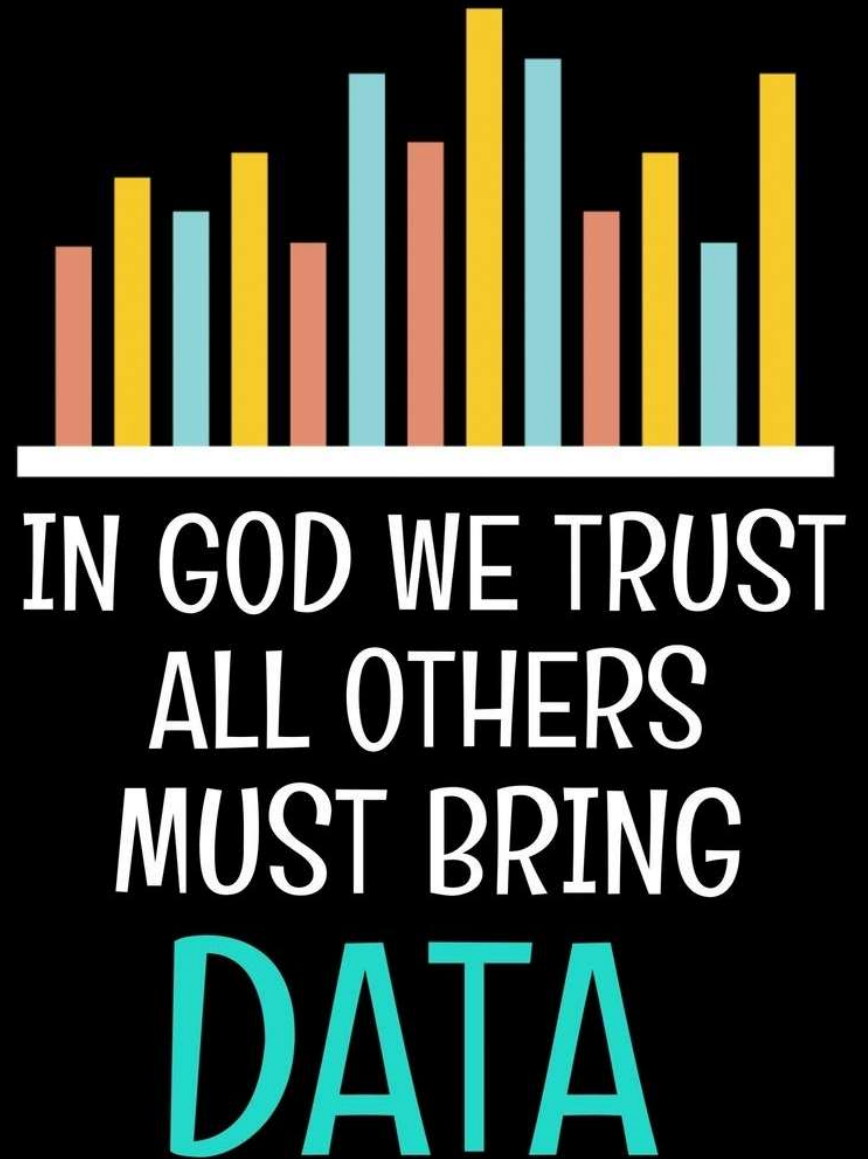


Introducción

Tema I

Edwards Deming



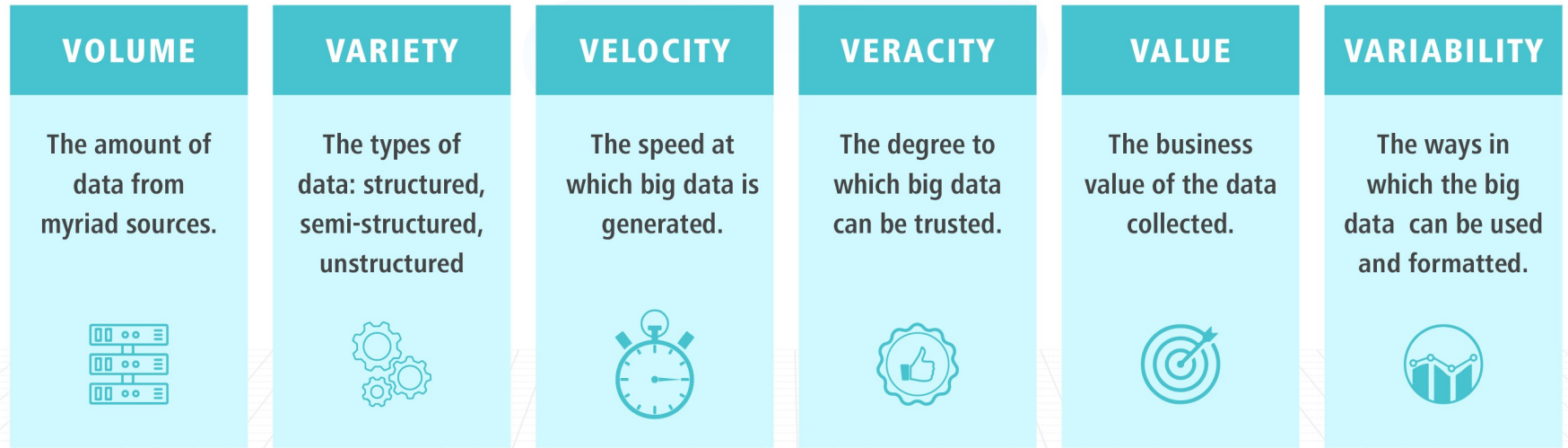
Algunos hechos del fenómeno digital

- Los analistas predecían que para 2020 habrá 5,200 gigabytes de datos por cada persona en el mundo.
- En promedio, la gente envía unos 500 millones de tweets por día.
- El cliente promedio de EE. UU. Utiliza 1.8 gigabytes de datos por mes en su plan de teléfono celular.
- En promedio, cada persona que usa el correo electrónico recibe 88 correos electrónicos por día y envía 34. Eso suma más de 200 mil millones de correos electrónicos cada día.
- Walmart procesa un millón de transacciones de clientes por hora y Amazon vende 600 artículos por segundo.
- MasterCard procesa 74 mil millones de transacciones por año.

(Christopher Tozzi, <https://blog.syncsort.com>)

The Six V's of Big Data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: Volume, Variety and Velocity. Over time, other Vs have been added to descriptions of big data.



¿Qué revelan las 6 V de Big Data?

- Qué es enorme.
- Que se expande rápidamente.
- Que está desordenada y en diferentes formatos.
- Que debe ser correcta.
- Que siempre está en proceso de cambio.
- Y que no vale nada si no se la utiliza de forma práctica.

Es por eso que **datos** y **procedimientos de análisis** son interdependientes: ninguno sirve sin el otro, pero el poder que tienen juntos no tiene límites.

¿Sabes qué país europeo tiene más ciudades con nombre de santo? Y no es el que te imaginas

Datos de la base de datos de la Agencia Nacional de Inteligencia Geoespacial (NGA; anteriormente conocida como Agencia Nacional de Imágenes y Cartografía, NIMA).

