NERSC Requirements for Enabling CS Research in Exascale Data Analytics

Nagiza F. Samatova
samatovan@ornl.gov
Oak Ridge National Laboratory
North Carolina State University

Oak Ridge National Laboratory

January 5, 2011

NC STATE UNIVERSITY
Department of Computer Science

ACKNOWLEDGEMENTS

- **DOE ASCR for funding CS research in exascale data analytics**
- **Arie Shoshani, Alok Choudhary, Rob Ross, etc**
  - PI's and co-PI's on the projects
- **LCF Facilities at ORNL**
- **Application Scientists:**
  - CS Chang, ORNL
  - Stephane, PPPL
  - Fred Semazzi, NCSU
  - Many others

HARDWARE FOR DATA ANALYTICS IS A FOSTER CHILD

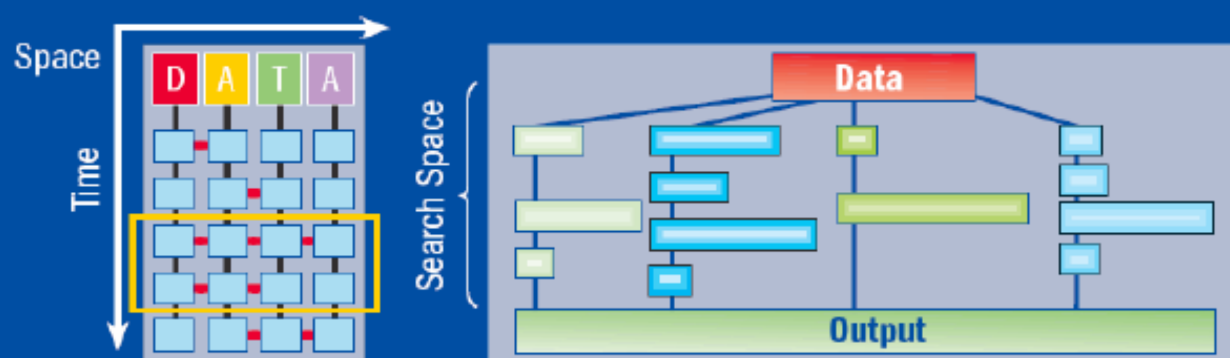
- **HW configuration for DA is an after-thought:**
  - Has been traditionally optimized for running simulations
  - Whatever is left over is what data analyst should live with
  - DA-driven HW must become the first class citizen on the agenda if we are serious about the exascale

- **Infrastructure depends on the DA modality:**
  - In situ?
  - Distributed or streamline fashion?
  - Local or global context analysis?
  - Shared among a group of collaborators?
  - Linked to experimental and/or other data archives?

DISTINCT DATA ACCESS PATTERNS

In contrast to simulations, Data Analytics requires a different mix of memory, disk storage, & communication trade-offs.

