



UNIVERSIDAD  
DE GRANADA

# Aprendizaje predictivo en Nutrición

Trabajo Fin de Máster

**Andrea Morales Garzón**

**Juan Gómez Romero**

**María J. Martín Bautista**

# Contenido

**1** Introducción

**2** Objetivos

**3** Planificación

**4** Arquitectura del sistema y experimentación

**5** Aplicación

**6** Conclusiones

# Contenido

## 1 **Introducción**

## 2 Objetivos

## 3 Planificación

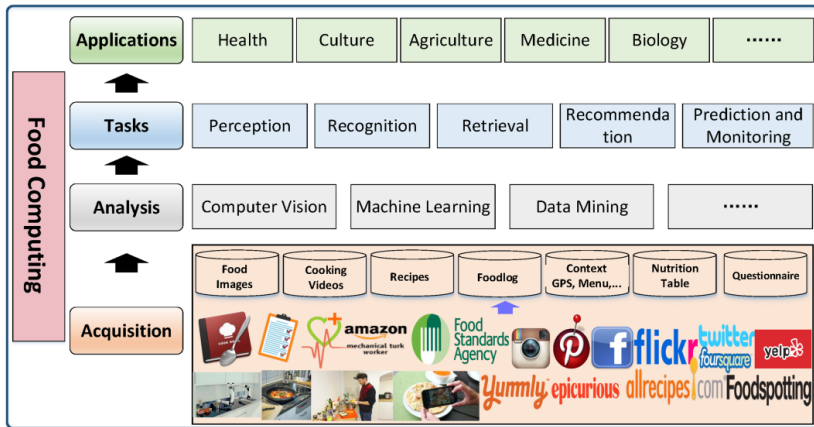
## 4 Arquitectura del sistema y experimentación

## 5 Aplicación

## 6 Conclusiones

# Food Computing

- Área de la computación que abarca problemas sobre nutrición
- Suele combinar más de una fuente alimenticia
- La **predicción** es una de las tareas más relevantes



# Fuentes de datos heterogéneas

- Food Computing combina distintas fuentes de datos: nutricionales, de recetas, etc.
- Problemas al trabajar con datos heterogéneos:
  - Difieren en contenido, estructura y origen
  - Requieren tratamiento previo para usarlos de forma agregada



# Nuestro problema a resolver

## ¿Qué queremos hacer?

- Desarrollar una herramienta para combinar información heterogénea
  - Detectar términos comunes a través de sus **descripciones textuales**

## ¿Cómo lo resolvemos?

- Comparando representaciones numéricas de las descripciones textuales
  - **Modelos predictivos** para aprender las representaciones
  - **Medidas de distancia** para identificar equivalencias
- Mostrando su funcionamiento en un problema real
  - Adaptación de recetas a restricciones (vegetarianas y veganas)

# Objetivos

## Objetivo principal:

*Estudio, diseño e implementación de técnicas predictivas para resolver un problema de fusión y consulta sobre datos textuales*

## Objetivos específicos:

- 1 *Identificar los elementos a los que representan los datos que intervienen en el problema*
- 2 *Fusionar información heterogénea*
- 3 *Mostrar la eficacia y el alcance de la herramienta desarrollada*

