

Técnicas avanzadas de analítica visual

Ignacio Díaz Blanco



Grupo de Supervisión y Diagnóstico
de Procesos Industriales

Analítica Visual

Tres pilares básicos

VD → Visualización de Datos (VD)

IDA → Análisis Inteligente de Datos (IDA)

I → Interacción (I)

Visualización de datos

interiorizar la información

“Una imagen vale más que mil palabras”



Visualización de datos

parte importante del proceso cognitivo

Lo *veo*...

Está muy *claro*

Muéstrame por qué...

Ella *desveló*
el misterio

Esto arroja *luz*
al problema

Este autor es un
visionario

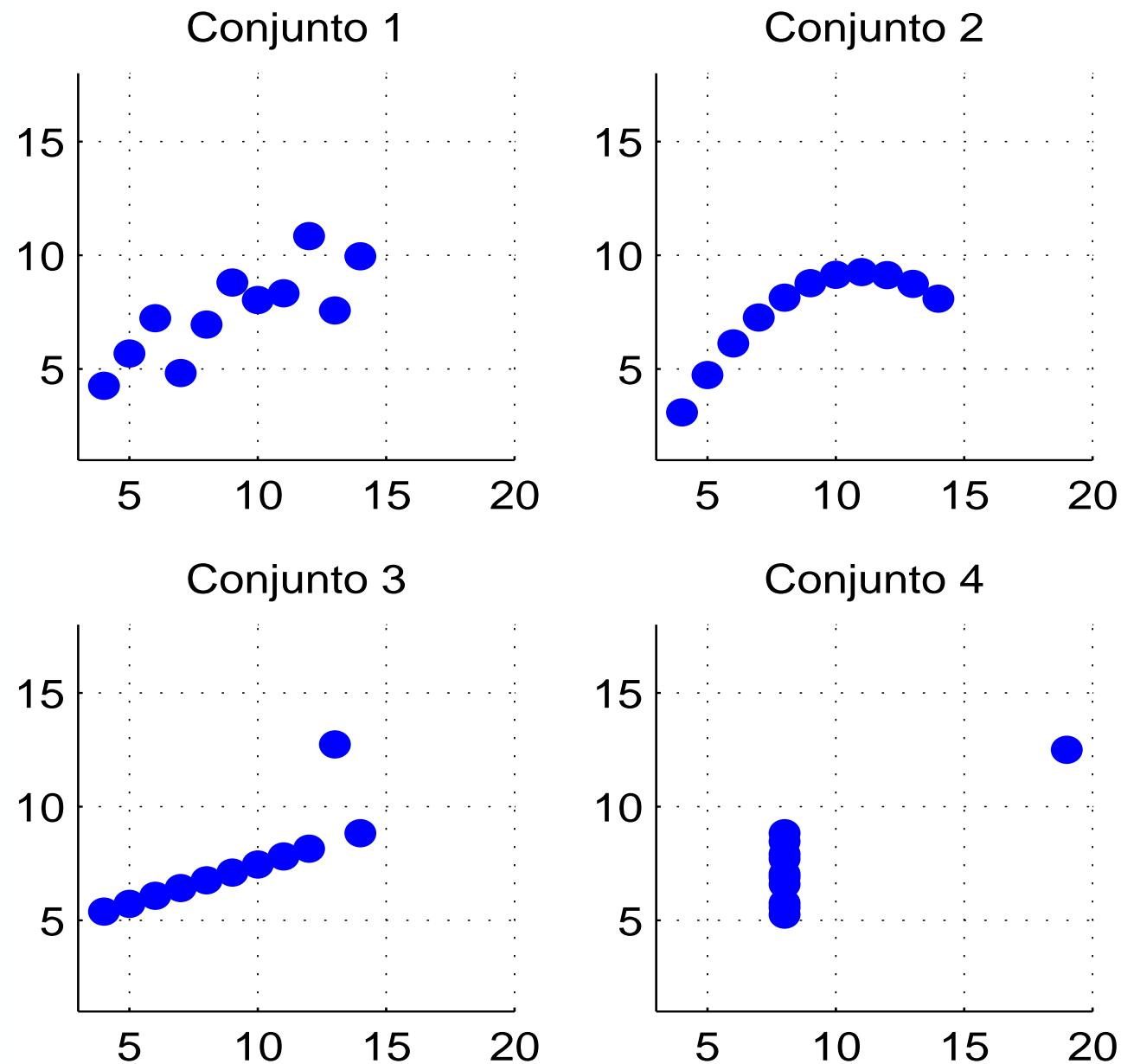
Su explicación
resultó algo *oscura*

Él siempre dice la verdad:
es una persona muy *transparente*

Visualización de datos

parte importante del proceso cognitivo

el 40 % del cerebro
está dedicado a **tareas visuales**



¡Los 4 conjuntos tienen los mismos estadísticos!

Matrices de covarianzas

```
>> for k = 1:4, C{k}, end;
ans =
    11.0000    5.5010
    5.5010    4.1273

ans =
    11.0000    5.5000
    5.5000    4.1276

ans =
    11.0000    5.4970
    5.4970    4.1226

ans =
    11.0000    5.4990
    5.4990    4.1232
```

Medias

```
>> for k = 1:4, M{k}, end;
ans =
    9.0000    7.5009

ans =
    9.0000    7.5009

ans =
    9.0000    7.5000

ans =
    9.0000    7.5009
```

```
clear;
close all;

% Datos del cuarteto Anscombe (http://en.wikipedia.org/wiki/Anscombe's\_quartet)
p = [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];

% Separamos los cuatro conjuntos
x{1} = p(:,1:2);
x{2} = p(:,3:4);
x{3} = p(:,5:6);
x{4} = p(:,7:8);

figure(1);
clf;

for k = 1:4,
    subplot(2,2,k);
    scatter(x{k}(:,1),x{k}(:,2),50,'filled');
    axis equal;
    axis([3 20 1 18]);
    grid on;
    C{k} = cov(x{k});
    M{k} = mean(x{k});
    R{k} = corrcoef(x{k});
    title(sprintf('Conjunto %d',k));
end
```

¿Cuántos “sietes” encuentras?

1093840192384901834093281481083459830985560984095809456894
0680946804986049860458094802943209385029348509898690285096
7890956890456809348560394563956034560584560935050358609358
6904690586034870909289091001012948019238401983401923109234
0480923489043859023850923896508309683098670395860394586039
4860934869034586093860938093468093458038034680934568039458
6034560386034563945839586358568039539568309458038505609568
0349568345809358045860958603758609458603583405684305680349
5680395680395840958034860394680938609348609685309456039465
0439580349758603496093458620943580281039803481894104391092

¿Cuántos “sietes” encuentras?

1093840192384901834093281481083459830985560984095809456894
0680946804986049860458094802943209385029348509898690285096
7890956890456809348560394563956034560584560935050358609358
6904690586034870909289091001012948019238401983401923109234
0480923489043859023850923896508309683098670395860394586039
4860934869034586093860938093468093458038034680934568039458
6034560386034563945839586358568039539568309458038505609568
0349568345809358045860958603758609458603583405684305680349
5680395680395840958034860394680938609348609685309456039465
0439580349758603496093458620943580281039803481894104391092

¿Cuántos “sietes” encuentras?

1093840192384901834093281481083459830985560984095809456894
0680946804986049860458094802943209385029348509898690285096
7890956890456809348560394563956034560584560935050358609358
6904690586034870909289091001012948019238401983401923109234
0480923489043859023850923896508309683098670395860394586039
4860934869034586093860938093468093458038034680934568039458
6034560386034563945839586358568039539568309458038505609568
0349568345809358045860958603758609458603583405684305680349
5680395680395840958034860394680938609348609685309456039465
0439580349758603496093458620943580281039803481894104391092

los has detectado en < 200 ms
¡¡entre cientos de números!!
“cero esfuerzo” para el cerebro

¿Cuántos “sietes” encuentras?

1093840192384901834093281481083459830985560984095809456894
0680946804986049860458094802943209385029348509898690285096
7890956890456809348560394563956034560584560935050358609358
6904690586034870909289091001012948019238401983401923109234
0480923489043859023850923896508309683098670395860394586039
4860934869034586093860938093468093458038034680934568039458
6034560386034563945839586358568039539568309458038505609568
0349568345809358045860958603758609458603583405684305680349
5680395680395840958034860394680938609348609685309456039465
0439580349758603496093458620943580281039803481894104391092

además, la visualización ha aportado “contexto”

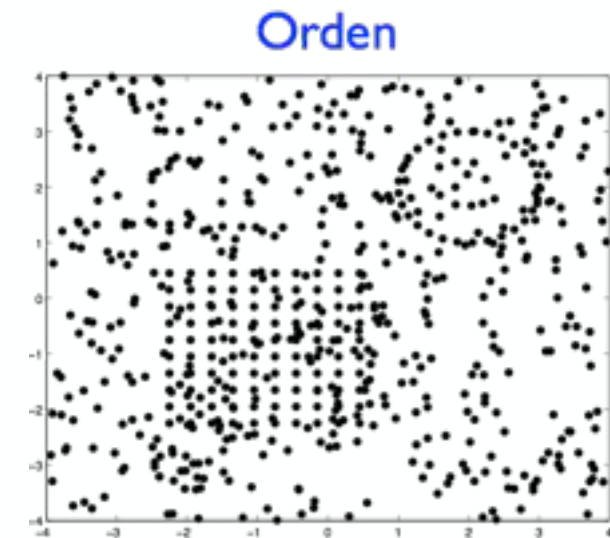
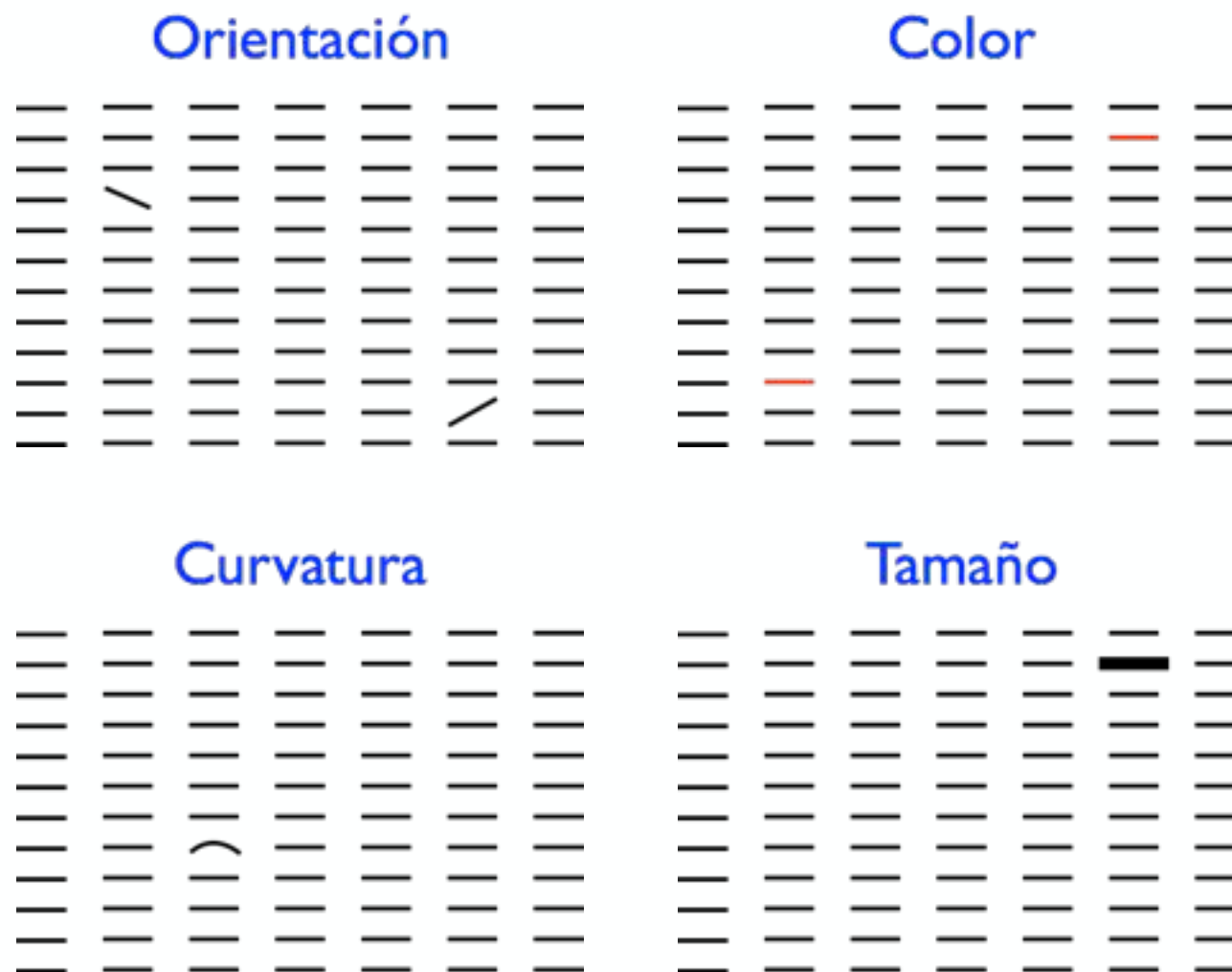
localización, cadencia, patrones de ocurrencias, proporción, ...