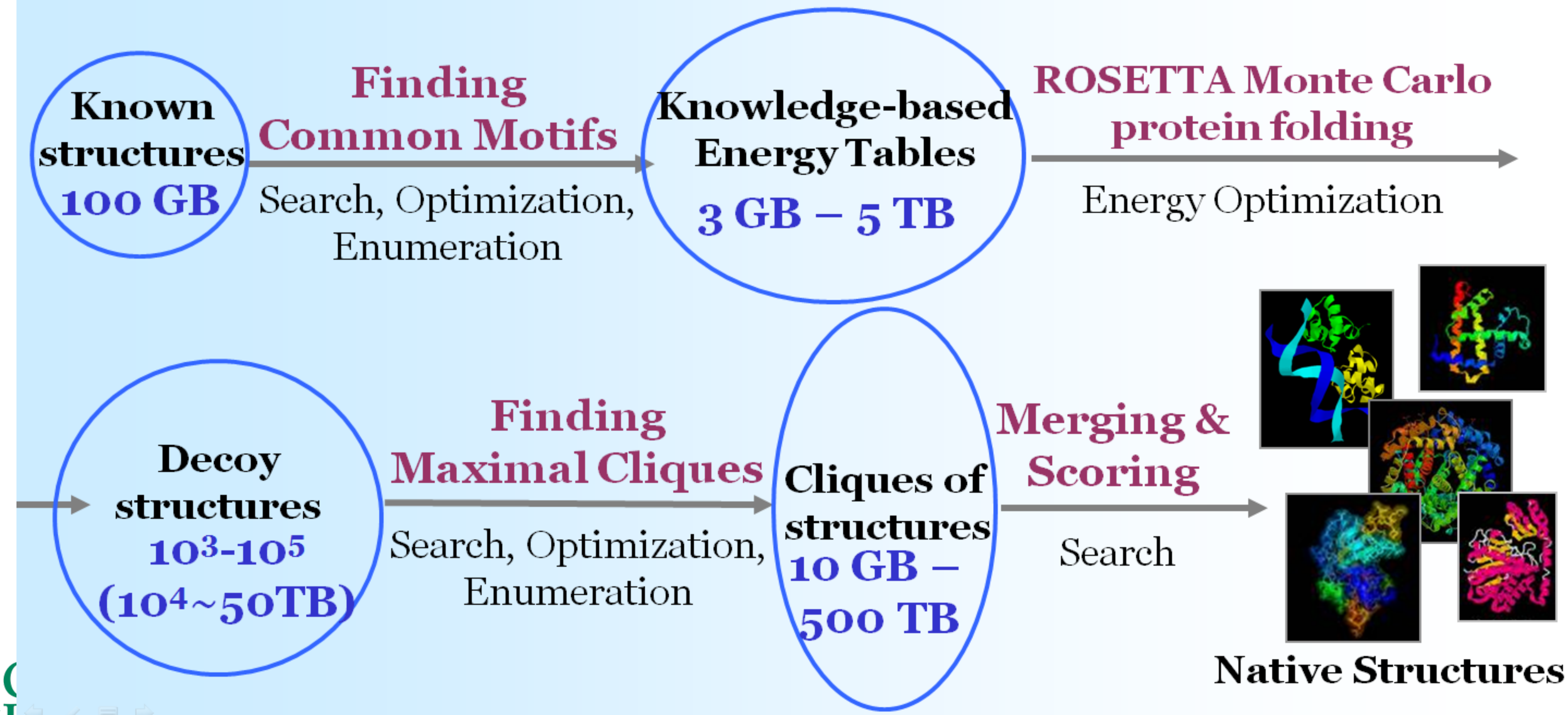
	Simulation	Search	Enumeration
Input	Medium	Huge	Medium
Memory Access	Local (2 Time Steps)	Global (Entire Database)	Exponential Irregular
Output to Disk	Iterative	Irregular	Irregular, Huge
Communication	Intensive	For Scoring	Load Balancing
Arithmetic	Float	Integer (Float)	Integer (Float)



Ex: Scientific Data Flow in Structure Modeling

Each step is a *combinatorial optimization problem* with different data access patterns.

Pipeline: *Ab Initio* Prediction of Protein 3-d Structure



Ex: Scientific Data Flow in MS Proteomics

Twice a month production runs

Archive Data (~2TB) (grows exponentially):
-- 100MB files
-- 100-1000s scans in each file

