# CPSC 304 – Administrative notes Nov. 29 and Dec. 3, 2024

- Today is optional – no material will be on the final
  - Clickers today will not be for points
- Next class: Apriori exercise & review (no new material)
- Project: don't forget to go to your demo!
- Regular office hours end with the end of classes
  - Additional office hours will be posted
- Final exam: December 16 at 12pm! Osborne A

# Today's class

- Ethics in Data Mining
- What's after 304 at UBC?
- Database research
- Office hours with remaining time

# Research? See ACM SIGKDD, etc. Conferences and Proceedings

- Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)
  - Related organizations: VLDB (Very Large Data Bases) and ACM SIGMOD (Special Interest Group on the Management of Data)
- SIGKDD has participants from 3 major disciplines:
  - Artificial Intelligence (especially Machine Learning)
  - Statistics
  - Databases
- All 3 disciplines deal heavily with algorithms.
  - AI has an emphasis on supervised and unsupervised learning.
  - Statistics has an emphasis on exploratory data analysis, probabilities, inference, and validation.
  - DB has an emphasis on managing large volumes of disk-resident data, especially with respect to I/Os and scalability.

# Tuesday is finishing data mining & review. What topics would you like to have covered?

- normalization (especially 3NF, especially minimal covers)
- sql: aggregations, group bys, having, etc.
- ER diagrams (but on midterm 1, so not prioritized)
- HRU algorithm for which views to materialize in data warehousing
- relational algebra (especially division)
- common mistakes
- division in SQL was requested, but will not be on the final

# CPSC 304 – December 7, 2021 Administrative Notes

- Tutorial this Monday & Tuesday: office hours
- Project Milestone 6 DONE INDIVIDUALLY
  Due today
- Goal: have all project milestones graded by December 10
  - Reminder: you have a week after a grade is released to request a regrade
- Regular office hours end today. More posted on Piazza
  Final exam: December 14 @3:30pm
  - Cumulative, but currently weighted to material past midterm #1
  - We have three rooms – rooms will be posted on Piazza

# There's more to data mining than just algorithms

Let's take a bit to think about how data mining is impacting the world around us

# An example data mining quote from the NY Times:

"We have the capacity to send every customer an ad booklet, specifically designed for them, that says, 'Here's everything you bought last week and a coupon for it,' " one Target executive told me. 'We do that for grocery products all the time.' But for pregnant women, Target's goal was selling them baby items they didn't even know they needed yet."

https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

# Target can identify pregnant women and send them individual mailings

Is this cool, creepy, or both?

A. Cool
B. Creepy
C. Both
D. Neither

# Target: Cool, creepy or both

| Cool | Creepy |
| --- | --- |
| Deep insight to take from not a lot of information. | Creepy because people don't really consent |

# Okay. What if I told you…

People pay different insurance (life or medical) rates based on being pregnant. Insurance companies may now want to pay Target for this information. Does this change your opinion?

A. It was already creepy
B. It wasn't creepy before, but this is
C. No, still not creepy

# On the other hand…

Grocery store loyalty cards help trace hepatitis A outbreak:

http://o.canada.com/health-2/grocery-store-loyalty-cards-help-b-c-disease-detectives-trace-hepatitis-a-outbreak

# Clicker question: Loyalty cards and credit cards

After thinking about this, are you more or less likely to use a credit card/loyalty card for purchases:

A. More likely
B. Less likely
C. The same

# And then there's Facebook

**Full page apology in many US and British papers**

"You may have heard about a quiz app built by a university researcher that leaked Facebook data of millions of people in 2014. This was a breach of trust, and I'm sorry we didn't do more at the time. We're now taking steps to make sure this doesn't happen again.

We've already stopped apps like this from getting so much information. Now we're limiting the data apps get when you sign in using Facebook.

We're also investigating every single app that had access to large amounts of data before we fixed this. We expect there are others. And when we find them, we will ban them and tell everyone affected.

Finally, we'll remind you which apps you've given access to your information — so you can shut off the ones you don't want anymore.

Thank you for believing in this community. I promise to do better for you.

Mark Zuckerberg"

# Okay, they're sorry. Obviously won't happen again, right?

## Text anyone in your phone

Continuously upload info about your contacts like phone numbers and nicknames, and your call and text history. This lets friends find each other on Facebook and helps us create a better experience for everyone.
Learn More.

TURN ON

NOT NOW

Manage your contacts

# What actually happened?

"*Ars Technica* reports that Facebook has been requesting access to contacts, SMS data, and call history on Android devices to improve its friend recommendation algorithm and distinguish between business contacts and your true personal friendships."

