Introducción a la Bioinformática

Conceptos elementales de Computación

Algoritmos

Fernán Agüero

fernan@iib.unsam.edu.ar

Instituto de Investigaciones Biotecnológicas

UNSAM

Ciencias de la Computación

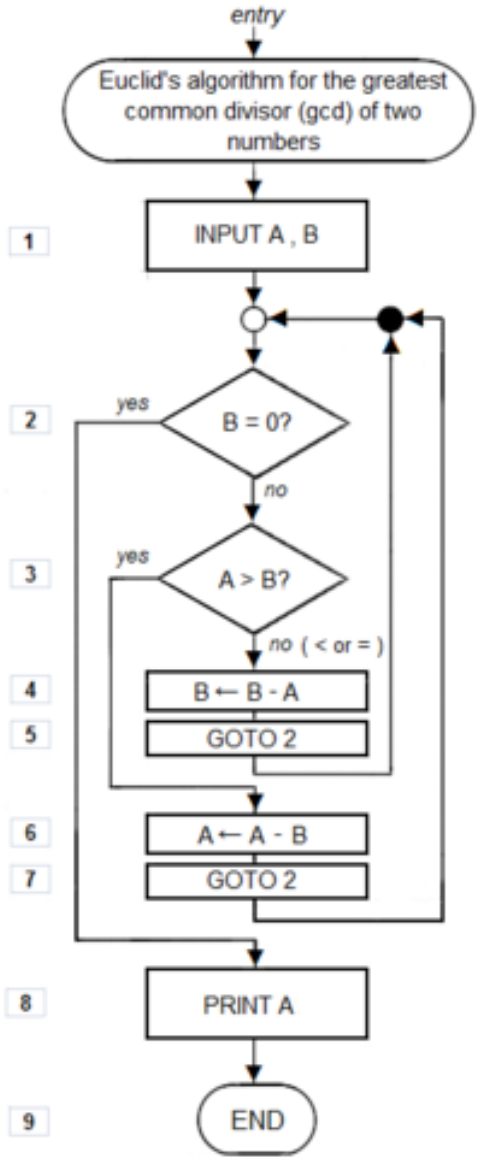
- Es la disciplina que estudia como resolver problemas con computadoras
- Deriva de las Matemáticas
  - Resolución de problemas
  - Algoritmos

- Qué es un Algoritmo?
  - **Informalmente:** *"una serie de reglas que definen en forma precisa una secuencia de operaciones"*
  - **Formalmente:** *"un método, expresado como una lista finita de instrucciones bien definidas para calcular una función"*

Qué es un algoritmo?

Comenzando en un **estado o *"input"*** inicial
Las instrucciones describen **cómputos**
Que al ser ejecutados procederán por una serie de **estados bien definidos**
Eventualmente produciendo un ***"output"***
Terminando en un estado final

http://en.wikipedia.org/wiki/Algorithm



Qué es un algoritmo? Cómo se especifica e implementa?

Un algoritmo puede ser especificado

- En inglés, español, francés, etc.
- En un lenguaje formal: matemático, o de programación
  - C++, Rust, Java, Perl, Python
- En forma de un diseño de hardware

A qué se parece?

- A un protocolo
- A una receta de cocina

Algoritmo ChocoCookies

Qué es un algoritmo? Cómo se especifica e implementa?

Especificación vs Implementación:
Un programa ejecutable es una implementación del algoritmo!
Distintas implementaciones del mismo algoritmo pueden ser mejores o peores.

Más fáciles o intuitivas de usar
Más rápidas
Con más funcionalidades adicionales

Ejemplo:
BLAST, algoritmo para realizar búsquedas de similitud entre secuencias biológicas.

