Introduction to Virtual Data Intensive Summer School

July 8-10, 2013
http://www.vscse.org/summerschool/2013/bigdata.html

Robert Sinkovits
Scientific Applications Lead
San Diego Supercomputer Center

Overview of topics

Monday

- The nuts and bolts of data – moving and managing data, Hadoop and a few random topics that don't fit in well anywhere else

Tuesday

- Introduction to R, statistical analysis of data, predictive analytics

Wednesday

- Visualization of data using R, text analytics

This is not a class on R and R is not the only language or tool for statistical analysis and visualization. But, it is free, widely used in communities ranging from the social sciences to genomics and has a friendly syntax that avoids bogging you down in the low-level details (contrast, for example, with C)

Food analogy #1 – this school will be like eating tapas

Three six hour days is not enough time to turn you into an expert in data intensive computing. Think of this as an introduction or something to hold you over until the next big meal.

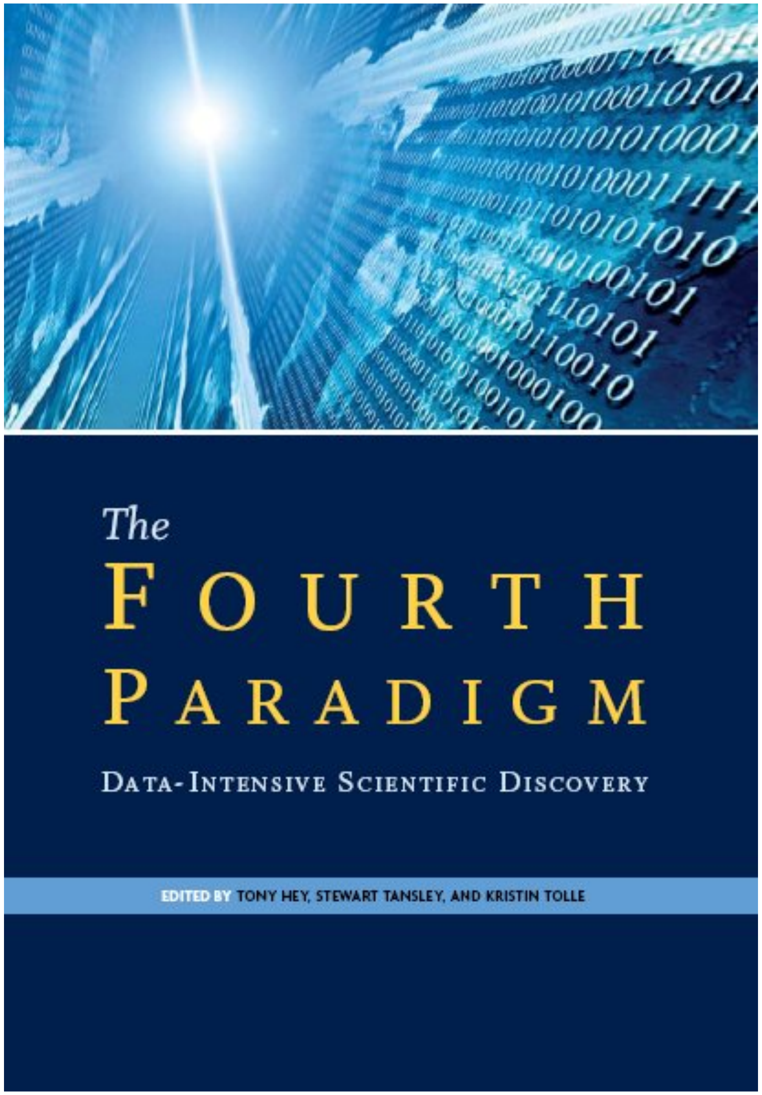
Food analogy #2 – compute is meat, data is rice

1970s American meal was based on a big portion of meat. Rice was a side dish and usually came out of an 8 ounce box.

Emphasis was on computing. Storage came for free with your HPC allocation

2000s Asian/Indian meal often has rice as the canvas or main ingredient. Probably came out of 20 pound sack.

Data management, processing and I/O dominates. Storage explicitly allocated just like compute time.

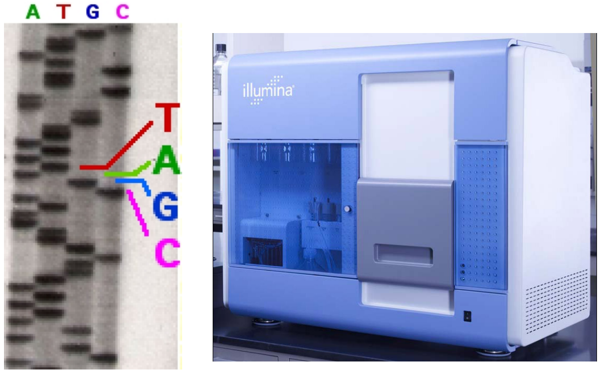

The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

So what exactly is data intensive computing?

