Big Data
Desde los datos al conocimiento

SOCIEDAD ECUATORIANA DE ESTADISTICA

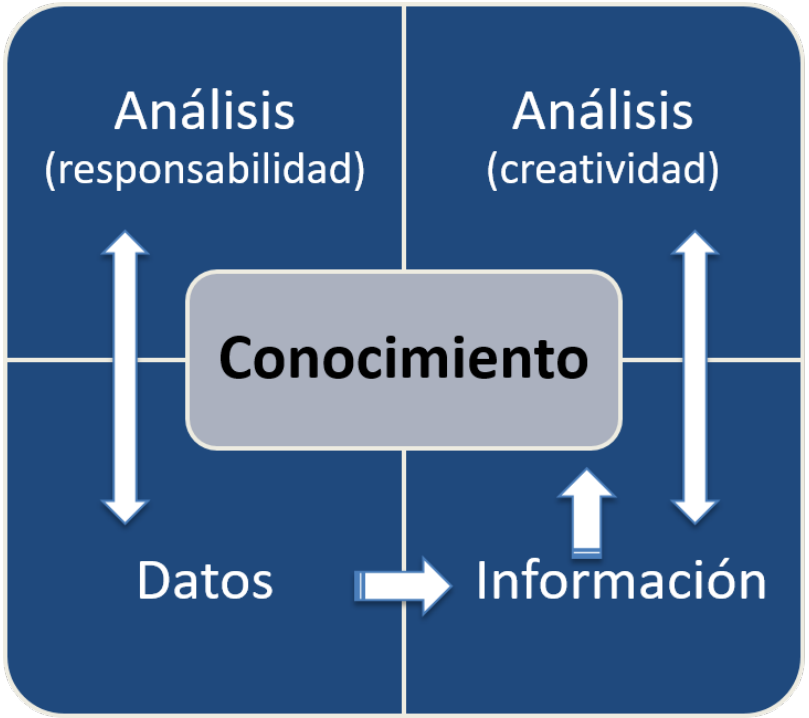
Lilia Quituisaca-Samaniego[§]

