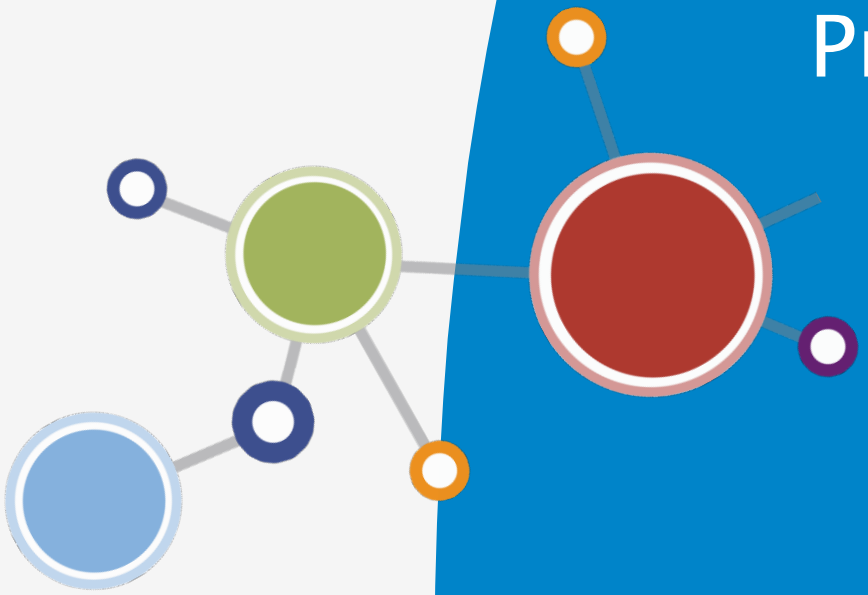
ICAP
Instituto Centroamericano de Administración Pública

Inteligencia de Datos

Prof. Catalina Artavia Pereira



Temas relevantes

01 **Contexto**
Programa del curso

02 **Introducción a inteligencia de Datos**

03 **Introducción a R**

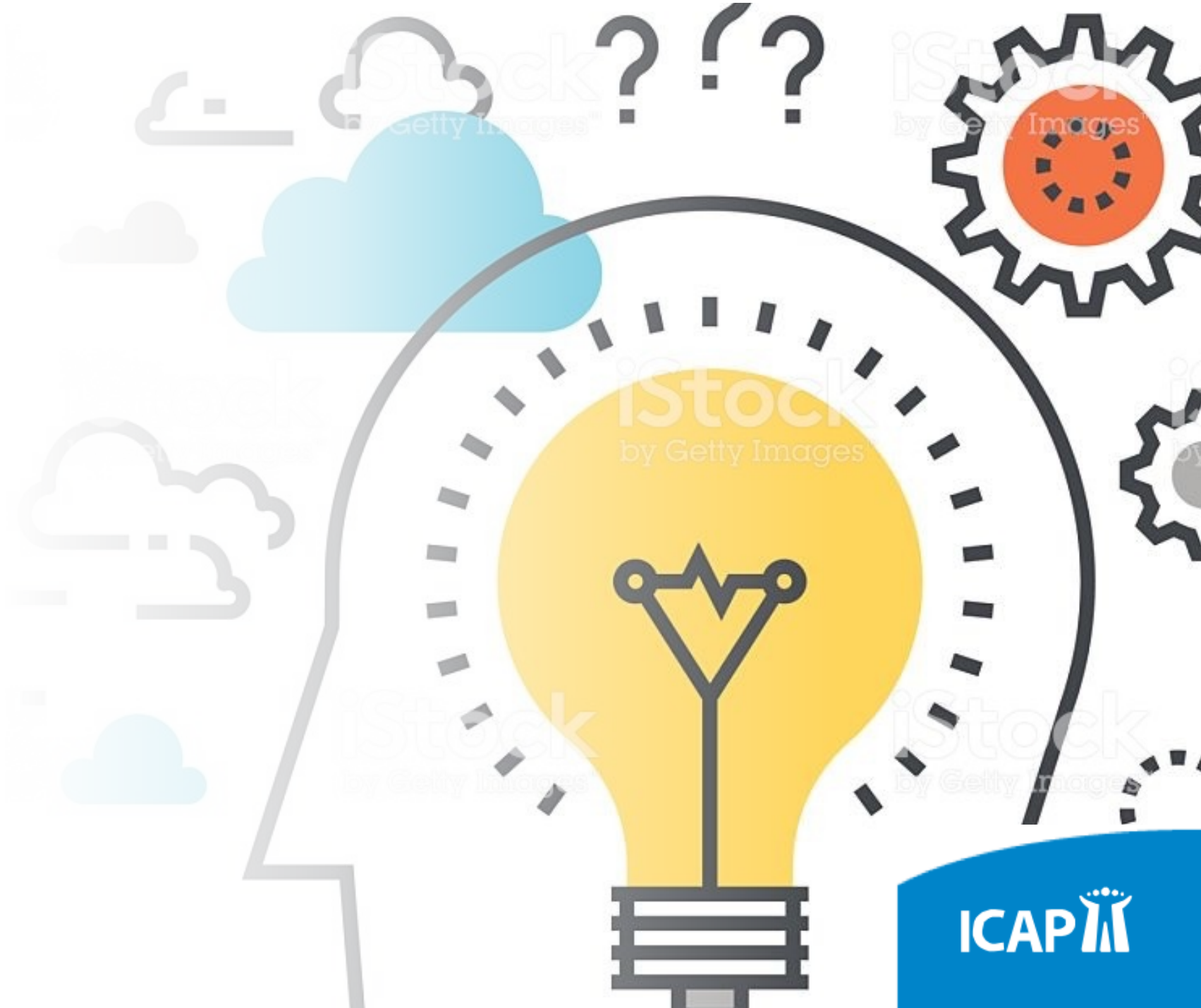
Tarea

Conceptos básicos

Minería de Datos:

