A use case of humanware and cloud-based CI: Time-series data classification using ML

Yongwook Song, Xu Fu, Chris Richards

University of Kentucky
Center for Computational Sciences
Computational Research Engineer
ywsong2@uky.edu

HARC at PEARC 19

Outlines

1. Introduction
2. Using CI
3. Our Science Challenge
4. Implementations
5. Results
6. Humanware Discussion
7. Conclusions
8. References

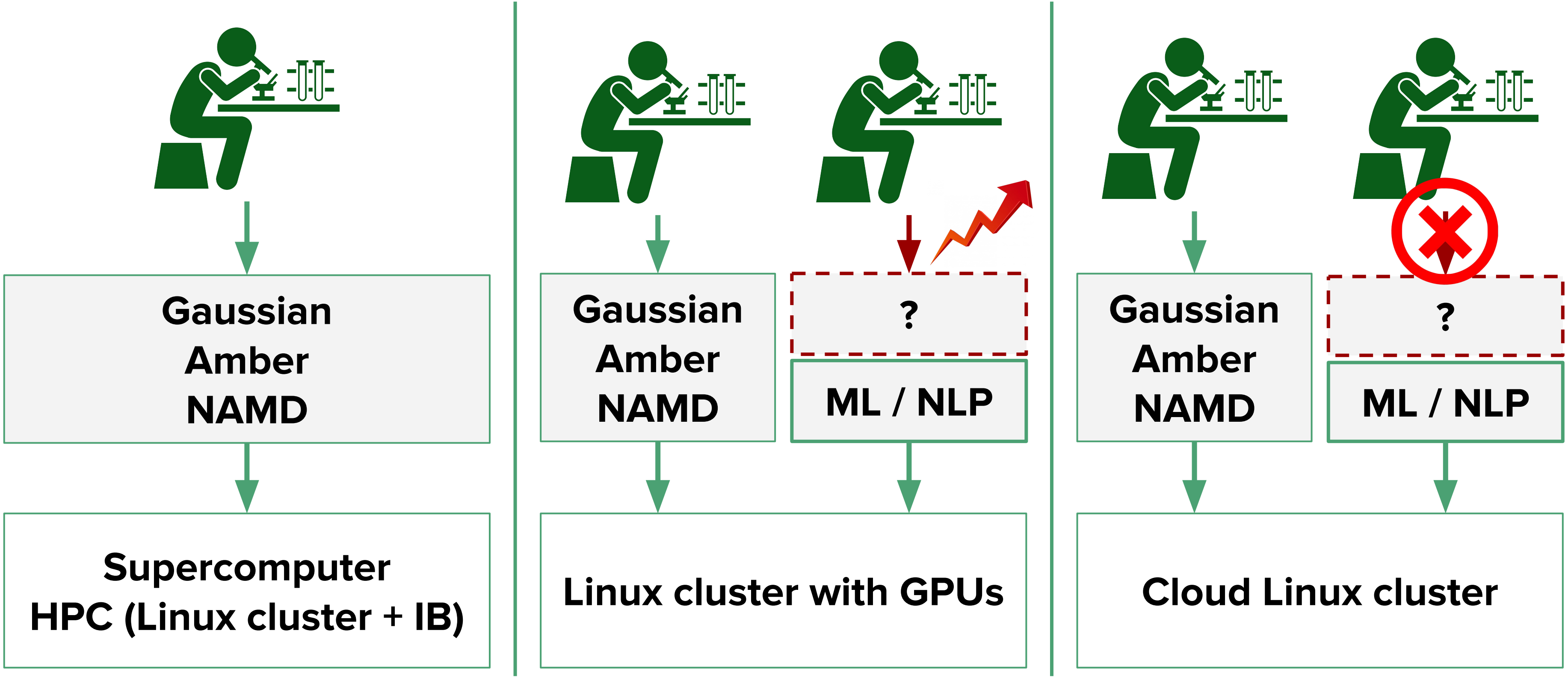
1.1 What is Cyberinfrastructure (CI) and Humanware?

Cyberinfrastructure (CI) can be defined as consisting of “… *computing systems, data storage systems, advanced instruments and data repositories, visualization environments,* ***and people****, all linked together by software and high-performance netw* *ductivity and enable breakthroughs not other*

Humanware[2]

- ➔ **Administering** physical component of CI
- ➔ **Support** researchers to utilize CI
- ➔ **Collaborate** with researchers
- ➔ Increase **efficiency**
- ➔ **Maximize** Return-On-Investment (ROI) of CI [6]
- ➔ Make **breakthrough** / Find **innovative** solutions

2.1 CI for researchers


Gaussian
Amber
NAMD

Supercomputer
HPC (Linux cluster + IB)

Gaussian
Amber
NAMD

?

