



Tecnológico  
de Monterrey

# 01. Presentación

**Ciencia de datos para la toma de decisiones II**

Jorge  
Juvenal  
Campos Ferreira

 [juvenal.campos@tec.mx](mailto:juvenal.campos@tec.mx)

# Programa de la clase

- Presentación
- Revisión del temario
- Diagnóstico programación
- Verificación de instalación de R y Python
- Revisión mecanismos evaluación
- Recomendaciones uso IA
- Uso de LLMs

# Sobre mí



**M.C. JORGE JUVENAL CAMPOS FERREIRA.**

- \* **Analista de datos, México, ¿Cómo vamos? Y Fundación Novagob México**
- \* **Columnista en Atiempo.TV Coahuila**

## Educación Formal:



### **Licenciatura:**

Ingeniería en Irrigación por la Universidad Autónoma Chapingo. (2009-2014).



### **Maestría:**

Maestría en Economía por El Colegio de México (2016-2018).

## Contacto:



**GitHub:** JuveCampos



**LinkedIn:** Jorge Juvenal Campos Ferreira



**Twitter:** @JuvenalCamposF



**IG:** juvenalcampos.dataviz

# Experiencia profesional

- **Analista de datos.**
  - CIDE - Laboratorio Nacional de Políticas Públicas.
  - México ¿Cómo vamos?
  - Fundación Novagob México
- **Profesor**
  - Periodismo de datos - Maestría en Periodismo del CIDE
  - Tableros en R/Shiny - Datacrunchers
- **Periodista**
  - Columna semanal en Atiempo.TV
  - Escritor y colaborador en Nexos y Animal Político

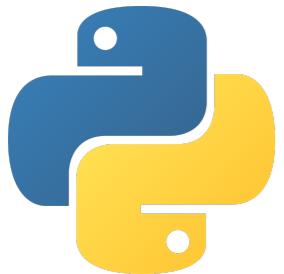
# Instalación R y Python

- Vamos a verificar que tengan R y Python instalado.



**Instalar R**

<https://cran.r-project.org/>



**Instalar Python**

<https://www.python.org/downloads/>



**Instalar RStudio**

[https://posit-co.translate.goog/download/rstudio-desktop/?  
\\_x\\_tr\\_sl=en&\\_x\\_tr\\_tl=es&\\_x\\_tr\\_hl=es&\\_x\\_tr\\_pto=tc](https://posit-co.translate.goog/download/rstudio-desktop/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc)

# Instalación R y Python

- Vamos a verificar que tengan R y Python instalado.



## Instalar R

<https://cran.r-project.org/>



## Instalar Python

<https://www.python.org/downloads/>



## Instalar RStudio

[https://posit-co.translate.goog/download/rstudio-desktop/?  
\\_x\\_tr\\_sl=en&\\_x\\_tr\\_tl=es&\\_x\\_tr\\_hl=es&\\_x\\_tr\\_pto=tc](https://posit-co.translate.goog/download/rstudio-desktop/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc)

**El curso se va a dar principalmente en R y RStudio, por lo que sí es importante que verifiquen que el programa esté instalado y funcionando.**

## Sobre ustedes

- ¿Cómo se llaman?
- ¿De donde son?
- ¿Por que están en esta carrera?
- ¿Que tanto saben de Ciencia de Datos?
- ¿Qué herramientas saben usar?
- ¿Qué esperan de este curso?



# Propósito del curso

## **El curso busca que el estudiante:**

Diseñe, implemente y comunique soluciones de ciencia de datos de forma integral, combinando el uso de programación en R, análisis estadístico, IA generativa, procesamiento de texto, análisis geoespacial y manejo de datos a gran escala, de tal forma que le ayuden a la toma de decisiones estratégicas basadas en evidencia.

