What is Big Data and what can we do with it?

An applications and higher education policy view

Women's Institute for Summer Enrichment
Cornell University, June 18, 2014

Amy Apon, Ph.D., Professor and Chair
Computer Science Division, School of Computing
Clemson University

Main Ideas of Talk

- There are a lot of data out there
- Solutions to many problems important to society can be advanced with Big Data
- University researcher's use of Big Computing is correlated with higher research productivity
- Work is needed in the development of curricula and course materials for Big Data

CLEMSON UNIVERSITY

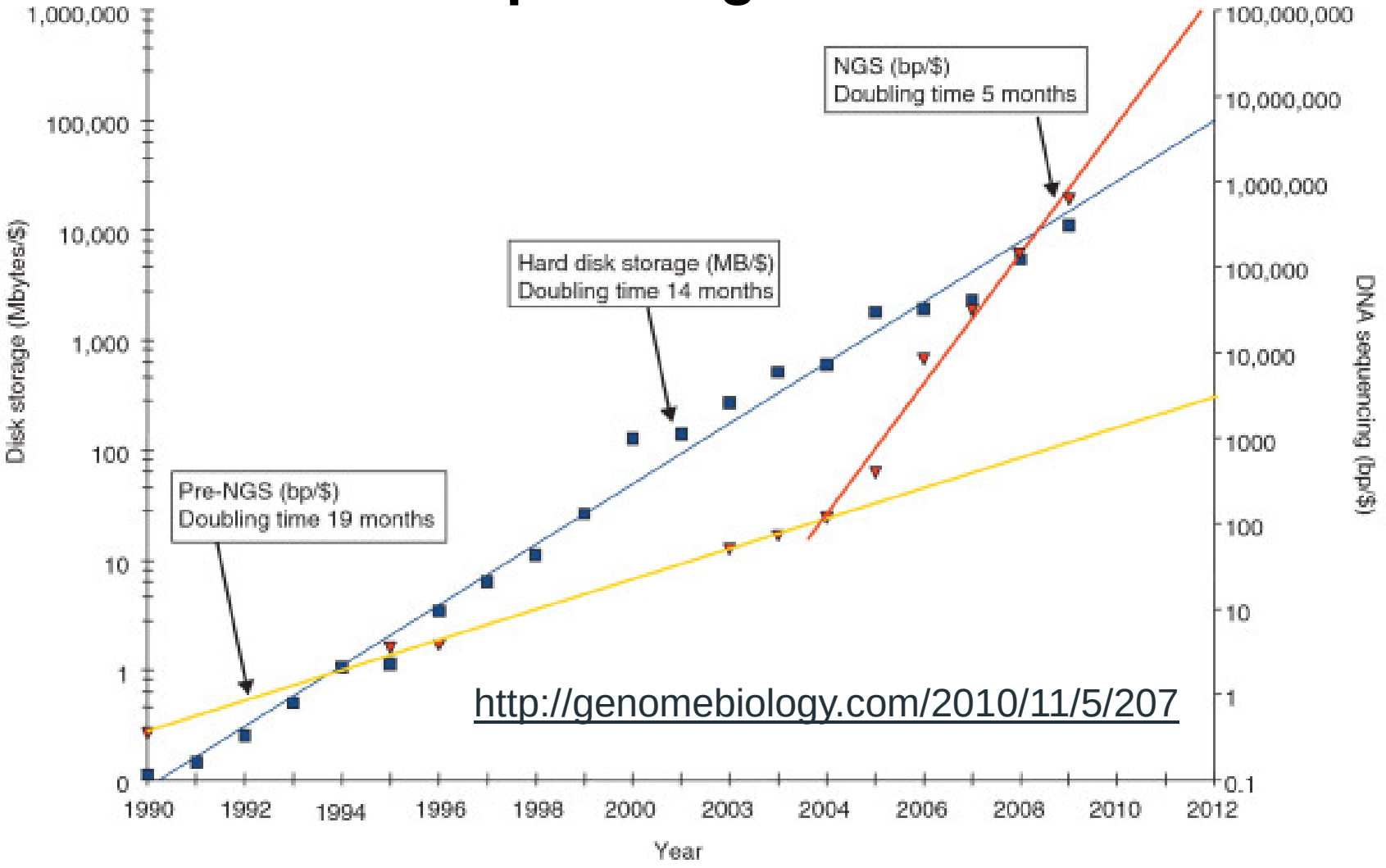
First Takeaway:
There are a lot of data out there.

