

## Programa de Curso

Segundo Semestre, 2022

Profesor: Juan Reutter  
jreutter@ing.puc.cl  
Clases: martes, módulos 5,6, sala javier pinto  
Atención de alumnos: por cita concertada via correo electrónico

### Descripción

Si bien los sistemas de bases de datos han logrado sobreponerse, en cuanto eficiencia, al tsunami de datos de este siglo, las formas de trabajar estos datos no han sabido actualizarse de la misma manera.

A medida que la capacidad de recopilar datos va creciendo, tareas de análisis de datos que antes eran simples hoy se vuelven engorrosas. Las aplicaciones y sus usuarios demandan a los sistemas de bases de datos para que exista la capacidad de analizar datos complejos con la misma facilidad que se construye un cubo o una vista en SQL.

Durante los últimos años hemos visto una multiplicación de iniciativas que aspiran entender cuáles son los límites de lo que se puede ofrecer a los usuarios de una forma relativamente declarativa. Esto requiere saber resolver los nuevos desafíos algorítmicos que se presentan, estudiar formas de expresar nuevas tareas analíticas, y lograr que estas tareas analíticas sean procesadas (y optimizadas) por los sistemas de bases de datos de la misma forma como un sistema toma una consulta SQL.

Buscando presentar esta problemática, y algunos de los avances y preguntas claves que actualmente tiene el área, este curso tiene dos propósitos. Por un lado, proveer una perspectiva más realista de la naturaleza de las tareas de análisis de datos que son requeridas hoy en día en la práctica, y por otro, discutir sobre los desafíos que supone levantar sistemas de bases de datos que integren soporte para estas tareas de forma nativa. La discusión debe ser técnica y formal, pues de otra forma nos veríamos obligados a navegar entre palabras al aire y conceptos inventados con el único propósito de vender.

### Objetivo General

En cuanto a tecnología, el curso se enfoca principalmente en el paradigma de bases de datos de grafos, y cómo este paradigma es capaz de integrar los últimos avances en cuanto a algoritmos, aproximaciones y aprendizaje de máquinas.

El objetivo de este curso es, entonces, doble. Por un lado se busca que los alumnos sean capaces de usar de la tecnología de bases de datos de grafos, comprender los desafíos actuales de esta tecnología en materia de eficiencia algorítmica, y elaborar pequeños proyectos de investigación que apunten a clasificar y analizar las distintas alternativas. Por el otro lado, se busca que los alumnos estén preparados para integrar nuevos algoritmos, o desarrollar mejoras, en un contexto realista. Finalmente, los alumnos conocerán como diversas tareas de análisis empujan algunos de los avances teóricos más importantes en el área.

### Metodología

Los contenidos del curso se aprenden mediante una mezcla de clases presenciales, videos, lecturas y guías de laboratorio, en el que se trabajará con tecnología de punta en el contexto de datos. Existen además tres tareas de dos semanas en las que los alumnos podrán profundizar en temas que sean de su elección, más un proyecto final.

Los alumnos deberán realizar sus tareas en grupos de tres personas, y cumpliendo además la condición de que los grupos de cada alumno no pueden repetir personas. La forma de entrega de las tareas depende de la naturaleza de estas, aunque por lo general corresponderá a una presentación de unos 20 minutos.

**Qué se espera de los alumnos.** A lo largo del semestre, los alumnos deberán

- Mostrar iniciativa e interés científico, el que se debe traducir en un buen trabajo con las tareas, y poder presentar sus resultados.
- Estar atentos al avance de sus compañeros, y demostrar la capacidad para aprender de sus pares.
- Contribuir a la evaluación del trabajo de sus compañeros (se entregará un formulario).
- Trabajar en un proyecto final, presentando un extended abstract y un video de sus resultados.
- Inscribirse para ver, evaluar y discutir sobre algunos de los proyectos finales (usaremos una plataforma especial para ello).

## Evaluación

Se realiza en base a

- El cumplimiento de las tareas que se esperan de los alumnos.
- El desarrollo y posterior presentación de las tareas.
- El desarrollo y posterior presentación del proyecto.

Para aprobar el curso, los alumnos necesitan colaborar en la evaluación de los videos, discutir los trabajos de sus compañeros, presentar sus videos y obtener una nota igual o superior a 4 en TODAS las actividades del curso.

De aprobarse el ramo, la nota final se calcula como  $= (0,4 \cdot NP + 0,2 \cdot NT1 + 0,2 \cdot NT2 + 0,2 \cdot NT3)$ , donde NP es la nota del proyecto y NT1, NT2 y NT3 son las notas de las tareas. De reprobar el ramo, la nota final es el mínimo entre la ecuación de arriba y un 3,9.

## Contenidos

### Introducción

Distintas soluciones para requerimientos de datos.  
¿Cómo poder combinarlas?

### Algoritmos discretos para bases de datos

1. Breve repaso de conceptos clásicos: índices, sort-merge join, planes de consulta
2. Cota AGM
3. Algoritmos óptimos en el peor caso
4. Lectura de tópicos de investigación reciente

### Algoritmos probabilísticos para datos masivos

6. Filtros de bloom

7. Sketches

8. Laboratorio: aplicaciones prácticas

### Análisis de grafos basado en embeddings

11. Embeddings de nodos
12. Mecanismos de paso de mensaje
13. Redes neuronales de grafos
14. Laboratorio: aplicaciones prácticas

## Calendario de Clases

El calendario podrá estar sujeto a cambios durante el semestre, en cuyo caso se avisará oportunamente.

Fecha	Clase	Descripción
09/08	A1:	Introducción, conceptos clásicos, cota AGM
16/08	A2:	Continuación cota AGM, algoritmos óptimos en el peor caso, otros tópicos
23/08	A3:	lectura de artículos
30/08	A4:	presentaciones
06/09	A5:	presentaciones
13/09	P1:	Estructuras probabilísticas para el manejo de datos
20/09	P2:	Laboratorio práctico
27/09	P3:	trabajo / consulta
04/10	P4:	Presentación de proyectos
11/10	G1:	Embeddings de grafos, paso de mensaje
18/10	G2:	Laboratorio práctico
25/10	G3:	Laboratorio práctico / trabajo
08/11	G4:	Presentación de proyectos
15/11	PF1:	trabajo / consulta proyectos finales
22/11	PF2:	trabajo / consulta proyectos finales
por definir	PF3:	Presentación de proyectos finales

## Bibliografía

Dado el carácter del curso, existe poca bibliografía formal sobre el tema. Durante el curso se irán entregando papers de lectura obligatoria, y papers de lectura complementaria.

## Otros

El Departamento de Ciencias de la Computación adopta una política de tolerancia-cero frente a copias o plagios. Se sugiere revisar las política y penalidades que el departamento establece ante estas acciones.

En particular, se pide a los alumnos **citar** todo el material que usen para el desarrollo de las tareas y del proyecto. La inclusión de material de terceros sin la cita apropiada se considera como plagio, y es penalizado con una nota 1.1 en el curso.

El curso tiene dos canales de comunicación oficiales: Las clases y la página Web (en GitHub). Se asume que que toda la información que es entregada por ambos canales llega a todos los alumnos. Por lo mismo, se sugiere a los alumnos revisar la página Web constantemente.