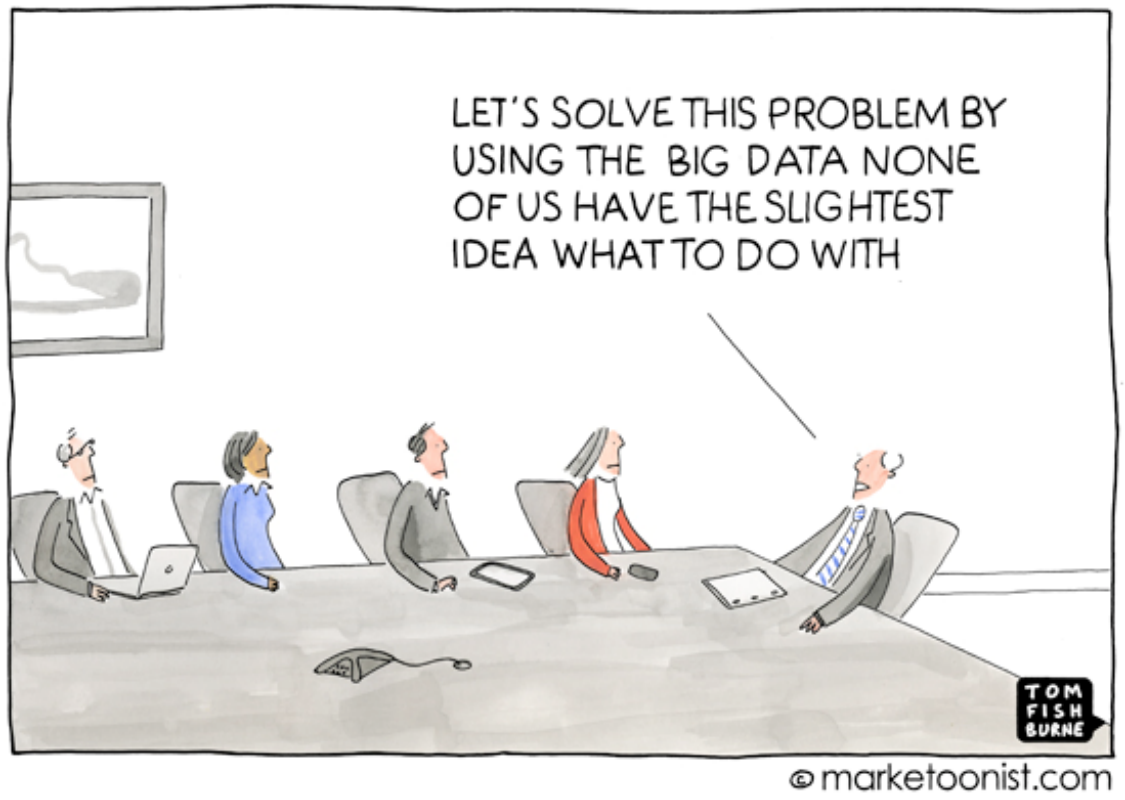
What is data science and how can it help my research?

Lance A. Waller, Ph.D.
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
lwaller@emory.edu
March 2021

How common is this?



https://marketoonist.com/wp-content/uploads/2014/01/140113.bigdata.jpg

What do these mean to you?

- Data Science
- Big data
- Cloud computing
- Informatics
- Reproducible Research

