Automating Data Science: Think About the Human-Machine Interface

Rebecca Nugent
Carnegie Mellon Statistics & Data Science

NASEM Opportunities for Accelerating Scientific Discovery

Data Science, A View

Thought of as an process or workflow; solving real problems with real data

J. Wing (2019), Harvard Data Science Review

- *Management* includes security, elements of data engineering
- *Interpretation* includes communication

In practice, move roughly from left to right but with loops and iterations; experts often focus on specific pieces; project managers oversee pipeline

Carnegie Mellon University

- Private university in Pittsburgh, PA
- R1 research university designation
- $\approx$ 7000 undergrads, 7000 grads
- Seven colleges (admission is college-specific) College of Fine Arts, Dietrich College of Humanities & Social Sciences, College of Engineering, Heinz College of Information Systems and Public Policy, Mellon College of Science, School of Computer Science, Tepper School of Business
- *In Dietrich*: Center for the Neural Basis of Cognition, Economics (joint in Tepper), English, History, Information Systems, Institute for Politics and Strategy, Modern Languages, Philosophy, Psychology, Social and Decision Science, Statistics & Data Science
- Heinz College; Computer Science Department and Machine Learning Department in School of Computer Science; Mathematical Sciences in Mellon College of Science; related degrees use courses across colleges

Statistics & Data Science

PhD programs

- primarily Statistics, joint in Stat/Public Policy, Stat/Machine Learning, Stat/PIER, etc; about 65 students total
- Master's in Statistics affiliated with PhD first year

Master's in Statistical Practice program

- one year terminal master's focusing on data science, applied statistics, industry preparation
- about 35 students a year

Master's in Computational Finance program

- 1.5 year terminal master's; jointly across StatDS, Math, Tepper
- about 100 students a year (50 in Pittsburgh, 50 on NYC campus)

Undergraduate programs

- Statistics, Economics-Statistics, Statistics and Machine Learning, Mathematical Statistics (track), Statistics and Neuroscience (track)
- over 550 majors; Statistics and Machine Learning rapidly growing

(Some) Data Science-Related Grants/Projects

- *Teaching the Teachers: Data Science for STEM Educators*
  "Intro to Data Science" Workshop for CMU graduate students
  Data Science Summer Workshops in partnership with Libraries: Carpentry Workshops followed by Data Analysis projects/presentations
- *Seeding and Synergizing a Data Science Community Corps Network*
- *Tackling Complex Data Challenges Across Tipping Point Areas in the Physical Sciences*; beginning phases of an institute/hub
- *Carnegie Mellon Sports Analytics:*
  Faculty/PhD research, conferences, workshops, heavy outreach component, summer research programs
- *Women in Data Science Pittsburgh @ CMU* : networking and research events across multiple campuses; strong engagement within Data Science/AI industry community, local schools, etc
- *Digital Humanities* - partnerships across Libraries, Statistics & Data Science, English, School of Computer Science (largely LTI)

All part of larger experiential learning and research framework

Data Science Experiential Learning

Corporate Capstone Program

- Recent job market has strongly pulled students toward industry
- Impact of experiential learning can be large; real-world experience, learning to work with outside clients, etc
- Match teams of students to projects with clients
- Students have both faculty supervision and PhD assistance
- Educational Project Agreement (semester-long, year, etc);

Data for Good: Expanding a separate branch for non-profits/social services/government organizations; piloting this summer. "Undergraduate Arm" of Block Center for Technology and Society.

Campus-wide effort to build a "Data Science Corps"

Examples from Industry

- Forecasting/nowcasting the flu using disparate data sources (`https://delphi.cmu.edu/`)
- Global disruption of supply chains; determine effect of social and political events on deliveries; predict need to re-route earlier
- Use structures/unstructured financial data to predict market behavior
- Building predictive models to characterize injury/incident rates at construction sites
- Using scanned receipts to describe consumer behavior; improve marketing, item availability
- "Tinder for Brands" - based on opinion surveys, match celebrities to brand endorsements
- Build models to predict expected win probability play-by-play: `nflscrapR`

Lessons Learned

- By far and away, the important issues have NOT been the (AI) modeling
- All about the questions, the data, their location, the data integration, more data integration, and communication
- Without context, the solutions aren't actionable;
  starts with people, ends with people
- Only can really automate portions of workflow
- If data scientists aren't working hand-in-hand with the content experts, the results are less impactful

The Science of Data Science

Huge emphasis on having reproducible and/or replicable results

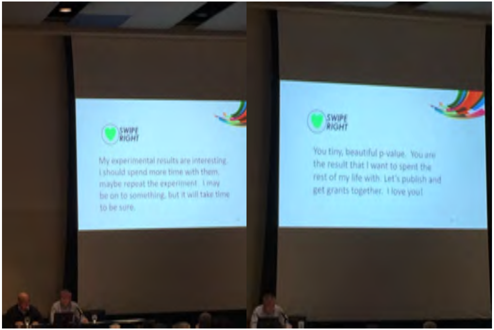
- **Reproducibility**: ability to implement the same experiment/code/procedures with the same data to obtain the exact same results
- **Replicability**: obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data (NAS)

Most can agree on need to carefully document all code, analyses, algorithms; slightly smaller group would add requirements to public post/disseminate all work, code, data sets, etc.

