Big Data Analytics

Profesor: José M. Luna

KDIS Lab

UNIVERSIDAD DE CÓRDOBA

Contenido

- **Introducción a BDA**
- Tipos de datos de entrada
- Tipos de BDA
- Empleo
- Casos de estudio

The Economist

MAY 6TH–12TH 2017

Theresa May v Brussels

Ten years on: banking after the crisis

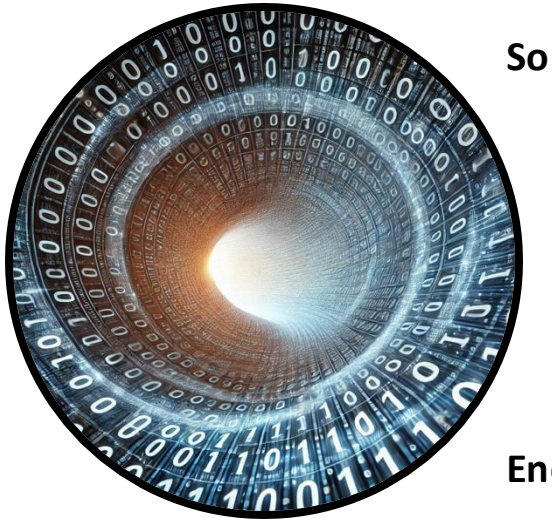
South Korea's unfinished revolution

Biology, but without the cells

The world's most valuable resource

Data and the new rules of competition

Introducción a BDA



No son finitos

Son reutilizables

Bajo coste (transporte y almacenamiento)

