

HANJO ODENDAAL

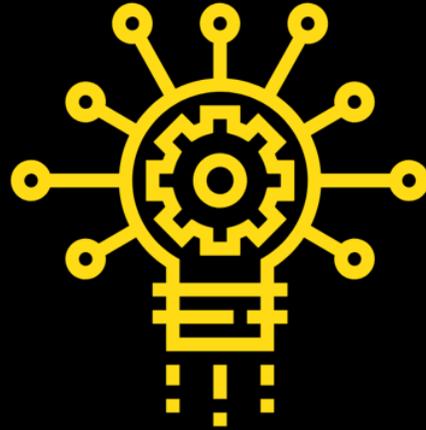


H₂O ai

Scalable Analytics with H2O and AWS

SatRday 2019 - Johannesburg

www.daeconomist.com



DATA SCIENCE GOALS



*If you take ONE thing away from today's talk, it should be this:
Take R100 (\$7.14), spin up BIG AWS machine (r4.16xlarge) and
watch for 10 hours light up like a christmas tree while telling
your line manager your tuning you ML model. #geekgoals*

H₂O.ai



NO BULLSH*T DATA SCIENCE

Szilard Pafka's¹ R/Finance 2017 presentation had a big influence on how I saw the hype around the industry

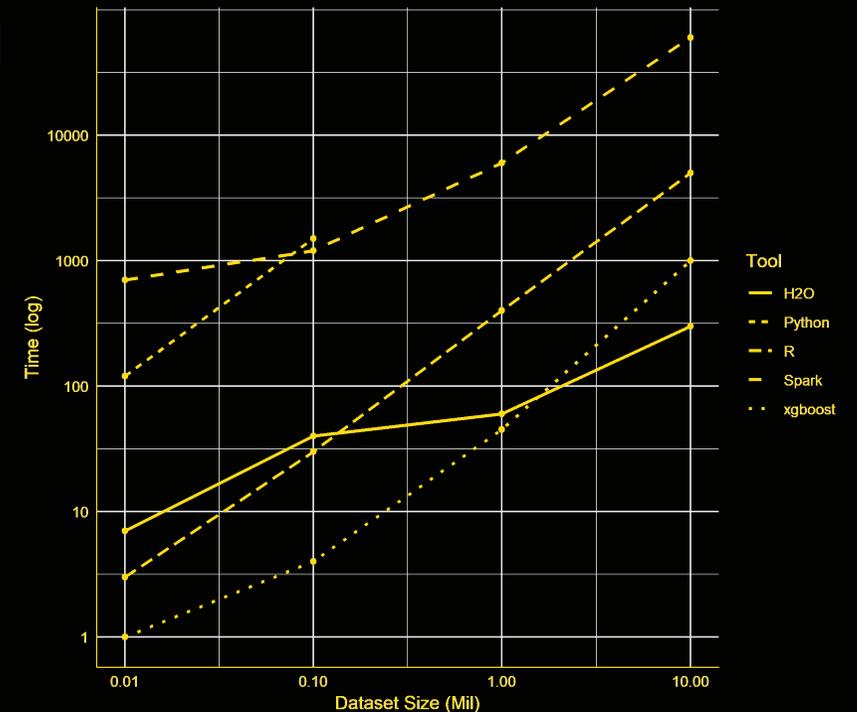
- Don't always believe the hype!
- Scaling shouldn't be an exercise in fighting your tools

Stop wasting time - if you interested in machine learning. Learn all you can about XGBoost and Random Forests²

[1] Hope I am pronouncing this correctly

[2] Which means you can concentrate on the stats and maths behind each model

Talk source: <https://bit.ly/2ErsBtB>



NO BULLSH*T DATA SCIENCE

Szilard Pafka's¹ R/Finance 2017 presentation had a big influence on how I saw the hype around the industry

