

Visualización para ciencia de datos

¿Qué es la visualización de datos y por qué la hacemos?

Contenido

1

Definiciones

2

Limitaciones de los recursos

3

Analítica visual

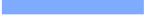
Visualización (vis)

“Sistemas de visualización por computadora proporcionan representaciones visuales de conjuntos de datos diseñados para ayudar a las **personas** a llevar a cabo las **tareas** con mayor **eficacia**” (y ética).

“La visualización es adecuada cuando hay una necesidad de aumentar capacidades humanas en vez de reemplazar a las personas con capacidades computacionales en los métodos de toma de decisiones. ”

—Tamara Muzner

¿Cuándo no usar vis?

- 
- ✓ Cuando la gente tiene preguntas bien definidas que hacer sobre los datos (Puede utilizar técnicas puramente computacionales de campos como la estadística y el aprendizaje automático)
 - ✓ Cuando existe una solución totalmente automática y se confía en ella.

¿Cuándo usar vis?

- ✓ Uso a largo plazo para los usuarios finales (por ejemplo, análisis exploratorio de datos científicos)
- ✓ Presentación para conocer resultados
- ✓ Un paso para comprender mejor los requisitos antes de desarrollar los modelos
- ✓ Ayudar a los usuarios finales de las soluciones automáticas a verificar, crear confianza.

**¿Por qué usar
representación
externa?**

Las representaciones externas aumentan la capacidad humana al permitirnos superar las limitaciones de nuestra propia cognición y memoria internas.

Es decir, reemplaza cognición con percepción.

**¿Por qué usar
computación
en bucle?**

- Más fácil
- Más allá de la paciencia humana
- Escala a grandes conjuntos de datos
(mejor distribución)
- Dinámica e interactiva
- Tiempo real

¿Por qué depende de la visión?

- **Visión:** El sistema visual humano es un canal de gran ancho de banda hacia el cerebro
- **Sonido:** menor ancho de banda y diferente semántica
- **Tacto:** empobrecida capacidad de grabación/reproducción
- **Sabor, Olor:** no hay dispositivos de grabación/reproducción viables.



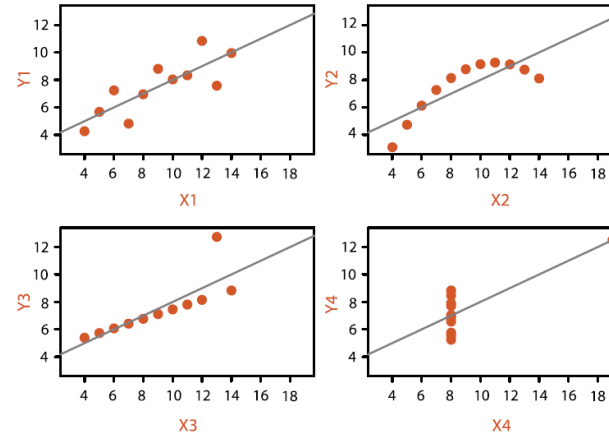
¿Por qué mostrar los datos en detalle?

Los resúmenes pierden información

- Confirmar lo esperado y encontrar patrones inesperados
- Evaluar la validez del modelo estadístico

Anscombe's Quartet: Raw Data

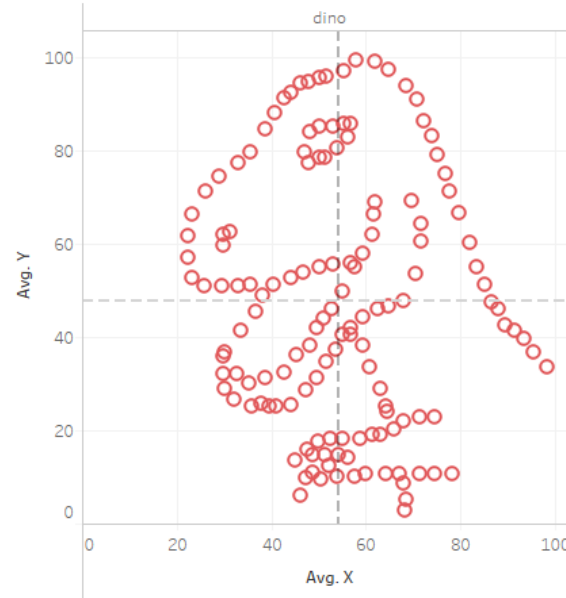
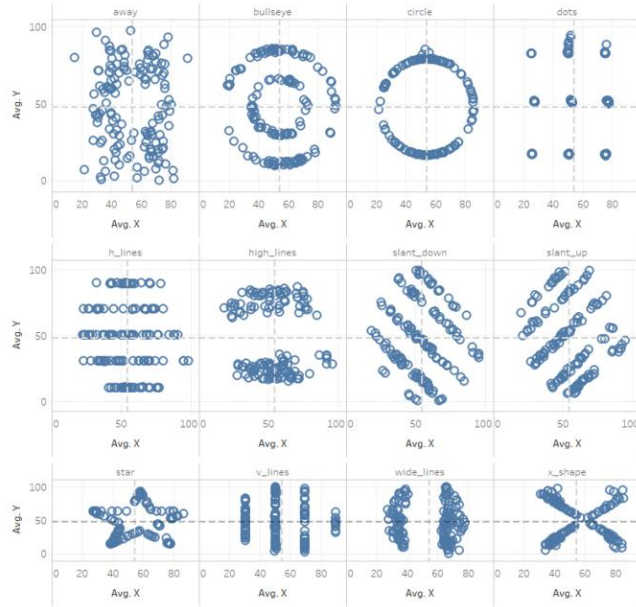
	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



Mr. Anscombe in 1973

¿Por qué mostrar los datos en detalle?