↓

SEQUEST

14-24 hours per file
One scan access at a time

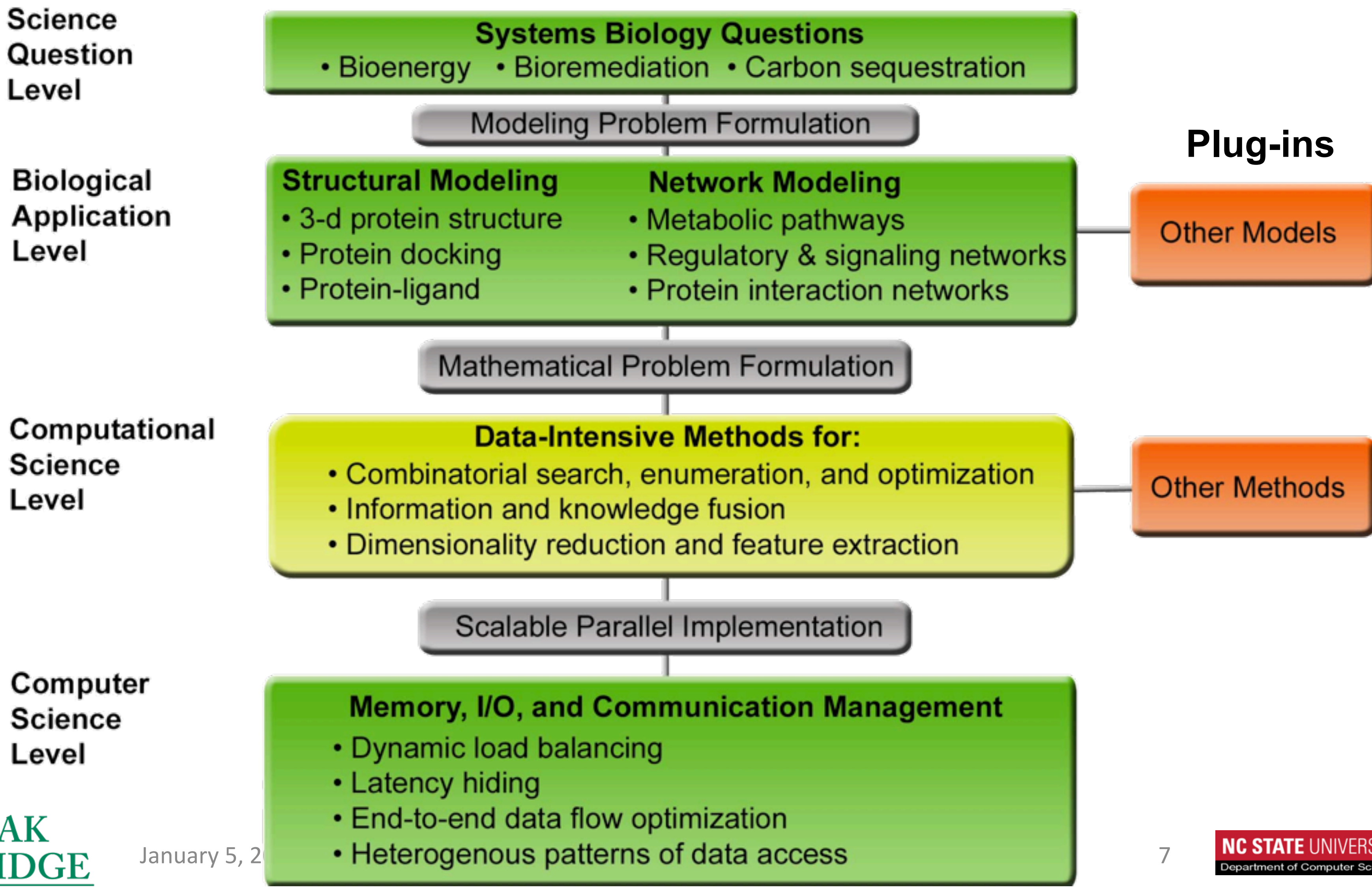
↓

Results (~2TB)

