Enfoque Estadístico del Aprendizaje y el Descubrimiento 2021

Sofía Perini – sofiaperini9@gmail.com
Juan Barriola – jmbarriola@gmail.com
Andrés Farall – afarall@hotmail.com

Motivación

- ¿ Para Qué el Enfoque Estadístico ?

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Tendencias en Ciencia de Datos

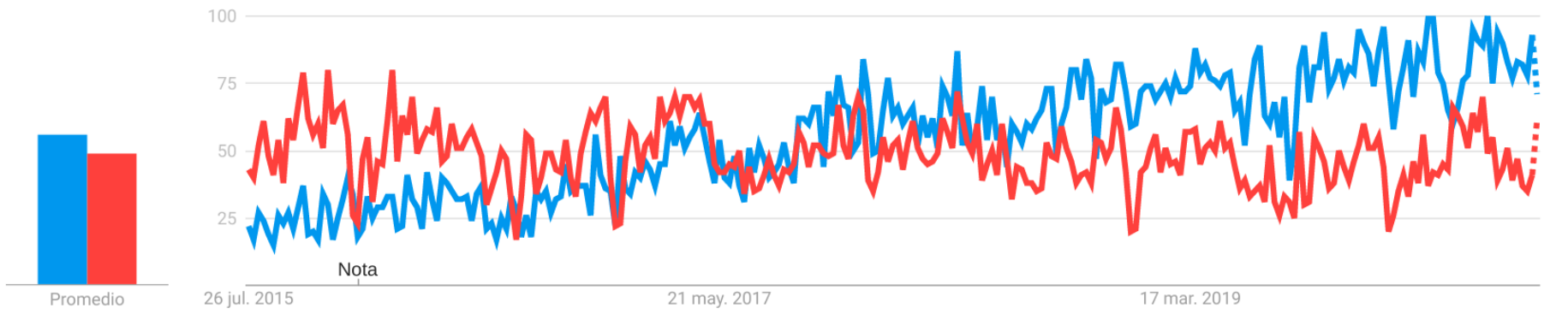
machine learning algorithm
Término de búsqueda

statistical model
Término de búsqueda

+ Agregar comparación

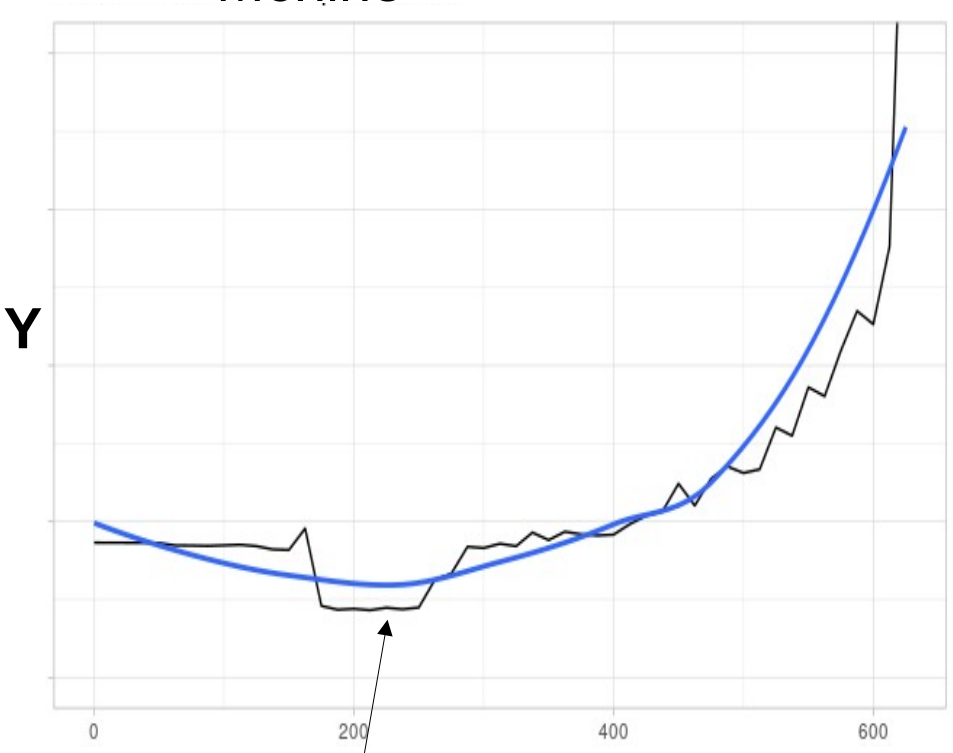
En todo el mundo | En los últimos cinco años | Todas las categorías | Búsqueda web

Interés a lo largo del tiempo



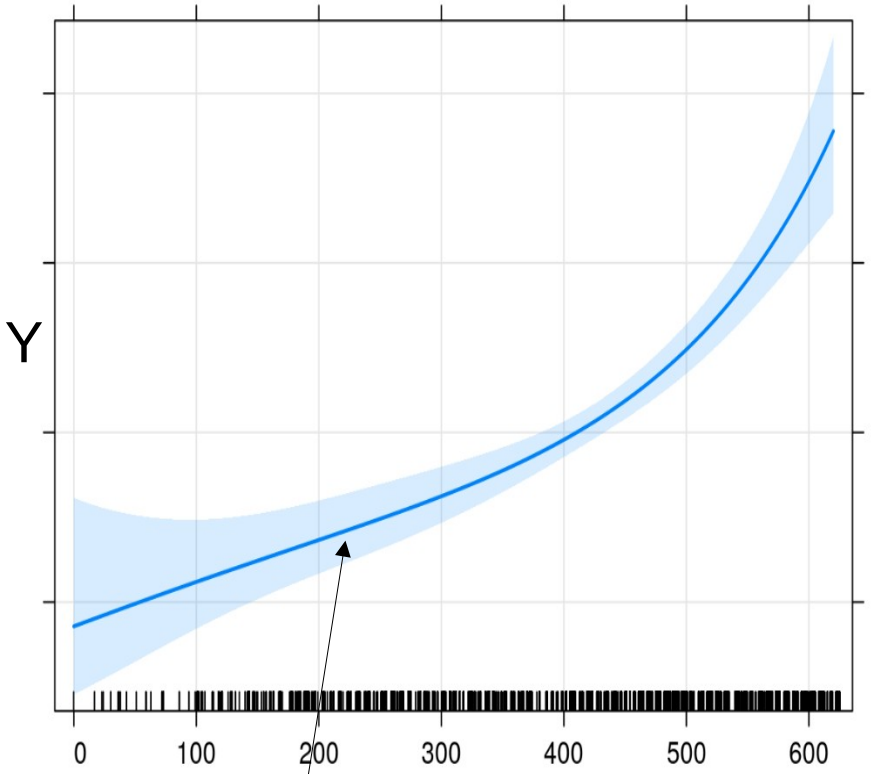
Predicción Versus Explicación

Gradient Boosting Mchine



Y = GBM (X , Z)

Modelo Lineal Generalizado



E(Y) = P(X) + P(Z)

Programa

- EEA20
  - Participantes
  - Insignias
  - General
  - Introducción: Inferencia Estadística.
  - Modelo de Regresión Lineal Simple y Múltiple
  - Estadística Bayesiana
  - Modelo Lineal Generalizado
  - Regularización
  - Modelos Aditivos y No Paramétricos
  - Redes Neuronales Artificiales
  - Supprt Vector Machines
  - Gradient Boosting Machines

- Enfoques de la inferencia estadística
  - Modelado Estadístico
  - Significatividad Estadística (p-valor)
  - Máxima Verosimilitud e Inferencia Bayesiana.
- El problema de predicción.
  - Aprendizaje Supervisado
  - Medidas de Bondad de Ajuste
  - Trade-off sesgo-varianza
  - Sobreajuste

Regresión lineal simple y múltiple.
- Estimación por Cuadrados Mínimos.
- Multicolinealidad.
- Transformaciones.
- Variables dummy. Interacción.
- Métodos de ajuste paso a paso.
- Alternativas Robustas

Modelos Lineales Generalizados (GLM)
Regresión logística.
Regresión de Cuantiles (QR)
Introducción a los Modelos Lineales Mixtos (LMM)
Regresión Lasso y Ridge
Regresión No Paramétrica
- Técnicas de suavizado
- Modelos Aditivos
- Projection Pursuit Regression

Redes Neuronales Artificiales Multicapa (ANN – MLP). Regresión, Clasificación y Reducción de Dimensión (Autoencoders).
Regresión/Clasificación con SVM y Gradient Boosting (Xgboost).
Benchmarking, Comparación y selección de modelos: AIC, BIC, Enfoque Multimodel, Model Tuning(Caret).