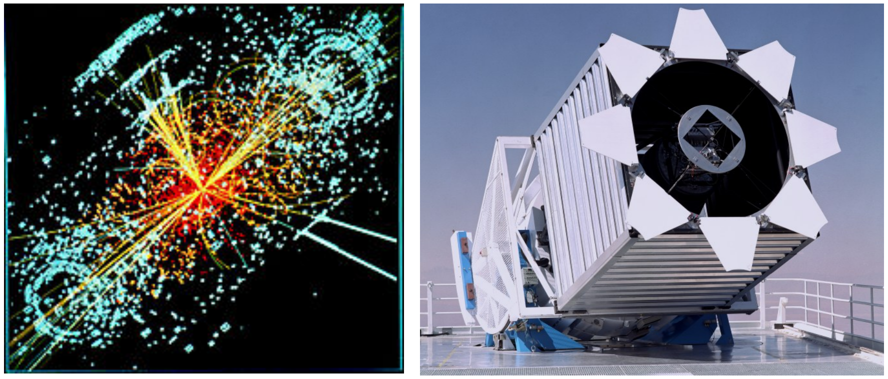
***The Fourth Paradigm** does a great job of introducing the topic, but still never manages to quite provide a definition*

I would define it broadly as any problem that is limited by the memory or storage subsytems rather than the compute power

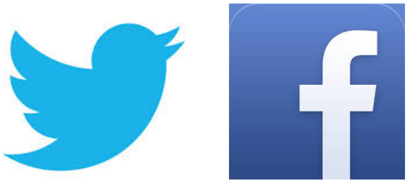
Why did data start to dominate?



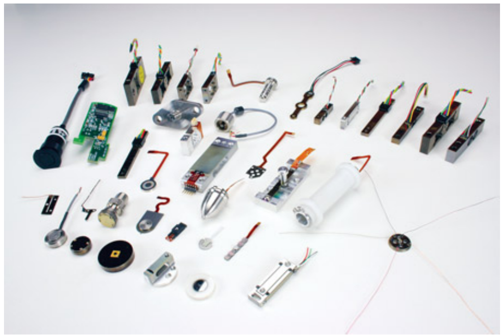
Changes in sequencing technology



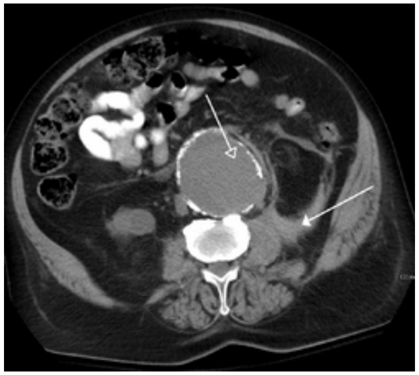
Scientific instruments (LHC, astronomy, etc.)



