MASSACHUSETTS GENERAL HOSPITAL

HARVARD MEDICAL SCHOOL

BROAD INSTITUTE

GINGER On-Site Training
Day 1: Refresher

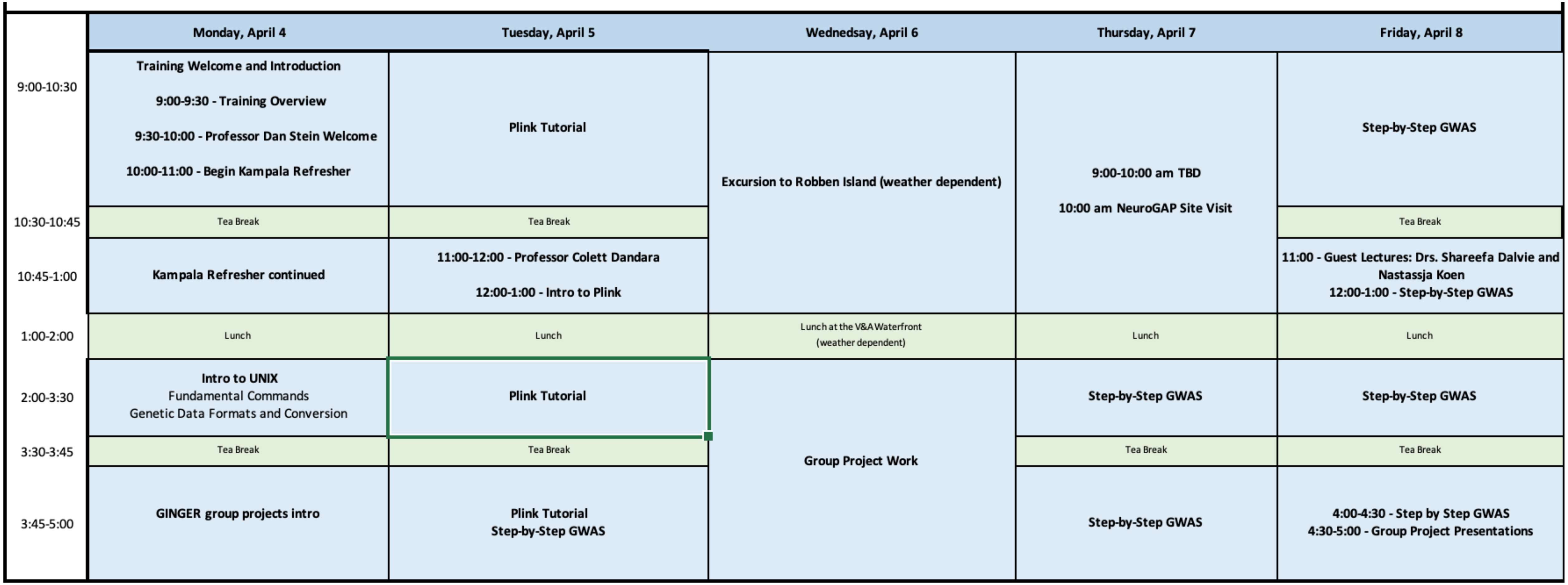
GINGER Program 2022

University of Cape Town

Teaching Fellows:

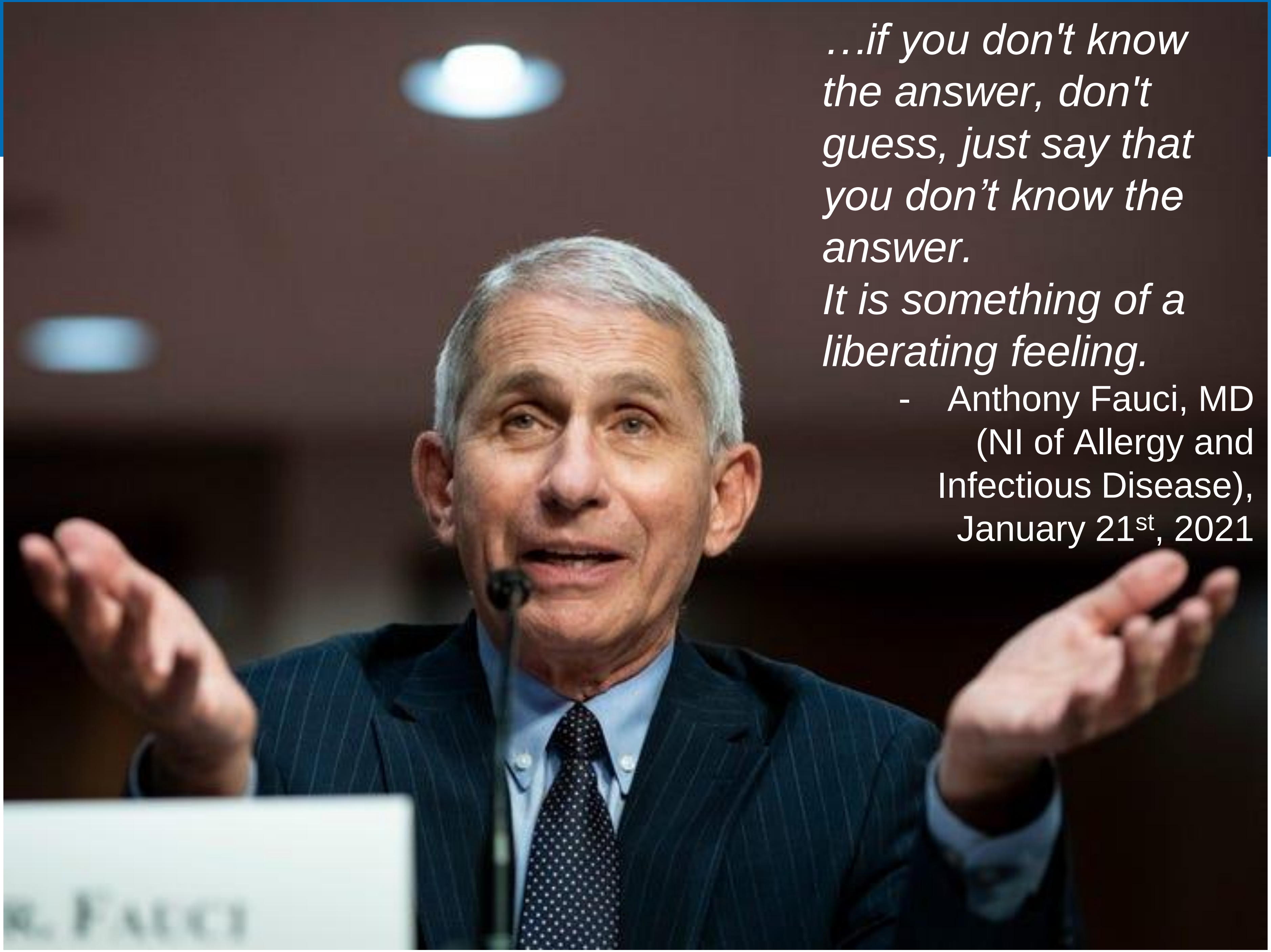
Kumar Veerapen, Ph.D.
Senior Expert I Data Science
Kumar.Veerapen@Novartis.com

Carla Marquez-Luna, Ph.D.
Postdoctoral Research Fellow
carlamarquezluna@gmail.com



	Monday, April 4	Tuesday, April 5	Wednedsay, April 6	Thursday, April 7	Friday, April 8
9:00-10:30	**Training Welcome and Introduction** **9:00-9:30 - Training Overview** **9:30-10:00 - Professor Dan Stein Welcome** **10:00-11:00 - Begin Kampala Refresher**	**Plink Tutorial**	**Excursion to Robben Island (weather dependent)**	**9:00-10:00 am TBD** **10:00 am NeuroGAP Site Visit**	**Step-by-Step GWAS**
10:30-10:45	Tea Break	Tea Break			Tea Break
10:45-1:00	**Kampala Refresher continued**	**11:00-12:00 - Professor Colett Dandara** **12:00-1:00 - Intro to Plink**			**11:00 - Guest Lectures: Drs. Shareefa Dalvie and Nastassja Koen** **12:00-1:00 - Step-by-Step GWAS**
1:00-2:00	Lunch	Lunch	Lunch at the V&A Waterfront (weather dependent)	Lunch	Lunch
2:00-3:30	**Intro to UNIX** Fundamental Commands Genetic Data Formats and Conversion	**Plink Tutorial**	**Group Project Work**	**Step-by-Step GWAS**	**Step-by-Step GWAS**
3:30-3:45	Tea Break	Tea Break		Tea Break	Tea Break
3:45-5:00	**GINGER group projects intro**	**Plink Tutorial** **Step-by-Step GWAS**		**Step-by-Step GWAS**	**4:00-4:30 - Step by Step GWAS** **4:30-5:00 - Group Project Presentations**

GWAS all the time*



…if you don't know the answer, don't guess, just say that you don’t know the answer.
It is something of a liberating feeling.

- Anthony Fauci, MD (NI of Allergy and Infectious Disease), January 21st, 2021

Outline

- Cloud Computing
- Rstudio on the Cloud

But first… Let’s reintroduce ourselves

You have 30 seconds

1) Name
2) Where are you from?
3) What is your genetics training?
4) What do you want from this week in UCT?