# Contenido

1 Introducción

2 Objetivos

**3 Planificación**

4 Arquitectura del sistema y experimentación

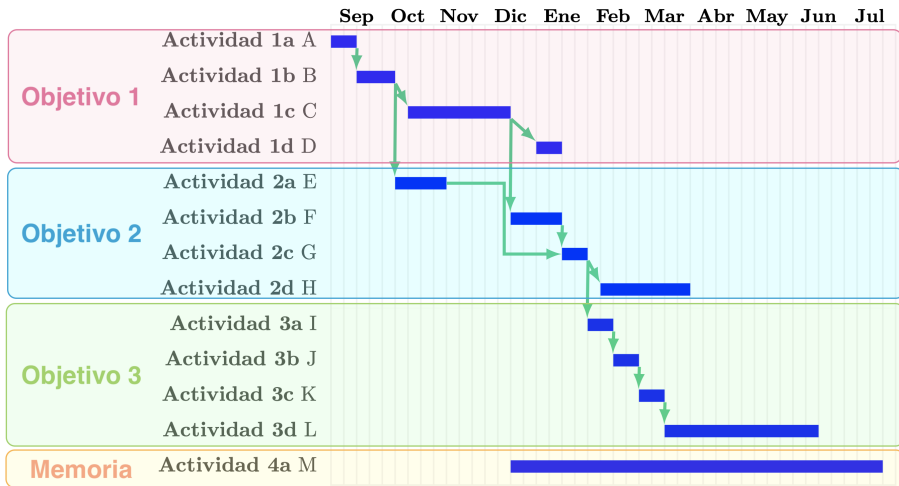
5 Aplicación

6 Conclusiones



# Planificación Temporal

## 1 Descomposición de los objetivos en actividades



# Estimación de costes

- Coste asociado al uso de recursos computacionales
- Costes asociados al personal
  - Estimación de costes por descomposición de actividades

Actividad	Plan	Análisis	Diseño	Desarrollo	Test	Total
<b>Actividad 1</b>	<b>0.4</b>	<b>0.4</b>	<b>1.4</b>	<b>1.1</b>	<b>0.7</b>	<b>4.0</b>
Actividad 1a	0.1	0.0	0.1	0.4	0.0	0.6
Actividad 1b	0.1	0.1	0.5	0.1	0.2	1.0
Actividad 1c	0.1	0.2	0.7	0.4	0.2	1.6
Actividad 1d	0.1	0.1	0.1	0.2	0.3	0.8
<b>Actividad 2</b>	<b>0.4</b>	<b>1.2</b>	<b>0.7</b>	<b>0.35</b>	<b>0.9</b>	<b>3.55</b>
Actividad 2a	0.1	0.4	0.2	0.0	0.2	0.9
Actividad 2b	0.1	0.5	0.2	0.0	0.2	1.0
Actividad 2c	0.1	0.2	0.2	0.25	0.2	0.95
Actividad 2d	0.1	0.1	0.1	0.1	0.3	0.7
<b>Actividad 3</b>	<b>0.3</b>	<b>0.35</b>	<b>1.0</b>	<b>0.95</b>	<b>0.75</b>	<b>3.35</b>
Actividad 3a	0.1	0.1	0.1	0.1	0.1	0.5
Actividad 3b	0.1	0.1	0.2	0.15	0.2	0.75
Actividad 3c	0.0	0.05	0.1	0.1	0.15	0.4
Actividad 3d	0.1	0.1	0.6	0.6	0.3	1.7
<b>Total</b>	<b>1.1</b>	<b>1.95</b>	<b>3.1</b>	<b>2.4</b>	<b>2.35</b>	<b>10.9</b>

**Tabla:** Estimación de costes en p.m. por descomposición de actividades

# Contenido

1 Introducción

2 Objetivos

3 Planificación

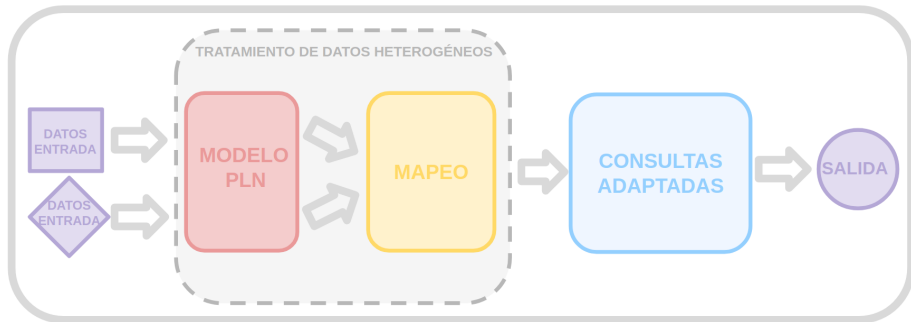
**4 Arquitectura del sistema y experimentación**

