From Zero to RAPIDS: Accelerating Data Science and Machine Learning Workflows on NVIDIA GPUs

Marty Kandes, Ph.D.
Computational & Data Science Research Specialist
High-Performance Computing User Services Group
San Diego Supercomputer Center
University of California, San Diego

HPC User Training Series
Friday, January 31st, 2020
2:00PM - 3:00PM PT

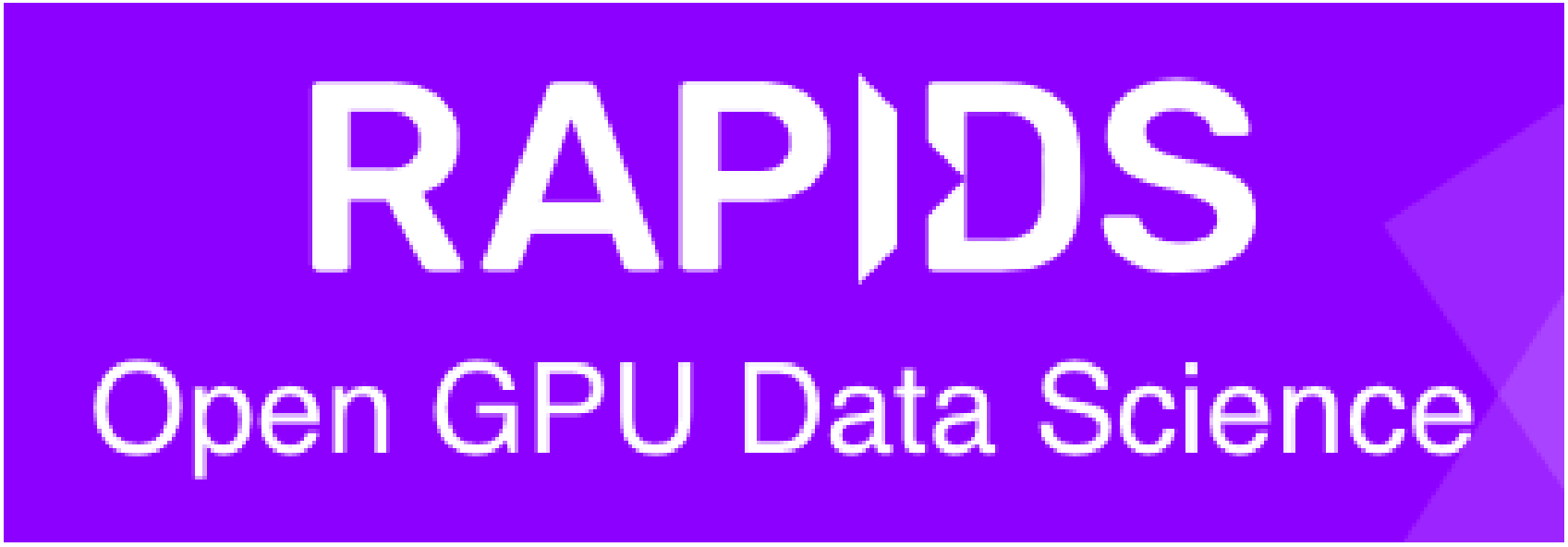
About Me

- **Work**: HPC User Services Group @ SDSC, Comet
- **Background**: Computational Physics, Applied Math, HPC
- **Disclaimer**: *Not* a Data Science or Machine Learning expert

From Zero to RAPIDS: An Overview

- A Quick Overview of RAPIDS
- Jupyter Notebook Demo
- Fannie Mae Single-Family Loan Performance Data
- Additional Resources and References

What is RAPIDS?



