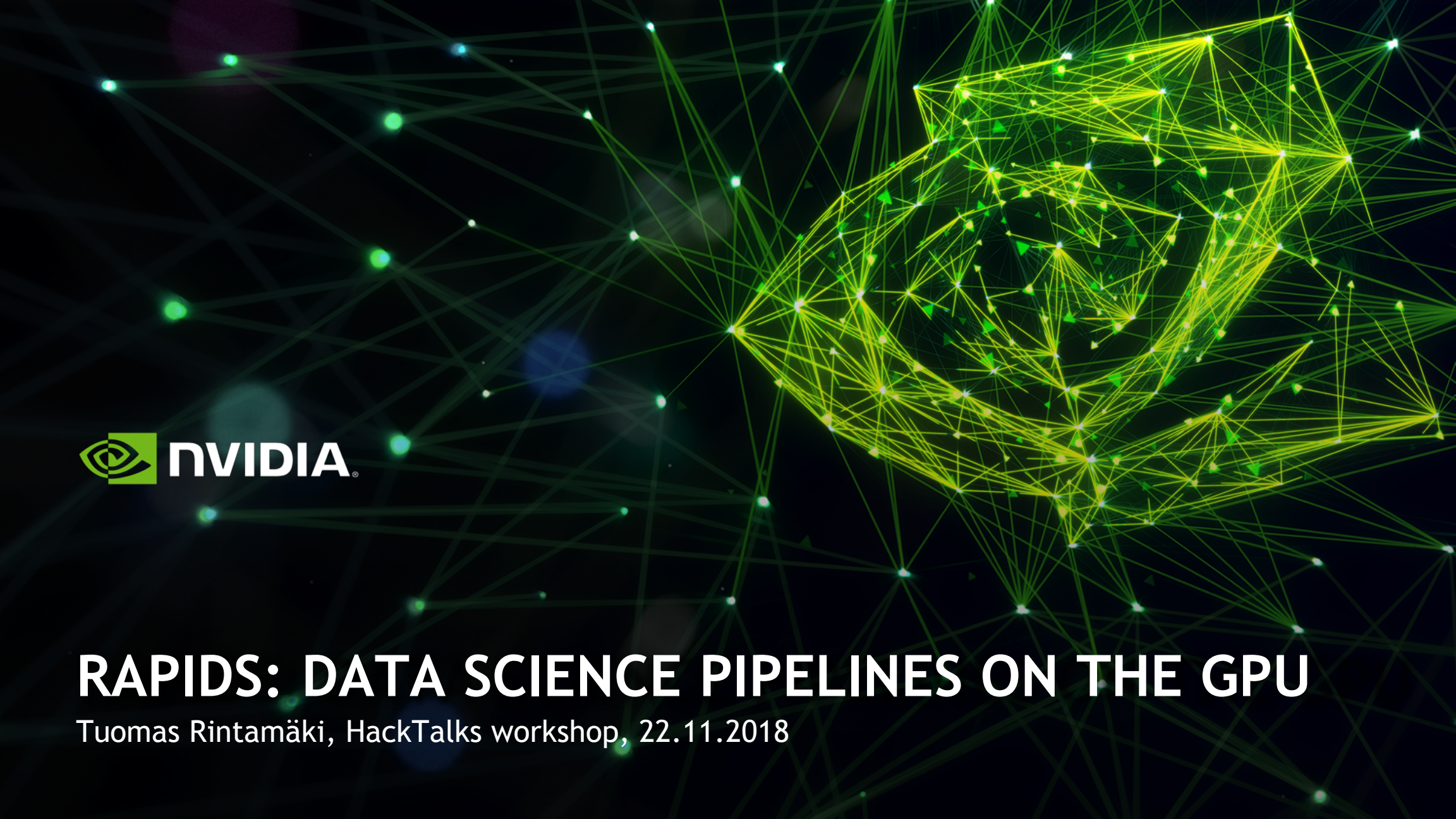


RAPIDS: DATA SCIENCE PIPELINES ON THE GPU

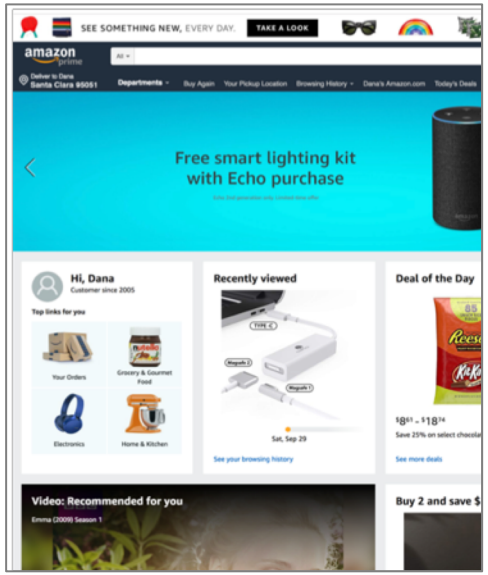
Tuomas Rintamäki, HackTalks workshop, 22.11.2018

