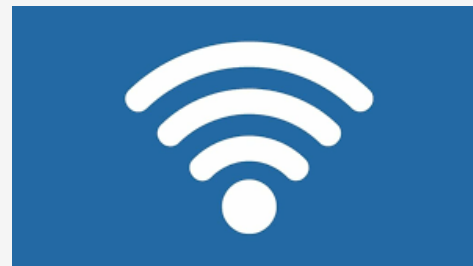
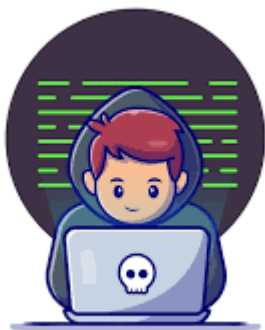
Source: Sun et al., Median Group via Epoch (2022)

Note: FLOP/s values refer to 32-bit (full) precision.

OurWorldInData.org/artificial-intelligence • CC BY

1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

El contexto: importancia del capital humano



Las **oficinas nacionales de estadística** deben adaptarse a esta nueva realidad



¿Qué significa exactamente adaptarse?

Combinando ambos mundos

Encuestas por muestreo

- Tecnología de recolección muy robusta

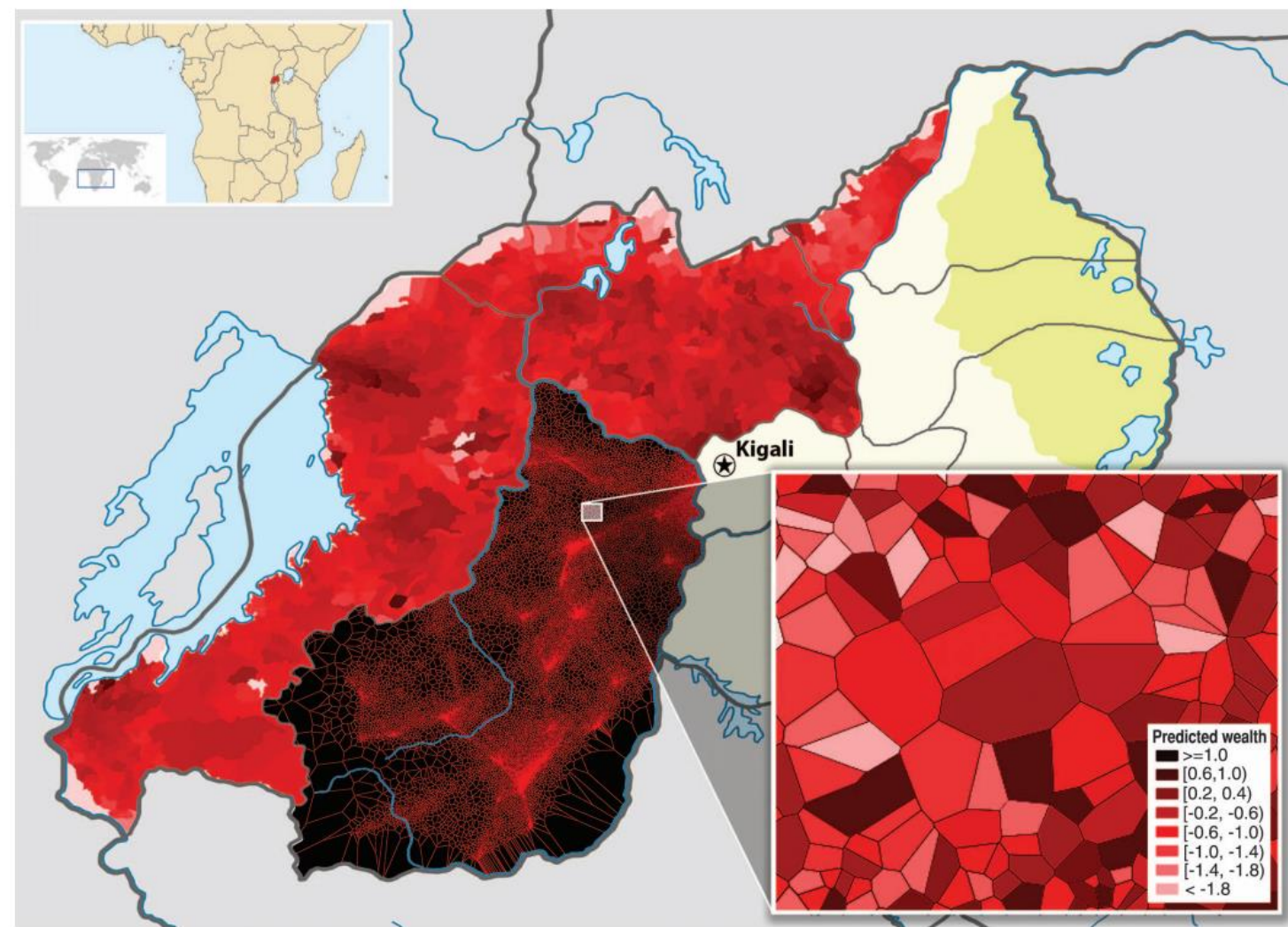
Teoría estadística

- Áreas pequeñas
- Muestreo
- Post estratificación
- Imputación
- Corrección por no respuesta

