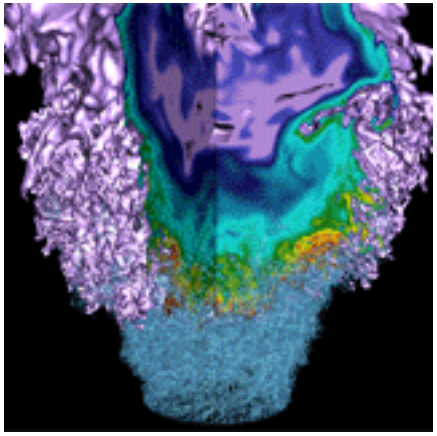
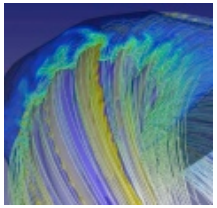
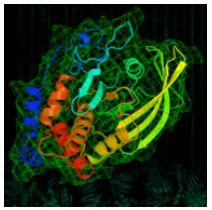
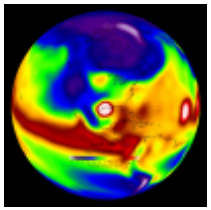
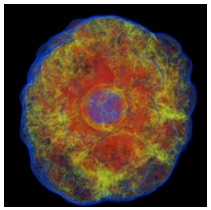
Data and Analytics Services

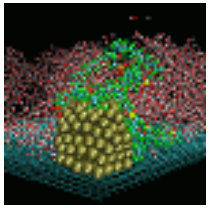












Prabhat

Data and Analytics Services Group Lead

March 21, 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Talk Overview

- **DAS Team and Goals**
- **Big Data Hardware**
- **Big Data Software**
- **User Engagement**

DAS Goal: "Enable Data-Intensive Science at Scale"

- **Internal Goals**
  - Provide world-class, production quality software services for all major Data capabilities:
    - Analytics, Management, Workflows, Transfer, Access, Visualization
  - Pioneer evaluation, research and deployment of Big Data technologies
    - Focusing on productivity and performance
  - Engage with stakeholders to enable scientific discovery in a data-driven world
    - Users
    - Computing Sciences Staff
    - Vendors
    - Researchers

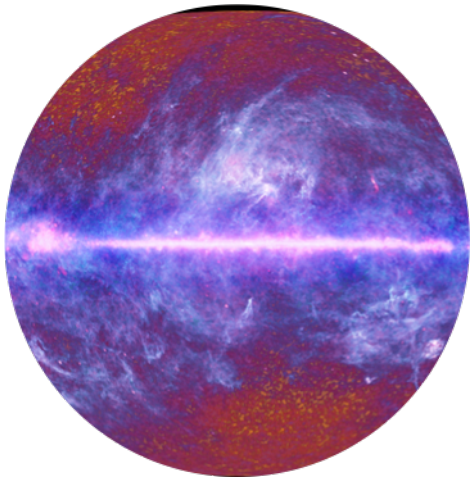
DAS Team

DAS member	DAS capability area	Science expertise
Debbie Bard	Analytics, Workflows	Cosmology, Astronomy
Wahid Bhimji	Management, Transfer	High Energy Physics
Shane Canon	Workflows	KBase, JGI
Shreyas Cholia	Access, Management	
Lisa Gerhardt	Management, Workflows	Particle Physics
Annette Greiner	Access	Systems Biology
Quincey Koziol	Management	
Jialin Liu	Management	
Jeff Porter	Transfer	Nuclear Physics
Prabhat	Analytics, Management	Climate
Evan Racah	Analytics	
Rollin Thomas	Analytics, Access	Astrophysics, Cosmology