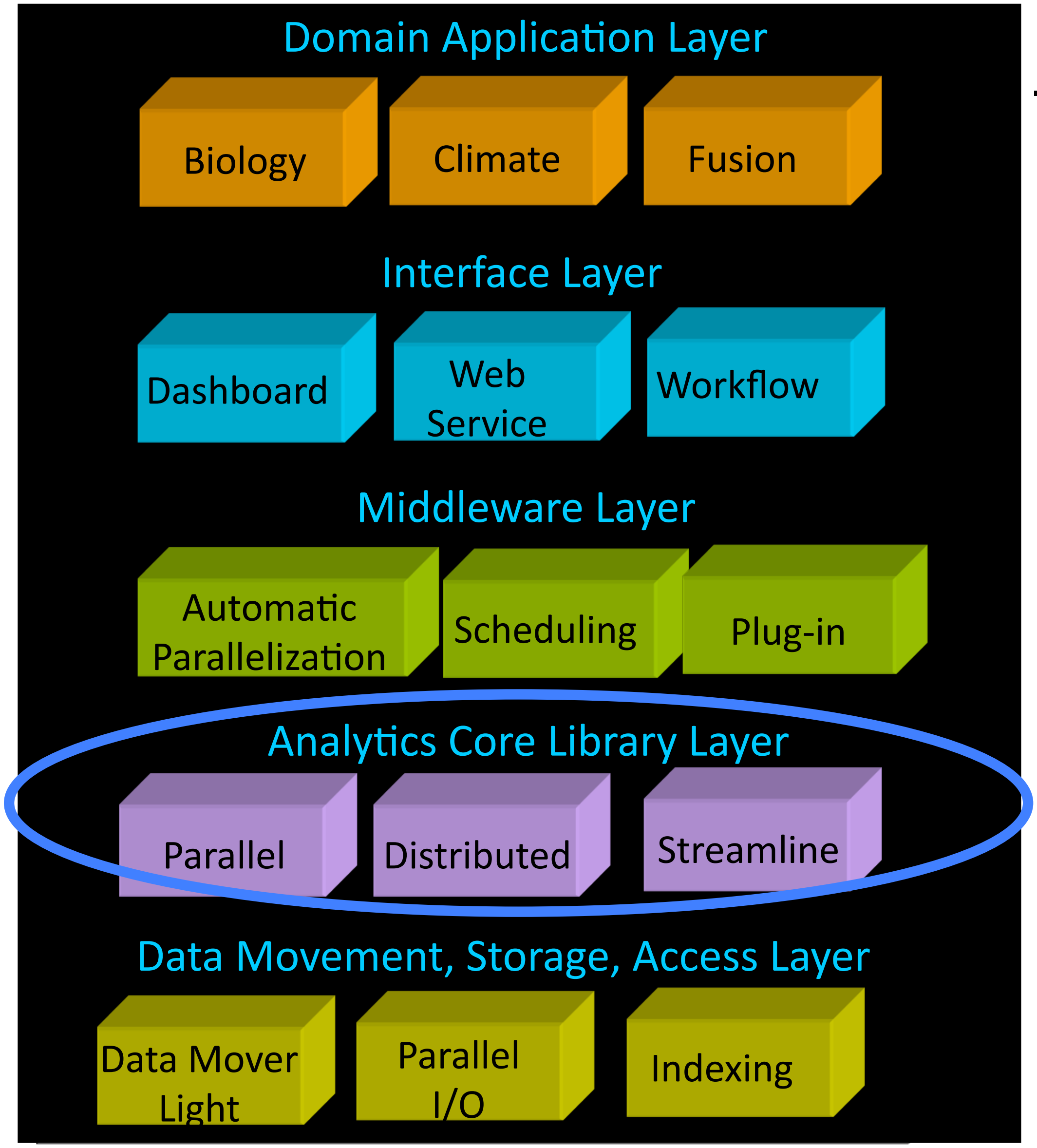
Oak Ridge National Laboratory

NC STATE UNIVERSITY
Department of Computer Science

DOMAIN-SPECIFIC REALIZATION OF THE SW STACK



End-to-End Data Analytics Software Stack is Complex: Generic (All Applications) Perspective



Focus of my talk

Exemplars

CDAT

Kepler

pR

RScaLAPACK

FastBit

pnetCDF

Oak Ridge National Laboratory

NC State University Department of Computer Science

The Lesson Learned from Linear Algebra

"Dwarfs:"
Sparse, Dense, etc.

- ScaLAPACK
- PETSc
- etc

Pre-conditioner

Solver

"…conditions a given problem into a form that is more suitable for numerical solution."

highly optimized computational kernel

SOFTWARE FOR DATA ANALYTICS IS MORE AD HOC

- **Should we adopt this approach from Linear Algebra to Data Analytics at extreme scale? If so, then**
  - What are the “Dwarfs” for data analytics?
  - What about the “Preconditioners?”
  - What are the “Computational kernels?”

- **Do we/should we have a ScaLAPACK-like library for Exascale Data Analytics?**

- **What NERSC should/could offer for enabling this activity?**

"Computational Kernels" Concept is Promising

The frequency of kernel operations in illustrative data mining algorithms and applications.

Application	Top 3 Kernels (%)			Sum %
	Kernel 1 (%)	Kernel 2 (%)	Kernel 3 (%)	
K-means	Distance (68)	Center (21)	minDist (10)	99
Fuzzy K-means	Center (58)	Distance (39)	fuzzySum (1)	98
BIRCH	Distance (54)	Variance (22)	redist.(10)	86
HOP	Density (39)	Search (30)	Gather (23)	92
Naïve Bayesian	probCal (49)	Variance (38)	dataRead (10)	97
ScalParC	Classify (37)	giniCalc (36)	Compare (24)	97
Apriori	Subset (58)	dataRead (14)	Increment (8)	80
Eclat	Intersect (39)	addClass (23)	invertC (10)	72
SVMlight	quotMatrix(57)	quadGrad (38)	quotUpdate(2)	97

Alok Choudhary, NWU, *NU-Minebench*

WHAT ABOUT "PRECONDITIONERS" FOR DATA ANALYTICS?

- **How to define a "preconditioner" for data analytics?**

Solve a Problem P_{hard}

Directly

Indirectly (via "Preconditioner"):

Reduce a Hard Problem P_{hard} to a "Better" Problem P_{better}

$$\boxed{P_{hard} \rightarrow \text{Preconditioner} \rightarrow P_{better}}$$

"Better" in terms of:

- Increased throughput
- Faster time-to-solution
- More accurate solution
- Higher data compression rate
- Approximate but real-time solution

If we are lucky…

- **and Jack Dongara did most of the work for us:**
  - Some data analysis routines call linear algebra functions
  - In R, they are built on top of LAPACK library
- **RScaLAPACK is an R wrapper library to ScaLAPACK**

```
A = matrix(rnorm(256),16,16)
b = as.vector(rnorm(16))
```

Using RScaLAPACK:	Using R:
library (RScaLAPACK)	
sla.solve (A,b)	solve (A,b)
sla.svd (A)	La.svd (A)
sla.prcomp (A)	prcomp (A)

What if we are not that lucky?

