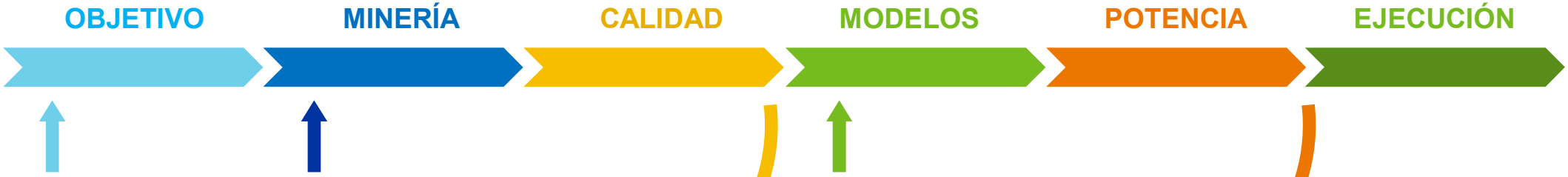
Generación Ontología - Suceso

INSTITUTO IBERMÁTICA DE INNOVACIÓN Ibermática

DESARROLLO DEL MODELO

Metodología: CRISP-DM

CRISP-DM es una metodología de desarrollo de sistemas de explotación de información que permite el **aseguramiento de la calidad técnica**. Los datos deben ser extraídos, depurados y preparados para su uso e interpretación. Este método se divide en 6 fases: conocimiento del negocio, conocimiento de los datos, preparación de los datos, modelización, evaluación y desarrollo.



OBJETIVO DE NEGOCIO Y ROI

Determinar la relación entre dosis de antipsicótico y disminución de la tensión arterial (TA) en personas expuestas a ejercicio físico de alta intensidad.

OBJETIVO DE MINERIA

- Modelado estadístico para determinar la validez de la hipótesis planteada.
- Perfilado de pacientes en base contexto físico, bioquímico y médico
- Determinar la interrelación entre los perfiles de cada contexto

MEDIR LA CALIDAD

Si la calidad de los datos no supera un umbral mínimo, incluso aplicando asunciones para mejorar los datos actuales, se tendrá que volver a etapas anteriores del proceso.

MODELOS ANALÍTICOS Y FOCO:

- Modelado MANOVA de 2 vías (medicamento y dosis)
- Modelado MANOVA de 1 vía (variables calculadas)
- Perfilado de pacientes y generación de reglas de conocimiento
- Modelado ANOVA de medidas repetidas
- Perfilado en base a series temporales y generación de reglas de conocimiento

MEDIR LA POTENCIA

Si la potencia de los modelos no supera un umbral, se tendrá que volver a etapas anteriores del proceso porque los datos no explican la hipótesis a demostrar.

INSTITUTO IBERMÁTICA DE INNOVACIÓN

Ibermática

OBJETIVO DE NEGOCIO Y DE MINERÍA

OBJETIVO DE NEGOCIO Y DE MINERÍA

ANALITICA AVANZADA EN CIBERSEGURIDAD INDUSTRIAL

OBJETIVO DE NEGOCIO

1. Descubrimiento de las relaciones existentes entre las distintas capas en las tramas de protocolos de transporte y comunicación en entornos industriales (IT/OT)
2. Detección de Vulnerabilidades y Ataques

OBJETIVO DE MINERÍA

• No existe un objetivo de minería claro, dada que la información de origen no contiene, “a priori”, anotaciones de ataques.

• Determinar de forma automática cuáles son los diferentes niveles de información a analizar, en una estrategia top-down.

•Dichas jerarquías serán nuestros objetivos de minería.

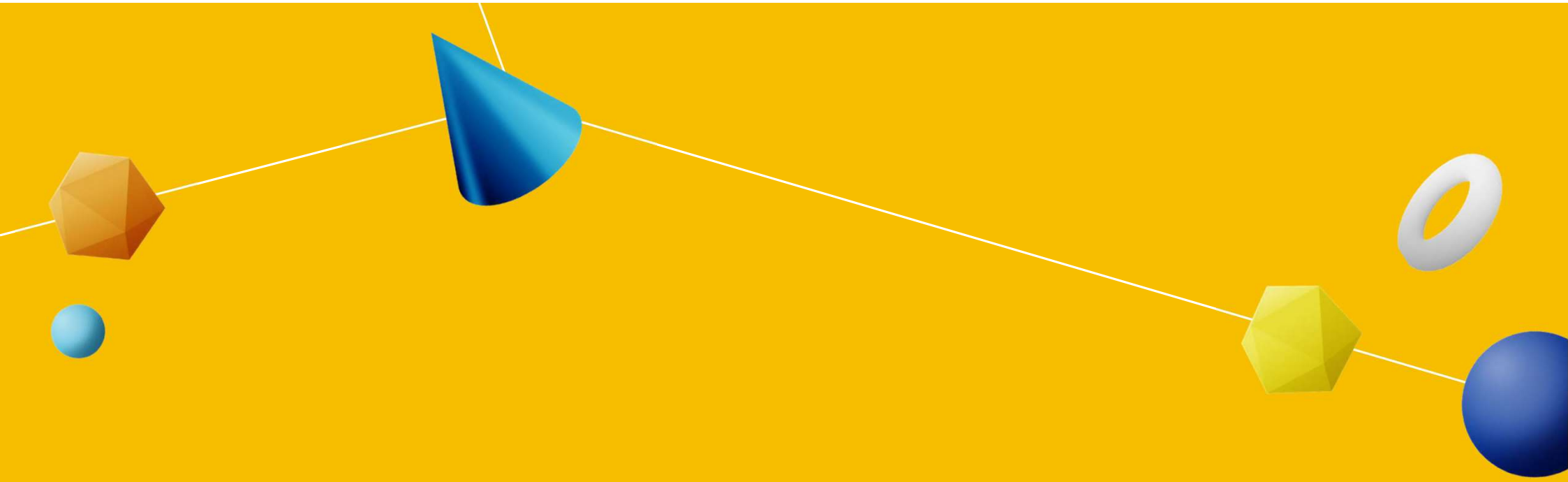
Objetivos de modelado

• Caracterizar, prevenir y detectar anomalías en distintas capas de análisis.

- •Análisis Descriptivo
- •Análisis Proactivo
- •Análisis Predictivo

INSTITUTO IBERMÁTICA DE INNOVACIÓN

Ibermática



CALIDAD DEL DATO

Ibermática

CALIDAD

METODOLOGÍA DE CALIDAD DEL DATO (Pendiente de rellenar por Patxi).

Integridad

COMPLETITUD

Conteo de registros y comparativa con registros esperados

CONSISTENCIA

Tipo de distribución de los datos

PRECISIÓN

EXACTITUD

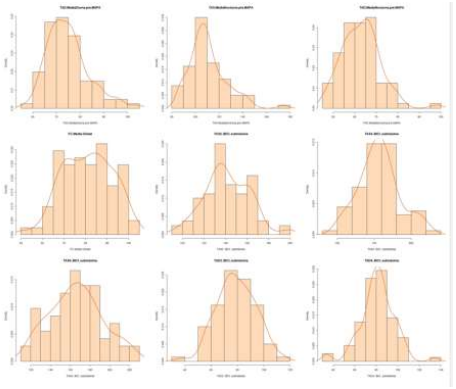
Desviación típica / Media cercana al cero

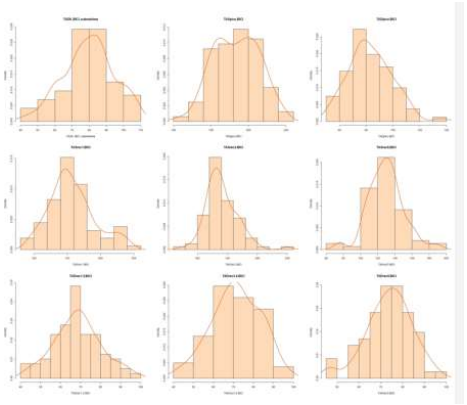
ATÍPICOS

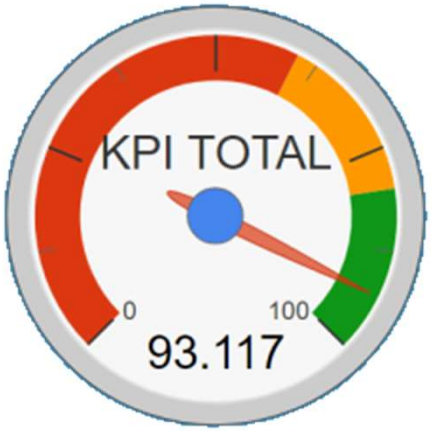
Existen outliers o datos anómalos?