5 Aplicación

6 Conclusiones

# Arquitectura del sistema

- **Entrada:** datos heterogéneos
- **Salida:** consulta sobre los datos **ya combinados**

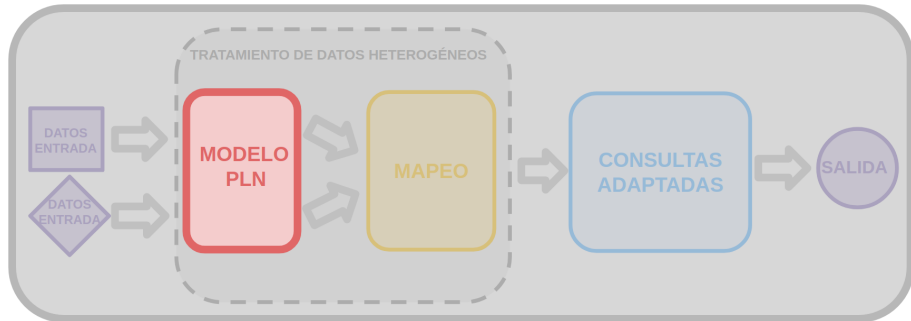


# Sistema adaptado a Food Computing

- Adaptación de recetas a restricción
  - Se combinan **ingredientes** de la receta con **información nutricional**
- **Entrada:** recetas y base de datos nutricional
- **Salida:** receta adaptada a la restricción



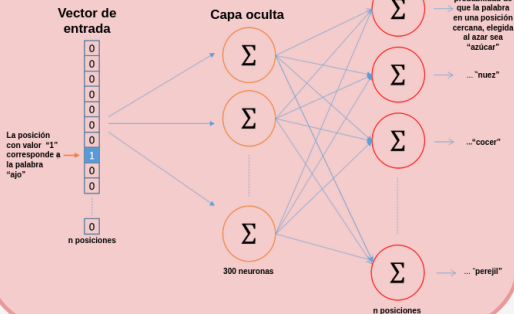
# Módulo de Procesamiento del Lenguaje



# Módulo de Procesamiento del Lenguaje

## TRATAMIENTO DE DATOS HETEROGÉNEOS

### MODELO PLN



MAPEO

# Modelo predictivo del lenguaje

## Word Embedding

- **Aprendizaje predictivo** (recuentos de co-ocurrencia de palabras)
- Representación de palabras en el **espacio semántico**

- Palabras
- Distancias significativas
- Relaciones significativas
- Espacio multidimensional





# Modelo predictivo del lenguaje

## 1 Creación del corpus para entrenar el modelo

- Colección de recetas obtenidas de webs en inglés
- Usamos las instrucciones de preparación para montar el corpus

Procedencia	Número de recetas
BBC Food Recipe	10,679
Epicurious	20,111
Cookstr	225,602
AllRecipes	10,679
<b>Total de recetas</b>	<b>267,071</b>

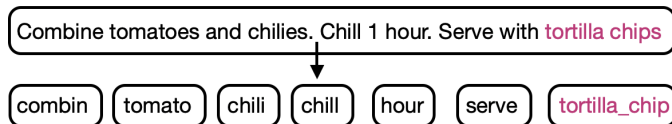
**Tabla:** Corpus de recetas: origen y tamaño

# Modelo predictivo del lenguaje

1 Creación del corpus para entrenar el modelo

2 Preprocesamiento del corpus

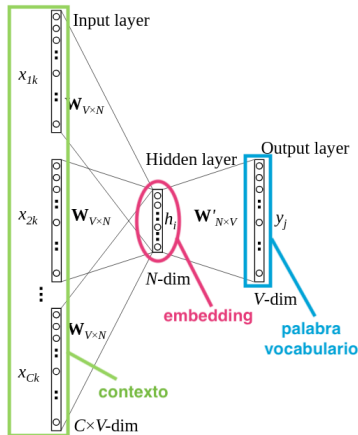
- Tokenizar el texto de preparación de las recetas
- Realizar un preprocesamiento típico
  - Convertir a minúscula
  - Eliminar signos de puntuación, dígitos y caracteres especiales
  - Eliminar *stop words*
  - Aplicar lematización
- Entrenar un modelo de bigramas para detectar palabras compuestas



