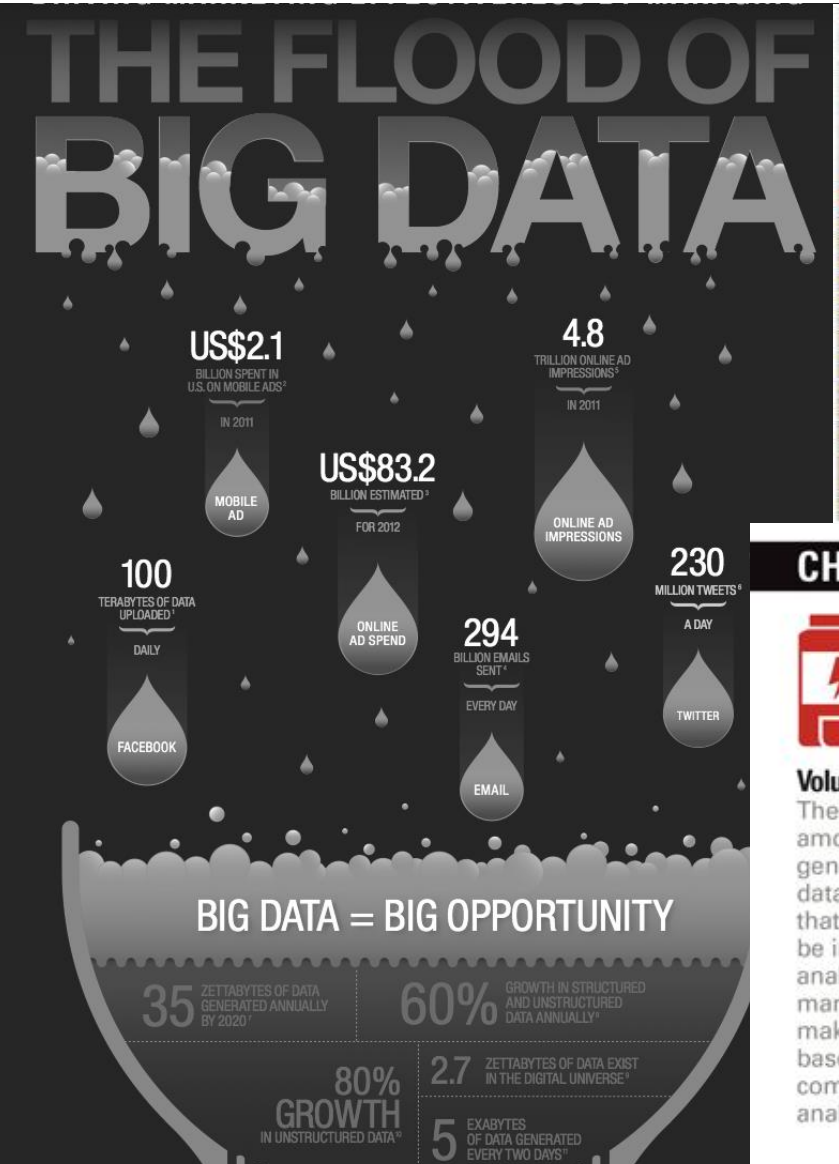
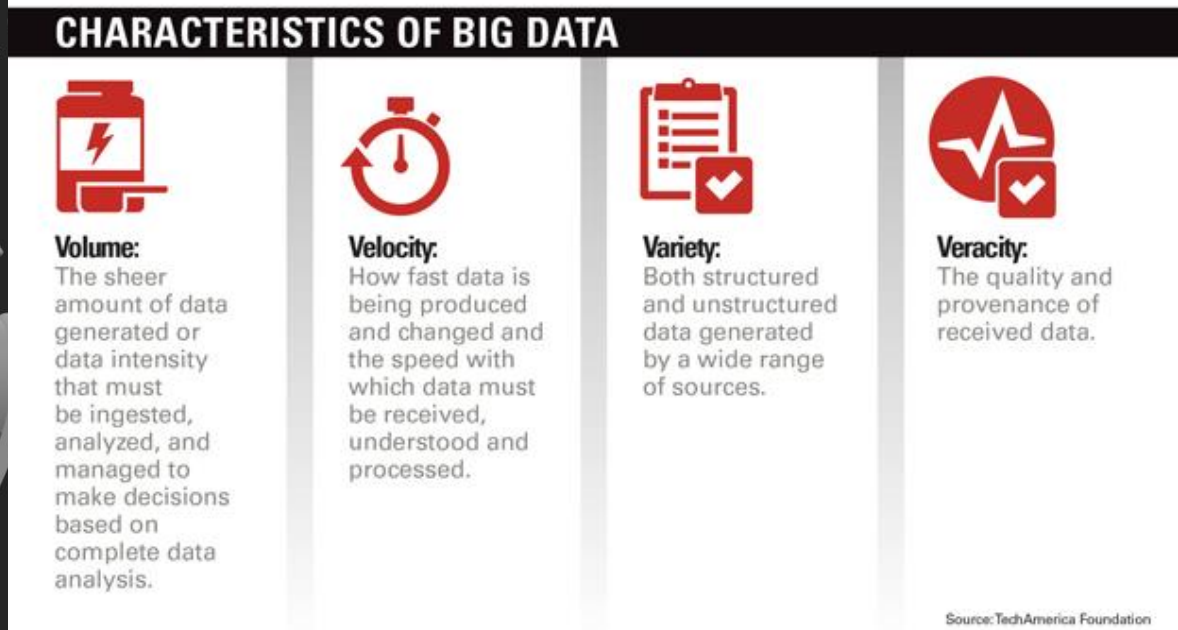
Big Data: Data Science Research and Education

