

# DATOS MASIVOS I

## CONCEPTOS BÁSICOS

---

Gibran Fuentes-Pineda  
Febrero 2021

# MUCHOS MÁS DATOS Y A MAYOR VELOCIDAD

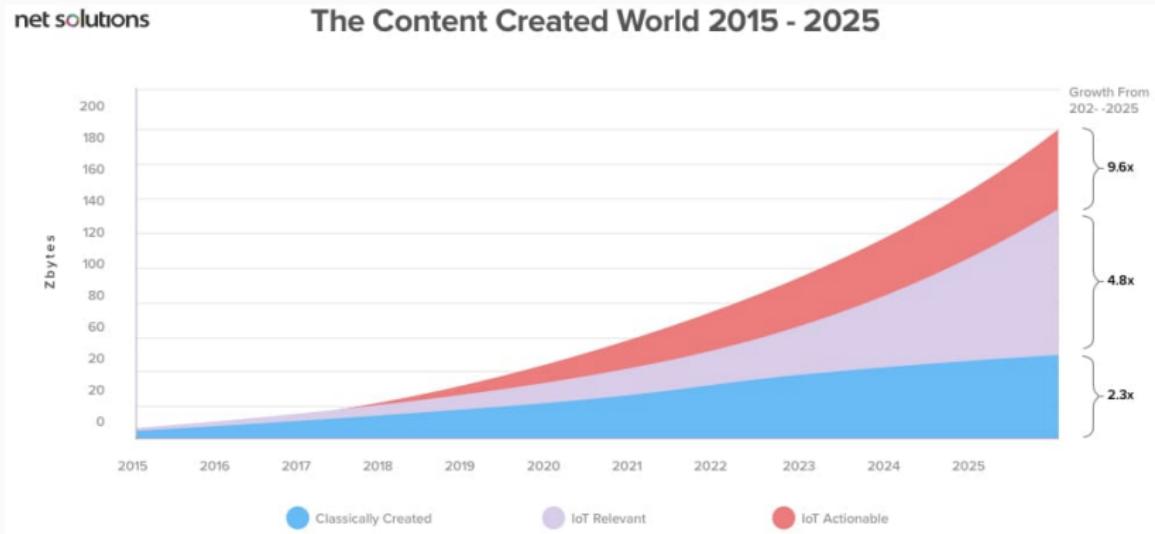


Imagen tomada de <http://www.tech-dynamics.com/wp-content/uploads/2014/02/BigDataChart.png>