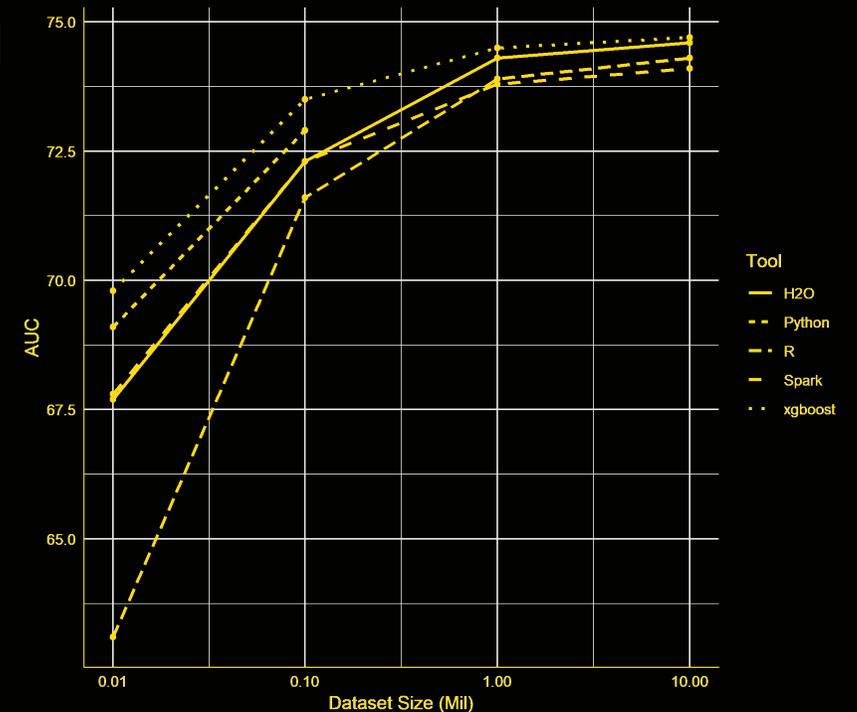
- Don't always believe the hype!
- Scaling shouldn't be an exercise in fighting your tools

Stop wasting time - if you interested in machine learning. Learn all you can about XGBoost and Random Forests²

[1] Hope I am pronouncing this correctly

[2] Which means you can concentrate on the stats and maths behind each model

Talk source: <https://bit.ly/2ErsBtB>



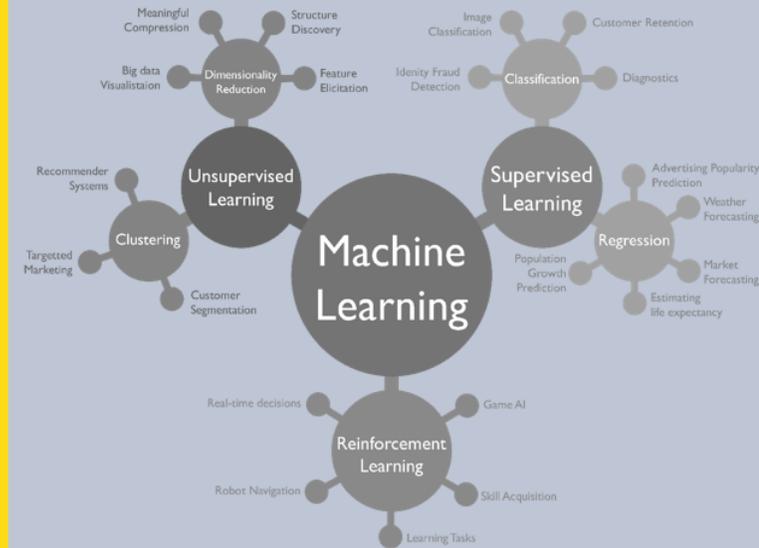
H2O

Leading open source platform for machine learning and artificial intelligence. Multiple API interface such as [REST](#), [SOAP](#), [Python](#), [R](#), [Java](#), [C++](#).

Well developed interfaces around GPUs.

AWS.EC2

Amazon Elastic Compute. For analytics, their [EMR](#) are perfect as they come at an enourmously discounted price as well as delivering bag for buck.



