The Current State of Data Science at Chico State

Robin Donatello

15th SVCCM Conference

March 3, 2018

Last time….

- What is Big Data and Data Science?
- Why should Mathematicians care about these fields?
- What can Community College instructors do to train students to enter these fast growing fields?

This time…

Presenting the
Data Science Initiative
at Chico State!

Who/What is the Data Science Initiative (DSI)?

- A group of faculty who are interested in building a Data Science community and teaching DS topics.
- Semi-organized/ quasi-funded
- Me: Self Appointed Project Coordinator
  - Future proposal to formalize this position
- Others: Steering committee? Advisory board? Parties of interest?
  - Edward Roualdes – Statistics
  - Todd Gibson, Kevin Buffardi – Computer Science,
    - Essia Hamouda (Now at UC Riverside)
  - Rick Hubbard, Arash Neghban – College of Business

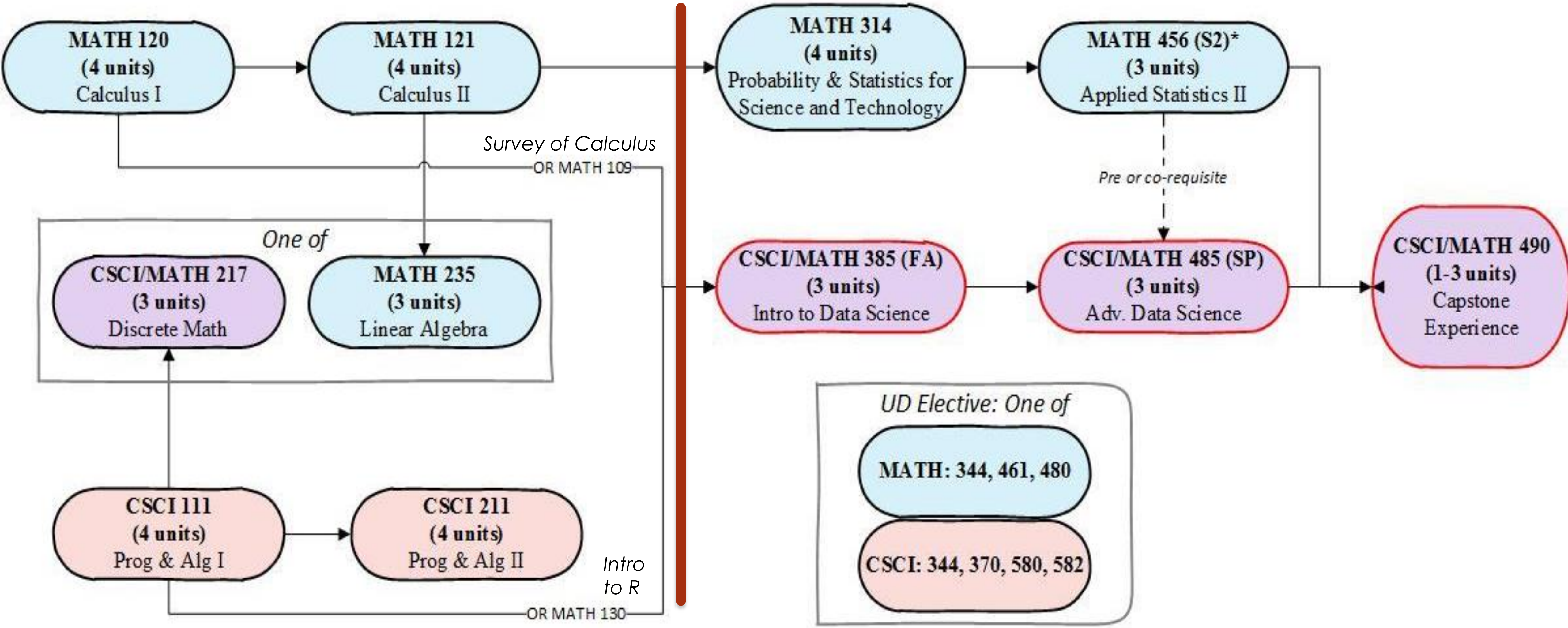
The DSI offers and sponsors:

- Traditional Classes – pathway to a Certificate
- Short courses
- Workshops
- Seminars
- Community Coding sessions
- Student events: Predictive Modeling Competitions / Hackathons

Three new courses in Data Science

- Introduction to Data Science
- Advanced Topics in Data Science
- Capstone in Data Science

