Visualización de Información
Abstracción de Datos

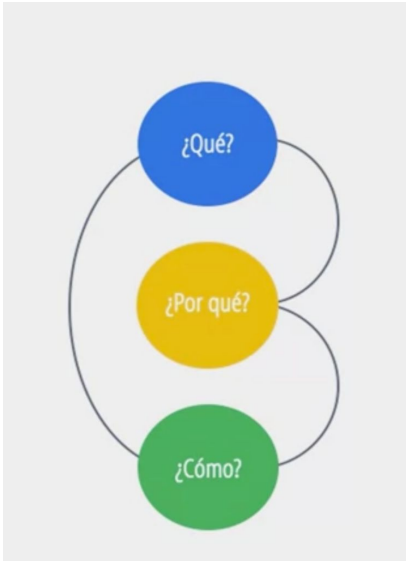
Daniela Opitz
dopitz@udd.cl
Data Science Institute, Universidad del Desarrollo
Edición 2023

Modelo de Muzner

¿Qué? Esta parte del modelo se refiere a la acción de identificar y derivar datos y atributos a visualizar.

¿Por qué? En esta etapa se identifican las tareas visuales que la visualización debe cumplir.

¿Cómo? Esta es la etapa en la que se identifica la o las codificaciones visuales o gráficos a utilizar.



Créditos: Denis Parra

¿**Qué** visualizar?

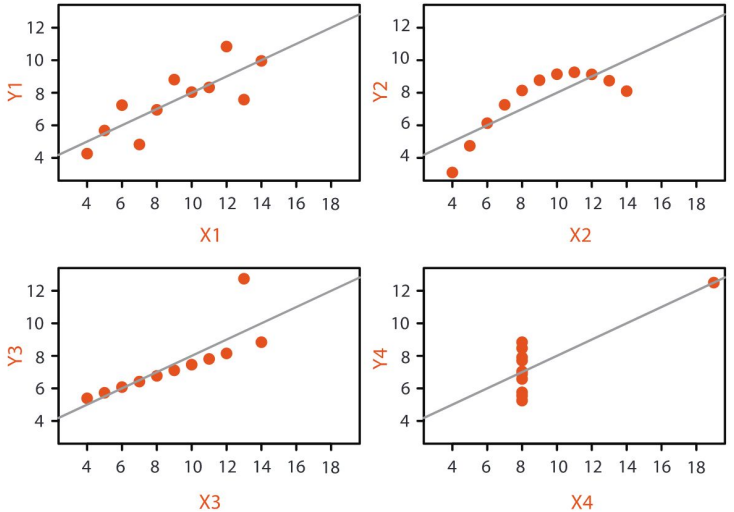
Dataset

Un dataset es un conjunto de observaciones, registros, etc., que contiene variables y atributos, usualmente numéricos.

Derecha: El Cuarteto de Anscombe. Presenta 4 subdatasets que tienen las mismas propiedades estadísticas básicas.

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	



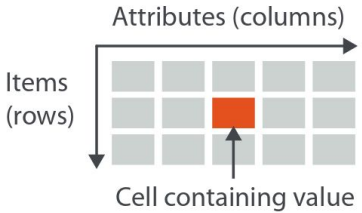
Tipos de Datasets

- Tabulares
- Redes
- Geometricos
- Campos

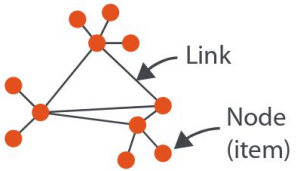
Tipos de Datasets

Dataset Types

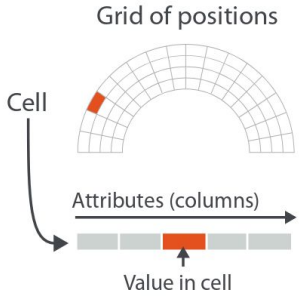
Tables



Networks

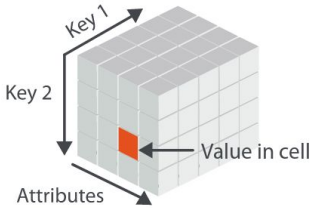


Fields (Continuous)

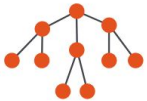


Geometry (Spatial)

Multidimensional Table



Trees



Tablas

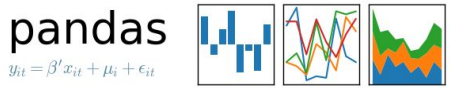
Una tabla es una colección de elementos (filas, observaciones, etc.).

Es un dataset base, en tanto todo lo podemos expresar como una tabla. Incluso una tabla multidimensional (un tensor) puede ser convertida a una tabla bidimensional (una matriz).