Departamento de Defensa de los Estados Unidos.

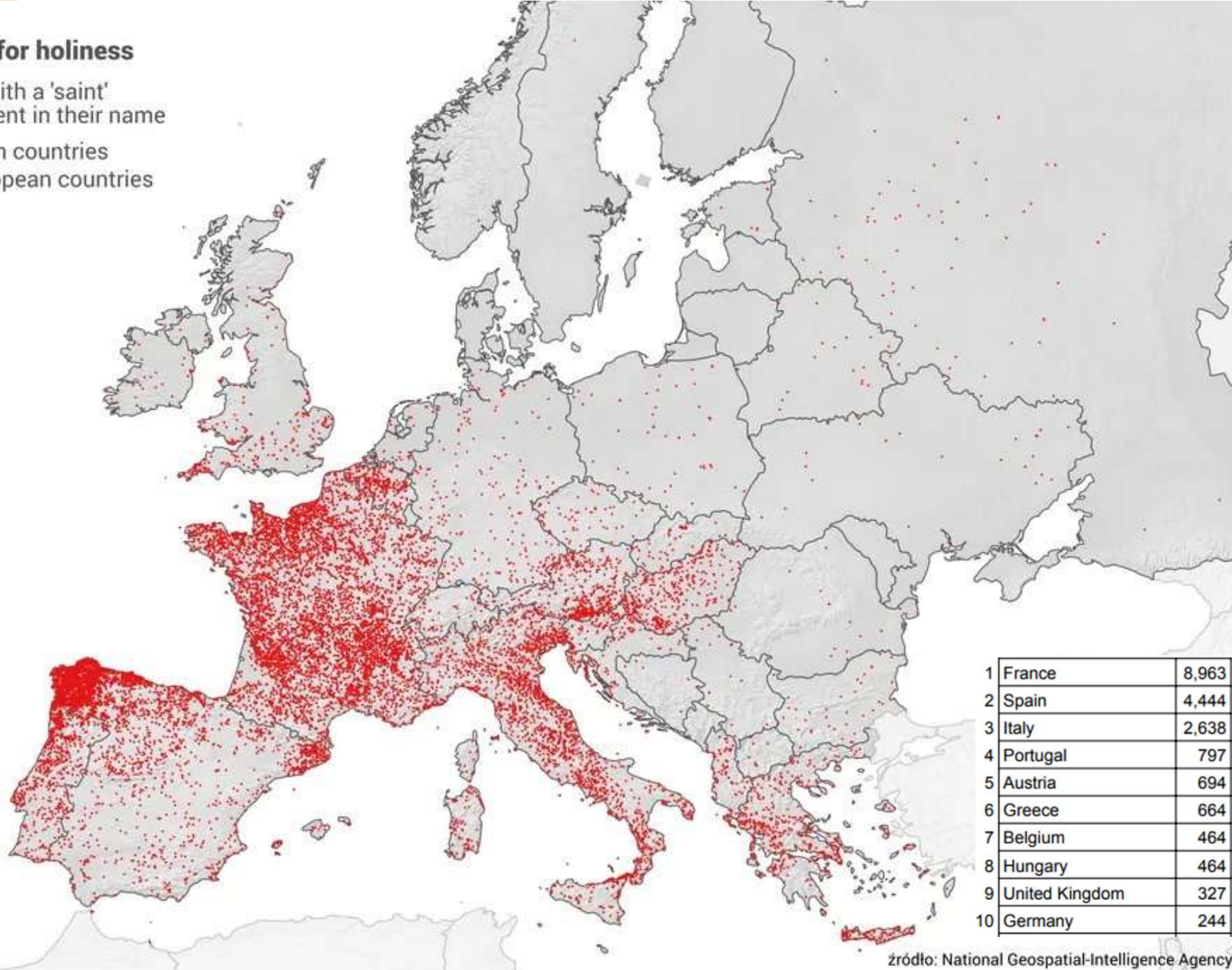
Responsable de recopilar, analizar y difundir la inteligencia geoespacial mediante imágenes vía satélite.

Biqdata.pl buscó como palabras clave los nombres de lugares que llevan en sus idiomas locales la raíz o la palabra «santo».

In search for holiness

- places with a 'saint' component in their name

■ European countries
■ non-European countries

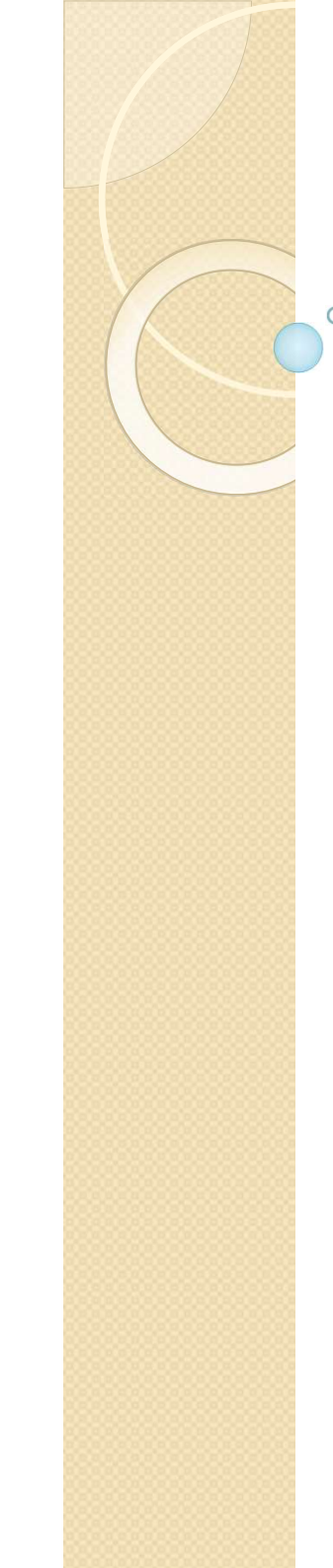


1	France	8,963
2	Spain	4,444
3	Italy	2,638
4	Portugal	797
5	Austria	694
6	Greece	664
7	Belgium	464
8	Hungary	464
9	United Kingdom	327
10	Germany	244



BIQdata.pl

źródło: National Geospatial-Intelligence Agency



Esquema

- La Geografía, Ciencia de Datos
- La Estadística, la estadística y las estadísticas
- Algunos conceptos básicos



