duolingo

Ohio LinuxFest
2023

Max Blaze

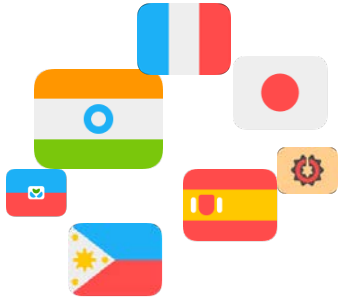
When Clouds Stop Raining Discount$

Surviving the Drought

our mission is
to develop the
best education in
the world
and make it
universally
available

#1 education app in the world

74M+ MAUs



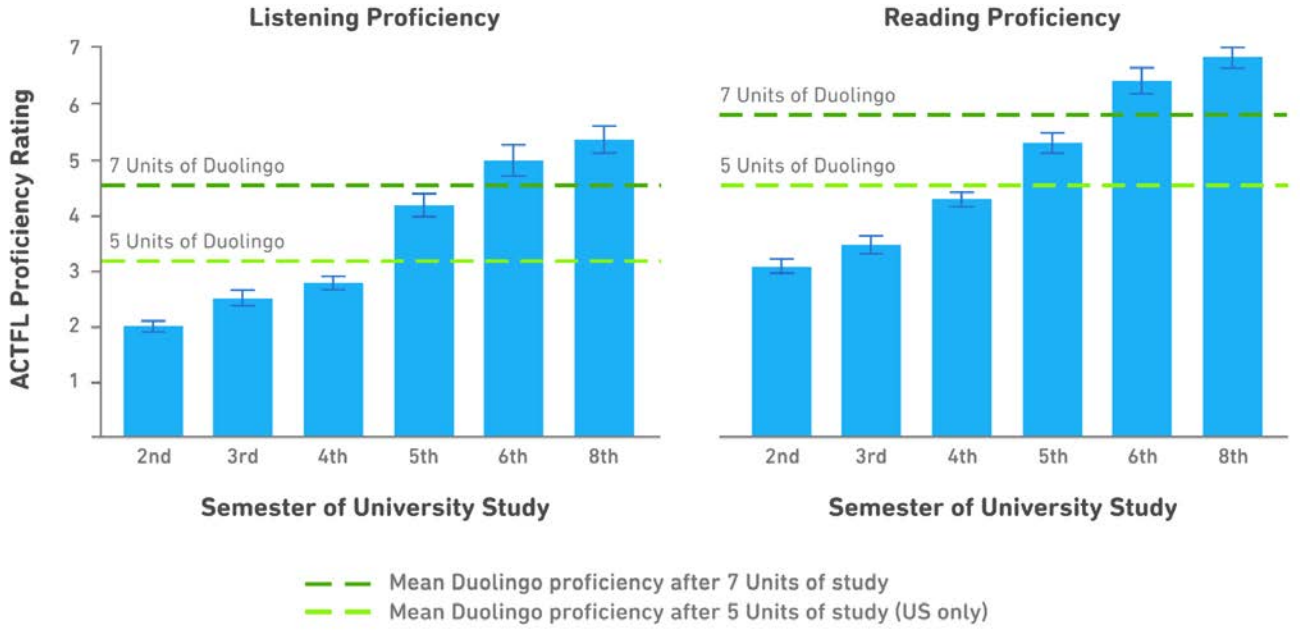
100+ Courses

190+ Countries

1 billion exercises completed each day

Duolingo is incredibly effective

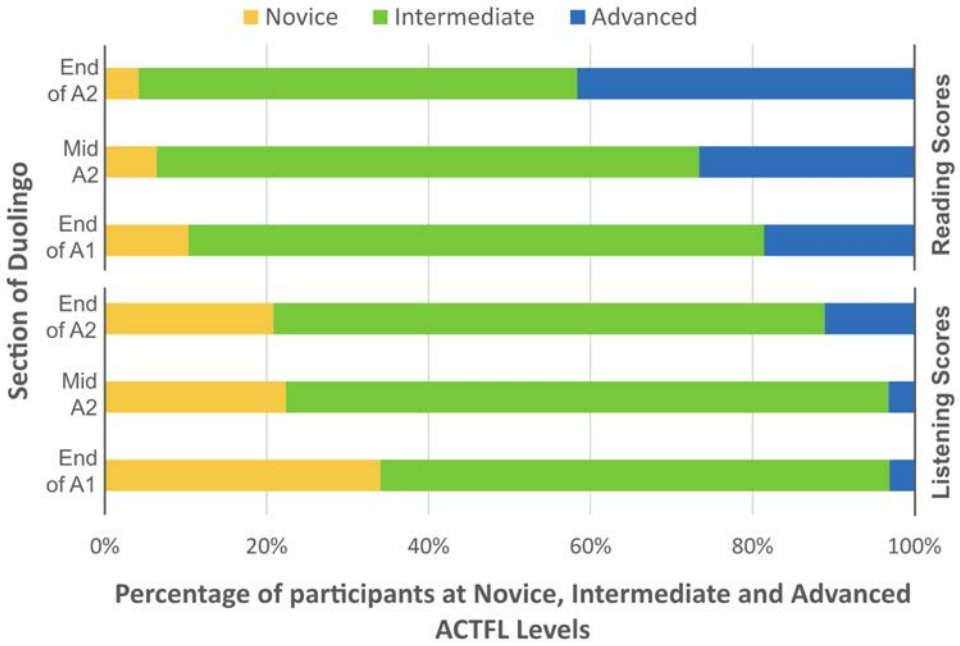
7 Units of Duolingo in French or Spanish = 5 semesters of university education in reading and listening



