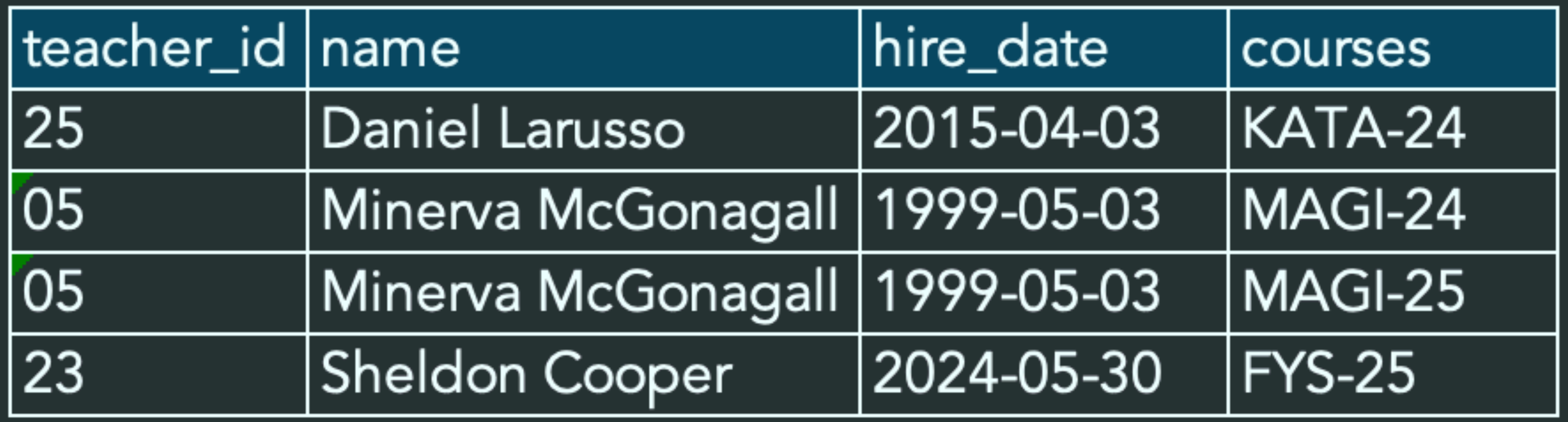
kokchun giang

# breaking down large tables and **normalizing** them to reduce data redundancy and improve data integrity

# a company called **teachy** teaches various cool topics

TeacherCourse



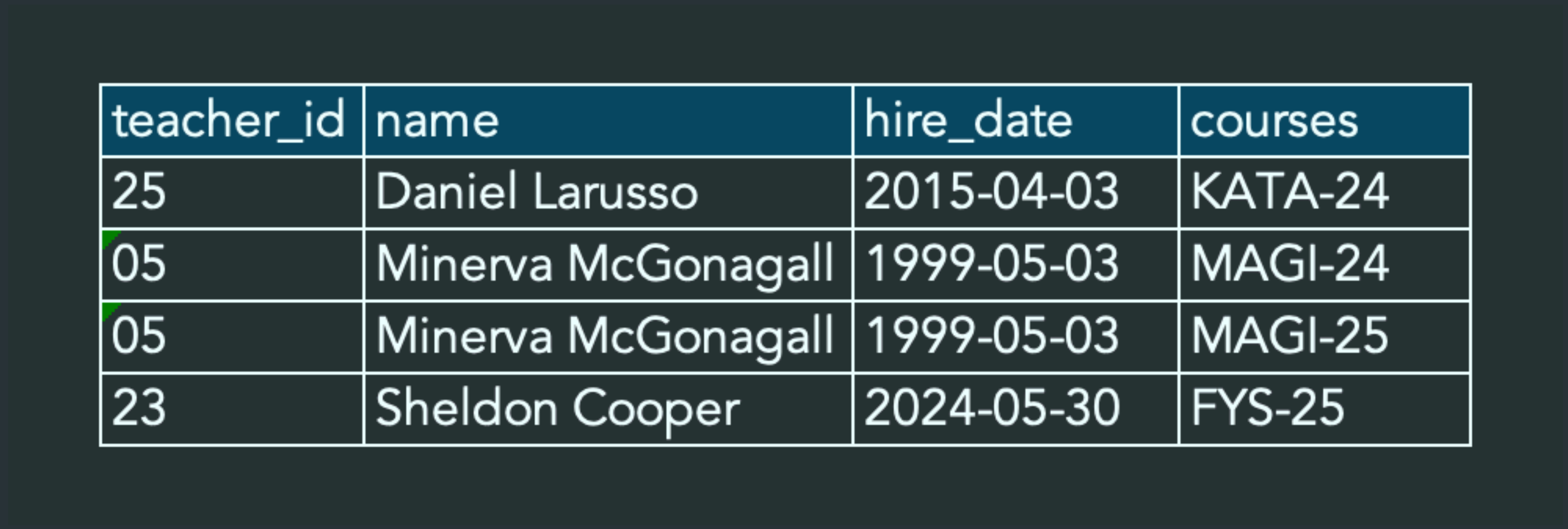
| teacher_id | name | hire_date | courses |
|---|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 | KATA-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-25 |
| 23 | Sheldon Cooper | 2024-05-30 | FYS-25 |

this is a table for teachers and their courses

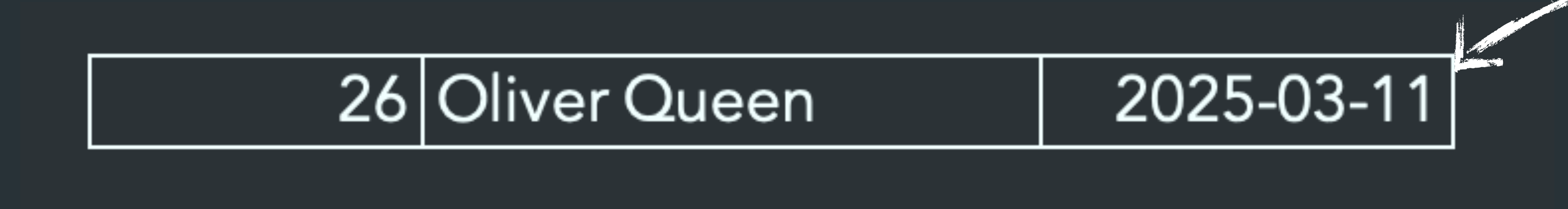
are there any issues with this table?

# an **insertion anomaly** in teachy

can’t insert data due to missing additional data



| teacher_id | name | hire_date | courses |
|---|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 | KATA-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-25 |
| 23 | Sheldon Cooper | 2024-05-30 | FYS-25 |



