



# **Data Science Foundations**

## **Unidad 3**

### **Metodología de ciencia de datos**

**Sesión Sincrónica 4,5,6**

# Comencemos con estas frases



**“Si tuviera sólo una hora para salvar el mundo, pasaría cincuenta y cinco minutos definiendo las preguntas, y sólo cinco minutos encontrando las respuestas”**

*Albert Einstein (se cree por lo menos)*

**“El valor de los datos esta en la historia que obtenemos a partir de ellos”**

*Luis Alcántara*

**“Los datos no son el fin sino el medio”**

*Giorgia Lupi*



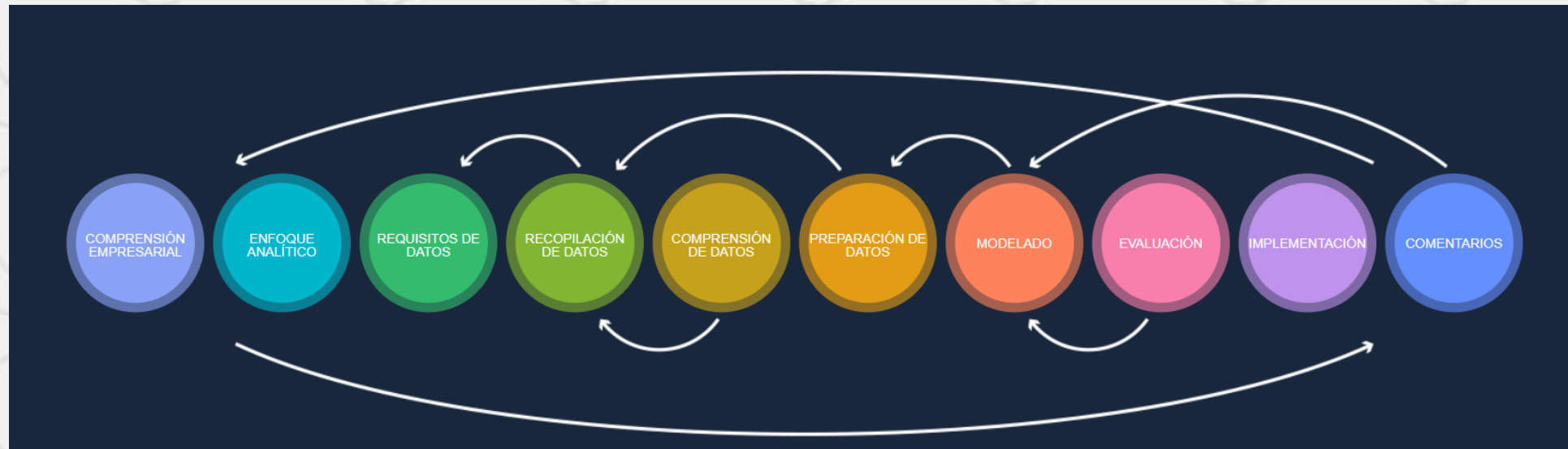
# Objetivos de aprendizaje



Al finalizar la Unidad el estudiante prepara un esquema de trabajo para el desarrollo de un proyecto de ciencia de datos indicando los entregables de cada etapa y proponiendo un modelo de ciencia de datos para solucionar un problema en entorno de negocios



# Metodología Data Science



<https://mym.cdn.laureate-media.com/2dett4d/IBM/LIDS0000/ES/infographic/index.html>

# Metodología de Data Science



Un sistema que garantiza que los datos utilizados en la resolución de problemas sean relevantes y se manipulen de forma adecuada a fin de abordar la pregunta de investigación.

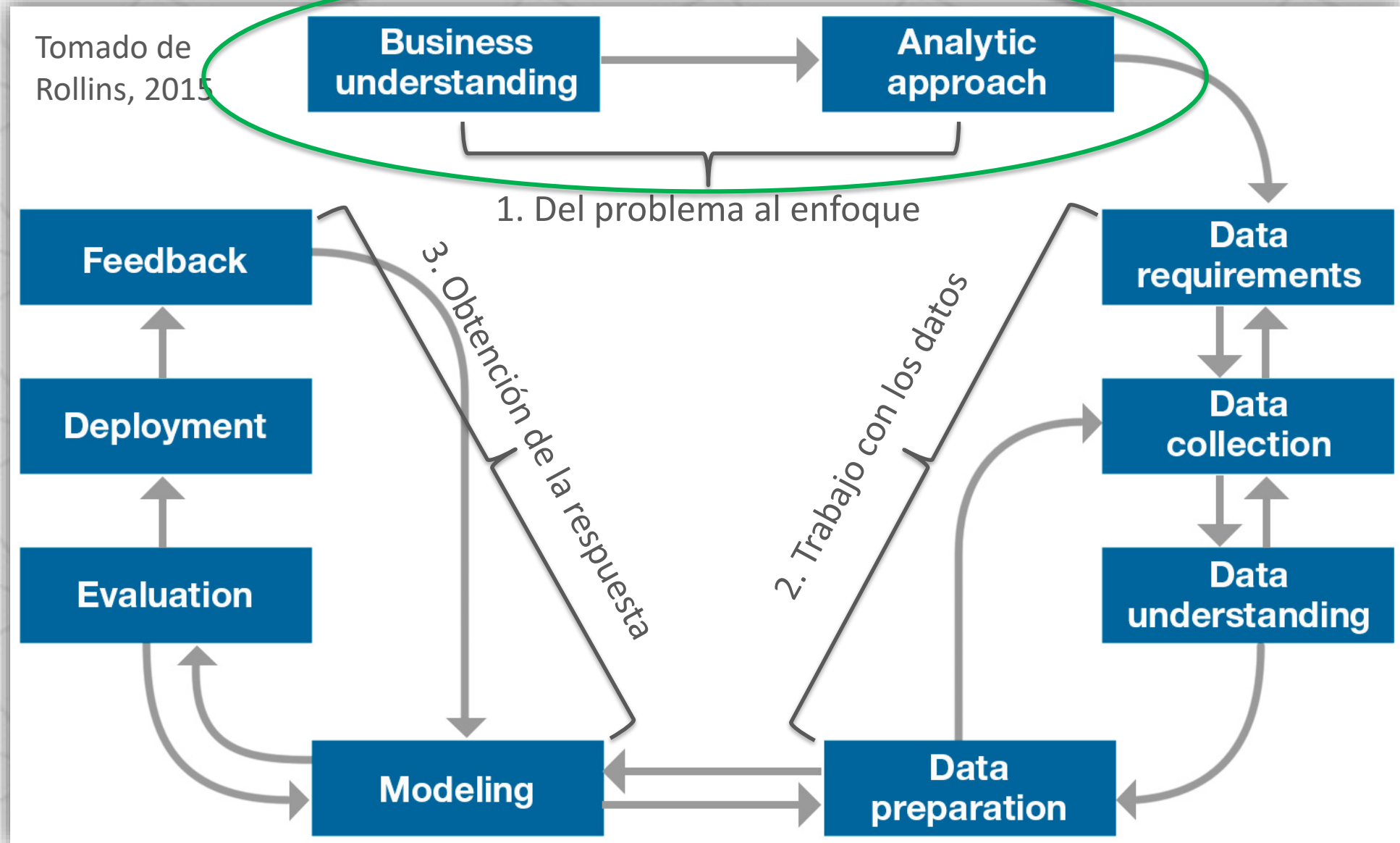
Se divide en tres etapas:

1. Del problema al enfoque
2. Trabajo con los datos
3. Obtención de la respuesta



John Rollins, PhD.  
Científico de datos,  
IBM Analytics, IBM

# Metodología Data Science



# Comprensión empresarial



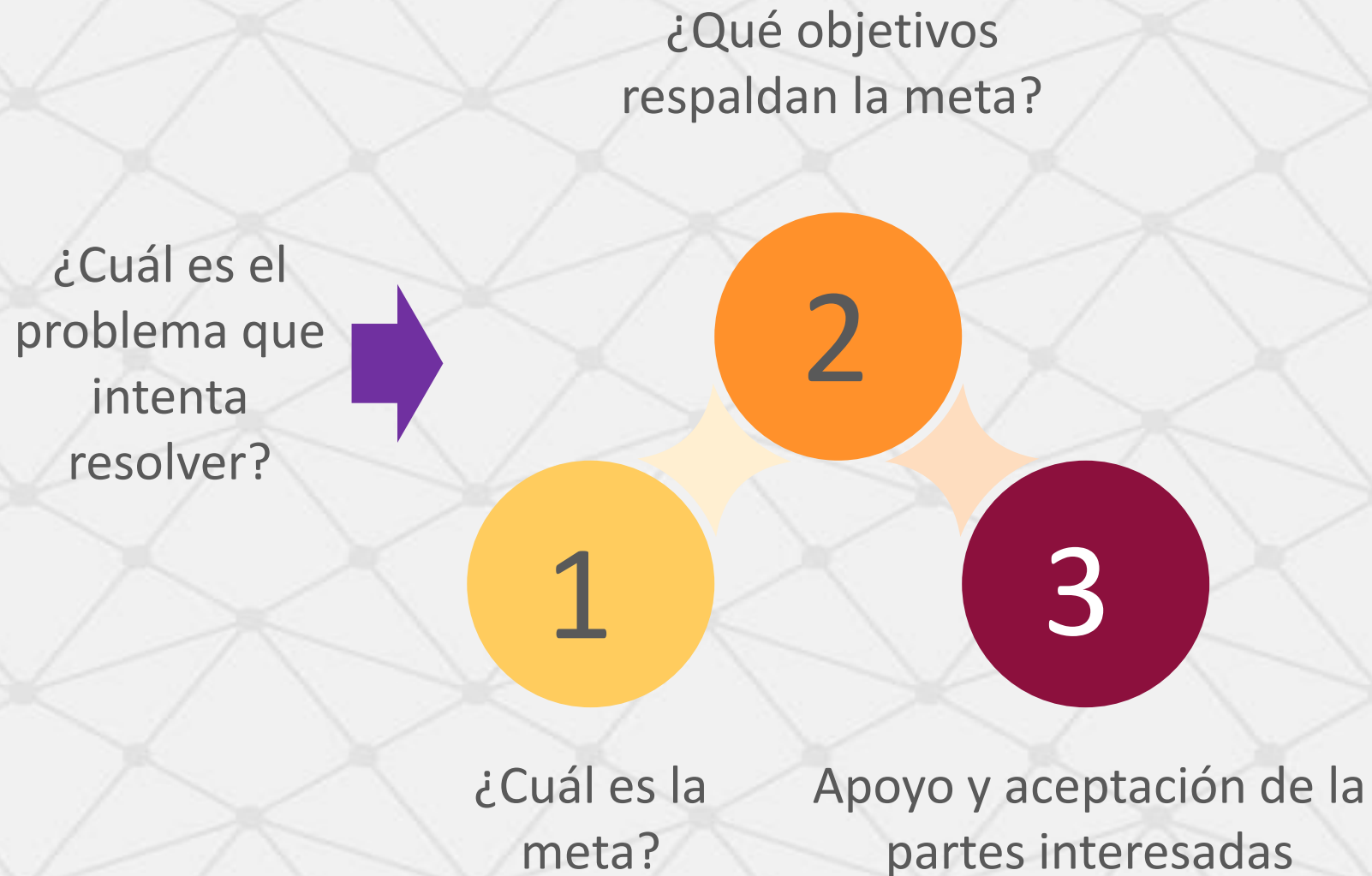


# ¿Porqué se debe entender los modelos de negocio?





# Comprensión empresarial





**Piense en una situación que usted considere un problema y en base a ello reflexione en lo siguiente:**

**¿Qué entienden por problema?**

**¿Porqué es un problema?**

**¿Qué técnicas utilizaría para entender el problema y ver posibles caminos para solucionarlo?**

# Entendiendo el problema a partir de las relaciones de causa - efecto



Redman, 2014



“En lugar de suponer una causalidad, trate las correlaciones como pistas que se deben combinar con otras evidencias para llegar a una conclusión. La correlación no es una causalidad. Pero es un excelente punto de partida”. Redman, 2014.



# Enfoque analítico



# Enfoques analíticos



¿Cómo puede  
usar los datos  
para  
responder la  
pregunta?