EXTREMOS

Existen valores extremos en los datos?



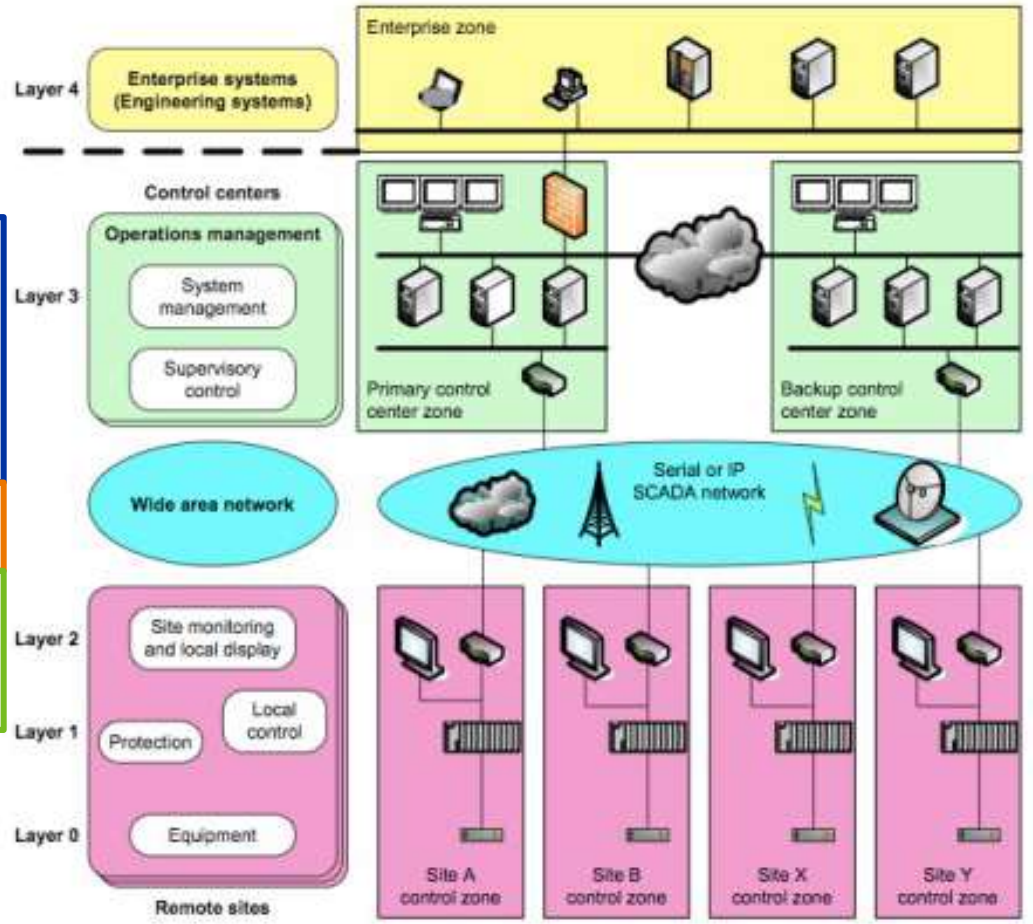




INSTITUTO IBERMÁTICA DE INNOVACIÓN

CONTEXTUALIZACIÓN DE DATOS
ESPECIFICACIONES ESTÁNDAR (ISO99)

- Arquitectura Industrial ISA-99
- Los niveles del proceso industrial determinan la segmentación.
- Filtrado muy básico
- Tres zonas principales.



INSTITUTO IBERMÁTICA DE INNOVACIÓN

CONTEXTUALIZACIÓN DE DATOS

ESPECIFICACIONES ESTÁNDAR (ISO99) . Nuestro Objetivo de Minería.

Zona 1

Zona 2

Zona 3

PROTOCOLO		CIFRADO	AUTENTICACION	NIVEL IP / TRANSPORTE	NIVEL APLICACIÓN	SEGURIDAD Y RECOMENDACIONES
Common Industrial Protocol (CIP)	DeviceNet	NO	NO	Protocolo propietario ***DeviceNet™***	CIP	CIP cuenta con la tecnología a nivel de aplicación ***CIP Safety™*** . Es recomendable complementar con medidas generales de segmentación y aislamiento de las redes de control
	ControlNet	NO	NO	Protocolo propietario ***ControlNet™***		
	Componet	NO	NO	Protocolo propietario ***Componet™***		
	Ethernet/IP	NO	NO	TCP/IP		
MODBUS	Modbus Serie	NO	NO	NO APLICA (transmisión serie)	Modbus (no seguro)	Adoptar, si es posible, cifrado (SSL,VPN) o medidas de inspección de tráfico como IDS (Snort), IPS (Tofino), etc.
	Modbus TCP	NO	NO	TCP/IP	Modbus (no seguro)	
DNP3		SOLO DNP Secure	SOLO DNP Secure	DNP Secure	DNP Secure	Se recomienda la implementación de DNP Secure
Profibus		NO	NO	NO APLICA (transmisión serie)	Sin medidas de seguridad propias	Aplicar las recomendaciones generales propuestas de segmentación, inspección de tráfico y cifrado.
Profinet		NO	NO	TCP/IP UDP/IP	Sin medidas de seguridad propias	Profinet Security Guide
Powerlink Ethernet		NO	NO	No incorpora medidas propias	No incorpora medidas propias	PowerLink es un protocolo basado en Ethernet y orientado a control en tiempo real. Se recomiendan medidas de segmentación en la arquitectura.
OPC		OPC UA	OPC UA	Base TCP/IP OPC UA	OPC UA	Implementar OPC UA
EtherCAT		NO	NO	No incorpora medidas propias	No incorpora medidas propias	Se recomiendan medidas de segmentación y seguridad perimetral