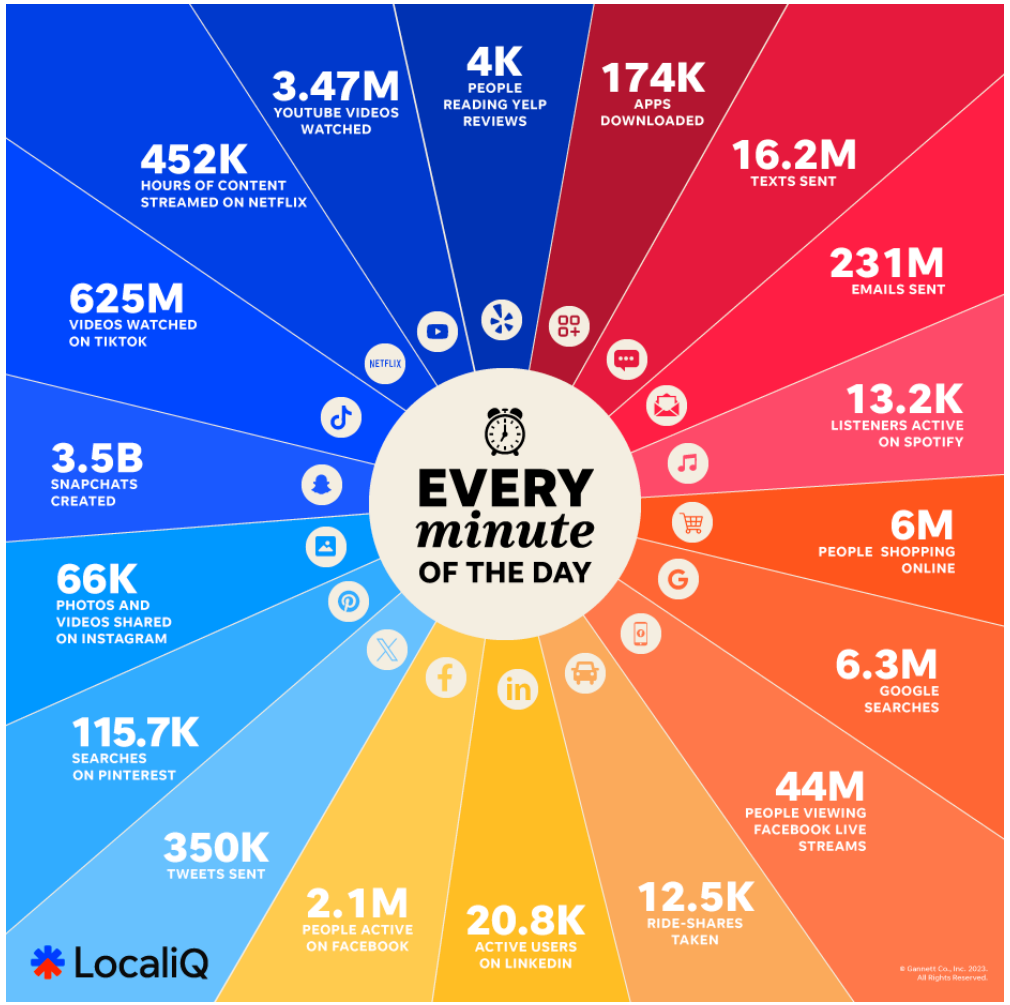
No requieren transporte

Si se procesan, aparecen nuevas aplicaciones

Enorme variabilidad

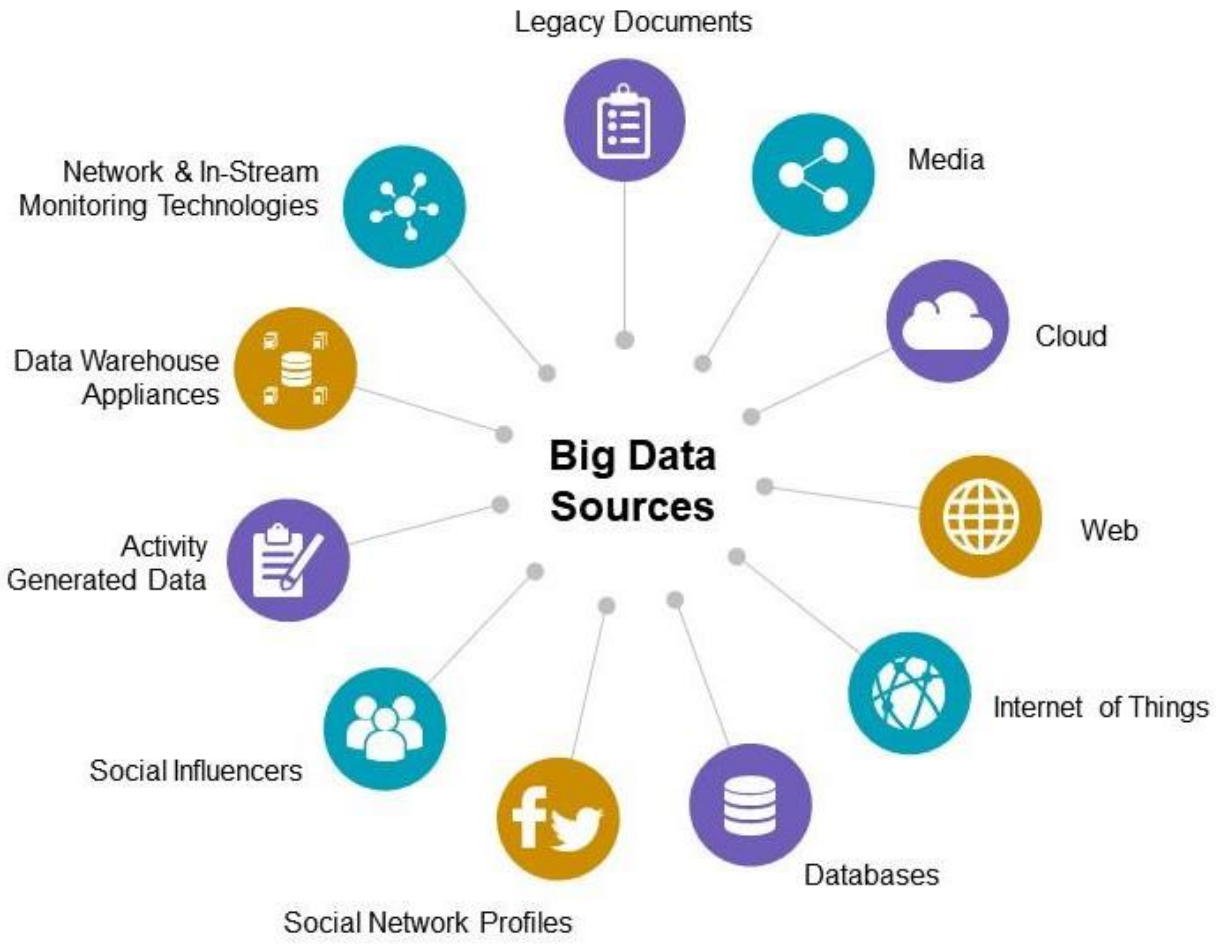
DATOS

Introducción a BDA



José María Luna

Introducción a BDA



José María Luna

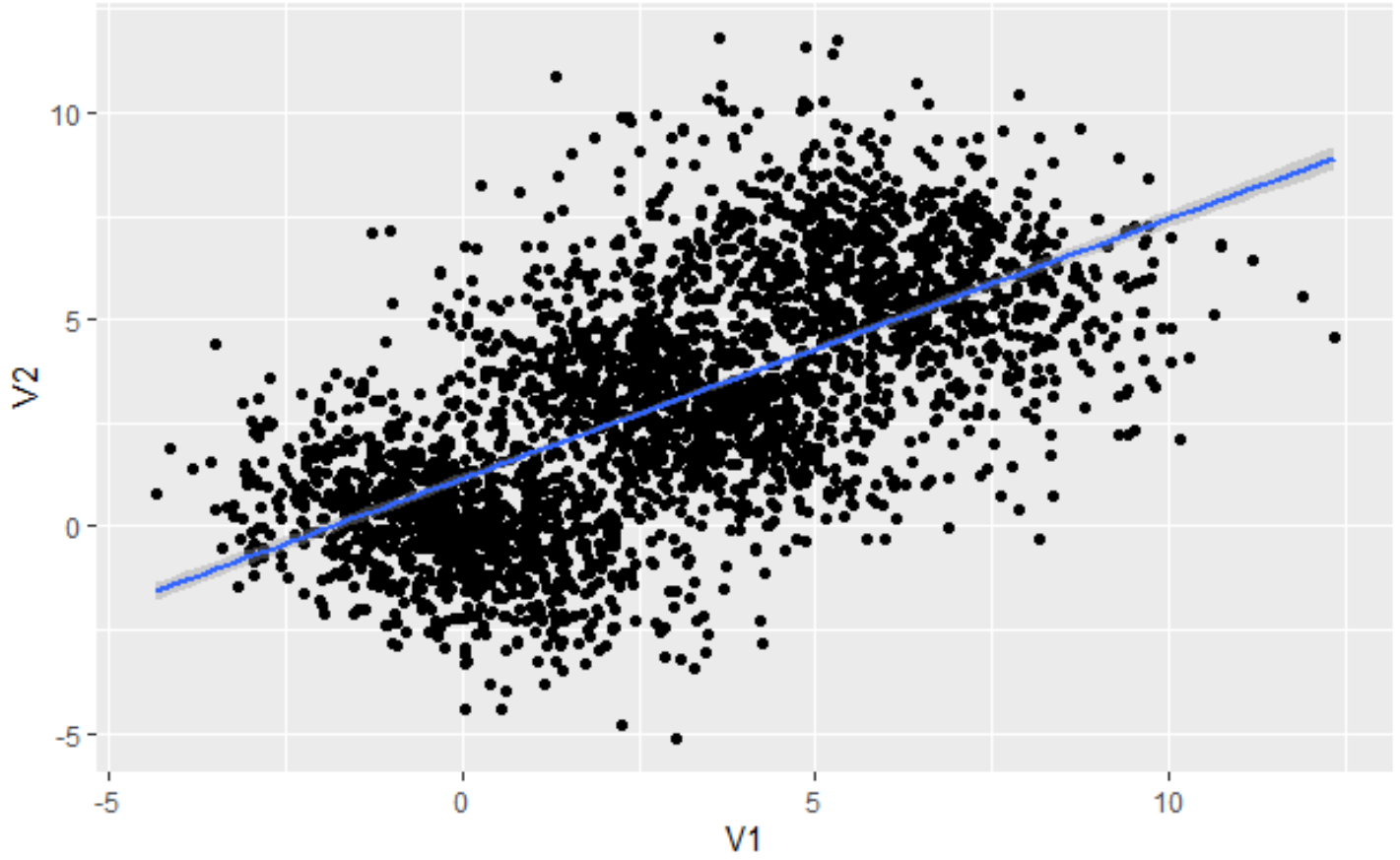
Introducción a BDA

KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Introducción a BDA

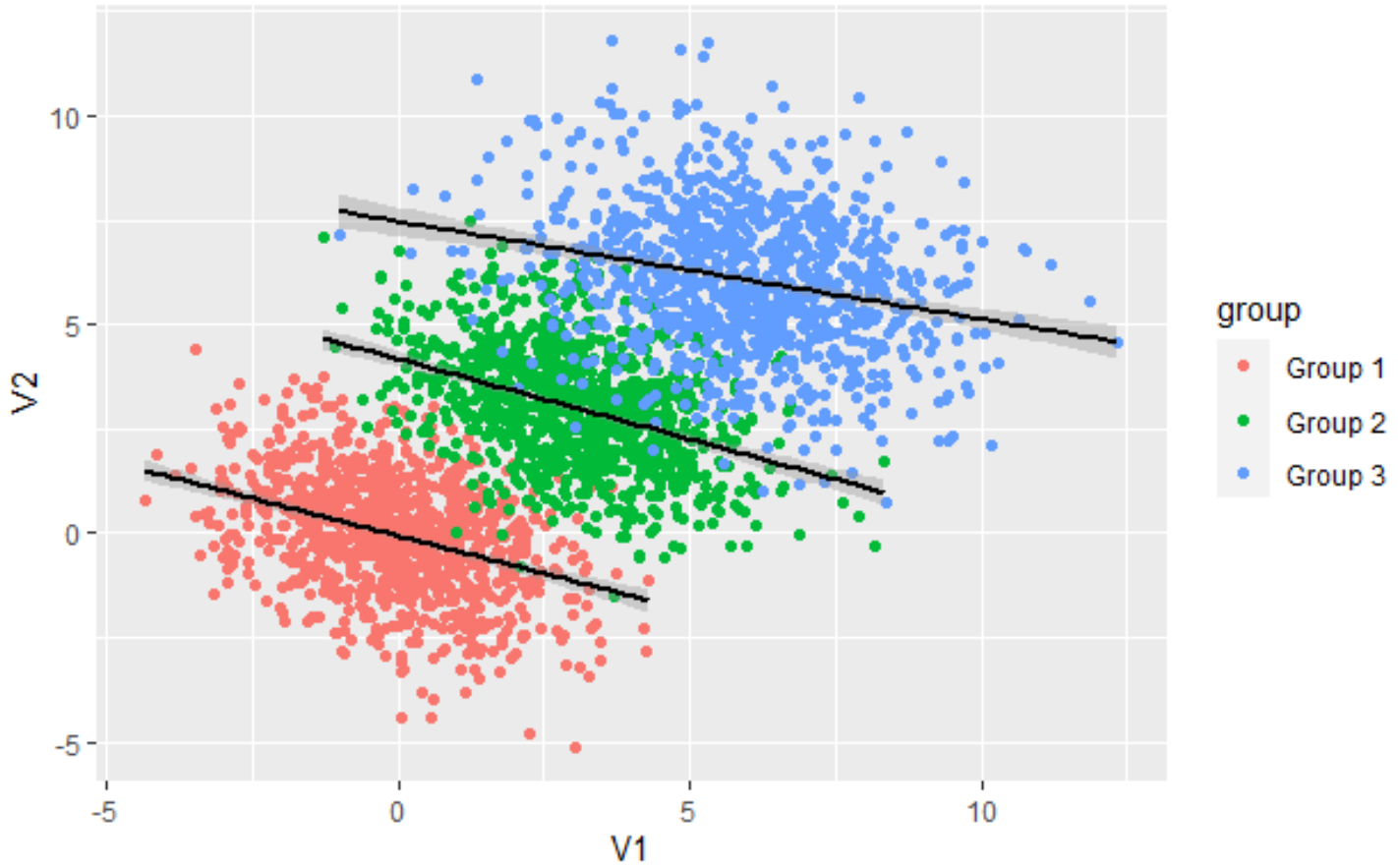


KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Introducción a BDA



Paradoja Simpson

La tendencia que aparece en varios grupos de datos desaparece cuando estos grupos se combinan y en su lugar aparece la tendencia contraria para los datos agregados

KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

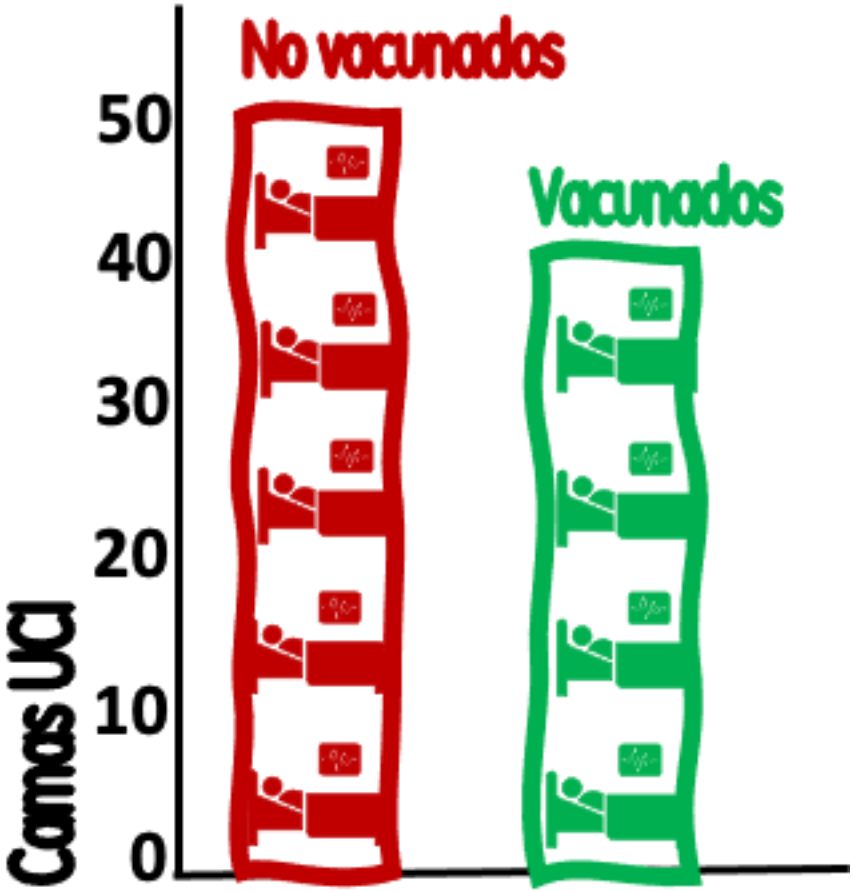
Introducción a BDA

Caso real

Efectividad vacunas COVID19

El 44% de los pacientes COVID ingresados en UCI estaban vacunados

El 56% de los de los pacientes COVID ingresados en UCI no estaban vacunados



KDIS Lab

José María Luna

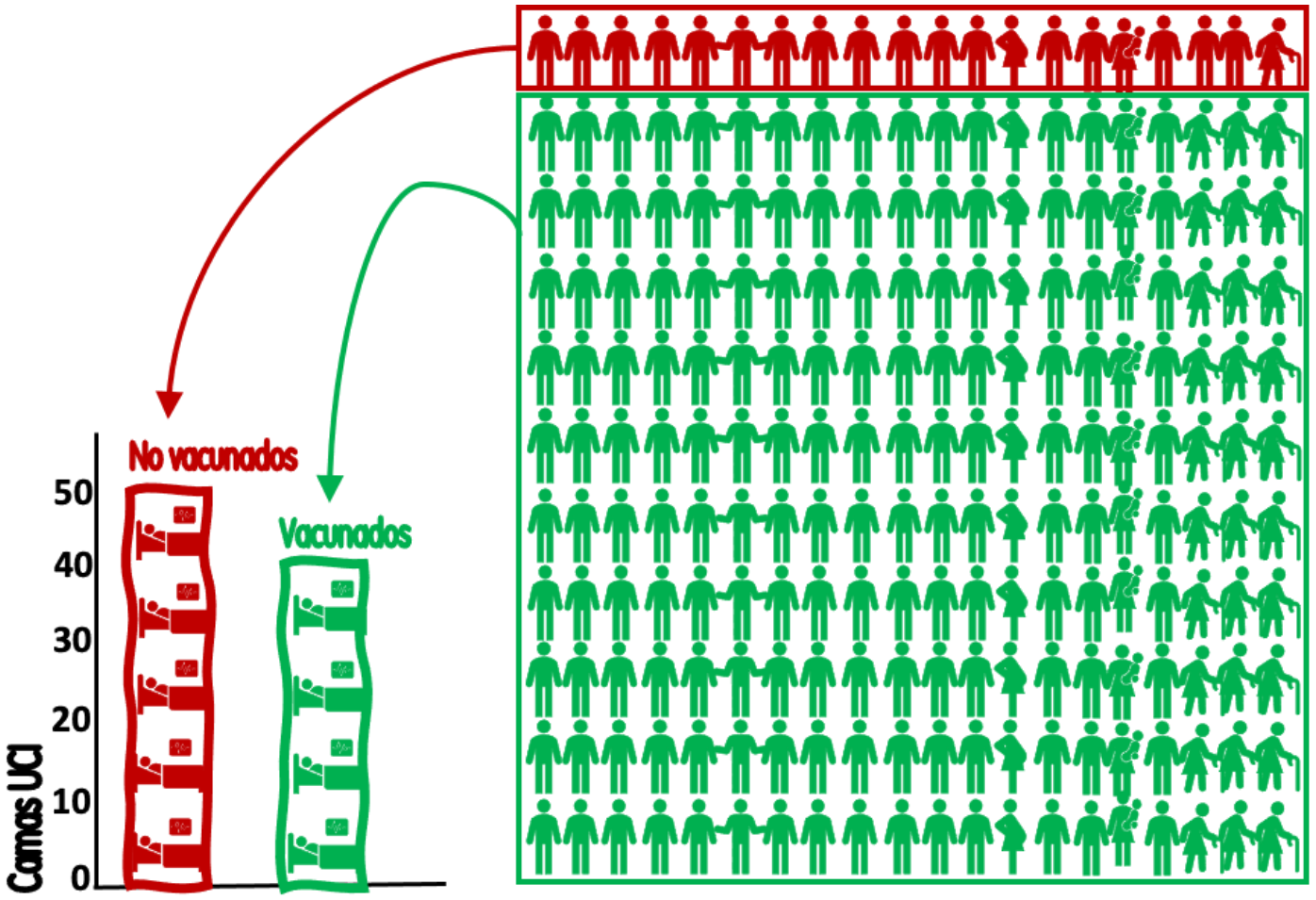
UNIVERSIDAD DE CÓRDOBA

Introducción a BDA

Añadimos más información

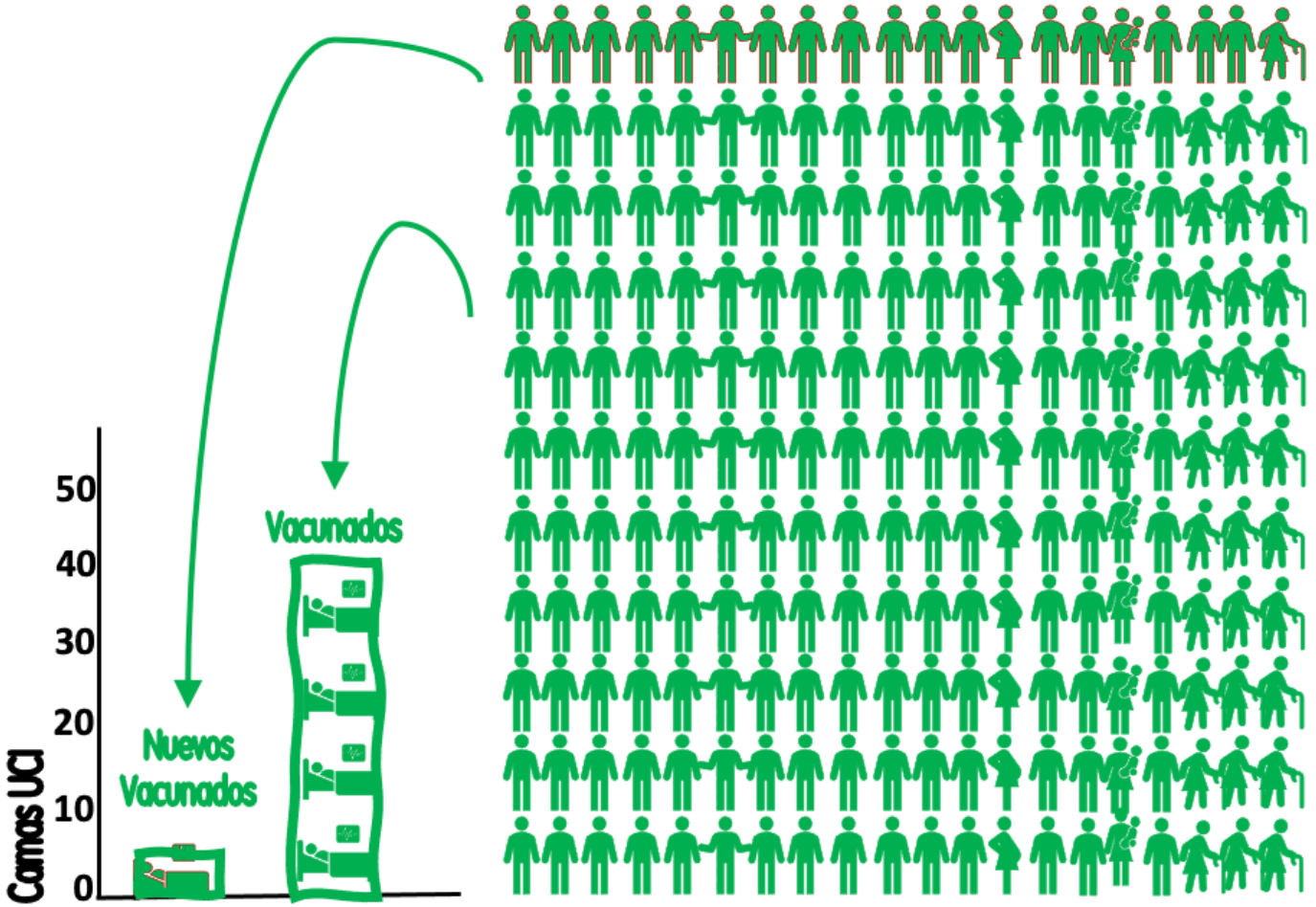
- 90 ingresos por COVID semanales, 40 vacunados (44,4 %) y 50 en no vacunados (55,6 %).
- Población de la región es de 5,5 millones de personas >12 años.
- La tasa total de ingresados en UCI sería de 1,6 ingresados semanales por 100.000 personas > 12 años.
- El 91 % de la población >12 años (5 millones) está vacunada y 0,5 millones no lo está.
- Los 500.000 no vacunados han generado 50 ingresos/semana en UCI, con una tasa de (50*100.000/500.000) **10 ingresos por 100.000 no vacunados a la semana**.
- Los 5 millones de vacunados han generado 40 ingresos, con una tasa de (40*100.000/5.000.000) de **0,8 ingresos en UCI por 100.000 vacunados**.

