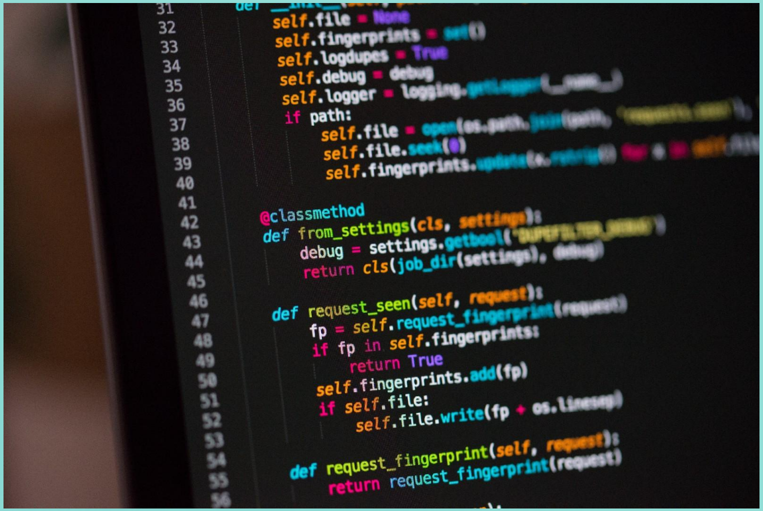
Leveraging data technologies + SQL to bring bigger data into the classroom

Nicholas J. Horton, Amherst College

May 20, 2025, nhorton@amherst.edu

Image source: SQL Humor



```
    def __init__(
        self.file = None
        self.fingerprints = set()
        self.logdupes = True
        self.debug = debug
        self.logger = logging.getLogger(__name__)
        if path:
            self.file = open(os.path.join(path, 'requests.seen'), 'a+')
            self.file.seek(0)
            self.fingerprints.update(x.rstrip() for x in self.file)

    @classmethod
    def from_settings(cls, settings):
        debug = settings.getbool('DUPEFILTER_DEBUG')
        return cls(job_dir(settings), debug)

    def request_seen(self, request):
        fp = self.request_fingerprint(request)
        if fp in self.fingerprints:
            return True
        self.fingerprints.add(fp)
        if self.file:
            self.file.write(fp + os.linesep)

    def request_fingerprint(self, request):
        return request_fingerprint(request)
```

Image source: Wikicommons

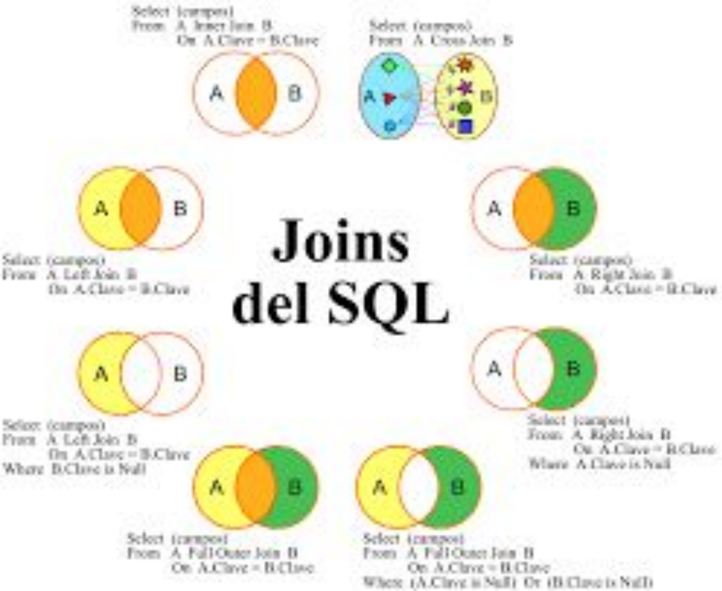


Image source: Wikimedia Commons

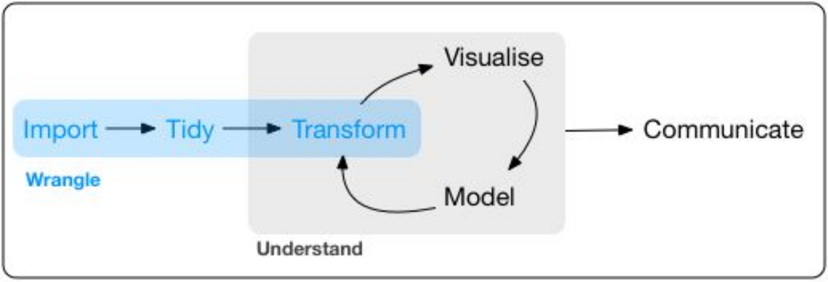


Image source: Hadley Wickham and Garrett Grolemund

Links at https://nicholasjhorton.github.io/K12-Data-Tools/icerm.html

DATA SCIENCE FOR UNDERGRADUATES

Opportunities and Options

consensus report published in 2018
free download from
https://nas.edu/envisioningds

