KIBERNUM
IT Academy

Módulo 4 – Distribuciones de Probabilidad
Clase 2

# Inferencia Estadística

## Especialización en Ciencia de Datos 2023

# Objetivos

- **Reconocer las etapas del método científico.**
- **Describir los elementos de un experimento simple.**
- **Reconocer las características principales de la ciencia de los datos.**
- **Investigación reproducible.**

# Contenido

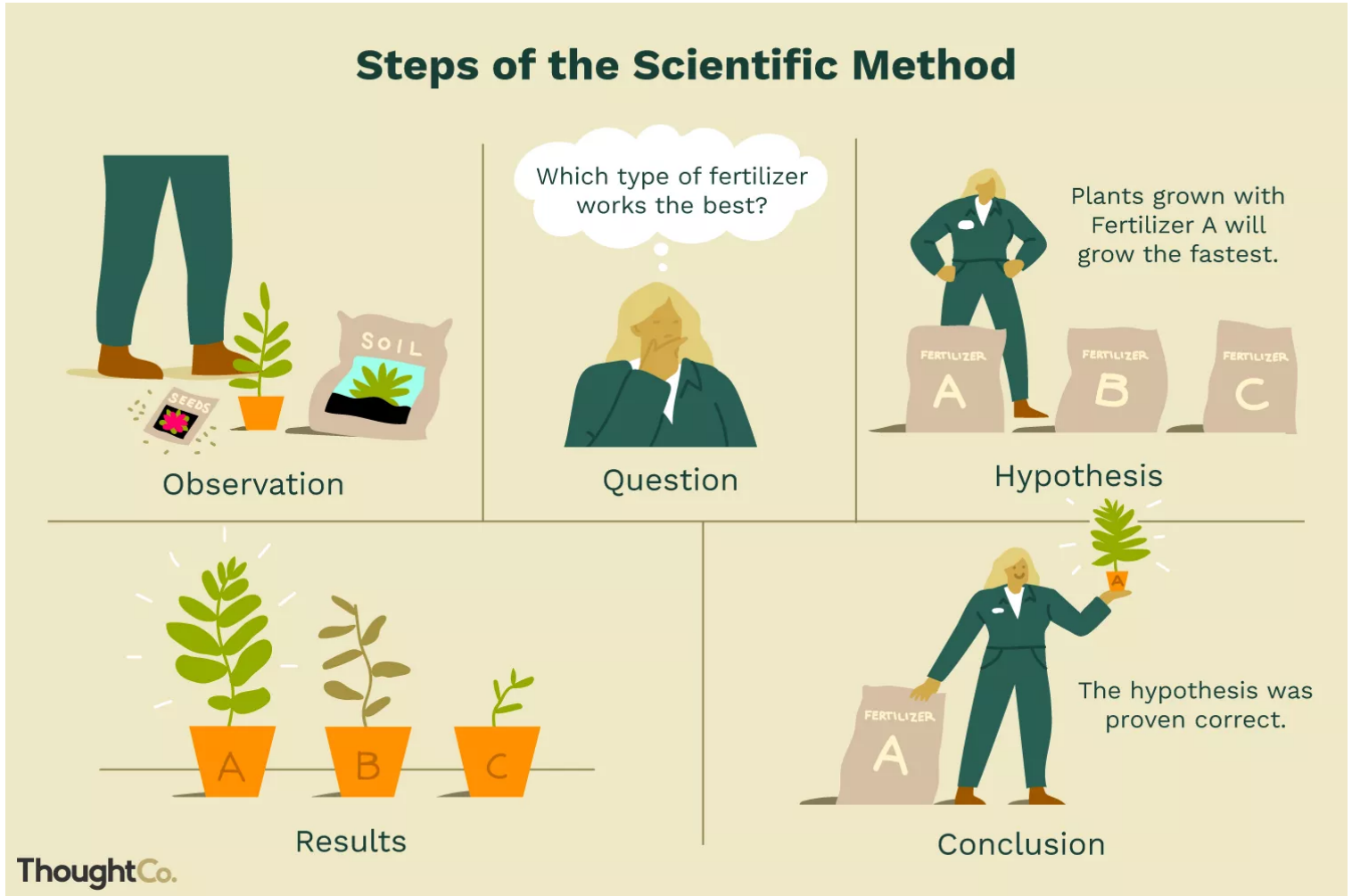
- El método científico.
- Elementos de un experimento simple.
- La ciencia de los datos.

# El Método Científico

**El método científico es una serie de pasos seguidos por investigadores científicos para responder preguntas específicas sobre el mundo natural. Implica hacer observaciones, formular una hipótesis y realizar experimentos científicos. La investigación científica comienza con una observación seguida de la formulación de una pregunta sobre lo que se ha observado.**


KIBERNUM

# El Método Científico



# El Método Científico

**Los pasos del método científico son los siguientes:**

- **Observar lo que vas a investigar**
- **Formular una pregunta investigativa**
- **Plantea una hipótesis de trabajo y recolecta datos**
- **Examina los resultados y elabora conclusiones**
- **Presenta y comparte los resultados y conclusiones**

**También, se podría considerar un paso adicional consistente en desarrollar nuevas preguntas investigativas a partir de los hallazgos.**

# Observar lo que Vas a Investigar

**Antes de que un investigador pueda comenzar, debe elegir un tema para estudiar. Una vez que se ha elegido un área de interés, los investigadores deben realizar una revisión exhaustiva de la literatura existente sobre el tema. Esta revisión proporcionará información valiosa sobre lo que ya se ha aprendido sobre el tema y qué preguntas quedan por responder.**

**La información relevante recopilada por el investigador se debe presentar en la sección de introducción de los resultados finales obtenidos del estudio. Este material de antecedentes también, ayudará al investigador en el primer paso importante para realizar un estudio: formular una hipótesis.**

# Formular una Pregunta Investigativa

**Una vez que un investigador ha observado algo y ha obtenido información básica sobre el tema, el siguiente paso es hacer una pregunta. El investigador formulará una hipótesis, que es una conjetura informada sobre la relación entre dos o más variables.**

**Por ejemplo, un investigador podría hacer una pregunta sobre la relación entre el sueño y el rendimiento académico: ¿Los estudiantes que duermen más obtienen mejores resultados en las pruebas escolares?**

**Para formular una buena hipótesis, es importante pensar en diferentes preguntas que pueda tener sobre un tema en particular.**

**También, debe considerar cómo podría investigar las causas. La falsabilidad es una parte importante de cualquier hipótesis válida. En otras palabras, si una hipótesis es falsa, debe haber una manera de que los científicos demuestren que es falsa.**

# Probar Hipótesis y Recolectar Datos

**Una vez que tenga una hipótesis sólida, el siguiente paso del método científico es poner a prueba esta corazonada mediante la recopilación de datos. Los métodos exactos utilizados para investigar una hipótesis dependen exactamente de lo que se esté estudiando. Hay dos formas básicas de investigación que se podría utilizar: investigación descriptiva o investigación experimental.**