The Science of Data Science

About 15 years ago, there was a loud cry for the end of the p-value *(Ioannidis, PLOS, 2005)*. That cry is calmer now but gaining momentum.

"Moving to a World Beyond $p < 0.05$" *(Wasserstein (ASA); Schirm (Math Pol Res)*

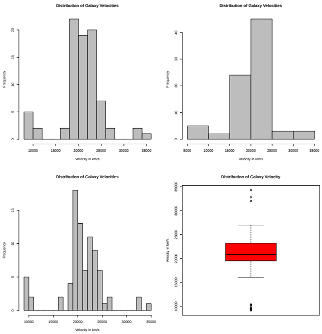


How a p-value of 0.03 is really like swiping right on Tinder - not so much a lifelong commitment but more just a sign of interest.....

The Science of Data Science

While much of data science relies on extracting signal/structure using machine learning algorithms, much is based on human subjective decisions.

Velocities of 82 galaxies; multimodality - voids and superclusters (Roeder, JASA, 1990)



Don't really understand how people do data analysis. Have to study people.

The Science of Data Science

Many analysts, one dataset *(Silberzahn, et al 2018)*

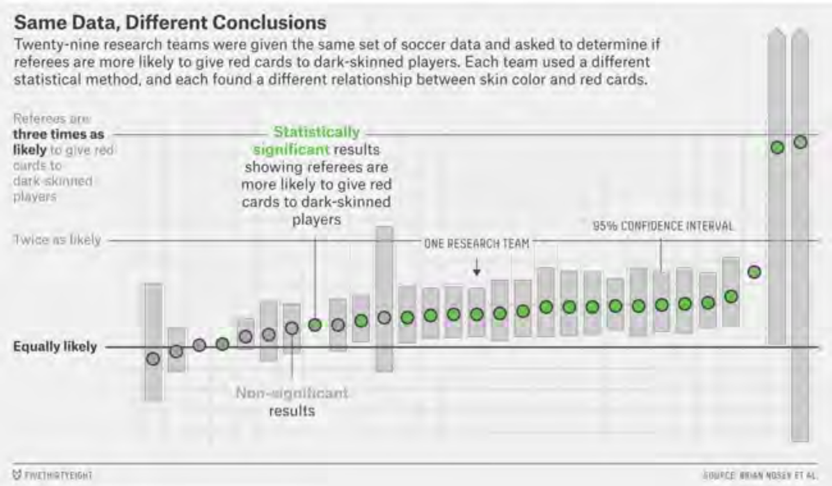
29 teams of analysts, same dataset, same question:

> *Are soccer referees more likely to give red cards to players with dark skin than to players with light skin?*

Analysis stages:

- Teams worked independently
- Peer-review, exchanged information and analysis
- Revisions and submit final conclusions

The Science of Data Science



Impact of EDA on Modeling Decisions

Study on CMU early/advanced regression students (T Lee, Mejia, Nugent)

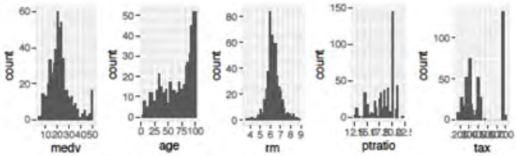


Figure 1: Univariate EDA plots given to group A participants

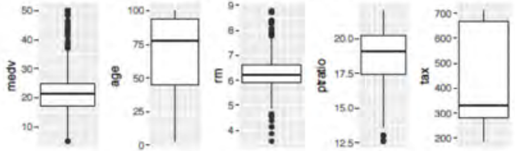


Figure 2: Univariate EDA plots given to Group B participants

Downstream we saw wide range of final predictions, p-values

The Science of Data Science

Common issues that we face teaching and working with these concepts:

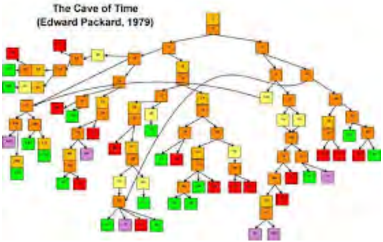
- **Reproducibility**:
  - How do people keep track of their work?
  - Writing (commented) code is one way;
    what happens with tasks without code?
  - More importantly, how do you keep track of decisions?
- **Replicability**:
  - People commonly work on one piece at a time
  - See variation in simulation-based/sampling distributions activities,
    not in real data analysis decisions
  - People rarely all work on/present the same project

Some current actions/questions:

- *Think-Alouds*: recording what you're thinking while doing your work
- *Crowd-Sourcing*: have groups work independently on same problem;
  how do you reconcile differences in data analysis variations?
- *Data Analysis Population*: Is our one data analysis is "different"?