Hay varias implementaciones:

- **NCBI (C, obsolete)**
- **WU-BLAST, Washington University (obsolete)**
- **NCBI (C++)**
- **Blast-rs (Rust)**

Problemas computacionales

- Vamos a ver varios ejemplos a lo largo del curso!
- **Tienen que cumplir con las siguientes condiciones**
  - **Tiene que estar bien definido**
  - **Tiene que tener una solución**

- **"Un problema es una colección *infinita* de instancias, junto con una solución para cada una de esas instancias"**

Ejemplo: ordenar números en forma creciente

- **Input:** una secuencia de **n** números (a1, a2, …, a**n**)
- **Output:** una permutación de la secuencia original (a'1, a'2, …, a'**n**) tal que a'1 > a'2 > … > a'**n**
- **Una instancia del problema:**
  - **Input:** (31, 41, 59, 26, 41, 58)
  - **Output:** (26, 31, 41, 41, 58, 59)

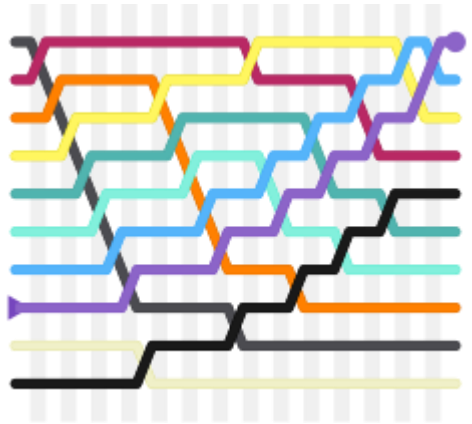
Diversidad de Algoritmos

Hay muchas maneras de resolver el mismo problema computacional.

Ejemplo: Ordenar números

Hay muchos algoritmos para ordenar:

- **Insertion sort - https://en.wikipedia.org/wiki/Insertion_sort**
- **Merge sort - https://en.wikipedia.org/wiki/Merge_sort**
- **Selection sort - https://simple.wikipedia.org/wiki/Selection_sort**
- **Bubble sort - https://en.wikipedia.org/wiki/Bubble_sort**
- **…**



- Cuál es el mejor? **Hay que analizar el algoritmo!**

Corrección

Un algoritmo es *"correcto"* si para cada instancia de un input, termina con el output correcto (la solución exacta).

Hay algoritmos incorrectos!

- **Aproximados**
- **Heurísticos**

Eficiencia

Una medida posible de eficiencia es la *"velocidad"*: cuánto demora un algoritmo en llegar al output/solución.

Pero la *"velocidad"* es engañosa (depende del hardware).

Tal vez la mejor definición sea:

"Cuántos recursos utiliza el algoritmo para realizar su tarea"

Recursos = tiempo, espacio de almacenamiento, cantidad de memoria.

Problema genérico

- Dado un conjunto de números, existe algún subconjunto cuya suma sea N?
- Ejemplo:
  - Conjunto: 1 2 4 5 8 9 10 11 23 76 89
  - N = 119