# MÁS PARÁMETROS Y MÁS DATOS NO ESTRUCTURADOS

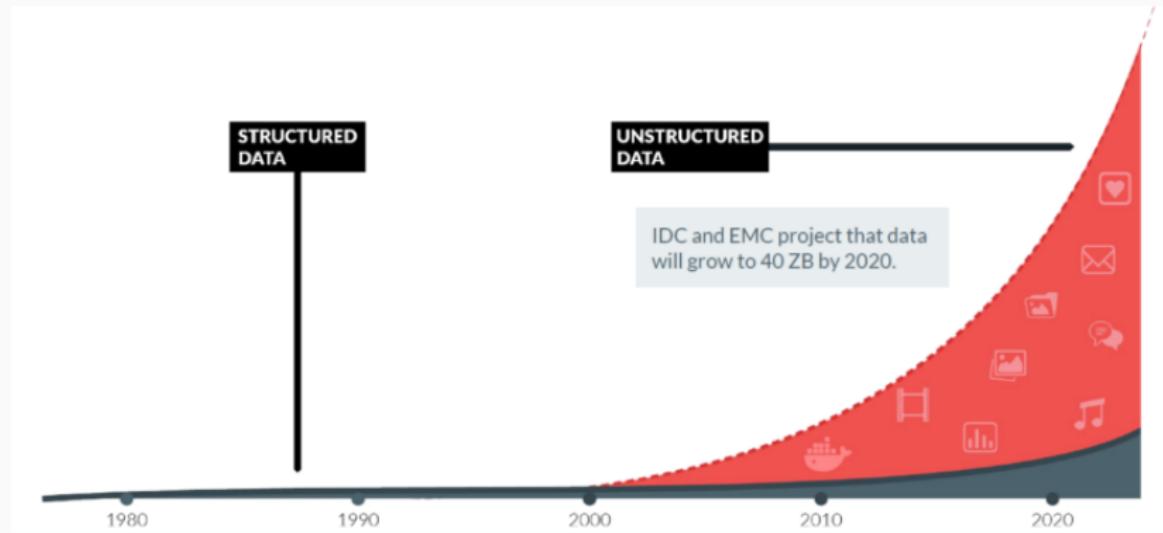
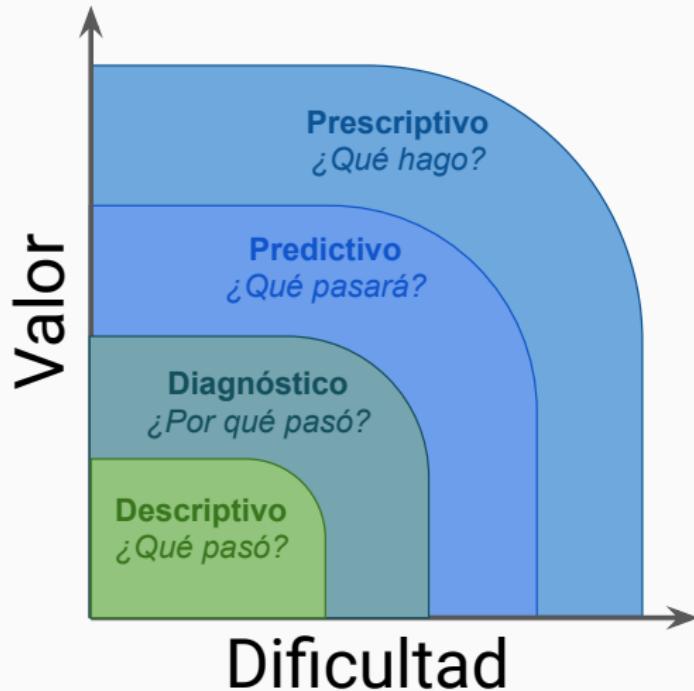


Imagen tomada de <https://www.datanami.com/2017/02/01/solving-storage-just-beginning-minio-ceo-periasamy/>

# ANÁLISIS MENOS DESCRIPTIVO, MÁS PREDICTIVO/PRESCRIPTIVO



## MÁS APLICACIONES (1)

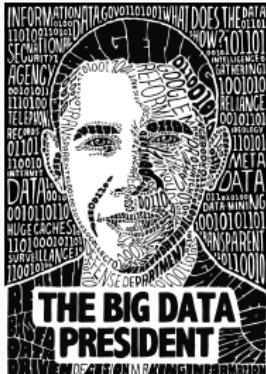


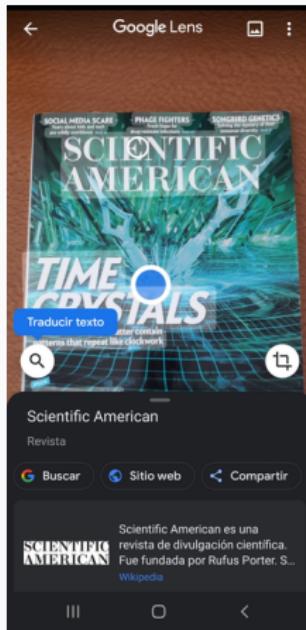
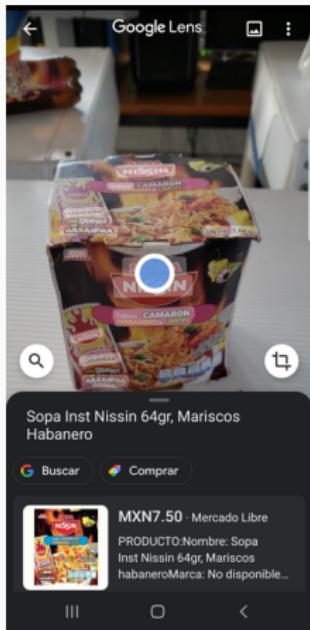
Imagen tomada de The Washington Post