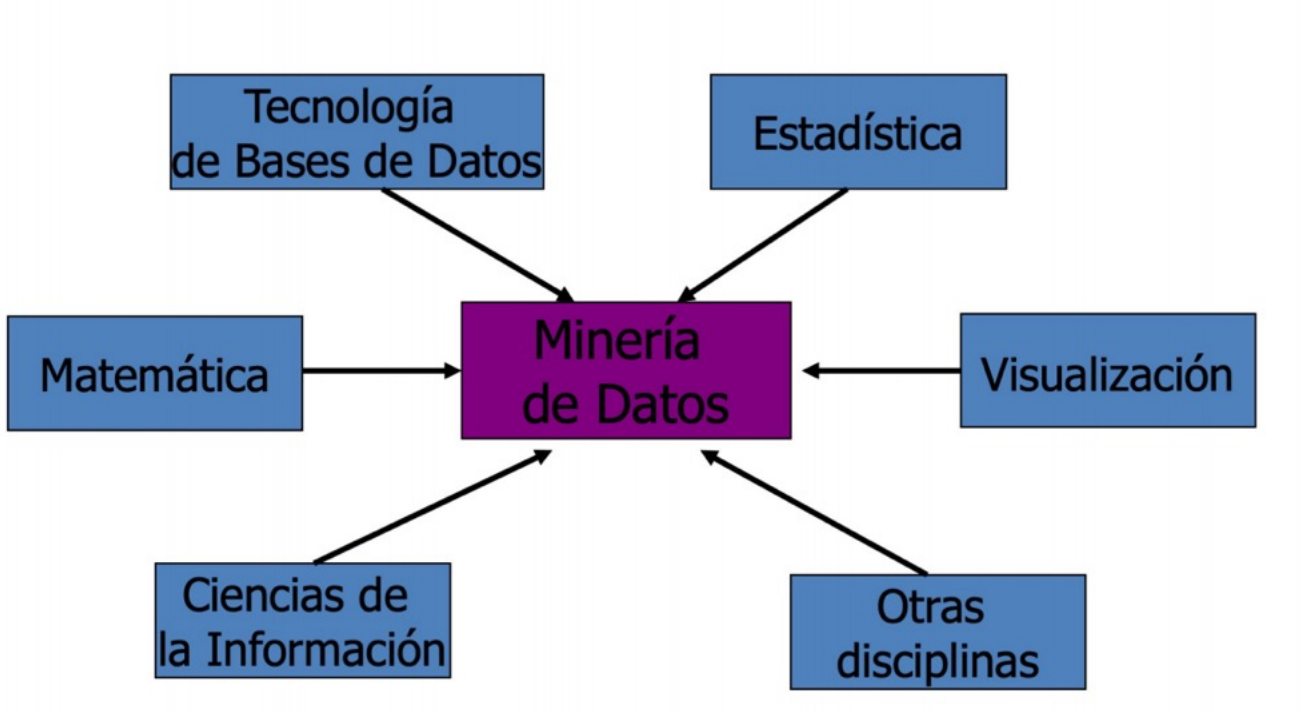
Extracción de información o de patrones (no trivial, implícita, previamente desconocida y potencialmente útil) de grandes bases de datos.

Es analizar datos para encontrar patrones ocultos usando medios automatizados.




¿Qué es ser un
Minero de Datos?

Confluencia de múltiples disciplinas



Crear un proceso automatizado que toma como punto de partida los datos y cuya meta es la ayuda a la toma de decisiones.

Minería de Datos

-Pretende buscar información útil usando toda la base datos.

-Usa técnicas mucho más exploratorias que vienen de la IA

Estadística

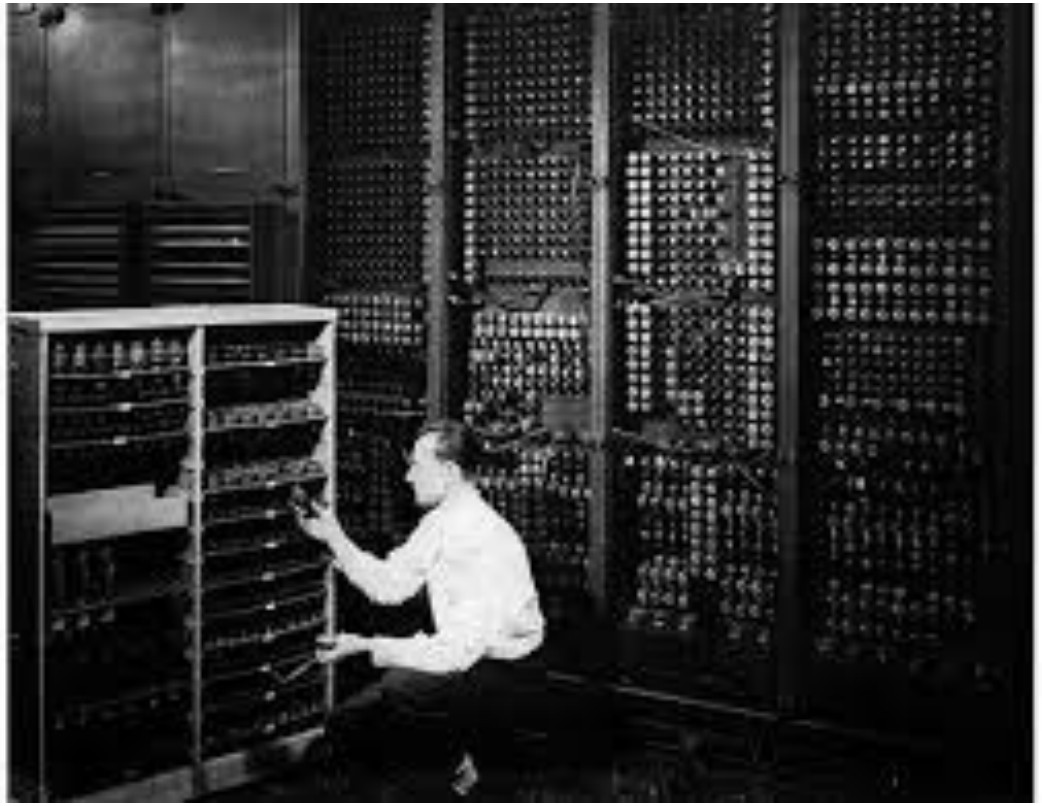
-Analiza muestras de datos para luego hacer inferencia a toda la población.