Cloud Computing Refresher

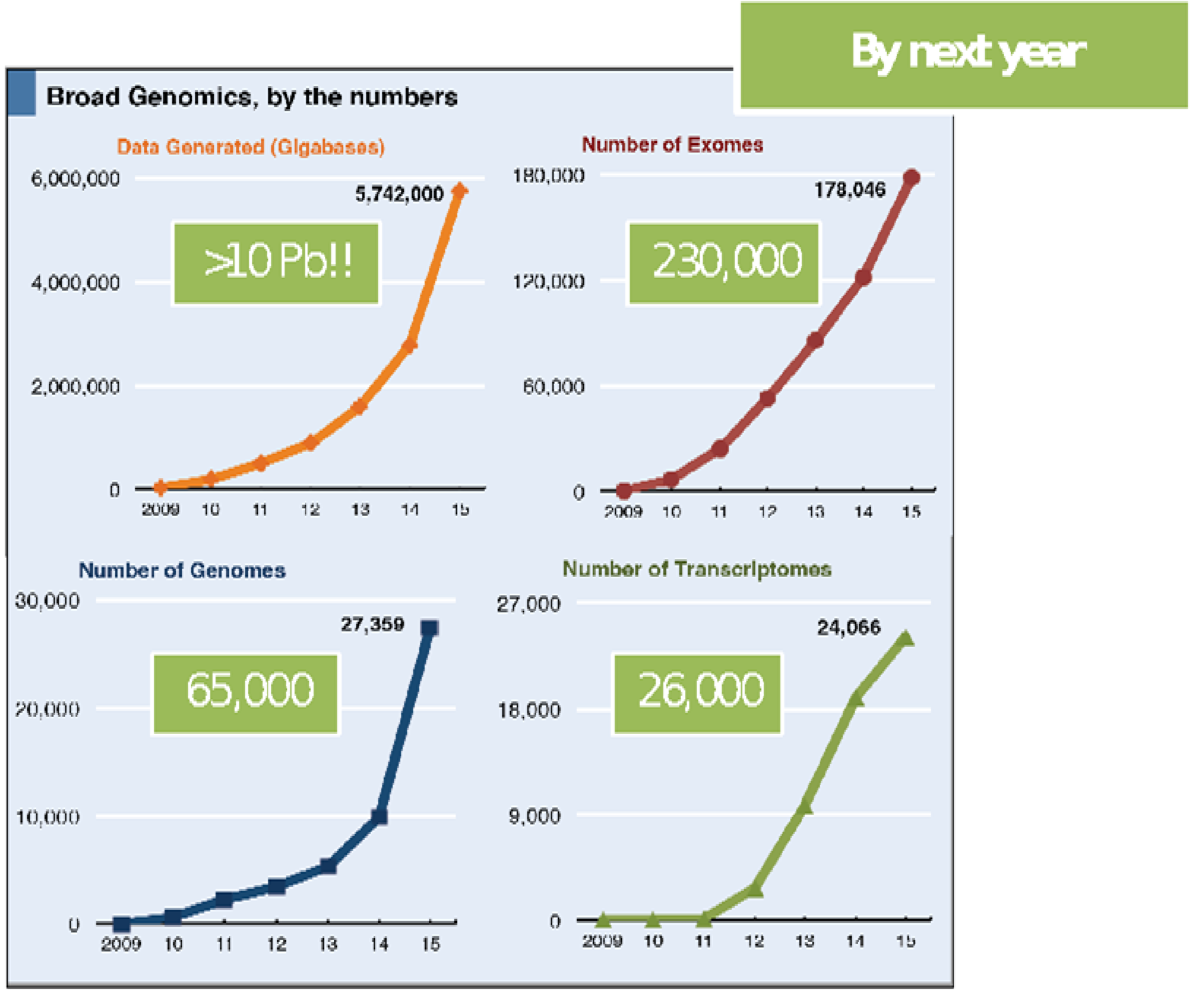
Learning objectives:

- Be able to set up virtual machines
- Be able to spin up Rstudio instance for analytical purposes

Adapted from Konrad Karczewski

But… Y’all have used this.
Why do ***you*** think we should
switch to cloud computing
resources for analytics?

Technological Growth in Genetics and Genomics



By next year

>10 Pb!!