# Modelo predictivo del lenguaje

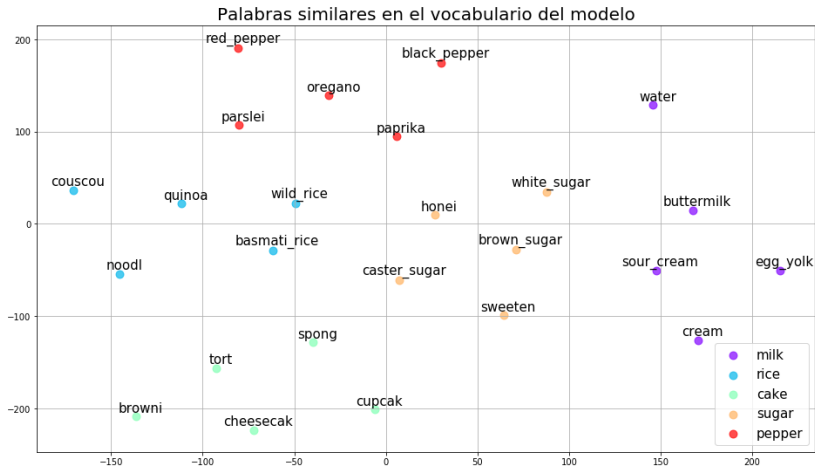
- 1 Creación del corpus para entrenar el modelo
- 2 Preprocesamiento del corpus
- 3 Entrenamiento del modelo

- Word2vec - CBOW
- Dimensión: 300
- Contexto: 5
- Épocas: 30
- Vocabulario: 11288



# Experimentación con el modelo

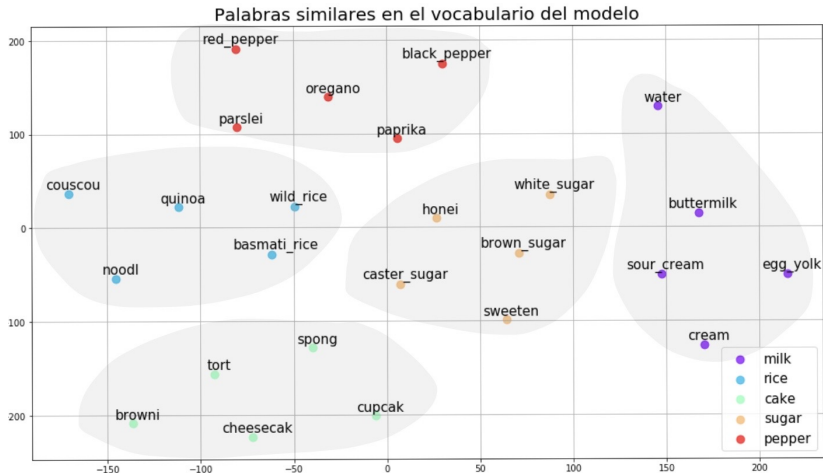
## 1 Visualización del vocabulario del modelo predictivo