-En la mayoría de los casos supone que los datos se comportan de acuerdo a ciertas distribuciones de probabilidad (normal, binomial, geométrica, Poisson, etc),

Análisis de datos

-Con el advenimiento de las computadoras, aproximadamente en 1960, un nuevo concepto surgió del “matrimonio” entre la informática y la estadística: *El Análisis de Datos*

-Es analizar los datos con un objetivo decisional usa mucho más la informática y los métodos analíticos (el análisis de factorial, la clasificación automática, la discriminación, etc.) que los métodos estadísticos clásicos, las pruebas de hipótesis, que parten de supuestos matemáticos muy difíciles de verificar en la práctica.



A diferencia de la minería de datos, el análisis de datos usualmente no es automatizado, ni trata con volúmenes de datos tan grandes.

¿Y Machine Learning?

– *“Machine Learning”*: es un área de la Inteligencia Artificial (IA) que trata sobre como escribir *programas puedan aprender*.

– En “Data Mining” es usualmente usado para predicción y clasificación.

Big Data

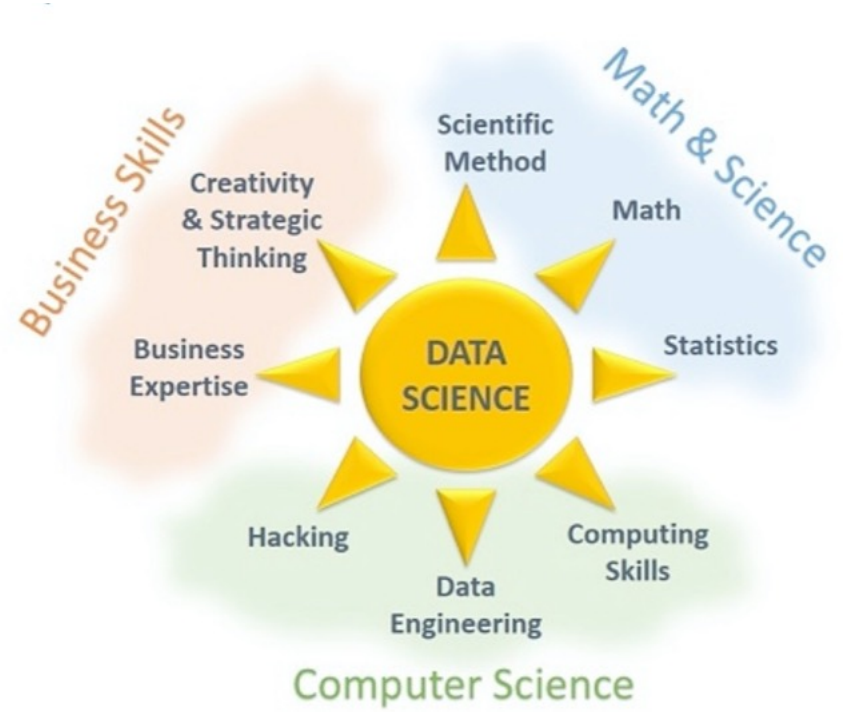
Big Data: termino de moda en el mundo de la informática y de la Administración de Negocios (MBA)

Se populariza con el concepto de:

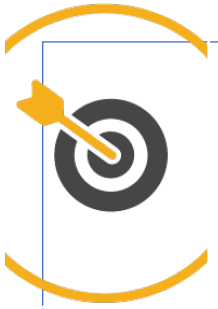
Conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos tradicionales.

Ciencia de Datos

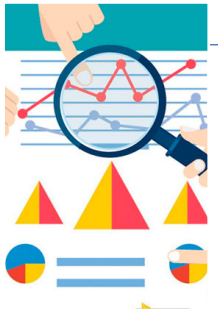
campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva.



¿Qué es un científico de datos?



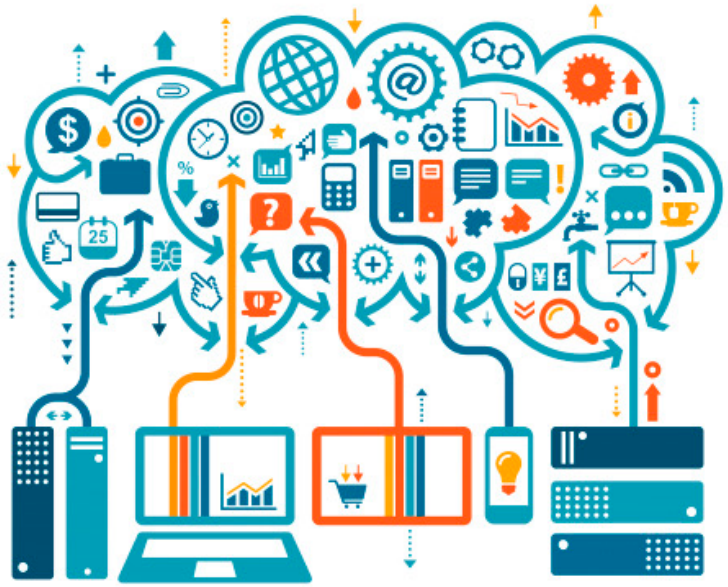
Profesional dedicado a analizar e interpretar grandes bases de datos.



Capacitado para crear sus propios modelos dado un juego de datos.

Conoce fundamentos de los métodos y no sólo usa “cajas negras”

Persona que sabe más de estadística que cualquier programador y que a la vez sabe más de programación que cualquier estadístico.