Introducción a BDA



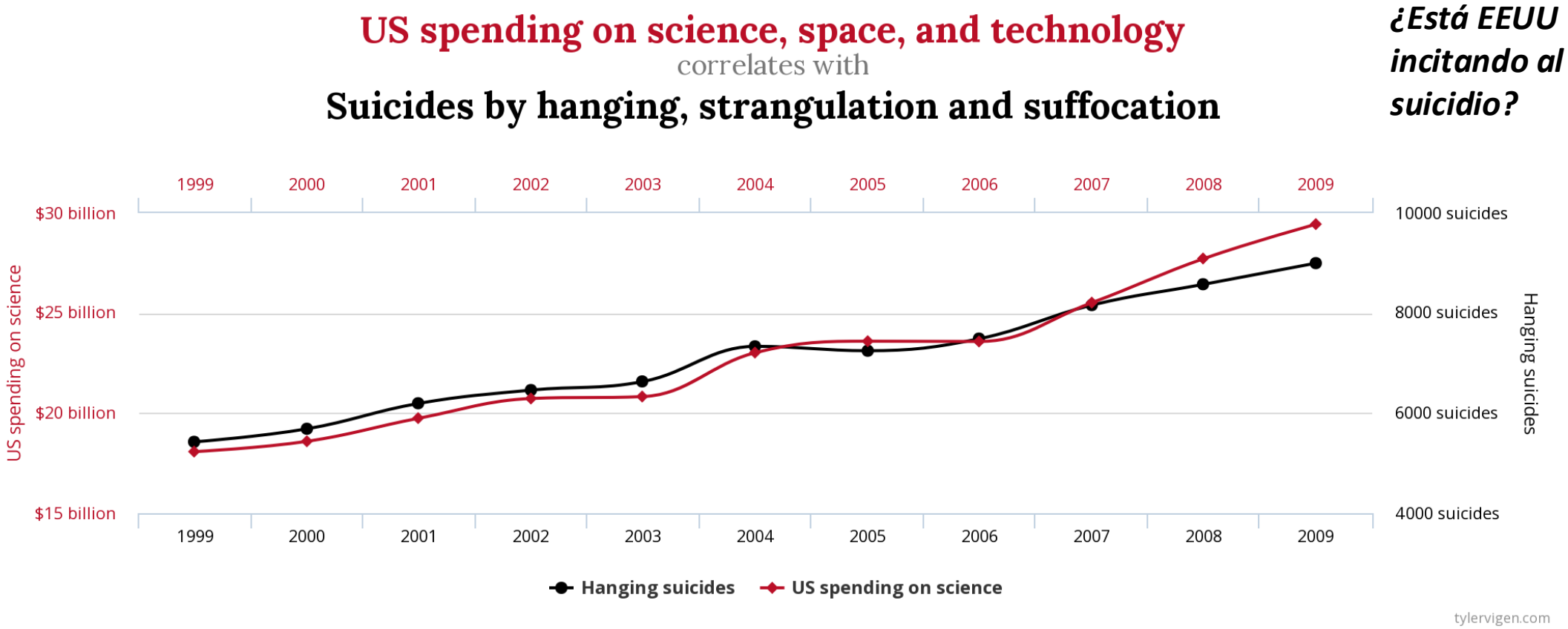
José María Luna

Introducción a BDA



José María Luna

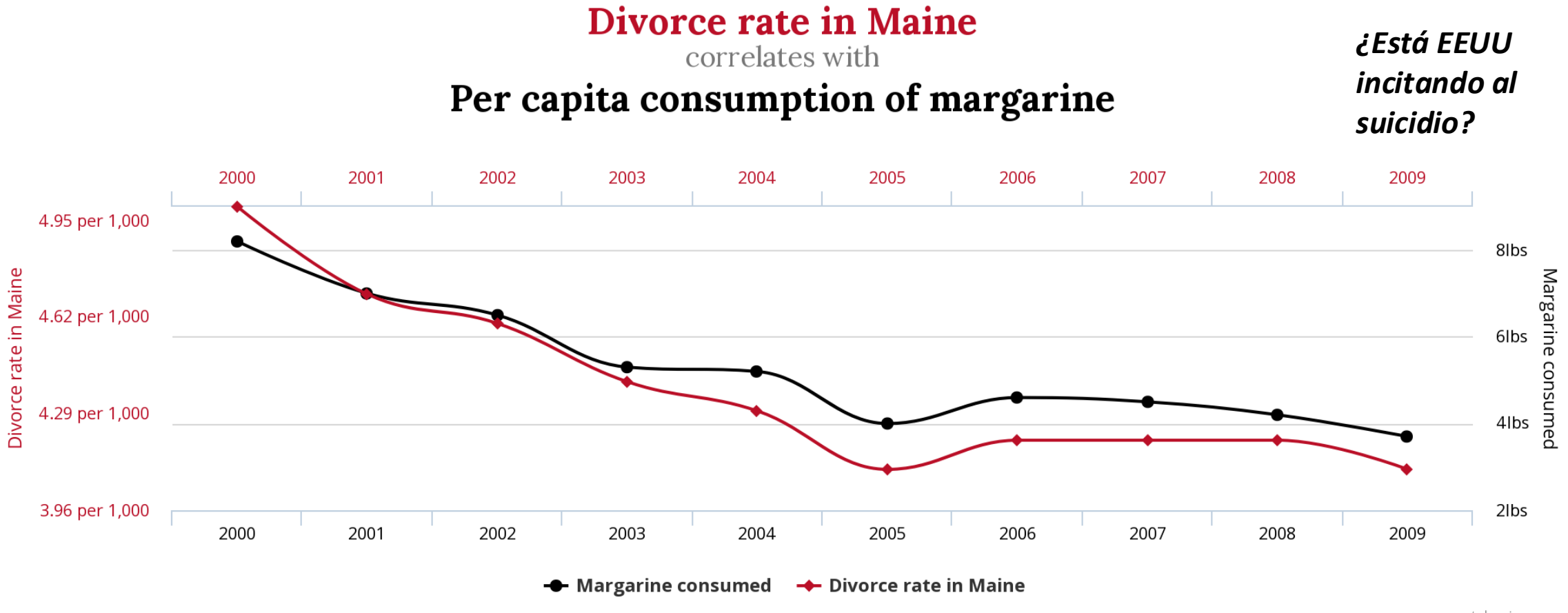
Introducción a BDA



¿Está EEUU incitando al suicidio?

José María Luna

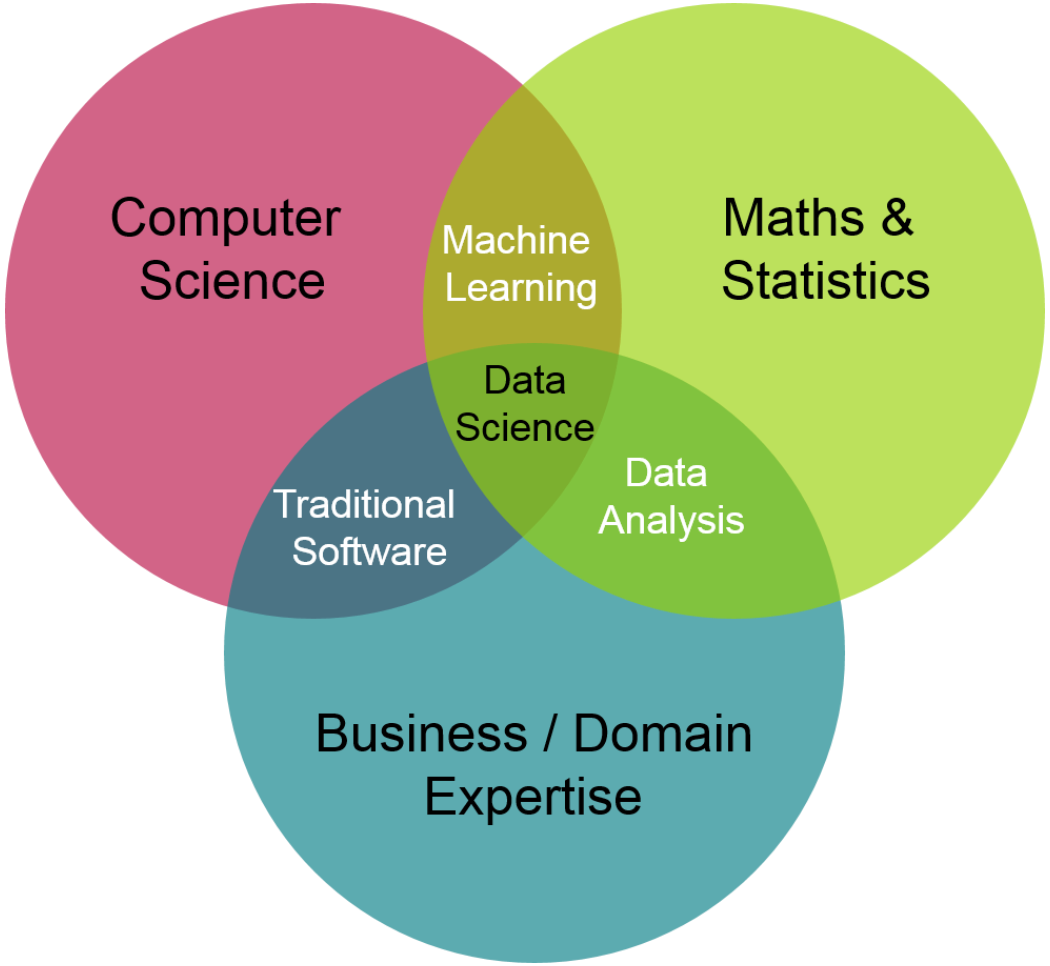
Introducción a BDA



¿Está EEUU incitando al suicidio?