**La *investigación descriptiva* generalmente se usa cuando sería difícil o incluso imposible manipular las variables en cuestión. Los ejemplos de investigación descriptiva incluyen estudios de casos, observación naturalista y estudios de correlación. Las encuestas telefónicas que suelen utilizar los especialistas en marketing son un ejemplo de investigación descriptiva.**

**La *investigación experimental* se utiliza para explorar las relaciones de causa y efecto entre dos o más variables. Este tipo de investigación implica manipular sistemáticamente una variable independiente y luego, medir el efecto que tiene sobre una variable dependiente definida.**

**Un experimento simple es bastante básico pero permite a los investigadores determinar las relaciones de causa y efecto entre las variables. La mayoría de los experimentos simples utilizan un grupo de control (aquellos que no reciben el tratamiento) y un grupo experimental (aquellos que sí reciben el tratamiento).**

# Examinar Resultados y Elaborar Conclusiones

**Una vez que un investigador ha diseñado el estudio y recopilado los datos, es hora de examinar esta información y sacar conclusiones sobre lo que se ha encontrado. Usando estadísticas, los investigadores pueden resumir los datos, analizar los resultados y sacar conclusiones basadas en esta evidencia.**

**Entonces, ¿cómo decide un investigador qué significan los resultados de un estudio? El análisis estadístico no solo puede respaldar (o refutar) la hipótesis del investigador; también se puede utilizar para determinar si los hallazgos son estadísticamente significativos. Cuando se dice que los resultados son estadísticamente significativos, significa que es poco probable que estos resultados se deban al azar.**

**¿Qué sucede si los resultados de un experimento no respaldan la hipótesis del investigador? ¿Significa esto que el estudio fue inútil? El hecho de que los hallazgos no respalden la hipótesis no significa que la investigación no sea útil o informativa. De hecho, dicha investigación juega un papel importante para ayudar a los científicos a desarrollar nuevas preguntas e hipótesis para explorar en el futuro.**

**Una vez extraídas las conclusiones, el siguiente paso es compartir los resultados con el resto de la comunidad científica. Esto puede ayudar a otros científicos a encontrar nuevas vías de investigación para explorar.**

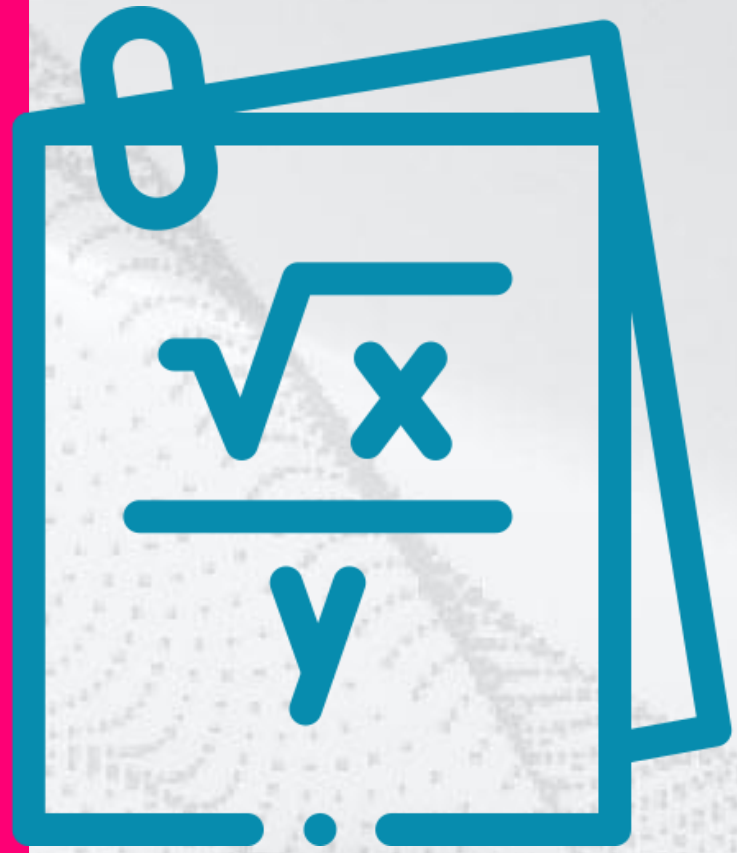
# Presentación de Resultados

**El paso final en un estudio es presentar los hallazgos. Esto a menudo se hace escribiendo una descripción del estudio y publicando el artículo en una revista académica o profesional, o bien, presentarlos en la organización a las áreas de negocio involucradas.**

**La estructura de un artículo de investigación, podrían llevar los siguientes elementos:**

- **Proporcione una breve historia y antecedentes sobre investigaciones anteriores.**
- **Presentar la hipótesis.**
- **Identificar quiénes participaron en el estudio y cómo fueron seleccionados.**
- **Proporcionar definiciones operativas para cada variable.**
- **Describir las medidas y procedimientos que se utilizaron para recopilar datos.**
- **Explicar cómo se analizó la información recopilada.**
- **Discutir lo que significan los resultados.**

**¿Por qué es importante seguir esta estructura? Al explicar claramente los pasos y procedimientos utilizados a lo largo del estudio, otros investigadores no solamente pueden conocer los resultados finales sino que también pueden replicar los resultados y referirse a la metodología utilizada. A esto, le llamamos *reproducibilidad* de la investigación.**



# El Método Científico (Ejemplo)

**Paso 1**
**Observar lo que vas a investigar**

Los investigadores eligen enfocar su estudio en adultos de 25 a 40 años con trastorno de ansiedad generalizada

**Paso 2**
**Formular una pregunta investigativa**

La pregunta que quieren responder en su estudio es: ¿Las sesiones semanales de psicoterapia reducen los síntomas en adultos de 25 a 40 años con trastorno de ansiedad generalizada?

**Paso 3**
**Plantear una hipótesis y recolectar datos**

Se trabajan con terapeutas para crear un programa consistente al que se someten todos los participantes. El grupo 1 puede asistir a terapia una vez por semana, mientras que el grupo 2 no asiste a terapia.

**Paso 4**
**Examinar resultados y elaborar resultados**

Participantes registran sus síntomas y cualquier cambio durante un período de estudio. Finalmente, las personas del grupo 1 informan mejoras significativas en sus síntomas de ansiedad, mientras que las del grupo 2 no informan cambios significativos.

