IIC2026
Visualización de Información

Hernán F. Valdivieso López
(2022 - 2 / Clase 25)

Antes de empezar... Revisión de contenidos (RC)

1. ¡No se olviden del último control publicado!
2. Se acaba de publicar una actividad obligatoria. Consiste en **inventar una pregunta de alternativas** de algunos de los contenidos vistos en clases. En el enunciado se detalla cuáles contenidos pueden ser.
    - **Duración**: 2 semanas a partir de hoy.
    - **Intentos para responder**: ilimitados.
    - **Condición para obtener el punto RC**: cumplir las condiciones indicadas en el enunciado.
    - Pueden acercarse a **cualquier miembro del cuerpo docente** para tener *feedback* de esta actividad (solo *feedback*, no que le inventen la pregunta a ustedes).

Temas de la clase - Privacidad de Datos en Visualización

1. Motivación
2. Privacidad en el dataset
3. Privacidad en la visualización
4. Examen

Motivación

Motivación

- A veces nos toca trabajar con datos que tienen información sensible para el usuario
  - Enfermedad.
  - Sueldo.
  - Deudas.
  - Entre otros (dependerá del contexto).
- 🤔 ¿Cómo aseguramos que no puedan identificar a nuestros usuarios?

Motivación

- Los datos se pueden catalogar en diferentes tipos (no excluyentes):
  - **Identificadores**. Información que permite identificar al individuo detrás de los datos como el rut, mail uc, nombre completo.
  - **Potenciales identificadores**. Información general del individuo que eventualmente pueden servir para identificar al individuo como la edad.
  - **No identificadores**. Información que por sí sola no permite identificar al individuo.
  - **Sensibles**. Información privada que se espera desvincular del individuo.

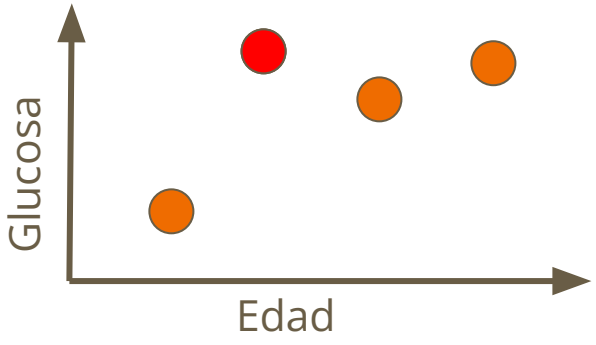
Name	Age	Glucose	Diabetes
Alice Smith	25	237	Yes
Bob Taylor	34	186	Yes
Frank Jones	33	165	Yes
Ivy Smith	42	80	No
Jim Davies	12	190	Yes
…	…	…	…

Fuente: Privacy preserving data visualizations, Avraam et al.

Motivación - ¿Y qué pasa en la visualización?

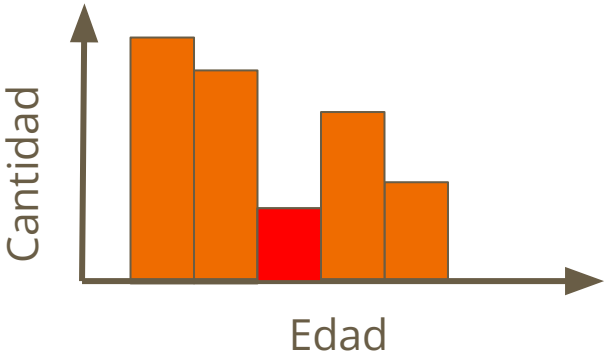
Name	Age	Glucose	Diabetes
Alice Smith	25	237	Yes
Bob Taylor	34	186	Yes
Frank Jones	33	165	Yes
Ivy Smith	42	80	No
Jim Davies	12	190	Yes
…	…	…	…

Scatterplot Glucosa VS Edad



Conozco la edad de Alicia, ahora sé que tiene diabetes

Histograma por edad de personas diabéticas



Fuente: Privacy preserving data visualizations, Avraam et al.

¿Qué se hace?

¿Qué se hace?

- **Objetivo**: Asegurar que no sea posible vincular un dato sensible al usuario.
- **Acciones**:
  - Se evita mantener identificadores en la visualización.
  - **Intentar generalizar o modificar posibles identificadores.**

¿Qué se hace?

- Se han creado diferentes nociones de privacidad:
  - ***k-anonymity*** → Para cada fila, hay al menos ***K-1*** filas que tienen los mismos datos no identificadores o potenciales identificadores.
  - ***l-diversity*** → Para cada grupo de datos iguales, hay al menos ***L*** datos sencible diferentes
  - ***t-Closeness*** → Para cada grupo, la distancia entre la distribución de datos sensibles no supera el valor ***T***.

Veamos un ejemplo de estas 2 nociones

Fuente: t-closeness: Privacy beyond k-anonymity and l-diversity, Li et al.

¿Qué se hace?

- Se han creado diferentes nociones de privacidad:
  - ***k-anonymity*** → Para cada fila, hay al menos k-1 filas que tienen los mismos datos no identificadores o potenciales identificadores.

Company	Position	Nationality	Zip	Age	Disease
Alpha	Director	Japanese	10001	32	Galactosemia
Beta	Manager	Indian	11049	53	Cancer
Gamma	Associate	American	10011	38	Galactosemia
Beta	Manager	Russian	10004	43	Fatty Liver
Alpha	Manager	Japanese	10014	48	Hepatitis B
Delta	Consultan	Indian	10017	34	Galactosemia
Gamma	Associate	American	11042	57	Hepatitis B
Delta	Manager	American	10007	42	Hepatitis B
Gamma	Director	Japanese	11043	51	Galactosemia
Beta	Manager	Russian	10009	35	Galactosemia
Delta	Associate	Indian	10019	42	Fatty Liver
Gamma	Manager	Japanese	11047	63	Fatty Liver

Company	Position	Nationality	Zip	Age	Disease
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	110**	>=50	Galactosemia
*	*	*	110**	>=50	Cancer
*	*	*	110**	>=50	Hepatitis B
*	*	*	110**	>=50	Fatty Liver
*	*	*	100**	4*	Hepatitis B
*	*	*	100**	4*	Fatty Liver
*	*	*	100**	4*	Fatty Liver
*	*	*	100**	4*	Hepatitis B

4-anonymity

Fuente: Privacy-Preserving Data Visualization: Reflections on the State of the Art and Research Opportunities, Bhattacharjee et al.

