CSC 369: Distributed Computing

Alex Dekhtyar

April 17

Day 6: The Algebra Of Data Transformations Part II

db.collection.aggregate()

Housekeeping

- Lab 2:
  - Submit from unix1-2-3-4-5
  - `handin dekhtyar lab02 <files>`
- **Lab 3:**
- **Lab 4: Python application, teams of 2.**

Recall from last class

```
{name:”Alex”,

teaches:[“CSC 369”, “DATA 452”],

department:”CSSE”,

enrollments:[28,20],

position: “professor”,

office:{building:14, room:210}

}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Decomposition into atomic operations in “Generalized” Data Algebra

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

Keep information about only "CSC", "CPE", and "DATA" courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

Express using these operations….

Relational Algebra	Generalized Algebra
Selection	Filtering
Projection	Projection/Transformation
Set Operations	
Join	Join
Grouping/Aggregation	Grouping/Aggregation
Sort	Sort

Very Tersely

`Filtering`	Given a condition - keep only objects that satisfy it
`Projection` `Transformation`	Modify the contents of its object based solely on what's in the object itself
`Grouping`	Break collection into groups, each representing objects with same values of some keys
`Aggregation`	Compute an aggregate value over a set of objects
`Join`	Combine objects from two different collections based on matches in values of some keys
`Sort`	Return objects in a specific order

… and a few more

`Ungrouping` `Unwinding`	Opposite of grouping - build an object for each element of an array
`Limit`	Return a specific number of documents
`Skip`	Return documents after skipping a specified number
`Sample`	Return a random sample of documents

Let’s use our intuition

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Exercise Time!!

File posted to Slack/chat.

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Filtering

Remove unnecessary data

Projection

Find the total enrollment for each CSSE instructor and number of sections taught

Projection

Aggregation

Find the largest total enrollment for a CSSE instructor

Aggregation

Projection

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Filtering

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

Exercise Time!!

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

Unwinding

Filtering

Projection

Grouping

Sort

Now, let’s learn all this for real...

```
db.collection.aggregate(<aggregation
                                pipeline>)
```

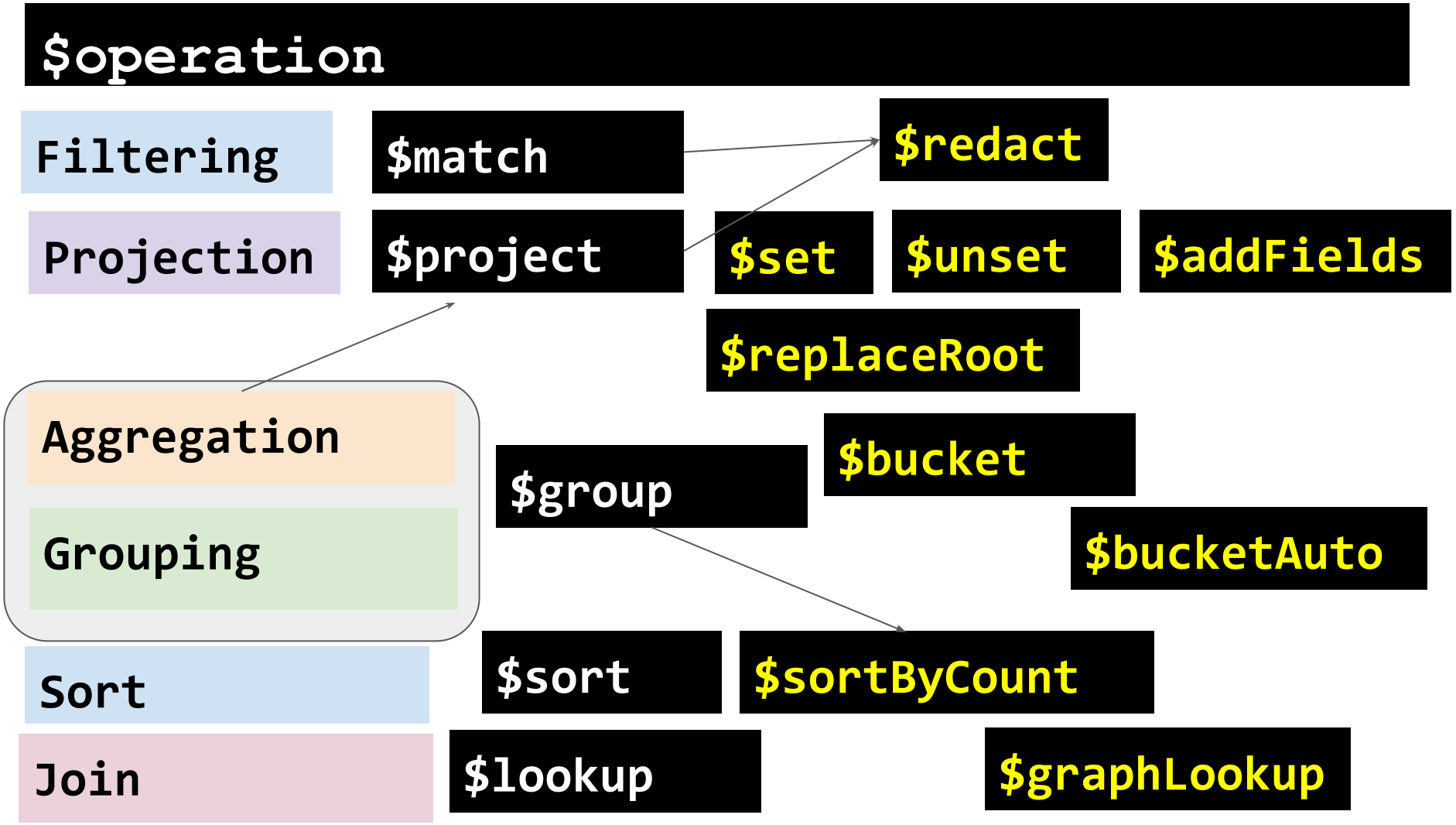
Now, let’s learn all this for real...

