NGS Data Management

Introduction to Next Generation Sequencing

2022 Workshop

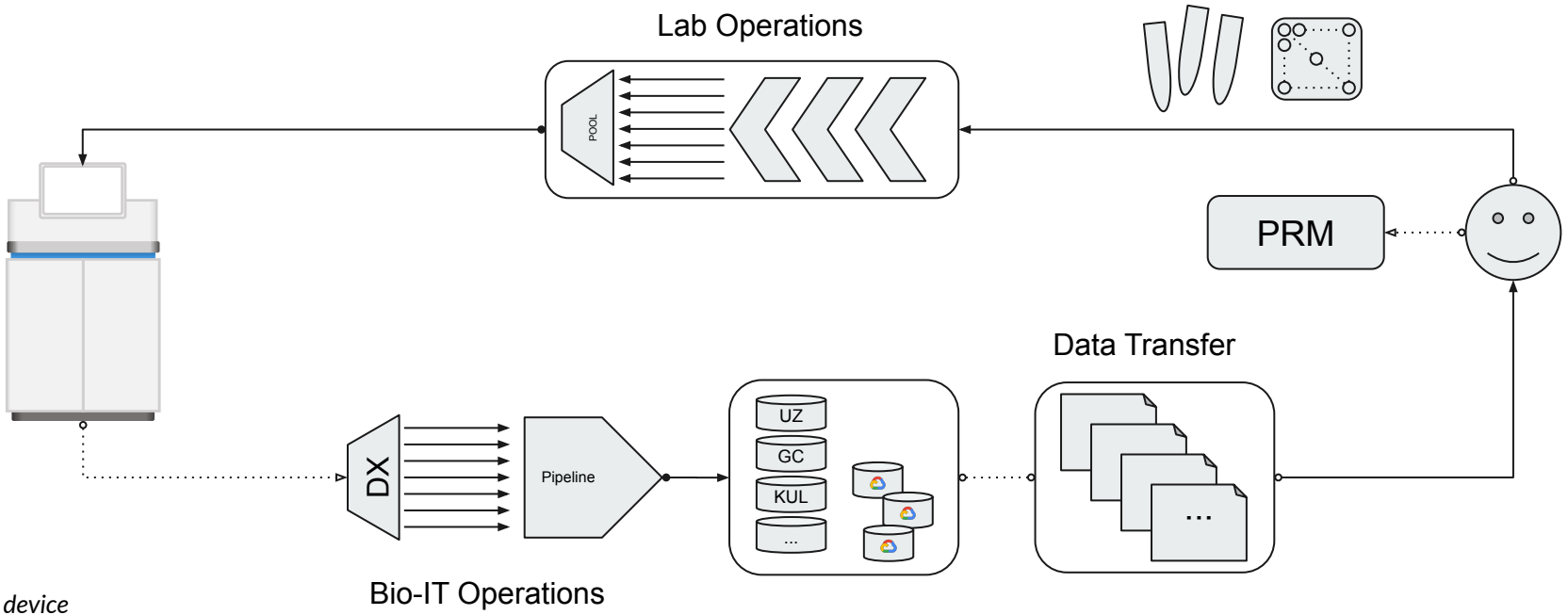
Karel De Schepper

Overview

NGS Data Management

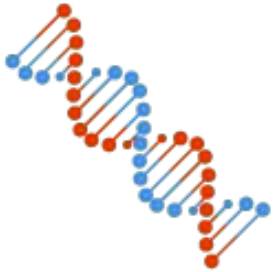
1. Generating Data
    a. Data Flow
    b. Typical Datasets
    c. Some Numbers
2. Delivering Data - Tier Based
    a. Silver Tier
    b. Gold Tier
3. Dos and Don'ts with Data

Data flow*



(*) illumina device workflow

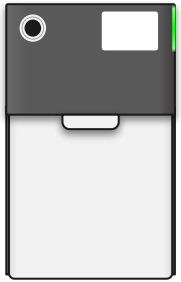
In other words…

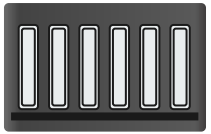


NGS

Synthetic DNA »