Talk Overview

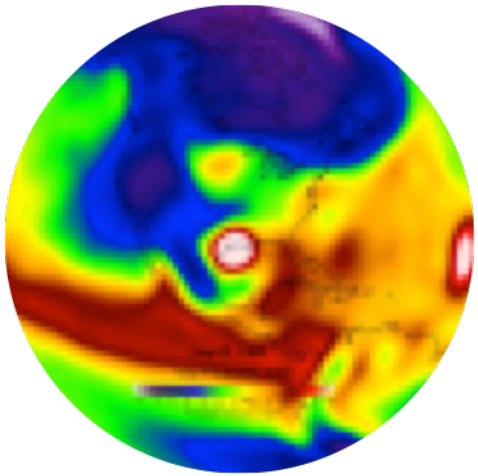
- DAS Team and Goals
- **Big Data Hardware**
- Big Data Software
- User Engagement

DOE Facilities are Facing a Data Deluge

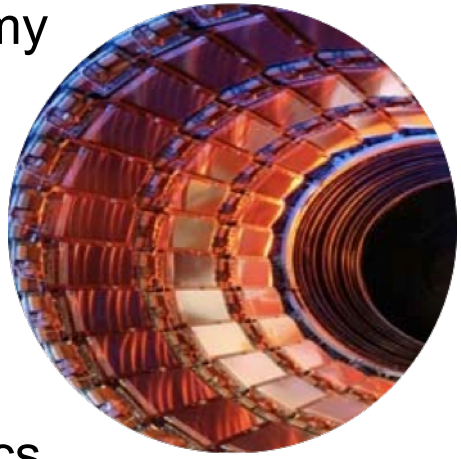


Astronomy

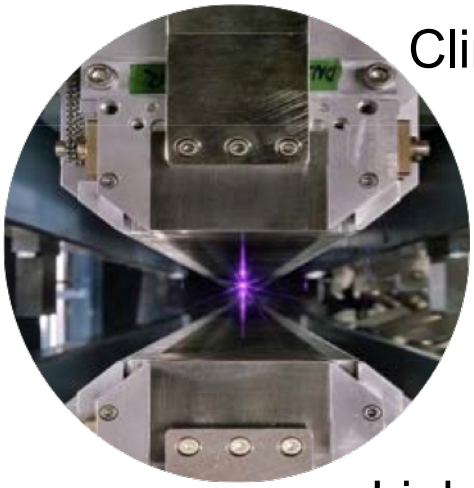
Genomics



Climate

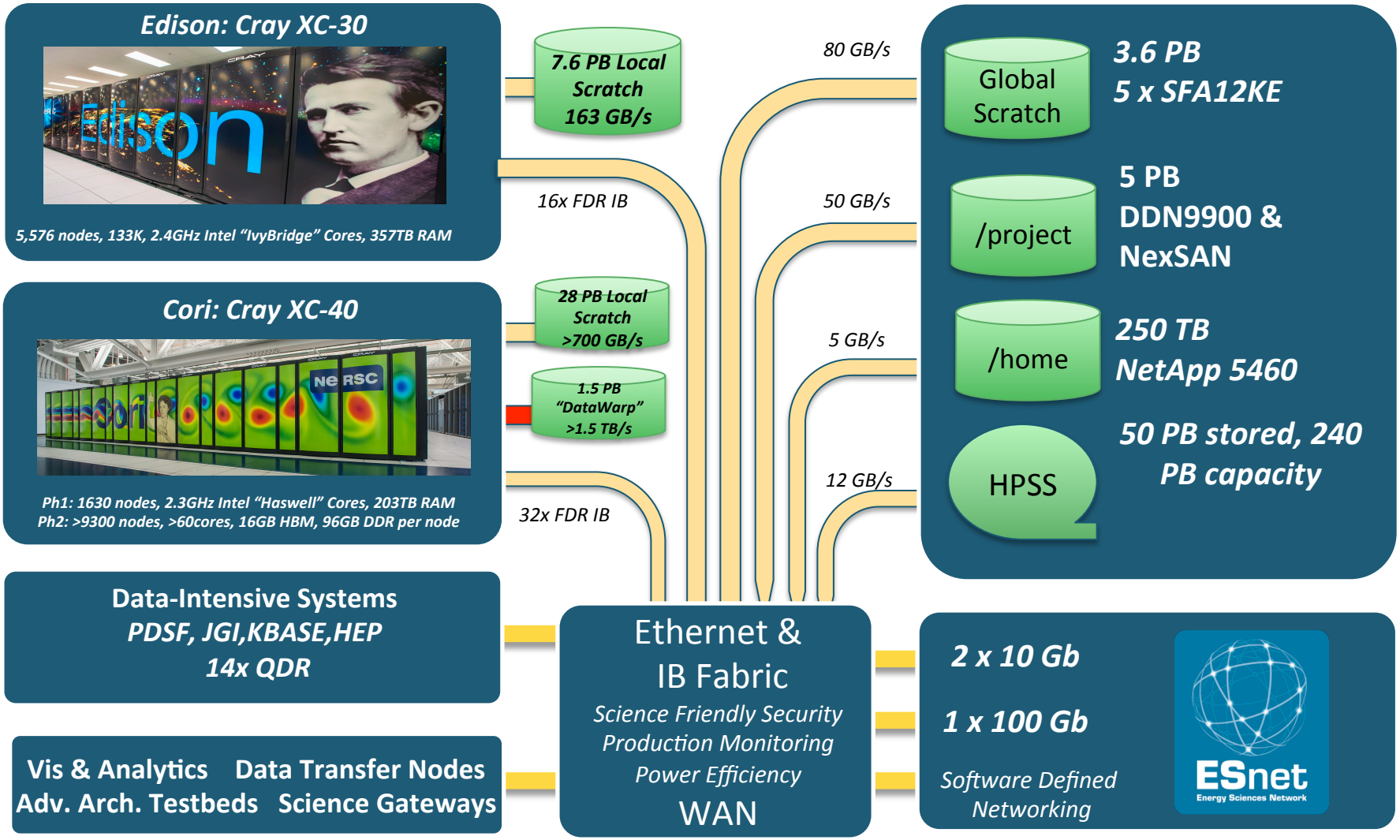


Physics



Light Sources

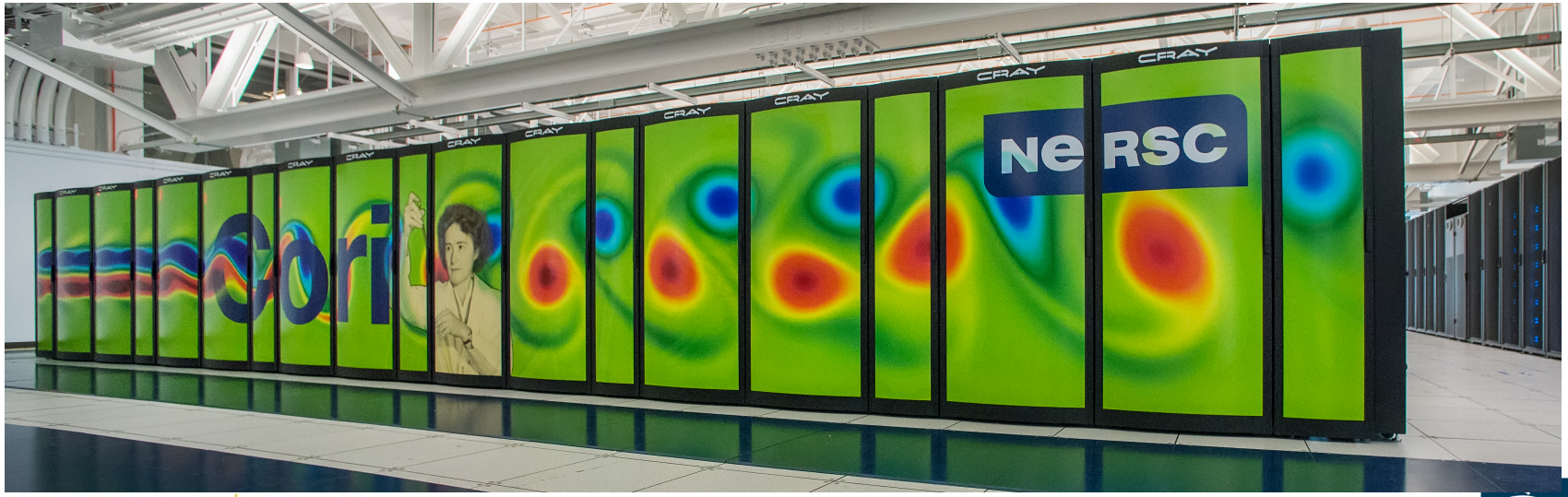
NERSC - 2016


Edison: Cray XC-30
Edison
5,576 nodes, 133K, 2.4GHz Intel "IvyBridge" Cores, 357TB RAM
Cori: Cray XC-40
Cori
NERSC
Ph1: 1630 nodes, 2.3GHz Intel "Haswell" Cores, 203TB RAM
Ph2: >9300 nodes, >60cores, 16GB HBM, 96GB DDR per node
7.6 PB Local Scratch 163 GB/s
28 PB Local Scratch >700 GB/s
1.5 PB "DataWarp" >1.5 TB/s
16x FDR IB
32x FDR IB
80 GB/s
50 GB/s
5 GB/s
12 GB/s
Global Scratch
3.6 PB
5 x SFA12KE
/project
5 PB
DDN9900 & NexSAN
/home
250 TB
NetApp 5460
HPSS
50 PB stored, 240 PB capacity
Data-Intensive Systems
PDSF, JGI,KBASE,HEP
14x QDR
Vis & Analytics
Data Transfer Nodes
Adv. Arch. Testbeds
Science Gateways
Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN
2 x 10 Gb
1 x 100 Gb
Software Defined Networking
ESnet
Energy Sciences Network

The Cori System

- **Cori will transition HPC and Data-intensive workloads to energy efficient architectures**

System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.



Data-friendly features on Cori

Data Intensive Workload Need	Cori Solution