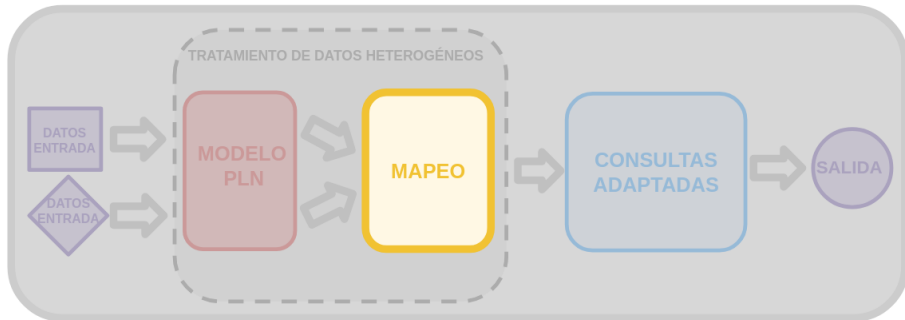
# Experimentación con el modelo

1 Visualización del vocabulario del modelo predictivo

2 La separación espacial es relevante

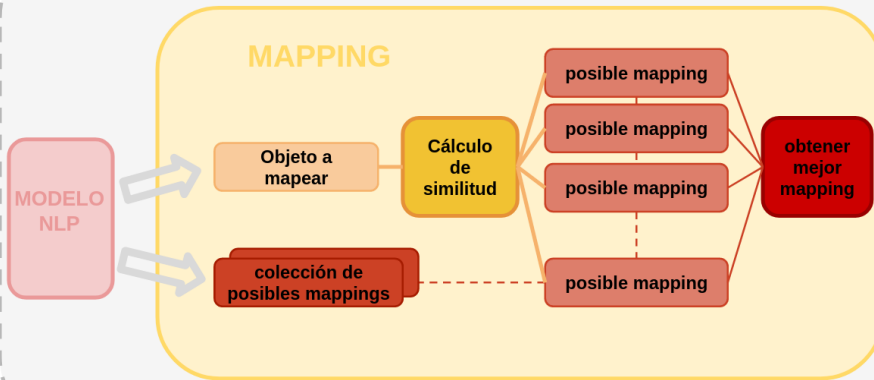


# Módulo de Mapeo



# Módulo de Mapeo

## TRATAMIENTO DE DATOS HETEROGÉNEOS



# Medidas de distancia

## ■ Distancia de Jaccard

- Distancia sintáctica
- Intersección entre los conjuntos asociados a las descripciones

## ■ Distancia Word's Mover

- Distancia semántica
- Coste de *viajar* de una descripción a la otra

## ■ Distancia híbrida

- Distancia sintáctica y semántica
- Combinación ponderada de Jaccard y Word's Mover



# Medidas de distancia

## ■ Distancia de Jaccard

- Distancia sintáctica
- Intersección entre los conjuntos asociados a las descripciones

## ■ Distancia Word's Mover

- Distancia semántica
- Coste de *viajar* de una descripción a la otra

## ■ Distancia híbrida

- Distancia sintáctica y semántica
- Combinación ponderada de Jaccard y Word's Mover

## ■ Distancia difusa de Jaccard

- Intersección **difusa** entre conjuntos (sintáctica)

## ■ Distancia difusa entre documentos

- Intersección **difusa** (semántica)
- Descripciones como un todo (y no elemento a elemento)


# Mapeos: experimentación y resultados

## ■ Problema de mapeo entre bases de datos: i-Diet y USDA

ID	Description	...	Food Code	Main Food Description	...
290	Apple, raw	...	396	Onion, mature, raw	...
96	Onion, raw	...	599	Fruit juice, average	...
...	...	...	...	...	...

i-Diet

USDA



## ■ Las medidas difusas mejoran a las clásicas

Medida de distancia	Top 1	Top 2	Top 3	Top 5	Top 10
Distancia Jaccard	16.75	20.16	22.20	25.20	27.52
Word Mover's Distance	30.65	35.55	36.92	40.87	44.82
Distancia híbrida	32.15	37.12	40.19	43.05	47.41
Distancia Jaccard difusa	23.84	29.70	33.37	39.23	45.64
Distancia difusa entre documentos	35.55	40.46	43.46	47.00	53.26