In Situ Preconditioners for Scientific Data Compression

- **Myth: "Scientific data is almost uncompressible."**

GTS Fusion Simulation Data (Stephane, PPPL)

C&R Data

- ~2TB per C&R
- Every 1 hour
- Two copies
- Keep the last copy

Analysis Data

- ~2TB per run (now)
- Every 10^{th} time step
- Cannot afford storing all b/s of
- Analysis routines and I/O reads
- Matlab analysis routines

V&V Data

- Small
- Every 2^{nd} time step

Expected: 10-fold increase by 2012-2014

Computing and Storage Resources

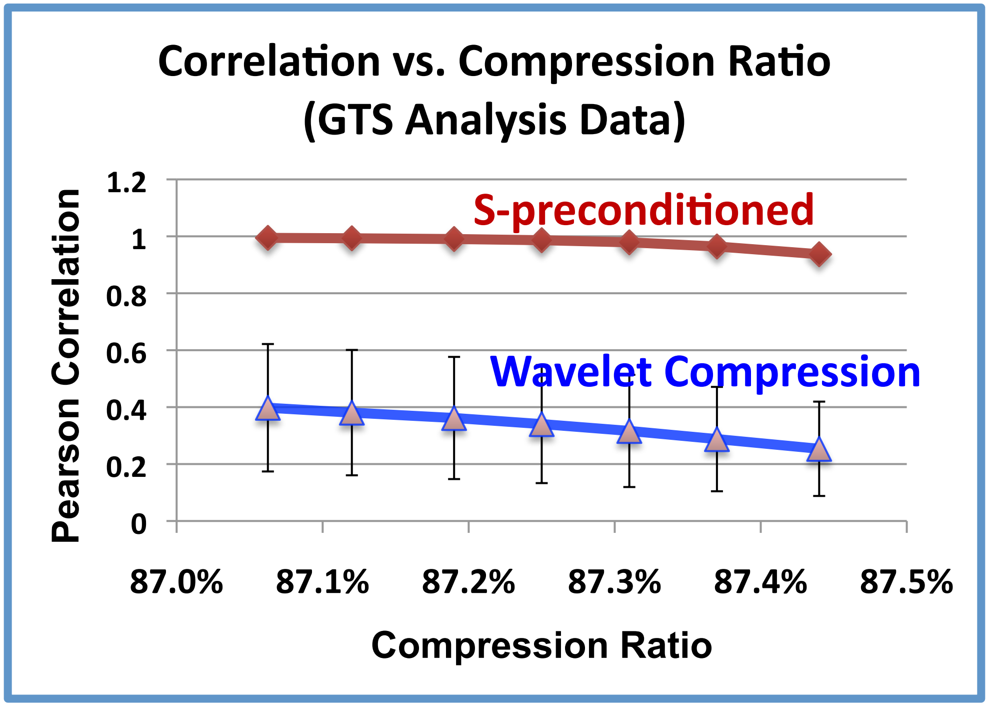
	GYRO		GTS		XGC1*	
Facilities	NERSC/OLCF		NERSC/OLCF		NERSC/OLCF	
Architectures	XT5,Power,Cluster		XT5		XT5	
Years	Present	In 5 yrs	Present	In 5 yrs	Present	In 5 yrs
Hrs used/year	30M	50M	24M	50M	65M	500M
NERSC'09 used	1.2Mhrs		~2Mhrs		~8M hrs	
#Cores per run	512	512	8-98K	32-130K	10-223K	1M
Wall clock/run	12	24	72 Hrs	72 Hrs	20-100hrs	20-100hrs
Memory/run	512GB	1.024TB	16-100T	32-160TB	40 TB	100 TB
Min Memory/core	1GB	2GB	1GB	1GB	0.3GB	0.1GB
Read/Write data			2.5TB	8TB	5TB	25TB
Checkpoint size	4GB	8GB	1-8GB	1-10 GB	1TB	5TB
Data in/out nersc			5GB/run	10GB/run	10GB/day	50GB/day
On-line storage			4TB/10K	8TB/10K	4TB/3K	5TB/3K
Off-line storage			25GB	100GB	1TB/30	10TB/100

FROM C.S. CHANG'S TALK AT NERSC

***Unstructured mesh**

S-PRECONDITIONER FOR ANALYSIS DATA COMPRESSION

- **Analysis data is stored every N-*th* time step:**
  - Lossy data reduction
  - Data is almost random—hard/impossible to compress; <10% lossless
  - N is defined ad hoc (N=10 for GTS, N=100 for Supernova)



Stephane, PPPL: "**With this data quality and data reduction rate, I can test many more hypothesis using my analysis tools**."