## **Large Language Models**

0.1 Fundamentos teóricos de LLMs

0.2 Utilización de herramientas de LLMs para generación de textos y código

## **MÉTODOS NO SUPERVISADOS DE APRENDIZAJE DE MÁQUINA.**

1.1 INTRODUCCIÓN A MODELOS BASADOS EN DISTANCIA.

1.2 NEAREST NEIGHBOURS.

1.3 K-MEANS.

1.4 K-MEDOIDS.

1.5 HIERARCHICAL CLUSTERING.

1.6 Evaluar la calidad y utilidad de clusters en contextos reales

1.7 Traducir los resultados en segmentos, grupos o patrones accionables

## **ANÁLISIS GEO-ESPACIAL.**

2.1 INTRODUCCIÓN A DATOS GEOESPACIALES.

2.1.1 Introducción al trabajo con datos espaciales (coordenadas, polígonos, shapefiles, etc.).

2.2 MÉTODOS DE MEDICIÓN DE DISTRIBUCIONES GEOGRÁFICAS.

2.3 IDENTIFICACIÓN DE PATRONES Y CLUSTERS EN ESPACIOS GEOGRÁFICOS.

## **PROCESAMIENTO DE LENGUAJE NATURAL.**

3.1 INTRODUCCIÓN AL PROCESAMIENTO NATURAL DE TEXTO (NLP)

3.2 ESTADÍSTICAS BÁSICAS DE TEXTO

3.3 REGULAR EXPRESSIONS

3.4 PROCESAMIENTO DE HTML CON BEAUTIFUL SOUP (Web Scraping)

3.5 REDES Y TEXTO NO ESTRUCTURADO (grafos, enlaces).

3.6 DESCARGA DE TEXTO NO ESTRUCTURADO DE INTERNET

3.7 Descarga de datos desde APIs

## **ARQUITECTURA DE DATOS.**

4.0 Diseñar bases de datos

4.1 BASES DE DATOS: RELACIONALES (SQL), NO RELACIONALES (MONGO, DYNAMO), CSV Y JSON.

4.2 INSTANCIAS, CÓMPUTO DISTRIBUIDO, PROCESAMIENTO EN PARALELO, CONTENEDORES Y MICROSERVICIOS.

## **DESARROLLO DE UNA NARRATIVA DE COMUNICACIÓN.**

4.5.1 Presentación de hallazgos con datos para audiencias no técnicas

4.5.2 Integrar resultados de modelos, análisis espaciales y textuales en informes estratégicos

4.5.3 Argumentar decisiones en base en evidencia y métricas.

# Evaluación

## Criterios de evaluación:

- \* **(30%)** Reto del socio formador
- \* **(35%)** Tareas
- \* **(35%)** Examenes

Se realizará un mínimo de dos examenes, y una serie de tareas relativas a los temas que se revisarán. Tanto Manuel como yo nos encargaremos de construir el examen.

### Reto:

Proporcionar herramientas de datos útiles para el área de datos de Global Fund for Women.

**Contacto en la contraparte:** María de los Ángeles Lasa, PhD

**Ejemplo:** análisis y detección de discurso machista o misógino.

## Reto

Durante este semestre, trabajaremos en colaboración con **Global Fund for Women** en un reto enfocado en el uso de ciencia de datos para potenciar su misión de financiar y fortalecer movimientos sociales alrededor del mundo.

El desafío consistirá en **desarrollar un modelo que tome como insumo principal texto — por ejemplo, documentos, narrativas o descripciones de proyectos— y que pueda aportar valor a la organización en la identificación, clasificación o análisis de información relevante para sus objetivos estratégicos**. Este modelo podrá ser supervisado o no supervisado, según la naturaleza del problema planteado y la estrategia de abordaje del equipo.

Contaremos con la **Dra. María de los Ángeles Lasa**, Gerente del Data Gender Hub, como socia formadora, quien nos guiará para asegurar que la solución propuesta esté alineada con las necesidades reales y el impacto social que busca la organización.



## Reglas de clase

- Mantener el respeto hacia compañeros y profesores
- Participar de forma ordenada y respetuosa
- Usar dispositivos electrónicos solo cuando el profesor lo indique
- Realizar actividades correspondientes al curso en desarrollo

# Uso de la IA



✓ Se permite usar modelos de IA para el apoyo en la resolución de trabajos, siguiendo los **principios éticos** del uso de la IA en el Tec de Monterrey.

