Reproducibility in Computational Research

“ A Predictive DASH QoE Approach Based on Machine Learning at Multi-access Edge Computing ”

Md Tariqul Islam
MSc Student
FEEC, UNICAMP
mtarislam@gmal.com

Agenda

- What is Reproducible Research
- Importance of Reproducibility
- How to Achieve Reproducibility
- Project Overview
- Project Reproducibility
- Challenges and Lessons

What is Reproducible Research?

To reproduce original work

Getting same result by independent data, code and condition

Gold Standard

Always Not Feasible
(e.g., cost, complexity)

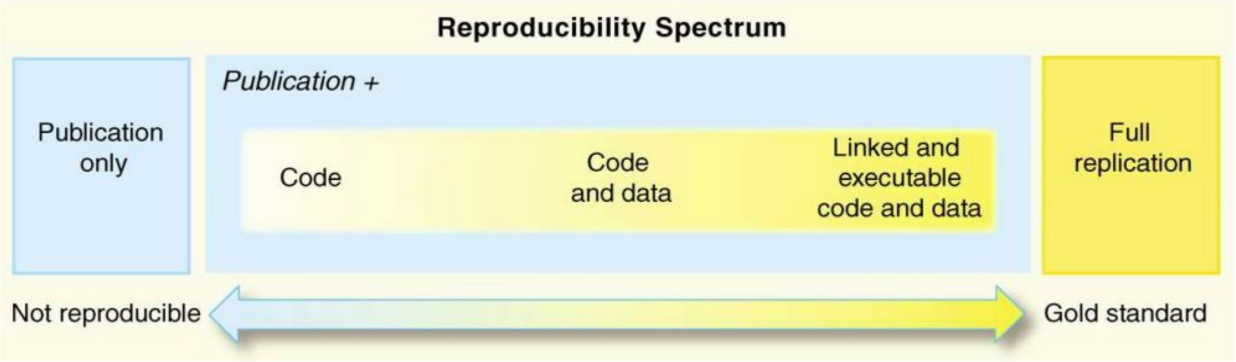
Replicability

Getting same result by same data , code and condition

Minimum Standard

Always Feasible

Reproducibility



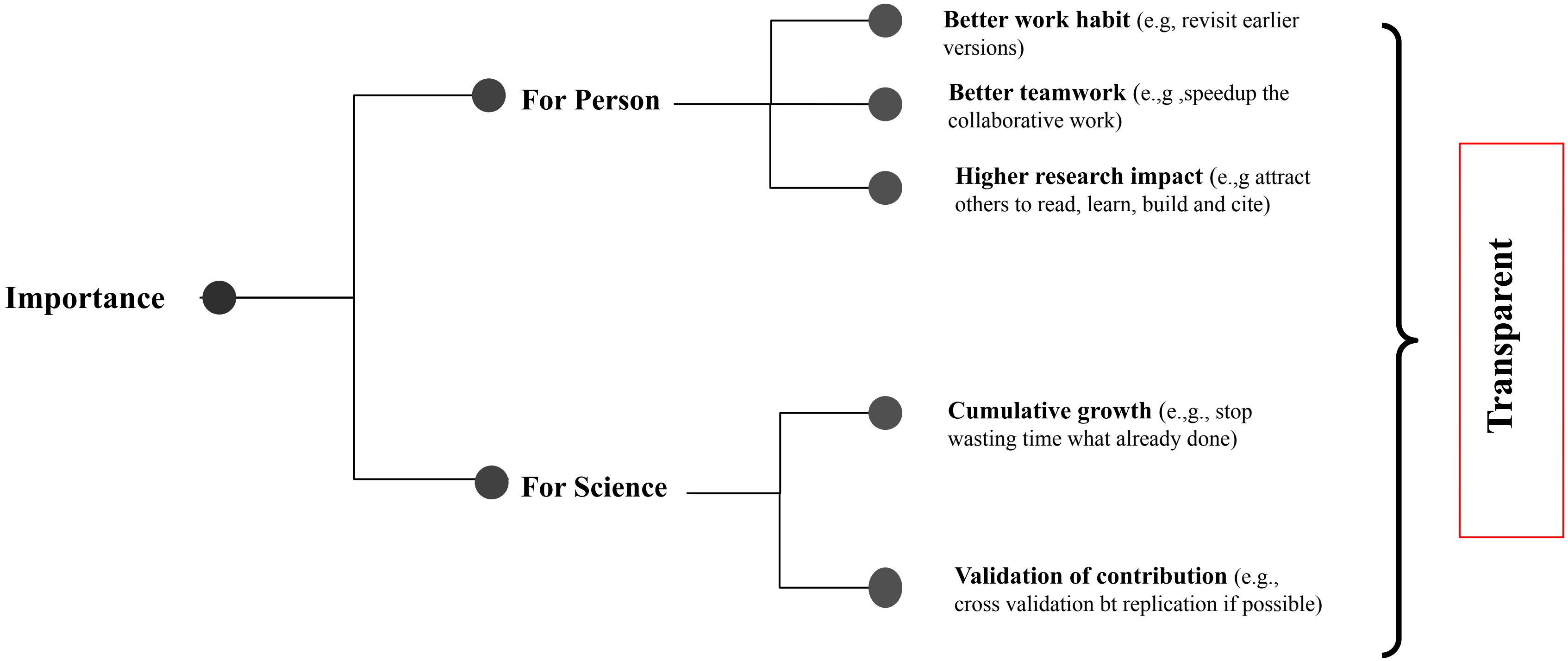
"Again, and Again, and Again …" BR Jasny et. al. Science, 2011. 334(6060) pp. 1225 DOI: 10.1126/science.334.6060.1225