**Study funded by the
National Science Foundation**

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

Data management concepts

Key **data management and curation** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Data provenance;
- Data preparation, especially data cleansing and data transformation;
- Data management (of a variety of data types);
- Record retention policies;
- Data subject privacy;
- Missing and conflicting data; and
- **Modern databases**.

SQL + Databases

Common response seen from reflections from statistics graduates (program review):

“You should have taught us SQL: I needed it for my job”.

SQL + Databases

Common response seen from reflections from statistics graduates (program review):

“You should have taught us SQL: I needed it for my job”.

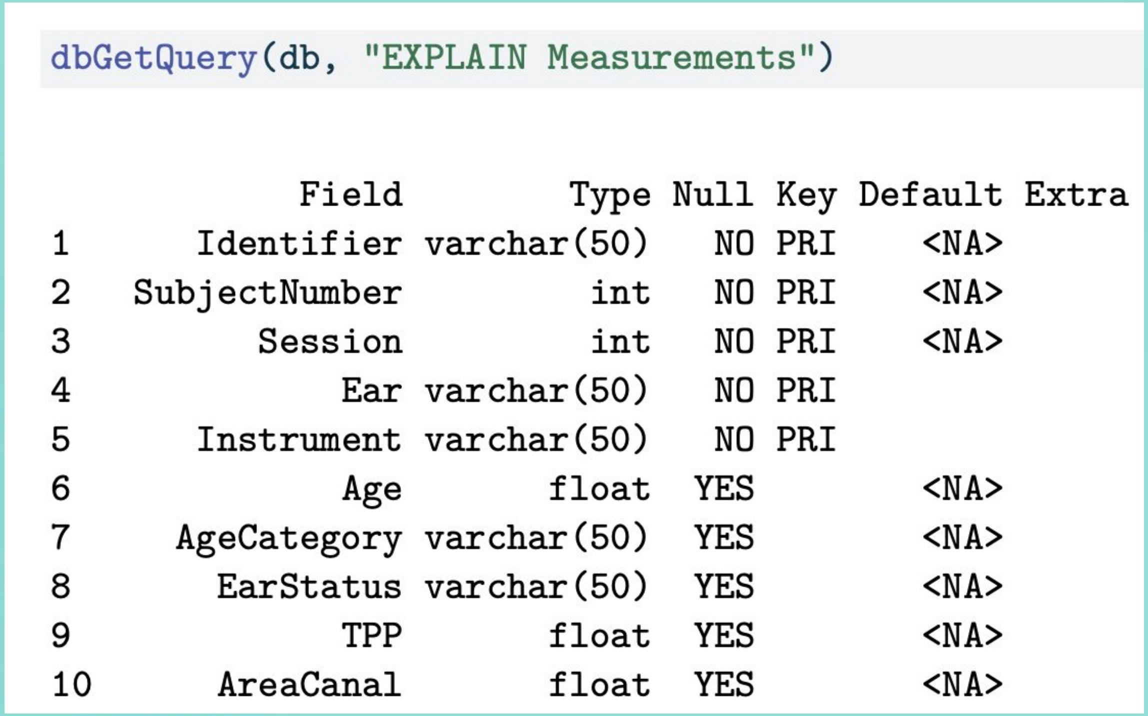
“But the good news is that I taught it to myself over a weekend.”

SQL + Databases

- Structured Query Language (SQL) implements Codd's relational model
- Since 1970 has provided a framework for relational databases, now lingua franca for large data stores
- Relatively easy to learn to access
- Lets the highly optimized database do much of the work
- "Ensuring that Mathematics is Relevant in a World of Data Science" (Hardin and Horton, *Notices of the American Mathematical Society*, 2017)

SQL + Databases

- Want to explore? See the sample Quarto file (+ associated pdf) https://nicholasjhorton.github.io/K12-Data-Tools/icerm.html



