

Today

```
Welcome |
    | Course Information |
    | Data Modeling |
    | Memory |
    | File |
    | Database
```

Title	Fundamentals of Data Engineering
<u>Code</u>	COE848
Term	Winter 2019
Time	Tuesday 09AM-12PM
Location	KHE225

<u>Name</u> Hossein Fani

Education Ph.D. Candidate

Faculty of Computer Science

University of New Brunswick, NB., Canada

Office ENG324

Hour Tuesday 12PM-01PM

Email hossein.fani@ryerson.ca

Grading	5	
Labs and Final Project	35%	
Midterm Exam (W08: Feb. 26, 2019)	25%	
Final Exam	40%	

Location

Assignments

Deadlines

ENG411

Steps of one final project

7 days

Late submission receives none!

Laboratory × Deadline

Lab#	No. of Weeks	Due Date
Lab 1	1	Jan. 23
Lab 2	1	Jan. 30
Lab 3	2	Feb. 13
Lab 4	2	Mar. 06
Lab 5	2	Mar. 20
Lab 6	2	Apr. 03
All sections	have same submission dead	dline, i.e., Thursdays Midnight

Welcome | Course Information | Data Modeling | Memory | File | Database

Laboratory × TAs

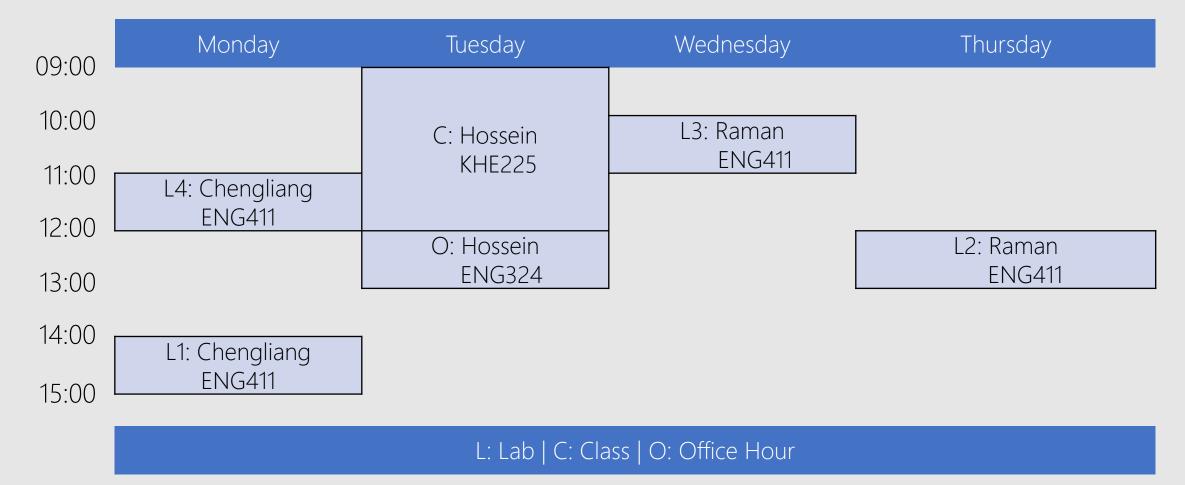


Chengliang Huang c28huang@ryerson.ca



Raman Deep Singh raman.singh@ryerson.ca

Laboratory × Time



Welcome | Course Information | Data Modeling | Memory | File | Database

Book 10

Title

Authors

Publisher

Edition

<u>ISBN</u>

Database Systems: The Complete Book Hector Garcia-Molina

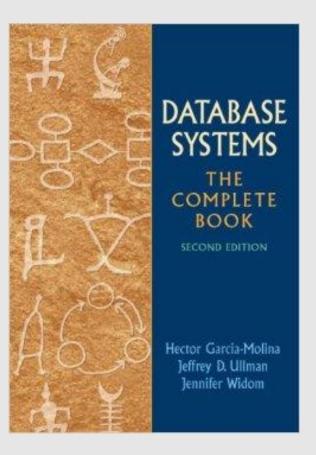
Jeffrey D. Ullman

Jennifer D. Widom

Prentice Hall

2nd Edition, 2008

0131873253



Course Outline Done

Lab Guide Done

Lab Deadlines Done

Slides End of Each Week

Announcements Check D2L for course updates & information!

12

Running Example

Title

Use Case

Test Case

Moviesion

Similar Movie Recommendation

Input: "2001: A Space Odyssey"

Output: 1. "Interstellar"

2. "Her"

3. "Metropolis"

4. ...

Name

IDE

Code Convention

Naming Convention

Java (why?)