"Reproducible Research in Computational Science". RD Peng Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

"Reproducible Research in Computational Science". RD Peng Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

Reproducibility in Computational Research is an exercise to make available of all data, code, and required tools for others to reproduce the same results discussed in original research work.

Importance of Reproducibility



Importance

For Person

Better work habit (e.g, revisit earlier versions)

Better teamwork (e.,g ,speedup the collaborative work)

Higher research impact (e.,g attract others to read, learn, build and cite)

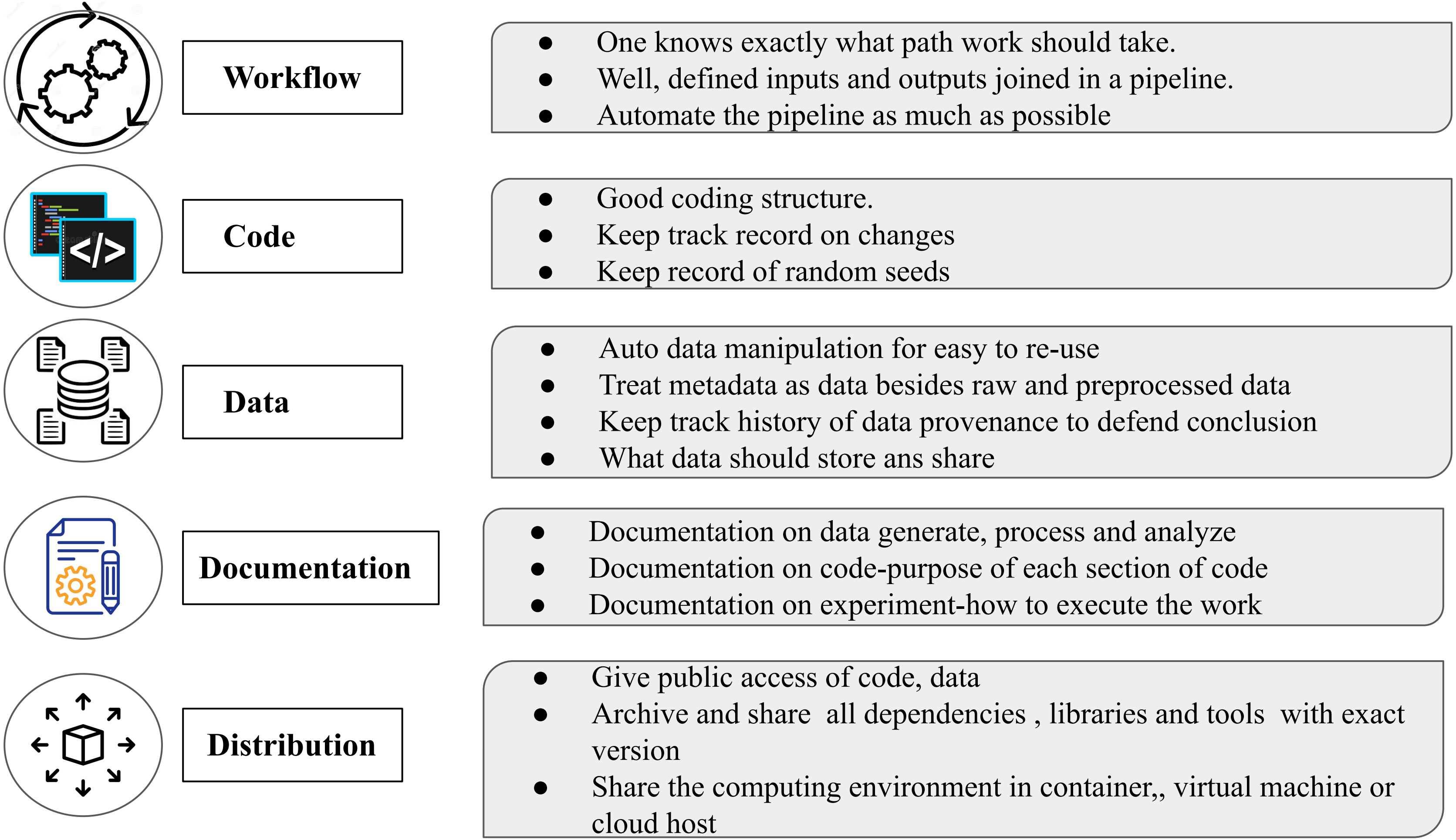
For Science

Cumulative growth (e.,g., stop wasting time what already done)

Validation of contribution (e.g., cross validation bt replication if possible)

Transparent

How to Achieve Reproducibility?



Workflow

- One knows exactly what path work should take.
- Well, defined inputs and outputs joined in a pipeline.
- Automate the pipeline as much as possible

Code

- Good coding structure.
- Keep track record on changes
- Keep record of random seeds

Data

- Auto data manipulation for easy to re-use
- Treat metadata as data besides raw and preprocessed data
- Keep track history of data provenance to defend conclusion
- What data should store ans share