Son costosas

- No se pueden hacer todo el tiempo
- No pueden ser representativas a cualquier nivel





Muestra de ~850 personas:

- Características de la vivienda
- Índice de bienestar material

$$y = f(\text{DatosTelefono})$$

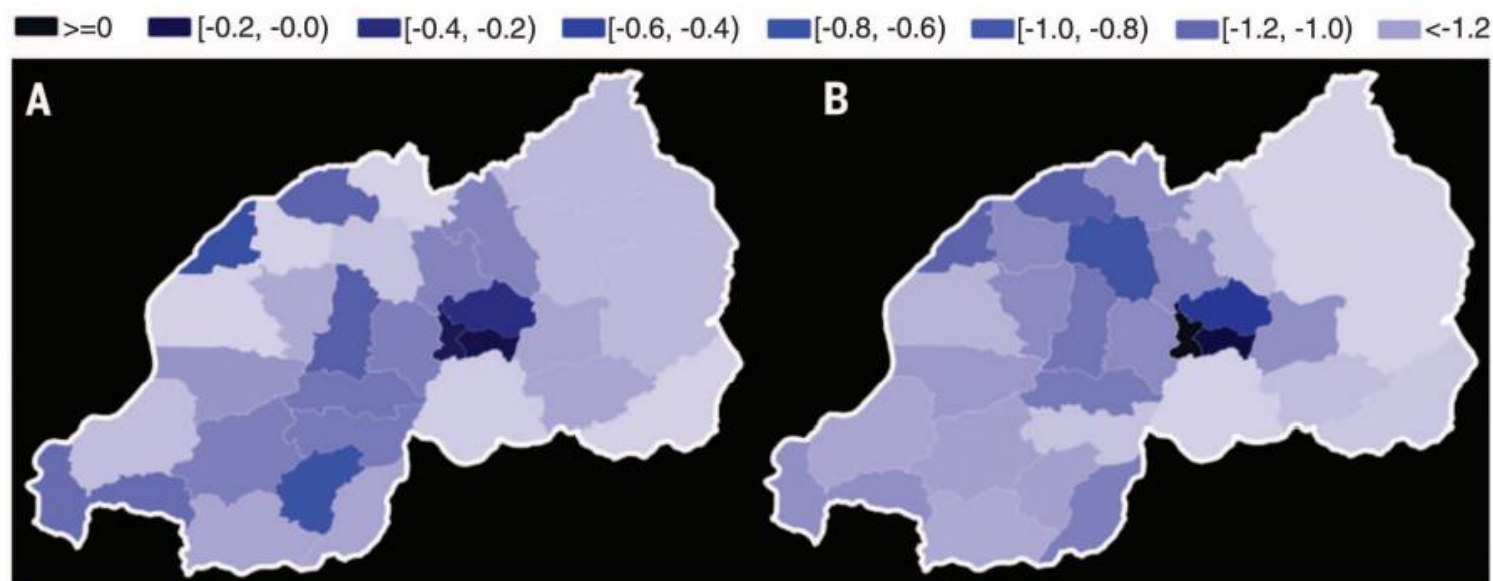


Fig. 3. Comparison of wealth predictions to government survey data. (A) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (B) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households. (C) Comparison of actual and predicted district wealth, for each of the 30 districts, with dots sized by population. (D) Comparison of actual and predicted rates of electrification, for each of the 30 districts. (E) Comparison of actual and predicted cluster wealth, for each of the 492 DHS clusters. CDR, call detail records.

Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning



Paso 1: Predecimos la pobreza con datos satelitales

Paso 2: Usamos esas predicciones como una covariable
para SAE



El INE genera datos de alta calidad

La ciencia de datos permite potenciar la oferta estadística

Es un camino que recién comenzamos

Les voy a contar sobre algunas experiencias

Codificación automática de ocupación

b1_1 ¿Cuál es el oficio, labor u ocupación que [nombre de persona] realizó la semana pasada en su actividad principal?

☐ 1 Analista socioeconómico

☐ 88 No sabe
☐ 99 No responde

b1_2 ¿Qué tareas realizó en esta ocupación?