**Respeto a la dignidad humana:** No manipular ni influir indebidamente en personas.

**No maleficencia:** Evitar daños físicos, psicológicos, reputacionales o a la privacidad.

**Promoción de la autonomía:** Fomentar que las personas tomen decisiones informadas.

**Equidad:** Acceso inclusivo y beneficios compartidos.

**Seguridad:** Proteger datos y usar entornos seguros.

**Veracidad:** Contrastar y validar la información generada.

**Explicabilidad y transparencia:** Entender y declarar el uso de IA.

**Responsabilidad:** Evaluar consecuencias y actuar con reflexión.

**Bienestar social y medioambiental:** Uso para el bien común y la sostenibilidad.

# Uso de la IA



## Sugerencias:

1. Declarar siempre su uso en trabajos individuales y colaborativos
2. Emplear IA para potenciar el aprendizaje, resolver dudas y organizar ideas
3. Citar y referenciar cuando sea necesario (seguir formato APA para Chatbots)

Estudiantes



Docentes



# **Diagnostico uso herramientas Ciencia de Datos**

# Test diagnóstico

Saque lápiz y papel. Escriban en una hoja su nombre y respondan a las siguientes preguntas.



**¿Haz utilizado alguna de estas herramientas?**

**¿Alguna otra?**

- R/Rstudio
- Python (que IDE)
- QGIS / ARCGIS o alguna herramienta para hacer mapas
- Algún modelo LLM

Del uno al 10, ¿Cómo te consideras de experto en el manejo de R/Python? ¿Por qué?

Del uno al 10, **¿Cómo te consideras en el uso de IA / LLMs? ¿Por qué?**

¿Qué es lo más **complejo/complicado o elaborado** que has hecho con IA?

**¿Que temas de datos te interesan? (Educación, Finanzas, Estadísticas nacionales, manejo de texto, datos geográficos, etc).**

# Test diagnóstico

6/7

¿Qué métodos de **Machine Learning** conoces?  
¿Para que sirven o para que los usas/has usado?

# Test diagnóstico

7/7

¿Haz escuchado alguna noticia sobre la IA últimamente? ¿Cuales? ¿Qué **opinión** te generan?

# **Modelos LLM**

# Introducción a los Modelos de Lenguaje

Un sistema de IA que **procesa y genera texto** en lenguaje natural, aprendiendo patrones a partir de **grandes** cantidades de datos.

Ejemplo: Predecir la palabra que sigue en una frase.