Datasaurus Dozen



<https://dabblingwithdata.wordpress.com/2017/05/03/the-datasaurus-a-monstrous-anscombe-for-the-21st-century/>

Modismos

¡Encuentra la mayor cantidad de maneras de visualizar los siguientes números en 5 minutos!

103

13

31

Modismos

Un enfoque distinto para crear o manipular representaciones visuales. El espacio de diseño de los posibles modismos visuales es enorme, e incluye las consideraciones de cómo crear y cómo interactuar con las representaciones visuales.

- **Cómo dibujarlo: el lenguaje de la codificación visual**
 - Muchas posibilidades de cómo crear
- **Cómo manipularlo: el lenguaje de la interacción**
 - Aún más posibilidades
 - Hacer que un solo idioma sea dinámico
 - Vincular múltiples expresiones idiomáticas a través de la interacción

¿Por qué centrarse en las tareas y la eficacia?

Los sistemas de visualización computarizada proporcionan representaciones visuales de conjuntos de datos diseñados para ayudar a las personas a llevar a cabo las tareas con mayor eficacia.

**Las tareas sirven de limitación en el diseño
(al igual que los datos)**

Ventajas: Los modismos no sirven para todas las tareas por igual

Desafío: convertir las tareas de vocabulario específico del dominio a formas abstractas

La mayoría de las posibilidades son ineficaces

Ventajas: Aumenta la posibilidad de encontrar buenas soluciones si se entiende el espacio completo de posibilidades

Desafío: La validación es necesaria, pero es difícil

Limitaciones en los recursos

Los diseñadores de Vis deben tener en cuenta tres tipos muy diferentes de limitaciones de recursos: las de las **computadoras**, las de los **humanos** y las de las **pantallas**.

Limitaciones en recursos

Límites computacionales

- Tiempo de procesamiento
- Memoria del sistema

Límites humanos

- Atención humana
- Memoria
- Retención de información

Límites en pantalla

- Los píxeles son un recurso precioso, el recurso más limitado
- Proporción de espacio usado para codificar información vs. espacio blanco no usado
- Intercambio entre el desorden y el desperdicio de espacio

Analítica visual

¿Cómo hacer análisis de datos?

- Análisis estadístico
- Aprendizaje automático e inteligencia artificial
- **Analítica visual (y análisis de datos)**

Análisis de datos

Análisis de datos descriptivo y exploratorio

- Búsqueda de patrones conocidos
- Mostrar los resultados utilizando técnicas tradicionales

Pros:

- Muchas soluciones
- Más fácil de implementar

Contras:

- No se puede buscar lo inesperado

Minería de datos / ML

- Basado en estadística clásica
- El enfoque de la caja negra
- Valores atípicos de salida y correlaciones
- El humano está fuera del circuito

Pros:

- Escalable

Contras:

- Los analistas tienen que dar sentido a los resultados
- Hace suposiciones sobre los datos

Analítica visual

- Interfaces visuales interactivas
- El humano en el bucle

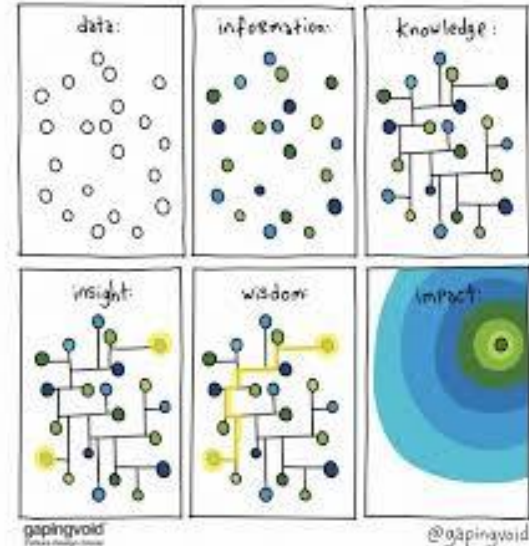
Pros:

- El ancho de banda visual es enorme
- Identificar patrones desconocidos y errores en los datos

Contras

- La escalabilidad puede ser un problema

Cadena de valor y modos de pensamiento



En analítica
visual buscamos
insights basados
en datos

Un profundo
conocimiento

Significativo

No sea obvio

Accionable

Un **insight** es algo
que el usuario
puede aprender
de los datos
usando el dataviz

No esperaba

Útil y
necesario

No sabía
absolutamente
nada de eso

Pueda sacarle
provecho

Gracias

¿Preguntas?

