



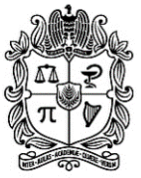
UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# Introducción al Análisis de Datos

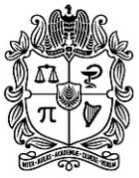
**Felipe Restrepo Calle, PhD.**

<https://dis.unal.edu.co/~ferestrepoca/>  
[ferestrepoca@unal.edu.co](mailto:ferestrepoca@unal.edu.co)

Departamento de Ingeniería de Sistemas e Industrial  
Facultad de Ingeniería  
Universidad Nacional de Colombia  
Sede Bogotá



1. ¿Qué es?
2. Big Data vs. Data Science
3. Herramientas
4. Aplicaciones y Retos
5. Metodología



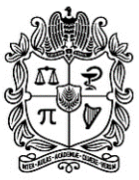
## 1. ¿Qué es?

2. Big Data vs. Data Science

3. Herramientas

4. Aplicaciones y Retos

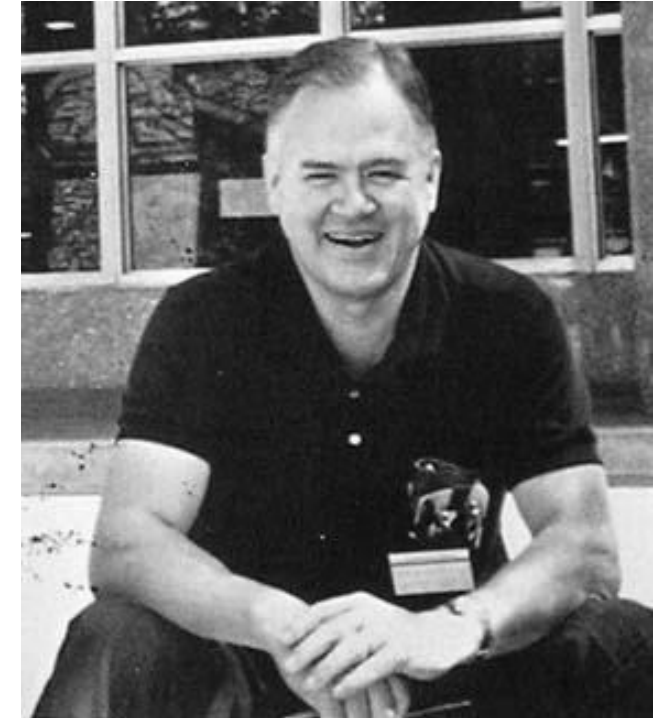
5. Metodología



## ¿Qué es el análisis de datos?

John Tukey definió el análisis de datos en 1961 así:

*“Son los procedimientos para analizar datos, las técnicas para interpretar los resultados de dichos procedimientos, las formas de planear la recolecta de datos para hacer el análisis más fácil, más preciso o más exacto.”*

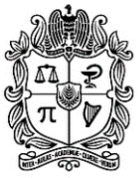


John Tukey

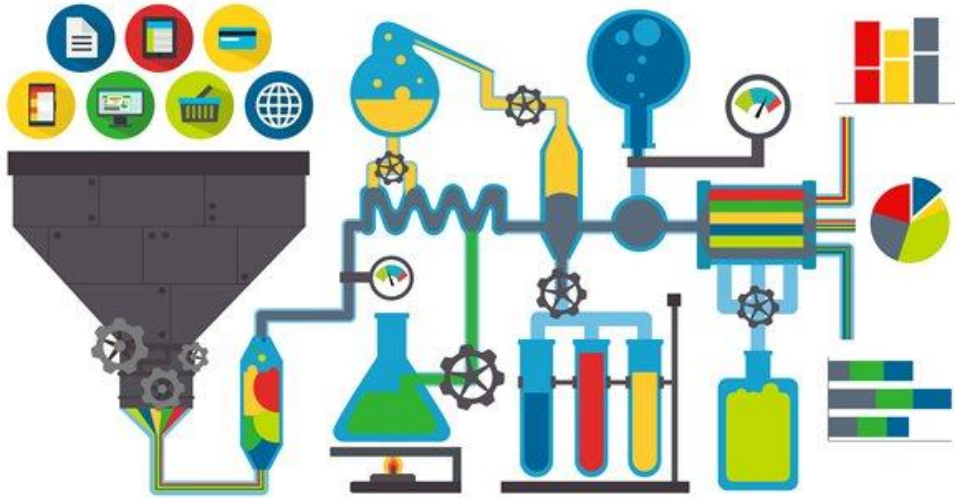
***The Future of Data Analysis***

July 1961

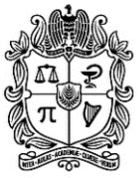
[https://projecteuclid.org/download/pdf\\_1/euclid.aoms/1177704711](https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711)



## ¿Qué es el análisis de datos?



- Es el proceso de: inspeccionar, limpiar y transformar datos.
- Con el objetivo de resaltar y extraer información útil para apoyo en la toma de decisiones.



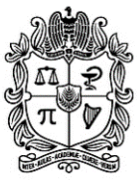
1. ¿Qué es?

**2. Big Data vs. Data Science**

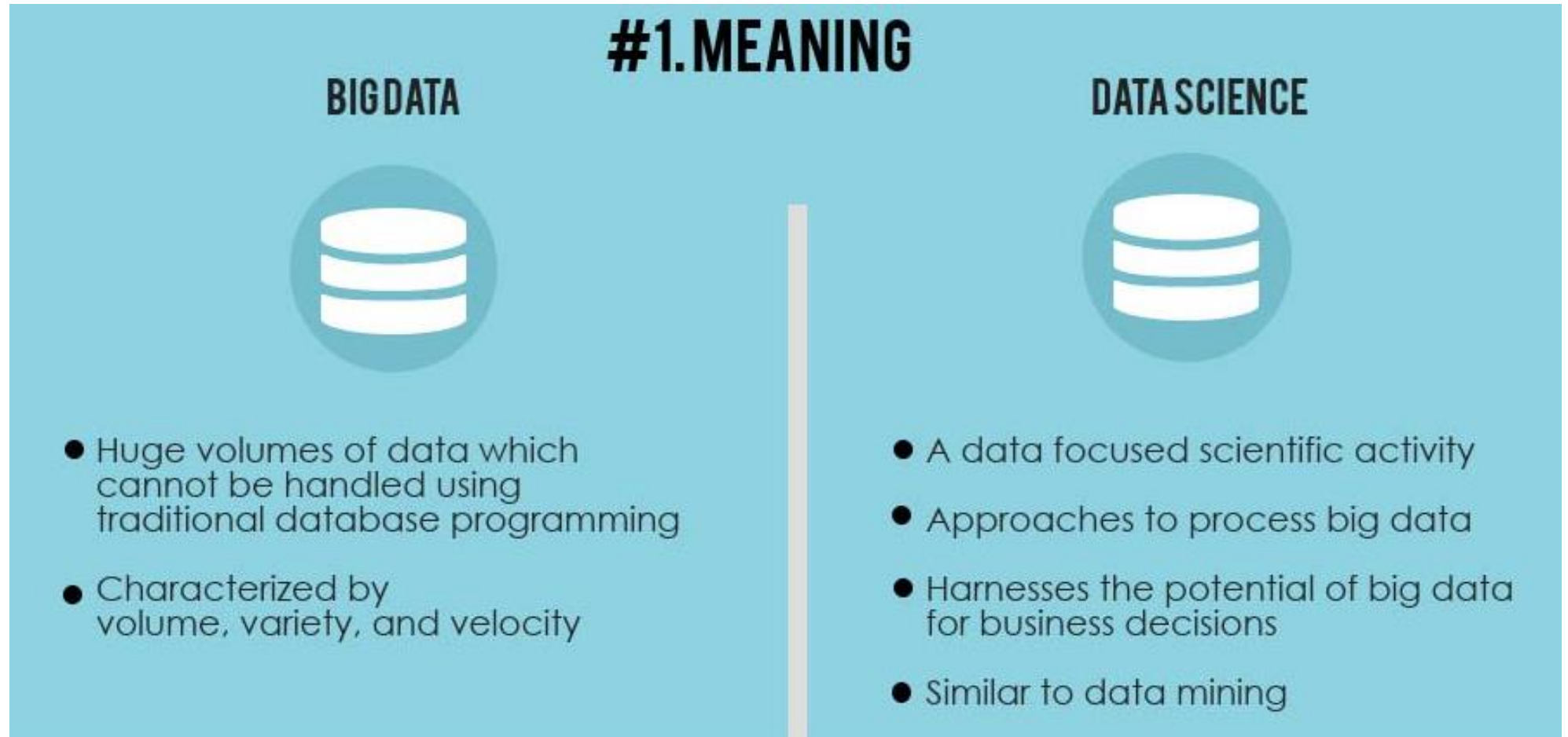
3. Herramientas

4. Aplicaciones y Retos

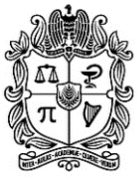
5. Metodología



## Big Data vs. Data Science



Tomado de: <https://www.educba.com/big-data-vs-data-science/>



# Big Data vs. Data Science

## #2. Concept

### BIGDATA



- Diverse data types generated from multiple data sources
- Includes all types and formats of data

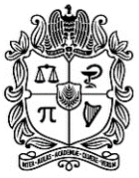
### DATA SCIENCE



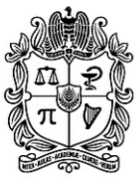
- A specialized area involving scientific programming tools, models and techniques to process big data
- Provides techniques to extract insights and information from large data sets
- Supports organizations in decision making

Tomado de: <https://www.educba.com/big-data-vs-data-science/>

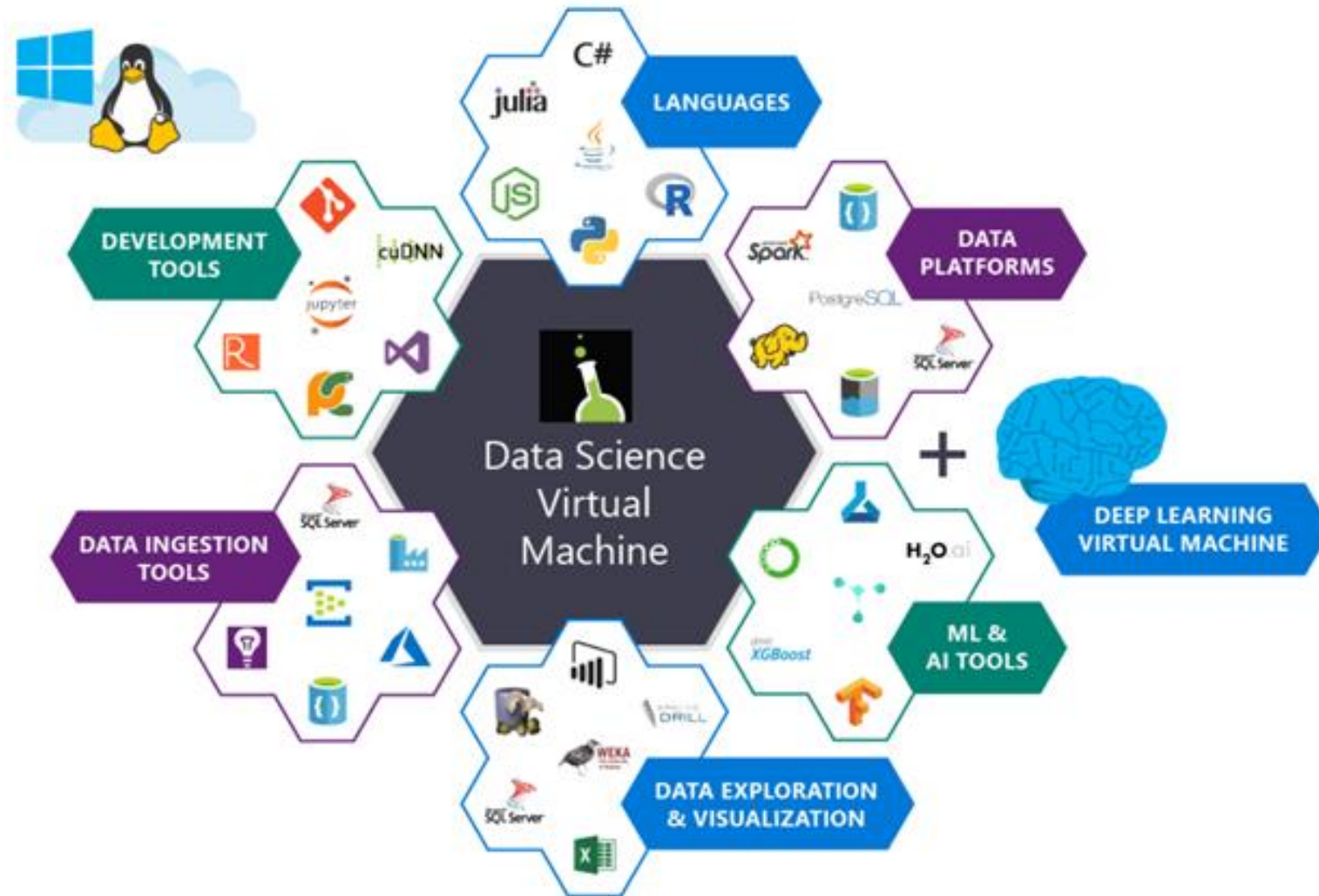


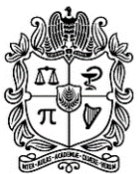


1. ¿Qué es?
2. Big Data vs. Data Science
- 3. Herramientas**
4. Aplicaciones y Retos
5. Metodología

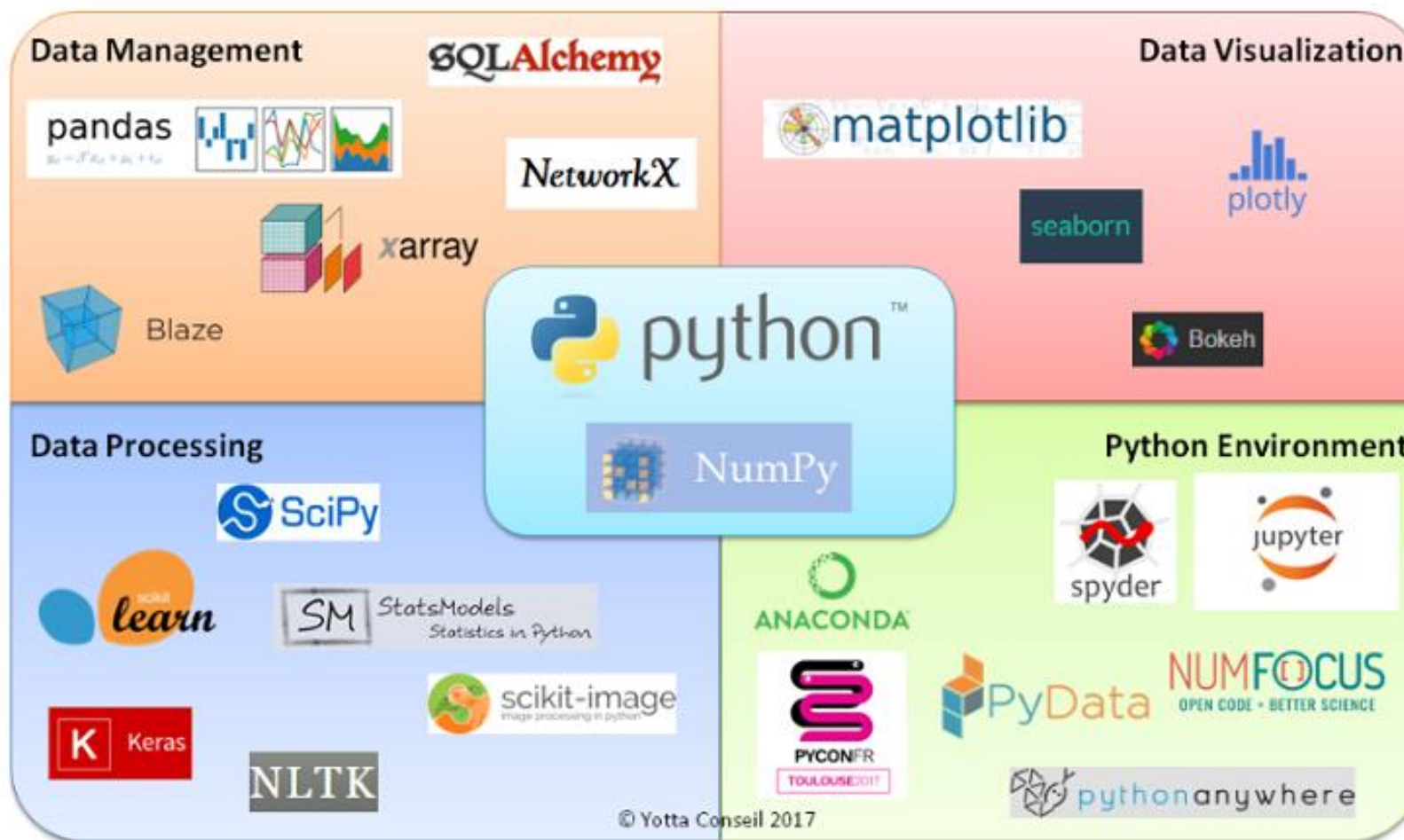


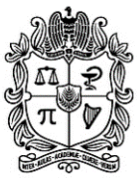
## Herramientas



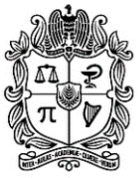


## Herramientas: Ecosistema SciPy

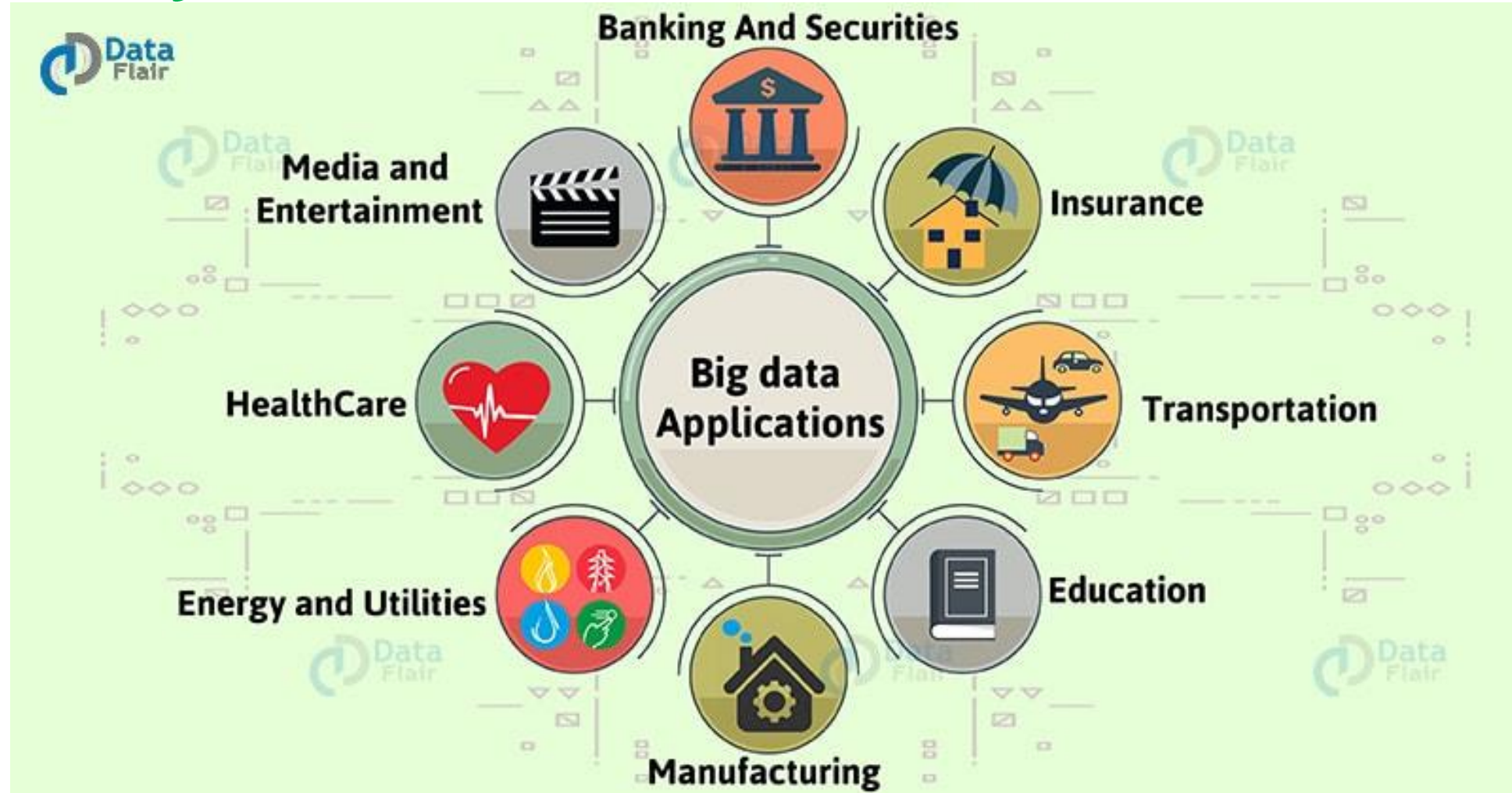




1. ¿Qué es?
2. Big Data vs. Data Science
3. Herramientas
- 4. Aplicaciones y Retos**
5. Metodología

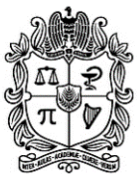


## Aplicaciones y Retos



Tomado de:  
data-flair.training

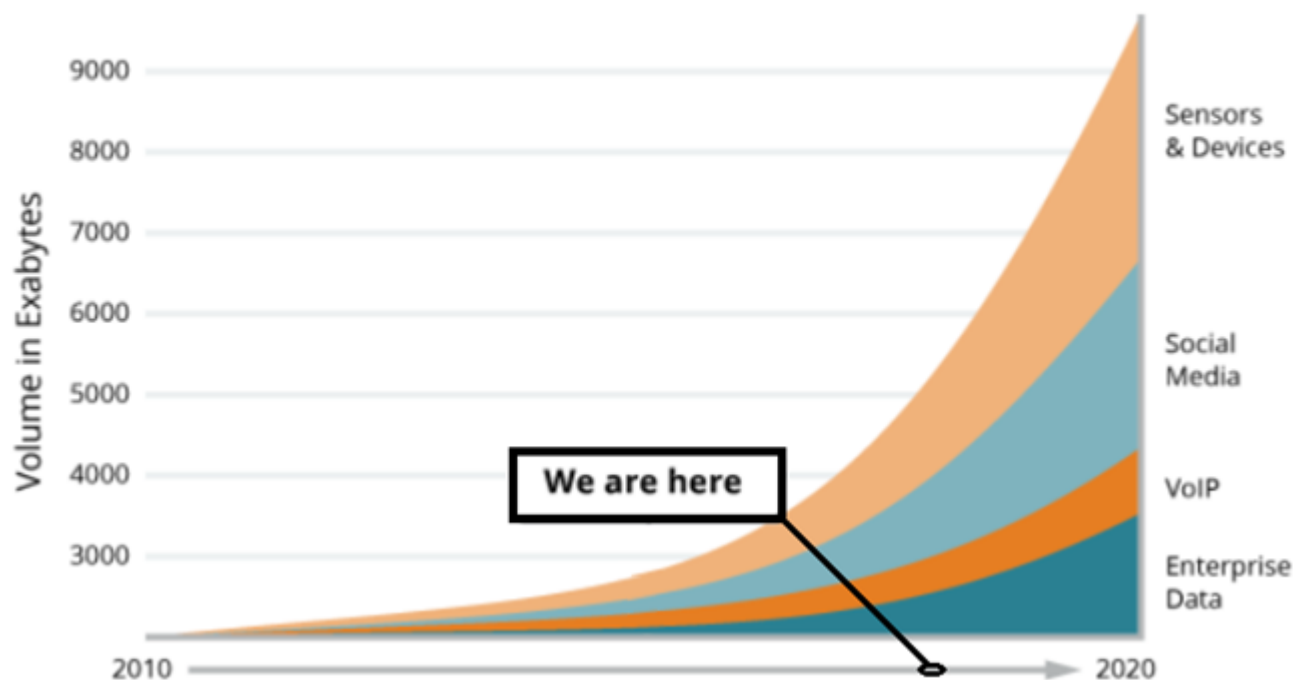




# Aplicaciones y Retos

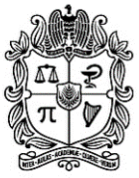
## Hecho: crecimiento de los datos

The Dawn of Big Data: the uncertainty of new information is growing alongside its complexity



Según un artículo de [Forbes](#), los datos están creciendo más rápido que nunca antes. Para el año 2020, se crearán aproximadamente 1.7 MB de nueva información por segundo para cada ser humano en el planeta.

1 EB =  $10^3$  PB =  $10^6$  TB =  $10^9$  GB

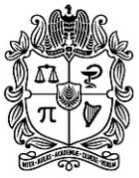


# Introducción al Análisis de Datos

## Aplicaciones y Retos

### Retos:

- Almacenar
- Procesar
- Acceder
- Entender
- Seguridad y privacidad



# Aplicaciones y Retos

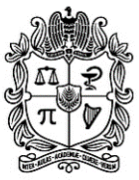
## Retos:

- Almacenar
- Procesar
- Acceder
- Entender
- Seguridad y privacidad



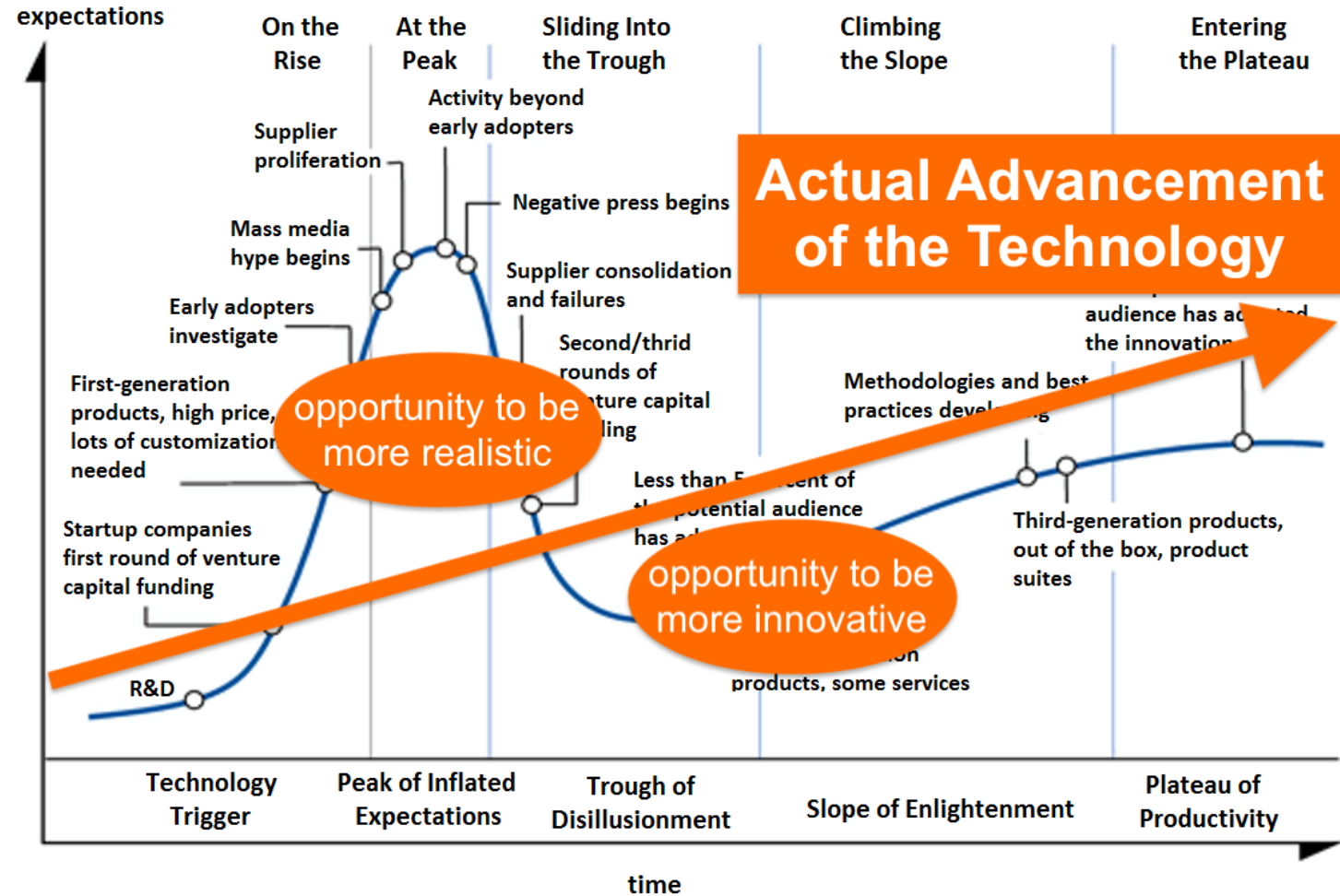
- Analizar
- Cuantificar
- Comparar
- Visualizar
- Predecir

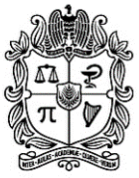




# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

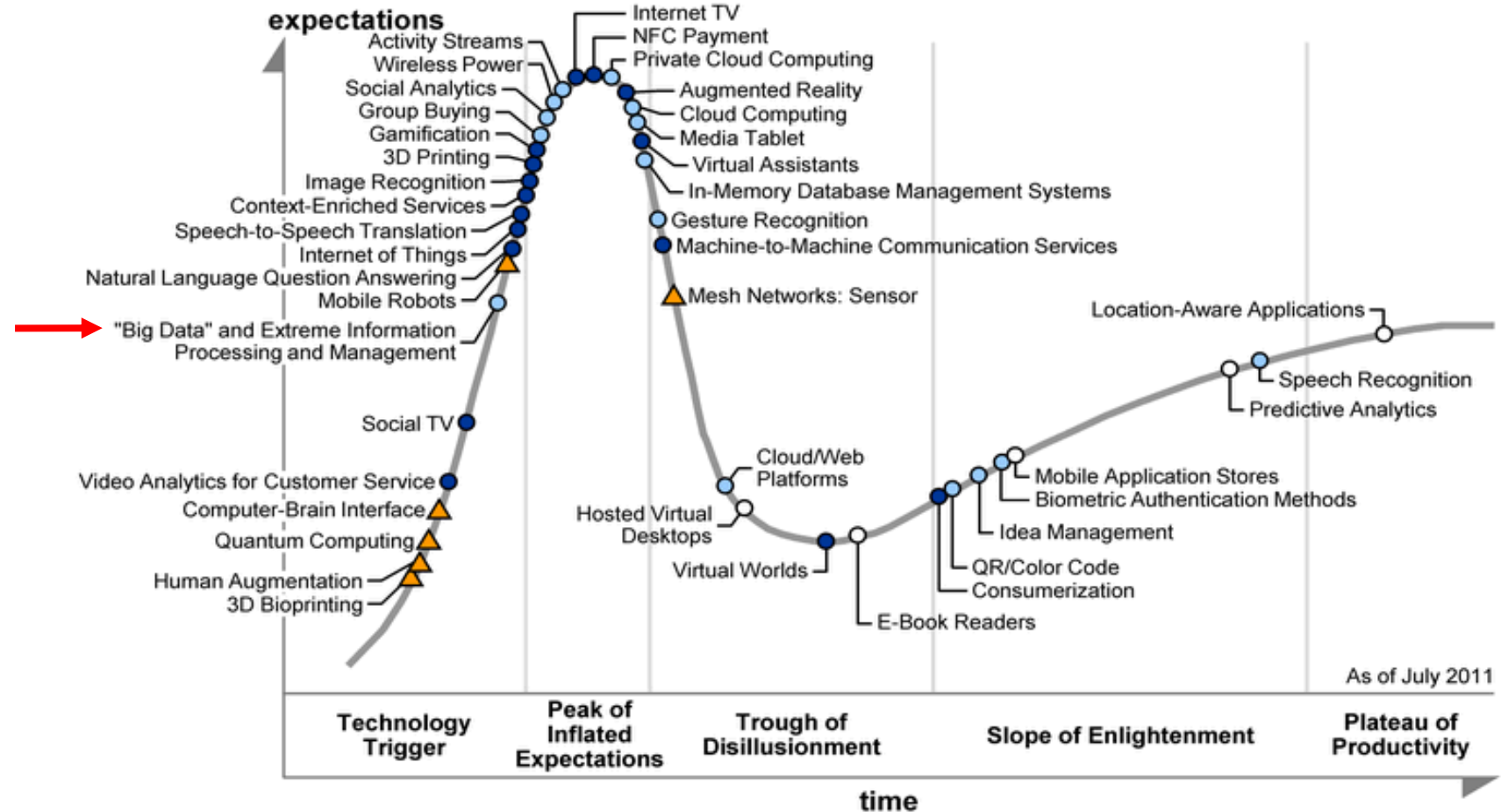




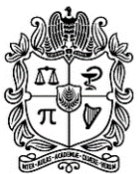
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2011



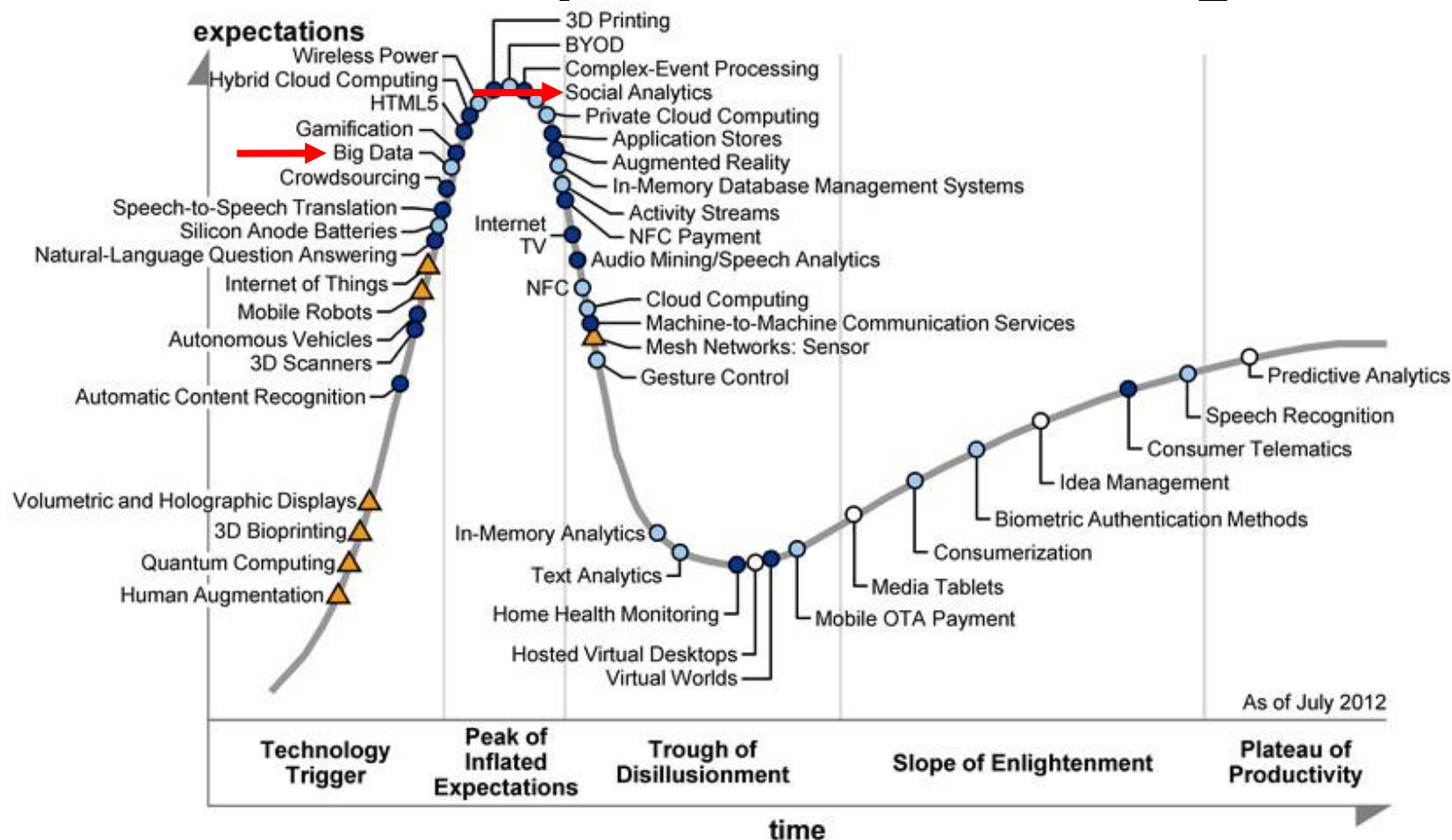
<https://www.gartner.com>



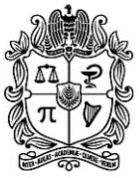
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2012



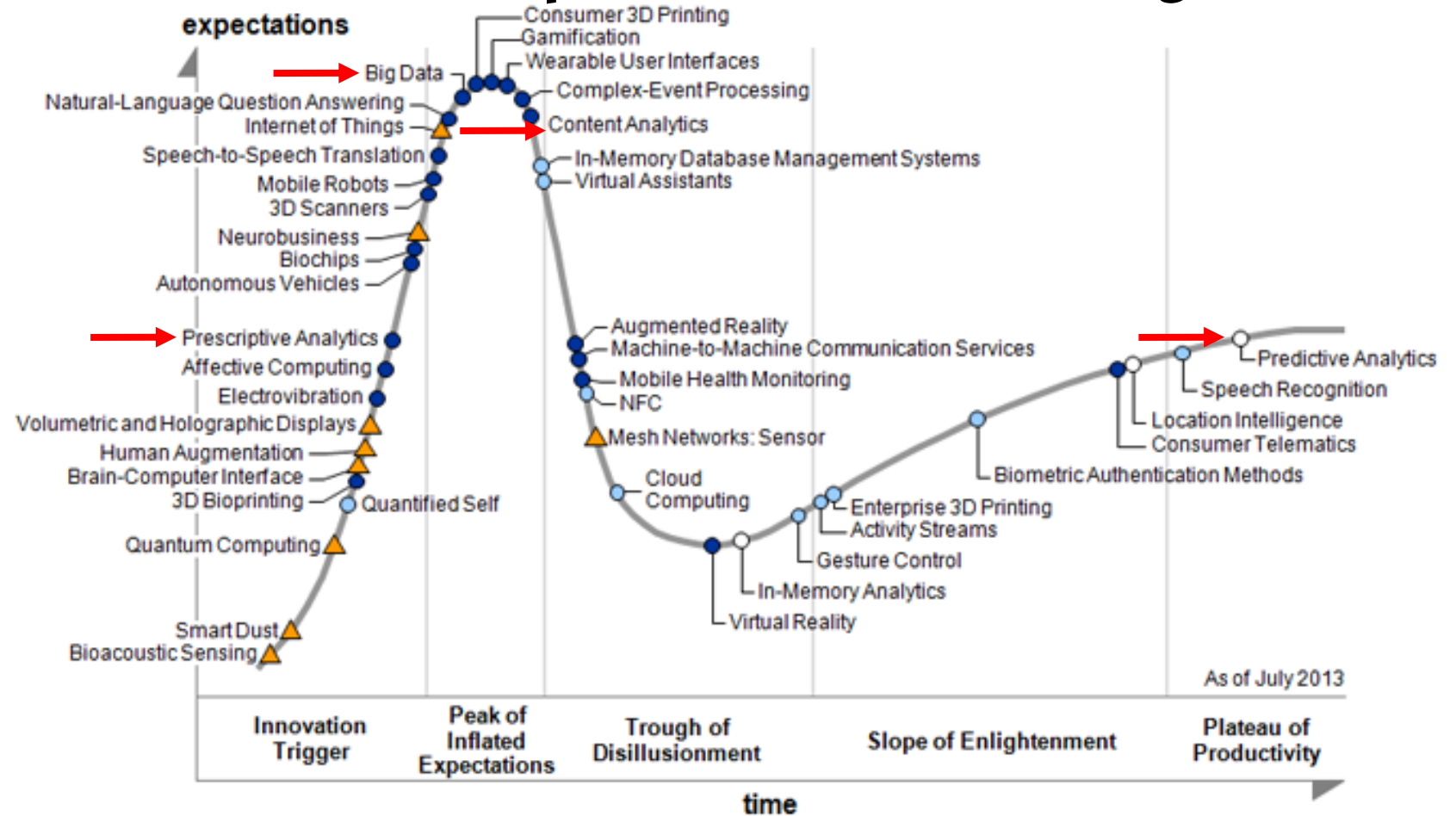
<https://www.gartner.com>



# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2013



<https://www.gartner.com>

Plateau will be reached in:

○ less than 2 years

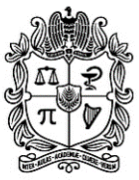
● 2 to 5 years

● 5 to 10 years

▲ more than 10 years

○ obsolete before plateau





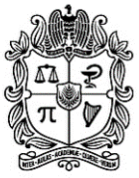
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2014



<https://www.gartner.com>



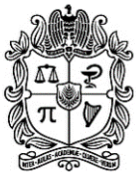
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2015



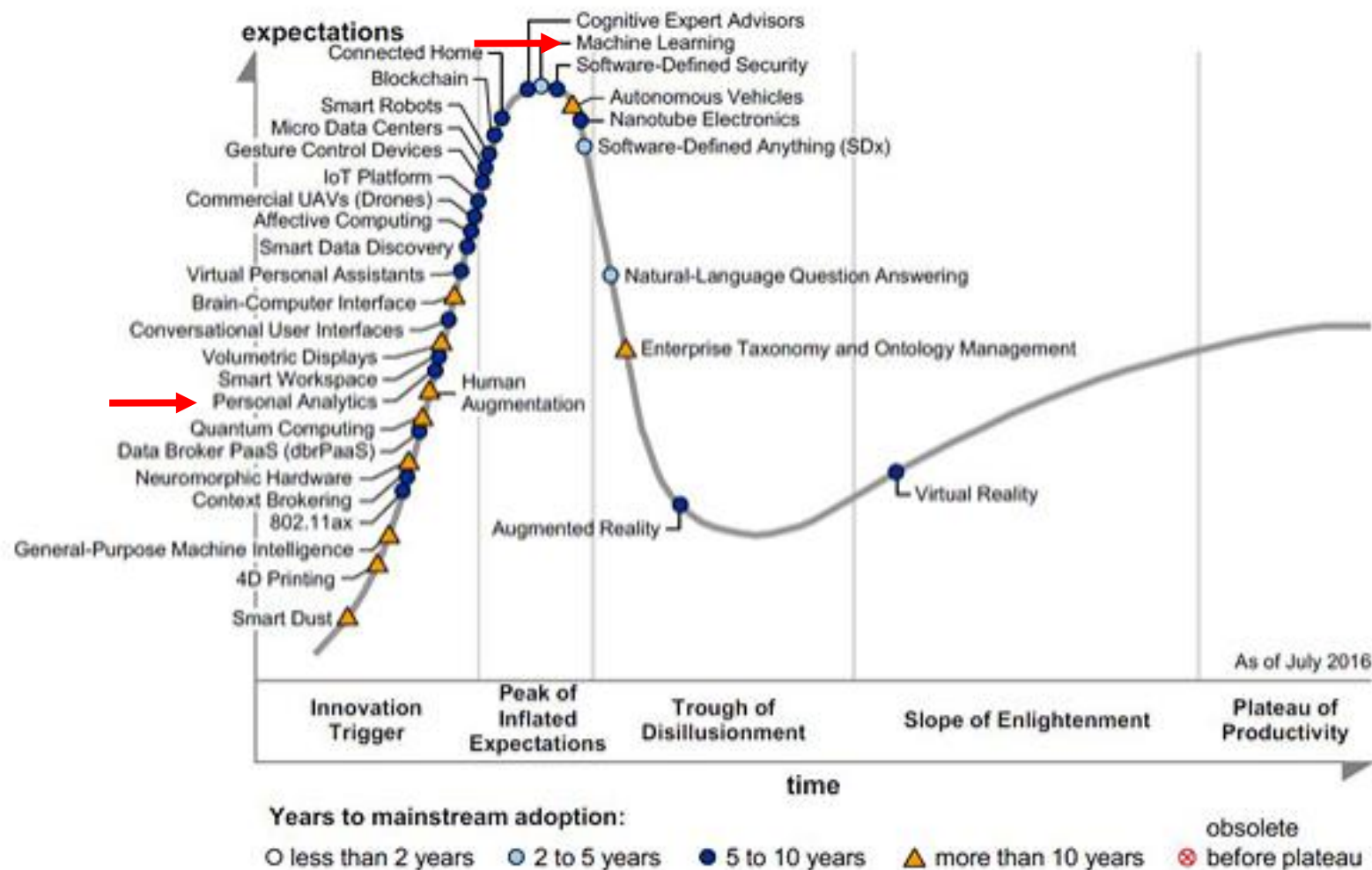
<https://www.gartner.com>



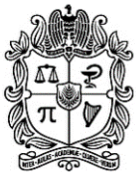
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2016



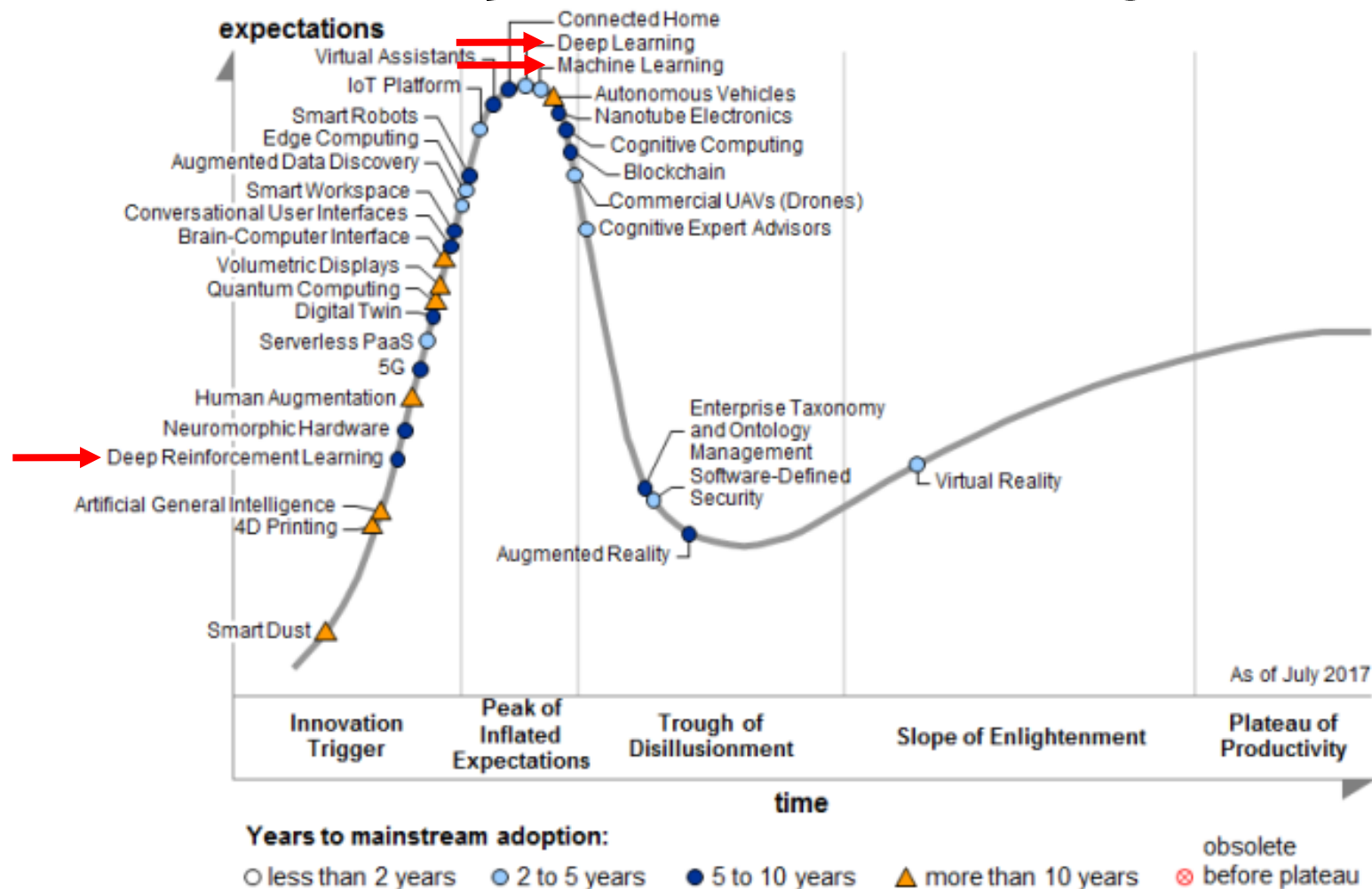
<https://www.gartner.com>



# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

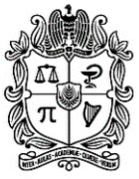
## Gartner's hype cycle

2017



<https://www.gartner.com>

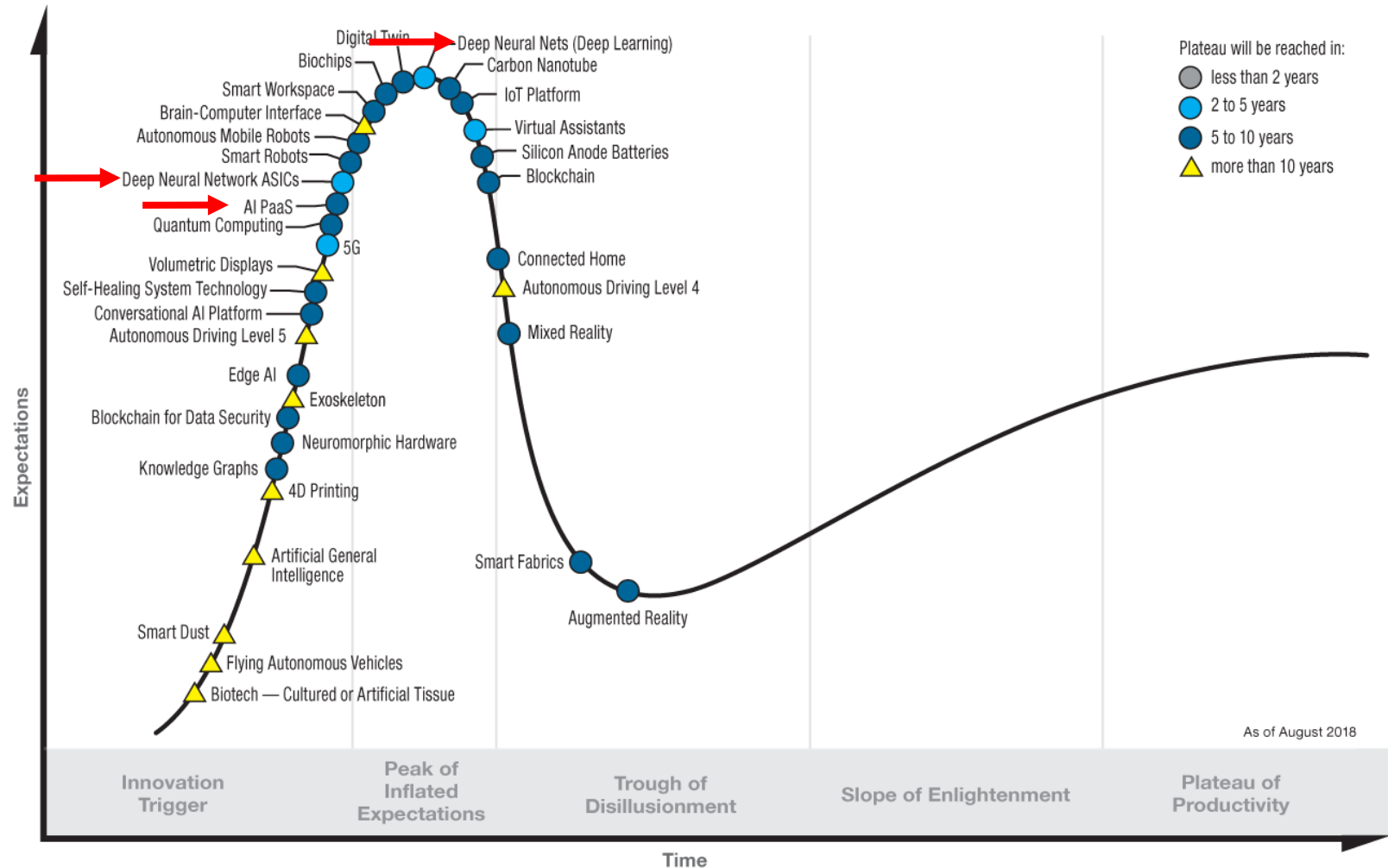




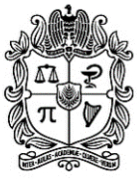
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2018



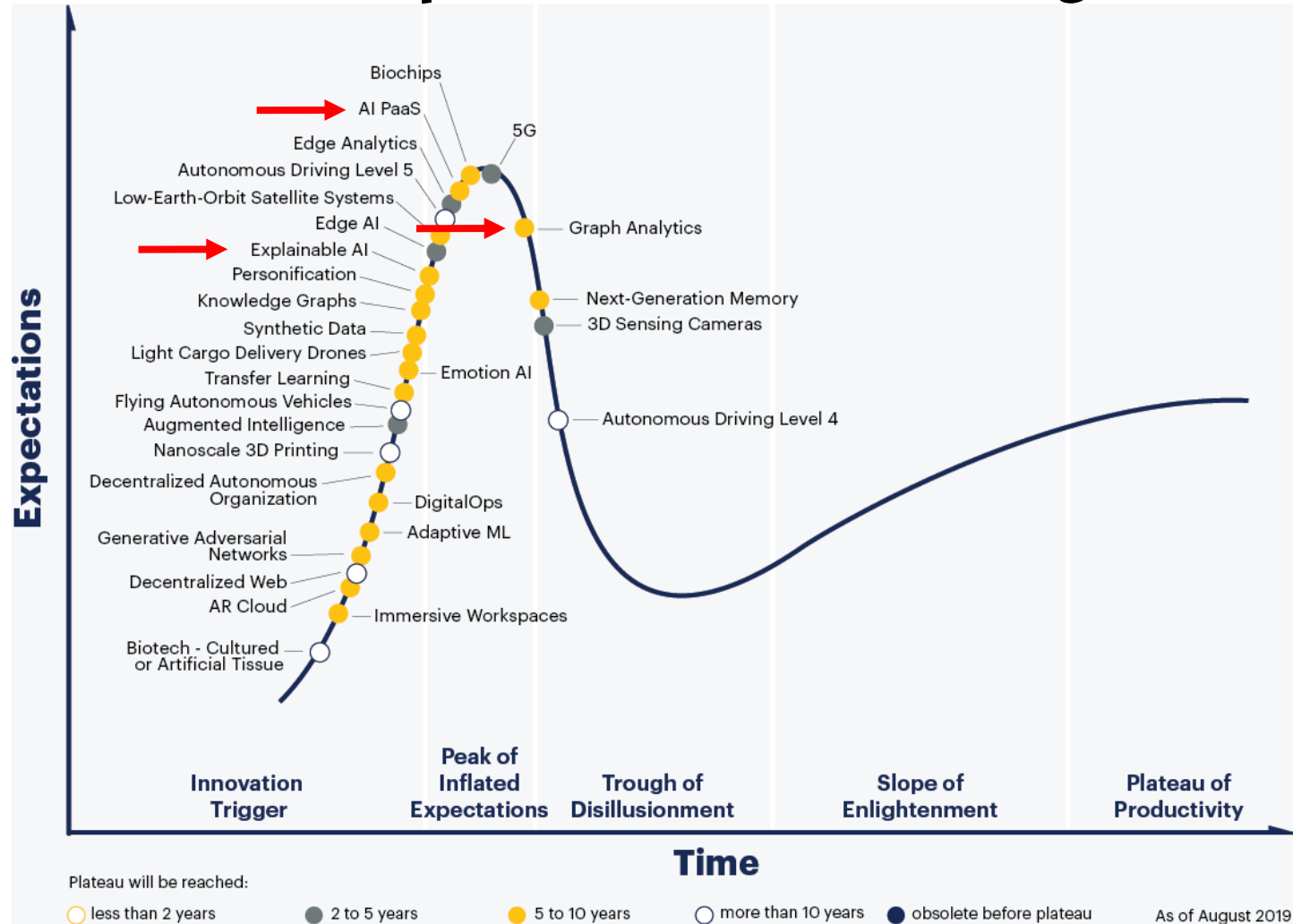
<https://www.gartner.com>



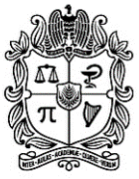
# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

## Gartner's hype cycle

2019



<https://www.gartner.com>

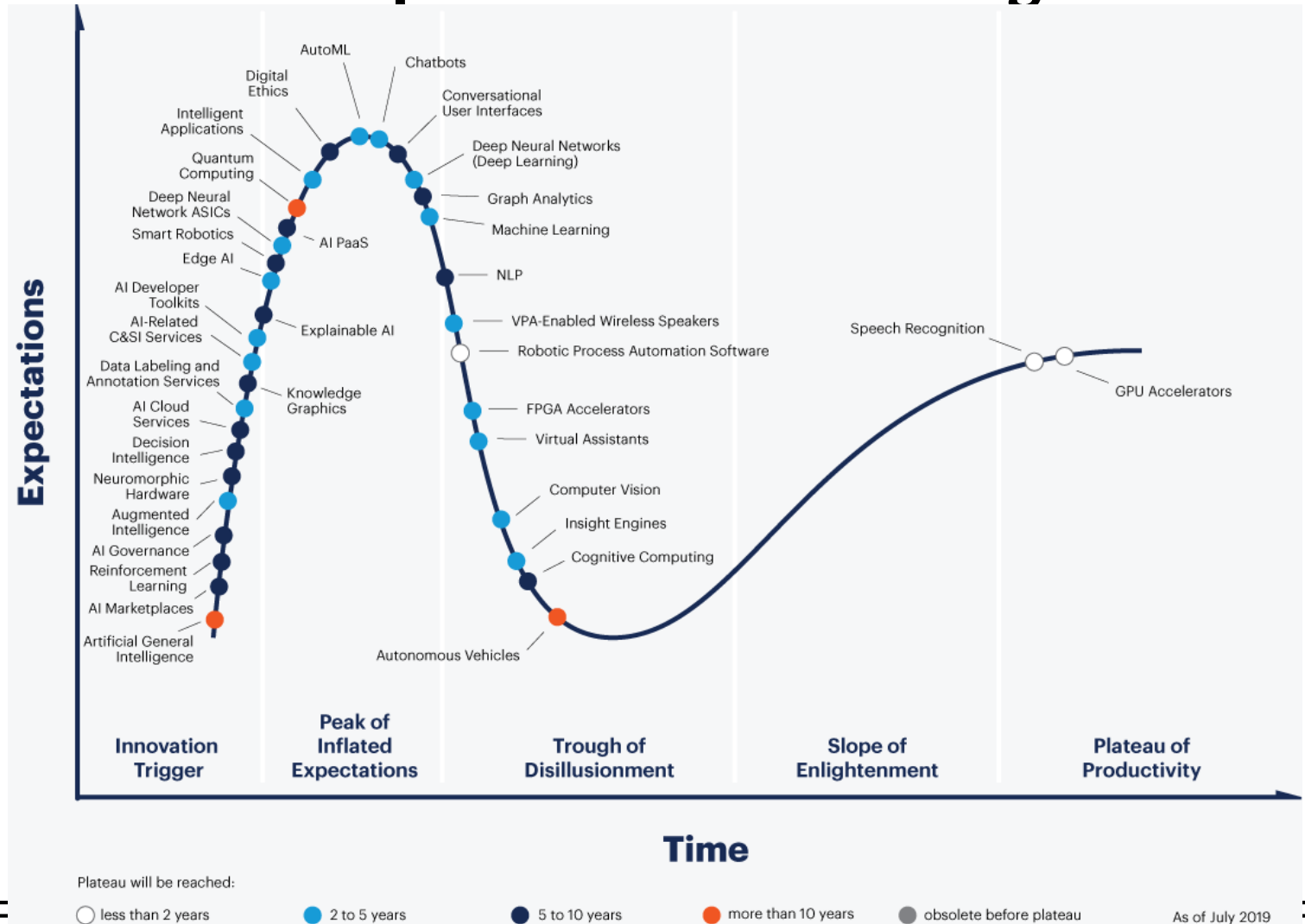


# Aplicaciones y Retos: ciclo de sobre-expectación de tecnologías

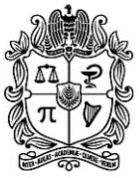
## Gartner's hype cycle

2019

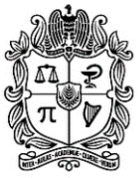
for Artificial Intelligence



<https://www.gartner.com>

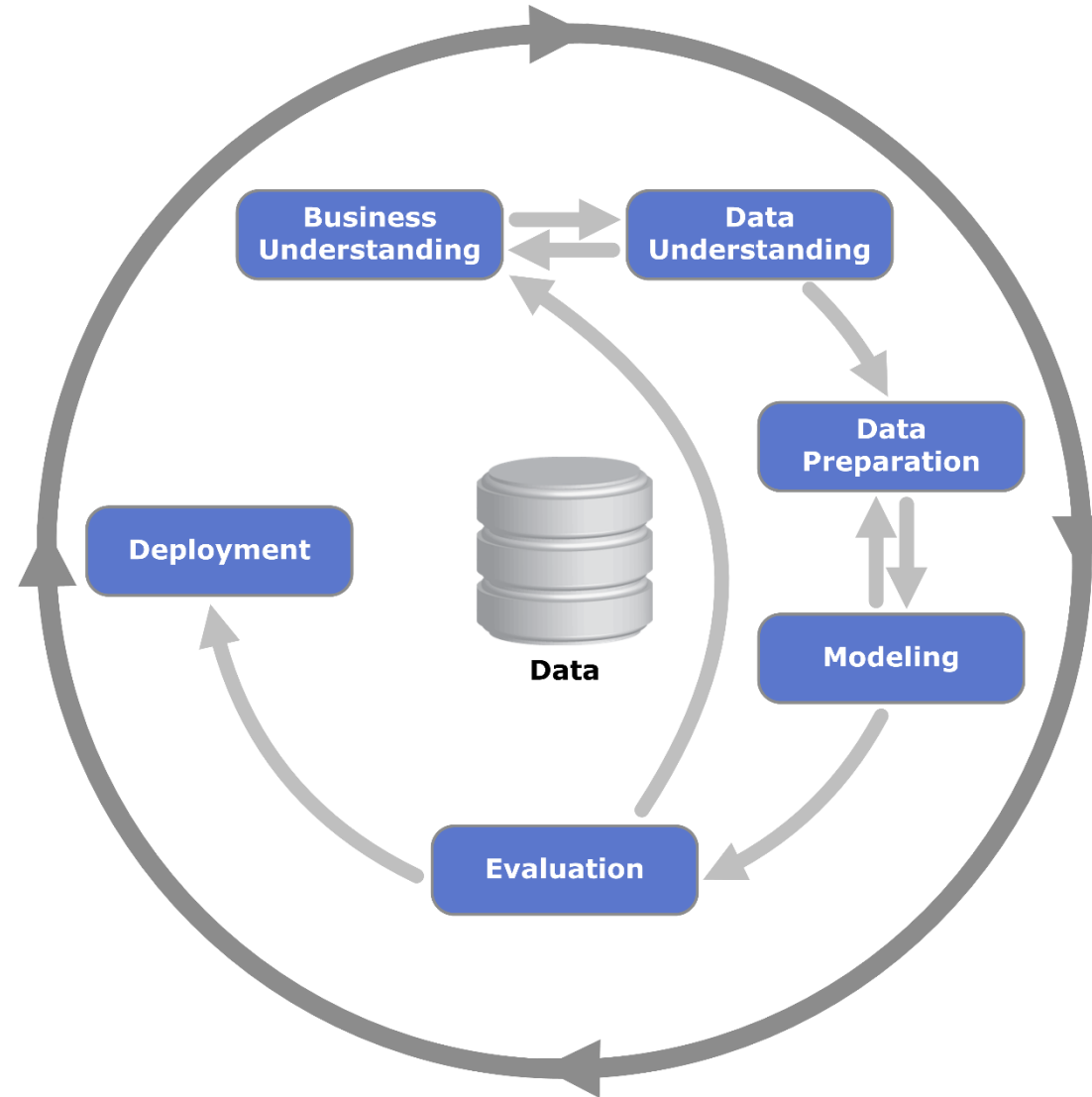


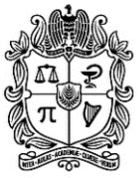
1. ¿Qué es?
2. Big Data vs. Data Science
3. Herramientas
4. Aplicaciones y Retos
- 5. Metodología**



# Metodología: CRISP-DM

## Cross Industry Standard Process for Data Mining

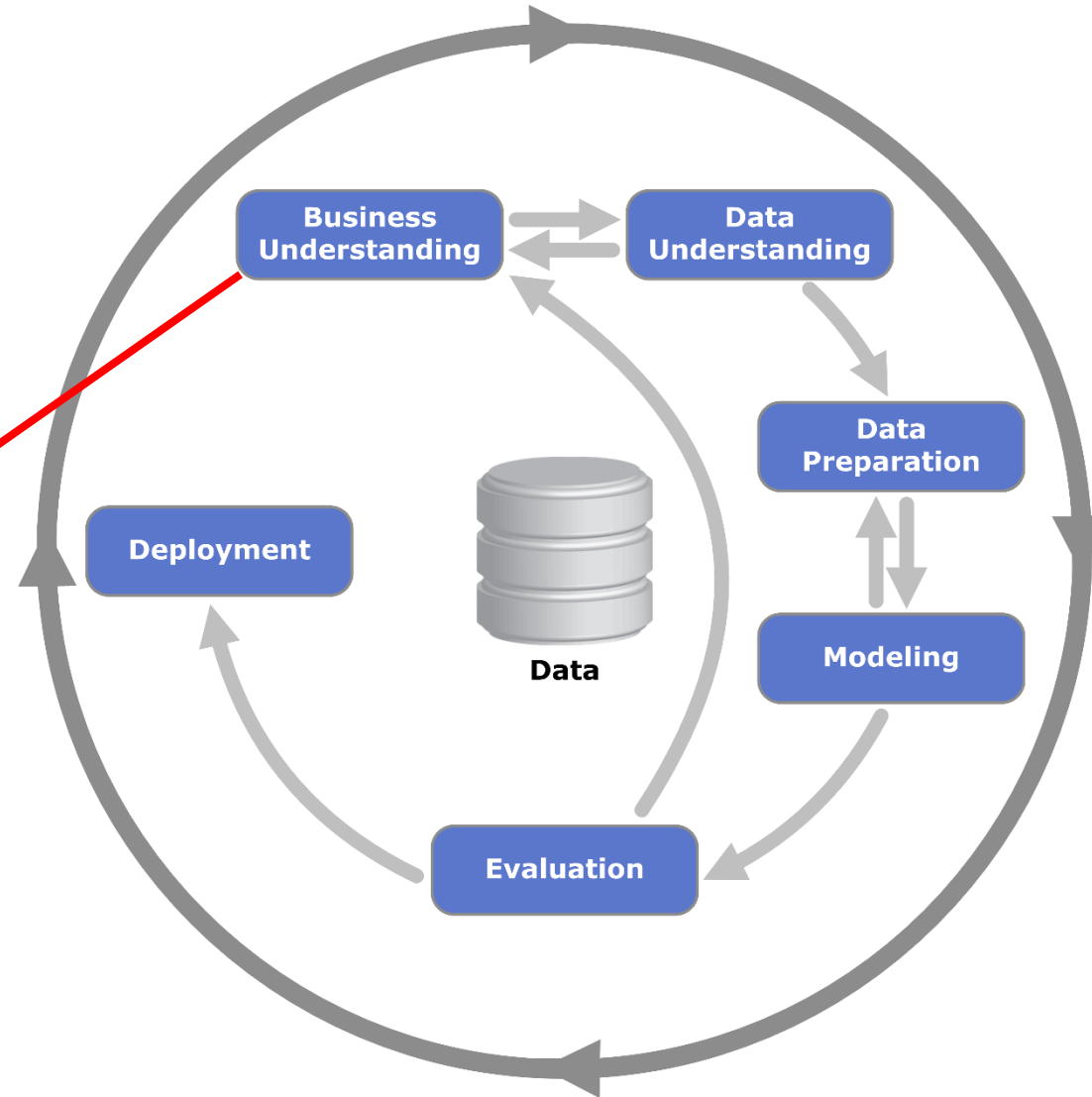


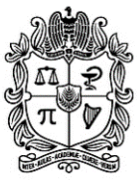


## Metodología: CRISP-DM

### Cross Industry Standard Process for Data Mining

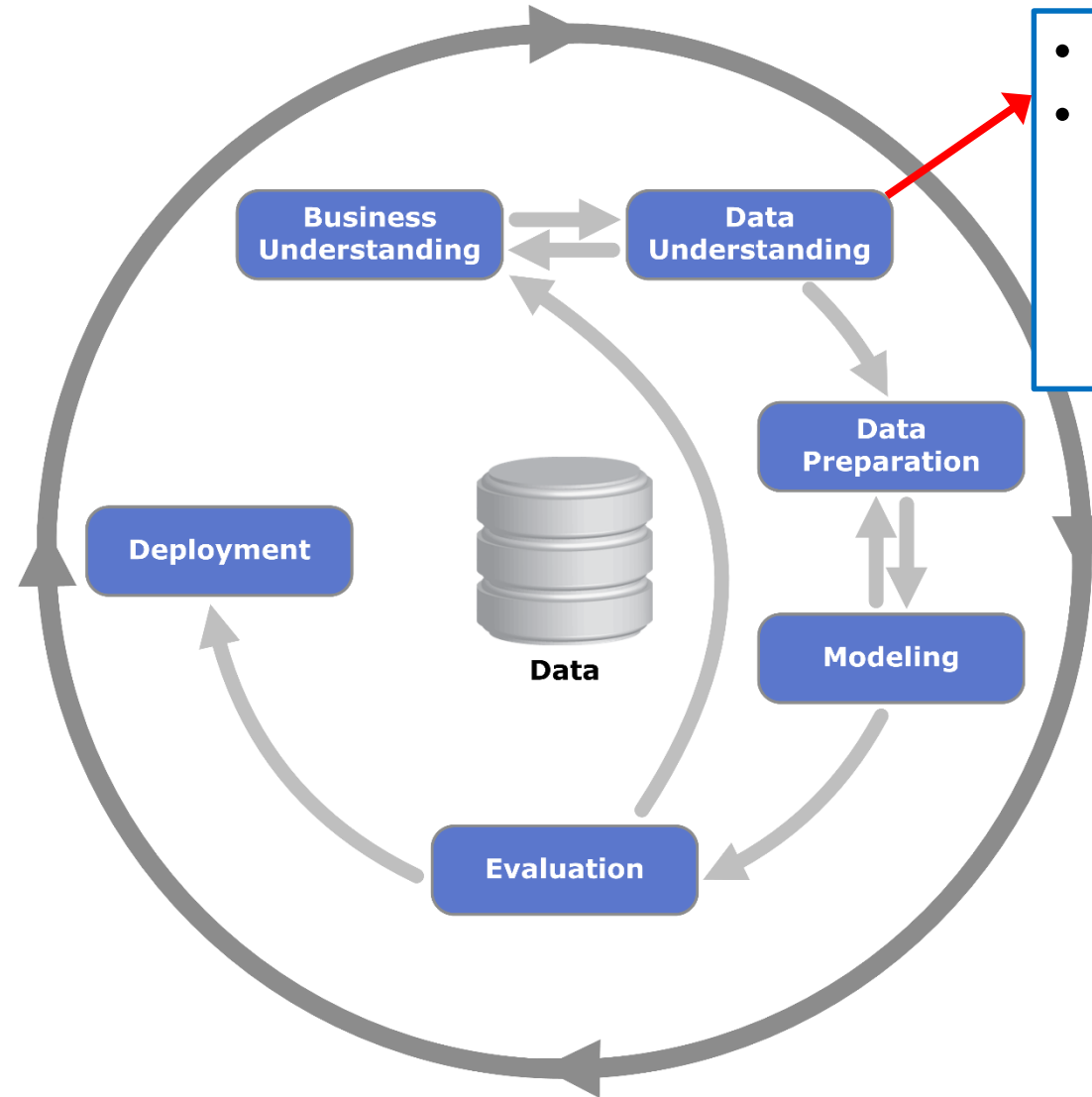
- Definir problema u oportunidad
- Evaluar la situación
- Formular objetivos



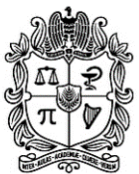


## Metodología: CRISP-DM

### Cross Industry Standard Process for Data Mining

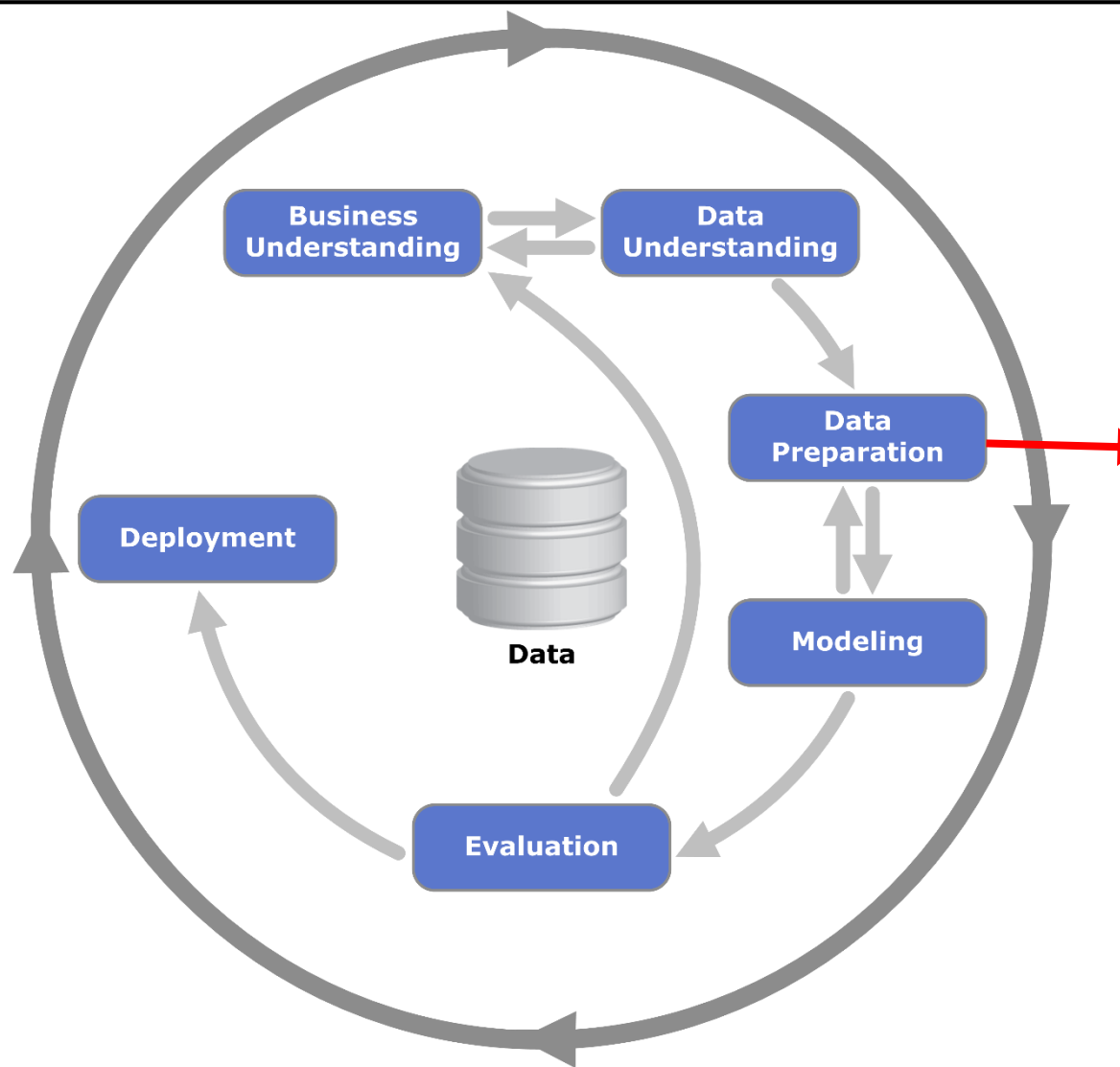


- Adquisición de datos
- Exploración de datos:
  - Estadística descriptiva
  - Distribuciones
  - Visualizaciones



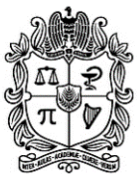
## Metodología: CRISP-DM

### Cross Industry Standard Process for Data Mining



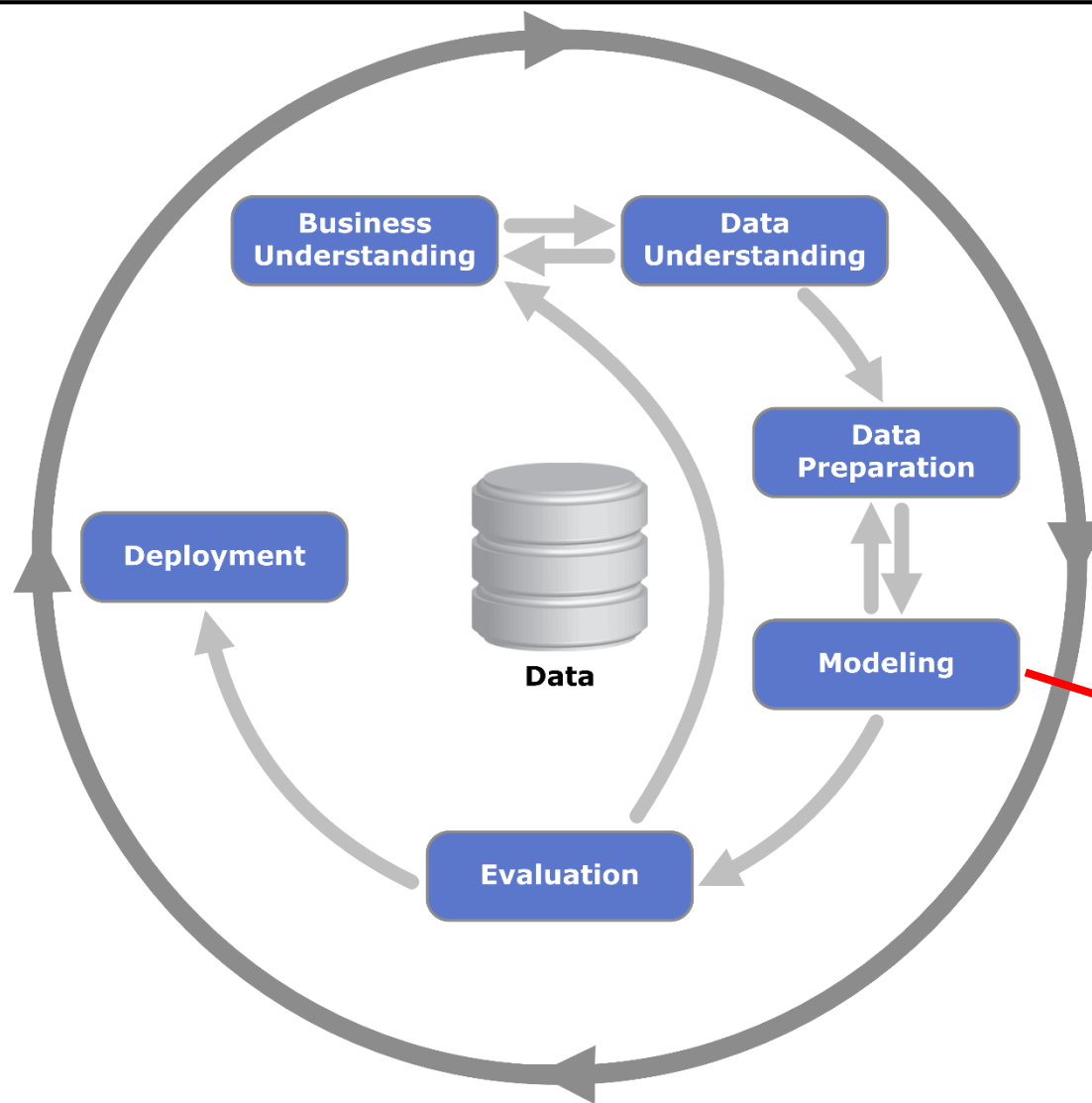
- Preparar datos para modelar
- Abordar problemas de calidad
- Encontrar y caracterizar relaciones
- Seleccionar características
- Procesar datos para modelar



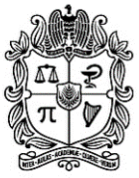


## Metodología: CRISP-DM

### Cross Industry Standard Process for Data Mining

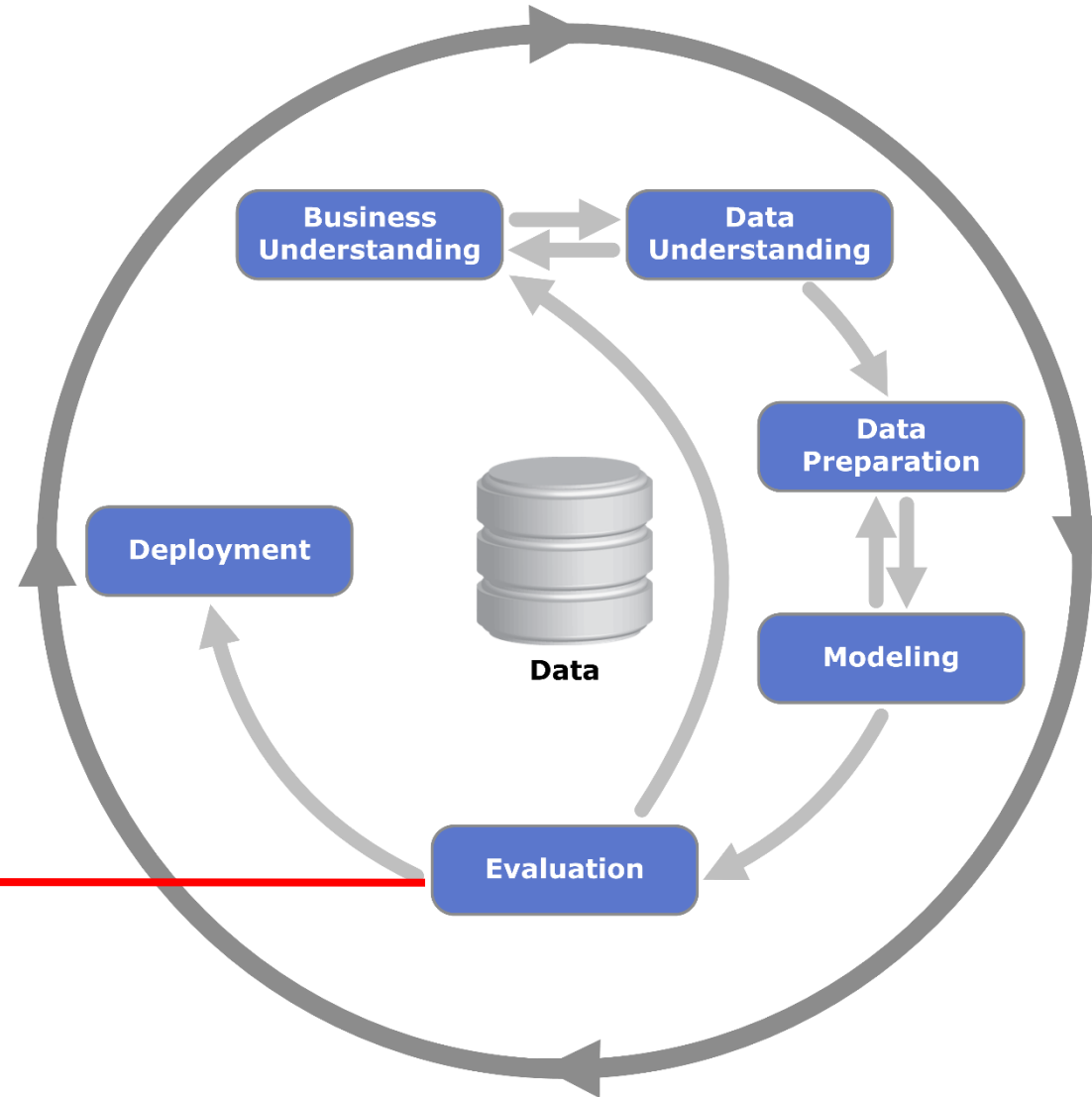


- Determinar el tipo de problema
- Seleccionar las técnicas de modelado
- Construir modelos

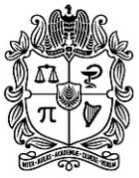


# Metodología: CRISP-DM

## Cross Industry Standard Process for Data Mining



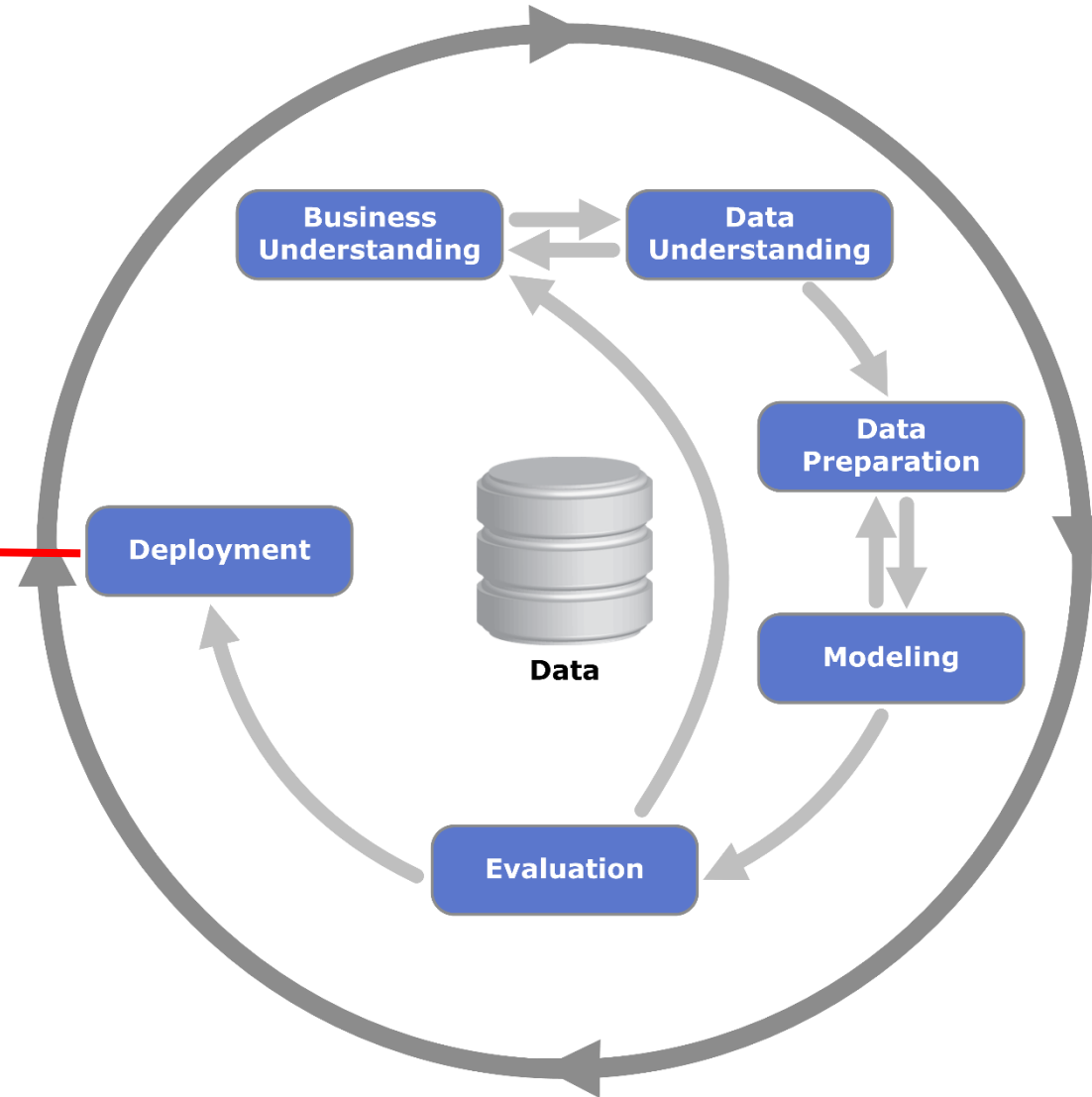
- Evaluar el desempeño del modelo
- Evaluar los resultados con respecto a los criterios de éxito

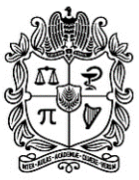


## Metodología: CRISP-DM

### Cross Industry Standard Process for Data Mining

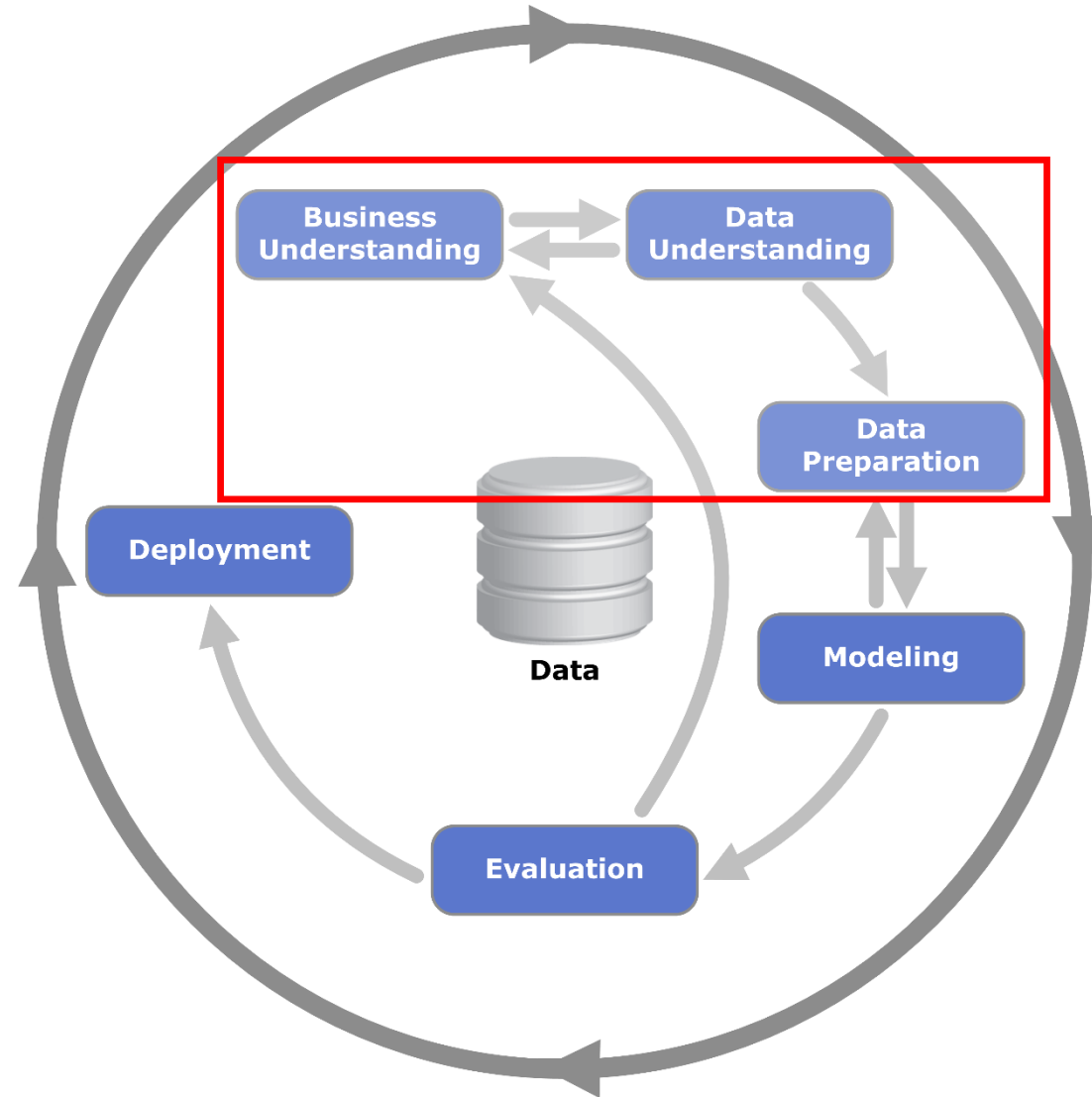
- Desplegar el modelo
- Informe final
- Monitorear el modelo





# Metodología: CRISP-DM

## Cross Industry Standard Process for Data Mining





UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# Introducción al Análisis de Datos

**¡Gracias por su atención!**

**Felipe Restrepo Calle, PhD.**

<https://dis.unal.edu.co/~ferestrepoca/>  
[ferestrepoca@unal.edu.co](mailto:ferestrepoca@unal.edu.co)

Departamento de Ingeniería de Sistemas e Industrial  
Facultad de Ingeniería  
Universidad Nacional de Colombia  
Sede Bogotá