

Programación para Data Science

Sesión 1: Introducción

Raimundo Sánchez, PhD

- Profesor de Modelamiento Matemático, Advanced Analytics, Data Science.
- Ingeniero Industrial, U. Adolfo Ibáñez (2008).
- Doctor en Sistemas Complejos, U. Adolfo Ibáñez (2015).
- Gerente de Revenue Management Analytics en LATAM Airlines (2015 – 2019)
- 10 años de experiencia en consultoría de Data Science.

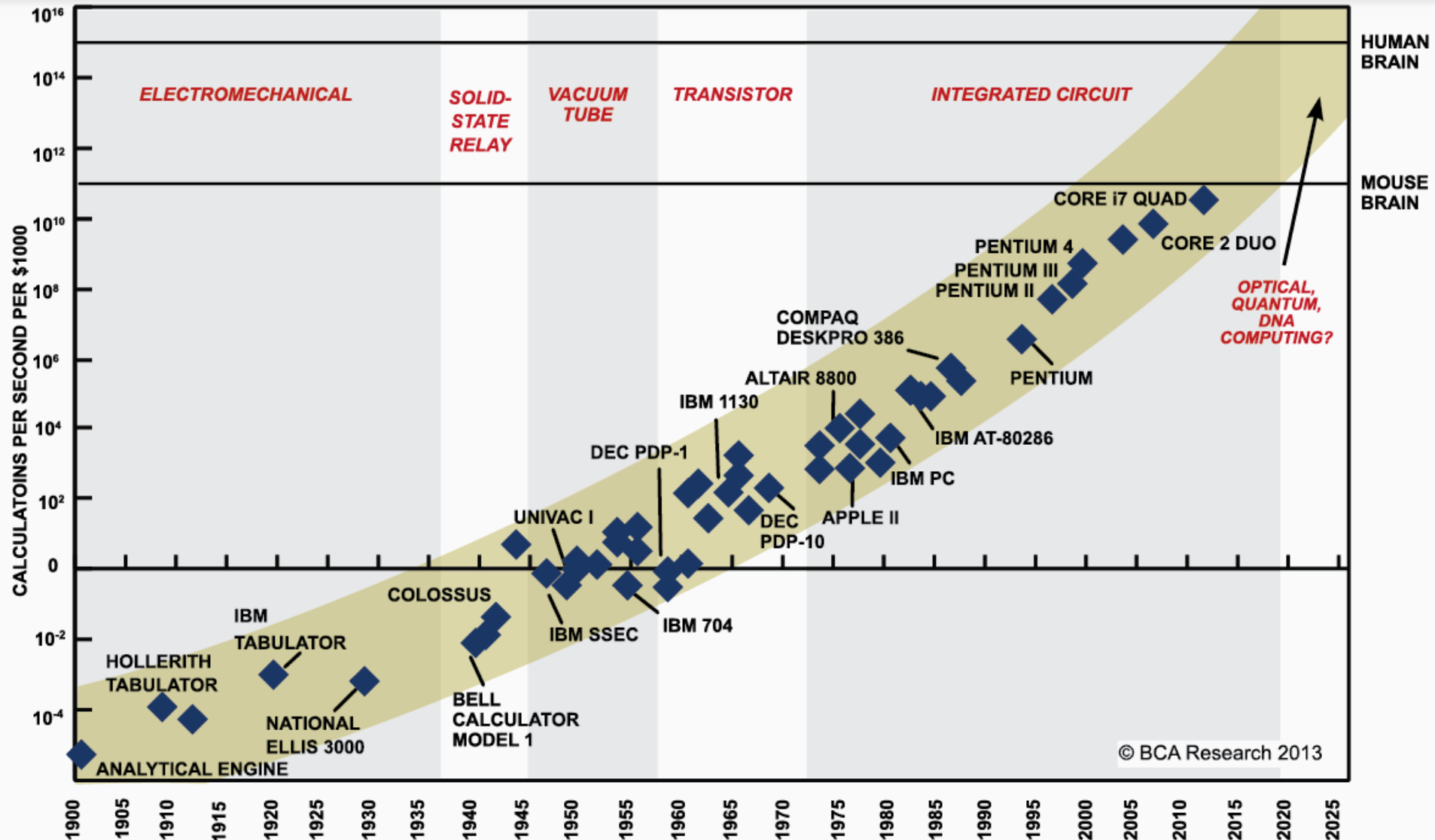
Agenda del curso de programación

- 7 módulos:
 - 1 modulo introducción
 - 3 módulos R
 - 3 módulos Python (Miguel Carrasco)



¿Se puede hacer ciencia de datos sin programación?

Singularidad cognitiva



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, THE VIKING PRESS, 2006. DATAPPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