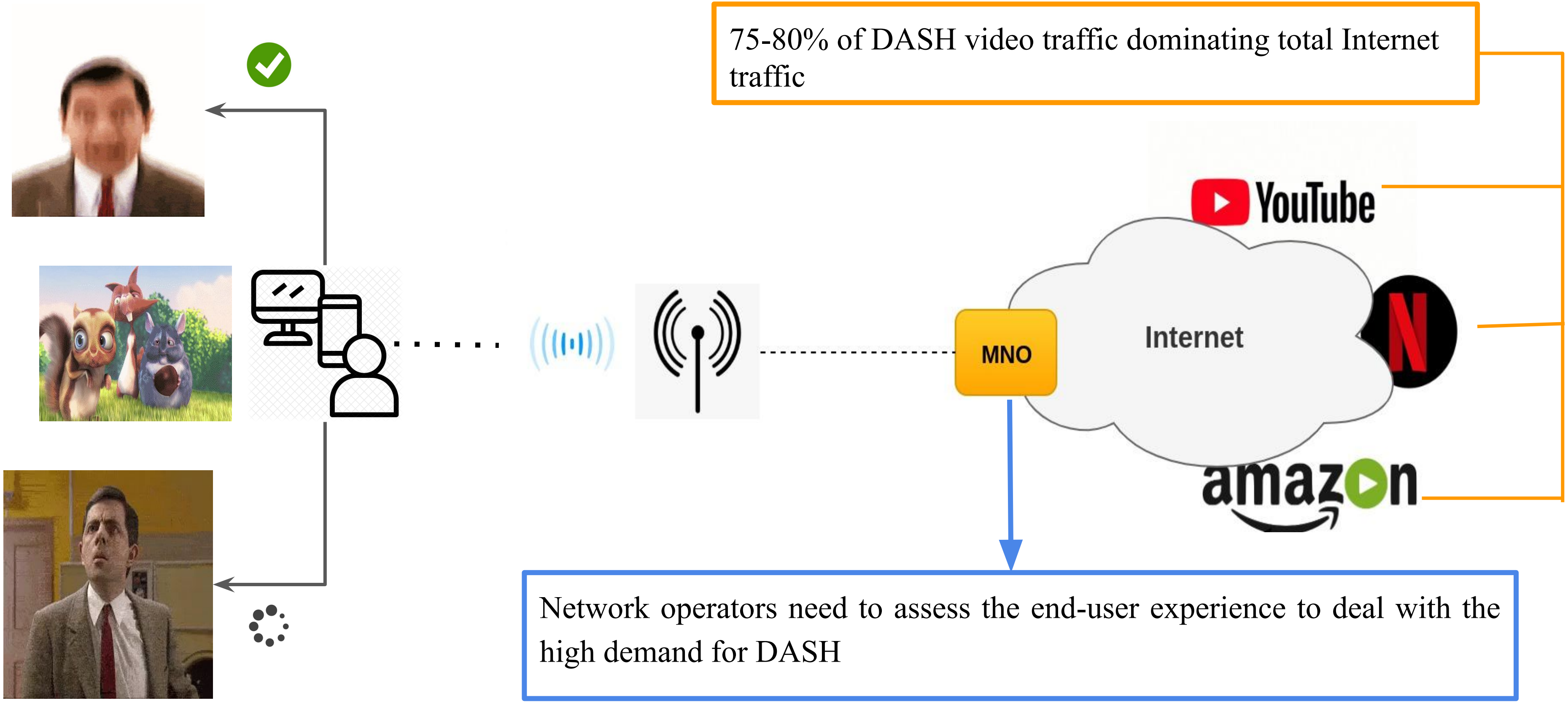
Documentation

- Documentation on data generate, process and analyze
- Documentation on code-purpose of each section of code
- Documentation on experiment-how to execute the work