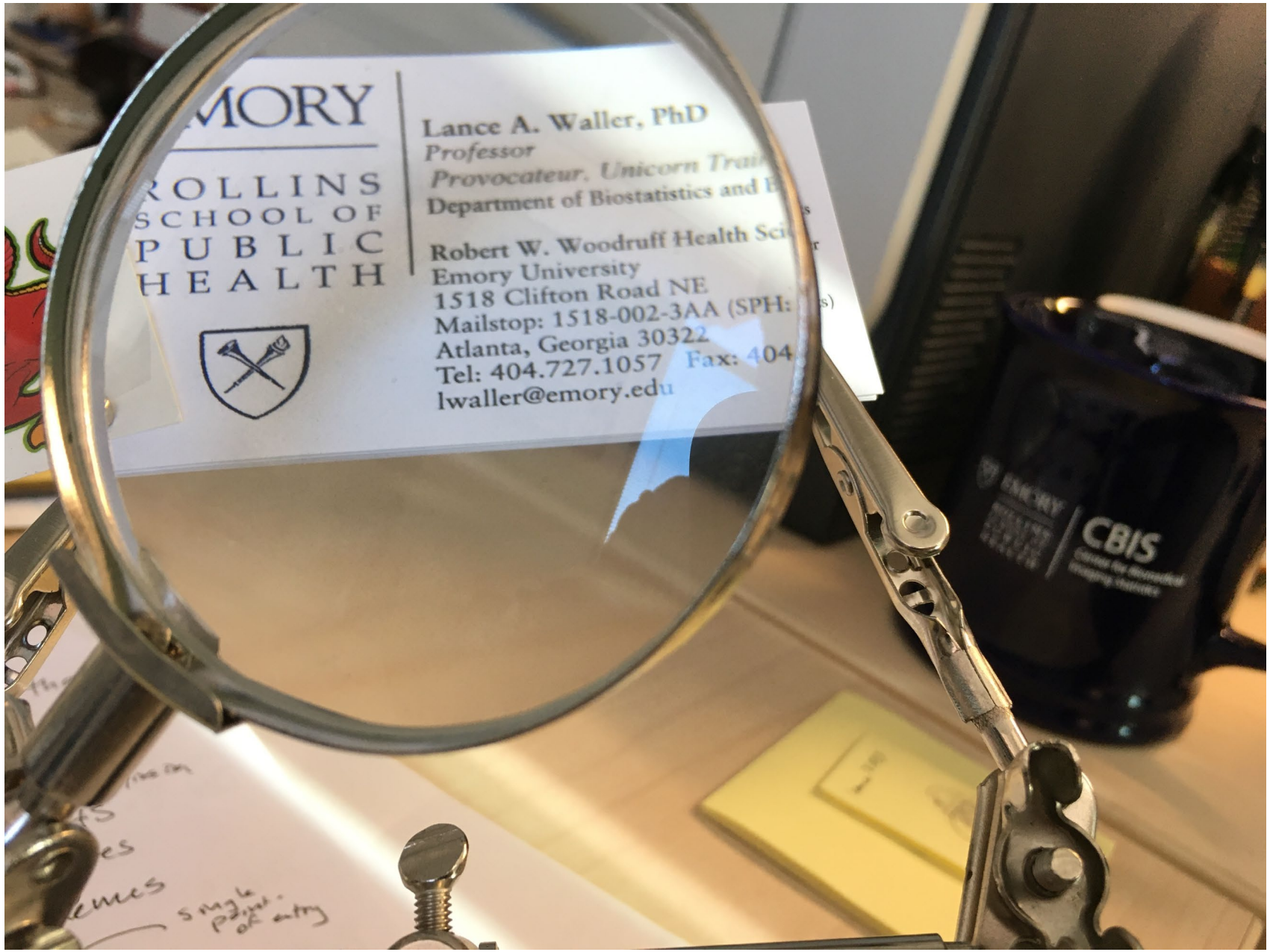
Big Data

- Big data not from one project, one lab, one investigator.
- Big data gets bigger by linking to data from other projects, other labs, other investigators.
- NIH has public release requirements on funded research.
- How can you use other people's data or other people's computers?

- **Big data are more data than you know what to do with.**
- **Data are always greener if someone else collects them.**

Data Science

- “A data scientist is a data analyst who lives in San Francisco.”
- “Data Science is statistics on a Mac.”
- “A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.”
- http://priceonomics.com/whats-the-difference-between-data-science-and/


MORY
ROLLINS
SCHOOL OF
PUBLIC
HEALTH
Lance A. Waller, PhD
Professor
Provocateur, Unicorn Trai
Department of Biostatistics and
Robert W. Woodruff Health Sci
Emory University
1518 Clifton Road NE
Mailstop: 1518-002-3AA (SPH:
Atlanta, Georgia 30322
Tel: 404.727.1057 Fax: 404
lwaller@emory.edu
CBIS

