

Introducción a Data Science



Copyright (C) DBlandIT SRL. Todos los derechos reservados.



UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Qué es Data Science?



Copyright (C) DBlandIT SRL. Todos los derechos reservados.

Definición Data Science

Según [Wikipedia](#):

“Campo interdisciplinario que involucra métodos científicos, procesos y sistemas para *extraer conocimiento o insights de los datos.*”

“Emplea técnicas y conceptos de otros campos como *estadística, matemática, ciencias de la computación, data mining & machine learning, base de datos y visualización.*”



WIKIPEDIA
The Free Encyclopedia

Definición Data Science

Según [New York University](#):

“At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.”

“With such automated methods turning up everywhere from genomics to high-energy physics, data science is helping to create new branches of science, and influencing areas of social science and the humanities.”



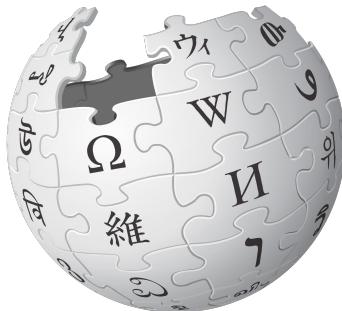
NEW YORK UNIVERSITY



Definición Business Intelligence

Según [Wikipedia](#):

“Conjunto de estrategias, aplicaciones, datos, productos, tecnologías y arquitectura técnicas, los cuales están enfocados a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa..”



WIKIPEDIA
The Free Encyclopedia

Definición Business Intelligence

Según [Microsoft](#):

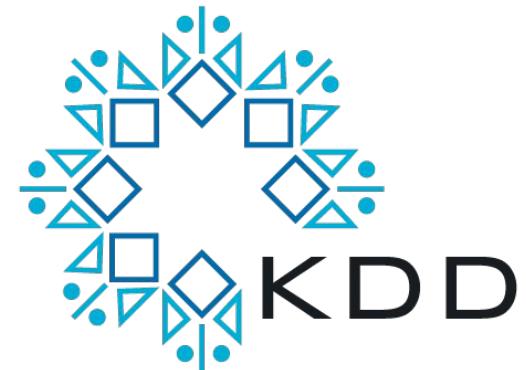
“Business intelligence (BI) simplifies information discovery and analysis, making it possible for decision-makers at all levels of an organization to more easily access, understand, analyze, collaborate, and act on information, anytime and anywhere.”



Definición Data Mining

Según SIGKDD.org:

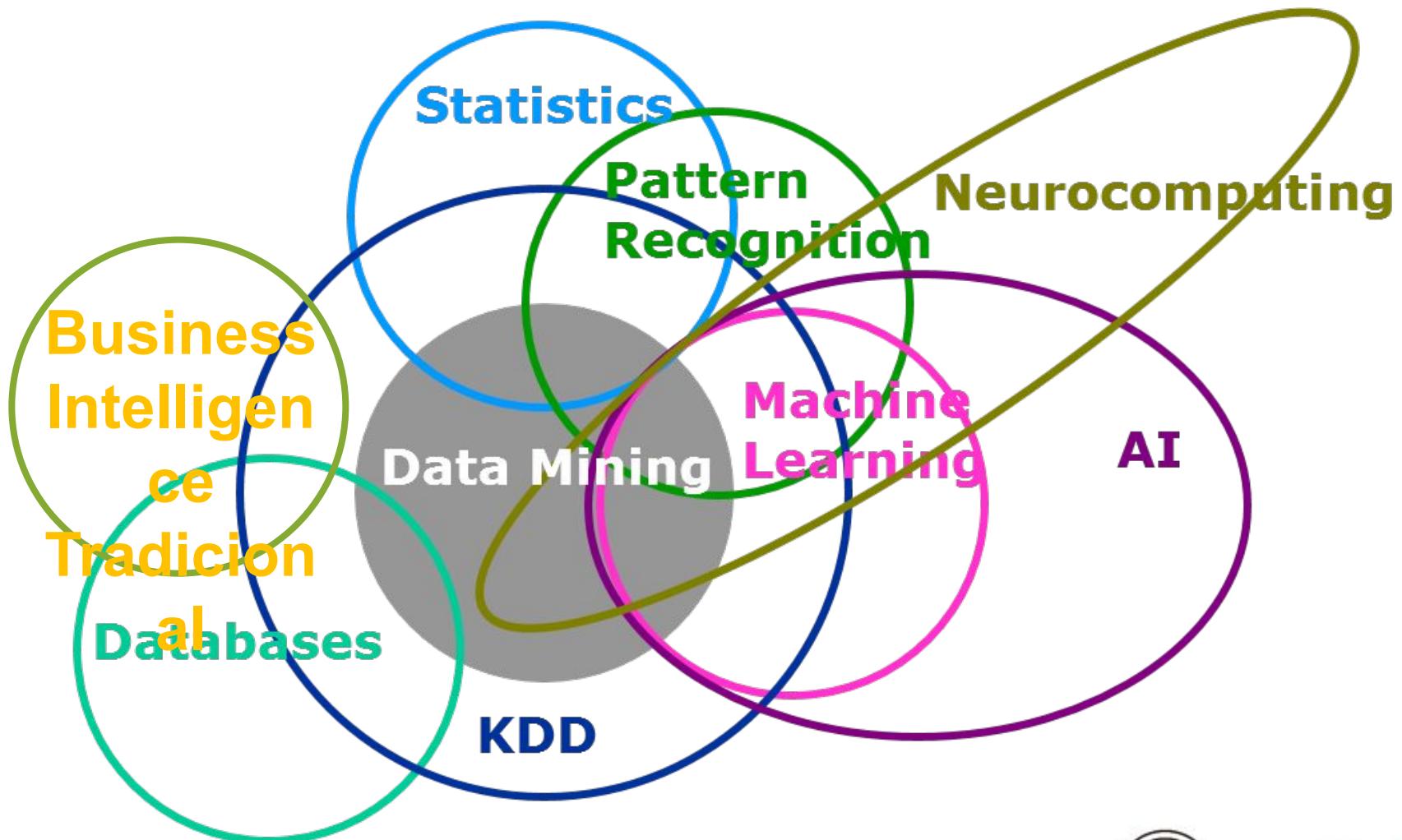
“The science of extracting useful knowledge or insights from data.”



Definición Data Science

- No existe un consenso sobre definición clara y formal ampliamente aceptada sobre qué es data science, qué no es y dónde comienza una disciplina y termina otra.
- Si hay consenso en que no hay una definición clara y formal.
- Es un término muy utilizado estos días (Incluso a veces considerado un Buzzword).

Analytics



Campos & disciplinas

Bastante similar a otras disciplinas...

- BI o Inteligencia del negocio.

- Estadística Aplicada

Se ocupa de inferir resultados sobre una población a partir de una o varias muestras.

- Machine Learning

Desarrollar técnicas que permitan a las computadoras aprender sin programarlas explícitamente).

- Analytics

Descubrimiento, interpretación y comunicación de patrones significativos en los datos.

- Inteligencia artificial:

La construcción de tecnología y sistemas que se comportan como los humanos tales como *Self-driving cars*, asistentes como *Siri*, casas inteligentes.

Data Science - Habilidades

- Conocimiento del negocio
- Estadística & Matemática
- Cs. Computación
- Comunicación / Story Telling

Data Science & BI Tradicional

BI Tradicional (Sistemas OLAP + Agregación) brinda información sobre hechos pasados.

Data Science + ML + DM estima y predice resultados futuros.

PASADO

FUTURO

- ¿Cuál fue la tasa de respuesta a la campaña?
- ¿Cuántas unidades del producto nuevo vendimos a nuestros clientes?
- ¿Cuáles fueron los top 10 clientes el año pasado?
- ¿Qué clientes dejaron de pagar sus prestamos?
- ¿Qué porcentaje de las piezas producidas en los últimos 10 meses fueron defectuosas?