La Geografía, Ciencia de Datos

La Geografía, Ciencia de datos

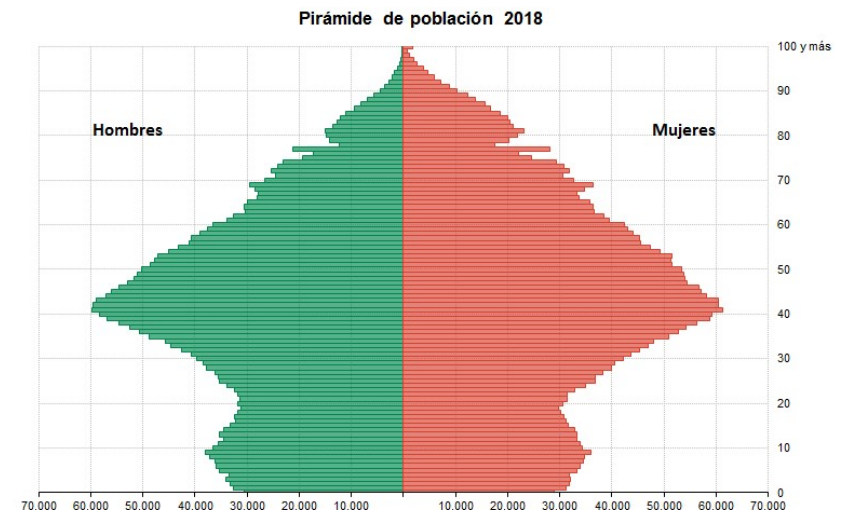
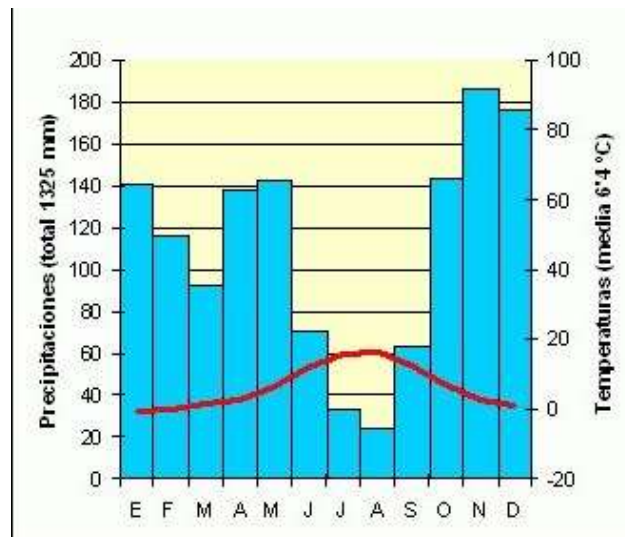
- ¿Qué es la **GEOGRAFÍA**? → Una **Ciencia**
- **Objetivo:** explicar, comprender y predecir la localización y dinámica de diferentes fenómenos sobre la superficie terrestre.
 - ¿Cómo difieren de un lugar a otro?
 - ¿En qué difieren de un lugar a otro?
 - ¿Cómo cambian a través del tiempo?
- Ciencia con unas características y una especificidad muy concretas → **información estructurada y georreferenciada** espacialmente.

La Geografía, Ciencia de datos

- **Datos, ¿para qué?**

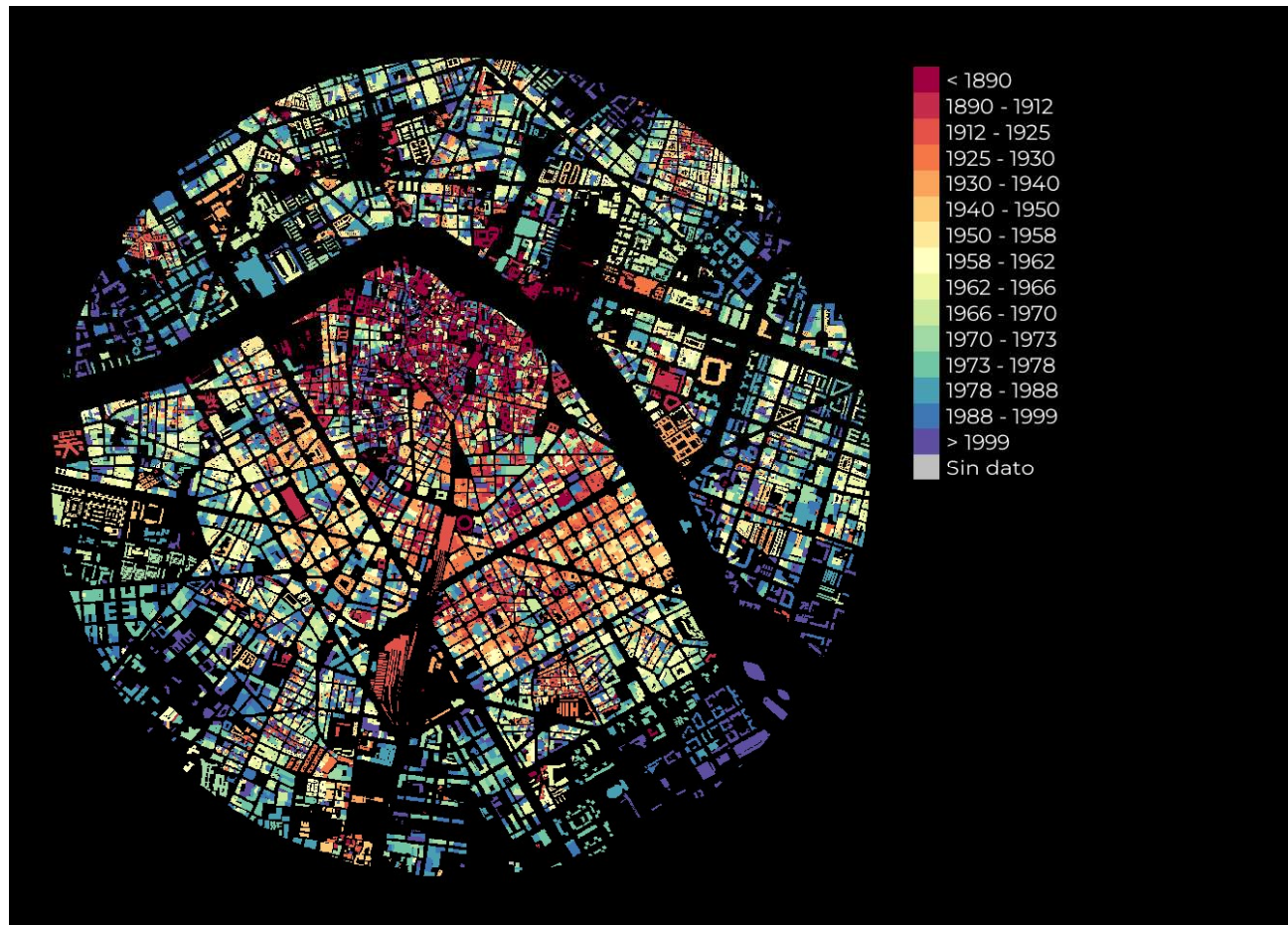
- **División tradicional de los fenómenos geográficos:**

- Fenómenos **naturales** → litología, relieve, clima, vegetación, suelo, hidrografía, etc.)
- Fenómenos **humanos** → características sociales, económicas, políticas, culturales, etc
 - ¿Qué datos utilizamos para elaborar estos gráficos?



La Geografía, Ciencia de datos

- **Datos, ¿para qué?**
 - ¿Qué datos utilizamos para elaborar estos gráficos?

