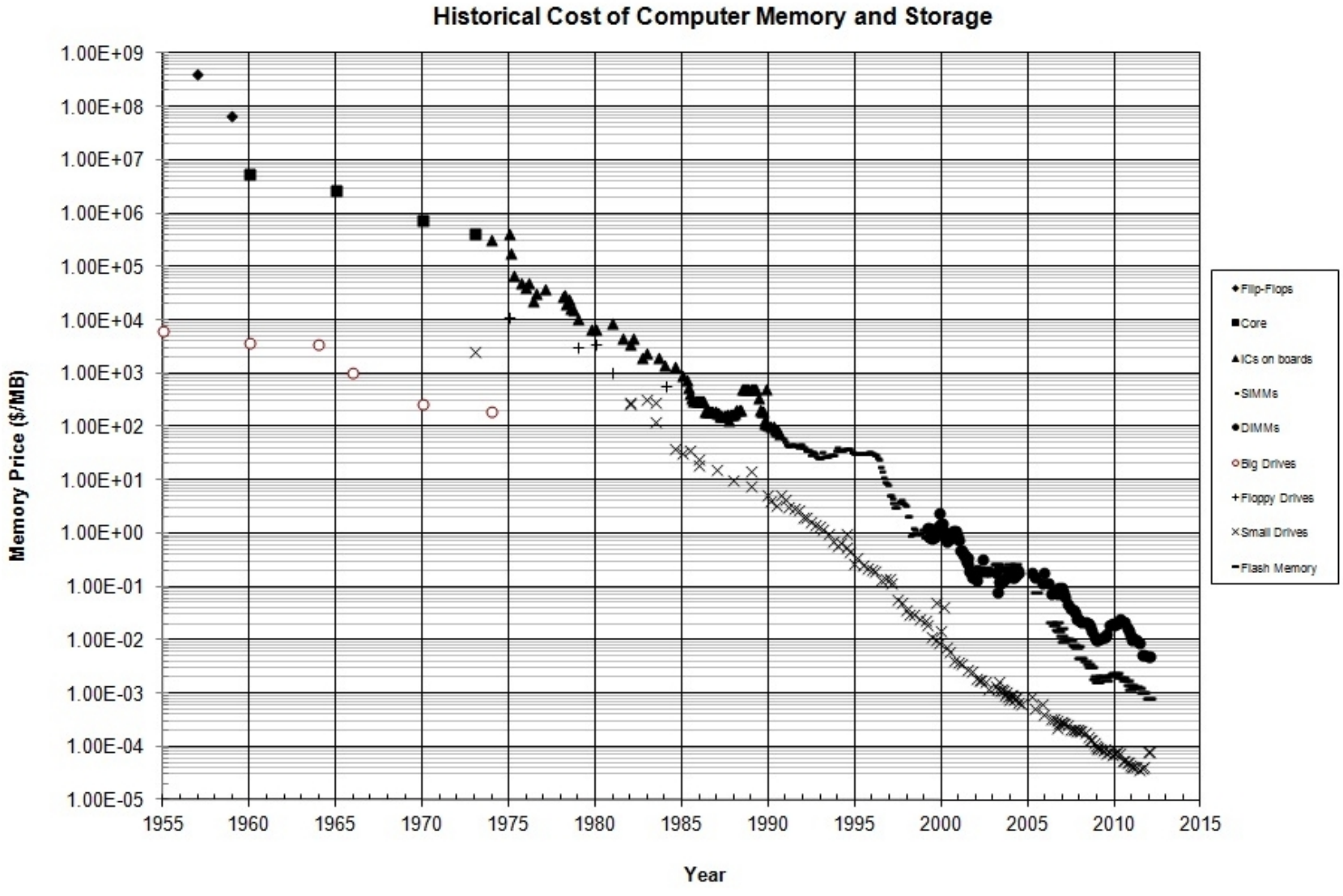
Computer Science 1000: Part #7

Computer Databases

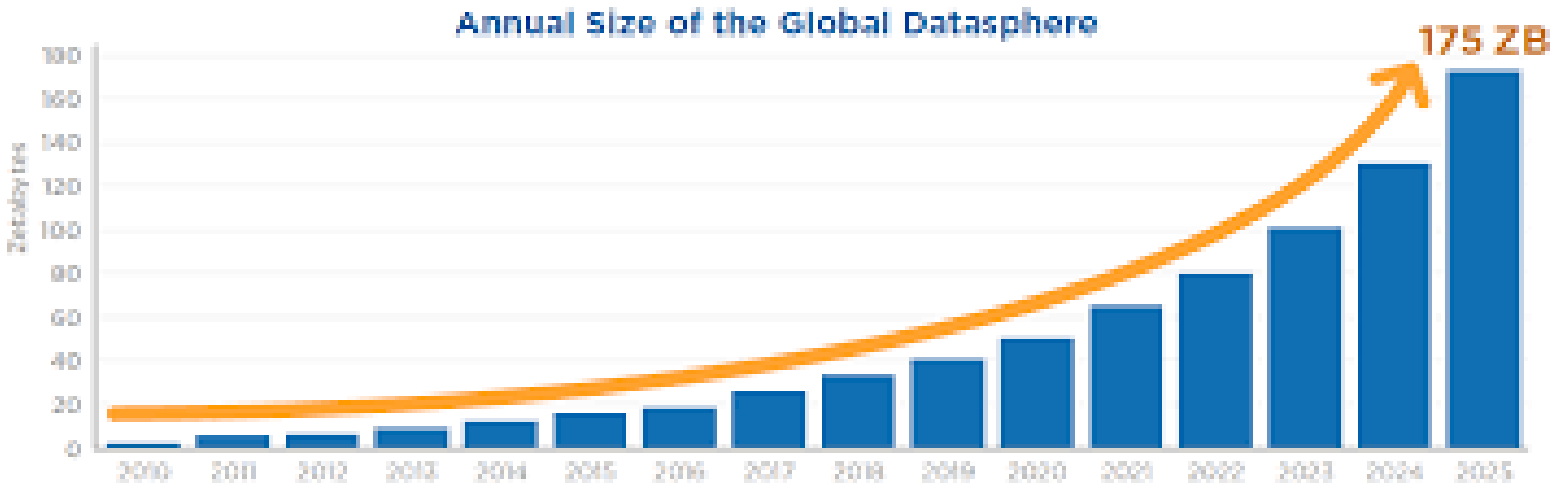
- Stored Data: An Overview
- Relational Databases
- Big Data and Data Mining
- Data Privacy

Stored Data: An Overview
The Computer Memory Cost Implosion



Stored Data: An Overview (Cont'd)
The Stored Data Explosion

Figure 1 – Annual Size of the Global Datasphere



(ZB (Zettabyte) $= 10^{21}$ (sextillion) bytes)

Stored Data: An Overview (Cont'd)
The Stored Data Explosion

MEMORY $\neq$ DATA

- Raw bytes require both context and accessibility to become data.
- Data must be accessible and usable, but not *too* accessible and usable (e.g., the Cambridge Analytica scandal).
- Focus first on accessibility and usability and then on privacy.

Relational Databases

- A **database management system (DBMS)** imposes an organization on information stored in memory, i.e.,

  Bytes are combined to form fields
  Fields are combined to form records
  Records are combined to form files
  Files are combined to form databases

  A DBMS also incorporates operations by which data can be accessed and manipulated.
- Each DBMS is based on an abstract database model. We will here consider the most popular of these models, the **Relational Model** proposed by E. F. Codd at IBM in 1970.