Social media

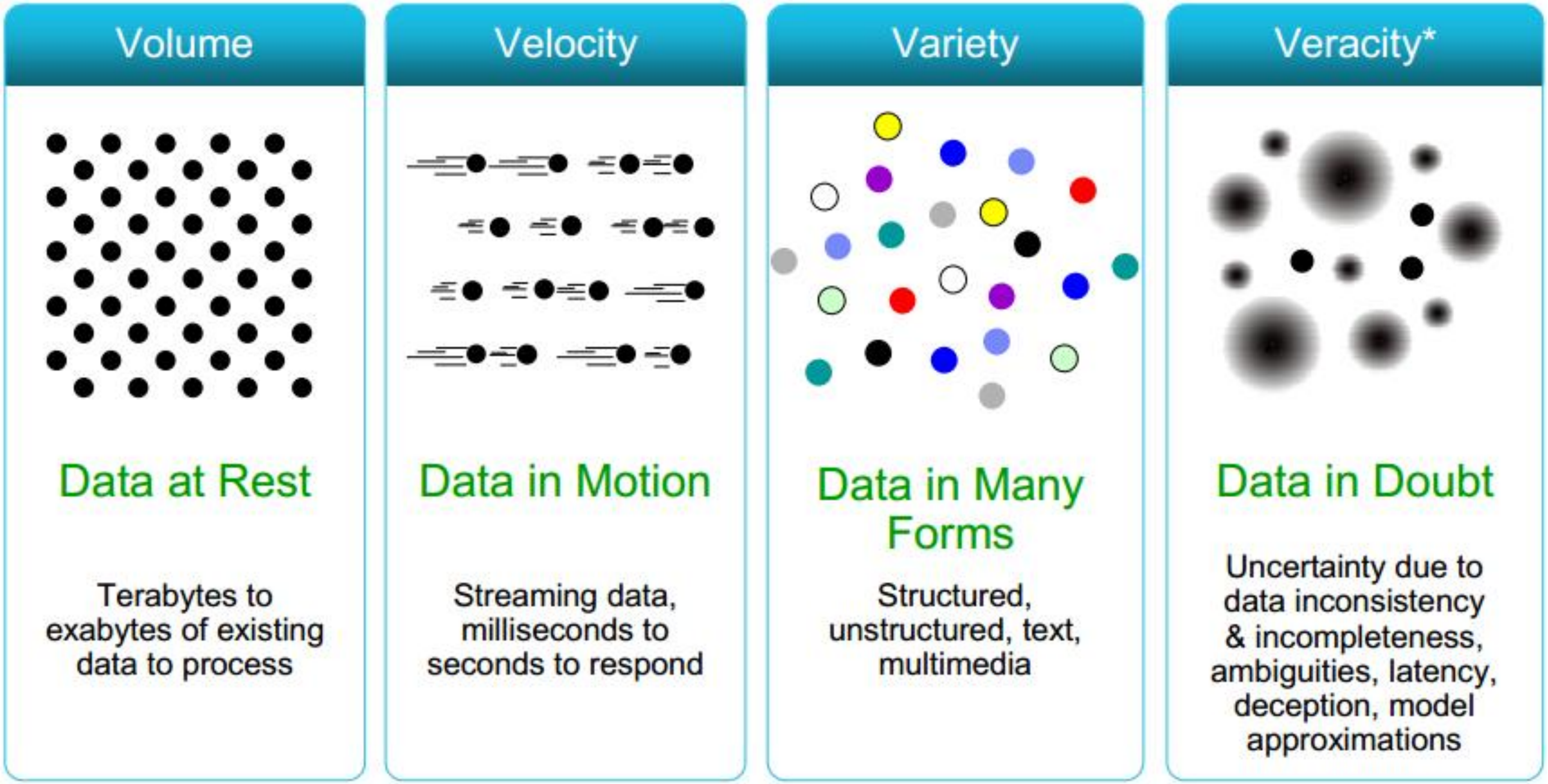


Sensor networks



Medical imaging

The 4 V's – Big data is about more than just volume



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

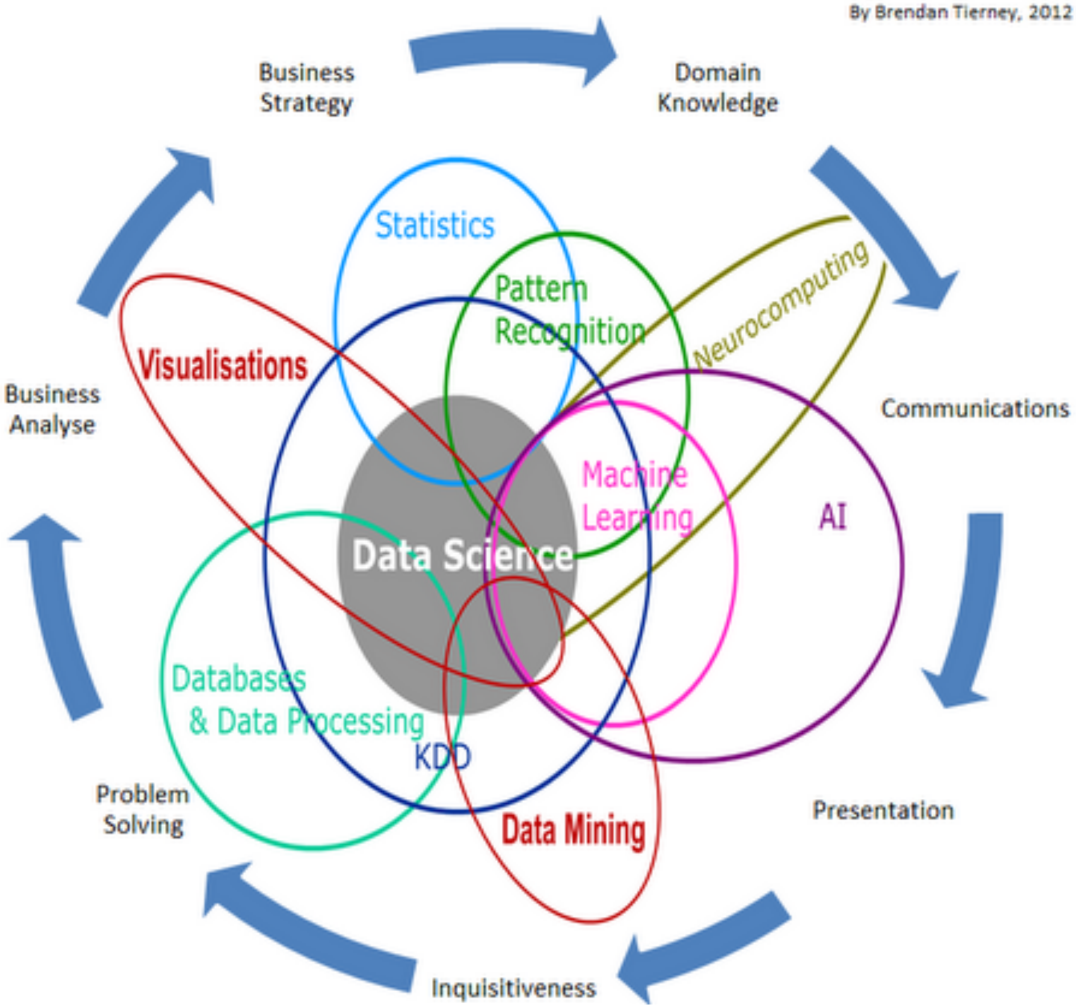
Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

