



USING THE DATA YOU COLLECT: ACCELERATING CYBERSECURITY APPLICATIONS WITH RAPIDS

Bianca Rhodes (Senior Full-Stack Engineer, RAPIDS)

Bartley Richardson, PhD (AI Infrastructure Manager / Senior Data Scientist)

GTC SJ 2019 (18 March 2019)

CYBERSECURITY PRESENTS UNIQUE CHALLENGES

Combination of factors lead to the need for fast iteration and quick exploration

Data velocity higher than most transactional systems and organizations

Data volume at a larger scale than most other industries

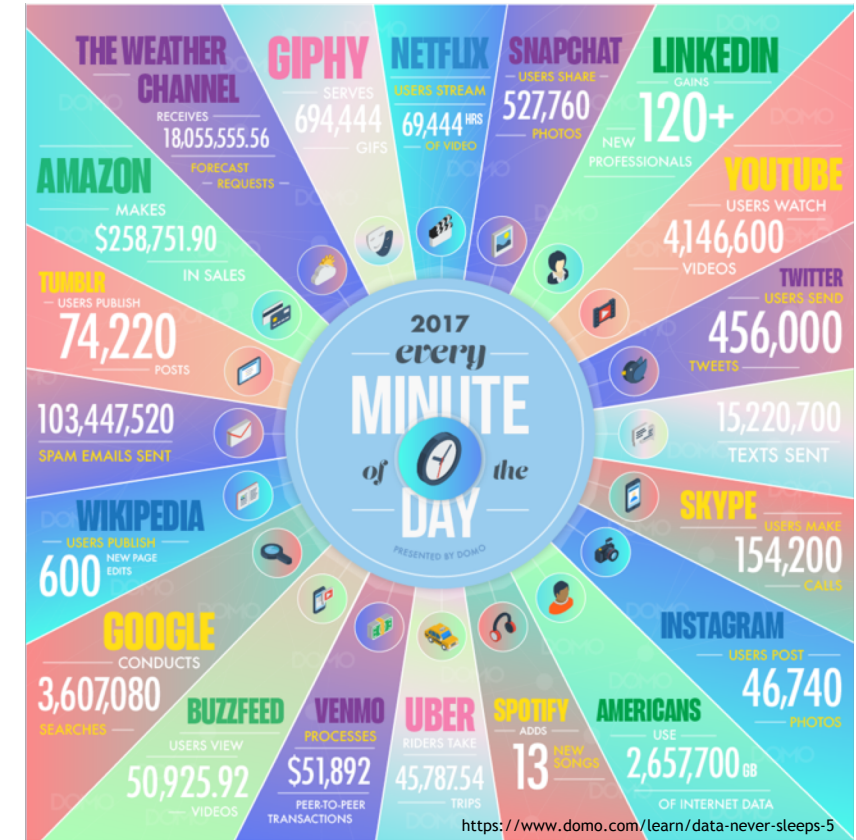
Privacy concerns abound

Decentralized IT, BYOD

User expectations

Unfilled cyber security jobs expected to reach 3.5 million by 2021¹

2.5 quintillion bytes of data created each day²



[1] <https://www.csoonline.com/article/3200024/security/cybersecurity-labor-crunch-to-hit-35-million-unfilled-jobs-by-2021.html>
[2] <https://www.domo.com/learn/data-never-sleeps-5>



RAPIDS

WHAT IS RAPIDS?

