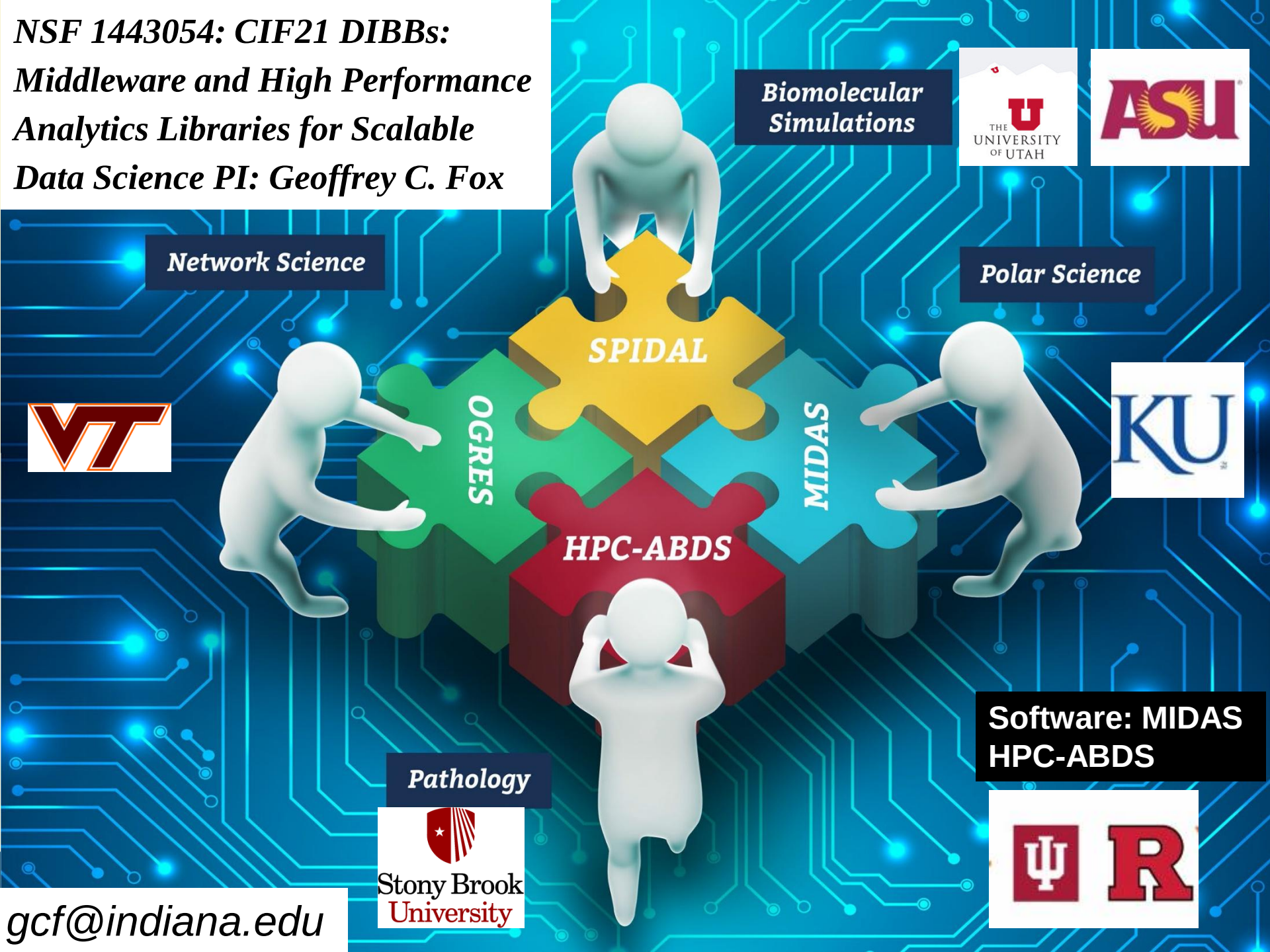
NSF 1443054: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science PI: Geoffrey C. Fox



gcf@indiana.edu

Status of NSF 1443054 Project

- **Big Data Application Analysis** identifies features of data intensive applications that need to be supported in software and represented in benchmarks. This analysis was started for proposal and has been extended to support HPC-Simulations-Big Data convergence.
- The project is a collaboration between computer and domain scientists in **application areas** in Biomolecular Simulations, Network Science, Epidemiology, Computer Vision, Spatial Geographical Information Systems, Remote Sensing for Polar Science and Pathology Informatics.
- **HPC-ABDS** as Cloud-HPC interoperable software with performance of HPC (High Performance Computing) and the rich functionality of the commodity Apache Big Data Stack was a bold idea developed for proposal. We have successfully delivered and extended this approach, which is one of ideas described in Exascale Big Data report.

