2014

2014 Big Data Conference Program

Big Data Conference

APRIL 25,2014

2ND ANNUAL BIG DATA CONFERENCE

SPONSORED BY
THE PROVOST'S OFFICE
THE CENTER FOR SCHOLARLY
AND CREATIVE EXCELLENCE AND
THE PEW FACULTY TEACHING
AND LEARNING CENTER

GRAND VALLEY
STATE UNIVERSITY

2014

2ND ANNUAL BIG DATA CONFERENCE

CONFERENCE COMMITTEE

Ed Aboufadel
Maria Cimitile
Paul Leidig
Shaily Menon
Carlos Rodriguez
Jerry Scripps
Greg Wolffe

Glossary of Sizes

Kilo	1,000
Mega	1,000,000
Giga	1,000,000,000
Tera	1,000,000,000,000
Peta	1,000,000,000,000,000
Exa	1,000,000,000,000,000,000
Zetta	1,000,000,000,000,000,000,000

"It is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions." – Harvard Business Review, October 2012.

Thank you to the Provost's Office, the Center for Scholarly Creative Excellence and the Pew Faculty Teaching and Learning Center for their support for today's conference.

8:30 - 9:00 **Poster and Networking Session**

2201 KC

Aaron Clark, Statistics, Engineering
Visualizing Lake Michigan Wind

Tyson Spoelma, Statistics, Engineering
Staged Data Warehousing with Gigabytes of Longitudinal Data

Heather Carpenter, Public, NonProfit and Health, Administration
Connections and Overlap between Capacity Building Measures, Nonprofit Management Competencies and Training Needs of Nonprofit Managers

Logan Westrick, Computing and Information Systems
Building a Better Stockbroker: Constructing an Ontology-Based Financial Knowledge Base

Nicholas Vogel, Computing and Information Systems
Mining Enormous Mobile Datasets to Improve Mitigation Strategies for Limiting the Spread of Infectious Disease

9:00 - 9:45 **Keynote Address**

Pere Marquette

Barbara O'Brien, MSU College of Law
Finding Data: The Politics and Magic of Accessing Capital Punishment Data

Presentations

9:50 - 10:10 **Mary E. Winn**, Van Andel Institute, Pere Marquette
Big Data Challenges in Bioinformatics – The Collision of Technology and Biology

Laura Ring Kapitula, Statistics 2215/2216KC
Graphing Big Data Using the SAS© System

10:15 - 10:35	**Shaily Menon**, Biology Biodiversity Informatics – Big Data for Biodiversity Conservation and Ecological Forecasting	Pere Marquette
	Szymon Machajewski, CIS Computing and Information Systems Open Source Analytics – Blackboard Learn	2215/2216KC
10:40 - 11:00	**Andy Van Solkema**, Visual Thinking	Visualhero, Pere Marquette
	Mario Adkins, Education Tutorial: NVivo 10	2215/2216KC
11:05 - 11:25	**David Zeitler**, Statistics High Performance Statistical Computing	Pere Marquette
	Jerry Scripps, CIS Data Mining the Liberal Arts	2215/2216KC
11:30 - 11:50	**Paul Stephenson**, Statistics A Statistical Analysis on the Likelihood of Independent Occurrences of Serious Residential Fires	Pere Marquette
	Robert Deaner, Psychology The Tortoise and the Hare: Men are More Likely than Women to Slow in the Marathon	2215/2216KC
11:55-12:15	**Edward Aboufadel**, Mathematics Creating “Baseball Motion Graphs”	Pere Marquette
	Paul Leidig Computing and Information Systems New Curriculum Development in Data Science Programs	2215/2216KC

12:25 - 1:10 **Keynote Address and Lunch**

Pere Marquette

Jonathan White, Meijer Honors College
Big Data, Big Intelligence, Big Security: The Challenge of Protecting Rights

Birds of a Feather (breakout)

1:15 - 2:00	**Andy Van Solkema,** Visualhero Can We Measure Visual Literacy?	2201 KC
	Edward Aboufadel, Mathematics Big Data Where You Might Not Expect It (English, History, …)	2215/16

KEYNOTE 1

9:00 - 9:45
PERE MARQUETTE

Finding Data: The Politics and Magic of Accessing Capital Punishment Data

Barbara O'Brien

MSU COLLEGE OF LAW

This talk provides an overview of the primary data sources for death penalty research and a glimpse at some of the limitations of working with these sources. Examining each stage of a capital prosecution—from the police investigation of a potentially capital murder to a governor's decision to grant clemency sparing the life of a death-row inmate—is necessary to understand the process as a whole. But arbitrary factors like local politics, luck, and institutional legacies limit researchers' ability to collect data about each decision point, making researchers' attempts to follow cases across multiple stages exceptionally difficult.