Sources: Duolingo Efficacy Study (November 2021), Duolingo Efficacy Study (August 2020)

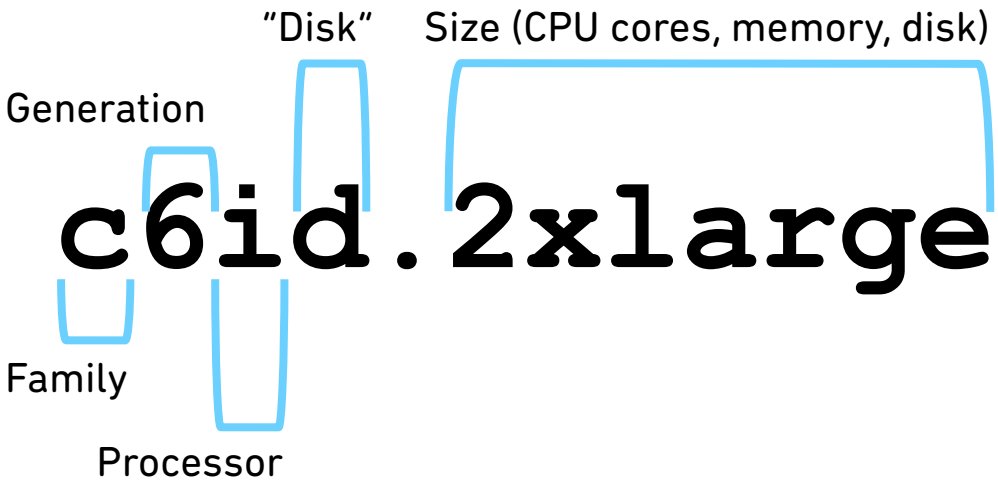
Duolingo is incredibly effective

Most beginner students in the English course for Spanish speakers were able to achieve intermediate reading and listening proficiency in English



Source: Duolingo Efficacy Study (August 2022)

What is an EC2 instance?



Intel Virtual Machine (VM) with 8 vCPUs,16 GiB of memory, and up to 12.5 Gbps of bandwidth

See https://instances.vantage.sh/ for all of the options

What is an EC2 instance?