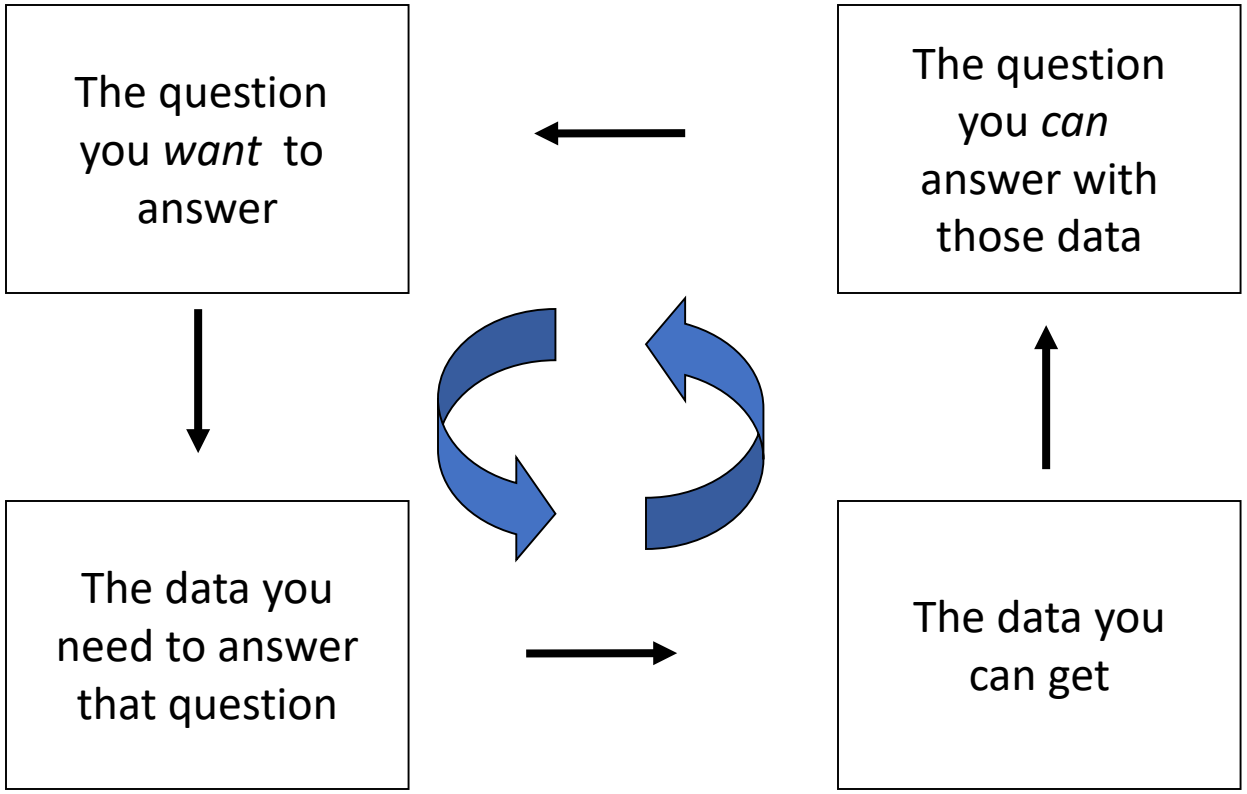
Lance Waller, Provocateur

- June 2019, “Provocateur” for an NIH-sponsored Innovation Lab on Data Science Challenges in Rural Health and Environmental Exposures
- https://clic-ctsa.org/events/2019-innovation-lab-data-science-challenges-rural-health-and-environmental-exposures

- I presented a series of provocative statements regarding data science collaborations.
- This took some training since…

The last time I was provocative…May 1983 (but after some training…here are a few)

1. Framework: Whirling vortex of analysis



Original source: John Richardson, Toxicologist EPA Region IV

2. Data Science:
Statistics is dead, long live statistics

- Data science is not just statistics.
- Data science is not just computer science.
- Data science is not just informatics.

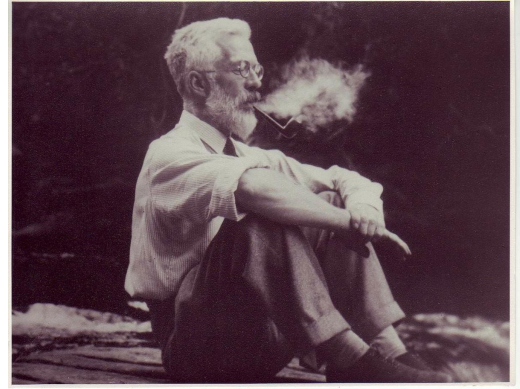
- Donoho (2017) 50 years of data science. *Journal of Computational and Graphical Statistics* **26**, 745-766.

History of statistics

- 1700s-1800s: Probability to understand gambling
- 1920s-1930s: Experimental design:
  - Collecting data to understand processes.

- Data were expensive.
- Careful design yields most information from data you collect.

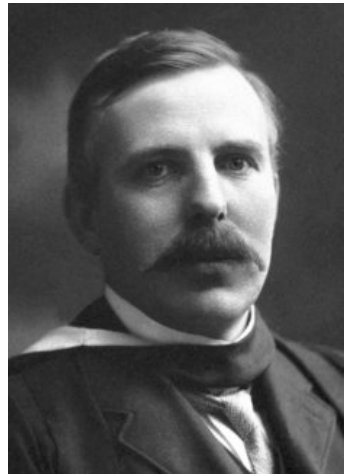
Design-centric thinking



- “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.” R.A. Fisher

https://priceonomics.com/why-the-father-of-modern-statistics-didnt-believe/

- "If your experiment needs statistics, you ought to have done a better experiment." Ernest Rutherford