An Early Big Data System

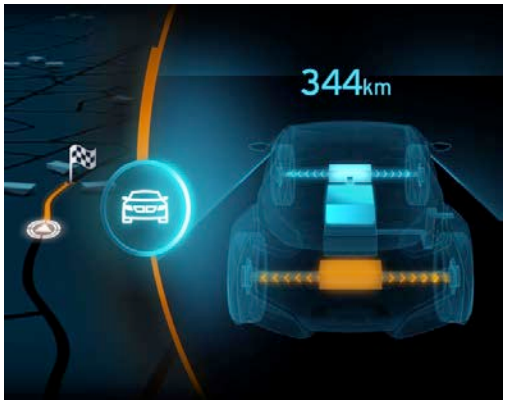

http://landsat.gsfc.nasa.gov/

Social Media

Historical trends in storage prices versus DNA sequencing costs



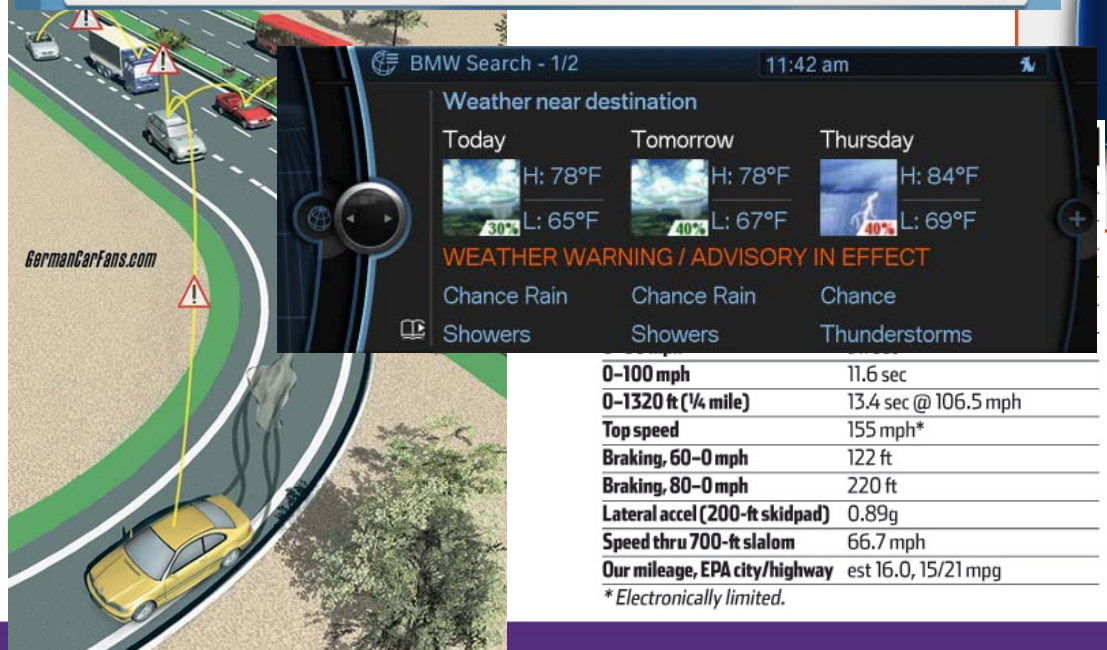
Big Automotive Data



- Data in every car!
  - On-board diagnostic data
  - Up to 85 electronic Control Units (ECSs)
  - Up to 1,500 individual selection options make practically each car electrically unique
  - Up to 25GB software,
  - About 2,000 customer functions
  - Up to 13,000 diagnostic readouts
- Data about the car starts in the assembly plant or earlier!

CLEMSON UNIVERSITY

Tremendous Volume
Extreme Velocity
"Endless" Variety



0–100 mph	11.6 sec
0–1320 ft (¼ mile)	13.4 sec @ 106.5 mph
Top speed	155 mph*
Braking, 60–0 mph	122 ft
Braking, 80–0 mph	220 ft
Lateral accel (200-ft skidpad)	0.89g
Speed thru 700-ft slalom	66.7 mph
Our mileage, EPA city/highway	est 16.0, 15/21 mpg

** Electronically limited.*

Data in the car
Data from the car
Data between cars

What can we do with more data?

- Predictions using simple techniques and large data can outperform predictions using complex techniques and fewer data
  - "The unreasonable effectiveness of data," [Havely 2009]

