SG VIRTUAL CONFERENCE
7ma edición

Data Mining: Torturando los datos hasta que confiesen

Presentado por:
Luis Carlos Molina

Curriculum

- Desde 1996 se ha dedicado de manera ininterrumpida a temas de análisis de información, en especial de minería de datos (*data mining*). Sus estudios de maestría los realizó en la Universidad de São Paulo, Brasil y de doctorado en la Universidad Politécnica de Cataluña, España. Ha sido investigador huésped en la Universidad Atrás dos Montes, Portugal. Cuenta con más de 20 publicaciones científicas y más de 80 conferencias sobre temas analíticos, limpieza de datos, minería de datos, visualización de datos y big data.
- Dentro de sus publicaciones importantes destacan:
  - Libro: **Data Mining: Una Introducción.** Luis Carlos Molina y Ramón Sanguesa. UOC. Barcelona, España. 2001.
  - **Data Mining: Torturando los datos hasta que confiesen.** UOC. Barcelona, España. 2002.
  - **Representing a relation between porosity and permeability based on inductive rules.** Luis Carlos Molina and Luis Belanche, Journal of Petroleum, Volume 47, Issue 1-2, May 2005, Pages 23-34. Top 25 Hottest Articles (3rd Place).
  - **Del Data Mining al Big Data**. Luis Carlos Molina. Power Builders. México. 2013.
- En el ámbito laboral se ha desempeñado como consultor del sector bancario, retail, gobierno, educación, energía y telefonía celular en México, España, Portugal, Brasil y Colombia.
- Entre sus trabajos relevantes se destacan:
  - Implementación de un **modelo analítico para telefonía celular** que fue presentado en el Congreso Mundial de Telefonía Celular en Singapur en el 2012.
  - Desarrollo de un **modelo para mitigar el fraude con tarjeta de crédito** que se hizo un caso de éxito en una entidad bancaria mexicana.
  - Responsable del primer proyecto analítico en una empresa mundial de **retail**.
  - Diseño de una metodología exitosa para **limpieza de grandes volúmenes de información** probada en varios proyectos del gobierno mexicano.

Contacto:

Luis Carlos Molina Félix

luiscarlos.molina@powerbuilders.com.mx

Cel. +(521) 5523008882

¿Cual es la problemática?

- El volumen de datos es enorme:
  - Problemas de dimensionalidad (100-10,000 atributos)
  - Número de observaciones (Varios Servidores)
- Análisis de datos es crucial para tomar decisiones rápidas de negocio.
- Las empresas necesitan conocer mejor a sus clientes.
- Dificultad para aplicar técnicas tradicionales.
- Solamente entre un 5% a 10% de la información es analizada (Gartner Group).

"Knowledge is the only competitive Advantage"
Jack Welch, ex-CEO, General Electric

Una primera aproximación

- Ejemplo: Pañales y Cerveza.
- Con técnicas de MD se encontró que un grupo de clientes compraban pañales junto con cerveza después de las 7 de la noche en días laborables.
- El perfil del consumidor eran hombres casados entre 25 y 35 años.
- Wal-Mart optó por una adecuación de los estantes en sus puntos de venta colocando los pañales al lado de las cervezas.
- Resultado: El consumo de cerveza creció 30% con ese cambio. Colocaron papas fritas en medio y las ventas de los 3 productos se incrementaron.

Ejemplo en México

- En las tiendas con formato de clubes de precios.
- En los productos que contengan la palabra “orgánico”, por ejemplo:
  - Tomate orgánico
  - Lechuga orgánica
  - Huevo orgánico
- El producto que se llevan junto con esto es:
  - Brownie de chocolate

Minería de Datos

Conjunto de áreas que tienen como propósito la identificación de conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacía la toma de decisión. [Molina 2000]

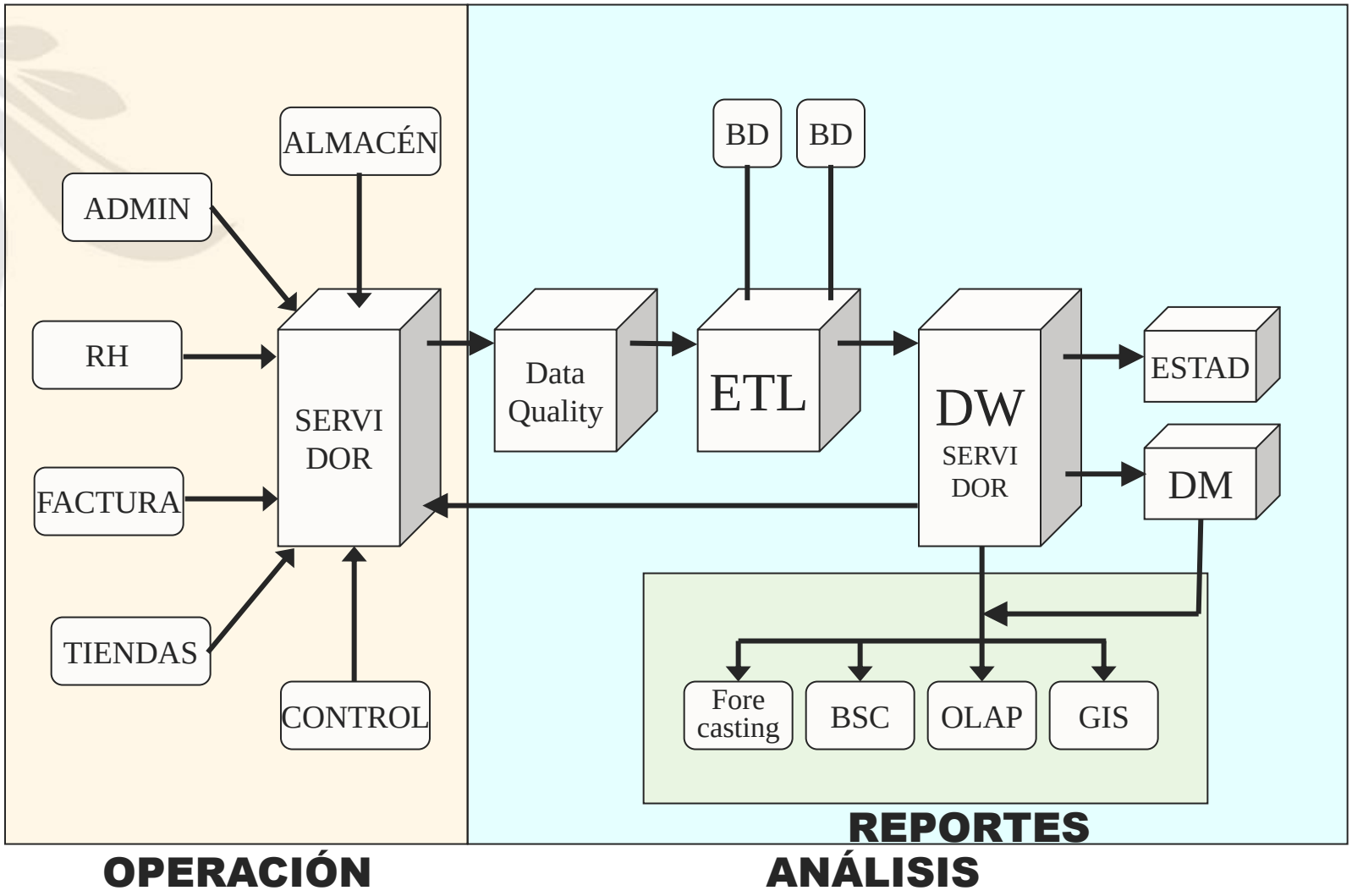
¿Que áreas?

- **Estadística**
- **Inteligencia Artificial**
- **RP**
- **Computación Gráfica**
- **Bases de Datos**

Donde trabaja

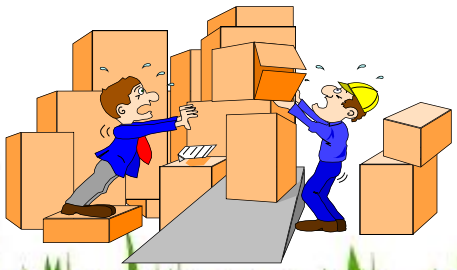
•[Molina 1998

Componentes de un Ambiente Analítico

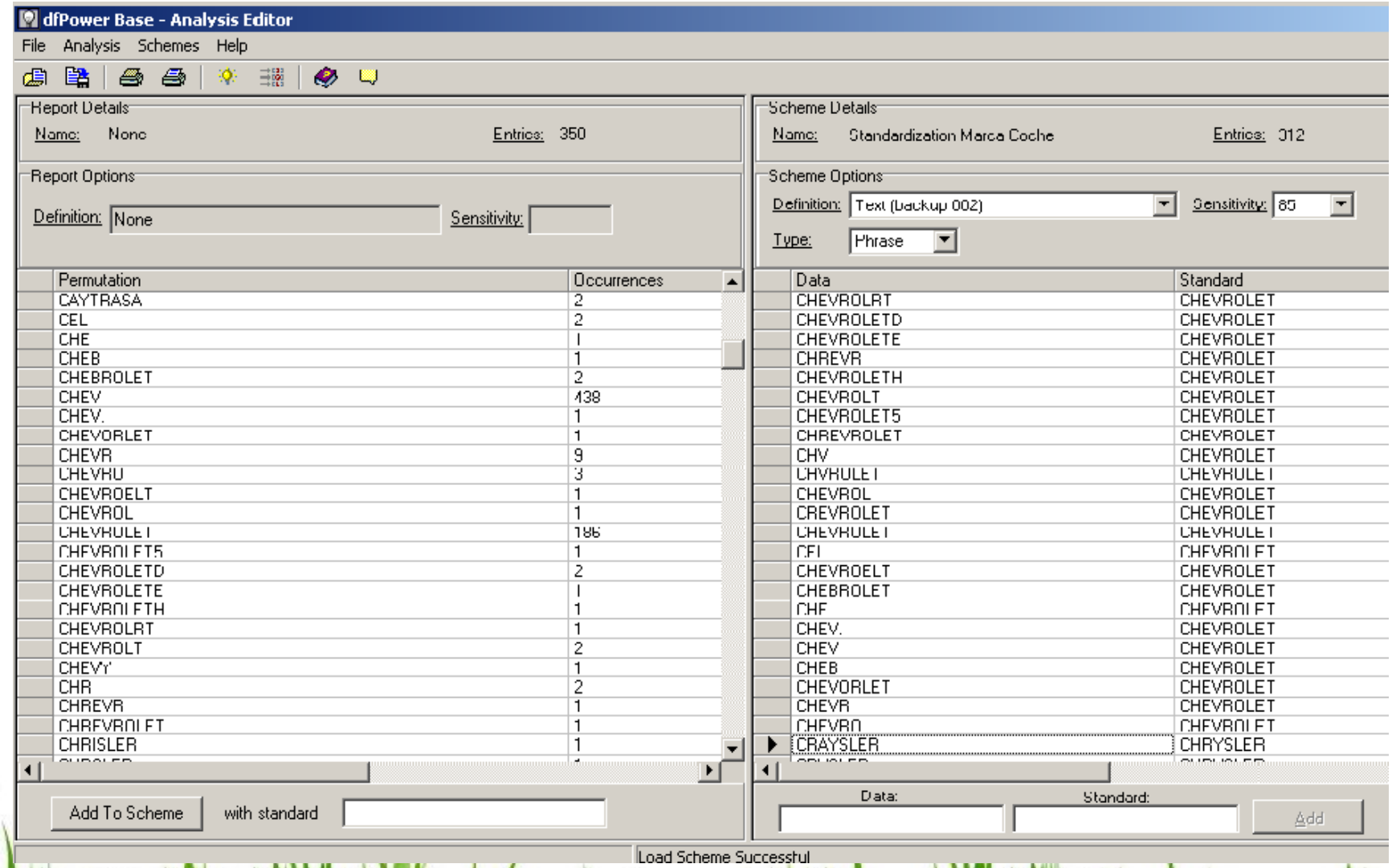


¿Qué conforma a MD?