`u-2t1b1.112xlarge`

448 vCPUs, 12288 GiB of memory, and 100 Gbps of bandwidth

See https://instances.vantage.sh/ for all of the options

Most common ways to purchase EC2 instances

- **On-Demand Instances**
  - Highest fixed price with no commitment
- **Reserved Instances (RIs)**
  - Up to 74% off on-demand price with a 1 or 3 year commitment
  - Can sell unused RIs via a marketplace for a fee
- **Spot Instances**
  - Comes from spare capacity on AWS
  - Up to 90% off on-demand price
  - Can shut down with only a 2 minute warning
- **Savings Plans**
  - Commit to a consistent amount of flexible usage
  - Up to a 72% discount with a 1 or 3 year commitment

The Spot Market

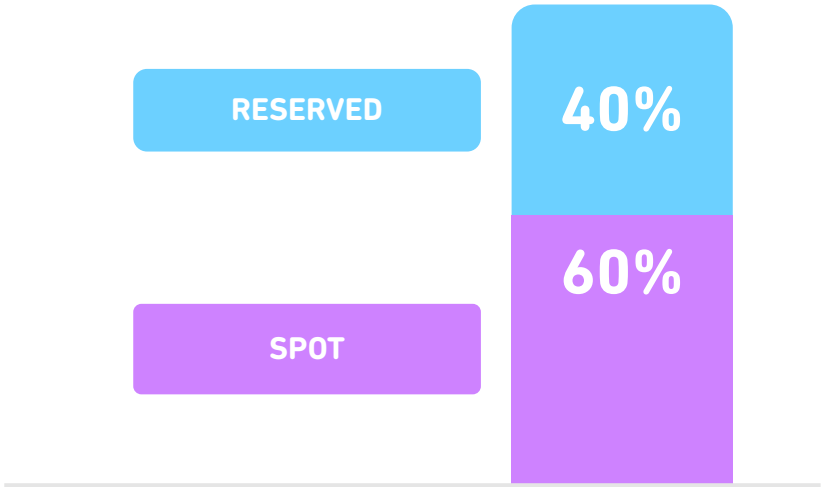
- Based on the supply/demand of AWS spare capacity
  - Bid the maximum that you are willing to pay
  - If the bid is above the market price, you are given an instance at the current market rate (not the bid price)
  - If the market price goes above the bid price while the instance is running, it is shut down after a 2 minute warning

- For practical reasons, the bid is generally set to the on-demand price

- Historical note: prior to 2018, prices could go above the on-demand rate and were subject to the whims of pure market dynamics! https://aws.amazon.com/blogs/compute/new-amazon-ec2-spot-pricing/
  - The market has been **relatively** stable since

Compute cluster mix

Duolingo ECS compute clusters typically run on a mix of Spot and Reserved instances, with on-demand instances being added as-needed when Spot capacity is low



