Taller de Proyecto

Diplomado de Ciencia e Ingeniería de Datos
Departamento de Ciencias de la Computación
Universidad de Chile

Juan Manuel Barrios
juan.barrios@impresee.com

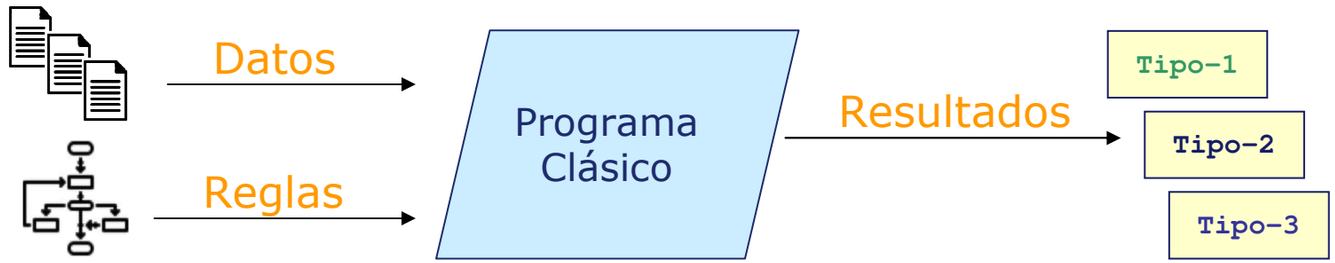
Camila Álvarez
camila.alvarez@impresee.com

6 de agosto de 2018

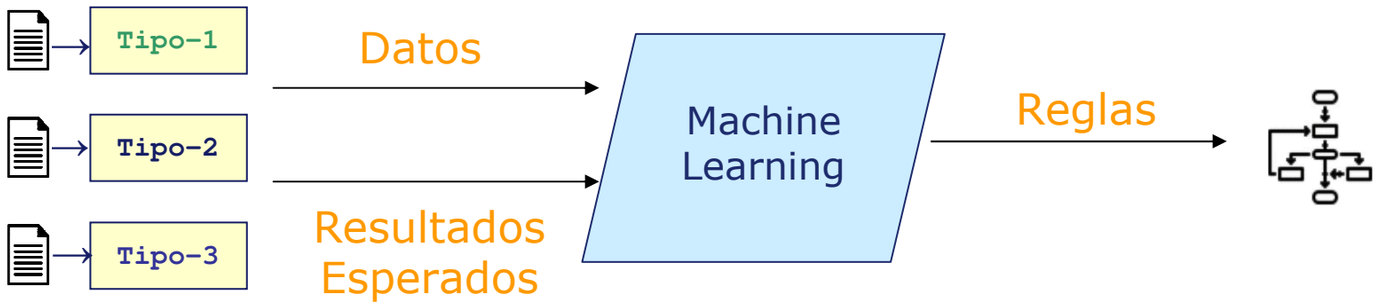
Introducción

Machine Learning

- Programación Clásica: Aplicar reglas sobre datos

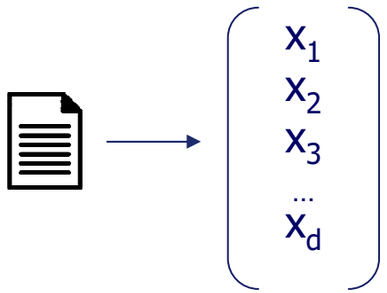


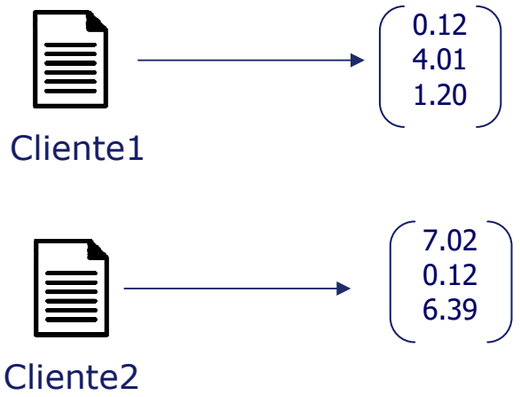
- Machine Learning: Aprender reglas desde los datos