IBM, 2012

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Installing software on your laptops

My original intention was to have participants run exercises on the XSEDE systems, but the interest was just too great for this to be practical (400+ registrants plus many others at streaming sites). Instead we'll be doing exercises on our laptops

Monday (afternoon)

- Google Chrome browser: https://www.google.com/intl/en/chrome/browser/

Tuesday (morning)

- R, version 3.0.1 (Good Sport) or later http://www.r-project.org/

Wednesday

- KNIME: http://www.knime.org/
- R packages: ggplot2 (version 0.9.3), scales, ggmap, googleVis, and igraph
- Download uber cars data set from http://www.infochimps.com/datasets/uber-anonymized-gps-logs

XSEDE resources – quick start guide

Much of what you learn here can be scaled up. We strongly encourage those of you with larger compute/storage needs to apply for XSEDE resources. All users will need an XSEDE portal account, only PI needs to apply for resources.

Everyone

- Go to https://www.xsede.org and create portal account (follow SIGN IN link at upper right hand corner of page)

PI

- From XSEDE home page, navigate to Allocations > Submit/Review Request https://www.xsede.org/group/xup/submit-request
- Click on “Click to Enter or View Request” to go to POPS page
- Follow link near bottom of page to “Startup”
- Fill in forms (fairly self-explanatory), note you can “Save to Date” at any time
- On Abstract/FOS page enter a short abstract (1-2 paragraphs) that describes your research and why you need supercomputing/storage resources

XSEDE resources – what's available

High performance computing resources ranging from 100-6000 TFlop peak

High throughput computing

Open Science Grid

Visualization

Future Grid

Distributed test beds

4-11 PB parallel file systems, XSEDE Wide File System, 60-170 PB tape archives

Steve Tuecke – Globus Online for Research Data Management

Abstract: The goal of the tutorial is to introduce researchers and systems administrators to the easy-to-use Globus Online services for moving and sharing large amounts of data. Increasingly computational- and data-intensive science make data movement and sharing across organizations inevitable. The cloud-hosted Globus Online service makes such research tasks as easy as Netflix makes streaming movies. In this tutorial, attendees will learn: how to perform fire-and-forget file transfer and synchronization between their local machine, campus clusters, regional supercomputers and national cyberinfrastructure using Globus Transfer via both Web and command line interfaces; how to store, version and share data using the new Globus Storage cloud storage service; how to integrate their security and identity infrastructure so that users may access Globus services seamlessly with their organization’s standard account (e.g., campus login via InCommon) or existing accounts with OAuth and OpenID identity providers (e.g., XSEDE, Google); and how to connect their own systems to the Globus Online services.

Steve Tuecke – Globus Online for Research Data Management

Steven Tuecke is Deputy Director at The University of Chicago's Computation Institute (CI), where he is responsible for leading and contributing to projects in computational science, high-performance and distributed computing, and biomedical informatics. His specific focus is on co-leading the Globus project, with Dr. Ian Foster.

Prior to CI, Steven was co-founder, Chief Technology Officer, and on the board of Univa Corporation from 2004-2008, and also served as Univa's first Chief Executive Officer. Univa provided open source and proprietary software for the high-performance computing and cloud computing markets. Steven helped lead Univa through several new product launches, multiple venture capital investment rounds, and the acquisition of United Devices. He continues to serve on Univa's board and as CTO advisor. Prior to Univa, Steven co-founded the Globus Project, with Dr. Ian Foster and Dr. Carl Kesselman. He was responsible for managing the architecture, design, and development of Globus software, as well as the Grid and Web Services standards that underlie it.

He began his career in 1990 as a software engineer for Foster in the Mathematics and Computer Science division at Argonne National Laboratory. In 1995, Tuecke helped create the Distributed Systems Laboratory at Argonne which, under his management and technology leadership, became the premier Grid research and development group in the world. In 2001, Tuecke focused on Globus architecture and design, creating Grid and Web Services standards, and expanding corporate relationships. Tuecke graduated summa cum laude with a B.A in mathematics and computer science from St. Olaf College.