Big Data

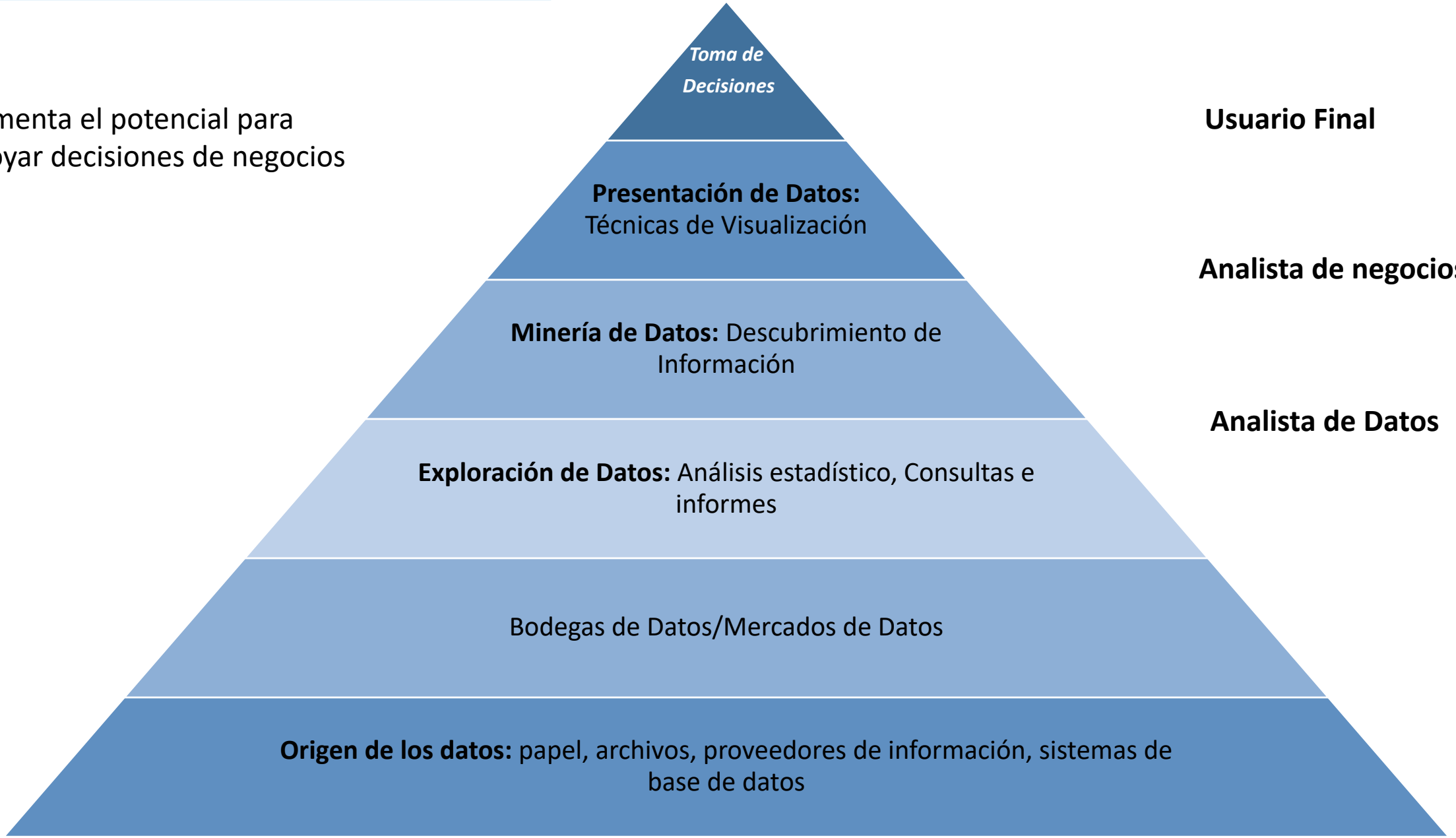
- **Evolución de la capacidad de almacenamiento**

- **Explosión de los datos**

"Big data" son activos de información caracterizados por su alto ***volumen, velocidad*** y ***variedad***, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones.

- Volumen: …
- Velocidad: …
- Variedad: …
- ***Veracidad:*** Por último el ***Big Data*** ha de ser capaz de tratar y analizar inteligentemente este inmenso volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar la toma de decisiones en las organizaciones.

Aumenta el potencial para apoyar decisiones de negocios



Usuario Final

Analista de negocios

Analista de Datos

Aplicaciones para la toma de decisiones

Su aplicación abarca todas las disciplinas y enfoques:

Retención de Clientes

Patrones de Compra

Detección de Fraude

Manejo del Riesgo

Segmentación de clientes

Predicción de Ventas

¿Cuáles clientes se van ir para la competencia?

¿Cuándo un cliente compra un producto cuál otro le podría interesar?

¿Cuáles transacciones son fraudulentas?

¿A qué clientes les doy un préstamo?

¿Quiénes son mis clientes?

¿Cuánto voy a vender el próximos mes?

¿Por qué es importante?

- Desde un punto de vista comercial
  - Muchos datos están siendo generados y almacenados, datos de la Web, comercio electrónico.
  - Las compras
  - Bancos / tarjeta de crédito
  - Millones de transacciones
- Las computadoras se han vuelto más baratas y más potentes
- La presión para la calidad es fuerte:
  - Proporcionar mejores y más servicios personalizados

APLICACIÓN PARA LA TOMA DE DECISIONES

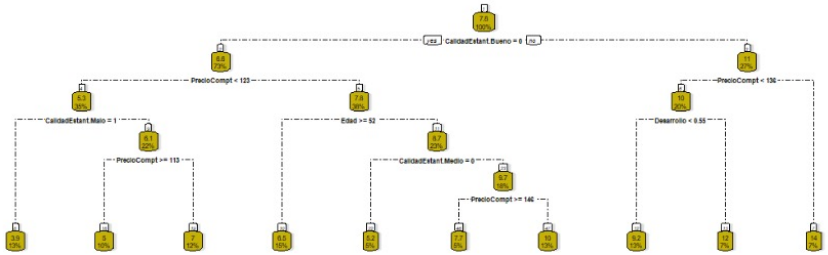
¿CÓMO TOMAR DECISIONES BASADO EN DATOS?