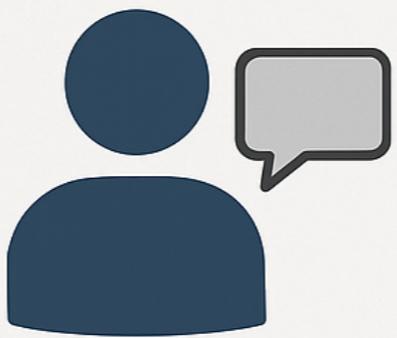


## Usos comunes de los LLMs

### Usos Comunes de los LLMs



Redacción y  
edición de textos



Explicaciones  
y tutorias



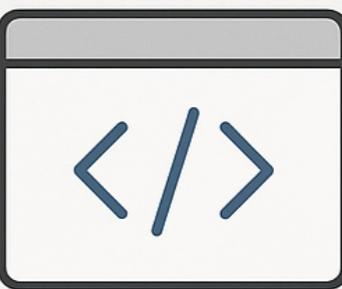
Resumen de  
documentos



Simulación de  
diálogos o roles



Resolución  
de dudas

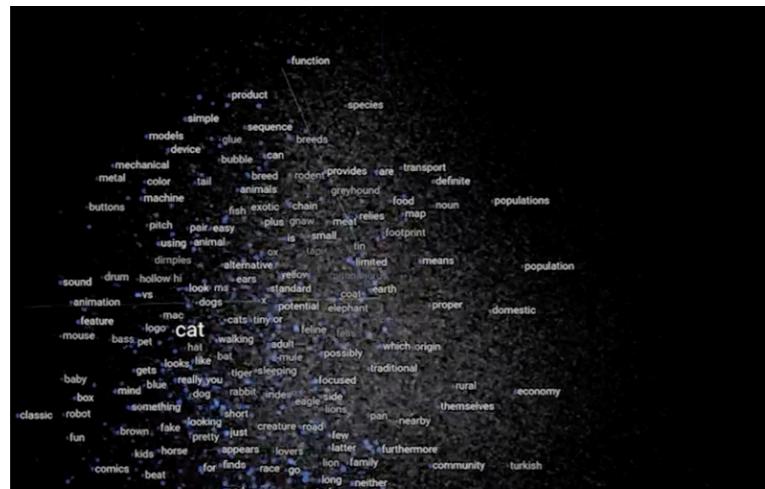


Generación  
de código

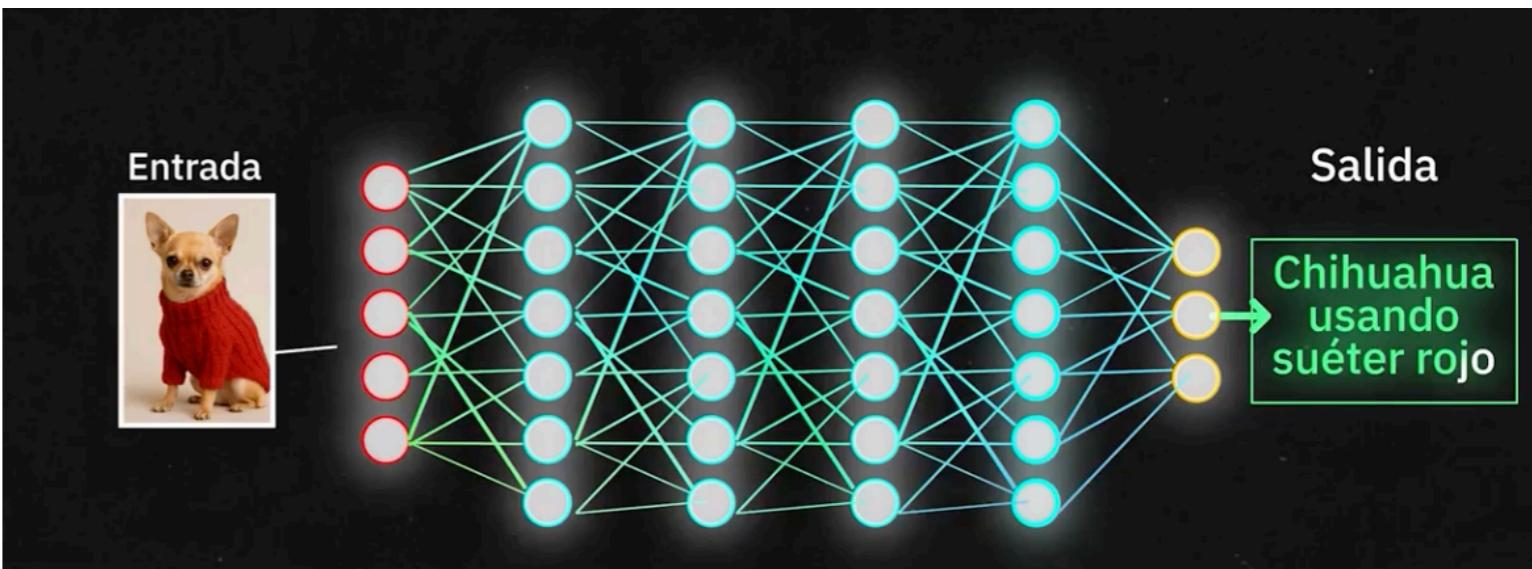
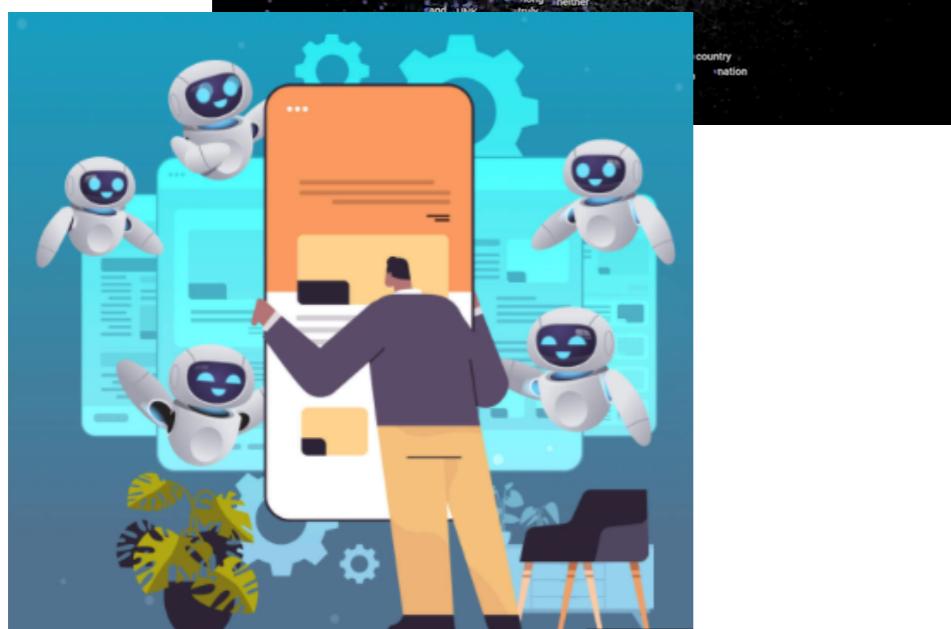
*\*Las imágenes no se generan estrictamente en un modelo LLM, sino que se generan en otro tipo de modelos.*

