AI from Space using Azure

Christos Charmatzis

@christoscharmatzis

https://tageoforce.com

Center of the Milky Way Galaxy

Source: Wikipedia

Agenda

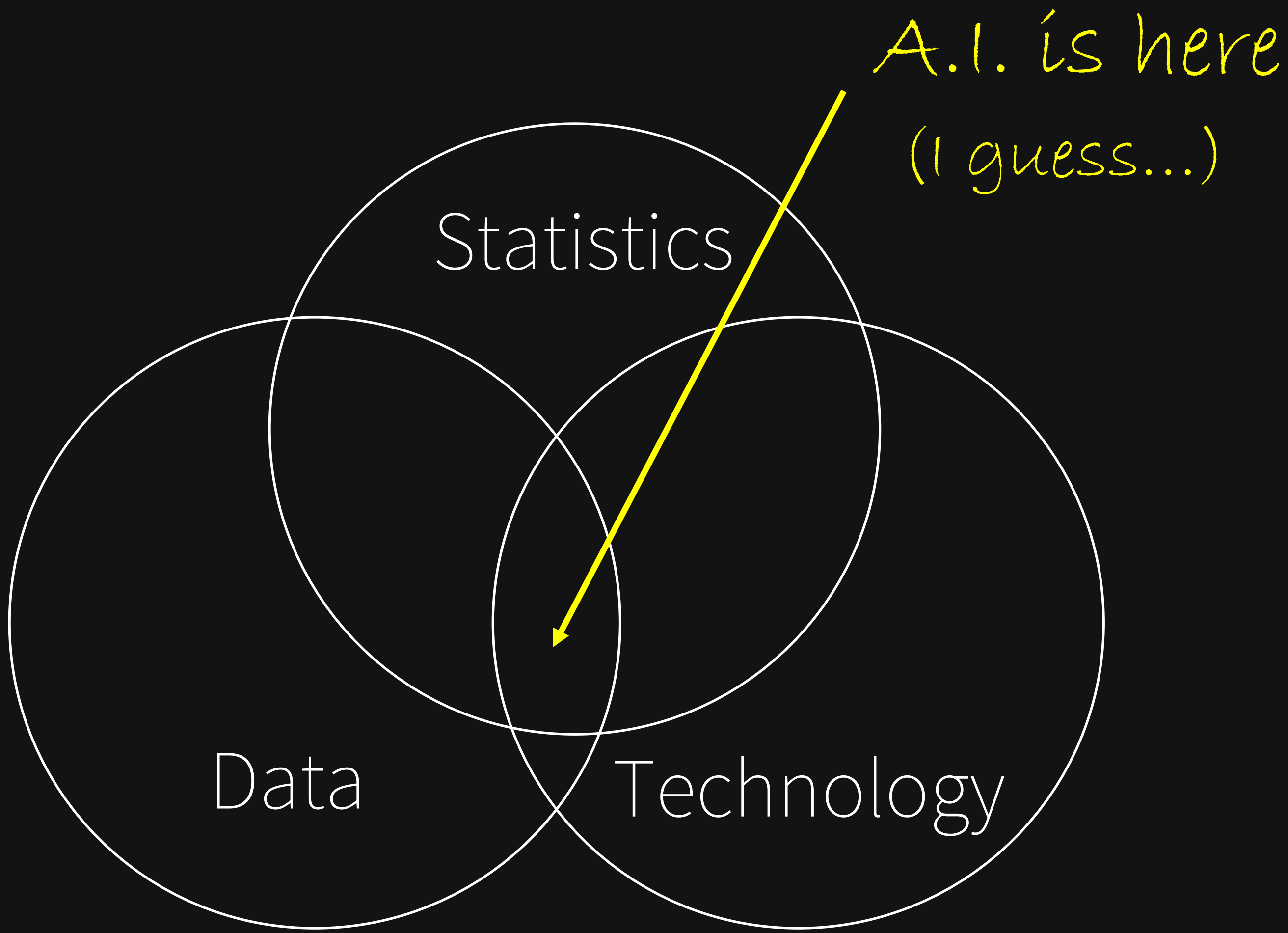
- Few things about me
- Introduction to AI
- Earth Observation Data and where to find them
- Choosing the right way with the right tools
- Going to Azure for full power
- Conclusions

