STATS 604
Lecture 1

Prediction, estimation, data science, and statistics.

Jonathan Terhorst

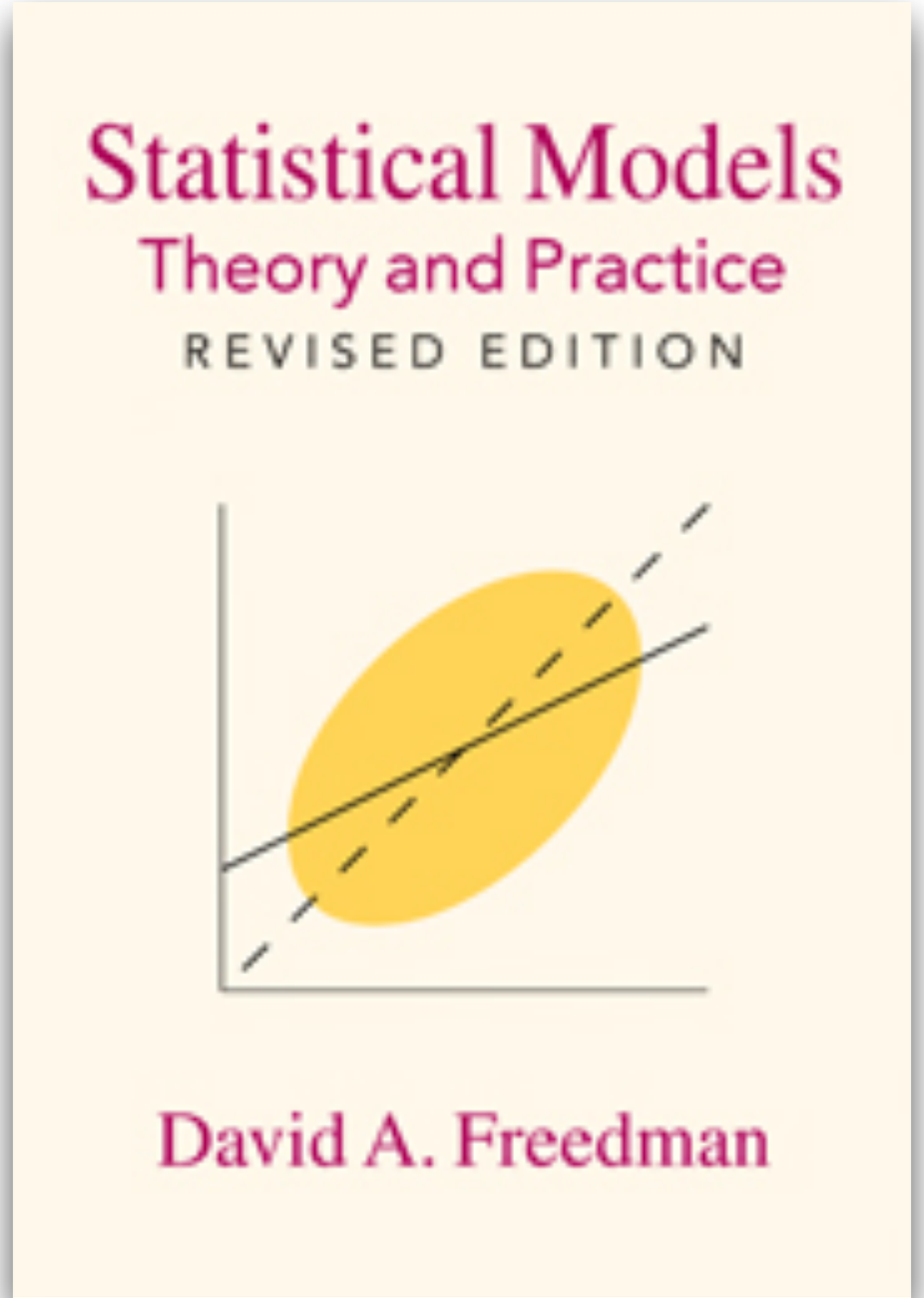
August 29, 2023

Administrativia

- Instructor: Jonathan Terhorst
  - Office hours: 252 WH W1-3p or by appointment
- GSI: Jesse Wheeler
  - Lab: W4-5:30p, B760EH
  - Office hours: TBD

Textbooks

- **Required**: Statistical Models: Theory and Practice (2e)
  - David A. Freedman
  - Can download PDF free from library website.
- **Optional**:
  - Statistics (4e) (Freedman, Pisani, Purves) (Some exercises are taken from this book but you don't need it.)
  - ESL (Hastie, Tibshirani, Friedman)
  - PRML (Bishop)