## MÁS APLICACIONES (2)



Imagen tomada de Snavely et al. Photo Tourism: Exploring Photo Collections in 3D, *ACM Transactions on Graphics*, 2006.

## MÁS APLICACIONES (3)



# MEJORES MODELOS (1)

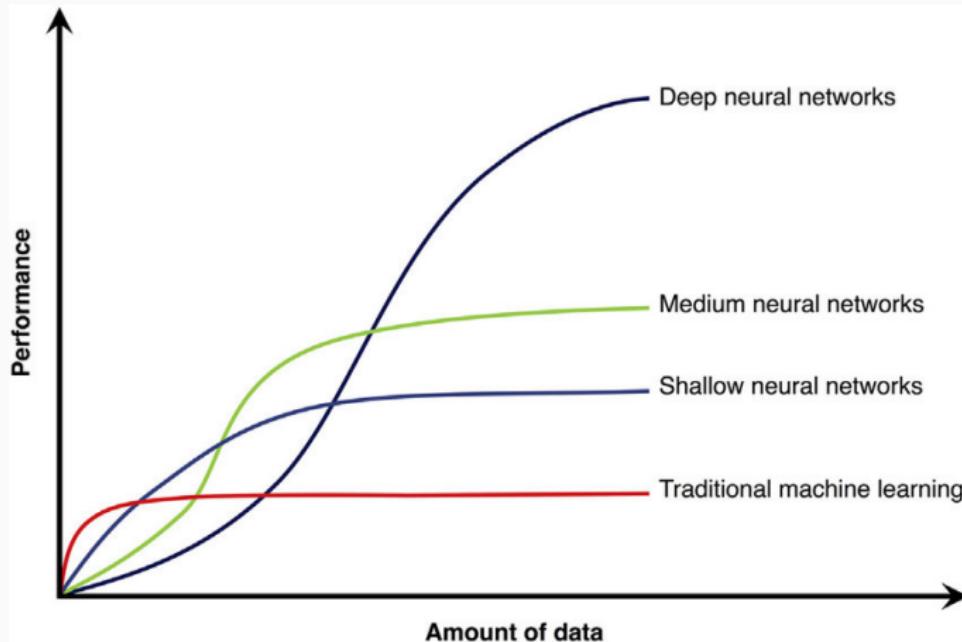
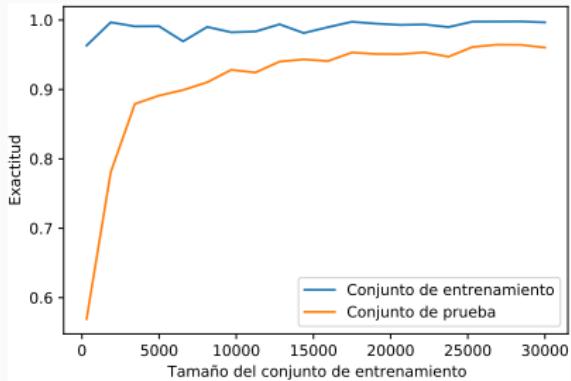
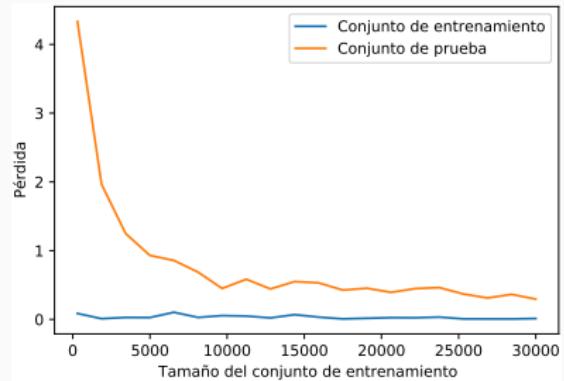
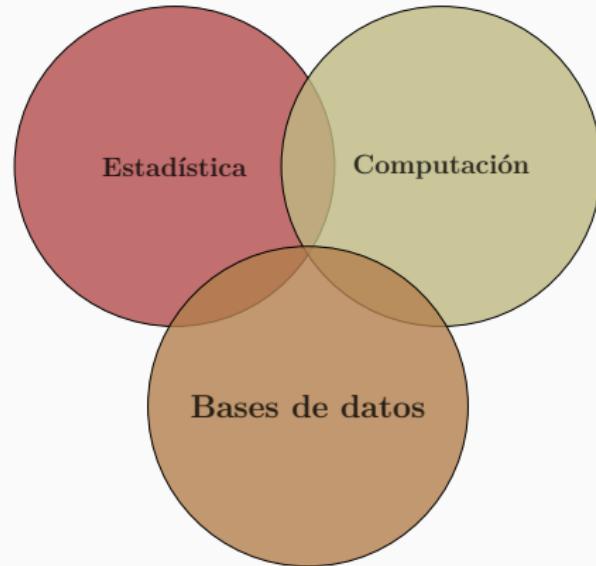


