Motivation

- end-of-term projects, and even capstone experiences, can become run-of-the-mill

- students seeking to please the instructor, sometimes assuming instructor already knows how to do the analysis; and so their input isn’t really important.

- depending on project, the best students might not get to “shine”

- grades are important, but can discourage risk-taking

- there are few opportunities to wrestle with complex, real data
What is ASA DataFest?

• A celebration of data!
• A competition for teams of undergraduates to find insight and meaning in a rich and complex data set.
• A data hackathon

A typical DataFest

• Friday Night
  • Meet the data
• Friday night through Sunday afternoon
  • Work furiously. Eat.
  • Talk to roving statisticians from industry and academics
• Sunday afternoon
  • 5 minute presentations to judges
  • Winners announced
Data

• 2011: Los Angeles Police Department Arrest Reports
  Make a policy recommendation to reduce crime in Los Angeles.

• 2012: kiva.com lending data
  What motivates people to lend money, and what factors are associated with paying loans?

• 2013: eHarmony dating data
  What qualities do people look for in prospective dates?

• 2014: GridPoint energy consumption data
  How can clients best save money and energy?

• 2015: Edmunds.com
  Detect insights into the process of car shopping to make shopping process easier for visitors.

• 2016: Ticketmaster
  How can fans be better connected to the concerts they wish to attend?

Prizes

• Best Insight

• Best Visualization

• Best Use of External Data
Not StatsFest

• Not about statistical modeling
• No pre-defined “correct” outcome
• Many access points for students at different levels
• Emphasis on data
• “fast” analysis

Why DataFest?

• Friendly competition brings out best
• “Group work” in a setting that actually requires teamwork
• Access to complex data that isn’t available to (most) classrooms
• Cultural indoctrination (the “secret sauce”?)
Choosing the data

• The data must have a personality!
  • a spokesperson explains why the data are important and what they hope to learn

• Many variables (p more important than n)

• Aim for about 1 GB

• Context is key: accessible, interesting, cool

• 5-6 months time working with data donor to prep data

“Secret Sauce”

• “To my mind, the crucial but unappreciated methodology driving predictive modeling’s success is...the Common Task Framework”
  — D. Donoho “50 Years of Data Science”
CTF Key Features

• Shared data
• A set of competitors
• Judges

In Donoho’s setting, the goal is prediction. But more generally, DF encourages improvement through shared information between communities.

community

Alumni  Professionals
Undergrads
Faculty  Grad students
hosting your own

amstat.org/education/datafest

Hosting an Official ASA DataFest

To be an official ASA DataFest, organizers must agree to use the official data set and abide by a few ground rules. For more information, contact npoul@stat.ucla.edu and join the listserv by sending a blank email to sympa@sympa.cts.ucla.edu with subject header subscribe stat_datafestinabox.

Participating Institutions

- March 18-20 — University of Virginia
- April 1-3 — Duke University (Host), NCSU, UNC Chapel Hill, North Carolina A&T State, Meredith College, Elon, and North Carolina School of Science and Math
- April 1-3 — Five Colleges at the University of Massachusetts (host) with Smith College, Hampshire College, Amherst College and Mt. Holyoke College
- April 1-3 — BOW at Olin College (host) with Babson College and Wellesley College
- April 1-3 — Macalester College
- April 1-3 — Mannheim/Munich
- April 1-3 — Purdue University
- April 8-10 — University of California, Berkeley
- April 8-10 — Loyola University Chicago
- April 8-10 — The Ohio State University
- April 8-10 — Vassar College
- April 8-10 — Vermont at St. Michaels College (host) with Middlebury College
- April 8-10 — Wesleyan University
- April 8-10 — DC Summit: University of Maryland (host)
- April 15-17 — Oregon State University
- April 22-24 — Penn State University
- April 22-24 — Virginia Tech
- April 23-24 — Brown University
- April 29-May 1 — Miami University of Ohio (host) with University of Cincinnati
- April 29-May 1 — UCLA (host) with the University of California, Riverside; University of Southern California; Pomona College; and Cal Poly Pomona
- April 29-May 1 — University of Toronto
More

Video


Articles

- A Celebration of Data  CHANCE Magazine
- A Festival of Data: Student Perspectives  Amstat News
- Big Data Goes to College  Amstat News
- The Students Most Likely to Take Our Jobs  FiveThirtyEight
- DataFest draws hundreds to solve data challenges  Duke Chronicle
- Got Data? 200+ Crunch Numbers for Duke DataFest  Duke Research Blog
- Duke Wraps Fourth Annual DataFest State Competition  Campus Technology
- Edmunds.com sponsors data set for fifth ASA DataFest at UCLA  UCLA Daily Bruin
- DataFest 2015  Data Skeptic Podcast
- DataFest: Finding and Communicating the Meaning in Data  Pearson Higher Ed blog
- DC Area Undergrads Launch Successful 2015 ASA DataFest  Summit Consulting blog
- DataFest Competition Comes to DC  The Hoya (Georgetown University)