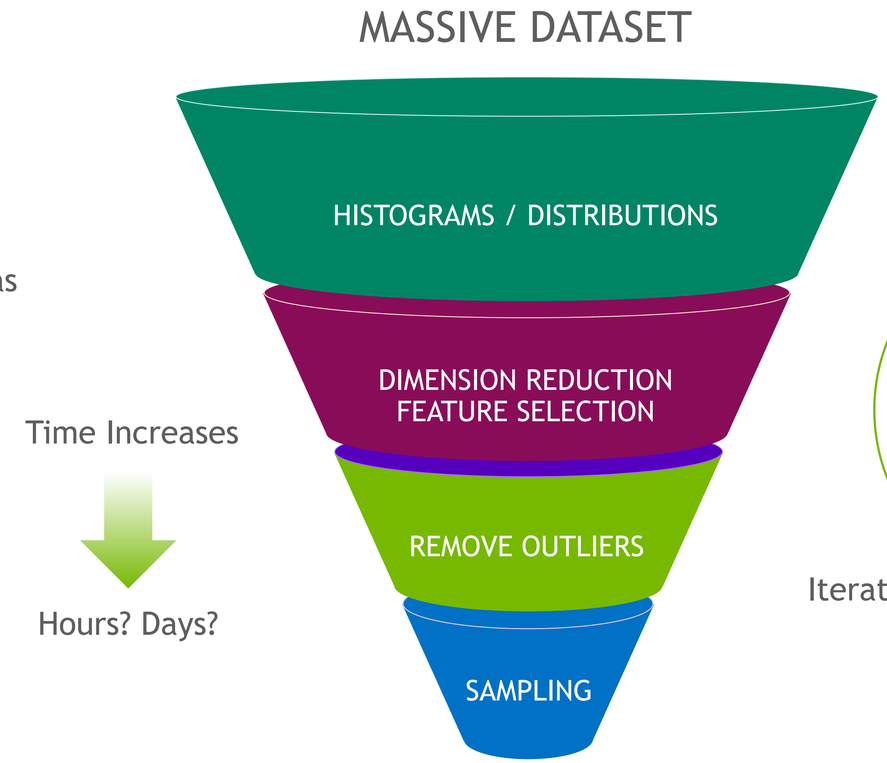
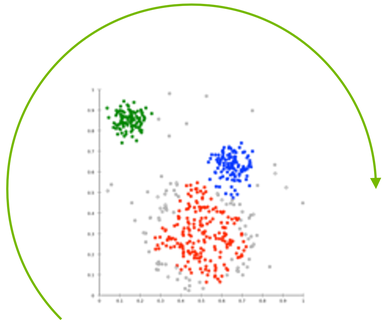

NVIDIA
cuML

Problem

Data Sizes Continue to Grow

Better to start with as much data as possible and explore / preprocess to scale to performance needs.





Iterate. Cross Validate & Grid Search.
Iterate Some More.

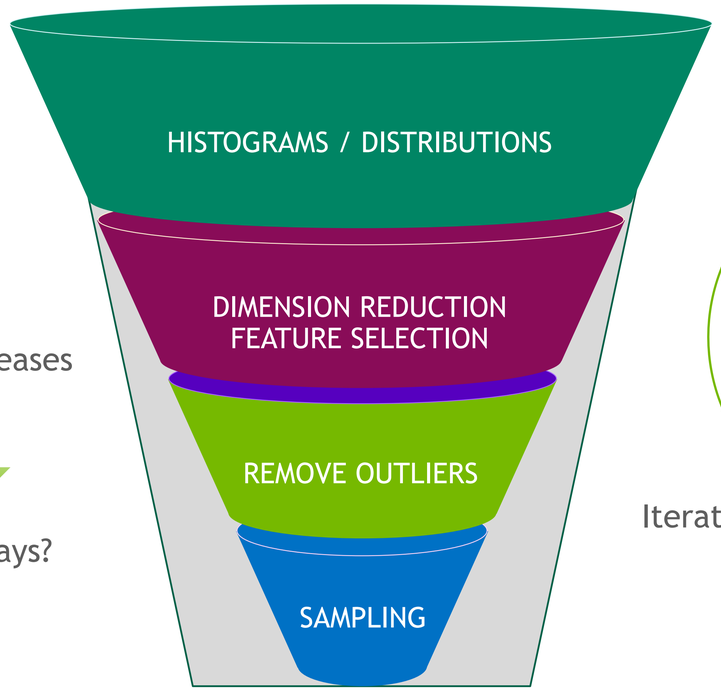
Meet Reasonable Speed vs Accuracy Trade-off