☐ 1 Procesar y analizar datos de encuestas

☐ 88 No sabe
☐ 99 No responde

Inicio > Calidad Estadística > Clasificaciones > API codificación

Clasificaciones

> Directrices
Metodológicas

> Código de Buenas
Prácticas

Tutorial API de codificación automática

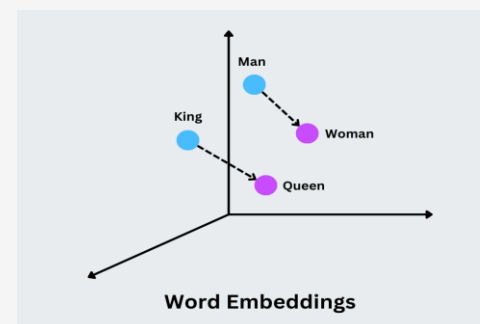
Con el objeto de hacer más eficiente el uso de recursos y mejorar la calidad de los datos publicados por el INE, durante los últimos años la institución ha avanzado en estrategias automatizadas de codificación, principalmente basadas en técnicas de aprendizaje de máquinas (machine learning). Este trabajo se encuentra a la base de la API de codificación automática que el presente tutorial busca acercar a las personas usuarias.

Esta API para la codificación automática pone a disposición de los usuarios y usuarias modelos para clasificar rama de actividad económica (CAENES) y ocupación (CIUO-08.CL) de las personas, al nivel de desagregación de 1 y 2 dígitos, de acuerdo a como sea parametrizada. Los datos de entrenamiento provienen principalmente de la coyuntura de la Encuesta Nacional de Empleo, de modo que los modelos deberían ser utilizados sobre glosas cuya recolección tenga características similares a las implementadas en el trabajo de campo de dicha encuesta.

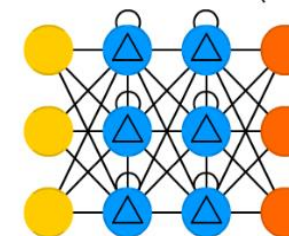
El etiquetado de los datos y el entrenamiento de los modelos fueron realizados en el marco del Proyecto Estratégico Servicios Compartidos para la Producción Estadística, radicado en la Subdirección Técnica del INE. Para mayor información acerca del proceso de etiquetado manual y de la arquitectura de los modelos, diríjase al documento "Codificación automática de clasificadores CIUO-08.CL y CAENES a partir de técnicas de machine learning. Creación de sets de entrenamiento y optimización de algoritmos", disponible en este mismo sitio para su consulta y descarga.

En la presente viñeta se muestra, a partir de algunos ejemplos, la forma de interactuar con la API de codificación automática mediante R y Python. Esta guía está orientada a usuarios y usuarias con un manejo intermedio de R y/o Python y con conocimientos básicos de machine learning. Para una aproximación más formal a los métodos de la API, diríjase al siguiente sitio https://rapps.ine.cl:9292/_docs_/

Documento metodológico servicio de codificación automática:
Codificación automática de clasificadores CIUO-08.CL y CAENES a partir de técnicas de machine learning



Gated Recurrent Unit (GRU)



Codificación automática de ocupación

```
library(httr)

glosa <- "manipulador de alimentos prepara colaciones"

request <- httr::POST("https://rapps.ine.cl:9292/predict",
                      encode = "json",
                      body = list(text = glosa,
                                  classification = "ciuo",
                                  digits = 2))
```

API de clasificación CIUO y CAENES del INE 2.1.3 OAS 3.1

[/openapi.json](#)