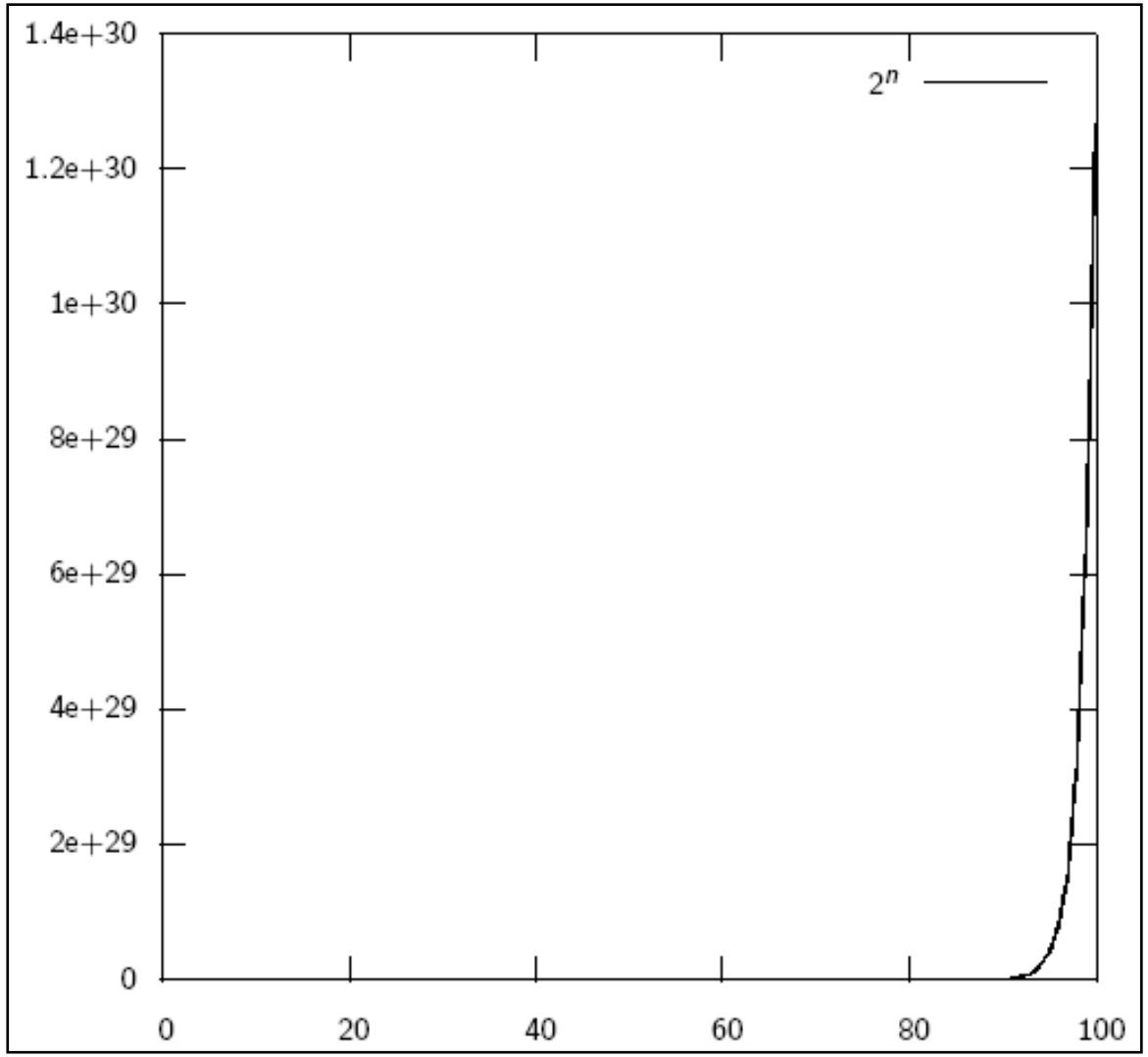
Está bien definido?
Tiene solución?

Problema ejemplo (continuacion)

- Evaluar
  - **Viabilidad**
    - Existen impedimentos teóricos?
  - **Solución (algoritmo)**
    - Tomar subconjuntos de S y evaluar la suma
    - Responder SI/NO
  - **Análisis del algoritmo**
    - Eficiencia: cómo escala con el input
    - Problema = hay 2^S subconjuntos de S!
      - Si cada subconjunto se verifica en 0.0001 segundo
      - Para 100 elementos tardaríamos
      - 2^{100} * 0.0001s / 60 * 60 * 24 * 360 * 100 = 6.34 10^{86} siglos