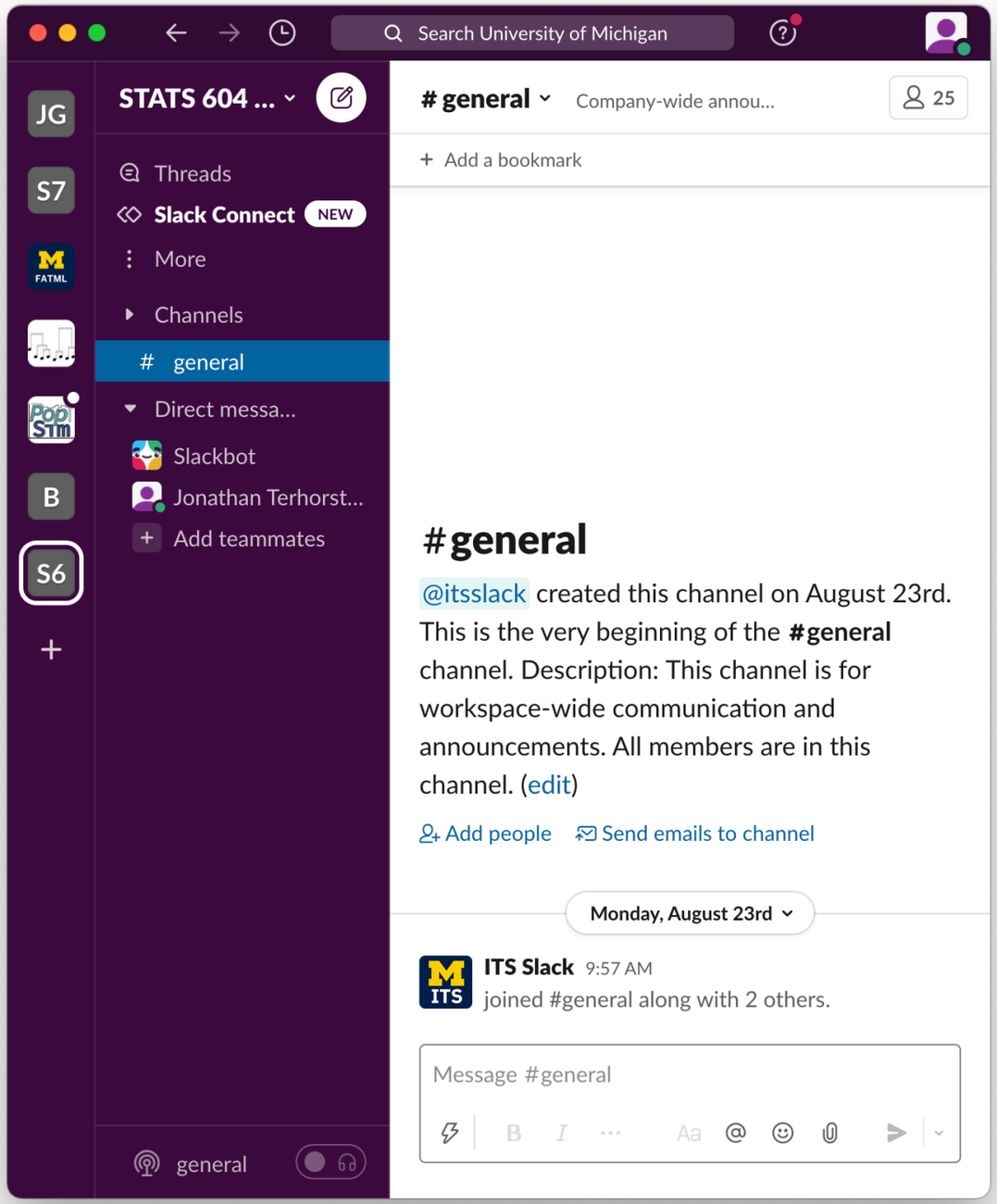


Software

- R is the preferred language.
    - Slides and assignments will be distributed in R.
    - If you absolutely must use another language, discuss with me first.
- Also required is "notebook" software (either RStudio or Jupyter).
- Docker and Git will be required for producing reproducible analyses.
    - We'll go over using these later in semester.

Slack

- You should all have access to the Slack channel.
- This is the best venue for discussing course-related matters, seeking help, etc.
- I prefer that questions go here instead of e-mail, so that answers are visible to all.
- For confidential matters, use DM or e-mail.



Course Structure

Two-tracks

- Tuesdays:
  - More "theoretical".
  - First 1/2 of semester will roughly follow the textbook.
  - Second 1/2 of semester will critically analyze applied papers.
- Thursdays:
  - Focus more on the active "doing" of statistics.
  - Largely project based.
  - Develop presentation and other communication skills.
  - Bring your laptop!

Assignments and Grading

- Homework:
  - About 5 HW assignments, staggered to avoid too much overlap with projects.
  - Focus primarily on material from Tuesday classes.
  - You may collaborate, but each person must write up their own solutions.
- Projects:
  - Projects 1-3 each will take ~3 weeks.
  - Final Project will take ~5 weeks.
  - You'll work in randomly-assigned groups of 3-4 (different for each project).
  - Grade based on peer-reviewed writeup + in-class presentation.

Component	Weight
Projects 1-3	30% (10% each)
Final Project	20%
Midterm	20%
Homeworks	15%
Participation	15%

- Midterm:
  - Primarily based on Tuesday lectures and assigned reading.
  - Closed-book, no notes.
  - You may use a calculator.

Participation

- Class discussions are central to this course.
- **You are expected to participate meaningfully in the discussions.**
- (This means doing the reading ahead of time.)
- Participation grade will be determined holistically at the end of the semester.

This portion of your grade is <u>not</u> "freebie points".

How to do well

- Ask questions
    - Ask questions in lecture, ask questions in lab, ask questions in office hours, and ask questions online. We are here to help...
    - **Asking questions = participation.**
- Don't fall behind. This course is a lot of work and moves very quickly. Start on the projects immediately after they are introduced.
- Do the reading. Especially in the second half of the semester, when we will be critically evaluating applied papers during lecture. In particular, it is essential do the reading so that you can...
- Participate in class discussions. Class discussions are a central part of the course.

Introductions

- Name
- What is your statistics background?
- Research area or interest
- What would you like to get out of this course?
- Favorite food
- Random fact about yourself!
- (I'll start)

What is STATS 604

- Last year you learned all about:
  - Classical statistics / regression (STATS 600)
  - Modern statistics / ML (STATS 601)
- In this class we study how these/other techniques are utilized in real-world applications, via:
  - Studying/discussing examples from the scientific literature.
  - Conducting your own hands-on analyses.

What is STATS 604

- Not much math/theory will be taught.
- Familiarity with the methods is assumed based on 600/601.
- If you need a refresh, come to OH/lab.
- Or consult supplemental reading.

Tukey on reading vs. studying

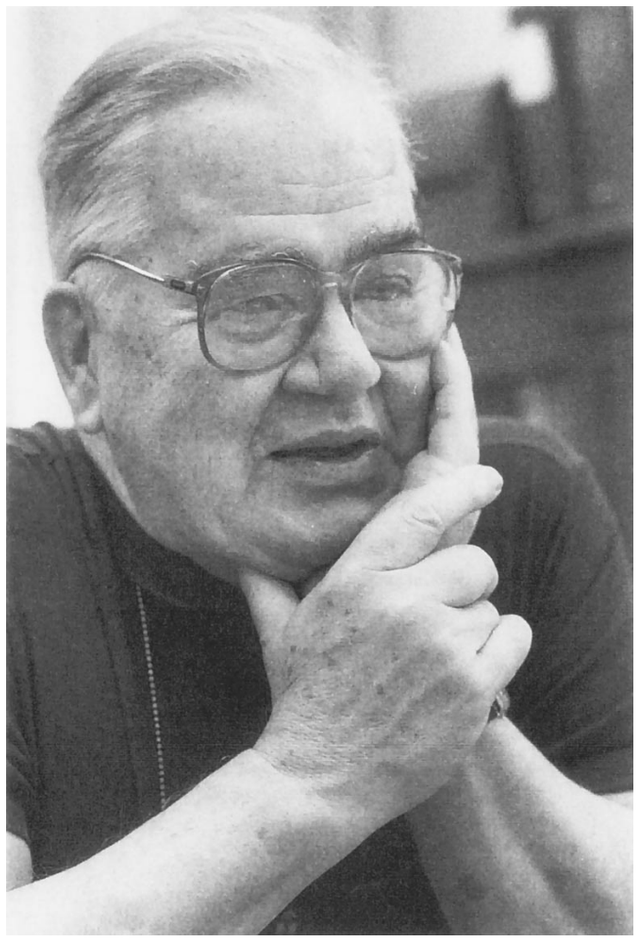


FIG. 1. *John Tukey, date and place unknown.*

J: I'm not sure that that's what happened early on. My first quasistatistical paper is probabilistic. It's the one about the fractional part of a statistical variable. I had read a fair amount of statistics because I read a fair amount of many of the things that were in the math library at Brown. I read them rather than studied them. Let me get a bibliography [gets a bibliography from the bookcase]. The

FIG. 3. *John W. Tukey receiving the National Medal of Science from President Nixon in 1973.*

involved in the working of that. That's not just a decorative appearance on the list of authors.

Q: Would you say that you read most of the literature that was published? As it came out, you read it?