NetBeans

www.oracle.com/technetwork/java/codeconventions-150003.pdf

www.oracle.com/technetwork/java/codeconventions-135099.html

Package

Class

Variable

Method

Constant

```
all.lower.case
```

```
package ca.ryerson.ee.coe848
```

Singular Noun, PascalCase

```
public class Movie{}
public class MovieProductionCompany{}
```

Noun, camelCase

```
Movie oscarBestPicture = new Movie();
ArrayList<Movie> oscarWinners = new ArrayList<>();
```

Verb, camelCase

```
public void addMovie(Movie newMovie);
```

ALL UPPER CASE

```
final int MAX_MOVIE_TITLE_WIDTH = 255;
```

Policy KINDLY DO NOT DISTRACT OTHERS!

Phone Of

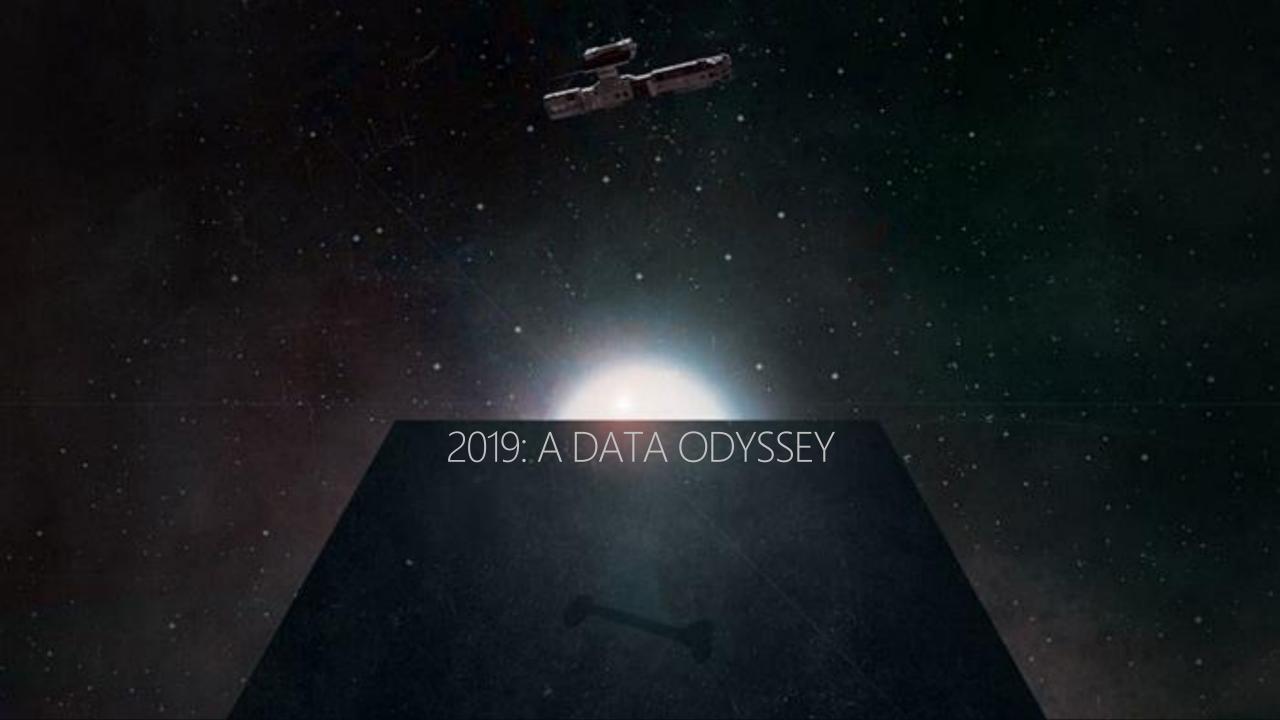
<u>Silent</u>

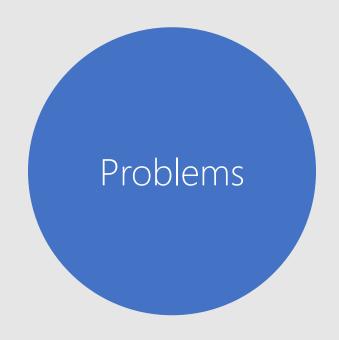
Laptop Closed

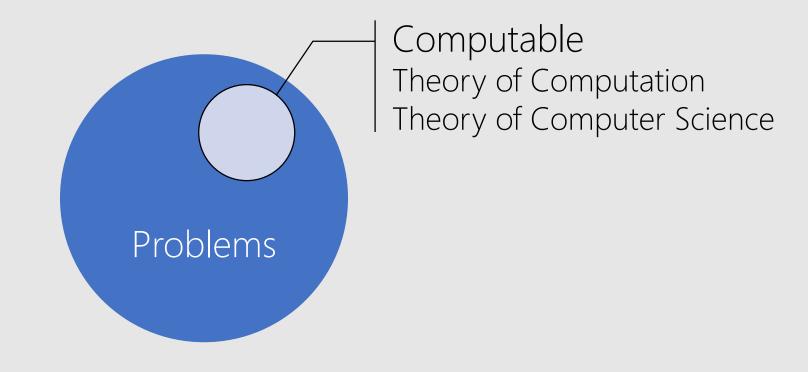
Far End of Each Row

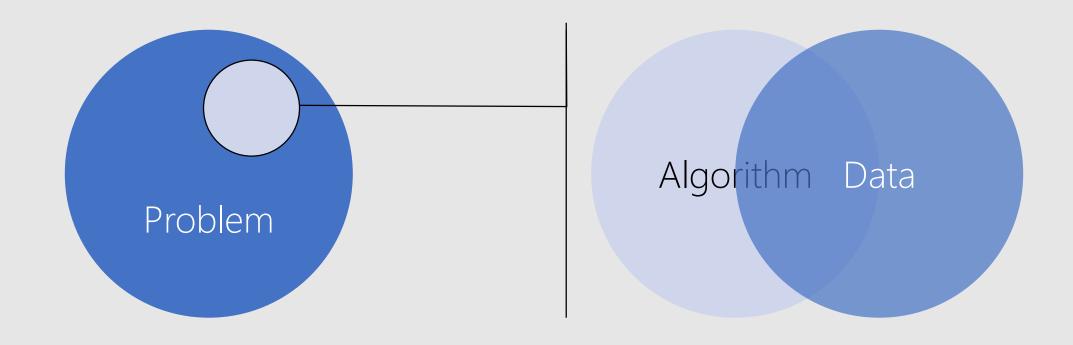
Break 10:30 - 10:45

Others www.ryerson.ca/studentguide/learning_experience/

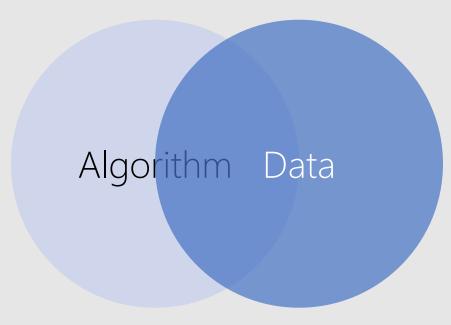








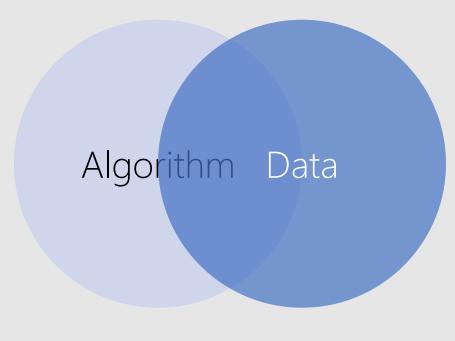
There is interaction between <u>data engineers</u> and <u>algorithm designers</u>.



Data gets bigger!