KEYNOTE 2

12:25 - 1:10
PERE MARQUETTE

Big Data, Big Intelligence, Big Security: The Challenge of Protecting Rights

Jonathan R. White

MEIJER HONORS COLLEGE

This address will focus on the changing nature of conflict in the 21st century and the future role of intelligence gathering and analysis. Big Data offers numerous opportunities for the Intelligence Community, national defense, and law enforcement. At the same time it threatens individual liberties. The role of Big Data and the parameters under which it operates should be standardized through public discourse and legal limitations.

8:30- 9:00
PERE MARQUETTE

Aaron Clark

STATISTICS,
ENGINEERING

Visualizing Lake Michigan Wind

A wind resource assessment buoy, residing in Lake Michigan, uses a pulsing laser wind sensor to measure wind speed and direction offshore up to a wind turbine hub-height of 175m and across the blade span every second. Understanding wind behavior would be tedious and fatiguing with such large data sets. However, SAS/GRAPH® 9.4 helps the user grasp wind characteristics over time and at different altitudes by exploring the data visually. This paper covers graphical approaches to evaluate wind speed validity, seasonal wind speed variation, and storm systems to inform engineers about the energy potential of Lake Michigan offshore wind farms.

8:30- 9:00
PERE MARQUETTE

Logan Westrick

COMPUTING AND
INFORMATION SYSTEMS

Building a Better Stockbroker: Constructing an Ontology-Based Financial Knowledge Base

Decision support systems are a rapidly growing class of computer programs used to assist middle and upper management in making decisions and planning. There are essentially three components to any decision support system: the knowledge base, the model (that is, the part that processes the data to make an informed decision), and the user interface. This project focused on building a knowledge base for a financial decision support system. It did so by creating an ontology, a formal representation (and abstract model) of knowledge as concepts within a domain such that it can be understood by a computer. By adding this layer of abstraction around Big Data, an ontology relieves users of the burden of managing enormous data complexity, freeing them to concentrate on the problem at hand.

8:30- 9:00
PERE MARQUETTE

Tyson Spoelma

STATISTICS,
ENGINEERING

Staged Data Warehousing with Gigabytes of Longitudinal Data

The GVSU Offshore Wind Assessment Project buoy generates approximately 20 gigabytes of raw, longitudinal data each year. It has been running from the end of 2011 to early 2014 with potentially continuing operation for several years. Our data warehousing measures receives the data as unindexed files with loosely defined variable sets to an organized database with standardized message types and date span folders. In order to utilize and share our large data most efficiently, we created a streamlined database file structure for internal analytical use and a web based data retrieval system for portable external environment use.

8:30- 9:00
PERE MARQUETTE

Heather Carpenter

PUBLIC NONPROFIT AND
HEALTH ADMINISTRATION

Connections and Overlap between Capacity Building Measures, Nonprofit Management Competencies and Training Needs of Nonprofit Managers

Studies conducted in the past twenty years have focused on the capacity building needs of nonprofit organizations, competencies of nonprofit managers as well as the training needs of nonprofit managers. However, many of these studies have been discussed in silos, separate journal articles and studied in different areas of the nonprofit management literature. There has been a scarcity of studies that document and track common terms across the nonprofit management literature. The author used social network analysis and identified 12 nonprofit management terms most commonly discussed within nonprofit capacity building, management competencies and training needs literature.

8:30- 9:00
PERE MARQUETTE

Nicholas Vogel

COMPUTING AND
INFORMATION SYSTEMS
ENGINEERING

Mining Enormous Mobile Datasets to Improve Mitigation Strategies for Limiting the Spread of Infectious Disease

We secured access to a dataset containing the entire anonymized call detail records (phone/text) of 5 million mobile phone subscribers in Cote d'Ivoire, tracked over a 5-month period. The goal of this work-in-progress is to analyze the dataset for information that could help public health officials develop more effective strategies for limiting the spread of infectious disease. Using antenna (cell tower) proximity data to situate subscribers, clustering algorithms were applied to identify groups of individuals expressing similar mobility patterns. Incorporating this knowledge of dynamic population densities could lead to better-informed quarantine/isolation decisions.

