SCASA: SOUTHERN CALIFORNIA CHAPTER OF THE AMERICAN STATISTICAL ASSOCIATION

E-Tidings Newsletter

SCASA Events and News

VOLUME 6, ISSUE 8 AUGUST 2017

UPCOMING EVENTS

- ***R & SPARK: TOOLS FOR DATA SCIENCE WORK-FLOWS,*** Instructor: E. James Harner, Sept. 30 - Oct. 1, 9am to 5pm, at UCR (NISS workshop, flyer on the next page).
- ***QUANTILE REGRESSION IN PRACTICE,*** Instructor: Yonggang Yao, Oct. 20, UCLA, and Oct. 21, Orange County (ASA Traveling Course, details are being clarified).
- ***FIVE THINGS YOU NEED TO KNOW ABOUT QUANTILE REGRESSION***, Presenter: Bob Rodriguez, Nov. 9 or 16, at UCR (6th Annual Gokhale Undegraduate Lecture, details are forthcoming, given title is approximate).
- ***STATISTICAL MODEL BUILDING*** (talk 1) and ***LEADERSHIP IN STATISTICS*** (talk 2), Presenter: Bob Rodriguez, November 11 or 18 (SCASA Kick-off, details TBA, given titles are approximate).

NISS WORKSHOP

WEST COAST

R & SPARK: TOOLS FOR DATA SCIENCE WORKFLOWS

DATE: September 30 and October 1, 2017, 9 A.M to 5 P.M.

VENUE: Genomics Auditorium, University of California, 900 University Avenue, Riverside, CA 92521

CLASS CAPACITY: 35

FEES: The registration fee includes continental breakfast, lunch and afternoon snacks for each day

- For NISS Affiliate Institutions: $150 for regular registration, $100 for students.
- For Non-NISS Affiliate Institutions: $200 for regular registration, $125 for students.

COURSE OUTLINE: R is a flexible, extensible statistical computing environment, but it is limited to single-core execution. Spark is a distributed computing environment which treats R as a first-class programming language. This course introduces data structures in R and their use in functional programming workflows relevant to data science.

The course covers the initial steps in the data science process:

- Extracting data from source systems
- Transforming data into tidy form
- loading data into distributed file systems, distributed data warehouses, and NoSQL databases, i.e., ETL.

This workflow is illustrated by using the SparkR and sparklyr package frontends to Spark from R.

SparkR and sparklyr are then used as interfaces for modeling big data using regression and classification supervised learning methods. Unsupervised learning methods, such as clustering and dimension reduction, are also covered. Additional methods, such as gradient boosting and deep learning, are illustrated using the h2o and rsparkling R packages. Finally, methods for analyzing streaming data are presented. The course finishes with an in-depth example. The infrastructure and content is containerized for easy download to your laptop using Docker.

PREREQUISITES FOR THIS COURSE: Differential calculus, basic matrix algebra, a statistics course covering regression, basic R.

OPERATING SYSTEMS: MacOS 10.11 (El Capitan) or higher or Windows 10 Professional. Students must bring their own laptops.

CONTACT US: Direct questions about this course to the Instructor E. James Harner at **eharner@mail.wvu.edu** or call him on his cell phone at **304-376-4170**.

To enroll in this shortcourse, contact **officeadmin@NISS.org**

NISS

INSTRUCTOR:

E. JAMES HARNER

E. James Harner is Professor Emeritus of Statistics at West Virginia University (WVU). He was the Chair of the Department of Statistics for 17 years and the Director of the Cancer Center Bioinformatics Core for 15 years at WVU. Currently, he is the Chairman of the Interface Foundation of North America which has partnered with the American Statistical Association to organize the annual Symposium on Data Science and Statistics (SDSS), May 16-19, 2018 in Reston,VA. The areas of his technical and research expertise include: bioinformatics, high-dimensional modeling, high-performance computing, streaming and big data modeling and statistical machine learning.