Fast I/O (local disk)	NVRAM 'burst buffer' Lustre filesystem
Large memory nodes	128 GB/node on Haswell High memory 775 GB/ login node
Massive serial jobs	NERSC serial queue prototype on Edison; MAMU
Complex workflows	14 external login nodes Spark/JupyterHub/workflow software
Communicate with databases from compute nodes	Improved RSIP handles most cases; actively exploring SDN
Stream Data from observational facilities	Software Defined Networking
Easy to customize environment	User-defined images w/ Shifter
Policy Flexibility	Real-time; High throughput; Serial queues Reservations

Talk Overview

- DAS Team and Goals
- Big Data Hardware
- **Big Data Software**
- User Engagement

Big Data Landscape 2016

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, CAZENA, TREASURE DATA, splice MACHINE, bluedata, jethro, altiscale, Qubole, xplenty

Hadoop in the Cloud
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, memsql, nuoDB, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, GIRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, kognitio, dremio

Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
alteryx, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk, DATADOG, rocana, Anodot

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, Qumulo

App Dev
apigee, CASK, Keen IO, Typesafe, CONCURRENT

Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Dataameer, interana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, MODE, plotly, ADATAO, dataiku, nutonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Roambi, ZOOMDATA, Qlik, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, Wave Analytics, GoodData, birst, platfora, looker, atscale, ARCADIA, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
NETBASE, DATASIFT, tracx, bitly, synthesio, bottlenose, simplereach

Real-Time
amazon web services, METAMARKETS, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, amazon web services, H2O.ai, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfi.sh

Speech & NLP
NarrativeScience, api.ai, NUANCE, Gridspace, semantic machines, cortical.io, MindMeld, IDIBON, yseop

Horizontal AI
IBM Watson, Cortana, sentient, VIV, vicarious, nervana, Numenta, nara, HyperScience, MetaMind, clarifai, DEXTRO, Geometric Intelligence

Search
hp, Autonomy, ORACLE, ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
uc, Mu Sigma, OPERA SOLUTIONS, DATASCIENCE, DATA MOTH, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

SMB / Commerce
Google Analytics, AMPLITUDE, RJMetrics, BLUECORE, sumall, granify, Airtable, retention science, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, livefyre, blueyonder, kahuna, Lattice, SAILTHRU, persado, infer, 6sense, AVISO, ACTIONIQ, QUANTIFIND, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, STELLAService, NGDATA, Preact, Wise.io, DigitalGenius, appuri, fuse|machines