CONTEXTUALIZACIÓN DE DATOS

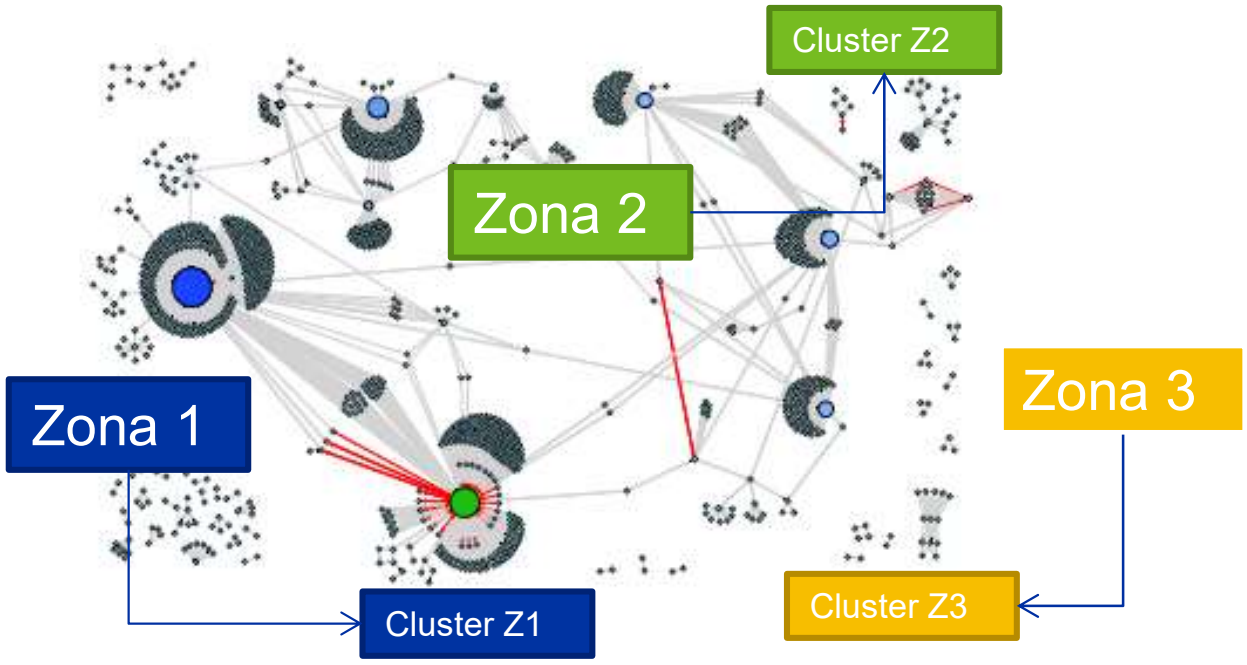
Objetivo de negocio a conseguir:

1. **Identificar de forma automática los tres niveles.**
2. **Caracterizar los patrones típicos y atípicos de cada nivel de forma independiente.**
3. **Caracterizar los patrones típicos y atípicos de los niveles de forma combinada.**

Zona 1 Ethernet

Zona 2 TCP/ IP

Zona 3 OPCUA

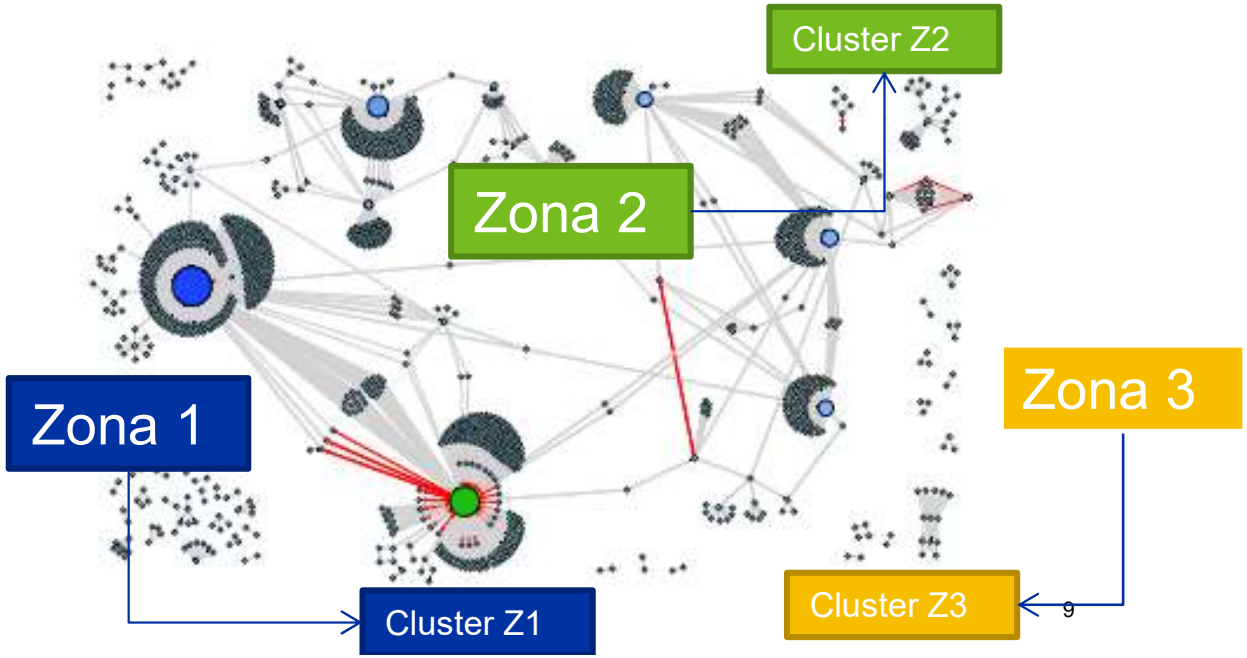


INSTITUTO IBERMÁTICA DE INNOVACIÓN

OPORTUNIDAD DEL PROYECTO

¿Para qué?

Tener una ***herramienta*** que nos permita dividir la información de origen en ***niveles independientes*** para ***optimizar el análisis de la información*** en una estrategia ***top-down*** (de menor a mayor especificidad). Esto permite ***detectar patrones anormales*** en distintos niveles de información de forma ***paralela, independiente, y proactiva***, evitando "obviar" comportamientos muy específicos que se "ocultarían" en modelos más generalistas.



Ventajas:

- Conocimiento de la estructura de datos ***independientemente del dominio experto***
- Nos basamos en la ***hipótesis*** que una normalización matemática del espacio vectorial va a mejorar la disminución de la entropía y, por lo tanto, la mejora en la detección
- Aplicable a cualquier contexto en ciberseguridad, independientemente de protocolos, servicios, redes o sistemas de monitorización (modbus, scada, MTU, RTU, etc…)

INSTITUTO IBERMÁTICA DE INNOVACIÓN

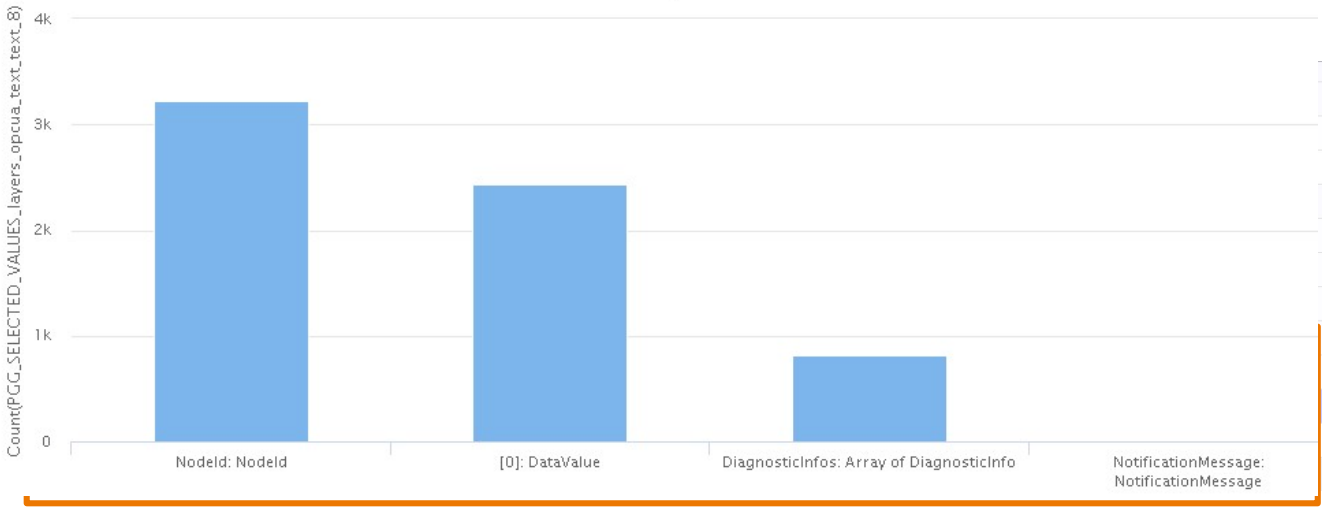
MODELOS ANALÍTICOS

Ibermática

MODELOS EXPLORATORIO
MATRIZ DE COVARIANZA

¿Existe alguna forma matemática que diferencie los distintos niveles de información de una manera estadísticamente significativa?