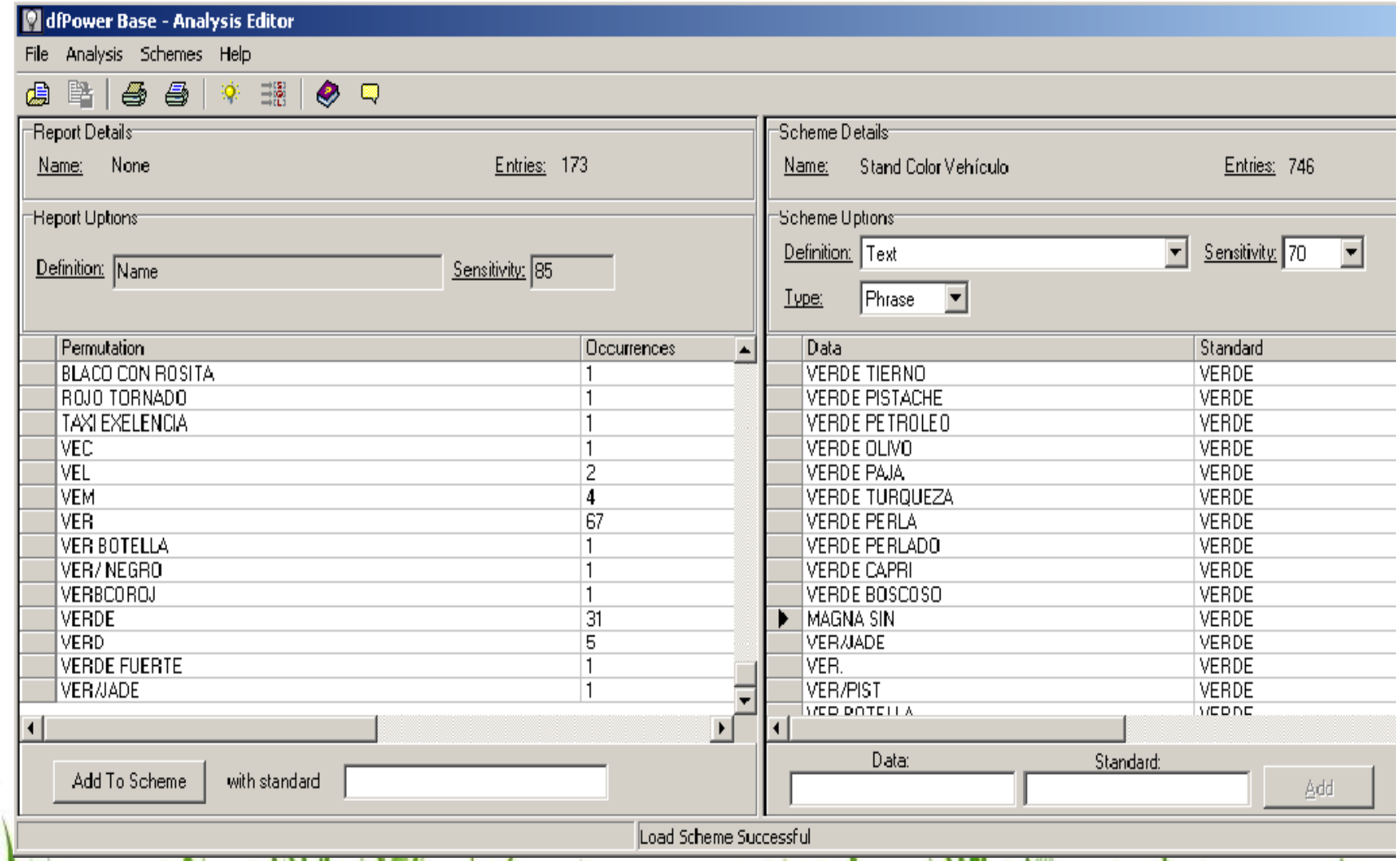
- Una gran base de datos.
- Un especialista de dominio.
  - Unos objetivos.
- Un minero de datos.
  - Un software.
  - Un especialista.
- Una metodología.
- Herramientas de MD.
  - Técnicas de MD.
  - Técnicas de validación.



Homogenización de Datos



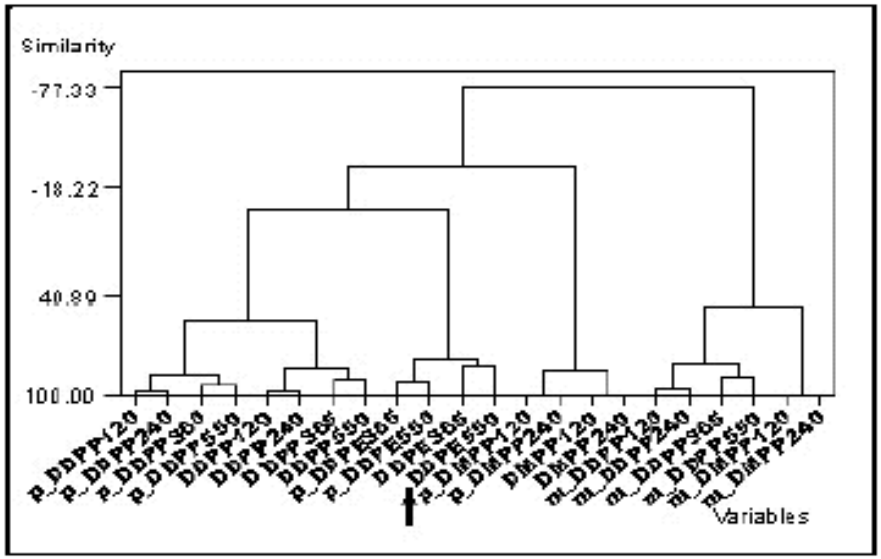
Homogenización de Datos



Estadística

- Rango
- Media
- Moda
- Mediana
- Varianza
- Correlación
- Dendogramas
- Histogramas
- Desviación Estándar

$$\overline{X} = \frac{1}{n}\sum_{i-1}^{n} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i-1}^{N} (x_i - \mu)^2 . \qquad s^2 = \frac{1}{(n-1)}\sum_{i-1}^{n} (x_i - \overline{X})^2 .$$

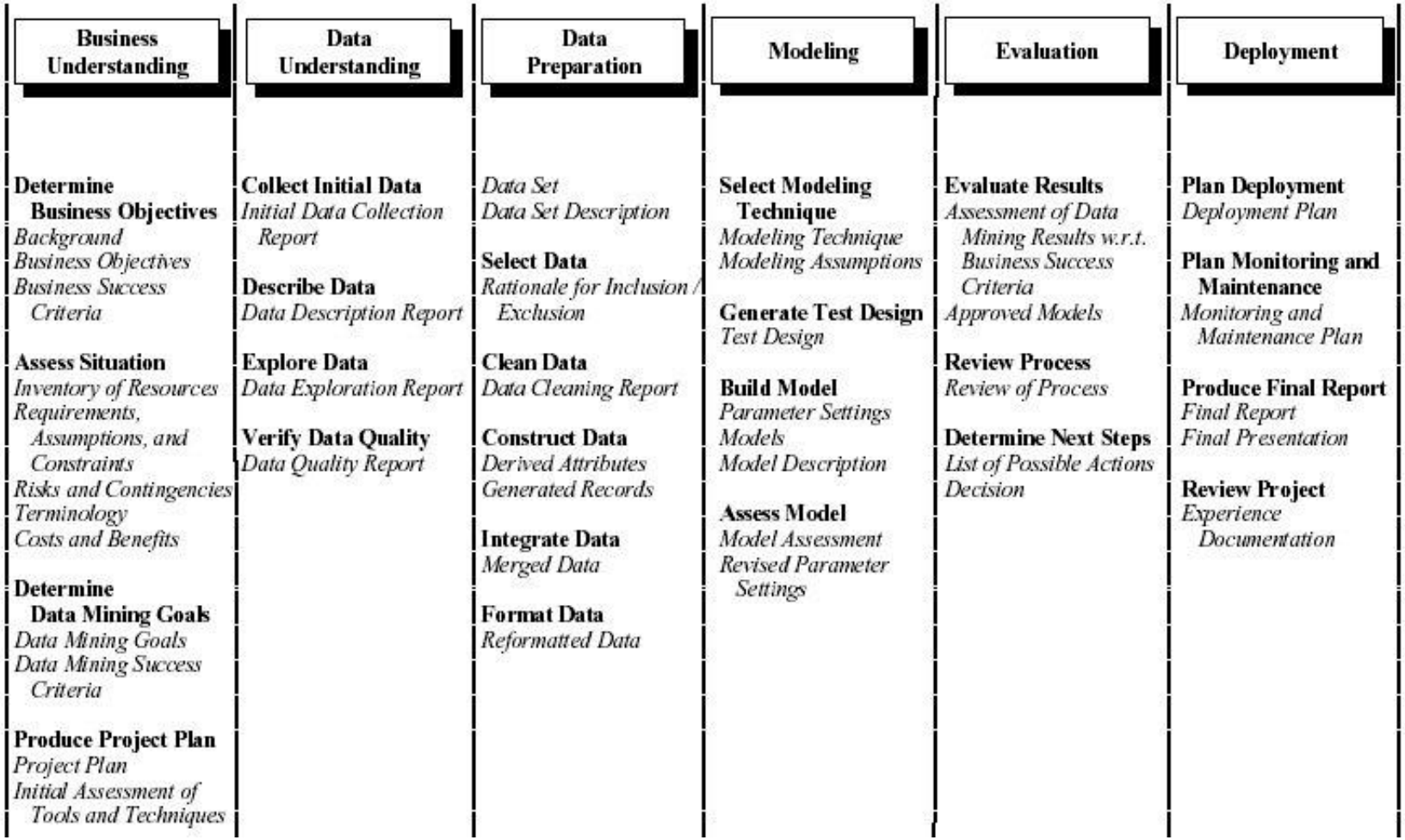


$$c_{ij} = \frac{1}{n}\sum_{k=1}^{n} [(x_{ki} \quad \overline{x}_i) \cdot (x_{kj} \quad \overline{x}_j)].$$

