Applications of HPCC Systems at Clemson University

Amy Apon, PhD • Linh Ngo, PhD • Michael Payne
Big Data Systems Laboratory
Clemson University

CLEMSON UNIVERSITY

BDSLab

LexisNexis™

Applications of HPCC Systems at Clemson University

Clemson Strengths and Opportunities

People

PhD-level faculty & research staff
Talented students
Significant industry collaborators

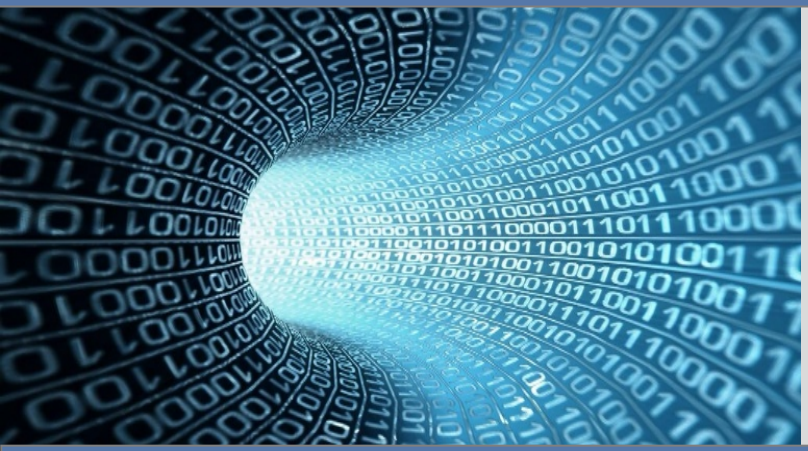
Facilities

Palmetto – Top 5 in US Academic Supercomputers
~2000 nodes, 20K cores, 600 GPUs
100Gb Internet connectivity

Applications of HPCC Systems at Clemson University

Big Data Systems Lab Overview

Big Data Systems Lab Vision



Perform World Class Research on the Systems and Enabling Information Technology for Advanced Data Analytics