Yoav Freund
Map-reduce, Hadoop and The communication bottleneck

Abstract: "Big Data" is one of the trendiest terms of the last few years. "Hadoop" is one of the most popular approaches for dealing with big data. Both terms represent a rapidly increasing trend in computation. One important shift that corresponds to this trend is that the bottleneck in big data analysis is communication speed, rather than computation speed. In this talk I will explain the roots causes of this shift. I will then describe Map-Reduce and the distributed file system, which are the basis of Hadoop.

Yoav Freund is a professor of Computer Science and Engineering at UC San Diego. His work is in the area of machine learning, computational statistics and their applications. Dr. Freund is an internationally known researcher in the field of machine learning, a field which bridges computer science and statistics. He is best known for his joint work with Dr. Robert Schapire on the Adaboost algorithm. For this work they were awarded the 2003 Gödel prize in Theoretical Computer Science, as well as the Kanellakis Prize in 2004.

Richard Marciano and Jeff Heard
Managing and interacting with big data

Abstract Session 1: CI-BER is a research collaboration that brings together experts in computer science, engineering, and archival science. It extends earlier work done under TPAP (Transcontinental Persistent Archives Prototype) and its goal is to further the understanding of infrastructure that scales and provide insights into the management of scientific data in general. We will look at very high scale collections and visual analytics techniques, in order to enhance the value of government records that can lead to generalizable infrastructure and technology. This 1 hour session, will introduce the project and the underlying technologies and techniques and allow virtual attendees to interact with one of the early interfaces.

Abstract Session 2: The Big Board is a collaborative environment on top of RENCI's Geoanalytics framework that allows researchers to synthesize diverse data in real-time collaborative. It is a web-based environment that lets researchers overlay data from diverse sources, integrate domain specific applications, and annotate, link, and attach files to features on a geographic map. The underlying technology, Geoanalytics enables the real-time work and provides the ability to overlay sources from larger data streams and repositories not typically fully accessible to GIS. In this one hour session we will discuss the applications of collaborative technology to big geographic data and allow students to interact with the Big Board interface over a sample problem domain.

Richard Marciano and Jeff Heard
Managing and interacting with big data

Richard Marciano is a professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill (UNC). He serves as chair of the CDHI Cyberinfrastructure Task Force, a $5M Mellon Foundation / UNC award in the digital humanities and "big data", director of Sustainable Archives and Leveraging Technologies (SALT), principle investigator for the Cyberinfrastructure for Billions of Electronic Records (CI-BER) project and a number of other "big data" collaboratives, and is a 2012 recipient of the JISC Digging Into Data Challenge Grant. Richard holds a BS in Avionics and Electrical Engineering, and an M.S. and Ph.D. in Computer Science, and has worked as a Postdoc in a Computational Geography. He conducted interdisciplinary research at the San Diego Supercomputer at UC San Diego for over a decade, working with teams of scholars in sciences, social sciences, and humanities.

Jeff Heard is a Senior Visualization Researcher at the Renaissance Computing Institute (RENCI) at the University of North Carolina at Chapel Hill. His research interests include scalable interactive visualization systems, geographic visualization, visual analytics, and information visualization. Jeff is the architect and principle developer of RENCI's Geoanalytics framework, which has been used to support multiple funded projects in the environmental, health, and social sciences by data integration and collaboration.

Chris Fariss – Introduction and statistical analysis using R

Abstract: R is the most widely used software environment for statistical computing and graphics. It is used across across the sciences. This three hour workshop is designed to introduce users to the syntax of R and to help them get started analyzing data.

Christopher Fariss will join the Department of Political Science at Penn State University in Fall 2013. He recently graduated with a Ph.D. in political science from the University of California, San Diego (2013). His core research interest is in the politics of human rights, violence, and repression. He uses computational methods to understand why governments around the world choose to torture, maim, and kill individuals within their jurisdiction. Other projects cover a broad array of themes, ranging from foreign aid to American voting behavior, but share a focus on computationally intensive methods and research design. These methodological tools, essential for analyzing "big data", open up new insights into the micro-foundations of state repression.

Charles Elkan – The landscape of analytics: A personal view

Abstract: "Analytics" is a popular name for a research area that can also be called data science or data mining. This talk will provide an evaluation of current research and trends in this field. I will discuss applications and research questions involving structured and unstructured data, and big and fast data. Topics will include text mining, social network and social media analysis, search engines, reinforcement learning, and research and start-up opportunities.

Dr. Charles Elkan is a professor in the Department of Computer Science and Engineering at the University of California, San Diego. In the past he has been a visiting associate professor at Harvard and a researcher at MIT. Dr. Elkan is known for his research in machine learning and data mining. The MEME bioinformatics algorithm that he developed with his Ph.D. student Tim Bailey has been used in over 2000 published research projects in biology. Dr. Elkan has won several best paper awards and data mining contests, and some of his graduate students have become leaders at companies including Google, Yahoo, Zillow, and IBM, while others have held faculty positions at Columbia University, the University of Washington, and other universities inside and outside the U.S.