The New GPU Data Science Pipeline

rapids.ai

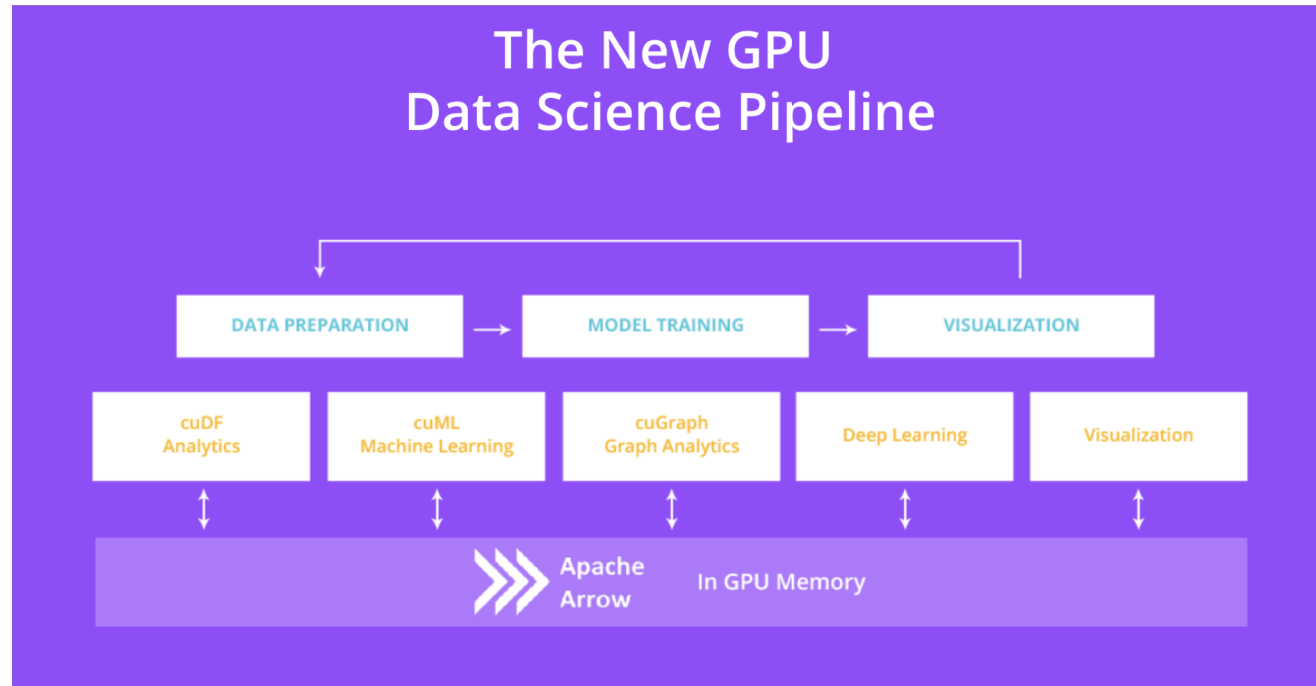
Suit of open-source, end-to-end data science tools

Built on CUDA

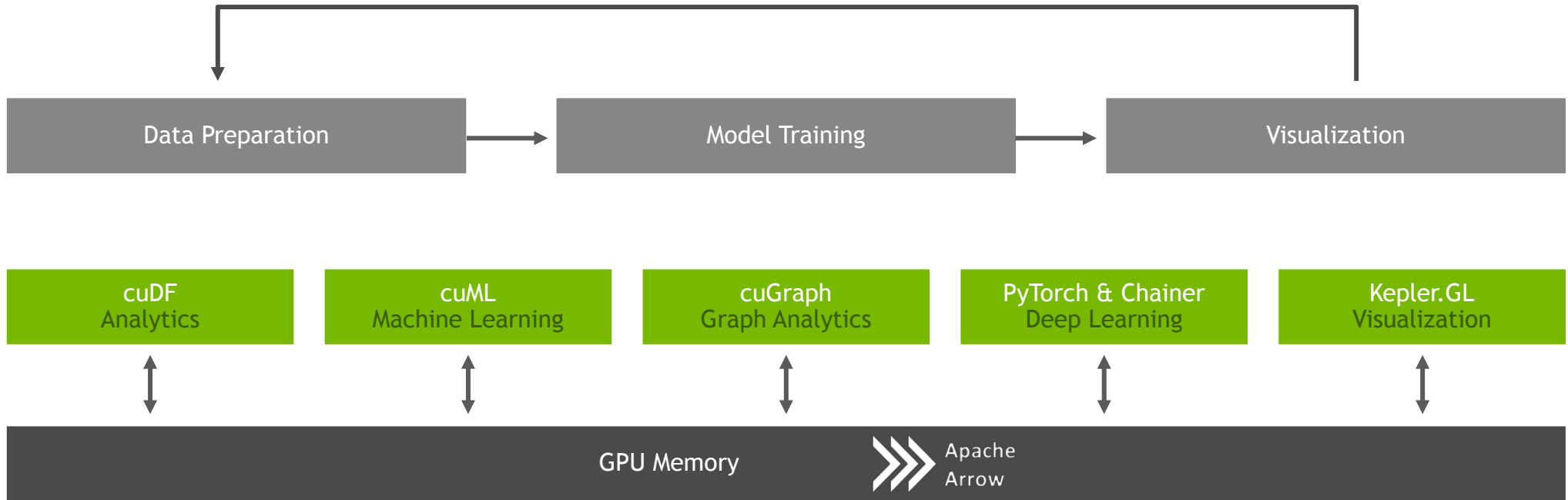
Pandas-like API for data cleaning and transformation