- ¿Cuál es el perfil de los clientes que probablemente respondan a la futura campaña?
- ¿Cuáles de nuestros clientes son los que probablemente compren nuestro nuevo producto?
- ¿Qué 10 clientes representan las mayores posibilidades de ingresos futuros?
- ¿Cuáles son las probabilidades de que éste cliente deje de pagar?
- ¿Qué piezas van a presentar un defecto en los próximos 10 meses?

¿Por qué es importante?

Tomar Decisiones

~~Como se toman las decisiones?~~

~~Intuición vs. Realidad.~~

Tomar decisiones en escala

Realidad implica...

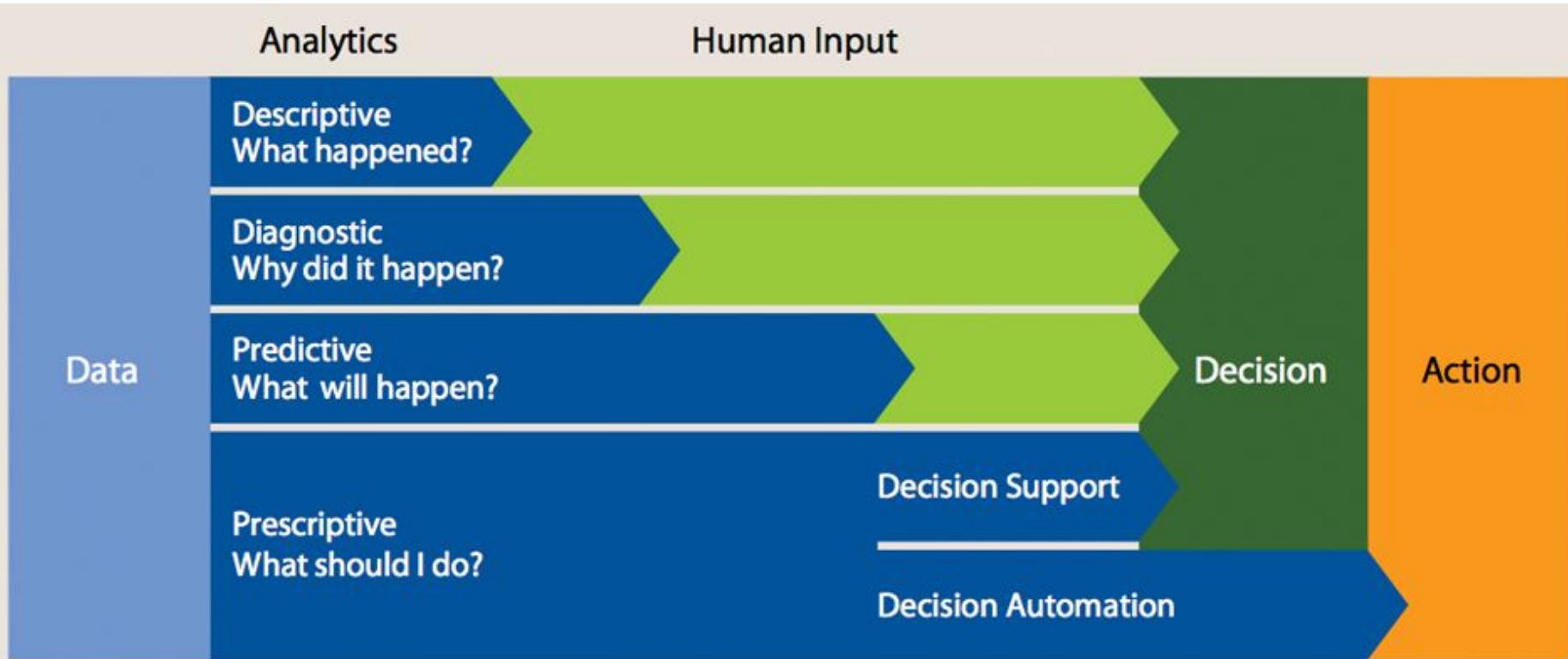
1. Identificar el problema
2. Encontrar alternativas
3. Evaluar resultados de cada una



Información y conocimiento forman la base del proceso de toma de decisiones



Hacia dónde vamos?



Source: Gartner, #G00254653 (September 2013)

Ejemplos



Customers who viewed this item also viewed these products



Dualit Food XL1500
Processor
\$560

Add to cart



Kenwood kMix Manual
Espresso Machine
 \$250

Select options



Weber One Touch Gold
Premium Charcoal
Grill-57cm
\$225

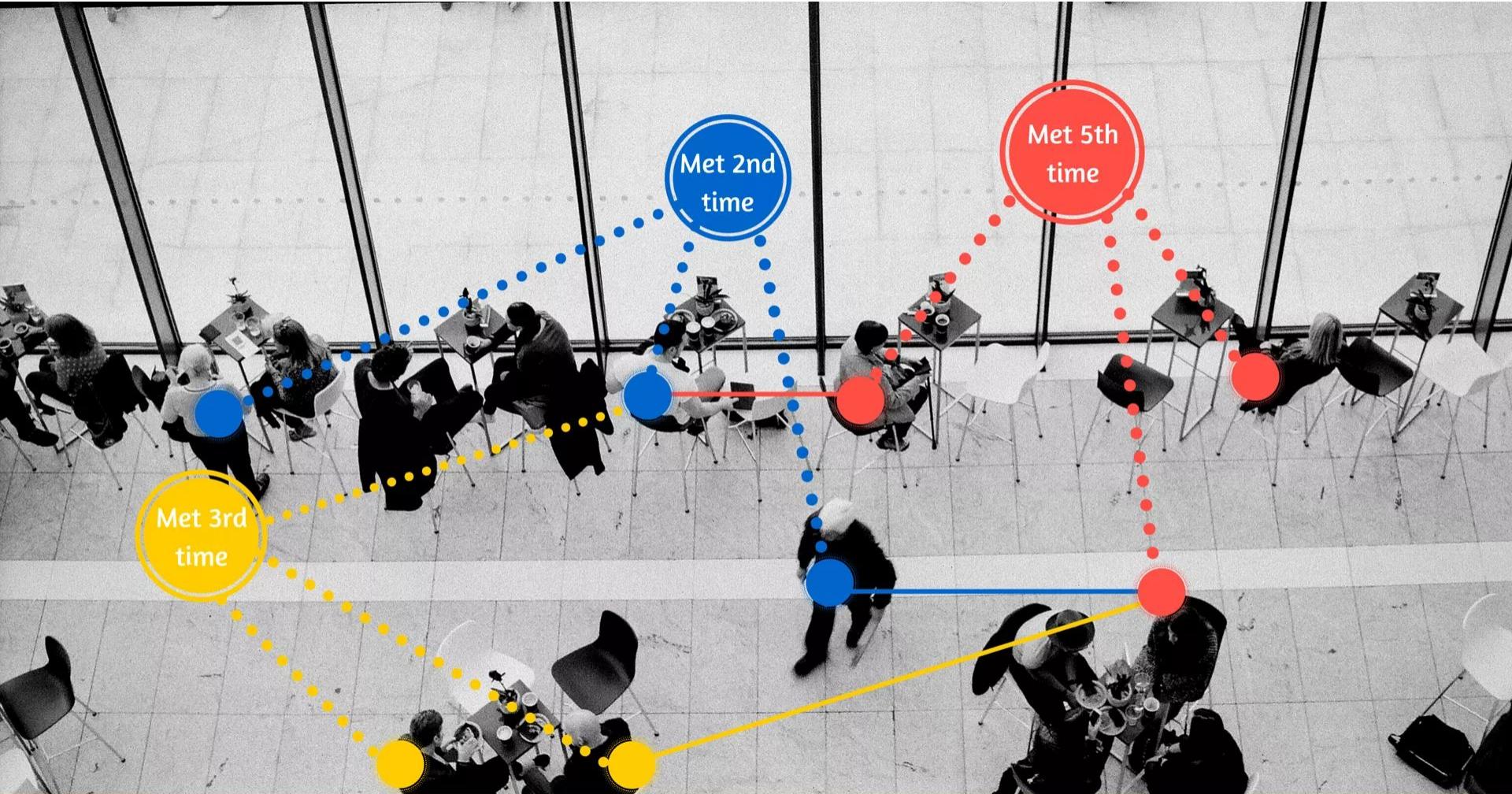
Add to cart



NoMU Salt Pepper and
Spice Grinders
\$3

View options

facebook®



Introducing Discover Weekly: your ultimate personalised playlist

Posted on July 20, 2015

- Our best-ever recommendations delivered to you as a weekly mixtape of fresh music
- New discoveries and deep cuts, based on what fans are playlisting around the songs you love