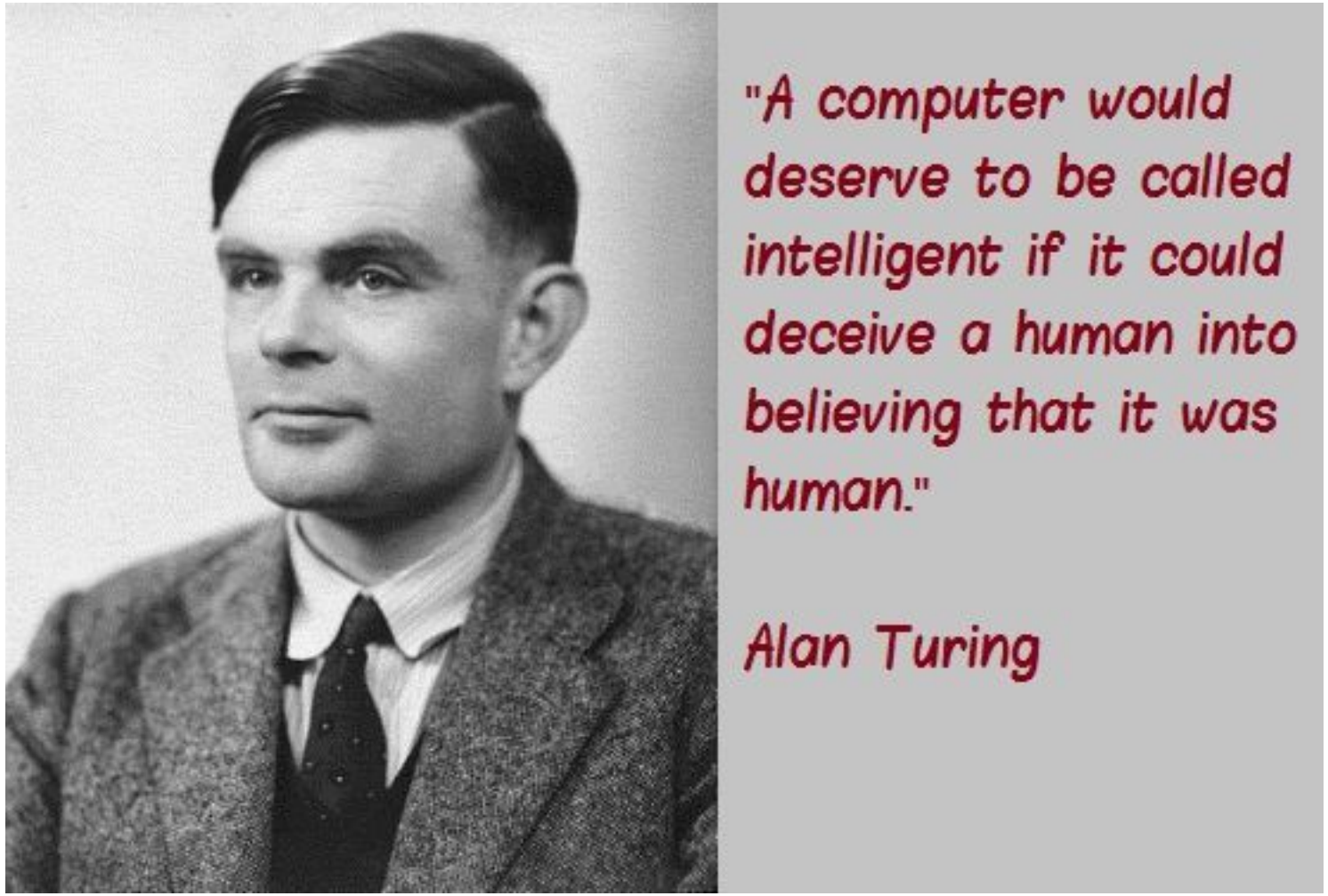
https://www.nobelprize.org/prizes/chemistry/1908/rutherford/biographical/

Contrast with the setting of Data Science

- Data are plentiful.
- Perhaps not exactly what you need, but there is a lot of it.
- And a lot of types.
- And computers are fast.

- Can’t we do something with ths?

Alan Turing Quote:



Artificial Intelligence

- Complex, human-like decisions
  - Based on a LOT of small decisions.
  - Based on a LOT of data.
  - Done quickly.
  - And faster.
  - And better?

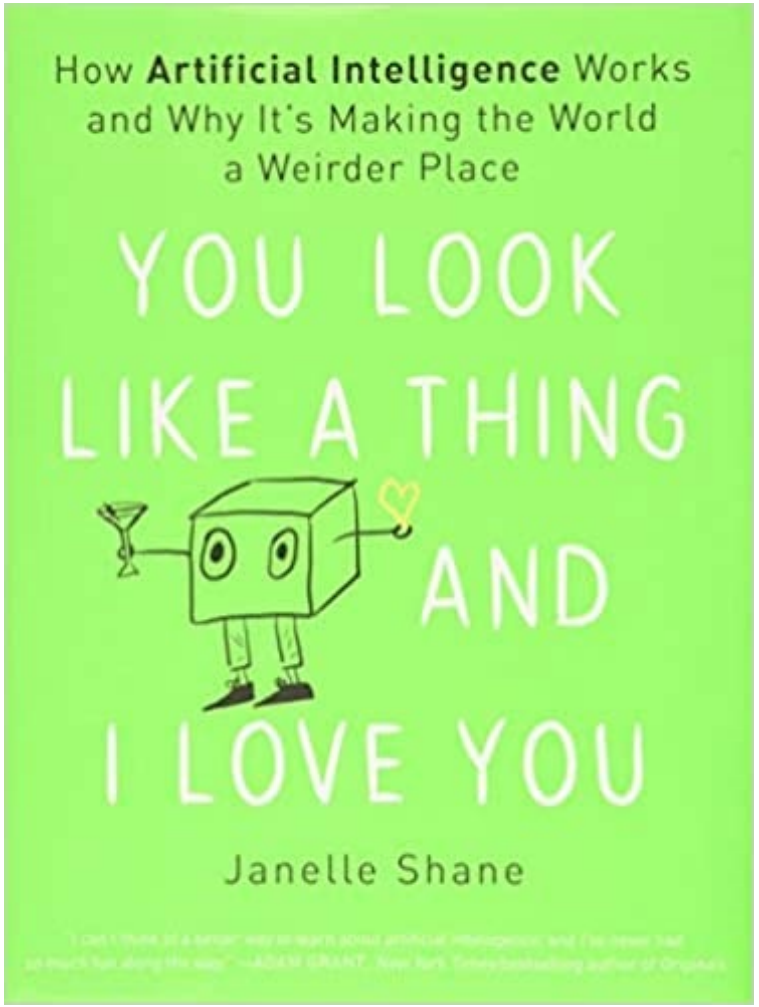
- A lot of what we hear about AI now, relates to *machine learning*.