la visualización combina formas sutiles de conocimiento

conocimiento conectado = más conocimiento

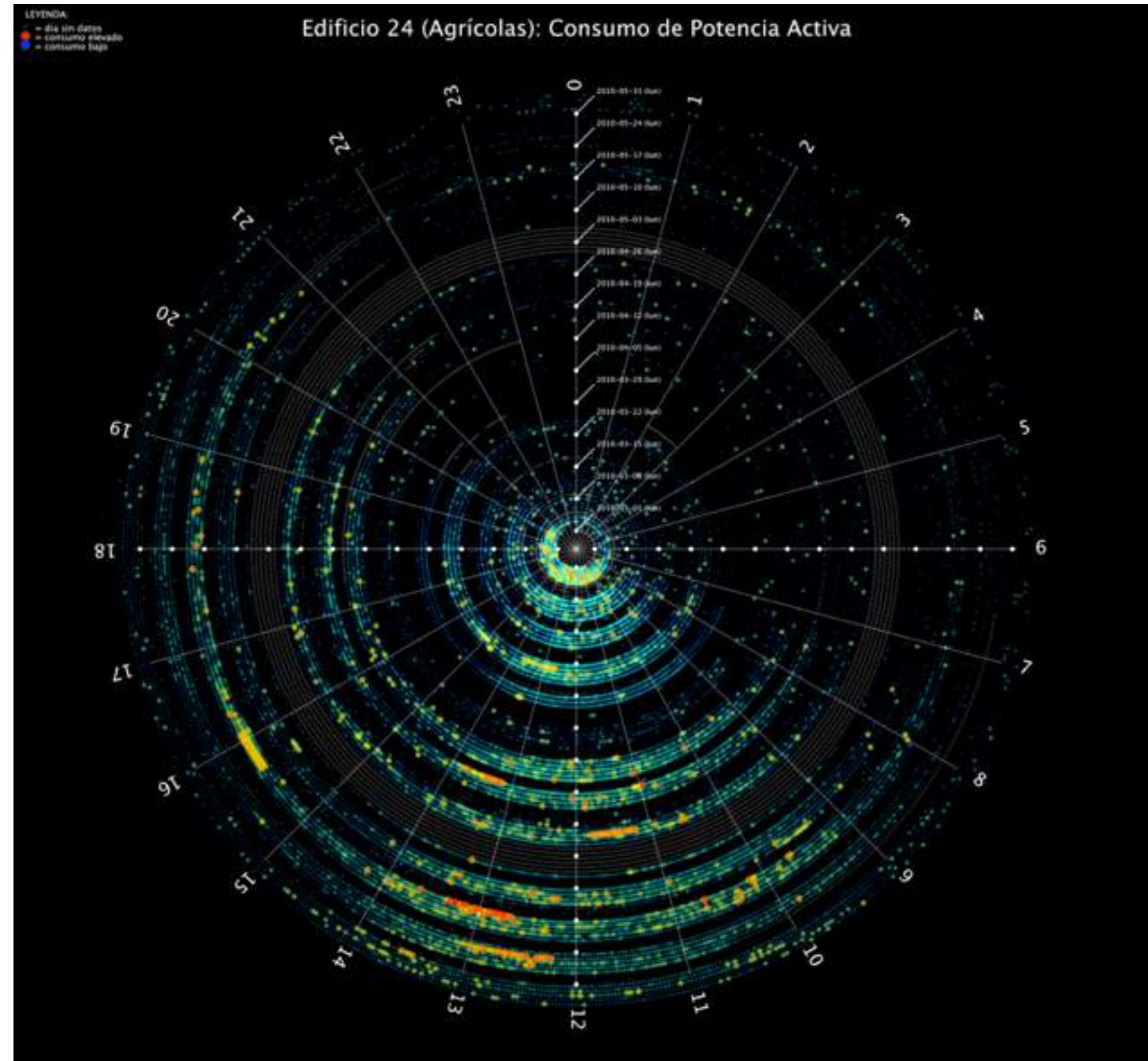
Mecanismos pre-atención (cortex VI)

- Baja latencia (muy rápidos ~ 200 ms),
- Masivamente paralelo: no depende de distractores
- “Cero” esfuerzo: no afectan a proceso cognitivo superior



Visualización de datos

reorganizando datos → emergen patrones



datos de consumos eléctricos en Unileón
desarrollado en el marco de DPI2015-69891-C2-2-R

Visualización de datos

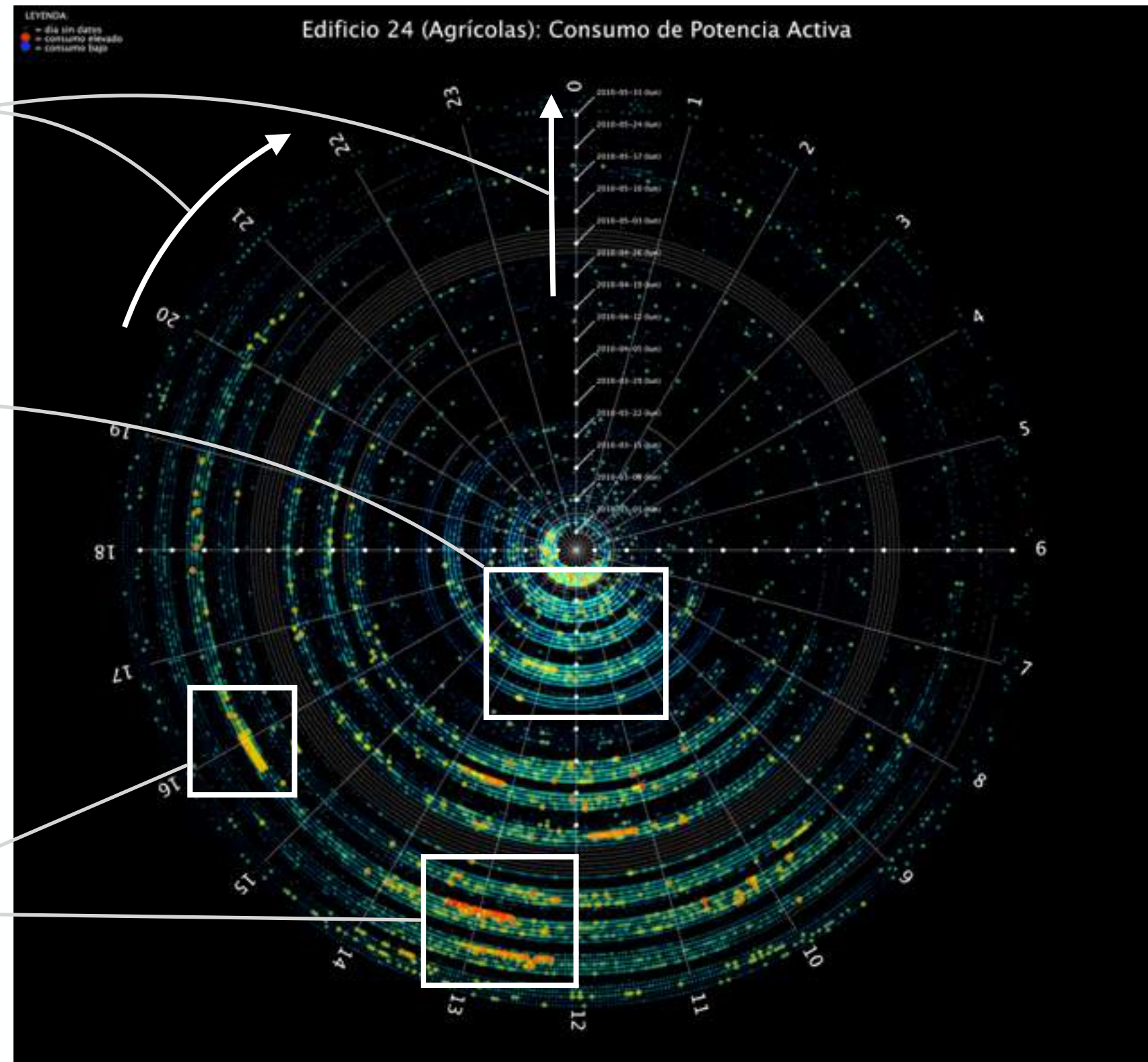
reorganizando datos → emergen patrones

reorganización
espacial
de la información
basada en
conocimiento *a priori*
(periodicidades
temporales)

el “orden” produce
patrones reconocibles

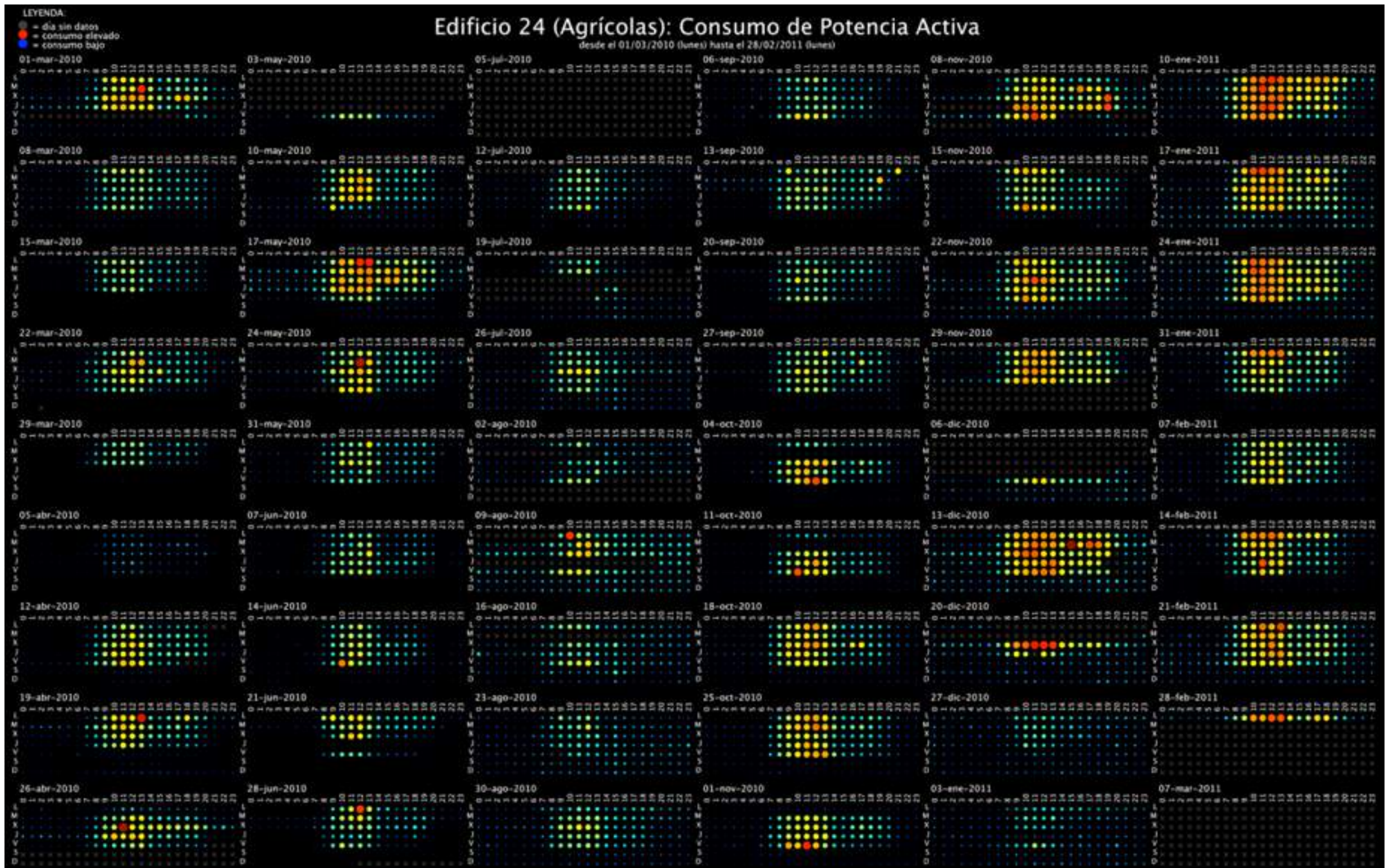
el **contexto temporal**
permite intepretarlos

preatención
detección visual
inmediata
de patrones
inusuales



Visualización de datos

reorganizando datos → emergen patrones



Visualización de datos

aporta contexto, conecta dominios

Mapas

fuerte conocimiento previo
lugares, rincones, recuerdos ...

+

Datos

consumos, ventas, contaminación,
ruido, ...