Updated every Monday morning, Discover Weekly brings you two hours of custom-made music recommendations, tailored specifically to you and delivered as a unique Spotify playlist.

For the first time ever, we're combining your personal taste in music with what similar fans are enjoying right now. This means every song in Discover Weekly is based both on your own listening as well as what others are playlisting and listening to around the songs you love – making your playlist completely unique and full of deep cuts and new discoveries.

It's like having your best friend make you a personalised mixtape every single week.

As your music taste evolves, so will Discover Weekly. In fact, it's designed to grow with you. Because it's a playlist you can access and listen to it across devices, sharing it with friends or making it available offline for your Monday morning commute.

"*High Fidelity*'s Rob Gordon had it right – the making of a Discover Weekly is the result of a collective effort," says Gustav Söderström, VP of Product. "It's a much simpler, more personalised way to discover music, with everyone's taste in mind."

You'll soon find your Discover Weekly playlist at the top of your library. Refreshed with new music every week, remember to save it to your profile for easy access.

It's time for Spotify to soundtrack your week! Who knows where it might lead.



Netflix

NETFLIX [Browse ▾](#) [Search](#)



Billions
94% Match 2017 16+ 2 Seasons
They're both ruthless, relentless, and refuse to lose. With a gargantuan fortune at stake, neither one is backing down.





Because you watched House of Cards



Binge-worthy US TV Thrillers



Copyright (C) DBlandIT SRL. Todos los derechos reservados.



UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

ORIGINAL DE NETFLIX HOUSE of CARDS

- Apuesta: De distribuidor a generador de contenido.
- Netflix's Chief Communications Officer:

“Because we have a direct relationship with consumers, we know what people like to watch and that helps us understand how big the interest is going to be for a given show. It gave us some confidence that we could find an audience for a show like House of Cards.”

- Qué sabían:
 - Gran cantidad de usuarios que vieron películas de David Fincher.
 - Versión inglesa de “House of Cards” fue muy vista.
 - Muchos de esos usuarios también vieron películas de Kevin Spacey y/o David Fincher.
- 10 trailers distintos.. a media.



Comercio & Marketing

- Identificar patrones de compra
- Buscar asociaciones entre clientes y características demográficas
- Segmentación de clientes para campañas de marketing
- Análisis de canasta de compra

Bancos

- Detectar patrones de uso fraudulento de tarjetas
- Identificar clientes leales
- Identificar clientes con probabilidad de cambiar de categoría
- Encontrar correlaciones entre indicadores financieros
- Identificar reglas de mercados de valores

Seguros & Salud

- Análisis de procedimientos médicos solicitados en conjunto
- Identificar clientes para nuevos servicios
- Identificar patrones de comportamiento en clientes con riesgo
- Detectar comportamiento fraudulento

Medicina

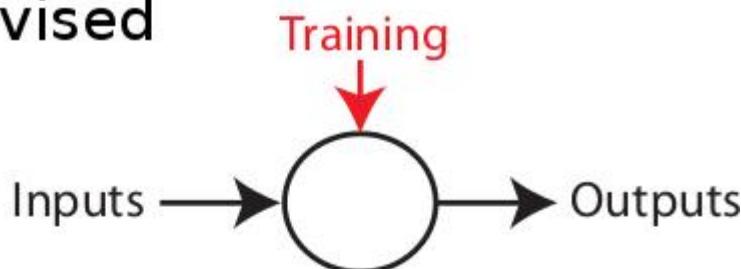
- Identificación de terapias médicas satisfactorias para distintas enfermedades
- Asociación de síntomas y patologías
- Estudio de factores de riesgo/salud
- Segmentación de pacientes para atención inteligente del grupo
- Estudios epidemiológicos
- Análisis de rendimientos de campaña de información (prevención).
- Predicción de requerimientos de los centros asistenciales para la asignación óptima de recursos.

¿Cómo modelar un problema?

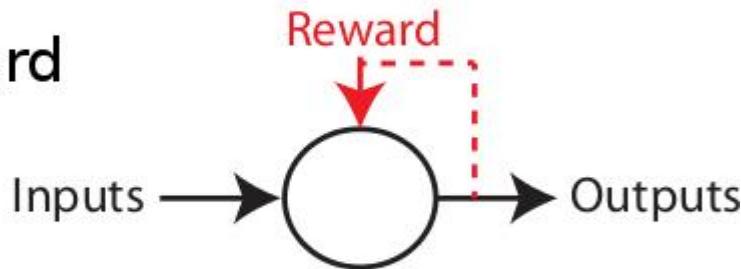
Machine Learning

Algoritmos Supervisados vs No Supervisados

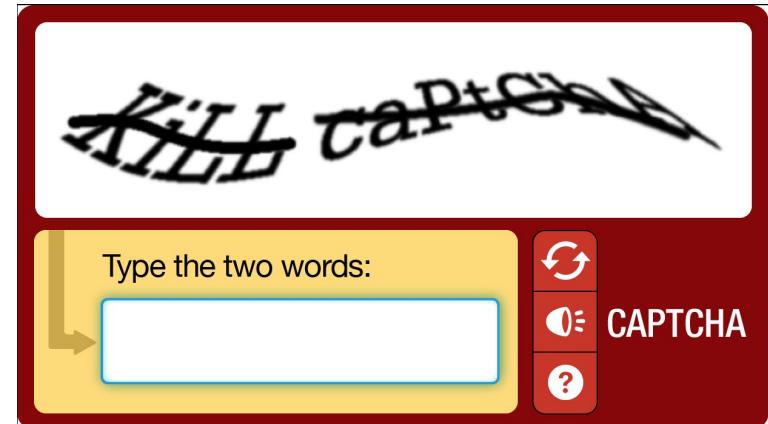
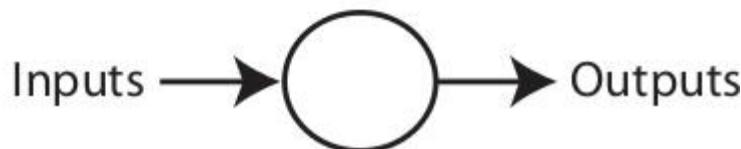
Supervised



