

# Big Data Analysis

## Setting up an AWS with RStudio Server

*Marcel Neunhoeffler - University of Mannheim*

*June 03 2019*

### Setting up a AWS account

To get started you need an AWS account. Even though everything will be in the free tier, you will need to give AWS some payment information.

### Choosing the right RStudio Server Amazon Machine Image (AMI)

Academia is great, there are super helpful people out there. Louis Aslett is one of those great people. He created (and maintains) RStudio Server Amazon Machine Images so that it is really easy for us to get started. Go to his homepage [http://www.louisaslett.com/RStudio\\_AMI/](http://www.louisaslett.com/RStudio_AMI/) and choose the right image for you (depending on your location).

### Launching the instance

#### Choosing the right instance type

- Today we will work with the free tier eligible `t2.micro` instance.
- If you want to tackle big data problems you will have to choose an appropriate instance type for your problem. The main thing for big data applications will often be a big enough memory (especially since R will load all data directly into memory). You will find instances with 0.5 GB up to instances with 3904 GB of memory (that seems a bit excessive though). The second dimension to choose from is the number of CPU cores aka computing power. Especially if you can parallelize code such that it can run on multiple cores simultaneously it might make sense to pick an instance with more cores. You will find instances with 1 core up to instances with 96 cores. If you want to implement a deep learning solution it might be useful to add a GPU to your instance.
- After you chose the appropriate instance, click **Review and Launch**.

#### Editing the Security Group Settings

- We need to edit the security groups slightly to work with RStudio Server later on. Specifically we need to open the following ports:
  - 22 for SSH
  - 80 for HTTP
  - 443 for HTTPS
  - 8787 for RStudio Server
- Click **Review and Launch**
- Click **Launch**

## Creating (or choosing) a key pair

- On your Desktop create a folder called **AWS**
- Choose **Create a new key pair** and call it **rstudio**
- Click **Download Key Pair**
- Place the downloaded file in the **AWS** folder on your desktop
- Click **Launch Instance**

## Working with the instance

- Click **View Instances**
- Wait until it is done **Initializing** (this might take some minutes).

## Connecting to Rstudio Server

- Go to the Public DNS of your instance.
- Open the public DNS in your favorite browser.
- Copy your 'Instance ID'
- Login to Rstudio Server.
  - Your Username: **rstudio**
  - Your Password: **<Your Instance ID>**

## Done!

- Happy Big Data Analysis!

## After your Analysis

- Don't forget to terminate your instance!
- Before you do, download all your results. Or they will be lost.