lilia.quituisaca.samaniego@gmail.com

info@liliaquituisacasamaniego.com

[§]**Sociedad Ecuatoriana de Estadística, Quito, Ecuador**

Introducción

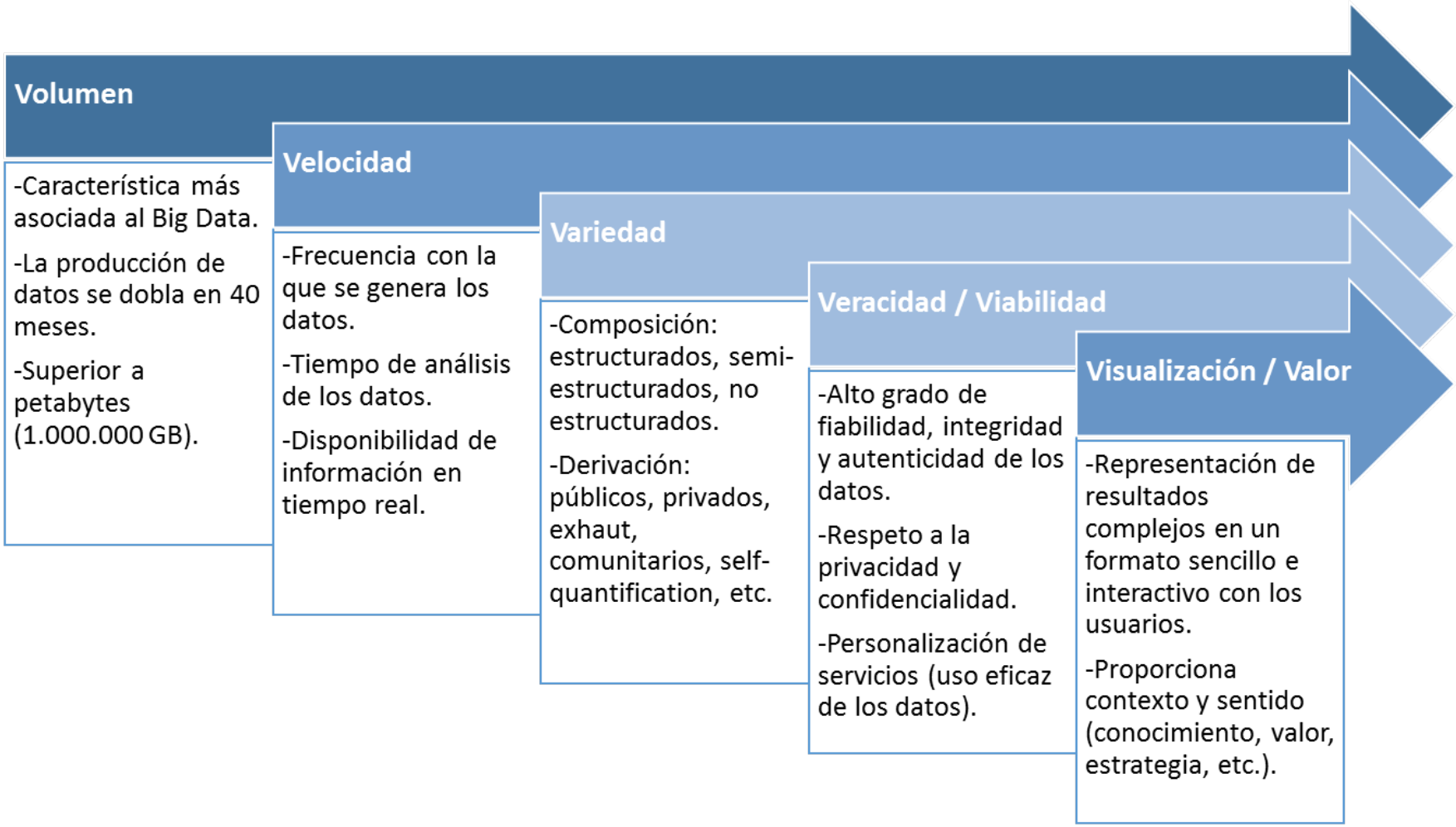


Es importante:

disponer de datos y conocer su relación entre variables; sin embargo, su valor está en transformarlos en conocimiento.

- "Transformar y traducir los datos en respuestas" es parte de la innovación, y ésta debe ser aplicada con responsabilidad y creatividad [4].
- Uno de los siete pasos en la visualización de datos es la extracción y transformación, ya que permite "aplicar métodos de análisis para distinguir patrones"; es aquí donde recobra aún más importancia el contexto matemático[3].
- Algunos métodos nos permitirán evaluar su complejidad mediante Inteligencia Artificial (IA), minería de datos o específicamente la aplicación de técnicas de aprendizaje automático (machine learning).
- Luego se realiza comparaciones y obtención de conclusiones previas con la extensión de los datos estadísticos mediante "índices" representativos.

Conceptos relacionados


Volumen
-Característica más asociada al Big Data.
-La producción de datos se dobla en 40 meses.
-Superior a petabytes (1.000.000 GB).
Velocidad
-Frecuencia con la que se genera los datos.
-Tiempo de análisis de los datos.
-Disponibilidad de información en tiempo real.
Variedad
-Composición: estructurados, semi-estructurados, no estructurados.
-Derivación: públicos, privados, exhaut, comunitarios, self-quantification, etc.
Veracidad / Viabilidad
-Alto grado de fiabilidad, integridad y autenticidad de los datos.
-Respeto a la privacidad y confidencialidad.
-Personalización de servicios (uso eficaz de los datos).
Visualización / Valor
-Representación de resultados complejos en un formato sencillo e interactivo con los usuarios.
-Proporciona contexto y sentido (conocimiento, valor, estrategia, etc.).

Objetivo:

Devolver a la sociedad sus datos públicos y fomentar la transparencia y la reutilización.

Características:

- Formatos digitales, estandarizados y abiertos.
- Estructura clara que permita la comprensión.
- Fácil acceso y permanencia en repositorios.

Principales beneficios:

- Empoderar al ciudadano de la información para que pueda ser interpretada o reinterpretada.
- Fomentar un mayor grado de transparencia y reutilización.

Inquietudes

- Si los datos son generados en un espacio público ¿por qué estos datos deben mantenerse en privado?
- ¿A quién/es le pertenece los datos generados: a las administraciones, al ciudadano, al mundo?
- ¿Qué pasa con la evolución de la privacidad?
- ¿Dónde están los límites de la información sobre nosotros, que nosotros no generamos?
- ¿Qué competencias se deben desarrollar?

En base a la tendencia actual tenemos algunos perfiles [2]:

- **Chief Data Officer (Oficial Jefe de Datos)**, lidera la gestión de datos y analítica asociada por el negocio (responsable de los diferentes equipos especialidades en datos).

- **Data Scientists (Científico de Datos)**, extraen conocimiento e información valiosa de los datos.