Distribution

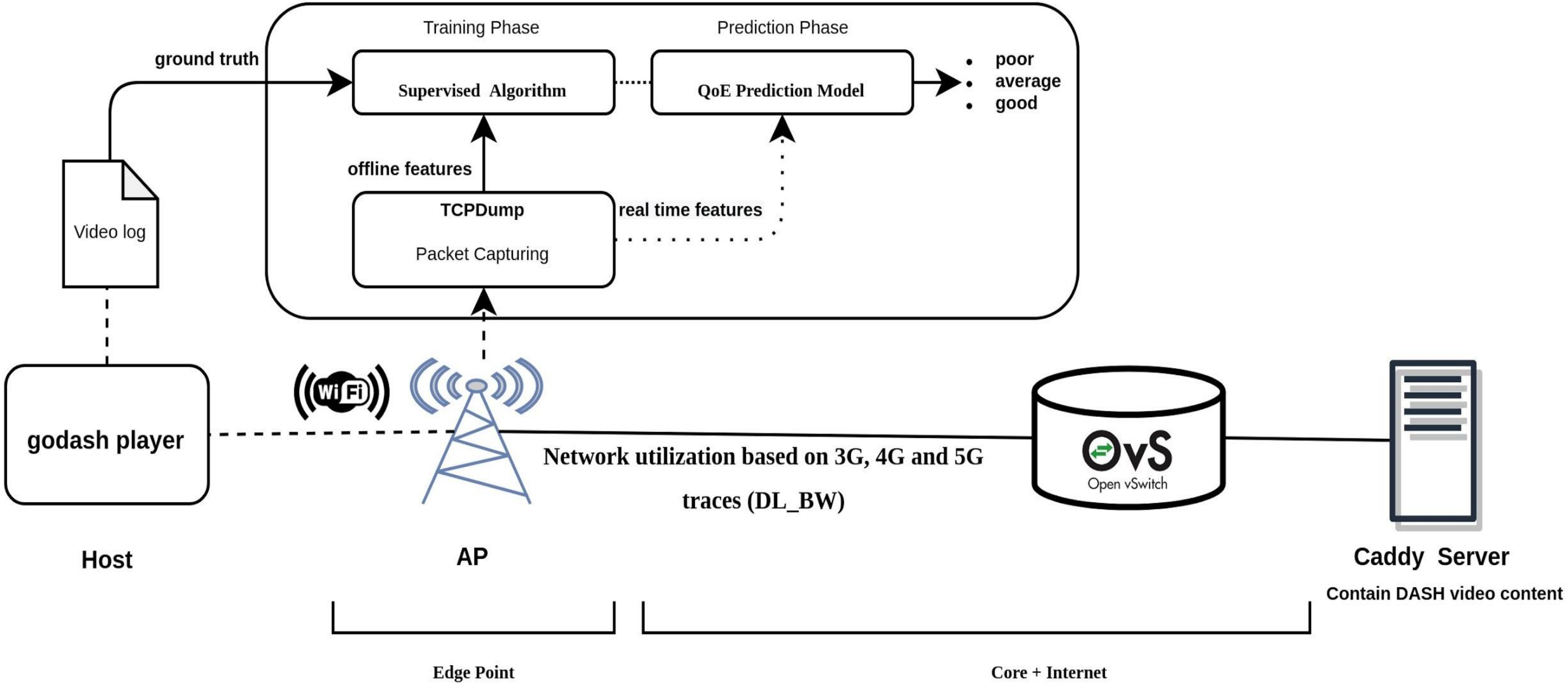
- Give public access of code, data
- Archive and share all dependencies , libraries and tools with exact version
- Share the computing environment in container,, virtual machine or cloud host

