Some Issues in Data Science and Statistics Education

Jessica Utts, Ph.D.
Professor Emerita of Statistics
University of California, Irvine

Outline of Talk

- Part 1: Developing a data science major at University of California, Irvine
  - My experience developing the major in 2015 when I was chair of the Statistics department.
  - At the same time I was President-elect of the American Statistical Association (ASA), and helped by ASA reports
- Part 2: Including ethics in statistics education
  - My interest in statistical literacy for consumers includes this topic
  - Ideas about how to include ethics for both producers and consumers of statistics

Some Sources for Part 1 of This Talk

- American Statistical Association (ASA) Data Science Statement, Aug 2015
  - *The Role of Statistics in Data Science*
- ASA Curriculum Report, Nov 2014
  - *Curriculum Guidelines for Undergraduate Programs in Statistical Science*
- UC Irvine Data Science major*

*Thanks to Stacey Hancock for some of the DS major slides.

Some History of Statistics and Data Science

- Statistics has evolved with technology and the growth of data
  - Statistics from the 1990s ≠ Statistics today!
- Foundational *goal* is the same
  - ASA's vision statement says it well: "A world that relies on data and statistical thinking to drive discovery and inform decisions"
- But *methods* for achieving that goal have changed and expanded

ASA Statement on the Role of Data Science in Statistics, 2015

- Identifies foundational data science fields as:
  - Statistics and machine learning
  - Database management
  - Distributed and parallel systems
- Encourages greater, mutually beneficial collaboration across these three fields
- Data Science intersects with numerous disciplines and related research areas

ASA Statement, Continued

- Notes that statistics education must evolve to meet needs
- Discusses the role of statistics in data science
- "New problem-solving strategies are needed to develop 'soup to nuts' pipelines that start with managing raw data and end with user-friendly efficient implementations of principled statistical methods and the communication of substantive results."

Links Related to the ASA Statement on the Role of Data Science in Statistics

- Statement

  https://www.amstat.org/docs/default-source/amstat-documents/pol-datasciencestatement.pdf

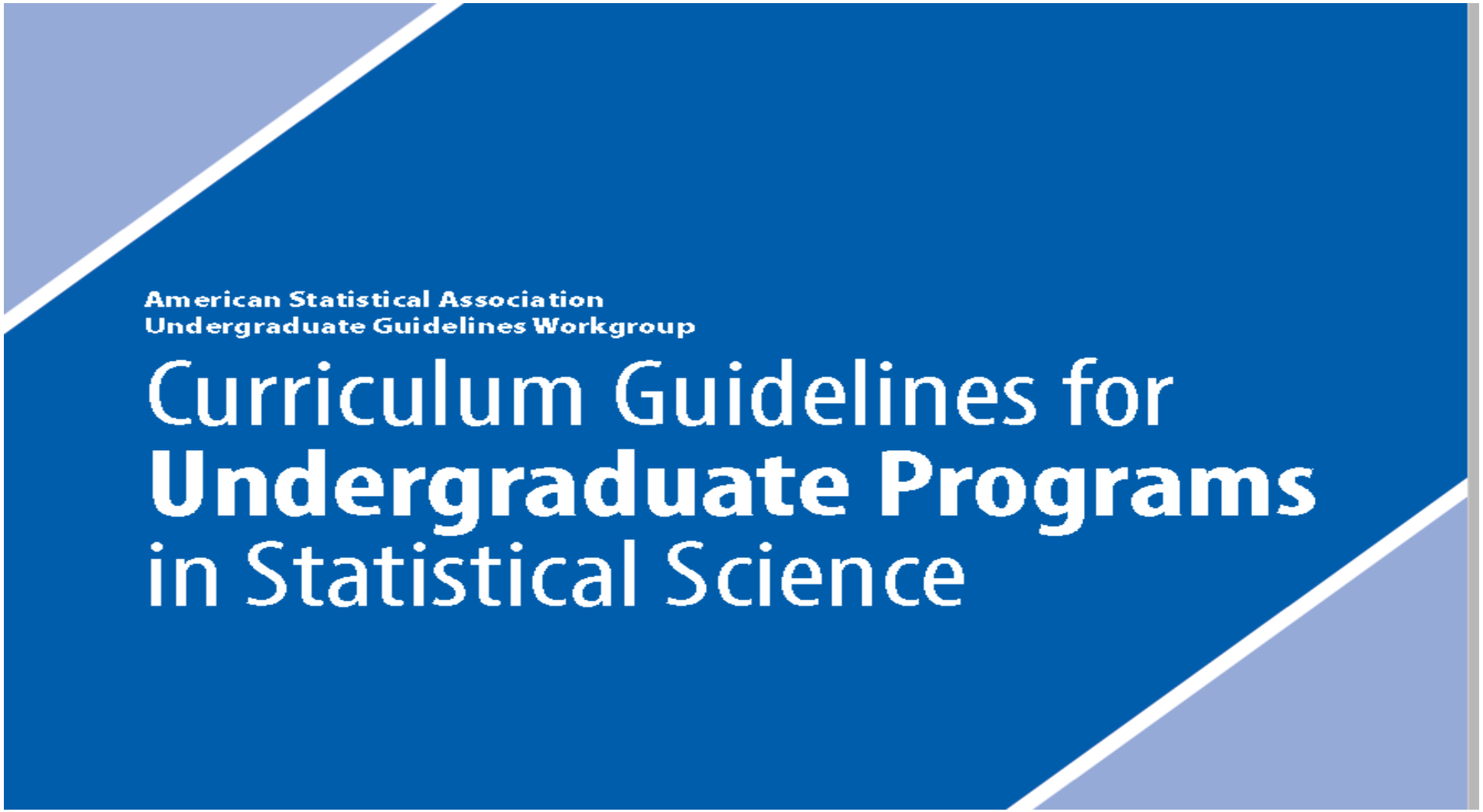
- Press Release:

