Visualización de información

IIC1005 — Exploratorio de Computación

Invitado de hoy: **Nebil Kawas** (nebil@uc.cl)

¿Por qué necesito este curso? (I)

- Es una introducción al aprendizaje de los principios del diseño gráfico y de técnicas interactivas para visualizar datos.
- Está diseñado para entregar las herramientas necesarias para entender el estado del arte en *Visualización de información*.

¿Por qué necesito este curso? (II)

- Es un curso multidisciplinario que incorpora subcampos de la computación, de la estadística, del diseño gráfico (*e.g.* teoría de color, tipografía), y de la psicología cognitiva.

¿Por qué necesito este curso? (III)

- Busca explicar cómo las representaciones visuales son una ayuda en el análisis y entendimiento de *datasets* altamente complejos, y cómo, además, diseñar e implementar visualizaciones efectivas usando modernas librerías *web-based*.

Competencias del curso (I)

- Aplicar un proceso de diseño para crear visualizaciones efectivas,
- Llevar ideas a prototipos concretos, con la ayuda de bosquejos,
- Utilizar principios de percepción y cognición humana en visualización,
- Exponerse a distintos dominios de datos (*e.g.* redes, textos, cartografía),

Competencias del curso (II)

- Aplicar distintos métodos de visualización para un rango variado de *datasets*,
- Evaluar una visualización de forma crítica, pudiendo además sugerir e implementar mejoras,
- Trabajar como miembro en un equipo para sacar adelante un proyecto.

El propósito de visualizar información

¿Cuántos datos hemos producido?

- Al 2013: 4,4 zetabytes (1 zetabyte = 10^{21} bytes = 10^{9} terabytes)

https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Sensores físicos

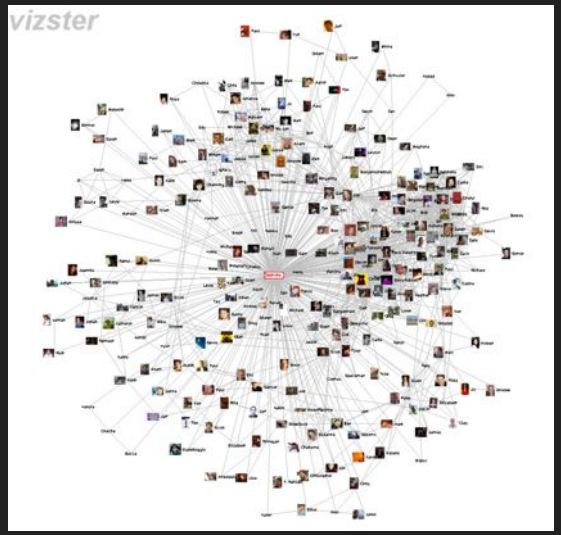
- Datos de sensores GPS de taxis en NYC (cabspotting.org)

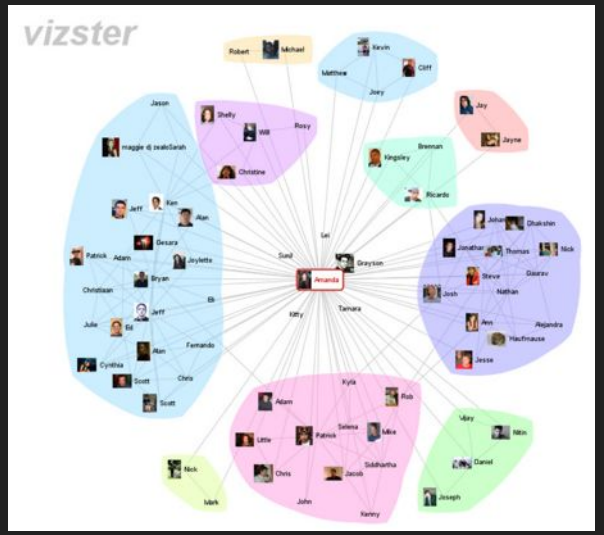
Sensores físicos (II)

- Datos en vivo de vuelos (www.flightradar24.com)

Registros de actividad humana

- Actividad en redes sociales, comunidades en linea, videos *online*, etcétera.

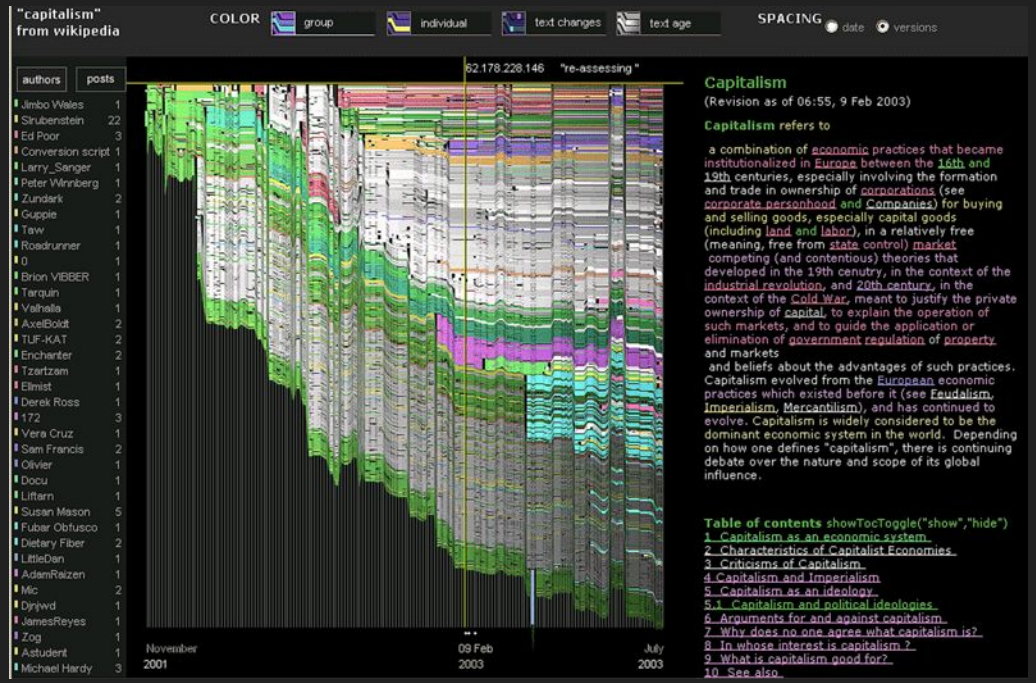




Heer, J., & Boyd, D. (2005, October). Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (pp. 32-39). IEEE.

Registros de actividad humana (II)

- IBM History Flow (2004)
- Visualización de ediciones en Wikipedia



Viégas, F. B., Wattenberg, M., & Dave, K. (2004, April). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 575-582). ACM.

¡Sobrecarga de información!

“What information consumes is rather obvious: it consumes the attention of its recipients.
Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.”

Herbert Alexander Simon (1916–2001)
Premio Nobel de Economía
Premio Turing

¡Sobrecarga de información! (II)

“The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”

Hal Varian
Google’s Chief Economist (2009)

¿Qué es visualización?

- Según el *Diccionario de la lengua española*, visualizar es…
  - Visibilizar.
  - Representar mediante imágenes ópticas fenómenos de otro carácter; p. ej., el curso de la fiebre o los cambios de condiciones meteorológicas mediante gráficas, los cambios de corriente eléctrica o las oscilaciones sonoras con el oscilógrafo, etc.

¿Qué es visualización? (II)

- Según el *Diccionario de la lengua española*, visualizar es…
  - Formar en la mente una imagen visual de un concepto abstracto.
  - Imaginar con rasgos visibles algo que no se tiene a la vista.
  - Hacer visible una imagen en un monitor.

¿Qué es visualización? (III)

- “Transformación de lo simbólico a lo geométrico.” [McCormick et al. 1987]
- “[...] encontrar la memoria artificial que mejor apoya nuestros medios naturales de percepción.” [Bertin 1967]
- “El uso de representaciones visuales de datos, generados por computador, interactivos, para amplificar nuestra cognición.” [Card, Mackinlay, & Shneiderman 1999]

¿Por qué crear visualizaciones?

- Comprender las relaciones entre conjuntos de datos
- Entender algo sobre los datos
- Resaltar información importante
- Plantear un argumento convincente
- A nadie le gusta leer registros (*e.g. web logs*)

```
127.0.0.1 - mbostock [12/Mar/2018:18:55:32- 0300] "GET /vis.gif HTTP/1.0" 200
```

¿Por qué crear visualizaciones? (II)

- Encontrar *outliers*
- Descubrir datos faltantes
- Comunicar información
- Reducir carga cognitiva para procesar información

Tres funciones de las visualizaciones

- Registrar información
  - Fotografías, planos, etcétera
- Apoyar razonamientos sobre la información (analizar)
  - Procesar y calcular
  - Razonar acerca de los datos
  - Retroalimentación e interacción
- Transmitir información a otras personas
  - Compartir y persuadir
  - Colaborar y revisar
  - Enfatizar aspectos importantes de los datos

Visualizar para registrar información

Registrar movimiento

Gallop, Bay Horse “Daisy” [Muybridge 1884-86]

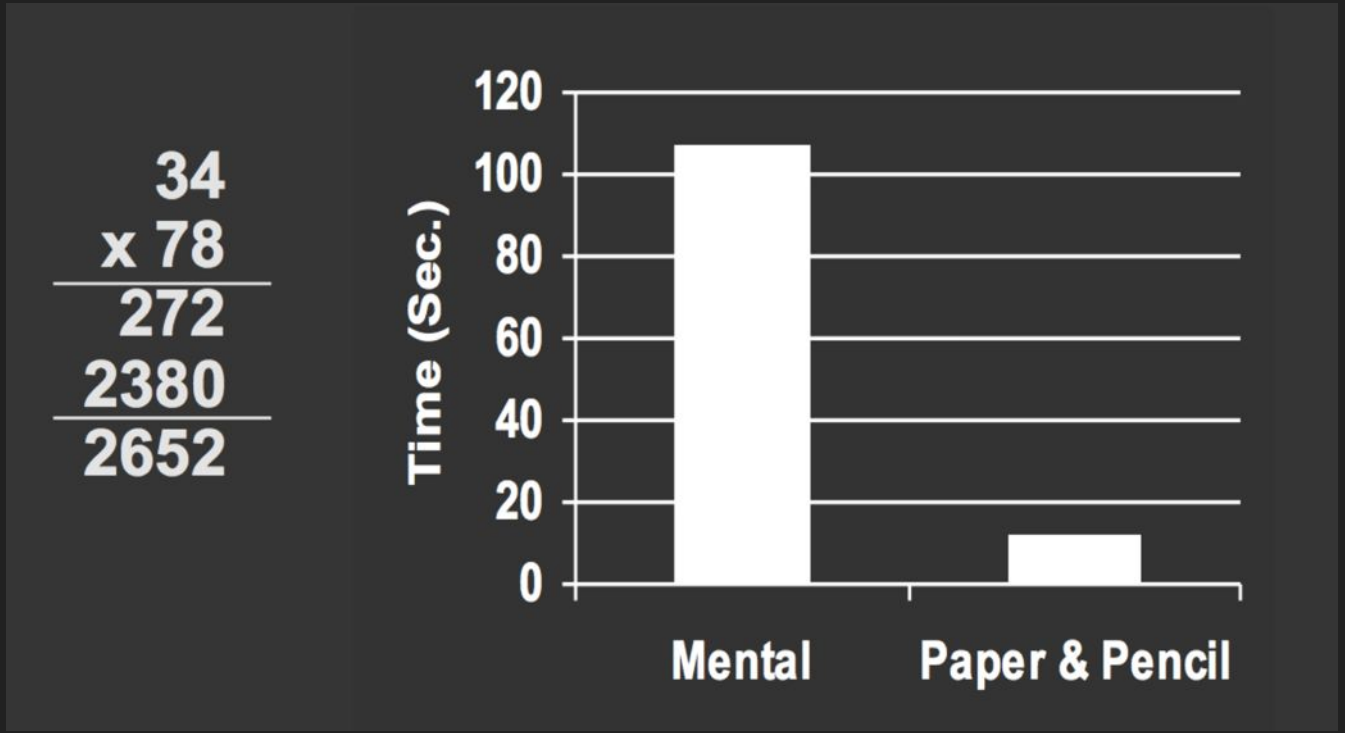
Registrar movimiento (II)

Visualizar para
apoyar razonamiento

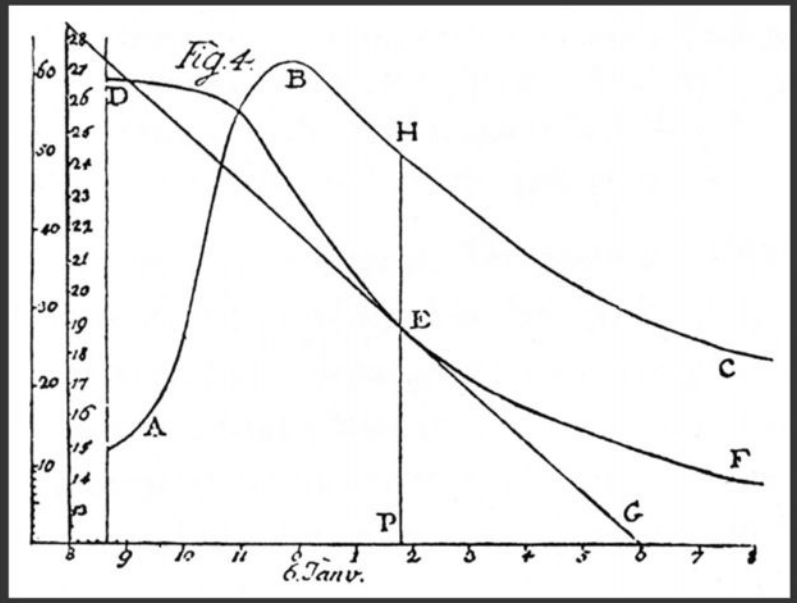
1854: cólera en Londres

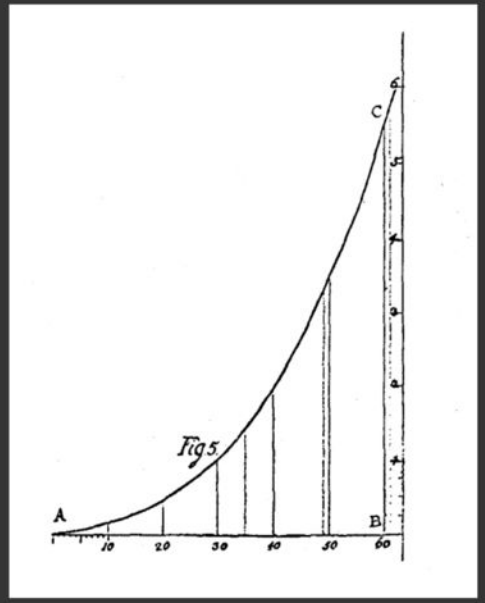
- Dr. John Snow usa análisis espacial para apoyar su hipótesis.