Objetivos Principales del Curso

- Ofrecer un enfoque Estadístico de las técnicas de Regresión
- Balancear el compromiso Explicación Vs. Predicción.
- Brindar herramientas aplicadas
- Posicionarse en un contexto científico e interdisciplinario
- Enseñar una amplia variedad de técnicas implementadas en R
- Utilizar conjuntos de datos reales
- No profundizar en la matemática sobre la cual se basan los métodos

“Un Modelo Lineal No Se Le Niega a Nadie”

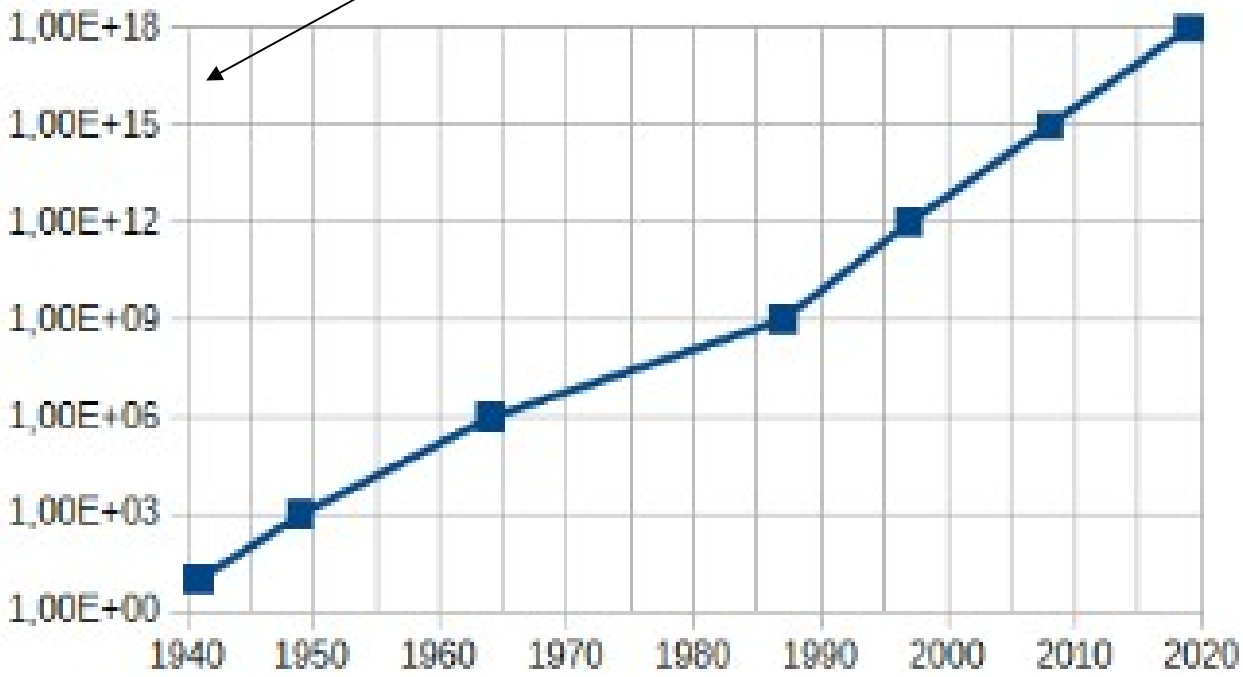
- Avanzamos de lo simple a lo complejo.
- Un modelo simple sirve como “benchmark” contra el que comparar el resultado de modelos más complejos.
- Un modelo simple permite interpretar la mecánica de las relaciones entre variables.

El Contexto Tecnológico

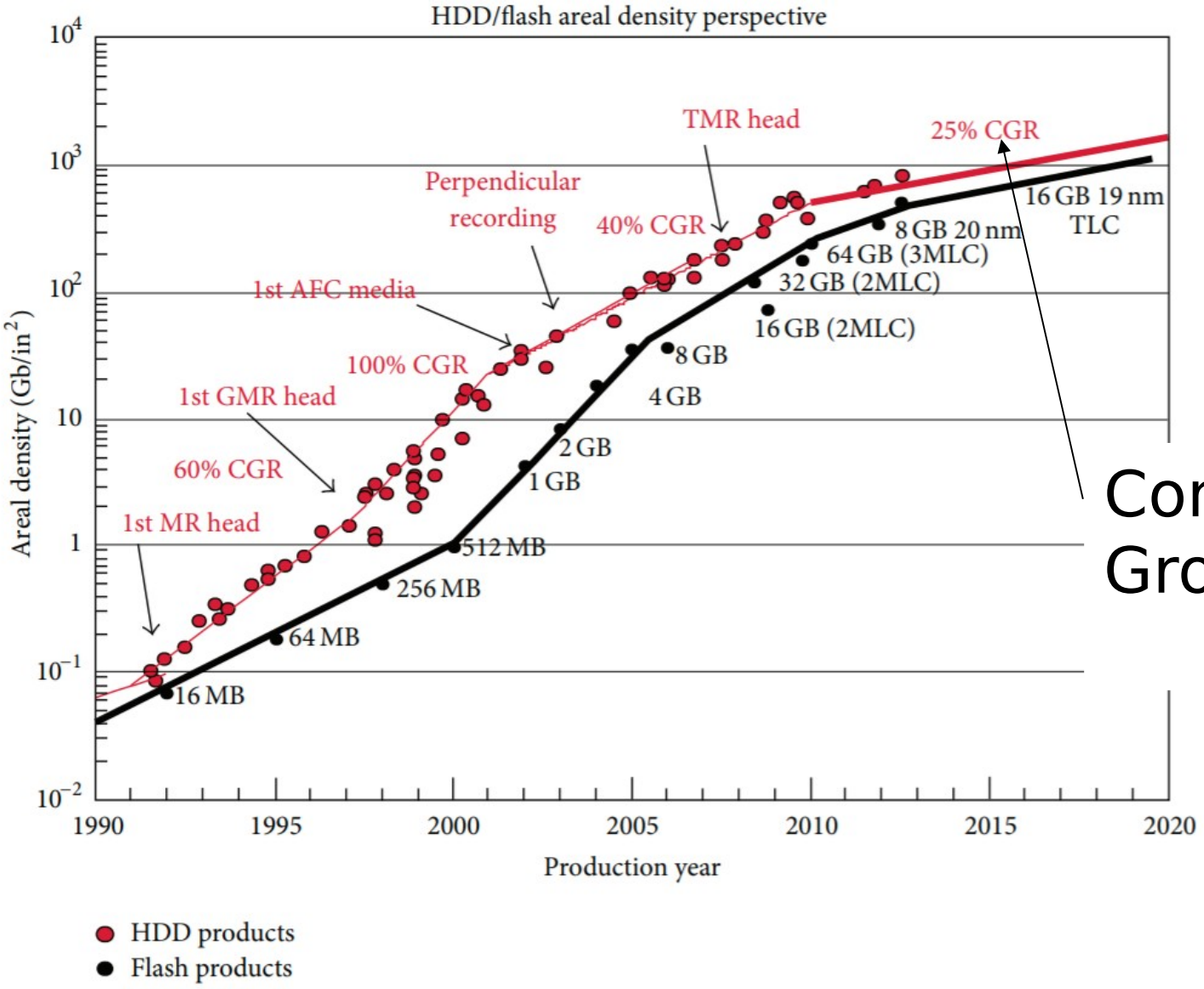
- Capacidad de Cálculo
- Capacidad de Almacenamiento
- Velocidad en la Transmisión de Datos
- Ciencia de Datos
- Machine Learning
- Data Mining
- Big Data
- Optimización

Que pasó con la Capacidad de Cálculo ?