230,000

65,000

26,000

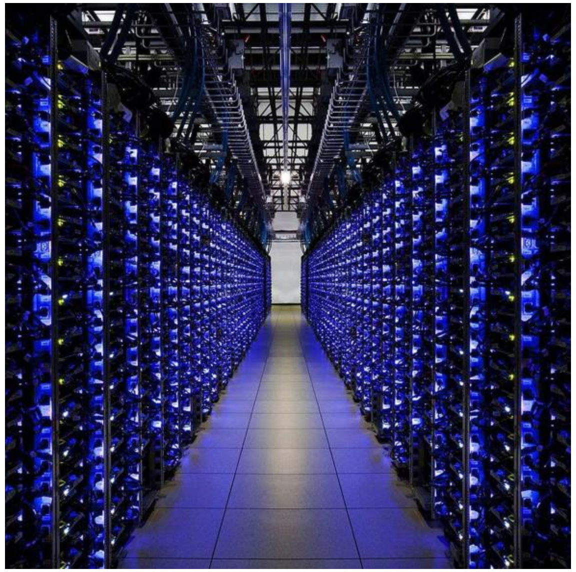
Generally speaking…!

- Cloud computing: storing and accessing data and programs over the Internet instead of your computer's hard drive

- Why is it "the cloud"?: A metaphor for the internet
  - Goes back to presentations showing connections going in and out of a cloud

Local machine

Server

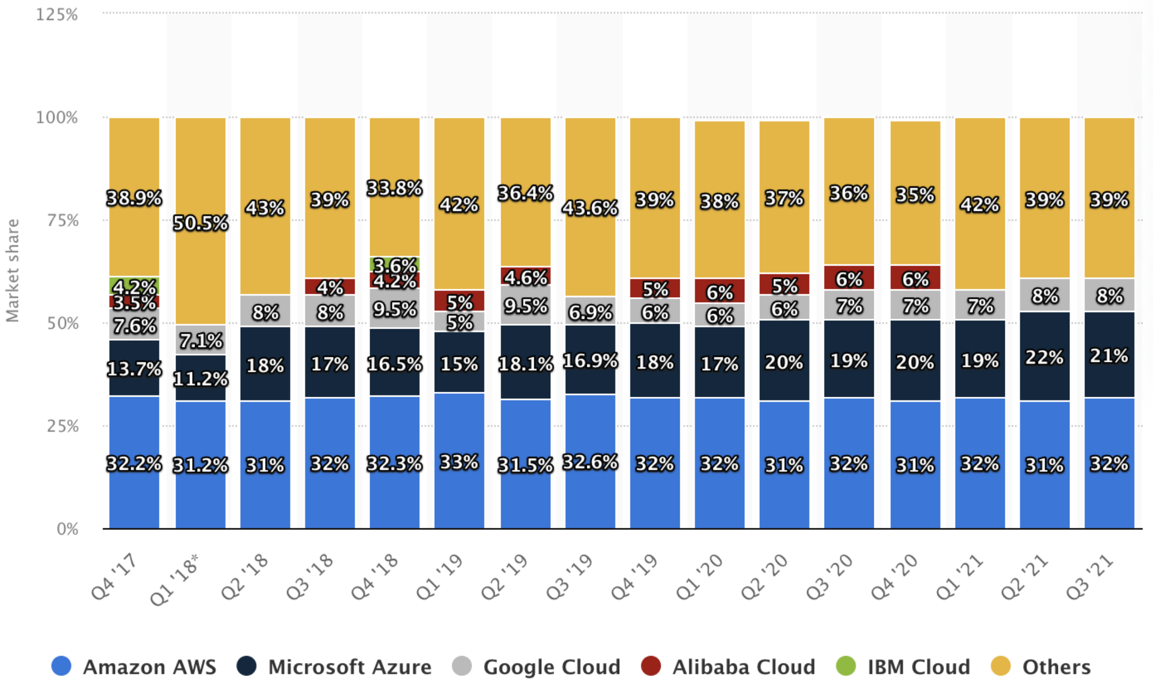


Cloud (remote servers)

What do we know of this now?

- **Internet-based computing services**
  - *Remotely hosted*: data and compute are hosted on remote infrastructure
  - *Commodified*: pay-as-you-go, like other utilities (e.g. electricity)
    - What happens if you don’t use it?
  - Tends to be cheaper
    - Broad Prem cluster $75/TB/month vs $25/TB/month
    - Cheaper if cold storage