Expandir memoria: multiplicación



Expandir memoria: cálculo apoyado por gráficos



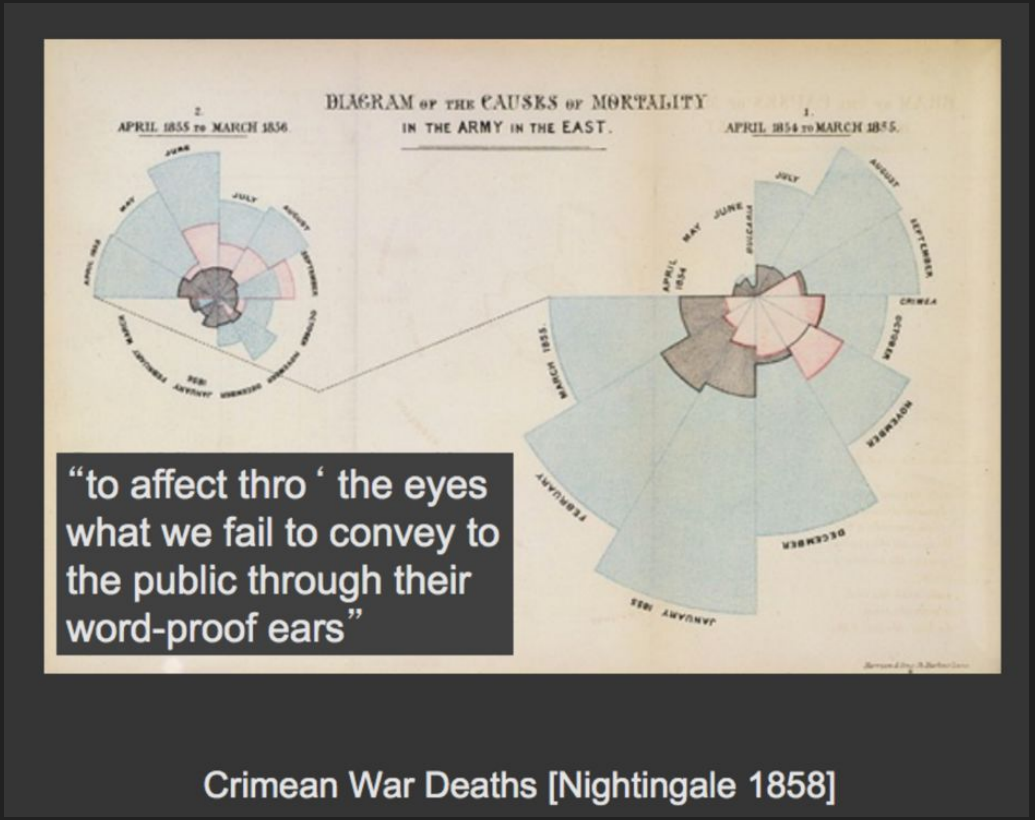


Johannes Lambert used graphs to study the rate of water evaporation as function of temperature [from Tufte 83]

Visualizar para
presentar información a otros

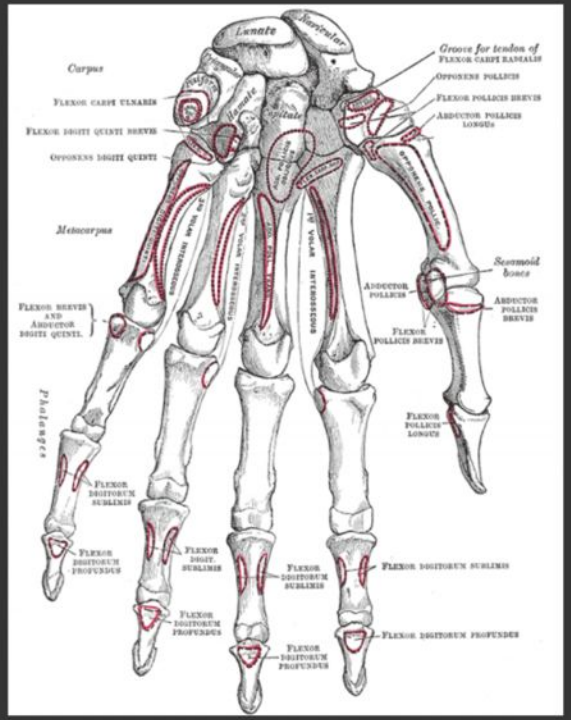
Argumentar

- “Las muertes en la Guerra de Crimea” por Florence Nightingale (trabajo hecho en 1858)
- El diseño escogido: *coxcomb* (o bien, *polar area diagram*)



Crimean War Deaths [Nightingale 1858]

Inspirar



Bones in hand [from 1918 edition]


Double helix model [Watson and Crick 53]

Retomando definición…

¿Qué es?

Definiciones [Munzner, 2014]

- *“Computer-based visualization systems provide **visual representations of datasets** designed to help people carry out tasks more **effectively**.”*

- *“Visualization is suitable when there is a need to **augment human capabilities** rather than replace people with computational decision-making methods.”*

¿Por qué…?

- ¿Por qué usar humanos en esto?
- ¿Por qué usar computadores en esto?
- ¿Por qué usar representaciones externas?
- ¿Por qué depender de la visión?
- ¿Por qué mostrar los datos en detalle?
- ¿Por qué usar interactividad?
- ¿Por qué enfocarse en la efectividad?
- ¿Por qué la mayoría de los diseños son inefectivos?
- ¿Por qué validar es un proceso difícil?
- ¿Por qué existe una limitación de recursos?

¿Por qué usar humanos en esto?

- La visualización permite analizar datos cuando todavía no sabemos qué preguntas formular.
- Si encontramos soluciones aceptables a problemas que no necesiten juicio humano, entonces una herramienta de visualización no será necesaria. (ejemplo: *stock market trading*)
- Sin embargo, hay problemas que no están bien definidos: allí necesitamos al humano con su poderosa detección de patrones. Lo que buscamos es, entonces, aumentar las capacidades humanas.

¿Por qué usar computadores en esto?

- Con la ayuda de un computador, es posible construir herramientas que permiten explorar o presentar *datasets* enormes, algo prácticamente imposible de *dibujar a mano*.
- Trabajar con un *dataset* de cientos de ítems podría tomar horas (o incluso días), por lo que una herramienta basada en un computador nos ahorra mucho esfuerzo humano en relación a la creación manual.

¿Por qué usar representaciones externas?

- Las representaciones externas aumentan la capacidad humana, al permitirnos superar las limitaciones de nuestra cognición interna.
- Estas pueden tomar muchas formas, como objetos tangibles (*e.g.* un ábaco, un quipu), aunque en este curso nos enfocaremos en lo que puede ser mostrado en la pantalla bidimensional de un computador.
- Estos diagramas son diseñados para apoyar inferencias perceptuales, que tienen como ventaja la posibilidad de organizar información en el espacio, lo que acelera tanto la búsqueda como en el reconocimiento.

¿Por qué depender de la visión?

- La visualización, como el nombre lo dice, se basa en explotar el sistema visual humano como un medio de comunicación. Nos enfocaremos en este canal en vez de otros sistemas sensoriales, ya que es apropiado para transmitir información.
- El sistema visual ofrece un canal *de banda ancha* hacia nuestros cerebros, a diferencia de nuestro oído que no funciona bien cuando tenemos una experiencia simultánea de sonidos durante un periodo largo de tiempo.
- Los otros canales tienen limitaciones tecnológicas: el olfato y el gusto no tienen una manera factible de ser grabados. Y el sentido del tacto es una parte bastante limitada de lo que podemos sentir.

¿Por qué mostrar los datos en detalle?

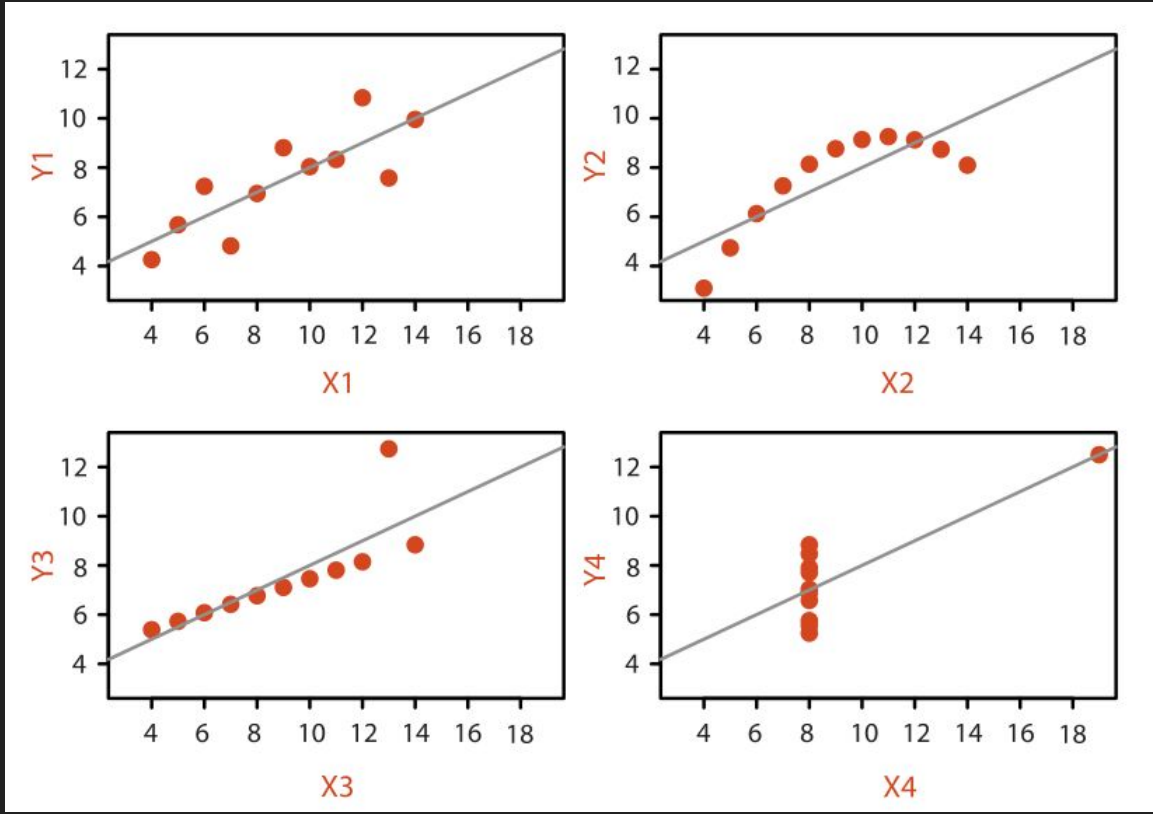
- Las herramientas de visualización son útiles cuando es necesario conocer la estructura de un *dataset* en detalle, y no sólo un breve resumen de él.
- Una de estas situaciones ocurre cuando exploramos datos para buscar patrones: tanto confirmar algo esperado, como encontrar algo inesperado.
- Obtener una caracterización estadística de los datos es un *approach* poderoso; no obstante, tiene inconvenientes intrínsecos al perder información mientras esta se resume.

Anscombe's Quartet

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	

Anscombe's Quartet



¿Por qué usar interactividad?

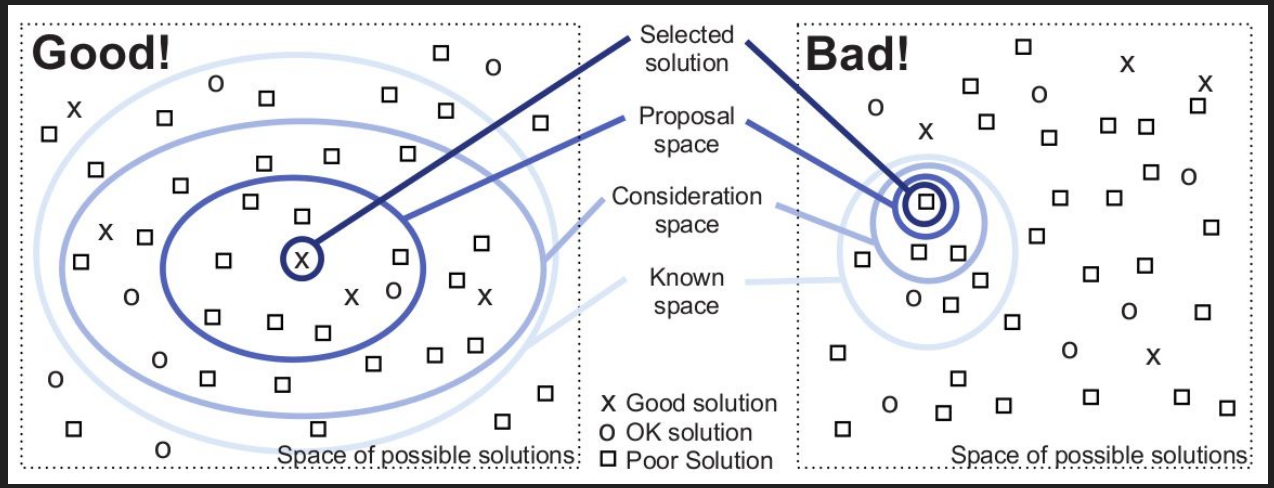
- La interactividad es **clave** para construir herramientas de visualización que manejen un *dataset* complejo.
- Este tipo de *datasets* tiene dos limitantes: tanto la percepción humana como los *displays* no nos permiten mostrar todo en un vistazo.
- Además, una vista estática sólo nos muestra un aspecto de los datos.
  Un *display* interactivo nos permite exponer distintos ***encodings***.
- En este curso, veremos técnicas de visualización relacionadas a las representaciones estáticas; sin embargo, la interacción será parte intrínseca a lo largo del curso.

¿Por qué enfocarse en la efectividad?

- El objetivo de la efectividad va asociado con correctitud, precisión y verdad, que juegan un papel esencial en visualización.
- El énfasis que tiene la visualización es diferente con respecto a otras disciplinas que también utilizan imágenes (*e.g.* arte, películas, *marketing*).
- Un diseñador de visualizaciones no tiene una licencia artística: no se trata de hacer “imágenes bonitas”. No sirve si es “bonito”, pero no efectivo.

¿Por qué la mayoría de los diseños son inefectivos?

- Dada la cantidad de combinaciones posibles al momento de crear una visualización, el espacio de posibles diseños es enorme.
- No hay un método claro de cómo optimizar, pero en este curso entregaremos algunos *guidelines*. Existen pocas verdades en esta disciplina.



¿Por qué validar es un proceso difícil?

- El problema de la validación en un diseño de visualización es complejo; podemos hacer muchas preguntas para saber si cumplimos con los objetivos:
  - ¿Cómo sabes si es que funciona? ¿Cómo argumentas que este diseño es mejor que otro para los usuarios? Espera, ¿qué significa *mejor*? ¿Los usuarios logran sus objetivos de manera más rápida? ¿O se divierten más al hacerlo? ¿O trabajan de manera más efectiva? Pero, ¿cómo mides la efectividad? ¿Cómo mides el *engagement*, o los *insights*? ¿Y esto es mejor que hacerlo de forma manual? ¿Y es mejor que hacerlo de forma automática? Ya, ¿y qué tipo de tarea realiza mejor?
  - ¿Cómo decides qué tipo de tareas debe hacer un usuario al momento de testear el sistema? ¿Y quién es el *usuario*? ¿Un experto que ha hecho esto durante décadas? ¿O un novato que necesita que le explique qué debe hacer antes de comenzar? ¿Y los usuarios están limitados por la rapidez de su pensamiento, la habilidad de mover el *mouse* o la del computador?
  - Incluso únicamente a nivel computacional: ¿la complejidad del algoritmo depende del número de datos a mostrar o del número de pixeles a dibujar? ¿Existe un *trade-off* entre la velocidad del computador y el uso de la memoria de este mismo?

Rules of thumb

¿Qué es un *rule of thumb*?

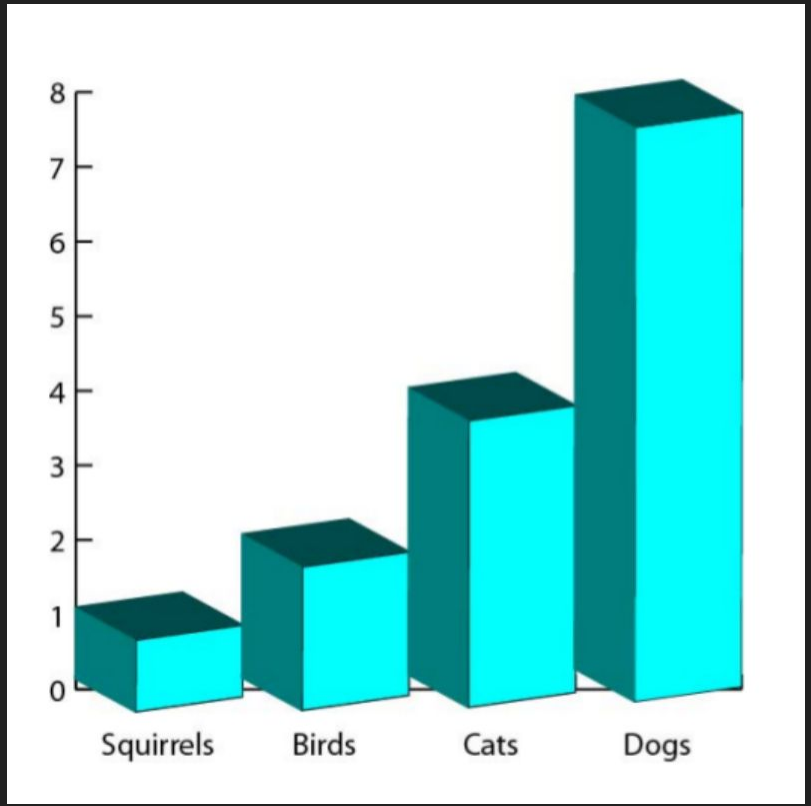
- Es posible definirlo como un principio o una guía, basado en la experiencia/práctica más que en la teoría.
- La idea de esta clase es sintetizar el conocimiento, sacado desde estudios empíricos que se desarrollan en esta área.