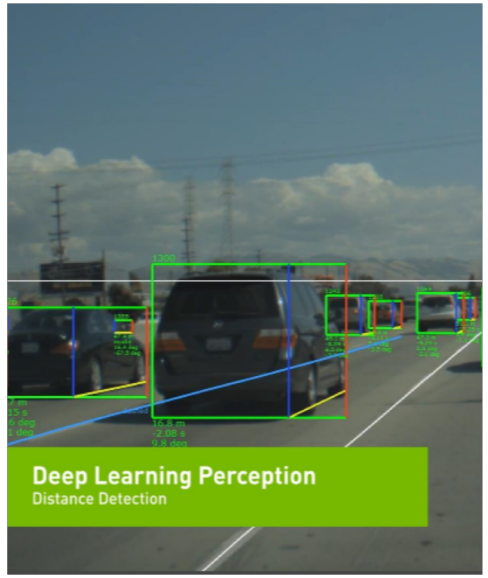
OUTLINE OF THE WORKSHOP

- A quick introduction to why RAPIDS was built and how it works
- Hands-on demo
  - Installation and setup
  - Tour of cudf
  - Tour of cuml
  - End-to-end pipeline example
- Looking to the future

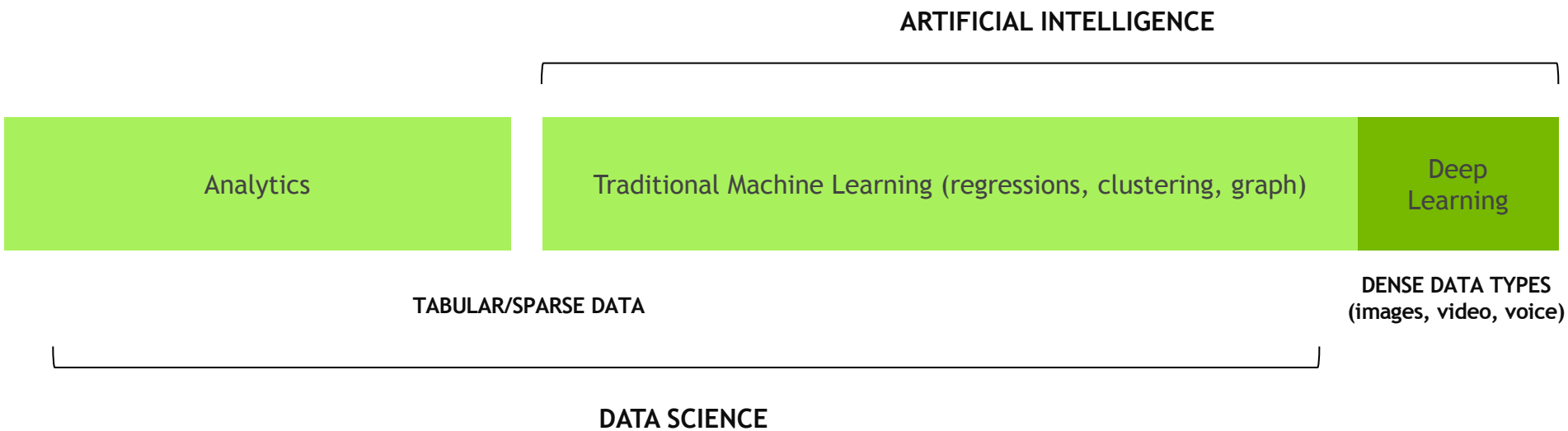
BIG DATA IS DRIVING AI

...and AI is everywhere





RAPIDS: EXTENDING DL → BIG DATA ANALYTICS



RAPIDS TARGET PERSONAS

High Performance, Easy-to-use

Data Scientist

Explore Data, Build Models

Self-serve.

Responsible for taking cleaned data, exploring it and providing insights and building models. Typically uses tools like Pandas, sci-kit learn, python, R, etc.

Pain Points

Slow data preparation and model training. Wants faster tools.

Example Titles

Data scientist, data analyst

Data Science Leader

Responsible Business ROI

Is responsible for analytics and machine learning. Provides business insights that improve business economics.

