

NLP

Deploy y servicios

Msc. Rodrigo Cardenas Szigety
rodrigo.cardenas.sz@gmail.com

Programa de la materia



Clase 1: Introducción a NLP, Vectorización de documentos.

Clase 2: Preprocesamiento de texto, librerías de NLP y Rule-Based Bots.

Clase 3: Word Embeddings, CBOW y SkipGRAM, representación de oraciones.

Clase 4: Redes recurrentes (RNN), problemas de secuencia y estimación de próxima palabra.

Clase 5: Redes LSTM, análisis de sentimientos.

Clase 6: Modelos Seq2Seq, traductores y bots conversacionales.

Clase 7: Celdas con Attention. Transformers, BERT & ELMo, fine tuning.

Clase 8: Cierre del curso, deploy y servicios, NLP hoy y futuro.

*Unidades con desafíos a presentar al finalizar el curso.

*Último desafío y cierre del contenido práctico del curso.



Conjunto de herramientas o actividades



Resuelven una necesidad

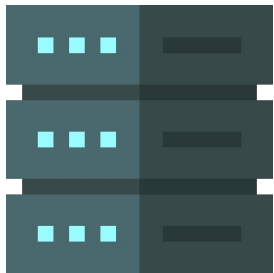


La industria del software se basa principalmente en ofrecer servicios

¿Cómo podemos ofrecer/consumir un servicio?



IaaS



architect, build



PaaS



developer, deploy



SaaS



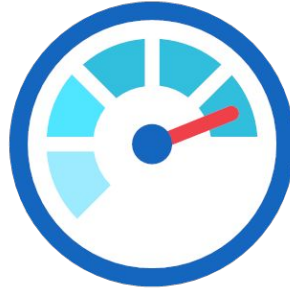
user, product



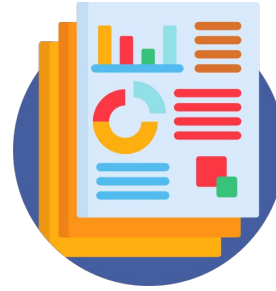
¿Cuáles son los servicios más ofrecidos?



Base de datos



Monitoreo



Reportes



Inteligencia
Artificial

Arquitectura de un servicio

[LINK](#)



Aplicación monolítica 2003 / 2005



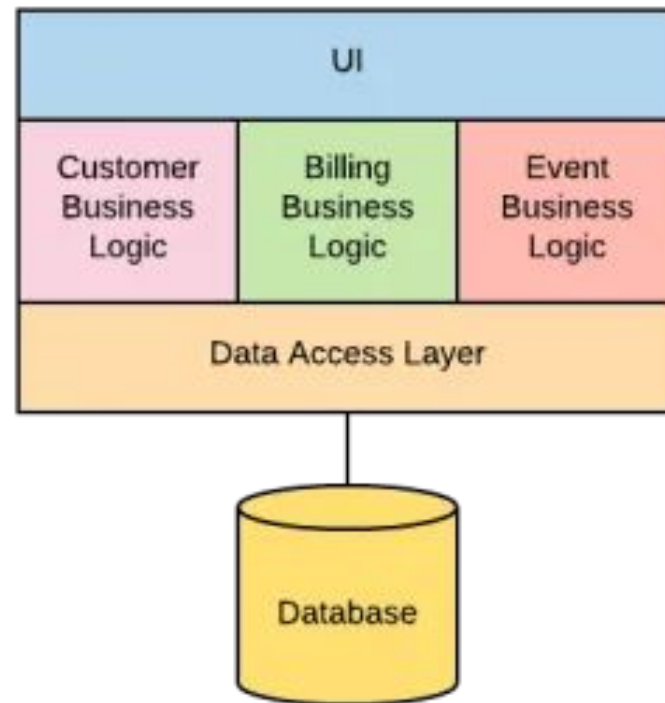
Simple y favorables en una primera etapa del proyecto