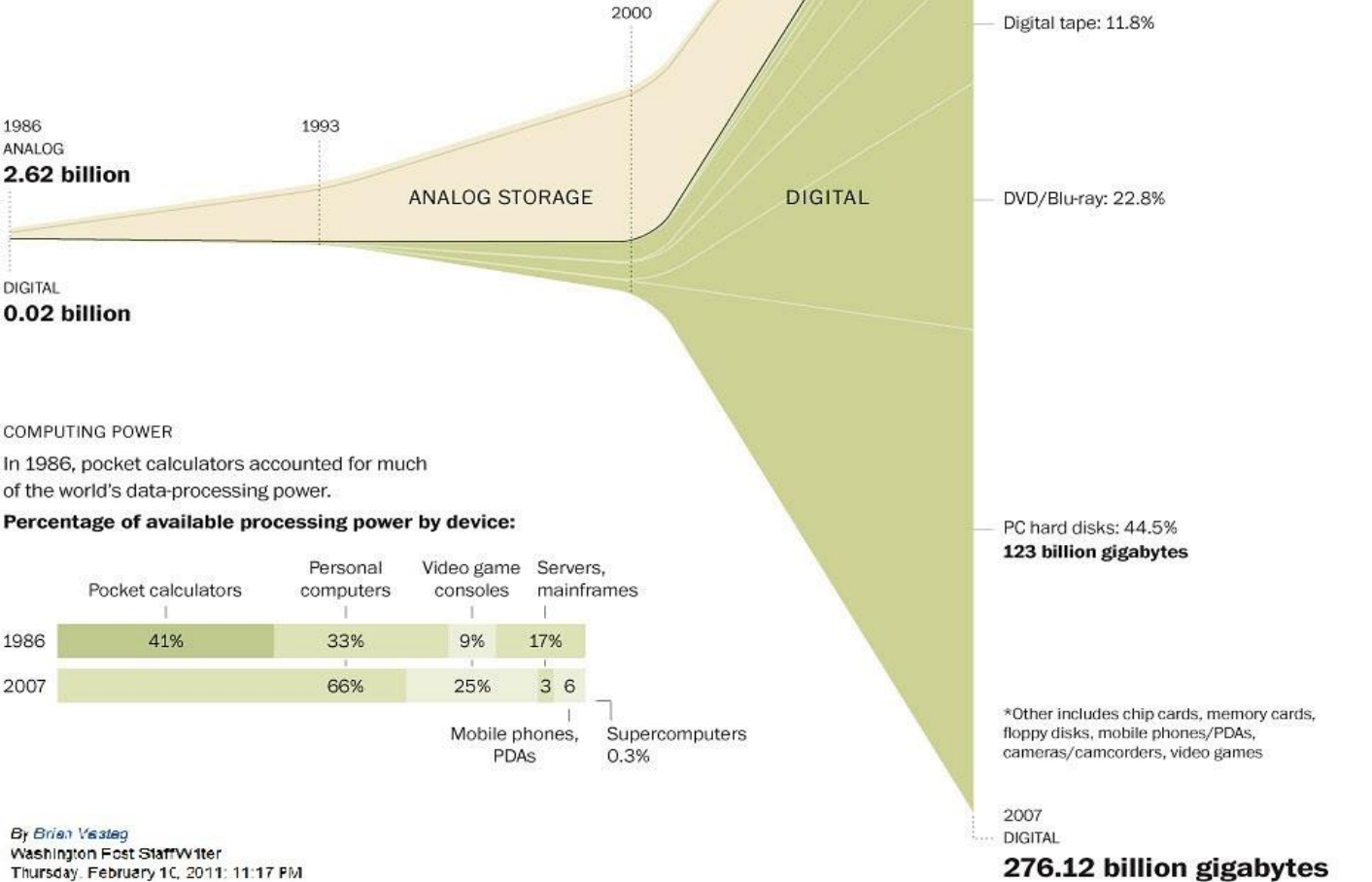
The Washington Post

Exabytes: Documenting the 'digital age' and huge growth in computing capacity

THE WORLD'S CAPACITY TO STORE INFORMATION

This chart shows the world's growth in storage capacity for both analog data (books, newspapers, videotapes, etc.) and digital (CDs, DVDs, computer hard drives, smartphone drives, etc.)

In gigabytes or estimated equivalent



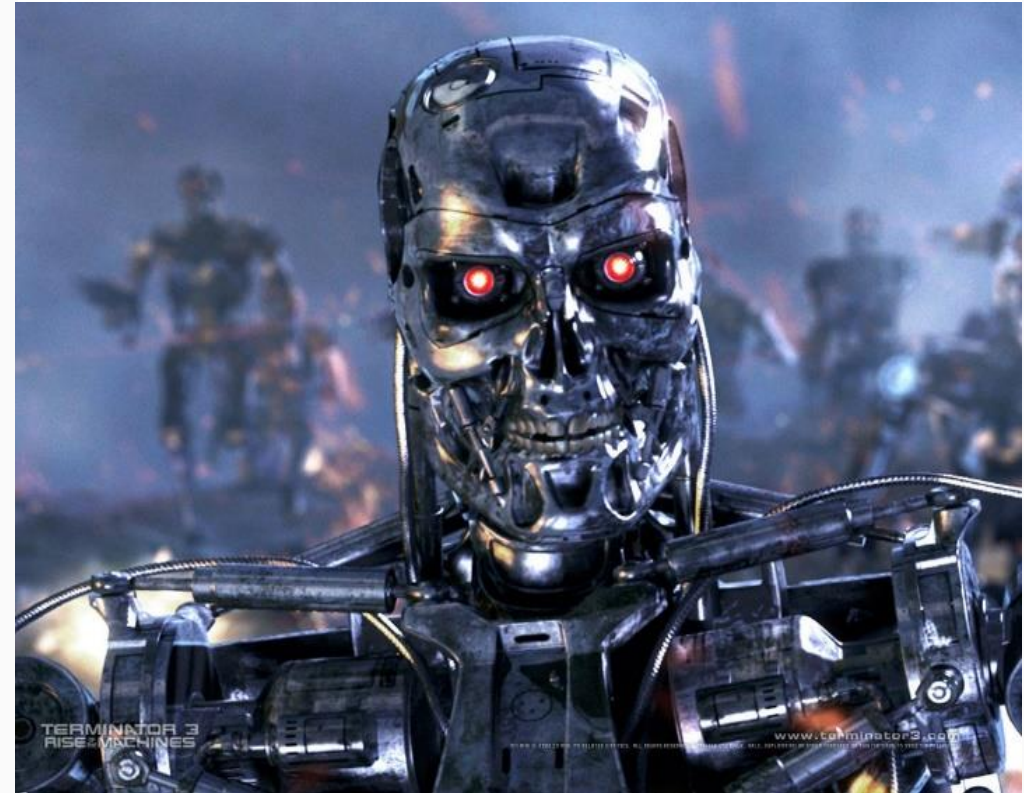
El futuro del trabajo: The rise of the machines