Metodología CRISP-DM

- Desarrollada por compañías que trabajan en Data Mining (SPSS, NCR, OHRA, ChryslerDaimler)
- Fundada por la Comisión Europea
- Herramienta-independiente / industria-independiente
- Modelo por proceso jerárquico
  - De lo general a lo particular

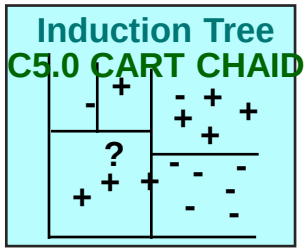
Metodología CRISP-DM

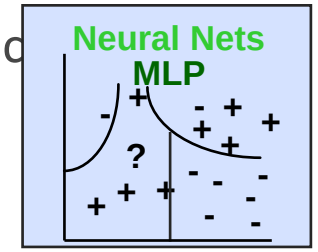


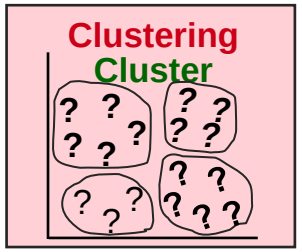
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Técnicas usadas en MD

- Agrupación
  - Agrupación de objetos similares
- Clasificación y Regresión
  - Agrupación de objetos similares considerando una estruc
    de clases conocidas
- Modelos Predictivos
  - Identificar las variables más predictivas
  - Anticiparse a los eventos
- Descubrimiento de secuencias
  - Agrupa un tipo especial de objetos: secuencias
- Asociación
  - Encuentra relaciones entre productos







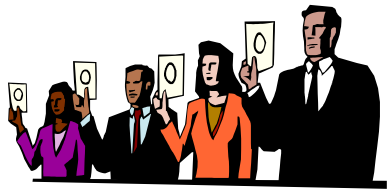
Los 10 algoritmos más usados (ICDM 06)

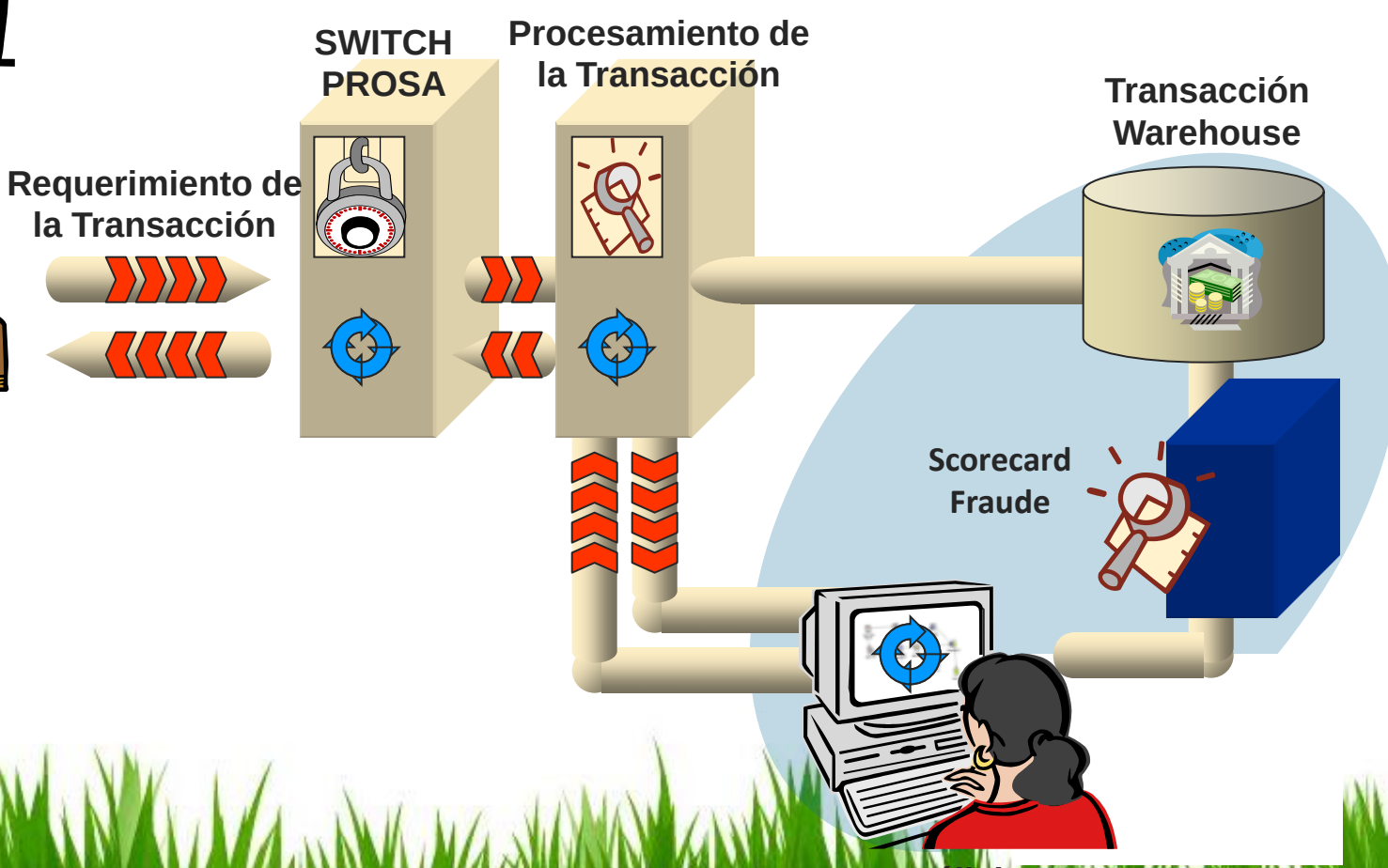
- #1: C4.5 Decision Tree - Classification (61 votes)
- 2: K-Means - Clustering (60 votes)
- 3: SVM – Classification (58 votes)
- 4: Apriori - Frequent Itemsets (52 votes)
- 5: EM – Clustering (48 votes)
- 6: PageRank – Link mining (46 votes)
- 7: AdaBoost – Boosting (45 votes)
- 7: kNN – Classification (45 votes)
- 7: Naive Bayes – Classification (45 votes)
- 10: CART – Classification (34 votes)