Representación del Contenido

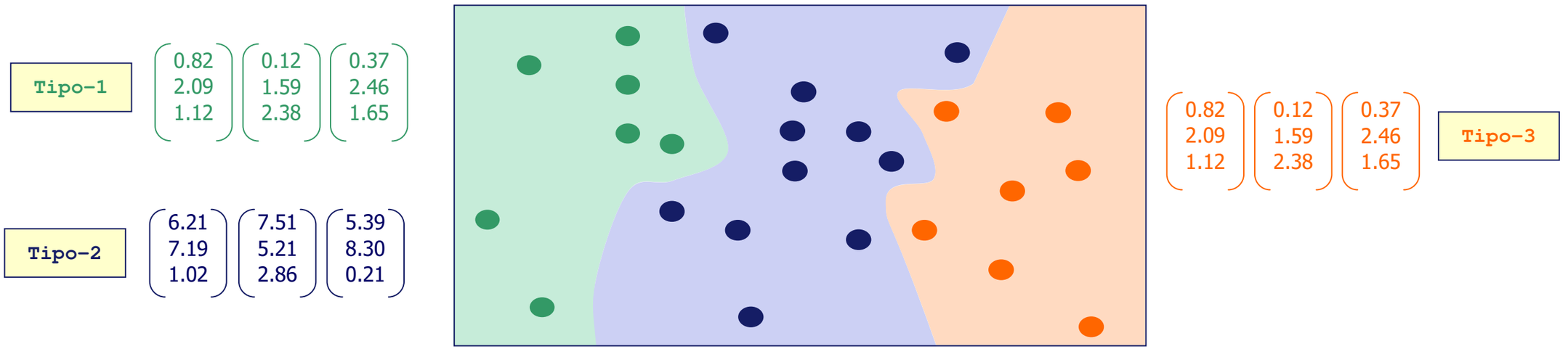
- **Entidad**: Se refiere a un **objeto de interés**, ej.: una persona, un producto, una página web, una foto, etc.
- **Descriptor**: Los datos de la entidad deben ser modelados como un **vector** de dimensión fija
  - Vector Característico o Descriptor





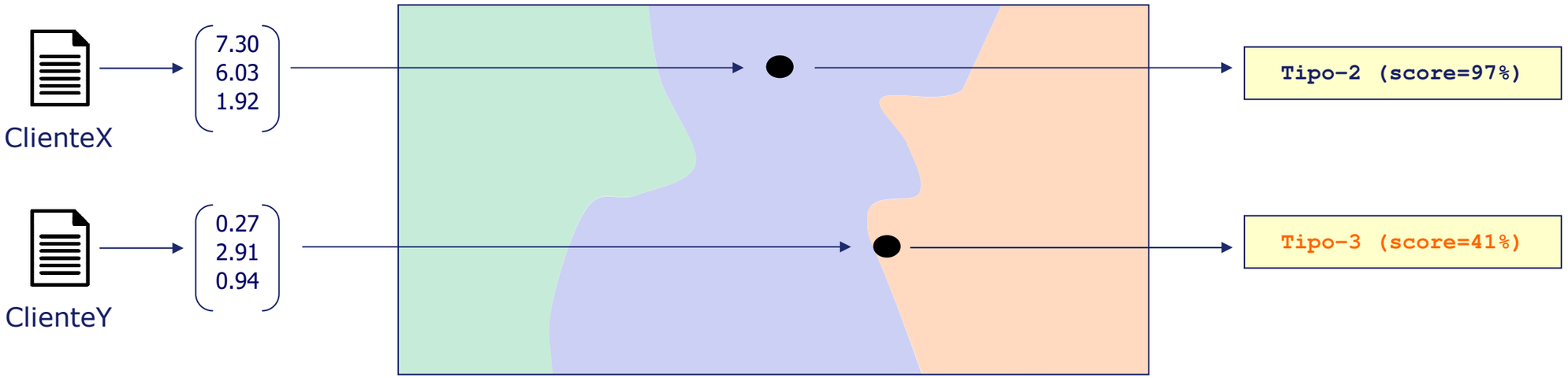
Métodos Supervisados

- La etapa de **entrenamiento** consiste en, dado un conjunto de **vectores** característicos y **etiquetas** asociadas, calcular una **función** que asocia cada vector con su etiqueta
  - Define **zonas** que agrupan vectores con una misma etiqueta
  - Define **fronteras** de separación entre zonas



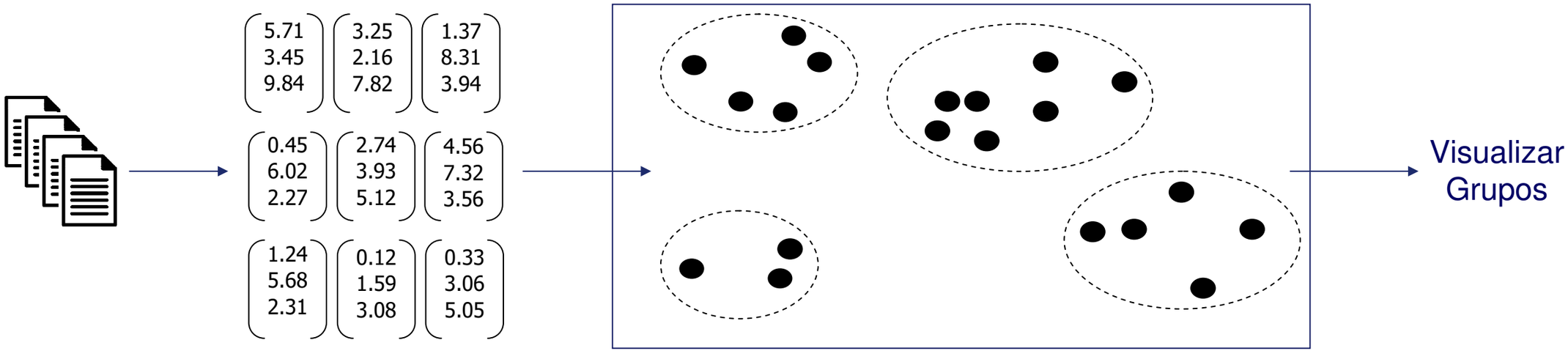
Métodos Supervisados

- La etapa de **clasificación** consiste en **asignar** o **predecir** la etiqueta para una entidad nueva
  - La entidad se **representa** por su vector característico
  - El clasificador determina la **zona** donde está el vector
  - Retorna la **etiqueta** de esa zona junto con un **valor de confianza**



