Stat 214 - Spring 2026

Berkeley
UNIVERSITY OF CALIFORNIA

Data Analysis and ML for Real-World Decision Making

A new framework for practicing responsible
data analysis and decision making

Instructor: **Bin Yu**

GSIs: **Anqi Wang, Sean Richardson, Sequoia Andrade, Zach Rewolinski**




Stat 214 is a stat MA course approved by campus in fall 2024

This is an MA class in statistics. Students will be engaged in open-ended data projects for decision making to solve domain problems. It mirrors the entire data science life cycle in practice, including problem formulation, data cleaning, exploratory data analysis, statistical and machine learning modeling and computational techniques, and interpretation of results in context. It is guided by the Predictability-Computability-Stability (PCS) framework for veridical data science and emphasizes critical thinking and documenting human judgment calls and code. It coaches not only the technical but also communication and teamwork skills in order to obtain responsible and reliable data-driven conclusions for solving complex real-world problems.

214 design work started in summer 2024

- 214 uses much material from 215A that Bin has been teaching every year since 2008 except for two years
- We added materials on deep learning, cut down on linear and generalized linear models, and got GPUs on NSF ACCESS platform

Many thanks to the wonderful prep team and support form the dept.

Anthony Ozerov

Rodrigo Palmaka

Zach Rewolinski

Substantial Additional Work by TAs in Spring 2025

Many thanks to them as well

GSIs: **Abhi Agarwal, Zach Rewolinski, Austin Zane**

Who are we?

Bin, Anqi, Sean, Sequoia, and Zach: quick self introductions

Questions?

Class: introduce yourselves to your neighbors

More questions?

Why are you here?

What is your goal?

Our goal:

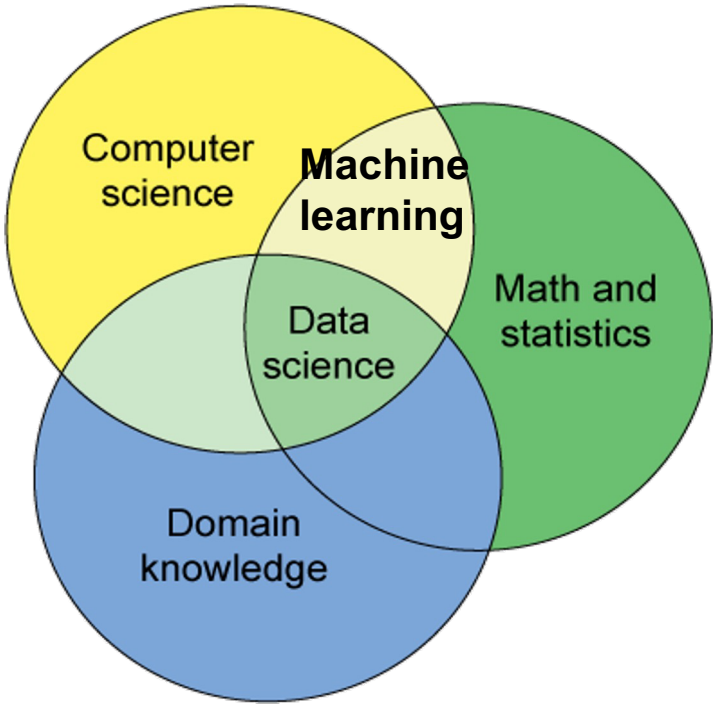
to help you learn the process of drawing trustworthy data conclusions
in a reliable and transparent manner
so that you are confident to do the same and pick up new ideas and techniques on your own in the future

How are we going to work together?