José María Luna

Introducción a BDA



José María Luna

Introducción a BDA

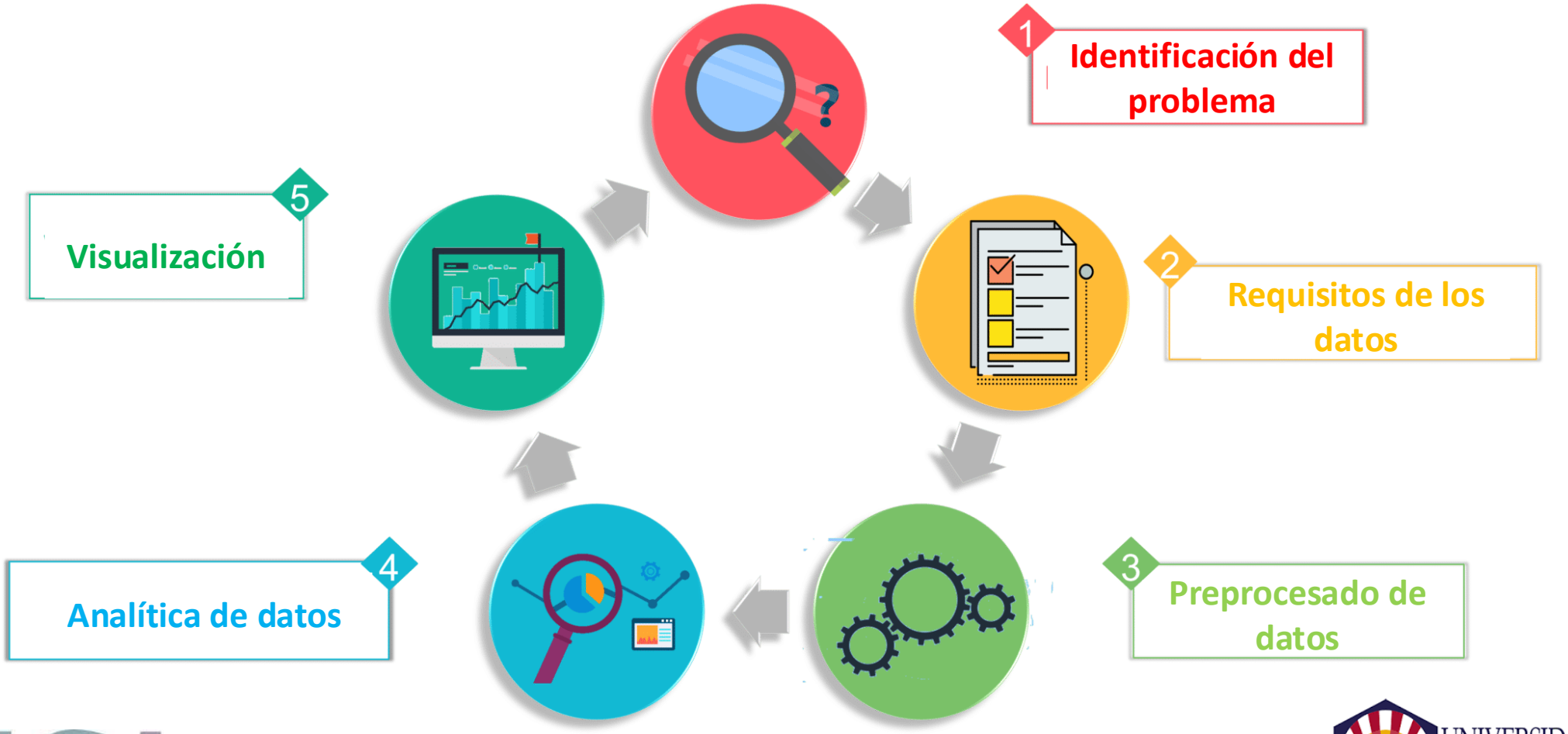
Big Data Analytics examina enormes cantidades de datos y de diferentes tipos para descubrir patrones ocultos, correlaciones y así como cualquier otra tipo de conocimiento

KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Introducción a BDA



José María Luna

Introducción a BDA

Capacidad de procesamiento

- 1974 Intel 8080 IPS 640,000
- 2005 AMD Athlon 64 IPS 84,000,000,000
- 2020 AMD Ryzen 3990x IPS 2,356,230,000,000

100 Km/h

37,000,000 Km/h

KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Introducción a BDA

Una vuelta a la Tierra 40,000 Km

925 vueltas cada hora

15 vueltas por minuto

KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Introducción a BDA

APACHE PIG

kafka

splunk>

Big Data Analytics

APACHE Spark™

Spark MLlib
The Machine Learning Library

+ableau

KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Contenido

- Introducción a BDA
- **Tipos de datos de entrada**
- Tipos de BDA
- Empleo
- Casos de estudio

Tipos de datos de entrada

Datos en cualquier organización:

- **Datos conocidos**
  - Utilizados: usados con fines analíticos o cualquier otro propósito que aporte valor a la organización
  - No utilizados: no son empleados, bien por falta de tiempo, presupuesto, o por desconocimiento de cómo emplearlos
  - Desorganizados: relacionado con Big Data. Su uso queda relegado a un trabajo futuro
- **Datos desconocidos**. Suelen ser datos no estructurados.

Tipos de datos de entrada

KDIS Lab

José María Luna

UNIVERSIDAD D CÓRDOBA

Tipos de datos de entrada

Ejemplos de Dark Data:

- Datos de Recursos Humanos
- Datos extraídos en encuestas y estudios que nunca fueron procesados
- Notas o presentaciones antiguas
- Versiones obsoletas de documentos actualizados
- Correos electrónicos
- Cambios registrados en negociaciones de pedidos
- Vídeos, imágenes y grabaciones
- Registros de actividad

Tipos de datos de entrada

Tres ejemplos de información que las empresas almacenan y no usan que pueden aportar datos muy interesantes:

- Los **archivos de registro** del servidor pueden hablar de un **comportamiento escondido que tienen los usuarios** de un sitio web

- Los **registros de llamadas comerciales** que se hayan generado **pueden revelar sentimientos y opiniones de los clientes** que no se reflejan en las encuestas de calidad de servicio al cliente

- Si se conservan **datos de ubicación móvil**, pueden dar muchas pistas sobre **patrones de consumo de los clientes**

Contenido

- Introducción a BDA
- Tipos de datos de entrada
- **Tipos de BDA**
- Empleo
- Casos de estudio

Tipos de BDA

Análisis descriptivo

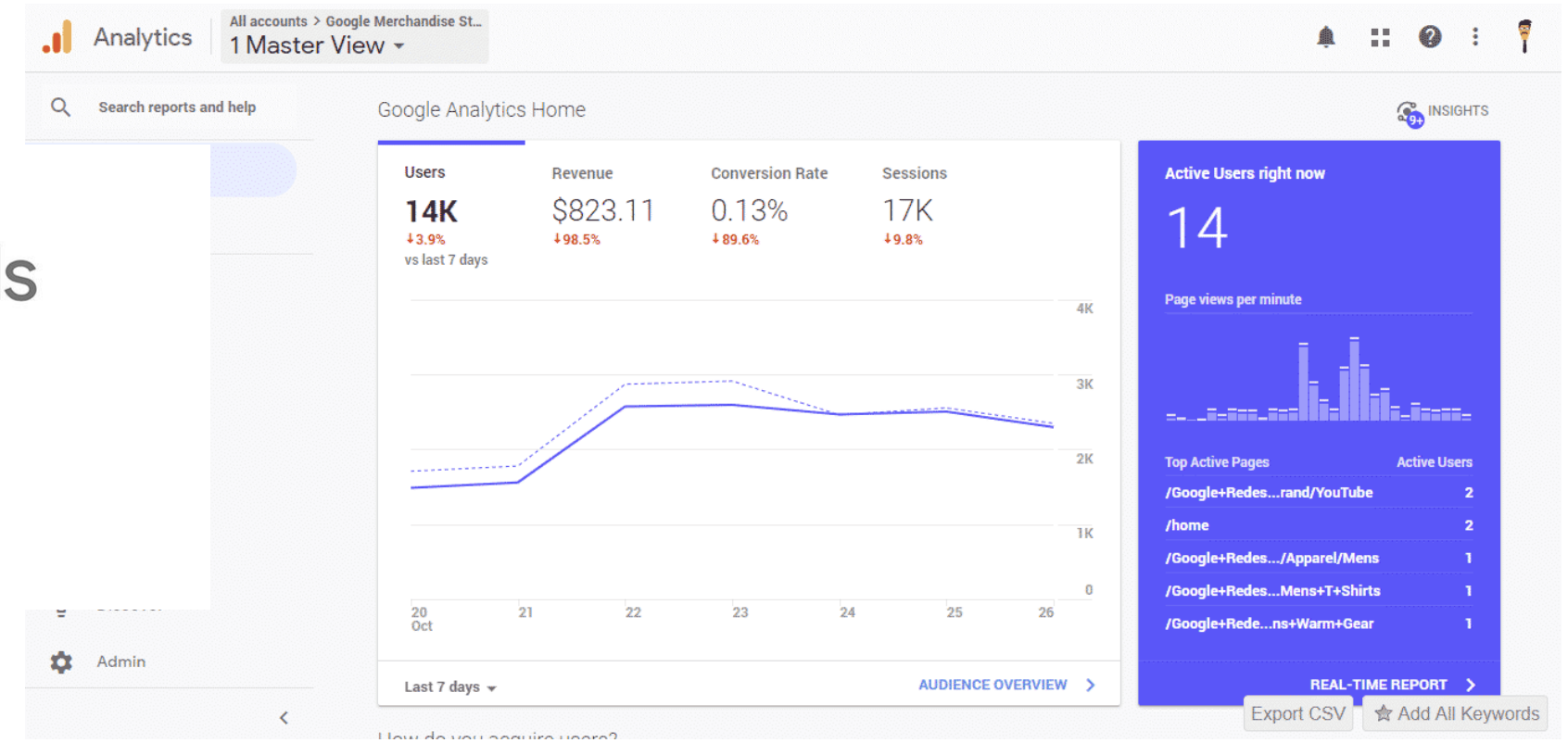
Análisis predictivo

Tipos de BDA

- Análisis descriptivo
  - Qué está ocurriendo en mis datos
- Análisis predictivo
  - Qué ocurrirá en mis datos
- Análisis prescriptivo
  - Qué acción debe ser tomada
- Analítica de diagnóstico
  - Por qué ocurrió algo