Sanjay Ranka, CISE Department, ranka@cise.ufl.edu, 352 514 4213





Data Science Challenge: Solutions that use extant hardware and leverage open source software

Falling hardware prices

Storage

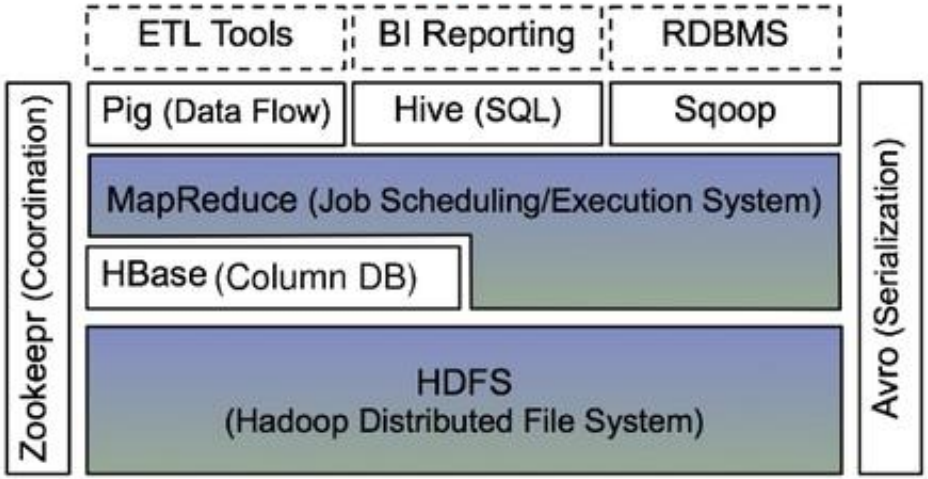
100 Terabyte (Cost $10000)
Tens of multi terabyte disks
Read speed: several GB/s