Casos para la toma de decisiones

Caso 1: Predicción de buen o mal pagador
Matriz de confusión

		Predicción	
		Mal Pagador	Buen Pagador
Valor Real	Mal Pagador	800	200
	Buen Pagador	500	1500

- 800 predicciones de Mal Pagador fueron realizadas correctamente, para un 80%, mientras que 200 no, para un 20%.
- 1500 predicciones de Buen Pagador fueron realizadas correctamente, para un 75%, mientras que 500 no (para un 25%).
- En general 2300 de 3000 predicciones fueron correctas para un 76,6% de efectividad en las predicciones. ***Cuidado***, este dato es a veces engañoso y debe ser siempre analizado en la relación a la dimensión de las clases.



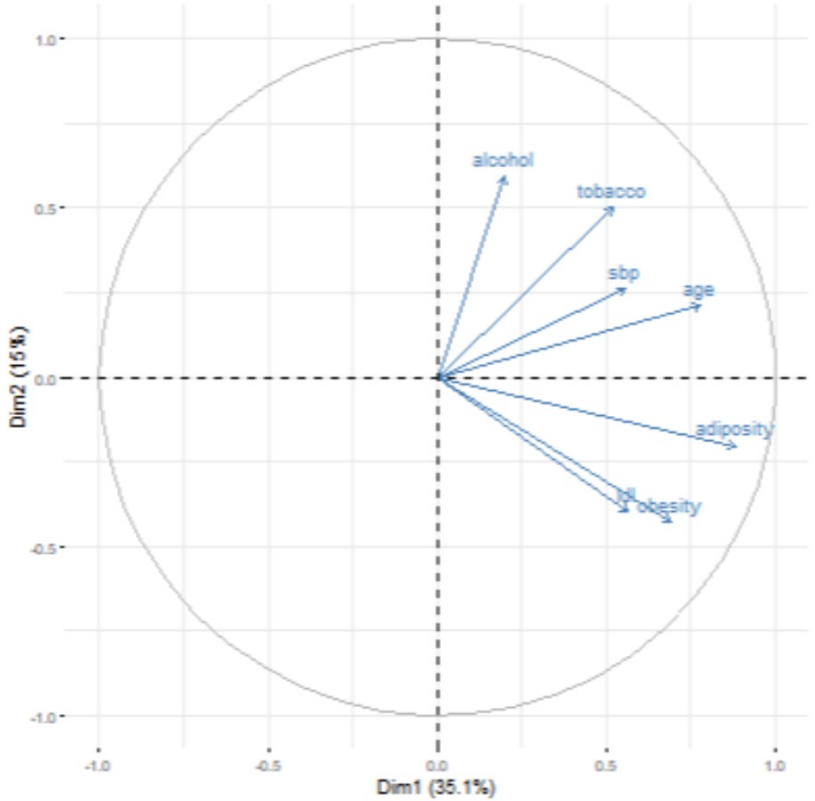
```
Rule number: 10 [Ventas=6.46666666666667 cover=9 (15%)]
  CalidadEstant.Bueno< 0.5
  PrecioCompt>=123
  Edad>=52

Rule number: 12 [Ventas=9.20875 cover=8 (13%)]
  CalidadEstant.Bueno>=0.5
  PrecioCompt< 135.5
  Desarrollo< 0.553

Rule number: 47 [Ventas=10.445 cover=8 (13%)]
  CalidadEstant.Bueno< 0.5
  PrecioCompt>=123
  Edad< 52
  CalidadEstant.Medio>=0.5
  PrecioCompt< 145.5
```

```
Rule number: 8 [Ventas=3.87125 cover=8 (13%)]
  CalidadEstant.Bueno< 0.5
  PrecioCompt< 123
  CalidadEstant.Malo>=0.5

Rule number: 19 [Ventas=7.05 cover=7 (12%)]
  CalidadEstant.Bueno< 0.5
  PrecioCompt< 123
  CalidadEstant.Malo< 0.5
  PrecioCompt< 112.5

Rule number: 18 [Ventas=4.99833333333333 cover=6 (10%)]
  CalidadEstant.Bueno< 0.5
  PrecioCompt< 123
  CalidadEstant.Malo< 0.5
  PrecioCompt>=112.5
```

Regla 10: El valor promedio de las ventas de asientos de niños para autos en cada localidad va a ser de 6,46 unidades con un cover de 9 casos que son el 15%, si la calidad estándar buena es menor a 0,5, si el precio promedio es mayor o igual a 123 y si las edades son mayores o guales a 52 años.