Attributes	PGG_SE...	PGG_SE...	PGG_SE...	PGG_SE...	PGG_SE...	PGG_SE...
PGG_SELECTED_VALUES_layers_eth_eth_dst_eth_addr	1.000	-0.329	0.041	0.850	-0.876	-0.492
PGG_SELECTED_VALUES_layers_opcua_text_text_10	-0.329	1.000	-0.004	-0.053	0.285	0.166
PGG_SELECTED_VALUES_layers_opcua_text_text_11	0.041	-0.004	1.000	-0.137	-0.066	0.018
PGG_SELECTED_VALUES_layers_opcua_text_text_8	0.850	-0.053	-0.137	1.000	-0.730	-0.450
PGG_SELECTED_VALUES_layers_ip_ip_ip_checksum	-0.876	0.285	-0.066	-0.730	1.000	0.826
PGG_SELECTED_VALUES_layers_ip_ip_ip_id	-0.492	0.166	0.018	-0.450	0.826	1.000



$$K_{\mathbf{XX}} = \operatorname{cov}[\mathbf{X}, \mathbf{X}] = \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^{\mathrm{T}}] = \mathrm{E}[\mathbf{X}\mathbf{X}^{T}] - \mu_{\mathbf{X}}\mu_{\mathbf{X}}{}^{T} \quad \text{(Eq.1)}$$

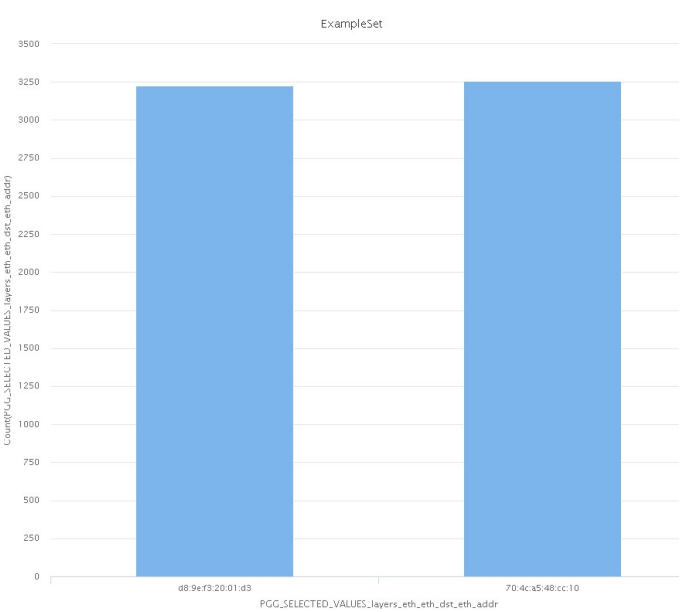
where $\mu_{\mathbf{X}} = \mathrm{E}[\mathbf{X}]$.

- La matriz de covarianza es una herramienta a partir de la cual se puede encontrar una base matemática para representar los datos de forma óptima
- Se aprecia agrupamiento de varianzas entre distintos grupos de atributos (jerarquías).
- La matriz de covarianzas nos da una representación visual, no un modelizado matemático
- La matriz de covarianzas no nos suministra información al respecto de la dependencia de niveles.