Millones de Instrucciones en Punto Flotante por Segundo

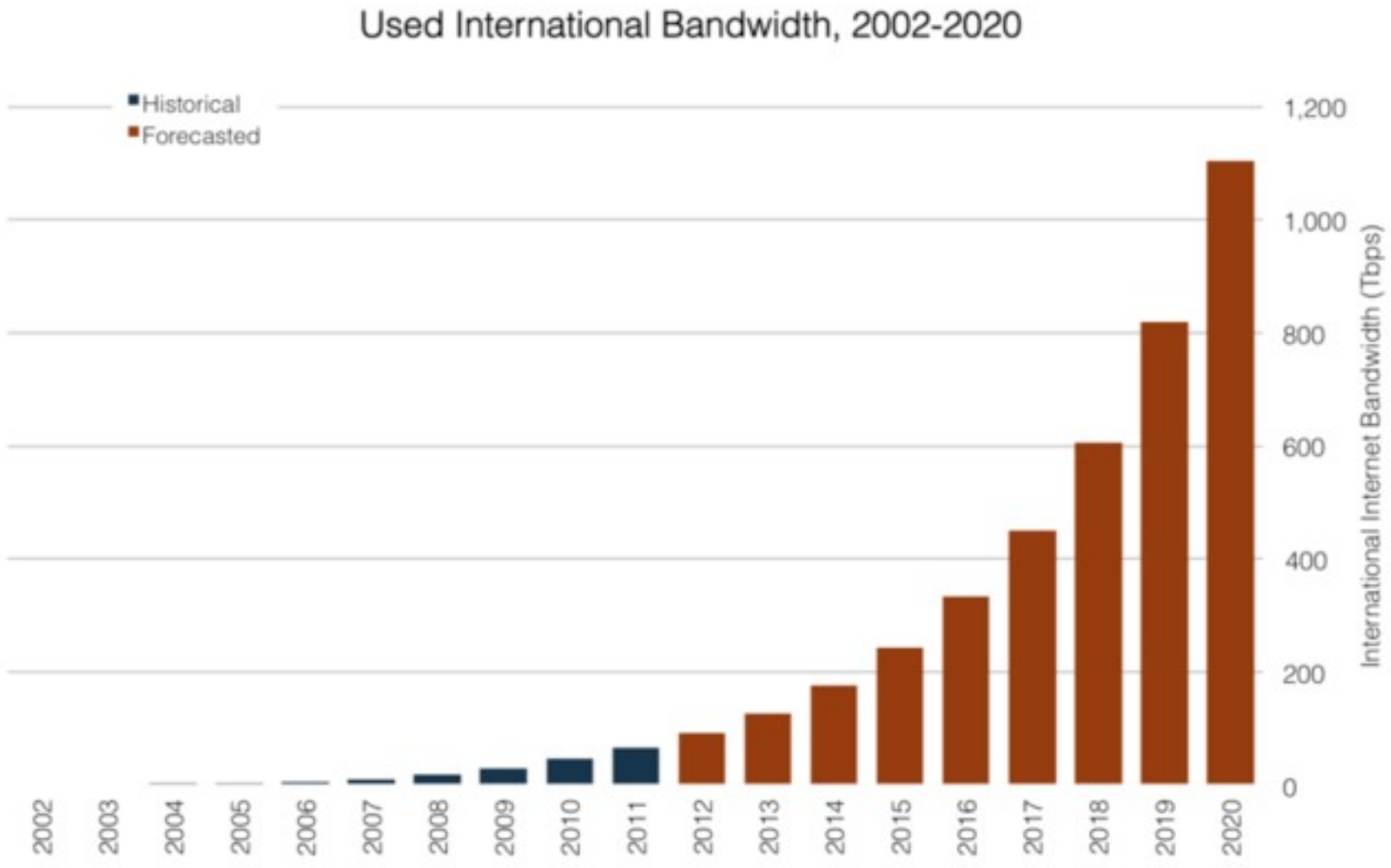


Que pasó con la Capacidad de Almacenamiento ?