Problem

Data Sizes Continue to Grow

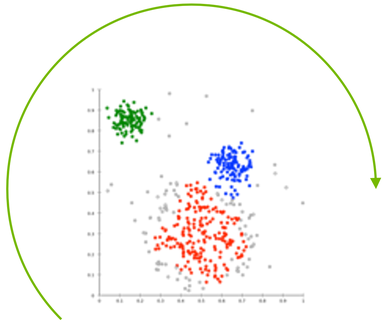
RAPIDS helps simplify that process

Better to start with as much data as possible and explore / preprocess to scale to performance needs.



Time Increases

Hours? Days?



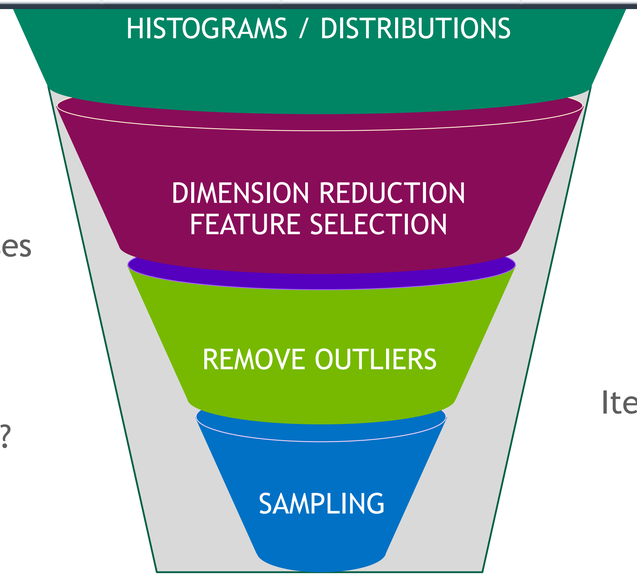
Iterate. Cross Validate & Grid Search.
Iterate Some More.

Meet Reasonable Speed vs Accuracy Trade-off

Mike McCarty Capital One, GTC DC 2019 talk on XGBoost Training

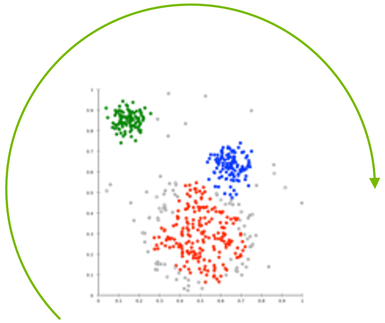
dataset	# nodes	node type	# GPUs	training size	test size	time	cost*
40 GB	1	x1.32xlarge	0	13 GB	6 GB	2 - 3 hrs	$26.68 - $40.01
40 GB	1	p3dn.24xlarge	8	13 GB	6 GB	2m 1s	$1.02
40 GB	1	p3.16xlarge	8	13 GB	6 GB	2m 23s	$0.82
1.15 TB	12**	p3dn.24xlarge	96	288 GB	127 GB	52m 21s	$326.85
1.15 TB	2**	p3dn.24xlarge	16	288 GB	127 GB	22m 47s	$54.47

Better to start with as much data as possible and explore / preprocess to scale to performance needs.



Time Increases

Hours? Days?