  https://www.amstat.org/asa/files/pdfs/POL-ASAStatementonRoleofStatsinDataScience.pdf

- *Amstat News* Story:

  https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/

ASA Curriculum Guidelines, Nov 2014



Advice from the ASA Curriculum Guidelines

The statistical analysis process involves:

1. Formulating good questions
2. Considering whether available data are appropriate for addressing the problem
3. Choosing from a set of different tools
4. Undertaking the analyses in a reproducible manner
5. Assessing the analytic methods
6. Drawing appropriate conclusions

And, last and very important …. Next slide

Advice from the ASA Curriculum Guidelines, Last Step

The final step of the statistical analysis process:

7. Communicating results

"This scientific approach to statistical problem-solving is important for all data analysts. It needs to start in the first course and be a consistent theme in all subsequent courses."

Links Related to the ASA Curriculum Guidelines

- Report:

  https://www.amstat.org/docs/default-source/amstat-documents/guidelines2014-11-15.pdf

- Description:

  https://www.amstat.org/education/curriculum-guidelines-for-undergraduate-programs-in-statistical-science-

- *Amstat News* Story:

  https://magazine.amstat.org/blog/2015/01/01/guidelines-updated/

UC Irvine Department of Statistics

- In the Donald Bren School of Information and Computer Sciences (unusual for a statistics department)
- Three departments in the School: Statistics, Computer Science, Informatics
- Close working relationships among the departments

UCI Department of Statistics History

- Started with one person in 2002, who was given the job of creating a department.
- At the beginning, offered:
  - MS and PhD graduate degrees in Statistics
  - Undergraduate Statistics minor, but no undergraduate major
- In Fall 2015 we created a Data Science major within the department (instead of a Statistics major).
  - We had 11 faculty members. Not enough to offer a statistics major. (Now there are 13 faculty + 2 Emeriti + 5 Lecturers.)
  - Worked closely with Computer Science to create the major

Bachelor of Science (BS) in Data Science at UCI

- UCI Data Science Initiative and Major:

  Two Statistics (Stacey Hancock and Jessica Utts) and two Computer Science (Padhraic Smyth and Mike Carey) faculty developed the major proposal

datascience.uci.edu

UCI Data Science Initiative

Overview of Courses for BS in Data Science

- Dual emphasis in statistics and computer science
- Core statistics courses include:
  - Introductory statistics
  - *Exploratory data analysis and statistical computing*
  - One-year statistical methods sequence
  - One-year probability and mathematical statistics sequence
  - Bayesian statistics

Overview of Courses, continued

- Core computer science courses include:
  - Introductory and intermediate programming
  - Data structure implementation and analysis
  - Logic and discrete structures
  - Design and analysis of algorithms
  - Software engineering
  - Machine learning and data mining
  - Data management
  - Information visualization

Overview of Courses, Continued

Additional courses:

- *First year seminar in data science*
- Calculus
- Linear algebra
- Discrete math
- Critical writing on information technology
- Three elective courses from statistics or CS
- *Two quarter senior capstone project course*

Ideas Taken from the ASA Curriculum Guidelines

Some key points of focus:

1. Increased importance of data science – the ability to "think with data"
2. Real applications as a major component available early in the curriculum
3. More diverse models and approaches
4. Ability to communicate.

ASA Curriculum Guidelines: Key Point #1

- Increased importance of data science – the ability to “think with data”
  - New course in exploratory data analysis and statistical computing
  - Our entire Data Science curriculum was built with this point in mind

ASA Curriculum Guidelines: Key Point #2

- Real applications as a major component available early in the curriculum
  - First-year seminar in data science, speakers from across campus who “do” data science applications
  - Senior year two-quarter capstone project course; team taught by Statistics and CS faculty
  - Real applications throughout the curriculum, starting with the introductory statistics course.