Spot.io

Shifting to Graviton

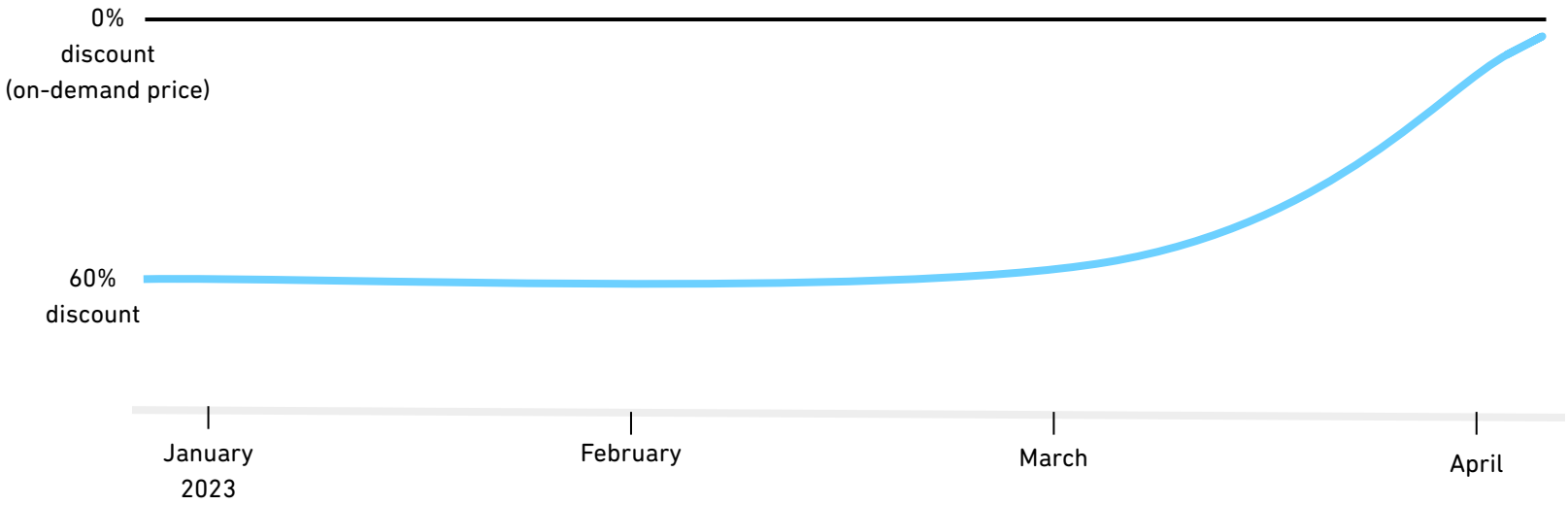
`c5d -> c6gd`

- We shifted our largest services to AWS Graviton2 (ARM-based) instances to get an additional 20% savings over our preferred Intel instances
- Graviton is similar the Apple M1 processor, but geared towards the cloud (very good price/performance)
- We ended up with even more savings than expected because there were more Graviton instances available on the Spot market at the time than Intel machines


duolingo
When Clouds Stop Raining Discounts

The crash

Increasing spot prices



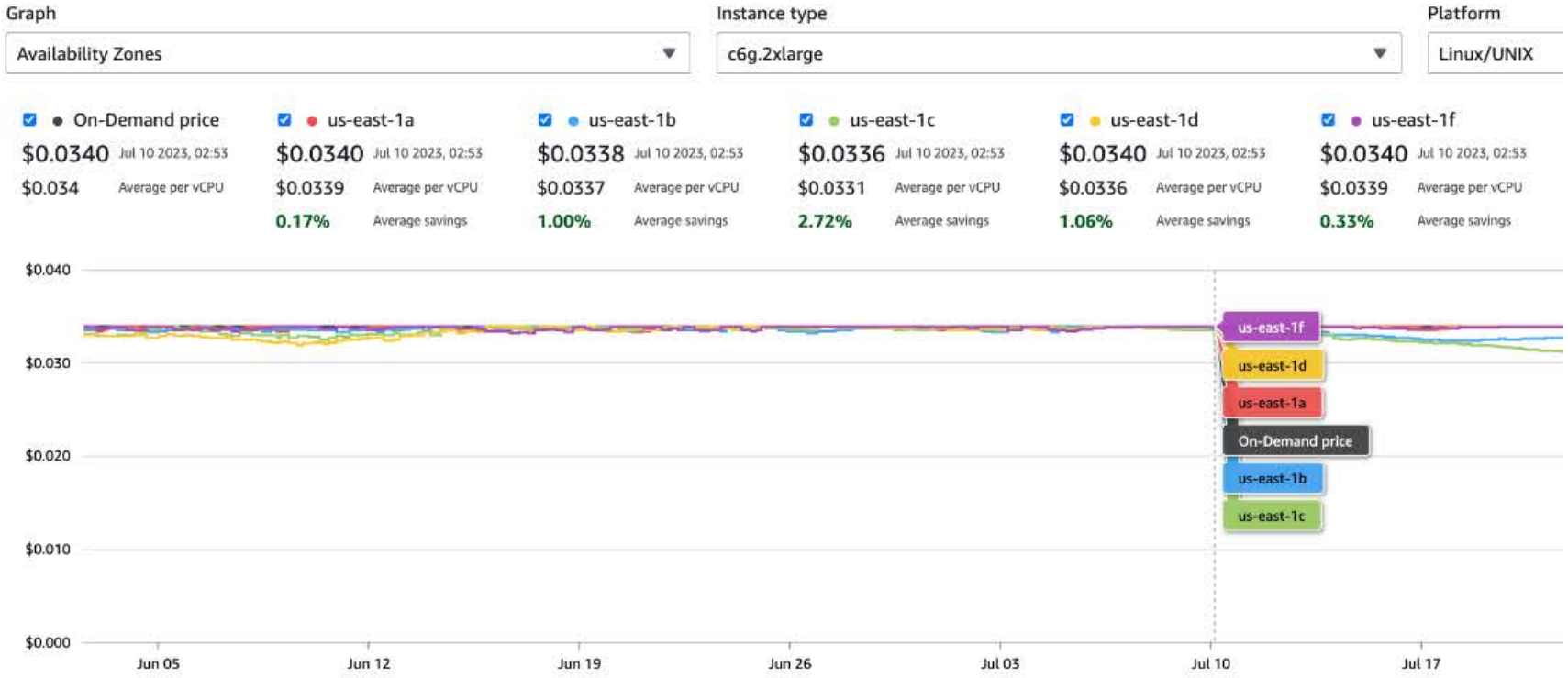
Who else noticed and when?