Data ink ratio (Tufte)

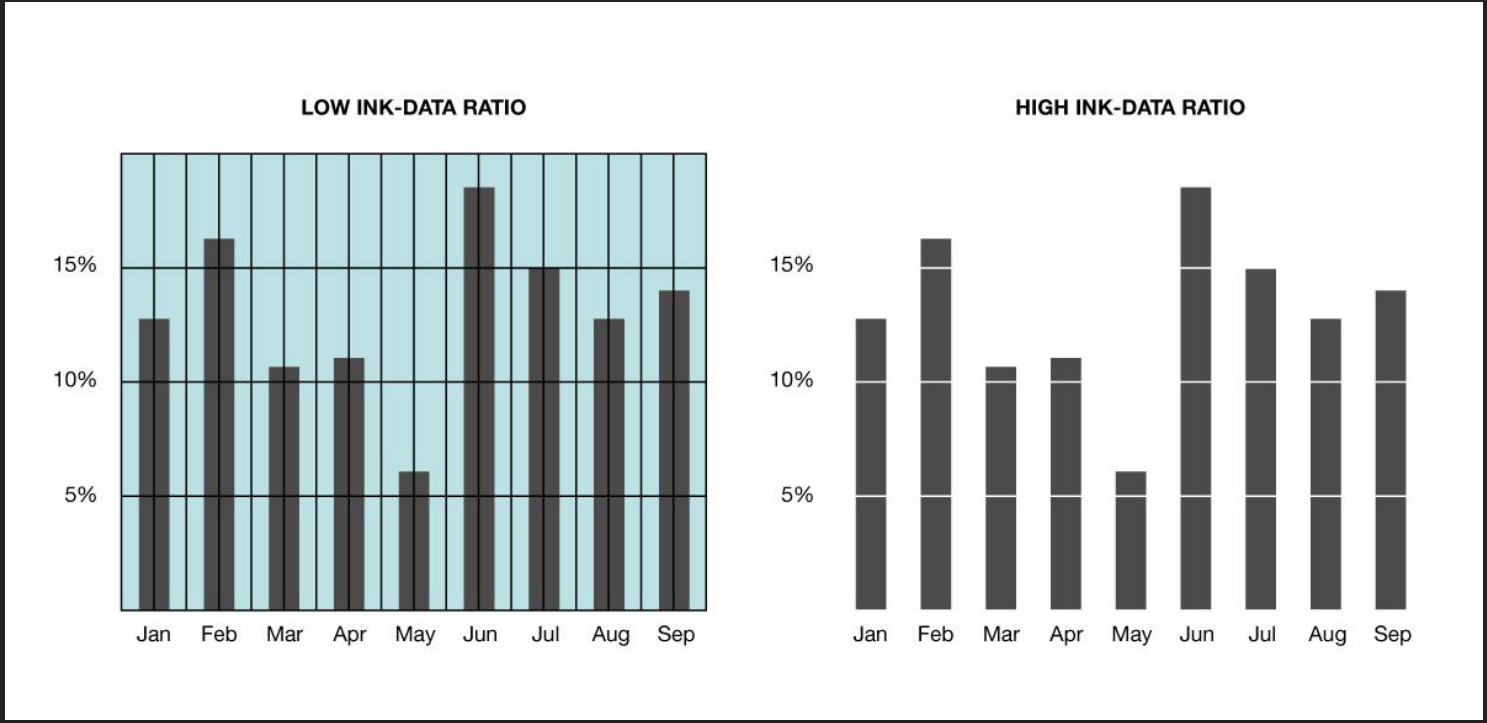
$$\text{data ink ratio} = \frac{\text{data ink}}{\text{total ink used}}$$

En nuestras visualizaciones, buscaremos **maximizar** este *ratio*, para que cada marca/canal que usemos tenga una razón de existir.
Llevándolo a un extremo (i.e. *ratio* = 1), cada pixel debe estar justificado.

Data ink ratio (Tufte)

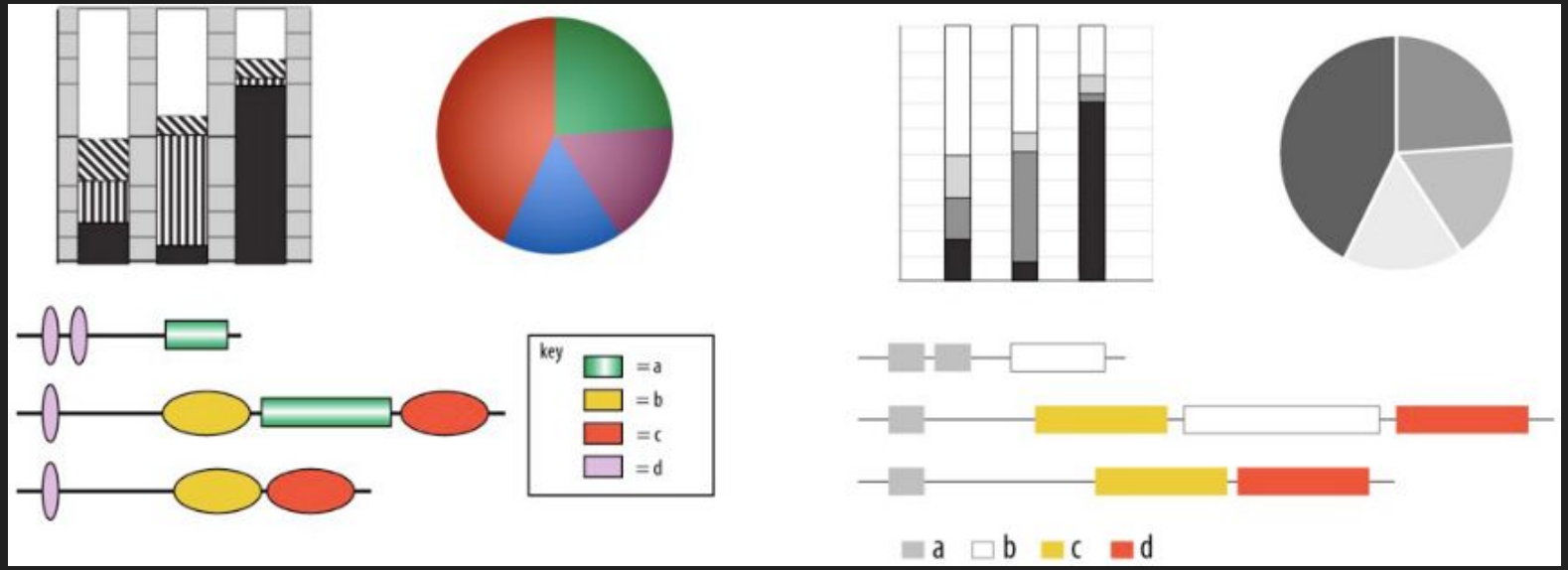


Data ink ratio (Tufte)



Data ink ratio (Tufte)

Más ejemplos



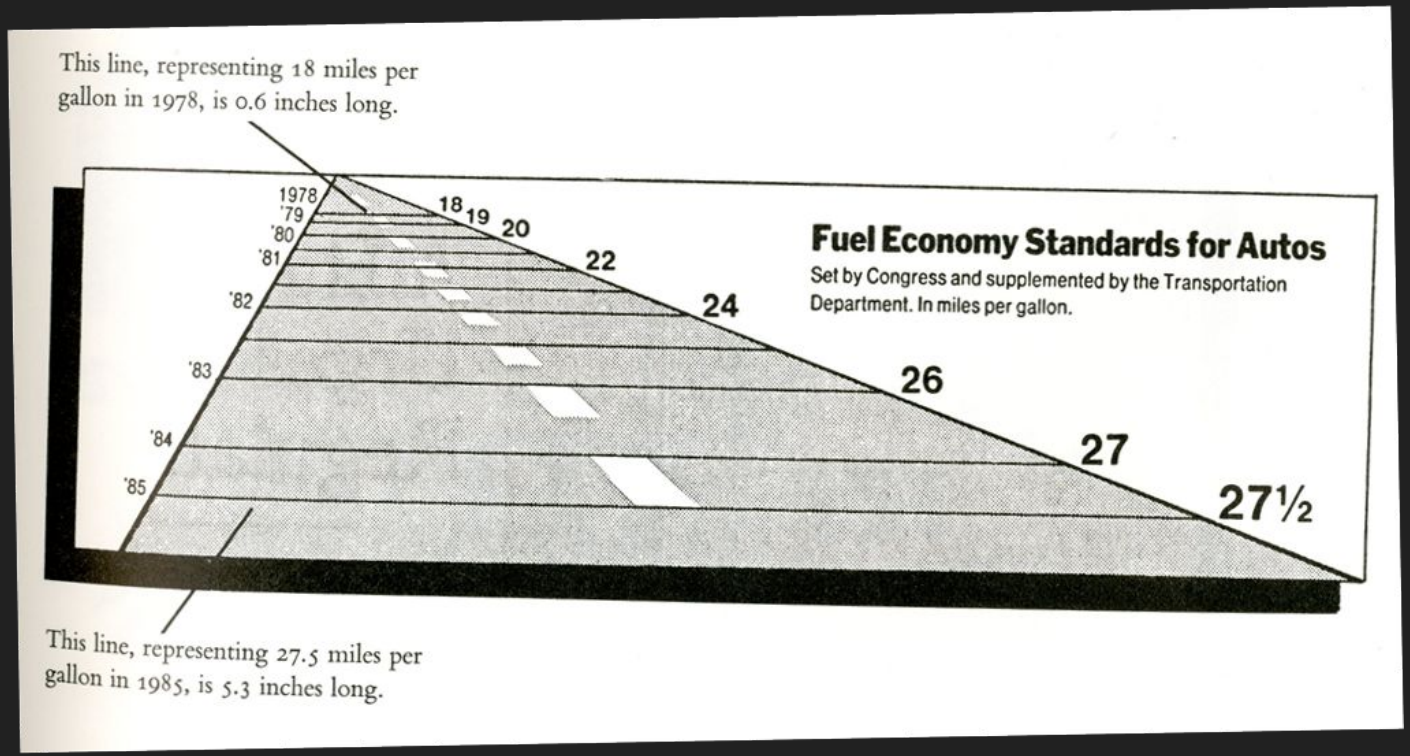
Quitamos colores, formas, texturas que no aportan información.

Lie factor (Tufte)

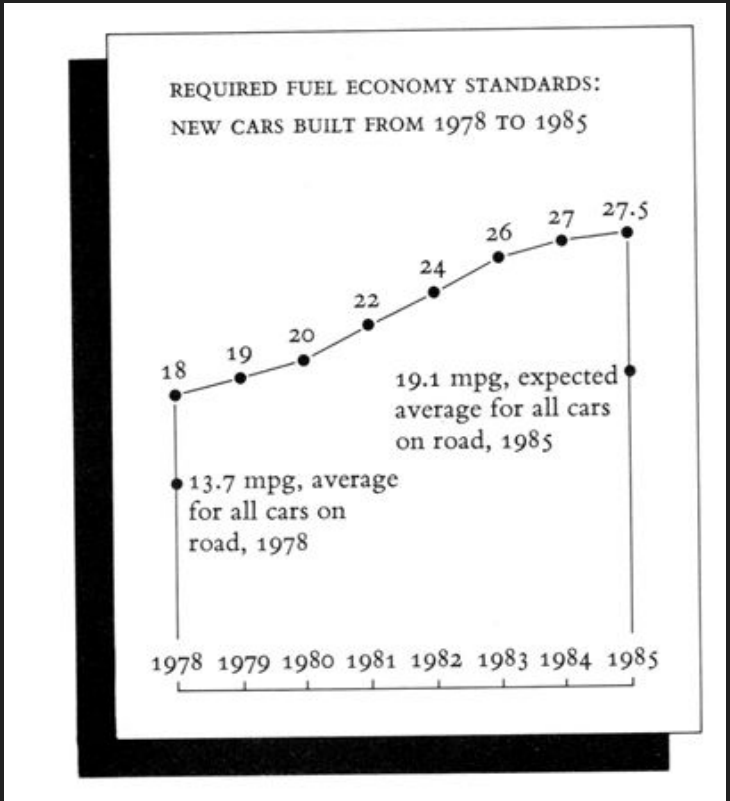
$$\text{lie factor} = \frac{\text{size effect in graphic}}{\text{size effect in data}}$$

- La tasa de cambio entre los datos debe ser fielmente reflejada por el efecto que se muestra gráficamente.
- En este caso, deberíamos apuntar a un factor de 1 (*i.e.* mismo efecto).

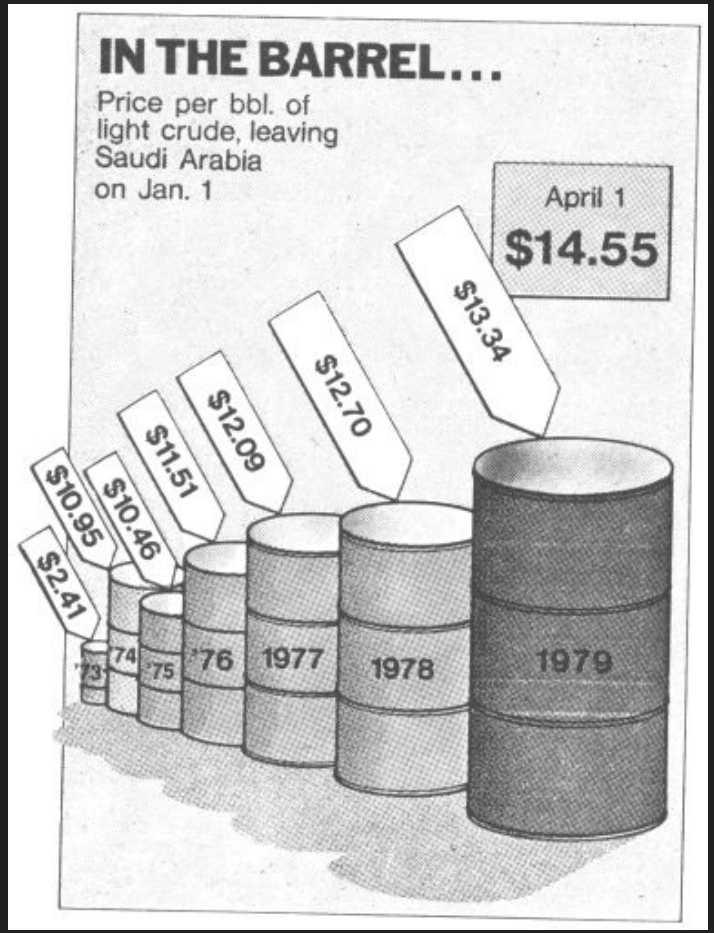
Lie factor (Tufte)