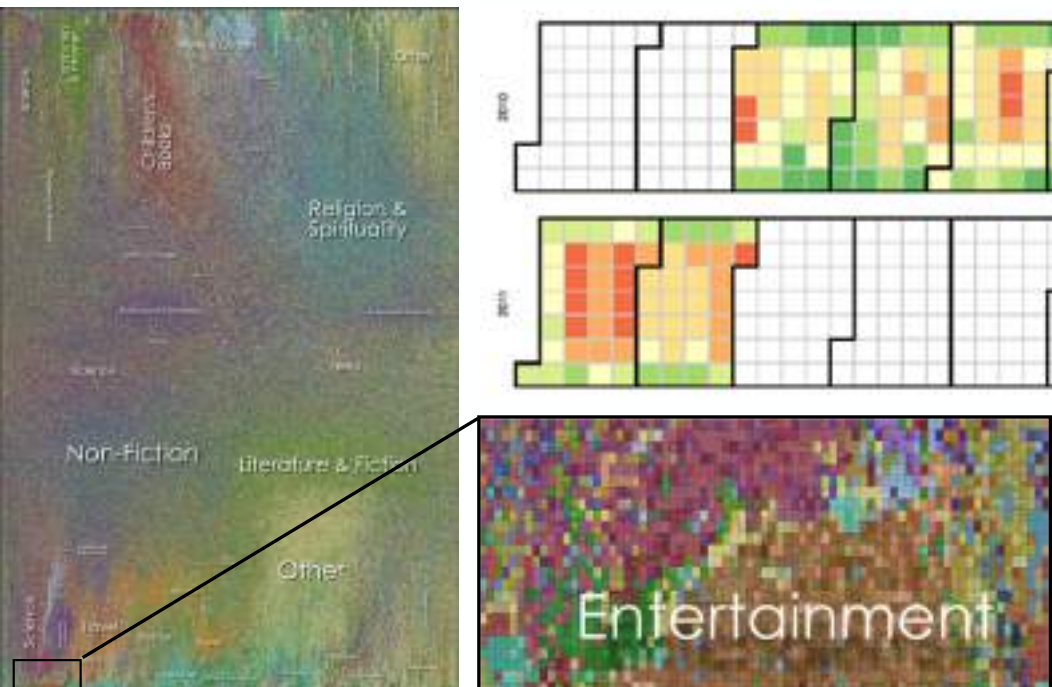
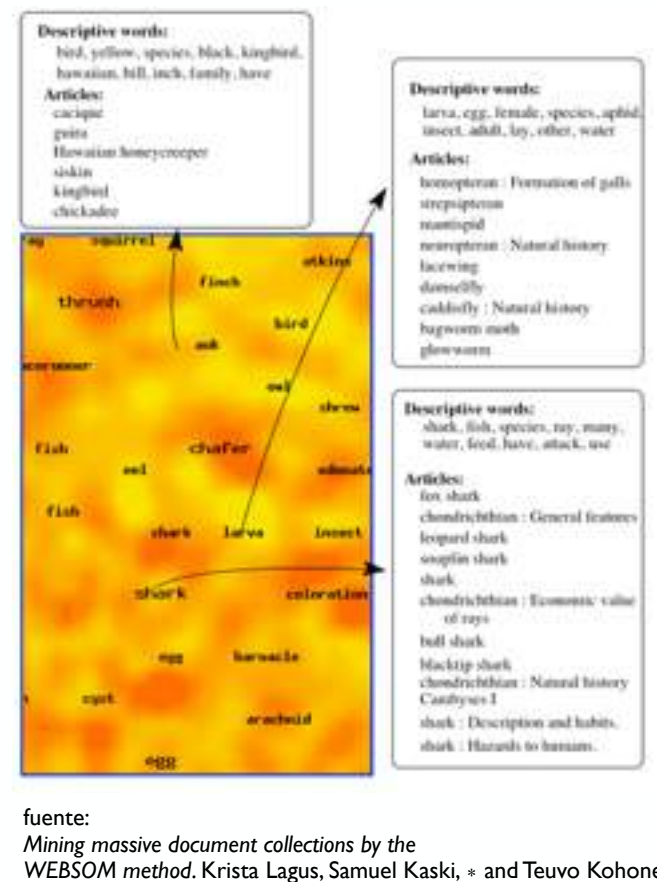
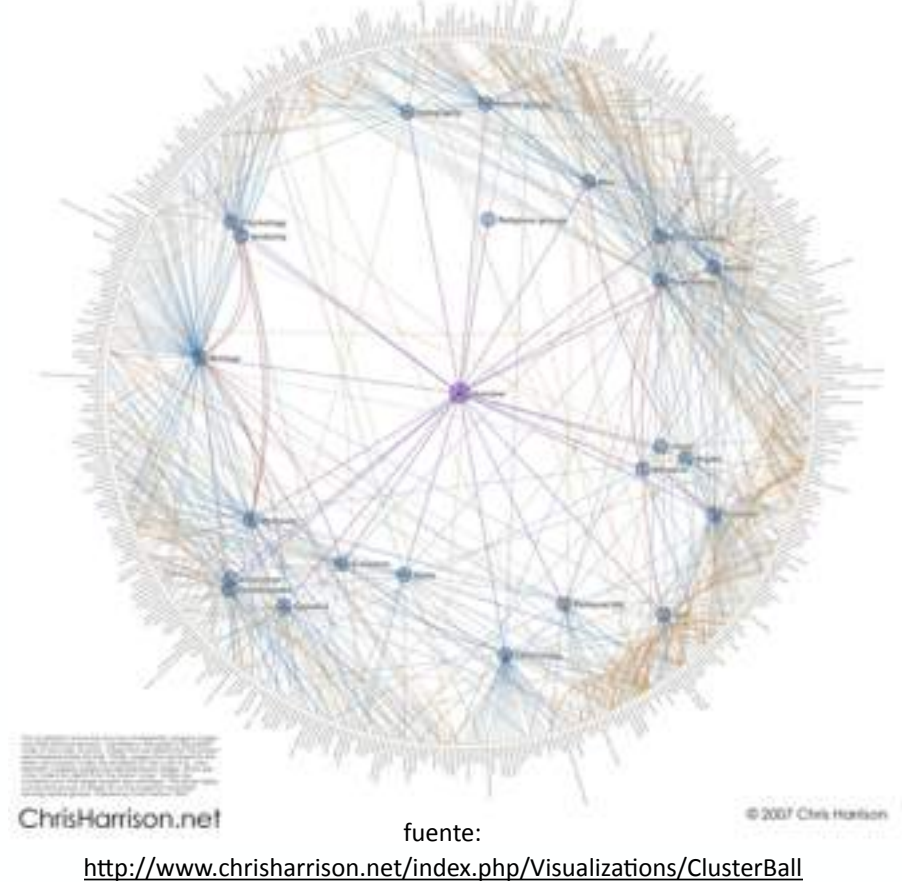
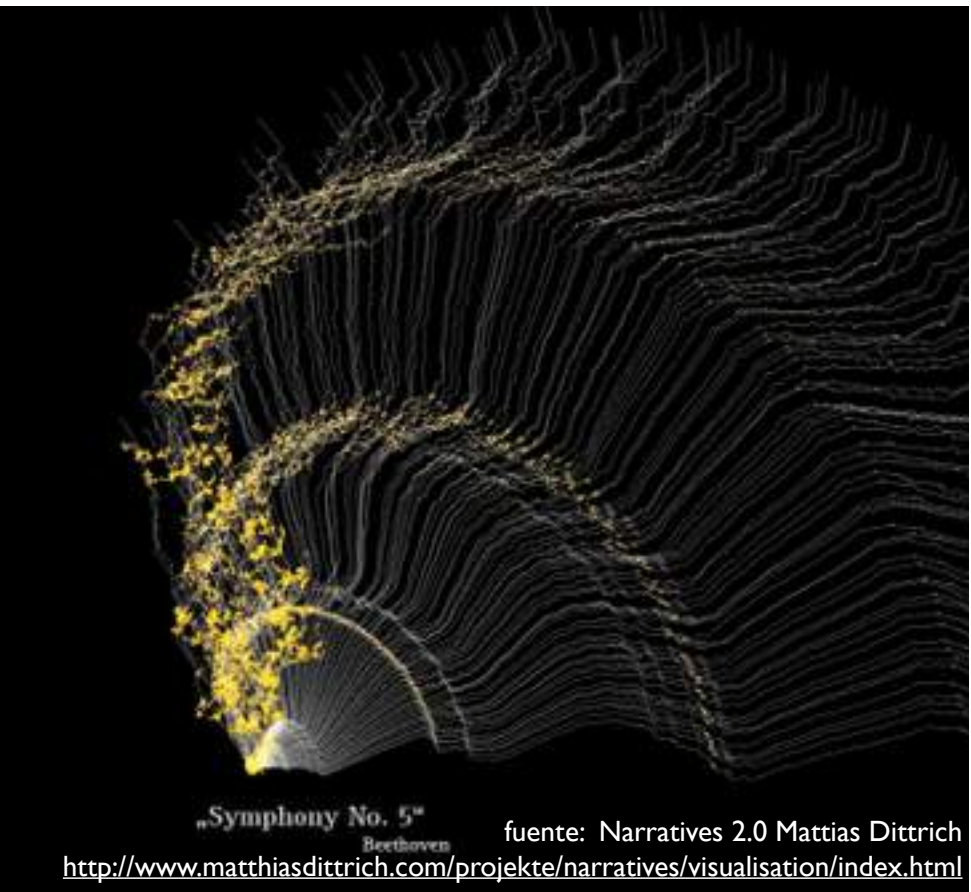
→

juntos valen más
que su suma

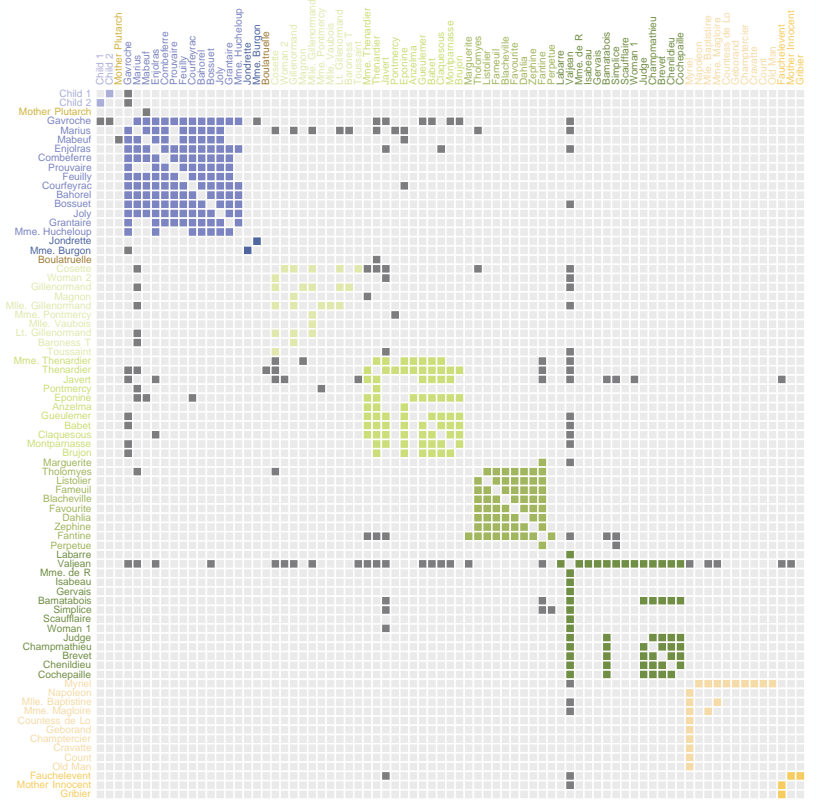


Visualización de datos

otros ejemplos



datos de consumos eléctricos en UniLeón
desarrollado en el marco de DPI2015-69891-C2-2-R



Análisis de Datos Inteligente (IDA)

detectando estructura en los datos

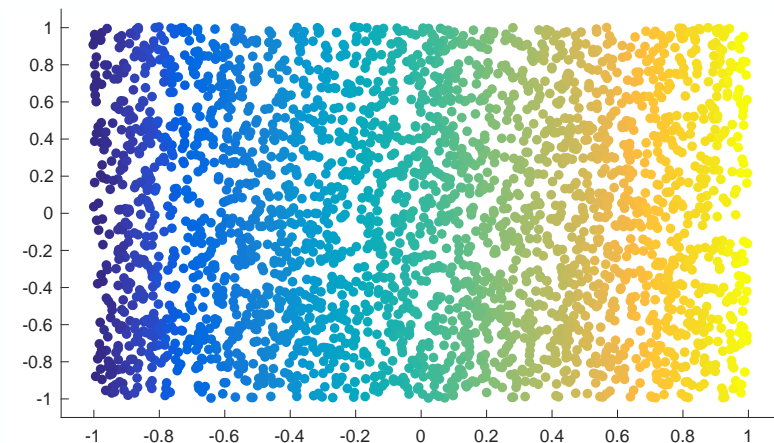
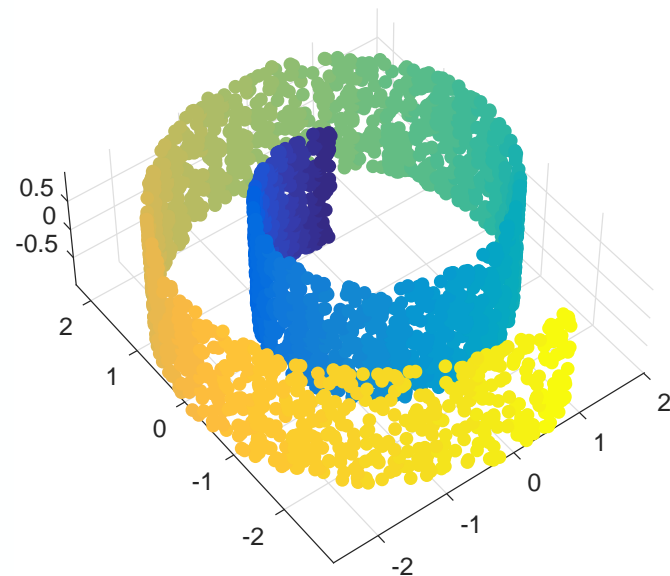
Análisis inteligente de datos

estructura en los datos → conocimiento

Búsqueda “algorítmica” de estructura en los datos

Desenredo de datos en espacios multidimensionales (“disentangling”)

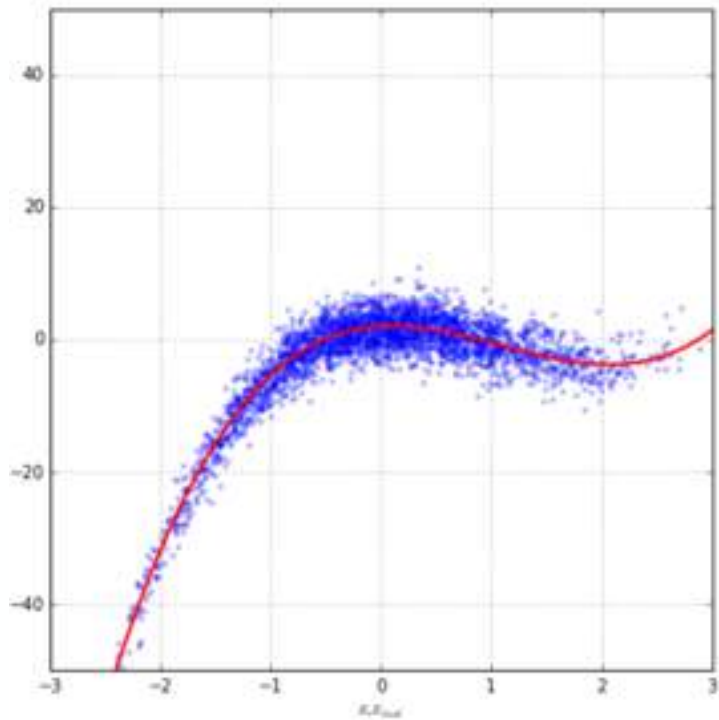
Técnicas de reducción de la dimensionalidad... pero no solo ellas!