J: I don't know. What was maybe more important was that I read *Series B*, then called the *Supplement to the Journal of the Royal Statistical Society*—read, again, rather than studied—from volume 1 on. And I read through *Biometrika*, so that I had a reasonably good feel for what people were doing or had been doing—for 40 years in the case of *Biometrika*.

Fernholz, Luisa T., and Stephan Morgenthaler. "A conversation with John W. Tukey and Elizabeth Tukey." *Statistical Science* 15.1 (2000): 79-94.

Concrete goals

Active participation

- I want you to participate actively in labs and lectures.
- The more participants, the more enjoyable it is for all.
- Be an active listener too.
- Goals for improved communication:
  - Speakers get better at listening.
  - Listeners get better at speaking.

To speak or not to speak

- Pros:
  - Practice communication skills.
  - Clarify your own thoughts
  - Getting to know one another.
  - You learn more if you are engaged.
- Cons:
  - Perceived embarrassment. (We're all here to learn.)
  - Risk of confusing yourself or others. (I get confused, everyone gets confused, it's a part of learning.)

Concrete goals

Working with others

- Modern data problems are complex and need multi-disciplinary teams to solve.
- Communication and interpersonal skills are more important than ever.
- What are your suggestions for:
  - Making sure your voice gets heard?
  - Inviting introverts into the discussion?
  - Pushing back on those who take too much space.

Concrete goals

Critical thinking & problem solving

- To be good at statistics, you have to have good technical understanding *and* a healthy dose of common sense.

- What does critical thinking mean?
  - Don’t rush to judgement.
  - Continually ask “why”, look for negative examples/controls.
  - Seek the opinions of others.

Preventing confirmation bias

Ineffective strategies

1. Knowing that c.b. exists won't cure it.

2. Looking for evidence that your initial hypothesis is wrong won't cure it.

3. Exposing yourself to a lot of information won't necessarily help you discover the right answer.

Effective strategies

1. Stick to your guns—don't abandon your first guess too readily.

2. Learn to think of the unlikely, and look for evidence thereof.

3. Embrace surprises—when something didn't go as planned, ask why.

> “If you adopt a strategy that is one part sticking to your guns, one part considering far-out ideas, and, one part paying attention to surprises, you’re ready to adapt to whatever the world throws at you in the way of evidence.

Figuring out how complex things work is a lot like fishing. If you don’t know which lure works for the fish you’re after, start with your best guess and experiment from there. This strategy might come in especially handy if you plan to go fishing on Mars.”

https://www.globalcognition.org/confirmation-bias-3-cures/

Based on research by:
Mynatt, Clifford R., Michael E. Doherty, and Ryan D. Tweney. "Confirmation bias in a simulated research environment: An experimental study of scientific inference." Quarterly Journal of Experimental Psychology 29.1 (1977): 85-95.

Concrete goals

How to give and receive criticism

- In addition to examining your own work critically, it is important to be able to criticize the work of others in a respectful and constructive manner.

- You will practice this by peer-reviewing each other's assignments.

- To reduce bias and encourage honest feedback, peer reviews are doubly blind.

- Each lab is graded on a scale of 0-10 (rubric will be provided).

  - Be polite but *critical*—most assignments should not receive a 10.

  - The GSI will ensure that grading is ultimately fair and uniform.

Concrete goals

Improved communication

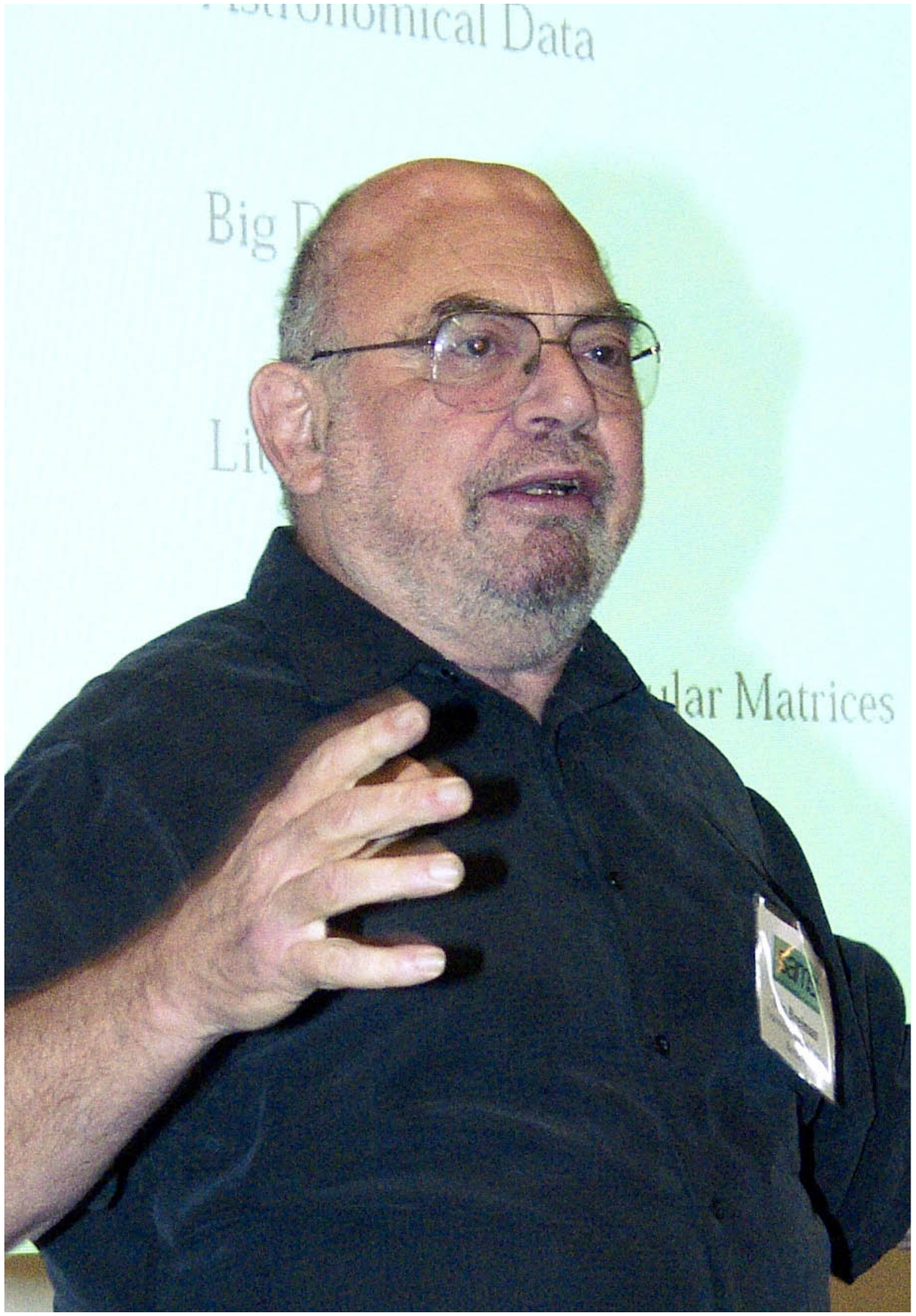
- Improve your ability to communicate (both verbally and in writing) statistics to others.
- Probably the hardest thing to teach!
- Which is why it’s often not taught!!
- One of the best ways to learn is by reading well-written papers.
- Random sampling of statisticians who write well (IMO): Jerzy Neyman, Leo Breiman, Peter Hall, David Freedman, Radford Neal, Brad Efron, Andrew Gelman, David Donoho, Mike Jordan, Rob Tibshirani, Jim Berger.