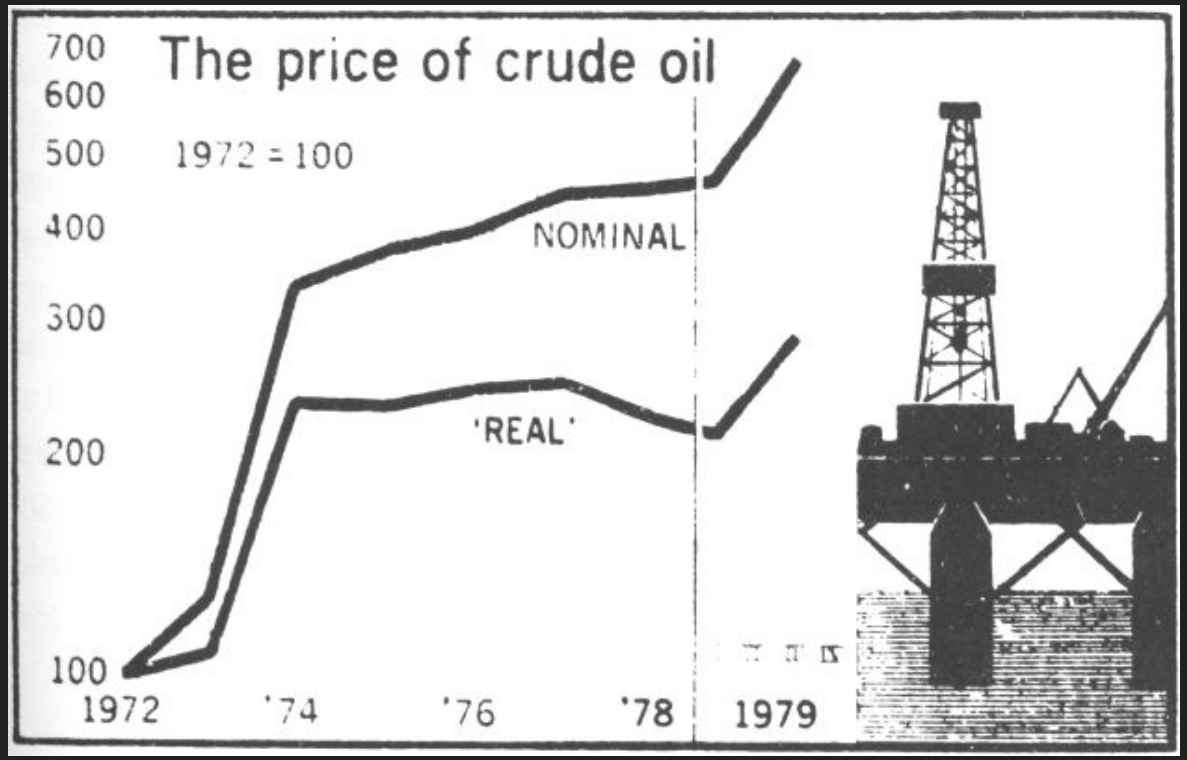
Lie factor (Tufte)



Lie factor (Tufte)

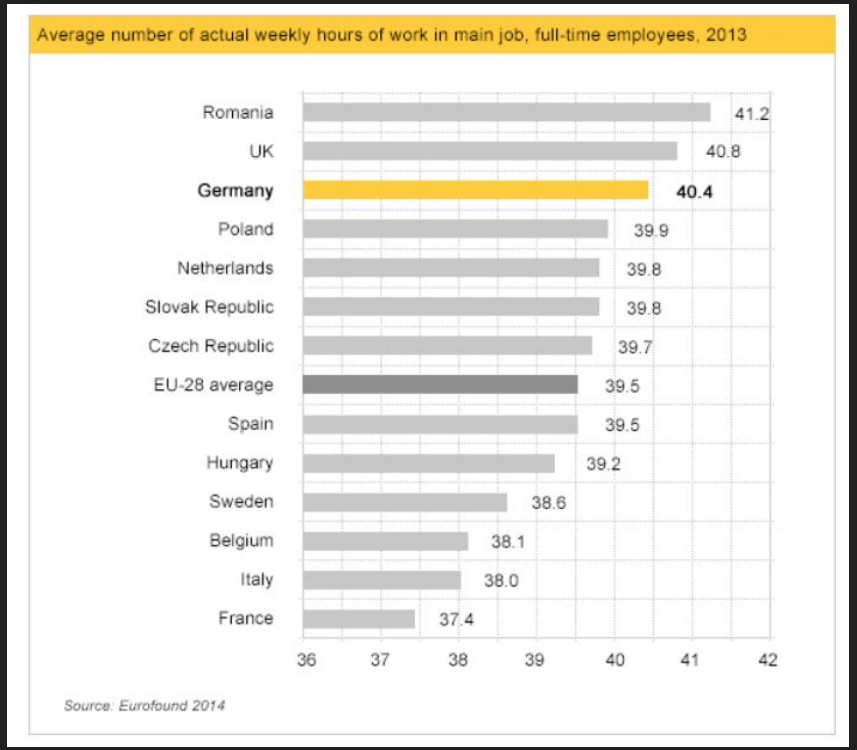


Lie factor (Tufte)



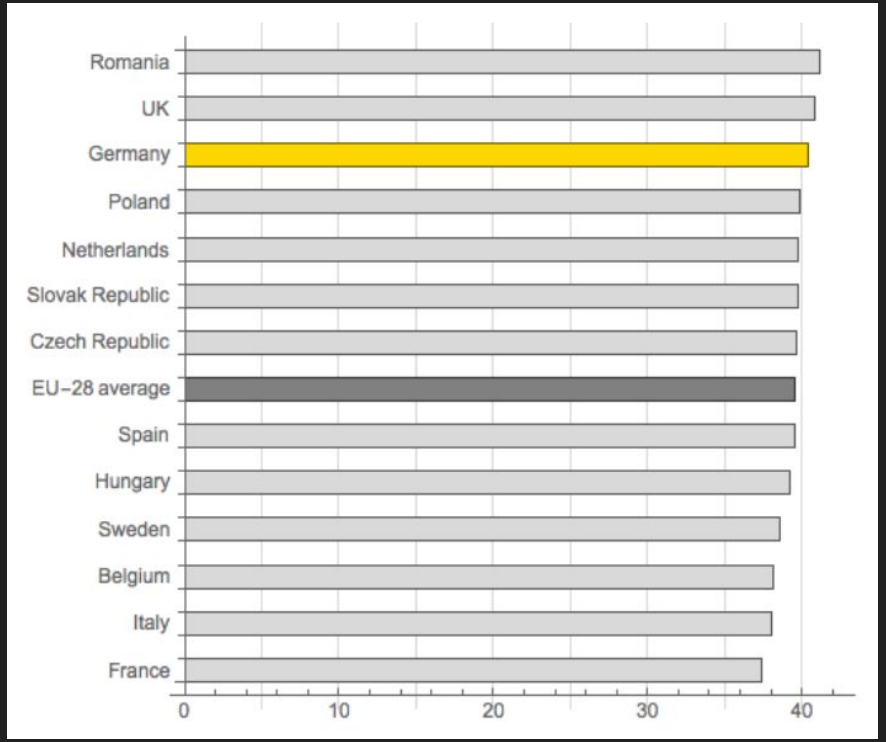
Ejes engañosos

- Relacionado con el *lie factor*, los ejes de este diagrama de barras no comienzan en cero.
- A raíz de esto, pareciera que los alemanes trabajan casi el triple que los franceses, siendo que sólo hay un factor de 1,08 de diferencia.

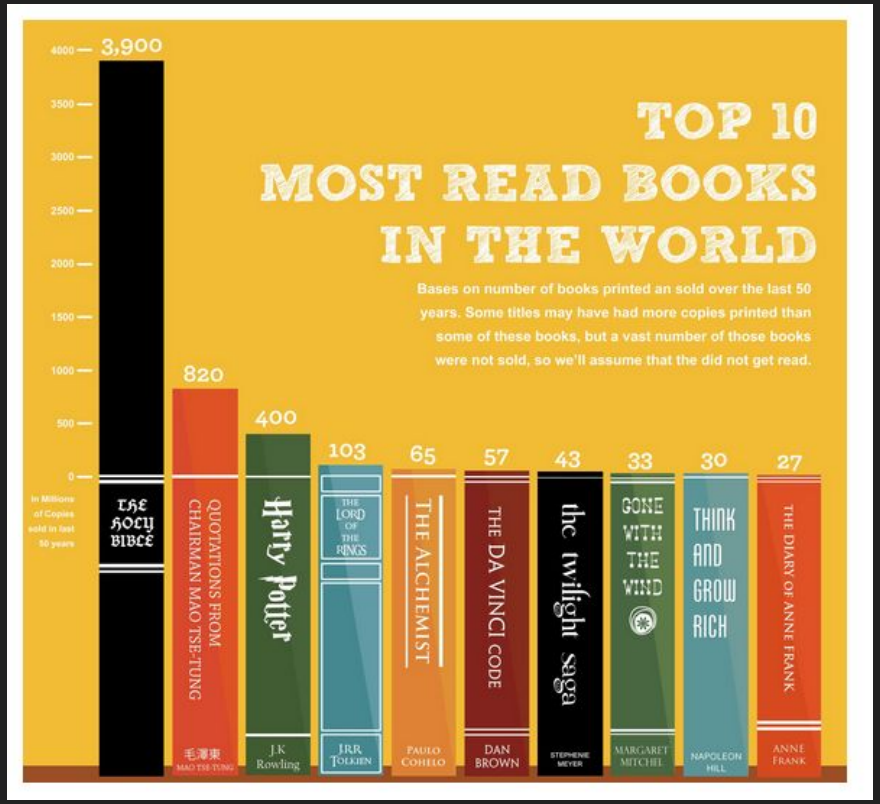


Ejes engañosos

- Aquí tenemos el mismo *dataset* representado con ejes que parten desde cero.
- Ahora las diferencias de las barras se ajustan acorde a las diferencias en el *dataset*.

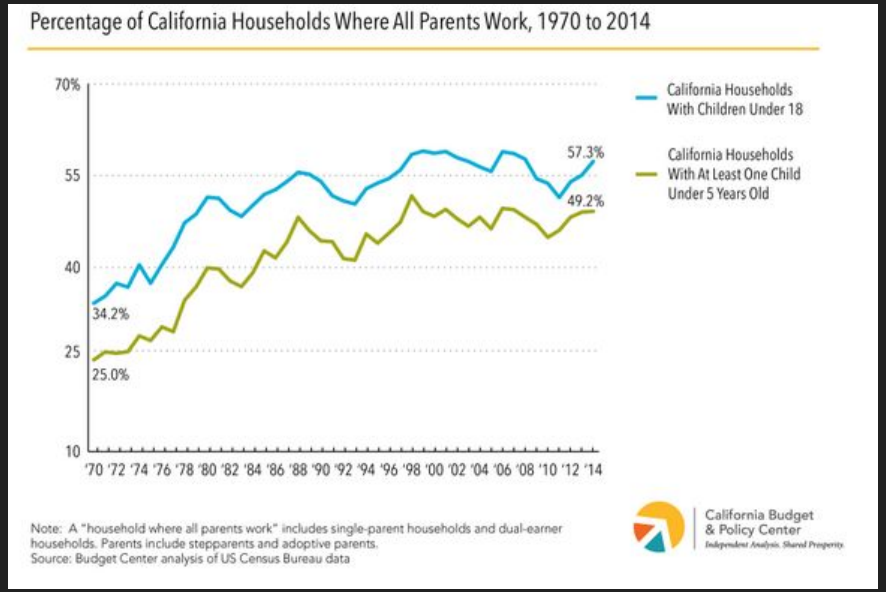


Ejes engañosos (más ejemplos)



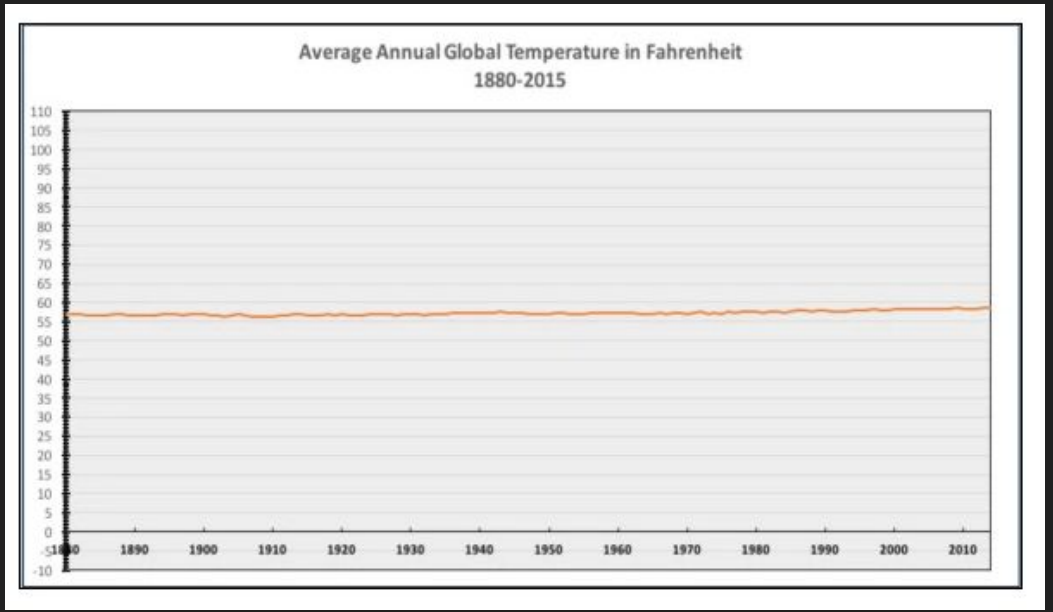
Ejes engañosos (más ejemplos)

- Este gráfico de línea no parte con su eje en cero. Sin embargo, no hay problema con eso, puesto que, a diferencia de las barras, este *line graph* busca contar **otra historia**.
- El diagrama de barra se enfoca en la diferencia de magnitud entre las categorías, mientras que acá buscamos mostrar el cambio de la variable dependiente (eje y).



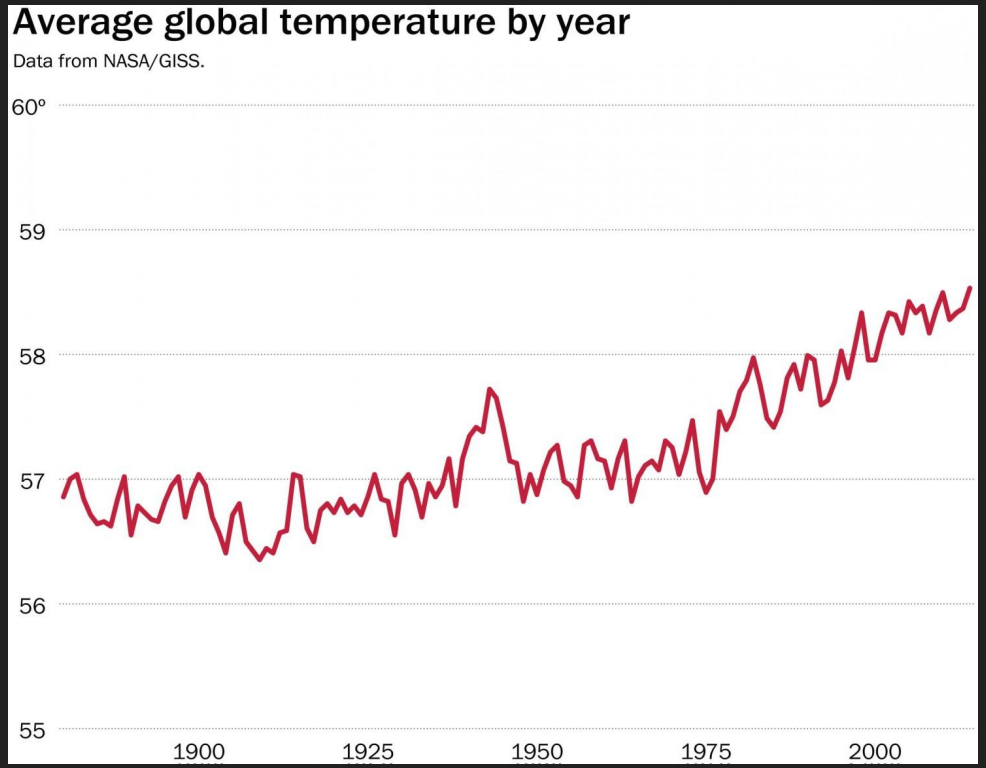
Ejes engañosos (más ejemplos)

- Incluso, mostrar el cero en el eje puede ser engañoso, ya que se intentan ocultar la tasa con que ocurren los cambios.
- En este caso, este gráfico no muestra los cambios, sino la magnitud absoluta que es inconsistente con la historia que estaban tratando de narrar.



Ejes engañosos (más ejemplos)

- Esta es una representación más correcta de lo que ocurre con la temperatura promedio global de nuestro planeta, ya que no es importante la magnitud, sino el cambio.