ML / NLP

Linux cluster with GPUs

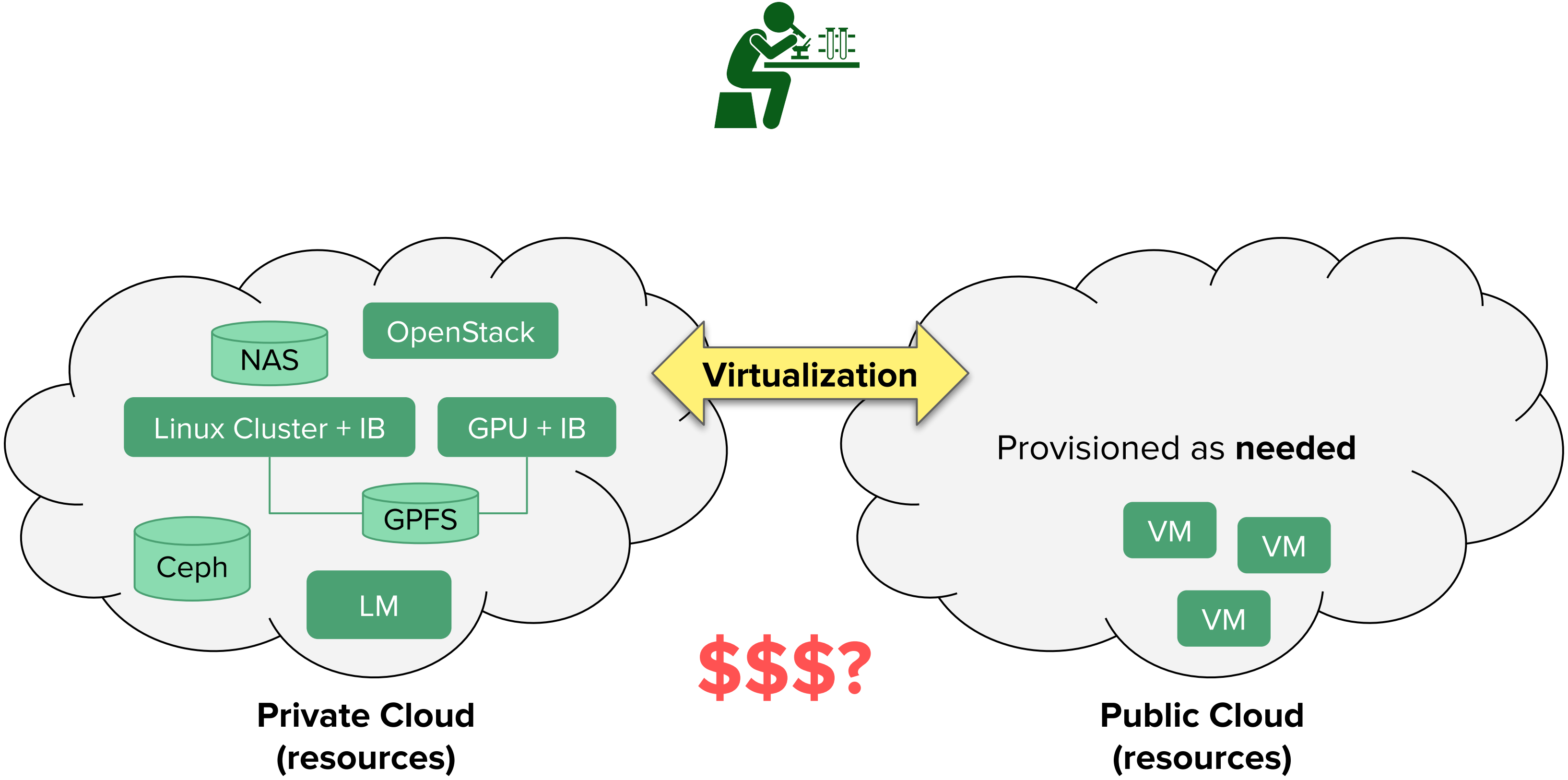
Gaussian
Amber
NAMD

?

ML / NLP

Cloud Linux cluster

2.2 Clouds complicate things


NAS
OpenStack
Linux Cluster + IB
GPU + IB
GPFS
Ceph
LM
Virtualization
Provisioned as **needed**
VM
VM
VM
$$$?
Private Cloud (resources)
Public Cloud (resources)

2.3 Cloud provisioning challenges

➔ Virtual Machine for Cloud-based CI

- ◆ **Operating System:** Windows? Linux? Distribution? Version?
- ◆ **CPU:** Number of CPUs?
- ◆ **GPU:** Number of GPUs? Nvidia-CUDA enabled?
- ◆ **RAM:** How large RAM?
- ◆ **Storage:** SSD? HDD? How large?
- ◆ **Network:** Requirement? How fast?

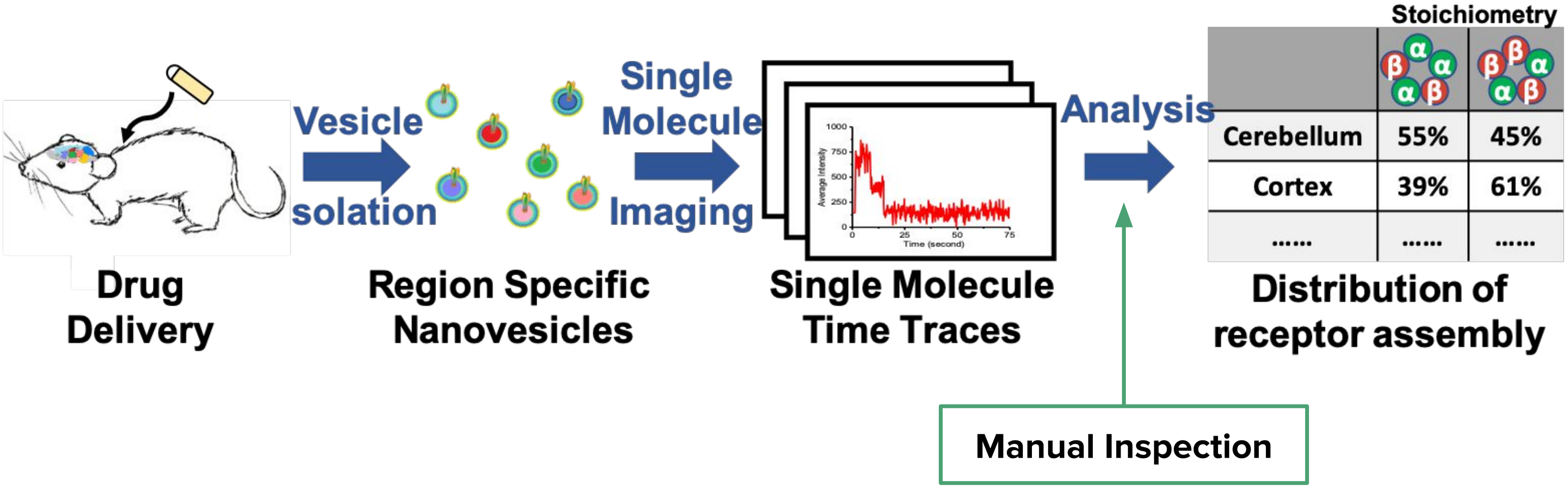
2.4 Cloud workflow challenges

- ➔ Move **data** in & out
  - ◆ scp / ssh / rclone
  - ◆ Git server
  - ◆ Dropbox / Google Drive
- ➔ **Sync** codes and data
  - ◆ Version control -- git / SVN / CVS
  - ◆ ~~**No version control!! -- Local disk / file servers / USB**~~
- ➔ Documentation
  - ◆ **Wiki** pages / How-to / Reports -- Google drive / Web pages / Project tracking tool
- ➔ Launching **pipeline** on data (stream processing)

2.5 Cloud programming challenges

- ➔ **Writing** machine learning codes
- ➔ What ML framework? (**Tensorflow / Keras / Pytorch / Theano**)
- ➔ Python, C++, R?
- ➔ Data **pre-processing**
- ➔ Data **post-processing**
- ➔ User interface?
- ➔ How to **manage** data, codes & results?
- ➔ How to **visualize** results?
- ➔ **Learning curve** for researchers

3.1 Our Science Challenge: Drug delivery (Chem/Pharm)



	Stoichiometry	
	α β α α β	β β α α β
Cerebellum	55%	45%
Cortex	39%	61%
......	......	......