Statistics in 2023

Leo Breiman

1928—2005

- Invented CART, random forests, bagging, …
- (From those who knew him):
  - Original and daring thinker.
  - Not afraid to challenge the statistics establishment.
  - Artistic (designed his own home, talented sculptor).
  - Nice guy.



Statistical Science
2001, Vol. 16, No. 3, 199–231

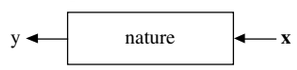
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

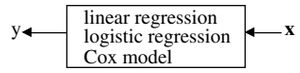
The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

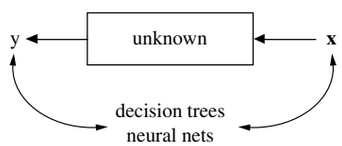
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

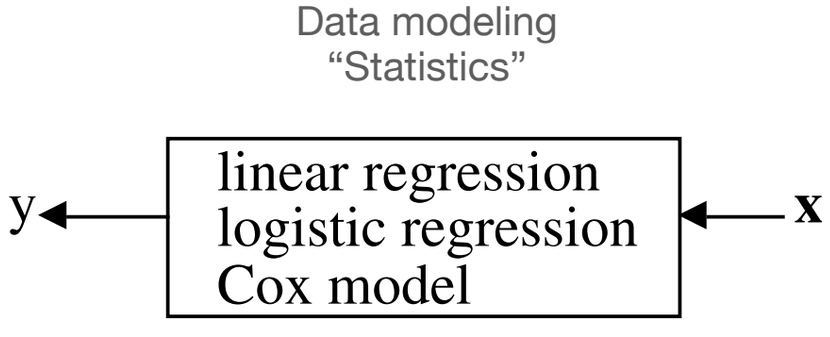
In this paper I will argue that the focus in the statistical community on data models has:

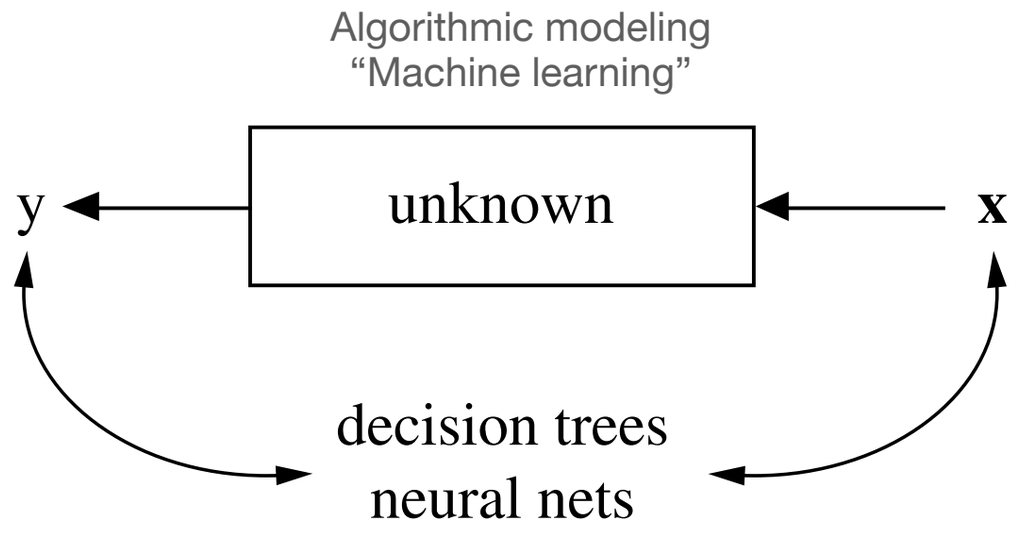
- Led to irrelevant theory and questionable scientific conclusions;

- One of my all-time favorite papers.
- Continues to generate much discussion, even today.

  See e.g. *Observational Studies* 7.1 (2021).
- Amazingly prescient—predicted exactly how the next 20 years would unfold.
- Sharp, incisive writing.

Breiman's two cultures

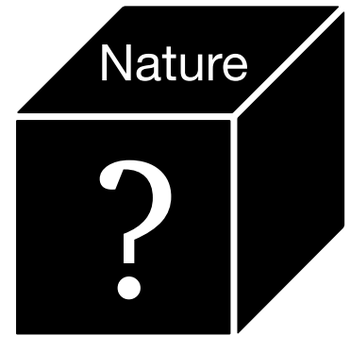




- More observations: good
- More predictors: bad
- p-values: yes
- Interpretable: yes
- Verification: "goodness of fit"
- Computation: generally easy

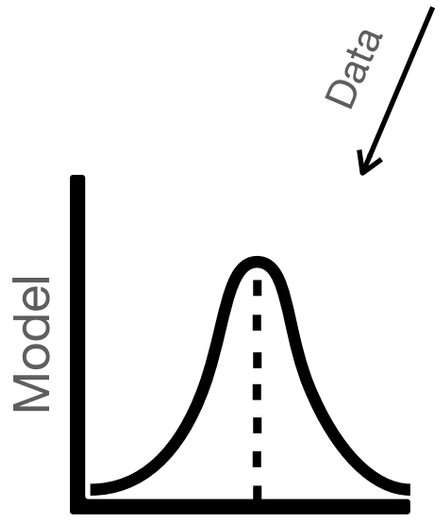
- More observations: good
- More predictors: good
- p-values: no
- Interpretable: currently*, no
- Verification: predictive error
- Computation: challenging

The difficulty with data modeling



- The conclusions we draw are about the model (not nature!)

- Example: p-value = “the probability of observing a value at least this extreme under the assumed model”



- Hence, they are useful *only if* the model is a good representation of reality.

- Often, this is un-checkable, or at least requires a lot of hard-won experience and domain knowledge.