Project: *A Predictive DASH QoE Approach Based on Machine Learning at Multi-access Edge Computing*



75-80% of DASH video traffic dominating total Internet traffic

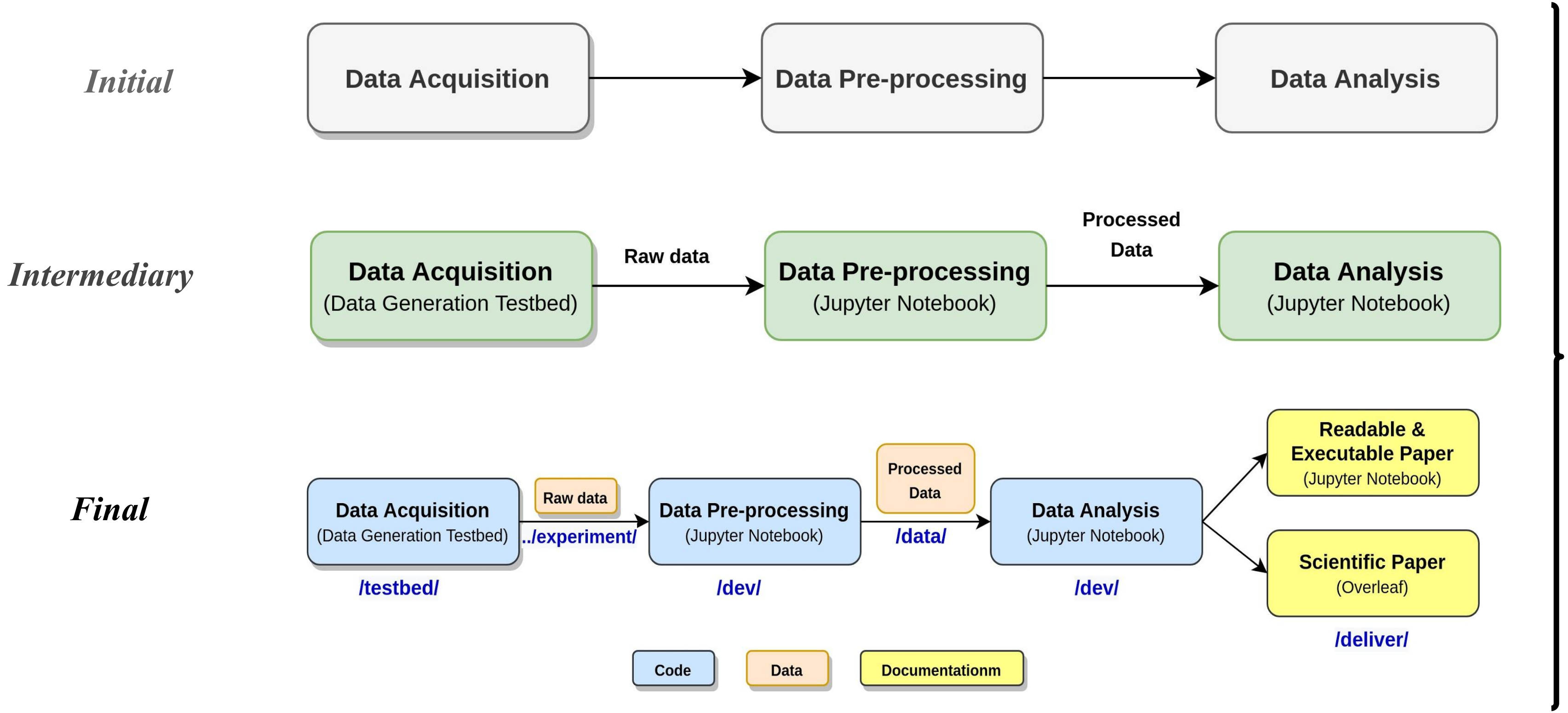
Network operators need to assess the end-user experience to deal with the high demand for DASH