Relational Databases (Cont'd)

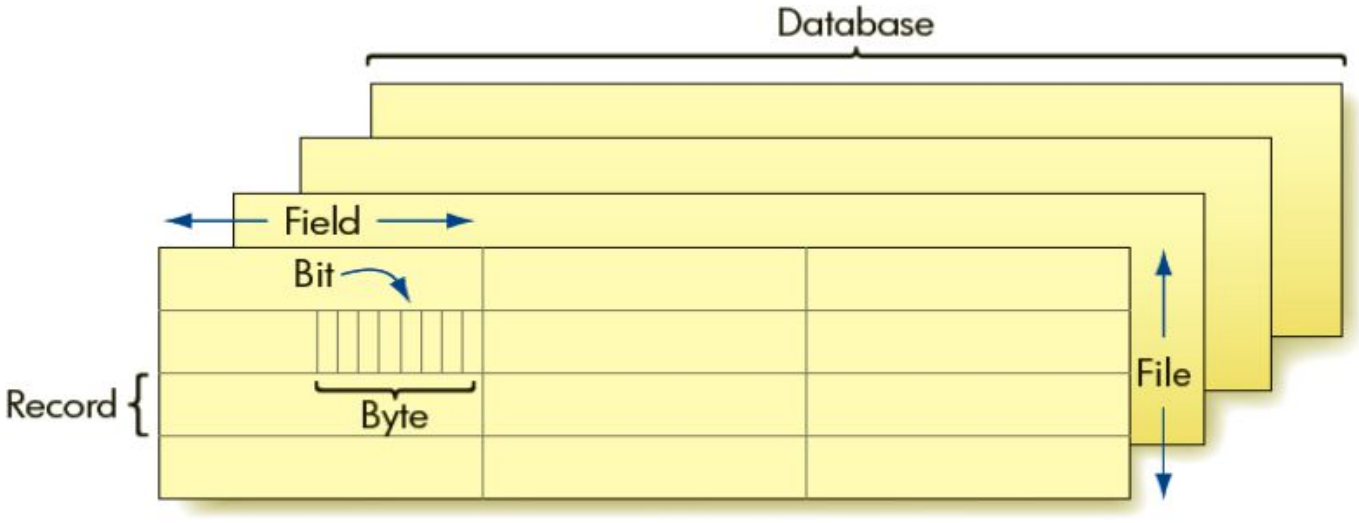


Figure 14.3 Data Organization Hierarchy

Relational Databases (Cont'd)

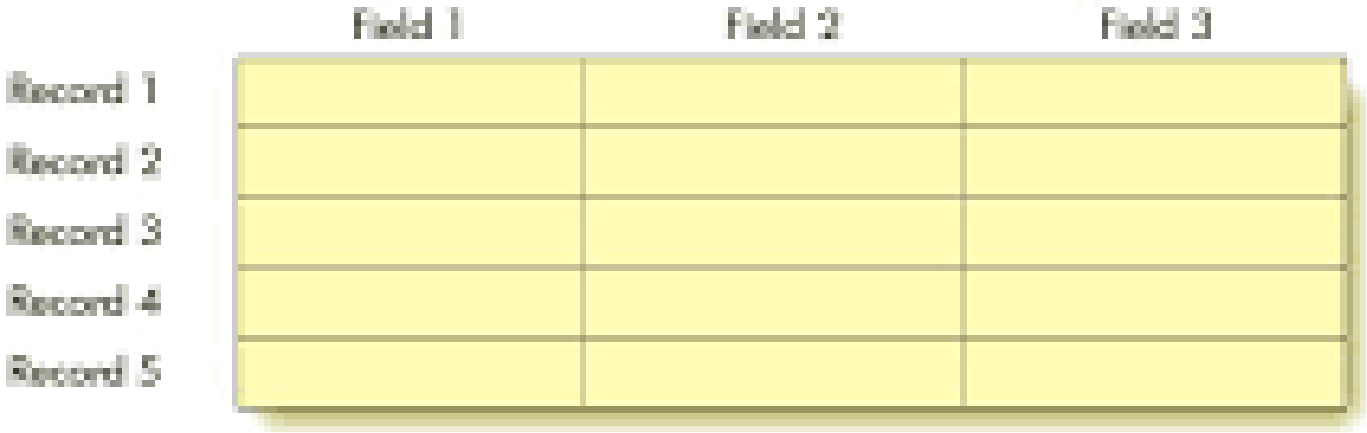


Figure 14.4 Records and Fields in a Single File

Relational Databases (Cont'd)