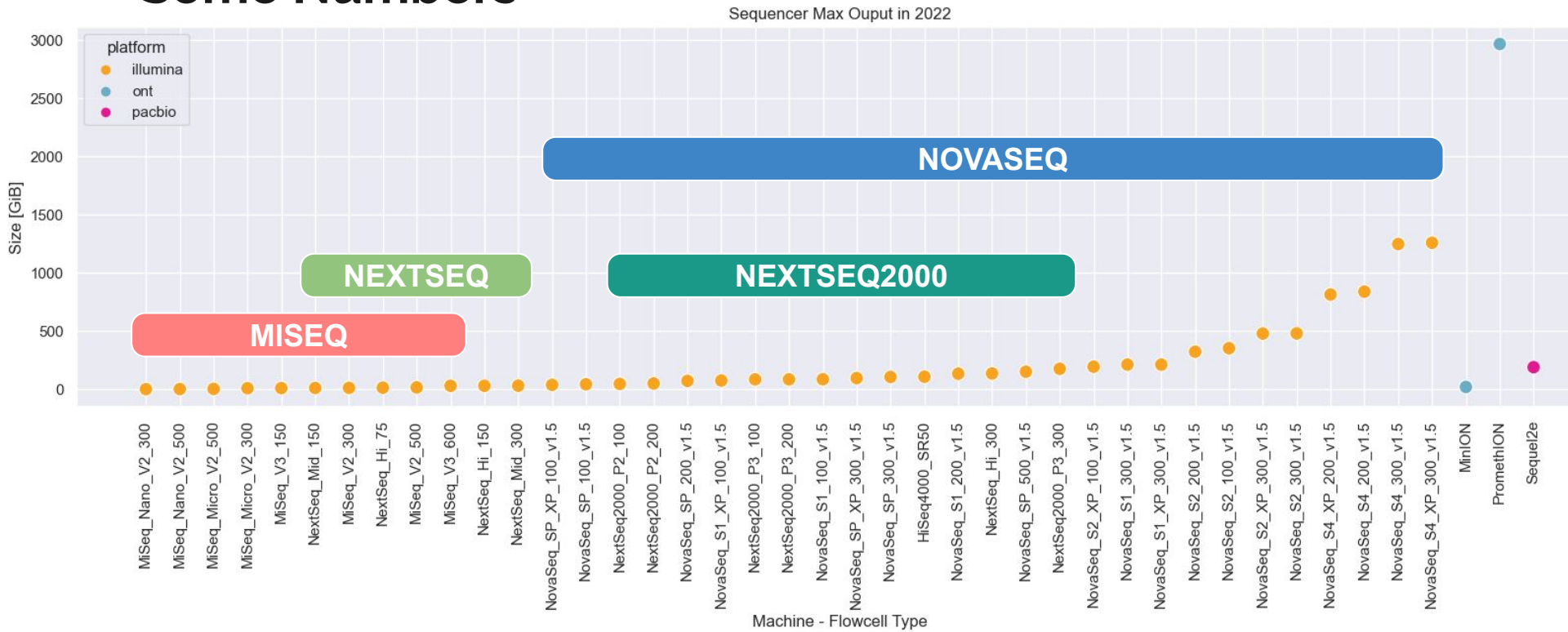
Typical Datasets



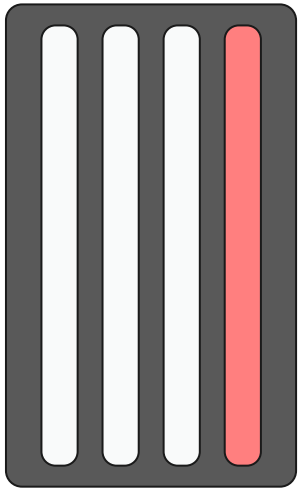


Device Output	BCL	CCS Reads	FAST5 / FASTQ
Demultiplexed	FASTQ		
Pipeline	Pipeline output	Pipeline output	Pipeline output

Some Numbers



Some Numbers: Example



NovaSeq S4
2 x 150 bp »

Device Output	Demultiplexed	Pipeline Output
BCL	FASTQ	Pipeline output
2 TB	2 TB / 4 lanes ~= 500 GB	?