Operaciones: filtrado, agrupación, agregación.

```
# https://pandas.pydata.org/
import pandas as pd

census = pd.read_csv('census.csv')
census.head(5)
```



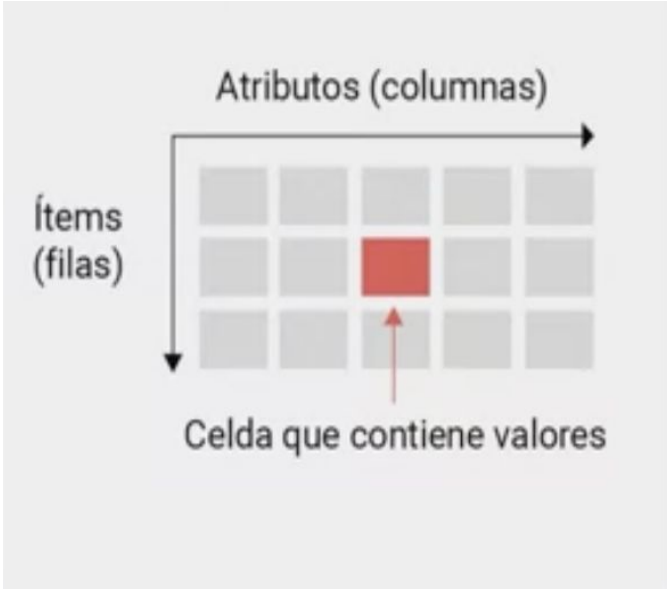
	REGION	PROVINCIA	COMUNA	DC	AREA	ZC_LOC	ID_ZONA_LOC	NVIV	NHOGAR	PERSONAN	P07	P08	P09	P10	P10COMUNA	P10PAIS	P11	P11COMUNA	P11PAIS	P12
0	15	152	15202	1	2	6	13225	1	1	1	1	1	73	1	98	998	3	15101	998	1
1	15	152	15202	1	2	6	13225	3	1	1	1	1	78	1	98	998	2	98	998	1
2	15	152	15202	1	2	6	13225	3	1	2	2	2	78	1	98	998	2	98	998	1
3	15	152	15202	1	2	6	13225	3	1	3	5	2	52	1	98	998	2	98	998	1
4	15	152	15202	1	2	6	13225	3	1	4	11	1	44	1	98	998	2	98	998	1

Datos Tabulares

- El dataset de tipo tabular es el más común. Lo encontramos en las tradicionales planillas de cálculo como Excel, donde cada fila es un ítem y cada columna representa un atributo.

- Se le llama celda a la intersección entre una fila, un ítem, y un atributo en particular, una columna, como en una matriz.

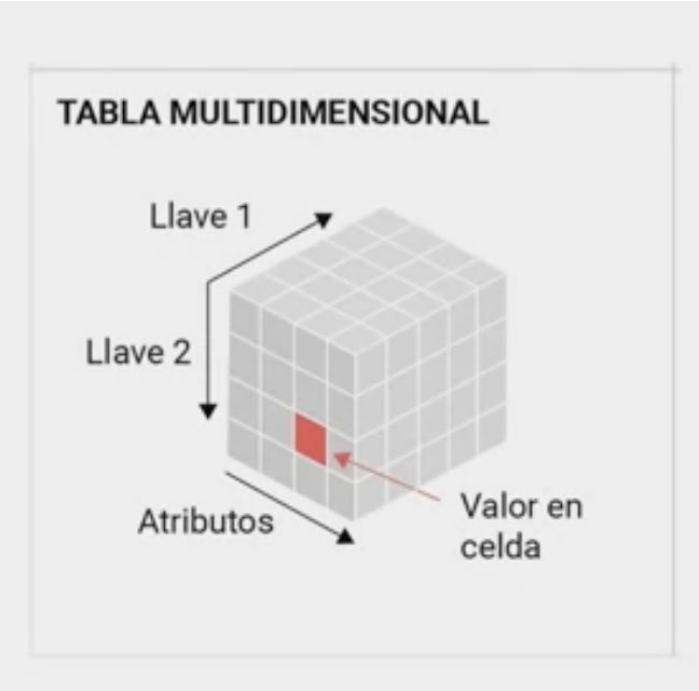
- También es importante el concepto de llave. Se trata de un atributo o columna particular que actúa como eje central para la búsqueda de valores en otros atributos.



Créditos: Denis Parra

Datos Tabulares Multidimensionales

- Hay más de una llave
- Ejemplo: informaciones de países en más de un año. La llave 1 sería país y la llave 2 sería el año.



