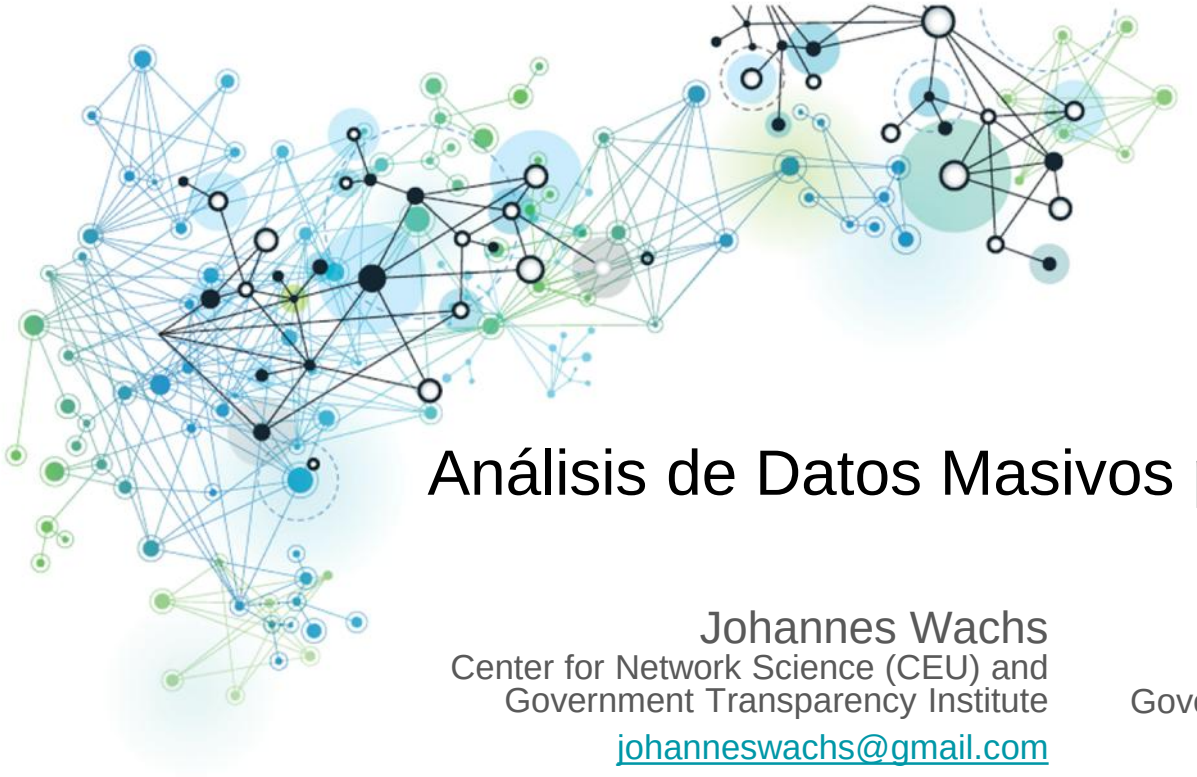


digiwhist

Análisis de Datos Masivos para el Sector Público

Johannes Wachs
Center for Network Science (CEU) and
Government Transparency Institute
johanneswachs@gmail.com
@johannes_wachs

Mihály Fazekas
University of Cambridge and
Government Transparency Institute
mf436@cam.ac.uk

Government Transparency Institute

CNS CEU

UNIVERSITY OF CAMBRIDGE
Department of Sociology

This project has received funding from the European Union's Horizon 2020 research and innovation Programme under grant agreement No 645852