Algorithms get slower!

Better Data Modeling -> Faster Algorithm e.g., ?

There is interaction between <u>data engineers</u> and <u>algorithm designers</u>.



Data gets bigger!
Algorithms get slower!

Better Data Modeling >> Faster Algorithm e.g., Sorting in Arrays vs. Trees e.g., Searching in Arrays vs. Trees e.g., Sparse Matrix in Arrays vs. Linked Lists

Algorithm Design Algorithm Analysis Artificial Intelligence (AI) Machine Learning Data Mining



Data Structure (Memory)
File Structure (File Systems)
Database Management Systems
Data Warehouse
Big Data
Cloud

24

Algorithm Design Algorithm Analysis Artificial Intelligence (Al) Machine Learning Data Mining



Data Structure (Memory)
File Structure (File Systems)
Database Management Systems
Data Warehouse
Big Data
Cloud

25

```
Data Modeling: Real World Entity

Conceptual Level |

Logical Level |

Physical Level

Computable Entity
```

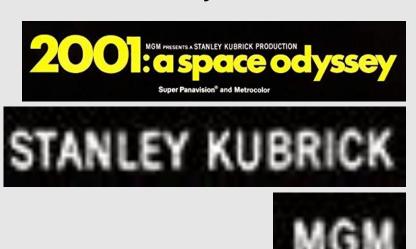
- 1. Identify real world <u>entities</u>, <u>attributes</u>, <u>relationships</u>
- 2. Create Schema

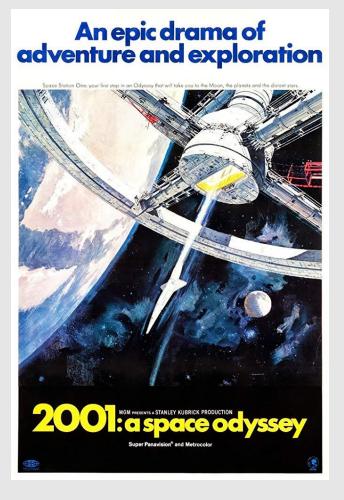
Entities for a movie recommender system:

Movie

Director

Company





Movie Attributes:

Title

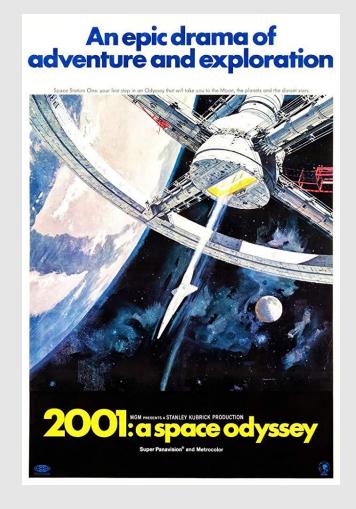
Genre

Language

RunningTime

Poster

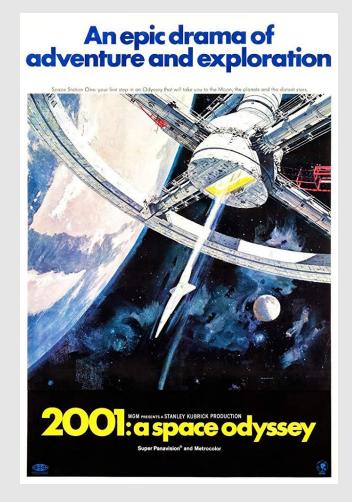
• • •



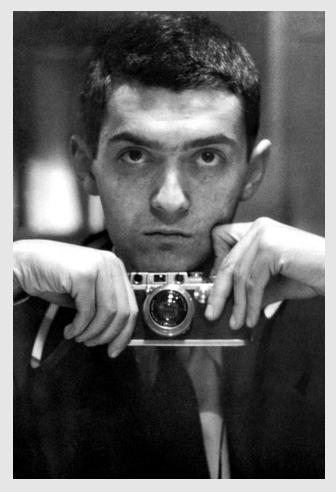
```
Movie Attributes' Type (More Optional Details):
   Title (string)
   Genre (string)
   Language (string)
   RunningTime (decimal)
   Poster (?)
...
```



```
Movie Attributes' Type (More Optional Details):
   Title (string)
   Genre (string)
   Language (string)
   RunningTime (decimal)
   Poster (byte[])
...
```



```
Director Attributes:
Name (string)
DateOfBirth (date)
PlaceOfBirth (string)
Photo (byte[])
```



```
Company Attributes:
Name (string)
Address (string)
DateOfEstablishment (date)
Logo (byte[])
```



Movie Schema

Movie Data <u>Instance</u>

Title (string)

2001: A Space Odyssey

Genre (string)

Sci-fi

Language (string)

English

RunningTime (decimal)

142

Poster (byte[])



Poster (byte[])

Movie Schema

Title (string)

Genre (string)

Language (string)

RunningTime (decimal)

Movie Data Instance

Rosemary's Baby

Horror

English

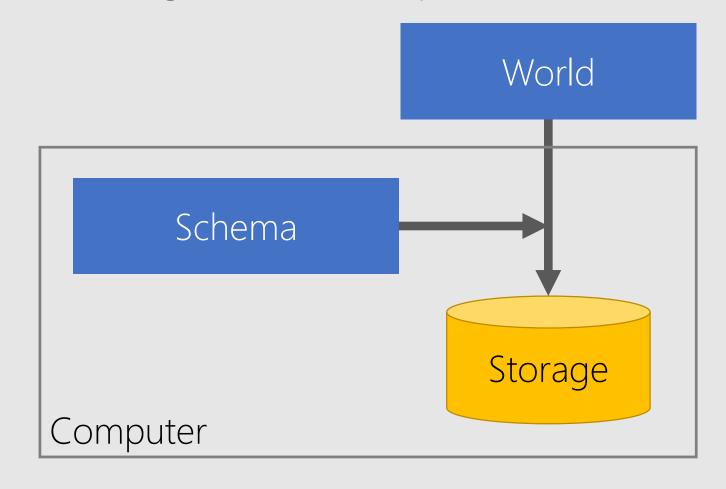
35

Movie Schema	Title (string)	Genre (string)	, ,	RunningTime (decimal)	Poster (byte[])
Movie Instance 1	2001: A Space Odyssey	Sci-fi	English	142	
Movie Instance 2	Rosemary's Baby	Horror	English	136	

Director Schema	Name (string)	DateOfBirth (date)	PlaceOfBirth (date)	Photo (byte[])
Director Instance 1	Stanley Kubrick	1949	USA	
Director Instance 2	Roman Polanski	1933	Poland	

Company Schema	Name (string)	Address (string)	DateOfEstablishment (date)	Logo (byte[])
Company Instance 1	Metro-Goldwyn-Mayer Studios Inc.	Beverly Hills, CA, USA	1924	THE COLUMN AND SHARE
Company Instance 2	Warner Bros	null	1923	