9:50 - 10:10
PERE MARQUETTE

Big Data Challenges in Bioinformatics – The Collision of Technology and Biology

Mary E. Winn

CORE MANAGER
BIOINFORMATICS &
BIOSTATISTICS CORE
VAN ANDEL RESEARCH
INSTITUTE

With advances in technology comes the ability of biologists to generate large amounts of data quickly and cost-effectively. This data explosion presents unique challenges not only in data storage, management, transfer, and analysis, but in arming biologists with the tools and knowledge to effectively use technology and interpret data.

9:50 - 10:10
2215/2216 KC

Graphing Big Data Using the SAS© System

Laura Ring Kapitula

STATISTICS

Have you ever wondered how to make informative and attractive graphics and reports in a time efficient manner? Do you prefer writing code over point-and-click or are you comfortable with point-and-click but would like to automate the creation of statistical reports? If so then this talk is for you. In this talk I will share some applied examples where data stored in multiple files is downloaded from the internet, prepared for analysis and used to make informative and attractive reports and graphics. I will also show how even a novice can exploit powers of SAS© Enterprise Guide to write SAS© graph code and input statements that are easily reused and edited.

10:15 - 10:35
PERE MARQUETTE

Shaily Menon

BIOLOGY

Biodiversity Informatics – Big Data for Biodiversity Conservation and Ecological Forecasting

Biodiversity informatics is the creation, integration, analysis, and understanding of information about biological diversity. The emphasis is on primary data of historical and current species occurrence points, which serve as inputs for predictive modeling and ecological forecasting of the likely effects of local and global change. Applications have included predicting the spread of invasive species and disease agents, and changes in species distributions due to climate change. I will share sources of big data as it relates to biodiversity informatics and some examples of applications of these data.

10:15 - 10:35
2215/2216 KC

Szymon Machajewski

COMPUTING AND INFORMATION SYSTEMS

Open Source Analytics – Blackboard Learn

Big Data is often addressed by small open source projects. A few of them include Apache Hadoop, Oracle Cluster File System, Storm, and Drill. BbStats is a Blackboard module, authored by Szymon Machajewski, to predict system performance and graph trends for course usage and user activity. The software is the second most frequently downloaded module for Blackboard at projects.oscelot.org. The BbStats project leverages the data generated by student activity and provides Business Intelligence reports powered by JavaScript libraries of Timeplot (based on MIT Simile project).

10:40 - 11:00
PERE MARQUETTE

Visual Thinking

Andy Van Solkema
VISUALHERO

Solving a problem cannot be done purely by thinking. Cognition is enhanced when it is embodied with action or graphics. Visual thinking becomes an extension of our brain and allows us to share and advance thought. Recent increase in information graphics and visual design is enabled by ease in technology and visual resources. I will look at patterns and how they may enhance and shape our understanding for thought. Andy Van Solkema, a graduate of GVSU in 2002, is the Principal Designer and Founder of Visualhero, a design studio focusing on design research, information and user experience design.

11:40 - 11:00
2215/2216 KC

Tutorial: NVivo 10

Mario Adkins
EDUCATION

My expertise is qualitative data analysis, and evaluation and assessment. I do this by using a program called NVivo 10. I believe this could be a vital tool for student services assessment, especially for coding textual data from students. I believe this could be an optimal talk into tutorial session. I have access to the software and have been trained on it at an out-of-town conference. I have also presented posters/papers in Las Vegas, NV and Jacksonville, FL this semester via visualized data obtained from NVivo 10. Note that the talk will also include general recommendations on processing/coding qualitative and quantitative (descriptive statistics) data analytics, regardless of if NVivo 10 is the program doing the processing.

11:05 - 11:25
PERE MARQUETTE

David Zeitler

STATISTICS

High Performance Statistical Computing

This past semester I've been working three separate big data projects while on sabbatical. This talk will highlight the activities and some preliminary outcomes. The projects are: Lake Michigan Wind Assessment, data quality and modeling with dozens of variables collected once a second for roughly 9 months each during 2012 and 2013. Meijer Marketing Analytics, working with trillions of transactions and hundreds of variables using SAS and SQL on their new Teradata high performance computational facilities. Empirical Spectral Test, collaborative research with WMU Computer Science and Statistics faculty using multidimensional spatial Fast Fourier Transforms in R on our analytics server and at the WMU High Performance Computational Science Laboratory.