You Request DATA

We Generate & Deliver DATA

Overview

NGS Data Management

1. Generating Data
    a. Data Flow
    b. Typical Datasets
    c. Some Numbers
2. **Delivering Data - Tier Based**
    a. **Silver Tier**
    b. **Gold Tier**
3. Dos and Don'ts with Data

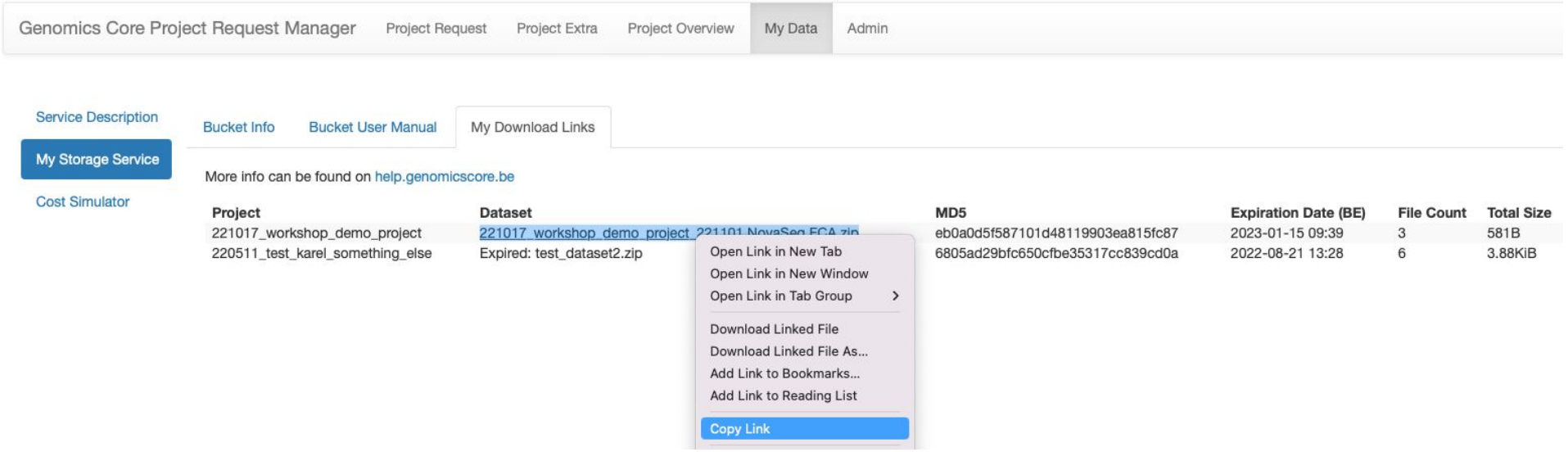
Data Delivery: A Tier Based Service

TIER	Silver (default)	Gold

Data Delivery: Silver Tier

TIER	Silver (default)	Gold
DELIVERY	Download link	
ACCESS	Through PRM	
POLICY	Valid for 90 days	
PRICING	Included in Project	

Silver Tier: Where?



Silver Tier: How?

> **Warning**
>
> Always ensure you understand the commands you execute to avoid unintended behaviour.