Data Schema	Data Instance
Data about Data: Meta-data	Actual Data
Defined at Setup Time	Inserted at Running Time
Rarely Change	Rapidly Change
	Must Conform to Schema



40

Movie Schema	Title (string)	Genre (string)	Language (string)	RunningTime (decimal)	Poster (byte[])
Movie Instance 1	2001: A Space Odyssey	Sci-fi	English	142	And the second s
Movie Instance 2	Rosemary's Baby	Horror	English	136	
Director Schema	Name (string)	DateOfBirth (date)	PlaceOfBirth (date)	Photo (byte[])	
Director Instance 1	Stanley Kubrick	1949	USA		
Director Instance 2	Roman Polanski	1933	Poland		
Company Schema	Name (string)	Address (string)	DateOfEstablishment (date)	Logo (byte[])	
Company Instance 1	Metro-Goldwyn-Mayer Studios Inc.	Beverly Hills, CA, USA	1924	TALE TO SAFETY AND A SAFETY AND	
Company Instance 2	Warner Bros	null	1923		

1
-

Name (string) No R	Date Of Birth elationship	PlaceOfBirth (date)	
	1949		
Roman Movie × Dir	rector × C	ompany	

Welcome | Course Information | Data Modeling | Memory | File | Database

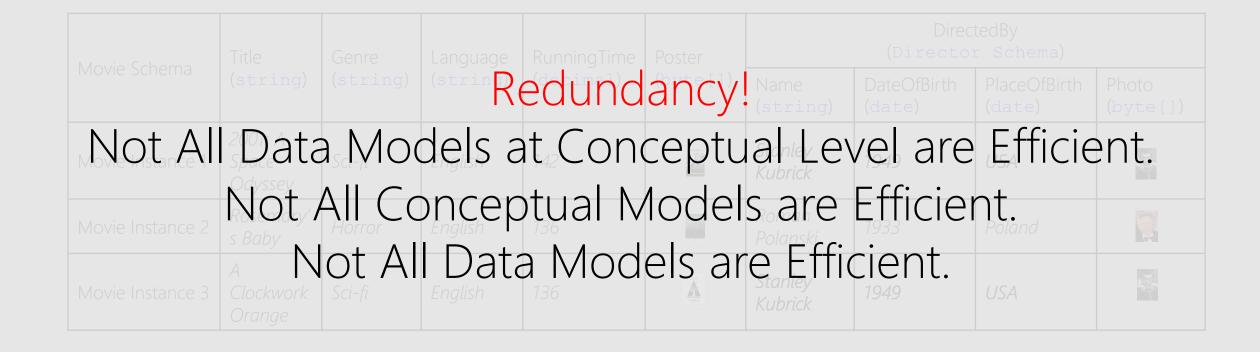
42

Movie × Director × Company Relationships:

```
Director <u>makes</u> Movie | Movie is made by Director
Company <u>distributes</u> Movie | Movie is distributed by Company
Company <u>budgets</u> Movie | Movie is budgeted by Company
```

Mayia Cahama	Title Genre Language RunningTime Poster		Language RunningTime		Poster	DirectedBy (Director Schema)			
Movie Schema	(string)	(string)	(string)	(decimal)	(byte[])	Name (string)	DateOfBirth (date)	PlaceOfBirth (date)	Photo (byte[])
Movie Instance 1	2001: A Space Odyssey	Sci-fi	English	142	2	Stanley Kubrick	1949	USA	
Movie Instance 2	Rosemary' s Baby	Horror	English	136		Roman Polanski	1933	Poland	9

May in Calapan	Title	Genre	Language (string)	Language RunningTime		DirectedBy (Director Schema)			
Movie Schema	(string)	(string)) (string) (decimal) (byte[]) Name	ng) (decimal)	Name (string)	DateOfBirth (date)	PlaceOfBirth (date)	Photo (byte[])
Movie Instance 1	2001: A Space Odyssey	Sci-fi	English	142		Stanley Kubrick	1949	USA	
Movie Instance 2	Rosemary' s Baby	Horror	English	136	40	Roman Polanski	1933	Poland	9
Movie Instance 3	A Clockwork Orange	Sci-fi	English	136	**************************************	Stanley Kubrick	1949	USA	4.0



46

Practice1: Data Modeling for Ryerson <u>Library</u> at Conceptual Level

Entities: ?

Attributes: ?

Relationships: ?

Practice1: Data Modeling for Ryerson <u>Library</u> at Conceptual Level

Entities: Book, Staff, Person, ...

Attributes:

Book: ISBN, Title, ...

Staff: Name, Salary, Rank, ...

Student: Studentld, Name, ...

Relationships: Student borrows Book from Staff

48

Practice2: Data Modeling for Ryerson <u>University</u> at Conceptual Level

Entities: ?

Attributes: ?

Relationships: ?