"Several Twitter users have reported finding months or years of call history data in their downloadable Facebook data file"

"Facebook has responded to the findings, but the company appears to suggest it's normal for apps to access your phone call history when you upload contacts to social apps."

https://www.theverge.com/2018/3/25/17160944/facebook-call-history-sms-data-collection-android

# With great power comes great responsibility…

Is this okay? Why or why not?

A. Yes, it's okay.  B. No, it's not okay

# And there's this

An executive order signed by President Trump in January 2017 states:

> "Agencies shall, to the extent consistent with applicable law, ensure that their privacy policies exclude persons who are not United States citizens or lawful permanent residents from the protections of the Privacy Act regarding personally identifiable information."

# But even if you trust Facebook with this data, there are still concerns

"**The Trump administration has said it wants to start collecting the social media history of nearly everyone seeking a visa to enter the US.**

The proposal, which comes from the state department, would require most visa applicants to give details of their Facebook and Twitter accounts.

They would have to disclose all social media identities used in the past five years.

About 14.7 million people a year would be affected by the proposals."

http://www.bbc.com/news/world-us-canada-43601557

# A question to you

Do you think the US should be able to collect social media history of those entering the US? Why or why not?

A. Yes B. No

# Let's make this more personal

Imagine you are working at Facebook as a developer. What are your responsibilities w.r.t. people's privacy when collecting, storing, and mining data?

A. It's not my responsibility
B. I have some responsibility
C. I have a large responsibility

Not right to put the pressure on individuals.
What makes the corporation is the employees

# Learning Goals Revisited

- Define the term *knowledge discovery*.
- Explain the general steps involved in the *knowledge discovery in databases* (KDD) process
- Comment on the benefits and challenges that data mining has when dealing with imperfect data quality, especially in large datasets (e.g., data mining can point out anomalies (outliers), optimize the use of human time, detect patterns in data (including patterns that are there just by chance).
- Explain the value of finding frequent itemsets and association rules. Provide some real-world examples of their use (e.g., retailing, biology).
- Explain the purpose of association rules.
- Apply the Apriori Algorithm and compute frequent itemsets and association rules (by hand, for a small dataset).

# Reading list (from 3rd edition)

- Chapter 1: Intro
- Chapter 2: ER diagrams
- Chapter 3: The relational model
- Chapter 19: Normal forms: Sections 19.1 – 19.6
- Chapter 4: Relational Algebra all except 117-122, and No DRC
- Chapter 24: Sections 24.1 - 24.3 (Datalog)
- Chapter 5: SQL (No syntax for check constraints, assertions, or triggers)
- Chapter 25: Data Warehousing
- Chapter 26: Data mining

# What's next if you liked 304?

- For classes, you can take:
  - CPSC 404 (Database internals)
  - CPSC 504 (Data management research)
  - CPSC 534L (Information Networks)

# CPSC 404

- The role of an RDBMS in an organization's data management strategy.
- The relationship among bytes, pages, disks, buffer pools, data tables, indexes, metadata, etc.
- Indexing strategies
- SQL query evaluation and optimization
- Transaction Processing and Concurrency Control
- Crash Recovery and Application Recovery

# CPSC 504 -Data Management Research (My grad class)

- Relational roots (history, what happened, query optimization research, query optimizer research, query execution)
- Adaptive execution
- Object-oriented & object-relational databases
- XML
- What Goes Around Comes Around
- No-SQL/New SQL
- Student requests

# CPSC 534L: Information Networks (Laks Lakshmanan's grad class)

- Modeling, Prediction, & Optimization
- Finding Communities
- Information/Influence/Infection Propagation: modeling and optimization.
- Recommender Systems: models, composite/holistic recs.
- Knowledge Graph Completion, Question Answering, & Fact Checking

# What else can you do to learn more about an area?

- Reading groups!
  - Faculty are usually happy to have you join – send them email
- Independent studies (CPSC 448/449)
- USRAs and other paid options
- Volunteering

# What is data management research?

- Research about managing data including:
  - Traditional (relational) database management systems
    - What they are, how to make them work
  - Other kinds of databases
    - Object-oriented, XML, No SQL
  - Other data management applications
    - OLAP, data mining, etc.

# What is *my* research

- How to help people understand and explore their data,
- How to manage data that is currently not well supported by databases
- How data can be managed in situations where there are multiple databases.

# Your TA Jianhao's research

- Data lakes/open data are often not designed to be easy to accessed
  - Use LLMs to generate labels/column names for tables in data lakes
  - LLMs hallucinate. Knowledge graphs are error-prone and hard to write. Combine LLMs & knowledge graphs to help answer queries in data lakes.
  - Develop a data lake query interface to help users navigate through their data