Datos Masivos I

Profa. Alondra Berzunza

Alondra Vanianinetl Berzunza Rodríguez

Licenciada en Matemáticas Aplicadas y Computación

Diplomado en Estadística y Ciencia de Datos

Insaite - Científico de Datos

CitiBanamex - Científico de Datos

Aptude - Python Developer

alondraberzunza09@gmail.com

cleg@acatlan.unam

5556231500 ext 38839

Planta baja edificio 9

Atención psicológica y jurídica

Evaluación

alondraberzunza09@gmail.com

Actividad	Porcentaje
Exámenes Parciales	20 %
Exposiciones	10 %
Participaciones y tareas	10 %
Prácticas (OBLIGATORIAS)	20 %
Proyecto Final (OBLIGATORIO)	40 %
Total	**100 %**
Examen Final	100 %

Calendario

- Última clase 24 de mayo
- 26 de Mayo 1er final
- 2 de Junio 2do final

Las tareas se revisarán a la siguiente clase

Los exámenes se realizarán al término de cada unidad

Descanso

Las prácticas, exposiciones y el proyecto se realizarán en equipos*

- **24 de mayo entregas y exposiciones del proyecto final.**
- **24 de mayo examen general de conocimientos**

Unidad 1 – Conceptos Básicos

1.1 Definición y características

1.2 Generación, procedencia y preparación de datos

1.3 Consideraciones estadísticas y computacionales de los datos masivos

1.4 El principio de Bonferroni

1.5 Privacidad y riesgo

1.6 Modelos de computación para datos masivos

1.1 Definición y Características

1.1 Definición y Características

La definición de big data son datos que contienen una mayor **variedad** y que se presentan en **volúmenes crecientes** y a mayor **velocidad**. Esto se conoce también como "las tres V". - Oracle