```
db.collection.aggregate({$operation:{<doc>}},
                        {$operation:{<doc>}},
                          …,
                       })
```

Now, let's learn all this for real...

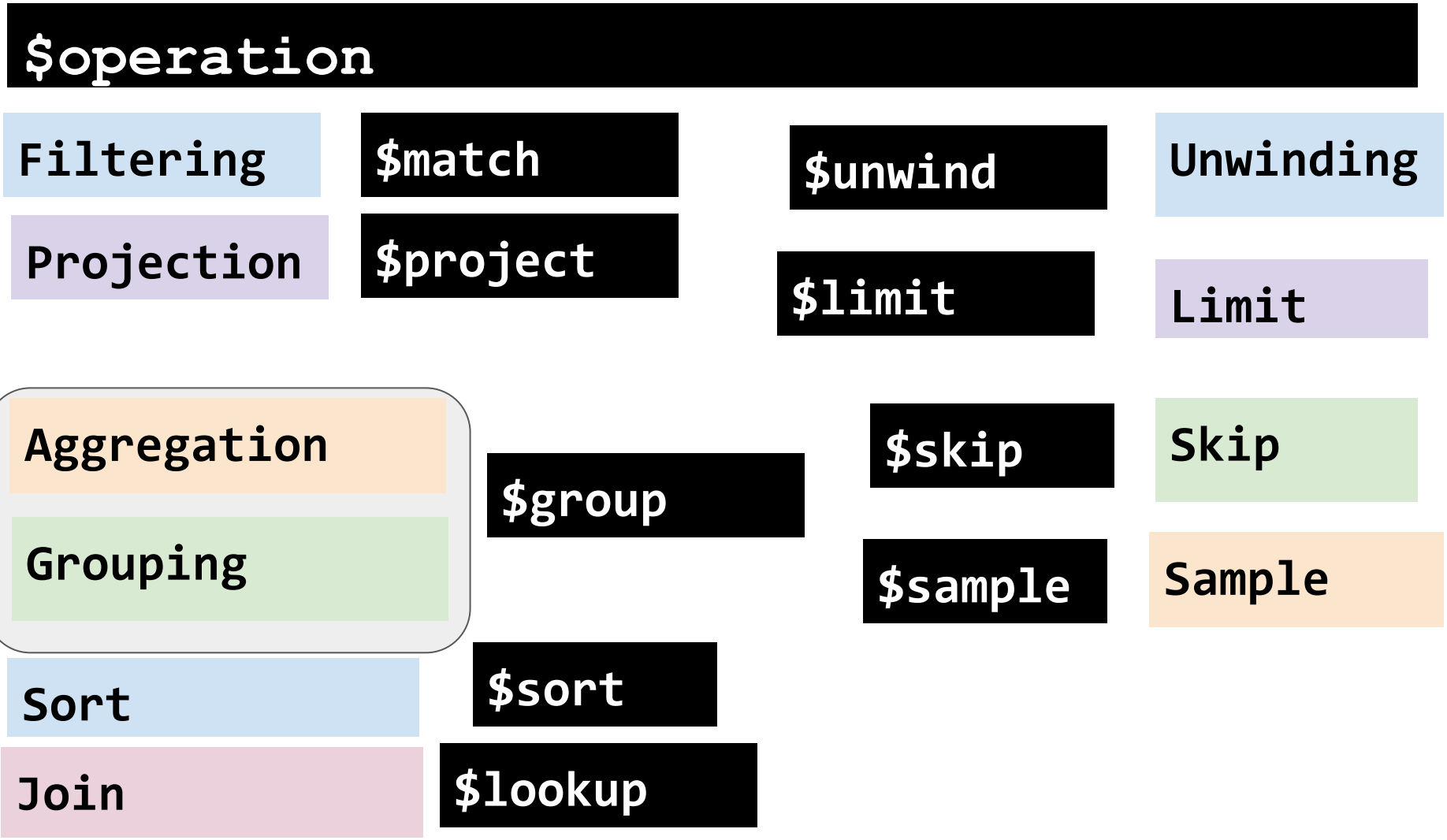
```
db.collection.aggregate({$operation:{<doc>}},
                        {$operation:{<doc>}},
                          …,
                       })
```

In Lecture: basic ideas
In Lab: all the syntax you can handle


$operation
Filtering
$match
$redact
Projection
$project
$set
$unset
$addFields
$replaceRoot
Aggregation
Grouping
$group
$bucket
$bucketAuto
Sort
$sort
$sortByCount
Join
$lookup
$graphLookup

$operation

Filtering	$match
Projection	$project
Aggregation	$group
Grouping	
Sort	$sort
Join	$lookup


$operation
Filtering
$match
$unwind
Unwinding
Projection
$project
$limit
Limit
Aggregation
$group
$skip
Skip
Grouping
$sample
Sample
Sort
$sort
Join
$lookup

Let’s Learn By Doing

Query 2 first (it is simpler)

Query 1 second (it has **layers**)

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

Unwinding

Filtering

Projection

Grouping

Sort

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

`$unwind`