Everyone has questions and ideas

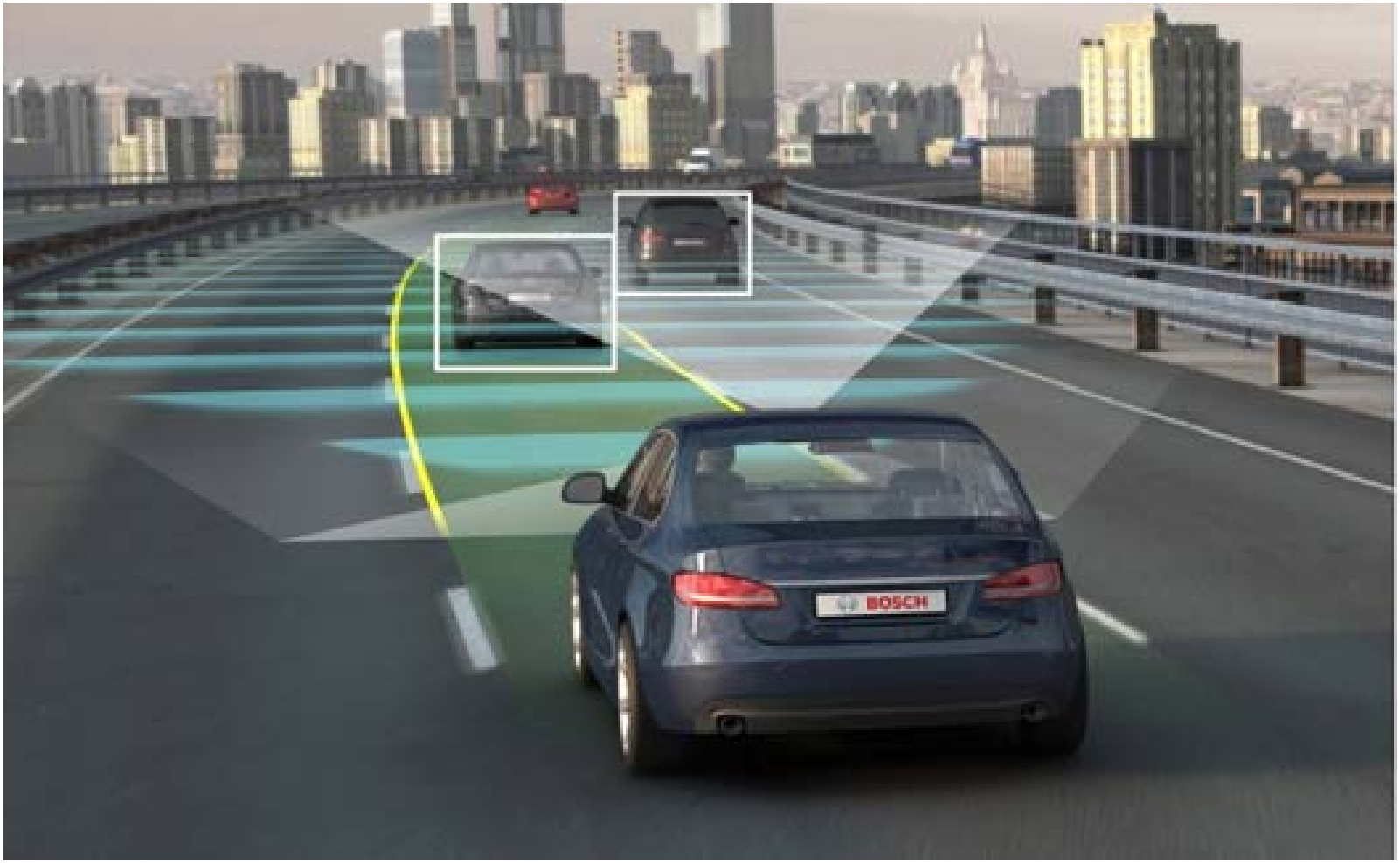
- What research can we accomplish?
- Do we understand the question?
- Is our data accurate? If not, how can I use that information?
- Can I use what I have?

CLEMSON UNIVERSITY

Second Takeaway:
Solutions to many problems that are important to society can be advanced with Big Data

CLEMSON UNIVERSITY

Intelligent Transportation and Automotive Systems



Find the trends and anomalies in warranty and assembly data

Predict failures *before* they become widespread problems.

Goal is to fuse data from several sources to build a model of anomalies and trends

salesforce chatter

TOYOTA

Mining Customer Satisfaction

JD Powers measures new vehicle quality:

- 83,000 questionnaires in total
- 4,636 questionnaires for BMW GROUP vehicles

Defines two types of issues:

- Defects and Malfunctions
- Design

New technologies and usability causes more problems: in average 2013 vehicle report 20% more problems.