Training Phase
Prediction Phase
ground truth
Supervised Algorithm
QoE Prediction Model
poor
average
good
offline features
TCPDump
Packet Capturing
real time features
Video log
godash player
Wi Fi
Network utilization based on 3G, 4G and 5G traces (DL_BW)
OvS
Open vSwitch
Host
AP
Caddy Server
Contain DASH video content
Edge Point
Core + Internet

Project Reproducibility

Workflow



Initial

Data Acquisition → Data Pre-processing → Data Analysis

Intermediary

Data Acquisition (Data Generation Testbed)

Raw data

Data Pre-processing (Jupyter Notebook)

Processed Data

Data Analysis (Jupyter Notebook)

Final

Data Acquisition (Data Generation Testbed)

/testbed/

Raw data

../experiment/

Data Pre-processing (Jupyter Notebook)

/dev/

Processed Data

/data/

Data Analysis (Jupyter Notebook)

/dev/

Readable & Executable Paper (Jupyter Notebook)

Scientific Paper (Overleaf)

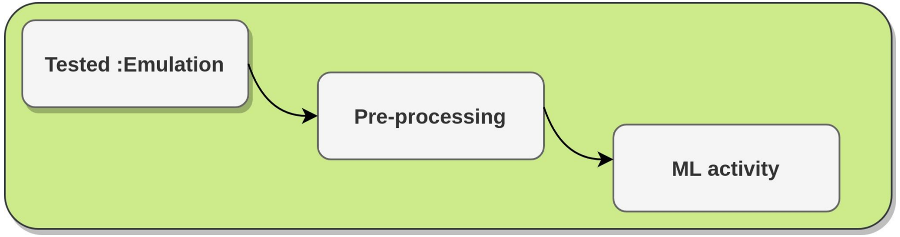
/deliver/

Code

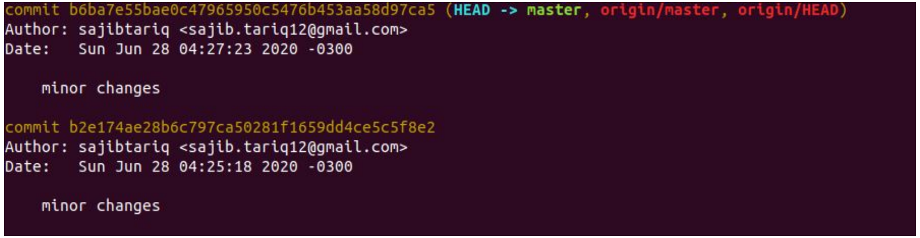
Data

Documentationm

Code



Literate Programming =
human readable (text) +
machine readable (code)