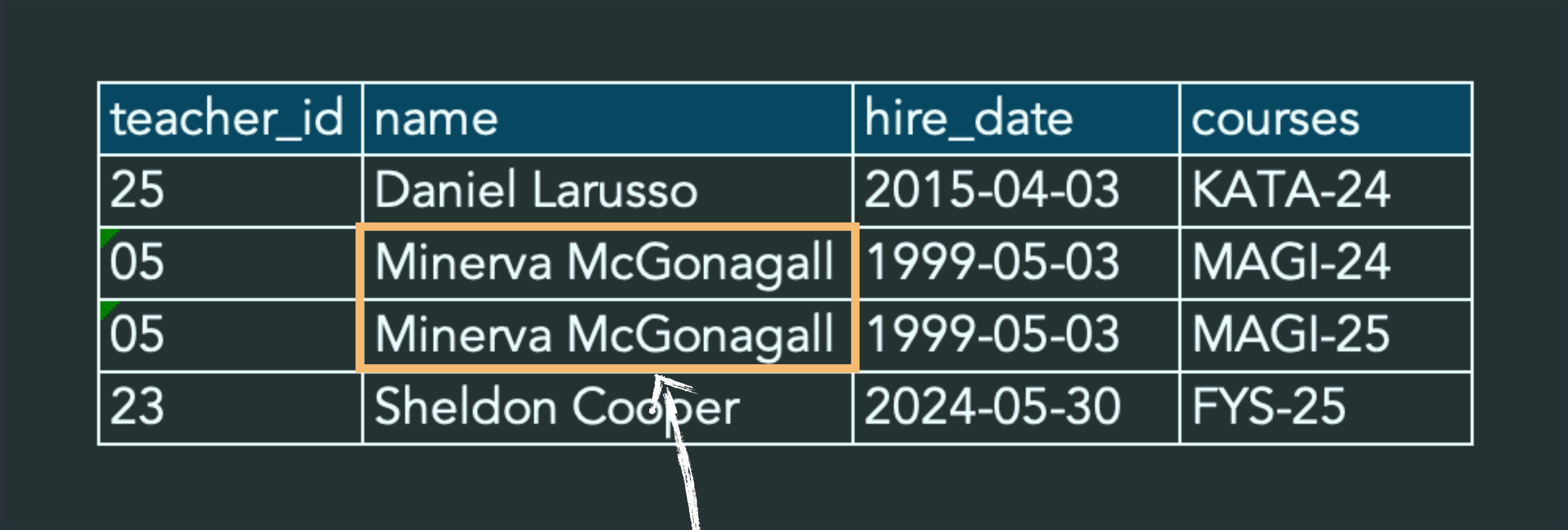
| | | |
|---|---|---|
| 26 | Oliver Queen | 2025-03-11 |

we want to hire Oliver Queen for teaching bow courses in the future but the courses don’t exist yet

this causes an **insertion anomaly** as we must insert a NULL value for course

# an **update anomaly** in teachy

modifying data can result in inconsistencies



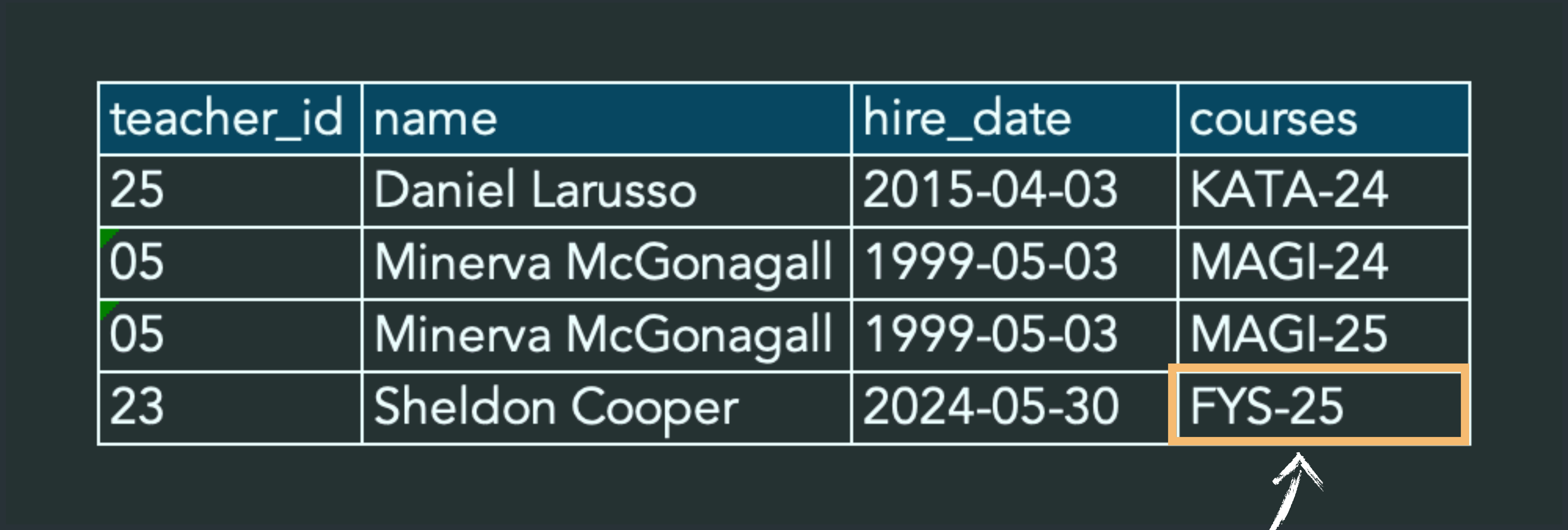
| teacher_id | name | hire_date | courses |
|---|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 | KATA-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-25 |
| 23 | Sheldon Cooper | 2024-05-30 | FYS-25 |

duplicate data, when updating one, we need to update all other entries as well

this can cause an **update anomaly** is if for some reason the update is not completed correctly

# an **deletion anomaly** in teachy

deletion of a record leads to unintentional loss of data



| teacher_id | name | hire_date | courses |
|---|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 | KATA-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-25 |
| 23 | Sheldon Cooper | 2024-05-30 | FYS-25 |

FYS-25 has ended and
needs will be deleted

unfortunately we have also deleted our beloved teacher, which is a **deletion anomaly**

# **normalization** will fix these anomalies

organizing data to different normal forms

- first normal form (1NF)
- second normal form (2NF)
- third normal form (3NF)
- boyce-codd normal form (BCNF)
- fourth normal form (4NF)
- fitfh normal form (5NF)
- sixth normal form (6NF)

↓ safer against redundant data

in most cases 3NF or BCNF is already normalized enough and a good database design

# **normalization** will fix these anomalies

### first normal form

- row order doesn't matter
- primary key in each table
- no repeating groups
- uniform column data

### second normal form

- 1NF
- non-prime attributes must be functionally dependent on entire primary key and not just part of it

### third normal form

- 2NF
- non-prime attributes depends on the key, the whole key and nothing but the key

we won't go further than 3NF

# normalization will fix these anomalies

## first normal form

- row order doesn’t matter
- primary key in each table
- no repeating groups
- uniform column data

## second normal form

- 1NF
- non-prime attributes must be functionally dependent on entire primary key and not just part of it