Undergraduate Certificate in Data Science



MATH 120
(4 units)
Calculus I

MATH 121
(4 units)
Calculus II

Survey of Calculus
OR MATH 109

One of

CSCI/MATH 217
(3 units)
Discrete Math

MATH 235
(3 units)
Linear Algebra

CSCI 111
(4 units)
Prog & Alg I

CSCI 211
(4 units)
Prog & Alg II

Intro to R
OR MATH 130

MATH 314
(4 units)
Probability & Statistics for Science and Technology

MATH 456 (S2)*
(3 units)
Applied Statistics II

Pre or co-requisite

CSCI/MATH 385 (FA)
(3 units)
Intro to Data Science

CSCI/MATH 485 (SP)
(3 units)
Adv. Data Science

CSCI/MATH 490
(1-3 units)
Capstone Experience

UD Elective: One of

MATH: 344, 461, 480

CSCI: 344, 370, 580, 582

Introduction to Data Science

- The analytics life cycle
- Data integration and modeling in R/Python
- Relational databases and SQL
- Text processing and sentiment analysis
- Statistical models for clustering and classification
- Data visualization

Emphasis is placed on

- reproducible research,
- code sharing using version control
- communicating results to a non-technical audience
    - Writing tutorials
    - YouTube video tutorials!

Nitty gritty extracting, munging, wrangling and visualizing data.
Quality data is critical for proper insights.

Community driven student projects

"One thing I was always curious about was stats comparing the pre and post pitching and hitting stats for the steroid era. Many hitters took and are still taking mass criticism for steroids and have been banned from the hall of fame but I always wondered if the pitchers were affected as well."

"Meaning, did the hitters not actually gain an advantage by being on steroids because the pitchers were also on steroids?"