Créditos: Denis Parra

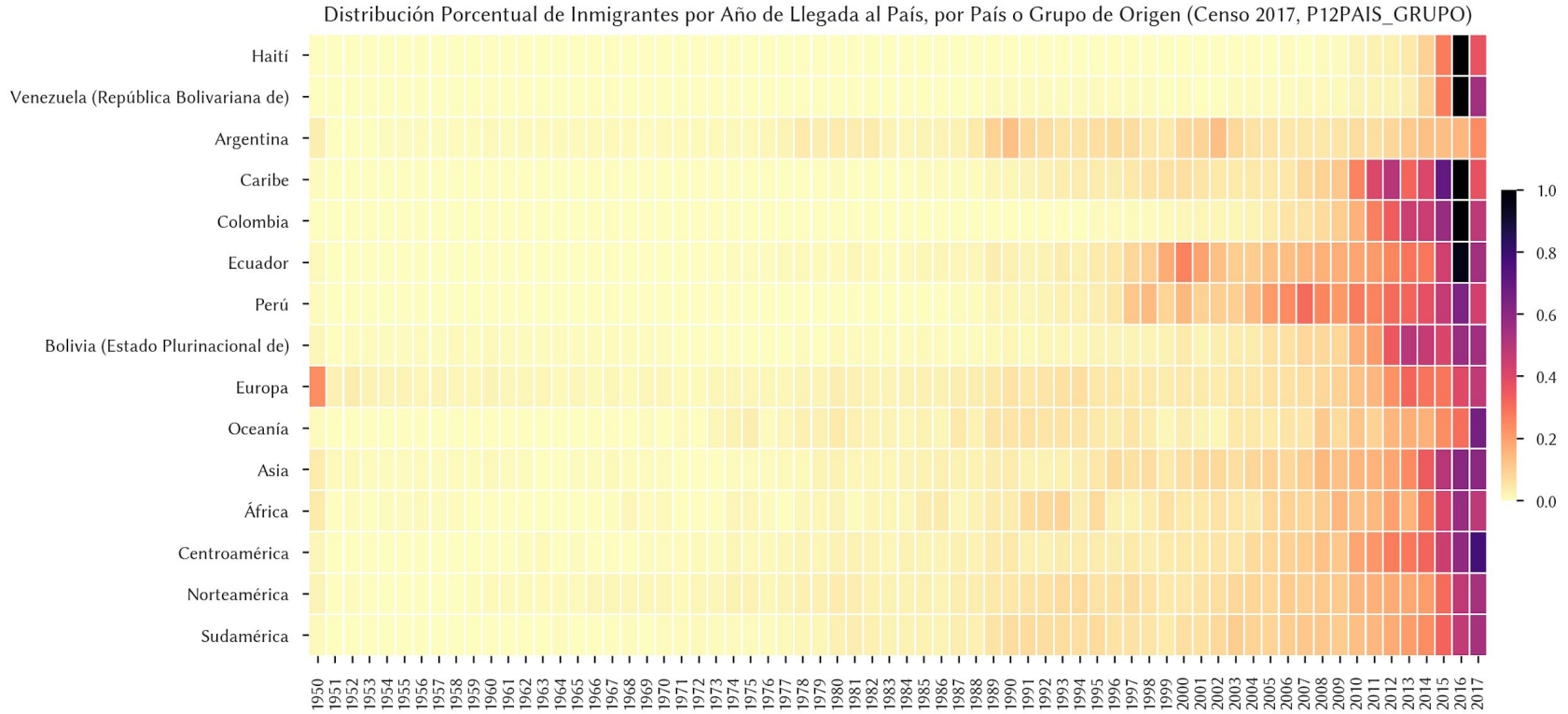
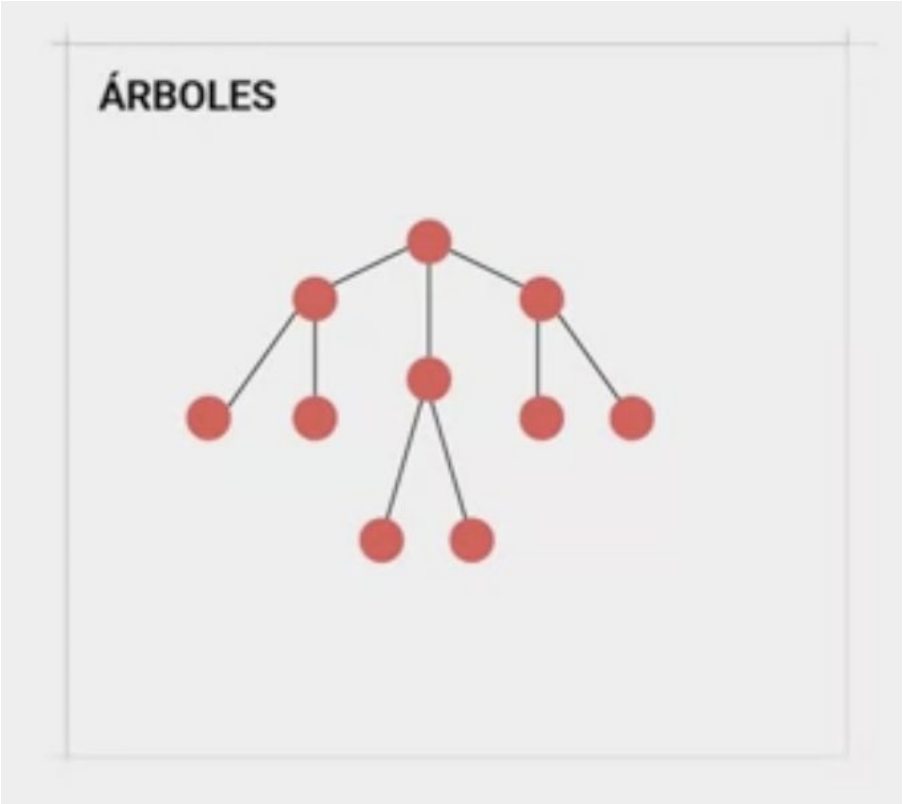