We do not have all the answers

Individual learning + Collaboration is our approach

Data science (DS) is a pillar of AI



Conway's Venn Diagram

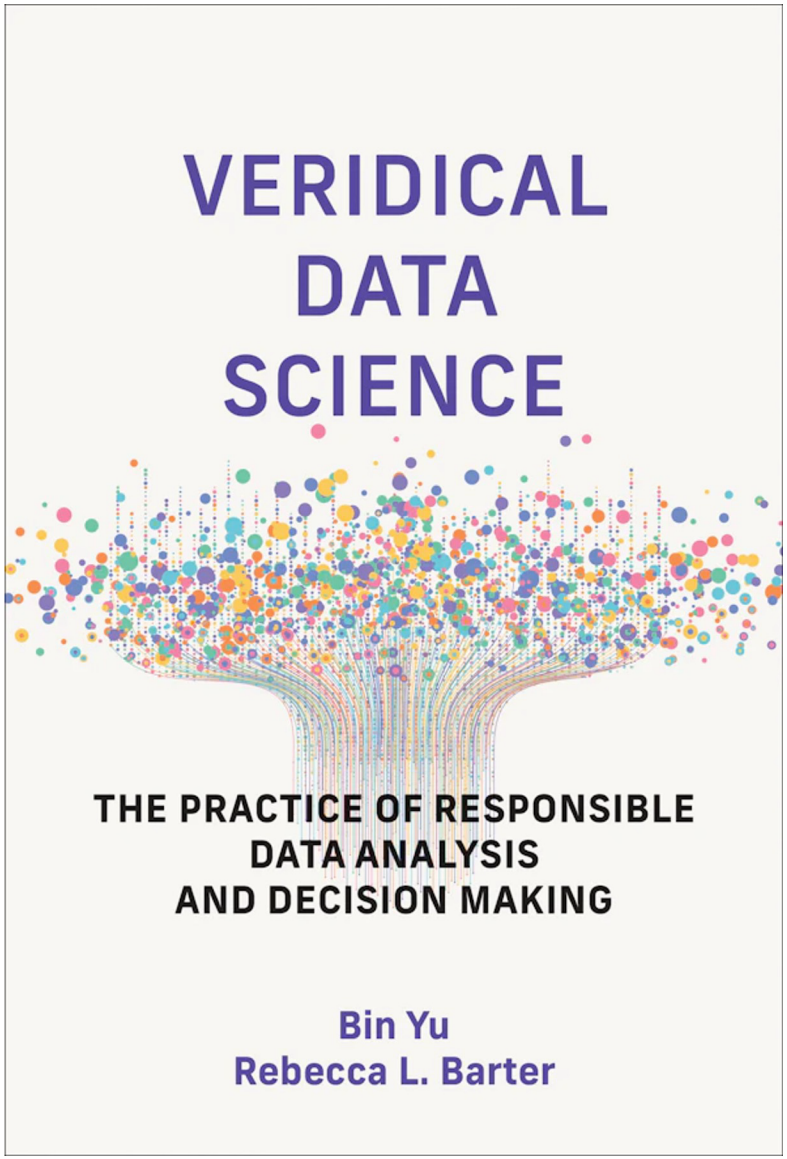
Co-author

R. Barter

Free version at

vdsbook.com

VDS book took 9 years to write.



MIT Press (ML Series)

Oct. 15, 2024

Veridical

Definitions

Definitions from Oxford Languages · Learn more

adjective **FORMAL**

truthful.
"Pilate's attitude to the veridical"

- coinciding with reality.
  "such memories are not necessarily veridical"

What does “veridical” mean in VDS?

- Veridical means “truthful” in English, and is a common word in Spanish where it implies verified truth.

- A more precise articulation of “veridical” in VDS hinges on both:

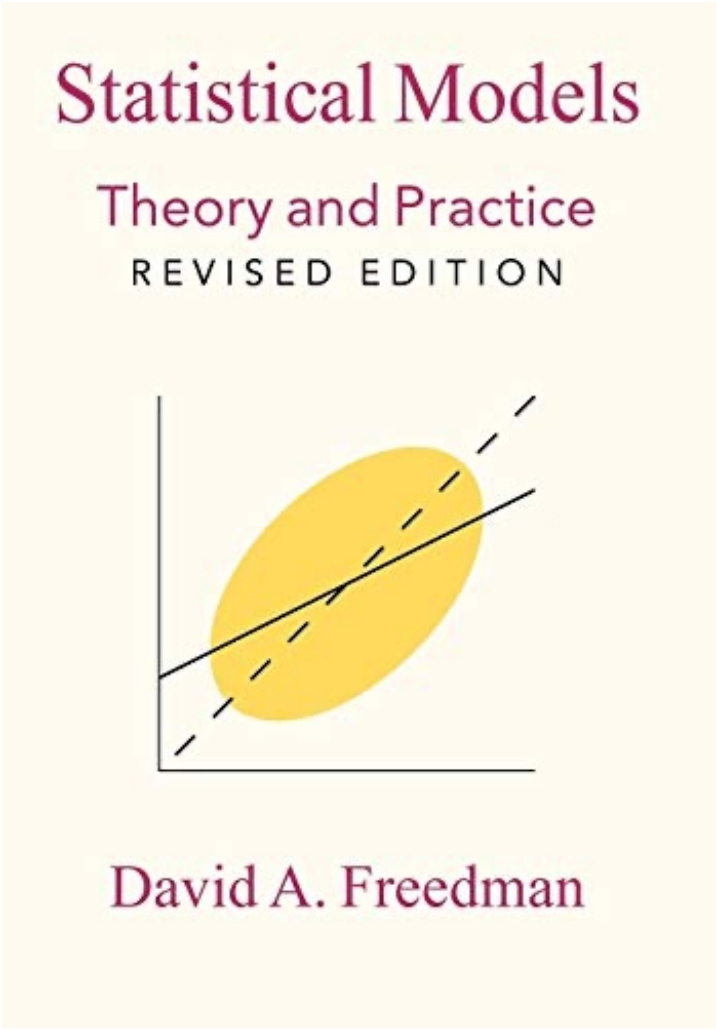
1. It seeks truth in data conclusions.
2. It goes through a transparent (or truthful) DSLC guided by PCS towards verification.

Veridical data science

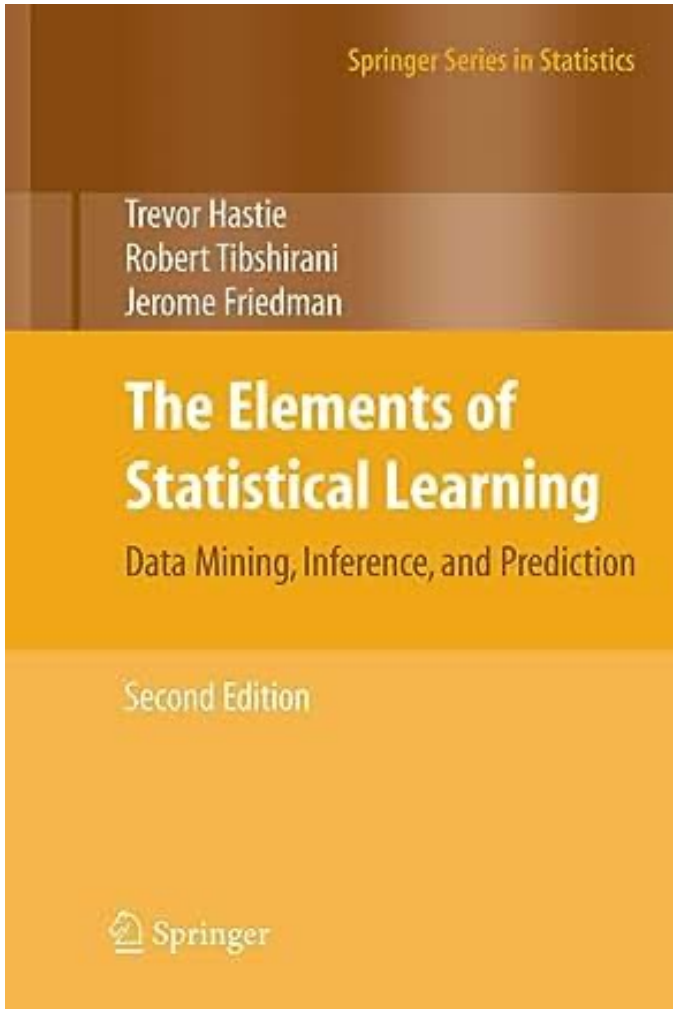
Veridical Data Science (VDS) is a philosophical and conceptual framework for practicing data science responsibly with a documentation requirement.