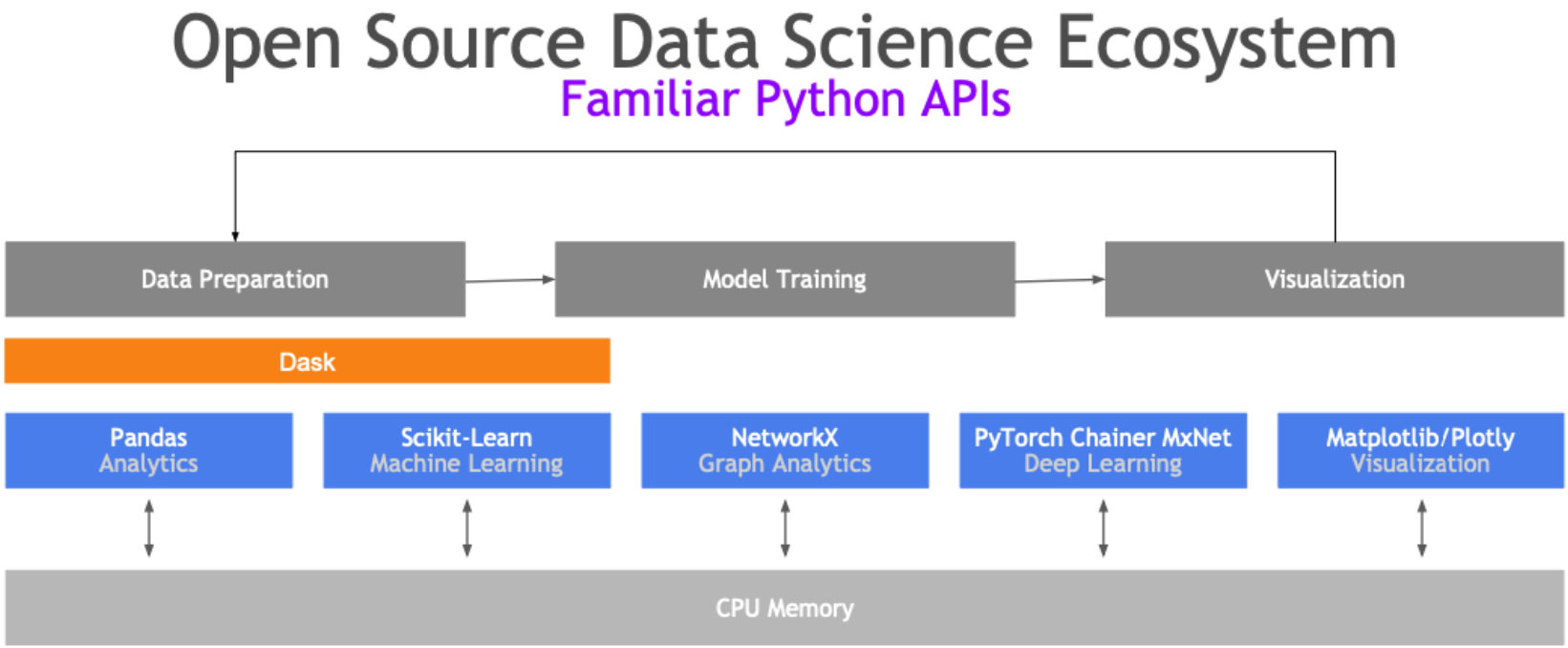
RAPIDS is NVIDIA's new suite of open source software libraries and application programming interfaces (APIs) that aim to give you the ability to accelerate (classical) data science, analytics, and machine learning workflows on NVIDIA GPUs.