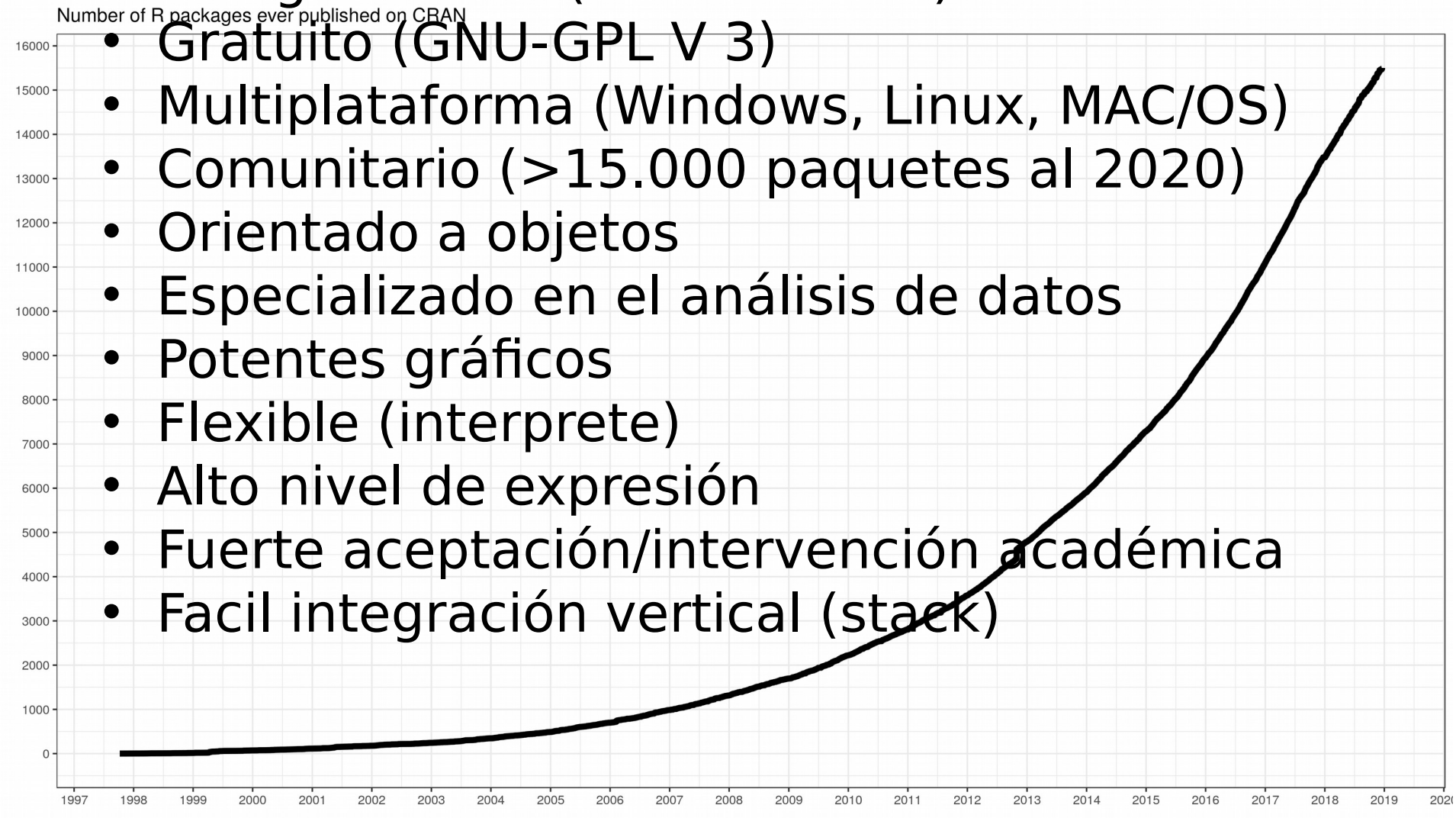
Compound Growing Rate

Que pasó con la Capacidad de Transmisión de Datos (Ancho de Banda) ?



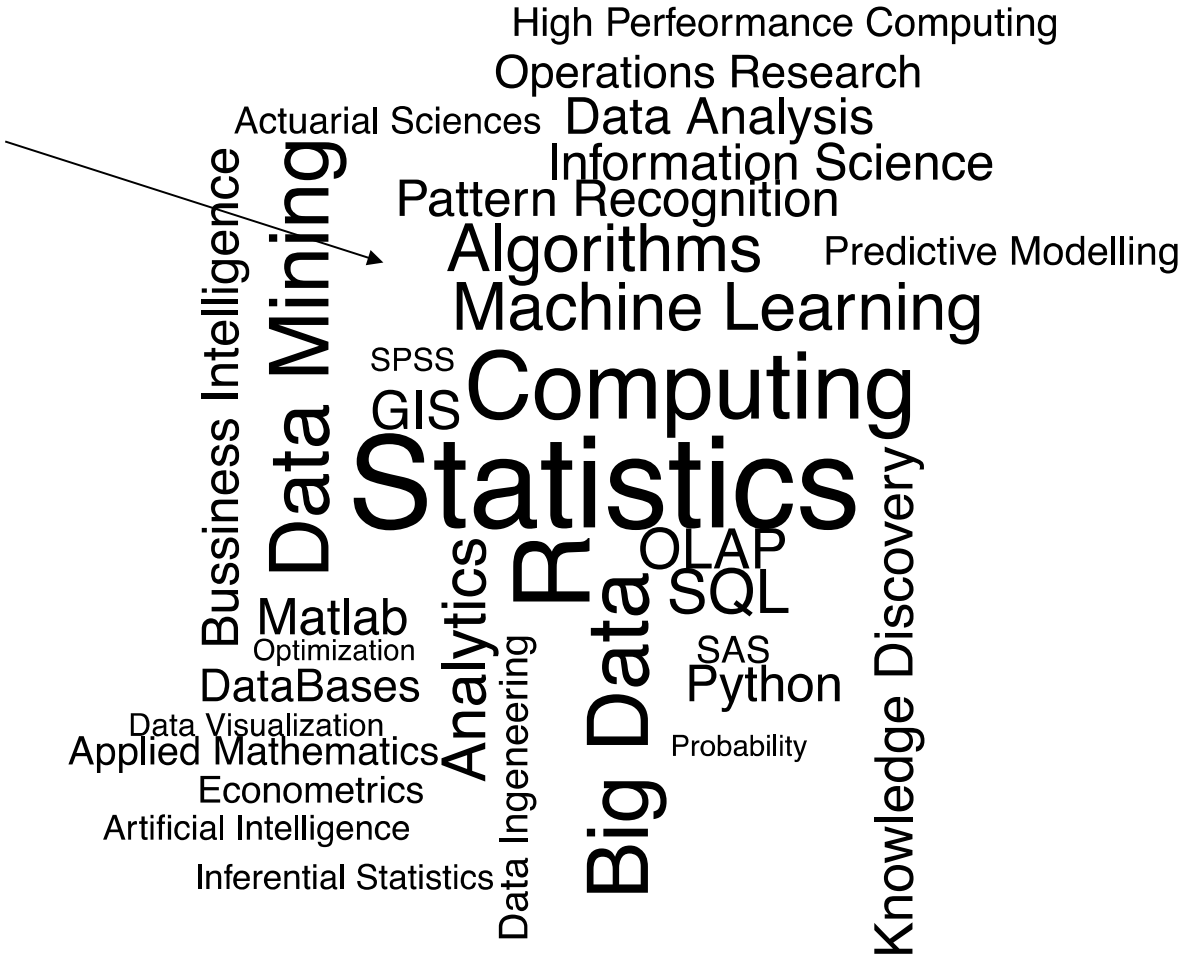
Porque R ?

- Código Abierto (GNU-GPL V 3)
- Gratuito (GNU-GPL V 3)
- Multiplataforma (Windows, Linux, MAC/OS)
- Comunitario (>15.000 paquetes al 2020)
- Orientado a objetos
- Especializado en el análisis de datos
- Potentes gráficos
- Flexible (interprete)
- Alto nivel de expresión
- Fuerte aceptación/intervención académica
- Facil integración vertical (stack)



Que es Ciencia de Datos ?