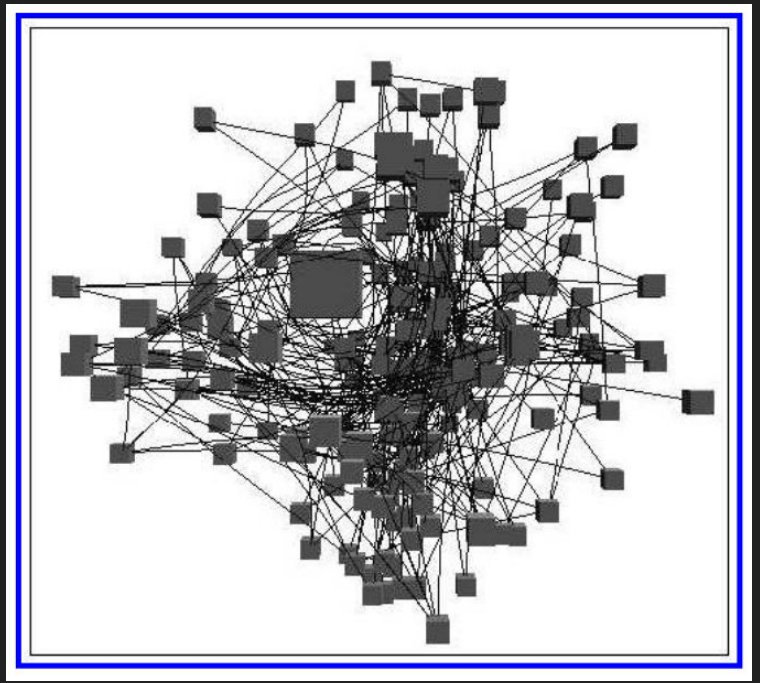


No al 3D injustificado

- Generalmente, las personas creen que si en dos dimensiones se ve bien, entonces en tres debe ser mejor aún —claro, después de todo, vivimos en un mundo tridimensional.
- Sin embargo, existen muchas **dificultades** relacionadas al *encoding* de información usando una tercera dimensión espacial (i.e. profundidad), ya que tiene diferencias importantes con respecto a las otras dos dimensiones.
- El uso de una tercera dimensión sí se justifica cuando el usuario debe ejecutar tareas (*e.g.* percepción de formas 3D) que están relacionadas con una estructura que inherentemente tiene tres dimensiones.

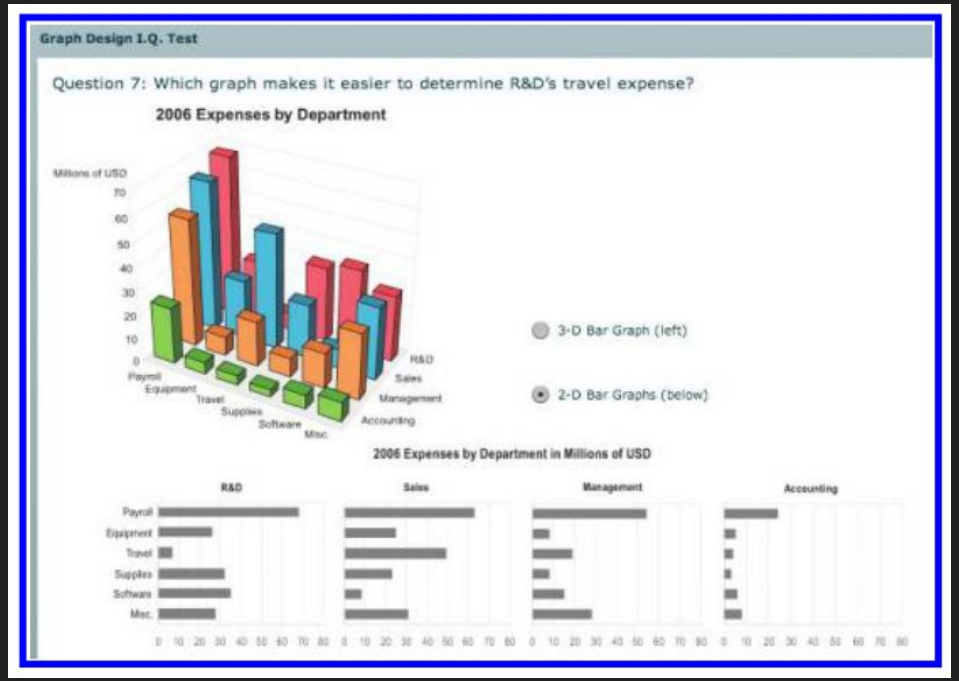
No al 3D injustificado (oclusión)

- El problema de este grafo con nodos y aristas es el de oclusión, en donde nodos quedan ocultos detrás de otros.
- Si bien es posible agregar algún tipo de navegación interactiva, el costo asociado a esto es el tiempo.

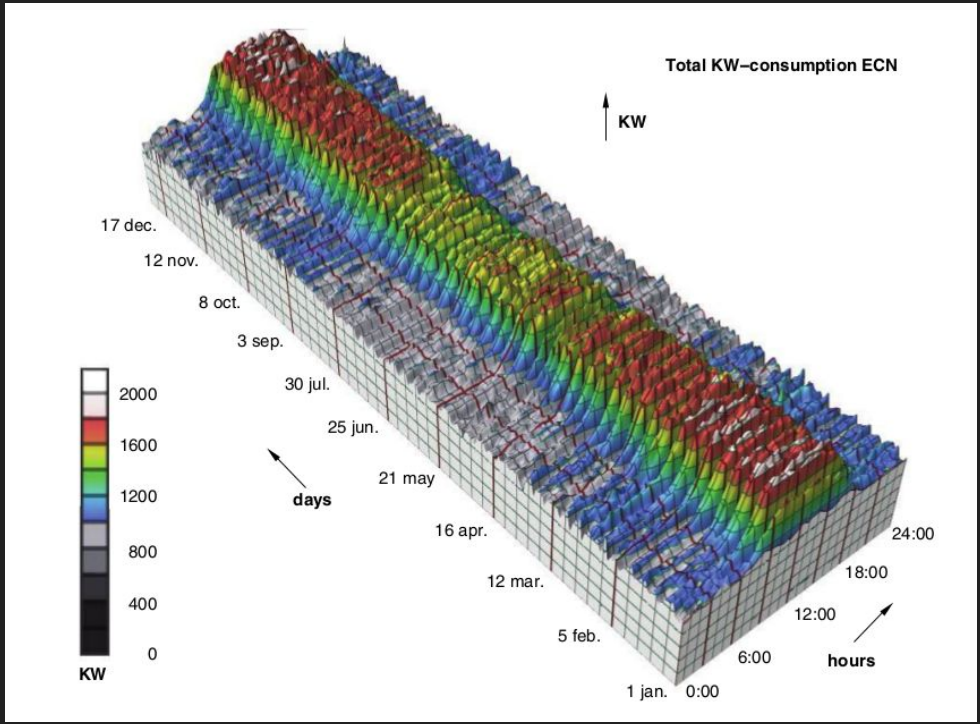


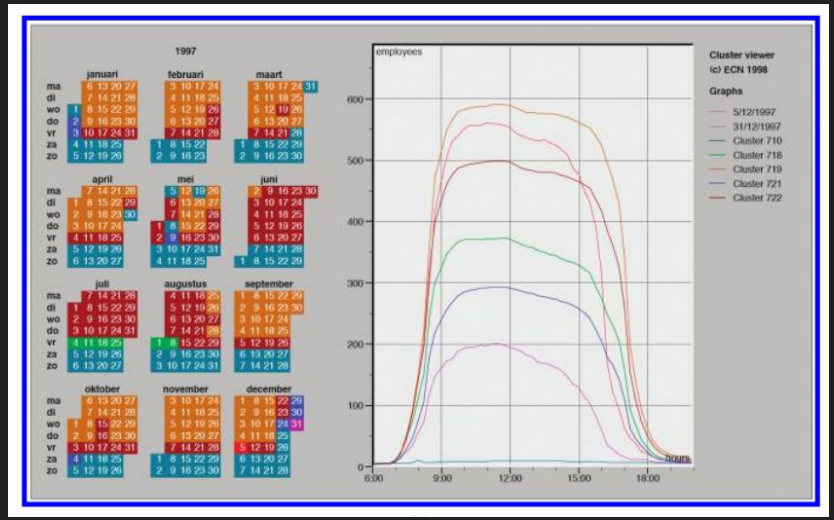
No al 3D injustificado (distorsión por perspectiva)

- Los objetos que están a mayor distancia se perciben más pequeños (*e.g.* línea de tren).
- En este ejemplo, tenemos dos variables categóricas y otra numérica; sin embargo, usar 3D no es una buena opción.
- Por la perspectiva (y también por oclusión), cuesta comparar los tamaños de las barras.



No al 3D injustificado (buscar alternativas)





No al 3D injustificado (buscar alternativas)

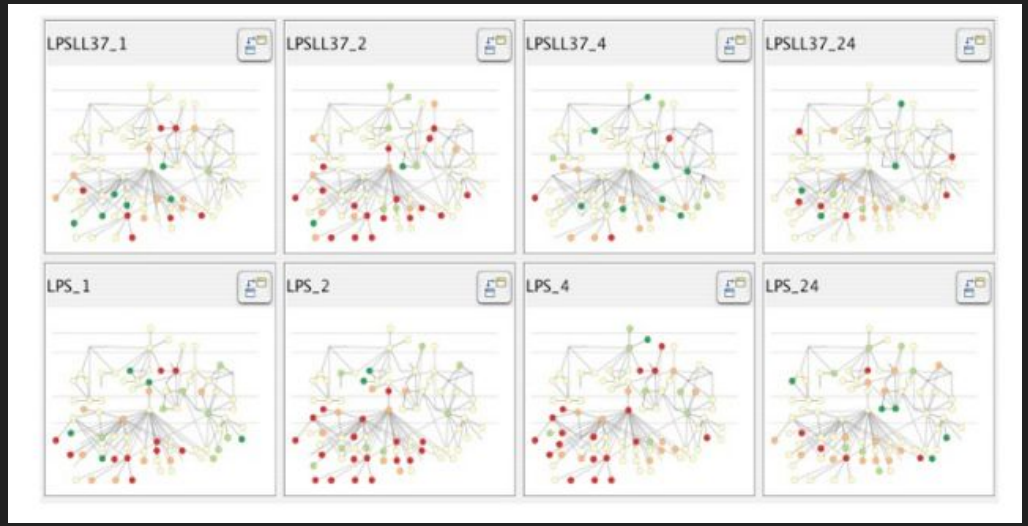
- En el ejemplo de la izquierda, hay problemas de **oclusión** y **distorsión**. Además, sólo es posible notar que existe un cambio en las horas de trabajo y la variación por estación entre verano e invierno.
- En el de la derecha, se crearon nuevos datos a partir de un *clustering* sobre las curvas más parecidas, logrando un promedio entre ellas. Aquí no hay oclusión ni distorsión entre las curvas, lo que permite una rápida comparación de ellas.
- Asimismo, el calendario ya es una marca tradicional y exitosa para mostrar patrones temporales.

No al 2D injustificado

- De forma análoga, mostrar datos en plano también debe ser justificado, comparado con la alternativa de una única dimensión (*e.g.* una lista).
- Las listas tienen varias ventajas: pueden exponer información (como etiquetas de texto) en un espacio mínimo; además, las listas son una herramienta excelente para tareas de *lookup*, cuando están ordenadas de forma apropiada.
- Por ejemplo, buscar un *label* específico será más fácil en una lista, que una representación 2D de *node-link*, en donde el usuario tendrá que buscar nodo por nodo. Sin embargo, si la tarea realmente requiere entender la estructura topológica de la red, entonces, en ese caso, sí vale la pena mostrar las relaciones en el plano.

Eyes beat memory

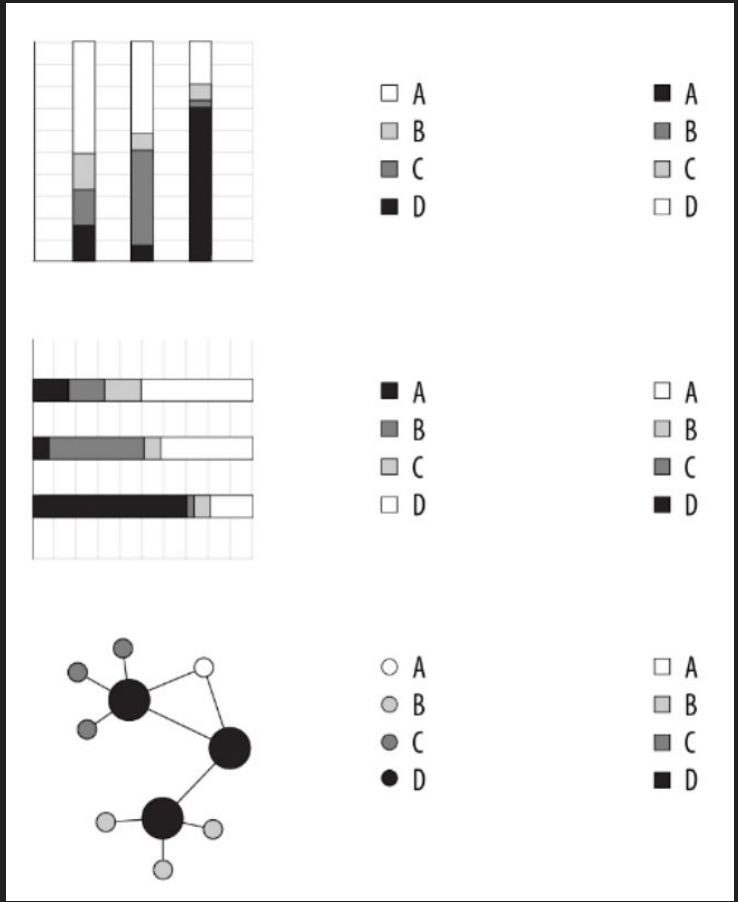
- Es más fácil usar ***external cognition*** que nuestra **memoria interna**.
- Por lo tanto, es más fácil comparar, moviendo nuestros ojos de lado a lado, que hacerlo tratando de recordar algo que vimos recientemente.
- Ejemplo: un gráfico con diferentes instancias, variando el color según el experimento.



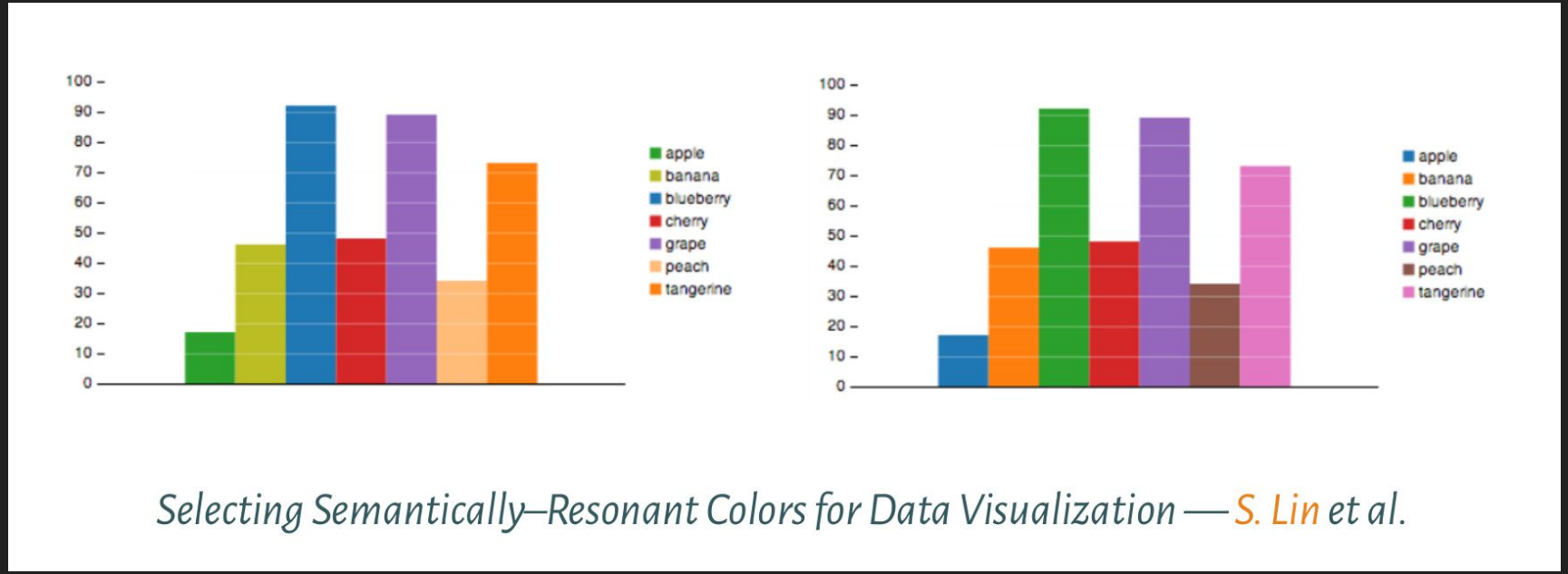
Primero el fondo, luego la forma

- Las mejores visualizaciones deben destacar tanto en funcionalidad como en forma: deben ser efectivas y agradables al ojo humano.
- Sin embargo, es mejor enfocarse, primero, en conseguir un diseño efectivo y quizá tosco, porque es posible refinarlo en forma más tarde, mientras se mantiene la efectividad.
- Al contrario, dado un diseño bello pero inefectivo, probablemente se tendrá que rehacer desde cero.