Drug Delivery

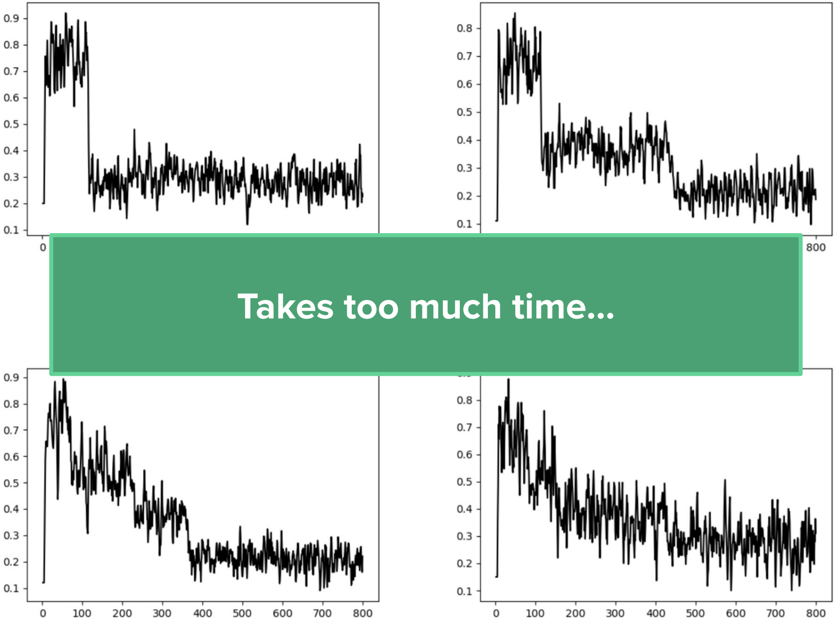
Region Specific Nanovesicles

Single Molecule Time Traces

Distribution of receptor assembly

Manual Inspection

3.2 Ambiguous Time-series Data



Takes too much time...