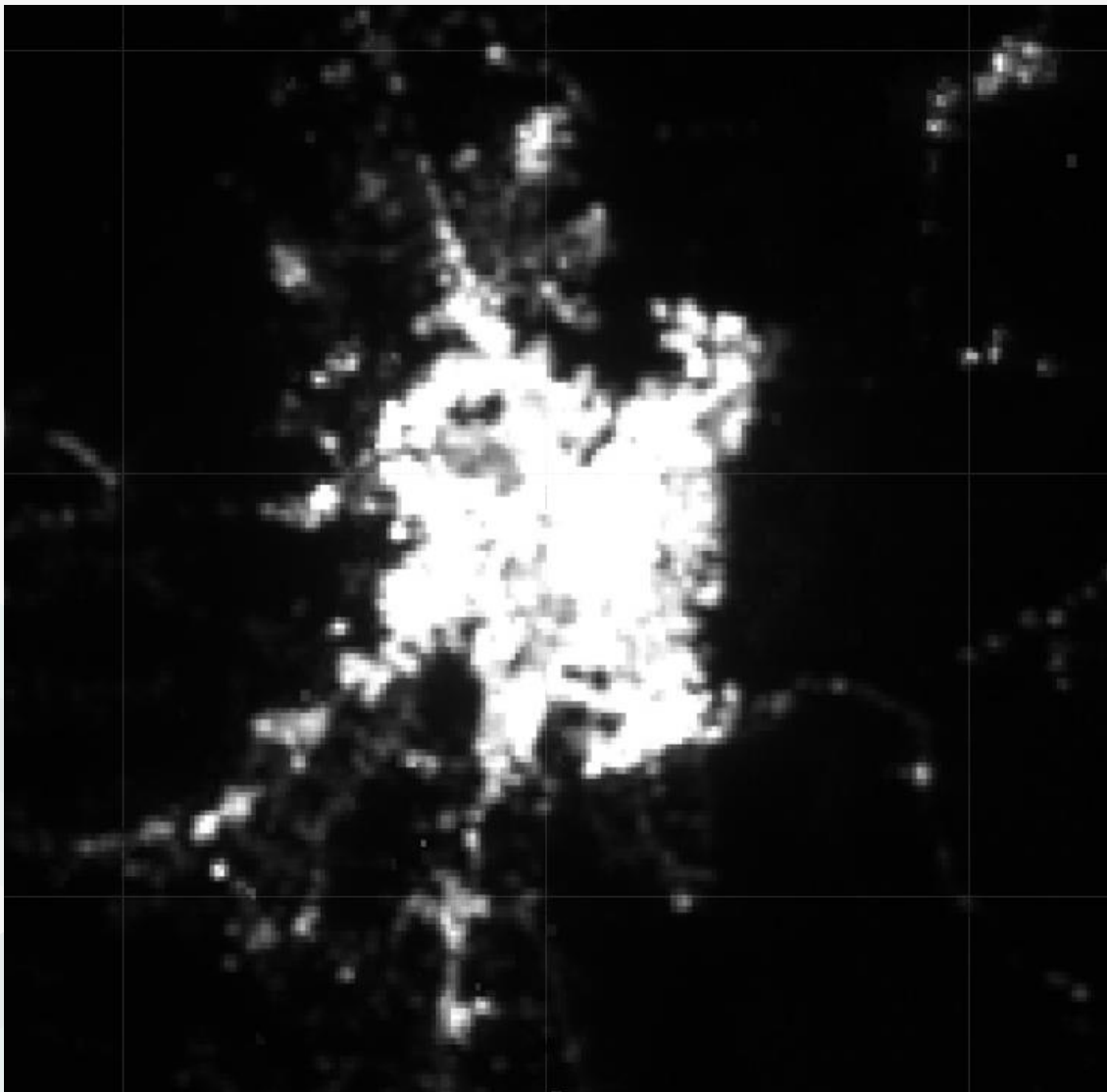
Clasificador CIUO y CAENES

obtener predicciones

POST

/predict Predice para CIUO o CAENES a 1 ó 2 dígitos.

Luces nocturnas para SAE



Google Earth Engine



Portal Transparencia

¿Qué es Transparencia? | Buscar Organismos | Solicitar información | Seguimiento de Solicitudes | Reclamos | Datos e informes | Iniciar Sesión

Acceda a la información pública de organismos del Estado

¿Qué Organismo busca? Por ejemplo Municipio de ...

Se estima que anonimizar uno de estos documentos manualmente tarda alrededor de 2 min:

	N documentos	Tiempo
Enero 2024	256	~8.8 hrs.
Junio 2024	1519	~50.6 hrs.



INE INSTITUTO NACIONAL DE ESTADÍSTICAS

INFORME DE CUMPLIMIENTO
CONVENIO A HONORARIOS A SUMA ALZADA
CON PERSONAS NATURALES

N° DEL INFORME:

FECHA INFORME:

IDENTIFICACIÓN DEL PRESTADOR DE SERVICIO

NOMBRE:

RUT:

PERÍODO DE DESEMPEÑO: DEL AL:

REGIÓN:

MOTIVO DE LA CONTRATACIÓN

ACTIVIDADES EFECTIVAMENTE DESARROLLADAS	OBJETIVO CONTRACTUAL AL QUE SE VINCULA
1. Entrenamiento de modelo de deep learning para innominación de información en los informes de cumplimiento del personal honorarios a suma alzada de la institución. 2. Elaborar presentación sobre la innominación de informes para visita de INEGI.	INVESTIGAR E IMPLEMENTAR METODOLOGÍAS DE MACHINE LEARNING Y DEEP LEARNING RELATIVAS PRINCIPALMENTE AL RECONOCIMIENTO VISUAL CON DEEP LEARNING, PARA GENERAR MODELOS Y ALGORITMOS PARA SU APLICACIÓN EN LA ALIANZA CHILE-MEXICO

Encuesta de Seguridad Ciudadana

* B* COORDINADOR DEL CONVENIO



Visión computacional

Input



python
ultralytics
YOLOv5

- ✓ Informes innominados
- ✗ Informes dudosos
- Informes no procesados

Output

Responder Responder a todos Favoritar 10/10/2024 14:57
ciencia-de-datos@ine.gob.cl
Informes October
Para: Javier Pineda Prieto, J. Pineda

Estimado/a,
Los informes anonimizados ya están listos. Puedes descargarlos en la [siguiente url](#).
El link estará disponible por 24 horas. Caso contrario puedes obtenerlos en el Lago de Datos, en la carpeta "anonimizacion_informez", en el archivo informes_10-10-24_17_46_58.zip.