Tabla del censo 2017, filtrada para analizar inmigración, agrupada por lugar de origen, agregada contando personas, normalizada por filas.
Fuente: https://github.com/zorzalerrante/mapas_censo_2017

Árboles

Arbol: es una estructura de datos que imita una estructura jerárquica de la vida real, similar a un árbol genealógico o los directorios en una computadora. Son un tipo especial de red.



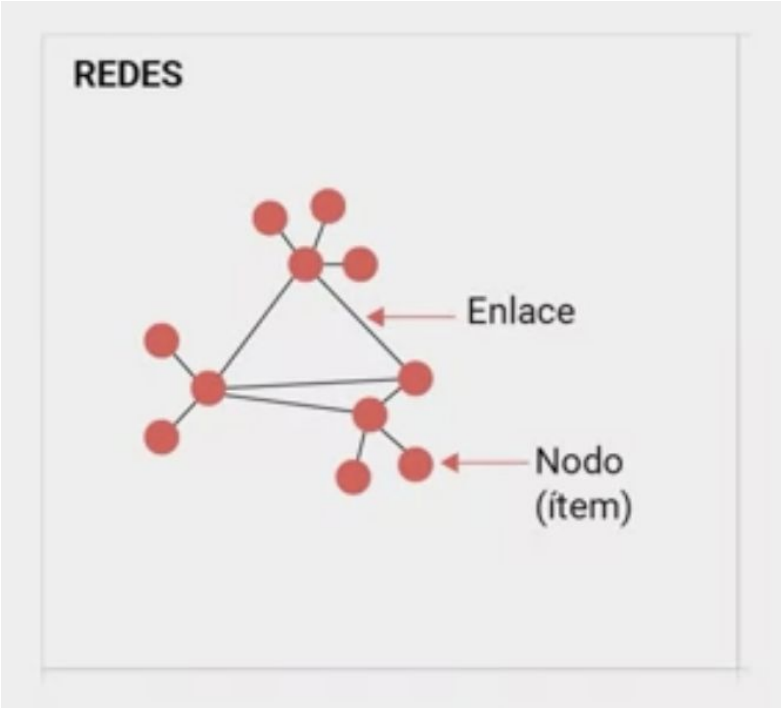
Créditos: Denis Parra

Redes