Few things about me

- Project manager @TA-Geoforce
- GIS Specialist 10+ years
- AI professional
- Open Source enthusiasm
- Piano player

Chopin – Heroic Polonaise
Source: youtube.com/Rouseau

What they show us



What's AI for me

AI ≡ Knowledge

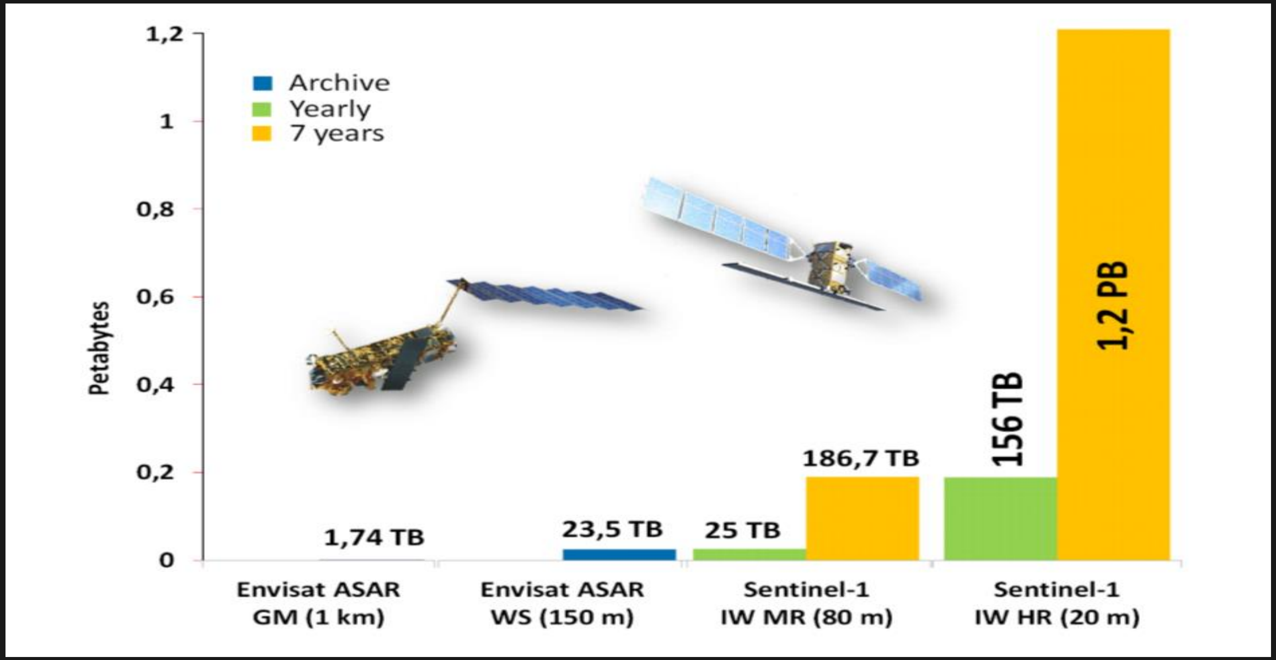
Data

All of us

Earth Observation Data

Only ESA satellites produces around 150 terabytes per day!!

(source:https://www.esa.int/Applications/Observing_the_Earth/Working_towards_AI_and_Earth_observation)



Growth of data volume from ENVISAT ASAR to Sentinel-1.

Source: Big Data Infrastructures for Processing Sentinel Data - Wolfgang Wagner, Vienna - 2015

Did you know that Azure has an Open Data Catalogue?