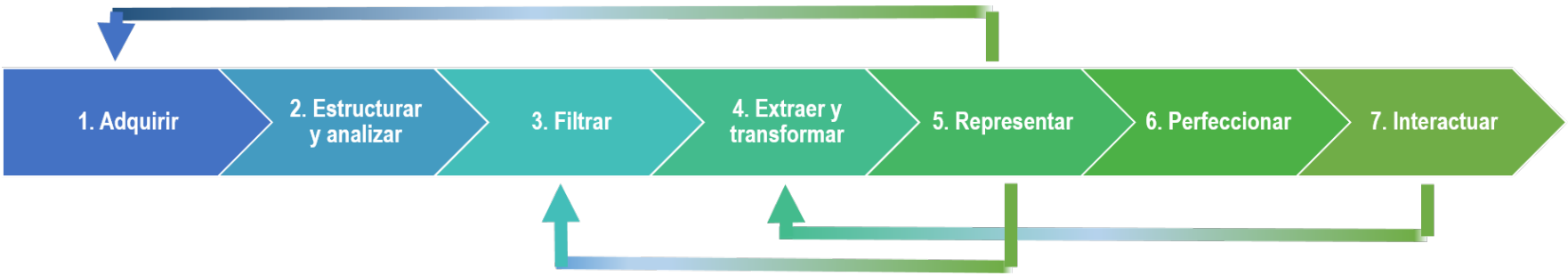
- **Citizen Data Scientist (Ciudadano Científico de Datos)**, no está formado específicamente para ser Data Scientist, pero puede extraer valor, a través de su experiencia, explorando los datos, desde las unidades de negocio.

- **Data Engineer (Ingeniero de datos)**, proporciona datos de una manera accesible y apropiada a los usuarios y Data Scientists.
- **Data Steward (Administrador de datos)**, mantiene la calidad, disponibilidad y seguridad de los datos.
- **Business Data Analyst (Analista de datos comerciales)**, recoge las necesidades de los usuarios de negocio para los Data Scientist y presenta resultados obtenidos.
- **Data Artist (Profesional Creativo)**, expertos en Business Analytics, crean gráficos, infografías y otras herramientas visuales.

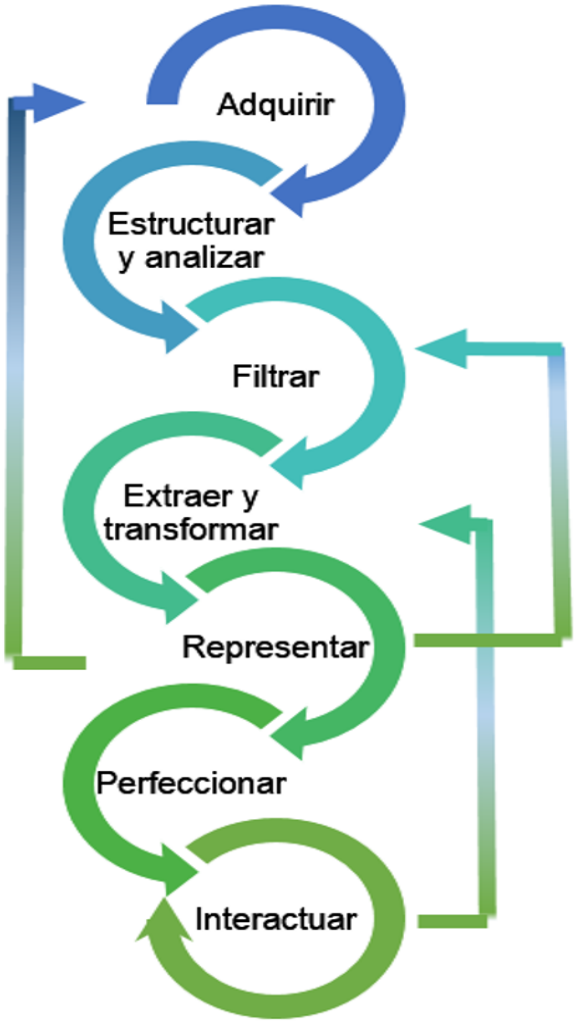
Visualización de datos [1]

Debemos saber que:

- La información reside en las relaciones, no en los datos.
- Visualizar es representar gráficamente esas relaciones aprovechando nuestra enorme capacidad de analítica visual.



La visualización de datos no es hacer visible la información ante nuestros ojos, sino ante nuestro entendimiento.



1 2 3 4 5 6 7

Obtención de datos, mediante fuentes directas desde un archivo en un disco o de una fuente a través de una red.

Preparación de alguna estructura para el significado de los datos y ordenación por categorías.

Eliminación de todos los datos excepto los de interés, es decir, permanecen los datos seleccionados.

Aplicación de métodos de análisis, para distinguir patrones o colocar los datos en el contexto matemático.

Elección de un modelo visual básico de representación, como un gráfico de barras, una tabla, un árbol, etc.

Optimización de la representación básica para hacerla más clara y visualmente más atractiva.

Adición de métodos para manipular los datos o controlar las características que serán visibles.

