

Data & Text Mining



Exponentes:

- Dr. Jorge Bacallao Guerra
- Ing. Ramón Martínez
- Ing. José Villanueva

III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Introducción

- Porqué es necesario utilizar data mining
- Que és data mining y qué no es?
- Sobre que tipo de datos se puede aplicar?
- Que es text mining
- Como trabajan y como se complementan

Data & text mining

- Modelos y Técnicas de uso
- Procesos de uso
- Ciclo de data mining
- Data mining vs análisis estadístico / Data Warehousing
- Metodos (JB): Método Estadístico y del Método de árboles de regresión y clasificación

Casos Prácticos

- Sobre tema de salud materna infantil (JB)
- Sobre tema de enfermedades transmisibles-Malaria (JV)
- Uso de la plataforma de información e inteligencia en salud de OPS (RM)

Introducción

...

III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Porqué es necesario utilizar Data Mining?

Porque todo el tiempo existe un crecimiento explosivo de datos como por ejemplo:

1. Transacciones bancarias
2. Datos científicos
3. Datos astronomicos
4. Datos biologicos
5. Datos fisicos
6. La misma Web
7. Documentos y archivos de textos
8. Comercio electronico, etc.

...



III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Qué és y qué no es data mining

Data Mining es el análisis y exploración, por medios automáticos o semiautomáticos de grandes cantidades o volúmenes de datos para descubrir patrones significativos (útiles), y reglas.



La meta es permitir a la organización mejorar sus ventas, sus campañas de marketing, las operaciones de soporte a los clientes, a través de una mejor comprensión de sus clientes

Data Mining no es:

- Una búsqueda del número telefónico en un directorio telefónico.
- Realizar una búsqueda acerca de la palabra “Salud”, entre otros.

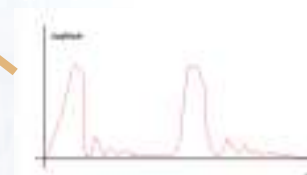


III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Sobre que tipo de datos se puede aplicar data mining

1. Data warehouses.
2. Base de datos de transacciones
3. Sistemas avanzados de base de datos como:
 1. Datos espaciales y temporales
 2. Series de tiempo
 3. Datos de tipo multimedia
 4. Textos
 5. Web
 6. Etc....



Metadatos

```
GeneFilter Comparison F
GeneFilter 1 Name:      GeneFilter
O2#1 8-20-99adjfinal    N2#1final
INTENSITIES
RAW  NORMALIZED
ORF NAME  GENE NAME  CHRM  F    G    R
YAL001C   TFC3  1      1  A  1  212.037.3
YBL080C   PET112 2      1  A  1  353.2
YBR154C   RPB5  2      1  A  1  479.2678.1
YCL044C   3      1  A  1  553.2244.1
```

III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Que es text mining

Text mining permite a la organización la posibilidad de explorar grandes cantidades de textos, no organizados en forma de datos, establecer patrones y extraer conocimientos útiles.



Data & Text Mining

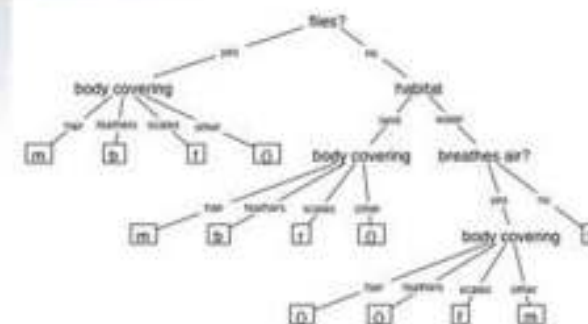
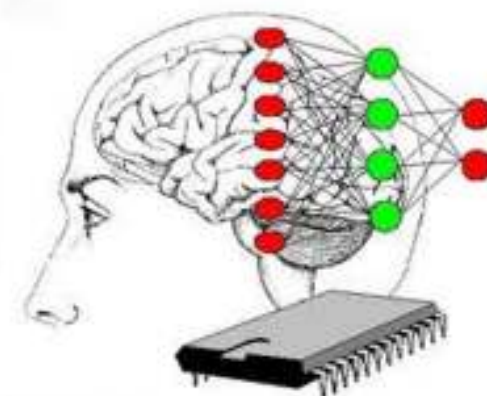
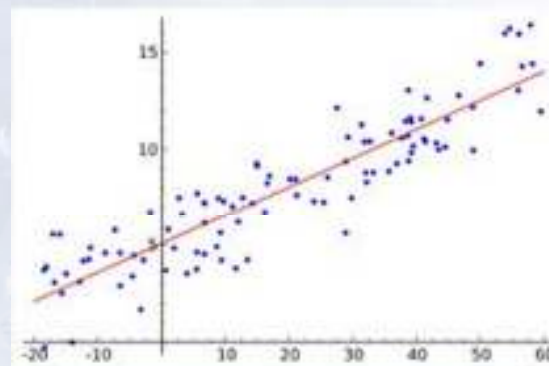
...

III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Modelos / Técnicas de uso:

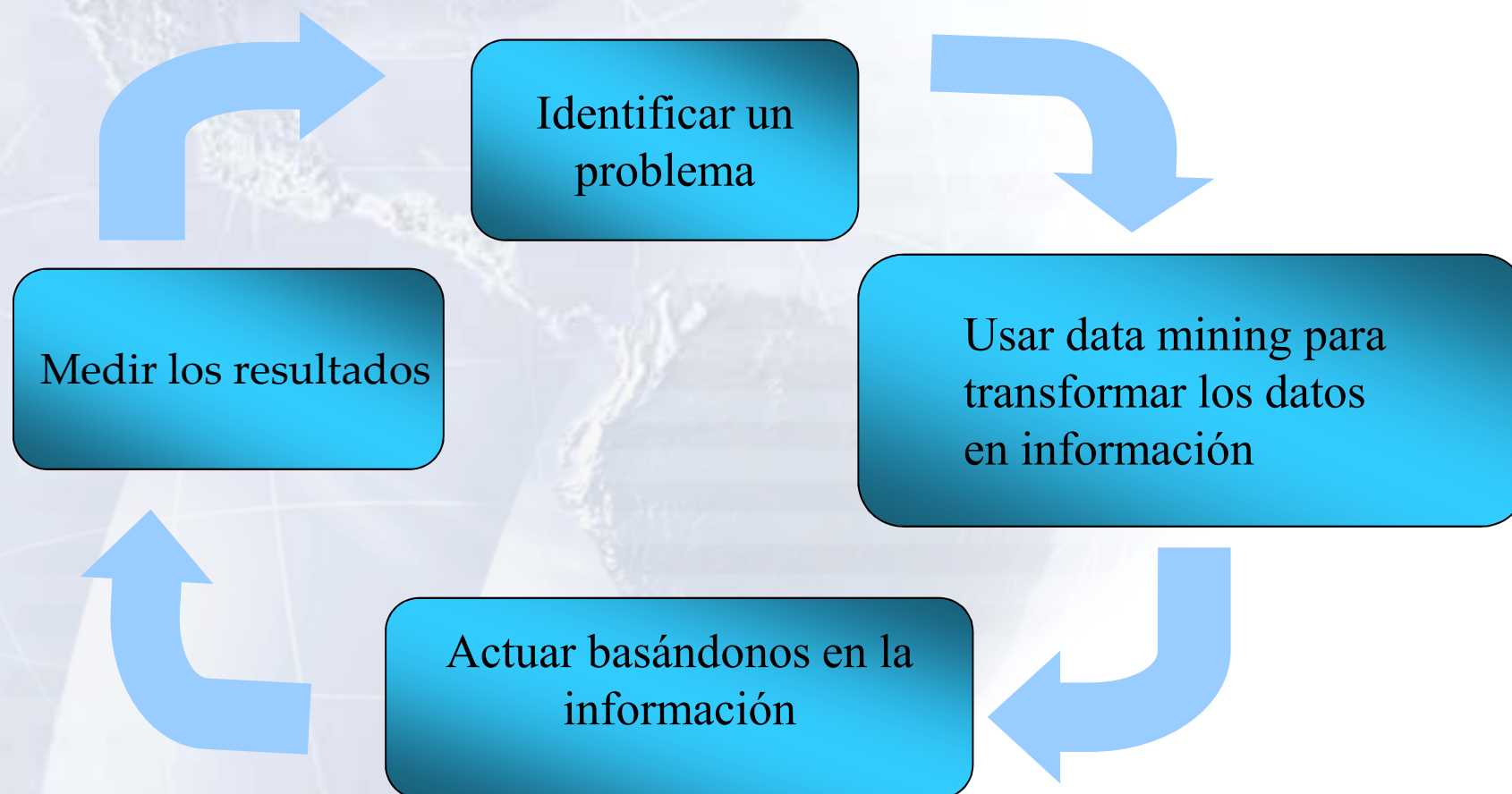
1. Redes neuronales artificiales
2. Regresión lineal
3. Árboles de decisión
4. Modelos estadísticos
5. Algoritmos genéticos
6. Agrupamiento o clustering
7. Reglas de asociación
8. Regla de inducción



III Reunión de la RedDES (GT de Tecnología de la BVS6)

Ciclo de vida:

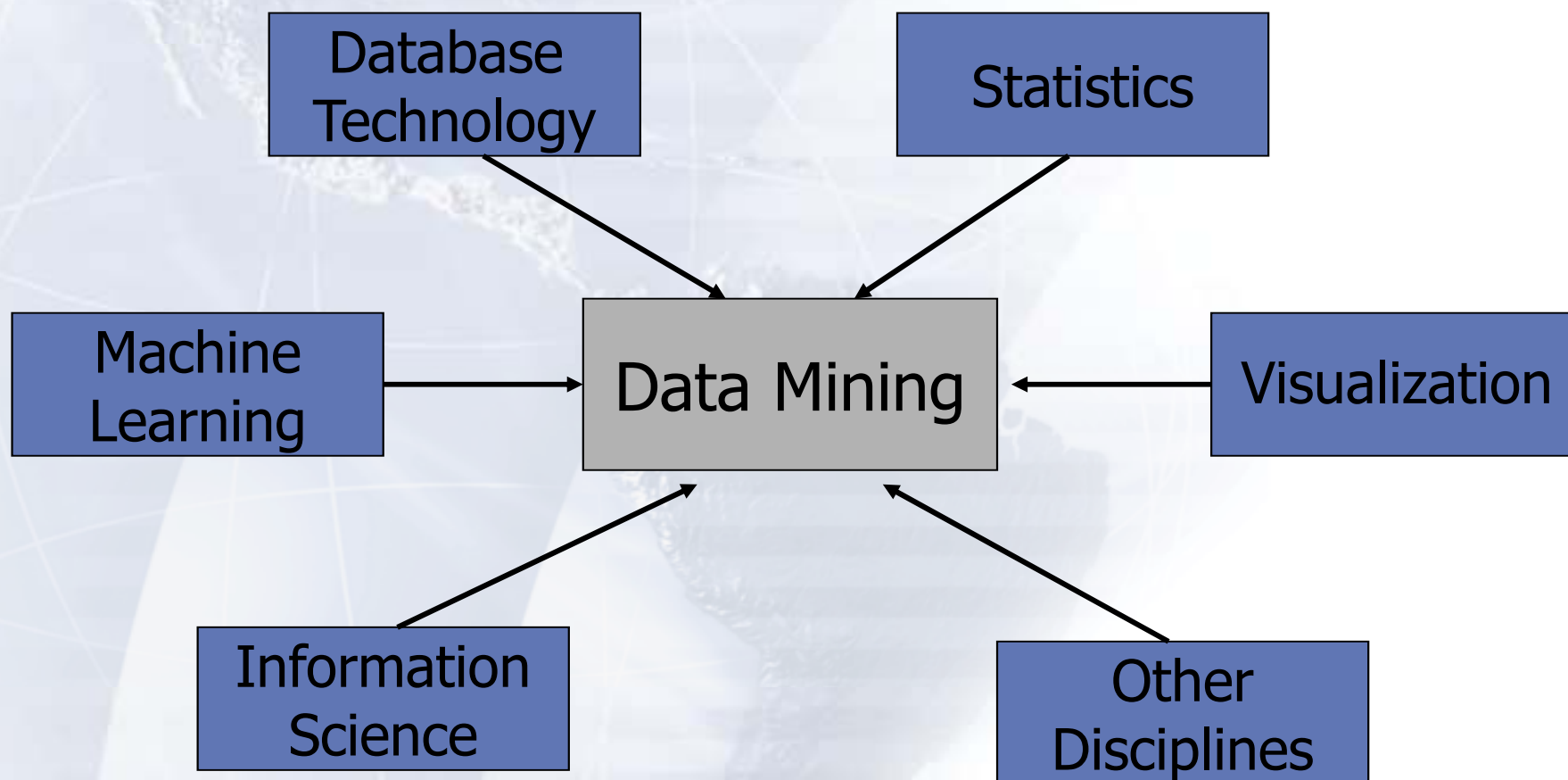
DATA & TEXT MINING



III Reunión de la RedDES (GT de Tecnología de la BVS6)

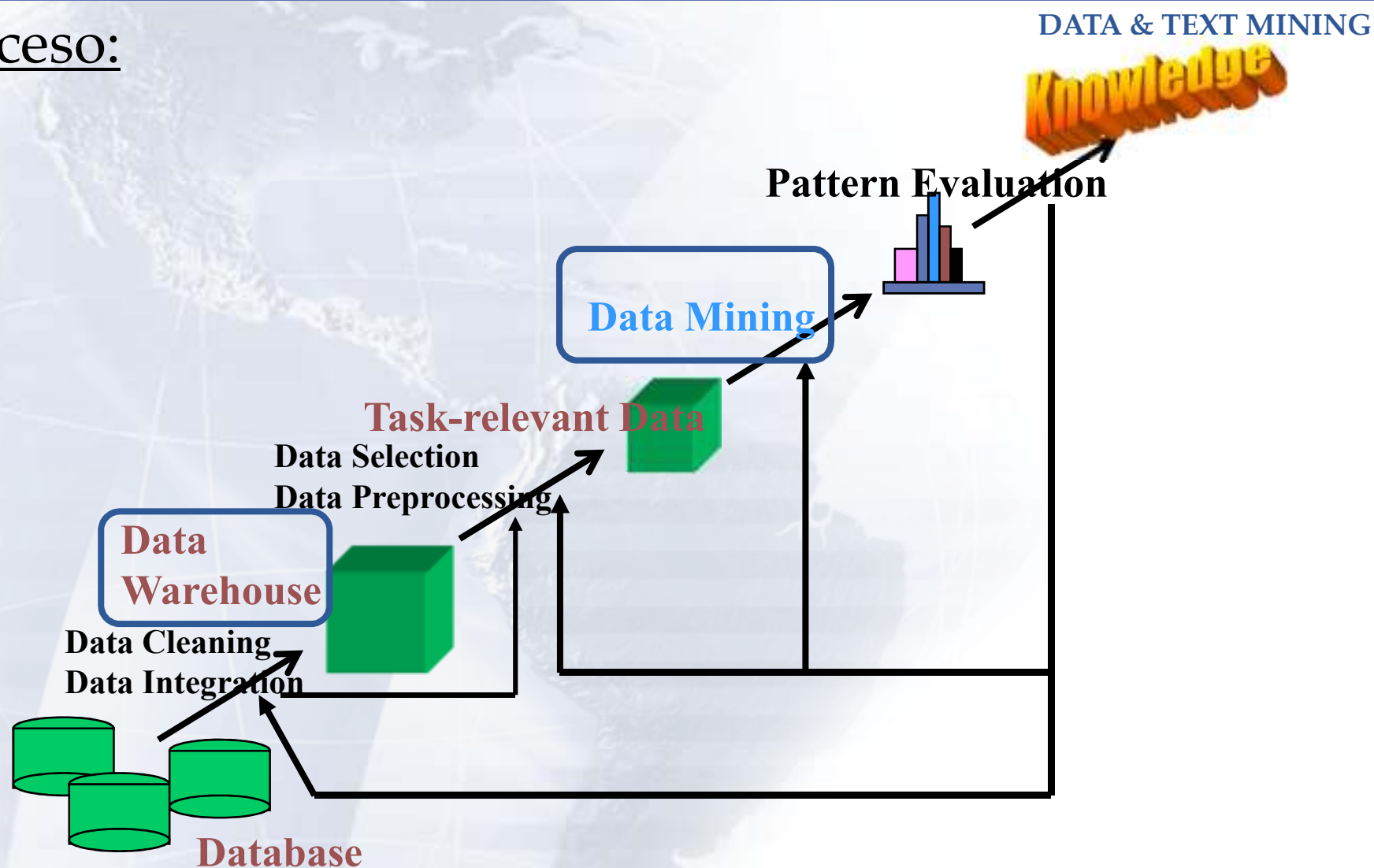
Confluencia de multiples disciplinas:

DATA & TEXT MINING



III Reunión de la RedDES (GT de Tecnología de la BVS6)

Proceso:



III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Data Mining vs Análisis Estadístico:

Análisis Estadístico:

- Poco adecuado para los tipos de datos estructurales y nominales
- Difícil y desalentador la interpretación de los datos
- Requiere apoyo de un experto.