VDS provides a framework for **producing trustworthy data-driven** results, and **critically assessing** the **trustworthiness of data-driven results** in the context of domain science and reality.

Original PNAS article: Yu and Kumbier (2020), *Veridical Data Science*

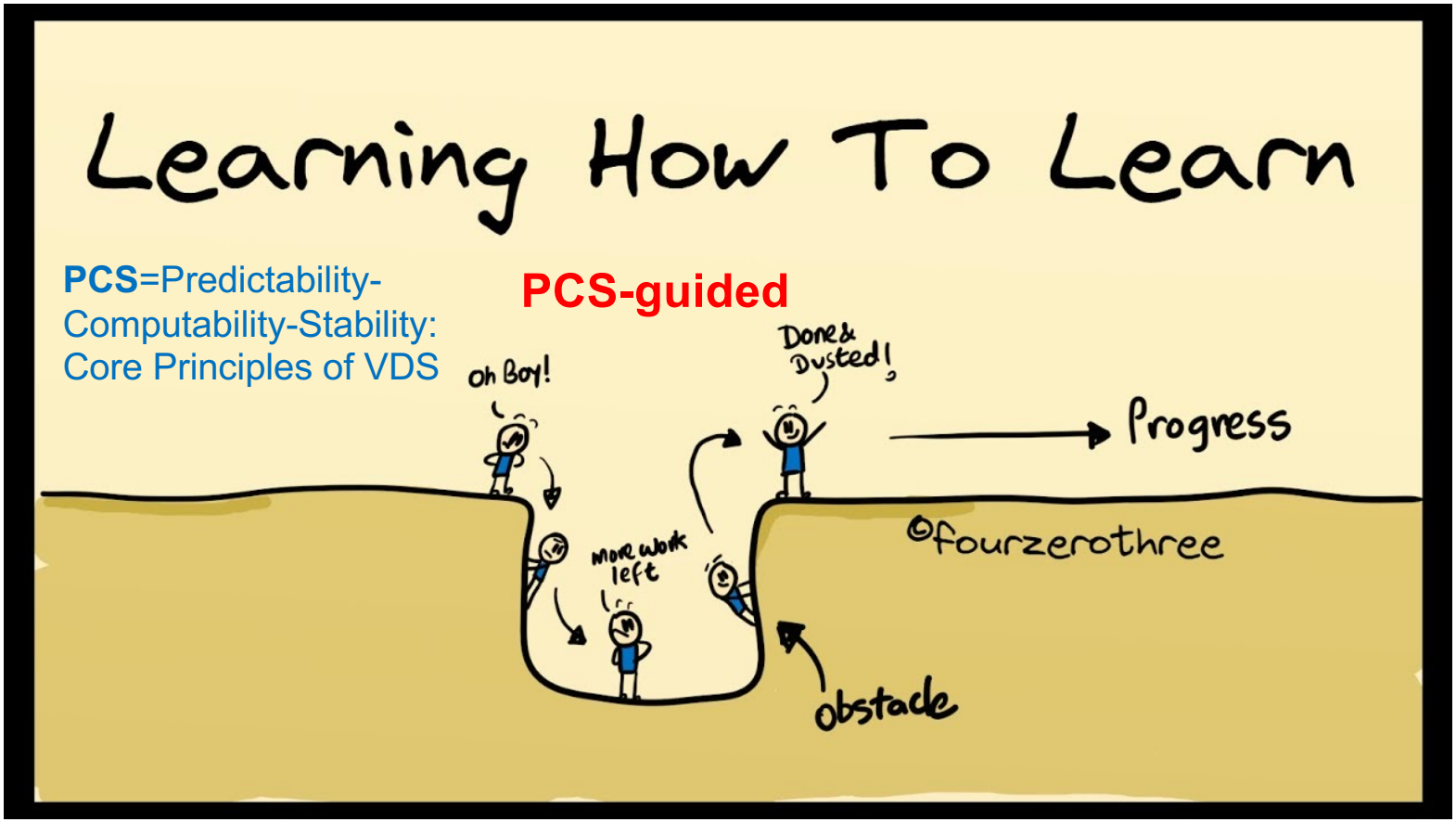


Recommended



Recommended

Teaching philosophy

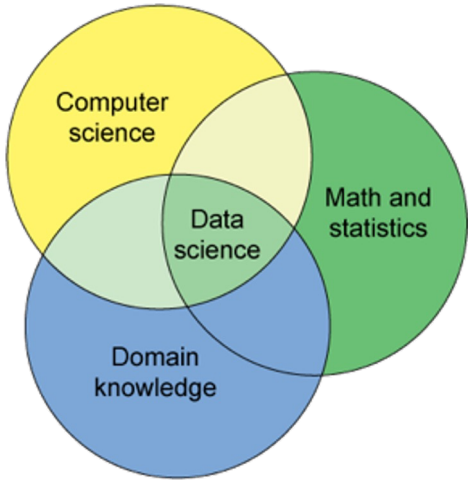


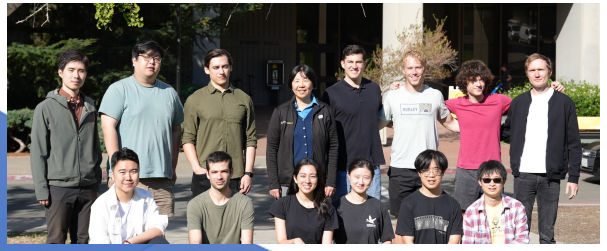
https://www.fourzerothree.in/p/learning-how-to-learn

People make “veridical” happen

New name for applied stats+ML

Data Science

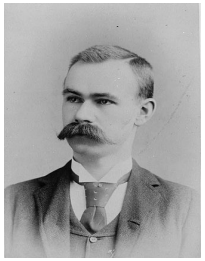




VDS follows a long tradition in statistics research, driven by solving a real problem

One writes a paper when there is something worth communicating

Statistician, Inventor
H. Hollerith



1890's Hollerith
Tabulating Machine

Data science is the re-merging of **computational** and **statistical** thinking in the context of domain problems.

ML is part of modern statistics.

Neyman came to Berkeley Math in 1938, started stat lab, …, and became the founding chair of the new Statistics Department in 1955

(1894-1981)

Neyman came to Berkeley Math in 1938, started stat lab, …, and became the founding chair of the new Statistics Department in 1955

Neyman started "***a cell of statistical research and teaching… not being hampered by any existing traditions and routines***"

- -Speed, Pitman and Rice (2000) "*A Brief History of the Statistics Department …at Berkeley*"

