```
-- controlling-duplicates

-- In SQL, select-from-where (SFW) queries use bag semantics by default
-- but we can override this by saying DISTINCT.
-- UNION/INTERSECT/EXCEPT do they opposite: they is set semantics by default
-- but we can overrid this by saying ALL.
-- This demo explores these concepts in more detail.


dbsrv1% psql csc343h-dianeh
psql (9.1.14)
Type "help" for help.

csc343h-dianeh=> set search_path to university;
SET

-- A query with duplicate results
csc343h-dianeh=> select oid from Took where grade > 95;
 oid
-----
  16
  11
  13
  39
  11
  13
  16
  22
   1
  14
(10 rows)

-- Order by so we can see it more clearly.
csc343h-dianeh=> select oid
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade > 95
csc343h-dianeh-> order by oid;
 oid
-----
   1
  11
  11
  13
  13
  14
  16
  16
  22
  39
(10 rows)

-- Introduce distinct to say we don't want the duplicates.
csc343h-dianeh=> select distinct oid
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade > 95
csc343h-dianeh-> order by oid;
 oid
-----
   1
  11
  13
  14
  16
  22
```

```
  39
(7 rows)

-- Aside:
-- We have seen DISTINCT before, inside the brackets on an aggregation.
-- Example: SELECT count(DISTINCT sid) FROM ...
-- In that case, we are asking that only the distinct sid values contribute
-- to the count.
-- Now, we are putting the DISTINCT outside of any aggregation.

-- Back to our example.
-- Here we add another column.  It has duplicates in different places
-- than oid did.
csc343h-dianeh=> select distinct oid, grade
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade > 95
csc343h-dianeh-> order by oid;
 oid | grade
-----+-------
   1 |    99
  11 |    99
  13 |    98
  13 |    99
  14 |    98
  16 |   100
  16 |    98
  22 |    96
  39 |    97
(9 rows)

-- We can't ask for both to be distinct.  That couldn't work because the duplicates in
-- column oid occur in different rows than the duplicates in column grade.
-- So SQL won't let us write a query that attempts to do this.
csc343h-dianeh=> select distinct oid, distinct grade
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade > 95
csc343h-dianeh-> order by oid;
ERROR:  syntax error at or near "distinct"
LINE 1: select distinct oid, distinct grade from Took where grade > ...
                             ^

-- Distinct actually works at the level of the row, not individual cells.
-- It turns the result of the query into a set, rather than a bag.

-- We can only say distinct once, right before we list the columns that
-- we want in the result.
csc343h-dianeh=> select oid, distinct grade
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade > 95
csc343h-dianeh-> order by oid;
ERROR:  syntax error at or near "distinct"
LINE 1: select oid, distinct grade from Took where grade > 95 order ...
                    ^

-- Let's try another query with >1 column where we can get a non-set back.
csc343h-dianeh=> select sid, grade
csc343h-dianeh-> from took
csc343h-dianeh-> order by sid, grade;
  sid  | grade
-------+-------
   157 |    39          <-- There are repeated sids, such as 157
   157 |    59          <-- There are repeated grades per sid such as <157, 59>
   157 |    59              ie., entire repeated rows.
   157 |    62
   157 |    71
```

```
   157 |    71
   157 |    72
   157 |    75
   157 |    82
   157 |    82
   157 |    89
   157 |    90
   157 |    91
   157 |    98
   157 |    99
 11111 |     0
 11111 |    17
 11111 |    40
 11111 |    45
 11111 |    46
 98000 |    54
 98000 |    72          <-- There are grades, such as 72, that are repeated
 98000 |    78              across sids
 98000 |    78              (i.e., repeats that are in the grades column only)
 98000 |    79
 98000 |    79
 98000 |    79
 98000 |    82
 98000 |    89
 98000 |    89
 98000 |    89
 98000 |    92
 98000 |    93
 98000 |    97
 98000 |    98
 99132 |    39
 99132 |    62
 99132 |    75
 99132 |    79
 99132 |    82
 99132 |    98
 99132 |    99
 99999 |    52
 99999 |    70
 99999 |    71
 99999 |    76
 99999 |    78
 99999 |    89
 99999 |    91
 99999 |    94
 99999 |    96
 99999 |    99
 99999 |    99
 99999 |   100
(54 rows)

-- With DISTINCT, we lose only entire repeated rows.
csc343h-dianeh=> select distinct sid, grade
csc343h-dianeh-> from took
csc343h-dianeh-> order by sid, grade;
  sid  | grade
-------+-------
   157 |    39          <-- We still have repeated sids, such as 157
   157 |    59          <-- But <157, 59> occurs in only one row
   157 |    62
   157 |    71
   157 |    72
   157 |    75
   157 |    82
   157 |    89
```

```
   157 |    90
   157 |    91
   157 |    98
   157 |    99
 11111 |     0
 11111 |    17
 11111 |    40
 11111 |    45
 11111 |    46
 98000 |    54
 98000 |    72          <-- We still have grades, such as 72, that are
 98000 |    78              repeated across sids
 98000 |    79
 98000 |    82
 98000 |    89
 98000 |    92
 98000 |    93
 98000 |    97
 98000 |    98
 99132 |    39
 99132 |    62
 99132 |    75
 99132 |    79
 99132 |    82
 99132 |    98
 99132 |    99
 99999 |    52
 99999 |    70
 99999 |    71
 99999 |    76
 99999 |    78
 99999 |    89
 99999 |    91
 99999 |    94
 99999 |    96
 99999 |    99
 99999 |   100
(45 rows)

-- So SFW queries by default include duplicates.
-- Set ops do the opposite.

-- Here are two SFW queries that we'll union together in a sec.
-- The first one has duplicates (and they're left in).
csc343h-dianeh=> select sid
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade > 95;
  sid
-------
 99132
 99132
 98000
 98000
 99999
 99999
 99999
 99999
   157
   157
(10 rows)

-- The second one does too.
csc343h-dianeh=> select sid
csc343h-dianeh-> from Took
csc343h-dianeh-> where grade < 50;
```

```
  sid
-------
 99132
   157
 11111
 11111
 11111
 11111
 11111
(7 rows)

-- But when we do union, we don't get all 17 rows.  The duplicates are
-- eliminated, by default.
csc343h-dianeh=> (select sid from Took where grade > 95)
csc343h-dianeh-> union
csc343h-dianeh-> (select sid from Took where grade < 50);
  sid
-------
 98000
 99132
 99999
   157
 11111
(5 rows)

-- UNION ALL says we want them.
csc343h-dianeh=> (select sid from Took where grade > 95)
csc343h-dianeh-> union all
csc343h-dianeh-> (select sid from Took where grade < 50);
  sid
-------
 99132
 99132
 98000
 98000
 99999
 99999
 99999
 99999
   157
   157
 99132
   157
 11111
 11111
 11111
 11111
 11111
(17 rows)
```