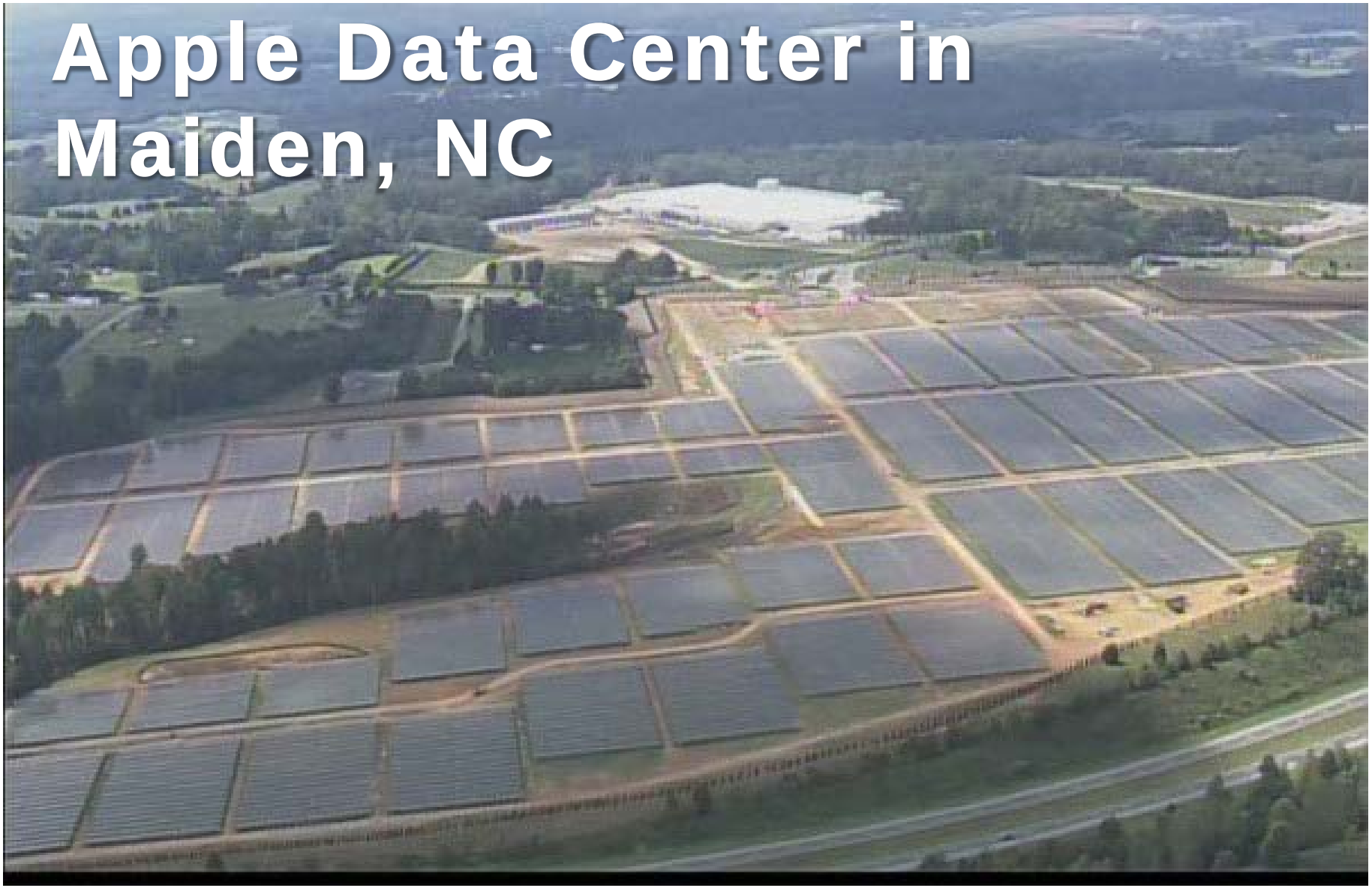
But, can we get Big Social Data to find these answers more quickly?

What is needed to make effective use of Big Data?

- Access to computing resources as a starting point
- Tools for data infrastructure and analysis
- Education and training