Data Mining:

- Trabaja sobre grandes volúmenes de datos
- Maneja eficientes algoritmos y a la vez son escalables.
- Usa datos del mundo real
- Disponibilidad de métodos eficaces para la recuperación de datos.

III Reunión de la RedDES (GT de Tecnología de la BVS6)

Métodos Estadísticos:

DATA & TEXT MINING

- Existen numerosos métodos estadísticos emparentados con la naturaleza y los propósitos del Data Mining
- Estos métodos tienen características comunes en general que son: no son **inferenciales**, tienen carácter **exploratorio**, manejan **grandes volúmenes** de datos, no descansan sobre **suposiciones distribucionales**
- Cluster Analysis, Classification and Regression Trees (CART)

Árboles de Regresión y Clasificación

- Sus objetivos son clasificar, explicar y predecir (pariente de la regresión)
- Se parte de una base de datos en donde está definida una variable objetivo
- Usando las variables de la base que se consideren necesarias se parte la base en sub-bases que tienen el nombre de nodos terminales
- En cada paso, el algoritmo selecciona la variable y el punto de corte de ésta, que determinen los grupos más homogéneos (que los individuos de los grupos se parezcan mucho entre sí y a la vez se diferencien mucho los grupos entre sí)

Árboles de Regresión y Clasificación

- Se debe definir un criterio de distancia o similaridad
- Se debe considerar si podar el árbol o no hacerlo
- Se debe decidir qué método usar y los tamaños de los nodos terminales
- Se debe elegir correctamente el conjunto de variables predictoras
- Una variable puede entrar al árbol más de una vez con distintos puntos de corte

Ventajas de los Árboles de Regresión y Clasificación

- Pueden ser utilizados en muchos contextos
- Sirven para predecir, explicar y clasificar
- No tienen supuestos distribucionales
- Son una excelente herramienta descriptiva y exploratoria
- Tienen un costo computacional relativamente bajo
- Están implementados en la mayoría de los softwares de análisis de datos

CASOS PRÁCTICOS

...

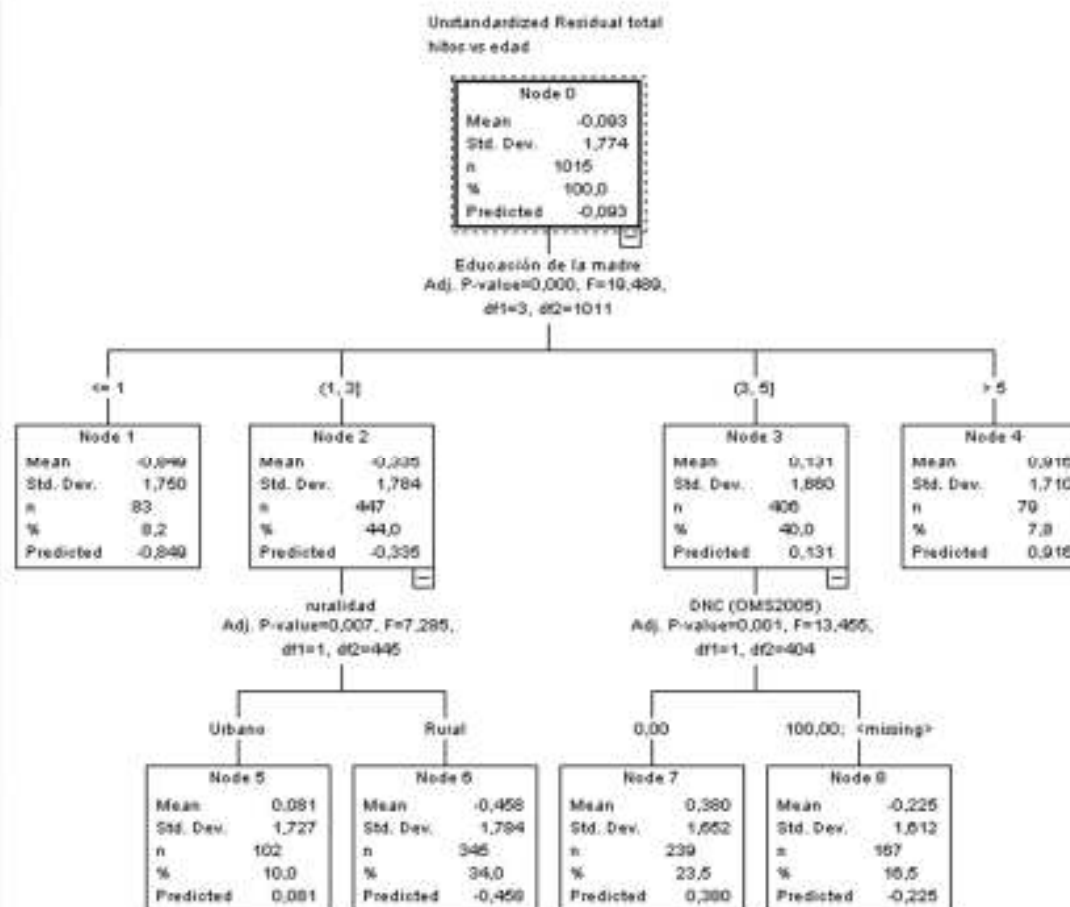
Caso 1: Nutrición y Desarrollo

- Estudio sobre desarrollo infantil en donde se consideran 13 hitos de desarrollo, 5 motores y 8 de lenguaje
- Se tienen datos de niños de distintas zonas de Perú (desarrollo, nutrición, escolaridad de los padres, condición socioeconómica, estimulación en el hogar)
- Interesa conocer como influyen las variables del contexto en el desarrollo de los niños con vista a futuras intervenciones
- Bases grandes, con datos mixtos, continuos, categóricos, etc
- Variable objetivo: cantidad de hitos alcanzados para la edad

III Reunión de la RedDES (GT de Tecnología de la BVS6)