Ejemplos

Caso de Estudio:
Detección de Fraudes en Tarjeta de Crédito

Antecedentes

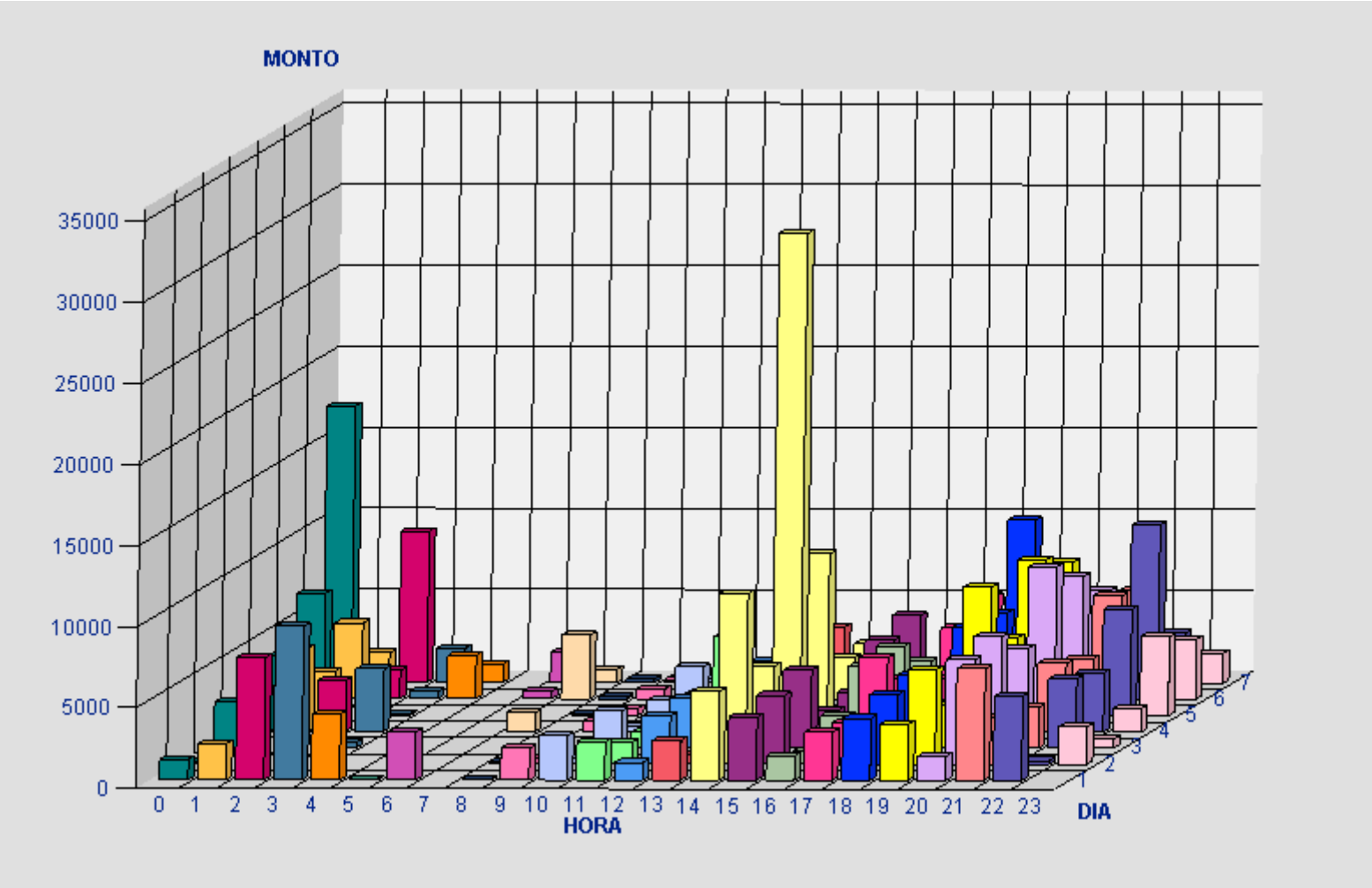




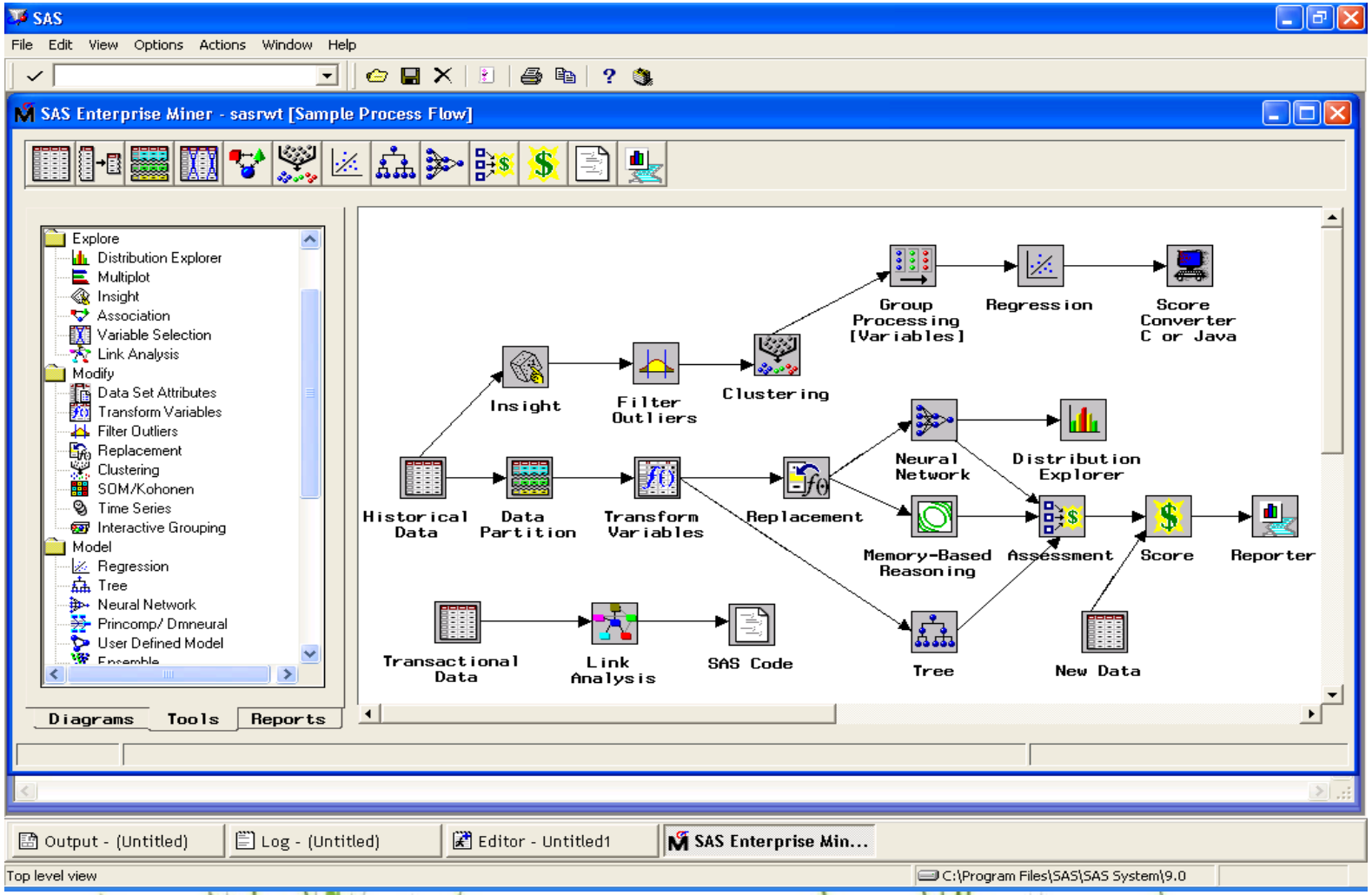
Estructura de la BD

- FOLIO
- AFILIACIÓN
- BANCO ADQUIRIENTE
- CIUDAD
- CÓDIGO POSTAL
- COMERCIO
- CUENTA
- ESTADO
- FECHA
- HORA
- MONTO
- AUTORIZACIÓN
- SCORE
- SIC (GIRO)
- TIPO FRAUDE
- FRAUDE STATUS
- RESULTADO
- FECHA RESULTADO
- FECHA BONIFICACIÓN

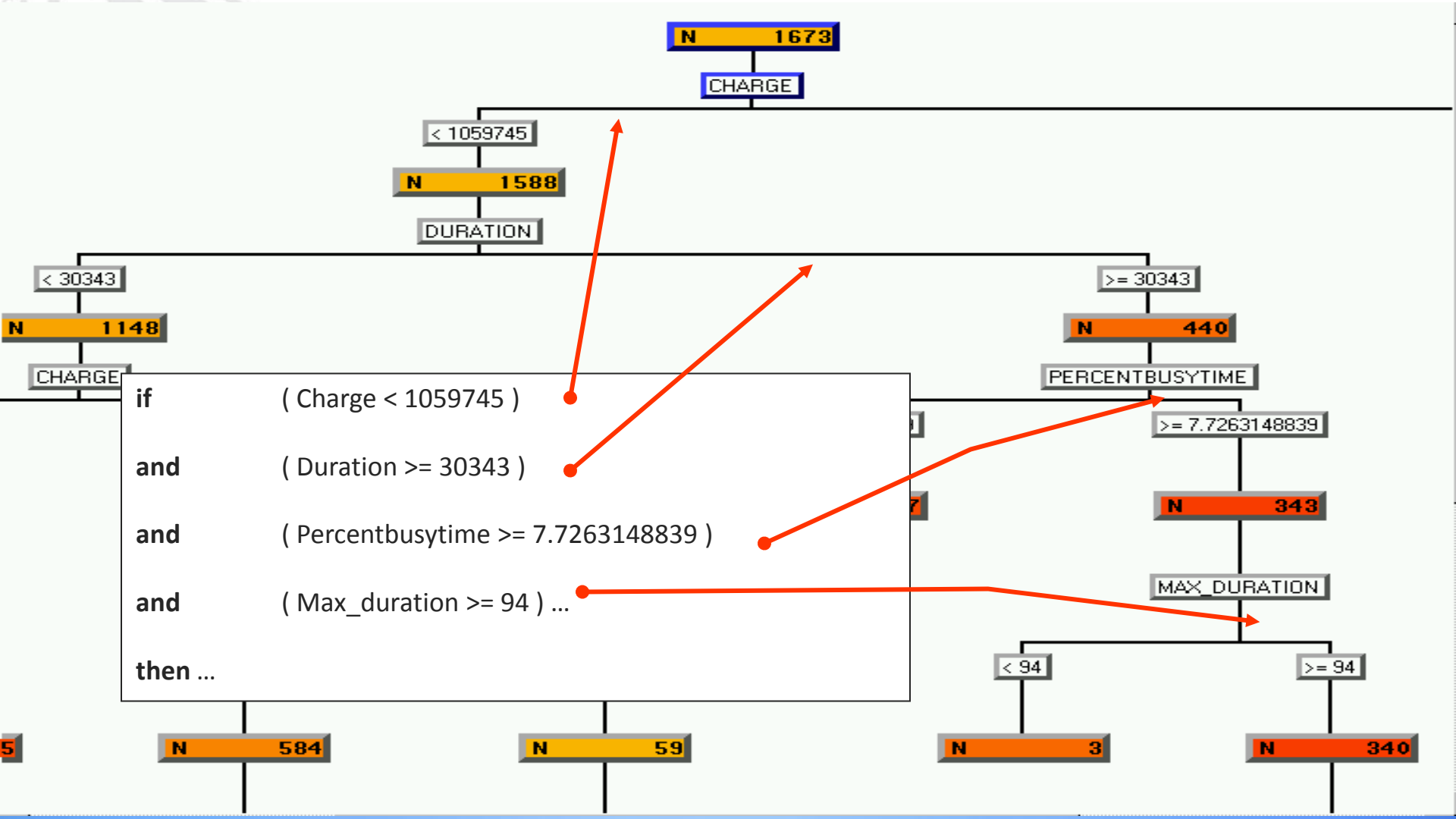
Fraude en Restaurantes



La Herramienta



Pasando de un Árbol de Decisión a Reglas

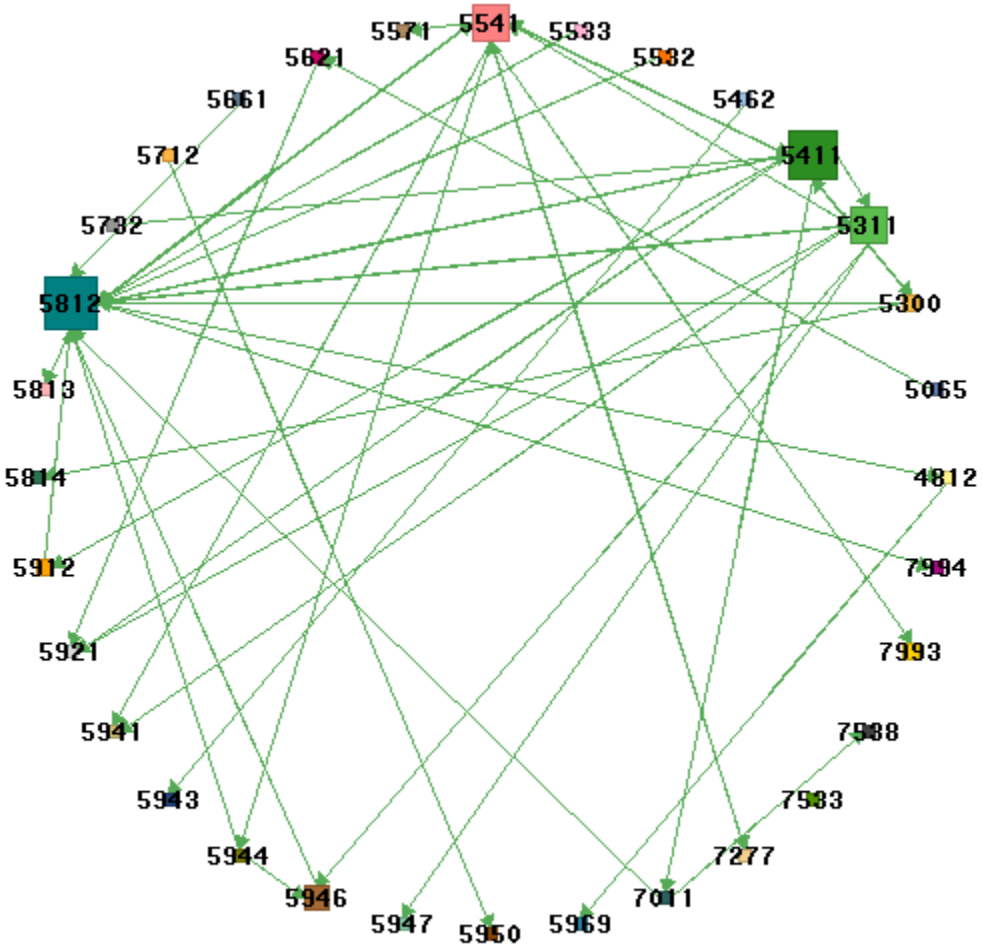


Resultados al negocio

IF MONTO < 4500
AND MODO = DESLIZADA
AND ESTADO = OAXACA
AND COMERCIO = “ZAPATERIA XXX”
THEN FRAUDE

N : 29
ROBO : 100%

Link Analysis por Actividad Comercial



Conocimientos Obtenidos

- Se tuvo un mejor conocimiento de como opera el defraudador.
- Se encontraron nuevas variables que son muy relevantes y que juegan un factor muy importante.
- Se conoció el modus operandi de las bandas delictivas.
- Se descubrio que algunos clientes se hacen “auto-fraude”
- En un banco, antes de iniciar el proyecto era el número 1 en fraude (Reporte de VISA), actualmente ocupa el último lugar.