2005
Luca Bruno / AP
2013
NBC NEWS
Michael Sohn / AP

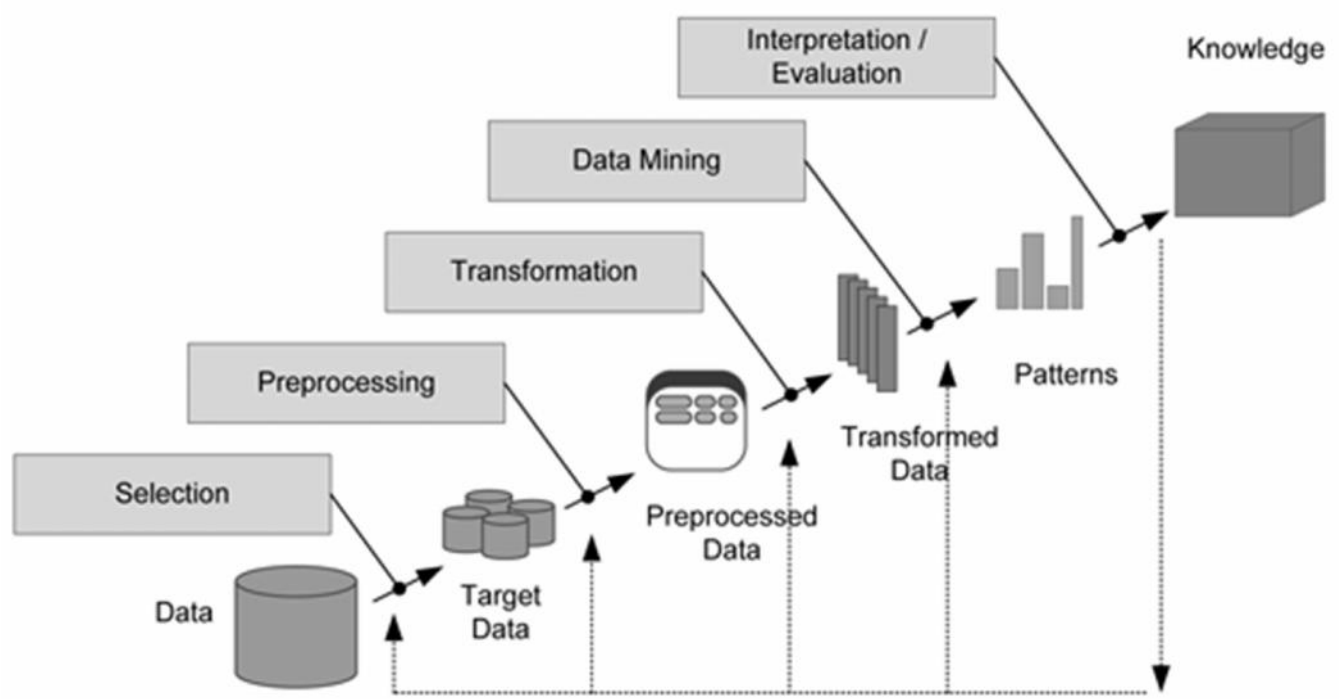
Descripción

- Principios del Análisis: El proceso KDD
- ¿Por qué KDD?
- Aplicación de KDD a Datos de Contratación Pública

Conocimiento Descubierto en Bases de Datos (KDD)

KDD: Metodología de cinco pasos para extraer información útil de los datos.

Conocimiento Descubierto en Bases de Datos (KDD)



IEEE

¿Por qué KDD?

KDD es más que una lista de verificación – es una forma para que nosotros organicemos nuestras ideas sobre la data, la cual se vuelve más y más difícil de manejar. ¿Cómo así?

¿Por qué KDD?

KDD es más que una lista de verificación – es una forma para que nosotros organicemos nuestras ideas sobre la data, la cual se vuelve más y más difícil de manejar. ¿Cómo así?

1. Tamaño: Datos de Contratos del Gobierno Federal de EEUU (sin texto) > 100 GB para 15 años de series de tiempo.

¿Por qué KDD?

KDD es más que una lista de verificación – es una forma para que nosotros organicemos nuestras ideas sobre la data, la cual se vuelve más y más difícil de manejar. ¿Cómo así?

1. Tamaño: Datos de Contratos del Gobierno Federal de EEUU (sin texto) > 100 GB para 15 años de series de tiempo.
2. Automatización: nos importa el análisis de datos entrantes.

¿Por qué KDD?