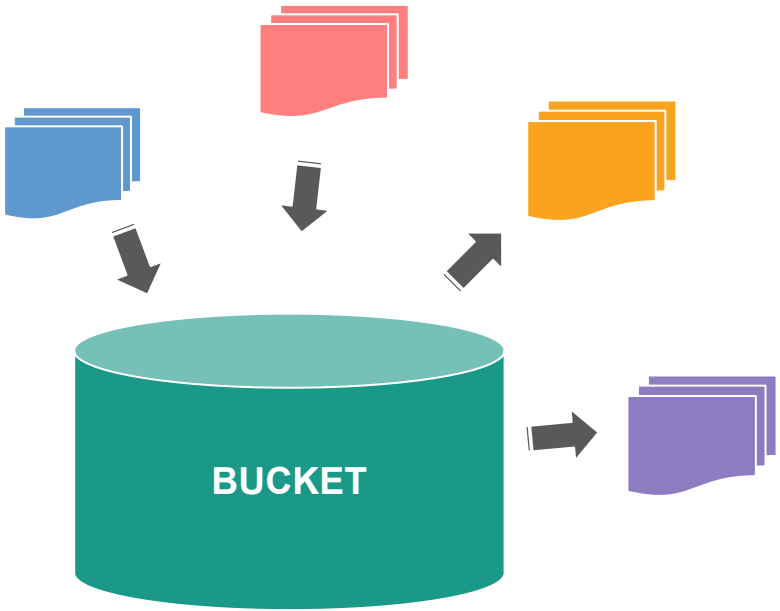
macOS | Linux | Windows

```
# Update package sources list
sudo apt update

# Install curl
sudo apt -y install curl

# Install zip and unzip
sudo apt install zip unzip

# Downloading
curl -o {/My/Location/some_file.zip} '{DOWNLOAD_LINK}'

# Verify MD5 hash to ensure no corruption occured
md5sum {/My/Location/some_file.zip}

# Extracting
unzip {/My/Location/some_file.zip}
```

Data Delivery: Gold Tier

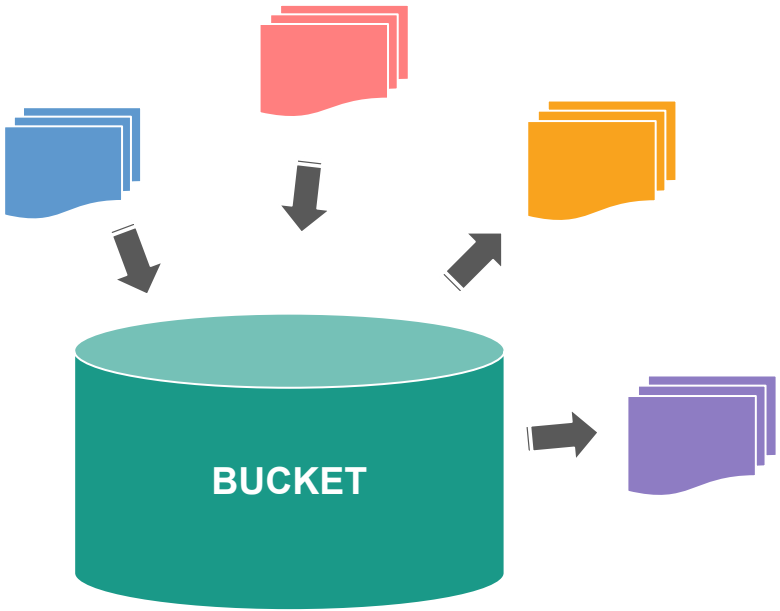
TIER	Silver (default)	Gold
DELIVERY	Download link	Google Cloud Bucket
ACCESS	Through PRM	
POLICY	Valid for 90 days	
PRICING	Included in Project	

Gold Tier: What?



- Google **Cloud**
- Virtual data "container"
- A dedicated bucket per pi
- Under GC's administration
- Access through share accounts (@share.uzleuven.be + 2FA enforced)
- Archiving capabilities - Set through PRM
  - STANDARD (HOT)
  - NEARLINE
  - COLDLINE
  - ARCHIVE (COLD)

Gold Tier: Why?