https://rapids.ai

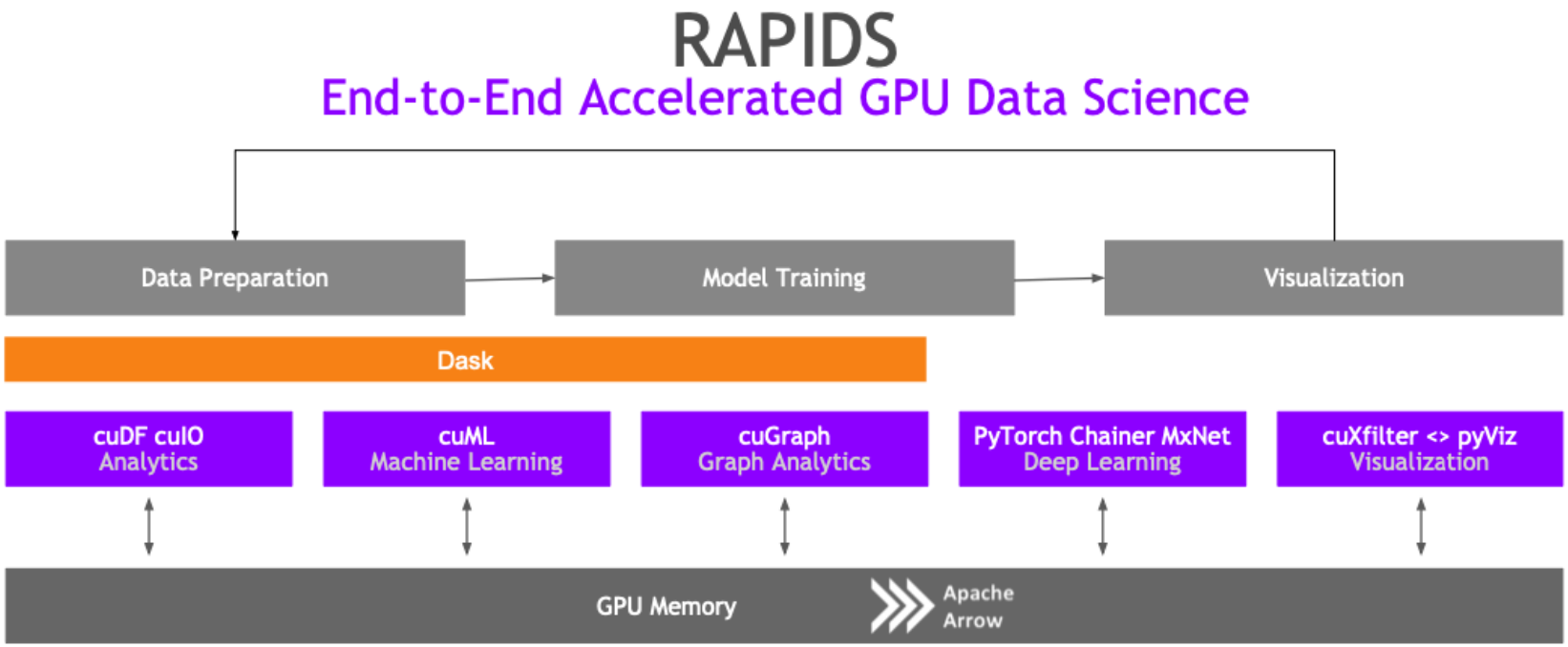
Why RAPIDS?

- **Fast**: Leverages CUDA primitives to provide you with out-of-the-box low-level GPU compute optimizations
- **User-Friendly**: Familiar Python interfaces for each library allows you to (quickly) integrate GPU parallelism and high-bandwidth memory speeds into your existing workflows
- **Scalable**: Integrates with Dask to allow you to scale-out across multiple-node, multi-GPU systems (more easily)
- **Open Source**: Licensed under Apache 2.0

Traditional (CPU-based) Data Science Software Stack



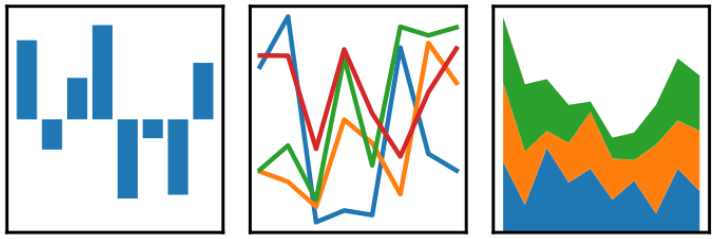
RAPIDS (GPU-based) Data Science Software Stack



RAPIDS: cuDF

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



- cuDF is NVIDIA's GPU-accelerated version of Pandas.
- It provides you with a GPU-native DataFrame library for loading, joining, aggregating, filtering, and otherwise manipulating data.
- It provides you with a familiar Pandas-like Python API that aims to easily accelerate your Pandas-based workflows on NVIDIA GPUs without having to understand the details of CUDA programming.

RAPIDS: cuML

- cuML is NVIDIA's GPU-accelerated version of Scikit-learn.
- It provides you with a GPU-acclerated library of standard statistical and (classical) machine learning algorthms: Linear Regression, K-Means, SVD, etc.