WordCloud de los Componentes de la Ciencia de Datos



Estadística

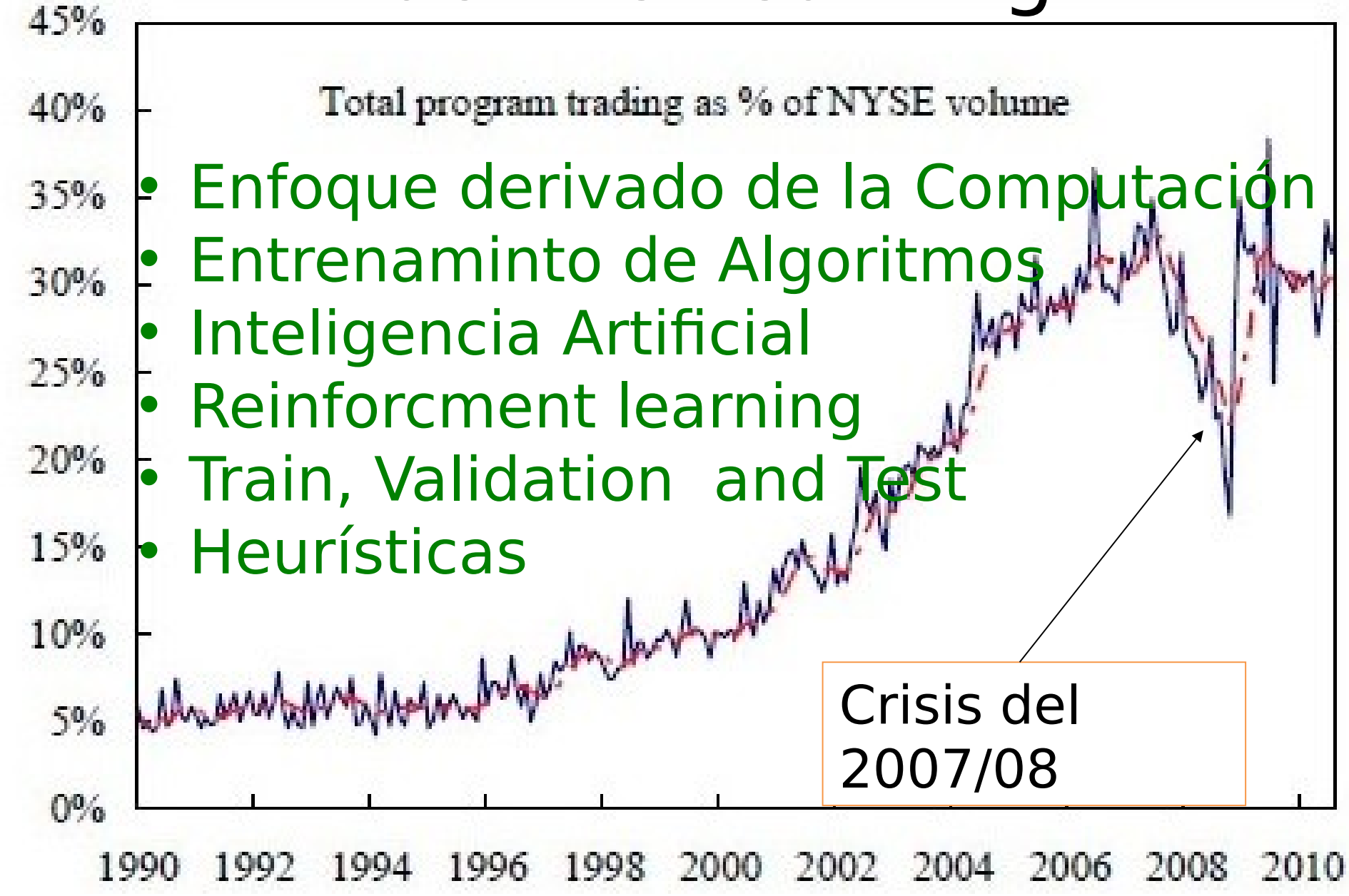
- Basada en la Teoría de Probabilidades.
- Formalizó el concepto de incertidumbre en las mediciones/estimaciones.
- Condicionada por la escacez de datos (N > 30 ?)
- Herramientas/conceptos básicos utilizados:
  - Modelo probabilístico
  - Población / Muestra
  - Variable Aleatoria
  - Verosimilitud
  - Inferencia
  - Significancia / P-valor
  - Intervalos de Confianza
  - Test de Hipótesis
  - Interpretabilidad

$$X = \mu + \epsilon$$

Componente Determinístico

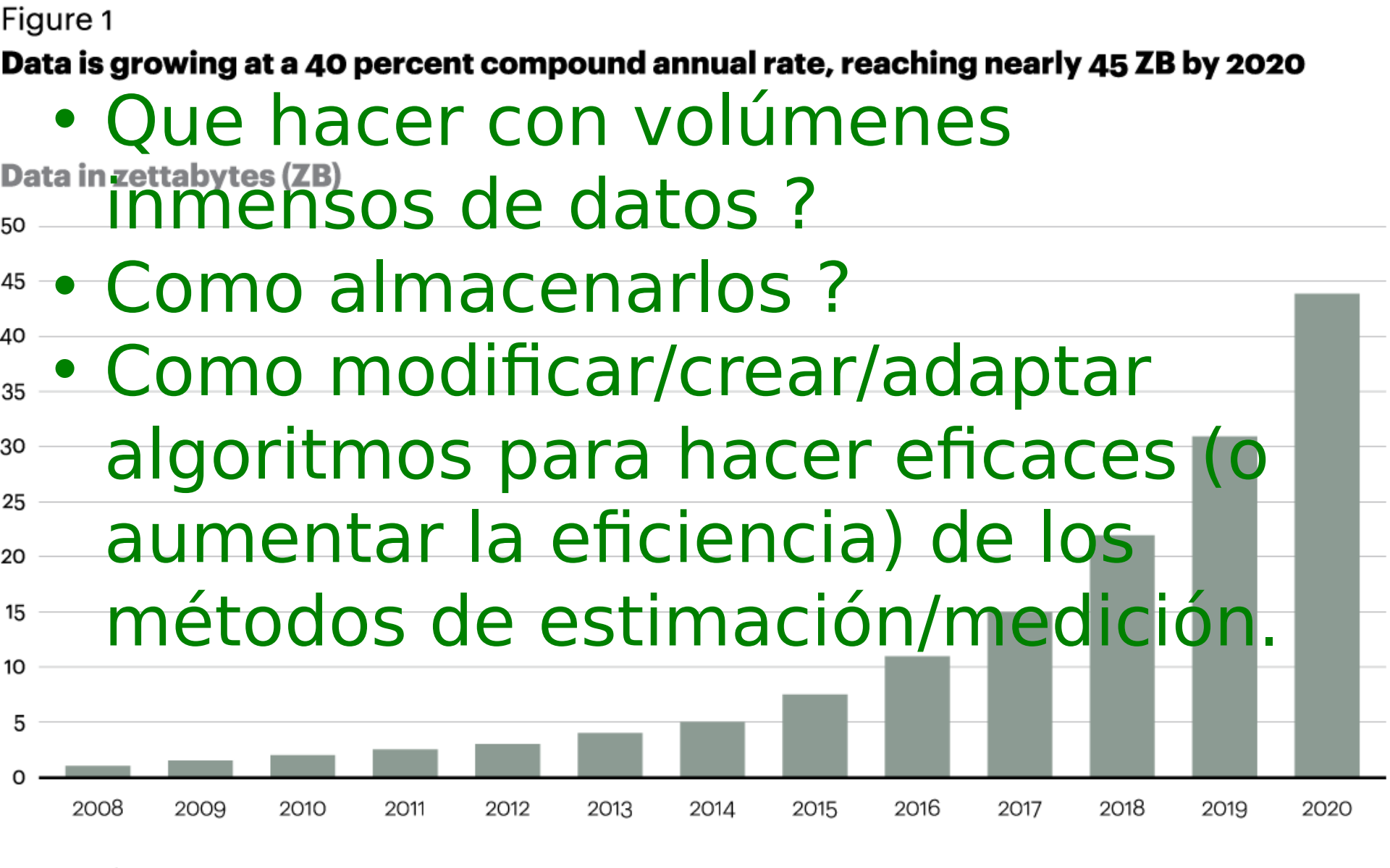
Componente Aleatorio

Machine Learning



- Enfoque derivado de la Computación
- Entrenaminto de Algoritmos
- Inteligencia Artificial
- Reinforcment learning
- Train, Validation and Test
- Heurísticas