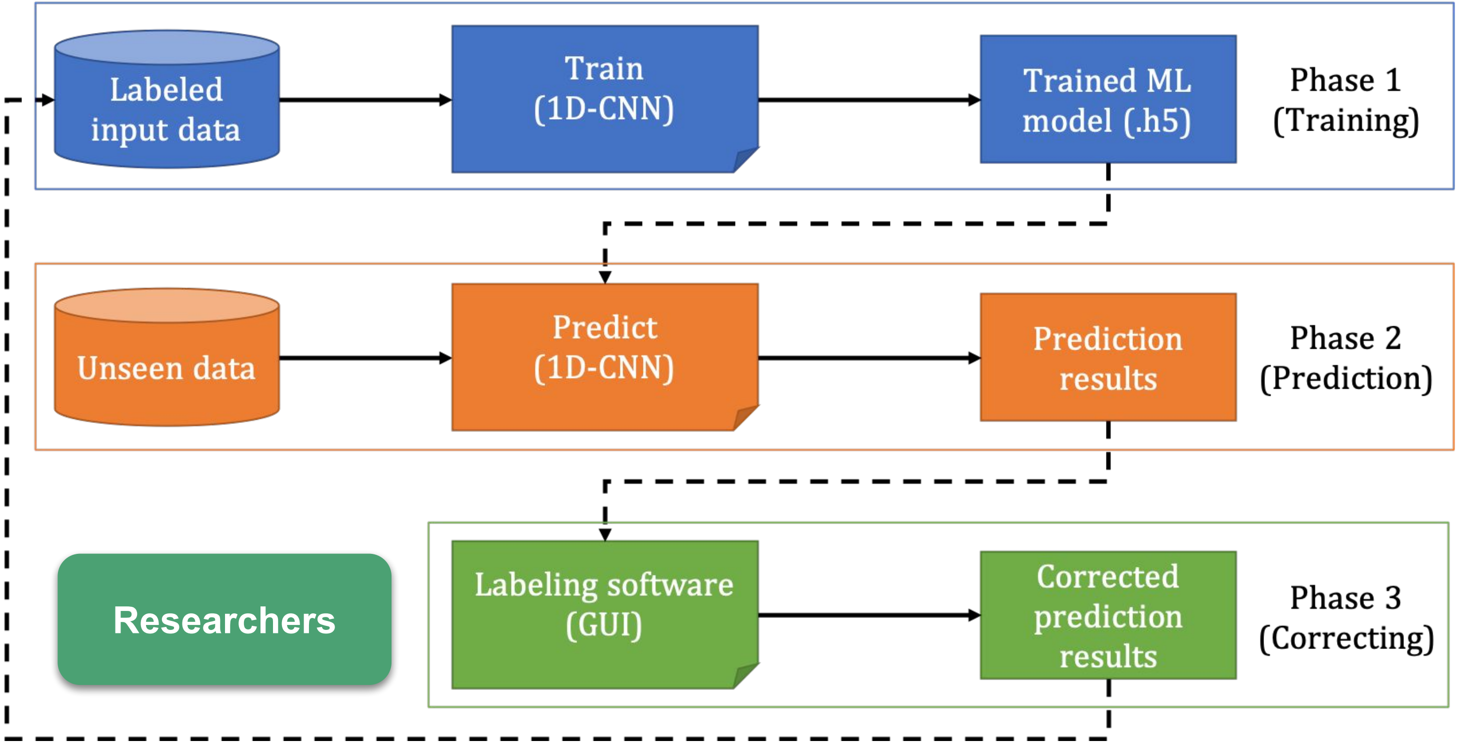
(c) 3 steps

(d) 0 steps

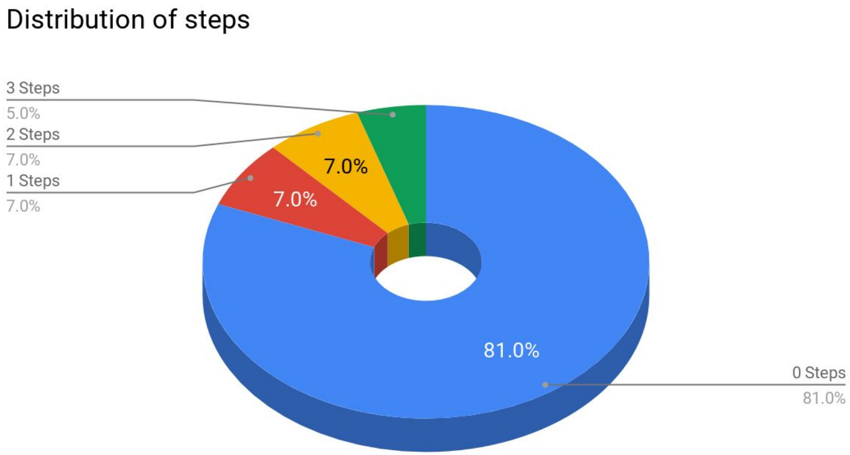
Figure 1: Example of actual data with correct labels

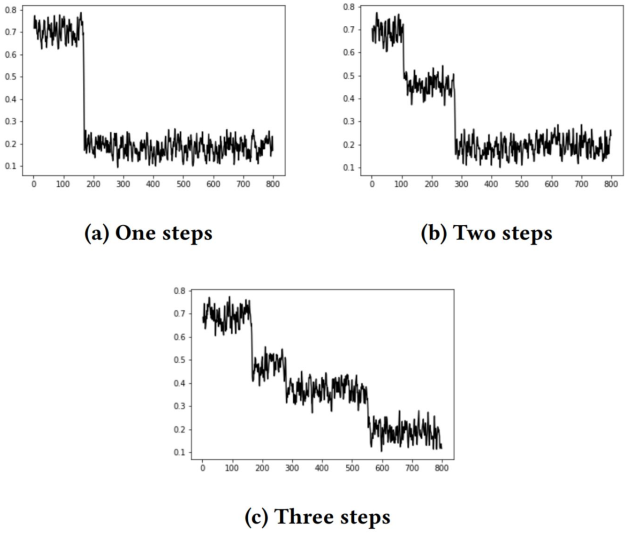
Machine Learning

4.1 Processing Pipeline



4.2 Augmentation for training phase





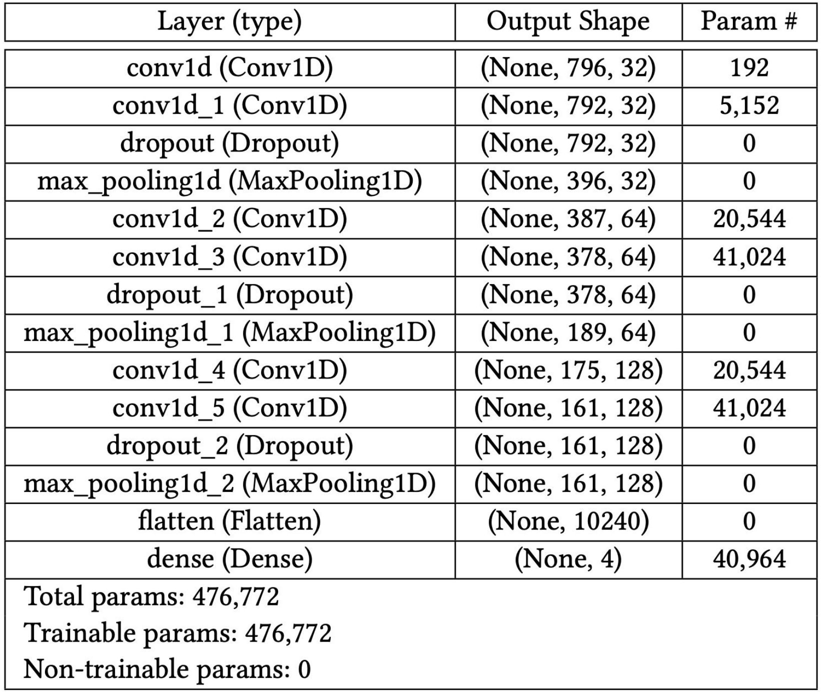
(a) One steps

(b) Two steps

(c) Three steps

Figure 4: Example of augmented data for 1, 2, and 3 steps

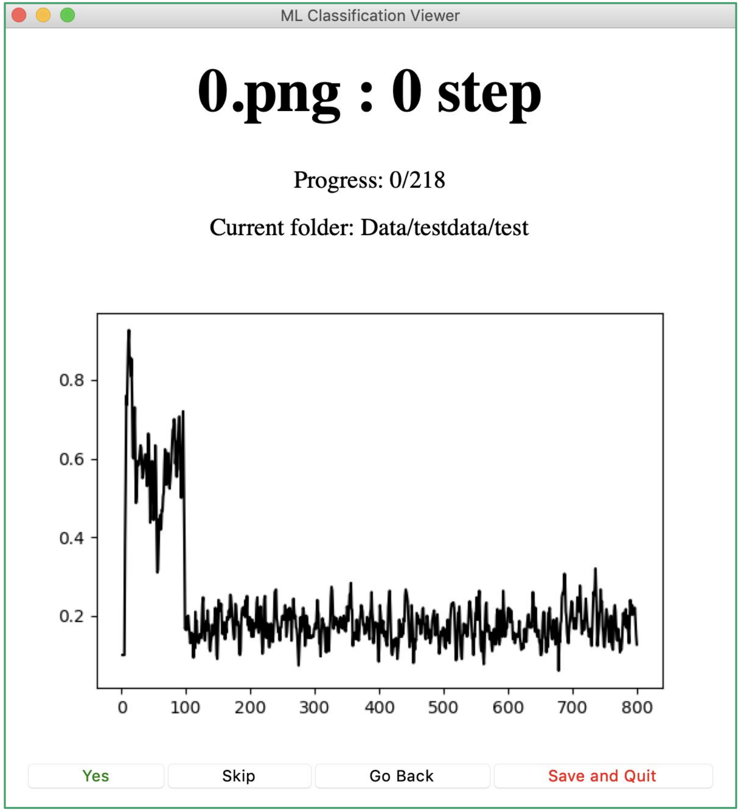
4.3 ML approaches / algorithm (1D-CNN)



Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 796, 32)	192
conv1d_1 (Conv1D)	(None, 792, 32)	5,152
dropout (Dropout)	(None, 792, 32)	0
max_pooling1d (MaxPooling1D)	(None, 396, 32)	0
conv1d_2 (Conv1D)	(None, 387, 64)	20,544
conv1d_3 (Conv1D)	(None, 378, 64)	41,024
dropout_1 (Dropout)	(None, 378, 64)	0
max_pooling1d_1 (MaxPooling1D)	(None, 189, 64)	0
conv1d_4 (Conv1D)	(None, 175, 128)	20,544
conv1d_5 (Conv1D)	(None, 161, 128)	41,024
dropout_2 (Dropout)	(None, 161, 128)	0
max_pooling1d_2 (MaxPooling1D)	(None, 161, 128)	0
flatten (Flatten)	(None, 10240)	0
dense (Dense)	(None, 4)	40,964

Total params: 476,772
Trainable params: 476,772
Non-trainable params: 0

4.4 Labeling GUI software



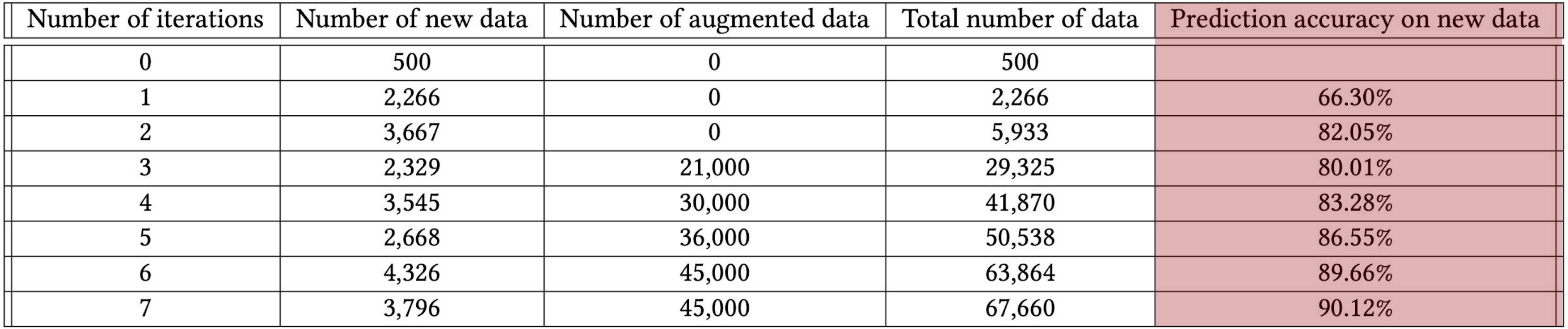
ML Classification Viewer

0.png : 0 step

Progress: 0/218

Current folder: Data/testdata/test

5.1 Records of training data and prediction accuracy

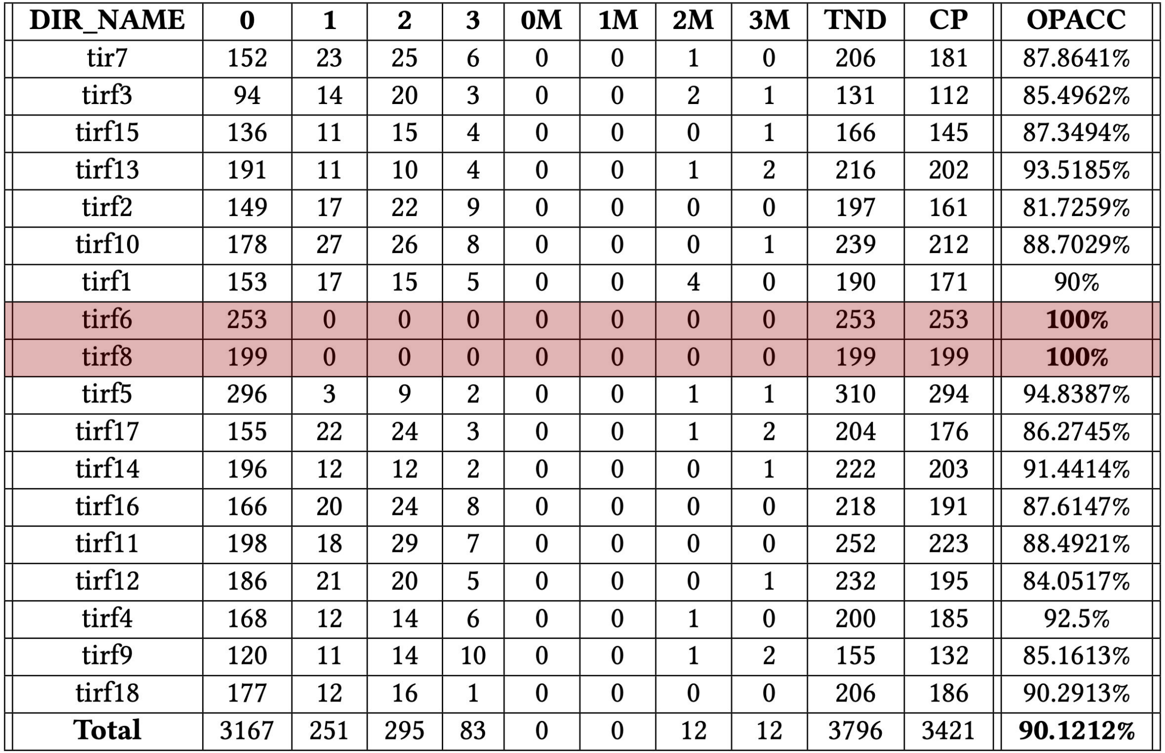


Number of iterations	Number of new data	Number of augmented data	Total number of data	Prediction accuracy on new data
0	500	0	500	
1	2,266	0	2,266	66.30%
2	3,667	0	5,933	82.05%
3	2,329	21,000	29,325	80.01%
4	3,545	30,000	41,870	83.28%
5	2,668	36,000	50,538	86.55%
6	4,326	45,000	63,864	89.66%
7	3,796	45,000	67,660	90.12%