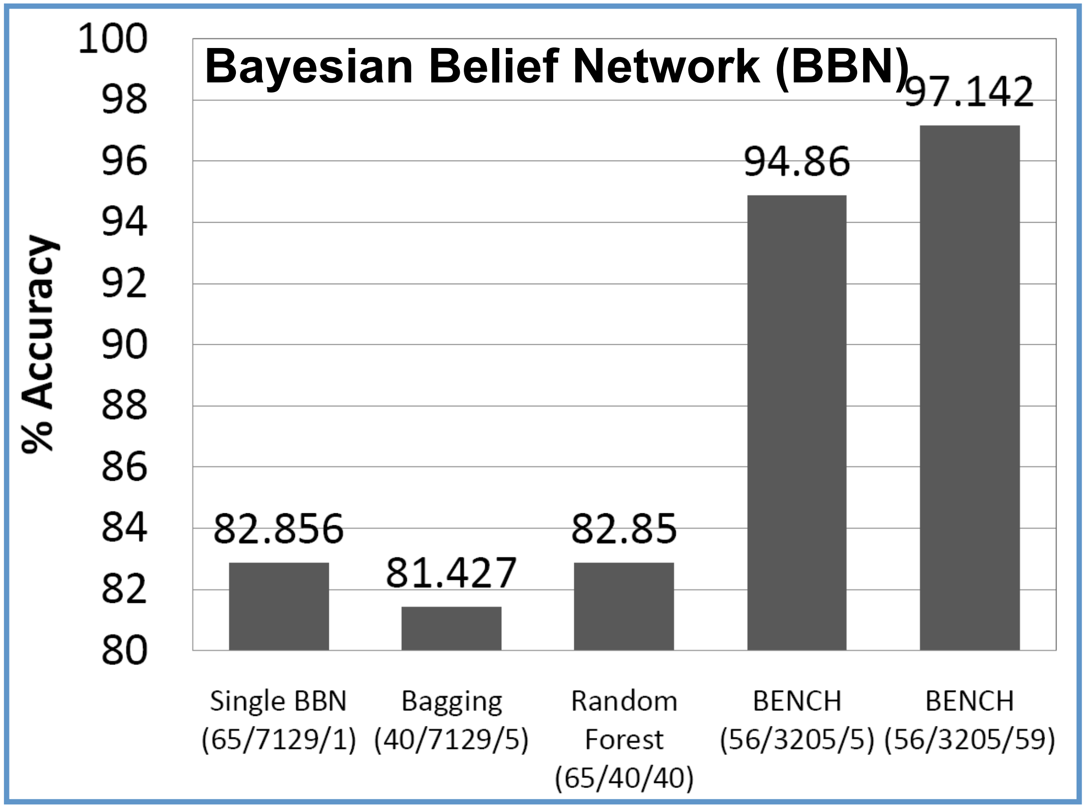
While Compression Ratio is growing up from 87.06% to 87.44%, the Pearson Correlation dropped from 0.994 to 0.937.

BFA-PRECONDITIONER FOR C&R DATA COMPRESSION

- **C&R Data Compression:**
  - Must be lossless
  - Must be fast

- **Impact of BFA-preconditioner:**
  - **8x throughput increase for bzip**
  - **4x throughput increase for gzip**
  - **1.41 compression ratio (CR) for zpaq with BFA-precond**
  - **1.33 vs. 1.17 CR for bzip2 with vs. w/o BFA-precond.**
  - **1.32 vs. 1.19 CR for zlib with vs. w/o BFA-precond.**

BC-PRECONDITIONER FOR UNDERDETRMINED CLASSIFICATION PROBLEM (BENCH)



- **Accuracy increase by 13%-16%**
- **Across different classifiers**
- **On data with <100 samples >d=4,000-7,000 dimensions—underdetermined problems**
- **When applied to seasonal hurricane prediction (d>35K), correlation with observed improved from 0.64 to 0.92-0.96**