Data Modeling: Real World Entity

Conceptual Level |

| Logical Level |

| Physical Level Computable Entity

50

- 1. How entities, attributes, relationships should be <u>represented</u>.
- 2. <u>Update</u> Schema

We already have seen one representation, haven't we?

Data Modeling × Logical Level

We already have seen one representation, haven't we? Tabular!

Movie Schema	Title (string)	Genre (string)		RunningTime (decimal)	Poster (byte[])
Movie Instance 1	2001: A Space Odyssey	Sci-fi	English	142	2
Movie Instance 2	Rosemary's Baby	Horror	English	136	· ·

Data Modeling × Logical Level

There are other representations as well.

There are other logical models as well.

There are other data models at logical level as well.

~1960: Object Oriented

1969: Relational: Mathematical Relationships

1996: XML

Data Modeling × Logical Level

There are other representations as well.

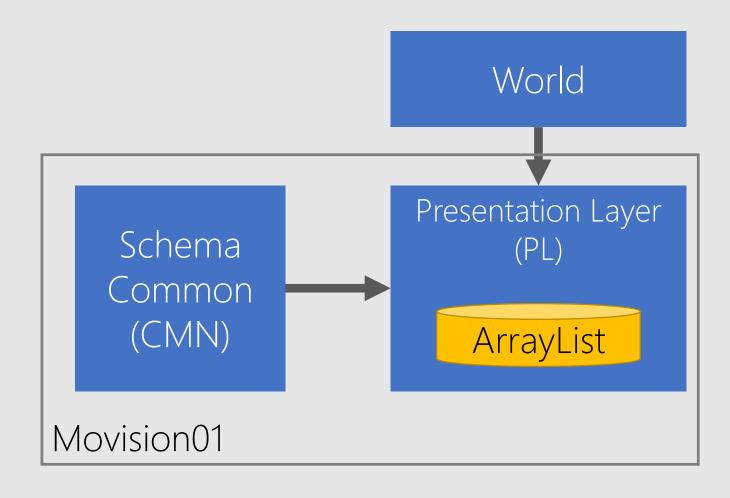
There are other logical models as well.

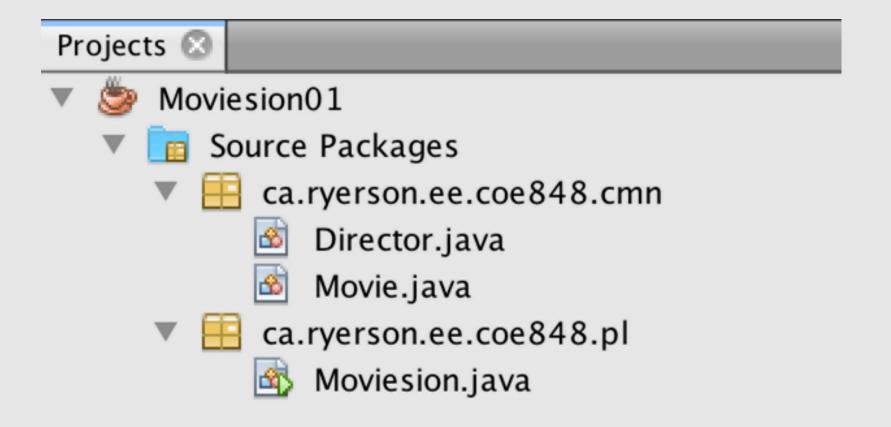
There are other data models at logical level as well.

~1960: Object Oriented

1969: Relational: Mathematical Relation

1996: XML





Conceptual Model	Transformation	Logical Model by OO
Data Schema	DDL	Class
Data Instance	DML	Object

Data Definition Language (DDL): Commands for setting up schema of data Data Manipulation Language (DML): Commands to manipulate data, also called "query language"

```
Conceptual Model
                                       Logical Model by OO
                             public class Movie {
Movie Schema
                               String title;
  Title
                               String genre;
  Genre
                               String language;
  Language
  RunningTime
                               Decimal runningTime;
  Poster
                               Byte[] poster;
```

Conceptual Model

Logical Model by OO

Movie Instance

Movie Object

Title	Genre	Language	RunningTime	Poster
2001: A Space Odyssey	Sci-fi	English	142	Annajir drama of adventure and appleration adventure and appleration

```
Movie newMovie = new Movie();
newMovie.setName("2001: A Space Odyssey");
newMovie.setGenre("Sci-fi");
newMovie.setLanguage("English");
newMovie.setRunningTime(142);
newMovie.setPoster(?);
```

Cons

```
Pros
Time | Random Access Memory (RAM) | Fast
Variety of Data Structures
Array, Linked List, Stack, Queue, Tree, HashTable, ...
Easy DML in Memory
SELECT: E element = ArrayList<E>.get(int index);
INSERT: ArrayList<E>.add(int index, E element);
DELETE: ArrayList<E>.remove(int index);
UPDATE: ArrayList<E>.set(int index, E element);
Any Computation
Sorting, Searching, ...
```

Space Far Too Small Expensive <u>NOT DURABLE</u> Volatility | Transient | Short-term Retention Portability

61

```
Data Modeling: Real World Entity

Conceptual Level |

Logical Level |
```

| Physical Level Computable Entity

1. Where and how are the <u>data instances</u> stored in <u>physical storage</u>.

We already have seen one physical storage, haven't we?