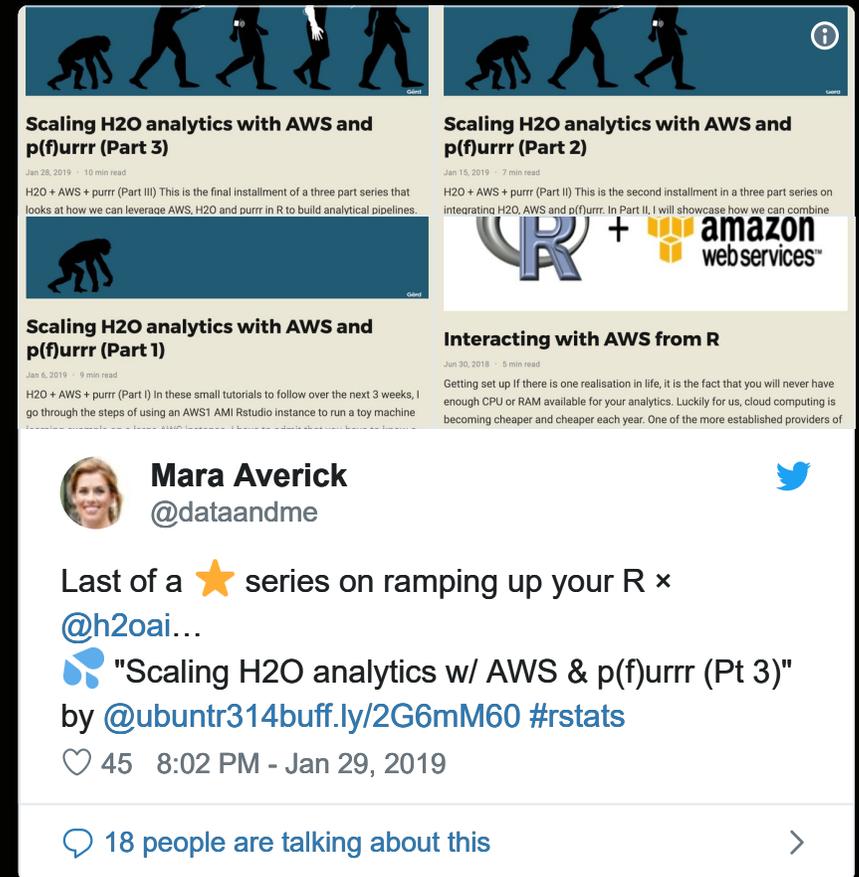


80% exploratory research still in caret

But, running H2O in live systems.

WHERE DO YOU START?

- Very nice high level functions and methods
- It really does scale nicely on large datasets
- Code interface smells good



Scaling H2O analytics with AWS and p(f)urrr (Part 3)
Jan 28, 2019 · 10 min read
H2O + AWS + purrr (Part III) This is the final installment of a three part series that looks at how we can leverage AWS, H2O and purrr in R to build analytical pipelines.

Scaling H2O analytics with AWS and p(f)urrr (Part 2)
Jan 15, 2019 · 7 min read
H2O + AWS + purrr (Part II) This is the second installment in a three part series on integrating H2O, AWS and p(f)urrr. In Part II, I will showcase how we can combine

Scaling H2O analytics with AWS and p(f)urrr (Part 1)
Jan 6, 2019 · 9 min read
H2O + AWS + purrr (Part I) In these small tutorials to follow over the next 3 weeks, I go through the steps of using an AWS1 AMI Rstudio instance to run a toy machine

Interacting with AWS from R
Jun 30, 2018 · 5 min read
Getting set up If there is one realisation in life, it is the fact that you will never have enough CPU or RAM available for your analytics. Luckily for us, cloud computing is becoming cheaper and cheaper each year. One of the more established providers of

Mara Averick
@dataandme

Last of a ★ series on ramping up your R ×
[@h2oai...](#)
"Scaling H2O analytics w/ AWS & p(f)urrr (Pt 3)"
by [@ubuntr314buff.ly/2G6mM60](#) #rstats

45 8:02 PM - Jan 29, 2019

18 people are talking about this

NO EXCUSE NOT TOO PLAY

r3.xlarge

- 4 cores
- 30.5 GB RAM
- \$0.0379 (R0,53/h)