**Tabla:** Resultados del mapeo (%) con las medidas de distancia

# Mapeos: experimentación y resultados

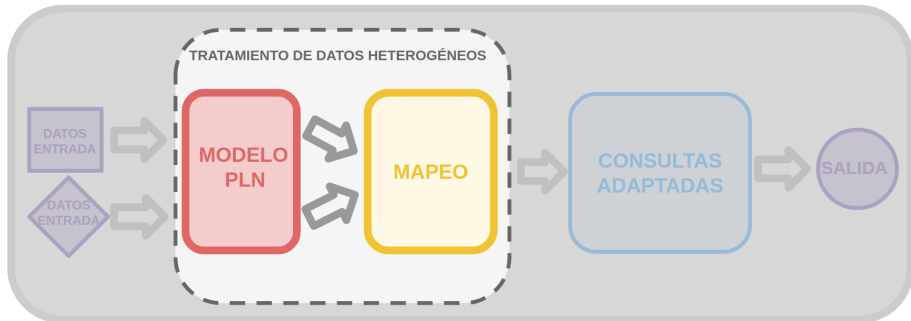
## Distancia difusa entre documentos

Alimento a mapear (i-Diet)	Alimento mapeado (USDA)	Mejor mapeo posible (USDA)	
(1) Sausage Bratwurst	Bratwurst	Bratwurst	✓
(2) Chicory	Chicory beverage	Chicory beverage	✓
(3) Chocolate and cream pudding, Chamburcy	Pie, chocolate cream	Pie, chocolate cream	✓
(4) Swett potatoes	Potato, NFS	Sweet potato NFS	✗
(5) Cocoa and hazelnut butter, Nocilla, Nutela	Hazelnuts	No matches	✗
(6) Sobrasada mallorquina	No matches	No matches	✗

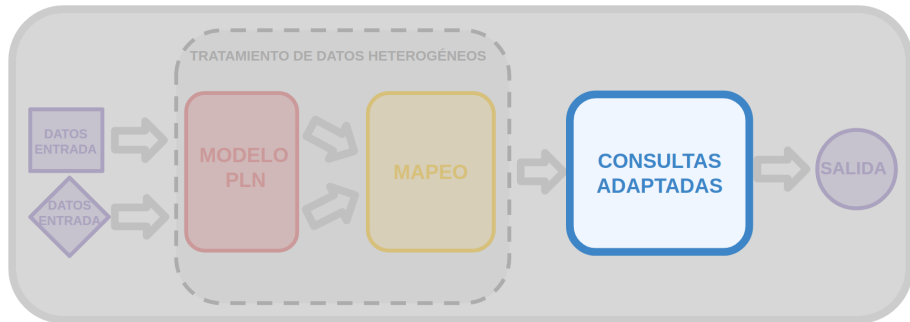
**Tabla:** Resultados obtenidos con la distancia difusa entre documentos

Morales-Garzón, Andrea, et al. *A Word Embedding Model for Mapping Food Composition Databases Using Fuzzy Logic*  
International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer  
International Publishing, pp. 635–647, 2020

# Tratamiento de datos heterogéneos



# Módulo de Consultas Adaptadas



# Módulo de Consultas Adaptadas



# Adaptación de recetas a restricciones

## Datos

- 1 Receta con ingredientes
- 2 Base de datos de composición de alimentos

## Procedimiento

- 1 Se combinan los ingredientes de la receta con la información nutricional de la base de datos de alimentos
- 2 Se detectan los ingredientes que no cumplan la restricción
- 3 **Se adapta la receta:** se cambian los ingredientes restringidos por unos válidos
- 4 Se devuelve la receta adaptada

# Contenido

1 Introducción

2 Objetivos

3 Planificación

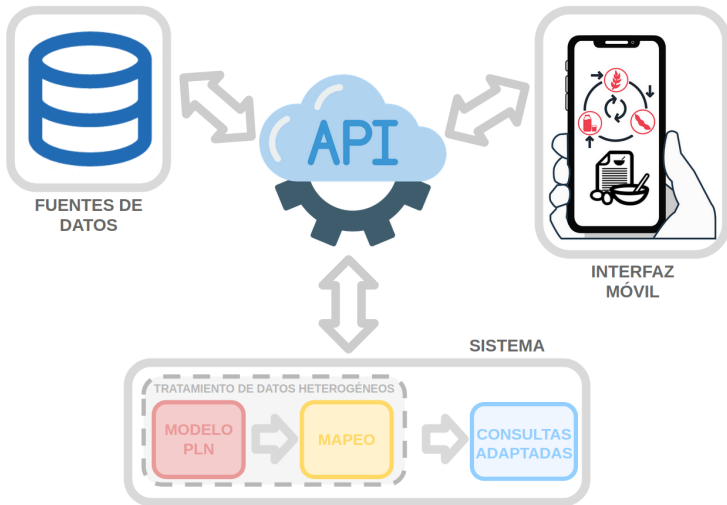
4 Arquitectura del sistema y experimentación