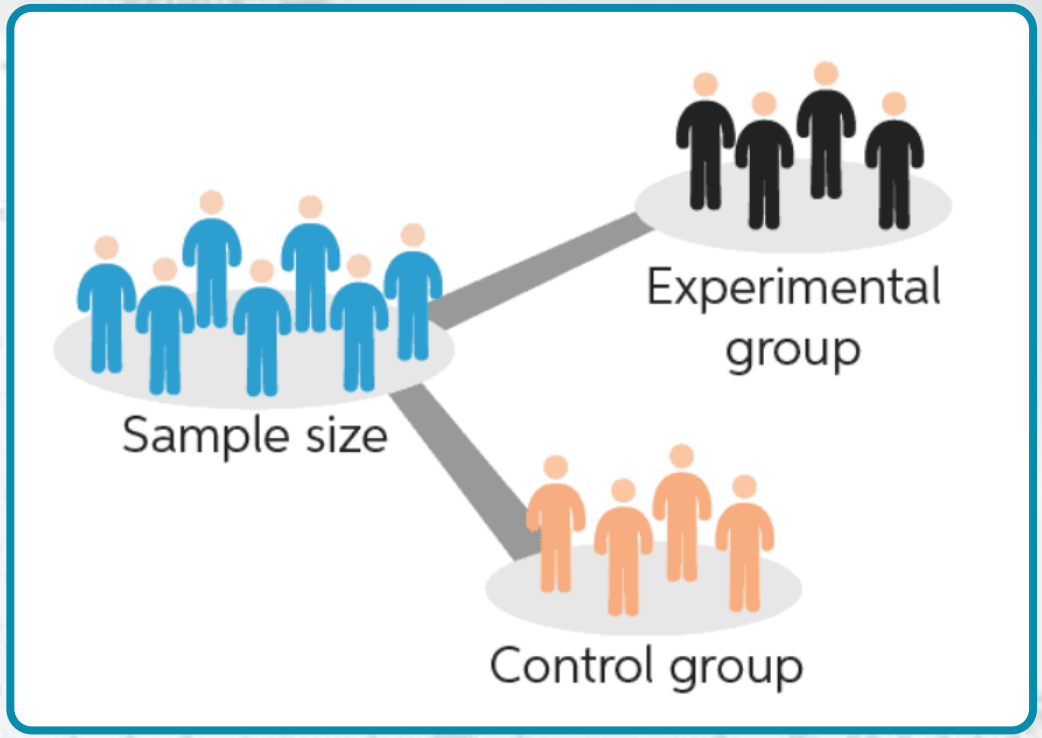
**Paso 5**
**Presentar y compartir conclusiones**

Se confecciona un informe que incluye hipótesis, información sobre los participantes, variables, procedimiento y conclusiones extraídas del estudio. En este caso, afirman que “se ha demostrado que las sesiones semanales de terapia reducen los síntomas de ansiedad en adultos de 25 a 40 años”.

# Elementos de un Experimento Simple

## EXPERIMENTO SIMPLE

**Un experimento simple es uno que los investigadores usan a menudo para determinar si los cambios en una variable pueden conducir a cambios en otra variable; en otras palabras, para establecer causa y efecto. En un experimento simple que analiza la eficacia de un nuevo medicamento, por ejemplo, los participantes del estudio pueden ser asignados aleatoriamente a uno de dos grupos: uno de ellos sería el grupo de control y no recibiría tratamiento, mientras que el otro grupo sería el grupo experimental que recibe el tratamiento en estudio.**



# Elementos de un Experimento Simple

**Un experimento simple está compuesto por varios elementos claves.**

***Hipótesis experimental***
**Esta es una declaración que predice que el tratamiento causará un efecto y, por lo tanto, siempre se expresará como una declaración de causa y efecto. Por ejemplo, los investigadores podrían formular una hipótesis de esta manera: "La administración del medicamento A resultará en una reducción de los síntomas de la enfermedad B".**

***Hipótesis nula***
**Esta es una hipótesis de que el tratamiento experimental no tendrá efecto sobre los participantes o las variables dependientes. Es importante tener en cuenta que no encontrar un efecto del tratamiento no significa que no haya efecto. El tratamiento podría afectar otra variable que los investigadores no están midiendo en el experimento actual.**

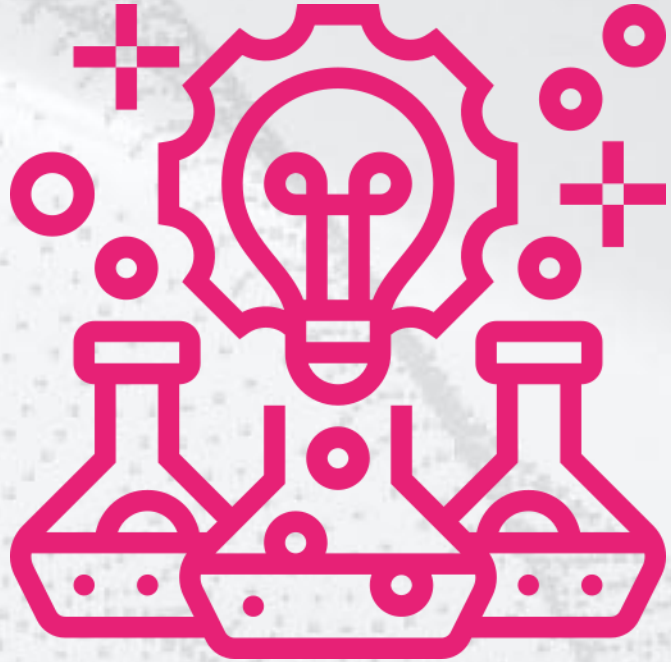
# Elementos de un Experimento Simple

***Variable independiente***
**La variable de tratamiento que es manipulada por el experimentador. Por ejemplo, la dosis del medicamento a aplicar.**

***Variable dependiente***
**Se refiere a la respuesta que miden los investigadores. Por ejemplo, el nivel de colesterol en la sangre.**

***Grupo de control***
**Estos son los individuos que se asignan aleatoriamente a un grupo pero no reciben el tratamiento. Las medidas tomadas del grupo de control se compararán con las del grupo experimental para determinar si el tratamiento tuvo algún efecto.**

***Grupo experimental***
**Este grupo de participantes del estudio está compuesto por sujetos seleccionados al azar que recibirán el tratamiento que se está probando.**