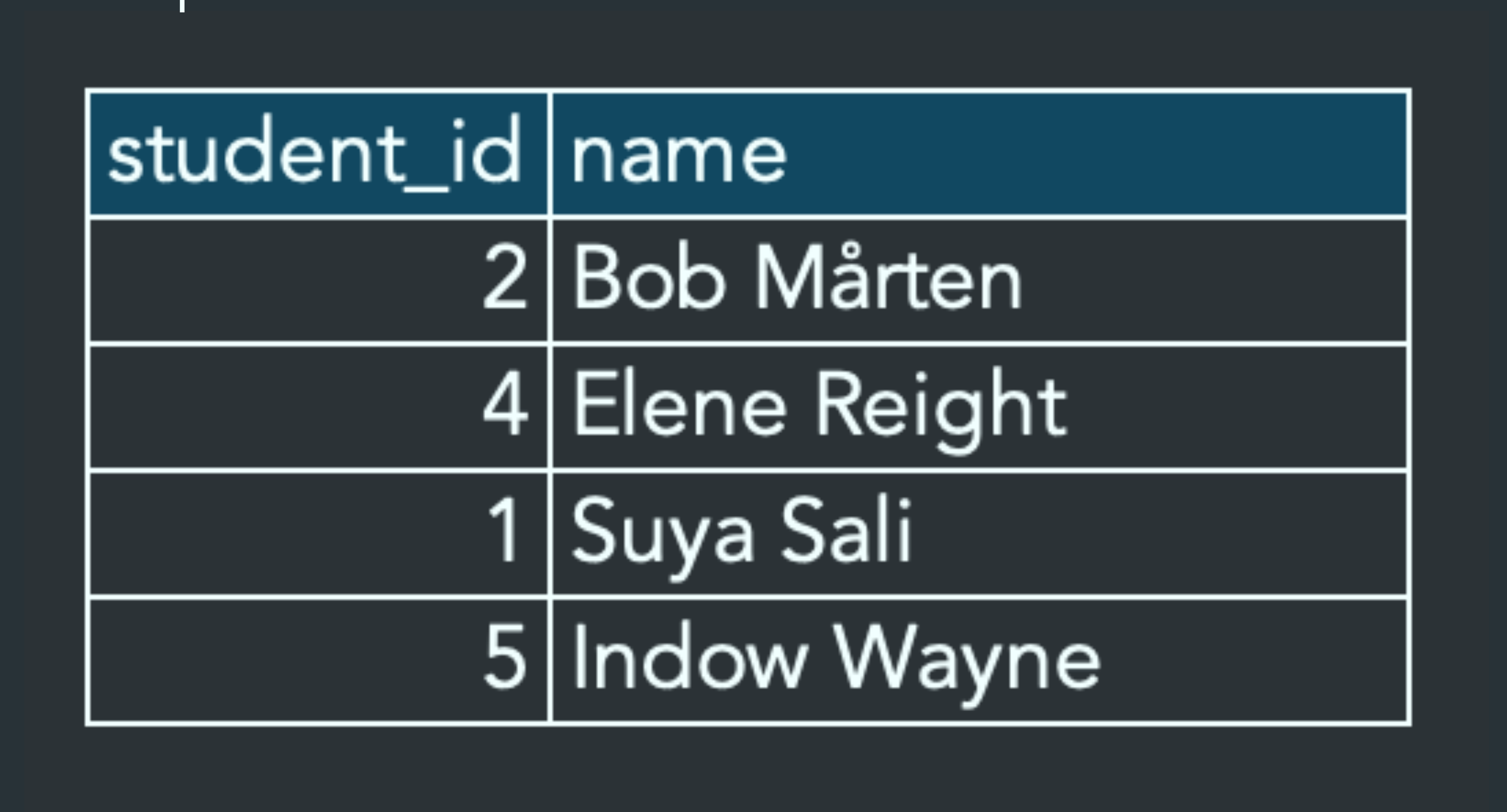
## third normal form

- 2NF
- non-prime attributes depends on the key, the whole key and nothing but the key

we won’t go further than 3NF

# violating **first normal form (1NF)** by conveying row information

TopStudentResults



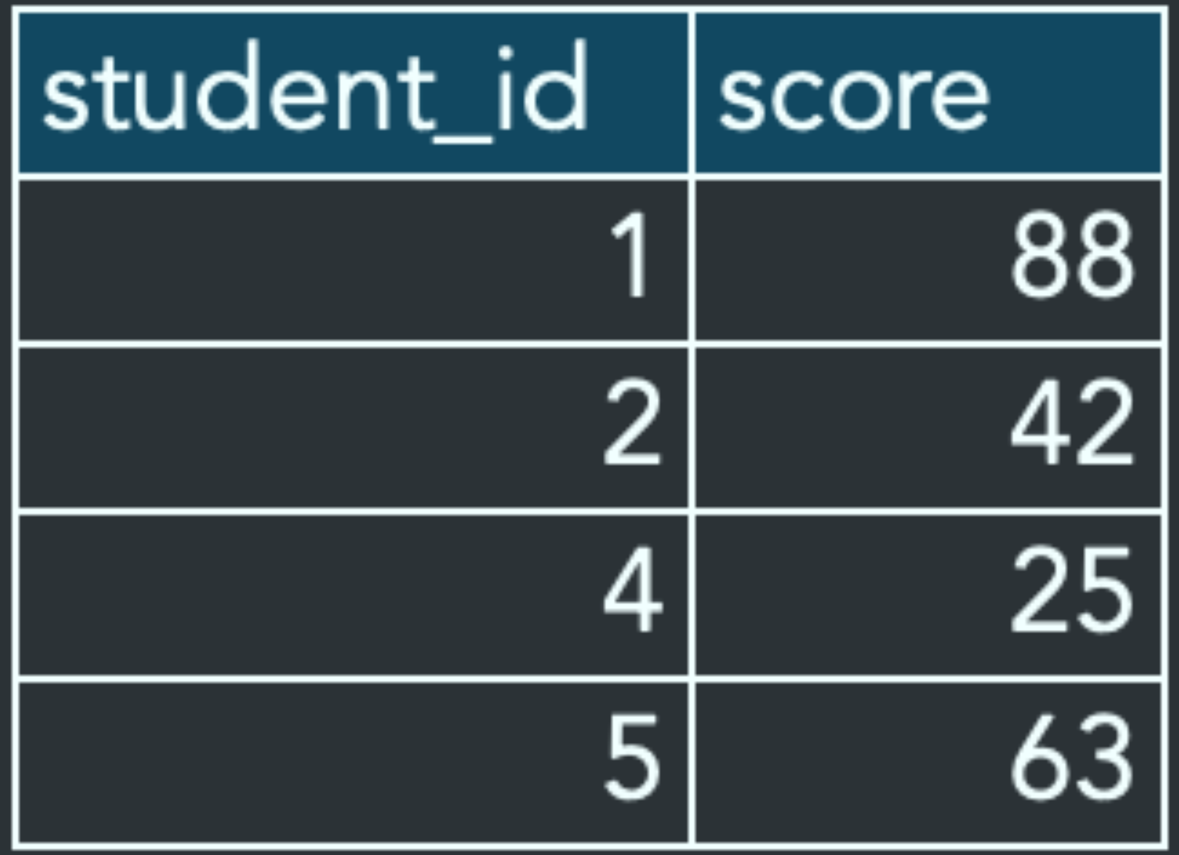
| student_id | name |
|---:|---|
| 2 | Bob Mårten |
| 4 | Elene Reight |
| 1 | Suya Sali |
| 5 | Indow Wayne |

this table give information on students ranking based on their ordering

violates 1NF as it conveys information through row ordering

# we create a new table to satisfy **1NF**

StudentScore table



| student_id | score |
|---|---|
| 1 | 88 |
| 2 | 42 |
| 4 | 25 |
| 5 | 63 |

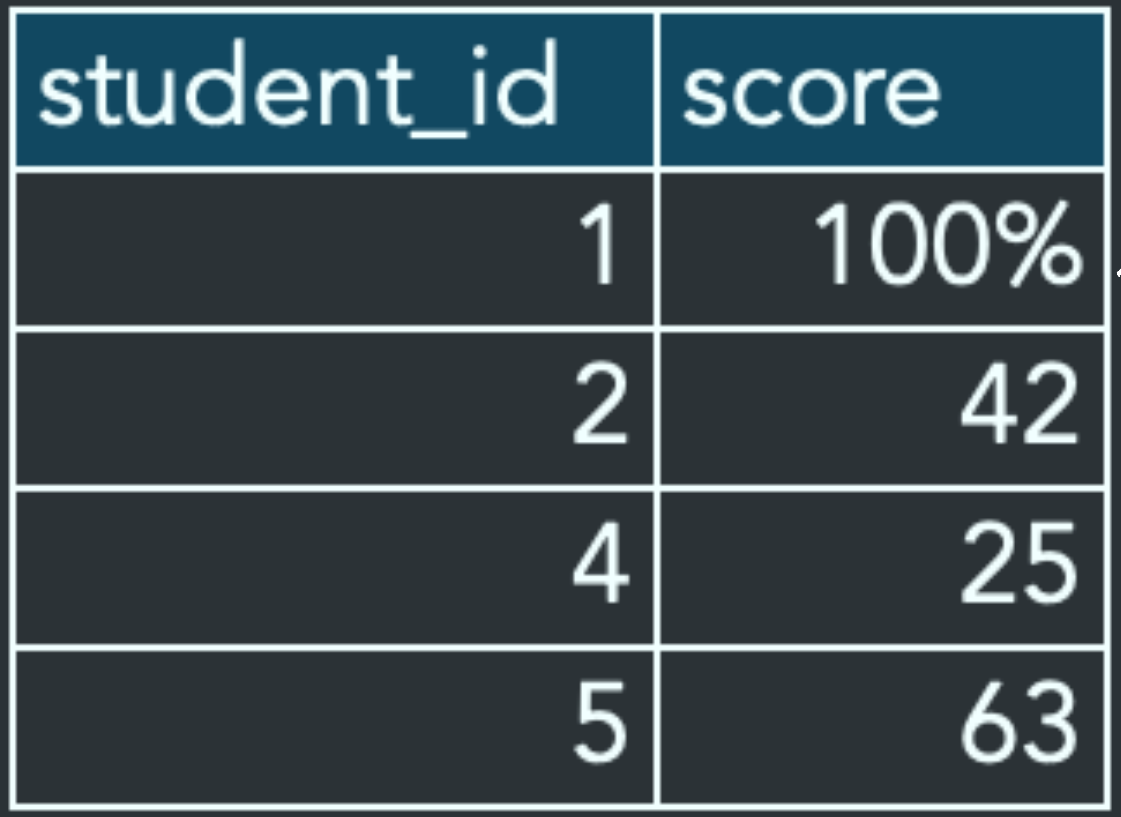
instead we could have a separate score table together with the student table