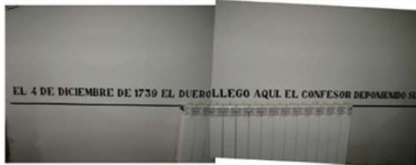


La Geografía, Ciencia de datos

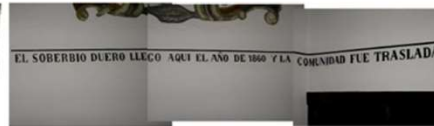
- **Datos, ¿para qué?**
 - ¿Qué datos utilizamos para elaborar estos gráficos?



Marcas del calado de inundaciones históricas en el convento de clausura de St^a M^a de Dueñas

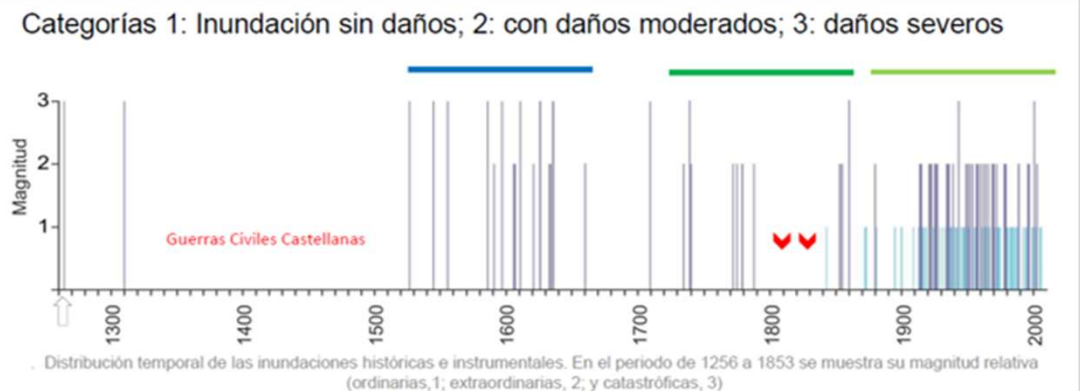


Limnimarca 5/12/1739



Limnimarca 25/12/1860

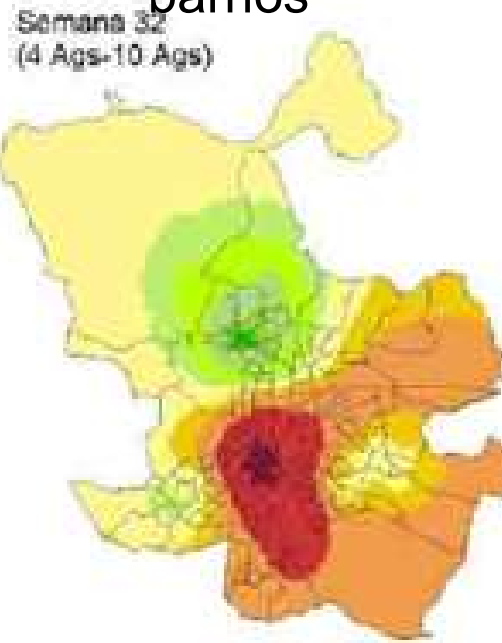
3. Resultados: inundaciones históricas en Zamora III



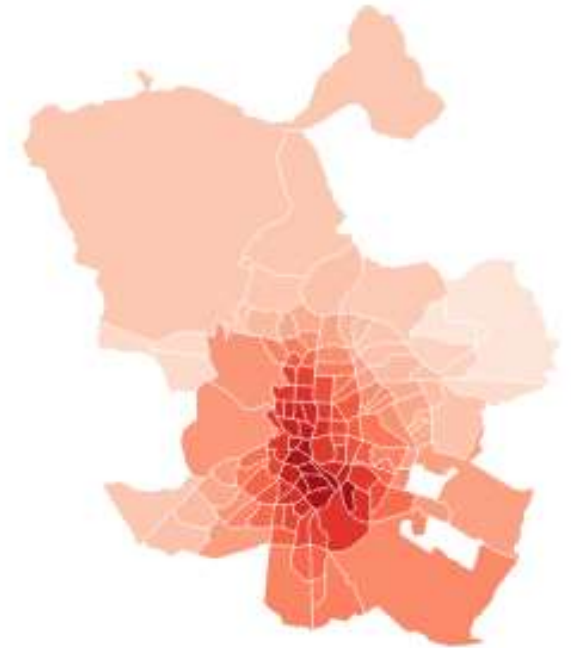
- Ejemplos de la **interacción** entre los fenómenos en Geografía: mortalidad por olas de calor