# ¿Cómo funciona un LLM?

1. Texto de entrada → **Tokenización** (palabras a números).
2. **Procesamiento** con redes neuronales (capas de **atención**).
3. Predicción de la **siguiente palabra** según el contexto.
4. Respuesta generada en texto natural.



$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



# ¿Cómo funciona un LLM?

1.Revisar video:

<https://www.youtube.com/watch?v=awGfhmsN7Lc>



# **Fundamentos de Ingeniería de Prompts**

# Fundamentos de ingeniería de prompts

La “*ingeniería de prompts*” trata de incidir en las instrucciones que se le brindan a un LLM, de tal manera que:

- 1) Se transmita al modelo de manera inequívoca lo que se espera de él.
- 2) Lograr que el modelo genere respuestas consistentes y predecibles ante inputs similares
- 3) Obtener los mejores resultados con la menor cantidad de tokens posibles, optimizando tanto costos como tiempos de procesamiento.

La **calidad** de la respuesta depende directamente de la claridad, el contexto y la estructura de la instrucción.



# Fundamentos de ingeniería de prompts

★ **Prompt:** Es la instrucción o pregunta que le damos al modelo para que genere una respuesta. La calidad del *prompt* influye directamente en la calidad de la respuesta.

## Ingredientes mínimos de un buen prompt

**Enfoque:** La actividad que queremos que realice el modelo.

**Contexto:** La información previa que queremos que el modelo tome en consideración al momento de procesar el prompt

**Límites:** Lo que el modelo debe o no debe incluir en su respuesta

**Rol:** El rol que debe asumir el modelo al momento de dar la respuesta.

# Ejemplos de Prompts

## Prompt sencillo

Realiza una investigación completa sobre el INFONAVIT de México. Incluye su historia, funcionamiento actual, noticias recientes y cualquier dataset disponible público. Proporciona información detallada y actualizada.

¿Cuál dará el mejor resultado?

## Prompt más elaborado

**<rol>**  
Actúa como un investigador especializado en instituciones de vivienda de México con experiencia en análisis de políticas públicas y manejo de datos gubernamentales.  
**</rol>**

**<enfoque>**  
Realizar una investigación exhaustiva y estructurada sobre el Instituto del Fondo Nacional de la Vivienda para los Trabajadores (INFONAVIT), abarcando múltiples dimensiones: histórica, institucional, operativa y de disponibilidad de datos.  
**</enfoque>**

**<contexto>**  
El INFONAVIT es una institución tripartita mexicana creada en 1972 que administra el fondo de vivienda de los trabajadores formales. Es crucial para el sector habitacional mexicano y maneja grandes volúmenes de datos sobre créditos, viviendas y trabajadores cotizantes. La investigación debe considerar su evolución, reformas recientes, impacto socioeconómico y transparencia en datos.  
**</contexto>**

