Procesos en Paralelo en R

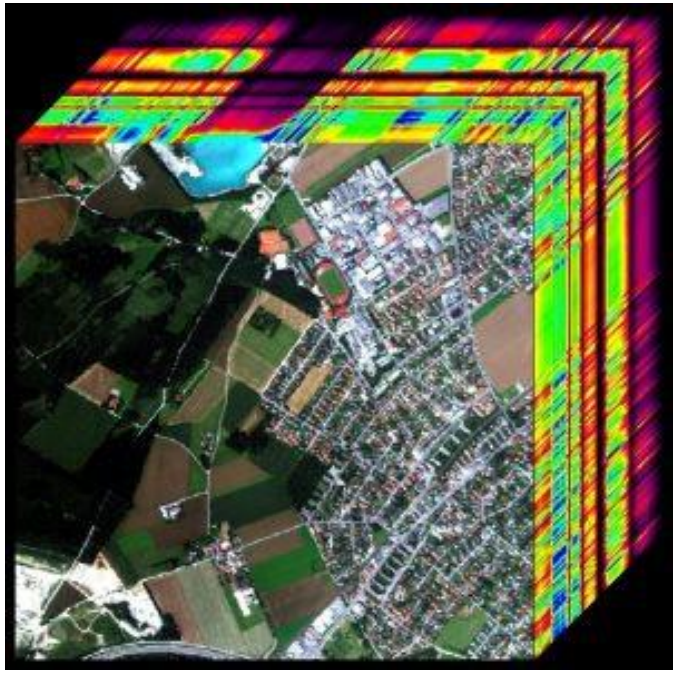
Hugo Andrés Dorado

Científico de datos

hugo.doradob@gmail.com

Introducción

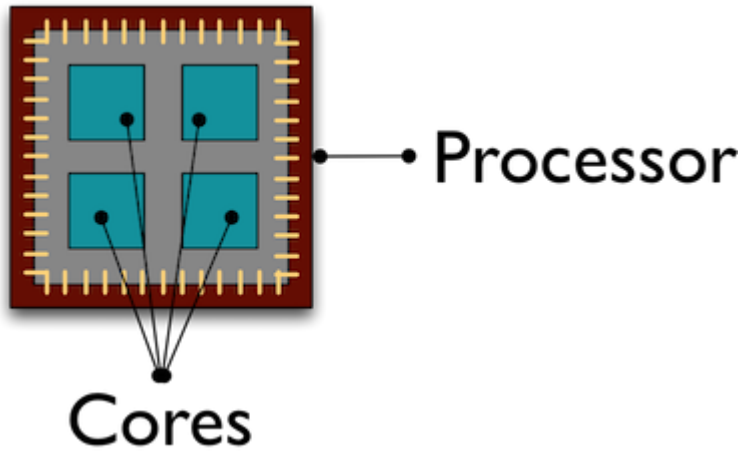
- Procesar grandes cantidades de datos puede tomar mucho tiempo.
- La escala de los datos que hay para procesar es masiva.

