

# Algoritmos para Big Data

García Prado, Sergio  
sergio@garciparedes.me

16 de marzo de 2017

## Resumen

*En este documento se expone una breve descripción acerca de las distintas disciplinas de estudio relacionadas con el ámbito del tratamiento de grandes cantidades de información (Big Data) desde una perspectiva algorítmica.*

## 1. INTRODUCCIÓN

El procesamiento de grandes cantidades de información presenta un gran reto a nivel computacional debido al elevado coste originado por el elevado tamaño del conjunto de datos. Para solventar dicha problemática se prefieren algoritmos cuya principal característica es un orden de complejidad sublineal ( $o(N)$ ), tanto en tiempo como en espacio. Dichas técnicas se llevan a cabo sobre paradigmas de computación paralela, que permiten aprovechar en mayor medida las restricciones actuales a nivel de hardware.

## 2. ALGORITMOS PARA STREAMING

Los algoritmos para streaming se caracterizan por procesar los datos de una forma secuencial dependiente del orden de llegada. La ventaja que estos presentan respecto de otras alternativas en tiempo real es que utilizan propiedades estadísticas para reducir su coste, lo que añade una cierta tasa de error. El descubrimiento de métodos altamente eficiente para estimar *Momentos de Frecuencia* ha sido un gran hito en esta categoría algorítmica.

## 3. ESTRUCTURAS DE DATOS DE RESUMEN

[TODO ]

### 3.1. SAMPLING

[TODO ]

### 3.2. HISTOGRAM

[TODO ]

### 3.3. WAVELETS

[TODO ]

### 3.4. SKETCH

[TODO ][Count-Sketch, CountMin-Sketch, AMS Sketch...]

## 4. ALGORITMOS PARA GRAFOS

[TODO ]

## 5. TÉCNICAS DE MINERÍA DE DATOS

[TODO ]

### 5.1. CLASIFICACIÓN VS REGRESIÓN

[TODO ]

### 5.2. SUPERVISADAS VS NO SUPERVISADAS

[TODO ]

## 6. PARALELIZACIÓN A GRAN ESCALA

[TODO ]

### 6.1. MODELO MAPREDUCE

[TODO ]