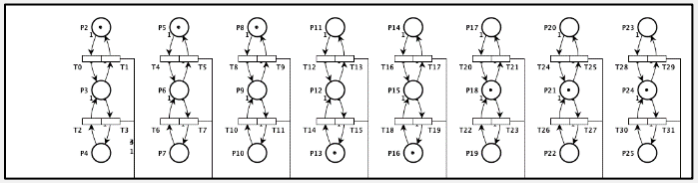
Big Data Systems Lab Research Areas

Systems and Architectures

Tools and Operations

LexisNexis HPCC Systems

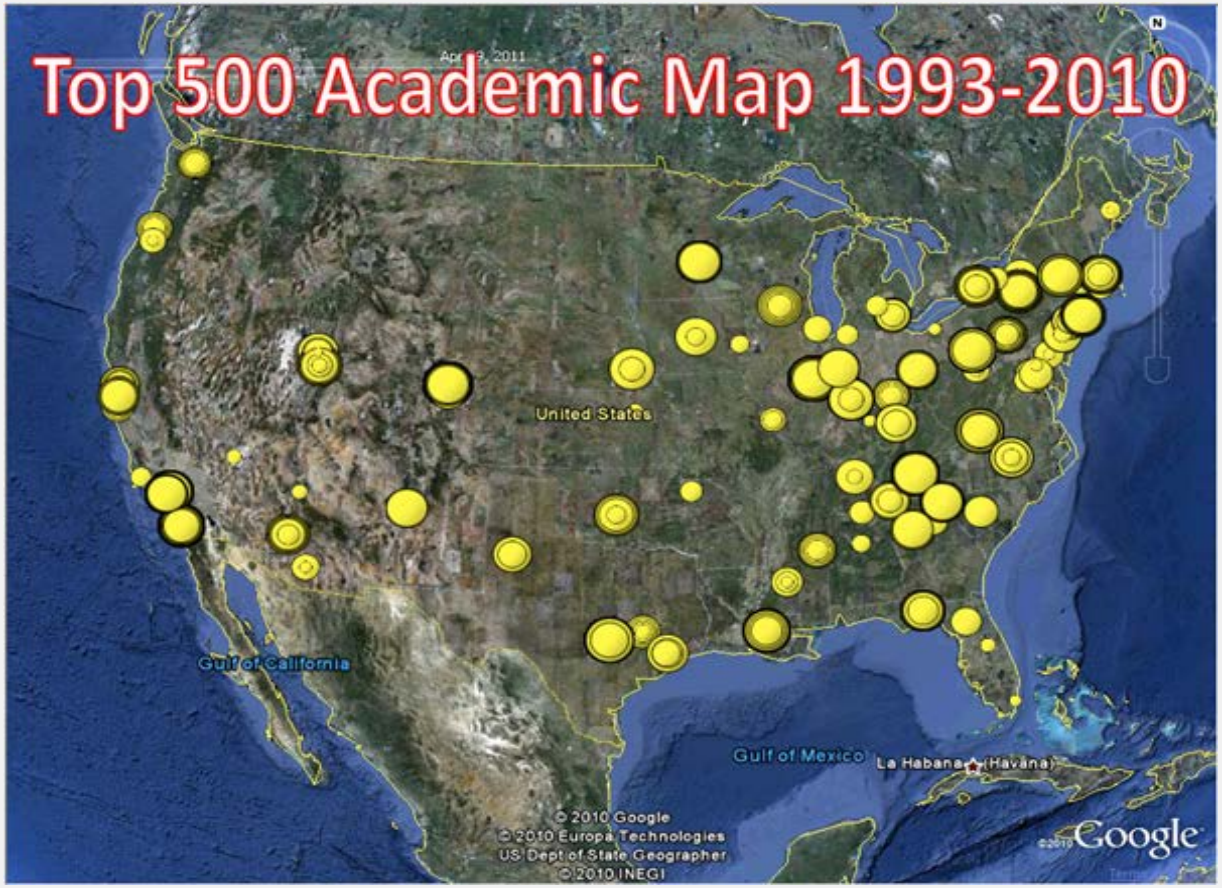
Data Analytics and Applications



CLEMSON UNIVERSITY BDSLab LexisNexis™

Applications of HPCC Systems at Clemson University

Effect of High Performance Computing on Academic Research Productivity



Motivation: There is a lot of pressure on federal funding

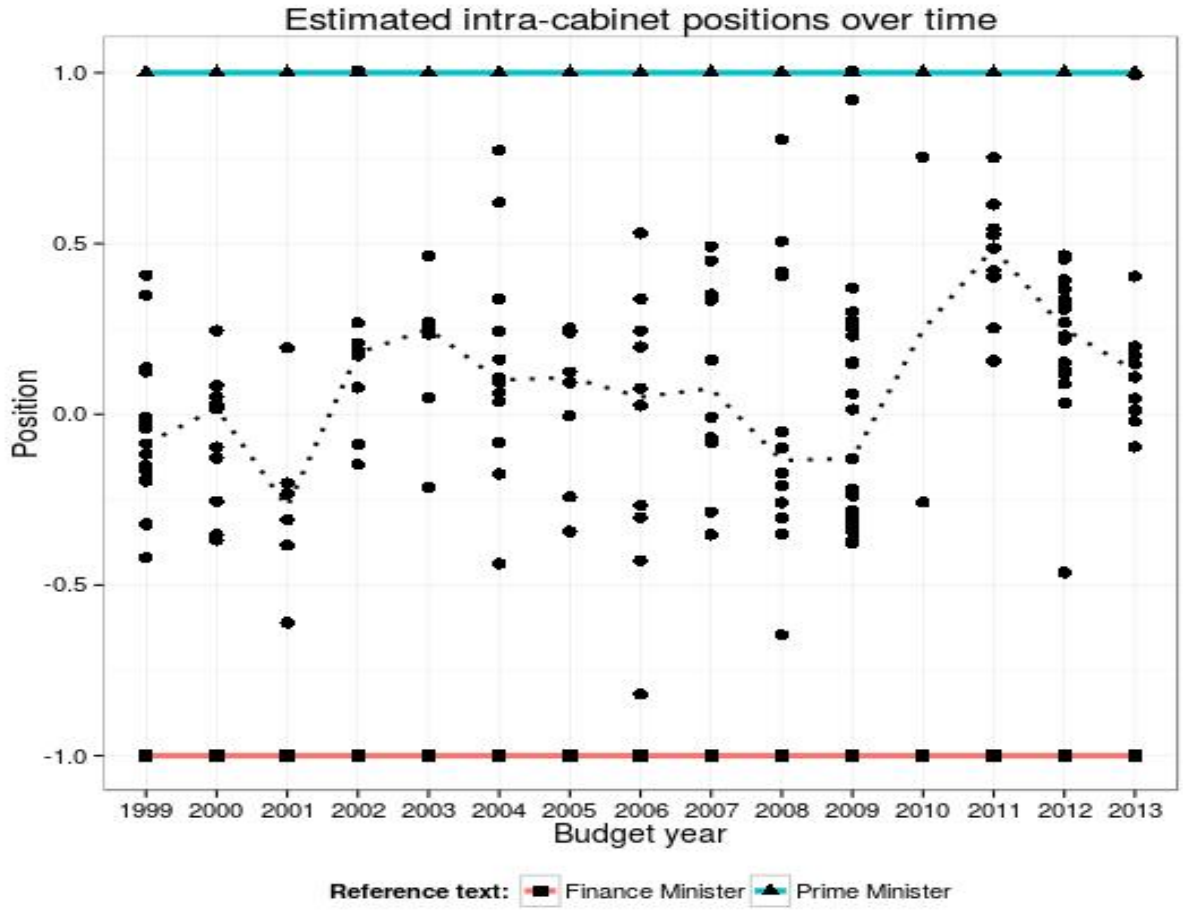
We propose *efficiency* as a measure from which to gain insights on return on investment

We show that locally-available HPC has a positive effect on the ability of a university to do research

CLEMSON UNIVERSITY BDSLab LexisNexis™

Applications of HPCC Systems at Clemson University

Text mining of news reports and social media for business intelligence

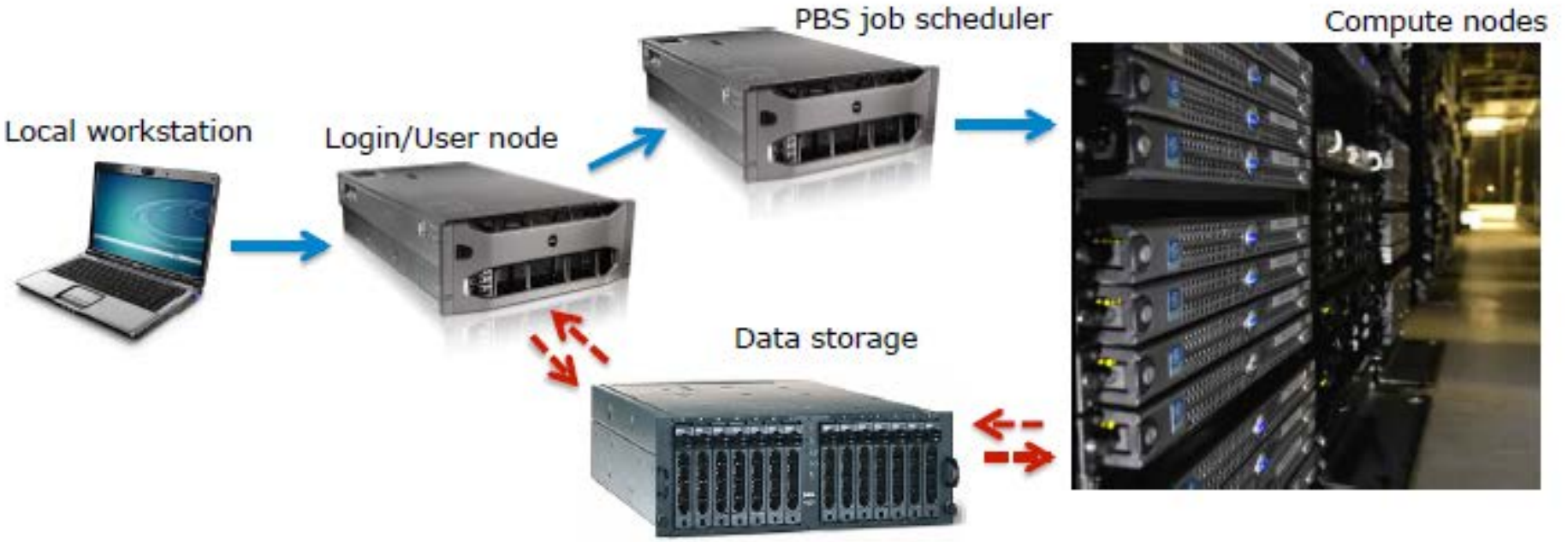


Motivation: Government and business need information about public sentiment.