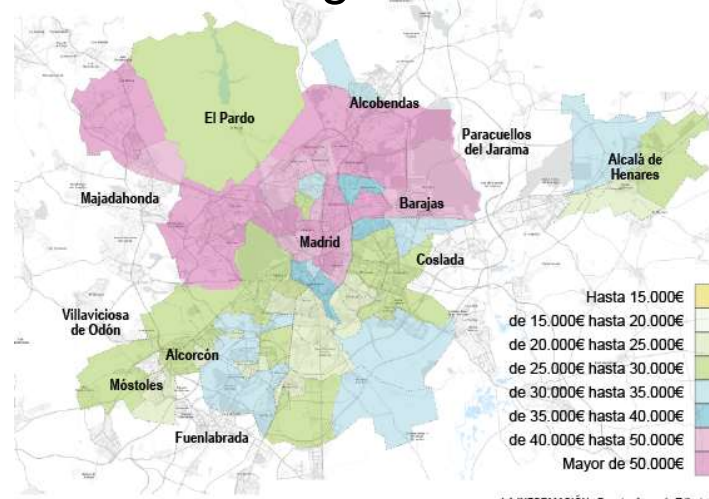
Mortalidad según barrios



Isla de Calor

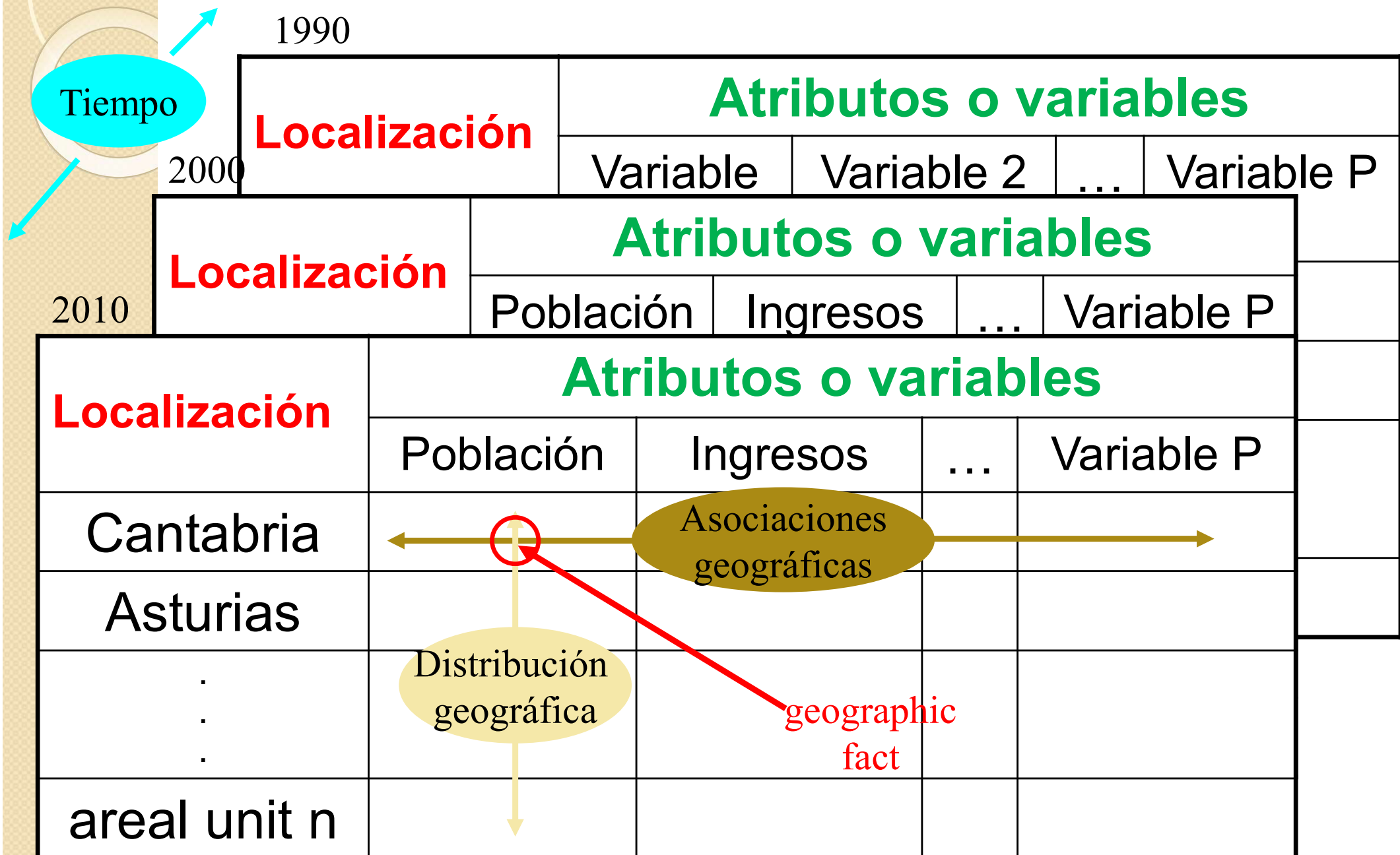


Renta según barrios



La Geografía, Ciencia de datos

- Los fenómenos geográficos se han resumido tradicionalmente en 3 **dimensiones**:
 - Dimensión **especial**.
 - Características (o atributos): dimensión **temática**.
 - El tiempo: dimensión **temporal**.
- La **matriz de información geográfica**, Berry 1964.



La Geografía, Ciencia de datos

Las matrices de información geográfica.

Estadísticamente, el geógrafo analiza simultáneamente 2 dimensiones, manteniendo fija la tercera → diferentes puntos de vista

En la misma (única) localidad

Tiempo		Atributos	
Atributos	1... n	Tiempo	1... n
	.		.
	.		.
	.		.
	N		N

Localidades		Tiempo	
Tiempo	1... n	Localidades	1... n
	.		.
	.		.
	.		.
	N		N

El mismo (único) atributo

Atributos		Localidades	
Localidades	1... n	Atributos	1... n
	.		.
	.		.
	.		.
	N		N

En el mismo (único) momento de tiempo

La Geografía, Ciencia de datos

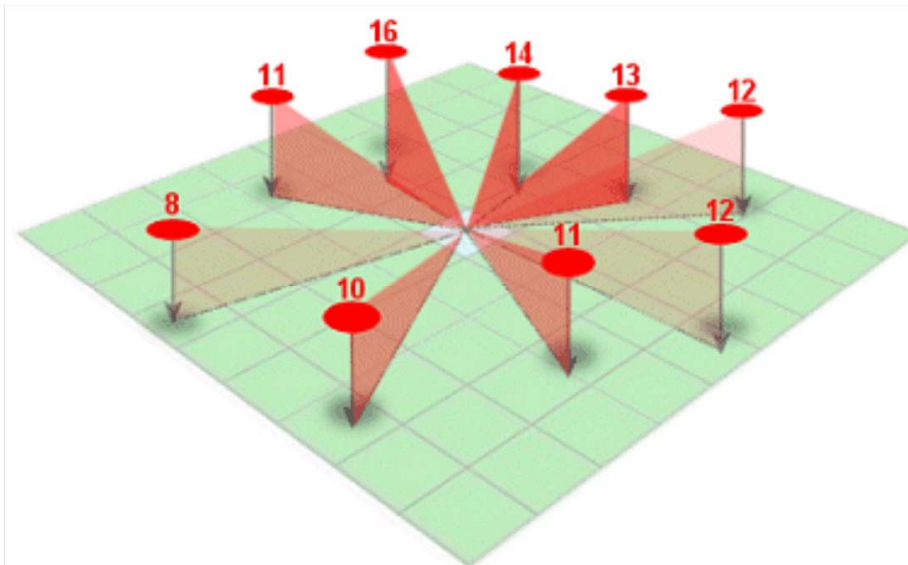
- Visión moderna → los fenómenos geográficos poseen propiedades...
 - **Locacional**: tienen una posición sobre la superficie terrestre (sistema de coordenadas).
 - **Espacial**: ocupan un lugar en el espacio (forma y volumen geométrico).
 - **Temática**: tienen atributos o propiedades que expresan sus características.
 - **Temporal**: pueden cambiar en el transcurso del tiempo.

La Geografía, Ciencia de datos

- La diversidad de propiedades → diversidad de métodos para el análisis espacial.
 - **Estadística convencional**
 - Prescinde de la localización de los objetos.
 - Se centra en las **componentes temática y temporal.** 😊
 - **Estadística espacial:**
 - Procedimientos específicos que analizan fenómenos con estructuras espaciales no aleatorias;
 - Se centra en las componentes locacional y espacial.

La Geografía, Ciencia de datos

- La **primera ley de la Geografía** (Waldo Tobler)
 - “*Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes*”.
 - Fundamental para conceptos como
 - Principio de Autocorrelación espacial (dependencia espacial)
 - Formulación de métodos de interpolación especial (kriging...)



La Geografía, Ciencia de datos

- La **segunda ley de la geografía** →
 - “Un fenómeno que ocurre fuera de un área de interés afecta a lo que sucede dentro” (Tobler, 1999)
 - "Las cosas que saben dónde están pueden actuar sobre su conocimiento de localización (Foresman y Luscombe, 2017)



La Estadística, la estadística y las estadísticas

La Estadística, la estadística y las estadísticas

- Definiciones:

- La **estadística**/las **estadísticas**:

- Colección de datos clasificados según un determinado criterio → *estadísticas de natalidad, de empleo, de precios* etc.

- La **Estadística**: **rama** de la Matemática aplicada que

- Utiliza números para el estudio de las leyes que dependen del **azar**.
 - Trata de descubrir la(s) causa(s) que explica los fenómenos analizados mediante el **razonamiento inductivo**: un patrón que se repite sirve para llegar a una conclusión.