https://norcalbiostat.github.io/EDA/index.html

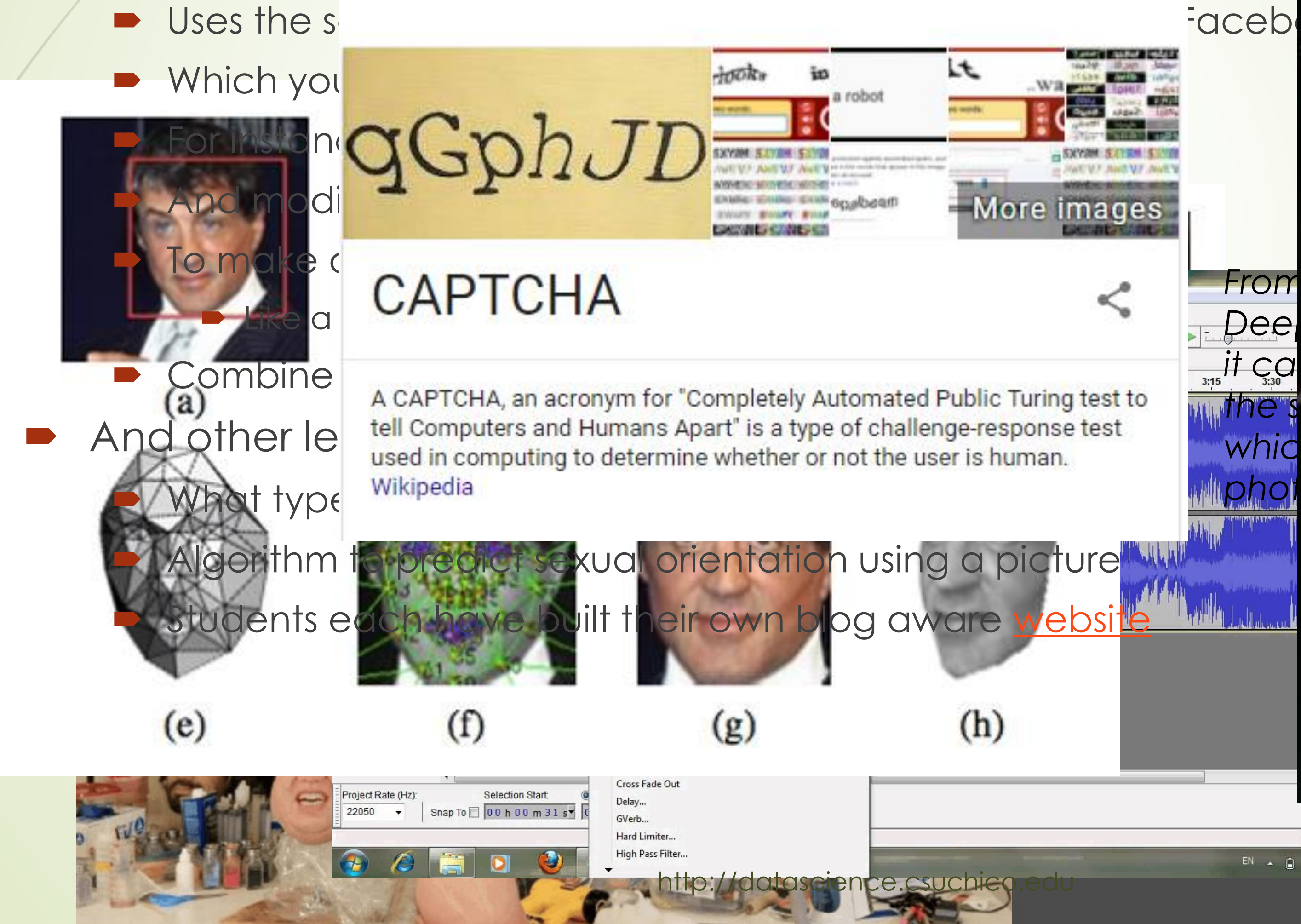
Advanced Topics in Data Science

- Train and validate machine learning predictive models for a variety of situations.
- Apply basic principles of parallel computing.
- Analyze data too large to fit on a personal computer.
- Identify the characteristics of a predictive model that make it potentially dangerous to society.
- Work effectively with a remote team using modern collaboration tools.
- Build and update a professional online presence.
- Create a self-service interactive dashboard for user-driven data questions.

Six weeks into the pilot course…

- We've talked about how letter pattern identification
  - Uses the s… Faceb…
  - Which you…
  - For instanc…
  - And modi…
  - To make c…
    - Like a
  - Combine
- And other le…
  - What type…
  - Algorithm to predict sexual orientation using a picture
  - Students each have built their own blog aware website

From
Dee…
it ca…
the s…
whic…
phot…
…n



CAPTCHA

A CAPTCHA, an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart" is a type of challenge-response test used in computing to determine whether or not the user is human.
Wikipedia

Data Science Capstone

- Students will work independently to provide a service in the form of a data product to a local business, researcher, or community member.
- Students will generate a public facing data product that can be used to show their skills to potential employers. Students will experience the cycle of innovation, critique and revision from peers and external stakeholders.

Short Courses

- Introduction to R (MATH 130)
  - Just enough to ~~make them dangerous~~ get them going.
- Introduction to Python *(planned)* & SAP Hana (*Business, planned*)
- 2-day Carpentry Workshops (graduate students & faculty initially)
  - Software Carpentry teaching researchers the computing skills they need to get more done in less time and with less pain.
  - Data Carpentry develops and teaches workshops on the fundamental data skills needed to conduct research. Our mission is to provide researchers high-quality, domain-specific training covering the full lifecycle of data-driven research.