INSERT HAIL CLOUD SLIDE

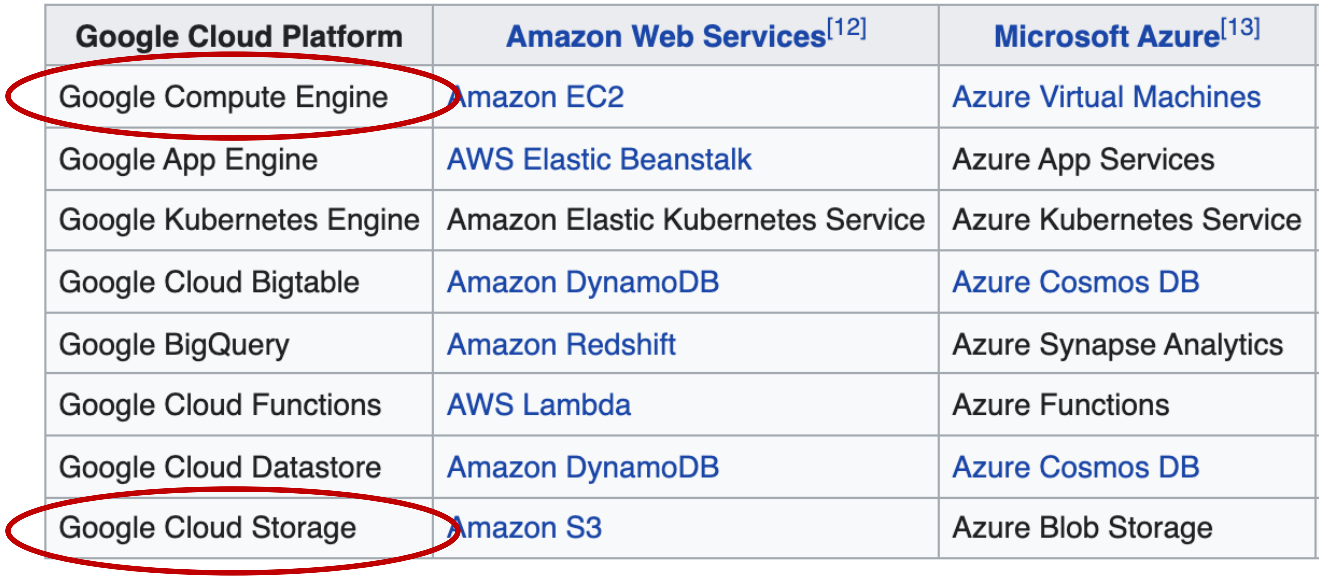


- Most offer the same set of products with slightly different configurations, pros/cons

We use Google Cloud Platform (GCP)

- Hundreds of products (it can be *overwhelming*!)
  - All of these are "the cloud" – you may talk to someone about the cloud and they may be referring to an entirely different ecosystem

- Most fundamental ones (most "infrastructure"-like)
  - **Google Compute Engine**
  - **Google Cloud Storage**



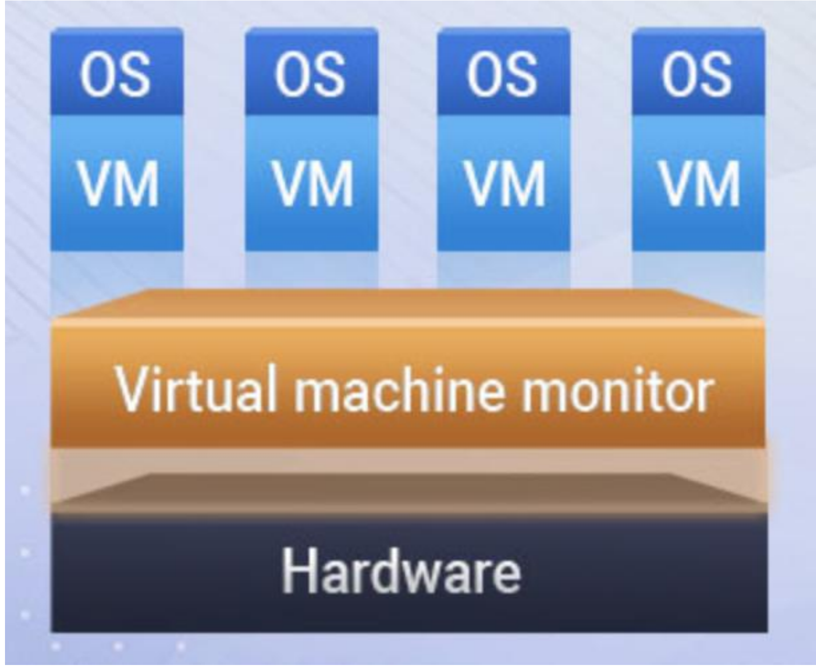
Google Cloud Platform	Amazon Web Services[12]	Microsoft Azure[13]
Google Compute Engine	Amazon EC2	Azure Virtual Machines
Google App Engine	AWS Elastic Beanstalk	Azure App Services
Google Kubernetes Engine	Amazon Elastic Kubernetes Service	Azure Kubernetes Service
Google Cloud Bigtable	Amazon DynamoDB	Azure Cosmos DB
Google BigQuery	Amazon Redshift	Azure Synapse Analytics
Google Cloud Functions	AWS Lambda	Azure Functions
Google Cloud Datastore	Amazon DynamoDB	Azure Cosmos DB
Google Cloud Storage	Amazon S3	Azure Blob Storage