Neyman came to Berkeley Math in 1938, started stat lab, …, and became the founding chair of the new Statistics Department in 1955

Neyman started "***a cell of statistical research and teaching…not being hampered by any existing traditions and routines***"

- -Speed, Pitman and Rice (2000) "*A Brief History of the Statistics Department …at Berkeley*"

*"Neyman's **theoretical research** in Berkeley was largely motivated by his **consulting work**,…*

– Lehmann (1994) in "Jerzy Neyman's NAS Biographical Memoir"

Neyman came to Berkeley Math in 1938, started stat lab, …, and became the founding chair of the new Statistics Department in 1955

Neyman started "***a cell of statistical research and teaching…not being hampered by any existing traditions and routines***"

- -Speed, Pitman and Rice (2000) *"A Brief History of the Statistics Department …at Berkeley"*

*"Neyman's **theoretical research** in Berkeley was largely motivated by his **consulting work**,…*

*… His major research efforts in Berkeley were devoted to several large-scale applied projects. … **competition of species** …, **accident proneness** .. , … **galaxies and the expansion of the universe** … , … **cloud seeding**, …*

– Lehmann (1994) in "Jerzy Neyman's NAS Biographical Memoir"

In-context research: developing methods & theory while solving a domain problem

Neyman's in-context research and teaching vision and his leadership were instrumental for Berkeley statistics to become a top statistics department in the world.