Saludos,
Proyecto Ciencia de Datos



informes_dudosos
informes_no_procesados
informes_innominados



Anonimizador de informes de cumplimiento

Esta aplicación detecta automáticamente firmas y ruts en los informes de cumplimiento.

Correo de recepción

Subir archivo comprimido

Realizar carga en formato .ZIP



INFORME DE CUMPLIMIENTO CONVENIO A HONORARIOS A SUMA ALZADA CON PERSONAS NATURALES

N° DEL INFORME

1

FECHA INFORME

28-mar-24

IDENTIFICACIÓN DEL PRESTADOR DE SERVICIO

NOMBRE: Javiera Eduarda

RUT : [REDACTED]

PERÍODO DE DESEMPEÑO: DEL

01-mar-24

AL:

31-02-2024

REGIÓN:

Metropolitana

MOTIVO DE LA CONTRATACIÓN

ACTIVIDADES EFECTIVAMENTE DESARROLLADAS	OBJETIVO CONTRACTUAL AL QUE SE VINCULA
1. Entrenamiento de modelo de deep learning para innominación de información en los informes de cumplimiento del personal honorarios a suma alzada de la institución. 2. Elaborar presentación sobre la innominación de informes para visita de INEGI.	INVESTIGAR E IMPLEMENTAR METODOLOGÍAS DE MACHINE LEARNING Y DEEP LEARNING RELATIVAS PRINCIPALMENTE AL RECONOCIMIENTO VISUAL CON DEEP LEARNING, PARA GENERAR MODELOS Y ALGORITMOS PARA SU APLICACIÓN EN LA ALIANZA CHILE-MEXICO

Encuesta de Seguridad Ciudadana



V-B COORDINADOR DEL CONVENIO



FIRMA PRESTADOR DE SERVICIO



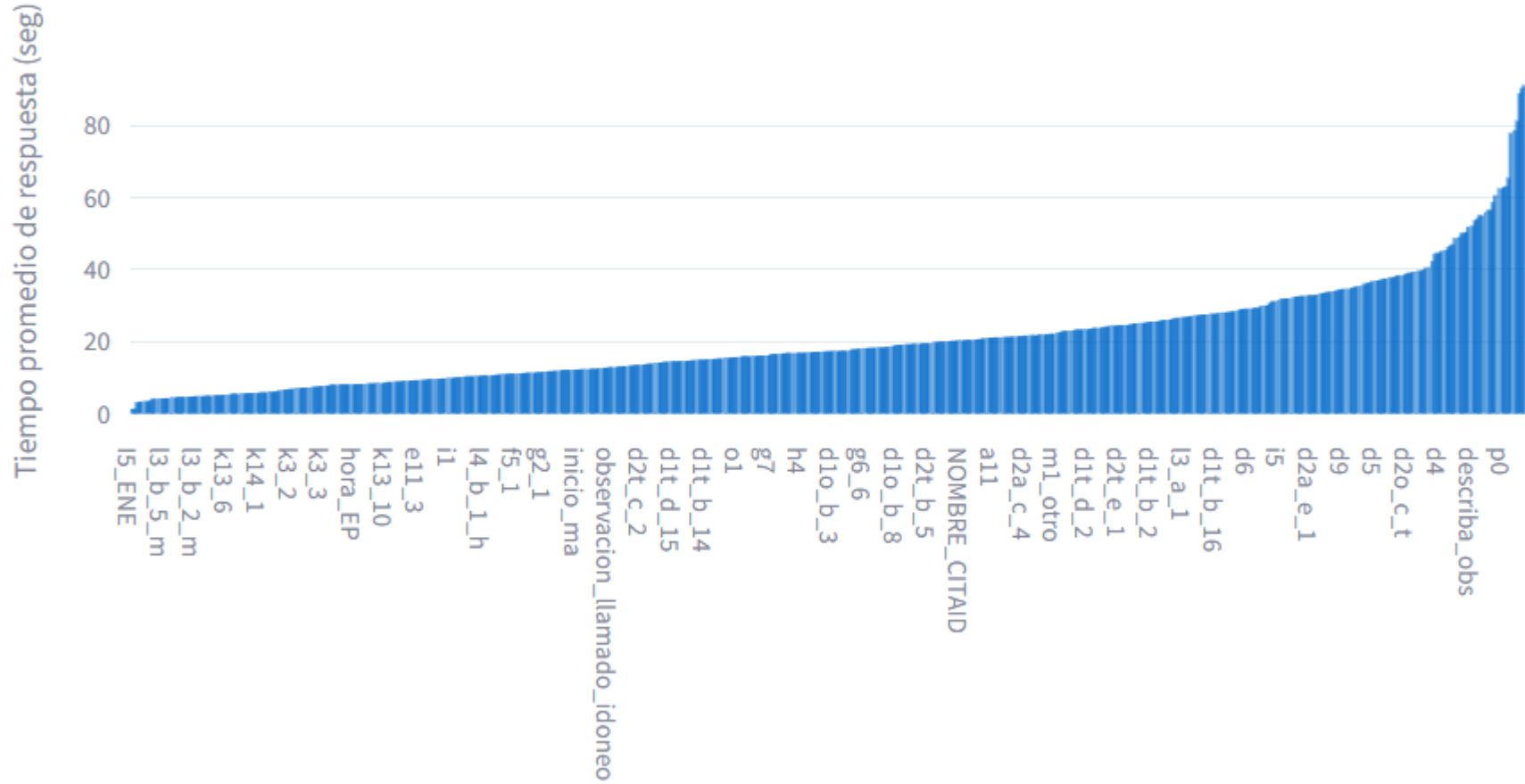
ABC interview_id	ABC event	timestamp_utc	ABC parameters
0f8360bc00df4afa81eaf902d5095e3f	InterviewCreated	2023-11-15 02:33:36.000	nan
0f8360bc00df4afa81eaf902d5095e3f	SupervisorAssigned	2023-11-15 02:33:36.000	nan
0f8360bc00df4afa81eaf902d5095e3f	InterviewModeChanged	2023-11-15 02:33:36.000	CAP
0f8360bc00df4afa81eaf902d5095e3f	InterviewerAssigned	2023-11-15 02:33:36.000	152247125
0f8360bc00df4afa81eaf902d5095e3f	KeyAssigned	2023-11-15 02:33:36.000	28-67-14-95
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	FOLIO 52059
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	REGION 13
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	PROVINCIA 131
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	COMUNA 13102
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	RPC 13102
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	UPM 1310210029
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	VIVIENDA 0
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	DIRECCION 12 DE FEBRERO 6011
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	I_kish S
0f8360bc00df4afa81eaf902d5095e3f	AnswerSet	2023-11-15 02:33:36.000	selec_aleat 0
0f8360bc00df4afa81eaf902d5095e3f	VariableDisabled	2023-11-15 02:33:36.000	Auxa_M1
0f8360bc00df4afa81eaf902d5095e3f	VariableDisabled	2023-11-15 02:33:36.000	Auxa1_M1
0f8360bc00df4afa81eaf902d5095e3f	VariableDisabled	2023-11-15 02:33:36.000	periodo_M1
0f8360bc00df4afa81eaf902d5095e3f	VariableDisabled	2023-11-15 02:33:36.000	mes_levantamiento_M1
0f8360bc00df4afa81eaf902d5095e3f	VariableDisabled	2023-11-15 02:33:36.000	extra_fono2