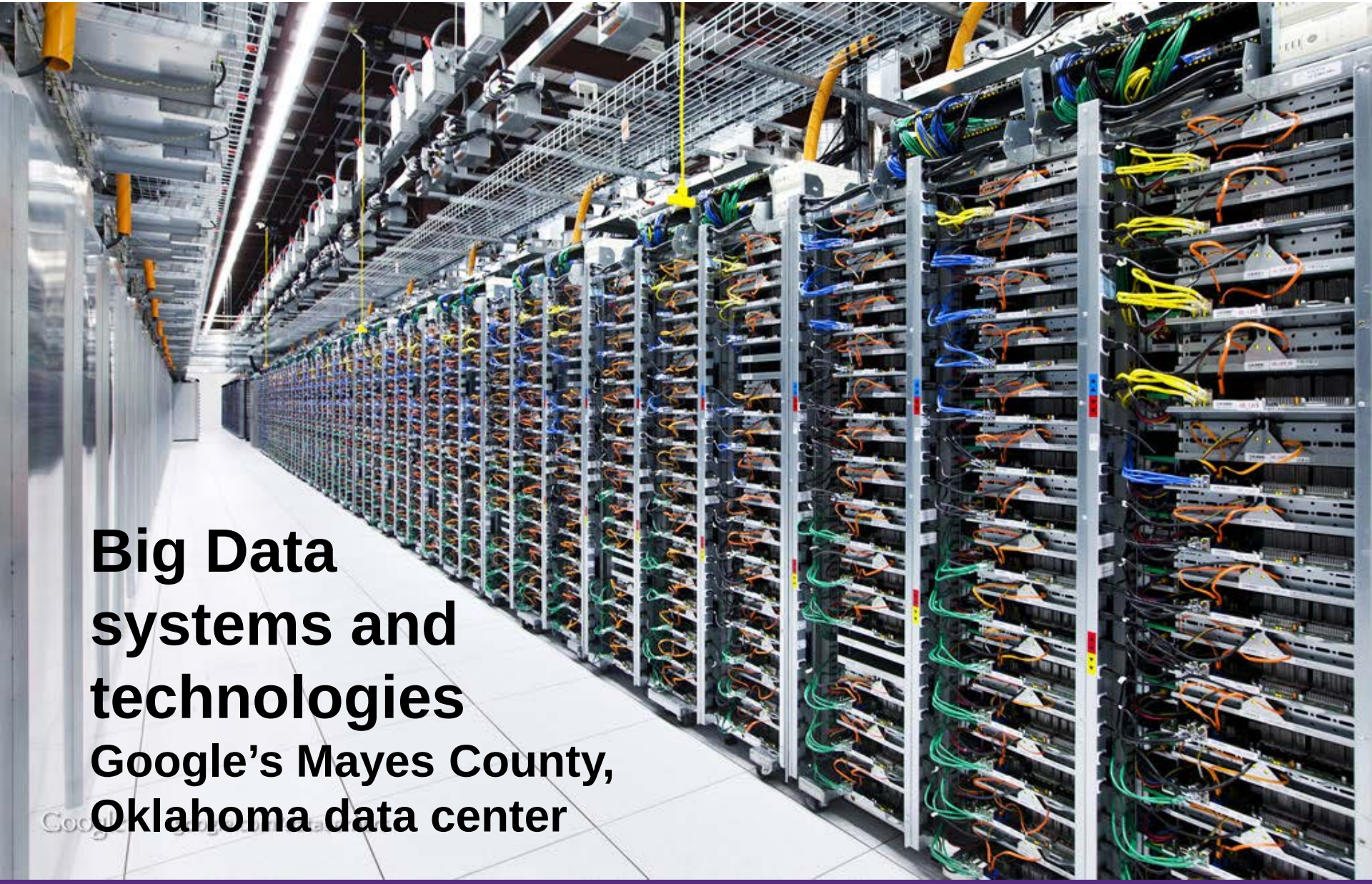
CLEMSON UNIVERSITY

Apple Data Center in Maiden, NC



http://appleinsider.com/articles/12/09/14/aerial_photos_show_apples_massive_nc_solar_farm_near_completion

CLEMSON UNIVERSITY



Big Data systems and technologies

Google's Mayes County, Oklahoma data center

CLEMSON UNIVERSITY

What are the effects of HPC resources on research productivity?

- Are departments with local access to HPC instrumentation more efficient at producing research than those without local access?
- We used Big Data to study Big Data ☺
- Data sources include:
  - Data from the National Research Council – counts of faculty, publications
  - Top 500 listing
  - Award data from the National Science Foundation

What are the effects of HPC resources on research productivity?

- Statistical methods
  - Data envelopment analysis
  - Inputs: count of faculty, GRE scores of new graduate students
  - Output: count of graduates and publications
- Results show that departments in universities that have local access to Top 500 HPC resources are more efficient for Chemistry, Physics, and Civil Engineering. This is not true for Computer Science.