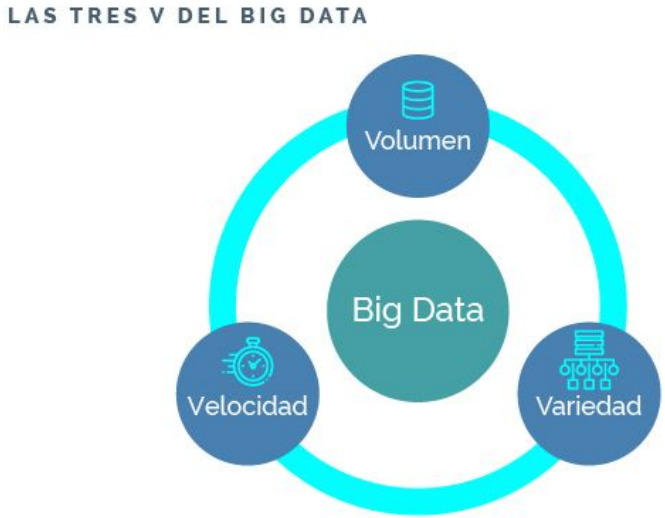
Captura	Gestión	Procesamiento	Análisis

1.1 Definición y Características

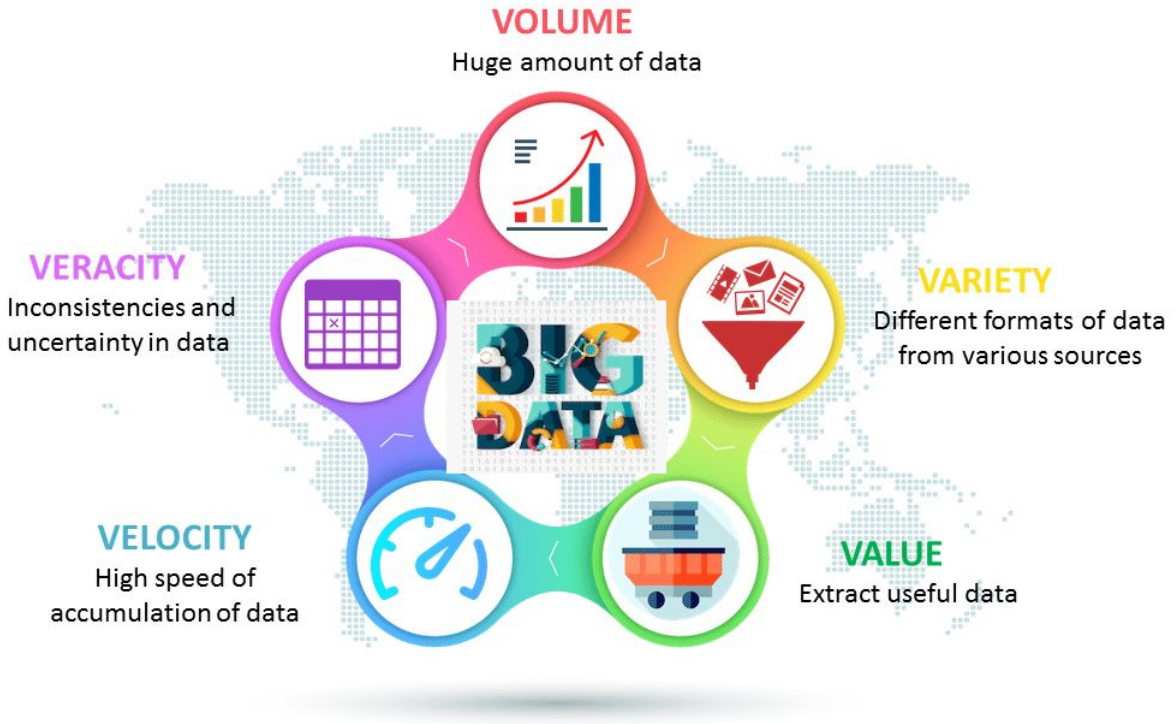
Las 3 V de Big Data

El acto de acceder y almacenar grandes cantidades de información para la analítica ha existido desde hace mucho tiempo.

Pero el concepto de big data cobró impulso a principios de la década de 2000 cuando el analista de la industria, Doug Laney, articuló la definición de grandes datos como las tres V.



1.1 Definición y Características



1.1 Definición y Características

Casos de Uso

- Desarrollo de productos
- Mantenimiento predictivo
- Experiencia del cliente
- Fraude y cumplimiento
- Machine learning / Deep learning
- Eficiencia operativa
- Impulse la innovación

Exposiciones

Equipo 1 - Plan de gobernanza de datos para Big Data y los posibles tipos de archivos con los que se puede generar big data.

Equipo 2 - Herramientas para trabajar Big Data

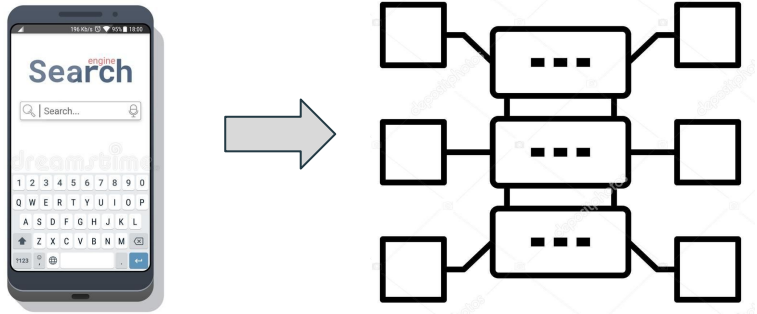
Equipo 3 - Arquitecturas/ modelos computacionales de datos

Equipo 4 - ¿Cuándo debe utilizarse el método de Bonferroni y cuándo no? Ejemplos.

1.2 Generación, procedencia y preparación de datos

El Big Data consiste en la recopilación, acumulación y búsqueda de patrones similares de información de toda clase.

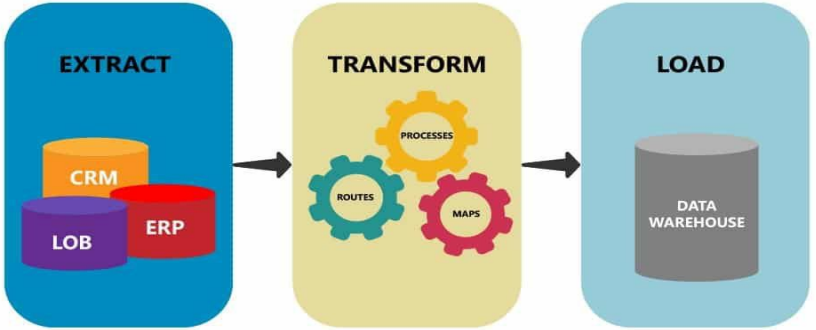
1. Obtención de la información:
    - fiable
    - de calidad
    - coherente