¿Qué se hace?

- Se han creado diferentes nociones de privacidad:
  - ***l-diversity*** → Para cada grupo de datos iguales, hay al menos L datos sencible **diferentes**

Company	Position	Nationality	Zip	Age	Disease
Alpha	Director	Japanese	10001	32	Galactosemia
Beta	Manager	Indian	11049	53	Cancer
Gamma	Associate	American	10011	38	Galactosemia
Beta	Manager	Russian	10004	43	Fatty Liver
Alpha	Manager	Japanese	10014	48	Hepatitis B
Delta	Consultan	Indian	10017	34	Galactosemia
Gamma	Associate	American	11042	57	Hepatitis B
Delta	Manager	American	10007	42	Hepatitis B
Gamma	Director	Japanese	11043	51	Galactosemia
Beta	Manager	Russian	10009	35	Galactosemia
Delta	Associate	Indian	10019	42	Fatty Liver
Gamma	Manager	Japanese	11047	63	Fatty Liver

Company	Position	Nationality	Zip	Age	Disease
*	*	*	1000*	<50	Galactosemia
*	*	*	1000*	<50	Fatty Liver
*	*	*	1000*	<50	Hepatitis B
*	*	*	1000*	<50	Galactosemia
*	*	*	1104*	>=50	Hepatitis B
*	*	*	1104*	>=50	Galactosemia
*	*	*	1104*	>=50	Fatty Liver
*	*	*	1104*	>=50	Cancer
*	*	*	1001*	<50	Galactosemia
*	*	*	1001*	<50	Hepatitis B
*	*	*	1001*	<50	Galactosemia
*	*	*	1001*	<50	Fatty Liver

3-diversity

Fuente: Privacy-Preserving Data Visualization: Reflections on the State of the Art and Research Opportunities, Bhattacharjee et al.

Métodos para intentar asegurar privacidad

- Generalización
- Filtrado
- Aplicar una máscara
- Reemplazar valores
    - Vecinos cercanos
    - Ruido

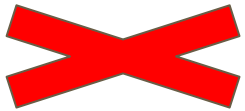
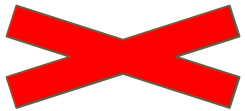
Métodos para intentar asegurar privacidad

- **Generalización**
- Filtrado
- Aplicar una máscara
- Reemplazar valores
  - Vecinos cercanos
  - Ruido

Edad	Edad Generalizada
11	< 50
69	>= 50
42	< 50
5	< 50
69	>= 50

Métodos para intentar asegurar privacidad

- Generalización
- **Filtrado**
- Aplicar una máscara
- Reemplazar valores
  - Vecinos cercanos
  - Ruido

Edad	Edad Generalizada
	69
69	42
42	69
	
69	

Métodos para intentar asegurar privacidad

- Generalización
- Filtrado
- **Aplicar una máscara**
- Reemplazar valores
  - Vecinos cercanos
  - Ruido

Edad	Edad con máscara
11	**
69	6&
41	4*
5	++
69	6&

Fuente: Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking, Ragan et al.

Métodos para intentar asegurar privacidad

- Generalización
- Filtrado
- Aplicar una máscara
- Reemplazar valores
  - **Vecinos cercanos**
  - Ruido

Edad	Reemplazar Valores *Promedio Vecinos cercanos*
11 (cercano al 5)	**8**
69	69
41	41
5 (cercano al 11)	**8**
69	69

Métodos para intentar asegurar privacidad

- Generalización
- Filtrado
- Aplicar una máscara
- Reemplazar valores
  - Vecinos cercanos
  - **Ruido**

Edad	Reemplazar Valores Ruido
11	12
69	68
41	40
5	5
69	70

Decisiones en visualización para facilitar la privacidad

Decisiones en visualización para facilitar la privacidad

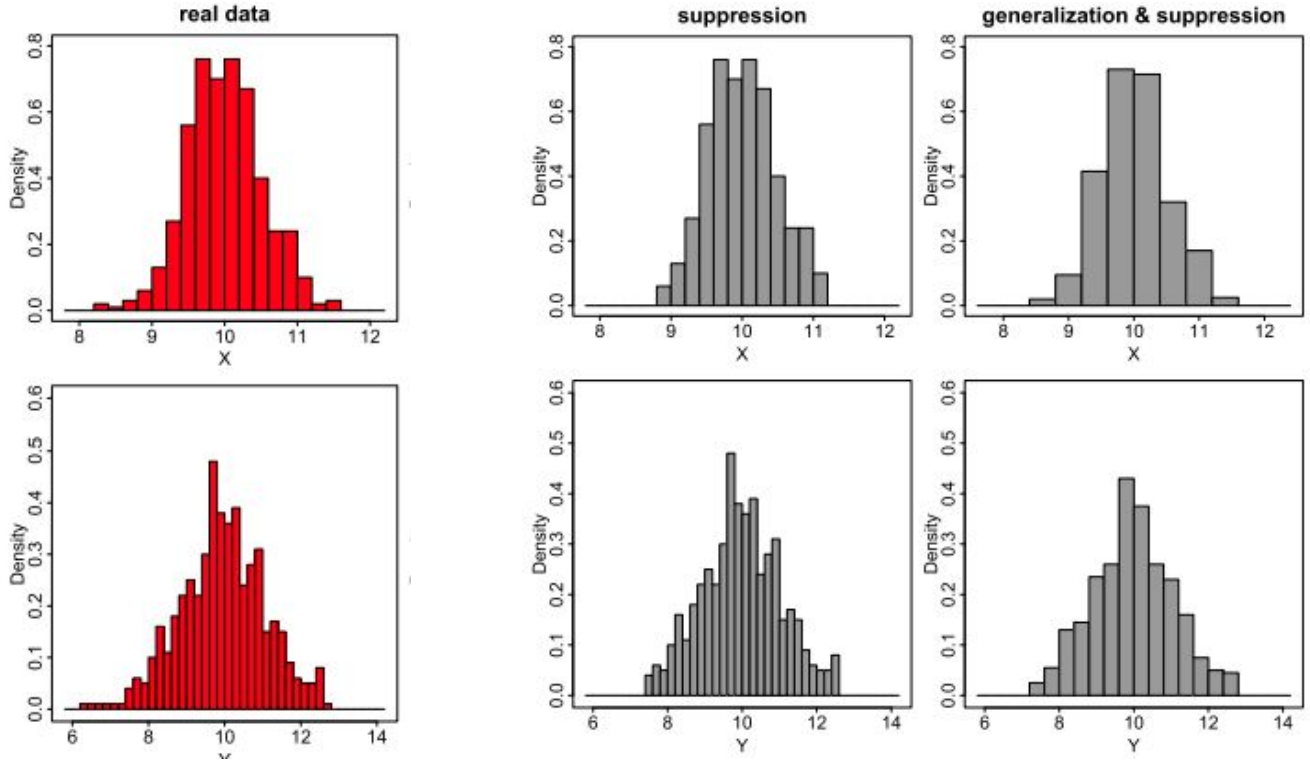
Hay casos donde no se puede modificar el *dataset* y es bueno pensar en formas que la visualización modifique *inline* la información antes de comenzar a dibujar.