```
{$unwind: "$teaches"}
```

Keep information about only “CSC”, “CPE”, and “DATA” courses.

`$match`

Remove unnecessary data

`$project`

For each course, combine instructors teaching it into a list

`$group`

Sort?

`$sort`

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

`$unwind`

```
{$unwind: "$teaches"}
```

Keep information about only "CSC", "CPE", and "DATA" courses.

`$match`

```
{$match:
 {teaches: {$in: ["CSC",
                  "CPE","DATA"]}
}}
```

Remove unnecessary data

`$project`

For each course, combine instructors teaching it into a list

`$group`

Sort?

`$sort`

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

`$unwind`

```
{$unwind: "$teaches"}
```

Keep information about only "CSC", "CPE", and "DATA" courses.

`$match`

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}
}}
```

Remove unnecessary data

`$project`

For each course, combine instructors teaching it into a list

`$group`

Sort?

`$sort`

Regular Expressions!!!

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

$unwind

```
{$unwind: "$teaches"}
```

Keep information about only "CSC", "CPE", and "DATA" courses.

$match

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}}}
```

Remove unnecessary data

$project

For each course, combine instructors teaching it into a list

$group

Sort?

$sort

```
{"name" : "Alex",
 "teaches" : "CSC 369",
 "department" : "CSSE",
 "enrollments" : [28,20],
 "position" : "professor",
 "office" : {"building" : 14,
             "room" : 210 }
}
```

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

```
{$unwind: "$teaches"}
```

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}}}
```

Deconstruct “teaches” arrays, create one object per instructor-course pairing

$unwind

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$match

Remove unnecessary data

$project

```
{$project:{_id:0,
           Name:1, department:1,
           course:"$teaches"}
}
```

For each course, combine instructors teaching it into a list

$group

Sort?

$sort

```
 "office" : {"building" : 14,
             "room" : 210 }
}
```

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

$unwind

Keep information about only "CSC", "CPE", and "DATA" courses.

$match

Remove unnecessary data

$project

For each course, combine instructors teaching it into a list

$group

Sort?