Reward



Unsupervised





I'm not a robot



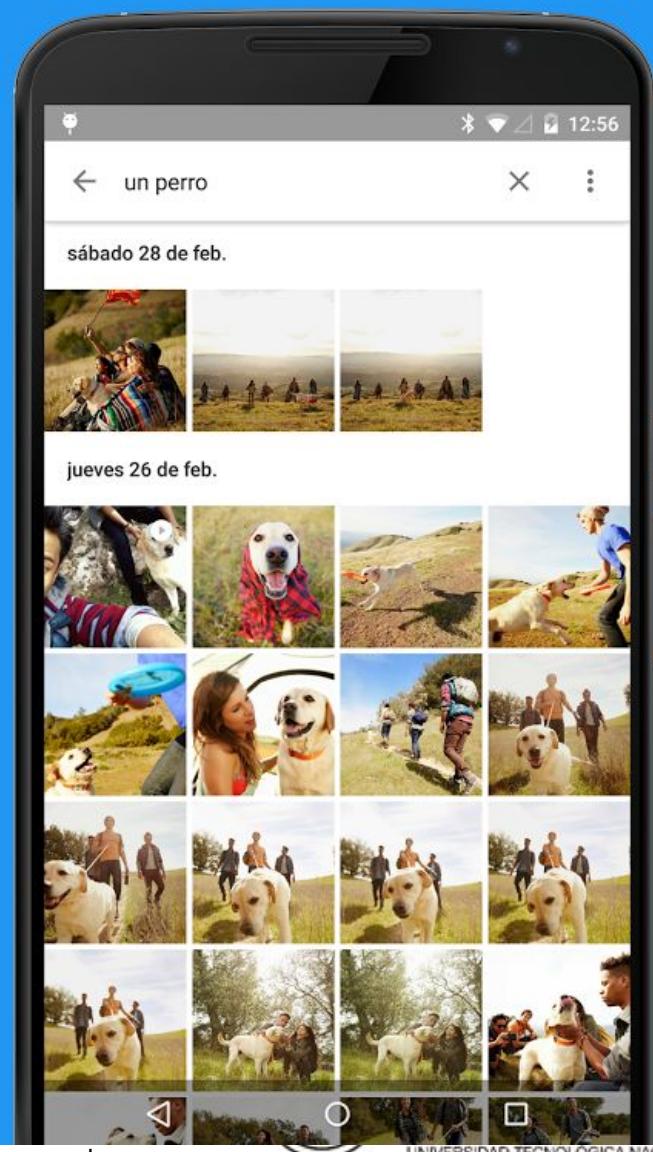
Select all dogs below. A sample image is on the right.



Verify

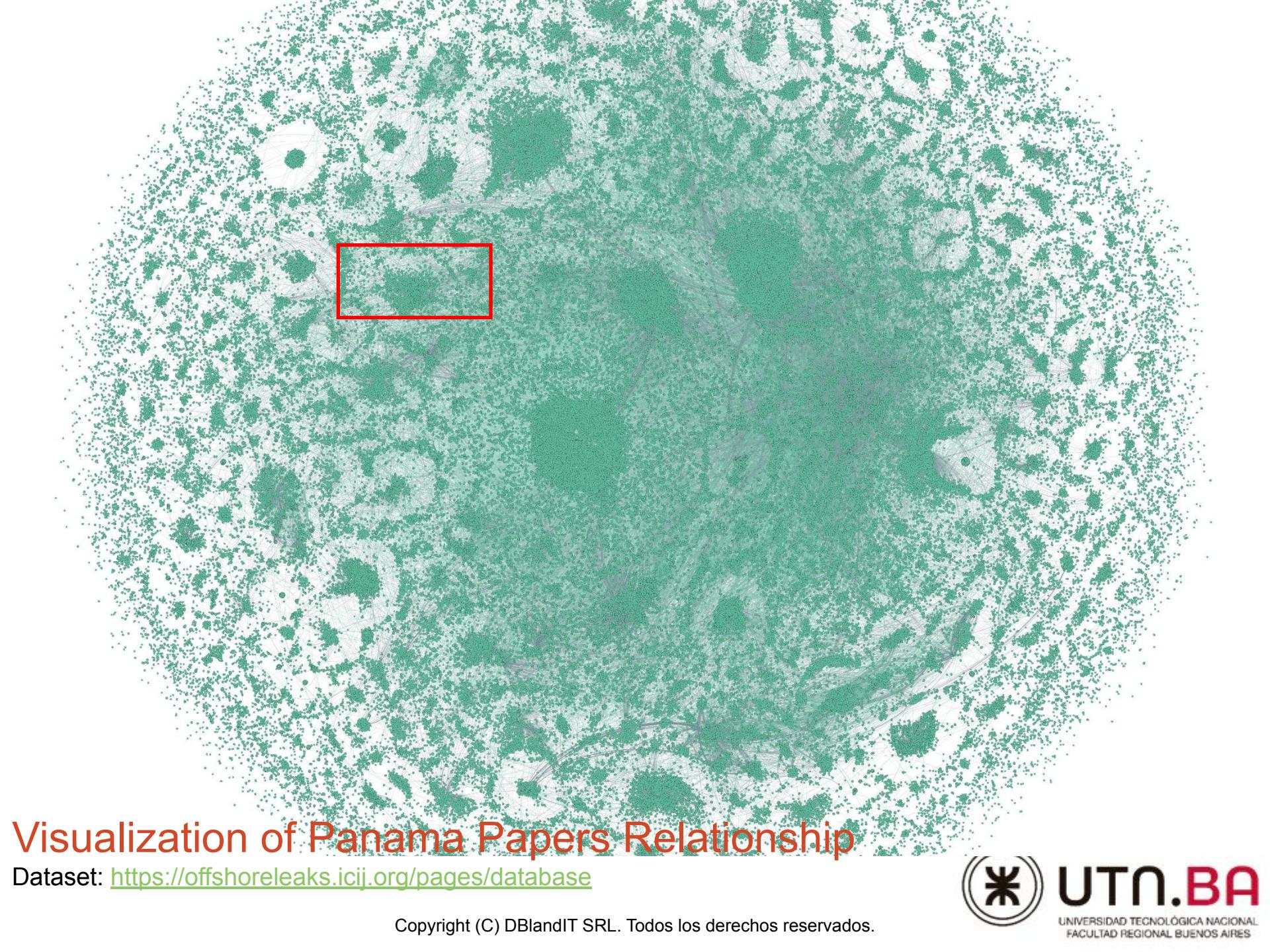
Copyright (C) DBlandIT SRL. Todos los derechos reservados.

Haz búsquedas según
lo que recuerdes



UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES





Visualization of Panama Papers Relationship

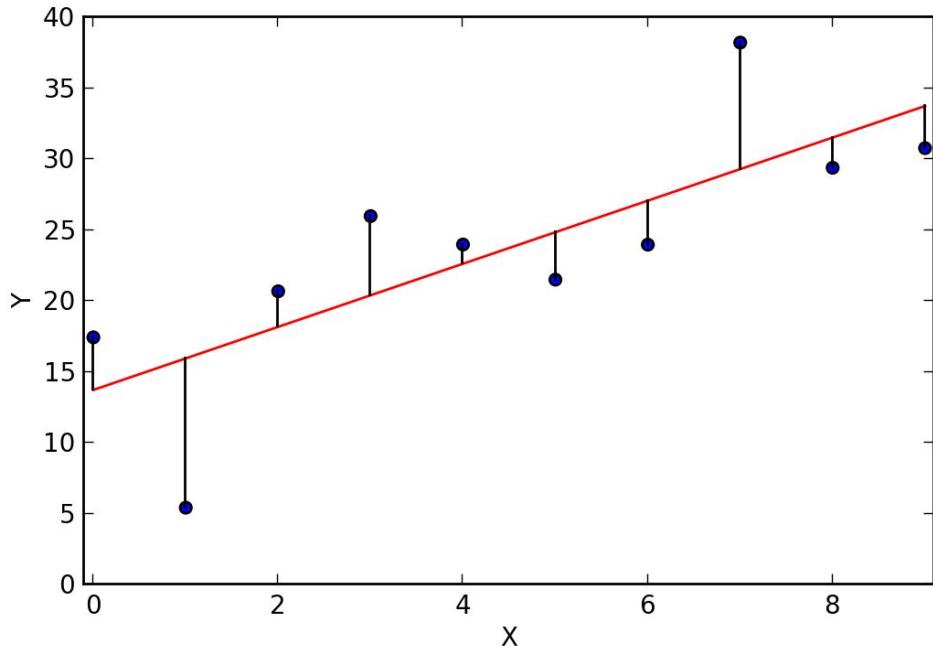
Dataset: <https://offshoreleaks.icij.org/pages/database>

Copyright (C) DBlandIT SRL. Todos los derechos reservados.