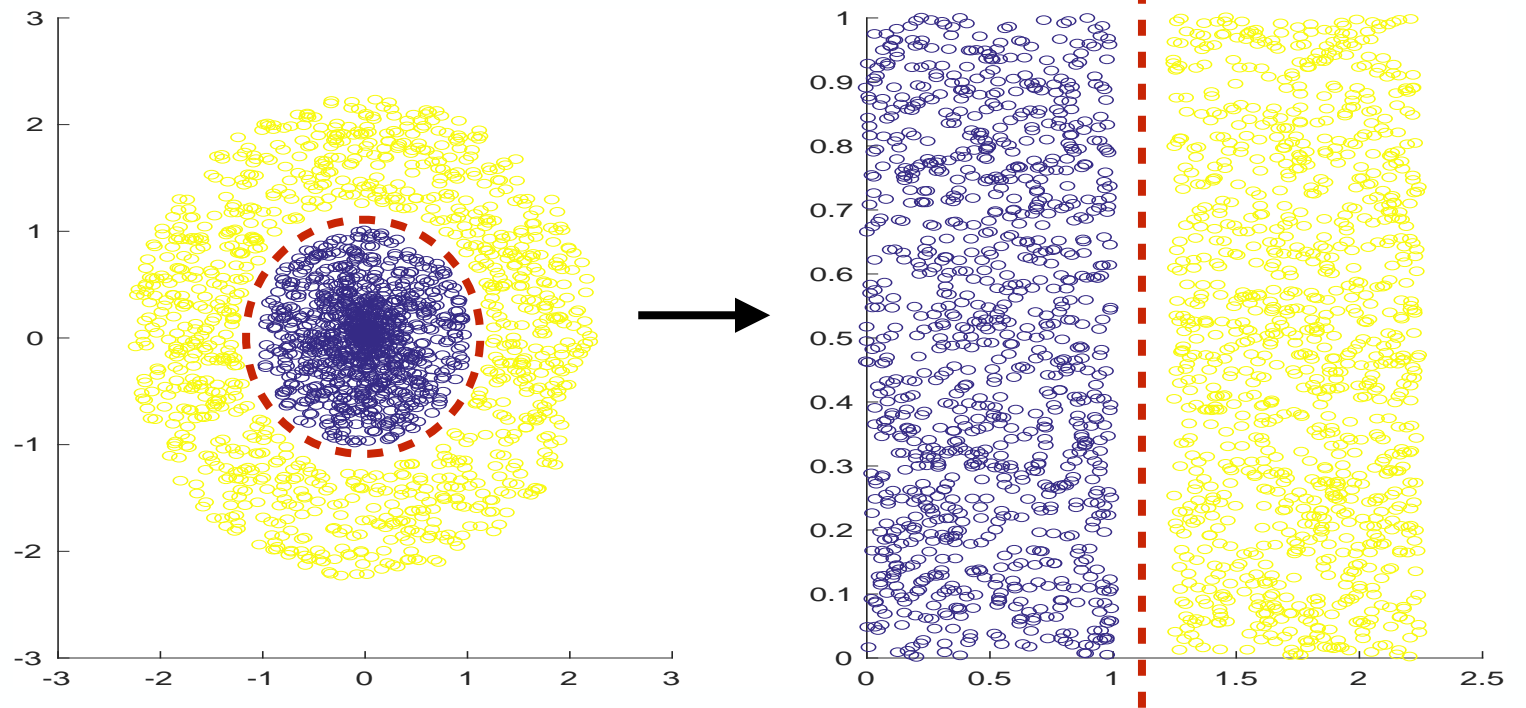
Análisis inteligente de datos

estructura en los datos → conocimiento

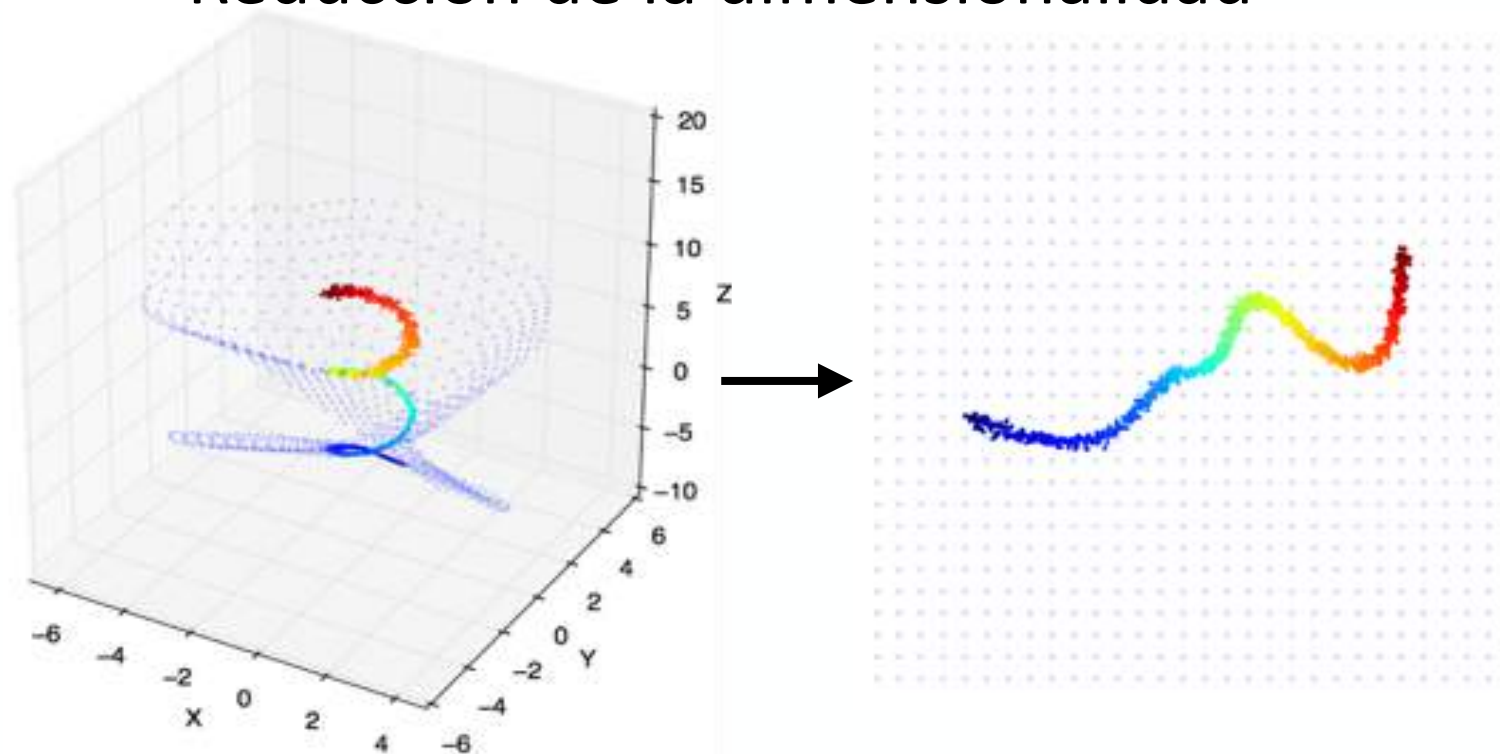
Regresión



Clasificación

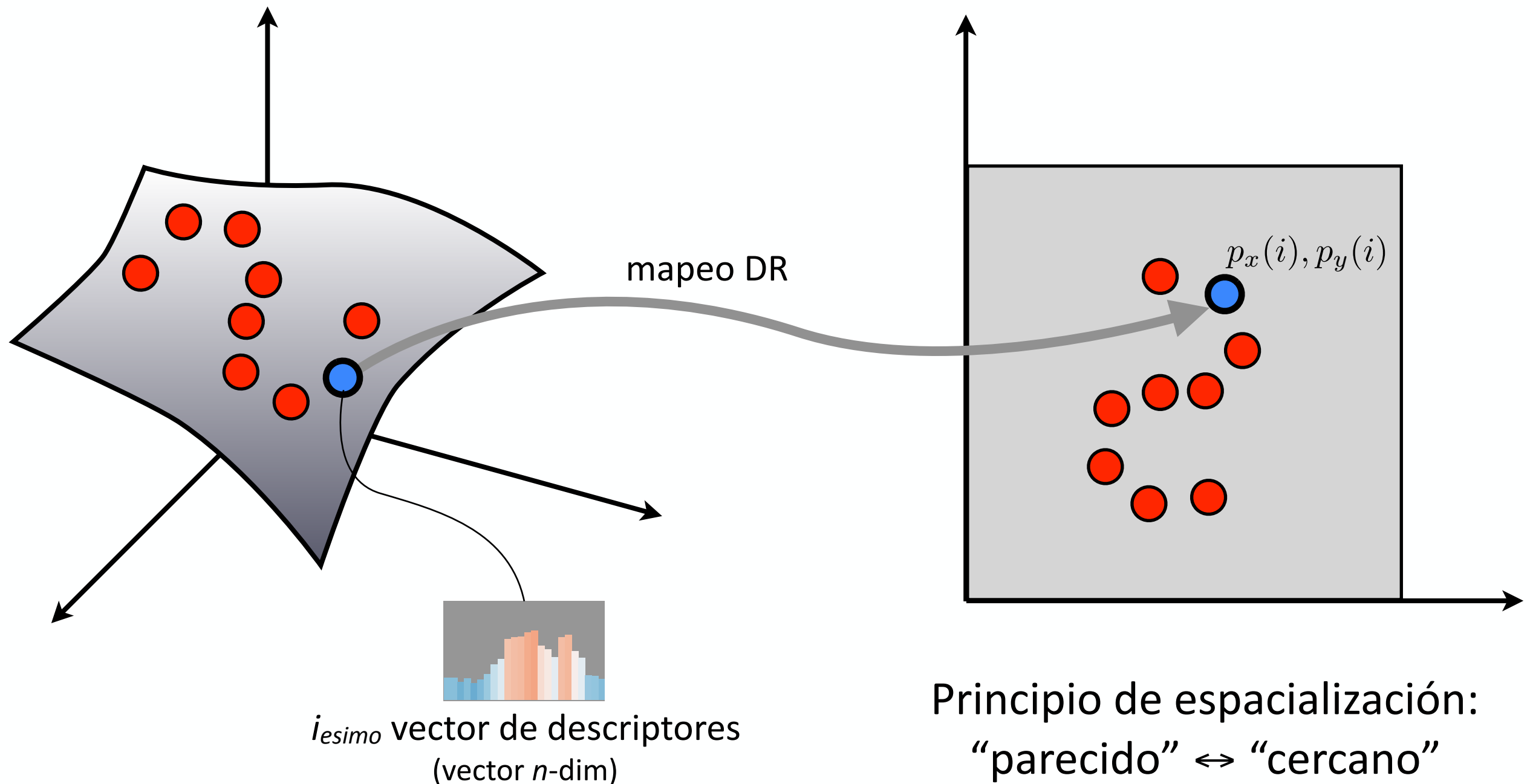


Reducción de la dimensionalidad



**espacio multidimensional
con vectores de descriptores R^n**

**espacio 2D
de visualización**

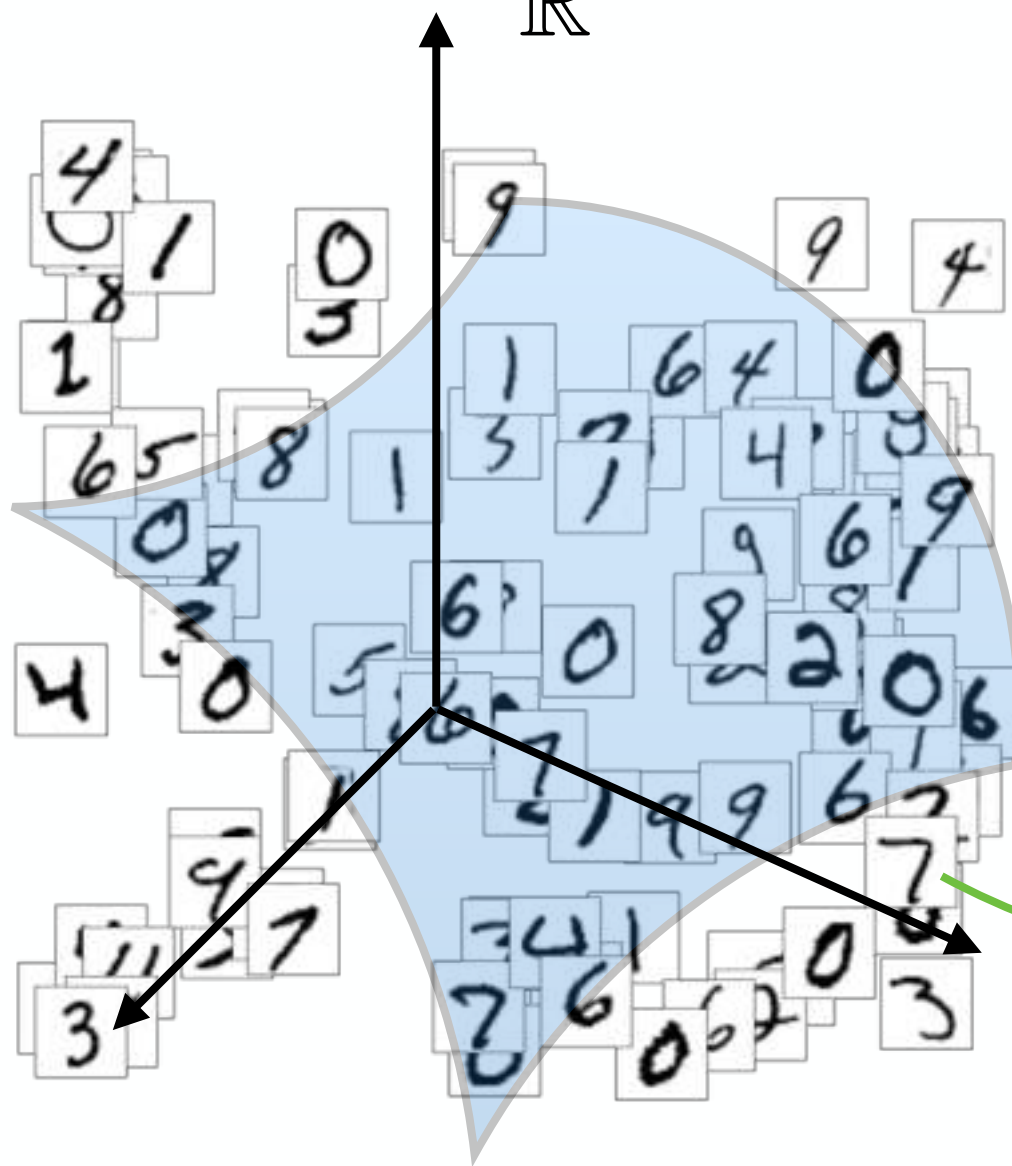


Principio de espacialización:
“parecido” ↔ “cercano”