Cada cambio implica auditar todo el sistema



No es flexible, no puede adaptarse cambios tecnologías



Arquitectura de un servicio

Microservicios

[LINK](#)



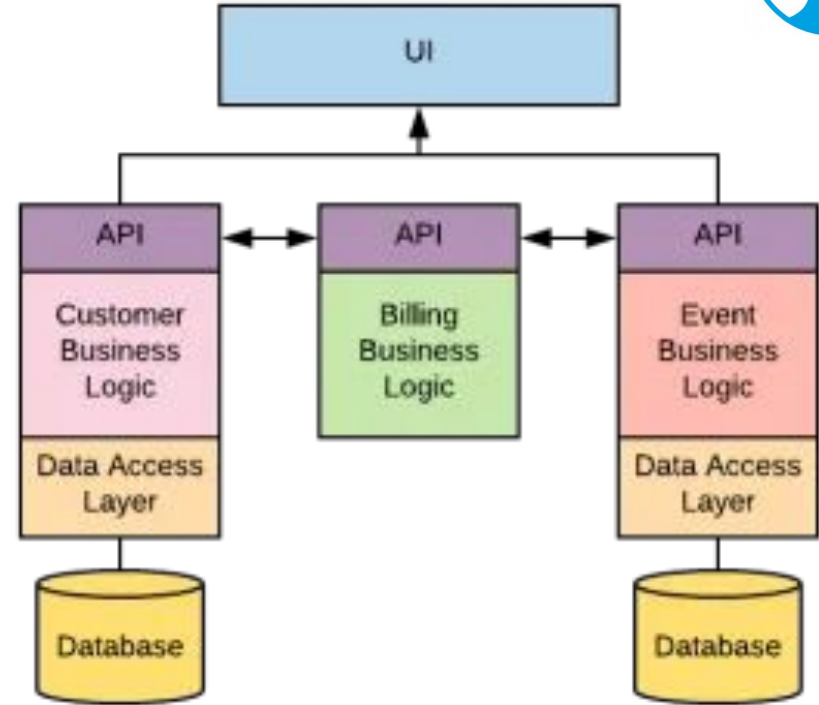
Responsabilidades separadas



Permite agregar o eliminar funcionalidades con riesgo acotado



Utilizar diferentes tecnologías



¿Qué es una API? ----->

¿Qué es una API?

[LINK](#)



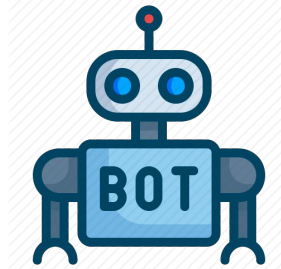
API → Interfaz de programación estándar



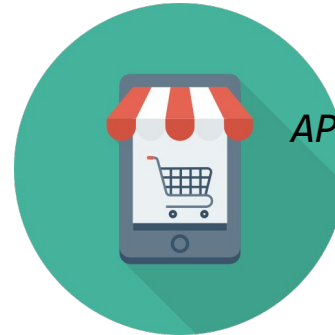
API BANK

APIs de entidades bancarias

*APIs de automatización
de software*



APIs de entidades comerciales



¿Cómo consumimos la REST API de nuestro servicio?

[LINK](#)



```
@app.route("/predict/<input_text>")  
def predict(input_text):
```



La forma correcta sería que los datos no viajen en la URL, sino que se encapsulen en un HTTP POST y JSON

¿Cómo desplegamos nuestro modelo en nuestro servicio?