1 Petabyte Storage System
(Cost $100,000)

Processing

Server ($10,000)
4-8 processors, 512GB
Billion tuples/second

Open Source Ecosystem

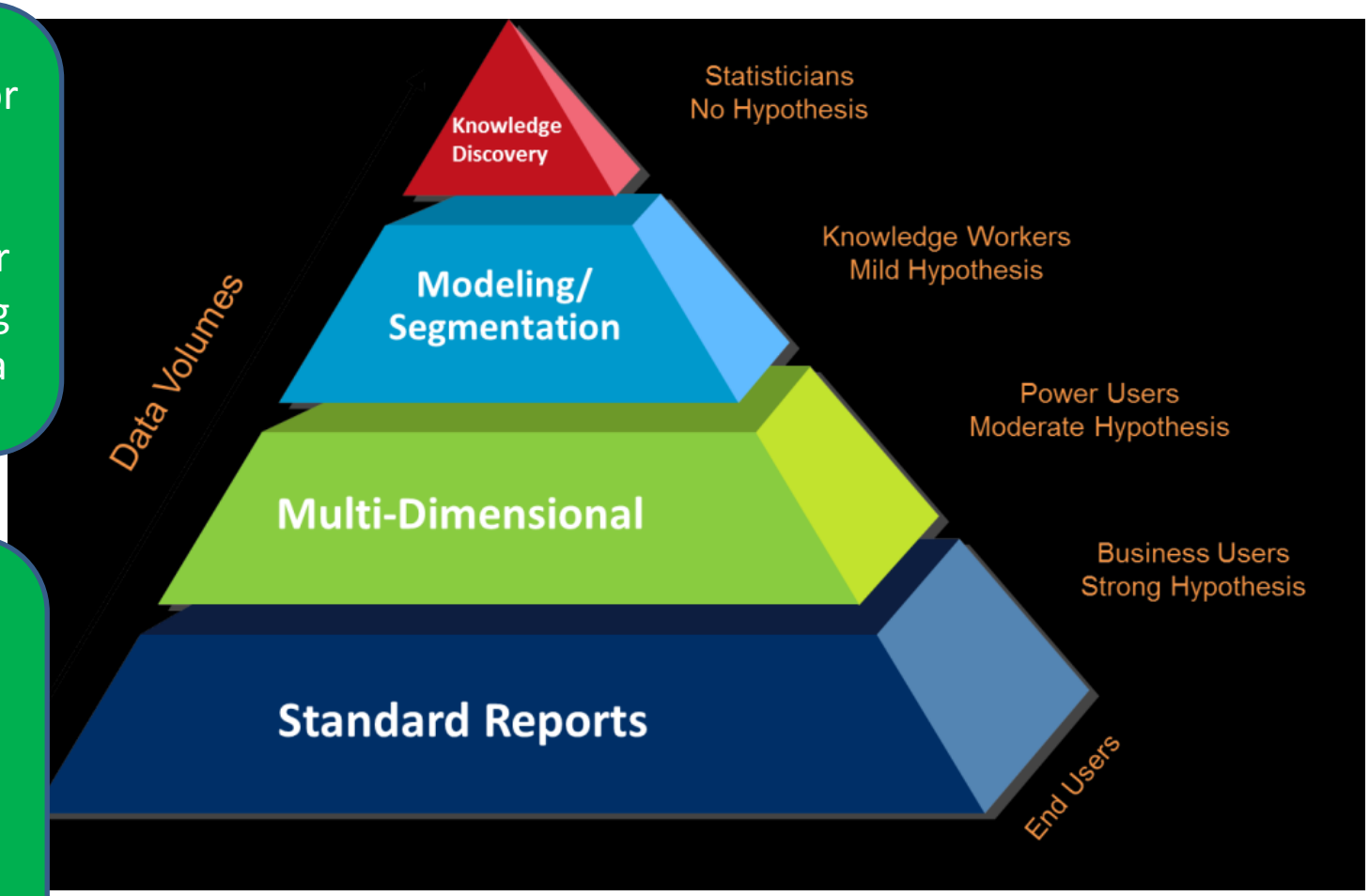


cloudera

Data Science Opportunity: Enable Novel Applications

Novel techniques for processing and storing data
Novel technique for modeling, analyzing and visualizing data

Background in core algorithmic, statistical, architectural concepts a requirement



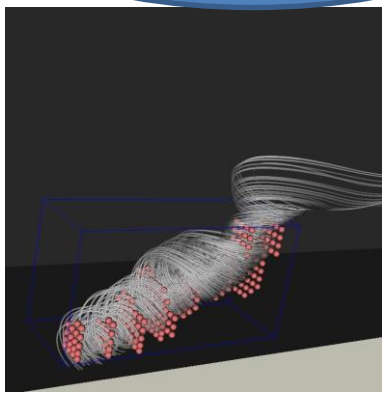
Machine Learning for Spatio-temporal Datasets (with Rangarajan)

Remote Sensing for Climate Modeling

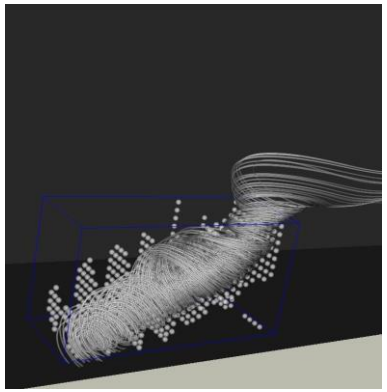
Semi Supervised Learning for Expert In the Loop