11:05 - 11:25
2215/2216 KC

Jerry Scripps

COMPUTING AND
INFORMATION SYSTEMS

Data Mining the Liberal Arts

Data mining is one of the tool sets that has been enlisted to help make sense of the massive amount of data that has been collecting over the past few years. Most people associate its use with targeted marketing and government oversight. However, they can also provide scholars in the liberal arts with powerful tools to discover new insights and questions. In this talk, I will discuss some ways in which researchers have recently used data mining to uncover new knowledge in some diverse areas.

11:30 - 11:50
PERE MARQUETTE

Paul Stephenson

STATISTICS

A Statistical Analysis on the Likelihood of Independent Occurrences of Serious Residential Fires

At the request of Federal Prosecutor Mike MacDonald, the author was asked to investigate the likelihood of a series of fires that were reported by a landlord in West Michigan. In this presentation the author will discuss how data from the National Fire Protection Agency and the American Housing Survey (conducted by the U.S. Census Bureau) can be used to obtain estimates for the number of residential fires, the number of housing units in the US, and the proportion of fires that are "serious." More specifically, the talk will present how a model can be developed that quantifies the probability of the random occurrence of "X" number of serious fires in properties owned by the same individual over a "Y" month time period.

11:30 - 11:50
2215/2216 KC

Robert Deaner

PSYCHOLOGY

The Tortoise and the Hare: Men are More Likely than Women to Slow in the Marathon

Pacing in endurance races has long been of interest to sports scientists. However, research has been traditionally limited to small data sets of elite or sub-elite performances. To address pacing in non-elites, including the possibility of marked gender differences, we acquired data from 14 marathons encompassing 91,929 performances; for 2,929 individuals, we obtained running experience data from a race-aggregating website. Even after controlling for age, experience, finishing time, and differences in maximal oxygen uptake, men were three times as likely as women to slow markedly (running the second half of the race more than 30% slower than the first half).

11:55 - 12:15
PERE MARQUETTE

Creating “Baseball Motion Graphs”

Edward Aboufadel
MATHEMATICS

During 2013, I worked with a college baseball player from Oregon to create a video presentation using baseball statistics – RBIs, ERA, and payrolls. We were inspired by the videos of Hans Rosling to create “Baseball Motion Graphs”, using a free data visualization tool from Google. In this talk, I will explain how to use this tool to create animated data graphs. The baseball video can be found here: gvsu.edu/s/AP.

11:55 - 12:15
2215/2216 KC

New Curriculum Development in Data Science Programs

Paul M. Leidig
COMPUTING AND
INFORMATION SYSTEMS

Data Science is gaining recognition as an emerging multidisciplinary field often associated with big data. The McKinsey report projects “… a need for 1.5 million additional managers and analysts in the United States …” and “… that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions …”. They project that demand will exceed supply by as much as 50-60 percent by 2018. In response, faculty in the Department of Statistics and the School of Computing and Information Systems are proposing interdisciplinary programs. The two units have developed an undergraduate Minor in Data Science. The Masters in Computer Information Systems will offer a graduate certificate providing computing skills used in big data analytics. This presentation is intended to show how students in different disciplines can enhance their skill sets with these programs.

1:15 - 2:00
2201 KC

Andy Van Solkema

VISUALHERO

Can We Measure Visual Literacy?

Creativity in our culture is expanded through our visual mind. With intentional learnings, exercise and measurement of our visual literacy can we enable understanding, therefore enhance cognition and the pursuit and application of creativity. Is there an opportunity for measurement or are visuals too subjective? What could the ability to measure and teach achieve in our evolving world?

1:15 - 2:00
2215/2216 KC

Edward Aboufadel

MATHEMATICS

Big Data Where You Might Not Expect It (English, History ...)

Projects involving large data sets have emerged the past few years in a variety of fields, including some you would not expect. For instance, there is a new book entitled *Who's Bigger? Where Historical Figures Really Rank* which creates a list of significant people based on an analysis of Internet data. This session will begin with the sharing of a collection of these projects, and then move to a group discussion of the viability and appropriateness of these projects in teaching and scholarship.

NOTES

Please watch your e-mail for a suvey about today's conference to help us make the 2015 conference even better.

NOTES

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

GRAND VALLEY STATE UNIVERSITY