Algoritmos

Regression

Regression



Regresión lineal

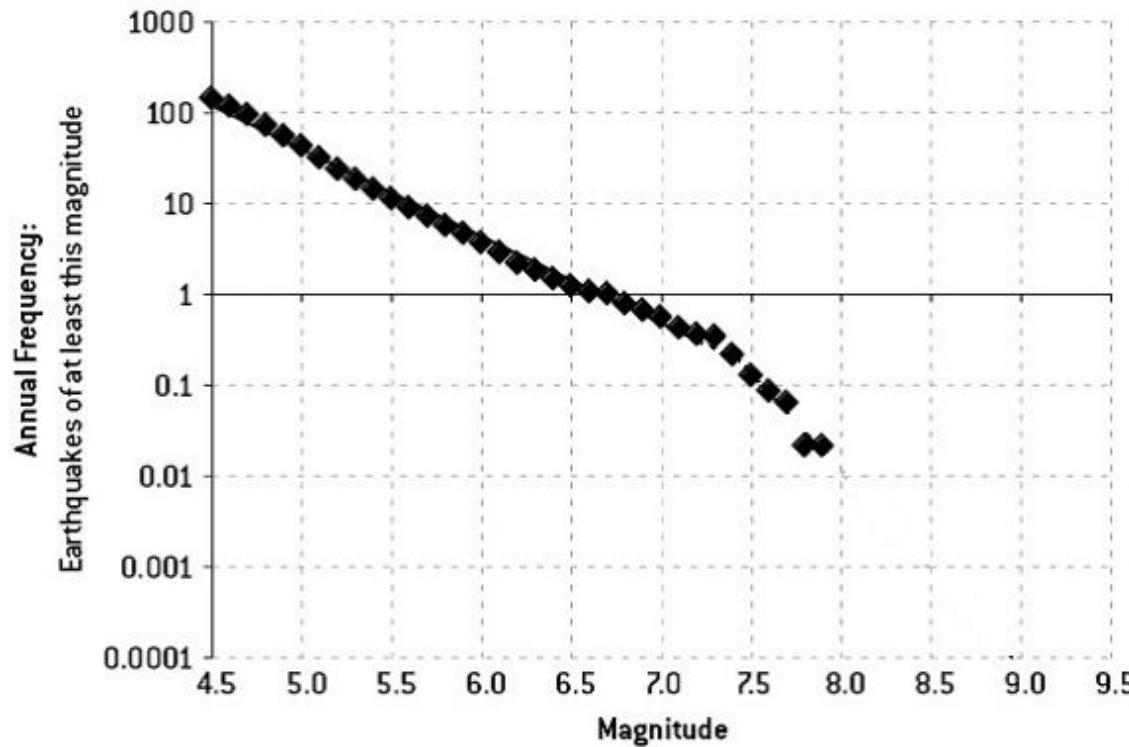
- Busca obtener una función que modela los datos con el menor error posible.
- Muestra como una variable dependiente cambia en función de una o mas variables independientes.
- Correlación no implica causalidad!

Un ejemplo...