Scikit-learn-like API

A unifying framework for GPU data science



RAPIDS OPEN SOURCE SOFTWARE



Get RAPIDS

RAPIDS is available as conda or pip packages, docker images, and from source builds. Use the tool below to select your preferred method, packages, and environment to install RAPIDS. Certain combinations may not be possible and are dimmed automatically. Be sure to review the [prerequisites](#) section for more details about requirements to use RAPIDS.

↓ Preferred ↓ ↓ 🚧 Beta ↓

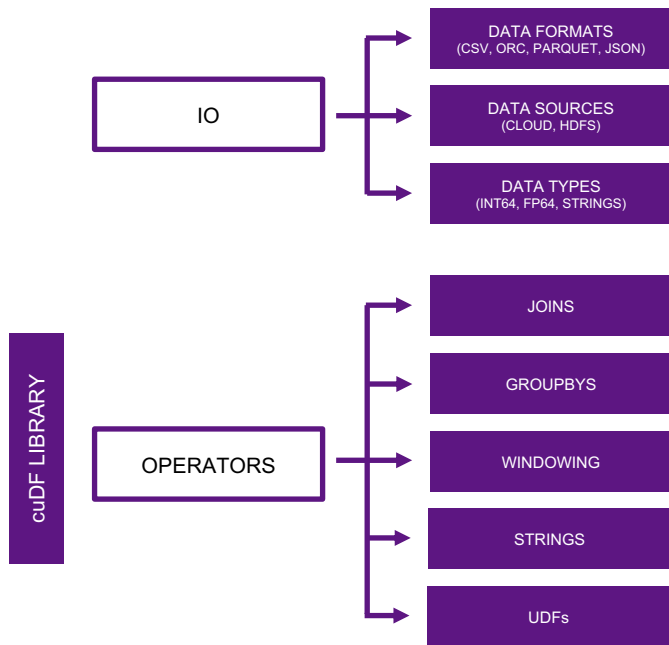
Method	Conda	Docker	Pip	Source
Packages	cuDF	cuML	cuDF & cuML	
Linux	Ubuntu 16.04	Ubuntu 18.04	CentOS 7	
Python	Python 3.6		Python 3.7	
CUDA	CUDA 9.2		CUDA 10.0	
Command	<pre>conda install -c nvidia -c rapidsai -c pytorch -c numba -c conda-forge \ -c defaults cudf=0.5 cuml=0.5 python=3.6</pre>			