RAPIDS: cuGraph

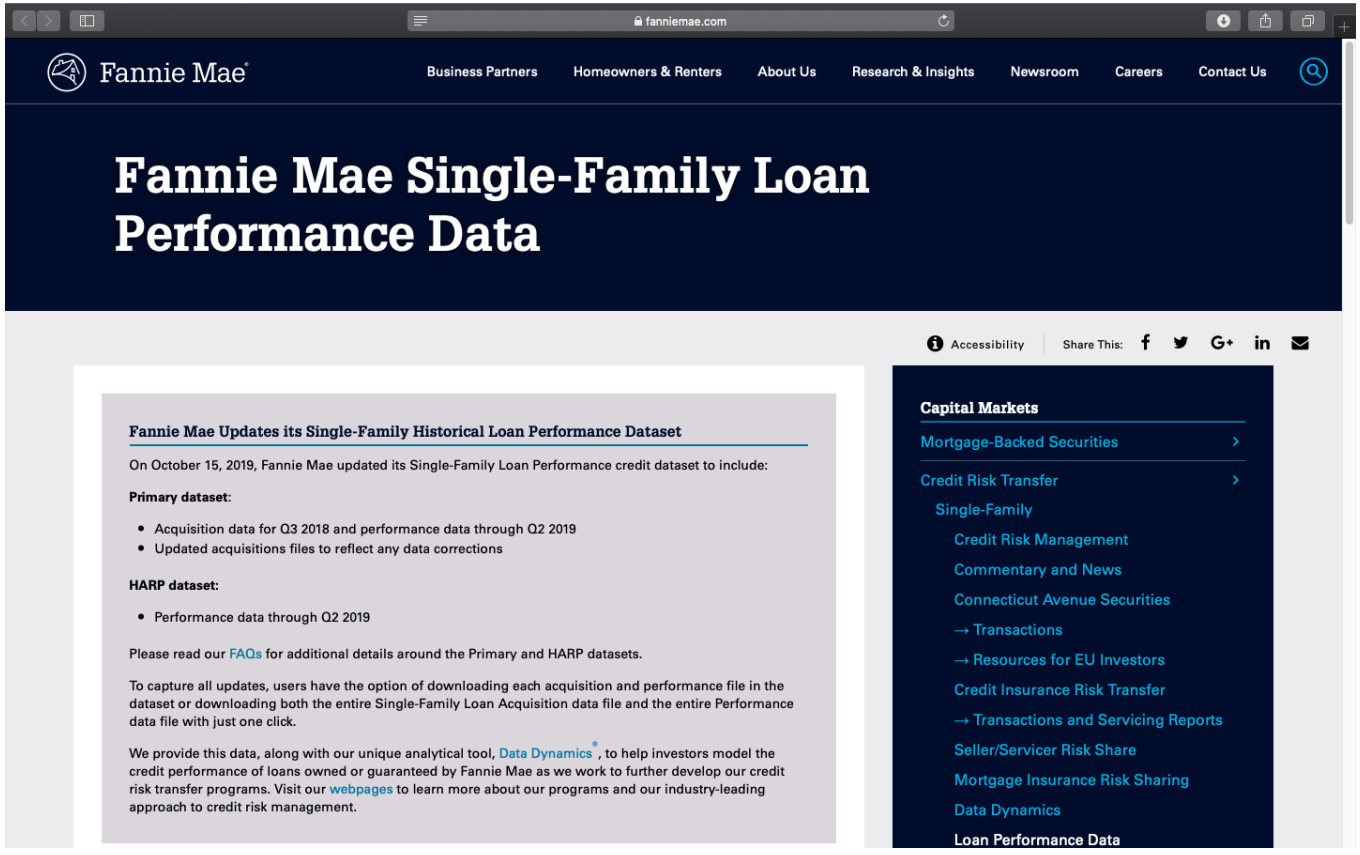
- cuGraphs is NVIDIA's GPU-acclerated version of NetworkX.
- It provides you with a collection of graph analytics algorithms that can be used to process data found in cuDF DataFrames: Page Rank, Breadth First Search, etc.
- Up to 500M edges on a single 32 GB NVIDIA GPU; scales to billions of edges on multi-GPUs.

RAPIDS: cuSpatial

- cuSpatial is NVIDIA's GPU-acclerated library for geospatial and spatiotemporal data processing.
- It provides you with a collection of common spatial and spatiotemporal operations: point-in-polygon tests, distances between trajectories, trajectory clustering, etc.

Jupyter Notebook Demo

Fannie Mae Single-Family Loan Performance Data



https://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html

Why the Fannie Mae SFLP Dataset?

- **Large in Size**: The dataset contains more than 1.9 billion records on 37 million 30-year fixed rate mortgages. It is approximately 200 GB uncompressed.
- **Growing**: The dataset is updated quarterly by Fannie Mae.
- **Well-documented**: Fannie Mae provides an in-depth tutorial on analyzing the dataset with R/SAS code examples. NVIDIA has also featured the dataset in previous RAPIDS demonstrations.
- **Easily accessible**: The dataset can be downloaded for free from Fannie Mae. Several subsets are available from NVIDIA on the RAPIDS website.