- MODIS
- NAIP
- NOAA Global Forecast System (GFS)
- Harmonized Landsat Sentinel-2
- NOAA Integrated Surface Data (ISD)
- Daymet

And the best part is that are FREE OF CHARGE!

https://azure.microsoft.com/en-us/services/open-datasets/catalog

Let’s get started

1. Talk with the client about the goal of the AI project.
2. Split the question that needs to be answered in small questions.
3. Search for datasets
4. Create the project structure.

You just hit the wall

The problem in every single AI project is ONE (1) WRANGLING with the DATA.

Visualize your data

Use ready examples

Azure Notebooks

Package: blob storage

Language: Python

Download Notebook

Demo notebook for accessing MODIS data on Azure

This notebook provides an example of accessing MODIS data from blob storage on Azure, including (1) finding the MODIS tile corresponding to a lat/lon coordinate, (2) retrieving that tile from blob storage, and (3) displaying that tile using the rasterio library.

This notebook uses the MODIS surface reflectance product as an example, but data structure and access will be the same for other MODIS product.

MODIS data are stored in the East US data center, so this notebook will run most efficiently on Azure compute located in East US. We recommend that substantial computation depending on MODIS data also be situated in East US. You don't want to download hundreds of terabytes to your laptop! If you are using MODIS data for environmental science applications, consider applying for an AI for Earth grant to support your compute requirements.

Imports and environment

In [3]:

```
import os
import tempfile
```

Spatial data are special data?

Tensorflow and Pytorch are specialized Deep Learning frameworks, that are developed for specific needs. E.g. Image recognition

Things don’t go well when you try to use them outside their comfort zone.

My favorite deep learning framework

Raster-Vision

https://rastervision.io/

Raster Vision is an open source framework for Python developers building computer vision models on satellite, aerial, and other large imagery sets (including oblique drone imagery)

Spatial Data vs Big Data

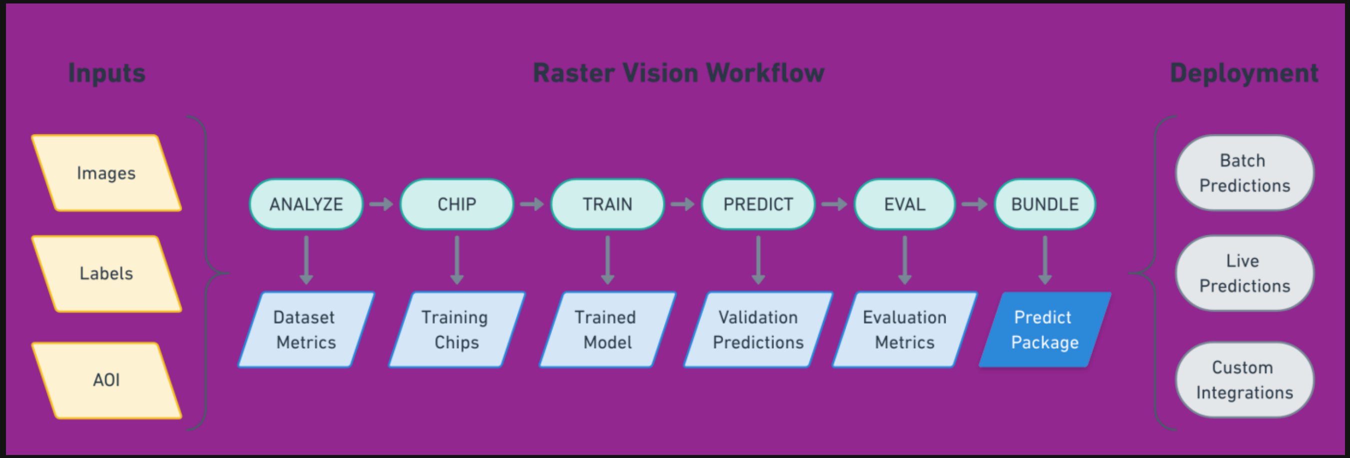
Since everything is relative:

- If you are studying (labeling) small features (e.g. roofs, cars, parking places) you are OK!!! There is nothing to worry about Big Data

- If you are studying (labeling) large features (e.g. lakes, oil spills, forests)

You are in Big (Trouble) Data!!!!