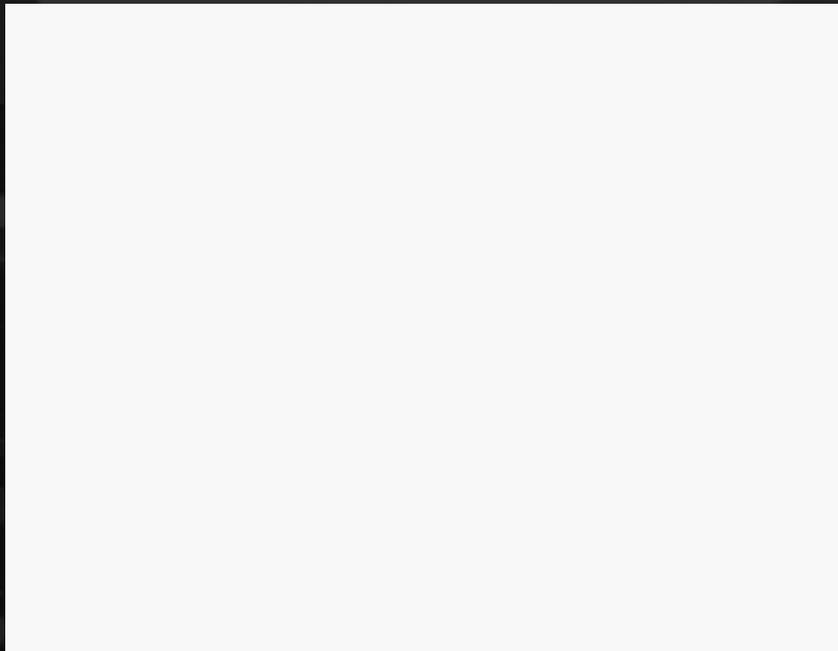
r4.16xlarge

- 64 cores
- 488 GB RAM
- \$0.6974 (R9,76/h)

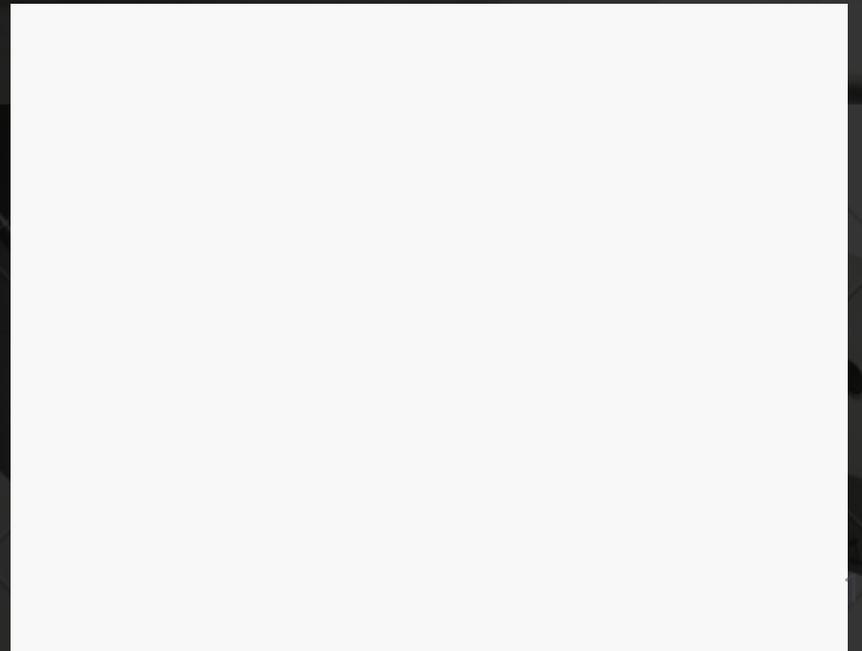


H2O HAS A SIGNIFICANT AMOUNT OF TUNING OPTIONS:

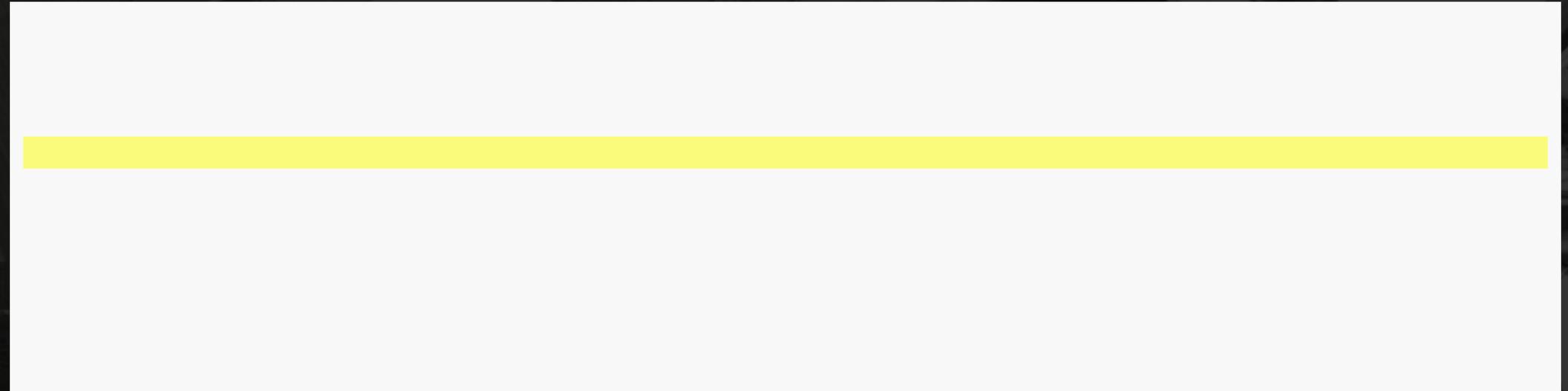
XGBoost:



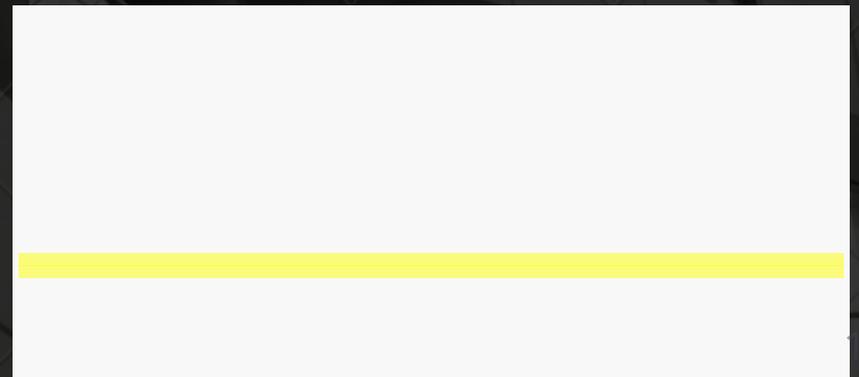
Random Forest:



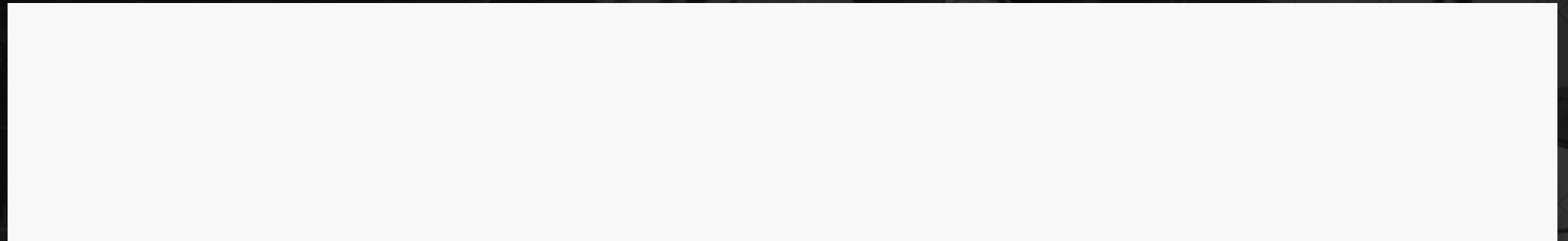
HELLO WORLD H2O:



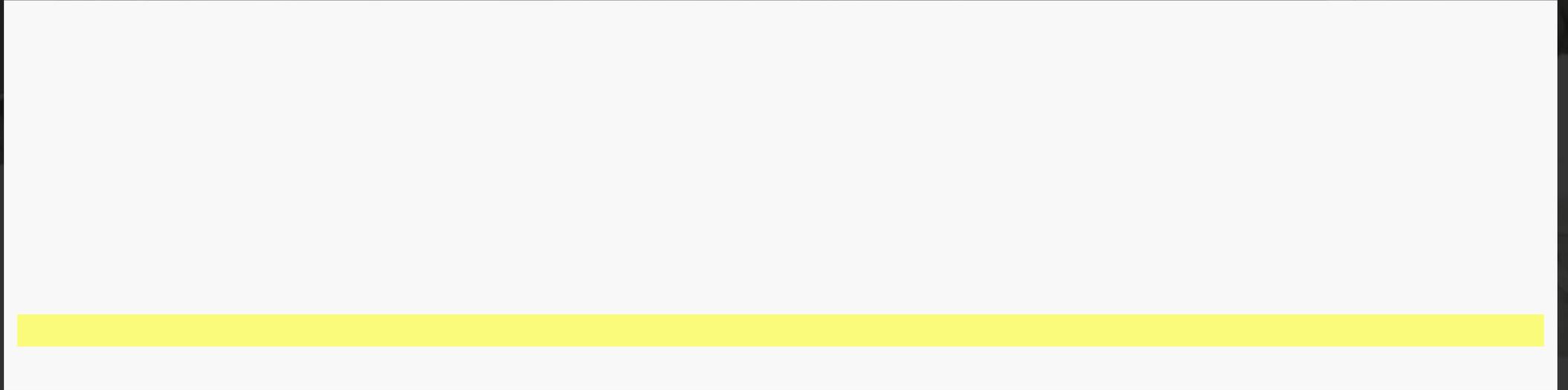
Messaging confirming the server has been initialized successfully