Pain Points

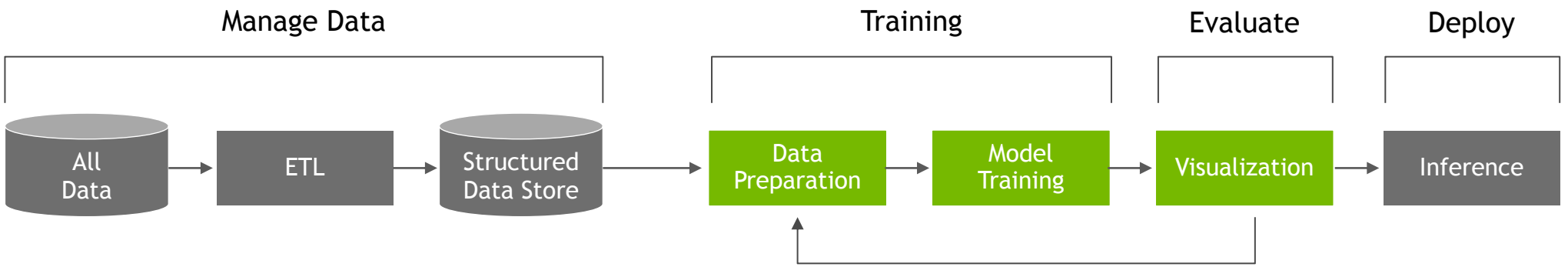
Needs faster time to insights and automated decision making.

