

# 2013 IEEE Conference on Semantic Computing (ICSC 2013)

*Special Session on Data Science*

*September 19, 2013*

*Organizer*

Dr. Joseph Barr, San Diego State University and University of California, Irvine

## **Morning Session**

***Data Analytics with Python (30 minutes each.)***

*Instructor: Joe Barr with graduate assistants*

8:45-9:00 Greetings from the organizers

Module 1: The basics of Python (Data types, logic, function, classes)

Module 2: Linear and Logistic Regression (with Python)

Module 3: Neural Networks

Module 4: Support Vector Machines

Module 5: Boosting

*Class material will be provided by organizers*

## **Afternoon session: Invited Speakers: Aspects of Analytics**

1:00 – 1:50

**Charles Elkan**, Professor of Computer Science, University of California San Diego

### **The past and future of analytics**

*Abstract:* This talk will explain some of the foundations of the field of analytics, and in particular will describe a prescient application from 1948 at the Sears Roebuck & Co. catalog. The talk will continue by linking this pioneering application to new methods from academic research, on what is called reinforcement learning, that will enable disruptive advances in customer lifetime value optimization, and in energy conservation and production. The talk will also describe the history and future of sentiment analysis, where new methods from academic research will also enable radical advances soon.

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 academic work in machine learning and data mining. The MEME algorithm developed by his group has been used in over 2000 research projects in biology and computer science. He has won several best paper awards and data mining contests. His M.S. and Ph.D. students have become leaders at companies including Google, Facebook, Amazon, and at many universities.

2:00 – 2:40

**Stephen Coggeshall**, Chief Technology Officer, ID Analytics

### **Identity Dynamics and the Future of Remote Authentication**

*Abstract:* Identity is a core concept in business applications with many nontrivial aspects and dynamics. What is identity? Can a person have multiple identities and under what circumstances? How do we determine which identity is truly behind a remote transactions, such as a purchase or change of information request? In this presentation we discuss problems and solutions with identity resolution and remote authentication solutions of today and tomorrow.

Dr. Stephen Coggeshall is the founding Chief Technology Officer at ID Analytics where he leads the Analytics team doing modeling, algorithms and R&D around very large data sets. He has spent the past 20 years building algorithms around large data flows to provide practical solutions to intractable business problems. He has built and managed teams of scientists at the Los Alamos National Laboratory, HNC Software, and Morgan Stanley. As an entrepreneur he has cofounded and successfully sold three companies: CASA, Los Alamos Computational Group and ID Analytics. Coggeshall was named a Technology Executive of the Year by the San Diego Business Journal in 2008 and awarded the Lifetime Achievement Award in 2012.

2:50 – 3:30

**Matthias Blume**, Sr. Director of Analytics, CoreLogic

**Real Estate Analytics** *Abstract:* This talk provides an introduction to analytics in the real estate industry, including automated valuation models, mortgage-backed security valuation, mortgage application fraud detection, home price index computation, HPI forecasts, and quantitative analysis of public policy and investor behavior. Specific examples of semantic computing include entity

resolution, correlating sentiment detected in Beige Books with market movements, and utilizing information extracted from Twitter. Finally, the presentation highlights the CoreLogic Academic Research Council (CLARC) data grant program, which enables academic research to support housing policy development and improved understanding of market dynamics.

Dr. Matthias Blume is Senior Director of Analytics at CoreLogic, the nation's largest real estate data provider. His team develops solutions for mortgage fraud detection, consumer credit scoring, automated valuation models, and more. Previously, he worked in marketing optimization, text analytics, and the gamut of financial services analytics at Redlign, Covario, and HNC/FICO. He received his PhD in Electrical and Computer Engineering from UCSD and a BS from Caltech.

3:40 – 4:20

**Richard Levine**, Professor & Chair, Department of Mathematics and Statistics, SDSU

***Survival Trees as Analytics Tools in Biomedical Applications***

*Abstract:* Survival tree methods for correlated observations have shown potential for developing objective biomedical prognosis systems, however the current technology suffers either from prohibitive computational expense or unrealistic simplifying assumptions to overcome computational demands. In this talk, we will introduce a suite of tree methods we developed for correlated survival data, relying on computationally feasible, yet flexible, frailty and joint longitudinal-correlated binary models. As part of the computational machinery, we will discuss Bayesian stochastic search and goodness-of-split methods for tree construction, (posterior) ensemble averaged variable importance ranking, and amalgamation procedures. We will motivate and illustrate our methods in two applications in dental and ophthalmic practice: assigning teeth to periodontal prognosis categories and assessing the effects of clinical factors and genetic polymorphisms in predicting tooth loss; detecting glaucomatous progression and studying risk factors of glaucomatous visual field deficits.

Dr. Rich Levine has worked as a statistical consultant since 1991 and Professor of Statistics since 1996, currently Professor and Chair of the Department of Mathematics and Statistics at San Diego State University. He is a Fellow of the American Statistical Association and the Past-Editor of the Journal of Computational and Graphical Statistics. He has served as lead-statistician on projects for the National Institutes of Health, Office of Naval Research, Department of Energy/Lawrence Livermore National Laboratory, CA Cancer Detection Service, CA Department of Corrections, and numerous University research projects, and has developed and offered statistical computing seminars and courses for scientists, engineers, and managers/business at all levels of quantitative acumen. His expertise includes general biostatistics design and analysis, computationally intensive statistical methods, and Bayesian hierarchical modeling, having developed and published internationally recognized data mining methods and software for Monte Carlo/simulation intensive statistical analyses. He has vast experience particularly with problems in cancer epidemiology, ophthalmology (glaucoma), genetic disorders, geophysics and climatology, ecological systems/ population dynamics, learning analytics, and sports.

4:20 – 4:30

Concluding Remarks (Organizers)