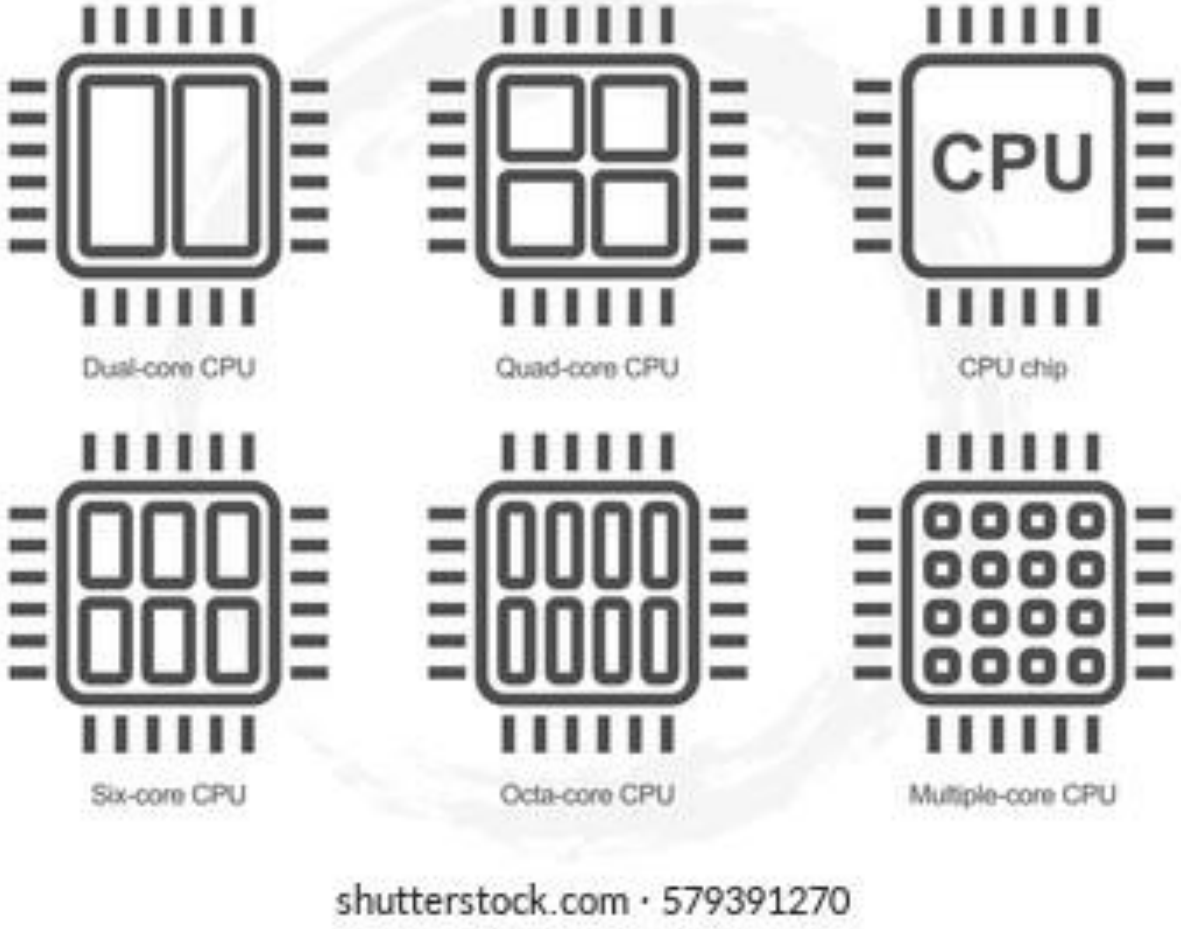


- Una imagen de celular tiene tres bandas primarias (rojo, verde y azul)
- Las imágenes hiperespectrales (HSI) usan miles de bandas específicas

https://www.telematica.com.pe/imagenes-hiperespectrales-ver-para-creer/

Procesadores





Limitaciones en ejecutar en una sola gpu

Ejecutar procesos con un solo procesador R.

