

# Normal forms

COMS10012 Software Tools

# Aim

How not to make a mess of your data.

- User interface: should be useful for users.
- In memory: objects, pointers – in a way that is useful for programmers.
- On disk / in the database: store each piece of information only once.

# In memory

```
class Lecturer {  
    String name;  
    String username;  
    List<Unit> units;  
}
```

```
{  
    name: "David",  
    username: "csxdb",  
    units: [COMS20700, COMS20002]  
}
```

# Never do this!

name	username	units
"David"	"csxdb"	"COMSM0016,COMS10010"
"Alice"	"csxaw"	""
"John"	"csxjs"	"COMS20002"
...	...	...

# 1NF

## Definition.

A database schema is in first normal form (1NF) if there are no collection-valued attributes (e.g. lists, sets etc.).

# Do it like this – achieving 1NF

<b>name</b>	<b>*username</b>
David	csxdb
Alice	csxaw
John	csxjs
...	...

<b>*unit</b>	<b>director</b>
COMSM0016	csxdb
COMS10010	csxdb
COMS20002	csxjs
...	...

# Try not to do this

student	e-mail	department	faculty	society
bb16801	bilbo@...	modern lang.	arts	Cheese and Wine
bb16801	bilbo@...	modern lang.	arts	Hoverboard
mm16280	mickey@...	civil eng.	engineering	Hoverboard
cl16343	calvin@...	drama	arts	PantoSoc
ht16991	hobbes@...	modern lang.	arts	PantoSoc
...	...	...	...	...

# Candidate key

A set of attributes (columns) is a **candidate key** if it has two properties:

1. Unique – no two rows in the table can ever be the same across all these columns.
2. Minimal – removing any of the columns destroys the uniqueness.

(A unique set is also called a **superkey**.)



# Key attribute

A **key attribute** (column) is one that is part of any candidate key.

A non-key column is one that is not part of any candidate key.

# Example

student	e-mail	department	faculty	society
bb16801	bilbo@...	modern lang.	arts	Cheese and Wine
bb16801	bilbo@...	modern lang.	arts	Hoverboard
mm16280	mickey@...	civil eng.	engineering	Hoverboard
cl16343	calvin@...	drama	arts	PantoSoc
ht16991	hobbes@...	modern lang.	arts	PantoSoc
...	...	...	...	...

# Functional dependency

$$\{A\} \rightarrow B$$

A **functional dependency** exists from a set of columns A to a column B if knowing the A-values uniquely identifies the B-value, for any row.

If B is part of A, the dependency is **trivial**.

# Example

student	e-mail	department	faculty	society / club
S	E	D	F	C
bb16801	bilbo@...	modern lang.	arts	Cheese and Wine
bb16801	bilbo@...	modern lang.	arts	Hoverboard
mm16280	mickey@...	civil eng.	engineering	Hoverboard
cl16343	calvin@...	drama	arts	PantoSoc
ht16991	hobbes@...	modern lang.	arts	PantoSoc
...	...	...	...	...

# 2NF

## Definition.

A schema is in 2NF if it is in 1NF and there are no partial functional dependencies, that is  $\{A\} \rightarrow B$  where

1. A is part of a candidate key (but not whole)
2. B is a non-key attribute.

# Example

student	e-mail	department	faculty	society / club
*S	*E	D	F	*C
bb16801	bilbo@...	modern lang.	arts	Cheese and Wine
bb16801	bilbo@...	modern lang.	arts	Hoverboard
mm16280	mickey@...	civil eng.	engineering	Hoverboard
cl16343	calvin@...	drama	arts	PantoSoc
ht16991	hobbes@...	modern lang.	arts	PantoSoc
...	...	...	...	...

# Achieving 2NF

Split off non-key attributes that depend on part of a CK into a new table, together with "their" part of the candidate key:

Membership		
student	e-mail	society
bb16801	bilbo@...	Cheese and Wine
bb16801	bilbo@...	Hoverboard
mm16280	mickey@...	Hoverboard
...	...	...

Student		
student	department	faculty
bb16801	modern lang.	arts
mm16280	civil eng.	engineering
ht16991	modern lang.	arts
...	...	...



# Another problem

Student		
student	department	faculty
bb16801	modern lang.	arts
mm16280	civil eng.	engineering
ht16991	modern lang.	arts
...	...	...