Research: We develop and apply methods to analyze large amounts of textual data to enable inquiry of social and business problems.

CLEMSON UNIVERSITY BDSLab LexisNexis™

Applications of HPCC Systems at Clemson University
Shared Computing Resources among Researchers



Shared Execution Environment

Temporary Local Storage

User Privileges Only

Applications of HPCC Systems at Clemson University

Linh Ngo, PhD

HPCC Systems in a Shared Research Computing Environment

CLEMSON UNIVERSITY BDSLab LexisNexis™

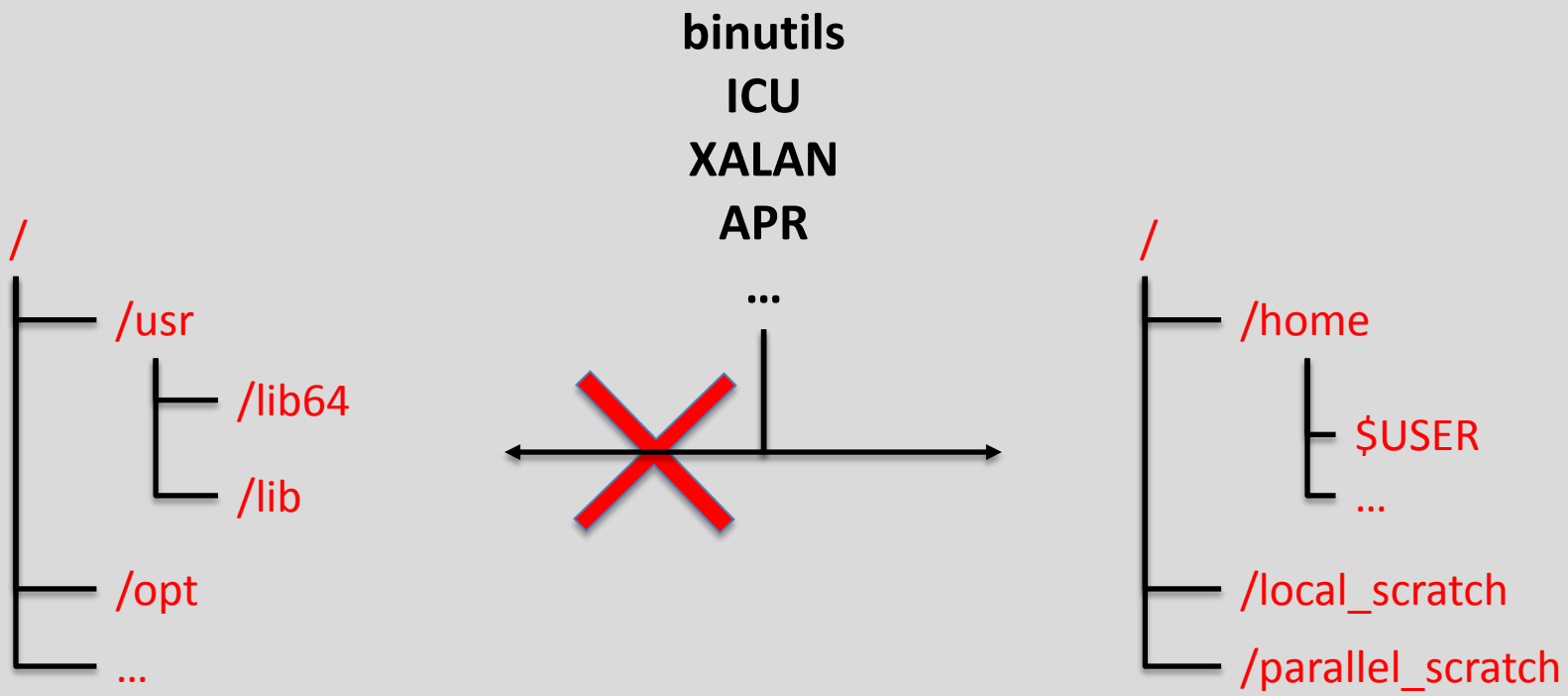
Applications of HPCC Systems at Clemson University
Shared Computing Resources among Researchers

How to provision and configure an HPCC cluster dynamically for research purposes?

- **Step 1: Configure, install, and deploy HPCC as a non-root user**
- **Step 2: Dynamically provision HPCC cluster in a shared research environment**

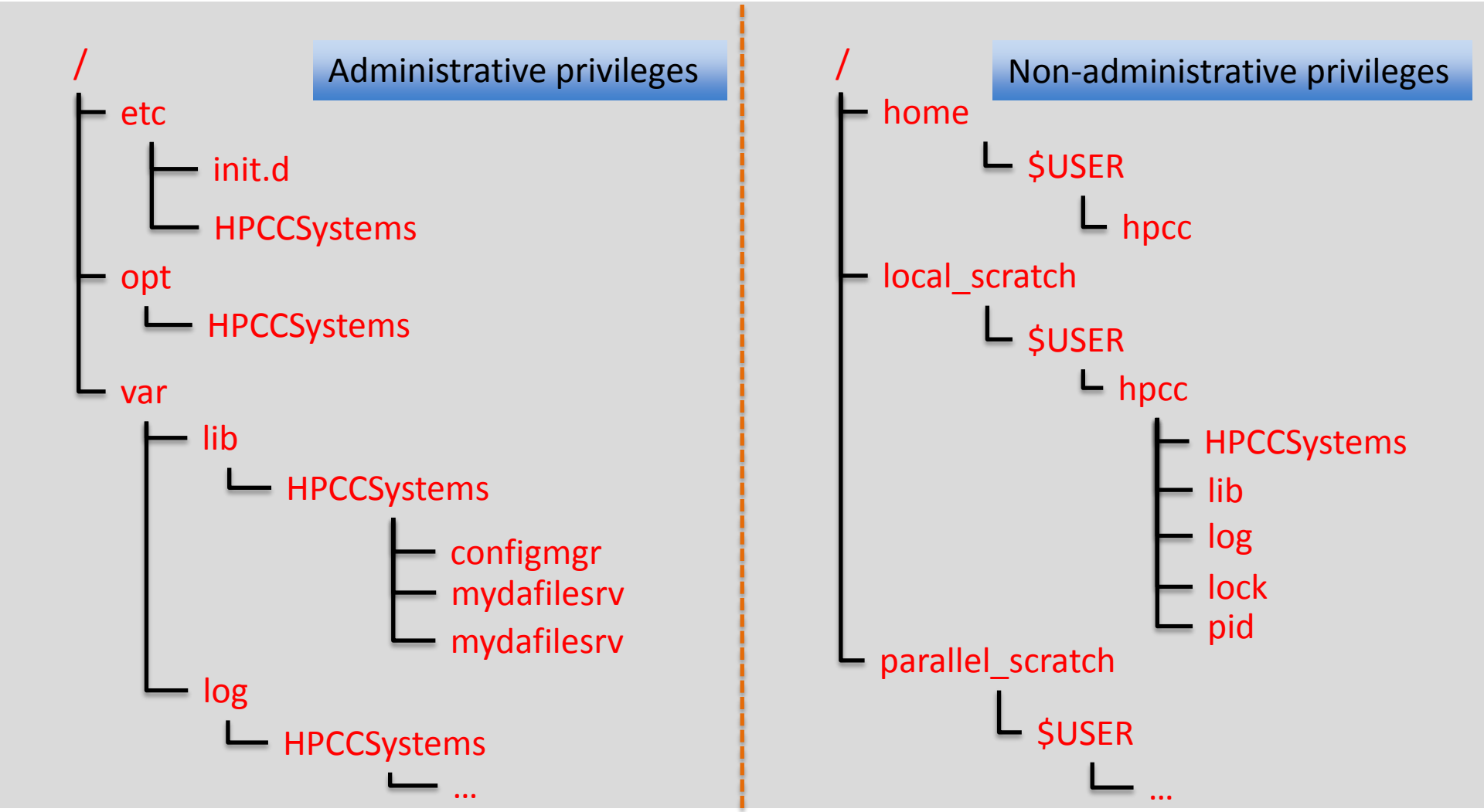
Applications of HPCC Systems at Clemson University