Data Science, Machine Learning, AI

- Machine Learning
  - Split data: training and validation sets
  - Assess classification of training data, apply to validation set.
- Big successes:
  - Image classification
  - Book, video recommendations based on past selections
  - Language translation
- Machine learning, deep learning, AI largely driven by a competition mindset…

Data Science, Machine Learning, AI

- For an entertaining overview: Janelle Shane's *You Look Like a Thing and I Love You* and her blog https://aiweirdness.com/
- AI generated
  - Knock-knock jokes
  - Recipes
  - Pick-up lines
- Lots of thoughtful explanations and examples.



How does fast, accurate classification help?

- Image classification, automated radiology
- Natural language processing and electronic health record notes
- Classify patients (phenotypes), syndromes
- Recruiting patients across:
  - A health system
  - Different health systems
  - International networks
- Prediction

Classification to creation

- If you can classify fast enough and with enough data you can create:
  - Automated help desks
  - Robocalls
  - Detailed directions
  - Estimated travel time
  - Updated directions based on current traffic conditions.
  - Self-driving cars
  - Realistic deep fake photos/videos
  - Short news articles
  - Pop songs
- What are the potential mistakes and what are the consequences?
- What are the potential mistakes in healthcare and biomedical research?

Generations of science

- First generation: Doing something *because you can*.
- Second generation: Doing something *how you should*.

- How do we decide what we *should* do?
- How do we find collaborators?

Pause for feedback

- What comments do you have?
- What questions do you have?

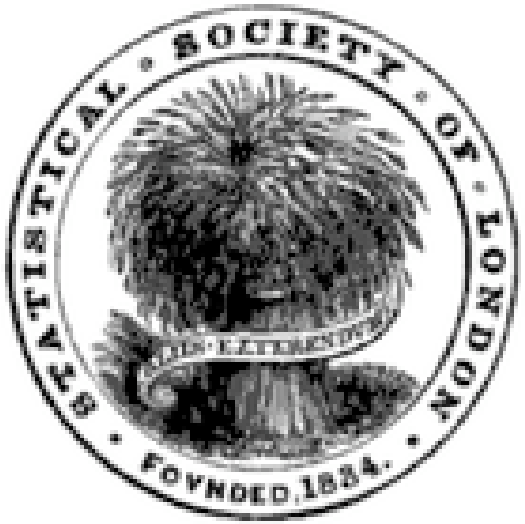
- What does all of this have to do with your work and where can you find help?

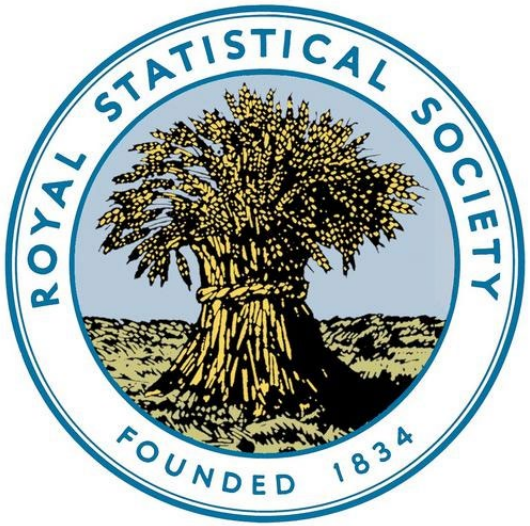
Mindset: It's data *science*, not data alchemy

- Data Science:
  - Documented
  - Citable
  - *Reproducible*
- Provide and document contributions to science:
  - *Citable* peer-reviewed publications (DOI)
  - *Citable* curated data sets (DOI)
    - e.g., Dryad repository, https://datadryad.org/stash
  - *Citable* peer-reviewed data descriptors (DOI)
    - e.g., Nature Publishing Group's *Scientific Data*), https://www.nature.com/sdata/
  - *Citable* code
    - R, Python, R markdown, Jupyter notebooks, GitHub
- Waller (2018, *American Statistician*)