- Parte del trabajo en Data Science es supervisar, desarrollar o mejorar procesos automatizados.
- Todo lo que pueda ser automatizado, será automatizado.
- La creatividad humana no es automatizable y se volverá muy valiosa.
- Compartir y colaborar es cada vez mas importante.



3 Niveles de la “Inteligencia” Artificial

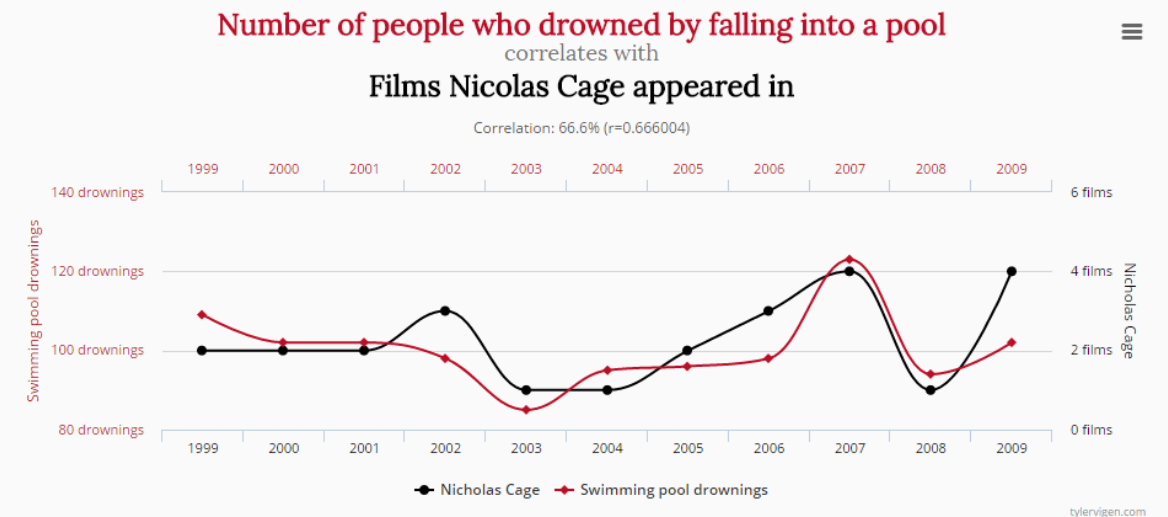
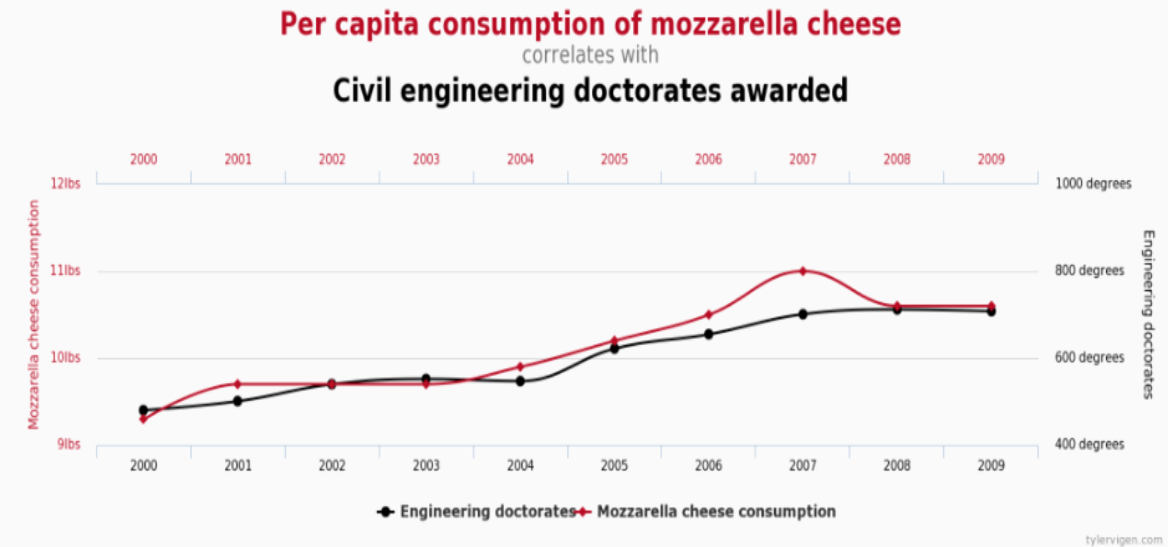
- Sistemas automáticos: Programas que manipulan conocimiento simbólico explícito
- Redes Neuronales Artificiales: Programas que imitan (vagamente) el comportamiento neuronal.
- Robotica Cognitiva: Programas capaces de interactuar con entornos reales de manera autónoma.