Construir una central nuclear en una zona sísmica

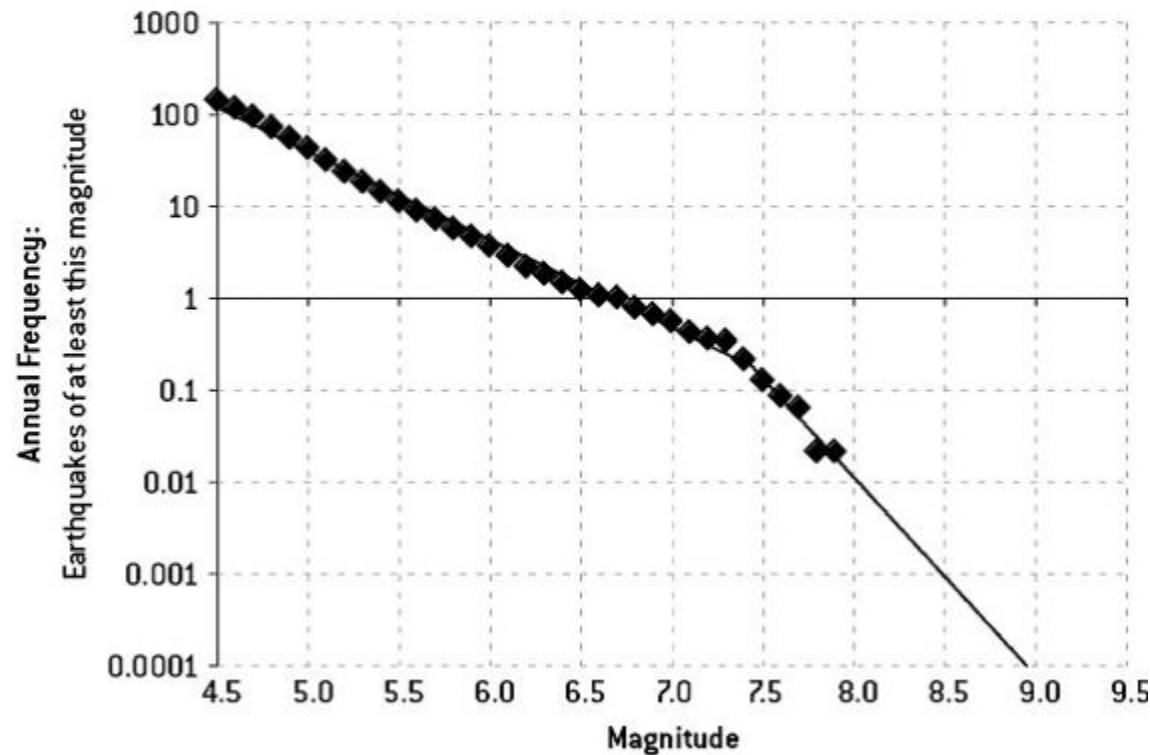


Modelo Predicción de Terremotos



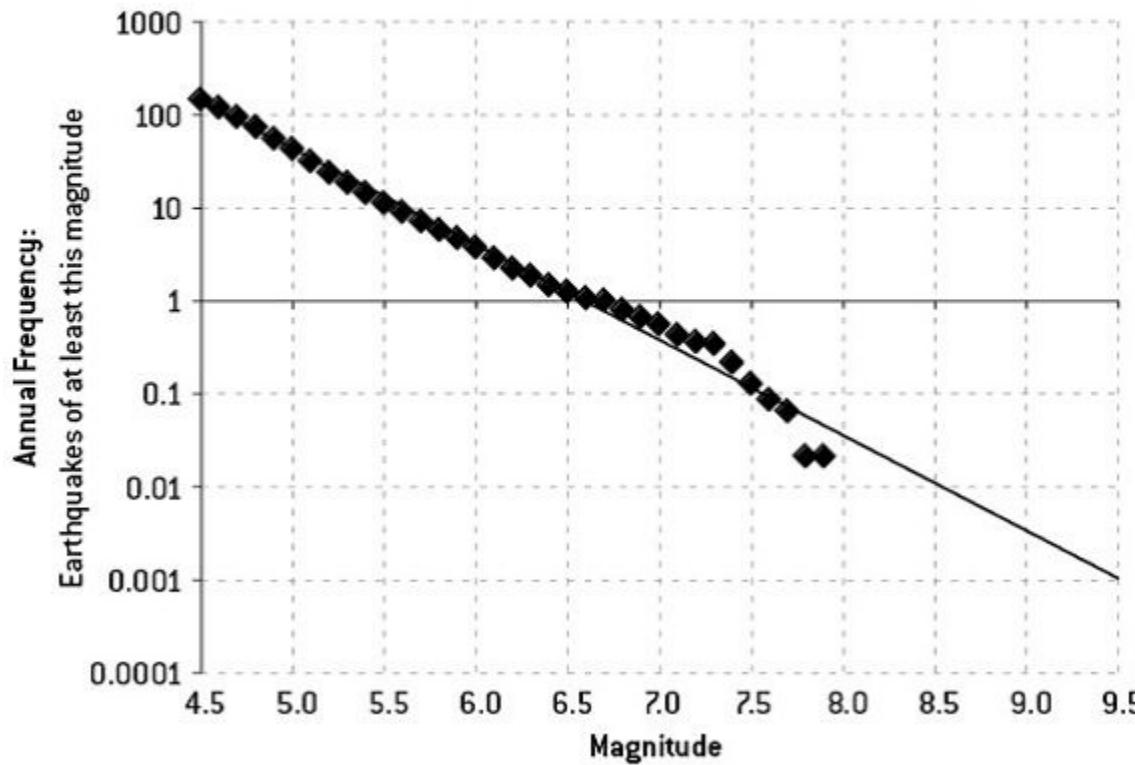
Alternativa 1

FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
CHARACTERISTIC FIT



Alternativa 2

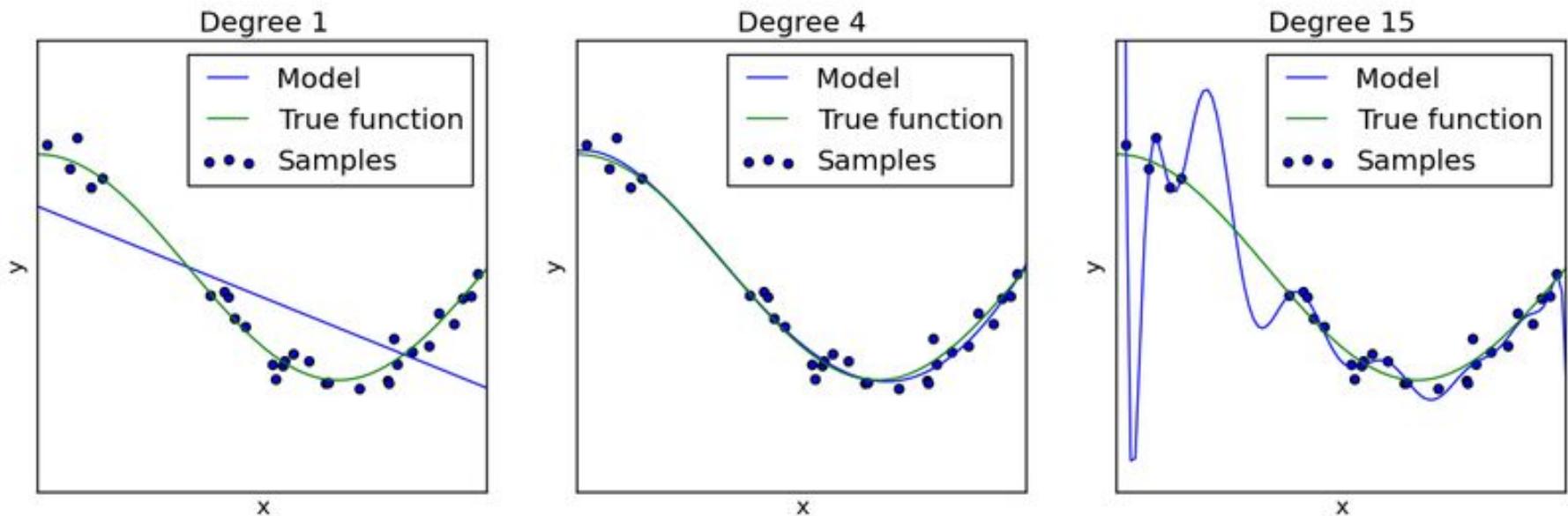
FIGURE 5-7B: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
GUTENBERG-RICHTER FIT



Que paso?



Regression: Overfitting vs. Underfitting



Clasification



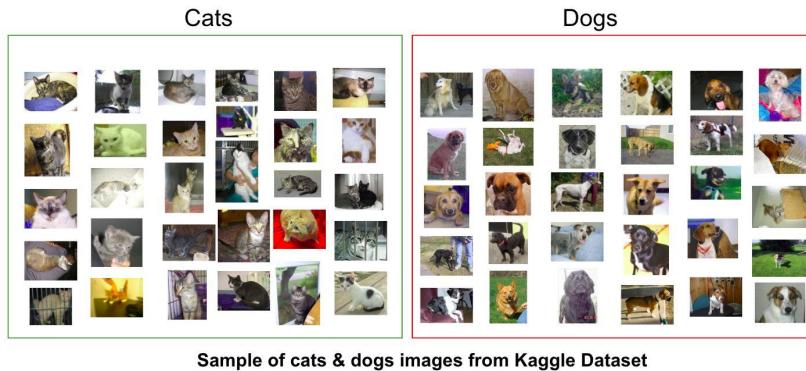
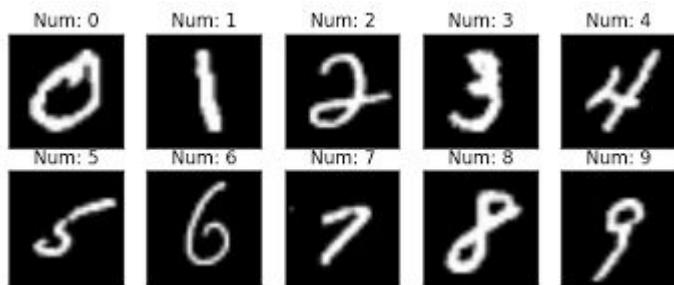
Clasification

Dividir en grupos predefinidos o categorías (categorial labels).

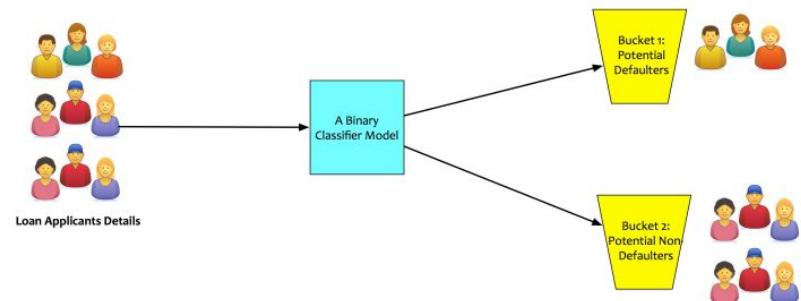
- Datos de entrenamiento para los cuales se conoce a que grupo pertenecen.
- Modelo es entrenado y luego predice a que grupo pertenecen nuevos elementos.

Algunos usos:

- Identificar transacciones fraudulentas.
- Categorizar clientes riesgosos.