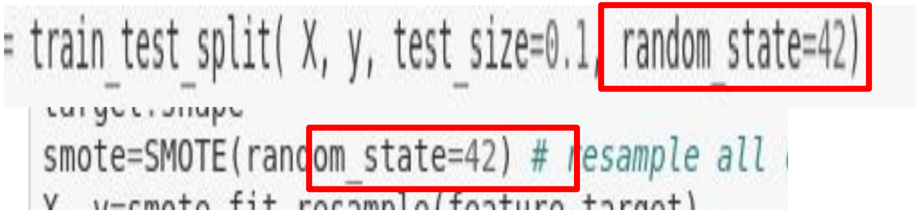
```
commit b6ba7e55bae0c47965950c5476b453aa58d97ca5 (HEAD -> master, origin/master, origin/HEAD)
Author: sajibtariq <sajib.tariq12@gmail.com>
Date:   Sun Jun 28 04:27:23 2020 -0300

    minor changes

commit b2e174ae28b6c797ca50281f1659dd4ce5c5f8e2
Author: sajibtariq <sajib.tariq12@gmail.com>
Date:   Sun Jun 28 04:25:18 2020 -0300

    minor changes
```

Track over changes =
Version control (git) + store
repository (github)



```
= train_test_split( X, y, test_size=0.1, random_state=42)
 target.shape
 smote=SMOTE(random_state=42) # resample all
 X, y=smote.fit_resample(feature,target)
```

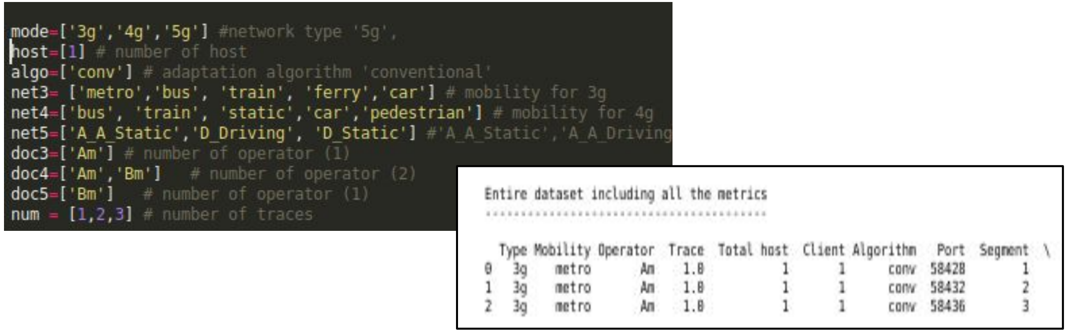
Control randomness =
Seeds (fixed)

Data

Auto Data Manipulation Script
data preprocess, plotting graph

Treat Metadata as data
input attributes of raw data generation and preprocessed dataframe header

Keep a copy of raw data
network traffic and video log for transparent on data provenance



```
mode=['3g','4g','5g'] #network type '5g',
host=[1] # number of host
algo=['conv'] # adaptation algorithm 'conventional'
net3= ['metro','bus', 'train', 'ferry','car'] # mobility for 3g
net4=['bus', 'train', 'static','car','pedestrian'] # mobility for 4g
net5=['A_A_Static','D_Driving', 'D_Static'] #'A_A_Static','A_A_Driving
doc3=['Am'] # number of operator (1)
doc4=['Am','Bm']   # number of operator (2)
doc5=['Bm']   # number of operator (1)
num = [1,2,3] # number of traces
```