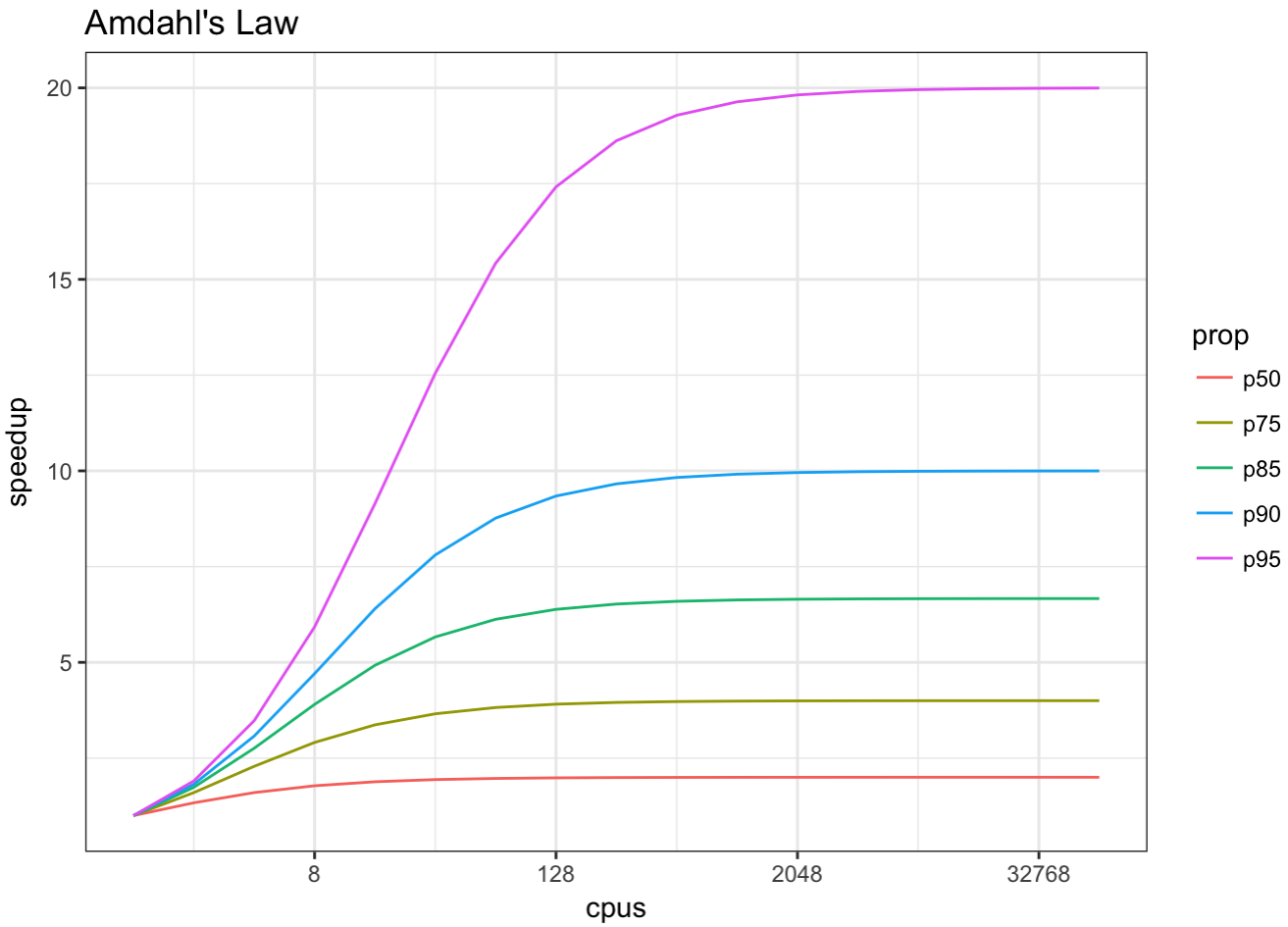
- CPU: toma demasiado tiempo de CPU
- Memoria: toma demasiada memoria
- E / S: toma demasiado tiempo para leer / almacenar desde el disco
- Red: toma demasiado tiempo para transferir.

https://nceas.github.io/oss-lessons/parallel-computing-in-r/parallel-computing-in-r.html

Para paralelizar tener en cuenta que:

- Los datos, funciones y otros inputs deben ser cargados a cada core.
- El sistema operativo debe de crear otros procesos y subprocesos.
- El desempeño es menor que el teórico en cuanto a reducción de tiempo.
- No se puede paralelizar toda una tarea. Dependiendo de la proporción, la aceleración esperada se puede reducir significativamente