Caso de Estudio:
En las Tiendas de Conveniencia

Tiendas de Conveniencia

Hábitos de Consumo entre Ciudades

México, Septiembre sabado y domingo, horario mañana			
N	PRODUCTO	FRECUENCIA	PORCENTAJE
1	CAPUCCINO 16OZ MEX	29884	3.7
2	COCA-COLA NR 600ML	27150	3.36
3	MARLBORO ROJO C DURA	23214	2.88
4	VASO CAFE MEDIANO	10575	1.31
5	VASO CAFE GRANDE	9453	1.17
6	NESCAFE CAPUCC.20OZ	8750	1.08
7	CAMEL CAJ DURA 20 PZ	8356	1.03
8	CAPUCC.MOKA 16OZ MEX	7864	0.97
9	COCA-COLA NR 1 LT	7394	0.92
10	AMIGO TELCEL 100PESO	7211	0.89

Guadalajara, Septiembre sabado y domingo, horario mañana			
N	PRODUCTO	FRECUENCIA	PORCENTAJE
1	CAPUCCINO ORIG. 12OZ	33624	3.85
2	COCACOLA 500ML NORET	22500	2.58
3	MARLBORO ROJO C DURA	14949	1.71
4	VASOCAFE GRANDE	14647	1.68
5	VASO CAFÉ MEDIANO	14531	1.66
6	COCA COLA 710 ML	10164	1.16
7	VASO CAFE CAP 20 OZ	10126	1.16
8	PER DIARIO DEPRECORD	9158	1.05
9	AMIGO TELCEL 100PESO	8925	1.02
10	CAPUCCINO MOKA 12OZ	8708	1

Monterrey, Septiembre sabado y domingo, horario mañana			
N	PRODUCTO	FRECUENCIA	PORCENTAJE
1	PERIODICO EL METRO	61128	3.08
2	REFRESCO PEPSI 600 M	51534	2.6
3	COCA-COLA NR 600ML	44651	2.25
4	EL NORTE DOMINICAL	33476	1.69
5	COCA- COCA 500 ML NR	31499	1.59
6	BARRILITO 750 ML BOT	29509	1.49
7	COCACOLA 355 ML RET.	24637	1.24
8	EL NORTE ORDINARIO	23460	1.18
9	LECHE LALA 1 LT	22990	1.16
10	MARLBORO LIGHTS SUAV	21457	1.08

León Bara, Septiembre sabado y domingo, horario mañana			
N	PRODUCTO	FRECUENCIA	PORCENTAJE
1	LECHE BOLSA 1L SELLO	7798	1.92
2	CLORO CHINITO 950 ML	5002	1.23
3	HIG VOGUE 500'S	4549	1.12
4	DET BLANCA NIEVES 1K	4301	1.06
5	HIG AZALEA 500HJ 4S	4285	1.05
6	COCA-COLA NR 600ML	4050	1
7	ACEITE CRISTAL 1 LT	3663	0.9
8	ARIELOXIAAZUL950G	2788	0.69
9	BOLILLO CHICO 1 PZ	2759	0.68
10	COCA-COLA NR 2.5 LT	2759	0.68

Tiendas de Conveniencia

Hábitos de Consumo entre Ciudades

México, Septiembre de lunes a viernes, horario madrugada				
N.	Soporte(%)	Confianza(%)	Conteo	Regla
1	2.92	30.05	7022	Vinos y Licores ==> Refrescos
2	2.92	11.99	7022	Refrescos ==> Vinos y Licores
3	2.5	10.26	6011	Refrescos ==> Botanas
4	2.5	33.69	6011	Botanas ==> Refrescos
5	2.08	23.08	5004	Reposteria ==> Leche fresca
6	2.08	29.15	5004	Leche fresca ==> Reposteria
7	1.88	20.87	4525	Reposteria ==> Beb Calientes No Emp
8	1.88	22.42	4525	Beb Calientes No Emp ==> Reposteria
9	1.6	6.57	3849	Refrescos ==> Com Premp con no in
10	1.6	27.83	3849	Com Premp con no in ==> Refrescos

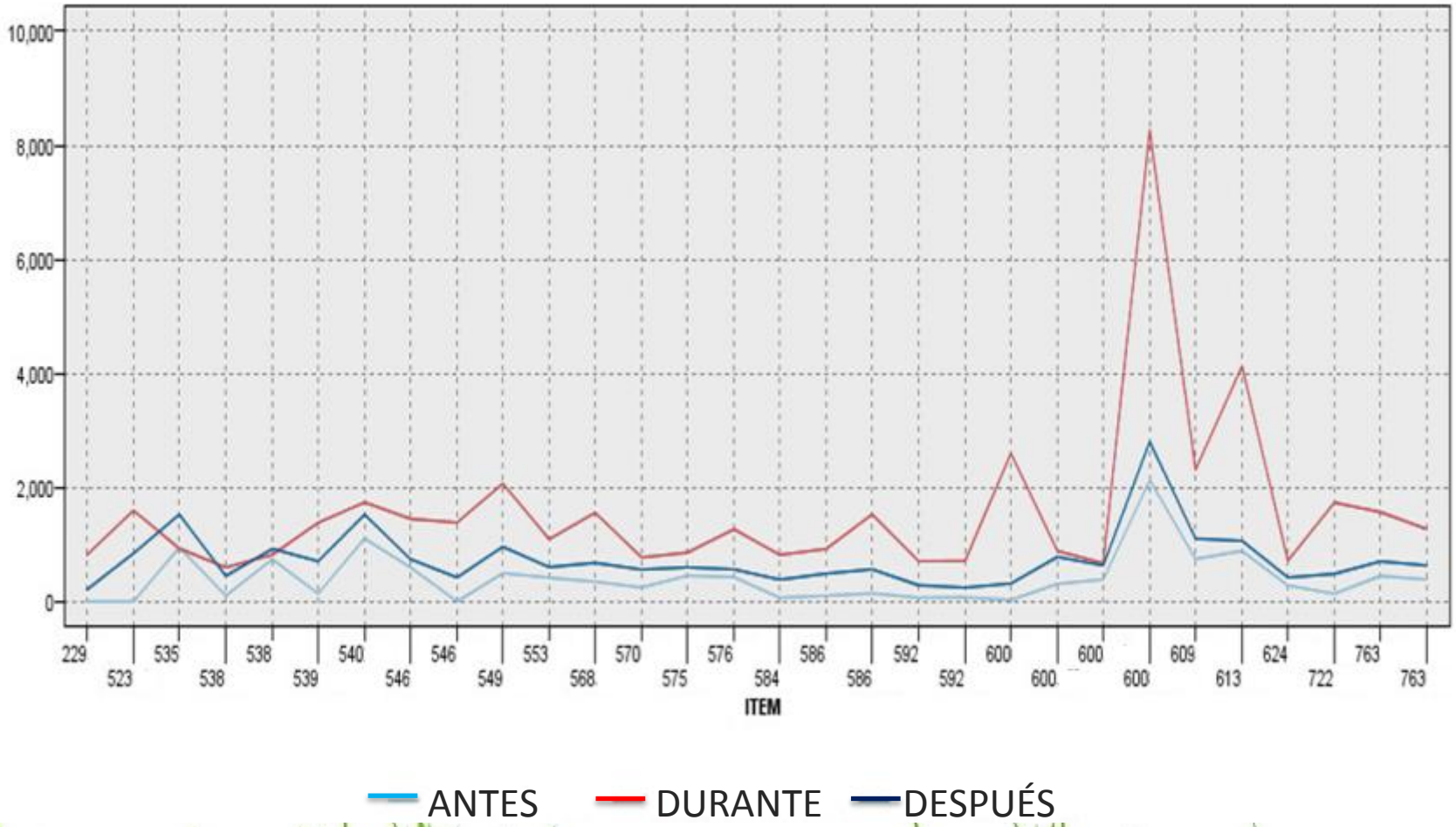
México, Septiembre de lunes a viernes, horario mañana				
N.	Soporte(%)	Confianza(%)	Conteo	Regla
1	3.47	34.03	34919	Reposteria ==> Beb Calientes No Emp
2	3.47	13.82	34919	Beb Calientes No Emp ==> Reposteria
3	3.38	28.66	34011	Galletas ==> Beb Calientes No Emp
4	3.38	13.46	34011	Beb Calientes No Emp ==> Galletas
5	1.45	25.72	14656	Yoghurt ==> Galletas
6	1.45	12.35	14656	Galletas ==> Yoghurt
7	1.38	29.84	13919	Com Premp con no in ==> Beb Calientes
8	1.38	5.51	13919	Beb Calientes No Emp ==> Com Premp
9	1.28	10.93	12937	Dulces ==> Agua Purificada
10	1.28	15.15	12937	Agua Purificada ==> Dulces

México, Septiembre de lunes a viernes, horario tarde				
N.	Soporte(%)	Confianza(%)	Conteo	Regla
1	4.17	13.02	50220	Refrescos ==> Botanas
2	4.17	35.5	50220	Botanas ==> Refrescos
3	1.58	10.88	19091	Dulces ==> Agua Purificada
4	1.58	14.71	19091	Agua Purificada ==> Dulces
5	1.57	32.75	18897	Reposteria ==> Refrescos
6	1.57	4.9	18897	Refrescos ==> Reposteria
7	1.3	16.77	15707	Jugos ==> Botanas
8	1.3	11.1	15707	Botanas ==> Jugos
9	0.81	12.59	9733	Galletas ==> Botanas
10	0.81	6.88	9733	Botanas ==> Galletas