ASA Curriculum Guidelines: Key Points #3 and #4

- More diverse models and approaches
  - Bridging computer science and statistics courses emphasizes both statistical and algorithmic models
- Ability to communicate
  - Many required courses already require student presentations, group projects, and written reports
  - Senior year two-quarter capstone project course to tie it all together
  - Required course: Critical Writing on Information Technology

ASA Curriculum Guidelines: Specific Skills

1. Statistical methods and theory
2. Data management and computation
3. Mathematical foundations
4. Statistical practice
5. Discipline-specific knowledge

We Included the ASA Guidelines on Specific Skills:

1. Statistical methods and theory
   - Core Statistics courses (methods and theory)
2. Data management and computation
   - Core computer science courses
3. Mathematical foundations
   - Calculus (through multivariable), linear algebra

ASA Guidelines Continued: More Skills

4. Statistical practice
   - Projects, presentations, reports included in many courses
   - Writing course: Critical writing on information technology
   - Senior year two-quarter capstone project course
5. Discipline-specific knowledge
   - Some experience through first-year seminar and senior capstone course
   - Would like to do more, but just not enough room in the major!

Discussion of Topics vs. Skills

What *topics* should an undergraduate program in statistics or data science cover?

vs.

What should a student who completes a major in statistics or data science be able to *do*?

Focus on Learning Objectives

- Learning objectives are "…statements of what a student is expected to know, understand and/or be able to demonstrate after completion of a process of learning." (Kennedy, 2007)
- For basic knowledge, use words such as "know" or "understand"; for skills, use active words such as "design", "apply" or "develop".
- Aim for 5-10 outcomes at the program level.

Educational Goals

- Provide students with a foundation in mathematical and statistical aspects of data analysis;
- Provide students with a foundation in the general principles of computer science;
- Teach students how to utilize their knowledge of statistical and computing principles to develop algorithms, and software for solving real-world data analysis problems;
- Provide students with practical experience in applying their knowledge of theories, methods, and tools, to a variety of data analysis problems;
- Teach students how to communicate effectively using data.

Translate Goals to Learning Objectives

Demonstrate understanding of…

1. Foundational mathematical concepts relevant to data analysis
2. Basic principles in computer science
3. Foundational statistical concepts
4. Basic principles in statistical computing.

More Learning Objectives

Demonstrate the ability to…

5. take a real-world data analysis problem, formulate a conceptual approach to the problem, match aspects of the problem to previously learned theoretical and methodological tools, break down the solution into a step-by-step approach, and implement a working solution in a modern software language
6. communicate effectively in data analysis projects.

Example of Curriculum Mapping: Match Learning Objectives to Courses

Learning Objectives

Year 4	LO1 Understand foundational mathematical principles	LO2 Understand foundational computer science principles	LO3 Understand foundational statistical principles	LO4 Understand the principles of statistical computing	LO5 Develop solutions for real-world data analysis problems	LO6 Demonstrate effective communication skills
ICS 51		x				
Stats 115			x	x	x	
Stats 170A				x	x	x
Stats 170B				x	x	x

Required Courses

Data Science vs. Statistics Major

- B.S. in Statistics planned for future in our department
  - Less computer science, more statistics
    - Possible future electives such as experimental design, time series analysis, etc.
  - Better preparation for graduate study in Statistics
    - More mathematics/theory (e.g., real analysis)
  - Include discipline-specific knowledge (e.g., 2-3 courses in applied area)
- B.S. in Data Science seen as more applied, possibly a terminal degree (no graduate degree to follow it)

Number of Data Science Majors

Enrollment in Fall Quarter Each Year

Year	2015	2016	2017	2018	2019	2020	2021
Freshman	2	6	58	13	11	29	14
Sophomore	2	12	24	40	25	25	42
Junior		9	17	24	36	38	51
Senior		2	12	29	41	65	71
Total	4	29	111	106	113	157	178

New report after we developed our major; ASA endorsed

Park City Group Report (2016): Curriculum Guidelines for Undergraduate Programs in Data Science (DeVeaux + 24 other authors)

- Data science as science
- Interdisciplinary nature
- Data at the core
- Analytical (computational and statistical) thinking and problem-solving
- (New pathways for) mathematical foundations
- Flexibility

http://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

Big Picture: What do statisticians bring to data science?

- Importance of context
- Accounting for variability
- Design, confounding, and analysis of found (observational) data
- Understanding of inference, multiplicity and reproducibility issues
- Statistical analysis cycle
- Long history of making decisions with data
- Experience working on multidisciplinary teams

These are important ideas to include in data science major!