Algunas decisiones implican:

1. Eliminación y/o agrupación de datos.
2. Aplicar ruido en función de vecinos cercanos o de forma probabilística.
3. Superposición de enlaces en el caso de grafo.

Decisiones en visualización para facilitar la privacidad

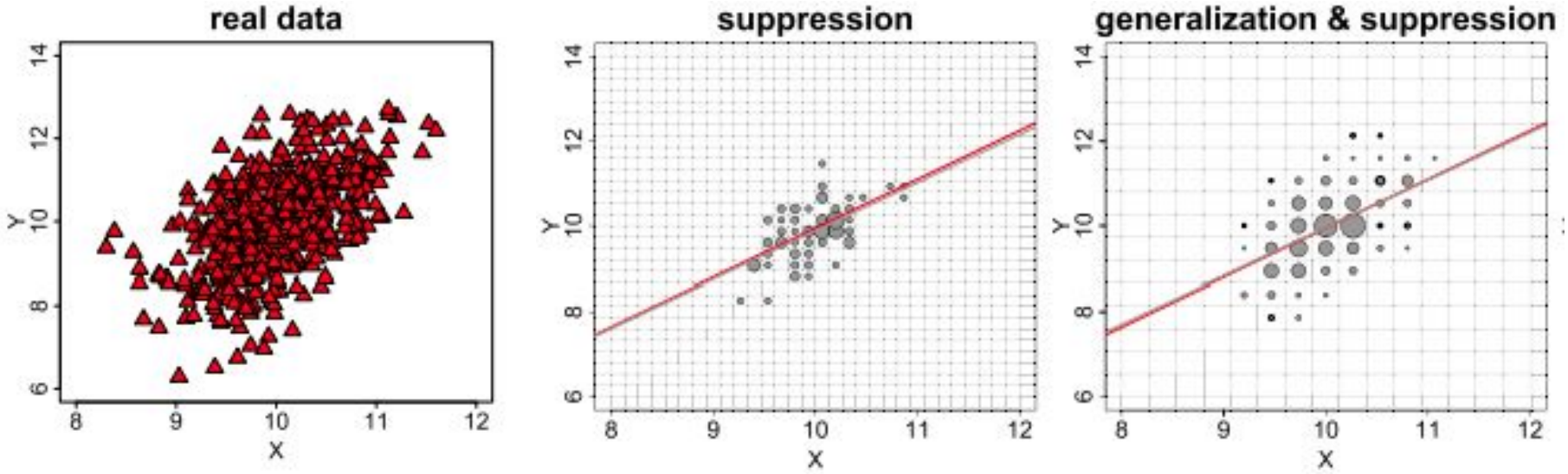
Eliminación y/o agrupación de datos.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

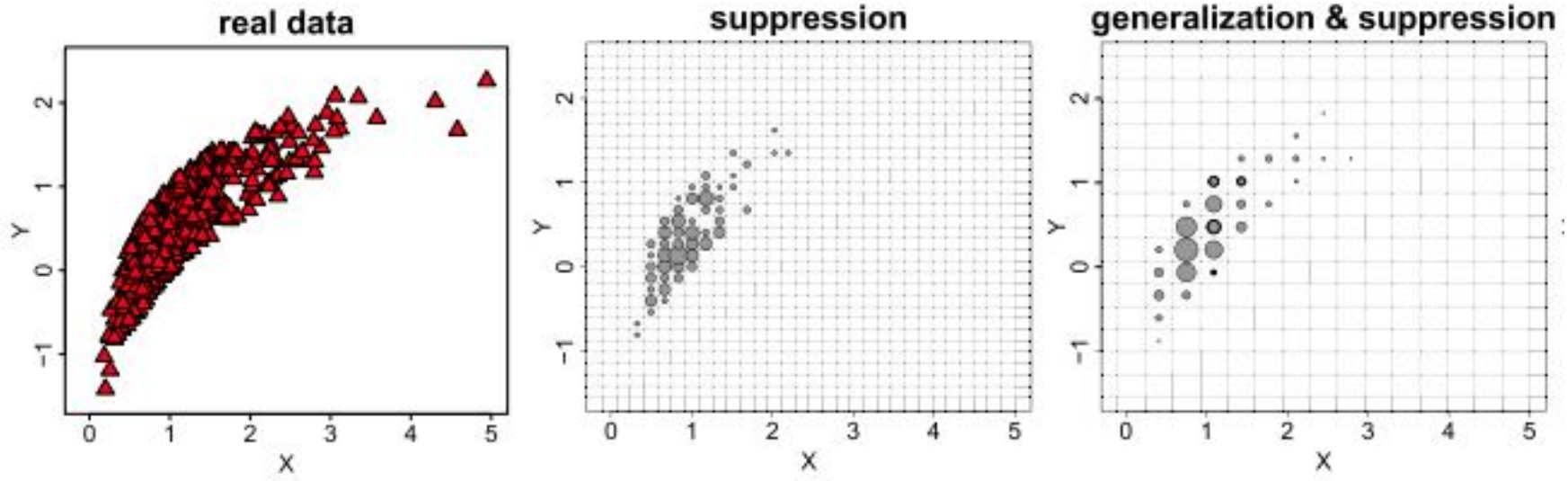
Eliminación y/o agrupación de datos.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