COPY COMMAND 

CONDA DETAILS

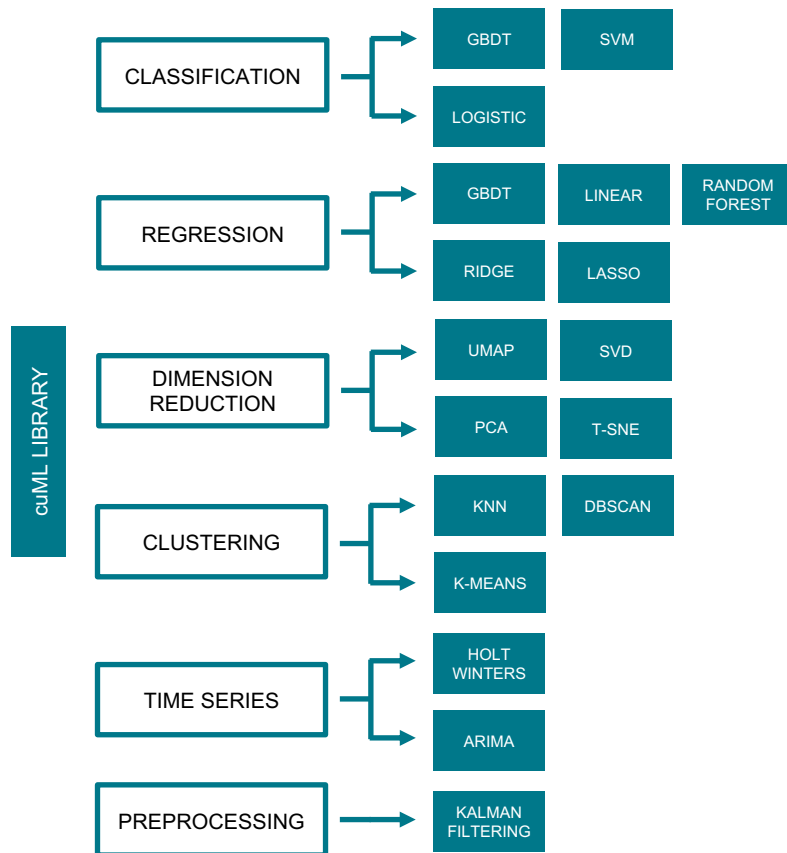
RAPIDS ROADMAP

DATA ANALYTICS



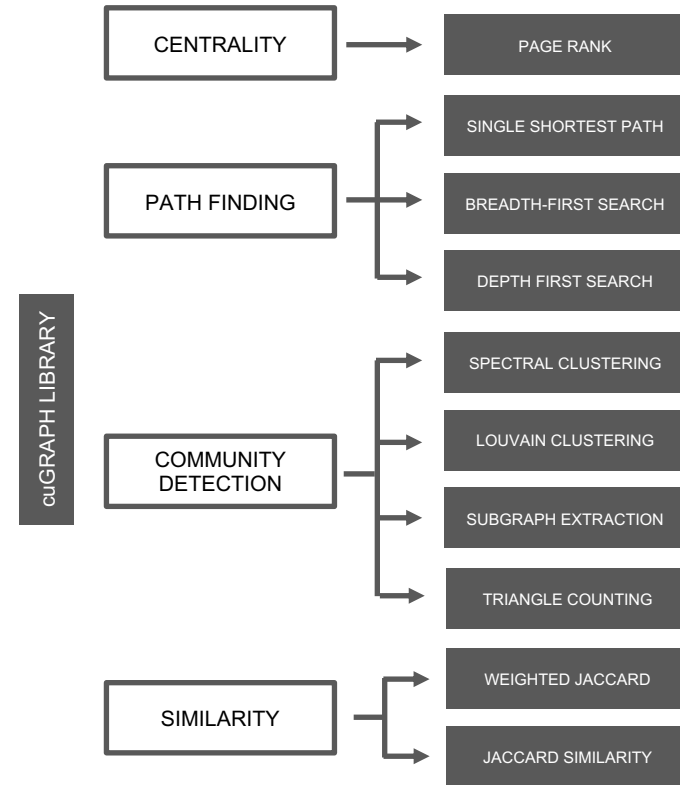
UP TO 5-15X SPEEDUP

MACHINE LEARNING



UP TO 10-20X SPEEDUP

GRAPH ANALYSIS



UP TO 100-500X SPEEDUP

RAPIDS PREREQUISITES


See more at rapids.ai

- NVIDIA Pascal™ GPU architecture or better
- CUDA [9.2](#) or [10.0](#) compatible NVIDIA driver
- Ubuntu 16.04 or 18.04

- Docker CE v18+
- [nvidia-docker](#) v2+

GOALS FOR THIS TUTORIAL

What to expect. We welcome questions along the way!

- ▶ Demonstrate how to load cybersecurity data types into RAPIDS using cuDF
- ▶ Learn how to feature engineer data with cuDF, including dealing with dataframes that have mixed column types (numeric and strings)
- ▶ Apply machine learning and graph analytics to the data
- ▶ Evaluate model results
- ▶ Visualize the output on an interactive graph
- ▶ Hands-on access to the tutorial notebooks courtesy of  Google Cloud
- ▶ Learn from you about your use cases, pain points, and necessities

START YOUR JUPYTER NOTEBOOK SERVER

Connect and start up Jupyter Notebook

Connect to your instance

```
Login:          ssh pydata@<IP>  
Password:      gtc2019
```