Sample of cats & dogs images from Kaggle Dataset



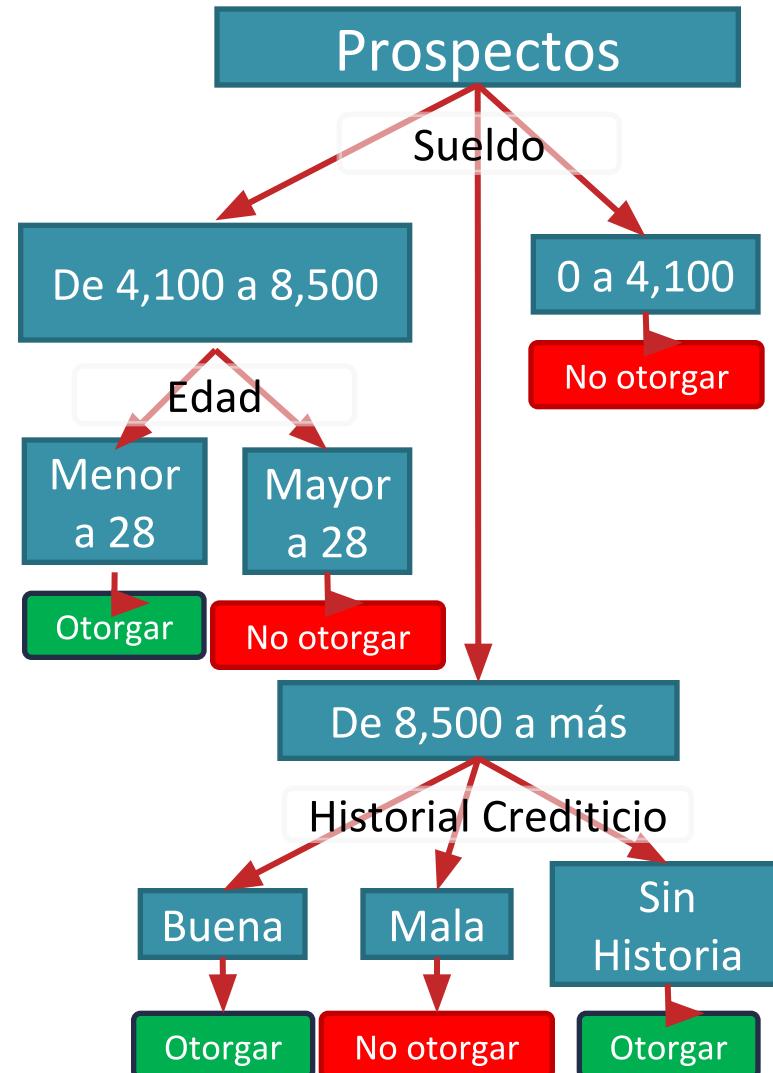
C4.5: Un árbol de decisión...

Clasifica mediante la construcción un árbol de decisión basado en las características de los elementos.

Explicación de resultados.

Fácil implementación de reglas resultantes:

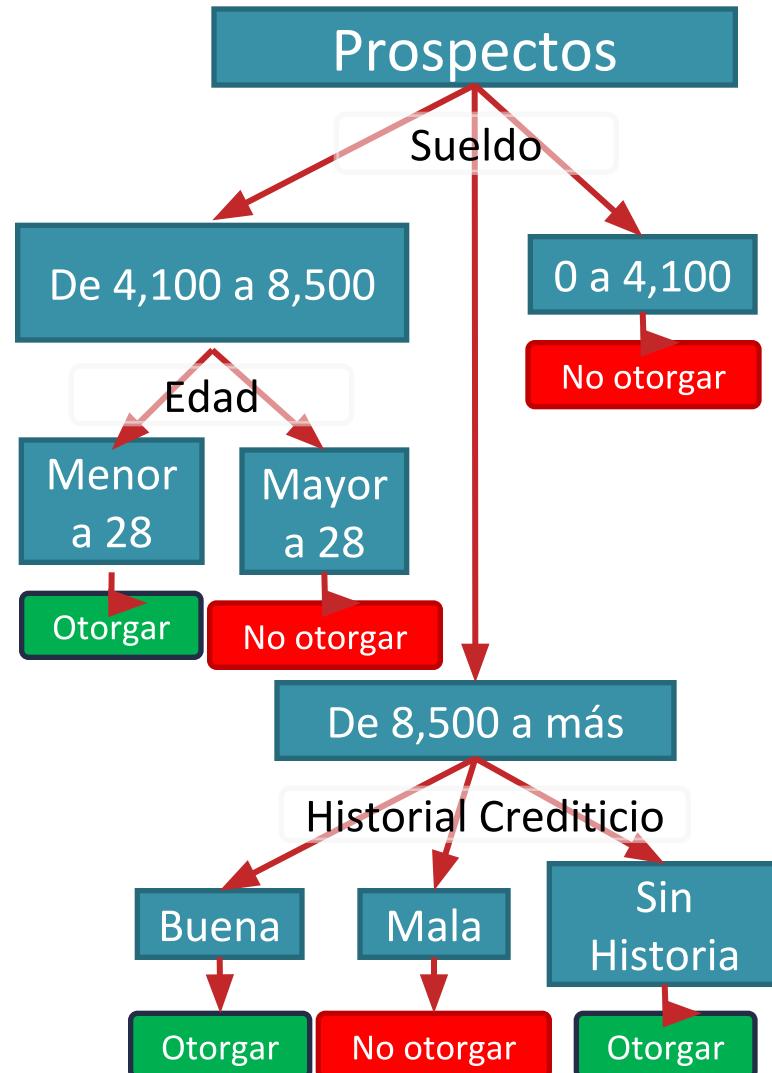
```
Si Ingreso < 4100
Entonces 'No'
Si Ingreso > 4100 Y Edad > 28
Entonces 'No'
Si Ingreso > 4100 Y PerfilCreditico =
Mala
Entonces 'No'
En otro, Caso 'Si'
```



C4.5: Un árbol de decisión...

Pseudocódigo general para generar árboles de decisión:

1. Comprobar los casos base.
2. Para cada atributo a
 1. Encontrar la ganancia de información normalizada de la división de a .
3. Dejar que a_{best} sea el atributo con la ganancia de información normalizada más alta.
4. Crear un nodo de decisión que divida a_{best} .
5. Repetir en las sublistas obtenidas por división de a_{best} , y agregar estos nodos como hijos de nodo.



Clustering



Clustering

Agrupar elementos en grupos (llamados clusters) en base a su similitud.

Parámetros:

- Número de clusters esperados
- Umbral de densidad
- Función de distancia

Relación entre clusters

- Hard clustering (Excluyentes)
- Fuzzy clustering (Grado de asociación)

Clustering Models

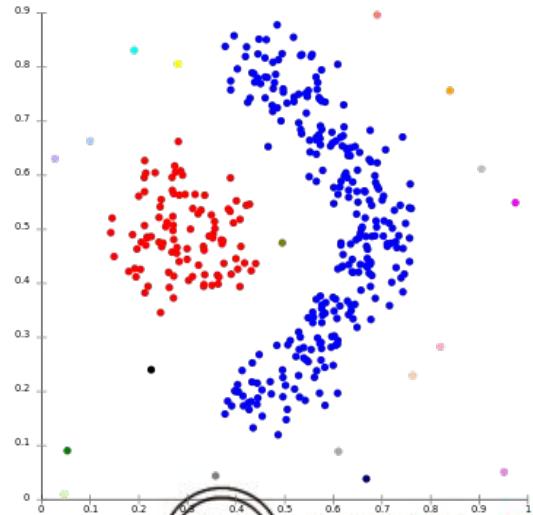
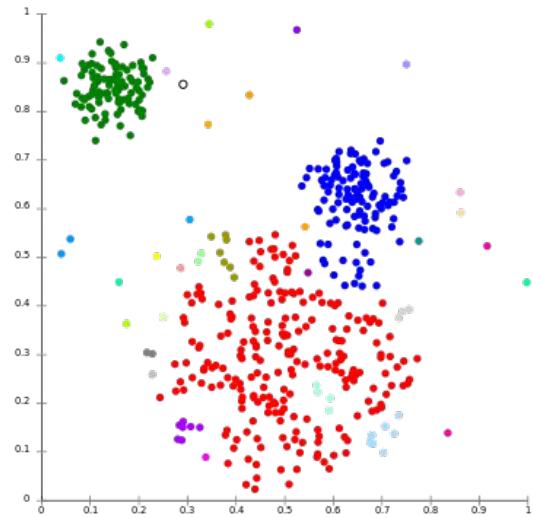
Definición de similitud varía según el algoritmo:

- Connectivy model (Ej. Hierarchical clustering)
- Centroid models (Ej. K-means algorithm)
- Density models (Ej. DBSCAN)
- Distribution models (Ej. Expectation-maximization)

Elección del algoritmo se basa en el problema a resolver, los datos y prueba/error.

