Data Modeling

After this module you will be able to...

- Describe at least 5 different types in which Big Data may appear in an application problem
- Identify the genre of the data you encounter in an analytics application

What is a Data Model?

After this video you will be able to…

- Distinguish between structured and unstructured data
- Describe four basic data operations, namely, selection, projection, union, join
- Enumerate different types of data constraints like type, value and structural constraints
- Explain why constraints are useful to specify the semantics of data

Data Models Describe Data Characteristics

```
Person {
    firstName: string,
    lastName; string,
    DOB: date
}
```

Structure

Data Models Describe Data Characteristics

All DOB before 1990.

Operations

Data Models Describe Data Characteristics

Today's Date 'minus' DOB must be greater than 18 years.

Constraints

pause

Structure

File 1

(John, Smith, 10-12-1989)
(Liz, Spencer, 09-29-1980)
(Marie, Bishop, 11-07-1992)

File 2

(John, Smith, 10-12-1989, Mechanical, 70000)
(Liz, Spencer, 09-29-1980, Electrical, 65000)
(Marie, Bishop, 11-07-1992, Driver,)
(Steve, Richards, 04-16-1958, 140000)

structure

$A_1, A_2, \ldots, A_k$
$B_1, B_2, \ldots, B_k$
. . .

. . .

$X_1, X_2, \ldots, X_k$

Unstructured Data

কার কোথায়থাকা
উচিতবোঝাযাচ্ছে
না ইদা&#24#2472;ীং! ঘরে
থাকবে কে,
আরবাইরেই বা কে,
বর্ধমানে কার
থাকা দরকার, কার
চলে যাওয়া দরকার
মালদহ থেকে— সব
কেমন গুলিয়ে
যাচ্ছে। সরকার
যাঁকে ঘরের আসনে
বসিয়ে রাখে,
নির্বাচন কমিশন
তাকেই বাইরের
দরজা দেখায়।
অতএবপ্রত্যাঘাত
ে #2478;ুখ্যমন্ত্রীর
মুখ থেকে বেরোয়,
ঘরে ফেরানোর
নিশ্চিত আশ্বাস।