De diagnóstico



Predictivo



Descriptivo



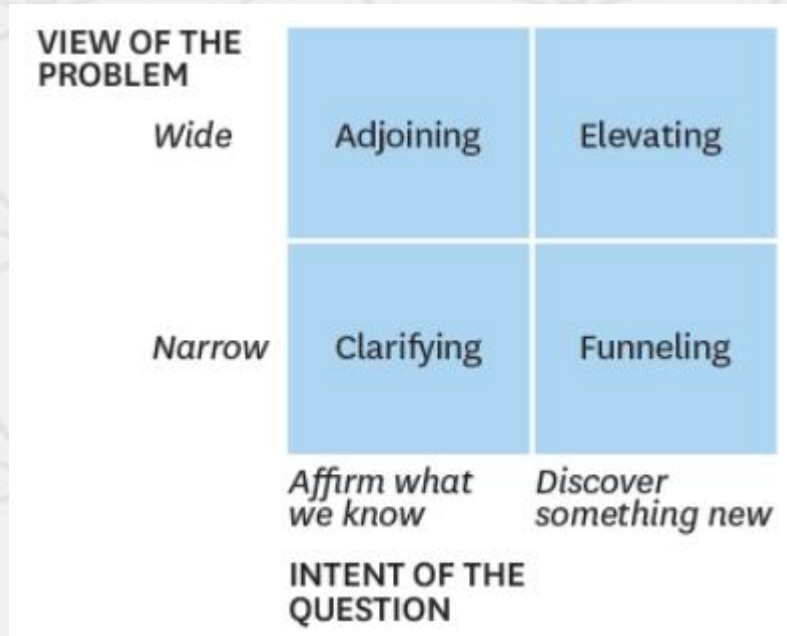
Prescriptivo

# Volvamos a los enfoques analíticos, pero veamos las preguntas





# El arte de hacer preguntas



Tomado de: Relearning the Art of Asking Questions. HBR. 2015  
Recuperado de: <https://hbr.org/2015/03/relearning-the-art-of-asking-questions>

## Explorando aspectos relacionados

- ¿Qué aspectos están relacionados con .....?

## Clarificando

- ¿Por qué esta sucediendo esto?

## Ampliando

- Ante lo que conocemos ¿cuáles son las tendencias más importantes que deberían preocuparnos?

## Profundizando

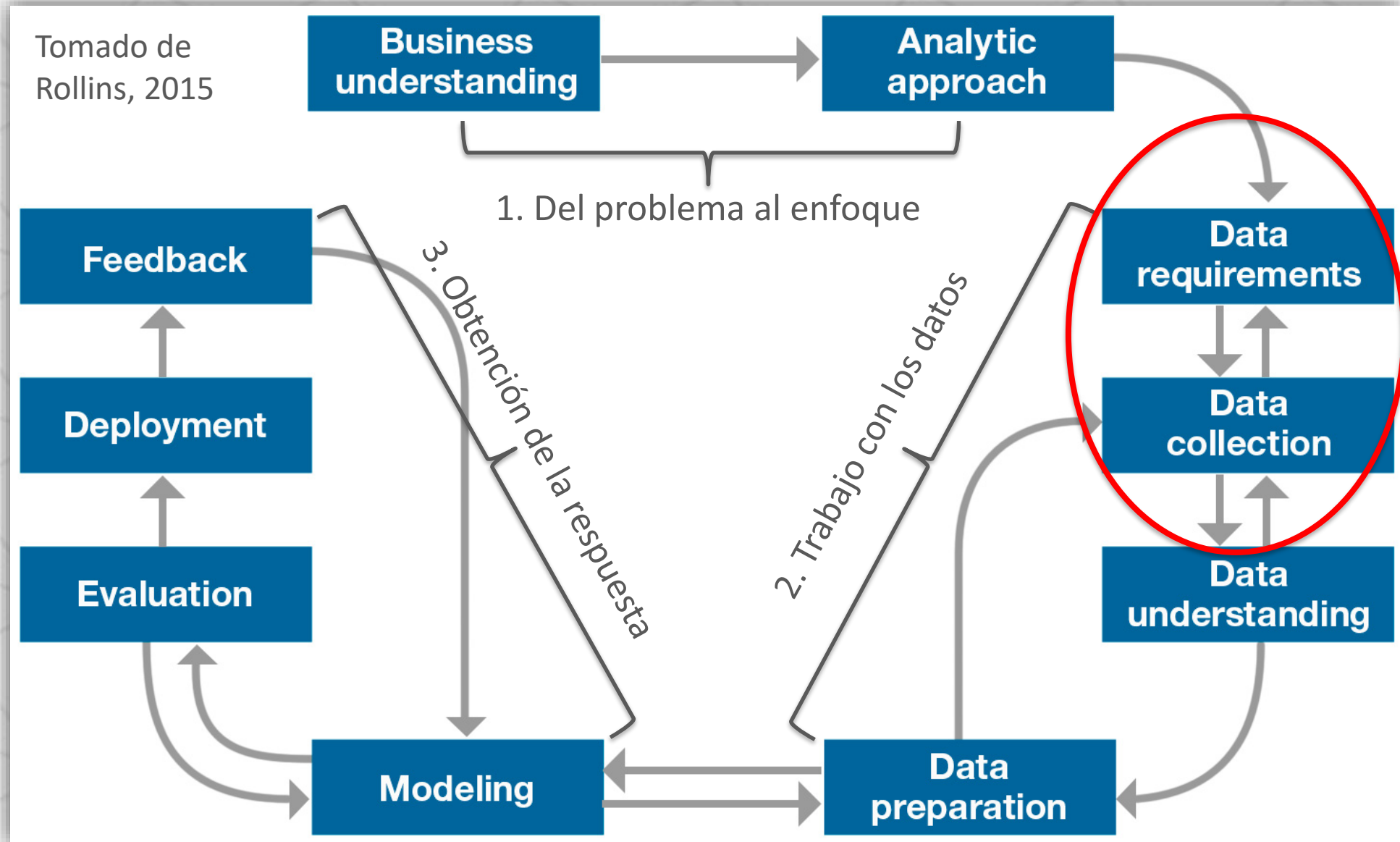
- ¿Por qué incluyeron este factor?

# Las 5 preguntas que responde el Data Science



01	¿Esto es A o B?: algoritmos de clasificación
02	¿Es esto normal?: algoritmos de detección de anomalías
03	¿Cuánto o cuántos?: algoritmos de regresión
04	¿Cómo está organizado?: algoritmos de clustering
05	¿Qué debería hacer?: algoritmos de refuerzo del aprendizaje

Tomado de: [Blogthinkbig.com](http://Blogthinkbig.com) - Telefónica

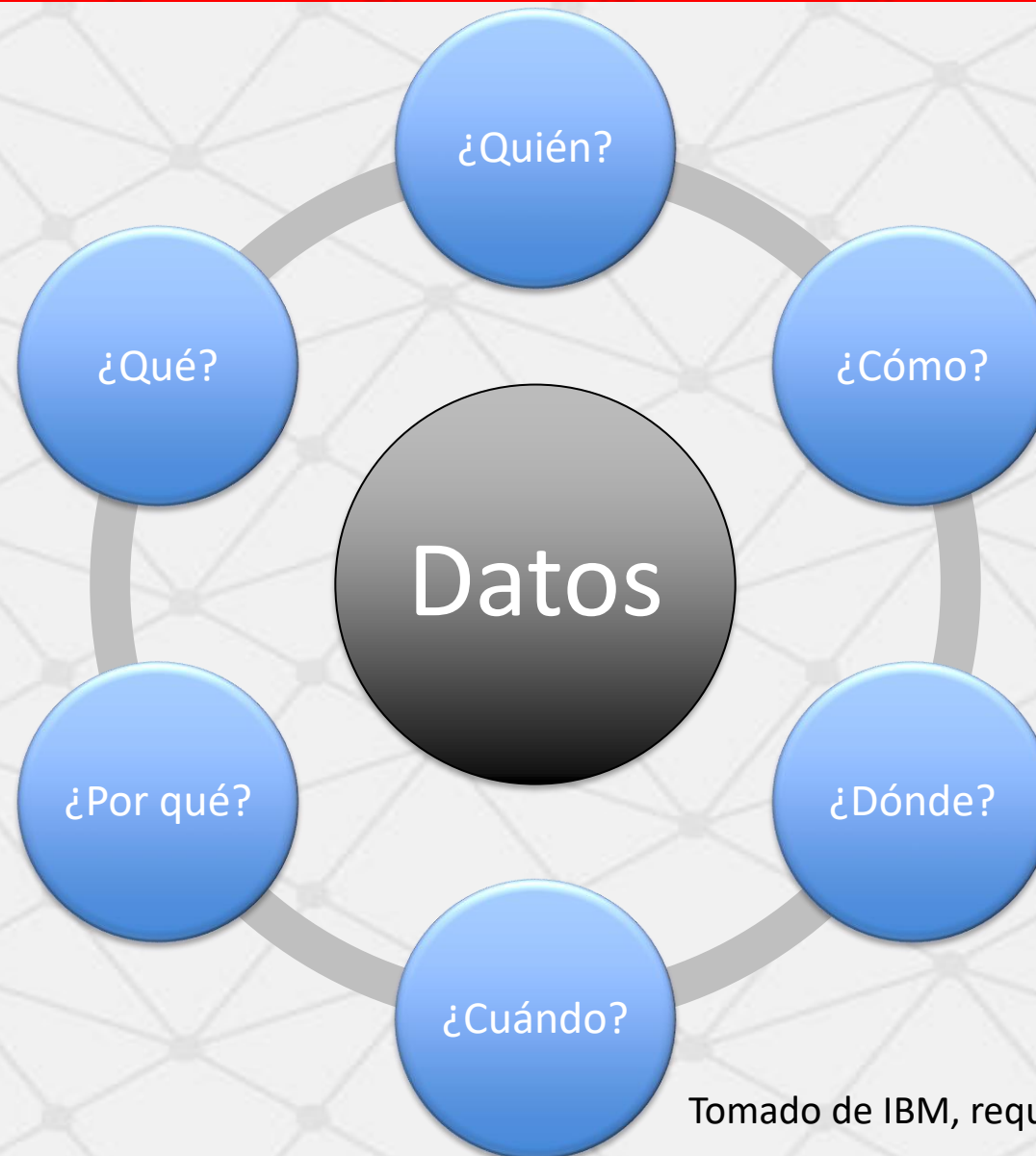




# Requisitos de datos



# ¿Cuáles son los requisitos de datos?



Tomado de IBM, requisitos de datos, 2018

# Recopilación de datos





# ¿Cómo se recopilan los datos? Ideas clave



Determine si los datos requeridos son de origen interno o externo

1

Identifique la fuente de los datos

2

Verifique la confiabilidad de los datos

3

Si los datos no existen en los registros, evalúe la posibilidad de recogerlos del origen

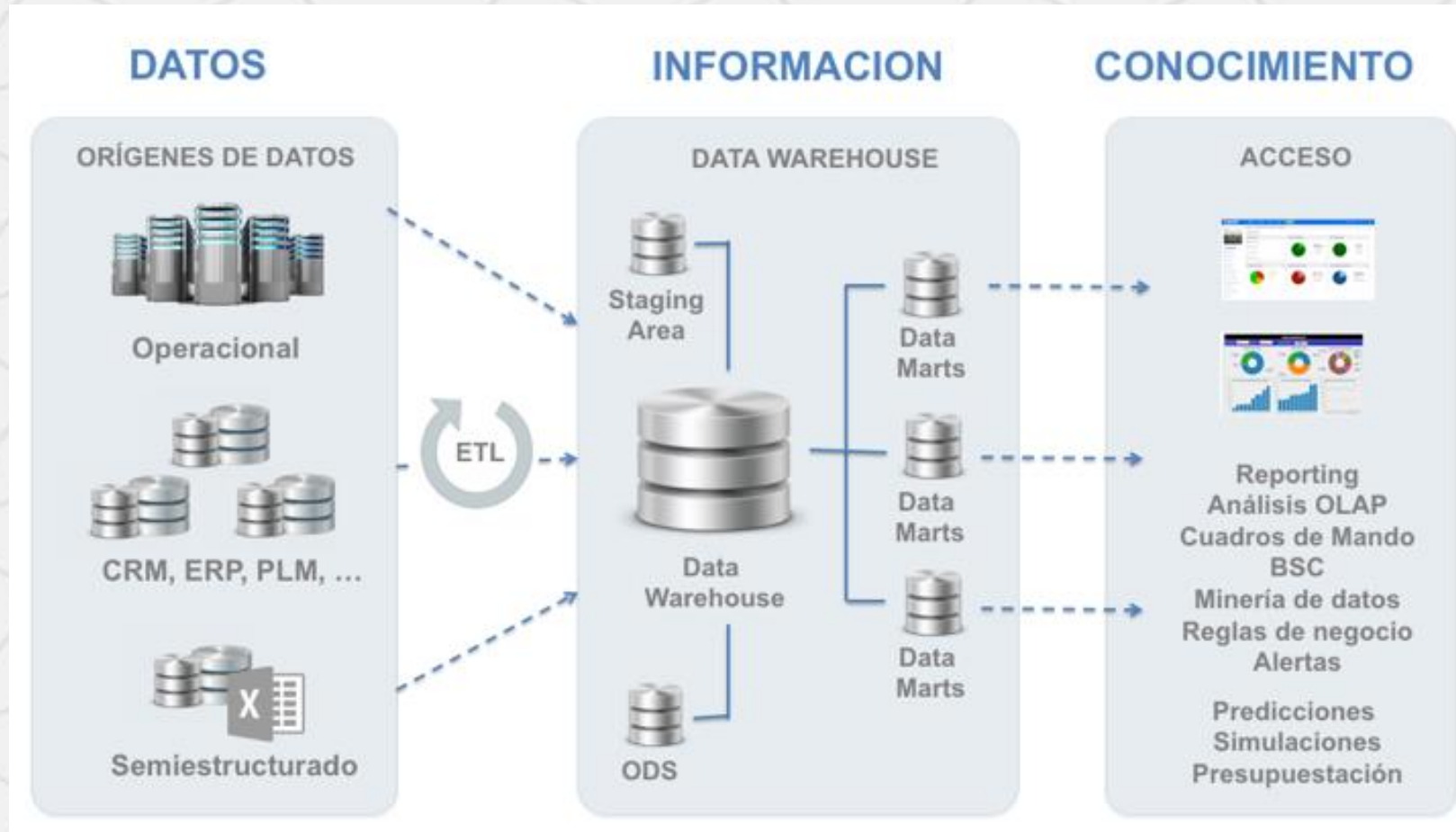
4

Recuerde que la recopilación de datos se hará cada vez que se identifique la necesidad

5

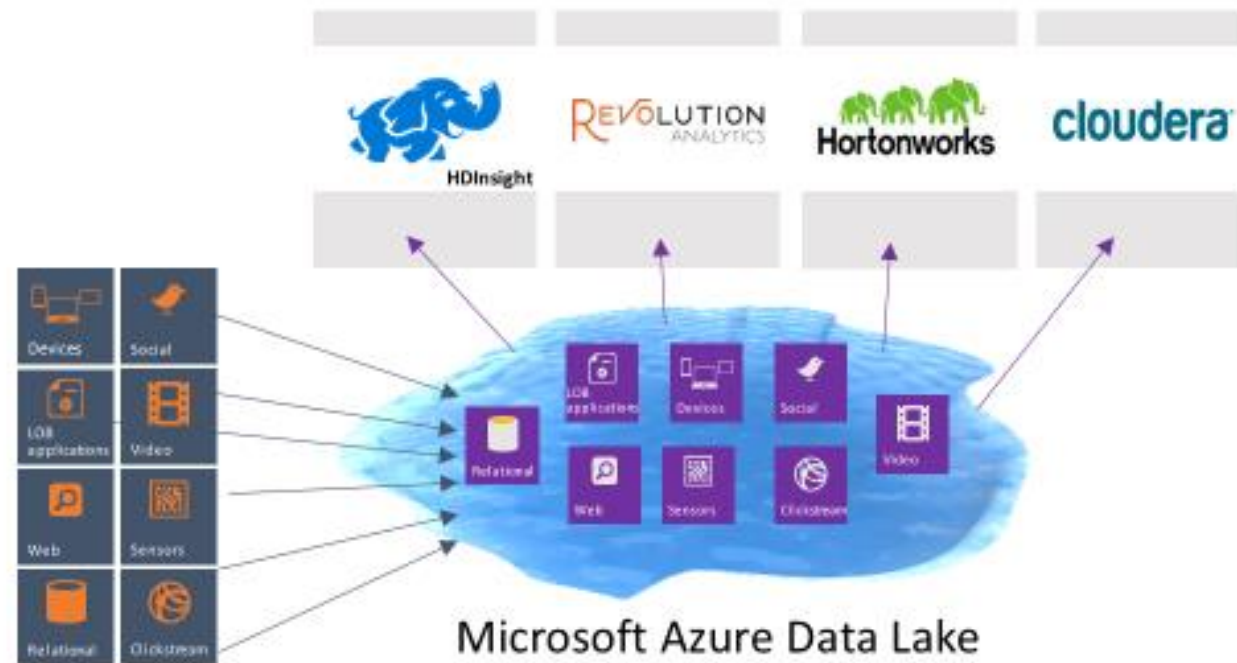
“No se necesita Big Data sino los datos correctos”. Wessel, 2016.

# Datawarehouse

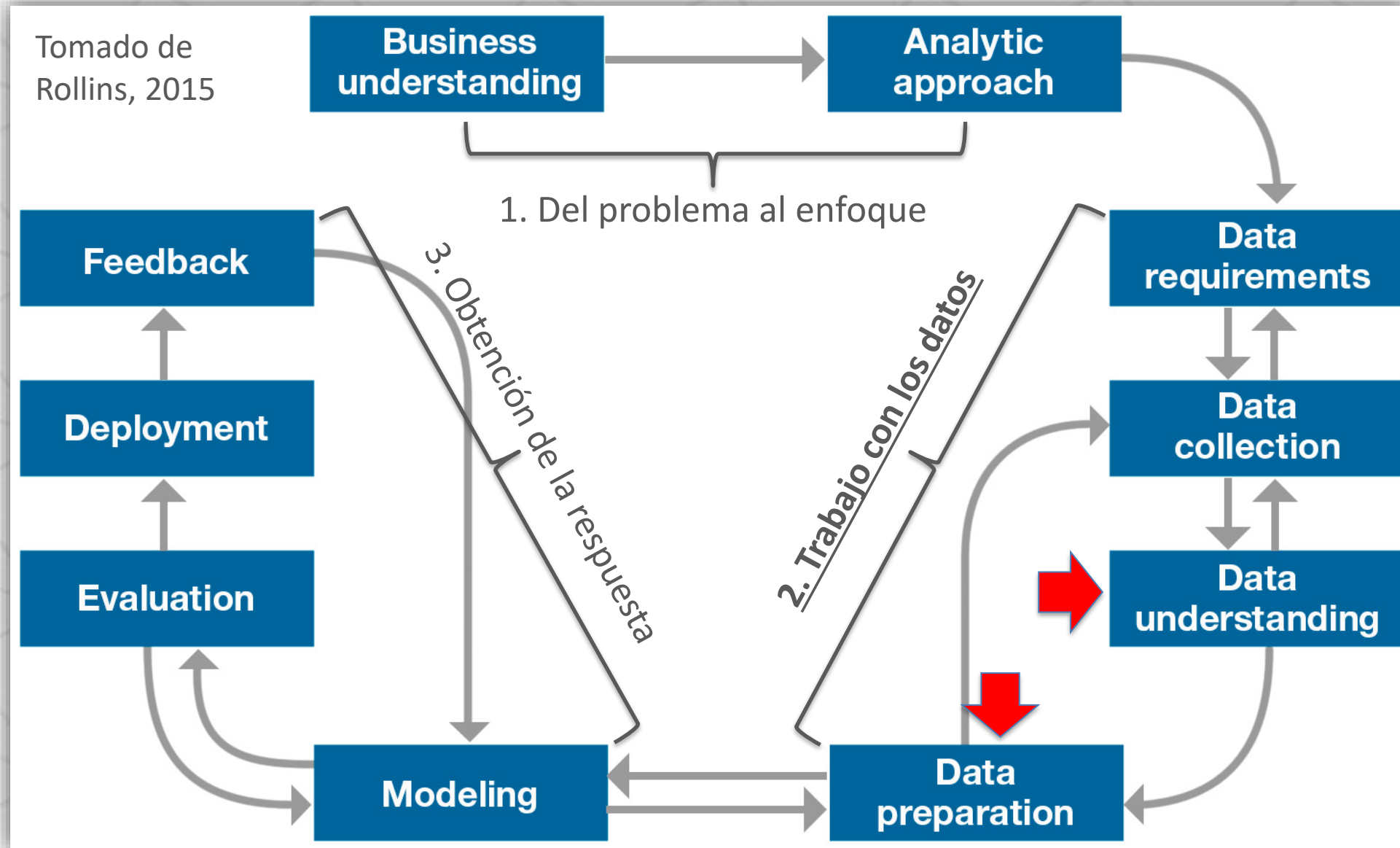


Fuente: Empresa Diego Calvo - España

## Introducing Azure Data Lake



Fuente: Microsoft







1

## Estadística descriptiva

Estadística univariable  
Correlaciones por pares  
Histogramas

2

## Calidad de los datos

Valores faltantes  
Valores no válidos o  
engañosos

3

## Proceso iterativo

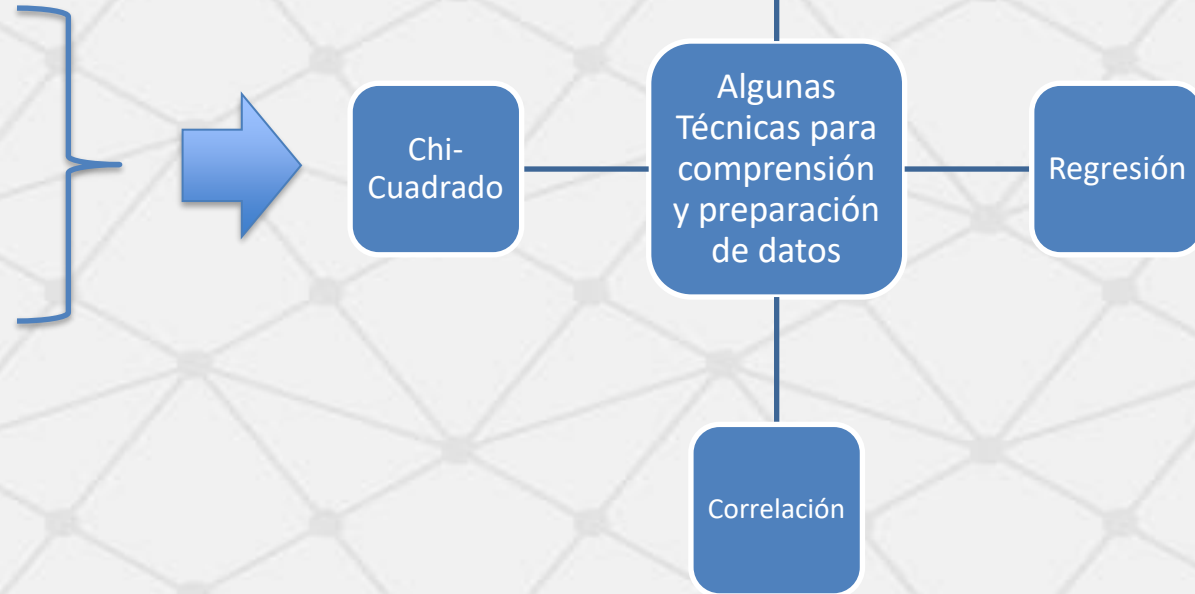
Recopilación y comprensión  
de datos iterativos

# Comprensión de datos





# Pasos para Comprensión de datos



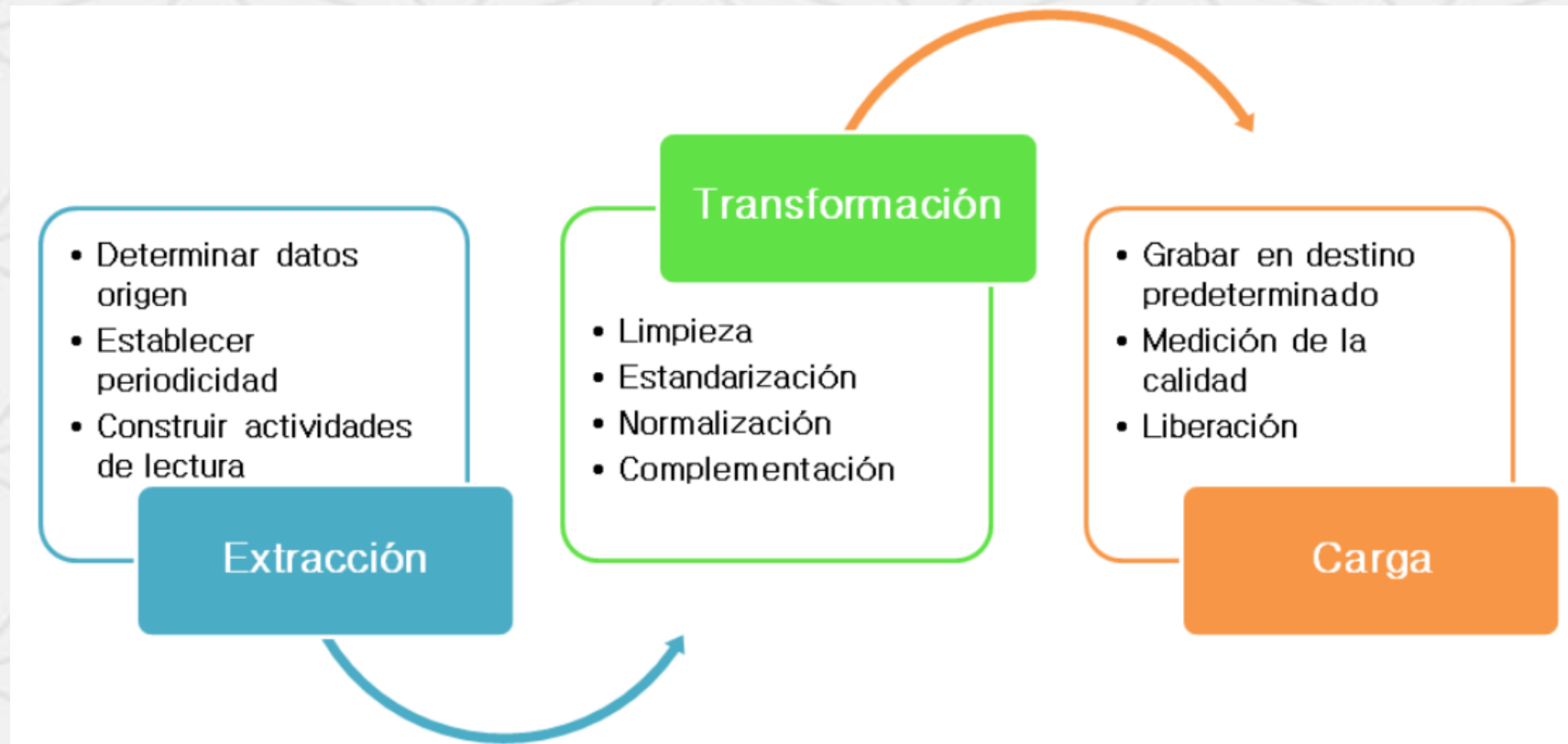
Adaptado de : Bizmetriks.com

# Preparación de datos

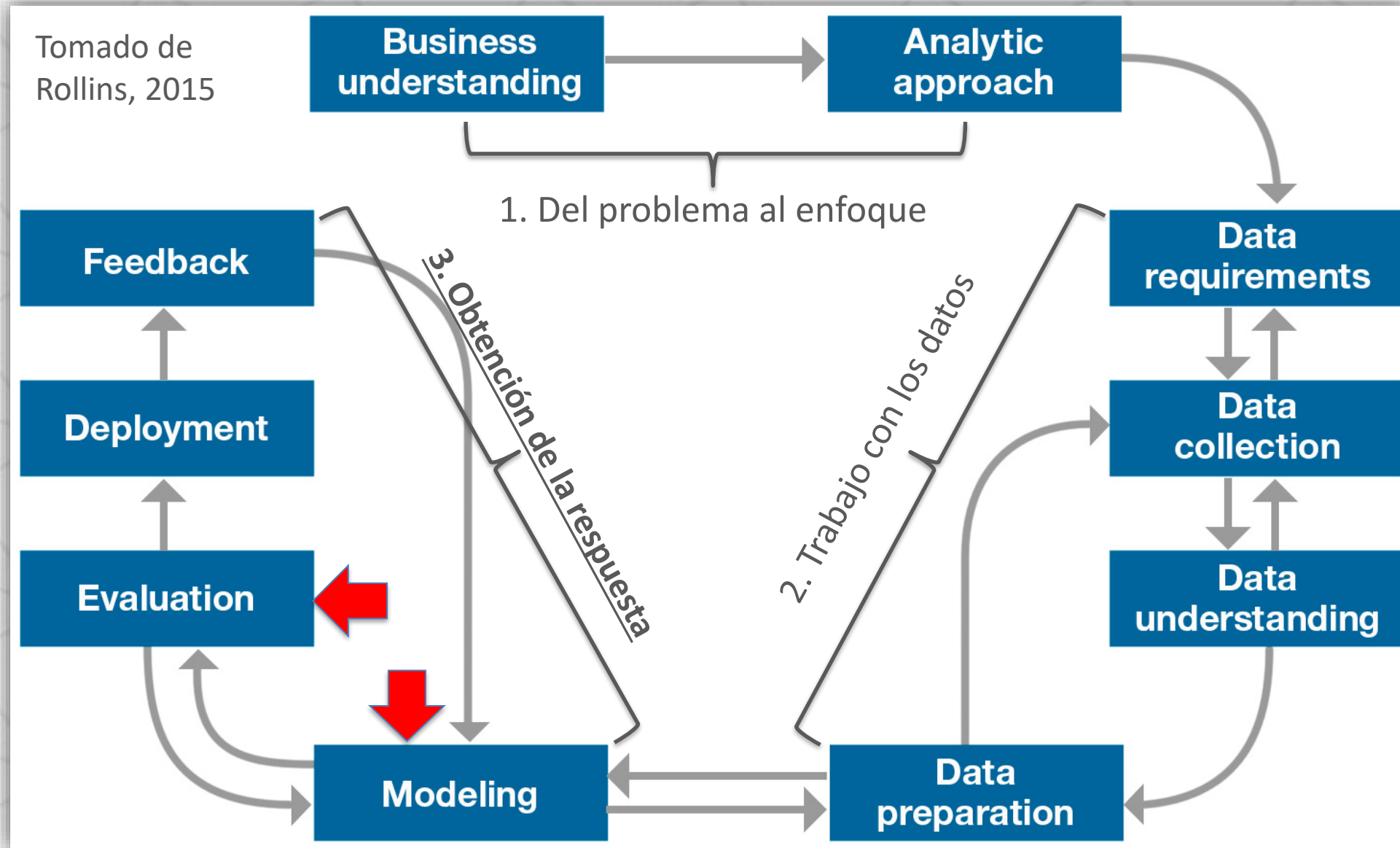




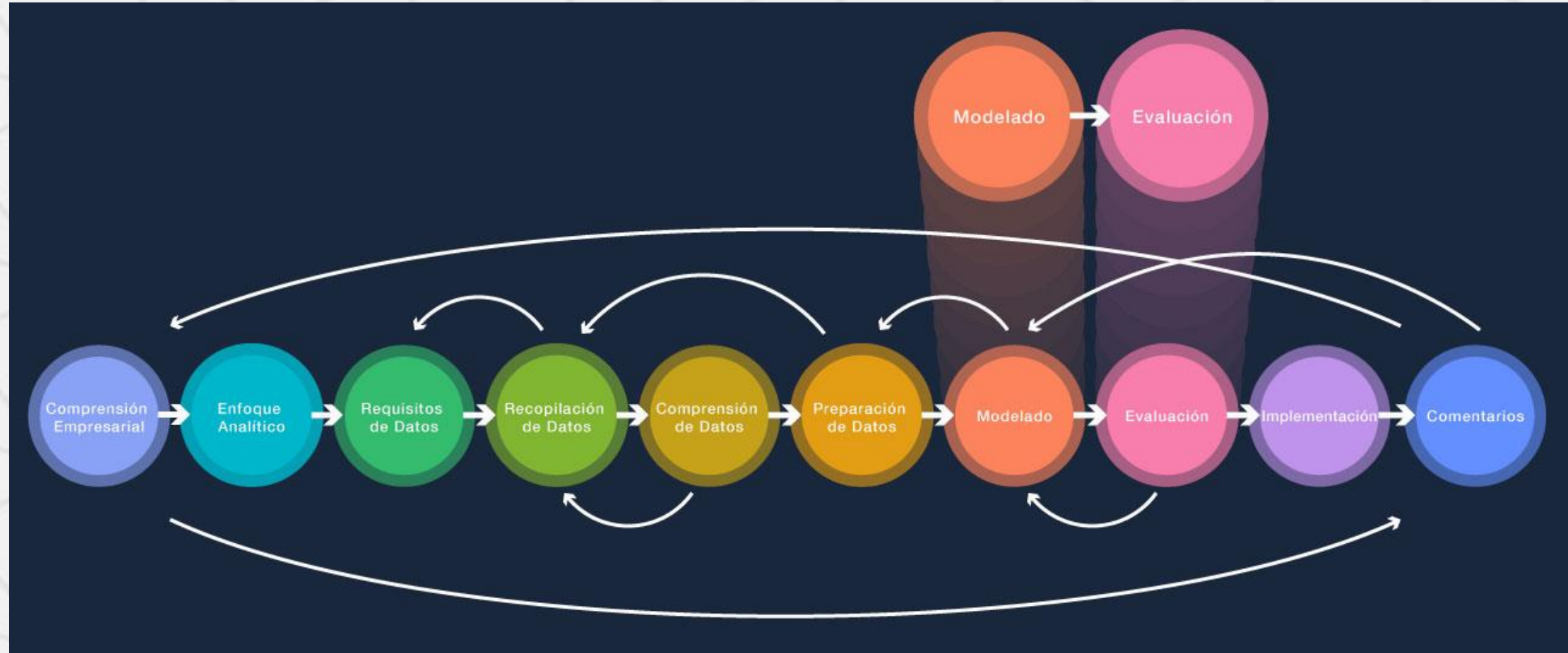
# Preparación de datos



Tomado de: dbadixit.com



# El modelado y la evaluación



<https://mym.cdn.laureate-media.com/2dett4d/IBM/LIDS0000/ES/infographic/index.html>

# Modelado de datos

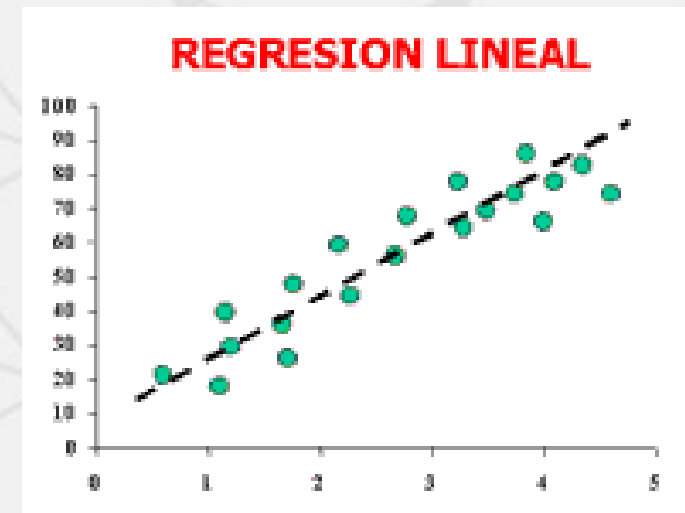




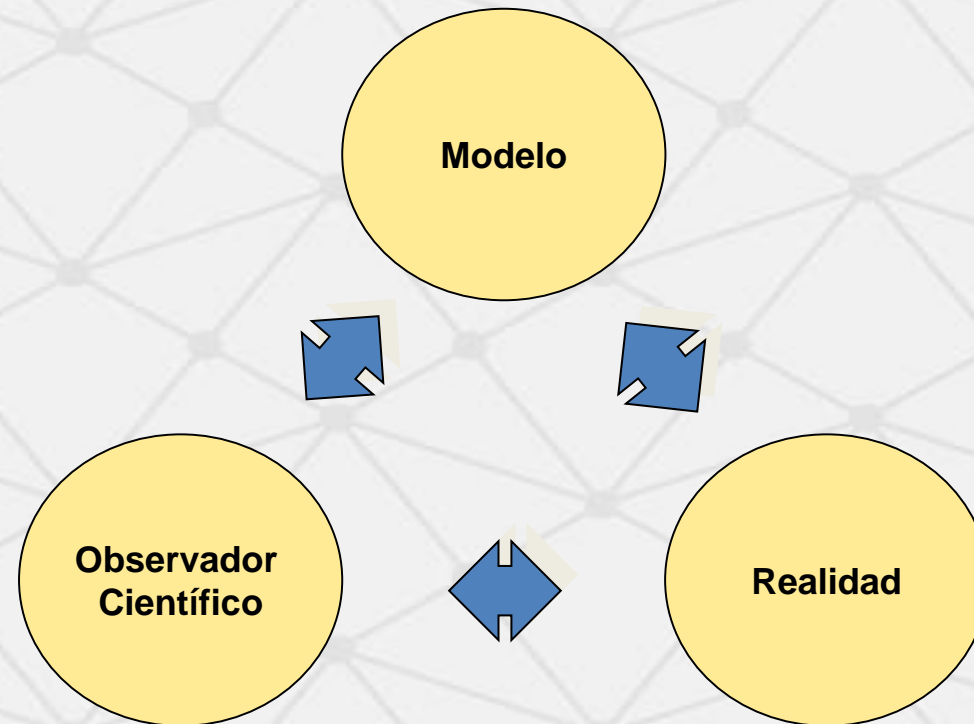
# ¿Qué es un modelo?



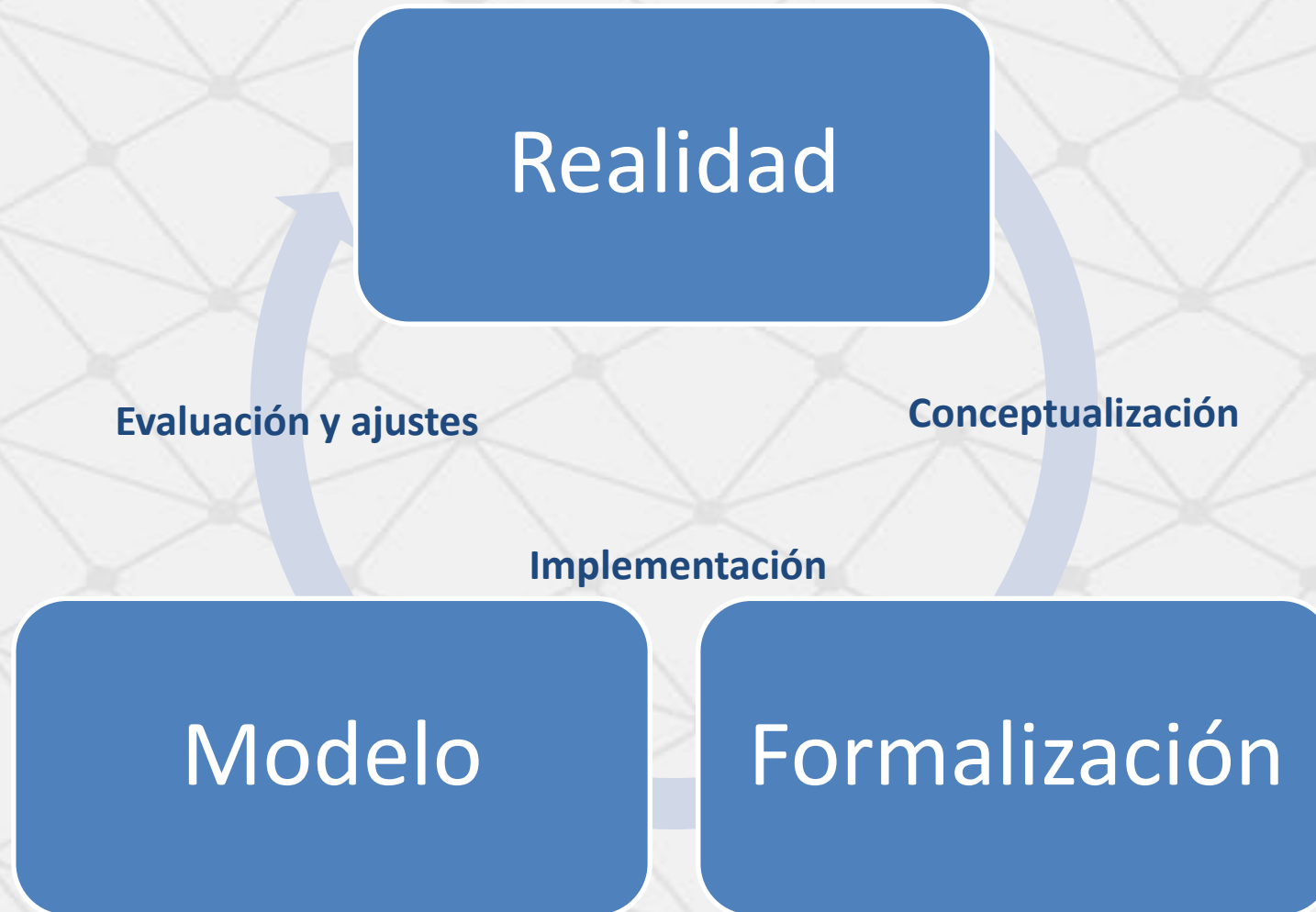
- Una ***abstracción*** de la realidad.
- Una ***representación*** de la realidad que ayuda a entenderla.
- Una ***construcción intelectual***, descriptiva o predictiva



# El triángulo: Modelo – Realidad - Científico



# ¿Qué hacemos cuando creamos un modelo?



# Modelos descriptivos y predictivos



Tomado de: <https://www.lainnovacionnecesaria.com/>



# Modelos descriptivos y predictivos



## Modelo Actual

Descriptivo  
e inferencial

Modelo que analiza datos, los describe, busca explicar la realidad y puede inferir posibles comportamientos

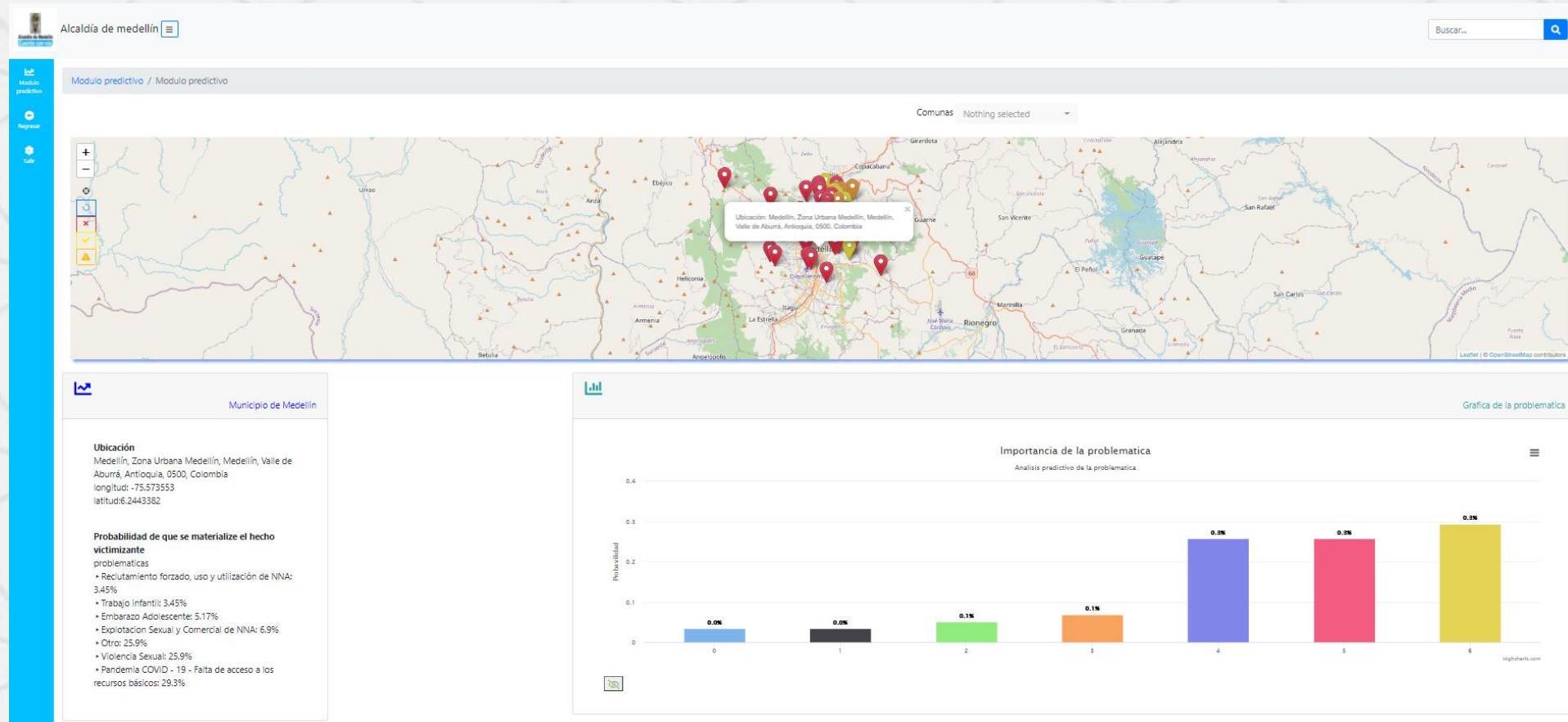


## Modelo Futuro

Predictivo e  
Inteligente

Modelo que analiza datos, identifica relaciones entre ellos, aprende de las relaciones y a partir de ello predice comportamientos, ante situaciones iguales o diferentes a las originales comportamientos

# Modelos Predictivos: El caso del SATMED Sistemas Alertas Tempranas Medellín



Recuperado de : <http://alertapp.satmed.co/mp/probabilidad>

# Alguno ejemplos de modelos de ciencia de datos usando machine learning



## Hospital Universitario de San Juan de Alicante

**Objetivo:** Reducción de costes al permitir un óptimo consumo de recursos.

**Problema planteado:** Las pruebas preparatorias que se llevan a cabo antes de cualquier intervención suelen ser excesivas, invasivas, caras y generan listas de espera para la intervención.

**Solución:** Analizando los datos del histórico de operaciones y aplicando técnicas de Minería de Datos se descubren aquellos casos en que dichas pruebas son prescindibles. El sistema de calidad proporciona información detallada del resultado de la operación, de forma que aquellos casos en los que no se han realizado las pruebas y sí hubiesen sido necesarias permiten al sistema seguir aprendiendo y mejorar la identificación de los patrones adecuados.



## Goldcar

**Objetivo:** Reducción de gastos anticipándose a problemas derivados de la demanda.

**Problema planteado:** Las reservas a través de su portal online que no terminan en alquiler generan grandes gastos. Como no se requiere pago previo para realizar una reserva, muchos usuarios no se presentan a recoger el coche reservado. Esto genera grandes gastos por los coches que quedan esperando a esos usuarios que nunca llegan.

**Solución:** Por medio del análisis de los datos de las reservas se identifican perfiles concretos que terminan en casos de reservas canceladas o clientes no presentados, en función de la procedencia, la temporada, antelación y otros factores clave.



## Portal B2B Neumáticos Soledad

**Objetivo:** Aumentar las ventas a través del portal.

**Problema planteado:** Cómo modificar el portal de compra online que usan los talleres asociados para aumentar las ventas por este canal.

**Solución:** Extraer patrones de comportamiento de los usuarios sobre el motor de búsquedas del portal, analizando aquellas búsquedas que terminan en pedido y las que no.



# El valor del Big Data: contar la historia.



Las estrellas del fútbol ya no solo renuevan por el dinero: De Bruyne contrató a analistas de datos para valorar su futuro en el Manchester City



Tomado de: <https://www.xataka.com/investigacion/estrellas-futbol-no-solo-renuevan-dinero-bruyne-contrato-a-analistas-datos-para-valorar-su-futuro-manchester-city>



# Evaluación (del modelo de datos)



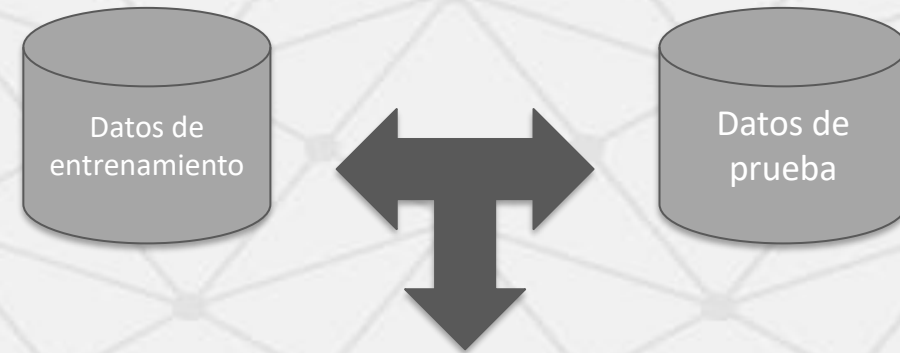


La evaluación de un modelo de datos se realiza en forma iterativa con el modelado

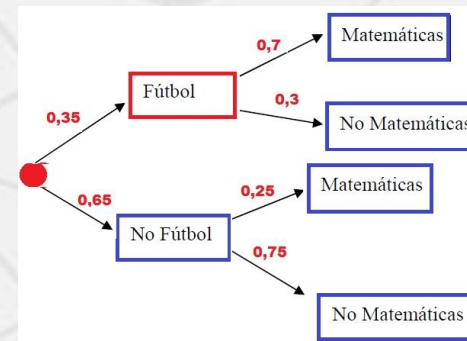
Se evalúa el modelo para comprender su calidad y garantizar que aborde de manera adecuada y completa el problema del negocio.

Evaluación:

¿El modelo utilizado realmente responde a la pregunta inicial o deben de realizarse ajustes?



Validación:



**Modelo predictivo::** Evalúa si el resultado de las pruebas luego del entrenamiento concuerda con los datos iniciales

# Algunas frases para comenzar



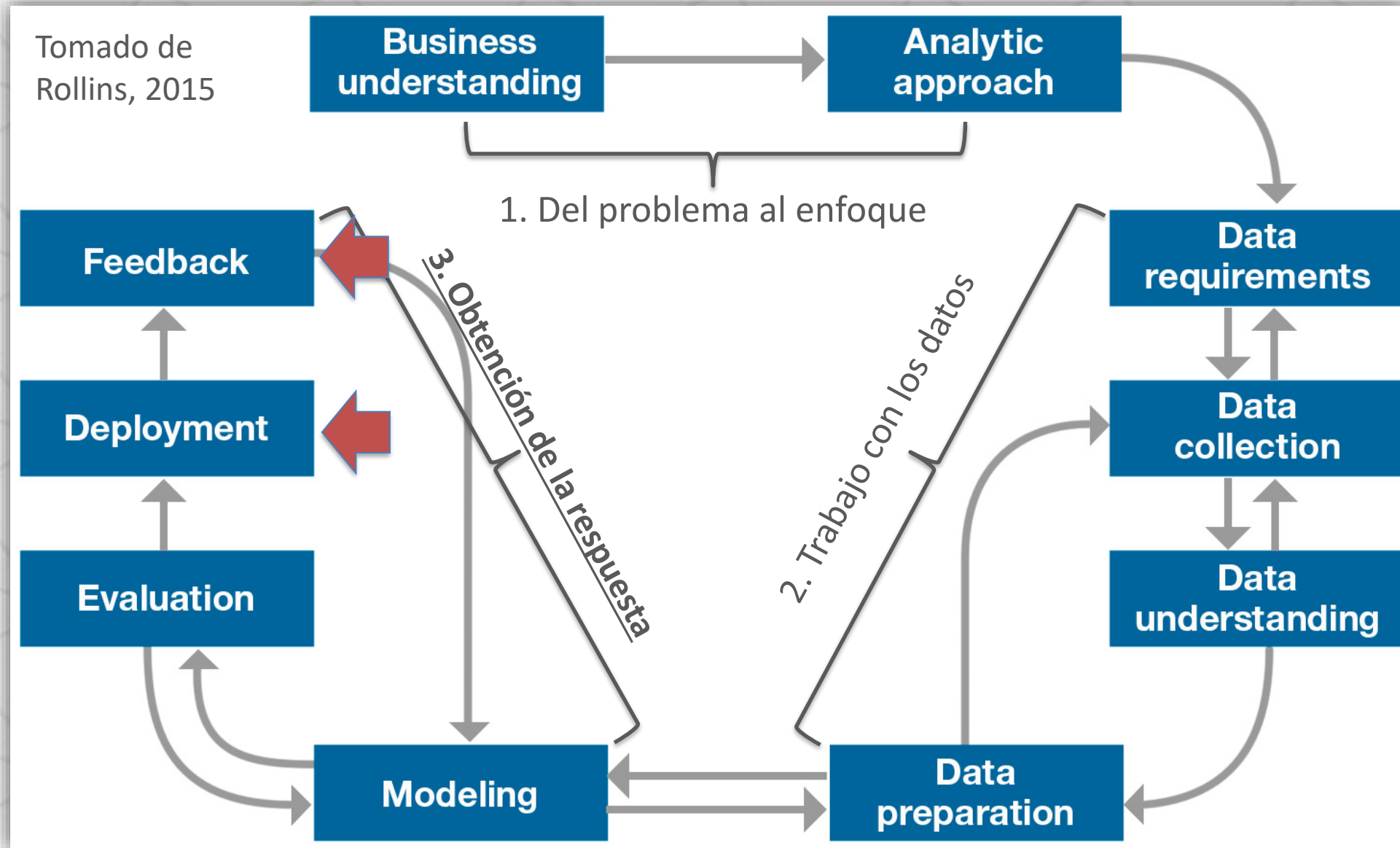
El valor del modelo de ciencia de datos lo define el negocio y la implementación es el momento de saberlo.

*Luis Alcántara*



La estrategia de implementación es tan importante como el modelo en sí.

*Luis Alcántara*





# Implementación



<https://mym.cdn.laureate-media.com/2dett4d/IBM/LIDS0000/ES/infographic/index.html>

# La estrategia mirando al usuario



***Por supuesto, la implementación se realizará si los aspectos técnicos ya han sido resueltos***

*Tomado de: Luis Alcántara – Merlin Data Technology*

# Los pasos a seguir



*Tomado de: Luis Alcántara – Merlin Data Technology*

# Los pasos a seguir



Tomado de: Luis Alcántara – Merlin Data Technology



# Los pasos a seguir



*Tomado de: Luis Alcántara – Merlin Data Technology*

# Los pasos a seguir



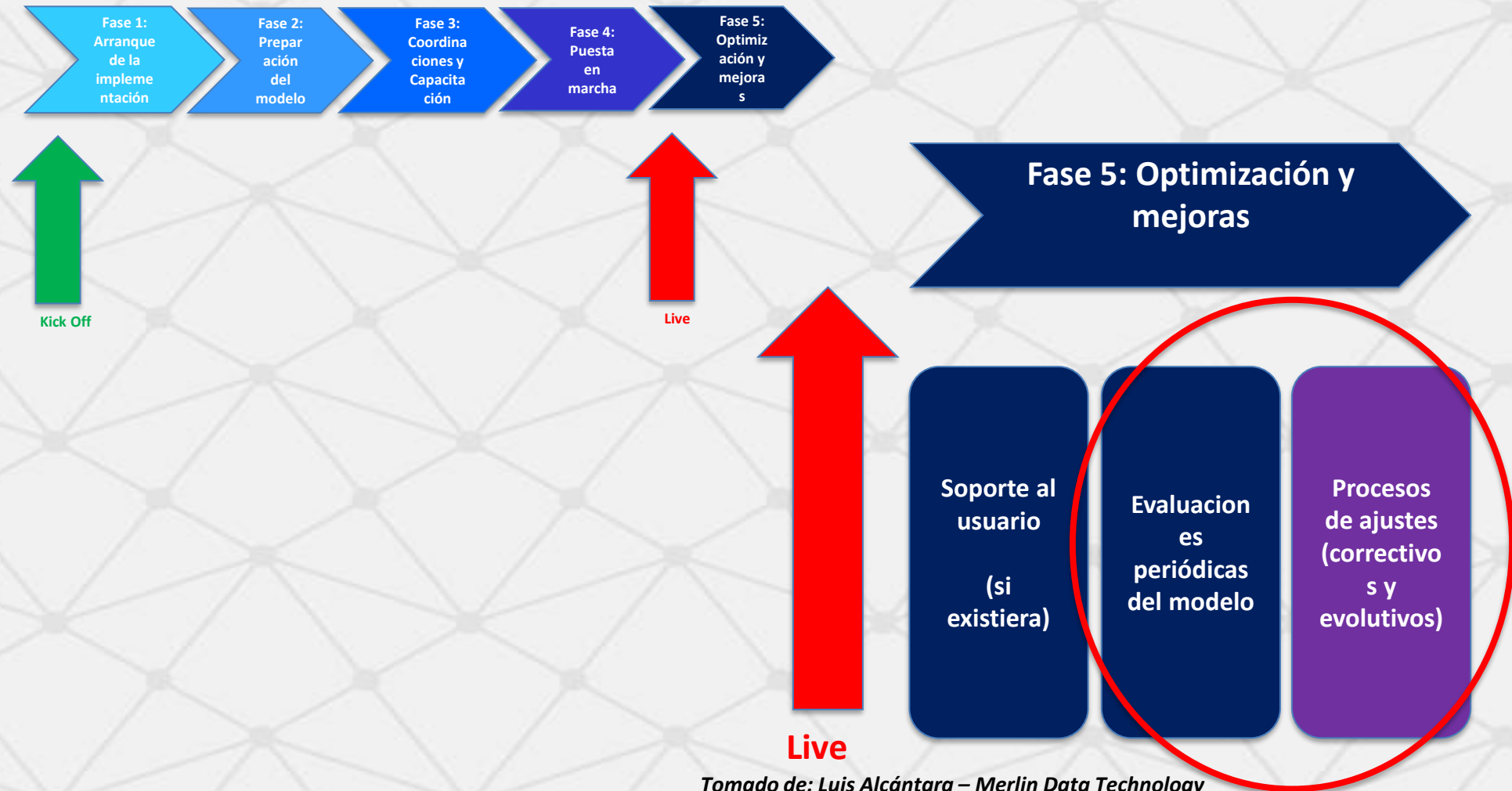
*Las actividades de estas fases dependerán de la estrategia de implementación y despliegue y de la madurez en ciencia de datos de la organización*

Tomado de: Luis Alcántara – Merlin Data Technology



<https://mym.cdn.laureate-media.com/2dett4d/IBM/LIDS0000/ES/infographic/index.html>

# Partamos de este proceso y concentrémonos en la última fase



Tomado de: Luis Alcántara – Merlin Data Technology

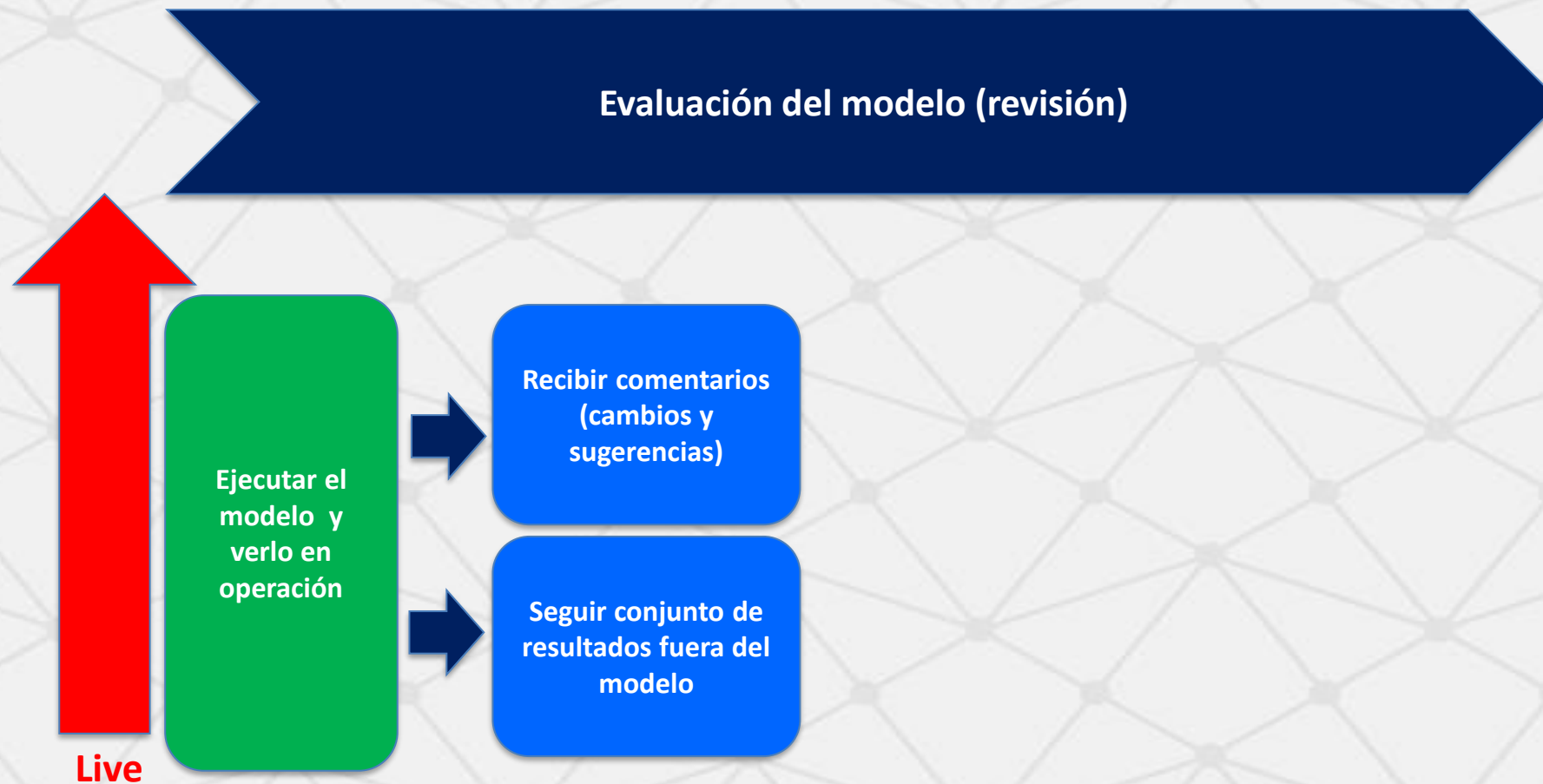


# La evaluación (revisión) del modelo



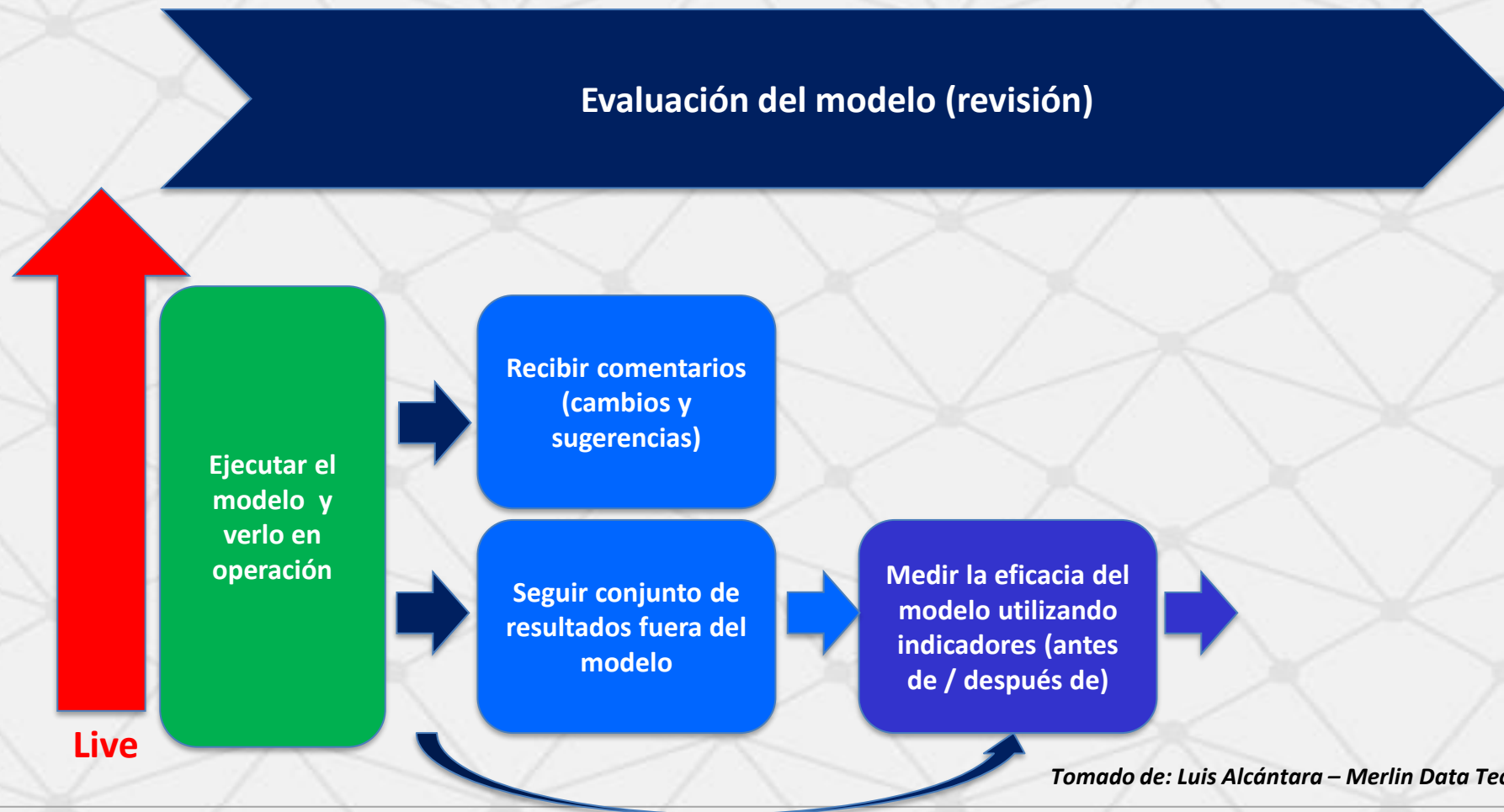
*Tomado de: Luis Alcántara – Merlin Data Technology*

# La evaluación (revisión) del modelo



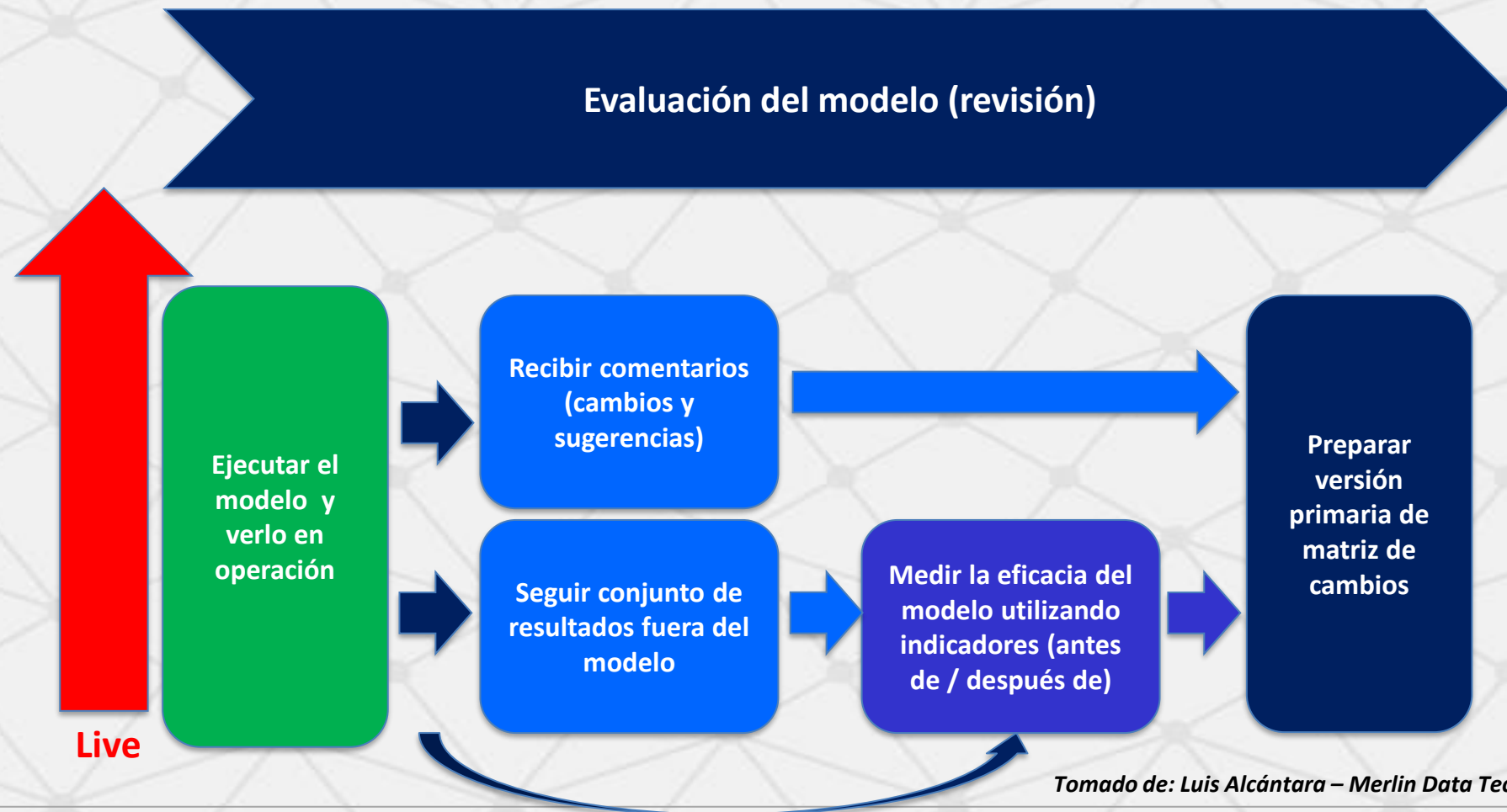
Tomado de: Luis Alcántara – Merlin Data Technology