Caso 1: Nutrición y Desarrollo

DATA & TEXT MINING



III Reunión de la RedDES (GT de Tecnología de la BVS6)

Caso 2: Enfermedades Transmisibles: Malaria

DATA & TEXT MINING

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	DEPARTAMENTO	PROVINCIA	DISTRITO	LOCALIDAD	Nivel_Riesgo_1 IPA	IPA	PRESENCIA VECTORIAL	ACCESIBILIDAD A SERVICIOS	ESPECIE PLASMODIAL	MIGRACION	SUSCEPTIBILIDAD A LAS ADICIONES	PLUVIOSIDAD	CAPACIDAD MICROSCOPIA	TIPO DE CIUDAD	PROTECTOR DE DESGARRO	TIPO DE VITANDA	TOTAL	Escenario, Nivel de Factores de Riesgo	Escenario 1
1																			
2	TUMBES	ZARUVILLA	AGUAS VERDES	AGUAS VERDES	Alto Riesgo	2	1	0	1	1	1	2	0	1	0	2	11	IPA: ≥ 10, Fact. Riesgo: 14 a 17	
3	TUMBES	ZARUVILLA	AGUAS VERDES	AGUAS VERDES	Alto Riesgo	2	1	0	1	1	1	2	1	2	0	2	14	IPA: ≥ 10, Fact. Riesgo: 14 a 17	
4	TUMBES	ZARUVILLA	AGUAS VERDES	AMH ALBERTO FUIMOR	Alto Riesgo	2	1	0	1	1	1	2	1	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
5	TUMBES	ZARUVILLA	AGUAS VERDES	LA CURVA	Alto Riesgo	2	1	0	1	1	1	2	1	1	0	2	12	IPA: < 1, Fact. Riesgo: 10 a 13	
6	TUMBES	ZARUVILLA	AGUAS VERDES	COMPLEJO	Alto Riesgo	2	1	0	1	1	1	2	1	1	0	2	12	IPA: < 1, Fact. Riesgo: 10 a 13	
7	TUMBES	ZARUVILLA	AGUAS VERDES	DE DE JAU	Alto Riesgo	2	1	0	1	1	1	2	1	1	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
8	TUMBES	ZARUVILLA	AGUAS VERDES	VILLA PRIMAVERA	Alto Riesgo	2	1	0	1	1	1	2	1	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
9	TUMBES	ZARUVILLA	ZARUVILLA	AMH EDMUNDO ROVER	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
10	TUMBES	ZARUVILLA	ZARUVILLA	MRAFLORES	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
11	TUMBES	ZARUVILLA	ZARUVILLA	ZARUVILLA CENTRO	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
12	TUMBES	ZARUVILLA	PAPAYAL	LA PALMA	Alto Riesgo	2	1	0	1	1	1	2	0	2	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
13	TUMBES	ZARUVILLA	PAPAYAL	PAPAYAL	Alto Riesgo	2	1	0	1	1	1	2	0	2	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
14	TUMBES	ZARUVILLA	PAPAYAL	LECHUGAL	Alto Riesgo	2	1	0	1	1	1	2	0	2	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
15	TUMBES	ZARUVILLA	TUMBES	ANDRES ARAUJO MORAN	Mediano Riesgo	1	1	0	1	1	1	2	0	1	0	2	10	IPA: < 1, Fact. Riesgo: 10 a 13	
16	TUMBES	ZARUVILLA	TUMBES	AMH MIGUEL ORAU	Alto Riesgo	2	1	0	1	1	1	2	0	2	0	2	14	IPA: ≥ 10, Fact. Riesgo: 14 a 17	
17	TUMBES	ZARUVILLA	TUMBES	SAN JOSE	Alto Riesgo	2	1	0	1	1	1	2	0	2	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
18	TUMBES	ZARUVILLA	TUMBES	PAMPAGRANDE	Mediano Riesgo	1	1	0	1	1	1	2	0	1	0	2	10	IPA: < 1, Fact. Riesgo: 10 a 13	
19	TUMBES	ZARUVILLA	TUMBES	TUMBES	Alto Riesgo	2	1	0	1	1	1	2	0	1	0	2	11	IPA: < 1, Fact. Riesgo: 10 a 13	
20	TUMBES	ZARUVILLA	TUMBES	SALAMANCA	Alto Riesgo	2	1	0	1	1	1	2	0	1	0	2	11	IPA: < 1, Fact. Riesgo: 10 a 13	
21	TUMBES	ZARUVILLA	TUMBES	LOS JARDINES	Alto Riesgo	2	1	0	1	1	1	2	0	1	0	2	11	IPA: < 1, Fact. Riesgo: 10 a 13	
22	TUMBES	ZARUVILLA	TUMBES	AMH MIGUEL ORAU	Mediano Riesgo	1	1	0	1	1	1	2	0	1	0	2	10	IPA: < 1, Fact. Riesgo: 10 a 13	
23	TUMBES	ZARUVILLA	CORRALES	LA GARITA	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
24	TUMBES	ZARUVILLA	CORRALES	EL TAGLAZO	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
25	TUMBES	ZARUVILLA	CORRALES	BUENA VISTA ALTA	Mediano Riesgo	1	1	0	1	1	2	2	0	2	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
26	TUMBES	ZARUVILLA	CORRALES	BUENA VISTA BAJA	Mediano Riesgo	1	1	0	1	1	2	2	0	2	0	1	11	IPA: < 1, Fact. Riesgo: 10 a 13	
27	TUMBES	ZARUVILLA	CORRALES	CERCADO	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
28	TUMBES	ZARUVILLA	CORRALES	MALVAL	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
29	TUMBES	ZARUVILLA	CORRALES	SAN PEDRO	Alto Riesgo	2	1	0	1	1	2	2	0	2	0	1	12	IPA: < 1, Fact. Riesgo: 10 a 13	
30	PIURA	MORROPON	FAQUE	LOMA LARGA	Bajo Riesgo	0	1	1	1	1	1	2	0	2	1	2	13	IPA: < 1, Fact. Riesgo: 10 a 13	IPA: < 1, NU
31	PIURA	SULLANA	BELLAVISTA	BELLAVISTA	Alto Riesgo	2	1	1	1	1	1	2	0	2	0	2	13	IPA: ≥ 10, Fact. Riesgo: 10 a 13	NULL NU
32	PIURA	SULLANA	BELLAVISTA	TUPAC AMARU	Alto Riesgo	2	1	1	1	0	1	2	1	2	0	2	13	IPA: ≥ 10, Fact. Riesgo: 10 a 13	NULL NU