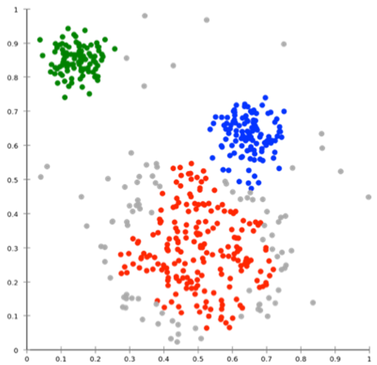
Iterate. Cross Validate & Grid Search.
Iterate Some More.

Meet Reasonable Speed vs Accuracy Trade-off

X1 32xlarge is 128 vCPUs and up to 1.9TB of memory

cuML - Algorithms

GPU-accelerated Scikit-Learn



Cross Validation

Hyper-parameter Tuning

Classification / Regression	Decision Trees / **Random Forests** Linear/Lasso/Ridge/**LARS**/ElasticNet Regression Logistic Regression K-Nearest Neighbors (**exact or approximate**) Support Vector Machine Classification and Regression Naive Bayes
Inference	Random Forest / GBDT Inference (FIL)
Preprocessing	Text vectorization (TF-IDF / Count) Target Encoding Cross-validation / splitting
Clustering Decomposition & Dimensionality Reduction	K-Means **DBSCAN** Spectral Clustering Principal Components (including iPCA) Singular Value Decomposition UMAP Spectral Embedding T-SNE
Time Series	Holt-Winters Seasonal ARIMA / Auto ARIMA

New Developments in ML

RAPIDS cuML

- Multi-node, multi-GPU support for DBSCAN, UMAP, kNN, Naive Bayes, and more
- Sparse data support for many core estimators
- Scikit-learn compatible preprocessing support plus target encoding
- Expanded features for SVM, forest inference, and random forests

dmlc XGBoost

- Seamless, drop-in GPU acceleration
- Zero-copy data loading from cuDF data frames
- Native Dask API for MNMG scaling
- Improved experience on k8s

SHAP

- GPU-accelerated model explainability
- Speedups of 20x to 300x for explanations
- Specialized GPUTreeSHAP for XGBoost and RF
- Black-box SHAP for any ML pipeline
- See the GPUTreeSHAP blog