# 3NF

## Definition.

A schema is in 3NF if it is in 2NF and there are no transitive dependencies  $\{A\} \rightarrow B, \{B\} \rightarrow C$  where both B and C are non-key attributes.

Most of the time this means:  $\{B\} \rightarrow C$  is bad if B is non-key.

# Achieving 3NF

Student	
student	department
bb16801	modern lang.
mm16280	civil eng.
ht16991	modern lang.
...	...

Department	
department	faculty
modern lang.	arts
civil eng.	engineering
...	...



# Heath's Theorem

If you can divide the columns of a table into three sets  $A, B, C$  where  $\{A\} \rightarrow \{B\}$ , then you can split it into two tables  $\{A, B\}$  and  $\{A, C\}$  without gaining or losing information.

This is called **lossless decomposition** and gets you 2NF and 3NF.

# Are we there yet?

student	e-mail	society
bb16801	bilbo@...	Cheese and Wine
bb16801	bilbo@...	Hoverboard
mm16280	mickey@...	Hoverboard
cl16343	calvin@...	PantoSoc
ht16991	hobbes@...	PantoSoc
...	...	...



# BCNF

We've followed all the rules, but the schema still forces us to repeat things.

Boyce and Codd: then we need more rules!



# BCNF (3.5NF)

## Definition.

A schema is in BCNF if it is in 3NF and the following rule applies: **the determinant of every nontrivial FD is a superkey.**

determinant: left-hand side

trivial:  $\{A, \dots\} \rightarrow A$

superkey: unique (minimal not required)



# BCNF notes

- The BCNF condition on its own implies 2NF and 3NF (but not 1NF).
- A schema in 3NF is automatically in BCNF, except if there are overlapping composite candidate keys.
- Not every schema can be losslessly transformed into BCNF.



# Finally ...

Student
*stuid email dept

Member
*stuid *socName

Society
*name

Dept
*name faculty





# Join table attributes

Student
*stuid email

Enrol
*stuid *code grade

Unit
*code name



# Example

city	c_pop	region	r_pop
Bristol	0.44 M	SW	5.2 M
Bath	88.8 K	SW	5.2 M
Manchester	0.52 M	NW	7 M
...			



# Example

house	postcode	city
MVB	BS8 1UB	Bristol
Flat D.01	BS8 4UN	Bristol
Flat D.02	BS8 4UN	Bristol
House of Commons	SW1 0AA	London
30	EC3A 8BF	London
...	...	...

# Example

<b>*postcode</b>	<b>city</b>	<b>region</b>
BS8 1UB	Bristol	SW
BS8 4UN	Bristol	SW
SW1 0AA	London	L
EC3A 8BF	London	L

<b>*postcode</b>	<b>city</b>
BS8 1UB	Bristol
BS8 4UN	Bristol
SW1 0AA	London
EC3A 8BF	London

<b>*city</b>	<b>region</b>
Bristol	SW
London	L



# Elections

from `/vagrant/sampleddata/elections` :

ward	electorate	candidate	party	votes
Avonmouth	9185	Patrick Dorian Hulme	Trade Unionists ...	16
Avonmouth	9185	Ian Humfrey Campion-Smith	Liberal Democrat	108
Avonmouth	9185	Stephen Kenneth James Norman	Independent	135
Avonmouth	9185	Justin Michael Quinnell	Green	149
Bishopston	10303	Owen James Evans	Conservative	511
Filwood	8299	Christopher David Jackson	Labour	1334



# Elections

Ward
*id name electorate

Candidate
*id name party ward votes

Party
*id name



# Table definitions

```
CREATE TABLE Ward (  
    id INTEGER PRIMARY KEY AUTO_INCREMENT,  
    name VARCHAR(100) UNIQUE NOT NULL,  
    electorate INTEGER NOT NULL  
);
```



# Importing the data

```
-- Rawdata: date, ward, electorate,  
-- candidate, party, votes, percent
```

```
LOAD DATA INFILE ... INTO TABLE Rawdata ...;
```

```
INSERT INTO Ward (name, electorate)  
SELECT DISTINCT ward, electorate FROM Rawdata;
```

```
INSERT INTO Party (name)  
SELECT DISTINCT party FROM Rawdata;
```

```
INSERT INTO Candidate (name, party, ward, votes)  
SELECT candidate, Party.id, Ward.id, votes FROM Rawdata  
INNER JOIN Party ON Party.name = Rawdata.party INNER JOIN  
Ward ON Ward.name = Rawdata.ward;
```