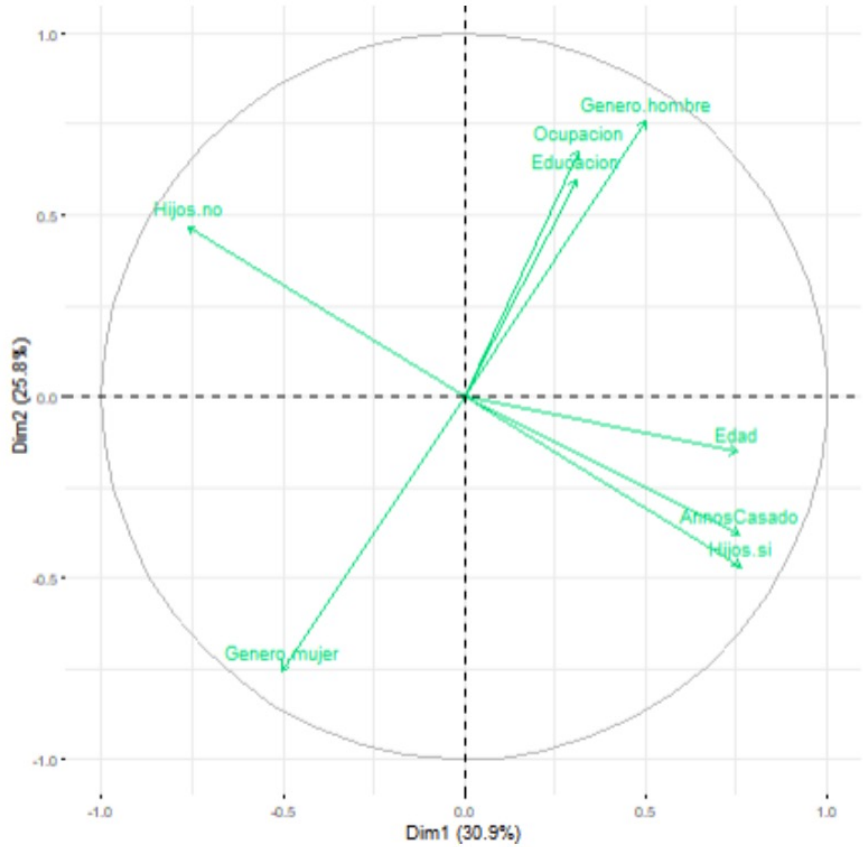
Regla 12: El valor promedio de las ventas de asientos de niños para autos en cada localidad va a ser de 9,20 unidades con un cover de 8 casos que son el 13%, si la calidad estándar buena es mayor o igual a 0,5, si el precio promedio es menor a 135,5 y si el desarrollo es menor a 0,553.

Regla 47: El valor promedio de las ventas de asientos de niños para autos en cada localidad va a ser de 10,44 unidades con un cover de 8 casos que son el 13%, si la calidad estándar buena es menor a 0,5, si el precio promedio es mayor o igual a 123 pero menor a 145,5, si la edad es menor a 52 años, y si la calidad estándar medio es mayor o igual a 0,5.



- Se presenta una correlación fuerte y positiva entre las variables **alcohol y tabaco**, esto debido a que se presenta un ángulo relativamente pequeño, es decir, a mayor consumo de alcohol de los hombres en una región de alto riesgo de enfermedad cardíaca de la Provincia Occidental del Cabo, Sudáfrica, mayor es el consumo de tabaco en kilogramos. O bien a mayor consumo de tabaco, mayor consumo de alcohol.
- Se presenta una correlación fuerte y positiva entre las variables **tabaco y edad**, esto debido a que se presenta un ángulo relativamente pequeño, es decir, a mayor consumo de tabaco de los hombres en una región de alto riesgo de enfermedad cardíaca de la Provincia Occidental del Cabo, Sudáfrica, mayor es la edad de esos individuos. O bien a mayor edad de los individuos, mayor consumo de tabaco.

Caso 3:Análisis de estudios

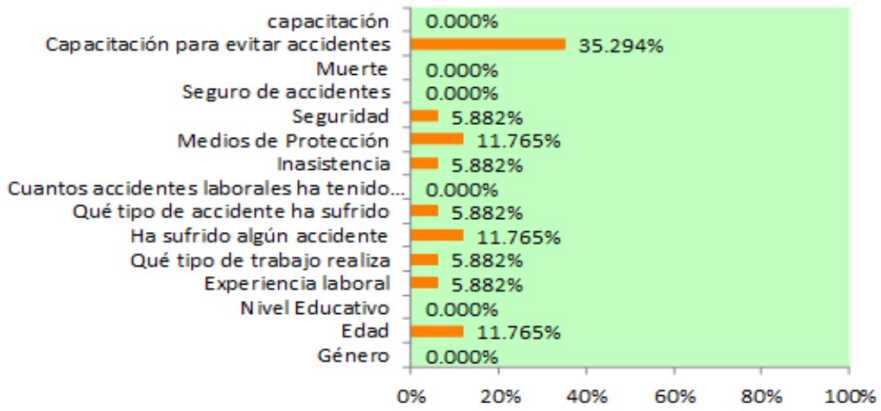


- Se presenta una correlación fuerte positiva entre los años de casados y la edad de los individuos, ya que se presenta un ángulo relativamente pequeño, es decir, a mayor años de casados de los individuos que incurrieron en una infidelidad registradas en la base de datos, mayor es la edad de esos individuos. O bien a mayor es la edad de los individuos que incurrieron en una infidelidad registradas en la base de datos, mayor es la cantidad de años de casados.
- Se presenta una correlación fuerte positiva entre los años de casados de los individuos y los hijos con los que cuentan en el matrimonio. Esto ya que se presenta un ángulo relativamente pequeño, es decir, a mayor cantidad de años de casados de los individuos

Casos para la toma de decisiones

Caso 4: Investigación aplicada sobre los accidentes de la compañía COSMO ASTRAL mediante el díselo de redes neuronales de inteligencia artificial para su uso en la toma de decisiones.

Calcular las probabilidades de ocurrencia de un accidente de cualquier empleado para que la empresa pueda focalizar sus esfuerzos de manera individual o grupal en aquellos empleados que representan mayor riesgo.



¿Qué es R?

¿Qué es R?

- Es un lenguaje enfocado en el análisis de datos, estadística, minería de datos y visualización de modelos.

- Es un software libre (gratuito).

- Se destaca por ser un programa relativamente sencillo porque permite compartir código de forma fácil.

- “Es un lenguaje creado por un estadístico para hacer estadística”. *John Chambers.*

¿Y qué es RStudio?

ICAP

¿Qué es R y RStudio?