Installation and Configuration of Dependencies



Applications of HPCC Systems at Clemson University

Resolving Non-default Installation Path Conflicts



Applications of HPCC Systems at Clemson University

Non-root Deployment

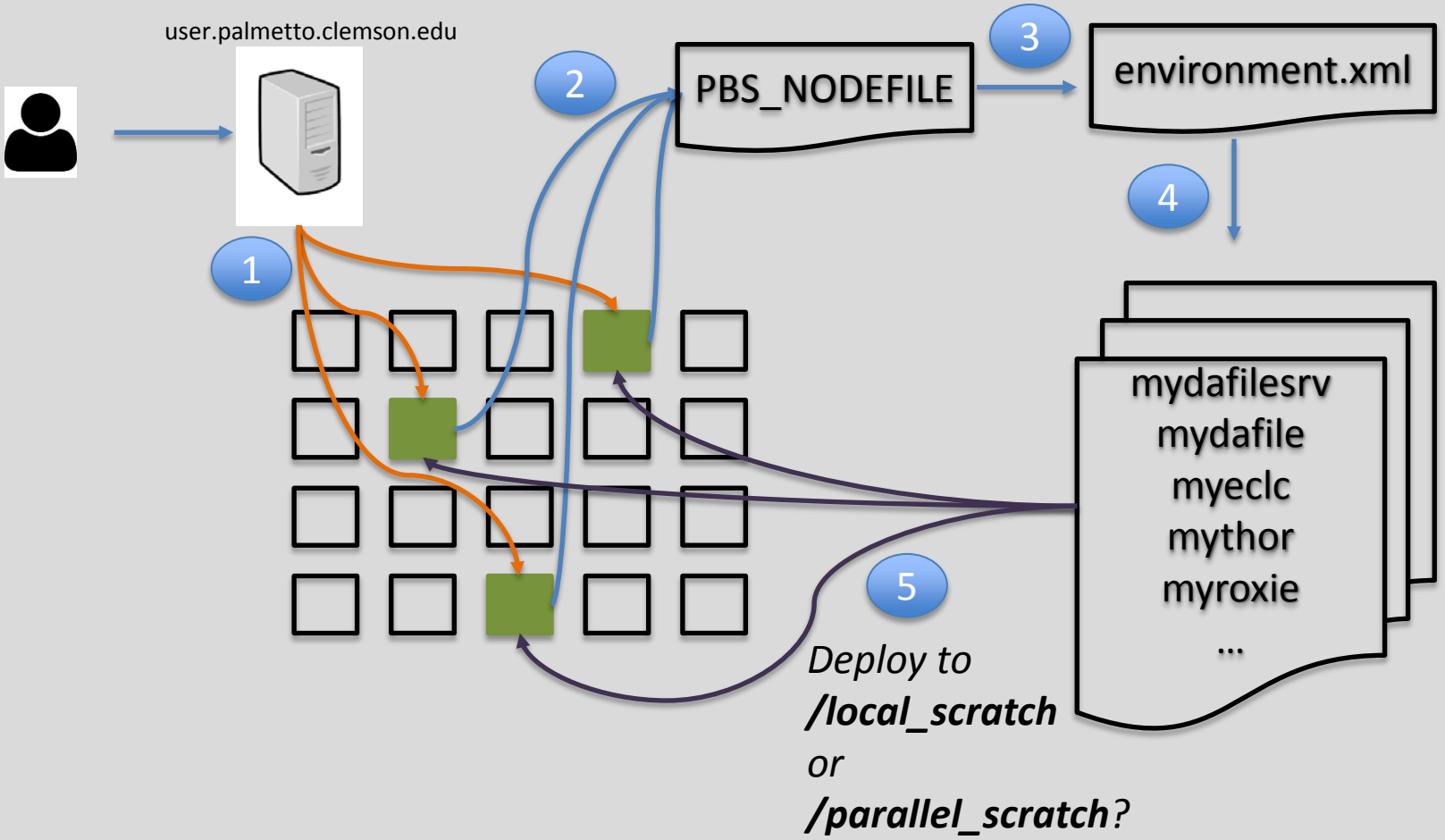
Remove/relax root-level settings:

i.e.: is_root

Reduce default configuration settings for resource requirements:

depended on resource allocation requests

Applications of HPCC Systems at Clemson University
Dynamic Provisioning



Applications of HPCC Systems at Clemson University

Michael Payne

Using HPCC Systems to Manage Academic Data

LexisNexis Summer 2014 Internship

LexisNexis
HPCC Systems

CLEMSON UNIVERSITY BDSLab LexisNexis

Applications of HPCC Systems at Clemson University

Using HPCC Systems to Manage Academic Data

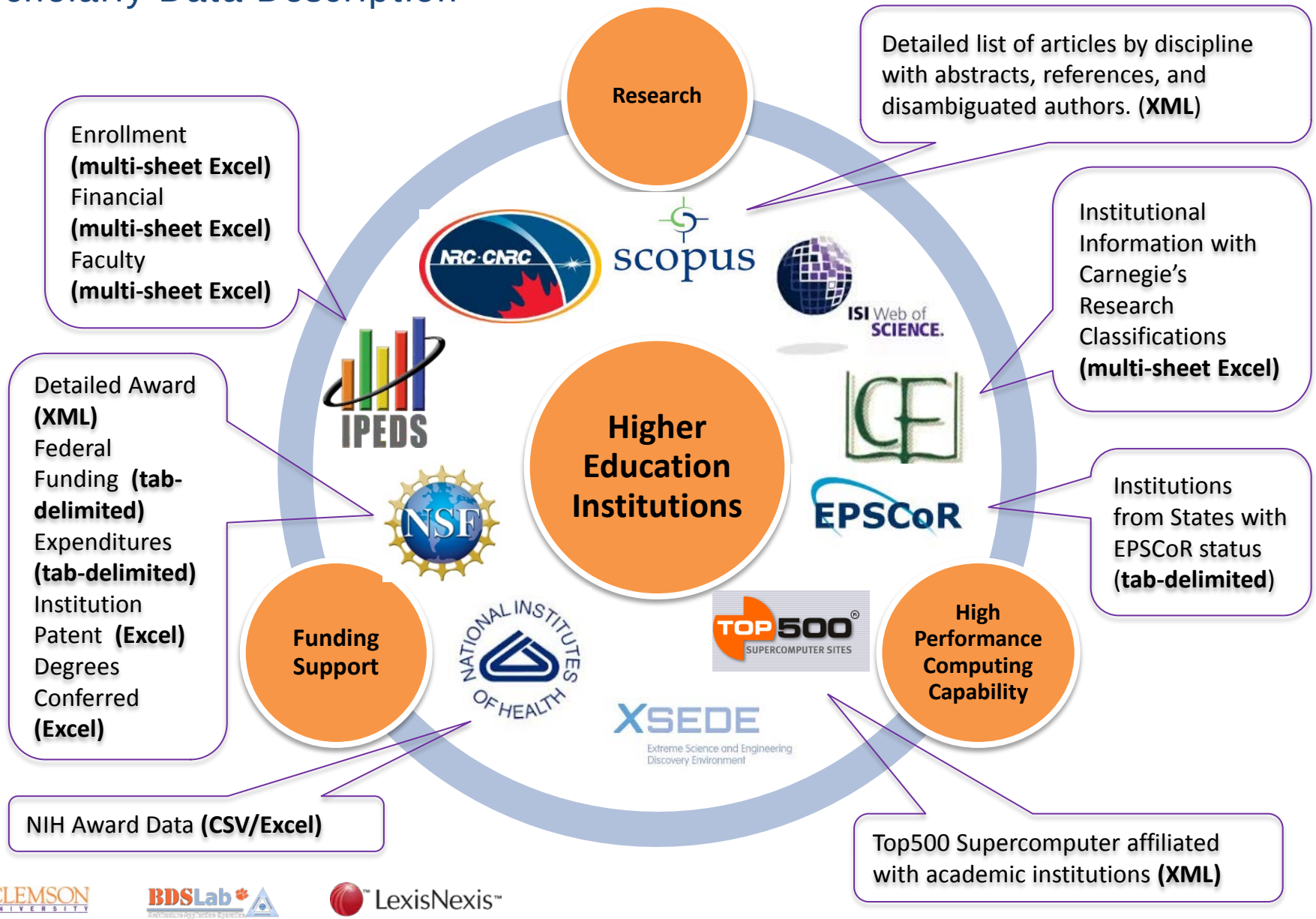
- Research in Scholarly Data requires academic data from many different sources, which store data under various formats
- Aggregating these sources into a useful and cohesive structure requires a data-intensive approach to preprocessing, integration and analysis
- HPCC Systems is a platform to streamline this process

