# Fall 2024 Database Systems (CSCI-3615 - 10112)

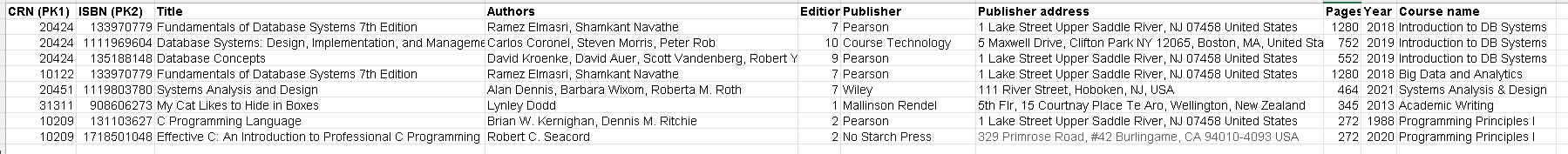
# Assignment 2. Normalization

# Nilufar Babayeva

**Content of the report:**

1. What is Normalization and Boyce-Codd Normal Form? Why do we need it? (p.3)
2. Unnormalized data, its analysis regarding the structure, problems, content of the data, E-R diagram of it. (p.4)
3. Converting the data into 1st Normal Form of BCNF, its analysis and E-R diagram. (p.5-6)
4. Converting the data into 2nd Normal Form of BCNF, its analysis and E-R diagram. (p.6-9)
5. Converting the data into 3rd Normal Form of BCNF, its analysis and E-R diagram. (p.10-13)
6. AI’s approach to normalization of the same data and its analysis. (p.14-15)
7. Comparison of human-written solution to AI generated solution. Its sufficiency and deficiencies. (p.16)
8. **What is Normalization and Boyce-Codd Normal Form? Why do we need Normalization?**

Normalization is organizing the data in a way that redundancy is reduced, and the data integrity is improved. The main goal of normalization is to structure the data so that each piece of data is stored in only one place which prevents issues like data anomalies. It is done by dividing large tables into several small tables and creating relationships among them, ensuring that the database is efficient to work on. Boyce-Codd Normal Form is an advanced level of normalization in database management system. It is about ensuring that the data has minimal redundancy, and the dependency is grouped accordingly. It has six normal forms, all of them having their special rules and regulations. We are required to normalize till the 4th form. We are going to analyze each form during the normalization. It is essential for the data to be normalized, since it makes the data to be less redundant and has less anomalies, more reliable, accurate, and consistent. Therefore, it is easier to maintain and work with normalized data.

1. **Unnormalized data**

This is the unnormalized data that we are given. It is about resources that are used during several courses. There are unique identifiers in two columns which are **CRN (for courses)** and **ISBN (for books)**. Resources are divided into **Title, Authors, Edition, Publisher, Publisher address, Pages, Year**. Courses are divided into CRN and Course names. How do we know that this is unnormalized?

* There are multiple values ​​in the **Authors** column.
* **Publisher address** is divisible.

They prevent rows from being unique which makes it difficult when searching for exactly one data. Also, its **E-R diagram** looks quite messy, and it is difficult to work with.

A screenshot of a computer

Description automatically generated

Because we have only one table, we will have only one E-R diagram for that. The first two attributes are different than others. They are underlined because they are **primary keys** of the entity. The other attributes, on the other hand, are **non-key attributes**. As you can see, this E-R diagram does not tell us much about attributes, how the data is structured etc.

1. **Converting into 1NF and its analysis. E-R diagram.**

1NF Criteria:

A table is in 1NF if:

1. Each column contains **atomic** which means **indivisible values**.
2. Each row should have **single values**, no multiple values.
3. There are **no repeating** rows.

**Applying 1NF to the previous unnormalized data:**

* **Publisher address** is divisible since it contains **street, postal code, state and country**. Therefore, we need to divide it into sections.
* In **Authors** column, there are more than one name. The table should include only one name for each row.
* A screenshot of a computer

  Description automatically generatedAfter applying the first two criteria, we will check whether there are **repeating rows** or not. If yes, we will remove them.