Imagen tomada de Tang et al., Canadian Association of Radiologists Journal 69(2), 2018

# MEJORES MODELOS (2)



# DEFINICIÓN



# CARACTERÍSTICAS: 3Vs



# CARACTERÍSTICAS: 4Vs

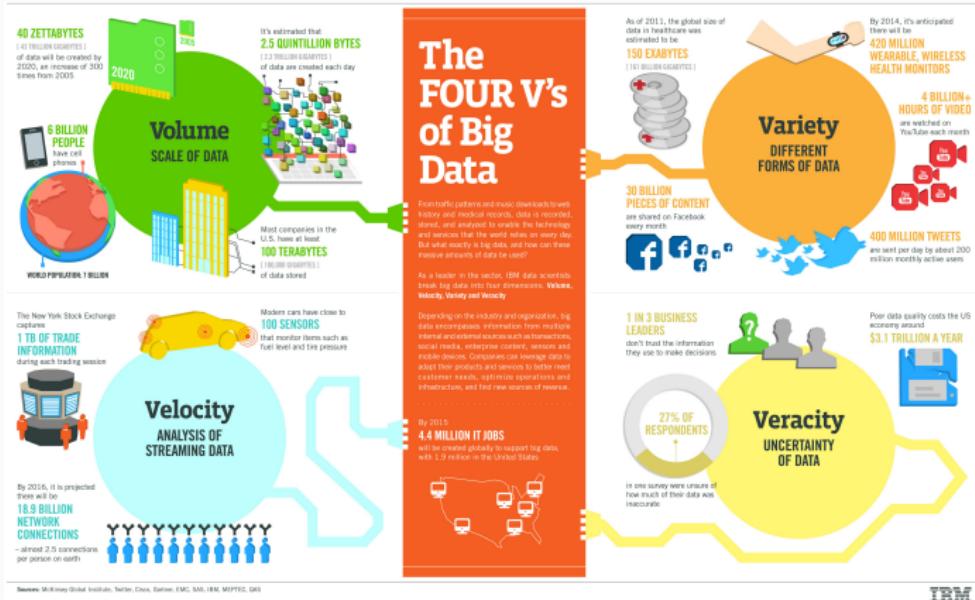


Imagen tomada de <https://www.ibmbigdatahub.com/infographic/four-vs-big-data>

# QUIZÁS MUCHAS Vs



- Concepciones: cuestión simplemente implementacional, lo mismo pero con más datos
- Muchos retos
  - Algoritmos, estructuras de datos y modelos escalables
  - Representaciones eficientes
  - Dimensionalidad, diversidad y variabilidad extremas
  - Respuesta en línea o tiempo real
  - Paralelización
  - Modelado
  - Búsqueda, organización y exploración