CLEMSON UNIVERSITY

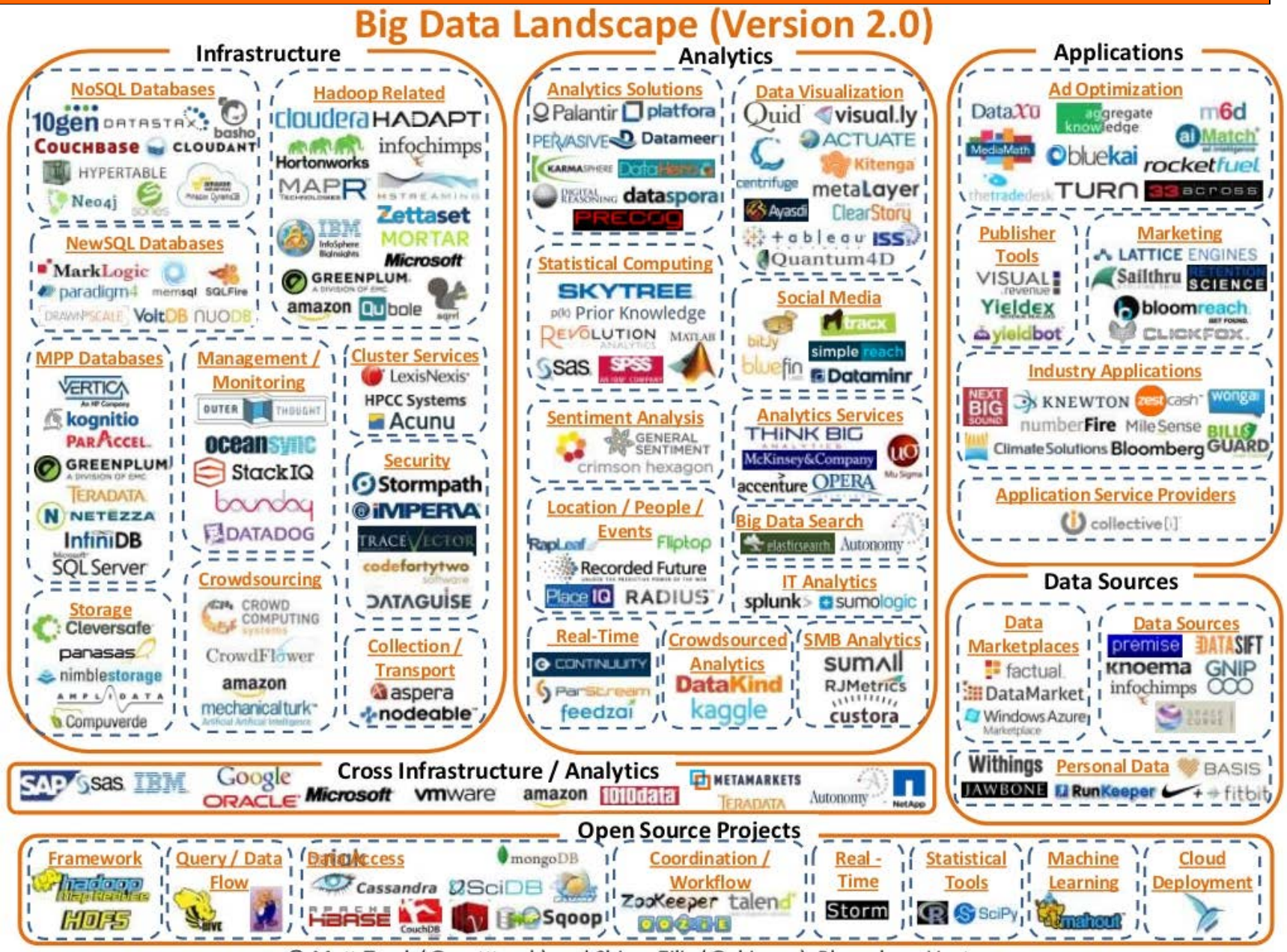
Third Takeaway:
Use of Big Computing at universities is correlated with higher research productivity

Typical Large-Data Approach

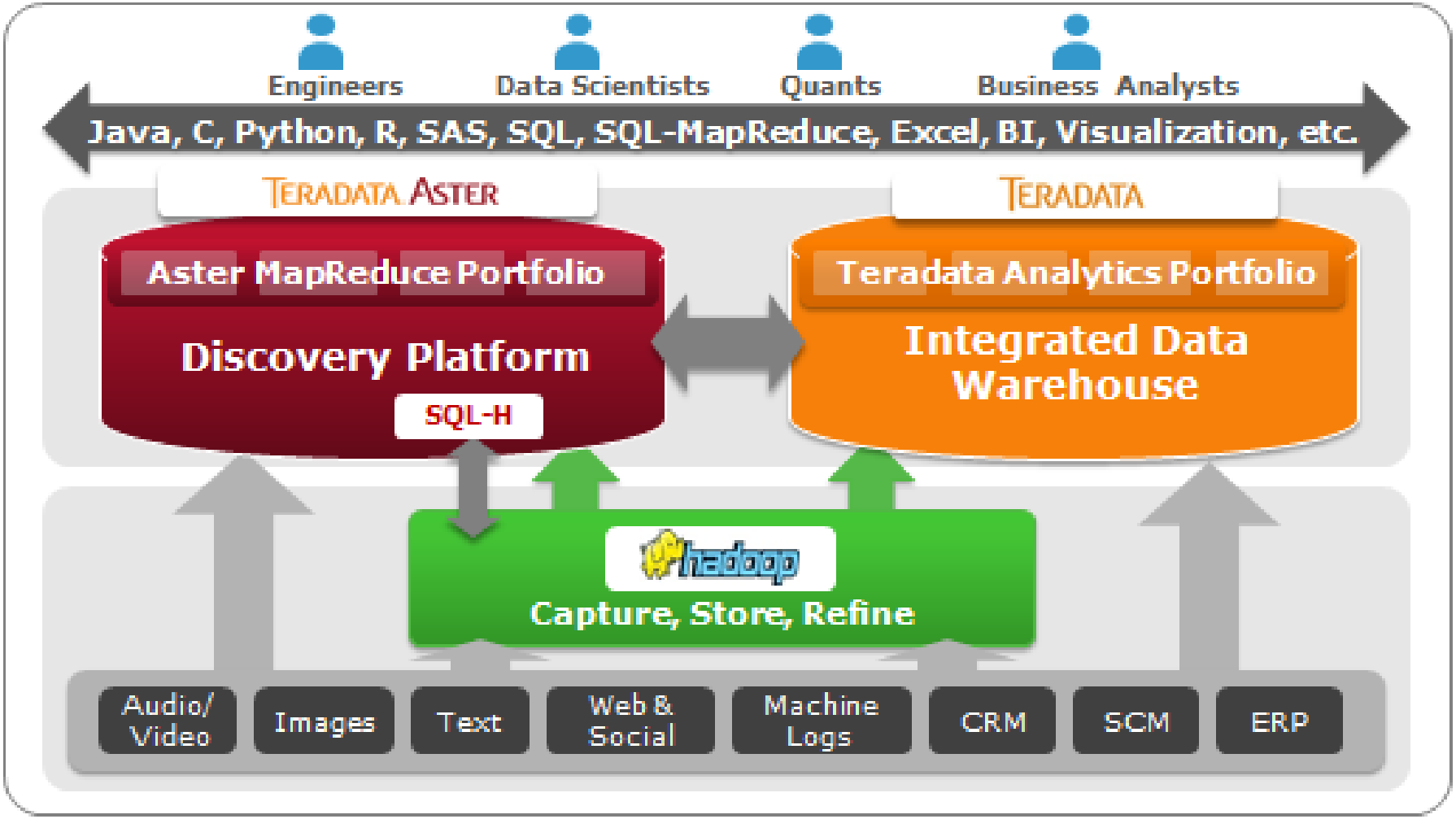
- Iterate over a large number of records
- Extract data of interest from each record
- Shuffle and sort intermediate results
- Aggregate intermediate results
- Generate final output