**5 Aplicación**

6 Conclusiones



# Aplicación



# Aplicación

## Fuentes de datos

### ■ Colección de recetas

- Recetas que se pueden adaptar
- Tiene **ingredientes**, pasos de preparación, etiquetado de recetas...
- Hemos añadido datos extra para una mejor experiencia de uso (imágenes y clasificación de recetas)

### ■ Base de datos de composición nutricional (i-Diet)

- Tiene **alimentos con información nutricional**
- Necesaria para consultar si los ingredientes cumplen las restricciones

# Aplicación

## Interfaz móvil

- Permite visualizar el funcionamiento del sistema
- Prototipo funcional
  - Diseño conceptual: primera iteración en el desarrollo

# Aplicación

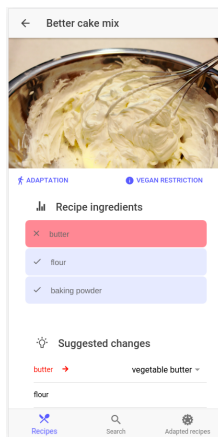
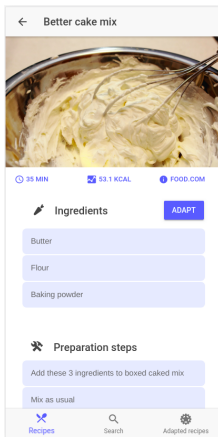
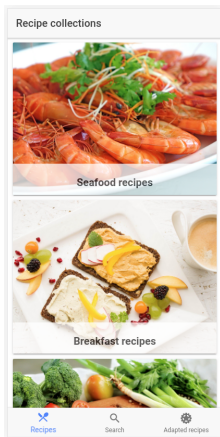
## Interfaz móvil

- Permite visualizar el funcionamiento del sistema
- Prototipo funcional
  - Diseño conceptual: primera iteración en el desarrollo

## Interfaz de Programación de Aplicaciones

- API REST
- Operaciones CRUD sobre las recetas
- Conecta el sistema con las fuentes de datos y la interfaz móvil

# Demostración de la aplicación



**Enlace:** <https://drive.google.com/file/d/1nFgGFtud37NkMOWEgnFR3a-2Ab8KJL6G/view?usp=sharing>

# Contenido

1 Introducción

2 Objetivos

3 Planificación

4 Arquitectura del sistema y experimentación

5 Aplicación

**6 Conclusiones**

# Conclusiones y trabajo futuro

- Word Embeddings capturan la semántica de datos alimenticios
  - Identifica alimentos equivalentes/sustitutos
  - Abarca marcas comerciales
  - Dificultades: problemas en el vocabulario
  - Ausencia de vocabulario: *out-of-vocabulary words*
  - Mejoras: calidad y alcance del corpus, modelos predictivos del lenguaje más sofisticados: fasttext, BERT
- Módulo de mapeo es capaz de detectar equivalentes
  - Medidas sintácticas y semánticas
  - Flexible al nivel de detalle en descripciones (medidas difusas)
  - Resultados aproximados cuando no hay correspondencia exacta
  - Mejoras: combinar con otros atributos no textuales
- Consultas adaptadas
  - Potencia del sistema implementado
  - Adaptaciones intuitivas
  - Mejoras: siguientes pasos en el desarrollo de la app

**¡Gracias por vuestra  
atención!**

**¿Preguntas?**