Table 1: Records of training data size and prediction accuracy for iterations

- ➔ Each iteration takes a week (takes time to correct data by researchers)
- ➔ After 3rd iteration, we added augmented data set

5.2 Prediction results of the last iteration

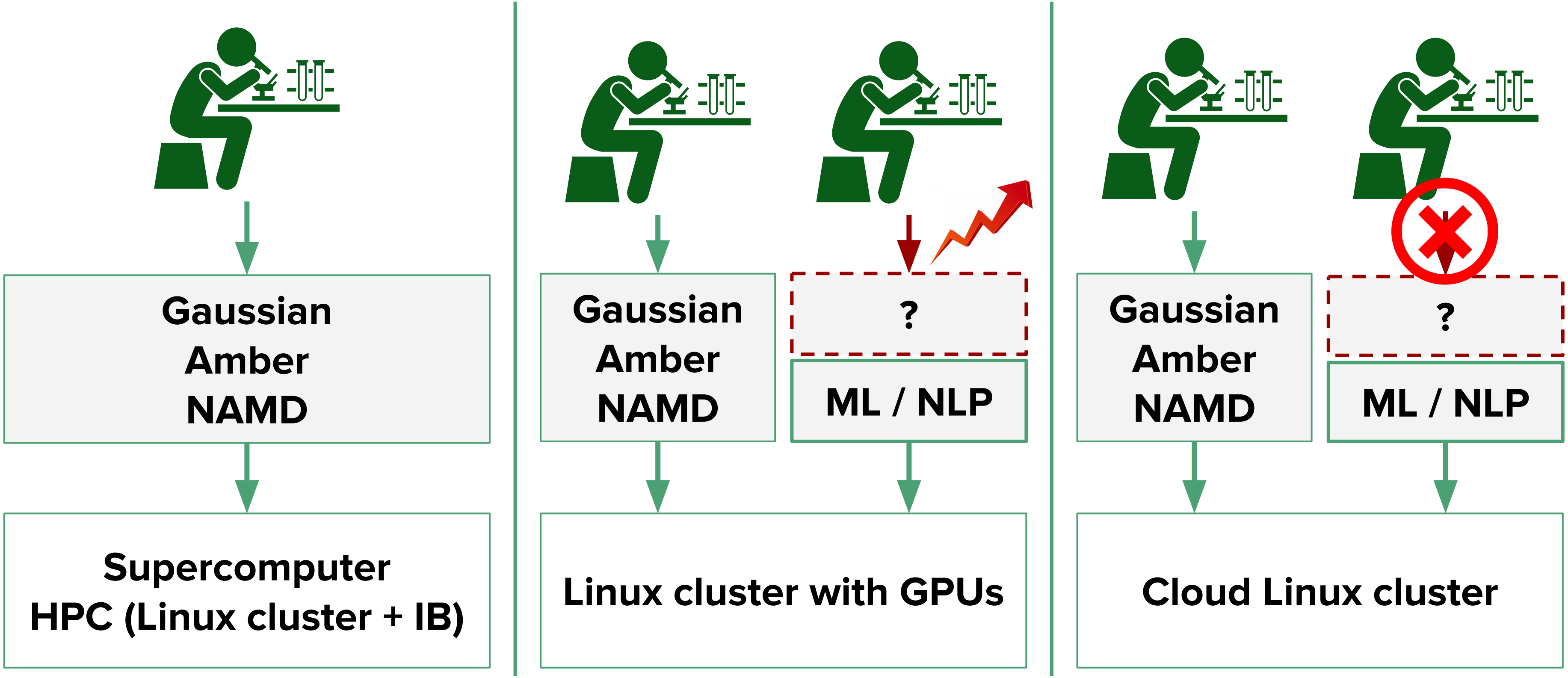


DIR_NAME	0	1	2	3	0M	1M	2M	3M	TND	CP	OPACC
tir7	152	23	25	6	0	0	1	0	206	181	87.8641%
tirf3	94	14	20	3	0	0	2	1	131	112	85.4962%
tirf15	136	11	15	4	0	0	0	1	166	145	87.3494%
tirf13	191	11	10	4	0	0	1	2	216	202	93.5185%
tirf2	149	17	22	9	0	0	0	0	197	161	81.7259%
tirf10	178	27	26	8	0	0	0	1	239	212	88.7029%
tirf1	153	17	15	5	0	0	4	0	190	171	90%
tirf6	253	0	0	0	0	0	0	0	253	253	**100%**
tirf8	199	0	0	0	0	0	0	0	199	199	**100%**
tirf5	296	3	9	2	0	0	1	1	310	294	94.8387%
tirf17	155	22	24	3	0	0	1	2	204	176	86.2745%
tirf14	196	12	12	2	0	0	0	1	222	203	91.4414%
tirf16	166	20	24	8	0	0	0	0	218	191	87.6147%
tirf11	198	18	29	7	0	0	0	0	252	223	88.4921%
tirf12	186	21	20	5	0	0	0	1	232	195	84.0517%
tirf4	168	12	14	6	0	0	1	0	200	185	92.5%
tirf9	120	11	14	10	0	0	1	2	155	132	85.1613%
tirf18	177	12	16	1	0	0	0	0	206	186	90.2913%
Total	3167	251	295	83	0	0	12	12	3796	3421	**90.1212%**

Table 3: Prediction results of the new data set in the seventh iteration

Humanware discussion

6.1 CI for researchers



Gaussian
Amber
NAMD

Supercomputer
HPC (Linux cluster + IB)

Gaussian
Amber
NAMD

?