to find out which person scored how much, we can join the two tables

also if we want to sort them, we can do ORDER BY

# violating **first normal form (1NF)** by mixing data types

StudentScore table



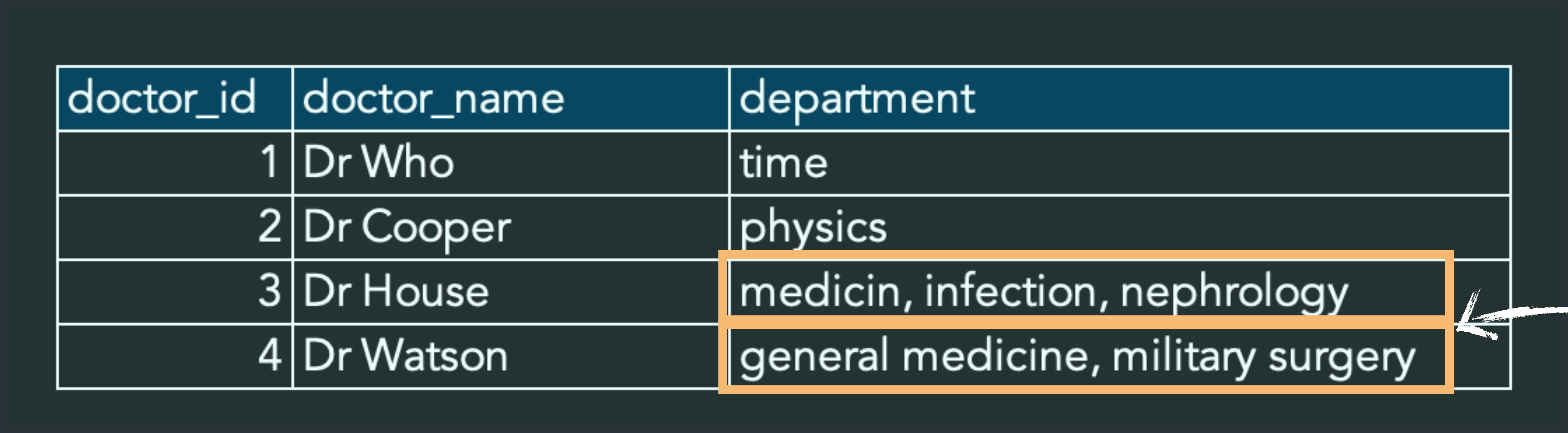
| student_id | score |
|---:|---:|
| 1 | 100% |
| 2 | 42 |
| 4 | 25 |
| 5 | 63 |

string and
other are
integer

a data column must have uniform data, that is we can’t mix data types

# violating **1NF** by repeating groups

Score



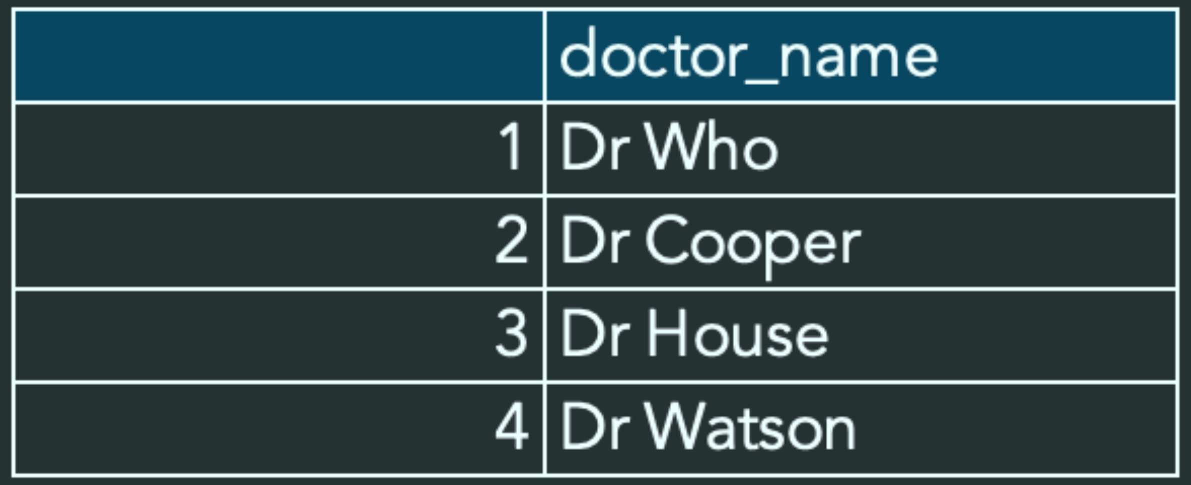
| doctor_id | doctor_name | department |
|---|---|---|
| 1 | Dr Who | time |
| 2 | Dr Cooper | physics |
| 3 | Dr House | medicin, infection, nephrology |
| 4 | Dr Watson | general medicine, military surgery |

repeating groups

repeating groups are attributes with more than one value in a row and it is violating 1NF