- Conjunto de programas integrados para el manejo de datos, simulación.
- *Es el "motor".*

- *Es el "Dashboard o el volante"*
- *Permite usar el lenguaje de programación.*

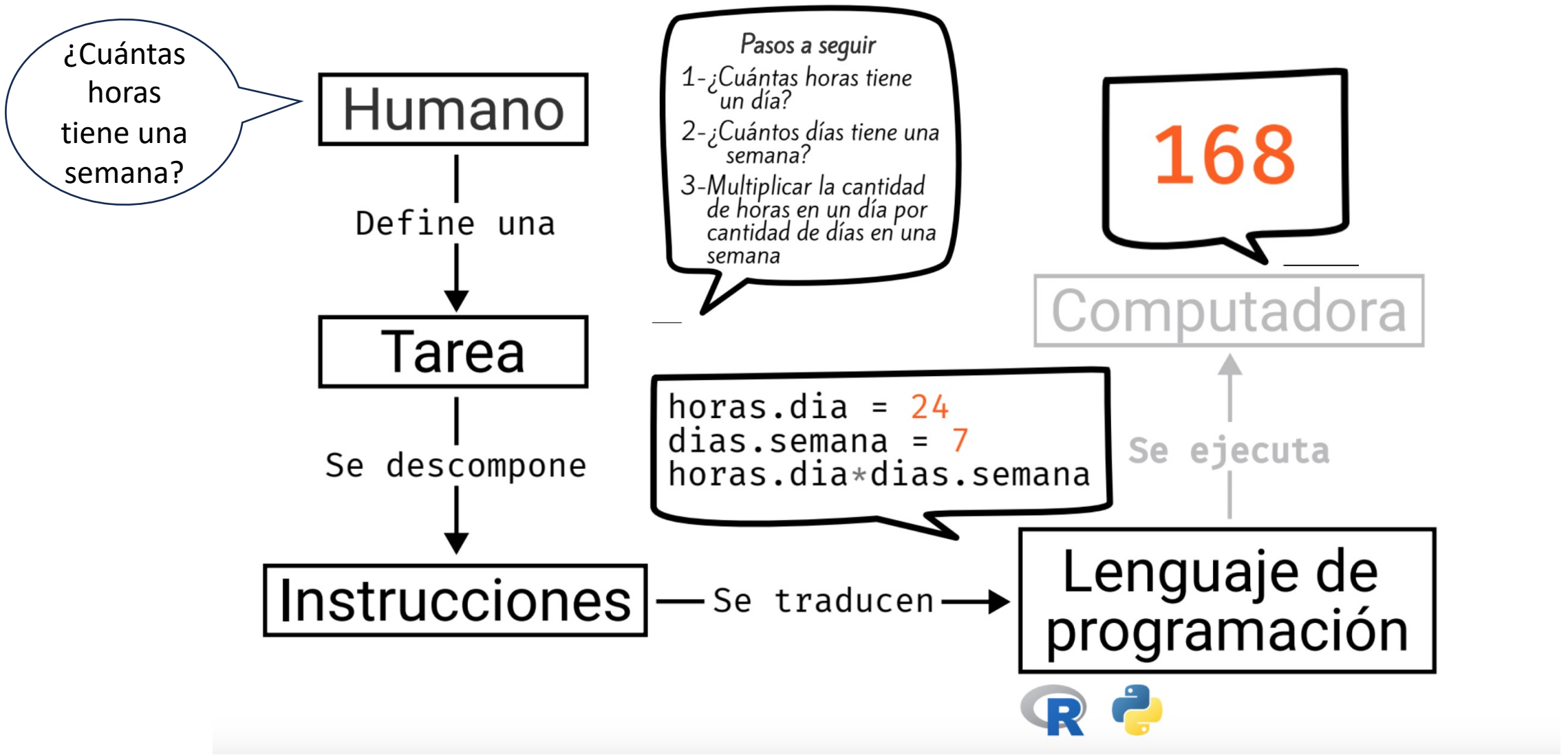
ICAP

¿Cuál es su utilidad?

- Permite hacer análisis estadístico.
- Visualización de datos, graficar.
- Utilizado por analistas de datos, actuarios, economistas, estadísticos, académicos, analistas financieros, estudiantes en investigaciones.

ICAP

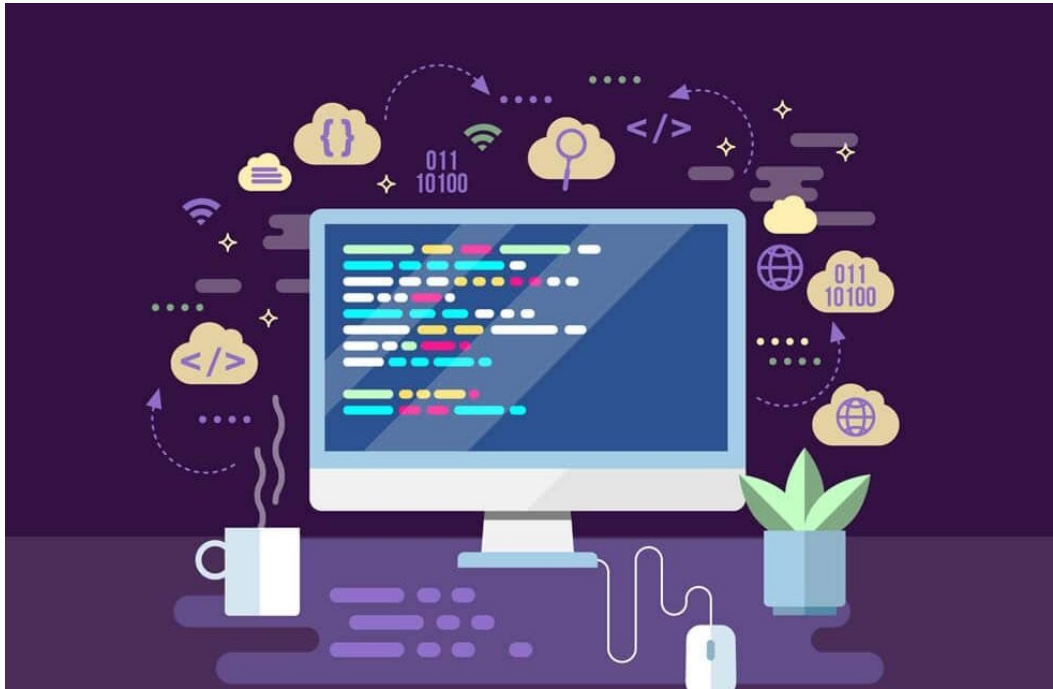
¿Qué es Programar?