[LINK](#)



TensorFlow

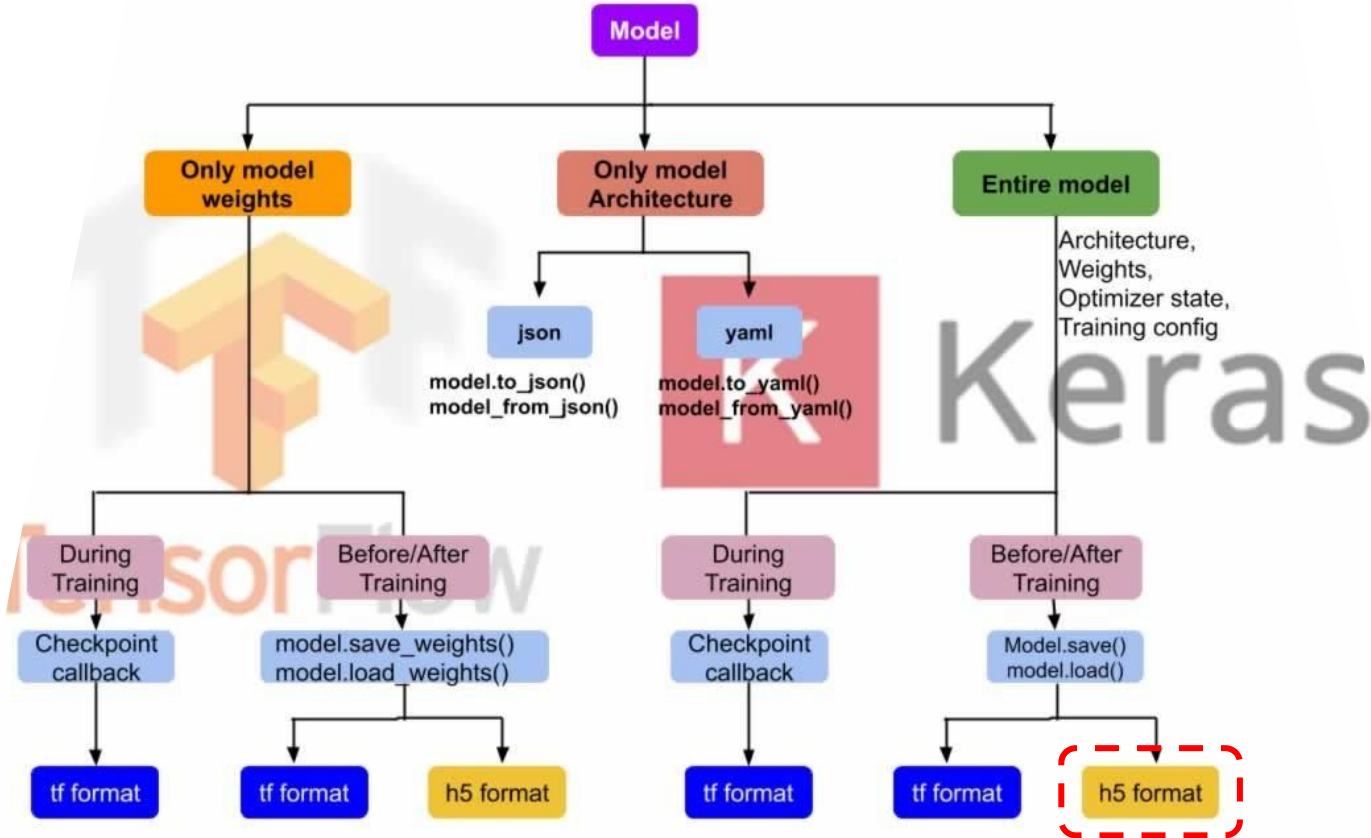


PyTorch



Caffe2

Formas de exportar un modelo TF/Keras



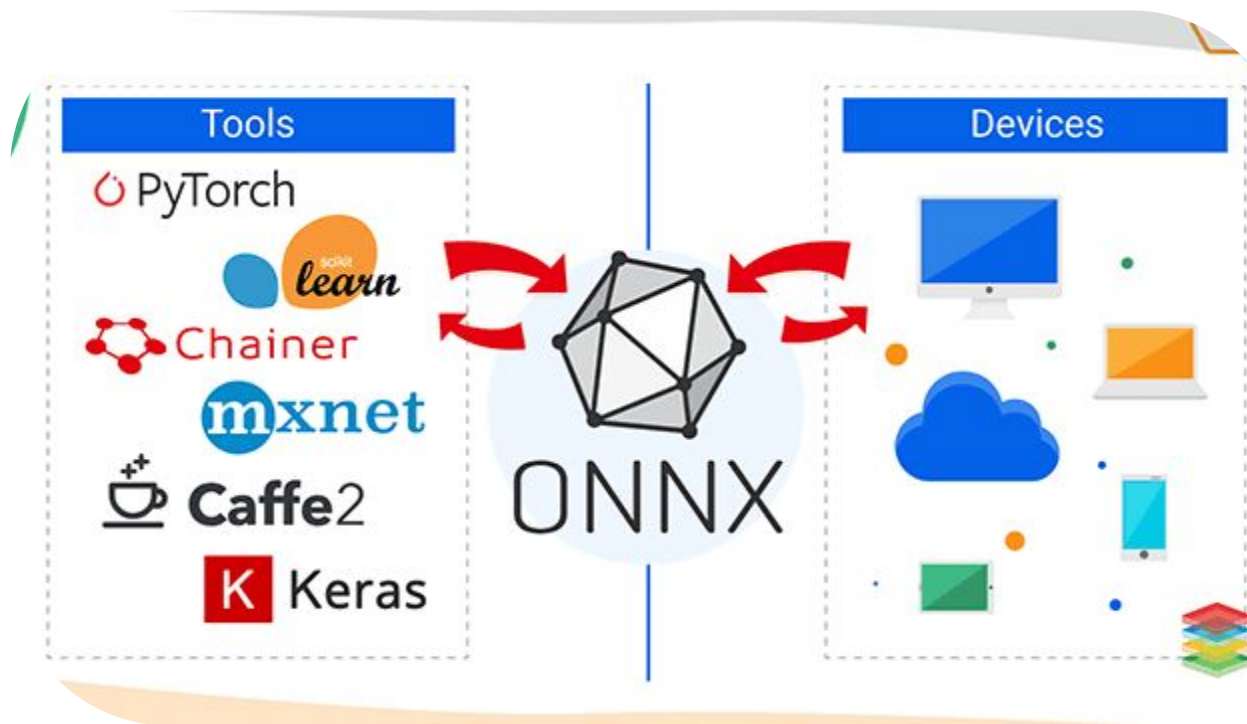