Variables derivadas

Nombre	Edad	Nivel de estudios	Ciudad	Latitud	Longitud	Salario
Pedro	34	Graduado escolar	Madrid	40.41	-3.69	21000
María	45	Licenciatura	Madrid	40.41	-3.69	27000
Jaime	40	Diplomatura	Barcelona	41.38	2.17	30000
Mercedes	45	Licenciatura	Madrid	40.41	-3.69	28000
Vanesa	36	Graduado escolar	Barcelona	41.38	2.17	19000
Carmen	28	Licenciatura	Málaga	36.71	-4.42	19500
Javier	25	Licenciatura	Madrid	40.41	-3.69	19000
Pablo	30	Graduado escolar	Málaga	36.71	-4.42	22000

Agregación de registros

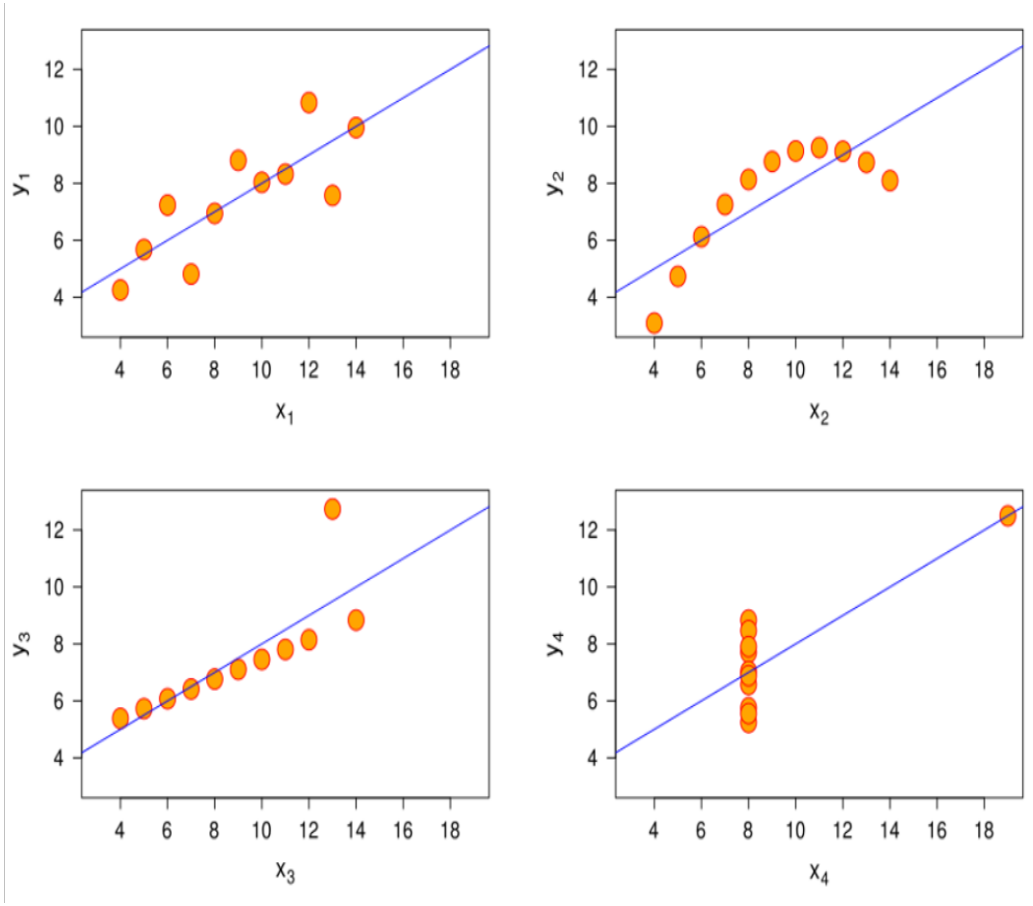
Nivel de estudios	Frecuencia	Salario (sumatorio)
Graduado escolar	3	62000
Licenciatura	4	93500
Diplomatura	1	30000

Agregación de registros + Cálculo de métricas

Nivel de estudios	Frecuencia	Salario medio
Graduado escolar	3	20666
Licenciatura	4	23375
Diplomatura	1	30000

Cuarteto de Anscombe [5]

Propiedad	Valor
Media de X	9,0
Varianza de X	11,0
Media de Y	7,5
Varianza de Y	4,12
Correlación entre X e Y	0,812
Recta de regresión	y=3+0,5x