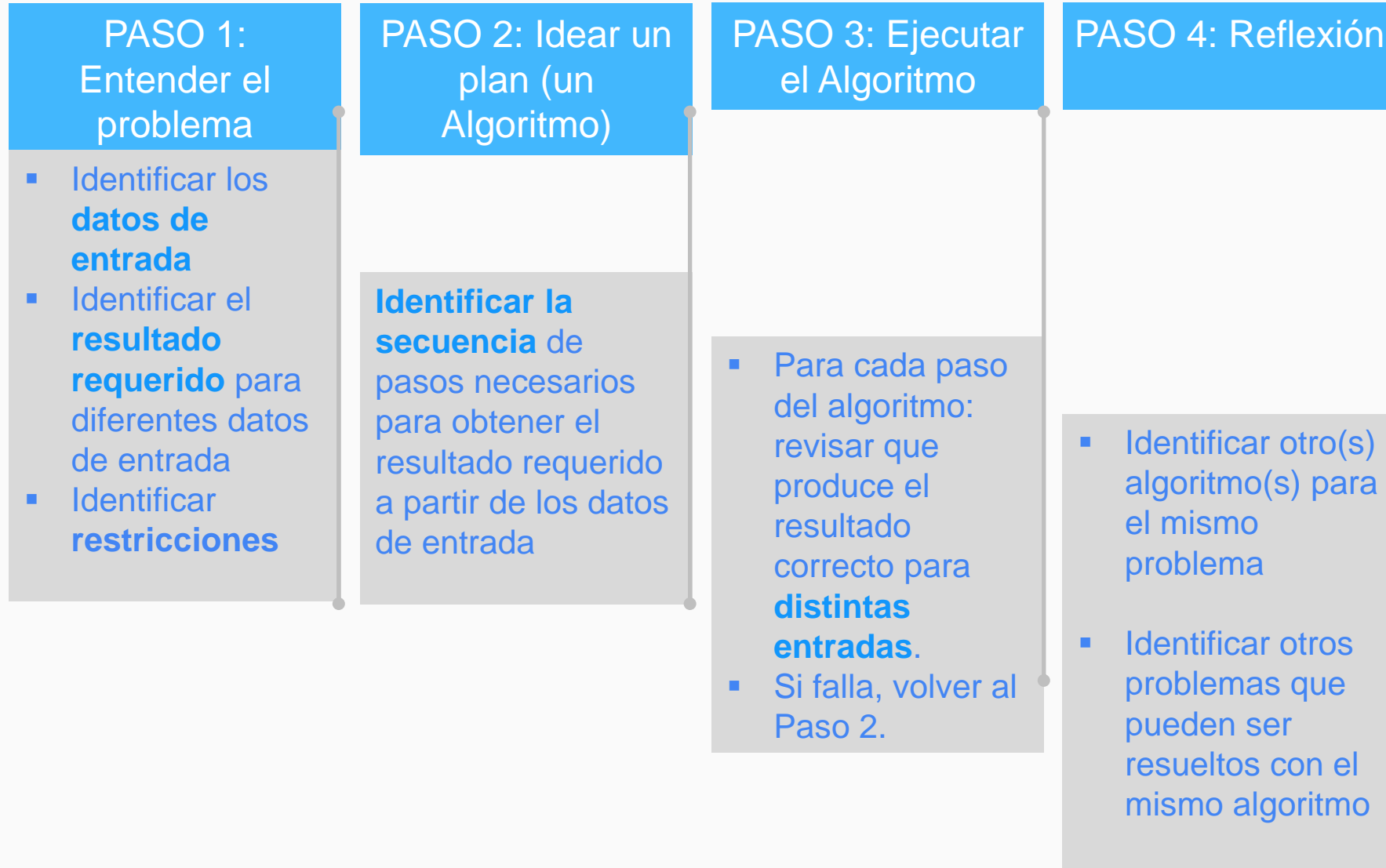
No basta con robots

- Sistemas de inteligencia artificial son excelentes en encontrar relaciones espurias
- Modelos de caja negra pueden llevar a conclusiones incorrectas

Los humanos aún debemos hacer las preguntas!