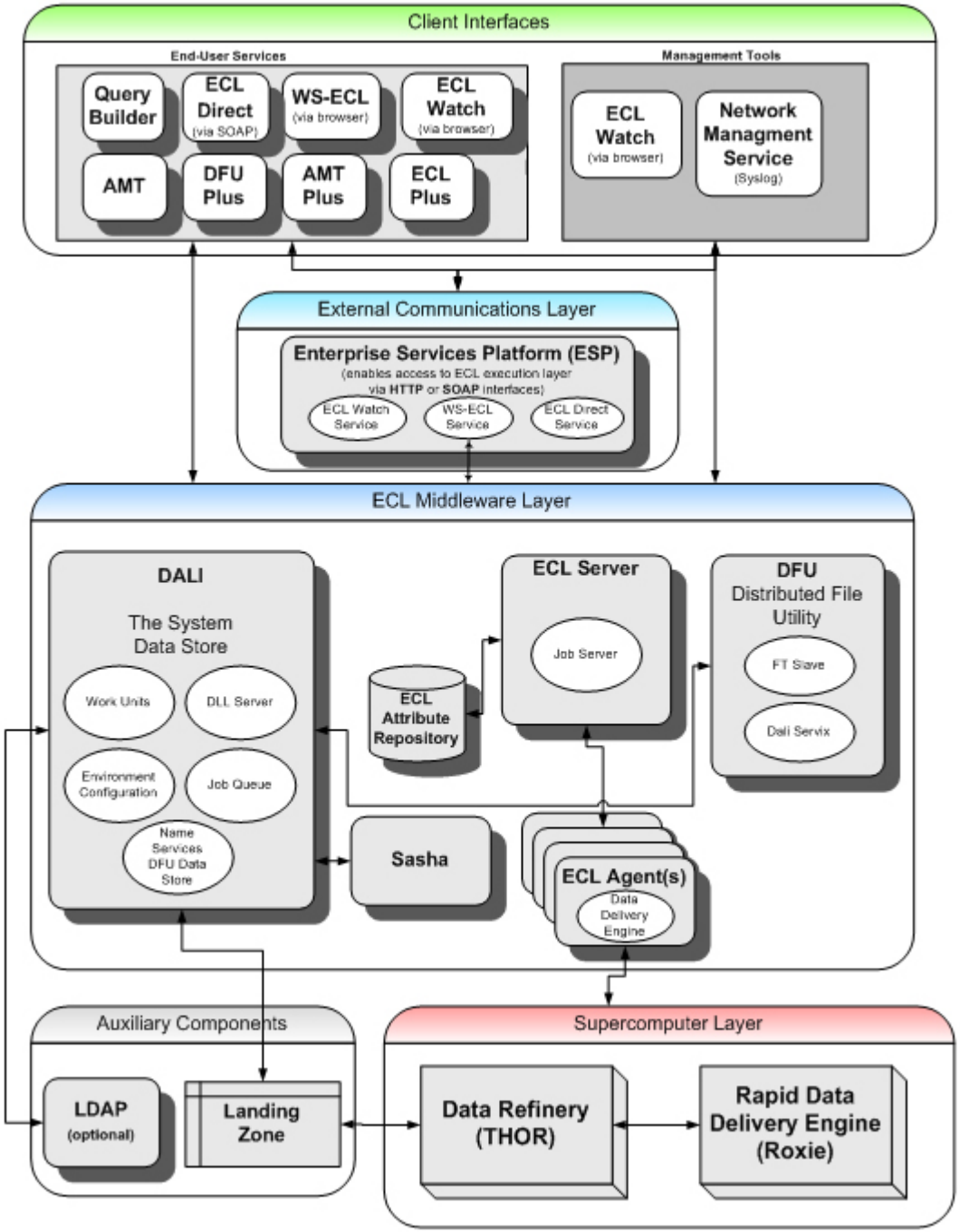
But, there is more to it than this.