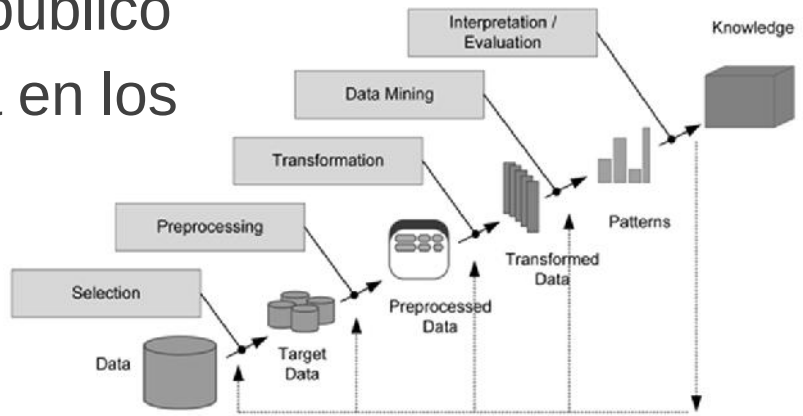
KDD es más que una lista de verificación – es una forma para que nosotros organicemos nuestras ideas sobre la data, la cual se vuelve más y más difícil de manejar. ¿Cómo así?

1. Tamaño: Datos de Contratos del Gobierno Federal de EEUU (sin texto) > 100 GB para 15 años de series de tiempo.
2. Automatización: nos importa el análisis de datos entrantes.
3. Descubrimiento: Las computadoras nos enseñan a encontrar nuevos patrones

KDD y Contratación Pública

Queremos mejorar los resultados públicos

1. Mejor costo por precio unitario
2. Mayor transparencia para el público
3. Mejoramiento de la confianza en los servicios del gobierno

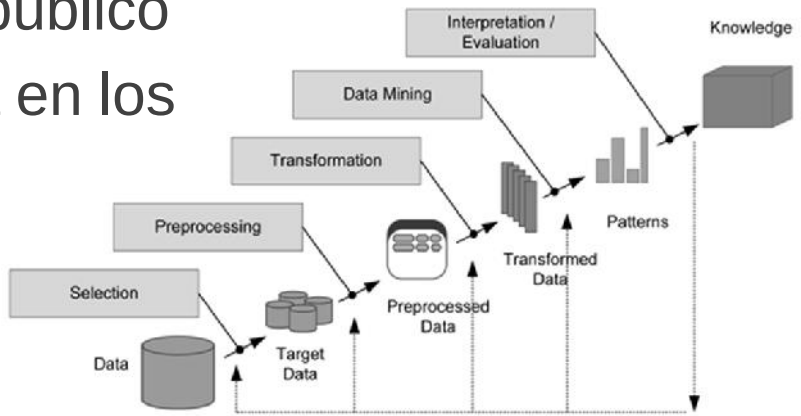


KDD y Contratación Pública

Queremos mejorar los resultados públicos

1. Mejor costo por precio unitario
2. Mayor transparencia para el público
3. Mejoramiento de la confianza en los servicios del gobierno

¿Cómo puede ayudar KDD?



Ejemplo: Mercado de la construcción Húngaro

Un funcionario del Servicio de Contratación Húngaro se topa con un contrato sospechoso: era muy costoso y solo una empresa ingresó su oferta. El quiere saber sobre los contratos pasados de la firma ganadora.

Afortunadamente, puede acceder a una bonita base de datos construida por los investigadores de Digiwhist. El funcionario puede revisar todos los contratos desde 2009 – 2014.

.