Espacio de datos

(alta dimensionalidad)

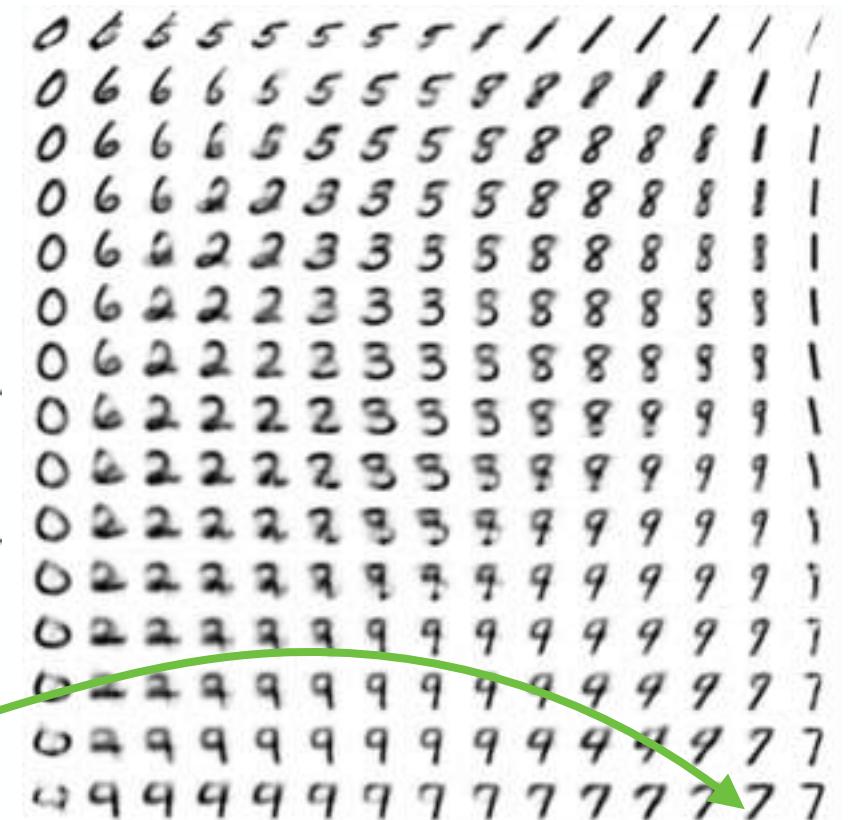
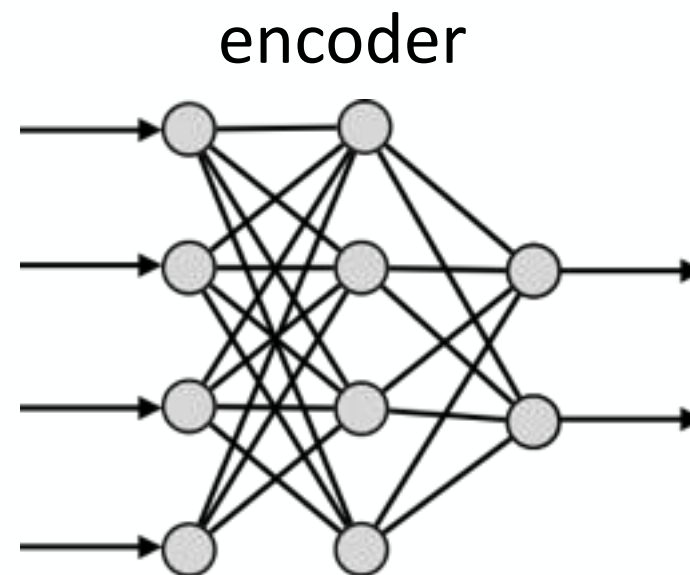
\mathbb{R}^{784}



Espacio latente

(baja dimensionalidad)

\mathbb{R}^2



Interacción

feedback en el proceso de análisis

Interacción

“feedback” en el proceso de análisis

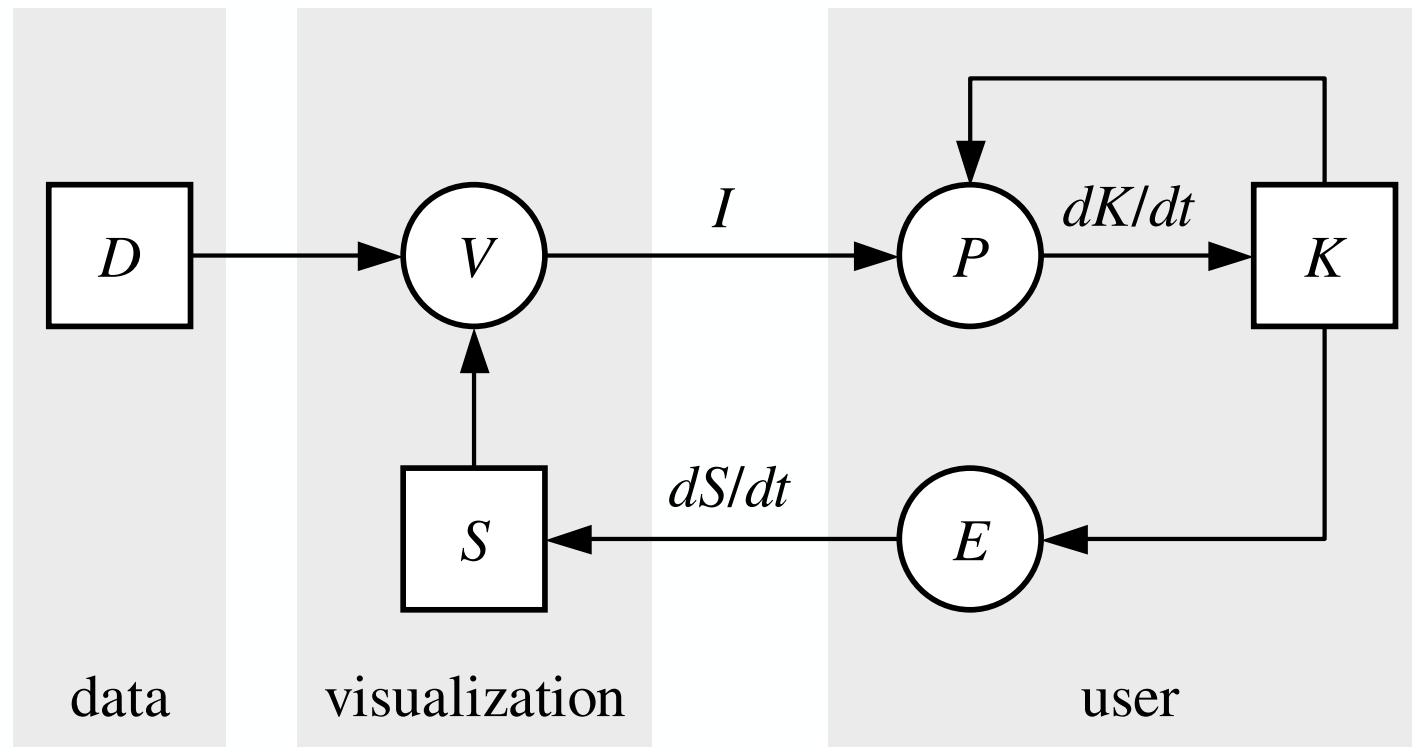


Figure 1: A simple model of visualization

K = conocimiento
P = percepción
V = interfaz visual
D = datos
S = especificaciones de la vis.
I = imagen
E = Exploración interactiva

La interfaz visual V produce una imagen I en función de los datos D y las especificaciones S

$$I(t) = V(D, S, t)$$

Evolución del conocimiento durante el proceso de visualización interactiva

$$K(t) = K_0 + \int_0^t P(I, K, t) dt$$

El usuario “modula” las especificaciones de la visualización en función de su conocimiento

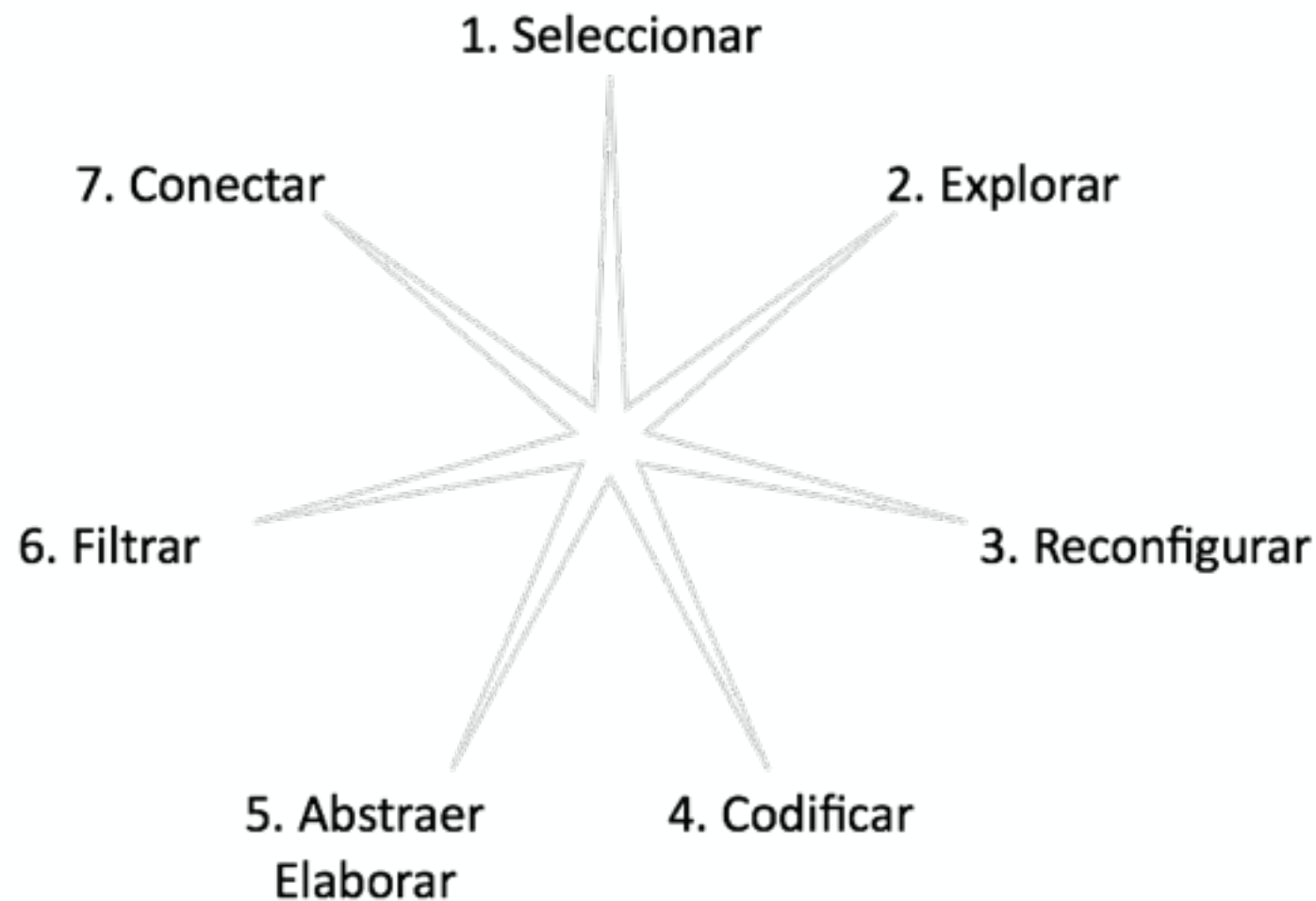
$$S(t) = S_0 + \int_0^t E(K) dt$$

Fuente

J. Van Wijk. The value of visualization. In *16th IEEE Visualization 2005 (VIS 2005)*. IEEE Computer Society, 2005.

Interacción

las 7 categorías de interacción



1. Seleccionar

marcar algo como interesante

2. Explorar

mostrar algo más

3. Reconfigurar

ordenar las cosas de otra forma

4. Codificar

mostrar una representación diferente

5. Abstraer/Elaborar

mas detalle o menos detalle (big picture)