Google Compute Engine a.k.a Virtual Machine

- A.k.a. computers
  - CPU,
  - some memory,
  - some disk space, and
  - an operating system (OS)

- You get a VM from a pool of machines

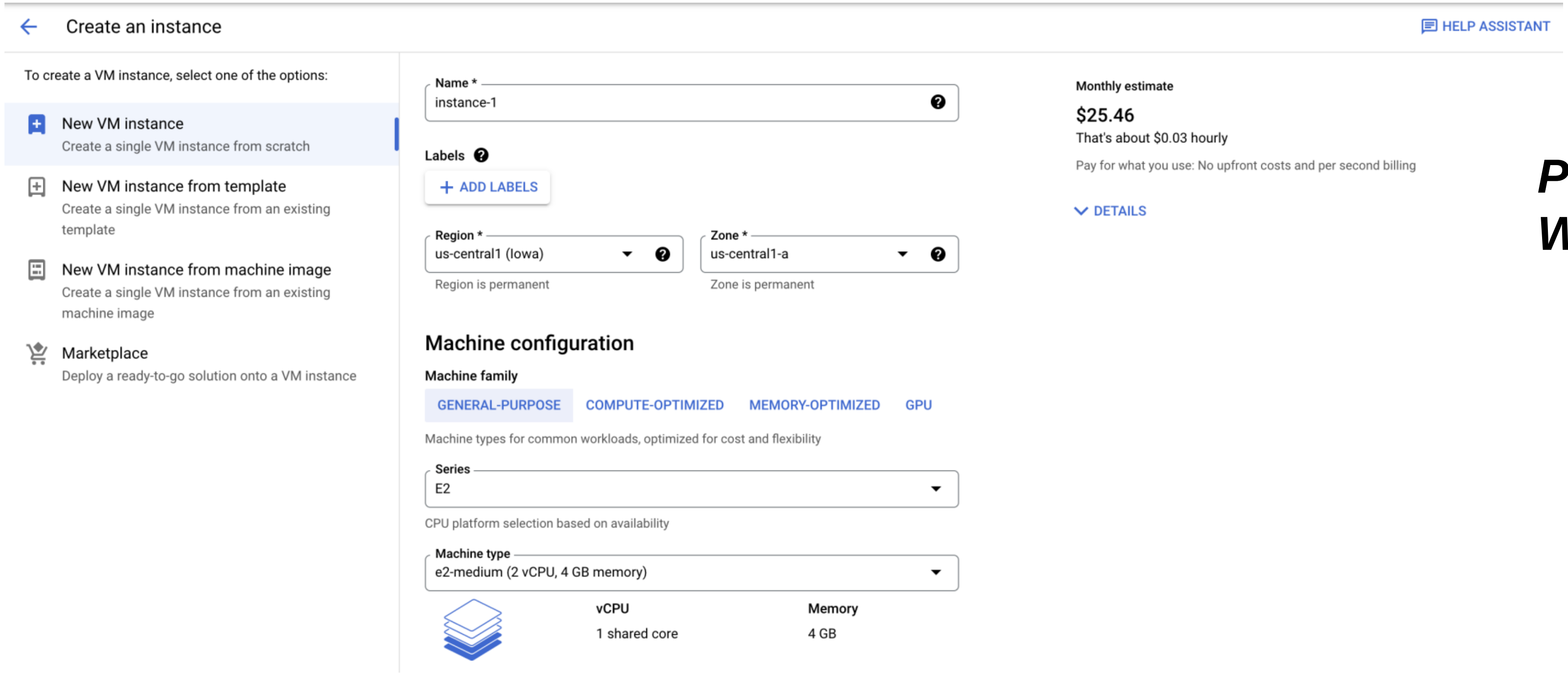
- When you delete your VM, they go back into the pool



Google Compute Engine a.k.a. Virtual Machine

- Fully customizable (e.g. number of CPUs, memory, disk space)

- Rule of thumb: a CPU costs about $0.04 or $1 per day
    - There is a way to get this down to $0.01/hour (pre-emptible)



To create a VM instance, select one of the options:

New VM instance
Create a single VM instance from scratch

New VM instance from template
Create a single VM instance from an existing template

New VM instance from machine image
Create a single VM instance from an existing machine image

Marketplace
Deploy a ready-to-go solution onto a VM instance

Name *: instance-1

Labels

Region *: us-central1 (Iowa)
Region is permanent

Zone *: us-central1-a
Zone is permanent

Machine configuration

Machine family: GENERAL-PURPOSE

Machine types for common workloads, optimized for cost and flexibility

Series: E2
CPU platform selection based on availability

Machine type: e2-medium (2 vCPU, 4 GB memory)

vCPU	Memory
1 shared core	4 GB

Monthly estimate
$25.46
That's about $0.03 hourly
Pay for what you use: No upfront costs and per second billing

Preemptible vs full node?
Why the cost differential?