Raster-Vision workflow



Dockerize everything

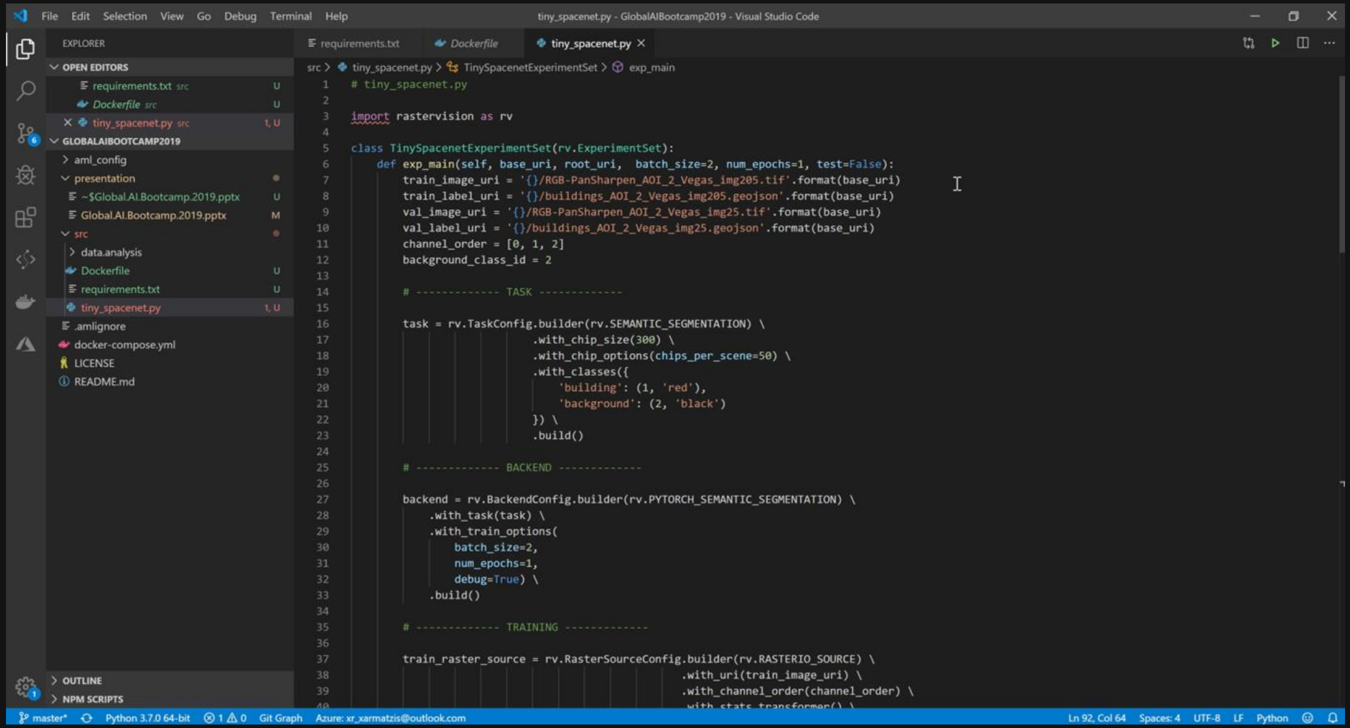
//requirements.txt

```
azure
azure-storage
azure-storage-blob
```

#Dockerfile

```
FROM quay.io/azavea/raster-vision:pytorch-0.10

COPY requirements.txt /
RUN pip install -r /requirements.txt

COPY  tiny_spacenet.py

ENV PATH=$PATH:/src
ENV PYTHONPATH /src

ADD ./ /src
WORKDIR /src/
```

Write experiments



Build & run

//Build docker image

docker build -t charmatzis/raster_vision_azure_batch_demo .