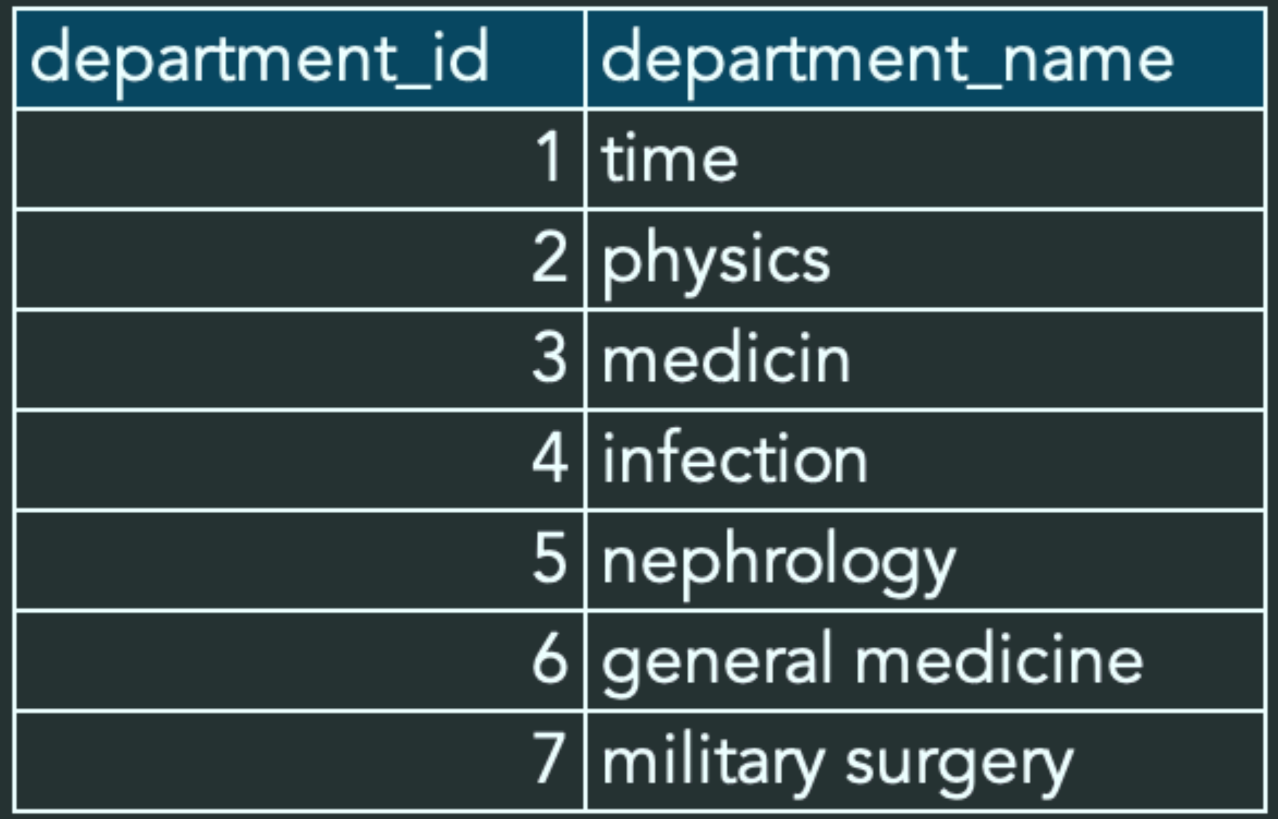
# divide into several tables to avoid **repeating groups**

Doctor table



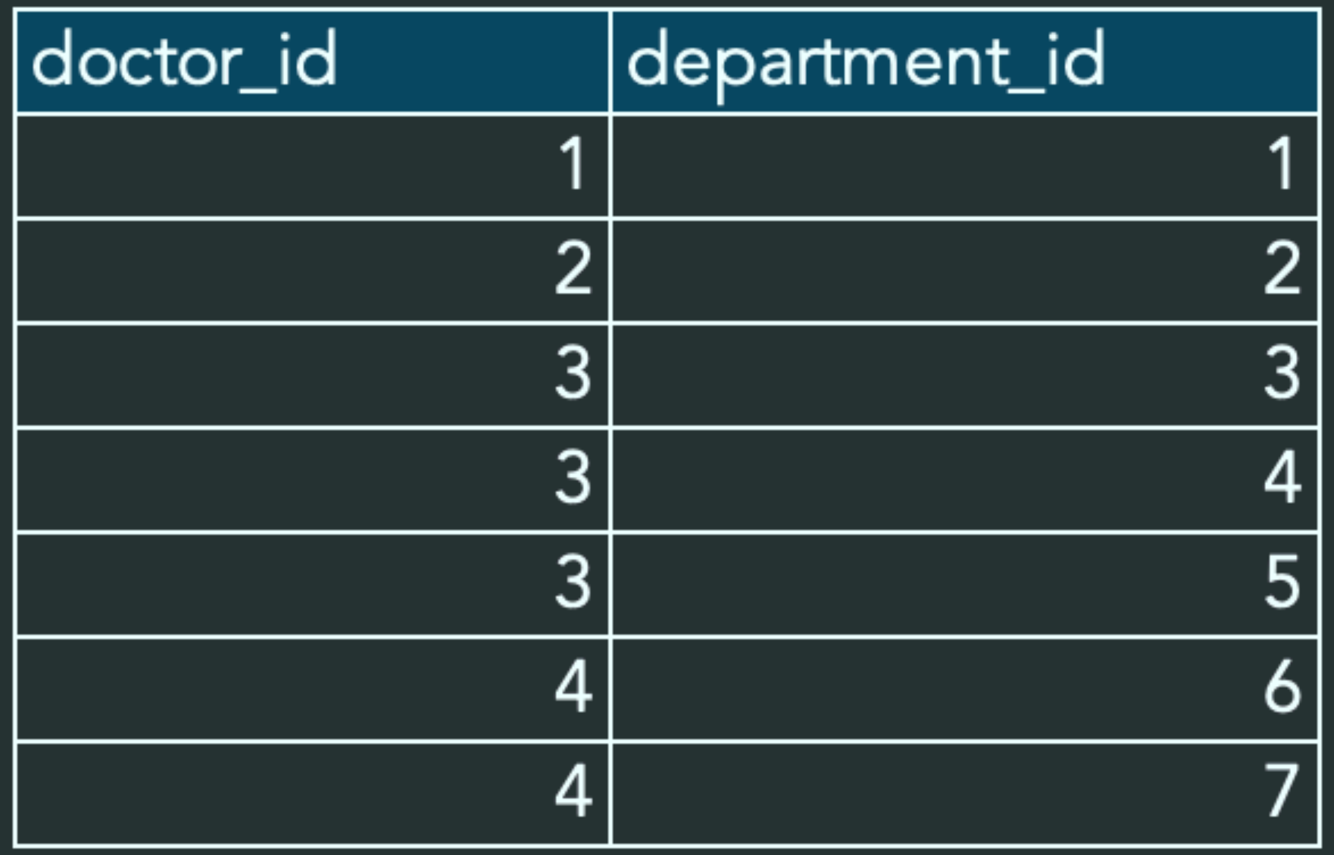
| | doctor_name |
|---:|---|
| 1 | Dr Who |
| 2 | Dr Cooper |
| 3 | Dr House |
| 4 | Dr Watson |

Department table



| department_id | department_name |
|---:|---|
| 1 | time |
| 2 | physics |
| 3 | medicin |
| 4 | infection |
| 5 | nephrology |
| 6 | general medicine |
| 7 | military surgery |

DoctorDepartment table



| doctor_id | department_id |
|---:|---:|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 3 | 4 |
| 3 | 5 |
| 4 | 6 |
| 4 | 7 |

# different **normal forms**

**first normal form**

- row order doesn’t matter
- primary key in each table
- no repeating groups
- uniform column data

**second normal form**

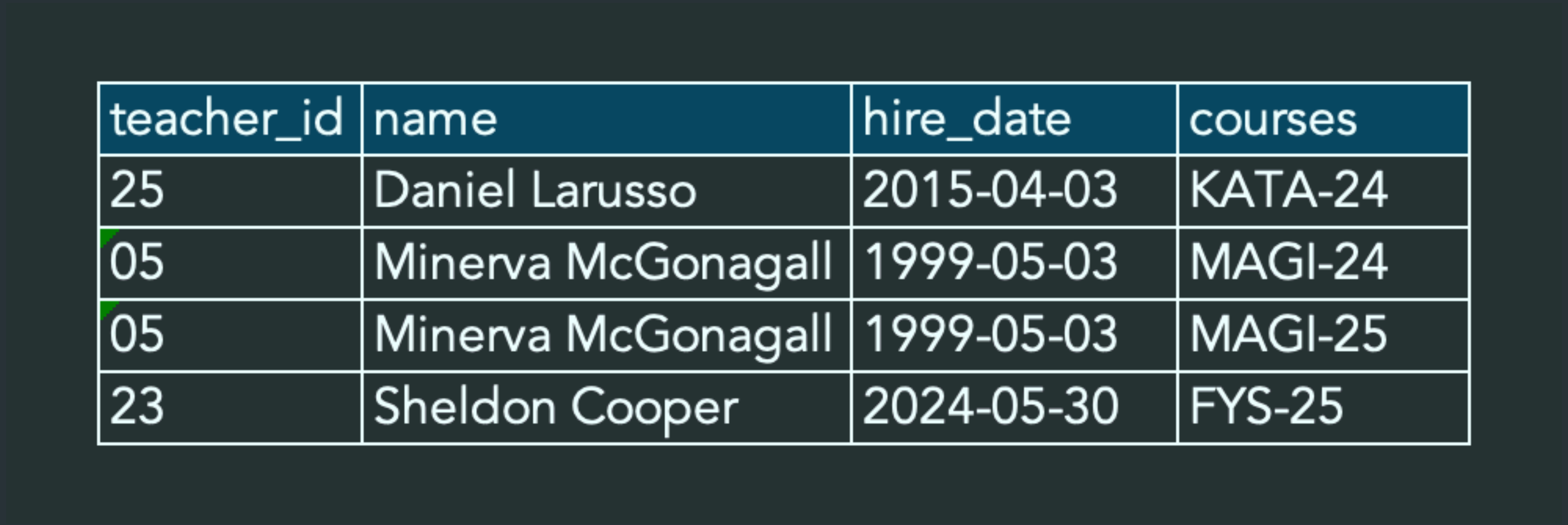
- 1NF
- non-prime attributes must be functionally dependent on entire primary key and not just part of it