Método científico

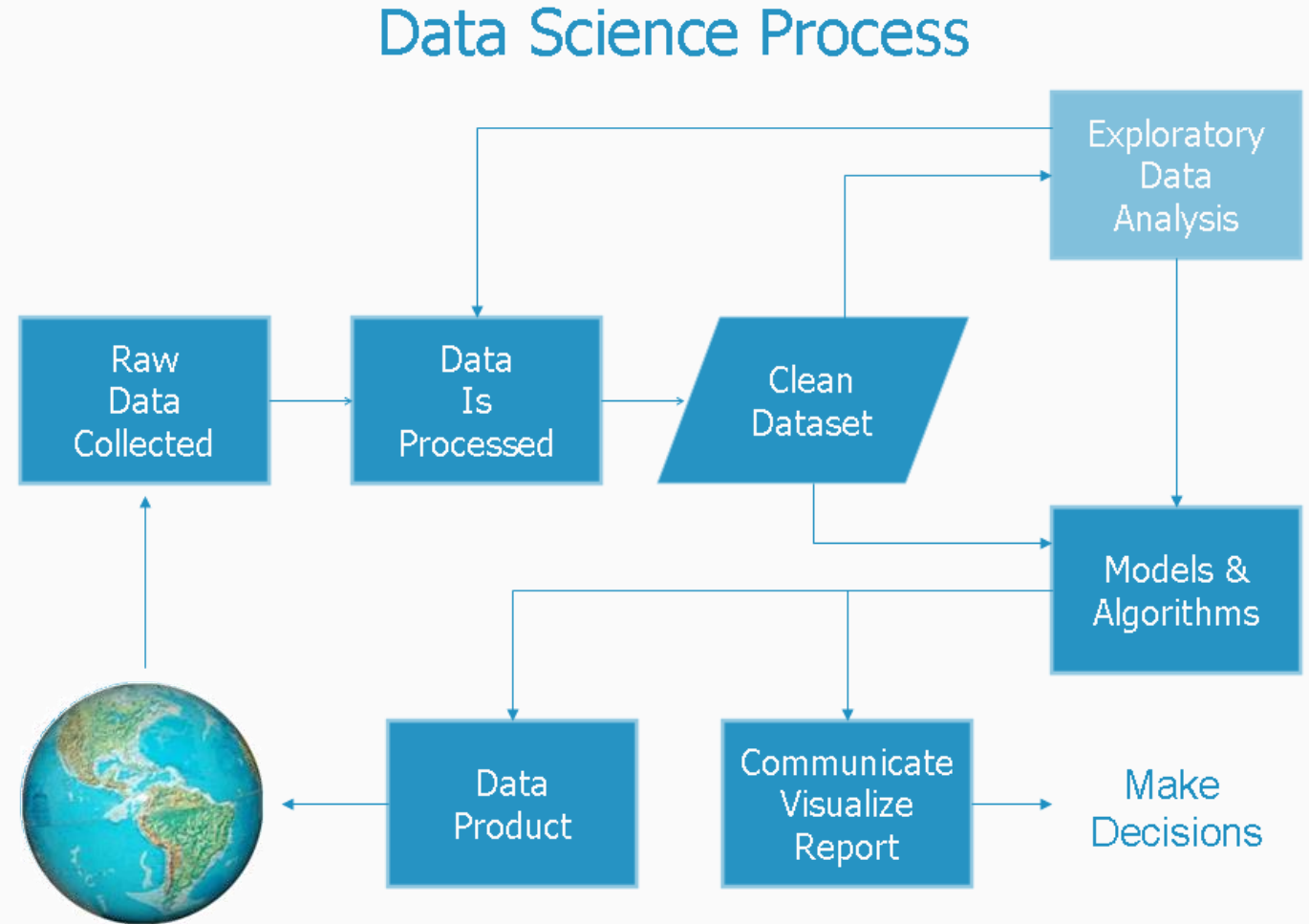


Principios de la ciencia de datos

1. Los datos nunca están limpios
2. Pasarás la mayor parte del tiempo limpiando y preparando datos.
3. En el 80% de los casos un modelo lineal hará el truco.
4. Bigdata es solo una herramienta
5. La presentación es clave
6. Todos los modelos están mal, pero algunos son útiles.
7. **Necesitas ensuciarte las manos**

Para trabajar en Data Science necesitamos herramientas que nos permitan:

1. Recolectar datos
2. Procesar altos volúmenes de manera eficiente
3. Análisis exploratorio de datos
4. Utilizar modelos de diversa complejidad
5. Visualizar y exportar datos / resultados




¿Programación?

Algoritmo + Sintaxis

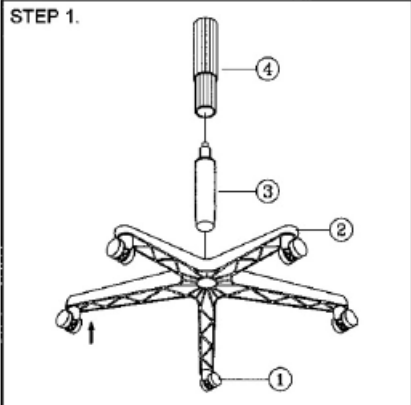
Algoritmo

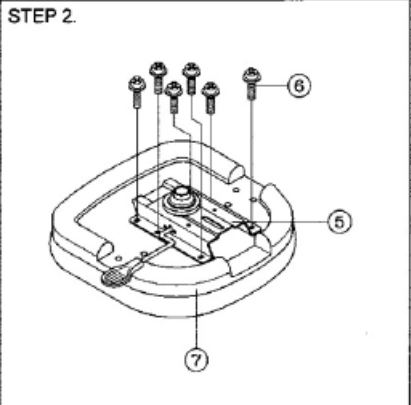
Secuencia de pasos
necesarios para
ejecutar una tarea

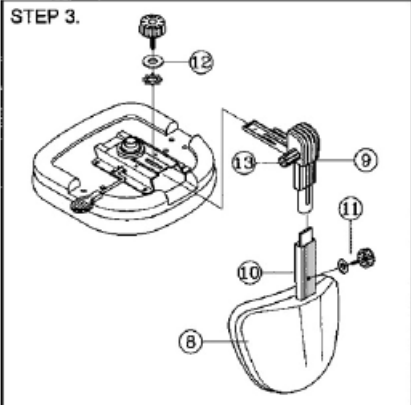
**Assembly Instruction for
Ergonomic Swivel Office / Task Chair
“PREMO” No. CH444**


PART LIST


KEY	QTY	DESCRIPTION	KEY	QTY	DESCRIPTION	KEY	QTY	DESCRIPTION
1	5	Caster	7	1	Seat	12	1	Backrest Depth Adj. Knob + Washers
2	1	Base	8	1	Backrest	13	1	Knob for Inward & Outward
3	1	Seat Post	9	1	Connector Bar + Bellow			
4	1	Seat Post Cover	10	1	Backrest Bellow			
5	1	Mechanism	11	1	Backrest Height Adj. Knob + Washer			
6	6	Mechanism Screw						

STEP 1.