# Resultados de un Experimento Simple

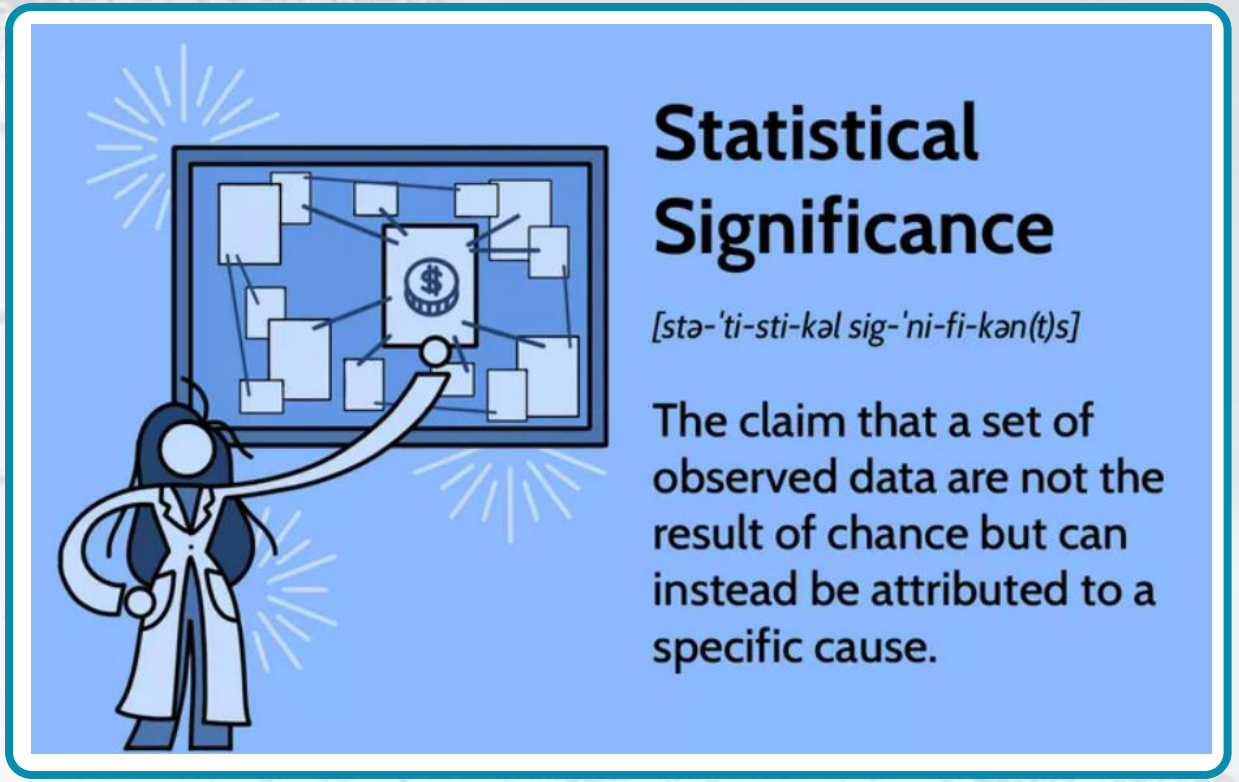
**Una vez que se han recopilado los datos del experimento simple, los investigadores comparan los resultados del grupo experimental con los del grupo de control para determinar si el tratamiento tuvo algún efecto. Debido a la posibilidad siempre presente de errores, no es posible estar 100 por ciento seguro de la relación entre dos variables. Podría, por ejemplo, haber variables desconocidas en juego que influyan en el resultado del experimento.**

**A pesar de este desafío, hay formas de determinar si lo más probable es que exista una *relación significativa.* Para hacer esto, los científicos usan *estadísticas inferenciales,* una rama de la estadística que se ocupa de sacar inferencias sobre una población con base en medidas tomadas de una muestra representativa de esa población.**

**La clave para determinar si un tratamiento tuvo efecto es *medir la significación estadística.* La significancia estadística muestra que la relación entre las variables probablemente no se deba al mero azar y que lo más probable es que exista una relación real entre las dos variables.**

# Significancia Estadística

**Una vez que se han recopilado los datos del experimento simple, los investigadores comparan los resultados del grupo experimental con los del grupo de control para determinar si el tratamiento tuvo algún efecto.**



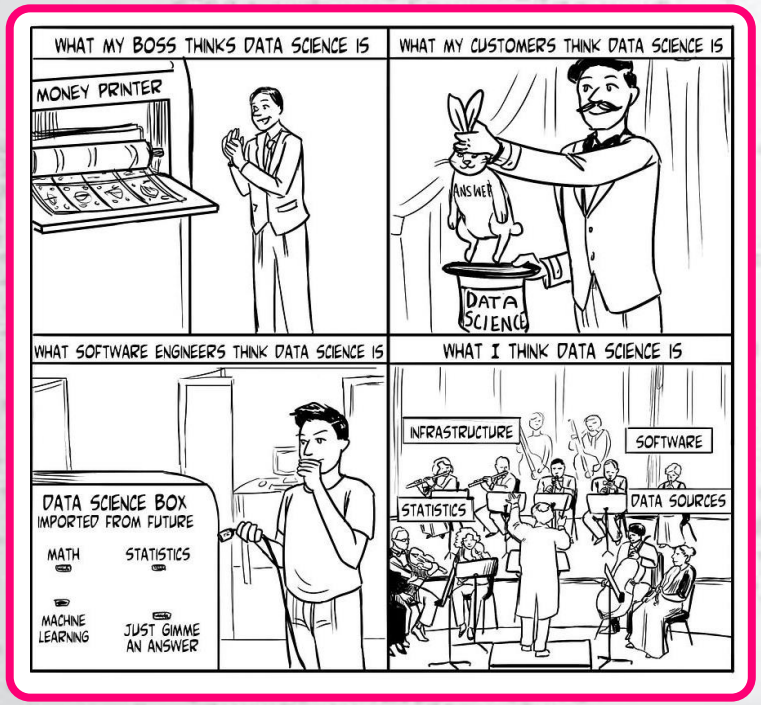
# Significancia Estadística

**La significación estadística puede considerarse fuerte o débil. Al analizar un conjunto de datos y realizar las pruebas necesarias para discernir si una o más variables tienen un efecto en un resultado, una fuerte significación estadística ayuda a respaldar el hecho de que los resultados son reales y no causados por la suerte o el azar.**

- **La importancia estadística se refiere a la afirmación de que es probable que un resultado de los datos generados por pruebas o experimentación sea atribuible a una causa específica.**
- **Un alto grado de significación estadística indica que es poco probable que una relación observada se deba al azar.**
- **El cálculo de la significación estadística está sujeto a un cierto grado de error.**
- **La significación estadística puede malinterpretarse cuando los investigadores no usan el lenguaje con cuidado al informar sus resultados.**
- **Se utilizan varios tipos de pruebas de significancia dependiendo de la investigación que se realice.**

# La Ciencia de los Datos

**¿Qué pasa si a nuestro proceso investigativo incorporamos habilidades fuertes con el manejo de datos desde distintas fuentes y técnicas computacionales?**

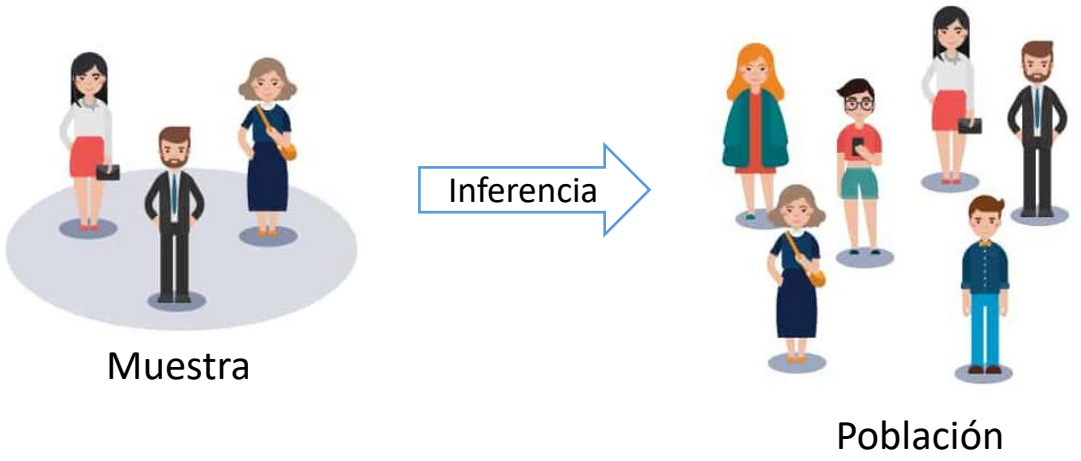




KIBERNUM

# La Ciencia de los Datos

**En la ciencia de datos, buscamos generar conclusiones sobre una población a partir de una muestra, por lo general, ruidosa.**




¿Quién ganará las elecciones?