Physics-based feature detectors for CFD applications

Terabyte Size Dataset per Simulation

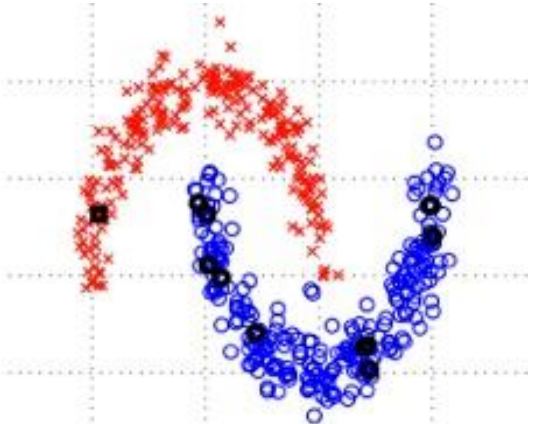


Machine learning

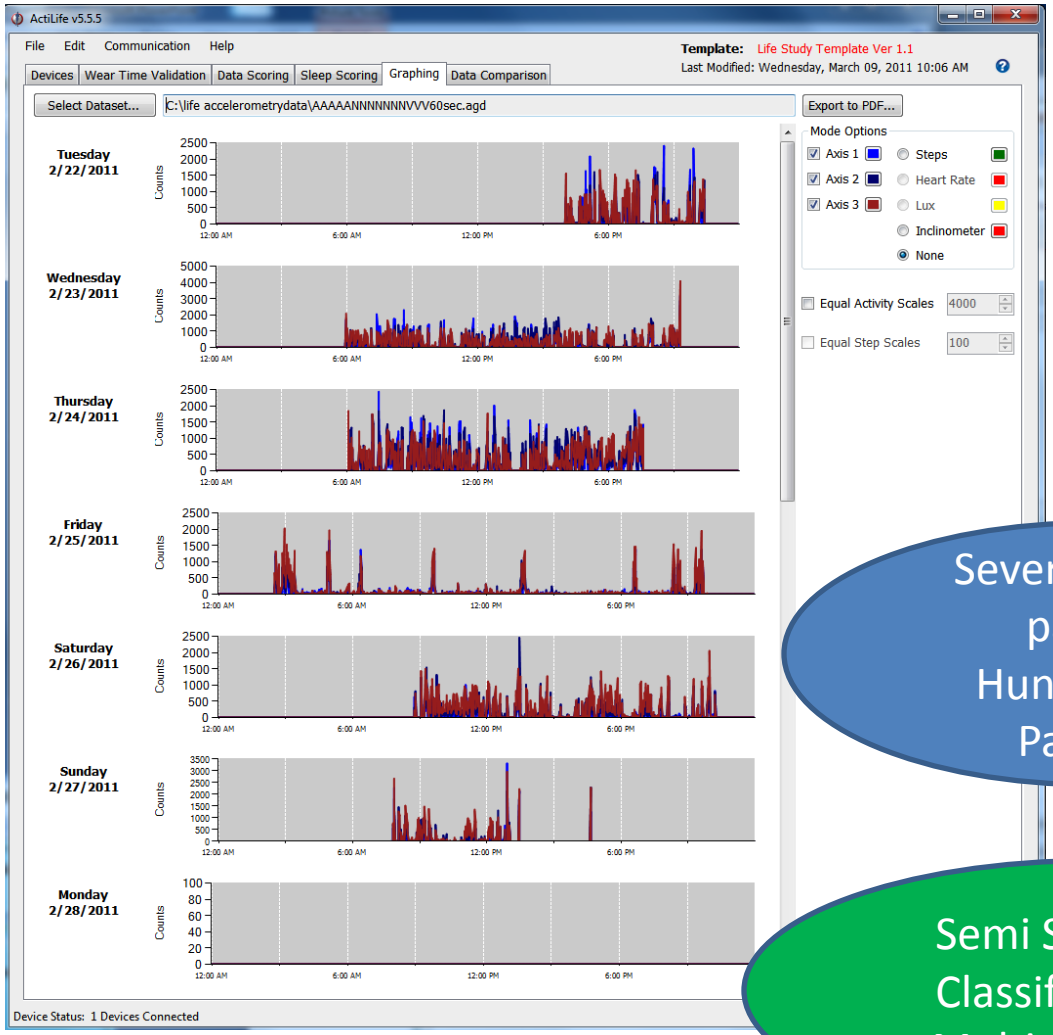


Physics-based

- Expert labels a small fraction of the data
- Construct graph to propagate labels
- Label prediction – weighted combination of neighbors

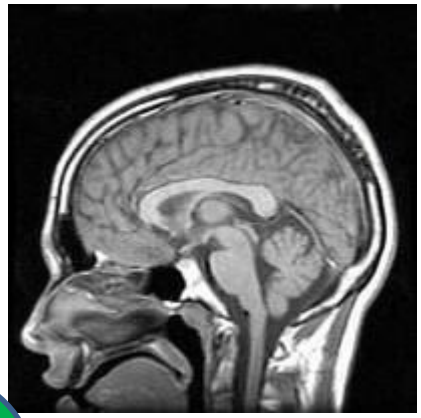


Understand Relationship between Aging, Mobility and Physical Activity (Manini)