Link al Colab



LINK

¿Hay alguna forma de exportar modelos entre frameworks o dispositivos?



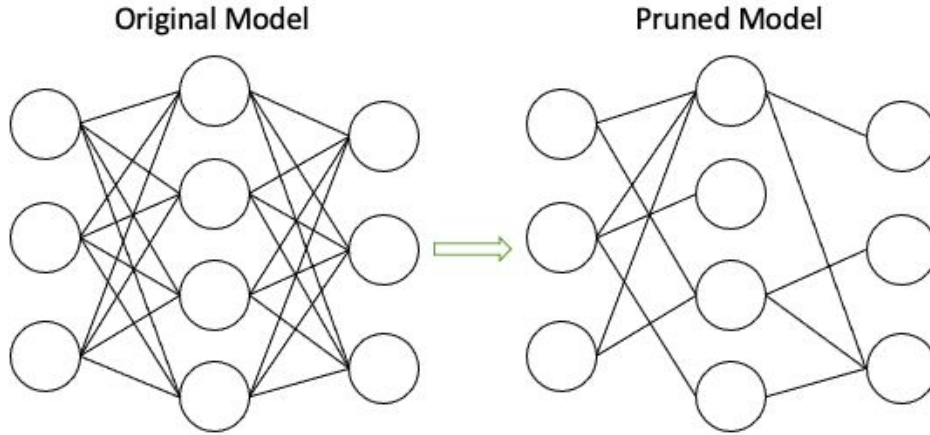
¿Puedo optimizar los modelos?

[LINK](#)

[LINK](#)

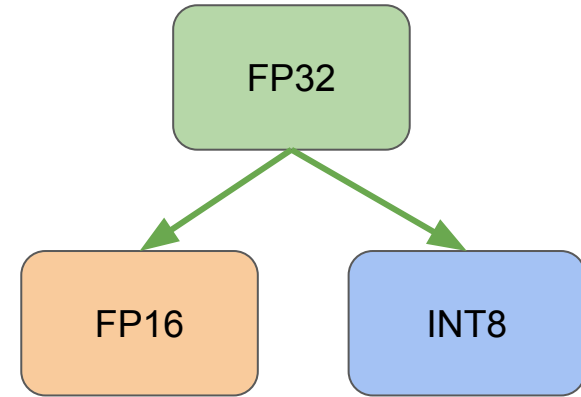


Prune (podar)



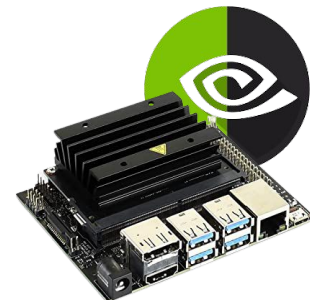
Eliminar los lazos con pesos muy bajos
(bajo aporte al resultado final)
Mejora el "size" y "speed" perdiendo
muy poca precisión

Quantization (cuantización)



Se reemplaza los pesos en float32
por una representación reducida
(float16) o int8. Se reduce "size"
pero se puede perder precisión.

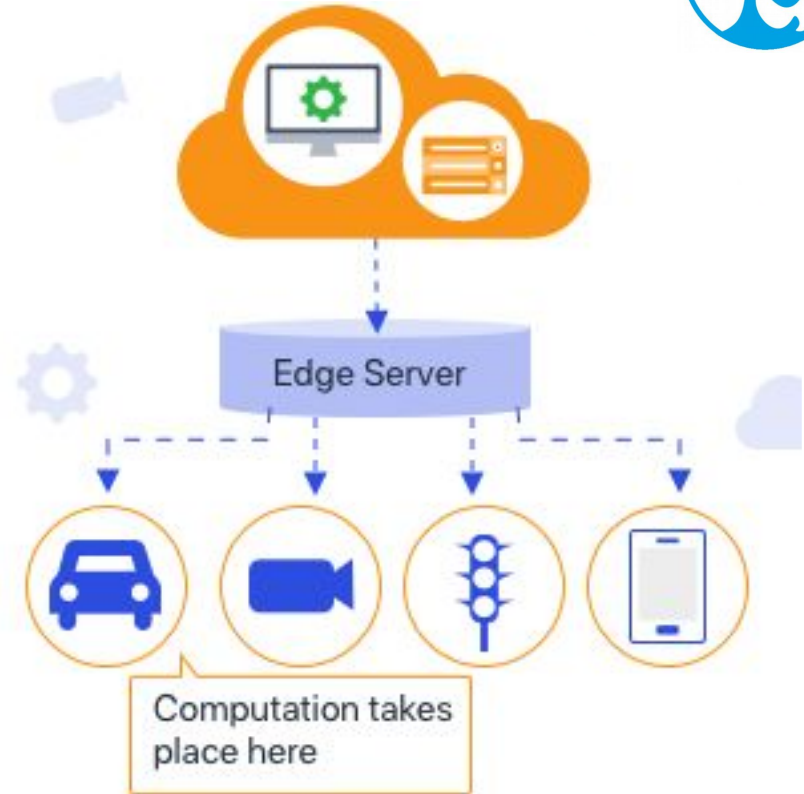
¿En qué plataforma puedo deployar el modelo?