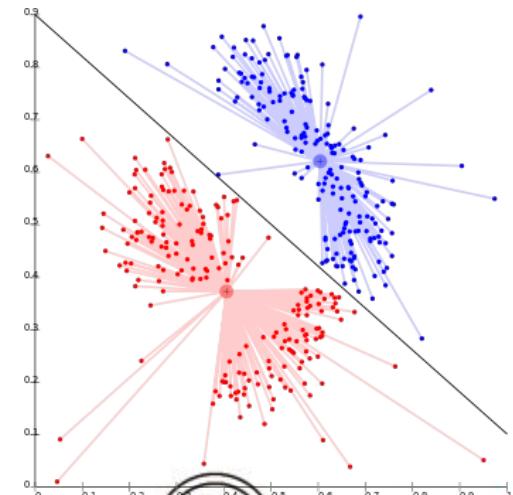
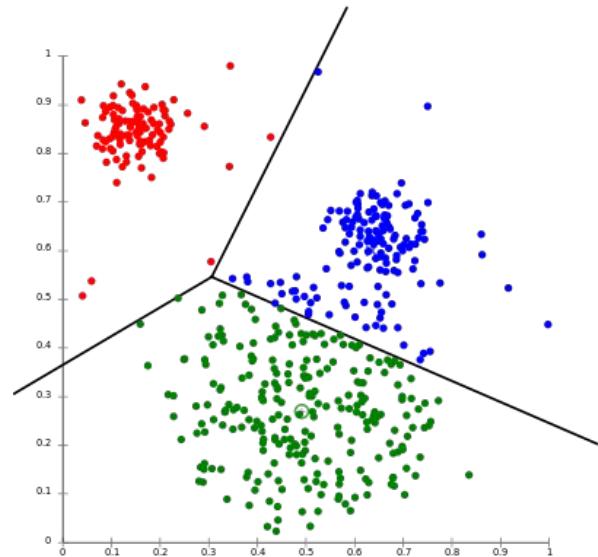
Hierarchical Clustering

- Elementos están más relacionados a elementos cercanos que elementos lejanos.
- Algoritmo conecta elementos basado en la distancia entre sí para formar clusters.
- Un cluster puede ser descripto por la máxima distancia requerida para conectar dos elementos.



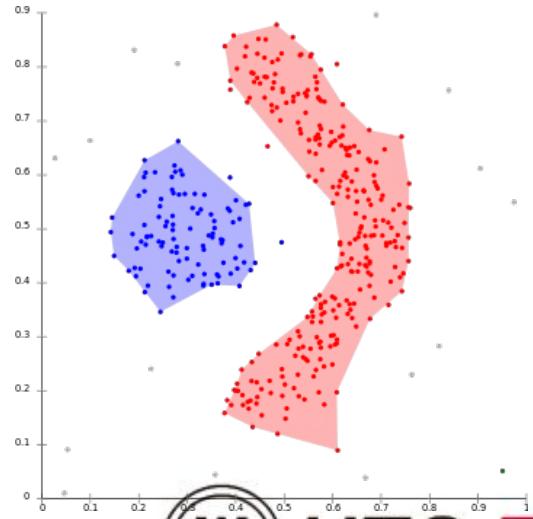
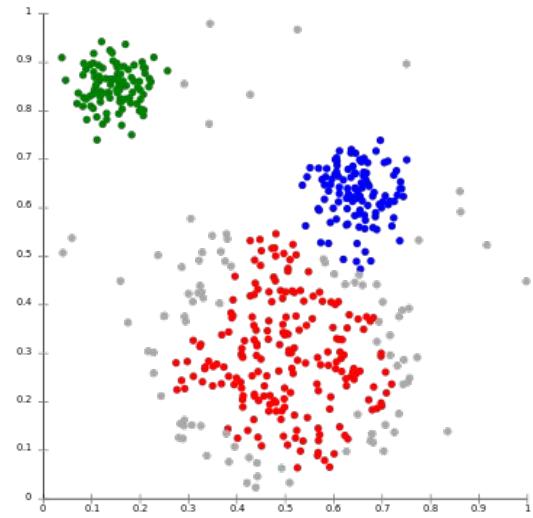
Centroid-based Clustering

- Cluster es representado por un vector central.
- Alto costo computacional.
- Suele requerir fijar k .
- Buscar los centros de los k clusters, asignándoles los objetos cercanos de forma que la distancia cuadrada al centro del cluster se minimice.



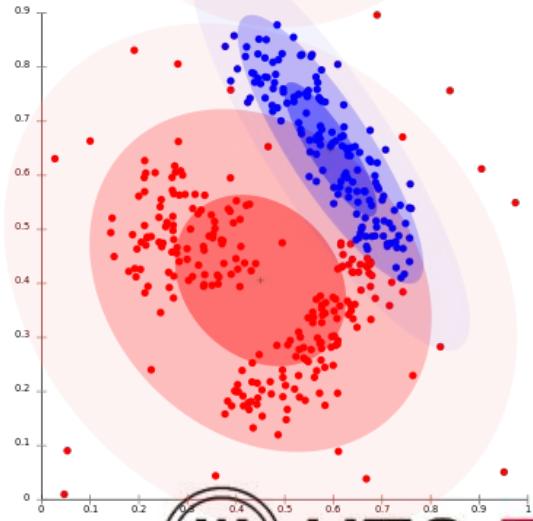
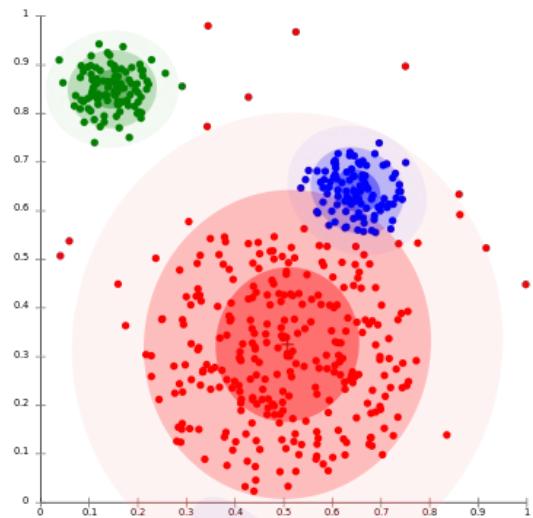
Density-based Clustering

- Puede generar clusters de contornos arbitrarios.
- Bajo costo computacional.
- Un cluster se define como un área de alta densidad comparado al resto del data set.
- Elementos en área de baja densidad son considerados como ruido o casos excepciones (outliners).
- Requiere una caída de densidad para detectar las fronteras del clusters.



Distribution-based Clustering

- Basado en distribuciones estadísticas.
- Susceptibles a overfitting. (Requieren limitar la complejidad del modelo).
- Un cluster puede ser descripto como elementos que probablemente tengan la misma distribución estadística.



Preguntas

MY HOBBY: EXTRAPOLATING