Consistencia



Consistencia



Selecting Semantically–Resonant Colors for Data Visualization — S. Lin et al.

Más *guidelines*

- Las unidades deben ser estandarizadas (por ejemplo, el dinero)
- Las dimensiones del gráfico no deberían exceder la de los datos
- Los datos deben ser mostrados en su contexto
- Ojo con el daltonismo
- Ojo con la tipografía
- Ofrecer *responsiveness*
- Gráficos autoexplicativos (mensajes claros, sin muchas abrevaciones)

Framework

Tres preguntas: qué, por qué, cómo

Partamos con **qué**

- Tipos de datos
- Tipos de *datasets*

Tres preguntas: qué, por qué, cómo

- Muchos aspectos que guían el diseño de una visualización son impulsados por el tipo de datos que tenemos a nuestra disposición.
- Hay que preguntarse, entonces, qué tipo de datos tenemos, qué información podemos obtener directamente, y qué sentido tienen realmente.

Semántica de los datos

14, 2.8, 30, 30, 15, 1001

Semántica de los datos

Santiago, 3, N, Nacimiento

Semántica de los datos

- Para salir de las adivinanzas, es necesario saber dos tipos de información: la semántica y el tipo de dato.
- La semántica es su significado en el mundo real (¿qué es? ¿un nombre de una persona, una ciudad, una abreviación de un punto cardinal?)
- El tipo de dato es interpretación estructural o matemática del dato (¿es un ítem, un enlace o un atributo?)
- Por ejemplo, si tenemos un número que representa cajas de azúcar, sí hace sentido sumarlas, ya que estamos hablando de una cantidad. Por otra parte, si el número fue el código postal, no tiene sentido sumarlos, ya que no es una cantidad, sino un código.
- A veces, se necesita leer información adicional (conocida como *metadata*) para poder interpretar correctamente un dato.

Dataset and data types

Data and Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Data Types

→ Items → Attributes → Links → Positions → Grids

Dataset Availability

→ Static → Dynamic

Tipos de dato (*data types*)

Según Munzner (2014), hay cinco tipos básicos de datos:

- Atributos
- Ítems
- Vínculos
- Posiciones
- Grillas

Atributos

- Es una propiedad específica que puede ser medida, observada o registrada.
- Por ejemplo: temperatura, salario, precio, número de ventas, etcétera.
- También se le conoce como *variable* o *dimensión*.

Ítems

- Es una entidad discreta (e.g. fila en una tabla, nodo en un grafo).
- Por ejemplo: personas, ciudades, tiendas de computación.

Vínculos

- Es una relación entre los ítems, generalmente en un grafo.

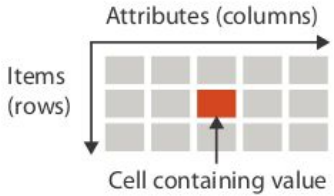
Posiciones

- Es un dato *espacial*, que provee una ubicación en un espacio 2D o 3D.
- Por ejemplo: un par latitud-longitud mostrando una ubicación en la Tierra, o también podría ser la ubicación en la región de un escáner médico.

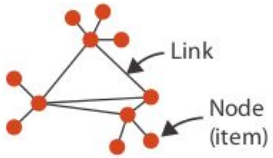
Tipos de *datasets*

Dataset Types

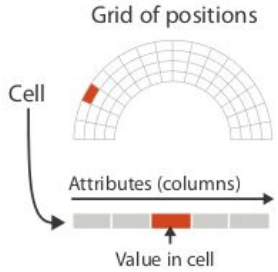
→ Tables



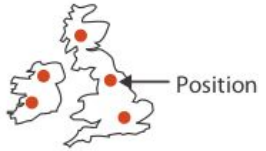
→ Networks



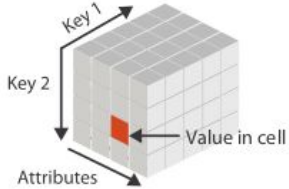
→ Fields (Continuous)



→ Geometry (Spatial)



→ *Multidimensional Table*



→ *Trees*

Tipos de *dataset* (*dataset types*)

Según Munzner, hay cuatro tipos básicos de *datasets*:

- Tablas
- Redes (grafos) y árboles
- Campos (*fields*)
- Geometría

Cada uno de ellos, está compuesto por los cinco tipos de dato recién vistos.

Tablas

- Es el tipo de *dataset* más común.
- Viene en forma de filas y columnas (*e.g. spreadsheet*).
- Los tipos de datos son: ítems y atributos
  - Generalmente, una fila representa un ítem,
  - Y una columna representa un atributo.
- Cada celda de la tabla es un **valor** para la combinación ítem-atributo.
- Además, existen las tablas multidimensionales, que tienen múltiples llaves.

Redes y árboles

- Este tipo de *dataset* es apropiado para mostrar que existe algún tipo de **relación** entre dos o más ítems.
- Un ítem en una red es llamado nodo o vértice.
- Una relación entre dos o más nodos se llama enlace o vínculo.
- Por ejemplo, las personas pueden ser representadas como nodos y su relación de amistad entre ellas como vínculos.
- Adicionalmente, es posible asociar atributos a cada nodo y enlace.
- Un árbol es un caso específico de un grafo, en donde no existen ciclos (*e.g.* árbol de jerarquía en una organización).
- Es importante distinguir que nos referimos al concepto abstracto de una red y no a un *layout* en particular (con las posiciones en el espacio) de esta red.

Geometría

- Habla sobre la forma de ítems con posiciones explícitas.
- Los ítems pueden ser puntos, curvas, superficies o volúmenes.
- Los *datasets* geométricos son intrínsecamente espaciales.
- Este tipo de *dataset* puede que no tenga atributos, a diferencia del resto.
  - Aquí es interesante saber cómo codificar información
- También es necesario saber con qué nivel de detalle se generan las formas (*shapes*) desde datos geográficos crudos.
  - Por ejemplo, la frontera de un bosque, o de una ciudad, o también la curva de una carretera

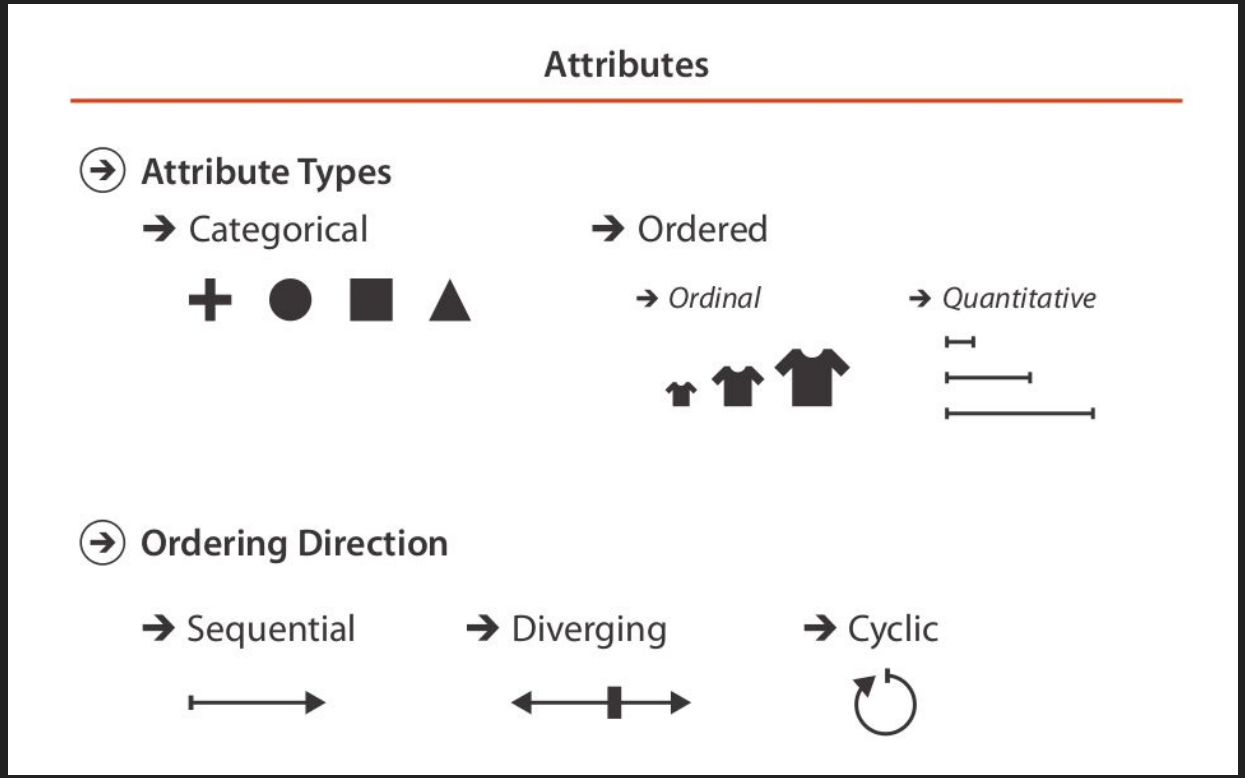
Otros tipos de *dataset*

- Existen múltiples formas de agrupar ítems, además de una tabla.
  - Un conjunto (*set*) es grupo sin orden de ítems
  - Una lista (*list*, *array*) es un grupo ordenado de ítems
  - Un clúster (*cluster*) es un grupo basado en la similaridad de un atributo específico
- También se pueden construir estructuras a partir de un grafo.
  - Por ejemplo, se pueden mostrar **caminos**, que son listas de vínculos que conectan nodos
  - O podríamos tener también un ***compound network***, que es una red que tiene asociado un árbol: todos los nodos de la red son las hojas del árbol, mientras que los nodos interiores del árbol proveen cierta estructura jerárquica para estos nodos de la red.
- Además, es posible crear estructuras híbridas y más complejas que intentan modelar aplicaciones de la vida real: esto es sólo un punto inicial del análisis de *data abstraction*.

Disponibilidad del *dataset*

- Existen dos categorías: *datasets* **estáticos** y *datasets* **dinámicos**
  - Estático (*offline*) es cuando el *dataset* está disponible *all at once* (*i.e.* todo en un instante)
  - Dinámico (*online*) es cuando nueva información llega a través del tiempo (*streaming data*)
- Cuando el *dataset* es dinámico, nuevos datos pueden ser agregados, otros eliminados o también actualizados.
- Esto agrega complejidad en varios aspectos al proceso de visualización comparado a un dataset estático.

Tipos de atributos



Tipos de atributos: categóricos

- La primera distinción que haremos entre los datos son los de tipo categóricos (o también conocidos como nominales).
- No tienen un orden implícito, pero generalmente sí existe una jerarquía.
- Podrían, eso sí, ser ordenados de forma arbitraria por datos externos.
- Ejemplo: nombres de frutas.

Tipos de atributos: ordenados

- Los datos que no son categóricos, se conocen como datos ordenados.
- Esto puede ser subdividido en: datos ordinales y datos cuantitativos.
- En los datos ordinales, no existe una aritmética bien definida entre sus componentes, pero sí es posible ofrecer un orden (*e.g.* tallas de poleras)
- Por otra parte, en los cuantitativos, existe una magnitud que sí permite una comparación aritmética. Ejemplos: altura, peso, temperatura, etcétera.

Tipos de atributos: secuencial, divergente o cíclico

- Entre los datos ordenados, podemos distinguir los datos secuenciales, en donde existe un rango homogéneo desde un valor mínimo hasta uno máximo (ejemplo: altura de montañas, que va desde el nivel del mar hasta el Everest).
- Por otra parte, también podemos hablar de datos divergentes, que puede ser descompuesto en dos secuencias que van en direcciones opuestas, que se encuentran en un punto en común: el cero (ejemplo: un *dataset* de elevación, en donde los valores van hacia arriba para las montañas y hacia abajo para los valles submarinos, siendo el nivel del mar, el valor cero).
- Por último, podrían ser cíclicos, en donde los valores *wrap around* hacia el punto inicial, en vez de crecer indefinidamente.

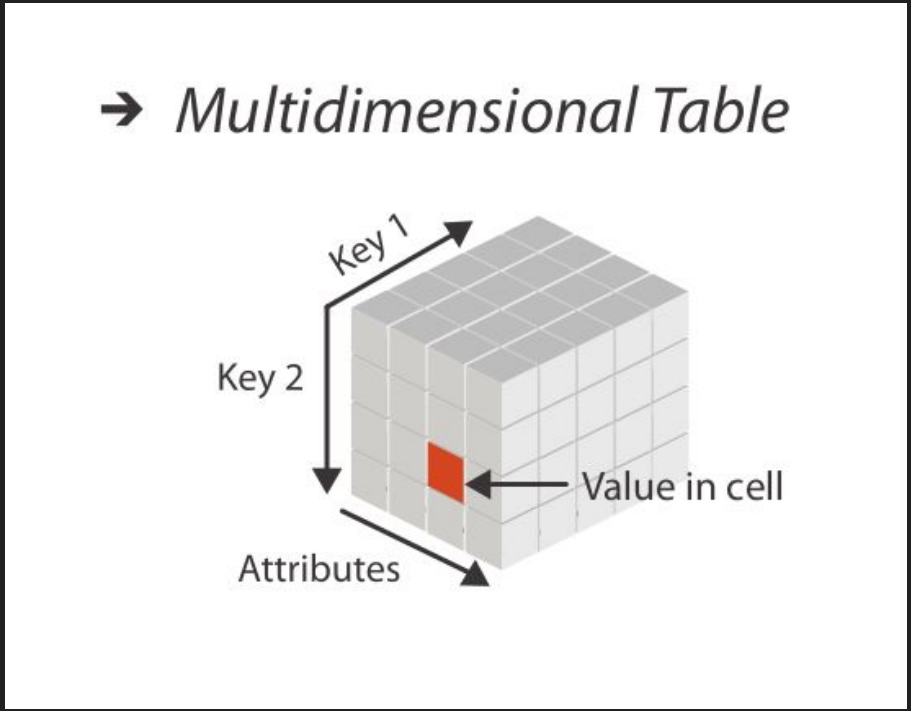
Atributos jerárquicos

- Puede existir una estructura jerárquica entre uno o múltiples atributos.
- Por ejemplo, los precios de acciones recolectados a lo largo de una década es un ejemplo de un *time-series dataset*, en donde uno de los atributos es el tiempo. Este atributo puede ser agregado de forma jerárquica.
- Muchos tipos de datos tienen esta propiedad: por ejemplo, el atributo geográfico de un código postal podría ser agregado a nivel de ciudades, como de regiones, o incluso países.

Semántica

- Saber el tipo de dato de un atributo no nos habla de su semántica, ya que son preguntas independientes: uno no impone el significado del otro.
- *Key versus value*: una llave se considera como un atributo independiente, en donde esta distinción es importante en un *dataset* tabular. Por otra parte, el valor vendría siendo el valor dependiente de la llave.
- En una tabla plana, tenemos una única llave, en donde cada ítem corresponde a una fila de la tabla. En este caso, la llave puede estar implícita.
- Generalizando a una tabla multidimensional, la llave puede ser considerada como múltiples atributos, en donde cada combinación debe ser única.
- En los campos (si bien no son discretos como las tablas) también podemos hablar de llaves y valores: tenemos campos escalares, vectoriales y tensores.