Tipos de BDA

- Análisis descriptivo



KDIS Lab

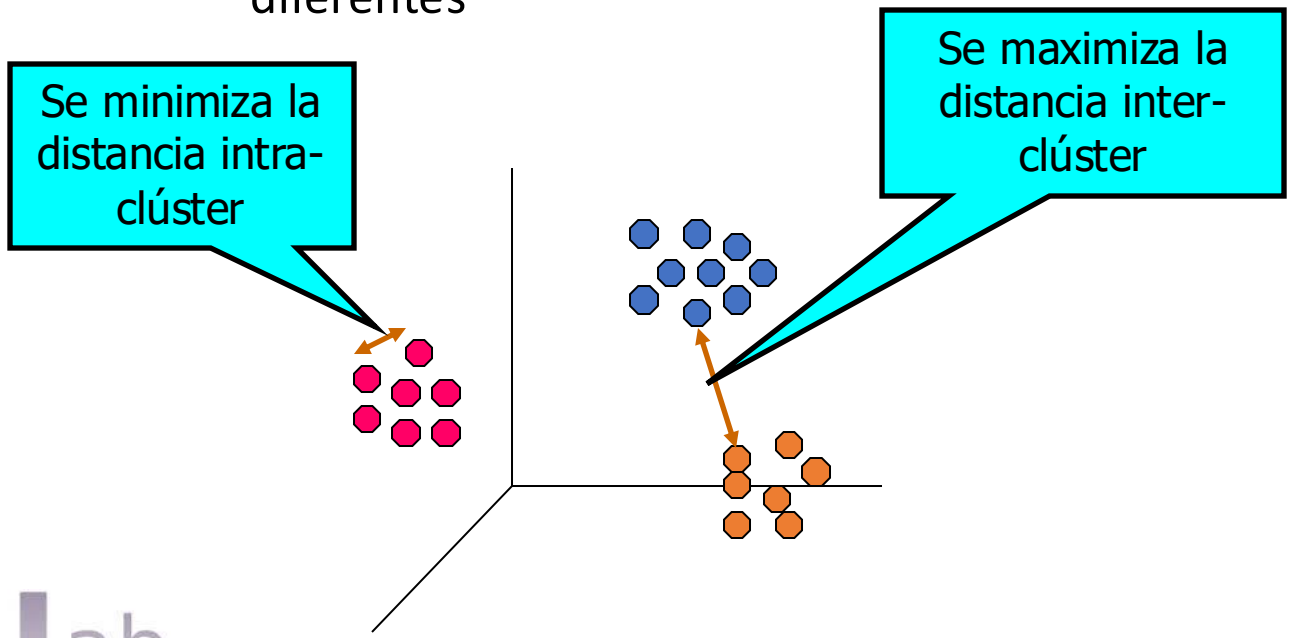
José María Luna

UNIVERSIDAD DE CÓRDOBA

Tipos de BDA

- Análisis descriptivo
  - *Clustering*

Dividir los datos en grupos (clústers) de tal forma que los datos que pertenecen al mismo clúster son similares, y datos que pertenecen a diferentes clústers son diferentes



Tipos de BDA

- Análisis descriptivo
  - *Clustering*

Ambigüedad en *clustering*

¿Cuántos clústers?

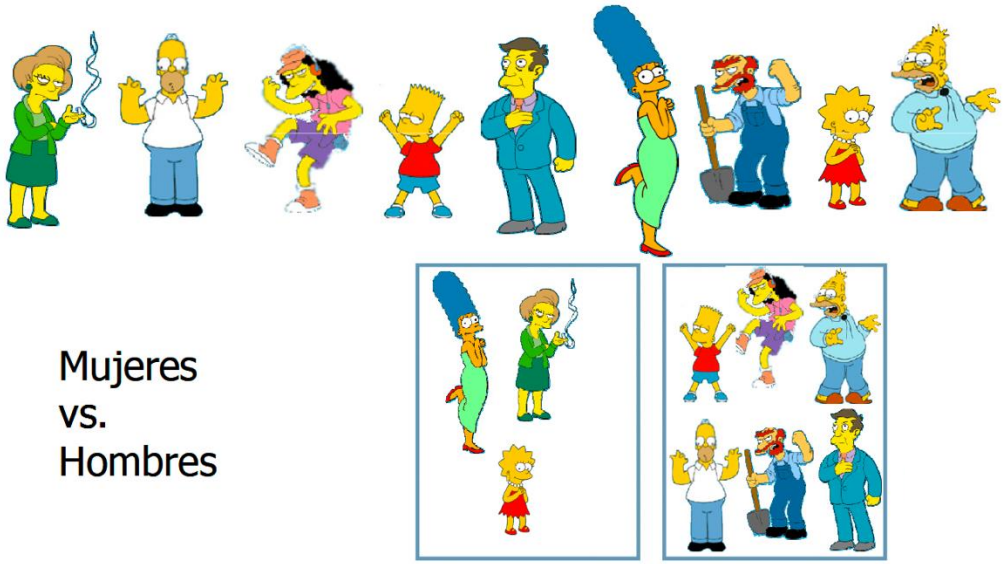
Dos clústers

Cuatro clústers

Seis clústers

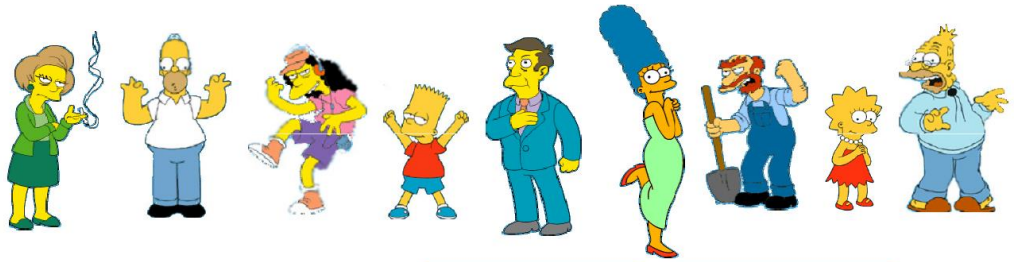
Tipos de BDA

- Análisis descriptivo
  - *Clustering* ¿Cuál es la forma natural de agrupar los personajes?

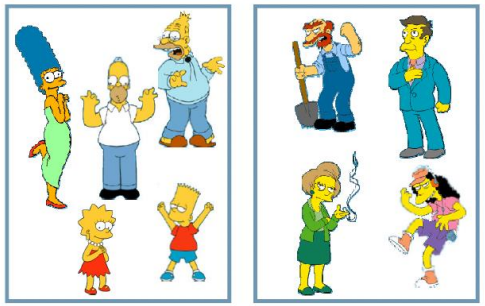


Tipos de BDA

- Análisis descriptivo
  - *Clustering* ¿Cuál es la forma natural de agrupar los personajes?



Simpsons
vs.
Empleados de
la escuela de
Springfield



Tipos de BDA

- Análisis descriptivo
  - *Clustering* ¿Cuál es la forma natural de agrupar los personajes?

El *clustering* es subjetivo!!!

Tipos de BDA

- Análisis descriptivo
  - *Clustering*

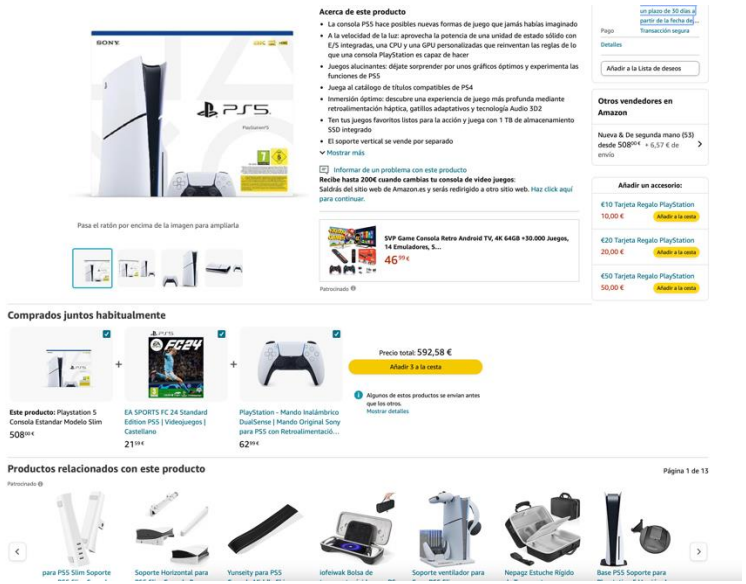
Usemos la MLlib…..

Documentación: *https://spark.apache.org/docs/latest/ml-clustering.html*

EJEMPLO en clase de k-means

Tipos de BDA

- Análisis descriptivo
  - Reglas de asociación



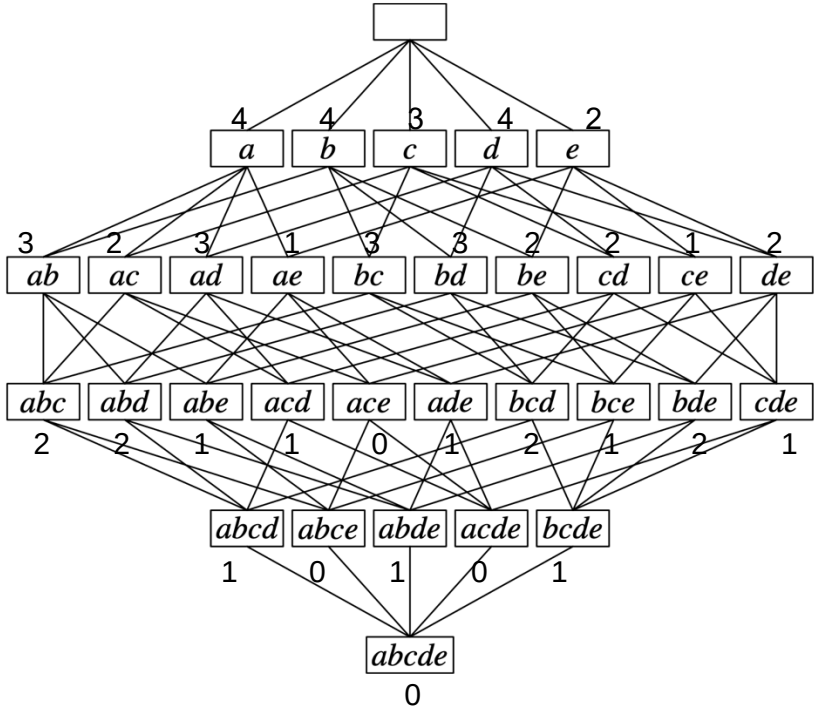
José María Luna

Tipos de BDA

- Análisis descriptivo
  - Reglas de asociación

a = Pan
b = Pañales
c = Cerveza
d = Leche
e = Coca-cola

TID	Items
1	Pan, Leche
2	Pan, Pañales, Cerveza
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola



Búsqueda de asociaciones

<Pan, Pañales> frecuencia 3
Si <Pan> entonces <Pañales> Exactitud 100%
Si <Pañales> entonces <Pan> Exactitud 75%

El orden de la implicación es muy importante

José María Luna

Tipos de BDA

- Análisis descriptivo
  - Reglas de asociación

Usemos la MLlib…..