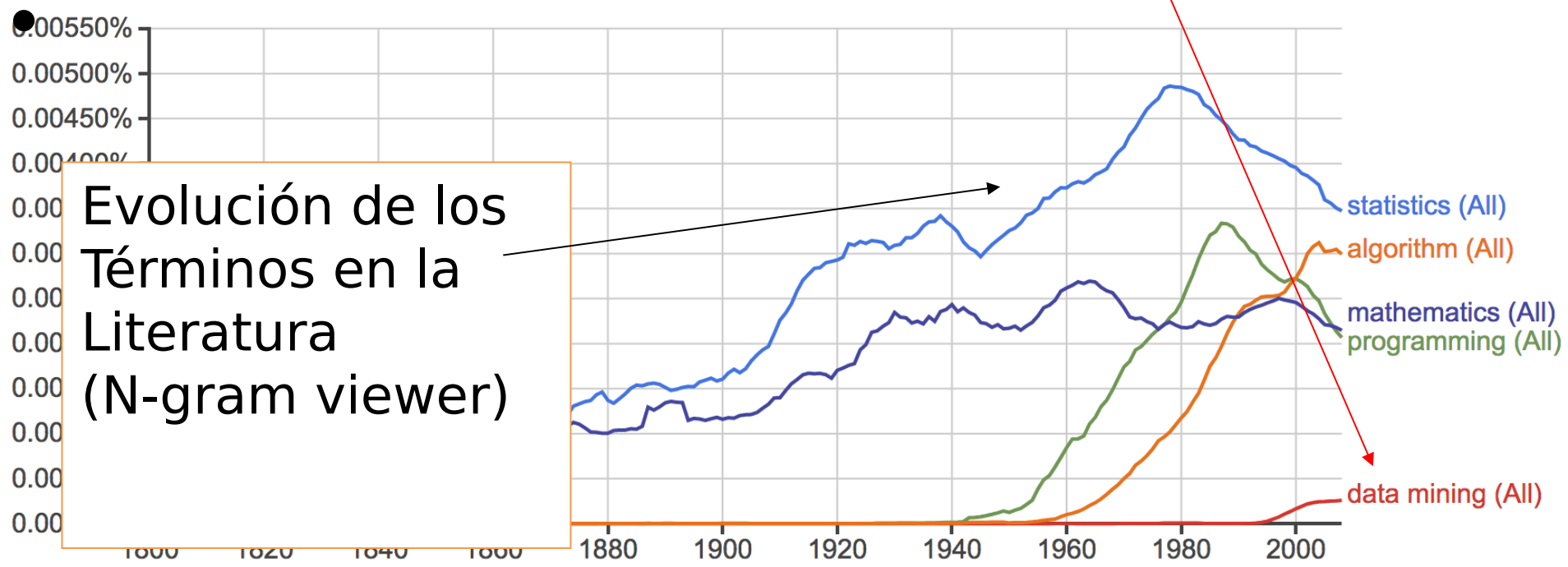
Big Data



- Que hacer con volúmenes inmensos de datos ?
- Como almacenarlos ?
- Como modificar/crear/adaptar algoritmos para hacer eficaces (o aumentar la eficiencia) de los métodos de estimación/medición.

Data Mining

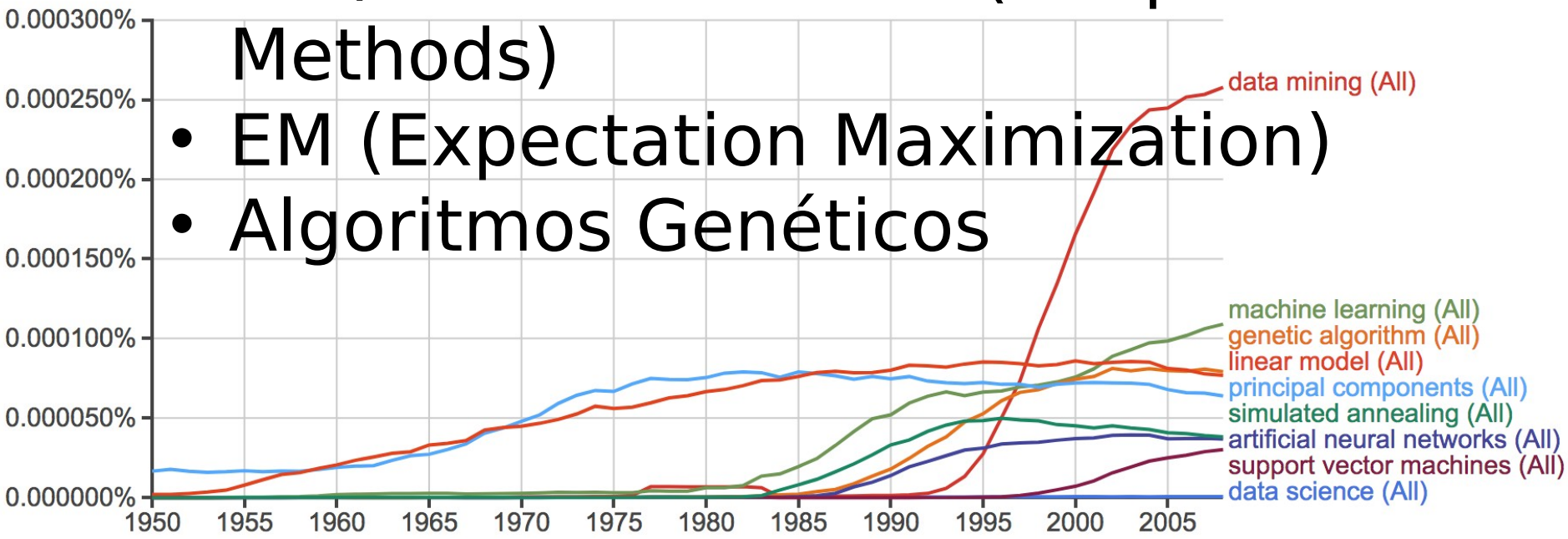
- Que patrones pueden extrarse de los datos ?
- Data Analysis y Analytics en gran escala
- Versión antigua del Data Science



Evolución de los Términos en la Literatura (N-gram viewer)

Optimización

- Fuerza Bruta
- Random Optimization (Luus-Jaakola)
- Gradient Descent
- Newton-Rapson (Quasi)
- Simulated Annealing
- Optimización Lineal/Cuadrática con/sin restricciones (Simplex Like Methods)
- EM (Expectation Maximization)
- Algoritmos Genéticos



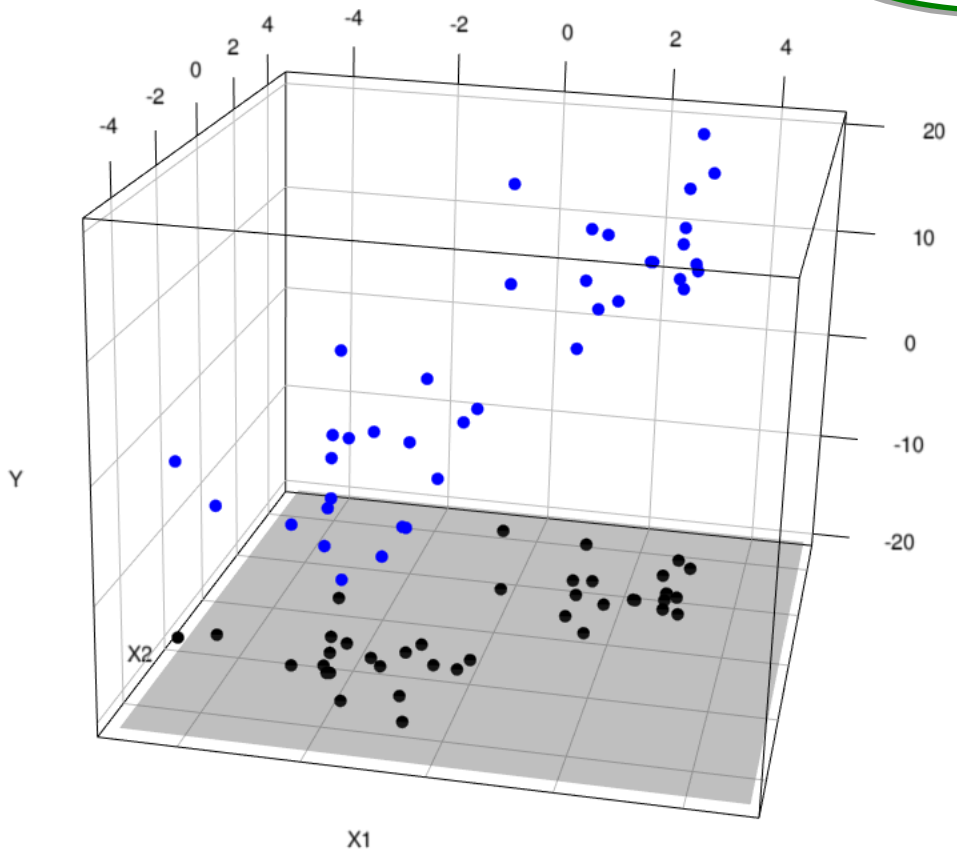
Taxonomía Basica de los Métodos en la Ciencia de Datos