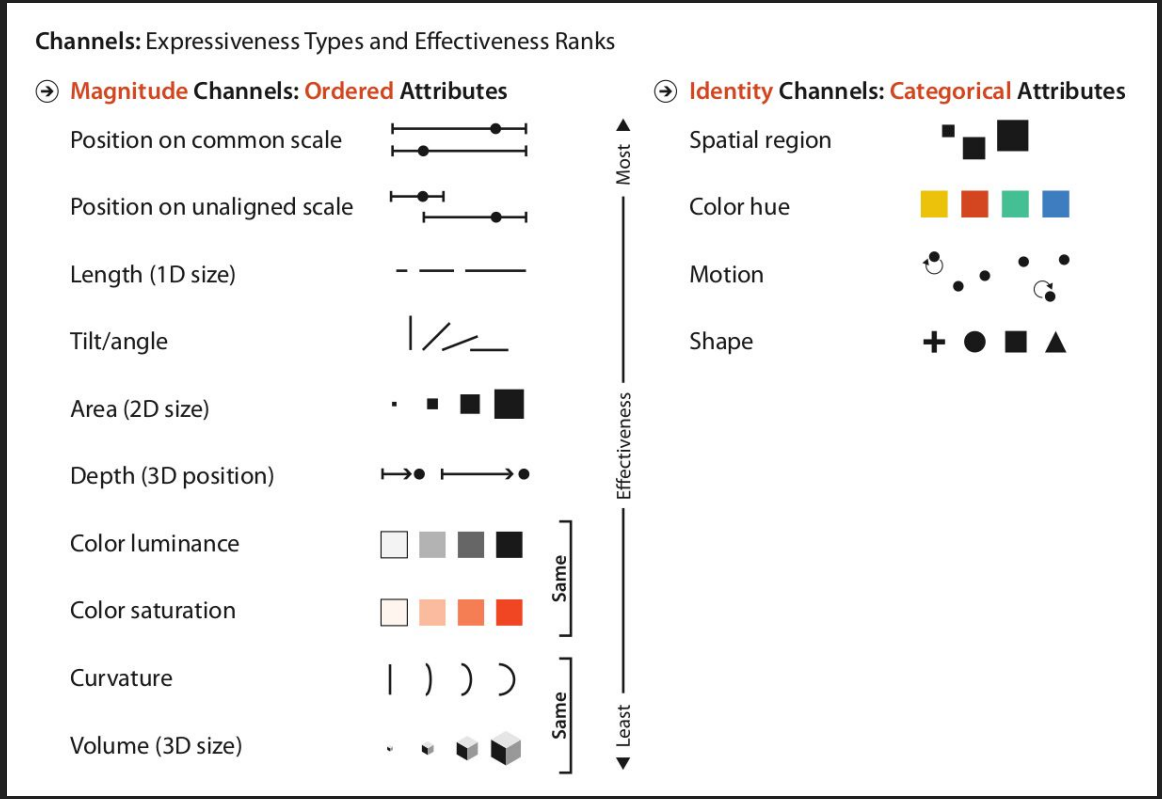
Semántica


➔ Multidimensional Table
Key 1
Key 2
Value in cell
Attributes

Semántica temporal

- Igualmente, es importante distinguir una semántica temporal en los datos, que es cualquier tipo de información que se relacione con el tiempo.
- No es sencillo manejar un *dataset* con una semántica temporal, dada la riqueza jerárquica que tiene el tiempo, tanto como la posible periodicidad
- Además, también existen algunos problemas con las escalas, ya que no calzan perfectamente (*e.g.* semanas en un mes).
- Puede ser considerado como un atributo cuantitativo (ya que es posible hacer aritmética con el tiempo), pero si la duración no es de interés, entonces podemos tratarlo como un atributo ordenado.

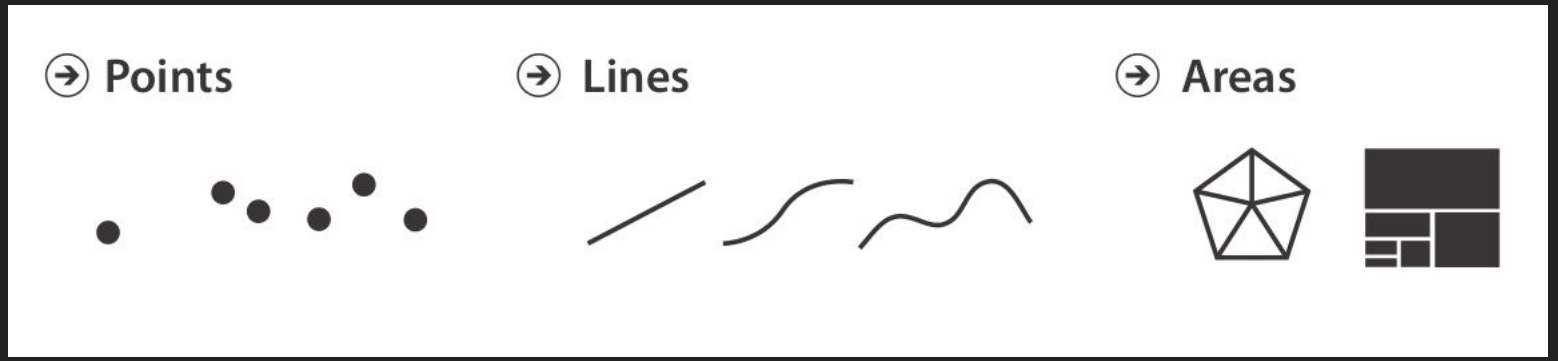
Overview


Channels: Expressiveness Types and Effectiveness Ranks
Magnitude Channels: Ordered Attributes
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Most
Effectiveness
Least
Identity Channels: Categorical Attributes
Spatial region
Color hue
Motion
Shape

Marcas & canales

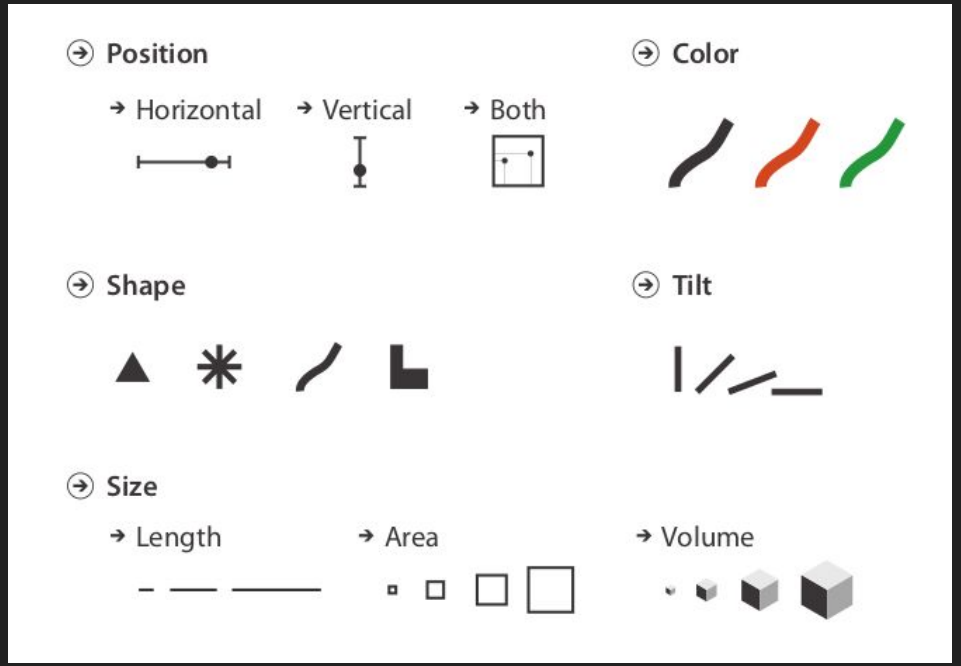
Definiciones

- Una **marca** es un elemento geométrico básico, que puede ser clasificado según el número de dimensiones espaciales que requiera.



Definiciones

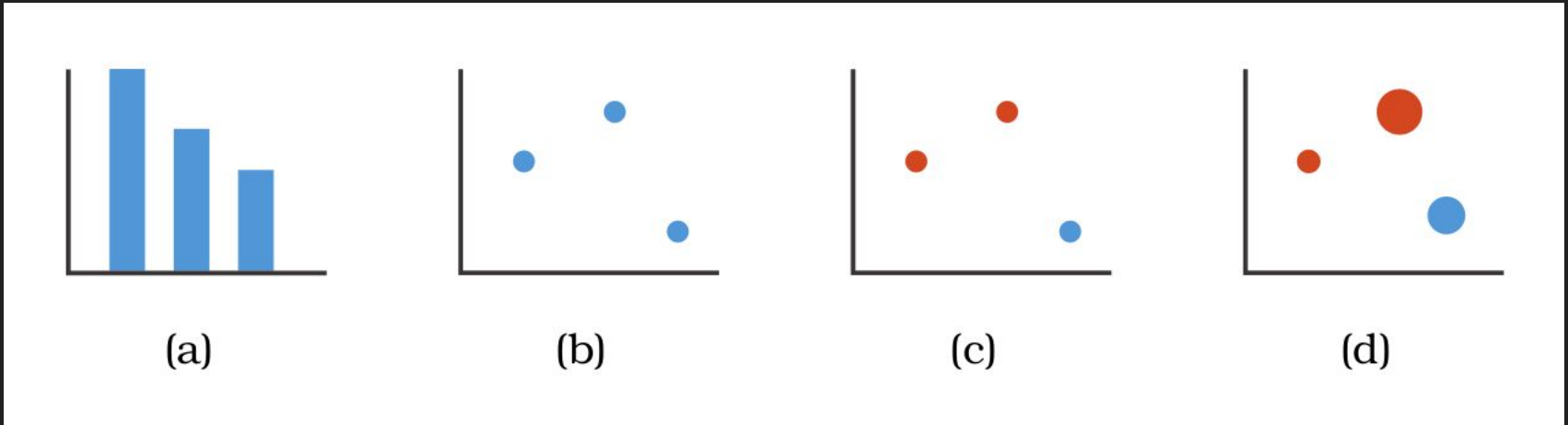
- Un **canal** visual permite controlar la apariencia de las marcas, independientemente de la dimensionalidad de este elemento primitivo.
- Entre los canales más comunes, tenemos: color (*i.e.* saturación, brillo, *hue*), tamaño, ángulo, curvatura, forma, entre otras más.



¿Por qué?

- La idea de razonar en términos de *marcas* y *canales* nos entrega los **bloques elementales** para analizar los *visual encodings*.
- El diseño de estos *visual encodings* pueden, entonces, ser descritos como una **combinación ortogonal** de ambos aspectos: elementos gráficos (marcas) y sus apariencias (canales).
- De esta forma, incluso los *encodings* complejos pueden ser **desglosados en componentes** más simples que, a su vez, pueden ser analizados en términos de sus marcas y canales.

Ejemplos



Ejemplos

- En los ejemplos anteriores, cada atributo fue codificado con un único canal.
- Múltiples canales pueden ser combinados de forma redundante para mostrar el mismo atributo; sin embargo, esto gasta innecesariamente canales que podrían ser utilizados para denotar futuros atributos.
- Por otra parte, existen marcas que no *deberían* recibir ciertos canales debido a su naturaleza.
    - Por ejemplo, el área de una comuna en un mapa, generalmente, está restringida a su forma geográfica. Sin embargo, existen excepciones como el *cartogram*.
    - O también en el ejemplo a), no es posible agregarle un *encoding* de tamaño vertical a las barras, porque ese canal ya *está tomado*.

Tipos de canales

- El sistema de percepción humano tiene dos tipos de modalidades:
  - El ***identity channel*** permite discernir información sobre **qué** es algo o **dónde** se encuentra;
  - El ***magnitude channel***, por otra parte, nos permite saber **cuánto** de ese algo existe.
- Con estas dos modalidades, podemos saber, por ejemplo:
  - ¿**qué** figura es? ¿un círculo, un triángulo, una cruz o un heptágono? **[identity]**
  - ¿de **qué** *hue* es? ¿rojo, verde, caqui o gris? **[identity]**
  - ¿**cuánta** saturación tiene ese azul? ¿celeste, azul marino o turquí? **[magnitude]**
  - ¿**dónde** está? ¿en qué región se encuentra la marca? **[identity]**
  - ¿**qué tan** larga es aquella línea con respecto a esta? **[magnitude]**
  - ¿**cuánto** espacio hay entre ambos rectángulos? **[magnitude]**

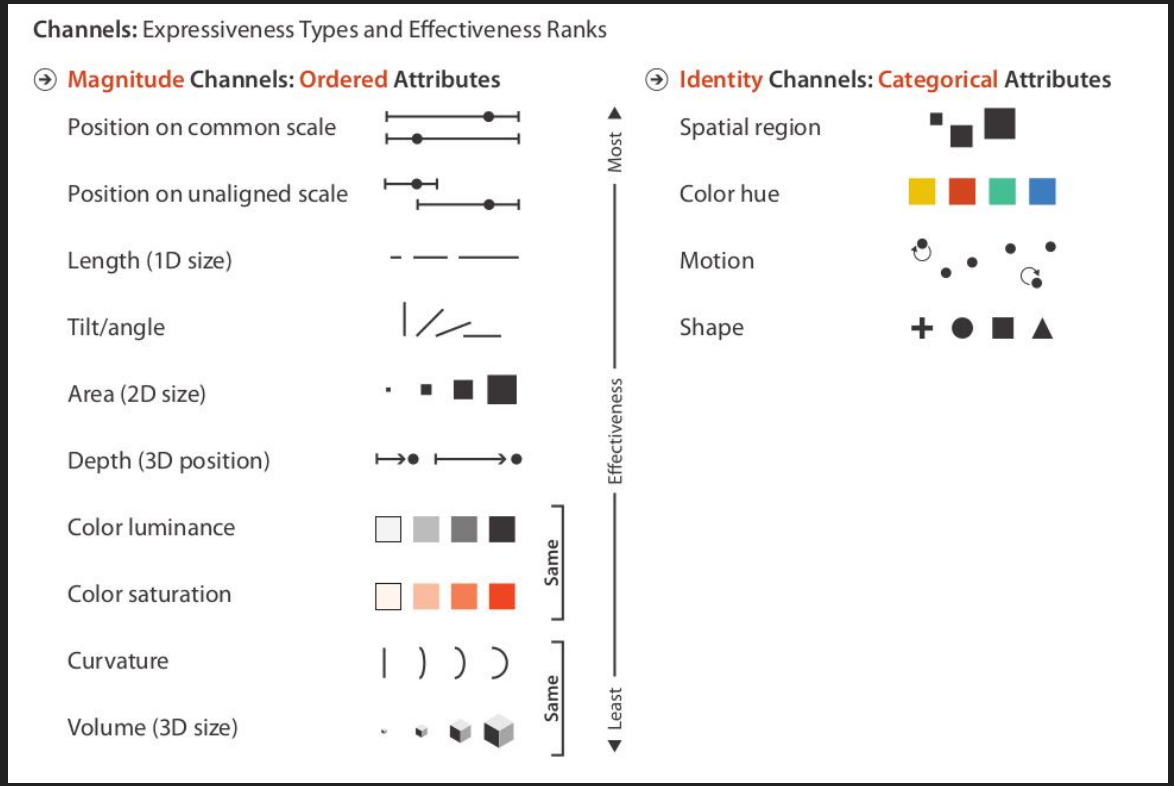
Tipos de marcas

- En los ejemplos vistos hasta ahora, cada marca ha representado un ítem de un *dataset* tabular. Sin embargo, en *datasets* de redes, también podemos usar marcas para representar ítems (con nodos) o sus conexiones (con enlaces). Aquí tenemos dos tipos de enlaces: *containment* y *connection*.