Desafío: Procesar entre 60 y 200 millones de registros

MINIO



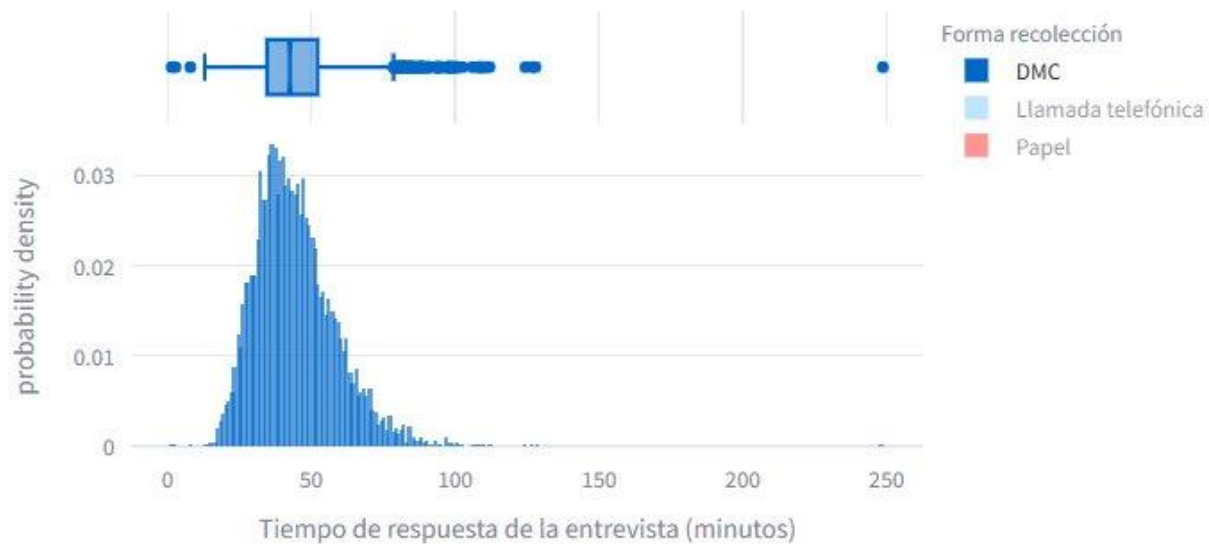
Tiempo promedio de respuesta por pregunta



