



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



Escola Tècnica  
Superior d'Enginyeria  
Informàtica

# Simulación de datos de sensores industriales

Máster en Big Data Analytics

*Autor:* Pedro Henrique Mano Figueiredo Fernandes

*Tutores:* José Ramón Navarro Cerdán, Francisco Sánchez Cid

# Índice

---

1. Motivación
2. Dataset
3. Reducción de dimensionalidad - PCA
4. Primera simulación - distribución Gaussiana
5. Series temporales - ARIMA
6. Búsqueda del punto más cercano
7. Simulación con datos controlados
8. Arquitectura Big Data
9. Demostración

# Motivación

---

## Problemática:

- **IoT** (*Internet of Things*) en la industria.
- **Sensores** que efectúan mediciones en la cadena de producción.
- **Machine Learning** para extracción de conocimiento de los datos.
- **Gran volumen** de datos para crear modelos matemáticos más fiables.
- **Conocimiento extraído** para anticipación de anomalías o mejoras de rendimiento.
- **Cantidad de datos pequeña** al principio de los proyectos.

## Objetivo:

- **Aprender y simular** datos de sensores industriales.
- Nuevos datos simulados de **mayor volumen**.
- Probar y depurar los algoritmos de **detección de averías** y de **predicción** en entornos Big Data.

# Dataset

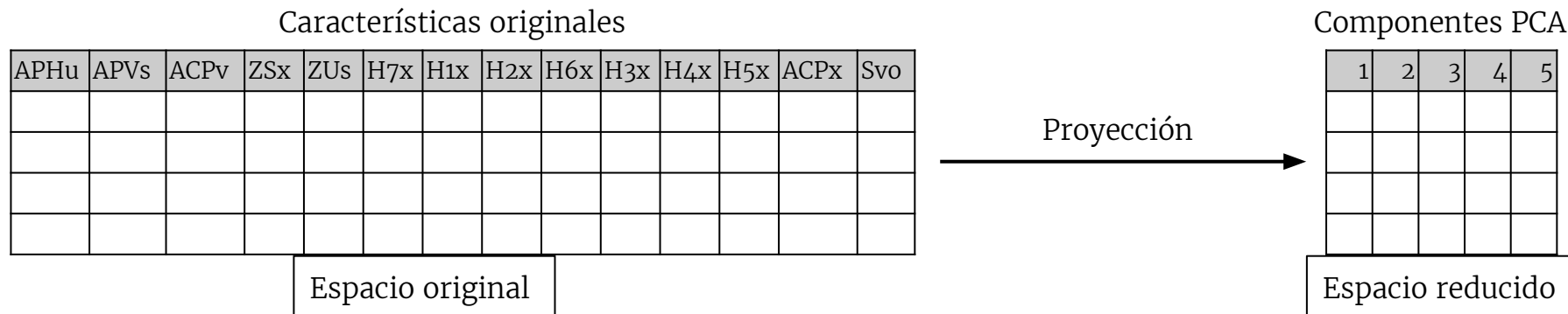
---

- Dataset de reducida dimensión (5MB).
- Poco estructurado.
- Campos separados por espacios y tabulaciones.
- Formatos de fechas inválidos.
- Mediciones en días distintos y con discrepancias.

Tiempoinicio	APHu	APVs	ACPv	ZSx	ZUs	H7x	H1x	H2x	H6x	H3x	H4x	H5x	ACPx	Svo		
06-oct-2015 21:57:03		44.6	69.3	3.81	0.60	8.81	3276.7		44.7	33.2	39.6	37.6	38.5	39.5	3.27	36.00
06-oct-2015 21:57:12		45.1	69.0	3.80	0.60	8.82	3276.7		44.7	33.2	39.6	37.5	38.5	39.5	3.26	36.01
06-oct-2015 21:57:21		44.8	69.8	3.80	0.60	8.84	3276.7		44.7	33.2	39.5	37.5	38.5	39.5	3.40	36.01
06-oct-2015 21:57:30		45.2	68.8	3.81	0.60	8.82	3276.7		44.7	33.2	39.4	37.5	38.5	39.5	3.40	36.00
...																

- Librería para lectura del dataset: pandas.
- Cabecera: Tiempoinicio APHu APVs ACPv ZSx ZUs H7x H1x H2x H6x H3x H4x H5x ACPx Svo
- Campo Tiempoinicio: formato fecha.
- Otros campos: formato decimal.

# Reducción de dimensionalidad - PCA



Aportación de PCA en este estudio:

- Simulación
- Reducción de ruido
- Reducción de redundancia

- Estrategia PCA: NIPALS
- Numero de componentes a usar basado en los resultados del test T2 de Hotelling

# Primera simulación - distribución Gaussiana

## Componentes PCA

1	2	3	4	5

# Simulación con distribución Gaussiana

## Componentes PCA simuladas

[illegible]

- Las componentes son variables independientes, se pueden simular nuevos valores de forma independiente para cada componente.

## Primera simulación - reproyección al espacio original

## Componentes PCA simuladas

[illegible]

## Reproyección al espacio original



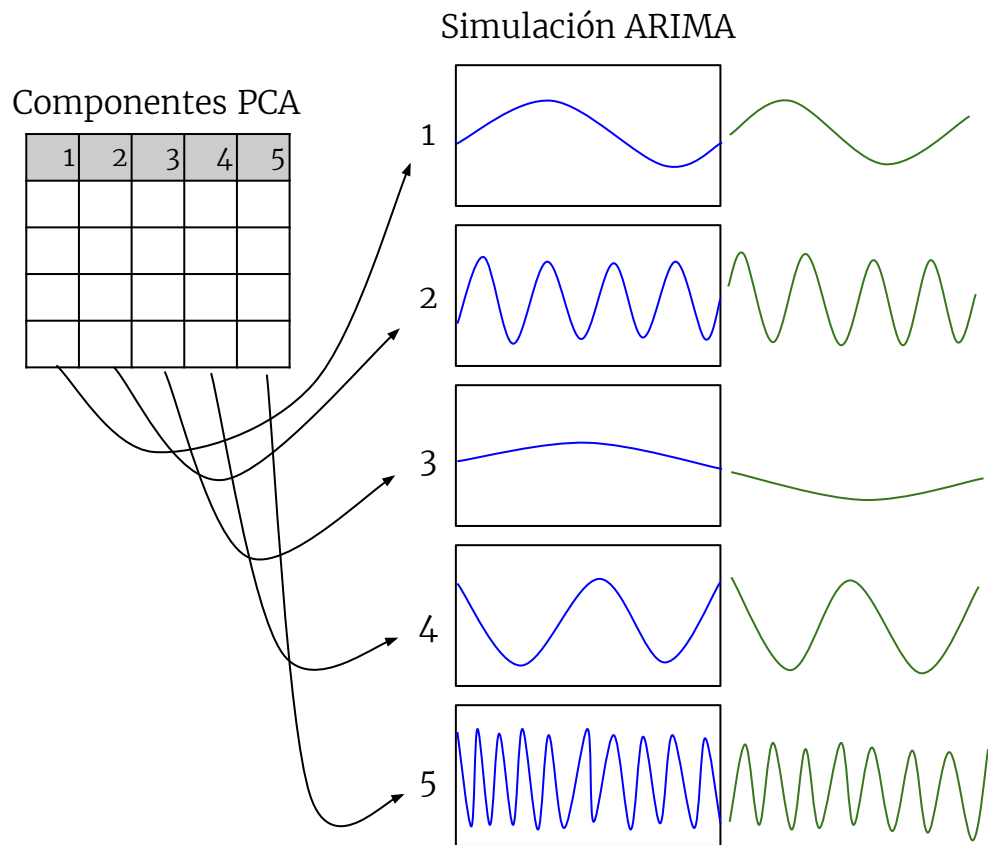
## Características originales simuladas

[illegible]

## Test T2 de Hotelling



# Series temporales - ARIMA



Aportación de ARIMA en este estudio:

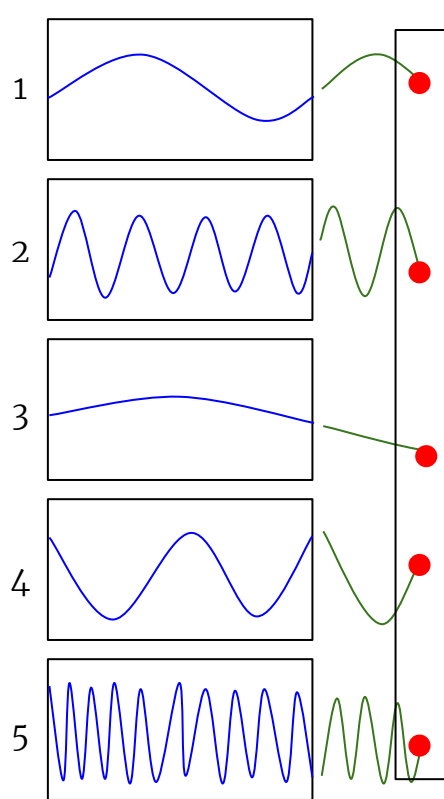
- Dotar la simulación de un patrón temporal aprendido previamente.

- Librería genérica para series temporales: `pandas`
- Librería para entrenamiento y predicción ARIMA: `statsmodels`



# Búsqueda del punto más cercano

Simulación ARIMA



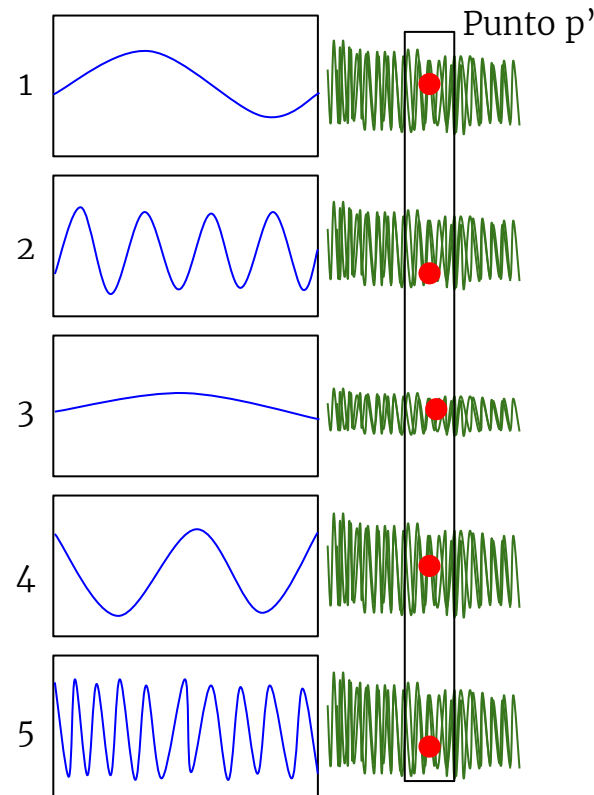
Punto p

Buscar el punto más cercano a p

Reemplazar el punto p por p'

• Distancia Euclídea

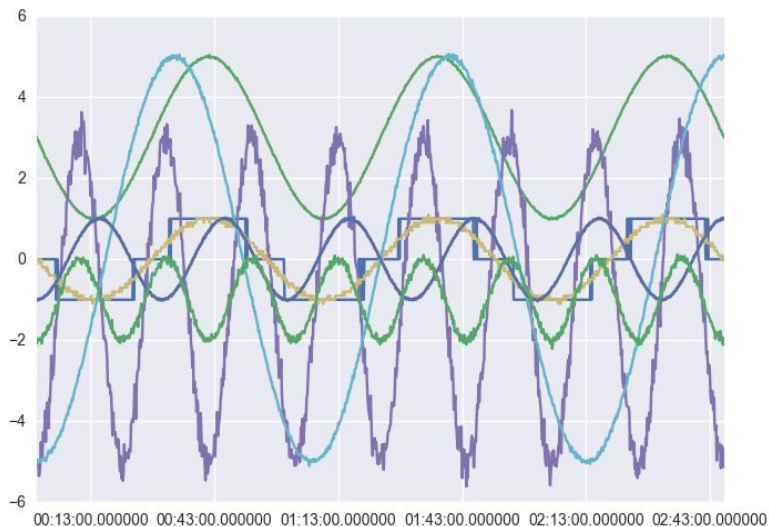
Simulación Gaussiana



Punto p'

# Simulación con datos controlados - PCA

Datos de funciones de senos y cosenos



Espacio original

Proyección

Componentes PCA

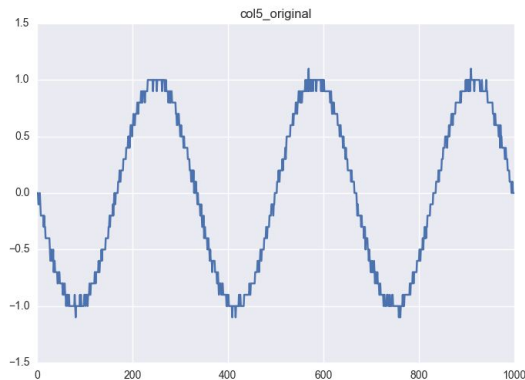
1	2	3	4	5

Espacio reducido

# Simulación con datos controlados - PCA

Ejemplo:  $\sin-1$  con ruido 0.05

$\sin-1$  con ruido 0.05



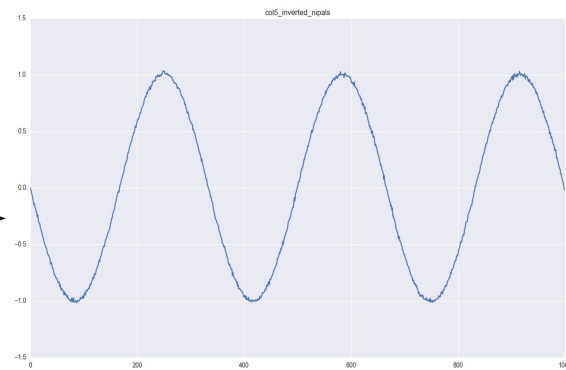
Proyección

Componentes PCA

1	2	3	4	5

Espacio reducido

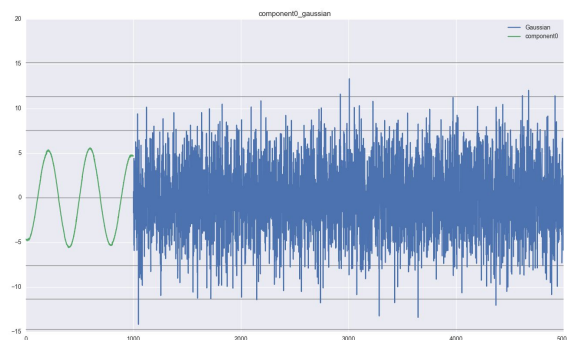
Reproyección



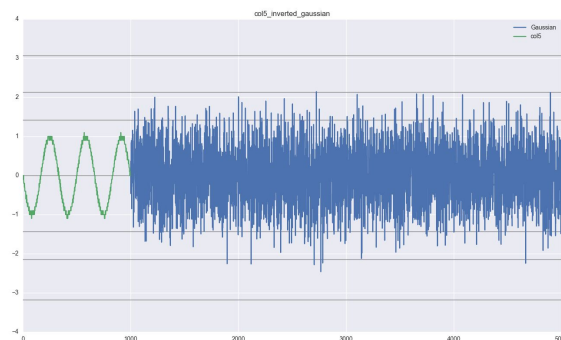
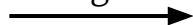
- La inversión se hace correctamente
- Se reduce el ruido de la señal.

# Simulación con datos controlados - PCA

Ejemplo:  $\sin-1$  con ruido 0.05



Reproyección  
al espacio  
original



Test T2 de  
Hotelling

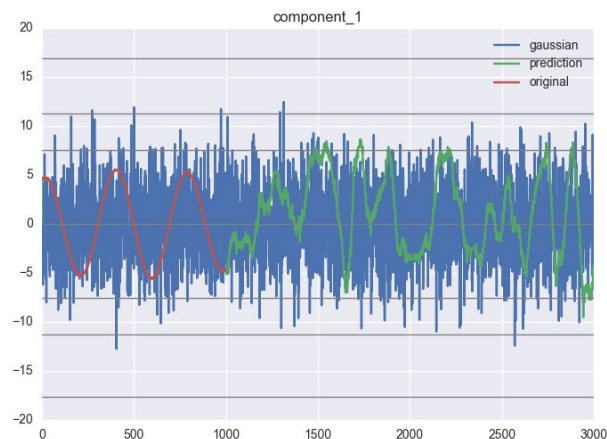


Simulación de la primera componente,  
usando distribución Gaussiana

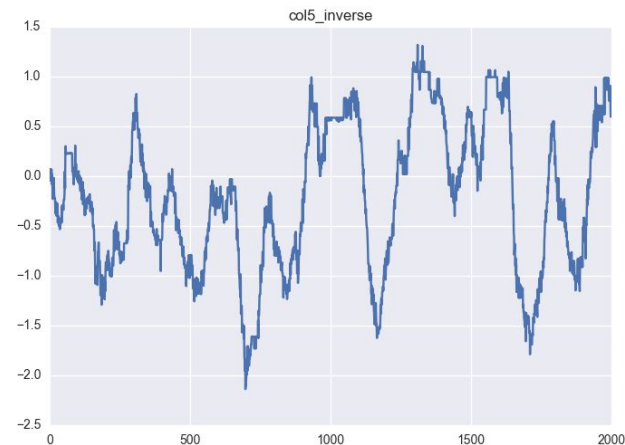
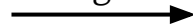
Inversión de  $\sin 1$  con ruido 0,05,  
usando distribución Gaussiana

# Simulación con datos controlados - ARIMA

Ejemplo:  $\sin^{-1}$  con ruido 0.05



Reproyección  
al espacio  
original

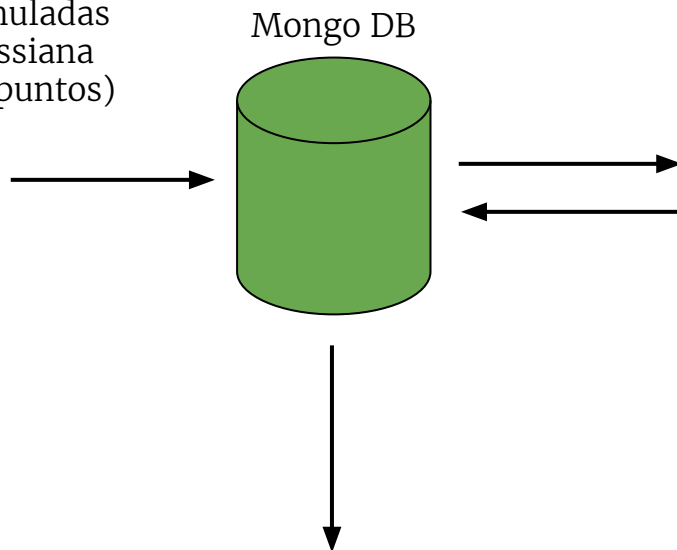
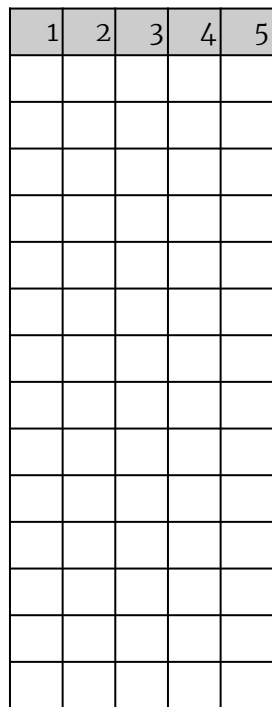


Simulación de la componente 1,  
usando ARIMA y búsqueda de  
puntos en la simulación  
Gaussiana

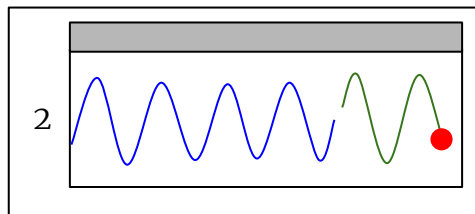
Inversión de  $\sin^{-1}$  con ruido 0.05,  
usando ARIMA y búsqueda de  
puntos en la simulación Gaussiana

# Arquitectura Big Data

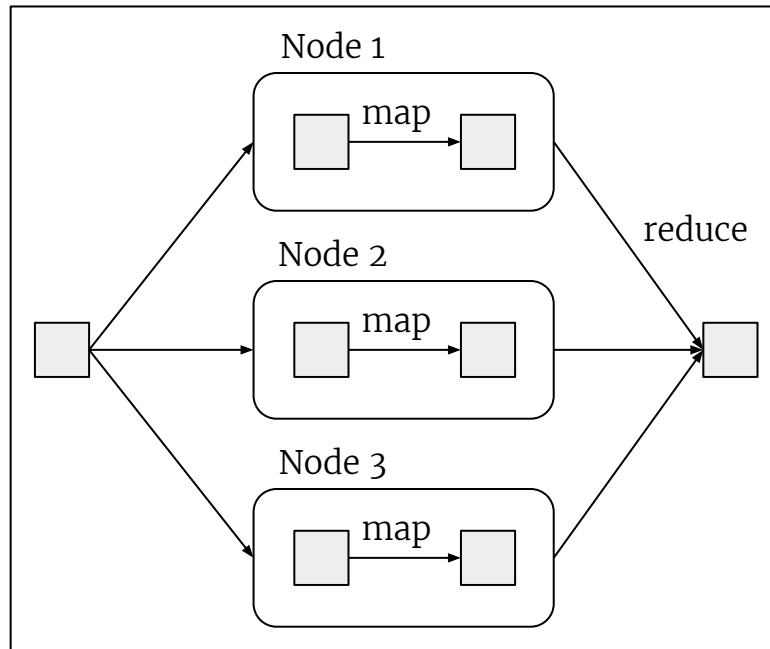
## Componentes PCA simuladas con distribución Gaussiana (miles de millones de puntos)



# Python + Matplotlib



# Python + Spark



# Conclusiones

---

## Conclusiones:

- **PCA** se puede usar para simulación.
- Cuantos más componentes:
  - Mayor **similitud** entre los datos invertidos y los datos originales.
  - Mayor **ruido** de la señal.
- Datos de simulación **Gaussiana** permiten contener la predicción ARIMA.
- Cuantos **más datos** de simulación Gaussiana mejor.
- **Big Data** viabiliza la solución de búsqueda de puntos cercanos.

## Trabajos futuros:

- Simulación con diferente número de componentes.
- Pruebas con distancia de Mahalanobis.
- Mejoras de la herramienta de visualización con matplotlib.
- Estudio de tiempos de ejecución y consumo de memoria.

# Demostración

- Visualización del proceso de simulación en tiempo real.
- Librería de gráficos: `matplotlib`
- MongoDB como *buffer* de datos.
- Datos de simulación de componentes PCA y de la respectiva inversión.