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization
MediaMath, Integral Ad Science, rocketfuel, OpenX, theTradeDesk, Adgorithms, LiveIntent, dstillery, DataXu, Appier, TAPAD

Security
CYLANCE, CounterTack, cybereason, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Guardian Analytics, Recorded Future, FORTSCALE, siftscience, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
facebook, X., Clara, KASISTO, lumiata

Publisher Tools
outbrain, mixpanel, Chartbeat, yieldbot, Yieldmo

Govt/ Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tidemark, Feeff, INSIKT, zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient

Education/ Learning
KNEWTON, Clever, declara, PANORAMA, knowre

Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon web services, Google, Microsoft, IBM, SAP, sas, hp, Autonomy, vmware, talend, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Apache Kylin, Flink, CDAP

Query / Data Flow
SLAMDATA, HIVE, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, APACHE HBASE, ACCUMULO, mongoDB, CouchDB, riak, SciDB, OPENTSDB, kafka, nifi

Coordination
talend, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, Scala, NumPy, SciPy

Machine Learning
mllib, mahout, Aerosolve, Apache SINGA, MADlib, Caffe, torch, TensorFlow, CNTK, FeatureFu, VELES, WEKA, DIMSUM, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

Financial & Economic Data
Bloomberg, DOW JONES, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location/People/Entities
GARMIN, foursquare, InsideView, esri, STREETLINE, CARTODB, factual, PlaceIQ, Crimson Hexagon, placemeter, BASIS TECHNOLOGY, Sense 360

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools

GA, DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

Big Data Software Portfolio

Capabilities	Technology Areas	Tools, Libraries
Data Transfer + Access	Globus, Grid Stack, Authentication	Globus Online, Grid FTP
	Portals, Gateways, RESTful APIs	Drupal/PHP, Django/Python, NEWT
Data Processing	Workflows	Swift, Fireworks
Data Management	Formats, Models	HDF5, NetCDF, ROOT
	Databases	MongoDB, SciDB, PostgreSQL, MySQL
Data Analytics	Statistics, Machine Learning	python, R, ROOT, Spark iPython/Jupyter
	Imaging	OMERO, Fiji, MATLAB
Data Visualization	SciVis InfoVis	VisIt, Paraview

Talk Overview

- DAS Team and Goals
- Big Data Hardware
- Big Data Software
- **User Engagement**

Engaging NERSC Users

- **Broad user support: O(100)-O(1000)**
  - Documentation:
    - http://www.nersc.gov/users/data-analytics/
  - Startup/troubleshooting questions:
    - Trouble ticket system (consult@nersc.gov)
- **Strategic Projects: O(10)**
  - 1-1 in-depth support
    - Architecting and implementing Data Analytics, Management, Access, Workflow and Visualization solutions
    - E-mail: Prabhat, prabhat@lbl.gov
  - Data Science Engagement Team
    - JGI, Light Sources, PDSF
    - E-mail: Kjiersten Fagnan, kmfagnan@lbl.gov