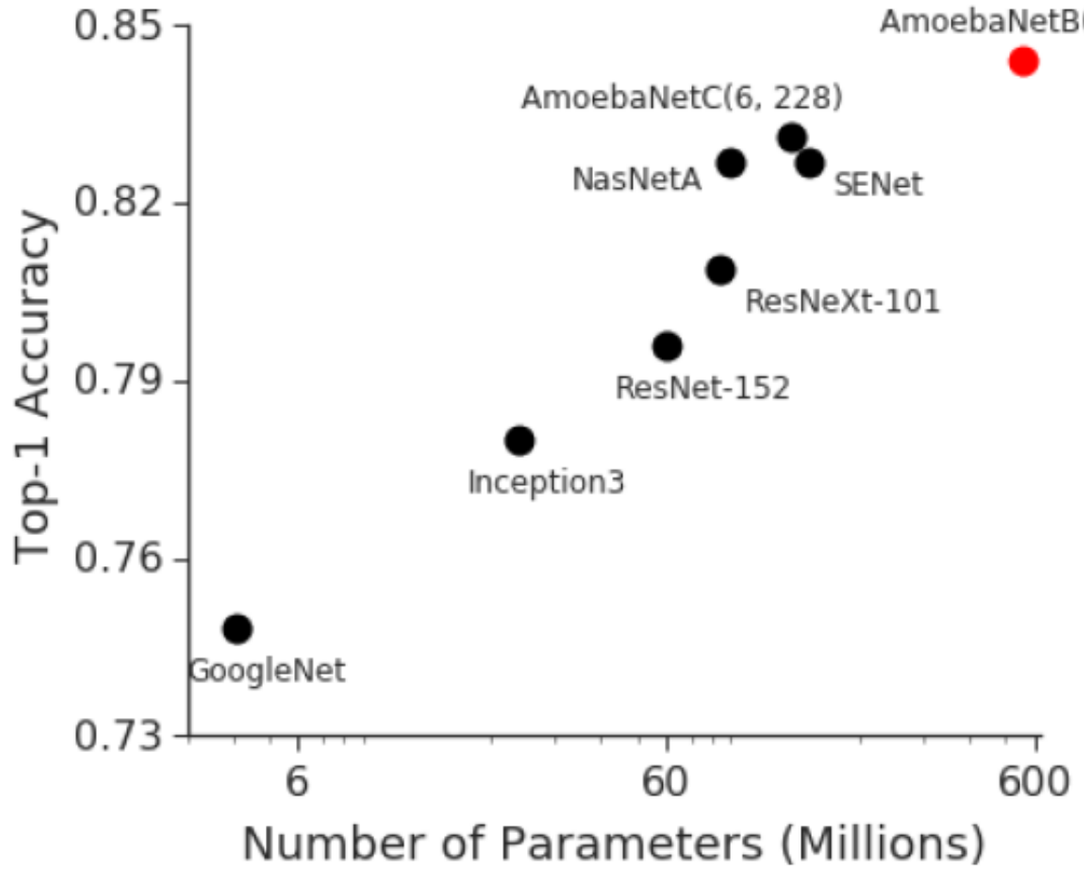
Rashomon effect

- Often, many different models perform (roughly) equally well.
- This happens all the time in practice.
- How to choose between them?
- One of Breiman's insights was to exploit this phenomenon via bagging, leading to more stable models.

Prediction and science

- Central to Breiman's argument is the idea that data sets have gotten increasingly complicated.
- Old/simplistic models are no longer adequate.
- If black box models are our only hope, what does this mean for science?
- What has been your experience?

Strong correlation between ImageNet accuracy and model size for recently developed representative image classification models



https://ai.googleblog.com/2019/03/introducing-gpipe-open-source-library.html

Counterarguments

- To be fair, statisticians have been aware of this issue for a long time.
- Many scientists would rather have a wrong model that they can reason about.
- There is some evidence that this has worked in the past.
- But maybe the situation has changed?

THE NEED FOR SIMPLE SCIENTIFIC MODELS - PARSIMONY

The scientist, studying some physical or biological system and confronted with numerous data, typically seeks for a model in terms of which the underlying characteristics of the system may be expressed simply.

For example, he might consider a model of the form

$$y_u = f^{(p)}(\underset{\sim}{\xi}_u \underset{\sim}{\theta}) + \varepsilon_u \qquad (u = 1,2,\ldots,n) \qquad (1)$$

in which the expected value η_u of a measured output y_u is represented as some function of k inputs $\underset{\sim}{\xi}$ and of p parameters $\underset{\sim}{\theta}$, and ε_u is an "error". One important measure of simplicity of such a model is the number of parameters that it contains. When this number is small we say the model is parsimonious.

Parsimony is desirable because (i) when important aspects of the truth are simple, simplicity illuminates, and complication obscures; (ii) parsimony is typically rewarded by increased precision (see Appendix 1); (iii) indiscriminate model elaboration is in any case not a practical option because this road is endless*.

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do

*Suppose for example that in advance of any data we postulated a model of the form of (1) with the usual normal assumptions. Then it might be objected that the distribution of ε_u might turn out to be heavy-tailed. In principle this difficulty could be allowed for by replacing the normal distribution by a suitable family of distributions showing varying degrees of kurtosis. But now it might be objected that the distribution might be skew. Again, at the expense of further parameters to be estimated, we could again elaborate the class of distribution considered. But now the possibility might be raised that the errors could be serially correlated. We might attempt to deal with this employing, say, a first order autoregressive error model. However, it could then be argued that it should be second order or that a model of some other type ought to be employed. Obviously these possibilities are extensive, but they are not the only ones: the adequacy of the form of the function $f(\xi,\theta)$ could be called into question and elaborated in endless ways; the choice of input variables $\underset{\sim}{\xi}$ might be doubted and so on.

Box, George EP. "Robustness in the strategy of scientific model building." *Robustness in statistics*. Academic Press, 1979. 201-236.

Counterarguments

- Breiman's paper was (intentionally) provocative.

thesis testing and asymptotics. There is a wide spectrum of opinion regarding the usefulness of the theory published in the *Annals of Statistics* to the field of statistics as a science that deals with data. I am at the very low end of the spectrum. Still, there

- Several of the discussants (Cox, Efron) are not altogether sold.

- Is prediction scientifically useful?

Comment

Brad Efron

At first glance Leo Breiman's stimulating paper looks like an argument against parsimony and scientific insight, and in favor of black boxes with lots of knobs to twiddle. At second glance it still looks that way, but the paper *is* stimulating, and Leo has some important points to hammer home. At the risk of distortion I will try to restate one of those points, the most interesting one in my opinion, using less confrontational and more historical language.

Prediction, estimation, attribution

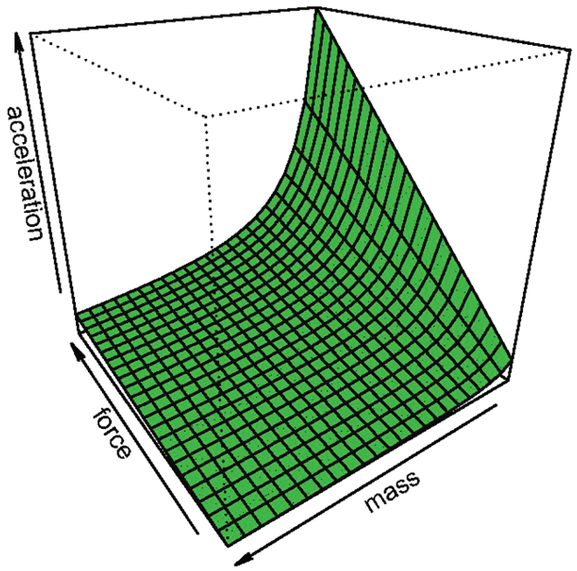
- One definition of science: separating eternal (or at least, long-lasting) structure from ephemeral noise.