Aliis exterendum: Let others thrash it out

- The motto chosen for the Statistical Society of London in 1834 was *Aliis exterendum:* 'Let others thrash it out.'
- "Their aim was thus simply to gather the facts, leaving it to others to draw whatever conclusions might be warranted." Hilts (1978)

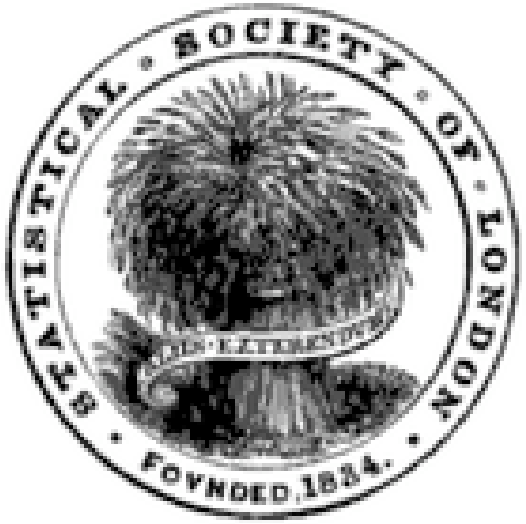




Aliis exterendum: To avoid this remember…

- Formulae are facts, not findings, but
- Calculations have consequences.

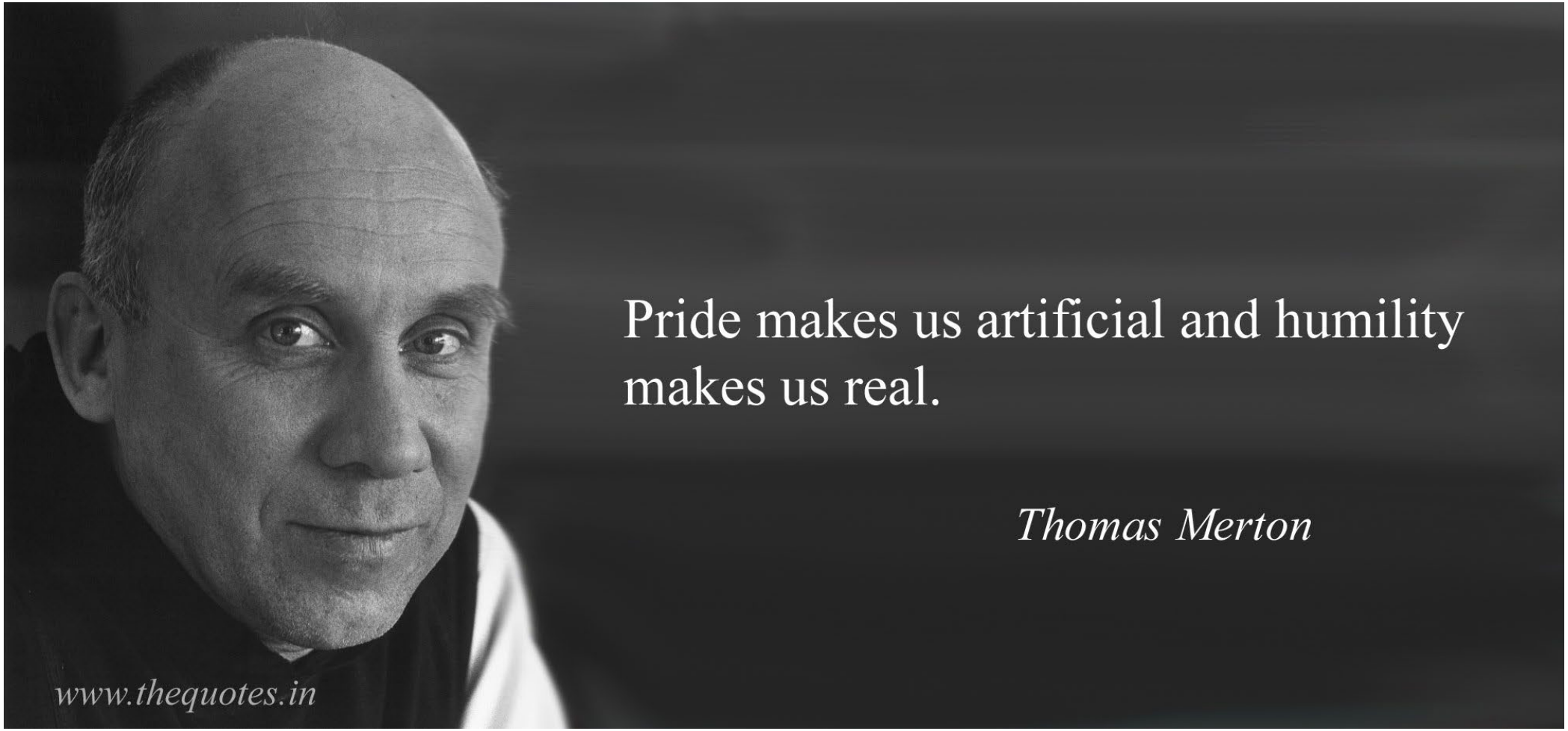
- What question did you answer?
- Under what circumstances: Win a competition or solve a challenge?