Part 2: Ethics in Data Science Education

Outline:

- Why it is important to include ethics in data science education
- Recent examples of ethical issues in artificial intelligence (AI)
- How statisticians can help
- Teaching statistical literacy as an ethical obligation

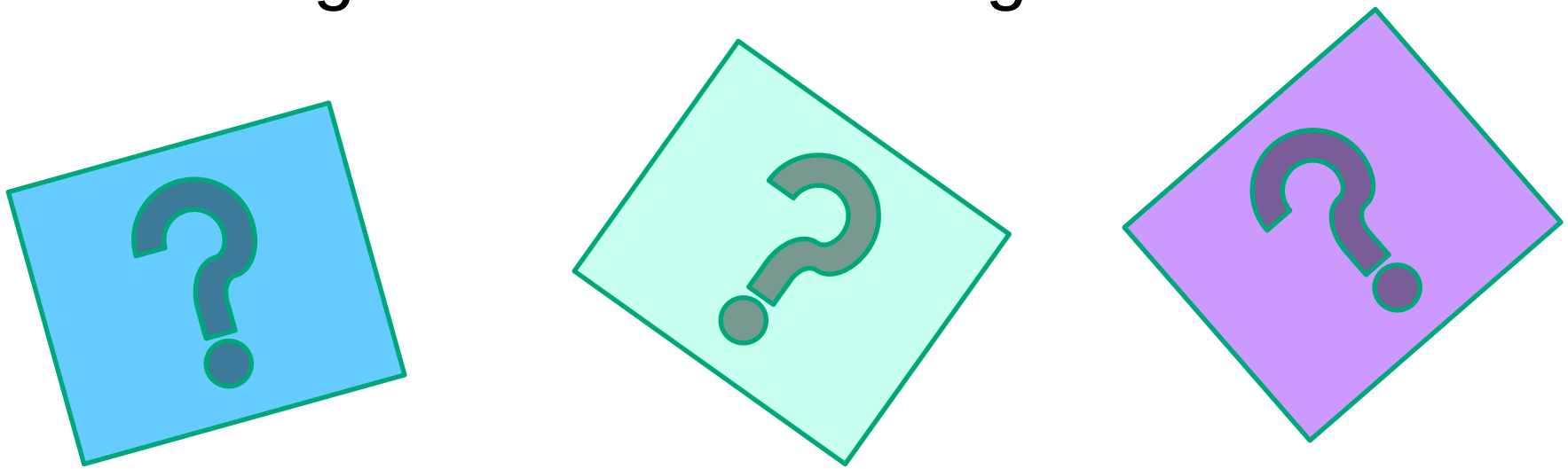
Related reference: Utts, Jessica (2021) "Enhancing Data Science Ethics Through Statistical Education and Practice," *International Statistical Review*, 89.1: 1-17.

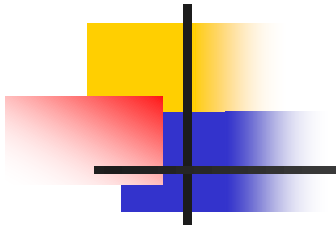
A story from my early teaching years

- An engineering professor gave students an assignment to design a pipeline to send blood from a poor developing nation to a rich developed one.
- The students got to work, discussing the optimal diameter for the pipe, how to go under a body of water, methods for keeping the blood fresh, etc.
- After letting them discuss for awhile, the professor demanded to know why not one of them had questioned the ethics of the assigned task.
- “This is a class in engineering not ethics,” was the answer the students gave.

Moral of the Story

- Train students (and practitioners) to ask *WHY* before asking *how*.
- Is the task ethical? Are there pros and cons?
- Who might benefit? Who might suffer?





Example: GPS Map program

Is it ethical…

- To clog roads by sending everyone on the same route when leaving a large event?
- To send cars through high-crime areas?
- To even identify high-crime areas?
- To send pedestrians through high-crime areas?
- To increase traffic in residential areas?
- What about school zones?