ML / NLP

Linux cluster with GPUs

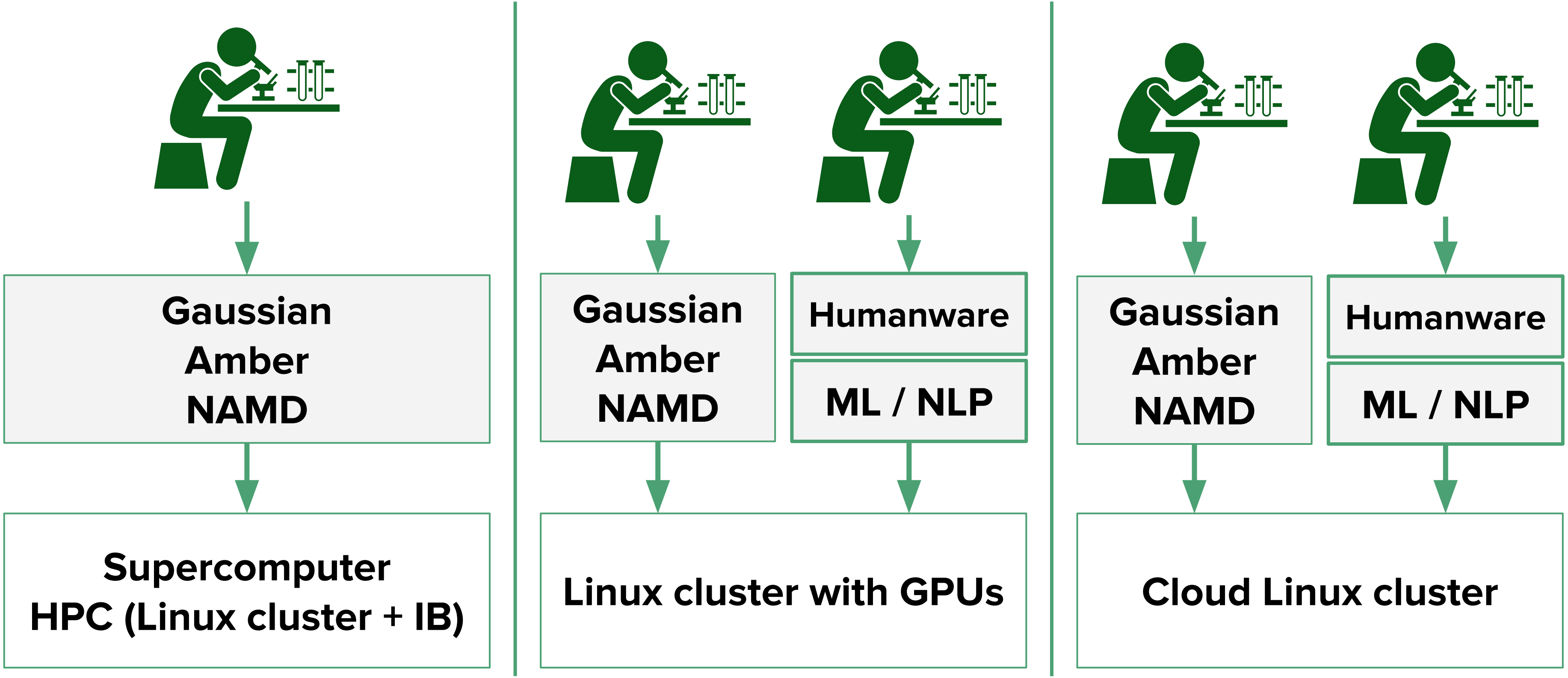
Gaussian
Amber
NAMD

?

ML / NLP

Cloud Linux cluster

6.2 Humanware in the loop



Gaussian
Amber
NAMD

Supercomputer
HPC (Linux cluster + IB)