La Estadística, la estadística y las estadísticas

- **¿Para qué y cómo usamos la Estadística en Geografía?**
 - **Estadística descriptiva**
 - Resume los datos.
 - Proporcionar mediciones de sus características.
 - No se pretenden conclusiones generales.
 - **La estadística inferencial**
 - Formula afirmaciones sobre las características de una población basándose en datos de una muestra representativa
 - Estimar las probabilidades de que algo ocurra.

La Estadística, la estadística y las estadísticas

- **¿Para qué y cómo usamos la Estadística en Geografía?**
 - **Predicción y modelización**
 - Las generalizaciones sobre patrones sirven para formular predicciones (a futuro) o explicaciones (afirmaciones sobre procesos sucedidos en el algún momento pasado).
 - Verificar si el patrón real coincide con un patrón esperado o teórico.
 - ¿Qué son los modelos?
 - Abstracciones del mundo real
 - Versiones simplificadas de la realidad

La Estadística, la estadística y las estadísticas

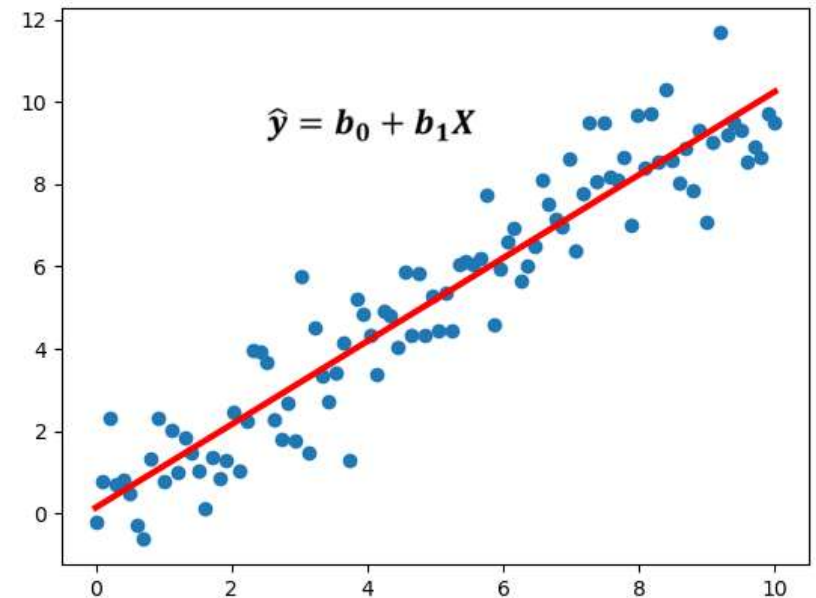
- **¿Para qué y cómo usamos la Estadística en Geografía?**
 - **Superación de la división entre Estadística Descriptiva e Inferencial.**
 - Desarrollo de la informática en la segunda mitad del siglo XX: grandes volúmenes de datos y gran rapidez de procesamiento
 - Disminución de la importancia de los estudios muestrales: ya no hay que limitarse a métodos basados en distribuciones conocidas (su principal aplicación eran las pequeñas muestras).
 - ¿Por qué restringirse a analizar unas pocas variables? ¿Por qué no utilizar toda la información disponible?
 - Desarrollo de una serie de tipos de análisis de datos que se sitúan entre la estadística descriptiva y la inferencia o estadística teórica, pe. **EDA** (análisis exploratorio de datos) desarrollado por Tukey.

La Estadística, la estadística y las estadísticas

- ¿Para qué y cómo usamos la Estadística en Geografía?

- **Análisis Exploratorio de datos.**

- Los datos están constituidos por dos partes:
 - Una “**regularidad**” o tendencia: la estructura simplificada de un conjunto de observaciones (en una nube de puntos, por ejemplo, la recta de ajuste).
 - Las “**desviaciones**” o variabilidad (residuos) son las diferencias de los datos con respecto a esta estructura: usualmente no tienen por qué presentar una estructura determinada.