Amy Szczepanski – Introduction to Visualization with R

Abstract: The statistical package R is becoming more and more popular for data analysis and visualization. This software can be easily extended by add-on packages, and some of these packages provide extremely sophisticated options for visualization. In this session, we will cover an overview of visualization and an introduction to visualizing several different types of data in R. Most of the session will focus on the use of the ggplot2 package, and we will also cover use of Google Motion Charts, mapping geographic data, and visualizing graphs. This session assumes an awareness of basic R syntax, a beginner level of programming experience, and a basic familiarity with working with data.

Amy F. Szczepanski is a mathematician by training, having received her Ph.D. in mathematics from the University of California, San Diego. She is a Research Assistant Professor of Electrical Engineering and Computer Science at the University of Tennessee and is the coordinator of Education, Outreach, and Training for UT's Remote Data Analysis and Visualization Center at the National Institute for Computational Sciences. Her research interests include understanding how researchers use high-performance computing in their computational workflows and extracting knowledge from data sets generated by automated observations of human activity.

Dean Abbott – Text Mining

Abstract: With many experts claiming that unstructured data comprises more than 80% of the stored business information (primarily as text), text mining has emerged as a critical leading-edge technology. This course will provide an overview of practical techniques for text extraction and text mining in a data mining context, meaning using processed text as features for supervised and unsupervised learning. Techniques for text extraction will include basic concepts of tokenization, part-of-speech tagging, stemming, and feature creation will be covered. The open source source text mining software tool KNIME will be used for in-class demonstrations and work, but tie-ins to other tools such as RapidMiner, R, IBM Modeler, and Megaputer's Polyanalyst will be included.

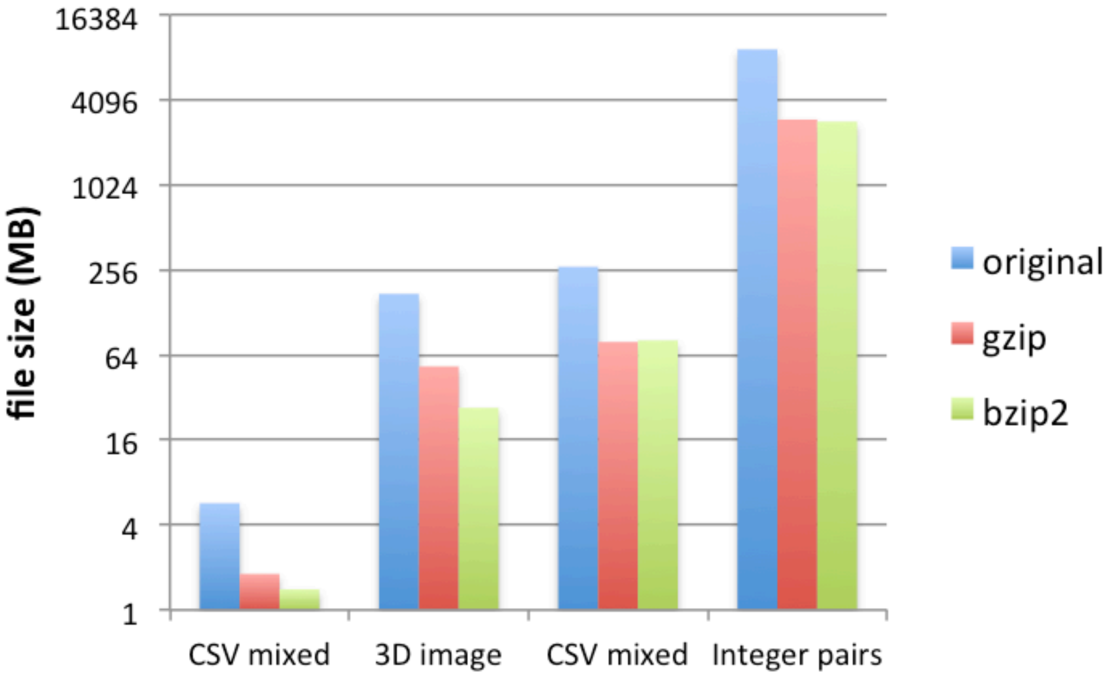
Dean Abbott is President of Abbott Analytics, Inc. in San Diego, California. Mr. Abbott has over 21 years of experience applying advanced data mining, data preparation, and data visualization methods in real-world data intensive problems, including fraud detection, response modeling, survey analysis, planned giving, predictive toxicology, signal process, and missile guidance. In addition, he has developed and evaluated algorithms for use in commercial data mining and pattern recognition products, including polynomial networks, neural networks, radial basis functions, and clustering algorithms, and has consulted with data mining software companies to provide critiques and assessments of their current features and future enhancements.