Applications of HPCC Systems at Clemson University

Categories of Scholarly Data

Applications of HPCC Systems at Clemson University

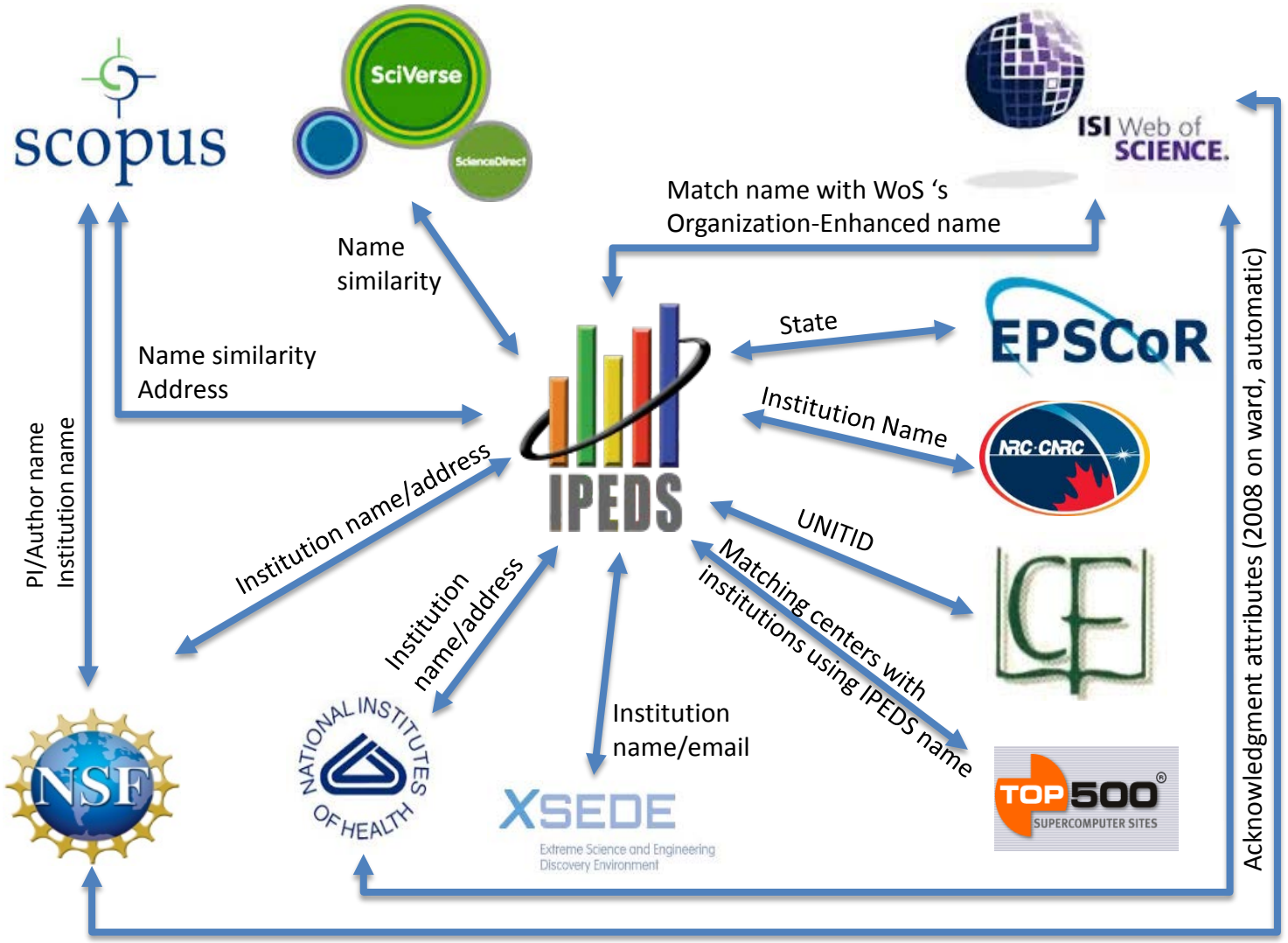
Scholarly Data Description



CLEMSON UNIVERSITY BDSLab LexisNexis™

Applications of HPCC Systems at Clemson University

Examples of Scholarly Data Links



Applications of HPCC Systems at Clemson University
Ongoing Work

- Porting data analytic processes to ECL
- Applying Machine Learning techniques for article abstract classification

Applications of HPCC Systems at Clemson University

Summer 2014 Internship - Logistic Regression for Dense Matrices

LexisNexis Internship
Machine Learning

Manager
Timothy Humphrey

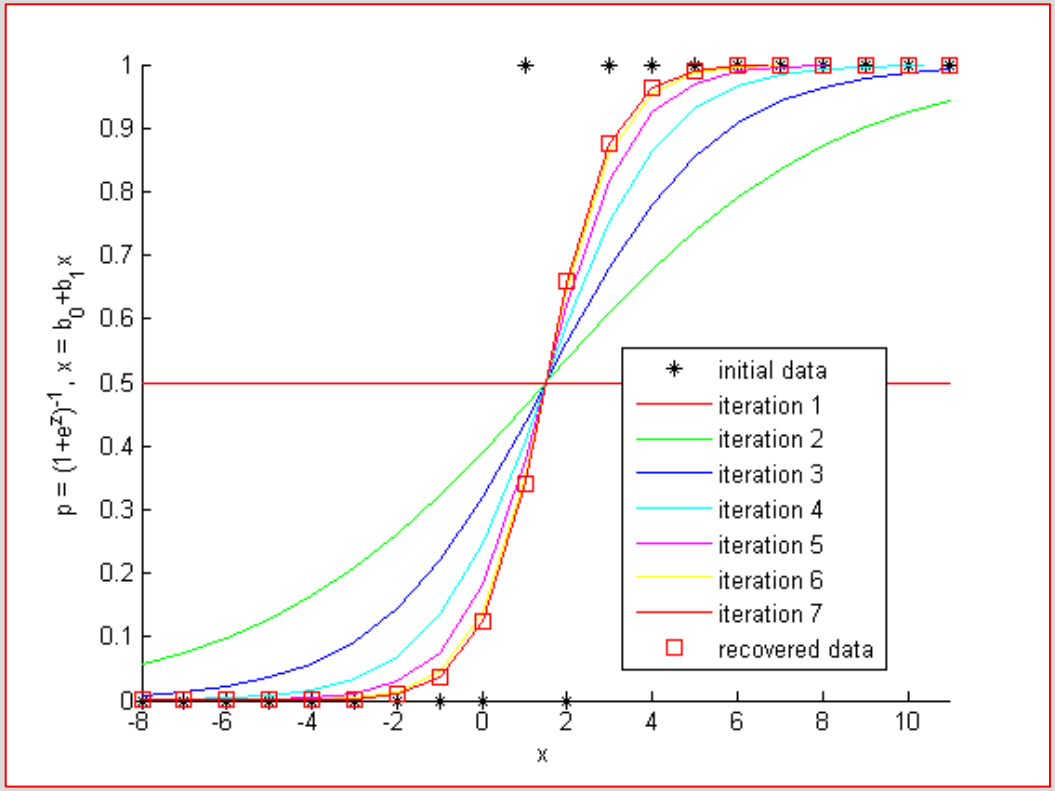
Mentor
Arjuna Chala

CLEMSON UNIVERSITY BDSLab LexisNexis™

Applications of HPCC Systems at Clemson University

Logistic Regression

- Prediction using continuous and discrete values
- No distributional assumptions on the predictors
  - May not be normally distributed or linearly related
- Relationship between the discrete variable and the predictor is non-linear



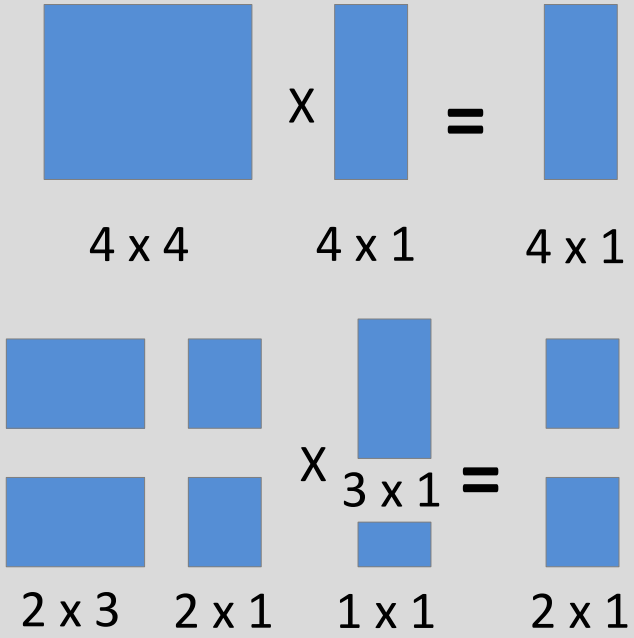
Applications of HPCC Systems at Clemson University