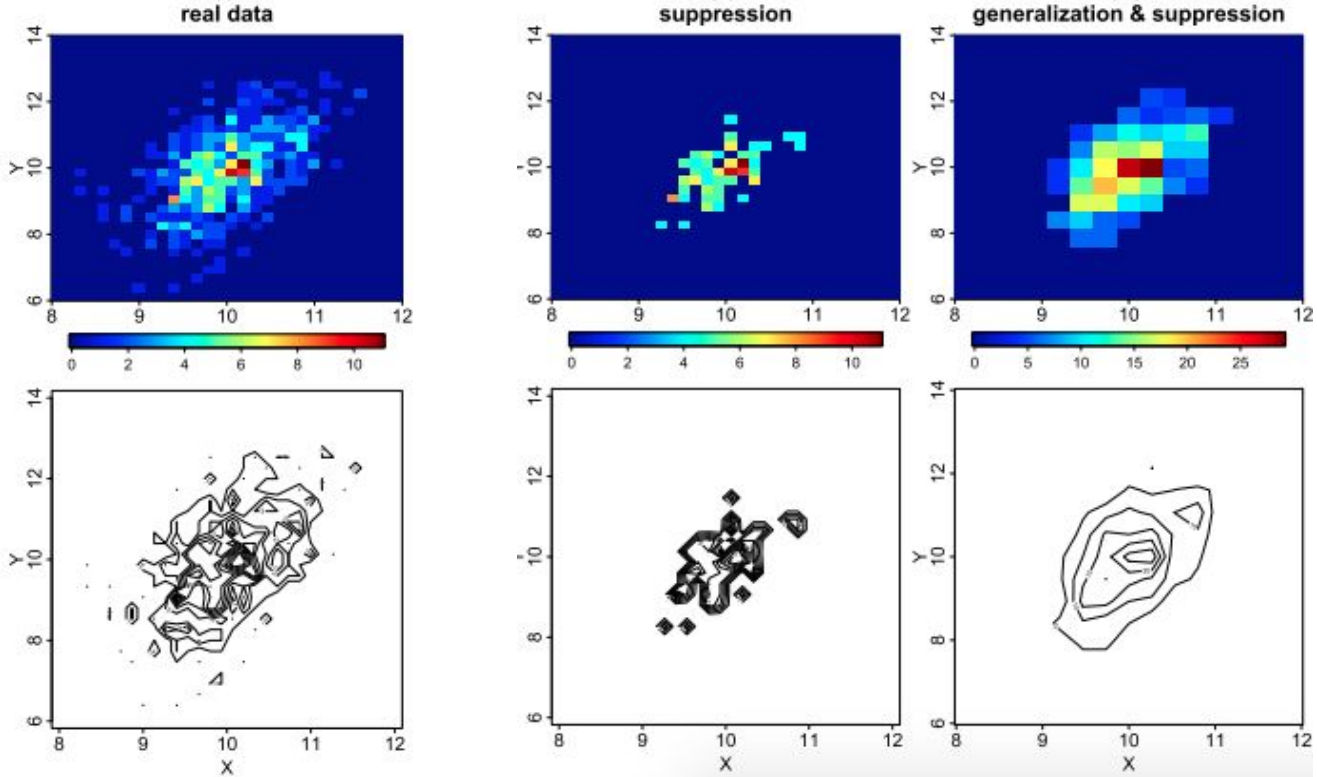
Eliminación y/o agrupación de datos.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

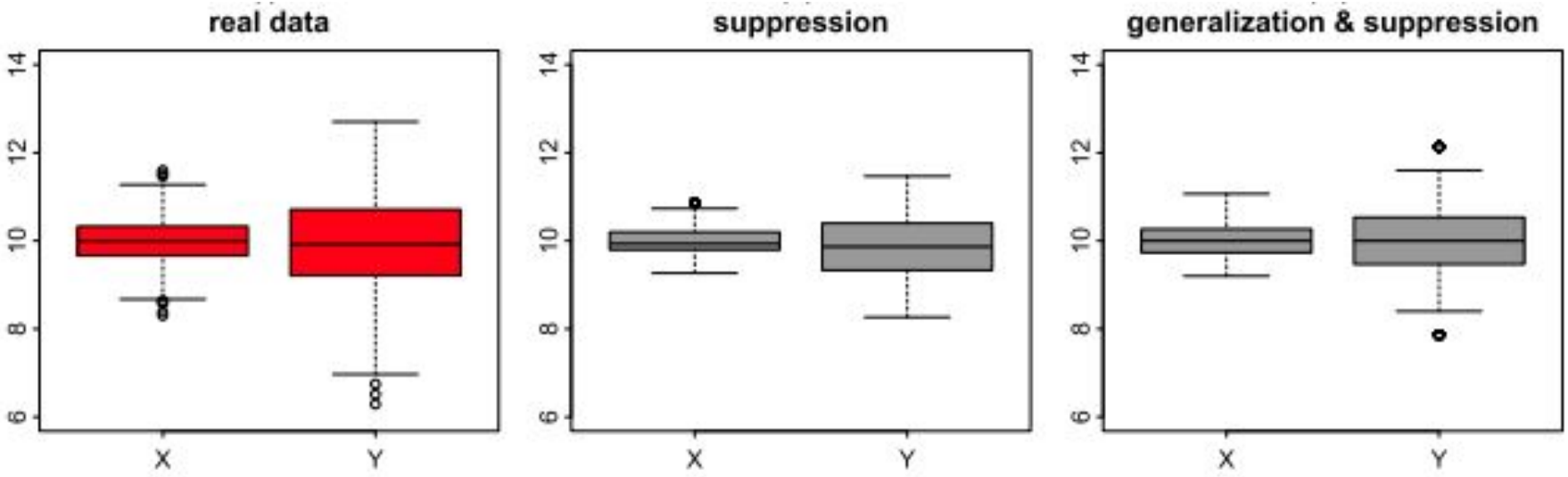
Eliminación y/o agrupación de datos.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