**<tareas\_específicas>**  
1. Investigar la historia institucional del INFONAVIT desde 1972 hasta la actualidad  
2. Analizar noticias y desarrollos recientes (últimos 2 años)  
3. Identificar y catalogar todos los datasets públicos disponibles  
4. Examinar programas actuales y productos crediticios  
5. Evaluar el marco regulatorio y reformas recientes  
**</tareas\_específicas>**

**<limites>**  
- NO incluir información de otras instituciones de vivienda sin relación directa con INFONAVIT  
- NO especular sobre datos no públicos o confidenciales  
- NO incluir opiniones políticas partidistas  
- Sí incluir únicamente fuentes verificables y oficiales  
- Sí proporcionar enlaces directos a datasets cuando sea posible  
- Sí mantener un enfoque analítico y objetivo  
**</limites>**

**<formato\_salida>**  
Estructura la investigación en las siguientes secciones:  
1. Resumen ejecutivo  
2. Historia institucional  
3. Marco legal y regulatorio actual  
4. Programas y productos actuales  
5. Noticias y desarrollos recientes (2023-2025)  
6. Datasets disponibles (con URLs directas)  
7. Análisis de transparencia y acceso a información  
8. Conclusiones y recomendaciones para investigadores  
**</formato\_salida>**

**<fuentes\_prioritarias>**  
- Sitio oficial de INFONAVIT  
- Portal de datos abiertos del gobierno mexicano  
- INAI (Instituto Nacional de Transparencia)  
- Informes anuales oficiales  
- Boletines de prensa institucionales  
- Bases de datos del IMSS relacionadas  
**</fuentes\_prioritarias>**

## Ejercicio

Redacte un *prompt* para investigar un tema que le sea de interés.

Redáctelo tomando en cuenta **Enfoque, Contexto, Límites y Rol**.

Utilice la sintaxis XML del ejemplo anterior (*envolver cada elemento en los símbolos “>” y “<”*). ¿Considera que redactar este prompt le representa alguna ventaja?

# LLMs para generar código

# Modelos de generación de código

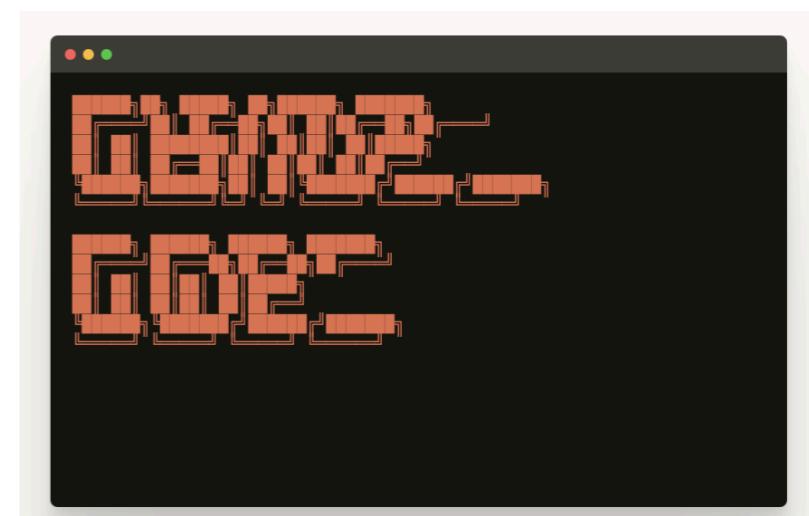
Los LLMs pueden escribir, corregir y optimizar código porque fueron entrenados no solo con lenguaje natural, sino también con **repositorios de código** y documentación técnica.

## Aplicaciones:

- Autocompletar y escribir funciones enteras.
- Detectar y corregir errores (*debugging*).
- Explicar código existente.
- Convertir descripciones en código (prompt → script).
- Traducir entre lenguajes de programación.

# Modelos de generación de código

- Hay muchas herramientas para generar código con IA. ChatGPT, Claude, Gemini o Grok funcionan bien
- También hay herramientas especializadas para mezclar IA y código, como Cursor, o CLIs (*Command Line Interface*), como Claude Code o Gemini CLI.
- Las herramientas especializadas en código facilitan el trabajo con este, y en algunos casos ejecutan y prueban el código (no hay que estar copiando y pegando, por ejemplo).