# ENCONTRANDO CORRELACIONES

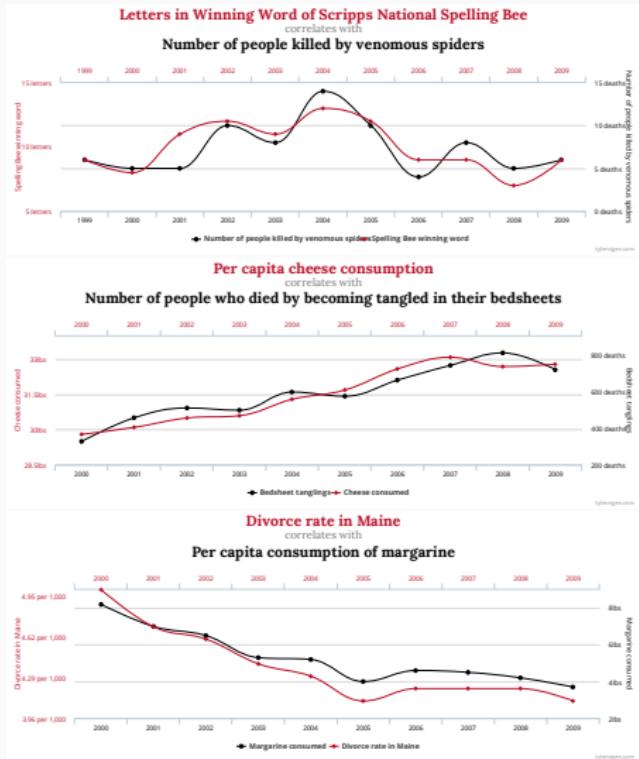


Figura de Tyler Vigen, tomada de <https://www.tylervigen.com/spurious-correlations>

# EL PRINCIPIO DE BONFERRONI

---

*Calcula el número esperado de ocurrencias de un evento bajo la suposición que es aleatorio. Si el número es mucho mayor al de las ocurrencias reales, entonces las conclusiones que puedas sacar a partir de estos eventos probablemente sean falsas.*

## PRINCIPIO DE BONFERRONI: DETECTANDO TERRORISTAS (1)

---

- Hay 1000 millones de personas
- Cada persona va a un hotel una vez cada 100 días
- Un hotel hospeda 100 personas y hay 100,000 hoteles (capaces de hospedar al 1% del total de personas)
- Para detectar un terrorista buscamos pares de personas que en 2 días distintos en una ventana de 1000 días fueron al mismo hotel

## PRINCIPIO DE BONFERRONI: DETECTANDO TERRORISTAS (2)

- La probabilidad de que 2 personas decidan ir a un hotel cualquiera de los 100 días es  $0.01 \times 0.01 = 0.0001$
- La probabilidad de que además elijan el mismo hotel es

$$\frac{0.0001}{10^5} = 10^{-9}$$

- La probabilidad de que 2 personas visiten el mismo hotel en 2 días distintos es  $10^{-9} \times 10^{-9} = 10^{-18}$

## PRINCIPIO DE BONFERRONI: DETECTANDO TERRORISTAS (3)

- El número total de posibles pares de personas es

$$\binom{10^9}{2} \approx 5 \times 10^{17}$$

- El número de pares de días es

$$\binom{1000}{2} \approx 5 \times 10^5$$

- Por lo tanto, el número esperado de personas que visitan el mismo hotel en 2 días distintos es

$$(5 \times 10^{17}) \times (5 \times 10^5) \times 10^{-18} = 250,000$$

# RIESGOS

---

- Vigilancia
- Sesgos y parcialidad
- Baja exploración (filtro burbuja)