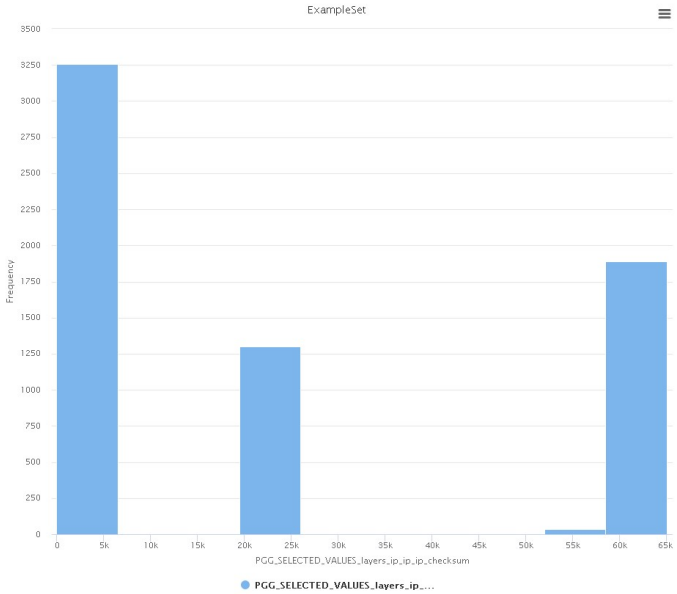
INSTITUTO IBERMÁTICA DE INNOVACIÓN

Variables Representativas y su información semántica

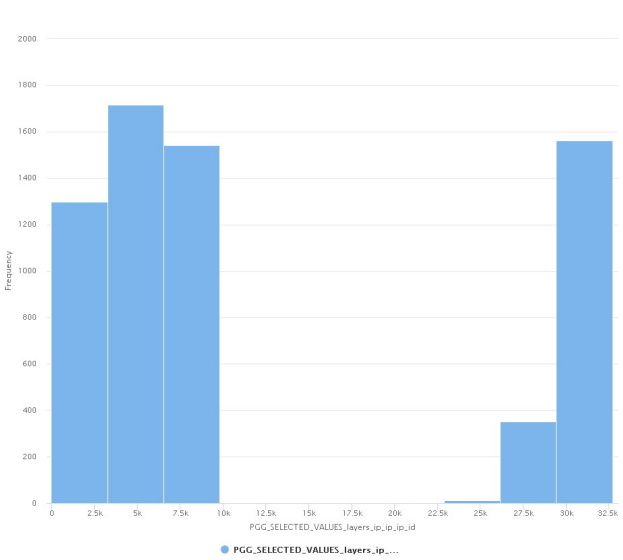
Contexto ETH, IP



PGG_SELECTED_VALUES_layers
_eth_eth_dst_eth_addr

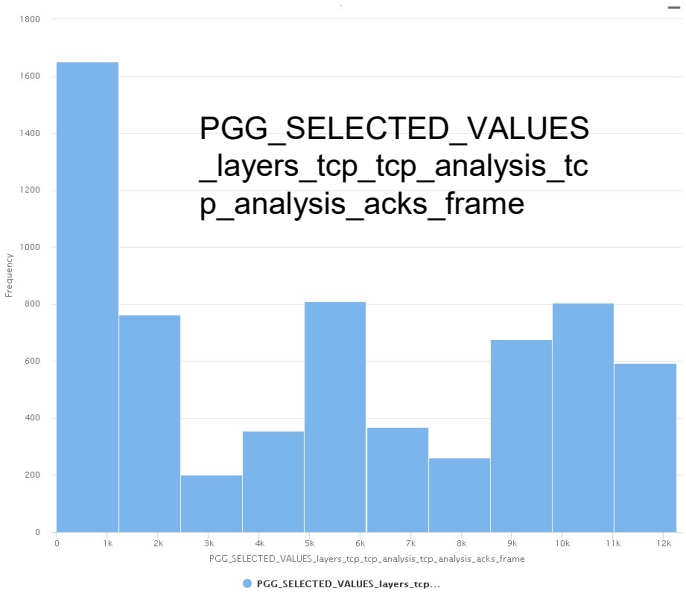


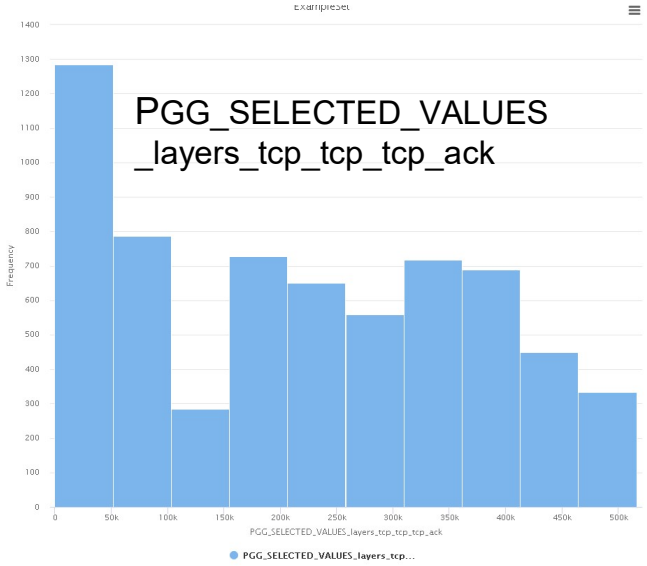
PGG_SELECTED_VALUES_layers_ip_ip_ip_checksum

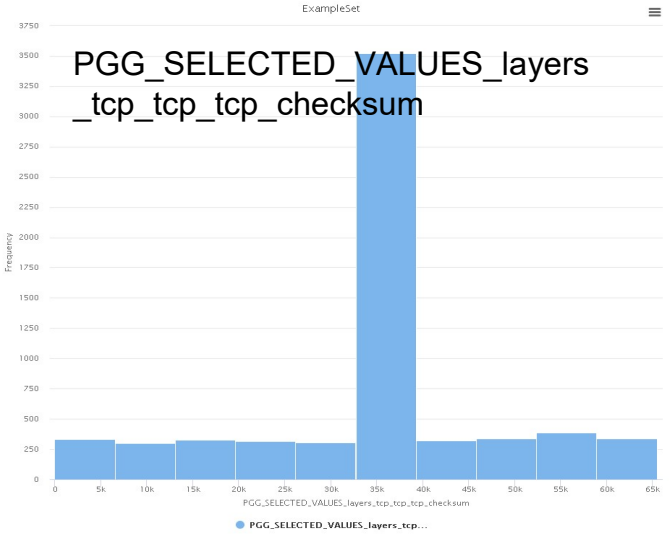


PGG_SELECTED_VALUES_layers_ip_ip_ip_id

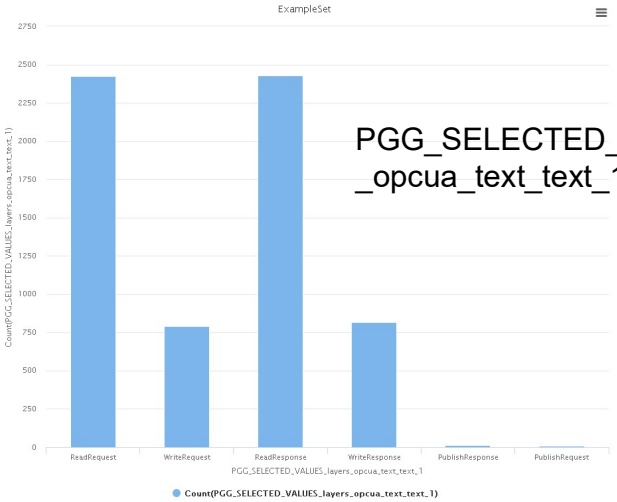
Variables Representativas y su información semántica
Contexto TCP



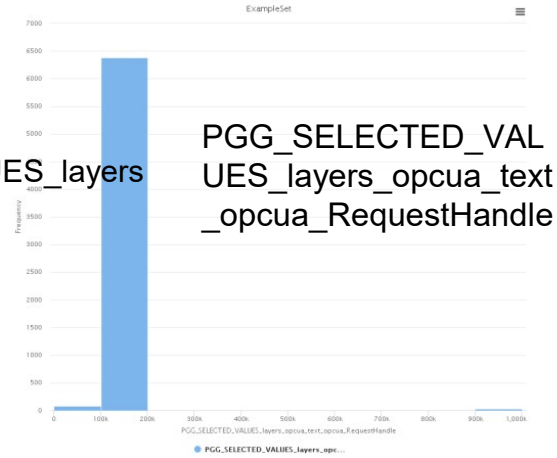




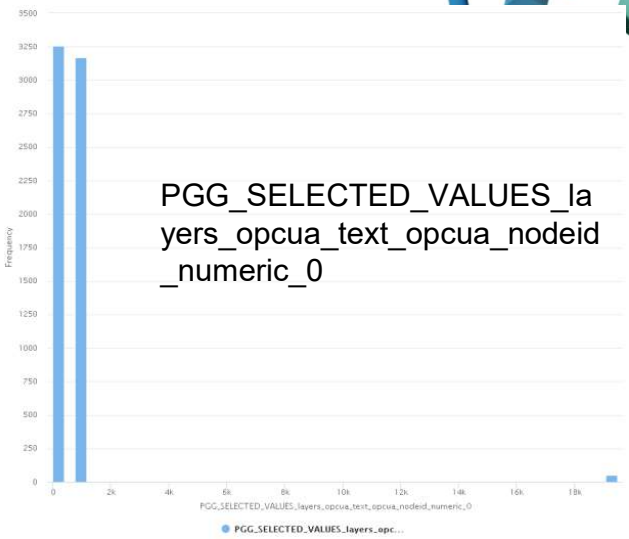
Variables Representativas y su información semántica Contexto OPCUA



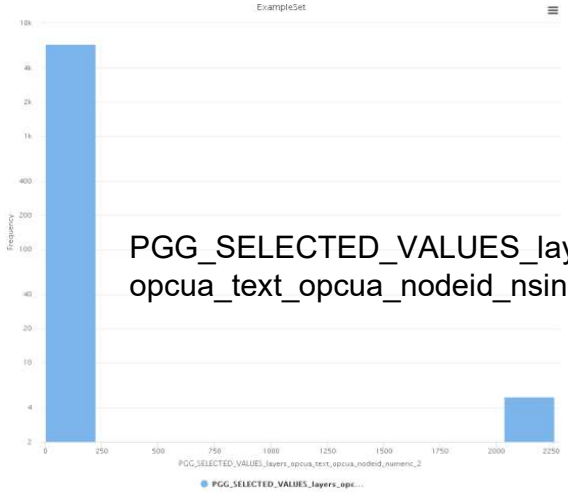
PGG_SELECTED_VALUES_layers_opcua_text_text_1



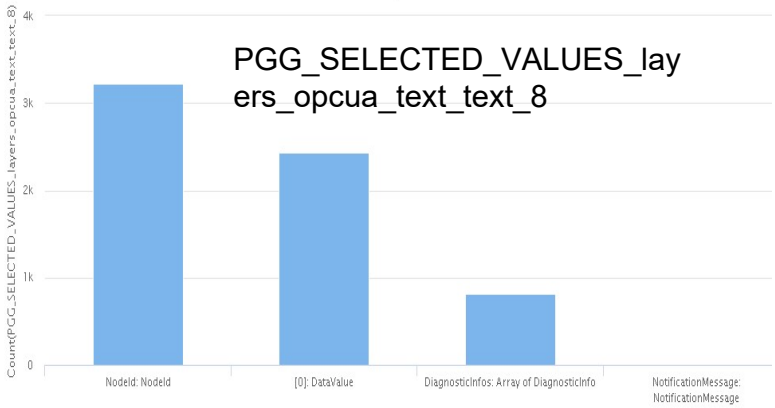
PGG_SELECTED_VALUES_layers_opcua_text_opcua_RequestHandle



PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_numeric_0

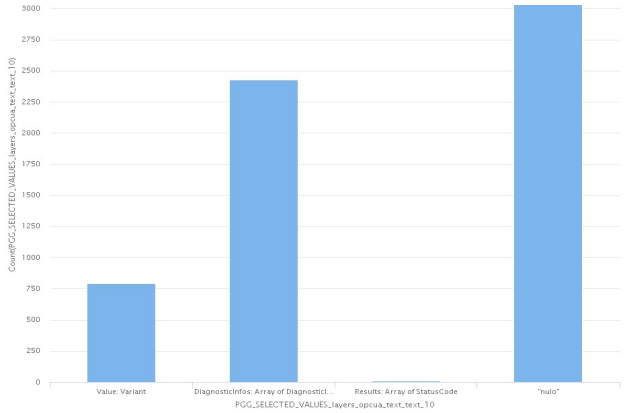


PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_nsindex_1



PGG_SELECTED_VALUES_layers_opcua_text_text_8

PGG_SELECTED_VALUES_layers_opcua_text_text_10



Métodos y Materiales

De la Visual a la Matemática: Transformar Cov(X) a una Ontología

Covariance-based Clustering

Assumption 1 *We assume that observations drawn from families P_X and P_Y constituting the data set D may be distinguished on the basis of their covariances, but not of their means, i.e. $\mu_1^{(0)} = \mu_2^{(0)}$ and $\mathcal{C}_1^{(0)} \neq \mathcal{C}_2^{(0)}$, and therefore $\|\mu_1^{(0)} - \mu_2^{(0)}\| = 0$ and $d(\mathcal{C}_1^{(0)}, \mathcal{C}_2^{(0)}) \gg 0$.*

$$\widehat{\mathcal{C}}_1^{(i)} = \frac{\sum_{j=1}^{K} w_{j,1} X_j \otimes X_j + \sum_{k=1}^{K} v_{k,1} Y_k \otimes Y_k}{K},$$

$$S = \frac{1}{N-1} \sum_{i=1}^{n} (\mathbf{x} - \bar{\mathbf{x}}_N)(\mathbf{x} - \bar{\mathbf{x}}_N)^\top$$

Cov. Matrix K_{XX} → Kmeans → Hierarchical clustering (Bottom-up) → Flatten

$$\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

Correlation clustering

We here propose a method to clusterize data streams, using a *local metric* based on the SU (Symmetrical Uncertainty) distance of the data points from each cluster center **ci, computed using an estimator of the covariance** matrix of the corresponding ith cluster. In the following we will always represent vectors as column vectors and we will assume that data are vectors in Rp.

$$\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^p, N \geq 2$$

$$\hat{S}_P = \frac{n_{h_1} \hat{S}_{h_1} + n_{h_2} \hat{S}_{h_2} + \cdots n_{h_K} \hat{S}_{h_K}}{n_{h_1} + n_{h_2} + \cdots + n_{h_K}},$$

Kmeans → Corr. Matrix K_{XX} → Hierarchical clustering (Bottom-up) → Flatten

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|} \quad SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X) + H(Y)}\right] \quad \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

Métodos y Materiales

De la Visual a la Matemática: Transformar Cov(X) a una Ontología

Covariance-based Clustering

- root
  - cluster_0
    - PGG_SELECTED_VALUES_layers_opcua_text_text_10
    - PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_checksum
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_servicenodeid_numeric
    - PGG_SELECTED_VALUES_layers_opcua_text_text_11
    - PGG_SELECTED_VALUES_layers_tcp_tcp_analysis_tcp_analysis_ack_rtt
  - cluster_1
    - PGG_SELECTED_VALUES_layers_opcua_opcua_opcua_security_rqid
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_numeric_2
    - PGG_SELECTED_VALUES_layers_tcp_text_tcp_time_delta
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_TimeoutHint
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_numeric_0
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_RequestHandle
  - cluster_2
    - PGG_SELECTED_VALUES_layers_tcp_tcp_analysis_tcp_analysis_acks_frame
    - PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_ack
    - PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_nxtseq
  - cluster_3
    - PGG_SELECTED_VALUES_layers_eth_eth_dst_eth_addr
    - PGG_SELECTED_VALUES_layers_ip_ip_ip_addr_0
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_nsindex_1
    - PGG_SELECTED_VALUES_layers_opcua_text_text_1
    - PGG_SELECTED_VALUES_layers_opcua_text_text_8
  - cluster_4
    - PGG_SELECTED_VALUES_layers_ip_ip_ip_checksum
    - PGG_SELECTED_VALUES_layers_ip_ip_ip_id

OPCUA

TCP

ETH

IP

Correlation clustering

- root
  - cluster_0
    - PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_nsindex_1
    - PGG_SELECTED_VALUES_layers_eth_eth_dst_eth_addr
    - PGG_SELECTED_VALUES_layers_opcua_text_text_10
  - cluster_1
    - PGG_SELECTED_VALUES_layers_opcua_text_text_1
    - PGG_SELECTED_VALUES_layers_opcua_text_text_8
  - cluster_2
    - PGG_SELECTED_VALUES_layers_tcp_tcp_analysis_tcp_analysis_acks_frame
    - PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_nxtseq
    - PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_ack
  - cluster_3
    - PGG_SELECTED_VALUES_layers_ip_ip_ip_checksum
  - cluster_4
    - PGG_SELECTED_VALUES_layers_ip_ip_ip_id

ETH

OPCUA

TCP

IP

Métodos y Materiales

De la Visual a la Matemática: Validación

¿Cómo medir qué distribución de clusters es la mejor?

The Gini coefficient provides a quick and intuitive way to evaluate the degree of the heterogeneity of the collection of clusters, which is useful to explain how well the cluster collection reveal the underlying true cluster patterns. The value of the Gini coefficient is between 0 and 1, with higher values indicating higher disparity in the clusters (a low Gini coefficient indicates a more equal distribution).

$$G = \sum_{i=1}^{I+1} (n_i c_{i-1} - n_{i-1} c_i)$$

Method	Gini Coefficient
Covariance-based Clustering	0.952 (**)
Correlation clustering	**0.867 (***)**
Taxonomía Manual (clustering)	**0.999 (*)**

Modelo de control:
Taxonomía Manual
•Nivel 1:.*tcp.*|.*eth.*
•Nivel 2: .*ip.*
•Nivel 3: .*opcua.*