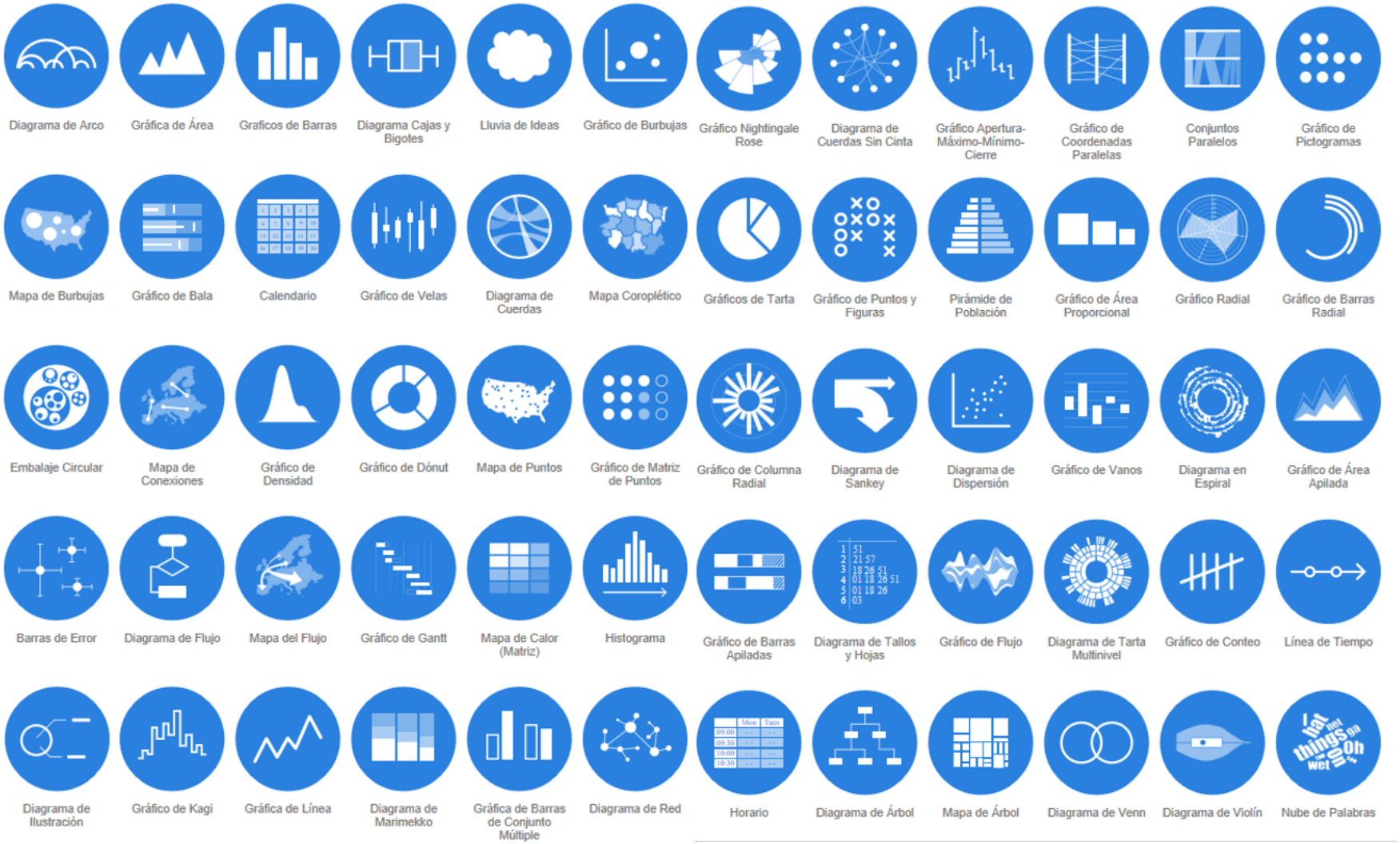
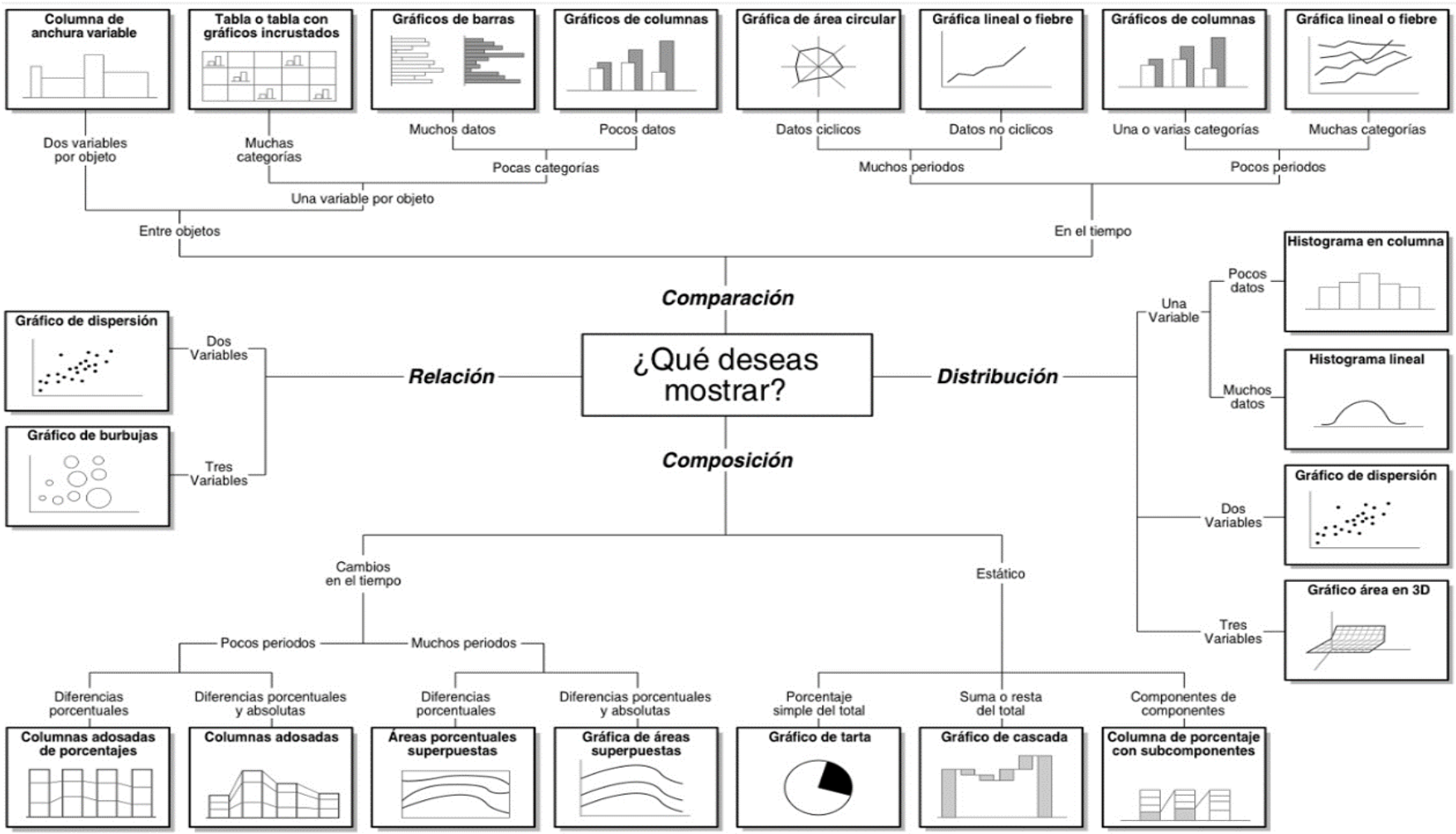