Actigraph GT3X

Several GB per patient
Hundreds of Patients



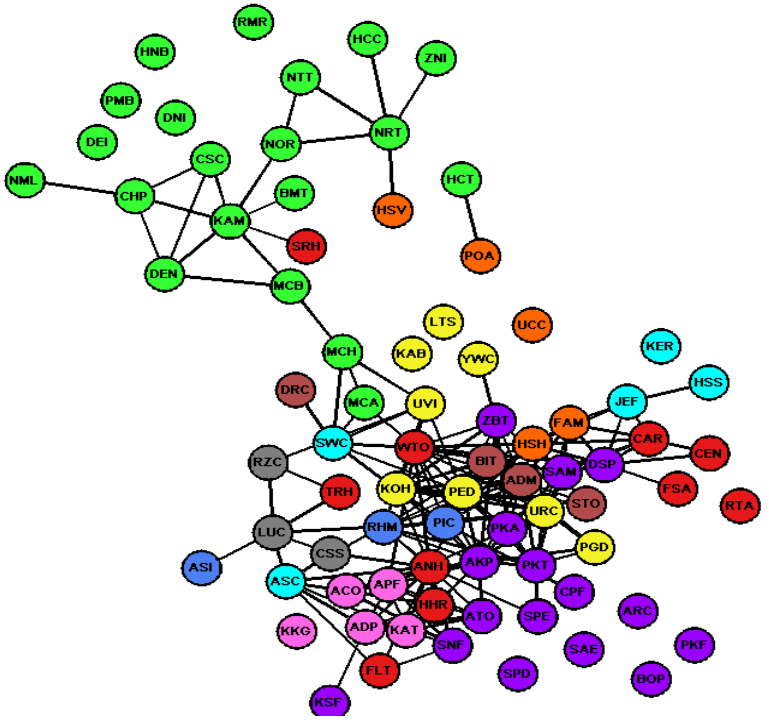
Semi Supervised Classification for Multimodal Data

Modeling Mobility Behavior (with Helmy)

Terabytes per week

User Internet Information:

1. Spatial and location-based information (buildings)
2. Temporal information (Sessions times and duration)
3. Interest-based information (web domains visited)
4. Load and traffic information (flow rate and packet rate)

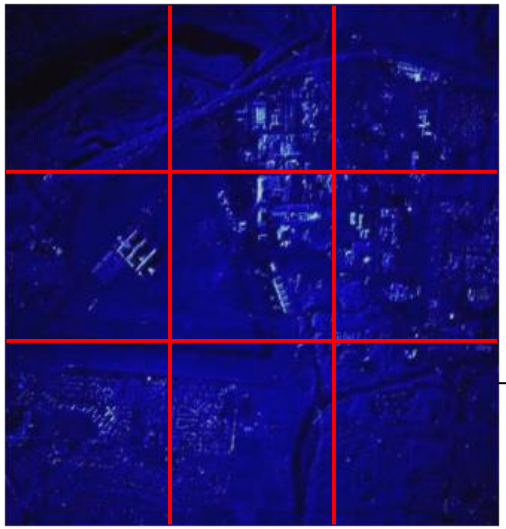


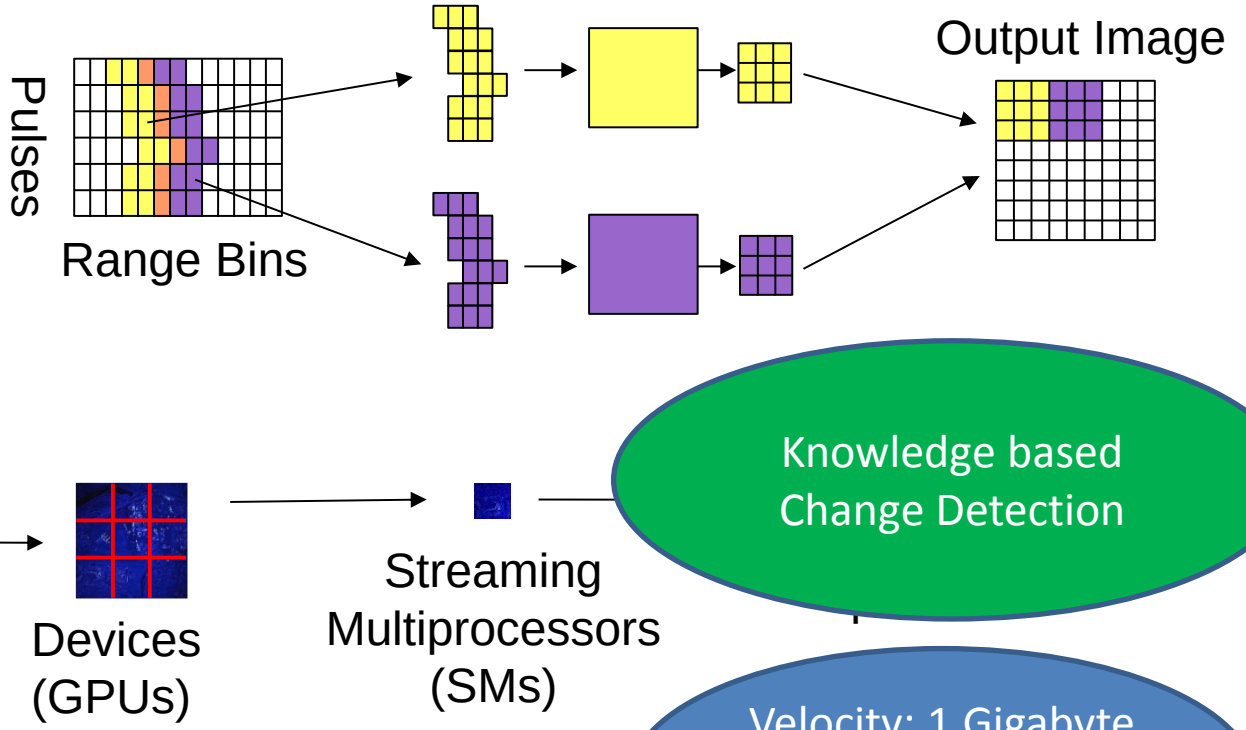
Hierarchical Clustering Change Detection

Table 1. Netflow sample

[illegible]nish Timestamp	Source IP	Source Port	Dest IP	Dest Port	Protocol Num	ToS	Packet Count	Flow Size
0618.00:00:07.184	0618.00:00:07.184	128.125.253.143	53	207.151.245.121	64209	17	0	1	469
0618.00:00:07.184	0618.00:00:07.472	207.151.241.60	52759	74.125.19.17	80	6	0	4	1789
0618.00:00:07.188	0618.00:00:07.188	193.19.82.9	31676	207.151.238.90	43798	17	0	1	103

Real Time Change Detection using Synthetic Aperture Radar (with Sahni)

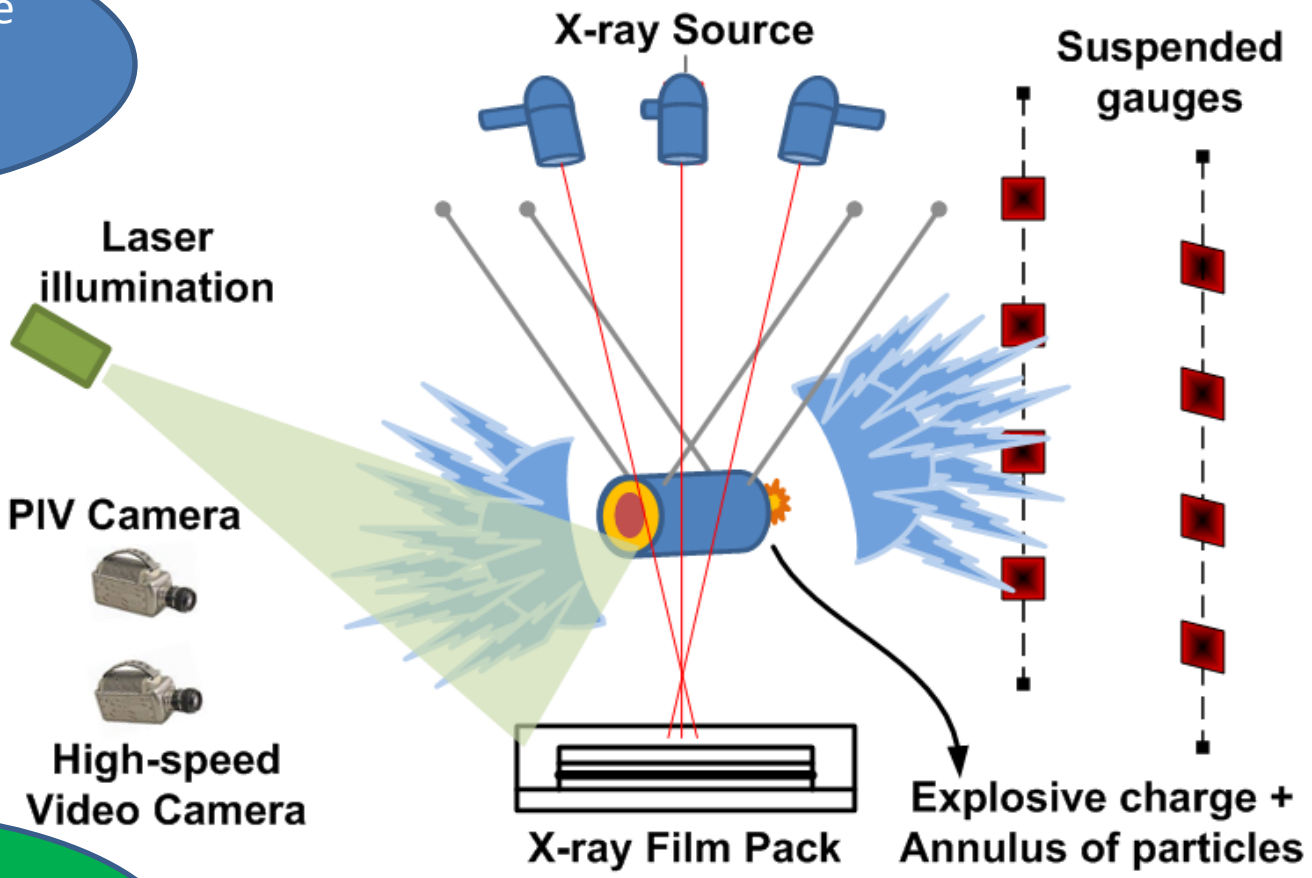




- Tiled data decomposition used to facilitate parallelism between GPUs, and also within the GPU.
- Throughput on cluster of 10 Tesla C2050s: 120 Gflop/s per GPU.

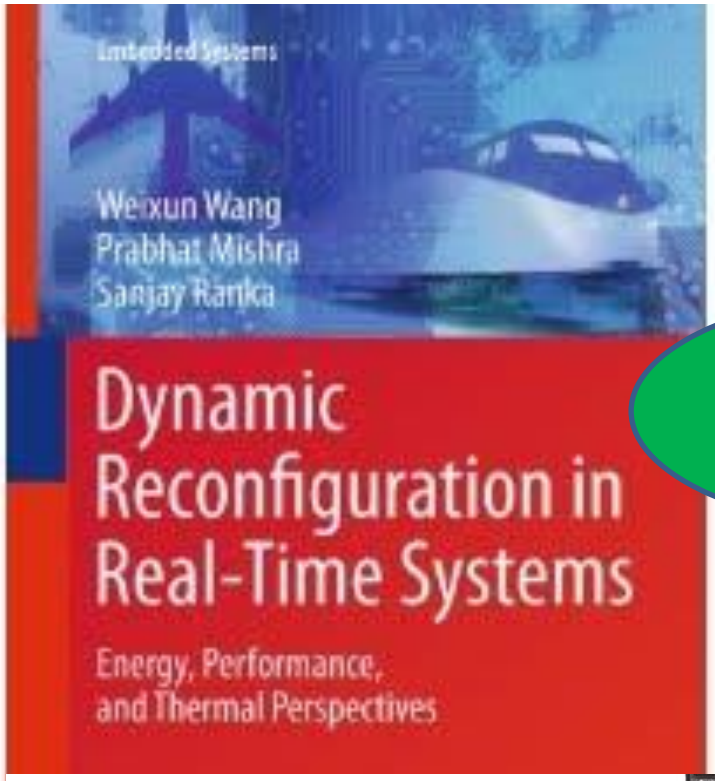
Hardware Software Co-design for Exascale Simulation (with Balachandar et. al.)

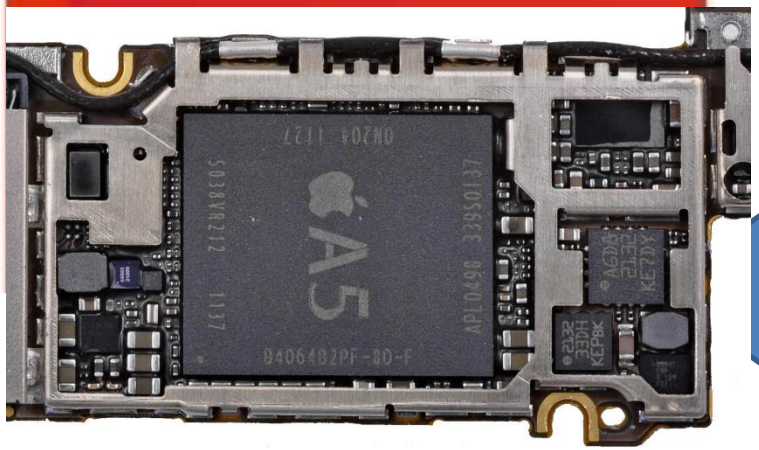
Terabyte to Exabyte Size Dataset per Simulation



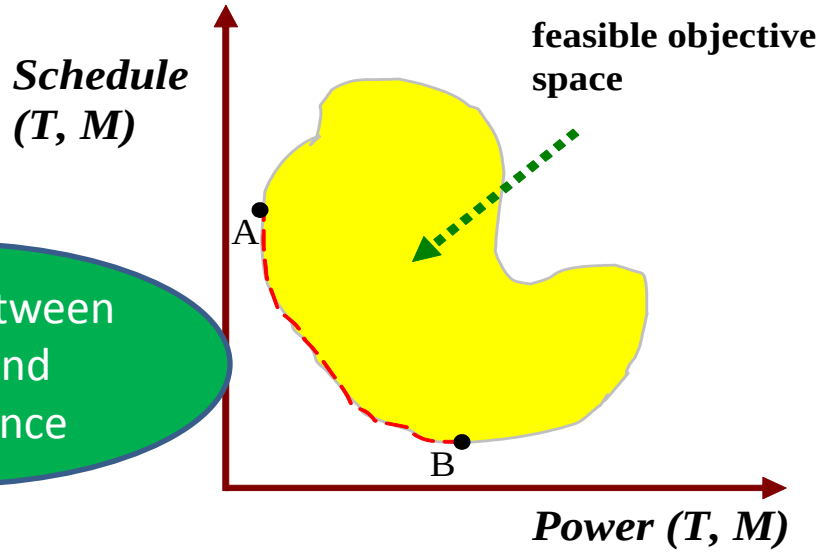
Multi level modeling of performance and energy requirements

Energy Minimization for Mobile Bigdata (with Mishra)

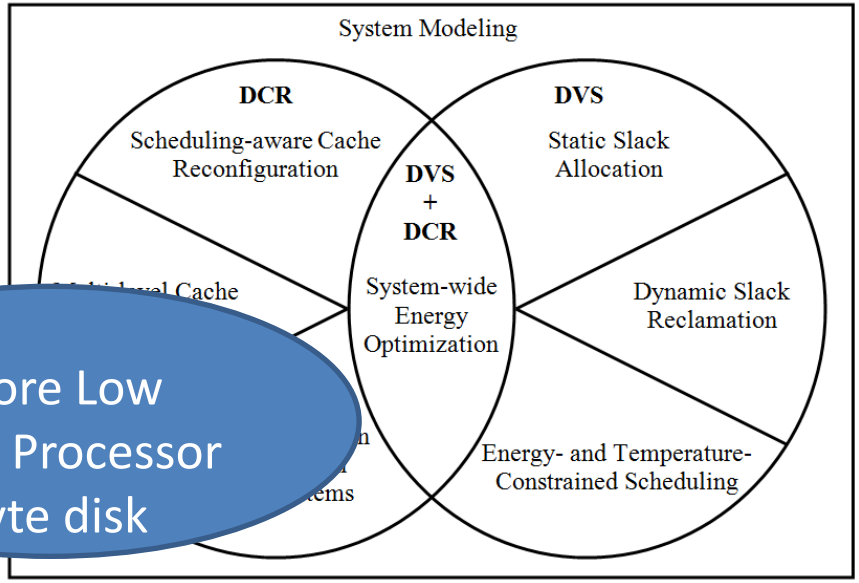




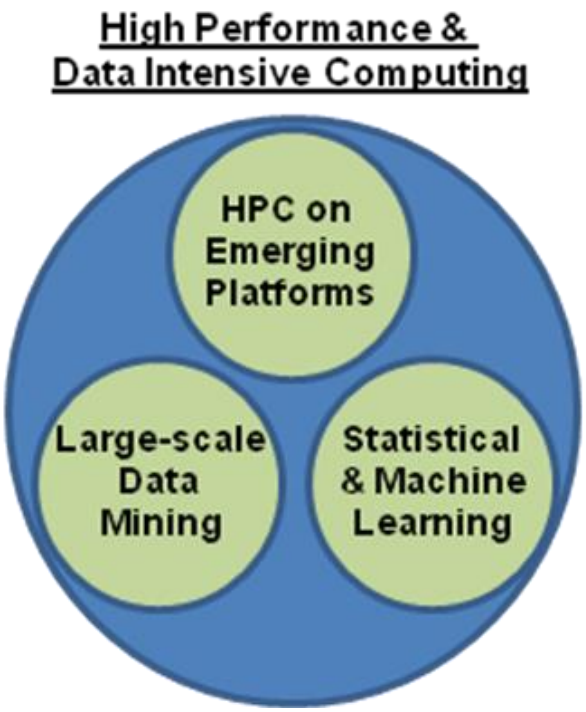
Tradeoff between Energy and Performance



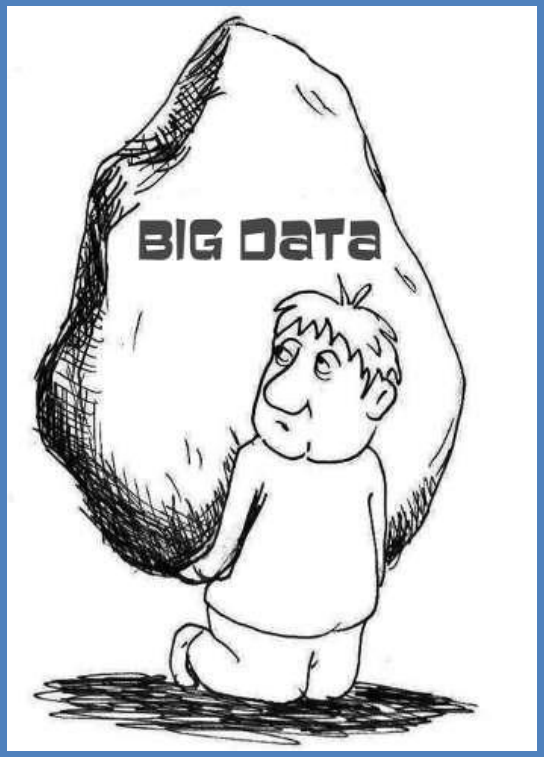
Multicore Low Energy Processor Terabyte disk



Data Science Curriculum (with Rangarajan and Wang)



Curriculum Design underway



Application Driven Project