Now this data is in 1NF. We created new rows for each author making them only one author for one row. This met the criteria for having single value for each record. Then, we divided **Publisher address** into **Street address, City, State, Zip code, Country.** This met the criteria for atomic (indivisible values). Now it is obviously seen that the data is **single-valued, has indivisible values for each record.**

****This is the **E-R diagram** for 1NF. **CRN** and **ISBN** remain as primary keys. We divided **Publisher address** into several sections. Then removed it from the E-R diagram of unnormalized and added already created sections of **Publisher address.** The remaining stays the same for now. We will go through the diagram during other NF’s.

1. **Converting into 2NF and its analysis. E-R diagram.**

2NF Criteria:

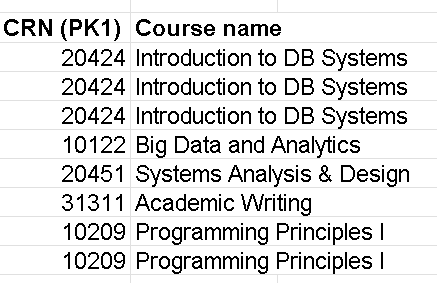
A table is in 2NF if:

1. It is already in 1NF

2. Each non-key attribute is fully functionally dependent on the entire primary key (not just part of it). If it does not depend on primary keys, move them to the separate tables. **Applying 2NF to the 1NF data:**

* As we previously converted the data into 1NF, it **meets the first criteria.**
* However, the current table does not meet second criteria. Given data has 2 primary keys: **CRN** and **ISBN**. According to the second criteria all non-key attributes must be dependent on the primary keys.

In our example, attributes partially depend on **CRN** and **ISBN**. Therefore, we need to group which attributes depend on which primary key. **CRN** is a unique identifier of **Course names**. **ISBN** is all about the book. Therefore, **Title, Authors, Edition, Publisher, Publisher address, Pages and Year** depends only on **ISBN**. Now it is crucial to separate them to the new table.

 This is the first table (**Courses**) of 2NF. As you may already know, **Course name** is only dependent on CRN. Each **Course name** has uniquely one CRN. Therefore, we separated **CRN** and **Course name.** However, this table still has some deficiencies. There are duplicate records for some CRNs. When we remove them we get a table like this:

A screenshot of a computer

Description automatically generatedBelow table is the **final** version of the **2NF-Courses**.

A close-up of a list

Description automatically generatedNow let’s analyze another partial dependent attribute table.

This is the second table (**Book**) of 2NF. **Title, Authors, Edition, Publisher, Street Address, City, State. Zip Code, Country, Pages** and **Year** are dependent only on **ISBN**. They have nothing to do with **CRN**! As we mentioned that there should not be any partial dependency, we moved them to the separate table.

Lastly, we need to show how **CRN** and **ISBN** tables connected to each other. In order to show that, we create another table that will help us to analyze their connection.

A screenshot of a computer

Description automatically generatedThis clearly displays the relationship between **CRN** and **ISBN.** This table will help us during the creation of E-R diagram.

A screenshot of a computer

Description automatically generatedNow the data is completely in 2NF. The final version of 2NF tables:

A close-up of a list

Description automatically generatedA screenshot of a computer

Description automatically generated

A diagram of a computer flowchart

Description automatically generatedLet’s analyze the E-R diagram of the 2NF:

This is the E-R diagram of 2NF. Between entities **Course** and **Book,** I created a relationship called **Resources**. There is a many-to-many relationship between them. It is because a course can use several books as a resource and a book can be taught in several courses. Therefore, two lines without an arrow display many-to-many relationship. Then, between **Book** and **Connection** entities I created **Linked to** relationship showing that a book can have only one **ISBN**. Therefore, I used lines with an arrow showing that this is a one-to-one relationship. Especially, I linked them accordingly – **ISBN** to **ISBN**. Lastly, the same relationship with **Course** and **Connection**. Also, in only our cases it also explains that a course can have only one **CRN**, making it as **Primary key**.