- In Codd's Relational Model, each file is represented as a table, which encodes information about a particular **entity**, e.g., employees, insurance plans purchased by employees.
- A row in a table encodes information about a particular instance of an entity, e.g., an employee's ID number, name, birthdate, pay rate per hour, and number of hours worked, and is called a **tuple**.
- A column in a table encodes a particular piece of information about an instance of an entity, e.g., employee last name, and is called an **attribute**.

Relational Databases (Cont'd)

ID	LastName	FirstName	Birthdate	PayRate	HoursWorked
149	Takasano	Frederick	5/23/1966	$12.35	250

Figure 14.5 One Record in the Rugs-For-You Employees File

Relational Databases (Cont'd)

ID	*LastName*	*FirstName*	*BirthDate*	*PayRate*	*HoursWorked*
116	Kay	Janet	3/29/1976	$16.60	94
149	Takasano	Frederick	5/23/1986	$19.35	250
171	Kay	John	11/17/1974	$17.80	245
165	Honou	Morris	6/9/1993	$6.70	53
123	Perreira	Francine	8/15/1989	$8.50	185

A **primary key** is an attribute or combination of attributes that uniquely identifies a tuple, e.g., employee ID number in the Employees table, employee ID number and plan type in the InsurancePolicies table (indicated by underlining)

Relational Databases (Cont'd)

FIGURE 14.7

InsurancePolicies

EmployeeID	PlanType	DateIssued
171	B2	10/18/1994
171	C1	6/21/2002
149	B2	8/16/2008
149	A1	5/23/2006
149	C2	12/18/2011

InsurancePolicies table for Rugs-For-You

Information about an instance of an entity can be split across multiple tables using **foreign keys**, e.g., employee ID number in the Employees and InsurancePolicies tables. This reduces the amount of redundant information that is stored.

Relational Databases (Cont'd)

Select information from a single table stored in a relational DBMS using the **Structured Query Language (SQL)**, e.g.,

```
SELECT LastName, PayRate
FROM Employee
WHERE LastName = "Perreira"
```

LastName	*PayRate*
Perreira	$8.50

Relational Databases (Cont'd)

```
SELECT *
FROM Employee
ORDER BY ID
```

ID	*LastName*	*FirstName*	*BirthDate*	*PayRate*	*HoursWorked*
116	Kay	Janet	3/29/1976	$16.60	94
123	Perreira	Francine	8/15/1989	$8.50	185
149	Takasano	Frederick	5/23/1986	$19.35	250
165	Honou	Morris	6/9/1993	$6.70	53
171	Kay	John	11/17/1974	$17.80	245

Relational Databases (Cont'd)

```
SELECT *
FROM Employee
WHERE PayRate > 15.00
```

ID	*LastName*	*FirstName*	*BirthDate*	*PayRate*	*HoursWorked*
116	Kay	Janet	3/29/1976	$16.60	94
149	Takasano	Frederick	5/23/1986	$19.35	250
171	Kay	John	11/17/1974	$17.80	245

Relational Databases (Cont'd)

Using foreign keys, can manipulate information stored across several tables, e.g.,

```
SELECT LastName, FirstName, PlanType
FROM Employees, InsurancePolicies
WHERE Lastname = "Takasano"
      AND FirstName = "Frederick"
      AND ID = EmployeeID
```

LastName	*FirstName*	*PlanType*
Takasano	Frederick	B2
Takasano	Frederick	A1
Takasano	Frederick	C2