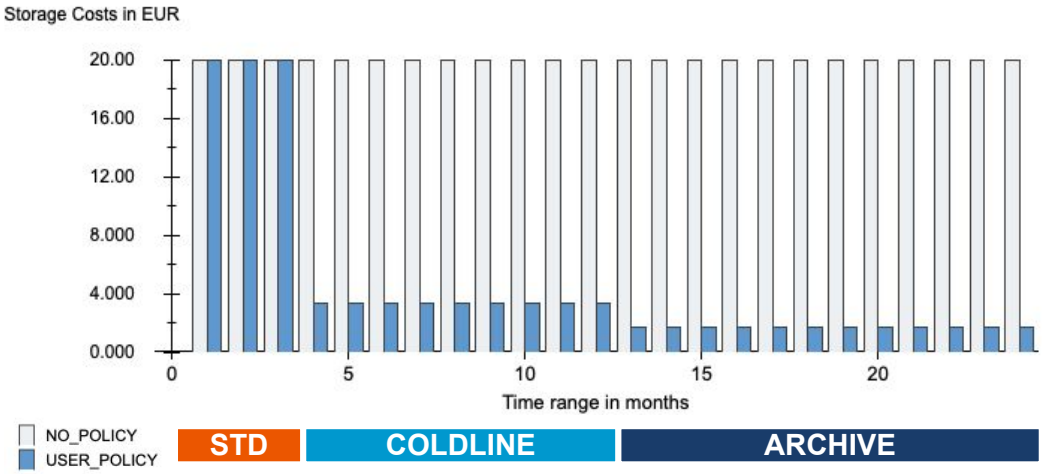


- Virtually unlimited in size
  - Cost is the limiting factor!
  - Reduction in storage admin
  - Avoids local infrastructure admin
- Increased flexibility
  - Partial downloads
  - Uploads
  - Cloud based analyses
  - Faster transfers

Gold Tier: Pricing & Policy

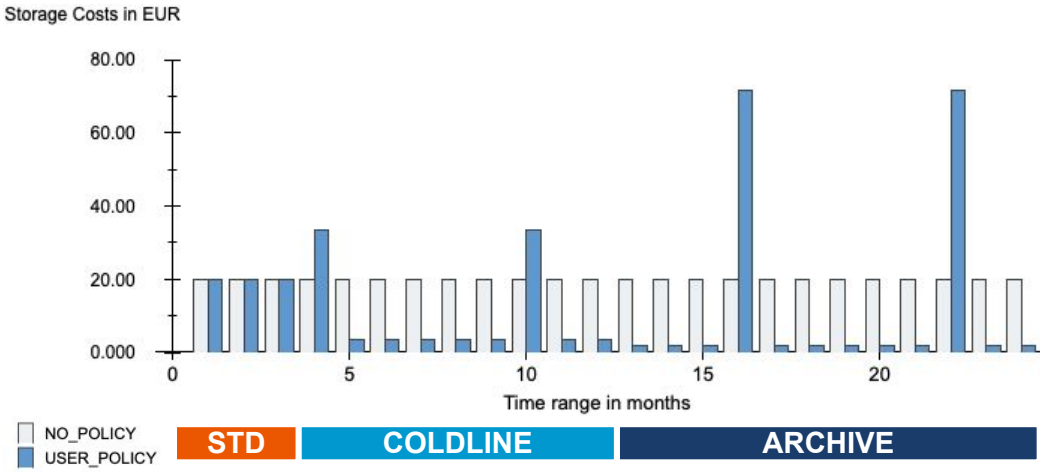
Prices may differ slightly due to precision,
, GB <> GiB (base 10 vs base 2)
and excludes VAT

STORAGE CLASS	STORAGE COST EUR/GB.month	RETRIEVAL COST EUR/GB	DOWNLOAD COST EUR/GB	MINIMUM RETENTION DAYS
STANDARD (STD)	0.02	0	0.13	0
NEARLINE (NL)	0.01	0.01	0.13	30
COLDLINE (CL)	0.0034	0.03	0.13	90
ARCHIVE (ARC)	0.0017	0.07	0.13	365