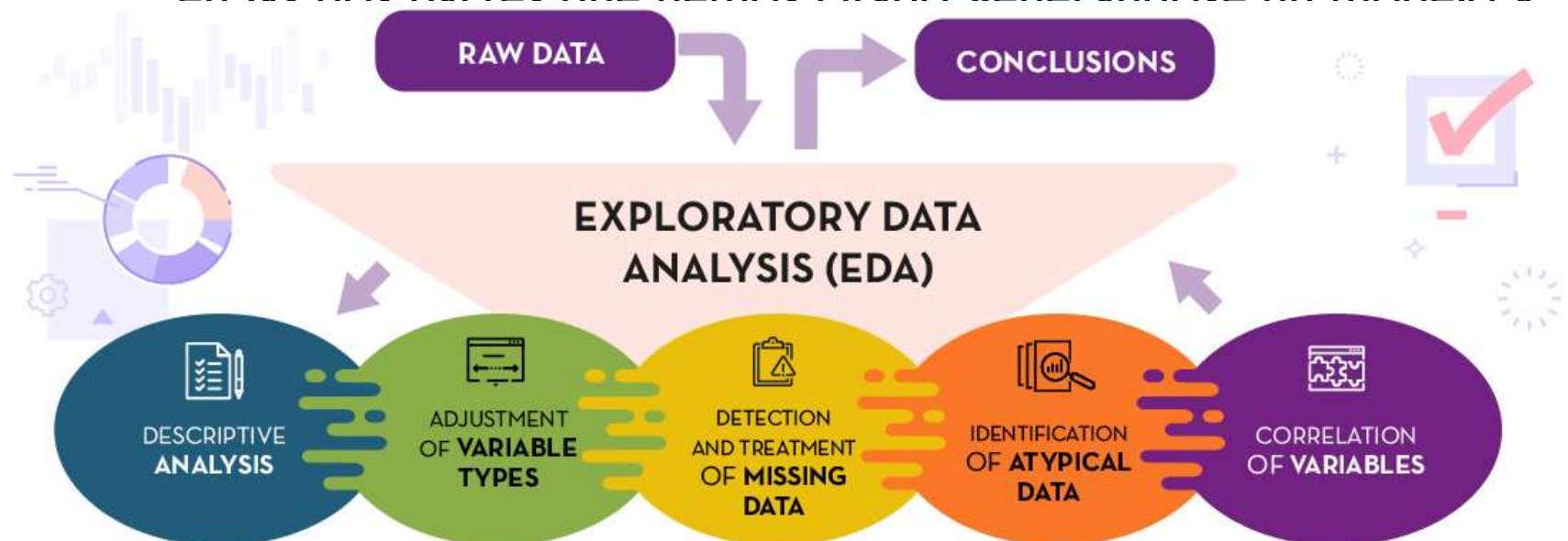


La Estadística, la estadística y las estadísticas

- **¿Para qué y cómo usamos la Estadística en Geografía?**

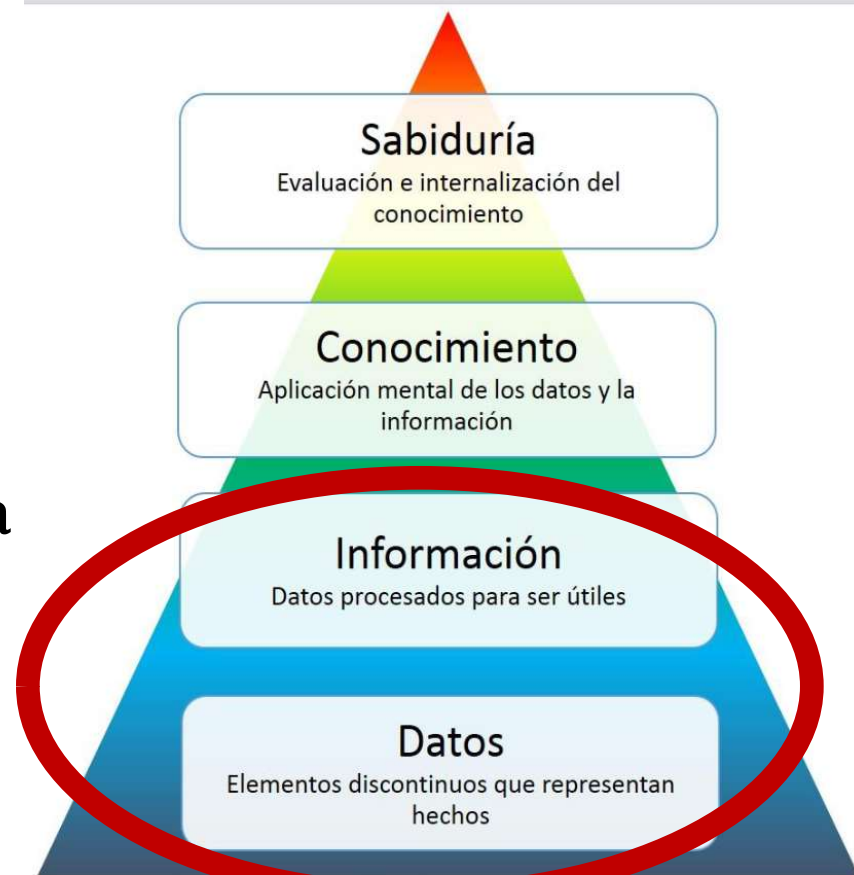
- **Análisis Exploratorio de datos.**

- Tradicionalmente la Estadística se concentró en la búsqueda de un modelo que exprese la regularidad de las observaciones.
 - Por el contrario, el EDA es básicamente el desglose de los mismos en las dos partes que hemos citado, generándose un modelo a

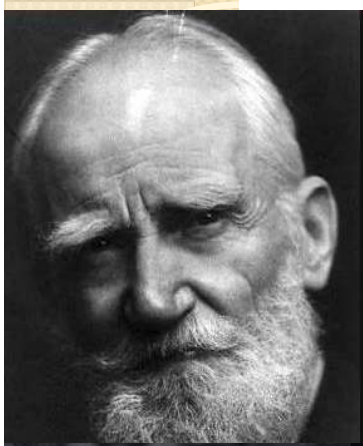


La Estadística, la estadística y las estadísticas

- La **Estadística** no es el **fin**, sino un **medio** para...
- **Jerarquía DIKW**", o "Pirámide del Conocimiento":
 - Modelo que representa las relaciones estructurales entre Datos, Información, Conocimiento, y en algunos casos Sabiduría



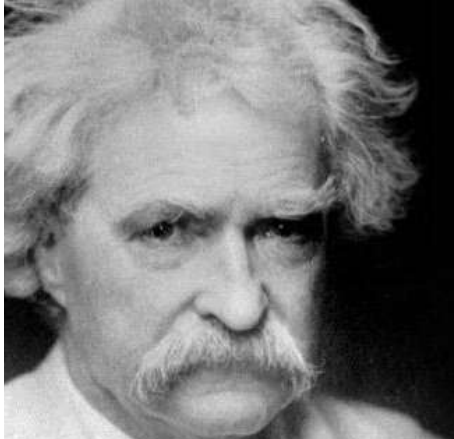
Fuente: Hey, J.: The Data, Information, Knowledge, Wisdom Chain: The Metaphorical Link



“La estadística es una ciencia que demuestra que, si mi vecino tiene dos coches y yo ninguno, los dos tenemos uno.”
— George Bernard Shaw —



Sólo me fío de las estadísticas que he manipulado
(Winston Churchill)



**“Hay tres clases de mentiras:
La mentira, la maldita mentira
y las estadísticas”**

Mark Twain (1835-1910)
Escritor estadounidense



Una única muerte es una tragedia, un millón de muertes es una estadística.
(Iósif Stalin)

La Estadística, la estadística y las estadísticas PERO

La manipulación de información es muy real.

Si crees todo lo que dicen entonces no estás prestando atención.

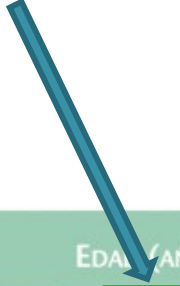
La evolución humana es directamente proporcional a su capacidad de cuestionamiento del mundo exterior. Cuestiona todo.



Conceptos básicos

Conceptos básicos

Esto es un dato



ID	EDAD (AÑOS)	COLESTEROL (MG/DL)	TIEMPO (MIN/MES)
1	59	155	41
2	65	215	24
3	78	235	21
4	84	266	14
5	65	176	41
6	69	181	37
7	76	241	25
8	81	257	22
9	58	215	26
10	67	188	29



*ID es el número identificativo de cada paciente.

Esto es una variable

Conceptos básicos

- **Dato:**
 - Elemento básico de información
 - Cuantifica (mide) una determinada propiedad o característica.
 - Tipos
 - **Primarios**
 - Adquiridos directamente de la fuente original.
 - El proceso de adquisición consume mucho tiempo e implica una decisión previa para la obtención de datos representativos.
 - **Secundarios**
 - Recopilados por algún organismo (Naciones Unidas, EU...).
 - Relativamente fácil acceso (INTERNET) pero su procedencia condiciona muchas de sus características.
 - Bases de datos: conjuntos masivos de datos

Conceptos básicos

- **Variable:**

- Cada una de las características o propiedades que poseen los individuos
- “Varían” de sujeto a sujeto (cada sujeto tiene un **valor** para cada variable)
- Suelen designarse con letras mayúsculas del final del abecedario
 - **(X, Y, Z)**
- Los **valores** que toman las variables se representan con letras minúsculas con subíndice
 - $X_1, X_2, X_3 \dots X_n$

Conceptos básicos

- **Valor:** cada uno de los distintos **resultados** de una variable
 - Si lanzamos una moneda al aire 5 veces sólo podemos obtener dos valores: cara y cruz.

ID	EDAD (AÑOS)	COLESTEROL (MG/DL)	TIEMPO (MIN/MES)
1	59	155	41
2	65	215	24
3	78	235	21
4	84	266	14
5	65	176	41
6	69	181	37
7	76	241	25
8	81	257	22
9	58	215	26
10	67	188	29

*ID es el número identificativo de cada paciente.