6. Filtrar

ver sólo lo que cumple una condición

7. Conectar

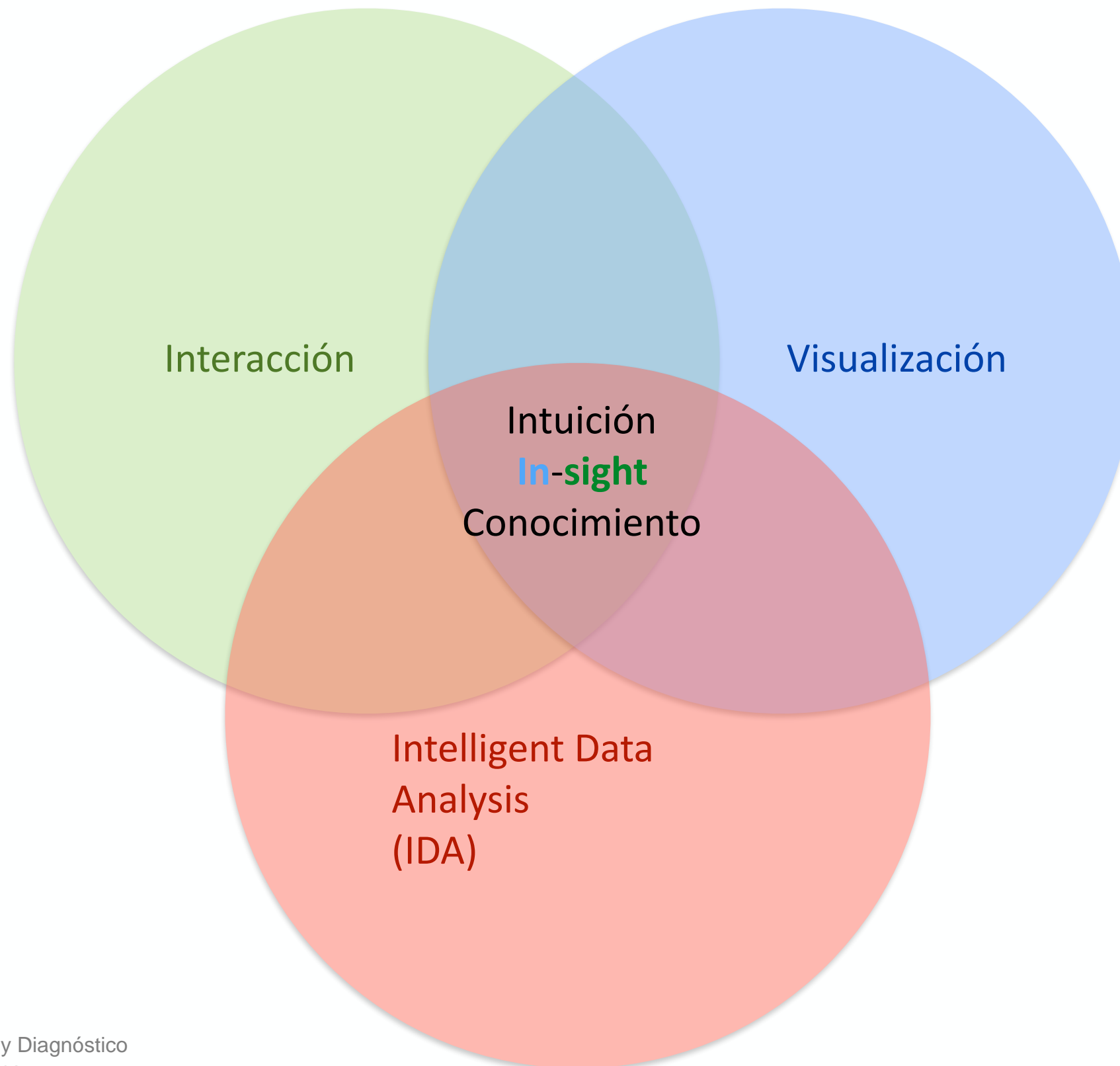
mostrar los items relacionados

Analítica Visual

sinergia: VD + IDA + I

Analítica Visual

aprendizaje humano + machine learning



Analítica Visual

varios niveles de integración VD + IDA + I

visualización estática de datos

Idea

Visualización datos en un póster
o estática en pantalla de ordenador



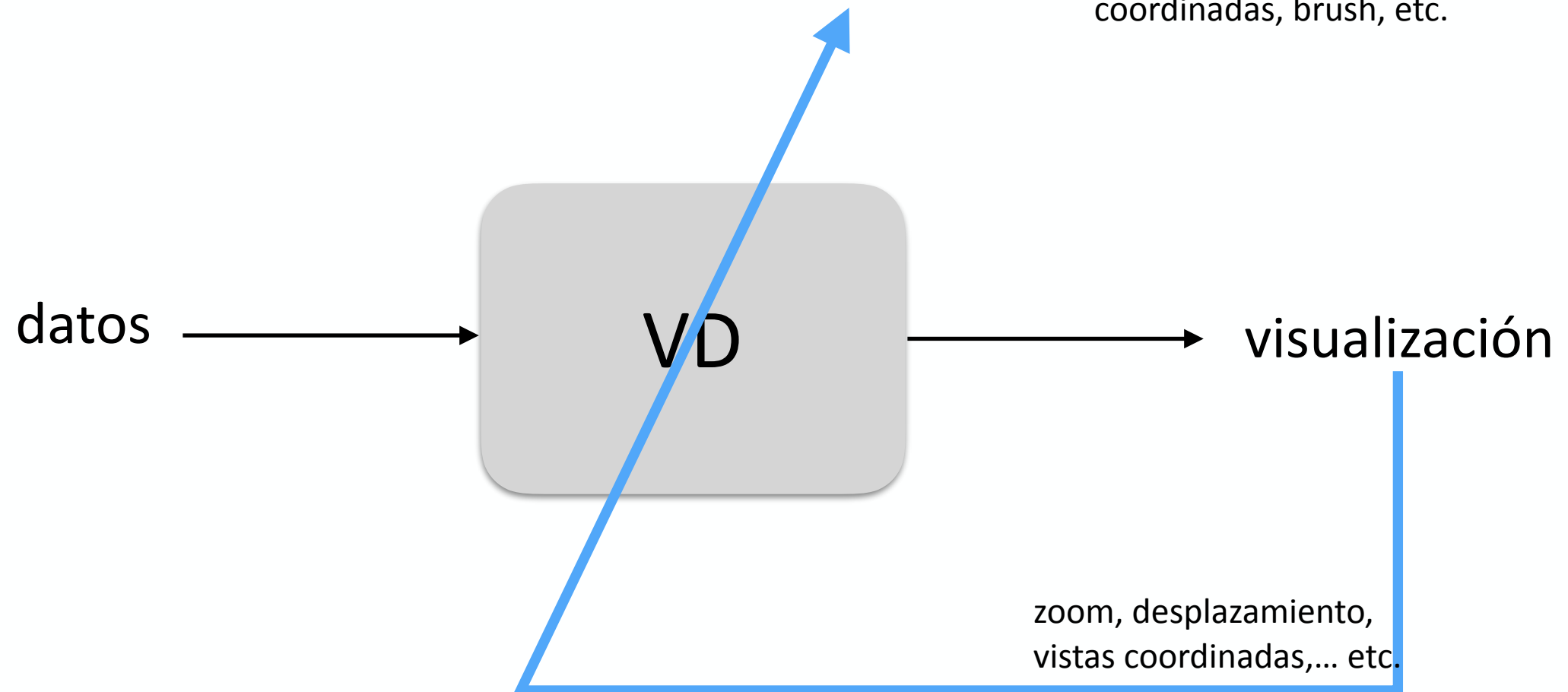
Analítica Visual

varios niveles de integración VD + IDA + I

visualización interactiva de datos

Idea

Visualización de datos en pantalla de ordenador con funciones de interacción típicas: zoom, desplazamiento, vistas coordinadas, brush, etc.



zoom, desplazamiento,
vistas coordinadas,... etc.

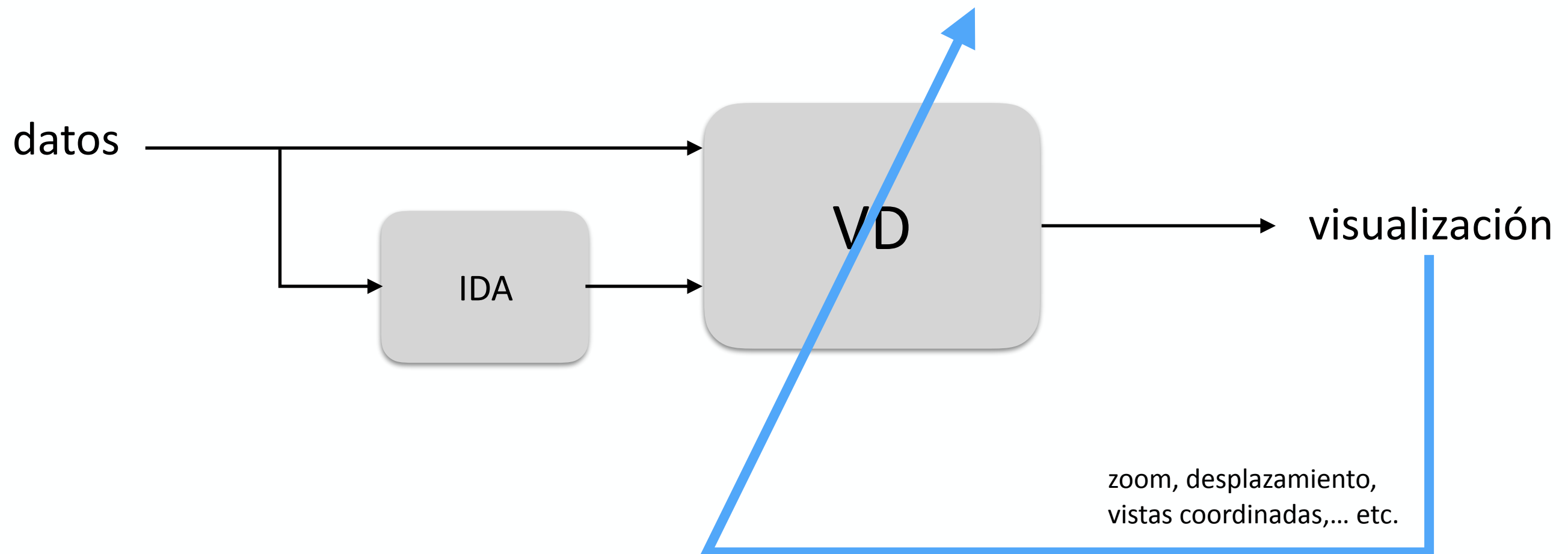
Analítica Visual

varios niveles de integración VD + IDA + I

visualización interactiva de datos + resultados de IDA

Idea

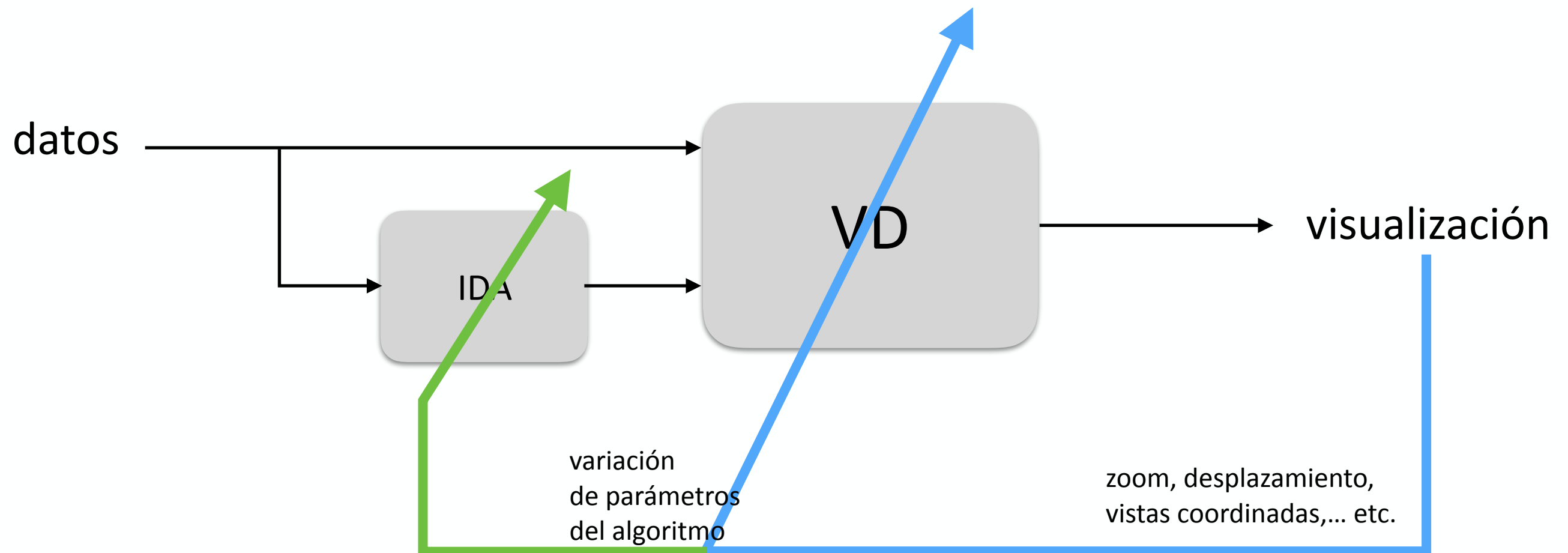
Visualización de datos
+ resultados de un algoritmo IDA
en pantalla de ordenador con funciones de
interacción típicas: zoom, desplazamiento,
vistas coordinadas, brush, etc



visualización interactiva de datos + manipulación interactiva de IDA

Idea

además de las funciones de interacción típicas (zoom, etc.) el usuario puede manipular el comportamiento del IDA variando sus parámetros para ver resultados en otros contextos



Algunos ejemplos

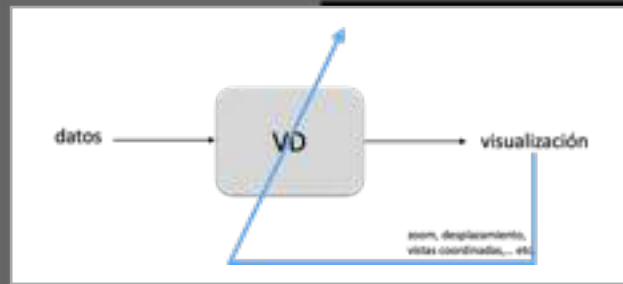


tabla de datos (csv, xls,...)