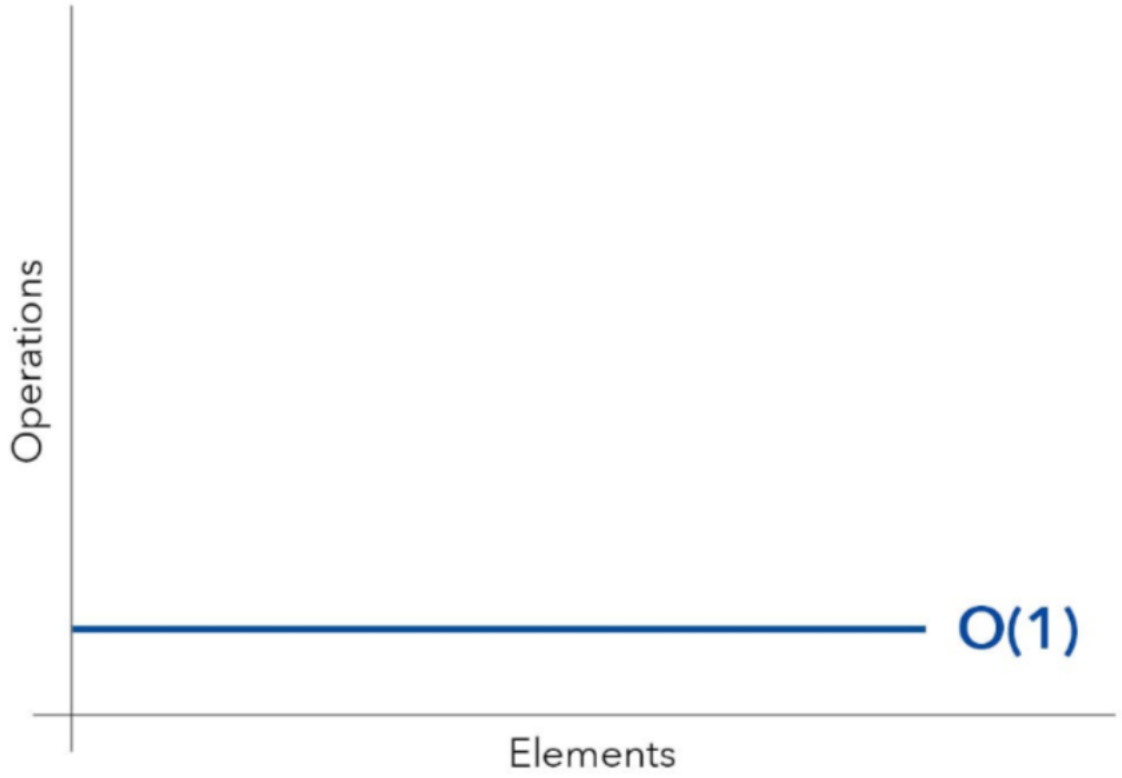
Complejidad

- Métrica para comparar algoritmos
- Relación entre cantidad de datos de entrada (input) y tiempo/espacio necesario para resolver el problema
- Normalmente se especifica para
  - θ(n), mejor caso
  - Ω(n), caso promedio
  - O(n), peor caso



Complejidad: Big O Notation

- Constante
  - O(1)
- Lineal
  - O(n)
- Logarithmic
  - O(log n)
- Polinomial
  - $O(n^2)$
- Exponencial
  - $O(2^n)$