$sort

```
{$unwind: "$teaches"}
```

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}}}
```

```
{$project:{_id:0,
           Name:1, department:1,
           course:"$teaches"}}
```

```
{$group:{_id:"$course",
   instructors:{$push: {
                  name:"$name",
     department:"$department"}}
}
```

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

$unwind

```
{$unwind: "$teaches"}
```

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$match

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}}}
```

Remove unnecessary data

$project

```
{$project:{_id:0,
           Name:1, department:1,
           course:"$teaches"}}
```

For each course, combine instructors teaching it into a list

$group

```
{$group:{_id:"$course",
   instructors:{$push: {
                  name:"$name",
     department:"$department"}}
}
{$project:{_id:0, instructors:1,
          $course:"$_id"}}
```

Sort?

$sort

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

$unwind

```
{$unwind: "$teaches"}
```

Keep information about only "CSC", "CPE", and "DATA" courses.

$match

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}}}
```

Remove unnecessary data

$project

```
{$project:{_id:0,
           Name:1, department:1,
           course:"$teaches"}}
```

For each course, combine instructors teaching it into a list

$group

```
{$group:{_id:"$course",
   instructors:{$push: {
                  name:"$name",
     department:"$department"}}
}
{$project:{_id:0, instructors:1,
          $course:"$_id"}}
```

Sort?

$sort

Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct "teaches" arrays, create one object per instructor-course pairing

$unwind

```
{$unwind: "$teaches"}
```

Keep information about only "CSC", "CPE", and "DATA" courses.

$match

```
{$match:
 {teaches: {$in: [/^CSC/,
                  /^CPE/,/^DATA/]}}}
```

Remove unnecessary data

$project

```
{$project:{_id:0,
           Name:1, department:1,
           course:"$teaches"}}
```

For each course, combine instructors teaching it into a list

$group

```
{$group:{_id:"$course",
   instructors:{$push: {
                  name:"$name",
     department:"$department"}}
}
{$project:{_id:0, instructors:1,
          $course:"$_id"}}
```

Sort?

$sort

```
{$sort: {course:1}}
```

```
db.spring.aggregate(
{$unwind:"$teaches"},
{$match: {teaches: {$in:[/^CSC/, /^CPE/, /^DATA/]}
         }},
{$project: {_id:0,
            name:1, department:1,
            course:"$teaches"}
},
{$group:{_id:"$course",
         instructors:{$push: {name:"$name",
                              department:"$department"}
                     }}
},
{$project:{_id:0, instructors:1,
           course:"$_id"}},
{$sort:{course:1}}
)
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Filtering

Remove unnecessary data

Projection

Find the total enrollment for each CSSE instructor and number of sections taught

Projection

Aggregation

Find the largest total enrollment for a CSSE instructor

Aggregation

Projection

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Filtering

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

`$match`

Remove unnecessary data

`$project`

Find the total enrollment for each CSSE instructor and number of sections taught

`$project`

Find the largest total enrollment for a CSSE instructor

`$group`

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

`$match`

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

`$match`

```
{$match: {department:"CSSE"}}
```

Same as for db.collection.find()

Remove unnecessary data

`$project`

Find the total enrollment for each CSSE instructor and number of sections taught

`$project`

Find the largest total enrollment for a CSSE instructor

`$group`

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

`$match`

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

Remove unnecessary data

$project

Find the total enrollment for each CSSE instructor and number of sections taught

$project

Find the largest total enrollment for a CSSE instructor

$group

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

$match

```
{$match: {department:"CSSE"}}
```

```
{$project:{... }
}
```

```
{name:"Alex",
teaches:["CSC 369", "DATA 452"],
department:"CSSE",
enrollments:[28,20],
position: "professor",
office:{building:14, room:210}
}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

$project

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taught

$project

Find the largest total enrollment for a CSSE instructor

$group

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

$match

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

`$match`

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

`$project`

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taught

`$project`

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"}
}
```

Find the largest total enrollment for a CSSE instructor

`$group`

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

`$match`

```
{ "name" : "Alex",
  "enrollments" : [ 28, 20 ]
}
```

28+20

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

`$match`

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

`$project`

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taught

`$project`

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"}
}
```

Find the largest total enrollment for a CSSE instructor

`$group`

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

`$match`

```
{ "name" : "Alex",
  "enrollments" : 48
}
```

```
'Kirsten",
nts" : 108
}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

$project

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"}
}
```

Find the total enrollment for each CSSE instructor and number of sections taugh

$project

Find the largest total enrollment for a CSS instructor

$group