STEP 2.

STEP 3.

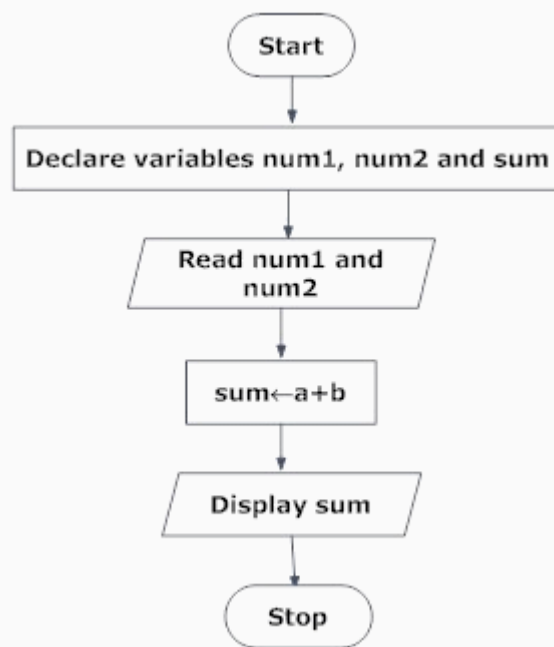


**Alvin & Company, Inc.** Bloomfield, Ct Grand Prairie, TX www.alvinco.com

Lenguaje para graficar procesos / algoritmos

Symbol	Purpose	Description
	Flow line	Used to indicate the flow of logic by connecting symbols.
	Terminal(Stop/Start)	Used to represent start and end of flowchart.
	Input/Output	Used for input and output operation.
	Processing	Used for arithmetic operations and data-manipulations.
	Desicion	Used to represent the operation in which there are two alternatives, true and false.
	Database	Used to represent data origins or destinations.

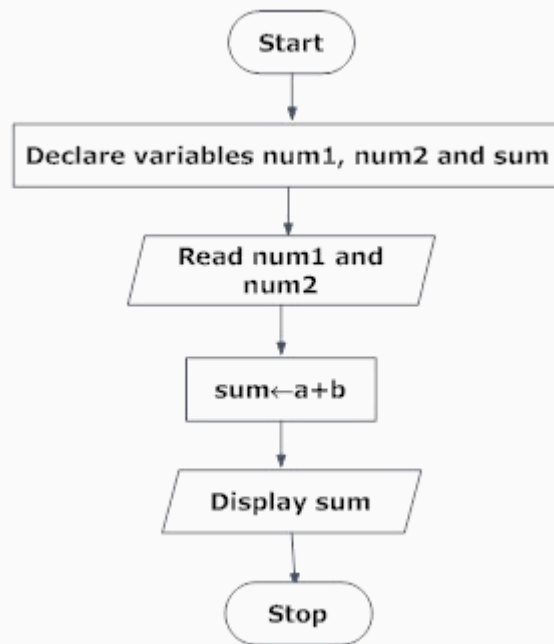
Dibuje un diagrama de flujo para sumar dos números ingresados por el usuario.



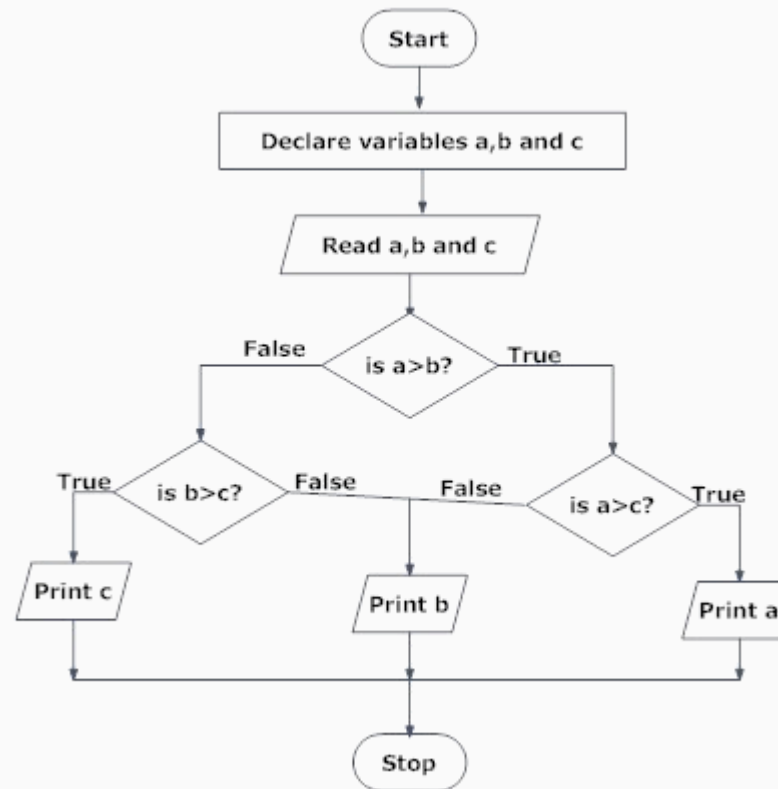
Variables

- Los datos (en R) puede tener variables de diversas clases:
 - Numeric, Character, Logical, Integer, Factor, Complex
 - Vector, Matrix, Data.frame, List
- Algunos tipos de variable se pueden transformar en otros, pero no siempre
- Se pueden crear nuevos tipos de variables.