III Reunión de la RedDES (GT de Tecnología de la BVS6)

Caso 2: Enfermedades Transmisibles: Malaria DATA & TEXT MINING

The image shows a screenshot of the 'Anexo OPS Formulario 2012' for Malaria. The form is titled 'Organización Panamericana de la Salud' and 'Anexo OPS Formulario 2012'. It contains several sections with tables for data entry, including 'Estratificación por área de riesgo', 'Estratificación por edad y sexo', 'Estratificación por sub-grupos', and 'Estratificación según inicio de tratamiento'. The tables have columns for 'Área', 'Edad', 'Sexo', 'Sub-grupo', and 'Inicio de tratamiento'. The form is designed for data collection and reporting on malaria cases.

- Estratificación por área de riesgo
 - Láminas examinadas
 - Casos confirmados
 - Población
- Estratificación por edad y sexo
 - < 5 años
 - 5 – 14 años
 - 15 -49 años
 - >= 50 años
- Estratificación por sub-grupos
 - Mujeres embarazadas
 - Casos importados (viajeros)
 - Grupos indígenas,
- Estratificación según inicio de tratamiento
- Política actual de antimaláricos
- Rociado interno residual
- Etc...

III Reunión de la RedDES (GT de Tecnología de la BVS6)

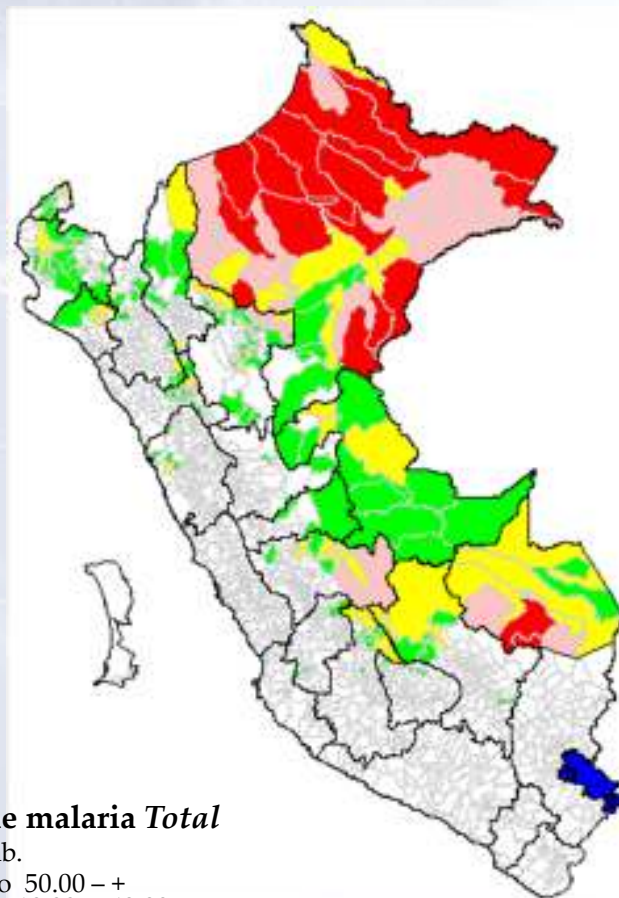
Caso 2: Enfermedades Transmisibles: Malaria

DATA & TEXT MINING



III Reunión de la RedDES (GT de Tecnología de la BVS6)

Caso 2: Enfermedades Transmisibles: Malaria DATA & TEXT MINING



Estratificación de malaria *Total*

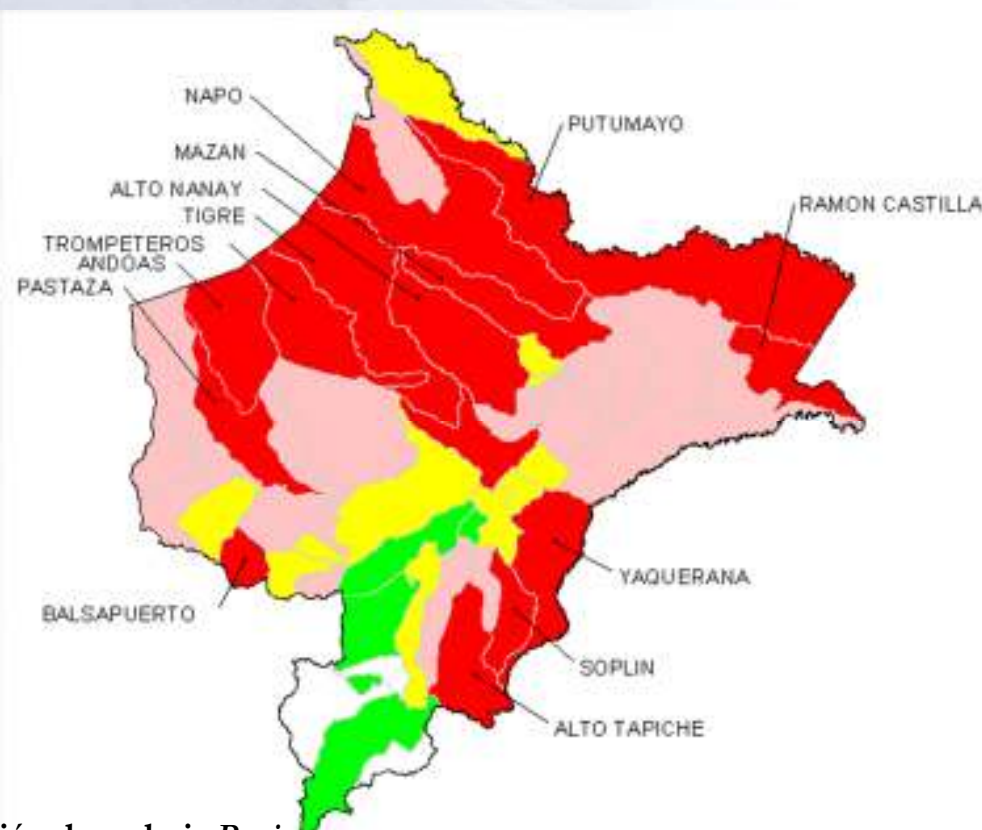
Incidencia x 1000 hab.