Collaboration: Who are you trying to impress?

- Collaboration and international travel
- Hubris and Humility
- **Good analyses are technical marvels, great analyses include humility**
- Speak truth to power, but who speaks truth to your power?

Thomas Merton Quote:



Ethics

- Another consideration of the difference between:
  - Something you *can* do.
  - Something you *should* do.

- What have you read, what do you think about ethics and data science?

Ethics in Biomedical Research

- The Nuremberg Code (1947): **INFORMED CONSENT**
  - http://history.nih.gov/research/downloads/nuremberg.pdf
  - "the voluntary consent of the human subject is absolutely essential"
- Declaration of Helsinki (1964): **INSTITUTIONAL REVIEW BOARD (IRB)**
  - http://history.nih.gov/research/downloads/helsinki.pdf
  - “research protocols should be reviewed by an independent committee prior to initiation”
  - “research with humans should be based on results from laboratory animals and experimentation”
- The Belmont Report (1979): **BENEFICENCE**
  - http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html
  - Respect for persons, beneficence (do not harm, maximize benefit, minimize possible harm), justice.

Ian Malcolm, *Jurassic Park*, Quote:

- “If I may... Um, I'll tell you the problem with the scientific power that you're using here, it didn't require any discipline to attain it. You read what others had done and you took the next step. You didn't earn the knowledge for yourselves, so you don't take any responsibility for it. You stood on the shoulders of geniuses to accomplish something as fast as you could, and before you even knew what you had, you patented it, and packaged it, and slapped it on a plastic lunchbox, and now you're selling it, you wanna sell it.”

American Statistical Assoc. Ethical Guidelines

- http://www.amstat.org/about/ethicalguidelines.cfm
- Recommended in 1949, approved in 1999, latest approval now.
- Eight Principles:
  - Professional Integrity and Accountability
  - Integrity of Data and Methods
  - Responsibilities to Science/Public/Funder/Client
  - Responsibilities to Research Subjects
  - Responsibilities to Research Team Colleagues
  - Responsibilities to Other Statisticians or Statistics Practitioners
  - Responsibilities Regarding Allegations of Misconduct
  - Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients

Data Science Ethics

- A lot written about:
  - Bias in facial recognition
  - Privacy and confidentiality
- Important but is that all there is?
- Guidelines for Ethical Statistical Practice are about decisions *people* make.
- Some discussions of Ethical Practice in Data Science related to decisions *algorithms* make.
- What's the difference?

Privacy and Confidentiality

- Privacy: control over the extent, timing, and circumstances of sharing oneself with others.
  - **Privacy is about *your control over your data*.**
- Confidentiality: *treatment of information* disclosed in a relationship of trust and with the expectation that it will not be shared without permission.
  - **Confidentiality is *managing data about others*.**
- Health Insurance Portability and Accountability Act (HIPAA) of 1996.

- http://www.research.uci.edu/compliance/human-research-protections/researchers/privacy-and-confidentiality.html

Deidentified data…

- Can anything be completely deidentified?
  - Connected data
  - Linkage algorithms
  - Roche et al. (2019,*Nature Communications*) claim:
    - "99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes"
  - https://www.nature.com/articles/s41467-019-10933-3.pdf

- Example: GPS locations collected by apps
  - NY Times: "Your apps know where you were last night, and they are not keeping it secret"
  - https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