**third normal form**

- 2NF
- non-prime attributes depends on the key, the whole key and nothing but the key

we won’t go further than 3NF

# this table conforms with **1NF but not 2NF**



| teacher_id | name | hire_date | courses |
|---|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 | KATA-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-24 |
| 05 | Minerva McGonagall | 1999-05-03 | MAGI-25 |
| 23 | Sheldon Cooper | 2024-05-30 | FYS-25 |

it conforms with 1NF if choosing either teacher_id together with courses as **composite primary key** or by creating a surrogate key teacher_courses_id

non-prime attributes (name and hire_date) depends partially on the primary key and not the entire primary key

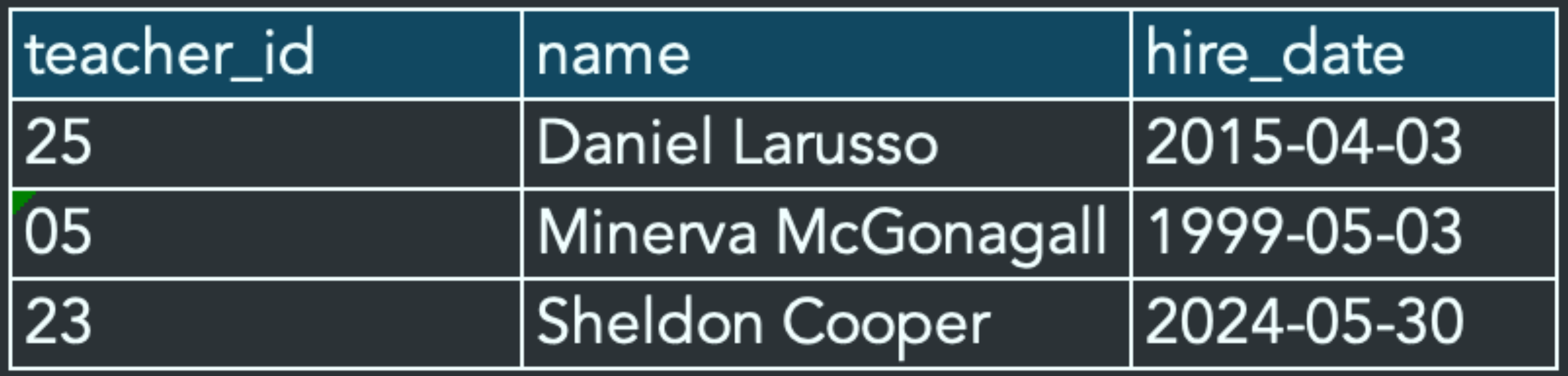
teacher_id → name

teacher_id → hire_date

teacher_id **functionally determines** name

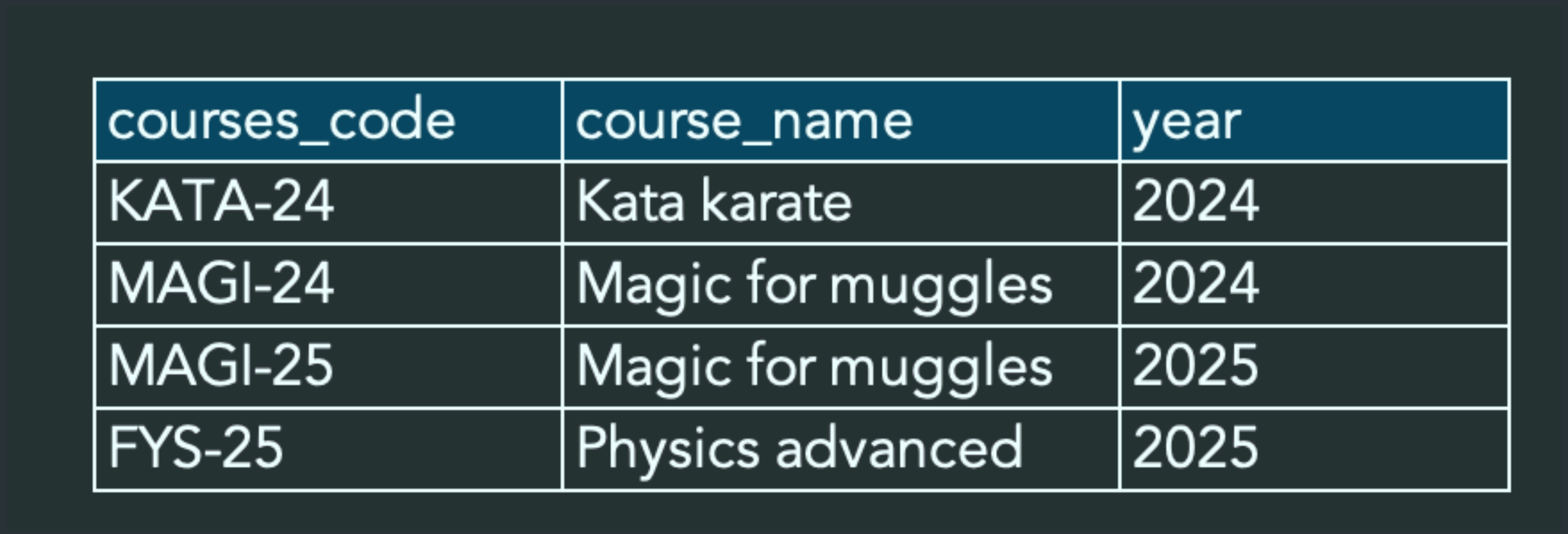
so for each value in teacher_id there is a corresponding value in name