Ejemplo: Las primeras columnas

	year	issuer_town	issuer_postcode	issuer_type	sector	winner_id	issuer_id	log_value
0	2009	Budapest	1024.0	regionális/helyi szintü	construction work	12017340.0	735650	NaN
1	2009	Szeged	6728.0	közjogi szervezet	sewage-, refuse-, cleaning-, and environmental...	21629712.0	18455058	17.58222
2	2009	Harta	6326.0	regionális/helyi szintü	construction work	10553602.0	724463	13.57612
3	2009	PUTNOK	3630.0	regionális/helyi szintü	architectural, construction, engineering and i...	12736384.0	726159	15.62373

Ejemplo : Mercado de Construcción Húngaro

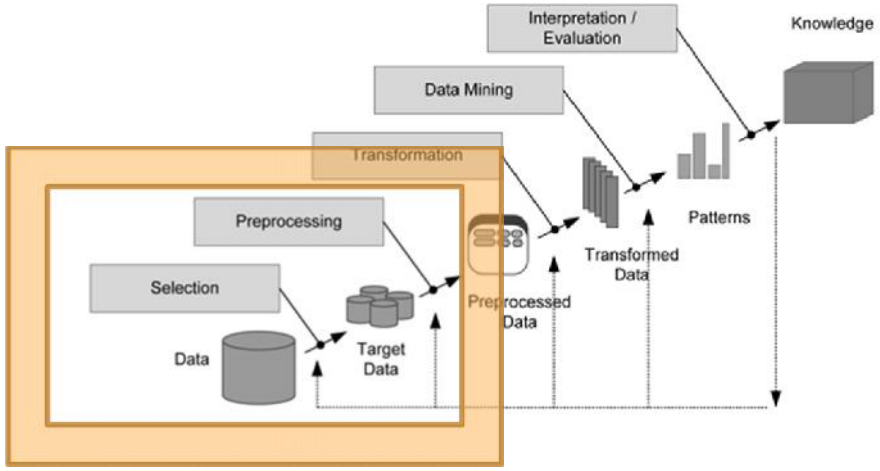
El funcionario sabe que la firma está involucrada en contratos de construcción. También sabe que la firma fue constituida en 2011.

Ejemplo: Filtros

El funcionario sabe que la firma está involucrada en contratos de construcción. También sabe que la firma fue constituida en 2011.

El busca contratos:

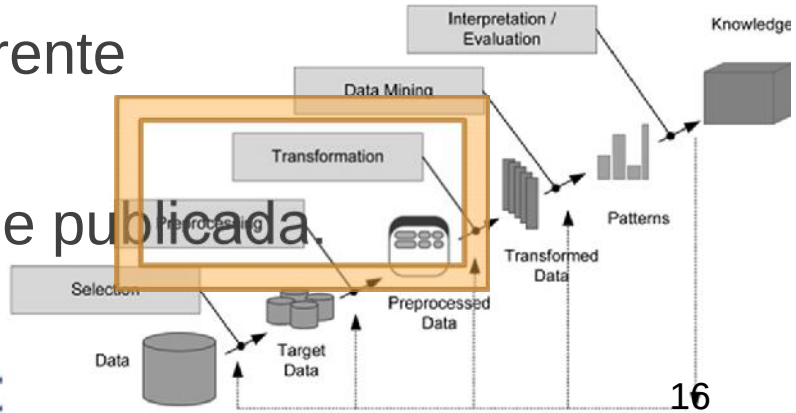
1. Desde 2011-2014
2. Construcción



Ejemplo: Procesamiento y Transformación

El quiere comparar los resultados de una firma con otras. El llena los valores faltantes con promedios y luego lo hace para todas las firmas en conjunto filtrado, y calcula,

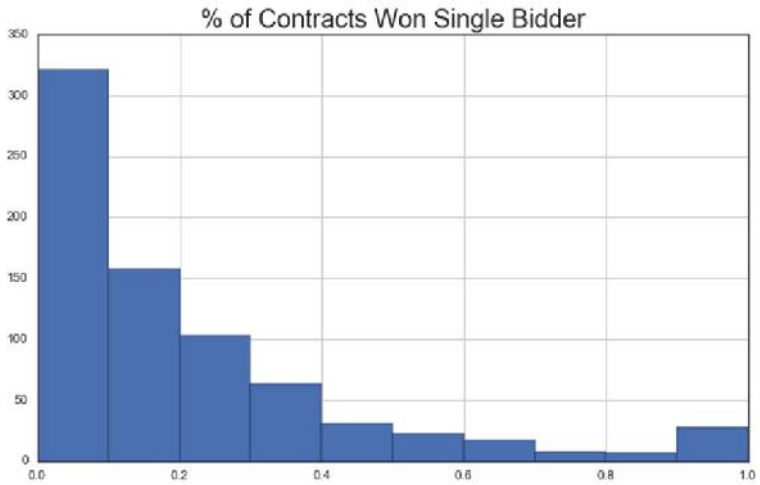
1. Cuántos contratos ganaron cada año (al menos 5)
2. Cuantos fueron de un solo oferente
3. El valor total que ganaron.
4. Si la convocatoria a subasta fue publicada.