- Public tweet referencing the increase (March 23, 2023) https://twitter.com/jonathannorris/status/1640727214013530114
- Discussions in FinOps Community Slack (April, 2023) https://www.finops.org/community/community-slack/
- Farewell to the Era of Cheap EC2 Spot Instances (May 2, 2023) https://pauley.me/post/2023/spot-price-trends/
- Thoughts on the current state of EC2 Spot pricing (And what you can do about it) (May 10, 2023) https://leanercloud.beehiiv.com/p/thoughts-current-state-ec2-spot-pricing
- The Rise and Fall of Spot Instance pricing (August 3, 2023) https://cast.ai/blog/the-rise-and-fall-of-spot-instance-pricing/

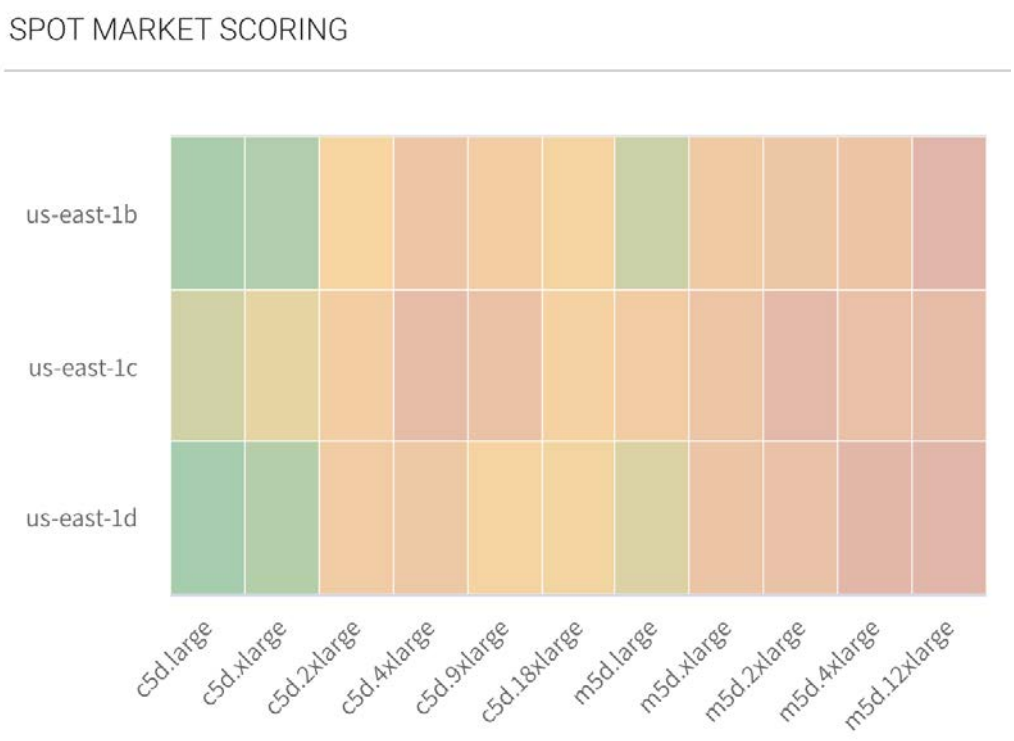
What happened?!