Other statistics departments also thrived across the US in late 40's, 50's and 60's.

Bell Labs Statistics (1925-90's) and "R": a forward looking "data science" group

Established 100 years ago, "**Bell Labs** made a great contribution to advancing both **fundamental science** and **technology**."

Bringing back the golden days of Bell Labs - PMC

Bell Labs Statistics Group was a top statistics place in industry, called "Dept. of Statistics and Data Analysis" during my time (98-00) at Lucent Bell Labs (on leave from Berkeley).

Bell Labs Statistics (1925-90's) and "R": a forward looking "data science" group

Established 100 years ago, "**Bell Labs** made a great contribution to advancing both **fundamental science** and **technology**."

Bringing back the golden days of Bell Labs - PMC

Bell Labs Statistics Group was a top statistics place in industry, called "Dept. of Statistics and Data Analysis" during my time (98-00) at Lucent Bell Labs (on leave from Berkeley).

Prominent alums: **Sherwart, Tukey, Chambers, Mallows, Cleveland, Lambert, Pregibon, Nair, Hastie, Hansen…**

Birthplace of the hugely impactful **Control Chart, EDA, "S",** upon which **"R"** was developed by a consortium, and **"Listening Post"**

Abraham Wald (1902-1950)

Research contributions

Statistics:

Decision theory, Sequential analysis, Game theory

Economics
Linear Programming
Geometry

Education contributions

lucid lectures, books, great mentor

Service and leadership contributions

First department chair at Columbia, IMS President

Image source: ps://www.boredpanda.com/world-war-2-aircraft-survivorship-bias-abraham-wald/?utm_source=google&utm_medium=organic&utm_campaign=organic

How did Wald develop sequential analysis?

https://en.wikipedia.org/wiki/World_War_II

Wald developed it with Jacob Wolfowitz, Allen Wallis, and Milton Friedman in the Statistical Research Group (SRG) at Columbia University (1942-45), which was focused on solving military problems during WWII.

Others brought to Wald a pretty formulated problem about increasing bomb testing efficiency and some initial ideas for a solution …

Wallis (JASA, 1980)

How did SRG accomplish so much?

8 out of 17 principals were economists including Milton Friedman and George Stigler

> "My recollections have less to do with specific accomplishments by specific individuals than with the feeling that the **interchange of ideas among members of SRG in their attempts to solve practical problems brought forth a new balance among the concepts which give statistics its vitality.** Each member of SRG had his own version of this balance, but there was much in common."

K. Arnold as quoted in Wallis (1980)

What was the key to Wald's success?

``Wald not only posed his statistical problems clearly and precisely, but he posed them to fit the practical problem and to accord with the decisions the statistician was called on to make.

This, in my opinion, was the key to his success – **a high level of mathematical talent of the most abstract sort, and a true feeling for, and insight into, practical problems**."

From Wald's biography by J. Wolfowitz (AoMS, 1952)

Wald's sequential probability ratio test is a **simple** and effective solution.

The in-context tradition that our class follows:

*Start with a **“practical problem”**,*

*and develop **“a true feeling for, and insight into it”***

*before bringing in **“a high level of mathematical talent of the most abstract sort”**.*

This approach led to Wald’s sequential probability ratio test, a **simple** and effective solution.

How does statistics thrive?