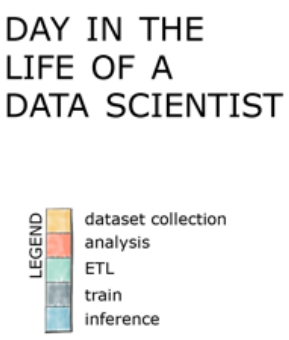
Example Titles

Director of machine learning, Director of Analytics

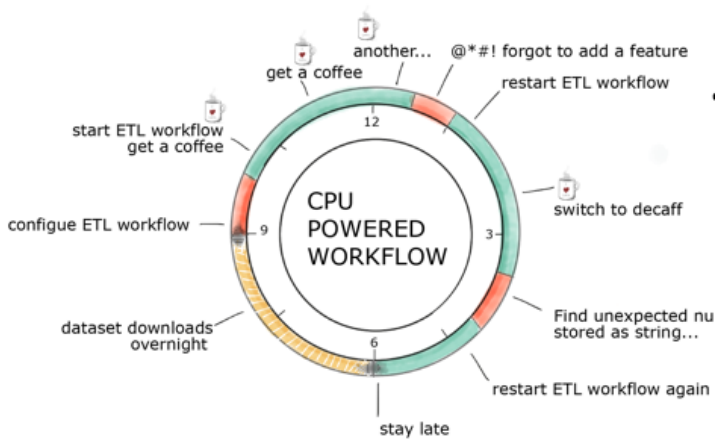
TODAY'S DATA SCIENCE PIPELINE

I.e.: HURRY UP AND WAIT



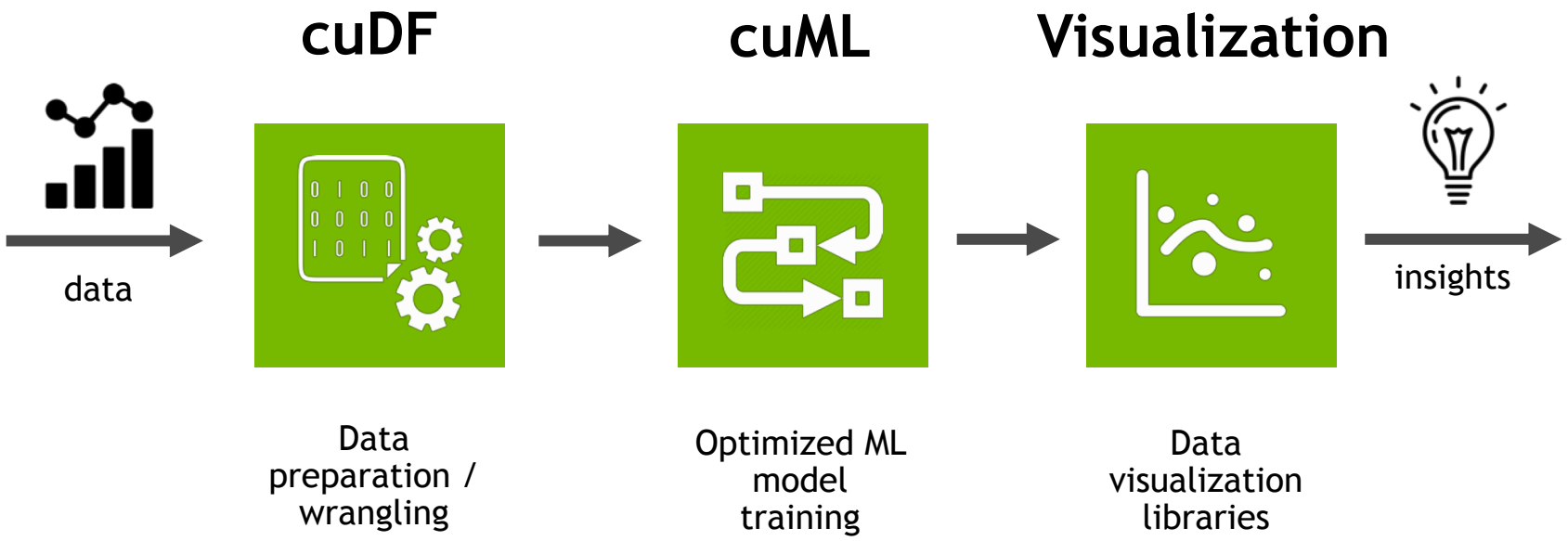


Slow Training Times for Data Scientists

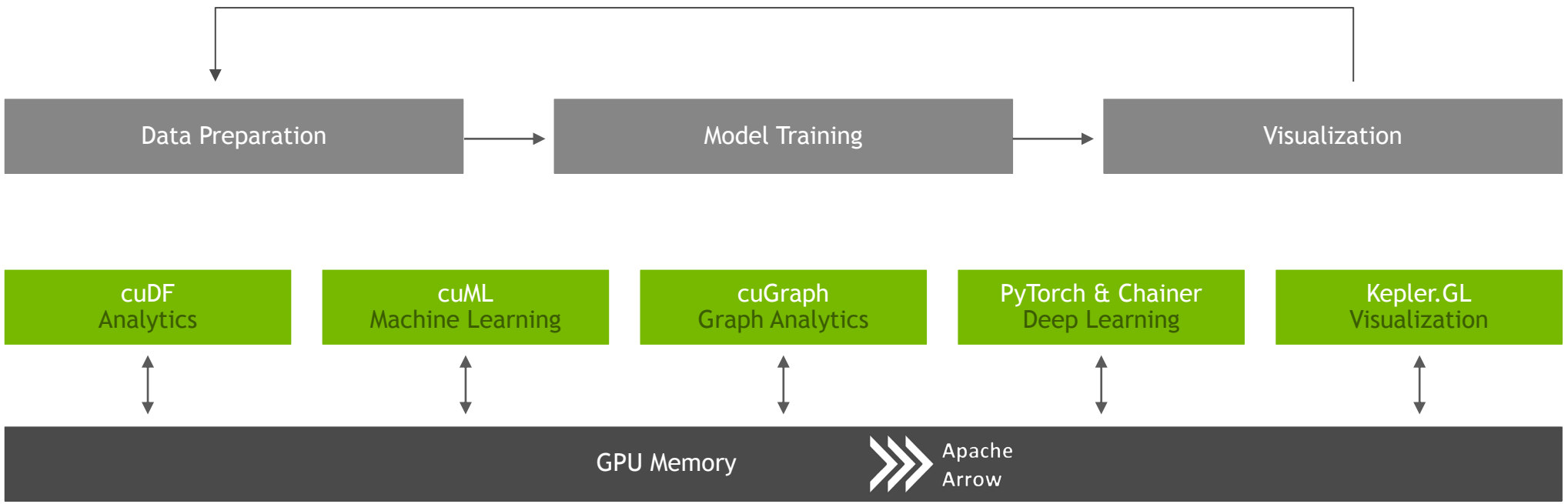


RE-IMAGINING DATA SCIENCE WORKFLOW