Big Data and Data Mining: Overview

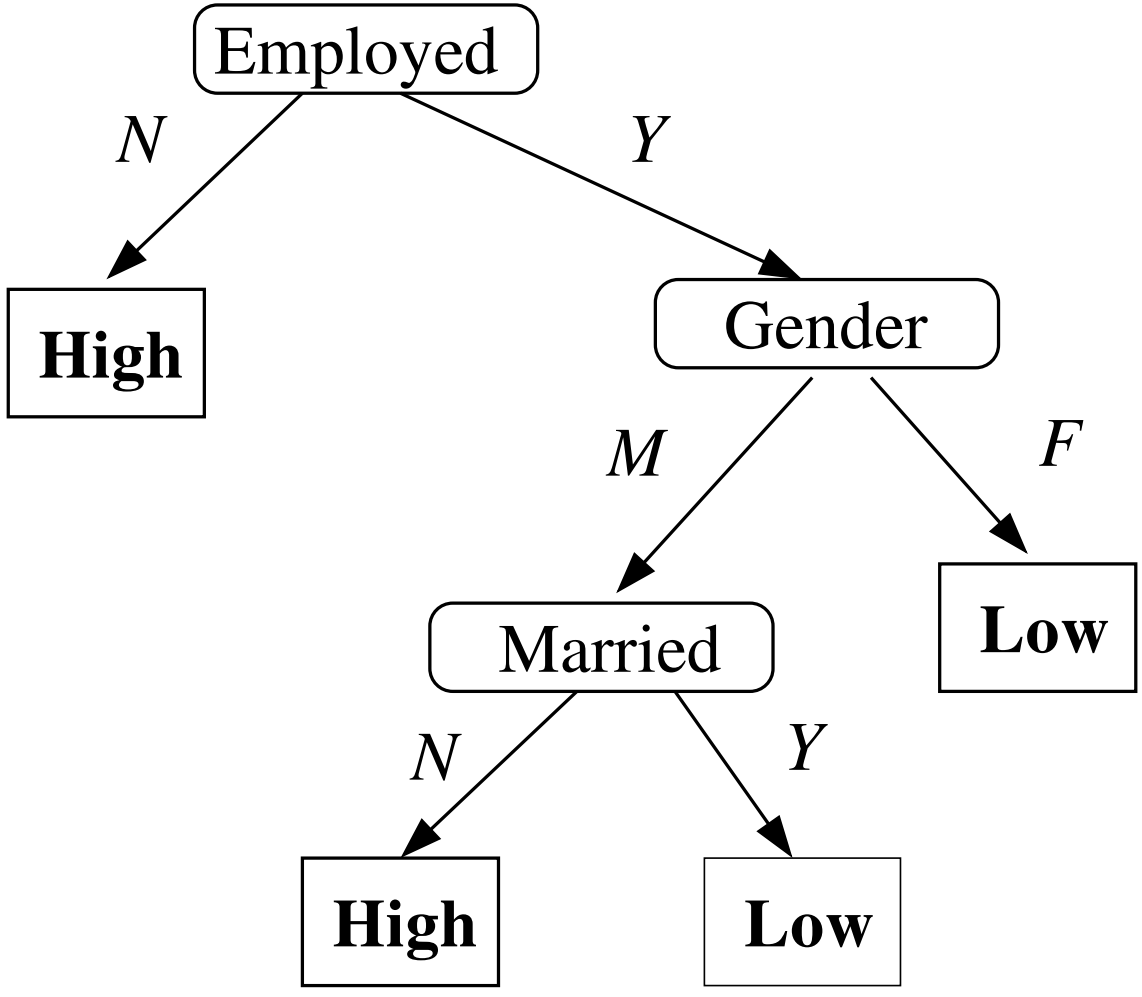
- Classical data analysis involves finding the right data, formatting that data in an appropriate database, and writing queries to answer specific questions, e.g., relational databases and SQL queries.
- Finding the right data and formatting it for examination is a problem in itself with the advent of truly massive datasets (**Big Data**), e.g., detailed multi-media health data for large populations, social media and online retailer datasets (Facebook, Twitter, Amazon).
- Modern data analysis (**data science** / **data analytics**) also incorporates advanced statistics, visualization, and pattern-finding capabilities (**data mining**); can not only answer specific questions but can also find (hopefully useful) trends and patterns in data ("pattern fishing").