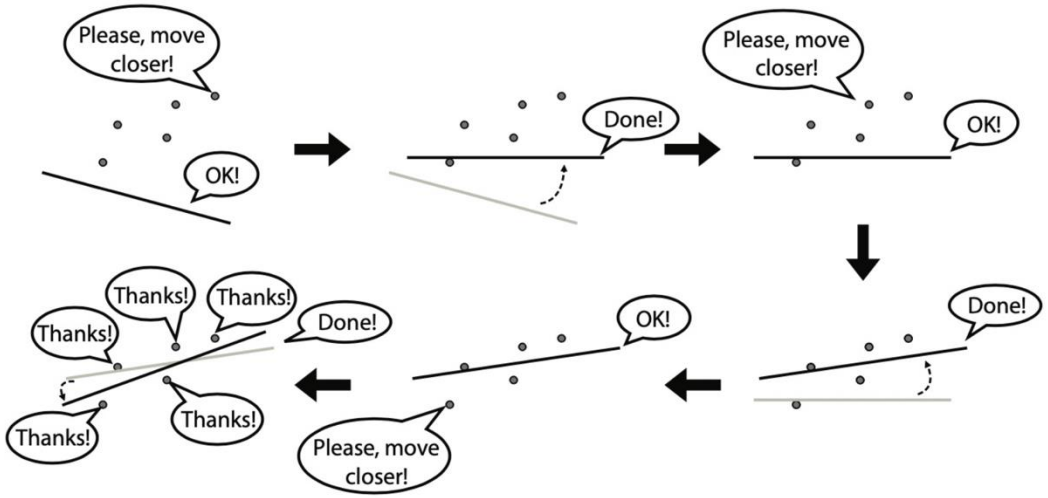
Documentación: *https://spark.apache.org/docs/latest/ml-frequent-pattern-mining.html*

EJEMPLO en clase de FP-Growth

Tipos de BDA

- Análisis predictivo
  - Regresión

"Generamos una línea al azar. Cogemos un punto al azar y movemos ligeramente la línea hacia este punto. Repetimos el proceso varias veces"



Tipos de BDA

- Análisis predictivo
  - Regresión

Usemos la MLlib…..

Documentación: https://spark.apache.org/docs/latest/ml-classification-regression.html

EJEMPLO en clase de Regresión Lineal

Tipos de BDA

- Análisis predictivo

Programa COINCIDENTE

El proyecto MANPREDIC

El proyecto MANPREDIC es un **proyecto de investigación y desarrollo tecnológico en el ámbito del mantenimiento predictivo de las plataformas terrestres del ejército de tierra**. El proyecto fue seleccionado como uno de los 10 proyectos de I+D interés para la Defensa a incluir en el **Programa de Cooperación en Investigación Científica y de Desarrollo de Tecnologías Estratégicas** (COINCIDENTE). El equipo integrante del proyecto, que está coordinado por el Prof. Sebastián Ventura

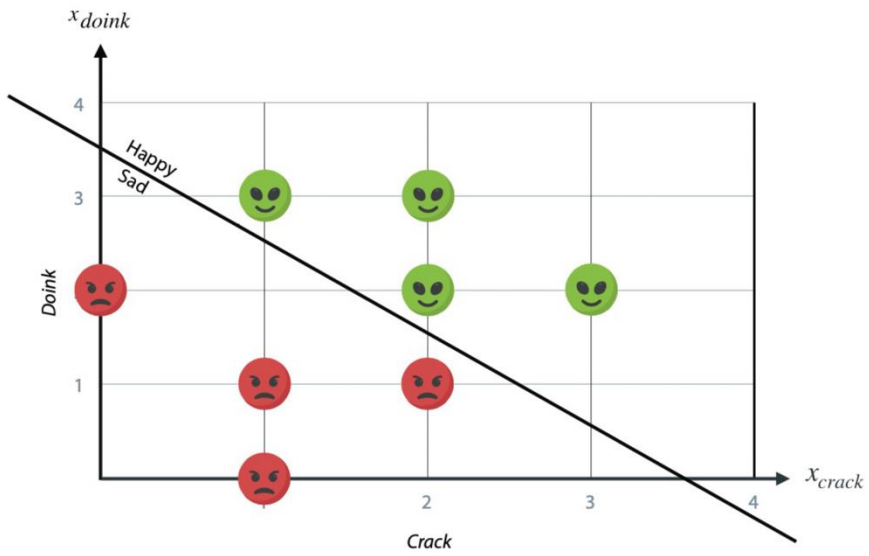
fuentes de datos (SIGLE, Información ITV, etc.). Para todas las plataformas objeto de estudio, se analizarán los datos que afectan a seguridad y vida del motor, huella logística y pautas de conducción.

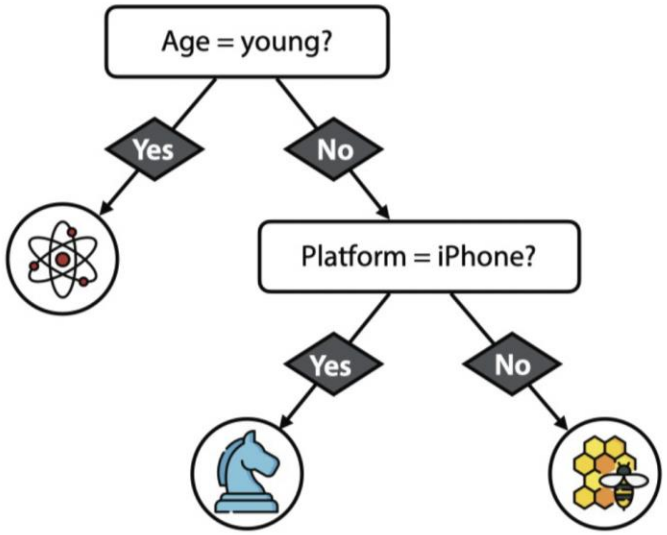
KDIS Lab

UNIVERSIDAD D CÓRDOBA

Tipos de BDA

- Análisis predictivo
  - Clasificación





Tipos de BDA

- Análisis predictivo
  - Regresión

Usemos la MLlib…..

Documentación: https://spark.apache.org/docs/latest/ml-classification-regression.html