Fuente: Promidat, 2020.

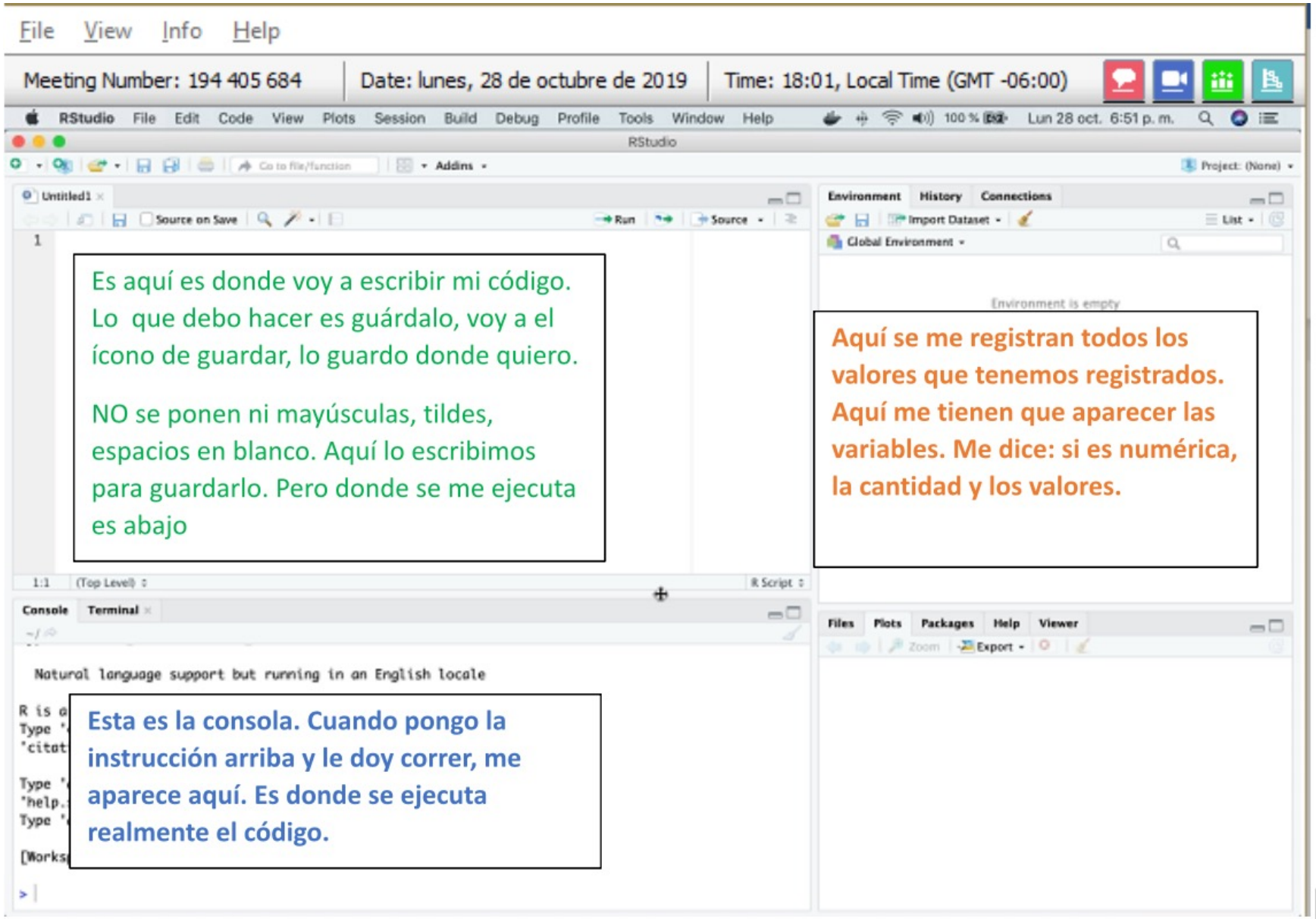
¿Qué es Programar?

- Programar es tomar instrucciones y traducirla a un lenguaje de programación.
- Es pasar instrucciones humanas a un lenguaje que la computadora puede procesar y dar un resultado.



Fuente: Promidat, 2020.

Comprendiendo R…



DESCARGAR

- Descarga e instalación de R
- Para descargar R podemos acceder al siguiente enlace:
- https://cran.r-project.org/
- En este enlace vamos a poder encontrar los instaladores respectivos para Linux, Windows o Mac.

DESCARGAR

- Para descargar RStudio podemos acceder al siguiente enlace:
- https://www.rstudio.com/products/rstudio/download/

Tarea Individual

1. Realizar la descarga e instalación de R.
2. Realizar la descarga e instalación de Rstudio.
3. Ingresar a Rstudio y realizar las siguientes actividades:

3.1 Crear las variables con los siguientes valores:
Pacientes_enero<- 150
Pacientes_febrero<- 100
Pacientes_marzo<- 50

3.2 Realizar la sumatoria de las tres variables creadas, de forma que se cuente con la suma de los pacientes de enero, febrero y marzo.

3.3 Realizar el promedio de las tres variables creadas de forma que se cuente con el promedio de pacientes atendidos en los meses de enero, febrero y marzo.

3.4 Crear un R Markdown y copiar los códigos realizados en los puntos anteriores.

3.5 Descargue el reporte en formatos: Html y Word y suba ambos archivos en el Campus Virtual.

Fecha de entrega: 13 de enero, **Hora:** 23:59pm.

Muchas gracias

La coyuntura actual presenta un panorama sin precedentes, nunca antes las comunicaciones, la información y las tecnologías han estado tan cerca de los individuos, por eso la información generada por todos y todas; los datos, son la una de las principales herramientas para encontrar soluciones a las nuevas urgencias (Vargas, Elizondo y Bonilla, 2019, p.16).

"Los datos tienen un poder casi especial para describir el mundo".
"La ciencia de datos es la disciplina de hacer que los datos sean útiles"