Real-World Speed-Ups

NVIDIA Kaggle Grand Masters win with RAPIDS

Booking.com

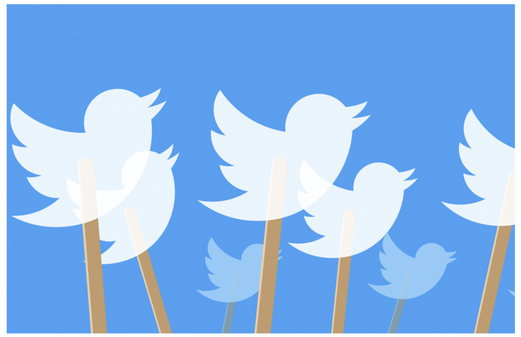
Booking.com RecSys 01/28/2021

1st

The Royal Australian and New Zealand College of Radiologists®

RANZCR CLiP 3/16/21

1st



Twitter RecSys 06/15/2020

1st

black box

NeurIPS Black Box Optimization 10/15/2020

2nd

TReNDS

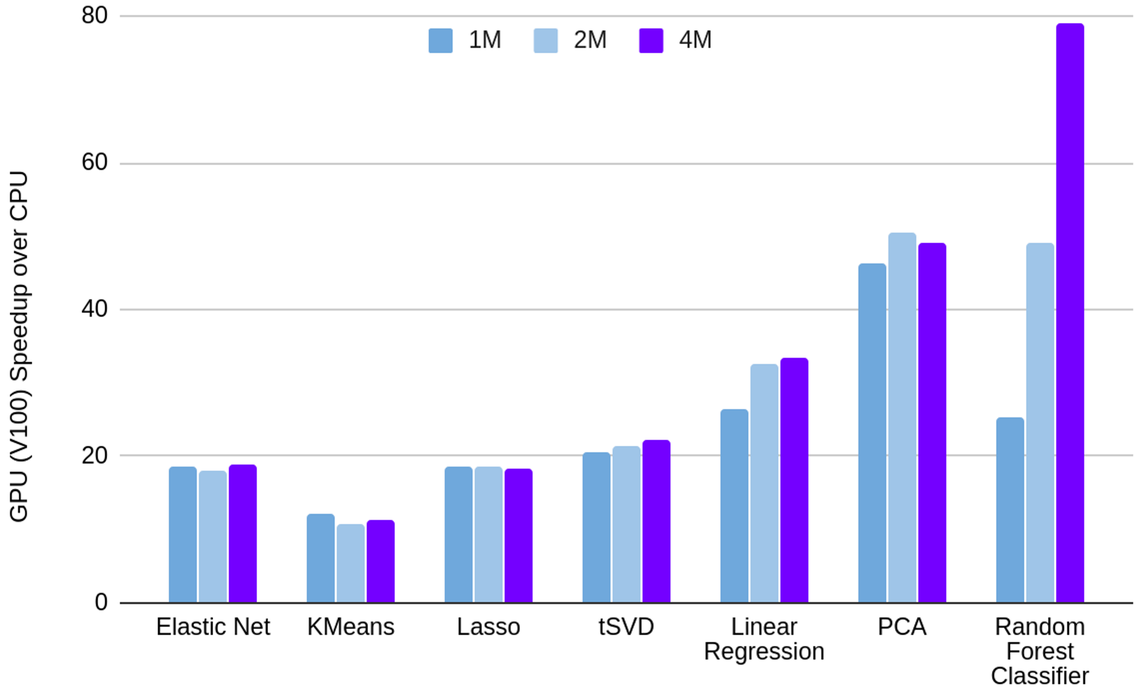
TReNDS Neuroimaging 6/29/2020

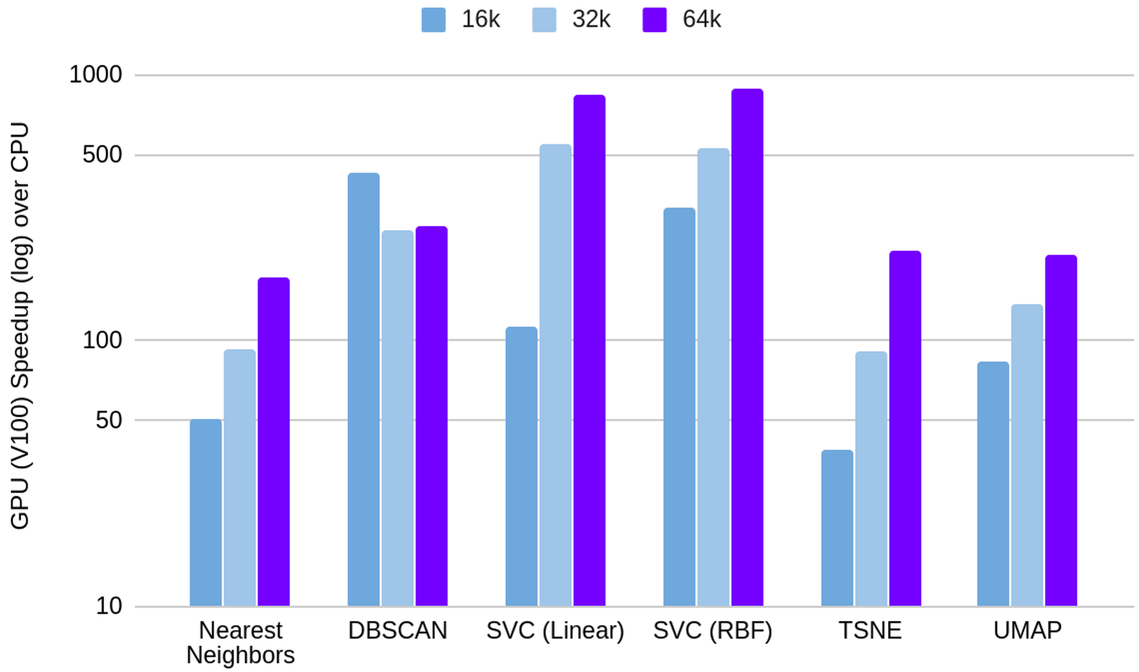
2nd

cuML

Accelerated Machine Learning with a scikit-learn API

30+ GPU-Accelerated Algorithms & Growing





1x V100 vs. 2x 20 Core CPUs (DGX-1, RAPIDS 0.15)

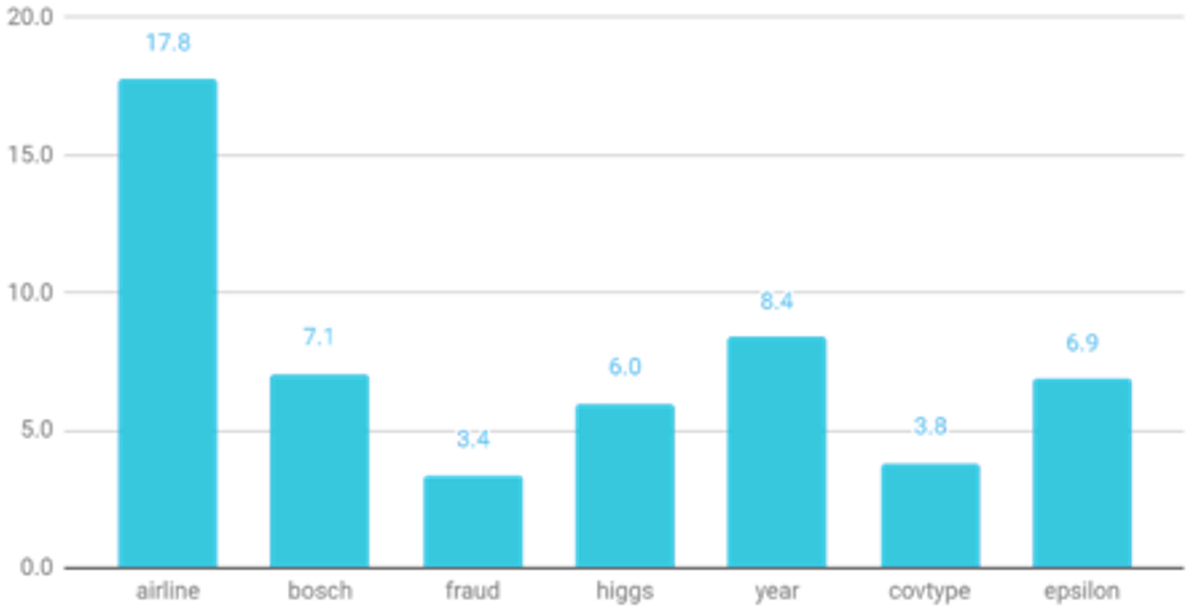
NVIDIA

XGBoost + RAPIDS: Better Together

“XGBoost is All You Need” – Bojan Tunguz, 4x Kaggle Grandmaster

- RAPIDS comes paired with XGBoost 1.6.0
- XGBoost now builds on the GoAI interface standards to provide zero-copy data import from cuDF, cuPy, Numba, PyTorch and more
- Official Dask API makes it easy to scale to multiple nodes or multiple GPUs
- GPU tree builder delivers huge perf gains
- Now supports Learning to Rank, categorical variables, and SHAP Explainability
- Use models directly in Triton for high-performance inference

All RAPIDS changes are integrated upstream and provided to all XGBoost users – via pypi or RAPIDS conda



XGBoost speedup on GPUs comparing a single NVIDIA V100 GPU to a dual 20-core Intel Xeon E5-2698 server


NVIDIA
WORKFLOW