Activate your Conda environment

```
$source activate rapids
```

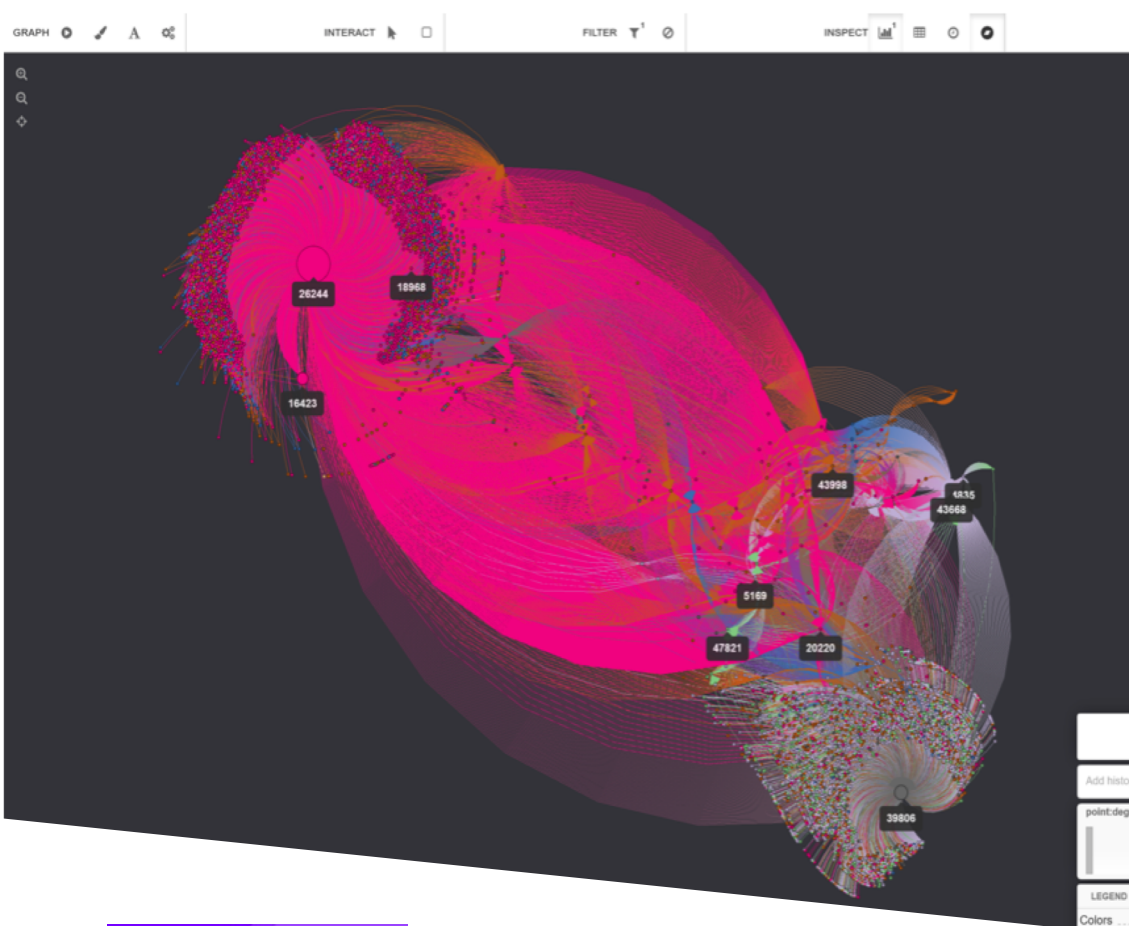
Start your Jupyter Notebook server

```
$jupyter-notebook --allow-root --ip=0.0.0.0 --port 8888  
--no-browser --NotebookApp.token='rapids'
```

Connect to your Jupyter notebook in your browser - navigate to:

```
<your.ip.address>:8888
```