¿Quiénes comprarán nuestros productos?

¿Cómo estará el tiempo mañana?

# La Ciencia de los Datos

**Algunos desafíos y dificultades que enfrentamos:**

- **¿Estamos tomando una muestra de datos representativa de la población sobre la cual queremos hacer inferencias?**
- **¿Hemos considerado todas las variables? ¿Hay variables que contaminen nuestras conclusiones?**
- **¿Hay sesgos sistemáticos en nuestros datos que desvíen los resultados?**
- **¿Qué aleatoriedad hay en los datos?¿Cómo lidiar con aquello?**
- **¿Estamos tratando de estimar un modelo subyacente del fenómeno estudiado?**

**Hay dos fuentes de aleatoriedad e incerteza:**

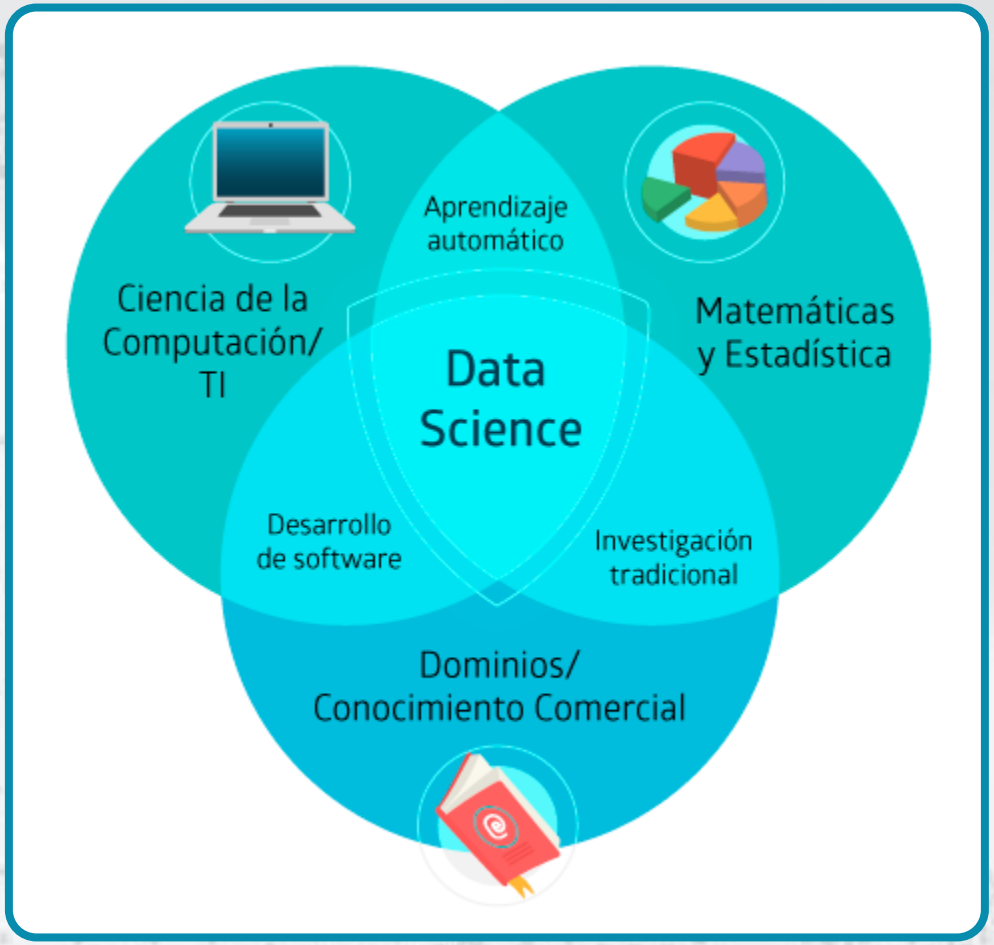
- **La aleatoriedad e incerteza propia del proceso**
- **La incerteza asociada a los métodos de recolección de datos**

**Se hace necesario procedimientos, métodos y teoremas que nos permitan extraer significado e información a partir de la data generada por procesos estocásticos.**

# La Ciencia de los Datos

**La Ciencia de Datos es considerada una ciencia debido a que utiliza un enfoque sistemático y riguroso para la obtención, análisis e interpretación de datos con el fin de descubrir patrones, tendencias y relaciones en ellos.**

**A través del método científico, la Ciencia de Datos utiliza una combinación de matemáticas, estadísticas, informática y habilidades de análisis para obtener conocimientos y comprensión de los datos, y para formular y probar hipótesis sobre los fenómenos observados.**



# La Ciencia de los Datos

**Además, la Ciencia de Datos se basa en la recopilación y análisis de datos empíricos, lo que significa que se apoya en la evidencia que se obtiene a partir de la observación y experimentación en el mundo real. De esta manera, los datos se utilizan para validar y refutar las teorías y modelos que se construyen para explicar los fenómenos observados.**

# Proceso Data Science

## Definición del problema

1

En esta etapa, se identifica el problema que se va a abordar y se define claramente el objetivo del análisis de datos. También, se identifican los datos necesarios para resolver el problema.

## Recopilación de datos

2

En esta etapa, se recopilan los datos necesarios para el análisis a partir de diversas fuentes, como bases de datos, archivos y fuentes externas. También, se limpian y transforman los datos para asegurarse de que sean precisos y completos.

## Preparación de datos

3

En esta etapa, se seleccionan y preparan los datos para su uso en el modelo de análisis de datos. Esto incluye la selección de variables relevantes, la normalización y la estandarización de datos, y la división de datos en conjuntos de entrenamiento y prueba.

## Análisis exploratorio

4

En esta etapa, se realizan diversas técnicas de visualización y análisis de datos para explorar patrones y tendencias. Esto ayuda a obtener una comprensión básica de los datos antes de proceder a técnicas de modelado más avanzadas.

## Modelamiento

5

En esta etapa, se construye y ajusta un modelo de análisis de datos que permita resolver el problema identificado. Esto puede implicar el uso de algoritmos de aprendizaje automático o de estadística para predecir o clasificar datos.