Cantidad de respuestas por entrevista según recolector

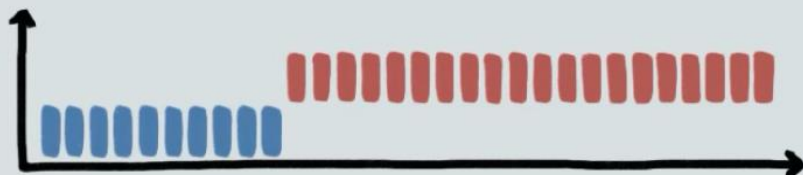


Tiempo de respuesta de entrevistas según recolector



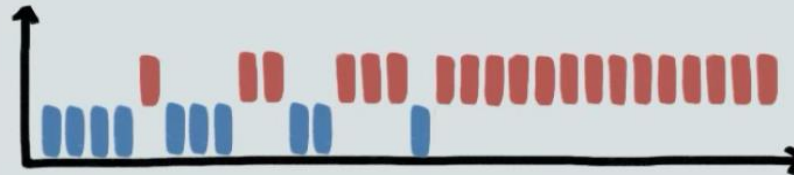
Para terminar: desafío de mediano plazo

sudden drift - a new concept occurs within a short time



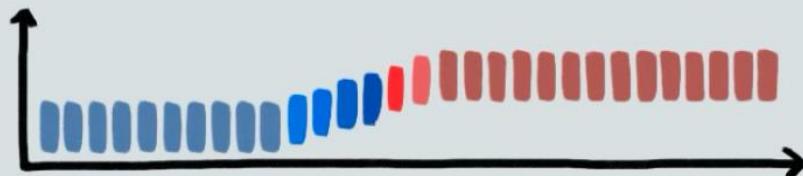
e.g. a language model can't interpret a new slang term that becomes viral overnight.

gradual drift - a new concept gradually replaces an old concept



e.g. user preferences shift slowly over time from preferring physical books to e-books.

incremental drift - an old concept incrementally changes to a new concept



e.g. climate patterns change slowly, such as the increase in averages temperatures.

recurring drift - an old concept reoccurs after some time



e.g. electricity demand experiences recurring drift due to seasonal changes.

Para terminar: desafío de mediano plazo

Quiero saber cuál era el PIB per cápita de Chile en 2020

◆ El PIB per cápita de Chile en 2020 fue de \$13,163 dólares estadounidenses.

Datos listos para la IA

Quiero el producto interno de Chile para el año 2020

◆ En 2020, el Producto Interno Bruto (PIB) de Chile se contrajo un 5,8%. Esta fue la mayor caída desde 1982, y estuvo fuertemente influenciada por el impacto de la pandemia de COVID-19.