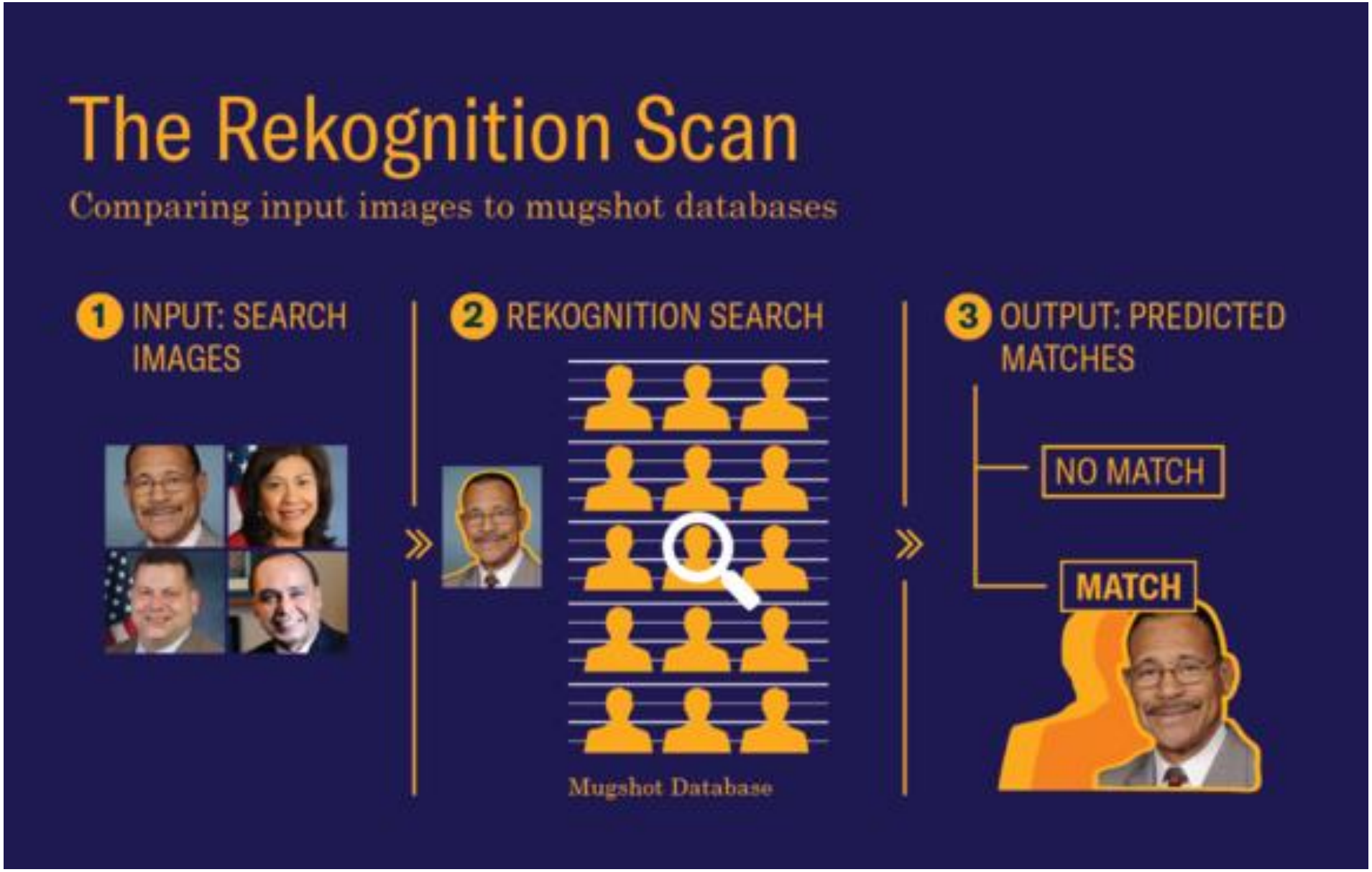
Why this topic? Why now?

- Lines are blurring for data science:

  Statistics/machine learning/artificial intelligence
- Our students' jobs reflect this cross-over
- Traditional ethical issues for statisticians
  - See for instance Ethical Guidelines from ASA, RSS, ISI
- Not enough. Complexity => new ethical issues
- Need to educate students on ethics of decisions and interpretations
  - As consumers
  - As data scientists

Example: ACLU Congress Face Recognition Study

- Used facial recognition system Amazon offers to public (Rekognition), which anyone could use.
- Running the entire test cost $12.33.
- Built a face database and search tool using 25,000 publicly available arrest photos. Then searched that database against public photos of every current member of the United States Congress.
- Used default settings Amazon sets for Rekognition.
- Found 28 (false) matches; disproportionately people of color


The Rekognition Scan
Comparing input images to mugshot databases
1 INPUT: SEARCH IMAGES
2 REKOGNITION SEARCH
Mugshot Database
3 OUTPUT: PREDICTED MATCHES
NO MATCH
MATCH

AI ethics examples: Facial recognition

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of Northern California

JULY 26, 2018 | 8:00 AM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology

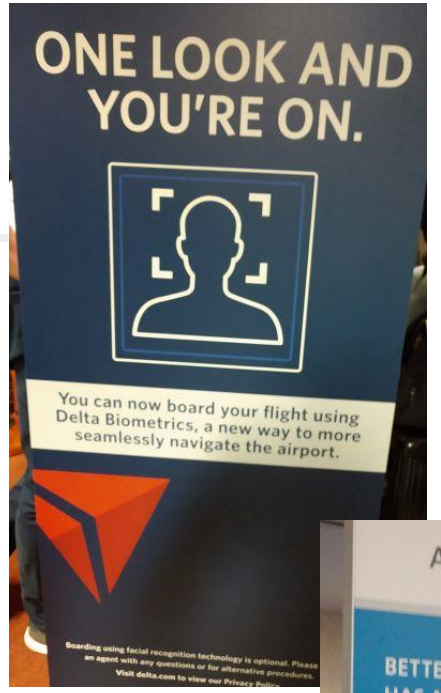
28/535 = 5.2%

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

https://aws.amazon.com/rekognition/

Facial recognition software: Benefit/risk tradeoffs

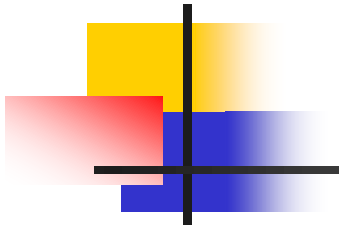
- Used for people boarding airplanes, verifying banking customers, etc.
- Used for finding celebrities in surveillance videos
- Used by police to find [possible] criminals
- Less reliable for people of color and women. Especially bad for children.
- Has resulted in false convictions, and even deaths.