¿Cómo usarlos? (expresividad y efectividad)

- No todos los canales son iguales: los mismos datos codificados con dos canales visuales distintos resultará en *información* diferente.
- Dos principios guían el uso de canales visuales: expresividad y efectividad.
- El principio de expresividad dicta que el *encoding* visual debe representar toda (y sólo) la información de los atributos del *dataset*.
  - Los datos ordenados deben ser mostrados de tal forma que nuestro sistema perceptual los perciba como ordenados; inversamente, debe ocurrir lo mismo con los datos no ordenados.
  - Esta es la razón de por qué clasificamos los atributos como ordenados o como categóricos.
  - Los canales de magnitud funcionan bien con los atributos ordenados, mientras que los de identidad son el *match* correcto con los atributos categóricos.
- El principio de efectividad dicta que los atributos más importantes deben ser codificados con los canales más efectivos, para que sean más perceptibles.

¿Cómo usarlos? (*ranking* de canales)

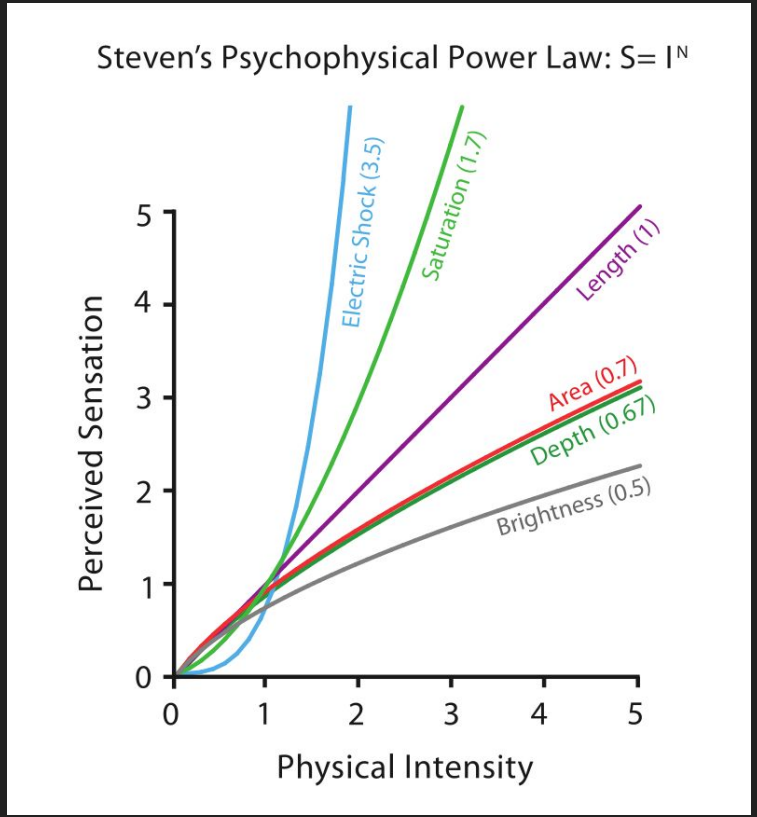


Efectividad de un canal

- Para analizar el espacio de *encodings* posibles, hay que entender ciertas características de estos canales visuales.
  - ¿Cómo se justifica este *ranking*?
  - ¿Por qué hay canales mejores que otros?
  - ¿Cuánta información puede codificar un canal?
  - ¿Pueden ser usados de forma independiente o podría haber interferencia entre ellos?
- Responderemos a estas preguntas, estudiando ciertos criterios:
  - el criterio de *accuracy*,
  - el criterio de *discriminability*,
  - el criterio de *separability*,
  - la habilidad de ofrecer *visual popout*.

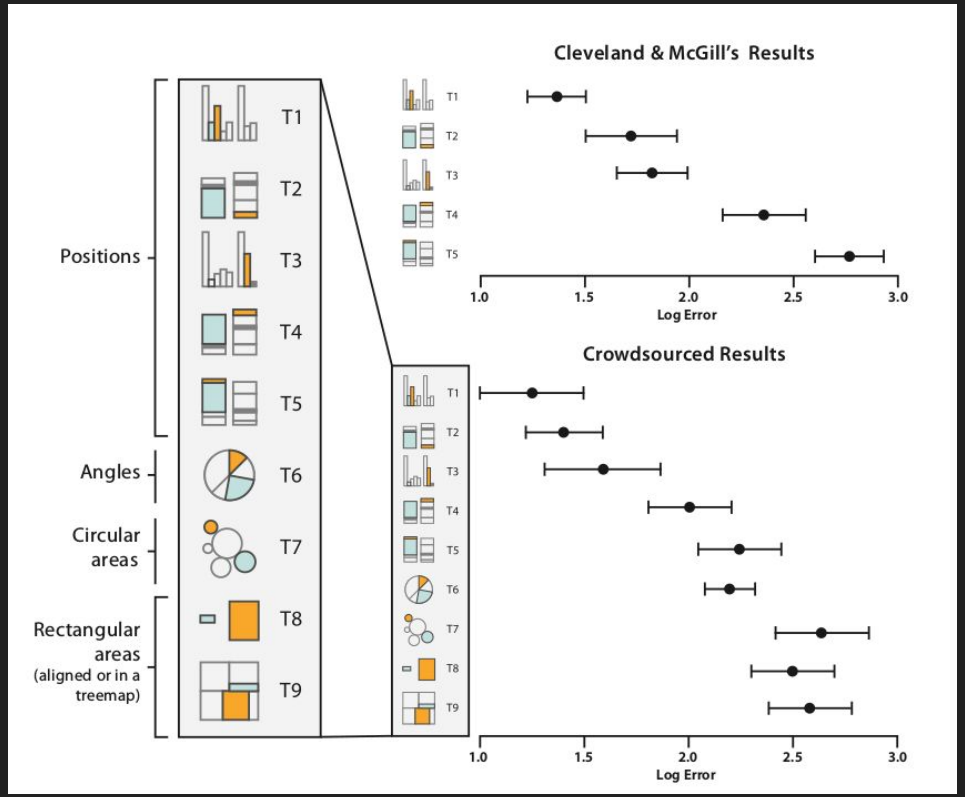
Efectividad de un canal (*accuracy*)

- *Stevens's power law* (1975)



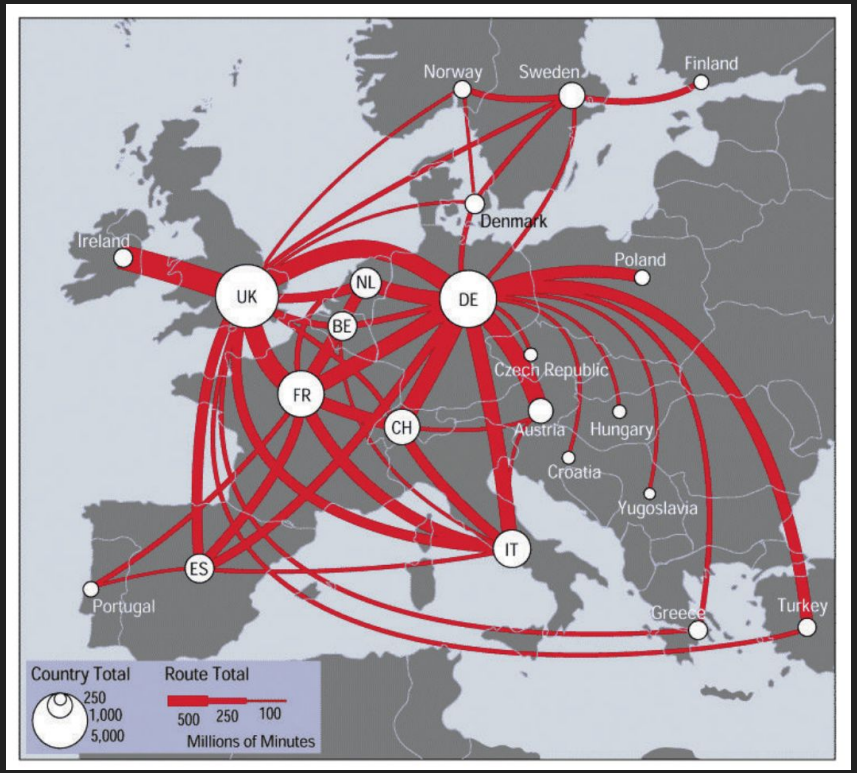
Efectividad de un canal (*accuracy*)

- Cleveland & McGill (1984)
- Heer & Bostock (2010)



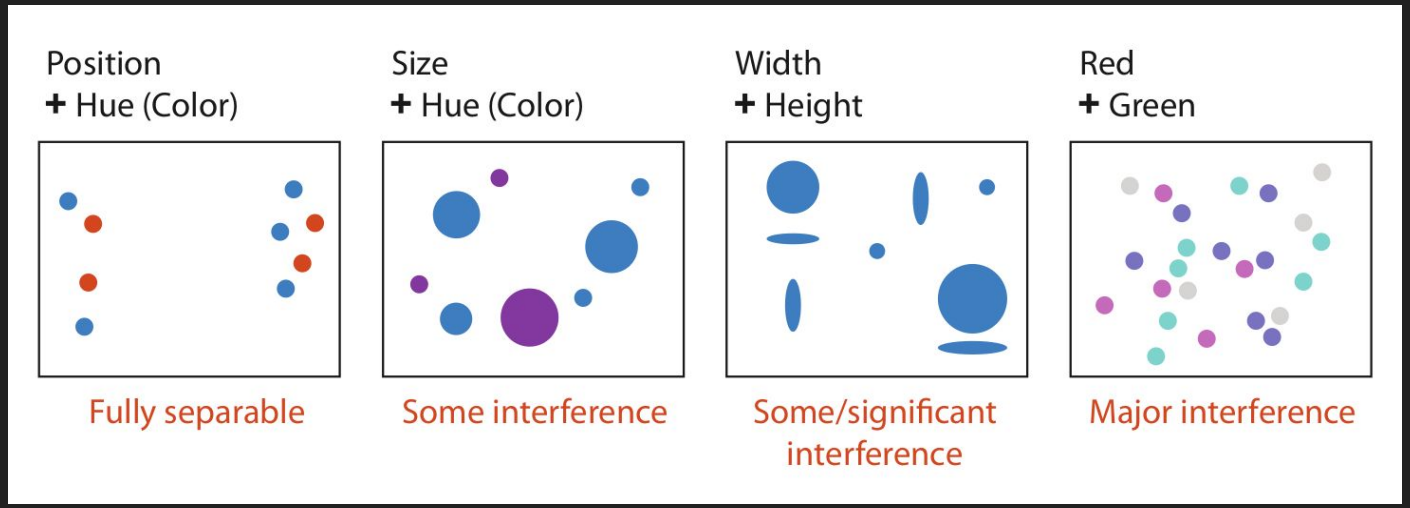
Efectividad de un canal (*discriminability*)

- Es importante considerar también cuántos *bins* están disponibles para ser usados en un canal visual, en donde cada *bin* es un paso (o nivel) distinguible del anterior o siguiente.
- Ejemplo: los *line widths*.



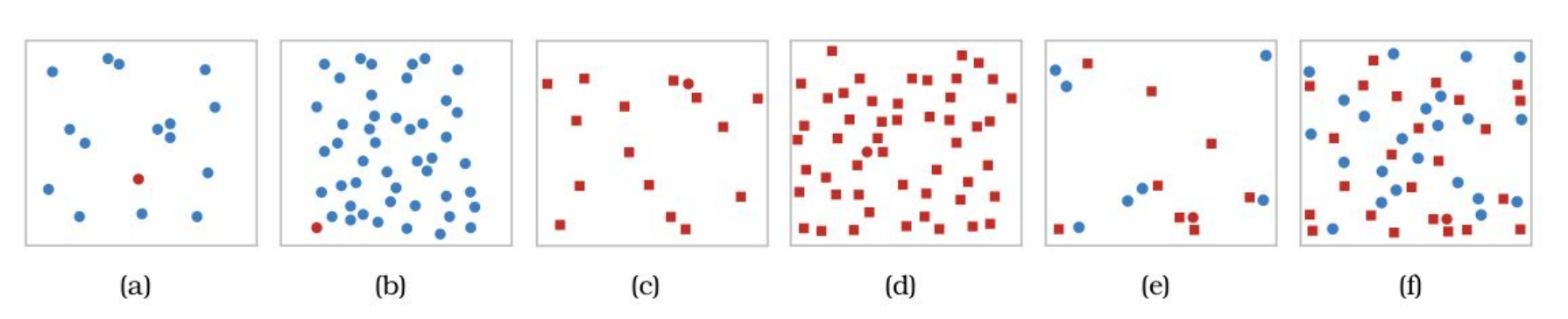
Efectividad de un canal (*separability*)

- No es posible tratar a los canales de forma independiente, puesto que generalmente tendremos dependencias e interacciones entre ellos.
- Existe un espectro de potenciales interacciones entre cada par de canales, que oscilan desde canales separables hasta canales integrales.

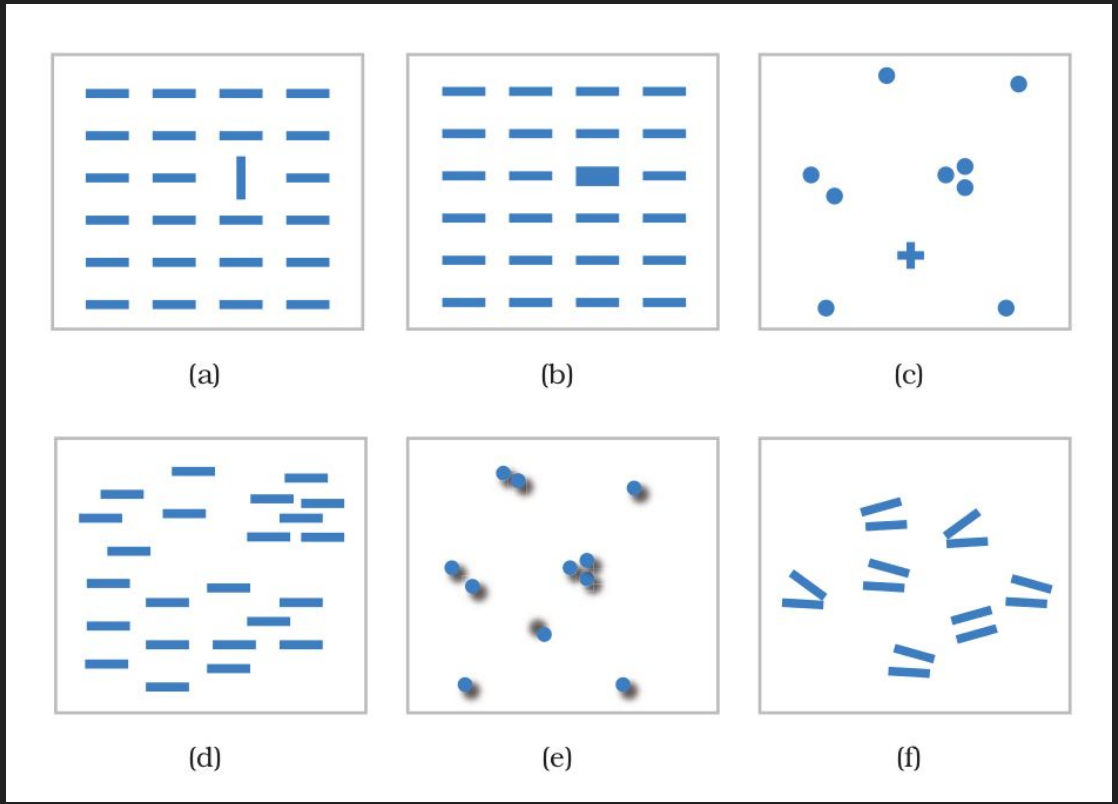


Efectividad de un canal (*visual popout*)

- Muchos canales ofrecen un efecto de *popout*, donde un elemento distinto se diferencia de forma inmediata.
- El valor del *popout* es que el tiempo que nos toma encontrar el objeto diferente (casi) no depende de la cantidad de los distractores.
  ¿Dónde está el círculo rojo?



Efectividad de un canal (*visual popout*)



Referencias

- El material es un extracto del curso *Visualización de información* (IIC2026)
- Munzner, T. (2014). *Visualization analysis and design*. CRC Press.