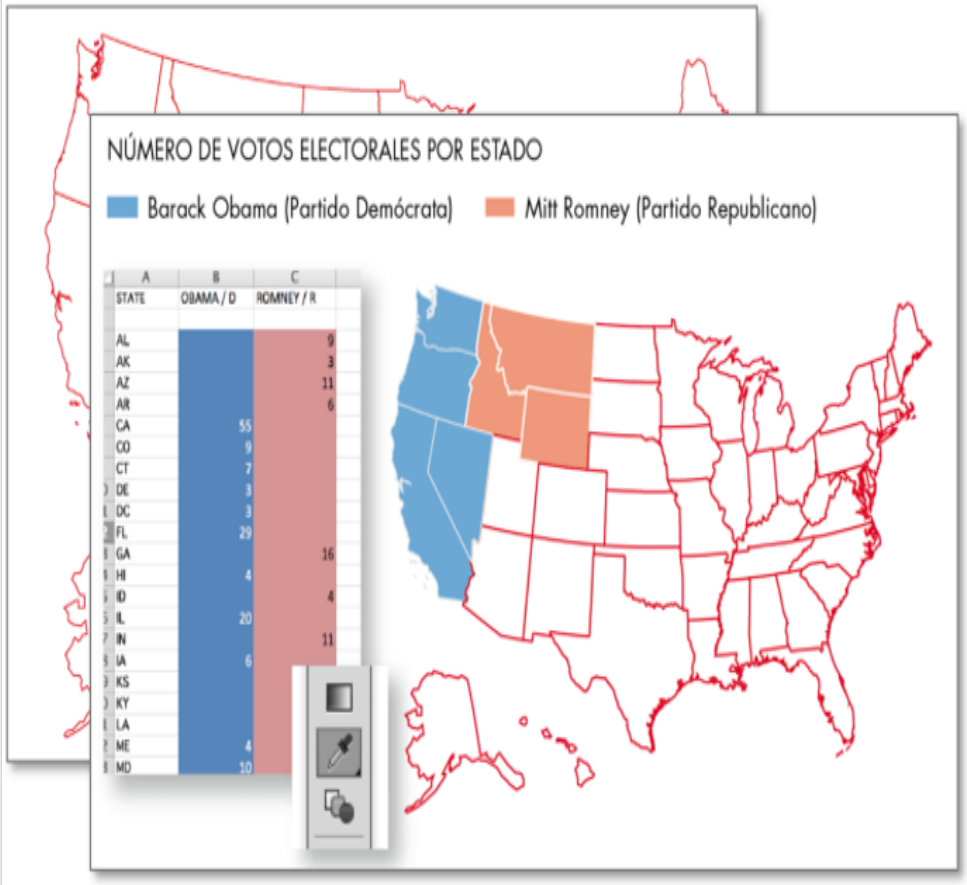