Table 1: dimension attributes used in the interface

id	label	attribute
1	week	week of year
2	hour	hour
3	Pact (KW)	active power (kW)
4	month	month
5	T(C)	temperature
6	Edificio	building id
7	dayofyear	day of year
8	weekday	day of the week
9	THD Van (%)	THD (%)
10	cosPhi	power factor (cos ϕ)

Table 2: list of buildings

id	Building name
5	Filosofía
15	Animalario
16	INCAFD
17	Pabellón Deportivo
18	Frontón
19	Biblioteca Central
21	Cafetería II
22	Molecular
23	Complejo Agrícolas
24	Colegio Mayor
25	Complejo Rectorado
27	Minas
28	Centro Idiomas

		Edificio	Pact (KW)	cosPhi	THD Van (%)	T(C)	hour	week	weekday	month	dayofyear
2010-03-01	00:00:00	5	14.425620	0.953855	0.0	7.823652	0	9	0	3	60
2010-03-01	01:00:00	5	14.665796	0.954259	0.0	7.126155	1	9	0	3	60
2010-03-01	02:00:00	5	12.007591	0.933131	0.0	7.488719	2	9	0	3	60
2010-03-01	03:00:00	5	12.608106	0.939845	0.0	7.710455	3	9	0	3	60
2010-03-01	04:00:00	5	16.145092	0.957809	0.0	7.114933	4	9	0	3	60
2010-03-01	05:00:00	5	21.638385	0.952554	0.0	6.273484	5	9	0	3	60
2010-03-01	06:00:00	5	19.207911	0.938453	0.0	5.830762	6	9	0	3	60
2010-03-01	07:00:00	5	29.815537	0.943666	0.0	5.586618	7	9	0	3	60
2010-03-01	08:00:00	5	62.869654	0.905892	0.0	5.677350	8	9	0	3	60
2010-03-01	09:00:00	5	96.999449	0.893683	0.0	7.221909	9	9	0	3	60
2010-03-01	10:00:00	5	115.899338	0.902373	0.0	8.914559	10	9	0	3	60
2010-03-01	11:00:00	5	115.750337	0.898621	0.0	10.785449	11	9	0	3	60
2010-03-01	12:00:00	5	113.689223	0.887986	0.0	11.807558	12	9	0	3	60
2010-03-01	13:00:00	5	97.973818	0.873983	0.0	12.523772	13	9	0	3	60
2010-03-01	14:00:00	5	62.232332	0.894678	0.0	13.008751	14	9	0	3	60
2010-03-01	15:00:00	5	47.559666	0.886473	0.0	12.905749	15	9	0	3	60
2010-03-01	16:00:00	5	61.865053	0.877000	0.0	12.307785	16	9	0	3	60
2010-03-01	17:00:00	5	72.128616	0.881834	0.0	11.948900	17	9	0	3	60
2010-03-01	18:00:00	5	76.023250	0.899110	0.0	11.038489	18	9	0	3	60
2010-03-01	19:00:00	5	67.845450	0.899951	0.0	10.443579	19	9	0	3	60
2010-03-01	20:00:00	5	45.527110	0.926689	0.0	9.490753	20	9	0	3	60
2010-03-01	21:00:00	5	19.650929	0.967985	0.0	8.646537	21	9	0	3	60
2010-03-01	22:00:00	5	13.698341	0.945193	0.0	7.629666	22	9	0	3	60
2010-03-01	23:00:00	5	12.817378	0.946495	0.0	6.866202	23	9	0	3	60
2010-03-02	00:00:00	5	13.934914	0.954740	0.0	5.837988	0	9	1	3	61
2010-03-02	01:00:00	5	13.270300	0.950812	0.0	4.818204	1	9	1	3	61
2010-03-02	02:00:00	5	12.724814	0.946554	0.0	4.202187	2	9	1	3	61
2010-03-02	03:00:00	5	12.614464	0.947774	0.0	4.148451	3	9	1	3	61
2010-03-02	04:00:00	5	13.122296	0.950718	0.0	3.681740	4	9	1	3	61
2011-02-28	17:00:00	28	32.808913	0.929433	2.587	9.584414	17	9	0	2	59
2011-02-28	18:00:00	28	35.892543	0.917200	2.366	8.495864	18	9	0	2	59
2011-02-28	19:00:00	28	31.729217	0.922733	2.340	7.332595	19	9	0	2	59
2011-02-28	20:00:00	28	27.129239	0.889200	2.340	6.175993	20	9	0	2	59
2011-02-28	21:00:00	28	23.596643	0.933500	2.392	5.719882	21	9	0	2	59
2011-02-28	22:00:00	28	7.263327	0.984700	2.366	5.617063	22	9	0	2	59
2011-02-28	23:00:00	28	5.602116	0.985567	2.353	4.965097	23	9	0	2	59

La tabla equivale a un
hipercubo de datos

(aquí sólo mostramos tres atributos)

$$C(a_1, a_2, \dots, a_n)$$

registros

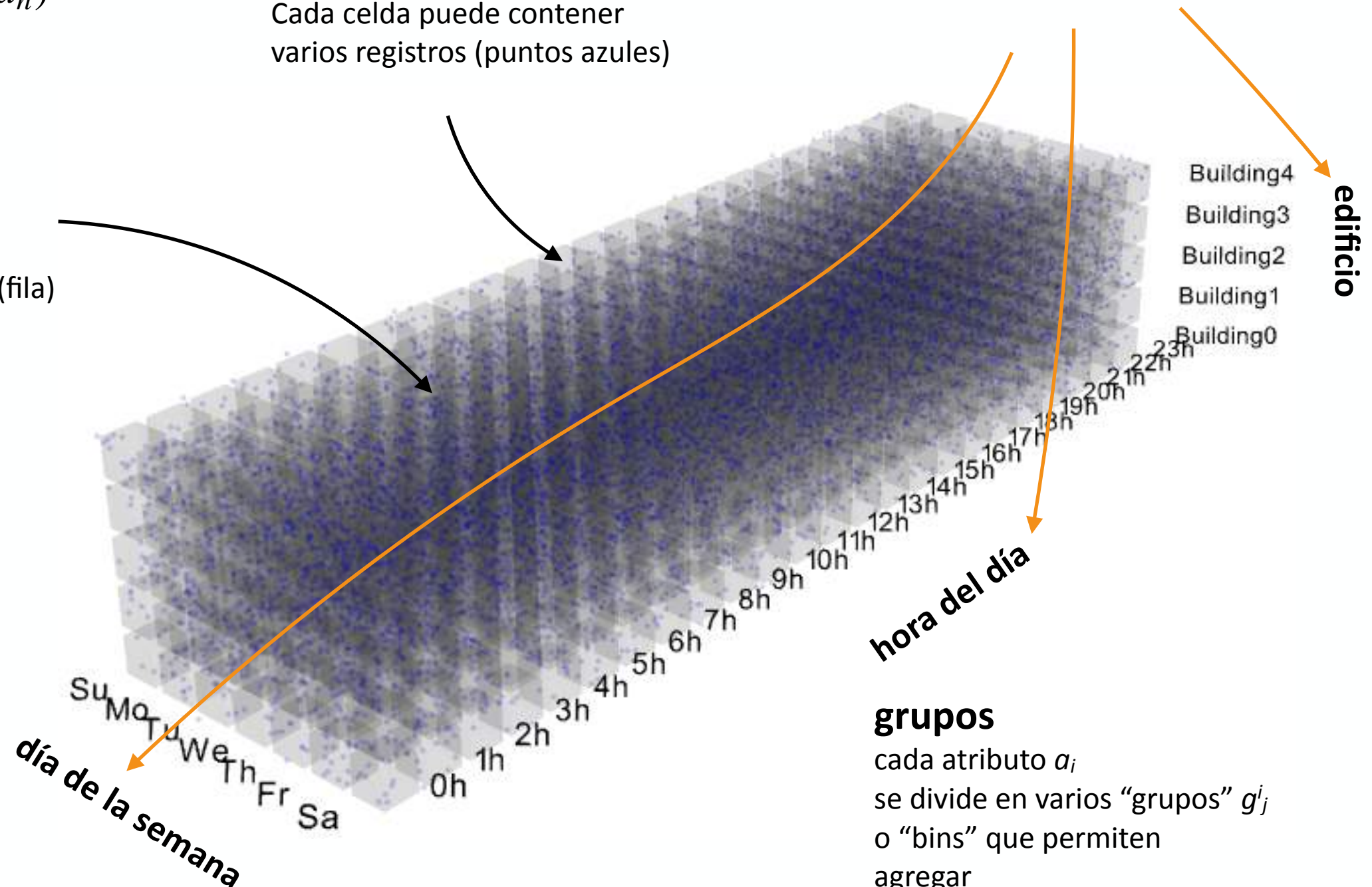
cada punto azul
representa un registro (fila)
de la tabla

celdas

cada celda (dado gris) viene definida
por una tupla de n grupos, i.e.
un grupo de cada uno
de los atributos del cubo
Cada celda puede contener
varios registros (puntos azules)

atributos

el hipercubo tiene
varias dimensiones
una por cada atributo
 a_1, a_2, \dots, a_n



grupos

cada atributo a_i
se divide en varios "grupos" g_j^i
o "bins" que permiten
agregar

Cubos de datos interactivos

Operaciones típicas: selección y proyección

Selección σ

toma un cubo de datos
y devuelve otro cubo
conteniendo un
subconjunto
de los registros

$$\sigma_{\substack{a_k=a'_k \\ a_j=a'_j}} [C(a_1, a_2, \dots, a_n)] \longrightarrow$$

$$\longrightarrow C(a_1, \dots, \underbrace{\{a, b\}}_{a'_k}, \dots, \underbrace{\{p, q, r\}}_{a'_j}, \dots, a_n)$$

La selección equivale a una
operación de **filtrado**

Casos particulares

Operaciones *dice* y *slice*

records #	attributes					measures		
	a1	a2	a3	a4	a5	μ1	μ2	μ3
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								

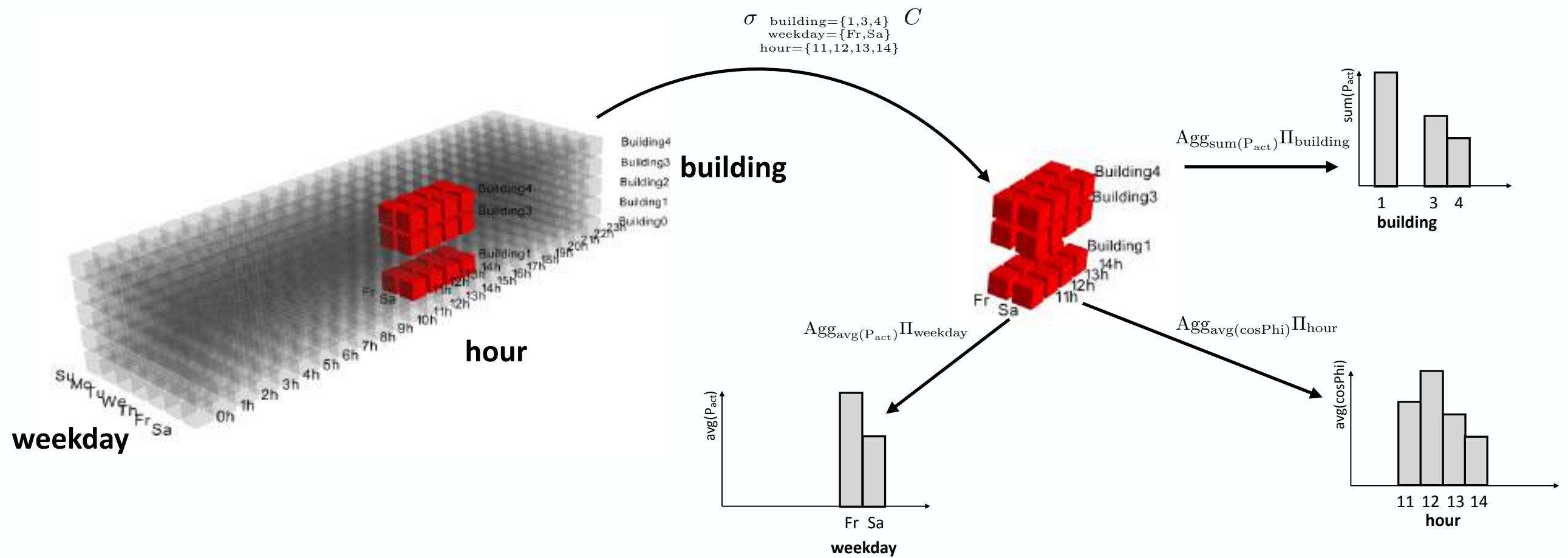
Proyección Π

toma un cubo de datos
y devuelve otro cubo
con un subconjunto de los atributos

$$\Pi_{a_{i_1}, \dots, a_{i_p}} [C(a_1, a_2, \dots, a_n)] \longrightarrow C(a_{i_1}, a_{i_2}, \dots, a_{i_p})$$

Cubos de datos interactivos

Operaciones típicas: visión geométrica



iHistograms: interactive Histograms

Problem description

This demo shows interactive histograms of several process variables. It may also shows other context or helper variables such as weekday, day hour, process state, revealing the time sequence correlations with the help of fluid interaction.

Interaction mechanisms

The user can have interaction by filtering histograms. Upon changes in the filters, the histograms are fluidly updated in a seamless way with low latency. The user can modify the filter range (limits) in a number of ways with a *brush* behavior. Ranges for several variables can be set simultaneously, resulting in a combined filter selection. The results displayed "on the fly" upon any change, allowing the user to explore the behavior of the whole system under changes in ranges of any variable/s.

Basic usage

To use it, drag the mouse on any of the lower axes. You can modify the extent of the range/s by dragging the limits (left and right). The user can also preserve the extent and move its position by dragging the mouse on the center of the current range.