Other classic AI ethics examples

- Bias in hiring algorithms, based on bias in training data.
- Algorithms used by judges to decide who is likely to commit (another) crime.
- Deciding who should get loans based on geographic and other aggregate data.
- Medical diagnostic algorithms trained on data excluding certain sub-populations…
- But, are they better or worse than humans??

Some Important AI Ethical Guidelines



- Transparency. “Black box” allure gives more credibility than justified. Ignores uncertainty.
- Consider likely biases in data sources.
- Everyone on an interdisciplinary team should take responsibility for ethical issues.
- Humans should always be involved in decisions based on algorithms.
- Correlation does not imply causation. And algorithms can suggest intervention action.

What statisticians can (uniquely?) offer

- Input on data issues, for example:
  - How to get high-quality data
  - How to assess bias in data
  - How conclusions depend on data sources
- Input on analysis issues, for example:
  - Intelligent use of modeling; considering assumptions
  - Multi-variable thinking; but pitfalls of multiple analyses
  - Dealing with outliers
- Input on reporting results, for example:
  - Practical vs statistical significance; what p-values mean
  - When cause and effect conclusions can be made

General ethical concerns for statisticians

- Ethics in data collection, quality and uses
- Ethical implementation of details in a study
- Issues of ethics during the analysis
- Ethics of reporting results
  - To clients
  - To the media and the public
- Teaching statistical literacy in all introductory statistics courses is an ethical obligation.

Ethics of data issues; questions to ask

- Informed consent for all uses of data and/or all interventions?
- Individual anonymity, and likely to remain so when merged with other datasets?
- Possible biases in the data? Sub-populations under-represented?
- Missing data or drop outs for reasons related to the research questions?

Examples of ethics in implementing a study

- Ecological validity of intervention – are conditions in the study the same as "real world"?
  - Example: Time of day electric use study
- Ethics of interventions without consent
  - Cornell/Facebook emotion study (more next)
- Power analysis to make sure a large enough sample is used to detect a meaningful effect
  - Consider power for sub-groups too.

Facebook/Cornell Emotion Study

- 2012 study, randomly selected 689,003 Facebook users, assigned to 4 groups.
- No informed consent!
- One group had negative news feed reduced; another had positive news feed reduced. Control groups had news feed randomly omitted. Study lasted one week.
- Use of negative and positive words used in subjects' own posts were measured.

Results from Cornell press release

"News feed: Emotional contagion sweeps Facebook"

- "People who had positive content experimentally reduced on their Facebook news feed for one week used more negative words in their status."
- "When news feed negativity was reduced the opposite pattern occurred. Significantly more positive words were used in peoples' status updates."

https://news.cornell.edu/stories/2014/06/news-feed-emotional-contagion-sweeps-facebook

More about the Emotion study

- Published in the *Proceedings of the National Academy of Sciences*, June 2014
- According to Altmetric data it got lots of attention:
  - Mentioned by 337 news outlets
  - 136 blogs
  - 4164 tweeters
  - “In top 5% of all research outputs scored by Altmetric”

BUT, the actual results…

"When positive posts were reduced in the News Feed, the percentage of positive words in people's status updates decreased by 0.1% compared with control [t(310,044) = −5.63, P < 0.001, Cohen's d = 0.02], whereas the percentage of words that were negative increased by 0.04% (t = 2.71, P = 0.007, d = 0.001)."

Results, continued…

"Conversely, when negative posts were reduced, the percent of words that were negative decreased by 0.07% [t(310,541) = −5.51, P < 0.001, d = 0.02] and the percentage of words that were positive, conversely, increased by 0.06% (t = 2.19, P < 0.003, d = 0.008)."

Control
Experimental
Positive Words (per cent)
5.0
5.1
5.2
5.3
5.4
Negativity Reduced
Positivity Reduced
Negative Words (per cent)
1.50
1.60
1.70
1.80

Authors' justification…

- "The effect sizes from the manipulations are small (as small as d = 0.001).
- Given the massive scale of social networks such as Facebook, even small effects can have large aggregated consequences.
- And after all, an effect size of d = 0.001 at Facebook's scale is not negligible: In early 2013, this would have corresponded to hundreds of thousands of emotion expressions in status updates per day."

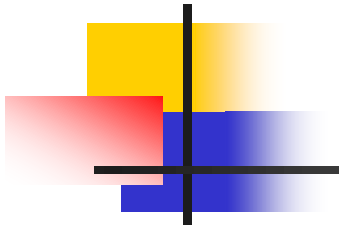
Ethical Issues from this Study

- No informed consent.
- Misleading graphs.
- Confusion of statistical significance with practical significance (importance)
- Justification of small effect size as being of practical importance because of large population affected.