Métodos y Materiales

De la Visual a la Matemática: Agrupamiento en niveles

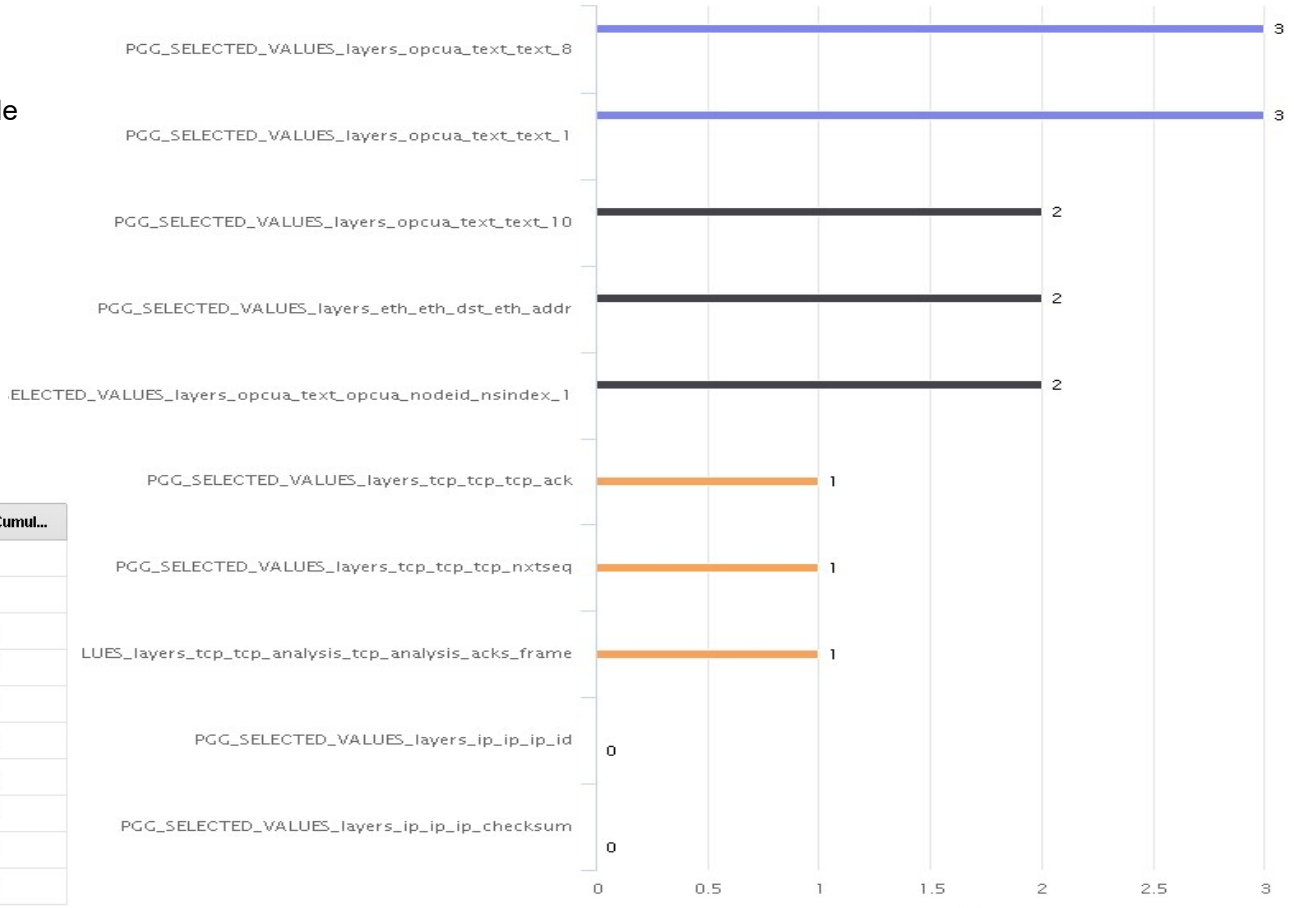
¿Cómo medir qué clusters son más generales o más específicos?

Principal component analysis (**PCA**): the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible).

"Los primeros componentes principales describen la mayor parte de la varianza de los datos (más cuanto más correlacionadas estuvieran las variables originales). Estos componentes de bajo orden a veces contienen el aspecto "más importante" de la información"

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\hat{\mathbf{X}}_k \mathbf{w}\|^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^T \hat{\mathbf{X}}_k^T \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

Attribute	Nivel	cluster	sum(Cumul...
PGG_SELECTED_VALUES_layers_ip_ip_ip_checksum	0	cluster_3	1
PGG_SELECTED_VALUES_layers_ip_ip_ip_id	0	cluster_4	1
PGG_SELECTED_VALUES_layers_tcp_tcp_analysis_tcp_analysis_acks_frame	1	cluster_2	2.864
PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_nxtseq	1	cluster_2	2.864
PGG_SELECTED_VALUES_layers_tcp_tcp_tcp_ack	1	cluster_2	2.864
PGG_SELECTED_VALUES_layers_opcua_text_opcua_nodeid_nsindex_1	2	cluster_0	2.685
PGG_SELECTED_VALUES_layers_eth_eth_dst_eth_addr	2	cluster_0	2.685
PGG_SELECTED_VALUES_layers_opcua_text_text_10	2	cluster_0	2.685
PGG_SELECTED_VALUES_layers_opcua_text_text_1	3	cluster_1	1.974
PGG_SELECTED_VALUES_layers_opcua_text_text_8	3	cluster_1	1.974



Métodos y Materiales

De la Visual a la Matemática: Validación

¿Cuál es la estrategia que mejor predice la segmentación con los datos originales?

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees

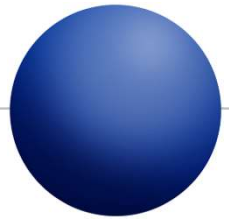
$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x), \quad \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma).$$

Method	Gini Coefficient	GBT
Covariance-based Clustering	0.952 (**)	0.9974 (**)
Correlation clustering	0.867 (***)	0.9977 (***)
Taxonomía Manual (clustering)	0.999 (*)	0.9873 (*)

Modelo de control:
Taxonomía Manual
- Nivel 1:.*tcp.*|.*eth.*
- Nivel 2: .*ip.*
- Nivel 3: .*opcua.*

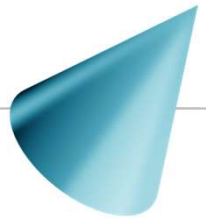
MODELOS ANALÍTICOS

METODOLOGÍA DE ANALÍTICA AVANZADA



SEGMENTACIÓN K-MEANS

- Segmentación de todos los datos en grupos similares.
- Automática y no supervisada
- Formación de grupos homogéneos de datos

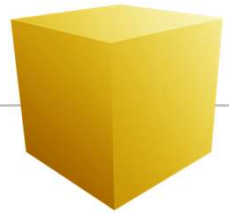


PRINCIPAL COMPONENT ANALYSIS

- Transformación del sistema de coordenadas original.
- Determinación de la varianza de los datos.

GRADIENT BOOSTED TREES

- Obtención de la importancia de las variables en las segmentaciones.
- Validación de la clasificación de los segmentes.



MATRIZ DE COVARIANZA

- Visualización de variables como variables distintas se comportan de forma parecida.
- Sirve para encontrar comportamientos similares en las variables

HIERARCHICAL CLUSTERING

- Transforma matrices de clusters, covarianzas o correlaciones en estructuras en árbol

INSTITUTO IBERMÁTICA DE INNOVACIÓN

CONCLUSIONES Y SIGUIENTES PASOS

Ibermática

CONCLUSIONES

Hipótesis Inicial

•Desarrollar un método matemático que identifique las variables principales que “desdoblen” en distintos niveles jerárquicos la información de tramas de origen.

Métodos

•Chequear los resultados automáticos con un modelo de estratificación “heurístico”, confrontándolo a un evaluador analítico.