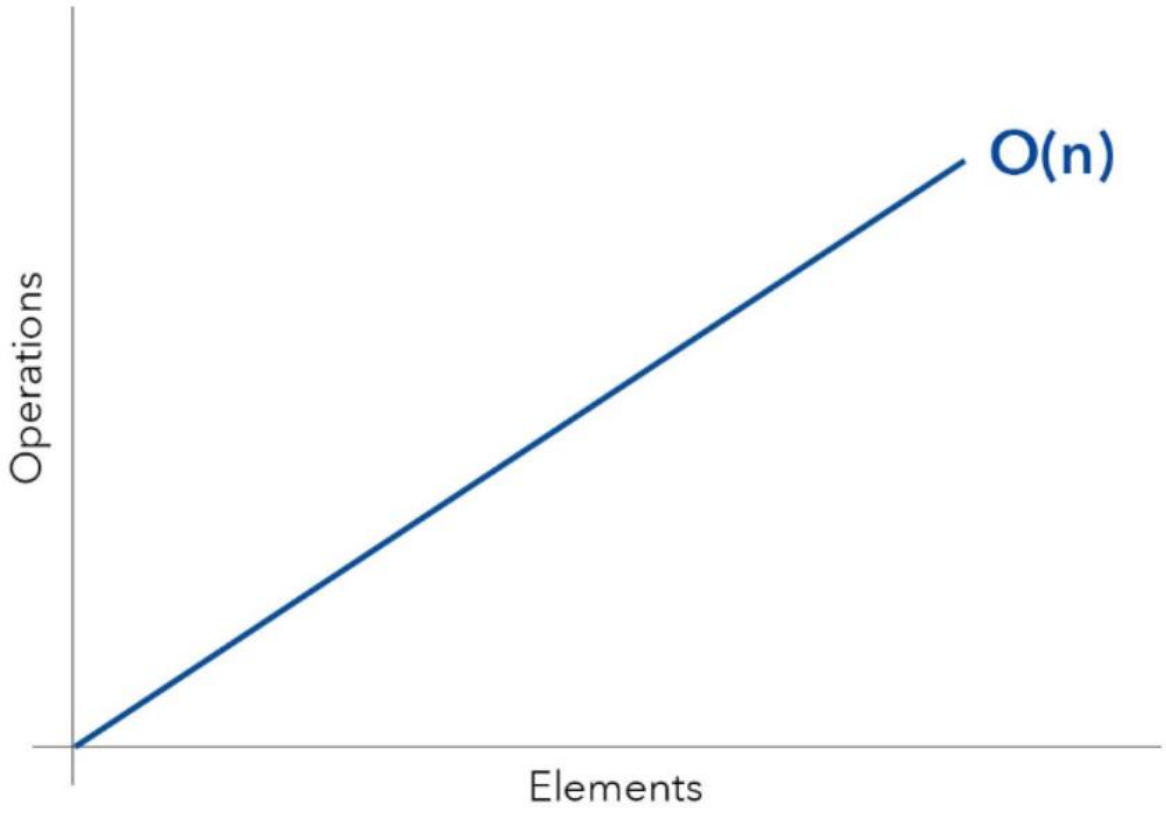
Big O Notation = Cota superior asintótica

Notación matemática que describe el tamaño aproximado de una función sobre un dominio.