```
{$group:{_id:"1"
  mEnr: {$max: "$enrollments"}}
}
```

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

$match

```
{ "name" : "Alex",
  "enrollments" : 48
}
```

'Kirsten",
nts" : 108
}

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

$project

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taug

$project

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"}
}
```

Find the largest total enrollment for a CS instructor

$group

```
{$group:{_id:"1"
 data:{$push:{name:"$name",
              enr:"$enrollments"}},
mEnr: {$max: "$enrollments"}}
}
```

Compare each instructor's total enrollme the largest; keep only instructors with la enrollment

$match

```
}                                    nts" : 108
}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

$project

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taug

$project

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"}
}
```

Find the largest total enrollment for a CS instructor

$group

```
{$group:{_id:"1"
 data:{$push:{name:"$name",
              enr:"$enrollments"}},
mEnr: {$max: "$enrollments"}}
},
{$unwind:"$data"}
```

Compare each instructor's total enrollme the largest; keep only instructors with la enrollment

$match

```
}                                    nts" : 108
}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

$project

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taught

$project

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"} }}
```

Find the largest total enrollment for CSSE instructor

$group

```
{$group:{_id:"1"
 data:{$push:{name:"$name",
         enr:"$enrollments"}},
mEnr: {$max: "$enrollments"}}
},
{$unwind:"$data"}
```

Compare each instructor's total enr the largest; keep only instructors w enrollment

$match

```
{$match:{
  maxEnrollment:"$data.enrollments"
 }}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

$project

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

Find the total enrollment for each CSSE instructor and number of sections taught

$project

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"} }}
```

Find the largest total enrollment for CSSE instructor

$group

```
{$group:{_id:"1"
 data:{$push:{name:"$name",
         enr:"$enrollments"}},
mEnr: {$max: "$enrollments"}}
},
{$unwind:"$data"}
```

Compare each instructor's total enr… the largest; keep only instructors w… enrollment

$match

```
{$match:{
  maxEnrollment:"$data.enrollments"
 }}
```

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"} }}
```

Find th
instruc

Find the
instruct

Compar
the larg
enrollment

```
{$project:{diff: {$subtract:["$maxEnrollment",
                             "$data.enrollments"]}
}

{$match:{
  maxEnrollment:"$data.enrollments"
 }}
```

$match

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

$match

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"} }}
```

Find th
instruc

Find the
instruct

Compar
the larg
enrollment

```
{$project:{diff: {$subtract:["$maxEnrollment",
                             "$data.enrollments"]},
           _id:0,
           name:"$data.name",
           enrollments:"$data.enrollments"
}
{$match:{
  maxEnrollment:"$data.enrollments"
 }}
```

$match

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

`$match`

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"} }}
```

Find th
instruc

Find the
instruct

Compar
the larg
enrollment

```
{$project:{diff: {$subtract:["$maxEnrollment",
                             "$data.enrollments"]},
           _id:0,
           name:"$data.name",
           enrollments:"$data.enrollments"
}
{$match:{ diff:0} }
```

`$match`

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

`$match`

```
{$match: {department:"CSSE"}}
```

Remove unnecessary data

```
{$project:{_id:0, name:1,
           enrollments:1 }
}
```

```
{$project:{name:1,
 enrollments:{$sum:"$enrollments"} }}
```

Find th
instruc

Find the
instruct

Compar
the larg
enrollment

```
{$project:{diff: {$subtract:["$maxEnrollment",
                             "$data.enrollments"]},
           _id:0,
           name:"$data.name",
           enrollments:"$data.enrollments"
}
{$match:{ diff:0} }

{$project: {diff:0}}
```

`$match`

```
db.spring.aggregate(
 {$match: {department:"CSSE"}},
 {$project:{_id:0, name:1, enrollments:1}                    //cleaning
 },
 {$project:{name:1,                                          //transformation
            enrollments: {$sum: "$enrollments"}}
  },
 {$group:{ _id:"1",
          maxEnrollment:{$max:"$enrollments"},
          data: {$push:{name:"$name", enrollments:"$enrollments"}}
          }
 },
 {$unwind:"$data"},
 {$project:{_id:0,
            diff: {$subtract:["$maxEnrollment", "$data.enrollments"]},
            name:"$data.name",
            enrollments:"$data.enrollments"
           }
 },
 {$match:{diff:0}},
 {$project:{diff:0}}
)
```