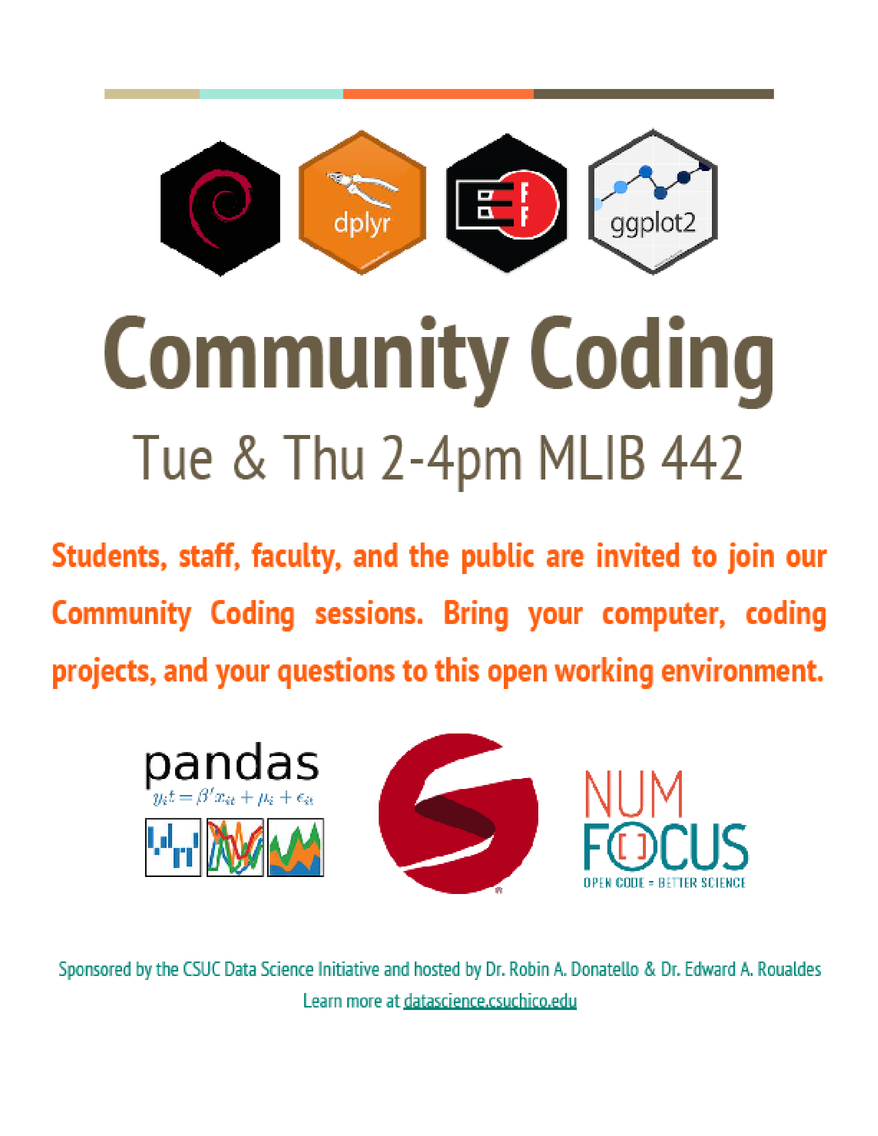
Workshops & Seminars

- (Semi) Weekly
- Presenters are faculty, staff, students, and the public.
  - Anyone with a tip, program, tool, or some interesting analysis that they want to share.
- 30 min to 2 hours
- Centralized location (Library)
- Open to the public

Date & Time	Topic	Presenter
Th 3/08: 3pm	Database Joins using SQL	Robin Donatello (STAT)
Tue 3/13: 2pm	BIOL 482: Bioinformatics Progress Report	Gordon Wolfe (BIOL) & Dave Keller (BIOL)
Thu 3/15: 3pm	Introduction to Bayesian Modeling	Edward Roualdes (STAT)
Thu 3/29: 2pm	Chico State Data Science Certificate: Informational Session	Robin Donatello
Thu 4/05: 3pm	Getting Started with Python	Grant Esparza (CSCI)
Thu 4/12: 3pm	Data Wrangling & Analysis with Python	Edward Roualdes
Thu 4/19: 3pm	Data Visualization with ggplot	Robin Donatello
Tue 4/24: 3pm	Data Fest Prep!	Robin Donatello
Thu 4/26: 2pm	SAP Lumeria and Predictive Analytics	Arash Negahban (BSIS)
Thu 5/03: 2pm	Marketing Analysis	Laurie McConville (IR)

Community Coding

- Open to everyone sessions
- Independent / collaborative work
- “Meet and analyze data”
- Quasi-office hours
- Social coding... but not too social
- Same time/place as the workshops
- Establishing a habit / place
    - “Oh you need help in XYZ? You should go to [here] on [day/time]. There’s always someone who can help”


dplyr
ggplot2
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
NUM FOCUS
OPEN CODE = BETTER SCIENCE