https://en.wikipedia.org/wiki/Big_O_notation

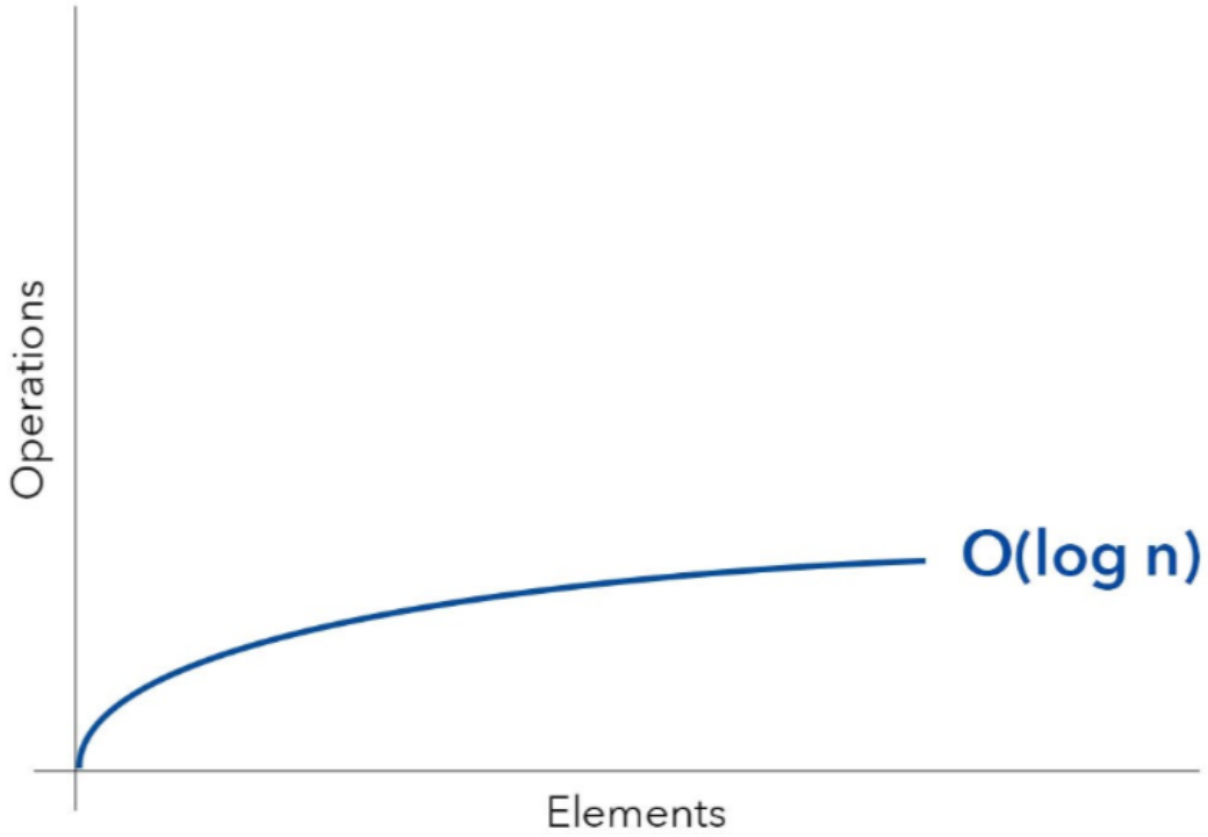
Complejidad: Big O Notation

- Constante
  - O(1)
- Lineal
  - O(n)
- Logarithmic
  - O(log n)
- Polinomial
  - $O(n^2)$
- Exponencial
  - $O(2^n)$



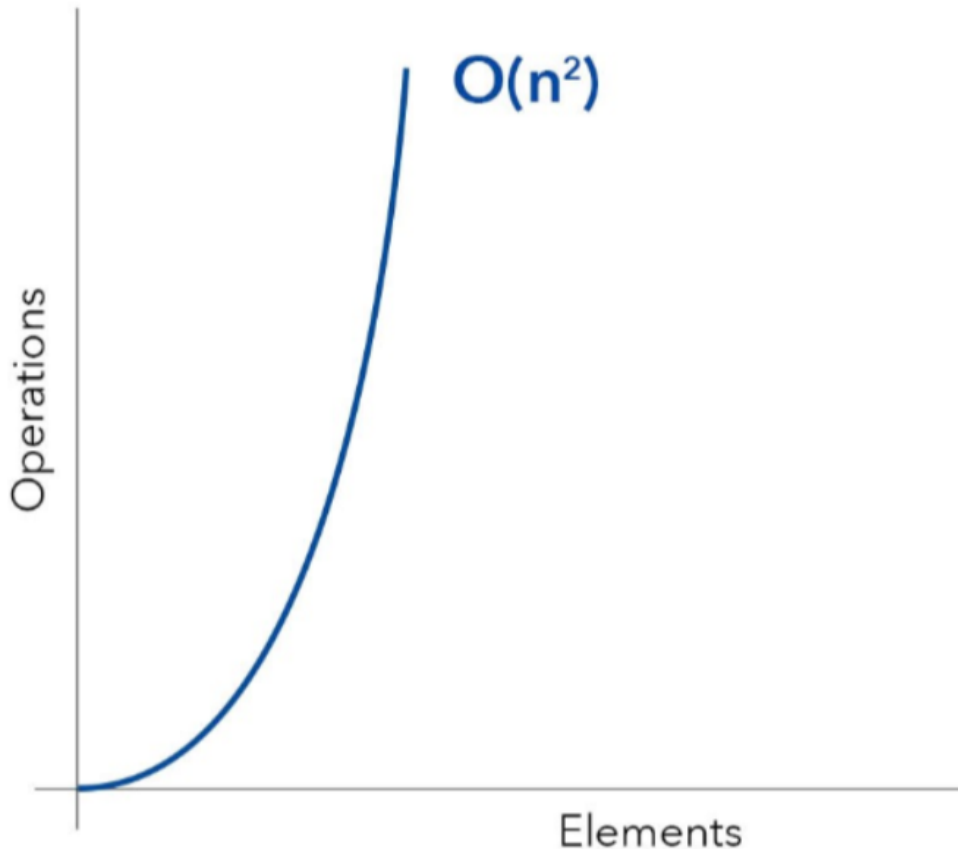
Complejidad : Big O Notation

- Constante
  - O(1)
- Lineal
  - O(n)
- Logarithmic
  - O(log n)
- Polinomial
  - $O(n^2)$
- Exponencial
  - $O(2^n)$



Complejidad : Big O Notation

- Constante
  - O(1)
- Lineal
  - O(n)
- Logarithmic
  - O(log n)
- Polinomial
  - $O(n^2)$
- Exponencial
  - $O(2^n)$



“quadratic time”

Complejidad: Big O Notation

Mas info online.