TensorFlow Extended


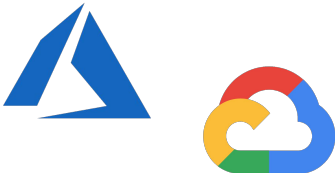
PYTORCH

Edge vs Cloud computing

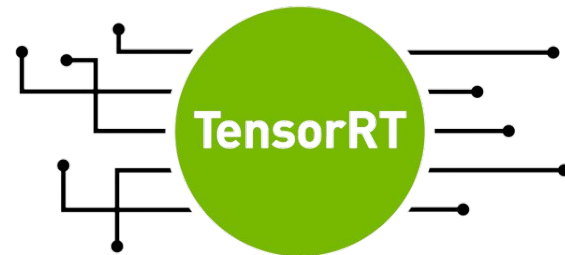
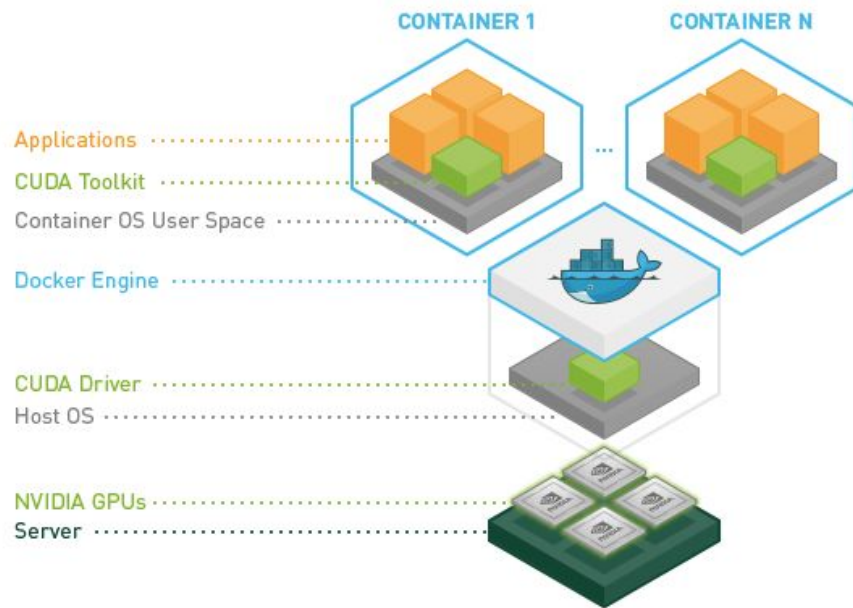


Edge vs Cloud computing

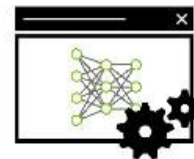


	Pros	Cons
Edge 	<ul style="list-style-type: none">• Más control de tu aplicación.• No requiere conexión a internet.• Menor latencia.• Más seguro	<ul style="list-style-type: none">• Un dispositivo por cliente o solución.• Responder a fallas o problemas con el hardware (reemplazo).
Cloud 	<ul style="list-style-type: none">• Los recursos pueden ser compartidos entre aplicaciones.• No hay que mantener una plataforma o hardware.• No hay que reemplazar hardware dañado.	<ul style="list-style-type: none">• Costos mensuales asociados a la infraestructura.• Costos por tráfico de red (internet).• Costo por uso de storage (disk).

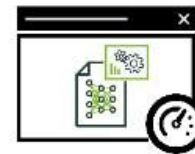
¿Puedo optimizar el modelo para la plataforma seleccionada (ej: NVIDIA)?



Trained
Neural
Network



TensorRT
Optimizer



TensorRT
Runtime
Engine

¿Cómo se puede vender/ofrecer nuestro servicio?



PaaS



developer



Utilizando a Flask/Django como plataforma, que gestione usuarios con tokens de acceso.



Brindando una API con documentación para desarrolladores.

SaaS



user



Creando un plugin para una SaaS utilizada (como wordpress) que consuma nuestras APIs por debajo.



Brindar una interfaz de configuración (GUI) para no programadores



Link al Colab



[LINK](#)



Link al github



LINK



¡Muchas gracias!