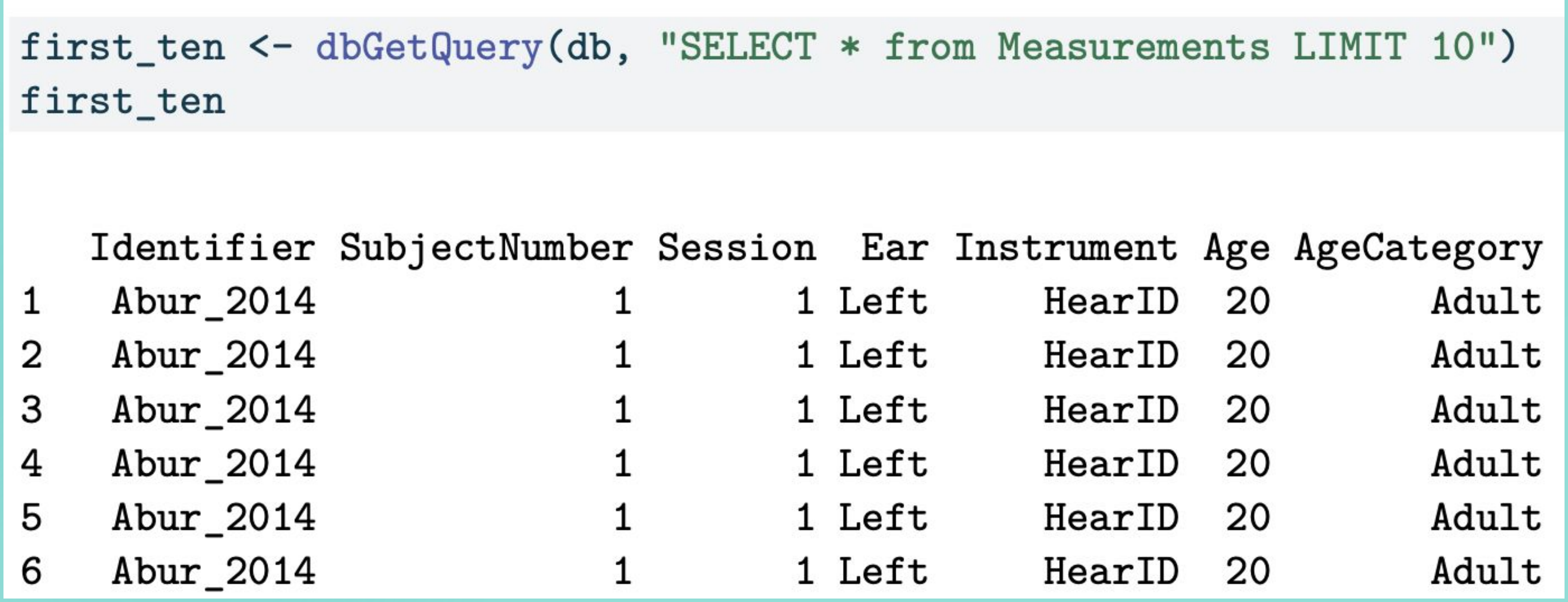
```
dbGetQuery(db, "EXPLAIN Measurements")
```

```
           Field        Type Null Key Default Extra
1     Identifier varchar(50)   NO PRI    <NA>
2  SubjectNumber         int   NO PRI    <NA>
3        Session         int   NO PRI    <NA>
4            Ear varchar(50)   NO PRI
5     Instrument varchar(50)   NO PRI
6            Age       float  YES        <NA>
7    AgeCategory varchar(50)  YES        <NA>
8      EarStatus varchar(50)  YES        <NA>
9            TPP       float  YES        <NA>
10     AreaCanal       float  YES        <NA>
```

- see more: "Modern Data Science with R (2e+)" (Baumer, Kaplan, and Horton, 2024, https://mdsr-book.github.io/mdsr3e/)

SQL + Databases

- Want to explore? See the sample Quarto file (+ associated pdf) https://nicholasjhorton.github.io/K12-Data-Tools/icerm.html



```
first_ten <- dbGetQuery(db, "SELECT * from Measurements LIMIT 10")
first_ten
```

```
  Identifier SubjectNumber Session  Ear Instrument Age AgeCategory
1  Abur_2014             1       1 Left     HearID  20       Adult
2  Abur_2014             1       1 Left     HearID  20       Adult
3  Abur_2014             1       1 Left     HearID  20       Adult
4  Abur_2014             1       1 Left     HearID  20       Adult
5  Abur_2014             1       1 Left     HearID  20       Adult
6  Abur_2014             1       1 Left     HearID  20       Adult
```

- see more: "Modern Data Science with R (2e+)" (Baumer, Kaplan, and Horton, 2024, https://mdsr-book.github.io/mdsr3e/)