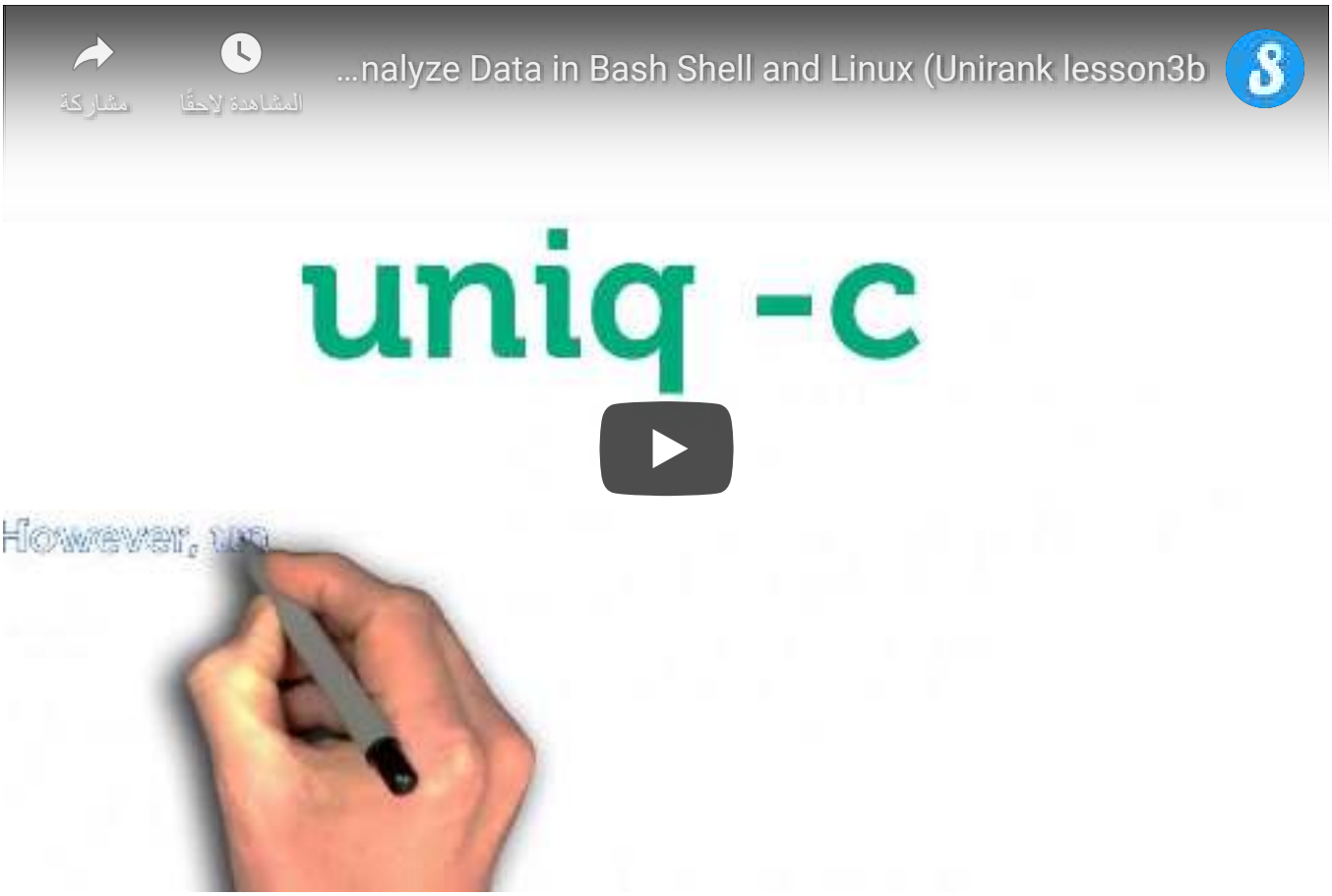
Finding the number of Institutes from each state

WE'LL COVER THE FOLLOWING

- Do you want to know more?

At this point we want to calculate how many Institutes have been ranked from each of the US states in the dataset. Let's watch the following video lecture to understand the lesson goal fast!