Lugar	Hora	Búsqueda	Resultado Consultado

1.2 Generación, procedencia y preparación de datos

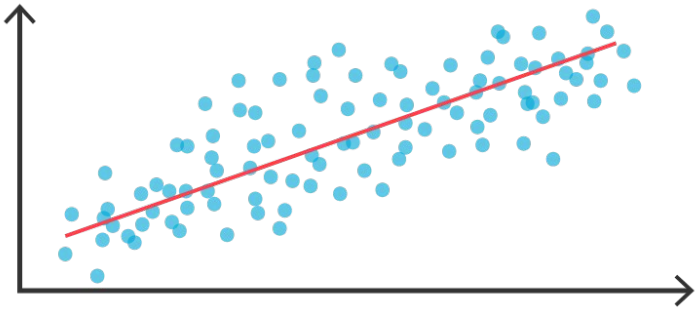
-
2. Procesamiento
   - transformar los datos recogidos para su posterior almacenamiento
3. Almacenamiento
   - grandes bases de datos.
4. Análisis



Lugar	Hora	Búsqueda	Resultado Consultado	Género / Sexo

F, M, Fem, Female, femenino, MUJER → fem

1.3 Consideraciones estadísticas y computacionales de los datos masivos

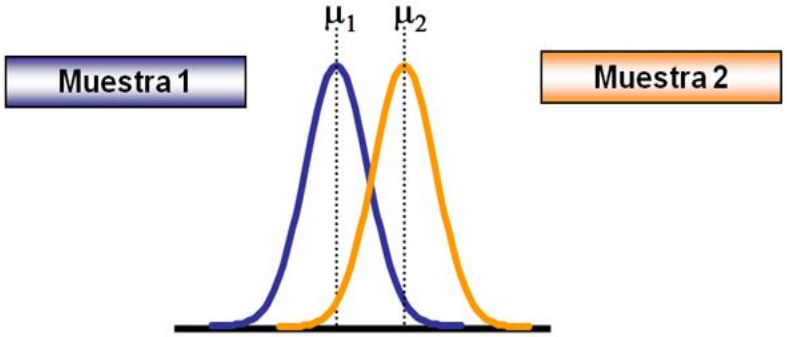


Análisis convencionales que causan problemas

- Media aritmética
- Desviación estándar

Algunas soluciones

- Muestreo
- Regresiones
- Pruebas de hipótesis



1.3 Consideraciones estadísticas y computacionales de los datos masivos

- Aumenta el peligro de ver patrones donde ninguno realmente existe.
- Los datos más grandes no siempre son mejores datos.
- Las estadísticas son cruciales para obtener conocimiento de conjuntos de datos cada vez más grandes.

Se requieren habilidades estadísticas serias para evitar errores como:

- Responder la pregunta incorrecta.
- Recolectar los datos incorrectos.
- Usar la técnica estadística incorrecta.
- Malinterpretar los resultados.

Otras consideraciones

En la Antigüedad, vivían seis hombres ciegos que pasaban las horas compitiendo entre ellos para ver quién era el más sabio. Exponían sus saberes y luego decidían entre todos quién era el más convincente.

Un día, discutiendo acerca de la forma exacta de un elefante, no conseguían ponerse de acuerdo. Como ninguno de ellos había tocado nunca uno, decidieron salir al día siguiente a la busca de un ejemplar, y así salir de dudas. Puestos en fila, con las manos en los hombros de quien les precedía, emprendieron la marcha enfilando la senda que se adentraba en la selva. Pronto se dieron cuenta que estaban al lado de un gran elefante. Llenos de alegría, los seis sabios ciegos se felicitaron por su suerte. Finalmente podrían resolver el dilema.

El más decidido, se abalanzó sobre el elefante con gran ilusión por tocarlo. Sin embargo, las prisas hicieron tropezar y caer de bruces contra el costado del animal. “El elefante –exclamó– es como una pared de barro secada al sol”.

El segundo avanzó con más precaución. Con las manos extendidas fue a dar con los colmillos. “¡Sin duda la forma de este animal es como la de una lanza!”.

Entonces avanzó el tercer ciego justo cuando el elefante se giró hacía él. El ciego agarró la trompa y la resiguió de arriba a abajo, notando su forma y movimiento. “Escuchad, este elefante es como una larga serpiente”.

Era el turno del cuarto sabio, que se acercó por detrás y recibió un suave golpe con la cola del animal, que se movía para asustar a los insectos. El sabio agarró la cola y la resiguió con las manos. No tuvo dudas, “Es igual a una vieja cuerda” exclamo.

El quinto de los sabios se encontró con la oreja y dijo: “Ninguno de vosotros ha acertado en su forma. El elefante es más bien como un gran abanico plano”.

El sexto sabio que era el más viejo, se encaminó hacia el animal con lentitud, encorvado, apoyándose en un bastón. De tan doblado que estaba por la edad, pasó por debajo de la barriga del elefante y tropezó con una de sus gruesas patas. “¡Escuchad! Lo estoy tocando ahora mismo y os aseguro que el elefante tiene la misma forma que el tronco de una gran palmera”.

Satisfecha así su curiosidad, volvieron a darse las manos y tomaron otra vez la senda que les conducía a su casa. Sentados de nuevo bajo la palmera que les ofrecía sombra retomaron la discusión sobre la verdadera forma del elefante. Todos habían experimentado por ellos mismos cuál era la forma verdadera y creían que los demás estaban equivocados.

Otras consideraciones

- Debemos conservar el contexto. → Nuestra verdad es solo la porción de realidad que percibimos.
- Solo porque sea accesible no lo hace ético.
- Se crean nuevas brechas digitales/sociales.

Plan de Gobernanza de Big Data

Herramientas fundamentales en las empresas que necesitan entender de mejor forma sus datos.

- procesos
- funciones
- normas
- políticas
- mediciones

Garantizar el uso eficiente y eficaz de la información proveniente de una fuente con el fin de ayudar a lograr los objetivos planificados por una empresa.

Asegurar la calidad y la seguridad de los datos a analizar.

Plan de Gobernanza de Big Data

- Garantizar acceso y autorización granular a datos.
- Protección de datos y autenticación integrada
- Encriptación y Tokenización de Datos
- Auditoría y Análisis
- Diseñar una arquitectura de datos unificada

1.4 El principio de Bonferroni

- Test de comparaciones múltiples de eventos “raros”.
- Asumimos datos aleatorios e independientes.
- Si el número de potenciales eventos es mayor que el de instancias reales que se tenía la esperanza de hallar, entonces el modelo de búsqueda necesita ser reformulado.

1.4 El principio de Bonferroni

Supongamos que estamos en búsqueda de un grupo de maleantes. La información sugiere que éstos se reúnen periódicamente en un hotel para organizar sus planes. Supongamos también que:

1. Tenemos una lista de mil millones de personas consideradas como posibles maleantes.
2. Cualquiera de estas personas visita un hotel un día de cada cien.
3. Un hotel en promedio alberga cien personas. Por lo tanto, existen 100,000 hoteles (lo suficiente para albergar al 1% entre los mil millones de personas que visitan un hotel en cualquier día).
4. Podemos examinar los registros correspondientes a los últimos mil días de los hoteles.

Para encontrar a los maleantes en estos datos debemos buscar a las personas que en dos días diferentes estuvieron en el mismo hotel. Supongamos primero que las personas se comportan de manera aleatoria, entonces tenemos una probabilidad de 0.01 de visitar un hotel en cualquier día dado. El grupo de personas que decidieron ir a un hotel, seleccionarán uno entre

los 10^5 que existen.

¿Qué tan probable es encontrar pares de personas que tomaron la misma decisión?

¿Cuántos eventos indicarían la presencia de maleantes?

¿Qué pasaría si se incrementara el número de días de observación a 2000?

¿Si el número de personas observadas fuese 2 mil millones (incrementando también el número de hoteles a 200,000)?

1.4 El principio de Bonferroni

¿Qué tan probable es encontrar pares de personas que tomaron la misma decisión?

La probabilidad de que dos personas decidan ir a un hotel en cualquier día, 0.0001

La probabilidad de que vayan al mismo hotel, es 0.0001/ 10^5 = 10^{-9}

La probabilidad de que además sea en días diferentes 10^{-18}

¿Cuántos eventos indicarían la presencia de maleantes?

El número de pares posibles en una lista de mil millones de personas $\binom{10^9}{2} = 5 \times 10^{17}$

El número de pares de días en mil días es $\binom{1000}{2} = 5 \times 10^5$

1.4 El principio de Bonferroni

Entonces el número esperado de eventos

$$5\,x\,10^{17} * 5\,x\,10^{5} * 10^{-18}$$

250,000

Un cuarto de millón de personas tiene un comportamiento atípico.

1.4 El principio de Bonferroni

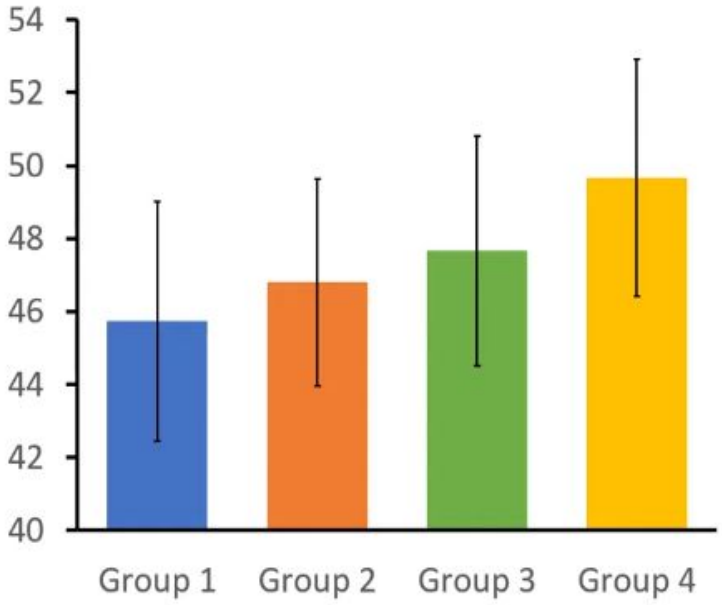
$$FWER = 1 - (a)^m$$

a tasa de error, m número de pruebas

$$FWER = 1 - (0.05)^1 = 0.05 = 5\%$$

1.4 El principio de Bonferroni

ANOVA
p-value 0.01



Comparison

Group 1 v Group 2

Group 1 v Group 3

Group 1 v Group 4

Group 2 v Group 3

Group 2 v Group 4

Group 3 v Group 4

1.4 El principio de Bonferroni

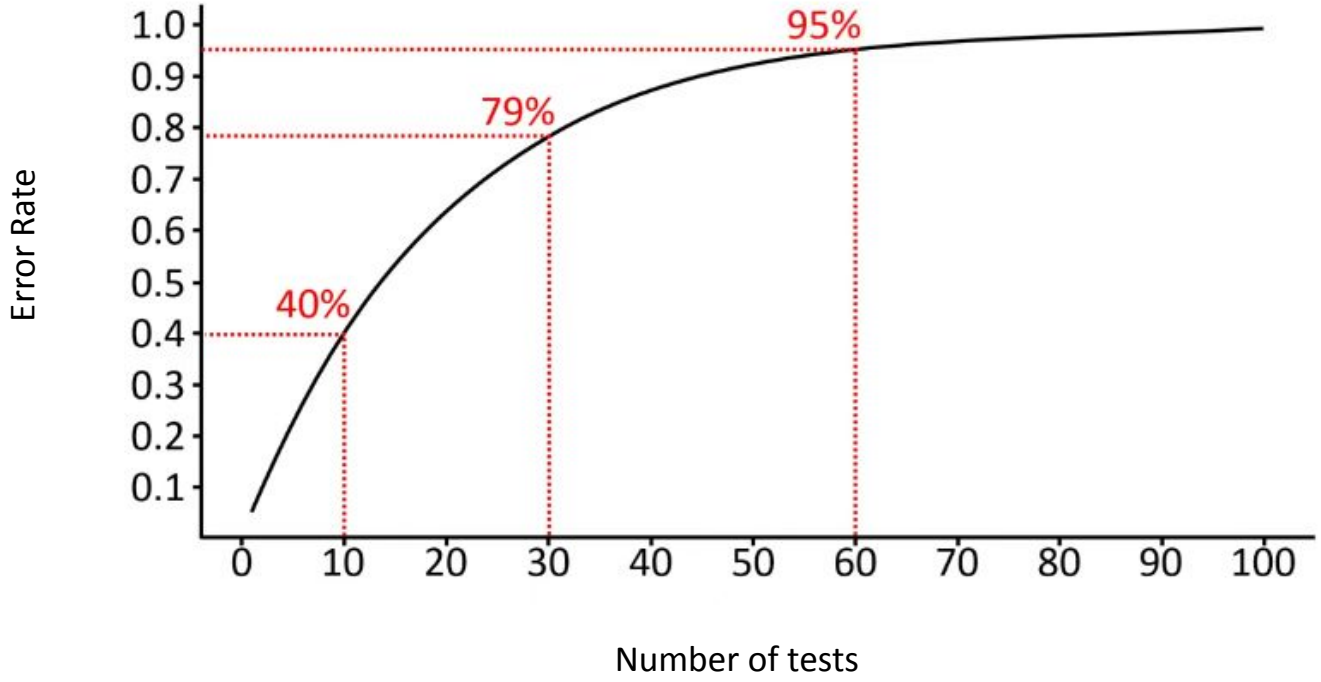
¿Cuál sería la tasa de error de tipo 1 para el ejemplo?

$$FWER = 1 - (0.05)^6 = 1 - 0.95^6 = 0.265$$

Así que haciendo seis pruebas, hay un 26.5% de oportunidad de encontrar al menos un falso positivo.

Mayor número de pruebas, más grande la tasa de error.

1.4 El principio de Bonferroni



1.4 La corrección de Bonferroni

$$a_2 = \frac{a_1}{k}$$

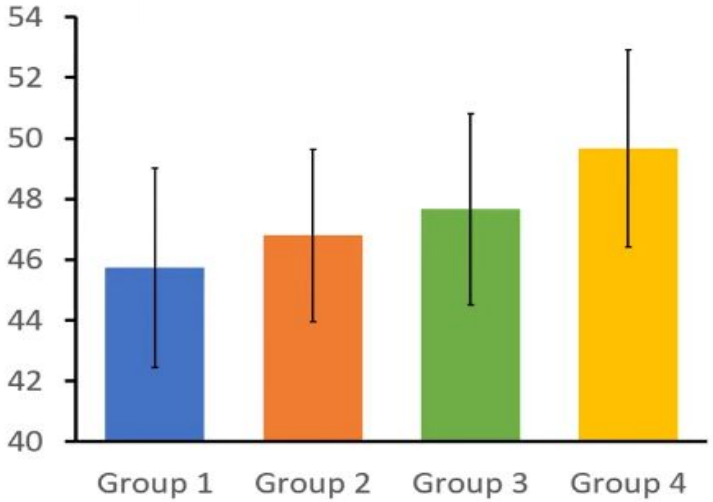
a_1 número de significancia
k número de pruebas

para el ejemplo

$$a_2 = \frac{0.05}{6} = 0.008 \longrightarrow 1-(1-0.008)^6 \longrightarrow 1-0.992^6 \longrightarrow 0.047$$

$$a_2 < 0.05\,\%$$

$\alpha_{\text{Bonferroni-corrected}} = 0.008$



Comparison	P value	
Group 1 v Group 2	0.349	✗
Group 1 v Group 3	0.111	✗
Group 1 v Group 4	0.003	✓
Group 2 v Group 3	0.435	✗
Group 2 v Group 4	0.016	✗
Group 3 v Group 4	0.098	✗

Práctica 1 - Equipos

Buscar un set de datos (reales) grande.

- Planteamiento del problema
- Cálculo del p-value
- Aplicación del principio
- **Conclusiones ***

* Deben incluir si fue bueno o malo aplicar el principio y porqué, además de las conclusiones de aplicarlo para sus casos específicos.

1.5 Privacidad y riesgo

- Asegurar la **máxima protección de los datos** es un elemento básico en cualquier proyecto Big Data.
  - Los datos deben ser tratados con transparencia (para el cliente).
  - Creación de perfiles.
  - Privacidad desde el diseño.
  - Aplicación de técnicas que permitan trabajar data sensible.
  - Consentimiento.

- Riesgos
  - Programas de computación no fiables.
  - Necesidad de recopilar información de muchas fuentes distintas - Problemas de validación de entrada y filtrado.
  - No existen controles de acceso granular a los datos.
  - Sesgos y privacidad
  - Filtro burbuja