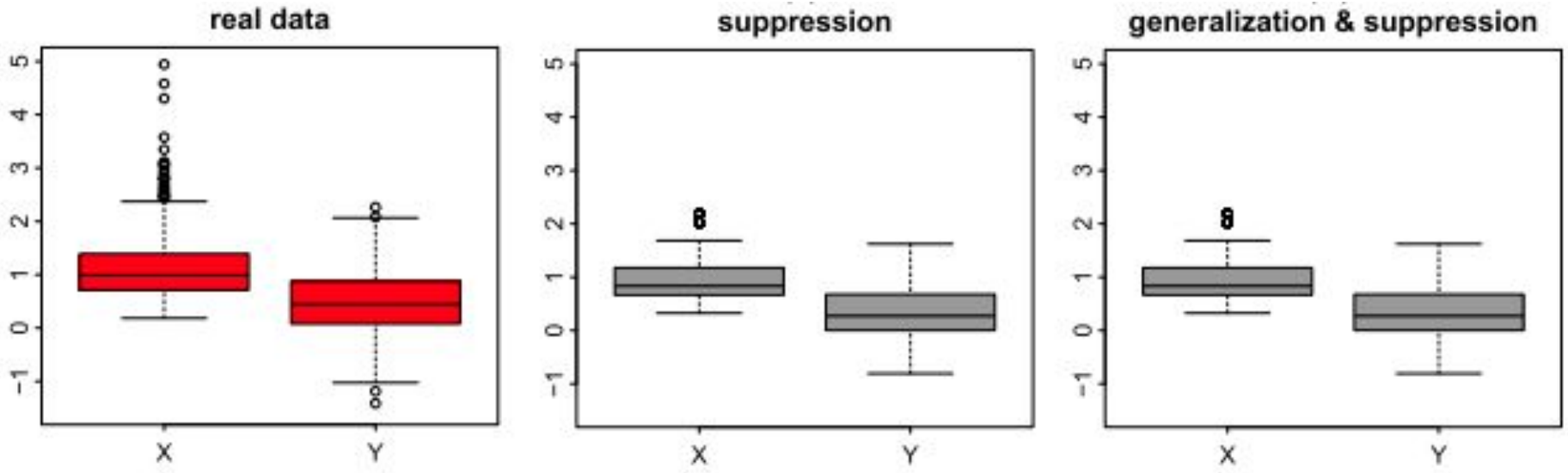
Eliminación y/o agrupación de datos.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

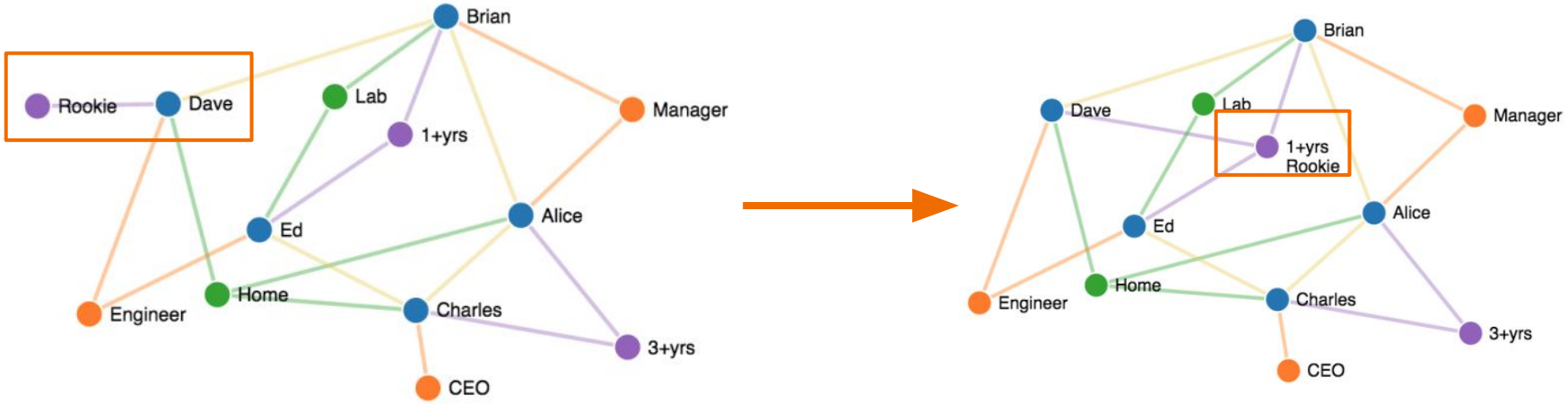
Eliminación y/o agrupación de datos.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