- Many companies became more cost-conscious and moved workloads to Spot and Graviton instances?
- " During Prime Day 2023, Amazon used tens of millions of normalized AWS Graviton-based Amazon EC2 instances, 2.7x more than in 2022" https://aws.amazon.com/blogs/aws/prime-day-2023-powered-by-aws-all-the-numbers/?
- Another large AWS customer bringing more capacity online?
- AWS itself running out of physical machines and/or becoming more efficient with spare capacity?

The current situation



The current situation



Data from Spot.io

GPT-4

https://blog.duolingo.com/duolingo-max/

Getting pricing information

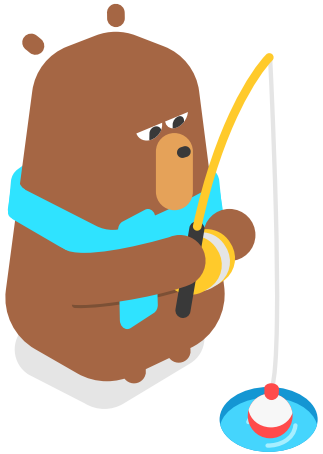
“Give me a bash script that lists all of the current on-demand and spot prices for all instance types in us-east-1”

```
#!/bin/bash

# Get the on-demand prices
echo "On-Demand Prices:"
aws ec2 describe-instance-types --query 'InstanceTypes[*].[InstanceType,
OnDemandPrice]' --output table --region us-east-1

# Get the spot prices
echo "Spot Prices:"
aws ec2 describe-spot-price-history --start-time=$(date -u +"%Y-%m-%dT%H:%M:
%SZ") --query 'SpotPriceHistory[*].[InstanceType, SpotPrice]' --output table
--region us-east-1
```

Getting pricing information

Not exactly what we want…but it's a starting point

```
On-Demand Prices:
---------------------------------
|    DescribeInstanceTypes      |
+--------------------+----------+
|  inf2.48xlarge     |  None    |
|  c4.4xlarge        |  None    |
|  m7i.16xlarge      |  None    |
…
Spot Prices:
-----------------------------------
|    DescribeSpotPriceHistory     |
+-------------------+-------------+
|  i3en.xlarge      |  0.195600   |
|  i3en.xlarge      |  0.195600   |
|  i3en.xlarge      |  0.195600   |
…
```