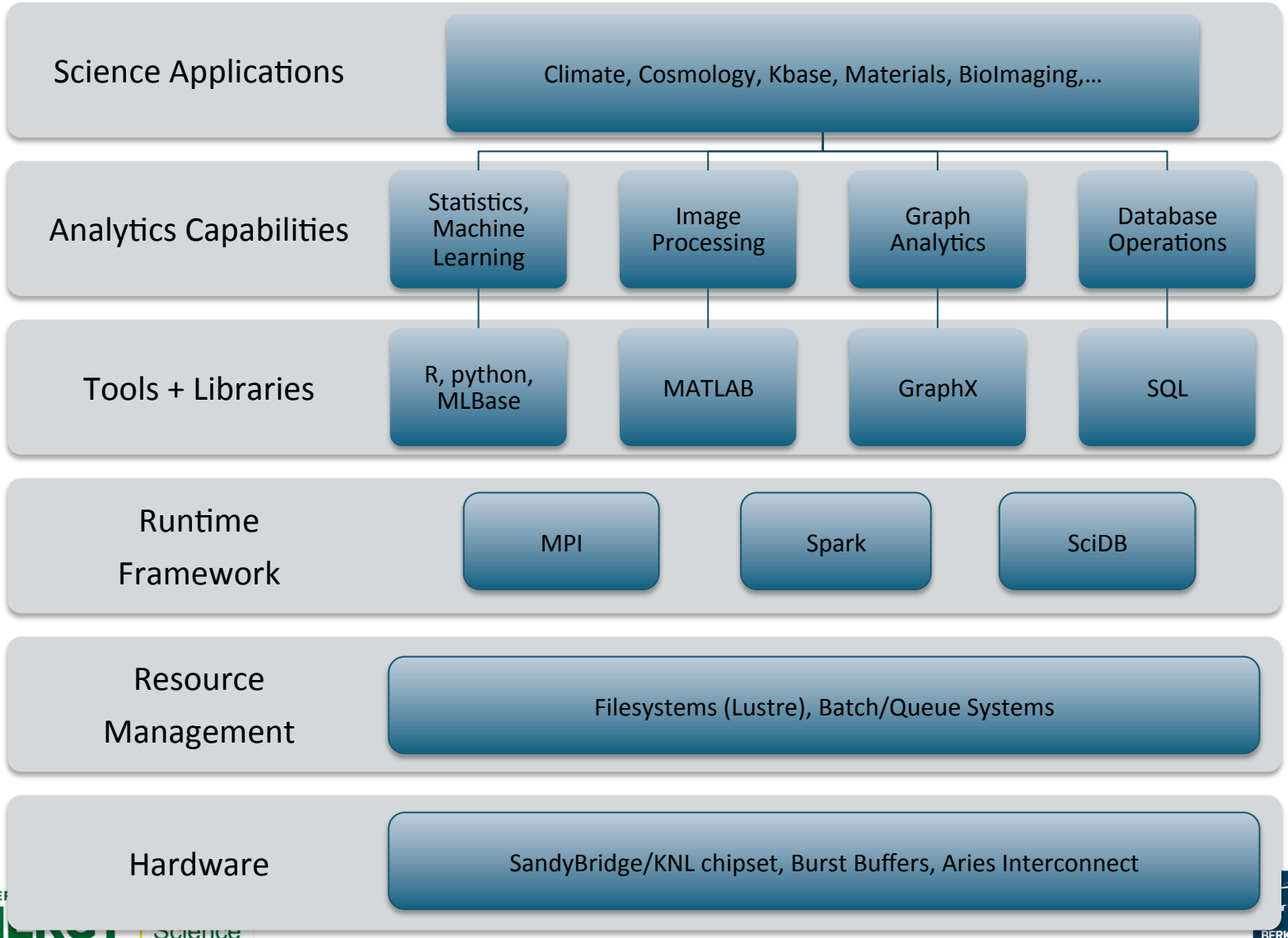
We need your input!

- **NERSC is leading the discussion on Data**
  - Cori platform is stable
  - Software
    - Productivity
      - Easy-to-use solutions across all major capability areas
    - Performance (HPC and HTC)
      - Processing 1TB datasets on 1,000 cores should be possible *now*
      - 2016 goal: enable processing of 10TB datasets on 10,000 cores
  - We need your leading science use cases
    - World-class team is looking forward to working with you

Questions? Comments?

Contact: prabhat@lbl.gov

Analytics Software Strategy



	Astronomy	Cosmology	Climate	Systems Biology	Neuroscience	BioImaging	Mass-spec Imaging	Personalized Toxicology	Materials	Particle Physics
Classification	✗		✗		✗	✗	✗			✗
Regression								✗	✗	
Clustering		✗	✗		✗		✗			✗
Dimensionality Reduction			✗		✗		✗			
Inference	✗						✗			✗
Model Estimation	✗				✗			✗		
Design of Experiments		✗	✗						✗	
Semantic Analysis			✗	✗					✗	
Feature Learning			✗		✗		✗	✗	✗	✗
Anomaly Detection	✗		✗							✗