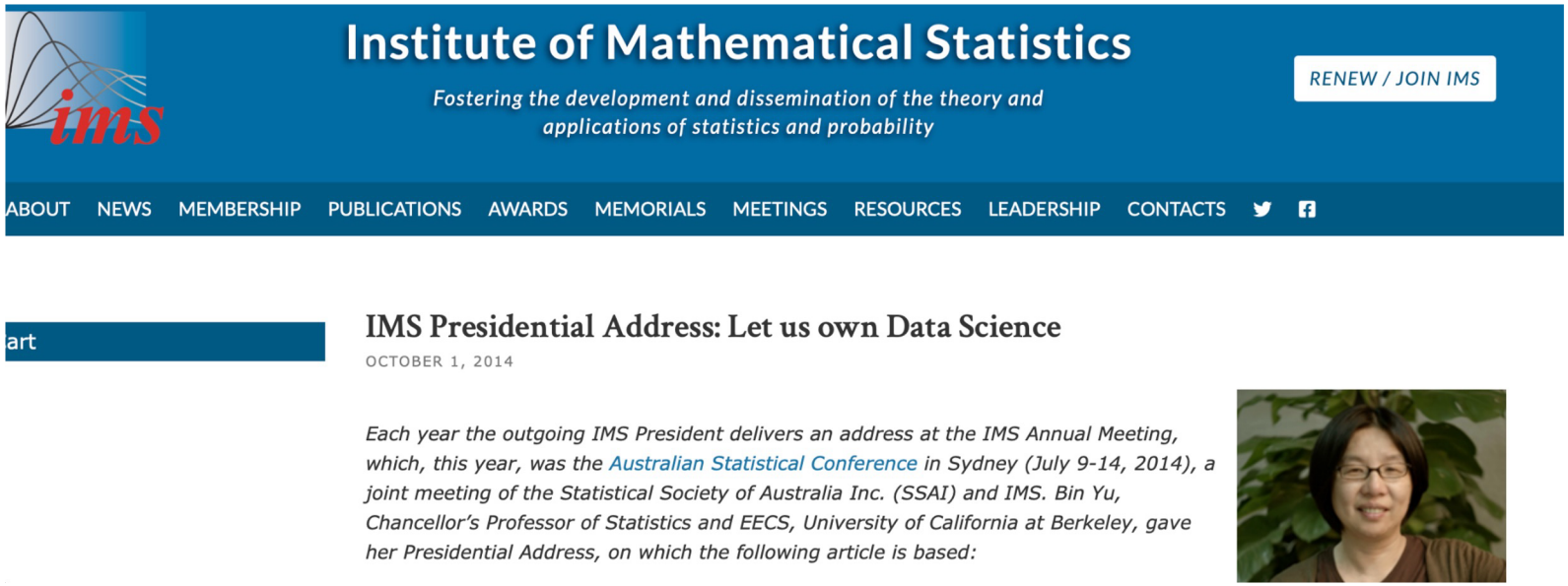
How does statistics thrive?

"According to Darwin's *Origin of Species*, **it is not the most intellectual of the species that survives; it is not the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself.**"

– Megginson, L. C. (1963)

Statistics thrives by adapting to the changing environment…

2014 IMS Presidential Address:
"Let Us Own Data Science"



https://imstat.org/2014/10/01/ims-presidential-address-let-us-own-data-science/

IMS-MSR Data Science Conference in 2015
IMS Data Science Conference in 2018
ICSDS in 2022, 2023, 2024, 2025, …

Veridical Data Science

… is Data Science done right

How does learning happen in 214?

(I'd appreciate English being used related to the class.)

- **Lectures: by Bin (concepts, methods, algorithms, critical thinking), guest lectures, (case studies in industry), and student presentations (on classical and current papers)**

Three talks from Nvidia, Google, and MA alums (on careers) committed
Plan is to have a panel on finance.

- **Labs: computing skills, critical thinking, discussions on lecture materials**
- **Office hours by Bin and GSIs**
- **Assignments**

Active participation is key for this class

- Speak one's mind in a respectful way
- Be an active listener

Goals for improved communication:

Speakers get better at listening; listeners get better at speaking.

Good writers or coders become better at editing or code-review; good editors or code reviewer become better at writing or coding (maybe with AI tools' help)

Finding common grounds and embracing differences

Modern data problems are extremely complex and need multi-disciplinary teams to solve.

To allow everyone a voice on a team, what are your suggestions?

- How to get your voice heard
- How to be aware of quiet people and invite them in
- How to push back on people who take much “space” or be an ally

Assignments:

no late submissions in general, except under special circumstances.

- **Weekly VDS book (and sometimes paper) reading and selected problems**

Assigned every Tuesday in class and reading summary and problem solutions due next Tuesday 11:59 pm on gradescope).