Parallel Block Basic Linear Algebra Subprograms (PB-BLAS)

Matrices can be partitioned

Schemes must be compatible

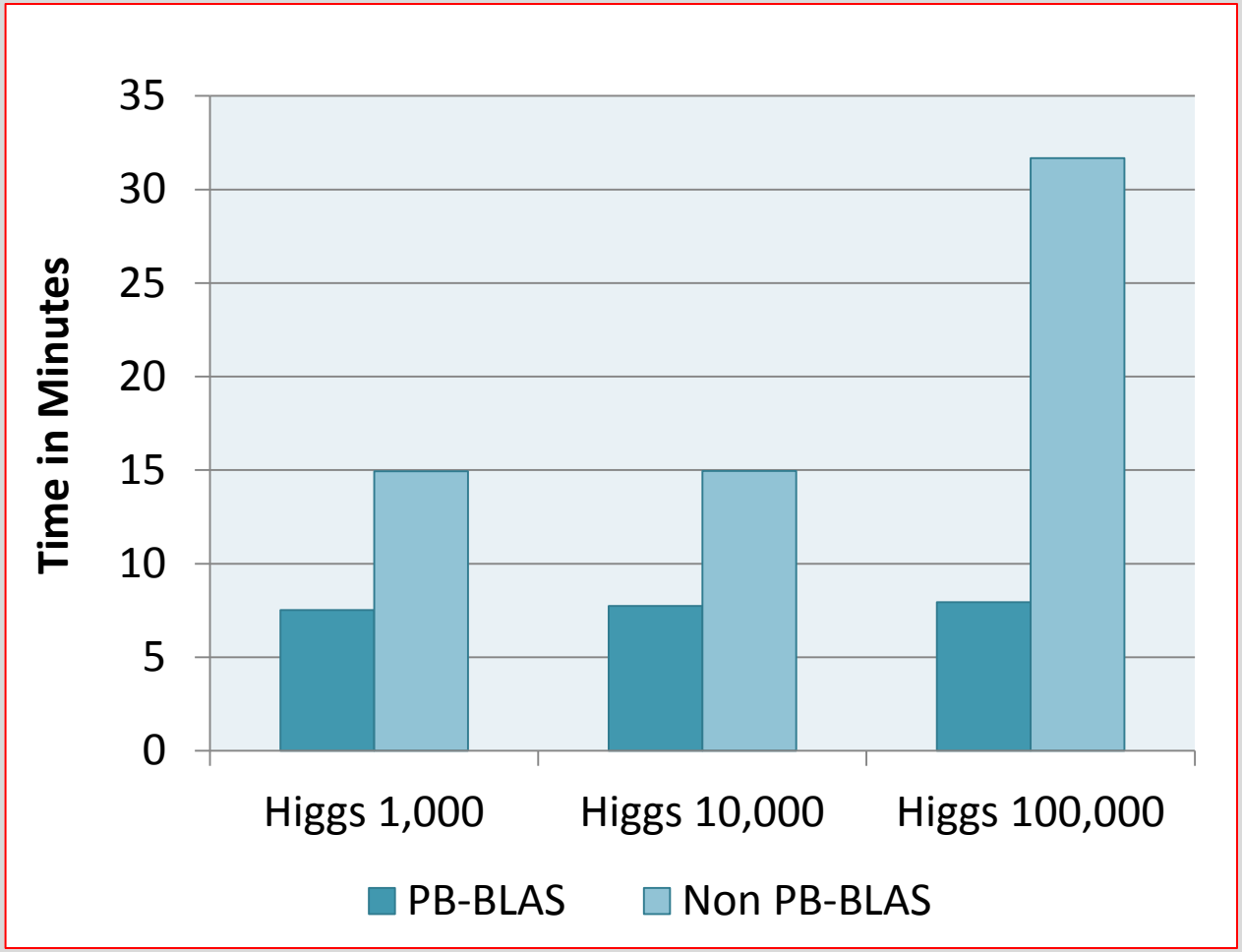
There are multiple choices!



Applications of HPCC Systems at Clemson University

Machine Learning in ECL

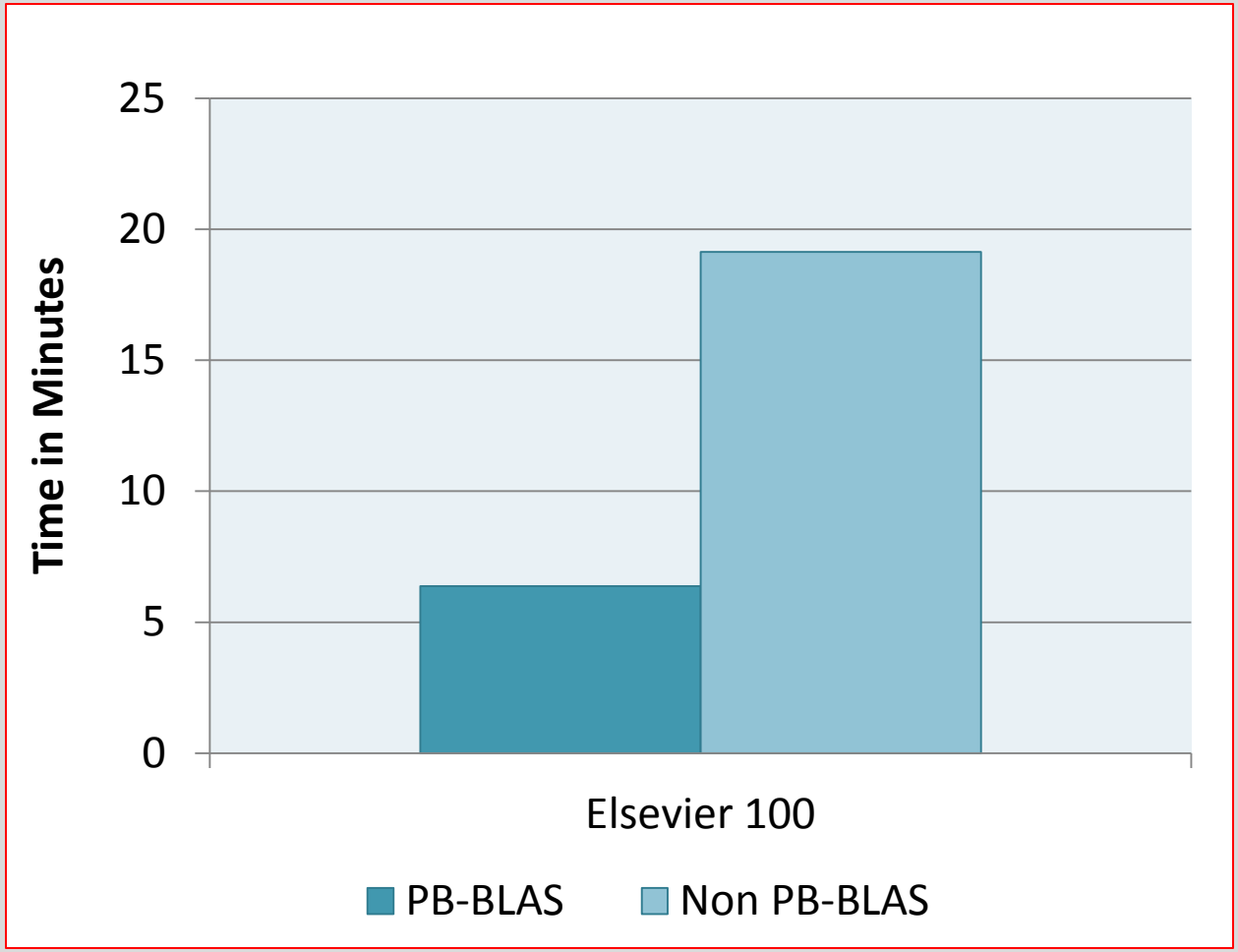
- Logistic Runtimes
- Hard Coded Mapping
- Full Higgs Dataset 11,000,000 x 28