Métodos No Supervisados

- ¿Qué hacer cuando **no existen** o se **desconocen** las **etiquetas**?
  - Obtener vectores característicos de **todas** las entidades
  - **Buscar grupos** de vectores cercanos (clusters)
  - **Revisar** las entidades asignadas a cada grupo (visualización) y **decidir** si la agrupación encontrada tienen algún sentido



Taller de Proyecto

Taller de Proyecto

- **Objetivo del Taller**: Desarrollar un proyecto práctico de análisis de datos reales, aplicando las técnicas y herramientas aprendidas en los cursos del diplomado
- Grupos de **cuatro** personas
- ¿Qué se espera del proyecto?
  - **Analizar** un conjunto de datos de tamaño razonable
  - **Aplicar** conceptos vistos en el diplomado
  - **Extraer información** interesante de los datos

Objetivo del Proyecto

- Cada grupo debe definir un **conjunto de datos** sobre los que desea trabajar
  - Idealmente sobre un dominio donde tengan experiencia
- Se debe definir una **entidad** interesante de estudiar
  - Ejemplo: paciente, cliente, alumno, factura, etc.
- De esa entidad escoger algún **atributo**, **evento** o **comportamiento** que se desee entender mejor
  - Ejemplo: clientes que compran una sola vez, pacientes con cierta enfermedad, facturas sin pagar, etc.
- El objetivo general del proyecto es **analizar y caracterizar** de mejor manera esa entidad y ese comportamiento específico

Datos del Proyecto

- Los datos pueden ser **públicos** o **privados**
  - En caso de datos privados al menos se debe poder mostrar estadísticas, análisis y conclusiones
- Los datos se deben **enriquecer** con datos externos públicos
  - Ejemplo: información de comunas, ingresos promedios, noticias, variables macroeconómicas del país, etc.
- Notar que el proyecto se enfoca en **buscar** y **descubrir** patrones en vez de comprobar hipótesis preconcebidas

Herramientas

- ¿Qué herramientas usar para el proyecto?
  - Es posible usar **cualquier herramienta** a elección del grupo

"Tus herramientas **no** deben determinar cómo respondes una pregunta. **Tus preguntas** son las que **deben determinar** las **herramientas** que usas."

Entregas del Proyecto

- Se realizarán tres entregas del proyecto:
  - Entrega 1: **Preparación** y **exploración** de datos
    - Definir la estructura de los datos, corregir y limpiar datos
  - Entrega 2: Métodos **no supervisados** y **visualización**
    - Encontrar grupos en los datos y evaluarlos visualmente
  - Entrega 3: Métodos **supervisados**
    - Implementar y evaluar un clasificador
- No será parte de este Taller:
  - Procesar grandes volúmenes de datos
  - Implementación de métodos de regresión

Evaluación

- Cada entrega se **evaluará** con dos notas:
  - Nota de trabajo en clases y presentación oral
  - Nota de informe escrito
- Cada **presentación oral** es una exposición en horario de clases
  - Deben estar presentes todos los integrantes del grupo
- El **informe escrito** es un documento PDF que se entrega al final del taller
  - Las mismas slides de la presentación oral con correcciones y mejoras
- La nota final del Taller es el promedio de las seis notas
- **Requisito de Asistencia: 100%**

Sesiones

- Cinco sesiones entre cursos:
  - **Jueves** 30-Agosto (después de programación estadística)
  - **Martes** 25-Septiembre (después de minería de datos)
  - **Miércoles** 17-Octubre (después de aprendizaje de máquinas)
  - **Martes** 6-Noviembre (después de recuperación de información)
  - **Martes** 28-Noviembre (después de visualización de datos)
- Tres sesiones al final del diplomado:
  - **Lunes** 17, **Martes** 18 y **Jueves** 20-Diciembre
- En cada sesión:
  - Trabajo grupal, Responder dudas, Presentaciones

Próxima Sesión

- **Jueves 30 de Agosto 2018**
- Presentación de grupos y proyectos
  - **Integrantes** del grupo
  - **Objetivos** del proyecto
    - Entidad(es) relevante(s)
    - Comportamiento/Evento que se desea estudiar
  - **Fuente de datos** a usar
    - Cantidad de registros, tamaños
  - Fuentes de datos **externas** a integrar