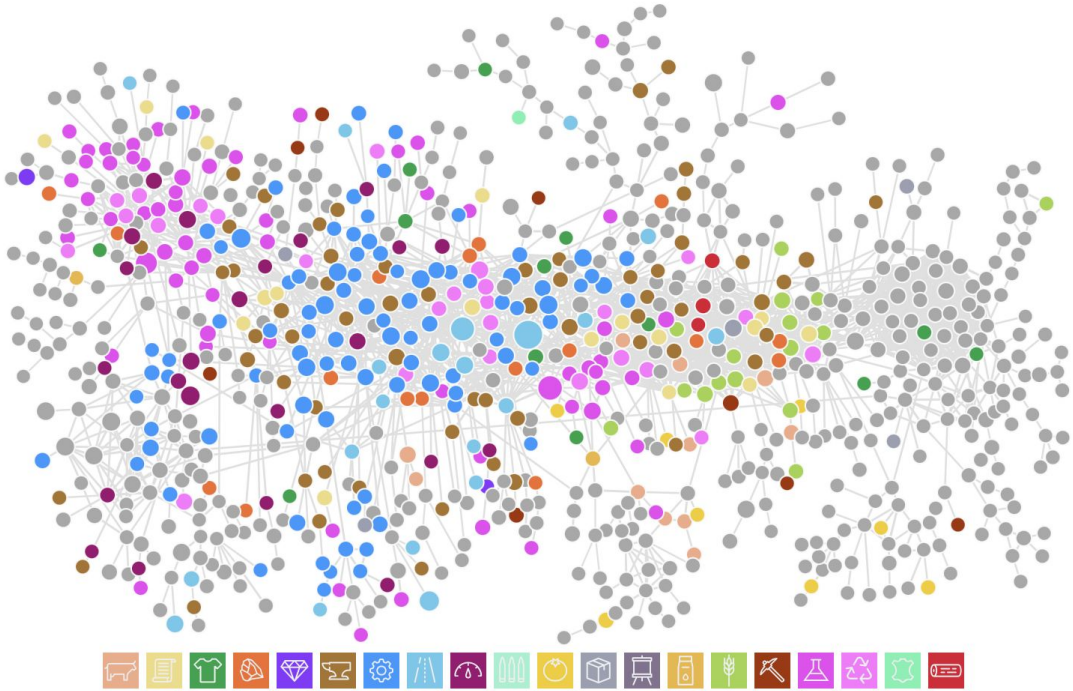
- El tipo de dataset de red se caracteriza por los enlaces, también llamados conexiones o aristas, que se establecen entre los ítems, también llamados nodos o vértices.

- Es posible tener atributos asociados tanto a los nodos como a los enlaces.

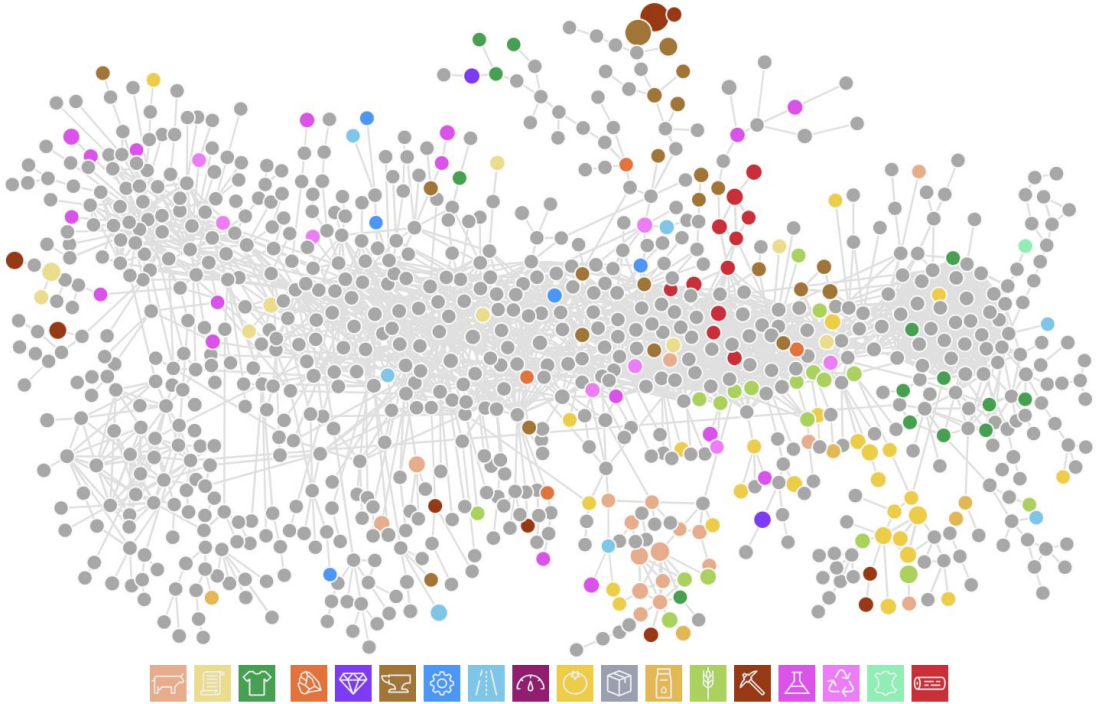
Operaciones: cálculo de importancia de nodos y aristas, caminos de un nodo a otro, detección de comunidades, predicción de atributos.