HELLO WORLD H2O:



HELLO WORLD H2O:

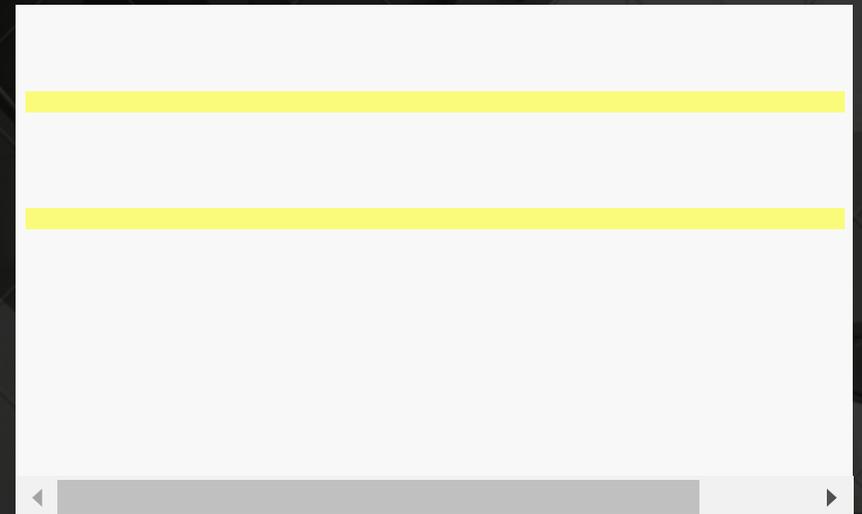
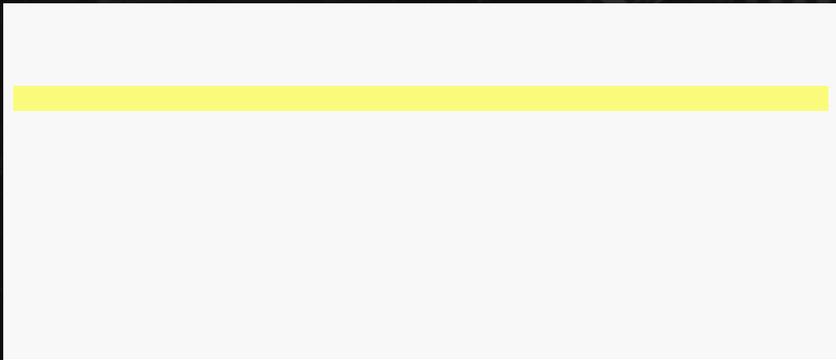


FURRR NEVER LOOKED SO GOOD

Interact with EC2 through

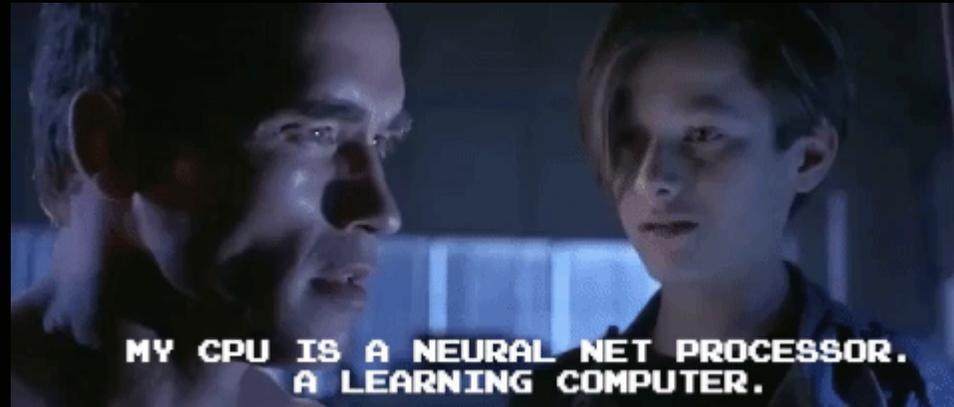
:

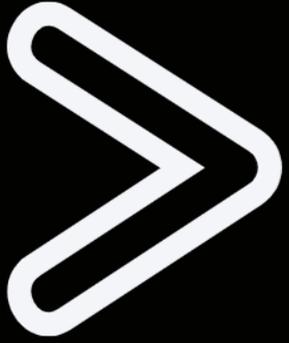
to connect to server





ON TO AN EXAMPLE





THANK YOU