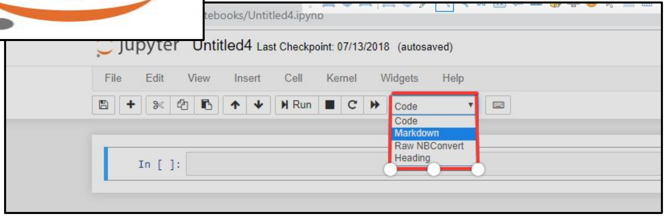
```
Entire dataset including all the metrics
****************************************

  Type Mobility Operator  Trace  Total host  Client Algorithm   Port  Segment  \
0   3g    metro       Am    1.0           1       1      conv  58428        1
1   3g    metro       Am    1.0           1       1      conv  58432        2
2   3g    metro       Am    1.0           1       1      conv  58436        3
```

Documentation

Readable & Executable Paper
Jupyter notebook using Markdown feature

NOTEBOOK

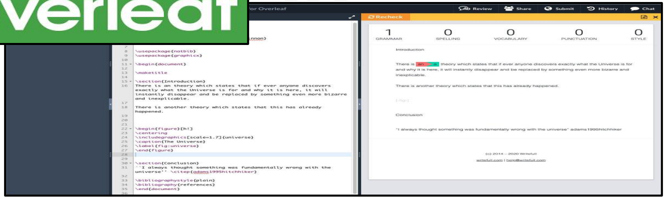


tebooks/Untitled4.ipynb

jupyter Untitled4 Last Checkpoint: 07/13/2018 (autosaved)

Scientific Paper
PDF format

PDF



```
\usepackage{natbib}
\usepackage{graphicx}

\begin{document}

\maketitle

\section{Introduction}
There is an theory which states that if ever anyone discovers
exactly what the Universe is for and why it is here, it will
instantly disappear and be replaced by something even more bizarre
and inexplicable.

There is another theory which states that this has already
happened.



\begin{figure}[h!]
\centering
\includegraphics[scale=1.7]{universe}
\caption{The Universe}
\label{fig:universe}
\end{figure}


\section{Conclusion}
``I always thought something was fundamentally wrong with the
universe'' \citep{adams1995hitchhiker}

\bibliographystyle{plain}
\bibliography{references}
\end{document}
```

Introduction

There is an a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable.

There is another theory which states that this has already happened.

[-fig-]

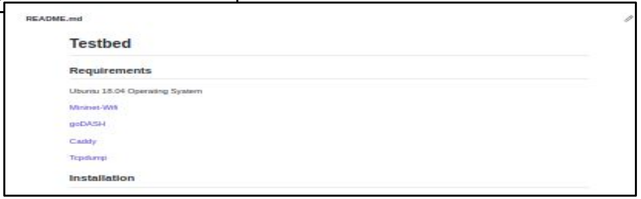
Conclusion

"I always thought something was fundamentally wrong with the universe" adams1995hitchhiker

(c) 2014 – 2020 Writefull
writefull.com | help@writefull.com

Github Readme
workflow, requirements, folder structure scheme, and how to use codes and data

README



README.md

Testbed

Requirements

Ubuntu 18.04 Operating System

Mininet-Wifi

goDASH

Caddy

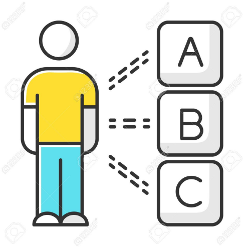
Tcpdump

Installation

Distribution

Public access to code and data

Provide multiple options to reproduce the work in the shared project



Computational environment wrapped with all everything

- ✓ Code
- ✓ Data
- ✓ Workflow
- ✓ Documentation
- ✓ Distribution

Embrace Reproducibility All Keys

Project Shared on Github

Challenges

😰 To fulfill every reproducibility criteria

😰 Make an understandable document for others

😰 Wrapping the computational environment in virtual machine

😰 Store large data set

Lessons

Current:

- ✓ Maintain documentation (code and data)
- ✓ Version control (control)
- ✓ Archiving all dependencies with the exact version

Future:

- ○ Docker to wrapping all packages as a lightweight container
- ○ Data version control and sharing (e.g., zenodo, kaggle)

Questions?

Thanks!