Status of NSF 1443054 Project

- **MIDAS** integrating middleware that links HPC and ABDS now has several components including an architecture for Big Data analytics, an integration of HPC in communication and scheduling on ABDS; it also has rules to get high performance Java scientific code.
- **SPIDAL** (Scalable Parallel Interoperable Data Analytics Library) now has 20 members with domain specific (general) and core algorithms.
- **Benchmarks**. We reached out to database community with keynote and paper at WBDB2015 Benchmarking Workshop.
- **Language:** SPIDAL **Java** runs as fast as **C++**
- Designed and Proposed **HPCCloud** as hardware-software infrastructure supporting
  - **Big Data Big Simulation** Convergence
  - **Big Data Management** via Apache Stack ABDS
  - **Big Data Analytics** using SPIDAL and other libraries

Current Challenge: Allow Broad Deployment at Scale

- Classic **Software Engineering**:
  - apply **lessons like SPIDAL Java** uniformly to each SPIDAL library member
  - **Package** and **testing** for each component of SPIDAL and MIDAS
- Offer HPC-ABDS Capabilities as Platform as a Service with each capability specified as Ansible role giving **function as a service**
  - Will allow HPC, Cloud and converged infrastructure **HPCCloud**
  - Define good **API**'s to each function – current Apache libraries not well designed
- **Demonstrate test applications** as software as a service on virtual clusters
- Show that **service providers** (XSEDE) can deploy on variety of hardware
- Not clear if **HPCCloud infrastructure** available as must support roles from HPC and Big Data
  - Comet one of best for us but need broader use of platforms to allowfor example biomolecular simulations on Blue Waters

HPC and/or Cloud 1.0 2.0 3.0

- **Cloud 1.0**: IaaS PaaS
- **Cloud 2.0**: DevOps
- **Cloud 3.0**: Insight (Solution) as a Service from IBM; serverless computing; event driven function as a service

- **HPC 1.0** and **Cloud 1.0** separate ecosystems
- **HPCCloud** or **HPC-ABDS**: Take performance of HPC and functionality of Cloud (Big Data) systems
- **HPCCloud 2.0** Use DevOps to invoke HPC or Cloud software on VM, Docker, HPC infrastructure
- **HPCCloud 3.0** Automate Solution as a Service using HPC-ABDS on varied infrastructure suitable for HPC and Big Data Management and Analytics