Créditos: Denis Parra



Red de complejidad económica. Los nodos son productos y los enlaces o aristas establacen la conexion entre productos https://oec.world/en/profile/country/deu#product-spac



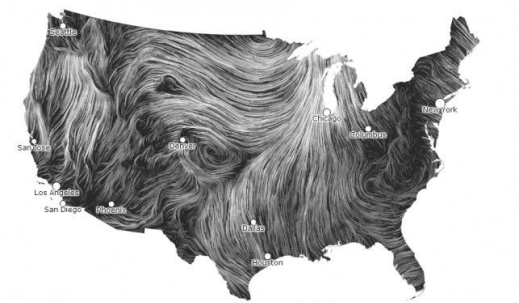
¿Qué país es este?

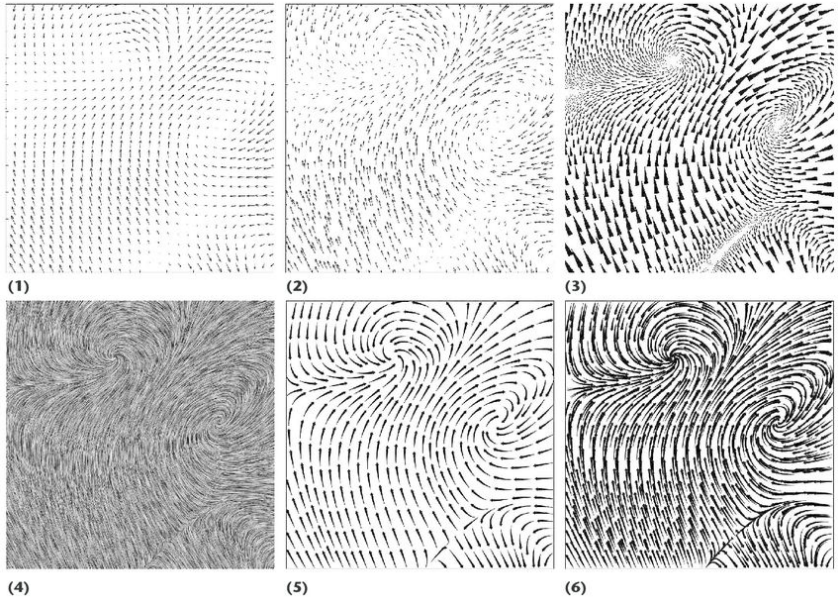
Campos

Son datos continuos en el espacio que se discretizan en grillas. Las grillas pueden ser regulares o irregulares.

Usualmente son datos provenientes de simulaciones (ej., física), de sensores (ej., viento), de exámenes médicos (ej., escáner).

Un buen ejemplo es http://hint.fm/wind/





Visualización de Campos vectoriales.

R. Moorhead, P. Rheingans, C. Johnson, T. S. Yoo, H. Pfister and T. Munzner, "Visualization Research Challenges: A Report Summary," in Computing in Science & Engineering, vol. 8, no. , pp. 66-73, 2006.

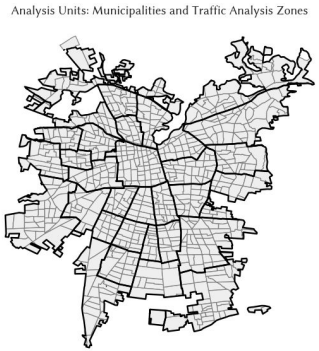
Geografía

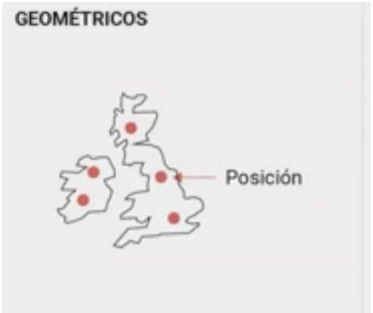
Los datos geográficos usualmente los podemos representar de manera directa (ítems en posiciones).