Community Coding

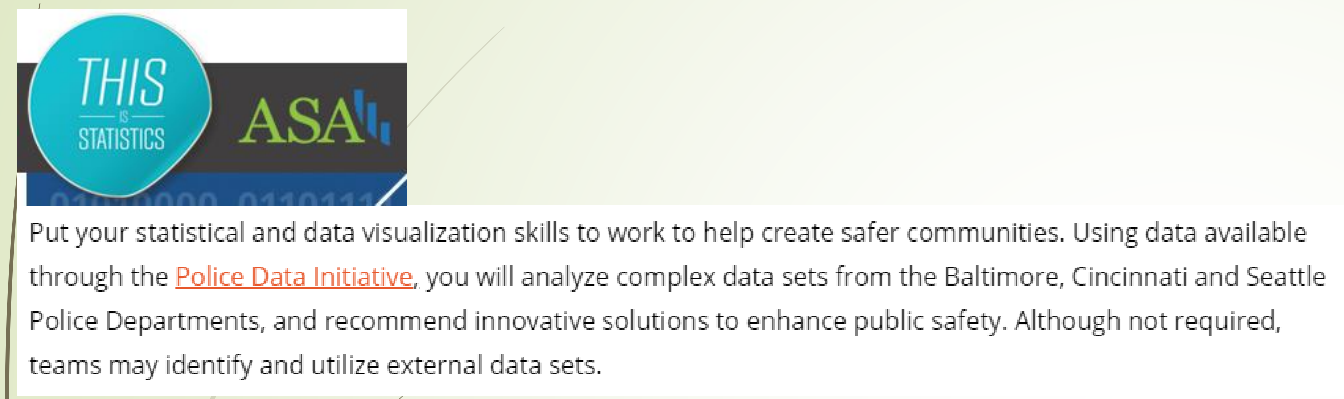
Tue & Thu 2-4pm MLIB 442

Students, staff, faculty, and the public are invited to join our Community Coding sessions. Bring your computer, coding projects, and your questions to this open working environment.

Sponsored by the CSUC Data Science Initiative and hosted by Dr. Robin A. Donatello & Dr. Edward A. Roualdes