Input textbox

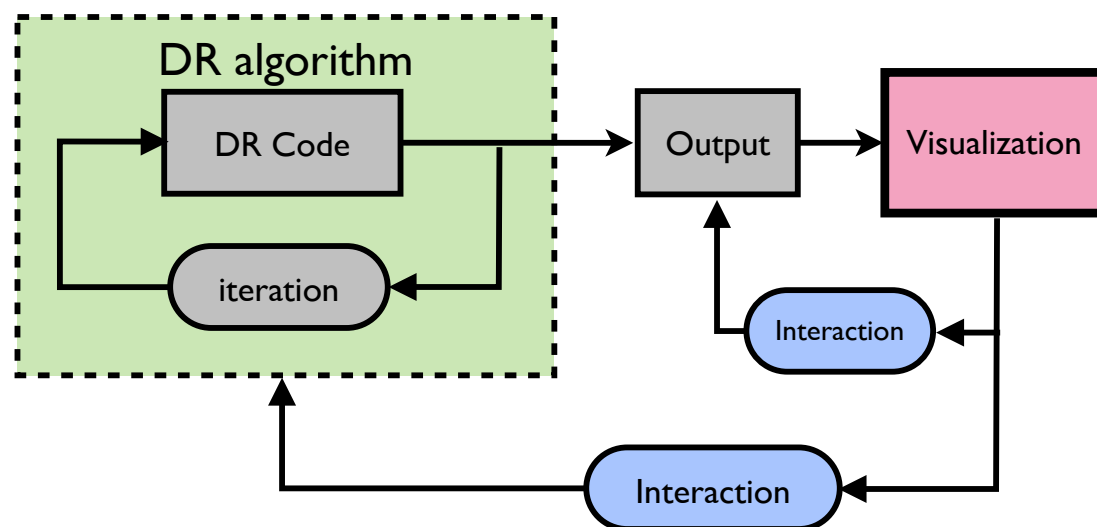
The input textbox allows the user to modify the behavior of the barcharts, by setting different types of aggregations, such as "avg" or "sum", or a simple "count" to setup a histogram. The user can also set the measures (variables being aggregated or represented)

hola

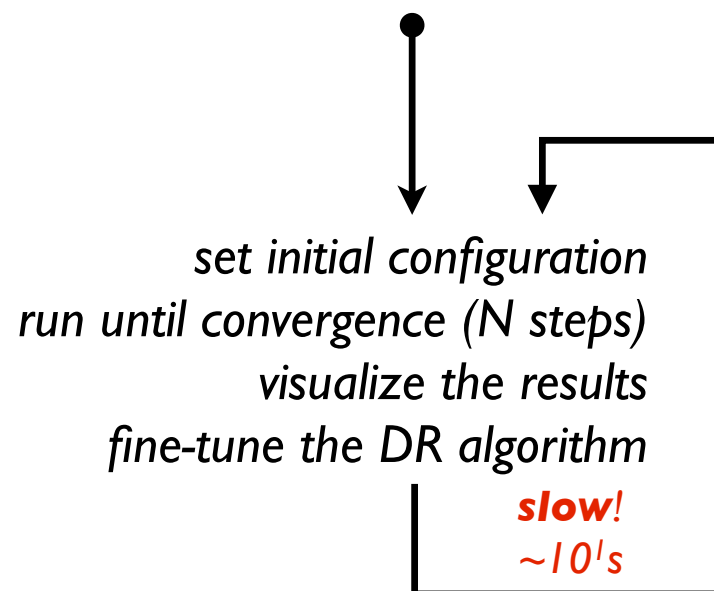
syntax: "dimension,aggregation,measure" (examples: "week,count,samples", "hour,avg,cosPhi", "Edificio,sum,activePower").

Reducción de la dimensionalidad interactiva (iDR)

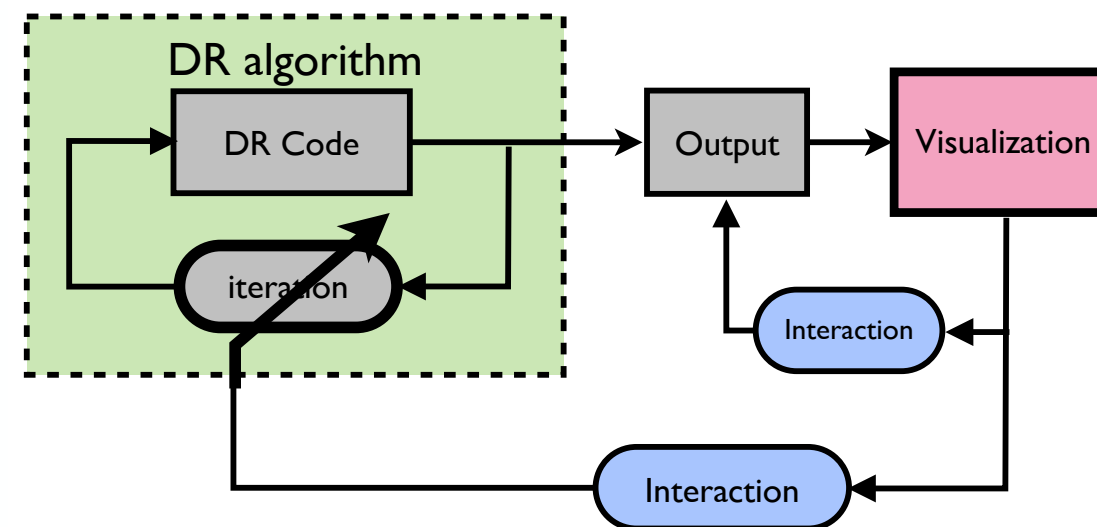
DR básica



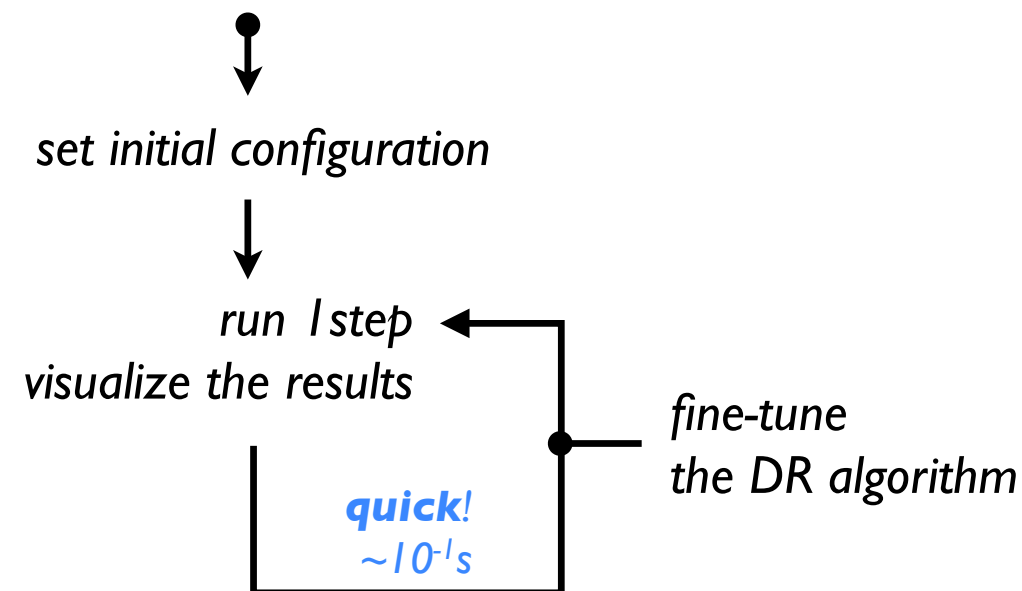
batch mode interaction scheme



DR interactiva



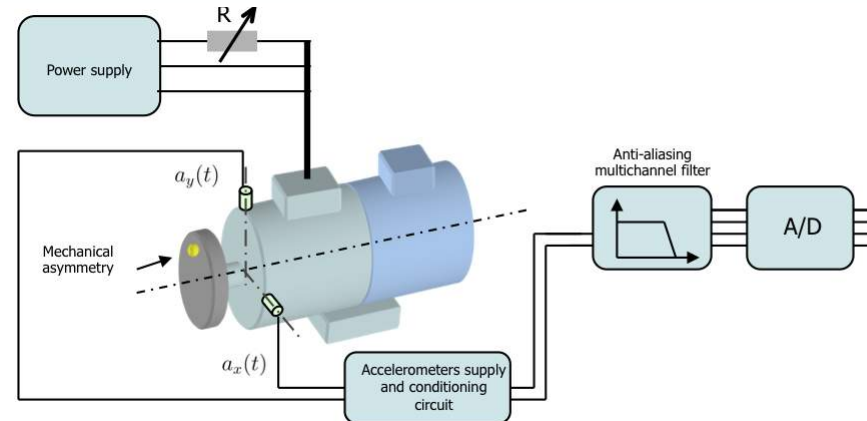
iDR interaction scheme



enables a quick feedback to the user and hence a much better user integration in the exploration process

Reducción de la dimensión interactiva

análisis de fallos en motor AC

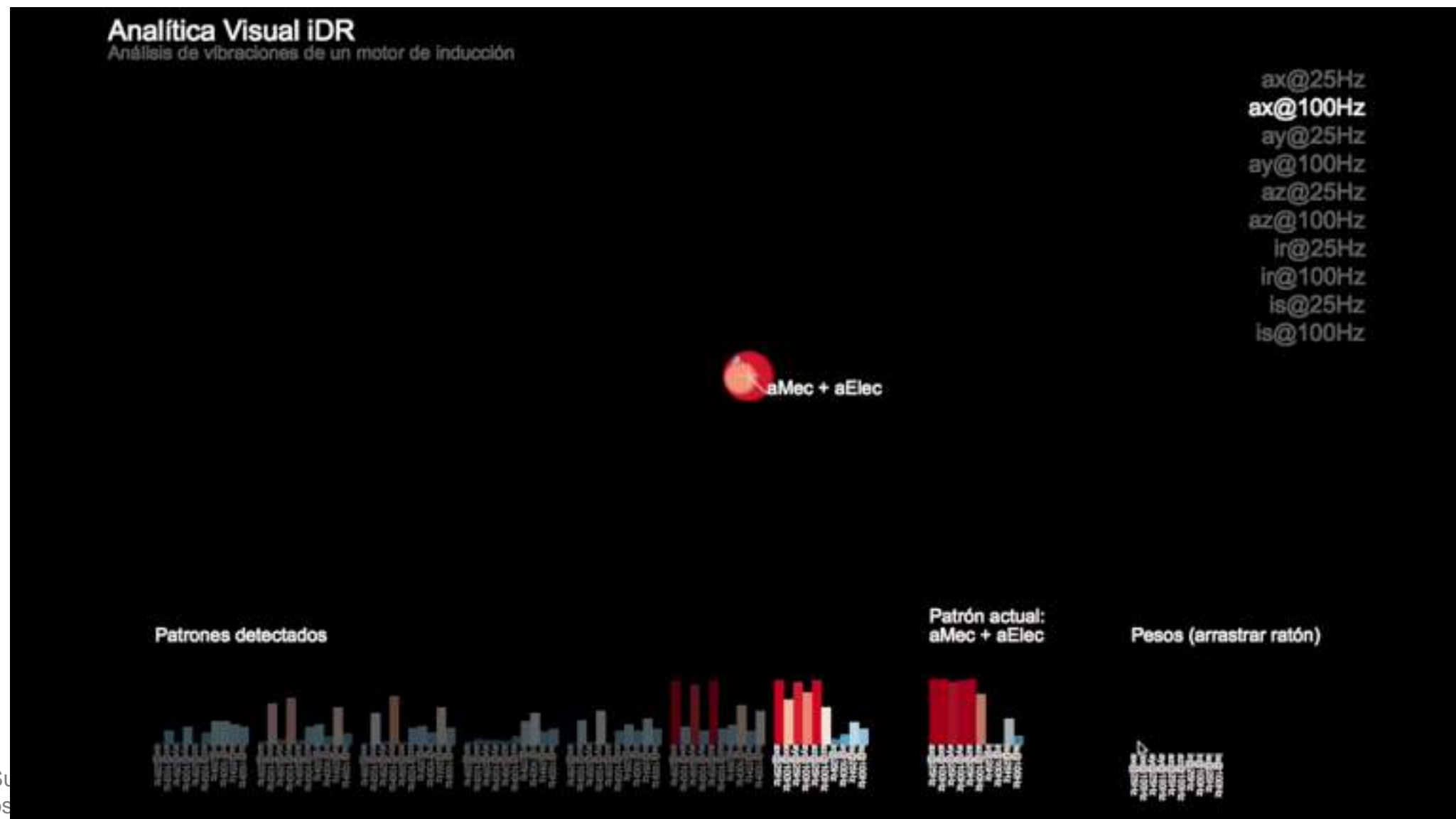


Application interface with iDR user-driven modification of the input metric space

Javascript application using processingjs (<http://processingjs.org>)

Analysis of three vibration signals $a_x(t)$ $a_y(t)$ $a_z(t)$ and two phase currents $i_R(t)$, $i_S(t)$

of a 4kW 2 pole-pair asynchronous motor <http://isa.uniovi.es/~idiaz/demos/iDR-vibracionesMotor/>



- visualización
análisis inteligente de datos
interacción
→ tienen ventajas complementarias
- La integración de los tres paradigmas
tiene un gran potencial
 - técnicas sencillas “brillan”... parecen otras
cuando se le añade interacción y visualización
 - importante nicho de innovación aprovechando tecnologías web
- Soluciones especialmente interesantes...
 - en problemas multivariable (ej. factores estudiantes)
 - cuando son importantes el análisis cualitativo, la intuición y la interpretación

Gracias a la colaboración estrecha y feedback continuo del **grupo SUPPRESS de la Universidad de León** con quienes constantemente estamos discutiendo nuevas ideas y aplicaciones

Financiación del MINECO

Gran parte de las ideas y material presentados se han desarrollado en el marco del proyecto DPI2015-69891-C2-2-R, financiada por el Ministerio de Economía y Competitividad (MINECO), Programa Estatal de I+D+i Orientada a los Retos de la Sociedad y por el Fondo Europeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa"



Unión Europea

Fondo Europeo
de Desarrollo Regional
"Una manera de hacer Europa"