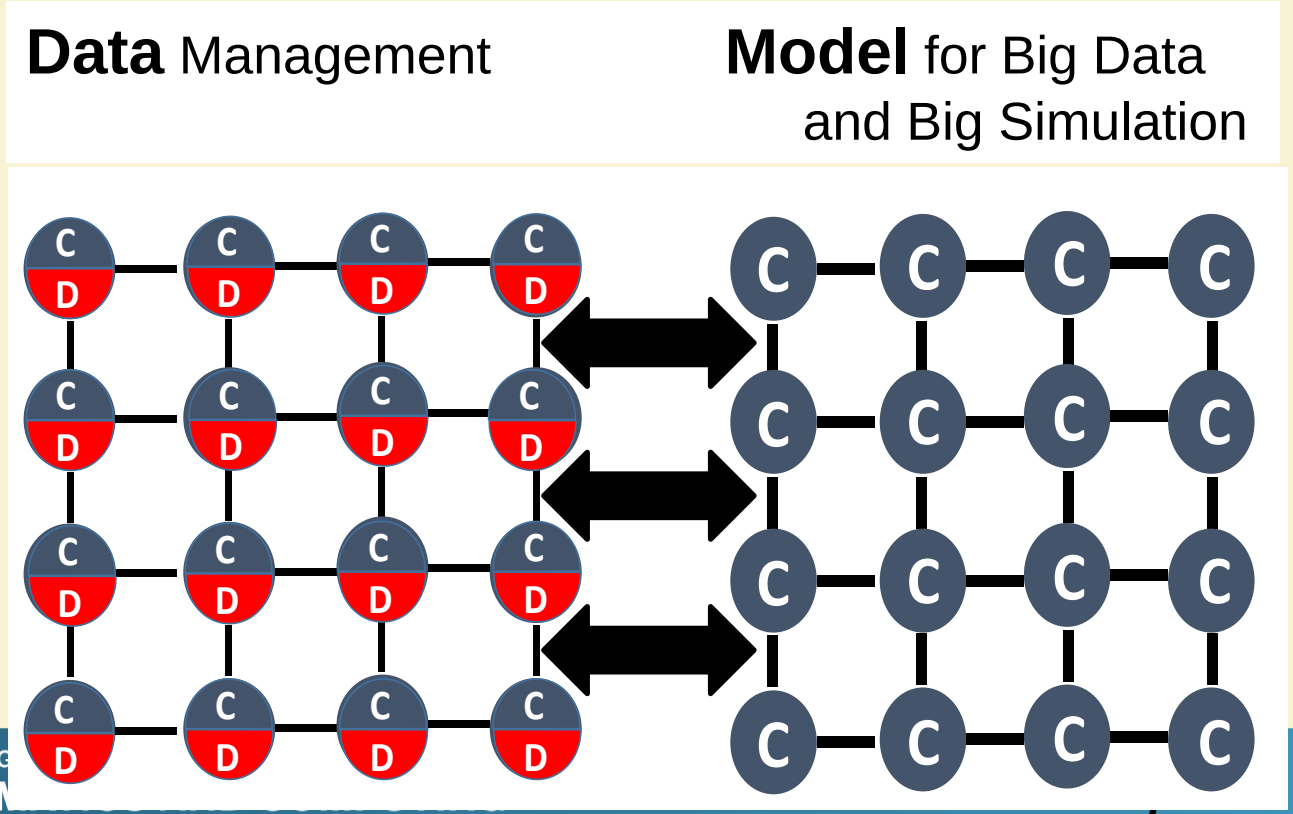
27 Ansible Roles and Re-use in 6 NIST use cases

ID	6 NIST Use Cass	Hadoop	Mesos	Spark	Storm	Pig	Hive	Drill	HDFS	HBase	Mysql	MongoDB	RethinkDB	Mahout	D3, Tableau	nltk	MLlib	Lucene/Solr	OpenCV	Python	Java	maven	Ganglia	Nagios	spark supervisord	zookeeper	AlchemyAPI	R
1	NIST Fingerprint Matching	x		x			x	x		x	x										x	x	x	x	x	x		
2	Human and Face Detection		x	x															x	x						x		
3	Twitter Analysis				x					x		x			x	x				x	x					x	x	x
4	Analytics for Healthcare Data/Health Informatics	x		x						x				x	x		x	x			x					x		
5	Spatial Big Data/Spatial Statistics/Geographic Information Systems	x		x										x	x		x				x	x						
6	Data Warehousing and Data Mining	x		x		x	x			x		x		x	x		x	x			x					x		
	count	4	1	5	1	1	2	1	0	4	1	2	0	3	4	1	3	2	1	2	5	2	3	1	1	5	1	1

HPCCloud Convergence Architecture

- Running **same HPC-ABDS software** across all platforms but data management machines has different balance in I/O, Network and Compute from "model" (data analytics, simulation) machine
  - Note data storage approach: HDFS v. Object Store v. Lustre style file systems is still rather unclear
- **The Model** behaves similarly whether from **Big Data or Big Simulation**.

HPCCloud
Operational Model matches hardware features with application requirements



Future Challenge: Support and Encourage Deployment

- Even as the middleware and analytics are being developed and properly packaged we need to address
  - **Supported deployment**
  - User **training**
  - proactive outreach to **users** and **service providers**.
- We need to consider traditional library as well as PaaS/FaaS and SaaS deployments.
- We will give a 6 hour tutorial on MIDAS and SPIDAL at a European winter school in February 2017 and this will increase our work in this area.