- Métodos Supervisados
  - Clasificación
    - CART
    - Support Vrctor Machines
  - Regresión
    - Modelos Lineales
    - Redes Neuronales
- Métodos No Supervisados
  - Análisis Factorial
    - Componentes Principales
    - Análisis de Correspondencia
  - Segmentación
    - K-medias
    - Clusterización Jerarquica
    - GMM

Supervisado Vs. No Supervisado

Espacio de probabilidad conjunto

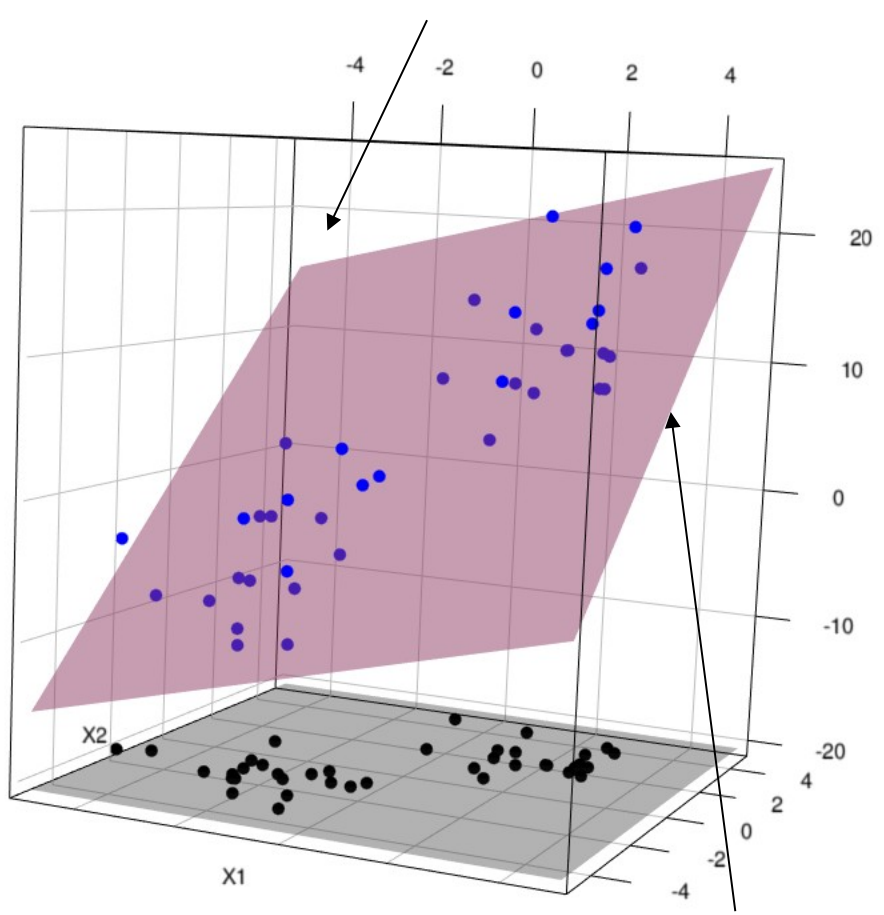
$$\Pr(X, Y) = \Pr(Y|X) \cdot \Pr(X)$$

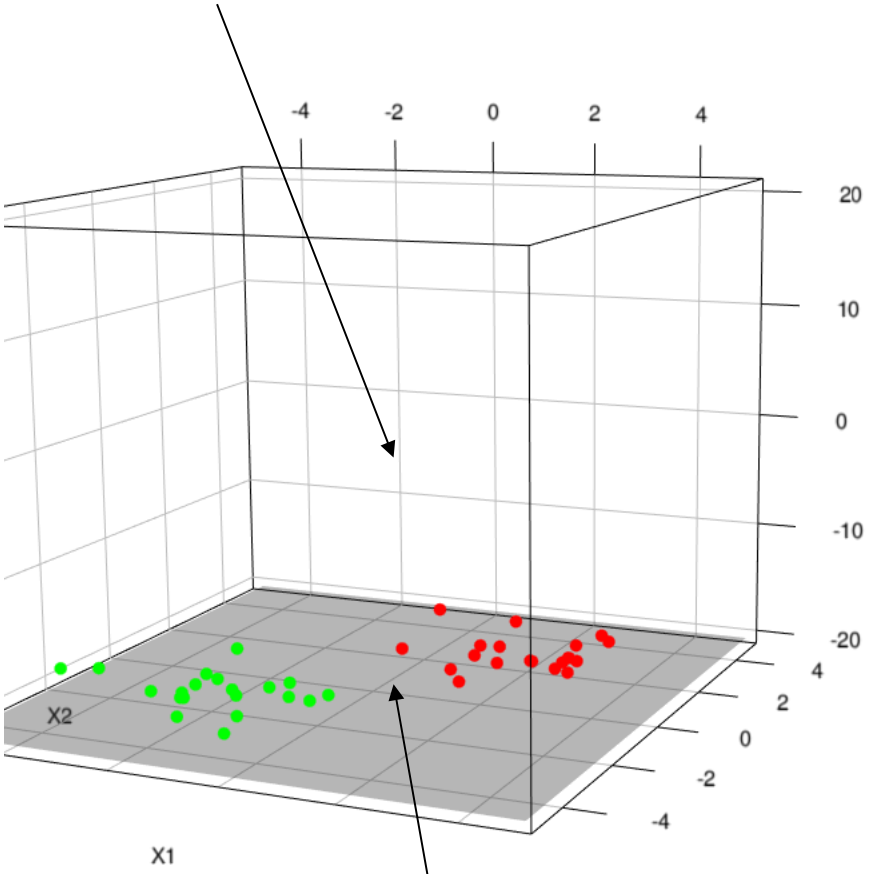


Variables Explicativas o Features

Variables Respuesta o Target

Regresión Vs. Clasificación

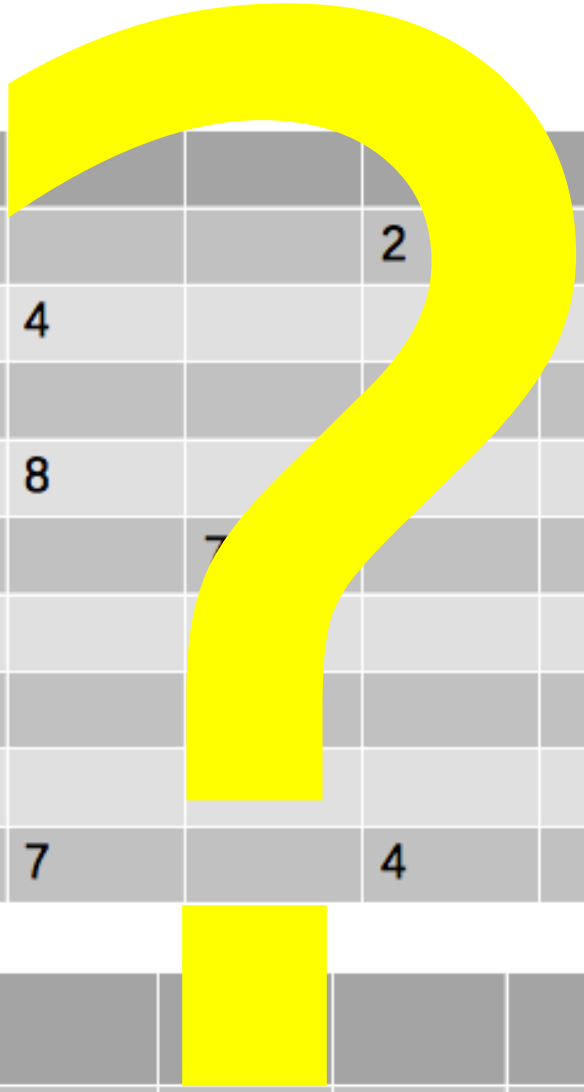




Relación entre Target y Features

Sólo Features

Sistema de Recomendaciones



Indiv.	Item 1	Item 2		Item j					Item p
1		3		9			2		
2	7	2			4				10
3			5						
4					8				3
5		1				7			
i			3	9					
N					7		4		