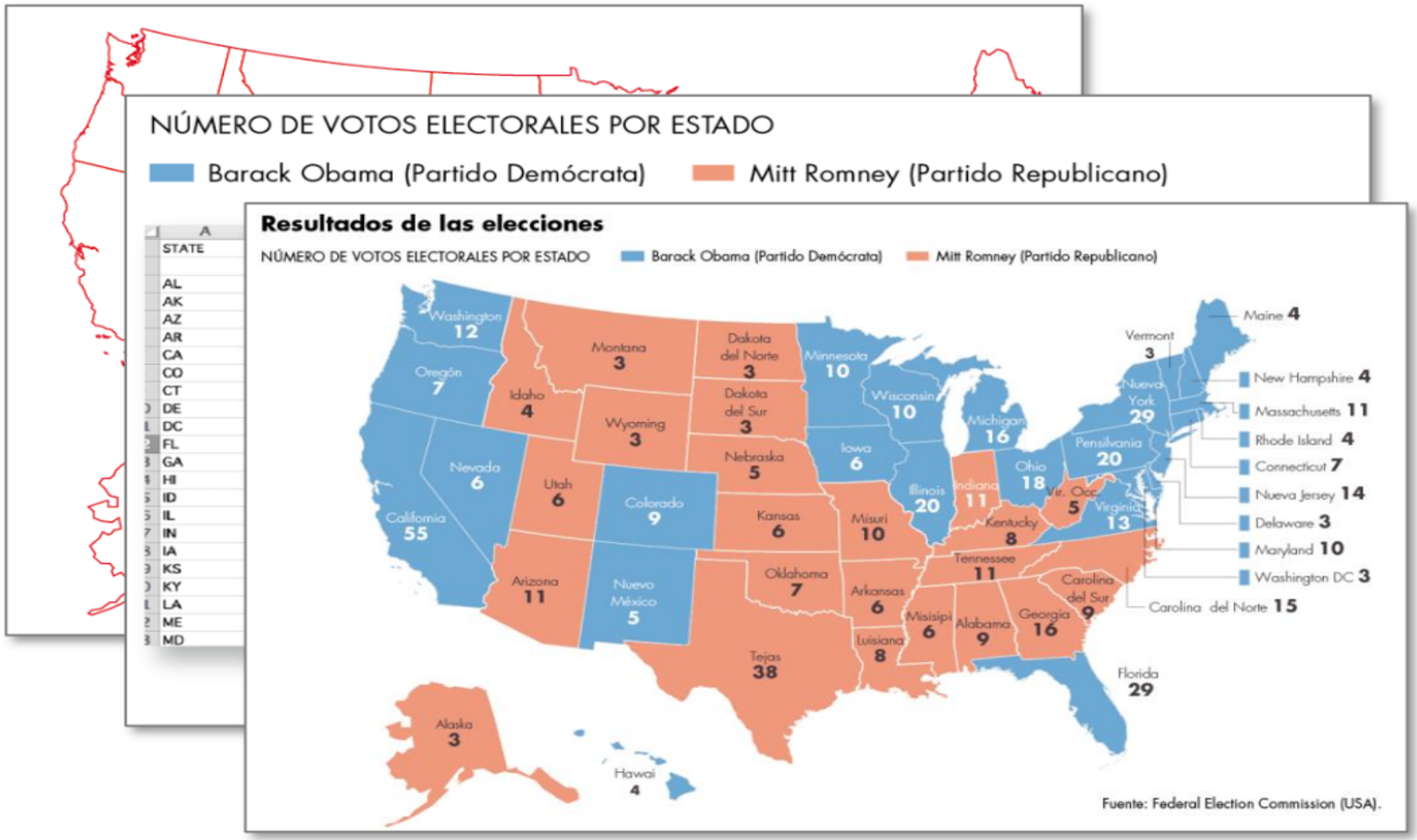
Diagrama de Arco
Gráfica de Área
Graficos de Barras
Diagrama Cajas y Bigotes
Lluvia de Ideas
Gráfico de Burbujas
Gráfico Nightingale Rose
Diagrama de Cuerdas Sin Cinta
Gráfico Apertura-Máximo-Mínimo-Cierre
Gráfico de Coordenadas Paralelas
Conjuntos Paralelos
Gráfico de Pictogramas
Mapa de Burbujas
Gráfico de Bala
Calendario
Gráfico de Velas
Diagrama de Cuerdas
Mapa Coroplético
Gráficos de Tarta
Gráfico de Puntos y Figuras
Pirámide de Población
Gráfico de Área Proporcional
Gráfico Radial
Gráfico de Barras Radial
Embalaje Circular
Mapa de Conexiones
Gráfico de Densidad
Gráfico de Dónut
Mapa de Puntos
Gráfico de Matriz de Puntos
Gráfico de Columna Radial
Diagrama de Sankey
Diagrama de Dispersión
Gráfico de Vanos
Diagrama en Espiral
Gráfico de Área Apilada
Barras de Error
Diagrama de Flujo
Mapa del Flujo
Gráfico de Gantt
Mapa de Calor (Matriz)
Histograma
Gráfico de Barras Apiladas
Diagrama de Tallos y Hojas
Gráfico de Flujo
Diagrama de Tarta Multinivel
Gráfico de Conteo
Línea de Tiempo
Diagrama de Ilustración
Gráfico de Kagi
Gráfica de Línea
Diagrama de Marimekko
Gráfica de Barras de Conjunto Múltiple
Diagrama de Red
Horario
Diagrama de Árbol
Mapa de Árbol
Diagrama de Venn
Diagrama de Violín
Nube de Palabras