- Data generation

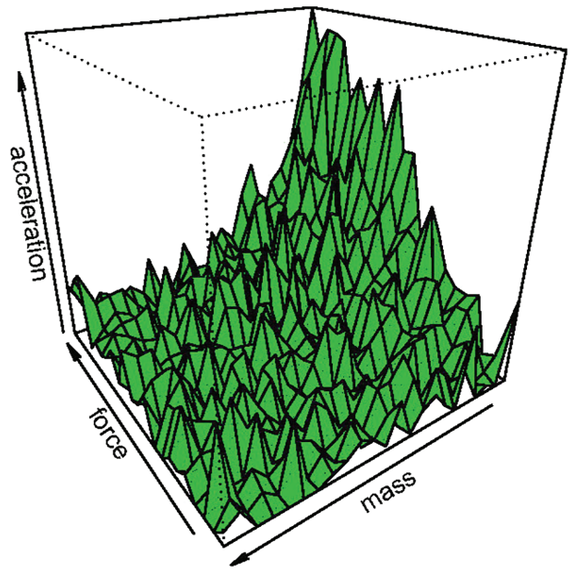
  - Traditional: (usually regular) surface perturbed by noise.

  - Predictive: Irregular, must be learned from data.

$F = ma$



$F = ma + \varepsilon$



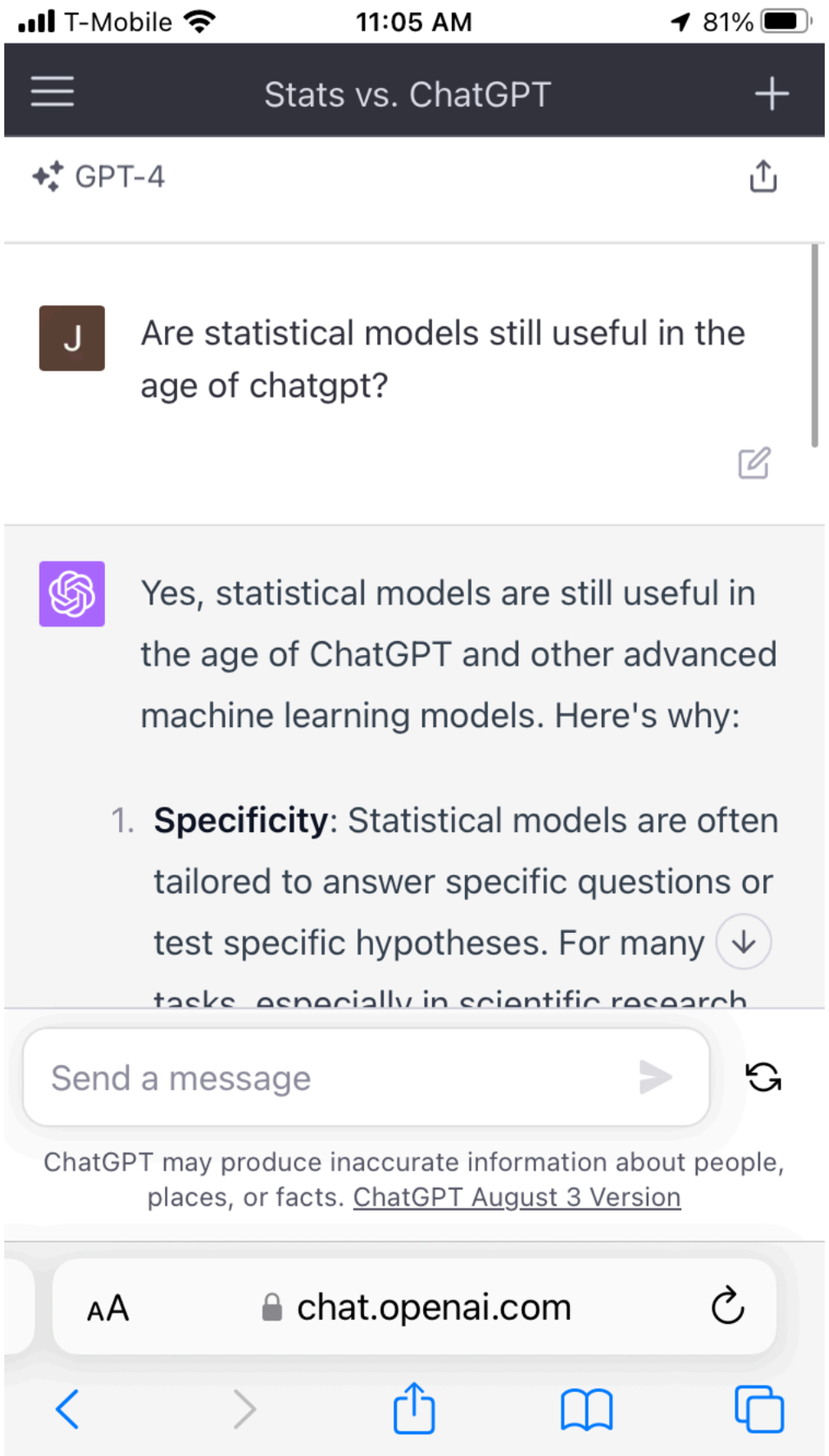
Efron, Bradley. "Prediction, estimation, and attribution." *International Statistical Review* 88 (2020): S28-S59.

Efron's two cultures

Table 5. *A comparison checklist of differences between traditional regression methods and pure prediction algorithms.*

	Traditional regressions methods	Pure prediction algorithms
1.	Surface plus noise models (continuous, smooth)	Direct prediction (possibly discrete, jagged)
2.	Scientific truth (long-term)	Empirical prediction accuracy (possibly short-term)
3.	Parametric modeling (causality)	Nonparametric (black box)
4.	Parsimonious modeling (researchers choose covariates)	Anti-parsimony (algorithm chooses predictors)
5.	$\mathbf{x}p \times n$: with $p \ll n$ (homogeneous data)	$p \gg n$, both possibly enormous (mixed data)
6.	Theory of optimal inference (mle, Neyman–Pearson)	Training/test paradigm (Common Task Framework)

Efron, Bradley. "Prediction, estimation, and attribution."
International Statistical Review 88 (2020): S28-S59.


T-Mobile
11:05 AM
81%
Stats vs. ChatGPT
GPT-4
J
Are statistical models still useful in the age of chatgpt?
Yes, statistical models are still useful in the age of ChatGPT and other advanced machine learning models. Here's why:
1. **Specificity**: Statistical models are often tailored to answer specific questions or test specific hypotheses. For many
Send a message
ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version
AA
chat.openai.com

Discussion

🤔 Do you identify with one culture more than the other?

🤔 Which is more useful to your work?

🤔 Can you give an example of a successful application of:

- Data modeling?
- Algorithmic modeling?

🤔 Do you agree with Efron's assertion that regression is more "scientific"? Do you find his data examples compelling?