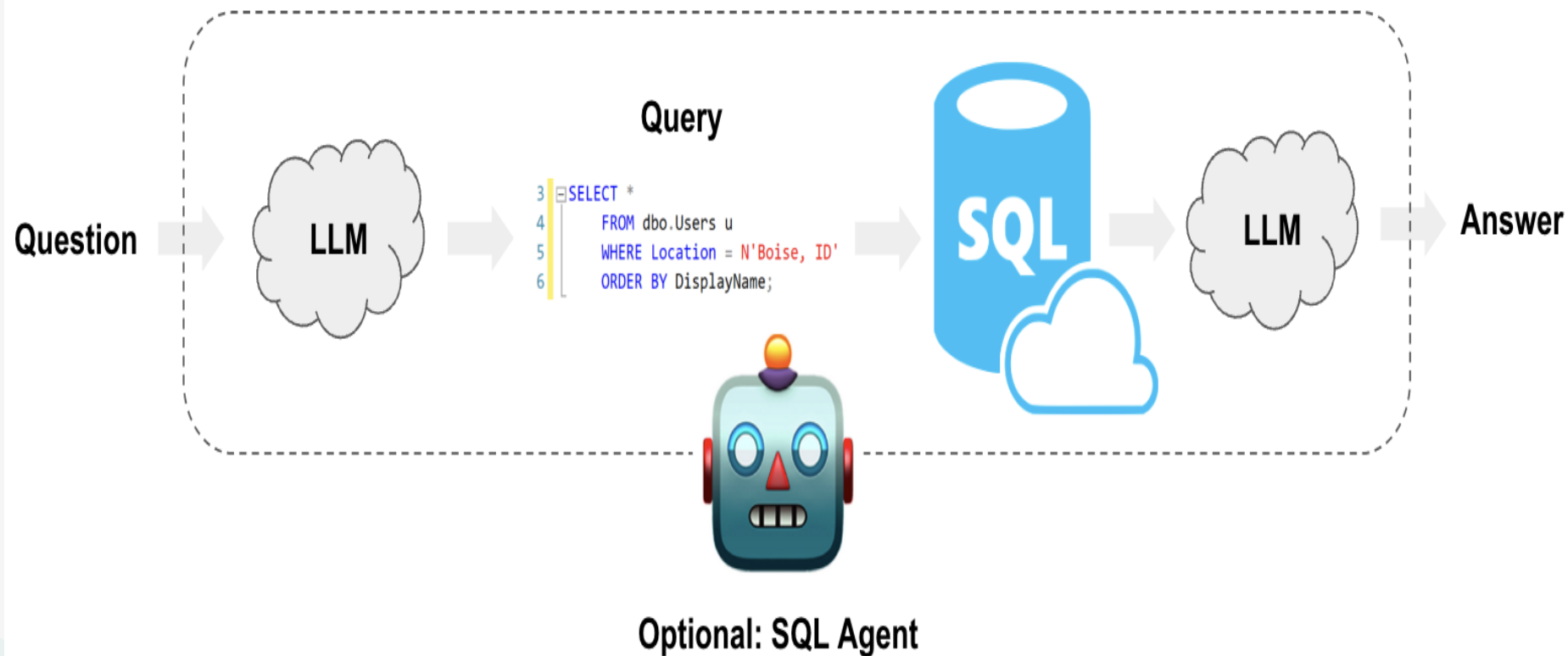
dame el valor en dólares

◆ El valor del Producto Interno Bruto (PIB) de Chile en 2020 fue de aproximadamente 277.674 millones de dólares estadounidenses.

per cápita

◆ En 2020, el Producto Interno Bruto (PIB) per cápita de Chile fue de aproximadamente 13.088 dólares estadounidenses.

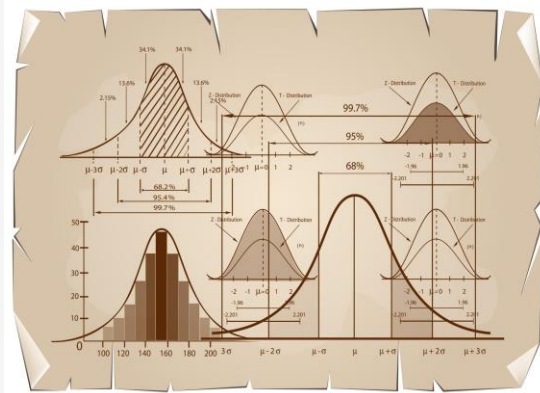
Para terminar: desafío de mediano plazo



El foco no es la automatización de tareas manuales, sino la mejora en la oferta estadística



“La investigación social en la era digital no es solo diferente en grado; es diferente en naturaleza”.



ENCUESTA

Nombre: _____ Fecha de nacimiento: _____

Sexo: ☐ Masculino ☐ Femenino ☐ Otro ☐ Prefiero no responder

Dirección: _____

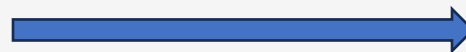
INSTRUCCIONES

En una escala de 1 a 5, ¿qué tan satisfecho estás con nuestros productos y servicios?

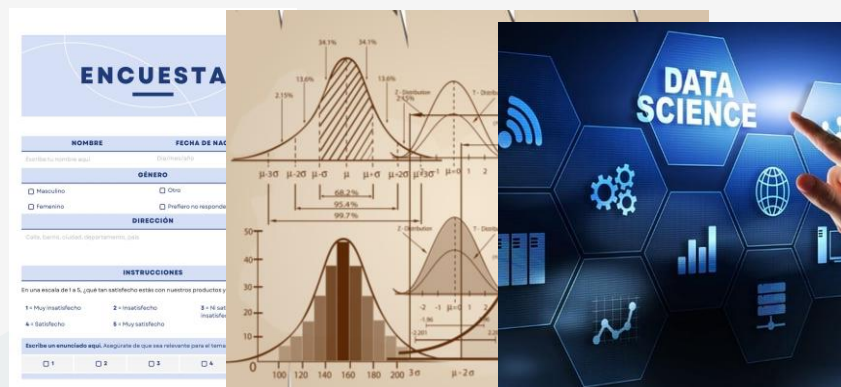
1 = Muy insatisfecho 2 = Insatisfecho 3 = Ni satisfecho ni insatisfecho
4 = Satisfecho 5 = Muy satisfecho

Escribe un comentario aquí. Responde de que sea relevante para el tema de la encuesta.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5



Las encuestas de ayer es la ciencia de datos de hoy



Síntesis entre nuevos y viejos métodos para mejorar y ampliar la producción de estadística oficiales



Gracias

ine.gob.cl