Applications of HPCC Systems at Clemson University

Machine Learning in ECL

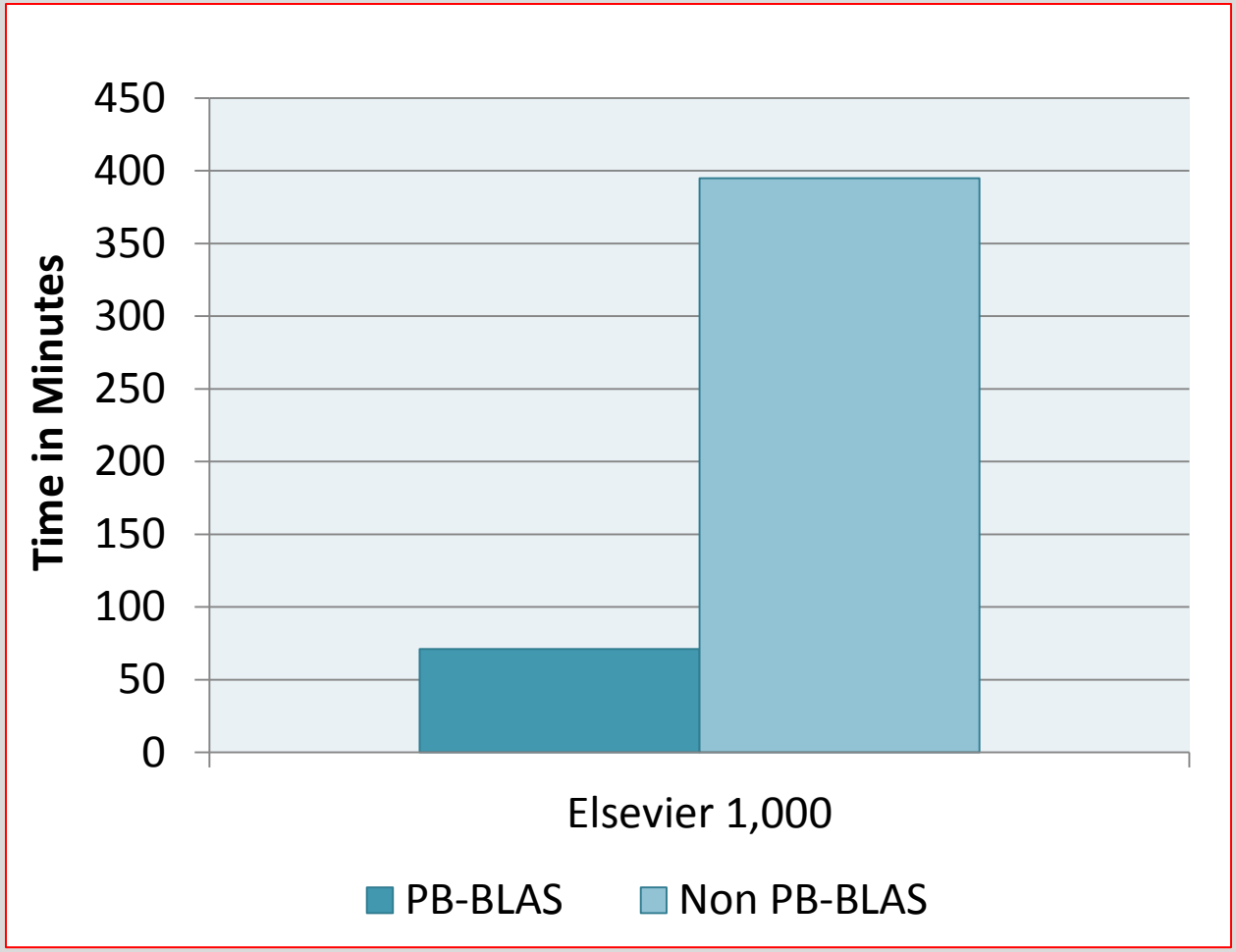
- Logistic Runtimes
- Auto Mapping
- Full Elsevier Dataset 100,000 x 3,291



Applications of HPCC Systems at Clemson University

Machine Learning in ECL

- Logistic Runtimes
- Auto Mapping
- Full Elsevier Dataset 100,000 x 3,291



Applications of HPCC Systems at Clemson University

Project Summary

- Logistic Regression code and supporting functions have been documented and merged to ECL-ML GitHub repository
- Auto block vector mapping function for any user that wants to use PB-BLAS
- Ready to use element wise multiplication in PB-BLAS
- Updated debugging statements that a clear understanding of errors
- Test functions for both block vector mapping function
- Sample code for using logistic regression
- Currently working on K-means implementation that utilizes PB-BLAS

Linh Ngo, PhD • Alex Herzog, PhD • Michael Payne • Amy Apon, PhD
{lngo, aherzog, mpayne3, aapon}@clemson.edu

Big Data Systems Laboratory
Clemson University