Gaussian
Amber
NAMD

Humanware

ML / NLP

Linux cluster with GPUs

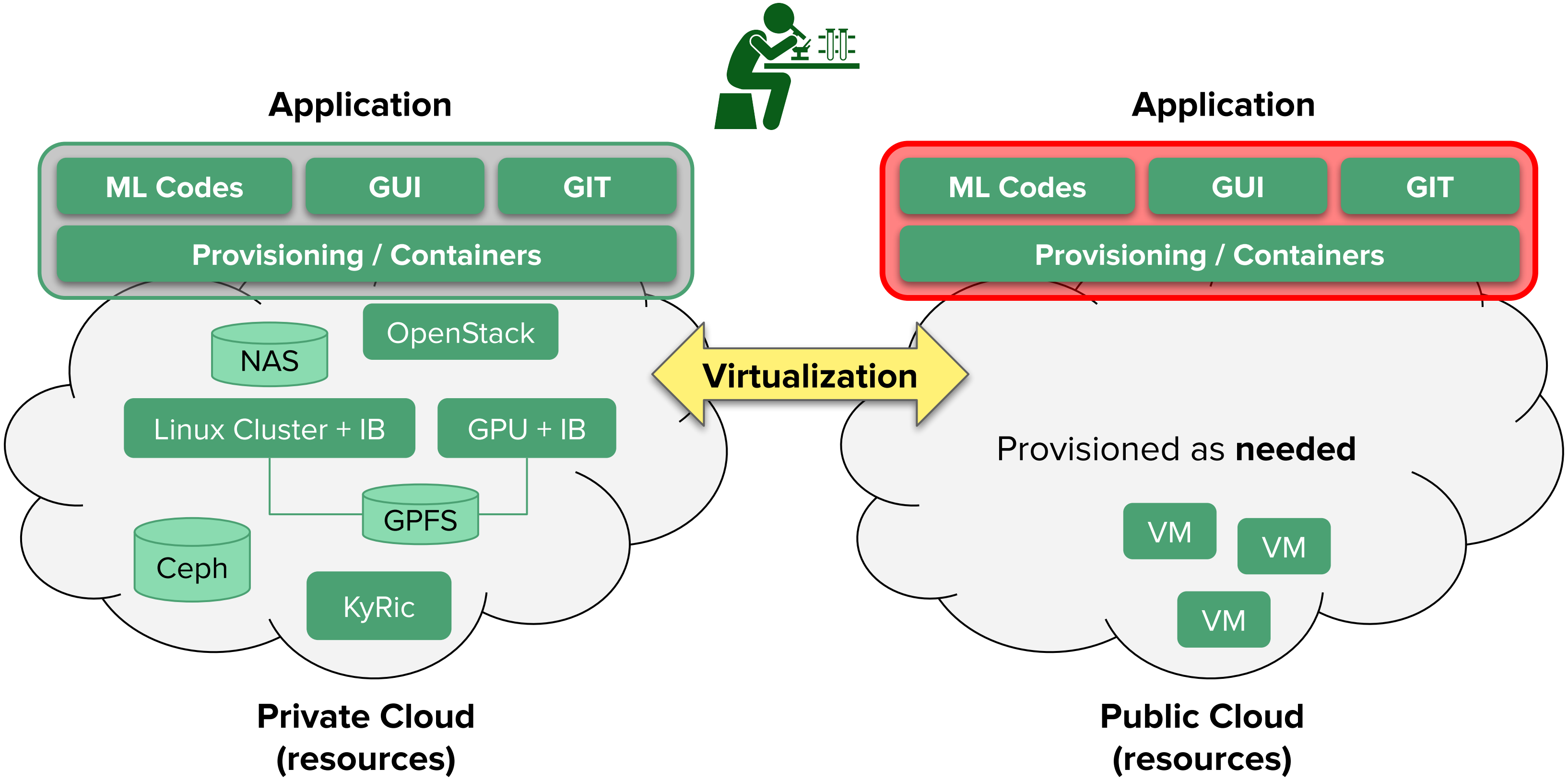
Gaussian
Amber
NAMD

Humanware

ML / NLP

Cloud Linux cluster

6.3 Moving to cloud CI



6.4 Cloud provisioning solutions

- Virtual Machine for Cloud-based CI
  - **Operating System:** Linux 16.04 LTS, 18.04 LTS
  - **CPU:** 6 x vCPUs
  - **GPU:** 1 x Nvidia Tesla M60 (8GB GPU memory)
  - **RAM:** 56GB
  - **Storage:** 340GB HDD + 512GB HDD
  - **Network:** Normal
  - **Cost**: NV6 Promo ($0.721 per hour + extra HDD = $200 per month)

6.5 Cloud workflow solutions

- Move **data** in & out
  - **Git server**
- **Sync** codes and data
  - **Version control -- git**
- Documentation
  - **Wiki** pages / How-to / Reports
  - Microsoft Azure Devops (https://ywsong2.visualstudio.com/Chem_ML_GUI)
- Launching **pipeline** on data (stream processing)
  - Training, Prediction, and Correction phases

6.6 Cloud programming solutions

- ➔ **Writing** machine learning codes (**1D-CNN**)
- ➔ What ML framework? (**Tensorflow / Keras / Nvidia CUDA**)
- ➔ Python? (**Anaconda** virtual environment)
- ➔ Data **pre- and post-processing** (Custom software)
- ➔ User interface? (Custom GUI software)
- ➔ How to **manage** data, codes & results? (Custom software)
- ➔ How to **visualize** results? (Custom software)

7.1 Conclusions

- ➔ **Clear need** for humanware as result of **new CI** and **research challenges**
- ➔ Many aspects of humanware component (**provisioning, workflow, programming, interfaces**)
- ➔ Possible to build **applications** that **hide details** for researchers
- ➔ Public cloud was **sufficient** and **usable** platform for our problem
- ➔ We could run in the **private cloud** with a few changes
- ➔ **Humanware** collaborates with researchers to **maximize** ROI of evolving CIs and make **breakthroughs**

QUESTIONS
ANSWERS

References

[1] Craig A. Stewart, Stephen Simms, Beth Plale, Matthew Link, David Y. Hancock, and Geoffrey C. Fox. 2010. What is Cyberinfrastructure. In Proceedings of the 38th Annual ACM SIGUCCS Fall Conference: Navigation and Discovery (SIGUCCS '10). ACM, New York, NY, USA, 37–44. https://doi.org/10.1145/1878335.1878347

[2] Brian D. Voss. 2015. The Critical Importance of People in Cyberinfrastructure. https://www.linkedin.com/pulse/critical-importance-peoplecyberinfrastructure-brian-d-voss/.

[3] Microsoft.2019. MicrosoftAzure. https://azure.microsoft.com/en-us/.

[4] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, CraigCitro,GregS.Corrado,AndyDavis,JeffreyDean,MatthieuDevin,SanjayGhemawat,IanGoodfellow,AndrewHarp,GeoffreyIrving,MichaelIsard,YangqingJia,RafalJozefowicz,LukaszKaiser,ManjunathKudlur,JoshLevenberg,DanMané,RajatMonga,SherryMoore,DerekMurray,ChrisOlah,Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker ,VincentVanhoucke, VijayVasudevan,FernandaViégas,OriolVinyals, PeteWarden,MartinWattenberg,MartinWicke,YuanYu,andXiaoqiangZheng. 2015. TensorFlow:Large-ScaleMachineLearningonHeterogeneousSystems. http://tensorflow.org/ Softwareavailablefromtensorflow.org.

[5] FrançoisCholletetal.2015. Keras. https://github.com/fchollet/keras.

[6] CraigAStewart, DavidYHancock, JulieWernert, ThomasFurlani,DavidLifka, AlanSill, NicholasBerente, DonaldFMcMullen, ThomasCheatham, AmyApon, etal.2019. Assessmentoffinancialreturnsoninvestmentsincyberinfrastructure facilities:Asurveyofcurrentmethods. (2019).

[7] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, MartinFowler, JamesGrenning, JimHighsmith, AndrewHunt,Ron Jeffries,JonKern,BrianMarick,RobertC.Martin,SteveMellor,KenSchwaber,Jeff Sutherland,andDaveThomas.2001. ManifestoforAgileSoftwareDevelopment. http://www.agilemanifesto.org/