Traducción a cargo de Víctor Caballero con autorización del autor
www.Vectart.com / contact@vectart.com

www.ExtremePresentation.com
© 2009 A. Abela - a.v.abela@gmail.com


NÚMERO DE VOTOS ELECTORALES POR ESTADO
Barack Obama (Partido Demócrata)
Mitt Romney (Partido Republicano)
STATE
OBAMA / D
ROMNEY / R


NÚMERO DE VOTOS ELECTORALES POR ESTADO
Barack Obama (Partido Demócrata)
Mitt Romney (Partido Republicano)
A
STATE
AL
AK
AZ
AR
CA
CO
CT
DE
DC
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
Resultados de las elecciones
NÚMERO DE VOTOS ELECTORALES POR ESTADO
Barack Obama (Partido Demócrata)
Mitt Romney (Partido Republicano)
Washington 12
Oregón 7
California 55
Idaho 4
Nevada 6
Montana 3
Wyoming 3
Utah 6
Arizona 11
Colorado 9
Nuevo México 5
Dakota del Norte 3
Dakota del Sur 3
Nebraska 5
Kansas 6
Oklahoma 7
Tejas 38
Minnesota 10
Iowa 6
Missouri 10
Arkansas 6
Luisiana 8
Wisconsin 10
Illinois 20
Michigan 16
Indiana 11
Ohio 18
Kentucky 8
Tennessee 11
Misisipi 6
Alabama 9
Georgia 16
Florida 29
Carolina del Sur 9
Carolina del Norte 15
Virginia 13
Vir. Occ. 5
Pensilvania 20
Nueva York 29
Vermont 3
Maine 4
New Hampshire 4
Massachusetts 11
Rhode Island 4
Connecticut 7
Nueva Jersey 14
Delaware 3
Maryland 10
Washington DC 3
Alaska 3
Hawai 4
Fuente: Federal Election Commission (USA).

Conclusiones

- La clave del éxito del Big Data no son los datos en sí, sino los modelos de interpretación (extracción del conocimiento).
- "Todas las opciones por defecto de Excel son erróneas: los ejes, el hecho de que todo se represente mediante efectos 3D, los colores por defecto, las etiquetas, los estilos gráficos por defecto..." (Noah Iliinsky)
- "Los modelos mentales a menudo están construidos sobre evidencias incompletas, sobre un escaso conocimiento acerca de lo que está ocurriendo, y con un tipo de psicología ingenua que postula causas, mecanismos y relaciones, incluso cuando no existen." (Don Norman)

Referencias

A. Cairo, *El arte funcional*, Fareso S.A., 2011.

L. Quituisaca-Samaniego, *Big data: visión general*, Numérica Resumiendo, (2017).

——, *Visualización de datos: siete pasos*, Numérica Resumiendo, (2017).

The Software Alliance, *Por qué son tan importantes los datos?*, 2015.

E. Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press, 2001.

Gracias

lilia.quituisaca.samaniego@gmail.com

info@liliaquituisacasamaniego.com

SOCIEDAD ECUATORIANA DE ESTADISTICA