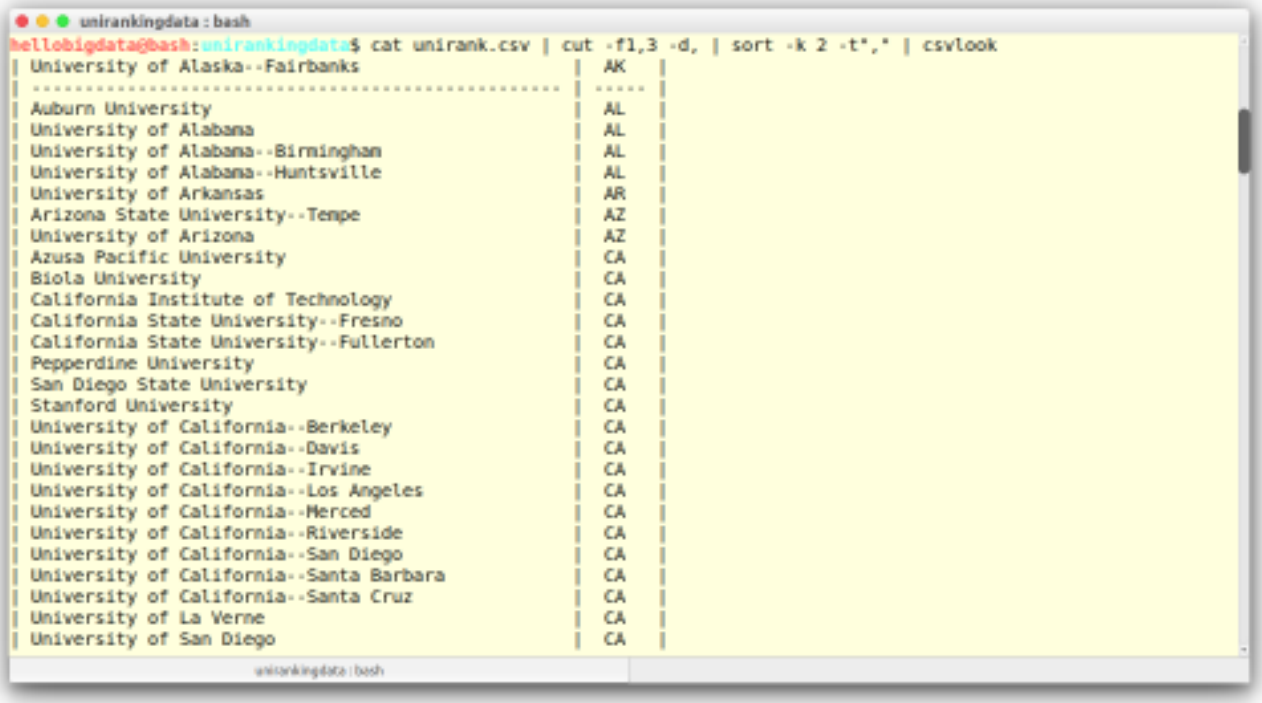
Video lecture: Finding the number of Institutes from a given state

Let's start by extracting only the part of each line that is relevant to us. In our case, notice that we are interested in column `#1` and `3` (`university` and `state` names, respectively). To extract these columns, we can make use of a command called `cut` as follows:

```
cat unirank.csv | cut -f1,3 -d,
```

Here, the command-line option `-f` specifies which field (column) to extract or `cut` out from the file and the option (`d,`) tells that we want delimit the cuts by comma (`,`). When you run that command, you should see that the output consist only of lines such as university `names` and `states`. Note that, despite its name, the `cut` command does not modify the original file it acts on. Now onto the last part. We would like to count how many unis came from each state. However, this is a complex procedure and there isn't one command that can do all that; we will have to use two commands. Here we need the command `uniq -c` to count (hence the -c) how many unique appearances of each state. However, `uniq -c` requires the input to be sorted, so the first step is to `sort` the list of universities and states. We can do this very easily with a command that is conveniently called `sort` :

```
cat unirank.csv | cut -f1,3 -d, | sort -k 2 -t","
```



Institutes by states

The `sort` options: `k 2` tells sort function to select the column 2 as a key and `t","` option tells that the delimiter is a comma (`,`).

```
cat unirank.csv | cut -f1,3 -d, | csvlook
```

Notice that, as a result of our list being sorted, all the lines with same state are right next to each other. Now, as mentioned in our plan above, we'll use uniq -c to "condense" neighboring lines that are the same and in the process, count how many of each are seen:

```
cat unirank.csv | cut -f3 -d, | sort | uniq -c | csvlook
```

22 Institutes from the CA (California) state!

Do you want to know more?

'cut' man page

'cat' man page

'sort' man page

'uniq' man page