Selected problems are T/F and conceptual – Stat 230 by S. Pimentel will use some of the coding the project problems in the VDS book.

Assignments (cont):

no late submissions in general, except under special circumstances.

Three data labs and lab 0 (not graded) (in Python) (PCS overlay)
Computing on SCF and NSF Access Platform (GPUs)

Lab 1: data cleaning and medical prediction using non-DL supervised learning methods of your choice (single-author) (**Zach** leads)

Lab 2: Remote sensing cloud detection problem, EDA, unsupervised and supervised learning, ML and deep learning (representation) (**Anqi** leads)

Lab 3 (final project): Text-fMRI problem to understand the brain (LLMs and ML and interpretation) (**Sean and Sequoia** lead)

Academic integrity

- **Why is it important?**
- **How do we cultivate integrity, honesty, and fairness in the class?**

Every lab report will be turned in with statements from students on their contributions and about whether and how they used AI tools such as chatGPT.

How to use AI tools such as ChatGPT?

Potential harm of inappropriate use: students stop learning writing and coding skills that are essential for their future jobs.

Example: surgeons who are trained with telesurgery technologies are not as good with their hands as the ones trained without.

- For writing, use it as a paid editor (cost is on our environment)
- For coding, use it as a paid editor (e.g. Claude?).

Should we use it more? How?

Attendance policy

Email notices to Zach and Bin are required for missing lectures or discussion sessions.

Attendances will be taken at lectures and discussion sessions.

No exams.

More logistics details in the handout

Questions?

Real world is messy, but symbols are clean

- This class bridges between the messy and the clean – the goal is to help you impact the world
- It differs in many ways from a traditional math/methods class

1. The organization is unique – reflecting practice
2. Students are active partners

Pedagogy aim for this class:
“learn how to learn in context”

- Symbols have meanings in context just like in art

“Untitled” (1989-1990)

Doris Salcedo (born 1958)
Colombian-born sculptor who lives and works in Bogotá.

.
http://www.nytimes.com/2015/02/15/arts/design/doris-salcedo-whose-art-honors-lives-lost-gets-a-retrospective-in-chicago.html?_r=0

Doris Salcedo

- “The Colombian sculptor's work is about mourning. Influenced by the horrific violence she observed throughout the world, but especially in her native Bogotá, Colombia, Salcedo wanted to find a way to bring humanity to the losses. She worries that society has become hardened to violence and that victims have become mere statistics or headlines.”

- “These eleven "Untitled" sculptures (1989-90), composed of white cotton shirts in plaster and impaled by steel rebar, are Salcedo's response to two 1988 massacres that took place on banana plantations in La Negra and La Honduras. The shirts represent the standard dress of the plantation workers while alluding to the absence of their bodies as well as the funerary dress for the dead”

http://www.chicagonow.com/show-me-chicago/2015/02/who-is-doris-salcedo/

Our stats/ML work impacts real world

- Reality

What we do have to take into account reality through narratives, codes, algorithms and visualization. We also have to connect with the audience

- Responsibility

There are consequences that could lift the world and those that could harm

- Professional responsibility/ethics and critical thinking to get there

Honesty (no copying), reproducibility, fairness to others and oneself (fair share of credits in your lab reports)

What is professional responsibility to you?

Week 1 Assignments

(due Tuesday 1/27 at 11:59 pm on gradescope)

1. Read and summarize without using genAI

First 4 chapters of VDS book at vdsbook.com and
Box (1976): Science and statistics; https://www.jstor.org/stable/2286841

Extra reading: no summary needed
A conversation with George Box by M. DeGroot (1987)
https://projecteuclid.org/journals/statistical-science/volume-2/issue-3/A-Conversation-with-George-Box/10.1214/ss/1177013223.full

2. Selected problems from VDS book:
All T/F problems in Ch. 1-4
Conceptual problems:
Ch. 1, Ex. 10; Ch. 2 Ex. 10-11; Ch. 3 Ex. 11; Ch. 4 Ex. 14.