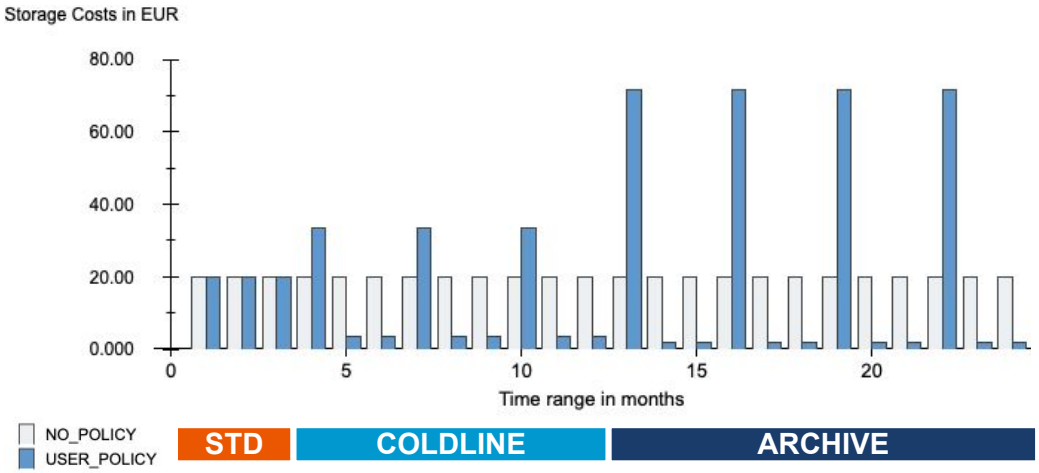
- **Size:** 1 TB
- **Time:** 2 years

		STORAGE	RETRIEVAL	DOWNLOAD	TOTAL
NO POLICY	24 months of standard	€ 480			
USER POLICY	3 months of standard > 9 months of coldline > 12 months of archive	€ 110			



- **Size:** 1 TB
- **Time:** 2 years
- **Full Downloads:** 5
  (1 STD, 2 in CL, 2 in ARC)

		STORAGE	RETRIEVAL	DOWNLOAD	TOTAL
NO POLICY	24 months of standard	€ 480	NA	5 x € 130	€ 1130
USER POLICY	3 months of standard > 9 months of coldline > 12 months of archive	€ 110	CL 2 x € 30 + ARC 2 x € 70	5 x € 130	€ 960



- **Size:** 1 TB
- **Time:** 2 years
- **Full Downloads:** 8
  (1 STD, 3 in CL, 4 in ARC)

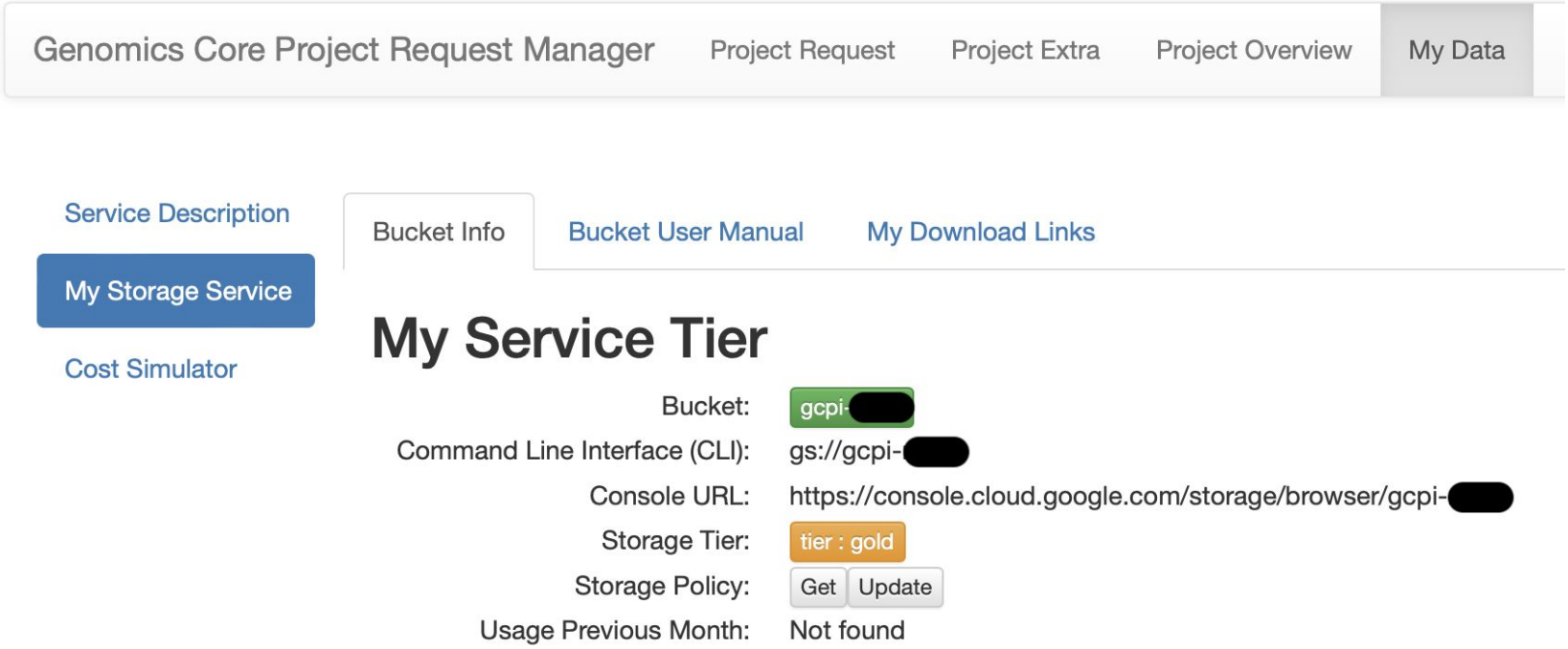
		STORAGE	RETRIEVAL	DOWNLOAD	TOTAL
NO POLICY	24 months of standard	€ 480	NA	8 x € 130	€ 1520
USER POLICY	3 months of standard > 9 months of coldline > 12 months of archive	€ 110	CL 3 x € 30 + ARC 4 x € 70	8 x € 130	€ 1520