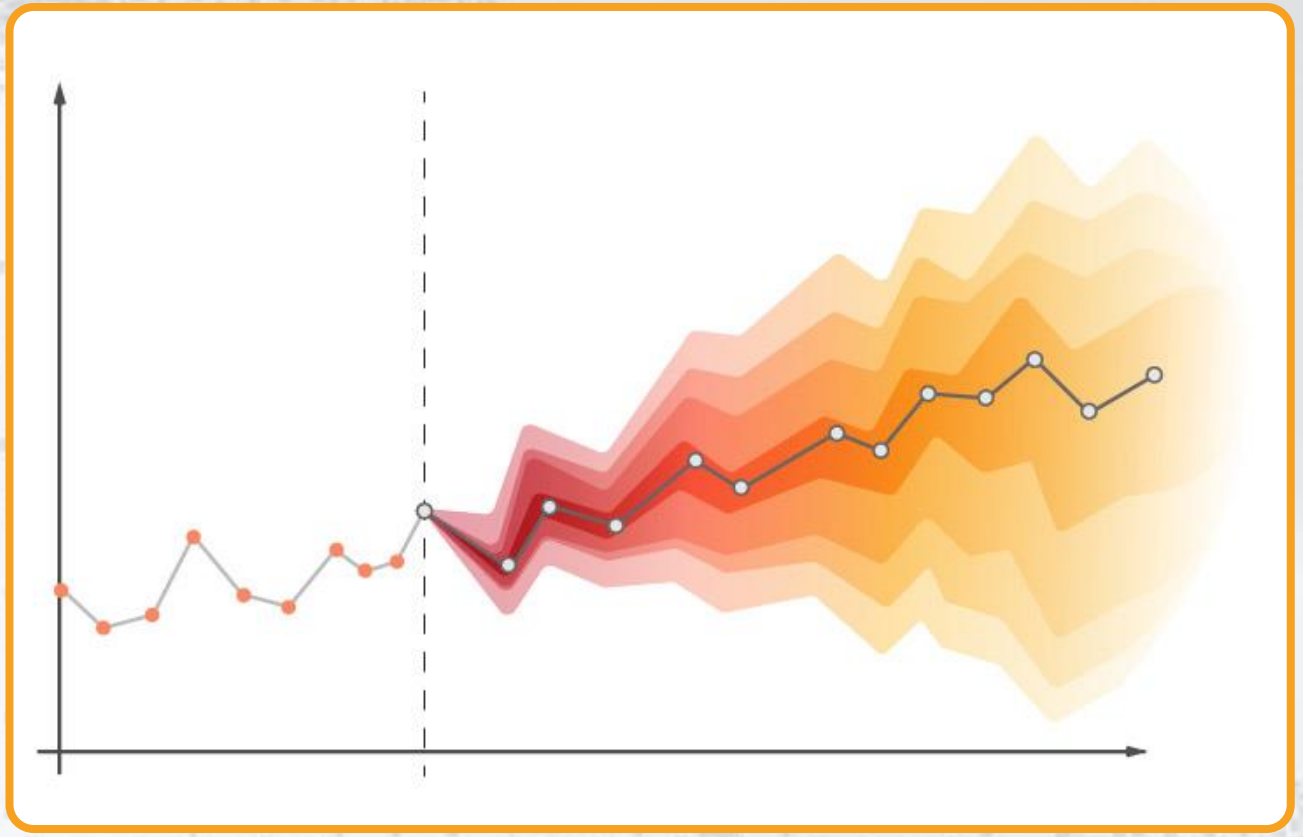
## Interpretación y comunicación

6

En esta etapa, se comunican las conclusiones y hallazgos a la organización, quienes pueden decidir realizar una explotación del modelo en el negocio.

# La Ciencia de los Datos

**En resumen, la Ciencia de Datos es una ciencia debido a que sigue un proceso riguroso y sistemático para la obtención, análisis e interpretación de datos, utiliza el método científico para formular y probar hipótesis, y se basa en datos empíricos para validar y refutar teorías y modelos.**



# Investigación Reproducible

## ¿QUÉ ES INVESTIGACIÓN

**La investigación reproducible se refiere a la práctica de publicar y compartir investigaciones científicas de tal manera que otros investigadores puedan reproducir los resultados y verificar la validez de las conclusiones.**

**La idea detrás de la investigación reproducible es que los resultados de una investigación deben ser capaces de ser verificados y validados por otros investigadores. Esto implica hacer que los datos, el software, los métodos y los análisis utilizados en una investigación sean accesibles y transparentes para que otros puedan replicar el estudio y llegar a los mismos resultados.**

KIBERNUM

# ¿Qué es Investigación Reproducible?

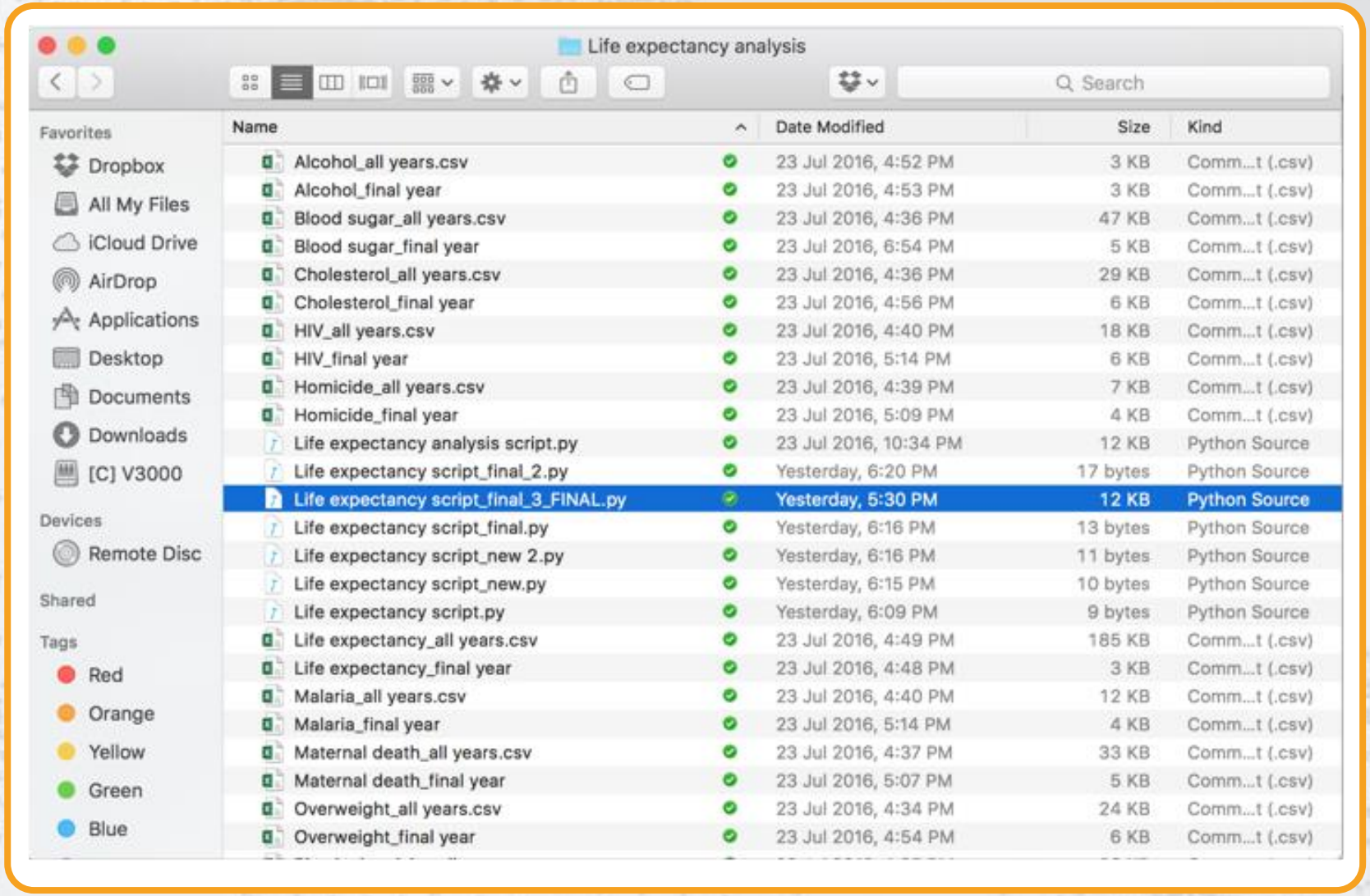
**Para lograr una investigación reproducible, se pueden utilizar diversas herramientas y prácticas, como la documentación detallada de los métodos y los procedimientos, el uso de software de control de versiones, la publicación de los datos y el código en línea, y la utilización de herramientas como los notebooks Jupyter para registrar el análisis y los resultados.**