//Run it

docker run charmatzis/raster_vision_azure_batch_demo python /src/tiny_spacenet.py -- base_uri

wasbs://demo@charmatzis.blob.core.windows.net/ --root_uri wasbs://demo@charmatzisdata.blob.core.windows.net/results

The Don’t’s

1. Never use Windows, always Linux
2. Don’t use the CPU versions, always the GPU
3. Never run it in your local computer.

- Bonus –

4. Don’t go to your supervisor for a new Alienware laptop…. ;-)

If not local, then what?

Good choice (generally) Azure Machine Learning and good match with VS Code.

If you are working in special stuff (as we always do) use Azure Batch.

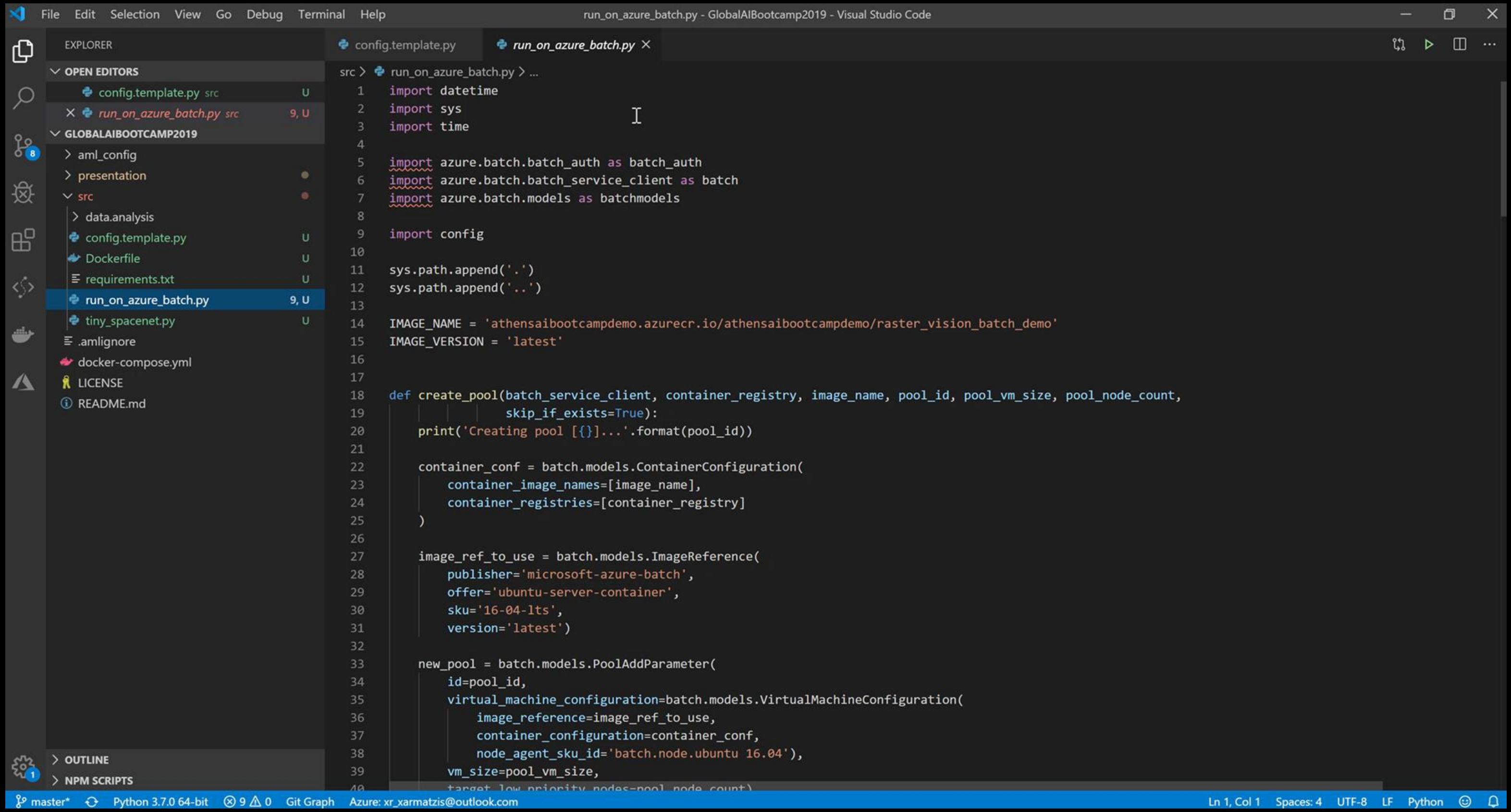
Move images to Azure with 3 simple moves