You should see a Jupyter notebook directory listing



```

File Edit View Insert Cell Kernel Help Trusted Python 3
+ -> Run C Markdown

Data Investigation

Let's first see some of the data. We'll load a PCAP file in using Scapy.

In [2]: from scapy.all import *
        cap = rdpcap(DATA_PCAP)

In [3]: eth_frame = cap[3]
        ip_pkt = eth_frame.payload
        segment = ip_pkt.payload
        data = segment.payload

In [4]: eth_frame.show()

##[ Ethernet ]##
dst      = 14:cc:20:51:33:ea
src      = 30:8c:fb:2f:e4:b2
type     = 0x800
##[ IP ]##
version  = 4
ihl      = 5
tos      = 0x0
len      = 142
id       = 55213
flags    = DF
frag     = 0
ttl      = 64
proto    = 6
chksum   = 0x7ab3
src      = 192.168.1.106
dst      = 52.87.241.159
\options \
##[ TCP ]##
sport    = 40767
dport    = 443
seq      = 91355963
ack      = 679917399
dataofs  = 8
reserved = 0
flags    = PA
window   = 2549
chksum   = 0xd703
urgptr   = 0
options  = [('NOP', None), ('NOP', None), ('Timestamp', (23325908, 4049739322))]
##[ Raw ]##
load     = '\x17\x03\x01\x00 \xa0\xa9'\xe1\x7f>\x88\xccP=\xb6\x1c\xdf\xa9u\x82z\xf2G\xaaH\x82mV\xac\x97\xe7l\x19\x17\x03\x01\x000\x95\xd9\x85wV\xb9u\xf7N\xb6>\x14\xae\xcf\x9f\xbbD\xae\x11"\xc0\xedw\x1c\xfa0\xal:\x85\xad0w+...\x06\x14M\x01\x13\xf4\x88\xf4\x19\xa4\x11e:'

```



CYBER TUTORIALS USING RAPIDS WITH Google Cloud

We'll illustrate two sample use cases, each working with a different type of cyber data to answer a cybersecurity question

SESSION WRAP-UP

Now what?

Shown how you can work with multiple types of cybersecurity log data (host and network) in RAPIDS

Look for the tutorial notebooks to be posted to the RAPIDS notebooks GitHub repo shortly after GTC concludes - github.com/rapidsai/notebooks

We're interested in your cybersecurity use cases and how you'd use RAPIDS in R&D and production environments

Want to hear about your experiments and how things are going

Many RAPIDS platform and RAPIDS cyber-focused talks at GTC this year

LEARN MORE DURING GTC

Want to see detailed results using RAPIDS or speak with us more?
Check out these sessions.

Connect with the Experts: Accelerated DS and ML for Cybersecurity Applications (CE9139)

Tuesday, March 19 - 12:00-1:00pm // SJCC Hall 3 Pod A

Bianca Rhodes (NVIDIA)

Mike Geide (PUNCH Cyber Analytics Group)

Aaron Sant-Miller (Booz Allen Hamilton)

Bartley Richardson (NVIDIA)

Context-Aware Network Mapping and Asset Classification (S9802)

Thursday, March 21 - 10:00-10:50 // SJCC Room 212A

Bartley Richardson (NVIDIA)

Detecting the Unknown: Using Unsupervised Behavior Models to Expose Malicious Network Activity (S9794)

Thursday, March 21 - 3:00-3:50pm // SJCC Room 212A

Aaron Sant-Miller (Booz Allen Hamilton)

JOIN THE MOVEMENT

Everyone can help!



APACHE ARROW

<https://arrow.apache.org/>

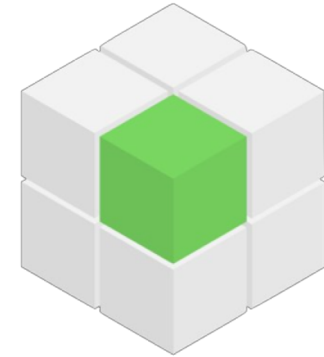
@ApacheArrow



RAPIDS

<https://rapids.ai>

@RAPIDSAI



GPU Open Analytics Initiative

<http://gpuopenanalytics.com/>

@GPUOAI

Integrations, feedback, documentation support, pull requests, new issues, or code donations welcomed!

THANK YOU TO GOOGLE CLOUD PLATFORM

Kubeflow also has a RAPIDS container!



Kubeflow

Google kindly donated the instances for
this tutorial at GTC SJ 2019!

GETTING STARTED RESOURCES

Rapids.ai

cuDF Documentation: <https://rapidsai.github.io/projects/cudf/en/latest/>

cuML Documentation: <https://rapidsai.github.io/projects/cuml/en/latest/>

Github: <https://github.com/RAPIDSai>

Twitter: @rapidsai

THANK YOU

Bianca Rhodes
brhodes@nvidia.com

Bartley Richardson, PhD
brichardson@nvidia.com

@bartleyr



Eli Fajardo

Bhargav Suryadevara

Randy Gelhausen

Nick Becker

Keith Kraus



NVIDIA[®]