**La investigación reproducible es una práctica importante en la ciencia y la investigación, ya que ayuda a garantizar la transparencia y la validez de los resultados de la investigación, lo que a su vez promueve la confianza en la ciencia y mejora la calidad de la investigación en general.**

KIBERNUM

# ¿Qué es Investigación Reproducible??

**Para entender qué es un análisis reproducible, partamos primero entendiendo *¿qué es un análisis no reproducible?*.**



# ¿Qué es Investigación Reproducible??

**Suponga que acaba de llegar a la organización y el anterior investigador le dejó una carpeta con esos archivos. ¿Por dónde partiría? No solamente hay múltiples versiones del script de análisis sino también, cerca de 30 datasets diferentes.**

**Pareciera que hay un código que es el que tiene la versión final. Si lo inspeccionamos, ¿logramos saber cómo se condujo la investigación? Inspeccionemos segmentos del código.**

```
# Check for missingness
totaldf.isnull().sum()
        # Lots of missing data for Smoking, Physical Activity, Malaria and HIV, so won't use
        # 14 missing values in Life Expectancy so will delete them
        # Will delete all non-complete rows once I get rid of the above 4 columns

totaldf = totaldf.drop(['Smoking', 'PhysicalActivity', 'Malaria', 'HIV'], axis=1)
totaldf = totaldf.dropna()
totaldf.isnull().sum()
totaldf.shape

# Explore the data a bit
totaldf.ix[totaldf['AlcConsumption'].idxmax()]
totaldf.ix[totaldf['ImprovedWater'].idxmin()]

Series.mean(totaldf['Suicide'])
print(totaldf.loc[totaldf['Country'].isin(['Panama', 'Guatemala', 'Australia'])])
```

# ¿Qué es Investigación Reproducible??

**Pareciera que no hay un dirección hacia donde va la investigación, o al menos el código no es capaz de reflejarlo.**

```
151  import math
152  #math.sqrt(totaldf['LifeExpectancy'])
153
154  plt.hist(np.sqrt((max(totaldf['LifeExpectancy']) + 1) - totaldf['LifeExpectancy']))
155  plt.title("Life Expectancy Histogram")
156  plt.xlabel("Value")
157  plt.ylabel("Frequency")
158  plt.show()
159
160  # Check back-transform
161  (max(totaldf['LifeExpectancy']) + 1) - (totaldf['TransformedLife']**2)[:5]
162
163  # Create the variable
164  totaldf['TransformedLife'] = np.sqrt((max(totaldf['LifeExpectancy']) + 1) - totaldf['LifeExpectancy'])
165
166  # Standardise the predictors
167  for i in list(totaldf.columns.values)[2:14]:
168      totaldf['%s' % i] = (totaldf['%s' % i] - totaldf['%s' % i].mean()) / totaldf['%s' %i].std()
169
170  # Try out the ridge/LASSO table of results from blog post
171  from matplotlib.pylab import rcParams
172  rcParams['figure.figsize'] = 12, 10
173
174  x = np.array([i*np.pi/180 for i in range(60,300,4)])
175  np.random.seed(10)  #Setting seed for reproducability
```