Classifier	Single classifier	BENCH ensemble
BBN	82.856	97.142
Decision Tree	82.856	95.714
SVM	91.426	97.142

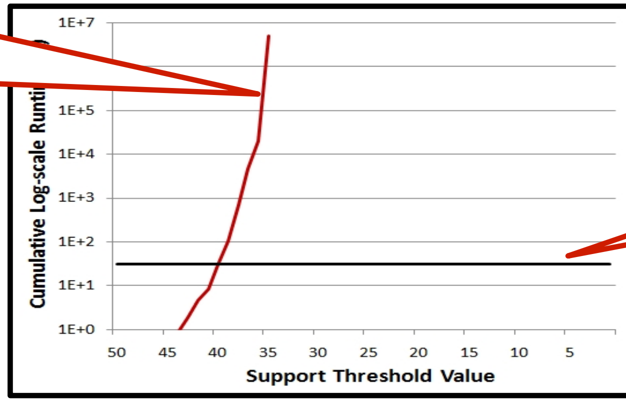
Oak Ridge National Laboratory

PhD students: Ann Pansombut, William Hendrix, Zhengzhang Chen, Doel Gonzalez

SE-Preconditioner for Contrasting Frequent Subgraph Mining (NIBBS-Search)

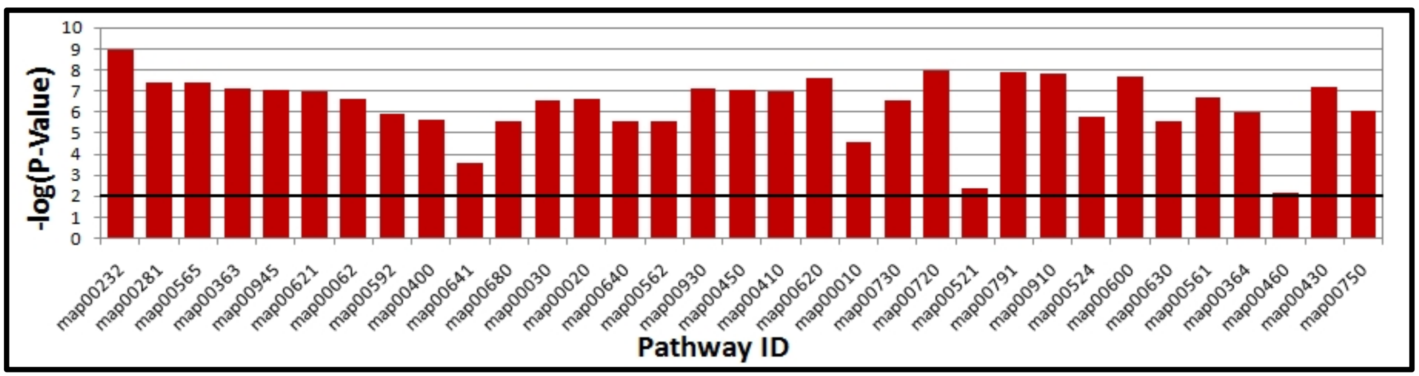
Exact Algorithm versus NIBBS-Search
(98 Genome-Scale Metabolic Networks, 49 Positive, 49 Negative)

Runtime of exact algorithm grows exponentially (unable to complete run)



The NIBBS-Search algorithm completes in a matter of seconds

Empirical tests show that the NIBBS-Search subgraphs are significantly close approximations of maximally-biased subgraphs



PhD students: Matt Schmidt, Andrea Rocha

Oak Ridge National Laboratory

NC State University
Department of Computer Science

Dark Fermentative Bio-hydrogen Production Pathways are Identified with NIBBS-Search