# first attempt to normalize to **2NF**



| teacher_id | name | hire_date |
|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 |
| 05 | Minerva McGonagall | 1999-05-03 |
| 23 | Sheldon Cooper | 2024-05-30 |

is in 2NF as
teacher_id → name
teacher_id → hire_date

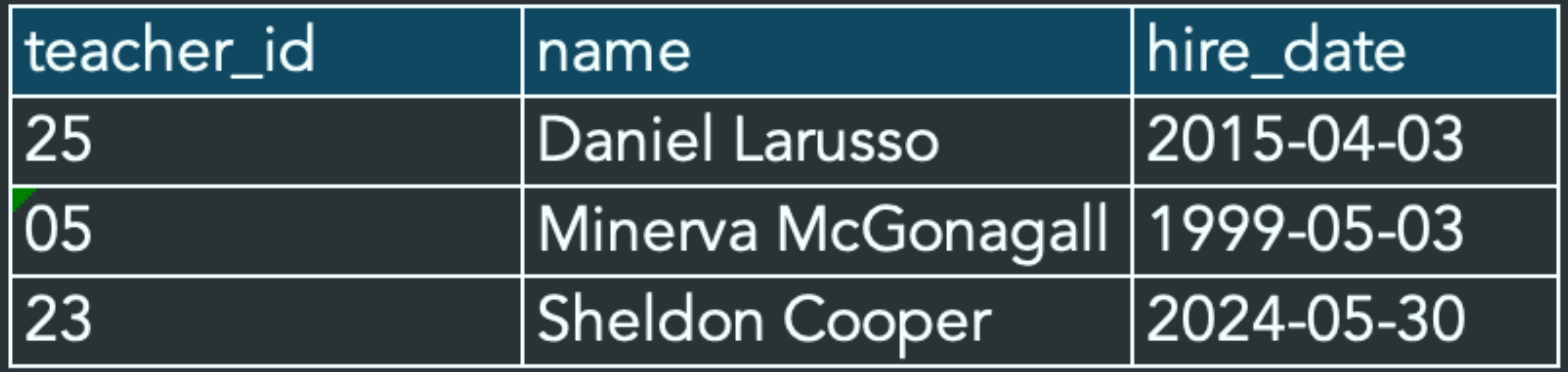


| courses_code | course_name | year |
|---|---|---|
| KATA-24 | Kata karate | 2024 |
| MAGI-24 | Magic for muggles | 2024 |
| MAGI-25 | Magic for muggles | 2025 |
| FYS-25 | Physics advanced | 2025 |

not in 2NF as
course_code if chosen as primary key has the year encoded in it so course_name and year are **partially dependent** on course_code and not the entire primary key

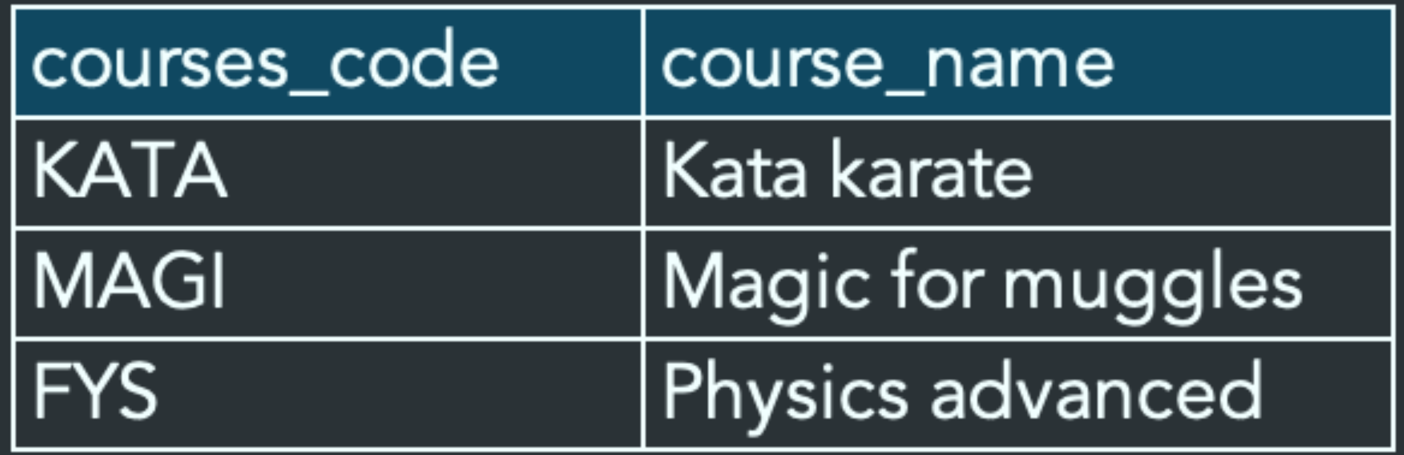
# normalize it to **2NF** by removing partial dependencies

Teacher



| teacher_id | name | hire_date |
|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 |
| 05 | Minerva McGonagall | 1999-05-03 |
| 23 | Sheldon Cooper | 2024-05-30 |

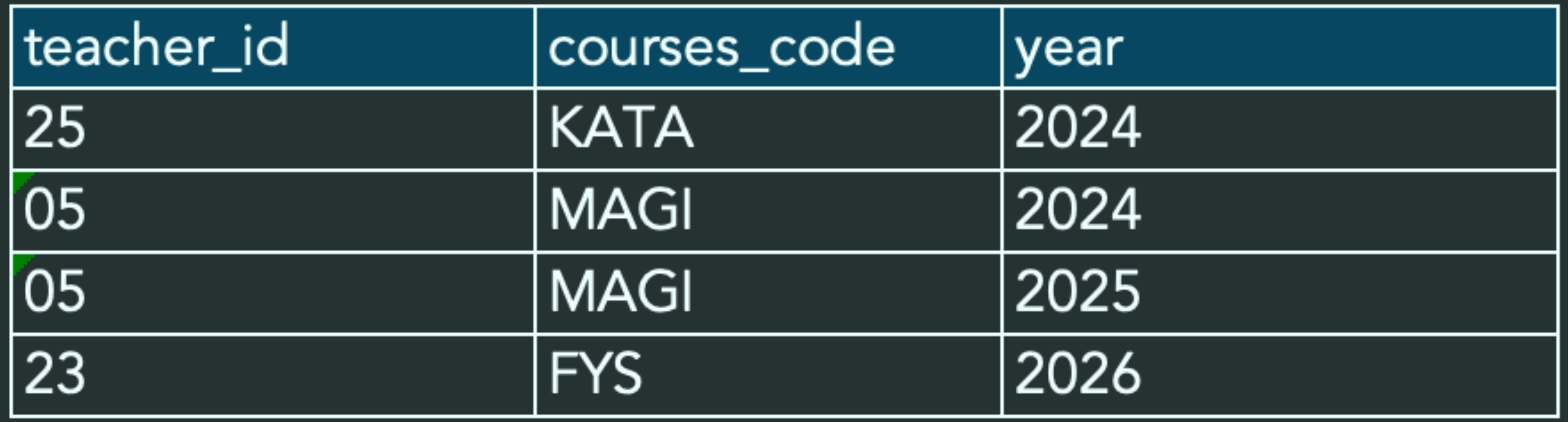
Courses



| courses_code | course_name |
|---|---|
| KATA | Kata karate |
| MAGI | Magic for muggles |
| FYS | Physics advanced |

course_code → course_name

TeacherCourseOfferings



| teacher_id | courses_code | year |
|---|---|---|
| 25 | KATA | 2024 |
| 05 | MAGI | 2024 |
| 05 | MAGI | 2025 |
| 23 | FYS | 2026 |

no non-prime attributes exist if all three attributes chosen as primary key, therefore no partial dependencies, and it conforms to 2NF trivially

# **normalization** will fix these anomalies

### first normal form

- row order doesn’t matter
- primary key in each table
- no repeating groups
- uniform column data

### second normal form

- 1NF
- non-prime attributes must be functionally dependent on entire primary key and not just part of it

### third normal form

- 2NF
- non-prime attributes depends on the key, the whole key and nothing but the key

therefore can’t have **transitive dependencies**

we won’t go further than 3NF

# this design violates **3NF** with transitive dependency

here we'll use relational schema
notation

TeacherCourse(teacher_course_id, teacher_id, course, course_name)

this conforms to 2NF as
teacher_course_id → teacher_id, course, course_name

however
course → course_name
$\Rightarrow \boxed{\text{teacher\_course\_id} \rightarrow \text{course} \rightarrow \text{course\_name}}$

**transitive dependency**
course_name dependent on
non-prime course which is
dependent on prime
teacher_course_id

# normalize it to **3NF** by removing transitive dependency

Teacher(teacher_id, name, hire_date)

this conforms to 2NF and 3NF as
teacher_id → name, hire_date

TeacherCourse(teacher_course_id, teacher_id, course_id)