The Science of Data Science

The Ultimate Choose Your Own Adventure Book (where hopefully the data analysis doesn't lead to being trapped in a cave forever):



With apologies to Edward Packard

Which Experiment to Run and When

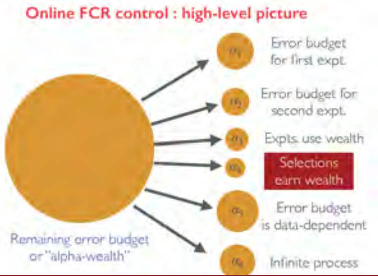
Aaditya Ramdas, StatDS/ML CMU (current NSF CAREER)

- Issues within each experiment (tech industry)
  - Continuous monitoring
  - Unknown experiment horizon
  - Flexible stopping rules
- Issues across experiments (tech + pharma)
  - Selection bias (multiplicity)
  - Dependence across experiments
  - Don't know future outcomes

Which Experiment to Run and When

The foundations of “doubly-sequential experimentation"

- Inner sequential process
  - Confidence Sequence
  - Valid (post-hoc) inference at different stopping times
  - No pre-specified sample size; can extend/stop exp adaptively
- Outer sequential process
  - Controlling False Coverage Rate online
  - Assign target conf level; end of exp., decide whether to report



The Democratization of Data Science

- Explosion of Stat & Data Science programs, courses, materials
- The People's Science
- Initial phase has largely been concentrated on "low-hanging fruit" disciplines (Stat, CS, IS, etc)
- Unleash the real potential of Data Science and Data Literacy when accessible and equitable for all
- Data Analytics Platform: Integrated Statistics Learning Environment (ISLE)
- Tackling accessibility and demand through interactive browser-based platforms that provide "one-stop shop" for data science
- Thinking about population of possible data analysis workflows; how can we adapt data science for different groups to enhance skill sets
- ISLE being used by hundreds of CMU students, in beta at other universities, professional training in multiple industries, and UN pilot initiative supporting data science in developing countries

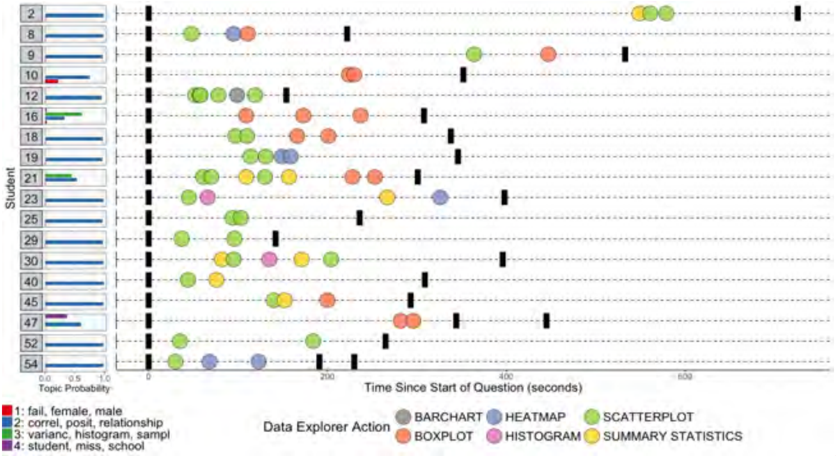
Integrated Statistics Learning Environment (ISLE)

- No coding syntax. Coding *concepts*.
  More direct interaction with the material
- Easy to adapt new hypotheses, problems during class/lab
- Can collect answers from students; propagate through labs
- "Data Set Explorer": upload (formatted) data, variables
- Students can save graphs and work to editors that create reproducible websites/documents for a portfolio
- We collect information on clicks, decisions, times, text, anything
  How/why do students analyze data?
- Combining tools like Java Script with RMarkdown;
  built in modular form; can "mix-and-match"
- Initial development by Philipp Burckhardt

Hard to know best practices if you have no idea what they're doing

Integrated Statistics Learning Environment (ISLE)

Topic models linking answers to timelines of their actions



ISLE

http://www.stat.cmu.edu/isle

When we create tools and automation, what is the effect on the workforce? What is the societal impact? Who is left out?

rnugent@stat.cmu.edu, http://www.stat.cmu.edu/∼rnugent; @CMU_Stats