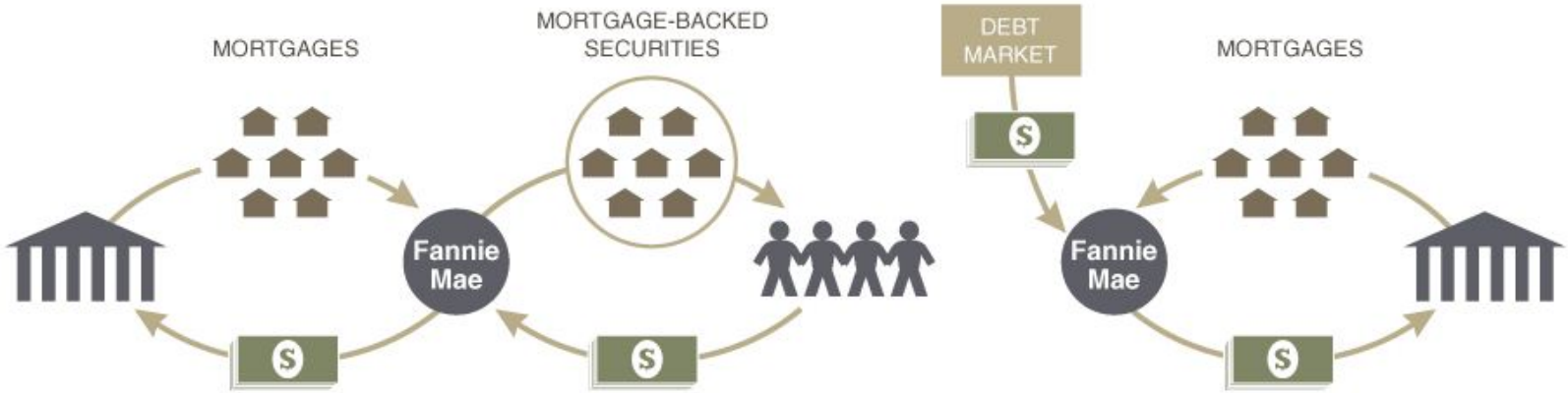
What is Fannie Mae?

- Federal National Mortgage Association (FNMA)
- Government-Sponsored Enterprise
- Founded in 1938
- Created the secondary mortgage market in the U.S.
- Became a publicly traded company in 1968
- Placed into conservatorship by U.S. federal government in 2008 along with Freddie Mac

What does Fannie Mae do?

How Does Fannie Mae Work?

Fannie Mae was originally formed by the federal government in 1938 in order to supply liquidity to the mortgage market. Since 1968, it has been a private corporation. Here is how it works:



Fannie Mae **takes mortgage loans** from banks, in order to repackage them in the form of **mortgage-backed securities**. There are limits on the types and size of loans it can guarantee.

Those **mortgage-backed securities are sold** to investors, and Fannie Mae guarantees that the loans will be repaid.