Storage is getting cheaper, but …

As disk becomes cheaper, it's easy to get cavalier regarding storage. Try to minimize data requirements, even if you think that there is a lot of excess capacity

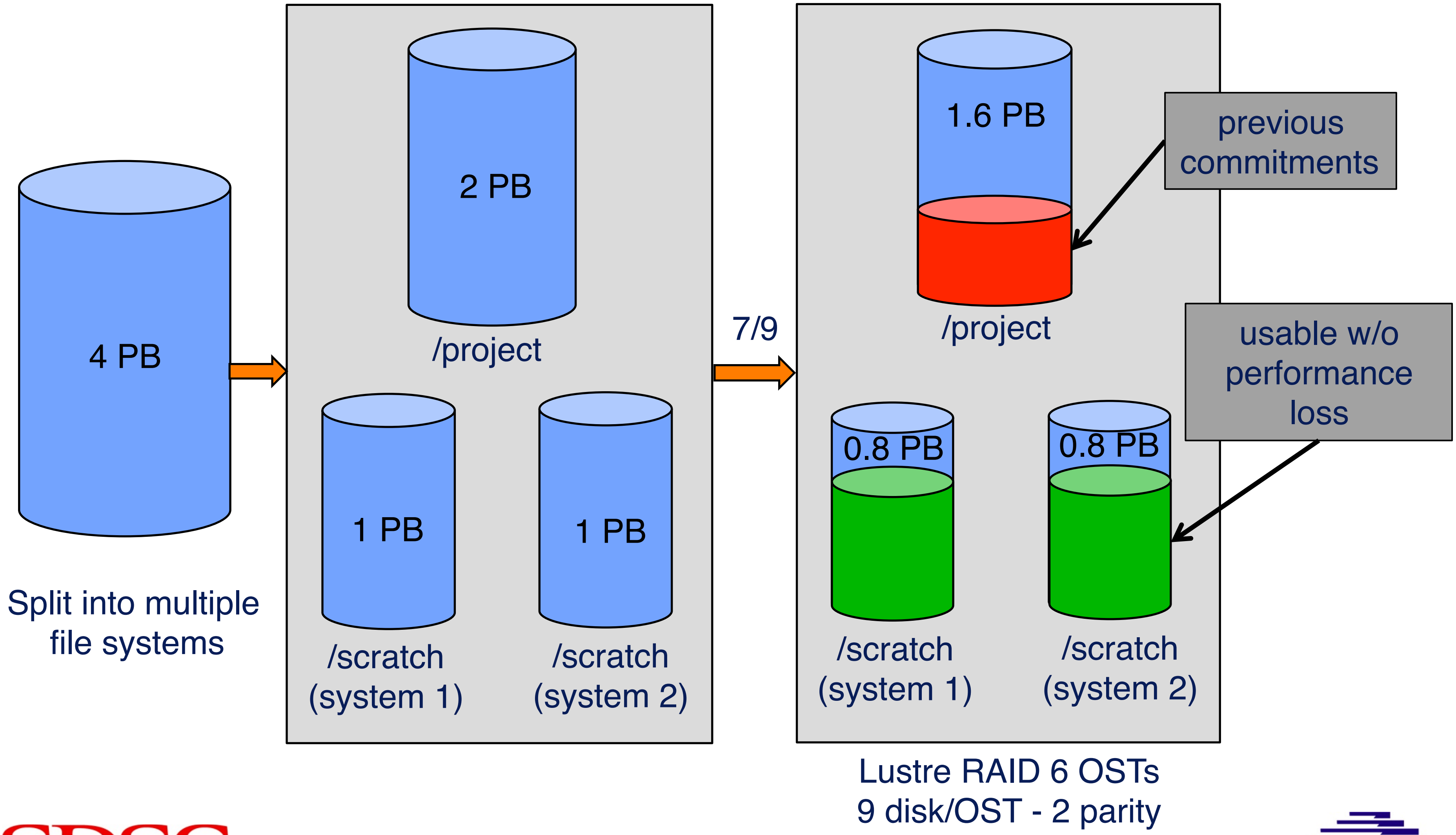
- Avoid unnecessary duplication of data sets
- Write binary or native formats rather than ASCII
- Compress using gzip or bzip2. Consider more aggressive compression if storage is more valuable than CPU time
- Create archives of (e.g. tar) large numbers of small files before applying compression (or storing to disk/tape or transferring over network)
- Consider what data really needs to be retained and what can easily be recreated on the fly

gzip and bzip2 performance



The bzip2 time is often 2-4x longer than gzip, but can often yield much better compression ratios. Also consider pbzip2 (parallel bzip2) when available

So your center just bought 4 PB of disk …



Even the most inexpensive, high latency storage options add up if you have enough data