Ejemplos

Ver:

https://www.educative.io/blog/a-big-o-primer-for-beginning-devs

Big O Notation examples

O(1)

```
 void printFirstItem(const vector<int>& items)
{
    cout << items[0] << endl;
}
```

This function runs in O(1) time (or “constant time”) relative to its input. This means that the input array could be 1 item or 1,000 items, but this function would still just require one “step.”

O(n)

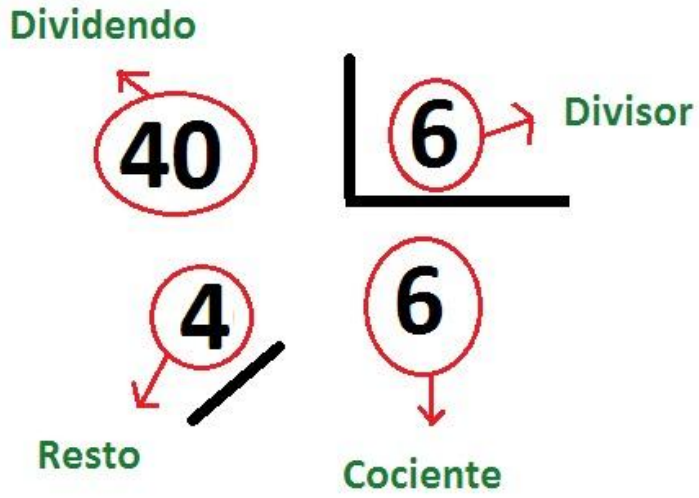
```
void printAllItems(const vector<int>& items)
{
    for (int item : items) {
        cout << item << endl;
    }
}
```

This function runs in O(n) time (or “linear time”), where n is the number of items in the vector. If the vector has 10 items, we have to print 10 times. If it has 1,000 items, we have to print 1,000 times.

O(n^2)

```
void printAllPossibleOrderedPairs(const vector<int>& items)
{
    for (int firstItem : items) {
        for (int secondItem : items) {
            cout << firstItem << ", " << secondItem << endl;
        }
    }
}
```

Algoritmo de la division entre dos números



Algoritmo = reglas o pasos (mecánica)

Es exacto este algoritmo?

https://en.wikipedia.org/wiki/Division_algorithm

Algoritmos exactos, aproximados, heurísticas

- Existen varias maneras de resolver un problema
- La elección final puede depender del tiempo que estamos dispuestos a esperar para obtener una respuesta
- Algoritmo de la división
  - **10 / 5 = 2**
  - **133 / 4 = 33,25**
  - **10 / 3 = 3,33333…**
- Opciones del algoritmo
  - **Termina cuando no hay más resto**
  - **Termina cuando llega a colocar la coma**
  - **Termina cuando llega al n-ésimo dígito.**

Algoritmos

- Algoritmo exacto
  - Da la solución exacta (óptima)
- Algoritmo aproximado
  - Da una solución que se encuentra a una distancia calculable (el error) de la solución óptima.
- Heurística
  - Da una solución aceptable sin error calculable.