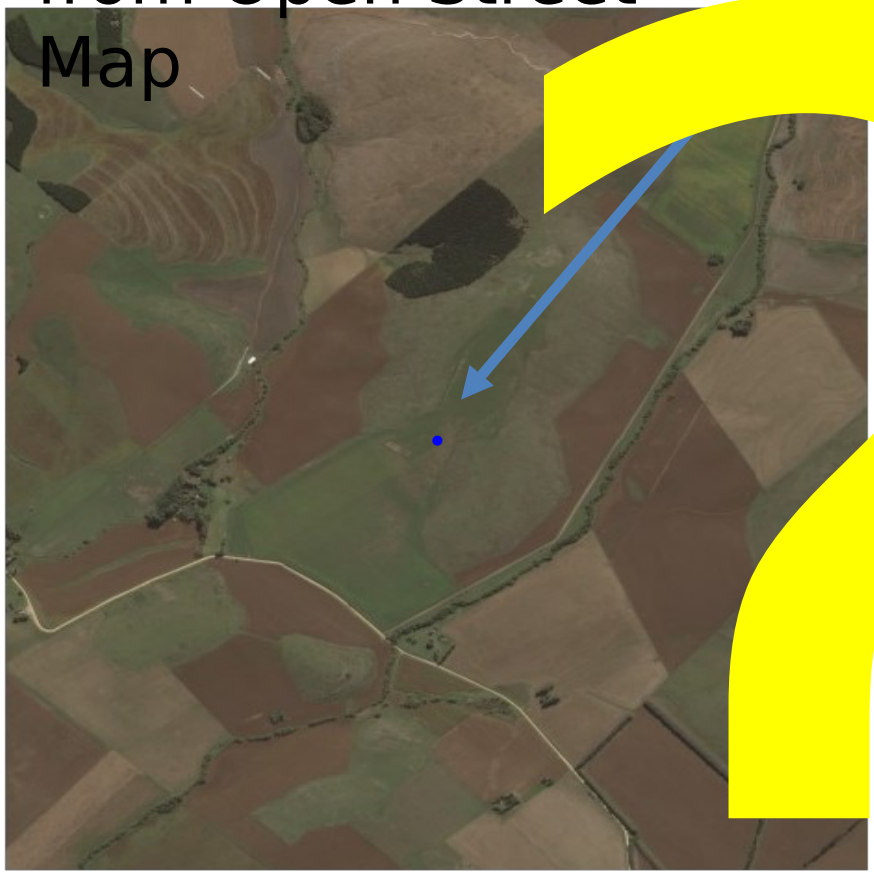
Indiv.	Item 1	Item 2		Item j					Item p
i*	?	6	?	?	?	8	?	3	?

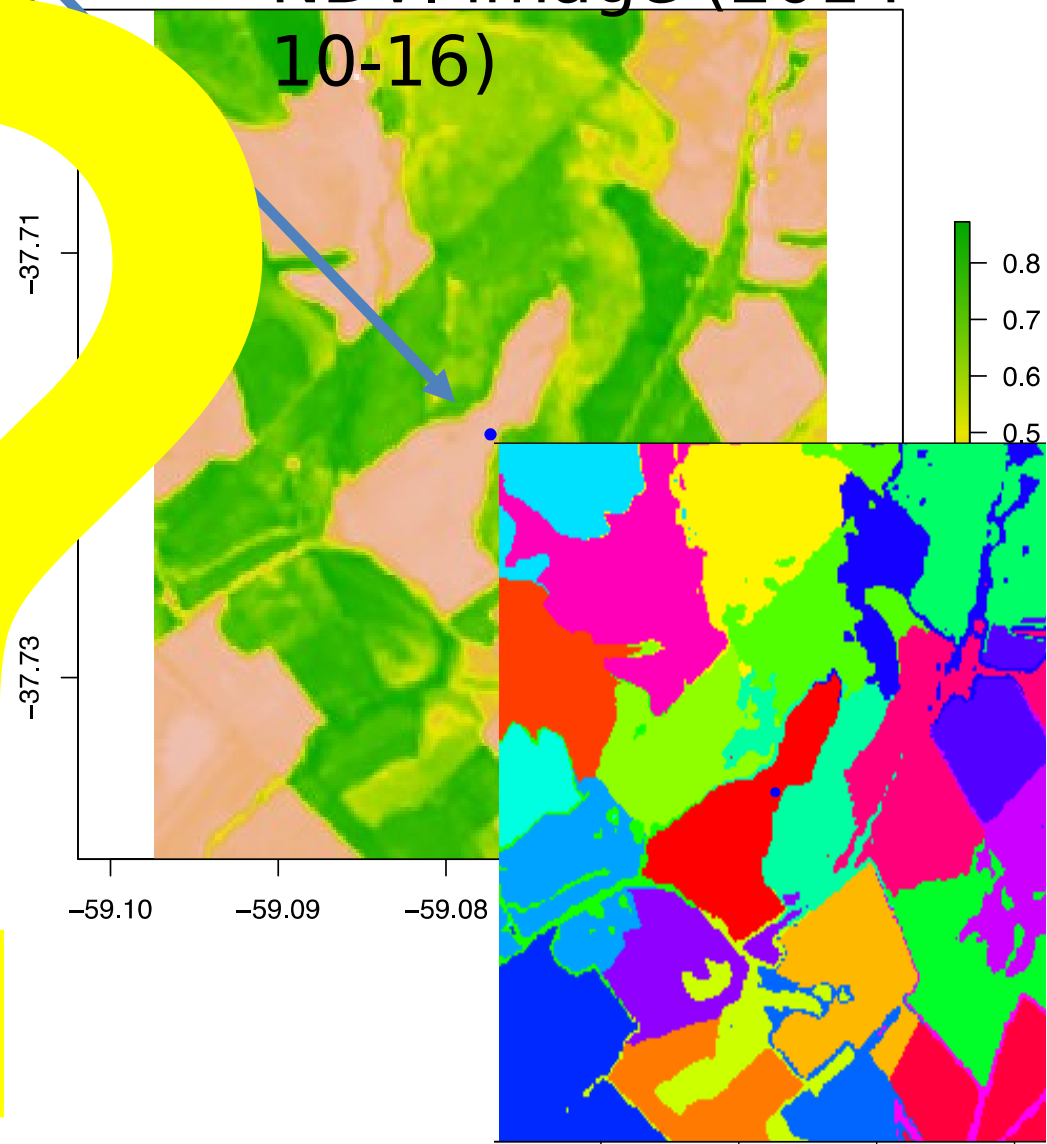
GeoReferenciación Automática

Visible image from Open Street Map

Point of interest

NDVI image (2014-10-16)





Symbolic Data Analysis
(Estadística de Objetos?)