Examples of Ethics in Analysis

- Multiple tests, type 1 errors, p-hacking, "questionable research practices"
- Collinearity and (mis)interpretation of individual regression coefficients
- Hidden, unrealistic Bayesian priors
- Ignoring data not missing at random
  - Ex: Dropping out of drug trial due to side effects

Ethics of Reporting Results



- Focus on magnitude, not p-values.
  - With big data, small effects have tiny p-values
- Include clear explanation of uncertainty.
- Don’t overstate the importance of results.
- Graphics should be clear, not misleading.
- Media coverage should include all relevant results, not just most interesting or surprising.
- Don’t imply causal connection if not justified.

Example: Reporting to client & media

- Suppose a client asks you to evaluate an online game for boosting children's math skills.
- Data provided include pre-post math and language scores, time spent studying each one.
- Results: Math scores went up but language scores went down, and game was addictive.
- Are you ethically bound to report the negative consequences of using the game…
  - To the client?
  - In media requests?

Example of misleading reporting: Hormone replacement therapy

- Women's Health Initiative, randomized study comparing hormones with placebo.
- Surprising result was *increase* in risk of coronary heart disease in hormone group.
- Trial was stopped early, and millions of women were advised to stop taking HRT (hormone replacement therapy) immediately.
- Large scale media attention on risks of heart disease and breast cancer from HRT.

But… Results from the original article

"Absolute excess risks per 10,000 person-years attributable to [hormones] were 7 more CHD [coronary heart disease] events, 8 more strokes, 8 more PEs [pulmonary embolism], 8 more invasive breast cancers, while absolute risk reductions per 10,000 person-years were 6 fewer colorectal cancers and 5 fewer hip fractures."

More results..

- 231 out of 8506 = 2.72% of women taking hormones died of any cause during the study.
- 218 of the 8102 = 2.69% of women taking placebo died of any cause during the study.
- Adjusted for the time spent in the study, the death rate was slightly lower in the hormone group, with an annualized rate of 0.52% compared with 0.53% in the placebo group.

Ethical issue for reporting these results

- The media and medical community focused on the surprising heart disease results
- In fact the hormone group fared *better* in many ways, including adjusted death rate.
- Were millions of women misled?
- If full results had been reported in the media, women could decide for themselves, for instance based on family or personal medical history.

Ethics and statistical literacy in education

- For training statisticians:
  - Include ethical considerations throughout their training.
  - Idea: Ask for a discussion of ethical issues as part of all data analysis projects, possibly dissertations as well.
- For educating all students:
  - Statistical literacy involves recognizing ethical issues.
  - Emphasize topics students can use in their lives, to help them make informed decisions and recognize statistical errors.

Example: Confusion of the inverse
$P(A|B) \neq P(B|A)$

- P(positive test | disease) > P(disease | positive test)

  especially for rare disease (doctors get this wrong)

- Prosecutor's fallacy in courtroom:

  P(innocent | evidence) vs P(evidence | innocent)

- COVID Example: UK, Delta variant (early data)

  P(Double Vaccinated | Death) ≈ 0.43 = 50/117

  But P(Death | Double Vaccinated) is still very, very low!

If everyone was vaccinated, P(Vaccinated | Death) = 1.0!

Suggestions for Statistics Educators

- Train all students to think about ethics:
  - Propose and discuss ethical issues in class
  - Include ethics section in all data analysis assignments
  - Ask *why* before *how* for all assignments
- For non-statistics students, teach literacy issues such as:
  - Recognizing multiple analyses and selective reporting
  - Trade off in risks
  - Trade offs in society benefits versus personal rights
  - Confusion of the inverse

Preparing Statistics Majors for Jobs

- It's not true that numbers are just numbers – learning subject matter knowledge is crucial!
- Often the statistician will be the most objective member of a multi-disciplinary team.
  - Train students to speak up as a member of a team.
- Ask questions!
  - Including "why" before "how"
  - Consider ethical implications throughout the cycle of data collection, analysis, reporting

Summary and Conclusions

- Statisticians have a major role to play in data science ethics.
- Need to speak up as a member of an interdisciplinary team.
- Teach our statistics students ethics alongside technical issues.
- Teach all students to identify ethical issues and mistakes in reports based on statistical studies.

THANK YOU

Question/Answer?

(Move to additional literacy topics if there is time.)

Contact info:

jutts@uci.edu

http://www.ics.uci.edu/~jutts

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

My Top 10 Important Literacy Topics

1. Observational studies, confounding, causation
2. The problem of multiple testing
3. Sample size and statistical significance
4. Why many studies fail to replicate
5. Does decreasing risk actually increase risk?
6. Personalized risk versus average risk
7. Poor intuition about probability and risk
8. Using expected values to make decisions
9. Surveys and polls – good and not so good
10. Confirmation bias

Observational studies, confounding, causation

Example headline from observational study:

"Breakfast Cereals Prevent Overweight in Children"

Worldhealthme.com, 4/12/13

The article continues:

> *"Regularly eating cereal for breakfast is tied to healthy weight for kids, according to a new study that endorses making breakfast cereal accessible to low-income kids to help fight childhood obesity."*