Big Data technologies are complex today



Teradata Unified Data Architecture



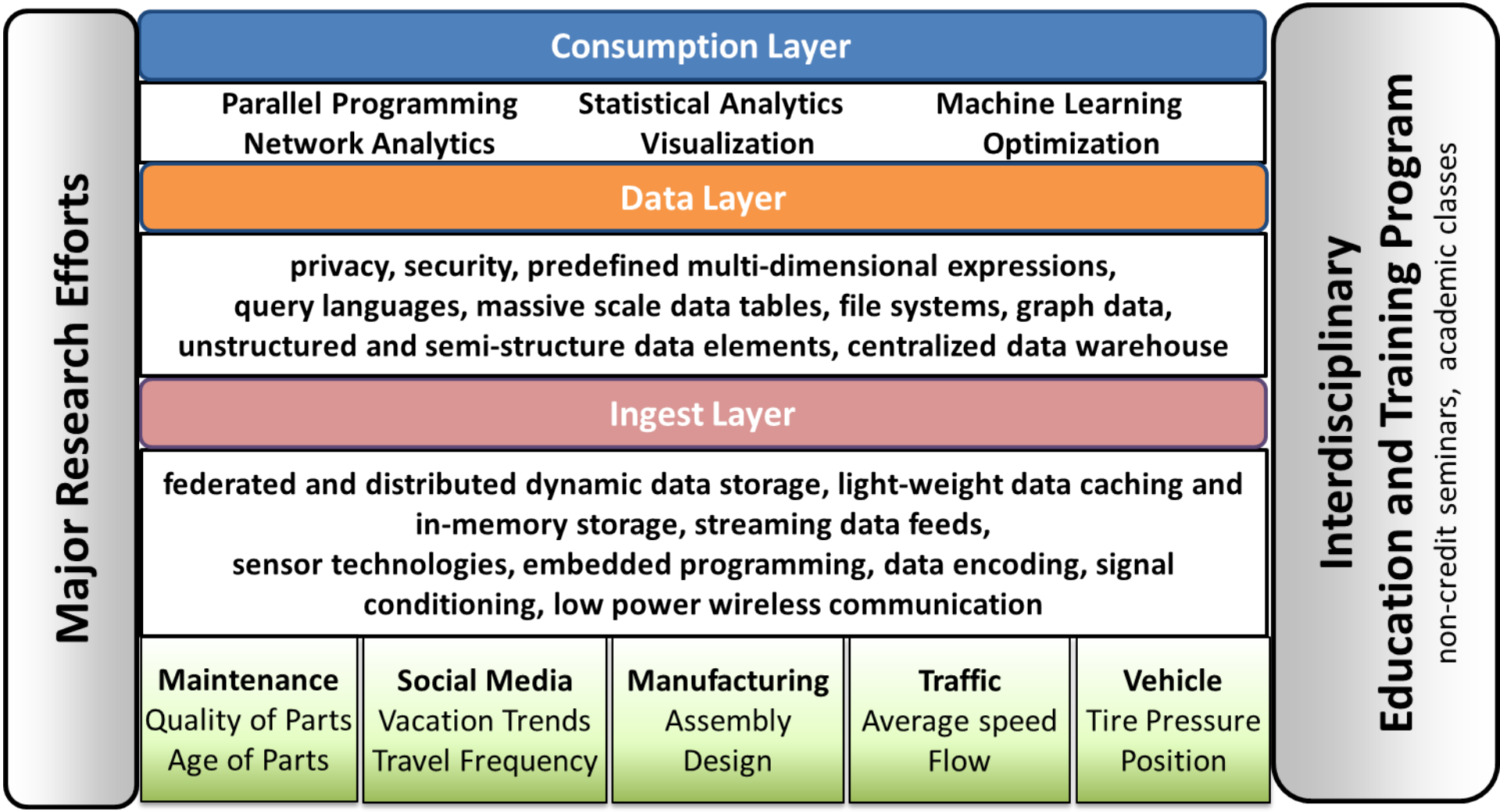


HPCC Systems

http://hpccsystems.com/

LexisNexis®

Data-Enabled Engineering Reference Architecture for Research and Education



Fourth Takeaway:
Work is needed to educate faculty and students in the use of tools

And, in the development of curricula and courses in these areas

CLEMSON UNIVERSITY

Summary

There are a lot of data out there

Solutions to many problems important to society can be advanced with Big Data

University researcher's use of Big Computing is correlated with higher research productivity

Work is needed in the development of curricula and course materials for Big Data

Amy Apon

aapon@clemson.edu

THANK YOU