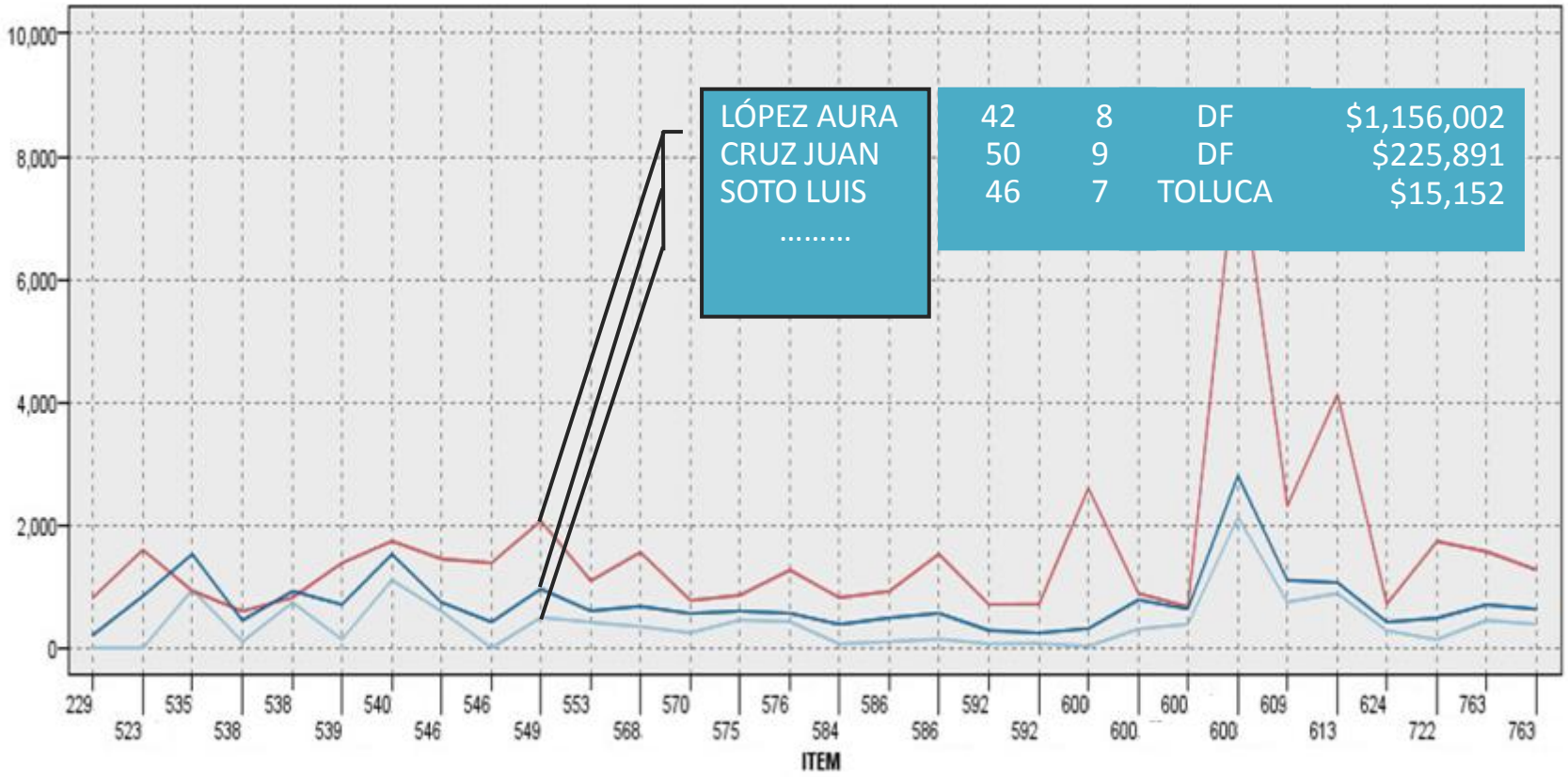
México, Septiembre de lunes a viernes, horario noche				
N.	Soporte(%)	Confianza(%)	Conteo	Regla
1	3.27	14.55	38062	Refrescos ==> Botanas
2	3.27	29.75	38062	Botanas ==> Refrescos
3	1.88	22.8	21847	Reposteria ==> Leche fresca
4	1.88	20.4	21847	Leche fresca ==> Reposteria
5	1.72	12.28	19994	Dulces ==> Botanas
6	1.72	15.63	19994	Botanas ==> Dulces
7	1.65	28.14	19261	Vinos y Licores ==> Refrescos
8	1.65	7.36	19261	Refrescos ==> Vinos y Licores
9	1.21	13.16	14092	Leche fresca ==> Galletas
10	1.21	17.42	14092	Galletas ==> Leche fresca

Caso de Estudio:
Impacto de las ofertas en una tienda del sector retail

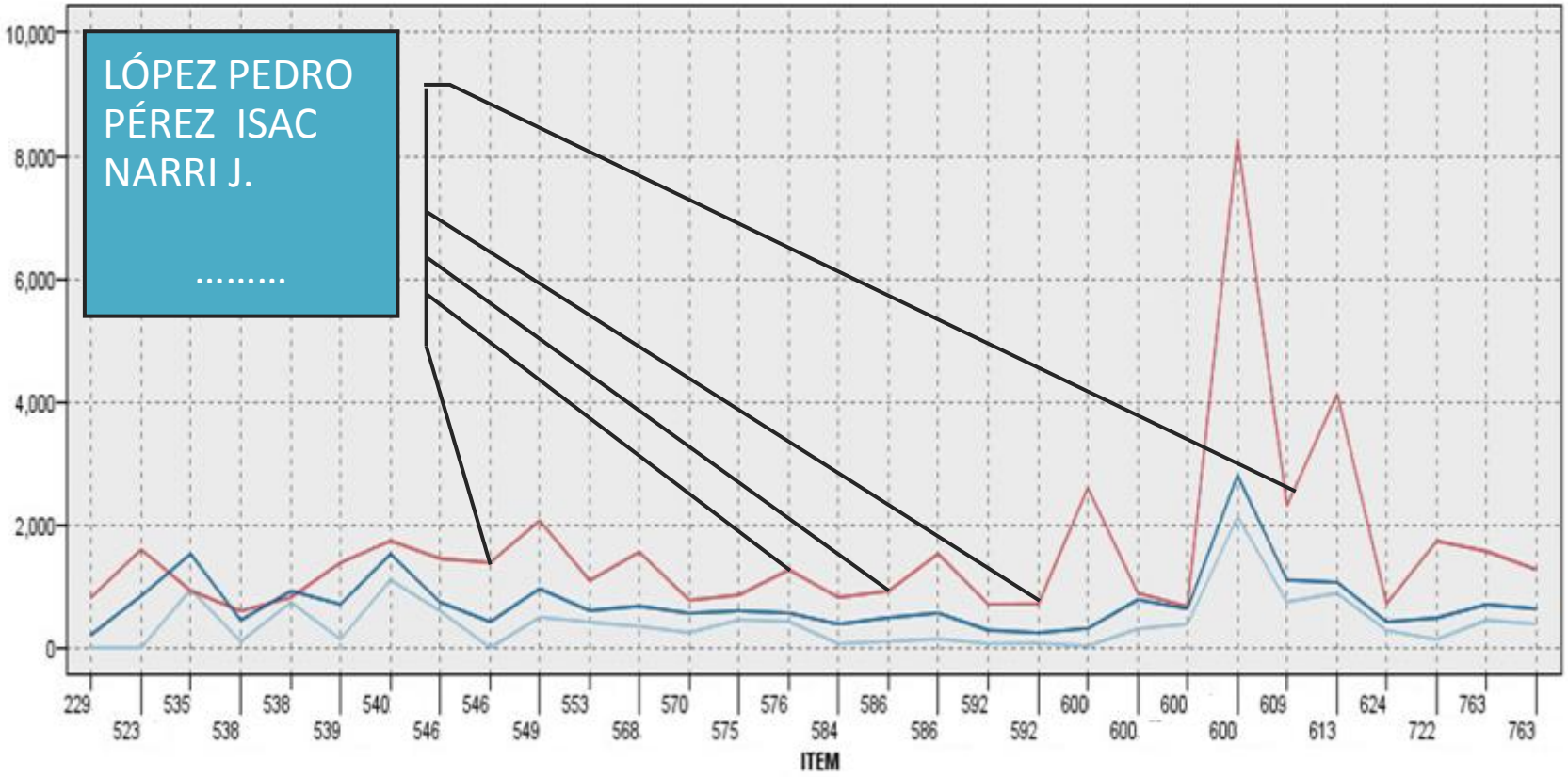
Impacto de las ofertas



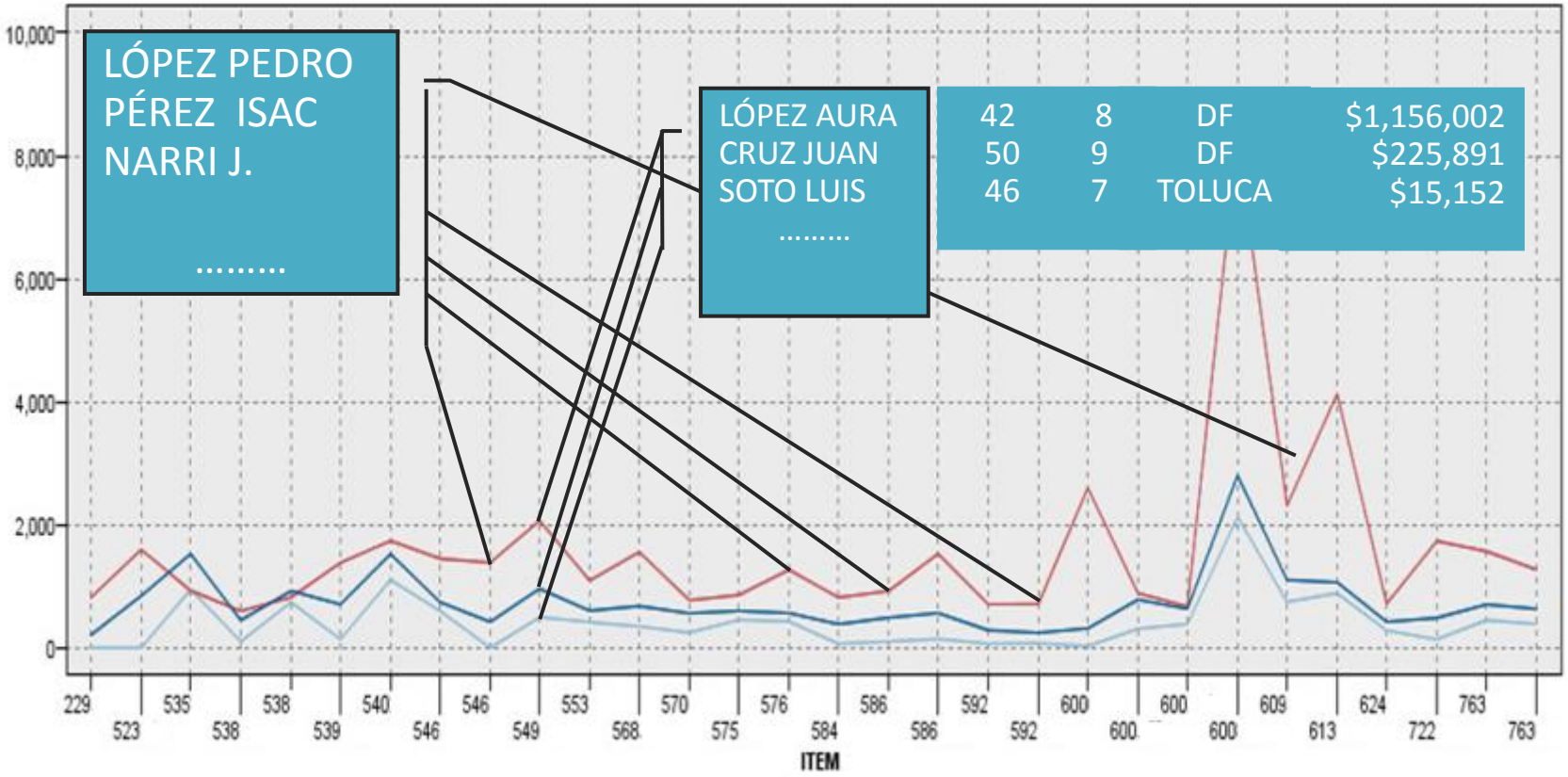
Impacto de las ofertas



Impacto de las ofertas

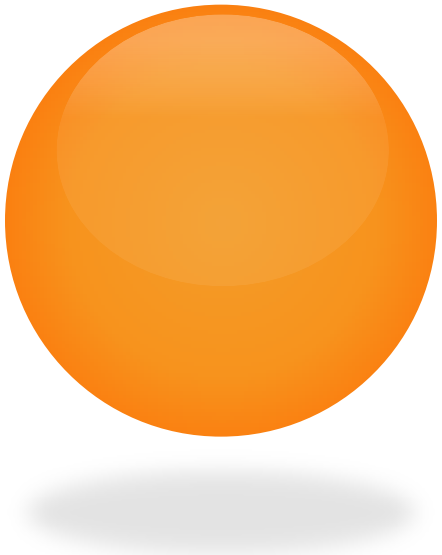


Impacto de las ofertas



¿Porqué no es una técnica tradicional?

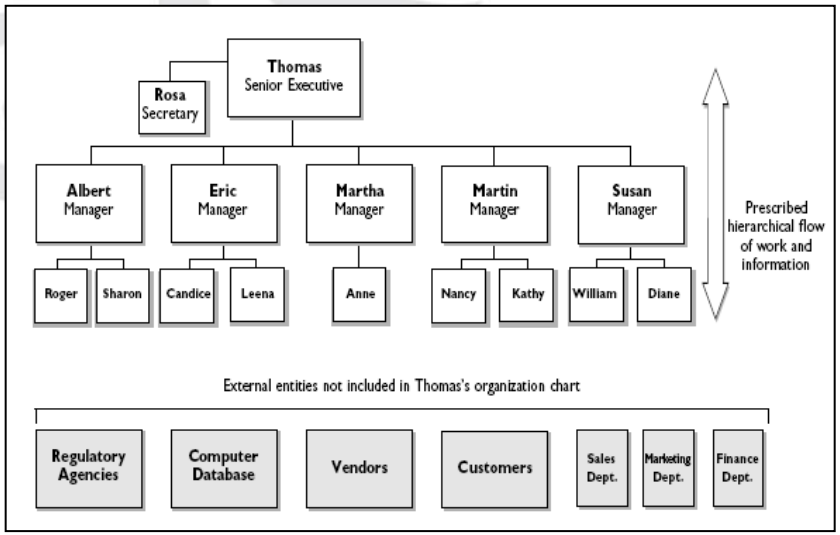
- Grandes cantidades de datos
  - Los algoritmos deben de ser altamente escalables para trabajar con terabytes de información
- Alta dimensionalidad en las bases de datos
  - Cientos o miles de variables
- Alta complejidad en los datos
  - Data streams y datos de sensores
  - Series de tiempo, Datos temporales, secuencias
  - Grafos, redes sociales y datos multi-ligados
  - Bases de datos heterogéneas
  - Datos espaciales, espacio-temporal, multimedia, texto y de Web
  - Simulaciones científicas
- Nuevas y sofisticadas aplicaciones



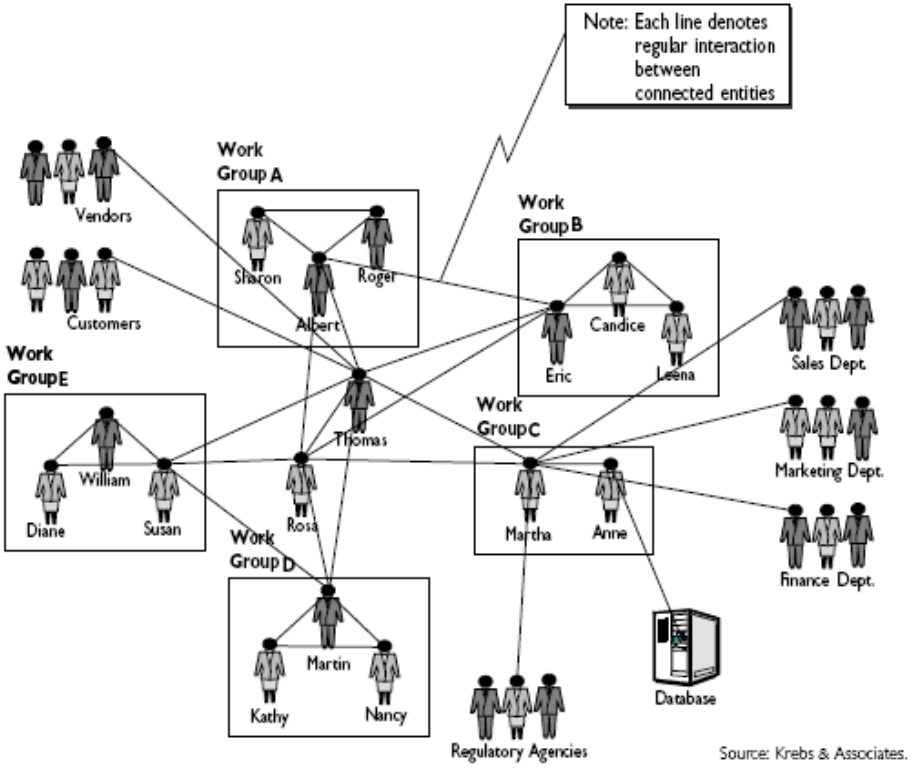
NUEVOS RETOS EN DATA MINING

LAS REDES SOCIALES
GRAFICACIÓN DE DATOS

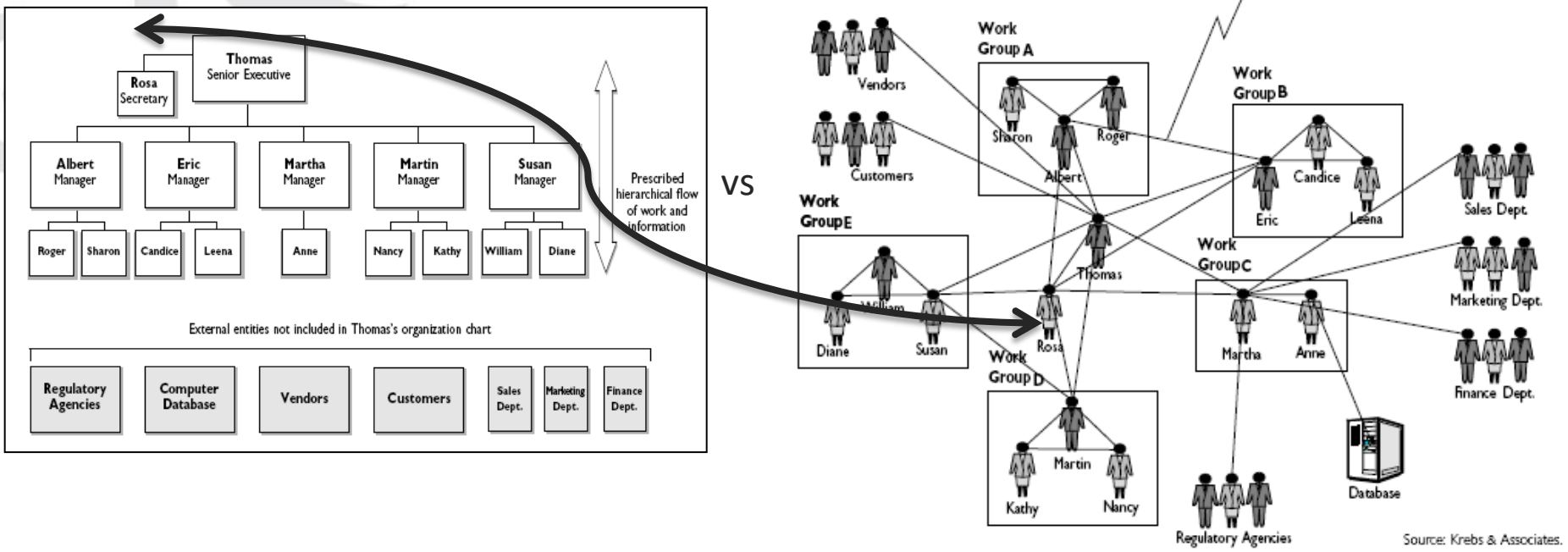
El organigrama vs Lo real



vs

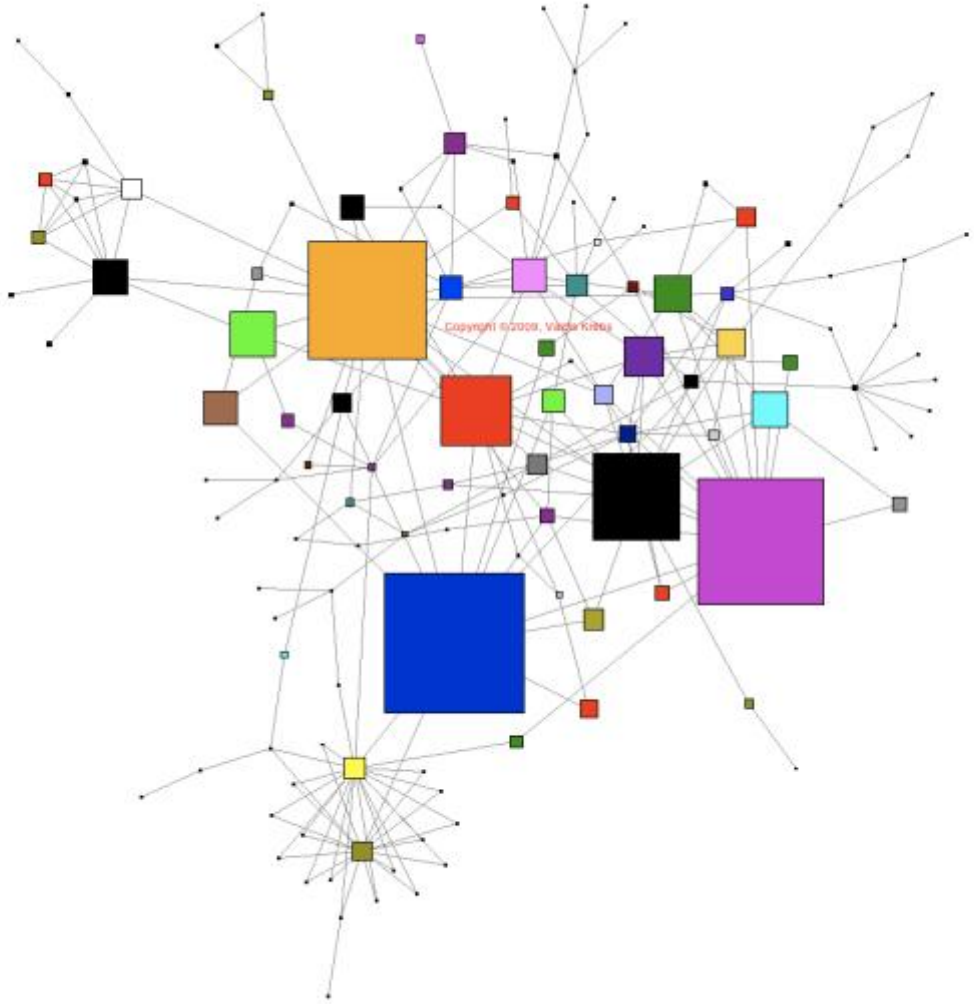


El organigrama vs Lo real

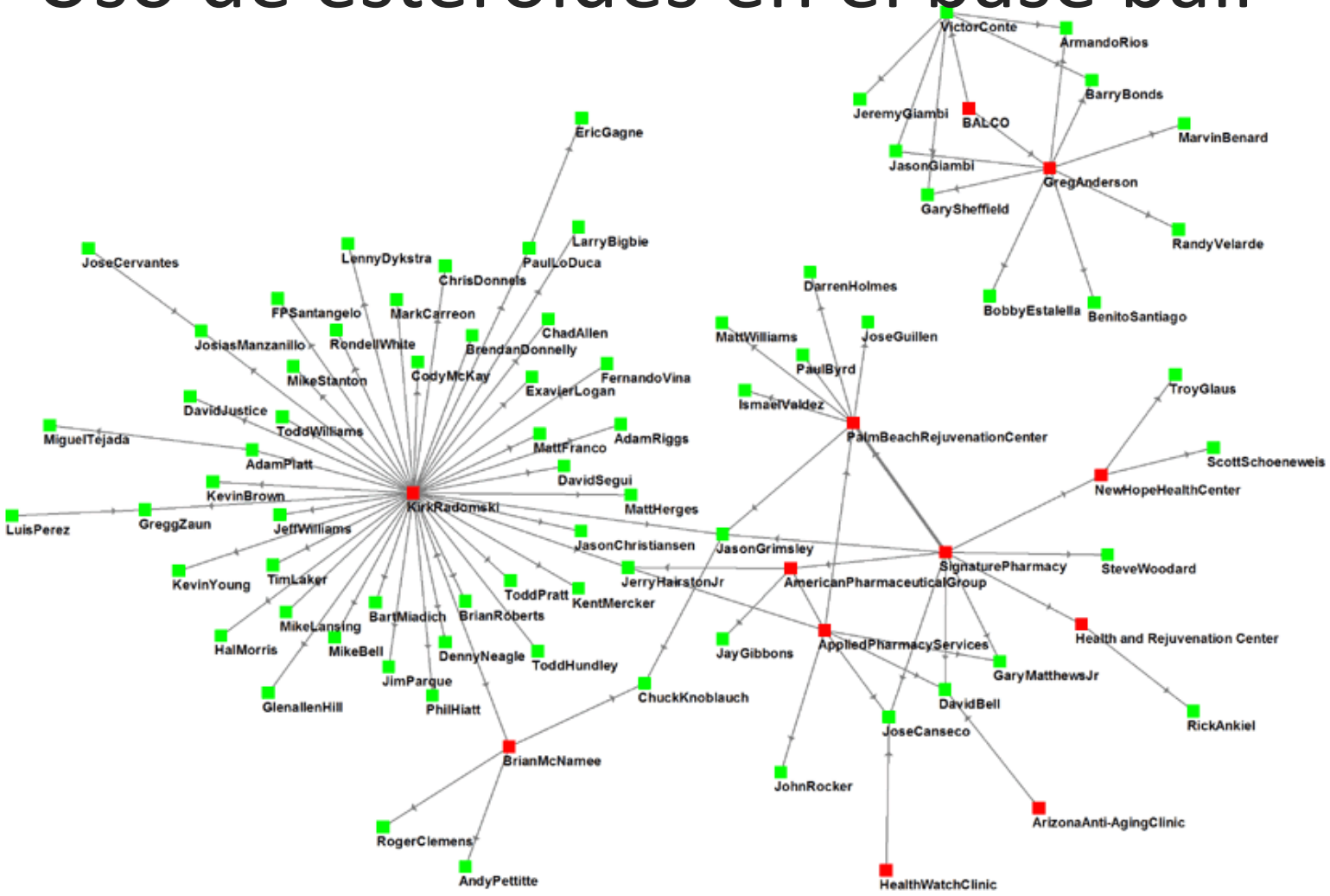


http://wordle.net

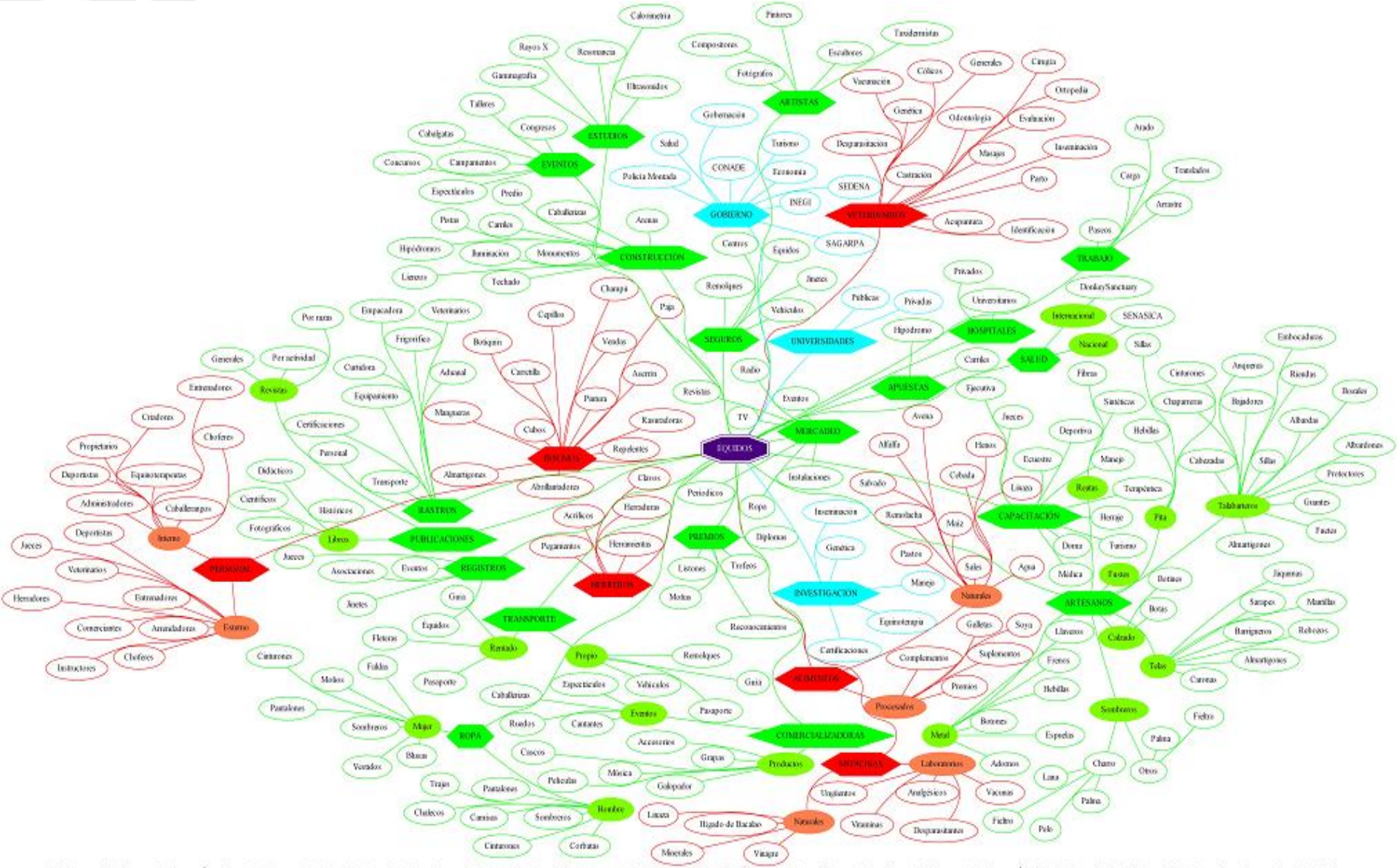
Comunicación entre sucursales



Uso de esteroides en el base ball



La industria del caballo en México



Eslabones del Sistema Producto Équido en México / Rojo = Primario / Verde = Secundario / Azul =Terciario / Autor: Luis Carlos Molina Félix lcmolinaf@hotmail.com / Versión=1.0 / Sistema Producto Équido del Estado de México / 2014 Derechos Reservados ©. México.

Conclusiones

- En México cada vez se le está dando una mejor importancia al análisis de la información.
- Gracias a la metodología CRISP-DM nos ofrece una importante guía en el desarrollo de un proyecto.
- La visualización de datos es fundamental para mostrar los resultados en Data Mining.
- Una buena herramienta pública para Data Mining es knime.
- Faltan muchos profesionales en está área.

Referencias

- Artículo: Data Mining: Torturando los datos hasta que confiesen
  - http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html
- Artículo: Del Data Mining al Big Data
  - http://www.powerbuilders.com.mx / Artículos

SG VIRTUAL CONFERENCE
7ma edición

Luis Carlos Molina

luiscarlos.molina@powerbuilders.com.mx