	Muy alto riesgo	50.00 – +
	Alto riesgo	10.00 – 49.99
	Mediano riesgo	1.00 – 9.99
	Bajo riesgo	0.01 – 0.99

DIRESA: 23
Departamento: 18
Provincia: 72
Distrito: 250
Casos: 36886
Defunciones: 2

III Reunión de la RedDES (GT de Tecnología de la BVS6)

Caso 2: Enfermedades Transmisibles: Malaria DATA & TEXT MINING



Estratificación de malaria *P. vivax*

Incidencia x 1000 hab.

■	Muy alto riesgo	50.00 – +
■	Alto riesgo	10.00 – 49.99
■	Mediano riesgo	1.00 – 9.99
■	Bajo riesgo	0.01 – 0.99

Departamento: Loreto

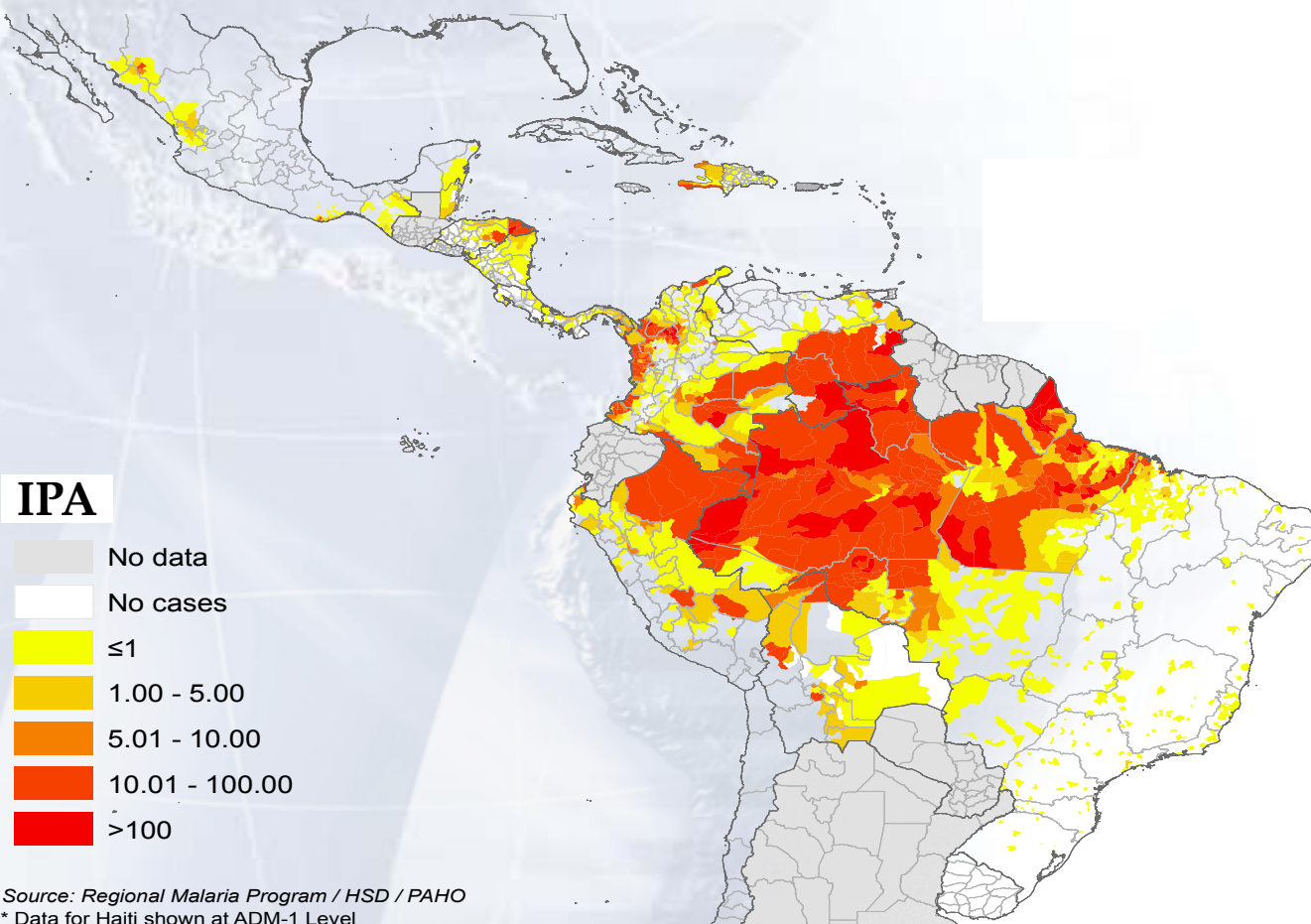
Provincia: 07

Distrito: 48

Casos: 22032

III Reunión de la RedDES (GT de Tecnología de la BVS6)