National Institute of Statistical Sciences

1150 Connecticut Avenue NW, 9th Floor, Washington, DC 20036;
Tel: (202) 862-4316;
Fax: (202) 828-4130

Joint Statistical Meetings 2017 Summary Report

Article and Photography by Harold Dyck, CSU San Bernardino

The Joint Statistical Meetings are always exciting. Over 7,000 statisticians converge to present and discuss their ideas, renew acquaintances, dance, run, watch a ball game, eat (crab cakes!) and have a good time. This year's theme was "Promoting the Practice and Profession of Statistics". For me, Baltimore was a new experience which I thoroughly enjoyed with my wife, Cass. My schedule included Council of Section meetings, a JMP friends and users' reception, a reception for long-standing members, the ASA president's invited address, the Deming Lecture, sessions for special interest groups and various invited sessions, and meeting with vendors. JSMs have so many concurrent activities that you just have to pick and choose what interests you most—it's impossible to do it all.

Camden Yards from the Baltimore Convention Center

Of particular interest to SCASA members were the meetings for the Council of Chapters (COC) and the International Science and Engineering Fair (ISEF) meeting, where I was representing Southern California. At the COC business meeting, President-elect Karen Kafadar talked about her initiatives: leadership, personal communication and developing case studies. Executive Director Ron Wasserstein discussed training workshops ASA is developing; how statisticians can have global impacts by reviewing National Academy initiatives and advocating for the Greek chief statistician and former IMF economist, Andreas Georgiou, who was (again) found guilty of breach of duty (for being honest and professional); and that ASA officers are available to visit chapters. Chapters were encouraged to take advantage of the Traveling Speakers program as well as the initiative to provide $1,000 to chapters as a stimulus for putting on chapter events. (SCASA has been diligent in these areas, but, in my opinion, we need to continue to take advantage of both. We used stimulus money to help with catering at our Applied Statistics Workshop and for a microphone for our new book discussion club.) Announcements include: Dan Jeske is the new editor of American Statistician; print subscription rates on journals are increasing; member dues for students and registration for JSM are increasing; our work on statistical literacy continues; and the COC is looking for examples/stories of how statistics plays a role in society to create vignettes for radio broadcasts. I had the opportunity to share our Chapter's experiences with ISEF and to encourage the COC and ASA's support.

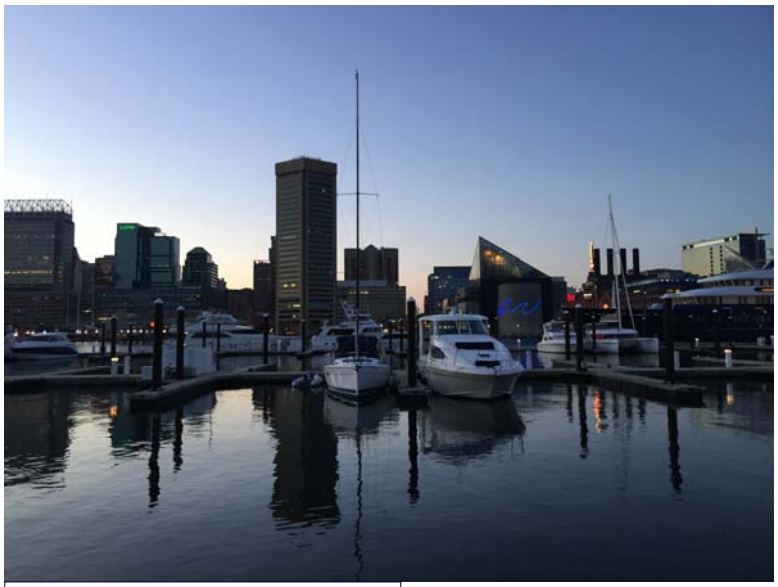
Continues on the next page....

Joint Statistical Meetings 2017 Summary Report

...Continued from the previous page.