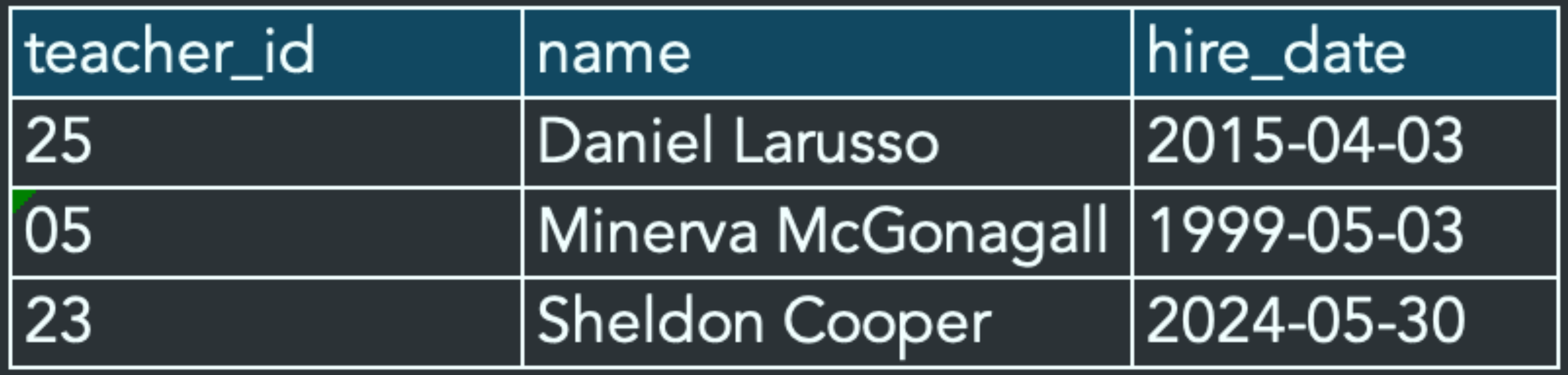
this conforms to 2NF and 3NF as there are
no non-prime attributes

Course(course_id, course_name, credits)

no transitive dependency, 3NF
course_id → course_name, credits

# the previous teacher data model also conforms to **3NF**

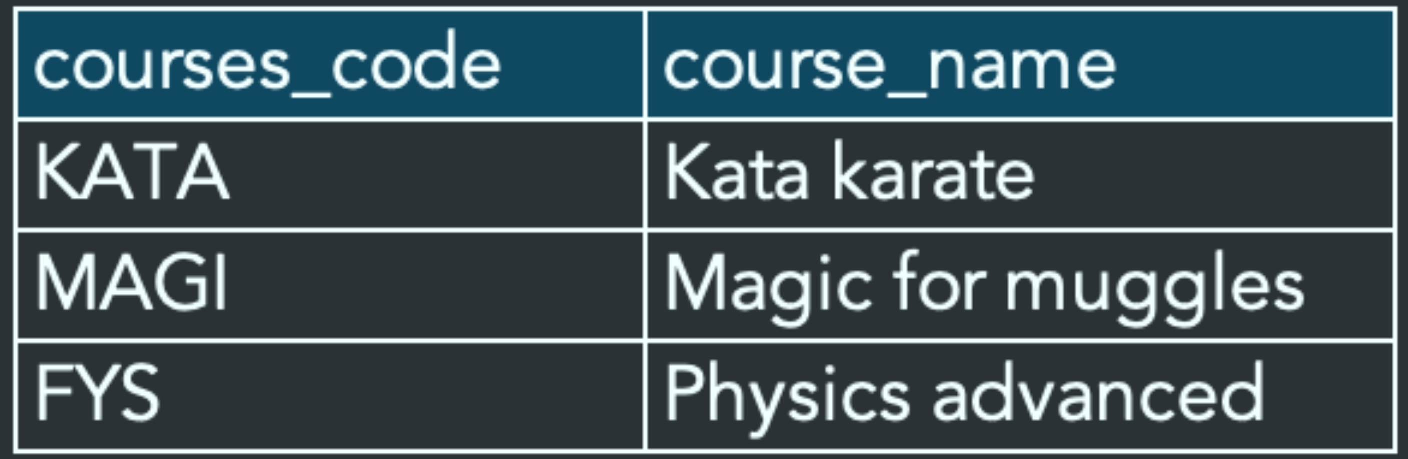
Teacher



| teacher_id | name | hire_date |
|---|---|---|
| 25 | Daniel Larusso | 2015-04-03 |
| 05 | Minerva McGonagall | 1999-05-03 |
| 23 | Sheldon Cooper | 2024-05-30 |

non-key attributes depend only on the key, name and hire_date depend only  on teacher_id

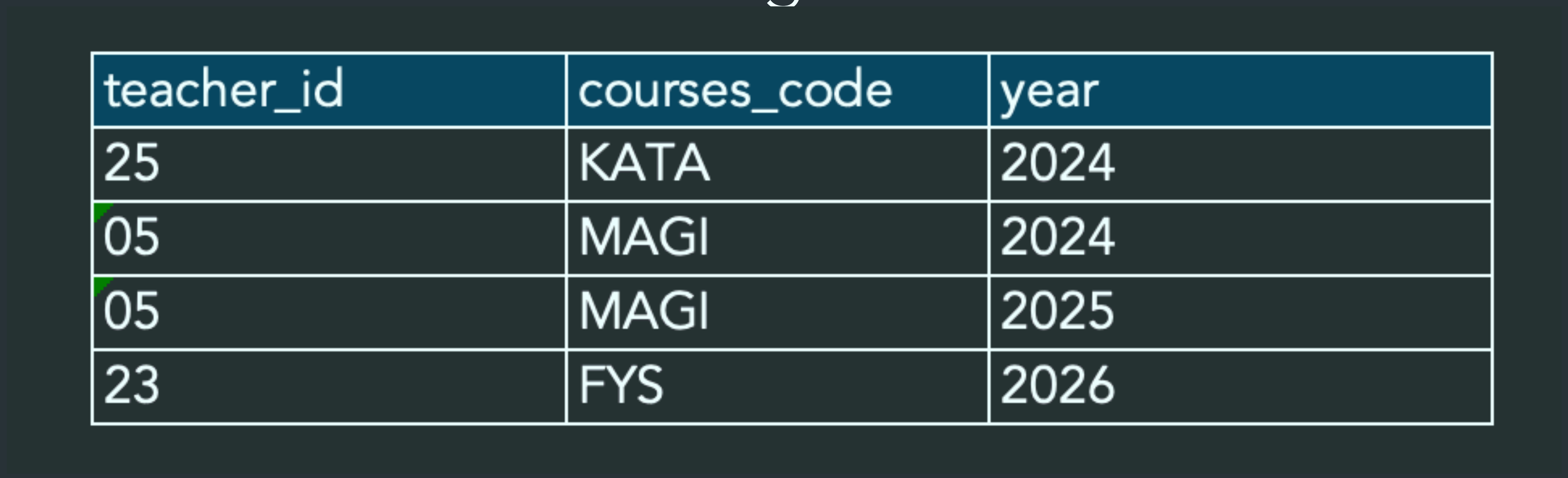
Courses



| courses_code | course_name |
|---|---|
| KATA | Kata karate |
| MAGI | Magic for muggles |
| FYS | Physics advanced |

course_name depend only on course_code

TeacherCourseOfferings



| teacher_id | courses_code | year |
|---|---|---|
| 25 | KATA | 2024 |
| 05 | MAGI | 2024 |
| 05 | MAGI | 2025 |
| 23 | FYS | 2026 |

all attributes part of composite primary key (teacher_id, course_code, year)