🤔 Do you agree with Breiman's assertion: "the soundest path is to go for predictive accuracy first, then try to understand why?"

Data science lifecycle

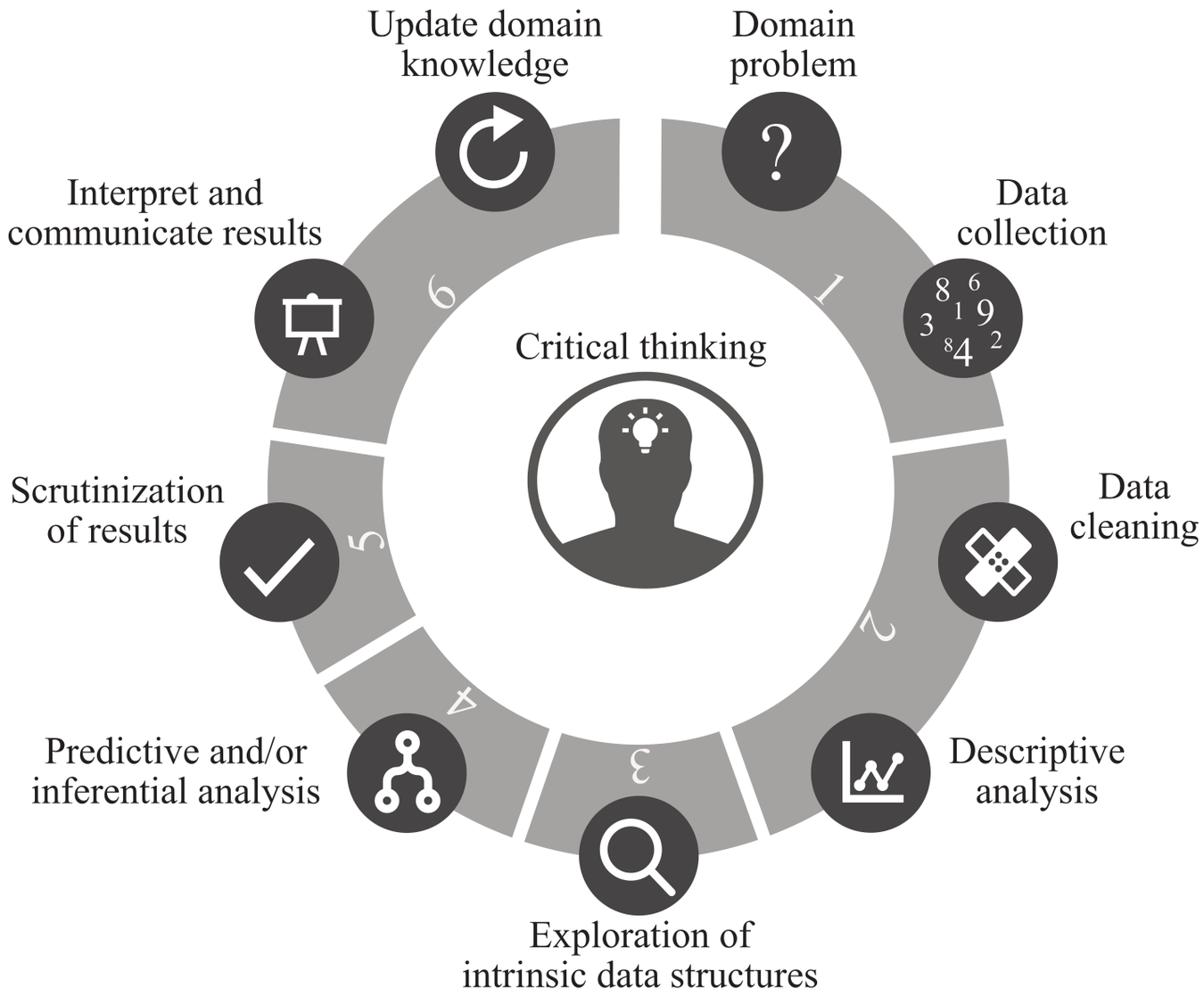


Image courtesy of Bin Yu

DSLC 1

Question formulation

- The first stage of the DSLC involves identifying a domain problem, and formulating a question whose answer will help you understand it better.
- Many interesting questions are nevertheless unanswerable!
- It’s up to you to refine the question to the point where:
  A. It can be answered using the available data; and
  B. Remains interesting/useful for answering the original question.

Formulating an answerable question

Example

- “Will I contract COVID today?”
  - Definitely interesting (at least to me).
  - Not answerable as stated.
- Refinements:
  - “What are my odds of getting COVID today?”
  - “What is the positivity rate on campus today?”
  - “What was the positivity rate on campus last week?”

Exercise

- With your table, agree on an interesting data question. Start general.
- Using Google, try and identify a source of data that could potentially answer this question.
- If none exists, refine question and try again.
- Record your progress.