- Azure Container Registry
  docker login athensaibootcampdemo.azurecr.io

- Tag your docker container
  docker tag charmatzis/raster_vision_azure_batch_demo:latest athensaibootcampdemo.azurecr.io/charmatzis/raster_vision_azure_batch_demo :latest

- Upload it to ACR
  docker push athensaibootcampdemo.azurecr.io/athensaibootcampdemo / raster_vision_azure_batch_demo :latest

Run it on Azure Batch



Run it on Azure Batch…
But how?

- It connects to your container registry and uses those docker images
- Create a Pool if it doesn’t exist yet. Here, you can configure which kind of VMs and how many of them you want in your pool. And more importantly, you can specify that it are Low Prio VMs, which are cheap.
- Create a Job within the Pool
- Create a separate task to process each year of data. In a real-life situation, you would have a task for each day of data.

$Pricing$

NC-series

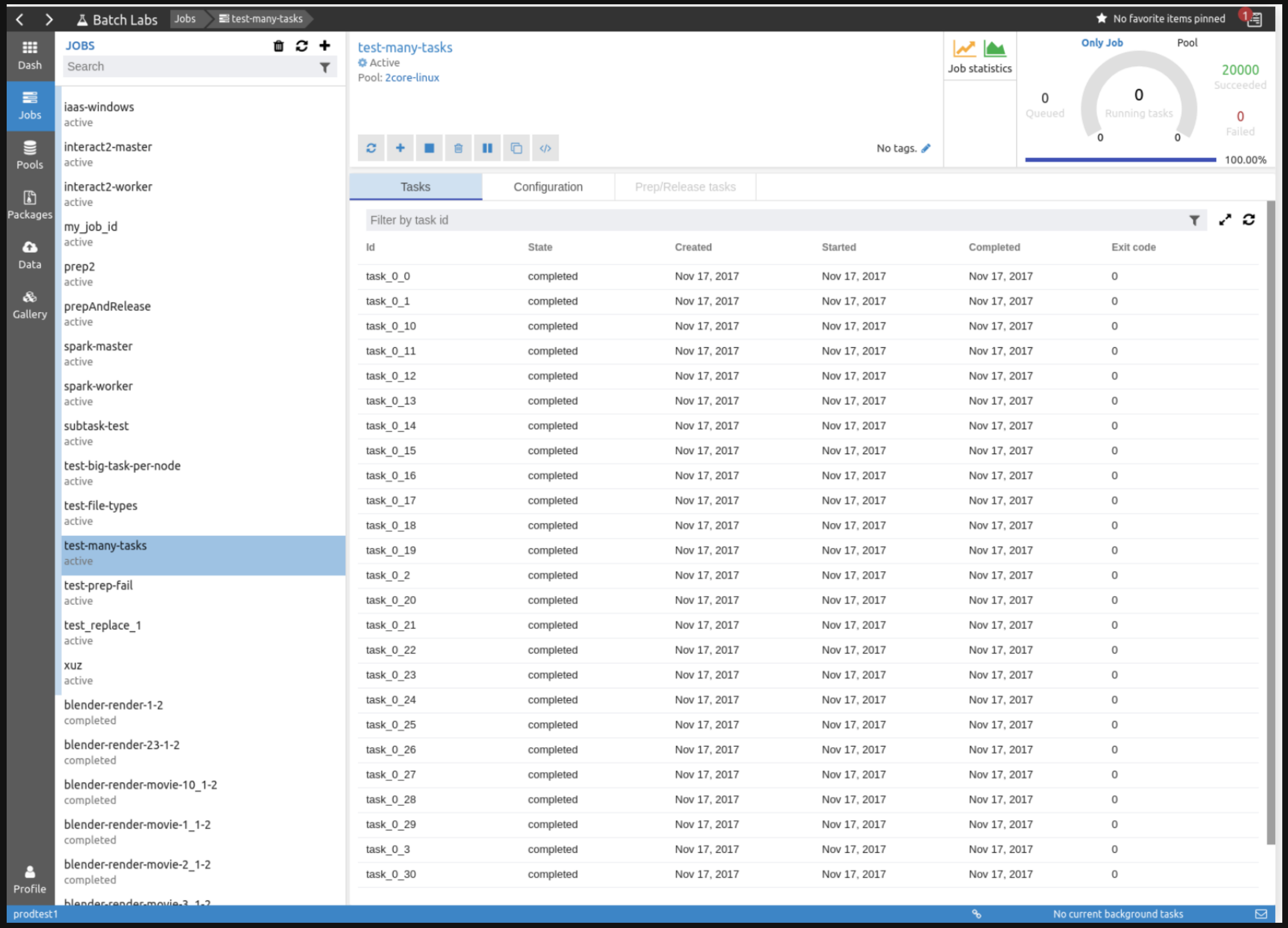
Add to estimate	Instance	Core	RAM	Temporary storage	GPU	Pay as you go (Low priority)	Pay as you go (normal priority)	1 year reserved (% Savings)	3 year reserved (% Savings)	Spot (% Savings)
⊕	NC6	6	56 GiB	340 GiB	1X K80	$0.18/hour	$0.90/hour	$0.5733/hour (~36%)	$0.3996/hour (~56%)	$0.18/hour (~80%)
⊕	NC12	12	112 GiB	680 GiB	2X K80	$0.36/hour	$1.80/hour	$1.1466/hour (~36%)	$0.7991/hour (~56%)	$0.36/hour (~80%)