Pricing: https://gcpinstances.doit-intl.com/

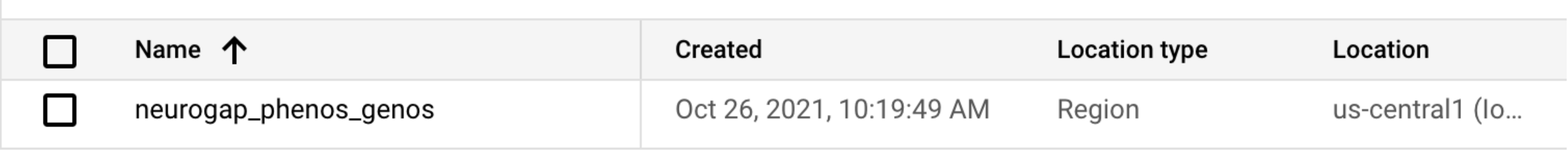
Google Cloud Storage

- Files are stored on a distributed system (”object store”) rather than a traditional file system (*I found out the hard way*)

- Different storage types
    - Multi-regional (accessible quickly in multiple regions) – most expensive
    - Regional (normal)
    - Nearline, Coldline, Archive: cheaper per month, but costs $ to access

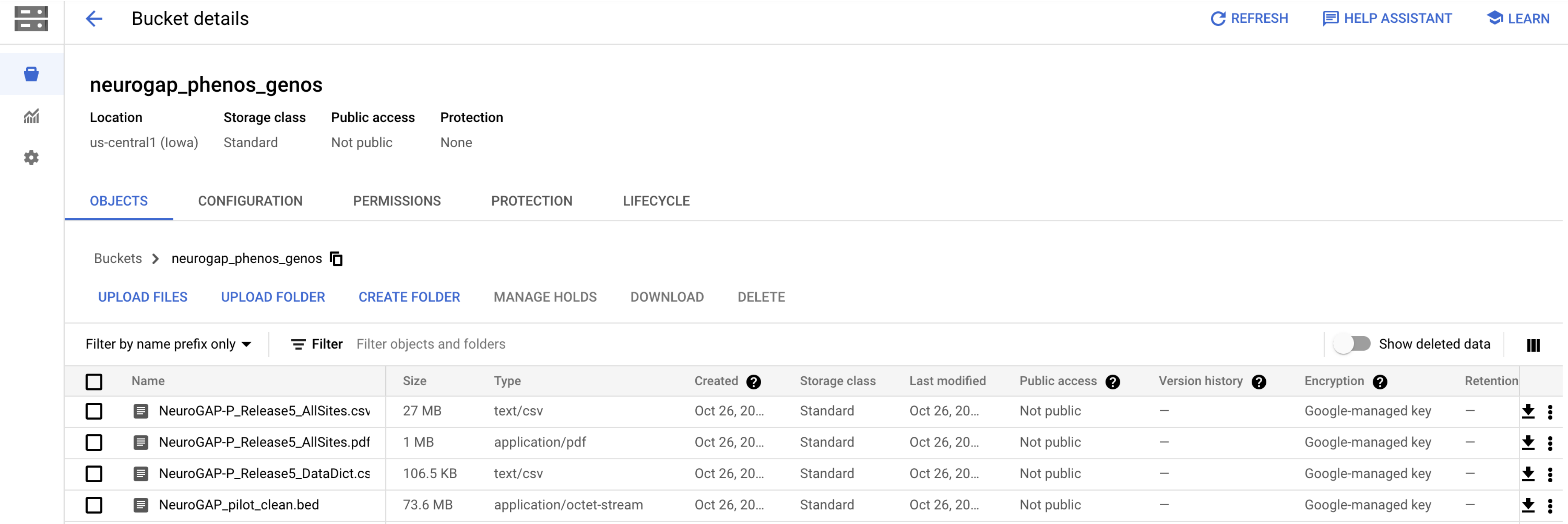
Google Cloud Storage

- Bucket list:



□	Name ↑	Created	Location type	Location
□	neurogap_phenos_genos	Oct 26, 2021, 10:19:49 AM	Region	us-central1 (lo…

- Files within buckets:



neurogap_phenos_genos

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	None

Buckets > neurogap_phenos_genos

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention
□	NeuroGAP-P_Release5_AllSites.csv	27 MB	text/csv	Oct 26, 20…	Standard	Oct 26, 20…	Not public	–	Google-managed key	–
□	NeuroGAP-P_Release5_AllSites.pdf	1 MB	application/pdf	Oct 26, 20…	Standard	Oct 26, 20…	Not public	–	Google-managed key	–
□	NeuroGAP-P_Release5_DataDict.cs	106.5 KB	text/csv	Oct 26, 20…	Standard	Oct 26, 20…	Not public	–	Google-managed key	–
□	NeuroGAP_pilot_clean.bed	73.6 MB	application/octet-stream	Oct 26, 20…	Standard	Oct 26, 20…	Not public	–	Google-managed key	–

How do you pay for the cloud?