DSLC

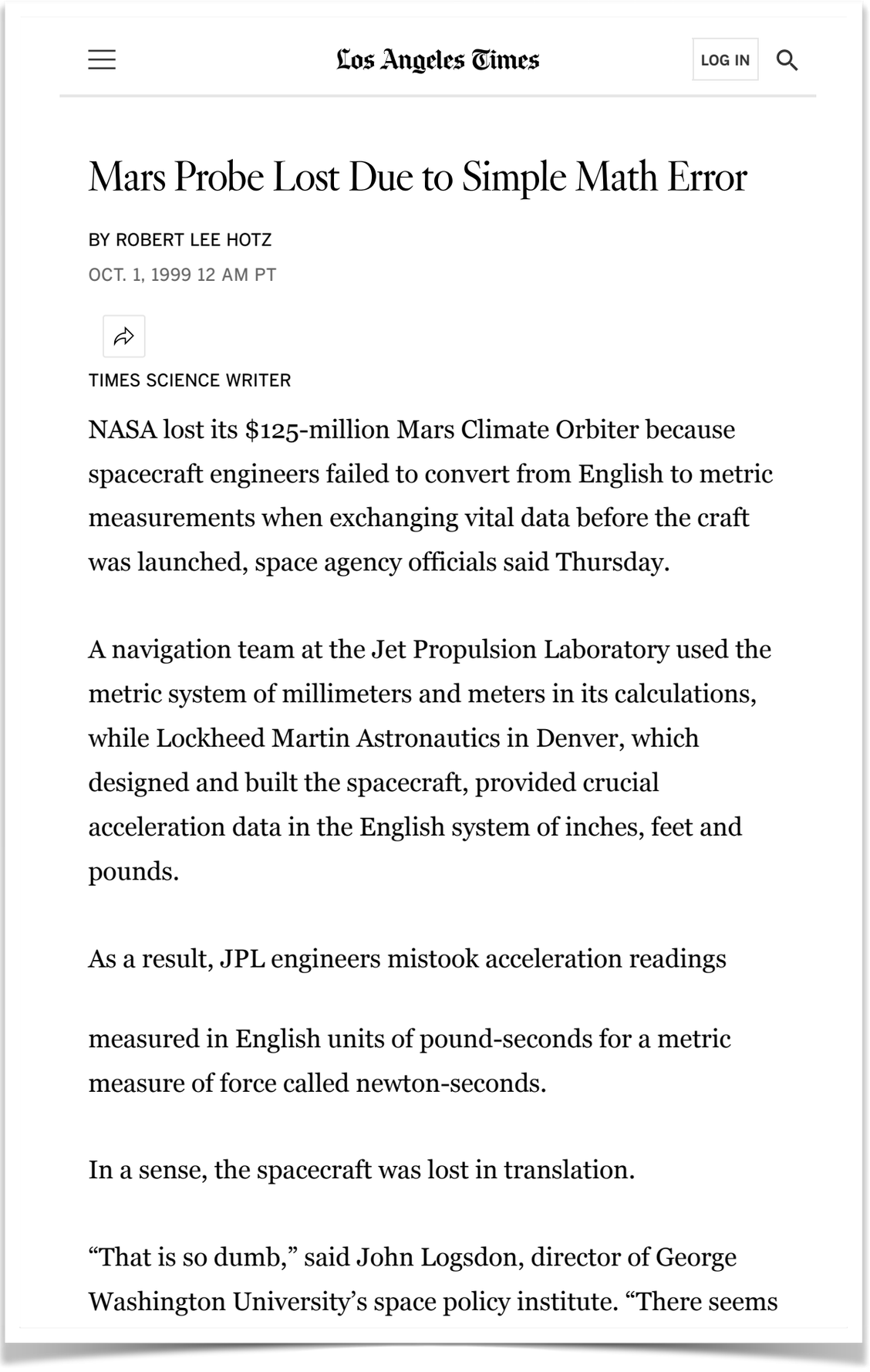
Data collection

- Most of the time, you will be working with publicly available data collected from repositories.
    - Try to understand as much about how the data were collected as possible.
    - Sometimes, the researchers who originally collected the data are available for questions.
    - Otherwise, rely on READMEs, original publication(s), possibly other researchers you know who have worked with the data in the past.

DSLC

Data collection

- Serious errors have been committed due to users not understanding certain key aspects of their data.



Los Angeles Times

LOG IN

Mars Probe Lost Due to Simple Math Error

BY ROBERT LEE HOTZ

OCT. 1, 1999 12 AM PT

TIMES SCIENCE WRITER

NASA lost its $125-million Mars Climate Orbiter because spacecraft engineers failed to convert from English to metric measurements when exchanging vital data before the craft was launched, space agency officials said Thursday.

A navigation team at the Jet Propulsion Laboratory used the metric system of millimeters and meters in its calculations, while Lockheed Martin Astronautics in Denver, which designed and built the spacecraft, provided crucial acceleration data in the English system of inches, feet and pounds.

As a result, JPL engineers mistook acceleration readings

measured in English units of pound-seconds for a metric measure of force called newton-seconds.

In a sense, the spacecraft was lost in translation.

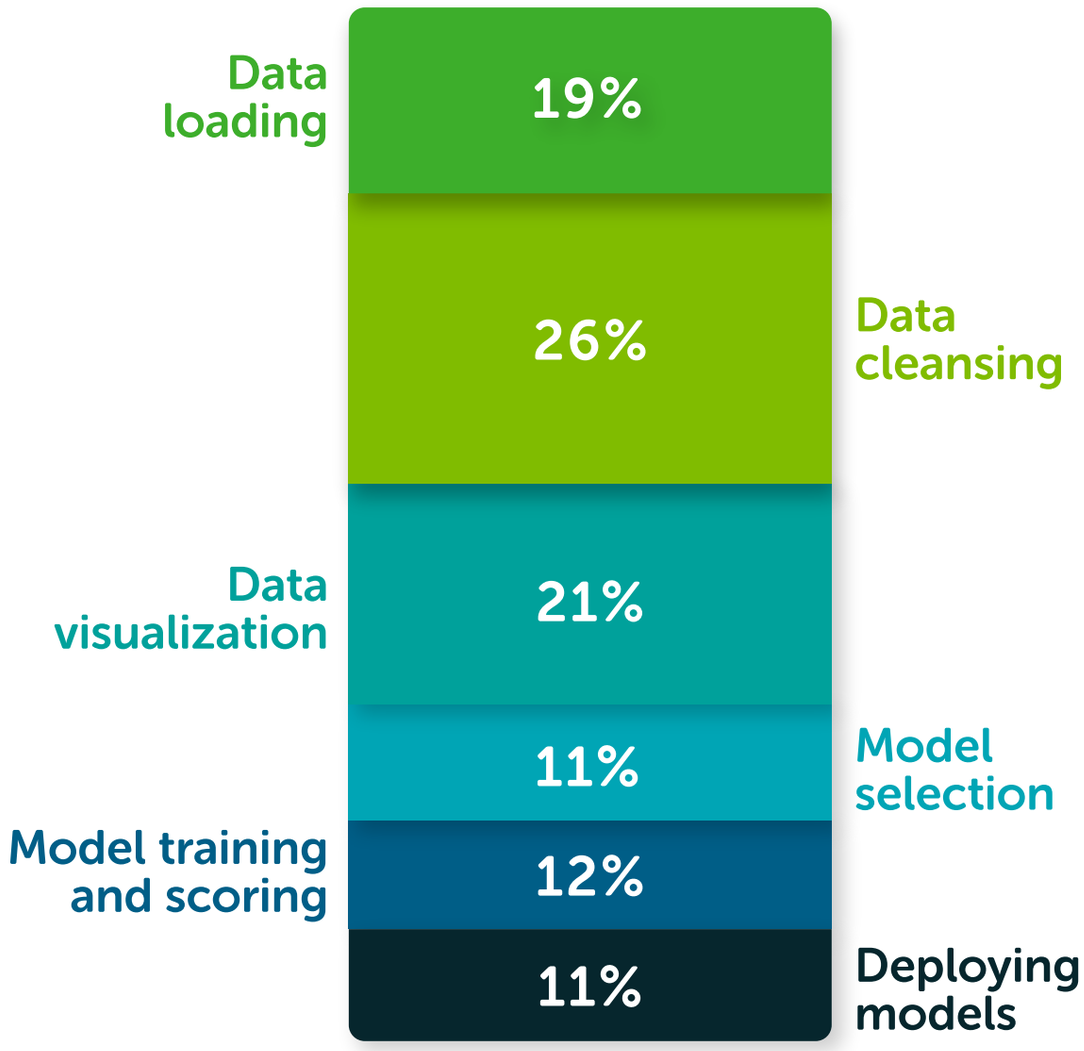
“That is so dumb,” said John Logsdon, director of George Washington University’s space policy institute. “There seems

DSLC

Data cleaning and exploration

- Acquiring, cleaning, and exploring data occupies 2/3rds of the data scientist's time.
- This process involves a large number of judgment calls—document them!
- Feel free to explore your data. Stability will be assessed later on.

HOW MUCH OF YOUR TIME IS SPENT IN EACH OF THE FOLLOWING TASKS?



The State of Data Science 2020