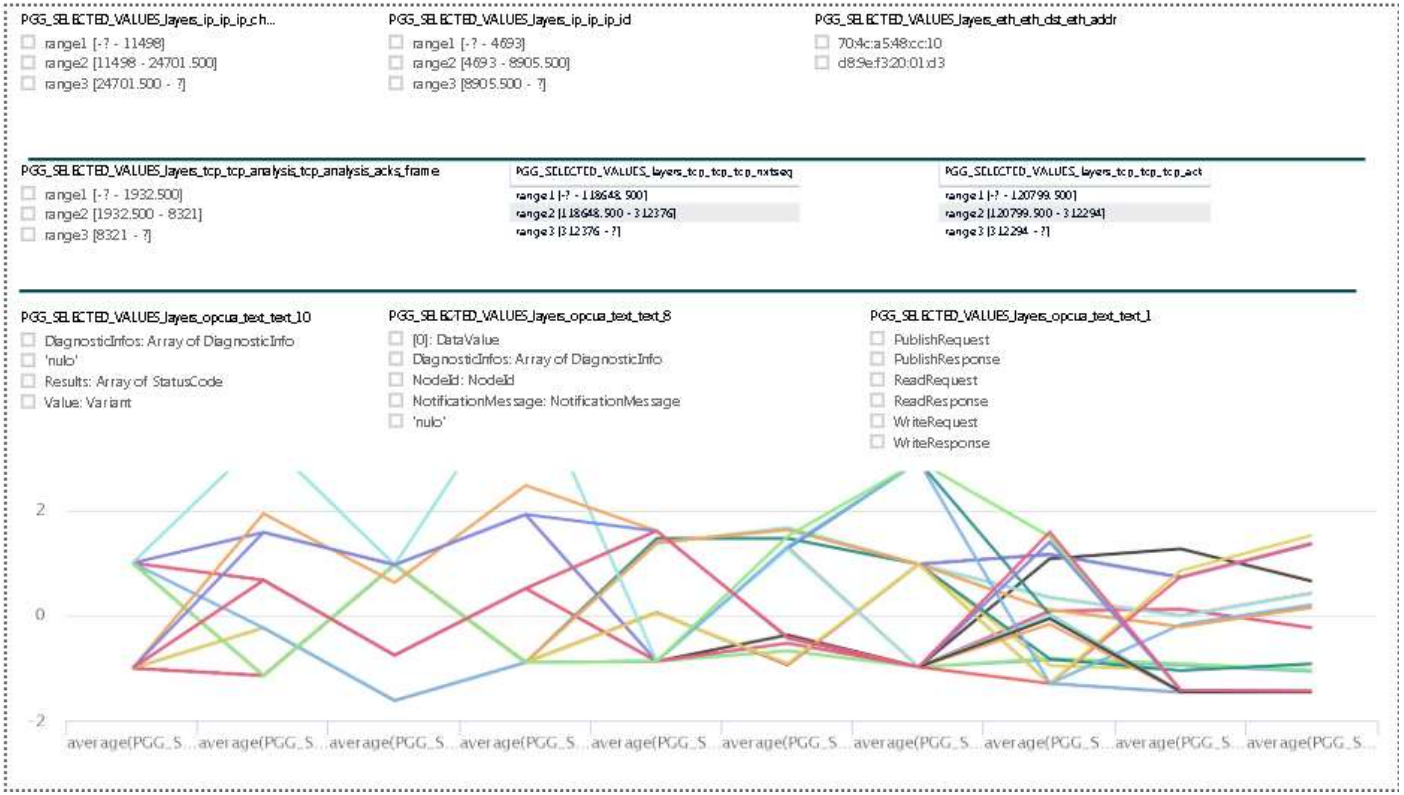
Discusión

• EL modelo heurístico de jerarquización manual da muy buenos resultados, lo que demuestra que el conocimiento experto es válido.

•Se han realizado dos aproximaciones matemáticas en función de las distribuciones de la covarianza y la medias, dando mejor resultado el análisis por distribución de las medias.

•Ambos métodos automática mejoran los resultados heurísticos tanto en adherencia a los perfiles como en su naturaleza predictiva (objetivo secuencia de clusters en niveles).

Propuesta de Plataforma



Siguientes Pasos

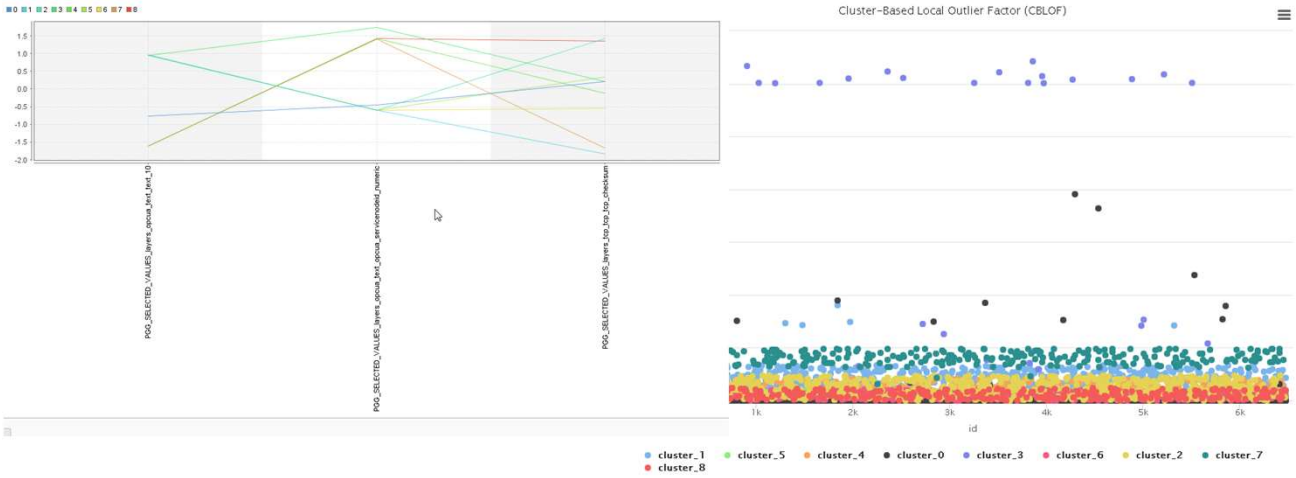
•Refactorizar el método “Correlation clustering” en spark.

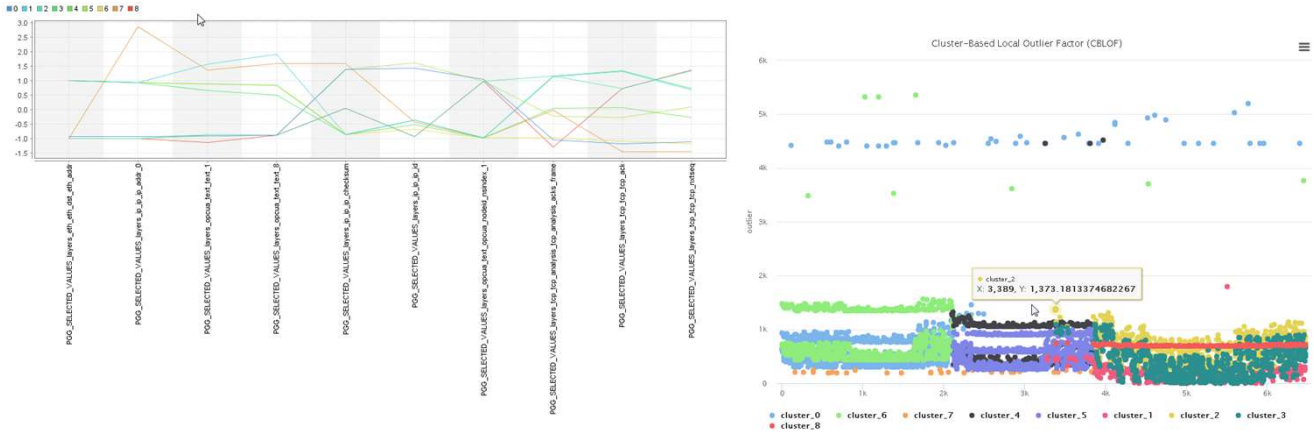
• Analizar métodos “econder” de DL en la mejora de la división entre niveles y la naturaleza predictiva.

•Validación final de los distintos métodos en base a la captura de anomalías y la potencia predictiva de cada método.

MODELOS ANALÍTICOS

Patrones típicos y atípicos de cada nivel de forma independiente





- Para cada nivel, Obtener una segmentación.
- Los centroides caracterízan las tramas típicas.
- Medición de la distancia de las tramas con respecto al centroide del segmento al que pertenecen.
- Determinación de outliers, que serían las tramas “extrañas” en cada nivel y que habría que analizar

INSTITUTO IBERMÁTICA DE INNOVACIÓN