Some Details

- Observational study
- 1024 children, only 411 with usable data
  - Mostly low-income Hispanic children in Texas, USA
  - Control group for a larger study on diabetes
- Asked what foods they ate for 3 days, in each of 3 years (same children for 3 years) looked at number of days they ate cereal = 0 to 3 each year.
- Lead author = Vice President of Dairy MAX, a regional dairy council

More Details: The analysis

- Multiple regression was used
  - Response variable = BMI percentile each year (BMI = body mass index)
  - Explanatory variable = days of eating cereal in each year (0 to 3), modeled as linear relationship with BMI!
- Did not differentiate between other breakfast or no breakfast (if not cereal)
- Also adjusted for age, sex, ethnicity and some nutritional variables

Some problems

- Observational study – no cause/effect.
- Obvious possible confounding variable is general quality of nutrition in the home
  - Unhealthy eating for breakfast (non-cereal breakfast or no breakfast), probably unhealthy for other meals too.
- Possible reversed cause/effect: High metabolism could cause low weight and the need to eat breakfast. Those with high metabolism require more frequent meals.

More of my favorite (!) headlines

- 6 cups a day? Coffee lovers less likely to die, study finds *NBC News website*, 5/16/12
- Joining a Choir Boosts Immunity *Woman's World*, 6/27/16
- Walk faster and you just might live longer *NBC News website*, 1/4/11
  - Researchers find that walking speed can help predict longevity
  - The numbers were especially accurate for those older than 75

Poor intuition about probability & risk

- William James was first to suggest that we have an *intuitive* mind and an *analytical* mind, and that they process information differently.
- Example: People feel safer driving than flying, when probability suggests otherwise.
- Psychologists have studied many ways in which we have poor intuition about probability assessments.
  - Recommended reading: *Thinking, Fast and Slow* by Daniel Kahneman

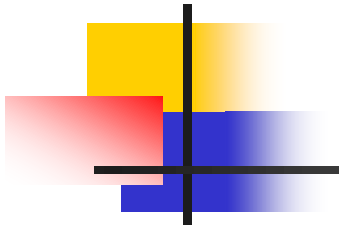
Which do you think is more probable?

- A massive flood somewhere in North America next year, in which more than 1,000 people drown.
- An earthquake in California sometime next year, causing a dam to burst resulting in a flood in which more than 1,000 people drown.

The Representativeness Heuristic and the Conjunction Fallacy

- **Representativeness heuristic:** People assign higher probabilities than they should to scenarios that are *representative* of how they *imagine* things would happen.
- This leads to the **conjunction fallacy** … when detailed scenarios involving the conjunction of events are given, people assign *higher* probability assessments to the *combined event* than to statements of one of the simple events alone.
- But *P(A* and *B) = can't exceed P*(*A*)

Confusion of the inverse: DNA Example



- Dan's DNA matches DNA at a crime scene. Only 1 in a million people have this specific DNA.
- There are 6 million people in the local area, so about 6 have this DNA.
- Is Dan almost surely guilty?

Is Dan Guilty??

Let's look at hypothetical 6 million people.

Only 6 have a DNA match

	Guilty	Innocent	Total
DNA match	1	5	6
No match	0	5,999,994	5,999,994
Total	1	5,999,999	6,000,000

How to illustrate conditional probabilities

- P(DNA match | Dan is innocent)

  ≈ 5 out of almost 6 million, extremely low!

  - *Prosecutor would emphasize this (wants to show Dan is guilty)*

	Guilty	Innocent	Total
DNA match	1	5	6
No match	0	5,999,994	5,999,994
Total	1	5,999,999	6,000,000

DNA Example, continued

- But… P(Dan is innocent | DNA match)

≈ 5 out of 6, fairly high!

- *Defense lawyer would emphasize this (wants to show Dan is not guilty)*

	Guilty	Innocent	Total
DNA match	1	5	6
No match	0	5,999,994	5,999,994
Total	1	5,999,999	6,000,000

Prosecutor's Fallacy

- P(DNA match | innocent) = 5/5,999,999 *very low*

  P(match evidence | innocent)

  Prosecutor would emphasize this

- P(innocent | DNA match) = 5/6 *high*

  P(innocent | match evidence)

  Defense lawyer would emphasize this

- Jury needs to understand this difference!

More general version

Example: Facial recognition data finds 6 people who match video of a criminal from crime scene

	Guilty	Innocent	Total
Match rare evidence	1	5	6
No match	0	5,999,994	5,999,994
Total	1	5,999,999	6,000,000

- P(Innocent | evidence match) = 5/6
- P(Evidence match | innocent) = 5/5,999,999

My Top 10 Important Literacy Topics

1. Observational studies, confounding, causation
2. The problem of multiple testing
3. Sample size and statistical significance
4. Why many studies fail to replicate
5. Does decreasing risk actually increase risk?
6. Personalized risk versus average risk
7. Poor intuition about probability and risk
8. Using expected values to make decisions
9. Surveys and polls – good and not so good
10. Confirmation bias

THANK YOU

Contact info:

jutts@uci.edu

http://www.ics.uci.edu/~jutts

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE