ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Diplomado en Big Data y Ciencia de Datos
Ciencia de Datos y sus Aplicaciones

Clase 05:

Modelos Predictivos y Series de Tiempo

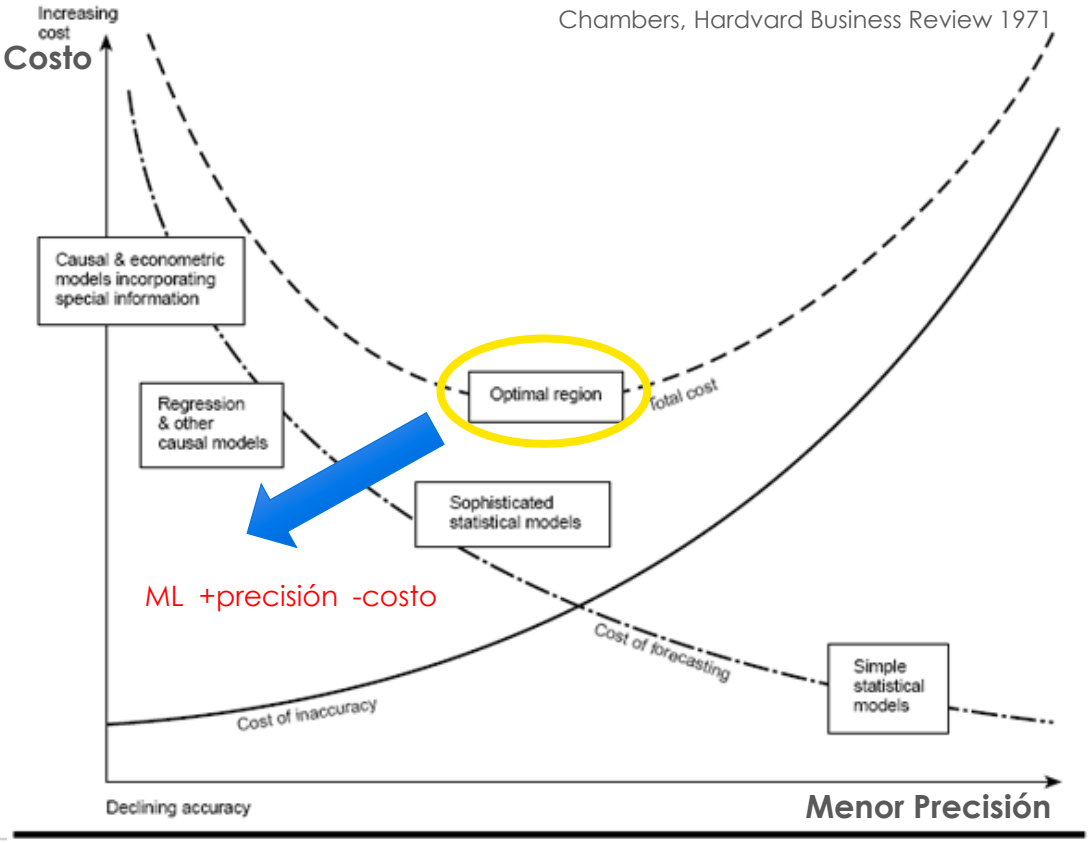
Roberto González

regonzar@uc.cl

Forecasting y Series de Tiempo

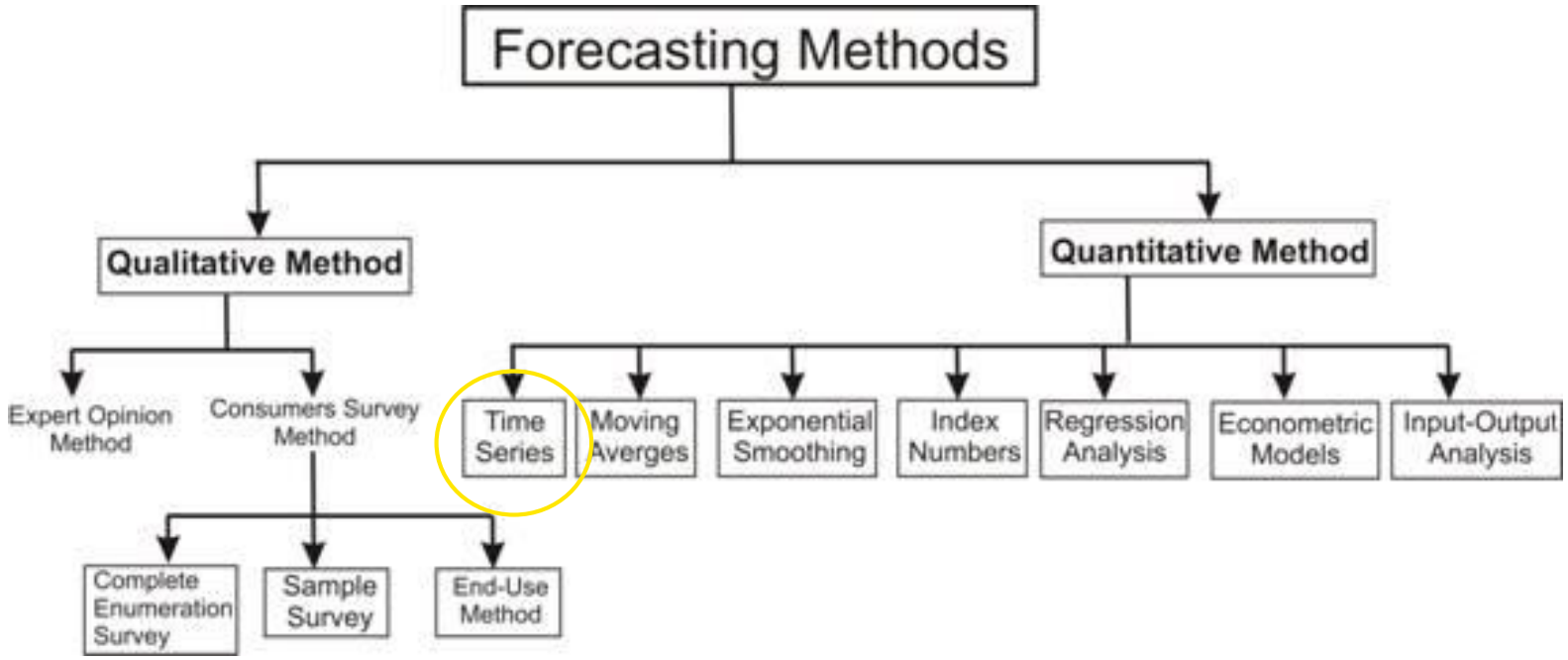
- *I think there is a world market for maybe five computers.* (Chairman of IBM, 1943)
- *Computers in the future may weigh no more than 1.5 tons.* (Popular Mechanics, 1949)
- *There is no reason anyone would want a computer in their home.* (President, DEC, 1977)

Exhibit I **Cost of Forecasting Versus Cost of Inaccuracy For a Medium-Range Forecast, Given Data Availability**



Forecasting y Series de Tiempo

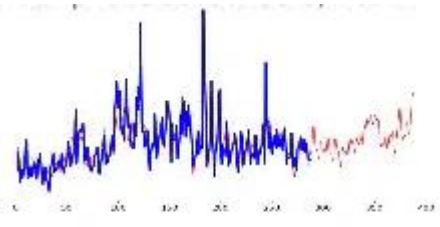
Que modelos predictivos utilizar y su eficacia dependerán de la cantidad y calidad de los datos, horizontes de tiempo considerados(corto-largo plazo), objetivos y planificación.
Hay diversas formas de abordar modelos predictivos, nos enfocaremos en series de tiempo, en donde técnicas de machine learning han progresado y se han masificado para múltiples aplicaciones.



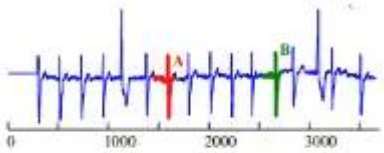
Series de Tiempo

Datos recolectados en intervalos a través del tiempo. Hay múltiples formas de estudiar series de tiempo, sin embargo nos enfocaremos en las orientadas a técnicas modernas en donde la idea es entrenar un modelo que utilice datos del pasado para predecir variables en el futuro.

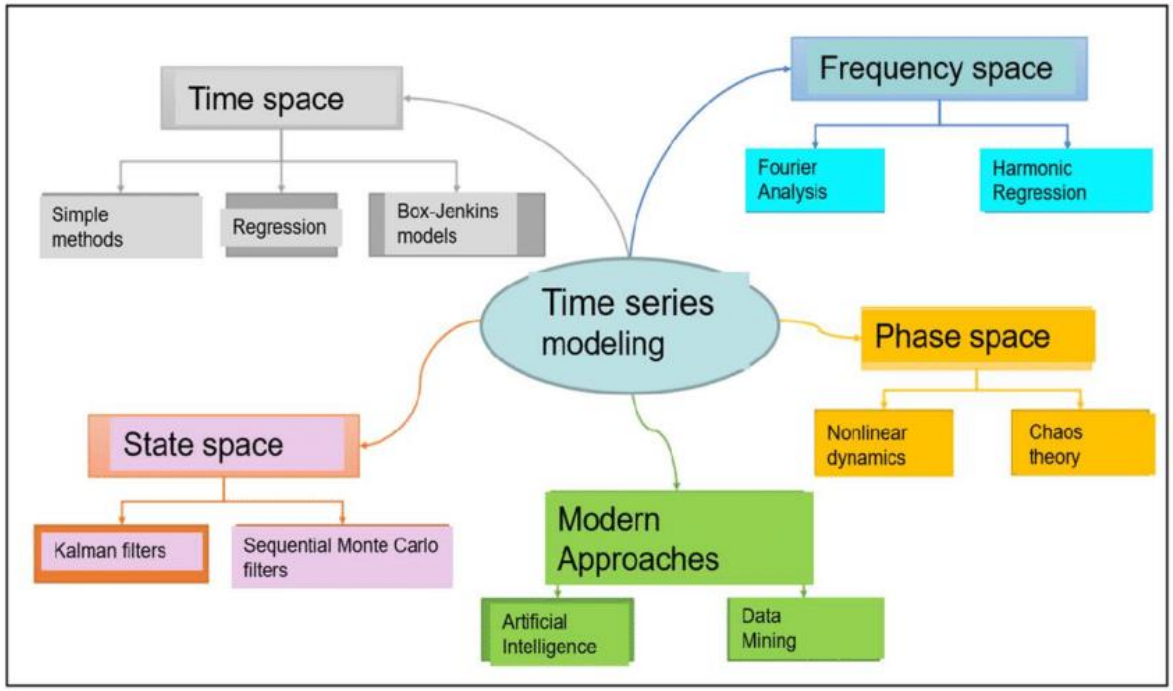
Economía: Indicadores económicos, valor de acciones en la bolsa
Agricultura: lluvias, heladas, temperatura
Industria: fallos maquinaria
Retail: Stock, demanda, ventas



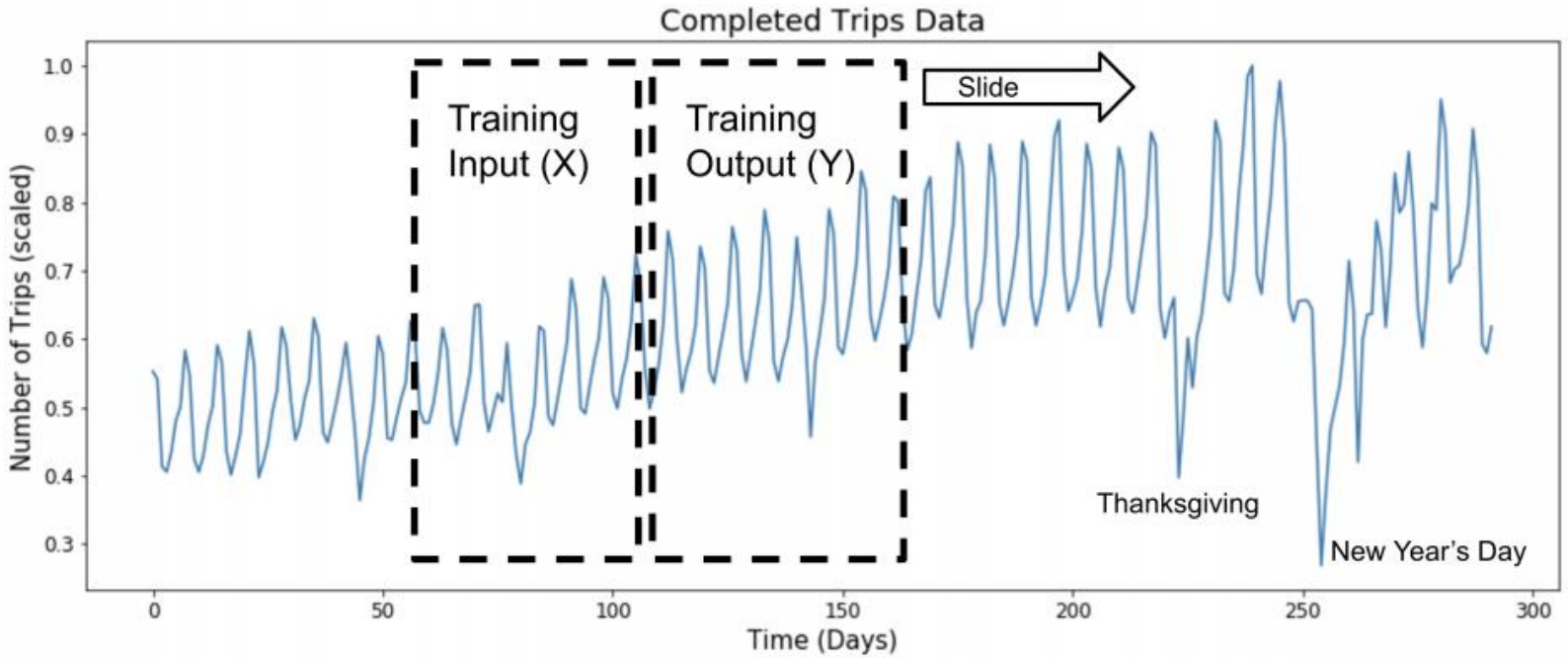
Time series forecasting
Energy demand prediction
Stock market prediction



Time series classification
ECG anomaly detection
Human activity recognition

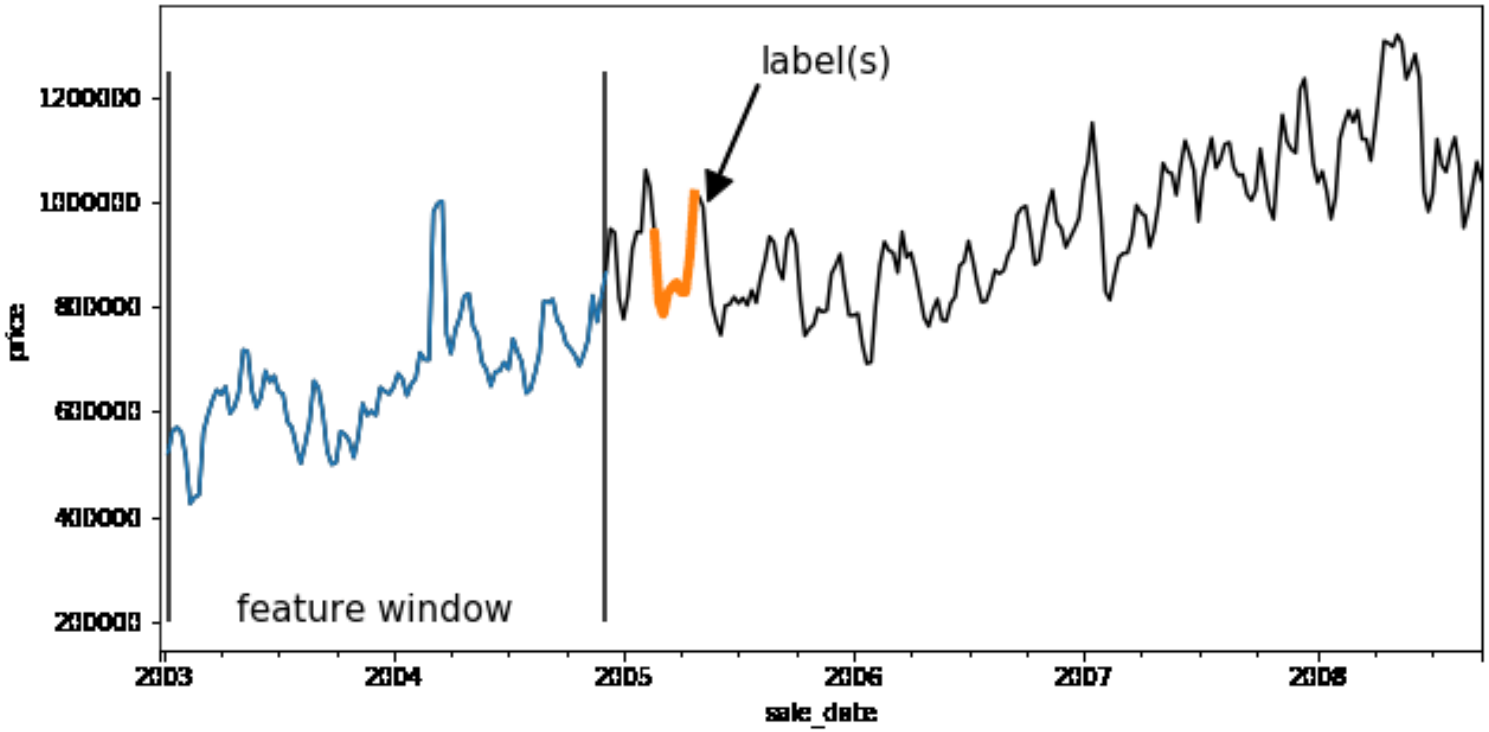


Conceptos: Sliding Window

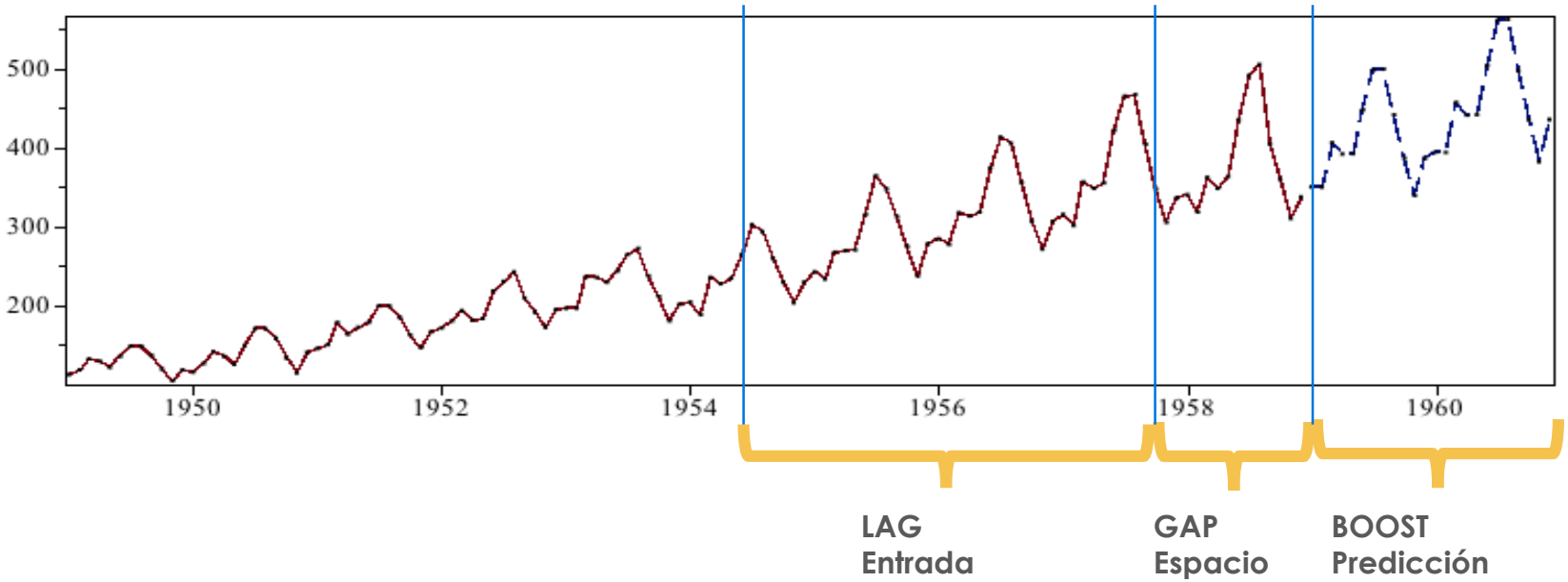


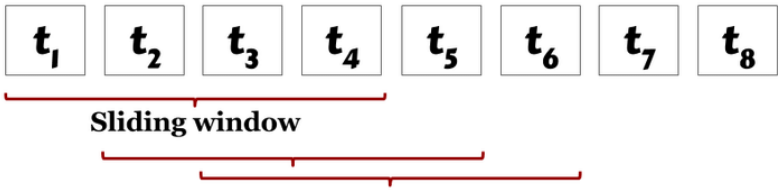
Aparte de separar la serie de tiempo en un set de entrenamiento y de prueba, la idea es recortar y agrupar en datos históricos(X) que servirán para predecir datos futuros(Y), a medida que uno recorre el set de datos.

Conceptos: Sliding Window

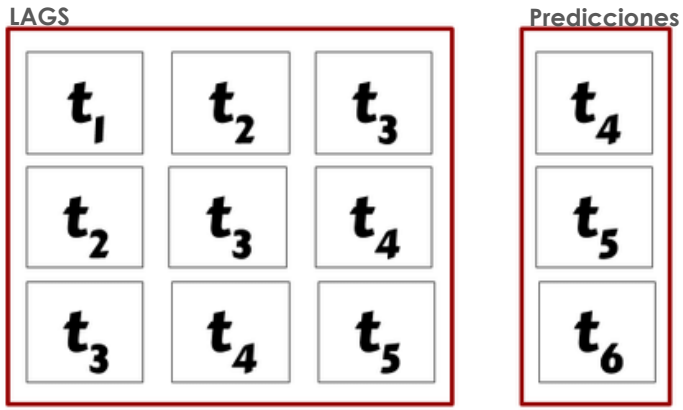


Conceptos: Sliding Window

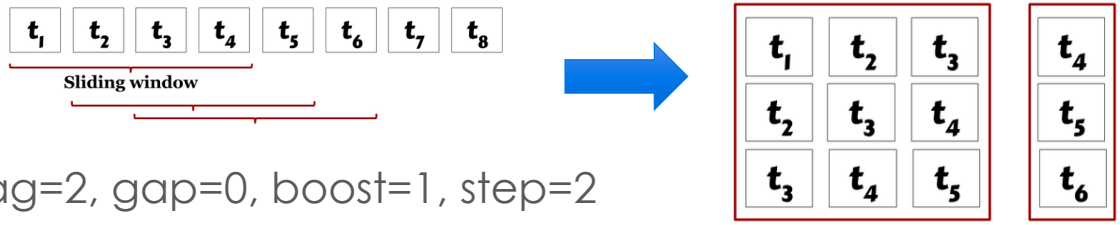




Step: saltos para formar cada lag
en este caso step=1



Conceptos: Sliding Window

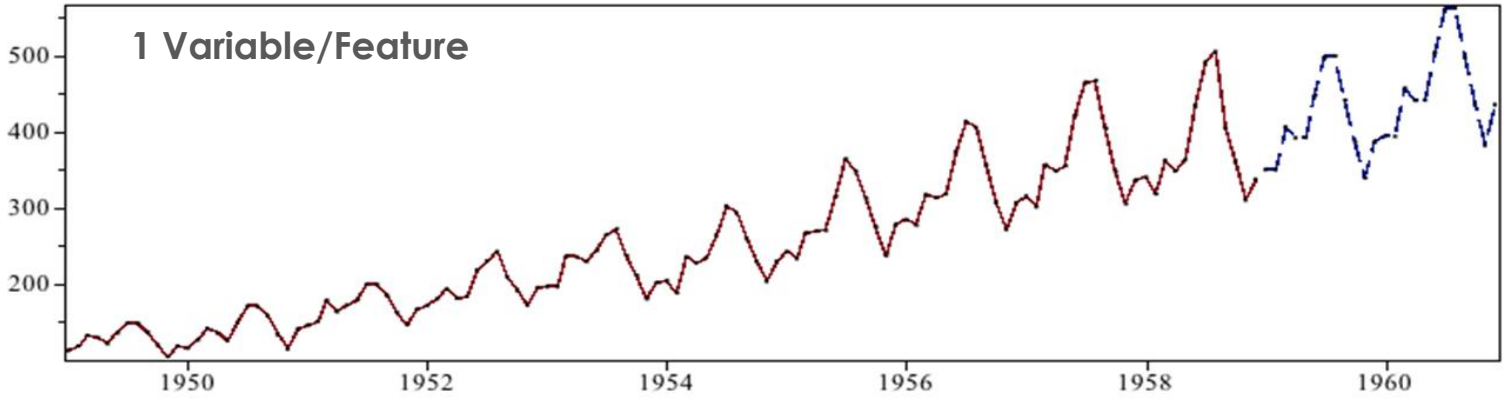


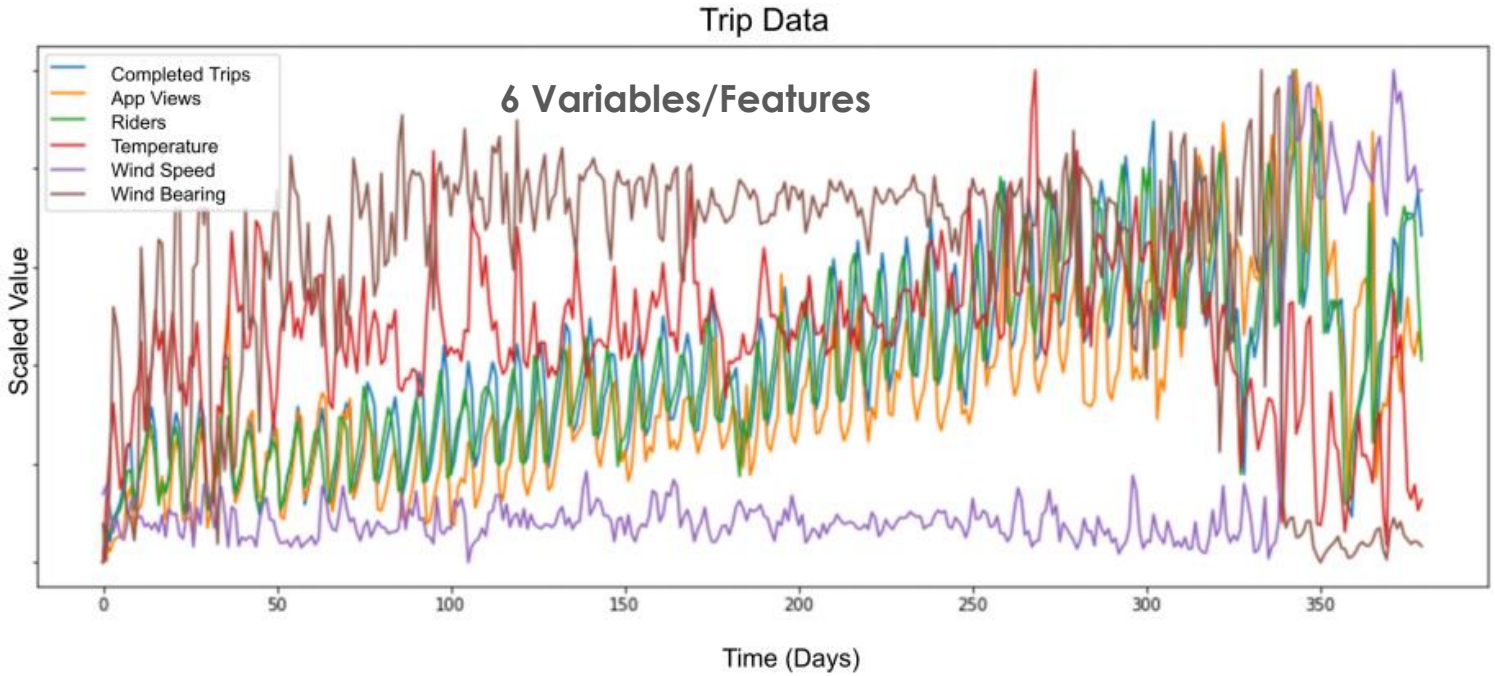
Ejemplo
Serie de tiempo con Lag=2, gap=0, boost=1, step=2

Tiempo	variable
1	0.5
2	0.7
3	0.8
4	1.5
5	1.2
6	1.7
7	1.8
8	2.0
9	2.2

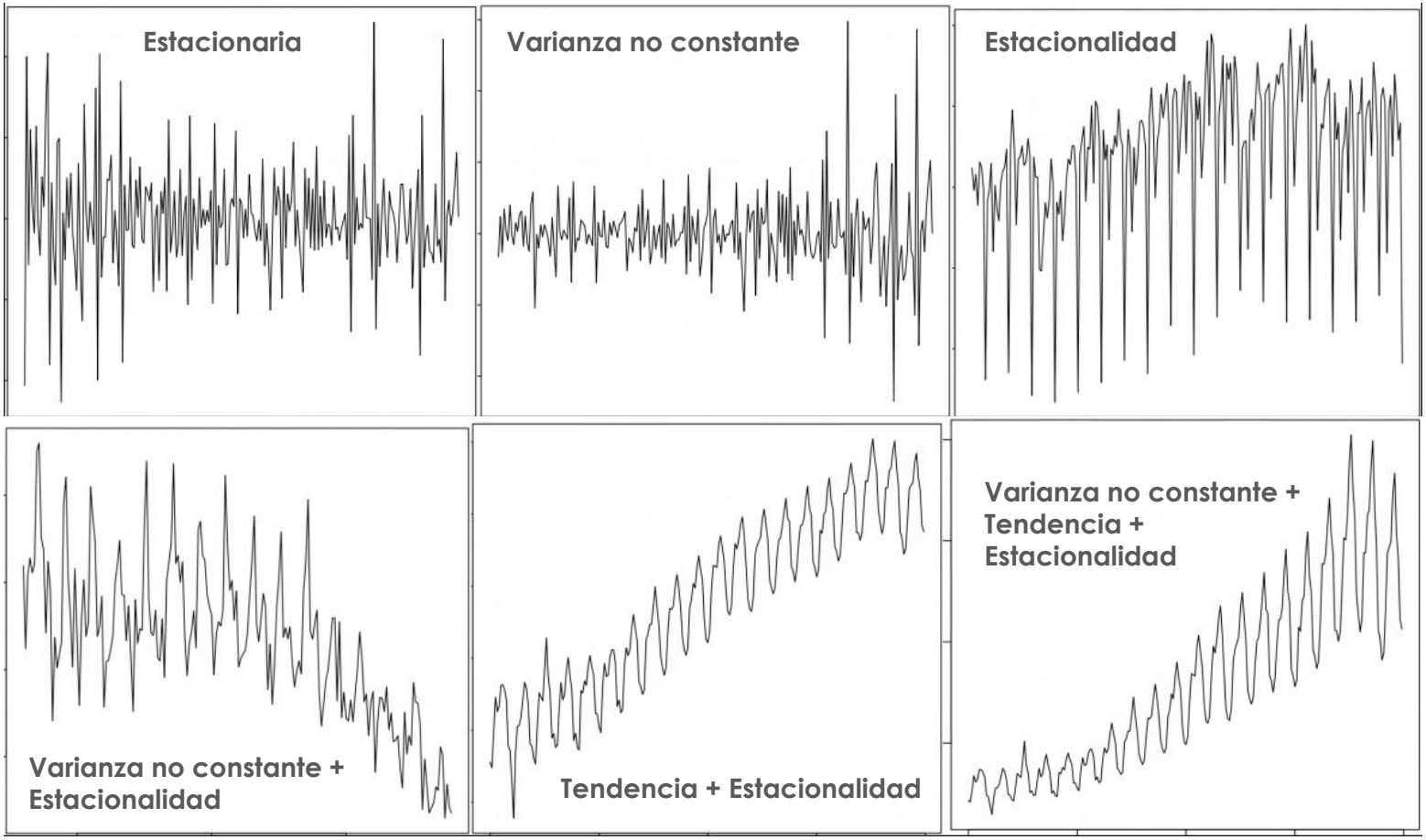
	X		Y
	Lag 1	Lag 2	Predicción
Muestra 1	0.5	0.7	0.8
Muestra 2	0.8	1.5	1.2
Muestra 3	1.2	1.7	1.8
…	…	…	…

Univariadas / Multivariadas



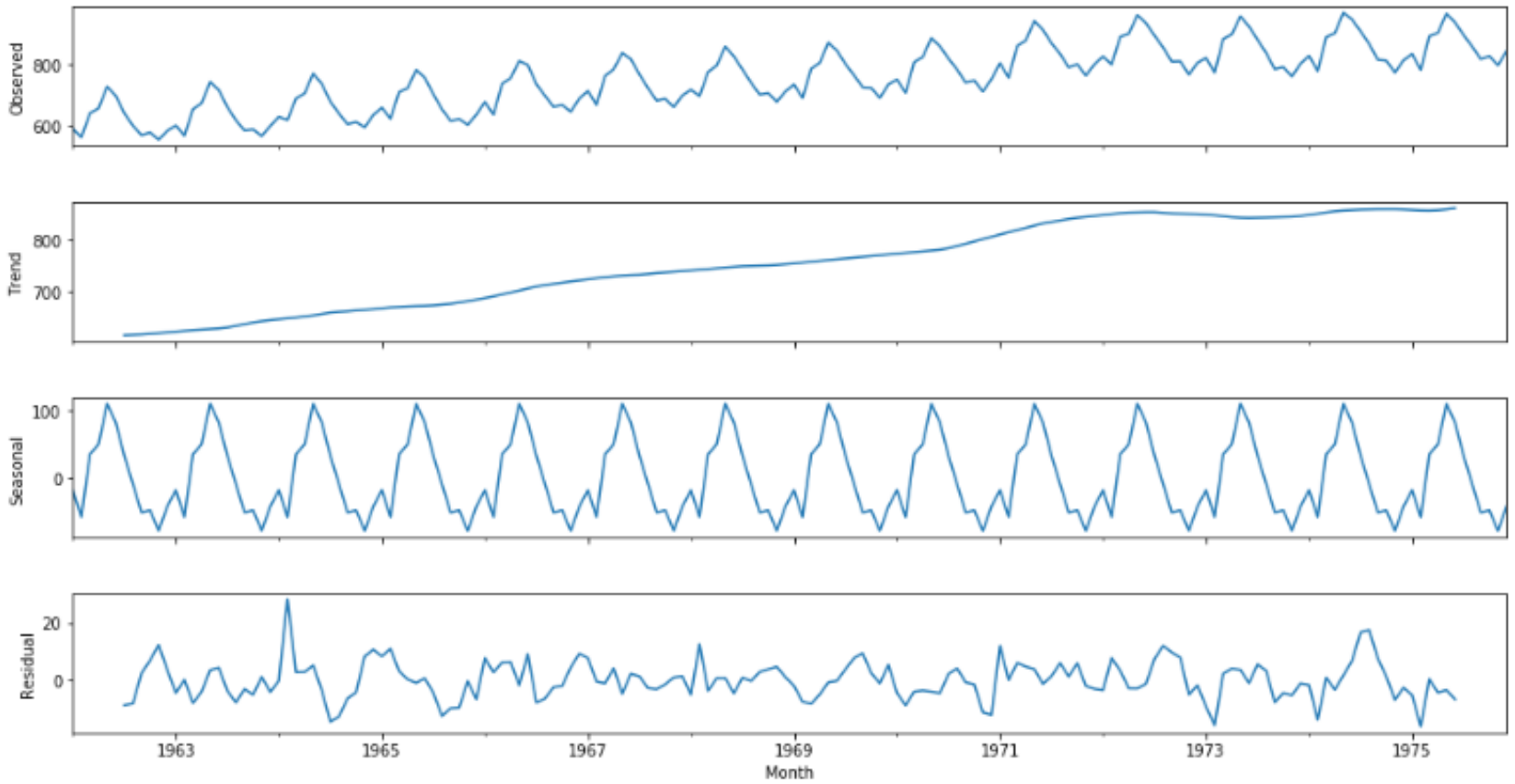


Estacionariedad



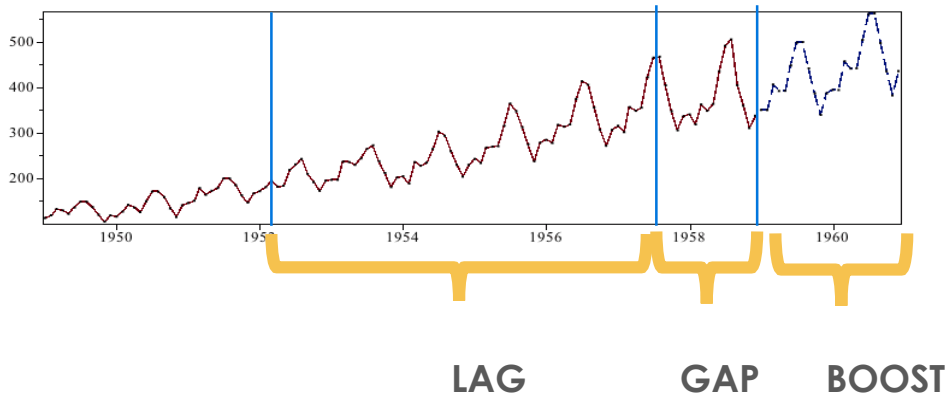
Tendencia: En promedio la variable se incrementa/reduce en función del tiempo?
Estacionalidad: Hay un patrón que se repite regularmente? i.e. aumento ventas los fines de semana
Varianza: Cambia la varianza en función del tiempo?

Descomponer serie



Para modelos de ML en general se puede entrenar la serie observada directamente, en donde el modelo debería aprender la estacionalidad, tendencia y varianza. Sin Embargo, con fines exploratorios y entendimiento de los datos conviene descomponer la serie de tiempo y en algunos casos entrenar modelos en donde se remueven tendencias o estacionalidad(ya entendidos y modelados) puede entregar mejores resultados.

Consideraciones Generales



Consideraciones antes de entrenar un modelo predictivo de ML con series de tiempo:

- **Sampling**: mis datos están medidos en intervalos de tiempo regulares? Missing values? Sincronizacion? Suavizado?
- **Univariado/Multivariado**: Cuantas variables son relevantes?
- **Lag**: cuanto tiempo hacia atrás debo considerar? , cuantas muestras puedo formar?
- **Gap**: Necesito un transiente de tiempo para "decantar" la predicción?
- **Boost**: Que deseo predecir? A mayor tiempo y mas variables de salida menor precisión…
- **Estacionalidad, Tendencia, Varianza**: Como influyen en Lag/Boost? Hay suficientes ciclos o el tamaño del dataset es suficiente para aprender la estacionalidad? Debo descomponer la serie?
- **Step**: importante para la separación del train/test set y evitar correlaciones indeseadas.

Clase 05: Modelos predictivos y series de tiempo

MÉTODOS ACTUALES DE ML

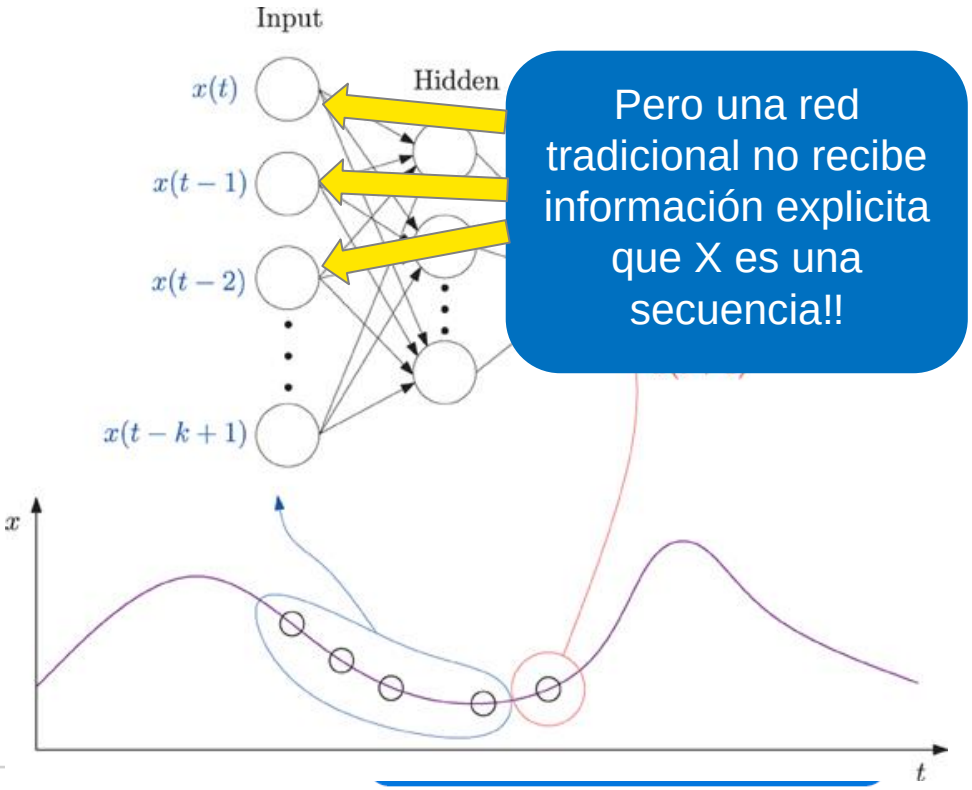
Profesor: **Roberto González**

Redes Neuronales

Cualquier red neuronal tradicional puede ser utilizada para predecir usando series de tiempo, como las MLP, CNN, etc, en donde:
Entrada(X) = Matriz (Features x Lags) que puede ser aplanada en un vector
Salida(Y) = Vector (Boost), para una sola predicción puede ser un escalar, o ser convertido a una categoría para convertirlo en un problema de clasificación. i.e. Temperatura (Kelvin continua) → Alta, media, baja (Categórica 3 clases)

X Y

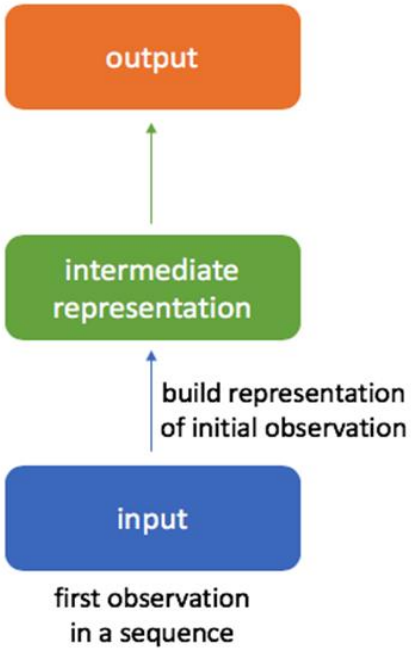
	Lag 1	Lag 2	Predicción
Muestra 1	0.5	0.7	0.8
Muestra 2	0.8	1.5	1.2
Muestra 3	1.2	1.7	1.8
…	…	…	…



Pero una red tradicional no recibe información explicita que X es una secuencia!!

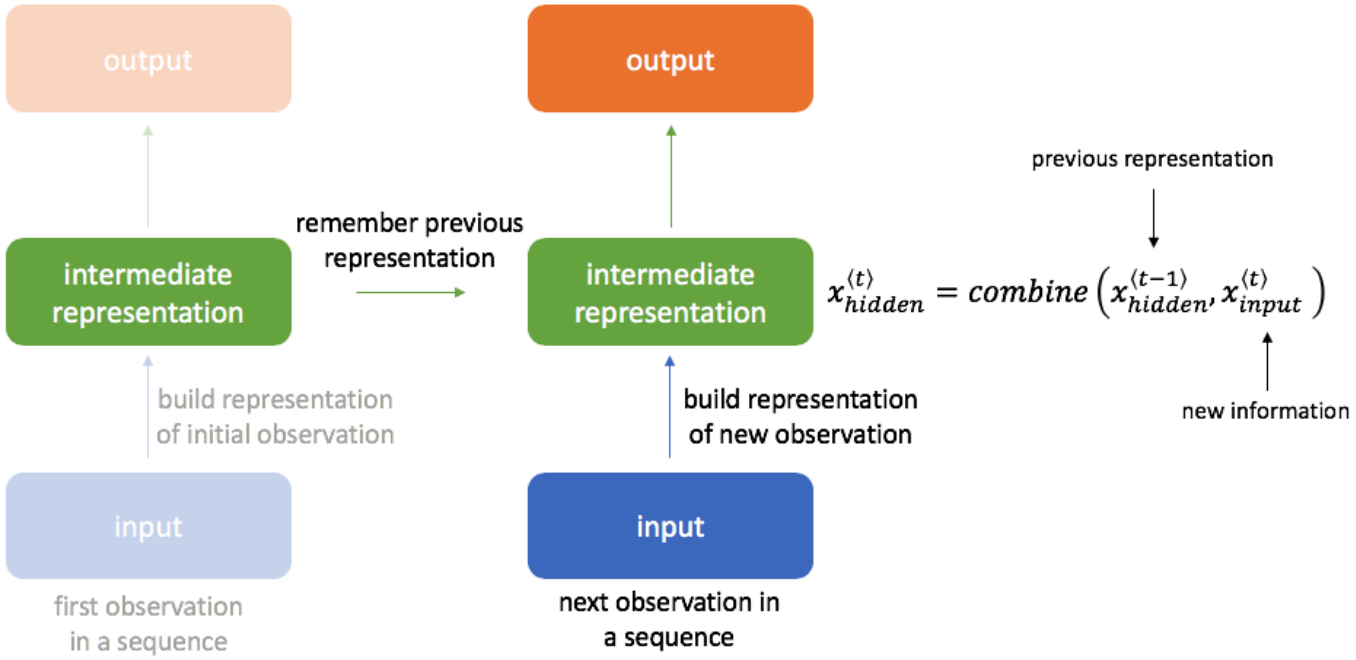
Redes recurrentes RNN

Son redes neuronales especializadas para datos secuenciales como las series de tiempo, en donde las muestras pueden contener un contexto temporal.
Las RNN toman información de la salida del ultimo paso como información de entrada para el siguiente y lo propaga en un flujo secuencial que contiene información sobre los pasos anteriores.
Esto las hace ideales para series de tiempo y procesamiento de lenguaje natural(NLP)



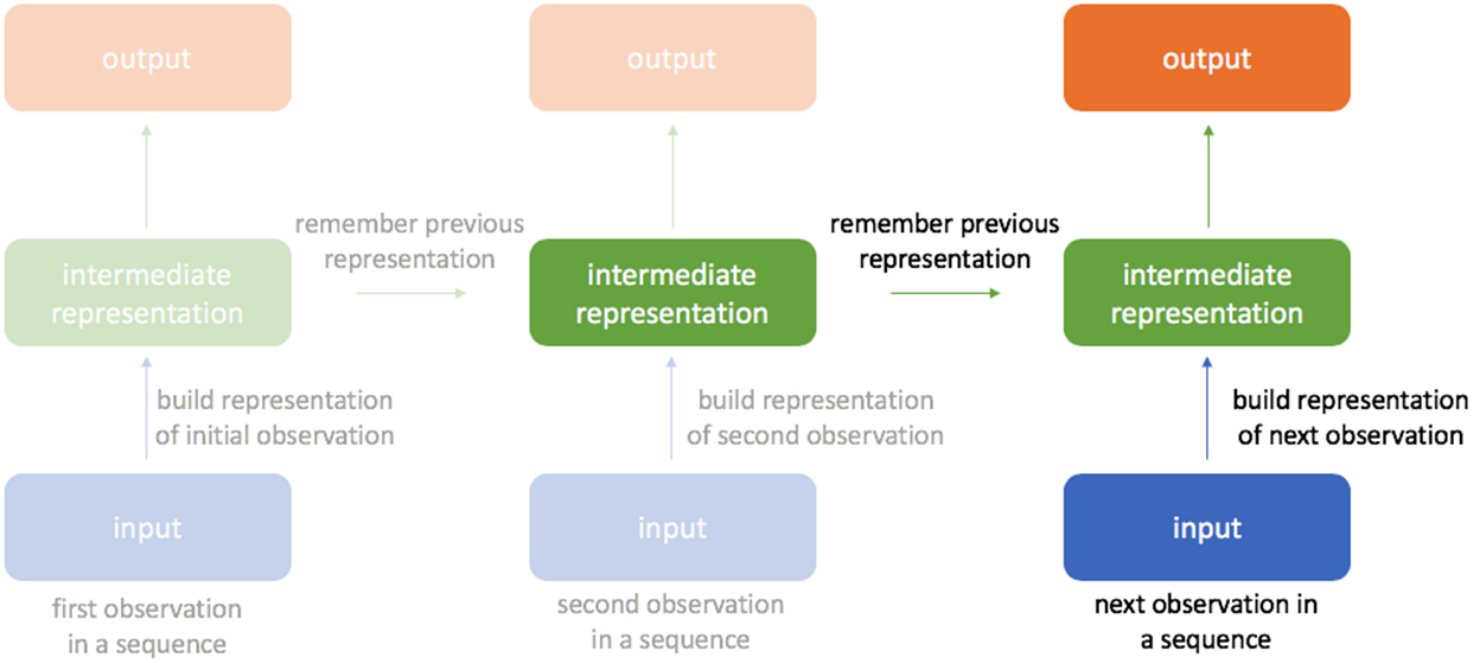
Redes recurrentes RNN

Son redes neuronales especializadas para datos secuenciales como las series de tiempo, en donde las muestras pueden contener un contexto temporal.
Las RNN toman información de la salida del ultimo paso como información de entrada para el siguiente y lo propaga en un flujo secuencial que contiene información sobre los pasos anteriores.
Esto las hace ideales para series de tiempo y procesamiento de lenguaje natural(NLP)

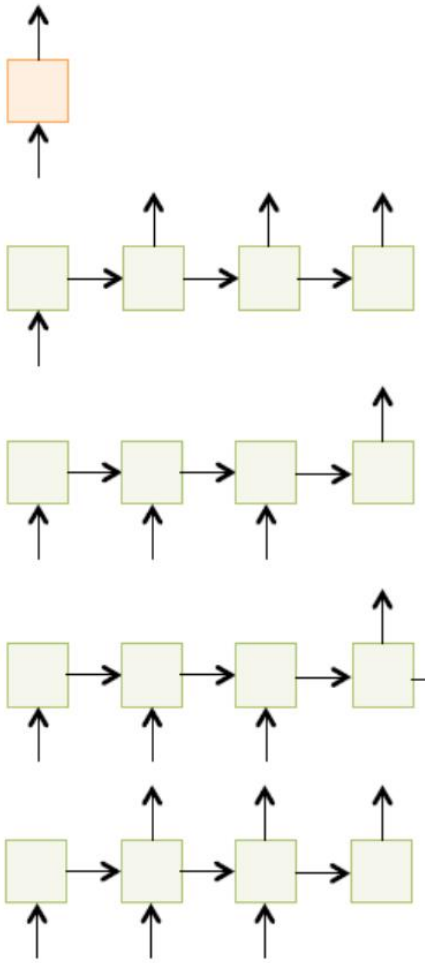


Redes recurrentes RNN

Son redes neuronales especializadas para datos secuenciales como las series de tiempo, en donde las muestras pueden contener un contexto temporal.
Las RNN toman información de la salida del ultimo paso como información de entrada para el siguiente y lo propaga en un flujo secuencial que contiene información sobre los pasos anteriores.
Esto las hace ideales para series de tiempo y procesamiento de lenguaje natural(NLP)



Entrada/Salida RNN



Single - Single — Feed-forward Network

Single - Multiple — Image Captioning

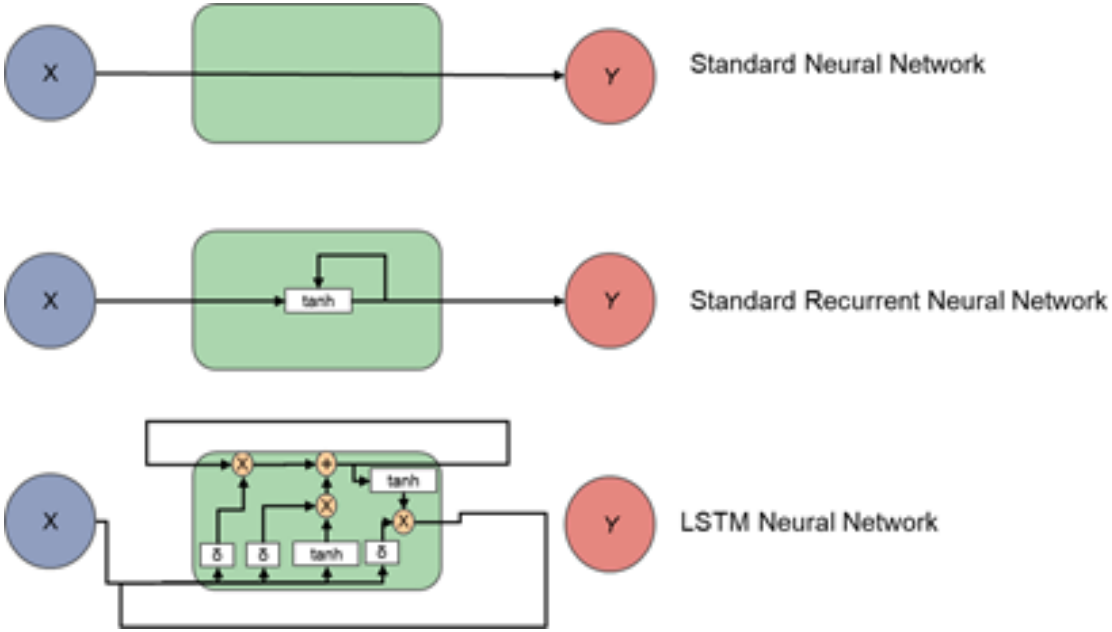
Multiple - Single — Sentiment Classification
Time Series Single Prediction

Multiple - Multiple — Translation
Time Series Multi Prediction

Image Captioning

LSTM y GRU

- RNN sufren de "memoria a corto plazo" es decir para secuencias largas, la información de los pasos mas antiguos se va diluyendo. (Vanishing Gradient Problem, Backpropagation)
- **Gated Recurrent Unit** (GRU) es una primera solución para establecer "memoria a largo plazo" en donde a las neuronas de la red RNN se le agregan "gates"(compuertas).
- **Long Short-term Memory units** (LSTM) son la continuación de las GRU con propiedades especiales para regular el flujo de información del pasado.
- Actualmente las redes LSTM son usadas con bastante éxito en un múltiples aplicaciones para predecir series de tiempo



Comparativas

Técnicas de ML/DL como MLP y LSTM presentan mejor rendimiento en general cuando comparamos con los métodos tradicionales

Table 1: Summary of model accuracy for Up Prediction

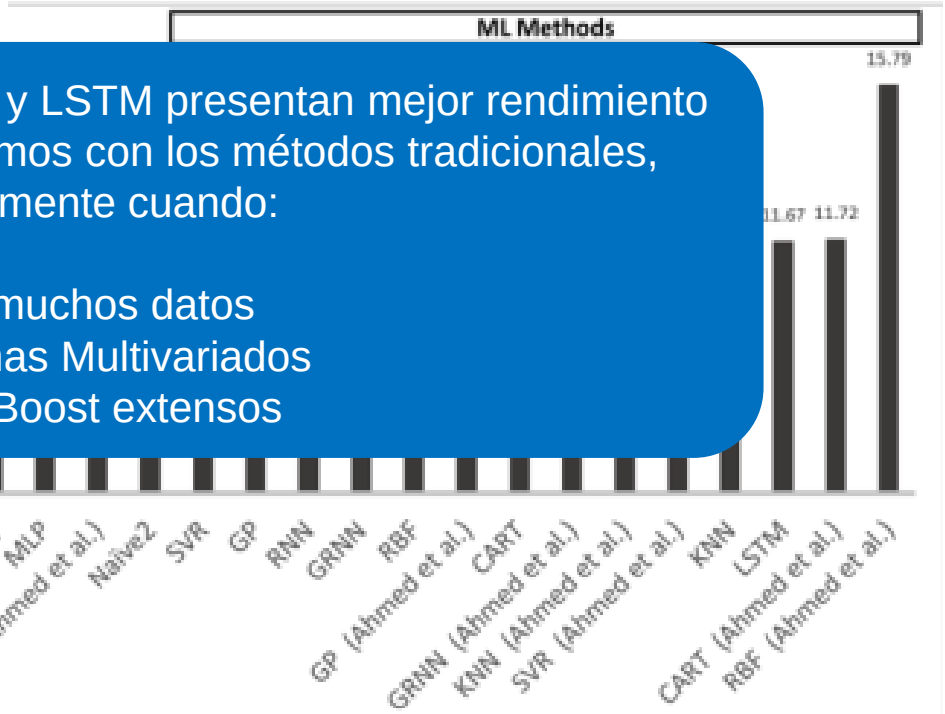
Model	Features	Train Acc.	Validation Acc.	Test Acc.
ARIMA	N/A	N/A	59.16%	N/A
Shallow LSTM	4	59.95%	62.02%	74.16%
Deep LSTM	4	79.26%	88.35%	62.85%

Financial Data			
N225			
IXIC			
HSI			
GSPC			
DJI[1]	516.979	77.643	84.981
DJI[2]	287.6	30.61	-89.356
Average	***511.481***	***64.213***	***-87.445***

[1]. Monthly; [2]. Weekly

Técnicas de ML/DL como MLP y LSTM presentan mejor rendimiento en general cuando comparamos con los métodos tradicionales, especialmente cuando:

- Son muchos datos
- Problemas Multivariados
- Lag y Boost extensos

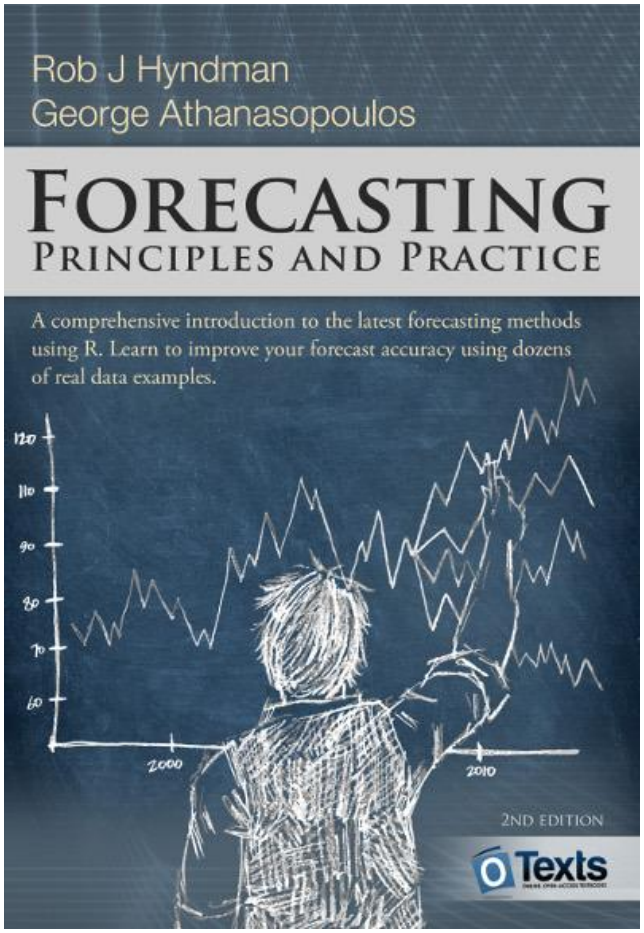


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870978/

Material Complementario

Introducción a Forecasting con ejemplos en R

https://otexts.com/fpp2/



Proyecto 1 – ejemplo R

https://colab.research.google.com/drive/1_rKKFrcB8y-k9evp1MNcxVnB7uVlqRGd?usp=sharing

Proyecto 1
Plan Actividades y Estado de Avance

Con el fin de optimizar y consolidar avances para el proyecto, durante el trabajo grupal cada grupo debe preparar una breve lamina y mostrar al profesor:

- Plan de Actividades definidas para realización del proyecto
- Miembros responsables de cada actividad pasada y futura
- Marcar en las actividades el estado de avance actual
- Dificultades encontradas hasta ahora

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Diplomado en Big Data y Ciencia de Datos
Ciencia de Datos y sus Aplicaciones

Clase 05:

Modelos Predictivos y Series de Tiempo

Roberto González

regonzar@uc.cl