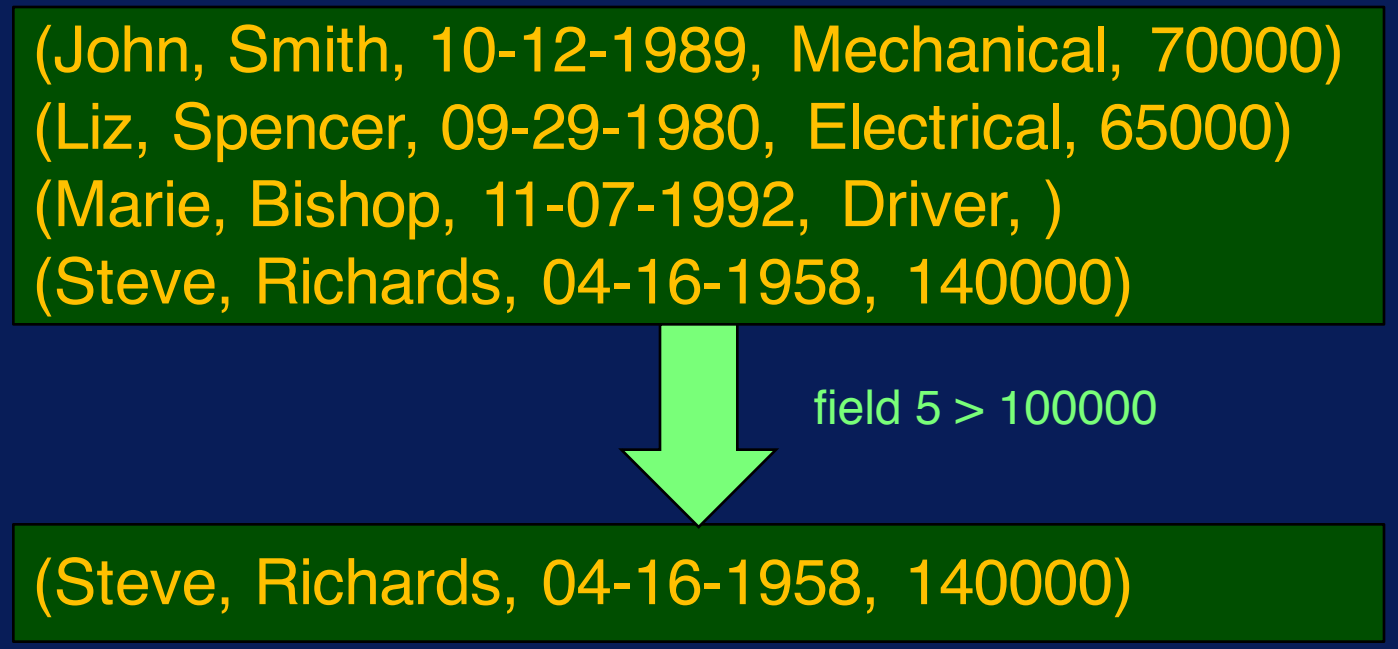
pause

Operations

- **“Subsetting”**
  - Example: Given a collection of data, and a condition
    - Find a subset of data from the collection so that each element in the subset satisfied

Operations

- “Subsetting”



Operations

- **“Substructure extraction”**
  - Given a data collection with some structure, extract from each data item a part of the structure as specified by a condition

Operations

- **“Substructure extraction”**

(John, Smith, 10-12-1989, Mechanical, 70000)
(Liz, Spencer, 09-29-1980, Electrical, 65000)
(Marie, Bishop, 11-07-1992, Driver,)
(Steve, Richards, 04-16-1958, 140000)

field 1, field 2

(John, Smith)
(Liz, Spencer)
(Marie, Bishop)
(Steve, Richards)

Operations

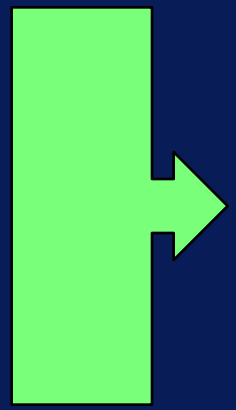
- **“Union”**
  - Given two data collections, create a new one with elements of the two input collections
  - Duplicate elimination

Operations

- “Union”

(John, Smith, 10-12-1989)
(Liz, Spencer, 09-29-1980)
(Marie, Bishop, 11-07-1992)

(Lance, Holt, 04-02-1976)
(Liz, Spencer, 09-29-1980)



(John, Smith, 10-12-1989)
(Liz, Spencer, 09-29-1980)
(Marie, Bishop, 11-07-1992)
(Lance, Holt, 04-02-1976)

Operations

- **“Join”**
  - Given two data collections, create a new one with elements of the two input collections
  - Duplicate elimination

Operations

- “Join”

(12, John, Smith, 10-12-1989)
(14, Liz, Spencer, 09-29-1980)
(18, Marie, Bishop, 11-07-1992)
(20, Sue, Daveson, 03-16-1986)

(12, Mechanical, 70k)
(14, Electrical, 65k)
(18, Driver, 45k)
(23, Student, 30k)

(12, John, Smith, 10-12-1989, Mechanical, 70k)
(14, Liz, Spencer, 09-29-1980, Electrical, 65k)
(18, Marie, Bishop, 11-07-1992, Driver, 45k)

pause

Constraints

- Constraints are logical statements that must hold for data

A movie has only one title

Constraints

- Constraints are logical statements that must hold for data

A movie has only one title

- Different data models have different ways to express constraints

Types of Constraints

- Value constraint
  - Age is never negative
- Uniqueness constraint
  - A movie can have only one title
- Cardinality constraint
  - A person can take between 0 and 3 blood pressure medications at a time

Types of Constraints

- Type constraint

Last Name is alphabetical

Lname:string, not(isNumeric(Lname))

Types of Constraints

- Type constraint

  Last Name is alphabetical

  Lname:string, not(isNumeric(Lname))

- Domain constraint

  Day in (1 … 31)

  Month in (1 … 12) or Month in (‘Jan’, ‘Feb’, … ‘Dec’)

Structural Constraints

- A structural constraint puts restrictions on the structure of the data rather than the data values themselves

Structural Constraints

10	3	23	-3
43	9	86	5
20	-56	0	-16
65	38	36	29

i	j	value
1	1	10
1	2	3
1	3	23
1	4	-3
2	1	43
2	2	9
…	…	…