digiwhist

Ejemplo: Visualización y Minería de Datos

El funcionario revisa las características a nivel de la empresa:



Ejemplo: Hipótesis

El realiza una hipótesis: si el número del porcentaje de contratos de un solo oferente ganados por esta firma es mucho más grande que el promedio, el funcionario realizará una investigación a más profundidad.

Mucho más grande que el promedio significa más de una desviación estándar sobre el promedio.

Ejemplo: Mercado de la Construcción Húngaro

La firma gana 40.6% de sus contratos como único oferente.

El promedio del mercado para los mismos años es de cerca del 20%, siendo una desviación estándar 42%. El índice de la firma de ganancias como único contratista es alta, pero no llega al punto que definimos.

Resultado: este comportamiento no está fuera de lo común.

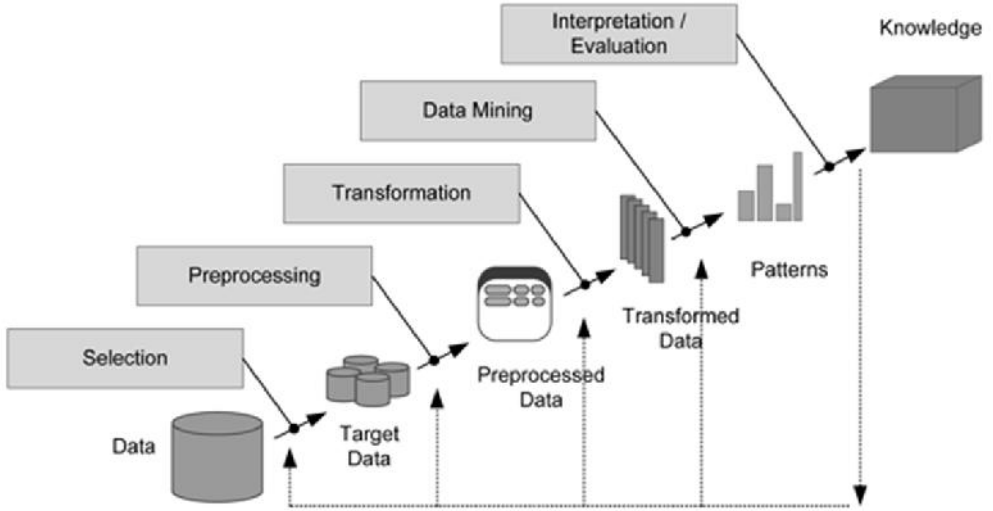
Ejemplo: Lecciones y Advertencias

- 1 desviación estándar es arbitraria. Construya su método de revisión con base en la teoría!
- Los valores absolutos si importan: un índice del 20% como oferente único es alto por si solo.
- Oferente único es solo un indicador. Un sistema robusto incluye mucho más.

Ejemplo: Resumen

Las flechas son importantes:

1. Anote la hipótesis antes de probarla.
2. Regrese a pasos anteriores antes de hacer cambios.



Conclusión

KDD nos proporciona una forma rigurosa de trabajar los datos.

Los resultados de un proceso KDD no son solamente una respuesta informada para nuestra pregunta original, sino más bien una fórmula para futuras investigaciones y replicación.

Considere al funcionario: si hubiera descubierto algo, su equipo podría haber transformado su procedimiento en una prueba automatizada estándar.

Gracias!

johanneswachs@gmail.com
twitter: johannes_wachs