Open Source, End-to-end GPU-accelerated Workflow Built On CUDA

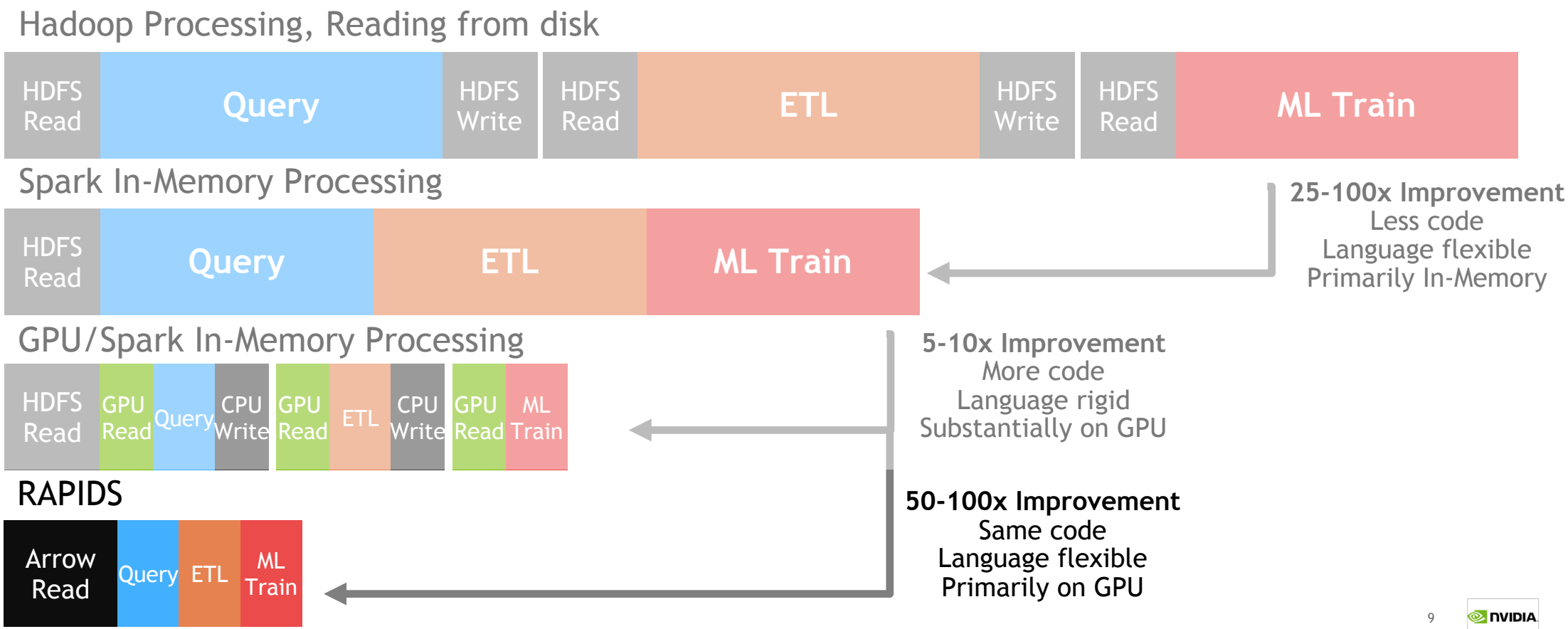


RAPIDS OPEN SOURCE SOFTWARE



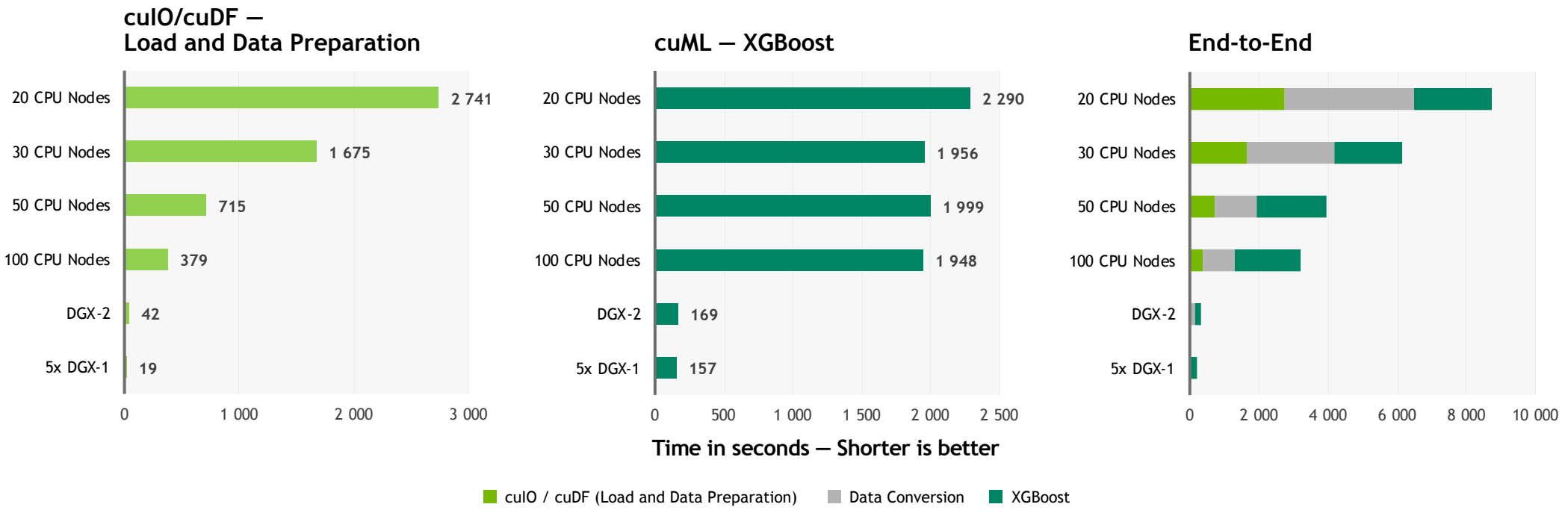
DATA PROCESSING EVOLUTION

Faster Data Access Less Data Movement



NVIDIA

FASTER SPEEDS, REAL WORLD BENEFITS



Benchmark

200GB CSV dataset; Data preparation includes joins, variable transformations.