At the ISEF meeting, Teri Utlaut and Tom Short gave an update on ASA's involvement in awarding special prizes for the best use of statistics in a project. Teri has been the ASA liaison to ISEF for a number of years and Tom is now transitioning into that slot. Many SCASA members have volunteered as judges every three years when the event was held at the LA Convention Center and our participation was noted and thanked. The rotation of LA, Pittsburgh and Phoenix is coming to an end after nine years, but the ISEF is expected to be in Anaheim in 2020. It is rumored to be in D.C. in 2021. Other changes: the SAO (Special Awards Organization) affiliation fee is going up from $1,000 to $1,500; each SAO is required to develop specific award selection criteria; and the SAO student award package minimum is now $5,000 (in dollars or experiences). In my opinion ISEF is a great opportunity for outreach and the Southern California contingent of statisticians has done an extraordinary job at increasing that outreach each time it has come our way. We need to continue our excellent progress. This is an opportunity to make an impact on students and our profession. To read more about ISEF, see the July issue of AmStat News (http://magazine.amstat.org/blog/2017/07/01/isefcomp/) and the forthcoming October issue, as well as past issues of SCASA's eNewsletter.

Next year's JSM will be in Vancouver. I encourage all of you to start making plans to attend!



Baltimore's Inner Harbor & Skyline

Calypso, a 14-years old (rescue amputee) sea turtle at National Aquarium in Baltimore, MD

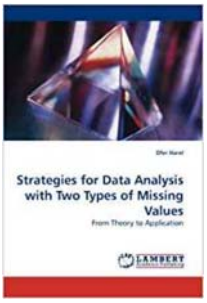
Announcing the 37th Annual Applied Statistical Workshop (ASW)

By James Joseph, Chair of ASW organizing committee, President-elect of SCASA, and VP for Professional Affairs

The SCASA is glad to announce that we've secured **Dr. Ofer Harel, Ph.D.** for the 37th Annual Applied Statistical Workshop to offer special insights on the Problem of Dealing with Incomplete Data. Once again, **City of Hope**, a dedicated sponsor for a very long time, has agreed to accommodate our Chapter in their spacious and professional venue, the **Arnold and Mabel Beckman Center in the Argyros Auditorium**, on **Saturday, April 21, 2018**.

ABOUT THE SPEAKER Ofer Harel, Ph.D. is a Professor and Director of Admissions at the Department of Statistics and a (past) Principal Investigator in the Institute for Collaboration on Health, Intervention, and Policy (InCHIP) at the University of Connecticut.

Dr. Harel received his Doctorate in Statistics in 2003 from the Pennsylvania State University; where he developed his methodological expertise in the areas of missing data techniques, diagnostic tests, longitudinal studies, Bayesian methods, sampling techniques, mixture models, latent class analysis, and statistical consulting. Dr. Harel received his post-doctoral training at the University of Washington, Department of Biostatistics, where he worked for the Health Services Research & Development (HSR&D) Center of Excellence, VA Puget Sound Healthcare System, and the National Alzheimer's Coordinating Center (NACC). Dr. Harel has served as a Biostatistical Consultant nationally and internationally since 1997. Through his collaborative consulting, Dr. Harel has been involved with a variety of research fields including, but not limited to Alzheimer's, diabetes, nutrition, HIV/AIDS, and alcohol and drug abuse prevention. Dr. Harel's book "Strategies for Data Analysis with Two Types of Missing Values From Theory to Application" came out in 2009.



BRIEFLY ON THE TOPIC Biased results and inefficient estimates are just some of the risks of incorrectly dealing with incomplete data, a common problem in applied research. This course will emphasize practical implementation of proposed strategies for dealing with missing data, including discussion of software to implement recommended procedures.