⊕	NC24r	24	224 GiB	1,440 GiB	4X K80	$0.792/hour	$3.96/hour	$2.5224/hour (~36%)	$1.7578/hour (~56%)	$0.792/hour (~80%)
⊕	NC24	24	224 GiB	1,440 GiB	4X K80	$0.72/hour	$3.60/hour	$2.2932/hour (~36%)	$1.5981/hour (~56%)	$0.72/hour (~80%)

NCsv2-series

Add to estimate	Instance	Core	RAM	Temporary storage	GPU	Pay as you go (Low priority)	Pay as you go (normal priority)	1 year reserved (% Savings)	3 year reserved (% Savings)	Spot (% Savings)
⊕	NC6s v2	6	112 GiB	736 GiB	1X P100	$0.36/hour	$2.07/hour	$1.3187/hour (~36%)	$0.9189/hour (~56%)	$0.36/hour (~83%)
⊕	NC12s v2	12	224 GiB	1,474 GiB	2X P100	$0.72/hour	$4.14/hour	$2.6371/hour (~36%)	$1.8378/hour (~56%)	$0.72/hour (~83%)
⊕	NC24rs v2	24	448 GiB	2,948 GiB	4X P100	$1.584/hour	$9.108/hour	$5.8015/hour (~36%)	$4.0430/hour (~56%)	$1.584/hour (~83%)
⊕	NC24s v2	24	448 GiB	2,948 GiB	4X P100	$1.44/hour	$8.28/hour	$5.2742/hour (~36%)	$3.6755/hour (~56%)	$1.44/hour (~83%)

https://azure.microsoft.com/en-us/pricing/details/batch/

How can I monitor my Batch?



Conclusions

- If you have normal experiments, use Azure Machine Learning
- If you are working in some crazy stuff go straight to Azure Batch using containers.
- Never, use your laptop for deep learning…

Thank U

&

Questions