Amazon Glacier storage pricing:

- $0.01/GB/month
- $120/TB/year
- $120,000/PB/year

Transfers from Glacier to Internet

- 10 TB/month: $0.12/GB ($1200)
- next 40 TB/month: $0.09/GB ($3600)
- next 100 TB/month: $0.07/GB ($7000)
- next 350 TB/month: $0.05/GB ($17,500)

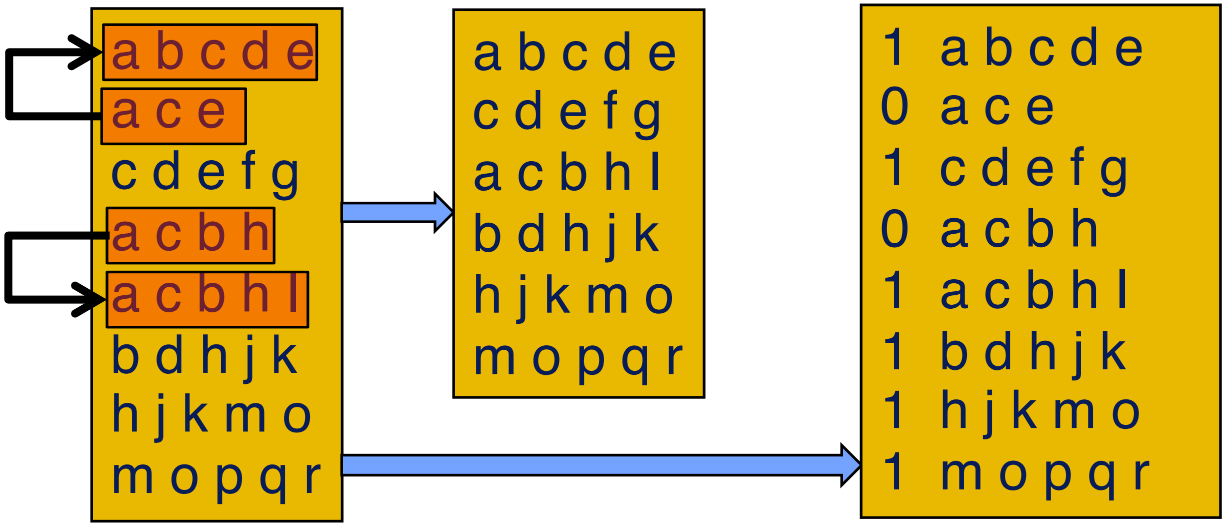
A few words about memory – double precision

- Can cut memory footprint in half by using float (4-byte) rather than double (8-byte)
- Benchmark your application for a variety of problems to determine how precision affects outcome

- Large linear algebra problems, iterative solvers, ODEs and PDEs, simulations where solution evolves over time and problems in general where a subsequent result depends on a previous result probably need double precision
- Many Monte Carlo problems where each configuration or trial is independent can probably use single precision

A few words about memory – data structures

Example from the LSST Moving Object Pipeline System (MOPS). Identifying tracks of near earth objects that are subsets of other tracks and eliminating from data set before further processing



a b c d e
a c e
c d e f g
a c b h
a c b h l
b d h j k
h j k m o
m o p q r

a b c d e
c d e f g
a c b h l
b d h j k
h j k m o
m o p q r

1 a b c d e
0 a c e
1 c d e f g
0 a c b h
1 a c b h l
1 b d h j k
1 h j k m o
1 m o p q r

Associating a flag (one bit or byte) with each track adds minimal memory overhead. Negligible if the size of the tracks is much larger than one byte.

Concluding remarks

- Enjoy the virtual summer school and feel free to ask questions
- Use this summer school as a jumping off point for deeper study
- Give us your feedback at the end of the summer school. This will help us to do an even better job next time
- Finish your software installations as soon as possible. At the very least
  - Google Chrome before this afternoon
  - R before tomorrow morning
  - KNIME before Wednesday morning
- Please pay your $100 registration fee if you haven’t done so already and your site has not provided a waiver (e.g. UCLA is not charging students, streaming sites may not be charging)

The 2013 VSCSE Data Intensive Summer School is made possible by the generous support of:

- The local organizing teams at each site
- The National Science Foundation under award number OCI-1041313
- The VSCSE organizing team, including Tricia Barker (NCSA), Karen Coulter, (U Michigan), Maytal Dahan (TACC), Prof. Sharon Glotzer (U Michigan), Dr. Steve Gordon (Ohio Supercomputing Center), Erik Hofer (U Michigan), Scott Lathrop (U Illinois), Meagan Lefebvre (U Michigan), Michael Miller (U Illinois), Sam Moore (TACC), Paul Ponder (NCSA), Andrew Schuh (U Illinois), Robert Sinkovits (SDSC), Dan Stanzione (TACC), Trevor Walker (SDSC)