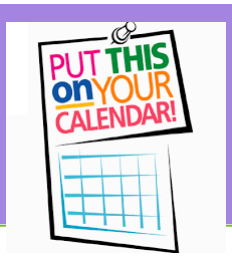
EVENT DAY SCHEDULE This all-day networking and course-style event will feature an early check-in option, Late Registration booth, catered breakfast, speaker session with discussion, lunch options, and a raffle.

SPECIAL RATES Join us on **Facebook.com/AmStatSoCal** or **Twitter.com/AmStatSoCal** and we will send you a personal message including a special thanks!

Additional announcements regarding ASW2018 to come.

Continues on the next page....

Announcing the 37th Annual Applied Statistical Workshop (ASW)



...Continued from the previous page.

SPONSORSHIP BENEFITS We can only strengthen our professional communities with your dedicated support! Our past sponsors include City of Hope & Amgen - if you are interested in returning as sponsor for our 37th Annual Statistical Workshop, becoming a new sponsor, or recommending a potential sponsor, then please contact me at ***james@offloft.com*** to discuss new benefits and initiatives.

STATISTICS BOOK DONATIONS If you haven't embraced stat books as a curious source of enjoyment and wisdom, try joining Daniel Jeske and guest Sherman Rizzo for our first installment of our reading club podcast, 'ScasaCast', where they discuss Nate Silver's 'The Signal and The Noise'.

Want to stay inspired and grow your knowledge? Join the club! Contact ***danielJeske@ucr.edu***.

BOOK RAFFLE DONATIONS To donate your Like-New stat books to our Raffle cause, please contact me at ***james@offloft.com*** and I will arrange delivery or pick-up.

ASA DataFest 2018 at Chapman University

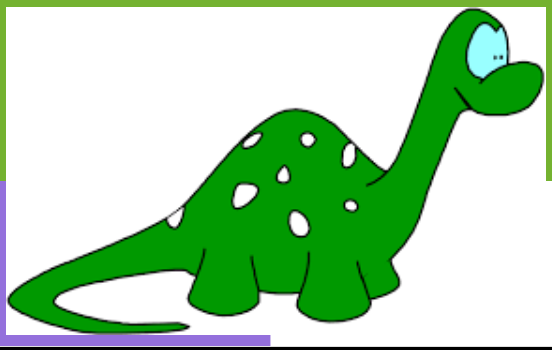
The **ASA DataFest** at **Chapman University** will be held **April 27-29, 2018**.

In the 7 years since Professor Rob Gould started DataFest in 2011 at UCLA with 30 students, it has grown to being held at 31 sites in 3 countries (see the May 24, 2017 blog "Growth of DataFest over the years" by Mine Çetinkaya-Rundel **https://rviews.rstudio.com/2017/05/24/growth-of-datafest-over-the-years/**)

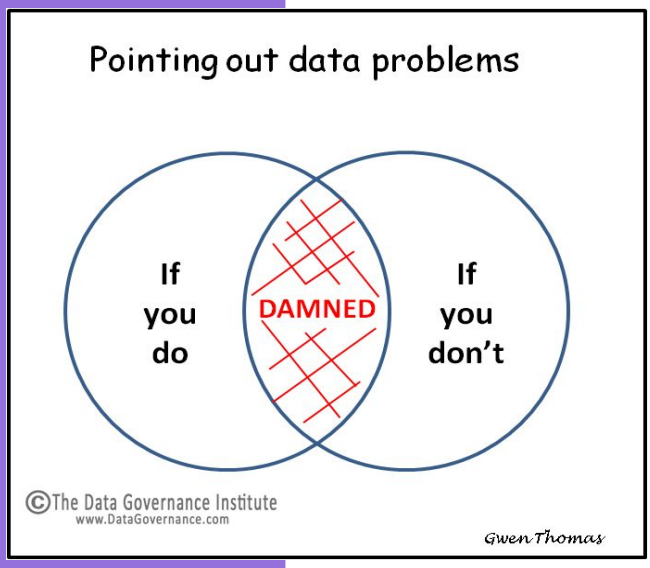
The UCLA site has not been able to keep up with the growing interest and demand in the Southern California area. Students from surrounding universities would come to the OCLB/SCASA Careers Day and ask statisticians how they could participate! This year, ASA DataFest was held at UCLA and also at Chapman University, thanks to Michael Fahy, Professor of Mathematics and Computer Science and Associate Dean, Schmid College of Science & Technology at Chapman University.