# La evaluación (revisión) del modelo



Tomado de: Luis Alcántara – Merlin Data Technology

# La evaluación (revisión) del modelo



Tomado de: Luis Alcántara – Merlin Data Technology



# Ajustes al modelo



Ajustes al modelo

Revisión de la  
matriz de  
cambios

*Tomado de: Luis Alcántara – Merlin Data Technology*

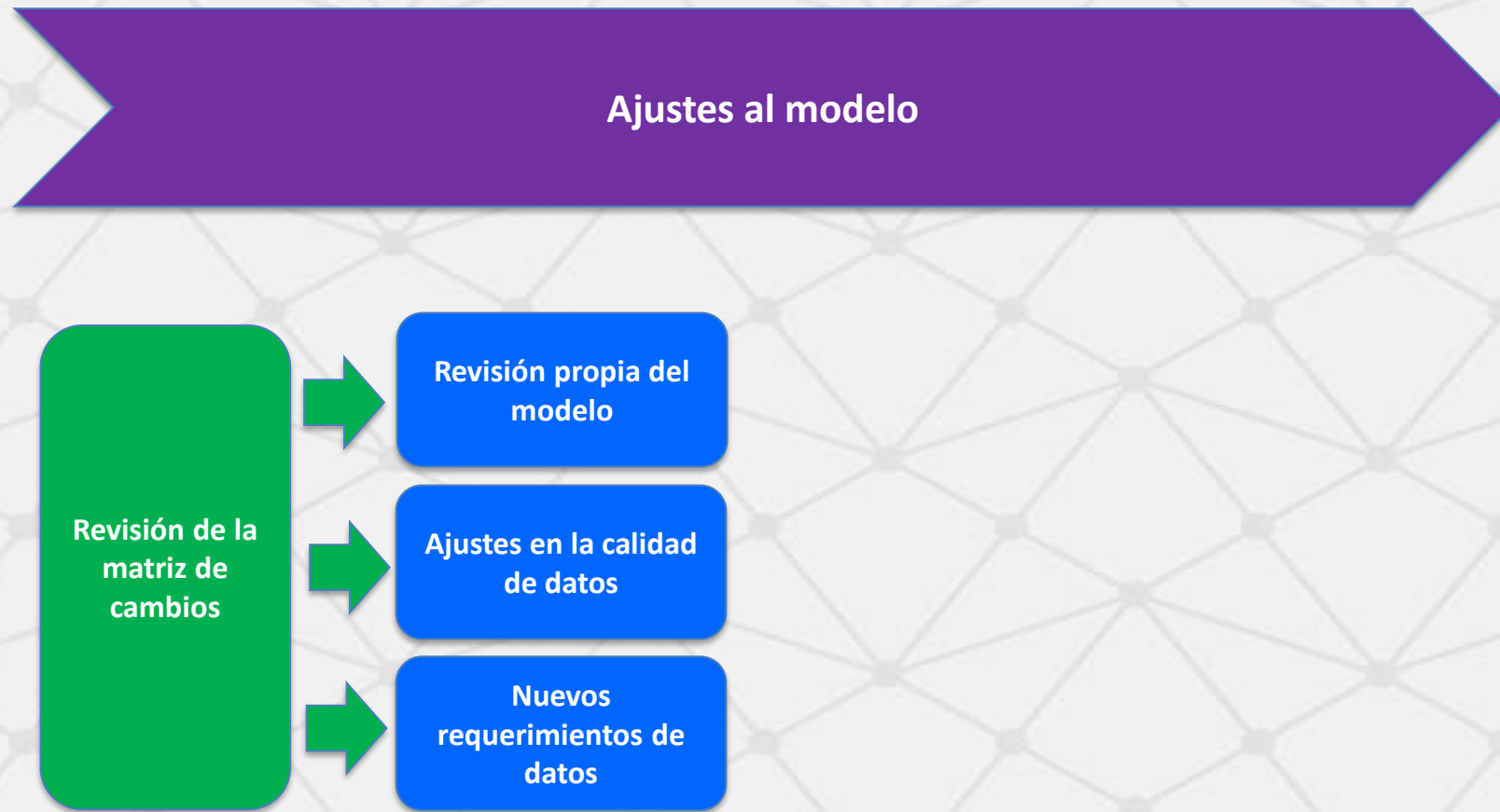


**LAUREATE**  
INTERNATIONAL  
UNIVERSITIES®



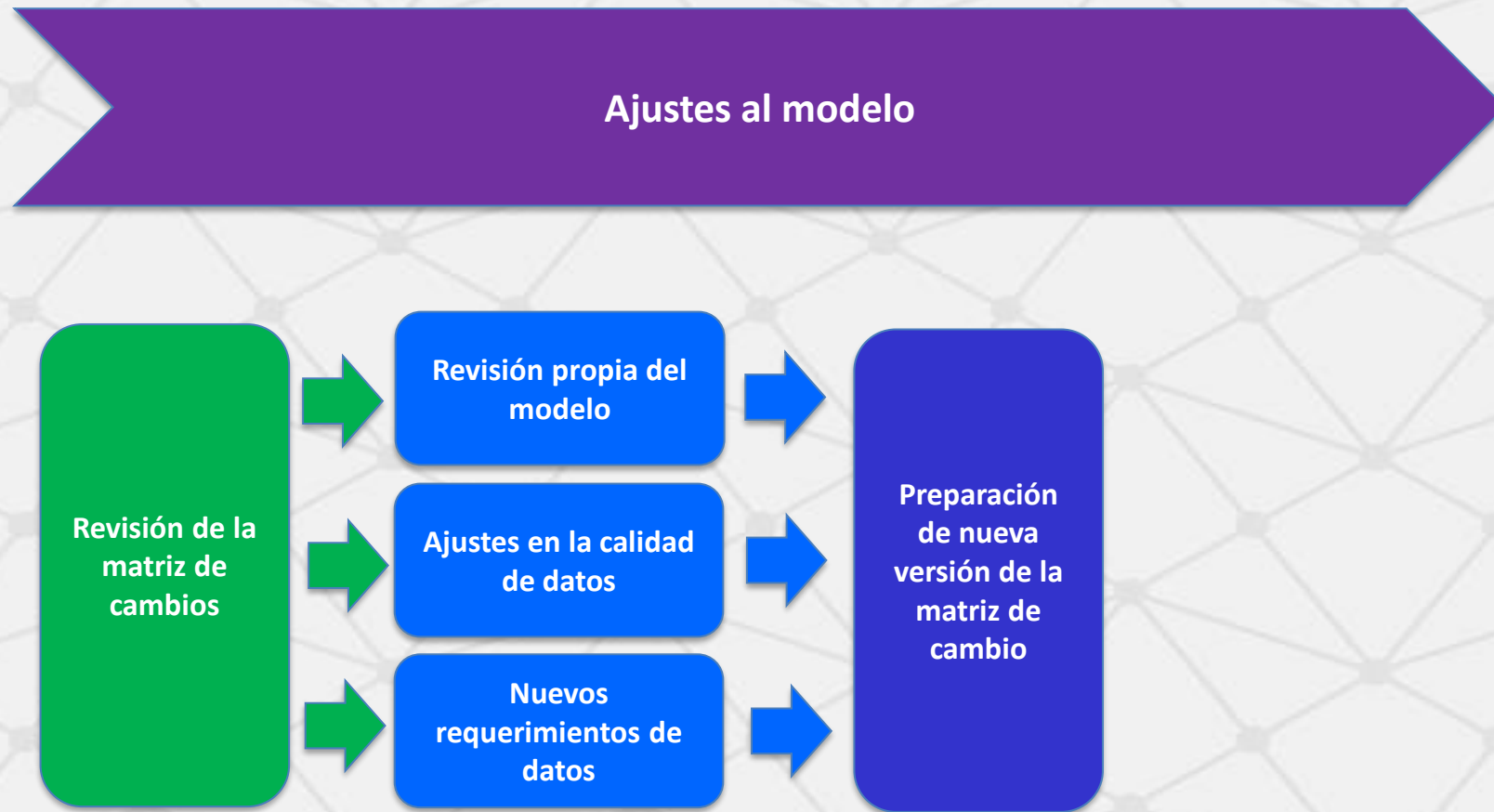
Facultad de Negocios UPC

# Ajustes al modelo



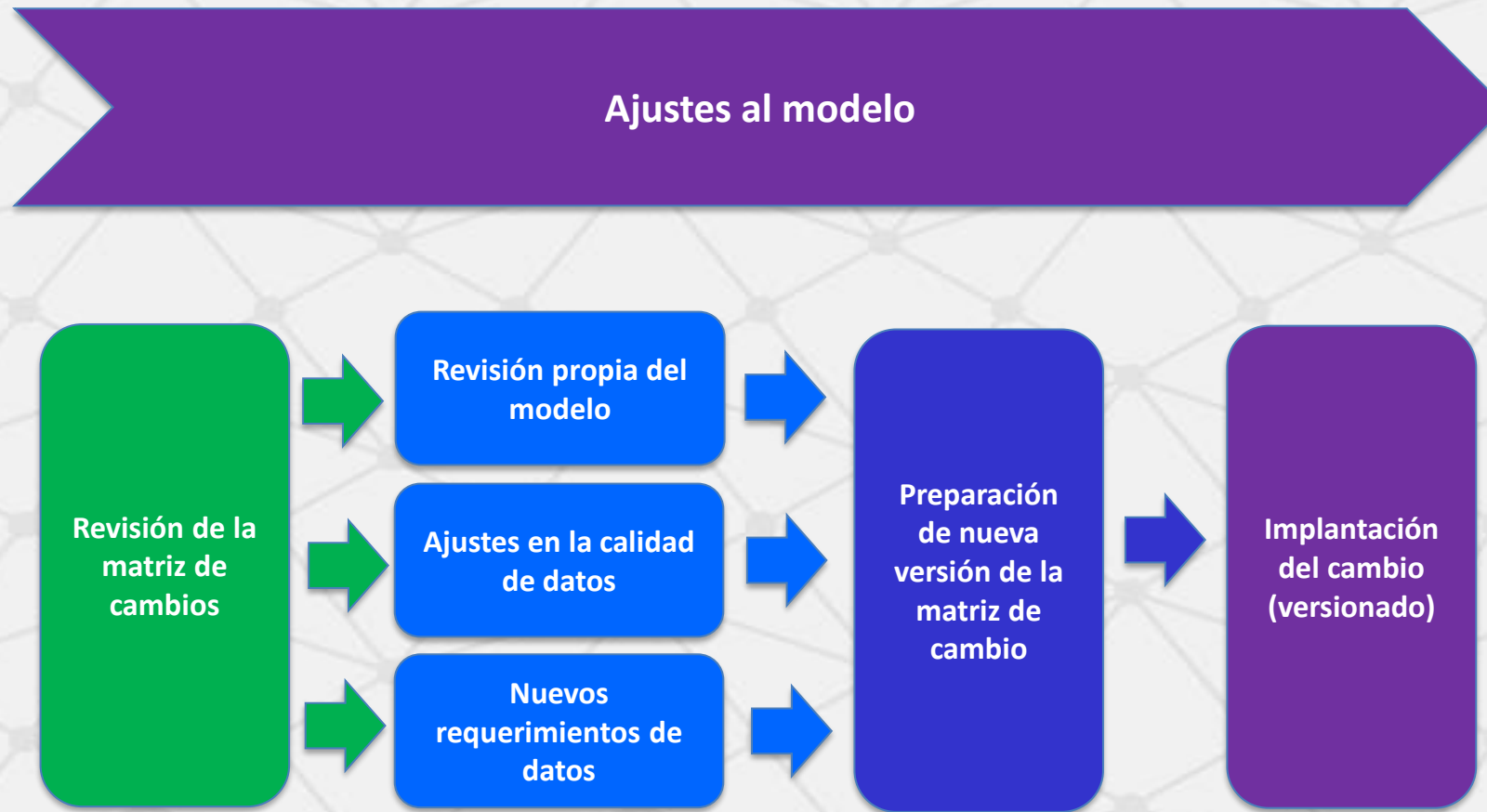
*Tomado de: Luis Alcántara – Merlin Data Technology*

# Ajustes al modelo



*Tomado de: Luis Alcántara – Merlin Data Technology*

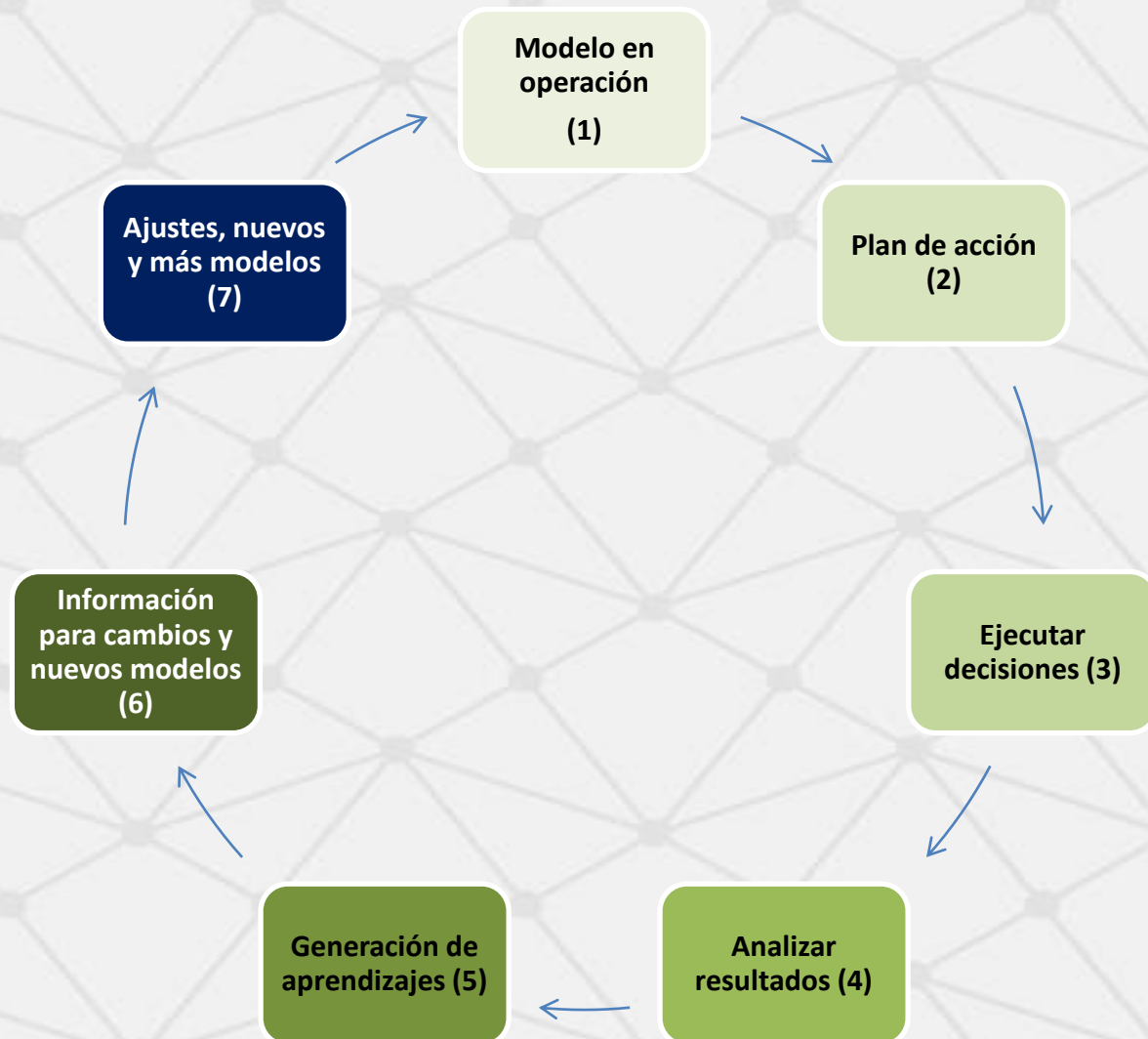
# Ajustes al modelo



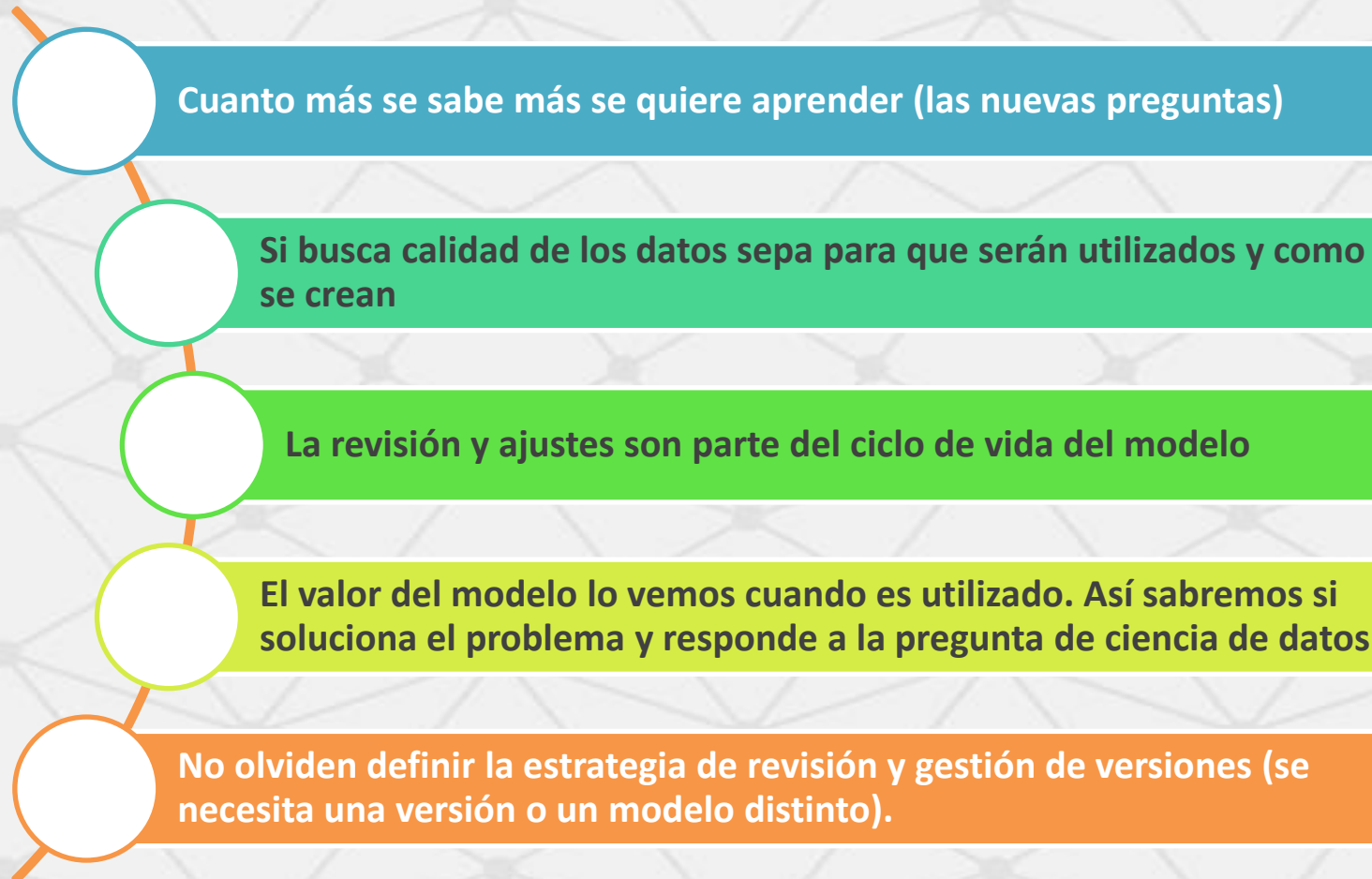
Tomado de: Luis Alcántara – Merlin Data Technology



# Ejecutando el modelo y poniéndolo en operación – Tomando decisiones



# Apuntes sobre retroalimentación y ajustes del modelo





GRACIAS POR LA ATENCIÓN