CPU Cluster Configuration

CPU nodes (61 GiB of memory, 8 vCPUs, 64-bit platform), Apache Spark

DGX Cluster Configuration

5x DGX-1 on InfiniBand network

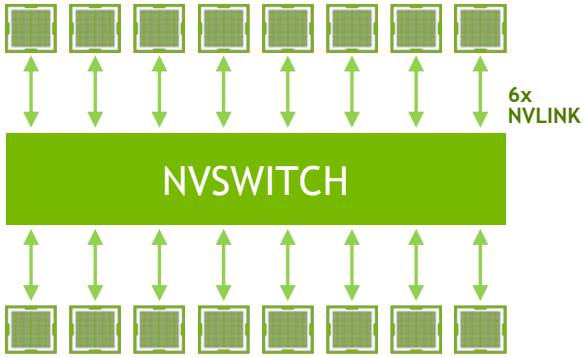
PILLARS OF RAPIDS PERFORMANCE

DASK

CUDA Architecture

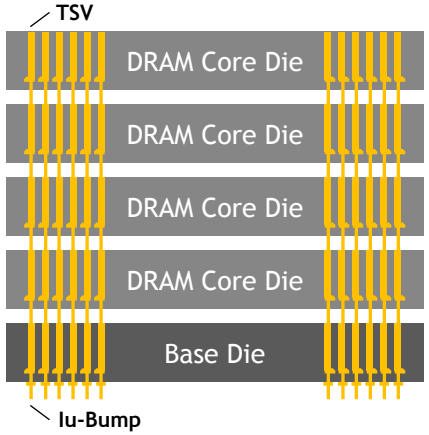
Massively parallel processing

NVLink/NVSwitch



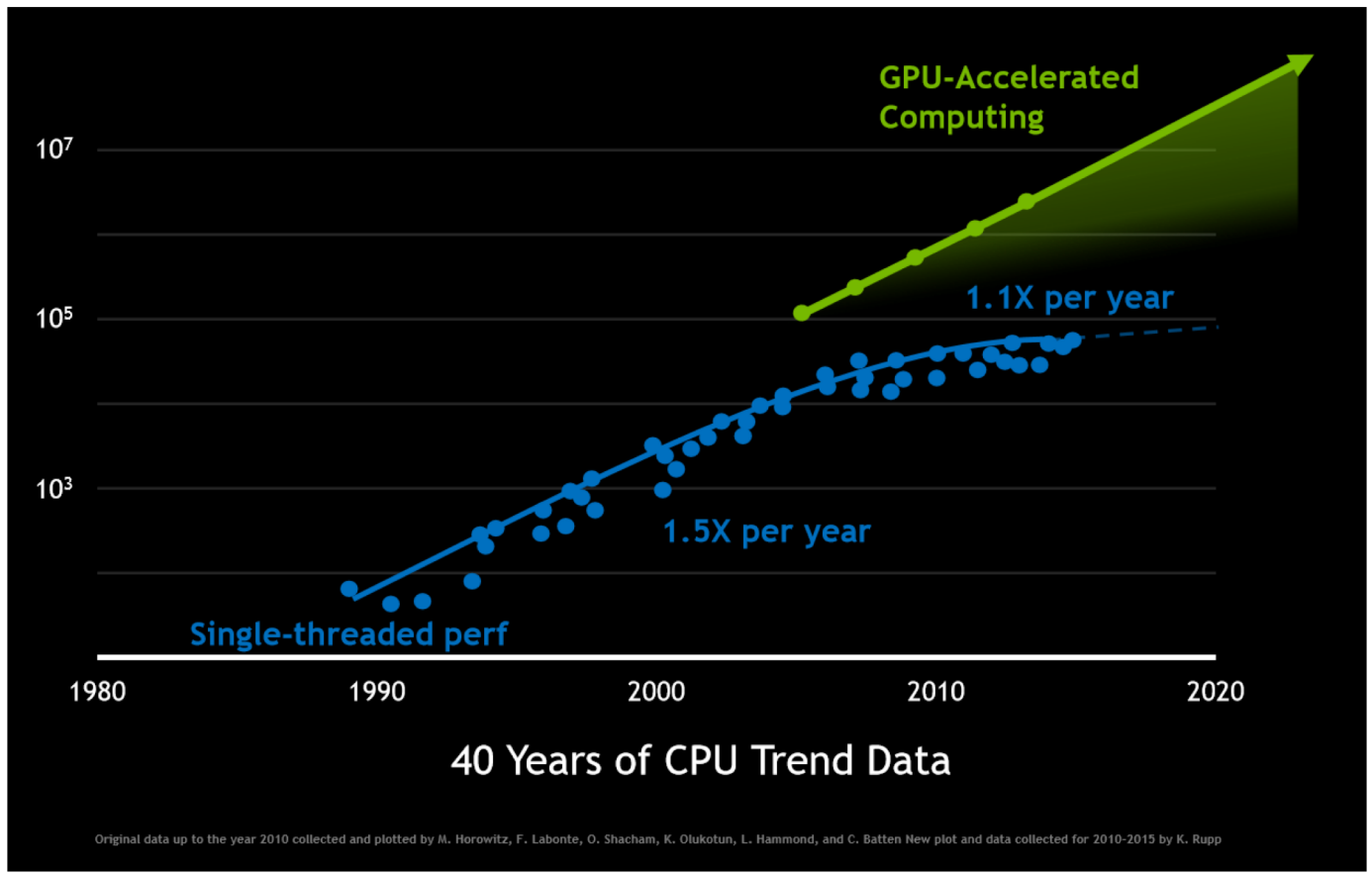
High speed connecting between GPUs for distributed algorithms