# Buenas Prácticas

**Para realizar análisis reproducibles, podemos partir utilizando las siguientes herramientas:**

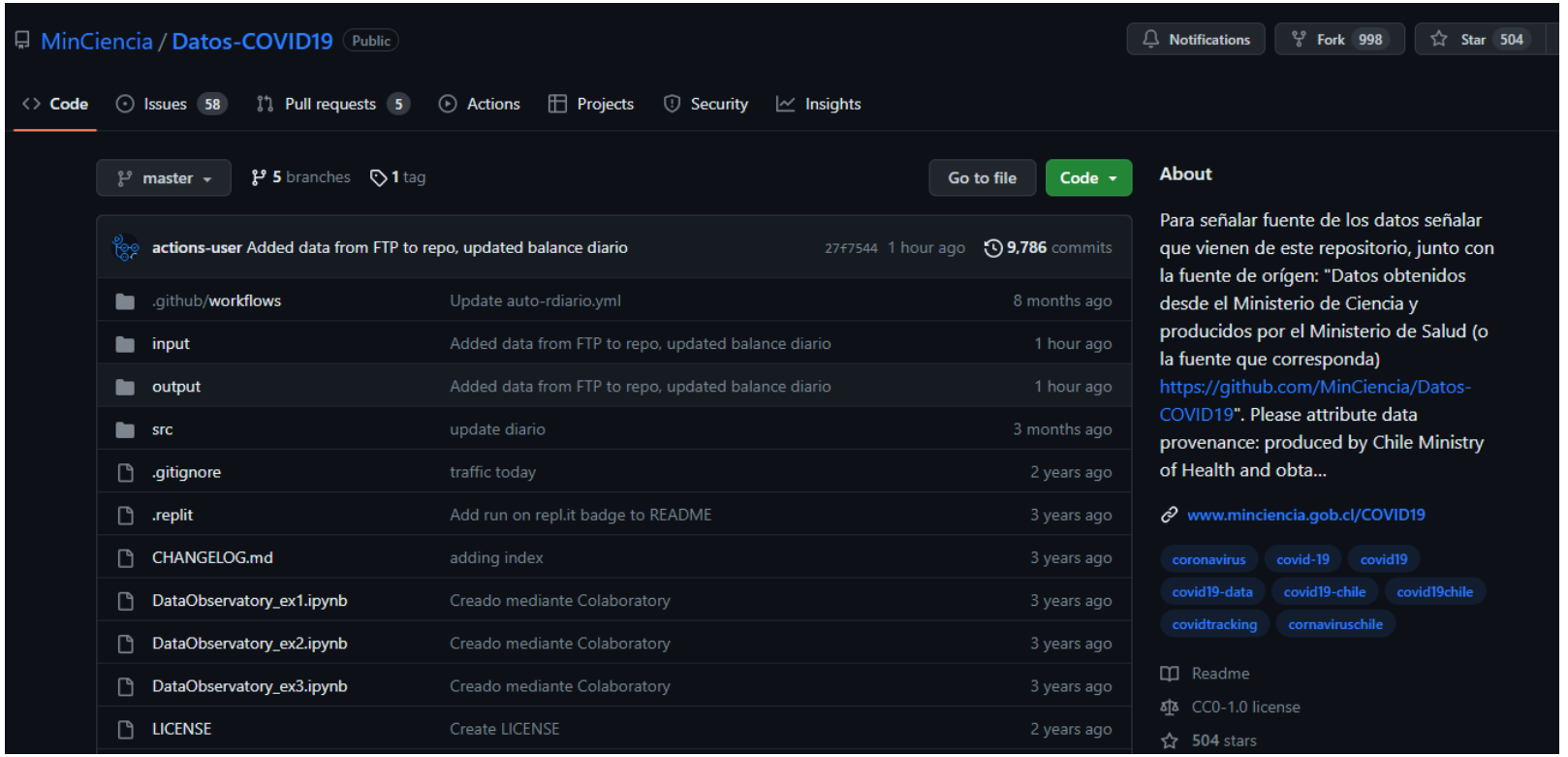

**Jupyter Notebook** es una aplicación web de código abierto que permite crear y compartir documentos interactivos que contienen código, visualizaciones, texto explicativo y otros elementos multimedia.

Además, los notebooks Jupyter son una herramienta importante para la investigación reproducible, ya que permiten compartir código, datos y resultados en un formato accesible y fácil de reproducir.


**GitHub** es una plataforma de alojamiento y colaboración de proyectos de software que utiliza el sistema de control de versiones Git. GitHub permite a los desarrolladores compartir y colaborar en proyectos de software de manera fácil y eficiente.

# Buenas Prácticas

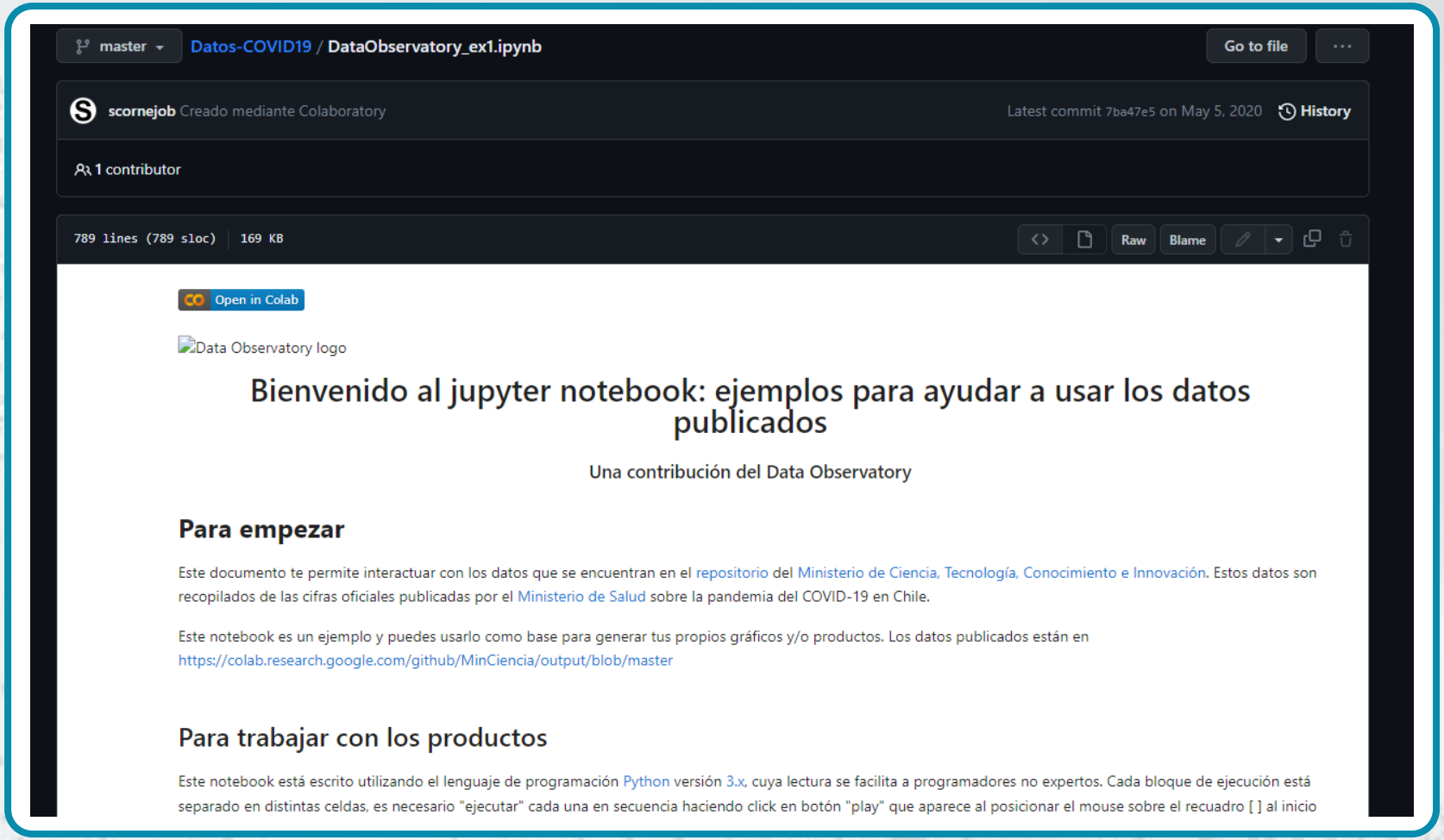
**El siguiente, es un ejemplo de investigación reproducible. Se trata del repositorio Github del ministerio de ciencias y tecnología, el cual comparte datos de las estadísticas de contagios Covid en Chile y almacena notebooks reproducibles para facilitar el análisis a los investigadores.**




KIBERNUM

# Buenas Prácticas

**Este es un ejemplo de un notebook reproducible. Recomendamos lo puedas revisar en detalle en el link.**




**https://github.com/MinCiencia/Datos-COVID19/blob/master/DataObservatory_ex1.ipynb**

KIBERNUM

Dudas y consultas
¡Gracias!