Velocidad vs Cpus

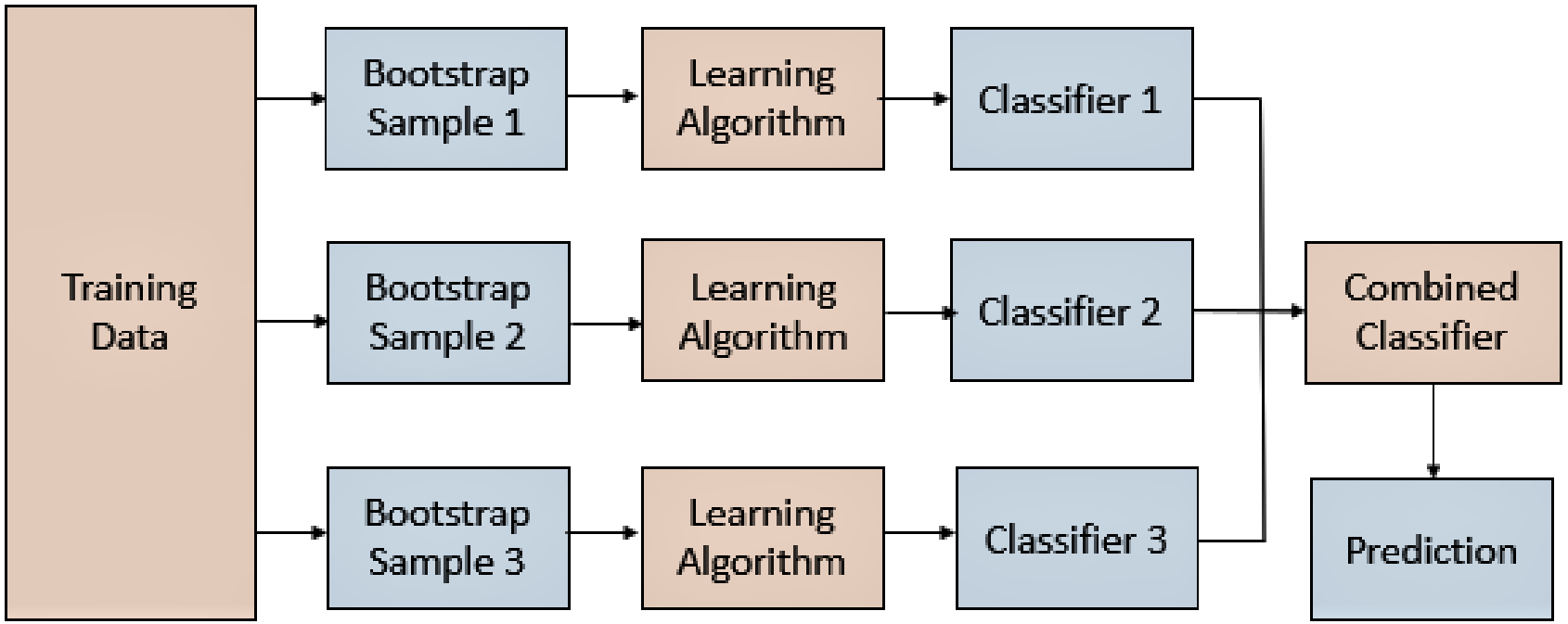


Medir tiempo en R

- User time es el tiempo de la CPU dedicado a la ejecución del las instrucciones del proceso
- *System time* es el tiempo de la CPU empleado por el sistema operativo (el núcleo o *kernel*) siguiendo las instrucciones del proceso (abrir ficheros, iniciar otros procesos o mirar al reloj del sistema, etc.).
- *Elapsed ti*me es el tiempo transcurrido 'real' desde que se inició el proceso.

Si dos tareas de 15 segundos se ejecutan en paralelo en dos núcleos, el *user time* computa la suma de las dos: 30 segundos. Mientras que *elapsed time* no incluye la suma del tiempo de las dos tareas en paralelo, solo el tiempo 'real' que la CPU tardó: 15 segundos.

Ejemplo con boostrap regression



Más información

- https://www.r-bloggers.com/2015/02/how-to-go-parallel-in-r-basics-tips/
- https://psu-psychology.github.io/r-bootcamp-2018/talks/parallel_r.html
- http://www.john-ros.com/Rcourse/parallel.html

Gracias!

Hugo Andrés Dorado.

Científico de datos

hugo.doradob@gmail.com

Conocimiento generado a partir de proyectos de:

Alianza

Bioversity International

CIAT
Centro Internacional de Agricultura Tropical
Desde 1967 Ciencia para cultivar el cambio

CLIMATE CHANGE AGRICULTURE AND FOOD SECURITY