Memory Architecture



Large virtual GPU memory, high-speed memory

RISE OF GPU COMPUTING



NVIDIA

IMPROVING DEMAND FORECASTS

Accurate demand forecasting is a critical but challenging science for retailers requiring massive amounts of data and compute cycles. Walmart is optimizing machine learning with NVIDIA RAPIDS open-source software on GPUs.

GPUs deliver 50x faster processing speed allowing Walmart to benefit from more sophisticated algorithms, reduce forecasting errors, and increase the efficiency of its supply chain.

NVIDIA

Walmart

THE RAPIDS VALUE PROPOSITION

High Performance, Easy-to-use

Data Scientist

Reduced Data Preparation and Training Time

Drastically improve your productivity with near-interactive data science

Hassle-Free Integration

Accelerate your Python data science toolchain with minimal code changes and no new tools to learn

Open Source

Customizable, extensible, interoperable — the open-source software is supported by NVIDIA and built on Apache Arrow

Data Science Leader

Top Model Accuracy

Increase machine learning model accuracy by iterating on models faster and deploying them more frequently

Increased Data Scientist Productivity

Reduce training time, allow data scientists to be more productive

TCO Reduction

Decrease the server costs, footprint, power consumption of your ML workloads reducing the TCO

RAPIDS DOWNLOAD AND DEPLOY

Source available on Github | Container available on NGC and Dockerhub | PIP available at a later date

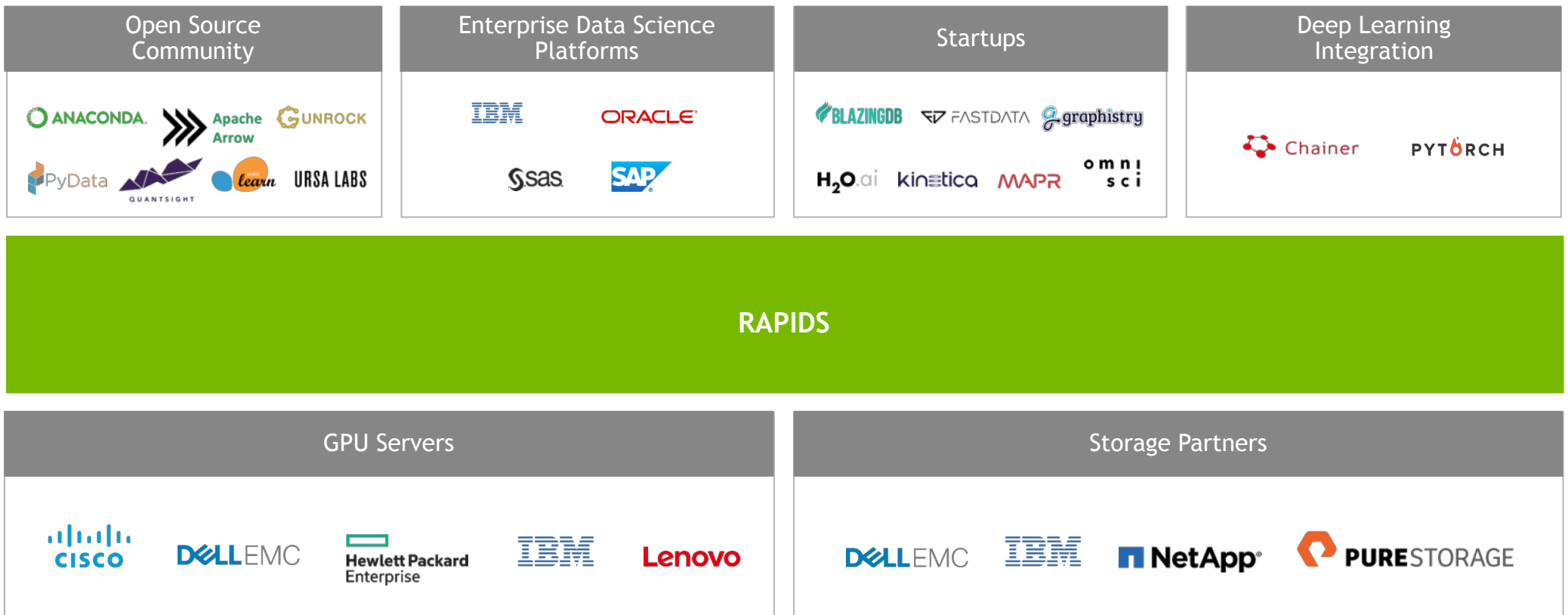
GitHub NGC docker CONDA python Package Index