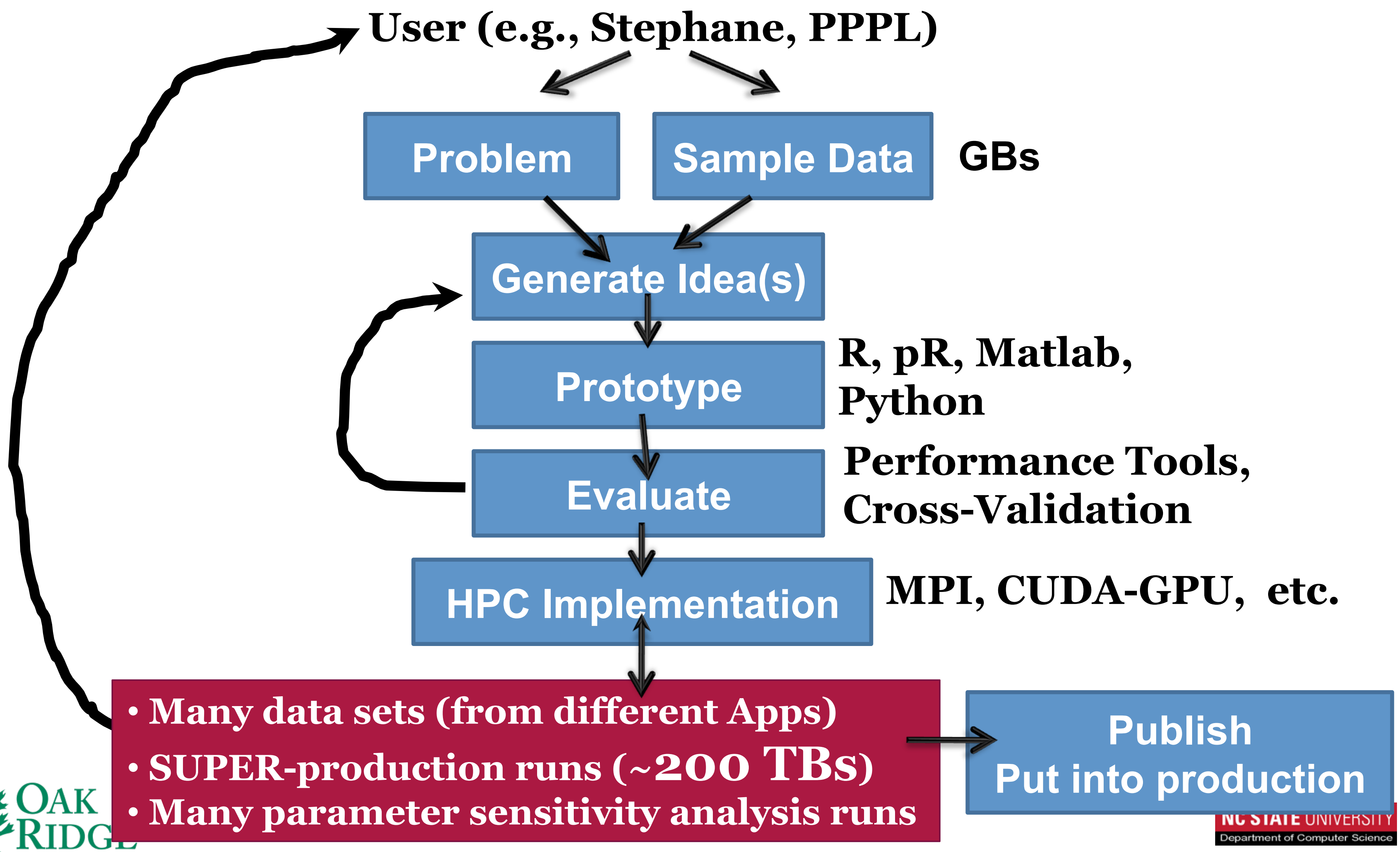
EC Number	Enzyme Name	T-Test	NIBS	Mutual Information
Acetate Pathway				
2.7.2.1	acetate kinase;		TRUE	
2.3.1.8	phosphotransacetylase	TRUE	TRUE	
4.2.1.55	crotonase	TRUE	TRUE	
2.3.1.54	pyruvate formate lyase		TRUE	
Butyrate Pathway				
1.3.99.2	butyryl-CoA dehydrogenase;		TRUE	
2.7.2.7	butyrate kinase	TRUE	TRUE	
1.1.1.157	3-hydroxybutyryl-CoA dehydrogenase;		TRUE	
2.3.1.19	phosphate butyryltransferase;	TRUE	TRUE	
2.3.1.9	acetyl-CoA C-acetyltransferase;		TRUE	
2.3.1.54	pyruvate formate lyase		TRUE	
4.2.1.55	crotonase	TRUE	TRUE	
Formate Pathway				
1.12.1.2	formate dehydrogenase	TRUE	TRUE	
1.2.7.1	pyruvate ferredoxin oxidoreductase		TRUE	
1.12.7.2	ferrodoxin hydrogenase			

PhD students: Andrea Rocha and Kanchana; Undergrad: Katie Shpanskaya

Oak Ridge National Laboratory

E UNIVERSITY
Computer Science

CS Data Analysis Research "Workflow"— Iterative Process w/ Significant Resource Needs



Oak Ridge National Laboratory

NC State University
Department of Computer Science