Caso 2: Enfermedades Transmisibles: Malaria DATA & TEXT MINING



III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Caso 3: Plataforma de Información e Inteligencia en Salud (PHIP) de la OPS

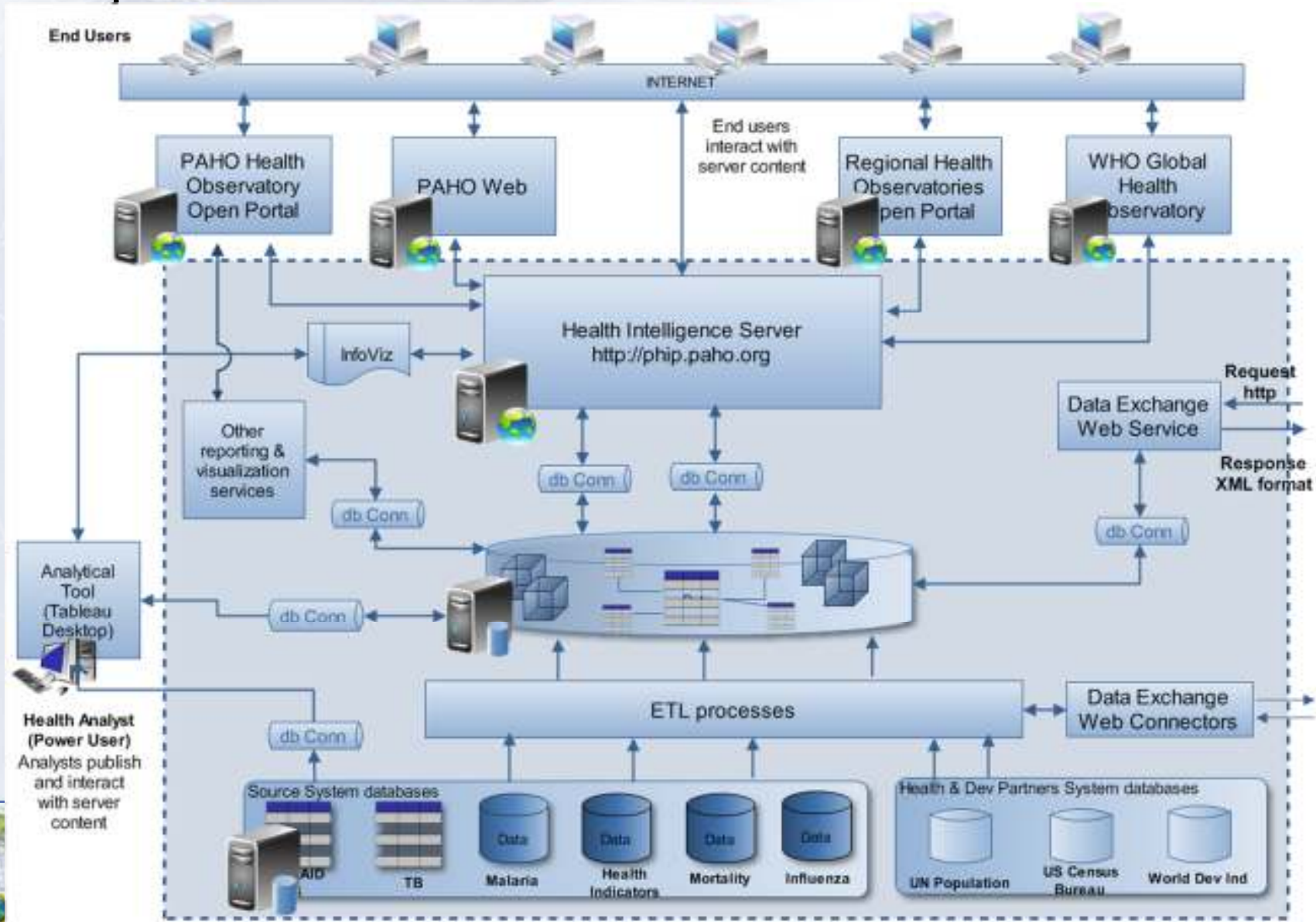
Contenido del tema

- Plataforma de Información de e Inteligencia en Salud
 - Arquitectura de PHIP
 - Data Warehouse: datos disponibles y parámetros de conexión
- Demostración práctica de análisis exploratorio de datos de mortalidad

III Reunión de la RedDES (GT de Tecnología de la BVS6)

Arquitectura de PHIP

DATA & TEXT MINING



III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Caso 3: Plataforma de Información e Inteligencia en Salud (PHIP) de la OPS

Data Warehouse de Salud: Datos disponibles

- **Base de datos regional de Mortalidad:** datos nominales de defunciones, y defunciones agregadas por las dimensiones: país, año, sexo, edad, causa básica de muerte.
- **Estimaciones de población:** División de Población de Naciones Unidas y Buró de Censo de USA.
- **Indicadores Básicos de Salud.** 114 indicadores por país y año.
- **Datos de programas técnicos:** Tuberculosis
- **Ayuda Internacional para el Desarrollo de la Salud**

Data Warehouse de Salud: Parámetros de conexión

Data Engine: Microsoft SQL Server

Servidor: HA-DBSERV-02

Username: DWreader

Password: dwreader

Base de datos: HealthDataWarehouse

Más información: <http://sites.paho.org/rho/SitePages/Data%20and%20Metadata.aspx>

III Reunión de la RedDES (GT de Tecnología de la BVS6)

DATA & TEXT MINING

Caso 3: Plataforma de Información e Inteligencia en Salud (PHIP) de la OPS

Demostración práctica: Análisis exploratorio de mortalidad

- **Preguntas de análisis:**
 - De los grupos amplios de causas de muerte -- Enfermedades No Transmisibles, Enfermedades Transmisibles, y Causas Externas -- **cual** es el grupo más frecuente?
 - Cual es la distribución por país, año, edad y sexo de la mortalidad por causas externas? Puede identificarse algún patrón en la distribución?