Dibuje un diagrama de flujo para sumar dos números ingresados por el usuario.



Dibuje un diagrama de flujo para encontrar el más grande entre tres números diferentes ingresados por usuario.



Estructuras de control: Condicional

Condicional simple

Si (condición),
(acción):
Sino, no hacer nada

Si (condición), (acción):

Cuando no hay acción asociada al no cumplimiento de la condición, solo se escribe la parte del Si...(se omite la parte del Sino).

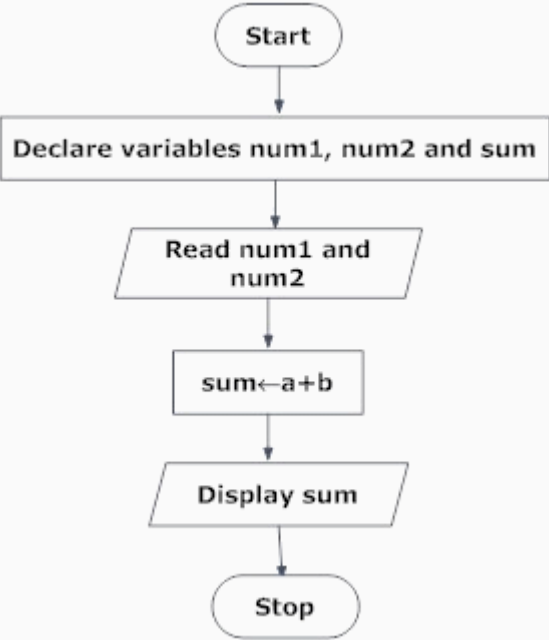
Condicional estándar

Si (condición), (acción):
Sino, (acción)

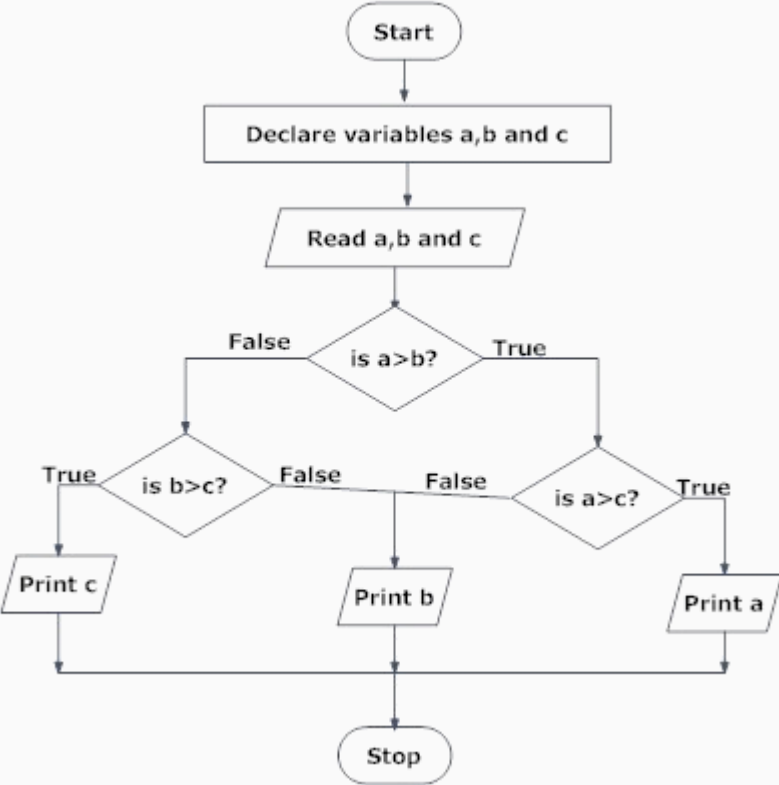
Condicional anidado

Si (condición1), (acción).
Sino (la acción depende de una segunda condición):
 Si (condición2), (acción).
 Sino (acción).

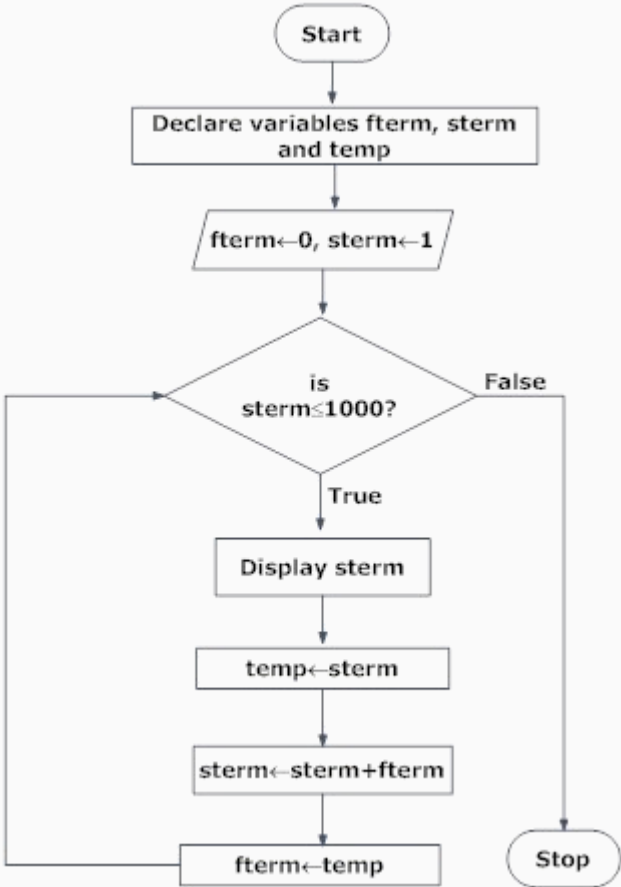
Dibuje un diagrama de flujo para sumar dos números ingresados por el usuario.