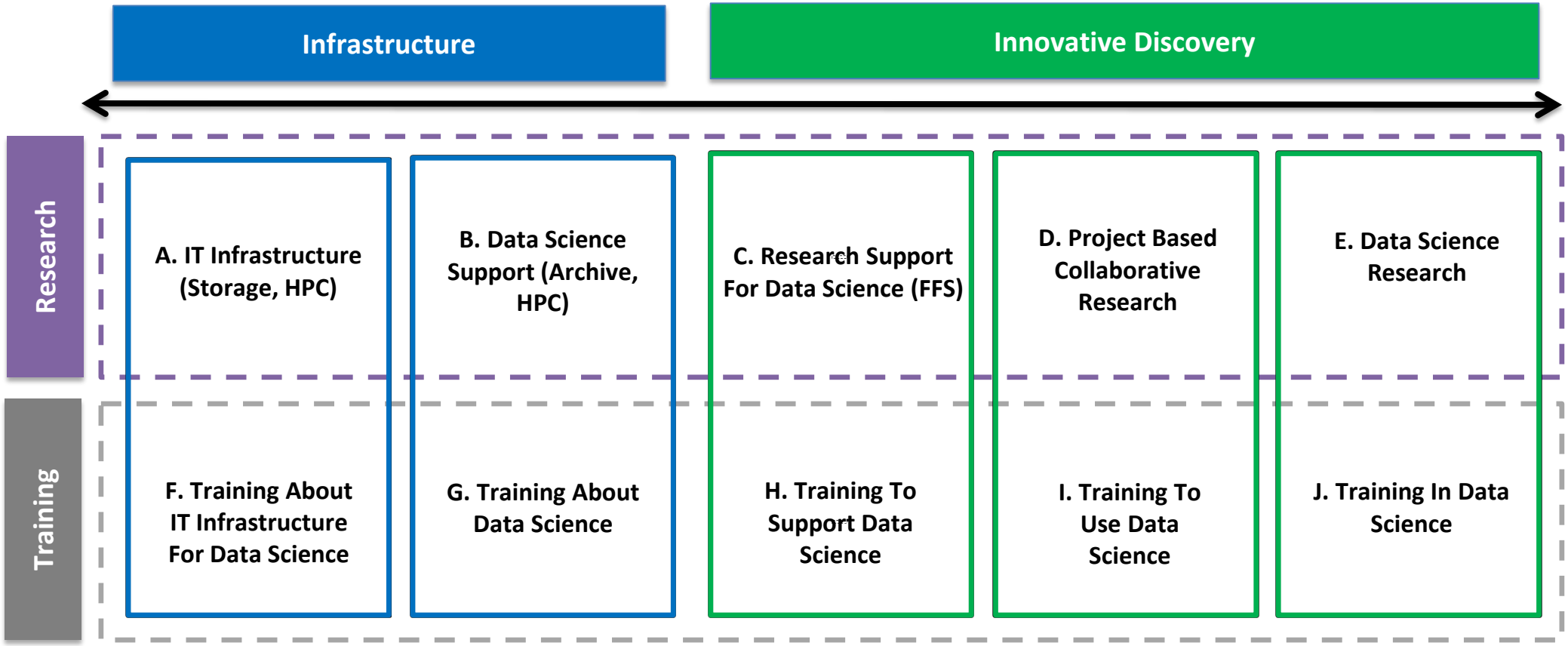
This two-day virtual public workshop will discuss the challenges and opportunities for researchers, institutions, and other stakeholders to establish effective data management and sharing practices. The workshop will also provide a neutral venue for stakeholders across sectors to discuss possible data management and sharing implementation strategies and researcher needs in light of the final NIH Policy on Data Management and Sharing (effective January 2023).

For more information about this workshop, including the agenda, please visit the workshop webpage.

Data Science in WHSC at Emory

- What is going on, how can we find collaborators?

Innovation, Incubation, Integration, Investigation: The Data Science Continuum



DSI Executive Committee Membership

Lance Waller, PhD
Rollins School of Public Health

Madhu Behera, PhD
Winship Cancer Institute

Gari Clifford, DPhil
School of Medicine

Adam Ericsen, PhD
Yerkes National Primate Research Center

Vicki Hertzberg, PhD
Nell Hodgson Woodruff School of Nursing

Nabile Safdar, MD
Emory Healthcare

Data Science @ Emory

- Departments
  - BIOS, BMI, CS, Math, QTM, SON
  - Radiology, Winship, RSPH, Yerkes
- Training Programs
  - About Data Science: BIOS (PhD, MSPH, MPH), DATA (MSPH), CSI (PhD)
  - CTSA/MSCR training in informatics
  - Using Data Science: GDBBS, EPI, EHS, GH, HPM, BSHES,…, everyone
- Student/Postdoc/Researcher Groups
  - Data Science for Scientists, etc.
  - Rigor and Reproducibility webinars

Opportunities to engage

- DSI Executive Cmt
- CTSA Studio consults with BERD or Informatics Cores
- Biostatistics Collaboration Core
- Emory's Rigor and Reproducibility webinar series
  - https://guides.libraries.emory.edu/rigor-rep
  - Past topics (recordings and slides available)
    - **Data Publication and Citation: How do I get credit for promotion?**
    - **Making Reproducibility Practical: Using R and R Markdown**
    - **Rigor and Reproducibility at eLife**
    - **An Introduction to Open Science Principles and Tools**
    - Organizer: Jeremy Kupsco, Emory Libraries, jkupsco@emory.edu
- Data Science for Scientists ATL
  - https://emory.campuslabs.com/engage/organization/data-science-for-scientists-atl
  - https://data-science-for-scientists-atl.github.io/

Summary

- Whirling vortex: Questions, data, methods, answers
- *Aliis exterendum*
- Hubris and humility
- Ethical Guidelines for Statistics
- Ethical Guidelines for Data Science
- Emory Resources

In conclusion:

- "…Nature's dice are always loaded…in her heaps and rubbish are concealed sure and useful results." Ralph Waldo Emerson, *Nature*

https://www.thoughtco.com/ralph-waldo-emerson-1773667

Questions/Comments/Suggestions?