Heurísticas

- Abundan en biología
  - Soluciones ‘casi’ óptimas
  - Costo computacional razonable
  - No garantiza factibilidad del resultado ni da idea de a qué distancia se encuentra uno del resultado óptimo
  - Pero se pueden hacer análisis estadísticos sobre los datos

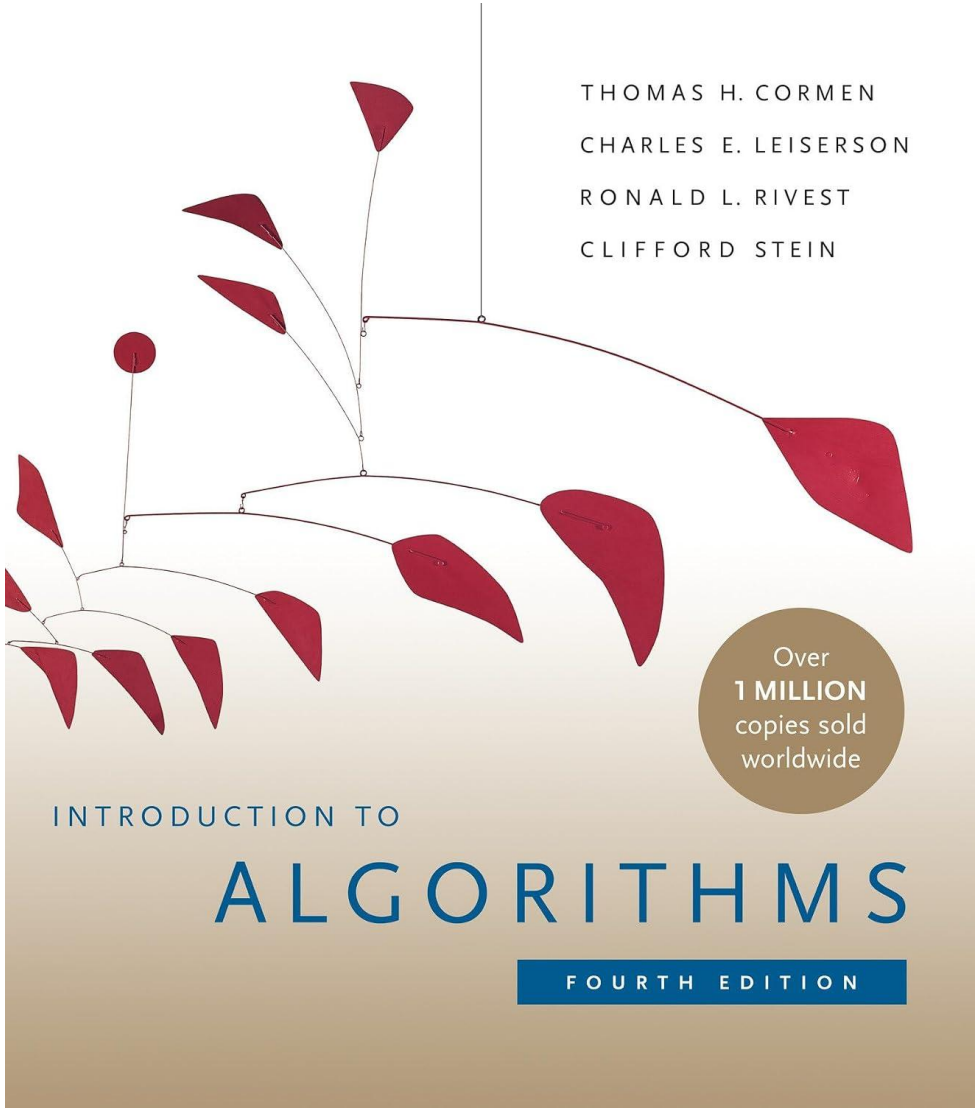
Aprendizaje automático

- Inteligencia artificial, Machine learning
  - **Se entrenan programas sobre un set de datos "de entrenamiento"**
  - **El programa entrenado se enfrenta con un set de datos desconocido**
    - Hidden Markov Models
    - Neural networks

- Ejemplos en biología:
  - **Reconocimiento de patrones**
    - SignalP, NetOGlyc, DGPI
    - Identificación de genes
    - Búsquedas de similitud de alta sensibilidad

Bibliografía opcional

Introduction to Algorithms (2022). Cormen, Leiserson, Rivest, Stein. 4th Edition, MIT Press.



INTERVALO

Fin del Slot 1