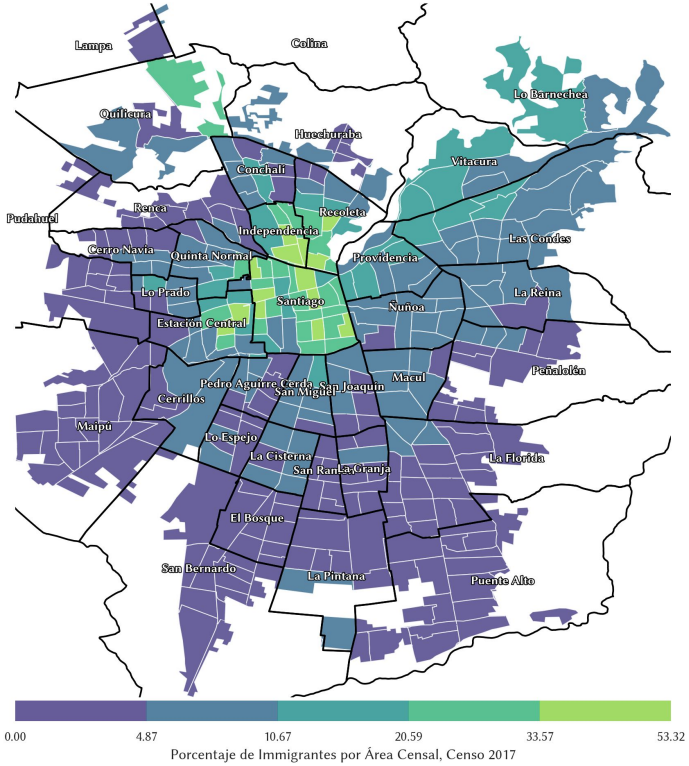
Operaciones: encontrar, comparar, explorar, contextualizar, manipular.

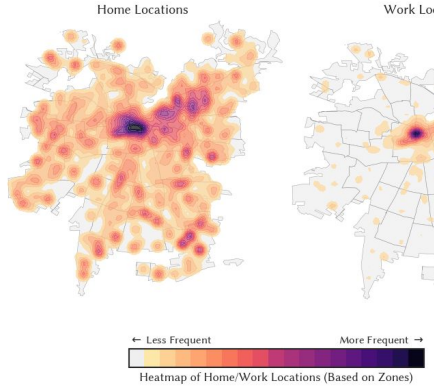
En el curso utilizaremos la biblioteca geopandas que, construida sobre pandas, permite hacer análisis, visualización, y transformación.

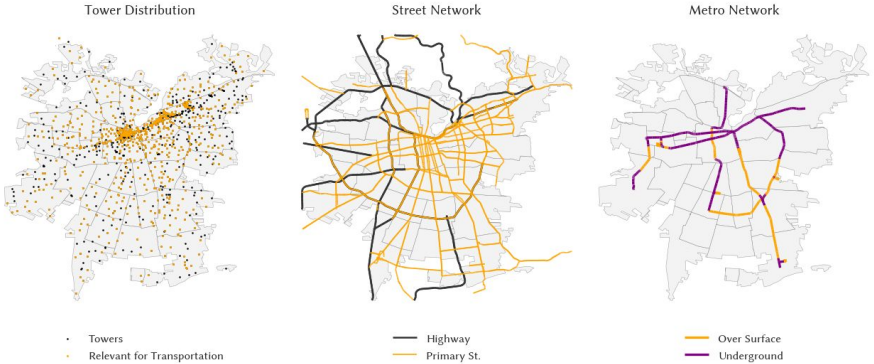
```
# http://geopandas.org
import geopandas as gpd
shape_1 =
gpd.read_file("municipalidades/")
shape_2 = gpd.read_file("zonas/")
ax = shape_1.plot()
shape_2.plot(ax=ax)
```











Visualizaciones de ejemplo. Izquierda: mapa de **coropletas** representando la cantidad de inmigrantes por área censal en Santiago. Derecha: **heatmaps** de hogar y trabajo en Santiago; infraestructura urbana y de telefonía. Todo esto hecho con **geopandas** (y esfuerzo).

Tipos de Atributos

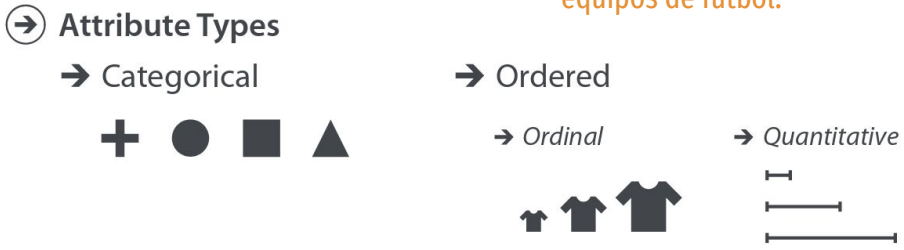
- **Categoricos**: género, tipos de fruta, nacionalidad

- **Ordenados**:
  - Ordinales: tallas de poleras, ranking
  - Cuantitativos: altura, precio

Tipos de Atributos

Aunque hay un orden, no podemos hacer operaciones aritméticas. Ejemplo: tallas de poleras, o equipos de fútbol.

¿Perro o gato? No hay operaciones aritméticas, ni orden.



Aquí sí podemos operar. Podemos calcular algo como: (a + b) * 0.5.

Hay un punto de origen, y de ahí sólo se crece. Ejemplo: edad.

Hay un punto de origen, y de ahí se puede avanzar en dos direcciones opuestas. Ejemplo: temperatura en ° C.