- Compute engine:
  - $0.04/CPU-hour
  - Very easy to rack up large bills
    - 1,000 CPUs running for a week = $7,000
    - 10 CPUs running for a year = $3,500

- Storage:
  - $0.02/GB/month
  - Harder to rack up large bills at these prices, but still possible with huge datasets
  - $25/TB/month

The cloud is awesome (Karczewski, 2017)

- Pay for the hardware you use
  - Be careful about spending!
  - Stop/delete your VMs, and be careful of storage!

- Scale analyses to thousands of machines

- Easier reproducible and shareable workflows

Word of Advice

Evernote

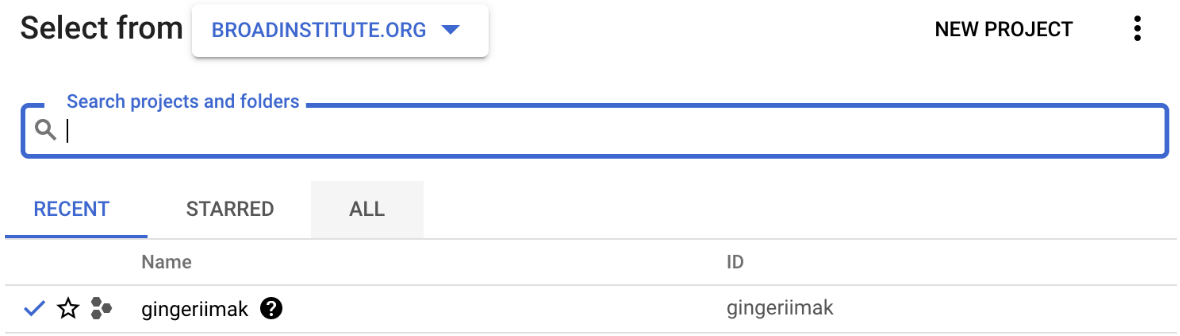
Tasks for today’s refresher. You do, we help!: GCP

- Pair up! Preferably someone who is computationally savvy with someone who feels less savvy
- Link to cloud console: console.cloud.google.com

1. Check your project that it is listed as `gingeriimak`
2. Create your own VM

   Name: [your first name]
   Zone: us-central1-b
   In Boot disk, please change:
       Operating system: Ubuntu
       Version (if not already selected): Ubuntu 18.04 LTS

   Access scopes: Allow full access to all cloud APIs



	Name	ID
✓ ☆	gingeriimak	gingeriimak

3. What’s in your current cloud storage on the API? Clue: https://github.com/atgu/GINGER_cloud/blob/main/Console.md

How do you code?

- There are many ways to code.
- Beginners, don’t worry about how short your code is. As long as IT WORKS!
- Later, get someone who is more experienced to do code review with you.

Tasks for today’s refresher. You do, we help! : RStudio

1. Create an Rstudio instance. You remember how to do this right…?

https://github.com/atgu/GINGER_cloud/blob/main/RStudio.md

1. Get your data. Using your Rstudio instance:
   - Click Terminal
   - Here we have our gsutil program already installed and it is already set up with permissions
   - Let's see what files we can use:
     - `gsutil ls`
   - Let's grab the phenotype files:
     - `gsutil cp gs://neurogap_phenos_genos/NeuroGAP-P_Release5_*.csv .`

Rstudio Exercise

But first:

Load in R library tidyverse and the NeuroGAP dataset we have been using
You may or may not need tidyverse but you will need the data. Obvi

```
~~library(tidyverse)~~
data = read.csv('NeuroGAP-P_Release5_AllSites.csv')
theme_set(theme_classic())
```

Rstudio Exercise Items

1. What were the top consent languages by country?

2. What is the proportion of cases by language + country

3. Number of HIV+ patients

4. Number of missing data from the HIV+ column
   1. By country, what's the distribution and plot out a histogram

There are many ways to code. You do you, boo!

GINGER On-Site Training
Day 1: Refresher
QUESTIONS? ☺

GINGER Program 2022

University of Cape Town

Teaching Fellows:

Kumar Veerapen, Ph.D.
Senior Expert I Data Science
Kumar.Veerapen@Novartis.com

Carla Marquez-Luna, Ph.D.
Postdoctoral Research Fellow
carlamarquezluna@gmail.com