Big Data and Data Mining: An Example

ID	*Employed*	*Gender*	*Married*	*Risk*
1	Y	M	Y	Low
2	Y	F	N	Low
3	N	M	N	High
4	Y	M	Y	Low
5	Y	F	Y	Low
6	N	F	N	High
7	Y	M	N	High
8	N	M	N	High
9	Y	F	N	Low
10	Y	M	Y	Low

Existing data for bank loan risk (Figure 14.9, Textbook)

Big Data and Data Mining: An Example (Cont'd)



Decision tree for bank loan risk (Figure 14.10, Textbook)

Data Privacy: Overview

- Big Data and data mining allow the extraction of patterns that are of public (e.g., detection and tracking of disease outbreaks, evaluation of treatment outcomes) and commercial (e.g., targeted advertising, product recommendations) use.
- Need to balance access to data with personal privacy.
- To service commercial needs, **data brokers** have emerged which accumulate and integrate publicly-available and commercial datasets, with disconcerting results, e.g., Latanya Sweeney medical records reconstruction.
- Personal rights often obscure wrt original collection of data, and become moreso with secondary aggregation (“You aren’t the customer, you’re the product”).

Data Privacy:
The Evolution of Stored Data

local	$\Rightarrow$	networked / distributed
use-specific	$\Rightarrow$	detailed / overall
short-term	$\Rightarrow$	(very) long-term
user-accessible	$\Rightarrow$	anyone*-accessible
bulky	$\Rightarrow$	(very) portable
one copy	$\Rightarrow$	(very) many copies
hard to copy	$\Rightarrow$	(very) easy to copy
authority-verified	$\Rightarrow$	anyone*-verified

Data Privacy: Joys and Perils of Stored Data

Joys	Characteristics	Perils
		Store false / misleading easily
Store anything easily	Storage easy	Find false / misleading easily
Find anything easily	Store anything	Integrate / reconstruct easy
Spread anything easily	Store anytime	Steal anything easily
Everything remembered	Store forever	Spread impossible to stop
Personal customization		Nothing forgotten
		Personal commercialization

In addition to the use of certain technologies (e.g., fake detection, k-anonymization), appropriate governance and laws are critical in mitigating the perils above; so is responsible behaviour by individual people.

Data Privacy: k-anonymity

"Blur" data-identifying fields such that each original entity is indistinguishable from at least $(k-1)$ others for $k \geq 2$, e.g.,

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	Diarrhea
4	25	15	Fever
5	22	28	Flu
6	32	35	Fever
7	38	32	Flu
8	35	25	Diarrhea

(a) Sensitive table

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	0-20	10-30	Flu
2	0-20	10-30	Fever
3	20-30	10-30	Diarrhea
4	20-30	10-30	Fever
5	20-30	10-30	Flu
6	30-40	20-40	Fever
7	30-40	20-40	Flu
8	30-40	20-40	Diarrhea

(b) 2-anonymous Table

Surviving and Thriving with Big Data

- Limit degree of personal (esp. commercial) exposure online
  - Know privacy settings and use appropriately
- Limit types of personal exposure online
- Learn crap detection and online research skills (Rheingold)
- Be aware of what's going on privacy-wise both technologically and commercially

"Don't Panic" – *The Hitchhiker's Guide to the Galaxy*
"Let's be careful out there" – *Hill Street Blues*

. . . And If You Liked This . . .

- MUN Computer Science courses on this area:
  - COMP 2007: Introduction to Data Management
  - COMP 3401: Introduction to Data Mining
  - COMP 4754: Database Systems
- MUN Computer Science professors teaching courses / doing research in in this area:
  - Mark Hatcher
  - Lourdes Pena-Castillo
  - Jian Tang