Learn more at datascience.csuchico.edu

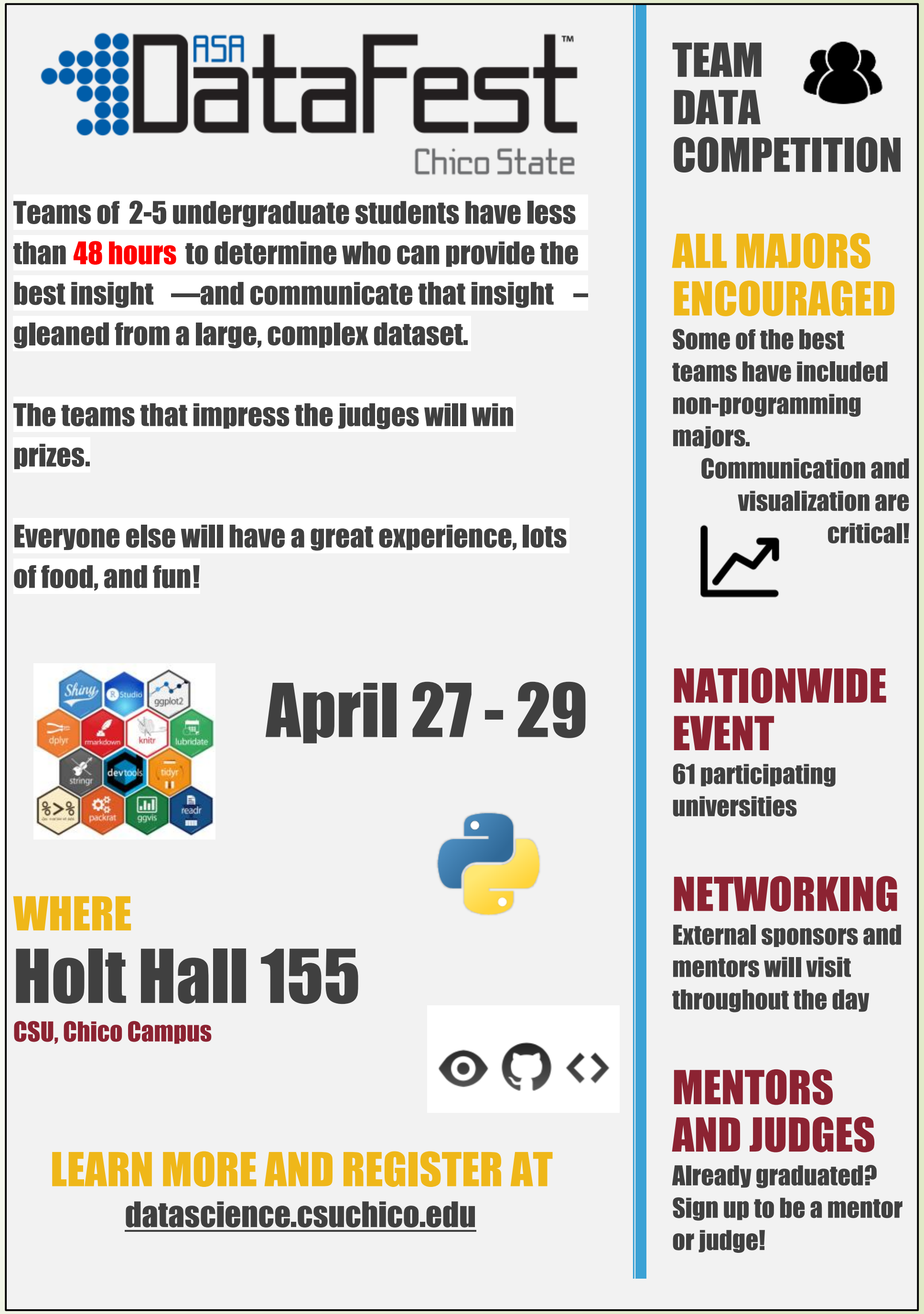
Student Events



Put your statistical and data visualization skills to work to help create safer communities. Using data available through the Police Data Initiative, you will analyze complex data sets from the Baltimore, Cincinnati and Seattle Police Departments, and recommend innovative solutions to enhance public safety. Although not required, teams may identify and utilize external data sets.

Predictive Modeling Competition at UC Davis iiData 2017 conference (2nd place!)

Detecting Basketball Shots from Motion Tracking data
Aaron Shaffer
Ricardo Aguilar
Nicholas Eisemann
Advisor: Dr. Robin Donatello, DrPH
California State University, Chico
March 25, 2017



ASA DataFest™
Chico State

Teams of 2-5 undergraduate students have less than **48 hours** to determine who can provide the best insight —and communicate that insight – gleaned from a large, complex dataset.

The teams that impress the judges will win prizes.

Everyone else will have a great experience, lots of food, and fun!

April 27 - 29

WHERE

Holt Hall 155

CSU, Chico Campus

LEARN MORE AND REGISTER AT

datascience.csuchico.edu

TEAM DATA COMPETITION

ALL MAJORS ENCOURAGED

Some of the best teams have included non-programming majors.

Communication and visualization are critical!

NATIONWIDE EVENT

61 participating universities

NETWORKING

External sponsors and mentors will visit throughout the day

MENTORS AND JUDGES

Already graduated? Sign up to be a mentor or judge!

What’s next?

In order of implementation

- Care and feeding of the certificate program.
- More events, outreach, student, faculty and staff training.
- Programming workshops open to the entire community.
- College of Business
  - BS in Business Information Systems
    - Option in Management Information Systems
    - Option in Operations and Supply Chain Management
    - **Option in Business Analytics**
- Center for Data Science Research?
- Professional Certificate / Masters Program? Online?

Elsewhere,

What are other campuses doing?

University of California

- Irvine: Data Science Initiative, Certificate (Data Science, Predictive Analytics), BS
- Davis: Data Science Initiative, MS Business Analytics, BS Statistical Data Science, upcoming Data Science unit
- Berkeley: Division of Data Sciences, DS Professional certificate, a DS Major and Minor are in the approval process now.
- San Francisco: Data Science Initiative (Library) MS in DS, BS in DS
- San Diego: Data Science program BS, DS Minor, MS in Data Science and Engineering, online MicroMasters program
- Los Angeles: MS Business Analytics, Extension Specialization, courses in Engineering and Statistics
- Riverside: Online Masters
- Santa Barbara: Data Science student club
- Santa Cruz: Center for Data, Discovery and Decisions. *("loosely affiliated group of students, researchers and faculty who are data science enthusiasts")*
- Merced: BS Applied Math Computational and DS Emphasis

Interest ranges from fully funded research and teaching oriented divisions, to a combination of computational math & applied stat classes.

North State CSU’s

- Humboldt: none
- Sacramento: Center for Business Analytics, MS Certificate in Data Mining, Data Science Student Club *(*started by a CC xfer!)*
- San Jose: MS Data Analytics, DS Specialization (MS Software Engineering) Undergraduate pathway (School of Info), DS Post-masters certificate
- San Francisco: MS In Data Science, MS in Business Analytics,
- Sonoma: None
- Fresno: Certificate in Applied Data Analysis
  - 22 weeks + final project. That’s it.

Primarily Masters programs
Few UG programs & Continuing Education

Community Colleges

- Two-Year College Data Science Summit May 10-11, 2018
  - NSF funded workshop to bring together a diverse group of participants to make recommendations for two-year college data science programs, keeping in mind the needs of each of three student populations
    - Those seeking employment following an associate's degree
    - Those seeking transfer to four-year programs
    - Those seeking certificate programs and college-level courses in data science for professional development
- Steve Pierson @ ASA Generating a list of DS programs
  - Business, Computer Science, Statistics

Questions?

http://datascience.csuchico.edu