Dibuje un diagrama de flujo para encontrar el más grande entre tres números diferentes ingresados por usuario.



Dibuje un diagrama de flujo para encontrar la serie de Fibonacci hasta el término ≤1000.



Estructuras de control: Ciclos

- Existen 2 tipos de ciclos:

Ciclos While

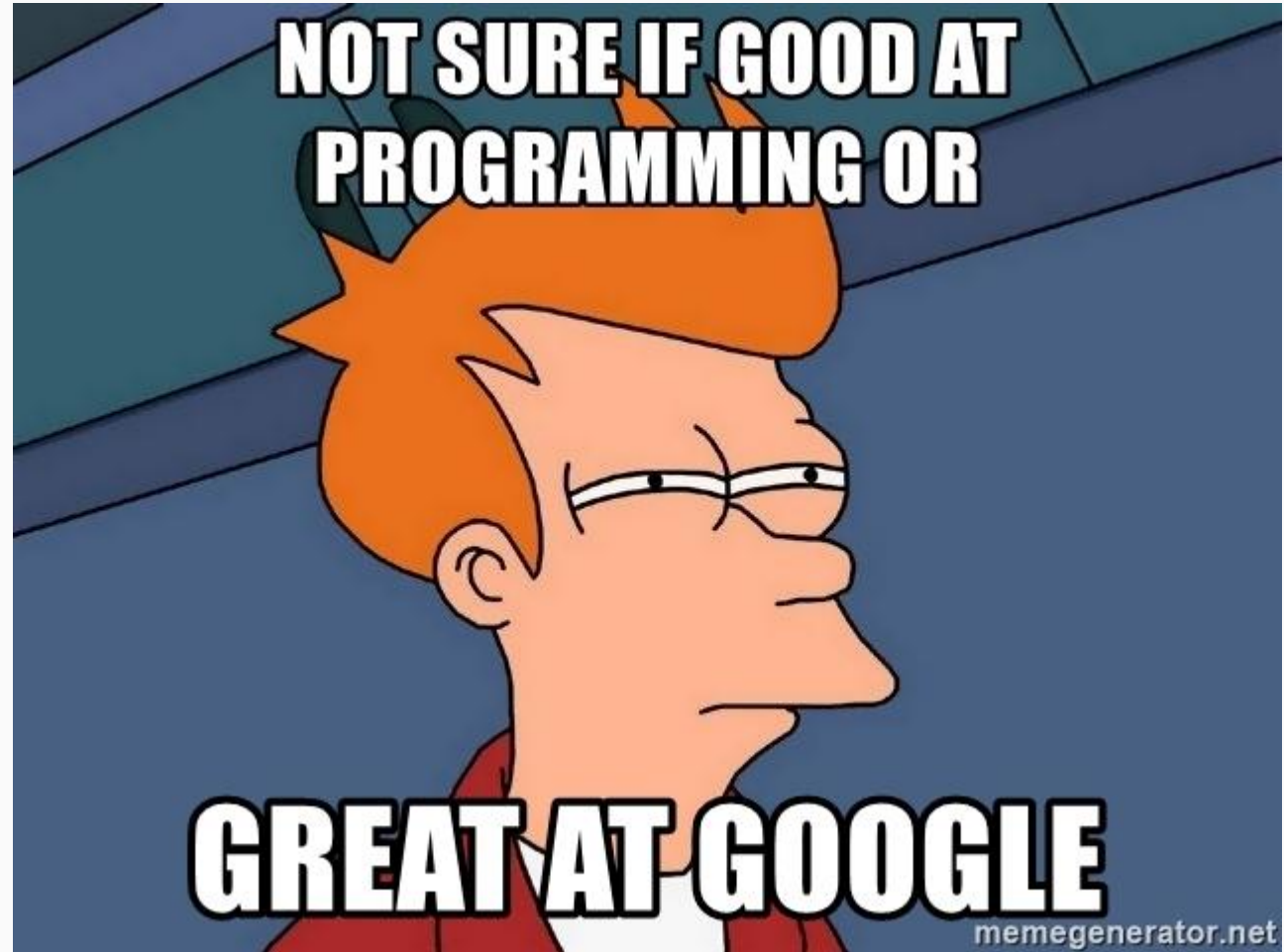
Mientras (condición),
(acción):

Se ejecutan hasta que se
cumpla condición

Ciclos For

Para cada elemento en (lista),
(acción):

Se ejecutan hasta que se
recorra la lista completa



RStudio

- La IDE para R mas utilizada.
- Equipo de RStudio soporta diversas librerías para manejo de datos, como tidyverse, lubridate, ggplot2 o shiny.

Programación para Data Science

Sesión 1: Introducción