# Modelos de generación de código

- Mi herramienta favorita (ahorita a agosto del 2025) es Claude Code.
- Es un CLI que permite acceder a carpetas de trabajo, leer y escribir archivos y también generar código.
- Claude Code ha tenido un buen desempeño al momento de trabajar con código de R, y particularmente en generar aplicaciones shiny básicas.

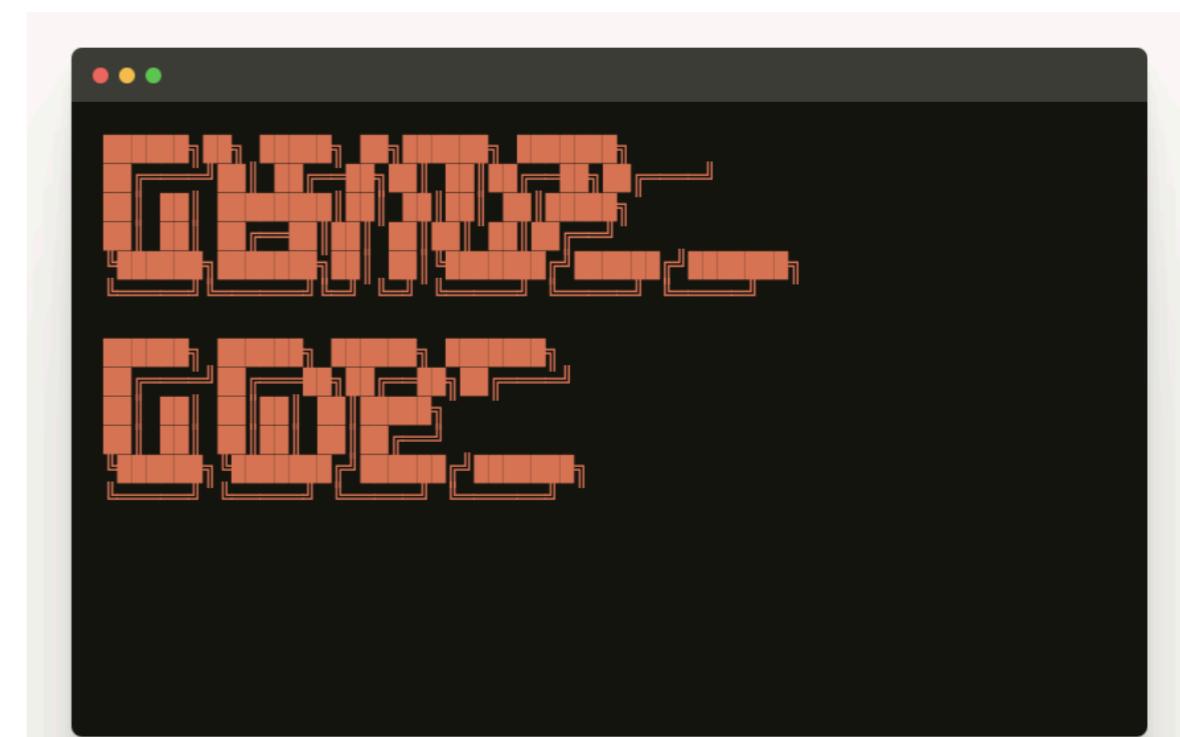


# Claude Code

- Instrucciones de instalación: [https://www-anthropic-com.translate.goog/claude-code?  
\\_x\\_tr\\_sl=en&\\_x\\_tr\\_tl=es&\\_x\\_tr\\_hl=es&\\_x\\_tr\\_pto=tc](https://www-anthropic-com.translate.goog/claude-code?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc)

**Se necesita usar terminal.**

Antes de la instalación, hay que instalar Node.js 18+



## Ejercicio guiado

Instale Claude Code o Gemini en su computadora  
Ejecutelo y corra un prompt que genere un archivo para generar una  
gráfica básica en ggplot.  
Ejecute el código y verifique que funcione.

## Tarea

Investigue que es “Zero-shot prompting”, “Few shot prompting” y “Prompt chaining” y trate de explicarlo en una guía generada para usted. Entregar guía impresa o en archivo al inicio de la clase de la próxima semana.