Conceptos básicos

- ¿Por qué es importante el tipo de variable?
 - El proceso para **definir y medir las variables es crucial**
 - Una definición inapropiada o una medición incorrecta compromete el resto de decisiones
 - Algunas variables no son difíciles (ejemplo “sexo”, “edad”).
 - Otras variables, aparentemente “obvias”, no lo son tanto: ejemplo “estado civil”
 - Prueba y error → para definir variables que captan características como “estatus social”, “nivel educativo”, “ideología política”, “religiosidad”...

Conceptos básicos

- ¿Por qué es importante el tipo de variable?
 - **Escala:**
 - Distintos modos de organizar, medir o asignar valores a los datos recopilados de una muestra.
 - Existen diferentes tipos de escalas, dependiendo de la naturaleza de las variables que se estudian o de la precisión que requiera la organización de los datos.
 - Distinguimos
 - **Cualitativas:** nominales y ordinales
 - **Cuantitativa:** discretas y continuas

Conceptos básicos

ESCALAS DE MEDICIÓN



Conceptos básicos

- Tipos de variables:
 - **Cualitativas (categóricas):**
 - Sus características o cualidades que no pueden ser medidas con números.
 - Sus valores son “categorías” → son diferentes por una cualidad (no por una cantidad).
 - Propiedades de las categorías
 - Exhaustivas: deben clasificar a todas las unidades de análisis (suele aparecer una categoría –“otros”- cuando hay individuos difíciles de clasificar).
 - Mutuamente excluyentes: sin ambigüedad (un único criterio de clasificación), de manera que cada individuo sea asignado a una categoría y sólo a ella.

Conceptos básicos

- Tipos de variables
 - **Cualitativas**
 - **Nominal:** no admiten un orden.
 - No podemos calcular estadísticos (pe. promedio) de estado civil, o de partido político
 - Podríamos transformar una variable cuantitativa en escala ordinal, o en cualitativa.
 - Usos del territorio: urbano, industrial, forestal...

Conceptos básicos

- Tipos de variables
 - **Cualitativas**
 - **Ordinal** (cuasicuantitativa): sus valores son “categorías” pero cada una puede ser “mayor” o “menor que” las restantes → no numéricas pero con orden
 - *Nota de un examen* (suspense, aprobado, notable, sobresaliente)
 - *Medallas* (oro, plata, bronce), de una prueba deportiva
 - *Clase social* (baja, media, alta)
 - *Opinión sobre una propuesta política* (muy en contra, más bien en contra, indiferente, más bien a favor, muy a favor)

Conceptos básicos

- **Atributo (variables cualitativas)**
 - Característica no susceptible de medida
 - Se designa con las primeras letras mayúsculas de abecedario (**A, B, C, D, ...**), y para las distintas modalidades, la correspondiente letra minúscula con subíndices (**$a_1, a_2..a_n$**)
 - Atributo A: color del cabello.
 - Modalidades: a_1 cabello negro, a_2 cabello rubio etc..

Conceptos básicos

- Tipos de variables
 - **Cuantitativa:** se expresan mediante un número
 - Cada posible valor es menor o mayor que otro valor.
 - *Ej: edad, ingresos, nota en un examen, número de años de educación, kilómetros de distancia entre trabajo y residencia...*
 - Podemos realizar operaciones aritméticas.
 - El conjunto de valores forman una escala de intervalo
 - **Intervalo:** distancia entre valores → podemos calcular la distancia o intervalo entre cualquier par de valores de la variable. Ejemplo: ingresos. 10.000, 30.000, 40.000 €
 - **ATENCIÓN:** “números” que son “etiquetas”: código postal, número de teléfono, código de una asignatura

Conceptos básicos

- Tipos de variables
 - **Cuantitativas:** según el número de valores que tengan en la escala
 - **Discretas:**
 - Valores son números enteros, sin intermedios entre dos valores específicos.
 - Resultado de contar pe. *personas en un hogar.*
 - El número de valores posibles entre dos valores dados es finito

Conceptos básicos

- Tipos de variables
 - **Cuantitativas:** según el número de valores que tengan en la escala
 - **Continuas:**
 - Puede tomar cualquier valor entre dos números (número de valores posibles es infinito).
 - Resultado de medir: *altura, peso, tamaño del piso, edad*
 - En la práctica medimos la altura con dos decimales, pero también se podría dar con tres decimales

Conceptos básicos

- Tipos de variables
 - **Cualitativas Ordinal** → Tratamiento estadístico:
 - Opción 1: como variables cualitativas
 - Opción 2: como variables cuantitativas:
 - “cambiando” las etiquetas por números
 - Los resultados del análisis pueden variar según qué números asignemos a los valores: métodos específicos para escalas ordinales

Conceptos básicos

- Tipos de variables
 - **Cuantitativas**
 - **Escala de medida**
 - Escala de intervalo
 - La distancia entre las unidades de medida es uniforme (D es el doble que A) y permite realizar operaciones matemáticas, como suma, resta, multiplicación o división.
 - El cero es arbitrario, no indica la ausencia de atributo.
 - Temperatura °C
 - Escala de razón
 - El cero sí indica la ausencia de atributo, es cero absoluto.
 - Altura (cm), peso (gr).

Conceptos básicos

- Tipos de variables

- **Variables cuantitativas:**

- La diferencia está difuminada
 - Variables que son resultado de “medir” (continuas), pero redondeamos y convertimos en número finito de valores enteros. *Ej: escala likert*



- Variables que son resultado de contar (discretas), pero que tienen muchísimos valores diferentes. *Ej ingresos, población de un municipio*

Conceptos básicos

- Variables cuantitativas:
 - En la práctica:
 - **Variables “discretas”:**
 - Cuantitativas con pocos valores distintos (ejemplo: escala ideológica).
 - Cualitativas ordinales.
 - **Variables “continuas”:** cuantitativas con muchos valores distintos (ejemplo: ingresos)

Conceptos básicos

- **Población:**

- Conjunto de elementos sometidos a un estudio estadístico
- Sinónimo: colectivo o universo.
- **Individuo:** cada uno de los elementos que componen una población
- **Tamaño:** número de elementos que la forman
 - Representada por la letra N .
 - El tamaño puede algunas veces considerarse infinito
→
 - Tamaño original tan grande que conviene tratarlo como compuesta por infinitos elementos.

Conceptos básicos

- **Muestra:** subconjunto representativo de una población de referencia
 - **Tamaño muestral** → número de valores de una variable menor que el número de valores correspondiente a su población

Conceptos básicos

- **Muestreo:** acción de seleccionar una muestra. Puede ser
 - **Aleatorio:** los elementos de la muestra se seleccionan al azar.
 - **No aleatorio (estratificado):** los elementos de la muestra se seleccionan siguiendo criterios (edad, nivel económico, procedencia) reproduciendo los caracteres de una población.
 - **Ventajas:** reducción de tiempo y costes. Imprescindible cuando la población es infinita o la prueba es destructiva (duración de un determinado tipo de bombilla).
 - **Inconvenientes** Conclusiones sometidas a un cierto grado de error, tanto mayor cuanto más pequeño es el tamaño de la muestra.