Source code, libraries, packages

On-premises

Cloud

THE RAPIDS ECOSYSTEM



NVIDIA HELSINKI IS HIRING

Drop your application to jobs-finland@nvidia.com

We are looking for

- Developers to build distributed systems for data ingestion, training and evaluation of deep learning and data science models at TB to PB scale
- Developers to optimize and deploy models for self-driving cars and embedded devices
- DNN and data science researchers to improve on the state-of-the-art models
- Also interns in these fields for next summer


LOOKING TO THE
FUTURE

GPU DATAFRAME

Next few months

- Continue improving performance and functionality
  - Single GPU
  - Single node, multi GPU
  - Multi node, multi GPU
- String Support
  - Support for specific “string” dtype with GPU-accelerated functionality similar to Pandas
- Accelerated Data Loading
  - File formats: CSV, Parquet, ORC – to start

CUML

In development right now or roadmapped

Kalman filter

Collaborative filtering

Generalized linear models

Random Forests

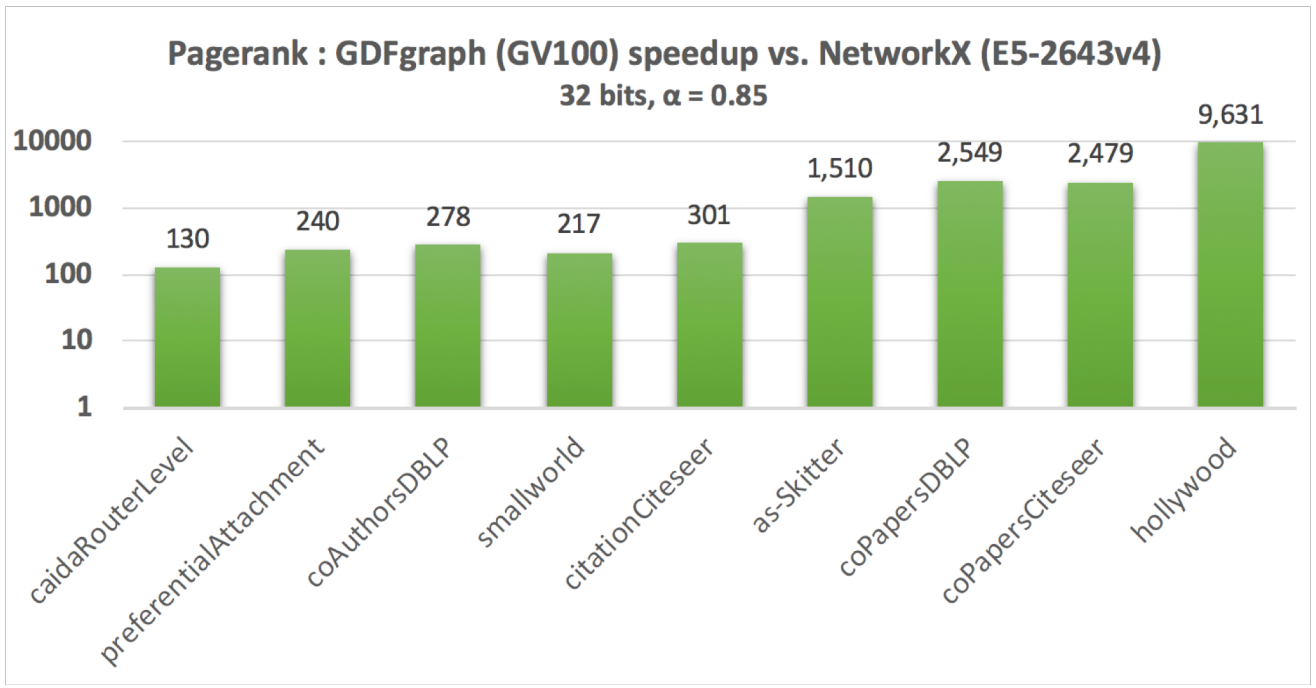
ARIMA

Holt-Winters smoothing

Bayesian methods

CUGRAPH

GPU-Accelerated Graph Analytics Library



Coming Soon:
PageRank, BFS, Triangle Counting with NetworkX-like API

JOIN THE MOVEMENT

Everyone Can Help!

APACHE ARROW

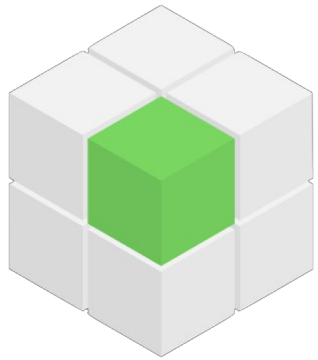
https://arrow.apache.org/

@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
Tuomas Rintamäki
@tuomars
NVIDIA