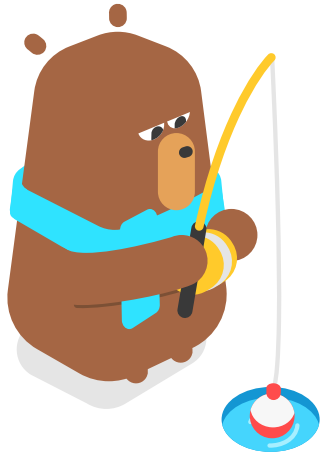
Getting pricing information

```
#!/bin/bash

# Set AWS_REGION variable
export AWS_REGION="us-east-1"

# Get the on-demand prices
echo "On-demand prices:"
aws pricing get-products --service-code "AmazonEC2" --filters
"Type=TERM_MATCH,Field=location,Value=US East (N. Virginia)"
"Type=TERM_MATCH,Field=operatingSystem,Value=Linux"
"Type=TERM_MATCH,Field=preInstalledSw,Value=NA" "Type=TERM_MATCH,Field=tenancy,Value=Shared"
--output json | jq '.PriceList[] | fromjson | .product.attributes.instanceType + ": " +
(.terms.OnDemand[].priceDimensions[].pricePerUnit.USD | tostring) + " USD"'

# Get the spot prices
echo "Spot prices:"
aws ec2 describe-spot-price-history --start-time=$(date +%Y-%m-%dT%H:%M:%S)
--product-descriptions="Linux/UNIX" --output json | jq '.SpotPriceHistory[] | .InstanceType
+ ": " + .SpotPrice + " USD"'
```

Getting pricing information

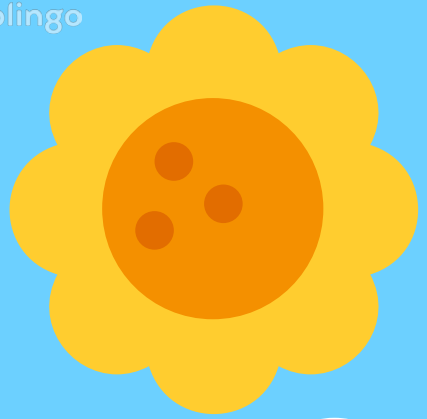
The information is correct this time and the formatting is much more useful

```
On-demand prices:
"m5d.12xlarge: 2.7120000000 USD"
"dl1.24xlarge: 0.0000000000 USD"
"r6g.medium: 0.0504000000 USD"
"r6g.12xlarge: 0.0000000000 USD"
…
Spot prices:
"x2idn.24xlarge: 3.596400 USD"
"c6in.xlarge: 0.128100 USD"
"c6in.16xlarge: 2.052200 USD"
"r5a.16xlarge: 2.051100 USD"
…
```