EJEMPLO en clase de Árbol de decisión

Tipos de BDA

- Análisis prescriptivo

Tipos de BDA

- Analítica de diagnóstico

KDIS Lab

José María Luna

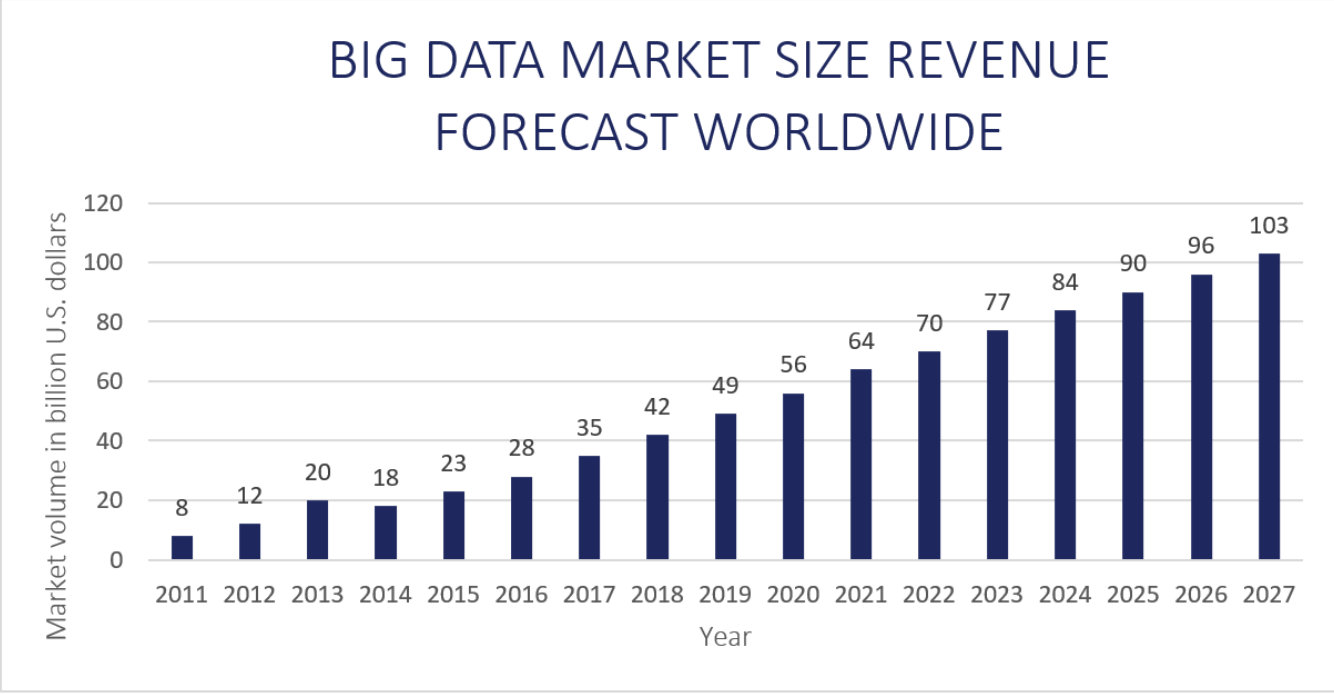
UNIVERSIDAD DE CÓRDOBA

Contenido

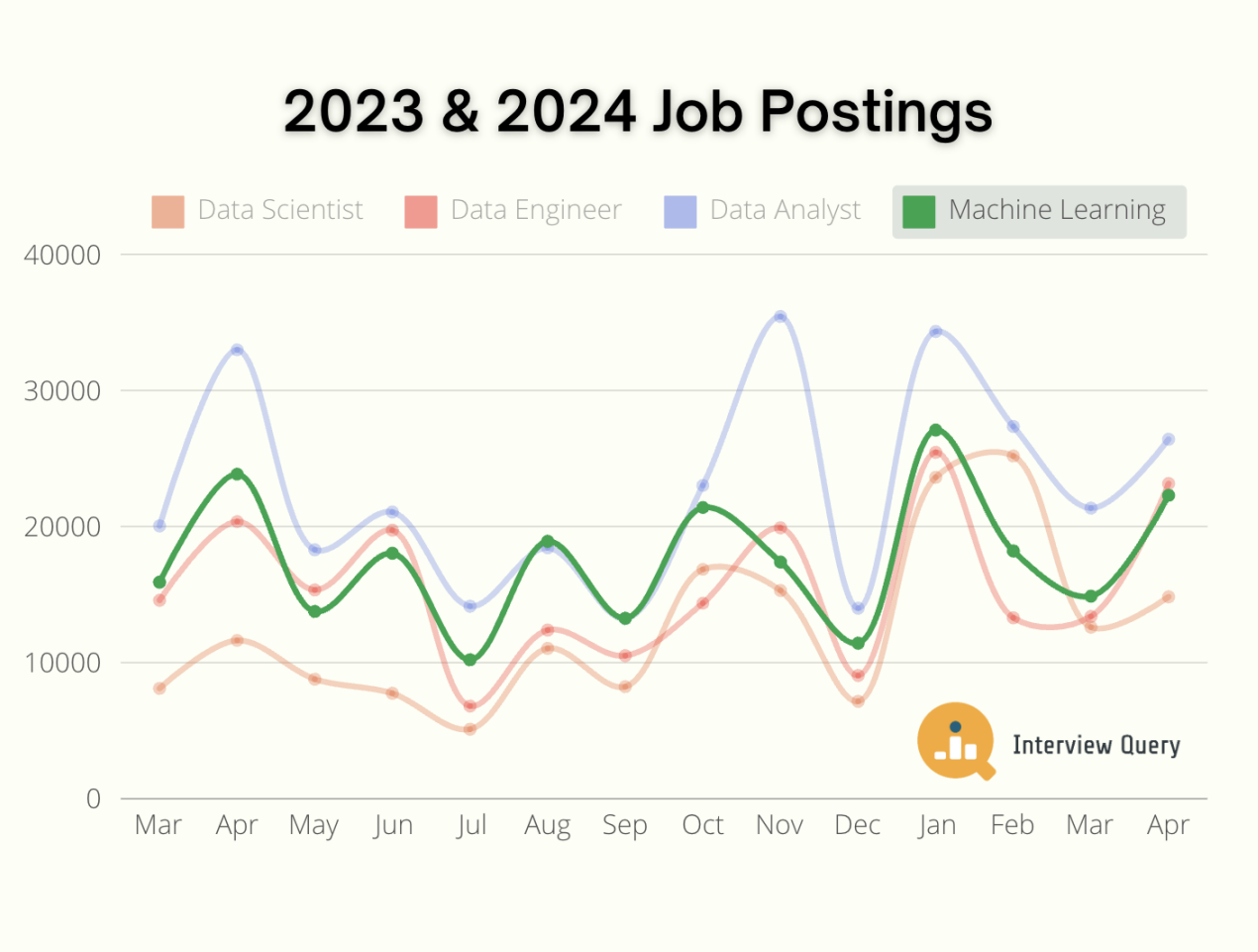
- Introducción a BDA
- Tipos de datos de entrada
- Tipos de BDA
- **Empleo**
- Casos de estudio

Empleo

- 96% de las empresas internacionales contratan personal con conocimientos en Big Data Analytics
- Es el tipo de empleo más demandado según el informe de *Monster Annual Trends*



Empleo



KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA

Empleo


Datamites
Global Institute for Data Science
Worldwide Data Analyst Salaries in 2024
UK
GBP 34,996
Canada
CAD 68,165
USA
USA 107,928
UAE
AED 73,430
India
INR 6,25,341
Germany
EUR 60,500
South Africa
ZAR 337,169
Saudi Arabia
SAR 215,884
(Data Source: Glassdoor, Indeed & Salary Expert)

Empleo

- Habilidades que tiene un *Big Data Scientists*
  - Programación
  - Estadística
  - Aprendizaje automático
  - *Data warehousing*: SQL, No SQL
  - Visualización de datos
  - Frameworks de computación: Spark, Hadoop, etc
  - Conocimiento de dominio de aplicación

Contenido

- Introducción a BDA
- Tipos de datos de entrada
- Tipos de BDA
- Empleo
- **Caso de estudio. NETFLIX**

Caso de estudio. NETFLIX

- Si estas viendo una serie, Netflix sabe el ratio en el que los usuarios terminarán esa serie
- Eventos analizados
  - Qué búsquedas realizan
  - Qué dispositivos usan
  - Cuál es su día preferido
  - Cuánto tiempo emplean y en cada uno de los contenidos
  - Si ven los capítulos enteros y qué fragmentos se ven más
  - Sus valoraciones
  - Preferencias de sus amigos o situación geográfica
  - etc

Caso de estudio. NETFLIX

If we can get each user to watch at least 15 hours of content each month, they are 75% less likely to cancel. If they drop below 5 hours, there is a 95% chance they will cancel

How do we help users watch at least 15 hours of content per month?

KDIS Lab

UNIVERSIDAD DE CÓRDOBA

Caso de estudio. NETFLIX

Éxito de *House of Cards*

- Hicieron 10 versiones del trailer dirigidas a diferentes audiencias, segmentadas en función de su comportamiento en la plataforma
  - Si has visto muchos programas de televisión protagonizados por mujeres, se te muestra un avance centrado en los personajes femeninos
- Se garantizan cumplir con el conocimiento que tenían: los subscriptores no cancelan porque consumen más de 15 horas

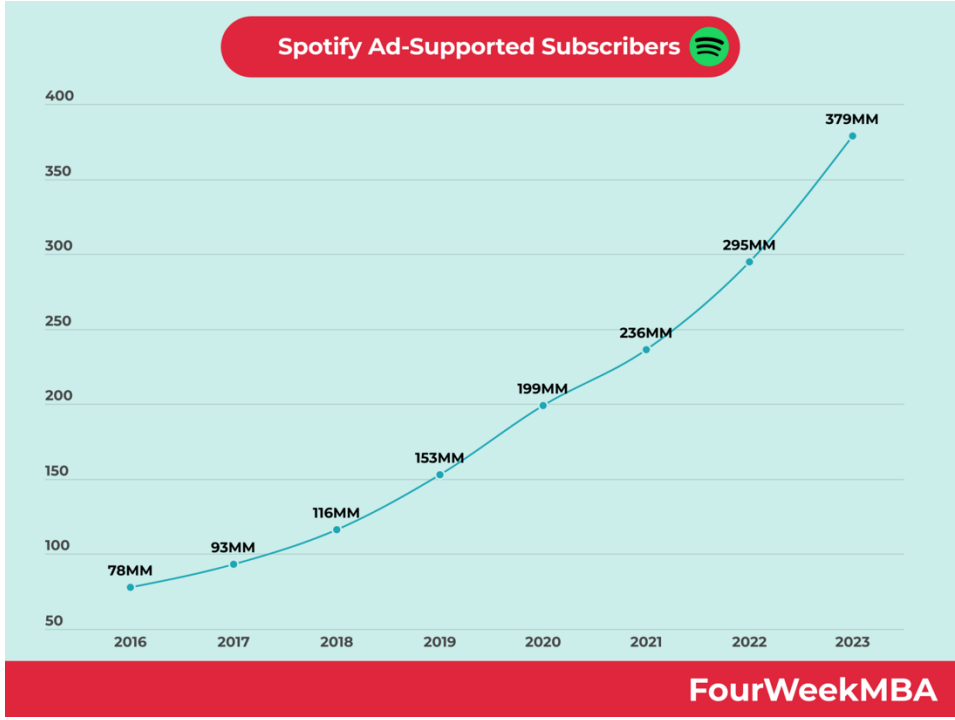
Contenido

- Introducción a BDA
- Tipos de datos de entrada
- Tipos de BDA
- Empleo
- **Caso de estudio. Spotify**

Caso de estudio. Spotify

- Análisis de tus gustos
  - Propone listas de gustos similares con el fin de que sigas escuchando

  - Prepara la lista que predice que vas a escuchar con el fin de que cargue más rápido



KDIS Lab

José María Luna

UNIVERSIDAD DE CÓRDOBA