- If Gold Tier → **setup bucket policy**!
  - Cost **can be reduced significantly,** especially if you perform **many** NGS experiments

- Visit www.genomicscore.be/request » > Login > My Data > Cost Simulator

- REMARK: Objects can only go **down** in storage class, **not up**!

Gold Tier: Where?



Gold Tier: How?

> **Warning**
>
> Always ensure you understand the commands you execute to avoid unintended behaviour.

Console / User Interface (UI) | Terminal / Command Line Interface (CLI)

CLI

Prerequisites

If you want to access the data through the Command Line Interface (CLI) please install gsutil on your device through the Google Cloud CLI. More info on the gsutil usage can be found here.

Syncing

One can use gsutil rsync to match source and destination (unidirectional)

```
gsutil -m rsync -r src/ dst/
```

Listing files

```
# List
gsutil ls gs://{BUCKET_NAME}
```

Other

```
# Delete
gsutil rm gs://{BUCKET_NAME}/{FILE_NAME}
```

```
# Bucket size
gsutil du -sh gs://{BUCKET_NAME}
```

```
# File size
gsutil du -sh gs://{BUCKET_NAME}/{FILE_NAME}
```

Data Delivery: A Tier Based Service

TIER	Silver (default)	Gold
DELIVERY	Download link	Google Cloud Bucket
ACCESS	Through PRM	Through ***share accounts*** (@share.uzleuven.be)
POLICY	Valid for 90 days	Custom - Customer Controlled
PRICING	Included in Project	Billed Quarterly

Modify Tier → info@genomicscore.be

Set / Modify Storage Policy → www.genomicscore.be/request» > Login > My Data > Bucket Info

Overview

NGS Data Management

1. Generating Data
    a. Data Flow
    b. Typical Datasets
    c. Sequencers Overview
2. Delivering Data - Tier Based
    a. Silver Tier
    b. Gold Tier
3. **Dos and Don'ts with Data**

Dos and Don'ts with Data

Please **DON'T** 🙅 …

… download data at home

… download on your personal computer

… rely on GC for backups
… choose Gold Tier without cost reducing policy

Rather **DO** 👌 …

… download at **the office**

… download on a **server** or server like machine **with ample storage capacity**

… devise a **data management strategy + infrastructure**

→ store your data
in a **structured** manner
with **scalability** in mind
and with the appropriate **access/permission**

📧 info@genomicscore.be

🌐 www.genomicscore.be

🐦 @GC_Leuven

⛳ Centre for Human Genetics

UZ – KU Leuven
Herestraat 49 PO box 606
B-3000 Leuven
Belgium

Conclusion

- You **request**, we **generate** data
- We **deliver** data
  - **Silver Tier** (Download Link)
  - **Gold Tier** (Google Cloud Bucket)
- For **technical** support: help.genomicscore.be 🚀