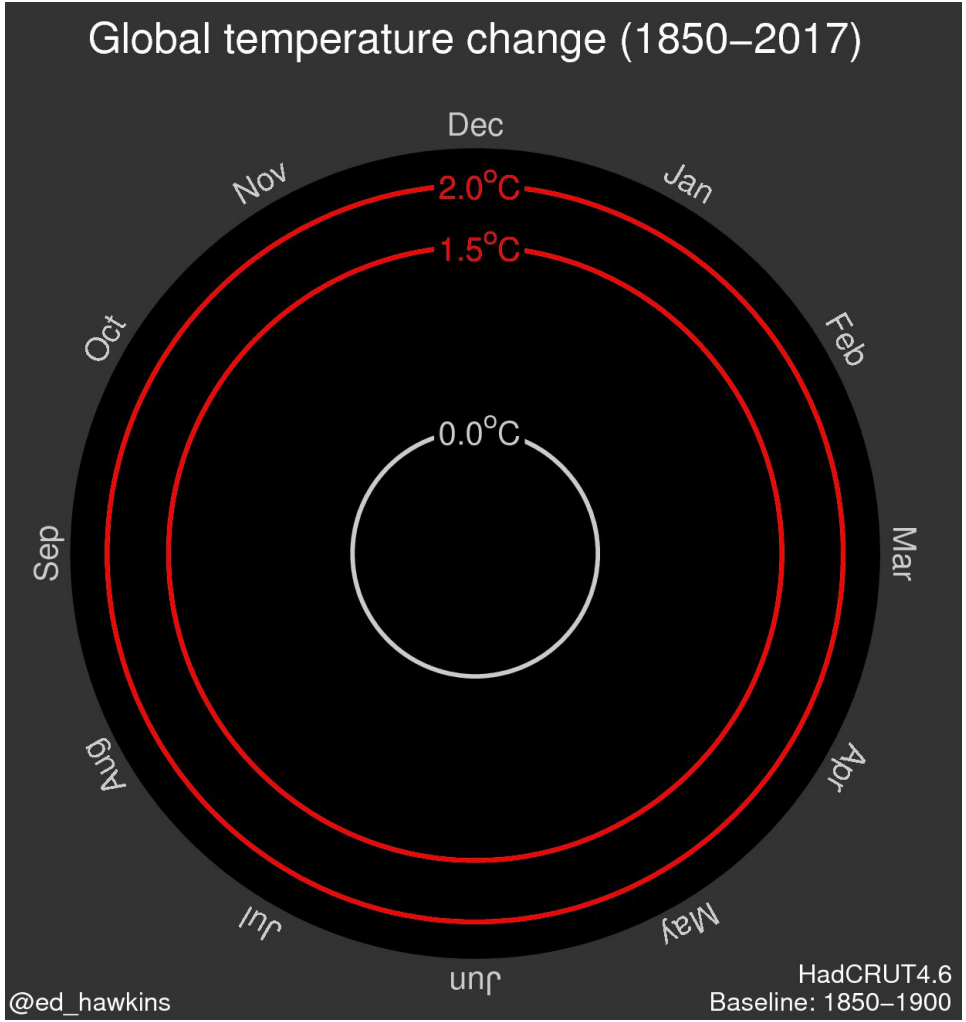
Hay un punto de origen, y de ahí se puede avanzar hacia adelante, pero hay atributos que se repiten en ciclos (ejemplo: los meses del año).

Atributo Cíclico

Estos atributos se pueden desplegar en gráficos con coordenadas polares, o usando espirales.

La fecha es un atributo interesante, puesto que puede ser **secuencial** (desde el *Big Bang*), **divergente** (en algún punto específico, como el año 0 en occidente), o **cíclica**, poniendo énfasis en los meses más que en los años.

Fuente de la visualización: http://www.climate-lab-book.ac.uk/spirals/



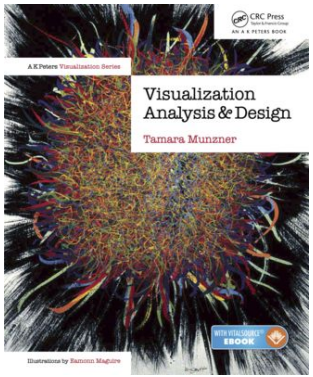
¿Dónde encontrar datos?

Esta página tiene enlaces a muchos datasets: Recursos

Preguntas

Mentimiter: 82 46 16 0

¿Preguntas?



Esta clase incluye material del libro **Visualization Analysis & Design** de Tamara Munzner.
http://www.cs.ubc.ca/~tmm/vadbook/