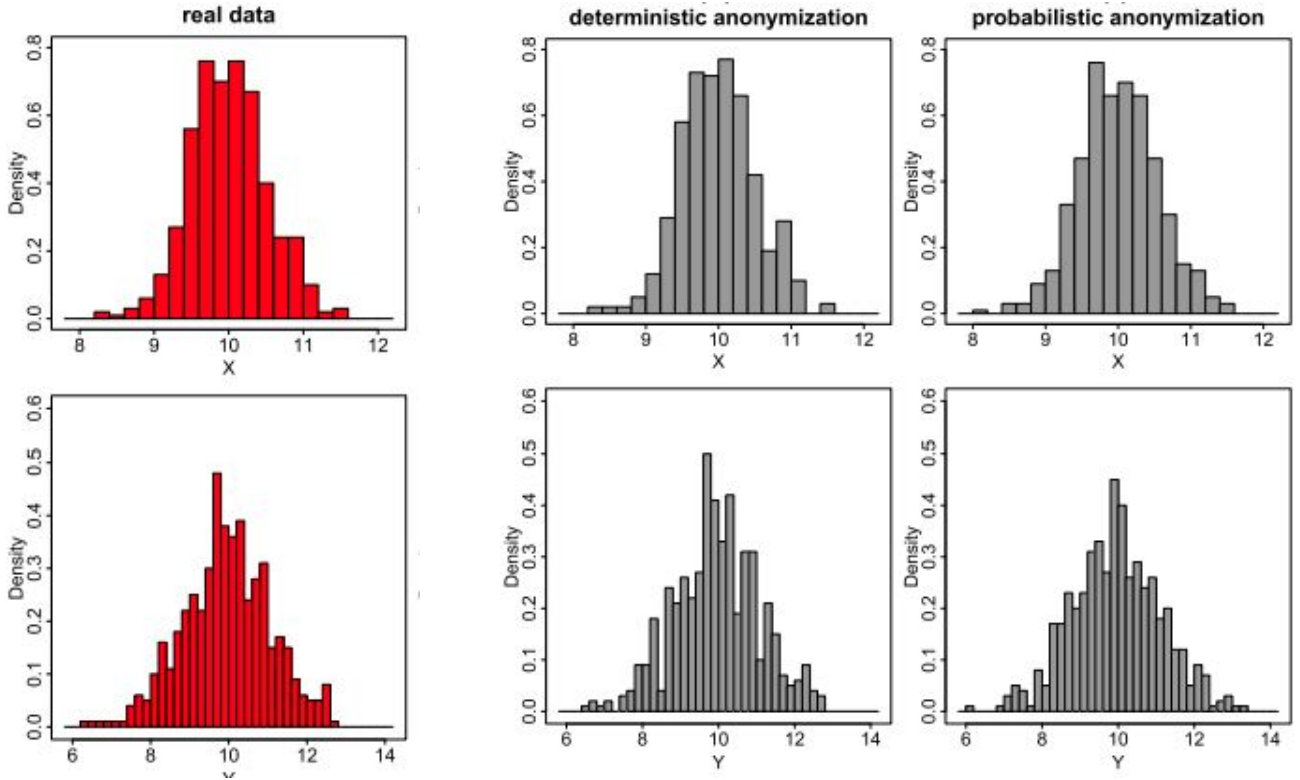
Eliminación y/o agrupación de datos.



Fuente: Privacy Preserving Visualization for Social Network Data with Ontology Information, Chou et al.

Decisiones en visualización para facilitar la privacidad

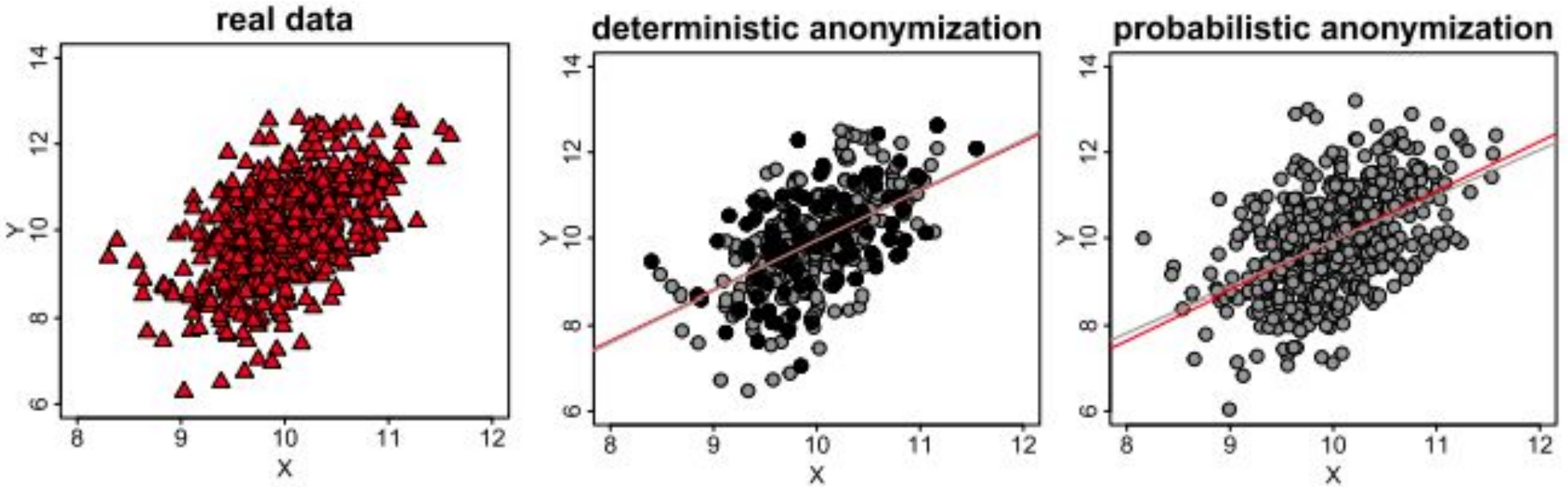
Aplicar ruido en función de vecinos cercanos o de forma probabilística.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

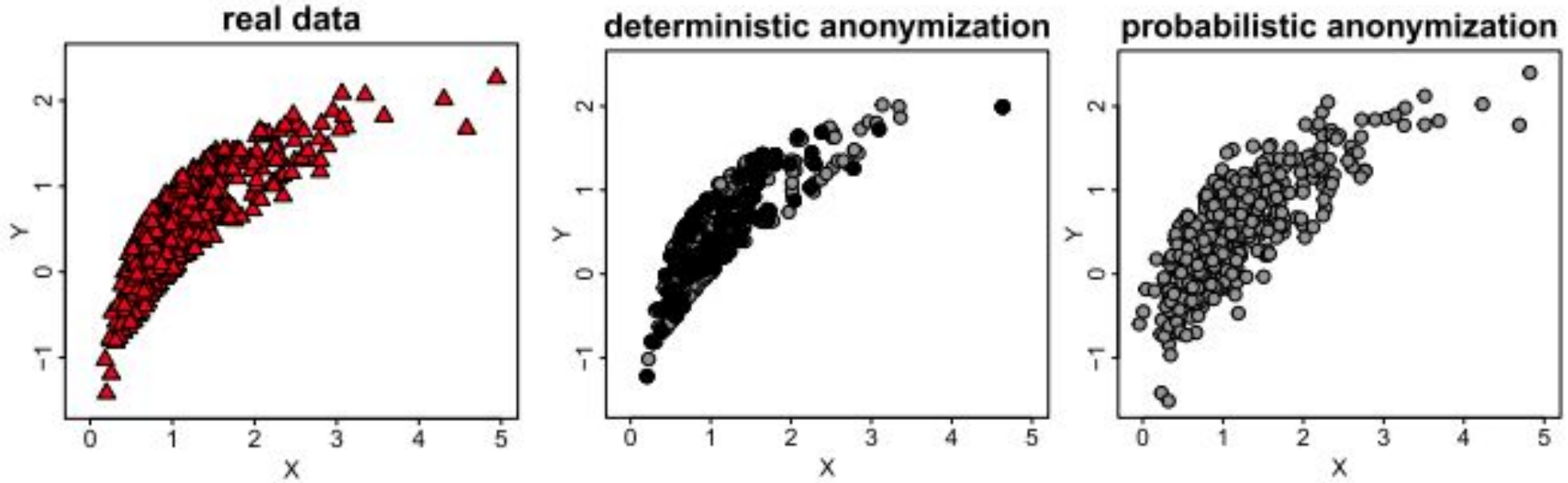
Aplicar ruido en función de vecinos cercanos o de forma probabilística.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

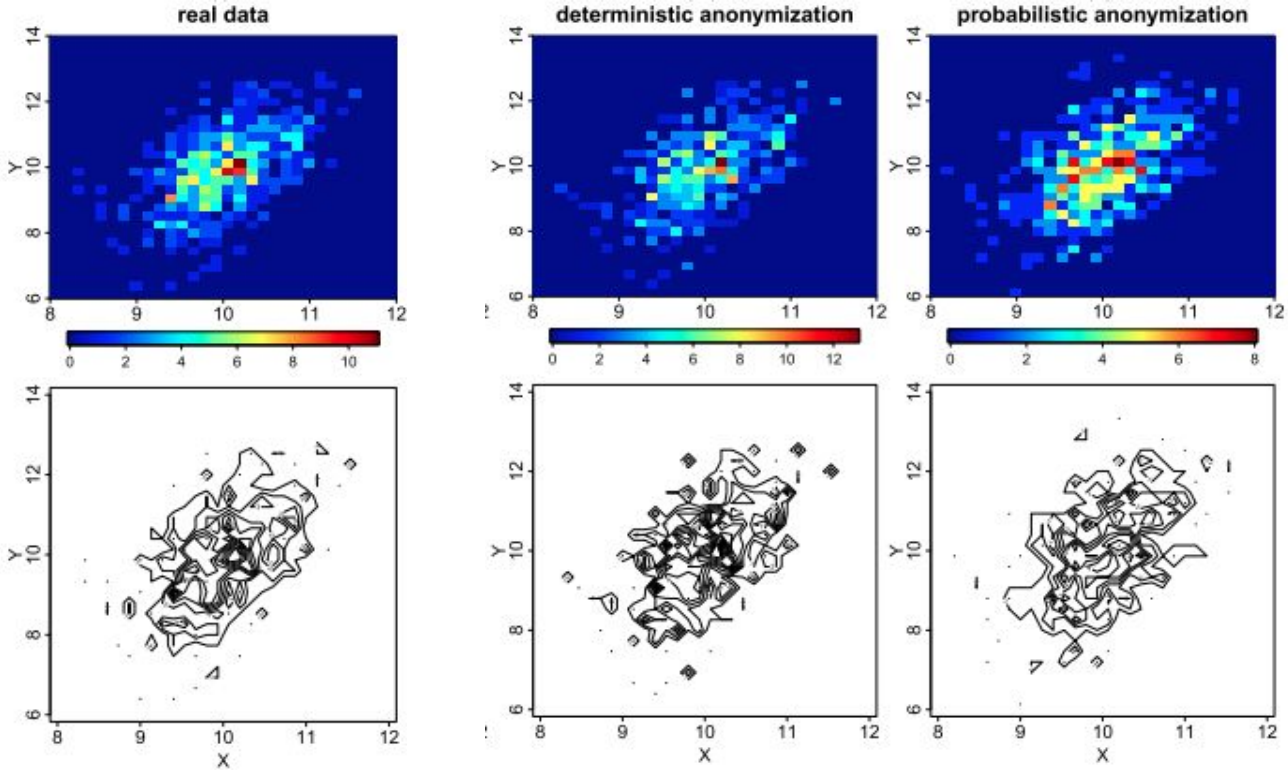
Aplicar ruido en función de vecinos cercanos o de forma probabilística.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

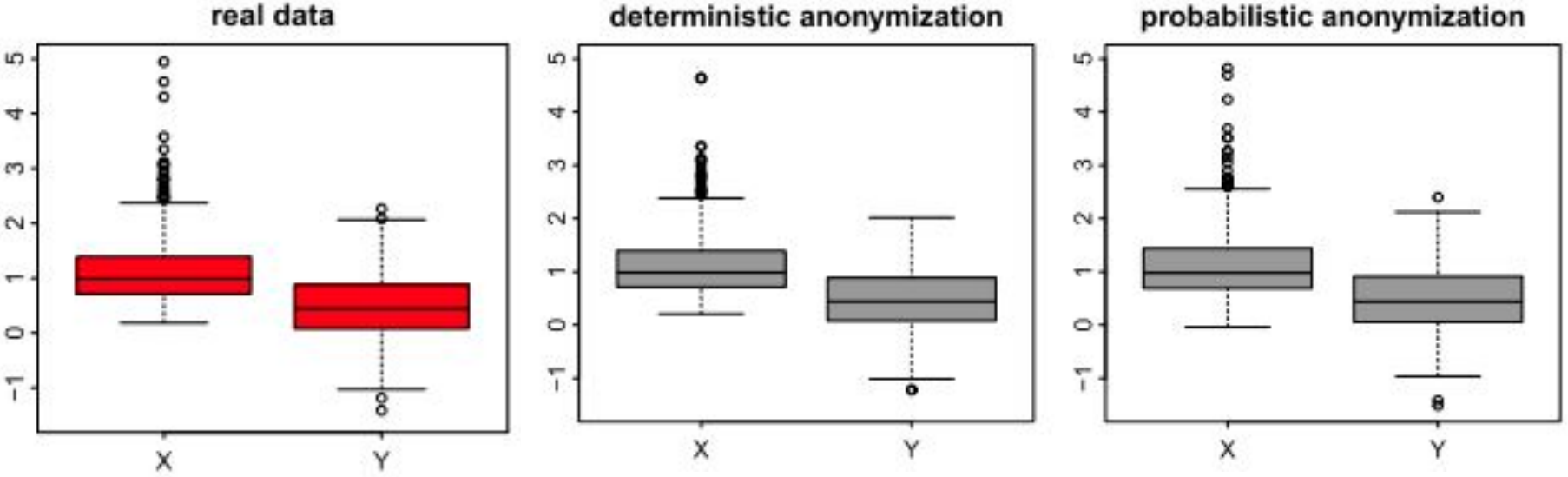
Aplicar ruido en función de vecinos cercanos o de forma probabilística.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

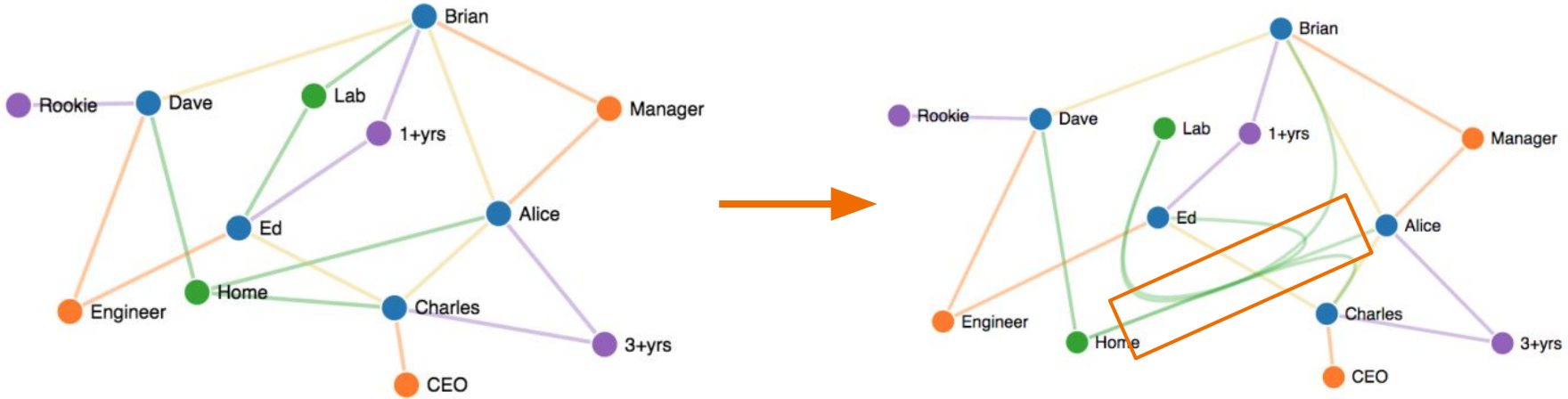
Aplicar ruido en función de vecinos cercanos o de forma probabilística.



Fuente: Privacy preserving data visualizations, Avraam et al.

Decisiones en visualización para facilitar la privacidad

Superposición de enlaces en el caso de grafo.



Fuente: Privacy Preserving Visualization for Social Network Data with Ontology Information, Chou et al.

Examen

Examen

- Trabajo final e integrador de los contenidos del curso.
- 2 entregables: informe en HTML y herramienta programada en D3.
- Gran parte de la herramienta se corrige en función de lo indicado en el informe.
    - No entregar informe hará que la herramienta no pueda ser corregida adecuadamente.
- Reprobatorio con nota menor a 2.95.
- 2 de diciembre, examen bonus para ganar décimas adicionales o recuperar puntos RC.

Examen

- Trabajo final e integrador de los contenidos del curso.
- **2 entregables**: informe en HTML y herramienta programada en D3.
- Gran parte de la herramienta se corrige en función de lo indicado en el informe.
    - **No entregar informe hará que la herramienta no pueda ser corregida adecuadamente.**
- **Reprobatorio con nota menor a 2.95.**
- 2 de diciembre, examen bonus para ganar décimas adicionales o recuperar puntos RC.

Examen - Informe

- **Situación de dominio.**
  - Explicar la situación en la que se enmarca la herramienta, el uso que le dará el usuario final y la naturaleza de los datos.
- **Abstracción de datos y tareas**
  - Utilizar el framework para describir los datos.
  - Utilizar el framework para indicar **3 tareas visuales.**
- **Decisiones de diseño.**
  - Indicar cada decisión de diseño utilizada.
  - Vincular las decisiones de diseño a las tareas visuales definidas antes.
  - Justificar, brevemente, que esa decisión es correcta con los principios y criterios vistos en el curso.

Examen - Herramienta

- **Algoritmo**
  - Programar las decisiones de diseño indicadas en el informe.
- **Requisitos mínimos a implementar y justificar**
  - Al menos 2 visualizaciones. Una de ellas no debe pertenecer al conjunto baneado.
  - Implementar navegación (`zoom`) o multiselección (`brush`).
  - Implementar vistas coordinadas entre 2 visualizaciones.
  - Implementación de filtro para reducir la cantidad de información.
  - Implementación de selección de 1 elemento con el *mouse*.
  - Uso correcto de `Data Join` para actualizar la visualización.

Examen - Datasets

- Find Open Datasets and Machine Learning Projects | Kaggle
- *Spreadsheet* incluido en el enunciado
- GitHub - awesomedata/awesome-public-datasets: A topic-centric list of HQ open datasets.
- OpenCorporates
- Data | FiveThirtyEight
- UNdata

Próximos eventos

Clase de martes

- Pandas
- Última clase de contenidos

Próxima ayudantía

- Grafos y árboles en D3
- Última ayudantía con temas nuevos de D3

Clase del jueves

- Trabajar en el examen
- Resolver dudas

IIC2026
Visualización de Información

Hernán F. Valdivieso López
(2022 - 2 / Clase 25)