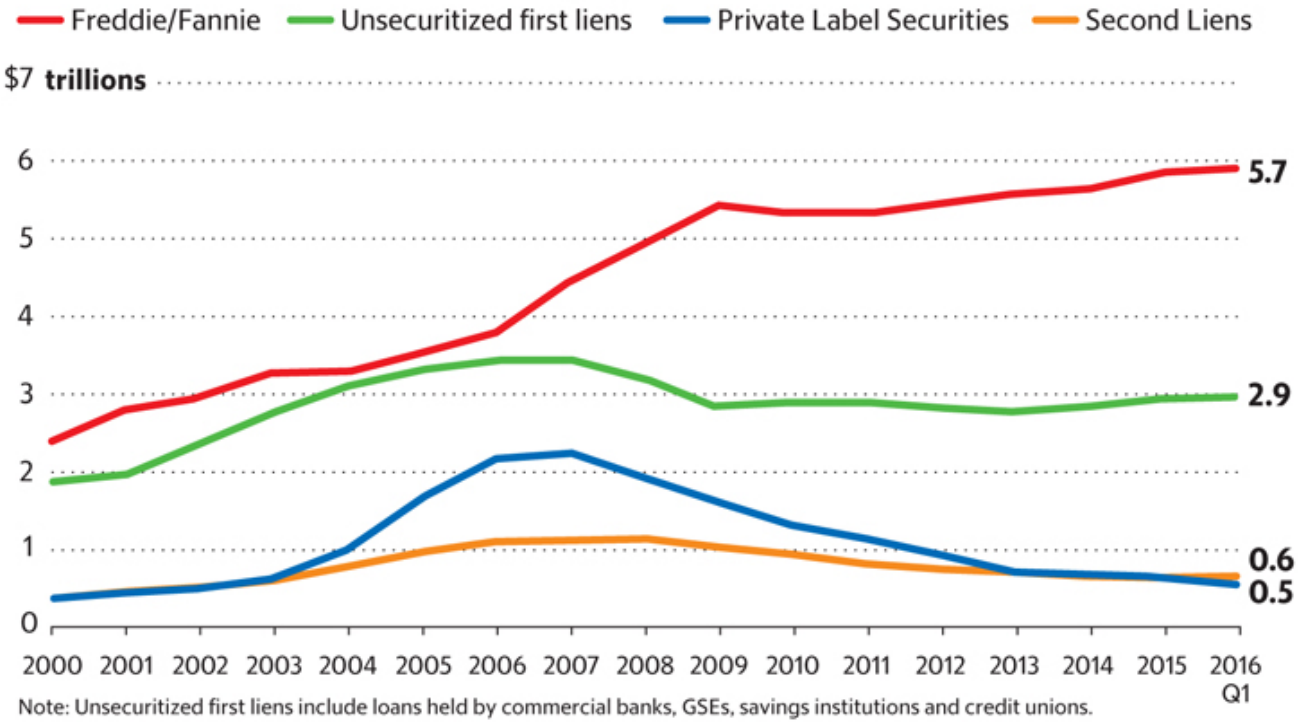
Fannie Mae also **borrows money from the debt markets**, traditionally at a rate much lower than other banks, and uses it to **buy mortgages it holds as its own investments**. By buying these loans, Fannie injects new money into the housing economy.

Size of U.S. Residential Mortgage Market

Private lenders a bit player

Government-sponsored enterprises Freddie Mac and Fannie Mae dominate the residential mortgage-backed securities market. The Obama administration wants private lenders, who right now have a fraction of the market share, to step in.

Size of the US residential mortgage market



Note: Unsecuritized first liens include loans held by commercial banks, GSEs, savings institutions and credit unions.

Sources: Federal Reserve Flow of Funds, Inside Mortgage Finance, Fannie Mae, Freddie Mac, CoreLogic Servicing and Urban Institute.

Conforming Mortgage Loans

- Primary type of mortgage loans that Fannie Mae and other GSEs can purchase.
- Limits maximum loan amount for a given type of property
- In 2019, the standard conforming loan limit for a single family home was $484,350.
- *Jumbo conforming loan limits*: In 2008, conforming loan limits in high cost areas were raised to $729,750 or 125% of the median home value within the metropolitan statistical area, whichever is the lesser.

Okay, let's go to the Jupyter Notebook demo ...

https://github.com/mkandes/presentations/blob/master/2020/01/31/from-zero-to-rapids/notebooks/fannie-mae-sflp.ipynb

How to run the Jupyter Notebook on Comet

1. Obtain an interactive session on one of Comet's GPU nodes.

```
srun --account=use300 --partition=gpu-shared --nodes=1
    --ntasks-per-node=7 --gres=gpu:p100:1 --time
    =01:00:00 --pty --wait=0 /bin/bash
```

2. Navigate to the directory where your notebook is located.

```
cd /oasis/scratch/comet/${USER}/temp_project/
```

3. Shell into the RAPIDS Singularity container.

```
singularity shell --bind /oasis,/run,/scratch --nv /
    share/apps/gpu/singularity/images/rapids/rapids.img
```

4. Start up Jupyter Lab from inside the container.

```
jupyter lab --no-browser --ip="$(hostname)"
```

5. Copy the (remote) URL and token into your browser.
   http://comet-3X-XX.sdsc.edu:8888/?token=
   4f8c8b19748146a2f5b53d0223880d5363ee31b8efee2cb5

Additional Resources and References

- **Homepage**: https://rapids.ai
- **Blog**: https://medium.com/rapids-ai
- **Intro Tutorials**: https://docs.rapids.ai/start
- **GitHub**: https://github.com/rapidsai

Jupyter Notebooks

- **NVIDIA Notebooks**: https://github.com/rapidsai/notebooks
- **Community Contributed Notebooks**: https://github.com/rapidsai/notebooks-contrib
- **Using RAPIDS + PyTorch on the SFLP Dataset**: https://github.com/rapidsai/notebooks-contrib/tree/branch-0.12/blog_notebooks/mortgage_deep_learning

RAPIDS API Documentation

- https://rapidsai.github.io/projects/cudf/en/0.11.0/api.html
- https://rapidsai.github.io/projects/cuml/en/0.11.0/api.html
- https://rapidsai.github.io/projects/cugraph/en/0.11.0/api.html

Recent and Upcoming RAPIDS-Related Training Events

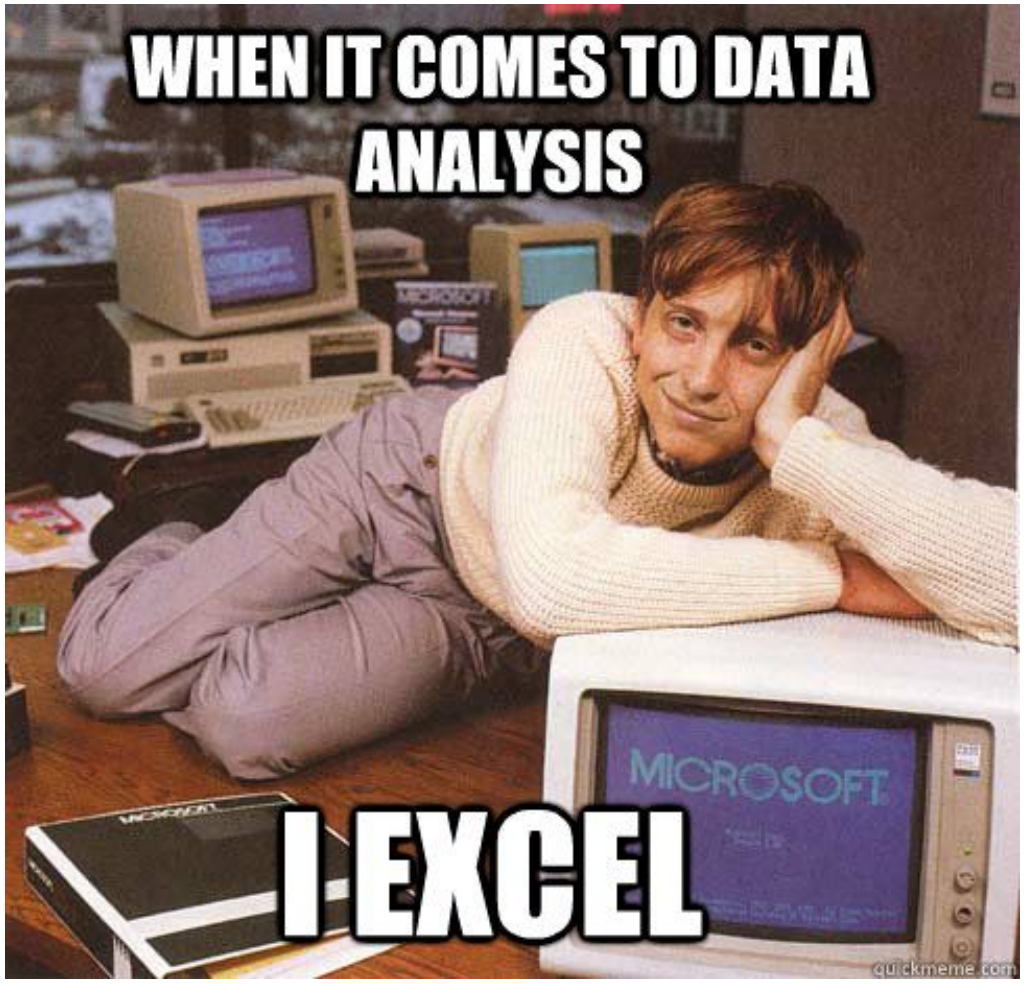
CUDA-Python and RAPIDS for blazing fast scientific computing

Abraham Stern, Solutions Architect, NVIDIA

- **XSEDE ECSS Symposium**
  Tuesday, January 21st, 2020
  https://youtu.be/NdKWEkV9X34

- **XSEDE Webinar**
  Thursday, February 20, 2020
  https://www.sdsc.edu/education_and_training/training/202002_gpu_accelerated_computing_with_cuda_python.html

Questions?


WHEN IT COMES TO DATA ANALYSIS
MICROSOFT
I EXCEL
quickmeme.com