1. **Converting the data into 3rd Normal Form of BCNF, its analysis and E-R diagram.**

**3NF Criteria:**

1. It is already in 2NF

2. It has no transitive dependency which means a non-key attribute cannot depend on another non-key attribute.

This means all non-key attributes should directly depend on the primary key and not on other non-key attributes. Our data is already in 2NF. We only need to analyze for the second criteria.

A screenshot of a computer

Description automatically generatedIf we deeply look at **Book** table, there is no partial dependency. There is only one non-key attribute, and it is directly dependent on the **CRN** (primary key).

A close-up of a list

Description automatically generatedLet’s analyze **Book** table.

From the table it is obvious that Street Address, City, State, Zip Code, Country are dependent on Publisher and Publisher depends directly on ISBN. This is where transitive dependency starts. In order to prevent this, we need to remove them to the separate table. Note: You should move not only the attributes that are dependent, but also the attribute that they are dependent on.

Table: **Publishing**. Now to distinguish **Publisher** and their information, you need to create a primary key. It uniquely identifies each row. We will use this **ID** on the **Book** table to match the records properly.A white background with black text

Description automatically generated

A white sheet with black text

Description automatically generated

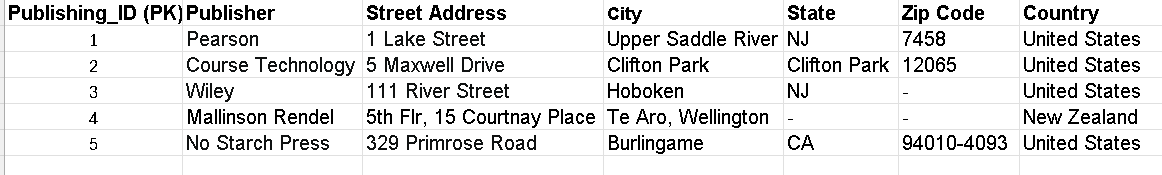
In this updated version of Book table, you can see **Publishing\_ID** used as a foreign key here leading analyst to the Publishing table directly. There is not any transitive dependency on this table anymore.

A screenshot of a computer

Description automatically generatedAs you can see, **Course name** is directly dependent on **CRN**, therefore, we do not need to create new table or change the existing one. This table will remain the same.

A screenshot of a computer

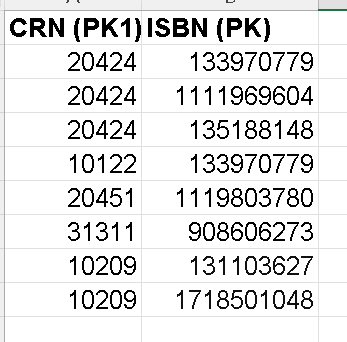
Description automatically generatedAs well as here. We are not going to change anything here since there is **not** any transitive dependency.

Thereby, the **updated version** **of 3NF** is below:

A white sheet with black text

Description automatically generated

A screenshot of a computer

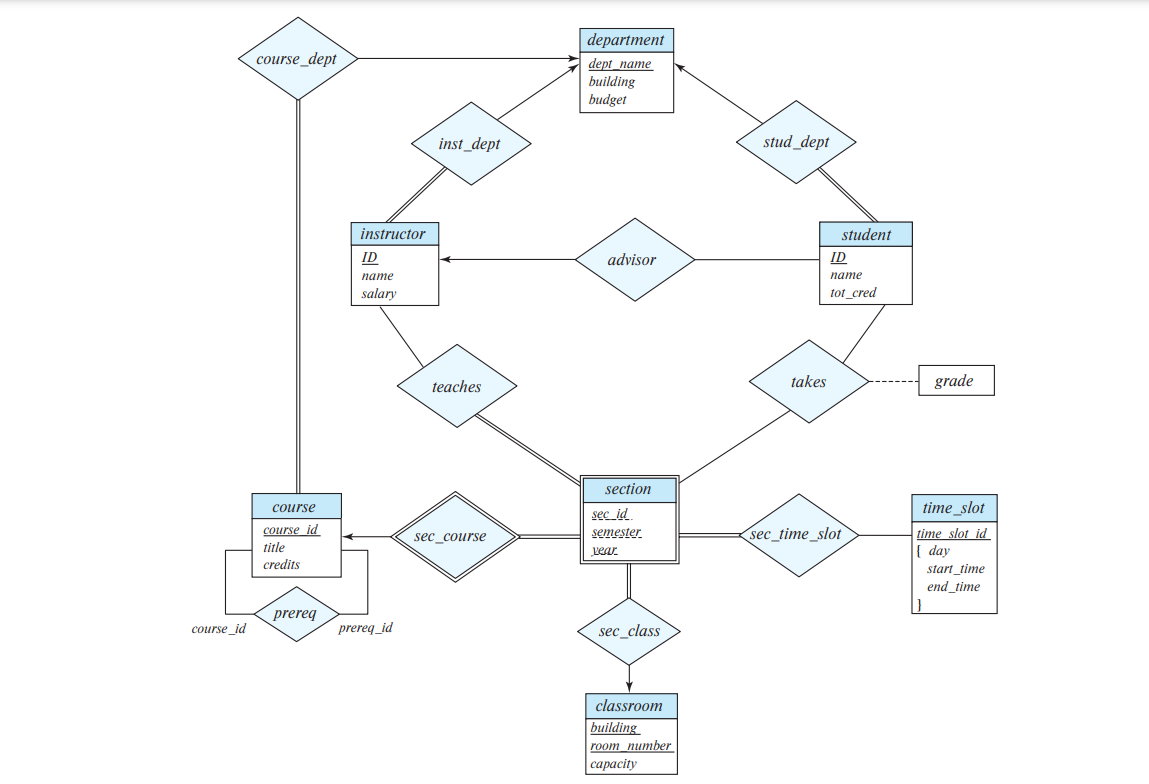
Description automatically generated

A diagram with lines and dots

Description automatically generatedLet’s analyze the E-R diagram of the 3NF:

This is the E-R diagram for 3NF. As we only changed **Book** table in 2NF- moving Publishing information to separate table (in this diagram **Publishing** table), I added a new entity here called **Publishing.** Now **Publishing\_ID** contains information about **Publisher** that we have separated while converting to 3NF. Then using **Publishing\_ID** as a foreign key in **Book** entity, we have created a smooth connection between two entities. The relationship between **Book** and **Publishing\_ID** is one-to-one which means a book can have only one ISBN. The remaining stays the same with the same purposes.

1. **AI’s approach to normalization of the same data and its analysis.**

I sent my excel file that includes all the tables of each normalization form (the tables that I used in this report) and one sample E-R diagram from the book.

A screenshot of a computer

Description automatically generatedThen asked ChatGPT to draw E-R diagrams based on the sample. However, it failed to draw as in the sample. It can normalize data into 1NF, 2NF, 3NF somehow truly, having small deficiency, yet it lacks the necessary capabilities to handle ER diagram drawings.

A screenshot of a computer

Description automatically generatedThis is the first reply of ChatGPT to my previous request. Then I asked him, much more clearly, and also with a very simple explanation, mentioning every simple step several times making sure that instructions are clear. Despite clear explanations, it gave me the same answer again.

A screenshot of a computer

Description automatically generatedThis is my next instructions to ChatGPT about ER diagrams. Though it says I will be provided in the style-shown in my photo provided, the provided is completely different to the one I sent.

1. **Comparison of human-written solution to AI generated solution. Its sufficiency and deficiencies.**

Finally, when we compare human-written and AI generated solutions, we can clearly claim that ChatGPT is not intelligent enough in drawings. Despite the clear instructions, AI solutions are completely wrong for E-R diagram; however, for converting normalization AI somehow understands the task and gives some tables right. Therefore, I would not consider ChatGPT as a way to complete this assignment. At least, you need to know basics of Normalization and E-R diagrams, so that when you ask ChatGPT you can work on its solutions, reduce deficiencies and add some clarifications.