There are other physical storages as well.

There are other physical models as well.

There are other data models at physical level as well.

File

Database

Cloud

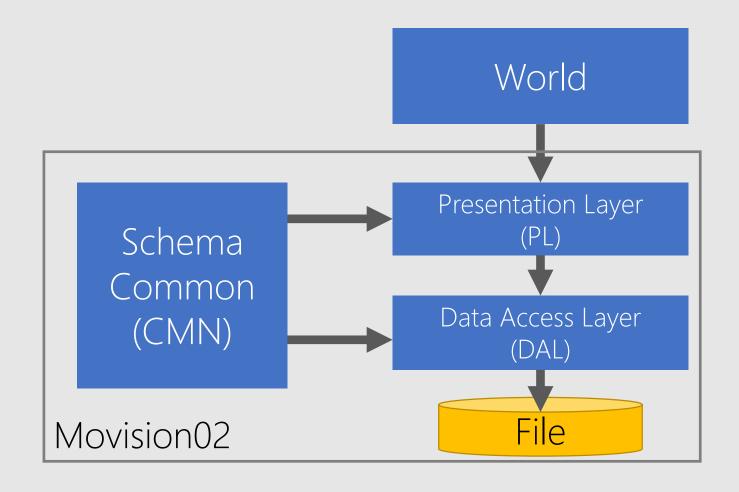
There are other physical storages as well.

There are other physical models as well.

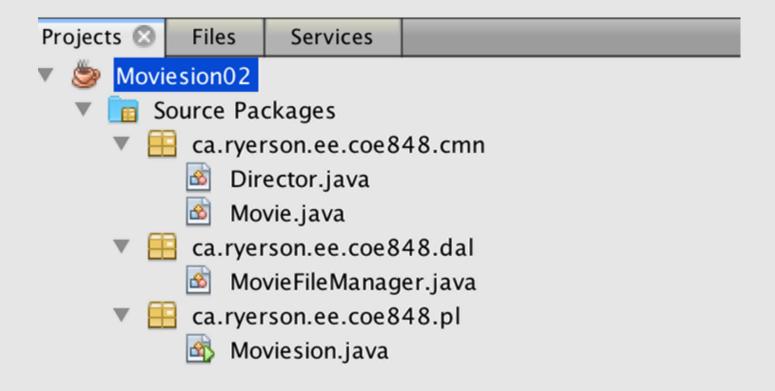
There are other data models at physical level as well.

File Database Cloud

Physical Level × File



Physical Level × File



Conceptual Model	Logical Model by OO	Physical Model by File
Data Schema	Class	CSV File: nothing
		JSON File: nothing
		XML File: XSD
Data Instance	Object	Record

```
69
```

```
Real World Entity

Conceptual Level | Instance

| Logical Level | Object
| Physical Level | Record

Computable Entity
```

Pros Cons

Space

Time Hard Disk Drive (HDD) Sequential Access Method

Electromechanical

DURABLE
Persistency | Long-term Retention

Not Easy DML DELETE | UPDATE

No ACID Properties

Portable

Solid State Drive (SSD) | Random Access Memory | Expensive

Physical Level × File × Cons

Not Easy DML

DELETE

Strategy#1: Create a new file without the deleted record.

Strategy#2: Shift all following over the deleted record.

Strategy#3: Logical Deletion | Invalidate Record

UPDATE

Strategy#1: Replace bytes of old value with bytes of new value

Strategy#2: Delete Record + Insert Record

Physical Level × File × ACID

Transaction

A group one or more operations (DDL|DML) into a single unit of work.

```
BEGIN TRANSACTION

INSERT Movie
INSERT Movie's Director IF NOT EXIST
INSERT Movie's distributor Company IF NOT EXIST
END TRANSACTION
```

Physical Level × File × <u>A</u>CID

Atomicity

All-or-nothing execution of transaction

BEGIN TRANSACTION

INSERT Movie

INSERT Movie's Director IF NOT EXIST

INSERT Movie's distributor Company IF NOT EXIST

END TRANSACTION

Transaction Committed.

3 records affected.

```
Atomicity
```

All-or-nothing execution of transaction

BEGIN TRANSACTION

INSERT Movie

Error!

Partial Execution of Transaction!

ROLLBACK any changes, i.e., Movie record!

Atomicity

All-or-nothing execution of transaction

Another example in online banking?