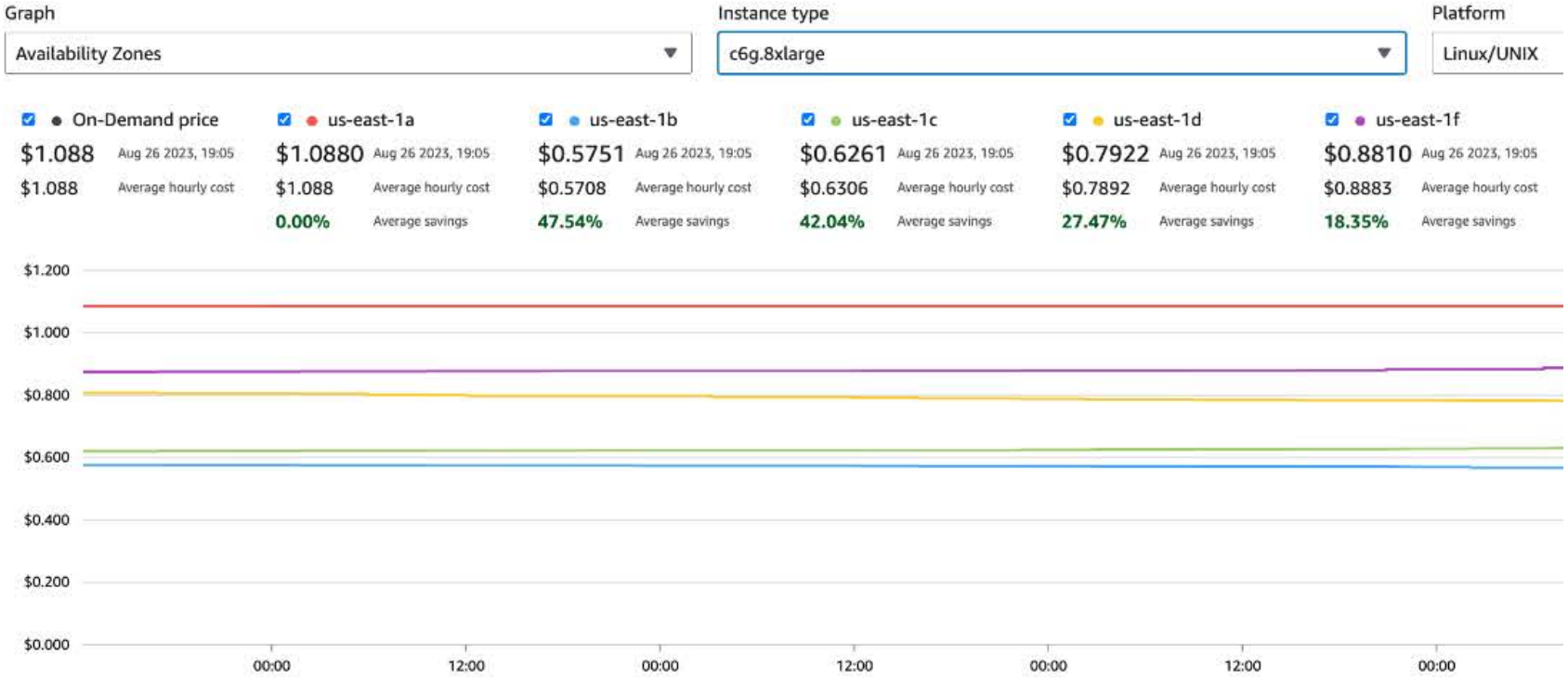
Lessons learned

- The GPT-4 model is non-deterministic – the same inputs will not generate the same outputs
- Generated script quality can vary widely – from good, usable code with useful comments, to completely unrunnable lines with made up parameters
- You must have domain knowledge in the area that is being generated in order to filter out the potential noise
- As a templating and autocomplete system it works pretty well – just don't throw out your technical documentation (yet)



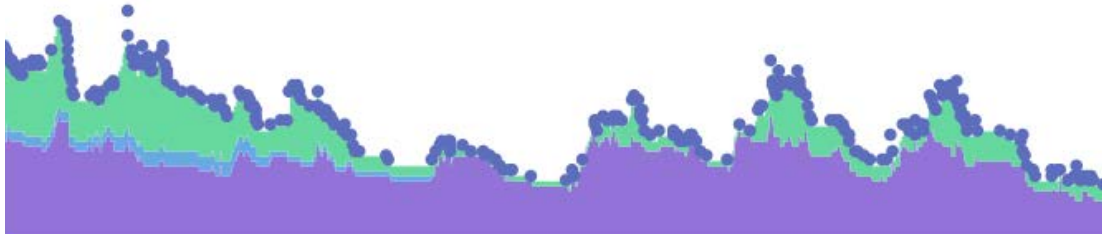
Surviving the Drought

Some glimmers of hope



Survival strategies

- If compute costs are known and relatively fixed, consider Savings Plans
- If compute costs are variable* consider “expanding” existing RIs
  - Reserved Instances can be modified to increase the number of covered instances while maintaining the end date
- Move to a range instance types/sizes with more discounts (diversify)
- Look for potential savings elsewhere to make up for the increase in compute costs (e.g. utilizing S3 storage tiers)



What about other clouds?

- From 2022 to 2023, **Azure** experienced an average **108% increase** in spot pricing compared to an average **21% increase in AWS**
- **GCP** experienced a **26% decrease** in spot prices in the same time period

https://cast.ai/blog/the-rise-and-fall-of-spot-instance-pricing/

gracias!