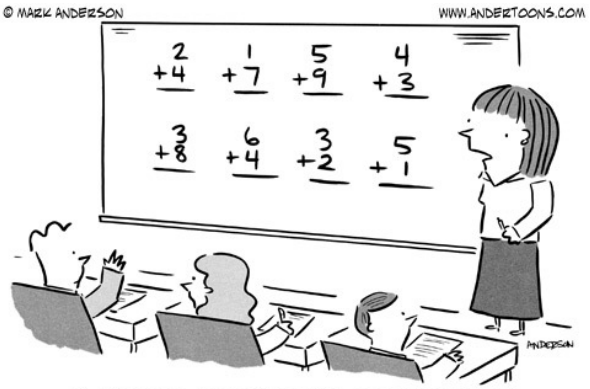
2018 DataFest registration for students and sign-up for statisticians and data scientists willing to mentor the students is available now at: **http://www.chapman.edu/datafest**

Now is the time for faculty to start organizing your student teams!



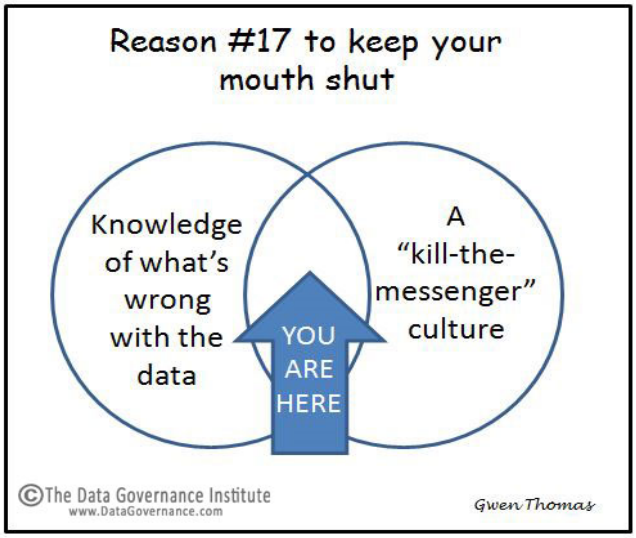
Dr. Normalcurvesaurus, Ph.D. presents

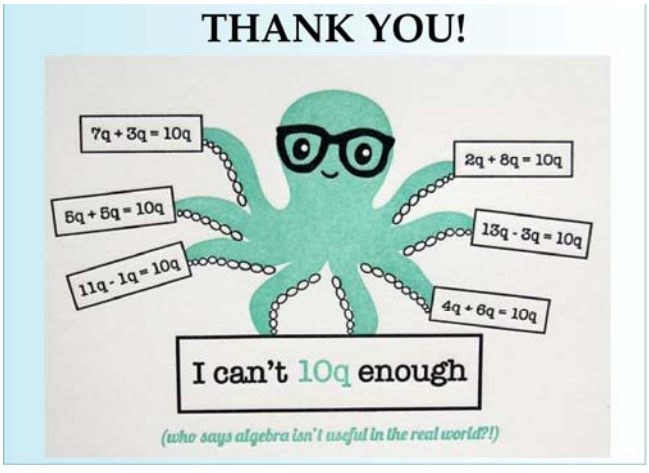




"Let's solve these first. We can worry about data mining later."

"It's important to remember that correlation does not Imply causation. Besides, we all know it was Brian."





If you would like to submit an entry to the next issue, please contact me at *Olga.Korosteleva@csulb.edu*.

Yours Truly,

Olga Korosteleva,

Your Editor-in-Chief