Physical Level × File × <u>A</u>CID

Atomicity

All-or-nothing execution of transaction

BEGIN TRANSACTION

Withdraw Money from Hossein's Saving Account

Deposit Money to Hossein's Checking Account

END TRANSACTION

77

Consistency

Respect constraints or expectations among data instances

CONSTRAINT#1: All Movies Must Have ReleaseDate ≥ 1890

CONSTRAINT#2: All Movies Must Have at Least One Director

CONSTRAINT#3: All Movies Must Have UNIQUE Title

Consistency

Respect constraints or expectations among data instances

BEGIN TRANSACTION

INSERT Movie

INSERT Movie's distributor Company IF NOT EXIST

END TRANSACTION

Transaction <u>Conflicts</u> with CONSTRAINT#2. ROLLBACK any changes.

Consistency

Respect constraints or expectations among data instances

BEGIN TRANSACTION

INSERT Movie

Error: INSERT Movie's Director IF NOT EXIST

INSERT Movie's distributor Company IF NOT EXIST

FND TRANSACTION

Transaction Committed.

2 records affected.

Partial Execution of Transaction → Transaction Conflicts with CONSTRAINT#2

Inconsistency was due to lack of atomicity.

ACID properties are <u>not orthogonal</u> (independent).

Isolation

Transaction <u>appear to be</u> executed <u>as if</u> no other transaction is executing at the same time

Multi-user environment!

Isolation

Transaction <u>appear to be</u> executed <u>as if</u> no other transaction is executing at the same time

Transaction Queue (Sequential Execution, One-at-a-time)				
T1	W1	T1.1. INSERT 2001: A Space Odyssey		
	W2	T1.2. INSERT Stanley Kubrick IF NOT EXIST		
	W3	T1.3. INSERT MGM IF NOT EXIST		
T2	W4	T2.1. INSERT A Clockwork Orange		
	W5	T2.2. INSERT Stanley Kubrick IF NOT EXIST		
	W6	T2.3. INSERT Warner Bros IF NOT EXIST		
T3	W7	T3.1. UPDATE Director SET PlaceOfBirth = "United States of America" WHERE PlaceOfBirth = "USA"		
T4	W8			

Isolation

Transaction <u>appear to be</u> executed <u>as if</u> no other transaction is executing at the same time

Transaction Queue (Parallel Execution)				
W1	T1.1. INSERT 2001: A Space Odyssey	T2.1. INSERT A Clockwork Orange		
W2	T1.2. INSERT Stanley Kubrick IF NOT EXIST	T1.3. INSERT MGM IF NOT EXIST	T2.3. INSERT Warner Bros IF NOT EXIST	
W3	T2.2. INSERT Stanley Kubrick IF NOT EXIST			
W4	T3.1. UPDATE Director SET PlaceOfBirth = "United States of America" WHERE PlaceOfBirth = "USA"			

Isolation

Transaction <u>appear to be</u> executed <u>as if</u> no other transaction is executing at the same time

Transaction Queue (Parallel Execution)				
W1	T1.1. INSERT 2001: A Space Odyssey	T2.1. INSERT A Clockwork Orange		
W2	T1.2. INSERT Stanley Kubrick IF NOT EXIST	T2.2. INSERT Stanley Kubrick IF NOT EXIST		
W3	T1.3. INSERT MGM IF NOT EXIST	T2.3. INSERT Warner Bros IF NOT EXIST		
W4	T3.1. UPDATE Director SET PlaceOfBirth = "United States of America" WHERE PlaceOfBirth = "USA"			

What's the Result? Duplicate Stanley Kubrick!

Durability

Once a transaction has committed, the effect must never be lost

Function Exception

Program Crash

System Crash: files might not be durable at this level.

"Apocalypse Now"



Database Management Systems (DBMS) 86

A Software System
Provides <u>General Purpose</u>, <u>Efficient</u>, <u>Convenient</u>, & <u>Safe Multi-user</u> Storage of and Access to Massive Amounts of Persistent Data

ACID Properties Guaranteed!

<u>Database Management Systems (DBMS)</u> 87

General Purpose Different Applications: Moviesion, Library,

Safe From System Failures, From Malicious Users

Convenient Simple DDL & DML

Efficient Don't Search All Files in Order to Get a Record

ACID Atomicity, Consistency, Isolation, Durability

<u>Database Management Systems (DBMS)</u> 88

	People (Role)
Is this course about building a DBMS?	No!
	DBMS Implementors
Is this course about maintaining a DBMS?	No!
	DB Administers (DBA)
Is this course about <u>using</u> a DBMS?	<u>Yes</u> .
	DB Designers & User

Company	DBMS
Oracle	Oracle, <u>MySQL</u>
Microsoft	MySQLServer
IBM	DB2
OpenLink Software	Virtuoso Universal Server
Apache Software Foundation	CouchDB

